บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

แนวคิดและงานวิจัยที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ผู้วิจัยได้นำเสนอเป็น 2 ตอน ดังนี้

ตอนที่ 1 การทำหน้าที่ต่างกันของข้อสอบ

1.1 ประเภทของการทำหน้าที่ต่างกันของข้อสอบ

1.2 หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

ตอนที่ 2 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบและงานวิจัยที่เกี่ยวข้อง

2.1 วิธีซิปเทสท์

2.2 วิธีการวิเคราะห์องค์ประกอบจำกัด

2.3 วิธีถดถอยโลจิสติก

ตอนที่ 1 การทำหน้าที่ต่างกันของข้อสอบ

การศึกษาเรื่องผลการสอบของกลุ่มผู้เข้าสอบย่อยของผู้เข้าสอบทั้งหมดมีมานานแล้ว เพิ่งมีการศึกษาเรื่องความยุติธรรมในการสอบระหว่างผู้เข้าสอบย่อยต่างกลุ่มอย่างจริงจังในช่วงปลายทศวรรษ 1960 โดยมีการเสนอวิธีการต่าง ๆ เพื่อนำไปใช้ตรวจสอบความลำเอียงของแบบทดสอบ (Test Bias) และความลำเอียงในการคัดเลือกผู้เข้าสอบ (Selection Bias) ขึ้นหลายวิธี ในช่วงเวลานั้นนักพัฒนาแบบทดสอบมีความสนใจวิธีการจำแนกข้อสอบที่ไม่เหมาะสมกับผู้เข้าสอบบางกลุ่มออกจากแบบทดสอบ ก่อนที่จะพัฒนาเป็นแบบทดสอบฉบับสมบูรณ์ ทำให้มีการพัฒนาวิธีการตรวจสอบความลำเอียงของข้อสอบ (Item Bias) เพื่อใช้ในการจำแนกข้อสอบที่ลำเอียงกับผู้เข้าสอบบางกลุ่มที่มีลักษณะบางอย่างแตกต่างกัน เช่น เชื้อชาติ ศาสนา วัฒนธรรม ภูมิลำเนา สังคม เพศ ภาษา อายุ ประสบการณ์ เป็นต้น เพื่อพัฒนาแบบทดสอบให้มีคุณภาพเหมาะสมสำหรับนำไปใช้ทดสอบต่อไป

การศึกษาเรื่องผลการสอบในสมัยแรก ๆ มีจุดประสงค์ เพื่อคัดเลือกคนเข้าศึกษาต่อหรือเข้าทำงาน ปรากฏว่ามีหลักฐานแสดงว่ามีความลำเอียงกับกลุ่มคนต่างชาติ ต่างเพศ ทำให้ต้องมีการศึกษา"ความลำเอียงในการคัดเลือกผู้เข้าสอบ" เพื่อให้การศึกษาเรื่องนี้มีความชัดเจนยิ่งขึ้น ในเวลาต่อมาจึงได้ศึกษาในระดับข้อสอบ (Item Level) ที่เรียกว่า "ความลำเอียงของข้อสอบ (Item Bias)" ซึ่งในปัจจุบันนักวิจัยส่วนใหญ่ใช้คำว่า "ข้อสอบทำหน้าที่ต่างกันกับกลุ่มผู้เข้าสอบย่อย

ต่างกลุ่ม" หรือเรียกสั้น ๆ ว่า "ข้อสอบทำหน้าที่ต่างกัน (Differential Item Functioning: DIF)" เนื่องจากเห็นว่าเป็นคำที่มีความหมายกลาง ๆ และมีความเหมาะสมในเชิงวิชาการมากกว่าคำว่า "ความลำเอียง (Bias)" ซึ่งเป็นคำที่ใช้กันในทางสังคมและมีความหมายในเชิงลบ อย่างไรก็ตาม คำสองคำนี้มีจุดเน้นที่แตกต่างกัน โดยคำว่า ความลำเอียงของข้อสอบ เน้นที่อิทธิพลที่สังเกตได้ของกลุ่มผู้เข้าสอบย่อยที่มุ่งศึกษา ส่วนคำว่า ข้อสอบที่ทำหน้าที่ต่างกัน เน้นที่ลักษณะทางสถิติของข้อสอบที่ตรวจสอบได้ด้วยวิธีวิเคราะห์ทางสถิติ ซึ่งเป็นส่วนประกอบหนึ่งของสิ่งที่แสดงถึงความลำเอียงของข้อสอบ (Scheuneman & Bleistein, 1989; Angoff, 1993; Hambleton & Others, 1993; Zieky, 1993; Camilli & Shepard, 1994) จากจุดเน้นนี้แสดงให้เห็นว่า วิธีการทางสถิติที่ใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นเงื่อนไขจำเป็น (Necessary Condition) ในการประเมินความลำเอียงของข้อสอบ แต่ถ้าใช้เฉพาะวิธีการทางสถิติอย่างเดียว ผลการตรวจพบข้อสอบทำหน้าที่ต่างกันที่ได้ไม่อาจสรุปได้ว่าข้อสอบข้อนั้นลำเอียงหรือไม่ เนื่องจากการประเมินความลำเอียงของข้อสอบยังต้องรวมไปถึงการใช้วิธีการตัดสินข้อสอบ (Judgmental Method) โดยให้ผู้เชี่ยวชาญพิจารณาเนื้อหาสาระของข้อสอบและจุดมุ่งหมายในการวัดของแบบทดสอบ ก่อนที่จะสรุปว่าข้อสอบข้อนั้นลำเอียงหรือไม่ (Angoff, 1993; Linn, 1993; Ramsay, 1993; Zieky, 1993; Camilli & Shepard, 1994 อ้างถึงใน เสรี ชัดแช้ม, 2539, หน้า 1-2)

ปัจจุบันปรากฏว่า นักวัดผลการศึกษาส่วนใหญ่ใช้คำว่าการทำหน้าที่ต่างกันของข้อสอบแทนคำว่า ความลำเอียงของข้อสอบ ซึ่งมีผู้ให้ความหมายการทำหน้าที่ต่างกันของข้อสอบ ไว้ดังนี้

เคเดอร์แมน และมาครีดี (Kederman & Macready, 1990) กล่าวว่า การทำหน้าที่ต่างกันของข้อสอบ หมายถึง คะแนนที่ได้จากกลุ่มผู้เข้าสอบที่มีความสามารถเท่ากัน แต่มาจากต่างกลุ่มกันมีความแตกต่างกันอย่างเป็นระบบ

ฮอลแลนด์ และไวเนอร์ (Holland & Wainer, 1993) กล่าวว่า การทำหน้าที่ต่างกันของข้อสอบ หมายถึง สารสนเทศทางสถิติของข้อสอบที่ได้จากผลการตอบของผู้เข้าสอบต่างกลุ่มกันและมีความสามารถเท่ากัน แต่มีโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกัน

คามิลลิ และเชฟเพอร์ด (Camilli & Shepard, 1994) กล่าวว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เป็นการตรวจสอบความเป็นพหุมิติของข้อสอบ โดยจะแสดงการทำหน้าที่ต่างกันระหว่างกลุ่มผู้เข้าสอบตั้งแต่สองกลุ่มขึ้นไปที่มีความสามารถหลัก (Primary Abilities) เท่ากัน แต่มีความสามารถรอง (Secondary Abilities) แตกต่างกัน

โปเทนซา และโดรันส์ (Potenza & Dorans, 1995) กล่าวว่า การทำหน้าที่ต่างกันของข้อสอบ หมายถึง ผลการตอบข้อสอบระหว่างกลุ่มผู้เข้าสอบสองกลุ่มที่นำมาเปรียบเทียบมีความแตกต่างกัน

การเปรียบเทียบกลุ่มผู้เข้าสอบเป็นสิ่งสำคัญ ที่จะอธิบายถึงความแตกต่างระหว่างการทำหน้าที่ของข้อสอบกับคุณลักษณะแฝงของกลุ่มผู้เข้าสอบ

นารายานัน และสวามินาทาน (Narayanan & Swaminathan, 1996) กล่าวว่า การทำหน้าที่ต่างกันของข้อสอบ หมายถึง ผู้เข้าสอบมีความสามารถระดับเดียวกัน แต่มาจากกลุ่มย่อยต่างกัน มีโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกัน

มีผู้ให้ความหมายของคำว่า "การทำหน้าที่ต่างกันของข้อสอบ" (Differential Item Functioning: DIF) ไว้หลายความหมายดังกล่าวแล้วข้างต้น แต่ความหมายที่เป็นที่ยอมรับกันอย่างกว้างขวางก็คือ ข้อสอบทำหน้าที่ต่างกันภายใต้เงื่อนไขผู้เข้าสอบมีความสามารถเท่ากัน แต่มาจากกลุ่มผู้เข้าสอบย่อยที่มีลักษณะต่างกัน มีความน่าจะเป็นในการตอบข้อสอบข้อนั้นไม่เท่ากัน (เสรี ชัดแช้ม, 2540, หน้า 42)

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นการเปรียบเทียบผลการตอบระหว่างผู้เข้าสอบ 2 กลุ่ม คือ กลุ่มอ้างอิง (Reference Group หรือกลุ่ม R) และกลุ่มเปรียบเทียบ (Focal Group หรือกลุ่ม F) กลุ่มอ้างอิงเป็นกลุ่มที่คาดว่าจะได้ประโยชน์จากการตอบข้อสอบ กล่าวคือ มีโอกาสในการตอบข้อสอบถูกได้มากกว่าผู้เข้าสอบกลุ่มเปรียบเทียบ ส่วนกลุ่มเปรียบเทียบเป็นกลุ่มที่คาดว่าจะเสียประโยชน์จากการตอบข้อสอบ กล่าวคือ มีโอกาสตอบข้อสอบถูกได้น้อยกว่าผู้เข้าสอบกลุ่มอ้างอิง เนื่องจากคุณลักษณะเฉพาะของบุคคลกับเนื้อหาของข้อสอบนั้น เช่น การศึกษาการทำหน้าที่ต่างกันของข้อสอบระหว่างผู้เข้าสอบต่างสถานที่ กลุ่มเปรียบเทียบ ได้แก่ กลุ่มผู้เข้าสอบในชนบท กลุ่มอ้างอิง ได้แก่ กลุ่มผู้เข้าสอบในเมือง เป็นต้น ในการเปรียบเทียบจะศึกษาปัจจัยอันเกิดจากผู้เข้าสอบซึ่งส่งผลให้เกิดการได้ประโยชน์และเสียประโยชน์ระหว่างกลุ่มผู้เข้าสอบ เช่น เพศ สีผิว เชื้อชาติ ภาษา สถาบันการศึกษา ประสบการณ์ เป็นต้น ต่อมาระยะหลังได้มีการศึกษาเปรียบเทียบวิธีการต่าง ๆ ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

วิธีการตรวจสอบการทำหน้าที่ต่างกันมีหลายวิธี ทั้งนี้เพราะมีการศึกษาและคิดค้นวิธีการต่างๆ เพื่อให้สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้อย่างมีประสิทธิภาพมากที่สุด แฮมเบิลตันและคณะ (Hambleton & Other, 1993 อ้างถึงใน เสรี ชัดแช้ม, 2539, หน้า 4-6) จำแนกวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบออกเป็น 3 กลุ่มใหญ่ ๆ ดังนี้

1. กลุ่มวิธีที่ใช้ทฤษฎีการทดสอบแบบดั้งเดิม (Methods Using Classical Test Theory: CTT) วิธีในกลุ่มนี้พัฒนามาจากหลักการของทฤษฎีการทดสอบแบบดั้งเดิม โดยปกติแล้วใช้คะแนนที่สังเกตได้ของผู้เข้าสอบแต่ละคนเป็นเกณฑ์การจับคู่กลุ่มผู้เข้าสอบย่อย และเปรียบเทียบค่าความยากของข้อสอบแต่ละข้อระหว่างกลุ่มผู้เข้าสอบย่อยที่สนใจศึกษา วิธีการในกลุ่มนี้ ได้แก่ การวิเคราะห์ความแปรปรวน (Analysis of Variance) วิธีสหสัมพันธ์ (Correlational Methods)

(Green & Draper, 1972 cited in Scheuneman & Bleistein, 1989) วิธีแปลงค่าความยากของข้อสอบ (Transformed Item Difficulty Method, TID) หรือวิธีการกำหนดจุดเดลต้า (Delta Plot Method) (Angoff, 1982) การวิเคราะห์ตัวลวง (Distractor Analysis) (Scheunemam, 1982) วิธีสหสัมพันธ์บางส่วน (Partial Correlation Methods) (Stricker, 1982) และวิธีการทำให้เป็นมาตรฐาน (Standardization Method) (Dorans & Kulick, 1983)

ข้อได้เปรียบของวิธีในกลุ่มนี้ คือ กระบวนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบไม่ยุ่งยาก เสียค่าใช้จ่ายไม่สูงนัก ใช้ตรวจสอบกับกลุ่มตัวอย่างขนาดเล็กได้ และสามารถอธิบายให้คนทั่วไปเข้าใจได้ง่าย ส่วนข้อเสียเปรียบก็คือ ค่าสถิติของข้อสอบเปลี่ยนไปตามกลุ่มตัวอย่าง เมื่อกลุ่มตัวอย่างเปลี่ยนไปผลการตรวจพบข้อสอบทำหน้าที่ต่างกันก็เปลี่ยนไป ทำให้การสรุปอ้างอิงผลการศึกษาไปยังกลุ่มประชากร อาจเชื่อถือได้น้อยลง

2. กลุ่มวิธีที่ใช้ทฤษฎีการตอบสนองข้อสอบ (Methods Using Item Response Theory: IRT) วิธีการในกลุ่มนี้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ตามกรอบแนวคิดของทฤษฎีการตอบสนองข้อสอบ โดยปกติแล้วใช้การเปรียบเทียบเส้นโค้งลักษณะข้อสอบ (Item Characteristic Curves: ICCs) ของกลุ่มผู้เข้าสอบย่อยตามระดับความสามารถของผู้เข้าสอบ ถ้าเส้นโค้งลักษณะข้อสอบของกลุ่มผู้เข้าสอบย่อยสองกลุ่ม มีรูปร่างเหมือนกัน แสดงว่าข้อสอบข้อนั้นทำหน้าที่ไม่ต่างกัน แต่ถ้าเส้นโค้งลักษณะข้อสอบของกลุ่มผู้เข้าสอบย่อยสองกลุ่มมีรูปร่างแตกต่างกัน แสดงว่าข้อสอบข้อนั้นทำหน้าที่ต่างกัน วิธีการในกลุ่มนี้ได้แก่ วิธี Analysis of fit (Durovic, 1975 cited in Hambleton & Others, 1993) วิธี Difficulty shift (Wright, Mead & Draba, 1976 cited in Hambleton & Others, 1993) ซึ่งใช้โมเดล IRT แบบ หนึ่งพารามิเตอร์ วิธี IRT area (Ironson & subkoviak, 1979: Raju, 1988, 1990) วิธี Two-Stage (Lord, 1980) ซึ่งใช้โมเดล IRT แบบ สอง หรือสามพารามิเตอร์ วิธี Plot (Hambleton & Rogers, 1991 cited in Hambleton & Others, 1993) และวิธีซิบเทสท์ (SIBTEST) (Shealy & Stout, 1993)

ข้อได้เปรียบของวิธีการในกลุ่มนี้คือ การแก้ไขข้อบกพร่องของทฤษฎีการทดสอบแบบดั้งเดิม ทำให้ค่าสถิติของข้อสอบไม่เปลี่ยนไปตามกลุ่มตัวอย่างที่สุ่มมาจากประชากรเดียวกัน การประมาณค่าความสามารถของผู้เข้าสอบเป็นอิสระจากค่าความยากของแบบทดสอบ โมเดลทางคณิตศาสตร์ง่ายต่อการจับคู่เส้นโค้งลักษณะข้อสอบตามระดับความสามารถของผู้เข้าสอบ ทำให้สามารถศึกษาความแตกต่างของผลการตอบข้อสอบตามระดับความสามารถของกลุ่มผู้เข้าสอบย่อยได้ ไม่ต้องมีข้อตกลงเบื้องต้นเรื่องแบบทดสอบคู่ขนานในการหาค่าสัมประสิทธิ์ความเที่ยงของแบบทดสอบ และถ้าผลการตอบข้อสอบของกลุ่มผู้เข้าสอบสอดคล้องกับข้อตกลงเบื้องต้นของโมเดล IRT แล้ว วิธีในกลุ่มนี้ก็น่าจะเป็นวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ได้ผลดี

เนื่องจากเป็นวิธีที่มีทฤษฎีการตอบสนองข้อสอบสนันสนุนและใช้ค่าประมาณค่าความสามารถที่แท้จริงของผู้เข้าสอบแทนคะแนนที่สังเกตได้ ส่วนข้อเสียเปรียบของวิธีการในกลุ่มนี้ก็คือ กระบวนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสลับซับซ้อน เสียค่าใช้จ่ายในการดำเนินการสูง และต้องการกลุ่มตัวอย่างขนาดใหญ่

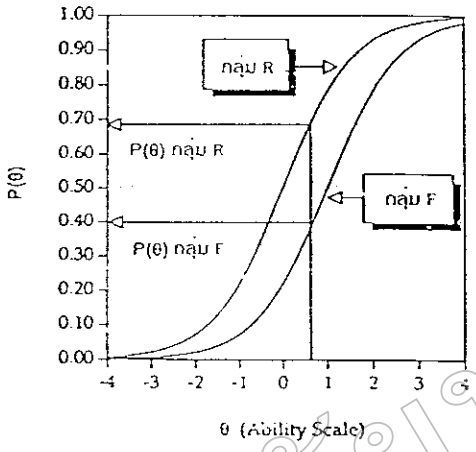
3. กลุ่มวิธีที่ใช้วิธีไค-สแควร์ (Methods Using Chi-Square Methods) วิธีในกลุ่มนี้บางครั้งก็เรียกว่า กลุ่มวิธีไค-สแควร์ เนื่องจากใช้ค่าสถิติไค-สแควร์ แสดงการทำหน้าที่ต่างกันของข้อสอบ และใช้คะแนนของแบบทดสอบหรือคะแนนของแบบทดสอบที่ทำให้บริสุทธิ์ เป็นเกณฑ์การจับคู่กลุ่มผู้เข้าสอบย่อยสองกลุ่มที่ศึกษา ก่อนการเปรียบเทียบผลการตอบข้อสอบ วิธีการในกลุ่มนี้ ได้แก่ วิธีตารางการณ์จร (Contingengy Table Method) (Scheuneman, 1975; 1979) วิธีตารางการณ์จรปรับใหม่ (Modified Contingency Table Mothod) (Veale, 1977 cited in Hambleton & Others, 1993) วิธีลอก-ลิเนียร์ (Log-Linear Mothods) (Mellenbergh, 1982) วิธีแมนเทล-แฮนส์เซล (Mantel-Haenszal Method: MH) (Holland & Thayer, 1986; 1988) และวิธีถดถอยโลจิสติก (Logistic Regression Methods: LR) (Swaminathan & Rogers, 1990) และวิธีการวิเคราะห์องค์ประกอบจำกัด (Restricted Factor Analysis Methods: RFA) (Oort, 1998)

ข้อได้เปรียบของวิธีในกลุ่มนี้คือ กระบวนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบไม่ยุ่งยาก เสียค่าใช้จ่ายในการวิเคราะห์ข้อมูลไม่สูง ใช้ขนาดกลุ่มตัวอย่างไม่ใหญ่นัก และบางวิธีมีหลักการที่ดีในการจับคู่กลุ่มผู้เข้าสอบย่อยตามความสามารถของผู้เข้าสอบ และมีการทดสอบนัยสำคัญ ส่วนข้อเสียเปรียบของวิธีในกลุ่มนี้ก็คล้าย ๆ กับวิธีที่ใช้ทฤษฎีการทดสอบแบบดั้งเดิม

1. ประเภทของการทำหน้าที่ต่างกันของข้อสอบ

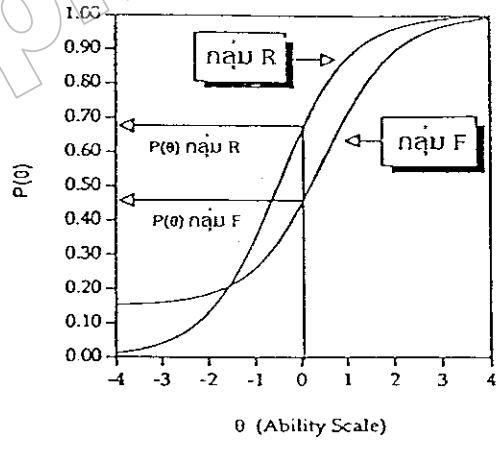
ข้อสอบสามารถทำหน้าที่แตกต่างกันได้ 2 ประเภท (Mellenbergh, 1982) ดังนี้

1.1 ข้อสอบทำหน้าที่ต่างกันแบบเอกรูป (Uniform DIF) หมายถึง ข้อสอบที่ทำหน้าที่ให้ผู้เข้าสอบกลุ่มหนึ่งมีโอกาสในการตอบข้อสอบถูกมากกว่าผู้เข้าสอบอีกกลุ่มหนึ่งสม่ำเสมอกันในทุกระดับความสามารถ เมื่อพิจารณาเส้นโค้งคุณลักษณะข้อสอบของผู้เข้าสอบ 2 กลุ่ม จะพบว่าไม่มีปฏิสัมพันธ์ระหว่างเส้นโค้งคุณลักษณะข้อสอบในทุกระดับความสามารถ



ภาพที่ 1 ข้อสอบทำหน้าที่ต่างกันแบบเอกรูป (Uniform DIF)

1.2 ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป (Nonuniform DIF) หมายถึง ข้อสอบที่ทำให้โอกาสในการตอบข้อสอบถูกของผู้เข้าสอบระหว่างกลุ่มไม่สม่ำเสมอกันในทุกระดับความสามารถ เมื่อพิจารณาเส้นโค้งคุณลักษณะข้อสอบของผู้เข้าสอบ 2 กลุ่ม พบว่ามีปฏิสัมพันธ์ร่วมกันระหว่างเส้นโค้งคุณลักษณะ เช่น ที่ระดับความสามารถสูง กลุ่มอ้างอิงมีโอกาสในการตอบข้อสอบถูกมากกว่ากลุ่มเปรียบเทียบ แต่ที่ระดับความสามารถต่ำ กลุ่มเปรียบเทียบมีโอกาสในการตอบข้อสอบถูกมากกว่ากลุ่มอ้างอิง เป็นต้น



ภาพที่ 2 ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป (Nonuniform DIF)

ตามทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) สามารถพิจารณา "ปฏิสัมพันธ์" ดังกล่าวได้จากความแตกต่างของค่าพารามิเตอร์อำนาจจำแนกของข้อสอบ ระหว่างผู้เข้าสอบกลุ่มย่อยสองกลุ่ม กล่าวคือ ถ้าข้อสอบทำหน้าที่ต่างกันแบบเอกรูป แล้วเส้นโค้งลักษณะข้อสอบ (Item Characteristic Curves: ICCs) ระหว่างผู้เข้าสอบกลุ่มย่อยสองกลุ่มจะขนานกัน หรือมีฟังก์ชันการตอบสนองข้อสอบ (Item Response Functions: IRFs) เหมือนกัน แต่ถ้าข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปแล้ว เส้นโค้งลักษณะข้อสอบระหว่างผู้เข้าสอบกลุ่มย่อยสองกลุ่มจะไม่ขนานกัน หรือมีฟังก์ชันการตอบสนองข้อสอบต่างกัน

2. หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เป็นการเปรียบเทียบผลการตอบข้อสอบเป็นรายข้อระหว่างกลุ่มผู้เข้าสอบอย่างน้อย 2 กลุ่ม ที่มีความสามารถหลัก (Primary Ability) ที่มุ่งวัดเท่ากัน แต่คาดว่าจะมีความได้เปรียบหรือเสียเปรียบกัน โดยกลุ่มหนึ่งถือเป็น กลุ่มอ้างอิง ซึ่งคาดว่าน่าจะได้เปรียบในการตอบข้อสอบข้อนั้น หรือมีโอกาสตอบข้อสอบได้ถูกต้องมากกว่า เพราะมีคุณลักษณะเฉพาะต่างกัน ส่วนอีกกลุ่มคือ กลุ่มเปรียบเทียบ ซึ่งเป็นกลุ่มที่สนใจศึกษา และคาดว่าน่าจะเป็นกลุ่มที่เสียเปรียบในการตอบข้อสอบข้อนั้นเพราะมีคุณลักษณะเฉพาะต่างกัน หรือมีโอกาสตอบข้อสอบได้ถูกต้องน้อยกว่า

ในการเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบจำเป็นต้องมีการจับคู่ (Matching) ผู้เข้าสอบ ซึ่งเป็นเงื่อนไขสำคัญของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

เกณฑ์การจับคู่ (Matching Criteria) ที่นิยมใช้กันมี 2 วิธี ดังนี้

1. เกณฑ์ภายนอก (External Criteria) การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยใช้เกณฑ์ภายนอกนี้ สามารถนำไปใช้ได้ทั้งข้อสอบรายข้อและแบบทดสอบทั้งฉบับ โดยการใช้คะแนนจากแบบทดสอบอื่นเป็นเกณฑ์ภายนอก แล้วใช้เทคนิคการวิเคราะห์การถดถอย (Regression Analysis) เพื่อเปรียบเทียบเส้นกราฟความสัมพันธ์ระหว่างตัวแปรเกณฑ์ กับตัวแปรทำนายระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ

หลักการนี้มีจุดมุ่งหมาย เพื่อสร้างสมการทำนายตัวแปรเกณฑ์ ซึ่งเป็นคะแนนของแบบทดสอบอื่นจากตัวแปรทำนายที่เป็นคะแนนรายข้อหรือคะแนนแบบทดสอบ ระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ จะใช้คะแนนรายข้อเป็นตัวแปรทำนาย แต่ถ้าเป็นการวิเคราะห์การทำหน้าที่ต่างกันของแบบทดสอบ จะใช้คะแนนรวมของแบบทดสอบทั้งฉบับเป็นตัวแปรทำนาย สำหรับตัวแปรเกณฑ์ที่ใช้เป็นเกณฑ์ภายนอก อาจใช้

คะแนนรวมทั้งฉบับหรือเกรดเฉลี่ย หรือคะแนนจากงานที่เกี่ยวข้องของผู้เข้าสอบ (Cronbach, 1970) สมการทำนายสำหรับกลุ่มอ้างอิงและกลุ่มเปรียบเทียบแสดงได้ดังนี้

กลุ่มอ้างอิง (R) $Y_i = A_R + B_R X_i$ (1)

กลุ่มเปรียบเทียบ (F) $Y_i = A_F + B_F X_i$ (2)

เมื่อ Y_i = คะแนนของตัวแปรเกณฑ์ภายนอก

X_i = คะแนนของตัวแปรทำนาย

A = ค่าคงที่หรือจุดตัดแกน y (Intercept)

B = ค่าความชัน (Slope)

จากฟังก์ชันการทำนายทั้ง 2 สมการ สามารถเปรียบเทียบค่าตัดแกน (A) และค่าความชัน (B) ของเส้นกราฟระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบได้ ถ้าเส้นกราฟดังกล่าวมีค่าความชัน หรือค่าตัดแกนแตกต่าง สำหรับข้อสอบใดหรือแบบทดสอบใด แสดงว่าข้อสอบหรือแบบทดสอบนั้นมีการทำหน้าที่ต่างกัน โดยเข้าข้างกลุ่มผู้เข้าสอบที่มีค่าตัดแกนหรือค่าความชันที่สูงกว่า

การใช้เกณฑ์ภายนอกมีข้อดี คือ เกณฑ์ที่ใช้มีความเป็นอิสระจากข้อสอบ และแบบทดสอบที่ต้องการตรวจสอบ แต่มีจุดอ่อนตรงที่ความเหมาะสมของเกณฑ์ที่จะนำมาใช้ ในทางปฏิบัติเป็นการยากที่จะหาเกณฑ์ภายนอกจากแบบทดสอบฉบับอื่นที่มีความตรงเชิงทำนาย และมีความยุติธรรมสำหรับกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ถ้าเกณฑ์ภายนอกขาดคุณสมบัติดังกล่าว จะทำให้ผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบหรือแบบทดสอบขาดความแม่นยำ

2. เกณฑ์ภายใน (Internal Criteria) การวิเคราะห์การทำหน้าที่ต่างกัน โดยใช้เกณฑ์ภายในเป็นการนำวิธีการทางสถิติมาตรวจสอบการทำหน้าที่ต่างกันของข้อสอบหรือแบบทดสอบ เป็นการเน้นการพิจารณาจากโครงสร้างภายในของแบบทดสอบเป็นหลัก ด้วยการวิเคราะห์ผลจากการตอบข้อสอบ และความสามารถหรือคะแนนจริงของผู้เข้าสอบที่ได้จากแบบทดสอบฉบับนั้น เพื่อนำมาเปรียบเทียบระหว่างผู้เข้าสอบจากกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ ที่มีความสามารถหรือคะแนนจริงเท่ากัน ว่าจะมีผลการตอบหรือโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกันหรือไม่ เพื่อบ่งชี้ถึงการทำหน้าที่ต่างกันของข้อสอบ การวิเคราะห์ในลักษณะนี้นิยมใช้ค่าสถิติต่าง ๆ เป็นตัวบ่งชี้การทำหน้าที่ต่างกันของข้อสอบ ค่าสถิติทดสอบที่นิยมนำมาใช้มีดังนี้

2.1 การทดสอบปฏิสัมพันธ์ ในระยะแรกของการศึกษาความลำเอียงของข้อสอบมีการใช้สถิติทดสอบเอฟ (F-test) ในการวิเคราะห์ความแปรปรวน (ANOVA) เพื่อทดสอบปฏิสัมพันธ์ระหว่างกลุ่มผู้เข้าสอบกับข้อสอบ ถ้าการทดสอบมีนัยสำคัญเป็นสิ่งที่แสดงว่ามีการทำหน้าที่ต่างกัน จากนั้นจึงวิเคราะห์ต่อด้วยวิธีการเปรียบเทียบภายหลัง เพื่อระบุข้อสอบที่มีผลต่อการเกิดปฏิสัมพันธ์ ซึ่งเป็นข้อของข้อสอบที่ทำหน้าที่ต่างกัน

วิธีการนี้มีข้อดีคือ สามารถศึกษาผู้เข้าสอบหลาย ๆ กลุ่มได้ แต่จุดอ่อนในเรื่องการทำให้กลุ่มผู้เข้าสอบต่าง ๆ มีความสามารถที่ทัดเทียมกัน ขนาดกลุ่มตัวอย่างของกลุ่มผู้เข้าสอบแต่ละกลุ่ม และอัตราความคลาดเคลื่อนประเภทที่ 1 จะสูงขึ้น ถ้าจำนวนข้อสอบเพิ่มมากขึ้น

2.2 การวัดความเบี่ยงเบนสัมพัทธ์ การคำนวณค่าความยากของข้อสอบ เมื่อคำนวณแยกระหว่างกลุ่มผู้เข้าสอบ และแปลงให้เป็นค่าความยากมาตรฐาน (Δ) สามารถนำมาเขียนกราฟเปรียบเทียบเป็นรายข้อ ถ้าข้อใดเบี่ยงเบนไปจากแกนหลักที่คาดหมาย หรือเบี่ยงเบนเกินกว่าความคลาดเคลื่อนมาตรฐานของค่าความยากที่กำหนด แสดงถึงการทำหน้าที่ต่างกันของข้อสอบ รวมทั้งสามารถคำนวณค่าสหสัมพันธ์ระหว่างค่าความยากรายข้อระหว่างกลุ่ม เพื่อแสดงถึงการทำหน้าที่ต่างกันของแบบทดสอบ ถ้าสหสัมพันธ์เข้าใกล้ 1.00 แสดงว่าค่าความยากสัมพัทธ์ของข้อสอบมีค่าใกล้เคียงกันระหว่างกลุ่ม ดังนั้นแบบทดสอบวัดคุณลักษณะคล้ายกันระหว่างกลุ่ม

วิธีการนี้มีข้อดีและข้อเสียคล้ายกับการทดสอบปฏิสัมพันธ์ นอกจากนี้ค่าความยากของข้อสอบ (p) มิใช่ตัวแทนของค่าความยากที่แท้จริงของข้อสอบ และอาจได้รับอิทธิพลจากตัวแปรอื่น ได้แก่ ค่าอำนาจจำแนก และความสามารถของผู้เข้าสอบ

2.3 การเปรียบเทียบน้ำหนักองค์ประกอบ การวิเคราะห์องค์ประกอบ (Factor Analysis) เป็นเทคนิคทางสถิติที่นิยมใช้ในการตรวจสอบความตรงเชิงโครงสร้าง เมื่อนำการวิเคราะห์องค์ประกอบมาใช้ในการวิเคราะห์โครงสร้างของแบบทดสอบแยกตามกลุ่มผู้เข้าสอบ ความไม่สอดคล้องกันระหว่างน้ำหนักองค์ประกอบบนคุณลักษณะสำคัญที่มุ่งวัดหรือความแตกต่างของค่าเฉลี่ยคะแนนองค์ประกอบ (Factor Scores) ระหว่างกลุ่มผู้เข้าสอบ ย่อมสะท้อนการทำหน้าที่ต่างกันของข้อสอบและแบบทดสอบ

การใช้เทคนิคการวิเคราะห์องค์ประกอบเชิงสำรวจ (Exporatory Factor Analysis: EFA) สำหรับศึกษาการทำหน้าที่ต่างกันของข้อสอบ มีจุดอ่อนในเรื่องความไม่สอดคล้องระหว่างน้ำหนัก องค์ประกอบอาจเกิดจากความแตกต่างของความสามารถระหว่างกลุ่มก็ได้ แนวทางที่เหมาะสมจึงควรใช้เทคนิคการวิเคราะห์องค์ประกอบเชิงยืนยัน (Confirmatory Factor Analysis: CFA)

2.4 การเปรียบเทียบโอกาสตอบข้อสอบถูก การวิเคราะห์โอกาสตอบข้อสอบถูกของผู้เข้าสอบจากกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่มีความสามารถเท่ากัน เป็นแนวทางสำคัญที่นิยมใช้กันและเป็นที่ยอมรับในปัจจุบัน สำหรับบ่งชี้การทำหน้าที่ต่างกันของข้อสอบ มีการคำนวณค่าสถิติ 2 แนวทางดังนี้

2.4.1 เปรียบเทียบค่าสัดส่วนหรือความน่าจะเป็นในการตอบข้อสอบถูกของผู้เข้าสอบต่างกลุ่มที่มีความสามารถเท่ากัน

2.4.2 เปรียบเทียบค่าฟังก์ชันการตอบสนองข้อสอบหรือเส้นโค้งลักษณะข้อสอบระหว่างกลุ่มที่มีความสามารถเท่ากัน เป็นวิธีที่อยู่บนพื้นฐานของทฤษฎี IRT

วิธีการนี้มีข้อดีที่สำคัญ ได้แก่ การคำนวณค่าสถิติของข้อสอบมีความน่าเชื่อถือ มีกลไกควบคุมความสามารถของผู้เข้าสอบโดยการจับคู่กลุ่มความสามารถ เพื่อเปรียบเทียบ ณ ตำแหน่งต่าง ๆ ที่มีความสามารถเท่ากัน จึงเป็นวิธีการที่ยอมรับกันทั่วไป การวิเคราะห์ต้องใช้โปรแกรมคอมพิวเตอร์เฉพาะ

ตอนที่ 2 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบและงานวิจัยที่เกี่ยวข้อง

1. วิธีซิปเทสท์ (SIBTEST)

เชียลีและสเตาท์ (Shealy & Stout, 1993) ได้เสนอวิธีซิปเทสท์ (Simulataneous Item Bias Test: SIBTEST) เพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) การทำหน้าที่ต่างกันของแบบทดสอบ (Differential Test Functioning: DTF) และการทำหน้าที่ต่างกันของกลุ่มข้อสอบ (Differential Bundle Functioning: DBF) โดยสามารถใช้ได้กับแบบทดสอบเอกมิติ (Unidimensional Test) และแบบทดสอบพหุมิติ (Multidimensional Test) (Stout, Li & Nandakumar, 1997) วิธีซิปเทสท์เป็นสถิติแบบนันพาราเมตริก (Nonparametric) พัฒนามาจากพื้นฐานของทฤษฎี IRT ชนิดพหุมิติ แต่ไม่ต้องใช้ฟังก์ชันการตอบสนองข้อสอบหรือการประมาณค่าความสามารถแฝง วิธีซิปเทสท์ได้รับการออกแบบมาสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป (Uniform DIF) ดังนั้น จึงไม่มีความไวในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป (Nonuniform DIF) (Li & Stout, 1996) จุดเด่นของวิธีซิปเทสท์ คือ คำนวณได้ง่าย ไม่ซับซ้อน ประหยัดค่าใช้จ่าย และไม่จำเป็นต้องใช้กับกลุ่มตัวอย่างที่มีขนาดใหญ่ อีกทั้งใช้สถิติทดสอบนัยสำคัญ (Narayanan & Swaminathan, 1996) นอกจากนี้ ยังสามารถนำไปประยุกต์ใช้กับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนแบบพหุวิภาค (Polytomous DIF) (Chang, Mazzeo & Roussos, 1995)

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีซิปเทสท์ในแบบทดสอบเอกมิติต้องมีข้อตกลงเกี่ยวกับแบบทดสอบ กล่าวคือ ข้อสอบในแบบทดสอบจะต้องมุ่งวัดคุณลักษณะหรือความสามารถแฝงเพียงลักษณะเดียว ความสามารถแฝงประเภทหนึ่งเรียกว่า *ความสามารถเป้าหมายที่ต้องการวัด* (Target Ability: θ) แต่จะมีความสามารถแฝงอีกประเภทหนึ่งที่มีอิทธิพลต่อผลการตอบข้อสอบซึ่งเรียกว่า *ความสามารถแทรกซ้อนที่ไม่ต้องการวัด* (Nuisance Ability: η) ตัวอย่างเช่น แบบทดสอบคำศัพท์ในวิชาภาษาต่างประเทศ ข้อสอบบางข้ออาจถามความรู้สำหรับผู้ชายโดยเฉพาะ เช่น ความรู้เกี่ยวกับอิเล็กทรอนิกส์ เป็นต้น ในขณะที่ข้อสอบบางข้ออาจถามความรู้

สำหรับผู้หญิงโดยเฉพาะ เช่น ความรู้เกี่ยวกับการเย็บปักถักร้อย เป็นต้น จากสถานการณ์ดังกล่าว ทักษะความรู้เกี่ยวกับคำศัพท์ในวิชาภาษาต่างประเทศ เป็นความสามารถเป้าหมายที่ต้องการวัด ซึ่งแทนด้วย θ ส่วนความรู้ทางด้านอิเล็กทรอนิกส์และการเย็บปักถักร้อย เป็นความสามารถแทรกซ้อนที่ไม่ต้องการวัด ซึ่งแทนด้วย η_1 และ η_2 ตามลำดับ ข้อสอบทุกข้อในแบบทดสอบจะวัดความสามารถเป้าหมาย ส่วนข้อสอบทำหน้าที่ต่างกันจะวัดทั้งความสามารถเป้าหมาย และความสามารถแทรกซ้อน (Nandakumar, 1993)

ความสามารถ θ และ η ได้มาจากผลการตอบข้อสอบระหว่างผู้เข้าสอบกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ ส่วนผลการตอบข้อสอบได้มาจากการสุ่มจากกลุ่มผู้ตอบในแต่ละกลุ่ม ซึ่งแทนด้วย $U = (U_1, U_1,..., U_N)$ โดยที่ U_i เท่ากับ 1 ถ้าตอบข้อสอบข้อที่ i ถูก และ U_i เท่ากับ 0 ถ้าตอบข้อสอบผิด ส่วน N เป็นจำนวนข้อสอบ ในการสร้าง U ตามโมเดล IRT โดยทั่ว ๆ ไป ต้องประกอบด้วย 2 องค์ประกอบคือ *(1) พารามิเตอร์ความสามารถของผู้สอบจำนวน d มิติ* และ *(2) ฟังก์ชันการตอบสนองข้อสอบ (IRFs)* ซึ่งฟังก์ชันการตอบสนองข้อสอบในแต่ละข้อจะเป็นตัวกำหนดความน่าจะเป็นในการตอบข้อสอบได้ถูกต้อง สำหรับในที่นี้ $d = 2$ เพราะมีพารามิเตอร์ความสามารถสองชนิดคือ ความสามารถเป้าหมายที่ต้องการวัด (θ) และความสามารถแทรกซ้อนที่ไม่ต้องการวัด (η) ส่วนเวกเตอร์ความสามารถที่กำหนดจากผู้เข้าสอบในแต่ละกลุ่มแทนด้วย (θ, η) โดยที่การแจกแจงของ (θ, η) ถูกเลือกขึ้นมาอย่างสุ่มจากกลุ่มผู้เข้าสอบ และความสามารถที่สุ่มมาจากกลุ่มผู้เข้าสอบแทนด้วย (Θ, η) โดยสมมติว่าข้อสอบทุกข้อของแบบทดสอบจะวัดความสามารถ θ ส่วนข้อสอบบางข้อที่ทำหน้าที่ต่างกันจะวัดทั้งความสามารถ θ และ η ซึ่งความสามารถ η อาจมีเพียงความสามารถเดียวหรือมากกว่าหนึ่งความสามารถก็ได้ สำหรับ IRF ข้อที่ i ซึ่งขึ้นอยู่กับความสามารถ θ เพียงอย่างเดียว แทนด้วย $P_i(\theta)$ ส่วน IRF ข้อที่ i ซึ่งขึ้นอยู่กับความสามารถ θ และ η แทนด้วย $P_i(\theta, \eta)$ ดังนี้ (Shealy & Stout, 1993)

$$P_i(\theta) = c_i + \frac{(1 - c_i)}{1 + \exp\left[-1.7 a_{i0}(\theta - b_{i0})\right]} \quad , i = 1,...,N$$

$$P_i(\theta, \eta) = c_i + \frac{(1 - c_i)}{1 + \exp\left\{-1.7\left[a_{i0}(\theta - b_{i0}) + a_{i\eta}(\eta - b_{i\eta})\right]\right\}} \quad , i = 1,...,N$$

ค่า IRF ทั้ง 2 ลักษณะดังกล่าวเป็นโมเดลการตอบสนองข้อสอบแบบ 3 พารามิเตอร์ ซึ่งมีคุณสมบัติไม่แปรเปลี่ยน (Invariance) ไปตามกลุ่มผู้สอบ ทั้งยังมีข้อตกลงเกี่ยวกับความเป็นอิสระ

ต่อกันในการตอบข้อสอบ (Local Independence) ของ U เมื่อให้ IRF แทนด้วย $P_i(\theta, \eta)$ สามารถกำหนดในรูปของความน่าจะเป็นดังนี้ (Li & Stout, 1996)

$$P[U/(\Theta = \theta, \eta = \eta)] = \prod_{i=1}^{N} P_i(\theta,\eta)^{u_i}\left[1 - P_i(\theta,\eta)\right]^{1-u_i}$$

เชียลี และสเตาท์ (Shealy & Stout, 1993) ได้ใช้ Marginal IRFs อธิบายการทำหน้าที่ต่างกันของข้อสอบ ดังนี้

$$M_{ig}(\theta) = \int_{\eta} P_i(\theta,\eta) f_g(\eta/\theta)\, d\eta$$

เมื่อ $M_{ig}(\theta)$ แทน marginal IRF สำหรับความสามารถเป้าหมายที่ต้องการวัด θ ของผู้เข้าสอบกลุ่มอ้างอิงหรือกลุ่มเปรียบเทียบ

$P_i(\theta,\eta)$ แทน IRF ของข้อสอบข้อที่ i

$f_g(\eta/\theta)$ แทน การแจกแจงแบบมีเงื่อนไขของกลุ่มผู้เข้าสอบ

การทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป (Uniform DIF) เกิดขึ้นเมื่อโอกาสของการตอบข้อสอบถูกจากผู้สอบกลุ่มหนึ่งมีค่ามากกว่าผู้สอบอีกกลุ่มหนึ่งอย่างสม่ำเสมอในทุกระดับความสามารถ ซึ่งตามทฤษฎี IRT สามารถแสดงได้ในรูปโค้งลักษณะข้อสอบของผู้สอบสองกลุ่มไม่ตัดกัน (Noncrossing ICCs) ข้อสอบจะเข้าข้างผู้สอบกลุ่มใดนั้นให้พิจารณาค่า Marginal IRFs กล่าวคือ ถ้า $M_{iF}(\theta) < M_{iR}(\theta)$ ทุกค่าความสามารถ (θ) แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบเอกรูป โดยข้อสอบจะเข้าข้างผู้เข้าสอบกลุ่มอ้างอิง และ ถ้า $M_{iF}(\theta) > M_{iR}(\theta)$ ทุกค่าความสามารถ (θ) แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบเอกรูป โดยข้อสอบจะเข้าข้างผู้เข้าสอบกลุ่มเปรียบเทียบ การทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป อาจเรียกอีกอย่างหนึ่งว่า "*การทำหน้าที่ต่างกันแบบไม่ตัดกัน*" (Noncrossing DIF)

หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปด้วยวิธี SIBTEST ของเชียลี และสเตาท์ (Shealy & Stout, 1993) จะเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ โดยแบ่งแบบทดสอบออกเป็น 2 ชุดย่อย (Subtest) คือ (1) ชุดแบบทดสอบที่มีความตรง (Valid Subtests) หรือชุดแบบทดสอบที่ใช้เป็นเกณฑ์ในการเปรียบเทียบ (Matching Subtests) แบบทดสอบชุดนี้ประกอบด้วยข้อสอบที่ทำหน้าที่ไม่ต่างกัน และ (2) ชุดแบบทดสอบที่ต้องการศึกษา (Studied Subtests) ประกอบด้วยข้อสอบที่สงสัยว่าทำหน้าที่ต่างกัน ถ้าแบบทดสอบ

ชุดแรกมีจำนวน n ข้อ (ข้อที่ 1 ถึง n) แล้วแบบทดสอบชุดที่สองจะมีจำนวน N-n ข้อ (ข้อที่ n+1 ถึง N) เมื่อ N เป็นจำนวนข้อสอบทั้งหมด

ฟังก์ชันการตอบสนองข้อสอบของแบบทดสอบที่ต้องการศึกษา จากผู้เข้าสอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ กำหนดได้ ดังนี้

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

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

เมื่อ $M_{SR}(\theta)$ แทน ผลรวมของ Marginal IRFs ของข้อสอบที่ต้องการศึกษา จากผู้เข้าสอบกลุ่มอ้างอิง ณ ระดับความสามารถ θ

เมื่อ $M_{SF}(\theta)$ แทน ผลรวมของ Marginal IRFs ของข้อสอบที่ต้องการศึกษา จากผู้เข้าสอบกลุ่มเปรียบเทียบ ณ ระดับความสามารถ θ

n แทน จำนวนข้อสอบในชุดแบบทดสอบที่มีความตรง

N แทน จำนวนข้อสอบทั้งหมด

ขนาดของความสามารถแตกต่างระหว่าง $M_{SR}(\theta)$ กับ $M_{SF}(\theta)$ แสดงถึงปริมาณของการทดสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป หรือการทำหน้าที่ต่างกันแบบไม่ตัดกันจากชุดแบบทดสอบที่ต้องการศึกษา ณ ระดับความสามารถ θ ซึ่งสามารถคำนวณในรูปการอินทิเกรท ดังนี้

$$\beta_{uni} = \int_{\theta}[M_{SR}(\theta) - M_{SF}(\theta)]f_p(\theta)d\theta$$

เมื่อ β_{uni} แทน ดัชนีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป

$f_p(\theta)$ แทน ฟังก์ชันความหนาแน่นของโอกาสการแจกแจงความสามารถ θ ทั้ง 2 กลุ่ม

ดัชนี β_{uni} ที่คำนวณได้จากสูตรดังกล่าวข้างต้น นำมาทดสอบสมมติฐานของการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป ดังนี้

$$H_0 : \beta_{uni} = 0$$

$$H_1 : \beta_{uni} > 0$$

สมมติฐานอื่น (H_1) มีลักษณะทิศทางเดียว ซึ่งใช้ทดสอบการทำหน้าที่ต่างกันของข้อสอบที่เข้าข้างผู้เข้าสอบกลุ่มเปรียบเทียบ สำหรับค่าประมาณของ β_{uni} คำนวณได้จากคะแนนรวมของชุดแบบทดสอบที่มีความตรงและชุดแบบทดสอบที่ต้องการศึกษา ซึ่งกำหนดด้วยสัญลักษณ์ดังนี้

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

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

เมื่อ X แทน คะแนนรวมของชุดแบบทดสอบที่มีความตรง (ใช้เป็นเกณฑ์ในการเปรียบเทียบ)

Y แทน คะแนนรวมของชุดแบบทดสอบที่ต้องการศึกษา

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

นำคะแนนเฉลี่ยจากผลการตอบข้อสอบชุดแบบทดสอบที่ต้องการศึกษาระหว่างผู้เข้าสอบกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบที่มีความสามารถระดับเดียวกันมาจับคู่เปรียบเทียบกัน ซึ่งพิจารณาจากคะแนนรวมที่เท่ากันของชุดแบบทดสอบที่มีความตรง ($X = k$) เขียนในรูปสัญลักษณ์ได้ ดังนี้

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

เมื่อ $\bar{Y}_{Rk}$ แทน ค่าเฉลี่ยของคะแนนรายข้อ จากชุดแบบทดสอบที่ต้องการศึกษาของผู้เข้าสอบกลุ่มอ้างอิง ซึ่งได้คะแนน $X = k$

$\bar{Y}_{Fk}$ แทน ค่าเฉลี่ยของคะแนนรายข้อ จากชุดแบบทดสอบที่ต้องการศึกษาของผู้เข้าสอบกลุ่มเปรียบเทียบ ซึ่งได้คะแนน $X = k$

k แทน คะแนนรวมจากชุดแบบทดสอบที่มีความตรง

ค่า $\bar{Y}_{Rk} - \bar{Y}_{Fk}$ ดังกล่าวเป็นความแตกต่างของผลการตอบข้อสอบในชุดแบบทดสอบที่ต้องการศึกษา ระหว่างผู้เข้าสอบกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบที่มีความสามารถระดับเดียวกัน ถ้า $\bar{Y}_{Rk} - \bar{Y}_{Fk} = 0$ ทุกคะแนน k แสดงว่าข้อสอบที่ต้องการศึกษาทำหน้าที่ไม่ต่างกัน และถ้า $\bar{Y}_{Rk} - \bar{Y}_{Fk} > 0$ ทุกคะแนน k แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบเอกรูป โดยจะลำเอียงเข้าข้างผู้เข้าสอบกลุ่มอ้างอิง ค่าความแตกต่างของผลการตอบข้อสอบสามารถประมาณค่าในรูป β_{uni} ได้ดังนี้

$$\hat{\beta}_{uni} = \sum_{k=0}^{n} \hat{P}_k \left(\bar{Y}_{Rk} - \bar{Y}_{Fk} \right)$$

โดยที่ $$\hat{P}_k = \frac{\left(J_{Rk} + J_{Fk} \right)}{\sum_{k=0}^{n} \left(J_{Rk} + J_{Fk} \right)}$$

เมื่อ P_k แทน สัดส่วนของจำนวนผู้เข้าสอบทั้งหมด (กลุ่มอ้างอิงและกลุ่มเปรียบเทียบ) ซึ่งตอบชุดแบบทดสอบที่มีความตรงแล้วได้คะแนนรวม $X = k$

J_{Fk} แทน จำนวนผู้เข้าสอบกลุ่มเปรียบเทียบซึ่งตอบชุดแบบทดสอบที่มีความตรงแล้วได้คะแนนรวม $X = k$

J_{Rk} แทน จำนวนผู้เข้าสอบกลุ่มอ้างอิงซึ่งตอบชุดแบบทดสอบที่มีความตรงแล้วได้คะแนนรวม $X = k$

สำหรับการทดสอบสมมติฐานศูนย์ของข้อสอบทำหน้าที่ไม่ต่างกัน (NO DIF) ใช้สถิติ β_{uni} ดังนี้

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

โดยที่ $$\hat{\sigma}(\hat{\beta}_{uni}) = \sqrt{\sum_{k=0}^{n} \hat{P}_k^2 \left[\frac{1}{J_{Rk}} \hat{\sigma}^2 (Y / k, R) + \frac{1}{J_{Fk}} \hat{\sigma}^2 (Y / k, F) \right]}$$

เมื่อ $\sigma(\beta_{uni})$ แทน ค่าประมาณความคลาดเคลื่อนมาตรฐานของ β_{uni}

$\sigma^2(Y/k,g)$ แทน ค่าประมาณความแปรปรวนของคะแนนจากชุดแบบทดสอบที่ต้องการศึกษา สำหรับผู้เข้าสอบกลุ่ม g (R หรือ F) ซึ่งมีคะแนนรวมเท่ากับ k

J_{gk} แทน จำนวนผู้เข้าสอบกลุ่ม g (R หรือF) ซึ่งตอบชุดแบบทดสอบที่มีความตรงแล้วได้คะแนนรวม $X = k$

สถิติที่ใช้ในการทดสอบ β_{uni} มีการแจกแจงในลักษณะปกติมาตรฐาน [$N(0,1)$] เมื่อ $\beta_{uni} = 0$ และถ้าผลการทดสอบปรากฏว่า $\beta_{uni} > Z_\alpha$ อย่างมีนัยสำคัญที่ระดับ α โดยที่ $\alpha = P[N(0,1) > Z_\alpha]$ แสดงว่า ปฏิเสธ H_0 นั่นคือ ข้อสอบที่นำมาตรวจสอบการทำหน้าที่ต่างกัน เมื่อ β_{uni} มีค่าเป็นบวก

โดยจะลำเอียงเข้าข้างผู้เข้าสอบกลุ่มอ้างอิง และเมื่อ β_{uni} มีค่าเป็นลบ จะลำเอียงเข้าข้างผู้เข้าสอบกลุ่มเปรียบเทียบ

อย่างไรก็ตาม สถิติที่ใช้ในการทดสอบสำหรับสรุปอ้างอิง การทำหน้าที่ต่างกันของข้อสอบดังกล่าวมักจะมีปัญหาในกรณีที่มีความแตกต่างของการแจกแจงความสามารถเป้าหมาย ระหว่างกลุ่มผู้เข้าสอบ กล่าวคือ ถ้าผู้เข้าสอบกลุ่มอ้างอิงมีความสามารถเป็นหมายสูงกว่าผู้เข้าสอบกลุ่มเปรียบเทียบ จะเกิดผลกระทบทำให้สถิติ β_{uni} มีค่าเฟ้อ (Inflate) หรือมีค่าสูงผิดปกติ ถึงแม้ว่าในความเป็นจริงข้อสอบทำหน้าที่ไม่ต่างกัน ดังนั้นจึงมีความจำเป็นที่จะแก้ไขความแตกต่างของการแจกแจงความสามารถเป้าหมายด้วยวิธีการปรับแก้ค่าการถดถอย (Regression Correction) เพื่อกำจัดอิทธิพลค่าเฟ้อของผลกระทบดังกล่าว โดยปรับแก้ค่า $\bar{Y}_{Rk}$ และ $\bar{Y}_{Fk}$ เป็นรายคู่ก่อนการคำนวณ β_{uni}

งานวิจัยที่เกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีซิปเทสท์

กายจนา วัธนสุนทร (2537) ได้พัฒนาเกณฑ์ตัดสินข้อสอบลำเอียงทางเพศด้วยข้อมูลเชิงประจักษ์สำหรับดัชนี 4 ตัว คือ พื้นที่ระหว่างโค้งการตอบข้อสอบชนิดคิดเครื่องหมาย (SA) และไม่คิดเครื่องหมาย (UA) จากวิธีทฤษฎีการตอบข้อสอบแบบ 2 พารามิเตอร์ ดัชนีแอลฟา (α_{MH}) จากวิธีแมนเทล-แฮนส์เซล และดัชนีเบต้า (β_{SIB}) จากวิธีซิปเทสท์ โดยใช้ข้อมูลการตอบข้อสอบคัดเลือกบุคคลเข้าศึกษาในสถาบันอุดมศึกษาของทบวงมหาวิทยาลัย ปีการศึกษา 2535 ใช้แบบทดสอบขนาด 20, 30 และ 40 ข้อ สำหรับวิชาคณิตศาสตร์ และ 50, 60, 70 และ 80 ข้อ สำหรับวิชาภาษาอังกฤษ ใช้กลุ่มผู้เข้าสอบ 6 ขนาด คือ 100, 200, 400, 600, 800 และ 1,000 คน

การพัฒนาเกณฑ์กระทำโดยคำนวณค่าดัชนีทั้ง 4 ตัว จากข้อมูลการตอบข้อสอบของผู้สอบเพศเดียวกัน เพศละ 50 ค่า สำหรับแต่ละความยาวของแบบทดสอบและขนาดของผู้สอบ จากนั้นนำค่าดัชนีที่ได้ทั้งหมดมาวิเคราะห์ค่าเฉลี่ยและกำหนดเกณฑ์จากค่าเฉลี่ย 2 ลักษณะ คือ เกณฑ์ที่กำหนดจากค่าเฉลี่ย ซึ่งรวมค่าดัชนีทุกข้อโดยไม่พิจารณาความแตกต่างในด้านความยาวของแบบทดสอบและขนาดของผู้เข้าสอบ และเกณฑ์ที่กำหนดจากค่าเฉลี่ยที่พิจารณาถึงความยาวของแบบทดสอบและขนาดผู้สอบด้วย จากนั้นนำเกณฑ์ที่กำหนดไว้ไปตัดสินค่าดัชนีที่ได้จากการวิเคราะห์ระหว่างผู้สอบเพศหญิงกับชาย ปรากฏว่าความสอดคล้องของการตัดสินภายในดัชนีเดียวกันมีความไม่คงที่ข้ามขนาดผู้เข้าสอบ อย่างไรก็ตาม ความสอดคล้องมีแนวโน้มสูงขึ้นที่ขนาดผู้สอบตั้งแต่ 600 คน ขึ้นไป ผลการศึกษาปรากฏดังต่อไปนี้

1. เกณฑ์ที่พัฒนาจากข้อมูลเชิงประจักษ์เพื่อใช้ในการตัดสินความลำเอียงของข้อสอบระหว่างผู้สอบหญิงกับชาย เป็นดังนี้

(1) $|SA| > .80$ และ $UA > .05$ กรณีความยาวของแบบทดสอบต่ำกว่า 50 ข้อ

(2) |SA| > .40 และ UA > 1.20 กรณีความยาวของแบบทดสอบ 50 ข้อ ขึ้นไป

(3) .60 > α_{MH} > 1.40 และ |β_{SIB}| > .60 ทุกความยาวของแบบทดสอบและทุกขนาดผู้เข้าสอบ ทั้งนี้ควรใช้ขนาดผู้เข้าสอบอย่างน้อย 800 คน สำหรับดัชนี SA และ UA และ 600 คน สำหรับดัชนี α_{MH} และ β_{SIB}

2. การตรวจค้นข้อสอบลำเอียงทางเพศมีความไม่คงที่ข้ามขนาดผู้เข้าสอบ และความยาวของแบบทดสอบ

3. ความสอดคล้องในการตรวจค้นข้อสอบลำเอียงภายในวิธีเดียวกันข้ามขนาดผู้สอบค่อนข้างต่ำ แต่จะสูงขึ้นที่ขนาดผู้เข้าสอบตั้งแต่ 600 คน ขึ้นไป

4. ข้อสอบลำเอียงวิชาคณิตศาสตร์ส่วนใหญ่ลำเอียงเข้าข้างผู้สอบชาย และวิชาภาษาอังกฤษลำเอียงเข้าข้างผู้สอบหญิง เมื่อใช้ดัชนี SA และ α_{MH} แต่ดัชนี β_{SIB} ให้ผลตรงข้าม

จิติมา วรรณศรี (2539) ได้เปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันระหว่างวิธีแมนเทล-แฮนส์เซลกับวิธีซิปเทสท์ โดยศึกษาจากข้อมูลจำลอง ตัวแปรที่ศึกษาได้แก่ ความยาวของแบบทดสอบ 3 ขนาด คือ 30 ข้อ 60 ข้อ และ 90 ข้อ ขนาดกลุ่มตัวอย่าง 3 ขนาด คือ 200 คน 600 คน และ 1,000 คน โดยแต่ละขนาดมีอัตราส่วนระหว่างผู้เข้าสอบกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบต่างกัน คือ 1: 1 1: 0.9 1: 0.75 และ 1: 0.5 ผลการศึกษาปรากฏว่า วิธีแมนเทล-แฮนส์เซล กับวิธีซิปเทสท์ มีประสิทธิภาพเท่าเทียมกันในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในทุกขนาดกลุ่มตัวอย่างและทุกอัตราส่วนภายใต้ความยาวของแบบทดสอบเดียวกัน และเมื่อใช้แบบทดสอบที่มีความยาวปานกลาง(60 ข้อ) วิธีทั้งสองสามารถตรวจสอบได้อย่างมีประสิทธิภาพที่สุด นอกจากนี้เมื่อใช้ขนาดกลุ่มตัวอย่างมากขึ้นจะสามารถตรวจสอบข้อสอบทำหน้าที่ต่างกันได้ถูกต้องมากขึ้น โดยส่วนมากวิธีซิปเทสท์มีอัตราความคลาดเคลื่อนประเภทที่ 1 มากกว่าวิธีแมนเทล-แฮนส์เซล เล็กน้อย

พรรณี จินตมาศ (2540) ได้ศึกษาผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับเพศจากแบบทดสอบคณิตศาสตร์โจทย์ปัญหา ที่ผู้วิจัยสร้างขึ้นเอง โดยใช้วิธีวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ 3 วิธี คือ วิธีแปลงค่าความยาก วิธีแมนเทล-แฮนส์เซล และวิธีซิปเทสท์ ในกลุ่มผู้เข้าสอบคือ ขนาดกลุ่มผู้เข้าสอบ 500 คน และขนาดกลุ่มผู้เข้าสอบ 1,000 คน โดยเปรียบเทียบจำนวนข้อสอบทำหน้าที่ต่างกัน และเปรียบเทียบค่าความเชื่อมั่นแบบแบ่งครึ่งฉบับของแบบทดสอบหลังคัดเลือกข้อสอบทำหน้าที่ต่างกันออกแล้ว กลุ่มตัวอย่างเป็นนักเรียนชั้นมัธยมศึกษาปีที่ 1 ภาคเรียนที่ 2 ปีการศึกษา 2539 ของโรงเรียนสังกัดกรมสามัญศึกษา ส่วนกลาง จำนวน 2,200 คน ซึ่งได้มาโดยการสุ่มแบบแบ่งชั้น มีขนาดโรงเรียนเป็นชั้น และโรงเรียนเป็นหน่วยการสุ่ม ผลการศึกษาปรากฏว่า เมื่อวิเคราะห์จากกลุ่มผู้เข้าสอบขนาด 500 คน และวิธีซิปเทสท์พบข้อสอบทำหน้าที่ต่างกันมากที่สุด

และวิธีแปลงค่าความยากพบข้อสอบที่ทำหน้าที่ต่างกันน้อยที่สุด โดยจำนวนข้อสอบทำหน้าที่ต่างกันแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติทุกวิธีวิเคราะห์ และเมื่อวิเคราะห์จากกลุ่มผู้เข้าสอบ 1,000 คน วิธีแมนเทล-แฮนส์เซล พบข้อสอบทำหน้าที่ต่างกันมากที่สุด วิธีแปลงค่าความยากไม่พบข้อสอบทำหน้าที่ต่างกัน โดยจำนวนข้อสอบที่ทำหน้าที่ต่างกันจากการวิเคราะห์ด้วยวิธีแปลงค่าความยากกับวิธีแมนเทล-แฮนส์เซล และวิธีแปลงค่าความยากกับวิธีซิปเทสท์แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นอกนั้นแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ สำหรับจำนวนข้อสอบทำหน้าที่ต่างกันจากการวิเคราะห์ด้วยวิธีแปลงค่าความยากระหว่างกลุ่มผู้เข้าสอบ 500 คน และขนาดกลุ่มผู้เข้าสอบ 1,000 คน จะมีจำนวนข้อสอบทำหน้าที่ต่างกัน แตกต่างกันมีนัยสำคัญทางสถิติที่ระดับ .05 นอกนั้นมีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

อารี วัชรโสตถิกุล (2543) ได้เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้รูปแบบต่างกันคือ รูปแบบคะแนนรวมทั้งฉบับ แยกตามเนื้อหา และแยกตามระดับพฤติกรรม ด้วยใช้วิธีการตรวจสอบต่างกัน คือ วิธีซิปเทสท์ และวิธีถดถอยโลจิสติก แล้วคัดเลือกข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบ เพื่อเปรียบเทียบค่าความเชื่อมั่น ผลการศึกษาปรากฏว่า ในรูปแบบรวมทั้งฉบับพบข้อสอบทำหน้าที่ต่างกันโดยใช้วิธีการตรวจสอบต่างกันแตกต่างกัน ส่วนรูปแบบแยกตามเนื้อหา และแยกตามระดับพฤติกรรมไม่แตกต่างกัน

วลีมาศ แซ่อึ้ง (2543) ได้เปรียบเทียบอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูประหว่างวิธีซิปเทสท์ปรับใหม่ วิธีซิปเทสท์ วิธีแมนเทล-แฮนส์เซล และวิธีถดถอยโลจิสติก โดยจำลองข้อมูลด้วยโมเดลโลจิสติกชนิด 3 พารามิเตอร์ ในเงื่อนไข 324 เงื่อนไข ผลการศึกษาปรากฏว่า อำนาจการทดสอบของวิธีซิปเทสท์ปรับใหม่และวิธีถดถอยโลจิสติกมีค่าเท่าเทียมกันเกือบทุกเงื่อนไข และทั้งสองวิธีมีอำนาจการทดสอบสูงกว่าวิธีซิปเทสท์และวิธีแมนเทล-แฮนส์เซลเกือบทุกเงื่อนไข อัตราความคลาดเคลื่อนประเภทที่ 1 ที่ตรวจพบด้วยวิธีการทั้ง 4 มีค่าอยู่ภายในเกณฑ์ของอัตราความคลาดเคลื่อนที่ระดับ 10 % เกือบทุกเงื่อนไข

วิภา จำมั่น (2544) ได้เปรียบเทียบผลการตรวจสอบข้อสอบที่ลำเอียงของแบบทดสอบวัดความสามารถด้านภาษา เมื่อตรวจสอบด้วยวิธีซิปเทสท์กับวิธีแมนเทล-แฮนส์เซล ในกลุ่มข้อสอบที่มีระดับความง่ายต่ำ กลุ่มข้อสอบที่มีระดับความง่ายปานกลางและกลุ่มข้อสอบที่มีความง่ายสูง โดยเปรียบเทียบข้อสอบที่มีความลำเอียงและค่าความเที่ยงของแบบทดสอบหลังจากคัดเลือกข้อสอบที่มีความลำเอียงออกแล้ว กลุ่มตัวอย่างเป็นนักเรียนชั้นมัธยมศึกษาปีที่ 3 ภาคเรียนที่ 1 ปีการศึกษา 2544 ของโรงเรียนสังกัดคณะกรรมการการศึกษาขั้นพื้นฐาน จังหวัดลพบุรี จำนวน 1,041 คน ซึ่งได้มาโดยการสุ่มแบบแบ่งชั้น ผลการศึกษาปรากฏว่า

1. ข้อสอบที่มีความลำเอียง เมื่อตรวจสอบด้วยวิธีซิปเทสท์ ระหว่างกลุ่มข้อสอบที่มีระดับความง่ายต่ำ กลุ่มข้อสอบที่มีระดับความง่ายปานกลาง และกลุ่มข้อสอบที่มีความง่ายสูง มีจำนวนข้อแตกต่างกันอย่างมีนัยสำคัญทางสถิติระดับ .05 ส่วนข้อสอบที่มีความลำเอียงระหว่างกลุ่มข้อสอบที่มีระดับความง่ายต่ำกับกลุ่มข้อสอบที่มีระดับความง่ายปานกลาง และระหว่างกลุ่มข้อสอบที่มีระดับความง่ายปานกลางกับกลุ่มข้อสอบที่มีความง่ายสูง มีจำนวนข้อแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ เมื่อตรวจสอบด้วยวิธีแมนเทล-แฮนส์เซล พบว่าข้อสอบที่มีความลำเอียงระหว่างกลุ่มผู้สอบที่มีระดับความง่ายต่ำ กลุ่มข้อสอบที่มีระดับความง่ายปานกลาง และกลุ่มข้อสอบที่มีความง่ายสูง มีจำนวนข้อแตกต่างกันอย่างไม่มีนัยสำคัญ

2. จำนวนข้อสอบที่มีความลำเอียงระหว่างการตรวจสอบด้วยวิธีซิปเทสท์กับวิธีแมนเทล-แฮนส์เซล ในกลุ่มข้อสอบที่มีระดับความง่ายต่ำ กลุ่มข้อสอบที่มีระดับความง่ายปานกลางและกลุ่มข้อสอบที่มีความง่ายสูง มีจำนวนข้อแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

3. ค่าความเที่ยงของแบบทดสอบหลังจากคัดเลือกข้อสอบที่มีความลำเอียงออกแล้ว ระหว่างกลุ่มข้อสอบที่มีระดับความง่ายต่ำ กลุ่มข้อสอบที่มีระดับความง่ายปานกลาง และกลุ่มข้อสอบที่มีความง่ายสูง วิธีซิปเทสท์กับวิธีแมนเทล-แฮนส์เซล มีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

สิริรัตน์ วิภาสศิลป์ (2545) ได้เปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบหมวดข้อสอบ และแบบทดสอบ จากข้อมูลการตอบข้อสอบที่ใช้ความสามารถหลายมิติ ระหว่างวิธีซิปเทสท์กับวิธีดีเอฟไอที ภายใต้เงื่อนไข ความยาวของแบบทดสอบ 30, 40 และ 50 ข้อ กลุ่มตัวอย่างขนาด 50, 100, 200, 500 และ 1,000 คน กลุ่มตัวอย่างในการศึกษาได้มาจากการสุ่มแบบใส่คืนจากประชากรเทียม ซึ่งกำหนดจากนักเรียนชาย และนักเรียนหญิง ชั้นมัธยมศึกษาปีที่ 1 จังหวัดนนทบุรี แต่ละขนาดสุ่มกลุ่มตัวอย่าง 50 ครั้ง เครื่องมือที่ใช้ในการวิจัยเป็นแบบทดสอบวิชาคณิตศาสตร์ที่ผู้วิจัยสร้างขึ้นซึ่งเป็นข้อสอบแบบหลายตัวเลือก 5 ตัวเลือก จำนวน 50 ข้อ มีข้อสอบที่ผู้เชี่ยวชาญพิจารณาว่าเป็นข้อสอบที่ทำหน้าที่ต่างกันต่อเพศชาย จำนวน 16 ข้อ หลังจากเก็บรวมรวบข้อมูลแล้วคัดเลือกข้อสอบตามสัดส่วนในตารางกำหนดข้อสอบ จัดเป็นแบบทดสอบที่มีความยาว 30 และ 40 ข้อ แล้วตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ หมวดข้อสอบ และแบบทดสอบในเงื่อนไขต่าง ๆ โดยโปรแกรม SIBTEST และ DFIT จากนั้นนำผลที่ได้ไปเปรียบเทียบความถูกต้อง และการระบุผิดพลาดในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีเดียวกันและต่างกัน โดยการวิเคราะห์ความแปรปรวนแบบตัวแปรพหุ คำนวณความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบ และแบบทดสอบด้วยวิธีซิปเทสท์และวิธีดีเอฟไอที โดยใช้สถิติ Z-test ผลการศึกษาปรากฏว่า

1. เมื่อแบบทดสอบประกอบด้วยข้อสอบ 30, 40 และ 50 ข้อ กลุ่มตัวอย่างขนาด 50, 100 และ 200 คน ส่งผลต่อความถูกต้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีซิปเทสท์ไม่แตกต่างกัน กลุ่มตัวอย่างขนาด 500 และ 1,000 คน ส่งผลต่อความถูกต้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีซิปเทสท์สูงกว่ากลุ่มตัวอย่างขนาด 50, 100 และ 200 คน แต่การระบุผิดพลาดในการตรวจสอบสูงกว่าด้วย เมื่อตรวจสอบด้วยวิธีดีเอฟไอที พบว่า กลุ่มตัวอย่างขนาด 50, 100, 200, 500 และ 1,000 คน ส่งผลต่อความถูกต้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบไม่แตกต่างกัน

2. ทุกเงื่อนไขความยาวแบบทดสอบและกลุ่มตัวอย่างขนาดแตกต่างกัน วิธีซิปเทสท์มีความถูกต้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบน้อยกว่าวิธีดีเอฟไอที และพบว่าความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีทั้งสองมีค่าต่ำกว่าร้อยละ 1

3. วิธีซิปเทสท์มีความถูกต้องในการตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบมากกว่าวิธีดีเอฟไอที เมื่อแบบทดสอบมีข้อสอบ 30 ข้อ กลุ่มตัวอย่าง 1,000 คน และเมื่อแบบทดสอบมี 40 ข้อ กลุ่มตัวอย่างขนาด 500 คน

4. วิธีซิปเทสท์มีความถูดต้องในการตรวจสอบการทำหน้าที่ต่างกันของแบบทดสอบมากกว่าวิธีดีเอฟไอที เมื่อแบบทดสอบมีข้อสอบ 50 ข้อ กลุ่มตัวอย่างขนาด 100, 200 และ 1,000 คน

ศุภวัฒน์ มะลิเผือก (2548) ได้ศึกษาอิทธิพลของการทำหน้าที่ต่างกันของข้อสอบที่ส่งผลต่อคุณภาพของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ วิชาภาษาไทย ชั้นประถมศึกษาปีที่ 6 ปีการศึกษา 2546 ด้วยวิธีซิปเทสท์กับวิธีฉดถอยโลจิสติก โดยการเปรียบเทียบค่าความเที่ยง ความตรงเชิงโครงสร้าง ความคงที่ของโครงสร้างองค์ประกอบ และค่าสัมประสิทธิ์สหสัมพันธ์อันดับของผู้สอบ ระหว่างแบบทดสอบฉบับก่อนกับฉบับหลังตัดข้อสอบทำหน้าที่ต่างกันออก ผลการศึกษาปรากฏว่า พบข้อสอบทำหน้าที่ต่างกัน เมื่อจำแนกกลุ่มผู้สอบตามตัวแปรเพศ จำนวน 12 ข้อ คิดเป็นร้อยละ 30 แบบทดสอบฉบับก่อนกับฉบับหลังตัดข้อสอบทำหน้าที่ต่างกันมีค่าความเที่ยงแตกต่างกัน ความตรงเชิงโครงสร้างไม่แตกต่างกัน โครงสร้างองค์ประกอบคงที่ และค่าสัมประสิทธิ์สหสัมพันธ์อันดับของผู้สอบ มีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05

เชียลี และสเตาท์ (Shealy & Stout, 1993, pp. 159-194) ได้ศึกษาประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธีซิปเทสท์กับวิธีแมนเทล-แฮนส์เซล ในกรณีที่มีข้อสอบทำหน้าที่ต่างกันข้อเดียว โดยใช้ข้อมูลจากผลการตอบแบบทดสอบวิชาคณิตศาสตร์ของ ACT และแบบทดสอบของ ASVAB ใช้ขนาดกลุ่มตัวอย่างของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ

ต่างกัน ผลการศึกษาปรากฏว่า ในกรณีที่มีข้อสอบทำหน้าที่ต่างกันข้อเดียวทั้งวิธีแมนเทล-แฮนส์เซล มีอำนาจในการตรวจสอบดีเท่าเทียมกัน วิธีซิปเทสท์มีประสิทธิภาพดีในการตรวจสอบกรณีที่มีข้อสอบทำหน้าที่ต่างกันหลาย ๆ ข้อ และทั้งวิธีซิปเทสท์กับวิธีแมนเทล-แฮนส์เซล มีประสิทธิภาพดีเมื่อใช้กับแบบทดสอบที่มีความยาวพอสมควร ($\geq$ 25 ข้อ)

นารายานัน และสวามินาทาน (Narayanan & Swaminathan, 1994, pp. 315-328) ได้เปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป ด้วยวิธีแมนเทล-แฮนส์เซล และวิธีซิปเทสท์ โดยใช้ข้อมูลที่จำลองภายใต้เงื่อนไข 1,296 เงื่อนไข (9x3x2x4x6) แล้วคำนวณซ้ำ 100 ครั้ง แต่ละเงื่อนไขที่ศึกษาประกอบด้วย

1. ขนาดกลุ่มตัวอย่างที่ต่างกัน 3 ขนาด คือกลุ่มเปรียบเทียบใช้ขนาด 100, 200 และ 300 คน กลุ่มอ้างอิงใช้ขนาด 300, 500 และ 1,000 คน วางไขว้กันได้กลุ่มตัวอย่างทั้งหมด 9 กลุ่ม

2. ความแตกต่างในการแจกแจงความสามารถ 3 ระดับ ระดับที่ 1 ให้ทั้ง 2 กลุ่มมีค่าเฉลี่ยการแจกแจงความสามารถอยู่ที่ 0.0 ระดับที่ 2 ให้ค่าเฉลี่ยของความสามารถในกลุ่มอ้างอิงและกลุ่มเปรียบเทียบเป็น 0.0 และ 0.5 ตามลำดับ และระดับที่ 3 ค่าเฉลี่ยความสามารถในกลุ่มเปรียบเทียบเป็น 0.0 และ –1.0 ตามลำดับ ส่วนเบี่ยงเบนมาตรฐานของทั้ง 3 ระดับเป็น 1.0

3. ร้อยละของข้อสอบทำหน้าที่ต่างกันมี 2 ขนาด คือ ร้อยละ 10 และร้อยละ 20 จากแบบทดสอบ จำนวน 40 ข้อ

4. ขนาดอิทธิพลของข้อสอบที่ทำหน้าที่ต่างกัน 4 ระดับ คือ 0.4, 0.6, 0.8 และ 1.0

5. ลักษณะของข้อสอบ 6 ระดับ (ลักษณะ b ต่ำกับ a ปานกลาง, b ต่ำกับ a สูง, b ปานกลางกับ a ต่ำ, b ปานกลางกับ a สูง, b สูงกับ a ต่ำ, b สูงกับ a ปานกลาง)

ผลการศึกษาปรากฏว่า วิธีแมนเทล-แฮนส์เซล และวิธีซิปเทสท์ สามารถตรวจพบข้อสอบทำหน้าที่ต่างกันได้ใกล้เคียงกัน โดยตรวจพบได้ดีในขนาดกลุ่มตัวอย่าง 600 คน (กลุ่มอ้างอิง และกลุ่มเปรียบเทียบกลุ่มละ 300 คน) ขนาดของข้อสอบทำหน้าที่ต่างกันทั้ง 4 ขนาดมีผลต่อการตรวจพบการทำหน้าที่ต่างกันของข้อสอบทั้ง 2 วิธีอย่างมีนัยสำคัญทางสถิติ ความแตกต่างในการแจกแจงความสามารถไม่มีผลต่อวิธีการวิเคราะห์การทำหน้าที่ต่างกันวิธีซิปเทสท์ แต่มีผลต่อวิธี แมนเทล-แฮนส์เซล ที่เป็นเช่นนี้อาจเป็นเพราะวิธีซิปเทสท์ ได้มีการปรับแก้การถดถอย ขนาดของแบบทดสอบและร้อยละของข้อสอบทำหน้าที่ต่างกันไม่มีผลต่อวิธีทั้งสอง สำหรับการเกิดความคลาดเคลื่อนแบบที่ 1 วิธีแมนเทล-แฮนส์เซล จะมีอัตราการเกิดอยู่ในระดับปกติ แต่วิธีซิปเทสท์ จะเพิ่มเล็กน้อยในกลุ่มที่มีความสามารถเท่ากัน แต่ในกลุ่มที่มีความสามารถไม่เท่ากันการเกิดความคลาดเคลื่อนแบบที่ 1 จะมีอัตราการเกิดสูงขึ้นทั้ง 2 วิธี ซึ่งการเพิ่มความยาวของแบบทดสอบจะทำให้เกิดความเชื่อมั่นสูงขึ้นซึ่งมีผลให้ความคลาดเคลื่อนประเภทที่ 1 น่าจะมีอัตราการเกิดลดลง

นารายานัน และสวามินาทาน (Narayanan & Swaminathan, 1996, pp. 257-274) ได้เปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปด้วยวิธีแมนเทล-แฮนส์เซล วิธีถดถอยโลจิสติก และวิธีซิปเทสท์ โดยศึกษาอำนาจการตรวจสอบและการจำแนกผิดพลาดโดยการจำลองข้อมูล ตัวแปรที่ศึกษา ได้แก่ ขนาดกลุ่มตัวอย่าง ความแตกต่างของการกระจายความสามารถระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ สัดส่วนของข้อสอบทำหน้าที่ต่างกันที่มีภายในแบบทดสอบ ขนาดของพื้นที่ระหว่างโค้งคุณลักษณะข้อสอบของผู้เข้าสอบ 2 กลุ่ม ค่าความยากและค่าอำนาจจำแนกของแบบทดสอบ ผลการศึกษาปรากฏว่า วิธีซิปเทสท์ และวิธีถดถอยโลจิสติก มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปได้เท่าเทียมกันในทุกเงื่อนไขที่ศึกษา ส่วนวิธีแมนเทล-แฮนส์เซล ไม่สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูป และการจำแนกผิดพลาด วิธีแมนเทล-แฮนส์เซล จะสูงกว่าวิธีถดถอยโลจิสติก และวิธีซิปเทสท์

รุสโซ และสเตาท์ (Roussos & Stout, 1996, pp. 215-230) ได้ศึกษาสถานการณ์จำลองของผลกระทบของกลุ่มตัวอย่างขนาดเล็กและค่าพารามิเตอร์ของข้อสอบที่มีต่ออัตราความคลาดเคลื่อนประเภทที่ 1 ด้วยวิธีซิปเทสท์และวิธีแมนเทล-แฮนส์เซล โดยจำลองข้อมูล 2 ครั้ง ครั้งแรก ศึกษากลุ่มตัวอย่างขนาดเล็ก ใช้กลุ่มตัวอย่าง 4 ขนาด (100, 200, 500 และ1,000) และความแตกต่างของการแจกแจงค่าความสามารถระหว่างผู้เข้าสอบกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ 3 ขนาด (0, 0.5 และ 1) ใช้แบบทดสอบ 25 ข้อ ครั้งที่ 2 ศึกษาค่าพารามิเตอร์ของข้อสอบ ใช้กลุ่มตัวอย่าง 3 ขนาด (500, 1,000 และ 3,000) และความแตกต่างของการแจกแจงค่าความสามารถระหว่างผู้เข้าสอบกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ 2 ขนาด (0 และ 1) เลือกค่าพารามิเตอร์อำนาจจำแนก 3 ค่า (0.4, 1.0 และ 2.5) พารามิเตอร์ความยาก 5 ค่า (-1.5, -0.5, 0, 0.5 และ 1.5) และพารามิเตอร์การเดา 1 ค่า (.20) ผลการศึกษาพบว่า ครั้งที่ 1 เมื่อศึกษากลุ่มตัวอย่างขนาดเล็ก พบว่า อัตราความคลาดเคลื่อนประเภทที่ 1 ระหว่างวิธีซิปเทสท์และวิธีแมนเทล-แฮนส์เซล มีค่าไม่แตกต่างกัน ครั้งที่ 2 เมื่อศึกษาค่าพารามิเตอร์ของข้อสอบ พบว่า เมื่อความแตกต่างของการแจกแจงค่าความสามารถระหว่างผู้เข้าสอบกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบไม่แตกต่างกัน และอัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีซิปเทสท์มีค่าต่ำกว่าวิธีแมนเทล-แฮนส์เซล ในทุกเงื่อนไข

ดักกลัส รุสโซ และสเตาท์ (Douglas, Roussos, & Stout, 1996 cited in Roussos & Stout, 1996) ได้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยพิจารณาข้อสอบพร้อม ๆ กันครั้งละหลาย ๆ ข้อ โดยนำเสนอวิธีการตรวจสอบข้อสอบที่คาดว่าจะแสดงการทำหน้าที่ต่างกันเมื่อข้อสอบเหล่านั่นถูกนำมาพิจารณาพร้อมกัน 2 วิธี คือ ใช้เฉพาะความคิดเห็นของผู้เชี่ยวชาญและใช้วิธีการทางสถิติและตรวจสอบซ้ำด้วยความเห็นของผู้เชี่ยวชาญ พร้อมทั้งแสดงตัวอย่างการตรวจสอบการทำหน้าที่ต่างกัน 3 ตัวอย่าง ดังนี้ ตัวอย่างที่ 1 เลือกหมวดข้อสอบโดยใช้ความเห็นของผู้เชี่ยวชาญอย่างเดียว

ใช้ผู้เชี่ยวชาญชาย 3 คน และหญิง 1 คน พิจารณาแบบทดสอบย่อยการใช้เหตุผลเชิงตรรกศาสตร์ (Logical Reasoning Subtest) ที่ได้จากการดำเนินการสอบในเดือนธันวาคม ค.ศ. 1991 จำนวน 49 ข้อ ในการตรวจสอบให้เพศชายเป็นกลุ่มอ้างอิงและเพศหญิงเป็นกลุ่มเปรียบเทียบ ผู้เชี่ยวชาญจัดหมวดข้อสอบเป็น 8 หมวด และพิจารณาว่าหมวดใดน่าจะให้ประโยชน์แก่เพศชายหรือเพศหญิง หลังจากนั้นวิเคราะห์ด้วยโปรแกรมคอมพิวเตอร์สำเร็จรูป SIBTEST ใช้กลุ่มตัวอย่างเพศชาย 3,000 คน เพศหญิง 3,000 คน พบว่าความคิดเห็นของผู้เชี่ยวชาญและผลการวิเคราะห์ด้วยวิธีซิปเทสท์สอดคล้องกัน จึงทำการวิเคราะห์ปริมาณ DIF ของแต่ละข้อ พบว่าแต่ละข้อแสดง DIF ด้วยปริมาณที่น้อยมาก จึงไม่มีการคัดเลือกข้อสอบออกจากแบบทดสอบ ตัวอย่างที่ 2 ใช้วิธีการทางสถิติในการตรวจสอบหมวดข้อสอบ คือวิธี HCA (Agglomerative Hireachical Cluster Analysis; Jain & Kubes, 1988) และ DIMTEST (Nandakumar & Stout, 1993, Stout, 1997) แล้วจึงตรวจสอบด้วยวิธีซิปเทสท์ ในการตรวจสอบใช้ข้อสอบ Nation Assessment of Education Progress (NAEP) จำนวน 36 ข้อ ใช้กลุ่มตัวอย่างกลุ่มละ 50 คน ผลการตรวจสอบปรากฏว่า การตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบประสบความสำเร็จเป็นอย่างดี ตัวอย่างที่ 3 ใช้แบบทดสอบความเข้าใจการอ่าน ที่สอบในเดือนธันวาคม ค.ศ. 1991 ซึ่งเป็นส่วนหนึ่งของการทดสอบ LSAT มีข้อสอบ 28 ข้อ แบ่งตามบทความที่กำหนดให้อ่านเป็น 4 บทความ แต่ละบทความมีข้อสอบ 5-8 ข้อ แล้วทำการตรวจสอบหมวดข้อสอบซ้ำโดยใช้ HCA และ DIMTEST หลังจากนั้นจึงวิเคราะห์ด้วย SIBTEST ใช้กลุ่มตัวอย่างกลุ่มละ 1,000 คน พบว่าทั้ง 4 บทความแสดงการทำหน้าที่ต่างกันของหมวดข้อสอบ ข้อสอบ 2 หมวดให้ประโยชน์แก่เพศชาย และอีก 2 หมวดให้ประโยชน์แก่เพศหญิง

ชาง และคณะ (Chang et al., 1996, pp. 333-353) ได้ศึกษาผลของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนหลายค่า โดยประยุกต์วิธีซิปเทสท์มาใช้เปรียบเทียบกับวิธีแมนเทล-แฮนส์เซล และวิธี SMD แบ่งการศึกษาออกเป็น 2 ตอน ตอนที่ 1 ใช้ข้อมูลจำลองภายใต้เงื่อนไขเดียวกับงานวิจัยของ ซวิค และคณะ (Zwick et al., 1993) เพื่อเปรียบเทียบวิธีซิปเทสท์แบบประยุกต์วิธีแมนเทล-แฮนส์เซล และวิธี SMD ผลการศึกษาปรากฏว่า วิธีซิปเทสท์มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบดี แต่วิธีแมนเทล-แฮนส์เซล และวิธี SMD มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบค่อนข้างดีกว่า ตอนที่ 2 ใช้ข้อมูลจำลอง คือ ข้อสอบที่ศึกษามีอำนาจจำแนกแตกต่างกัน 11 ค่า ตั้งแต่ 1.5 ถึง 2.00 ขนาดกลุ่มตัวอย่างต่างกัน คือ 500 1,000 และ 2,000 คน ข้อสอบมีความยาว 24 ข้อ สำหรับ วิธีซิปเทสท์ และ 25 ข้อสำหรับวิธีแมนเทล-แฮนส์เซล และวิธี SMD ผลการศึกษาปรากฏว่า ทั้ง วิธีแมนเทล-แฮนส์เซล และวิธี SMD มีค่าความคลาดเคลื่อนประเภทที่ 1 ค่อนข้างสูง เมื่อค่าอำนาจจำแนกของข้อสอบที่ศึกษามี

ค่าต่างกันจากค่าเฉลี่ยของอำนาจจำแนกของแบบทดสอบที่มีความตรงและอัตราการปฏิเสธ (Refection Rates) ของทั้ง 3 วิธีจะมีค่าสูงขึ้น เมื่อมีค่าอำนาจจำแนกสูงขึ้น

ซวิค และคณะ (Zwick et al., 1997, pp. 321-344) ได้ศึกษาการทำหน้าที่ต่างกันของแบบทดสอบที่มีการตรวจให้คะแนนแบบหลายค่า โดยใช้วิธีวิเคราะห์ 5 วิธีด้วยกัน คือ SMD จำนวน 2 วิธี ได้แก่ SMD-H และ SMD-M วิธีแมนเทล-แฮนด์เซล และวิธีซิปเทสท์ 2 วิธี ได้แก่ Standard SIBTESTS และวิธี Modified SIBTEST โดยใช้เงื่อนไขแบบทดสอบที่มีการตรวจสอบให้คะแนนแบบสองค่า จำนวน 50 ข้อ และแบบทดสอบที่มีการตรวจให้คะแนนแบบหลายค่า จำนวน 18 ข้อ ใช้กับกลุ่มตัวอย่าง 1,000 คน แบ่งเป็นกลุ่มอ้างอิง 500 คน และกลุ่มเปรียบเทียบ 500 คน โดยมีการจับคู่ระหว่างข้อสอบแบบให้คะแนนสองค่า กับแบบให้คะแนนหลายค่า โดยใช้โมเดลโลจิสติก 3 พารามิเตอร์ จากข้อสอบจำนวน 75 ข้อ ที่มีค่าอำนาจจำแนกอยู่ระหว่าง 0.74 ถึง 1.0 ค่าความยากอยู่ระหว่าง -1.95 ถึง 1.95 ค่าการเดาอยู่ที่ 0.15 และสถานการณ์ที่จำลองขึ้นมากับกลุ่มตัวอย่างทั้ง 2 กลุ่ม ที่มีการแจกแจงดัชนีการปฏิบัติที่ดี เมื่อความแตกต่างของค่าเฉลี่ยของกลุ่มที่มีความคลาดเคลื่อนมาตรฐานเป็น 1 วิธี Modified SIBTEST วัดผลกระทบของขนาดกลุ่มตัวอย่างค่อนข้างดี จาก 5 วิธี ในทางปฏิบัติไม่สามารถมองเห็นความแตกต่างของกลุ่มย่อยทั้ง 2 กลุ่ม จะมีการแจกแจงเบ้ไปด้านใดด้านหนึ่ง เมื่อกลุ่มมีการแจกแจงที่แตกต่างกัน และข้อสอบที่ศึกษามีค่าอำนาจจำแนกสูง วิธี SIBTEST ดีที่สุด รองลงมาวิธี SMD และวิธีแมนเทล-แฮนส์เซล เมื่อพิจารณาจากความคลาดเคลื่อนประเภทที่ 1 ถ้าแบบทดสอบตอบสั้น ๆ เมื่อดูค่าอำนาจจำแนก ทั้ง 5 วิธี แตกต่างกัน เนื่องจากเป็นความแตกต่างในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ และองค์ประกอบอื่น ๆ วิธีแมนเทล-แฮนส์เซล และวิธี SMD เป็นวิธีที่ใช้ตรวจสอบการทำหน้าที่ต่างกันของแบบทดสอบแบบการให้คะแนนหลายค่า (Polytomous) เมื่อมีเกณฑ์ การจับคู่กันระหว่างการตอบแบบฝึกหัดกับอัตราความคลาดเคลื่อนประเภทที่ 1 และวิธี SIBTEST ไม่สามารถใช้กับการจับคู่ได้ เนื่องจากใช้กับแบบทดสอบแบบถูกผิดไม่ได้ และปัจจุบันไม่นิยมใช้กัน

ไรอัน และชิว (Ryan & Chiu, 2001, pp. 73-90) ได้ศึกษาการทำหน้าที่ต่างกันของหมวดข้อสอบ เมื่อมีการเปลี่ยนแปลงตำแหน่งของข้อสอบ โดยใช้แบบทดสอบวิชาคณิตศาสตร์ระดับชั้นอุดมศึกษาปีที่ 1 จำนวน 40 ข้อ จัดทำเป็น 2 รูปแบบ รูปแบบที่ 1 เรียงลำดับข้อสอบตามเนื้อหาจากง่ายไปยาก รูปแบบที่ 2 เรียงลำดับข้อสอบแบบสุ่มกลุ่มตัวอย่างที่ตอบข้อสอบรูปแบบ 1 เป็นเพศชาย จำนวน 546 คน และเพศหญิง จำนวน 520 คน กลุ่มตัวอย่างที่ตอบข้อสอบ รูปแบบ 2 เป็นเพศชาย จำนวน 554 คน และเพศหญิง จำนวน 511 คน วิเคราะห์ข้อสอบโดยใช้โปรแกรม SIBTEST ผลการศึกษาปรากฏว่า การเปลี่ยนแปลงตำแหน่งข้อสอบในแบบทดสอบไม่มีผลต่อการตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบ

2. วิธีการวิเคราะห์องค์ประกอบจำกัด (Restricted Factor Analysis: RFA)

หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยวิธีการวิเคราะห์องค์ประกอบจำกัด ใช้โมเดลองค์ประกอบร่วมเชิงเส้น โดยสมมติว่าแบบทดสอบชุดหนึ่งสร้างขึ้นเพื่อวัดคุณลักษณะแฝงคุณลักษณะหนึ่ง ซึ่งมีข้อสอบทั้งหมด p ข้อ และการทำหน้าที่ต่างกันของข้อสอบจะศึกษาได้จากตัวฝ่าฝืน (Potential Violator) r ตัว สำหรับการเลือกข้อสอบแบบสุ่มคะแนนของข้อสอบข้อที่ i แทนได้ด้วยค่า X_i ซึ่งค่า X_i แสดงได้ดังสมการที่ 1

$$X_i = m_i + a_i T + \sum_{k=1}^{r} (b_{ik} V_k) + D_i \qquad (1)$$

เมื่อ T เป็นคะแนนที่แท้จริง (True Score) ของคุณลักษณะที่ศึกษา

V_k เป็นคะแนนจริง (True Score) ของคุณลักษณะตัวฝ่าฝืนตัวที่ k

D_i เป็นองค์ประกอบของคะแนนความคลาดเคลื่อนของการทำข้อสอบข้อที่ i

m_i เป็นค่าจุดตัดแกน

a_i เป็นสัมประสิทธิ์การถดถอยของข้อสอบข้อที่ i ของคุณลักษณะ T_i

b_{ik} เป็นสัมประสิทธิ์การถดถอยของข้อสอบข้อที่ i ของคุณลักษณะ V_k

จากสมการที่ 1 ข้อสอบข้อที่ i จะทำหน้าที่ต่างกันตามตัวฝ่าฝืนที่ k ถ้ามีอิทธิพลทางตรงจากตัวฝ่าฝืนตัวที่ k มายังข้อสอบข้อที่ i แล้วส่งผลให้ค่าน้ำหนักองค์ประกอบไม่เท่ากับศูนย์ ($b_{ik} \neq 0$) ลักษณะเช่นนี้เรียกว่า การทำหน้าที่ต่างกันจากตัวฝ่าฝืนตัวที่ k บนข้อสอบข้อที่ i เพราะฉะนั้นการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการวิเคราะห์องค์ประกอบจำกัดจะตั้งสมมติฐานหลัก (H_0) ว่า $b_{ik} = 0$ แล้วทำการตรวจสอบสมมติฐานหลักสำหรับข้อสอบทุกข้อและสำหรับตัวฝ่าฝืนทุกตัว

จากโมเดลการวิเคราะห์องค์ประกอบ ความสัมพันธ์ระหว่างข้อสอบ คุณลักษณะ T กับตัวฝ่าฝืน r ตัวที่ถูกวัดด้วยข้อสอบหรือตัวบ่งชี้ q ตัว ($q \geq r$) คะแนน Y_j จากข้อสอบหรือตัวบ่งชี้ตัวที่ j สามารถคำนวณได้จากสมการที่ 2

$$Y_j = n_j + \sum_{k=1}^{r} (C_{jk} V_k) + E_j \qquad (2)$$

เมื่อ Y_j เป็นคะแนนจากข้อสอบหรือตัวบ่งชี้ตัวที่ j

n_j เป็นค่าเฉลี่ยของข้อสอบหรือตัวบ่งชี้ที่ j

C_{jk} เป็นสัมประสิทธิ์การถดถอยของข้อสอบหรือตัวบ่งชี้ที่ j บนตัวฝ่าฝืนตัวที่ k

E_j เป็นองค์ประกอบของความคลาดเคลื่อน

โดยกำหนดให้ i, j, k, D_i และ E_j ทุกค่าเป็นอิสระจากกันและเป็นอิสระจาก T และ V_k

จากสมการที่ 1 และสมการที่ 2 สามารถนำมาใช้ในโมเดลการวิเคราะห์องค์ประกอบความแปรปรวนร่วมในสมการที่ 3

$$S = LFL' + U^2 \qquad (3)$$

เมื่อ S เป็นเมทริกซ์ความแปรปรวน-ความแปรปรวนร่วมของคะแนนสังเกตได้ X_i

Y_i, L เป็นเมทริกซ์ของน้ำหนักองค์ประกอบของตัวแปรสังเกตได้จากตัวแปรแฝงหรือองค์ประกอบร่วม T

V_k, F เป็นเมทริกซ์ความแปรปรวน-ความแปรปรวนร่วมของตัวแปรแฝง

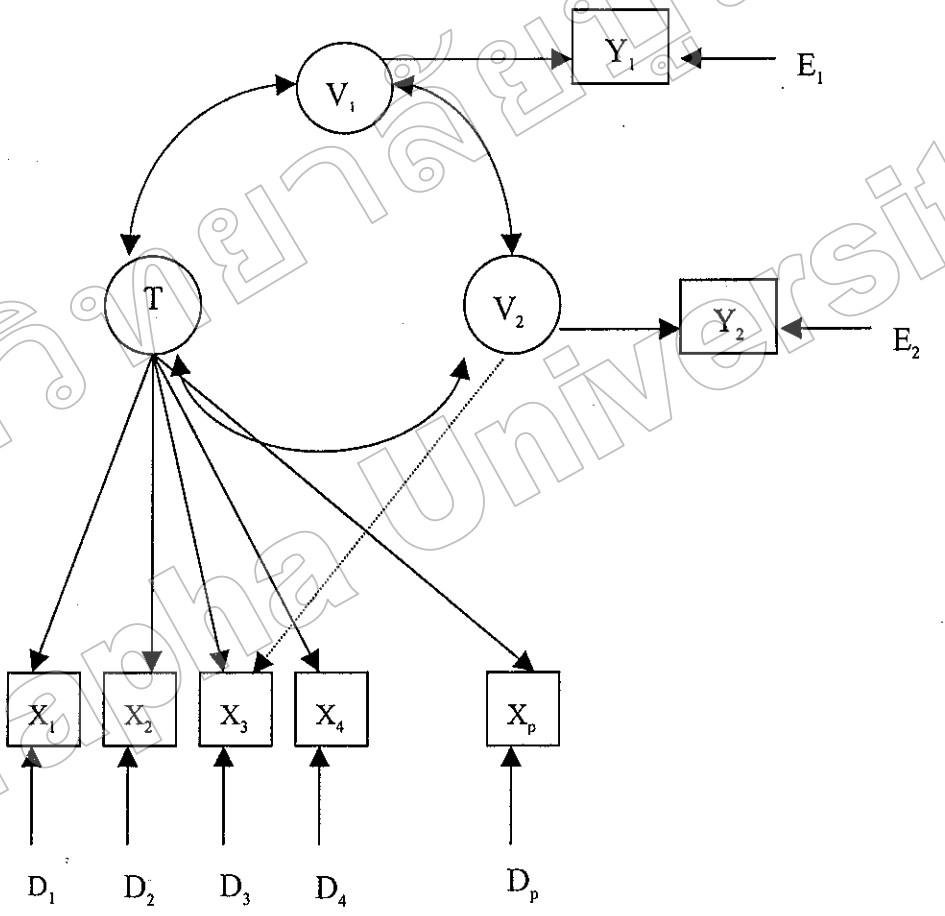
U^2 เป็นเมทริกซ์แนวทแยง (Diagonal Matrix) ของค่าความแปรปรวนขององค์ประกอบที่คลาดเคลื่อน

การวิเคราะห์องค์ประกอบ ตัวแปรทุกตัวจะวัดได้จากค่าเฉลี่ย จุดตัดแกนในสมการที่ 1 และสมการที่ 3

ในกรณีที่มีข้อสอบหรือตัวบ่งชี้ 1 ตัวของแต่ละตัวฝ่าฝืน ($q = r$) ตัวฝ่าฝืน V_k ถูกแทนที่ด้วยตัวแปรสังเกตได้ Y_j ซึ่งหมายความว่าในโมเดลองค์ประกอบนี้ตัวฝ่าฝืนสันนิษฐานว่าสามารถวัดได้โดยปราศจากความคลาดเคลื่อนในการวัด และความแปรปรวนที่อธิบายไม่ได้ ($E_i = 0$) ของตัวฝ่าฝืนใน U^2 จะถูกกำหนดให้เป็น 0 เพื่อความสะดวกมากขึ้น ตัวแปรสังเกตได้เหล่านี้อาจจะเป็นคะแนนมาตรฐานก็ได้ และถ้าใช้เมทริกซ์สหสัมพันธ์แทนเมทริกซ์ความแปรปรวนร่วมจะทำให้มีประสิทธิภาพมากขึ้น เพราะค่าสัมประสิทธิ์คือ คะแนนมาตรฐานของค่าความแปรปรวนร่วมนั่นเองและตัวแปรแฝงอาจจะทำให้เป็นคะแนนมาตรฐานได้โดยการกำหนดสมาชิกในแนวทะแยงของ F ให้เหมือนกัน นอกจากนั้นน้ำหนักองค์ประกอบของตัวแปร Y_j จาก V_k อาจจะกำหนดให้เหมือนๆ กัน ($C_{ik} = 0$ ของทุกๆ j และทุกๆ k เมื่อ $j = k$) และกำหนดให้ตัวอื่นๆ เป็น 0 ซึ่งตัวที่จะถูกประมาณค่าคือ น้ำหนักองค์ประกอบของข้อสอบจากคุณลักษณะแฝง T ความแปรปรวนของความคลาดเคลื่อนของข้อสอบ และสหสัมพันธ์ระหว่างตัวฝ่าฝืนและคุณลักษณะแฝง T

การทำหน้าที่ต่างกันของข้อสอบสามารถตรวจสอบได้จากค่าสัมประสิทธิ์การถดถอยของข้อสอบจากตัวฝ่าฝืนต่าง ๆ โมเดลการวิเคราะห์องค์ประกอบจำกัด ค่าสัมประสิทธิ์การถดถอยจะถูกนำเสนอในรูปน้ำหนักองค์ประกอบของตัวแปรสังเกตได้ X_i จากตัวฝ่าฝืน V_k โมเดลองค์ประกอบ

ที่มีค่าน้ำหนักองค์ประกอบเป็น 0 ถูกกำหนดให้เป็นโมเดลหลัก (H_0) ซึ่งเป็นโมเดลที่ไม่มีความลำเอียง คือ ค่า $b_{ik} \neq 0$ ของทุก ๆ i และทุก ๆ k จากภาพที่ 3 โมเดลการทดสอบความลำเอียงของข้อสอบข้อที่ 3 ตามตัวฝ่าฝืนตัวที่ 2 โมเดลองค์ประกอบอื่นก็ต้องถูกตรวจสอบด้วย ซึ่งโมเดลอื่น ๆ นี้จะมีลักษณะเหมือนกับโมเดลหลัก (H_0) ยกเว้นน้ำหนักองค์ประกอบของข้อที่ 3 จากตัวฝ่าฝืนตัวที่ 2 (b_{32}) ซึ่งจะถูกกำหนดให้เป็นพารามิเตอร์อิสระ ถ้าโมเดลนี้มีค่านัยสำคัญมากกว่า โมเดลหลัก (H_0) ข้อสอบข้อที่ 3 ก็จะเป็นข้อสอบข้อที่มีความลำเอียงตามตัวฝ่าฝืนตัวที่ 2 และจะต้องถูกขจัดออกจากแบบทดสอบ



ภาพที่ 3 โมเดลองค์ประกอบแสดงความสัมพันธ์ระหว่างข้อสอบแต่ละข้อ คุณลักษณะ T กับตัวฝ่าฝืน (Oort, 1996, p. 46)

ภาพที่ 3 แสดงรูปแบบที่อธิบายได้จากสมการที่ 2 และ 3 เมื่อมีตัวฝ่าฝืน 2 ตัว (V_1, V_2) ซึ่งแต่ละตัววัดจากตัวบ่งชี้เพียงตัวเดียวและคุณลักษณะ T มีความสัมพันธ์กับคะแนน Y_1, Y_2 และ V_1,V_2 ซึ่งคุณลักษณะ T วัดได้จากคะแนนข้อสอบ X_i นอกจากนี้ คุณลักษณะ T, V_k และ X_i อาจมี

ความสัมพันธ์กันด้วย จากภาพที่ 3 เส้นประจากตัวฝ่าฝืนตัวที่ 2 ไปยังข้อสอบข้อที่ 3 (X_3) แสดงถึงอิทธิพลทางตรงจากตัวฝ่าฝืนตัวที่ 2 ไปยังข้อสอบข้อที่ 3 ซึ่งอาจเป็นอิทธิพลทางอ้อมผ่านตัวแปร T ไปยัง X_3 ก็ได้ ค่าอิทธิพลทางตรงนี้บอกเป็นนัยว่าคะแนนที่วัดจาก T ที่มีค่าสูง ๆ ไม่จำเป็นต้องสัมพันธ์กับคุณลักษณะแฝง T ซึ่งมีความเป็นไปได้ที่ผู้เข้าสอบที่มีคะแนน T เท่ากัน อาจจะได้คะแนนในข้อสอบข้อที่ 3 (X_3) ไม่เท่ากัน ทั้งนี้เพราะคะแนนดังกล่าวได้รับอิทธิพลจากตัวฝ่าฝืนตัวที่ 2 (V_2) แตกต่างกัน ดังนั้น ข้อสอบข้อที่ 3 (X_3) จึงไม่จำเป็นที่จะต้องวัดเพียงคุณลักษณะแฝง T อย่างเดียว แต่วัดตัวฝ่าฝืนตัวที่ 2 (V_2) ด้วย กล่าวได้ว่าข้อสอบข้อที่ 3 (X_3) จะมีความลำเอียงตามตัวฝ่าฝืนตัวที่ 2 (V_2)

โมเดลหลัก (H_0) สามารถตรวจสอบด้วยโปรแกรม LISREL โดยพิจารณาจากค่าดัชนีดัดแปรโมเดล (Modification Indices: MI) ซึ่งปกติแล้วการกระจายของดัชนี MI จะมีการแจกแจงเป็นไค-สแควร์ที่มีองศาอิสระเท่ากับ 1 ถ้าดัชนี MI ของข้อสอบข้อใดมีค่ามากและมีนัยสำคัญแสดงว่าข้อสอบข้อนั้นทำหน้าที่ต่างกันต้องตัดออกจากแบบทดสอบ นอกจากนี้ยังสามารถตรวจสอบขนาดและทิศทางของการทำหน้าที่ต่างกันของข้อสอบได้ โดยการพิจารณาค่าพารามิเตอร์ที่คาดหวัง (Expected Parameter Change: EPC) ซึ่งค่า EPC นี้เป็นการประมาณค่าการเปลี่ยนแปลงที่คาดหวังของค่าพารามิเตอร์กำหนด เมื่อโมเดลสมมติฐาน (H_0) กำหนดให้เป็นค่าพารามิเตอร์อิสระ ถ้าข้อสอบข้อที่ i เป็นข้อสอบที่ทำหน้าที่ต่างกันและค่า EPC เป็นค่าสัมประสิทธิ์การถดถอยของข้อสอบข้อที่ i จากตัวฝ่าฝืน k เป็นบวกแสดงว่าข้อสอบข้อที่ i ทำหน้าที่ต่างกันต่อกลุ่มผู้ตอบที่มีคะแนนของตัวฝ่าฝืน k มาก แคบแลน (Kaplan, 1990 cited in Oort, 1998) เสนอแนะว่า การที่จะขจัดข้อสอบที่ทำหน้าที่ต่างกันออกจากแบบทดสอบ สามารถพิจารณาโดยใช้ดัชนี MI และค่า EPC ที่แตกต่างจาก 0 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 โดยไม่ต้องพิจารณาขนาดและทิศทางของค่า EPC ก็ได้

การใช้วิธี RFA ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมีข้อดีอยู่หลายประการดังนี้ (Oort, 1996)

1. สามารถตรวจสอบการทำหน้าที่ต่างกันของแบบทดสอบที่มีรูปแบบการตอบที่หลากหลาย เช่น ถ้าระดับการตอบข้อสอบเป็นแบบต่อเนื่อง จะสามารถวิเคราะห์โดยใช้สหสัมพันธ์แบบเพียร์สันและความแปรปรวนร่วมได้ ถ้าระดับการตอบข้อสอบแบ่งเป็นสองหรือแบ่งแบบพหุ จะสามารถวิเคราะห์โดยใช้สหสัมพันธ์เตตระคอริกหรือสหสัมพันธ์โพลีคอริกได้

2. ตัวฝ่าฝืนเป็นตัวแปรที่แสดงความเป็นสมาชิกของกลุ่ม คือ เป็นตัวแปรนามบัญญัติ แต่ในวิธี RFA สามารถตรวจสอบตัวฝ่าฝืนที่มีลักษณะใดก็ได้ โดยไม่คำนึงถึงระดับของการวัด

3. วิธี RFA สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีตัวฝ่าฝืนหลายตัวได้ แม้ว่าจะเป็นคนละชนิดกันก็ตาม

4. วิธี RFA ไม่จำเป็นต้องแบ่งกลุ่มตัวอย่างออกเป็นกลุ่มย่อย ๆ ก็ได้

5. หลีกเลี่ยงปัญหาการประมาณค่าพารามิเตอร์ของกลุ่มประชากรที่แตกต่างกันให้อยู่ในระดับเดียวกันได้

6. กลุ่มตัวอย่างที่ใช้ในการวิเคราะห์ด้วยวิธี RFA ไม่จำเป็นต้องมีขนาดใหญ่เหมือนกับวิธี IRT

งานวิจัยที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการวิเคราะห์องค์ประกอบจำกัด

นิคม กีรติวรางกูร (2542) ได้เปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธีการวิเคราะห์องค์ประกอบจำกัด วิธีแมนเทล-แฮนส์เซล กับวิธีการตอบสนองข้อสอบแบบ 2 พารามิเตอร์ เมื่อขนาดกลุ่มตัวอย่างแบ่งออกเป็น 2 ขนาด คือ กลุ่มตัวอย่างขนาดเล็ก (300 คน) และขนาดใหญ่ (1,000 คน) ความยาวของแบบทดสอบแบ่งออกเป็น 2 ขนาด คือ แบบทดสอบสั้น (25 ข้อ) และแบบทดสอบยาว (75 ข้อ) ค่าความยากของข้อสอบแบ่งออกเป็น 3 ระดับ คือ กลุ่มข้อสอบที่มีความยากมาก ปานกลาง และน้อย ค่าอำนาจจำแนกของข้อสอบแบ่งออกเป็น 3 ระดับ คือ กลุ่มข้อสอบที่มีค่าอำนาจจำแนกสูง ปานกลาง และต่ำ ซึ่งข้อสอบทำหน้าที่ต่างกัน แบ่งออกเป็น 2 กลุ่ม คือ กลุ่มข้อสอบที่มีข้อสอบทำหน้าที่ต่างกันจำนวนมาก และจำนวนน้อย ผลการศึกษาปรากฏว่า โดยภาพรวมวิธี RFA มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบดีที่สุด รองลงมา คือ วิธีแมนเทล-แฮนส์เซล และ วิธี IRT แบบ 2 พารามิเตอร์ และ วิธี IRT มีอัตราความคลาดเคลื่อนประเภทที่ 1 สูงกว่าวิธี แมนเทล-แฮนส์เซล และวิธีการวิเคราะห์องค์ประกอบจำกัด

ออร์ท (Oort, 1998, pp. 107-124) ได้ใช้วิธีการวิเคราะห์องค์ประกอบจำกัด (RFA) ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยการจำลองข้อมูล ใช้คะแนนรวมจากการสอบเป็นตัวแทนความสามารถ และเปรียบเทียบประสิทธิภาพในการตรวจสอบด้วยวิธี RFA กับวิธี IRT แบบ 1 พารามิเตอร์ ปรากฏว่าในข้อสอบที่มีระดับการตอบแบบแบ่งเป็นสอง เมื่อกำหนดขนาดของการทำหน้าที่ต่างกันของข้อสอบออกเป็น 3 ระดับ คือ ข้อสอบทำหน้าที่ต่างกันมาก (0.8 SD) ปานกลาง (0.5 SD) และน้อย (0.2 SD) ขนาดกลุ่มตัวอย่างแบ่งเป็น 2 ขนาด คือ ขนาดใหญ่ (2,000 คน) และขนาดเล็ก (200 คน) และค่าเฉลี่ยของความสามารถของทั้งสองกลุ่มไม่เท่ากัน ผลการศึกษาปรากฏว่า เมื่อข้อสอบทำหน้าที่ต่างกันปานกลาง และข้อสอบที่ทำหน้าที่ต่างกันมาก วิธี RFA ให้ผลการตรวจสอบได้ดีกว่าวิธี IRT และในกลุ่มตัวอย่างขนาดเล็กการตรวจสอบด้วยวิธี RFA ให้ผลการตรวจสอบที่ดีกว่า และทั้งสองวิธีจะให้ผลการตรวจสอบสมบูรณ์ที่สุดเมื่อกลุ่มตัวอย่างมีขนาดใหญ่

3. วิธีถดถอยโลจิสติก (Logistic Regression: LR)

สวามินาทาน และโรเจอร์ (Swaminathan & Rogers, 1990) ได้พัฒนาวิธีถดถอยโลจิสติกเพื่อใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนแบบสองค่า (Dichotomous) วิธีนี้มีแนวคิดมาจากวิธีตารางการณ์จร โดยดัดแปลงมาจากวิธีลอกลิเนียร์ (Loglinear) ของเมลเลนเบอร์ก (Mellenberg, 1982) และวิธีแมนเทล-แฮนส์เซลของ ฮอลแลนด์ และเทเยอร์ (Holland & Thayer, 1988) หลักการตรวจสอบด้วยวิธีถดถอยโลจิสติกจะใช้โมเดลการถดถอยโลจิสติกทำนายโอกาสของผลการตอบข้อสอบถูก โมเดลดังกล่าวใช้ตัวแปรความสามารถแบบต่อเนื่องซึ่งมีเทอมที่ใช้คำนวณปฏิสัมพันธ์ระหว่างการเป็นสมาชิกของกลุ่มผู้สอบกับระดับความสามารถ จึงทำให้สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ทั้งแบบเอกรูป (Uniform DIF) และแบบอเนกรูป (Nonuniform DIF) นอกจากนี้ยังสามารถนำไปประยุกต์กับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีผู้สอบหลายกลุ่ม และการให้คะแนนข้อสอบแบบพหุวิภาค (Polytomous) (Miller & Spray, 1993)

การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธีถดถอยโลจิสติกจะใช้สมการมาตรฐานของโมเดลการถดถอยโลจิลติกคำนวณผลการตอบข้อสอบถูก ดังนี้ (Swaminathan & Rogers, 1990)

$$P(U_{ij}=1/\theta_{ij})=\frac{\exp^{(\beta_{0j}+\beta_{1j}\theta_{1j})}}{1+\exp^{(\beta_{0j}+\beta_{1j}\theta_{1j})}} \qquad , i=1,2,\ldots,n : j=1,2$$

เมื่อ U_{ij} แทน ผลการตอบข้อสอบของผู้เข้าสอบคนที่ 1 ในกลุ่ม j

θ_{ij} แทน ค่าความสามารถที่สังเกตได้ของผู้เข้าสอบคนที่ i ในกลุ่ม j

β_{0j} แทน ค่าพารามิเตอร์จุดตัด (Intercept Parameter)

β_{ij} แทน ค่าพารามิเตอร์ความชันสำหรับกลุ่ม j (Slope Parameter)

จากโมเดลดังกล่าว ถ้า $\beta_{01}=\beta_{02}$ และ $\beta_{11}=\beta_{12}$ แล้ว ฟังก์ชันการถดถอยโลจิสติกของผู้เข้าสอบสองกลุ่มเหมือนกัน แสดงว่าข้อสอบทำหน้าที่ไม่ต่างกัน (No DIF) ถ้า $\beta_{11}=\beta_{12}$ แต่ $\beta_{01}\neq\beta_{02}$ แล้ว ฟังก์ชันการถดถอยโลจิสติกของผู้เข้าสอบสองกลุ่มขนานกันแต่ไม่ทับกัน แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบเอกรูป (Uniform DIF) และถ้า $\beta_{01}=\beta_{02}$ แต่ $\beta_{11}\neq\beta_{12}$ แล้ว ฟังก์ชันการถดถอยโลจิสติกของผู้เข้าสอบไม่ขนานกัน แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป

(Nonuniform DIF) นอกจากนี้โมเดลการถดถอยโลจิสติกดังกล่าวสามารถเปลี่ยนเป็นโมเดลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปและแบบอเนกรูป ดังนี้

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

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

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

θ_{ij} แทน ความสามารถของผู้เข้าสอบคนที่ i ในกลุ่ม j

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

$\theta_{ij}G_j$ 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

$$\log\left[\frac{P}{1-P}\right]=Z_{ij}=\tau_0+\tau_1\theta_{ij}+\tau_2G_j+\tau_3(\theta_{ij}G_j)$$

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

ในการตัดสินว่าข้อสอบทำหน้าที่ต่างกันแบบเอกรูปหรืออเนกรูป จะพิจารณาพารามิเตอร์ 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

สำหรับการประมาณค่าพารามิเตอร์ตามโมเดลโลจิสติก ของข้อสอบแต่ละข้อของโมเดล Z_{ij} 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 τ และเมทริกซ์ความแปรปรวน – ความแปรปรวนร่วมในรูป Σ ในขณะที่ Σ^{-1} เป็นเมทริกซ์สารสนเทศกำหนด ดังนี้

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

เมื่อ E และ lnL แทนค่าความคาดหมายของเมทริกซ์และลอการิทึมของฟังก์ชันความควรตามลำดับ ดังนั้นการแจกแจงของการประมาณค่าพารามิเตอร์ด้วยวิธี MLE จะอยู่ในรูปดังนี้

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

โดยที่ $\tau = [\tau_0, \tau_1, \tau_2, \tau_3]$ ส่วนความคลาดเคลื่อนมาตรฐานเชิงเส้นกำกับของค่าประมาณของ τ_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 2 x 4 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

งานวิจัยที่เกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีถดถอยโลจิสติก

รัชรินทร์ มุคดา (2540) ได้เปรียบเทียบประสิทธิภาพระหว่างวิธีแมนเทล-แฮนส์เซลกับวิธีถดถอยโลจิสติก ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป โดยศึกษาจากข้อมูลจำลองขึ้นด้วยโปรแกรม IRTDATA เงื่อนไขที่ศึกษา ได้แก่ (1) กลุ่มความสามารถผู้เข้าสอบ 3 ระดับ คือ กลุ่มผู้เข้าสอบที่มีความสามารถสูง ปานกลาง และต่ำ (2) ค่าความยากของข้อสอบ 3 ระดับ คือ กลุ่มข้อสอบที่มีค่าความยากสูง ปานกลาง และต่ำ (3) ค่าอำนาจจำแนกของข้อสอบ 3 ระดับ คือ กลุ่มข้อสอบที่มีค่าอำนาจจำแนกสูง ปานกลาง และต่ำ รวมเงื่อนไขที่ศึกษาทั้งหมด 27 เงื่อนไข ผลการศึกษาปรากฏว่า ในกลุ่มผู้เข้าสอบที่มีความสามารถสูง ปานกลางและต่ำ วิธีแมนเทล-แฮนส์เซลและวิธีถดถอยโลจิสติกมีประสิทธิภาพเท่าเทียมกันในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป สำหรับปัจจัยของลักษณะของข้อสอบที่เกี่ยวกับค่าความยากของข้อสอบ พบว่า ข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปซึ่งตรวจพบมากที่สุดในกลุ่มผู้เข้าสอบที่มีความสามารถ สูง ปานกลาง และต่ำ เป็นข้อสอบที่มีค่าความยากสูง ปานกลาง และต่ำตามลำดับ ส่วนลักษณะของข้อสอบที่เกี่ยวกับค่าอำนาจจำแนก พบว่า ข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูป ซึ่งตรวจพบมากที่สุดในทุกกลุ่มผู้เข้าสอบเป็นข้อสอบที่มีค่าอำนาจจำแนกสูง

นพมาศ พิพัฒนสุข (2541) ได้เปรียบเทียบประสิทธิภาพระหว่างวิธีแมนเทล-แฮนส์เซลกับวิธีถดถอยโลจิสติกในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อใช้เกณฑ์จับคู่เปรียบเทียบแตกต่างกันในแบบทดสอบชนิดพหุมิติ ผลการศึกษาปรากฏว่า วิธีแมนเทล-แฮนส์เซล มีประสิทธิภาพมากกว่าวิธีถดถอยโลจิสติก ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติเมื่อใช้เกณฑ์จับคู่คะแนนรวม และมีประสิทธิภาพไม่แตกต่างกัน เมื่อใช้เกณฑ์จับคู่คะแนนแบบทดสอบย่อย ส่วนวิธีถดถอยโลจิสติกเมื่อใช้เกณฑ์จับคู่เปรียบเทียบคะแนนจาก

แบบทดสอบย่อยหลายฉบับมีความเหมาะสมในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบชนิดพหุมิติ

อารี วัชรโสตถิกุล (2543) ได้เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้รูปแบบต่างกันคือ รูปแบบคะแนนรวมทั้งฉบับ แยกตามเนื้อหา และแยกตามระดับพฤติกรรม ด้วยวิธีการตรวจสอบต่างกัน คือ วิธี ซิปเทสท์ และวิธีถดถอยโลจิสติก แล้วทำการตัดข้อสอบที่ทำหน้าที่ต่างกันออกจากแบบทดสอบ เพื่อเปรียบเทียบค่าความเชื่อมั่น ผลการศึกษาปรากฏว่า จำนวนข้อสอบที่ทำหน้าที่ต่างกันโดยใช้วิธีการตรวจสอบต่างกันแตกต่างกัน ในรูปแบบรวมทั้งฉบับ ส่วนรูปแบบแยกตามเนื้อหา และแยกตามระดับพฤติกรรมไม่แตกต่างกัน

ทองอยู่ สาระ (2543) ได้เปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาดในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปและแบบอเนกรูป ระหว่างวิธีถดถอยโลจิสติกกับวิธีแมนเทล-แฮนส์เซล โดยใช้ความยาวของแบบทดสอบและขนาดกลุ่มตัวอย่างแตกต่างกัน กลุ่มตัวอย่างที่ใช้ในการศึกษาครั้งนี้เป็นนักเรียนชั้นมัธยมศึกษาปีที่ 3 แบบทดสอบที่ใช้วัดความสามารถทางสมองที่สร้างตามแนวโครงสร้างของโอติส-เลนนอน ขนาดแบบทดสอบ 3 ขนาด คือ 20 40 และ 60 ข้อ สุ่มกลุ่มตัวอย่างขนาดกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบเป็นดังนี้ คือ 300: 100, 300: 200, 300: 300, 500: 100, 500: 200, 500: 300, 1,000: 100, 1,000: 200, 1,000: 300 และ 1,000: 1,000 เงื่อนไขที่ศึกษา 30 เงื่อนไข ผลการศึกษาปรากฏว่า วิธีถดถอยโลจิสติกและวิธีแมนเทล-แฮนส์เซล มีอำนาจการตรวจสอบ และการจำแนกผิดพลาดในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปและแบบอเนกรูป ไม่แตกต่างกัน ความยาวของแบบทดสอบไม่มีผลต่ออำนาจการตรวจสอบและการจำแนกผิดพลาดในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปและแบบอเนกรูปด้วยวิธีถดถอยโลจิสติกและวิธีแมนเทล-แฮนส์เซล ขนาดกลุ่มตัวอย่างมีผลต่ออำนาจการตรวจสอบ แต่ขนาดกลุ่มตัวอย่างไม่มีผลต่อการจำแนกผิดพลาดด้วยวิธีถดถอยโลจิสติกและวิธีแมนเทล-แฮนส์เซล

สวามินาทาน และโรเจอร์ (Swaminathan & Rogers, 1990, pp. 361-370) ได้เปรียบเทียบวิธีถดถอยโลจิสติกกับวิธีแมนเทล-แฮนส์เซล ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปและอเนกรูป โดยใช้สถานการณ์จำลอง 6 เงื่อนไข คือ กลุ่มตัวอย่าง 2 ขนาด คือ 250 คน และ 500 คน ความยาวแบบทดสอบ 3 ขนาด คือ 40 ข้อ 60 ข้อ และ 80 ข้อ ซึ่งในแบบทดสอบแต่ละชุดประกอบด้วยสัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 20% โดยครึ่งเป็นข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูป และอีกครึ่งหนึ่งเป็นข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูป ผลการศึกษาปรากฏว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป วิธีถดถอยโลจิสติกและวิธีแมนเทล-แฮนส์เซล ให้ผลการวิเคราะห์ใกล้เคียงกัน แต่วิธีถดถอยโลจิสติกให้ผลดีกว่า

วิธีแมนเทล-แฮนส์เซล ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป สำหรับอัตราความคลาดเคลื่อนประเภทที่ 1 พบว่าวิธีแมนเทล-แฮนส์เซล มีอัตราความคลาดเคลื่อนร้อยละ 1 ส่วนวิธีถดถอยโลจิสติก มีอัตราความคลาดเคลื่อนร้อยละ1-6 และวิธีถดถอยโลจิสติกเสียค่าใช้จ่ายสูงกว่าวิธีแมนเทล-แฮนส์เซล ประมาณ 3-4 เท่า

โรเจอร์ และสวามินาทาน (Roger & Swaminathan, 1993, pp. 105-116) ได้เปรียบเทียบวิธีถดถอยโลจิสติกกับวิธีแมนเทล-แฮนส์เซล ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบซึ่งทดสอบเกี่ยวกับการกระจายของสถิติทดสอบและประสิทธิภาพของสถิติทดสอบแต่ละวิธี โดยใช้ข้อมูลจำลอง การศึกษาด้านการกระจายของสถิติทดสอบ ค่าพารามิเตอร์ที่แปรเปลี่ยนได้แก่ ขนาดกลุ่มตัวอย่าง ความเหมาะสมของข้อมูลกับโมเดลค่าความยาก ค่าอำนาจจำแนก ความยาวแบบทดสอบ 40 ข้อ ส่วนการศึกษาด้านประสิทธิภาพของแต่ละวิธีมีค่าพารามิเตอร์ที่แปรเปลี่ยนได้แก่ ขนาดกลุ่มตัวอย่าง ความเหมาะสมของโมเดลกับข้อมูล ขนาดของแบบทดสอบ การกระจายของคะแนนสอบ อัตราส่วนของข้อสอบที่ทำหน้าที่ต่างกัน ค่าความยาก ค่าอำนาจจำแนกและพื้นที่ระหว่างเส้นโค้งคุณลักษณะข้อสอบระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ ผลการศึกษาปรากฏว่าการกระจายของค่าสถิติเป็นไปตามสมมติฐานที่ตั้งไว้เกือบทั้งหมดในทั้งสองวิธี กรณีที่การกระจายของค่าสถิติของวิธีถดถอยโลจิสติกไม่เป็นไปตามที่คาดไว้ เนื่องมาจากข้อสอบมีค่าความยากสูงและมีค่าอำนาจจำแนกสูง ด้านประสิทธิภาพพบว่าทั้งสองวิธีมีประสิทธิภาพไม่ต่างกันในการตรวจสอบ DIF แบบเอกรูป (Uniform DIF) แต่วิธีถดถอยโลจิสติกมีประสิทธิภาพมากกว่าในการตรวจสอบ DIF แบบอเนกรูป (Nonuniform DIF) ขนาดกลุ่มตัวอย่างเป็นปัจจัยที่มีผลกระทบอย่างมากต่ออัตราการตรวจสอบการทำหน้าที่ต่างกันของทั้งสองวิธีนี้กล่าวคือ เมื่อเพิ่มขนาดกลุ่มตัวอย่าง อัตราการตรวจสอบจะเพิ่มขึ้น ส่วนขนาดของแบบทดสอบและการกระจายของคะแนนไม่มีผลกระทบต่ออัตราการตรวจสอบ

เมเซอร์ และเคลาเซอร์ (Mazor & Clauser, 1995, pp. 131-144) ได้เปรียบเทียบวิธีถดถอยโลจิสติกกับวิธีแมนเทล-แฮนส์เซล ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อใช้เกณฑ์ภายนอกหรือความสามารถหลากหลาย (Multiple Ability) เข้ามาร่วมเป็นเกณฑ์เปรียบเทียบโดยศึกษาจากข้อมูลจริงซึ่งเป็นผลการตอบข้อสอบวัดผลสัมฤทธิ์ทางการเรียนวิชาประวัติศาสตร์และวิชาเคมีของนักเรียนระดับชั้นมัธยมศึกษา ความยาวแบบทดสอบ 75 ข้อ จำแนกกลุ่มผู้เข้าสอบตามเพศและความสามารถทางภาษาสำหรับเกณฑ์ที่ใช้จับคู่ ได้แก่ คะแนนรวมของแบบทดสอบ คะแนนรวมของแบบทดสอบร่วมกับเกณฑ์ภายนอก คือ ตัวแปรความถนัดทางภาษาและความถนัดทางคณิตศาสตร์ ผลการศึกษาปรากฏว่า วิธีถดถอยโลจิสติกตรวจพบข้อสอบทำหน้าที่ต่างกันมากกว่าวิธีแมนเทล-แฮนส์เซล ในทุกเงื่อนไข ส่วนใหญ่เมื่อใช้เกณฑ์ภายนอกทั้งสองตัวแปรเข้ามาเป็น

เกณฑ์จับคู่ร่วมกับคะแนนรวมจากแบบทดสอบจะทำให้พบข้อสอบที่ถูกระบุว่าทำหน้าที่ต่างกันน้อยลง โดยพบว่า เมื่อใช้คะแนนรวมเป็นเกณฑ์ร่วมกับตัวแปรความถนัดทางภาษาจะตรวจพบน้อยกว่าใช้คะแนนรวมเป็นเกณฑ์ร่วมกับตัวแปรความถนัดทางคณิตศาสตร์

นารายานัน และสวามินาทาน (Narayanan & Swaminathan, 1996, pp. 257-274) ได้เปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปด้วยวิธีแมนเทล-แฮนส์เซล วิธีถดถอยโลจิสติก และวิธีซิปเทสท์ โดยศึกษาอำนาจการตรวจสอบและการจำแนกผิดพลาดโดยการจำลองข้อมูล ตัวแปรที่ศึกษา ได้แก่ ขนาดกลุ่มตัวอย่าง ความแตกต่างของการกระจายความสามารถระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันที่มีภายในแบบทดสอบ ขนาดของพื้นที่ระหว่างเส้นโค้งคุณลักษณะข้อสอบของผู้เข้าสอบ 2 กลุ่ม ค่าความยาก และค่าอำนาจจำแนกของแบบทดสอบ ผลการศึกษาปรากฏว่า วิธีถดถอยโลจิสติก และวิธีซิปเทสท์มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปได้เท่าเทียมกันในทุกเงื่อนไขที่ศึกษา ส่วนวิธีแมนเทล-แฮนส์เซล ไม่สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปและการจำแนกผิดพลาด วิธีแมนเทล-แฮนส์เซล จะสูงกว่าวิธีถดถอยโลจิสติกและวิธีซิปเทสท์

จากการศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ในระยะประมาณ 10 ปีที่ผ่านมา พบว่าวิธีที่นิยมใช้ ได้แก่ วิธีแมนเทล-แฮนส์เซล วิธีถดถอยโลจิสติก และวิธีซิปเทสท์ นอกจากนี้ยังมีการพัฒนาวิธีอื่น ๆ อีก เช่น วิธีการวิเคราะห์องค์ประกอบจำกัด เป็นต้น ซึ่งมีการศึกษาทั้งข้อมูลจำลองและข้อมูลจริง รวมทั้งกำหนดเงื่อนไขในการศึกษาแตกต่างกันไป เช่น ความยาวของแบบทดสอบต่างกัน ขนาดของกลุ่มตัวอย่างต่างกัน เกณฑ์การจับคู่ต่างกัน และจำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบทดสอบต่างกัน แม้ว่าจะมีการศึกษาอยู่บ้าง แต่ยังไม่มีข้อสรุปที่ชัดเจนว่าวิธีใดมีประสิทธิภาพดีกว่ากันในเงื่อนไขของการศึกษาที่กำหนด

ดังนั้น ผู้วิจัยสนใจที่จะศึกษาการเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างวิธีการวิเคราะห์องค์ประกอบจำกัดกับวิธีถดถอยโลจิสติก ภายใต้เงื่อนไข 18 เงื่อนไข (3×3×2) คือ ขนาดของกลุ่มตัวอย่าง 3 ขนาด (2,000 คน 1,000 คน และ 300 คน) ความยาวของแบบทดสอบ 3 ขนาด (40 ข้อ 30 ข้อ และ 20 ข้อ) และเกณฑ์การจับคู่ 2 เกณฑ์ (คะแนนรวมทั้งฉบับ และคะแนนแบบทดสอบย่อย)