06

Item analysis of Multiple-choice questions

การวิเคราะห์ข้อสอบปรนัย

รศ. ดร. นพ.เชิดศักดิ์ ไอรมณีรัตน์

ผู้อำนวยการศูนย์ความเป็นเลิศด้านการศึกษาวิทยาศาสตร์สุขภาพ คณะแพทยศาสตร์ศิริราชพยาบาล มหาวิทยาลัยมหิดล

การวิเคราะห์ข้อสอบ (item analysis) เป็นการใช้วิธีการทางสถิติเพื่อวิเคราะห์คำตอบที่ผู้เข้าสอบตอบเพื่อประเมินคุณสมบัติของข้อสอบว่าทำงานได้ตามที่คาดหวังหรือไม่ การวิเคราะห์ข้อสอบเป็นศาสตร์ที่ได้รับการพัฒนาอย่างต่อเนื่องมาตั้งแต่ต้นศตวรรษที่ 20 และมีการเสนอเทคนิคใหม่ออกมาเป็นระยะๆ เพื่อศึกษาคุณสมบัติที่ลึกซึ้งขึ้นของข้อสอบ อย่างไรก็ตาม เนื่องจากเทคนิคที่มีการใช้งานในวงการแพทยศาสตรศึกษากันอย่างกว้างขวางกันในปัจจุบันเป็นการวิเคราะห์ข้อสอบแบบพื้นฐาน ด้วยแนวคิดของ Classical test theory ในบทความนี้ผู้เขียนจะขอกล่าวถึงเฉพาะเทคนิคการวิเคราะห์ข้อสอบตาม Classical test theory เท่านั้น เพื่อมุ่งหวังให้ผู้อ่านทุกท่านสามารถแปลผลรายงานการวิเคราะห์ข้อสอบปรนัยที่ใช้กันทั่วไปในโรงเรียนวิทยาศาสตร์สุขภาพไทย และนำไปสู่การพัฒนาการสอบปรนัยในปัจจุบันได้

การวิเคราะห์ข้อสอบปรนัย โดยทั่วไปสามารถแยกออกได้เป็นสองส่วนหลักๆ ได้แก่ การวิเคราะห์คุณสมบัติของข้อสอบทั้งชุด (test statistics) และการวิเคราะห์คุณสมบัติข้อสอบรายข้อ (item statistics)

การวิเคราะห์คุณสมบัติของข้อสอบทั้งชุด (Test statistics)

สิ่งที่ผู้จัดสอบต้องตรวจสอบขั้นพื้นฐานสำหรับการสอบทุกครั้ง คือ **ความเที่ยง (reliability)** ของคะแนนสอบ ซึ่งเป็นดัชนีที่บอกว่า คะแนนที่ได้มานั้นมีความคลาดเคลื่อนมากน้อยเพียงใด หากวัดผลซ้ำในผู้เข้าสอบที่มีความสามารถเท่าเดิม คะแนนที่ได้จะคงเดิมหรือไม่ การวัดความเที่ยงของคะแนนสอบปรนัยที่นิยมทำกันที่สุดคือการหา internal consistency reliability โดยมีหลักการพื้นฐานคือ ในการสอบข้อสอบชุดหนึ่งๆ เป้าหมายคือ การวัดความสามารถของผู้เข้าสอบหนึ่งอย่าง (unidimensional construct) แม้ว่ารายละเอียดของสิ่งที่ทำการวัดผลนั้นจะมีหลายองค์ประกอบย่อย แต่ทุกข้อในข้อสอบชุดเดียวกันนั้นมุ่งไปหาวัตถุประสงค์ใหญ่อันเดียวกัน ดังนั้นคะแนนจากข้อสอบแต่ละข้อในชุดเดียวกันควรมีความสัมพันธ์เชิงบวกกัน (positive correlation) ซึ่งสูตรที่ใช้ในการหา internal consistency reliability สำหรับข้อสอบปรนัย ที่ใช้กันอย่างกว้างขวางคือ Kuder-Richardson Formula 20 (KR-20)

$$KR-20 = (\frac{n}{n-1})(1-\frac{\sum pq}{Var})$$

เมื่อ n คือ จำนวนข้อสอบ
Var คือ Variance ของคะแนนสอบทั้งชุด
p คือ สัดส่วนของผู้สอบที่ตอบข้อสอบข้อนั้นถูก
q คือ สัดส่วนของผู้สอบที่ตอบข้อสอบข้อนั้นผิด

โดยค่า KR-20 มีค่าระหว่าง 0 – 1 โดยค่าสูงแสดงถึงคะแนนสอบมีความเที่ยงสูง แสดงว่าข้อสอบทุกข้อในชุดข้อสอบนั้นวัดผลไปในทางเดียวกัน ประเด็นที่สำคัญคือ ค่าความเที่ยงสูงเพียงใดจึงจะพอ ซึ่งผู้เชี่ยวชาญในการวัดผลจะให้พิจารณาว่าการสอบนั้นๆ มีความสำคัญต่อผู้เข้าสอบมากน้อยเพียงใด หากเป็นการสอบที่มีความสำคัญมาก (high-stakes examination) เช่น การสอบขอรับใบประกอบวิชาชีพเวชกรรม หรือประกาศนียบัตรแพทย์ผู้เชี่ยวชาญเฉพาะสาขา มักต้องการความเที่ยงไม่ต่ำกว่า 0.9 หากเป็นการสอบที่มีความสำคัญปานกลาง (medium-stakes examination) เช่น การสอบ summative test ปลายภาคเพื่อตัดสินเกรดและเลื่อนชั้น มักต้องการความเที่ยง 0.8 – 0.89 ส่วนการสอบที่มีความสำคัญน้อย (low-stakes examination) เช่น การสอบ formative test มักต้องการความเที่ยง 0.7 – 0.79

การวิเคราะห์คุณสมบัติของข้อสอบรายข้อ (Item statistics)

การวิเคราะห์ข้อสอบรายข้อโดยทั่วไป พิจารณาคุณสมบัติสามประการ ได้แก่ ความยากง่าย ความสามารถในการจำแนก และการทำงานของตัวลวง

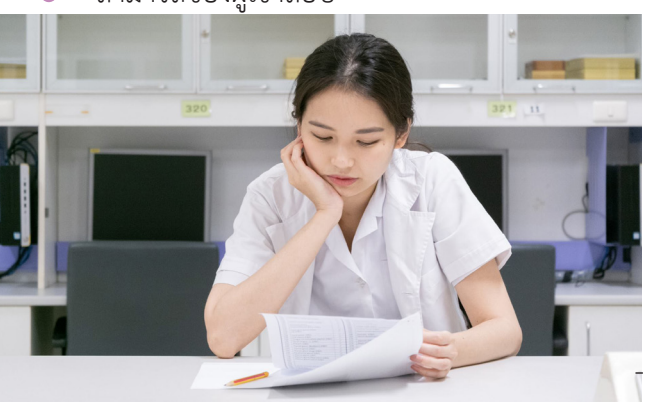
1 ความยากง่าย Item difficulty (p)

การวัดความยากง่ายของข้อสอบทำโดยหาสัดส่วนของผู้สอบที่ตอบข้อสอบถูกต่อผู้ที่ตอบข้อสอบทั้งหมด (proportion of examinees answering items correctly, p)

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

เมื่อ C คือ จำนวนผู้สอบที่ตอบข้อสอบถูก (correct)
I คือ จำนวนผู้สอบที่ตอบข้อสอบผิด (incorrect)

หากข้อสอบข้อนั้นง่าย มีจำนวนผู้ตอบข้อสอบถูกมาก ค่า p จะสูง เพื่อให้ได้ข้อมูลที่จะเป็นประโยชน์ต่อการตัดสินผลสอบมากที่สุด ข้อสอบที่ดีควรมีค่า p อยู่ในช่วง 0.45 – 0.75 ข้อสอบที่ค่อนข้างง่ายแต่ยังได้ข้อมูลที่มีประโยชน์สูงพอควร จะมีค่า p ในช่วง 0.76 – 0.91 ในทางตรงข้าม ข้อสอบที่ค่อนข้างยากแต่ยังน่าจะพอยอมรับได้จะมีค่า p ในช่วง 0.25 – 0.44 ส่วนข้อสอบที่มีค่า p ที่ต่ำกว่า 0.25 เป็นข้อสอบที่ยากมาก และอาจเป็นไปได้ว่าอาจเฉลยคำตอบผิด ส่วนข้อสอบที่ค่า p สูงกว่า 0.91 เป็นข้อสอบที่ง่ายมาก จนอาจไม่มีประโยชน์มากนักในการแยกแยะความสามารถของผู้เข้าสอบ



2 ความสามารถในการจำแนก Item discrimination (r)

ความสามารถในการจำแนกคือ ความสามารถของข้อสอบในการแยกผู้สอบที่ทำคะแนนรวมได้ดีจากผู้สอบที่ทำคะแนนรวมได้น้อย ข้อสอบที่มีความสามารถในการจำแนกสูงคือ ข้อสอบที่ผู้สอบที่ตอบถูกมักได้คะแนนรวมสูง แต่ถ้าตอบผิดมักได้คะแนนรวมต่ำ ดัชนีในการจำแนกที่ใช้กันมากที่สุดสำหรับข้อสอบปรนัยในปัจจุบันคือ point-biserial correlation (r)

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

เมื่อ Mp คือ คะแนนรวมเฉลี่ยของผู้สอบที่ตอบข้อสอบถูก
Mq คือ คะแนนรวมเฉลี่ยของผู้สอบที่ตอบข้อสอบผิด
SD คือ ค่าเบี่ยงเบนมาตรฐาน (standard deviation) ของคะแนนสอบ
p คือ สัดส่วนของผู้สอบที่ตอบข้อสอบถูกต่อผู้สอบทั้งหมด
q คือ สัดส่วนของผู้สอบที่ตอบข้อสอบผิดต่อผู้สอบทั้งหมด

ค่า point-biserial correlation (r) มีค่าอยู่ในช่วง -1 ถึง 1 โดยค่าที่ติดลบหมายถึงผู้ที่ตอบข้อสอบข้อนั้นถูกต้อง มักมีคะแนนรวมต่ำ แต่ผู้ที่ตอบข้อสอบผิด กลับได้คะแนนสูง ในทางตรงข้าม ข้อสอบที่มีค่า r สูง เป็นข้อสอบที่ผู้ตอบถูกมีคะแนนรวมสูง แต่ผู้ตอบผิดมีคะแนนรวมต่ำ ข้อสอบที่ดีควรมีค่า r สูงกว่า 0.20 ข้อสอบที่พอใช้ได้ค่าอยู่ในช่วง 0.1 – 0.19 ส่วนข้อสอบที่มีค่า r เป็น 0 หรือติดลบเป็นข้อสอบที่ไม่ค่อยดี และอาจเฉลยคำตอบผิด

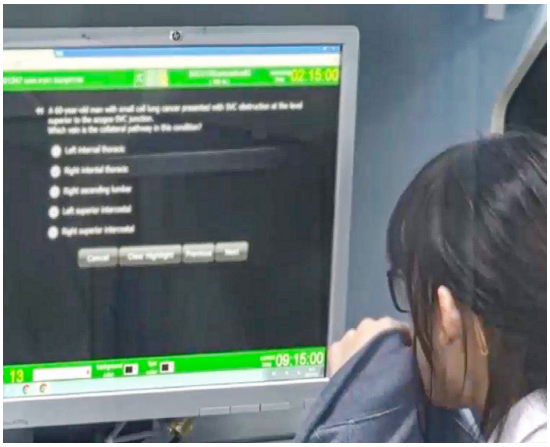
3 การทำงานของตัวลวง
Distractor functionality

ตัวลวงที่มีประสิทธิภาพดี ควรมีคุณสมบัติสองประการคือ

3.1 มีผู้สอบเลือกตัวลวงนั้นไม่ต่ำกว่าร้อยละ 5 ของจำนวนผู้สอบทั้งหมด

3.2 มีค่า point-biserial correlation (r) ของตัวลวงนั้นเป็นลบ กล่าวคือ ผู้สอบที่เลือกตัวลวงเป็นคำตอบ เป็นผู้ที่มีคะแนนต่ำ ในขณะที่ผู้สอบที่ไม่เลือกตัวลวงเป็นคำตอบ (มีความรู้ดี) มีคะแนนรวมสูง หากตัวลวงใดมีค่า r เป็นบวก โดยเฉพาะเมื่อ r สูงกว่าตัวเลือกที่เฉลยเป็นคำตอบ ควรได้รับการทบทวนว่ามีการเฉลยคำตอบผิดหรือข้อสอบนั้นมีคำตอบที่ถูกต้องมากกว่าหนึ่งตัวเลือกหรือไม่

ตัวลวงใดที่มีผู้สอบเลือกน้อย หรือลวงให้ผู้ที่มีความรู้ดีมาเลือก จัดเป็นตัวลวงที่ทำหน้าที่ไม่ดี ควรได้รับการทบทวนในการปรับเปลี่ยนหรือตัดทิ้ง



การประยุกต์ใช้ผลการวิเคราะห์ข้อสอบในทางปฏิบัติ

อาจารย์ผู้รับผิดชอบดูแลการจัดสอบ MCQ สามารถนำผลการวิเคราะห์ข้อสอบมาใช้ประโยชน์ได้หลายประการ เช่น

(1) ตรวจสอบความถูกต้องของผลสอบก่อนประกาศคะแนน

ในการจัดสอบบางครั้งอาจมีข้อสอบที่เฉลยคำตอบผิด หรือมีตัวเลือกที่ถูกต้องมากกว่าหนึ่งได้ อาจารย์สามารถตรวจสอบหาข้อสอบที่มีค่า p ต่ำมาก หรือมีค่า r ติดลบ และนำมาทบทวนเนื้อหาข้อสอบว่า โจทย์มีความคลุมเครือ หรือตัวเลือกไม่เหมาะสมหรือไม่ ควรมีคณะกรรมการที่มีผู้มีความเชี่ยวชาญในเนื้อหาวิชาที่ทำการจัดสอบพิจารณาให้ความเห็นว่าจะดำเนินการอย่างไรกับข้อสอบข้อดังกล่าว หากข้อสอบเฉลยผิด กรรมการทำการแก้เฉลยให้ถูกต้องแล้วตรวจให้คะแนนข้อสอบนั้นใหม่ ในข้อสอบที่มีคำตอบที่ถูกต้องมากกว่าหนึ่งตัวเลือก กรรมการสามารถปรับเพิ่มคะแนนให้กับผู้ที่เลือกตัวเลือกที่ไม่ได้เฉลยเป็นคำตอบไว้แต่แรกแต่พบว่าสามารถเป็นคำตอบที่เหมาะสมได้ หากข้อสอบข้อใดที่มีความคลุมเครือมากจนไม่สามารถตัดสินใจเลือกคำตอบที่ดีที่สุดได้ กรรมการอาจตัดข้อสอบข้อนั้นออกจากการคิดคะแนน และปรับเกณฑ์ผ่านลดลงตามความเหมาะสม อาจารย์ที่ดูแลการสอบ MCQ ควรทำการวิเคราะห์ข้อสอบเพื่อตรวจสอบความผิดพลาดของผลสอบเหล่านี้ก่อนทำการประกาศคะแนนสอบทุกครั้ง

(2) ปรับปรุงคุณภาพข้อสอบ

การวิเคราะห์ข้อสอบเป็นโอกาสอันดีที่อาจารย์ผู้ออกข้อสอบจะได้รับ feedback ถึงคุณภาพข้อสอบ หากผลวิเคราะห์แสดงให้เห็นว่าข้อสอบยากเกินไป (p ต่ำ) หรืออำนาจจำแนกไม่ดี (r ต่ำ) อาจเกิดจากโจทย์มีความคลุมเครือ อาจารย์สามารถปรับแก้โจทย์ให้มีความชัดเจนขึ้น นอกจากนี้การวิเคราะห์การทำงานของตัวลวง อาจพบว่าตัวลวงบางตัวไม่ทำงานอย่างเหมาะสม (ไม่มีผู้เลือกเลย หรือค่า point-biserial เป็นบวก) ก็เป็นข้อมูลที่อาจารย์อาจใช้ปรับตัวเลือกของข้อสอบได้

(3) บริหารคลังข้อสอบ

ข้อมูลผลการวิเคราะห์ข้อสอบทำให้เห็นได้ชัดเจนว่าข้อสอบใดเป็นข้อสอบที่ดี มีระดับความยากง่ายเหมาะสม มีความสามารถในการจำแนกที่ดี ควรได้รับการเก็บเข้าคลังข้อสอบเพื่อจะนำมาใช้ใหม่ในอนาคต ส่วนข้อสอบที่มีปัญหาก็ควรได้รับการปรับปรุงให้ดีก่อนที่จะนำมาใช้จัดสอบใหม่ การเก็บข้อสอบเข้าระบบคลังข้อสอบพร้อมด้วยข้อมูลคุณสมบัติของข้อสอบจะเป็นประโยชน์อย่างมากต่ออาจารย์ที่ทำการคัดเลือกข้อสอบมาใช้งาน

นอกจากนี้การติดตาม item statistics อย่างต่อเนื่องของข้อสอบ อาจทำให้อาจารย์วินิจฉัยปัญหาการใช้ข้อสอบซ้ำมากเกินไป (item overexposure) ได้ด้วย ข้อสอบที่ใช้ซ้ำจนผู้เข้าสอบรู้ข้อสอบข้อดังกล่าวล่วงหน้าก่อนทำข้อสอบ จะมีค่า p สูงขึ้นผิดไปจากค่า p เดิมชัดเจน เป็นการแสดงให้เห็นว่าผู้สอบเปลี่ยนพฤติกรรมการทำข้อสอบข้อดังกล่าวจากเดิมที่เป็น problem solving กลายเป็น simple recall แทน เมื่อพบลักษณะเช่นนี้อาจารย์ควรหยุดพักการใช้ข้อสอบดังกล่าว

(4) พัฒนาคุณภาพการสอน

ข้อสอบที่สถิติมีปัญหา (p ต่ำ r ต่ำ) ที่ชี้บ่งให้อาจารย์ไปทบทวนเนื้อหาข้อสอบว่า เฉลยผิดหรือไม่ บ่อยครั้งตรวจสอบแล้วพบว่า โจทย์มีความชัดเจน เฉลยถูกต้อง ปัญหาไม่ได้อยู่ที่ข้อสอบเลย แต่เหตุที่ผู้สอบทำข้อสอบผิดเยอะเป็นเพราะผู้สอบขาดความรู้ความเข้าใจในหัวข้อดังกล่าว อาจารย์สามารถใช้ข้อมูลนี้เป็นตัวช่วยชี้แนะว่าหัวข้อการเรียนการสอนใดที่ควรมีการปรับปรุง เปลี่ยนแปลง เพื่อแก้ปัญหาความเข้าใจผิดของนักศึกษา ทำให้การสอนมีคุณภาพดีขึ้นได้

- ผู้เขียนหวังว่าเนื้อหาเกี่ยวกับการวิเคราะห์ข้อสอบปรนัยที่นำเสนอมาในบทความนี้จะเป็นประโยชน์ต่ออาจารย์ผู้ดูแลการสอน MCQ ทุกท่าน ที่จะทำให้อาจารย์สามารถใช้สถิติเหล่านี้ในการวินิจฉัยปัญหาที่เกิดขึ้นในการสอบปรนัย นำไปสู่การพัฒนาคุณภาพการสอบให้ดีขึ้น

หากอาจารย์ลองทำการวิเคราะห์ข้อสอบตามแนวทางที่เสนอในบทความนี้แล้วประสบปัญหาใดๆ สามารถติดต่อมาทางศูนย์ SHEE เพื่อขอรับคำปรึกษาได้นะครับ