บทที่ 2

วรรณคดีที่เกี่ยวข้อง

บทนี้กล่าวถึงหลักการ แนวคิด ทฤษฎีและผลงานการศึกษาวิจัยที่เกี่ยวข้อง โดยแบ่งเป็น 3 ตอน ตอนที่หนึ่งเกี่ยวกับข้อสอบแบบเลือกตอบ ตอนที่สองเกี่ยวกับรูปแบบการตอบสนองข้อสอบ (Item Response Patterns) และตอนที่สามเกี่ยวกับทฤษฎีการตอบสนองข้อสอบ (Item Response Theory)

ตอนที่หนึ่ง

ตอนที่หนึ่งนี้เป็นเรื่องเกี่ยวกับข้อสอบแบบเลือกตอบอันประกอบด้วยลักษณะทั่วไปของข้อสอบแบบเลือกตอบ จุดเด่นของข้อสอบแบบเลือกตอบ ข้อจำกัดของข้อสอบแบบเลือกตอบ การเขียนข้อสอบแบบเลือกตอบและงานวิจัยเกี่ยวกับผลของลักษณะตัวเลือกที่มีต่อคุณภาพของแบบสอบ

ลักษณะทั่วไปของข้อสอบแบบเลือกตอบ

ข้อสอบแบบเลือกตอบเป็นข้อสอบแบบปรนัยชนิดหนึ่งซึ่งเป็นที่นิยมมากทั้งในแบบสอบมาตรฐานและแบบสอบวัดผลสัมฤทธิ์ที่ใช้ในการทดสอบตามสถานศึกษาต่าง ๆ ลักษณะทั่วไปของข้อสอบแบบเลือกตอบก็คือ แต่ละข้อจะประกอบด้วยตัวคำถามและตัวเลือกชุดหนึ่งซึ่งอาจจะเป็น 3 4 หรือ 5 ตัวเลือกให้ผู้สอบเลือกว่าตัวเลือกใดเป็นคำตอบที่ถูกต้องหรือเป็นคำตอบที่ดีที่สุด ตัวคำถามอาจอยู่ในรูปประโยคคำถามที่สมบูรณ์ เช่น **ใครเป็นผู้แต่ง "พระอภัยมณี"** หรืออาจจะเขียนในรูปประโยคที่ไม่สมบูรณ์ก็ได้ เช่น **เมืองหลวงของรัฐโอเรกอนคือ** ในกรณีที่เป็นประโยคคำถามที่สมบูรณ์นั้นตัวเลือกจะเป็นคำตอบที่เป็นไปได้สำหรับคำถามนั้น ส่วนตัวคำถามซึ่งเป็นประโยคที่ไม่สมบูรณ์ตัวเลือกจะเป็นคำหรือข้อความที่เมื่อนำไปต่อท้ายตัวคำถามแล้วจะได้ประโยคที่สมบูรณ์

จุดเด่นของข้อสอบแบบเลือกตอบ

การที่ข้อสอบแบบเลือกตอบเป็นที่นิยมมากนั้นก็เนื่องมาจากจุดเด่นหลายประการของข้อสอบแบบนี้ (Marshall and Hales 1971 : 93 - 96 ; Tuckman 1975 : 92 ;

Mehren and Lehman 1984 : 154 - 155) อันได้แก่

1. ข้อสอบแบบเลือกตอบที่สร้างดี ๆ สามารถวัดวัตถุประสงค์ของการเรียนการสอนทางด้านพุทธิพิสัย(Cognitive Domain)ได้ทุกระดับ ตั้งแต่ระดับความรู้ความจำไปจนถึงระดับการประเมินค่า

2. สามารถใช้ได้กับเนื้อหาวิชาทุกสาขาและใช้ได้กับนักเรียนทุกระดับชั้น

3. ภายในเวลาของการสอบที่จำกัดข้อสอบแบบเลือกตอบจะถามได้มากข้อกว่าข้อสอบแบบอัตนัยจึงสามารถออกข้อสอบครอบคลุมวัตถุประสงค์ของการเรียนการสอนได้หลายข้อ ถ้าสร้างตารางกำหนดข้อสอบ (Table of Specifications) ดี ๆ และใช้ตารางนี้เป็นกรอบในการสร้างข้อสอบแล้วการสอบก็จะมีความคลาดเคลื่อนอันเกิดจากการสุ่มเนื้อหาน้อย

4. เมื่อเปรียบเทียบกับข้อสอบแบบเติมคำข้อสอบแบบเลือกตอบจะมีความคลาดเคลื่อนในการตรวจให้คะแนนน้อยกว่า นอกจากนี้ก็ยังสามารถตรวจให้คะแนนได้อย่างรวดเร็ว ถูกต้องและเป็นปรนัยแม้จะให้ผู้อื่นหรือเครื่องจักรเป็นผู้ตรวจก็ตาม ยิ่งกว่านั้น การตรวจให้คะแนนก็ยังเป็นอิสระจากอิทธิพลของผลการทดสอบที่มีมาก่อนหน้านั้น ตลอดจนหน้าตา ท่าทาง หรือความเรียบร้อยของผู้สอบ

5. เมื่อเปรียบเทียบกับข้อสอบแบบถูกผิด ข้อสอบแบบเลือกตอบจะมีความแปรปรวนของคะแนนอันเนื่องมาจากการเดาน้อยกว่า เพราะว่าความน่าจะเป็นในการเดาถูกนั้นขึ้นอยู่กับจำนวนตัวเลือก เช่น ในกรณี 3 ตัวเลือก ความน่าจะเป็นในการเดาถูกคือ 1 ใน 3 ส่วน 5 ตัวเลือกก็จะเป็น 1 ใน 5

6. จากการศึกษาของ Knapp (1968 อ้างถึงใน Marshall and Hales 1971 : 155) พบว่า ตัวเลือกในข้อสอบแบบเลือกตอบนั้นมีผลต่อความยากของข้อสอบ ดังนั้น เราจึงสามารถควบคุมความยากของข้อสอบได้ด้วยการปรับระดับความเป็นเอกพันธ์ของตัวเลือก

7. แม้ว่าข้อสอบทุกชนิดต้องประสบกับปัญหาเรื่องความกำกวมซึ่งผู้พัฒนาแบบสอบจะต้องพยายามขจัดให้หมดไปหรือให้มีอยู่น้อยที่สุด อย่างไรก็ตาม เมื่อเปรียบเทียบกับข้อสอบแบบอื่น ๆ แล้วข้อสอบแบบเลือกตอบจะสามารถควบคุมความกำกวมได้มากกว่า

8. ข้อสอบแบบเลือกตอบสามารถให้สารสนเทศสำหรับการวินิจฉัยนักเรียนได้ โดย

เฉพาะอย่างยิ่งถ้าคำตอบที่ให้เลือกนั้นมีระดับของความถูกต้องต่าง ๆ กัน

9. ในทัศนะของนักเรียนข้อสอบแบบเลือกตอบนี้ตอบง่ายจึงชอบมากกว่าข้อสอบแบบถูกผิดเพราะรู้สึกว่ากำกวมน้อยกว่าแบบถูกผิด

ข้อจำกัดของข้อสอบแบบเลือกตอบ

แม้ว่าข้อสอบแบบเลือกตอบจะมีจุดเด่นอยู่หลายประการดังกล่าวแล้ว แต่ก็มีข้อจำกัดอยู่บ้างเช่นเดียวกับข้อสอบแบบอื่น ๆ (Marshall and Hales 1971 : 92 - 93 ; Tuckman 1975 : 92 ; Mehrens and Lehman 1984 : 155 - 156) ข้อจำกัดบางประการของข้อสอบแบบเลือกตอบ ได้แก่

1. เปิดโอกาสให้ผู้สอบตอบถูกได้โดยการเดา ในบางครั้งผู้สอบสามารถตอบถูกได้โดยไม่ต้องมีความรู้หรือเคยเรียนมาก่อน เนื่องจากข้อสอบนั้นมีสิ่งที่ช่วยแนะคำตอบ หรือมีตัวเลือกที่ไม่สมเหตุผลมากเกินไป หรืออาจเป็นเพราะว่าข้อสอบนั้นทดสอบความถนัดของผู้สอบในการเผชิญกับสถานการณ์แปลกใหม่มากกว่าทดสอบความรู้ของผู้สอบ

2. ผลจากการวิจัยพบว่าผู้สอบที่มีความฉลาดในการตอบข้อสอบ (Test-wise) จะทำข้อสอบแบบเลือกตอบได้ดีกว่าผู้สอบที่ไม่มีความฉลาดในการตอบข้อสอบ เช่นเดียวกับผู้สอบที่มีทักษะในการจับความกำกวมจะทำข้อสอบได้ดีกว่าผู้ที่ไม่มีทักษะนี้ อีกทั้งข้อสอบแบบเลือกตอบยังเป็นที่ชื่นชอบของผู้สอบที่มีความกล้าเสี่ยงสูง (High-risk-taking) (Rowley 1974 ; Alker and others 1967 อ้างถึงใน Mehrens and Lehman 1984 : 156)

3. ในบรรดาข้อสอบแบบปรนัยที่มีคำตอบให้นั้น ข้อสอบแบบเลือกตอบใช้เวลาในการตอบมากที่สุด โดยเฉพาะอย่างยิ่งเมื่อมีความละเอียดในการจำแนกมาก ๆ ฉะนั้น ภายในเวลาอันจำกัดของการสอบจึงสามารถสุ่มเนื้อหามาถามได้น้อยกว่า นอกจากนั้นก็ยังสิ้นเปลืองวัสดุมากกว่าอีกด้วย

4. ข้อสอบแบบเลือกตอบไม่สามารถวัดความสามารถของผู้สอบในการเรียบเรียงหรือแสดงคำตอบด้วยภาษาของผู้สอบเองได้ หากต้องการวัดความสามารถดังกล่าวก็ต้องใช้ข้อสอบแบบอัตนัยแทน

5. ผู้เขียนข้อสอบมักมีแนวโน้มที่จะเขียนข้อสอบวัดเพียงระดับระลึกข้อเท็จจริงเท่านั้น

แม้ว่าแนวโน้มที่ว่านี้สำหรับข้อสอบแบบเลือกตอบจะน้อยกว่าข้อสอบปรนัยแบบอื่น ๆ ก็ตาม

6. ในการเขียนข้อสอบแบบเลือกตอบให้ได้ข้อสอบที่มีคุณภาพนั้น ผู้เขียนจะต้องมีความรู้ในเนื้อหาวิชา ตระหนักในวิธีวิทยาในการเขียนข้อสอบ มีทักษะในการใช้ภาษาและต้องมีความรู้เกี่ยวกับระดับพัฒนาการของผู้สอบ การเขียนข้อสอบให้ได้ข้อสอบที่ดีจริงเป็นเรื่องยากจะทำได้ง่ายก็แต่ผู้ที่มีทักษะเท่านั้น นอกจากนั้นยังต้องใช้เวลาและความพยายามอย่างสูง โดยเฉพาะอย่างยิ่งเมื่อต้องการวัดกระบวนการทางสมองในระดับสูง ๆ หากไม่ต้องการเลือกข้อสอบดี ๆ ไว้ใช้อีกก็ไม่ขอแนะนำให้สร้างข้อสอบแบบเลือกตอบสำหรับทดสอบคนกลุ่มเล็ก ๆ ซึ่งใช้ข้อสอบนั้นเพียงครั้งเดียวแล้วก็ไม่ได้ใช้อีก

การเขียนข้อสอบแบบเลือกตอบ

ข้อสอบแบบใดก็ตามจะดีหรือไม่นั้นอยู่ที่ความชัดเจน สำหรับข้อสอบแบบเลือกตอบจะต้องมีตัวคำถามที่ชัดเจนไม่กำกวม มีคำตอบถูกและตัวลวงที่เป็นไปได้ คุณค่าของข้อสอบแบบเลือกตอบขึ้นอยู่กับทักษะในการเขียนตัวลวงเป็นอย่างมาก ในระหว่างการเขียนข้อสอบมีคำถามอยู่ 2 คำถามที่ควรจะถามตนเองอยู่ตลอดเวลา คือ หนึ่ง ข้อสอบข้อนี้สื่อสารได้ดีแล้วหรือยัง และสอง ข้อสอบข้อนี้มีอะไรเป็นสิ่งชี้แนะคำตอบถูกหรือไม่

การเขียนข้อสอบแบบเลือกตอบอาจจะไม่มีวิธีใดเป็นวิธีที่ดีที่สุด อย่างไรก็ตาม เพื่อให้ได้ข้อสอบที่มีคุณภาพในการเขียนข้อสอบจึงควรประกอบด้วยกิจกรรม 3 อย่าง ได้แก่ การวิเคราะห์ผลการเรียนรู้ (Learning Outcome) ที่ต้องการวัด การร่างข้อสอบและการบรรณาธิกรณ์ข้อสอบ กิจกรรมทั้งสามมีรายละเอียดดังต่อไปนี้

1. การวิเคราะห์ผลการเรียนรู้ที่ต้องการวัด

เนื่องจากในการตอบข้อสอบแบบเลือกตอบนั้นผู้สอบจะต้องเลือกคำตอบจากตัวเลือกที่กำหนดให้ ฉะนั้น ก่อนลงมือเขียนข้อสอบผู้เขียนจะต้องพิจารณาว่าผลการเรียนรู้ที่ต้องการจะวัดนั้นสามารถจะแสดงออกด้วยวิธีการให้เลือกจากตัวเลือกที่กำหนดให้ได้หรือไม่ ถ้าผลการเรียนรู้นั้นต้องการให้ระลึก (Recall) ข้อเท็จจริง หรือผลของบางสิ่งบางอย่างแล้ว การให้เลือกจากตัวเลือกจะไม่เหมาะสมกับผลการเรียนรู้ที่ต้องการวัด แต่ถ้าผลการเรียนรู้นั้นให้เลือกสิ่งที่มีประสิทธิภาพมากที่สุดการให้เลือกจากตัวเลือกที่กำหนดก็จะเป็นวิธีที่เหมาะสม ยังมีผลการเรียนรู้มากมายที่ไม่สามารถบอกได้โดยง่ายว่าการให้เลือกเป็นวิธีที่เหมาะสมหรือไม่ ทางที่ดีคือขจัดผลการเรียนรู้ที่เห็นได้ชัดเจนว่าไม่เหมาะกับการให้เลือกออกไปก่อน แล้วพิจารณาผล

การเรียนรู้ข้อที่เหลืออยู่ว่าจะใช้วิธีให้เลือกได้อย่างไร หรือเพื่อให้ง่ายขึ้นอาจเปลี่ยนมาเป็นการถามว่า ผลการเรียนรู้ข้อนี้จะใช้ตัวเลือกว่าอย่างไร

2. การร่างข้อสอบ

ในกิจกรรมการร่างข้อสอบจะประกอบด้วยกิจกรรมย่อย 5 กิจกรรม ได้แก่ การเขียนตัวคำถาม การเขียนตัวคำตอบถูก การเขียนตัวลวง การจัดเรียงตัวเลือก และการตรวจสอบความกำกวมและการแนะคำตอบ

2.1 การเขียนตัวคำถาม

ตัวคำถามของข้อสอบแบบเลือกตอบอาจเป็นประโยคคำถามที่สมบูรณ์ หรือเป็นข้อความที่ไม่ใช่ประโยคที่สมบูรณ์ก็ได้ หน้าที่หลักของตัวคำถามมีอยู่ 2 ประการ คือ เสนอปัญหาและกำหนดกรอบของปัญหา เมื่ออ่านตัวคำถามแล้วผู้สอบควรจะรู้ว่าต้องระลึก (Recall) ถึงอะไรบ้าง แล้วพิจารณาเลือกคำตอบจากตัวเลือกที่กำหนดให้ ในที่นี้ขอเสนอข้อควรคำนึงบางประการสำหรับการเขียนตัวคำถาม ดังต่อไปนี้

2.1.1 ตัวคำถามจะต้องเสนอปัญหาเพื่อให้ผู้ตอบรู้ว่าเขาจะตอบสนองอย่างไร

2.1.2 การเขียนคำถามให้แคบเข้าจะช่วยในการวัดความเข้าใจของผู้สอบ

2.1.3 พยายามใช้คำถาม **ทำไม อย่างไร** มากกว่า **ใคร เมื่อไร ที่ไหน** เพราะสองคำถามแรกผู้ตอบต้องใช้ความเข้าใจและการประยุกต์ความรู้ไปใช้ตอบคำถามนี้ ส่วนสามคำถามหลังนั้นเป็นการวัดเพียงระดับความรู้ความจำเท่านั้น

2.1.4 ข้อสอบเกี่ยวกับนิยามศัพท์ไม่ควรใช้นิยามเป็นตัวคำถาม ตัวคำถามควรเป็นคำศัพท์และตัวเลือกเป็นนิยามของคำศัพท์

2.1.5 ไม่ควรลอกข้อความในหนังสือเรียนหรือสื่อการสอนอื่นมาเป็นคำถาม

เมื่อเขียนตัวคำถามแต่ละข้อเสร็จแล้วควรทบทวนดูว่า ตัวคำถามได้เสนอปัญหาชัดเจนแล้วหรือยัง การใช้ถ้อยคำภาษาชัดเจนเข้าใจง่ายสื่อความหมายได้ตรงตามที่ตั้งใจไว้หรือไม่ ตัวคำถามควรจะเพิ่มสารสนเทศอะไรอีกบ้างไหม หรือว่ามีสารสนเทศที่เกินความจำเป็นควรตัดออกหรือไม่

2.2 การเขียนคำตอบถูก

เมื่อเขียนตัวคำถามเสร็จแล้วควรจะเขียนคำตอบถูกไว้ก่อนที่จะเขียนตัวลวง การที่ต้องเขียนตามลำดับเช่นนี้ก็ด้วยเหตุผล 2 ประการ คือ ประการแรก เพื่อให้ผู้เขียนข้อสอบแน่ใจว่าตัวคำถามนั้นจะได้คำตอบที่สั้นแต่ชัดเจนซึ่งบางครั้งก็อาจจะมีการปรับตัวคำถามบ้าง ประการที่สอง เพื่อให้ความยากของข้อสอบนั้นเป็นผลจากการเขียนตัวลวงให้เข้าคู่กับคำตอบถูก

จุดมุ่งหมายของคำตอบถูกในข้อสอบแบบเลือกตอบก็เพื่อให้การตอบสนองที่ชัดเจนถูกต้องสำหรับคำถามนั้น คำตอบถูกจึงควรเป็นตัวเลือกที่ดีที่สุดตามความเห็นของผู้เชี่ยวชาญด้านเนื้อหาและเป็นตัวเลือกซึ่งนักเรียนที่มีผลสัมฤทธิ์สูงน่าจะเลือกมากที่สุด ในการเขียนคำตอบถูกมีข้อควรระวัง ดังนี้

2.2.1 หลีกเลี่ยงการคัดลอกข้อความจากหนังสือเรียน หรือสื่อการเรียนอื่น ๆ

2.2.2 คำตอบถูกควรจะสอดคล้องกับตัวคำถามทั้งในแง่ของไวยากรณ์และเชิงเหตุผล

2.2.3 คำตอบถูกควรปรากฏในตำแหน่งต่าง ๆ ของตัวเลือกด้วยความถี่พอ ๆ กัน

2.2.4 ใช้ตัวเลือกปลายเปิด **ไม่มีข้อใดถูก** และตัวเลือกปลายปิด **ถูกทุกข้อ** เป็นคำตอบถูกเฉพาะข้อสอบที่ต้องการคำตอบที่ถูกต้อง (Correct-answer) เท่านั้น ไม่ควรใช้กับข้อสอบที่ต้องการคำตอบที่ดีที่สุด (Best answer)

2.3 การเขียนตัวลวง

ตัวลวงมีบทบาทอย่างสำคัญในการกำหนดประสิทธิผลและความยากของ

ข้อสอบแบบเลือกตอบ จุดมุ่งหมายของตัวลวงคือการเสนอคำตอบที่น่าสนใจและสมเหตุผลสำหรับผู้สอบที่มีความเข้าใจผิด ๆ หรือผู้ที่ยังไม่สัมฤทธิ์ผลตามที่ต้องการพอที่จะระบุคำตอบที่ถูกต้องได้ ส่วนผู้ที่มีผลสัมฤทธิ์ตามที่ต้องการแล้วจะปฏิเสธตัวลวงนี้

ตัวลวงจึงควรสร้างจากความเข้าใจผิด ๆ หรือข้อสรุปจากการใช้ความรู้ที่ไม่เหมาะสม ไม่ควรใช้วิธีปรับข้อความที่ถูกต้องเพียงเล็กน้อย หรือใช้การแนะคำตอบอย่างผิด ๆ เพราะว่าจะทำให้เป็นที่สนใจของผู้สอบที่เตรียมตัวมาอย่างดี

หากมีตัวคำถามที่ดี คำตอบถูกดี และตัวลวงที่น่าสนใจแต่ไม่ใช่คำตอบถูกแล้ว ทั้งหมดนี้ประกอบกันเข้าจะได้ข้อสอบแบบเลือกตอบที่มีความตรง (Valid) และท้าทายความสามารถของผู้สอบ อย่างไรก็ตามมีข้อควรคำนึงในการเขียนตัวลวง ดังนี้

2.3.1 ข้อสอบแต่ละข้อควรมีตัวลวง 3 หรือ 4 ตัว

2.3.2 ตัวลวงทุกตัวควรมีความยาวใกล้เคียงกัน มีโครงสร้างแบบเดียวกัน และใช้ถ้อยคำที่กระชับพอ ๆ กัน

2.3.3 สร้างตัวลวงจากความเข้าใจผิด ๆ และข้อสรุปที่เป็นไปได้แต่ไม่ถูกต้อง

2.3.4 ตัวลวงไม่ควรคาบเกี่ยวกัน หรือคาบเกี่ยวกับคำตอบถูกทั้งในแง่ของภาษาและตัวเลข

2.3.5 ปรับตัวลวงให้สามารถควบคุมความยากของข้อสอบได้ ทั้งนี้เพราะว่าตัวเลือกต่าง ๆ ยิ่งมีความใกล้เคียงกันมากเท่าใดข้อสอบก็จะยิ่งยากเท่านั้น

2.4 การเรียงตัวเลือก

การเรียงตัวเลือกอาจทำได้หลายวิธีขึ้นอยู่กับลักษณะของตัวเลือก เช่น อาจเรียงตามจำนวนเลขจากมากไปน้อยหรือน้อยไปมาก เรียงตามลำดับเวลาหรือเหตุการณ์ที่เกิดขึ้น เรียงตามลำดับตัวอักษร เรียงตามความยาวของตัวเลือก หรือเรียงโดยวิธีสุ่ม ไม่ว่าจะเรียงโดยวิธีใดก็ตามผลสุดท้ายคำตอบถูกควรอยู่ในตำแหน่งต่าง ๆ ด้วยความถี่พอ ๆ กัน

2.5 การตรวจสอบความกำกวมและการแนะคำตอบ

เมื่อเขียนข้อสอบเสร็จแล้วก่อนที่จะไปสู่ขั้นตอนของการบรรณาธิกรณ์ข้อสอบ ผู้เขียนข้อสอบควรตรวจสอบข้อสอบแต่ละข้อเสียก่อนเพื่อให้แน่ใจในขั้นต้นว่าไม่มีข้อใดที่ยังกำกวม หรือมีส่วนที่ชี้แนะคำตอบถูกอยู่ หากพบว่ามีอยู่ก็จะได้ปรับปรุงแก้ไขทันที

3. การบรรณาธิกรณ์ข้อสอบ

การบรรณาธิกรณ์ข้อสอบ เป็นการทบทวนตรวจสอบข้อสอบที่เขียนขึ้นอีกชั้นหนึ่งโดยบุคคลอีกคนหนึ่งหรืออีกกลุ่มหนึ่งซึ่งมิใช่ผู้เขียนข้อสอบ การบรรณาธิกรณ์ข้อสอบนี้อาจแบ่งเป็นกิจกรรมย่อย ๆ ได้ 4 กิจกรรม คือ การบรรณาธิกรณ์ตัวคำถาม การบรรณาธิกรณ์ตัวเลือก การตรวจสอบความสัมพันธ์ระหว่างตัวคำถามและตัวเลือก และการบรรณาธิกรณ์ขั้นสุดท้าย

3.1 การบรรณาธิกรณ์ตัวคำถาม

การบรรณาธิกรณ์ตัวคำถามจะพิจารณาถึงความชัดเจน เข้าใจง่ายของถ้อยคำภาษาที่ใช้ ความสมบูรณ์เพียงพอของสารสนเทศในตัวคำถาม ตลอดจนพิจารณาว่ามีสิ่งใดที่เกินความจำเป็นในตัวคำถามซึ่งควรตัดออกหรือไม่

3.2 การบรรณาธิกรณ์ตัวเลือก

การบรรณาธิกรณ์ตัวเลือกจะพิจารณาเพื่อให้แน่ใจว่าถ้อยคำที่ใช้กระชับชัดเจน ตัวเลือกต่าง ๆ เทียบเท่ากันทั้งในแง่ของความยาวและโครงสร้าง ตัวเลือกไม่ยาวหรือซับซ้อนจนเกินไป ทุกตัวเลือกสัมพันธ์กับปัญหาที่ถาม

3.3 การตรวจสอบความสัมพันธ์ระหว่างตัวคำถามและตัวเลือก

การตรวจสอบความสัมพันธ์ระหว่างตัวคำถามและตัวเลือกนี้มีจุดที่ควรพิจารณาอยู่ 3 ประการ คือ ประการแรก ความสอดคล้องรับกันระหว่างตัวเลือกและตัวคำถามทั้งในเชิงเหตุผลและไวยากรณ์ ประการที่สอง ทุก ๆ ตัวเลือกเป็นการตอบปัญหาในตัวคำถามโดยตรง ประการที่สาม ตัวคำถามสะท้อนถึงธรรมชาติของตัวเลือก

3.4 การบรรณาธิกรณ์ขั้นสุดท้าย

การบรรณาธิกรณ์ขั้นสุดท้ายเป็นการทบทวนตรวจสอบข้อสอบโดยรวมเป็นครั้งสุดท้ายซึ่งมีแนวทางในการพิจารณา ดังนี้

3.4.1 ความสามารถที่วัดด้วยข้อสอบข้อนั้นควรสะท้อนถึงผลการปฏิบัติหลัก ๆ ที่ระบุไว้ในวัตถุประสงค์อย่างแท้จริง

3.4.2 คำถามต้องระบุถึงข้อเท็จจริงอย่างถูกต้อง พร้อมทั้งมีคำตอบที่ถูกต้องเพียงคำตอบเดียว

3.4.3 ตัวคำถามต้องเขียนอย่างชัดเจน รัดกุม และปัญหาที่ถามนั้นต้องเข้าใจได้ง่าย

3.4.4 ตัวเลือกที่เป็นคำตอบถูกควรจะกระชับและเป็นคำตอบที่ถูกต้องของปัญหาที่ถามหรือเป็นส่วนที่ทำให้ข้อความในตัวคำถามเป็นประโยคที่สมบูรณ์และถูกต้อง

3.4.5 ตัวเลือกที่เป็นคำตอบที่ผิดนั้นต้องผิดอย่างแท้จริง หรือมีความเหมาะสมน้อยกว่าตัวเลือกที่เป็นคำตอบถูกอย่างชัดแจ้ง

3.4.6 ทุกตัวเลือกควรจะคล้ายคลึงกันในทุก ๆ ด้าน ยกเว้นประเด็นที่เป็นคำถามของข้อนั้น ตัวเลือกที่ผิดควรจะมีความเป็นไปได้และน่าสนใจสำหรับผู้ที่ไม่รู้คำตอบที่ถูกต้อง

3.4.7 ตัวเลือกทุกตัวควรจะสอดคล้องกับตัวคำถาม และเป็นตัวเลือกที่มีความหมายไม่ใช่ตัวเลือกเหลวไหล ควรใช้คำที่น่าเชื่อไม่ควรใช้คำที่เฉพาะเจาะจง เช่น คำว่า **เสมอ ๆ เฉพาะ ไม่เคย ไม่มีเลย** เป็นต้น อีกทั้งไม่ควรมีตัวเลือกที่เหมือนกัน 2 ตัวเลือกเพราะจะมีค่าเท่ากับตัวเลือกเพียงตัวเดียวเท่านั้น

3.4.8 ข้อสอบทุกข้อควรจะธรรมดาที่สุดเท่าที่จะเป็นไปได้ แต่ยังคงสามารถประเมินวัตถุประสงค์ที่ต้องการได้

3.4.9 ข้อสอบทุกข้อในแบบสอบแต่ละฉบับควรเป็นอิสระต่อกัน กล่าว

คือ ไม่มีข้อใดที่เป็นตัวชี้คำตอบของข้ออื่นและการตอบข้อสอบข้อใดข้อหนึ่งได้ถูกต้องนั้นไม่ได้ขึ้นอยู่อยู่กับความสามารถในการตอบข้อสอบข้ออื่นได้ถูกต้อง

งานวิจัยเกี่ยวกับผลของลักษณะตัวเลือกที่มีต่อคุณภาพของแบบสอบ

การศึกษาวิจัยเกี่ยวกับลักษณะของตัวเลือกแบบต่าง ๆ ในข้อสอบแบบเลือกตอบที่ผ่านมานั้นล้วนศึกษาในแง่ของความยาก อำนาจจำแนก ความเที่ยงและความตรงเป็นประเด็น ๆ ไป ในที่นี้ขอแยกกล่าวเป็น 2 กลุ่ม คือ งานวิจัยของต่างประเทศและงานวิจัยในประเทศ

1. **งานวิจัยของต่างประเทศ**

ลักษณะตัวเลือกที่มีการศึกษาวิจัยในต่างประเทศหลายชิ้น คือ ตัวเลือกปลายเปิด **ไม่มีข้อใดถูก** (Non of These) ซึ่งได้ผลการวิจัยสอดคล้องกันบ้าง แตกต่างกันบ้าง ดังเช่น การศึกษาของ Wesman และ Bennett (1946 อ้างถึงใน Thorndike 1971 : 85) ในแบบสอบคำศัพท์และแบบสอบเลขคณิต โดยแต่ละแบบสอบจัดทำเป็น 2 ฉบับ ฉบับแรกมีตัวเลือกธรรมดา 5 ตัวเลือก และฉบับที่สองแทนคำตอบถูกหรือตัวลวงตัวใดตัวหนึ่งด้วยตัวเลือกปลายเปิด **ไม่มีข้อใดถูก** ผลการศึกษาพบว่า แบบสอบ 2 ฉบับนี้ไม่มีความแตกต่างกันในด้านความยากและสหสัมพันธ์กับเกณฑ์ภายนอกซึ่งสอดคล้องกับการศึกษาในทำนองเดียวกันของ Rimland (1960 อ้างถึงใน นวลน้อย แต่บรรพกุล 2520 : 14) ที่ศึกษากับแบบสอบคณิตศาสตร์แล้วไม่พบความแตกต่างทั้งในด้านอำนาจจำแนกและระดับความยากง่าย อีกทั้ง William และ Hopkins (1967 อ้างถึงใน Wood 1979 : 225) ก็รายงานว่า การใช้ตัวเลือก **ไม่มีข้อใดถูก** ไม่ทำให้ความเที่ยง ความตรงของแบบสอบแตกต่างออกไปแต่อย่างใด แต่ Boynton (1950 อ้างถึงใน Thorndike 1971 : 85) กลับพบว่าข้อสอบสะกดคำที่ใช้ตัวเลือกปลายเปิด **ไม่มีข้อใดถูก** ยากกว่าข้อสอบที่ไม่ใช้ตัวเลือกนี้

นอกจากการศึกษาเปรียบเทียบข้อสอบที่ใช้ตัวเลือกปลายเปิดกับข้อสอบที่ใช้ตัวเลือกธรรมดาดังกล่าวแล้ว Hughes และ Trimble (1965 อ้างถึงใน นวลน้อย แต่บรรพกุล 2520 : 16) ยังได้ศึกษาแบบสอบในวิชาจิตวิทยาโดยเปรียบเทียบแบบสอบที่มีตัวเลือกธรรมดากับแบบสอบที่มีตัวเลือกซับซ้อน (Complex Alternatives) ซึ่งได้แก่ตัวเลือกปลายเปิด **ไม่มีข้อใดถูก** ตัวเลือกปลายปิด **ถูกทุกข้อ** และตัวเลือกผสม คือ **ถูกทั้งข้อ 1 และข้อ 2** หรือ **ผิดทั้งข้อ 1 และข้อ 2** ผลการวิจัยพบว่า แบบสอบที่มีตัวเลือกซับซ้อนมีระดับความยากง่ายสูงกว่าแบบสอบที่มีตัวเลือกธรรมดาอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ส่วนค่าอำนาจจำแนกแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ ต่อมา Mueller (1975 อ้างถึงใน นวลน้อย แต่บรรพกุล

2520 : 17) ได้ศึกษาในทำนองเดียวกันด้วยแบบสอบสำหรับผู้ขอใบอนุญาตประกอบอาชีพพนักงานขาย พบว่า แบบสอบตัวเลือกธรรมดามีค่าอำนาจจำแนกสูงสุด ส่วนแบบสอบที่มีตัวเลือกปลายเปิด ตัวเลือกปลายปิดและตัวเลือกผสมมีค่าอำนาจจำแนกพอ ๆ กัน แบบสอบที่มีตัวเลือกผสมมีระดับความยากง่ายสูงสุดและแบบสอบที่มีตัวเลือกธรรมดามีระดับความยากง่ายต่ำสุด นอกจากนี้ยังพบว่า เมื่อใช้ตัวเลือก 4 แบบนี้เป็นตัวลวงนั้นนักเรียนเลือกตัวลวงแบบผสมมากที่สุด รองลงมาคือตัวเลือกปลายเปิด ตัวเลือกปลายปิดและตัวเลือกธรรมดาตามลำดับ

ตัวเลือกอีกแบบหนึ่งที่พบว่ามีผู้ศึกษาเพียงรายเดียว คือ ตัวเลือก **ไม่ทราบ** (Don't Know) ซึ่ง Friedman และ Fleishman (1956 อ้างถึงใน Thorndike 1971 : 85) ได้เพิ่มตัวเลือก **ไม่ทราบ** นี้จากตัวเลือกเดิมคือ **เหมือนกัน** และ **ต่างกัน** ของแบบสอบ Rhythm Discrimination Test of Seashore Measures of Musical Talent แล้วพบว่า แบบสอบที่เพิ่มตัวเลือก **ไม่ทราบ** เข้าไปนี้มีความเที่ยงสูงกว่าเล็กน้อย

2. งานวิจัยในประเทศ

การศึกษาวิจัยภายในประเทศเกี่ยวกับลักษณะตัวเลือกของแบบสอบเลือกตอบนั้นมีน้อยมากและเป็นการศึกษาเปรียบเทียบหลาย ๆ ลักษณะในคราวเดียวกัน ได้แก่ งานวิจัยของไพบูลย์ จิตร์โต (2514) ซึ่งศึกษาเปรียบเทียบแบบสอบ 4 แบบ คือ แบบสอบเลือกตอบชนิดตัวลวงเป็นตัวเฉียด ชนิดตัวลวงเป็นตัวดัก ชนิดปลายเปิดให้เติม และแบบสอบเติมคำ ในวิชาคณิตศาสตร์ ชั้น ป.4 และ ป.7 โดยแยกเป็นคณิตศาสตร์ทักษะและโจทย์ปัญหา ผลการวิจัยพบว่า แบบสอบเลือกตอบชนิดตัวลวงเฉียดให้ค่าความเที่ยง ความตรงและความยากมาตรฐานต่ำสุด ส่วนแบบสอบเติมคำให้ค่าความเที่ยง ความตรงและความยากมาตรฐานสูงสุด

ต่อมา นวลน้อย แต่บรรพกุล (2520) ได้ศึกษาเปรียบเทียบแบบสอบเลือกตอบที่มีตัวเลือกธรรมดากับแบบสอบเลือกตอบที่มีตัวเลือกซับซ้อนต่างกัน 3 แบบ คือ แบบปลายเปิด **ผิดทุกข้อ** แบบปลายปิด **ถูกทุกข้อ** และแบบผสม **ถูกทั้ง ก และ ข** หรือตัวเลือกที่มีลักษณะคล้ายกันได้แก่ **ถูกทั้ง ก และ ค ถูกทั้ง ก และ ง ถูกทั้ง ข และ ค ถูกทั้ง ข และ ง ถูกทั้ง ค และ ง** โดยศึกษากับแบบสอบสะกดคำ แบบสอบหลักภาษาและแบบสอบความเข้าใจในการอ่านภาษาไทยชั้น ป.7 พบว่า ความตรงของแบบสอบสะกดคำทั้ง 4 ชุดต่างกันอย่างไม่มีนัยสำคัญทางสถิติ ส่วนความตรงของแบบสอบหลักภาษาและแบบสอบความเข้าใจในการอ่านนั้นแบบสอบที่มีตัวเลือกธรรมดามีค่าความตรงสูงสุดและแบบสอบที่มีตัวเลือกผสมมีค่าความตรงต่ำสุด ส่วนค่าความเที่ยงของแบบสอบทั้ง 4 ชุด แตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ ไม่ว่าจะเป็นแบบสอบสะกดคำ หลักภาษาหรือความเข้าใจในการอ่าน เช่นเดียวกับค่าอำนาจจำแนก สำหรับค่า

ความยากมาตรฐานนั้นทั้ง 4 ชุด แตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติในแบบสอบสะกดคำและหลักภาษา ส่วนแบบสอบความเข้าใจในการอ่านนั้นแบบสอบที่มีตัวเลือกผสมมีค่าความยากมาตรฐานสูงสุดและแบบสอบที่มีตัวเลือกปลายเปิดมีค่าความยากมาตรฐานต่ำสุด ในเรื่องการเดาแบบสอบสะกดคำและแบบสอบหลักภาษาที่มีตัวเลือกธรรมดามีเปอร์เซ็นต์การเดาสูงสุด และตัวเลือกผสมมีเปอร์เซ็นต์การเดาต่ำสุด ส่วนแบบสอบความเข้าใจในการอ่านที่มีตัวเลือกปลายปิดมีเปอร์เซ็นต์การเดาสูงสุดและตัวเลือกผสมมีเปอร์เซ็นต์การเดาต่ำสุด

ล่าสุด นิรมล บุญตะรัตน์ (2524) ได้ศึกษาแบบสอบเลือกตอบที่มีรูปแบบตัวเลือกแตกต่างกัน 3 ชุด คือ แบบสอบที่ใช้ตัวเลือกธรรมดาทั้งฉบับ แบบสอบที่ใช้ตัวเลือกปลายเปิด **ไม่มีข้อใดถูก** ทั้งฉบับและแบบสอบที่ใช้ตัวเลือกผสมระหว่างตัวเลือกธรรมดากับตัวเลือกปลายเปิด โดยศึกษาจากแบบสอบในวิชาคณิตศาสตร์ ชั้น ม.1 พบว่า แบบสอบเลือกตอบทั้ง 3 ชุดนี้ให้ค่าความเที่ยง ความตรง ความยากและอำนาจจำแนกไม่แตกต่างกัน

การศึกษาเกี่ยวกับข้อสอบแบบเลือกตอบที่ใช้ตัวเลือกลักษณะต่าง ๆ ดังกล่าวข้างต้นนั้น จะเห็นว่า ผลของการศึกษาทั้งภายในประเทศและต่างประเทศแม้ว่าจะไม่สอดคล้องกันทั้งหมดแต่ก็พอจะเห็นแนวโน้มว่า ข้อสอบเลือกตอบที่ใช้ตัวเลือกธรรมดา ตัวเลือกปลายเปิด ตัวเลือกปลายปิดและตัวเลือกผสมนั้นจะให้ค่าความเที่ยง ความตรง ความยาก และอำนาจจำแนกพอ ๆ กันเมื่อใช้กับบางเนื้อหาวิชาที่มีธรรมชาติของวิชาค่อนข้างชัดเจนแน่นอน ดังเช่น คณิตศาสตร์ หลักภาษา เป็นต้น ฉะนั้นจึงเป็นข้อสังเกตประการหนึ่งว่า ผลของตัวเลือกลักษณะต่าง ๆ ที่มีต่อความเที่ยง ความตรง ความยาก และอำนาจจำแนกของแบบสอบนั้นอาจจะขึ้นอยู่กับลักษณะของเนื้อหาวิชาที่ต้องการวัดด้วย

ตอนที่สอง

ตอนที่สองนี้เป็นเรื่องเกี่ยวกับรูปแบบการตอบสนองข้อสอบ (Item Response Patterns) ประกอบด้วยความสำคัญของรูปแบบการตอบสนองข้อสอบ ดัชนีแสดงความเหมาะสมของรูปแบบการตอบสนองข้อสอบ และการศึกษาวิจัยเกี่ยวกับดัชนีแสดงความเหมาะสมของรูปแบบการตอบสนองข้อสอบ

ความสำคัญของรูปแบบการตอบสนองข้อสอบ

ข้อมูลการตอบสนองข้อสอบข้อหนึ่งของคน ๆ หนึ่งนั้นบอกอะไรได้น้อยมากทั้งในแง่ของกระบวนการภายใต้การตอบสนองข้อสอบข้อนั้น หรือความสอดคล้อง (Relevance) ของข้อสอบข้อนั้นกับคุณลักษณะแฝงที่ต้องการวัด แต่เราจะทราบคุณลักษณะเฉพาะของแบบสอบได้จากการตรวจสอบเมตริกซ์ของการตอบสนองข้อสอบ (Response Matrix) เมตริกซ์ดังกล่าวเป็นเมตริกซ์ของเลข 0 และ 1 ซึ่งประกอบด้วย N แถวและ n สดมภ์ อันหมายถึงข้อมูลการตอบสนองข้อสอบผิดและถูกของคน N คนต่อข้อสอบ n ข้อ

การตรวจสอบเมตริกซ์ของการตอบสนองข้อสอบนั้นพิจารณาได้ 2 แนวทาง แนวทางแรกเป็นการตรวจสอบเวคเตอร์ในแนวตั้ง (Column Vectors) ของผู้สอบ N คนต่อข้อสอบแต่ละข้อ และแนวทางที่สองเป็นการตรวจสอบเวคเตอร์ในแนวนอน (Row Vectors) ของการตอบสนองข้อสอบ n ข้อของแต่ละคน

แนวทางแรกของการตรวจสอบเมตริกซ์ของการตอบสนองข้อสอบนั้นใช้เปรียบเทียบการตอบสนองของกลุ่มผู้สอบต่อข้อสอบข้อเดียวกัน ผู้สอบที่มีความสามารถเหมือน ๆ กันแม้จะมาจากคนละกลุ่มก็ควรจะมีความน่าจะเป็นเท่า ๆ กันในการตอบข้อสอบข้อนั้นได้ถูกต้อง หากไม่เป็นเช่นนั้นแล้วก็กล่าวได้ว่าข้อสอบข้อนั้นลำเอียง (Biased)

ส่วนแนวทางที่สอง ใช้ในการหารูปแบบของการตอบสนอง (Patterns of Responses) ที่มีการเบี่ยงเบน (Deviance) ไปจากรูปแบบที่คาดหวัง (Expected Patterns) รูปแบบการตอบสนองที่คาดหวังนี้คือ ในกรณีของผู้สอบที่มีความสามารถสูงควรจะตอบข้อสอบข้อง่าย ๆ ได้ถูกต้องทุกข้อ ส่วนผู้ที่มีความสามารถต่ำไม่ควรจะตอบข้อสอบข้อยาก ๆ ได้ถูกต้อง ฉะนั้น เราจะพบความเบี่ยงเบนไปจากรูปแบบที่คาดหวังเมื่อผู้สอบที่มีความสามารถสูงตอบข้อสอบข้อง่าย ๆ ผิดหลายข้อ ผู้สอบที่มีความสามารถต่ำตอบข้อสอบข้อยาก ๆ ถูกหลายข้อ ผู้สอบเว้นไม่ตอบข้อง่าย ๆ หลายข้อ หรือผู้สอบตอบข้อสอบแบบสุ่มทั้งฉบับ (Hulin and

others 1983 : 110) ฉะนั้นการตรวจสอบเมตริกซ์การตอบสนองข้อสอบในแนวทางที่สองนี้จึงสามารถใช้ตรวจสอบความเหมาะสม(Appropriateness)ของคะแนนสอบ (Test score) กับความสามารถของผู้สอบได้ LevineและRubin(1979)เป็นผู้ริเริ่มใช้คำว่า Appropriateness Measurement ซึ่งหมายถึงวิธีการใด ๆ ก็ตามที่ใช้ในการตรวจสอบ (Detecting) ว่าผู้สอบคนใดมีคะแนนสอบที่ไม่ได้แสดงถึงการวัดลักษณะแฝงที่สนใจ หรือมีคะแนนสอบที่ไม่เหมาะสมกับความสามารถ

คะแนนสอบที่ไม่เหมาะสมกับความสามารถได้แก่ คะแนนซึ่งสูงกว่าที่ควรจะเป็น (Spuriously High) และคะแนนต่ำกว่าที่ควรจะเป็น (Spuriously Low) อันเนื่องมาจากสาเหตุหลายประการดังได้กล่าวไว้แล้วในบทที่ 1 การตรวจสอบคะแนนที่ไม่เหมาะสมมีความสำคัญมากในสถานการณ์ต่าง ๆ ดังเช่น ในการสอบคัดเลือกเข้าศึกษาผู้ที่ขาดคุณสมบัติแต่ได้คะแนนสูงกว่าที่ควรจะเป็นอาจได้รับการคัดเลือก หรือในโปรแกรมคัดสรรทางการศึกษา (Selective Academic Programs) การรับนักเรียนที่ได้คะแนนสูงกว่าที่ควรจะเป็นทำให้ผู้ที่สมควรจะได้เข้าโปรแกรมพลาดโอกาสไป นอกจากนี้คะแนนสอบที่ไม่เหมาะสมยังนำไปสู่ปัญหาในการสอบเพื่อการจ้างงานด้วย ดังเช่น คะแนนสูงกว่าที่ควรจะเป็นมีผลให้บุคคลที่ขาดคุณสมบัติได้รับเลือกเข้าทำงานหรือเข้ารับการฝึกอบรม บุคคลเหล่านี้อาจจะไม่ประสบความสำเร็จในงานหรือการฝึกอบรมก็ได้ ผู้ที่ทำงานไม่สำเร็จหรือไม่ประสบความสำเร็จในการฝึกอบรมย่อมทำให้หน่วยงานสูญเสียค่าใช้จ่ายโดยไร้ประโยชน์

สำหรับการสอบเพื่อวัดผลการเรียนนั้นผู้ที่ได้คะแนนสูงกว่าที่ควรจะเป็นอาจได้รับการประเมินว่าผ่านจุดประสงค์การเรียนรู้หรือผ่านรายวิชานั้น ๆ โดยที่มีความรู้ไม่เพียงพอผลเสียก็จะตกแก่นักเรียนคนนั้นในการเรียนจุดประสงค์หรือรายวิชาต่อ ๆ ไปที่จำเป็นต้องอาศัยความรู้ในจุดประสงค์หรือรายวิชานั้นเป็นพื้นฐาน ส่วนผู้ที่ได้คะแนนต่ำกว่าที่ควรจะเป็นอาจได้รับการประเมินว่าไม่ผ่านจุดประสงค์การเรียนรู้หรือรายวิชานั้น ๆ โดยที่แท้จริงแล้วเขามีความรู้อย่างพอเพียง ผลเสียที่จะเกิดขึ้นก็คือครูต้องเสียเวลามาสอนซ่อมเสริมให้แก่เขาโดยไม่จำเป็น ตัวนักเรียนเองก็เบื่อหน่ายที่จะต้องมาเรียนในสิ่งที่รู้แล้วซ้ำซาก ในที่สุดอาจทำให้นักเรียนคนนั้นเบื่อหน่ายต่อการเรียนหรือไม่สนใจเรียนแล้วก่อกวนคนอื่น ๆ ก็ได้

ดัชนีแสดงความเหมาะสมของรูปแบบการตอบสนองข้อสอบ

ประมาณปลายทศวรรษ 1970 เป็นต้นมา ได้เริ่มมีผู้สนใจวิเคราะห์รูปแบบการตอบสนองข้อสอบเพิ่มขึ้น ได้มีการคิดค้นและเสนอวิธีวิเคราะห์ต่าง ๆ ซึ่งอาจจัดอย่างกว้าง ๆ ตามลักษณะการตอบที่นำมาวิเคราะห์ได้เป็น 2 พวก (Harnisch 1983 : 193) พวกแรกเป็นการวิเคราะห์การตอบสนองที่ผู้ตอบเรียบเรียงคำตอบขึ้นเอง (Constructed Response) อีกพวกหนึ่งเป็นการวิเคราะห์การตอบสนองที่ผู้ตอบเลือกจากตัวเลือกที่กำหนดให้ (Selected Response)

1. การวิเคราะห์การตอบสนองที่ผู้ตอบเรียบเรียงคำตอบขึ้นเอง

การวิเคราะห์การตอบสนองที่ผู้ตอบเรียบเรียงคำตอบขึ้นเองนี้ ใช้การวิเคราะห์ความผิดพลาดตามความรู้สึกของผู้วิเคราะห์ (Intuitive Error Analysis) หรือใช้การสัมภาษณ์แบบเจาะลึก (Extensive Clinical Interviews) วิธีนี้ความถูกต้องเชื่อถือได้ขึ้นอยู่กับประสบการณ์และการตัดสินของผู้วิเคราะห์อีกทั้งเป็นวิธีที่ใช้เวลาและค่าใช้จ่ายสูง

2. การวิเคราะห์การตอบสนองที่ผู้ตอบเลือกจากตัวเลือกที่กำหนดให้

การวิเคราะห์การตอบสนองที่ผู้ตอบเลือกจากตัวเลือกที่กำหนดให้นี้ใช้การวิเคราะห์หาดัชนีจากรูปแบบการตอบถูกหรือตอบผิดเพื่อชี้ถึงขนาด (Degree) ที่รูปแบบการตอบสนองของผู้สอบคนหนึ่ง ๆ ผิดไปจากรูปแบบที่คาดหวัง ดัชนีที่แสดงถึงขนาดความผิดปกติของรูปแบบการตอบสนองของผู้สอบเป็นรายคนนี้อาจแบ่งได้เป็น 2 ประเภท คือ ดัชนีที่ขึ้นอยู่กับรูปแบบการตอบสนองของคนอื่น ๆ ในกลุ่ม (Group - Dependent Indicies) และดัชนีที่ใช้ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory - Based Indicies)

2.1 ดัชนีที่ขึ้นอยู่กับรูปแบบการตอบสนองของคนอื่น ๆ ในกลุ่ม

ดัชนีประเภทนี้วิเคราะห์จากข้อมูลการตอบสนองข้อสอบที่ได้จากกระดาษคำตอบและค่าสถิติเบื้องต้นของข้อมูลเหล่านั้น เช่น จำนวนหรือสัดส่วนของคนในกลุ่มอ้างอิง (Norm Group) ที่ตอบข้อสอบถูก เนื่องจากดัชนีในประเภทนี้จะชี้ว่ารูปแบบการตอบสนองของผู้สอบคนหนึ่งผิดไปจากรูปแบบการตอบสนองของผู้สอบคนอื่น ๆ ในกลุ่มอ้างอิงหรือไม่ ฉะนั้น Harnisch (1983 : 194) จึงเรียกดัชนีกลุ่มนี้ว่า Group - Dependent Indicies ซึ่งได้แก่

Caution Index ที่เสนอโดย Sato (1975)

Modified Caution Index เสนอโดย Harnisch และ Linn (1981)

U' Index เสนอโดย van de Flier (1977)

Personal Biserial เสนอโดย Donlon และ Fischer (1968)

Norm Conformity Index เสนอโดย Tatsuoka และ Tatsuoka (1982) และ Agreement and Disagreement Indicies ที่เสนอโดย Kane และ Brennan (1980)

เนื่องจากดัชนีหลายตัวในกลุ่มนี้พัฒนามาจาก S - P Curve Theory ฉะนั้นจึงขอกล่าวถึงทฤษฎีนี้ก่อนที่จะกล่าวถึงรายละเอียดของดัชนีในกลุ่มนี้ต่อไป

S - P Curve Theory

S - P Curve Theory นี้ Sato (1975 อ้างถึงใน Tatsuoka and Linn 1983 : 82 - 84) เป็นผู้คิดขึ้น เริ่มจากเมตริกซ์ข้อมูลการตอบถูกหรือตอบผิดของผู้สอบ (Data Matrix) โดยในแนวแถว (Row) หมายถึงผู้สอบ แนวสดมภ์ (Column) หมายถึงข้อสอบ เลข 1 หมายถึงตอบถูก 0 หมายถึงตอบผิด จัดเรียงแถวและสดมภ์สลับไปมาจนกระทั่งแถวเรียงลำดับจากบนลงล่างตามจำนวนข้อที่ตอบถูกจากมากมาน้อย ส่วนสดมภ์เรียงลำดับจากซ้ายไปขวาตามลำดับความยาก (สัดส่วนของผู้ตอบข้อนั้นถูก) ที่เพิ่มขึ้น เมตริกซ์ที่ได้นี้เรียกว่า Student - Problem (S - P) Table ของ Sato

ถ้าข้อสอบเหล่านั้นก่อตัวเป็นกัทท์แมนสเกลที่สมบูรณ์ (Perfect Guttman Scale) ในมุมบนซ้ายของ S - P Table จะประกอบด้วยเลข 1 ทั้งหมด และในมุมล่างขวาจะประกอบด้วยเลข 0 ทั้งหมดเหมือนกับมีเส้นหนึ่งแบ่งเขต 1 และ 0 ซึ่งหมายความว่า ใครก็ตามที่ตอบข้อยากถูกก็ควรจะตอบข้อที่ง่ายกว่าได้ถูกทั้งหมด แต่ในความเป็นจริงนั้นเราไม่อาจคาดหวังให้มีคะแนนที่สมบูรณ์เช่นนี้ในข้อสอบวัดผลสัมฤทธิ์ได้ ดังนั้น S - P Table ในกรณีดังกล่าว ในมุมบนซ้ายจะมีเลข 1 มากกว่าและในมุมล่างขวาจะมีเลข 0 มากกว่า

ตัวอย่างของ S - P Table ในตารางที่ 1 มีผู้สอบ 15 คน ข้อสอบ 10 ข้อ เส้นทึบและเส้นประหมายถึง S - Curve และ P - Curve ตามลำดับ S - Curve (เส้นทึบ) ได้จากการลากเส้นในแนวตั้งสำหรับแต่ละแถวทางขวามือของสดมภ์ที่ Y_i เมื่อ Y_i คือจำนวนข้อที่คนที่ i ตอบถูก แล้วลากเส้นในแนวนอนเชื่อมปลายบนของเส้นซ้ายกับปลายล่างของ

เส้นขวาที่อยู่ติดกันไปจนครบทุกเส้นตั้งจะได้เส้นชั้นบันไดของ S - Curve ส่วน P - Curve (เส้นประ) ได้จากการลากเส้นในแนวนอนสำหรับแต่ละสดมภ์ใต้แถวที่ $Y_{.j}$ เมื่อ $Y_{.j}$ คือ จำนวนคนที่ตอบข้อ j ถูก แล้วลากเส้นตั้งเชื่อมปลายขวาของเส้นทางซ้ายกับปลายซ้ายของเส้นนอนทางขวาที่อยู่ถัดกันไปจนครบทุกเส้นนอนจะได้ P - Curve

โดยอุดมคติเมื่อเป็นภัทท์แมนสเกลที่สมบูรณ์แล้ว S - Curve และ P - Curve จะทับกัน แต่ในความเป็นจริง S - Curve และ P - Curve จะแยกออกจากกัน ซึ่งขนาดของการแยกออกจากกันนี้ชี้ถึงขนาดของความเป็นเอกพันธ์ของรูปแบบการตอบสนองข้อสอบซึ่ง Sato ได้พัฒนาดัชนีที่อาศัยพื้นที่ระหว่าง S - Curve และ P - Curve นี้สำหรับประเมินความเป็นเอกพันธ์ของแบบสอบ ดัชนีนี้คือ Sato's Caution Index ซึ่งจะกล่าวถึงในลำดับต่อไป

ตารางที่ 1 เมตริกซ์ของคะแนนสมมุติ (Hypothetical Score Matrix (Y_{ij})) กับ S - Curve (เส้นทึบ) และ P - Curve (เส้นประ)

คนที่ i	ข้อที่ j										$Y_{i.}$	$P_{i.}$	$M_{i.}^{S}$
	1	2	3	4	5	6	7	8	9	10			
1	1	1	1	1	1	1	1	1	1	1	10	1.0	10
2	1	1	1	1	1	1	1	1	1	0	9	0.9	9
3	1	1	1	1	1	0	1	1	0	1	8	0.8	8
4	1	0	1	1	1	1	0	1	0	0	6	0.6	6
5	1	1	1	1	0	1	0	0	1	0	6	0.6	6
6	1	1	1	0	1	0	1	0	1	0	6	0.6	6
7	1	1	1	1	0	0	1	0	0	0	5	0.5	5
8	1	1	1	0	1	1	0	0	0	0	5	0.5	5
9	1	0	0	1	0	1	0	1	1	0	5	0.5	5
10	1	1	0	1	0	0	1	0	0	1	5	0.5	5
11	0	1	1	1	1	0	0	0	0	0	4	0.4	4
12	1	0	0	0	1	1	0	0	0	1	4	0.4	4
13	1	1	0	0	0	1	0	0	0	0	3	0.3	3
14	1	0	1	0	0	0	0	0	0	0	2	0.2	2
15	0	1	0	0	0	0	0	0	0	0	1	0.1	1
$Y_{.j}$	13	11	10	9	8	8	6	5	5	4	$Y_{..}$ = 79		
$P_{.j}$	.87	.73	.67	.60	.53	.53	.40	.33	.33	.27	$P_{..}$ = .527		
$M_{.j}^{P}$	13	11	10	9	8	8	6	5	5	4			

(Tatsuoka and Linn 1983 : 83)

ตารางที่ 2 S - Curve สมบูรณ์ที่ได้จากการเปลี่ยน 1 ข้างขวาของ S - Curve เป็น 0 และเปลี่ยน 0 ข้างซ้ายของ S - Curve เป็น 1

คนที่ i	ข้อที่ j										$Y_{i.}$	$M_{i.}^{S}$
	1	2	3	4	5	6	7	8	9	10		
1	1	1	1	1	1	1	1	1	1	1	10	10
2	1	1	1	1	1	1	1	1	1	0	9	9
3	1	1	1	1	1	1	1	1	0	0	8	8
4	1	1	1	1	1	1	0	0	0	0	6	6
5	1	1	1	1	1	1	0	0	0	0	6	6
6	1	1	1	1	1	1	0	0	0	0	6	6
7	1	1	1	1	1	0	0	0	0	0	5	5
8	1	1	1	1	1	0	0	0	0	0	5	5
9	1	1	1	1	1	0	0	0	0	0	5	5
10	1	1	1	1	1	0	0	0	0	0	5	5
11	1	1	1	1	0	0	0	0	0	0	4	4
12	1	1	1	1	0	0	0	0	0	0	4	4
13	1	1	1	0	0	0	0	0	0	0	3	3
14	1	1	0	0	0	0	0	0	0	0	2	2
15	1	0	0	0	0	0	0	0	0	0	1	1
$Y_{.j}$	13	11	10	9	8	8	6	5	5	4	79	79
$M_{.j}^{P}$	15	14	13	12	10	6	3	3	2	1		

(Tatsuoka and Linn 1983 : 84)

ตารางที่ 3 P - Curve สมบูรณ์ที่ได้จากการเปลี่ยน 1 ใต้ P - Curve เป็น 0 และเปลี่ยน 0 เหนือ P - Curve เป็น 1

คนที่ i	ข้อที่ j										$Y_{i.}$	$M_{i.}^{S}$
	1	2	3	4	5	6	7	8	9	10		
1	1	1	1	1	1	1	1	1	1	1	10	10
2	1	1	1	1	1	1	1	1	1	1	9	10
3	1	1	1	1	1	1	1	1	1	1	8	10
4	1	1	1	1	1	1	1	1	1	1	6	10
5	1	1	1	1	1	1	1	1	1	0	6	9
6	1	1	1	1	1	1	1	0	0	0	6	7
7	1	1	1	1	1	1	0	0	0	0	5	6
8	1	1	1	1	1	1	0	0	0	0	5	6
9	1	1	1	1	0	0	0	0	0	0	5	4
10	1	1	1	0	0	0	0	0	0	0	5	3
11	1	1	0	0	0	0	0	0	0	0	4	2
12	1	0	0	0	0	0	0	0	0	0	4	1
13	1	0	0	0	0	0	0	0	0	0	3	1
14	0	0	0	0	0	0	0	0	0	0	2	0
15	0	0	0	0	0	0	0	0	0	0	1	0
$Y_{.j}$	13	11	10	9	8	8	6	5	5	4	79	
$M_{.j}^{P}$	13	11	10	9	8	8	6	5	5	4	79	

(Tatsuoka and Linn 1983 : 85)

Caution Index

Caution Index ถูกนำมาใช้อย่างกว้างขวางในประเทศญี่ปุ่น สำหรับวินิจฉัยผลการปฏิบัติ (Performance) ของนักเรียน สำหรับตรวจสอบรูปแบบการตอบสนองข้อสอบที่ผิดปกติและสำหรับประเมินคุณภาพของแบบสอบและลำดับการสอน (Tatsuoka and Linn 1983 : 82) Caution Index ใช้ได้ทั้งกับผู้สอบและข้อสอบ กรณีที่ใช้กับผู้สอบเป็นดัชนีที่ให้สารสนเทศเกี่ยวกับผู้สอบคนหนึ่งซึ่งสารสนเทศนี้ไม่มีอยู่ในคะแนนรวม เป็นดัชนีที่บอกให้ระมัดระวังในการตีความหมายของคะแนนรวมตามปกติ ดัชนีนี้ยิ่งมีค่ามากก็ยิ่งต้องระมัดระวังในการตีความหมายคะแนนรวมมากขึ้น ชื่อของดัชนีนี้มาจากข้อสังเกตที่ว่า ค่ามากสัมพันธ์กับผู้สอบที่มีรูปแบบการตอบสนองข้อสอบที่ผิดปกติ รูปแบบที่ผิดปกตินี้อาจเป็นผลจากการเดา ความสะเพร่า ความวิตกกังวลสูง ประวัติการสอนที่ผิดธรรมดา หรือภูมิหลังประสบการณ์อื่น ๆ ความเข้าใจผิด การลอกคำตอบจากเพื่อน (Harnisch and Linn 1981 : 135)

นิยามของ Caution Index พัฒนามาจาก S - P Curve Theory (Tatsuoka and Linn 1983 : 82 - 84) โดยให้ Y_{ij} เป็นการตอบสนองถูกหรือผิด (Binary Response) ของนักเรียนคนที่(แถว) i ต่อข้อสอบข้อที่(สดมภ์) j ของ S - P Table ผลรวมของแถวและสดมภ์ให้เป็น $Y_{i.}$ และ $Y_{.j}$ ตามลำดับ จำนวนเลข 1 ทั้งหมดใน S - P Table แทนด้วย Y.. และสัดส่วนของการตอบถูกแทนด้วย $P_{i.}$, $P_{.j}$ และ P.. สำหรับแถว สดมภ์ และทั้งตาราง ตามลำดับ ดังที่เห็นในตารางที่ 1 S - Curve เป็น Step Function Ogive ของฟังก์ชันการแจกแจงสะสมของคะแนนรวม ($Y_{i.}$) สำหรับนักเรียน 15 คน ส่วน P - Curve เป็นฟังก์ชันที่สัมพันธ์กันของจำนวนคำตอบถูก ($Y_{.j}$) สำหรับข้อสอบ 10 ข้อ

ถ้าให้ S - Curve คงที่ (Invariant) และเลข 0 ทั้งหมดที่อยู่ทางซ้ายของ S - Curve เปลี่ยนเป็นเลข 1 และเลข 1 ทั้งหมดที่อยู่ทางขวาของ Curve เดียวกันนี้เปลี่ยนเป็นเลข 0 ผลคือ S - P Table ที่แสดงไว้ในตารางที่ 2 S - Curve ที่ได้นี้เรียกว่า Perfect S - Curve สมาชิกในตารางที่ 2 นี้แทนด้วย M_{ij}^{S} ในทำนองเดียวกันจะได้ Perfect P - Curve ซึ่งแสดงไว้ในตารางที่ 3 สมาชิกในตารางที่ 3 นี้ให้เป็น M_{ij}^{P} จะเห็นว่า $M_{i.}^{S} = Y_{i.}$ สำหรับทุก i ซึ่งสอดคล้องกับข้อเท็จจริงที่ว่า S - Curve ไม่เปลี่ยนอันเป็นผลลัพท์จากการเปลี่ยนเซลสมาชิกจาก Y_{ij} เป็น M_{ij}^{S}

Sato (1975 อ้างถึงใน Tatsuoka and Linn 1983 : 82) ได้นิยาม Caution Index สำหรับคนที่ i โดยใช้อัตราส่วนของความแปรปรวนร่วม(Covariance) 2 ตัว ตัวเศษของอัตราส่วนคือความแปรปรวนร่วมของเวคเตอร์แถวที่ i ที่สังเกตได้(Observed

Row Vector i) (Y_{ij}), j = 1,...,n และเวคเตอร์ของผลรวมในแนวตั้ง (Sum - of - Column Vector) ($Y_{.j}$), j = 1,...,n และตัวส่วนคือ ความแปรปรวนร่วมของคะแนนที่สัมพันธ์กัน สมมุติว่าเป็น Perfect S - Curve (M_{ij}^{S}), j = 1,...,n และ Column - Sum Vector ($Y_{.j}$), j = 1,...,n Caution Index C_i ของคนที่ i คำนวณได้จากสมการต่อไปนี้

$$C_i = 1 - \frac{\sum_{j=1}^{n} (Y_{ij} - P_{i.})(Y_{.j} - P..)}{\sum_{j=1}^{n} (M_{ij}^{S} - P_{i.})(Y_{.j} - P..)} \qquad (2\text{-}1)$$

และ Caution Index C_j สำหรับข้อที่ j คำนวณได้จากสมการต่อไปนี้

$$C_j = 1 - \frac{\sum_{i=1}^{N} (Y_{ij} - P_{.j})(Y_{i.} - P..)}{\sum_{i=1}^{N} (M_{ij}^{P} - P_{.j})(Y_{i.} - P..)} \qquad (2\text{-}2)$$

เทอมที่สองของ Caution Index สำหรับข้อที่ j เป็นอัตราส่วนของความแปรปรวนร่วมสองตัว ตัวเศษคือ ความแปรปรวนร่วมของ คอลัมน์เวคเตอร์ j (Y_{ij}) และ ($Y_{i.}$), i=1,...,N และตัวส่วนคือ ความแปรปรวนร่วมของ เวคเตอร์ ($Y_{i.}$) และ (M_{ij}^{P}), i=1,...,N ค่าของตัวส่วนจัดให้เป็นค่านอร์ม (Norm Value) เพื่อทำตัวเศษให้เป็นมาตรฐาน(Standardize)

กล่าวได้ว่า อัตราส่วนของ Caution Index ข้างต้นเท่ากับอัตราส่วนของดัชนีอำนาจจำแนกดั้งเดิม (Traditonal Discrimination Index) คือ r_j (Item - Total Correlation) ต่อ ดัชนีอำนาจจำแนกที่ทำให้เป็นมาตรฐาน (Standardized Discriminating Index) คือ r_j' สำหรับข้อ j นั่นคือ

$$\frac{\text{Cov}_j\,(Y_{ij},Y_{i.})}{\text{Cov}_j\,(M_{ij}{}^{P},Y_{i.})} = \frac{\dfrac{\text{Cov}_j\,(Y_{ij},Y_{i.})}{\sigma_j\,(Y_{ij})\,\sigma\,(Y_{i.})}}{\dfrac{\text{Cov}_j\,(M_{ij}{}^{P},Y_{i.})}{\sigma_j\,(M_{ij}{}^{P})\,\sigma\,(Y_{i.})}} = \frac{r_j}{r_j'} \qquad (2\text{-}3)$$

จะเห็นว่า $(Y_{ij} - P_{.j})^2 = (M_{ij}{}^{P} - P_{.j})^2$ เพราะว่าจำนวนเลข 1 ในสดมภ์ j นั้นไม่แปรเปลี่ยน (Invariant) ดังที่เห็นในตารางที่ 1 และ 3 ดังนั้น จำนวนเลข 1 ใน คอลัมน์เวคเตอร์ j $(M_{ij}{}^{P})$ และ (Y_{ij}) เท่ากัน ความแปรปรวนสองตัว คือ $\sigma_j{}^2\,(Y_{ij})$ และ $\sigma_j{}^2\,(M_{ij}{}^{P})$ จึงเท่ากันด้วย

หากไม่เทียบกับ Perfect S - Curve และ Perfect P - Curve ก็จะนิยาม Caution Index ได้ดังนี้ (Harnisch and Linn 1981 : 135)

Caution Index C_i สำหรับคนที่ i

$$C_i = \frac{\sum^{n1.}_{j=1}\,(1 - U_{ij})\,n_{.j} - \sum^{J}_{j=n1.+1} U_{ij}\,n_{.j}}{\sum^{n1.}_{j=1} n_{.j} - n_{i.}\,\{[\,\sum^{J}_{j=1} n_{.j}\,]/J\}} \qquad (2\text{-}4)$$

เมื่อ i = 1,2,...,I หมายถึงผู้สอบ

j = 1,2,...,J หมายถึงข้อสอบ

$$U_{ij} = \begin{cases} 1 & \text{ถ้าผู้สอบคนที่ i ตอบข้อ j ได้ถูกต้อง} \\ 0 & \text{ถ้าผู้สอบคนที่ i ตอบข้อ j ผิด} \end{cases}$$

$n_{i.}$ = จำนวนข้อที่คนที่ i ตอบถูก

$n_{.j}$ = จำนวนคนที่ตอบข้อ j ถูก

Caution Index C_j สำหรับข้อที่ j คือ

$$C_j = \frac{\sum_{i=1}^{n.j} (1 - U_{ij})\, n_{i.} - \sum_{i=n.j+1}^{I} U_{ij}\, n_{i.}}{\sum_{i=1}^{n.j} n_{i.} - n_{.j} \{[\sum_{i=1}^{I} n_{i.}]/I\}} \qquad (2\text{-}5)$$

นอกจากนี้ Harnisch และ Linn (1981 : 138) ยังได้เสนอการหาค่า C_i ที่ง่ายกว่าในสมการ (2-4) ดังนี้

$$C_i = \frac{J \sum_{j=1}^{ni.} n_{.j} - J \sum_{j=1}^{J} U_{ij}\, n_{.j}}{J \sum_{j=1}^{ni.} n_{.j} - n_{i.} \sum_{j=1}^{J} n_{.j}} \qquad (2\text{-}6)$$

Modified Caution Index

เนื่องจากในบางกรณี เช่น ผู้ที่มีความสามารถต่ำได้คะแนนรวมต่ำแต่ตอบข้อยาก ๆ ได้ถูกต้อง หรือผู้ที่มีความสามารถสูงได้คะแนนรวมสูงแต่ตอบข้อง่าย ๆ ผิด ในกรณีดังกล่าวจะได้ค่า Caution Index (C_i) สูงมาก คือเกิน 1 มาก เพื่อขจัดปัญหานี้โดยเฉพาะในกรณีผู้ที่ได้คะแนนสูงมาก ๆ แต่ผิดข้อง่าย ๆ เพียงข้อเดียว Harnisch และ Linn (1981 : 135) จึงได้ปรับ Caution Index (C_i) มาเป็น Modified Caution Index (C_i^*) เพื่อให้ดัชนีมีค่าตั้งแต่ 0 ถึง 1 โดยนิยาม C_i^* สำหรับผู้สอบคนที่ i ดังนี้

$$C_i^* = \frac{\sum_{j=1}^{ni.} (1 - U_{ij}) n_{.j} - \sum_{j=ni.+1}^{J} U_{ij} n_{.j}}{\sum_{j=1}^{ni.} n_{.j} - \sum_{j=J+1-ni.}^{J} n_{.j}} \qquad (2\text{-}7)$$

U' Index

van der Flier (ใน Poortinga ed. 1977 :31) ได้พัฒนาวิธีการอย่างหนึ่งขึ้นมาเพื่อประเมินความเปรียบเทียบกันได้ของคะแนนสอบของแต่ละคนโดยมีจุดมุ่งหมายที่การกำหนดขนาดที่รูปแบบคะแนนเฉพาะ(Specific Score Pattern)ของแต่ละคนเบี่ยงเบนไปจากรูปแบบคะแนนที่คาดหวัง(Expected Score Pattern) รูปแบบคะแนนที่คาดหวังนี้คำนวณได้จากคะแนนเฉลี่ยรายข้อ(Average Item Scores)ของกลุ่มตัวอย่าง ถ้าข้อสอบเป็นแบบตอบถูกได้ 1 ตอบผิดได้ 0 (Dichotomous) ก็คือค่า p นั่นเอง ถ้าข้อสอบเรียงตามลำดับความยาก (จากง่ายไปหายาก) • ในกรณีของผู้สอบที่ตอบถูก 50 % ของข้อสอบก็สามารถคาดได้ว่า ข้อสอบในครึ่งแรกส่วนมากจะตอบถูกและข้อสอบในครึ่งหลังจะตอบผิดเป็นส่วนมาก ถ้าความจริงเป็นตรงกันข้ามแล้วรูปแบบคะแนนจะเบี่ยงเบนไปจากรูปแบบที่คาดหวังมาก การวัดที่ชี้ถึงขนาดที่รูปแบบการตอบจริง ๆ (Actual Response Pattern) เบี่ยงเบนไปจากรูปแบบที่คาดหวังนี้โดยนิยาม รูปแบบ ก.จะกล่าวว่าเบี่ยงเบนมากกว่ารูปแบบ ข. ถ้าความน่าจะเป็นของรูปแบบ ก. ที่กำหนดคะแนนสอบให้นั้นน้อยกว่าความน่าจะเป็นของรูปแบบ ข.

เพื่อให้สามารถคำนวณความน่าจะเป็นของรูปแบบเฉพาะใด ๆ van der Flier จึงกำหนดข้อตกลงเบื้องต้นขึ้น 2 ประการเกี่ยวกับคุณสมบัติของแบบสอบในกลุ่มอ้างอิง คือกลุ่มที่ใช้พิจารณาค่าความยากของข้อสอบ ข้อตกลงเบื้องต้นนี้ คือ

1. คะแนนของข้อสอบเป็น Stochastically Independent หมายความว่า ความน่าจะเป็นของรูปแบบคะแนนเป็นผลคูณของความน่าจะเป็นของการตอบถูกหรือตอบผิดในข้อต่าง ๆ ดังแสดงในตารางที่ 4

2. เส้นแนวทาง (Trace Line) ของข้อสอบเป็นแบบสูงขึ้นหรือต่ำลงอย่างเดียว (Monotonic) เส้นแนวทางของข้อสอบข้อหนึ่ง ๆ ชี้ถึงสัดส่วนของผู้ตอบที่ตอบข้อนั้นถูกโดยเป็นฟังก์ชันของตำแหน่งของผู้ตอบบนความต่อเนื่อง(Underlying Continuum) หรือเป็นฟังก์ชันของความสามารถของเขา เช่นข้อสอบที่มีเส้นแนวทางสูงขึ้นอย่างเดียว(Monotonically Increasing) หมายความว่า ผู้สอบยิ่งมีความสามารถสูงความน่าจะเป็นในการตอบข้อสอบข้อนั้นถูกก็ยิ่งสูง

ตารางที่ 4 รูปแบบของคะแนนที่มีคะแนนรวม 2 พร้อมทั้งความน่าจะเป็นของรูปแบบและคะแนนรวม

รูปแบบของคะแนน (คะแนนรวม 2)	ข้อที่ (ความน่าจะเป็นในการตอบถูก)				ความน่าจะเป็นของรูปแบบคะแนน	คะแนนเบี่ยงเบน
	1(.70)	2(.60)	3(.40)	4(.30)		
ก	1	1	0	0	.176	1.000
ข	1	0	1	0	.078	.560
ค	1	0	0	1	.050	.365
ง	0	1	1	0	.050	.365
จ	0	1	0	1	.032	.115
ฉ	0	0	1	1	.014	.035
					= .400	

(van der Flier ใน Poortinga ed. 1977 : 33)

ตารางที่ 4 นี้แสดงข้อสอบ 4 ข้อซึ่งมีผู้ตอบถูก 70,60,40 และ 30 % ตามลำดับ รวมทั้งแสดงคำตอบแบบต่าง ๆ ที่มีคะแนนรวมเป็น 2 โดยเรียงตามลำดับความน่าจะเป็นของรูปแบบคะแนน ข้อที่ตอบถูกได้ 1 ตอบผิดได้ 0

ความน่าจะเป็นของรูปแบบคะแนนของแต่ละคนได้จากผลคูณของความน่าจะเป็นในการตอบถูกหรือตอบผิดของทุกข้อ ดังเช่น ความน่าจะเป็นของรูปแบบ ก คือ

$$(.70)(.60)(1 - .40)(1 - .30) = .176$$

คะแนนเบี่ยงเบนของแต่ละรูปแบบได้จากการรวมความน่าจะเป็นของรูปแบบที่มีความน่าจะเป็นเท่ากันหรือน้อยกว่าแล้วหารด้วยความน่าจะเป็นของทุกรูปแบบรวมกัน เช่น คะแนนเบี่ยงเบนของรูปแบบ ค คือ

$$(.050 + .050 + .032 + .014) / .400 = .365$$

คะแนนเบี่ยงเบนนี้ชี้ถึงขนาดของความเบี่ยงเบนของรูปแบบคะแนนที่เบี่ยงเบนไปจากรูปแบบคะแนนในกลุ่มอ้างอิงที่มีคะแนนรวมเท่ากัน ค่ายิ่งต่ำก็ยิ่งแสดงว่ามีความเบี่ยงเบนมาก

วิธีการดังกล่าวนี้ค่อนข้างยุ่งยากในกรณีที่มีข้อสอบจำนวนมากข้อ van der Flier จึงพัฒนาวิธีที่ง่ายกว่าขึ้นโดยใช้ลำดับที่ความยากแทนค่า p ของแต่ละข้อ แต่เนื่องจากลำดับความยากนั้นอาจแปรเปลี่ยนไปตามระดับความสามารถ เพราะฉะนั้น ลำดับที่ความยากของข้อสอบจึงควรพิจารณาแยกกันในแต่ละระดับคะแนน ดังแสดงในตารางที่ 5

ตารางที่ 5 รูปแบบคะแนนที่มีคะแนนรวม 2 ค่า U และ U'

รูปแบบ \ ข้อที่	1(.70)	2(.60)	3(.40)	4(.30)	U	U'
ก	1	1	0	0	0	0
ข	1	0	1	0	1	1/4
ค	1	0	0	1	2	2/4
ง	0	1	1	0	2	2/4
จ	0	1	0	1	3	3/4
ฉ	0	0	1	1	4	1

(van der Flier ใน Poortinga ed. 1977 : 34)

ตารางที่ 5 แสดงข้อสอบ 4 ข้อตามลำดับความยากและทุกรูปแบบของคะแนนที่มีคะแนนรวมเป็น 2 เรียงตามลำดับความน่าจะเป็นเช่นเดียวกับในตารางที่ 4 ความเบี่ยงเบนของคะแนนแต่ละรูปแบบจากรูปแบบที่คาดหวังหาได้โดยการรวมจำนวนเลข 1 ที่อยู่ทางขวาของเลข 0 ทุกตัว จะได้ค่า U (เหมือนกับ Mann - Whitney U) ค่าต่ำสุดของ U คือ 0 และค่าสูงสุดเท่ากับจำนวนข้อที่ถูกคูณกับจำนวนข้อที่ผิด พิสัยของค่า U ทำให้เท่ากันได้สำหรับคะแนนต่าง ๆ โดยการหารด้วยค่าสูงสุดนี้ ค่าที่ได้คือ U' ซึ่งมีค่าตั้งแต่ 0.0 ถึง 1.0 รูปแบบใดที่มีค่า U' มากแสดงว่าเบี่ยงเบนไปจากรูปแบบที่คาดหวังมาก

Personal Biserial Index

Personal Biserial Index (r_i^*) (Donlon and Fischer 1968 อ้างถึงใน Harnisch and Linn 1981 : 137) ของแต่ละคนจะตรงกับ Personal Point - Biserial Index (r_i) โดย r_i^* มีข้อตกลงเบื้องต้นว่า ตัวแปรภายใต้ U_{ij} นั้นเป็นตัวแปรต่อเนื่องที่มีการแจกแจงปกติ

Personal Point - Biserial Index

Personal Point - Biserial Index (r_i) เป็นค่าสหสัมพันธ์ (Product Moment Correlation) ระหว่างคะแนนรายข้อของคนที่ i (U_{ij} = 0 หรือ 1 สำหรับแต่ละข้อ) กับจำนวนคนที่ตอบแต่ละข้อถูก ($n_{.j}$) $n_{.j}$ แทนด้วยค่าความยากก็ได้โดย $P_{.j} = n_{.j}/I$ ถ้าค่า r_i สูงแสดงว่าคนที่ i ตอบข้อง่ายสำหรับคนในกลุ่มได้ถูกต้องและตอบผิดในข้อที่ยากสำหรับคนกลุ่มนั้น ดังนั้น Personal Point - Biserial Index จึงเป็นดัชนีแสดงความสอดคล้องระหว่างการตอบของแต่ละคนกับกลุ่มที่ใช้พิจารณาความยาก แต่จุดอ่อนของ r_i ก็คือ ขอบเขตของ r_i ขึ้นอยู่กับจำนวนข้อที่ตอบถูกของแต่ละคน เช่น รูปแบบการตอบของคน 2 คนมีความสอดคล้องกับค่าความยากมากที่สุดพอ ๆ กัน แต่เนื่องจากจำนวนข้อที่ตอบถูกต่างกัน ค่า r_i ที่ได้จะต่างกัน หากตีความหมายตามค่า r_i ที่ได้ก็จะผิดไปจากความเป็นจริง

Personal Point - Biserial Index กับ Modified Caution Index มีความสัมพันธ์กันมาก (Brennan 1980 อ้างถึงใน Harnisch and Linn 1981 : 137) ดังนี้

$$C_i^* = \frac{\text{Max } (r_i) - r_i}{\text{Max } (r_i) - \text{Min } (r_i)} \qquad (2\text{-}8)$$

Max (r_i) และ Min (r_i) เป็นค่า r_i สูงสุดและต่ำสุดที่เป็นไปได้สำหรับ $n_{i.}$ ที่กำหนด

Norm Conformity Index

Tatsuoka และ Tatsuoka (1982 : 215 - 231) ได้พัฒนาดัชนีขึ้นมาตัวหนึ่ง คือ Consistency Index โดยเริ่มจากชุดข้อมูลของรูปแบบการตอบของคนหนึ่ง คือ Row Vector S ซึ่ง Dominance Matrix ของรูปแบบการตอบนี้คือ

$$S'S = N = (n_{ij}) \; ; \; i,j = 1,2,\ldots,n \text{ (ข้อที่)} \qquad (2\text{-}9)$$

เมื่อ S' เป็น Transpose ของ Complement ของ S โดย $n_{ij} = 1$ เมื่อคนนั้นตอบข้อ i ผิด ข้อ j ถูก นอกนั้น $n_{ij} = 0$

ถ้าลำดับของข้อใน S เปลี่ยนไป Dominance Matrix ก็จะเปลี่ยนไปด้วย Consistency Index ที่สัมพันธ์กับรูปแบบการตอบ S นิยามได้ดังนี้

$$C = (2U_a / U) - 1 \qquad (2\text{-}10)$$

เมื่อ $U_a = \sum_{i<j} n_{ij}$ (ผลรวมสมาชิกของ N ที่อยู่เหนือเส้นทแยง)
และ $U = \sum_{i,j} n_{ij}$ (ผลรวมของสมาชิกทั้งหมดของN) ต่างเป็นฟังก์ชันของลำดับข้อ คือ O จึงเขียนได้เป็น C(O)

C(O) จะมีค่าเป็น 1 เมื่อ S เป็น Guttman Vector ซึ่ง 0 ทุกตัวอยู่ข้างหน้า 1 ทุกตัว

C(O) จะมีค่าเป็น -1 เมื่อ S เป็น Reversed Guttman Vector คือ 1 มาก่อน 0

ถ้าลำดับข้อตรงกันข้ามกับใน S ค่าสัมบูรณ์ของ C(O) จะไม่เปลี่ยนแต่เครื่องหมายจะกลับกัน

เมื่อเรียงข้อสอบตามลำดับความยากที่ลดลงสำหรับกลุ่มเฉพาะกลุ่มหนึ่ง (Norm Group) Consistency Index C(O) ที่สัมพันธ์กับรูปแบบการตอบ S ของแต่ละคน เรียกว่า Norm Conformity Index ,NCI NCI ชี้ถึงขนาดที่ Vector การตอบ S เป็น Guttman Vector ที่มีจำนวน 1 เท่ากัน (O ทุกตัวอยู่ด้านซ้ายของ 1)

Agreement Index และ Disagreement Index

Brennan (1980 อ้างถึงใน Harnisch and Linn 1981 : 137) ได้ชี้ว่า Modified Caution Index สัมพันธ์อย่างใกล้ชิดกับ Agreement Index (A_i) และ Disagreement Index (D_i) ดังนี้

$$C_i^* = [\text{Max}(A_i) - A_i] / [\text{Max}(A_i) - \text{Min}(A_i)] \qquad (2\text{-}11)$$

เมื่อ $A_i = \sum_{j=1}^{J} U_{ij} P_{.j}$ เป็นดัชนีความสอดคล้องระหว่างการตอบของแต่ละคนกับความยากที่กำหนดโดยกลุ่ม

Max (A_i) และ Min (A_i) เป็นค่ามากที่สุดและน้อยที่สุดของ A_i สำหรับ $n_{i.}$ ที่กำหนด

นอกจาก A_i จะสัมพันธ์กับ C_i^* แล้วยังสัมพันธ์อย่างสูงกับคะแนนรวมอีกด้วย ส่วน Disagreement Index สัมพันธ์กับ C_i^* ดังนี้

$$C_i^* = D_i / \text{Max}(D_i) \qquad (2\text{-}12)$$

D_i หมายถึง Disagreement Index

$D_i = \text{Max}(A_i) - A_i$

$\text{Max}(D_i) = \text{Max}(A_i) - \text{Min}(A_i)$

2.2 ดัชนีที่ใช้ทฤษฎีการตอบสนองข้อสอบ

ดัชนีในประเภทนี้ได้แก่ Squared Standardized Residual ที่พัฒนาขึ้นโดย Wright และดัชนีความเหมาะสม (Appropriateness Indicies) ที่เสนอโดย Levine และคณะ คือ ดัชนีในกลุ่ม l_o (The l_o class of appropriateness indicies) กับดัชนีความเหมาะสมที่ใช้ Gaussian Model (Hulin and others 1983 : 121 - 122) ดัชนีความเหมาะสมนี้เป็นการวัดความสอดคล้อง (Goodness of Fit) ของโมเดลการวัด (Psychometric Model) กับรูปแบบการตอบสนองข้อต่อข้อของผู้สอบแต่ละคน ดัชนีความเหมาะสมจะมีค่าสูงถ้าคำตอบของผู้สอบคนนั้นเหมือนกับของคนอื่น ๆ ที่มีความสามารถระดับเดียวกัน และจะมีค่าต่ำถ้าคำตอบแตกต่างไปจากคนอื่น ๆ ที่มีความสามารถระดับเดียวกัน ดังนั้น ดัชนีความเหมาะสมจึงเป็นดัชนีที่ชี้ว่า ผู้สอบคนนั้นทำแบบสอบนั้นได้เช่นเดียวกับผู้สอบคนอื่นๆ ที่มีความสามารถระดับเดียวกันหรือไม่

Squared Standardized Residual

Wright (1977 อ้างถึงใน Hulin and others 1983 :129) ได้เสนอว่าความเหมาะสมระหว่างโมเดลและการตอบสนองของผู้สอบนั้นสามารถประเมินได้โดยการตรวจสอบความแตกต่างระหว่างความน่าจะเป็นของการตอบถูกของโมเดลนั้นกับรูปแบบการตอบสนองถูกและผิดของผู้สอบ นั่นคือ Residual ของข้อ i

$$R_i = u_i - P_i(\Theta) \qquad (2 - 13)$$

เมื่อ u_i คือ 1 หรือ 0 และ $P_i(\Theta)$ คือความน่าจะเป็นของการตอบถูก

เมื่อรวม Residual จากข้อสอบทุกข้อจะได้การวัดความเหมาะสมของแบบสอบสำหรับแต่ละคน อย่างไรก็ตาม R_i ต้องทำให้เป็นมาตรฐาน(Standardized) เสียก่อน การทำให้เป็นมาตรฐานนี้ทำให้สามารถเปรียบเทียบ Residual ระหว่างข้อสอบได้ การรวม Residual ที่ยังไม่ได้ทำให้เป็นมาตรฐานเพื่อจะให้ได้ดัชนีความเหมาะสมนั้นก็เหมือนกับการรวมระดับคะแนนเฉลี่ย (Grade Point Average) ในระบบ 4 แต้ม กับคะแนนรวม GRE เพื่อให้ได้ผลการวัดความสามารถทางวิชาการ ความแปรปรวนของระดับคะแนนเฉลี่ยจะน้อยกว่าความแปรปรวนของคะแนนรวม GRE มาก ด้วยเหตุนี้คะแนนทั้งสองอย่างจึงควรทำให้เป็นมาตรฐานก่อนที่จะนำมารวมกัน ในทำนองเดียวกัน Residual R_i ก็ควรทำให้เป็นมาตรฐานก่อนที่จะนำมารวมกัน

ในการทำ R_i ให้เป็นมาตรฐานต้องทราบค่าเฉลี่ยและความแปรปรวนเสียก่อน Expectation ของ R_i จะมีค่าเป็น 0 โดยประมาณ เพราะว่า

$$E(u_i) = P_i(\Theta) = P_i(\hat{\theta}) \qquad (2 - 14)$$

ความแปรปรวนของ u_i ได้จากสูตรความแปรปรวนของไบโนเมียล ดังนี้

$$Var(u_i) = P_i(\theta)[1 - P_i(\theta)] = P_i(\hat{\theta})[1 - P_i(\hat{\theta})] \qquad (2 - 15)$$

ดังนั้น Standardized Residual สำหรับข้อ i คือ

$$SR_i = [u_i - P_i(\hat{\theta})] / [Var(u_i)]^{1/2} \qquad (2 - 16)$$

และ Wright's Squared Standardized Residual คือ

$$W = \Sigma_i SR_i^{\ 2} \qquad (2 - 17)$$

Wright กล่าวว่า W มีการแจกแจงแบบไคสแควร์ ที่มี Degree of Freedom $(n - 1)(N - 1)/N$ ซึ่ง N หมายถึง จำนวนผู้สอบในการประมาณค่าพารามิเตอร์ของข้อสอบ
n หมายถึง จำนวนข้อสอบในแบบสอบ

สำหรับโมเดลโลจิสติกที่มีพารามิเตอร์ตัวเดียว (One - Parameter Logistic Model)

$$W = \Sigma_i \{u_i \exp[- D (\hat{\theta} - b_i)] + (1 - u_i) \exp[D (\hat{\theta} - b_i)]\} \qquad (2 - 18)$$

เนื่องจาก W ที่มีค่ามากสัมพันธ์กับเวคเตอร์การตอบสนองที่ผิดปกติ ดังนั้นจึงใช้ W เป็นดัชนีความเหมาะสมของรูปแบบการตอบสนอง

ดัชนีความเหมาะสมที่ใช้ Gaussian Model

Levine และ Rubin (1979 อ้างถึงใน Hulin and others 1983 : 127) เรียก IRT Model ที่ถือข้อตกลงคุณลักษณะแฝงมีมิติเดียวว่า Standard Models และได้พัฒนาโมเดลใหม่ขึ้นมาอีกโมเดลหนึ่งซึ่งมีการสุ่มตัวอย่างระดับความสามารถ θ_1 ขึ้นมาใหม่สำหรับข้อสอบข้อที่ i ตัวอย่างเช่น ในแบบสอบที่มีข้อสอบ 60 ข้อ โมเดลของ Levine และ Rubin ยอมให้มีค่าของ θ ถึง 60 ค่า θ เหล่านี้ถือว่าสุ่มขึ้นมาอย่างอิสระจากการแจกแจงปกติซึ่งมีค่าเฉลี่ย θ_o และความแปรปรวน σ_g^2 Levine และ Rubin เรียกโมเดลนี้ว่า **Gaussian Model**

ในกรณีผู้สอบที่มีความสามารถต่ำซึ่งลอกคำตอบจากเพื่อนข้าง ๆ ที่มีความสามารถสูงกว่านั้นในกระดาษคำตอบของเขาจะประกอบด้วยการตอบสนองข้อสอบบางข้อที่สะท้อนถึงความสามารถจริง ๆ ของเขาและบางส่วนก็สะท้อนถึงความสามารถของเพื่อนของเขา ในกรณีเช่นนี้ โมเดลการตอบสนองข้อสอบที่ยอมให้มีการแปรเปลี่ยนระดับความสามารถน่าจะเหมาะกับรูปแบบการตอบสนองที่สังเกตได้นั้นมากกว่า Standard Model ที่มีความสามารถระดับเดียวตลอดทั้งแบบสอบ

วิธีการประมาณค่าพารามิเตอร์ของผู้สอบด้วย Gaussian Model จะเริ่มด้วยการประมาณโค้งลักษณะของข้อสอบด้วย Standard Model แล้วเขียนไลลิฮูดของเวคเตอร์ของการตอบสนองข้อสอบในรูปฟังก์ชันของ θ_o, σ_g^2 และ θ_1 ตัว θ_1 นั้นสามารถอินทิเกรทจากสมการไลลิฮูด

$$l_n = \log \text{Prob}(U \mid \theta_o, \sigma_g^2) \qquad (2 - 19)$$

ในที่สุดจะสามารถประมาณค่า θ_o และ σ_g^2 ด้วยวิธีการทางตัวเลข

Levine และ Rubin (1979 อ้างถึงใน Hulin and others 1983 : 128) ได้เสนอดัชนีความเหมาะสม 2 ตัวที่ใช้ Gaussian Model ดัชนีตัวแรก คือ ล็อกการิทึมของ Likelihood Ratio ที่เปรียบเทียบ Gaussian Model กับ Standard Model ดังนี้

$$LR = l_n - l_o \qquad (2 - 20)$$

สำหรับโค้งลักษณะของข้อสอบที่กำหนด (เช่น Three - Parameter Logistic) Gaussian Model จะเหมาะสมกับเวคเตอร์การตอบสนองอย่างน้อยที่สุดก็ดีเท่ากับ Standard Model ทั้งนี้เพราะว่า Standard Model เป็นกรณีเฉพาะของ Gaussian Model เมื่อ $\sigma_g^2 = 0$ จากการที่โมเดลซึ่งความสามารถแปรเปลี่ยนได้นี้เหมาะสมกว่าโมเดลที่ความสามารถคงที่ l_n จะมากกว่า l_o และจะได้ค่า LR มาก ค่าของ l_n ควรจะมากกว่า l_o สำหรับรูปแบบการตอบสนองที่ผิดปกติ

ดัชนีตัวที่สองที่ได้จาก Gaussian Model คือ การประมาณ $\hat{\sigma}_g^2$ ความแปรปรวนของการแจกแจง θ_i ดัชนีนี้ควรจะมีค่าใกล้ 0 เมื่อ Major Determinant ของเวคเตอร์การตอบสนองเป็นความสามารถเดียว (Single Ability) ถ้าองค์ประกอบอื่น ๆ (เช่น การทุจริต ความเข้าใจผิดในคำสั่ง เป็นต้น) มีอิทธิพลต่อการตอบสนองของผู้สอบแล้ว σ_g^2 จะมีค่ามากกว่า 0

ดัชนีในกลุ่ม l_o

ดัชนีในกลุ่ม l_o นี้ได้แก่ ดัชนี l_o, l_g และ l_z หรือ Z_3 ซึ่ง Levine และคณะได้พัฒนาดัชนี l_o ขึ้นมาก่อนแล้วปรับปรุงมาเป็น l_g และ l_z ตามลำดับ (Hulin and others 1983 : 122)

หลักการของดัชนี l_o ได้แนวคิดมาจากการประมาณค่าไลลิฮูดสูงสุด (Maximum Likelihood Estimation) ซึ่งมีจุดมุ่งหมายเพื่อเลือกค่าประมาณพารามิเตอร์ที่จะทำให้ข้อมูลชุดนั้นมีความเหมาะสมมากที่สุด Levine และคณะเกิดความคิดว่า ถ้าไม่มีค่า θ ที่ทำให้ไลลิฮูดของการตอบสนองของผู้สอบคนหนึ่งมีค่ามากได้แล้ว ทฤษฎีการตอบสนองข้อสอบที่ใช้ในการประมาณค่าพารามิเตอร์ของข้อสอบและประมาณค่าความสามารถของผู้สอบก็จะไม่เหมาะสมสำหรับผู้สอบคนนี้ คะแนนสอบที่ได้ก็จะไม่เป็นตัวแทนของการวัดความสามารถ คะแนนสอบที่ไม่เหมาะสมจะให้เวคเตอร์การตอบสนองซึ่งไม่มีค่า θ ใดที่จะให้ค่าฟังก์ชันไลลิฮูดที่มีค่ามากได้ ดัชนี l_o สามารถตรวจสอบเวคเตอร์การตอบสนองเช่นนี้ได้จึงทำหน้าที่เป็นดัชนีความเหมาะสมของรูปแบบการตอบสนองตามที่ต้องการ

ดัชนี l_o ตัวแรกที่ศึกษาโดย Levine กับ Rubin และ Levine กับ Drasgow (1982 อ้างถึงใน Hulin and others 1983 : 122) คือ

$$l_o = \log \max_{\theta} \text{Prob}(U \mid \theta) = \log \text{Prob}(U \mid \hat{\theta}) \qquad (2 - 21)$$

เมื่อ U คือเวคเตอร์ของการตอบสนองข้อสอบ

Prob(U I θ) คือ Likelihood Function

$\hat{\theta}$ คือ Maximum Likelihood Estimate ของ θ

โปรดสังเกตว่า l_o เป็นเพียง Natural Logarithm ของ Maximum Likelihood Function เท่านั้น

ตารางที่ 6 เวคเตอร์การตอบสนองและความน่าจะเป็นของการตอบสนอง สำหรับผู้สอบสมมุติ 2 คน

ข้อสอบ	$P_i(\hat{\theta})$	$1 - P_i(\hat{\theta})$	ผู้สอบ	
			1	2
1	.9	.1	1	0
2	.7	.3	1	0
3	.5	.5	1	1
4	.3	.7	0	1
5	.1	.9	0	1

(Hulin and others 1983 : 123)

ตารางที่ 6 เป็นข้อมูลสมมุติของผู้สอบ 2 คน พร้อมทั้งความน่าจะเป็นในการตอบถูกและตอบผิด สมมุติว่าข้อมูลของผู้สอบ 2 คนนี้มีค่าประมาณไลลิฮูดสูงสุด (Maximum Likelihood Estimate) ของความสามารถเท่า ๆ กัน

ฟังก์ชันไลลิฮูดสูงสุดของผู้สอบคนแรกคือ

Prob(U I θ) = (.9)(.7)(.5)(.7)(.9) = .198

และ l_o = log (.198) = - 1.62

ฟังก์ชันไลลิฮูดสูงสุดของผู้สอบคนที่สองคือ

Prob(U I θ) = (.1)(.3)(.5)(.3)(.1) = .00045

และ l_o = - 7.71

จะเห็นว่า l_o ประสบความสำเร็จในการตรวจสอบการตอบสนองที่ผิดปกติของผู้สอบคนที่สอง ยิ่งกว่านั้น l_o จะมีค่าเป็นเลขลบมาก ๆ สำหรับเวคเตอร์การตอบสนองที่ตอบข้อยาก ๆ ได้ถูกต้องหลายข้อและตอบข้อง่าย ๆ ผิดหลายข้อ ที่เป็นเช่นนี้เพราะว่าการตอบข้อยาก ๆ ได้หลายข้อแสดงว่า θ ไม่ควรจะต่ำ ขณะเดียวกันการตอบข้อง่าย ๆ ผิดหลายข้อก็แสดงว่า θ ไม่ควรจะสูง

แม้ว่า l_o จะตรวจสอบเวคเตอร์การตอบสนองได้ค่อนข้างดีแต่ก็ยังไม่แน่นอนสำหรับกรณีที่มีการเว้นไม่ตอบในบางข้อ ในกรณีดังกล่าว Levine และ Drasgow (1982) คำนวณ l_o จากข้อที่ตอบโดยไม่สนใจข้อที่เว้นไป สำหรับการสอบที่ผู้สอบเว้นจำนวนข้อต่าง ๆกัน การเปรียบเทียบค่า l_o อาจทำให้เกิดความเข้าใจผิดได้ ทั้งนี้เพราะว่า ผู้สอบที่ตอบมากข้อกว่าอีกคนหนึ่งมักจะมีเวคเตอร์การตอบสนองที่มีค่า l_o น้อยกว่าและมักปรากฏว่ามีความผิดปกติมากกว่า เพื่อให้เห็นชัดเจนก็จะเขียน Logarithm ของสมการ Likelihood เสียใหม่ ดังนี้

$$l_o = \log \text{Prob}(U \mid \hat{\theta}) = \Sigma_i \log \text{Prob}(u_i \mid \hat{\theta}) \qquad (2 - 22)$$

เนื่องจากล็อก (Logarithm) ของจำนวนที่อยู่ระหว่าง 0 และ 1 เป็นลบ ฉะนั้น $\log \text{Prob}(u_i \mid \hat{\theta})$ จึงเป็นลบสำหรับทุกข้อที่ตอบ ดังนั้น ผู้ที่ตอบมากข้อกว่าจึงได้ค่า l_o ที่น้อยกว่า

เพื่อปรับปรุงดัชนี l_o Drasgow (1982 อ้างถึงใน Hulin and others 1983 : 124) ได้คำนวณ l_o สำหรับทุกข้อที่ตอบแล้วคำนวณ Geometric Mean Likelihood ดังนี้

$$l_z = \exp (l_o / n) \qquad (2 - 23)$$

เมื่อ n คือจำนวนข้อที่ผู้สอบคนนั้นตอบ

ถ้า l_z มีค่าน้อยแสดงว่าเวคเตอร์การตอบสนองนั้นผิดปกติ

ตารางที่ 7 เวคเตอร์การตอบสนองและความน่าจะเป็นของการตอบสนอง สำหรับผู้สอบสมมุติ 3 คน

ข้อสอบ	$P_1(\hat{\theta})$	$1 - P_1(\hat{\theta})$	ผู้สอบ		
			1	2	3
1	.90	.10	1	1	*
2	.70	.30	1	1	0
3	.50	.50	0	0	*
4	.30	.70	1	1	1
5	.10	.90	0	0	*
6	.61	.39	*	1	0
7	.39	.61	*	0	*
	จำนวนข้อที่ตอบ		5	7	3

* หมายถึงข้อที่ไม่ตอบ

(Hulin and others 1983 : 124)

สมมุติว่าเวคเตอร์การตอบสนองของผู้สอบทั้งสามคนในตารางที่ 7 มีค่าประมาณไลลิฮูดสูงสุดของ θ เท่ากัน

สำหรับผู้สอบคนแรก

$$l_o = \log [(.9)(.7)(.5)(.3)(.9)] = -2.46$$

สำหรับผู้สอบคนที่สอง

$$l_o = \log [(.9)(.7)(.5)(.3)(.9)(.61)(.61)] = -3.45$$

ค่า l_o ที่ได้ชี้ว่า รูปแบบการตอบสนองของผู้สอบคนที่สองมีความผิดปกติมากกว่ารูปแบบการตอบสนองของผู้สอบคนแรก ผลที่ได้นี้นำไปสู่ความเข้าใจผิดเพราะว่ารูปแบบ

การตอบสนองของผู้สอบสองคนนี้เหมือนกันในข้อที่ผู้สอบคนแรกตอบ และผู้สอบคนที่สองได้ตอบสนองในแบบที่น่าจะเป็นไปได้ในข้อ 6 และข้อ 7

ส่วนผู้สอบคนที่สาม

l_o = log [(.3)(.3)(.39)] = - 3.34 ซึ่งชี้ว่าเวคเตอร์การตอบสนองของผู้สอบคนที่สองมีความผิดปกติมากกว่าผู้สอบคนที่สาม อันเป็นการนำไปสู่ความเข้าใจผิดอีกเช่นกัน

Geometric Mean Likelihood สำหรับผู้สอบคนแรก คือ

l_g = exp (- 2.46 / 5) = .61

และ l_g ของผู้สอบคนที่สองและคนที่สามคือ .61 และ .33 ตามลำดับ ค่าของ l_g เป็นดัชนีความเหมาะสม ดังที่ปรากฏคือ ผู้สอบคนแรกและคนที่สองมีค่าดัชนีเท่า ๆ กัน ขณะที่ผู้สอบคนที่สามมีค่าดัชนีต่ำกว่า ผลของอัตราการเว้นที่แตกต่างกันจะถูกขจัดออกไปโดยการแปลง l_o เป็น l_g

แต่ว่าทั้ง l_o และ l_g ก็นำไปสู่ความเข้าใจผิดได้ถ้าเปรียบเทียบผู้สอบที่มีระดับความสามารถแตกต่างกัน Drasgow, Levine และ Williams (1982 อ้างถึงใน Hulin and others 1983 : 125) ได้แสดงว่า Expected Value ของ l_o และ l_g แปรผันเป็นฟังก์ชันของ θ

ถ้าให้ $\hat{P}_{ik}$ = Prob(u_i = k I $\hat{\theta}$), k = 0, 1 เป็นความน่าจะเป็นของการตอบผิด (k = 0) และการตอบถูก (k = 1) สำหรับผู้สอบที่มีความสามารถที่ประมาณได้เป็น $\hat{\theta}$ แล้ว Drasgow และคณะได้แสดงว่า

$$E(l_o) = \Sigma_i [\hat{P}_{i1} \log \hat{P}_{i1} + \hat{P}_{io} \log \hat{P}_{io}] \qquad (2 - 24)$$

เพราะว่า $\hat{P}_{ik}$ เป็นฟังก์ชันของ $\hat{\theta}$ จึงเป็นที่ชัดเจนว่า E (l_o) ขึ้นอยู่กับระดับความสามารถ

Drasgow และคณะได้พัฒนาสูตรหาค่าโดยประมาณของความแปรปรวนของ l_o สำหรับผู้สอบที่มีความสามารถ θ เพื่อใช้ร่วมกับสมการ (2 - 24) เพื่อหาค่าของ Standardize l_o ดังนี้

$$\text{Var}(l_o) = \Sigma_i \hat{P}_{i1} \hat{P}_{io} [\log (\hat{P}_{i1} / \hat{P}_{io})]^2 \qquad (2 - 25)$$

และจะได้ l_o ในรูปของ Standardized ดังนี้

$$l_z = [l_o - E(l_o)] / [Var(l_o)]^{1/2} \qquad (2 - 26)$$

เราสามารถคำนวณ l_o ได้จากข้อที่ตอบ ส่วน l_z ทำให้สามารถเปรียบเทียบค่าดัชนี(Index Scores) ของผู้สอบที่ตอบข้อต่างกันได้ นอกจากนี้ l_z ยังสามารถขจัดความแตกต่างที่ขึ้นอยู่กับ θ ดังนั้นจึงสามารถเปรียบเทียบค่าดัชนีของผู้สอบที่มีความสามารถแตกต่างกันได้

การศึกษาวิจัยเกี่ยวกับดัชนีความเหมาะสมของรูปแบบการตอบสนองข้อสอบ

Tatsuoka และ Tatsuoka (1983 : 226 - 227) ได้อภิปรายถึงจุดอ่อนของ Norm Conformity Index และ Caution Index ว่าการคำนวณดัชนี 2 ตัวนี้ขึ้นอยู่กับการเรียงลำดับข้อตามความยากใน Norm Group เมื่อ Norm Group เปลี่ยนไปลำดับข้อตามความยากจะเปลี่ยนตามซึ่งมีผลกระทบต่อค่าดัชนีทั้งสองนี้สำหรับรูปแบบการตอบสนองเดียวกันของผู้สอบคนนั้น ดัชนีดังกล่าวจึงไม่เหมาะที่จะใช้วินิจฉัยความรุนแรงของความผิดพลาดของผู้สอบ Tatsuoka และ Tatsuoka จึงได้เสนอดัชนีที่เป็นอิสระจากกลุ่ม คือ Individual Consistency Index (ICI) ค่าดัชนีนี้จะขึ้นอยู่กับความยากของงานที่กำหนดโดยความสามารถของผู้สอบเอง แต่ ICI ก็มีข้อจำกัดอยู่ที่ต้องใช้แบบสอบคู่ขนานตั้งแต่ 2 ชุดขึ้นไปจึงจะสามารถคำนวณหาค่าดัชนีนี้ได้ซึ่งเป็นไปได้ยากในทางปฏิบัติ

ในปีเดียวกันนี้ Rudner (1983 อ้างถึงใน Tatsuoka 1984 : 96) ได้ทำการศึกษาโดยใช้ข้อมูลที่สร้างขึ้น(Monte Carlo Data) ผลการศึกษาพบว่า ดัชนีที่ใช้ IRT Model แสดงอัตราการตรวจจับ(Detection Rate) รูปแบบการตอบสนองที่ผิดปกติได้ดีกว่าดัชนีที่ใช้สถิติพื้นฐาน (Observed Standard Statistics) เล็กน้อย ในปีเดียวกันนั้น Tatsuoka และ Linn (1983 : 82) ได้นัมมาดัชนีขึ้นมาอีกโดยเชื่อมโยง Sato's Caution Index ซึ่งนิยามโดย S - P Curve Theory เข้ากับ IRT ได้เป็น Extended Caution Index หลายตัว และยังได้แสดงความสัมพันธ์ระหว่าง S - P Curve Theory กับ Test Response Curves และ Group Response Curves ที่นัมมาจาก IRT จากนั้น Tatsuoka (1984) ได้ศึกษาค้นคว้าต่อถึงคุณสมบัติเชิงสถิติของ Extended Caution Index รวมทั้งการทำดัชนีเหล่านี้ให้อยู่ในรูปมาตรฐานและได้ศึกษาถึงความสัมพันธ์กับ Guttman Scales ตลอดจนอภิปรายถึง Item และ Person Response Curves ด้วย

นอกจากการศึกษาเปรียบเทียบดัชนีต่าง ๆ โดย Rudner (1983) แล้ว Harnisch และ Tatsuoka (1983 อ้างถึงใน Drasgow and others 1987 : 60) ได้หาสหสัมพันธ์ระหว่างดัชนี 14 ตัวเพื่อดูว่าคู่ใดสัมพันธ์กันมากคู่ใดสัมพันธ์กันน้อย แต่การศึกษาครั้งนี้ก็ไม่ได้แสดงว่าดัชนีตัวใดดีพอที่จะนำไปใช้ได้ Drasgow และคณะ(1987)จึงได้ใช้เกณฑ์ 2 ประการในการประเมินประสิทธิภาพของดัชนีความเหมาะสมซึ่งได้แก่ ความเป็นมาตรฐาน (Standardization) และ อำนาจเชิงสัมพัทธ์ (Relative Power)

ความเป็นมาตรฐาน (Standardization) หมายถึง การที่ Conditional Distributions ของดัชนีไม่แปรผันไปตามระดับความสามารถ ดัชนีที่มีความเป็นมาตรฐานสูง (Well - Standardized Indicies) มีลักษณะเด่นตรงที่มีอัตราการตรวจจับได้ (Rate of Detection) สูง และค่าดัชนี (Index Scores) ของแต่ละคนที่ความสามารถแตกต่างกันสามารถเปรียบเทียบกันได้โดยตรงซึ่งตรงกันข้ามกับค่าดัชนี (Index Scores) ของดัชนีที่มีความเป็นมาตรฐานต่ำ (Poorly Standardized Indicies) ที่สามารถตีความได้เฉพาะในการแจกแจงจำกัด (Conditional Distribution) ของพวกเดียวกันเท่านั้น (Drasgow and others 1987 : 60)

ส่วนอำนาจเชิงสัมพัทธ์ (Relative Power) นั้น เมื่อกำหนดอัตราการจัดพวกผิดโดยจัดรูปแบบการตอบสนองปกติว่าเป็นพวกผิดปกติ (Type I Error Rate) ให้ ดัชนีตัวใดมีอัตราการจัดพวกรูปแบบการตอบสนองผิดปกติว่าผิดปกติได้ถูกต้องสูงที่สุดถือว่ามีอำนาจ (Power) สูงสุด หากดัชนีที่มีความเป็นมาตรฐานสูง (Well - Standardized Index) มีอำนาจ (Power) ที่ยอมรับได้แล้วก็สามารถนำไปใช้ในทางปฏิบัติได้

Drasgow และคณะ (1987 : 59) ได้เปรียบเทียบประสิทธิภาพของดัชนีความเหมาะสม 9 ตัวโดยพิจารณาในเชิงสัมบูรณ์ (Absolute Sense) ด้วยการเปรียบเทียบดัชนีเหล่านี้กับดัชนีที่เหมาะสม (Optimal Index) ซึ่งดัชนีที่เหมาะสมนี้คือดัชนีที่มีอำนาจสูงสุดสำหรับรูปแบบการตอบสนองผิดปกตินั้น ๆ ที่สามารถคำนวณได้จากการตอบสนองข้อสอบ ผลปรากฏว่า มีดัชนี 3 ตัวที่ให้อัตราการตรวจจับ (Rate of Detection) รูปแบบการตอบสนองของความสามารถต่ำมาก ๆ ที่ปรับเป็นการตอบถูก (Simulate Cheating) ได้ใกล้เคียงดัชนีที่เหมาะสมพอ ๆ กับรูปแบบการตอบสนองของความสามารถสูงมาก ๆ ที่ปรับเป็นการตอบผิดทำให้คะแนนต่ำกว่าที่ควรจะเป็น ดัชนีที่เหมาะสม ดังกล่าวมี อัตราการตรวจจับได้ (Detection Rates) ตั้งแต่ 50 % ถึง 200 % สูงกว่าดัชนีใด ๆ เมื่อเวคเตอร์การตอบสนอง (Response Vectors) ของความสามารถเฉลี่ยถูกจัดกระทำให้สูงกว่าที่ควรจะเป็นและต่ำกว่าที่ควรจะเป็น

ดัชนีตัวหนึ่งในสามตัวจากผลการศึกษาของ Drasgow และคณะ ดังกล่าวแล้วนั้นก็คือ Standardized l_o คือ l_z หรือ Z_3 ซึ่งเป็นดัชนีตัวหนึ่งที่มีความเป็นมาตรฐาน (Well - Standarded) และมีอำนาจ (Power) สูง จึงน่าจะนำไปทดลองใช้ในการตรวจสอบความเหมาะสมของรูปแบบการตอบสนองข้อสอบในการทดสอบทางการศึกษา

รายงานการศึกษาเกี่ยวกับการวัดความเหมาะสมของคะแนนสอบที่ผู้วิจัยได้ศึกษาค้นคว้ามานั้นเกือบทั้งหมดเป็นการศึกษาค้นคว้าเพื่อนำเสนอดัชนีและศึกษาประสิทธิภาพของดัชนีด้วยข้อมูลที่สร้างขึ้นจากเครื่องคอมพิวเตอร์ รวมทั้งได้มีการเสนอแนะการนำไปใช้ในแง่ต่าง ๆ ทางการศึกษา เช่น ในการศึกษาความแตกต่างระหว่างกลุ่ม (Group Differences) ความแตกต่างของการจัดการเรียนการสอน (Schooling Differences) การวิเคราะห์ข้อสอบ (Distractor Analysis) การจัดกลุ่มนักเรียน (Classification of Students) การระบุข้อสอบที่ผิดปกติ (Identification of Unusual Items) เป็นต้น (Harnisch 1983 : 192 - 203)

ส่วนการศึกษากับข้อมูลจริงนั้น สุนันท์ ศลโกสุม (2530) ได้นำไปใช้ศึกษาเปรียบเทียบผลการวิเคราะห์ผลการสอบด้วยวิธีต่าง ๆ และศึกษาความสัมพันธ์ระหว่างวิธีการเหล่านั้น ซึ่งได้แก่ วิธีวิเคราะห์ด้วยทฤษฎีการทดสอบแบบดั้งเดิม ดัชนีชี้นำของซาโต้ และทฤษฎีการตอบสนองข้อสอบซึ่งวิเคราะห์ด้วยค่าความสามารถ และดัชนีชี้นำของทาทซูโอกะ ข้อมูลในการศึกษาครั้งนี้ได้มาจากการตรวจสอบคุณภาพการศึกษา วิชาภาษาไทย ชั้นมัธยมศึกษาปีที่ 6 ปีการศึกษา 2528 ของกรมวิชาการ กระทรวงศึกษาธิการ สุนันท์ ศลโกสุม (2530 : 126) ได้สรุปว่า การวิเคราะห์ผลการสอบด้วยดัชนีชี้นำของทาทซูโอกะได้ผลไม่สอดคล้องกับผลการวิเคราะห์ด้วยดัชนีชี้นำของซาโต้ แสดงว่าเกิดความแตกต่างกันในเรื่องการเรียงลำดับข้อสอบตามความยากกับการแทนค่า P_{ij} ลงในตารางเมตริกซ์

ในปีต่อมา บุญมี พันธุ์ไทย (2531) ได้นำ Squared Standardized Residual ของ Wright ไปใช้ในการจำแนกนักเรียนที่มีรูปแบบการตอบสนองข้อสอบซึ่งสอดคล้องและไม่สอดคล้องกับความสามารถ พบว่า กลุ่มนักเรียนที่รับรู้ผลกระทบของการสอบเพื่อใช้เป็นส่วนหนึ่งของการตัดสินผลการเรียน (N = 263) มีรูปแบบของการตอบสนองข้อสอบที่ไม่สอดคล้องกับความสามารถจำนวนมากกว่ากลุ่มนักเรียนที่รับรู้ผลกระทบของการสอบเพื่อทำวิจัย (N = 266) ที่มีรูปแบบการตอบสนองข้อสอบไม่สอดคล้องกับความสามารถ อีกทั้งในจำนวนนักเรียนที่มีรูปแบบการตอบสนองข้อสอบไม่สอดคล้องกับความสามารถของกลุ่มนักเรียนที่รับรู้ผลกระทบของการสอบเพื่อใช้เป็นส่วนหนึ่งของการตัดสินผลการเรียนมีนักเรียนที่มีความสามารถสูงอยู่เป็นจำนวนมาก นอกจากนี้ผู้วิจัยยังได้ให้ข้อสังเกตว่า นักเรียนที่มีความสามารถสูงแต่ทำข้อสอบที่ง่ายมาก ๆ ผิดเพียงหนึ่ง

หรือสองข้อก็จะทำให้ดัชนีมีค่าสูงหรือนักเรียนที่มีความสามารถต่ำมาก ๆ แต่ทำข้อสอบที่ยากมาก ๆ ถูกหนึ่งหรือสองข้อค่าดัชนีก็จะสูงเหมือนกัน

อีกรายหนึ่ง คือ Tomsic (1986) ได้ศึกษา Extended Caution Index Two (ECI2) และ Extended Caution Index Four (ECI4) ในรูปที่ทำให้เป็นมาตรฐานและไม่ได้ทำให้เป็นมาตรฐาน โดยใช้ข้อมูลผลการสอบ Comprehensive Test of Basic Skills (CTBS) ของนักเรียนเกรด 3 และเกรด 7 แต่ผลสรุปที่ได้ยังไม่ชัดเจน

นอกจากงานวิจัยสามชิ้นนี้แล้วยังไม่พบงานวิจัยเกี่ยวกับดัชนีความเหมาะสมของรูปแบบการตอบสนองข้อสอบที่ใช้ข้อมูลจริงล้วน ๆ อีกโดยเฉพาะอย่างยิ่งในแง่ที่นำมาศึกษาประสิทธิภาพของแบบสอบ

ตอนที่สาม

ตอนที่สามนี้เป็นเรื่องเกี่ยวกับทฤษฎีการตอบสนองข้อสอบซึ่งกล่าวถึงข้อตกลงเบื้องต้นของทฤษฎีการตอบสนองข้อสอบ โมเดลของการตอบสนองข้อสอบ ฟังก์ชันสารสนเทศของข้อสอบและแบบสอบ ประสิทธิภาพเชิงสัมพัทธ์ของข้อสอบและแบบสอบ

ทฤษฎีการตอบสนองข้อสอบ

ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory) มีชื่ออื่น ๆ อีก เช่น ทฤษฎีความสามารถแฝง (Latent Trait Theory) ทฤษฎีโค้งลักษณะของข้อสอบ (Item Characteristic Curve Theory) ทฤษฎีนี้มีความเชื่อหลักอยู่ 2 ประการ (Hambleton and Swaminathan 1985 : 13) คือ **ประการแรก** ผลการทำแบบสอบของผู้สอบสามารถทำนายหรืออธิบายได้ด้วยองค์ประกอบ (Factors) ชุดหนึ่งที่เรียกว่าคุณลักษณะ (Traits) คุณลักษณะแฝง (Latent Traits) หรือความสามารถ (Abilities) **ประการที่สอง** ความสัมพันธ์ระหว่างผลการทำข้อสอบกับคุณลักษณะที่เชื่อว่ามีอิทธิพลต่อผลการทำข้อสอบนั้นสามารถบรรยายได้ด้วย Monotonically Increasing Function ที่เรียกว่า Item Characteristic Function ฟังก์ชันนี้ระบุว่า ผู้สอบที่มีคะแนนคุณลักษณะนั้น ๆ สูงกว่าย่อมมีความน่าจะเป็นที่คาดหวังสำหรับการตอบข้อสอบได้ถูกต้องมากกว่าผู้สอบที่มีคะแนนคุณลักษณะนั้น ๆ ต่ำกว่า ในทางปฏิบัติ ผู้ใช้ทฤษฎีการตอบสนองข้อสอบจะถือข้อตกลงว่ามีองค์ประกอบหรือความสามารถเด่น ๆ เพียงอย่างเดียวที่อธิบายผลการทำข้อสอบนั้น ในโมเดลคุณลักษณะเดียวหรือมิติเดียวนี้เรียก Item Characteristic Function ว่า Item Characteristic Curve (ICC) ซึ่งบอกถึงความน่าจะเป็นที่ผู้สอบจะตอบข้อสอบได้ถูกต้อง สำหรับผู้สอบที่มีความสามารถระดับต่าง ๆ บนสเกลของความสามารถ

เป้าหมายของทฤษฎีการตอบสนองข้อสอบก็คือการให้การประมาณค่าสถิติของข้อสอบและความสามารถของผู้สอบที่ไม่แปรเปลี่ยน (Invariant) ซึ่งจะเป็นไปได้ก็ต่อเมื่อโมเดลที่เลือกนั้นเหมาะสมกับข้อมูล ในกระบวนการของการประมาณค่านั้นข้อสอบและผู้สอบจะถูกวางบนสเกลความสามารถในลักษณะที่จะมีความสัมพันธ์กันมากที่สุดเท่าที่จะเป็นไปได้ระหว่างความน่าจะเป็นที่คาดหวังกับความน่าจะเป็นจริง ๆ ของผลการทำข้อสอบสำหรับผู้สอบที่อยู่ในตำแหน่งความสามารถแต่ละระดับ การประมาณค่าพารามิเตอร์ของข้อสอบและความสามารถของผู้สอบนั้นกระทำซ้ำ ๆ จนกระทั่งการทำนายจากความสามารถและพารามิเตอร์ของข้อสอบที่ประมาณได้กับการทำนายจากข้อมูลการทดสอบจริง ๆ นั้นสอดคล้องกันมากที่สุด

1. ข้อตกลงเบื้องต้นของทฤษฎีการตอบสนองข้อสอบ

ในทฤษฎีการตอบสนองข้อสอบมีข้อตกลงเบื้องต้นที่สำคัญอยู่ 3 ประการ คือ Dimentionality of Latent Space , Local Independence และ Item Characteristic Curve

1.1 Dimentionality of Latent Space

ทฤษฎีการตอบสนองข้อสอบเชื่อว่าพฤติกรรมการตอบแบบสอบสามารถอธิบายได้ด้วยความสามารถหรือคุณลักษณะแฝง (Latent Trait) หลายอย่าง คุณลักษณะแฝงแต่ละอย่างจะกำหนดมิติของ Latent Space ขึ้นมาหนึ่งมิติ ฉะนั้น Latent Space จะสมบูรณ์ถ้ามีการระบุถึงคุณลักษณะแฝงทุกอย่างที่สัมพันธ์กับคะแนนสอบของประชากรผู้สอบ แต่โมเดลการตอบสนองข้อสอบทั่วไปถือว่าพฤติกรรมการทำแบบสอบหรือผลการทำแบบสอบนั้นสามารถอธิบายได้ด้วยความสามารถหรือคุณลักษณะแฝงที่เด่น ๆ เพียงอย่างเดียว โมเดลที่ยึดถือคุณลักษณะแฝงอย่างเดียวนี้เรียกว่า Unidimentional Item Response Models ส่วนโมเดลที่เชื่อว่ามีความสามารถมากกว่าหนึ่งอย่างที่อธิบายการตอบสนองข้อสอบของผู้สอบคือ Multidimentional Item Response Model แม้โมเดลหลังจะมีความสมเหตุผลมากกว่าโมเดลแรกแต่ก็ยังมีข้อจำกัดในการที่จะนำไปประยุกต์ใช้ในขณะนี้ (Gruijter and Kamp 1984 : 160 ; Hambleton and Swaminathan 1985 : 17)

1.2 Local Independence

ข้อตกลงเบื้องต้นเกี่ยวกับ Local Independence กล่าวว่า การตอบสนองข้อสอบของผู้สอบคนหนึ่ง ๆ ต่อข้อสอบข้อต่าง ๆ ในแบบสอบฉบับหนึ่งนั้นจะเป็นอิสระต่อกันในเชิงสถิติ นั่นคือ ถ้าข้อตกลงเบื้องต้นข้อนี้เป็นจริงแล้วผลการทำข้อสอบข้อหนึ่งจะไม่มีผลให้การตอบข้อสอบข้ออื่น ๆ ในแบบสอบนั้นดีขึ้นหรือเลวลงแต่อย่างใด และความน่าจะเป็นของรูปแบบการตอบสนองต่อแบบสอบฉบับนั้นของผู้สอบแต่ละคนจะเท่ากับผลคูณของความน่าจะเป็นของการตอบสนองข้อสอบแต่ละข้อของผู้นั้น

1.3 โค้งลักษณะของข้อสอบ (Item Characteristic Curves)

โค้งลักษณะของข้อสอบ คือ ฟังก์ชันทางคณิตศาสตร์ที่แสดงความสัมพันธ์ระหว่างความน่าจะเป็นในการตอบข้อสอบข้อหนึ่งได้ถูกต้องกับระดับความสามารถที่วัดด้วยแบบสอบที่มีข้อสอบข้อนั้นอยู่ กล่าวอีกนัยหนึ่ง โค้งลักษณะของข้อสอบเป็นฟังก์ชันถดถอยที่ไม่ใช่เส้นตรง (Non Linear) ของคะแนนข้อสอบบนความสามารถหรือคุณลักษณะที่วัดด้วยแบบสอบนั้น

โค้งลักษณะของข้อสอบมีหลายรูปแบบขึ้นอยู่กับความเชื่อเกี่ยวกับความสัมพันธ์ระหว่างระดับความสามารถกับความน่าจะเป็นในการตอบถูก ทำให้มีโมเดลที่ใช้อธิบายความสัมพันธ์ดังกล่าวในทฤษฎีการตอบสนองข้อสอบหลายโมเดลด้วยกัน ดังจะกล่าวถึงต่อไป

2. โมเดลการตอบสนองข้อสอบ

โมเดลการตอบสนองข้อสอบทุกโมเดลที่ใช้อยู่ในขณะนี้มีข้อตกลงเบื้องต้นเกี่ยวกับ Dimentionality of Latent Space และ Local Independence เช่นเดียวกัน แต่มีข้อแตกต่างกันที่รูปแบบทางคณิตศาสตร์ที่ใช้อธิบายโค้งลักษณะของข้อสอบและลักษณะของข้อมูลหรือการตรวจให้คะแนน ถ้าจัดแบ่งโมเดลการตอบสนองข้อสอบตามลักษณะของข้อมูลหรือการตรวจให้คะแนนจะได้เป็น 3 กลุ่ม ดังนี้ (Hambleton and Swaminathan 1985 : 35)

2.1 โมเดลที่มีการตรวจให้คะแนนแบบถูกให้ 1 ผิดให้ 0 (Dichotomous) ได้แก่

2.1.1 Guttman Perfect Scale

2.1.2 Latent Distance Model

2.1.3 Latent Linear Model

2.1.4 One -, Two -, Three - Parameter Normal Ogive Model

2.1.5 One -, Two -, Three - Parameter Logistic Model

2.1.6 Four - Parameter Logistic Model

2.2 โมเดลที่มีการตรวจให้คะแนนแบบหลายประเภท (Multicategory Scoring) ได้แก่

2.2.1 Nominal Response Model

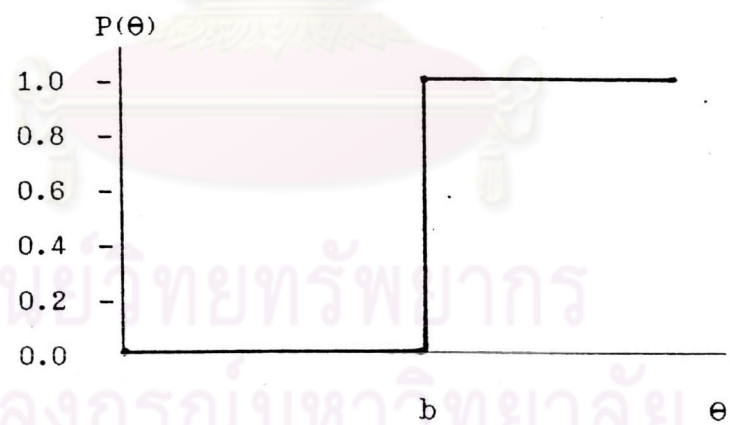
2.2.2 Graded Response Model

2.2.3 Partial Credit Model

2.3 โมเดลที่ข้อมูลเป็นแบบต่อเนื่อง (Continuous) ได้แก่ Continuous Response Model

Guttman Perfect Scale

Guttman Perfect Scale เป็นโมเดลทฤษฎีการตอบสนองข้อสอบในยุคแรก ๆ โมเดลนี้มีความเชื่อว่า ความน่าจะเป็นในการตอบข้อสอบข้อหนึ่งได้ถูกต้องนั้นสัมพันธ์กับความสามารถของผู้สอบในลักษณะของฟังก์ชันแบบขั้นบันได (Step Function) โค้งลักษณะของข้อสอบจะมีลักษณะดังภาพที่ 1



ภาพที่ 1 เส้นโค้งลักษณะของข้อสอบที่เป็น Guttman Perfect Scale

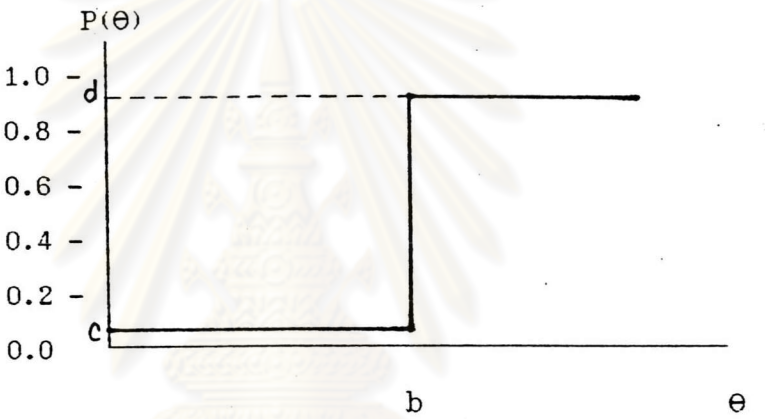
ภาพที่ 1 แสดงว่าผู้ที่มีความสามารถต่ำกว่า b ($\Theta < b$) ความน่าจะเป็นที่จะตอบข้อสอบข้อนี้ได้ถูกต้องจะเท่ากับ 0 ส่วนผู้ที่มีความสามารถเท่ากับหรือมากกว่า b ($\Theta \geq b$) ความน่าจะเป็นที่จะตอบข้อสอบข้อนี้ได้ถูกต้องจะเท่ากับ 1

Guttman Perfect Scale มีลักษณะเป็น Deterministic Model ซึ่งข้อมูลที่ได้จากการทดสอบมักจะไม่สอดคล้องกับโมเดลนี้จึงไม่สู้จะมีการนำไปประยุกต์ใช้

ในทางปฏิบัติ

Latent Distance Model

Latent Distance Model นี้ความสัมพันธ์ระหว่างความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องกับความสามารถยังคงมีลักษณะเป็นฟังก์ชันแบบขั้นบันได แต่เป็น Stochastic Model กล่าวคือ ความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องสำหรับผู้สอบที่มีความสามารถมากกว่า b จะน้อยกว่า 1 และความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องสำหรับผู้สอบที่มีความสามารถน้อยกว่า b จะมากกว่า 0 เส้นโค้งลักษณะของข้อสอบจะมีลักษณะดังภาพที่ 2



ภาพที่ 2 เส้นโค้งลักษณะของข้อสอบใน Latent Distance Model

ภาพที่ 2 แสดงว่า ผู้ที่มีความสามารถต่ำกว่า b จะมีความน่าจะเป็นในการตอบข้อสอบข้อนี้ได้ถูกต้องเท่ากับ c ($0 \leqslant c < d$) ส่วนผู้ที่มีความสามารถเท่ากับหรือมากกว่า b จะมีความน่าจะเป็นในการตอบข้อสอบข้อนี้ได้ถูกต้องเท่ากับ d ($c < d \leqslant 1$) (Hulin and others 1983 : 17)

แม้ว่าโมเดลนี้จะเป็น Stochastic Model แต่ข้อมูลจากการทดสอบก็มักจะไม่สอดคล้องกับโมเดลนี้อีกเช่นกัน เพราะว่าผู้ที่มีความสามารถมากกว่าหรือน้อยกว่า b จะแสดงพฤติกรรมการตอบสนองข้อสอบที่คงเส้นคงวาเช่นนี้คงจะเป็นไปได้ยาก

Latent Linear Model

Latent Linear Model เป็นโมเดลที่เชื่อว่า เส้นโค้งลักษณะ

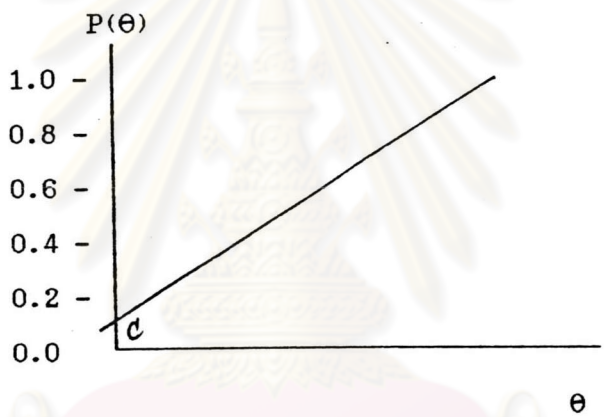
ของข้อสอบมีลักษณะเป็นเส้นตรงดังภาพที่ 3 โดยความสัมพันธ์ระหว่างความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องกับระดับความสามารถ เขียนเป็นสมการได้ดังนี้

$$P(\theta) = c + a\theta \qquad (2 - 27)$$

(Hulin and others 1983 : 18)

เมื่อ a คือความชันของโค้งลักษณะของข้อสอบที่แสดงความสัมพันธ์ระหว่างความน่าจะเป็นในการตอบถูกกับระดับความสามารถ

c คือ Intercept ของฟังก์ชัน



ภาพที่ 3 เส้นโค้งลักษณะของข้อสอบของ Latent Linear Model

ภาพที่ 3 แสดงโค้งลักษณะของข้อสอบตาม Latent Linear Model ซึ่งผู้ที่มีระดับความสามารถเป็น θ มีความน่าจะเป็นในการตอบข้อสอบถูกเท่ากับ c ผู้ที่มีความสามารถสูงขึ้นความน่าจะเป็นในการตอบข้อสอบถูกก็จะมากขึ้นด้วย

Latent Linear Model มีข้อจำกัดอยู่ว่า ถ้า $a \neq 0$ แล้ว ผู้ที่มีความสามารถระดับต่ำมาก ๆ ความน่าจะเป็นในการตอบข้อสอบถูกอาจจะมีค่าเป็นลบและผู้ที่มีความสามารถระดับสูงมากก็อาจจะมีความน่าจะเป็นในการตอบข้อสอบถูกมากกว่า 1 ได้ แต่ตามนิยามความน่าจะเป็นจะมีค่าอยู่ระหว่าง 0 ถึง 1 เท่านั้น

Normal Ogive Model

โมเดลนี้มีความเชื่อว่าความสัมพันธ์ระหว่างความน่าจะเป็นใน

การตอบข้อสอบถูกกับระดับความสามารถอยู่ในรูปของฟังก์ชันการแจกแจงสะสมปกติ (Cumulative Normal Ogive) ซึ่งแสดงได้ด้วย Two - Parameter Normal Ogive Model ดังนี้

$$P_i(\theta) = \int_{-\infty}^{a_i(\theta - b_i)} (1/\sqrt{2\pi})\, e^{-z^2/2}\, dz \qquad (2 - 28)$$

(Hambleton and Swaminathan 1985 : 35)

เมื่อ $P_i(\theta)$ คือความน่าจะเป็นที่ผู้สอบที่มีความสามารถ θ จะตอบข้อสอบข้อที่ i ได้ถูกต้อง

a_i คือพารามิเตอร์อำนาจจำแนกของข้อสอบซึ่งเป็นสัดส่วนกับความชันของ $P_i(\theta)$ ที่จุด $\theta = b_i$

b_i คือพารามิเตอร์ความยากของข้อสอบและแสดงถึงจุดบนสเกลของความสามารถซึ่งผู้สอบมีความน่าจะเป็นในการตอบข้อสอบข้อที่ i ได้ถูกต้อง 50 %

z คือความเบี่ยงเบนปกติ (Normal Deviate) จากการแจกแจงซึ่งมีค่าเฉลี่ย b_i และส่วนเบี่ยงเบนมาตรฐาน $1/a_i$

หากแปลงคะแนนความสามารถของผู้สอบกลุ่มหนึ่งไปจนกระทั่งค่าเฉลี่ยเป็น 0 และส่วนเบี่ยงเบนมาตรฐานเป็น 1 ค่าของ b จะอยู่ระหว่าง - 2.0 และ 2.0 ถ้าค่าของ b เข้าใกล้ - 2.0 แสดงว่าข้อสอบข้อนั้นง่ายมาก และถ้าค่า b เข้าใกล้ 2.0 แสดงว่าข้อสอบยากมากสำหรับผู้สอบกลุ่มนั้น

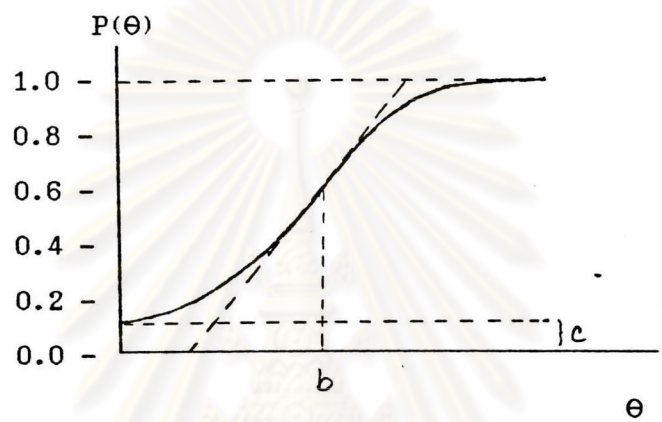
ส่วน a_i ซึ่งเป็นพารามิเตอร์อำนาจจำแนกของข้อสอบ โดยทฤษฎีมีค่าอยู่ระหว่าง $-\infty$ และ $+\infty$ แต่ข้อสอบที่มีอำนาจจำแนกติดลบนั้นไม่เป็นที่ต้องการ และมักไม่ค่อยพบค่า a_i ที่มากกว่า 2 ฉะนั้น โดยปกติค่าของ a_i จะอยู่ระหว่าง 0 และ 2 ถ้า a_i มีค่ามากโค้งลักษณะของข้อสอบก็จะยิ่งชันมาก

ในการทดสอบที่ใช้ข้อสอบแบบปรนัยผู้สอบที่มีความสามารถน้อย ๆ ก็อาจตอบข้อสอบถูกได้ด้วยการเดา ในกรณีที่ θ มีค่าน้อย ๆ นี้ ค่า $P_i(\theta)$ จะไม่เข้าใกล้ 0 แต่จะเข้าใกล้ค่า ๆ หนึ่งที่มากกว่า 0 โมเดลที่มีพารามิเตอร์ 2 ตัวไม่เหมาะกับสภาพการณ์เช่นนี้จึงได้มีการเพิ่มพารามิเตอร์ของข้อสอบอีกตัวหนึ่ง คือ c ได้เป็น Three - Parameter Normal Ogive Model ดังนี้ (Hambleton and Swaminathan 1985 : 50)

$$P_i(\theta) = c_i + (1 - c_i) \int_{-\infty}^{a_i(\theta - b_i)} (1/\sqrt{2\pi})\, e^{-z^2/2}\, dz \qquad (2 - 29)$$

เมื่อ c คือ Lower Assymtote ของโค้งลักษณะของข้อสอบ แสดงถึงความน่าจะเป็นที่ผู้ที่ไม่มีความสามารถเลย ($\theta = -\infty$) จะตอบข้อสอบได้ถูกต้อง จึงเรียกว่าเป็นค่าการเดา ซึ่งมีค่าอยู่ระหว่าง 0 ถึง 1 ถ้าข้อสอบข้อนั้นไม่สามารถตอบถูกได้ด้วยการเดาแล้ว c จะมีค่าเป็น 0

เส้นโค้งลักษณะของข้อสอบใน Three - Parameter Normal Ogive Model จะมีลักษณะเป็นรูปตัว S ดังภาพที่ 4



ภาพที่ 4 เส้นโค้งลักษณะของข้อสอบของ Three - Parameter Normal Ogive Model

Logistic Model

เส้นโค้งลักษณะของข้อสอบในโมเดลนี้มีลักษณะเป็นรูปตัว S เช่นเดียวกับ Normal Ogive Model แต่ใน Logistic Model ความสัมพันธ์ระหว่างความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องกับระดับความสามารถอยู่ในรูปของฟังก์ชันการแจกแจงสะสมแบบ Logist ดังแสดงในสมการต่อไปนี้ (Lord 1980 : 12)

$$P_i(\theta) = c_i + (1 - c_i)\,[1 + e^{-1.7a_i(\theta - b_i)}]^{-1} \qquad (2 - 30)$$

โดย a_i, b_i และ c_i เป็นค่าพารามิเตอร์ของข้อสอบซึ่งมีความหมายเช่นเดียวกับ Normal Ogive Model ส่วน e คือค่าคงที่มีค่าประมาณ 2.71828

เส้นโค้งลักษณะของข้อสอบของ Normal Ogive Model กับของ Logistic Model จะมีลักษณะใกล้เคียงกันมากเนื่องจากค่าฟังก์ชันแตกต่างกันเล็กน้อยเมื่อมีการปรับค่าตัวแปรด้วย Scaling Factor ซึ่งมีค่าเท่ากับ 1.7 แต่ Logistic Model

คำนวณได้ง่ายและสะดวกกว่ามากจึงเป็นที่นิยมใช้ในทางปฏิบัติมากกว่า

สมการ (2 - 30) มีพารามิเตอร์ของข้อสอบ 3 ตัว จึงเรียกว่า Three - Parameter Logistic Model ในบางสถานการณ์ของการสอบที่ผู้สอบมีความน่าจะเป็นในการตอบถูกด้วยการเดาน้อยมากอาจกำหนดให้พารามิเตอร์การเดาเป็น 0 ก็จะเหลือพารามิเตอร์เพียง 2 ตัว คือ a_i และ b_i สมการ (2 - 30) ก็จะเขียนใหม่ได้ดังสมการ (2 - 31) ซึ่งเรียกว่า Two - Parameter Logistic Model

$$P_i(\theta) = 1 / [1 + e^{-1.7a_i(\theta - b_i)}] \qquad (2 - 31)$$

ในกรณีที่กำหนดให้ $c_i = 0$ และถือว่าข้อสอบทุกข้อมีค่าอำนาจจำแนกเท่ากัน ($a_i = \bar{a}$) สมการก็จะเหลือพารามิเตอร์เพียงตัวเดียวดังในสมการ (2 - 32) ซึ่งเรียกว่า One - Parameter Logistic Model

$$P_i(\theta) = 1 / [1 + e^{-1.7\bar{a}(\theta - b_i)}] \qquad (2 - 32)$$

บางครั้งผู้สอบที่มีความสามารถสูงก็มิใช่ว่าจะตอบข้อสอบได้ถูกต้องเสมอไป อาจจะมีการตอบผิดด้วยความสะเพร่าหรืออื่น ๆ เพื่อแก้ปัญหานี้ Barton และ Lord (1981 อ้างถึงใน Hambleton and Swaminathan 1985 : 48) ได้เสนอ Four - Parameter Logistic Model ขึ้นดังในสมการ (2 - 33)

$$P_i(\theta) = c_i + (\gamma_i - c_i) [1 + e^{-Da_i(\theta - b_i)}]^{-1} \qquad (2 - 33)$$

โมเดลนี้แตกต่างจาก Three - Parameter Logistic Model ที่ γ_i นั้นถือว่ามีค่าต่ำกว่า 1 เล็กน้อย อย่างไรก็ตามโมเดลนี้ก็ยังเป็นเพียงโมเดลที่น่าสนใจในเชิงทฤษฎีเท่านั้น เพราะว่า Barton และ Lord ยังไม่สามารถหาผลได้ที่เพิ่มขึ้นจากโมเดลนี้ได้

Rasch Model

Rasch Model เป็นโมเดลที่ได้รับความนิยมมากโมเดลหนึ่ง Gorge Rasch ผู้พัฒนาโมเดลนี้ได้เสนอในรูปสมการดังต่อไปนี้ (Hambleton and Cook 1977 : 82 อ้างถึงใน จักรกฤษณ์ สำราญใจ 2531 : 37)

$$P_i(\theta^*) = \theta^* / (\theta^* + b_i{}^*) \qquad (2 - 34)$$

โดย $\theta^* = e^{1.7a\theta}$ และ $b_i{}^* = e^{1.7ab_i}$ จะเห็นว่า สมการ (2 - 34) ของ Rasch Model ก็คือสมการ (2 - 32) ของ One - Parameter Logistic Model นั่นเอง

Nominal Response Model

โมเดลที่เสนอมาก่อนหน้านี้เป็นโมเดลที่ใช้กับข้อสอบที่มีการให้คะแนนแบบตอบผิดได้ 0 ตอบถูกได้ 1 คะแนน ส่วน Nominal Response Model นี้ใช้กับข้อสอบที่มีการตรวจให้คะแนนมากกว่า 2 ค่า จุดมุ่งหมายก็เพื่อเพิ่มความแม่นยำในการประมาณค่าความสามารถของผู้สอบโดยใช้ประโยชน์จากสารสนเทศทั้งการตอบถูกและการตอบผิด โมเดลนี้แต่ละตัวเลือกจะมีโค้งลักษณะของตัวเลือก (Item Option Characteristic Curve) โค้งของตัวเลือกที่เป็นคำตอบถูกจะมีลักษณะเป็น Monotonic Increasing Function ส่วนโค้งของตัวลวงขึ้นอยู่กับการเลือกตัวเลือกนั้นของผู้สอบที่มีความสามารถในระดับต่าง ๆ รูปแบบทางคณิตศาสตร์ของโค้งลักษณะตัวเลือกมีหลายแบบ เช่น รูปแบบของ Bock ที่สมมุติว่าความน่าจะเป็นที่ผู้สอบที่มีความสามารถ θ จะเลือกตัวเลือก k จากตัวเลือกทั้งหมด m ตัวของข้อสอบข้อที่ i กำหนดโดยสมการ (2 - 35) (Hambleton and Swaminathan 1985 : 50)

$$P_{ik}(\theta) = [e^{b^*_{ik} + a^*_{ik}\theta}] / [\sum^{m}_{k=1} (e^{b^*_{ik} + a^*_{ik}\theta}) \qquad (2 - 35)$$

เมื่อ b^*_{ik} และ a^*_{ik} คือค่าพารามิเตอร์ของตัวเลือก k ที่ระดับความสามารถ θ ผลรวมของความน่าจะเป็นในการเลือกตัวเลือกทุกตัวจะเท่ากับ 1

Grade Response Model

Grade Response Model เป็นโมเดลที่ใช้กับการสอบที่พฤติกรรมการตอบสนองสามารถจัดประเภทเรียงตามลำดับได้ Samejima ได้เสนอโมเดลที่สมมุติว่าพฤติกรรมการตอบสนองข้อสอบสามารถจัดแบ่งประเภทได้เป็น $m_i + 1$ ประเภทและให้คะแนน $x_i = 0,1,\ldots,m_i$ ตามลำดับ ความน่าจะเป็นที่ผู้ที่มีความสามารถ θ จะได้คะแนนข้อนี้เป็น x_i คือ

$$P_{x_i}(\theta) = P_{x_i}{}^*(\theta) - P^*_{(x_i+1)}(\theta) \qquad (2 - 36)$$

(Hamblton and Swaminathan 1985 : 51)

เมื่อ $P_x^*(\theta)$ คือ ฟังก์ชันการตอบสนองข้อสอบของการให้คะแนนแบบ 2 ค่า โดยผู้ที่ได้คะแนนต่ำกว่า x_1 ถือเป็น 0 และผู้ที่ได้คะแนนเท่ากับ x_1 หรือมากกว่าถือเป็น 1 ส่วนรูปแบบของฟังก์ชัน $P_x^*(\theta)$ จะใช้โมเดลที่ใช้กับข้อสอบที่ให้คะแนน 2 ค่าโมเดลใดนั้นขึ้นอยู่กับผู้ใช้จะเลือก

3. ฟังก์ชันสารสนเทศของข้อสอบและแบบสอบ

ในสถานการณ์ทั่วไปถ้ามั่นใจค่อนข้างมากว่าเหตุการณ์อย่างหนึ่งจะเกิดขึ้นแสดงว่ามีข่าวสารข้อมูลหรือสารสนเทศเกี่ยวกับเหตุการณ์นั้นมากพอสมควร ในทางกลับกันถ้าไม่มีข่าวสารข้อมูลเกี่ยวกับเหตุการณ์นั้นหรือมีน้อยความมั่นใจก็จะมีน้อยตามไปด้วย ในการอ้างอิงเชิงสถิติความแม่นยำของการประมาณค่าพารามิเตอร์ของกลุ่มประชากรอาจดูได้จากช่วงกว้างของค่าประมาณ ถ้าไม่มีสารสนเทศใด ๆ เกี่ยวกับประชากรเลยก็อาจจะต้องประมาณค่าเป็นค่าใด ๆ ในช่วง $-\infty$ ถึง $+\infty$ แต่ถ้ามีสารสนเทศเกี่ยวกับประชากรบ้างช่วงของค่าประมาณจะแคบเข้า นั่นคือความแม่นยำในการประมาณเริ่มมีมากขึ้น ตามปกติความแม่นยำของการประมาณค่าพารามิเตอร์จะแสดงได้ด้วยค่าความคลาดเคลื่อนมาตรฐานของการประมาณค่า กล่าวคือ ถ้าความคลาดเคลื่อนมาตรฐานของการประมาณค่ามีมาก ความแม่นยำของการประมาณค่าก็จะมีน้อย เพราะช่วงของค่าประมาณจะกว้าง ในทางกลับกัน ถ้าความคลาดเคลื่อนมาตรฐานของการประมาณค่ามีน้อยความแม่นยำของการประมาณค่าก็จะมีมาก ช่วงของค่าประมาณจะแคบ แสดงว่าค่าสารสนเทศมีความสัมพันธ์กับความคลาดเคลื่อนมาตรฐานของการประมาณค่าโดยมีความสัมพันธ์ในทิศทางกลับกัน คือความคลาดเคลื่อนมาตรฐานของการประมาณค่าเท่ากับ $1 / \sqrt{\text{สารสนเทศ}}$ (จักรกฤษณ์ สำราญใจ 2531 : 40)

ในทฤษฎีการตอบสนองข้อสอบจะใช้ผลการตอบแบบสอบประมาณค่าความสามารถของผู้สอบ การประเมินคุณภาพของแบบสอบดูได้จากความถูกต้องแม่นยำในการประมาณค่าความสามารถโดยใช้คะแนนจากแบบสอบ ค่าสารสนเทศจากแบบสอบจะเป็นดัชนีชี้ถึงความถูกต้องแม่นยำของการประมาณค่า (Birnbaum 1968 : 418 อ้างถึงใน จักรกฤษณ์ สำราญใจ 2531 : 40) ค่าสารสนเทศคือปริมาณที่เป็นส่วนกลับของกำลังสองของความกว้างของช่วงความเชื่อมั่นของการประมาณค่าความสามารถด้วยคะแนนจากแบบสอบ (Lord 1980 : 65) ถ้าให้ y เป็นคะแนนที่ได้จากการตอบแบบสอบ ค่าสารสนเทศสำหรับคะแนน y แสดงได้ด้วยสมการต่อไปนี้ (Lord 1980 : 67)

$$I\{\theta, y\} = [d/d\theta\, \mu_{y|\theta}]^2 / \mathrm{Var}(y \mid \theta) \qquad (2-37)$$

จะเห็นว่าค่าสารสนเทศสำหรับคะแนน y ก็คือ กำลังสองของอัตราส่วนระหว่างความชันของเส้นถดถอยของ y บน θ กับความคลาดเคลื่อนมาตรฐานของการวัด y ที่ระดับความสามารถ θ นั่นเอง ดังนั้นค่าสารสนเทศของแบบสอบฉบับหนึ่งจึงมีได้หลายค่าแตกต่างกันไปตามระดับของ θ ที่ระดับความสามารถค่าใดค่าหนึ่ง ค่าสารสนเทศจะมากหรือน้อยขึ้นอยู่กับสิ่งต่อไปนี้ (Lord 1980 : 68)

(1) **ค่าความคลาดเคลื่อนมาตรฐานของการวัด y ที่ระดับความสามารถนั้น** ถ้ามีความคลาดเคลื่อนน้อยค่าสารสนเทศจาก y เกี่ยวกับ θ ก็จะมาก

(2) **ความชันของเส้นถดถอยของ y บน θ** ถ้ามีความชันมากค่าสารสนเทศจาก y เกี่ยวกับ θ ก็จะมากเช่นกัน

นอกจากนี้ค่าสารสนเทศของแบบสอบยังขึ้นอยู่กับ**คุณภาพและจำนวนข้อสอบ**ด้วย ค่าสารสนเทศของแบบสอบจะเพิ่มมากขึ้นเมื่อแบบสอบมีข้อสอบจำนวนมากขึ้น (Hambleton and Swaminathan 1985 : 105,107)

Birnbaum (1968 : 453 - 454 อ้างถึงใน จักรกฤษณ์ สำราญใจ 2531 : 41) ได้ให้สูตรทั่วไปสำหรับคำนวณค่าสารสนเทศของคะแนน y จากแบบสอบใด ๆ ไว้ดังนี้

$$\text{ถ้า } y = \Sigma^{n}_{i=1} w_i . u_i$$

ค่าสารสนเทศของแบบสอบคือ

$$I(\theta,y) = [\Sigma^{n}_{i=1} w_i P_i'(\theta)]^2 / \Sigma^{n}_{i=1} w_i^2 P_i(\theta)Q_i(\theta) \qquad (2 - 38)$$

เมื่อ $P_i'(\theta) = d P_i(\theta) / d \theta$

ส่วนค่าสารสนเทศของข้อสอบก็จะหาได้จากการแทนค่า y ด้วย u_i นั่นคือ w_i มีค่าเท่ากับ 1 และ n มีค่าเท่ากับ 1 สมการหาค่าสารสนเทศของข้อสอบจึงเป็นดังนี้

$$I(\theta,u_i) = [P_i'(\theta)]^2 / P_i(\theta)Q_i(\theta) \qquad (2 - 39)$$

ในกรณีของการให้คะแนนโดยการกำหนดน้ำหนักอย่างเหมาะสม (Optimal Weight) โดย

$$w_i = P_i'(\theta) / P_i(\theta)Q_i(\theta)$$

ค่าสารสนเทศของแบบสอบจะเป็นดังนี้

$$I(\theta) = \sum_{i=1}^{n} \{[P_i'(\theta)]^2 / P_i(\theta)Q_i(\theta)\} \qquad (2 - 40)$$

ซึ่งก็คือผลรวมของค่าสารสนเทศของข้อสอบแต่ละข้อในแบบสอบนั่นเอง ค่า $I(\theta)$ นี้จะเป็นค่าสารสนเทศที่มากที่สุดที่จะได้จากแบบสอบฉบับใดฉบับหนึ่ง นั่นคือ $I(\theta,y) < I(\theta)$ เสมอ

เนื่องจากฟังก์ชันลักษณะของข้อสอบ $P_i(\theta)$ ขึ้นอยู่กับค่าพารามิเตอร์ของข้อสอบข้อที่ i คือ a_i, b_i และ c_i ดังนั้นค่าพารามิเตอร์ของข้อสอบจึงเป็นตัวกำหนดค่าสารสนเทศของข้อสอบและแบบสอบ จึงอาจกล่าวได้ว่า ค่าสารสนเทศของข้อสอบหรือแบบสอบเป็นดัชนีผสม (Composite Index) ที่สร้างจากดัชนีบอกคุณลักษณะของข้อสอบหลายลักษณะรวมเป็นดัชนีเพียงตัวเดียวเพื่อชี้ถึงคุณภาพของข้อสอบหรือแบบสอบ อีกทั้งคุณสมบัติความไม่แปรเปลี่ยนของค่าพารามิเตอร์ของข้อสอบ ค่าสารสนเทศจึงเหมาะที่จะใช้เป็นดัชนีบอกคุณภาพของข้อสอบหรือแบบสอบได้ดีกว่าค่าสถิติหรือค่าดัชนีอื่น ๆ ตามแนวคิดของทฤษฎีการวัดแบบดั้งเดิม (จักรกฤษณ์ สำราญใจ 2531 : 42)

4. ประสิทธิภาพเชิงสัมพัทธ์ของข้อสอบและแบบสอบ

จุดมุ่งหมายของการสอบตามทฤษฎีการตอบสนองข้อสอบก็คือ การใช้ผลการสอบประมาณค่าความสามารถของผู้สอบซึ่งไม่สามารถสังเกตได้โดยตรง ดังนั้น การประเมินคุณภาพของแบบสอบจึงต้องพิจารณาที่ความถูกต้องแม่นยำของการใช้คะแนนสอบประมาณค่าระดับความสามารถ ดัชนีที่บอกถึงความถูกต้องแม่นยำในการประมาณค่าความสามารถก็คือ ค่าสารสนเทศของแบบสอบ เนื่องจากวิธีการให้คะแนนใด ๆ ที่ไม่ใช่การกำหนดน้ำหนักที่เหมาะสมแล้ว ค่าสารสนเทศที่ได้จะมีค่าน้อยกว่าค่าสารสนเทศที่ได้จากวิธีการให้คะแนนโดยการกำหนดน้ำหนักคะแนนที่เหมาะสมเสมอในทุกระดับความสามารถ Birnbaum (1968 : 471 - 472 อ้างถึงใน จักรกฤษณ์ สำราญใจ 2531 : 43) ได้เสนอแนวคิดในการวัดประสิทธิภาพของแบบสอบโดยถือว่าแบบสอบที่มีประสิทธิภาพมากคือแบบสอบที่คะแนนจากการสอบให้ค่าสารสนเทศสูงใกล้เคียงกับสารสนเทศที่ได้จากการให้คะแนนแบบกำหนดน้ำหนักที่เหมาะสม ดัชนีบอกประสิทธิภาพของแบบสอบก็คือ อัตราส่วนระหว่างค่าสารสนเทศที่ได้จากการให้คะแนนแบบ x กับค่าสารสนเทศที่ได้จากการให้คะแนนแบบกำหนดน้ำหนักที่เหมาะสม นั่นคือ

$$Eff(\theta,x) = I(\theta,x) / I(\theta) \qquad (2 - 41)$$

ในกรณีที่ต้องการเปรียบเทียบประสิทธิภาพของแบบสอบ 2 ฉบับที่ใช้วัดลักษณะหรือความสามารถอย่างเดียวกันก็สามารถหาค่าดัชนีประสิทธิภาพเชิงสัมพัทธ์ (Relative Efficiency) ของแบบสอบทั้งสองได้ โดยหาอัตราส่วนระหว่างค่าสารสนเทศที่ได้จากแบบสอบทั้งสองฉบับ นั่นคือ

$$RE(\theta,x_1,x_2) = I_1(\theta,x_1) / I_2(\theta,x_2) \qquad (2 - 42)$$

ค่าดัชนีประสิทธิภาพเชิงสัมพัทธ์นี้ไม่ขึ้นอยู่กับมาตราที่ใช้วัดความสามารถ กล่าวคือ ค่าดัชนีจะไม่แปรเปลี่ยนไปตามการแปลงมาตราการวัดความสามารถ (Lord 1980 : 89)

มโนทัศน์เกี่ยวกับประสิทธิภาพเชิงสัมพัทธ์นอกจากจะประยุกต์ใช้ในการเปรียบเทียบประสิทธิภาพของแบบสอบได้แล้วยังสามารถนำไปใช้ในการเปรียบเทียบประสิทธิภาพของข้อสอบได้อีกด้วย (Hambleton and Swaminathan 1985 : 124)

4. การวิจัยเกี่ยวกับการเปรียบเทียบประสิทธิภาพของแบบสอบ

การนำแนวคิดเกี่ยวกับประสิทธิภาพเชิงสัมพัทธ์ไปใช้เปรียบเทียบประสิทธิภาพของแบบสอบยังไม่แพร่หลายนัก งานวิจัยที่มีอยู่ได้แก่ผลงานของ Lord (1974 : 247 - 254 อ้างถึงใน จักรกฤษณ์ สำราญใจ 2531 : 43) ซึ่งนำแนวคิดนี้ไปเปรียบเทียบประสิทธิภาพของแบบสอบ Metropolitan Reading Tests กับแบบสอบการอ่านอื่น ๆ อีก 6 ฉบับ นอกจากนี้ Lord ยังได้ศึกษาเปรียบเทียบคุณภาพของแบบสอบที่ประกอบด้วยข้อสอบเลือกตอบที่มีจำนวนตัวเลือกต่างกัน พบว่า การลดจำนวนตัวเลือกต่อข้อลงในขณะที่เพิ่มจำนวนข้อสอบให้มากขึ้นมีผลทำให้ประสิทธิภาพของแบบสอบเพิ่มขึ้นในกลุ่มผู้สอบที่มีความสามารถในระดับสูงและทำให้ประสิทธิภาพของแบบสอบลดลงในกลุ่มผู้สอบที่มีความสามารถในระดับต่ำ แบบสอบที่ประกอบด้วยข้อสอบที่มีจำนวนตัวเลือก 3 ตัวมีประสิทธิภาพสูงที่สุด (Lord 1976 : 36 - 39 อ้างถึงใน จักรกฤษณ์ สำราญใจ 2531 : 45)

ส่วนงานวิจัยภายในประเทศนั้น จักรกฤษณ์ สำราญใจ (2531) ได้ทำการศึกษาประสิทธิภาพเชิงสัมพัทธ์ของข้อสอบและแบบสอบโดยศึกษากับข้อสอบเลือกตอบชนิดแบบฉบับ ข้อสอบเลือกตอบชนิดตัดสินคำตอบทุกตัวเลือกที่ตรวจให้คะแนนแก่ความรู้ที่ถูกต้องสมบูรณ์ และข้อสอบเลือกตอบชนิดตัดสินคำตอบทุกตัวเลือกที่ตรวจให้คะแนนแก่ความรู้บางส่วน แบบสอบที่ใช้ศึกษาคือ

แบบสอบวัดผลสัมฤทธิ์ทางการเรียนวิชาการประเมินผลและการสร้างแบบทดสอบ ผลการวิจัยพบว่า ข้อสอบเลือกตอบชนิดตัดสินคำตอบทุกตัวเลือกที่ตรวจให้คะแนนแก่ความรู้บางส่วน(MTF.P) มีประสิทธิภาพสูงกว่าข้อสอบเลือกตอบชนิดแบบฉบับ(T.MC) และข้อสอบเลือกตอบชนิดตัดสินคำตอบทุกตัวเลือกที่ตรวจให้คะแนนแก่ความรู้ที่ถูกต้องสมบูรณ์ (MTF.A) สำหรับผู้สอบที่มีความสามารถในระดับต่ำมากไปจนถึงระดับสูง แต่ข้อสอบ MTF.P มีประสิทธิภาพต่ำกว่าข้อสอบ T.MC และข้อสอบ MTF.A ที่ระดับความสามารถสูงมาก ข้อสอบ MTF.A มีประสิทธิภาพสูงกว่าข้อสอบ T.MC สำหรับผู้สอบที่มีความสามารถในระดับสูงขึ้นไป แต่ข้อสอบ MTF.A มีประสิทธิภาพต่ำกว่าข้อสอบ T.MC ที่ระดับความสามารถปานกลางค่อนข้างสูงลงมาถึงระดับต่ำมาก