

การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบพหุมิติ (Multidimensional Computerized Adaptive Testing : MCAT)

เบญจมาภรณ์ เสนารัตน์¹

สมประสงค์ เสนารัตน์²

สมบัติ ท้ายเรือคำ³

การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (CAT) ได้รับความนิยมมาก ทั้งในการทดสอบทางการศึกษาและจิตวิทยา CAT เป็นวิธีการเฉพาะที่ใช้ในการประเมินความสามารถและคุณลักษณะแฝงอื่น ๆ ของผู้สอบ โดยการเลือกข้อสอบจากคลังข้อสอบให้ผู้สอบได้รับข้อสอบที่เหมาะสมกับความสามารถตามผลการตอบข้อสอบของข้อที่สอบผ่านไปแล้ว (Frey and Seitz. 2009 : 90 citing Frey. 2007) จุดมุ่งหมายของขั้นตอนการเลือกข้อสอบ คือ การเลือกข้อสอบให้เหมาะสมกับระดับลักษณะของผู้เข้าสอบ (Frey and Seitz. 2009 : 89)

ในปัจจุบันนี้ มีการทดสอบจำนวนมากที่มีลักษณะเหมือนกันกับ CAT เช่น การสอบภาษาอังกฤษในลักษณะเป็นภาษาต่างประเทศ (TOEFL) ข้อสอบวัดเชาวน์ปัญญาทั่วไปก่อนเข้าศึกษาในระดับบัณฑิตศึกษา (The Graduate Record Examinations : GREs) หรือข้อสอบที่ใช้วัดความสามารถของผู้ที่ต้องการเข้าศึกษาในระดับปริญญาโทและปริญญาเอกสาขาบริหารธุรกิจ (Graduate Management Admission Test : GMAT) ที่ใช้การบริหารการสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ นอกจากนี้ในปี 2012 องค์กรความร่วมมือและพัฒนาทางเศรษฐกิจ (OECD) มีแผนที่จะนำเอา CAT มาใช้ในการประเมินสมรรถนะของนักเรียนในโครงการประเมินนักเรียนในระดับนานาชาติ (Programme for International Student Assessment: PISA) (Frey and Seitz. 2009 : 89)

การใช้ CAT มีประโยชน์ที่โดดเด่น คือ ความสามารถในการเพิ่มประสิทธิภาพของการวัดอย่างมาก เนื่องจากในการใช้งานจะถูกกำหนดโดยการวัดอัตราส่วนของการวัดที่แม่นยำ และความยาวของแบบสอบ (Segall. 2005 ; Frey and Seitz. 2009 : 90 citing Frey. 2007) นอกจากนี้ CAT ยังสามารถลดจำนวนข้อสอบเพื่อให้ผู้เข้าสอบตอบข้อสอบด้วยจำนวนข้อที่น้อยลง แต่มีประสิทธิภาพในการวัดเพิ่มขึ้นเมื่อเปรียบเทียบกับ การทดสอบแบบดั้งเดิมที่มีจำนวนข้อสอบคงที่ในแบบสอบ (Fixed Item Test: FIT) โดยทั่วไป ในการทดสอบด้วย CAT สามารถที่จะลดจำนวนข้อสอบลงได้ครึ่งหนึ่ง โดยไม่มีการสูญเสียความแม่นยำ ในการวัด ดังนั้นการทดสอบด้วย CAT จึงช่วยให้ลดภาระงาน เวลาและค่าใช้จ่ายลงไปได้มาก (Frey and Seitz. 2009 : 89)

ดุชนิบัณฑิต สาขาวิชาวิจัยและประเมินผลการศึกษา มหาวิทยาลัยมหาสารคาม

อาจารย์ ดร. วิทยาลัยการศึกษา มหาวิทยาลัยราชภัฏร้อยเอ็ด

รองศาสตราจารย์ ดร. คณะศึกษาศาสตร์ มหาวิทยาลัยมหาสารคาม

การทดสอบแบบปรับเหมาะส่วนใหญ่ใช้คลังข้อสอบที่ได้รับการสอบเทียบด้วยทฤษฎีโมเดลการตอบสนองข้อสอบแบบมิติเดียว (Unidimensional IRT) (Frey and Seitz. 2009 : 89 citing Fischer & Molenaar. 1995 ; Hambleton & Swaminathan. 1985) แต่ภายใต้เงื่อนไขในเชิงทฤษฎีการประเมินทางการศึกษาและจิตวิทยาในปัจจุบันนี้มักจะเกี่ยวข้องกับหลายประเด็น เช่น PISA เกี่ยวข้องกับความสามารถทางคณิตศาสตร์ การอ่านและวิทยาศาสตร์ (OECD. 2006) ซึ่งเงื่อนไขดังกล่าวจะเหมาะสมกับการใช้วิธีการทดสอบแบบปรับเหมาะแบบพหุมิติ (MCAT) ด้วยโมเดลการตอบสนองข้อสอบแบบพหุมิติ (MIRT) มากกว่าการทดสอบแบบปรับเหมาะแบบมิติเดียว (CAT) ด้วยโมเดลการตอบสนองข้อสอบแบบมิติเดียว (IRT) (Frey and Seitz. 2009 : 89)

แนวคิดของ MCAT เป็นที่รู้จักกันมาในช่วง 20 ปีที่ผ่านมา ในปี 1987 Bloxom and Vale ได้นำเสนอวิธีการที่ปรับปรุงใหม่เกี่ยวกับ MCAT โดยใช้วิธีการคัดเลือกข้อสอบแบบเบสปรับปรุงใหม่ของ Owen (1975) หลังจากนั้นในปี 1992 Tam ได้ศึกษาเปรียบเทียบวิธีการคัดเลือกข้อสอบใน MCAT ระหว่างวิธีแบบเบสปรับปรุงใหม่ของ Owen (1975) กับวิธีการอื่น ๆ ในประเด็นความแม่นยำในการวัดสารสนเทศของข้อสอบและเวลาในการคำนวณ ต่อมา Segall (1996: 338-341) ได้นำเสนอวิธีการแบบเบสสำหรับ MCAT และในเวลาต่อมาวิธีการแบบเบสเป็นวิธีการที่ได้รับความสนใจมาก วิธีการนี้ใช้ความรู้เกี่ยวกับโครงสร้างของความแปรปรวนร่วมระหว่างมิติของการวัดสำหรับการเลือกข้อสอบ และการประมาณค่าความสามารถ ส่งผลให้จำนวนข้อสอบที่ลดลงซึ่งเป็นคุณสมบัติที่จำเป็นของการทดสอบแบบ CAT แนวคิดของ MCAT เป็นแนวทางใหม่ในการประเมินความสามารถด้วยเหตุผลสองประการ ประการแรก เมื่อเปรียบเทียบกับการใช้โมเดลการวัดมิติเดียว MCAT ยอมให้มีข้อจำกัดที่เหมาะสมระหว่างโมเดลเชิงทฤษฎีที่ประกอบกันขึ้นมาจากหลากหลายแนวคิดและโมเดลการวัดในเชิงสถิติ ประการที่สอง การลดลงของจำนวนข้อสอบที่สามารถเห็นได้จากการเปรียบเทียบการทดสอบด้วย MCAT กับ CAT และการทดสอบแบบดั้งเดิม (Frey and Seitz. 2009 : 89-90)

นอกจากนี้ Frey & Carstensen (2009 : 59) ยังได้กล่าวถึง MCAT ว่าเป็นการนำเอาหลักการของ CAT มาขยายขอบเขตการวัดของข้อสอบจากการวัดในมิติเดียวเป็นข้อสอบแต่ละข้อวัดพร้อมกันหลาย ๆ มิติ และถ้าใช้จำนวนข้อสอบที่เท่ากันการทดสอบด้วย MCAT จะเป็นการทดสอบที่มีประสิทธิภาพสูงกว่า CAT และการทดสอบแบบดั้งเดิม เนื่องจาก MCAT มีอัตราการวัดที่แม่นยำกว่าเมื่อใช้ข้อสอบเท่ากัน ดังนั้นการทดสอบด้วย MCAT จึงจะสามารถที่จะลดจำนวนข้อสอบลงได้มากกว่าการทดสอบด้วย CAT ประมาณ 30-50% และลดจำนวนข้อสอบได้มากกว่าการทดสอบแบบดั้งเดิมที่กำหนดจำนวนข้อสอบไว้ประมาณ 70% โดยไม่สูญเสียความแม่นยำ (Frey and Seitz. 2009: 93) และ Frey and Seitz (2008: 81) ได้กล่าวถึง MCAT ว่ามีประสิทธิภาพสูงกว่า CAT ถึง 1.3 เท่า อย่างไรก็ตาม Frey and Seitz (2009: 92) ได้ระบุว่าถ้าใช้การเลือกข้อสอบแบบสุ่ม MCAT จะมีประสิทธิภาพสูงกว่า CAT ถึง 3.7 เท่า

MCAT เป็นการมอง CAT จากมุมมองแบบพหุตัวแปร ประกอบไปด้วยองค์ประกอบสำคัญหลายองค์ประกอบ เช่น โมเดลอยู่ภายใต้ทฤษฎีตอบสนองข้อสอบแบบพหุมิติ วิธีการประมาณค่าความสามารถ วิธีการคัดเลือกข้อสอบ และเกณฑ์การยุติการสอบ ทางเลือกขององค์ประกอบเหล่านี้จะเป็นประโยชน์สำหรับสถานการณ์การใช้งานหลายสถานการณ์ เป็นกลยุทธ์สำหรับการกำหนดเนื้อหาสาระและกำหนดสิ่งที่ต้องการนำเสนอ ดังนี้ (Frey and Seitz. 2009 : 90)

โมเดลการตอบสนองข้อสอบแบบพหุมิติ (MIRT Models)

MCAT สามารถใช้งานร่วมกับโมเดล MIRT ได้หลากหลายโมเดล (Frey and Seitz. 2009 : 90) เช่น โมเดลโลจิสติกแบบพหุมิติ (Multidimensional Logistic Model) โมเดลปกติสะสมแบบพหุมิติ (Multidimensional Normal Ogive Model) โมเดลสำหรับการวัดมิติที่ไม่สามารถทดแทนกันได้ (Models for Noncompensatory Dimensional) โมเดลสำหรับการวัดการเรียนรู้และการเปลี่ยนแปลง (Models for Learning and Change) โมเดลในการระบุโครงสร้างของระดับคุณลักษณะ (Models with Specified Trait Level Structures) และโมเดลสำหรับการจำแนกกลุ่มบุคคล (Models for Distinct Classes of Persons)

โมเดลโลจิสติกแบบพหุมิติและโมเดลปกติสะสมแบบพหุมิติเป็นโมเดลที่มีการนำมาใช้งานอย่างกว้างขวางกว่าโมเดลอื่น ๆ เนื่องจากรูปแบบทางสถิติของโมเดลโลจิสติกแบบพหุมิติสามารถ บูรณาการไปใช้ ในโมเดลปกติสะสมแบบพหุมิติได้ แต่ไม่สามารถบูรณาการไปใช้กับโมเดลสำหรับการวัดมิติที่ไม่สามารถทดแทนกันได้ แม้ว่าจะมีคุณลักษณะคล้ายคลึงกัน อีกทั้งรูปแบบในเชิงสถิติของโมเดลสำหรับการวัดมิติ ที่ไม่สามารถทดแทนกันได้ยังไม่ได้รับการพัฒนาให้เพียงพอที่จะนำไปใช้งานได้ในปัจจุบันนี้ (Reckase. 2009 : 121) และโมเดลปกติสะสมมีความชันของ ICCs (Intraclass correlation coefficients) มากกว่าโมเดล โลจิสติก แต่เมื่อคูณ 1.702 เข้าไปใน Exponents ของโมเดลโลจิสติกจะทำให้ค่าความน่าจะเป็นในการตอบได้ถูกต้องของทั้งสองโมเดลมีค่าใกล้เคียงกันมากคือมีความแตกต่างกันไม่เกิน 0.01 (Reckase. 2009 : 26 citing Birnbaum. 1968 ; Zimowski. 2003 : 541) แสดงให้เห็นว่าโมเดลทั้งสองนี้สามารถบูรณาการไปใช้งานด้วยกันได้ด้วยการปรับแก้เพียงเล็กน้อย

โดยทั่วไปโมเดลการระบุความน่าจะเป็นของ MIRT สำหรับการตอบสนองกับข้อสอบที่ถูกต้อง i ($u_i = 1$) ขึ้นอยู่กับความสามารถแฝง p เมื่อ $\theta = (\theta_1, \theta_2, \dots, \theta_p)$ และคุณลักษณะของข้อสอบที่ครอบคลุม ด้วยพารามิเตอร์ข้อสอบหนึ่งพารามิเตอร์หรือมากกว่า (Frey and Seitz. 2009: 90) โดยโมเดลการระบุ ความน่าจะเป็นในการตอบที่ถูกต้องของโมเดลโลจิสติกแบบพหุมิติ มีสูตรคำนวณดังสมการ 1 (Segall. 1996 : 333; 2010 : 62)

$$P_i(\theta) \equiv Prob(U_{ij} = 1 | \theta_j) = c_i + \frac{1 - c_i}{1 + \exp[-Da_i(\theta - b_i1)]} \quad (1)$$

$$\begin{aligned} \text{เมื่อ } & -Da_i(\theta - b_i1) = -D\sum_{k=1}^p a_{ki}(\theta_k - b_i) \\ u_{ij} & = \text{ตัวแปรสุ่มที่มีสองค่า (0 และ 1) จากข้อสอบข้อที่ i จากผู้สอบคนที่ j} \\ D & = \text{ค่าคงที่ เท่ากับ 1.702} \\ a_i & = \text{เวกเตอร์ } 1 \times p \text{ ของพารามิเตอร์อำนาจจำแนกของข้อสอบข้อที่ i} \\ b_i & = \text{พารามิเตอร์ความยากของข้อสอบข้อที่ i} \\ c_i & = \text{พารามิเตอร์โอกาสการเดาของข้อสอบข้อที่ i} \\ 1 & = \text{เวกเตอร์ } p \times 1 \text{ เมื่อ p คือ จำนวนมิติ} \end{aligned}$$

โมเดลการระบุความน่าจะเป็นในการตอบที่ถูกต้องของโมเดลปกติสะสมแบบพหุมิติ ดังสมการ 2 (Bock and Schilling 2003 : 585 ; Reckase. 2009 : 95)

$$\begin{aligned} P_i(\theta) \equiv \text{Prob}(U_{ij} = 1|\theta_j) &= c_i + (1 - c_i) \frac{1}{\sqrt{2\pi}} \int_{-z_i(\theta_j)}^{\infty} \exp(-t^2/2) dt \quad (2) \\ &= c_i + (1 - c_i) \Phi[z_i(\theta_j)] \end{aligned}$$

$$\begin{aligned} \text{เมื่อ } z_i(\theta_j) &= d_i + a_i\theta_j = d_i + a_{i1}\theta_{j1} + a_{i2}\theta_{j2} + \dots + a_{ip}\theta_{jp} \\ \Phi &= \text{เมทริกซ์ความแปรปรวนร่วมของมิติขนาด } p \times p \\ c_i &= \text{พารามิเตอร์โอกาสการเดาของข้อสอบข้อที่ i} \\ d_i &= \text{Easiness Intercept ของข้อสอบข้อที่ i} \end{aligned}$$

การประมาณค่าความสามารถ (Proficiency estimation)

การประมาณค่าความสามารถของ MIRT มีเกณฑ์การประมาณค่าที่แตกต่างกันหลายแนวทาง เช่น Maximum Likelihood Criterion, Maximum a Posteriori Bayesian Criterion, Newton-Raphson Method และ Least Squares Criterion (Reckase. 2009 : 139) แต่โดยทั่วไปเกณฑ์การประมาณค่าความสามารถของ MCAT ที่นิยมมีอยู่ 2 วิธี คือ วิธีการประมาณค่าแบบความน่าจะเป็นสูงสุด (Maximum Likelihood Estimation) และวิธีการประมาณค่าแบบเบย์ (Bayesian Estimation) (Reckase. 2009 : 314) และจากข้อตกลงเบื้องต้นเกี่ยวกับความเป็นอิสระของข้อสอบและผู้สอบในทฤษฎีการตอบสนองข้อสอบ ถ้า ตัวแปรความสามารถมีความคงที่ การตอบสนองข้อสอบของผู้สอบต่อข้อสอบแต่ละข้อของแบบทดสอบถูกคาดหวังว่าเป็นอิสระจากกันในเชิงสถิติ การประมาณค่าความสามารถด้วยเทคนิคการประมาณค่าแบบความน่าจะเป็นสูงสุด (Maximum Likelihood Estimation) นำไปสู่ความไม่ชัดเจนของการประมาณค่า ของเวกเตอร์ความสามารถ $\hat{\theta} = \hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_p$ ณ จุดเริ่มต้นของการทดสอบ แต่วิธีการประมาณค่าแบบเบย์จะหลีกเลี่ยงข้อบกพร่องนี้ แต่ก็ยังไม่มีข้อสรุปเกี่ยวกับ

ปัญหาในการประมาณค่าความสามารถ ในทางปฏิบัติ จึงยังคงมีการใช้วิธีการประมาณค่าทั้งสองแบบ (Frey and Seitz. 2009: 90)

วิธีการประมาณค่าความสามารถมีหลากหลายวิธี แต่ในบทความนี้จะนำเสนอเพียง 4 วิธี ดังนี้

1. วิธีการประมาณค่าแบบความน่าจะเป็นสูงสุด (Maximum Likelihood Estimation) มีสูตรคำนวณดังสมการ 3 (segall. 2010 : 62 ; Reckase. 2009 : 140)

$$f(U_{i1} = u_{i1}, U_{i2} = u_{i2}, \dots, U_{in} = u_{in} | \theta) = L(u | \theta) = \prod_{i \in S_n} p_i(\theta)^{u_i} q_i(\theta)^{1-u_i} \quad (3)$$

เมื่อ $f(U_{i1} = u_{i1}, U_{i2} = u_{i2}, \dots, U_{in} = u_{in} | \theta)$

เป็นผลจากฟังก์ชันการตอบข้อสอบรายชื่อที่จะนำไปใช้ในการ
คำนวณในขั้นตอนการคัดเลือกข้อสอบและการให้คะแนน

$L(u | \theta)$ คือ ฟังก์ชันความน่าจะเป็น

S_n คือ ข้อสอบข้อที่ 1, 2, ..., n

$p_i(\theta)$ คือ ความน่าจะเป็นในการตอบข้อสอบถูก

$q_i(\theta)$ คือ $1 - p_i(\theta)$

วิธีการประมาณค่าแบบนี้มักจะมีปัญหาไม่สามารถประมาณค่าได้ถ้ามีจำนวนข้อของข้อสอบน้อย (Diao. 2009 : 2 ; Reckase. 2009 : 314-315) และจากการศึกษาของ Diao. (2009 : 3) โดยใช้ข้อสอบจำนวน 50 ข้อ พบว่าการประมาณค่าด้วยวิธีนี้ใช้เวลานานกว่าจะประสบความสำเร็จในการประมาณค่าความสามารถ

2. วิธีการประมาณค่าแบบเบส์ (Bayesian Estimation) มีสูตรคำนวณดังสมการ (Segall. 2010 : 60-64 ; Reckase. 2009 : 144-145)

$$f(\theta | u) = \frac{L(\theta | u)f(\theta)}{f(u)} = \frac{L(\theta | u)f(\theta)}{\int_{-\infty}^{\infty} f(u | \theta)f(\theta) d\theta} \quad (4)$$

เมื่อ $L(u | \theta)$ คือ ฟังก์ชันความน่าจะเป็น (สมการที่ 3)

$f(u)$ คือ ความน่าจะเป็นที่เกิดขึ้นภายหลังของ u (สมการที่ 6)

$f(\theta)$ คือ ความน่าจะเป็นที่เกิดขึ้นก่อนของ θ

โดยมีข้อตกเบื้องต้นว่ามีการแจกแจงเป็นแบบปกติหลายตัวแปร (Multivariate Normal) มีค่าเฉลี่ยเท่ากับ θ และมีเมทริกซ์ความแปรปรวนร่วมเท่ากับ Φ มีสูตรคำนวณดังสมการ 5

$$f(\theta) = (2\pi)^{-p/2} |\Phi|^{-1/2} \exp \left[-\frac{1}{2} (\theta - \mu) \Phi^{-1} (\theta - \mu) \right] \quad (5)$$

$$f(u) = \int_{-\infty}^{\infty} f(u|\theta) f(\theta) d\theta \quad (6)$$

จากการศึกษาของ Diao. (2009 : 3) โดยใช้ข้อสอบจำนวน 50 ข้อ พบว่าการประมาณค่าด้วยวิธีนี้ประสบความสำเร็จในการประมาณค่าด้วยเวลาที่รวดเร็ว

3. วิธีการประมาณค่าแบบนิวตันราฟ (Newton-Raphson Method) วิธีการนี้ใช้ควบคู่ไปกับการประมาณค่าวิธีแบบความน่าจะเป็นสูงสุด ในกรณีที่ไม่สามารถประมาณค่าความสามารถได้ (segall. 1996 : 335-340 ; Chen. 2009 : 1-2 ; Frey and Seitz. 2009 : 90) โดยมีวิธีการคำนวณ ได้ดังสมการ 7 (segall. 1996 : 335-340 ; Chen. 2009 : 1-2)

$$\theta^{(j+1)} = \theta^{(j)} - \delta^{(j)} \quad (7)$$

เมื่อ $\delta^{(j)}$ คือ เวกเตอร์ $p \times 1$ และคำนวณค่าได้ดังสมการ 8

$$\delta^{(j)} = -[H(\theta^{(j)})]^{-1} \left(\frac{\partial}{\partial \theta} \ln L(u/\theta) \right) \quad (8)$$

เมื่อ คือ เมทริกซ์ $p \times p$ เป็นค่าของอนุพันธ์อันดับสองของฟังก์ชันความน่าจะเป็นของ $\theta^{(j)}$ และ $H(\theta)$

มีสูตรคำนวณดังสมการ 9

$$H(\theta) = \begin{bmatrix} \frac{\partial^2 \ln L}{\partial \theta_1^2} & \frac{\partial^2 \ln L}{\partial \theta_1 \partial \theta_2} & \frac{\partial^2 \ln L}{\partial \theta_1 \partial \theta_p} \\ \frac{\partial^2 \ln L}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \ln L}{\partial \theta_2^2} & \frac{\partial^2 \ln L}{\partial \theta_2 \partial \theta_p} \\ \vdots & \vdots & \vdots \\ \frac{\partial^2 \ln L}{\partial \theta_p \partial \theta_1} & \frac{\partial^2 \ln L}{\partial \theta_p \partial \theta_2} & \frac{\partial^2 \ln L}{\partial \theta_p^2} \end{bmatrix} \quad (9)$$

และ $\frac{\partial}{\partial \theta} \ln L(u/\theta)$ เป็นค่าของอนุพันธ์อันดับหนึ่งของฟังก์ชันความน่าจะเป็นของ $\theta^{(j)}$ มีสูตรคำนวณ ดังสมการ 10

$$\frac{\partial}{\partial \theta} \ln L(\mathbf{u}|\theta) = \begin{bmatrix} \frac{\partial}{\partial \theta_1} \ln L(\mathbf{u}|\theta) \\ \frac{\partial}{\partial \theta_2} \ln L(\mathbf{u}|\theta) \\ \vdots \\ \frac{\partial}{\partial \theta_p} \ln L(\mathbf{u}|\theta) \end{bmatrix} \quad (10)$$

4. วิธีการประมาณค่าแบบกำลังสองน้อยที่สุด (Least Squares Estimation) วิธีการนี้พบได้ไม่บ่อยนักในการประมาณค่าพารามิเตอร์ของ MIRT แต่ก็ใช่วิธีที่ใช้ในโปรแกรม NOHARM มีสูตรคำนวณ ดังสมการ 11 (Reckase. 2009 : 147-148)

$$SS_\theta = \sum_{i=1}^n (u_i - P(u_i = 1 / \theta, a_i, d_i))^2 \quad (11)$$

เมื่อ SS_θ คือ ผลรวมของความแตกต่างยกกำลังสองของแต่ละเวกเตอร์ θ
 u_i คือ คะแนนที่ได้จากข้อสอบข้อที่ i (มีค่าเป็น 0 และ 1)

การคัดเลือกข้อสอบ (Item Selection)

เป็นการระบุการเลือกข้อสอบจากคลังข้อสอบเพื่อให้ผู้เข้าสอบได้ตอบข้อสอบ ซึ่งโดยหลักการการคัดเลือกข้อสอบมีเป้าหมายเพื่อลดความแปรปรวนของเวกเตอร์ความสามารถแบบพหุมิติชั่วคราว (Frey and Seitz. 2009 : 90) และวิธีการคัดเลือกข้อสอบที่ใช้กับแกนหลักของการทดสอบแบบปรับเหมาะด้วย คอมพิวเตอร์แบบพหุมิติเกือบทั้งหมดอาศัยความน่าจะเป็นสูงสุดหรือต่ำสุดจากการประมาณค่าความสามารถ θ และมีวิธีการคัดเลือกข้อสอบอยู่หลากหลายวิธี แต่ละวิธีสามารถใช้ร่วมกันกับวิธีการประมาณค่าทุกแบบ และยังสามารถใช้วิธีการคัดเลือกข้อสอบหลายวิธีร่วมกันได้อีกด้วย (Reckase. 2009 : 327) แต่ในการนำเสนอวิธีการคัดเลือกข้อสอบในบทความนี้ จะนำเสนอเพียง 3 วิธี คือ Maximize the Determinant of the Fisher Information Matrix, Largest Decrement in the Volume of the Bayesian Credibility Ellipsoid และ Maximize Kullback-Leibler Information ดังนี้

1. Maximize the Determinant of the Fisher Information Matrix การคัดเลือกข้อสอบด้วยเมทริกซ์สารสนเทศของฟิชเชอร์ อยู่ภายใต้เงื่อนไขที่ว่า การแจกแจงความสามารถที่ประมาณค่าเป็นค่าความสามารถจริงที่มีการแจกแจงแบบปกติหลายตัวแปรจากเมทริกซ์ ความแปรปรวนความแปรปรวนร่วม (Variance-Covariance Matrix) ที่เกี่ยวข้อง กับเมทริกซ์สารสนเทศของฟิชเชอร์ (Fisher Information Matrix) (Frey and Seitz. 2009 : 90 ;

Reckase. 2009 : 330) ในโครงสร้างของ MIRT สามารถมองเมทริกซ์สารสนเทศของพิชเชอร์เป็นเหมือนการวัดปริมาณของสารสนเทศการตอบสนองของข้อสอบและนำเสนอเกี่ยวกับความสามารถภายหลังการตอบข้อสอบข้อนั้นแล้ว (θ) ในทางคณิตศาสตร์ค่าความคาดหวังของอนุพันธ์ที่สองของความน่าจะเป็นลอกลิสต์ L ที่เกี่ยวข้องกับความสมารถจะมีค่าเป็นลบ (Frey and Seitz. 2009 : 90) สามารถแสดงในวิธีการทำงานแบบพหุมิติ ที่นำเสนอโดย Segall ในปี 1996 ได้ตั้งสมการ 12 (Frey and Seitz. 2009 : 90 ; Reckase. 2009 : 330 ; Segall. 1996 : 336)

$$I_{rs}(\theta, \hat{\theta}) = -E \left[\frac{\partial^2 \ln L}{\partial \theta_r \partial \theta_s} \right] \quad (12)$$

เมื่อ	θ	คือ	ความสามารถจริงของผู้สอบ
	$\hat{\theta}$	คือ	ค่าประมาณจากความน่าจะเป็นสูงสุด
	θ_r	คือ	องค์ประกอบของเวกเตอร์ความสามารถ (θ vector) ลำดับที่ r
	θ_s	คือ	องค์ประกอบของเวกเตอร์ความสามารถ (θ vector) ลำดับที่ s
	L	คือ	ฟังก์ชันความน่าจะเป็นของเซตการตอบข้อสอบจาก MCAT

เนื่องจากแต่ละองค์ประกอบของเมทริกซ์ $I(\theta, \hat{\theta})$ มาจากระดับข้อสอบจึงสามารถกำหนด เมทริกซ์สารสนเทศของข้อสอบได้ด้วย $I(\theta, u_i)$ ดังสมการ 13, 14 และ 15 (Segall. 1996 : 337)

$$I_{rr}(\theta, u_i) = \frac{\left[\frac{\partial P_i(\theta)}{\partial \theta_r} \right]^2}{P_i(\theta) Q_i(\theta)} \quad (13)$$

และ

$$I_{rs}(\theta, u_i) = \frac{\left[\frac{\partial P_i(\theta)}{\partial \theta_r} \right] \left[\frac{\partial P_i(\theta)}{\partial \theta_s} \right]}{P_i(\theta) Q_i(\theta)} \quad (14)$$

และ

$$I(\theta, u_{k+1}) = \frac{\left[\frac{\partial P_{k+1}(\hat{\theta}_k)}{\partial \theta_r} \right]^2}{P_{k+1}(\hat{\theta}_k) Q_{k+1}(\hat{\theta}_k)} \quad (15)$$

ด้วยลักษณะเฉพาะของเมทริกซ์ความแปรปรวนความแปรปรวนร่วมที่ใช้ในการประมาณค่าภายหลังการสอบข้อสอบข้อที่ k จากแบบทดสอบแบบปรับเหมาะ Σ_k เป็นค่าประมาณจากส่วนกลับของเมทริกซ์สารสนเทศ มีสูตรคำนวณดังสมการ 16 (Segall. 1996 : 338)

$$\Sigma_k = \{I(\theta, \hat{\theta}_k)\}^{-1} = [\sum_{i \in v} I(\theta, u_i)]^{-1} \quad (16)$$

และ $\Sigma_{k+(k+1)}$ เป็นการประมาณค่าของเมทริกซ์ความแปรปรวนร่วมชั่วคราวของข้อสอบข้อที่ k รวมกับ ข้อสอบข้อที่ $k+1$ มีสูตรคำนวณดังสมการ 17 (Segall. 1996 : 338)

$$\Sigma_{k+(k+1)} = [I(\theta, \hat{\theta}_k) + I(\theta, u_{k+1})]^{-1} \quad (17)$$

เกณฑ์การเลือกข้อสอบที่เสนอโดย Segall (1996) จะขึ้นอยู่กับความสัมพันธ์ระหว่างเมทริกซ์สารสนเทศของพิชเชอร์กับบริเวณความเชื่อมั่นที่ครอบคลุมการประมาณค่าความสามารถที่กำหนดโดย Anderson (1984) และ Anderson พบว่ามีความเท่าเทียมกับช่วงความเชื่อมั่นที่ครอบคลุมการประมาณค่าความสามารถจากตำแหน่งอ้างอิงจากรูปทรงรีในพหุพื้นที่และปริมาณของรูปทรงรีนี้ก็คือ ฟังก์ชัน Σ_k และ Segall (1996) แสดงให้เห็นว่าเมื่อ Σ_k มีค่ามากที่สุดแล้วปริมาณของช่วงความเชื่อมั่นของการประมาณค่าความสามารถบริเวณรูปทรงรีจะมีขนาดเล็กที่สุด

สมการ 18 เป็นดีเทอร์มิแนนต์ของเมทริกซ์เทอมทางด้านซ้ายมือคือเมทริกซ์สารสนเทศของข้อสอบที่ได้ทดสอบไปแล้วด้วยข้อสอบข้อที่ k ส่วนเทอมทางด้านขวาเป็นเมทริกซ์สารสนเทศของข้อสอบข้อที่จะถูกนำไปใช้สอบข้อต่อไป กระบวนการคัดเลือกข้อสอบที่จะนำไปสอบข้อถัดไปดำเนินการโดยการระบุว่าข้อสอบข้อใดจะมีเมทริกซ์สารสนเทศสูงสุดเมื่อรวมกับเมทริกซ์สารสนเทศของข้อสอบข้อที่ผ่านมา โดยจะสรุปจากผลรวมของการดีเทอร์มิแนนต์ที่มีค่ามากที่สุด

$$|I(\theta, (\hat{\theta}_k)) + I(\theta, u_{k+1})| \quad (18)$$

2. Largest Decrement in the Volume of the Bayesian Credibility Ellipsoid วิธีการประมาณค่าแบบเบส์ที่นำเสนอโดย Segall (1996 : 338-342, 2000 : 61-72, 2010 : 65-74) นำไปสู่การคัดเลือกข้อสอบที่จะลดปริมาณของข้อสอบได้อย่างมาก (Frey and Seitz. 2009 : 90 ; Reckase. 2009 : 331-332) ด้วยการเลือกข้อสอบที่มีค่าเมทริกซ์สารสนเทศภายหลังสูงสุด $I_{i/s_{k-1}}$ โดยใช้ข้อมูลจาก 3 แหล่ง คือ อินเวอร์สของเมทริกซ์ความแปรปรวนร่วมที่เกิดขึ้นก่อน (Inverse Prior Covariance Matrix : Φ^{-1}) , ผลรวมของเมทริกซ์ W ที่ได้มาจาก

ข้อสอบข้อที่ผ่านมา ($W_{s_{k-1}}$) และเมทริกซ์ W ที่ได้มาจากข้อสอบข้อที่จะถูกนำมาใช้สอบข้อต่อไป มีสูตรคำนวณดังสมการ 19

$$I_{i/s_{k-1}} = \Phi^{-1} + W_{s_{k-1}} + W_i \quad (19)$$

ในการประมาณค่าความสามารถชั่วคราวของ $\hat{\theta}_k$ (เมื่อ $k=0$) ด้วยการกำหนดเซตของค่าเฉลี่ยของการแจกแจงความสามารถเบื้องต้น และโดยทั่วไปมักจะกำหนดให้มีค่าเท่ากับ 0 ขั้นตอนแรกต้องดำเนินการคำนวณหาค่าอินเวอร์สของเมทริกซ์ความแปรปรวนร่วมที่เกิดขึ้นก่อน (Φ^{-1}) เนื่องจากต้องใช้ในขั้นตอนอื่น ๆ ที่จะดำเนินการต่อไป ขั้นตอนที่สองคำนวณหา $W_{s_{k-1}}$ จากสมการ 20

$$W_{s_{k-1}} = \sum_{j \in S_{k-1}} w_j \quad (20)$$

เมื่อ W_j ของข้อสอบข้อที่ j คำนวณได้จากสมการ 18 และผลรวมของเมทริกซ์ W ที่ได้มาจากข้อสอบข้อที่ผ่านมา $\sum_{j \in S_{k-1}} w_j$ คำนวณจากข้อสอบทุกข้อที่ได้รับการเลือกมาแล้ว ขั้นตอนที่สามคำนวณหา $W_{s_{k-1}}$ ของเมทริกซ์ ที่ได้มาจากข้อสอบข้อที่จะถูกนำมาใช้สอบข้อต่อไป จากสมการที่ 21

$$W_i = D^2 a_i a_i^* w_i^* \quad (21)$$

เมื่อ

$$w_i^* = \left[\frac{qi(\theta)}{pi(\theta)} \right] \left[\frac{pi(\theta) - ci}{1 - ci} \right]^2$$

นอกจากนี้ในการคัดเลือกข้อสอบข้อแรกสามารถเลือกข้อสอบที่มีค่าเมทริกซ์สารสนเทศสูงสุด (I_i) จากสมการ 22 ดังนี้ (Segall. 2010 : 70)

$$I_i = \Phi^{-1} + W_i \quad (22)$$

3. Maximize Kullback-Leibler Information

วิธีการ Posterior Expected Kullback-Leibler Information ถูกนำเสนอให้ใช้ในการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะที่เป็นมิติเดียว Chang and Ying (1996)

และต่อมา Veldkamp and van der Linden (2002) ได้เสนอทางเลือกให้ใช้วิธีการ Posterior Expected Kullback–Leibler Information ในการคัดเลือกข้อสอบแบบปรับเหมาะพหุมิติ (MCAT) โดยมีเกณฑ์การคัดเลือกข้อสอบจากระยะห่างของการกระจายของการตอบสนองจากข้อสอบข้อที่จะถูกเลือกให้ใช้ทดสอบข้อต่อไป เป็นตำแหน่งที่ใช้ประมาณค่าความสามารถของผู้สอบ $\hat{\theta}$ กับความสามารถที่แท้จริงของผู้สอบ θ ด้วยการคาดหมายเอาจากการตอบข้อสอบ (Mulder, van der Linden. 2010 : 85 ; Reckase. 2009 : 334-335) กรณีข้อสอบ 1 ข้อ คำนวณได้ดังสมการ 23 และ 24

$$K_i(\hat{\theta}; \theta) = -E \left[\log \frac{f(U_i|\hat{\theta})}{f(U_i|\theta)} \right] \quad (23)$$

$$K_i(\hat{\theta}; \theta) = P_i(\hat{\theta}) \log \frac{P_i(\hat{\theta})}{P_i(\theta)} + Q_i(\hat{\theta}) \log \frac{Q_i(\hat{\theta})}{Q_i(\theta)} \quad (24)$$

เมื่อ $P_i(*)$ เป็นความน่าจะเป็นในการตอบสนองข้อสอบข้อที่ i ได้ถูกต้อง

$$Q_i(*) = 1 - P_i(*)$$

$U_i(*)$ เป็นผลการตอบข้อสอบข้อที่ กรณีข้อสอบ n ข้อ คำนวณได้ดังสมการ 25

$$K_n(\hat{\theta}; \theta) = \sum_{i=1}^n K_i \quad (25)$$

ส่วนแนวทางในการคัดเลือกข้อสอบที่แนะนำโดย Veldkamp and van der Linden (2002) คือ เลือกจากข้อสอบที่มีค่าสารสนเทศ KLจากการคาดหวังภายหลังสูงสุด (Maximizes the Posterior Expected Kullback–Leibler Information) ที่ถ่วงน้ำหนักสารสนเทศด้วยฟังก์ชันความหนาแน่นภายหลังจากการประมาณค่าความสามารถ และ Veldkamp and van der Linden ใช้ Gauss–Hermite Quadrature ประเมินค่าที่ได้จากสมการที่ 26 แต่อย่างก็ตามวิธีการ Quadrature วิธีอื่น ๆ ก็สามารถใช้งานได้ดีเช่นกัน (Reckase. 2009 : 335) มีสูตรคำนวณดังสมการ 26

$$K_i^B(\hat{\theta}^{k-1}) = \int_{\theta} K_i(\theta, \hat{\theta}^{k-1}) f(\theta|u_1, \dots, u_{k-1}) d\theta \quad (26)$$

เมื่อ K_i^B คือ สารสนเทศความคาดหวังภายหลังแบบเบส์
(Bayesian Posterior Expected Information)
คือ ความสามารถที่ประมาณค่าหลังจากข้อสอบ ข้อที่ $k-1$ สอบไปแล้ว
 $f(\theta / u_1, \dots, u_{k-1})$ คือ ความหนาแน่นภายหลังจากการสอบข้อสอบ ข้อที่ $k-1$ ผ่านไปแล้ว

แต่อย่างไรก็ตามวิธีการนี้ค่อนข้างจะมีความซับซ้อนและจากการทดลองใช้งานของ Reckase โดยใช้โปรแกรม MATLAB ด้วยคุณลักษณะแฝง 3 คุณลักษณะพบว่า การคำนวณทำได้ช้ามาก (Reckase. 2009 : 335)

การยุติการทดสอบ (Test termination)

จุดสิ้นสุดการทดสอบ มีเกณฑ์ในการยุติการทดสอบที่นำเสนอหลากหลายเกณฑ์ เช่น หยุดเมื่อจำนวนข้อสอบที่ทดสอบครบตามจำนวน หรือเมื่อการวัดมีความแม่นยำตามเกณฑ์ หรือใช้ทั้งสองเกณฑ์ร่วมกัน ในการศึกษาจากข้อมูลจำลองส่วนใหญ่ การทดสอบแบบปรับเหมาะแบบพหุมิติจะยุติการทดสอบภายหลังการสอบครบตามจำนวนข้อสอบที่กำหนดไว้ล่วงหน้า อย่างไรก็ตาม การกำหนดจำนวนข้อสอบที่ใช้ในการทดสอบเป็นเกณฑ์การยุติการสอบจะส่งผลต่อความแม่นยำในการประมาณค่าความสามารถที่มีความแปรปรวนระหว่างผู้สอบสูงมาก ผู้สอบบางคนมีความคลาดเคลื่อนมาตรฐานของการวัดจากหนึ่งมิติหรือมากกว่าจะมีขนาดใหญ่ ในขณะที่คนอื่นจะมีขนาดเล็กกว่ามาก และความแม่นยำในการประมาณค่าความสามารถยังขึ้นอยู่กับคำตอบข้อสอบ และคุณลักษณะของคลังข้อสอบ (Frey and Seitz. 2009 : 91)

การยุติการทดสอบจากความแม่นยำในการวัดเป็นทางเลือกหนึ่งที่สามารถใช้เป็นเกณฑ์การยุติ การทดสอบได้ เช่น ในกระบวนการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบพหุมิติของ Segall (1996) การทดสอบดำเนินการต่อไปเรื่อย ๆ จนกว่าค่าความคลาดเคลื่อนมาตรฐานจะตกอยู่ในช่วงความเชื่อถือได้ (ต่ำกว่าค่าที่กำหนดไว้) การกำหนดเกณฑ์วิธีนี้เป็นที่ยอมรับของคนทั่วไปที่สามารถดำเนินการได้อย่างเหมาะสม การใช้ความแม่นยำในการวัดเป็นเกณฑ์ในการยุติการทดสอบจะมีประสิทธิภาพหรือไม่ขึ้นอยู่กับคุณลักษณะของคลังข้อสอบที่นำมาใช้งานด้วย ถ้าองค์ประกอบของความสามารถของผู้สอบตกอยู่ในขอบเขตภายใต้พื้นที่หลายพื้นที่และวัดด้วยข้อสอบจำนวนมากกว่าหนึ่งข้อที่เป็นตัวแทนของเนื้อหาสาระ จะทำให้แบบทดสอบนั้นมีความยาวมาก ดังนั้นการใช้การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบพหุมิติจึงเป็นการรวมเอาการวัดที่มีความแม่นยำและจำนวนข้อสอบที่มีมากข้อในกรณีที่บรรลุถึงระดับความแม่นยำที่ตั้งเอาไว้ซ้ำ และบางครั้งก็จะมีจำนวนข้อสอบน้อยมากในกรณีที่บรรลุถึงระดับความแม่นยำที่ตั้งเอาไว้รวดเร็ว (Frey and Seitz. 2009 : 91)

นอกจากนี้การทำข้อตกลงร่วมกันกับผู้สอบและผู้มีส่วนเกี่ยวข้อง ก็สามารถใช้เป็นเกณฑ์ในการยุติการทดสอบได้เช่นกัน แม้ว่าการใช้เกณฑ์การยุติทดสอบด้วยการบรรยายจะไม่ได้รับความ

มั่นใจจากผู้สอบทุกคนที่จำเป็นต้องใช้เวลาเท่ากันในการทำแบบทดสอบ เนื่องจากสภาพความเป็นจริงแบบทดสอบส่วนมากใช้เวลาในการทดสอบแตกต่างกัน และผู้สอบมีความเร็วในการตอบคำถามแตกต่างกัน โดยข้อเท็จจริงแล้วเวลาในการทดสอบมีความแตกต่างกันมากระหว่างผู้สอบ ดังนั้น ในการประยุกต์ใช้งานในอนาคตประเด็นเวลาสูงสุดที่ใช้ในการทดสอบจะนำมาเพิ่มเป็นเกณฑ์ในการยุติการทดสอบด้วย (Frey and Seitz. 2009 : 91)

สรุป

การทดสอบแบบปรับเหมาะแบบพหุมิติ (MCAT) มีองค์ประกอบที่คล้ายกันกับการทดสอบแบบ ปรับเหมาะกับความสามารถของผู้สอบ (CAT) ที่มีองค์ประกอบสำคัญ เช่น โมเดลอยู่ภายใต้ทฤษฎีตอบสนองข้อสอบแบบพหุมิติ มีการพัฒนาลังข้อสอบที่มีข้อสอบจำนวนมาก ครอบคลุมเนื้อหาที่ต้องการวัด มีการประมาณค่าความสามารถของผู้สอบ มีการคัดเลือกข้อสอบ และมีเกณฑ์ยุติการสอบ แต่สิ่งที่แตกต่างกันคือ ข้อสอบแต่ละข้อของ MCAT จะสามารถวัดความสามารถของผู้สอบได้หลายอย่างในขณะที่ CAT ข้อสอบ แต่ละข้อวัดความสามารถของผู้สอบเพียงด้านเดียว จึงทำให้การทดสอบแบบ MCAT ให้สารสนเทศได้มากกว่าการทดสอบแบบ CAT เมื่อใช้ข้อสอบจำนวนที่มีความยาวของแบบทดสอบเท่ากัน

บรรณานุกรม

- Bock, R.D. and Schilling, S.G. IRT based item factor analysis. In mathilda du Toit (ed). IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT. (pp. 584–591), Scientific Software International, Lincolnwood, IL, 2003.
- Chen, P. H. Comparison of adaptive Bayesian estimation and weighted Bayesian estimation in multidimensional computerized adaptive testing. In D. J. Weiss (Ed.), Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing. Retrieved on 8 February 2011 from www.psych.umn.edu/psylabs/CATCentral/ , 2009.
- Diao, Q., and Reckase, M. (2009). Comparison of ability estimation and item selection methods in multidimensional computerized adaptive testing. In D. J. Weiss (Ed.), Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing. Retrieved on 8 February 2011 from www.psych.umn.edu/psylabs/CATCentral/, 2009.
- Frey, Andreas and Carstensen, Claus H. Diagnostic Classification Models and Multidimensional Adaptive Testing: A Commentary on Rupp and Templin. Measurement: Interdisciplinary Research & Perspective. 7: 1, 58-61, 2009.

- Frey, A., & Seitz, N. N. Assessing the attainment of educational standards: Gains in measurement efficiency by multidimensional adaptive testing. International Journal of Psychology. 43,3/4. 81, 2008.
- Frey, A., & Seitz, N. N. Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. Studies in Educational Evaluation. 35, 89-94, 2009.
- Hambleton, R.K and Swaminathan, H. Item response theory: Principles and applications. (2nd. ed) Boston : Kluwer Nijhoff Publishing, 1985.
- Mulder, J., & van der Linden, W. J. & Pashley, P.J. Multidimensional Adaptive Testing with Kullback–Leibler Information Item Selection. In Wim J. van der Linden & Cees AW Glas (Eds.). Elements of adaptive testing. Springer: New York Dordrecht, Heidelberg London, 2010.
- OECD. Assessing scientific, reading and mathematical literacy: A framework for PISA 2006. Paris : OECD., 2006.
- Owen, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 70, 351–56, 1975.
- Reckase, M.D. Multidimensional item response theory. Springer Science+Business Media: New York, 2009.
- Segall D. O. Multidimensional adaptive testing. Psychometrika. 61(2), 331–354, 13, 1996.
- Segall, D.O. Computerized adaptive testing. In K.Kempf-Leonard (Ed.), Encyclopedia of Social Measurement. (pp.229-438). New York : Academic Press, 2005.
- Segall, D.O. Principles of Multidimensional Adaptive Testing. In Wim J. van der Linden & Cees AW Glas (Eds.). Elements of adaptive testing. Springer : New York Dordrecht Heidelberg , 2010.
- van der Linden, W. J. & Pashley, P.J. Item selection and ability estimation in adaptive Testing. In Wim J. van der Linden & Cees AW Glas (Eds.). Elements of adaptive testing. Springer : New York Dordrecht Heidelberg London, 2010.
- Zimowski, M. Multiple-group analysis. In mathilda du Toit (ed). IRT from SSI: BILOG- MG, MULTILOG, PARSCALE, TESTFACT. (pp. 584–591). Scientific Software International, Lincolnwood, IL, 2003.