

สรุปรายงานการฝึกอบรมเชิงปฏิบัติการ
การวัดและประเมินผลให้ถูกต้องและมีประสิทธิภาพ
(Assessment workshop for clinical teachers)

1. เนื้อหา ความรู้ ทักษะ ที่ได้เรียนรู้จากการฝึกอบรมเชิง

การประเมินผล (Assessment) เป็นกระบวนการบันทึกความรู้ ทักษะ ทศนคติ ความเชื่อ โดยปกติสามารถวัดได้หลากหลาย ประกอบด้วย

- (1) Placement** เพื่อพิจารณาความพร้อมของผู้เรียน สำหรับการเรียนการสอนที่วางแผนไว้
- (2) Formative** เพื่อให้ข้อเสนอแนะแก่ผู้เรียน เกี่ยวกับความสำเร็จและการพัฒนาปรับปรุงการเรียนรู้ต่าง ๆ เป็นระยะ ๆ ระหว่างที่ทำการเรียนการสอน จากงานวิจัยจึงควรประเมิน Formative บ่อย ๆ ทดสอบไว้ ประเมินไว้ ประเมินบ่อย เห็นข้อบกพร่อง จะช่วยให้นักเรียนเกิดการเรียนรู้ได้ดี
- (3) Summative** เพื่อพิจารณาถึงระดับที่บรรลุเป้าหมายการเรียนการสอน โดยส่วนใหญ่ใช้เพื่อกำหนดเกรด เพื่อเป็นการตัดสินผลว่าที่เรียนมานั้นผ่านหรือไม่ผ่าน การทดสอบบ่อย ๆ จะทำให้การเรียนรู้มีประสิทธิภาพมากขึ้น

Criteria for Good Assessment

(1) Validity การวัดที่มีความตรง จึงต้องแสดงหลักฐานและทฤษฎีสนับสนุนการตีความคะแนนการทดสอบที่เกี่ยวข้อง และผลลัพธ์การเรียนรู้ให้สอดคล้องกับสิ่งที่ต้องประเมิน สิ่งอื่นไม่พึงประสงค์ในการสอบ (Validity threats) ที่เราต้องคำนึงถึงในการจัดสอบ Validity นั้นคือการประเมินคุณค่าของการแปลผลและการนำผลสอบไปใช้ ดังนั้นสิ่งอื่นไม่พึงประสงค์ในการสอบ หรือ validity threats ก็คือสิ่งใดก็ตามที่เข้ามารบกวนการแปลผลสอบสิ่งรบกวนเหล่านี้แยกได้เป็น 2 ปัจจัยหลัก คือ construct underrepresentation และ construct-irrelevant variance

Construct underrepresentation หมายถึง การประเมินผลที่ไม่ครอบคลุมสิ่งที่ต้องการวัดอย่างเพียงพอ ทำให้ผลการสอบไม่สามารถบ่งบอกถึงความสามารถของผู้เรียน ผู้สอบ ในเรื่องที่ต้องการวัดผลอย่างครบถ้วน เช่น การสอบ OSCE เพื่อวัดความสามารถของแพทย์ประจำบ้านในการให้คำแนะนำปรึกษาแก่ผู้ป่วย หากเกณฑ์การให้คะแนนมีเพียงหัวข้อที่เกี่ยวกับการพูดกับผู้ป่วย แต่ไม่มีหัวข้อที่เกี่ยวกับการใช้ อวัยวะ ภาษา เช่น การใช้ท่าทาง น้ำเสียง การรับฟังปัญหา เป็นต้น ก็จัดว่าทำการประเมินไม่ครอบคลุมเนื้อหา ผลการประเมินก็นำไปใช้บอกได้เพียงว่าแพทย์ประจำบ้านให้ข้อมูลผู้ป่วยครบถ้วน แต่ไม่สามารถบอกได้ว่าแพทย์ประจำบ้านทำการสื่อสารกับผู้ป่วยได้ดีในทุกด้าน ในการสอบข้อเขียนสำหรับวัดความรู้ของผู้เรียน หากใช้ข้อสอบที่สั้นเกินไป มีจำนวนข้อสอบไม่กี่ข้อ ก็จะมีปัญหาที่ไม่สามารถวัดความรู้ของผู้เรียนได้ครอบคลุมเนื้อหาที่ต้องการวัดผล

Construct-irrelevant variance หมายถึง ปัจจัยอื่นที่นอกเหนือไปจากความรู้ความสามารถของผู้เรียนที่สามารถส่งผลต่อ คะแนนสอบของผู้เรียนได้ ปัจจัยที่อาจรบกวนคะแนนสอบ multiple-choice examination ได้แก่

- ข้อสอบที่ไม่มีคุณภาพ โจทย์คำถามกำกวม มีตัวเลือกที่ถูกมากกว่า 1 ตัวเลือก ทำให้ผู้เรียนที่มีความรู้ตอบผิด หรือโจทย์ คำถามบอกใบ้ให้นักเรียนตอบถูก โดยไม่ต้องใช้ความรู้ ข้อสอบเก่าที่รั่วไหลออกจากรั้วข้อสอบทำให้ผู้เรียนที่รู้ข้อสอบ มาก่อนสามารถตอบได้โดยไม่ต้องคิด
- ผู้เรียนที่ทุจริตในการสอบ ลอกข้อสอบของเพื่อน หรือใช้วิธีการอื่น ๆ ในการได้มาซึ่งคำตอบโดยที่ไม่ได้ใช้ความรู้ในเรื่องที่ทำการสอบ
- อาจารย์ที่บอกข้อสอบให้ผู้เรียนในการสอนทำให้ผู้เรียนที่ท่องคำตอบเข้าไปสอบทำข้อสอบได้โดยไม่ต้องคิด

สำหรับการสอบในรูปแบบอื่นที่ต้องใช้กรรมการให้คะแนน เช่น OSCE การสอบข้อสอบบรรยายหรือ การสอบปากเปล่า นั้นจะมีปัจจัยที่เกี่ยวข้องกับกรรมการผู้ให้คะแนนเข้ามารับทวนการแปลผลคะแนนสอบได้ด้วย เช่น

- ความไม่เสมอภาคของอาจารย์ในเกณฑ์การให้คะแนน ผู้เรียนที่สอบกับอาจารย์ที่กดคะแนน เสียเปรียบผู้เรียนที่สอบ กับอาจารย์ที่ใจดี และปล่อยคะแนน

- ความไม่สม่ำเสมอของอาจารย์ในการให้คะแนน อาจารย์บางท่านมีแนวโน้มจะให้คะแนนต่ำลงในกลุ่มผู้เรียนที่สอบตอนท้าย เนื่องด้วยความเหนื่อยล้า ในขณะที่อาจารย์บางท่านมีแนวโน้มจะให้คะแนนสูงขึ้นในตอนท้ายของการสอบ เนื่องจากได้เห็นความสามารถของผู้เรียนจำนวนหนึ่ง แล้วพบว่าเกณฑ์ที่ตั้งเป้าไว้นั้นสูง เกินความสามารถของผู้เรียน ส่วนใหญ่จึงปรับเกณฑ์การให้คะแนนให้ง่ายขึ้น ทำให้นักเรียนในกลุ่มหลังได้คะแนนง่ายขึ้น

- การจำกัดช่วงของคะแนนที่พบบ่อย คืออาจารย์บางท่านนิยมเดินสายกลางไม่ว่าผู้เรียนจะทำดีมาก หรือน้อยเพียงใดก็มักจะให้คะแนนอยู่ในเกณฑ์ปานกลาง ไม่กล้าให้คะแนน 0 ในรายที่ทำไม่ดี แต่ก็ไม่กล้าให้คะแนนเต็มในผู้เรียนที่ทำได้ดี

ปัจจัยต่าง ๆ เหล่านี้เป็นสิ่งที่ผู้จัดสอบต้องคำนึงถึงเสมอในการจัดสอบและตั้งมาตรการเพื่อควบคุมและกำจัดปัจจัยรบกวน เหล่านี้จากการสอบ เพื่อให้ได้ผลการสอบที่มีความเที่ยงตรง เป็นธรรม และสามารถใช้ออกความรู้ ความสามารถของผู้เรียนได้ตามที่ต้องการ

(2) Reliability

การประเมินความแม่นยำของคะแนนสอบ (Reliability) การตรวจสอบความแม่นยำของคะแนนสอบเป็นการตอบคำถามว่า หากทำการสอบซ้ำผู้เรียนจะได้คะแนนเท่าเดิมหรือไม่ ในการสอบทั่วไปมักรายงานความแม่นยำของคะแนนสอบด้วยค่า reliability coefficient ซึ่งมีค่าได้ตั้งแต่ 0 – 1 โดยค่ายิ่งสูง บ่งบอกว่าผลสอบมีความน่าเชื่อถือมาก ค่า reliability coefficient = 0 บอกละเอียดคะแนนสอบที่ขาดความแม่นยำโดยสิ้นเชิง เทียบได้กับการให้คะแนนผู้เรียนโดยการสุ่ม

ตัวเลขให้ส่วนค่า reliability coefficient = 1 บอกละเอียดคะแนนสอบที่มีความแม่นยำมาก หากให้ผู้เรียนสอบซ้ำก็จะได้คะแนนเท่าเดิมเพื่อยืนยันความเข้าใจ คุณลักษณะที่สำคัญของ reliability ได้แก่

1. Reliability เป็นคุณสมบัติของคะแนนสอบ ไม่ใช่ตัวข้อสอบ ข้อสอบชุดหนึ่งทำการสอบกับผู้เรียนกลุ่มหนึ่ง พบว่ามีความแม่นยำสูง แต่เมื่อเอาข้อสอบชุดเดียวกันไปทำการสอบผู้เรียนอีกกลุ่มหนึ่ง อาจมีความแม่นยำต่ำได้

2. Reliability มีด้วยกันหลายชนิด และค่า reliability coefficient ที่ได้จากการประเมินความแม่นยำแต่ละชนิดก็แปลผลแตกต่างกัน การประเมินความแม่นยำของคะแนนสอบเป็นการตรวจสอบว่าหากทำการสอบซ้ำจะได้คะแนนเท่าเดิมหรือไม่ ประเด็นสำคัญคือเราจะทำการสอบซ้ำอย่างไรจะสอบซ้ำด้วยข้อสอบชุดเดิมหรือข้อสอบชุดใหม่ที่ออกแบบให้ เปรียบเทียบได้กับข้อสอบชุดเดิม สอบซ้ำ ณ เวลาเดียวกันหรือใกล้เคียงกัน หรือเวลาห่างกันเป็นสัปดาห์ สอบซ้ำ โดยใช้กรรมการให้คะแนนคนเดิม หรือสอบซ้ำโดยเปลี่ยนกรรมการให้คะแนน จะเห็นได้ว่า วิธีการสอบซ้ำต่างกันบอกความแม่นยำของคะแนนในสถานการณ์ต่างกัน (ความแม่นยำเมื่อเปลี่ยนชุดข้อสอบ หรือความแม่นยำเมื่อเปลี่ยนเวลา หรือ ความแม่นยำเมื่อเปลี่ยนกรรมการให้คะแนน) ดังนั้นการแปลผลของค่า reliability coefficient ต้องทำความเข้าใจว่าค่าดังกล่าวบอกละเอียดความแม่นยำชนิดใด โดยทั่วไปในการวัดความแม่นยำของคะแนนสอบ multiple-choice examination จากการสอบครั้งเดียว มักเป็นการประเมิน internal consistency reliability ซึ่งบ่งบอกว่าข้อสอบทุกข้อที่ใช้ในการสอบผู้เรียนกลุ่มหนึ่ง ๆ ทำการวัดความรู้ในเรื่องเดียวกันหรือไม่

3. Reliability เป็นปัจจัยที่สำคัญเพียงปัจจัยหนึ่งในการประเมินคุณค่าของผลสอบ ผลสอบที่ไม่มีความแม่นยำนั้นเป็นผลสอบที่มีคุณค่าต่ำไม่สามารถให้ข้อมูลที่ เป็นประโยชน์เกี่ยวกับผู้เรียน ผู้สอบได้ แต่ผลสอบที่มีความแม่นยำสูงนั้นก็ไม่ว่าจะเป็นว่า จะเป็นผลสอบที่เราสามารถนำไปใช้ประโยชน์ได้เสมอไป จำเป็นต้องพิจารณาปัจจัยร่วมอื่น ๆ อีกหลายอย่าง เช่น หากมีผู้เรียนทุจริตในการสอบ คะแนนสอบที่ได้ก็อาจมีค่า

reliability coefficient สูง แต่ผลสอบนั้นก็จะเป็นผลสอบที่บิดเบือน ไม่สามารถบอกได้ว่าผู้เรียนที่ได้คะแนนสูง เป็นผู้เรียนที่มีความรู้ หรือเป็นผู้เรียนที่ไม่มีความรู้แต่ลอกข้อสอบเพื่อน

ประเด็นที่ได้รับความสนใจ คือค่า reliability coefficient ต้องสูงเพียงใดจึงจะเพียงพอที่จะนำผลสอบไปใช้ได้ โดยทั่วไปนั้นจำเป็นต้องพิจารณาควบคู่ไปกับการนำผลสอบไปใช้ หากผลสอบนั้นนำไปใช้ในการตัดสินใจที่สำคัญ เมื่อตัดสินใจไปแล้วผลเป็นที่สุดไม่สามารถเปลี่ยนแปลงได้ และส่งผลยาวนาน โดยเฉพาะการตัดสินใจที่ส่งผลกระทบต่อตัวบุคคล มักต้องการคะแนนสอบที่มีค่า reliability coefficient สูงมาก ในทางกลับกัน หากผลสอบนั้น ใช้ในการตัดสินใจที่ไม่ค่อยสำคัญ มีผลระยะสั้นและการตัดสินใจสามารถเปลี่ยนแปลงได้หลัง จากการสอบนี้ โดยพิจารณาจากการสอบอื่นที่จะจัดตามมาภายหลัง โดยเฉพาะการตัดสินใจที่มีผลต่อผู้เรียนเป็นกลุ่ม ไม่ส่งผลกระทบต่อตัวบุคคลมักไม่ต้องการค่า reliability coefficient ที่สูงมาก โดยทั่วไปสำหรับการสอบย่อย ๆ ในชั้นเรียน ควรให้ค่า reliability coefficient สูงกว่า 0.7 สำหรับการสอบลงกองของนักศึกษาแพทย์ การสอบปลายภาค หรือการสอบใหญ่ต่าง ๆ ในโรงเรียนแพทย์ควรให้ค่า reliability coefficient สูงกว่า 0.8 สำหรับการสอบที่มีความสำคัญมากเช่นการสอบคัดเลือกเข้าเรียนมหาวิทยาลัยการสอบใบอนุญาตประกอบวิชาชีพเวชกรรมการสอบวุฒิบัตรผู้เชี่ยวชาญเฉพาะทางมักต้องให้ reliability coefficient สูงกว่า 0.9

- **Classical Test Theory** โดยพิจารณาจากสูตร

$$\text{True score (T)} = \text{Observe score (O)} + \text{Error (e)}$$

Error มีสาเหตุได้จาก systematic error และ random error โดยผลกระทบคะแนนของผู้เรียนที่ไม่สามารถคาดเดาได้ อาจเกิดจากการจำเนื้อหาไม่ได้ ความแตกต่างของแรงจูงใจ สมาธิ ความประมาท โชคในการเดา เป็นต้น

- **Reliability of Test Score** โดยพิจารณาจาก

- ค่าสัมประสิทธิ์ความน่าเชื่อถือ / ดัชนีความน่าเชื่อถือ
- ระบุความสอดคล้องของคะแนนการทดสอบจากการวัดผลครั้งหนึ่งไปสู่อีกครั้งหนึ่ง
- ช่วงคะแนน คือ 0-1
- หากค่าที่ได้มีค่าสูง: คะแนนการทดสอบมีความสอดคล้องกันสูง

- **Internal Consistency Reliability**

- Split-half method โดย $\text{Reliability} = 2r / 1+r$ ซึ่ง $r = \text{Reliability for half test}$
- Cronbach's Alpha คือ ค่าเฉลี่ยของค่าสัมประสิทธิ์แบบแบ่งครึ่งทั้งหมดเมื่อแบ่ง

การทดสอบในทุกรูปแบบที่เป็นไปได้ โดยแบ่งช่วงคะแนน ดังนี้ ระดับสูง (0.9 - สูงกว่า) ระดับกลาง (0.80 - 0.89) ระดับต่ำ (0.70-0.79)

- **Improving Reliability** ปัจจัยที่ส่งผลต่อค่า reliability coefficient สิ่งเหล่านี้มี

ความสำคัญมากเมื่อต้องการอธิบายว่าเหตุใดคะแนนสอบที่ได้จึงไม่แม่นยำ และเราต้องทำอย่างไรจึงจะทำให้คะแนนสอบมีความแม่นยำมากขึ้น โดยทั่วไปปัจจัยที่สำคัญที่ส่งผลต่อความแม่นยำของคะแนนสอบมี 4 ปัจจัย คือ

1. จำนวนข้อสอบ ถ้าทำการสอบด้วยข้อสอบที่สั้น ประกอบด้วยคำถามไม่กี่ข้อ คะแนนสอบที่ได้มัก ไม่แม่นยำ วิธีเพิ่ม ความแม่นยำของคะแนนสอบที่ง่ายที่สุด คือการเพิ่มจำนวนข้อสอบ

2. การกระจายตัวของคะแนนสอบ ถ้าคะแนนสอบมีความแตกต่างกันมาก มีทั้งผู้เรียนที่ทำคะแนนได้สูง และผู้เรียนที่ทำคะแนนได้ต่ำ คะแนนสอบมักมีความแม่นยำสูง ในทางตรงข้ามหากผู้เรียนทำคะแนนใกล้เคียงกันคะแนน เกาะกลุ่มกันมาก คะแนนสอบมักมีความแม่นยำต่ำ วิธีการเพิ่มความแม่นยำของคะแนนสอบโดยการเพิ่มการกระจายตัวของคะแนนของผู้เรียนทำได้โดยใช้ข้อสอบที่มีความยากมากขึ้น

3. ปัจจัยรบกวนการสอบของผู้เรียน หากทำการจัดสอบไม่ดี มีสิ่งรบกวนผู้เรียนในขณะที่ทำการสอบ (เช่น เสียงดังรบกวน ห้องสอบร้อนอบอ้าวจนผู้เรียนไม่มีสมาธิ) คะแนนสอบมักมีความแม่นยำต่ำ ดังนั้น ผู้คุมสอบต้องจัดสถานที่สอบให้ดี เพื่อให้ผู้เรียนมีสมาธิในการทำข้อสอบ ซึ่งจะนำไปสู่คะแนนสอบมีความแม่นยำสูง

4. ลักษณะการให้คะแนนของข้อสอบ ข้อสอบที่ไม่ต้องใช้กรรมการตรวจ เช่น multiple-choice examination มักให้คะแนนที่มีความแม่นยำสูง ในทางตรงข้ามข้อสอบที่ต้องใช้กรรมการให้คะแนนเช่น ข้อสอบบรรยายข้อสอบ OSCE คะแนนที่ได้มักมีความแม่นยำไม่สูงนักเนื่องจากมีปัจจัยที่นอกเหนือไปจากความสามารถของผู้เรียน (เช่น ความเหนื่อยล้าของกรรมการ ความไม่สม่ำเสมอของการใช้เกณฑ์ให้คะแนน หรือ อารมณ์ของกรรมการตรวจข้อสอบ) เข้ามาส่งผลต่อคะแนนสอบ

(3) Equivalence การทดสอบหัวข้อเดียวกันกับผู้เรียนระดับชั้นเรียนเดียวกัน ที่ทดสอบกันต่างเวลาได้คะแนนที่เทียบเคียงกันได้

(4) Feasibility การประเมินจะต้องมีความเป็นรูปธรรม สมจริง และสมเหตุสมผล โดยต้องอยู่ในบริบทที่เหมาะสม ทั้งในด้านเวลา การเงิน ความเชี่ยวชาญ และการบริหาร เป็นต้น

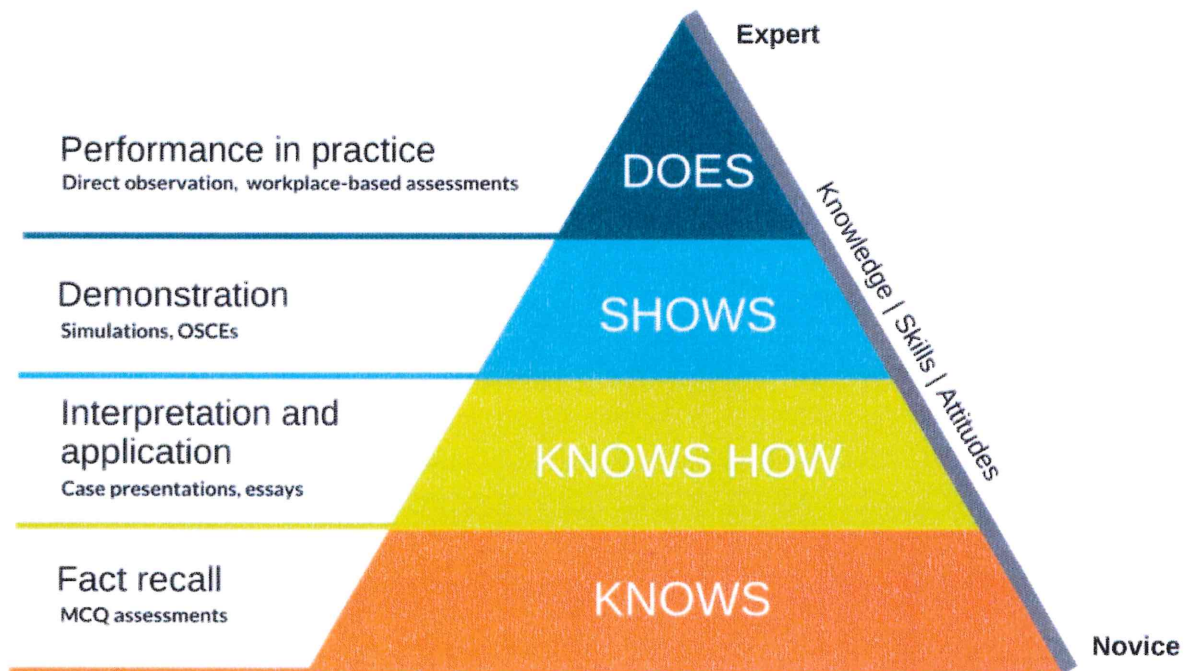
(5) Educational Effect การประเมินผลนั้นกระตุ้นให้ผู้เรียนมีการเรียนรู้ในเรื่องที่ควรเรียนรู้

(6) Catalytic Effect การประเมินผลก่อให้เกิดการนำผลของการสอบไปใช้ให้ feedback เพื่อสร้าง หรือส่งเสริม หรือสนับสนุนการเรียนรู้ของผู้เรียน

(7) Acceptability ผู้เกี่ยวข้อง (Stakeholders) ทั้งหมดยอมรับกับกระบวนการประเมินและเชื่อถือผลการตัดสิน

แนวทางการประเมินที่มีประสิทธิผล

1. การประเมินที่มีประสิทธิผลต้องอาศัยแนวคิดที่ชัดเจนเกี่ยวกับผลลัพธ์การเรียนรู้ที่ตั้งใจไว้ทั้งหมด
2. การประเมินที่มีประสิทธิผลต้องอาศัยขั้นตอนการประเมินที่หลากหลาย
3. การประเมินที่มีประสิทธิผลต้องคำนึงถึงความเกี่ยวข้องของขั้นตอนในการสอน
4. การประเมินผลที่มีประสิทธิผลต้องมีตัวอย่างผลการปฏิบัติงานของนักเรียนที่เพียงพอ
5. การประเมินผลที่มีประสิทธิผลต้องมีขั้นตอนที่ยุติธรรมต่อทุกคน
6. การประเมินผลที่มีประสิทธิผลต้องมีการกำหนดเกณฑ์สำหรับการตัดสินผลการปฏิบัติงานที่ประสบความสำเร็จ
7. การประเมินผลที่มีประสิทธิผลต้องอาศัยข้อเสนอแนะต่อนักศึกษาที่เน้นย้ำจุดแข็งของการปฏิบัติงานและจุดอ่อนที่ต้องแก้ไข
8. การประเมินผลที่มีประสิทธิผลต้องได้รับการสนับสนุนจากระบบการให้คะแนนและการรายงานที่ครอบคลุม



การสร้างข้อสอบปรนัย (multiple-choice question)

รูปแบบพื้นฐานของข้อสอบปรนัย

ข้อสอบปรนัย คือข้อสอบชนิดที่มีคำถามแล้ว มีตัวเลือกให้ผู้สอบเลือกตัวเลือกที่เหมาะสมเพื่อตอบคำถามดังกล่าว ข้อสอบปรนัยสามารถแบ่งออกได้เป็น 2 รูปแบบ ได้แก่

1. ข้อสอบถูกผิด (True/false item)

ข้อสอบประเภทนี้จะมีข้อความให้ผู้สอบ พิจารณาว่าถูกหรือผิด ผู้สอบตัดสินใจว่าข้อความแต่ละข้อถูกหรือผิดโดยไม่เกี่ยวข้องกับข้อความในข้ออื่น ต่อมาผู้พัฒนาข้อสอบเป็นชุดข้อความ (multiple true/false หรือ K-type item) โดยในแต่ละข้อจะมีข้อความ ผู้สอบต้องพิจารณาว่าแต่ละข้อความถูกหรือผิด แล้วทำการเลือกตัวเลือกที่บรรยายจำนวนข้อความ ที่ถูกต้องได้อย่างเหมาะสม

2. ข้อสอบเลือกคำตอบที่ถูกต้องที่สุด (one best response item) ข้อสอบประเภทนี้จะมีคำถามแล้วตามด้วย ตัวเลือกจำนวนหนึ่งให้ผู้สอบเลือกตัวเลือกที่เหมาะสมที่สุดเป็นคำตอบ ข้อสอบประเภทนี้เป็นที่นิยมกันมากที่สุด คือข้อสอบที่มีตัวเลือก 4 - 5 ตัวเลือก (A-type) แต่นอกจากข้อสอบมาตรฐานแล้วก็มีผู้ใช้ข้อสอบประเภทที่มีลักษณะเป็นการจับคู่ (extended matching item) โดยให้ผู้สอบเลือกตัวเลือกที่เหมาะสม (จากตัวเลือกจำนวน มาก 8 - 20 ตัวเลือก) ไปจับคู่กับโจทย์ (stem) ซึ่งมีหลายข้อ เช่น จับคู่ระหว่างคำบรรยายอาการของผู้ป่วยจำนวน 5 - 10 ราย กับการวินิจฉัยโรคที่เหมาะสม จำนวน 15 โรค เป็นต้น

องค์ประกอบของข้อสอบปรนัยชนิดเลือกคำตอบที่ถูกต้องที่สุด ข้อสอบปรนัยแต่ละข้อมีส่วนประกอบสำคัญ 2 ส่วน คือ

1. **โจทย์ (stem)** เป็นข้อมูลของโรค หรือภาวะ หรือผู้ป่วยตามด้วยคำถามหรือเว้นช่องว่างสำหรับเติมคำ หรือข้อความที่เหมาะสมลงไป

2. **ตัวเลือก (options)** คือคำ หรือข้อความที่ผู้ออกข้อสอบนำเสนอตามหลังจากโจทย์เพื่อให้ผู้สอบเลือกไปใช้ตอบคำถาม หรือเติมลงในช่องว่างในโจทย์

2.1 ตัวเลือกที่ถูกต้อง (correct option) เป็น คำตอบที่ถูกต้องมีเพียงตัวเลือกเดียวต่อข้อสอบข้อหนึ่ง

2.2 ตัวลวง (distractors) เป็นคำตอบที่ผิด หรือไม่เหมาะสม มีไว้ลวงให้ผู้สอบที่ไม่มีความรู้ หรือมีความเข้าใจไม่ถูกต้องในเนื้อหาที่นำมาออกข้อสอบเลือกตอบ ตัวลวงไม่จำเป็นต้องเป็นคำตอบที่ผิดชัดเจนเสมอไป ตัวลวงที่ดีมักมีส่วนถูกบ้าง แต่มีระดับของความถูกต้อง เหมาะสมน้อยกว่าคำตอบที่ถูก

ข้อเสนอแนะพื้นฐานของการเขียนข้อสอบปรนัย

1. เนื้อหาข้อสอบ

- 1.1 ข้อสอบหนึ่งข้อควรมุ่งเน้นประเมินความรู้เพียงเรื่องเดียว
- 1.2 หลีกเลี่ยงการถามความรู้ในรายละเอียดปลีกย่อยที่ไม่มีที่ใช้ทางคลินิก (trivial content)
- 1.3 หลีกเลี่ยงการถามความรู้ในเรื่องที่ยังมีความขัดแย้งกันในแนวปฏิบัติ
- 1.4 หลีกเลี่ยงการลอกประโยคหรือข้อความจากตำราโดยตรง
- 1.5 หลีกเลี่ยงการนำเสนอข้อสอบที่ประเมินความรู้ในเรื่องเดียวกันสองข้อในข้อสอบชุดเดียวกัน

2. การจัดรูปแบบข้อสอบ

- 2.1 เลือกใช้คำศัพท์หรือรูปประโยคที่ง่ายต่อการทำความเข้าใจ
- 2.2 หลีกเลี่ยงการนำเสนอข้อมูลที่ไม่เกี่ยวข้องกับการแก้ปัญหาของโจทย์ข้อนั้น
- 2.3 จัดให้มีการตรวจสอบเนื้อหา คำศัพท์ และรูปประโยคที่ใช้ในข้อสอบแต่ละข้อก่อนนำไปใช้

3. การเขียนโจทย์

- 3.1 เขียนโจทย์ให้มีความชัดเจน ผู้สอบทุกคนอ่านแล้วมีความเข้าใจตรงกัน
- 3.2 เรียบเรียงเนื้อหาให้ใจความสำคัญของข้อสอบอยู่ในโจทย์
- 3.3 หลีกเลี่ยงการเขียนโจทย์ที่มีรูปประโยคเชิงปฏิเสธ

4. การเขียนตัวเลือก

- 4.1 เขียนตัวเลือกที่มีประสิทธิภาพให้มีจำนวนมากที่สุดเท่าที่เหมาะสมกับบริบท
- 4.2 จัดให้ตัวเลือกที่ถูกต้องมีการกระจาย ตำแหน่งไปให้มีจำนวนพอ ๆ กันในทุกตัวเลือก
- 4.3 เขียนตัวเลือกแต่ละข้อให้เป็นอิสระ ไม่ขึ้นต่อกัน
- 4.4 เขียนตัวเลือกให้ทุกตัวเลือกมีความเป็นเนื้อเดียวกัน (homogeneous)
- 4.5 เขียนตัวเลือกแต่ละข้อให้มีความยาวพอ ๆ กัน
- 4.6 หลีกเลี่ยงการใช้ตัวเลือก “ถูกทุกข้อ” หรือ “ไม่มีข้อใดถูก”

การวิเคราะห์ข้อสอบ

1. การวิเคราะห์ข้อสอบรายข้อ (item analysis) การวิเคราะห์ข้อสอบแต่ละข้อให้พิจารณา 3 ปัจจัย คือ

1.1 ความยากง่ายของข้อสอบ (item difficulty, p)

ความยากง่ายของข้อสอบวัดโดยใช้ค่า p (proportion of examinees answering items correctly) คือสัดส่วนของผู้สอบที่ตอบข้อสอบข้อนั้นถูก ซึ่งหาได้จากการนำจำนวนผู้สอบที่ตอบข้อสอบข้อนั้นถูก ต้องหารด้วยจำนวนผู้สอบที่ตอบข้อสอบข้อนั้นทั้งหมด หากข้อสอบข้อนั้นเป็นข้อสอบที่ง่าย ผู้สอบทุกคนตอบถูก ค่า p ก็จะเป็น 1 หากไม่มีผู้สอบคนใดตอบถูกเลย ข้อสอบข้อนั้นก็จะมีความยาก ค่า p เป็น 0 หากมีคนตอบถูก 70 % ข้อสอบข้อนั้นก็มีค่า p เท่ากับ 0.7 ข้อสอบที่ดีมากจะมี ค่า p อยู่ในช่วง 0.45 – 0.75 ข้อสอบที่ดีจะ

มีค่า p อยู่ในช่วง $0.76 - 0.91$ ข้อสอบที่พอใช้ได้มีค่า p อยู่ในช่วง $0.25 - 0.44$ ข้อสอบที่มีค่า p ต่ำกว่า 0.25 เป็นข้อสอบที่ยากเกินไป และข้อสอบที่มีค่า p สูงกว่า 0.91 เป็นข้อสอบที่ง่ายเกินไป

1.2 ความสามารถในการจำแนกผู้สอบตาม ระดับความสามารถ (item discrimination, r)

ความสามารถในการจำแนกผู้สอบ หมายถึง ความสามารถของข้อสอบข้อหนึ่ง ๆ ในการแยกผู้สอบที่ทำคะแนนได้ดีออกจากผู้สอบที่ทำคะแนนได้ไม่ดี ข้อสอบที่มีความสามารถในการแยกแยะได้ดีนั้นผู้สอบที่ตอบข้อสอบข้อนั้นถูกมักจะได้คะแนนสูง และผู้สอบที่ตอบข้อสอบข้อนั้นผิดมักจะได้คะแนนต่ำ ดัชนีที่ใช้วัดความสามารถในการจำแนกผู้สอบที่ใช้กันมากที่สุดในปัจจุบัน คือค่า point-biserial correlation ซึ่งนิยมใช้อักษรย่อเป็น r

ค่า point-biserial correlation ที่คำนวณได้นี้มี ค่าอยู่ในช่วง -1 ถึง 1 โดยค่าที่ติดลบ หมายถึง ข้อสอบข้อนั้นผู้ที่ตอบถูกมักสอบได้คะแนนรวมต่ำ แต่ผู้ที่ตอบผิดมักสอบได้คะแนนรวมสูง ในทางตรงข้าม หากค่า point-biserial ยิ่งสูง แสดงถึงข้อสอบที่มีความสามารถในการ แยกแยะดี ผู้ที่ตอบข้อสอบข้อนั้นถูกมักทำคะแนน รวมได้สูง ข้อสอบที่ดีควรมีค่า point-biserial สูงกว่า 0.20 ข้อสอบที่พอใช้ได้ควรมีค่า point-biserial อยู่ใน ช่วง $0.1 - 0.19$ ข้อสอบที่มีค่า point-biserial ต่ำกว่า 0.1 เป็นข้อสอบที่ไม่ดีนัก โดยเฉพาะอย่างยิ่งข้อสอบ ที่มีค่า point-biserial ต่ำกว่า 0 ไม่ควรนำมาคิดคะแนน (โดยทั่วไปแล้วข้อสอบที่มีค่า point-biserial ติดลบ ให้ สงสัยว่าจะเฉลยผิด)

1.3 ประสิทธิภาพของตัวลวง (distractor functionality) ตัวลวงที่มีประสิทธิภาพนั้นมีคุณสมบัติ

2 ประการ

(1) มีผู้สอบเลือกตัวลวงนั้นไม่ต่ำกว่าร้อยละ 5 ของจำนวนผู้สอบทั้งหมด

(2) มีค่า point-biserial correlation ของตัวลวง นั้นเป็นลบ กล่าวคือตัวลวงที่ดีจะลวงให้ผู้สอบที่มีความรู้ไม่ดี (มีคะแนนต่ำ) มาเลือก แต่ไม่ลวงให้ผู้สอบที่มีความรู้ดี (มีคะแนนสูง) มาเลือก หากตัวลวงใดมีค่า point-biserial correlation เป็นบวก ให้ทบทวนข้อสอบข้อนั้นดูว่าอาจจะเฉลยผิดหรือมีคำตอบที่ถูกต้องมากกว่า 1 ตัวเลือก

ตัวลวงใดที่มีผู้สอบเลือกน้อย หรือลวงให้ผู้ที่มีความรู้ดีมาเลือกจัดเป็นตัวลวงที่ไม่ดี สมควรพิจารณาตัดทิ้งหรือปรับเปลี่ยน

2. การวิเคราะห์ข้อสอบโดยรวม (test analysis)

การวิเคราะห์ข้อสอบโดยรวมเป็นการพิจารณา ว่าเมื่อข้อสอบทั้งชุดทำงานร่วมกันแล้วผลสอบที่ได้ ออก มาเป็นอย่างไร มีระดับความยากง่ายเป็นอย่างไร มีการกระจายตัวของคะแนนเป็นอย่างไร มีความน่าเชื่อถือ ของคะแนนสอบมากน้อยเพียงใด ดัชนีต่าง ๆ ที่ต้องพิจารณา ได้แก่

2.1 ความเที่ยงตรงของคะแนนสอบ (internal consistency reliability) การประเมินความ

เที่ยงตรงของคะแนนสอบ เป็นการตรวจสอบว่าคะแนนที่ได้ออกมานั้นมีความน่าเชื่อถือเพียงใด เป็นการตอบคำถามว่าหากนำผู้สอบ มาสอบใหม่ในสภาวะการเดิม ด้วยข้อสอบที่มีระดับความยากง่ายเท่าเดิม และผู้สอบมีความรู้เท่าเดิมไม่ได้ ไปศึกษาหาความรู้เพิ่มเติม จะได้คะแนนสอบเท่าเดิม หรือไม่ ดัชนีชี้วัดความเที่ยงตรงของคะแนนสอบที่ นิยมใช้ในการรายงานผลสอบด้วยข้อสอบปรนัยคือ ค่าสัมประสิทธิ์ อัลฟา (Coefficient Alpha) ค่าสัมประสิทธิ์อัลฟานี้มีค่าอยู่ในช่วง $0-1$ ค่าต่ำแสดงว่าคะแนนที่ได้มีความเชื่อถือได้น้อย ไม่แตกต่างไปจากการเดาสุ่ม ค่าสูงแสดงว่าคะแนนที่ได้นั้นมีความน่าเชื่อถือมาก หากทำการทดสอบซ้ำคะแนนที่ได้ ก็จะใกล้เคียงเดิม โดยทั่วไประดับของความเที่ยงตรง ของคะแนนสอบที่ยอมรับได้นั้นขึ้นกับว่าต้องการนำเอา คะแนนสอบไปใช้ทำอะไร หากการตัดสินผลสอบนั้นมีความสำคัญมาก (high-stakes examination) เช่น การ ตัดสินผลสอบขอรับใบประกอบวิชาชีพเวชกรรม หรือ ประกาศนียบัตรแพทย์ผู้เชี่ยวชาญเฉพาะสาขามัก ต้องการคะแนนสอบที่มีค่าสัมประสิทธิ์อัลฟา ไม่ต่ำกว่า 0.9 หากการตัดสินผลสอบนั้นมีความสำคัญปานกลาง (medium-stakes examination) เช่น การสอบลงกอง การสอบเลื่อนชั้นเรียน มักต้องการคะแนนสอบที่มีค่าสัมประสิทธิ์อัลฟา อยู่ในช่วง $0.8 - 0.89$ หากการตัดสินผลสอบนั้นมีความสำคัญน้อย (low-stakes examination) เช่น การสอบย่อยในชั้นเรียน การสอบแบบ formative assessment มักต้องการคะแนนสอบที่มีค่า สัมประสิทธิ์อัลฟา อยู่ในช่วง $0.7 - 0.79$

ประเด็นสำคัญที่ต้องพิจารณาคือเมื่อได้ คะแนนสอบที่มีค่าสัมประสิทธิ์อัลฟาต่ำ จะต้องดำเนินการอย่างไร เพื่อพัฒนาให้การสอบครั้งต่อไปไม่ประสบปัญหาเรื่องความไม่น่าเชื่อถือของคะแนนสอบอีก ปัจจัยหลักที่จะช่วยเพิ่มความเที่ยงตรงของคะแนนสอบปรนัยมี 3 ปัจจัย คือ

- (1) เพิ่มจำนวนข้อสอบให้มากขึ้น ยังมีข้อสอบ มากข้อคะแนนที่ได้ก็จะมีคะแนนที่เที่ยงตรงเพิ่มมากขึ้น
- (2) ปรับให้ข้อสอบมีการคละกันของข้อสอบที่ ยากและง่ายอย่างเหมาะสม เพื่อปรับให้คะแนนมีการกระจายตัวมากขึ้น หากข้อสอบทั้งชุดประกอบไปด้วย ข้อสอบที่ง่ายหมด ผู้สอบเกือบทั้งหมดได้คะแนนสูงมาก จะทำให้มีความแตกต่างของคะแนนน้อย โอกาสที่จะแยกแยะผู้สอบที่มีความรู้ดีออกจากผู้ที่มีความรู้ปานกลาง หรือไม่สู้ดีได้อย่างมั่นใจก็เป็นไปได้น้อย ดังนั้น หากอาจารย์ปรับให้มีการคละกันของข้อสอบยากและง่ายอย่างเหมาะสม ก็จะทำให้ผู้สอบมีระดับคะแนนแตก ต่างกันมาก ค่าสัมประสิทธิ์อัลฟาก็จะสูงขึ้นด้วย
- (3) ปรับสภาวะแวดล้อมของการสอบให้ เหมาะสม กำจัดสิ่งรบกวนสมาธิของผู้สอบให้มากที่สุด เช่น เสียงรบกวน แสงไฟที่ไม่เพียงพอ หรือไฟที่ติด ๆ ดับ ๆ เป็นต้น

2.2 การกระจายตัวของคะแนน และคะแนนเฉลี่ย (standard deviation and mean score)

การตรวจดูลักษณะพื้นฐานของคะแนนสอบนี้ จะช่วยบอกได้คร่าว ๆ ว่าการเรียนการสอนมีประสิทธิภาพ เพียงใด หากอาจารย์สอนได้ดี ผู้เรียนทั้งชั้นเรียน เข้าใจเนื้อหาดี คะแนนสอบที่ได้ออกมาก็ควรจะ ไม่ กระจายตัวมากนัก (คะแนนเกาะกลุ่มกัน) และคะแนนเฉลี่ยก็ควรจะค่อนข้างสูงเมื่อเทียบกับผู้เรียนรุ่นอื่น ๆ หากคะแนนสอบของผู้เรียนมีการกระจายตัวมากผิดปกติ แสดงว่าอาจมีปัญหาบางประการในการเรียน การ สอนทำให้นักเรียนบางคนมีความรู้ความเข้าใจดี แต่มีนักเรียนบางกลุ่มที่ไม่ค่อยรู้เรื่อง

2.3 ค่าความยากง่ายเฉลี่ยของข้อสอบ (average difficulty)

จากการวิเคราะห์ข้อสอบรายข้อ ได้ค่าความยากง่ายของข้อสอบแต่ละข้อ (p) เมื่อนำค่า p ของข้อสอบทุกข้อมาหาค่าเฉลี่ย ก็จะได้ค่าความยากง่าย ของข้อสอบทั้งชุด ค่าที่ได้มานี้ใช้เป็นดัชนีชี้วัดว่าข้อสอบ ทั้งชุดโดยรวมแล้วมีระดับความยากง่ายเป็นอย่างไร หากผู้สอบเป็นผู้เรียนกลุ่มใหญ่พอที่เราจะตั้งสมมติฐาน ว่าระดับความสามารถมีการกระจายตัวอย่างเหมาะสม และไม่ต่างจากระดับความสามารถเฉลี่ยของกลุ่มผู้สอบ ปีก่อน ๆ เราก็สามารถนำค่าความยากง่ายของข้อสอบ ทั้งชุดนี้มาเทียบได้ว่าข้อสอบที่นำมาใช้ในปีนี้น่ายาก หรือ ง่ายกว่าข้อสอบปีก่อน ๆ ซึ่งอาจารย์อาจนำข้อมูลนี้มา ใช้พิจารณาปรับเกณฑ์การตัดเกรดด้วยว่าต้องมีการปรับระดับคะแนนที่ได้เกรดต่าง ๆ หรือไม่ อย่างไร

2.4 ค่าความสามารถในการแยกแยะผู้สอบ เฉลี่ย (average discrimination)

การนำค่า point-biserial correlation ของข้อสอบทั้งชุดมาหาค่าเฉลี่ย เป็นการบอกคร่าว ๆ ว่า โดยรวม แล้วข้อสอบชุดนี้มีความสามารถในการแยกแยะผู้สอบ ตามระดับความสามารถเพียงใด ยิ่งได้ค่าสูงก็ยิ่งดี แต่มีข้อควรระวังในการแปลผลในกรณีที่การเรียนการสอน เป็นไปได้ดี และผู้สอบทั้งหมด หรือเกือบทั้งหมดทำคะแนนได้สูง ค่า point-biserial correlation เฉลี่ยของ ข้อสอบทั้งชุดจะไม่สูงแต่ไม่ได้แปลว่า ข้อสอบที่ใช้มี คุณภาพไม่ดี

การประเมินการปฏิบัติงานทางคลินิก (Clinical performance assessment)

1. การประเมินความสามารถในห้องสอบ (competence) การประเมินในกลุ่มนี้เป็นการประเมินที่อาจารย์จัดขึ้นในสถานการณ์ที่มีการควบคุมตัวแปรที่อาจส่งผลกระทบต่อความสามารถของผู้เรียนโดยมุ่งหวังจะวัดระดับความรู้ ความสามารถสูงสุด ที่ผู้เรียนมีโดยไม่มีปัจจัยอื่นมารบกวน มักเป็นการจัดสอบในห้องสอบโดยมีการแจ้งผู้สอบให้มีการเตรียมตัวมาสอบในหัวข้อที่กำหนด ในวันและเวลาที่กำหนด สิ่งที่ได้จัดเป็นระดับความสามารถสูงสุด ที่ผู้สอบสามารถแสดงออกมาได้ การสอบส่วนใหญ่ในโรงเรียนแพทย์จะเป็นการประเมินผลในกลุ่มนี้ เช่น การสอบข้อสอบข้อเขียน (ปรนัย หรือ อัตนัย) การสอบปากเปล่า (oral examination) การสอบ Objective Structured Clinical Examination (OSCE) เป็นต้น

2. การประเมินความสามารถในการปฏิบัติงานจริง (performance) การประเมินในกลุ่มนี้เป็นการประเมินจากการสังเกตการปฏิบัติงานของผู้เรียนในสถานการณ์จริง ซึ่งระดับความรู้ความสามารถที่ผู้สอบแสดงออกมาให้อาจารย์เห็น นั้นอาจมีปัจจัยรบกวนอื่น ๆ มาเกี่ยวข้องด้วย เช่น ระบบการทำงาน สภาพแวดล้อม สภาพความสัมพันธ์ระหว่างผู้สอบกับคนรอบข้าง สภาพจิตใจของผู้เข้าสอบ ฯลฯ สิ่งที่ได้วัดได้นั้น

อาจขาดความเป็นมาตรฐานเดียวกัน ระหว่างผู้สอบแต่ละคนไปบ้าง แต่สิ่งที่ประเมินได้จากการประเมินความสามารถในกลุ่มนี้น่าจะสอดคล้องกับระดับความรู้ ความสามารถที่ผู้เรียน ใช้ทำงานจริงในชีวิตประจำวันมากกว่า

ข้อพิจารณาในการประเมินผล

1. ความถูกต้อง (Validity) ความถูกต้องของผลการประเมิน หมายถึง ระดับคะแนนที่ได้นั้นแสดงถึงระดับความรู้ ความสามารถของผู้เรียน ที่อาจารย์ต้องการวัดผลจริง ๆ กล่าวคือผู้ที่ได้คะแนนสูงแสดงถึงระดับความรู้ ความสามารถที่สูง ในทางกลับกันผู้ที่ได้คะแนนต่ำ คือผู้ที่มีระดับความรู้ ความสามารถที่ต่ำหากมีปัจจัยอื่นใด ที่มีผลรบกวนการแปลผลดังกล่าว (validity threats) ก็จะลดระดับความถูกต้องของผลการประเมินลงตัวอย่างปัจจัย รบกวนความถูกต้องของการประเมินการปฏิบัติงาน เช่น ความแตกต่างในมาตรฐานการให้คะแนนของอาจารย์ ความแตกต่างกันของลักษณะผู้ป่วยที่ผู้เรียนแต่ละคน ดูแล เป็นต้น

2. ความเที่ยงตรง (Reliability) ความเที่ยงตรงของคะแนน หมายถึง หากนำผู้เรียนคนเดิมที่มีระดับ ความรู้ ความสามารถเท่าเดิม มาทำการประเมินผลซ้ำ คะแนนที่ได้จะมีค่าใกล้เคียงกันหรือไม่ ผลการสอบที่มีความเที่ยงสูง คือผลการสอบที่เมื่อสอบซ้ำ คะแนนก็จะเท่าเดิมหรือใกล้เคียงเดิม โดยทั่วไปแล้วเรารายงานความเที่ยงของคะแนนสอบเป็นตัวเลขมีค่า 0 – 1 โดยค่าดัชนีความเที่ยงที่ใกล้เคียงศูนย์บ่งชี้ถึง คะแนนสอบไม่ค่อยเที่ยง แต่หากค่าดัชนีความเที่ยงใกล้ 1 แสดงว่าคะแนนสอบมีความเที่ยงสูง การประเมินการปฏิบัติงานทางคลินิกโดยทั่วไปจัดเป็นการประเมินผลที่มีความสำคัญปานกลาง มักต้องการระดับความ เที่ยงตั้งแต่ 0.8 ขึ้นไป

3. ความเสมอภาค (Equivalence) ความเสมอภาคของการประเมินผล หมายถึง ผลการประเมินผู้เรียนในความรู้ หรือทักษะเดียวกันที่ทำใน ต่างวัน เวลา หรือสถานที่กัน สามารถนำมาเปรียบเทียบกันได้โดยไม่มีการได้เปรียบหรือเสียเปรียบกันเกิดขึ้น เช่น การสอบข้อเขียนวิชาเดียวกันของผู้เรียนที่ปฏิบัติงานกันคนละกลุ่ม สอบกันคนละวัน ก็ต้องมีมาตรการในการควบคุม ให้ข้อสอบดังกล่าวมีระดับความยากง่ายใกล้เคียงกัน ในการประเมินการปฏิบัติงานของผู้เรียน อาจารย์ก็ควรวาง ระบบให้ผู้เรียนที่ปฏิบัติงานคนละกลุ่ม เกิดความมั่นใจได้ว่ามาตรฐานการประเมินมีความยุติธรรม ไม่มีกลุ่มใดได้เปรียบ

4. ความเป็นไปได้ (Feasibility) อาจารย์ผู้วางแผนการประเมินจำเป็นต้องศึกษาความเป็นไปได้ของการจัดประเมินผลด้วย ไม่ว่าจะเป็นในแง่ เวลา สถานที่ งบประมาณ บุคลากร เป็นต้น เนื่องจากการพัฒนาการประเมินผลให้มีคุณภาพดีตามปัจจัย สามข้อแรก มักต้องการการลงทุน ลงแรงเพิ่มขึ้น แต่หากขาดงบประมาณก็จำเป็นต้องมีการลดหย่อนมาตรการต่าง ๆ ที่วางแผนไว้บ้างเพื่อให้ สามารถดำเนินการได้

5. ผลกระทบทางการศึกษา (Educational effect) การประเมินผลที่ดีนั้นจะช่วยส่งเสริมให้ผู้เรียนกระตือรือร้นที่จะทำการศึกษาพัฒนาความรู้และทักษะของตนเอง มีพฤติกรรมการเรียนรู้ที่เหมาะสม ตัวอย่างของการประเมินผลที่มีผลกระทบทางการศึกษาที่ไม่ดี เช่น การออกข้อสอบปรนัยที่เน้นการท่องจำเป็นการประเมินผลหลัก โดยไม่มีการประเมินรูปแบบอื่นมาเสริม ผลกระทบที่เกิดขึ้น ก็คือผู้เรียนจะมุ่งเน้นท่องเนื้อหาในตำราโดยไม่ใส่ใจการดูแลผู้ป่วยมากเท่าที่ควรในทางตรงข้ามการประเมินการปฏิบัติงานบนหอผู้ป่วย เป็นสิ่งที่ช่วยส่งเสริมให้ผู้เรียนสนใจผู้ป่วย ให้เวลากับผู้ป่วยมากขึ้น เป็นการส่งเสริมพฤติกรรม การเรียนรู้ที่ต้องการ

6. ผลเร่งการเรียนรู้ (Catalytic effect) การประเมินผลที่ดีนั้นควรมีการนำเอาข้อมูลผลการประเมินนั้นมา ให้ feedback ให้แก่ผู้เรียน เพื่อหวังผลให้ ผู้เรียนนำไปพัฒนาปรับปรุงตัวให้มีความรู้ความสามารถขึ้นในการประเมินการปฏิบัติงานของผู้เรียน อาจารย์ได้มีโอกาสนักสังเกตผู้เรียนในหลายแง่มุมทั้งในด้านความรู้ทักษะและเจตคติข้อมูลที่อาจารย์ใช้เป็นพื้นฐานของการให้คะแนนในใบประเมินการปฏิบัติงานนับว่าเป็นข้อมูลที่เป็นประโยชน์ต่อตัว ผู้เรียนเองด้วยซึ่งหากอาจารย์ สามารถจัดเวลาพูดคุย กับผู้เรียนผู้ได้รับการประเมินเพื่อให้ข้อมูลย้อนกลับ (feedback) ได้จะทำให้ได้ผลเร่งการเรียนรู้ด้วย

7. การยอมรับได้ของทุกฝ่ายที่เกี่ยวข้อง (Acceptability) การประเมินผลที่ดีนั้น ควรนำไปสู่ผลการประเมินที่เป็นที่ยอมรับได้ของทุกฝ่ายที่เกี่ยวข้อง ไม่ว่าจะเป็นผู้สอบ อาจารย์ผู้ให้คะแนน เจ้าหน้าที่ เป็นต้น

ความคลาดเคลื่อนของคะแนนอันเนื่องมาจากผู้ให้คะแนน (Rater errors)

1. การพัฒนาอาจารย์ผู้ให้คะแนน

สาเหตุสำคัญประการหนึ่งของความคลาดเคลื่อนของคะแนน คืออาจารย์ผู้ให้คะแนนมีความเข้าใจเกณฑ์การให้คะแนนแตกต่างไปจากผู้พัฒนาแบบประเมิน การจัดให้มีการชี้แจงวิธีกรใช้แบบประเมินให้อาจารย์ผู้เกี่ยวข้องทราบ รวมทั้งเปิดโอกาสให้อาจารย์ได้ทดลองใช้แบบประเมินแล้วอภิปรายแลกเปลี่ยนความเห็นกัน จะทำให้อาจารย์ผู้เกี่ยวข้องกับการประเมินนี้มีความเข้าใจที่ตรงกันมากขึ้น หลังจากที่มีการชี้แจงแล้ว ก็ควรให้มีการตรวจสอบคะแนนที่ได้จากใบประเมินของอาจารย์แต่ละท่านว่ามีอาจารย์ท่านใดที่น่าจะใช้เกณฑ์การประเมินที่แตกต่างจากอาจารย์ท่านอื่นบ้าง หากพบว่ามีการประเมินของอาจารย์ท่านใดท่านหนึ่ง ที่มีความคลาดเคลื่อนของคะแนนมาก การให้ข้อมูลย้อนกลับ (feedback) แก่อาจารย์ท่านนั้น เพื่อให้เกิดการปรับเปลี่ยนแนวทางการให้คะแนนก็จะช่วยให้ความคลาดเคลื่อนของ คะแนนมีน้อยลงเรื่อย ๆ

2. การพัฒนาแบบประเมิน

การสร้างแบบประเมินที่ดีนั้น ควรปฏิบัติตามหลักการพื้นฐานต่าง ๆ ดังนี้

2.1 เริ่มต้นสร้างแบบประเมินโดยมีความชัดเจนในวัตถุประสงค์ว่าต้องการประเมินความรู้ทักษะ หรือเจตคติในด้านใดบ้าง ควรทำการค้นคว้าเพิ่มเติมว่ามีผู้รู้ท่านอื่นได้สร้างเครื่องมือเพื่อประเมินสิ่งเดียวกันนี้มาก่อนหรือไม่ มีองค์ทฤษฎี หรือสถาบันฝึกอบรมอื่นที่ได้พัฒนาแบบประเมินในเรื่องที่คล้ายคลึงกันมาก่อนหรือไม่ การได้ข้อมูลเพิ่มเติมเหล่านี้จะทำให้หัวข้อต่าง ๆ ที่จะทำการประเมินครบถ้วน

2.2 ข้อความในแต่ละข้อเขียนด้วยภาษาที่อ่านเข้าใจง่าย สั้น และ กระชับ ควรให้อาจารย์ท่านอื่น หรือผู้เรียนช่วยอ่านและแสดงความเห็นว่ามีส่วนใดของแบบประเมินที่อ่านไม่เข้าใจบ้าง และทำการปรับแก้ตามความเหมาะสม

2.3 ในแต่ละข้อให้ทำการประเมินความรู้หรือทักษะ หรือเจตคติ เพียงด้านใดด้านหนึ่งเท่านั้น

2.4 พยายามจัดกลุ่มหัวข้อที่ทำการประเมินให้ประเด็นที่มีความคล้ายคลึงกันอยู่ข้อใกล้ ๆ กัน จะทำให้อาจารย์ผู้ประเมินทำการกรอกใบให้คะแนนได้สะดวกกว่า

2.5 ตัวเลือกระดับคะแนน สามารถสร้างได้หลายรูปแบบ แต่รูปแบบที่สามารถลดความคลาดเคลื่อนของคะแนนได้มากที่สุด คือ behavioral-anchored rating scale (BARS) ซึ่งมีการแบ่งคะแนนที่จะให้เป็นระดับ จาก น้อยไปมาก โดยในแต่ละระดับคะแนนนั้น มีการเขียนบรรยายลักษณะพฤติกรรมของผู้ถูกประเมินอย่าง ชัดเจนว่าต้องมีพฤติกรรมอย่างไร จึงจะเหมาะสมกับการได้คะแนนในระดับ ดังกล่าว

2.6 ควรจำกัดระดับของคะแนนที่อาจารย์ผู้ประเมินสามารถให้ได้ อย่าให้มีจำนวนระดับมากเกินไปโดยทั่วไปแล้วระดับคะแนนที่อาจารย์สามารถแยกแยะความรู้ ความสามารถของผู้เรียนได้ควรอยู่ในช่วง 3 – 6 ระดับการมีจำนวนระดับที่มากเกินไปมักสร้างความลำบากใจแก่อาจารย์ผู้ประเมินว่าแยกคะแนนระดับที่ใกล้เคียงกันได้ยาก

2.7 หากจะจัดให้มีระดับ คะแนนที่อยู่กึ่งกลาง (เช่น มี 5 ระดับคะแนน จาก 1 – 5 ระดับคะแนน กึ่งกลางคือ 3) ต้องระมัดระวังว่าบางครั้งอาจารย์ผู้ให้คะแนนอาจให้คะแนนกึ่งกลางดังกล่าว โดยที่ผู้เรียนไม่ได้มีระดับความรู้ ความสามารถอยู่ที่ระดับกึ่งกลางจริง แต่อาจารย์ให้คะแนนดังกล่าวด้วย เหตุผลอื่น เช่น ไม่ทันได้สังเกตพฤติกรรมดังกล่าว ไม่แน่ใจ ไม่มีโอกาสให้ผู้เรียนได้แสดงความรู้ หรือทักษะ ในด้านดังกล่าว เป็นต้น วิธีการแก้ปัญหาคือการจัดทำช่องประเมินขึ้น มาอีกช่องหนึ่ง คือ “ไม่สามารถประเมินได้” ขึ้นมา เพื่อให้อาจารย์ที่ไม่สามารถประเมินความรู้ หรือทักษะของผู้เรียนในเรื่องดังกล่าวได้ไม่ต้องเลือก ระดับคะแนนกึ่งกลางโดยความจำใจ

ข้อเสนอแนะในการประเมินการปฏิบัติงาน

เพื่อให้การประเมินความสามารถจากการปฏิบัติงานจริงเป็นไปอย่างถูกต้อง ได้ผลการประเมินที่เที่ยงตรง และเป็นธรรม ตรงตามหลักการต่าง ๆ ควรปฏิบัติต่อไปนี้

1. ให้อาจารย์ผู้ประเมินทุกท่านศึกษาแบบประเมินก่อนเริ่มสังเกตพฤติกรรมการทำงานของผู้เรียน ให้อาจารย์จดจำให้ได้ก่อนว่ามีหัวข้อใดต้องทำการประเมินบ้าง เนื่องจากการประเมินในบางหัวข้อ อาจารย์ต้องกำหนดบทบาท หรือสร้างสถานการณ์ให้ผู้เรียนแสดงความรู้ ความสามารถออกมา จึงจะประเมินได้

2. ให้อาจารย์ทำการให้คะแนนในใบประเมินในขณะที่ยังจดจำนักศึกษาผู้ถูกประเมินได้ โดยทั่วไปแล้วคือวันสุดท้ายของการปฏิบัติงานของนักศึกษา

3. ในการปฏิบัติงานแต่ละช่วงเวลาของนักศึกษา ให้อาจารย์จัดให้นักศึกษาได้รับการสังเกต และประเมินโดยอาจารย์หลายท่าน ในหลายบริบท และหลายครั้งยังมีการประเมินมาครั้ง ในมากบริบท และมากผู้ประเมิน ผลการประเมินที่ได้มาจะช่วยยืนยันกันเองได้ ทำให้ความน่าเชื่อถือของคะแนนมีมากขึ้น หากระยะเวลาปฏิบัติงาน 4 สัปดาห์ของนักศึกษามีอาจารย์เพียงท่านเดียวทำการประเมินหนึ่งครั้ง แล้วนักศึกษาก็ได้คะแนนต่ำ อาจมีการร้องเรียนว่าเป็นเพราะอาจารย์ท่านที่ประเมินไปสังเกตพฤติกรรมเขาในวันที่เขามีปัญหาขึ้นมาพอดี และไม่ใช่พฤติกรรมปกติที่เขาทำในวันอื่น ๆ แต่หากมีอาจารย์หลายท่าน ประเมินหลายครั้ง และทุกครั้งนั้น ก็ได้ผลคะแนนที่ต่ำเช่นเดียวกันหมด ความน่าเชื่อถือของผลประเมินก็มากขึ้น ว่านักศึกษาคงดั่งกล่าวมี ระดับความรู้ ความสามารถหรือเจตคติที่ไม่ดีจริง ๆ

4. ให้อาจารย์บันทึกระดับคะแนนในใบประเมินในช่วงเวลาที่อาจารย์ไม่อยู่ในสภาวะอารมณ์หงุดหงิด หิว หรือ เหนื่อยล้า เนื่องจากอารมณ์ที่แปรปรวนมีผลต่อการตัดสินใจของผู้ประเมินได้ ดังนั้นหากอาจารย์ตระหนักดีว่า กำลังไม่พอใจนักศึกษาคณะใดซึ่งได้มีพฤติกรรมไม่เหมาะสมในการทำงาน ขอให้อาจารย์ชะลอการบันทึกคะแนนไว้ก่อน รอให้อารมณ์ และความรู้สึกของเรานั้น กลับสู่สภาวะปกติก่อน การตัดสินใจต่าง ๆ ในการให้คะแนนจะได้ทำได้อย่างปราศจากอคติ

5. ให้อาจารย์อ่านหัวข้อในใบประเมินและตัดสินคะแนนทีละข้อ เนื่องจากหัวข้อต่าง ๆ ที่อยู่บนแบบประเมินแต่ละหัวข้อนั้น จะได้รับการออกแบบให้วัดผลความรู้ ทักษะ หรือเจตคติที่แตกต่างกันไป จึงไม่มีความจำเป็นที่คะแนน ในแต่ละหัวข้อต้องสอดคล้องกัน อาจารย์สามารถให้คะแนนความรู้สูง แต่มนุษย์สัมพันธ์กับเพื่อนร่วมงานต่ำก็ได้ ขอให้อาจารย์หลีกเลี่ยงวิธีการให้คะแนนแบบที่ทุกข้อได้คะแนนเท่ากัน โดยไม่ได้พิจารณารายละเอียด

6. ให้อาจารย์ใช้มาตรฐานเดียวกันในการตัดสิน คะแนนของนักศึกษาในทุกกลุ่ม ที่ปฏิบัติงานพยายามอย่าให้ปัจจัยอื่นนอกเหนือไปจากเกณฑ์ที่ระบุไว้ในแบบประเมินมีอิทธิพลทำให้เกิดความยืดหยุ่น ในเกณฑ์การพิจารณาคะแนนไม่ว่าจะเป็นความสนิทสนมส่วนตัว หรือความสามารถในมิติอื่น ๆ นอกเหนือไปจากหัวข้อที่กำหนดในแบบประเมิน

7. ให้อาจารย์ตัดสินคะแนนโดยไม่จำกัดช่วงคะแนน แต่ให้ใช้เกณฑ์ประเมินเป็นหลัก หากนักศึกษามีระดับความรู้ ความสามารถไม่ผ่านเกณฑ์ประเมิน ก็ควรประเมินคะแนนอยู่ในระดับไม่ผ่าน การประเมินผลตามจริงจะทำให้ได้ คะแนนที่มีความเที่ยงสูง และแยกแยะระดับ ความรู้ ความสามารถของนักศึกษาได้ดีกว่าคะแนนที่มีค่าพอ ๆ กัน ในนักศึกษาทุกคนไม่ว่าจะปฏิบัติงานดีหรือไม่ก็ตาม

2. ประโยชน์ที่ได้รับ

1. ทำให้เข้าใจการวัดและการประเมินผลในทั้งภาคทฤษฎี และภาคปฏิบัติได้อย่างถูกต้อง
2. สามารถนำความรู้ที่ได้ไปประยุกต์ใช้ในการจัดการเรียนการสอนทั้งภาคทฤษฎี และภาคปฏิบัติ

ข้อคิดเห็นและข้อเสนอแนะ

การอบรมในครั้งนี้เป็นหัวข้อที่ดีและน่าสนใจเป็นอย่างมาก ทำให้ได้ความรู้เกี่ยวกับการวัดและการประเมินผล เข้าใจวิธีการวัดที่หลากหลาย สามารถต่อยอดความรู้นำมาใช้ในการจัดการเรียนการสอนได้เป็นอย่างดี

3. ความคิดเห็นต่อหลักสูตรการฝึกอบรมเชิงปฏิบัติการ

หลักสูตรการฝึกอบรมเชิงปฏิบัติการในครั้งนี้มีความน่าสนใจเป็นอย่างมาก ทำให้เข้าใจหลักการและวิธีการสำคัญในการวัดและประเมินผลนักศึกษาได้เป็นอย่างดี ทำให้สามารถออกแบบการวัดผลและประเมินผลการศึกษาให้เหมาะสมกับวัตถุประสงค์การเรียนการสอนในหัวข้อ/รายวิชา และสามารถนำความรู้ทางด้านการวัดและการประเมินผลได้อย่างมีประสิทธิภาพ

การตั้งเกณฑ์การสอบผ่าน (Standard Setting)

รศ.ดร.นพ.เชิดศักดิ์ ไอรมนรัตน์
อ.ดร.ปาริชาติ อภิเดชากุล
ศูนย์ความเป็นเลิศด้านการศึกษาวิทยาศาสตร์สุขภาพ
คณะแพทยศาสตร์ศิริราชพยาบาล
มหาวิทยาลัยมหิดล

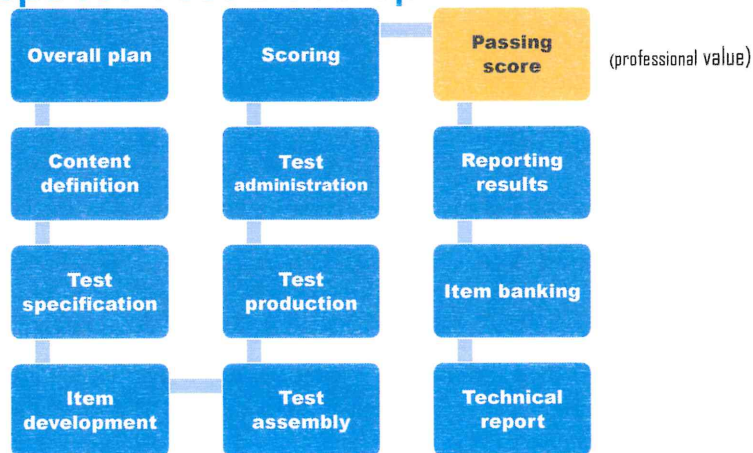
Outline

Explain the importance and principle of standard settings

Discuss different methods of setting the standard

Apply the standard setting method and use it in different health professions context

Twelve Steps for Test Development



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

Common Concepts in Standard Setting

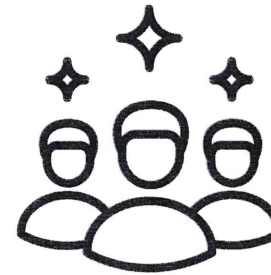
- Standard Setting
- Minimum Pass Level (MPL)
- Minimally Competent (Borderline) Student
- Panel

- Bandaranayake RC. Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37. Medical Teacher; 2008;30(9-10):836-45.
- Yudkowsky R, Downing SM, Tekian A. Standard setting. Assessment in health professions education. Vanderbilt Avenue, New York: Routledge; 2019. p. 86-105.
- De Champlain AF. Standard Setting Methods in Medical Education. In: Swanwick T, Forrest K, O'Brien BC, editors. Understanding Medical Education. West Sussex, UK: Wiley Blackwell; 2018. p. 347-59.

Steps in Setting Standards

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

Judgement



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

Steps in Setting Standards

1. Types of Standards

- ❖ Absolute standard (เกณฑ์ตัดสินอิงเกณฑ์)
- ❖ Relative standard (เกณฑ์ตัดสินอิงกลุ่ม)

Steps in Setting Standards

2. Methods for Setting Standards

- 2.1 **Test**-centered methods (แบบสอบเป็นศูนย์กลาง)
- 2.2 **Examinee**-centered methods (ผู้สอบเป็นศูนย์กลาง)
- 2.3 Compromised methods

2.1.1 Modified Angoff's Method

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

2.1 **Test-Centered** Methods

(เกณฑ์ตัดสินอิงเกณฑ์)



Judgmental 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

นักศึกษาคาบเส้น



2.1.1 Modified Angoff's Method (2)

ข้อที่	ความน่าจะเป็นในการตอบถูก
1	0
2	1
3	1
4	0
5	1
Passing score	

2.1.2 Nedelsky's Method

• The judgment (การตัดสิน)

- How many options a borderline examinee can **eliminate** from choosing in an item

• The passing score (เกณฑ์สอบผ่าน)

- 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

2.1.2 Nedelsky's Method (2)

ข้อที่	A	B	C	D	E	Not eliminated	ความน่าจะเป็นของตัวถูก
1			X	X	X	2	$\frac{1}{2} = 0.50$
2	X	X				3	$\frac{1}{3} = 0.33$
3	X					4	$\frac{1}{4} = 0.25$
4	X		X	X		2	$\frac{1}{2} = 0.50$
5	X				X	3	$\frac{1}{3} = 0.33$
Passing score							$\dots/5 \times 100 =$

2.1.3 Ebel's Method

	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

• The judgment (การตัดสิน)

- 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 (เกณฑ์สอบผ่าน)

- The sum of all the probability of correct answers for all items on the exam

2.1.3 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

ความน่าจะเป็นลดลงตามความยากง่าย ระดับความสำคัญเนื้อหา

2.1.3 Ebel's Method (2)

Item	Difficulty	Importance	ความน่าจะเป็นที่ นักเรียนคาบเส้นจะตอบ ข้อสอบถูก
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
คะแนนผ่านเกณฑ์			

The passing score (เกณฑ์สอบผ่าน)

– The sum of all the probability of correct answers for all items on the exam

2.2 Examinee-Centered Methods

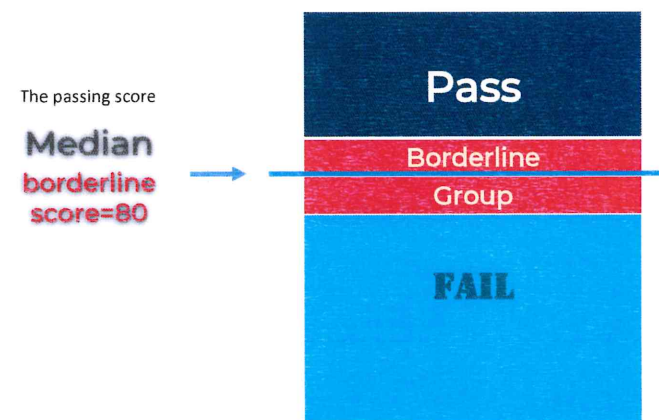
- The **judges** set a standard by reviewing the **overall performance of examinees** and determining **who should pass and who should fail**. The scores of examinees are reviewed and the passing score is set based on these judgments
 - Borderline-group method
 - Contrasting-groups method



2.2.1 Borderline-group method

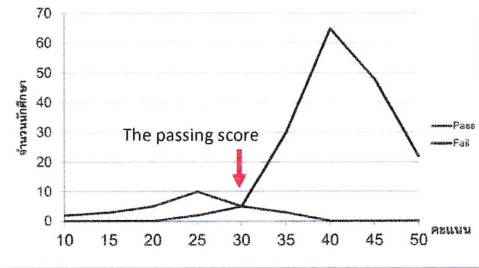
- The **judgment** (การตัดสินใจ)
 - Identify examinees who are “borderline”
(ยังไม่ทราบคะแนน)
- The **passing score** (เกณฑ์สอบผ่าน)
 - The median score of this “borderline group”

BORDERLINE-GROUP METHOD



2.2.2 Contrasting-Groups Method

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



Judgmental-empirical methods

2.3 Compromised Method

- Combining relative and absolute standard setting methods

2.3.1 Hofstee method

The judgment (การตัดสิน)

Minimum failure rate (f_{min})

Maximum failure rate (f_{max})

Minimum passing score (c_{min})

Maximum passing score (c_{max})

The passing score (เกณฑ์สอบผ่าน)

The intersection of test scores curves with a diagonal line drawn from the upper left to the lower right corner

2.3.1 Hofstee Method (2) Empirical-judgmental methods

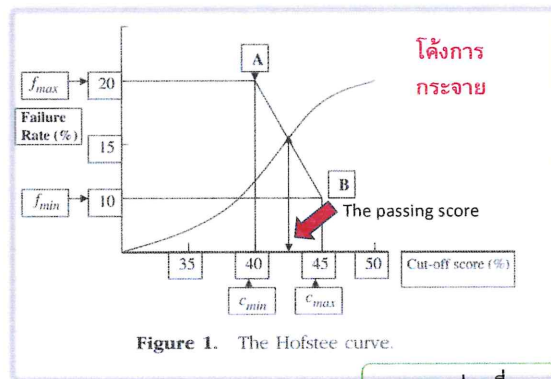


Figure 1. The Hofstee curve.

คะแนนผ่านที่ยอมรับได้ (%)

อัตราล้มเหลว (%)

วิธี	ประเด็นที่ใช้ในการตัดสิน	สังเกตการปฏิบัติของผู้สอบ	ช่วงเวลาที่ตัดสินเกณฑ์การตัดสินก่อน ระหว่าง หลังสอบ
Test-centered methods			
Examinee-centered methods			
Compromised methods			

วิธี	ประเด็นที่ใช้ในการตัดสิน	อาจารย์สังเกตการปฏิบัติของผู้สอบ	ช่วงเวลาที่ตัดสินเกณฑ์การตัดสิน
Test-centered methods (Pre-fixed cut-off score)	ข้อสอบรายข้อ	ไม่ใช่	ก่อนการสอบ Absolute standard
Examinee-centered methods (Standardized pass/fail rate)	ความสามารถของผู้สอบ	ใช่	ระหว่างหรือหลังการสอบ Relative standard
Compromised methods	แบบสอบทั้งชุด	ไม่ใช่	หลังการสอบ

Steps in Setting Standards

3. Selecting Judges

- The number of judges
- The qualification of judges

The qualification of judges



Standard
Setting in
Education



Self-monitoring

Face-to-face or virtual training is organized to align panel members on the purpose, methodological process, importance of standard setting, and definitions of performance levels.

Steps in Setting Standards

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 cut-point

6. Checking Standard

B. Maintenance of standards (Consistent)

1. Test equating

Standards can be maintained by test equating methods. “marker questions” from previous examination to determine **the relative difficulty of each examination**.

2. Develop the test to fit the standard

	Easy	Medium	Difficult
Essential			
Important			
Acceptable			
Questionable			

6. Checking Standard

A. Validation of two assumptions

1. External validity: Accurate reflection

1. Relationship with performance standards

2. Internal validity: Appropriateness

1. Stability of the passing standard
2. Selection, training, and monitoring of judges

Do we have to care about error?

• True score theory (ทฤษฎีคะแนนจริง)

- 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 (Standard Error of measurement)

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

SD = standard deviation

r = internal consistency reliability

↑SD (more spread of score): higher SEM

↑r (more accurate measures): smaller SEM

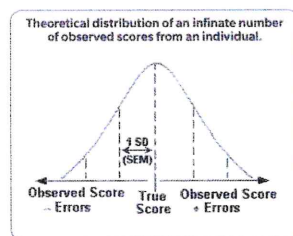


Figure 11: SEM Example

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

คำนวณคะแนนจริงของคะแนนที่วัดได้ในผู้สอบรายบุคคล

ตัวอย่าง

- กรณีคะแนนจุดตัดหรือเกณฑ์ผ่านขั้นต่ำ เท่ากับ 80
- หากมีผู้สอบ 2 คน สอบได้ 79, 78
- จะพิจารณาอย่างไรว่าจะยอมให้ผู้สอบคนใดผ่านได้

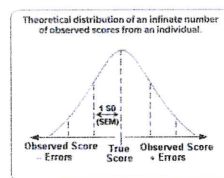


Figure 11: SEM Example

การคำนวณความคลาดเคลื่อนมาตรฐานของการวัด และหาช่วงความเชื่อมั่นของคะแนนจริงของผู้สอบครอบคลุมจุดตัด

ช่วงความเชื่อมั่นที่นิยมคือร้อยละ 90 (คะแนน $\pm 1.645SEM$) และร้อยละ

95 (คะแนน $\pm 1.96SEM$)

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

สมมติ MPL = 80, S.D. = 3, $r_{tt} = 0.9$

Take-Home Message

- While **developing standards** requires considerable effort, it is well worth it when used to certify safe medical doctors.
- Although there are several standard-setting methods available, the Angoff method is **the most popular** and straightforward to use.
- **Training judges** is the most critical step in any standard-setting process.
- There is **no gold standard-setting method**. For each test and its context, a standard-setting method could be appropriate.
- **All steps** in performing standard setting should be appropriately documented as evidence to support **the validity of decision-making of test results**.

Gasmalla HE, Ibrahim AA, Wadi MM, Taha MH, editors. *Written Assessment in Medical Education*. Springer; 2023 Mar 11.

Authentic Assessment - Foundation and application

Assessment workshop For Clinical teacher

1

Authentic Assessment (AA)

3 Main attributes of AA

- Formative
- Longitudinal
- Authenticity
 - Real-world situation
 - Related-competencies

Assessment workshop For Clinical teacher

2

Consensus of Good Assessment

Where the authentic assessment shine in the
Ottawa consensus of good assessment?

- | | |
|----------------|-----------------------|
| 1. Validity | 5. Educational effect |
| 2. Reliability | 6. Catalytic effect |
| 3. Equivalence | 7. Acceptability |
| 4. Feasibility | |

Assessment workshop For Clinical teacher

3

Authenticity vs Fidelity

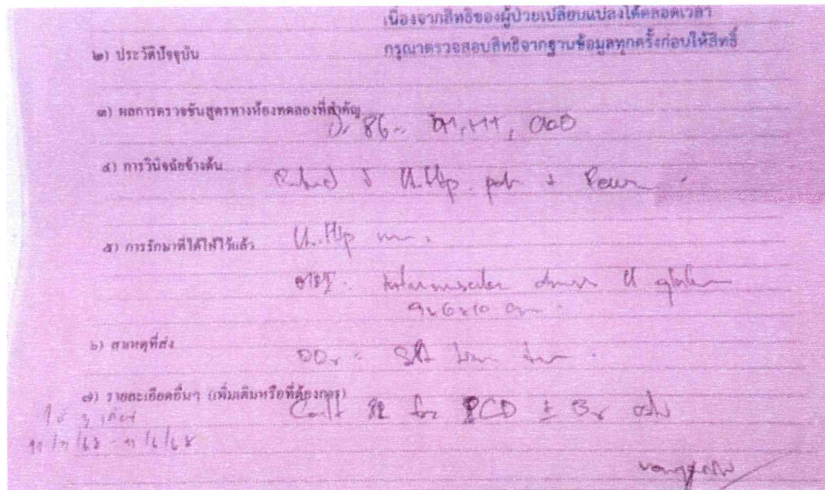
Higher Fidelity introduces Higher cognitive load

Low
Cognitive
Load

Assessment workshop For Clinical teacher

4

Authenticity? - Documentation



Assessment workshop For Clinical teacher

5

Authenticity? - OSCE

OSCE conversation

Patient - คุณสมชาย แข็งแรง	Doctor - นศพ.บุญญภัทร
	สวัสดีครับ คุณสมชาย ผม นศพ.บุญญภัทร ปี 6
สวัสดีครับหมอ ผมหายใจ เหนื่อยมา 2-3 วันแล้วครับ	อาการเหนื่อยเป็นยังไงหรือ ครับ
เวลาผมเดิน 30-40 นาที มันก็จะ เริ่มเหนื่อย	มีโรคประจำตัวอะไรบ้างครับ
โรคประจำตัว เช่น โรคอะไรบ้าง หรือหมอ	

Assessment workshop For Clinical teacher

6

Authenticity? - WPBA

Does authentic assessment undermine authentic learning?

- Yes, only if students perceive it as high-stake exam and judgmental
- No, if students could show uncertainty and vulnerability
- No, if the WPA is authentic and acknowledge complexity of case

Assessment workshop For Clinical teacher

7

Check Reliability

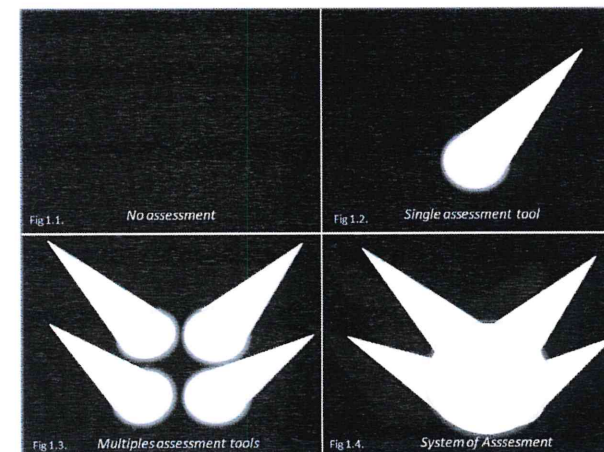


Figure 1. An illustration of the various states of assessment.

Assessment workshop For Clinical teacher

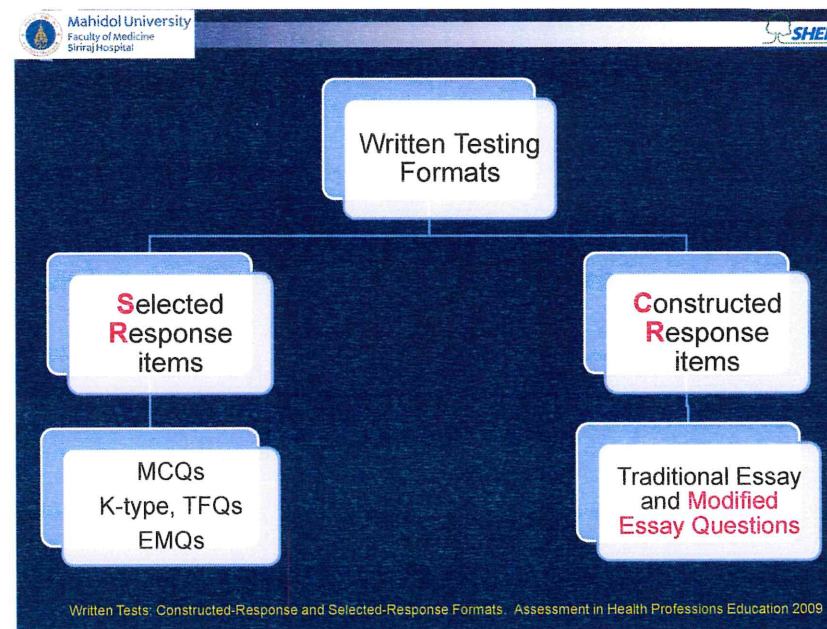
9

References

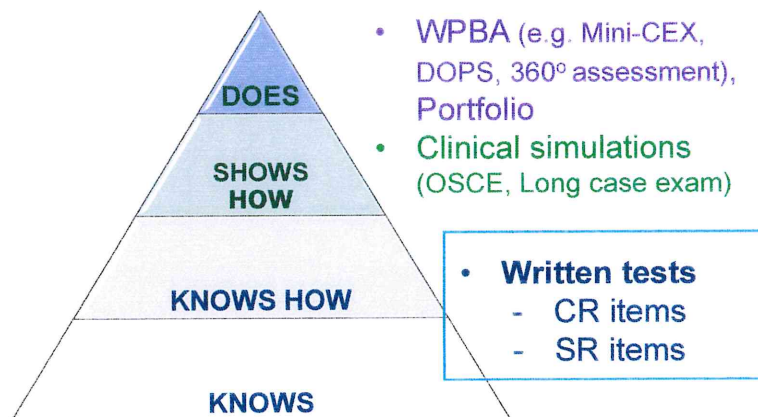
- Harden, Ronald M., and Geoffrey R. Norman, eds. 2016. *The Definitive Guide to the OSCE: The Objective Structured Clinical Examination as a Performance Assessment*. Edinburgh: Elsevier.
- Hatala, Rose, and Rachel H. Ellaway. 2024. "Does Authentic Assessment Undermine Authentic Learning?" *Advances in Health Sciences Education* 29 (4): 1067–70. <https://doi.org/10.1007/s10459-024-10361-z>.
- Norcini, John, M. Brownell Anderson, Valdes Bollela, Vanessa Burch, Manuel João Costa, Robbert Duvivier, Richard Hays, Maria Felisa Palacios Mackay, Trudie Roberts, and David Swanson. 2018. "2018 Consensus Framework for Good Assessment." *Medical Teacher* 40 (11): 1102–9. <https://doi.org/10.1080/0142159X.2018.1500016>.

Constructed Response Items Development

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



Miller's Pyramid and suitable methods of Assessment



Constructed Response Items

Traditional Essay Questions

- Long Essay
- Short Essay

Modified Essay Questions

- Standard Modified Essay Questions (MEQ)
- Key Feature Questions (KFQ)
- Patient Management Problem (PMP)
- Short Answer Questions (SAQ)

LOs: Cognitive domain Modified Bloom's taxonomy

- Level I: Recall, Recognition
- recall of information
- Level II: Comprehension, Interpretation
- understanding and being able to interpret data
- Level III: Application, Problem solving
- use of knowledge and understanding in new circumstances



Use of an educational taxonomy for evaluation of cognitive performance. *J Med Educ* 1981

Objectives

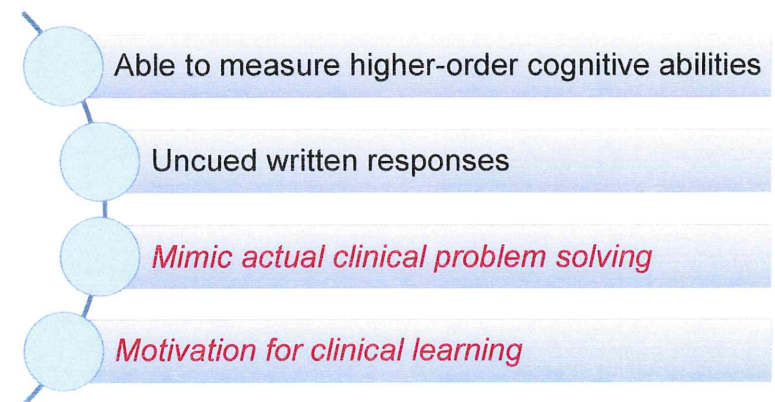
เมื่อสิ้นสุดการอบรม ผู้เข้าร่วมอบรมสามารถ

1. อธิบายข้อดีและข้อจำกัดของข้อสอบชนิด MEQ
2. อธิบายขั้นตอนและประเด็นสำคัญของการสร้างข้อสอบ MEQ
3. ร่วมในกระบวนการพัฒนาและทบทวนข้อสอบ MEQ สำหรับนักศึกษาในระดับคลินิกที่แต่ละท่านเกี่ยวข้องได้

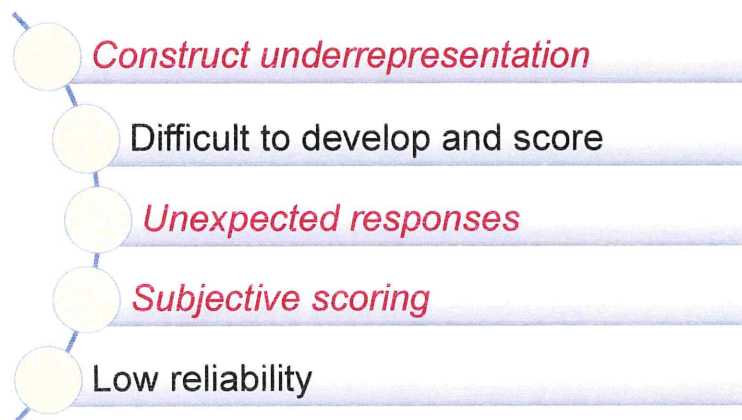
MEQ

- Strengths & Limitations
- Key principles of MEQ item development
 - Hypothesis-driven approach
 - Key Feature Questions
- Developing an MEQ

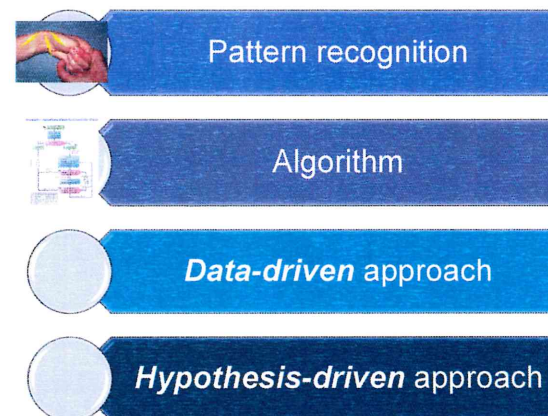
MEQ: Strengths



MEQ: Limitations

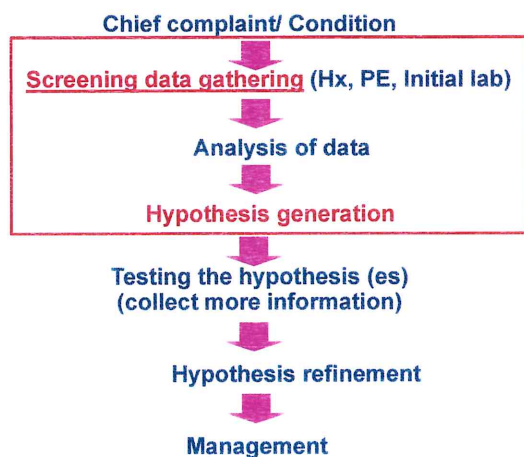


Clinical Problem-Solving Methods

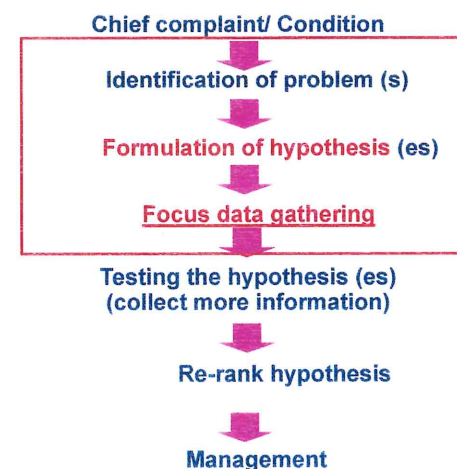


วิชญ์ ธรรมลิขิตกุล, การประเมินความถี่ในการแก้ปัญหาผู้ป่วยทางคลินิก, สารคดีราช 2534

Data-driven approach



Hypothesis-driven approach

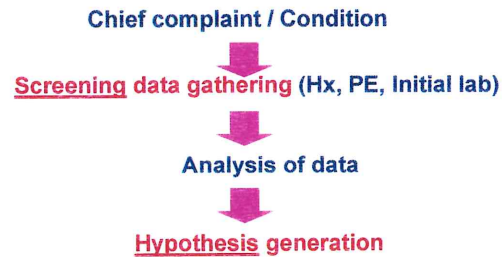


Data-driven approach

Clinical problem: เด็กชายอายุ 4 ปี มีอาการนอนกรนมา 1 ปี

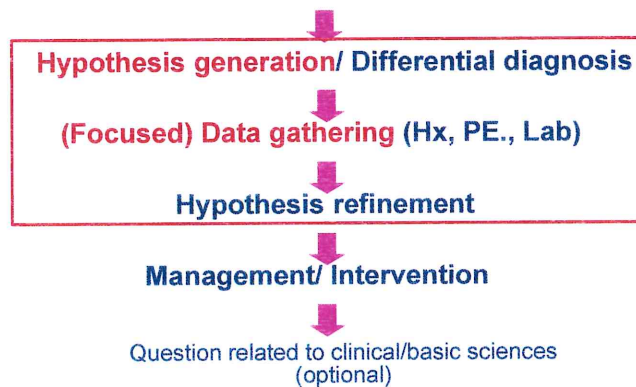
Question: ท่านจะซักประวัติอะไรบ้างเพื่อวินิจฉัยโรค

Data gathering



MEQ Process (Hypothesis-driven approach)

A brief clinical situation (scenario, vignette)



Hypothesis-driven approach

Clinical problem: เด็กชายอายุ 4 ปี มีอาการนอนกรนมา 1 ปี

ช่วง 3 เดือนนี้นอนกรนเกือบทุกคืน กระสับกระส่าย สะดุ้งตื่นบ่อย มี
หายใจสะดุดเป็นระยะ

Question: ระบุสาเหตุที่เป็นไปได้ของการนอนกรนในเด็กคนนี้มา 3 ข้อ

Hypothesis generation



MEQ: Serial Question Test

- เสริมการแก้ปัญหาผู้ป่วยในสถานการณ์จริง
- ประกอบด้วยหลายขั้นตอน
 - มีข้อมูลผู้ป่วยบางส่วนในช่วงแรก
 - สืบค้นหาข้อมูลเพิ่มเติมและวิเคราะห์ ตัดสินใจแก้ปัญหาที่
ละขั้นตอน
 - เมื่อทำแต่ละขั้นตอนแล้ว ไม่สามารถย้อนกลับไปแก้ไขสิ่งที่
ทำไปก่อนหน้านี้ได้

ประเมินทักษะ

- การแก้ปัญหา (problem solving)
- การตัดสินใจ (decision making)

Physician tasks / Competencies

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

Standard MEQ

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

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

KFQs: A key feature



A critical or **essential** step(s) in the resolution of a problem



A step(s) in which examinees are most likely to **make errors** in the resolution of the problem



A difficult or **challenging** aspect of **identifying** or **managing** the problem in practice

1. A practical guide to assessing clinical decision-making skills using the key features approach. *Med Educ* 2005
2. Key-feature questions for assessment of clinical reasoning: a literature review. *Med Educ* 2014
3. Twelve tips for developing key-feature questions (KFQ) for effective assessment of clinical reasoning. *Med Teach* 2018

Key-Feature Questions (KFQs)

- Provide a more focused assessment of **case-specific** clinical decision-making
- "In any clinical case, there are **2-5 key elements** in decision making."*
- Facilitate the inclusion of additional cases and items to assess a **broader content domain**
- Effectively distinguish between candidates with varying levels of clinical problem-solving ability

1. The Medical Council of Canada's key features project: A more valid written examination of clinical decision-making skills. *Acad Med* 1995
2. Twelve tips for developing key-feature questions (KFQ) for effective assessment of clinical reasoning. *Med Teach* 2018

KFQs: Validity & Reliability

- Provide a flexible approach to testing clinical decision-making skills with demonstrated validity and reliability when constructed according to the guidelines
 - Provide strong evidence for the content validity of the key features
 - Reliability of 0.8 in 4 hours of testing had been demonstrated

1. Content validation of key features on a national examination of clinical decision-making skills. *Acad Med* 1995
2. A practical guide to assessing clinical decision-making skills using the key features approach. *Med Educ* 2005

Developing an MEQ

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

Example of a Key Features Template for the Third Trimester Bleeding Problem.

Clinical problem	Third Trimester Bleeding
MCC objective	Vaginal Bleeding (112); Pregnancy - Obstetrical Complications (80-3)
Age group	Adult
Clinical situation	Single typical problem - life-threatening event
Site/setting	Emergency department

Key features: Given a pregnant woman experiencing third-trimester vaginal bleeding with no abdominal pain, the candidate will:

1. Consider placenta previa as a leading diagnosis;
2. Avoid performing a pelvic examination (because it may cause fatal bleeding);
3. Avoid discharging home from the emergency department; and
4. Order pelvic ultrasound (to confirm placenta location).

“Key features case is designed to specifically assess decision-making skills rather than simple recall of factual knowledge”

Guideline for the development of key feature problems & test cases. Medical council of Canada 2012

Assembling Problem-Writing Groups

Item writers

- Clinical expertise
- **Multidisciplinary approach/ combined expertise**

The written problem

- Well grounded in practice
- Represent a wide range of **real-life practice**

Review the content by a group of writers

Select a problem Define the key features

Define the domain of *clinical problems*

Develop an examination *blueprint*

Develop a *key feature* problem for each clinical problem

Defining Key Features (1/3)

KFs should be as *specific as possible*

KFs link to *physician tasks*

Select A Problem

- Refer to test specification table
- Select an appropriate clinical problem
 - พบบ่อยในเวชปฏิบัติ
 - ความยากง่ายเหมาะสมกับระดับของผู้เรียน
 - ประเมินทักษะการแก้ปัญหาและการตัดสินใจ
 - แพทย์มักตัดสินใจผิดพลาด
 - เกี่ยวข้องกับหลายระบบ
 - มีการบูรณาการของสาขาวิชา

Priority Topics and Key Features

Croup		
Key Feature	Skill	Phase
1 In patients with croup, a) <u>Identify the need for respiratory assistance</u> (e.g., assess ABCs, fatigue, somnolence, paradoxical breathing, in drawing) b) <u>Provide that assistance when indicated.</u>	Selectivity Clinical Reasoning Clinical Reasoning Selectivity	Hypothesis generation Diagnosis Treatment
2 Before attributing stridor to croup, <u>consider other possible causes</u> (e.g., anaphylaxis, foreign body (airway or esophagus), retropharyngeal abscess, epiglottitis).	Clinical Reasoning Selectivity	Hypothesis generation Diagnosis
3 In any patient presenting with respiratory symptoms, look specifically for the signs and symptoms that differentiate upper from lower respiratory disease (e.g., stridor vs. wheeze vs. whoop).	Clinical Reasoning	History Physical
4 In a child presenting with a clear history and physical examination compatible with mild to moderate croup, <u>make the clinical diagnosis without further testing</u> (e.g., do not routinely X-ray).	Clinical Reasoning Selectivity	Diagnosis Investigation
5 In patients with a diagnosis of croup, <u>use steroids</u> (do not under treat mild-to-moderate cases of croup).	Clinical Reasoning	Treatment
6 In a patient presenting with croup, <u>address parental concerns</u> (e.g., not minimizing the symptoms and their impact on the parents), acknowledging fluctuating course of the disease, providing a plan anticipating recurrence of the symptoms.	Clinical Reasoning Communication	Treatment Follow-up

Defining Key Features (2/3)

อภิปรายในกลุ่มผู้เขียนโจทย์จนได้ consensus

กำหนด significant steps

- ประเด็นสำคัญในการตัดสินใจ จัดการกับปัญหาผู้ป่วย
- ขั้นตอนที่ขาดไม่ได้ในการดูแลรักษาผู้ป่วย
- อาจเป็นประเด็นเกี่ยวกับ ethics, medico-legal, EBM

From A Problem to A Case: write case scenario

- Based on your own clinical experience
- Make it realistic
- Include data that will be required in order to answer the ensuing questions

Defining Key Features (3/3)

Typical KFs

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

ไม่จำเป็นต้องเริ่มต้นด้วยการถามประวัติ หรือ ตรวจร่างกาย

From A Problem to A Case

Decide a case scenario : define the context

Age, Gender

Setting of the encounter:

OPD, IPD, ER

KFs on diagnosis: brief case

KFs on management: longer case

Write Questions to Test KFs

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

Specify the Number of Required Answers

ระบุคำถามให้ชัดเจน

- บอกชื่อโรคที่ผู้ป่วยรายนี้น่าจะเป็นมากที่สุด 1 โรค
- บอกสิ่งที่ตรวจพบจากการตรวจร่างกายที่สำคัญที่จะช่วยยืนยันการวินิจฉัยโรค มา 3 ประการ
- เขียนคำสั่งการรักษาสำหรับผู้ป่วยรายนี้ในใบคำสั่งการรักษา

Preparing Scoring Keys (1/4)

- List of correct and incorrect options for answer
- Scores to be assigned to each response
 - **Multiple** acceptable answers

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

- **Only one** acceptable answer

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

Preparing Scoring Keys (2/4)

Complete score	คำตอบถูกต้องและสมบูรณ์
Partial score	คำตอบถูกต้องและสมบูรณ์ เพียงบางส่วน

No score

คำตอบไม่ถูกต้อง

"Partial-credit scoring"

Preparing Scoring Keys (3/4)

- Partial credit scoring

Investigation

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

Treatment

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

Preparing Scoring Keys (4/4)

Penalty/ Negative score

Absence of "must have" answers

- score of "0" *despite the presence of other less important answers*

Penalty/ Negative score

Presence of "unnecessary" investigations or "Harmful" Rx

- "0"
- negative score (but not cross items)

Time

- กำหนดเวลาให้เพียงพอสำหรับแต่ละคำถาม
 - อ่านข้อมูลเพิ่มเติมในแต่ละหน้า ที่อาจมีเนื้อหาหมก
 - วิเคราะห์คำถาม
 - เขียนคำตอบ
- เวลาที่นักศึกษาใช้ในการตอบคำถามนั้น ๆ จะมากกว่าเวลาที่อาจารย์ใช้ 30 – 50 %
 - ทดลองตอบด้วยตนเองและจับเวลา / ให้เพื่อนทดลองทำ

Validation and References

- Validation
 - pilot the problem* with colleagues new to the problem: discussion, revision
- References
 - Standard textbook, CPG
 - specially in the field of rapidly developing intervention and discovery*

Conclusion

- MEQ is a written test which can be used to measure ability of solving the clinical problems.
- Developing an MEQ should be based on hypothesis-driven and key feature (KF) approaches.
- The KF approach is based on case specificity

Workplace-Based Assessment

Assoc. Prof. Tripop Lertbunnaphong
Lecturer.Dr. Parichad Apidechakul

CONTENTS

1. Introduction
2. Competency Assessment Framework
3. Methodology
4. Strengths of WPBA
5. Types of WPBA
6. Feedback Importance
7. Challenges
8. Conclusion

01.

Introduction

What is Workplace-based assessment?

Definition of WPBA

WPBA as a Real-Time Evaluation Tool

A real-time evaluation tool used to assess healthcare professionals' performance in clinical settings.

It aligns with the competency-based medical education (CBME) model.

WPBA incorporates direct observation, immediate feedback, and structured methodologies to evaluate competencies in real-world scenarios.

Singh K, Singh A. Workplace-based assessment: a real-time assessment tool. MAMC Journal of Medical Sciences. 2021 Sep 1;7(3):204-10.

02

Competency Assessment Framework

Competency Assessment Framework

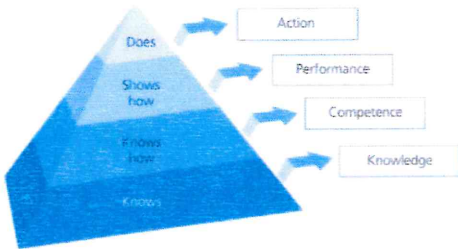


Figure 22.1 Miller's pyramid for assessing clinical competence

(Rethans et al. 1991 ; Kopelow et al. 1992).

Alignment with Miller's Pyramid

Miller's Pyramid Levels

Level 1
Knowledge ("Knows").

Level 2
Application of knowledge ("Knows how").

Level 3
Clinical skills competency ("Shows").

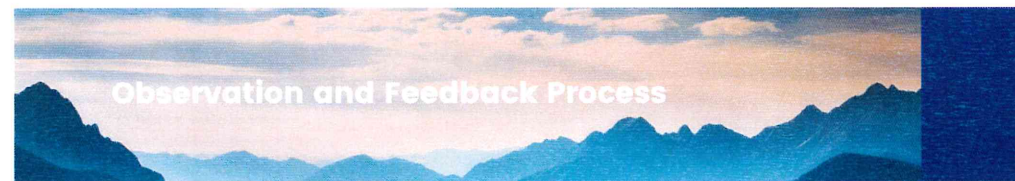
Level 4
Clinical performance in real settings ("Shows how").

WPBA focuses on the top level, emphasizing behavioral and practical skills.

03

Methodology

Structured Encounters and Feedback



WPBA involves structured encounters observed by skilled assessors using checklists and rating scales.

Feedback is provided immediately after each encounter, highlighting areas for improvement and creating an action plan.

Multiple encounters are conducted to ensure reliability and reduce bias.

Advantages

Benefits of WPBA

Provides **targeted feedback** to guide learning toward desired outcomes.

Reduces reliance on summative exams, which can be influenced by luck or examiner bias.

Conducted **during routine work** without requiring additional infrastructure.

Encourages **reflection and self-improvement** among trainees.

Helps **underperformers identify deficiencies early** while offering limited benefits for high achievers.

04

Strengths of WPBA

05

Types of WPBA

การอภิปรายรายงานผู้ป่วย (case-based discussion)

วิธีการ (Structured discussion)

ผู้เรียนนำเสนอรายงานผู้ป่วย สรุปและกำหนดประเด็นสำคัญ
อาจารย์ซักถาม เพื่อการใช้เหตุผลประกอบการตัดสินใจดูแลผู้ป่วย

ผู้ร่วมอภิปราย

ผู้เรียนระดับเดียวกัน/ประสบการณ์หลายระดับ

ผู้ประเมิน

สังเกตการณ์และให้ข้อมูลย้อนกลับ

สิ่งที่มุ่งประเมิน

ความรู้ ทักษะการวิเคราะห์ การแก้ปัญหา การคิดเป็นเหตุเป็นผล
การวินิจฉัยโรค วางแผนการรักษา



A single observation or assessor- assessee encounter

Methods

Direct Observation	Case Discussions	Feedback Linked Methods	Documentation Methods
Mini- clinical evaluation exercise (mini- CEX). Direct observation of procedural skills (DOPS).	Case- based discussions. Blinded patient encounters.	Multisource feedback (360° assessment).	Logbooks and portfolios.

แบบประเมิน Case-based Discussion (CBD)

ผู้เรียน: ☐ ผู้เรียน ☐ ผู้เรียน ☐ ผู้เรียน ☐ ผู้เรียน

อาจารย์: ☐ อาจารย์ ☐ อาจารย์ ☐ อาจารย์ ☐ อาจารย์

ผู้สังเกตการณ์: ☐ ผู้สังเกตการณ์ ☐ ผู้สังเกตการณ์ ☐ ผู้สังเกตการณ์ ☐ ผู้สังเกตการณ์

ผู้ประเมิน: ☐ ผู้ประเมิน ☐ ผู้ประเมิน ☐ ผู้ประเมิน ☐ ผู้ประเมิน

หัวข้อ	Before expectation	Before expectation	After expectation	After expectation	Unable to comment
1. การวินิจฉัยโรค (Medical record)	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	
2. การวินิจฉัยโรค (Clinical examination)	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	
3. การวินิจฉัยโรค (Investigation and referral)	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	
4. การวินิจฉัยโรค (Follow-up and future planning)	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	
5. การวินิจฉัยโรค (Professionalism)	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	
6. การวินิจฉัยโรค (Overall clinical judgment)	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	

การสังเกตการณ์และให้ข้อมูลย้อนกลับ

ผู้สังเกตการณ์: ☐ ผู้สังเกตการณ์ ☐ ผู้สังเกตการณ์ ☐ ผู้สังเกตการณ์

ผู้ประเมิน: ☐ ผู้ประเมิน ☐ ผู้ประเมิน ☐ ผู้ประเมิน

ผู้เรียน: ☐ ผู้เรียน ☐ ผู้เรียน ☐ ผู้เรียน ☐ ผู้เรียน

ผู้สังเกตการณ์: ☐ ผู้สังเกตการณ์ ☐ ผู้สังเกตการณ์ ☐ ผู้สังเกตการณ์ ☐ ผู้สังเกตการณ์

ผู้ประเมิน: ☐ ผู้ประเมิน ☐ ผู้ประเมิน ☐ ผู้ประเมิน ☐ ผู้ประเมิน

ผู้เรียน: ☐ ผู้เรียน ☐ ผู้เรียน ☐ ผู้เรียน ☐ ผู้เรียน

ผู้สังเกตการณ์: ☐ ผู้สังเกตการณ์ ☐ ผู้สังเกตการณ์ ☐ ผู้สังเกตการณ์ ☐ ผู้สังเกตการณ์

ผู้ประเมิน: ☐ ผู้ประเมิน ☐ ผู้ประเมิน ☐ ผู้ประเมิน ☐ ผู้ประเมิน

ผู้เรียน: ☐ ผู้เรียน ☐ ผู้เรียน ☐ ผู้เรียน ☐ ผู้เรียน

ผู้สังเกตการณ์: ☐ ผู้สังเกตการณ์ ☐ ผู้สังเกตการณ์ ☐ ผู้สังเกตการณ์ ☐ ผู้สังเกตการณ์

ผู้ประเมิน: ☐ ผู้ประเมิน ☐ ผู้ประเมิน ☐ ผู้ประเมิน ☐ ผู้ประเมิน

Case-based discussion

Mini-clinical evaluation exercise (mini-CEX)

วิธีการ

ประเมินทักษะผู้ป่วยแบบเฉพาะเจาะจง

ประเมินหลายทักษะ

ประเมินความสามารถในองค์รวม

ประเมิน 15-20 นาที และให้ข้อมูลย้อนกลับ 5 นาที

ผู้ประเมิน

สังเกตการณ์และให้ข้อมูลย้อนกลับ

สิ่งที่มุ่งประเมิน

การถามประวัติ การตรวจร่างกาย การให้คำปรึกษา การตัดสินใจทางคลินิก และการประเมิน

ความสามารถในองค์รวม การสื่อสาร

แบบประเมิน

แบบประเมินที่มีรูปแบบชัดเจน ส่วนมากใช้แบบประเมิน 9 ระดับ (9-point rating scale)

A single observation or assessor–assessee encounter



Liu C. An introduction to workplace-based assessments. Gastroenterol Hepatol Bed Bench 2012;5:24-8.
Bodgener S. Royal College of General Practitioner WPBA handbook; 2020.

แบบประเมิน Mini-Clinical Evaluation Exercise (mini-CEX)

ผู้ประเมิน: _____

แพทย์ผู้ถูกประเมิน: _____ วันที่: ____/____/____

ผู้ป่วย: _____

สถานที่: _____

เวลา: _____

หัวข้อ: _____

วัตถุประสงค์: _____

เกณฑ์การประเมิน: _____

ทักษะ	Unsatisfactory	Satisfactory	Superior	NA
1. การซักประวัติ (History taking)	1 2 3	4 5 6	7 8 9	
2. การตรวจร่างกาย (Physical examination)	1 2 3	4 5 6	7 8 9	
3. ความไว้วางใจ (Trust)	1 2 3	4 5 6	7 8 9	
4. การตัดสินใจทางคลินิก (Clinical judgment)	1 2 3	4 5 6	7 8 9	
5. การสื่อสาร (Communication)	1 2 3	4 5 6	7 8 9	
6. การประเมิน (Evaluation)	1 2 3	4 5 6	7 8 9	
7. การจัดการ (Management)	1 2 3	4 5 6	7 8 9	

หมายเหตุ: _____

ผู้ประเมิน: _____

แพทย์ผู้ถูกประเมิน: _____

ผู้ป่วย: _____

Mini-clinical evaluation exercise (mini-CEX)

ตัวอย่างของ Mini-CEX

- Medical interviewing skills
- Physical examination skills
- Humanistic qualities/professionalism
- Clinical judgment
- Counselling skills
- Organization and efficiency

Direct Observation of Procedural Skills (DOPS)

วิธีการ

การสังเกตการณ์ปฏิบัติหัตถการโดยตรง เป็นการประเมินที่ประกอบ

ด้วยการสังเกตความสามารถของผู้เรียน

ผู้ประเมิน

สังเกตการณ์และให้ข้อมูลย้อนกลับ

สิ่งที่มุ่งประเมิน

ทักษะหัตถการ

แบบประเมิน

Rating scale



A single observation or assessor–assessee encounter

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

Mini-Clinical Evaluation Exercise (CEX) – F1 Version

Please complete the following sections:

1. History taking

2. Physical examination skills

3. Communication skills

4. Clinical judgment

5. Professionalism

6. Organization/efficiency

7. Overall clinical care

Anything especially good?

Suggestions for development

Anything especially good?

Suggestions for development

Agreed action:

Anything especially good?

Suggestions for development

Liu C. An introduction to workplace-based assessments. Gastroenterol Hepatol Bed Bench 2012;5:24-8.

Direct Observation of Procedural Skills Form B
Department of Surgery, Faculty of Medicine Siriraj Hospital
Gowning and gloving

Part 1: To be completed by the student
Student's name _____ ID _____
Date of observation _____ Instructor _____
Patient's name _____ HN _____
Diagnosis _____
The number of previous attempts (circle the appropriate number): 0 1 2 3 4 5

Part 2: To be completed by the instructor
Please grade the students' skills using the following criteria

Gowning and gloving skills	Excellent	Adequate	Poor	Not applicable
1. Proper clipping of nails with no dirt beneath the nails				
2. Removal of watches, rings, and jewelry				
3. Constantly holding hands and forearms about chest level				
4. Correct usage of hand towel to dry hands and forearms				
5. Grasping the neck of gown and let the gown drip down properly				
6. Put both arms through the gown				
7. Put on sterile gloves (open/closed technique)				
8. Tie the rope that secures the gown to the body				
9. Proper sterile technique				
10. Proper disposal of towel and glove envelope				

Overall evaluation of students: ☐ Good pass ☐ Pass ☐ Borderline ☐ Fail
Total procedure time: _____ min Signature: _____

Direct Observation of Procedural Skills Form A
Department of Surgery, Faculty of Medicine Siriraj Hospital
Wound dressing

Part 1: To be completed by the student
Student's name _____ ID _____
Date of observation _____ Instructor _____
Patient's name _____ HN _____
Diagnosis _____ Type of wound _____
The number of previous attempts (circle the appropriate number): 0 1 2 3 4 5

Part 2: To be completed by the instructor
Please grade the students' skills using the following criteria

Wound dressing	Excellent	Adequate	Poor	Not applicable
1. Obtain verbal consent from the patient				
2. Remove old dressing material from the patient				
3. Open the dressing set				
4. Clean wound surrounding with appropriate solution				
5. Apply proper solution to the wound				
6. Remove tissue debris from the base of wound				
7. Apply appropriate dressing material to the wound				
8. Apply adhesive tape to secure the dressing in place				
9. Dispose used dressing material in proper way				
10. Proper sterile technique				

Overall evaluation of students: ☐ Good pass ☐ Pass ☐ Borderline ☐ Fail
Total procedure time: _____ min Signature: _____

การประเมินโดยผู้ร่วมงานและผู้เกี่ยวข้อง (Multisource feedback)

วิธีการ

การประเมินรอบด้านโดยผู้ร่วมงานและผู้ที่เกี่ยวข้องในหลายระดับ

การประเมินรายบุคคล

ผู้ประเมิน

สังเกตการณ์และให้ข้อมูลย้อนกลับ

สิ่งที่มุ่งประเมิน

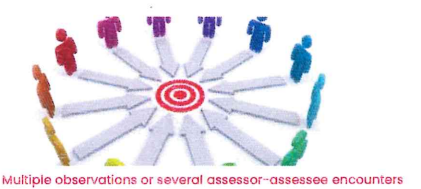
ความรู้ ควบคุมหรือประเมินความสามารถในการตัดสินใจทางคลินิก

ใช้ร่วมกับการประเมินอื่น ๆ ประเมินพฤติกรรมหลักด้านการสื่อสาร

ความเป็นมืออาชีพ การทำงานร่วมกับผู้อื่น

แบบประเมิน

แบบสอบถามที่ครอบคลุมพื้นฐานของการสังเกตพฤติกรรม



- healthcare professionals
- patients

Uu C. An introduction to workplace-based assessments. Gastroenterol Hepatol Bed Bench 2012;5:24-8.

Please refer to curriculum or some instructions for details of expected competencies for F1 and F2

mini-PAT (Peer Assessment Tool) – F1 Version

Student completed this assessment using a video ☒ Please use step 10 and 11 as a guide

Doctor: _____
Examiner: _____

OSCE Number: _____

How do you rate this Doctor in their:	Exceeds expectations for F1 completion	Meets expectations for F1 completion	Does not meet expectations for F1 completion	Does not meet expectations for F1 completion	Does not meet expectations for F1 completion
General Clinical Skills					
1. Ability to diagnose common conditions					
2. Ability to formulate appropriate management plan					
3. Awareness of their own limitations					
4. Ability to respond to professional aspects of illness					
5. Appropriate selection of resources and evidence synthesis					
Maintaining good medical practice					
6. Ability to manage time effectively					
7. Adherence to professional conduct					
Teaching and Learning, Supervising and Mentoring					
8. Anticipation and awareness of others' needs and interests					
9. Communication with patients					
10. Communication with colleagues					
11. Respect for privacy and confidentiality					
Working with colleagues					
12. Interpersonal communication skills					
13. Ability to integrate and use the contribution of others					
14. Responsibility					
15. Overall, how do you rate this doctor compared to a junior doctor in their specialty?					

Do you have any concerns about this doctor's safety or health? ☐ Yes ☐ No

- Professionalism
- Interpersonal and Communication Skills
- Interprofessional Teamwork
- Systems-based Practice

06

Feedback Importance

Role of Feedback

Improvement Through Feedback



Feedback is crucial for improving communication skills, professionalism, and interpersonal relationships.



Potential Issues

Requirements for Effective WPBA



WPBA requires skilled observers and structured frameworks to avoid becoming a "tick-box" exercise.



High achievers may experience the "expertise reversal effect," where step-by-step guidelines hinder their learning process.

07.

Challenges

Validity

- An assessment assesses what it purports to assess.
- Assessments are designed to assess major competencies that cannot be overtly measured, e.g., clinical reasoning.
- They are called constructs.
 - Several studies support the construct validity of mini-CEX.
 - Holmboe et al. (2004b) have shown that mini-CEX could differentiate the abilities of trainees at different stages of training.
 - A similar finding has been reported by Wilkinson et al. (2008) for both mini-CEX and DOPS, and by Beard (2005) for OSATS.
 - Durning et al. (2002) observe that mini-CEX ratings on specific competencies correlated with the monthly faculty ratings on corresponding competencies

Reliability

Reliability is the degree to which an assessment result (or observed score) matches with the true score or the actual candidate ability.

- Recent study has shown that five raters for mini-CEX, and two to three raters for DOPS are sufficient (Wilkinson et al. 2008).
- Extra raters may only be necessary for trainees with borderline ratings (Norcini 2010, p. 233).
- Pelgrim et al. (2011), based on a literature review, have generalized that 10 single observation encounters are sufficient to achieve acceptable reliability.

Van der Vleuten, C.P.M. (1996) The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ Theory Pract* 1: 41–67

Educational impact

Educational impact refers to the ability of an assessment to drive the learner towards desirable learning.

- Studies have indicated that mini-CEX and related assessment promote deep learning (Alves de Lima et al. 2005) and reflection (Cruess et al. 2006).
- There is also early work that indicates the diminishing educational benefit of mini-CEX, in the absence of reflective feedback (Soemantri et al. 2012).

Van der Vleuten, C.P.M. (1996) The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ Theory Pract* 1: 41–67

Table 46.2 Utility of WBA instruments

Assessment instrument	Validity	Reliability	Educational impact	Acceptability	Practicability	Cost-effectiveness
Mini-CEX	Construct validity.	With 4–6 raters per year.	High	Moderate to low—mainly due to the time taken.	Generally good, as they are in widespread use, but poor response rate is a concern.	Their widespread use suggests that they are cost effective, even when assessor time is accounted for.
DOPS	Construct validity.	With 3–5 raters per year.	High			
CBD (mainly for CSR)	Construct validity.	With 4–6 per year.	Low			
Peer assessment e.g. mini-PAT	Face, consequential, and construct validity.	With 8–13 raters per year.	High. Depends on specialty and quality of feedback.	Generally low—due to the high number of forms that needs to be collected.		
Patient surveys	May be.	With 20–40 patients per year.	Not much evidence.			
Portfolios	Face, content, and construct validity.	Adequate reliability possible, needs appropriate criteria.	High if reflection is practised.	Improving, but time taken to document evidence is a concern.		

Ponnamperuma, Gominda, 'Workplace based assessment', in Kieran Walsh (ed.), *Oxford Textbook of Medical Education*, Oxford Textbook (Oxford, 2013; online edn, Oxford Academic, 1 Oct. 2013), <https://doi.org/10.1093/med/9780199652679.003.0046>, accessed 24 Mar. 2024

08

Conclusion

Summary of WPBA Effectiveness

Value in Medical Education



WPBA is a valuable tool for assessing competencies in real- world clinical settings and addressing gaps in traditional education systems.



It emphasizes practical skills, communication, and professionalism while providing actionable feedback to improve performance.

• Comparison: WBAs vs Conventional Summative Assessment

01 WBAs

- **Assessment Tools**
- **Timing**
- **Impact on Trainees**
- **Generalization**
- **Environment**

02 Conventional Summative

- MCQs, essays, OSCE.
- End of training period.
- Limited information (marks/grades).
- Underperformance recognized late.
- Results limited to textbook knowledge.
- High- stakes.

Conclusion

- ❑ WBA, when implemented **properly**, has many benefits.
- ❑ Validity can be improved by better sampling of assessment material, using an assessment **blueprint**.
- ❑ Reliability could be improved by collecting a **sufficient number of ratings** before assessment decisions are made.
- ❑ The formative value of WBA should not be compromised in the pursuit of **summative requirements**.
- ❑ The major **research** focus in the coming decade should be on improving rating rubrics, assessor training, and feedback.
- ❑ The workplace is not a place for complex ratings. The success of WBA depends on **keeping it simple**.

Thank You