โมเดลการตอบสนองข้อสอบแบบพหุภาค (Polytomous IRT Model) และ การประยุกต์ใช้กับการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (Computerized Adaptive Testing)*

ศิริเดช สุชีวะ

1. ความนำ

ในการวัดทางจิตวิทยา รูปแบบของข้อคำถามในแบบวัดส่วนใหญ่จะเป็นข้อคำถามแบบหลายตัวเลือกชนิด 5 ตัวเลือก หรือไม่ก็มาตรประมาณค่า (rating scale) แบบ 5 จุดหรือ 7 จุด ในขณะที่โมเดลการตอบสนองข้อสอบ (Item Response Theory; IRT) ซึ่งใช้ Rasch model, two-parameter logistic model, normal ogive model หรือ three-parameter logistic model จะใช้ได้กับข้อสอบที่ให้คะแนนแบบทวิภาค (dichotomous) หรือแบบ 0-1 เท่านั้น ดังนั้นเมื่อใช้โมเดลการตอบสนองข้อสอบดังกล่าว (ซึ่งเรียกว่าเป็น dichotomous IRT model) วิเคราะห์ข้อสอบแบบหลายตัวเลือกหรือแบบมาตรประมาณค่า สารสนเทศเกี่ยวกับการตอบสนองข้อสอบส่วนหนึ่งจะถูกตัดทิ้งไป โดยถ้าเป็นการวิเคราะห์ข้อสอบแบบหลายตัวเลือก ก็จะต้องกำหนดให้คะแนนสำหรับตัวเลือกที่ถูกเป็น 1 ตัวเลือกนอกนั้นได้คะแนนเป็น 0 ทั้งหมด หรือถ้าจะวิเคราะห์ข้อที่เป็นมาตรประมาณค่าก็อาจกำหนดให้การตอบในรายการ (category) ที่ 1, 2 และ 3 ได้คะแนนเป็น 0 ส่วนการตอบในรายการที่ 4, 5, 6 และ 7

*บทความนำเสนอในการสัมมนาดุษฎีบัณฑิต ภาควิชาวิจัยการศึกษา ปีการศึกษา 2539

ได้คะแนนเท่ากับ 1 เป็นต้น จึงจะทำการวิเคราะห์โดยใช้โมเดลการตอบสนองข้อสอบแบบทวิภาค (dichotomous IRT model) ได้ ซึ่งทำให้สูญเสียสารสนเทศเกี่ยวกับแบบแผนการตอบส่วนหนึ่งไปอย่างเสียดาย

ในช่วงระยะ 25 ปีที่ผ่านมา ได้มีการคิดค้นและพัฒนาโมเดลการตอบสนองข้อสอบแบบพหุภาค (polytomous IRT model) ขึ้นมาอย่างต่อเนื่อง polytomous IRT model เข้ามาช่วยแก้ปัญหาความไม่เหมาะสมระหว่างข้อมูลกับโมเดลการวิเคราะห์ โดยทำให้การตอบแต่ละตัวเลือกมีฟังก์ชันการตอบสนอง (response function) ของแต่ละตัวเลือกเอง ที่เรียกว่า operating characteristic function สารสนเทศเกี่ยวกับการตอบสนองข้อสอบทั้งหมดจึงยังคงอยู่ในการวิเคราะห์ตลอดกระบวนการ

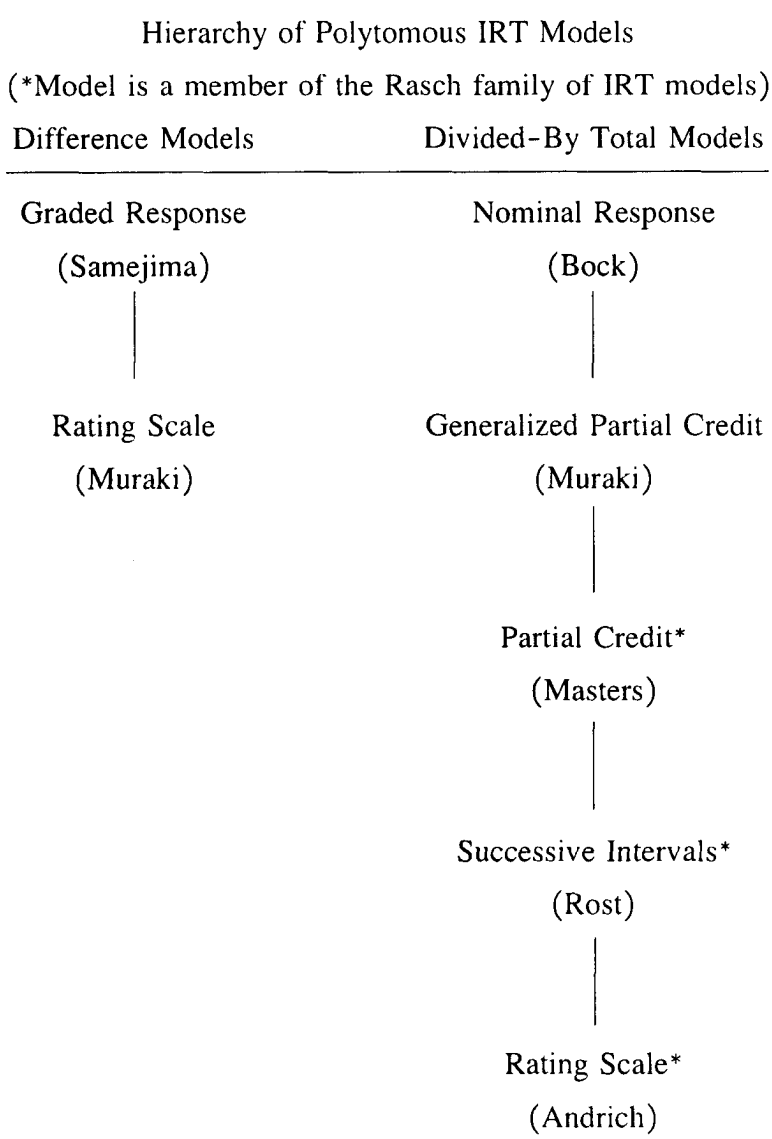
polytomous IRT model มีบทบาทในการวัดทางจิตวิทยามากขึ้น โดยนักวิจัยหลายท่านได้ใช้เครื่องมือวัดที่ให้คะแนนการตอบแบบพหุภาค มากกว่าแบบทวิภาคในการวัดคุณลักษณะภายในของบุคคล เช่น Bock (1972) Thissen (1986) แต่แม้ว่า polytomous IRT model จะมีประโยชน์กว้างขวาง แต่พัฒนาการเชิงทฤษฎีของโมเดลนี้ไปได้ช้ากว่า dichotomous IRT model นอกจากนั้นการประยุกต์ polytomous IRT model ในทางปฏิบัติเพื่อแก้ปัญหาทางการทดสอบที่สำคัญ ๆ ก็มีไม่มากนักในช่วงทศวรรษ 1970 และ 1980 แต่เมื่อมีการพัฒนาโปรแกรม MULTILOG โดย Thissen (1988) ขึ้นการใช้ polytomous IRT model ก็มีความเป็นไปได้มากขึ้นในหมู่นักวัดผลและแนวโน้มทางการศึกษาใหม่ ๆ เช่น authentic assessment ก็ให้ความสำคัญกับการให้คะแนนแบบพหุภาคมากขึ้น

อนึ่งในเรื่อง polytomous IRT model นี้ มีความเห็นที่หลากหลายว่าควรจะใช้คำว่า polytomous หรือ polychotomous จึงจะถูกต้องกันแน่ ข้อสรุปขั้นต้นในตอนนี้ ได้จากคำอธิบายของ Gideon Mellenberg (อ้างถึงใน Weiss, 1995) ที่ว่า คำว่า dichotomous มาจากภาษากรีกว่า dicho ซึ่งแปลว่า สอง กับ tomous ซึ่งแปลว่าการตัด และคำว่า polytomous ก็มาจากภาษากรีกว่า polus ซึ่งแปลว่า มาก กับคำว่า tomous เช่นกัน ดังนั้น คำที่ถูกต้องน่าจะเป็น dicho-tomous กับ poly-tomous มากกว่า

2. โมเดลการตอบสนองข้อสอบแบบพหุภาค (polytomous IRT model)

polytomous IRT model ได้มีการพัฒนาการมากว่า 25 ปี ซึ่ง Thissen และ Steinberg (1986) ได้จัดสารบบจำแนก polytomous IRT model ไว้ 3 รายการ อันได้แก่

difference model, divide-by-total model และ left-side added divide-by-total model ซึ่งประกอบด้วย multiple-choice model ของ Thissen and Steinberg, Samejima's multiple-choice model และ Sympson's model VI แต่โมเดลเหล่านี้ไม่ได้ใช้ในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (computerized adaptive testing : CAT) จึงไม่ขอกล่าว ณ ที่นี้ ส่วนรายละเอียดของ difference model และ divide-by-total model สรุปได้ดังภาพที่ 1 โดยในแต่ละรายการโมเดลที่อยู่บนสุดจะเป็นโมเดลที่มีความทั่วไปมากที่สุด ดังนี้



ภาพที่ 1

2.1 Difference Models

ใน difference model จะใช้วิธีการลบ เพื่อหาค่าโอกาสของการตอบในรายการต่าง ๆ ซึ่งได้แก่ Samejima's graded response model (GRM) และ Muraki's rating scale model (MRSM) เป็นต้น ใน difference model ทั้ง GRM และ MRSM โอกาสของการตอบในรายการหนึ่ง ๆ คำนวณจากการลบโอกาสของการตอบในรายการที่กำหนดจากโอกาสของการตอบในรายการที่ถัดมา สมการที่แสดงโอกาสของการตอบในรายการหนึ่ง ๆ จะใช้ operating characteristic function (OCF) โดยมีรายละเอียดในแต่ละโมเดล ดังนี้

2.1.1 Graded Response Model (GRM)

GRM เหมาะที่จะใช้เมื่อการตอบในข้อหนึ่ง ๆ มีรายการที่สามารถเรียงลำดับได้มากกว่า 2 รายการขึ้นไป ซึ่งแต่ละรายการแสดงถึงระดับของความสำเร็จในการแก้ปัญหาหรือระดับของความเห็นด้วยกับข้อความที่แสดงทัศนคติ ดังนั้นการตอบในรายการที่มีลำดับต่ำกว่าย่อมแสดงถึงการมีคุณลักษณะที่มุ่งวัดโดยข้อนั้นน้อยกว่าการตอบในรายการที่มีลำดับสูงกว่า คะแนนที่ให้ในแต่ละรายการจะเป็นจำนวนเต็ม แทนด้วย x โดย x = 0,1...., m_i Samejima ได้พัฒนากระบวนการสองขั้นตอนในการหาค่าโอกาสที่บุคคลที่มีระดับหนึ่งจะได้คะแนนที่กำหนดให้ ในขั้นที่หนึ่ง โอกาสที่บุคคลจะได้คะแนน x หรือสูงกว่าในการตอบข้อที่ i แสดงได้ดังสมการ

$$P^{*}_{ix}(\theta) = \frac{\exp[Da_i(\theta - b_{ix})]}{1+\exp[Da_i(\theta - b_{ix})]}$$

โดย D คือ ค่าคงที่ของมาตรซึ่งเท่ากับ 1.7
a_i คือ พารามิเตอร์อำนาจจำแนกของข้อที่ i
θ คือ พารามิเตอร์ระดับของคุณลักษณะ
b_{ix} คือ พารามิเตอร์ความยากของรายการ ที่ x ในข้อที่ i

ในแต่ละข้อจะมีการประมาณค่า a_i 1 ค่าและเซตของความยาก 1 ชุด ขั้นที่สองเป็นการหาโอกาสของการตอบในรายการหนึ่ง ๆ ($p_{ix}(\theta)$) โดยการลบโอกาสสะสมของการตอบรายการที่อยู่ถัดไปดังสมการ

$$P_{ix}(\theta) = P^{*}_{ix}(\theta) - P^{*}_{i,x+1}(\theta)$$

2.1.2 Muraki's Rating Scale Model (MRSM)

Muraki แสดงให้เห็นว่า MRSM เป็นกรณีเฉพาะของ GRM สำหรับมาตรวัดทัศนคติ Muraki รวมพารามิเตอร์ความยาก (b_{ix}) ของ GRM กับพารามิเตอร์ตำแหน่งของข้อบนมาตร (item location) และชุดของพารามิเตอร์ค่า threshold ของมาตร (t_x) ใน MRSM โอกาสที่ผู้สอบซึ่งมีระดับ θ ที่กำหนดจะตอบในรายการที่ x หรือสูงกว่า ในข้อที่ i เป็นดังสมการ

$$P^{*}_{ix}(\theta) = \frac{\exp[Da_i(\theta - b_i + t_x)]}{1+\exp[Da_i(\theta - b_i + t_x)]}$$

ดังนั้น ภายใต้ MRSM ข้อที่มีรายการ m_i+1 จะถูกกำหนดคุณลักษณะโดยตำแหน่งของข้อนั้นบนมาตร (b_i) ค่าอำนาจจำแนก (a_i) และชุดของ threshold (t_x) ของทั้งมาตร ข้อตกลงที่ว่า t_x จะคงที่ทุกข้อนั้นสอดคล้องกับการวิเคราะห์ rating scale ทั่วไป

ในกรณีของ GRM โอกาสของการตอบในรายการที่ x ของข้อที่ i หาได้โดยการลบฟังก์ชัน $P^{*}_{ix}(\theta)$ ที่ถัดมา เมื่อข้อคำถามมีแค่ 2 รายการ (ถูกกับผิด) GRM และ MRSM จะกลายเป็นโมเดลสองพารามิเตอร์ ทำให้ GRM สามารถใช้ได้กับแบบสอบที่ให้คะแนนทั้งแบบทวิภาคและพหุภาค

2.2 Divide-by-total Models

ในโมเดลนี้โอกาสของการตอบในรายการที่กำหนดหาได้จากการหารตัวเศษด้วยผลรวมของตัวเศษของโอกาสของการตอบในทุกรายการ โดยมีรายละเอียดดังนี้

2.2.1 Nominal Response Model (NRM)

NRM พัฒนาโดย Bock (1972) เป็นโมเดลที่มีความทั่วไปมากที่สุดใน divide-by-total model และ โมเดลอื่นที่จะกล่าวต่อไปก็เป็นกรณีเฉพาะของ NRM โดย NRM ต่างจาก difference model ตรงที่ NRM สามารถใช้ได้กับข้อที่มีตัวเลือก หรือรายการที่ไม่สามารถเรียงลำดับตามระดับของความถูกต้องหรือระดับของคุณลักษณะที่ข้อนั้นมุ่งวัดได้ NRM นั้นพยายามที่จะเพิ่มความแม่นยำในการประมาณค่าความสามารถของบุคคลโดยเฉพาะคนที่มีความสามารถในระดับต่ำ ๆโดยใช้สารสนเทศจากการเลือกตัวลวงของบุคคลเหล่านี้

NRM กำหนดโอกาสที่ผู้สอบที่มีระดับ θ ที่กำหนด จะตอบในรายการที่ x ของข้อที่ i ไว้ดังสมการ

$$P_{ix}(\theta) = \frac{\exp[c_{ix} + a_{ix}\theta]}{\sum_{h=1}^{ni} \exp[c_{ih} + \alpha_{ih}\theta]}$$

โดยที่ a_{ix} คือ slope หรือ อำนาจจำแนกของรายการที่ x ข้อที่ i

c_{ix} คือ intercept ของฟังก์ชันการตอบรายการที่ x ข้อที่ i

n_i คือ จำนวนรายการของข้อที่ i ($x = 1,.....n_i$)

ความสามารถในการจำแนกผู้สอบของแต่ละรายการแทนด้วย a_{ix} ส่วน c_{ix} จะสะท้อนปฏิสัมพันธ์ระหว่างความยากและประสิทธิภาพในการจำแนกของรายการหนึ่ง ๆ ผลที่เกิดขึ้น คือ ฟังก์ชันของข้อจะประกอบด้วยอำนาจจำแนกและจุดตัด (intercept) เมื่อข้อสอบมีแค่ 2 ตัวเลือก (ให้คะแนนแบบถูกกับผิด) NRM จะกลายเป็น 2-parameter logistic model (2PLM) โดยความยากของข้อจะเท่ากับ c_{ix} หารด้วย a_{ix}

2.2.2 Partial Credit Model (PCM)

Thissen and Steinberg (1986) แสดงให้เห็นว่าค่าความชันของ NRM เพิ่มขึ้นทีละ 1.0 ดังนั้นจึงเป็นได้ที่จะใช้ NRM สำหรับข้อที่ให้คะแนนตามรายการที่เรียงลำดับได้ Master (1982) จึงได้พัฒนาเป็น Partial credit model ซึ่งก็คล้ายกับ GRM ตรงที่ PCM เหมาะสำหรับข้อที่ให้คะแนนได้เป็นขั้น ๆ ตามตัวเลือก ซึ่งสามารถแสดงโอกาสที่ผู้สอบที่มีระดับ θ ที่กำหนด จะได้รับคะแนน x ในข้อที่ i ดังนี้

$$P_{ix}(\theta) = \frac{\exp[\sum_{i=0}^{x}(\theta - b_{ik})]}{\sum_{h=0}^{mi} \exp[\sum_{k=0}^{h}(\theta - b_{i})]}$$

โดย b_{ik} เป็นความยากประจำรายการที่ k ของข้อที่ i ซึ่งจะมีจำนวนขั้นความยากอยู่ m_i ค่าสำหรับข้อที่ i และเพื่อให้ง่ายขึ้น Master กำหนดให้ b_{ik} เท่ากับ 0.0 เมื่อ k = 0 ใน PCM ขั้นรายการในแต่ละขั้นไม่จำเป็นต้องเรียงตามค่าความยาก (เช่นรายการที่ 2 อาจง่ายกว่ารายการที่ 1) โดยมีข้อตกลงเบื้องต้นว่าทุกข้อมีค่าอำนาจจำแนกเท่าเทียมกัน ดังนั้นเมื่อให้คะแนนแบบทวิภาค PCM จะกลายเป็น Rasch model ซึ่งใช้ได้กับแบบสอบที่ประกอบไปด้วยข้อที่ให้คะแนนแบบทวิภาคและพหุภาคได้

2.2.3 Andrich's Rating Scale Model (ARSM)

เมื่อ PCM ใช้ได้กับข้อคำถามแบบ Likert ซึ่งมีช่วงของการประเมินที่แน่นอน (เช่น 7 จุด) PCM ก็สามารถทำให้ง่ายขึ้นโดยทำเป็น Andrich's rating scale model (ARSM) Master และ Wright ได้กล่าวไว้ในหนังสือ Rating Scale Analysis ว่า ค่าความยากในแต่ละรายการของข้อหนึ่ง ๆ จาก PCM สามารถแยกได้เป็น 2 องค์ประกอบ คือ

$$b_{ik} = b_i + t_k$$

โดย b_i คือตำแหน่งของข้อที่ i บนมาตร (ค่ามาตร) และ t_k คือ threshold สำหรับรายการที่ k ในข้อทั้งหมด

ARSM อนุพันธ์มาจาก PCM โดยโอกาสที่ผู้สอบที่มีระดับ θ ที่กำหนด จะได้รับคะแนน x ในข้อที่ i แสดงได้ดังนี้

$$P_{ix}(\theta) = \frac{\exp[K_x + x(\theta - b_i)]}{\sum_{h=0}^{mi} \exp[K_h + h(\theta - b_i)]}$$

โดย k_x เท่ากับลบของผลรวมของ threshold

ARSM เหมือนกับ MRSM ตรงที่ ค่า t_k ประมาณขึ้นสำหรับข้อสอบทั้งชุดในขณะที่ค่ามาตรของข้อ (b_i) จะประมาณขึ้นต่างหากสำหรับแต่ละข้อ แต่ต่างจาก MRSM ตรงที่ ARSM มีข้อตกลงเบื้องต้นว่าทุกข้อมีค่าอำนาจจำแนกเท่าเทียมกัน ซึ่งเหมือนกับข้อตกลงเบื้องต้นของ PCM

2.2.4 Successive Intervals Model (SIM)

Rost (1988) ได้พัฒนา SIM ขึ้นมาให้เหมาะสำหรับการวัดทัศนคติ โดยโอกาสที่ผู้สอบที่มี θ ระดับที่กำหนด จะได้รับคะแนน x ในข้อที่ i แสดงได้ดังนี้

$$P_{ix}(\theta) = \frac{\exp\{K_x + x\theta - [xb_i + x(m - x)d_i]\}}{\sum_{h=0}^{mi} \exp\{K_h + h\theta - [hb_i + h(m_i - h)d_i]\}}$$

โดย b_i คือ พารามิเตอร์ค่ามาตร (scale value)

d_i คือ พารามิเตอร์การกระจายของข้อที่ i ซึ่งแสดงถึงระยะทางของ threshold ของข้อที่เบี่ยงเบนไปจาก threshold ของทั้งมาตร และ

k_x คือ ลบของผลรวมของ threshold จากรายการที่ 1 ถึง x

SIM เป็นกรณีเฉพาะของ PCM เช่นเดียวกับ ARSM โดยประมาณค่ามาตรหรือตำแหน่งของข้อบนมาตรแต่ละข้อ และมีชุดของ threshold ของทั้งมาตร แต่ SIM ต่างจาก ARSM ตรงที่ SIM มีค่าพารามิเตอร์อีกตัวหนึ่ง คือ d_i ของแต่ละข้อ อันแสดงถึงขนาดความแตกต่างระหว่างระยะทาง threshold สำหรับข้อ และระยะทาง threshold ของทั้งมาตร แต่เมื่อกำหนดข้อตกลงว่าค่า d_i ทุกข้อเท่ากับ 0.0 SIM ก็จะง่ายลงกลายเป็น ARSM นั่นเอง

2.2.5 Generalized Partial Credit Model (GPCM)

Muraki (1992) ได้พัฒนา PCM โดยยกเลิกข้อตกลงเบื้องต้นที่ว่าข้อสอบทุกข้อมีค่าอำนาจจำแนกเท่ากัน GPCM ได้พัฒนามาคู่ขนานกับ PCM โดยใช้ 2-parameter logistic model แทน Rasch model ฟังก์ชันของ GPCM แสดงได้ดังสมการ

$$P_{ix}(\theta) = \frac{\exp[\sum_{k=0}^{x} a_i(\theta - b_{ik})]}{\sum_{h=0}^{mi} \exp[\sum_{k=0}^{h} a_i(\theta - b_{ik})]}$$

โดยที่ $P_{ix}(\theta)$ คือ โอกาสที่ผู้สอบที่มีระดับ θ ที่กำหนด จะได้รับคะแนน x ในข้อที่ i ซึ่งมี m_i+1 รายการ

a_i คือ พารามิเตอร์อำนาจจำแนกของข้อที่ i

b_{ix} คือ พารามิเตอร์ความยากของรายการที่ k ($k=1,...,m_i$)

Muraki กำหนดให้ $(-b_{ik}) = 0.0$ เมื่อ $k=0$ ซึ่งเหมือนกับ PCM ตรงที่ b_{ik} ไม่ต้องจำเป็นต้องเรียงตามลำดับ ในกรณีที่ $a_i = 1.0$ GPCM จะเหมือนกับ PCM และจากข้อตกลงที่ว่า b_{ik} สามารถแบ่งได้เป็น 2 องค์ประกอบ คือ ตำแหน่งของข้อ (item location; b_i) และ threshold ของทั้งมาตร (t_k) GPCM จะกลายเป็น ARSM นอกจากนั้น Muraki ยังแสดงให้เห็นอีกด้วยว่า GPCM เป็นกรณีเฉพาะของ NRM สำหรับข้อที่รายการสามารถเรียงลำดับไว้ด้วย

3. การเลือกใช้โมเดลการตอบสนองข้อสอบแบบพหุภาค

การเลือกใช้ polytomous IRT model ต้องคำนึงถึงองค์ประกอบหลายอย่าง เช่น ประเภทของข้อมูล ความสอดคล้องระหว่างข้อมูลกับโมเดล แนวคิดในการวิเคราะห์ และความประหยัด ถ้าข้อคำถามมีตัวเลือกที่ไม่สามารถเรียงลำดับได้ก็เหมาะที่จะใช้ NRM แต่ถ้าข้อนั้นมีรายการมากกว่า 2 รายการขึ้นไปที่สามารถเรียงรายการตามระดับของคุณลักษณะที่ข้อนั้นมุ่งวัดได้ ก็อาจจะใช้ GRM, GPCM, หรือ PCM แต่ถ้าข้อมูลที่เรียงลำดับนี้เป็น rating scale ก็น่าจะใช้โมเดลที่ค่อนข้างเฉพาะเจาะจงอย่าง MRSM, SIM หรือ ARSM หรืออาจเลือกใช้โมเดลจากการคำนวณ likelihood ratio เพื่อตรวจสอบความเหมาะสมระหว่างข้อมูลกับโมเดลแต่ละโมเดล และสามารถทดสอบความแตกต่างระหว่าง likelihood ratio ของแต่ละโมเดลได้ หากโมเดล 2 โมเดลมีความเหมาะสมกับข้อมูลไม่แตกต่างกันอย่างมีนัยสำคัญ เราก็ควรเลือกใช้โมเดลที่ง่ายกว่า เป็นต้น

4. งานวิจัยเกี่ยวกับ CAT ที่ใช้ polytomous IRT model

การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (Computerized Adaptive Testing: CAT) เป็นนวตกรรมใหม่ของการวัดผลโดยอาศัยการประยุกต์ IRT ซึ่งมีข้อได้เปรียบการสอบแบบใช้กระดาษ-ดินสอ (paper-pencil test) อยู่หลายประการ ที่สำคัญคือ CAT จะใช้การบริหารแบบสอบโดยออกแบบคัดเลือกข้อสอบให้มีความเหมาะสมระหว่างระดับความยากกับระดับของคุณลักษณะของผู้สอบที่ประมาณได้ CAT จึงใช้จำนวนข้อสอบน้อยกว่าแต่มีความแม่นยำในการวัดเท่ากับหรือสูงกว่าแบบสอบฉบับเต็ม CAT ส่วนมากพัฒนามาจากแบบสอบหลายตัวเลือกที่ให้คะแนนแบบทวิภาค ซึ่งเริ่มใช้แพร่หลายในหน่วยงานทางการทดสอบต่าง ๆ ของอเมริกา เช่น ETS ได้พัฒนา Graduate Record Examination แบบที่เป็น adaptive test ขึ้นมาใช้ American Society of Clinical Pathologist, National Council of State Board of Nursing และ American Board of Internal Medicine ก็ได้วิจัยเพื่อที่จะนำ CAT มาใช้ในการสอบเพื่อรับประกาศนียบัตร หรือกระทรวงกลาโหมของสหรัฐอเมริกาก็ได้นำ Armed Services Vocational Aptitude Battery แบบที่เป็น CAT มาใช้ และในการวิจัยเกี่ยวกับ CAT ในระยะหลัง ๆ มีแนวโน้มที่จะใช้ polytomous IRT model มากขึ้นเรื่อย ๆ

CAT มีองค์ประกอบหลักที่สำคัญ 4 องค์ประกอบ ได้แก่ คลังข้อสอบ กระบวนการคัดเลือกข้อสอบ กระบวนการประมาณค่าคุณลักษณะ และกฎการหยุด (stopping rule) งานวิจัยเกี่ยวกับเรื่องนี้จึงแบ่งออกเป็น 4 กลุ่มตามองค์ประกอบของ CAT ดังต่อไปนี้

4.1 งานวิจัยเกี่ยวกับคลังข้อสอบ

ขนาดของคลังข้อสอบและคุณลักษณะของข้อสอบในคลัง มีผลต่อคุณภาพของการทดสอบแบบปรับเหมาะมาก ถ้าข้อสอบมีการให้คะแนนแบบทวิภาค และใช้การวิเคราะห์ตาม 3-parameter logistic model ผลการวิจัยแนะนำว่าควรมีจำนวนข้อสอบอย่างน้อย 100 ข้อ (Dodd, 1995) ยิ่งถ้าเป็นการสอบที่มีความสำคัญด้วยแล้วคลังข้อสอบก็ควรจะมีขนาดใหญ่ประมาณ 500-1000 ข้อ

ปัญหาที่พบบ่อยในคลังข้อสอบขนาดเล็ก คือ การประมาณค่าจะไม่คงที่เมื่อประมาณโดยใช้ maximum likelihood หรือถ้าประมาณได้ก็จะมีความคลาดเคลื่อนมาตรฐานสูง แต่ถ้าเป็น polytomous IRT model ผลการวิจัยจะพบแตกต่างไป โดยพบว่าคลังข้อสอบที่มีข้อสอบเพียง 30 ข้อก็สามารถประมาณค่าได้อย่างมีความแม่นยำ เมื่อใช้โมเดล GRM PCM SIM และ ARSM และพบว่าในบริบทของการวัดทัศนคติแบบ Likert คลังข้อสอบที่มีข้อสอบเพียง 24 ข้อก็สามารถทำงานได้อย่างมีประสิทธิผลโดยใช้โมเดล PCM (Kock and Dodd, 1985 อ้างถึงใน Dodd, 1995) และ ARSM (Dodd, 1990)

อย่างไรก็ตามข้อค้นพบเหล่านี้มิได้แสดงว่า คลังข้อสอบควรมีข้อสอบ 30 ข้อหรือมากกว่า แล้วจะเพียงพอสำหรับ CAT ที่ใช้ polytomous IRT model คุณลักษณะของข้อสอบที่ประกอบกันขึ้นเป็นคลังจะมีผลต่อความสำเร็จของระบบ CAT มากกว่า ในปี 1993 Dodd และคณะ (อ้างถึงใน Dodd, 1995) รายงานว่าคลังข้อสอบที่มีข้อสอบ 30 ข้อ ได้ผลดีสำหรับ CAT ที่ใช้ PCM ถ้าฟังก์ชันสารสนเทศของคลังข้อสอบมีจุดสูงสุดอยู่ใกล้ ๆ กับตำแหน่ง $\theta = 0.0$ นอกจากนั้น คลังข้อสอบขนาดเล็กจะมีปัญหาในทางปฏิบัติเกี่ยวกับความตรงเชิงเนื้อหา การคัดเลือกข้อสอบ และความปลอดภัยของคลังข้อสอบ ดังนั้นในการสอบที่สำคัญ ๆ ควรใช้คลังข้อสอบขนาดใหญ่ดีกว่า

การที่พบว่าคลังข้อสอบขนาดเล็กใช้ได้ผลดีสำหรับ polytomous IRT model เนื่องจากความจริงที่ว่าการให้คะแนนแบบพหุภาคจะให้สารสนเทศมากกว่าการให้คะแนนแบบทวิภาคอยู่แล้ว ซึ่งไม่เพียงแต่จะให้ระดับสารสนเทศสูงกว่าเท่านั้น สารสนเทศที่ได้ก็ยังมีการแจกแจงที่กว้างครอบคลุมพิสัยของคุณลักษณะที่มุ่งวัด รายการที่อยู่ติดกันแต่ละคู่ของข้อที่ให้คะแนนแบบพหุภาคจะเหมือนกับข้อที่ให้คะแนนแบบทวิภาคข้อหนึ่ง ดังนั้นสารสนเทศที่ได้จากข้อที่ให้คะแนนแบบพหุภาคจึงมีส่วนร่วมในฟังก์ชันสารสนเทศของคลังข้อสอบรวมมากกว่าข้อที่ให้คะแนนแบบทวิภาค

ข้อจำกัดที่สำคัญของการวิจัยเกี่ยวกับคลังข้อสอบใน polytomous IRT model ก็คือคลังข้อสอบที่ศึกษามักได้จากการจำลองขึ้นมามากกว่าที่จะเป็นข้อมูลจริง แม้ว่าการจำลองคลังข้อสอบมาจะมีประโยชน์ในการจัดกระทำต่อพารามิเตอร์ที่รู้ค่าแล้วในการศึกษาตัวแปรที่สนใจได้สะดวก แต่ก็ควรจะมีการวิจัยโดยใช้ข้อสอบจริงและผู้สอบจริงด้วย

4.2 งานวิจัยเกี่ยวกับกระบวนการคัดเลือกข้อสอบ

เป้าหมายของการคัดเลือกข้อสอบใน CAT คือการคัดข้อสอบที่จะให้สารสนเทศสูงสุด ณ ระดับ θ ของผู้สอบที่ประมาณค่าได้ในขณะนั้นจากคลังข้อสอบ โดยระบบ CAT ส่วนใหญ่จะใช้ฟังก์ชันสารสนเทศของข้อสอบ (item information function) เป็นพื้นฐานของการคัดเลือกข้อ จากการศึกษาพบว่าใน CAT ที่ใช้ polytomous IRT model ประเภท GRM, NRM, และ PCM การคัดเลือกข้อสอบโดยใช้ฟังก์ชันสารสนเทศของข้อสอบจะได้ผลดีมาก การวิจัยของ De Ayala (1992) พบว่าการใช้สารสนเทศของ "รายการ" แทนสารสนเทศของ "ข้อ" ในกระบวนการคัดเลือก จะทำให้ได้ข้อสอบน้อยลง 1 ข้อโดยเฉลี่ย โดยใช้ NRM-CAT ที่จำลองขึ้น จนปัจจุบันก็ยังไม่มีการศึกษาเพื่อตรวจสอบการใช้สารสนเทศของรายการในการคัดเลือกข้อสอบใน CAT นอกจากนั้นก็ยังไม่ได้ศึกษากระบวนการคัดเลือกข้อสอบแบบอื่น ๆ นอกเหนือไปจากสารสนเทศของรายการหรือของข้อในโมเดลเหล่านี้

ทางเลือกใหม่ของการคัดเลือกข้อสอบใน ARSM และ SIM ได้แก่ การคัดเลือกด้วยค่ามาตรที่ใกล้กับระดับ θ ที่ประมาณได้ของผู้สอบมากที่สุด มีการศึกษา 2 เรื่องที่เปรียบเทียบระหว่างกระบวนการคัดเลือกด้วยค่ามาตรที่ใกล้กับระดับ θ ที่ประมาณได้ของผู้สอบมากที่สุด กับการคัดเลือกด้วยฟังก์ชันสารสนเทศของข้อสอบที่สูงที่สุด พบว่า การคัดเลือกข้อโดยถือหลักค่ามาตรที่ใกล้เคียงกับระดับ θ ของผู้สอบที่สุด ได้ผลไม่แตกต่างไปจากวิธีคัดเลือกตามสารสนเทศของข้อสอบสูงสุด แต่ Dodd (1995) เสนอว่าควรคัดเลือกตามค่ามาตรในโมเดลเหล่านี้มากกว่า เพราะง่ายกว่าและใช้เวลาคำนวณน้อยกว่าการคัดเลือกด้วยฟังก์ชันสารสนเทศของข้อสอบที่สูงที่สุด

อย่างไรก็ดี การคัดเลือกข้อด้วยค่ามาตรหรือการคัดเลือกด้วยฟังก์ชันสารสนเทศของข้อสอบอย่างเคร่งครัดก็มักส่งผลในทางลบต่อความตรงเชิงเนื้อหาเสมอ จึงควรมีการศึกษาถึงยุทธวิธีในการคัดเลือกข้อสอบให้สอดคล้องกับบริบทของความเป็นจริงด้วย

4.3 งานวิจัยเกี่ยวกับกระบวนการประมาณค่าคุณลักษณะ

กระบวนการของ CAT ที่ใช้ dichotomous IRT model จะประมาณค่า θ ด้วย

วิธี maximum likelihood หรือวิธีของ Bayesian แบบต่าง ๆ แต่ตอนนี้การประมาณค่าที่ใช้ใน polytomous IRT CAT ทั้ง PCM, ARSM, SIM และ GRM ล้วนใช้วิธี maximum likelihood เท่านั้น ซึ่งการประมาณค่าโดยวิธีนี้ จะไม่สามารถประมาณค่า maximum likelihood ในกรณีที่ผู้สอบเลือกรายการที่ต่ำสุดหรือสูงสุดในทุกข้อ หรือถ้าประมาณได้ก็มักจะไม่คงที่ และมีค่าความคลาดเคลื่อนมาตรฐานสูง มีการวิจัยเพียง 2 เรื่องเท่านั้นที่ศึกษาวิธีของ Bayesian ในการประมาณค่า ใน polytomous IRT CAT โดย De Ayala (1992) ใช้ expected a posteriori (EAP) ในการประมาณ θ ใน CAT ที่ใช้ NRM ในปี 1995 Chen และคณะ (อ้างถึงใน Dodd, 1995) ได้เปรียบเทียบการใช้ EAP และ maximum likelihood ในการประมาณค่าใน CAT ที่ใช้ ARSM ข้อได้เปรียบของ EAP ก็คือสามารถประมาณค่าได้แม้ว่าผู้สอบจะตอบรายการต่ำสุดหรือสูงสุดในทุกข้อ

4.4 งานวิจัยเกี่ยวกับกฎการหยุด

นอกเหนือไปจากการกำหนดกฎการหยุดโดยใช้จำนวนข้อสูงสุด-ต่ำสุดแบบคงที่แล้ว ได้มีการศึกษากฎการหยุดแบบอื่น 2 กฎซึ่งได้แก่กฎการหยุดเมื่อสารสนเทศต่ำสุด (minimum information stopping rule) ซึ่ง CAT จะหยุดเลือกข้อสอบเมื่อไม่มีข้อสอบในคลังที่มีระดับสารสนเทศของข้อเท่ากับระดับต่ำสุดที่กำหนดไว้สำหรับระดับ θ ของผู้สอบที่ประมาณได้ วิธีที่สองเป็นการหยุดเมื่อค่าความคลาดเคลื่อนมาตรฐาน ณ ระดับ θ ที่ประมาณได้อยู่สูงกว่าระดับที่กำหนดไว้ การใช้กฎการหยุดเหล่านี้แม้กระบวนการประมาณค่ายังไม่หยุดเลือกข้อสอบ แต่เมื่อมีข้อสอบถึงจำนวนข้อที่ตั้งไว้ (โดยทั่วไปจะตั้งไว้ประมาณ 20 ข้อ) CAT ก็จะหยุดเลือก การเปรียบเทียบการใช้กฎการหยุดใน CAT ที่ใช้ GRM, PCM, ARSM, และ NRM พบว่า การใช้กฎการหยุดตามความคลาดเคลื่อนมาตรฐาน (standard error stopping rule) จะได้ผลดีกว่า minimum information stopping rule ในเรื่องของจำนวนข้อสอบที่คัดเลือกได้โดยเฉลี่ยที่น้อยกว่า และความถี่ของกรณีที่เกิดการไม่คงที่ของการประมาณค่า ซึ่งมีน้อยกว่าด้วย

5. ทิศทางสำหรับการวิจัยในอนาคต

การวิจัยเกี่ยวกับ polytomous IRT model-based CAT ยังอยู่ในระยะเริ่มต้น เช่นเดียวกับที่ dichotomous IRT model-based CAT ที่ได้เริ่มต้นในปลายทศวรรษ 1970 ซึ่งประเด็นการวิจัยพื้นฐานเพื่อสร้างองค์ความรู้ในเรื่องนี้ยังมีความจำเป็นอยู่มาก ก่อนที่จะนำ polytomous IRT model-based CAT ไปใช้ในทางปฏิบัติ โดยมีประเด็นที่น่าจะวิจัยดัง

ต่อไปนี้

1) ควรมีการศึกษาวิธีการของ Bayesian ในการประมาณค่า สำหรับ polytomous IRT model ต่าง ๆ และเปรียบเทียบกับการประมาณค่าแบบ maximum likelihood แบบที่ใช้กันอยู่

2) ควรประเมินการทำงานของ CAT ในประเด็นความแม่นยำของการวัด จำนวนข้อสอบที่ใช้ ความลำเอียงในการประมาณค่า และความประหยัด

3) ควรมีการพัฒนา CAT ที่ยอมให้ข้อสอบในชุดเดียวกันมีวิธีการให้คะแนนแตกต่างกันไป

4) ควรศึกษาเรื่องความสมดุลของเนื้อหาของข้อสอบที่เป็น CAT โดยอาจใช้ table of specification เพื่อควบคุมความตรงเชิงเนื้อหา และเพิ่มจำนวนข้อสอบในคลังให้เพียงพอที่จะยังคงความสมดุลของเนื้อหาตามสัดส่วนที่กำหนดเมื่อคัดเลือกข้อสอบแล้ว

บรรณานุกรม

Andrich, D. (1978a). A rating formulation for ordered response categories. ***Psychometrika***, 43, 561-573.

Andrich, D. (1978b). Application of a psychometric rating model to ordered categories which are scored with successive integers. ***Applied Psychological Measurement***, 2, 581-594.

Baker, F.B. (1992). Equating tests under the graded response model. ***Applied Psychological Measurement***, 16, 87-96.

Baker, F.B. (1993). Equating tests under the nominal response model. ***Applied Psychological Measurement***, 17, 239-251.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. ***Psychometrika***, 37, 29-51.

Bock, R.D., & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. ***Applied Psychological Measurement***, 6, 431-444.

De Ayala, R.J. (1992). The nominal response model in computerized adaptive testing. ***Applied Psychological Measurement***, 16, 327-343.

De Ayala, R.J. Dodd, B.G., & Koch, W.R. (1992). A comparison of the partial credit and graded response models in computerized adaptive testing. ***Applied Measurement in Education***, 5, 17-34.

Dood, B.G. (1990). The effect of item selection procedure and stepsize on computerized adaptive attitude measurement using the rating scale model. ***Applied Psychological Measurement***, 17, 355-366.

Dodd, B.G. (1995). Computerized adaptive testing with polytomous item. ***Applied Psychological Measurement***, 19, 5-19.

Dodd, B.G., & Koch, W.R. (1987). Effects of variations in item step values on item and test information in the partial credit model. ***Applied Psychological Measurement***, 11, 371-384.

Dodd, B.G., Koch, W.R., & De Ayala, R.J. (1989). Operational characteristics of adaptive testing procedures using the graded response model. ***Applied Psychological Measurement***, 13, 129-143.

Drasgow, F. (1995). Introduction to the polytomous IRT special issue. ***Applied Psychological Measurement***, 19, 1-3.

Green, B.R., Bock, R.D., Humphreys, L.G., Linn, R.L., & Reckase, M.D. (1984). Technical guidelines for assessing computerized adaptive tests. ***Journal of Educational Measurement***, 21, 347-360.

Henly, S.J., Klebe, J.J., McBride, J.R.,& Cudeck, R. (1989). Adaptive and conventional versions of the DAT : The first complete test battery comparison. ***Applied Psychological Measurement***, 13, 363-371.

Weiss, D. (1995). Polychotomous or polytomous?. ***Applied Psychological Measurement***, 19,4.

Thissen, D. (1976). Information in wrong responses to the Raven Progressive Matrices. ***Journal of Educational Psychology***, 13, 201-214.

Thissen, D. (1988). ***MULTILOG : Multiple, categorical item analysis and test scoring using item response theory*** (Version 5.1) [Computer program]. Mooresville IN: Scientific Software.