บทที่ 2

เอกสารและงานวิจัยที่เกี่ยวข้อง

ผู้วิจัยได้ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ปรากฏดังนี้

1. แนวคิดที่เกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ
   1.1 ความเป็นมาของการทำหน้าที่ต่างกันของข้อสอบ
   1.2 ความหมายของการทำหน้าที่ต่างกันของข้อสอบ
   1.3 ประเภทของการทำหน้าที่ต่างกันของข้อสอบ
   1.4 หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ
   1.5 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ
2. โครงการประเมินผลสัมฤทธิ์ทางการเรียนระดับเขตพื้นที่การศึกษา
3. งานวิจัยที่เกี่ยวข้อง

1. แนวคิดที่เกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

1.1 ความเป็นมาของการทำหน้าที่ต่างกันของข้อสอบ

การศึกษาถึงคุณภาพของข้อสอบจากผลการตรวจข้อสอบของผู้สอบกลุ่มต่างๆในประชากรมีมานานแล้ว แต่การศึกษาคุณภาพด้านความยุติธรรมของข้อสอบหรือแบบสอบระหว่างผู้สอบกลุ่มต่างๆ เริ่มมีการศึกษากันอย่างจริงจังใน ช่วงปลายทศวรรษของปี ค.ศ. 1960 มีการเสนอวิธีการต่างๆ เพื่อตรวจสอบความลำเอียงของข้อสอบ (Item bias) ความลำเอียงของแบบสอบ (Test bias) และความลำเอียงในการคัดเลือก (Selection bias) ความพยายามของการตรวจสอบความลำเอียงดังกล่าวดำเนินไปเพื่อจำแนกข้อสอบที่ทำหน้าที่ไม่เหมาะสมหรือไม่ยุติธรรมสำหรับปรับปรุง หรือตัดข้อสอบนั้นออกจากแบบทดสอบเป็นการขจัดข้อสอบที่ไม่ทำให้เกิดปัญหาความไม่ยุติธรรมระหว่างกลุ่มต่าง ๆ ที่มีลักษณะบางอย่างแตกต่างกัน เช่น เชื้อชาติ ศาสนา วัฒนธรรม ภูมิลำเนา สังคม เพศ ภาษา อายุ ประสบการณ์ เป็นต้น เพื่อพัฒนาแบบสอบให้มีคุณภาพเหมาะสมสำหรับนำไปใช้ทดสอบต่อไป

ในเวลาต่อมา นักการวัดผลการศึกษาได้ทำการศึกษาความลำเอียงของข้อสอบ (Item bias) กันอย่างกว้างขวาง ทำให้เกิดความสับสนของการใช้คำและความหมาย มีประเด็นโต้แย้งกัน

ว่าความลำเอียงของข้อสอบเป็นผลการตัดสินว่าข้อสอบมีความยุติธรรมหรือไม่ อันส่งผลต่อการบรรลุจุดมุ่งหมายของการใช้แบบสอบหรือความลำเอียงของข้อสอบ เป็นสารสนเทศทางสถิติที่ได้จากข้อสอบเกี่ยวกับความสัมพันธ์ระหว่างคุณลักษณะที่ข้อสอบมุ่งวัด กับประสบการณ์ของผู้สอบกลุ่มต่าง ๆ ที่ทำการสอบเมื่อกลุ่มผู้สอบต่างกลุ่มกันตอบข้อสอบข้อเดียวกัน ความแตกต่างที่เกิดขึ้นอาจมาจากความไม่เหมาะสมของข้อคำถาม ซึ่งสามารถเกิดขึ้นได้หลายลักษณะ หรือประสบการณ์ของผู้สอบซึ่งอาจมีลักษณะพื้นฐานเดิมแตกต่างกันในหลายสถานการณ์จึงไม่เหมาะสมที่จะใช้คำว่า ข้อสอบลำเอียง (Biased item) เนื่องจากเป็นภาษาที่มีความหมายในเชิงลบประกอบกับเกณฑ์ที่ใช้สำหรับตัดสินความลำเอียงยังมีความคลุมเครือและค่อนข้างสับสน ดังนี้จึงควรเปลี่ยนมาใช้คำว่า การทำหน้าที่ต่างกันของข้อสอบ (Differential item functioning : DIF) ซึ่งเป็นคำที่มีความหมายเป็นกลางและเหมาะสมกว่า (Holland & Thayer, 1988 ; Holland & Wianer, 1993 อ้างถึงใน ศิริชัย กาญจนวาสี, 2550 : 115-116)

การทำหน้าที่ต่างกันของข้อสอบ เป็นเงื่อนไขที่จำเป็นในการประเมินความลำเอียงของข้อสอบแต่ไม่ใช่เงื่อนไขที่เพียงพอในการประเมินความลำเอียงของข้อสอบ ถ้าใช้เฉพาะวิธีการทางสถิติเพียงด้านเดียว ผลการตรวจพบข้อสอบทำหน้าที่ต่างกันที่ได้ไม่อาจสรุปได้ว่าข้อสอบนั้นมีความลำเอียงหรือไม่ เนื่องจากการประเมินความลำเอียงของข้อสอบยังต้องรวมไปถึงการพิจารณาเนื้อหาสาระของข้อสอบและจุดมุ่งหมายในการวัดของแบบสอบที่จะต้องพิจารณาโดยผู้เชี่ยวชาญที่เรียกว่า วิธีตัดสินข้อสอบ (Judgmental Method) ก่อนที่จะสรุปว่าข้อสอบนั้นลำเอียงหรือไม่ (Camilli & Shepard, 1994 อ้างถึงใน เสรี ชัดแช้ม, 2540 : 42)

1.2 ความหมายของการทำหน้าที่ต่างกันของข้อสอบ

นักวิจัยทางการวัดผลการศึกษาหลายท่านได้ให้ความหมายของความลำเอียงของข้อสอบและการทำหน้าที่ต่างกันของข้อสอบ ไว้ดังนี้

Kederman (1990) กล่าวว่า ความลำเอียงของข้อสอบ หมายถึง คะแนนข้อสอบของกลุ่มผู้สอบที่มีความสามารถเท่าเทียมกัน แต่มาจากต่างกลุ่มกัน มีความแตกต่างกันอย่างเป็นระบบ

Dorans & Kulick (1986) กล่าวว่า ความลำเอียงของข้อสอบ หมายถึง โอกาสในการตอบข้อสอบได้ถูกต้องของผู้สอบกลุ่มหนึ่งมีค่าต่ำกว่าหรือสูงกว่าผู้สอบอีกกลุ่มหนึ่งที่มีระดับความสามารถเดียวกัน

Shealy & Stout (1993) กล่าวว่า การทำหน้าที่ต่างกันของข้อสอบ หมายถึง ข้อสอบที่เข้าข้างผู้สอบกลุ่มหนึ่งมากกว่าผู้สอบอีกกลุ่มหนึ่งที่นำมาจับคู่เปรียบเทียบกัน ซึ่งทำให้ผู้สอบกลุ่มหนึ่งได้ประโยชน์ แต่ผู้สอบอีกกลุ่มหนึ่งเสียประโยชน์

Holland & Wainer (1993) กล่าวว่า การทำหน้าที่ต่างกันของข้อสอบ หมายถึง สารสนเทศทางสถิติของข้อสอบที่ได้จากผลการสอบของผู้สอบต่างกลุ่มกัน และมีความสามารถเท่ากัน แต่มีโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกัน

Camilli & Shepard (1994) กล่าวว่า การทำหน้าที่ต่างกันของข้อสอบ หมายถึง ความเป็นพหุมิติในการวัดของข้อสอบ ซึ่งแสดงได้จากการแจกแจงความสามารถหลัก (Primary ability) ของกลุ่มผู้สอบตั้งแต่ 2 กลุ่มขึ้นไปมีความเท่ากัน แต่มีการแจกแจงความสามารถรอง (Secondary ability) แตกต่างกัน

Potanza & Dorans (1995) กล่าวว่า การทำหน้าที่ต่างกันของข้อสอบ หมายถึง ถ้าข้อสอบทำหน้าที่ต่างกัน จะทำให้ผลการตอบข้อสอบระหว่างกลุ่มผู้เข้าสอบสองกลุ่มที่นำมาเปรียบเทียบกันจะแตกต่างกัน

Narayanan & Swaminathan (1996) กล่าวว่า การทำหน้าที่ต่างกันของข้อสอบ หมายถึง ฟังก์ชันการตอบสนองข้อสอบ ซึ่งคำนวณจากลุ่มผู้เข้าสอบกลุ่มย่อยที่ต่างกัน มีค่าไม่เท่ากัน

Scheuneman และ Bleistin (1999) กล่าวว่า การทำหน้าที่ต่างกันของข้อสอบ หมายถึง การดำเนินการเกี่ยวกับการตัดสินข้อสอบของแต่ละคนในทางที่เหมือนกันของคนสองกลุ่มโดยทั่วไปมักนิยามในส่วนที่เกี่ยวข้องกับเชื้อชาติ ชนชาติ ภูมิหลัง เพศ อายุ และประสบการณ์หรือเงื่อนไขอื่น ๆ ที่ทำให้เสียเปรียบ ลักษณะเช่นนี้ จะเกิดขึ้นเมื่อกลุ่มผู้สอบสนใจเกี่ยวกับความแตกต่างของระดับความสามารถ ความรู้ หรือทักษะที่ถูกวัดทำให้เกิดการเปรียบเทียบสำหรับผู้ที่เข้าสอบ

Anderson และ Demar (2002) ได้กล่าวว่า การทำหน้าที่ต่างกันของข้อสอบ หมายถึง การที่ผู้เข้าสอบซึ่งเป็นกลุ่มย่อย (Subgroup) และมีความสามารถเท่าเทียมกันกับกลุ่มประชากรหลักได้รับการตอบสนองในการตอบถูกจากข้อสอบหรือแบบสอบแตกต่างจากผู้เข้าสอบในกลุ่มประชากรหลัก โดยการทำหน้าที่ต่างกันของข้อสอบเช่นนี้จะเกิดขึ้นเมื่อประชากรมีความแตกต่างกันด้านเพศ เชื้อชาติ ภาษาพูด หรือวัฒนธรรม

McCallon และ Schumacker (2002) ได้กล่าวว่า การทำหน้าที่ต่างกันของข้อสอบ หมายถึง ลักษณะที่แสดงถึงความไม่ยุติธรรมของข้อสอบโดยจะเกิดขึ้นในกรณีที่ผู้สอบซึ่งมีความรู้ความสามารถเท่ากันแต่มาจากต่างกลุ่มกัน และได้รับการตอบสนองจากข้อสอบเกี่ยวกับโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกัน

ศิริชัย กาญจนวาสี (2550 : 117) กล่าวว่า การทำหน้าที่ต่างกันของข้อสอบ หมายถึง การที่ข้อสอบทำให้ผู้สอบจากกลุ่มต่างกันที่มีความสามารถหรือคุณลักษณะที่มุ่งวัดเท่ากัน มีโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกันหรือมีฟังก์ชันการตอบสนองข้อสอบแตกต่างกัน

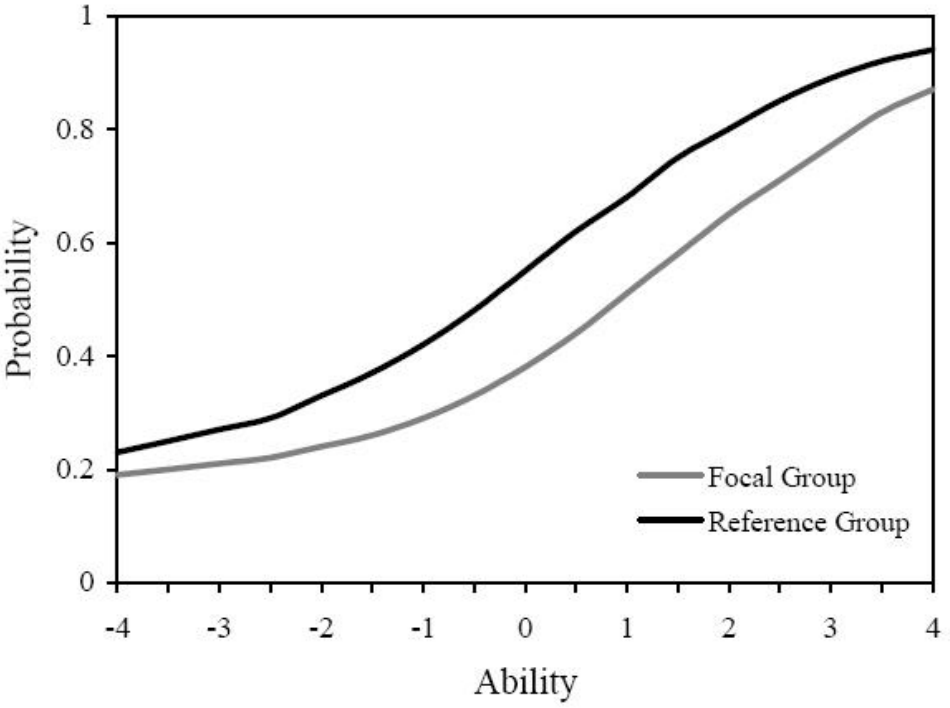
การทำหน้าที่ต่างกันของข้อสอบ เกิดขึ้นเมื่อนำข้อสอบไปทดสอบกับผู้สอบกลุ่มย่อยต่างกัน ที่มีความสามารถหลัก (Primary ability) ระดับเดียวกัน หรือมีคุณลักษณะแฝง (Latent trait) ที่ต้องการวัดเท่ากัน แต่มีความสามารถรอง (Secondary ability) แตกต่างกัน ทำให้ผู้สอบต่างกลุ่มที่นำมาจับคู่เปรียบเทียบมีโอกาสตอบ ข้อสอบถูกแตกต่างกัน

1.3 ประเภทของการทำหน้าที่ต่างกันของข้อสอบ

การทำหน้าที่ต่างกันของข้อสอบ เป็นการเปรียบเทียบผลการตอบข้อสอบระหว่างผู้เข้าสอบ 2 กลุ่ม คือ กลุ่มเปรียบเทียบ (Focal group หรือกลุ่ม F) เป็นกลุ่มที่สนใจศึกษาและคาดว่าจะเป็นกลุ่มที่เสียเปรียบในการตอบข้อสอบ กล่าวคือ มีโอกาสตอบข้อสอบถูกได้น้อยกว่าผู้เข้าสอบกลุ่มอ้างอิง และกลุ่มอ้างอิง (Reference group หรือ กลุ่ม R) เป็นกลุ่มที่คาดว่าจะได้เปรียบในการตอบข้อสอบได้ถูกต้อง

ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ จะพบว่า ข้อสอบสามารถทำหน้าที่แตกต่างได้ 2 ประเภท (Mellenbergh, 1982 อ้างถึงใน ศิริชัย กาญจนวาสี, 2550 : 118) ได้แก่ การทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป (Uniform) และแบบอเนกรูป (Nonuniform)

1) ข้อสอบทำหน้าที่ต่างกันแบบเอกรูป (Uniform DIF) หมายถึงข้อสอบที่ทำให้ผู้สอบกลุ่มหนึ่งมีโอกาสในการตอบข้อสอบถูกมากกว่าผู้สอบอีกกลุ่มหนึ่งอย่างสม่ำเสมอกัน ในทุกระดับความสามารถ เมื่อพิจารณาโค้งคุณลักษณะข้อสอบของผู้สอบ 2 กลุ่ม จะพบว่า ไม่มีปฏิสัมพันธ์ระหว่างระดับความสามารถของผู้สอบกับการเป็นสมาชิกของกลุ่ม (Group membership) ดังภาพที่ 1

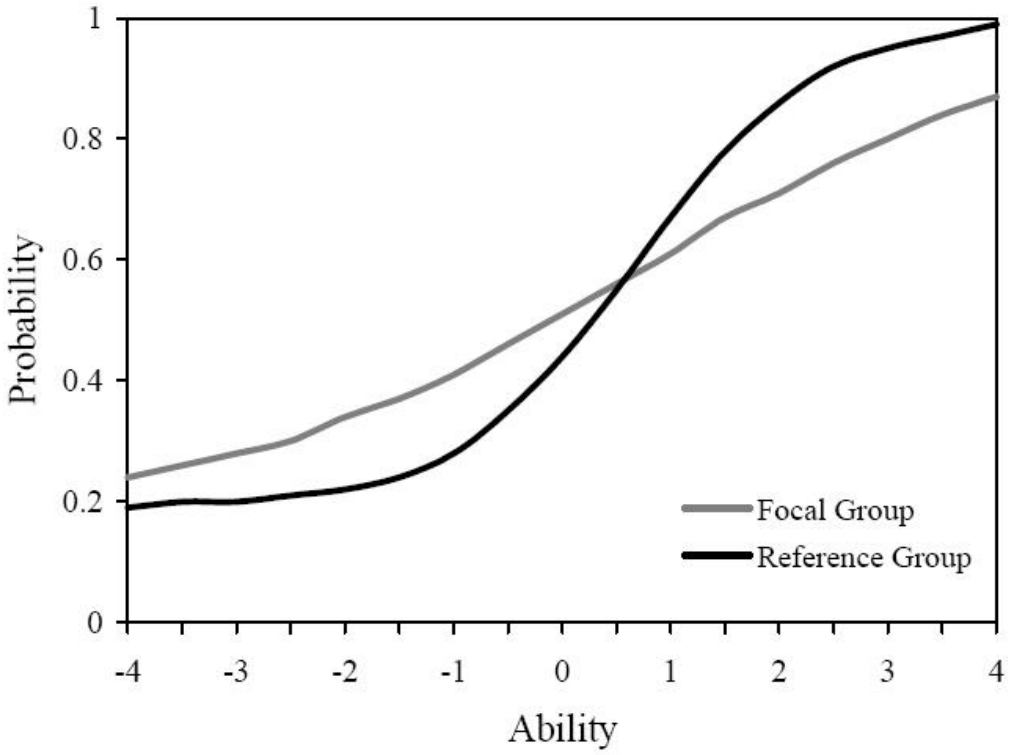


ภาพที่ 1 ข้อสอบทำหน้าที่ต่างกันแบบเอกรูป (Uniform DIF)

2) ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป (No-uniform differential item functioning) หมายถึง ข้อสอบที่ทำให้โอกาสในการตอบข้อสอบถูกของผู้สอบระหว่างกลุ่มแตกต่างกันอย่างไม่สม่ำเสมอกันในทุกระดับความสามารถ เมื่อพิจารณาโค้งคุณลักษณะข้อสอบของผู้สอบ 2 กลุ่ม พบว่ามีปฏิสัมพันธ์ร่วมกันระหว่างระดับความสามารถของผู้สอบ กับการเป็นสมาชิกของกลุ่ม เช่น ที่ระดับความสามารถหนึ่ง กลุ่มผู้สอบกลุ่ม R มีโอกาสในการตอบข้อสอบถูกมากกว่ากลุ่มผู้สอบกลุ่ม F แต่ที่ระดับความสามารถอีกระดับหนึ่งกลุ่มผู้สอบกลุ่ม F มีโอกาสในการตอบข้อสอบถูกมากกว่ากลุ่มผู้สอบกลุ่ม R ดังภาพที่ 2

ตามทฤษฎีการตอบสนองข้อสอบ (Item Response Theory : IRT) สามารถพิจารณา “ปฏิสัมพันธ์” ดังกล่าวได้จากความแตกต่างของพารามิเตอร์อำนาจจำแนกข้อสอบ ระหว่างผู้สอบกลุ่มย่อยสองกลุ่ม กล่าวคือ ถ้าข้อสอบทำหน้าที่ต่างกันแบบเอกรูป แล้วโค้งลักษณะข้อสอบ (Item characteristic curves : ICCs) ระหว่างกลุ่มผู้สอบย่อยสองกลุ่มจะขนานกัน หรือมีฟังก์ชันการตอบสนองข้อสอบ (Item response functions : IRFs) เหมือนกัน แต่ถ้าข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปแล้วโค้งลักษณะข้อสอบ (Item characteristic curves : ICCs) ระหว่างกลุ่มผู้สอบย่อยสองกลุ่มไม่ขนานกัน หรือมีฟังก์ชันการตอบสนองข้อสอบต่างกัน ดังนั้นความแตกต่างระหว่างโค้ง

ลักษณะข้อสอบทั้งสองแบบจะบ่งบอกถึงขนาด และทิศทางของข้อสอบที่ทำหน้าที่ต่างกัน ซึ่งสามารถคำนวณได้โดยใช้สูตรการคำนวณพื้นที่ของ Raju (1990)



ภาพที่ 2 ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป (No-uniform differential item functioning)

ข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูป สามารถจำแนกได้เป็น 2 ลักษณะ (Swaminathan & Rogers, 1990 อ้างถึงใน ศิริชัย กาญจนวาสี, 2550 : 119) ดังนี้

1) ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปโดยมีปฏิสัมพันธ์ไม่เป็นลำดับ (Disordinal interaction) เป็นการทำหน้าที่ต่างกันของข้อสอบสำหรับกลุ่มผู้สอบซึ่งเกิดขึ้น เมื่อโค้งลักษณะข้อสอบตัดกันระหว่างช่วงคะแนนความสามารถของผู้สอบหรือเรียกว่าข้อสอบทำหน้าที่ต่างกันแบบไม่มีทิศทาง (Non-Unidirectional differential item functioning)

2) ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป โดยมีปฏิสัมพันธ์เป็นลำดับ (Ordinal interaction) เป็นการทำหน้าที่ต่างกันสำหรับกลุ่มผู้สอบซึ่งเกิดขึ้น เมื่อโค้งลักษณะข้อสอบต่างกันอย่างไม่สม่ำเสมอ แต่ไม่ตัดกัน หรืออาจตัดกันข้างนอกช่วงความสามารถของผู้สอบตรงปลายสุดของช่วงความสามารถต่ำหรือสูง อาจเรียกข้อสอบนี้ว่า ข้อสอบทำหน้าที่ต่างกันแบบมีทิศทางเดียว (Unidirectional differential item functioning) ดังภาพที่ 3 (ศิริชัย กาญจนวาสี, 2550 : 117 -120)

1.4 หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (Differential item functioning detection) เป็นการเปรียบเทียบผลการตอบข้อสอบเป็นรายข้อระหว่างกลุ่มผู้สอบอย่างน้อยสองกลุ่ม ที่มีความสามารถหลัก (Primary ability) ที่มุ่งวัดเท่ากัน แต่คาดหวังว่าจะมีความได้เปรียบหรือเสียเปรียบกัน โดยอาศัยเกณฑ์จำแนก เช่น เพศ สีผิวเชื้อชาติ ศาสนา วัฒนธรรม ภูมิลำเนา เป็นต้น โดยกลุ่มหนึ่งถือว่าเป็นกลุ่มอ้างอิง (Reference group : R) ซึ่งคาดว่าน่าจะได้เปรียบในการตอบข้อสอบข้อนั้น หรือมีโอกาสตอบข้อสอบได้ถูกต้องมากกว่า ส่วนอีกกลุ่มหนึ่ง คือกลุ่มเปรียบเทียบ (Focal group : F) ซึ่งเป็นกลุ่มที่สนใจศึกษาและคาดว่าน่าจะเป็นกลุ่มที่เสียเปรียบในการตอบข้อสอบ

การเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบจำเป็นต้องมีการจับคู่ (Matching) ของผู้เข้าสอบตามความสามารถ ซึ่งเป็นเงื่อนไขสำคัญของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เกณฑ์การจับคู่ (Matching criteria) ที่นิยมใช้กันมี 2 วิธี ดังนี้

วิธีที่ 1 การใช้เกณฑ์ภายนอก (External criteria) สามารถนำมาใช้ได้ทั้งข้อสอบรายข้อและแบบสอบทั้งฉบับ โดยการใช้คะแนนจากแบบทดสอบอื่นเป็นเกณฑ์ภายนอกแล้วใช้เทคนิควิธีการวิเคราะห์ถดถอย (Regression analysis) เพื่อเปรียบเทียบเส้นกราฟความสัมพันธ์ระหว่างตัวแปรเกณฑ์กับตัวแปรทำนายระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ

หลักการนี้มีจุดมุ่งหมาย เพื่อสร้างสมการทำนายตัวแปรเกณฑ์ ซึ่งเป็นคะแนนของแบบทดสอบอื่นจากตัวแปรทำนายที่เป็นคะแนนรายข้อหรือคะแนนแบบทดสอบ ระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ ในการวิเคราะห์การทำหน้าที่ต่างกันของแบบทดสอบ จะใช้คะแนนรวมของแบบทดสอบทั้งฉบับเป็นตัวแปรทำนาย สำหรับตัวแปรเกณฑ์ที่ใช้เป็นเกณฑ์ภายนอก อาจใช้คะแนนรวมทั้งฉบับหรือเกรดเฉลี่ย หรือคะแนนจากงานที่เกี่ยวข้องของผู้สอบ (Cronbach, 1970) สมการทำนายสำหรับกลุ่มอ้างอิงและกลุ่มเปรียบเทียบแสดงได้ดังนี้

กลุ่มอ้างอิง $\quad Y_i = A_R + B_R X_i$

กลุ่มเปรียบเทียบ $\quad Y_i = A_F + B_F X_i$

เมื่อ Y_{I} = คะแนนของตัวแปรเกณฑ์ภายนอก

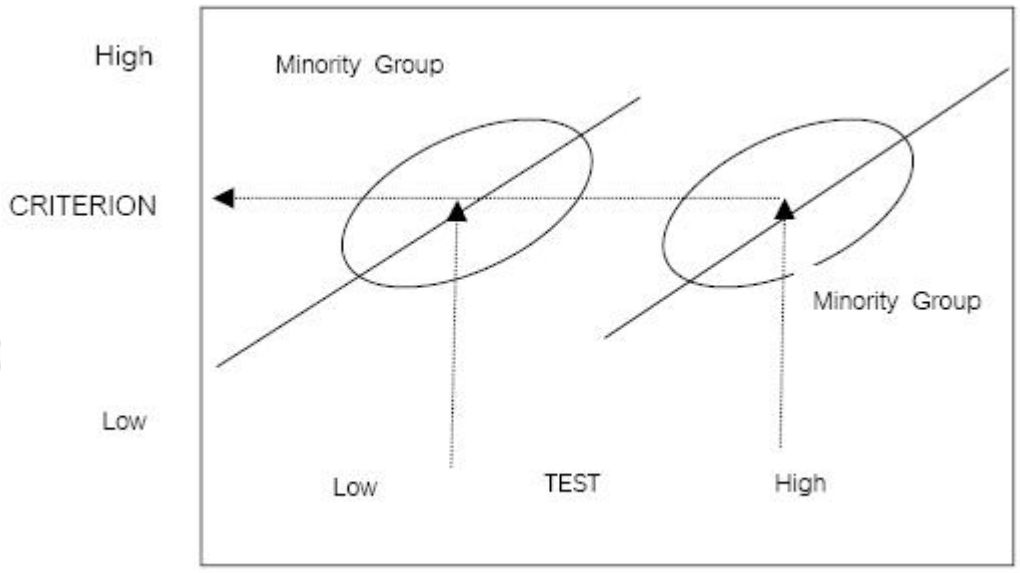
X_i = คะแนนของตัวแปรทำนาย

A = ค่าคงที่หรือจุดตัดแกน y (Intercept)

B = ค่าความชัน (Slope)

จากฟังก์ชันการทำนายทั้ง 2 สมการ สามารถเปรียบเทียบค่าตัดแกน (A) และค่าความชัน (B) ของเส้นกราฟระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบได้ ถ้าเส้นกราฟดังกล่าวมีค่าความชันหรือค่าตัดแกนแตกต่าง สำหรับข้อสอบใด แสดงว่าข้อสอบหรือแบบทดสอบนั้น มีการทำหน้าที่ต่างกัน โดยเข้าข้างกลุ่มผู้สอบที่มีค่าตัดแกนหรือค่าความชันที่สูงกว่า

การใช้เกณฑ์ภายนอกมีข้อดี คือเกณฑ์ที่ใช้ความเป็นอิสระจากข้อสอบ และแบบทดสอบที่ต้องการตรวจสอบ แต่มีจุดอ่อนตรงที่ความเหมาะสมของเกณฑ์ที่จะนำมาใช้ ในทางปฏิบัติเป็นการยากที่จะหาเกณฑ์ภายนอกจากแบบทดสอบฉบับอื่นที่มีความตรงเชิงทำนาย และมีความยุติธรรมสำหรับกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ถ้าเกณฑ์ภายนอกขาดคุณสมบัติดังกล่าว จะทำให้ผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบหรือแบบทดสอบขาดความแม่นยำ และขาดความสมบูรณ์



ภาพที่ 3 การทำหน้าที่ต่างกันของข้อสอบ เมื่อใช้วิธีการภายนอกตรวจสอบ (Camilli and Larried, 1994 อ้างถึงใน สุมาลี แกล้วทนงค์, 2547 : 18)

วิธีที่ 2 โดยใช้เกณฑ์ภายใน (Internal criteria) ในการวิเคราะห์การทำหน้าที่ต่างกันโดยใช้เกณฑ์ภายในเป็นการนำวิธีการทางสถิติมาตรวจสอบการทำหน้าที่ต่างกันของข้อสอบหรือแบบทดสอบเน้นการพิจารณาจากโครงสร้างภายในของแบบทดสอบเป็นหลัก ด้วยการวิเคราะห์ผลจากการตอบข้อสอบและความสามารถหรือคะแนนจริงของผู้เข้าสอบที่ได้จากแบบทดสอบฉบับนั้น เพื่อนำมาเปรียบเทียบระหว่างผู้เข้าสอบจากกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ ที่มีความสามารถหรือคะแนนจริงเท่ากัน ว่าจะมีผลการตอบหรือโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกันหรือไม่

เพื่อบ่งชี้ถึงการทำหน้าที่ต่างกันของข้อสอบ การคิดวิเคราะห์ในลักษณะนี้นิยมใช้ค่าสถิติต่าง ๆ เป็นตัวบ่งชี้การทำหน้าที่ต่างกันของข้อสอบค่าสถิติทดสอบที่นิยมมาใช้พอสรุปได้ ดังนี้

1. การทดสอบปฏิสัมพันธ์ (Interaction) ในระยะแรกของการศึกษาความลำเอียงของข้อสอบมีการใช้สถิติทดสอบเอฟ (F-test) ในการวิเคราะห์ความแปรปรวน (Analysis of variance : ANOVA) ทดสอบปฏิสัมพันธ์ระหว่างกลุ่มผู้เข้าสอบกับข้อสอบ ถ้าการทดสอบมีนัยสำคัญ แสดงว่า ข้อสอบมีการทำหน้าที่ต่างกันต่อจากนั้นจึงวิเคราะห์ต่อด้วยวิธี Post Hoc เพื่อระบุข้อสอบที่มีผลต่อการเกิด ปฏิสัมพันธ์ ซึ่งเป็นข้อที่ทำหน้าที่ต่างกัน

วิธีการนี้มีข้อดี คือ สามารถศึกษาผู้เข้าสอบหลาย ๆ กลุ่มได้ จุดอ่อนในเรื่องการทำให้กลุ่มผู้เข้าสอบต่าง ๆ มีความสามารถที่ทัดเทียมกัน ขนาดกลุ่มตัวอย่างของกลุ่มผู้เข้าสอบแต่ละกลุ่มและอัตราความคาดเคลื่อนประเภทที่ 1 จะสูงขึ้น ถ้าจำนวนข้อสอบเพิ่มมากขึ้น

2. การวัดความเบี่ยงเบนสัมพัทธ์ (Relative deviation) เป็นการคำนวณค่าความยากของข้อสอบ เมื่อคำนวณแยกระหว่างกลุ่มผู้เข้าสอบ และแปลงให้เป็นค่าความยากมาตรฐาน สามารถนำมาพล็อตกราฟเปรียบเทียบเป็นรายข้อ ถ้าข้อสอบข้อใดเบี่ยงเบนไปจากแกนหลักที่คาดหมายหรือเบี่ยงเบนเกินกว่าความคาดเคลื่อนมาตรฐานของค่าความยากที่กำหนด แสดงถึงการทำหน้าที่ต่างกันของข้อสอบ ทั้งนี้ยังสามารถคำนวณค่าสหสัมพันธ์เข้าใกล้ 1.00 แสดงว่า ค่าความยากสัมพันธ์ของข้อสอบที่ค่าใกล้เคียงกันระหว่างกลุ่ม ดังนั้นแบบทดสอบวัดคุณลักษณะคล้ายกันกับระหว่างกลุ่ม

วิธีการนี้มีข้อดีและข้อเสียคล้ายกับการทดสอบปฏิสัมพันธ์ นอกจากนี้ ค่าความยากของข้อสอบ (p) มิใช่ตัวแทนของค่าความยากที่แท้จริงของข้อสอบ และอาจได้รับอิทธิพลจากตัวแปรแทรกซ้อนอื่นเช่น ค่าอำนาจจำแนก และความสามารถของผู้เข้าสอบ

3. การเปรียบเทียบน้ำหนักองค์ประกอบ (Factor loading) ในการวิเคราะห์องค์ประกอบ (Factor analysis) เป็นเทคนิคทางสถิติที่นิยมใช้ในการตรวจสอบความตรงเชิงโครงสร้าง (Construct validity) เมื่อนำการวิเคราะห์องค์ประกอบมาใช้ในการวิเคราะห์โครงสร้างของแบบทดสอบแยกตามกลุ่มผู้เข้าสอบ ความไม่สอดคล้องกันระหว่างน้ำหนักองค์ประกอบบนคุณลักษณะสำคัญในสิ่งที่มุ่งวัดหรือความแตกต่างของค่าเฉลี่ยคะแนนองค์ประกอบ (Factor score) ระหว่างกลุ่มผู้เข้าสอบย่อมสะท้อนการทำหน้าที่ต่างกันของข้อสอบและแบบทดสอบ

ในการใช้เทคนิคการวิเคราะห์องค์ประกอบเชิงสำรวจ (Exploratory factor analysis : EFA) สำหรับศึกษาการทำหน้าที่ต่างกันของข้อสอบ มีจุดอ่อนในเรื่องความไม่สอดคล้องระหว่างน้ำหนักองค์ประกอบอาจเกิดจากความแตกต่างของความสามารถระหว่างกลุ่มก็ได้ แนวทางที่เหมาะสมจึงควรใช้เทคนิคการวิเคราะห์องค์ประกอบเชิงยืนยัน (Confirmatory factor analysis : CFA)

นอกจากนี้ยังสามารถใช้ CFA สำหรับตรวจสอบความแตกต่างระหว่างกลุ่มในลักษณะความสามารถหลักหรือความสามารถรองได้อีก

4. การเปรียบเทียบโอกาสตอบข้อสอบถูกในการวิเคราะห์โอกาสตอบข้อสอบถูกของผู้สอบจากกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่มีความสามารถเท่ากัน เป็นแนวทางสำคัญที่นิยมใช้กันและเป็นที่ยอมรับในปัจจุบัน สำหรับตัวบ่งชี้การทำหน้าที่ต่างกันของข้อสอบ มีการคำนวณค่าสถิติ 2 แนวทาง ดังนี้

4.1 เปรียบเทียบค่าสัดส่วนหรือความน่าจะเป็นในการตอบถูกของผู้สอบต่างกลุ่มที่มีความสามารถเท่ากัน เช่น วิธีแมนเทล-แฮนส์เซล เป็นต้น

4.2 เปรียบเทียบค่าฟังก์ชันการตอบสนองหรือโค้งลักษณะข้อสอบระหว่างกลุ่มที่มีความสามารถเท่ากัน เป็นวิธีที่อยู่บนพื้นฐานของทฤษฎี IRT เช่น วิธีการวัดความแตกต่างพื้นที่หรือวิธีการวัดความแตกต่างของพารามิเตอร์ความยาก เป็นต้น

วิธีการนี้มีข้อดีที่สำคัญคือ การคำนวณค่าสถิติของข้อสอบมีความน่าเชื่อถือ มีกลไกควบคุมความสามารถของผู้เข้าสอบโดยการจับคู่กลุ่มความสามารถ เพื่อเปรียบเทียบ ณ ตำแหน่งต่าง ๆ ที่มีความสามารถระดับเท่ากัน จึงเป็นวิธีการที่ยอมรับกันโดยทั่วไป แต่ก็ยังมีข้อจำกัดในด้านการสลับซับซ้อนของแนวคิดพื้นฐาน และจำต้องใช้โปรแกรมคอมพิวเตอร์ โดยเฉพาะในการวิเคราะห์เท่านั้น (ศิริชัย กาญจนวาสี, 2550 : 120 – 123)

1.5 เกณฑ์และวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

เกณฑ์การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF detection) จำแนกได้เป็น 2 ลักษณะ ดังนี้

1. ใช้เกณฑ์การให้คะแนนของข้อสอบ จะแบ่งออกเป็น 2 กลุ่ม คือกลุ่มวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนแบบทวิภาค (Dichotomous DIF methods) คือการให้คะแนน 0 – 1 และกลุ่มวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบพหุวิภาค (polytomous DIF methods) คือการให้คะแนนแบบหลายค่า และวิธีการตรวจสอบการทำหน้าที่ต่างกันทั้งสองแบบนี้ยังสามารถจำแนกได้ 2 มิติ ได้แก่ มิติลักษณะของตัวแปรเกณฑ์ซึ่งแบ่งเป็นกลุ่มที่ใช้คะแนนสังเกตได้ (Observed score) และกลุ่มวิธีที่ใช้คะแนนที่สังเกตไม่ได้หรือคะแนนของตัวแปรแฝง (Latent variable) และมิติลักษณะของสถิติวิเคราะห์ ซึ่งแบ่งเป็นกลุ่มที่ใช้สถิติพาราเมตริก (Parametric approach) และกลุ่มวิธีที่ใช้สถิตินันพาราเมตริก (Non-parametric approach) ดังรายละเอียดต่อไปนี้

1) วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนแบบทวิภาค ได้แก่

1.1 กลุ่มที่ใช้คะแนนที่สังเกตได้ วิธีการนี้มักวิเคราะห์ตามทฤษฎีการทดสอบแบบดั้งเดิม (Classical test theory : CTT) หรือกลุ่มที่ไม่ใช้ทฤษฎีการตอบสนองข้อสอบ (non-IRT approach) โดยใช้คะแนนรวมของผู้สอบเป็นเกณฑ์การจับกลุ่มผู้สอบ วิธีการตรวจสอบที่สำคัญในกลุ่มนี้ ได้แก่ การวิเคราะห์ความแปรปรวน (Analysis of variance : ANOVA) การวิเคราะห์การถดถอยโลจิสติก (Logistic regression : LR) วิธีแปลงค่าความยากของข้อสอบ (Transformed item difficulty : TID) วิธีแมนเทล-แฮนด์เซล (Mantel-Haenszel : MH)

1.2 กลุ่มที่ใช้คุณลักษณะหรือตัวแปรแฝง วิธีในกลุ่มนี้ใช้คุณลักษณะหรือตัวแปรแฝงซึ่งวิเคราะห์บนพื้นฐานของทฤษฎีการตอบสนองข้อสอบ (IRT) สำหรับใช้เป็นเกณฑ์จับคู่ กลุ่มผู้สอบ วิธีการตรวจสอบที่สำคัญในกลุ่มนี้ คือ วิธีวัดพื้นที่ความแตกต่างระหว่างโค้งการตอบสนองข้อสอบ (IRT-D^2) วิธีอัตราส่วนไลด์ลิฮูด ลอกลิเนียร์ (Loglinear IRT Likelihood Ratio) วิธีไคสแควร์ของลอร์ด (Lord's χ^2) และวิธีซิปเทสท์ (SIBTEST)

2) วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบพหุวิภาค ได้แก่

2.1 กลุ่มที่ใช้คะแนนสังเกตได้ ประกอบด้วย การวิเคราะห์ความแปรปรวน (ANOVA) การวิเคราะห์การถดถอยโลจิสติกพหุวิภาค (Polytomous logistic regression) วิธีดัชนีมาตรฐานพหุวิภาค (Polytomous Standardization) วิธีแมนเทล-แฮนด์เซลทั่วไป (General Mantel-Haenszel : GMH)

2.2 กลุ่มที่ใช้คุณลักษณะแฝง ได้แก่ วิธีอัตราส่วนไลค์ลิฮูดในรูปทั่วไป (General IRT Likelihood Ratio) วิธีการให้คะแนนบางส่วน (Partial Credit Model) วิธีซิปเทสท์พหุวิภาค (Polytomous SIBTEST) และวิธีการให้คะแนนบางส่วนทั่วไป (Generalized partial credit model : GPCM)

3) ใช้ข้อตกลงเบื้องต้นของโมเดลเป็นเกณฑ์ สามารถแบ่งออกเป็น 2 รูปแบบ คือ

3.1 รูปแบบพาราเมตริก (Parametric form) ซึ่งวิเคราะห์การทำหน้าที่ต่างกันโดยมีข้อตกลงเบื้องต้นของโมเดลสำหรับอธิบายความสัมพันธ์ระหว่างคะแนนของข้อสอบและการจับคู่ตัวแปร

3.2 รูปแบบนันพาราเมตริก (Non-parametric form) ซึ่งจะวิเคราะห์ดัชนีการทำหน้าที่ต่างกันของข้อสอบโดยจะไม่มีข้อตกลงเบื้องต้นของโมเดลและการจับคู่ตัวแปรดังกล่าว (สุมาลี แกล้วทนงค์, 2547 : 16-17)

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

จากการศึกษาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ตามแนวคิดทฤษฎีการทดสอบแบบดั้งเดิมและแนวคิดตามทฤษฎีการตอบสนองข้อสอบ วิธีการที่นักวัดผลยอมรับและให้การสนใจในปัจจุบัน สรุปได้ดังนี้

1. วิธีแมนเทล-แฮนเซล (Mantel – Haenzel : MH)

วิธีแมนเทล-แฮนส์เซล (Mantel – Haenzel : MH) เป็นเทคนิคที่แมนเทล-แฮนส์เซลได้เสนอขึ้นใช้ตั้งแต่ ปี 1959 แต่ฮอลแลนด์และเทเยอร์ (Holland & Thayer, 1988) ได้นำเสนอมาเพื่อใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในการบริการการทดสอบทางการศึกษา ของสหรัฐอเมริกาและเป็นที่ยอมรับของนักวิจัยอย่างมากมายเพราะเป็นวิธีที่ง่ายในการนำไปใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ไม่สลับซับซ้อน ปฏิบัติได้ง่ายกว่าวิธีการตอบสนองข้อสอบ เสียค่าจ่ายน้อย สามารถเข้าใจง่าย เทคนิคเชื่อถือได้ มีการทดสอบนัยสำคัญ และการแปรผลไม่ยุ่งยาก มีการทดสอบสถิติแบบนันพาราเมตริก ซึ่งไม่จำเป็นต้องใช้โมเดลการประมาณค่า เทคนิคนี้เป็นวิธีที่มีความคล้ายกับวิธีไคสแควร์ที่เสนอโดยชูเนแมน (Scheuneman) มาราซูโล (Marsscuilo) และสลาชเตอร์(Alaughter)

หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบของวิธีแมนเทล-แฮนส์เซล เป็นการเปรียบเทียบผลการตอบข้อสอบของกลุ่มผู้สอบ 2 กลุ่ม กลุ่มหนึ่งเรียกว่า กลุ่มเปรียบเทียบ (Focal group) ซึ่งเป็นกลุ่มที่สนใจศึกษา ถ้าข้อสอบทำหน้าที่ต่างกันคาดว่าผู้สอบกลุ่มนี้จะเสียเปรียบในการตอบข้อสอบ ส่วนผู้สอบอีกกลุ่มเรียกว่า กลุ่มอ้างอิง (Reference group) ซึ่งเป็นกลุ่มผู้สอบที่ใช้เป็นมาตรฐานในการเปรียบเทียบผลการตอบข้อสอบกับกลุ่มแรก ถ้าข้อสอบทำหน้าที่ต่างกันแล้วคาดว่าผู้สอบกลุ่มนี้จะได้เปรียบในการตอบข้อสอบ ในการเปรียบเทียบผลการตอบข้อสอบจะเปรียบเทียบทุกระดับความสามารถของผู้สอบกลุ่มย่อยทั้งสองกลุ่มที่ระดับความสามารถเท่ากัน ในทางปฏิบัติมักจะใช้คะแนนรวมของแบบสอบเป็นเกณฑ์ในการจับคู่กลุ่มผู้สอบ

การตรวจสอบด้วยวิธีการนี้ทำได้โดยแยกวิเคราะห์ข้อสอบเป็นรายข้อ เมื่อจับคู่กลุ่มผู้สอบแล้วจะนำข้อมูลผลการตอบข้อสอบ ระหว่างผู้สอบทั้ง 2 กลุ่ม มาจัดลงในตารางการณ์จรแบบ 2 x 2 (กลุ่มผู้สอบ 2 กลุ่ม x ผลการตอบ 2 แบบ) โดยที่ตารางการณ์จร 1 ตาราง แทนคะแนนรวม 1 ระดับ ดังนั้นถ้ามีคะแนนรวม K ระดับ จะต้องสร้างตารางการณ์จรแบบ 2 x 2 ทั้งหมด K ตาราง สำหรับตารางการณ์จรแบบ 2 x 2 ของข้อสอบแต่ละข้อที่มีคะแนนรวมระดับ j ดังตารางที่ 1

ตารางที่ 1 ผลการตอบข้อสอบข้อหนึ่งระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ที่มีคะแนนรวมอยู่ในช่วงคะแนน j

กลุ่ม	คะแนนผลของการตอบข้อสอบที่ศึกษา		
	ตอบถูก (1)	ตอบผิด (0)	รวม
กลุ่มอ้างอิง (R)	A_j	B_j	N_{Rj}
กลุ่มเปรียบเทียบ (F)	C_j	D_j	N_{Fj}
รวม	m_{1j}	m_{0j}	T_j

เมื่อ A_j แทน จำนวนผู้สอบในกลุ่มอ้างอิง ที่ระดับคะแนน j ซึ่งตอบข้อสอบถูก

B_j แทน จำนวนผู้สอบในกลุ่มอ้างอิง ที่ระดับคะแนน j ซึ่งตอบข้อสอบผิด

C_j แทน จำนวนผู้สอบในกลุ่มเปรียบเทียบที่ระดับคะแนน j ซึ่งตอบข้อสอบถูก

D_j แทน จำนวนผู้สอบในกลุ่มเปรียบเทียบที่ระดับคะแนน j ซึ่งตอบข้อสอบผิด

T_j แทน จำนวนผู้สอบทั้งหมดที่ระดับคะแนน j

m_{1j} แทน จำนวนผู้สอบทั้งหมดที่ระดับคะแนน j ที่ตอบข้อสอบถูก

m_{0j} แทน จำนวนผู้สอบทั้งหมดที่ระดับคะแนน j ที่ตอบข้อสอบผิด

N_{Rj} แทน จำนวนผู้สอบกลุ่มอ้างอิงที่ระดับคะแนน j

N_{Fj} แทน จำนวนผู้สอบกลุ่มเปรียบเทียบที่มีคะแนนรวม j

จากนั้นจึงนำคะแนนผลการตอบข้อสอบจากตารางที่ 1 มาคำนวณสัดส่วนของผลการตอบข้อสอบที่ตอบถูกและผิด ดังตารางที่

ตารางที่ 2 สัดส่วนของผู้สอบระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ที่ระดับคะแนน j

กลุ่ม	คะแนนของผลการตอบข้อสอบที่ศึกษา		
	1	0	รวม
กลุ่มอ้างอิง (R)	*pRj*	*qRj*	*nRJ*
กลุ่มเปรียบเทียบ (F)	*pFj*	*qFj*	*nFJ*

โดยที่ pRj คือ สัดส่วนของผู้ตอบกลุ่มอ้างอิงที่ระดับคะแนน j ซึ่ง ตอบถูก

qRj คือ สัดส่วนของผู้ตอบกลุ่มอ้างอิงที่ระดับคะแนน j ซึ่ง ตอบผิด

pFj คือ สัดส่วนของผู้ตอบกลุ่มเปรียบเทียบที่ระดับคะแนน j ซึ่ง ตอบถูก

qFj คือ สัดส่วนของผู้ตอบกลุ่มเปรียบเทียบที่ระดับคะแนน j ซึ่ง ตอบผิด

สำหรับการทดสอบสมมติฐานของการทำหน้าที่ต่างกันของข้อสอบจะกำหนดสมมติฐานศูนย์ และสมมติฐานอื่น ดังนี้

$$H_0 : \frac{pRj}{qrj} = \frac{pFj}{qFj} \quad ; \quad j = 1,2,3,...,K$$

$$H_1 : \frac{pRj}{qRj} = \alpha \frac{pFj}{qFj} \qquad j = 1.,2,3,...,K \qquad ,\alpha > 1$$

สมมติฐานศูนย์เป็นสมมติฐานที่เป็นอิสระอย่างมีเงื่อนไขของสมาชิกในกลุ่มผู้สอบและคะแนนจากข้อสอบที่ศึกษา ดังนั้นคะแนนที่ได้จากตารางที่ 2 และภายใต้สมมติฐานศูนย์สามารถสรุปเป็นค่าคาดหวัง (Expected values) ในแต่ละเซลล์ดังต่อไปนี้

$$E(A_j) = \frac{nR_j m_{1j}}{T_j}$$

$$E(B_j) = \frac{nR_j m_{0j}}{T_j}$$

$$E(C_j) = \frac{nR_j m_{1j}}{T_j}$$

$$E(D_j) = \frac{nR_j m_{0j}}{T_j}$$

พารามิเตอร์ α ภายใต้สมมติฐานอื่น เรียกว่า อัตราส่วนต่อร่วม (Common odds ratio) ซึ่งคำนวณได้จาก

$$\alpha = \frac{\frac{pRj}{qRj}}{\frac{pFj}{qFj}} = \frac{pRjqFj}{qRjpFj}$$

ค่า $\alpha = 1$ แสดงว่าโอกาสของการตอบข้อสอบถูกระหว่างกลุ่มผู้สอบทั้งสองกลุ่มมีค่าเท่ากัน ถ้า $\alpha > 1$ แสดงว่าผู้สอบกลุ่มอ้างอิงมีโอกาสของการตอบข้อสอบถูกมากกว่ากลุ่มเปรียบเทียบ และถ้า $\alpha < 1$ แสดงว่าผู้สอบกลุ่มเปรียบเทียบมีโอกาสของการตอบข้อสอบถูกมากกว่ากลุ่มอ้างอิงโดยแมนเทล-แฮนส์เซลได้ประมาณค่า α จากตารางไขว้แบบ 2x2 ดังนี้

$$\alpha_{MH} = \frac{\sum A_j D_j / N_j}{\sum B_j C_j / N_j}$$

ค่า α_{MH} เป็นค่าประมาณอิทธิพลของการทำหน้าที่ต่างกันของข้อสอบ (DIF effect size) จะมีค่าระหว่าง 0 และ ∞ นอกจากนั้น Education testing service's ได้เสนอให้แปลงค่า α_{MH} ให้เป็นคะแนนมาตรฐานในรูปเดลต้าที่มีค่าเฉลี่ยเป็น 0 โดเรนและฮอลแลนด์ (Dorans and Holland, 1993) เรียกค่าที่แปลงนี้ว่า MH DIF ที่มีสมการดังนี้

$$MH_{D-DTIF} = \frac{-4}{1.7} In(\alpha_{MH}) = -2.35 In(\alpha_{MH})$$

ค่า MH_{D-DIF} ดังกล่าวสามารถนำไปพิจารณาค่าความยากของข้อสอบ คือถ้า มีค่า MH_{D-DIF} มีค่าเป็นศูนย์ แสดงว่าข้อสอบของแต่ละกลุ่มยากเท่ากัน ถ้า MH_{D-DIF} มีค่าเป็นลบ แสดงว่าข้อสอบยากสำหรับกลุ่มเปรียบเทียบมากกว่ากลุ่มอ้างอิง และถ้า MH_{D-DIF} มีค่าเป็นบวก แสดงว่าข้อสอบยากสำหรับกลุ่มอ้างอิงมากกว่ากลุ่มเปรียบเทียบ

ในการทดสอบนัยสำคัญของสมมติฐาน จะนำค่า α_{MH} สถิติแมนเทล-แฮนส์เซลไคสแควร์ที่ระดับชั้นความเป็นอิสระเท่ากับ 1 (df=1) โดยนำค่า α_{MH} ไปเปรียบเทียบกับ 1 คำนวณโดยสูตร

$$MH - \chi^2 = \frac{\left(\left|\sum A_j - \sum E(A_j)\right| - 0.5\right)^2}{\sum Var(A_j)}$$

โดยที่ $E(A_j) = \frac{N_{Rj} m_{1j}}{T_j}$

$$Var(A_j) = \frac{N_{Rj} N_{Fj} m_{1j} m_{0j}}{T_j^2 (T_j - 1)}$$

เมื่อ $E(A_j)$ แทน ค่าคาดหวังของจำนวนผู้สอบกลุ่มอ้างอิงที่ระดับคะแนน j ซึ่งตอบข้อสอบถูก

$Var(A_j)$ แทน ค่าความแปรปรวนของจำนวนผู้สอบกลุ่มอ้างอิงที่ระดับคะแนน j ซึ่ง ตอบข้อสอบถูก

สำหรับ ค่า α_{MH} แปรผลการทดสอบดังนี้

1. ค่า $\alpha_{MH} = 1$ หรือไม่แตกต่างจาก 1 อย่างมีนัยสำคัญ แสดงว่าข้อสอบทำหน้าที่ไม่ต่างกัน (No DIF)

2. ค่า $\alpha_{MH} > 1$ แสดงว่าข้อสอบนั้นทำหน้าที่ต่างกันระหว่างกลุ่มโดยเข้าข้างกลุ่มอ้างอิง (DIF)

3. ค่า $\alpha_{MH} < 1$ แสดงว่าข้อสอบนั้นทำหน้าที่ต่างกันระหว่างกลุ่มโดยเข้าข้างกลุ่มเปรียบเทียบ (DIF)

ส่วนค่า MH DIF แปลผลการทดสอบได้ดังนี้

1) ค่า MH DIF เท่ากับ 0 หรือไม่แตกต่างจาก 0 อย่างมีนัยสำคัญทางสถิติ แสดงว่าข้อสอบนั้นทำหน้าที่ไม่แตกต่างกันระหว่างกลุ่ม (no DIF)

2) ค่า MH DIF แตกต่างจาก 0 อย่างมีนัยสำคัญทางสถิติ และมีค่าเป็นบวก (positive) แสดงว่าข้อสอบนั้นทำหน้าที่แตกต่างกันระหว่างกลุ่ม โดยจะเข้าข้างกลุ่มเปรียบเทียบ (F)

3) ค่า MH DIF แตกต่างจาก 0 อย่างมีนัยสำคัญทางสถิติ และมีค่าเป็นลบ (negative) แสดงว่าข้อสอบนั้นทำหน้ามี่แตกต่างกันระหว่างกลุ่ม โดยจะเข้าข้างกลุ่มอ้างอิง (R)

นอกจากนี้ขนาดของ | MH DIF | สามารถนำไปใช้แปลผลถึงระดับของการทำหน้าที่แตกต่างกันของข้อสอบได้ ถ้า 0 < |MH DIF | < 1.00 แสดงว่าข้อสอบทำหน้าที่แตกต่างกันเล็กน้อย ถ้า 1.00 $\leq$ | MH DIF | $\leq$1.50 แสดงว่าข้อสอบทำหน้าที่แตกต่างกันปานกลาง แต่ถ้า I MH DIF I > 1.50แสดงว่าข้อสอบทำหน้าที่แตกต่างกันมาก (Zieky. 1993)การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยวิธี MH นี้ ใช้คะแนนรวมของแบบสอบเป็นเกณฑ์การจับคู่ ซึ่งมีจุดอ่อนในด้านความไม่เป็นอิสระของคะแนนรวม กับคะแนนรายข้อที่ทำการศึกษาฮอลแลนด์และเทเยอร์ (Holland ; & Thayer. 1988) ได้เสนอวิธีแก้จุดอ่อนดังกล่าว เพื่อทำให้เกณฑ์การจับคู่ผู้สอบระหว่างกลุ่มมีความบริสุทธิ์ยิ่งขึ้น โดยใช้วิธีการ 2 ขั้นตอน ดังนี้

ขั้นตอนแรก นำคะแนนรวมของแบบสอบทั้งฉบับเป็นเกณฑ์การจับคู่ผู้สอบระหว่างกลุ่มย่อย 2 กลุ่ม แล้ววิเคราะห์การทำหน้าที่ต่างกันของข้อสอบเป็นรายข้อ เมื่อพบว่าข้อสอบข้อใดทำหน้าที่ต่างกันให้นำคะแนนของข้อสอบข้อนั้นออกจากคะแนนรวมของผู้สอบแต่ละคน

ขั้นตอนที่สอง ใช้คะแนนรวมของแบบสอบที่นำเอาคะแนนข้อสอบที่ทำหน้าที่ต่างกัน ซึ่งตรวจพบในขั้นตอนแรกออกไป เพื่อใช้เป็นเกณฑ์การจับคู่ แล้ววิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ ซ้ำอีกครั้งหนึ่ง สำหรับนำไปใช้สรุปผลการตรวจสอบ

เท่าที่ผ่านมามีผู้นำวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ วิธี Mantel-Haenszel มาศึกษา เช่นคลอสเซอร์และคนอื่น ๆ (Clauser; & others, 1991) ใช้วิธี Mantel-Haenszel วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบอย่างสม่ำเสมอ พบว่าวิธี Mantel-Haenszel จะใช้ได้ผลดีในกรณีที่ข้อสอบมีค่าอำนาจจำแนกสูง แต่จะไม่สามารถวิเคราะห์ข้อสอบที่มีค่าความยากมากได้ ซึ่งสอดคล้องกับเมเซอร์และ คนอื่นๆ (Mazor; & other, 1991 อ้างถึงใน จันทนา เปรมฤดีปรีชาชาญ, 2551 : 28)) ที่พบว่าข้อสอบ ที่ทำหน้าที่ต่างกันมักเป็นข้อสอบที่ยาก บากิและเฟอรารา (Baghi ; & Ferrara, 1990) พบว่า เมื่อใช้กลุ่มตัวอย่างขนาด 750 คนขึ้นไป วิธี MH ใช้แทนวิธีทฤษฎีการตอบข้อสอบ 3 พารามิเตอร์ได้ สวามินาธานและโรเจอร์ (Swaminathan; & Rogers, 1990) ได้ศึกษาเอาข้อมูลจำลองพบว่าวิธี Mantel-Haenszel วิเคราะห์ได้ดีกว่าการถดถอยแบบโลจีสเล็กน้อยโดยตรวจค้นได้ถูกต้องร้อยละ 75 กรณีใช้กลุ่มตัวอย่าง 250 คนและ ตรวจค้นได้ถูกต้องร้อยละ 100 กรณีกลุ่มตัวอย่าง 500 คน กรณีที่การทำหน้าที่ต่างกันของข้อสอบอย่างสม่ำเสมอ และกรณีที่ไม่สม่ำเสมอที่ติดกันปลายข้างใดข้างหนึ่ง แต่วิธี Mantel-Haenszel มีค่าใช้จ่ายง่ายน้อยกว่าวิธีโลจีสติกประมาณ 3- 4 เท่า แฮมเบิลตันและคนอื่นๆ (Hambleton and others, 1986) พบว่า Mantel-Haenszel ให้ค่าใกล้เคียงกับทฤษฎีการตอบข้อสอบ ทั้งที่ใช้ค่าความแตกต่างของค่าเฉลี่ยกำลังสองและการตรวจสอบความแตกต่างของพื้นที่รวมใต้โค้ง แต่วิธี Mantel-Haenszel มีค่าใช้จ่ายต่ำกว่าและใช้เวลาน้อยกว่า ริสเซน สไตน์เบอร์ก และวายเนอร์ (Thissen , Steinberg ; & Wainer,1988 อ้างถึงใน จันทนา เปรมฤดีปรีชาชาญ, 2551 : 28) พบว่าวิธี Mantel-Haenszel ให้ผลการวิเคราะห์คล้ายกับวิธีทฤษฎีการตอบข้อสอบแบบการถดถอยเชิงเส้น (IRT-LR) และอาจใช้วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบก่อนใช้วิธีทฤษฎีการตอบข้อสอบแบบการถดถอยเชิงเส้น (IRT-LR)

2. วิธีถดถอยโลจิสติก (Logistic regression : LR)

Swaminathan & Rogers (1990) ได้พัฒนาวิธีถดถอยโลจิสติก เพื่อใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนแบบสองค่า (Dichotomous) วิธีการนี้มีแนวคิดมาจากวิธีตารางการณ์จร โดยดัดแปลงมาจากวิธีแมนเทล-แฮนส์เซล และเทเยอร์ (Holland; & Thayer, 1988) และวิธีลอกลิเนียร์ของเมลเลนเบอร์ก (Mellenberg, 1982) หลักการตรวจสอบด้วยวิธีถดถอยโลจิสติกจะใช้โมเดลการถดถอยโลจิสติกทำนายโอกาสของผลการตอบข้อสอบถูก โมเดลดังกล่าวใช้ตัวแปรความสามารถแบบต่อเนื่อง ซึ่งมีเทอมที่ใช้คำนวณปฏิสัมพันธ์ระหว่างการเป็นสมาชิกของ

กลุ่มผู้สอบกับระดับความสามารถ จึงทำให้สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ทั้งแบบเอกรูป (Uniform DIF) และแบบอเนกรูป (Nonuniform DIF) นอกจากนี้ยังสามารถนำไปประยุกต์กับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีผู้สอบหลายกลุ่ม และการให้คะแนนข้อสอบแบบพหุวิภาค (Polytomous) (Miller & Spray, 1993 อ้างถึงใน ปิยะทิพย์ ตินวร, 2549 : 42)

การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธีถดถอยโลจิสติกจะใช้สมการมาตรฐานของโมเดลการถดถอยโลจิสติก คำนวณผลการตอบข้อสอบถูก ดังนี้ (Swaminathan; & Rogers, 1990 อ้างถึงใน ปิยะทิพย์ ตินวร, 2549 : 42)

$$P(U_{ij}-1/\theta)=\frac{\exp^{(\beta_{0j}+\beta_{1j}\theta_{ij})}}{1+\exp^{(\beta_0+\beta_{1j}\theta_{ij})}} \qquad \text{i = 1,2,3,…,n ; j=1.2}$$

เมื่อ U_{ij} แทน ผลการตอบข้อสอบของผู้เข้าสอบคนที่ 1 ในกลุ่ม j

θ_{ij} แทน ค่าความสามารถที่สังเกตได้ของผู้เข้าสอบคนที่ i ในกลุ่ม j

β_{0j} แทน ค่าพารามิเตอร์จุดตัด (Intercept parameter)

β_{1j} แทน ค่าพารามิเตอร์ความชันสำหรับกลุ่ม j (Slope parameter)

จากโมเดลดังกล่าวถ้า $\beta_{01}=\beta_{02}$ และ $\beta_{11}=\beta_{12}$ แล้ว ฟังก์ชันการถดถอยโลจิสติกของผู้เข้าสอบ 2 กลุ่มเหมือนกัน แสดงว่าข้อสอบทำหน้าที่ไม่ต่างกัน (No DIF) ถ้า $\beta_{11}=\beta_{12}$ แต่ $\beta_{01}\neq\beta_{02}$ แล้ว ฟังก์ชันการถดถอยของผู้เข้าสอบสองกลุ่มขนานกันแต่ไม่ทับกันแสดงว่าข้อสอบทำหน้าที่ต่างกันแบบเอกรูป (Uniform DIF) และถ้า $\beta_{01}=\beta_{02}$ แต่ $\beta_{11}\neq\beta_{12}$ แล้วฟังก์ชันการถดถอยโลจิสติกของผู้เข้าสอบไม่ขนานกัน แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป (Nonuniform DIF) นอกจากนี้โมเดลการถดถอยโลจิสติกดังกล่าวสามารถเปลี่ยนเป็นโมเดลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป และแบบอเนกรูป ดังนี้

$$P(U_{ij}=1/\theta_{ij})=\frac{\exp^{Z_{ij}}}{1+\exp^{Z_{ij}}}$$

โดยที่ $Z_{ij}=\tau_0+\tau_1\theta_1+\tau_2G_j+\tau_3(\theta_{ij}G_j)$

เมื่อ $P(U_{ij}=1/\theta_{ij})$ แทน โอกาสในการตอบข้อสอบถูกของผู้เข้าสอบคนที่ i ในกลุ่ม j

θ_{ij} แทน ความสามารถของผู้เข้าสอบคนที่ i กลุ่ม j

G_j แทน สมาชิกผู้เข้าสอบในกลุ่ม j (โดยกำหนดให้ $G_j = 1$ สมาชิกกลุ่ม 1 หรือกลุ่มเปรียบเทียบ , $G_j = 2$ สมาชิกกลุ่ม 2 หรือกลุ่มอ้างอิง)

θ_{ij}, G_j แทน ปฏิสัมพันธ์ของตัวแปรอิสระ 2 ตัว คือ θ_{ij} กับ G_j

τ_0 แทน พารามิเตอร์จุดตัด

τ_1 แทน สัมประสิทธิ์ของความสามารถของผู้เข้าสอบ

τ_2 แทน ความแตกต่างระหว่างกลุ่มผู้เข้าสอบในการตอบข้อสอบถูก โดย $\tau_2 = \beta_{01} - \beta_{02}$

τ_3 แทน ปฏิสัมพันธ์ระหว่างกลุ่มผู้เข้าสอบกับระดับความสามารถผู้เข้าสอบ โดย $\tau_3 = \beta_{11} - \beta_{12}$

โมเดลการถดถอยโลจิสติกข้างต้น สามารถเปลี่ยนเป็นโมเดลเชิงเส้นในเมทริกซ์โลจิท (Logit Metric) ซึ่งจะอยู่ในรูป log ของอัตราส่วนของโอกาสในการตอบข้อสอบถูกต่อโอกาสในการตอบข้อสอบผิด ดังนี้

$$\log\left[\frac{P}{1-P}\right] = Z_{ij} = \tau_0 + \tau_1\theta_{ij} + \tau_2 G_j + \tau_3(\theta_{ij}G_j)$$

จากโมเดลดังกล่าว เทอม $\theta_{ij}G_j$ เป็นผลคูณของตัวแปรอิสระ θ_{ij} และ G_j

ในการตัดสินว่าข้อสอบทำหน้าที่ต่างกันแบบเอกรูปหรืออเนกรูปจะพิจารณาพารามิเตอร์ τ_2 และ τ_3 ดังนี้

ถ้า $\tau_2 \neq 0$ และ $\tau_3 = 0$ แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบเอกรูป

และ $\tau_3 \neq 0$ และ $\tau_2 = 0$ แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป

สำหรับการประมาณค่าพารามิเตอร์ตามโลจิสติก ของข้อสอบแต่ละข้อของโมเดล ใช้วิธีประมาณค่าด้วยวิธีความควรจะเป็นสูงสุด (Maximum likelihood estimation : MLE) ซึ่งเขียนในรูปฟังก์ชันได้ดังนี้

$$L(U_{ij}/\theta) = \prod_{i=1}^{n}\prod_{j=1}^{k} P(U_{ij})^{U_{ij}}\left[1 - P(U_{ij})\right]^{1-U_{ij}}$$

โดยที่ n และ k แทนขนาดกลุ่มตัวอย่างและความยาวของแบบทดสอบตามลำดับ สำหรับค่าประมาณของพารามิเตอร์โดยใช้วิธีความควรจะเป็นสูงสุด มีการแจกแจงแบบปกติของตัวแปรพหุในรูปเชิงเส้นกำกับ (Asymptotically multivariate normal) ซึ่งมีค่าเฉลี่ยของเวคเตอร์ และเมทริกซ์ความแปรปรวน ความแปรปรวนร่วมในรูป ในขณะที่ เป็นเมทริกซ์สารสนเทศกำหนดดังนี้

$$\sum^{-1} = E\left[\frac{\partial^2}{\partial\tau_r \partial\tau_s} \ln L\right] \quad ; \ r, \ s = 0, 1, 2, 3$$

เมื่อ E และ ln L แทนค่าความคาดหมายของเมทริกซ์และลอกการิทึมของฟังก์ชันความน่าจะเป็นสูงสุดตามลำดับ ดังนั้นการแจกแจงของการประมาณค่าพารามิเตอร์ด้วยวิธี MLE จะอยู่ในรูปดังนี้

$$\tau \sim N(\tau, \Sigma)$$

โดยที่ $\tau = [\tau_0, \tau_1, \tau_2, \tau_3]$ ส่วนความคลาดเคลื่อนมาตรฐานเชิงเส้นกำกับของค่าประมาณของ $\tau_s (S = 0,1,2,3)$ เมื่อ S เป็นสมาชิกแนวเส้นทแยงมุมของ Σ สามารถคำนวณได้จากสูตรดังนี้

$$SE(\hat{\tau}_s) = \sqrt{\Sigma^{ss}}$$

ในการทดสอบสมมติฐานของการทำหน้าที่ต่างกันของข้อสอบจะทดสอบสมาชิกของ τ_s ซึ่งสมมติฐานที่สนใจคือ $H_0 : \tau_2 = 0$ และ $H_0 : \tau_3 = 0$ สมมติฐานทั้งสองสามารถทดสอบพร้อม ๆ กันไป ดังนี้

$$H_0 : C_\tau = 0$$
$$H_1 : C_\tau \neq 0$$

โดยที่ C เป็นเมทริกซ์ขนาด 2x4 ดังนี้

$$C = \begin{vmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{vmatrix}$$

ส่วนการทดสอบนัยสำคัญของสมมติฐานจะใช้สถิติไค-สแควร์ที่ระดับชั้นความเป็นอิสระเท่ากับ 2 (*df=2*) ดังนี้

$$\chi^2 = \hat{\tau}'C'(C\Sigma C')^{-1}C\hat{\tau}'$$

ถ้า χ^2 มีค่ามากกว่า $\chi^2_{(\alpha,2)}$ แสดงว่าปฏิเสธสมมติฐานของข้อสอบที่ทำหน้าที่ไม่ต่างกัน (No DIF) นั่นคือ ข้อสอบทำหน้าที่ต่างกัน นั่นเอง

3. วิธีซิปเทสท์ (SIBTEST)

ชีลลีและสเตาท์ (Shealy; & Stout. 1993) ได้เสนอวิธี “Simultaneous item bias test” หรือเรียกสั้นๆ ว่า “วิธีซิปเทสท์” (SIBTEST) เพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ และแบบทดสอบ (Differential item / test functioning : DIF/DTF) ในข้อสอบที่ให้คะแนนสองค่า หลักการตรวจสอบด้วยวิธีซิปเทสท์มีแนวคิดคล้ายกับวิธีการทำให้เป็นมาตรฐาน(Standardization : STD) (Dorans; & Kulick, 1986) โดยพัฒนามาจากโมเดลความลำเอียงของแบบทดสอบภายใต้ทฤษฎีการตอบข้อสอบแบบหลายมิติ (Multidimensional IRT) มีรูปแบบนันพาราเมทริก (Nonparametric form) ซึ่งไม่ต้องใช้ฟังก์ชันการตอบข้อสอบประมาณค่าความสามารถ วิธีซิปเทสท์เป็นวิธีที่ใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีทิศทางเดียวกัน (Unidirectional DIF) โดยเฉพาะ ดังนั้นจึงมีข้อจำกัดในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ไม่มีทิศทางเดียวกัน (Nonunidirectional DIF) (Li; & Stout. 1996 : 650; Shealy; & Stout. 1993 : 162) ส่วนข้อได้เปรียบของวิธีซิปเทสท์ก็คือ สามารถคำนวณได้ง่าย เสียค่าใช้จ่ายไม่มาก และไม่จำเป็นต้องใช้ตัวอย่างขนาดใหญ่ สามารถใช้สถิติทดสอบทดสอบนัยสำคัญ เพื่อตัดสินการทำหน้าที่ต่างกันของข้อสอบครั้งละหนึ่งข้อ หรือมากกว่าหนึ่งข้อพร้อมกัน (Simultaneous) ผลการวิเคราะห์ทำให้ทราบขนาดและทิศทางของการทำหน้าที่ต่างกันของข้อสอบ (Nandakumar. 1993 : 295) นักวิจัยหลายคนได้นำวิธีซิปเทสท์มาศึกษาและปรับขยายเพื่อใช้ตรวจสอบในประเด็นต่าง ๆ เช่น แนนดาคูมาร์ (Nandakumar. 1993) ได้ศึกษาการทำหน้าที่ต่างกันของข้อสอบในสองกรณี คือการขยายข้อสอบทำหน้าที่ต่างกัน (DIF amplification) และการหักล้างข้อสอบทำหน้าที่ต่างกัน (DIF cancellation) ลีและสเตาท์ (Li; & Stout. 1996) ได้พัฒนาวิธีครอซซิงซิปเทสท์ (Crossing SIBTEST) เพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบตัดกัน รุสโซและสเตาท์ (Roussos; & Stout. 1996a) ได้ศึกษาอัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีซิปเทสท์ในกลุ่มตัวอย่างขนาดเล็ก ชาง มาซซีโอ และรุสโซ (Chang; Mazzeo; & Roussos. 1996) ได้พัฒนาวิธีโพลี-ซิปเทสท์ (Poly-SIBTEST)

เพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนหลายค่า สเตาท์และคนอื่นๆ (Stout; et al., 1997) ได้พัฒนาวิธีมัลติซิป (MULTISIB) เพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบสองมิติ เป็นต้น

แนวคิดและหลักการ

Shealy; & Stout (1993 : 163-164) ได้อธิบายการทำหน้าที่ต่างกันของข้อสอบ โดยใช้ขอบของฟังก์ชันการตอบข้อสอบ (Marginal IRFs) ของความสามารถเป้าหมายที่ต้องการวัด สำหรับกลุ่ม g (กลุ่มอ้างอิง หรือกลุ่มสนใจ) ดังนี้

$$M_{ig}(\theta) = E\left[P_i(\Theta,\eta) \mid \Theta = \theta, G = d\right]$$

ถ้า $\eta \mid \Theta = \theta, G = g$ มีความหนาแน่นแบบมีเงื่อนไขของ η เมื่อกำหนดความสามารถ θ ของกลุ่ม g มีค่าคงที่ ซึ่งแทนด้วย $f_g(\eta \mid \theta)$ ดังนั้นการกำหนดนิยามในสมการ (1) สามารถคำนวณ ได้ดังนี้

$$M_{ig}(\theta) = \int_{-\infty}^{+\infty} P_i(\theta,\eta)\, f_g(\eta \mid \theta) d\eta$$

เมื่อ $M_{ig}(\theta)$ คือ Marginal IRT สำหรับความสามารถเป้าหมายที่ต้องการวัด (θ) ของผู้สอบกลุ่มอ้างอิงหรือกลุ่มเปรียบเทียบ

$P_i(\theta,n)$ คือ IRT ของข้อสอบข้อที่ i

$\int_g(\eta/\theta)$ คือ การแจกแจงแบบมีเงื่อนไขของกลุ่มผู้สอบ

การเปรียบเทียบ Marginal IRT ระหว่างกลุ่มอ้างอิง (R) กับกลุ่มเปรียบเทียบ (F) จะทำให้ทราบถึงทิศทางของการได้เปรียบหรือเสียเปรียบ กล่าวคือ ถ้า $M_{iR}(\theta) > M_{iF}(\theta)$ ทุกค่าของ θ แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบมีทิศทางเดียว โดยข้อสอบจะเข้าข้างกลุ่มอ้างอิงและถ้า $M_{iR}(\theta) < M_{iF}(\theta)$ ทุกค่าของ θ แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบมีทิศทางเดียว โดยข้อสอบจะเข้าข้างผู้สอบกลุ่มเปรียบเทียบ การทำหน้าที่ต่างกันของข้อสอบมีทิศทางเดียว อาจเรียกอีกอย่างหนึ่งว่า “การทำหน้าที่ต่างกันแบบไม่ตัดกัน” (Non-crossing DIF)

ในการเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบตามวิธีซิปเทสท์ จะแบ่งข้อสอบออกเป็นสองชุดย่อย (Subtests) คือ 1) ชุดแบบทดสอบที่มีความตรง

(Validity Subtests) หรือชุดแบบทดสอบที่ใช้ในการจับคู่เปรียบเทียบ (Matching Subtest) แบบทดสอบชุดนี้ประกอบด้วยข้อสอบที่ทำหน้าที่ไม่ต่างกัน 2) ชุดแบบทดสอบที่ต้องการศึกษา (Studied Subtests) ประกอบด้วยข้อสอบที่สงสัยว่าทำหน้าที่ต่างกัน ถ้าแบบทดสอบชุดแรกมีจำนวน n ข้อ (ข้อที่ 1 ถึง n) แล้วแบบทดสอบชุดที่สองจะมีจำนวน N-n ข้อ (ข้อที่ n + 1 ถึง N) เมื่อ N เป็นจำนวนข้อสอบทั้งหมด

ฟังก์ชันการตอบข้อสอบของแบบทดสอบที่ต้องการศึกษา สำหรับผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ กำหนดในรูปฟังก์ชัน Marginal ดังนี้

$$M_{SR}(\theta) = \sum_{i=n+1}^{N} M_{iR}(\theta)$$

$$M_{SF}(\theta) = \sum_{i=n+1}^{N} M_{iF}(\theta)$$

เมื่อ $M_{SR}(\theta)$ และ$M_{SF}(\theta)$ แทนผลรวม Marginal IRFs ของข้อสอบที่ต้องการศึกษา ณ ระดับความสามารถ θ จากผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ตามลำดับ สำหรับปริมาณของการทำหน้าที่ต่างกันของข้อสอบ (Amount of DIF) สามารถคำนวณจากความแตกต่างระหว่าง$M_{SR}(\theta)$ และ$M_{SF}(\theta)$ ดังนี้

$$B(\theta) = M_{SR}(\theta) - M_{SF}(\theta)\ |$$

ขนาดของความแตกต่างดังกล่าว แสดงถึงปริมาณของการทำหน้าที่ต่างกันของข้อสอบจากแบบทดสอบชุดย่อยที่ต้องการศึกษา ณ ระดับความสามารถ θ ซึ่งเข้าข้างกลุ่มอ้างอิง ชีลลีและสเตาท์ (Shealy; & Stout. 1993: 167) ได้คำนวณค่าเฉลี่ยของปริมาณการทำหน้าที่ต่างกันของข้อสอบที่มีทิศทางเดียวกัน (Unidirectional DIF) ดังนี้

$$\beta_{uni} = \int_{-\infty}^{+\infty} B(\theta)\, f_g(\theta) d\theta$$

เมื่อ β_{uni} แทนดัชนีการทำหน้าที่ต่างกันของข้อสอบที่มีทิศทางเดียวกัน และ $f_F(\theta)$ แทนฟังก์ชันความหนาแน่นความน่าจะเป็นของการแจกแจงความสามารถเป้าหมาย จากผู้สอบกลุ่มรวมทั้งหมด

ดัชนี β_{uni} ที่คำนวณได้จากสูตรดังกล่าว นำมาทดสอบสมมติฐานของการทำหน้าที่ต่างกันของข้อสอบแบบมีทิศทางเดียว ดังนี้

$$H_0 = \beta_{uni} = 0$$
$$H_1 = \beta_{uni} > 0$$

ในการทดสอบสมมติฐานของการทำหน้าที่ต่างกันของข้อสอบที่มีทิศทางเดียวกัน เมื่อข้อสอบเข้าข้างกลุ่มอ้างอิง นำดัชนี มากำหนด β_{uni} สมมติฐานศูนย์ (H_0) และสมมติฐานทางเลือก (H_1) ดังนี้

การทดสอบสมมติฐานการทำหน้าที่ต่างกันของข้อสอบที่มีทิศทางเดียวกันจะประมาณค่าดัชนี β_{uni} โดยคำนวณจากคะแนนของแบบทดสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบ และแบบทดสอบชุดย่อยที่ต้องการศึกษา ดังนี้

$$X = \sum_{i=1}^{n} U_i$$
$$Y = \sum_{i=n+1}^{N} U_i$$

เมื่อ

X แทน คะแนนรวมจากแบบทดสอบชุดย่อยที่มีความตรง

Y แทน คะแนนรวมจากแบบทดสอบชุดย่อยที่ต้องการศึกษา

U_i แทน ผลการตอบข้อสอบข้อที่ i (ตอบถูกได้ 1 คะแนน และตอบผิดได้ 0 คะแนน)

คำนวณคะแนนเฉลี่ยจากผลการตอบข้อสอบในแบบทดสอบชุดย่อยที่ต้องการศึกษาของผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่มีความสามารถระดับเดียวกัน แล้วนำคะแนนเฉลี่ยดังกล่าวมาจับคู่เปรียบเทียบ โดยพิจารณาได้จากคะแนนรวมที่เท่ากันของแบบทดสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบ ($X = k$) ดังนี้

$$\bar{Y}_{Rk} - \bar{Y}_{Fk} \qquad k = 0,1,2,...,n$$

เมื่อ $\bar{Y}_{Rk}$ และ $\bar{Y}_{Fk}$ แทนค่าเฉลี่ยของคะแนน Y จากการตอบแบบทดสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบ แล้วได้คะแนนรวม $X = k$ สำหรับผู้สอบในกลุ่มอ้างอิงและกลุ่มสนใจตามลำดับ คะแนนเฉลี่ยที่ใช้ในการเปรียบเทียบดังกล่าวอาจทำให้การตรวจสอบผิดพลาดจากความเป็นจริง กล่าวคือ เมื่อเกิดความแตกต่างของการแจกแจงค่าความสามารถ (Ability distribution)ของกลุ่มอ้างอิงและกลุ่มสนใจจะมีผลทำให้ $\bar{Y}_{Rk} - \bar{Y}_{Fk}$ มีค่าแตกต่างจาก 0 อย่างเป็นระบบ ทำให้ตรวจพบว่าข้อสอบทำหน้าที่ต่างกัน ซึ่งความเป็นจริงแล้วข้อสอบทำหน้าที่ไม่ต่างกัน ดังนั้นความแตกต่างของการแจกแจงค่าความสามารถของกลุ่มอ้างอิงและกลุ่มสนใจที่เกิดขึ้นสามารถปรับแก้ค่าการถดถอย (Regression correction) เพื่อกำจัดค่าที่สูงเกินปกติ (Inflate) (Shealy; & Stout, 1993 : 169) สำหรับค่าเฉลี่ย $\bar{Y}_{Rk}$ และ $\bar{Y}_{Rk}$ ที่ปรับแก้แล้วแทนด้วย $\bar{Y}^*_{Rk}$ และ $\bar{Y}^*_{Fk}$ ”ตามลำดับ

ค่า $\bar{Y}^*_{Rk} - \bar{Y}^*_{Fk}$ เป็นความแตกต่างของผลการตอบข้อสอบในแบบทดสอบชุดย่อยที่ศึกษาระหว่างกลุ่มผู้สอบที่มีความสามารถระดับเดียวกัน ถ้า $\bar{Y}^*_{Rk} - \bar{Y}^*_{Fk} = 0$ ทุกคะแนน k แสดงว่า ข้อสอบที่สงสัยในแบบทดสอบชุดย่อยทำหน้าที่ไม่ต่างกัน (No-DIF) และถ้า $\bar{Y}^*_{Rk} - \bar{Y}^*_{Fk} > 0$ ทุกคะแนน k แสดงว่าข้อสอบที่สงสัยในแบบทดสอบชุดย่อยทำหน้าที่ต่างกันที่มีทิศทางเดียวกัน (Unidirectional DIF) โดยข้อสอบเข้าข้างกลุ่มอ้างอิง ถ้า $\bar{Y}^*_{Rk} - \bar{Y}^*_{Fk} < 0$ ทุกคะแนน k แสดงว่าข้อสอบที่สงสัยในแบบทดสอบชุดย่อยทำหน้าที่ต่างกันที่มีทิศทางเดียวกัน โดยข้อสอบเข้าข้างกลุ่มสนใจ สำหรับค่าความแตกต่างของผลการตอบข้อสอบดังกล่าว สามารถนำมาประมาณค่าในรูป β_{uni} ดังนี้

$$\hat{\beta}_{uni} = \sum \hat{P}_k \left(\bar{Y}^*_{Rk} - \bar{Y}^*_{Fk} \right)$$

เมื่อ $\hat{P}_k$ แทนสัดส่วนของผู้สอบทั้งหมด (กลุ่มอ้างอิงและกลุ่มสนใจ) ซึ่งตอบแบบทดสอบชุดย่อยที่ใช้จับคู่เปรียบเทียบ แล้วได้คะแนนรวม $X = k$ สัดส่วนของผู้สอบดังกล่าวสามารถเขียนในรูปสัญลักษณ์ ดังนี้

$$\hat{P}_k = \frac{\left(J_{Rk} + J_{Fk} \right)}{\sum_{k=1}^{n} \left(J_{Rk} + J_{Fk} \right)}$$

เมื่อ J_{Rk} และ J_{Fk} แทนจำนวนผู้สอบซึ่งตอบแบบทดสอบชุดย่อยที่ใช้จับคู่เปรียบเทียบแล้วได้คะแนนรวม $X = k$ สำหรับกลุ่มอ้างอิงและกลุ่มสนใจ ตามลำดับ จากนั้นจึงนำค่าประมาณ $\hat{\beta}_{uni}$ มาทดสอบสมมติฐานศูนย์ (No-DIF) โดยใช้สถิติ β_{uni} ดังนี้

$$B_{uni} = \frac{\hat{\beta}_{uni}}{\hat{\sigma}(\beta_{uni})}$$

$\hat{\sigma}(\beta_{uni})$ เป็นค่าประมาณความคลาดเคลื่อนมาตรฐานของ β_{uni} คำนวณจาก

$$\hat{\sigma}(\hat{\beta}_{uni}) = \sqrt{\sum \hat{P}_k^2 \left[\frac{1}{J_{Rk}} \hat{\sigma}^2(Y \mid k, R) + \frac{1}{J_{Fk}} \hat{\sigma}^2(Y \mid k, F) \right]}$$

เมื่อ $\hat{\sigma}^2(Y \mid k, R)$ และ $\hat{\sigma}^2(Y \mid k, F))$ แทนค่าประมาณความแปรปรวนของคะแนนจากแบบทดสอบชุดย่อยที่ต้องการศึกษาในกลุ่มอ้างอิงและเปรียบเทียบ ตามลำดับ สำหรับสถิติ B_{uni} มีการแจกแจงใกล้เคียงการแจกแจงแบบปกติมาตรฐาน [$N(0, 1)$] เมื่อข้อสอบทำหน้าที่ไม่ต่างกันและถ้าผลการทดสอบพบว่า $B_{uni} > Z_{\alpha}$ อย่างมีนัยสำคัญที่ระดับ α โดยที่ $P[N(0, 1) > Z_{\alpha}] = \alpha$ แสดงว่า ปฏิเสธ H_0 นั่นคือ ข้อสอบทำหน้าที่ต่างกันที่มีทิศทางเดียวกัน (Unidirectional DIF) เมื่อ $B_{uni} > 0$ แสดงว่าข้อสอบเข้าข้างกลุ่มอ้างอิง และเมื่อ $B_{uni} < 0$ แสดงว่าข้อสอบเข้าข้างกลุ่มสนใจ รุสโซและสเตาท์ (Roussos; & Stout. 1996a : 220) ได้เสนอเกณฑ์เพื่อใช้จำแนกขนาดของการทำหน้าที่ต่างกันของข้อสอบไว้ดังนี้

1. DIF ระดับ A ขนาดเล็ก : ปฏิเสธสมมติฐานศูนย์ และ $|\hat{\beta}_{uni}| < 0.059$
2. DIF ระดับ B ขนาดปานกลาง : ปฏิเสธสมมติฐานศูนย์ และ $0.059 \leq |\hat{\beta}_{uni}| < 0.088$
3. DIF ระดับ C ขนาดใหญ่ : ปฏิเสธสมมติฐานศูนย์ และ $|\hat{\beta}_{uni}| \geq 0.088$

การปรับแก้ค่าการถดถอย

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจะมีปัญหาในเชิงสถิติ ซึ่งเกิดจากความแตกต่างของการแจกแจงความสามารถเป้าหมายระหว่างกลุ่มอ้างอิงและเปรียบเทียบ กล่าวคือ ถ้าการแจกแจงความสามารถเป้าหมายของผู้สอบกลุ่มอ้างอิงสูงกว่ากลุ่มสนใจจะเกิดเงื่อนไขที่เรียกว่า“ผลกระทบ” (Impact) ซึ่งทำให้สถิติ β_{uni} มีค่าเฟ้อ (Inflate) หรือมีค่าสูงเกินปกติ แล้วจะส่งผลให้การตรวจสอบเกิดความคลาดเคลื่อนประเภทที่ 1 (Type I error) เพราะในความเป็นจริงข้อสอบทำหน้าที่ไม่ต่างกันแต่ตรวจพบว่าข้อสอบทำหน้าที่ต่างกัน ดังนั้นจึงมีความจำเป็นที่จะต้อง

ปรับแก้ความแตกต่างของการแจกแจงความสามารถเป้าหมายโดยใช้การปรับแก้การถดถอย (Regression correction) เพื่อกำจัดอิทธิพลค่าเฟ้อของผลกระทบ ซึ่งจะแปลงค่า $\overline{Y}_{Rk}, \overline{Y}_{Fk}$ เป็น $\overline{Y}^*_{Rk}, \overline{Y}^*_{Fk}$ ละคู่ ดังนั้น $\overline{Y}^*_{gk}$ เป็นตัวประมาณค่าอิทธิพลของค่าเฉลี่ยคะแนนจริงจากแบบทดสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบในกลุ่มย่อย k กลุ่ม ซึ่งแบ่งมาจากแต่ละกลุ่มของกลุ่มอ้างอิงและกลุ่มสนใจโดยมีรายละเอียดดังนี้ (Shealy & Stout. 1993 : 190 -193)

กำหนดให้ $X = k$ แทนคะแนนรวมซึ่งเป็นคะแนนสังเกตจากแบบทดสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบ $V_g(k)$ แทนค่าการถดถอยของคะแนนจริงจากแบบทดสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบซึ่งได้คะแนนเท่ากับ k ในการประมาณค่า$Vg(k)$ สมมติว่าเป็นเส้นตรงดังนี้

$$V_g(k) = \alpha + \beta k$$

ในการประมาณค่า $V_g(k)$ จะใช้ข้อตกลงของโมเดลคะแนนจริง ดังนี้

$$X = T + e \qquad \text{(ก)}$$

$$E(e) = 0 \text{ และ } cov(T, e) = 0 \qquad \text{(ข)}$$

เมื่อ X, T และ e แทนคะแนนสังเกต คะแนนจริง และคะแนนความคลาดเคลื่อนจากแบบทดสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบ ตามลำดับ แล้วใช้ทฤษฎีการถดถอยมาตรฐานประมาณค่า$V_g(k)$ ดังนี้

$$V_g(k) = ET + \left(\frac{\rho_{XI}\sigma_T}{\sigma_X}\right)\left(k - EX\ |\right) \qquad \text{(ค)}$$

เมื่อ EX และ ET แทนค่าคาดหวังของ X และ T ในกลุ่ม g (R หรือ F) สำหรับ σ_X และ σ_T แทนส่วนเบี่ยงเบนมาตรฐานของ X และ T ในกลุ่ม g (R หรือ F) ส่วน ρ_{XT} แทนความสัมพันธ์ระหว่าง X และ T ในกลุ่ม g (R หรือ F) จากโมเดลคะแนนจริงสามารถคำนวณค่าความเชื่อมั่น ดังนี้

$$\frac{\rho_{XT}\sigma_T}{\sigma_X} = 1 - \frac{\sigma_e^2}{\sigma_X^2} \qquad \text{(ง)}$$

เมื่อ σ_e^2 และ σ_X^2 แทนความแปรปรวนของความคลาดเคลื่อนและความแปรปรวนของคะแนนสังเกตในกลุ่ม g (R หรือ F) ตามลำดับ จากสมการ (ก) และ (ข) แสดงว่า $ET = EX$ ดังนั้น สมการ (ค) และ (ง) สามารถเขียนใหม่ได้ดังนี้

$$V_g(k) = EX + \left(1 - \frac{\sigma_E^2}{\sigma_x^2}\right)(k - EX)$$

ค่าประมาณของ $v_g(k)$ สามารถคำนวณดังนี้

$$\hat{V}_g^2(k) = \bar{X}_g + \left(1 - | \frac{\hat{\sigma}^2(e \mid g)}{\sigma^2(X \mid g)}\right)(k - \bar{X}_g)$$

โดยที่

$$\sigma^2(x \mid g) = \frac{1}{(J_g - 1)} \sum_{j=1}^{Jg} (X_{gi} - \bar{X}_g)^2$$

และ

$$\hat{\sigma}^2(e \mid g) = \sum_{i=1}^{n} \bar{U}_{ig}(1 - \bar{U}_{ig})$$

เมื่อ

X_{gj} แทน คะแนนจากแบบทดสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบของผู้สอบคนที่ j ในกลุ่ม g

$\bar{X}_g$ แทน คะแนนเฉลี่ยจากแบบทดสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบของผู้สอบคนที่ j ในกลุ่ม g

$\bar{U}_{ig}$ แทน สัดส่วนการตอบข้อสอบถูกของผู้สอบกลุ่ม g ซึ่งตอบข้อสอบข้อที่ i จากแบบทดสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบ

ค่า $\hat{V}_g(k)$ เป็นค่าประมาณคะแนนจริงของความสามารถเป้าหมายสำหรับผู้สอบกลุ่ม g (R หรือ F) ที่ได้คะแนน $X = k$ จากแบบทดสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบ ซึ่งสอดคล้องกับคะแนนจากแบบทดสอบชุดย่อยที่ต้องการศึกษา โดยใช้วิธีอนุกรมของเทเลอร์ (Taylor, 1993) ปรับแก้คะแนนดังนี้

$$\bar{Y}^{*}_{gk} = \bar{Y}_{gk} + \hat{M}_{gk}\left[\hat{V}(k) - \hat{V}_{g}(k)\right]$$

เมื่อ $\bar{Y}_{gk}$ แทนค่าเฉลี่ยของคะแนนสังเกตจากแบบทดสอบชุดย่อยที่ต้องการศึกษาของผู้สอบกลุ่ม g (R หรือ F) ซึ่งได้คะแนน $X = k$ จากแบบทดสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบ สำหรับ $\hat{V}(k)$ และ $\hat{M}_{gk}$ คำนวณดังนี้

$$\hat{V}(k) = \frac{\hat{v}_{R}(k) + \hat{V}_{F}(k)}{2}$$

$$\hat{M}_{gk} = \frac{\bar{Y}_{g,k+1} - \bar{Y}_{g,k-1}}{\hat{V}_{g}(k+1) - \hat{V}_{g}(k-1)}$$

เมื่อ $\bar{Y}_{g,k+1}$ และ $\bar{Y}_{g,k-1}$ แทนค่าเฉลี่ยของคะแนนสังเกตจากแบบทดสอบชุดย่อยที่ต้องการศึกษาของผู้สอบกลุ่มย่อย จำนวน $k+1$ กลุ่มและ $k-1$ กลุ่ม ตามลำดับ $\hat{V}_{g}(k+1)$ และ $\hat{V}_{g}(k-1)$ แทนค่าประมาณของคะแนนจริงจากแบบสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบของผู้สอบกลุ่มย่อย จำนวน $k+1$ กลุ่มและ $k-1$ กลุ่ม ตามลำดับ สำหรับค่า $\bar{Y}^{*}_{gk}$ เป็นค่าประมาณของคะแนนจริงจากแบบทดสอบชุดย่อยที่ต้องการศึกษาของผู้สอบกลุ่มย่อย k ในกลุ่มg (F หรือ R) ซึ่งสมมติว่าเท่ากับค่าประมาณคะแนนจริงจากแบบทดสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบ $V(k)$ ของผู้สอบทั้งสองกลุ่ม

4. วิธีตัวแบบเชิงเส้นวางนัยทั่วไปลดหลั่น (Hierarchical Generalized Linear Model : HGLM)

คามาตะ (Kamata, 1998) ได้เสนอวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยใช้ตัวแบบเชิงเส้นวางนัยทั่วไประดับลดหลั่น (Hierarchical Generalized Linear Model, HGLM) มีพื้นฐานจากตัวแบบของราสช์ที่มีการกำหนดให้ ข้อสอบซ้อนทับในผู้สอบ นั่นคือข้อสอบจัดเป็นหน่วยในระดับที่ 1 (level-1) และผู้สอบจัดเป็นหน่วยในระดับที่ 2 (Level-2) หมายความว่าผู้สอบแต่ละคน (j=1,2,3,…,n) จะตอบข้อสอบแต่ล่ะข้อในแบบทดสอบ (i =1,2,3,…,k) ซึ่งข้อมูลที่มีการซ้อนทับกันจะเรียกว่า ข้อมูลพหุระดับ (ศิริชัย กาญจนวาสี, 2545 : 124)

ถ้ากำหนดให้ y_{ij} คือ ผลการตอบสนองของข้อสอบข้อที่ i และของผู้ตอบคนที่ j ที่มีค่าเป็น 0 ถ้าผู้ตอบข้อสอบไม่ถูกต้องและมีค่าเท่ากับ 1 เมื่อได้ตอบถูกต้องและกำหนดให้โอกาสที่ผู้สอบคนที่ j จะตอบข้อสอบข้อที่ i ได้ถูกต้องคือ p_{ij} ดังนั้นผลการตอบสนองของ y_{ij} เมื่อกำหนดค่า p_{ij} จะมีการแจกแจงแบบเบอร์นูลี หรือเขียนแทนด้วยสัญลักษณ์ได้ดังนี้

$$Y_{ij} \mid p_{ij} \sim B(1, p_{ij})$$

โดยที่ j=1,2,3,…,k ; n=1,2,…,k
ดังนั้นจะได้ว่า

$$E(Y_{ij} \mid p_{ij}) = p_{ij} \text{ และ } \mathrm{Var}(Y_{ij} \mid p_{ij}) = p_{ij}(1 - p_{ij})$$

ถ้ากำหนดให้ฟังชันโลจิท เป็นฟังก์ชันเชื่อมโยงแล้ว

$$n_{ij} = \log\left(\frac{p_{ij}}{1 - p_{ij}}\right)$$

สามารถเขียนตัวแบบของ HGLM ในแต่ละระดับได้ดังนี้

ตัวแบบระดับที่ 1 (Level-1 Models) สามารถเขียนเป็นสมการได้ดังนี้

$$\begin{aligned}\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) &= n_{ij} \\ &= \beta_{0j} + \beta_{1j} X_{1ij} + \beta_{2j} X_{2ij} + \ldots + \beta_{(k-1)ij} \\ &= \beta_{0j} + \sum \beta_{qj} X_{qij}\end{aligned}$$

เมื่อ p_{ij} คือความน่าจะเป็นที่ผู้สอบคนที่ j จะตอบข้อสอบข้อที่ i ได้ถูกต้อง η_{ij} คือ โครงสร้างของตัวแบบระดับที่ 1 หรือ สมการโลจิทของ p_{ij} . X_{qij} คือตัวแปรดัมมี่ตัวที่ q ของข้อสอบข้อที่ i ที่ตอบโดยผู้ตอบคนที่ j ซึ่ง $X_{qij} = -1$ เมื่อ q-1 และ $X_{qij} = 0$ เมื่อ $q \neq i$ โดยที่ q= 1,2,3,…,k-1 และ β_{0j} คือสัมประสิทธิ์ของค่าจุดตัดซึ่งสามารถใช้แสดงถึงความยากของข้อสอบที่ใช้อ้างอิง (Reference Item) ที่กำหนดให้ สัมประสิทธิ์ของข้อสอบเป็น 1 . β_{qj} คือ สัมประสิทธิ์ของตัวดัมมี่ตัวที่ q หรือ ค่าพารามิเตอร์ความยากของข้อสอบข้อที่ q ที่แตกต่างไปจากความยากของข้อสอบที่ใช้อ้างอิงโดยกำหนดให้มีค่าเป็น 0 เมื่อข้อสอบข้อนั้น มีการทำหน้าที่ต่างกัน (DIF) เกิดขึ้น ดังนั้นจะเห็นว่า ตัวแบบระดับที่ 1 เป็นตัวแบบที่เกี่ยวข้องกับ ค่าพารามิเตอร์ของข้อสอบ จึงเรียกตัวแบบระดับที่ 1 ว่า ตัวแบบของข้อสอบ

ตัวแบบระดับที่ 2 (Level-2 Model)

ในตัวแบบระดับนี้ค่าพารามิเตอร์ความยากของข้อสอบ แต่ละตัวจะสร้างโดยอาศัยข้อมูลของผู้สอบ ซึ่งในกรณีที่ใช้ในการตรวจสอบ DIF จะใช้กลุ่มของผู้สอบเป็นตัวแปรอิสระ ดังนั้น ตัวแบบระดับที 2 สามารถเขียนเป็นสมการได้ดังต่อไปนี้

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(GROUP) + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}(GROUP)$$
$$\beta_{2j} = \gamma_{20} + \gamma_{21}(GROUP)$$
.
.
.
$$\beta_{(k-1)j} = \gamma_{(k-1)0} + \gamma_{(k-1)1}(GROUP)$$

เมื่อ GROUP เป็นตัวแปรดัมมี่ โดยกำหนดให้ ถ้าผู้สอบอยู่ในกลุ่มที่ต้องการศึกษาหรือเปรียบเทียบ (Focal Group) ค่า GROUP =1 แต่ถ้าอยู่ในกลุ่มอ้างอิง (Reference Group) ค่า GROUP = 0 ค่า μ_{0j} ถือว่าเป็นค่าอิทธิพลสุ่มที่แสดงถึงความสามารถของผู้สอบคนที่ j และจะกำหนดให้มีการแจกแจงแบบปกติที่มีค่าเฉลี่ยเป็น 0 และความแปรปรวนเท่ากับ σ^2 ค่าอื่น ๆ ในตัวแบบถือว่าเป็นอิทธิพลกำหนด โดยที่ค่า γ_{00} หมายถึง ค่าเฉลี่ยของความสามารถของผู้สอบที่อยู่ในกลุ่มอ้างอิง ค่า γ_{01} หมายถึง ค่าเฉลี่ยความสามารถของผู้สอบที่อยู่ในกลุ่มเปรียบเทียบที่แตกต่างไปจากค่าเฉลี่ยของความสามารถของผู้สอบที่อยู่ในกลุ่มอ้างอิง สำหรับค่า γ_{q0} จะหมายถึง ความยากข้อของข้อสอบข้อที่ q ของผู้สอบที่อยู่ในกลุ่มอ้างอิง ค่า γ_{q1} คือความยากของข้อสอบข้อที่ q ของผู้สอบที่อยู่ในกลุ่มเปรียบเทียบที่แตกต่างจากผู้สอบที่อยู่ในกลุ่มอ้างอิง นั่นคือ ค่า DIF นั่นเอง

สมมติฐานของการทดสอบข้อสอบที่ทำหน้าที่ต่างกัน คือ

$$H_0 : \gamma_{q1} = 0$$
$$H_1 : \gamma_{q1} \neq 0$$

ตัวสถิติที่ใช้ในการทดสอบ คือ T-Ratio

$$t = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}}$$

ซึ่งข้อสอบข้อที่ q จะถือว่ามี DIF เกิดขึ้นเมื่อทดสอบสมมติฐาน $H_0 : \gamma_{q1} = 0$ แล้วปฏิเสธสมมติฐาน H_0 โปแกรม ศ?6.0 ที่ใช้ในการวิเคราะห์ตัวแบบ HGLM จะให้ค่า P-value ของตัวสถิติ T-ratio เป็นตัวสถิติที่ใช้ในการตัดสินใจที่ปฏิเสธหรือยอมรับสมมติฐาน $H_0 : \gamma_{q1} = 0$ ได้โดยตรง วิธีการประมาณค่าพารามิเตอร์ต่าง ๆ ในตัวแบบ HGLM ด้วยโปรแกรม HLM 6.0 นี้ สามารถเลือกใช้วิธีการ (Penalized Quasi-Likelihood : PQL) หรือ LaPlace6 ในการประมาณค่าความควรจะเป็นสูงสุด โดยมีข้อสมมุติว่าการแจกแจงความสามารถของผู้สอบจะต้องเป็นแบบปกติ

แต่ในสถานการณ์จริงของการทอสอบ ความสามารถของผู้สอบไม่จำเป็นที่จะต้องมีการแจกแจงแบบปกติเสมอไป บางครั้งอาจจะมีการแจกแจงที่มีการเบ้หรือการแจกแจงแบบอื่น ๆ

5. วิธีการทดสอบไคสแควร์ของลอร์ด (Lord's chi-square test)

วิธีการทดสอบไคสแควร์ของลอร์ด (Lord, 1980) เป็นวิธีที่ใช้เพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบภายใต้ทฤษฎีการตอบสนองข้อสอบ (IRT) โดยใช้โมเดลโลจิสติกแบบ 3 พารามิเตอร์ หลักการตรวจสอบด้วยวิธีนี้จะทดสอบความแตกต่างของค่าพารามิเตอร์ของข้อสอบระหว่างฟังก์ชันการตอบสนองข้อสอบ (IRTs) จากผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบจะเริ่มต้นด้วยการประมาณค่าพารามิเตอร์ a_i , bi และ c_i จากผู้สอบทั้งสองกลุ่มรวมกันแล้วแปลงค่าพารามิเตอร์ b_i ให้เป็นค่ามาตรฐานต่อจากนั้นจะประมาณค่าพารามิเตอร์ a_i และ b_i อีกครั้ง โดยประมาณค่าจากผู้สอบแต่ละกลุ่มส่วนพารามิเตอร์ c_i ยังคงใช้ค่าเดิมที่ประมาณได้ในครั้งแรก ข้อสอบที่มีค่า c_i ต่ำ และข้อสอบที่มีค่า b_i สูงหรือต่ำมากๆ จะถูกคัดออกไป แล้วแปลงค่าพารามิเตอร์ b_i ที่ได้ใหม่ให้เป็นค่ามาตรฐานอีกครั้งหนึ่ง ต่อจากนั้นจึงนำค่าพารามิเตอร์ a_i และb_i ของข้อสอบแต่ละข้อไปเปรียบเทียบความแตกต่างโดยใช้สถิติไคสแควร์ ซึ่งระดับชั้นของความเป็นอิสระเท่ากับ 2 สำหรับพารามิเตอร์ c_i ไม่ได้นำไปทดสอบด้วย ดังนั้นในการทดสอบสมมติฐานศูนย์ (H_0) และสมมติฐานอื่น (H_1) กำหนดดังนี้

$$H_0 : b_{iR} = b_{iF} \quad \text{และ} \quad a_{iF} = a_{iR}$$
$$H_1 : b_{iF} \neq b_{iR} \quad \text{และ} \quad a_{iF} \neq a_{iR}$$

สำหรับค่าสถิติไคสแควร์ $\left(\chi_i^2\right)$มีลักษณะดังนี้

โดยที่

$$\chi_i^2 = V_i^{'} \tau_i^{-1} V_i$$
$$V_i^{'} = \left(\hat{a}_{iF} - \hat{a}_{iR}, \hat{b}_{iF} - \hat{b}_{iR}\right)$$
$$\tau_i = \tau_{iF} + \tau_{iR}$$

เมื่อ $\hat{a}_{iF}$ แทน ค่าประมาณพารามิเตอร์อำนาจจำแนกของข้อสอบข้อที่ i จากผู้สอบกลุ่มเปรียบเทียบ

$\hat{a}_{iR}$ แทน ค่าประมาณพารามิเตอร์อำนาจจำแนกของข้อสอบข้อที่ i จากผู้สอบกลุ่มอ้างอิง

$\hat{b}_{iF}$ แทน ค่าประมาณพารามิเตอร์ค่าความยากของข้อสอบข้อที่ i จากผู้สอบกลุ่มเปรียบเทียบ

$\hat{b}_{iR}$ แทน ค่าประมาณพารามิเตอร์ค่าความยากของข้อสอบข้อที่ i จากผู้สอบกลุ่มอ้างอิง

$V_i^{'}$ แทน เวคเตอร์ความแตกต่างของค่าพารามิเตอร์ของข้อสอบที่ i จากผู้สอบกลุ่มเปรียบเทียบและกลุ่มอ้างอิง

τ_i แทน เมทริกซ์ความแปรปรวน – ความแปรปรวนร่วม และมีขนาด 2 x2

τ_i^{-1} แทน อินเวอส์ของเมทริกซ์

ถ้าผลทดสอบสมมติฐานแล้วปรากฏว่ามีค่าแตกต่างกันอย่างมีนัยสำคัญ แสดงว่า ข้อสอบดังกล่าวทำหน้าที่ต่างกัน สำหรับสถิติไคสแควร์จะมีประสิทธิภาพในการตรวจสอบเมื่ออยู่บนพื้นฐานของสมมติฐาน 3 ประการ คือ

1) การทดสอบทางสถิติมีการแจกแจงไคสแควร์เชิงเส้นกำกับ (Asymptotic)

2) ต้องรู้ค่าความสามารถของผู้สอบ

3) ใช้วิธีประมาณค่าแบบไลค์ลิฮู้ดสูงสุด (Maximum Likelihood Estimate)

นอกจากนี้วิธีการทดสอบไคสแควร์ของ Lord (1980) ยังสามารถนำไปประยุกต์ใช้กับโมเดลโลจิสติกแบบ 2 พารามิเตอร์ โดยกำหนดให้พารามิเตอร์การเดา c_i มีค่าเป็นศูนย์ ทั้งนี้เพราะว่าในการเปรียบเทียบค่าพารามิเตอร์ของข้อสอบจะทดสอบความแตกต่างเฉพาะพารามิเตอร์ความยาก b_i และค่าอำนาจจำแนก a_i เท่านั้น

6. วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วย IRT

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยตัวแบบ IRT ประกอบด้วย 2 ขั้นตอนที่สัมพันธ์กัน คือ ขั้นแรก ทำการวัดขนาดของการทำหน้าที่ของข้อสอบ (Measurement of DIF) และขั้นที่สอง ทำการทดสอบทางสถิติของการทำหน้าที่ต่างกันของข้อสอบ (Statistical Test of DIF)

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยตัวแบบ IRT ที่นิยมใช้กันมี ดังนี้

1) วิธีวัดความแตกต่างของพื้นที่ (Area Measures : AREA)

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการวัดพื้นที่ทำโดยการคำนวณค่าประมาณพื้นที่ระหว่างโค้งลักษณะข้อสอบระหว่างกลุ่มสองกลุ่มที่มีความสามารถระดับเดียวกันที่มีสูตรทั่วไปในการใช้คำนวณพื้นที่ ดังนี้

$$A = \int_{s} \left[P_R(\theta) - P_F(\theta) \right]$$

เมื่อ A แทน ดัชนีพื้นที่

$\int_{s}$ แทน ฟังก์ชันการตอบสนองของข้อสอบของผู้สอบที่มีความสามารถอยู่ในช่วง s ซึ่ง $s = (\theta_L, \theta_H)$

$P_R(\theta)$ แทน โอกาสในการตอบข้อสอบได้ถูกต้องของผู้สอบในกลุ่มอ้างอิงที่มีระดับความสามารถ θ

วิธีการวัดความแตกต่างของพื้นที่เป็นวิธีที่ทำความเข้าใจได้ง่ายสามารถวาดภาพแสดงได้ชัดเจน แต่มีข้อเสียคือ ค่าที่ได้ขาดความเชื่อถือ และมีความแตกต่างเมื่อใช้ช่วงของ θ ที่แตกต่างกัน วิธีการวัดความแตกต่างของพื้นที่นิยมใช้กันมี 2 วิธี คือ วิธีการวัดพื้นที่ของราจู (Raju, 1990) และวิธีการวัดพื้นที่ของ คิมและโคเฮน (Kim & Cohen, 1992)

2) วิธีวัดความแตกต่างของค่าพารามิเตอร์ความยาก (b Parameter difference)

วิธีนี้เป็นวิธีที่หาความแตกต่างระหว่างค่าพารามิเตอร์ความยากระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ทำโดยคิดเครื่องหมายเมื่อมีการควบคุมระดับความสามารถ นั่นคือ กำหนดให้

$$\begin{aligned} SBD - \theta &= b_F - b_R \\ &= \Delta b \end{aligned}$$

เมื่อ b_F, b_R คือ ค่าพารามิเตอร์ความยากของข้อสอบงกลุ่มเปรียบเทียบและกลุ่มอ้างอิงตามลำดับ ดังนั้น ถ้าค่า Δb มีค่าเป็นบวกแสดงว่าข้อสอบข้อนั้นมีค่าพารามิเตอร์ความยากของข้อสอบของกลุ่มเปรียบเทียบมีค่ามากกว่ากลุ่มอ้างอิง นั่นคือ ข้อสอบข้อนั้นเข้าข้างกลุ่มอ้างอิง แสดงว่ากลุ่มอ้างอิงมีโอกาสตอบข้อสอบข้อนั้นได้ถูกต้องมากกว่ากลุ่มเปรียบเทียบ ในทางตรงข้าม

ถ้าค่า Δb มีค่าเป็นลบ แสดงว่า กลุ่มเปรียบเทียบมีโอกาสตอบข้อสอบข้อนั้นได้ถูกต้องมากกว่ากลุ่มอ้างอิง

สำหรับการทดสอบนัยสำคัญของความแตกต่างระหว่างค่าพารามิเตอร์ความยากระหว่าง 2 กลุ่ม สำหรับทดสอบ $H_0 : \Delta b = 0$ ใช้สถิติการทดสอบดังนี้ (Lord, 1977)

$$d = \frac{\Delta\hat{b}}{S_{\Delta\hat{b}}}$$

$$\Delta\hat{b} = b_F - b_R$$

$$s_{\Delta b} = \sqrt{S_F^2 + S_R^2}$$

เมื่อ d แทน สถิติทดสอบซึ่งมีการแจกแจงปกติ

S_F แทน ความคลาดเคลื่อนมาตรฐานของ b_F

S_R แทน ความคลาดเคลื่อนมาตรฐานของ b_R

ผลการทดสอบพิจารณาได้จาก

1. การเปรียบเทียบค่า d ที่คำนวณได้กับค่าวิกฤติของค่าปกติมาตรฐาน
2. ถ้า $d > |Z|$ โดยที่ $d_{\text{ที่คำนวณได้}}$ แสดงว่าการทดสอบนัยสำคัญต่อข้อสอบนั้นที่ทำหน้าที่ต่างกัน

2. โครงการประเมินคุณภาพการศึกษาขั้นพื้นฐาน เพื่อการประกันคุณภาพผู้เรียน ปีการศึกษา 2550 สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน (สำนักทดสอบทางการศึกษา, 2550)

2.1. ความสำคัญของโครงการ

พระราชบัญญัติการศึกษาแห่งชาติ พ.ศ. 2542 มาตรา 47 กำหนดให้มีระบบการประกันคุณภาพการศึกษา เพื่อพัฒนาคุณภาพและมาตรฐานการศึกษาในทุกระดับ และมาตรา 48 ให้หน่วยงานต้นสังกัด และสถานศึกษา จัดให้มีระบบการประกันคุณภาพการศึกษาภายในสถานศึกษา และให้ถือว่าการประกันคุณภาพภายใน เป็นส่วนหนึ่งของกระบวนการบริหารการศึกษาที่ต้องดำเนินการอย่างต่อเนื่องโดยมีการจัดทำรายงานประจำปีเสนอต่อหน่วยงานต้นสังกัด หน่วยงานที่เกี่ยวข้อง และเปิดเผยต่อสาธารณชน เพื่อนำไปสู่การพัฒนาคุณภาพและมาตรฐานการศึกษา และเพื่อรองรับการประกันคุณภาพภายนอก การประเมินคุณภาพการศึกษาขั้นพื้นฐาน จึงเป็น

กระบวนการ วิธีการ เพื่อให้ได้ข้อมูลที่จะเป็นตัวบ่งชี้ถึงผลสำเร็จในการจัดการศึกษา ซึ่งเป็นส่วนประกอบสำคัญส่วนหนึ่งในการประกันคุณภาพภายใน หลักสูตรการศึกษาขั้นพื้นฐาน พ.ศ. 2544 จึงกำหนดแนวทางการวัดและประเมินผลการเรียนรู้ เพื่อให้ได้ข้อมูลสารสนเทศที่แสดงพัฒนาการ ความก้าวหน้า และความสำเร็จทางการเรียนของผู้เรียน ซึ่งสถานศึกษาต้องจัดให้มีการประเมินผลการเรียน ให้เป็นไปในมาตรฐานเดียวกัน ทั้งในระดับชั้นเรียน ระดับสถานศึกษา ระดับเขตพื้นที่การศึกษา และระดับชาติ ข้อมูลที่ได้จากการประเมินจะนำไปใช้ในการพัฒนาคุณภาพของผู้เรียน และคุณภาพการจัดการศึกษาของสถานศึกษาแต่ละแห่ง และเพื่อเป็นสารสนเทศรองรับบริบทของการประเมินภายนอก

โครงการประเมินคุณภาพการศึกษาขั้นพื้นฐาน เพื่อการประกันคุณภาพผู้เรียน ปีการศึกษา 2550 เป็นการตรวจสอบ ควบคุม กำกับดูแล และรักษาคุณภาพการศึกษาของสถานศึกษา ซึ่งสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐานมอบสำนักทดสอบทางการศึกษา จัดการประเมินผลสัมฤทธิ์ทางการเรียนของนักเรียน ในช่วงชั้นที่ 1 (ประถมศึกษาปีที่ 3) และช่วงชั้นที่ 3 (มัธยมศึกษาปีที่ 3) ทุกโรงเรียน ทุกคน ในแต่ละเขตพื้นที่การศึกษาทั่วประเทศ ซึ่งสถาบันทดสอบทางการศึกษาแห่งชาติ(องค์การมหาชน) ไม่ได้จัดการประเมิน และมอบสำนักงานเขตพื้นที่การศึกษา รับผิดชอบประเมินนักเรียนทุกคนในชั้นประถมศึกษาปีที่ 2 ประถมศึกษาปีที่ 5 และมัธยมศึกษาปีที่ 2 สำหรับผลการประเมินช่วงชั้นที่ 1 (ประถมศึกษาปีที่ 3) และช่วงชั้นที่ 3 (มัธยมศึกษาปีที่ 3) ที่ได้ จะสามารถนำมาพิจารณาเปรียบเทียบความก้าวหน้าของนักเรียนเป็นรายบุคคล จากผลการประเมินที่สำนักงานเขตพื้นที่การศึกษาได้ประเมินในปีการศึกษา 2549 ที่ผ่านมา ส่วนผลการประเมินของโรงเรียน และเขตพื้นที่การศึกษา จะเป็นตัวบ่งชี้คุณภาพการศึกษาขั้นพื้นฐานในภาพรวม และใช้เป็นข้อมูลประกอบการตัดสินใจในการกำหนดนโยบาย วางแผนในการพัฒนาคุณภาพการศึกษาระดับเขตพื้นที่การศึกษาและระดับสถานศึกษา ส่วนผลการประเมินนักเรียนทุกคนในชั้นประถมศึกษาปีที่ 2 ประถมศึกษาปีที่ 5 และมัธยมศึกษาปีที่ 2 เขตพื้นที่การศึกษาเป็นผู้ประเมิน จะเป็นข้อมูลสำคัญในการปรับปรุงเพื่อการพัฒนาตนเองของผู้เรียน และการจัดการเรียนการสอนของสถานศึกษาต่อไป

2.2 วัตถุประสงค์

1. เพื่อกำกับ ติดตาม และควบคุม คุณภาพการจัดการศึกษาขั้นพื้นฐานของประเทศ ในช่วงชั้นที่ 1 (ประถมศึกษาปีที่ 3) และช่วงชั้นที่ 3 (มัธยมศึกษาปีที่ 3) เพื่อให้เกิดการพัฒนาอย่างต่อเนื่อง ซึ่งเป็นส่วนหนึ่งของการประกันคุณภาพผู้เรียน

2. เพื่อให้ได้ข้อมูลย้อนกลับ สำหรับใช้ในกระบวนการตัดสินใจ และกำหนดแผนพัฒนา

คุณภาพการจัดการศึกษาขั้นพื้นฐานของประเทศ เขตพื้นที่การศึกษา และระดับสถานศึกษา

3. เพื่อให้เขตพื้นที่การศึกษามีข้อมูลผลสัมฤทธิ์นักเรียนชั้นประถมศึกษาปีที่ 2 ประถมศึกษาปีที่ 5 และมัธยมศึกษาปีที่ 2 เพื่อการปรับปรุงผู้เรียนเป็นรายบุคคล

2.3 ขอบเขตและแนวทางการประเมิน

สำนักงานเขตพื้นที่การศึกษา ประเมินนักเรียนทุกคนในชั้นประถมศึกษาปีที่ 2 ประถมศึกษาปีที่ 5 และมัธยมศึกษาปีที่ 2 โดยประเมินในสาระสำคัญที่ต้องเร่งปรับปรุงเพื่อพัฒนาผู้เรียนรายบุคคล ให้กำหนดสอบหลังจากส่วนกลาง 1 สัปดาห์

ที่	ระดับชั้น	สาระ	ข้อ	ผู้จัดทำ
1	ประถมศึกษาปีที่ 2	ภาษาไทย	30	สพท. ปัตตานี เขต 2
		คณิตศาสตร์	30	
รวม			60	
2	ประถมศึกษาปีที่ 5	ภาษาไทย	40	
		คณิตศาสตร์	40	
		ภาษาอังกฤษ	40	
รวม			120	
3	มัธยมศึกษาปีที่ 2	ภาษาไทย	40	
		คณิตศาสตร์	40	
		ภาษาอังกฤษ	40	
รวม			120	

2.4 คณะกรรมการดำเนินงาน

การประเมินคุณภาพการศึกษาขั้นพื้นฐาน เพื่อการประกันคุณภาพผู้เรียน ปีการศึกษา 2550 เป็นงานที่มีขอบข่ายกว้างขวาง เพราะเป็นการจัดสอบทุกเขตพื้นที่การศึกษา ทุกโรงเรียน ทุกคน ดังนั้นเพื่อให้การดำเนินงานการประเมินคุณภาพการศึกษาขั้นพื้นฐาน เพื่อการประกันคุณภาพผู้เรียน ปีการศึกษา 2550 เป็นไปด้วยความเรียบร้อยเช่นที่เคยปฏิบัติมา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐานจึงมีคำสั่ง แต่งตั้งคณะกรรมการดำเนินงานโครงการประเมินคุณภาพการศึกษาขั้นพื้นฐานเพื่อการประกันคุณภาพผู้เรียน ปีการศึกษา 2550 โดยกำหนดหน้าที่ไว้ดังนี้

2.4.1 คณะกรรมการอำนวยการ

คณะกรรมการอำนวยการ มีเลขาธิการคณะกรรมการการศึกษาขั้นพื้นฐาน เป็นที่ปรึกษา รองเลขาธิการคณะกรรมการการศึกษาขั้นพื้นฐาน (นายสมเกียรติ ชอบผล) เป็นประธานกรรมการ ผู้อำนวยการสำนักทดสอบทางการศึกษาเป็นกรรมการและเลขานุการ ผู้แทนหน่วยงานสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐานเป็นกรรมการ มีหน้าที่เสนอแนะ กำหนดรูปแบบ หลักการ แนวทางการประเมิน แนวทางการพัฒนาประสิทธิภาพการปฏิบัติงาน การพัฒนาคุณภาพและมาตรฐานการประเมิน ให้คำแนะนำ กำกับ ดูแล เร่งรัด ติดตาม ประเมินผลการปฏิบัติงานการประเมินคุณภาพการศึกษาระดับการศึกษาขั้นพื้นฐาน และมีอำนาจแต่งตั้งคณะกรรมการ (เพิ่มเติม) คณะอนุกรรมการหรือคณะทำงานเฉพาะเรื่องเพิ่มเติมตามความจำเป็นและเห็นสมควร เพื่อปฏิบัติงานตามหน้าที่ที่คณะกรรมการอำนวยการมอบหมาย

2.4.2 คณะกรรมการดำเนินงานส่วนกลางและเขตพื้นที่การศึกษา

คณะกรรมการดำเนินงานส่วนกลางและเขตพื้นที่การศึกษา มีรองเลขาธิการคณะกรรมการการศึกษาขั้นพื้นฐาน (นายสมเกียรติ ชอบผล) เป็นประธานกรรมการ ผู้อำนวยการสำนักงานเขตพื้นที่การศึกษา ผู้อำนวยการสำนักทุกสำนักในสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน หรือผู้แทน เป็นกรรมการ หัวหน้ากลุ่มพัฒนาการสอบ เป็นกรรมการและเลขานุการ มีหน้าที่กำหนดแผนดำเนินงานคุณภาพการศึกษา ระดับการศึกษาขั้นพื้นฐาน ในส่วนกลาง และเขตพื้นที่การศึกษาตามนโยบาย หลักการและแนวทางที่คณะกรรมการอำนวยการกำหนด ให้คำแนะนำ กำกับ ดูแลเร่งรัด ติดตามประเมินผลการปฏิบัติงานประเมินคุณภาพการศึกษาในส่วนกลาง และเขตพื้นที่การศึกษา ที่รับผิดชอบ แต่งตั้งคณะกรรมการ(เพิ่มเติม) คณะอนุกรรมการ คณะทำงานเฉพาะเรื่องตามความจำเป็นและเห็นสมควร เพื่อปฏิบัติงานตามหน้าที่ที่คณะกรรมการดำเนินงานส่วนกลาง และเขตพื้นที่การศึกษา มอบหมาย

2.5 การดำเนินงานของเขตพื้นที่การศึกษา

เพื่อให้การดำเนินงานประเมินคุณภาพการศึกษาขั้นพื้นฐาน เพื่อการประกันคุณภาพผู้เรียน ปีการศึกษา 2550 ของเขตพื้นที่การศึกษา มีมาตรฐานเดียวกัน ควรมีแนวดำเนินการประเมินดังนี้

1) จัดเตรียมแผน โครงการประเมิน ฯ สำรวจข้อมูลนักเรียนตามกลุ่มเป้าหมาย

2) แต่งตั้งคณะกรรมการดำเนินงานโครงการ ฯ ระดับ สพท. เพื่อกำหนดแผนดำเนินงานตามนโยบายหลักการ และแนวทางที่คณะกรรมการอำนวยการกำหนด

3) ร่วมประชุมกับสำนักทดสอบทางการศึกษา เพื่อจัดทำแผนการประเมินฯ

4) จัดประชุมคณะกรรมการดำเนินงานระดับเขตพื้นที่การศึกษาเพื่อชี้แจงรายละเอียด แผน ขั้นตอน และวิธีดำเนินการประเมิน

5) แต่งตั้งศูนย์ประสานการสอบภายในเขตพื้นที่การศึกษา พร้อมกำหนดบุคคลที่รับผิดชอบ ระบุขอบเขตและจำนวนสนามสอบ ที่จะต้องประสานการสอบระหว่างเขตพื้นที่การศึกษา กับสนามสอบ

6) แต่งตั้งกรรมการกำกับการสอบของสนามสอบ

7) แต่งตั้งคณะอนุกรรมการ/คณะทำงานที่จำเป็นและสำคัญเพื่อดำเนินงาน

8) ประชุมชี้แจงแนวปฏิบัติ และอำนาจหน้าที่ของ คณะกรรมการ/คณะทำงาน และคณะกรรมการกำกับการสอบ

9) ดำเนินการสอบตามตารางสอบ และวิธีการ ตามแนวดำเนินการงานโครงการประเมิน

10) ติดตาม ดูแล กำกับงาน ให้เป็นไปตามปฏิทินปฏิบัติงานอย่างมีประสิทธิภาพ

11) สรุปผล และรายงานผลการดำเนินงานของเขตพื้นที่การศึกษา ต่อสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน

2.6 การนำผลการประเมินไปใช้เพื่อการพัฒนาคุณภาพการศึกษา

การประเมินคุณภาพการศึกษาขั้นพื้นฐาน เพื่อการประกันคุณภาพผู้เรียน ปีการศึกษา 2550 มีจุดมุ่งหมายเพื่อนำผลการสอบไปใช้ในการพัฒนาการจัดการศึกษาให้มีคุณภาพและประสิทธิภาพสูงสุด ในการนำผลการประเมินไปใช้นั้น ได้วางแนวทางไว้ ดังนี้

ระดับกระทรวง

1. สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐานนำข้อมูลไปใช้ในการวางแผนพัฒนาหลักสูตรทั้งด้านเนื้อหาสาระและการจัดกิจกรรมที่เกี่ยวข้องกับหลักสูตร เช่น การจัดการเรียนการสอน การพัฒนาหนังสือและสื่อต่าง ๆ ทั้งนี้เพื่อให้การใช้หลักสูตรบรรลุเป้าหมายตามที่กำหนดไว้

2. สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐานเสนอผลการประเมินให้หน่วยงานต้นสังกัดของโรงเรียน เขตพื้นที่การศึกษาได้ทราบ เพื่อนำข้อมูลไปใช้วางแผนพัฒนาการจัดการศึกษาในส่วนที่รับผิดชอบ โดยหน่วยงานต่าง ๆ จะได้รับทราบผลการประเมิน ทั้งในรูปผลการวิเคราะห์ข้อมูลด้วยคอมพิวเตอร์และรายงานในรูปของเอกสาร

3. สำนักทดสอบทางการศึกษา จัดการประชุมสัมมนาสรุปผล และนำผลการสอบไปใช้เพื่อการพัฒนา โดยให้มีการแลกเปลี่ยนความคิดเห็น ระดมความคิดจากผู้บริหารโรงเรียน เขตพื้นที่

การศึกษา ตลอดจนผู้มีส่วนเกี่ยวข้องทุก ๆ ฝ่ายและผู้ทรงคุณวุฒิ เพื่อให้เกิดการนำผลการสอบไปใช้วางแผนพัฒนาการศึกษาให้คุ้มค่าที่สุด

ระดับเขตพื้นที่การศึกษา

1. จากการกระจายอำนาจการจัดการศึกษาไปสู่ท้องถิ่น หน่วยงานในระดับท้องถิ่นเขตพื้นที่การศึกษา จำเป็นจะต้องมีข้อมูลพื้นฐานประกอบการวางแผนและพัฒนางาน สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน จึงมีนโยบายที่จะส่งเสริมและกระตุ้นให้หน่วยงานระดับดังกล่าวจัดทำรายงานผลการสอบไปใช้วางแผนพัฒนาการจัดการศึกษาในระดับนั้นๆ

2. การประเมินคุณภาพการศึกษาขั้นพื้นฐานครั้งนี้ เป็นการจัดสอบในสถานศึกษา ทั้งประเทศ ข้อมูลที่ได้จึงเป็นสภาพการดำเนินการจัดการศึกษาในภาพรวมทั้งประเทศ การนำผลการสอบไปใช้ในการพัฒนาการเรียนการสอนจึงเป็นสิ่งจำเป็น สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน จึงกระตุ้นให้หน่วยงานระดับปฏิบัติ คือ โรงเรียน มีระบบการประเมินการใช้หลักสูตรด้วยตนเอง เพื่อนำข้อมูลไปปรับปรุงการดำเนินงานของโรงเรียนได้ตลอดเวลาและทันท่วงที ข้อมูลที่ได้จากการประเมินทั้ง 2 ลักษณะนี้ โรงเรียนและหน่วยงานที่เกี่ยวข้องควรนำมาประกอบกัน เพื่อนำไปใช้ในการพัฒนาการจัดการศึกษาให้มีคุณภาพดียิ่ง ๆ ขึ้นไป

ระดับสถานศึกษา

1. จากบริบทของการประกันคุณภาพการศึกษา โรงเรียนนำข้อมูลผลการสอบไปวางแผนพัฒนาคุณภาพการศึกษาของโรงเรียน (School improvement plan : SIP) ใช้เป็นข้อมูล เพื่อการปรับปรุง พัฒนาหลักสูตร การเรียนการสอน กำหนดเป้าหมายในการพัฒนาผลสัมฤทธิ์ของนักเรียนและจัดทำรายงานต่อสาธารณชน/ชุมชน คณะกรรมการสถานศึกษา เพื่อแสดงความรับผิดชอบและแสดงศักยภาพของโรงเรียนในการจัดการศึกษา

2. ครูผู้สอนนำข้อมูลผลการจัดสอบของนักเรียน มาศึกษา วิเคราะห์เปรียบเทียบเพื่อปรับปรุงแก้ไข พัฒนาวิธีการจัดการเรียนการสอน กำหนดจุดหมาย และเกณฑ์ที่จะทำให้นักเรียนมีผลสัมฤทธิ์สูงขึ้น

3. งานวิจัยที่เกี่ยวข้อง

งานวิจัยภายในประเทศ

เรวดี อินทสะระ (2539) ได้ศึกษาความเที่ยงเชิงพยากรณ์ของแบบทดสอบคัดเลือกที่วิเคราะห์ความลำเอียงต่อเพศด้วยวิธีใช้ทฤษฎีการตอบสนองข้อสอบ วิธีแมนเทล-แฮนส์เซล และวิธีซิปเทสท์ การตัดสินผลการสอบที่คิดคะแนนมาตรฐานทีปกติและคะแนนน้ำหนักความสามารถและสาเหตุความลำเอียงของข้อสอบ โดยศึกษาความลำเอียงของข้อสอบคัดเลือกเข้าศึกษาในชั้นปีที่ที่ 1 ประเภทรับตรง ปีการศึกษา 2538 ของมหาวิทยาลัยสงขลานครินทร์ในวิชาภาษาไทย ก วิชาสังคมศึกษา ก วิชาภาษาอังกฤษ กข วิชาละ 8,127 คน ชาย 2,722 คน หญิง 5,405 คน วิชาภาษาไทย กข วิชาสังคมศึกษา กข และวิชาภาษาอังกฤษ กขค วิชาละ 5,415 คน ชาย 1,454 คน หญิง 3,961 คน ความตรงเชิงพยากรณ์ศึกษาจากคะแนนสอบคัดเลือกกับเกรดภาคเรียนที่ 1 ปีการศึกษา 2538 ของนักเรียนที่ได้รับการคัดเลือกจากประเภทรับตรง สายวิทยาศาสตร์ 763 คน และสายศิลปะศึกษาศาสตร์ 281 คน และสาเหตุความลำเอียงของข้อสอบจากการระบุสาเหตุของนักวัดผลการศึกษาหรืออาจารย์ผู้สอน จำนวน 50 คน และนักศึกษาที่เรียนในสาขาวิชานั้น ๆ วิชาละ 30 คน

ผลการศึกษา พบว่า วิธีการการตรวจสอบความลำเอียงทั้ง 3 วิธีตัดสินจำนวนข้อสอบที่ลำเอียงแตกต่างกันในวิชาภาษาไทย ก ฉบับที่ 2 และวิชาสังคมศึกษา ก ฉบับที่ 1 ที่ระดับนัยสำคัญทางสถิติที่ .05 นอกจากนั้นต่างกันที่ระดับนัยสำคัญทางสถิติที่ .01 โดยวิธีที่ใช้ทฤษฎีตอบสนองข้อสอบ ตัดสินจำนวนข้อสอบที่ลำเอียงมากที่สุด ความสัมพันธ์ของลำดับที่ของการสอบไม่ว่าจะคิดคะแนนมาตรฐานทีปกติ หรือคิดคะแนนน้ำหนักความสามารถและใช้ข้อสอบจำนวนทั้งหมดหรือใช้เฉพาะข้อสอบที่ปราศจากความลำเอียงต่างมีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01

ประสิทธิภาพในการทำนายผลสัมฤทธิ์ทางการเรียนสายวิทยาศาสตร์ การคิดคะแนนน้ำหนักความสามารถทั้งใช้ข้อสอบจำนวนทั้งหมดและใช้เฉพาะข้อสอบที่ปราศจากความลำเอียงมีประสิทธิภาพในการทำนายสูงกว่าการคิดคะแนนมาตรฐานทีปกติ ส่วนสายศิลปะศาสตร์ การคิดคะแนนมาตรฐานปกติที ทั้งที่ใช้ข้อสอบจำนวนทั้งหมดและใช้เฉพาะข้อสอบที่ปราศจากความลำเอียงมีประสิทธิภาพในการทำนายสูงกว่าการคิดคะแนนน้ำหนักความสามารถและสาเหตุของความลำเอียงของข้อสอบต่อเพศทั้งชายและหญิงเกิดจากข้อสอบเป็นคำถามที่ผู้สอบได้รับการฝึกฝนเฉพาะจะมีโอกาสตอบถูกมากกว่าเป็นเรื่องราวที่กลุ่มนั้น ๆ สนใจและเป็นข้อสอบที่ถามความจำ

รัตนาพร วงศ์ช่วย (2541) ได้ศึกษาเปรียบเทียบผลการวิเคราะห์ความลำเอียงของข้อสอบด้วยวิธีไคสแควร์และวิธีแปลงค่าความยากของข้อสอบ โดยทำการทดสอบความแตกต่างของ

จำนวนข้อสอบความแตกต่างของจำนวนข้อสอบที่ลำเอียงด้วยวิธีไคสแควร์ กลุ่มตัวอย่าง คือ ผู้เข้าสอบคัดเลือกเพื่อเข้าศึกษาต่อระดับปริญญาตรีของสถาบันราชภัฏฉะเชิงเทรา ปีการศึกษา 2541 จำนวน 2,066 คน ผลกการวิจับพบว่า วิธีไคสแควร์พบข้อสอบที่ลำเอียงตามเพศ 65 ข้อ ตามเขตที่ตั้งของสถานศึกษา 29 ข้อ และตามสังกัดของสถานศึกษา 65 ข้อ วิธีแปลงค่าความยากของข้อสอบ พบข้อสอบที่ลำเอียงตามเพศ 41 ข้อ ตามเขตที่ตั้งของสถานศึกษา 14 ข้อ และตามสังกัดของสถานศึกษา 44 ข้อ โดยวิธีไคสแควร์มีจำนวนข้อสอบที่ลำเอียงมากกว่าวิธีแปลงค่าความยากของข้อสอบอย่างมีนัยสำคัญทางสถิติที่ระดับ .05

นพมาศ พิพัฒนสุข (2541) ได้เปรียบเทียบประสิทธิภาพระหว่างวิธีแมนเทล-แฮนส์เซลกับวิธีการถดถอยโลจิสติก ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อใช้เกณฑ์จับคู่เปรียบเทียบแตกต่างกันในแบบทดสอบหลายมิติ โดยใช้เกณฑ์การจับคู่ 3 เกณฑ์ คือ คะแนนรวมของแบบทดสอบ (Total test score) คะแนนแบบทดสอบชุดย่อย (Subtest score) และคะแนนแบบทดสอบหลายชุดย่อย (Multiple subtest scores) เครื่องมือที่ใช้ในการเก็บรวบรวมข้อมูลเป็นแบบทดสอบเลือกตอบชนิด 4 ตัวเลือก ประกอบด้วยแบบทดสอบชุดย่อย 2 ฉบับๆ ละ 40 ข้อ ซึ่งแต่ละฉบับวัดองค์ประกอบ 4 ด้าน คือ ความรู้ความเข้าใจทางคณิตศาสตร์ ความสามารถด้านการคิดคำนวณ ความสามารถเกี่ยวกับการพิจารณาผลลัพธ์อย่างสมเหตุสมผล ความสามารถด้านการแก้ปัญหา กลุ่มตัวอย่างเป็นนักเรียนชั้นประถมปีที่ 6 สังกัดสำนักงานการประถมศึกษากรุงเทพมหานคร จำนวน 1,076 คน การวิเคราะห์ข้อมูลประกอบด้วยการวิเคราะห์องค์ประกอบเชิงยืนยันเพื่อตรวจสอบความเที่ยงตรงเชิงโครงสร้างโดยใช้โปรแกรม LISREL การวิเคราะห์ด้วยวิธีซิปเทสท์ (เป็นวิธีเกณฑ์สำหรับการเปรียบเทียบ) และวิธีแมนเทล-แฮนส์เซล ภายใต้เกณฑ์คะแนนรวมและคะแนนแบบทดสอบชุดย่อยใช้โปรแกรม SIBTEST ส่วนวิธีการถดถอยโลจิสติกภายใต้คะแนนรวม คะแนนแบบทดสอบชุดย่อย และคะแนนแบบทดสอบหลายชุดย่อยใช้โปรแกรม SPSS แล้วคำนวณอัตราความถูกต้องและอัตราความคลาดเคลื่อนของวิธีแมนเทล-แฮนส์เซลและวิธีการถดถอยโลจิสติก จากนั้นทดสอบความแตกต่างโดยใช้สถิติ Z ผลการศึกษาพบว่า

1. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมนเทล-แฮนส์เซลเมื่อใช้คะแนนรวมและคะแนนแบบทดสอบชุดย่อยเป็นเกณฑ์การจับคู่เปรียบเทียบสามารถตรวจพบข้อสอบทำหน้าที่เบี่ยงเบนจำนวน 20% และ 18.67 % ตามลำดับ ส่วนวิธีการถดถอยโลจิสติก เมื่อใช้คะแนนรวม คะแนนแบบทดสอบชุดย่อย และคะแนนแบบทดสอบหลายชุดย่อยเป็นเกณฑ์การจับคู่สามารถตรวจพบข้อสอบทำหน้าที่ต่างกัน จำนวน 26.67 %, 22.67 % และ 17.33 % ตามลำดับภายใต้เงื่อนไขเกณฑ์การจับคู่โดยใช้คะแนนรวม

2. วิธีแมนเทล-แฮนส์เซลมีประสิทธิภาพสูงกว่าวิธีการถดถอยโลจิสติก และเมื่อใช้คะแนนแบบทดสอบชุดย่อยทั้งสองวิธีมีประสิทธิภาพไม่แตกต่างกัน

3. วิธีการถดถอยโลจิสติกมีความเหมาะสมในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบหลายมิติ เมื่อ ใช้คะแนนแบบทดสอบหลายชุดย่อยเป็นเกณฑ์การจับคู่ความสามารถ

สมศักดิ์ จั่นผ่อง (2542) ได้ศึกษาการเปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยใช้วิธีวิเคราะห์และขนาดกลุ่มผู้สอบต่างกัน ซึ่งใช้วิธีวิเคราะห์ 3 วิธี คือ วิธีโค้งลักษณะข้อสอบที่มี 3 พารามิเตอร์ วิธีแมนเทล-แฮนส์เซล และวิธีซิปเทสท์ กับขนาดกลุ่มผู้สอบ 2 ขนาด คือ 600 และ 1000 คน พบว่า วิธีซิปเทสท์ ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันได้มากที่สุด รองลงมาคือ วิธีแมนเทล-แฮนส์เซล และวิธีโค้งลักษณะข้อสอบ 3 พารามิเตอร์ โดยจำนวนข้อสอบที่ทำน้าที่ต่างกันทั้ง 3 วิธี แตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ เมื่อพิจารณาภายในวิธีเดียวกันแต่ต่างขนาดผู้สอบ พบว่า วิธีซิปเทสท์และวิธีแมนเทล-แฮนส์เซล มีจำนวนข้อสอบที่ทำหน้าที่ต่างกันอย่างมีนัยสำคัญ ส่วนวิธีโค้งลักษณะข้อสอบ 3 พารามิเตอร์จำนวนข้อสอบทำหน้าที่ต่างกันไม่แตกต่างกัน และ ค่าความสอดคล้องของวิธีวิเคราะห์ทั้ง 3 วิธี และขนาดผู้สอบ กับจำนวนข้อสอบที่ระบุว่าทำหน้าที่ต่างกันมีค่าอยู่ระหว่าง .60 - .85 และค่าความสัมพันธ์ของดัชนีการทำหน้าที่ต่างกันด้วยวิธีวิเคราะห์ 3 วิธี เมื่อพิจารณาจากกลุ่มผู้สอบ มีค่า ระหว่าง .10821 - .9249

ทองอยู่ สาระ (2543) ได้ศึกษาเปรียบเทียบอำนาจจำแนกการตรวจสอบ และการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูปและแบบอเนกรูป ระหว่างวิธีแมนเทล-แฮนส์เซล กับวิธีถดถอยโลจิสติก โดยใช้ความยาวของแบบทดสอบและขนาดกลุ่มตัวอย่างแตกต่างกัน กลุ่มตัวอย่างที่ใช้ในการศึกษาครั้งนี้เป็นนักเรียนชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2542 ในโรงเรียนสังกัดกรมสามัญศึกษา ส่วนกลาง กลุ่มที่ 3 จำนวน 3,242 คน เครื่องมือที่ใช้เป็นแบบทดสอบวัดความสามารถทางสมองที่ผู้วิจัยสร้างขึ้นตามแนวโครงสร้างของโอติส-เลนนอน ซึ่งเป็นแบบทดสอบเลือกตอบห้าตัวเลือก จำนวน 80 ข้อ วัดความสามารถทั่วไปสามด้าน คือ ความเข้าใจด้านภาษา เหตุผลด้านภาษา และเหตุผลด้านภาพ

ผลการศึกษาพบว่า

1. อำนาจการตรวจสอบ และการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันทั้งแบบเอกรูปและแบบอเนกรูป ระหว่างวิธีแมนเทล-แฮนส์เซล และวิธีถดถอยโลจิสติก ภายใต้ความยาวแบบทดสอบและขนาดกลุ่มตัวอย่างที่ศึกษาเกือบทุกเงื่อนไข มีค่าไม่แตกต่างกัน

2. ความยาวของแบบทดสอบไม่มีผลต่ออำนาจการตรวจสอบและการจำแนกผิดพลาดในการตรวจสอบด้วยวิธีวิธีแมนเทล-แฮนส์เซล และวิธีถดถอยโลจิสติก ทั้งการตรวจสอบที่ทำหน้าที่ต่างกันแบบเอกรูปและแบบอเนกรูป

3. ขนาดกลุ่มตัวอย่างมีผลต่ออำนาจการตรวจสอบ ในการตรวจสอบด้วยวิธีแมนเทล-แฮนส์เซล และวิธีถดถอยโลจิสติก เกือบทุกเงื่อนไขของการศึกษาทั้งการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูปและแบบอเนกรูป กล่าวคือ เมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้น อำนาจการตรวจสอบจะมีค่าเพิ่มขึ้น แต่พบว่าขนาดกลุ่มตัวอย่างไม่มีผลต่อการจำแนกผิดพลาด ในเกือบทุกเงื่อนไขที่ศึกษา ทั้งการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูปและแบบอเนกรูป

อารี วัชรโสตถิกุล (2543) ได้ศึกษาเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยใช้รูปแบบต่างกัน คือ รูปแบบคะแนนรวมทั้งฉบับ แยกตามเนื้อหา และแยกตามระดับพฤติกรรม ด้วยวิธีการตรวจสอบต่างกัน คือวิธีซิปเทสท์และวิธีถดถอยโลจิสติก แล้วทำการคัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบ เพื่อเปรียบเทียบค่าความเที่ยง กลุ่มตัวอย่าง เป็นนักเรียนระดับชั้นมัธยมศึกษาปีที่ 3 สังกัดกรมสามัญศึกษา กรุงเทพมหานคร ส่วนกลาง กลุ่ม 4 จำนวน 994 คน จำแนกเป็นนักเรียนชาย 480 คน และนักเรียนหญิง 514 คน เครื่องมือที่ใช้เป็นแบบทดสอบวัดผลสัมฤทธิ์วิชาคณิตศาสตร์ ชนิดเลือกตอบ 5 ตัวเลือก จำนวน 5 ข้อ ผลการศึกษาพบว่า

1. จำนวนข้อสอบที่ทำหน้าที่ต่างกันโดยใช้วิธีซิปเทสท์และวิธีถดถอยโลจิสติก แตกต่างกันอย่างมีมีนัยสำคัญทางสถิติ

2. จำนวนข้อสอบที่ทำหน้าที่ต่างกันโดยใช้วิธีซิปเทสท์และวิธีถดถอยโลจิสติกแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ในรูปแบบรวมทั้งฉบับ และแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติในรูปแบบ แยกตามเนื้อหา และรูปแบบแยกตามระดับพฤติกรรม

3. ความเที่ยงในแบบทดสอบหลังคัดข้อสอบที่ทำหน้าที่ต่างกันออกโดยใช้แบบรูปการตรวจสอบต่างกัน แล้วทำการปรับขยายค่าความเที่ยงของแบบทดสอบให้มีจำนวนข้อเท่ากันด้วยสูตรสเปียร์แมน-บราวน์แล้ว วิธีซิปเทสท์ พบว่า ค่าความเที่ยงแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ.05 ทดสอบรายคู่พบว่า ความเที่ยง แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ระหว่างรูปแบบแยกตามเนื้อหา และแยกตามระดับพฤติกรรม เมื่อตรวจสอบ โดยวิธีถดถอยโลจิสติก พบว่าค่าความเที่ยงแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

4. ความเที่ยงของแบบทดสอบหลังคัดเลือกข้อสอบที่ทำหน้าที่ต่างกันออกโดยใช้รูปแบบการตรวจสอบต่างกัน แล้วทำการปรับขยายค่าความเที่ยงของแบบทดสอบให้มีจำนวนข้อเท่ากันด้วยสูตรสเปียร์แมน-บราวน์ แล้ว เมื่อตรวจสอบโดยใช้รูปแบบคะแนนรวมทั้งฉบับพบว่า ค่าความ

เที่ยงแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 และเมื่อตรวจสอบโดยใช้รูปแบบแยกตามเนื้อหาพบว่า ค่าความเที่ยงแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 และเมื่อตรวจสอบโดยใช้รูปแบบแยกตามระดับพฤติกรรม พบว่า ค่าความเที่ยงแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

วิภา จำมั่น (2544) ได้ศึกษาเพื่อเปรียบเทียบผลการตรวจสอบข้อสอบที่ลำเอียงของแบบทดสอบวัดความสามารถด้านภาษา เมื่อตรวจสอบด้วยวิธีเซปเทสท์กับวิธีแมนเทล-แฮนส์เซล ในกลุ่มข้อสอบที่มีระดับความง่ายต่ำ กลุ่มข้อสอบที่มีระดับความง่ายปานกลางและกลุ่มข้อสอบที่มีความง่ายสูง โดย เปรียบเทียบข้อสอบที่มีความลำเอียงและค่าความเที่ยงของแบบทดสอบหลังจากคัดเลือกข้อสอบที่มีความลำเอียงออกแล้วกลุ่มตัวอย่างเป็นนักเรียนชั้นมัธยมศึกษาปีที่ 3 ภาคเรียนที่1 ปีการศึกษา 2544 ของโรงเรียนสังกัดคณะกรรมการศึกษาขั้นพื้นฐาน จังหวัดลพบุรี จำนวน 1,401 คน ซึ่งได้มาโดยการสุ่มแบบแบ่งชั้น ผลการศึกษา ปรากฏว่า

1. ข้อสอบที่มีความลำเอียง เมื่อตรวจสอบด้วยวิธีซิปเทสท์ ระหว่างกลุ่มข้อสอบที่มีระดับความง่ายต่ำกว่า กลุ่มข้อสอบที่มีระดับความง่ายปานกลาง และกลุ่มข้อสอบที่มีความง่ายสูงมีจำนวนข้อแตกต่างกันอย่างมีนัยสำคัญทางสถิติ .05 ส่วนข้อสอบที่มีความลำเอียงระหว่างกลุ่มข้อสอบที่มีระดับความง่ายต่ำกับกลุ่มข้อสอบที่มีระดับความง่ายปานกลาง และระหว่างกลุ่มข้อสอบที่มีระดับความง่ายปานกลางเทียบกับกลุ่มข้อสอบที่มีความง่ายสูง มีจำนวนข้อข้อแตกต่างกันอย่างมีนัยสำคัญทางสถิติเมื่อตรวจสอบด้วยวิธีแมนเทล-แฮนส์เซล พบว่า ข้อสอบที่มีความลำเอียงระหว่างกลุ่มผู้สอบที่มีระดับความง่ายต่ำ กลุ่มข้อสอบที่มีระดับความง่ายปานกลาง และกลุ่มข้อสอบที่มีความง่ายสูง มีจำนวนข้อแตกต่างกันอย่างไม่มีนัยสำคัญ

2. จำนวนข้อสอบที่มีความลำเอียงระหว่างการตรวจสอบด้วยวิธีแมนเทล-แฮนส์เซล ในกลุ่มข้อสอบที่มีระดับความง่ายต่ำ กลุ่มข้อสอบที่มีระดับความง่ายปานกลางและกลุ่มข้อสอบที่มีความง่ายสูง วิธีซิปเทสท์กับวิธีแมนเทล-แฮนส์เซล มีค่าแตกต่างกันอย่างไม่มีนัยสำคัญ

3. ค่าความเที่ยงของแบบทดสอบหลังจากคัดเลือกข้อสอบที่มีความลำเอียงออกแล้วระหว่างกลุ่มข้อสอบที่มีระดับง่ายต่ำ กลุ่มข้อสอบที่มีระดับความง่ายปานกลาง และกลุ่มข้อสอบที่มีความง่ายสูง วิธีซิปเทสท์กับวิธีแมนเทล- แฮนส์เซล มีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

สิริรัตน์ วิภาสศิลป์ (2545) ได้เปรียบเทียบวิธีซิปเทสท์และวิธีดีเอฟไอทีในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ หมวดข้อสอบและแบบทดสอบ จากข้อมูลการตอบข้อสอบที่ใช้ความสามารถหลายมิติ ภายใต้ปัจจัยที่ศึกษา 2 ปัจจัย คือ (1) ความยาวของแบบทดสอบ 3 ขนาด คือ 30, 40 และ 50 ข้อ และ (2) ขนาดตัวอย่าง 5 ขนาด คือ 50, 100, 200, 500 และ 1,000 คน กลุ่มตัวอย่างที่ใช้ในการศึกษาได้มาจากการสุ่มแบบใส่คืนจากประชากรเทียม ซึ่งกำหนดจากนักเรียนชายและนักเรียนหญิง ชั้นมัธยมศึกษาปีที่ 1 จังหวัดนนทบุรี โดยสุ่มแต่ละขนาด 50 ครั้ง เครื่องมือ

ที่ใช้ในการวิจัยเป็นแบบทดสอบวิชาคณิตศาสตร์ ชั้นมัธยมศึกษาปีที่ 1 ที่ผู้วิจัยสร้างขึ้น ประกอบด้วยข้อสอบแบบเลือกตอบชนิด 5 ตัวเลือก จำนวน 50 ข้อ โดยมุ่งวัดความสามารถหลายมิติ โดยให้ความสามารถทางคณิตศาสตร์เป็นความสามารถหลัก ส่วนความสามารถด้านอื่นๆ เป็นความสามารถรอง เช่นความสามารถในการอ่าน การแก้โจทย์ปัญหาที่ซับซ้อน ไหวพริบในการทำข้อสอบ เป็นต้น แล้วนำข้อสอบดังกล่าวให้ผู้เชี่ยวชาญตรวจสอบ ผลปรากฏว่า เป็นข้อสอบที่แสดงการทำหน้าที่ต่างกันของข้อสอบต่อเพศชาย จำนวน 16 ข้อ จากนั้นจึงเก็บรวบรวมข้อมูลแล้วคัดเลือกข้อสอบตามสัดส่วนในตารางกำหนดข้อสอบเพื่อจัดเป็นแบบทดสอบที่มีความยาว 40 และ 30 ข้อ ในการวิเคราะห์มิติของแบบทดสอบจะวิเคราะห์องค์ประกอบเพื่อพิจารณาค่าไอเกนโดยใช้โปรแกรม SPSS ส่วนการประมาณค่าพารามิเตอร์ของข้อสอบภายใต้โมเดลแบบ 2 พารามิเตอร์ (M2PL) ใช้โปรแกรม NOHARM การปรับเทียบมาตรใช้โปรแกรม IPLINK การประมาณค่าความสามารถหลักใช้โปรแกรม BILOG สำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบหมวดข้อสอบ และแบบทดสอบใช้โปรแกรม SIBTEST และ DFIT แล้วนำผลการตรวจสอบไปวิเคราะห์อัตราความถูกต้องและอัตราความคลาดเคลื่อนประเภทที่ 1 ของการตรวจสอบด้วยวิธีเดียวกันและวิธีต่างกัน จากนั้นจึงเปรียบเทียบความแตกต่างโดยใช้สถิติ MANOVA และ Z-test ผลการศึกษาในเงื่อนไขของข้อสอบพบว่า ปัจจัยความยาวของแบบทดสอบ 30, 40 และ 50 ข้อ และปัจจัยขนาดตัวอย่าง 50, 100 และ 200 คน ไม่มีผลต่ออัตราความถูกต้องของวิธีซิปเทสท์ ส่วนอัตราความถูกต้องของวิธีซิปเทสท์ เมื่อขนาดตัวอย่าง 500 และ 1,000 คน มีค่าสูงกว่าการตรวจสอบในขนาดตัวอย่าง 50, 100 และ 200 คน ส่วนอัตราความคลาดเคลื่อนประเภทที่ 1 ก็มีค่าสูงกว่าด้วย สำหรับการตรวจสอบด้วยวิธีดีเอฟไอทีพบว่า ปัจจัยขนาดตัวอย่างทั้ง 5 ขนาดไม่มีต่ออัตราความถูกต้องของวิธีดีเอฟไอที เมื่อเปรียบเทียบระหว่างวิธีซิปเทสท์และวิธีดีเอฟไอที พบว่า อัตราความถูกต้องของวิธีซิปเทสท์มากกว่าวิธีดีเอฟไอทีในทุกเงื่อนไข และความสอดคล้องในการตรวจสอบด้วยวิธีทั้งสองมีค่าต่ำกว่าร้อยละ 10 ส่วนผลการศึกษาในเงื่อนไขของหมวดข้อสอบพบว่า อัตราความถูกต้องของวิธีซิปเทสท์มากกว่าวิธีดีเอฟไอทีใน 2 เงื่อนไข คือ ความยาวของแบบทดสอบ 30 ข้อและขนาดตัวอย่าง 1,000 คน กับความยาวของแบบทดสอบ 40 ข้อและขนาดตัวอย่าง 500 คน ส่วนผลการศึกษาในเงื่อนไขของแบบทดสอบพบว่า อัตราความถูกต้องของวิธีซิปเทสท์มากกว่าวิธี ดีเอฟไอทีเมื่อความยาวของแบบทดสอบ 50 ข้อและขนาดตัวอย่าง100, 200 และ 1,000 คน

สุกัญญา ทองนาค (2549) ได้ศึกษาการวิเคราะห์ความลำเอียงของข้อสอบเข้าศึกษาต่อประเภทโควตา มหาวิทยาลัยเชียงใหม่ วิชาภาษาไทย วิชาสังคมและวิชาสามัญ 1 จำแนกตามกลุ่มเพศ ที่ตั้งและขนาดของโรงเรียนโดยวิธีการวิเคราะห์ 3 วิธี คือ วิธีวิเคราะห์ความแปรปรวน วิธีลอร์ดไคสแควร์ และวิธีแมนเทล -แฮนส์เซล เพื่อเปรียบเทียบจำนวนข้อสอบของแบบทดสอบ

ที่มีความลำเอียงและเพื่อวิเคราะห์ความสอดคล้องของความลำเอียงของข้อสอบระว่างวิธีทั้ง 3 วิธี ผลการวิจัยพบว่า

1. ผลการวิเคราะห์ความลำเอียงของข้อสอบแต่ละวิชา จากการวิเคราะห์โดยวิธีวิเคราะห์ 3 วิธี ได้ข้อสรุปดังนี้

1.1 ข้อสอบวิชาภาษาไทย จำนวน 100 ข้อ เมื่อวิเคราะห์ความลำเอียงจำแนกตามกลุ่มเพศ ปรากฏว่าข้อสอบมีความลำเอียงจำนวน 24 – 65 ข้อ เมื่อจำแนกตามกลุ่มที่ตั้งของโรงเรียน ข้อสอบมีความลำเอียง จำนวน 24 – 65 ข้อ และเมื่อจำแนกตามกลุ่มขนาดของโรงเรียนข้อสอบมีความลำเอียงจำนวน 16 – 65 ข้อ

1.2 ข้อสอบวิชาสังคม จำนวน 100 ข้อ เมื่อวิเคราะห์ความลำเอียงจำแนกตามกลุ่มเพศ ปรากฏว่าข้อสอบมีความลำเอียงจำนวน 15 – 55 ข้อ เมื่อจำแนกตามกลุ่มที่ตั้งของโรงเรียน ข้อสอบมีความลำเอียง จำนวน 40 – 63 ข้อ และเมื่อจำแนกตามกลุ่มขนาดของโรงเรียนข้อสอบมีความลำเอียงจำนวน 49 – 62 ข้อ

1.3 ข้อสอบวิชาสามัญ 1 จำนวน 100 ข้อ เมื่อวิเคราะห์ความลำเอียงจำแนกตามกลุ่มเพศ ปรากฏว่าข้อสอบมีความลำเอียงจำนวน 25 – 54 ข้อ เมื่อจำแนกตามกลุ่มที่ตั้งของโรงเรียน ข้อสอบมีความลำเอียง จำนวน 48 – 80 ข้อ และเมื่อจำแนกตามกลุ่มขนาดของโรงเรียนข้อสอบมีความลำเอียงจำนวน 35 – 74 ข้อ

2. ผลการเปรียบเทียบจำนวนข้อของแบบทดสอบที่มีความลำเอียงโดยใช้วิธีวิเคราะห์ความแปรปรวน วิธีแมนเทล-แฮนส์เซล และวิธีลอร์ดไคสแควร์ เมื่อจำแนกตามกลุ่มเพศ ที่ตั้งของโรงเรียน และขนาดของโรงเรียน ในข้อสอบวิชาภาษาไทย วิชาสังคม และวิชาสามัญ 1 พบว่าวิธีวิเคราะห์ความลำเอียงทั้ง 3 วิธี ให้ผลจำนวนข้อของแบบทดสอบที่ลำเอียงต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 โดยที่วิธีแมนเทล-แฮนส์เซล ตรวจพบจำนวนข้อสอบที่ลำเอียงในแต่ละวิชาสูงที่สุด รองลงมาคือ วิธีลอร์ดไคสแควร์

3. ผลการวิเคราะห์ความสอดคล้องของการวิเคราะห์ความลำเอียงของข้อสอบระหว่างวิธีวิเคราะห์ความแปรปรวน วิธีแมนเทล-แอนส์เซล และวิธีลอร์ดไคสแควร์ พบว่า วิเคราะห์โดยวิธีแมนเทล-แฮนส์เซลกับวิธีลอร์ดไคสแควร์มีความสอดคล้องกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 เมื่อจำแนกตามกลุ่มเพศ ที่ตั้งของโรงเรียน และขนาดของโรงเรียน ในข้อสอบวิชาภาษาไทย วิชาสังคมและวิชาสามัญ 1

งานวิจัยต่างประเทศ

เฟร็นซ์และมิลเลอร์ (French; & Miller. 1996: 315-332 อ้างถึงใน อรินทร์ อ่วมถนอม, 2549 : 93-94) ได้ใช้วิธีการถดถอยโลจิสติก ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่เป็นรูปแบบเดียวกัน (Uniform DIF) และที่ไม่เป็นรูปแบบเดียวกัน (Nonuniform DIF) โดยใช้ข้อมูลจำลองภายใต้โมเดลพาเชียลเครดิตทั่วไป (Generalized Partial Credit Model : GPCM) จำลองผลการตอบข้อสอบ 25 ข้อ ที่มี 4 รายการตอบ โดยให้คะแนนตั้งแต่ 0 ถึง 3 ในข้อสอบที่จำลองดังกล่าวมีเพียง 1 ข้อเป็นข้อสอบ DIF (ข้อ 25) ส่วนที่เหลืออีก 24 ข้อเป็นข้อสอบ No-DIF (ข้อ 1-24) การจำลองข้อสอบ DIF มี 4 เงื่อนไข คือ เงื่อนไข 1 ถึง 3 จำลองข้อสอบ Nonuniform DIF โดยกำหนดค่าพารามิเตอร์อำนาจจำแนก (*a*) มีค่าแปรเปลี่ยน ในเงื่อนไข 1, 2 และ 3 ของกลุ่มสนใจมีค่า *a* เท่ากับ 0.5 ทั้ง 3 เงื่อนไข ส่วนกลุ่มอ้างอิงมีค่า *a* เท่ากับ 1.0, 1.5 และ 2.0 ตามลำดับ ส่วนค่าพารามิเตอร์ความยาก (*b*) ในเงื่อนไข 1, 2 และ 3 มีค่าเท่ากันทั้ง 2กลุ่ม คือ $b1 = 0$, $b2 = -1$, $b3 = 0$ และ $b4 = 1$ สำหรับเงื่อนไข 4 จำลองข้อสอบ Uniform DIF โดยกำหนดพารามิเตอร์ความยาก (b) มีค่าแปรเปลี่ยน ในกลุ่มสนใจมีค่าพารามิเตอร์ความยาก $b1$, $b2$, $b3$ และ $b4$ เท่ากับ 0, -1, 0 และ 1 ตามลำดับ ส่วนกลุ่มอ้างอิงมีพารามิเตอร์ความยาก $b1$, $b2$, $b3$ และ $b4$ เท่ากับ 0, -2, 1 และ 2 ตามลำดับ ในการจำลองข้อสอบ No-DIF กำหนดให้พารามิเตอร์ของข้อสอบระหว่าง 2 กลุ่มมีค่าเท่ากัน การจำลองขนาดตัวอย่างมี 2 ขนาด คือ 500 คนและ 1,000 คน โดยมีการแจกแจงความสามารถแบบปกติ N(0, 1) ดังนั้นจะต้องจำลองข้อมูลทั้งหมด 4 เงื่อนไข (2 × 2) คือ รูปแบบของข้อสอบ DIF 2 รูปแบบ และขนาดตัวอย่าง 2 ขนาด โดยในแต่ละเงื่อนไขจำลองซ้ำ 100 ครั้ง แล้ววิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยประยุกต์ใช้โมเดลการถดถอยโลจิสติก 3 โมเดล คือ โมเดลโลจิตของอัตราส่วนแบบต่อเนื่อง (Continuation ratio logits) โมเดลโลจิตแบบสะสม (Cumulative logits) และโมเดลโลจิตของรายการคำตอบที่อยู่ติดกัน (Adjacent categories logits) โมเดลทั้งสามมีการลงรหัสข้อมูลในแบบแผนการวิเคราะห์แตกต่างกัน สำหรับข้อมูลที่ให้คะแนนหลายค่าซึ่งมีรายการตอบ 3 รายการจะลงรหัสข้อมูลสำหรับการวิเคราะห์ 3 โมเดล โดยที่โมเดลโลจิตแบบสะสมจะนำข้อมูลไปใช้วิเคราะห์ในทุกโมเดล ดังนั้นจึงไม่สูญเสียข้อมูล (Loss of data) ในแบบแผนการวิเคราะห์ของแต่ละโมเดล ส่วนการประมาณค่าพารามิเตอร์ของโมเดลการถดถอยโลจิสติกใช้วิธีความน่าจะเป็นมากสุดของ Newton-Raphson แล้วทดสอบนัยสำคัญด้วยสถิติไคสแควร์ ที่ระดับนัยสำคัญ .002 ผลการศึกษาพบว่า ขนาดกลุ่มตัวอย่าง ค่าพารามิเตอร์ของข้อสอบและการลงรหัสข้อมูลในแบบแผนของโมเดลการถดถอยโลจิสติก มีผลต่อความถูกต้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ในข้อสอบที่ให้คะแนนหลายค่า กล่าวคือ เมื่อขนาดตัวอย่างเพิ่มขึ้นจาก 500 คน เป็น 2,000 คน มีผลทำให้อัตราการตรวจสอบมีค่าเพิ่มขึ้น เมื่อเพิ่มขนาด

ตัวอย่างเป็น 2,000 คน วิธีโลจิตของอัตราส่วนแบบต่อเนื่องและวิธีโลจิตแบบสะสมมีอัตราความถูกต้องสูงสุด ภายใต้ทุกโมเดลการถดถอย ส่วนอัตราความถูกต้องของวิธีโลจิตของรายการคำตอบที่อยู่ติดกันมีค่าต่ำสุดในโมเดลการถดถอยที่หนึ่ง แต่ในโมเดลการถดถอยที่สองและสามยังคงมีความถูกต้องเพียงพอเมื่อขนาดตัวอย่างลดลงถึง 500 คน อัตราความถูกต้องในโมเดลแรกของวิธีโลจิตของอัตราส่วนแบบต่อเนื่องและวิธีโลจิตแบบสะสมมีค่าลดลง แต่วิธีโลจิตของรายการคำตอบที่อยู่ติดกันลดลงทุกโมเดล ดังนั้นวิธีโลจิตของอัตราส่วนแบบต่อเนื่องและวิธีโลจิตแบบสะสมให้ผลการตรวจสอบที่คล้ายกัน ซึ่งเป็นไปตามที่คาดหวัง ส่วนพารามิเตอร์อำนาจจำแนกของข้อสอบ เมื่อมีค่าเพิ่มขึ้นแล้วทำให้อัตราการตรวจสอบของทั้งสามวิธีมีค่าเพิ่มขึ้น และ ทั้งสามวิธีสามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่เป็นรูปแบบเดียวกัน และที่ไม่เป็นรูปแบบเดียวกัน

ซวิค และคณะ (Zwick et al., 1997, หน้า 321 – 344) ได้ศึกษาการตรวจสอบการทำหน้าที่ต่างกันของแบบทดสอบที่มีการตรวจให้คะแนนแบบหลายค่า โดยใช้วิธีวิเคราะห์ 5 วิธี คือ SMD จำนวน 2 วิธี วิธีแมนเทล-แฮนส์เซล และวิธีซิปเทสท์ 2 วิธี ได้แก่ วิธีซิปเทสท์ และวิธีซิปเทสท์ปรับใหม่ โดยใช้เงื่อนไขแบบทดสอบที่มีการตรวจสอบให้คะแนนแบบสองค่า (0,1) จำนวน 50 ข้อ และแบบทดสอบที่มีการตรวจให้คะแนนแบบหลายค่า จำนวน 18 ข้อ ใช้กับกลุ่มตัวอย่าง 1,000 คน แบ่งเป็นกลุ่มอ้างอิง 500 คน และกลุ่มเปรียบเทียบ 500 คน โดยมีการจับคู่ระหว่างข้อสอบแบบให้คะแนนสองค่า กับแบบให้คะแนนหลายค่า โดยใช้โมเดลโลจิสติก 3 พารามิเตอร์จากข้อสอบจำนวน 75 ข้อ ที่มีค่าอำนาจอำนาจอยู่ระหว่าง .74 ถึง 1.0 ค่าความยากอยู่ระหว่าง -1.95 ถึง 1.95 ค่าการเดาอยู่ที่ .15 และสถานการณ์ที่จำลองขึ้นมากับกลุ่มตัวอย่างทั้ง 2 กลุ่มที่มีการแจกแจงดัชนีการปฏิบัติที่ดี เมื่อความแตกต่างของค่าเฉลี่ยของกลุ่มที่มีความคลาดเคลื่อนมาตรฐานเป็น 1 วิธีวิธีซิปเทสท์ปรับใหม่ วัดผลกระทบของขนาดกลุ่มตัวอย่างค่อนข้าง จาก 5 วิธี ในทางปฏิบัติไม่สามารถมองเห็นความแตกต่างของกลุ่มย่อยทั้ง 2 กลุ่ม จะมีการแจกแจงเบ้ไปด้านใดด้านหนึ่งเมื่อกลุ่มมีการแจกแจงที่แตกต่างกันและข้อสอบที่ศึกษามีค่าอำนาจจำแนกสูง วิธีซิปเทสท์ดีที่สุด รองลงมาวิธี SMD และ วิธีแมนเทล-แฮนส์เซล เมื่อพิจารณาจากความคลาดเคลื่อนประเภทที่ 1 ถ้าแบบทดสอบตอบสั้นๆ เมื่อดูค่าอำนาจจำแนก ทั้ง 5 วิธี แตกต่างกัน เนื่องจากเป็นความแตกต่างในการวิเคราะห์การหน้าที่ต่างกันและองค์ประกอบอื่นๆวิธีแมนเทล-แฮนส์เซล และวิธี SMD เป็นวิธีที่ใช้เกณฑ์การจับคู่กันระหว่างการตอบแบบฝึกหัดกับอัตราความคลาดเคลื่อนประเภทที่ 1 วิธีซิปเทสท์ ไม่สามารถใช้กับการจับคู่ได้ เนื่องจากแบบทดสอบถูกผิดไม่ได้ และในปัจจุบันงานวิจัยไม่นิยมใช้กัน

แฮมเบิลตัน และ โจนส์ (Hambleton & other, 1993) ได้ศึกษาเปรียบเทียบความสอดคล้องของผลการตรวจสอบการทำหน้าที่ต่างกัน ระหว่างวิธีพิจารณาตัดสินข้อสอบ กับวิธี IRT Area และวิธีแมนเทล-แฮนส์เซล ผลการวิจัยพบว่า

1. ผลการตรวจสอบทำหน้าที่ต่างกัน ระหว่างวิธี IRT Area และ วิธีแมนเทล-แฮนส์เซล ไม่ค่อยคงเส้นคงวานัก และมีความสอดคล้องกันในระดับปานกลาง

2. ผลการตรวจพบข้อสอบทำหน้าที่ต่างกัน ระหว่างวิธีพิจารณาตัดสินข้อสอบ กับวิธีการทางสถิติ มีความสอดคล้องกันในระดับปานกลาง (5 ข้อ ใน 11 ข้อ)

3. วิธีการพิจารณาตัดสินข้อสอบ สามารถนำไปใช้จำแนกข้อสอบทำหน้าที่ต่างกันในทางปฏิบัติได้

เมเซอร์ และคณะ (mazer et al. 1991, 443- 451) ได้ศึกษาผลกระทบของขนาดของกลุ่มตัวอย่างที่มีต่อการวิเคราะห์ความลำเอียงด้วยวิธี MH โดยใช้ข้อมูลที่จำลองด้วยโปรแกรม DATAGEN แบบ 3พารามิเตอร์ ขนาดกลุ่มตัวอย่างที่ศึกษามี 5 คือ 2,000 คน 1,000 คน 500 คน 200 คน และ 100 คน ความยาวแบบทดสอบชุดล่ะ 75 ข้อ ผลการวิจัยพบว่า เมื่อใช้ขนาดกลุ่มตัวอย่าง 2,000 คน วิธี MH จะตรวจพบความลำเอียงได้ผิดพลาดร้อยล่ะ 50 และข้อสอบที่ไม่สามารถตรวจค้นความลำเอียงได้เป็นข้อสอบที่ยากมาก ข้อสอบที่มีค่าอำนาจจำแนกต่ำ และข้อสอบที่ค่าความยากต่างกันเล็กน้อย สำหรับ 2 กลุ่ม

โรเจอร์และสวามินาทาน (Rogers & Swaminatan. 1993, 105 – 116) ได้ศึกษาเปรียบเทียบผลวิเคราะห์ความลำเอียงของข้อสอบด้วยวิธีถดถอยโลจิสติก (LR) กับวิธีแมนเทล-แฮนส์เซล(MH) โดยใช้ข้อมูลที่จำลองขึ้นเงื่อนไขการศึกษาเป็น 4 แบบ คือ ขนาดกลุ่มตัวอย่าง 2 กลุ่ม ขนาด 250 คน และ 500 คนต่อกลุ่ม ระดับความเหมาะสมของโมเดลข้อมูลเป็น 2 ระดับ ระดับที่เหมาะสมใช้รูปแบบ 2 พารามิเตอร์ ระดับที่ไม่เหมาะสมใช้รูปแบบ 3 พารามิเตอร์โดยค่าการเดา (c) จะถูกกำหนดที่ 0.2 เครื่องมือที่ใช้เป็นข้อสอบจำนวน 40 ข้อที่จำลองขึ้นมา 7 แบบ ดังนี้ แบบที่ 1 ค่าความยากและค่าอำนาจจำแนกต่ำ แบบที่ 2 ค่าความยากสูงและค่าอำนาจจำแนกต่ำ แบบที่ 3 ค่าความยากและค่าอำนาจจำแนกต่ำ แบบที่ 4 ค่าความยากปานกลางค่าอำนาจจำแนกสูง แบบที่ 5 ค่าความยากสูงค่าอำนาจจำแนกต่ำ แบบที่ 6 ค่าความยากและค่าอำนาจจำแนกสูง แบบที่ 7 เป็นแบบผสมคือค่าอำนาจจำแนกและค่าความยากต่างกันสำหรับ 2 กลุ่ม โดยวิเคราะห์ร้อยละของความลำเอียงที่ตรวจค้นพบในความลำเอียงแบบยูนิฟอร์มและนอนยูนิฟอร์ม ผลการวิจัย พบว่า ร้อยละของความลำเอียงที่ตรวจค้นพบในความลำเอียง แบบยูนิฟอร์มสูงกว่าแบบนอนยูนิฟอร์มทั้ง 2 วิธีและในทุกเงื่อนไข โดยวิธีถดถอยโลจิสติก จะตรวจพบร้อยละของความลำเอียงสูงที่ระดับความยากปานกลาง

ค่าอำนาจจำแนกได้เสนอแนะว่า วิธีของแมนเทล-แฮนส์เซลนี้ใช้ได้ดีสำหรับข้อสอบที่มีค่าอำนาจจำแนกสูงและมีแนวโน้มจะไม่สามารถตรวจสอบความลำเอียงของข้อสอบที่มีค่าความยากสูงได้

จากการศึกษากรอบแนวคิดเชิงทฤษฎีเกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ตลอดจนงานวิจัยต่างๆ ที่เกี่ยวข้องพบว่า นักวิจัยตั้งแต่อดีตจนถึงปัจจุบัน พยายามคิดค้นและพัฒนาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เพื่อให้ผลการตรวจสอบมีความถูกต้องและแม่นยำมากที่สุด เริ่มต้นจากการตรวจสอบในข้อสอบที่วัดความสามารถมิติเดียวและให้คะแนนสองค่า แล้วพัฒนาต่อเนื่องมาจนถึงการตรวจสอบในข้อสอบที่ให้คะแนนหลายค่าต่อมามีการพัฒนาโมเดลการวัดความสามารถหลายมิติ และมีการพัฒนาโปรแกรมคอมพิวเตอร์ที่ใช้วิเคราะห์ค่าพารามิเตอร์ของข้อสอบที่วัดความสามารถหลายมิติ ตลอดจนสามารถประเมินมิติของแบบทดสอบ เช่น โปรแกรม SIBTEST, POLY-SIBTEST, IRTDIF, NOHARM, TESTFACT, DIMTEST และ Poly-DIMTEST เป็นต้น จากวิธีต่าง ๆ ดังกล่าวบางวิธีวิธีเป็นวิธีที่ให้ผลการวิเคราะห์ได้ถูกต้อง แม่นยำ แต่ไม่เหมาะที่จะนำมาใช้ในทางปฏิบัติ เพราะต้องใช้กลุ่มตัวอย่างขนาดใหญ่ การคำนวณและการแปลผลซับซ้อน ยุ่งยาก ต้องเสียค่าใช้จ่ายสูง ในขณะที่บางวิธีเป็นวิธีที่ใช้กลุ่มตัวอย่างไม่มาก การคำนวณและการแปลผลง่าย ไม่ซับซ้อน แต่ให้ผลการวิเคราะห์ที่มีความถูกต้องน้อยกว่าวิธีอื่น ๆ และบางวิธีใช้กลุ่มตัวอย่างไม่มาก การคำนวณและการแปลผลง่าย เหมาะที่จะนำมาใช้ในทางปฏิบัติได้และให้ผลการวิเคราะห์ที่มีความถูกต้อง สำหรับตัวแปรที่ทำหน้าที่ต่างกันที่นิยมมาศึกษาคือ ภาษา เพศ วัฒนธรรม ศาสนา สภาพภูมิศาสตร์เป็นต้น นอกจากนี้จากการศึกษางานวิจัยก็พบว่ามี ตัวแปรบางตัวที่ส่งผลต่อการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธีต่าง ๆ เช่น ภาษาที่ใช้พูดในครอบครัว เชื้อชาติ ศาสนา และตัวแปรจากแบบทดสอบ เช่น ความยาวแบบทดสอบ การแจกแจงความสามารถของผู้สอบในแต่ละกลุ่ม เป็นต้น

สำหรับการวิจัยครั้งนี้ผู้วิจัยเลือกวิธีถดถอยโลจิสติกและวิธีแมนเทล-แฮนส์เซล เนื่องจากทั้งสองวิธีนี้มีข้อดีและข้อเสียที่แตกต่างกัน ทั้งเป็นที่นิยมกันมากในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวัดผลสัมฤทธิ์ทางการเรียน ระดับเขตพื้นที่การศึกษา ชั้นมัธยมศึกษาปีที่ 2 เมื่อจำแนกตามกลุ่มเพศและการใช้ภาษาในชีวิตประจำวัน