Iramaneerat C. The Standards for Educational and Psychological Testing: Part I [Thai]. Medical Education Pamphlet 2006; 2(7): 3.

Iramaneerat C. The Standards for Educational and Psychological Testing: Part II [Thai]. Medical Education Pamphlet 2006; 2(8): 2.

มาตรฐานการประเมินคุณภาพของการสอบ

เชิดศักดิ์ ไอรมณีรัตน์

ในบทความนี้ผมจะขอแนะนำมาตรฐานที่ใช้ในการประเมินคุณภาพของการจัดสอบที่ใช้กันอย่างแพร่หลาย มาตรฐานนี้คือ The Standards for Educational and Psychological Testing ซึ่งตั้งขึ้นโดย The American Educational Research Association, The American Psychological Association, และ The National Council on Measurement in Education ฉบับที่เป็นที่ใช้กันอยู่ในปัจจุบันคือฉบับปี 1999 มาตรฐานนี้มีความละเอียดมากและครอบคลุมทุกประเด็นของการจัดสอบในทุกระดับ อย่างไรก็ตามบทความนี้จะไม่แจกแจงรายละเอียดของมาตรฐานนี้ แต่จะมุ่งเน้นประเด็นสำคัญที่เป็นหลักพื้นฐานในการพัฒนามาตรฐานการจัดสอบเพียงประเด็นเดียว หลักพื้นฐานนี้คือ “Validity”

Validity คือการประเมินคุณค่าของการแปลผลการสอบและการนำผลการสอบไปใช้ ประเด็นสำคัญคือเราประเมินคุณค่าของการนำผลสอบไปใช้ มิใช่ตัวข้อสอบ ข้อสอบที่วัดความรู้ทางภาษาอังกฤษได้ดี หากเอาคะแนนไปใช้ตัดเกรดวิชาออร์โธปิดิคส์ก็เป็นการใช้ผลสอบที่ไม่มีคุณค่า ข้อสอบที่วัดความรู้ทางศัลยศาสตร์ของนักศึกษาแพทย์ได้ดี อาจไม่เหมาะกับการวัดความรู้ทางศัลยศาสตร์ของแพทย์ประจำบ้าน ประเด็นสำคัญคือเราไม่สามารถประเมินคุณค่าของข้อสอบได้หากไม่พิจารณาสภาพแวดล้อมของข้อสอบนั้นๆร่วมด้วย

การประเมินคุณค่าของการแปลผลการสอบและการนำผลสอบไปใช้ทำโดยการตั้งสมมติฐานว่าผลสอบที่ได้นั้นสามารถบอกอะไรเราได้บ้าง แล้วทำการรวบรวมข้อมูลต่างๆที่เกี่ยวเนื่องกับการสอบนั้นๆมาสนับสนุนสมมติฐานที่ตั้งขึ้น ข้อมูลที่รวบรวมเพื่อการนี้เรารวมเรียกกันว่า validity evidence ในปัจจุบันการประเมินคุณภาพของการนำผลสอบไปใช้นั้นพิจารณาหลักฐานสำคัญจาก 5 แหล่งคือ (1) เนื้อหา (Content), (2) กระบวนการที่ใช้ในการตอบข้อสอบ (Response process), (3) โครงสร้างของข้อสอบ (Internal structure), (4) ความสัมพันธ์ของคะแนนสอบกับตัวแปรอื่นๆ (Relations to other variables), และ (5) ผลที่เกิดขึ้นจากการนำผลสอบไปใช้ (Consequences)

1. <u>การประเมินเนื้อหาข้อสอบ (Test content)</u>

หัวใจสำคัญของการประเมินเนื้อหาข้อสอบคือการพิจารณาความคลอบคลุมของเนื้อหาของสิ่งที่ต้องการประเมิน (Construct) ดังนั้นก่อนที่จะประเมินคุณค่าของเนื้อหาข้อสอบได้ต้องระบุให้แน่ชัดก่อนว่าต้องการประเมินความรู้หรือความสามารถของนักเรียนในด้านใด แล้วพิจารณาว่าข้อสอบได้วัดความรู้ครบในทุกประเด็นของเรื่องนั้นหรือไม่ โดยทั่วไปกระบวนการนี้มักทำโดยการศึกษาจาก Test blueprint ซึ่งเป็นโครงร่างของข้อสอบว่ามีข้อสอบในเรื่องต่างๆ อย่างละกี่ข้อ มีประเด็นสำคัญใดหรือไม่ที่ไม่มีข้อสอบครอบคลุม

นอกจากการพิจารณาความครอบคลุมของเนื้อหาโดยรวมแล้ว เรายังต้องพิจารณาในรายละเอียดของข้อสอบแต่ละข้ออีกด้วย การประเมินคุณค่าของข้อสอบแต่ละข้อนั้นทำโดยการตรวจว่าข้อสอบข้อนั้นๆสามารถวัดความรู้ในเนื้อหาที่ต้องการได้ตามที่ต้องการหรือไม่ มีปัจจัยอื่นที่ทำให้นักเรียนตอบถูกโดยไม่จำเป็นต้องใช้ความรู้หรือไม่ มีปัจจัยใดที่อาจทำให้นักเรียนที่มีความรู้ตอบข้อสอบผิดหรือไม่ ข้อสอบข้อนั้นยากหรือง่ายเกินไปหรือไม่ ในการสอบด้วยข้อสอบ multiple-choice question นั้นสิ่งต่างๆเหล่านี้สามารถตรวจสอบได้โดยการวิเคราะห์คำตอบของนักเรียนว่ามีนักเรียนตอบข้อสอบนั้นถูกกี่คน คนที่ตอบข้อสอบข้อ

นั้นถูกทำคะแนนรวมได้เท่าไร ในกลุ่มนักเรียนที่ตอบข้อสอบผิดนั้นเลือกตัวเลือกใดบ้าง ตัวลวงใดไม่มีนักเรียนเลือกเลย ตัวลวงใดมักลวงให้นักเรียนที่มีคะแนนสูงเลือก ตัวลวงใดมักลวงให้นักเรียนที่มีคะแนนต่ำเลือก เป็นต้น

2. <u>การประเมินกระบวนการที่ใช้ในการตอบข้อสอบ (Response process)</u>

การตรวจสอบกระบวนการที่ใช้ในการตอบข้อสอบนั้นกระทำเพื่อให้เกิดความมั่นใจว่านักเรียนได้ใช้ความรู้ที่ต้องการประเมินในการทำข้อสอบจริง ไม่ได้ใช้วิธีการอื่นใดในการได้มาซึ่งคำตอบ ในการสอบ multiple-choice examination การตรวจสอบนี้เริ่มตั้งแต่การตรวจสอบคำถามที่ใช้ในข้อสอบว่าไม่มีส่วนใดของโจทย์ที่บอกใบ้ให้นักเรียนรู้คำตอบโดยไม่ต้องใช้ความรู้ เช่นหากโจทย์ถามเกี่ยวกับการวินิจฉัยโรคจากลักษณะผู้ป่วยที่บรรยาย นักเรียนควรต้องใช้ความรู้เกี่ยวกับอาการและอาการแสดงของโรคนั้นๆในการเลือกตัวเลือกที่ถูกต้อง ไม่ใช่ใช้การพิจารณาจากรูปประโยคแล้วใช้ความรู้ทางภาษาในการเลือกคำตอบที่เข้าได้กับรูปประโยค นอกจากจะพิจารณาการตอบถูกอย่างไม่เหมาะสมแล้ว เรายังต้องคำนึงถึงการตอบผิดอย่างไม่เหมาะสมด้วย นักเรียนที่มีความรู้จริงไม่ควรตอบข้อสอบผิดเนื่องจากปัจจัยที่ไม่เกี่ยวเนื่องกับความรู้ที่ต้องการวัด เช่น โจทย์มีความกำกวม ตัวเลือกที่ถูกต้องพิมพ์ผิด ข้อสอบที่ใช้รูปประกอบ แต่รูปไม่ชัดเจน หรือข้อสอบที่มีตัวเลือกที่ถูกมากกว่า 1 ตัวเลือก เป็นต้น

นอกจากจะพิจารณาคุณภาพของข้อสอบแต่ละข้อแล้ว การประเมินคุณภาพจากกระบวนการที่ใช้ในการตอบข้อสอบยังหมายรวมถึงการจัดสอบด้วย การจัดสอบต้องมีมาตรการควบคุมที่ดี ไม่ให้มีปัจจัยอื่นมารบกวนการคิดของนักเรียน เช่น จัดให้มีแสงสว่างในห้องสอบเพียงพอ ไม่ให้ห้องสอบร้อนอบอ้าว หรือ ไม่หนาวเกินไป ไม่มีเสียงรบกวนจนนักเรียนไม่มีสมาธิในการสอบ การจัดเวลาที่ใช้สอบก็ต้องพอเหมาะ หากจัดเวลาให้นักเรียนน้อยเกินไป ถึงนักเรียนมีความรู้ก็ไม่สามารถตอบข้อสอบได้เนื่องจากถูกบีบด้วยเวลาที่จำกัดทำให้ต้องเดาข้อสอบในช่วงหลังเนื่องจากหมดเวลา ในทางกลับกันผู้จัดสอบก็ต้องควบคุมไม่ให้นักเรียนได้มาซึ่งคำตอบโดยวิธีการที่ไม่ถูกต้องด้วย โดยใช้มาตรการป้องกันการทุจริตในการสอบในรูปแบบต่างๆ

สำหรับการสอบในรูปแบบอื่นที่ต้องใช้กรรมการให้คะแนน เช่น การสอบ OSCE การสอบข้อสอบบรรยาย หรือการสอบปากเปล่า กระบวนการให้ได้มาซึ่งคะแนนสอบนั้นมีความซับซ้อนมากขึ้นเนื่องจากมี อาจารย์ผู้ให้คะแนนสอบเข้ามาร่วมด้วย ในการสอบรูปแบบเหล่านี้ต้องมีการควบคุมมาตรฐานของผู้ให้คะแนนให้มีความสม่ำเสมอ ไม่ให้มีปัจจัยอื่นมาบิดเบือนเกณฑ์การให้คะแนนเช่น ลายมือของนักเรียนที่ตอบข้อสอบ ความเหนื่อยล้าของกรรมการคุมสอบ อารมณ์ของผู้ตรวจข้อสอบ เป็นต้น

ดังได้กล่าวแล้วว่าการประเมินคุณภาพนี้เน้นที่การนำผลสอบไปใช้ ดังนั้นกระบวนการจึงไม่จบลงตรงที่การตอบข้อสอบแต่ยังครอบคลุมไปถึงกระบวนการให้คะแนน และการแปลผลจากคะแนนที่ได้ด้วย ในการให้คะแนนข้อสอบ multiple-choice examination ต้องมีการตรวจสอบที่ดีว่า key เฉลยคำตอบถูกต้อง หากมีข้อสอบหลายส่วนและต้องมีการคำนวนคะแนนรวมก็ต้องตรวจสอบว่ากระบวนการรวบรวมคะแนนทำอย่างถูกต้อง การรายงานผลสอบก็ต้องทำอย่างเหมาะสม ควรมีการบรรยายประกอบว่าคะแนนที่ได้สามารถบอกอะไรได้บ้างเพื่อป้องกันไม่ให้มีการนำผลสอบไปใช้ในทางที่ไม่เหมาะสม

3. <u>การประเมินโครงสร้างของข้อสอบ (Internal structure)</u>

การพิจารณาโครงสร้างของข้อสอบนั้นหมายถึงการตรวจสอบคุณลักษณะของคะแนนสอบด้วยวิธีการทางคณิตศาสตร์และสถิติบางประการเพื่อให้มั่นใจว่าคะแนนสอบที่ได้มานั้นเป็นมาตรวัดความรู้ที่ต้องการที่มีความถูกต้อง เที่ยงตรง และเป็นธรรม ไม่มีความผิดพลาด คลาดเคลื่อน หรือ บิดเบือน กระบวนการวิเคราะห์ในขั้นตอนนี้ค่อนข้างซับซ้อน และในบางประเด็นยังเป็นที่ถกเถียงกันในหมู่ผู้เชี่ยวชาญว่าวิธีการวิเคราะห์แบบใดจะเหมาะสมกว่ากัน แต่โดยสรุปแล้วมีวิธีการวิเคราะห์อยู่

หลักๆ 2 แนวทางด้วยกัน คือ (1) Classical test theory และ (2) Item response theory ทั้ง 2 ทฤษฎีนี้อธิบายความสัมพันธ์ระหว่างคะแนนที่ได้มาจากการสอบกับความสามารถของนักเรียนผู้สอบแตกต่างกัน ดังนั้นผลการประเมินที่ได้อาจมีความแตกต่างกันบ้าง Classical test theory เป็นทฤษฎีดั้งเดิมที่ใช้ง่าย ได้รับการคิดค้นขึ้นตั้งแต่ราว ปี 1907 และเป็นที่แพร่หลายมาเป็นระยะเวลานาน ส่วน Item response theory เป็นทฤษฎีใหม่ที่ได้รับการคิดค้นขึ้นราว ปี 1960 และเริ่มเป็นที่นิยมกันมากขึ้นในการคิดคะแนนสอบในปัจจุบัน โดยเฉพาะอย่างยิ่งในการจัดสอบด้วยระบบ computer นั้นล้วนใช้การวิเคราะห์คะแนนด้วย item response theory ทั้งสิ้น

สิ่งแรกที่ต้องคำนึงถึงในการวิเคราะห์คุณลักษณะของคะแนนสอบคือ ความแม่นยำของคะแนนสอบ (reliability) ซึ่งบ่งบอกว่า หากทำการสอบซ้ำนักเรียนจะได้คะแนนเท่าเดิมหรือไม่ ซึ่งมักบอกด้วยค่า reliability coefficient ซึ่งมีค่าตั้งแต่ 0 – 1 ยิ่งการสอบมีความสำคัญมาก เราก็ยิ่งต้องการคะแนนสอบที่มีความแม่นยำมาก ค่า reliability coefficient ก็ต้องสูงมากขึ้น โดยทั่วไปสำหรับการสอบย่อยๆ ในชั้นเรียนโดยทั่วไป มักยอมรับผลสอบที่มีค่า reliability coefficient มากกว่า 0.7 สำหรับการสอบที่มีความสำคัญปานกลางเช่น การสอบลงกองของนักศึกษาแพทย์ การสอบปลายภาค หรือการสอบใหญ่ต่างๆ ในโรงเรียนแพทย์ มักต้องการค่า reliability coefficient ที่มากกว่า 0.8 สำหรับการสอบที่มีความสำคัญมาก เช่น การสอบคัดเลือกเข้าเรียนมหาวิทยาลัย การสอบใบอนุญาตประกอบวิชาชีพเวชกรรม การสอบวุฒิบัตรผู้เชี่ยวชาญเฉพาะทาง มักต้องการค่า reliability coefficient ที่มากกว่า 0.9

นอกจากการพิจารณาความแม่นยำของคะแนนสอบแล้วยังมีปัจจัยอีกหลายอย่างที่ต้องคำนึงถึง เช่น dimensionality (คะแนนสอบบ่งบอกถึงความสามารถด้านเดียวหรือหลายด้าน), fit analysis (มีข้อสอบข้อใดหรือไม่ที่ให้คะแนนแปลกไป ไม่เข้ากับข้อสอบโดยรวม เช่นนักเรียนที่ได้คะแนนดีมักตอบผิด แต่นักเรียนที่ได้คะแนนไม่ดีมักตอบถูก), item dependency (คะแนนของข้อสอบข้อหนึ่งมีความสัมพันธ์กับคะแนนสอบข้ออื่นมากผิดปกติหรือไม่), differential item functioning (มีข้อสอบข้อใดหรือไม่ที่นักเรียนกลุ่มใดกลุ่มหนึ่งได้คะแนนมากกว่านักเรียนกลุ่มอื่นอย่างมีนัยสำคัญ) ฯลฯ

4. <u>การประเมินความสัมพันธ์ของข้อสอบกับตัวแปรอื่น (Relations to other variables)</u>

เราสามารถประเมินคุณภาพของคะแนนสอบได้ด้วยการศึกษาความสัมพันธ์ระหว่างคะแนนสอบกับตัวแปรอื่น การศึกษาความสัมพันธ์นี้ทำได้ใน 2 ลักษณะคือ

(1) convergent validity ซึ่งหมายถึงการตรวจพบความสัมพันธ์ระหว่างคะแนนสอบกับตัวแปรที่วัดความสามารถในด้านเดียวกัน เช่น หากพบว่า คะแนนสอบ in-training examination ของแพทย์ประจำบ้าน มีความสัมพันธ์กับคะแนนสอบ multiple-choice examination ของการสอบวุฒิบัตรแพทย์เฉพาะทาง ก็เป็นการบ่งบอกว่า in-training examination มีคุณภาพดี

(2) discriminant validity ซึ่งหมายถึงการตรวจพบว่าคะแนนสอบไม่มีความสัมพันธ์กับตัวแปรที่วัดความสามารถในด้านอื่น เช่น หากพบว่า คะแนนสอบ OSCE เพื่อประเมินความสามารถในการตรวจคนไข้ของนักเรียนแพทย์ ไม่มีความสัมพันธ์กับคะแนนสอบ multiple-choice examination ในการสอบลงกองของนักเรียน ก็เป็นการบ่งบอกว่า OSCE นั้นวัดความสามารถทางคลินิกของนักเรียนที่แตกต่างไปจากความรู้ที่วัดจากการสอบ multiple-choice examination

5. <u>การประเมินผลที่เกิดขึ้นจากการนำผลสอบไปใช้ (Consequences)</u>

การประเมินผลที่เกิดขึ้นจากการนำผลสอบไปใช้นั้นหมายรวมถึงทั้งการศึกษาผลที่เกิดขึ้นทั้งโดยตั้งใจและไม่ตั้งใจ ทั้งในระยะสั้น และ ระยะยาว การประเมินคุณภาพของการสอบในด้านนี้ทำได้โดยการวิเคราะห์ความสัมพันธ์ของผลพวงของการสอบกับดัชนีต่างๆ เช่น การศึกษาอัตราการสอบผ่าน การเปรียบเทียบอัตราการสอบผ่านของการสอบหนึ่งกับการสอบอื่นที่วัดความสามารถในด้านเดียวกัน ความแม่นยำของการตัดสินผลสอบได้ – สอบตกของนักเรียน เป็นต้น สิ่งที่สำคัญในประเด็นนี้คือการตรวจสอบว่าการสอบนั้นไม่ส่งผลเสียหายต่อผู้ใด หรือองค์กรใดอันเนื่องมาจากผลสอบที่ไม่ถูกต้อง ไม่แม่นยำ หรือไม่เป็นธรรม ยกตัวอย่างเช่น การนำผลสอบของนักเรียนที่มีค่า reliability coefficient เพียง 0.7 ไปตัดสินว่านักเรียนคนใดต้องเรียนซ้ำชั้น จัดว่าเป็นการนำผลสอบไปใช้ที่ไม่เหมาะสม เพราะจะมีนักเรียนจำนวนหนึ่งที่ได้รับผลเสียหายจากการสอบที่ไม่แม่นยำเพียงพอ หรือการคัดเลือกนักเรียนเข้าโรงเรียนแพทย์โดยดูจากคะแนนสอบวัดความรู้ของวิชาที่เรียนในระดับมัธยมปลายเพียงอย่างเดียวอาจเป็นการใช้ผลสอบที่ไม่เหมาะสมกับข้อสอบนัก เนื่องจากลักษณะเนื้อหาของวิชาที่เรียนในระดับมัธยมปลายมีความแตกต่างจากวิชาที่เรียนในโรงเรียนแพทย์ในหลายประการ นักเรียนที่ได้คะแนนสูงเนื่องจากความสามารถในการคำนวณทางคณิตศาสตร์และฟิสิกส์อาจประสบปัญหาในการเรียนวิชาในโรงเรียนแพทย์ที่ไม่เน้นความสามารถในการคำนวณ เป็นต้น นอกจากนี้ความสามารถทางวิชาการก็มิใช่คุณลักษณะประการเดียวที่จำเป็นในการเรียนแพทย์ ยังมีคุณลักษณะที่สำคัญอีกหลายด้านที่การสอบวัดความรู้ทางวิชาการไม่สามารถบอกได้ เช่น จริยธรรม ความรับผิดชอบ ความเสียสละ ฯลฯ การนำผลสอบวัดความรู้ระดับมัธยมปลายไปตัดสินว่านักเรียนคนใดควรเรียนแพทย์หรือไม่จึงอาจเป็นการนำผลสอบไปใช้มากเกินขอบเขตความสามารถของข้อสอบ