บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

ในการศึกษาครั้งนี้ ผู้วิจัยนำเสนอแนวคิดและทฤษฎีที่เกี่ยวข้องโดยแบ่งออกเป็น 5 ตอน ดังนี้

ตอนที่ 1 ความหมายของการทำหน้าที่ต่างกันของข้อสอบ

ตอนที่ 2 ประเภทของการทำหน้าที่ต่างกันของข้อสอบ

ตอนที่ 3 หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

ตอนที่ 4 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

ตอนที่ 5 งานวิจัยที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

ตอนที่ 1 ความหมายของการทำหน้าที่ต่างกันของข้อสอบ

การศึกษาเรื่องความยุติธรรมของข้อสอบหรือแบบสอบในกรณีที่จะทำให้ผู้สอบระหว่างกลุ่มย่อยมีความได้เปรียบหรือเสียเปรียบกันเดิมใช้คำว่า ***"ความลำเอียงของข้อสอบ"*** *(item bias)* หรือ ***"ความลำเอียงของแบบสอบ"*** *(test bias)* ซึ่งเป็นภาษาที่ใช้กันในทางสังคมและมีความหมายในเชิงลบ ส่วนการตัดสินว่าข้อสอบมีความลำเอียงหรือไม่นั้น มักจะพิจารณาอิทธิพลที่สังเกตได้ของกลุ่มผู้สอบย่อยที่นำมาศึกษาโดยไม่คำนึงถึงวิธีทางสถิติ จึงทำให้เกิดความคลุมเครือเกี่ยวกับเกณฑ์ที่ใช้ในการตัดสินความลำเอียง ต่อมาในระยะหลังนักวิธีวิทยาการวิจัยได้นำสารสนเทศทางสถิติมาใช้เป็นเกณฑ์ในการตัดสินความลำเอียงของข้อสอบ และได้เปลี่ยนไปใช้คำใหม่ว่า ***"การทำหน้าที่ต่างกันของข้อสอบ"*** *(differential item functioning; DIF)* และ ***"การทำหน้าที่ต่างกันของแบบสอบ"*** *(differential test functioning; DTF)* ซึ่งเป็นคำที่มีความเป็นกลางและเหมาะสมมากกว่า (Holland and Thayer, 1988; Holland and Wainner, 1993) สำหรับความหมายของการทำหน้าที่ต่างกันของข้อสอบและแบบสอบได้มีผู้ให้คำนิยามไว้หลายคน ดังนี้

การทำหน้าที่ต่างกันของแบบสอบ หมายถึง คะแนนเกณฑ์ที่ใช้ในการทำนายจากเส้นการถดถอยร่วมของสมาชิกผู้สอบกลุ่มย่อยมีความสอดคล้องกัน โดยสูงหรือต่ำเหมือนกัน (Cleary, 1968 cited in Camilli and Shepard, 1994)

การทำหน้าที่ต่างกันของข้อสอบ หมายถึง สัดส่วนของผู้สอบที่ตอบข้อสอบได้ถูกต้องไม่เท่ากันในแต่ละกลุ่มประชากรที่ใช้ในการพิจารณา เมื่อผู้สอบทั้งหมดที่มีคะแนนเท่ากันทำข้อสอบในชุดแบบสอบที่มีความเป็นเอกพันธ์ (Scheuneman, 1975 cited in Potenza and Dorans, 1995)

การทำหน้าที่ต่างกันของข้อสอบ หมายถึง โอกาสของการตอบข้อสอบได้ถูกต้องไม่เท่ากัน เมื่อผู้สอบทั้งหมดที่มีความสามารถระดับเดียวกัน แต่มาจากกลุ่มผู้สอบที่แตกต่างกัน (Pine, 1977 cited in S.-H. Kim, H.-O. Kim and Cohen, 1994)

การทำหน้าที่ต่างกันของข้อสอบ หมายถึง ข้อสอบที่มีค่าความยากสัมพัทธ์ในสมาชิกของผู้สอบกลุ่มหนึ่งมากกว่าสมาชิกของผู้สอบอีกกลุ่มหนึ่ง (Rudner, Getson and Knight, 1980)

การทำหน้าที่ต่างกันของข้อสอบ หมายถึง โอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกัน (การวัดความสามารถ) หรือโอกาสในการตอบข้อสอบในทางบวกแตกต่างกัน (การวัดเจตคติ) เมื่อผู้สอบที่มีคุณลักษณะของการวัดในปริมาณเท่ากัน แต่มาจากกลุ่มประชากรย่อยที่แตกต่างกัน (Hulin, Drasgow and Parson, 1983)

การทำหน้าที่ต่างกันของข้อสอบ หมายถึง โอกาสในการตอบข้อสอบได้ถูกต้องของผู้สอบกลุ่มหนึ่ง มีค่าต่ำกว่าหรือสูงกว่าผู้สอบอีกกลุ่มหนึ่งที่มีระดับความสามารถเดียวกัน (Dorans and Kulick, 1986)

การทำหน้าที่ต่างกันของแบบสอบ หมายถึง ความไม่ตรงหรือความคลาดเคลื่อนอย่างเป็นระบบในการวัด ซึ่งจะทำให้ผลของการวัดบิดเบือนสำหรับสมาชิกของกลุ่มผู้สอบบางกลุ่มโดยเฉพาะ (Camilli and Shepard, 1994)

การทำหน้าที่ต่างกันของข้อสอบ หมายถึง เทอมที่ใช้ในการอธิบายข้อสอบในแบบสอบซึ่งมีโอกาสของการตอบข้อสอบถูกแตกต่างกัน สำหรับผู้สอบสองกลุ่มที่มีความสามารถระดับเดียวกัน (Feinstein, 1995)

การทำหน้าที่ต่างกันของข้อสอบ หมายถึง ฟังก์ชันการตอบสนองข้อสอบซึ่งคำนวณจากสมาชิกของผู้สอบกลุ่มย่อยที่แตกต่างกันมีค่าไม่เท่ากัน (Narayanan and Swaminathan, 1996)

จากนิยามการทำหน้าที่ต่างกันของข้อสอบหรือแบบสอบดังกล่าวสามารถสรุปรวมได้ว่า ***"การทำหน้าที่ต่างกันของข้อสอบหรือแบบสอบ หมายถึง โอกาสของการตอบข้อสอบหรือแบบสอบได้ถูกต้องแตกต่างกัน สำหรับผู้สอบที่มีคุณลักษณะหรือความสามารถในระดับเดียวกัน แต่มาจากกลุ่มประชากรย่อยที่แตกต่างกัน"***

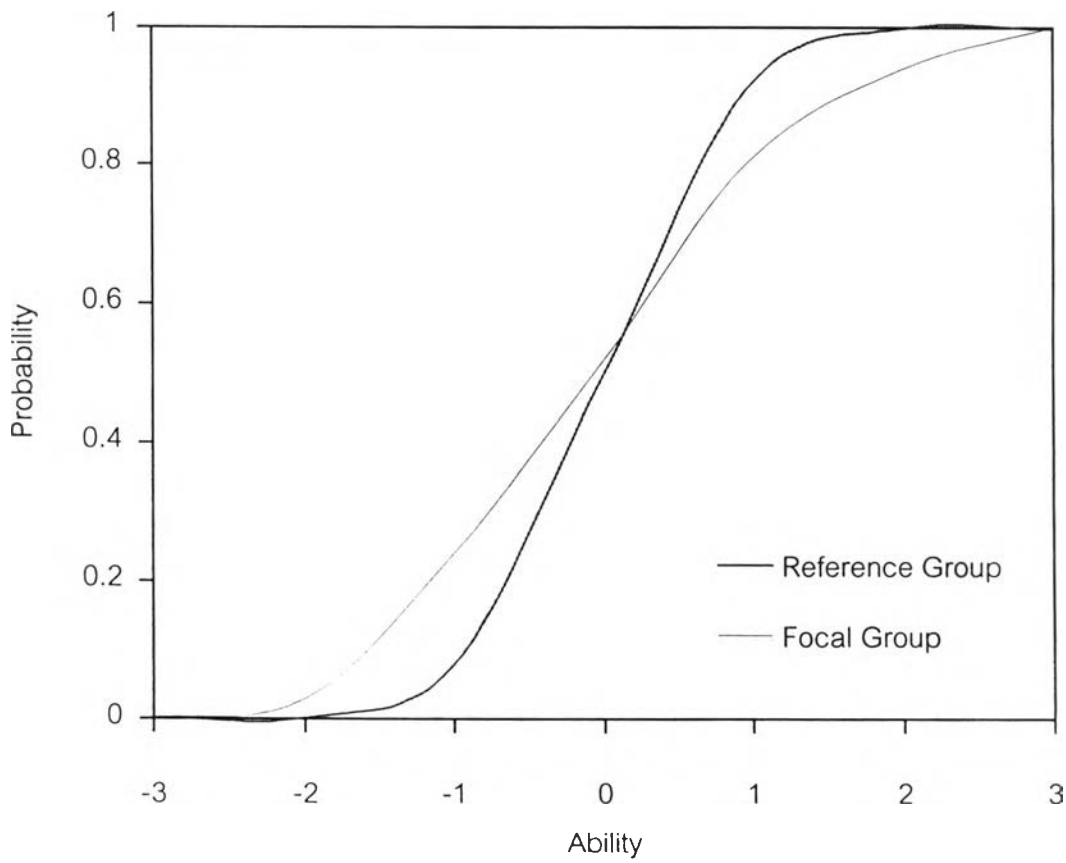
ตอนที่ 2 ประเภทของการทำหน้าที่ต่างกันของข้อสอบ

Mellenbergh (1982) ได้จำแนกรูปแบบของการทำหน้าที่ต่างกันของข้อสอบออกเป็น 2 ประเภท คือ ***การทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป*** *(uniform DIF)* และ***การทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป*** *(nonuniform DIF)* การทำหน้าที่ต่างกันของข้อสอบประเภทแรกจะเกิดขึ้นเมื่อไม่มีปฏิสัมพันธ์ (interaction) ระหว่างระดับความสามารถของผู้สอบและการเป็นสมาชิกของกลุ่มผู้สอบ (group membership) ส่วนการทำหน้าที่ต่างกันของข้อสอบประเภทหลังจะเกิดขึ้นเมื่อมีปฏิสัมพันธ์ระหว่างระดับความสามารถของผู้สอบและการเป็นสมาชิกของกลุ่มผู้สอบ ในวิธีวิทยาการวิจัยตามทฤษฎีการตอบสนองข้อสอบ (item response theory; IRT) สามารถพิจารณา "ปฏิสัมพันธ์" ดังกล่าวได้จากความแตกต่างของค่าพารามิเตอร์อำนาจจำแนกของข้อสอบระหว่างผู้สอบกลุ่มย่อยสองกลุ่ม กล่าวคือ ถ้าข้อสอบทำหน้าที่ต่างกันแบบเอกรูปแล้วโค้งลักษณะข้อสอบ (item characteristic curves; ICCs) ระหว่างผู้สอบกลุ่มย่อยสองกลุ่มจะขนานกัน แต่ถ้าข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปแล้วโค้งลักษณะข้อสอบระหว่างผู้สอบกลุ่มย่อยสองกลุ่มจะไม่ขนานกัน ดังนั้นความแตกต่างระหว่างโค้งลักษณะข้อสอบทั้งสองแบบจะบ่งบอกถึงขนาดและทิศทางของข้อสอบที่ทำหน้าที่ต่างกัน ซึ่งสามารถคำนวณได้โดยใช้สูตรการคำนวณพื้นที่ของ Raju (1990) โดยทั่ว ๆ ไป ในแบบสอบมาตรฐานมักจะมีข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูปมากกว่าข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูป แต่ในข้อมูลจริงจะมีข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปได้มากกว่า

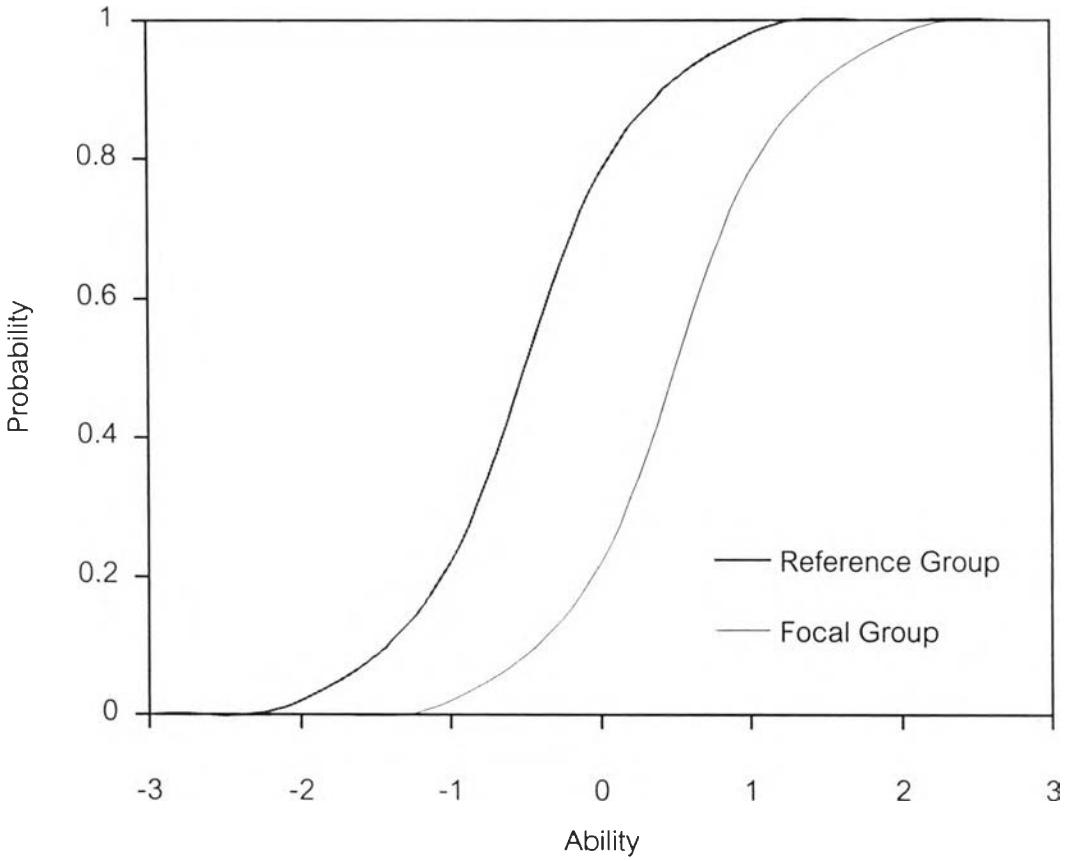
จากนิยามการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปที่ Mellenbergh (1982) กำหนดข้างต้นมีความหมายครอบคลุมในสถานการณ์ที่โค้งลักษณะข้อสอบตัดกันหรือไม่ตัดกันก็ได้ ทั้งนี้เนื่องจากผลการศึกษาของ Swaminathan และ Rogers (1990) พบว่า ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปมี 2 ลักษณะ คือ ***ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปโดยมีปฏิสัมพันธ์ไม่เป็นลำดับ*** *(disordinal interaction)* และ***ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปโดยมีปฏิสัมพันธ์เป็นลำดับ*** *(ordinal interaction)* ข้อสอบที่ทำหน้าที่ต่างกันลักษณะแรกเกิดขึ้นเมื่อโค้งลักษณะข้อสอบตัดกันตรงจุดกึ่งกลางของช่วงความสามารถ (ช่วงความสามารถมีค่าตั้งแต่ –3 ถึง +3) ส่วนข้อสอบที่ทำหน้าที่ต่างกันลักษณะหลังเกิดขึ้นเมื่อลักษณะข้อสอบตัดกันนอกช่วงความสามารถ ซึ่งอาจตัดกันตรงปลายสุดของช่วงความสามารถต่ำหรือสูง อย่างไรก็ตาม โค้งลักษณะข้อสอบที่ไม่ขนานกันอาจไม่ตัดกันก็ได้ สถานการณ์ดังกล่าวอาจเกิดขึ้นเมื่อใช้โมเดลแบบ 3 พารามิเตอร์ ดังนั้นโค้งลักษณะข้อสอบที่ไม่ขนานกันของข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปอาจจะตัดกันหรือ

ไม่ตัดกันก็ได้ ต่อมา Li และ Stout (1993 cited in Narayanan and Swaminathan, 1996) ได้เรียกข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปที่มีปฏิสัมพันธ์ไม่เป็นลำดับว่า ***"ข้อสอบทำหน้าที่ต่างกันแบบไม่มีทิศทาง"*** *(nondirectional DIF)* และเรียกข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปที่มีปฏิสัมพันธ์เป็นลำดับว่า ***"ข้อสอบทำหน้าที่ต่างกันแบบมีทิศทางเดียว"*** *(unidirectional DIF)* ดังนั้นจึงเป็นทางเลือกใหม่ในการใช้ถ้อยคำดังกล่าว

เมื่อพิจารณาถึงโค้งลักษณะข้อสอบอาจเรียกข้อสอบทำหน้าที่ต่างกันแบบไม่มีทิศทางอีกชื่อหนึ่งว่า ***"ข้อสอบทำหน้าที่ต่างกันแบบตัดกัน"*** *(crossing DIF)* ส่วนข้อสอบทำหน้าที่ต่างกันแบบมีทิศทางเดียวอาจเรียกอีกชื่อหนึ่งว่า ***"ข้อสอบทำหน้าที่ต่างกันของข้อสอบแบบไม่ตัดกัน"*** *(noncrossing DIF)* ซึ่งโค้งลักษณะข้อสอบที่ทำหน้าที่ต่างกันแบบตัดกันอาจตัดกันได้มากกว่า 1 จุดตลอดช่วงความสามารถ (Li and Stout, 1996) สำหรับข้อสอบที่ทำหน้าที่ต่างกันทั้งสองประเภทแสดงดังภาพที่ 1 และ 2



ภาพที่ 1 ข้อสอบทำหน้าที่ต่างกันแบบไม่ทิศทาง (nondirectional DIF)



ภาพที่ 2 ข้อสอบทำหน้าที่ต่างกันแบบมีทิศทางเดียว (unidirectional DIF)

ตอนที่ 3 หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจะเปรียบเทียบผลการตอบข้อสอบระหว่างผู้สอบสองกลุ่มที่มีความสามารถระดับเดียวกัน โดยกำหนดให้ผู้สอบกลุ่มหนึ่งเป็น ***"กลุ่มอ้างอิง"*** (*reference group; R*) และผู้สอบอีกกลุ่มหนึ่งเป็น ***"กลุ่มเปรียบเทียบ"*** (*focal group;* F) (Holland and Wainer, 1993) ผู้สอบกลุ่มแรกจะเป็นตัวแทนกลุ่มหลัก (majority group) ในประชากร ซึ่งเป็นกลุ่มที่ใช้อ้างหลักฐาน ส่วนผู้สอบกลุ่มหลังจะเป็นตัวแทนกลุ่มรอง (minority group) ในประชากร ซึ่งเป็นกลุ่มผู้สอบที่จะทำการศึกษาการทำหน้าที่ต่างกันของข้อสอบ (Angoff, 1993) ถ้าข้อสอบทำหน้าที่ต่างกันแล้วโอกาสในการตอบข้อสอบถูกของผู้สอบแต่ละกลุ่มจะไม่เท่ากัน โดยที่ผู้สอบกลุ่มแรกคาดว่าจะได้เปรียบในการตอบข้อสอบ ในขณะที่ผู้สอบกลุ่มหลังคาดว่าจะเสียเปรียบในการตอบข้อสอบ สำหรับเกณฑ์ที่ใช้ในการจำแนกผู้สอบเป็นกลุ่มเปรียบเทียบและกลุ่มอ้างอิงมีหลายลักษณะ เช่น เพศ สีผิว เชื้อชาติ ภาษา วัฒนธรรม ภูมิลำเนา เป็นต้น

ตอนที่ 4 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

การจัดประเภทของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ สามารถจำแนกได้หลายลักษณะขึ้นอยู่กับเกณฑ์ที่ใช้ในการจำแนก ดังเช่น ถ้าใช้เกณฑ์การให้คะแนนของข้อสอบจะแบ่งออกเป็น 2 กลุ่มวิธี คือ กลุ่มวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนแบบทวิภาค (dichotomous DIF methods) และกลุ่มวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนแบบพหุวิภาค (polytomous DIF methods) สำหรับวิธีการตรวจสอบในกลุ่มแรกจะให้คะแนนแบบ 0 – 1 ส่วนวิธีการตรวจสอบในกลุ่มหลังจะให้คะแนนแบบหลายค่า ถ้าใช้เกณฑ์ตามทฤษฎีของการวิเคราะห์ข้อมูลจะแบ่งออกเป็น 2 กลุ่มวิธีใหญ่ ๆ คือ กลุ่มวิธี non-IRT (non item response theory methods) และกลุ่มวิธี IRT (item response theory methods) ในกลุ่มวิธี non-IRT จะวิเคราะห์ดัชนีการทำหน้าที่ต่างกันของข้อสอบโดยใช้คะแนนที่สังเกตได้ภายใต้ทฤษฎีแบบดั้งเดิม ส่วนกลุ่มวิธี IRT จะวิเคราะห์โดยใช้คะแนนที่สังเกตไม่ได้หรือตัวแปรแฝงภายใต้ทฤษฎีการตอบสนองข้อสอบ แต่ถ้าใช้ข้อตกลงเบื้องต้นของโมเดลเป็นเกณฑ์สามารถแบ่งออกเป็น 2 รูปแบบ คือ รูปแบบพาราเมทริก (parametric form) และรูปแบบนันพาราเมทริก (nonparametric form) ในรูปแบบพาราเมทริกจะวิเคราะห์ดัชนีการทำหน้าที่ต่างกันของข้อสอบซึ่งมีข้อตกลงเบื้องต้นของโมเดลสำหรับอธิบายความสัมพันธ์ระหว่างคะแนนของข้อสอบและการจับคู่ตัวแปร ส่วนรูปแบบนันพาราเมทริกจะไม่มีข้อตกลงดังกล่าว สำหรับการจำแนกวิธีการตรวจสอบที่มีการให้คะแนนแบบทวิภาคสามารถสรุปรวมเป็นตารางได้ดังนี้ (Hulin and others, 1983; Millsap and Everson, 1993; Holland and Wainer, 1993; Feinstein, 1995; Potenza and Dorans, 1995)

ตารางที่ 1 วิธีการตรวจสอบการทำหน้าที่ต่างกันที่มีการให้คะแนนแบบทวิภาค จำแนกตามทฤษฎีที่ใช้ในการวิเคราะห์และรูปแบบของโมเดล

วิธีการตรวจสอบ DIF	รูปแบบพาราเมทริก (Parametric Form)	รูปแบบนันพาราเมทริก (Nonparametric Form)
กลุ่มวิธี Non-IRT	○ วิธีวิเคราะห์ความแปรปรวน ○ วิธีการถดถอยโลจิสติก	○ วิธีแปลงค่าความยากของข้อสอบ ○ วิธีตารางการณ์จร ❑ วิธีไค-สแควร์ ❑ วิธีลอก-ลิเนียร์ ❑ วิธีแมนเทล-แฮนส์เซล ○ วิธีการทำให้เป็นมาตรฐาน
กลุ่มวิธี IRT	○ วิธีการวัดพื้นที่ ❑ วิธีการวัดพื้นที่ของ Raju ❑ วิธีการวัดพื้นที่ของ Kim และ Cohen ○ วิธีการเปรียบเทียบค่าพารามิเตอร์ ❑ วิธีเปลี่ยนค่าความยาก ❑ วิธีการทดสอบ F ❑ วิธีการทดสอบไค-สแควร์ของ Lord ❑ วิธี IRT แบบเทียม ❑ วิธีการทดสอบอัตราส่วนไลค์ลิฮู้ด	○ วิธีซิปเทสท์

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนแบบทวิภาคในตารางที่ 1 มีรายละเอียดดังนี้

1. **กลุ่มวิธี** non-IRT

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในกลุ่มวิธี non-IRT มักจะใช้คะแนนรวมของผู้สอบเป็นเกณฑ์การจับคู่กลุ่มผู้สอบ ซึ่งเป็นเกณฑ์ภายใน (Angoff, 1993; Dorans and Holland, 1993) วิธีการวิเคราะห์ที่สำคัญในกลุ่มนี้ ได้แก่

1.1 **วิธีการวิเคราะห์ความแปรปรวน** (analysis of variance; ANOVA)

วิธีการวิเคราะห์ความแปรปรวนเป็นวิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในยุคเริ่มต้น (Cleary and Hilton, 1968 cited Camilli and Shepard, 1994) ซึ่งสามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจากผู้สอบกลุ่มย่อยตั้งแต่สองกลุ่มขึ้นไปที่ทำแบบสอบร่วมกัน ในการวิเคราะห์ด้วยวิธีนี้จะทดสอบผลของปฏิสัมพันธ์ (interaction effect) ระหว่างองค์ประกอบกลุ่มผู้สอบและข้อสอบ โดยที่กลุ่มผู้สอบและข้อสอบเป็นตัวแปรอิสระ ส่วนผลการตอบข้อสอบถูกเป็นตัวแปรตาม ถ้าผลการทดสอบมีนัยสำคัญแสดงว่าค่าความยากสัมพัทธ์ของข้อสอบจากผู้สอบกลุ่มหนึ่งสูงกว่าผู้สอบอีกกลุ่มหนึ่ง นั่นคือ ข้อสอบทำหน้าที่ต่างกัน (Rudner, Geston and Knight, 1980) ในการวิเคราะห์ความแปรปรวนจะแยกความแปรปรวนของคะแนนทั้งหมด (V_t) ออกเป็นความแปรปรวนย่อย ๆ ดังนี้ (Osterlind, 1983)

$$V_t = V_b + V_w + V_e$$

เมื่อ V_b แทน ความแปรปรวนระหว่างกลุ่ม

V_w แทน ความแปรปรวนภายในกลุ่ม

V_e แทน ความแปรปรวนของความคลาดเคลื่อน

ความแปรปรวนดังกล่าวสามารถแสดงในรูปกำลังสองของส่วนเบี่ยงเบนมาตรฐาน ($V = \sigma^2$) โดยที่ความแปรปรวนของคะแนนทั้งหมดสำหรับผู้สอบกลุ่ม j ตอบข้อสอบข้อที่ i แล้วได้ระดับคะแนน k (σ^2_{ijk}) ประกอบด้วยความแปรปรวนย่อย ๆ ดังนี้

$$\sigma^2_{ijk} = \sigma^2_i + \sigma^2_j + \sigma^2_{ij} + \sigma^2_{k(ij)}$$

เมื่อ σ^2_i แทน ความแปรปรวนของข้อสอบ

σ^2_j แทน ความแปรปรวนของกลุ่มผู้สอบ

σ^2_{ij} แทน ความแปรปรวนของปฏิสัมพันธ์ระหว่างกลุ่มผู้สอบและข้อสอบ

$\sigma^2_{k(ij)}$ แทน ความแปรปรวนของความคลาดเคลื่อน

สำหรับสมมติฐานของการทดสอบ กำหนดดังนี้

$$H_0 : \sigma^2_{ij} = 0$$

$$H_1 : \sigma^2_{ij} \neq 0$$

สมมติฐานศูนย์เป็นสมมติฐานที่ไม่มีปฏิสัมพันธ์ระหว่างกลุ่มผู้สอบและข้อสอบ ถ้าผลการทดสอบสมมติฐานด้วยสถิติ F ปรากฏว่าไม่มีนัยสำคัญ (ยอมรับ H_0) แสดงว่าข้อสอบทำหน้าที่ไม่ต่างกัน และถ้าผลการทดสอบมีนัยสำคัญ (ปฏิเสธ H_0) แสดงว่าในแบบสอบมีข้อสอบทำหน้าที่ต่างกัน แต่ยังไม่ทราบว่าข้อสอบข้อใดทำหน้าที่ต่างกัน ต่อจากนั้นจะนำข้อสอบมาทดสอบเป็นรายข้อด้วยวิธีการเปรียบเทียบภายหลัง (post hoc) ซึ่งอาจใช้วิธีของ Tukey หรือ Scheffe/ ก็ได้ ผลการเปรียบเทียบจะทำให้สามารถตัดสินได้ว่าข้อสอบข้อใดทำหน้าที่ต่างกัน

1.2 **วิธีแปลงค่าความยากของข้อสอบ** (transformed item difficulty; TID)

Angoff (1972 cited Camilli and Shepard, 1994) เสนอวิธีแปลงค่าความยากของข้อสอบเพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ หลักการตรวจสอบด้วยวิธีดังกล่าวจะพิจารณาค่าความยากสัมพัทธ์ของข้อสอบระหว่างผู้สอบกลุ่มย่อย 2 กลุ่ม หรือมากกว่า 2 กลุ่ม ซึ่งมีลักษณะคล้ายกับวิธีการวิเคราะห์ความแปรปรวน แต่มีข้อแตกต่างตรงที่ผลการตรวจสอบด้วยวิธีแปลงค่าความยากของข้อสอบจะมีดัชนีวัดขนาดการทำหน้าที่ต่างกันของข้อสอบ ในขณะที่วิธีการวิเคราะห์ความแปรปรวนจะไม่มีดัชนีดังกล่าว สำหรับการวิเคราะห์จะเริ่มต้นด้วยการคำนวณค่าความยากของข้อสอบ (item difficulty; p) ทุกข้อจากผู้สอบแต่ละกลุ่ม ต่อจากนั้นจะแปลงค่า p ให้เป็นคะแนนมาตรฐาน (standardized; Z) แล้วจึงแปลงค่า Z ให้เป็นค่าเดลตา (delta; Δ) อีกต่อหนึ่ง โดยใช้สมการดังนี้ (Osterlind, 1983)

$$\Delta_{ij} = 4Z_{ij} + 13$$

เมื่อ Δ_{ij} แทน ดัชนีความยากของข้อสอบข้อที่ i จากผู้สอบกลุ่ม j

Z_{ij} แทน คะแนนมาตรฐานของข้อสอบข้อที่ i จากผู้สอบกลุ่ม j ซึ่งมีการแจกแจงแบบปกติ (ค่าเฉลี่ยเท่ากับ 0 และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 1)

สำหรับค่าความคลาดเคลื่อนมาตรฐานของ Δ_{ij} คำนวณได้จากสูตร

$$SE_{\Delta_{ij}} = \frac{4}{N_j - 1}$$

เมื่อ $SE_{\Delta_{ij}}$ แทน ค่าความคลาดเคลื่อนมาตรฐานของ Δ_{ij}

N_j แทน จำนวนผู้สอบในกลุ่ม j

นำค่าเดลตาของข้อสอบระหว่างผู้สอบกลุ่มย่อย 2 กลุ่มมาพิจารณาการทำหน้าที่ต่างกันของข้อสอบ โดยการทดสอบสมมติฐาน ดังนี้

$$H_0 : \Delta_{i1} - \Delta_{i2} = 0$$
$$H_1 : \Delta_{i1} - \Delta_{i2} \neq 0$$

ในการพิจารณาว่าข้อสอบข้อใดทำหน้าที่ต่างกัน สามารถตรวจสอบได้โดยการนำค่าเดลตาแต่ละคู่มาลงจุดบนกราฟสองแกน ซึ่งให้แกนนอนแทนค่าเดลตาของผู้สอบกลุ่มหนึ่ง และแกนตั้งแทนค่าเดลตาของผู้สอบอีกกลุ่มหนึ่ง จุดต่าง ๆ ที่เขียนบนกราฟมีลักษณะเป็นรูปวงรีทำมุม 45 องศาจากจุดเริ่มต้น แล้วคำนวณระยะห่างของจุดดังกล่าวกับเส้นแกนหลัก (D_i) โดยใช้สูตรดังนี้

$$D_i = \frac{bx_i + a - y_i}{\sqrt{b^2 + 1}}$$ สำหรับข้อสอบข้อที่ i

โดยที่ $y = bx + a$

$$a = M_x - bM_y$$

$$b = \frac{\left(\sigma_y^2 - \sigma_x^2\right) \pm \sqrt{\left(\sigma_y^2 - \sigma_x^2\right)^2 + 4r_{xy}^2\sigma_x^2\sigma_y^2}}{2r_{xy}\sigma_x\sigma_x}$$

เมื่อ x แทน ค่าเดลตาแต่ละค่าของผู้สอบกลุ่ม 1

y แทน ค่าเดลตาแต่ละค่าของผู้สอบกลุ่ม 2

a แทน ค่าที่ตัดแกน y

b แทน ค่าความชันของเส้นกราฟ

M_x แทน ค่าเฉลี่ยเลขคณิตของค่าเดลตาแต่ละค่าของผู้สอบกลุ่ม 1

M_y แทน ค่าเฉลี่ยเลขคณิตของค่าเดลตาแต่ละค่าของผู้สอบกลุ่ม 2

σ_x แทน ค่าเบี่ยงเบนมาตรฐานของค่าเดลตาแต่ละค่าของผู้สอบกลุ่ม 1

σ_y แทน ค่าเบี่ยงเบนมาตรฐานของค่าเดลตาแต่ละค่าของผู้สอบกลุ่ม 2

r_{xy} แทน ค่าสัมประสิทธิ์สหสัมพันธ์ของค่าเดลตาแต่ละค่าของผู้สอบกลุ่ม 1 และกลุ่ม 2

เมื่อข้อสอบข้อใดที่จุดอยู่ห่างจากเส้นแกนหลักเกินกว่าเกณฑ์ที่กำหนดไว้ แสดงว่าข้อสอบข้อนั้นทำหน้าที่ต่างกัน

1.3 **วิธีตารางการณ์จร** (contingency table; CT)

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีตารางการณ์จร สามารถนำมาประยุกต์ใช้กับกลุ่มตัวอย่างขนาดเล็ก ในการวิเคราะห์จะสร้างตาราง 3 ทิศทาง ซึ่งประกอบด้วย กลุ่มผู้สอบ (กลุ่มอ้างอิง / กลุ่มเปรียบเทียบ) ผลการตอบข้อสอบ (ถูก / ผิด) และคะแนนรวม K ระดับ (1, 2, 3,...K) สำหรับการทดสอบทางสถิติมีรูปแบบนันพาราเมทริก ซึ่งไม่จำเป็นต้องใช้โมเดลในการประมาณค่า วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ใช้พื้นฐานวิธีตารางการณ์จรที่สำคัญมีดังนี้

1.3.1 **วิธีไค-สแควร์** (chi-square; χ^2)

Scheuneman (1979) ได้เสนอวิธีไค-สแควร์เพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ในการวิเคราะห์จะแบ่งคะแนนรวมที่ใช้แทนความสามารถของผู้สอบออกเป็นช่วง ๆ โดยปกติจะใช้ 3 ถึง 5 ช่วง แล้วใช้ผลการตอบข้อสอบถูกเพียงอย่างเดียวในแต่ละช่วงคำนวณดัชนีไค-สแควร์ ดังนี้

$$\chi^2 = \sum \frac{(B_e - B_o)^2}{B_e} + \sum \frac{(W_e - W_o)^2}{W_e}$$

เมื่อ χ^2 แทน ดัชนีไค-สแควร์ของ Scheuneman

B_e แทน ความถี่ที่คาดหวังของผลการตอบข้อสอบถูกในผู้สอบกลุ่ม B

W_e แทน ความถี่ที่คาดหวังของผลการตอบข้อสอบถูกในผู้สอบกลุ่ม W

B_o แทน ความถี่ที่สังเกตได้ของผลการตอบข้อสอบถูกในผู้สอบกลุ่ม B

W_o แทน ความถี่ที่สังเกตได้ของผลการตอบข้อสอบถูกในผู้สอบกลุ่ม W

สำหรับความถี่ที่คาดหวังของผลการตอบข้อสอบถูกคำนวณได้จากสูตร

$$E_{ij} = \frac{O_j}{N_j} N_{ij}$$

เมื่อ E_{ij} แทน ความถี่ที่คาดหวังของผลการตอบข้อสอบถูกในผู้สอบกลุ่ม i ที่มีช่วงคะแนน j

O_j แทน จำนวนผู้สอบที่ตอบข้อสอบถูก ซึ่งมีคะแนนรวมในช่วง j

N_j แทน จำนวนผู้สอบทั้งหมด ซึ่งมีคะแนนรวมในช่วง j

N_{ij} แทน จำนวนผู้สอบทั้งหมดในกลุ่ม i ซึ่งมีคะแนนรวมในช่วง j

ดัชนีไค-สแควร์ดังกล่าวจะทดสอบที่ระดับชั้นของความเป็นอิสระเท่ากับ $(k-1) \times (r-1)$ เมื่อ k เป็นจำนวนกลุ่มผู้สอบ และ r เป็นจำนวนคะแนนกลุ่มผู้สอบ

1.3.2 **วิธีลอก-ลิเนียร์** (log-linear; LL)

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีนี้ปรับขยายมาจากวิธีไค-สแควร์ ของ Scheuneman (1979) โดยนำมาประยุกต์กับโมเดลโลจิท (logit model) และโมเดลลอก-ลิเนียร์ (log-linear model) (Mellenbergh, 1982) ภายใต้วิธีการดังกล่าวจะนำผลการตอบข้อสอบ (ทั้งถูกและผิด) มาสร้างตารางการณ์จรแบบ 3 ทิศทาง (ระดับคะแนน × กลุ่มผู้สอบ × ผลการตอบข้อสอบ) โมเดลลอก-ลิเนียร์ในรูปลอกธรรมชาติ (natural logarithm; ln) ของผลการตอบข้อสอบประเภทคะแนน i (ถ้าตอบถูก $i = 1$ และถ้าตอบผิด $i = 2$) ผู้สอบกลุ่ม j ($j = 1, 2, 3, \ldots, g$) ในระดับคะแนน k ($k = 1, 2, 3, \ldots, s$) มีลักษณะดังนี้ (Bishop, Fienberg and Holland, 1975 cited in Millsap and Everson, 1993)

$$\ln F_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{23(jk)} + u_{123(ijk)}$$

สำหรับผลการตอบข้อสอบแบบทวิภาคตามโมเดลลอก-ลิเนียร์สามารถแปลงเป็นโมเดลโลจิทในรูปลอกธรรมชาติของอัตราส่วนของจำนวนข้อสอบที่ตอบถูกต่อจำนวนข้อสอบที่ตอบผิด ดังนี้

$$\ln\left[\frac{F_{1jk}}{F_{2jk}}\right] = \alpha + \beta_k + \delta_j + (\delta\beta)_{jk}$$

เมื่อ F_{1jk} แทน ความถี่ที่คาดหวังของผลการตอบข้อสอบถูก ในผู้สอบกลุ่ม j ณ ระดับคะแนน k

F_{2jk} แทน ความถี่ที่คาดหวังของผลการตอบข้อสอบผิด ในผู้สอบกลุ่ม j ณ ระดับคะแนน k

α แทน พารามิเตอร์อิทธิพลความยากของข้อสอบทั้งหมด

β_k แทน อิทธิพลของระดับคะแนนหลัก

δ_j แทน อิทธิพลของกลุ่มผู้สอบหลัก

$(\delta\beta)_{jk}$ แทน ผลของปฏิสัมพันธ์ระหว่างระดับคะแนนและกลุ่มผู้สอบ

ในการตัดสินการทำหน้าที่ต่างกันของข้อสอบจะใช้สถิติอัตราส่วนไลค์ลิฮู้ด (likelihood ratio; G^2) ซึ่งมีการแจกแจงแบบเชิงเส้นกำกับ (asymptotically) ทดสอบความเหมาะสมของโมเดลกับข้อมูล สถิติดังกล่าวมีลักษณะดังนี้

$$G^2 = 2\sum_{i=1}^{2}\sum_{j=1}^{g}\sum_{k=1}^{s} f_{ijk}\ln(f_{ijk}/\hat{F}_{ijk})$$

โดยที่ $$\hat{F}_{ijk} = \left(\sum_{j=1}^{g} f_{ijk}\right)\left(\sum_{i=1}^{2} f_{ijk}\right)\Big/\left(\sum_{j=1}^{g}\sum_{i=1}^{2} f_{ijk}\right)$$

เมื่อ f_{ijk} แทน ความถี่ที่สังเกตได้

$\hat{F}_{ijk}$ แทน ความถี่ของค่าประมาณที่คาดหวังภายใต้โมเดลที่ใช้ทดสอบ

จากโมเดลโลจิทดังกล่าว เมื่อตัดเทอม δ_j และ $(\delta\beta)_{jk}$ ออกไป แล้วนำไปทดสอบการทำหน้าที่ไม่ต่างกันของข้อสอบ (no DIF) ถ้าโมเดลมีความเหมาะสมกับข้อมูลแสดงว่าข้อสอบทำหน้าที่ไม่ต่างกัน ซึ่งตามปกติแล้วจะทดสอบโมเดลดังกล่าวก่อนเสมอ แต่เมื่อตัดเฉพาะเทอม $(\delta\beta)_{jk}$ แล้วโมเดลมีความเหมาะสมกับข้อมูล (โดยที่โมเดล no DIF ไม่มีความเหมาะสมกับข้อมูล) แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบเอกรูป (uniform DIF) และในโมเดลโลจิทที่ไม่ได้ตัดเทอมใด ๆ ออกไป แล้วโมเดลมีความเหมาะสมกับข้อมูลแสดงว่าข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป (nonuniform DIF)

1.3.3 วิธีแมนเทล-แฮนส์เซล (Mantel-Haenszel; MH)

วิธีแมนเทล-แฮนส์เซลพัฒนาโดย Mantel และ Haenszel (1959 cited in Camilli and Shepard, 1994) ซึ่งประกอบด้วยอัตราส่วนแต้มต่อร่วม (common odds ratio) และสถิติไค-สแควร์ (chi-square statistic) ต่อมา Holland (1985 cited in Holland and Thayer, 1988) ได้นำวิธีแมนเทล-แฮนส์เซลไปประยุกต์ใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในหน่วยงานการบริการทดสอบทางการศึกษา (Education Testing Service; ETS) ของประเทศสหรัฐอเมริกา จึงทำให้วิธีแมนเทล-แฮนส์เซลเป็นที่ยอมรับจากนักวิธีวิทยาการวิจัยอย่างกว้างขวาง วิธีแมนเทล-แฮนส์เซลเป็นวิธีตารางการณ์จรแบบไม่คำนวณทวนซ้ำ (noniterative contingency table) ซึ่งพัฒนามาจากวิธีไค-สแควร์แบบประเพณีนิยม (traditional) สามารถนำไปใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ง่าย มีขั้นตอนการคำนวณที่ไม่สลับซับซ้อน มีการทดสอบทางสถิติแบบนันพาราเมทริก (nonparametric) ซึ่งไม่จำเป็นต้องใช้โมเดลประมาณค่า

หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยวิธีแมนเทล-แฮนส์เซลจะเปรียบเทียบผลการตอบข้อสอบระหว่างผู้สอบสองกลุ่ม กลุ่มหนึ่งเรียกว่า ***"กลุ่มเปรียบเทียบ"*** ซึ่งเป็นกลุ่มผู้สอบที่ผู้วิจัยสนใจศึกษา ถ้าข้อสอบที่นำมาศึกษาทำหน้าที่ต่างกันคาดว่าผู้สอบกลุ่มนี้จะเสียเปรียบในการตอบข้อสอบ ส่วนผู้สอบอีกกลุ่มหนึ่งเรียกว่า ***"กลุ่มอ้างอิง"*** ซึ่งเป็นกลุ่มผู้สอบที่ใช้เป็นมาตรฐานในการเปรียบเทียบผลการตอบข้อสอบกับผู้สอบกลุ่มแรก ถ้าข้อสอบทำหน้าที่ต่างกันแล้วคาดว่าผู้สอบกลุ่มหลังนี้จะได้เปรียบในการตอบข้อสอบ สำหรับการเปรียบเทียบผลการตอบข้อสอบจะเปรียบเทียบทุกระดับความสามารถของผู้สอบกลุ่มย่อยสองกลุ่มที่มีระดับความสามารถเท่ากัน ในทางปฏิบัติมักใช้คะแนนรวมของแบบสอบเป็นเกณฑ์การจับคู่ของกลุ่มผู้สอบ เกณฑ์ในลักษณะดังกล่าวมีจุดอ่อน กล่าวคือ ถ้าในแบบสอบมีข้อสอบที่ทำหน้าที่ต่างกัน การใช้คะแนนรวมของแบบสอบที่ใช้เป็นเกณฑ์ในการจับคู่ของกลุ่มผู้สอบจะรวมคะแนนจากข้อสอบที่ทำหน้าที่ต่างกันดังกล่าวด้วย ซึ่งจะทำให้ได้เกณฑ์ที่มีความบิดเบือนจากความเป็นจริง ดังนั้น ผลการตรวจสอบจึงไม่ถูกต้อง ในการแก้ไขจุดอ่อนดังกล่าว Holland และ Thayer (1988) ได้เสนอแนะให้ใช้วิธีการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบแบบสองขั้นตอน เพื่อจะทำให้เกณฑ์ในการจับคู่ของกลุ่มผู้สอบมีความบริสุทธิ์ (purification of matching criterion) สำหรับขั้นตอนการวิเคราะห์มีดังนี้

ขั้นตอนแรก ใช้คะแนนรวมของแบบสอบทั้งฉบับเป็นเกณฑ์การจับคู่กลุ่มผู้สอบย่อย 2 กลุ่ม แล้ววิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ เมื่อพบว่าข้อสอบข้อใดที่ทำหน้าที่ต่างกันให้นำคะแนนของข้อสอบข้อนั้นออกจากคะแนนรวมของผู้สอบแต่ละคน

ขั้นตอนที่สอง ใช้คะแนนรวมของแบบสอบที่นำเอาข้อสอบที่ทำหน้าที่ต่างกันซึ่งตรวจพบในขั้นตอนแรกออกไป แล้ววิเคราะห์การทำหน้าที่ต่างกันของข้อสอบซ้ำอีกครั้งหนึ่ง

เมื่อจับคู่กลุ่มผู้สอบแล้วจะนำข้อมูลผลการตอบข้อสอบระหว่างผู้สอบย่อย 2 กลุ่มมาจัดลงในตารางการณ์จรแบบ 2 × 2 (กลุ่มผู้สอบ 2 กลุ่ม × ผลการตอบ 2 แบบ) โดยที่ตารางการณ์จร 1 ตารางแทนคะแนนรวม 1 ระดับ ดังนั้นถ้ามีคะแนนรวม K ระดับ จะต้องสร้างตารางการณ์จรแบบ 2 × 2 ทั้งหมด K ตาราง สำหรับตารางการณ์จรแบบ 2 × 2 ของข้อสอบแต่ละข้อที่มีคะแนนรวมระดับ j แสดงดังตารางที่ 2

ตารางที่ 2 ผลการตอบข้อสอบระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ณ ระดับคะแนน j

กลุ่ม	คะแนนของผลการตอบข้อสอบที่ศึกษา		รวม
	ตอบถูกได้ 1 คะแนน	ตอบผิดได้ 0 คะแนน	
R	A_j	B_j	nR_j
F	C_j	D_j	nF_j
รวม	m_{1j}	m_{0j}	T_j

เมื่อ A_j แทน จำนวนผู้สอบกลุ่มอ้างอิงที่ระดับคะแนน j ซึ่งตอบข้อสอบถูก

B_j แทน จำนวนผู้สอบกลุ่มอ้างอิงที่ระดับคะแนน j ซึ่งตอบข้อสอบผิด

C_j แทน จำนวนผู้สอบกลุ่มเปรียบเทียบที่ระดับคะแนน j ซึ่งตอบข้อสอบถูก

D_j แทน จำนวนผู้สอบกลุ่มเปรียบเทียบที่ระดับคะแนน j ซึ่งตอบข้อสอบผิด

m_{1j} แทน จำนวนผู้สอบทั้งหมดที่ระดับคะแนน j ที่ตอบข้อสอบถูก

m_{0j} แทน จำนวนผู้สอบทั้งหมดที่ระดับคะแนน j ที่ตอบข้อสอบผิด

nR_j แทน จำนวนผู้สอบกลุ่มอ้างอิงที่ระดับคะแนน j

nF_j แทน จำนวนผู้สอบกลุ่มเปรียบเทียบที่ระดับคะแนน j

T_j แทน จำนวนผู้สอบทั้งหมดที่ระดับคะแนน j

ต่อจากนั้นจึงนำผลการตอบข้อสอบจากตารางที่ 3 มาคำนวณสัดส่วนของผลการตอบข้อสอบถูกและผิด ดังตารางที่ 3

ตารางที่ 3 สัดส่วนของผลการตอบข้อสอบระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ณ ระดับคะแนน j

กลุ่ม	คะแนนของผลการตอบข้อสอบที่ศึกษา		รวม
	ตอบถูกได้ 1 คะแนน	ตอบผิดได้ 0 คะแนน	
R	pR_j	qR_j	nR_j
F	pF_j	qF_j	nF_j

เมื่อ pR_j แทน สัดส่วนของผู้สอบกลุ่มอ้างอิงที่ระดับคะแนน j ซึ่งตอบข้อสอบถูก

qR_j แทน สัดส่วนของผู้สอบกลุ่มอ้างอิงที่ระดับคะแนน j ซึ่งตอบข้อสอบผิด

pF_j แทน สัดส่วนของผู้สอบกลุ่มเปรียบเทียบที่ระดับคะแนน j ซึ่งตอบข้อสอบถูก

qF_j แทน สัดส่วนของผู้สอบกลุ่มเปรียบเทียบที่ระดับคะแนน j ซึ่งตอบข้อสอบผิด

สำหรับการทดสอบสมมติฐานของการทำหน้าที่ต่างกันของข้อสอบจะกำหนดสมมติฐานศูนย์ (H_0) และสมมติฐานอื่น (H_1) ดังนี้

$$H_0 : \frac{pR_j}{qR_j} = \frac{pF_j}{qF_j} \quad ; \ j = 1, 2, 3, \ldots, K$$

$$H_1 : \frac{pR_j}{qR_j} = \alpha \frac{pF_j}{qF_j} \quad ; \ j = 1, 2, 3, \ldots, K \text{ เมื่อ } \alpha \neq 1$$

สมมติฐานศูนย์เป็นสมมติฐานที่เป็นอิสระอย่างมีเงื่อนไขของสมาชิกในกลุ่มผู้สอบและคะแนนจากข้อสอบที่ศึกษา ดังนั้นคะแนนที่ได้จากตารางที่ 3 และภายใต้สมมติฐานศูนย์สามารถสรุปเป็นค่าคาดหวัง (expected values) ในแต่ละเซลได้ดังนี้

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

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

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

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

พารามิเตอร์ α ภายใต้สมมติฐานอื่นเรียกว่าอัตราส่วนแต้มต่อร่วม (common odds ratio) ซึ่งคำนวณได้จาก

$$\alpha = \frac{\frac{pR_j}{qR_j}}{\frac{pF_j}{qF_j}} = \frac{pR_j qF_j}{qR_j pF_j}$$

ถ้า α = 1 แสดงว่าโอกาสของการตอบข้อสอบถูกระหว่างผู้สอบทั้งสองกลุ่มมีค่าเท่ากัน ถ้า α > 1 แสดงว่าผู้สอบกลุ่มอ้างอิงมีโอกาสของการตอบข้อสอบถูกมากกว่าผู้สอบกลุ่มเปรียบเทียบ และถ้า α < 1 แสดงว่าผู้สอบกลุ่มเปรียบเทียบมีโอกาสของการตอบข้อสอบถูกมากกว่าผู้สอบกลุ่มอ้างอิง Mantel และ Haenszel (1959 cited in Holland and Thayer, 1988) ได้ประมาณค่า α จากตารางไขว้แบบ 2 × 2 ดังนี้

$$\hat{\alpha}_{MH} = \frac{\sum_{j=1}^{K} A_j D_j / T_j}{\sum_{j=1}^{K} B_j C_j / T_j}$$

$\hat{\alpha}_{MH}$ เป็นค่าประมาณขนาดอิทธิพลของการทำหน้าที่ต่างกันของข้อสอบ (DIF effect size) ซึ่งมีค่าอยู่ระหว่าง 0 ถึง ∞ Holland และ Thayer (1985 cited in Holland and Thayer, 1988) ได้เสนอให้แปลงค่า $\hat{\alpha}_{MH}$ เป็นสเกลเดลตา (delta scale; MH_{D-DIF}) ซึ่งเป็นค่าที่ใช้ในหน่วยงานการบริการทดสอบทางการศึกษาของประเทศสหรัฐอเมริกา ดังนี้

$$MH_{D-DIF} = -2.35\ln(\hat{\alpha}_{MH})$$

ค่า MH_{D-DIF} ดังกล่าวสามารถนำไปพิจารณาค่าความยากของข้อสอบ กล่าวคือ ถ้า MH_{D-DIF} มีค่าเป็นศูนย์ แสดงว่าข้อสอบของแต่ละกลุ่มยากเท่ากัน ถ้า MH_{D-DIF} มีค่าเป็นลบ แสดงว่าข้อสอบยากสำหรับผู้สอบกลุ่มเปรียบเทียบมากกว่ากลุ่มอ้างอิง และถ้า MH_{D-DIF}

มีค่าเป็นบวก แสดงว่าข้อสอบยากสำหรับผู้สอบกลุ่มกลุ่มอ้างอิงมากกว่ากลุ่มเปรียบเทียบ ส่วนการประมาณค่าความคลาดเคลื่อนมาตรฐาน (standard error; SE) ของ MH_{D-DIF} สามารถคำนวณได้จากสูตรดังนี้

$$SE(MH_{D-DIF}) = 2.35\sqrt{Var\left[\ln(\hat{\alpha}_{MH})\right]}$$

โดยที่ $$Var\left[\ln(\hat{\alpha}_{MH})\right] = \frac{\sum_{j=1}^{K} U_j V_j / T_j^2}{2\left[\sum_{j=1}^{K} A_j D_j / T_j\right]^2}$$

ขณะที่ $$U_i = A_j D_j + \hat{\alpha}_{MH}(B_j C_j)$$

และ $$V_i = (A_j + D_j) + \hat{\alpha}_{MH}(B_j + C_j)$$

ในการทดสอบนัยสำคัญของสมมติฐาน จะนำค่า $\hat{\alpha}_{MH}$ และค่า MH_{D-DIF} ไปทดสอบกับสถิติแมนเทล-แฮนส์เซลไค-สแควร์ (χ^2_{MH}) ที่ระดับชั้นความเป็นอิสระเท่ากับ 1 (df = 1) โดยนำค่า $\hat{\alpha}_{MH}$ ไปเปรียบเทียบกับ 1 ส่วนค่า MH_{D-DIF} จะเปรียบเทียบกับ 0 สำหรับสถิติ χ^2_{MH} มีสูตรในการคำนวณดังนี้

$$\chi^2_{MH} = \frac{\left[\left|\sum_{j=1}^{K} A_j - E(A_j)\right| - 0.5\right]^2}{\sum_{j=1}^{K} Var(A_j)}$$

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

$$Var(A_j) = \frac{nR_j nF_j m_{1j} m_{0j}}{T_j^2 (T_j - 1)}$$

เมื่อ $E(A_j)$ แทน ค่าคาดหวังของจำนวนผู้สอบกลุ่มอ้างอิงที่ระดับคะแนน j ซึ่งตอบข้อสอบถูก

$Var(A_j)$ แทน ค่าความแปรปรวนของจำนวนผู้สอบกลุ่มอ้างอิงที่ระดับคะแนน j ซึ่งตอบข้อสอบถูก

ถ้าค่า $\hat{\alpha}_{MH}$ ไม่แตกต่างจาก 1 อย่างมีนัยสำคัญ แสดงว่าข้อสอบทำหน้าที่ไม่ต่างกัน (no DIF) แต่ถ้าค่า $\hat{\alpha}_{MH}$ แตกต่างจาก 1 อย่างมีนัยสำคัญ แสดงว่าข้อสอบทำหน้าที่ต่างกัน (DIF) โดยจะเข้าข้างกลุ่มเปรียบเทียบเมื่อ $\hat{\alpha}_{MH}$ มีค่าเป็นบวก และจะเข้าข้างกลุ่มอ้างอิงเมื่อ $\hat{\alpha}_{MH}$ มีค่าเป็นลบ ส่วนการพิจารณาค่า MH_{D-DIF} ว่าแตกต่างจาก 0 อย่างมีนัยสำคัญทางสถิติหรือไม่ก็มีวิธีการพิจารณาลักษณะทำนองเดียวกัน

นอกจากนี้ หน่วยงานการบริการทดสอบทางการศึกษาของประเทศสหรัฐอเมริกาได้จัดกลุ่มข้อสอบที่ทำหน้าที่ต่างกันไว้ 3 กลุ่ม คือกลุ่ม A, B และ C โดยที่กลุ่ม A ประกอบด้วยข้อสอบทำหน้าที่ต่างกันเล็กน้อย หรือข้อสอบทำหน้าที่ต่างกันอย่างไม่มีนัยสำคัญ ส่วนกลุ่ม B และกลุ่ม C ประกอบด้วยข้อสอบทำหน้าที่ต่างกันอย่างมีนัยสำคัญ ซึ่งกลุ่ม B จัดเป็นข้อสอบทำหน้าที่ต่างกันเล็กน้อยถึงปานกลาง และกลุ่ม C จัดเป็นข้อสอบทำหน้าที่ต่างกันปานกลางถึงมาก สำหรับเกณฑ์ที่ใช้ในการจัดกลุ่มจะพิจารณา 2 องค์ประกอบ คือ ค่านัยสำคัญทางสถิติ และค่าสัมบูรณ์ของ MH_{D-DIF} ซึ่งมีรายละเอียดดังนี้ (Zieky, 1993)

กลุ่ม A MH_{D-DIF} มีค่าแตกต่างจาก 0 อย่างไม่มีนัยสำคัญ หรือ
$|MH_{D-DIF}|$ มีค่าน้อยกว่า 1

กลุ่ม B MH_{D-DIF} มีค่าแตกต่างจาก 0 อย่างมีนัยสำคัญ และ $|MH_{D-DIF}|$ มีค่าตั้งแต่ 1 ถึง 1.5 หรือ
MH_{D-DIF} มีค่ามากกว่า 1 อย่างไม่มีนัยสำคัญ

กลุ่ม C MH_{D-DIF} มีค่ามากกว่า 1 อย่างมีนัยสำคัญ และ $|MH_{D-DIF}|$ มีค่ามากกว่า 1.5

1.4 วิธีการทำให้เป็นมาตรฐาน (standardization; STND)

วิธีการทำให้เป็นมาตรฐานพัฒนาโดย Dorans และ Kulick (1986) สามารถนำมาใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ หลักการตรวจสอบจะวิเคราะห์ความแตกต่างของผลการตอบข้อสอบระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบในแต่ละระดับความสามารถที่ใช้จับคู่เปรียบเทียบ แล้วพิจารณาความแตกต่างระหว่างการถดถอยของข้อสอบในแบบสอบซึ่งเป็นข้อมูลเชิงประจักษ์จากผู้สอบ 2 กลุ่ม โดยใช้สูตรความแตกต่างของค่า p มาตรฐาน (standardized p-difference; D_{STD}) ดังนี้

$$D_{STD} = \frac{\sum_{k=1}^{s} W_k (P_{fk} - P_{rk})}{\sum_{k=1}^{s} W_k}$$

โดยที่ $P_{fk} = \frac{f_{1fk}}{n_{fk}}$

และ $P_{rk} = \frac{f_{1rk}}{n_{rk}}$

เมื่อ D_{STD} แทน ดัชนีการทำให้เป็นมาตรฐาน

$W_k / \sum(W_k)$ แทน น้ำหนักองค์ประกอบที่ระดับคะแนน k ซึ่งเป็นน้ำหนักความแตกต่างในสัดส่วนการตอบข้อสอบถูกระหว่างผู้สอบกลุ่มเปรียบเทียบ (P_{fk}) และกลุ่มอ้างอิง (P_{rk})

ดัชนี D_{STD} มีค่าตั้งแต่ –1 ถึง +1 ถ้ามีค่าเป็นบวกแสดงว่าข้อสอบเข้าข้างผู้สอบกลุ่มเปรียบเทียบ และถ้ามีค่าเป็นลบแสดงว่าข้อสอบเข้าข้างผู้สอบกลุ่มอ้างอิง ดัชนี D_{STD} เป็นดัชนีคิดเครื่องหมาย สำหรับดัชนีไม่คิดเครื่องหมายจะอยู่ในรูปรากกำลังสองของความแตกต่างถ่วงน้ำหนักเฉลี่ย (root mean weighted squared different; RMWSD) ดังนี้

$$RMWSD = \sqrt{\frac{\sum_{k=1}^{s} W_k (P_{fk} - P_{rk})^2}{\sum_{k=1}^{s} W_k}}$$

1.5 วิธีการถดถอยโลจิสติก (logistic regression; LR)

Swaminathan และ Rogers (1990) ได้พัฒนาวิธีการถดถอยโลจิสติกเพื่อใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนแบบทวิภาค (dichotomous) วิธีดังกล่าวมีแนวคิดมาจากวิธีตารางการณ์จร โดยดัดแปลงมาจากวิธีลอก-ลิเนียร์ของ Mellenbergh (1982) และวิธีแมนเทล-แฮนส์เซลของ Holland และ Thayer (1988) หลักการตรวจสอบด้วยวิธีการถดถอยโลจิสติกจะใช้โมเดลการถดถอยโลจิสติกทำนายโอกาสของผลการตอบข้อสอบถูกโมเดลดังกล่าวใช้ตัวแปรความสามารถแบบต่อเนื่องซึ่งมีเทอมที่ใช้คำนวณปฏิสัมพันธ์ระหว่างการเป็นสมาชิกของกลุ่มผู้สอบและระดับความสามารถ จึงทำให้สามารถตรวจสอบการทำหน้าที่

ต่างกันของข้อสอบได้ทั้งแบบเอกรูป (uniform DIF) และแบบอเนกรูป (nonuniform DIF) นอกจากนี้ยังสามารถนำไปประยุกต์กับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีผู้สอบหลายกลุ่ม และการให้คะแนนข้อสอบแบบพหุวิภาค (polytomous) (Miller and Spray, 1993)

ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการถดถอยโลจิสติกจะใช้สมการมาตรฐานของโมเดลการถดถอยโลจิสติกคำนวณผลการตอบข้อสอบถูก ดังนี้ (Bock, 1975 cited in Swaminathan and Rogers, 1990)

$$P(U_{pj}=1|\theta_{pj})=\frac{\exp(\beta_{0j}+\beta_{1j}\theta_{pj})}{1+\exp(\beta_{0j}+\beta_{1j}\theta_{pj})} \quad ; \quad p=1,...,n_j \ , j=1,2$$

เมื่อ $P(U_{pj}=1|\theta_{pj})$ แทน ผลการตอบข้อสอบของผู้สอบคนที่ p ในกลุ่ม j

β_{0j} แทน พารามิเตอร์จุดตัด

β_{1j} แทน พารามิเตอร์ความชันสำหรับกลุ่ม j

θ_{pj} แทน ค่าความสามารถที่สังเกตได้ของผู้สอบคนที่ p ในกลุ่ม j

จากโมเดลดังกล่าว ถ้า $\beta_{01}=\beta_{02}$ และ $\beta_{11}=\beta_{12}$ แล้วฟังก์ชันการถดถอยโลจิสติกของผู้สอบสองกลุ่มเหมือนกัน แสดงว่าข้อสอบทำหน้าที่ไม่ต่างกัน (no DIF) ถ้า $\beta_{11}=\beta_{12}$ แต่ $\beta_{01}\neq\beta_{02}$ แล้วฟังก์ชันการถดถอยโลจิสติกของผู้สอบสองกลุ่มขนานกันแต่ไม่ทับกัน แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบเอกรูป (uniform DIF) และถ้า $\beta_{01}=\beta_{02}$ แต่ $\beta_{11}\neq\beta_{12}$ แล้วฟังก์ชันการถดถอยโลจิสติกของผู้สอบไม่ขนานกัน แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป (nonuniform DIF) นอกจากนี้โมเดลการถดถอยโลจิสติกดังกล่าวสามารถเปลี่ยนเป็นโมเดลในรูปการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปและแบบอเนกรูป ดังนี้

$$P(U_{pj}=1)=\frac{\exp^{Z_{pj}}}{1+\exp^{Z_{pj}}}$$

โดยที่ $Z_{pj}=\tau_0+\tau_1\theta_{pj}+\tau_2 g_j+\tau_3(\theta_{pj}g_j)$

เมื่อ $P(U_{pj}=1)$ แทน โอกาสในการตอบข้อสอบถูกของผู้สอบคนที่ pในกลุ่ม j

θ_{pj} แทน ระดับความสามารถของผู้สอบ (คะแนนรวม) คนที่ p ในกลุ่ม j

τ_0 แทน พารามิเตอร์จุดตัด

τ_1 แทน สัมประสิทธิ์ของความสามารถ

τ_2 แทน ความแตกต่างในกลุ่มของผลการตอบข้อสอบ

τ_3 แทน ปฏิสัมพันธ์ระหว่างกลุ่มและระดับความสามารถ

g_j แทน สมาชิกผู้สอบในกลุ่ม j (โดยกำหนดให้ g_p = 1 ถ้าผู้สอบอยู่ในกลุ่ม 1 และ g_p = 0 ถ้าผู้สอบอยู่ในกลุ่ม 2)

โมเดลการถดถอยโลจิสติกข้างต้น สามารถเปลี่ยนเป็นโมเดลเชิงเส้นในเมทริกโลจิท (logit metric) ซึ่งจะอยู่ในรูป log ของอัตราส่วนของโอกาสในการตอบข้อสอบถูกต่อโอกาสในการตอบข้อสอบผิด ดังนี้

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

จากโมเดลดังกล่าว เทอม $\theta_{pj}\ g_j$ เป็นผลคูณของตัวแปรอิสระ θ_{pj} และ g_j ส่วนพารามิเตอร์ τ_2 และ τ_3 สอดคล้องกับเทอมของพารามิเตอร์ในสมการของโมเดลการถดถอยโลจิสติก ดังนี้

$$\tau_2 = \beta_{01} - \beta_{02}$$
$$\tau_3 = \beta_{11} - \beta_{12}$$

ในการตัดสินว่าข้อสอบทำหน้าที่ต่างกันแบบเอกรูปหรือแบบอเนกรูป จะพิจารณาพารามิเตอร์ τ_2 และ τ_3 กล่าวคือ ถ้า $\tau_2 \neq 0$ และ $\tau_3 = 0$ แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบเอกรูป และถ้า $\tau_3 \neq 0$ ซึ่ง $\tau_2 = 0$ หรือไม่ก็ได้ แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป สำหรับพารามิเตอร์ของข้อสอบในแต่ละข้อของโมเดล Z_{pj} จะประมาณค่าโดยใช้วิธีไลค์ลิฮู้ดสูงสุด (maximum likelihood estimation; MLE) ซึ่งกำหนดในรูปฟังก์ชันดังนี้

$$L(U_{pj}|\theta) = \prod_{p=1}^{N}\prod_{j=1}^{n} P(U_{pj})^{u_{pj}}\left[1 - P(U_{pj})\right]^{1-u_{pj}}$$

โดยที่ N และ n แทน ขนาดกลุ่มตัวอย่างและความยาวของแบบสอบตามลำดับ สำหรับค่าประมาณของพารามิเตอร์โดยใช้วิธีไลค์ลิฮู้ดสูงสุดมีการแจกแจงแบบปกติของตัวแปรพหุในรูปเชิงเส้นกำกับ (asymptotically multivariate normal) ซึ่งมีค่าเฉลี่ยของเวคเตอร์ τ และเมทริกซ์ความแปรปรวน-ความแปรปรวนร่วมในรูป $\sum$ ในขณะที่ $\sum^{-1}$ เป็นเมทริกซ์สารสนเทศ กำหนดดังนี้

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

เมื่อ E และ $\ln L$ แทนค่าคาดหมายของเมทริกซ์และลอการิทึมของฟังก์ชันไลค์ลิฮู้ด ตามลำดับ ดังนั้นการแจกแจงของการประมาณค่าพารามิเตอร์ด้วยวิธี MLE จะอยู่ในรูปดังนี้

$$\hat{\tau} \sim N(\tau, \Sigma)$$

โดยที่ $\hat{\tau}' = [\tau_0, \tau_1, \tau_2, \tau_3]$ ส่วนความคลาดเคลื่อนมาตรฐานเชิงเส้นกำกับของค่าประมาณของ $\tau_s (S = 0,1,2,3.)$ เมื่อ S เป็นสมาชิกแนวเส้นทแยงมุมของ $\sum$ สามารถคำนวณได้จากสูตรดังนี้

$$SE\hat{\tau}_s = \sqrt{\sum^{ss}}$$

ในการทดสอบสมมติฐานของการทำหน้าที่ต่างกันของข้อสอบจะทดสอบสมาชิกของ τ_s ซึ่งสมมติฐานที่สนใจคือ $H_0 : \tau_2 = 0$ และ $H_0 : \tau_3 = 0$ สมมติฐานทั้งสองสามารถทดสอบพร้อม ๆ กัน (simultaneously) ดังนี้

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

โดยที่ C เป็นเมทริกซ์ขนาด 2 × 4 ดังนี้

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

ส่วนการทดสอบนัยสำคัญของสมมติฐานจะใช้สถิติไค-สแควร์ที่ระดับชั้นความเป็นอิสระเท่ากับ 2 (df = 2) ดังนี้

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

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

2. **กลุ่มวิธี** IRT

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในกลุ่มวิธี IRT จะไม่ใช้คะแนนรวมเป็นเกณฑ์การจับคู่กลุ่มผู้สอบ แต่จะใช้ค่าประมาณระดับความสามารถ (θ) ของผู้สอบเป็นเกณฑ์การจับคู่กลุ่มผู้สอบ ซึ่งแตกต่างจากวิธีการตรวจสอบในกลุ่ม non-IRT ในการวิเคราะห์โดยใช้ทฤษฎี IRT จะต้องเป็นไปตามข้อตกลงเบื้องต้นที่สำคัญ 3 ประการ ดังนี้ (Hambleton and Swaminathan, 1985; Lord, 1980)

ประการแรก ความเป็นเอกมิติของแบบสอบ (unidimensionality test) หมายถึง แบบสอบจะต้องวัดความสามารถหรือคุณลักษณะเพียงลักษณะเดียว โดยทั่วไปแล้วเป็นการยากที่จะสร้างแบบสอบให้เป็นไปตามข้อตกลงเบื้องต้นข้อนี้ ลักษณะที่เป็นไปได้ก็คือพยายามเลี่ยงไปใช้ข้อตกลงที่ผ่อนคลายลง โดยพิจารณามิติเด่น (dominant dimension) เพียงมิติเดียวก็น่าจะเพียงพอแล้ว นั่นคือ ทำให้แบบสอบสามารถวัดคุณลักษณะที่ต้องการวัดเป็นส่วนใหญ่ ถึงแม้ว่าอาจจะมีคุณลักษณะอื่นเข้ามาเกี่ยวข้องบ้างก็เป็นสิ่งที่หลีกเลี่ยงไม่ได้ ความเป็นเอกมิติในลักษณะนี้เรียกว่า ***เอกมิติสำคัญ*** (*essential unidimension*) สำหรับการตรวจสอบแบบสอบว่ามีความเป็นเอกมิติหรือไม่นั้นมักนิยมใช้วิธีวิเคราะห์องค์ประกอบ (factor analysis) โดยวิเคราะห์องค์ประกอบสำคัญเพื่อพิจารณาความแปรปรวนสูงสุดจากองค์ประกอบหลักตัวแรกว่ามีสัดส่วนมากกว่าองค์ประกอบอื่น ๆ หรือไม่ ถ้ามีมากกว่าก็แสดงว่าแบบสอบนั้นมีแนวโน้มเป็นเอกมิติ

ประการที่สอง ความเป็นอิสระต่อกันในการตอบข้อสอบ (local independence) หมายถึง โอกาสของการตอบข้อสอบในแต่ละข้อได้ถูกต้องมีความเป็นอิสระทางสถิติ (statistically independence) และมีความเป็นอิสระจากตำแหน่ง (uncorrelated independence) กล่าวคือ การตอบข้อสอบข้อใดข้อหนึ่งไม่ว่าถูกหรือผิดจะไม่มีผลต่อการตอบข้อสอบข้ออื่น ๆ ในแบบสอบ และลำดับในการตอบข้อสอบจะไม่มีผลต่อการตอบข้อสอบของผู้สอบ ดังนั้นข้อสอบจะปรากฏอยู่ในตำแหน่งใดของแบบสอบก็ได้ การตรวจสอบความเป็นอิสระต่อกันในการตอบข้อสอบสามารถทำได้โดยการทดสอบไค-สแควร์ของข้อสอบครั้งละคู่ที่ระดับความสามารถเดียวกัน นอกจากนี้ยังสามารถพิจารณาได้จากความเป็นเอกมิติของแบบสอบ กล่าวคือ ถ้าแบบสอบฉบับใดมีความเป็นเอกมิติแล้วแบบสอบฉบับนั้นจะมีความเป็นอิสระต่อกันในการตอบข้อสอบโดยอัตโนมัติ

ประการที่สาม ฟังก์ชันการตอบสนองข้อสอบ (item response function ; IRF) หรือโค้งลักษณะข้อสอบ (item characteristic curve; ICC) เป็นโค้งแสดงความสัมพันธ์ระหว่างโอกาสของการตอบข้อสอบได้ถูกต้องกับระดับความสามารถของผู้สอบ โค้งลักษณะข้อสอบในยุคเริ่มต้นจะใช้ฟังก์ชันปกติสะสม (normal ogive function) ซึ่งมีลักษณะเป็นโค้งรูปตัว S ต่อมาได้เปลี่ยนมา

ใช้ฟังก์ชันโลจิสติก (logistic function) เนื่องจากมีสูตรคณิตศาสตร์ที่สามารถคำนวณได้ง่าย และให้ผลการประมาณค่าได้ใกล้เคียงกันมาก โค้งลักษณะข้อสอบมีหลายลักษณะ ทั้งนี้จะขึ้นอยู่กับโมเดลการตอบสนองข้อสอบ เช่น โมเดลปกติสะสม (normal ogive model) โมเดลโลจิสติก (logistic model) โมเดลการตอบสนองแบบเกรด (grade response model) โมเดลการให้คะแนนแบบบางส่วน (partial credit model) โมเดลการตอบสนองแบบคะแนนต่อเนื่อง (continuous response model) เป็นต้น

สำหรับข้อสอบที่มีการให้คะแนนแบบทวิภาค (dichotomous) ซึ่งเป็นข้อสอบที่ตรวจให้คะแนนแบบ 0 – 1 (ตอบถูกได้ 1 คะแนน และตอบผิดให้ 0 คะแนน) มักนิยมใช้โมเดลโลจิสติก (logistic model) แบบ 1, 2 และ 3 พารามิเตอร์ โดยที่โมเดลโลจิสติกแบบ 1 พารามิเตอร์ (1PLM) หรือโมเดล Rasch เป็นโมเดลที่อธิบายคุณลักษณะของข้อสอบด้วยพารามิเตอร์ความยาก (b_i) เพียงตัวเดียวเท่านั้น ส่วนพารามิเตอร์อำนาจจำแนก (a_i) จะกำหนดให้มีค่าคงที่ทุกข้อ ($\bar{a}$) ในขณะที่พารามิเตอร์การเดา (c_i) มีค่าเป็นศูนย์ โมเดลดังกล่าวสามารถเขียนในรูปคณิตศาสตร์ ดังนี้

$$P_i(\theta) = \left\{1 + \exp\left[-D\bar{a}_i(\theta - b_i)\right]\right\}^{-1}$$

ในโมเดลโลจิสติกแบบ 2 พารามิเตอร์ (2PLM) จะกำหนดให้พารามิเตอร์ความยาก (b_i) และพารามิเตอร์อำนาจจำแนก (a_i) มีค่าแปรเปลี่ยนไปตามระดับความสามารถของผู้สอบ ในขณะที่พารามิเตอร์การเดา (c_i) มีค่าเป็นศูนย์ ดังนี้

$$P_i(\theta) = \left\{1 + \exp\left[-Da_i(\theta - b_i)\right]\right\}^{-1}$$

เมื่อเพิ่มพารามิเตอร์การเดา (c_i) เข้าไปอีก 1 ตัว โมเดลโลจิสติกแบบ 2 พารามิเตอร์ก็จะกลายเป็นโมเดลโลจิสติกแบบ 3 พารามิเตอร์ (3PLM) ดังนี้

$$P_i(\theta) = c_i + (1 - c_i)\left\{1 + \exp\left[-Da_i(\theta - b_i)\right]\right\}^{-1}$$

เมื่อ $P_i(\theta)$ แทน โอกาสของผู้สอบซึ่งมีระดับความสามารถ θ จะตอบข้อสอบข้อที่ i ได้ถูกต้อง

θ แทน ค่าความสามารถของผู้สอบ

a_i แทน ค่าอำนาจจำแนกของข้อสอบข้อที่ i

b_i แทน ค่าความยากของข้อสอบข้อที่ i

c_i แทน ค่าการเดาของข้อสอบข้อที่ i

D แทน ค่าองค์ประกอบของการปรับสเกล ซึ่งมีค่าเท่ากับ 1.7

exp แทน ค่าคงที่ของลอการิทึมธรรมชาติ ซึ่งมีค่าประมาณ 2.71818...

พารามิเตอร์ความสามารถของผู้สอบและพารามิเตอร์ของข้อสอบ มีความหมายดังนี้ (Osterlind, 1983)

ค่าความสามารถของผู้สอบ (examinee ability; θ) หมายถึง ระดับความสามารถจริงของผู้สอบ ซึ่งประมาณค่าได้จากผลการตอบข้อสอบตามทฤษฎี IRT โดยปกติแล้วจะปรับค่าความสามารถให้เป็นคะแนนมาตรฐาน (ค่าเฉลี่ยเท่ากับ 0 และส่วนเบี่ยงเบนมาตรฐานท่ากับ 1) ในทางทฤษฎีแล้ว θ มีค่าอยู่ในช่วง $-\infty$ ถึง $+\infty$ แต่ในทางปฏิบัติมักใช้ค่าอยู่ในช่วง –3 ถึง +3

ค่าความยากของข้อสอบ (item difficulty; $\mathbf{b_i}$) หมายถึง ระดับความสามารถของผู้สอบตรงจุดเปลี่ยนโค้ง (inflexion point) หรือตรงที่โค้งลักษณะข้อสอบมีความชันมากที่สุด ซึ่งเป็นจุดที่ผู้สอบมีระดับความสามารถเท่ากับค่าความยากของข้อสอบ ($\theta = b_i$) จะมีโอกาสตอบข้อสอบข้อนั้น ๆ ได้ถูกต้อง [$P_i(\theta) = (1 + c) / 2$ สำหรับโมเดลแบบ 3 พารามิเตอร์ และ $P_i(\theta) =$ 0.5 สำหรับโมเดลแบบ 1 หรือ 2 พารามิเตอร์] ในทางทฤษฎีแล้ว b_i มีค่าอยู่ในช่วง $-\infty$ ถึง $+\infty$ แต่ในทางปฏิบัติมักนิยมใช้ค่า b_i อยู่ในช่วง –2.5 ถึง +2.5 ข้อสอบที่มีค่า b_i อยู่ใกล้ –2.5 แสดงว่าเป็นข้อสอบที่ง่าย ส่วนข้อสอบที่มีค่า b_i อยู่ใกล้ +2.5 แสดงว่าเป็นข้อสอบที่ยาก

ค่าอำนาจจำแนกของข้อสอบ (item discriminantion; $\mathbf{a_i}$) หมายถึง ค่าบนโค้งลักษณะข้อสอบตรงที่มีความชันมากที่สุด หรือที่จุด $\theta = b_i$ ในทางทฤษฎีแล้ว a_i มีค่าอยู่ในช่วง $-\infty$ ถึง $+\infty$ ข้อสอบที่มีค่า a_i เป็นลบแสดงถึงข้อสอบไม่มีคุณภาพ ดังนั้นในทางปฏิบัติมักนิยมใช้ค่า a_i อยู่ในช่วง 0.5 ถึง 2.5 ถ้าข้อสอบที่มีค่า a_i สูงแสดงว่าโค้งลักษณะข้อสอบมีความชันมาก นั่นคือ ข้อสอบข้อนั้นสามารถจำแนกผู้สอบที่มีระดับความสามารถต่ำและสูงได้อย่างชัดเจน

ค่าการเดาของข้อสอบ (item pseudo-guessing; $\mathbf{c_i}$) หมายถึง ค่าโอกาสของผู้สอบที่มีความสามารถต่ำจะตอบข้อสอบได้ถูกต้อง ในทางทฤษฎีแล้ว c_i มีค่าอยู่ในช่วง 0 ถึง 1 แต่ในทางปฏิบัติค่า c_i มักอยู่ในช่วง 0 ถึง 0.4 โดยที่ข้อสอบที่มีค่า c_i ต่ำยิ่งดี ถ้า c_i มีค่าตั้งแต่ 0.3 ขึ้นไป แสดงว่าเป็นข้อสอบที่ไม่ดี แต่ถ้า c_i มีค่าเท่ากับ 0.2 หรือน้อยกว่าแสดงว่าเป็นข้อสอบที่ดี สำหรับข้อสอบที่มีค่า c_i เป็น 0 แสดงว่าผู้สอบตอบข้อสอบข้อนั้นโดยไม่มีการเดา ซึ่งเป็นข้อสอบที่ดีมาก ตามปกติแล้วเมื่อแบบสอบมี 4 ตัวเลือก ค่า c_i จะมีค่าประมาณ 1/4 หรือ 0.25 แต่ในทางทฤษฎี IRT ค่า c_i จะไม่เท่ากับ 0.25 เสมอไป เนื่องจากเชื่อว่าผู้สอบตอบข้อสอบโดยไม่มีการเดาแบบสุ่ม

ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบภายใต้ทฤษฎี IRT ค่าพารามิเตอร์ของข้อสอบที่วิเคราะห์มาจากผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบไม่สามารถนำมาเปรียบเทียบกันได้ ทั้งนี้เนื่องจากวิเคราะห์มาจากผู้สอบคนละกลุ่ม ตามข้อตกลงเบื้องต้นของทฤษฎี IRT ค่าพารามิเตอร์จะต้องไม่แปรเปลี่ยนไปตามกลุ่มผู้สอบ ดังนั้นในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบจะต้องนำค่าพารามิเตอร์ของข้อสอบมาปรับเทียบ (equating) สเกลให้อยู่บนเมทริกซ์เดียวกัน ซึ่งสามารถทำได้ 2 วิธีใหญ่ ๆ (Camilli and Shepard, 1994) คือ ***วิธีแบบสอบร่วม*** (*anchor test method*) และ***วิธีกลุ่มตัวอย่างแยก*** (*separate sample method*) นักวิธีวิทยาการวิจัยมักใช้วิธีแบบหลังปรับเทียบค่าพารามิเตอร์อำนาจจำแนกและพารามิเตอร์ความยาก โดยใช้การแปลงค่าเชิงเส้น (linear transformation) ดังนี้

$$a_{iF}^{*} = a_{iF} \,/\, A$$
$$b_{iF}^{*} = Ab_{iF} \,/\, A + B$$

เมื่อ a_{iF}^{*} แทน ค่าอำนาจจำแนกของข้อสอบข้อที่ i จากผู้สอบกลุ่มเปรียบเทียบ ซึ่งได้แปลงค่าแล้ว

b_{iF}^{*} แทน ค่าความยากของข้อสอบข้อที่ i จากผู้สอบกลุ่มเปรียบเทียบ ซึ่งได้แปลงค่าแล้ว

a_{iF} แทน ค่าอำนาจจำแนกของข้อสอบข้อที่ i จากผู้สอบกลุ่มเปรียบเทียบ ซึ่งยังไม่ได้แปลงค่า

b_{iF} แทน ค่าความยากของข้อสอบข้อที่ i จากผู้สอบกลุ่มเปรียบเทียบ ซึ่งยังไม่ได้แปลงค่า

A แทน ค่าความชัน (slope)

B แทน ค่าจุดตัด (intercept)

ในการปรับเทียบค่าพารามิเตอร์ของข้อสอบภายใต้ทฤษฎี IRT จะต้องคำนวณค่าความชัน (A) และค่าจุดตัด (B) ซึ่งนิยมใช้วิธีการคำนวณ 3 วิธี ดังต่อไปนี้

1. **วิธีผลรวมและค่าเฉลี่ยถ่วงน้ำหนัก** (weighted mean and sigma method; WMS) (Linn และคณะ, 1981 cited in Kim and Cohen, 1992a) วิธีนี้จะคำนวณค่าความชัน (A) และค่าจุดตัด (B) โดยใช้ค่าประมาณถ่วงน้ำหนักของค่าความยากของข้อสอบ ดังนี้

$$b_{iF}^{w*} = Ab_{iF}^{w} + B$$

เมื่อ b_{iF}^{w*} แทน ค่าความยากของข้อสอบถ่วงน้ำหนักข้อที่ i จากผู้สอบกลุ่มเปรียบเทียบหลังการปรับเทียบ

b_{iF}^{w} แทน ค่าความยากของข้อสอบถ่วงน้ำหนักข้อที่ i จากผู้สอบกลุ่มเปรียบเทียบก่อนการปรับเทียบ

สำหรับการแปลงค่าความชัน (A) และค่าจุดตัด (B) ใช้สมการดังนี้

$$A = \sigma_{b_R^w} / \sigma_{b_F^w}$$

$$B = \mu_{b_R^w} - A\mu_{b_F^w}$$

เมื่อ $\sigma_{b_R^w}$ แทน ส่วนเบี่ยงเบนมาตรฐานของข้อสอบถ่วงน้ำหนักข้อที่ i จากผู้สอบกลุ่มอ้างอิง

$\mu_{b_R^w}$ แทน ค่าเฉลี่ยของค่าความยากของข้อสอบถ่วงน้ำหนักข้อที่ i จากผู้สอบกลุ่มอ้างอิง

$\sigma_{b_F^w}$ แทน ส่วนเบี่ยงเบนมาตรฐานของข้อสอบถ่วงน้ำหนักข้อที่ i จากผู้สอบกลุ่มเปรียบเทียบ

$\mu_{b_F^w}$ แทน ค่าเฉลี่ยของค่าความยากของข้อสอบถ่วงน้ำหนักข้อที่ i จากผู้สอบกลุ่มเปรียบเทียบ

2. **วิธีโค้งลักษณะแบบสอบ** (test characteristic curve method; TCC) (Stocking and Lord, 1983 cited in Kim and Cohen, 1992a) วิธีนี้จะคำนวณความแตกต่างของค่าประมาณโค้งลักษณะแบบสอบในแต่ละกลุ่มตัวอย่าง โดยใช้ฟังก์ชัน F ดังนี้

$$F = \frac{1}{N}\sum_{j=1}^{N}(T_{jF} - T_{jF}^{*})^2$$

เมื่อ T_{jF} แทน คะแนนจริงของผู้สอบคนที่ j จากกลุ่มเปรียบเทียบ ซึ่งยังไม่ได้แปลงค่า

T_{jF}^{*} แทน คะแนนจริงของผู้สอบคนที่ j จากกลุ่มเปรียบเทียบ ซึ่งได้แปลงค่าแล้ว

N แทน จำนวนผู้สอบ

ภายใต้โมเดลโลจิสติกแบบ 2 พารามิเตอร์ ค่า T_{jF} , T^*_{jF} และฟังก์ชัน F สามารถเขียนใหม่ ดังนี้

$$T_{jF} = \sum_{i=1}^{n} P(\theta_{jF}, a_{iR}, b_{iR})$$

$$T^*_{jF} = \sum_{i=1}^{n} P(\theta_{jF}, a^*_{iF}, b^*_{iF})$$

$$F = \frac{1}{N} \sum_{j=1}^{N} \left[\sum_{i=1}^{n} P(\theta_{jF}, a_{iR}, b_{iR}) - \sum_{i=1}^{n} P(\theta_{jF}, a^*_{iF}, b^*_{iF}) \right]^2$$

เมื่อ θ_{jF} แทน ระดับความสามารถของผู้สอบคนที่ j ในกลุ่มเปรียบเทียบ

a_{iR} แทน ค่าอำนาจจำแนกของข้อสอบข้อที่ i จากผู้สอบกลุ่มอ้างอิง ซึ่งยังไม่ได้แปลงค่า

b_{iR} แทน ค่าความยากของข้อสอบข้อที่ i จากผู้สอบกลุ่มอ้างอิง ซึ่งยังไม่ได้แปลงค่า

a^*_{iF} แทน ค่าอำนาจจำแนกของข้อสอบข้อที่ i จากผู้สอบกลุ่มเปรียบเทียบ ซึ่งได้แปลงค่าแล้ว

b^*_{iF} แทน ค่าความยากของข้อสอบข้อที่ i จากผู้สอบกลุ่มเปรียบเทียบ ซึ่งได้แปลงค่าแล้ว

n แทน จำนวนข้อสอบ

ค่าความชัน (A) และค่าจุดตัด (B) เป็นค่าที่ได้จากการแปลงค่า T^*_{jF} ไปยัง T_{jF} ซึ่งทำให้ฟังก์ชัน F มีค่าน้อยที่สุด ทั้งนี้เนื่องจาก F เป็นฟังก์ชันของ A และ B ซึ่งจะมีค่าน้อยที่สุดเมื่ออนุพันธ์ย่อย (partial derivative) $\partial F/\partial A = 0$ และ $\partial F/\partial B = 0$

3. **วิธีไค-สแควร์ที่มีค่าน้อยที่สุด** (minimum chi-square method; MCS) (Divgi, 1985 cited in Kim and Cohen, 1992a) เป็นวิธีผสมระหว่างวิธี TCC กับเมทริกซ์ความแปรปรวน-ความแปรปรวนร่วมขนาด 2 × 2 ของความคลาดเคลื่อนแบบสุ่มในแต่ละข้อจากวิธีการประมาณค่าพารามิเตอร์ของข้อสอบ ค่าความชัน (A) และค่าจุดตัด (B) คำนวณได้จากฟังก์ชัน Q_i ดังนี้

$$Q_i = (a_{iR} - a^*_{iF}, b_{iR} - b^*_{iF})(\Sigma_{iR} + \Sigma^*_{iF})^{-1}(a_{iR} - a^*_{iF}, b_{iR} - b^*_{iF})'$$

เมื่อ $\sum_{iR}$ แทน ค่าของเมทริกซ์ความแปรปรวน-ความแปรปรวนร่วมของข้อสอบข้อที่ i จากผู้สอบกลุ่มอ้างอิง

$\sum_{iF}^{*}$ แทน ค่าของเมทริกซ์ความแปรปรวน-ความแปรปรวนร่วมของข้อสอบข้อที่ i จากผู้สอบกลุ่มเปรียบเทียบ ซึ่งแปลงค่ามาจากเมทริกซ์ $\sum_{iF}$

ถ้ากำหนดให้ฟังก์ชัน Q มีค่าดังนี้

$$Q = \sum_{i=1}^{n} Q_i$$

ค่าความชัน (A) และค่าจุดตัด (B) เป็นค่าที่น้อยที่สุดของฟังก์ชัน Q ทั้งนี้เพราะว่าอนุพันธ์ย่อย $\partial Q / \partial B = 0$ เป็นเส้นตรงที่ใช้ในการพิจารณาค่า B ซึ่งสามารถคำนวณจากค่า A โดยใช้สูตรดังนี้

$$B = \left[\sum_{i=1}^{n} S_{iab}(a_{iR} - a_{iF} / A) + S_{ibb}(b_{iR} - Ab_{iF}) \right] \Big/ \sum_{i=1}^{n} S_{ibb}$$

โดยที่ S_{iab} และ S_{ibb} เหมือนกับสมาชิกเฉพาะจากเมทริกซ์ $S_i = (\sum_{iR} + \sum_{iF}^{*})^{-1}$ ค่าของ A และ B จะใช้การคำนวณแบบทวนซ้ำ (iteration) จนกระทั่งได้ค่าความแตกต่างระหว่าง A และ B ตามเกณฑ์ที่กำหนดไว้

เมื่อปรับเทียบสเกลค่าพารามิเตอร์ของข้อสอบแล้ว จึงสามารถนำไปวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ จากนิยามภายใต้ทฤษฎี IRT กำหนดไว้ว่า ***"ข้อสอบทำหน้าที่ต่างกันเมื่อผู้สอบที่มีความสามารถระดับเดียวกัน แต่มาจากกลุ่มผู้สอบที่แตกต่างกันมีโอกาสของการตอบข้อสอบได้ถูกต้องไม่เท่ากัน"*** (Mazor and others, 1994) ดังนั้นในการตรวจสอบการทำหน้าที่ต่างกันจะต้องทดสอบสมมติฐาน ดังนี้

$$H_0 : P_R(\theta) - P_F(\theta) = 0$$
$$H_1 : P_R(\theta) - P_F(\theta) \neq 0$$

เมื่อ $P_R(\theta)$ แทน ฟังก์ชันการตอบสนองข้อสอบของผู้สอบกลุ่มอ้างอิง

$P_F(\theta)$ แทน ฟังก์ชันการตอบสนองข้อสอบของผู้สอบกลุ่มเปรียบเทียบ

ในการทดสอบสมมติฐานดังกล่าว จะเปรียบเทียบค่าประมาณฟังก์ชันการตอบสนองข้อสอบ (IRFs) หรือค่าประมาณโค้งลักษณะข้อสอบ (ICCs) ระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ สำหรับวิธีการคำนวณสามารถแบ่งออกเป็น 3 วิธีใหญ่ ๆ คือ วิธีการวัดพื้นที่ วิธีการเปรียบเทียบค่าพารามิเตอร์ของข้อสอบ และวิธีซิปเทสท์ ซึ่งมีรายละเอียด ดังนี้

2.1 **วิธีการวัดพื้นที่** (area measures; AREA)

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการวัดพื้นที่จะเปรียบเทียบผลการตอบข้อสอบโดยตรง โดยการคำนวณค่าประมาณพื้นที่ระหว่างโค้งลักษณะข้อสอบ จากผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่มีความสามารถระดับเดียวกัน ซึ่งมีสูตรทั่วไปที่ใช้ในการคำนวณพื้นที่ดังนี้ (Millsap and Everson, 1993)

$$A = f_s\left[P_R(\theta) - P_F(\theta)\right]$$

เมื่อ A แทน ดัชนีพื้นที่

f_S แทน ฟังก์ชัน f ในช่วง S ซึ่ง $S = (\theta_L, \theta_H)$ โดยที่ L และ H เป็นค่าต่ำสุดและสูงสุดตามลำดับ

$P_R(\theta)$ แทน โอกาสของการตอบข้อสอบถูกที่ระดับความสามารถ θ จากผู้สอบกลุ่มอ้างอิง

$P_F(\theta)$ แทน โอกาสของการตอบข้อสอบถูกที่ระดับความสามารถ θ จากผู้สอบกลุ่มเปรียบเทียบ

ในการคำนวณวัดพื้นที่ดังกล่าว สามารถเลือกคำนวณได้หลายลักษณะ เช่น (1) คำนวณพื้นที่ชนิดคิดเครื่องหมาย (signed area) หรือชนิดไม่คิดเครื่องหมาย (unsigned area) (2) คำนวณพื้นที่ในช่วงที่มีขอบเขตจำกัด (bounded) หรือขอบเขตไม่จำกัด (unbounded) (3) คำนวณพื้นที่โดยใช้การอินทิเกรทแบบต่อเนื่อง (continuous integration) หรือการประมาณค่าแบบไม่ต่อเนื่อง (discrete approximation) (4) คำนวณพื้นที่โดยถ่วงน้ำหนักแบบเท่ากัน (equally weighted) หรือถ่วงน้ำหนักแบบไม่เท่ากัน (differentially weighted) เป็นต้น การคำนวณพื้นที่ในอดีตมักจะใช้การประมาณค่าแบบไม่ต่อเนื่อง (Rudner, 1977; Rudner, Getson and Knight, 1980) ดังเช่น Rudner (1977) ได้เสนอสูตรดัชนีชนิดไม่คิดเครื่องหมายดังนี้

$$R = \sum_{j=-3}^{3} \left| D_j \right| \Delta$$

โดยที่ $D_j = P_R(\theta) - P_F(\theta)$ และ $\Delta = .005$ ดัชนีชนิดไม่คิดเครื่องหมาย R ดังกล่าว เมื่อนำค่าสัมบูรณ์ออกก็จะเปลี่ยนเป็นดัชนีชนิดคิดเครื่องหมาย ซึ่งอาจจะมีเครื่องหมายเป็นบวกหรือลบก็ได้ ต่อมา Linn และคณะ (1981) ได้เสนอดัชนีชนิดไม่คิดเครื่องหมายในรูปรากของกำลังสองของความแตกต่างเฉลี่ย (root mean squared different; RMSD) ระหว่าง IRFs โดยแบ่งช่วงระดับความสามารถตั้งแต่ –3 ถึง +3 ออกเป็น 600 ช่วง ดังนี้

$$RMSD = \sqrt{\frac{1}{600} \sum_{j=1}^{N} \left[P_R(\theta_j) - P_F(\theta_j) \right]^2}$$

โดยที่ θ_j แทนค่าความสามารถของผู้สอบคนที่ j นอกจากนี้ Shepard, Camilli และ Williams (1984) ได้เสนอดัชนีชนิดคิดเครื่องหมายและไม่คิดเครื่องหมายในรูปผลรวมของกำลังสอง (sum of squares; SOS) ซึ่งมี 4 รูปแบบ ดังนี้

$$SOS_1 = \frac{1}{N} \sum_{j=1}^{N} D_j^2$$

$$SOS_2 = \frac{1}{N} \sum_{j=1}^{N} D_j^2 / \sigma_{D_j}^2$$

$$SOS_3 = \frac{1}{N} \sum_{j=1}^{N} |D_j| (D_j)$$

$$SOS_4 = \frac{1}{N} \sum_{j=1}^{N} |D_j| (D_j) / \sigma_{D_j}^2$$

โดยที่ N เป็นผู้สอบกลุ่มเปรียบเทียบและกลุ่มอ้างอิงรวมกัน ดัชนีการวัดพื้นที่จากที่กล่าวมาใช้การประมาณค่าแบบไม่ต่อเนื่อง แต่ในปัจจุบันนี้มักนิยมวัดพื้นที่โดยใช้การอินทิเกรทแบบต่อเนื่อง (Raju, 1990; Kim and Cohen 1991, 1995; Cohen and Kim, 1993; Feinstein, 1995) ซึ่งสามารถคำนวณพื้นที่ชนิดคิดเครื่องหมายและไม่คิดเครื่องหมาย สำหรับวิธีการคำนวณที่นิยมกันมากมี 2 วิธี คือ วิธีการวัดพื้นที่ของ Raju (1990) และวิธีการวัดพื้นที่ของ Kim และ Cohen (1991) รายละเอียดของทั้งสองวิธีมีดังนี้

2.1.1 **วิธีการวัดพื้นที่ของ** Raju

Raju (1990) ได้เสนอสูตรการคำนวณพื้นที่ชนิดคิดเครื่องหมายในช่วงเปิด (open-interval signed area or exact signed area; ESA) และพื้นที่ชนิดไม่คิดเครื่องหมายในช่วงเปิด (open-interval unsigned area or exact unsigned area; EUA) เพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนแบบทวิภาค ภายใต้ทฤษฎี IRT โมเดลโลจิสติกแบบ 1, 2 และ 3 พารามิเตอร์ สูตรทั่วไปที่ใช้ในการคำนวณพื้นที่ทั้งสองชนิดมีลักษณะดังนี้

$$ESA_{kl} = \int_{-\infty}^{+\infty} \left[P_R(\theta) - P_F(\theta) \right] d\theta$$

$$EUA_{kl} = \int_{-\infty}^{+\infty} \left| P_R(\theta) - P_F(\theta) \right| d\theta$$

โดยที่ k แทน โมเดลการตอบสนองข้อสอบ

ถ้า $k = 1$ แสดงว่าเป็นโมเดลแบบ 1 พารามิเตอร์

ถ้า $k = 2$ แสดงว่าเป็นโมเดลแบบ 2 พารามิเตอร์

ถ้า $k = 3$ แสดงว่าเป็นโมเดลแบบ 3 พารามิเตอร์

และ l แทน พารามิเตอร์อำนาจจำแนกของข้อสอบข้อที่ i

ถ้า $l = 0$ แสดงว่าเป็นดัชนีชนิดคิดเครื่องหมาย โดยมี a_i แบบเท่ากันหรือไม่เท่ากัน

ถ้า $l = 1$ แสดงว่าเป็นดัชนีชนิดไม่คิดเครื่องหมาย โดยมี a_i แบบเท่ากัน

ถ้า $l = 2$ แสดงว่าเป็นดัชนีชนิดไม่คิดเครื่องหมาย โดยมี a_i แบบไม่เท่ากัน

ค่า IRFs ระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบภายใต้โมเดลโลจิสติกแบบ 3 พารามิเตอร์ สามารถคำนวณได้จากสูตร

$$P(\theta) = c + (1-c)P^*(\theta)$$

โดยที่ $P^*(\theta) = \left\{1 + \exp\left[-Da(\theta - b)\right]\right\}^{-1}$

a, b และ c เป็นพารามิเตอร์ที่แสดงคุณลักษณะของข้อสอบ และ D เป็นค่าคงที่โดยปกติกำหนดให้เท่ากับ 1.7 ถ้าพารามิเตอร์ c เท่ากับ 0 แล้ว IRF จะมีเพียง 2 พารามิเตอร์ คือ a และ b และถ้าพารามิเตอร์ a เท่ากับ 1 แล้วจะเหลือเพียงพารามิเตอร์ b เท่านั้น การคำนวณพื้นที่ชนิด ESA และ EUA ภายใต้โมเดลโลจิสติกแบบ 1, 2 และ 3 พารามิเตอร์ มีรายละเอียดดังนี้

(1) **โมเดลโลจิสติกแบบ 1 พารามิเตอร์** (1PLM)

การคำนวณพื้นที่ซนิด ESA และ EUA ระหว่าง IRFs ภายใต้โมเดลโลจิสติกแบบ 1 พารามิเตอร์ หรือโมเดล Rasch สามารถคำนวณได้จากสูตร

$$ESA_{10} = \hat{b}_F - \hat{b}_R$$
$$EUA_{11} = \left|\hat{b}_F - \hat{b}_R\right|$$

เนื่องจากค่าประมาณความยากของข้อสอบมีการแจกแจงลักษณะเชิงเส้นกำกับ (asymptotically) ดังนั้นค่าเฉลี่ยและค่าความแปรปรวนของ ESA_{10} สามารถคำนวณได้จากสูตร

$$\mu(ESA_{10}) = b_F - b_R$$
$$\sigma^2(ESA_{10}) = Var(\hat{b}_F) + Var(\hat{b}_R)$$

โดยที่ $Var(\hat{b}_i) = \left[\sum_{j=1}^{N} P_i(\theta_j)Q_i(\theta_j)\right]^{-1}$

และ $Q_i(\theta_j) = 1 - P_i(\theta_j)$

เมื่อ $Var(\hat{b})$ แทน ความแปรปรวนของค่าประมาณความยากของข้อสอบข้อที่ i

N แทน จำนวนผู้สอบในแต่ละกลุ่ม

$P_i(\theta_j)$ แทน โอกาสของผู้สอบคนที่ j ที่มีระดับความสามารถ θ จะตอบข้อสอบข้อที่ i ถูก

$Q_i(\theta_j)$ แทน โอกาสของผู้สอบคนที่ j ที่มีระดับความสามารถ θ จะตอบข้อสอบข้อที่ i ผิด

ถ้ากำหนดให้ $X = \hat{b}_F - \hat{b}_R$ และ X มีข้อตกลงการแจกแจงแบบปกติ ดังนั้นค่าเฉลี่ยและค่าความแปรปรวนของ EUA_{11} สามารถคำนวณได้จากสูตร

$$\mu(EUA_{11}) = \mu(ESA_{10})[1 - 2\Phi(z_0)] + (2/\pi)^{1/2}\sigma(ESA_{10})\exp(-z_0^2/2)$$
$$\sigma^2(EUA_{11}) = \sigma^2(ESA_{10}) + \mu^2(ESA_{10}) - \mu^2(EUA_{11})$$

โดยที่ $\Phi(z_0) = \int_{-\infty}^{z_0} g(z)dz$

$$z_0 = [0 - \mu(ESA_{10})]/\sigma(ESA_{10})$$

เมื่อ g(z) แทน ฟังก์ชันความหนาแน่นของความน่าจะเป็นแบบปกติ

(2) **โมเดลโลจิสติกแบบ 2 พารามิเตอร์** (2PLM)

การคำนวณพื้นที่ชนิด ESA และ EUA ระหว่าง IRFs ภายใต้โมเดลโลจิสติกแบบ 2 พารามิเตอร์ สามารถคำนวณได้จากสูตร

$$ESA_{20} = \hat{b}_F - \hat{b}_R$$

$$EUA_{21} = \left|\hat{b}_F - \hat{b}_R\right| \qquad \text{เมื่อ } \hat{a}_R = \hat{a}_F$$

$$EUA_{22} = \left|\frac{2(\hat{a}_F - \hat{a}_R)}{D\hat{a}_R a_F} \ln\left\{1 + \exp\left[\frac{D\hat{a}_R\hat{a}_F(\hat{b}_F - \hat{b}_R)}{\hat{a}_F - \hat{a}_R}\right] - (\hat{b}_F - \hat{b}_R)\right\}\right| \qquad \text{เมื่อ } \hat{a}_R \neq \hat{a}_F$$

เนื่องจากค่าประมาณความยากของข้อสอบมีการแจกแจงลักษณะเชิงเส้นกำกับ (asymptotically) ดังนั้นค่าเฉลี่ยและค่าความแปรปรวนของ ESA_{20} สามารถคำนวณได้จากสูตร

$$\mu(ESA_{20}) = b_F - b_R$$

$$\sigma^2(ESA_{20}) = Var(\hat{b}_F) + Var(\hat{b}_R)$$

โดยที่ $$Var(\hat{b}_i) = \frac{I_{a_i}}{I_{a_i} I_{b_i} - I^2_{a_i b_i}}$$

และ $$I_{a_i} = D^2 \sum_{j=1}^{N} (\theta_j - b_i)^2 P_i(\theta_j) Q_i(\theta_j)$$

$$I_{b_i} = D^2 a_i^2 \sum_{j=1}^{N} P_i(\theta_j) Q_i(\theta_j)$$

$$I_{a_i b_i} = D^2 a_i \sum_{j=1}^{N} (\theta_j - b_i)^2 P_i(\theta_j) Q_i(\theta_j)$$

เมื่อ I_{a_i} แทน ฟังก์ชันสารสนเทศของ a_i

I_{b_i} แทน ฟังก์ชันสารสนเทศของ b_i

$I_{a_i b_i}$ แทน ฟังก์ชันสารสนเทศของ $a_i b_i$

ถ้ากำหนดให้ $X = \hat{b}_F - \hat{b}_R$ และ X มีข้อตกลงการแจกแจงแบบปกติ ดังนั้นค่าเฉลี่ยและค่าความแปรปรวนของ EUA_{21} เมื่อ $\hat{a}_R = \hat{a}_F$ สามารถคำนวณได้จากสูตร

$$\mu(EUA_{21}) = \mu(ESA_{20})[1-2\Phi(z_0)] + (2/\pi)^{1/2}\sigma(ESA_{20})\exp(-z_0^2/2)$$
$$\sigma^2(EUA_{21}) = \sigma^2(ESA_{20}) + \mu^2(ESA_{20}) - \mu^2(EUA_{21})$$

โดยที่ $z_0 = [0-\mu(ESA_{20})]/\sigma(ESA_{20})$

ถ้ากำหนดให้ $H = \frac{2(\hat{a}_F - \hat{a}_R)}{D\hat{a}_R a_F}\ln\left\{1+\exp\left[\frac{D\hat{a}_R\hat{a}_F(\hat{b}_F-\hat{b}_R)}{\hat{a}_F-\hat{a}_R}\right]-(\hat{b}_F-\hat{b}_R)\right\}$

และ X มีข้อตกลงการแจกแจงแบบปกติ ดังนั้นค่าเฉลี่ยและค่าความแปรปรวนของ EUA_{22} เมื่อ $\hat{a}_R \neq \hat{a}_F$ สามารถคำนวณได้จากสูตร

$$\mu(EUA_{22}) = \mu(H)[1-2\Phi(z_0)] + (2/\pi)^{1/2}\sigma(H)\exp(-z_0^2/2)$$
$$\sigma^2(EUA_{22}) = \sigma^2(H) + \mu^2(H) - \mu^2(|H|)$$

โดยที่ $z_0 = [0-\mu(H)]/\sigma(H)$

$$\mu(H) = \frac{2(a_F-a_R)}{Da_Ra_F}\ln\left\{1+\exp\left[\frac{Da_Ra_F(b_F-b_R)}{a_F-a_R}\right]-(b_F-b_R)\right\}$$

$$\begin{aligned}\sigma^2(H) &= Var(H)\\ &= B_R^2Var(\hat{b}_R) + B_F^2Var(\hat{b}_F) + A_R^2Var(\hat{a}_R) + A_F^2Var(\hat{a}_F)\\ &\quad + 2B_RA_RCov(\hat{b}_R,\hat{a}_R) + 2B_FA_FCov(\hat{b}_F,\hat{a}_F)\end{aligned}$$

$$B_R = 1 - \frac{2\exp(Y)}{1+\exp(Y)}$$

$$B_F = -B_R$$

$$A_R = \frac{2}{a_R^2}\left\{\frac{a_Ra_F(b_R-b_F)}{a_R-a_F}\cdot\frac{\exp(Y)}{1+\exp(Y)} - \frac{\ln[1+\exp(Y)]}{D}\right\}$$

$$A_F = -\frac{a_R^2}{a_F^2}A_R$$

$$Y = \frac{Da_Ra_F(b_F-b_R)}{a_F-a_R}$$

$$Var(\hat{b}_i) = \frac{I_{a_i}}{I_{a_i} I_{b_i} - I^2_{a_i b_i}}$$

$$Var(\hat{a}_i) = \frac{I_{b_i}}{I_{a_i} I_{b_i} - I^2_{a_i b_i}}$$

$$Cov(\hat{a}_i, \hat{b}_i) = \frac{-I_{a_i b_i}}{I_{a_i} I_{b_i} - I^2_{a_i b_i}}$$

เมื่อ $Var(\hat{b})$ แทน ความแปรปรวนของ $\hat{b}$

$Var(\hat{a})$ แทน ความแปรปรวนของ $\hat{a}$

$Cov(\hat{a}, \hat{b})$ แทน ความแปรปรวนร่วมระหว่าง $\hat{a}$ และ $\hat{b}$

(3) **โมเดลโลจิสติกแบบ 3 พารามิเตอร์** (3PLM)

Raju (1988 cited in Kim and Cohen, 1991) ได้เสนอให้กำหนดพารามิเตอร์ c เท่ากัน ($c_R = c_F = c$) ในการคำนวณพื้นที่ซนิด ESA และ EUA ระหว่าง IRFs ภายใต้โมเดลโลจิสติกแบบ 3 พารามิเตอร์ ทั้งนี้เพราะว่าถ้าพารามิเตอร์ c ระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบมีค่าไม่เท่ากันแล้ว ค่าที่คำนวณได้จะไม่มีที่สิ้นสุด (infinite) ดังนั้นค่า ESA และ EUA ภายใต้โมเดลโลจิสติกแบบ 3 พารามิเตอร์เมื่อกำหนด c คงที่ สามารถคำนวณได้จากสูตร

$$ESA_{30} = (1-c)ESA_{20}$$
$$EUA_{31} = (1-c)EUA_{21} \quad \text{เมื่อ } \hat{a}_R = \hat{a}_F$$
$$EUA_{32} = (1-c)EUA_{22} \quad \text{เมื่อ } \hat{a}_R \neq \hat{a}_F$$

ค่าเฉลี่ยและค่าความแปรปรวนของ ESA_{30} EUA_{31} และ EUA_{32} เมื่อกำหนดค่า c คงที่ สามารถคำนวณโดยใช้สูตรจากโมเดลแบบ 2 พารามิเตอร์ ดังนี้

$$\mu(ESA_{30}) = (1-c)\mu(ESA_{20})$$
$$\sigma^2(ESA_{30}) = (1-c)^2\sigma^2(ESA_{20})$$
$$\mu(EUA_{31}) = (1-c)\mu(EUA_{21})$$
$$\sigma^2(EUA_{31}) = (1-c)^2\sigma^2(EUA_{21})$$
$$\mu(EUA_{32}) = (1-c)\mu(EUA_{22})$$
$$\sigma^2(EUA_{32}) = (1-c)^2\sigma^2(EUA_{22})$$

สำหรับการคำนวณค่าฟังก์ชันสารสนเทศ (I) ใช้สูตรดังนี้

$$I_{a_i} = \frac{D^2}{(1-c)^2}\sum_{j=1}^{N}\left\{(\theta_j - b_i)^2\left[P_i(\theta_j)-c\right]^2\frac{1-P_i(\theta_j)}{P_i(\theta_j)}\right\}$$

$$I_{b_i} = \frac{D^2 a_i^2}{(1-c)^2}\sum_{j=1}^{N}\left\{\left[P_i(\theta_j)-c\right]^2\frac{1-P_i(\theta_j)}{P_i(\theta_j)}\right\}$$

$$I_{a_i b_i} = \frac{D^2 a_i}{(1-c)^2}\sum_{j=1}^{N}\left\{(\theta_j - b_i)\left[P_i(\theta_j)-c\right]^2\frac{1-P_i(\theta_j)}{P_i(\theta_j)}\right\}$$

ในการตัดสินว่าข้อสอบที่นำมาตรวจสอบทำหน้าที่ต่างกันหรือไม่ Raju (1990) ได้เสนอให้นำดัชนีการวัดพื้นที่มาทดสอบนัยสำคัญโดยใช้สถิติ Z ภายใต้สมมติฐานการแจกแจงแบบปกติ ในการทดสอบดังกล่าวแบ่งออกเป็น 2 ชนิด ดังนี้

ชนิดที่ 1 การทดสอบนัยสำคัญของ ESA

นำดัชนี ESA ของข้อสอบข้อที่ i ไปทดสอบความแตกต่างกับ 0 โดยใช้สถิติทดสอบ Z ดังนี้

$$Z_i(EUA) = \frac{\hat{b}_{iF} - \hat{b}_{iR}}{\sqrt{Var(\hat{b}_{iF}) + Var(\hat{b}_{iR})}}$$

ชนิดที่ 2 การทดสอบนัยสำคัญของ EUA

ในการทดสอบนัยสำคัญของดัชนี EUA สามารถแบ่งออกเป็น 2 กรณีคือ ***กรณีที่ 1*** เมื่อ $\hat{a}_{iR} = \hat{a}_{iF}$ จะทดสอบเหมือนกับดัชนี ESA สำหรับ ***กรณีที่ 2*** $\hat{a}_{iR} \neq \hat{a}_{iF}$ จะนำดัชนี EUA ของข้อสอบข้อที่ i ไปทดสอบความแตกต่างกับ 0 โดยใช้สถิติทดสอบ Z ดังนี้

$$Z_i(H) = \frac{H_i}{\sqrt{Var(H_i)}}$$

2.1.2 การวัดพื้นที่ของ Kim และ Cohen

Kim และ Cohen (1991) ได้พัฒนาสูตรการคำนวณพื้นที่ชนิดคิดเครื่องหมายในช่วงปิด (closed-interval signed area; CSA) และพื้นที่ชนิดไม่คิดเครื่องหมายในช่วงปิด (closed-interval unsigned area; CUA) โดยคำนวณพื้นที่ระหว่าง IRFs บนช่วงความสามารถ $[\theta_1, \theta_2]$ ซึ่งมีสูตรในรูปทั่วไปดังนี้

$$CSA = \int_{\theta_1}^{\theta_2} \left[P_R(\theta) - P_F(\theta) \right] d\theta = S_R(\theta_1, \theta_2) - S_F(\theta_1, \theta_2)$$

$$CUA = \int_{\theta_1}^{\theta_2} \left| P_R(\theta) - P_F(\theta) \right| d\theta = \left| S_R(\theta_1, \theta_2) - S_F(\theta_1, \theta_2) \right|$$

สำหรับการคำนวณ CSA และ CUA ภายใต้โมเดลโลจิสติกแบบ 1, 2 และ 3 พารามิเตอร์ มีรายละเอียดดังนี้

(1) **โมเดลโลจิสติกแบบ 1 พารามิเตอร์** (1PLM)

ภายใต้โมเดลโลจิสติกแบบ 1 พารามิเตอร์จะกำหนดให้ $a_R = a_F = 1$ และ $c_R = c_F = 0$ ดังนั้นในโมเดลลักษณะนี้จึงไม่มี IRFs ตัดกันบนสเกลความสามารถ สำหรับการคำนวณพื้นที่ชนิด CSA และ CUA สามารถคำนวณได้จากสูตร

$$CSA_{1PLM} = S_R(\theta_1, \theta_2) - S_F(\theta_1, \theta_2)$$

$$= \ln\left\{ \frac{\left[1 + \exp(\theta_2 - b_R)\right]\left[1 + \exp(\theta_1 - b_F)\right]}{\left[1 + \exp(\theta_1 - b_R)\right]\left[1 + \exp(\theta_2 - b_F)\right]} \right\}$$

$$CUA = \left| CSA_{1PLM} \right|$$

(2) **โมเดลโลจิสติกแบบ 2 พารามิเตอร์** (2PLM)

โมเดลในลักษณะนี้จะกำหนดให้พารามิเตอร์ $c_R = c_F = 0$ การคำนวณพื้นที่ชนิด CSA และ CUA แบ่งออกเป็น 2 กรณี คือ กรณีที่ 1 เมื่อ $a_R = a_F = a$ และกรณีที่ 2 เมื่อ $a_R \neq a_F$ ในการคำนวณพื้นที่ CSA ทั้ง 2 กรณีจะใช้สูตรเหมือนกันดังนี้

$$CSA_{2PLM} = S_R(\theta_1, \theta_2) - S_F(\theta_1, \theta_2)$$

$$= \ln\left\{ \frac{\left\{1 + \exp\left[Da_R(\theta_2 - b_R)\right]\right\}^{\frac{1}{Da_R}} \left\{1 + \exp\left[Da_F(\theta_1 - b_F)\right]\right\}^{\frac{1}{Da_F}}}{\left\{1 + \exp\left[Da_R(\theta_1 - b_R)\right]\right\}^{\frac{1}{Da_R}} \left\{1 + \exp\left[Da_F(\theta_2 - b_F)\right]\right\}^{\frac{1}{Da_F}}} \right\}$$

รายละเอียดของการคำนวณพื้นที่ทั้ง 2 กรณีมีดังนี้

กรณีที่ 1 เมื่อ $a_R = a_F = a$ แสดงว่าไม่มีจุดตัดของ IRFs บนสเกลความสามารถ ดังนั้นพื้นที่ CSA และ CUA สามารถคำนวณโดยใช้สูตรในรูปอย่างง่ายดังนี้

$$CSA = (Da)^{-1} \ln\left\{\frac{\left\{1+\exp\left[Da(\theta_2 - b_R)\right]\right\}\left\{1+\exp\left[Da(\theta_1 - b_F)\right]\right\}}{\left\{1+\exp\left[Da(\theta_1 - b_R)\right]\right\}\left\{1+\exp\left[Da(\theta_2 - b_F)\right]\right\}}\right\}$$

$$CUA = |\, CSA \,|$$

กรณีที่ 2 เมื่อ $a_R \neq a_F$ แสดงว่ามีจุดตัดของ IRFs บนสเกลความสามารถ ดังนั้นจะต้องคำนวณหาจุดตัด (θ_x) โดยใช้สูตรดังนี้

$$\theta_x = \frac{a_R b_R - a_F b_F}{a_R - a_F}$$

ค่า θ_x ที่คำนวณได้มี 2 ลักษณะ ดังนี้

ลักษณะที่ 1 ถ้า θ_x มีค่าอยู่นอกช่วง $[\theta_1, \theta_2]$ พื้นที่ CUA คำนวณได้จากสูตร

$$CUA = |\, CSA_{2PLM} |$$

ลักษณะที่ 2 ถ้า θ_x มีค่าอยู่ในช่วง $[\theta_1, \theta_2]$ พื้นที่ CUA คำนวณได้จากสูตร

$$\begin{aligned} CUA &= \left|S_R(\theta_1,\theta_x) - S_F(\theta_1,\theta_x)\right| + \left|S_R(\theta_x,\theta_2) - S_F(\theta_x,\theta_2)\right| \\ &= \left|\ln\left\{\frac{\left\{1+\exp\left[Da_R(\theta_x - b_R)\right]\right\}^{\frac{1}{Da_R}}\left\{1+\exp\left[Da_F(\theta_1 - b_F)\right]\right\}^{\frac{1}{Da_F}}}{\left\{1+\exp\left[Da_R(\theta_1 - b_R)\right]\right\}^{\frac{1}{Da_R}}\left\{1+\exp\left[Da_F(\theta_x - b_F)\right]\right\}^{\frac{1}{Da_F}}}\right\}\right| \\ &+ \left|\ln\left\{\frac{\left\{1+\exp\left[Da_R(\theta_2 - b_R)\right]\right\}^{\frac{1}{Da_R}}\left\{1+\exp\left[Da_F(\theta_x - b_F)\right]\right\}^{\frac{1}{Da_F}}}{\left\{1+\exp\left[Da_R(\theta_x - b_R)\right]\right\}^{\frac{1}{Da_R}}\left\{1+\exp\left[Da_F(\theta_2 - b_F)\right]\right\}^{\frac{1}{Da_F}}}\right\}\right| \end{aligned}$$

(3) โมเดลโลจิสติกแบบ 3 พารามิเตอร์ (3PLM)

การคำนวณพื้นที่ชนิด CSA และ CUA ภายใต้โมเดลโลจิสติกแบบ 3 พารามิเตอร์ แบ่งออกเป็น 4 กรณี คือ กรณีที่ 1 เมื่อ $c_R = c_F = c$ และ $a_R = a_F = a$ กรณีที่ 2 $c_R = c_F = c$ และ $a_R \neq a_F$ กรณีที่ 3 เมื่อ $c_R \neq c_F$ และ $a_R = a_F = a$ กรณีที่ 4 เมื่อ $c_R \neq c_F$ และ $a_R \neq a_F$ ในการคำนวณพื้นที่ชนิด CSA ทั้ง 4 กรณีจะใช้สูตรเหมือนกันดังนี้

$$CSA_{3PLM} = S_R(\theta_1,\theta_2) - S_F(\theta_1,\theta_2)$$
$$= (c_R - c_F)(\theta_2 - \theta_1)$$
$$+ \ln\left\{\frac{\left\{1+\exp\left[Da_R(\theta_2 - b_R)\right]\right\}^{\frac{(1-c_R)}{Da_R}}\left\{1+\exp\left[Da_F(\theta_1 - b_F)\right]\right\}^{\frac{(1-c_F)}{Da_F}}}{\left\{1+\exp\left[Da_R(\theta_1 - b_R)\right]\right\}^{\frac{(1-c_R)}{Da_R}}\left\{1+\exp\left[Da_F(\theta_2 - b_F)\right]\right\}^{\frac{(1-c_F)}{Da_F}}}\right\}$$

รายละเอียดของการคำนวณพื้นที่ทั้ง 4 กรณีมีดังนี้

กรณีที่ 1 เมื่อ $c_R = c_F = c$ และ $a_R = a_F = a$ แสดงว่าไม่มีจุดตัดของ IRFs บนสเกลความสามารถ ดังนั้นพื้นที่ CSA และ CUA สามารถคำนวณโดยใช้สูตรในรูปอย่างง่ายดังนี้

$$CSA = (1-c)(Da)^{-1}\ln\left\{\frac{\left\{1+\exp\left[Da(\theta_2 - b_R)\right]\right\}\left\{1+\exp\left[Da(\theta_1 - b_F)\right]\right\}}{\left\{1+\exp\left[Da(\theta_1 - b_R)\right]\right\}\left\{1+\exp\left[Da(\theta_2 - b_F)\right]\right\}}\right\}$$

$$CUA = |\,CSA\,|$$

กรณีที่ 2 เมื่อ $c_R = c_F = c$ และ $a_R \neq a_F$ แสดงว่ามีจุดตัดของ IRFs บนสเกลความสามารถ ดังนั้นการคำนวณหาจุดตัด (θ_x) จะใช้สูตรเหมือนกับโมเดลแบบ 2 พารามิเตอร์ ค่า θ_x ที่คำนวณได้มี 2 ลักษณะ ดังนี้

ลักษณะที่ 1 ถ้า θ_x มีค่าอยู่นอกช่วง $[\theta_1,\theta_2]$ พื้นที่ CUA คำนวณจากสูตร

$$CUA = |\,CSA_{3PLM}\,|$$

ลักษณะที่ 2 ถ้าค่า θ_x มีค่าอยู่ในช่วง $[\theta_1,\theta_2]$ พื้นที่ CUA คำนวณจากสูตร

$$CUA = \left|S_R(\theta_1,\theta_x) - S_F(\theta_1,\theta_x)\right| + \left|S_R(\theta_x,\theta_2) - S_F(\theta_x,\theta_2)\right|$$
$$= (1-c)\left|\ln\left\{\frac{\left\{1+\exp\left[Da_R(\theta_x - b_R)\right]\right\}^{\frac{1}{Da_R}}\left\{1+\exp\left[Da_F(\theta_1 - b_F)\right]\right\}^{\frac{1}{Da_F}}}{\left\{1+\exp\left[Da_R(\theta_1 - b_R)\right]\right\}^{\frac{1}{Da_R}}\left\{1+\exp\left[Da_F(\theta_x - b_F)\right]\right\}^{\frac{1}{Da_F}}}\right\}\right|$$
$$+(1-c)\left|\ln\left\{\frac{\left\{1+\exp\left[Da_R(\theta_2 - b_R)\right]\right\}^{\frac{1}{Da_R}}\left\{1+\exp\left[Da_F(\theta_x - b_F)\right]\right\}^{\frac{1}{Da_F}}}{\left\{1+\exp\left[Da_R(\theta_x - b_R)\right]\right\}^{\frac{1}{Da_R}}\left\{1+\exp\left[Da_F(\theta_2 - b_F)\right]\right\}^{\frac{1}{Da_F}}}\right\}\right|$$

กรณีที่ 3 เมื่อ $c_R \neq c_F$ และ $a_R = a_F = a$ แสดงว่ามีจุดตัดของ IRFs บนสเกลความสามารถ ดังนั้นจะต้องคำนวณหาจุดตัด (θ_x) โดยใช้สูตรดังนี้

$$\theta_x = (a)^{-1} \ln\left\{\frac{c_R - c_F}{\left[(1-c_R)/\exp(ab_F)\right] - \left[(1-c_F)/\exp(ab_R)\right]}\right\}$$

ค่า θ_x ที่คำนวณได้มี 3 ลักษณะ ดังนี้

ลักษณะที่ 1 ถ้า θ_x มีค่าอยู่นอกช่วง $[\theta_1, \theta_2]$ พื้นที่ CUA คำนวณได้จากสูตร $CUA = |\, CSA_{3PLM}|$

ลักษณะที่ 2 ถ้า θ_x ไม่สามารถคำนวณได้ ทั้งนี้อาจเนื่องมาจาก $c_R > c_F$ และ $b_R \leq b_F$; หรือ $c_R < c_F$ และ $b_R \geq b_F$ ซึ่งจะมีผลทำให้

$$\frac{c_R - c_F}{\left[(1-c_R)/\exp(ab_F)\right] - \left[(1-c_F)/\exp(ab_R)\right]} \leq 0$$

ดังนั้นจึงไม่มีจุดตัดที่แน่นอนบนสเกลความสามารถ การคำนวณพื้นที่ CUA ในสถานการณ์พิเศษดังกล่าวจะใช้สูตรดังนี้

$CUA = |\, CSA_{3PLM}|$

ลักษณะที่ 3 ถ้า θ_x มีค่าอยู่ในช่วง $[\theta_1, \theta_2]$ ค่า CUA สามารถคำนวณได้จากสูตรดังนี้

$$\begin{aligned}
CUA &= \left|S_R(\theta_1,\theta_x) - S_F(\theta_1,\theta_x)\right| + \left|S_R(\theta_x,\theta_2) - S_F(\theta_x,\theta_2)\right| \\
&= \left|(c_R - c_F)(\theta_x - \theta_1)\right. \\
&\quad \left. + \ln\left\{\frac{\left\{1+\exp\left[Da(\theta_x - b_R)\right]\right\}^{\frac{(1-c_R)}{Da}}\left\{1+\exp\left[Da(\theta_1 - b_F)\right]\right\}^{\frac{(1-c_F)}{Da}}}{\left\{1+\exp\left[Da(\theta_1 - b_R)\right]\right\}^{\frac{(1-c_R)}{Da}}\left\{1+\exp\left[Da(\theta_x - b_F)\right]\right\}^{\frac{(1-c_F)}{Da}}}\right\}\right| \\
&\quad + \left|(c_R - c_F)(\theta_2 - \theta_x)\right. \\
&\quad \left. + \ln\left\{\frac{\left\{1+\exp\left[Da(\theta_2 - b_R)\right]\right\}^{\frac{(1-c_R)}{Da}}\left\{1+\exp\left[Da(\theta_x - b_F)\right]\right\}^{\frac{(1-c_F)}{Da}}}{\left\{1+\exp\left[Da(\theta_x - b_R)\right]\right\}^{\frac{(1-c_R)}{Da}}\left\{1+\exp\left[Da(\theta_2 - b_F)\right]\right\}^{\frac{(1-c_F)}{Da}}}\right\}\right|
\end{aligned}$$

กรณีที่ 4 เมื่อ $c_R \neq c_F$ และ $a_R \neq a_F$ ในกรณีนี้จะใช้วิธีการประมาณค่าของ Newton-Raphson โดยคำนวณหาจุดตัดของ IRFs บนสเกลความสามารถ ซึ่งอาจมี IRFs ตัดกัน 0, 1 หรือ 2 จุด สำหรับวิธี Newton-Raphson มีขั้นตอนดังนี้

กำหนดให้ $f(\theta) = P_R(\theta) - P_F(\theta)$

ดังนั้น $f'(\theta) = \dfrac{df(\theta)}{d\theta}$

$= (1-c_R)Da_R P_R^*(1-P_R^*) - (1-c_F)Da_F P_F^*(1-P_F^*)$

โดยที่ $P^*(\theta) = \{1+\exp[-Da(\theta-b)]\}^{-1}$

เมื่อให้ θ_0 เป็นค่าเริ่มต้น จุดตัด θ_x ของ IRFs สามารถคำนวณได้จากสูตร

$$\theta_{n+1} = \theta_n - \frac{f(\theta_n)}{f'(\theta_n)}$$

โดยที่ $f'(\theta_n)$ เป็นค่าอนุพันธ์ของ $f(\theta)$ ที่ระดับความสามารถ θ_n

ค่า θ_x ที่คำนวณได้มี 3 ลักษณะ คือ

ลักษณะที่ 1 ถ้า θ_x มีค่าอยู่นอกช่วง $[\theta_1, \theta_2]$ ค่า CUA สามารถคำนวณได้จากสูตร

$CUA = |\, CSA_{3PLM}|$

ลักษณะที่ 2 ถ้า θ_x มีค่าอยู่ในช่วง $[\theta_1, \theta_2]$ และ IRFs ตัดกันเพียงจุดเดียวเท่านั้น ค่า CUA สามารถคำนวณได้จากสูตร

$$CUA = \left|S_R(\theta_1,\theta_x) - S_F(\theta_1,\theta_x)\right| + \left|S_R(\theta_x,\theta_2) - S_F(\theta_x,\theta_2)\right|$$

$$= \left|(c_R - c_F)(\theta_x - \theta_1) + \ln\left\{\frac{\left\{1+\exp\left[Da_R(\theta_x - b_R)\right]\right\}^{\frac{(1-c_R)}{Da_R}}\left\{1+\exp\left[Da_F(\theta_1 - b_F)\right]\right\}^{\frac{(1-c_F)}{Da_F}}}{\left\{1+\exp\left[Da_R(\theta_1 - b_R)\right]\right\}^{\frac{(1-c_R)}{Da_R}}\left\{1+\exp\left[Da_F(\theta_x - b_F)\right]\right\}^{\frac{(1-c_F)}{Da_F}}}\right\}\right|$$

$$+ \left|(c_R - c_F)(\theta_2 - \theta_x) + \ln\left\{\frac{\left\{1+\exp\left[Da_R(\theta_2 - b_R)\right]\right\}^{\frac{(1-c_R)}{Da_R}}\left\{1+\exp\left[Da_F(\theta_x - b_F)\right]\right\}^{\frac{(1-c_F)}{Da_F}}}{\left\{1+\exp\left[Da_R(\theta_x - b_R)\right]\right\}^{\frac{(1-c_R)}{Da_R}}\left\{1+\exp\left[Da_F(\theta_2 - b_F)\right]\right\}^{\frac{(1-c_F)}{Da_F}}}\right\}\right|$$

ลักษณะที่ 3 ถ้า θ_x ที่คำนวณได้มีค่าอยู่ในช่วง $[\theta_1, \theta_2]$ และ IRFs ตัดกัน 2 จุดที่ θ_{x_1} และ $\theta_{x_2}(\theta_{x_1} < \theta_{x_2})$ ค่า CUA สามารถคำนวณได้จากสูตร

$$
\begin{aligned}
CUA &= \left|S_R(\theta_1,\theta_{x_1}) - S_F(\theta_1,\theta_{x_1})\right| + \left|S_R(\theta_{x_1},\theta_{x_2}) - S_F(\theta_{x_1},\theta_{x_2})\right| \\
&\quad + \left|S_R(\theta_{x_2},\theta_2) - S_F(\theta_{x_2},\theta_2)\right| \\
&= \left|(c_R - c_F)(\theta_{x_1} - \theta_1)\right. \\
&\quad \left. + \ln\left\{\frac{\left\{1+\exp\left[Da_R(\theta_{x_1} - b_R)\right]\right\}^{\frac{(1-c_R)}{Da_R}}\left\{1+\exp\left[Da_F(\theta_1 - b_F)\right]\right\}^{\frac{(1-c_F)}{Da_F}}}{\left\{1+\exp\left[Da_R(\theta_1 - b_R)\right]\right\}^{\frac{(1-c_R)}{Da_R}}\left\{1+\exp\left[Da_F(\theta_{x_1} - b_F)\right]\right\}^{\frac{(1-c_F)}{Da_F}}}\right\}\right| \\
&\quad + \left|(c_R - c_F)(\theta_{x_2} - \theta_{x_1})\right. \\
&\quad \left. + \ln\left\{\frac{\left\{1+\exp\left[Da_R(\theta_{x_2} - b_R)\right]\right\}^{\frac{(1-c_R)}{Da_R}}\left\{1+\exp\left[Da_F(\theta_{x_1} - b_F)\right]\right\}^{\frac{(1-c_F)}{Da_F}}}{\left\{1+\exp\left[Da_R(\theta_{x_1} - b_R)\right]\right\}^{\frac{(1-c_R)}{Da_R}}\left\{1+\exp\left[Da_F(\theta_{x_2} - b_F)\right]\right\}^{\frac{(1-c_F)}{Da_F}}}\right\}\right| \\
&\quad + \left|(c_R - c_F)(\theta_2 - \theta_{x_2})\right. \\
&\quad \left. + \ln\left\{\frac{\left\{1+\exp\left[Da_R(\theta_2 - b_R)\right]\right\}^{\frac{(1-c_R)}{Da_R}}\left\{1+\exp\left[Da_F(\theta_{x_2} - b_F)\right]\right\}^{\frac{(1-c_F)}{Da_F}}}{\left\{1+\exp\left[Da_R(\theta_2 - b_R)\right]\right\}^{\frac{(1-c_R)}{Da_R}}\left\{1+\exp\left[Da_F(\theta_{x_2} - b_F)\right]\right\}^{\frac{(1-c_F)}{Da_F}}}\right\}\right|
\end{aligned}
$$

สำหรับการตัดสินการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการวัดพื้นที่ของ Kim และ Cohen (1991) จะไม่ใช้การทดสอบนัยสำคัญ แต่จะนำขนาดของพื้นที่ไปเปรียบเทียบกับเกณฑ์ที่กำหนดขึ้น

2.2 วิธีการเปรียบเทียบค่าพารามิเตอร์ของข้อสอบ

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีเปรียบเทียบค่าพารามิเตอร์ของข้อสอบไม่ได้นำค่า IRFs มาเปรียบเทียบโดยตรงเหมือนกับวิธีวัดพื้นที่ แต่เป็นการเปรียบเทียบทางอ้อมโดยนำค่าความยาก ค่าอำนาจจำแนก หรือค่าการเดาของข้อสอบ จากผู้สอบกลุ่มย่อยสองกลุ่มมาเปรียบเทียบแล้วใช้สถิติทดสอบนัยสำคัญ การตรวจสอบด้วยวิธีดังกล่าวแบ่งออกเป็นวิธีย่อย ๆ ดังนี้

2.2.1 **วิธีเปลี่ยนค่าความยาก** (difficulty shift)

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีเปลี่ยนค่าความยากจะใช้สถิติ Z_i ทดสอบการเท่ากันเฉพาะค่าความยากของข้อสอบ ระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ภายใต้โมเดลโลจิสติกแบบ 1 พารามิเตอร์ หรือโมเดลราชส์ (Rasch model) ซึ่งกำหนดให้พารามิเตอร์การเดา (c_i) เท่ากับ 0 และพารามิเตอร์อำนาจจำแนก (a_i) เท่ากันทุกข้อ ส่วนพารามิเตอร์ความยาก (b_i) จะแปรเปลี่ยนไปตามกลุ่มผู้สอบ สำหรับสมมติฐานศูนย์ (H_0) และสมมติฐานอื่น (H_1) ที่ใช้ในการทดสอบกำหนดดังนี้

$$H_0 : b_{iF} = b_{iR}$$
$$H_1 : b_{iF} \neq b_{iR}$$

หลักการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธีดังกล่าวจะประมาณค่าพารามิเตอร์ความยากจากผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบแล้วแปลงพารามิเตอร์ความยากให้เป็นค่ามาตรฐาน ต่อจากนั้นจะทดสอบความแตกต่างของพารามิเตอร์ความยากโดยใช้สถิติ Z_i ที่เสนอโดย Wright และคณะ (1976 cited in Hulin, Drasgow and Parson, 1983) ดังนี้

$$Z_i = \frac{\hat{b}_{iF} - \hat{b}_{iR}}{\sqrt{(SE\ of\ \hat{b}_{iF})^2 + (SE\ of\ \hat{b}_{iR})^2}}$$

โดยที่ $SE\ of\ \hat{b}_{ig} = \sqrt{I_{a_i} / (I_{a_i} I_{b_i} - I^2_{a_i b_i})}$

เมื่อ $SE\ of\ \hat{b}_{iF}$ แทน ค่าความคลาดเคลื่อนมาตรฐานของ $\hat{b}_{iF}$

$SE\ of\ \hat{b}_{iR}$ แทน ค่าความคลาดเคลื่อนมาตรฐานของ $\hat{b}_{iR}$

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

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

I_{b_i} แทน ฟังก์ชันสารสนเทศของ b_i

I_{a_i} แทน ฟังก์ชันสารสนเทศของ a_i

$I_{a_i b_i}$ แทน ฟังก์ชันสารสนเทศของ $a_i b_i$

ภายใต้สมมติฐานศูนย์ (H_0) ของการทำหน้าที่ต่างกันของข้อสอบ สถิติ Z_i มีการแจกแจงปกติแบบ asymptotically ซึ่งมีค่าเฉลี่ยเท่ากับ 0 และความแปรปรวนเท่ากับ 1 ถ้าผลการทดสอบปรากฏว่ามีนัยสำคัญ แสดงว่าข้อสอบเปลี่ยนค่าความยาก (difficulty shift) จากผู้สอบ

กลุ่มหนึ่งไปยังผู้สอบอีกกลุ่มหนึ่ง หรือกล่าวอีกนัยหนึ่งว่าข้อสอบมีความยากสัมพัทธ์ในผู้สอบกลุ่มหนึ่งมากกว่าผู้สอบในกลุ่มอื่น นั่นคือข้อสอบที่นำมาตรวจสอบดังกล่าวทำหน้าที่ต่างกัน นอกจากนี้ Wright และ Stone (1979 cited in Hulin, Drasgow and Parson, 1983) ได้เสนอวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบภายใต้โมเดลราชส์ โดยใช้สถิติ H_i ทดสอบความเหมาะสมของโมเดล (fit model) กับข้อมูลการตอบข้อสอบจากผู้สอบกลุ่มย่อยสองกลุ่ม สถิติดังกล่าวมีลักษณะดังนี้

$$H_i = \sum_{j=1}^{N}\left[\frac{\mathrm{U}_{ij} - \hat{P}_i(\hat{\theta}_j)}{\hat{P}_i(\hat{\theta}_j)\hat{Q}_i(\hat{\theta}_j)}\right]^2$$

โดยที่ $\hat{Q}_i(\hat{\theta}_j) = 1 - \hat{P}_i(\hat{\theta}_j)$

เมื่อ H_i แทน สถิติทดสอบความเหมาะสมของโมเดล

N แทน จำนวนผู้สอบทั้งหมด

U_{ij} แทน คะแนนของผลการตอบข้อสอบข้อที่ i ของผู้สอบคนที่ j (ตอบผิดได้คะแนนเท่ากับ 0 และตอบถูกได้คะแนนเท่ากับ 1)

$\hat{P}_i(\hat{\theta}_j)$ แทน ค่าประมาณโอกาสของผู้สอบคนที่ j ที่มีระดับความสามารถ θ จะตอบข้อสอบข้อที่ i ถูก

$\hat{Q}_i(\hat{\theta}_j)$ แทน ค่าประมาณโอกาสของผู้สอบคนที่ j ที่มีระดับความสามารถ θ จะตอบข้อสอบข้อที่ i ผิด

การประมาณค่าสถิติ H_i ดังกล่าวมีการแจกแจงแบบไค-สแควร์ และมีระดับของความเป็นอิสระเท่ากับ $N - 1$ ถ้าผลการทดสอบทางสถิติมีนัยสำคัญ แสดงว่าผลการตอบข้อสอบจากผู้สอบกลุ่มหนึ่งมีความเหมาะสมกับโมเดลมากกว่าผู้สอบอีกกลุ่มหนึ่ง นั่นคือ ข้อสอบทำหน้าที่ต่างกัน

2.2.3 **วิธีการทดสอบ** F

Hulin, Drasgow และ Komocar (1982) ได้พัฒนาวิธีการทดสอบทางอ้อมเพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยการนำทฤษฎี IRT มาประยุกต์ใช้กับการวิเคราะห์ข้อมูลทางเจตคติ และใช้โมเดลโลจิสติกแบบ 2 พารามิเตอร์ ซึ่งกำหนดให้พารามิเตอร์การเดา (c_i)

เท่ากับ 0 ส่วนพารามิเตอร์ความยาก (b_i) และพารามิเตอร์อำนาจจำแนก (a_i) จะแปรเปลี่ยนไปตามผู้สอบกลุ่มย่อย ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบจะเริ่มต้นด้วยการประมาณค่าพารามิเตอร์ความยาก (b_i) และพารามิเตอร์อำนาจจำแนก (a_i) จากผู้สอบกลุ่มย่อย 2 กลุ่ม แล้วแปลงพารามิเตอร์ความยาก (b_i) ให้เป็นค่ามาตรฐาน ต่อจากนั้นจึงแบ่งช่วงความสามารถของผู้สอบในแต่ละกลุ่มย่อยเพื่อคำนวณค่าสัดส่วนที่สังเกตได้ของผู้สอบ ซึ่งตอบข้อสอบได้ถูกต้องในแต่ละช่วงความสามารถ ขั้นตอนต่อไปจะนำค่าสัดส่วนมาเขียนกราฟบนสเกลความสามารถ โดยลงจุดตรงกึ่งกลางของค่าความสามารถในแต่ละช่วงที่แบ่งไว้ กราฟที่ได้จะเป็นโค้งลักษณะข้อสอบ (item characteristic curve; ICC) ที่ประมาณค่าได้จากข้อมูลเชิงประจักษ์ สำหรับขั้นตอนสุดท้ายจะเปรียบเทียบโค้งลักษณะข้อสอบระหว่างผู้สอบกลุ่มย่อยสองกลุ่ม โดยแปลงค่าสัดส่วนให้อยู่ในรูปโลจิท (logit) ดังนี้

$$L\left[P_i(\theta)\right] = \log\left[\frac{P_i(\theta)}{1-P_i(\theta)}\right]$$
$$= Da_i(\theta - b_i)$$

เมื่อ $L\left[P_i(\theta)\right]$ แทน ฟังก์ชันเชิงเส้นของค่าความสามารถ θ

$P_i(\theta)$ แทน ICC ของข้อมูลเชิงประจักษ์ หรือ สัดส่วนที่สังเกตได้ของผู้สอบที่ระดับความสามารถ θ ตอบข้อสอบข้อที่ i ได้ถูกต้อง

D แทน ค่าคงที่ปกติกำหนดให้มีค่าเท่ากับ 1.7

b_i แทน พารามิเตอร์ความยากของข้อสอบข้อที่ i

a_i แทน พารามิเตอร์อำนาจจำแนกของข้อสอบข้อที่ i

หลังจากแปลงค่าสัดส่วนเชิงประจักษ์ให้อยู่ในรูปโลจิท ค่าสัดส่วนดังกล่าวจะอยู่ในรูปฟังก์ชันเชิงเส้น ซึ่งฟังก์ชันเชิงเส้นนี้ถือว่าเป็นฟังก์ชันการถดถอย (regression function) ด้วยโดยที่จุดตัดและความชันของเส้นการถดถอยก็คือพารามิเตอร์ความยากและพารามิเตอร์อำนาจจำแนกตามลำดับ สำหรับการพิจารณาการทำหน้าที่ต่างกันของข้อสอบ จะทดสอบการเท่ากันของฟังก์ชันการถดถอย โดยใช้การทดสอบอัตราส่วน F ดังนี้

$$F = \frac{SSE(pooled) - [SSE(A)] + [SSE(B)]}{[SSE(A) + SSE(B)]} \cdot \frac{J_A + J_B - 4}{2}$$

เมื่อ $SSE(pooled)$ แทน ผลรวมของความคลาดเคลื่อนกำลังสองของเส้นการถดถอยรวม

$SSE(A)$ แทน ผลรวมของความคลาดเคลื่อนกำลังสองจากผู้สอบกลุ่ม A

$SSE(B)$ แทน ผลรวมของความคลาดเคลื่อนกำลังสองจากผู้สอบกลุ่ม B

J_A แทน จำนวนสัดส่วนจากผู้สอบกลุ่ม A

J_B แทน จำนวนสัดส่วนจากผู้สอบกลุ่ม B

การทดสอบอัตราส่วน F มีระดับของความเป็นอิสระเท่ากับ 2 และ $J_A + J_B - 4$ ถ้าผลการทดสอบอัตราส่วน F ปรากฏว่ามีนัยสำคัญทางสถิติ แสดงว่าเส้นการถดถอยที่ประมาณค่าจากผู้สอบกลุ่มย่อยสองกลุ่มมีค่าเท่ากัน นั่นคือ ข้อสอบทำหน้าที่ต่างกัน

2.2.4 วิธีการทดสอบไค-สแควร์ของ Lord (Lord's chi-squre test)

Lord (1980) ได้เสนอวิธีการทดสอบไค-สแควร์เพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบภายใต้ทฤษฎี IRT โดยใช้โมเดลโลจิสติกแบบ 3 พารามิเตอร์ หลักการตรวจสอบด้วยวิธีนี้จะทดสอบความแตกต่างของค่าพารามิเตอร์ของข้อสอบระหว่างฟังก์ชันการตอบสนองข้อสอบ (IRFs) จากผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบจะเริ่มต้นด้วยการประมาณค่าพารามิเตอร์ a_i, b_i และ c_i จากผู้สอบทั้งสองกลุ่มรวมกัน แล้วแปลงค่าพารามิเตอร์ b_i ให้เป็นค่ามาตรฐาน ต่อจากนั้นจะประมาณค่าพารามิเตอร์ a_i และ b_i อีกครั้งหนึ่ง โดยประมาณค่าจากผู้สอบแต่ละกลุ่ม ส่วนพารามิเตอร์ c_i ยังใช้ค่าเดิมที่ประมาณค่าได้ในครั้งแรก ข้อสอบที่มีค่า c_i ต่ำ และข้อสอบที่มีค่า b_i สูงหรือต่ำมาก ๆ จะถูกคัดออกไป แล้วแปลงพารามิเตอร์ b_i ที่ได้ใหม่ให้เป็นค่ามาตรฐานอีกครั้งหนึ่ง ต่อจากนั้นจึงนำค่าพารามิเตอร์ a_i และ b_i ของข้อสอบแต่ละข้อไปเปรียบเทียบความแตกต่างโดยใช้สถิติไค-สแควร์ ซึ่งมีระดับชั้นของความเป็นอิสระเท่ากับ 2 สำหรับพารามิเตอร์ c_i ไม่ได้นำไปทดสอบด้วย ดังนั้นในการทดสอบสมมติฐานศูนย์ (H_0) และสมมติฐานอื่น (H_1) กำหนดดังนี้

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

สำหรับสถิติไค-สแควร์ (χ_i^2) มีลักษณะดังนี้

$$\chi_i^2 = V_i' \Gamma_i^{-1} V_i$$

โดยที่ $V_i' = (\hat{a}_{iF} - \hat{a}_{iR}, \hat{b}_{iF} - \hat{b}_{iR})$

$\Gamma_i = \Gamma_{iF} + \Gamma_{iR}$

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

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

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

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

V_i' แทน เวคเตอร์ความแตกต่างของค่าประมาณพารามิเตอร์ของข้อสอบข้อที่ i จากผู้สอบกลุ่มเปรียบเทียบและกลุ่มอ้างอิง

Γ_i แทน เมทริกซ์ความแปรปรวน - ความแปรปรวนร่วมของ $(\hat{a}_{iF} - \hat{a}_{iR})$ และ $(\hat{b}_{iF} - \hat{b}_{iR})$ ที่มีขนาด 2×2

Γ_i^{-1} แทน อินเวอร์สของเมทริกซ์ Γ_i

ถ้าผลทดสอบสมมติฐานแล้วปรากฏว่ามีค่าแตกต่างกันอย่างมีนัยสำคัญ แสดงว่าข้อสอบดังกล่าวทำหน้าที่ต่างกัน สำหรับสถิติไค-สแควร์จะมีประสิทธิภาพในการตรวจสอบเมื่ออยู่บนพื้นฐานของสมมติฐาน 3 ประการ คือ (1) การทดสอบทางสถิติมีการแจกแจงไค-สแควร์แบบเชิงเส้นกำกับ (asymptotic) (2) ต้องรู้ค่าความสามารถของผู้สอบ และ (3) ใช้วิธีการประมาณค่าแบบไลค์ลิฮู้ดสูงสุด (maximum likelihood estimate) นอกจากนี้วิธีการทดสอบไค-สแควร์ของ Lord (1980) ยังสามารถนำไปประยุกต์ใช้กับโมเดลโลจิสติกแบบ 2 พารามิเตอร์ โดยกำหนดให้พารามิเตอร์การเดา (c_i) มีค่าเป็นศูนย์ ทั้งนี้เพราะว่าในการเปรียบเทียบค่าพารามิเตอร์ของข้อสอบจะทดสอบความแตกต่างเฉพาะพารามิเตอร์ความยาก (b_i) และอำนาจจำแนก (a_i) เท่านั้น

2.2.5 **วิธีการตอบสนองข้อสอบแบบเทียม** (pseudo-IRT)

ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบภายใต้ทฤษฎี IRT โดยใช้โมเดลแบบ 3 พารามิเตอร์ต้องใช้กลุ่มตัวอย่างที่มีขนาดใหญ่ Linn และ Harnisch (1981) จึงได้พัฒนาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามโมเดลดังกล่าว เพื่อให้สามารถวิเคราะห์ได้กับ

กลุ่มตัวอย่างที่มีขนาดเล็ก ซึ่งจะช่วยในการประหยัดค่าใช้จ่าย การตรวจสอบในลักษณะนี้เรียกว่า "วิธีการตอบสนองข้อสอบแบบเทียม" ในการวิเคราะห์จะเริ่มต้นด้วยการประมาณค่าพารามิเตอร์ของข้อสอบและพารามิเตอร์ความสามารถจากผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบรวมกัน โดยใช้โมเดล 3 พารามิเตอร์ แล้วเลือกค่าความสามารถที่ประมาณค่าได้จากผู้สอบกลุ่มเปรียบเทียบ ต่อจากนั้นจะแบ่งค่าความสามารถดังกล่าวออกเป็น 5 ช่วง เพื่อเปรียบเทียบผลการตอบข้อสอบที่เป็นอยู่จริงของผู้สอบกลุ่มเปรียบเทียบกับค่าประมาณโอกาสของผู้สอบที่ตอบข้อสอบถูกซึ่งทำนายโดยโมเดล 3 พารามิเตอร์ สำหรับสูตรที่ใช้ในการเปรียบเทียบจะคำนวณในรูปค่าเฉลี่ยของคะแนนความแตกต่างมาตรฐานในแต่ละช่วงความสามารถ ดังนี้

$$Z_{iq} = \frac{1}{n_q} \sum_{j \in q} \left[\frac{U_{ij} - P_{ij}}{\sqrt{P_{ij}(1 - P_{ij})}} \right]$$

เมื่อ Z_{iq} แทน ดัชนีการทำหน้าที่ต่างกันของข้อสอบข้อที่ i สำหรับผู้สอบที่มีระดับความสามารถในช่วง q (q = 1, 2, 3,..., m)

n_q แทน จำนวนผู้สอบกลุ่มเปรียบเทียบที่มีระดับความสามารถในช่วง q

U_{ij} แทน ผลการตอบข้อสอบที่สังเกตได้ของผู้สอบคนที่ j ซึ่งตอบข้อสอบข้อที่ i (ถ้าตอบถูก $U_{ij} = 1$ และถ้าตอบผิด $U_{ij} = 0$)

P_{ij} แทน ค่าประมาณโอกาสของผู้สอบคนที่ j จะตอบข้อสอบข้อที่ i ได้ถูกต้อง

ต่อจากนั้นจะคำนวณดัชนีการทำหน้าที่ต่างกันของข้อสอบข้อที่ i รวมทุกช่วงความสามารถ (J_i) ดังนี้

$$Z_i = \frac{\sum_{q=1}^{m} n_q Z_{iq}}{\sum_{q=1}^{m} n_q}$$

2.2.6 **วิธีการทดสอบอัตราส่วนไลค์ลิฮู้ด** (likelihood ratio test; LR)

Thissen และ คณะ (1993) ได้เสนอวิธีการทดสอบอัตราส่วนไลค์ลิฮู้ดสำหรับใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยการทดสอบความแตกต่างของผลการตอบข้อสอบระหว่างผู้สอบกลุ่มย่อย 2 กลุ่ม วิธีการทดสอบอัตราส่วนไลค์ลิฮู้ดแบ่งออกเป็น 3 วิธีย่อย ๆ คือ

(1) ***วิธีอัตราส่วนไลค์ลิฮู้ด IRT ในรูปทั่วไป*** *(general IRT-LR)* เป็นวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบซึ่งใช้การประมาณค่าพารามิเตอร์ในโมเดลการตอบสนองข้อสอบด้วยวิธีมาร์จินอลไลค์ลิฮู้ดสูงสุด (marginal maximum likelihood; MML) *(2)* ***วิธีอัตราส่วนไลค์ลิฮู้ด IRT ในรูปลอกลิเนียร์*** *(loglinear IRT-LR)* จะใช้การประมาณค่าพารามิเตอร์ในโมเดลการตอบสนองข้อสอบในรูปลอกลิเนียร์ด้วยวิธีไลค์ลิฮู้ดสูงสุด (maximum likelihood; ML) ***และ*** *(3)* ***วิธีอัตราส่วนไลค์ลิฮู้ด IRT ในรูปสารสนเทศที่มีขอบเขตจำกัด*** *(limited information IRT-LR)* สำหรับวิธีนี้จะใช้การประมาณค่าพารามิเตอร์ในโมเดลการตอบสนองข้อสอบแบบนอร์มัลออไจฟ์ (normal ogive) ด้วยวิธีกำลังสองน้อยที่สุดในรูปทั่วไป (generalized least squares; GLS) ทั้ง 3 วิธีดังกล่าวจะใช้การทดสอบอัตราส่วนไลค์ลิฮู้ด เพื่อทดสอบนัยสำคัญของการทำหน้าที่ต่างกันของข้อสอบ

หลักการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยวิธีการทดสอบอัตราส่วนไลค์ลิฮู้ดจะเปรียบเทียบระหว่าง 2 โมเดล คือ ***โมเดล compact*** และ***โมเดล augmented*** ในโมเดลแรกสมมติว่าไม่มีกลุ่มผู้สอบที่แตกต่างกัน ดังนั้นจึงบังคับให้พารามิเตอร์ของข้อสอบระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบมีค่าเท่ากัน นั่นคือ ในโมเดล compact จะประกอบด้วยข้อสอบที่ทำหน้าที่ไม่ต่างกัน สำหรับในโมเดลหลังจะประกอบด้วยข้อสอบที่มีค่าพารามิเตอร์ระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่มีค่าแปรเปลี่ยนไปตามกลุ่มผู้สอบ นั่นคือ ในโมเดล augmented อาจมีข้อสอบจำนวน 1 ข้อ หรือมากกว่าที่ทำหน้าที่ต่างกัน นอกจากนี้ในข้อสอบระหว่าง 2 โมเดลจะต้องมีข้อสอบร่วม (anchor items) ซึ่งเป็นข้อสอบที่สมมติว่าทำหน้าที่ไม่ต่างกัน ต่อจากนั้นจึงเปรียบเทียบระหว่าง 2 โมเดล ด้วยสถิติอัตราส่วนไลค์ลิฮู้ด ดังนี้

$$G_i^2 = -2\log\left[\frac{L_c}{L_a}\right]$$

เมื่อ G_i^2 แทน สถิติอัตราส่วนไลค์ลิฮู้ดของข้อสอบข้อที่ i

L_c แทน ฟังก์ชันไลค์ลิฮู้ดของโมเดล compact

L_a แทน ฟังก์ชันไลค์ลิฮู้ดของโมเดล augmented

โดยทั่วไปแล้ว $L_c < L_a$ และสถิติ $G_i^2 > 0$ มีการแจกแจงแบบไค-สแควร์ ซึ่งมีระดับของความเป็นอิสระเท่ากับจำนวนพารามิเตอร์ในโมเดล L_c และโมเดล L_a เช่น ถ้าใช้โมเดลแบบ 3 พารามิเตอร์แสดงว่า df = 3

2.4 **วิธีซิปเทสท์** (SIBTEST)

Shealy และ Stout (1993) ได้เสนอวิธีซิปเทสท์ (simullataneous item bias test; SIBTEST) เพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (differential item functioning; DIF) การทำหน้าที่ต่างกันของแบบสอบ (differential test functioning; DTF) และการทำหน้าที่ต่างกันของกลุ่มข้อสอบ (differential bundle functioning; DBF) โดยสามารถวิเคราะห์ได้ทั้งในแบบสอบเอกมิติ (unidimensional test) และแบบสอบพหุมิติ (multidimensional test) (Stout, Li and Nandakumar, 1997) วิธีซิปเทสท์มีรูปแบบนันพาราเมทริก (nonparametric) พัฒนามาจากโมเดลความลำเอียงของแบบสอบซึ่งมีพื้นฐานมาจากทฤษฎี IRT ชนิดพหุมิติ แต่ไม่ต้องใช้ฟังก์ชันการตอบสนองข้อสอบหรือการประมาณค่าความสามารถแฝง วิธีซิปเทสท์ถูกออกแบบมาสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบมีทิศทางเดียว (unidirectional DIF) ดังนั้นจึงไม่มีความไวในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบไม่มีทิศทาง (nondirectional DIF) (Li and Stout, 1996) จุดเด่นของวิธีซิปเทสท์ก็คือ สามารถคำนวณได้ง่ายไม่ซับซ้อน เสียค่าใช้จ่ายไม่มาก และไม่จำเป็นต้องใช้กลุ่มตัวอย่างที่มีขนาดใหญ่ ทั้งยังใช้สถิติทดสอบนัยสำคัญ (Narayanan and Swaminathan, 1996) นอกจากนี้ยังสามารถนำไปประยุกต์ใช้กับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนแบบพหุวิภาค (polytomous DIF) (Chang, Mazzeo and Roussos, 1995 cited in Potenza and Dorans, 1995)

ในการศึกษาการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีซิปเทสท์ของแบบสอบเอกมิติจะต้องมีข้อตกลงเกี่ยวกับแบบสอบ กล่าวคือ ข้อสอบในแบบสอบจะต้องมุ่งวัดคุณลักษณะหรือความสามารถแฝงเพียงลักษณะเดียว ความสามารถแฝงประเภทหนึ่งเรียกว่า ***ความสามารถเป้าหมายที่ต้องการวัด*** (*target ability* ; θ) แต่จะมีความสามารถแฝงอีกประเภทหนึ่งที่มีอิทธิพลต่อผลการตอบข้อสอบซึ่งเรียกว่า ***ความสามารถแทรกซ้อนที่ไม่ต้องการวัด*** (*nuisance ability* ; η) ตัวอย่างเช่น แบบสอบคำศัพท์ในวิชาภาษาอังกฤษ ข้อสอบบางข้ออาจถามความรู้สำหรับผู้ชายเป็นพิเศษ เช่น ความรู้เกี่ยวกับกีฬา ในขณะที่ข้อสอบบางข้ออาจถามความรู้เกี่ยวกับผู้หญิงโดยเฉพาะ เช่น ความรู้เกี่ยวกับงานในบ้าน จากสถานการณ์ดังกล่าวทักษะความรู้เกี่ยวกับคำศัพท์ในวิชาภาษาอังกฤษเป็นความสามารถเป้าหมายที่ต้องการวัด ซึ่งแทนด้วย θ ส่วนความรู้ทางด้านกีฬาและงานในบ้านเป็นความสามารถแทรกซ้อนที่ไม่ต้องการวัด ซึ่งแทนด้วย η_1 และ η_2 ตามลำดับ ข้อสอบทุกข้อในแบบสอบจะวัดความสามารถเป้าหมาย ส่วนข้อสอบบางข้อที่ทำหน้าที่ต่างกันจะวัดทั้งความสามารถเป้าหมายและความสามารถแทรกซ้อน (Nandakumar, 1993)

ความสามารถ θ และ η ได้มาจากผลการตอบข้อสอบระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ส่วนผลการตอบข้อสอบได้มาจากการสุ่มจากกลุ่มผู้ตอบในแต่ละกลุ่ม ซึ่งแทนด้วย $\mathrm{U} = (U_1, U_2, ..., U_N)$ โดยที่ U_i เท่ากับ 1 ถ้าตอบข้อสอบข้อที่ i ถูก และ U_i เท่ากับ 0 ถ้าตอบข้อสอบผิด ส่วน N เป็นจำนวนข้อสอบ ในการสร้าง U ตามโมเดล IRT โดยทั่ว ๆ ไป จะต้องประกอบด้วย 2 องค์ประกอบคือ (1) ***พารามิเตอร์ความสามารถของผู้สอบจำนวน d มิติ*** และ (2) ***ฟังก์ชันการตอบสนองข้อสอบ*** (*IRFs*) ซึ่งฟังก์ชันการตอบสนองข้อสอบในแต่ละข้อจะเป็นตัวกำหนดความน่าจะเป็นในการตอบข้อสอบได้ถูกต้อง สำหรับในที่นี้ $d = 2$ เพราะมีพารามิเตอร์ความสามารถสองชนิดคือ ความสามารถเป้าหมายที่ต้องการวัด (θ) และความสามารถแทรกซ้อนที่ไม่ต้องการวัด (η) ส่วนเวคเตอร์ความสามารถที่กำหนดจากผู้สอบในแต่ละกลุ่มแทนด้วย (θ,η) โดยที่การแจกแจงของ (θ,η) ถูกเลือกขึ้นมาอย่างสุ่มจากกลุ่มผู้สอบ และความสามารถที่สุ่มมาจากกลุ่มผู้สอบแทนด้วย (Θ,η) โดยสมมติว่าข้อสอบทุกข้อของแบบสอบจะวัดความสามารถ θ ส่วนข้อสอบบางข้อที่ทำหน้าที่ต่างกันจะวัดทั้งความสามารถ θ และ η ซึ่งความสามารถ η อาจมีเพียงความสามารถเดียว หรือมากกว่าหนึ่งความสามารถก็ได้ สำหรับ IRF ข้อที่ i ซึ่งขึ้นอยู่กับความสามารถ θ เพียงอย่างเดียวแทนด้วย $P_i(\theta)$ ส่วน IRF ข้อที่ i ซึ่งขึ้นอยู่กับความสามารถ θ และ η แทนด้วย $P_i(\theta,\eta)$ ดังนี้ (Shealy and Stout, 1993)

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

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

ค่า IRF ทั้ง 2 ลักษณะดังกล่าวเป็นโมเดลการตอบสนองข้อสอบแบบ 3 พารามิเตอร์ ซึ่งมีคุณสมบัติไม่แปรเปลี่ยน (invariance) ไปตามกลุ่มผู้สอบ ทั้งยังมีข้อตกลงเกี่ยวกับความเป็นอิสระต่อกันในการตอบข้อสอบ (local independence) ของ U เมื่อให้ IRF แทนด้วย $P_i(\theta,\eta)$ สามารถกำหนดในรูปของความน่าจะเป็นดังนี้ (Li and Stout, 1996)

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

Shealy และ Stout (1993) ได้ใช้ marginal IRFs อธิบายการทำหน้าที่ต่างกันของข้อสอบ ดังนี้

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

เมื่อ $M_{gi}(\theta)$ แทน marginal IRF ที่เกี่ยวกับความสามารถเป้าหมายที่ต้องการวัด θ ของผู้สอบกลุ่มอ้างอิงหรือกลุ่มเปรียบเทียบ

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

$f_g(\eta \mid \theta)$ แทน การแจกแจงแบบมีเงื่อนไขของกลุ่มผู้สอบ

การทำหน้าที่ต่างกันของข้อสอบแบบมีทิศทางเดียว (*unidirectional DIF*) เกิดขึ้นเมื่อโอกาสของการตอบข้อสอบถูกจากผู้สอบกลุ่มหนึ่งมีค่ามากกว่าผู้สอบอีกกลุ่มหนึ่งตลอดช่วงความสามารถ ซึ่งตามทฤษฎี IRT สามารถแสดงได้ในรูปโค้งลักษณะข้อสอบของผู้สอบสองกลุ่มไม่ตัดกัน (noncrossing ICCs) ข้อสอบจะเข้าข้างผู้สอบกลุ่มใดนั้นให้พิจารณาค่า marginal IRFs กล่าวคือ ถ้า $M_{iF}(\theta) < M_{iR}(\theta)$ ทุกค่าความสามารถ θ แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบมีทิศทางเดียว โดยข้อสอบจะเข้าข้างผู้สอบกลุ่มอ้างอิง และถ้า $M_{iF}(\theta) > M_{iR}(\theta)$ ทุกค่าความสามารถ θ แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบมีทิศทางเดียว โดยข้อสอบจะเข้าข้างผู้สอบกลุ่มเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบแบบมีทิศทางเดียวอาจเรียกอีกอย่างหนึ่งว่า ***"การทำหน้าที่ต่างกันแบบไม่ตัดกัน"*** (*noncrossing DIF*)

หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบมีทิศทางเดียวด้วยวิธี SIBTEST ของ Shealy และ Stout (1993) จะเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ โดยแบ่งแบบสอบออกเป็น 2 ชุดย่อย (subtests) คือ ***(1) ชุดแบบสอบที่มีความตรง*** (*valid subtests*) ***หรือชุดแบบสอบที่ใช้ในการจับคู่เปรียบเทียบ*** (*matching subtests*) แบบสอบชุดนี้ประกอบด้วยข้อสอบที่ไม่สงสัยว่าทำหน้าที่ต่างกัน และ ***(2) ชุดแบบสอบที่ต้องการศึกษา*** (*studied subtests*) ประกอบด้วยข้อสอบที่สงสัยว่าทำหน้าที่ต่างกัน ถ้าแบบสอบชุดแรกมีจำนวน n ข้อ แล้วแบบสอบชุดที่สองจะมีจำนวน $N - n$ ข้อ เมื่อ N เป็นจำนวนข้อสอบทั้งหมด

ฟังก์ชันการตอบสนองข้อสอบของแบบสอบที่ต้องการศึกษาจากผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ กำหนดในรูปฟังก์ชันดังนี้

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

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

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

$M_{SF}(\theta)$ แทน ผลรวมของ marginal IRFs ของข้อสอบที่ต้องการศึกษา จากผู้สอบกลุ่มเปรียบเทียบ ณ ระดับความสามารถ θ

n แทน จำนวนข้อสอบในชุดแบบสอบที่มีความตรง

N แทน จำนวนข้อสอบทั้งหมด

ปริมาณขนาดของความแตกต่างระหว่าง $M_{SR}(\theta)$ และ $M_{SF}(\theta)$ แสดงถึงปริมาณของการทดสอบการทำหน้าที่ต่างกันของข้อสอบแบบมีทิศทางเดียว หรือการทำหน้าที่ต่างกันแบบไม่ตัดกันจากชุดแบบสอบที่ต้องการศึกษา ณ ระดับความสามารถ θ ซึ่งสามารถคำนวณในรูปการอินทิเกรท ดังนี้

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

เมื่อ β_{uni} แทน ดัชนีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบมีทิศทางเดียว

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

ดัชนี β_{uni} ที่คำนวณได้จากสูตรดังกล่าว นำมาทดสอบสมมติฐานของการทำหน้าที่ต่างกันของข้อสอบแบบมีทิศทางเดียว ดังนี้

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

สมมติฐานอื่น (H_1) มีลักษณะทิศทางเดียว ซึ่งใช้ทดสอบการทำหน้าที่ต่างกันของข้อสอบที่เข้าข้างผู้สอบกลุ่มเปรียบเทียบ สำหรับค่าประมาณของ β_{uni} คำนวณได้จากคะแนนของชุดแบบสอบที่มีความตรงและชุดแบบสอบที่ต้องการศึกษา ซึ่งกำหนดด้วยสัญลักษณ์ดังนี้

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

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

เมื่อ X แทน คะแนนรวมของชุดแบบสอบที่มีความตรง

Y แทน คะแนนรวมของชุดแบบสอบที่ต้องการศึกษา

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

นำคะแนนเฉลี่ยจากผลการตอบชุดแบบสอบที่ต้องการศึกษาระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่มีความสามารถระดับเดียวกันมาจับคู่เปรียบเทียบกัน ซึ่งพิจารณาจากคะแนนรวมที่เท่ากันของชุดแบบสอบที่มีความตรง ($X = k$) โดยเขียนในรูปสัญลักษณ์ได้ดังนี้

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

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

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

k แทน คะแนนรวมจากชุดแบบสอบที่มีความตรง

ค่า $\bar{Y}_{Rk} - \bar{Y}_{Fk}$ ดังกล่าวเป็นความแตกต่างของผลการตอบข้อสอบในชุดแบบสอบที่ต้องการศึกษาระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่มีความสามารถระดับเดียวกัน ถ้า $\bar{Y}_{Rk} - \bar{Y}_{Fk} = 0$ ทุกคะแนน k แสดงว่าข้อสอบที่ต้องการศึกษาทำหน้าที่ไม่ต่างกัน และถ้า $\bar{Y}_{Rk} - \bar{Y}_{Fk} > 0$ ทุกคะแนน k แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบมีทิศทางเดียว โดยจะเข้าข้างผู้สอบกลุ่มเปรียบเทียบ ค่าความแตกต่างของผลการตอบข้อสอบสามารถประมาณค่าในรูป β_{uni} ได้ดังนี้

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

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

เมื่อ $\hat{P}_k$ แทน สัดส่วนของผู้สอบทั้งหมด (กลุ่มอ้างอิงและกลุ่มเปรียบเทียบ) ซึ่งตอบชุดแบบสอบที่มีความตรงแล้วได้คะแนนรวม $X = k$

J_{Fk} แทน จำนวนผู้สอบกลุ่มเปรียบเทียบซึ่งตอบชุดแบบสอบที่มีความตรงแล้วได้คะแนนรวม $X = k$

J_{Rk} แทน จำนวนผู้สอบกลุ่มอ้างอิงซึ่งตอบชุดแบบสอบที่มีความตรงแล้วได้คะแนนรวม $X = k$

สำหรับการทดสอบสมมติฐานศูนย์ของ no DIF ใช้สถิติ B_{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 $\hat{\sigma}(\hat{\beta}_{uni})$ แทน ค่าประมาณความคลาดเคลื่อนมาตรฐานของ β_{uni}

$\hat{\sigma}^2(Y|k,g)$ แทน ค่าประมาณความแปรปรวนของคะแนนจากชุดแบบสอบที่ต้องการศึกษาสำหรับผู้สอบกลุ่ม g (R หรือ F) ซึ่งมีคะแนนรวมเท่ากับ k

J_{gk} แทน จำนวนผู้สอบกลุ่ม g (R หรือ F) ซึ่งตอบชุดแบบสอบที่มีความตรงแล้วได้คะแนนรวม $X = k$

สถิติที่ใช้ในการทดสอบ B_{uni} มีการแจกแจงในลักษณะปกติมาตรฐาน $[N(0,1)]$ เมื่อ $B_{uni} = 0$ และถ้าผลการทดสอบปรากฏว่า $B_{uni} > Z_\alpha$ อย่างมีนัยสำคัญที่ระดับ α โดยที่ $\alpha = P[N(0,1) > Z_\alpha]$ แสดงว่าปฏิเสธ H_0 นั่นคือ ข้อสอบที่นำมาตรวจสอบทำหน้าที่ต่างกัน โดยจะเข้าข้างผู้สอบกลุ่มเปรียบเทียบเมื่อ B_{uni} มีค่าเป็นบวก และจะเข้าข้างผู้สอบกลุ่มอ้างอิงเมื่อ B_{uni} มีค่าเป็นลบ

อย่างไรก็ตาม **สถิติที่ใช้**ในการอ้างอิงการทำหน้าที่ต่างกันของข้อสอบมักจะมีปัญหาในกรณีที่มีความแตกต่างของการแจกแจงความสามารถเป้าหมายระหว่างกลุ่มผู้สอบ กล่าวคือ ถ้าผู้สอบกลุ่มอ้างอิงมีความสามารถเป้าหมายสูงกว่าผู้สอบกลุ่มเปรียบเทียบ จะเกิดเงื่อนไขที่เรียกว่า "***ผลกระทบ***" (*impact*) ซึ่งจะมีผลทำให้สถิติ B_{uni} มีค่าเฟ้อ (inflate) หรือมีค่าสูงผิดปกติ ถึงแม้ว่า

ในความเป็นจริงข้อสอบทำหน้าที่ไม่ต่างกัน ดังนั้นจึงมีความจำเป็นที่จะแก้ไขความแตกต่างของการแจกแจงความสามารถเป้าหมายด้วยวิธีการปรับแก้ค่าการถดถอย (regression correction) เพื่อกำจัดอิทธิพลค่าเฟ้อของผลกระทบโดยการแปลงค่า $\overline{Y}_{Rk}$, $\overline{Y}_{Fk}$ ให้เป็น $\overline{Y}^{*}_{Rk}$, $\overline{Y}^{*}_{Fk}$ ทีละคู่ดังนี้

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

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

$$\hat{V}(k) = \frac{1}{2}\left[\hat{V}_R(k) + \hat{V}_F(k)\right]$$

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

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

$$\hat{\sigma}^2(X|g) = \frac{1}{(J_g - 1)} \sum_{j=1}^{J_g} (X_{gj} - \overline{X}_g)^2$$

เมื่อ $\overline{Y}^{*}_{gk}$ แทน ค่าเฉลี่ยที่ปรับแก้แล้วของคะแนนรวมจากชุดแบบสอบที่ต้องการศึกษาในผู้สอบกลุ่ม g

$\hat{M}_{gk}$ แทน ค่าประมาณอย่างคร่าว ๆ ของค่าเบี่ยงเบนของคะแนนจริงจากชุดแบบสอบที่ต้องการศึกษาที่มีต่อฟังก์ชันของคะแนนจริงจากชุดแบบสอบที่มีความตรง

$\hat{V}(k)$ แทน ค่าประมาณของคะแนนจากชุดแบบสอบที่มีความตรง ซึ่งได้คะแนนสังเกต $X = k$

$\hat{V}_g(k)$ แทน ค่าประมาณการถดถอยของคะแนนจริงจากชุดแบบสอบที่มีความตรง ซึ่งได้คะแนนสังเกต $X = k$

$\overline{Y}_{gk}$ แทน ค่าเฉลี่ยของคะแนนสังเกตจากชุดแบบสอบที่ต้องการศึกษาของผู้สอบกลุ่ม g (R หรือ F) ซึ่งได้คะแนน $X = k$

X_{gj} แทน คะแนนจากชุดแบบสอบที่มีความตรงของผู้สอบคนที่ j ในกลุ่ม g

$\overline{X}_g$ แทน คะแนนเฉลี่ยจากชุดแบบสอบที่มีความตรงของผู้สอบกลุ่ม g

$\overline{U}_{ig}$ แทน สัดส่วนการตอบข้อสอบถูกของผู้สอบกลุ่ม g ซึ่งตอบข้อสอบข้อที่ i จากชุดแบบสอบที่มีความตรง

J_g แทน จำนวนผู้สอบกลุ่ม g

$\hat{\sigma}^2(e|g)$ แทน ค่าประมาณความแปรปรวนของความคลาดเคลื่อนในผู้สอบกลุ่ม g

$\hat{\sigma}^2(X|g)$ แทน ค่าประมาณความแปรปรวนของคะแนนสังเกตในผู้สอบกลุ่ม g

ต่อจากนั้น จึงนำค่า $\bar{Y}^*_{Rk}$ และ $\bar{Y}^*_{Fk}$ มาจับคู่เปรียบเทียบ $\bar{Y}^*_{Rk} - \bar{Y}^*_{Fk}$ ในลักษณะเดียวกับการเปรียบเทียบ $\bar{Y}_{Rk} - \bar{Y}_{Fk}$ ดังที่กล่าวมาแล้วข้างต้น เพื่อคำนวณดัชนี β_{uni} ต่อไป

ตอนที่ 5 งานวิจัยที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

1. งานวิจัยต่างประเทศ

Swaminathan และ Rogers (1990) ได้เปรียบเทียบระหว่างวิธีการถดถอยโลจิสติกและวิธีแมนเทล-แฮนส์เซล ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปและแบบอเนกรูป โดยศึกษาในสถานการณ์จำลอง 6 เงื่อนไข คือ ขนาดกลุ่มตัวอย่าง 2 ระดับ (250 คนต่อกลุ่ม และ 500 คนต่อกลุ่ม) และความยาวของแบบสอบ 3 ระดับ (40 ข้อ 60 ข้อ และ 80 ข้อ) ซึ่งในแบบสอบแต่ละชุดประกอบด้วยสัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันจำนวน 20% โดยครึ่งหนึ่งเป็นข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูป และอีกครึ่งหนึ่งเป็นข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูป สำหรับผลการตอบข้อสอบทั้งหมดจำลองโดยใช้โปรแกรม DATAGEN โมเดลการตอบสนองข้อสอบแบบ 3 พารามิเตอร์ ในการจำลองข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูปจะกำหนดให้พารามิเตอร์อำนาจจำแนกระหว่างผู้สอบ 2 กลุ่มมีค่าเท่ากัน ในขณะที่พารามิเตอร์ความยากจะมีค่าแปรเปลี่ยน สำหรับการจำลองข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปจะกำหนดให้พารามิเตอร์ความยากระหว่างผู้สอบ 2 กลุ่มมีค่าเท่ากัน ส่วนพารามิเตอร์อำนาจจำแนกจะมีค่าแปรเปลี่ยน ในการควบคุมขนาดของการทำหน้าที่ต่างกันของข้อสอบจะใช้พื้นที่ระหว่างโค้งลักษณะข้อสอบ ซึ่งคำนวณโดยใช้สูตรของ Raju

ผลการศึกษาพบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป วิธีการถดถอยโลจิสติกมีอำนาจการทดสอบสูงกว่าวิธีแมนเทล-แฮนส์เซล ส่วนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปทั้ง 2 วิธีมีอำนาจการทดสอบเท่าเทียมกัน สำหรับปัจจัยของขนาดกลุ่มตัวอย่าง พบว่า เมื่อขนาดกลุ่มตัวอย่าง 250 คนต่อกลุ่มผู้สอบ ทั้ง 2 วิธีมีความถูกต้องแม่นยำในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปประมาณ 75% และเมื่อขนาดกลุ่มตัวอย่าง 500 คนต่อกลุ่มผู้สอบ ทั้ง 2 วิธีมีความถูกต้องแม่นยำในการตรวจสอบการทำหน้าที่

ต่างกันของข้อสอบแบบเอกรูปประมาณ 100% สำหรับอัตราความคลาดเคลื่อนประเภทที่ 1 พบว่า วิธีแมนเทล-แฮนส์เซลมีอัตราความคลาดเคลื่อนร้อยละ 1 ในขณะที่วิธีการถดถอยโลจิสติกมีอัตราความคลาดเคลื่อนระหว่างร้อยละ 1 ถึง 6 เมื่อพิจารณาถึงค่าใช้จ่ายในการวิเคราะห์ ปรากฏว่าวิธีการถดถอยโลจิสติกมีค่าใช้จ่ายสูงกว่าวิธีแมนเทล-แฮนส์เซลประมาณ 3 – 4 เท่า

Kim และ Cohen (1991) ได้เปรียบเทียบวิธีการวัดพื้นที่ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้วิธีการวัดพื้นที่แตกต่างกัน 4 วิธี คือ วิธีการวัดพื้นที่ในช่วงเปิดของ Raju ชนิดคิดเครื่องหมาย (ESA) ชนิดไม่คิดเครื่องหมาย (EUA) วิธีการวัดพื้นที่ในช่วงปิดของ Kim และ Cohen ชนิดคิดเครื่องหมาย (CSA) ชนิดไม่คิดเครื่องหมาย (CUA) โดยใช้ข้อมูลเดิมของ Subkoviak, Mack, Ironson และ Craig (1984) ซึ่งประกอบด้วยกลุ่มตัวอย่างผิวดำ 1,008 คน และผิวขาว 1,021 คน แล้วสุ่มมากลุ่มละ 1,000 คน เครื่องมือที่ใช้ในการเก็บรวบรวมข้อมูลเป็นแบบสอบวัดคำศัพท์ชนิดเลือกตอบ 4 ตัวเลือก จำนวน 50 ข้อ ซึ่งประกอบด้วยข้อสอบที่เกี่ยวกับคำศัพท์ภาษาอังกฤษอเมริกันมาตรฐาน 40 ข้อ ส่วนอีก 10 ข้อ เป็นข้อสอบที่เกี่ยวกับคำศัพท์ง่าย ๆ สำหรับผู้สอบกลุ่มผิวดำ ในการประมาณค่าพารามิเตอร์ของข้อสอบจะใช้โปรแกรม BILOG โมเดลโลจิสติกแบบ 3 พารามิเตอร์ (3PLM) และแบบ 3 พารามิเตอร์เมื่อกำหนดค่าการเดาคงที่ (3PLM-c) โดยกำหนดค่า c เท่ากับ .23 สำหรับการปรับเทียบค่าพารามิเตอร์ของข้อสอบระหว่างกลุ่มผู้สอบใช้โปรแกรม EQUATE ส่วนการเปรียบเทียบดัชนี ESA, EUA, CSA และ CUA จะคำนวณค่าความสัมพันธ์ระหว่างวิธีการวัดพื้นที่และอัตราความคลาดเคลื่อน 2 ประเภท คือ อัตราความคลาดเคลื่อนประเภทที่ 1 (false positive; FP) และอัตราความคลาดเคลื่อนประเภทที่ 2 (false negative; FN) โดยทดสอบนัยสำคัญแบบทางเดียวที่ระดับ .05 และ .01

ผลการศึกษาพบว่า ค่าการวัดพื้นที่ในช่วงเปิดของ Raju และช่วงปิดของ Kim และ Cohen ภายใต้โมเดล 3PLM และ 3PLM-c มีความแตกต่างกันเล็กน้อย การวัดพื้นที่ด้วยวิธี CSA แบบ 3PLM มีความไวในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมากกว่าวิธีอื่น ๆ เล็กน้อย ส่วนผลการศึกษาอัตราความคลาดเคลื่อน พบว่า การวัดพื้นที่ด้วยวิธี CSA แบบ 3PLM มีอัตราความคลาดเคลื่อนประเภทที่ 1 และ 2 ต่ำกว่าวิธีอื่น ๆ สำหรับการวัดพื้นที่ด้วยวิธี CSA แบบ 3PLM ที่ระดับนัยสำคัญ .01 จะไม่พบอัตราความคลาดเคลื่อนทั้ง 2 ประเภท และที่ระดับนัยสำคัญ .05 จะไม่พบอัตราความคลาดเคลื่อนประเภทที่ 2 นอกจากนี้ยังพบว่า การวัดพื้นที่ด้วยวิธี CUA แบบ 3PLM จะมีอัตราความคลาดเคลื่อนประเภทที่ 1 สูงกว่าวิธีอื่น ๆ ที่ระดับนัยสำคัญ .01 และ .05

Cohen และ Kim (1993) ได้ศึกษาเปรียบเทียบวิธีการทดสอบไค-สแควร์ของ Lord และวิธีการวัดพื้นที่ชนิดคิดเครื่องหมายและชนิดไม่คิดเครื่องหมายของ Raju ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยใช้โปรแกรม GENIRV จำลองข้อมูลโมเดล 2 PLM จำนวน 140 เงื่อนไข (2 × 2 × 7 × 5) คือ ความยาวของแบบสอบ 2 ระดับ (20 ข้อและ 60 ข้อ) ขนาดกลุ่มตัวอย่าง 2 ระดับ (100 คนและ 500 คน) กลุ่มผู้สอบ 7 ระดับ (ผู้สอบกลุ่มอ้างอิงจำนวน 1 กลุ่ม ผู้สอบกลุ่มเปรียบเทียบที่จับคู่ความสามารถในเงื่อนไขของสัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน 0%, 10% และ 20% จำนวน 3 กลุ่ม และผู้สอบกลุ่มเปรียบเทียบที่ไม่ได้จับคู่ความสามารถในเงื่อนไขของสัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน 0%, 10% และ 20% จำนวน 3 กลุ่ม) โดยกระทำซ้ำ 5 ครั้งในแต่ละเงื่อนไข ในการประมาณค่าพารามิเตอร์ของข้อสอบและความสามารถของผู้สอบจะใช้โปรแกรม BILOG โดยใช้วิธี marginal maximum likelihood estimation (MMLE) และวิธี marginal Bayesian estimation (MBE) ส่วนการแปลงค่าพารามิเตอร์ของข้อสอบให้อยู่บนเมทริกซ์เดียวกันจะใช้วิธีฟังก์ชันการตอบสนองแบบสอบ (test response function; TRF) สำหรับการเปรียบเทียบวิธีการตรวจสอบทั้งสองจะคำนวณค่าความคลาดเคลื่อน 2 ประเภท คือ อัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราความคลาดเคลื่อนประเภทที่ 2 โดยทดสอบนัยสำคัญแบบทางเดียวที่ระดับ .05 และ .01

ผลการศึกษาพบว่า เมื่อสัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบมีจำนวนลดลงแล้วอัตราความคลาดเคลื่อนประเภทที่ 1 ของทั้ง 3 วิธีจะมีค่าเพิ่มขึ้น และเมื่อเพิ่มความยาวของแบบสอบแล้วอัตราความคลาดเคลื่อนประเภทที่ 1 จะมีค่าเพิ่มขึ้นเล็กน้อย โดยมีค่าเพิ่มขึ้นภายใต้เงื่อนไขที่ไม่ได้จับคู่ความสามารถมากกว่าภายใต้เงื่อนไขที่ได้จับคู่ความสามารถ นอกจากนี้เมื่อระดับนัยสำคัญมีค่าเพิ่มขึ้นแล้วอัตราความคลาดเคลื่อนประเภทที่ 1 จะมีค่าเพิ่มขึ้นด้วย ผลดังกล่าวจะเกิดขึ้นกับวิธีการวัดพื้นที่ของ Raju ภายใต้ทุกเงื่อนไขของการตรวจสอบ แต่จะเกิดขึ้นกับวิธีการทดสอบไค-สแควร์ของ Lord ภายใต้บางเงื่อนไขของการตรวจสอบ สำหรับผลการศึกษาอัตราความคลาดเคลื่อนประเภทที่ 2 พบว่า เมื่อสัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบมีจำนวนเพิ่มขึ้นและแบบสอบมีขนาดความยาวมากขึ้นแล้วอัตราความคลาดเคลื่อนประเภทที่ 2 ของทั้ง 3 วิธีจะมีค่าเพิ่มมากขึ้น โดยมีค่าเพิ่มขึ้นภายใต้เงื่อนไขที่ไม่ได้จับคู่ความสามารถมากกว่าภายใต้เงื่อนไขที่ได้จับคู่ความสามารถ สำหรับระดับนัยสำคัญของการทดสอบ พบว่า เมื่อระดับนัยสำคัญมีค่าเพิ่มขึ้นแล้วอัตราความคลาดเคลื่อนประเภทที่ 2 จะมีค่าลดลง ผลดังกล่าวจะเกิดขึ้นกับวิธีการวัดพื้นที่ของ Raju มากกว่าวิธีการทดสอบไค-สแควร์ของ Lord นอกจากนี้ยังพบว่า เมื่อ

ประมาณค่าพารามิเตอร์โดยใช้วิธี MBE จะส่งผลให้อัตราความคลาดเคลื่อนประเภทที่ 2 ของทั้ง 3 วิธีมีค่าน้อยกว่าประมาณค่าโดยใช้วิธี MMLE

Raju, Drasgow และ Slinde (1993) ได้ศึกษาเปรียบเทียบการประเมินการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธีการวัดพื้นที่ (ชนิดคิดเครื่องหมายและชนิดไม่คิดเครื่องหมาย) วิธีการทดสอบไค-สแควร์ของ Lord และวิธีแมนเทล-แฮนส์เซล โดยศึกษากับข้อมูลเชิงประจักษ์ ใช้กลุ่มตัวอย่างนักเรียนระดับ 10 และระดับ 12 จำนวน 839 คน การแบ่งกลุ่มตัวอย่างเป็นกลุ่มอ้างอิงและกลุ่มเปรียบเทียบใช้สีผิวและเพศเป็นเกณฑ์ ในกรณีใช้สีผิวเป็นเกณฑ์ได้นักเรียนผิวดำ 245 คน และผิวขาว 436 คน ส่วนที่เหลืออีก 158 คน แตกต่างไปจากทั้งสองกลุ่มจึงคัดออก ในกรณีใช้เพศเป็นเกณฑ์ได้นักเรียนหญิง 440 คน และนักเรียนชาย 399 คน ข้อมูลที่ใช้ในการศึกษาได้มาจากการทดลองสอบกับนักเรียนระดับ 10 และ 12 ในปี ค.ศ.1987 โดยใช้แบบสอบ Gates-MacGinitie Reading Tests (GMRT) ซึ่งเป็นชุดข้อสอบที่ใช้วัดความรู้เกี่ยวกับคำศัพท์จำนวน 45 ข้อ ในแต่ละข้อมี 5 ตัวเลือก สำหรับในการวิเคราะห์ข้อมูลจะเริ่มต้นด้วยการตรวจสอบความเป็นเอกมิติของแบบสอบโดยการวิเคราะห์องค์ประกอบหลักกับผลการตอบข้อสอบของกลุ่มตัวอย่างทั้งหมด แล้วจึงประมาณค่าพารามิเตอร์ของข้อสอบตามโมเดลโลจิสติกแบบ 2PLM โดยใช้โปรแกรม BILOG ด้วยวิธีการประมาณค่าของ Bayes ซึ่งจะประมาณค่าพารามิเตอร์แยกกลุ่มผู้สอบออกเป็น 4 กลุ่ม คือ กลุ่มผิวดำ กลุ่มผิวขาว กลุ่มหญิง และกลุ่มชาย ต่อจากนั้นจึงใช้โปรแกรม EQUATE ด้วยวิธีเทียบโค้งลักษณะข้อสอบของ Stocking และ Lord เพื่อแปลงค่าพารามิเตอร์ของข้อสอบระหว่างกลุ่มผู้สอบให้อยู่บนสเกลเดียวกัน หลังจากนั้นจึงคำนวณการทำหน้าที่ต่างกันของข้อสอบระหว่างกลุ่มผู้สอบดำ-ขาว และกลุ่มผู้สอบชาย-หญิง โดยใช้วิธีการวัดพื้นที่ชนิดคิดเครื่องหมายและชนิดไม่คิดเครื่องหมายของ Raju วิธีการทดสอบไค-สแควร์ของ Lord และวิธีแมนเทล-แฮนส์เซล สำหรับการทดสอบนัยสำคัญทางสถิติกับวิธีแรกใช้สถิติ Z ส่วนสองวิธีหลังใช้สถิติไค-สแควร์

ผลการศึกษาพบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกรณีการแบ่งกลุ่มตามเพศและสีผิว วิธีการวัดพื้นที่ชนิดคิดเครื่องหมาย วิธีการวัดพื้นที่ชนิดไม่คิดเครื่องหมาย และวิธีการทดสอบไค-สแควร์ของ Lord สามารถระบุข้อสอบที่ทำหน้าที่ต่างกันได้สอดคล้องกันอย่างมีนัยสำคัญ เฉพาะกรณีการแบ่งกลุ่มตามเพศ วิธีการวัดพื้นที่ชนิดคิดเครื่องหมาย การวัดพื้นที่ชนิดไม่คิดเครื่องหมาย วิธีการทดสอบไค-สแควร์ของ Lord และวิธีแมนเทล-แฮนส์เซล สามารถระบุข้อสอบที่ทำหน้าที่ต่างกันได้สอดคล้องกันสูง ส่วนในกรณีการแบ่งกลุ่มตามสีผิวจะระบุข้อสอบที่ทำหน้าที่ต่างกันได้แตกต่างกัน

Rogers และ Swaminathan (1993) ได้ศึกษาเปรียบเทียบวิธีการถดถอยโลจิสติกและวิธีแมนเทล-แฮนส์เซล ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปและแบบอเนกรูป โดยศึกษาในสถานการณ์จำลอง 2 ครั้ง คือ ครั้งที่ 1 ศึกษาการแจกแจงของสถิติที่ใช้ทดสอบ และครั้งที่ 2 ศึกษาอำนาจการทดสอบ ภายใต้การจำลองข้อมูล 32 เงื่อนไข (2 × 2 × 2 × 2 × 2) คือ ความเหมาะสมของข้อมูลกับโมเดล 2 ระดับ (เหมาะสมและไม่เหมาะสม) ขนาดกลุ่มตัวอย่าง 2 ระดับ (250 คนต่อกลุ่มและ 500 คนต่อกลุ่ม) ขนาดความยาวของแบบสอบ 2 ระดับ (40 ข้อและ 80 ข้อ) โค้งของการแจกแจงคะแนนการสอบ 2 ระดับ (ปกติและเบ้ซ้าย) สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน 2 ระดับ (15% และ 0%) ในแต่ละเงื่อนไขจะจำลองข้อสอบที่ทำหน้าที่ต่างกัน 2 ประเภท คือ แบบเอกรูปและแบบอเนกรูป โดยข้อสอบที่ทำหน้าที่ต่างกันในแต่ละประเภทและในแต่ละเงื่อนไขจะมีขนาดของข้อสอบที่ทำหน้าที่ต่างกันเท่ากับ .2, .4, .6 และ .8 ซึ่งขนาดของข้อสอบดังกล่าวคำนวณจากพื้นที่ระหว่าง IRFs สำหรับการจำลองข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูปจะมีข้อสอบ 4 ลักษณะ คือ b ต่ำกับ a สูง, b ปานกลางกับ a ต่ำ, b ปานกลางกับ a สูง และ b สูงกับ a สูง ส่วนการจำลองข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปจะมีข้อสอบ 4 ลักษณะ คือ b ต่ำกับ a ต่ำ, b ปานกลางกับ a ต่ำ, b ปานกลางกับ a สูง และ b สูงกับ a ต่ำ

ผลการศึกษากรณีการศึกษาการแจกแจงของสถิติ พบว่า วิธีการถดถอยโลจิสติกและวิธีแมนเทล-แฮนส์เซลให้ผลการตรวจสอบเป็นไปตามที่คาดไว้เกือบทุกเงื่อนไข เมื่อข้อสอบยากมากและค่าอำนาจจำแนกสูง การแจกแจงของสถิติที่ใช้ทดสอบวิธีการถดถอยโลจิสติกจะไม่เป็นไปตามที่คาดไว้ สำหรับผลการศึกษาอำนาจการทดสอบ พบว่า วิธีการถดถอยโลจิสติกและวิธีแมนเทล-แฮนส์เซลมีอำนาจการทดสอบเท่าเทียมกันในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป ส่วนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป วิธีการถดถอยโลจิสติกมีอำนาจการทดสอบสูงกว่าวิธีแมนเทล-แฮนส์เซล สำหรับปัจจัยของขนาดกลุ่มตัวอย่าง ลักษณะของข้อสอบ และขนาดอิทธิพลของข้อสอบที่ทำหน้าที่ต่างกันจะมีผลต่ออำนาจการทดสอบของวิธีการถดถอยโลจิสติกและวิธีแมนเทล-แฮนส์เซล กล่าวคือ เมื่อขนาดกลุ่มตัวอย่างและขนาดอิทธิพลของข้อสอบที่ทำหน้าที่ต่างกันมีค่าเพิ่มขึ้นจะมีผลทำให้อำนาจการทดสอบของทั้ง 2 วิธีดังกล่าวมีค่าเพิ่มขึ้น ภายใต้การทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปและแบบอเนกรูป และเมื่อลักษณะของข้อสอบมีค่าอำนาจจำแนกสูงและค่าความยากปานกลางจะมีผลทำให้อำนาจการทดสอบในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปและแบบอเนกรูปของทั้ง 2 วิธีดังกล่าวมีค่าสูงสุด แต่เมื่อลักษณะของข้อสอบมีค่าความยากปานกลางจะมีผลทำให้อำนาจการทดสอบในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปของวิธีแมนเทล-แฮนส์เซลมีค่าต่ำสุด

Mazor, Clauser และ Hambleton (1994) ได้ศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปด้วยวิธีแมนเทล-แฮนส์เซล โดยจำลองกลุ่มผู้สอบ 3 กลุ่ม ๆ ละ 1,000 คน ตามทฤษฎี IRT โมเดลแบบ 3PLM จำลองการแจกแจงค่าความสามารถของผู้สอบออกเป็น 2 แบบ คือ แบบเท่ากันและแบบไม่เท่ากัน สำหรับแบบสอบจำลองจำนวน 25 ฉบับ ในแต่ละฉบับประกอบด้วยข้อสอบที่ทำหน้าที่ไม่ต่างกันจำนวน 59 ข้อ และข้อสอบที่ศึกษาจำนวน 16 ข้อ รวมข้อสอบทั้งหมดจำนวน 75 ข้อ โดยข้อสอบทั้งหมดมีค่าการเดาเท่ากับ .20 ส่วนค่าความยากและค่าอำนาจจำแนกของข้อสอบที่ทำหน้าที่ไม่ต่างกันจำนวน 59 ข้อ สุ่มเลือกมาจากค่าสถิติของข้อสอบจากแบบสอบ GMAT ส่วนข้อสอบที่ศึกษาฉบับละ 16 ข้อ จำลองมาจากข้อสอบจำนวน 400 ข้อ ภายใต้เงื่อนไขการแปรเปลี่ยนของค่าพารามิเตอร์ของข้อสอบ ได้แก่ ค่าความยากสำหรับผู้สอบกลุ่มอ้างอิง 5 ระดับ (–1.5, –1.0, 0, 1.0, 1.5) ค่าอำนาจจำแนก 4 ระดับ (.25, .60, .90, 1.25) ความแตกต่างของค่าความยากระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ 4 ระดับ (0, .30, .60, 1.00) และความแตกต่างของค่าอำนาจจำแนกระหว่างกลุ่มผู้สอบทั้งสอง 5 ระดับ (0, .25, .50, .75, 1.0) ดังนั้นจะได้ข้อสอบที่จำลองขึ้นมาจำนวน 400 ข้อ แล้วสุ่มข้อสอบกลุ่มละ 16 ข้อ เพื่อสร้างแบบสอบจำนวน 25 ฉบับ แล้วนำข้อสอบที่เป็นแกนอีก 59 ข้อมารวมกับแบบสอบในแต่ละฉบับดังกล่าว ส่วนการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบจะใช้วิธีแมนเทล-แฮนส์เซลแบบ 2 ขั้นตอน โดยวิเคราะห์ข้อมูล 3 ครั้ง คือ ครั้งแรกวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบวิธีแมนเทล-แฮนส์เซลแบบมาตรฐานโดยใช้กลุ่มผู้สอบทั้งหมด ซึ่งใช้คะแนนรวมของผู้สอบเป็นเกณฑ์ในการจับคู่ระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ครั้งที่สองแบ่งกลุ่มผู้สอบออกเป็นกลุ่มที่มีคะแนนสูงและกลุ่มที่มีคะแนนต่ำ โดยใช้คะแนนเฉลี่ยของผู้สอบทั้งหมดเป็นเกณฑ์ในการแยกกลุ่มผู้สอบดังกล่าว แล้วจึงวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบเฉพาะกลุ่มผู้สอบที่มีคะแนนต่ำ และครั้งที่สามวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบทำนองเดียวกับครั้งที่สองแต่ใช้เฉพาะกลุ่มผู้สอบที่มีคะแนนสูง

ผลการศึกษาพบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปด้วยวิธีแมนเทล-แฮนส์เซลที่แบ่งกลุ่มผู้สอบออกเป็น 2 กลุ่ม คือ กลุ่มที่มีคะแนนสูงและกลุ่มที่มีคะแนนต่ำ มีอำนาจการทดสอบสูงกว่าวิธีแมนเทล-แฮนส์เซลแบบมาตรฐานที่รวมกลุ่มผู้สอบทั้งหมด ทั้งยังไม่ทำให้อัตราความคลาดเคลื่อนประเภทที่ 1 เพิ่มมากขึ้น นอกจากนี้ยังพบว่า ถ้าข้อสอบมีความแตกต่างของค่าอำนาจจำแนกและค่าความยากระหว่างกลุ่มผู้สอบมากขึ้นจะส่งผลให้ระบุข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปมากขึ้นด้วย

Narayanan และ Swaminathan (1994) ได้ศึกษาเปรียบเทียบอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป ระหว่างวิธีแมนเทล-แฮนส์เซลและวิธีซิปเทสท์ โดยใช้ข้อมูลจำลองภายใต้การจัดกระทำ 5 ปัจจัย คือ (1) ขนาดกลุ่มตัวอย่าง 9 ระดับ (กลุ่มอ้างอิงจำนวน 300 คน 500 คน และ 1,000 คน กลุ่มเปรียบเทียบจำนวน 100 คน 200 และ 300 คน) (2) ความแตกต่างของการแจกแจงค่าความสามารถ 2 ระดับ (แบบเท่ากันและแบบไม่เท่ากัน) (3) สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน 2 ระดับ (10% และ 20%) (4) ขนาดอิทธิพลของข้อสอบที่ทำหน้าที่ต่างกัน 4 ระดับ (พื้นที่ระหว่าง IRFs มีค่า .4, .6, .8, 1.0) และ (5) ลักษณะของข้อสอบ 6 ระดับ (b ต่ำกับ a ปานกลาง, b ต่ำกับ a สูง, b ปานกลางกับ a ต่ำ, b ปานกลางกับ a สูง, b สูงกับ a ต่ำ และ b สูงกับ a ปานกลาง) ข้อสอบที่จำลองมีความยาว 40 ข้อ ดังนั้นในการศึกษาครั้งนี้จะต้องจำลองข้อมูลทั้งหมด 1,296 เงื่อนไข ($9 \times 3 \times 2 \times 4 \times 6$)

ผลการศึกษาพบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป ขนาดกลุ่มตัวอย่าง สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน ขนาดอิทธิพลของข้อสอบที่ทำหน้าที่ต่างกัน และลักษณะของข้อสอบเป็นปัจจัยที่มีผลต่ออำนาจการทดสอบของวิธีแมนเทล-แฮนส์เซลและวิธีซิปเทสท์ ถ้าการแจกแจงค่าความสามารถระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบมีค่าเท่ากันแล้ววิธีแมนเทล-แฮนส์เซลและวิธีซิปเทสท์จะมีประสิทธิภาพเท่ากัน แต่ถ้าการแจกแจงค่าความสามารถระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบมีค่าไม่เท่ากันแล้ววิธีซิปเทสท์จะมีประสิทธิภาพสูงกว่าวิธีแมนเทล-แฮนส์เซล นอกจากนี้ยังพบว่า เมื่อขนาดกลุ่มอ้างอิงหรือกลุ่มเปรียบเทียบมีจำนวน 300 คน วิธีทั้งสองก็มีอำนาจการทดสอบเพียงพอในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ส่วนอัตราความคลาดเคลื่อนประเภทที่ 1 พบว่า ขนาดของกลุ่มตัวอย่างและลักษณะของข้อสอบไม่มีผลกระทบต่ออัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีแมนเทล-แฮนส์เซลและวิธีซิปเทสท์ ถ้าการแจกแจงค่าความสามารถระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบมีค่าเท่ากันแล้วอัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีซิปเทสท์มีค่าสูงกว่าวิธีแมนเทล-แฮนส์เซลเล็กน้อย และถ้าความแตกต่างของการแจกแจงค่าความสามารถมีค่าเพิ่มขึ้นจะมีผลทำให้อัตราความคลาดเคลื่อนประเภทที่ 1 ของทั้ง 2 วิธีมีค่าเพิ่มขึ้นด้วย

Uttaro และ Millsap (1994) ได้ศึกษาปัจจัยที่มีผลต่อวิธีแมนเทล-แฮนส์เซล ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปและแบบอเนกรูป โดยการจำลองข้อมูลตามโมเดลโลจิสติกแบบ 1PLM 2PLM และ 3PLM ภายใต้เงื่อนไขของค่าพารามิเตอร์ของข้อสอบและระดับความสามารถของผู้สอบ โดยออกแบบการจำลองภายใต้สถานการณ์ $3 \times 3 \times 2 \times 2$ เงื่อนไข

ตามตัวแปรต้นดังนี้ ค่าอำนาจจำแนก 3 ระดับ (.5, 1.0, 1.5) ค่าความยาก 3 ระดับ (0, .3, .5) ค่าการเดา 2 ระดับ (0, .2) และการแจกแจงค่าความสามารถระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ 2 ระดับ (0, –1.0) ตัวแปรเหล่านี้จะจำลองภายใต้ 2 เงื่อนไข คือ เงื่อนไข no-DIF และเงื่อนไข DIF โดยใช้กลุ่มตัวอย่างกลุ่มละ 500 คน และจำนวนข้อสอบ 20 ข้อ และ 40 ข้อ ภายใต้เงื่อนไข no-DIF ค่าพารามิเตอร์ของ IRFs ระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบจะมีค่าเท่ากัน ในขณะที่ภายใต้เงื่อนไข DIF ค่าพารามิเตอร์ของผู้สอบกลุ่มอ้างอิงจะกำหนดให้มีค่าคงที่ (a = 1.0, b = 0 และ c = 0.2) แต่ค่าพารามิเตอร์ของผู้สอบกลุ่มเปรียบเทียบจะแปรเปลี่ยนไปตามที่ศึกษา

ผลการวิจัยพบว่า ภายใต้เงื่อนไข no-DIF ปัจจัยที่เกี่ยวกับความยาวของแบบสอบ การแจกแจงค่าความสามารถของผู้สอบ ค่าอำนาจจำแนก และค่าการเดาของข้อสอบ จะมีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 1 และการประมาณค่า α_{MH} กล่าวคือ ในแบบสอบขนาด 20 ข้อ จะมีอัตราความคลาดเคลื่อนประเภทที่ 1 สูง และการประมาณค่า α_{MH} ผิดพลาด ซึ่งเป็นผลมาจากปฏิสัมพันธ์ระหว่างค่าพารามิเตอร์ของข้อสอบและความแตกต่างของค่าความสามารถ ส่วนในแบบสอบขนาด 40 ข้อ ไม่พบว่าอัตราความคลาดเคลื่อนประเภทที่ 1 สูง แต่การประมาณค่า α_{MH} ยังคงผิดพลาด สำหรับภายใต้เงื่อนไข DIF พบว่า อัตราความคลาดเคลื่อนประเภทที่ 2 ไม่สูง แต่การประมาณค่า α_{MH} ยังคงผิดพลาด แสดงว่าปฏิสัมพันธ์ระหว่างค่าพารามิเตอร์ของข้อสอบและค่าความสามารถจะมีผลต่อการประมาณค่า α_{MH} ของแบบสอบทั้ง 2 ขนาด แต่จะไม่มีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 2 นอกจากนี้ยังพบว่า ความแม่นยำในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบภายใต้เงื่อนไข DIF ยังขึ้นอยู่กับขนาดและลักษณะที่สม่ำเสมอของข้อสอบที่ทำหน้าที่ต่างกัน

Budgell, Raju และ Quartetti (1995) ได้วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในเครื่องมือการประเมินที่แปลเป็นสองภาษา โดยใช้วิธีการตรวจสอบ 4 วิธี (1) วิธีการวัดพื้นที่ชนิดคิดเครื่องหมาย (SA) (2) วิธีการวัดพื้นที่ชนิดที่ไม่คิดเครื่องหมาย (UA) (3) วิธีการทดสอบไค-สแควร์ของลอร์ด (LC) และ (4) วิธีแมนเทล-แฮนส์เซล (MH) วิธีการวัดพื้นที่ทั้งชนิดคิดเครื่องหมายและชนิดไม่คิดเครื่องหมายเป็นวิธีวัดพื้นที่ในช่วงเปิดของ Raju เครื่องมือที่ใช้ในการศึกษาเป็นแบบสอบชนิดเลือกตอบ ซึ่งวัดด้านตัวเลขจำนวน 15 ข้อ และวัดด้านเหตุผลจำนวน 18 ข้อโดยแยกเป็น 2 ฉบับ คือฉบับภาษาอังกฤษและฉบับภาษาฝรั่งเศส แบบสอบทั้ง 2 ฉบับพัฒนามาจากชุดแบบสอบวัดความรู้ความสามารถทั่วไปที่ใช้ในประเทศแคนาดาโดยคณะผู้เชี่ยวชาญของการแปลภาษา

ทั้งสอง สำหรับการดำเนินการสอบจะต้องควบคุมให้เป็นไปตามเงื่อนไข 2 ประการ คือ ประการแรก ผู้สอบต้องเลือกแบบสอบให้ตรงกับภาษาที่ใช้มาตั้งแต่กำเนิด และประการที่สอง ผู้สอบที่อาศัยอยู่ในประเทศแคนาดาจะต้องเลือกแบบสอบให้ตรงกับภาษาหลักที่ใช้กันในชุมชน มีผู้เข้าสอบทั้งหมด 16,362 คน ต่อจากนั้นจะสุ่มกลุ่มตัวอย่างมา 4 กลุ่ม ๆ ละ 1,000 คน กลุ่มผู้สอบรูปแบบภาษาอังกฤษ 2 กลุ่ม (E_1 และ E_2) และกลุ่มผู้สอบรูปแบบภาษาฝรั่งเศส 2 กลุ่ม (F_1 และ F_2) นำผู้สอบทั้ง 4 กลุ่มไปจับคู่เปรียบเทียบเพื่อวิเคราะห์หาดัชนี DIF โดยเปรียบเทียบกลุ่มผู้สอบ 4 เงื่อนไข คือ (1) E_1 กับ F_1 (2) E_2 กับ F_2 (3) E_1 กับ E_2 และ (4) F_1 กับ F_2 ในการวิเคราะห์แบบสอบทั้ง 2 ฉบับจะวิเคราะห์แยกกันทั้งรูปแบบภาษา (อังกฤษและภาษาฝรั่งเศส) และเนื้อหาที่ใช้วัด (ตัวเลขและเหตุผล) นำแบบสอบทั้ง 2 ฉบับ ไปวิเคราะห์องค์ประกอบโดยใช้การวิเคราะห์องค์ประกอบหลัก (principal component analysis; PCA) เพื่อตรวจสอบความเป็นเอกมิติก่อนที่จะนำไปวิเคราะห์ตามทฤษฎี IRT ใช้โปรแกรม BILOG ประมาณค่าพารามิเตอร์ของข้อสอบตามโมเดลโลจิสติกแบบ 2 พารามิเตอร์ และใช้โปรแกรม EQUATE เทียบค่าพารามิเตอร์ของข้อสอบจากผู้สอบกลุ่มย่อย 2 กลุ่ม สำหรับสถิติที่ใช้ทดสอบดัชนี DIF ประกอบด้วย สถิติ Z ทดสอบดัชนี SA และ UA ($Z < -2.58$ หรือ $Z > +2.58$ เมื่อ $\alpha = .01$) สถิติ χ^2 ทดสอบดัชนี LC (χ^2 เท่ากับ 9.21 เมื่อ $\alpha = .01$, $df = 2$) และสถิติ χ^2 ทดสอบดัชนี MH (χ^2 เท่ากับ 6.63 เมื่อ $\alpha = .01$, $df = 1$)

ผลการศึกษาพบว่า การเปรียบเทียบตามเงื่อนไข (1) และ (2) การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี SA, UA, LC และ MH สามารถระบุข้อสอบที่ทำหน้าที่ต่างกันได้สอดคล้องกันอย่างมีนัยสำคัญ โดยเฉพาะในแบบสอบวัดด้านตัวเลขและแบบสอบวัดด้านเหตุผลทั้ง 4 วิธี สามารถระบุข้อสอบที่ทำหน้าที่ต่างกันได้สอดคล้องกันสูง ยกเว้นการตรวจสอบด้วยวิธี UA ในแบบสอบวัดด้านเหตุผล สามารถระบุข้อสอบที่ทำหน้าที่ต่างกันต่ำกว่าการตรวจสอบด้วยวิธีอื่น ๆ สำหรับการเปรียบเทียบตามเงื่อนไข (3) และ (4) ปรากฏว่า ไม่พบข้อสอบที่ทำหน้าที่ต่างกัน

Narayanan และ Swaminathan (1996) ได้ศึกษาเปรียบเทียบวิธีแมนเทล-แฮนส์เซล (MH) วิธีโคร-ซิบ (CRO-SIB) และวิธีการถดถอยโลจิสติก (LR) ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป โดยศึกษาในสถานการณ์จำลองภายใต้เงื่อนไข 384 เงื่อนไข ($4 \times 2 \times 3 \times 4 \times 4$) ตามปัจจัยที่แปรเปลี่ยนดังนี้ ขนาดกลุ่มตัวอย่าง 4 ระดับ (กลุ่มอ้างอิงขนาด 500 คน และ1,000 คน กลุ่มเปรียบเทียบขนาด 200 คน และ 500 คน) สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน 3 ระดับ (0%, 10% และ 20%) ความแตกต่างของการแจกแจงค่าความสามารถ 2 ระดับ (แบบเท่ากันและแบบไม่เท่ากัน) ขนาดอิทธิพลของข้อสอบที่ทำหน้าที่ต่างกัน 4 ระดับ (พื้นที่ระหว่าง

IRFs มีค่าเท่ากับ .4, .6, .8 และ 1.0) ลักษณะของข้อสอบ 4 ระดับ (b ต่ำกับ a สูง, b ปานกลางกับ a ต่ำ, b ปานกลางกับ a สูง และ b สูงกับ a ต่ำ) สำหรับความยาวของแบบสอบใช้เพียง 40 ข้อ ข้อมูลที่ใช้ในการศึกษาจำลองตามโมเดลโลจิสติกแบบ 3 พารามิเตอร์ โดยใช้โปรแกรม DATAGEN สำหรับดัชนีการตรวจสอบด้วยวิธี MH และ วิธี LR ใช้โปรแกรม DICHODIF ส่วนวิธี CRO-SIB ใช้โปรแกรม CSIBTEST แล้วใช้การวิเคราะห์ ANOVA แบบ 5 ทิศทาง เพื่อทดสอบผลกระทบของปัจจัยที่ตรวจสอบด้วยวิธีทั้งสาม โดยทดสอบที่ระดับนัยสำคัญ .05 และ .01

ผลการศึกษาพบว่า วิธีโคร-ซิปและวิธีการถดถอยโลจิสติกมีประสิทธิภาพเท่าเทียมกันในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปภายใต้เกือบทุกเงื่อนไข ในขณะที่วิธีแมนเทล-แฮนส์เซลไม่มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนก-รูปที่ไม่มีทิศทางหรือมีปฏิสัมพันธ์แบบไม่เป็นลำดับ สำหรับปัจจัยของขนาดกลุ่มผู้สอบ ลักษณะของข้อสอบ และขนาดอิทธิพลของข้อสอบที่ทำหน้าที่ต่างกันมีผลต่ออำนาจการทดสอบของทั้ง 3 วิธี กล่าวคือ เมื่อขนาดกลุ่มตัวอย่างและขนาดอิทธิพลของข้อสอบที่ทำหน้าที่ต่างกันมีค่าเพิ่มขึ้นจะมีผลทำให้อำนาจการทดสอบของทั้ง 3 วิธีมีค่าเพิ่มมากขึ้น และเมื่อลักษณะของข้อสอบมีค่าอำนาจจำแนกสูงแล้วอำนาจการทดสอบของวิธีโคร-ซิปและวิธีการถดถอยโลจิสติกจะมีค่าเพิ่มขึ้น แต่เมื่อข้อสอบมีค่าความยากสูงหรือต่ำแล้วอำนาจการทดสอบของวิธีแมนเทล-แฮนส์เซลจะมีค่าเพิ่มขึ้น สำหรับอัตราความคลาดเคลื่อนประเภทที่ 1 พบว่า วิธีโคร-ซิปและวิธีการถดถอยโลจิสติกมีอัตราความคลาดเคลื่อนประเภทที่ 1 สูงกว่าวิธีแมนเทล-แฮนส์เซลภายใต้ทุกเงื่อนไข นอกจากนี้ยังพบว่า อัตราความคลาดเคลื่อนประเภทที่ 1 ของทั้ง 3 วิธีมีค่าสูง เมื่อลักษณะของข้อสอบมีค่าอำนาจจำแนกสูง

Roussos และ Stout (1996) ได้ศึกษาในสถานการณ์จำลองของผลกระทบของกลุ่มตัวอย่างขนาดเล็กและค่าพารามิเตอร์ของข้อสอบที่มีต่ออัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีซิปเทสท์และวิธีแมนเทล-แฮนส์เซล โดยจำลองข้อมูล 2 ครั้ง ครั้งที่ 1 เพื่อศึกษากลุ่มตัวอย่างขนาดเล็ก โดยใช้ขนาดกลุ่มตัวอย่าง 4 ระดับ (100 คน 200 คน 500 คน และ 1,000 คน) และความแตกต่างของการแจกแจงค่าความสามารถระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ (d_T) 3 ระดับ (0, 0.5 และ 1.0) ใช้แบบสอบจำนวน 25 ข้อ ซึ่งนำมาจากชุดแบบสอบ Armed Services Vocational Aptitude Battery (ASVAB) ประมาณค่าพารามิเตอร์ใช้โมเดลโลจิสติกแบบ 3 พารามิเตอร์ สำหรับการจำลองครั้งที่ 2 เพื่อศึกษาค่าพารามิเตอร์ของข้อสอบ โดยใช้ขนาดกลุ่มตัวอย่าง 3 ระดับ (500 คน 1,000 คน และ 3,000 คน) และความแตกต่างของการแจกแจงค่า

ความสามารถระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ (d_T) 2 ระดับ (0 และ 1.0) ส่วนแบบสอบใช้จากการศึกษาครั้งที่ 1 โดยเลือกค่าพารามิเตอร์อำนาจจำแนก 3 ระดับ (0.4, 1.0, 2.5) พารามิเตอร์ความยาก 5 ระดับ (–1.5, –0.5, 0, 0.5 และ 1.5) พารามิเตอร์การเดา 1 ระดับ (0.20) เมื่อ d_T = 0.0 และพารามิเตอร์การเดา 3 ระดับ (0.20, 0.10 และ 0.05) เมื่อ d_T = 1.0 สำหรับขนาดกลุ่มตัวอย่างระหว่างกลุ่มเปรียบเทียบและกลุ่มอ้างอิงจะใช้จำนวนเท่ากันทั้งสองครั้ง

ผลการศึกษาครั้งที่ 1 เมื่อศึกษากลุ่มตัวอย่างขนาดเล็ก พบว่า อัตราความคลาดเคลื่อนประเภทที่ 1 ระหว่างวิธีซิปเทสท์และวิธีแมนเทล-แฮนส์เซลมีค่าไม่แตกต่างกัน สำหรับผลการศึกษาครั้งที่ 2 เมื่อศึกษาค่าพารามิเตอร์ของข้อสอบ พบว่า เมื่อความแตกต่างของการแจกแจงค่าความสามารถระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบไม่แตกต่างกัน (d_T = 0) แล้วอัตราความคลาดเคลื่อนประเภทที่ 1 ระหว่างวิธีซิปเทสท์และวิธีแมนเทล-แฮนส์เซลมีค่าใกล้เคียงกัน แต่ถ้าความแตกต่างของการแจกแจงค่าความสามารถระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบแตกต่างกัน (d_T = 1.0) แล้วอัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีซิปเทสท์จะมีค่าต่ำกว่าวิธีแมนเทล-แฮนส์เซล ภายใต้ทุกเงื่อนไขของการตรวจสอบ

2. **งานวิจัยในประเทศ**

กาญจนา วัธนสุนทร (2537) ได้พัฒนาเกณฑ์ในการตัดสินข้อสอบลำเอียงทางเพศด้วยข้อมูลเชิงประจักษ์ โดยใช้วิธีการตรวจสอบ 4 วิธี คือ วิธีการวัดพื้นที่ชนิดคิดเครื่องหมาย (SA) วิธีการวัดพื้นที่ชนิดไม่คิดเครื่องหมาย (UA) วิธีแมนเทล-แฮนส์เซล (MH) และวิธีซิปเทสท์ (SIBTEST) ข้อมูลที่ใช้ในการศึกษาครั้งนี้ได้จากการตอบข้อสอบคัดเลือกบุคคลเข้าศึกษาในสถาบันอุดมศึกษาของทบวงมหาวิทยาลัย ปีการศึกษา 2535 โดยศึกษาภายใต้เงื่อนไขของปัจจัยที่แปรเปลี่ยน 2 ตัว คือ ความยาวของแบบสอบ ซึ่งมี 2 วิชา ได้แก่ แบบสอบวิชาคณิตศาสตร์ 3 ระดับ (20 ข้อ 30 ข้อ และ 40 ข้อ) แบบสอบวิชาภาษาอังกฤษ 4 ระดับ (50 ข้อ 60 ข้อ 70 ข้อ และ 80 ข้อ) และขนาดกลุ่มตัวอย่าง 6 ระดับ (100 คน 200 คน 400 คน 600 คน 800 คน และ 1,000 คน) ในการประมาณค่าพารามิเตอร์ของข้อสอบใช้โปรแกรม BILOG โมเดลโลจิสติกแบบ 2 PLM โดยใช้วิธีการประมาณค่าแบบ MMLE แล้วใช้โปรแกรม EQUATE ปรับเทียบสเกลพารามิเตอร์ของข้อสอบที่ประมาณค่าจากผู้สอบสองกลุ่ม คือ กลุ่มอ้างอิงและกลุ่มเปรียบเทียบซึ่งจำแนกตามเพศ ในการคำนวณดัชนี DIF ด้วยวิธีการวัดพื้นที่ชนิดคิดเครื่องหมายและชนิดไม่คิดเครื่องหมายใช้โปรแกรม

AREA ส่วนวิธี MH และวิธี SIBTEST ใช้โปรแกรม SIBTEST ในการพัฒนาดัชนีเพื่อใช้เป็นเกณฑ์สำหรับการตัดสินความลำเอียงของข้อสอบมี 4 ตัว คือ SA, UA, α_{MH} และ β_{SIB} โดยจะวิเคราะห์ค่าเฉลี่ยของดัชนีแต่ละตัวแล้วกำหนดเกณฑ์จากค่าเฉลี่ย 2 ลักษณะ คือ เกณฑ์ที่กำหนดจากค่าเฉลี่ยซึ่งรวมค่าดัชนีทุกข้อโดยไม่ได้พิจารณาถึงปัจจัยความยาวของแบบสอบและขนาดกลุ่มตัวอย่าง ส่วนเกณฑ์อีกลักษณะหนึ่งจะกำหนดจากค่าเฉลี่ยซึ่งได้พิจารณาถึงปัจจัยความยาวของแบบสอบและขนาดของกลุ่มตัวอย่าง แล้วจึงนำเกณฑ์ที่กำหนดไว้ไปตัดสินค่าดัชนีที่ได้จากการวิเคราะห์ระหว่างกลุ่มผู้สอบเพศชายและหญิง

ผลการวิจัยพบว่า ขนาดกลุ่มตัวอย่างมีผลต่อค่าเฉลี่ยของดัชนีทุกตัว ส่วนความยาวของแบบสอบมีผลต่อค่าเฉลี่ยของดัชนี SA และ UA แต่ไม่มีอิทธิพลต่อค่าเฉลี่ยของดัชนี α_{MH} และ β_{SIB} สำหรับเกณฑ์ที่พัฒนาเพื่อใช้ตัดสินความลำเอียงของข้อสอบระหว่างผู้สอบเพศชายและเพศหญิงเป็น ดังนี้

1) $|SA| > .80$ และ $UA > .50$ เมื่อความยาวของแบบสอบน้อยกว่า 50 ข้อ

2) $|SA| > .40$ และ $UA > 1.20$ เมื่อวามยาวของแบบสอบมากกว่า 50 ข้อ

3) $\alpha_{MH} < .60$, $\alpha_{MH} > 1.40$ และ $|\beta_{SIB}| > .06$ สำหรับทุกความยาวของแบบสอบและขนาดกลุ่มตัวอย่าง

เกษร หว่างจิตร (2539) ได้วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมนเทล-แฮนส์เซล โดยใช้เพศ ภูมิลำเนา ประสบการณ์ในการสอบ และสังกัดของสถานศึกษา เป็นเกณฑ์จำแนกกลุ่มผู้สอบเป็นกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ข้อมูลที่ใช้ในการศึกษาเป็นผลการสอบในวิชาสอบร่วมของศูนย์ทดสอบทางการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในส่วนเฉพาะที่เป็นข้อสอบชนิดเลือกตอบ 2 วิชา คือ วิชาภาษาไทยซึ่งมีผู้สอบจำนวน 506 คน และวิชาภาษาอังกฤษซึ่งมีผู้สอบจำนวน 501 คน แล้วนำผลการสอบมาวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปและแบบอเนกรูปโดยใช้โปรแกรม MH_{DIF} สำหรับการตรวจสอบความเที่ยงและความตรงของแบบสอบใช้โปรแกรม CTIA และ LISREL ตามลำดับ

ผลการศึกษาพบว่า ข้อสอบที่ถูกระบุว่าทำหน้าที่ต่างกันส่วนมากมีลักษณะเป็นแบบอเนกรูป โดยพบในกลุ่มผู้สอบย่อยที่จำแนกตามเพศมากที่สุด รองลงมาคือ จำแนกตามภูมิลำเนาสังกัดของสถานศึกษา และประสบการณ์ในการสอบตามลำดับ ข้อสอบที่ทำหน้าที่ต่างกันส่วนมากเป็นข้อสอบที่มีค่าอำนาจจำแนกค่อนข้างต่ำทั้งสองวิชา เมื่อพิจารณาลักษณะของข้อสอบ พบว่าในแบบสอบวิชาภาษาไทยข้อสอบที่ถูกระบุว่าทำหน้าที่ต่างกันส่วนมากจะเป็นข้อสอบที่ง่ายมาก

แต่ในแบบสอบวิชาภาษาอังกฤษข้อสอบที่ถูกระบุว่าทำหน้าที่ต่างกันส่วนมากจะเป็นข้อสอบที่ยากมาก นอกจากนี้ยังพบว่า ค่าความเที่ยงและความตรงของแบบสอบก่อนและหลังการตัดข้อสอบที่ทำหน้าที่ต่างกันออกจากแบบสอบไม่แตกต่างกันทั้งสองวิชา

เสรี ชัดแช้ม (2539) ได้เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป ระหว่างวิธีแมนเทล-แฮนส์เซลแบบปกติกับวิธีแมนเทล-แฮนส์เซลแบบแบ่งกลุ่มความสามารถของผู้สอบและความยากของข้อสอบ โดยใช้กลุ่มตัวอย่างขนาด 1,200 คน แบ่งเป็นเพศชายซึ่งเป็นกลุ่มเปรียบเทียบจำนวน 600 คน และเพศหญิงซึ่งเป็นกลุ่มอ้างอิงจำนวน 600 คน เครื่องมือที่ใช้ในการเก็บรวบรวมข้อมูลเป็นแบบสอบวัดความสามารถในการอ่านภาษาไทย 75 ข้อ ในการประมาณค่าพารามิเตอร์ของข้อสอบใช้โปรแกรม BILOG ตามโมเดลโลจิสติกแบบ 2 PLM แล้วใช้โปรแกรม EQUATE ปรับเทียบสเกลพารามิเตอร์ของข้อสอบระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ต่อจากนั้นจึงคำนวณดัชนี DIF ด้วยวิธี IRTarea แบบ 2 PLM เพื่อใช้เป็นเกณฑ์สำหรับการเปรียบเทียบระหว่างวิธีแมนเทล-แฮนส์เซล 2 วิธี ส่วนการวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปตามวิธีแมนเทล-แฮนส์เซลแบบปกติกับวิธีแมนเทล-แฮนส์เซลแบบแบ่งกลุ่มความสามารถของผู้สอบและความยากของข้อสอบใช้โปรแกรม MHDIF

ผลการศึกษาพบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป วิธีแมนเทล-แฮนส์เซลแบบปกติสามารถระบุข้อสอบที่ทำหน้าที่ต่างกันสอดคล้องกับวิธี IRTarea คิดเป็นร้อยละ 33.33 ส่วนวิธีแมนเทล-แฮนส์เซลแบบแบ่งกลุ่มความสามารถของผู้สอบและความยากของข้อสอบสามารถระบุข้อสอบที่ทำหน้าที่ต่างกันสอดคล้องกับวิธี IRT area คิดเป็นร้อยละ 61.11

จิติมา วรรณศรี (2539) ได้เปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมนเทล-แฮนส์เซลและวิธีซิปเทสท์ โดยศึกษาในสถานการณ์จำลอง ใช้โปรแกรม IRTDATA จำลองข้อมูลตามปัจจัยที่ศึกษา ได้แก่ ความยาวของแบบสอบ 3 ระดับ (30 ข้อ 60 ข้อ และ 90 ข้อ) ขนาดกลุ่มตัวอย่าง 3 ระดับ (200 คน 600 คน และ 1,000 คน) โดยในแต่ละขนาดมีสัดส่วนระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ 4 ลักษณะ คือ 1 : 1, 1 : .9, 1 : .75 และ 1 : .5 รวมเงื่อนไขที่ต้องจำลองทั้งหมดจำนวน 36 เงื่อนไข ($3 \times 3 \times 4$) ในการวิเคราะห์ข้อมูลใช้โปรแกรม BILOG ประมาณค่าพารามิเตอร์ของข้อสอบตามโมเดลโลจิสติกแบบ 2 PLM โดยเลือกวิธีการประมาณค่าแบบ MBE แล้วปรับเทียบสเกลพารามิเตอร์ของข้อสอบระหว่างผู้สอบ 2 กลุ่มโดยใช้โปรแกรม EQUATE ต่อจากนั้นจึงคำนวณดัชนี DIF ด้วยวิธีการวัดพื้นที่ของ Raju โดยใช้โปรแกรม AREA เพื่อใช้เป็นเกณฑ์สำหรับการเปรียบเทียบข้อสอบที่ทำหน้าที่ต่างกัน ส่วนการคำนวณดัชนี DIF ด้วยวิธีแมนเทล-แฮนส์เซลและวิธีซิปเทสท์ใช้โปรแกรม SIBTEST

ผลการศึกษาพบว่า วิธีแมนเทล-แฮนส์เซลและวิธีซิบเทสท์มีประสิทธิภาพเท่าเทียมกันในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ภายใต้เกือบทุกเงื่อนไขของการตรวจสอบ ส่วนปัจจัยของขนาดกลุ่มตัวอย่างและความยาวของแบบสอบ ปรากฏว่า มีผลต่ออัตราการตรวจสอบกล่าวคือ เมื่อขนาดกลุ่มตัวอย่าง 200 คน และ 600 คน สามารถระบุข้อสอบที่ทำหน้าที่ต่างกันได้ถูกต้องร้อยละ 50 แต่เมื่อขนาดกลุ่มตัวอย่าง 1,000 คน สามารถระบุข้อสอบที่ทำหน้าที่ต่างกันได้ถูกต้องร้อยละ 100 และเมื่อความยาวของแบบสอบ 60 ข้อ (ขนาดปานกลาง) จะมีผลทำให้วิธีตรวจสอบทั้งสองมีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบดีที่สุด

รัชรินทร์ มุคดา (2540) ได้เปรียบเทียบประสิทธิภาพระหว่างวิธีแมนเทล-แฮนส์เซลและวิธีถดถอยโลจิสติก ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป ข้อมูลที่ใช้ในการศึกษาเป็นข้อมูลจำลองโดยใช้โปรแกรม IRTDATA จำลองข้อมูลทั้งหมด 27 เงื่อนไข (3 × 3 × 3) คือ กลุ่มความสามารถผู้สอบ 3 ระดับ (สูง ปานกลาง และต่ำ) ค่าความยากของข้อสอบ 3 ระดับ (สูง ปานกลาง และต่ำ) ค่าอำนาจจำแนกของข้อสอบ 3 ระดับ (สูง ปานกลาง และต่ำ) ในการประมาณค่าพารามิเตอร์ของข้อสอบใช้โปรแกรม BILOG ตามโมเดลโลจิสติกแบบ 2 PLM แล้วใช้โปรแกรม EQUATE ปรับเทียบสเกลพารามิเตอร์ของข้อสอบระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ต่อจากนั้นจึงคำนวณดัชนี DIF ด้วยวิธีการวัดพื้นที่ของ Raju โดยใช้โปรแกรม AREA เพื่อใช้เป็นเกณฑ์สำหรับการเปรียบเทียบข้อสอบที่ทำหน้าที่ต่างกัน ส่วนการคำนวณดัชนี DIF ด้วยวิธีแมนเทล-แฮนส์เซลและวิธีการถดถอยโลจิสติกใช้โปรแกรม MH-DIF และโปรแกรม SPSS/PC^{+} ตามลำดับ

ผลการศึกษาพบว่า ในกลุ่มผู้สอบที่มีความสามารถสูง ปานกลางและต่ำ วิธีแมนเทล-แฮนส์เซลและวิธีถดถอยโลจิสติกมีประสิทธิภาพเท่าเทียมกันในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป สำหรับปัจจัยของลักษณะของข้อสอบที่เกี่ยวกับค่าความยากของข้อสอบ พบว่า ข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปซึ่งตรวจพบมากที่สุดในกลุ่มผู้สอบที่มีความสามารถสูง ปานกลาง และต่ำ เป็นข้อสอบที่มีค่าความยากสูง ปานกลาง และต่ำตามลำดับ ส่วนลักษณะของข้อสอบที่เกี่ยวกับค่าอำนาจจำแนก พบว่า ข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปซึ่งตรวจพบมากที่สุดในทุกกลุ่มผู้สอบเป็นข้อสอบที่มีค่าอำนาจจำแนกสูง