บทที่ 2

เอกสารและงานวิจัยที่เกี่ยวข้อง

ผู้วิจัยเสนอแนวคิดและทฤษฎีที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันโดยแบ่งการนำเสนอแบ่งออกเป็น 5 ตอน คือ

ตอนที่ 1 ความเป็นมาของการทำหน้าที่ต่างกันของข้อสอบ

ตอนที่ 2 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการตอบสนองข้อสอบ (IRT)

ตอนที่ 3 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมนเทล-แฮนส์เซล (MH)

ตอนที่ 4 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีวิเคราะห์องค์ประกอบจำกัด (RFA)

ตอนที่ 5 งานวิจัยที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

ตอนที่ 1 ความเป็นมาของการทำหน้าที่ต่างกันของข้อสอบ

การศึกษาเรื่องความยุติธรรมของข้อสอบมีการศึกษากันอย่างจริงจังในช่วงปลายทศวรรษ 1960 โดยมีการเสนอวิธีการต่าง ๆ เพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning : DIF) การทำหน้าที่ต่างกันของข้อสอบเดิมเรียกว่า "ความลำเอียงของข้อสอบ" (item bias) ในระยะหลังได้มีการศึกษาเรื่องความลำเอียงของข้อสอบกันอย่างกว้างขวาง และมีการเสนอวิธีการใหม่ ๆ ที่ใช้ในการตรวจสอบความลำเอียงของข้อสอบ วิธีการต่าง ๆ เหล่านี้จะเน้นไปที่ความแตกต่างระหว่างกลุ่มผู้สอบที่ตอบสนองต่างกันต่อข้อสอบข้อเดียวกัน ความแตกต่างที่เกิดขึ้นดังกล่าวอาจมาจากข้อคำถาม ประสบการณ์หรือพื้นฐานเดิมที่แตกต่างกันของกลุ่มผู้สอบ ซึ่งในบางสถานการณ์ก็ไม่เหมาะสมที่จะใช้คำว่า "ลำเอียง" จึงทำให้เกิดความสับสนในการใช้คำและความหมาย ด้วยเหตุนี้จึงควรใช้คำว่า "ข้อสอบทำหน้าที่ต่างกัน" เพราะเป็นคำที่มีความหมายกลาง ๆ มากกว่าและเหมาะสมกว่า (Holland and Thayer, 1988: Green ,1994 อ้างถึงใน จิติมา วรรณศรี, 2539) นอกจากนี้ Camilli และ Shepard (1994) เสนอแนวคิดว่า การทำหน้าที่ต่างกันของข้อสอบเน้นที่วิธีการทางสถิติที่นำมาตรวจสอบข้อสอบทำหน้าที่ต่างกันจากผู้สอบต่างกลุ่ม ส่วนความลำเอียงของข้อสอบถ้าเน้นที่วิธีการตรวจสอบทางสถิติเพียงอย่างเดียวยังสรุปไม่ได้ว่าข้อสอบลำเอียง อาจต้องรวมถึงการวิเคราะห์เชิงตรรกะ (logical analysis) โดยอาศัยผู้เชี่ยวชาญพิจารณาเนื้อหาสาระของข้อสอบและจุดมุ่งหมายในการวัดของแบบสอบก่อนที่จะระบุว่าข้อสอบนั้นลำเอียง

เนื้อหาสาระของข้อสอบและจุดมุ่งหมายในการวัดของแบบสอบก่อนที่จะระบุว่าข้อสอบนั้นลำเอียงหรือไม่ ดังนั้น ข้อสอบที่ระบุว่ามีความลำเอียงจึงเป็นกลุ่มข้อสอบย่อย (subset) ของข้อสอบที่ทำหน้าที่ต่างกัน ปัจจุบันพบว่านักวัดผลทางการศึกษาส่วนใหญ่นิยมใช้คำว่า "การทำหน้าที่ต่างกันของข้อสอบ" ซึ่งมีผู้ให้ความหมายไว้ดังนี้

Kederman และ Macready (1990) กล่าวว่า การทำหน้าที่ต่างกันของข้อสอบ เป็นคะแนนข้อสอบที่ได้จากกลุ่มผู้สอบที่มีความสามารถเท่ากันแต่มาจากต่างกลุ่มกันมีความแตกต่างกันอย่างมีระบบ

Camilli และ Shepard (1994) กล่าวว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นการตรวจสอบความเป็นพหุมิติของข้อสอบ โดยจะแสดงการทำหน้าที่ต่างกันระหว่างกลุ่มผู้สอบตั้งแต่สองกลุ่มขึ้นไปที่มีความสามารถหลัก (primary abilities) เท่ากัน แต่มีความสามารถรอง (secondary abilities) แตกต่างกัน

Potenza และ Dorans (1995) กล่าวว่า การทำหน้าที่ต่างกันของข้อสอบ หมายถึงผลการตอบข้อสอบระหว่างกลุ่มผู้สอบสองกลุ่มที่นำมาเปรียบเทียบมีความแตกต่างกัน การเปรียบเทียบกลุ่มผู้สอบเป็นสิ่งสำคัญที่จะอธิบายถึงความแตกต่างระหว่างการทำหน้าที่ของข้อสอบกับคุณลักษณะแฝงของกลุ่มผู้สอบ

Narayanan และ Swaminathan (1996) กล่าวว่า การทำหน้าที่ต่างกันของข้อสอบ หมายถึง ผู้สอบมีความสามารถระดับเดียวกันแต่มาจากกลุ่มย่อยต่างกัน มีโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกัน

ดังนั้น จึงสรุปได้ว่า การทำหน้าที่ต่างกันของข้อสอบ เกิดขึ้นเมื่อข้อสอบหรือแบบสอบวัดคุณลักษณะแฝงอื่นนอกเหนือจากคุณลักษณะแฝงที่ต้องการวัด ทำให้ผู้สอบแต่ละกลุ่มที่นำมาจับคู่เปรียบเทียบมีโอกาสตอบข้อสอบถูกแตกต่างกันทั้งๆ ที่มีความสามารถที่ต้องการวัดเท่ากัน

ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบดำเนินการโดยเปรียบเทียบผลการตอบข้อสอบข้อเดียวกันระหว่างผู้สอบ 2 กลุ่ม คือกลุ่มอ้างอิง (reference group) และกลุ่มเปรียบเทียบ (focal group) กลุ่มอ้างอิงเป็นกลุ่มที่คาดว่าจะได้ประโยชน์ในการตอบข้อสอบคือ มีโอกาสในการตอบข้อสอบถูกได้มากกว่าผู้สอบอีกกลุ่มหนึ่ง ส่วนกลุ่มเปรียบเทียบเป็นกลุ่มที่สนใจศึกษาเป็นกลุ่มที่คาดว่าจะเสียประโยชน์ในการตอบข้อสอบ เช่น การศึกษาการทำหน้าที่ต่างกันของข้อสอบระหว่างผู้สอบต่างเชื้อชาติ กลุ่มเปรียบเทียบได้แก่ กลุ่มผู้สอบผิวดำ ในขณะที่กลุ่มอ้างอิงได้แก่ กลุ่มผู้สอบผิวขาวเป็นต้น ในการเปรียบเทียบจะศึกษาปัจจัยอันเกิดจากผู้สอบซึ่งส่งผลให้เกิดการได้ประโยชน์และเสียประโยชน์ระหว่างกลุ่มผู้สอบ เช่น เพศ สีผิว เชื้อชาติ ภาษา สถาบันการศึกษา ประสบการณ์ เป็นต้น ต่อมาระยะหลังได้มีการศึกษาเปรียบเทียบวิธีการต่าง ๆ ในการตรวจสอบ

การทำหน้าที่ต่างกันของข้อสอบ ทั้งนี้เพราะมีวิธีการตรวจสอบหลายวิธีที่ถูกคิดค้นและพัฒนาปรับปรุง เพื่อให้สามารถตรวจสอบการทำหน้าที่ต่างกันได้อย่างมีประสิทธิภาพมากที่สุด

วิธีการในการตรวจสอบทำหน้าที่ต่างกันของข้อสอบมีอยู่ด้วยกันหลายวิธี ทั้งนี้เพราะมีการศึกษาและคิดค้นวิธีการต่าง ๆ เพื่อให้สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้อย่างมีประสิทธิภาพมากขึ้น ซึ่งสามารถแบ่งตามประเภทการวิเคราะห์ได้ดังนี้

กลุ่มวิธีการตอบสนองข้อสอบ (IRT) เป็นวิธีที่วิเคราะห์ความแตกต่างของฟังก์ชันการตอบสนองข้อสอบโดยเปรียบเทียบโค้งคุณลักษณะข้อสอบ (item characteristic curves : ICCs) ของกลุ่มผู้สอบตามระดับความสามารถของผู้สอบ ถ้าโค้งคุณลักษณะข้อสอบของกลุ่มผู้สอบสองกลุ่มมีรูปร่างเหมือนกันแสดงว่าข้อสอบนั้นทำหน้าที่ไม่ต่างกัน แต่ถ้าโค้งคุณลักษณะข้อสอบของผู้สอบสองกลุ่มมีรูปร่างต่างกันแสดงว่าข้อสอบนั้นทำหน้าที่ต่างกัน กระบวนการวิเคราะห์มีดัชนีบอกระดับการทำหน้าที่ต่างกันและทดสอบนัยสำคัญทางสถิติ ซึ่งได้รับการพัฒนาและแตกย่อยออกไปหลายวิธี เช่น วิธี General IRTLR วิธี Loglinear IRTLR วิธี IRT-D^2 และวิธี Lord's χ^2 จากการศึกษาของนักวัดผลทางการศึกษา พบว่าวิธี Lord's χ^2 เป็นวิธีที่มีความถูกต้องสูงสุด สามารถตรวจสอบพบจำนวนข้อสอบที่ทำหน้าที่ต่างกันมากที่สุด (Holland and Thater, 1988) ค่าสถิติของข้อสอบไม่เปลี่ยนไปตามกลุ่มตัวอย่าง และใช้การประมาณค่าความสามารถที่แท้จริงของผู้สอบแทนคะแนนที่สังเกตได้ แต่มีข้อจำกัดคือ ต้องใช้กลุ่มตัวอย่างขนาดใหญ่ ข้อมูลต้องเป็นไปตามข้อตกลงเบื้องต้น และถ้าใช้ข้อมูลจำลองในการศึกษาต้องสร้างขึ้นภายใต้ทฤษฎี IRT มีการคำนวณซับซ้อนหลายรอบ แปลผลยาก เสียค่าใช้จ่ายสูง (Ryan, 1991 ; Osterlind, 1993 ; Narayanan and Swaminatan, 1994 อ้างถึงใน สุรศักดิ์ อมรรัตนศักดิ์, 2531)

วิธีแมนเทล-แฮนส์เซล (MH) เป็นวิธีที่พัฒนามาจากวิธีไคสแควร์แบบดั้งเดิม ใช้คะแนนรวมจากแบบสอบเป็นตัวแทนความสามารถของผู้สอบ การวิเคราะห์จะวิเคราะห์ทีละระดับความสามารถ มีดัชนีบอกระดับการทำหน้าที่ต่างกันของข้อสอบและการทดสอบนัยสำคัญทางสถิติ เป็นวิธีที่มีความสอดคล้องกับวิธี IRT อีกทั้งสามารถใช้วิธี MH แทนวิธี IRT (Hambleton et al., 1986 อ้างถึงใน กาญจนา วัชนสุนทร, 2537) ข้อดีของวิธีนี้ คือคำนวณง่าย ใช้ได้กับกลุ่มตัวอย่างขนาดเล็กและประหยัดค่าใช้จ่าย

วิธีวิเคราะห์องค์ประกอบจำกัด (RFA) ใช้คะแนนรวมจากการสอบเป็นตัวแทนความสามารถ วิธีนี้จะมีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการตอบเป็นลักษณะต่อเนื่องหรือเป็นมาตรวัดที่มีหลายระดับ แต่ในแบบสอบที่มีการตอบแบบ 2 ค่า (ตอบถูกได้ 1 และตอบผิดได้ 0) วิธีนี้ให้ผลการวิเคราะห์ได้ใกล้เคียงกับวิธี IRT จึงน่าจะใช้แทนวิธี IRT ได้ (Oort, 1998) และวิธีนี้ไม่มีข้อจำกัดในเรื่องขนาดของกลุ่มตัวอย่าง ประหยัดเวลาและค่าใช้จ่ายในการวิเคราะห์

ตอนที่ 2 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีทฤษฎีการตอบสนองข้อสอบ (Item Response Theory procedure : IRT)

ทฤษฎี IRT ถูกคิดขึ้นมาเพื่อแก้ปัญหาเกี่ยวกับการแปรเปลี่ยนของค่าสถิติ ซึ่งเป็นจุดด้อยของทฤษฎีการวัดแบบคลาสสิคโดย Ferguson (1942) และ Lawleg (1943) เป็นผู้เสนอแนวคิดทฤษฎีการตอบสนองข้อสอบ มีหลักการว่า ผลการทดสอบของผู้สอบจากแบบสอบใด ๆ ขึ้นอยู่กับความสามารถของผู้สอบ และต่อมาในปี 1952 Lord ได้เสนอทฤษฎีนี้ขึ้นใหม่ในรูปโค้งลักษณะข้อสอบ (Item Characteristic Curve : ICC) ต่อมาเรียกว่า Ogive Model ซึ่งจะกล่าวถึงพารามิเตอร์ 2 ตัว คือ ค่าความยากและค่าอำนาจจำแนก แต่ในโมเดลดังกล่าวนี้มีความยุ่งยากในการคำนวณและขาดแคลนโปรแกรมคอมพิวเตอร์ในการวิเคราะห์ข้อมูล จึงทำให้ Lord หยุดความสนใจในทฤษฎีนี้ไประยะหนึ่ง ในปี ค.ศ. 1960 Rasch ได้ศึกษาเกี่ยวกับทฤษฎีดังกล่าวและได้เสนอแนวคิดในรูปพารามิเตอร์ตัวเดียว คือ ค่าความยาก ซึ่งบางครั้งเรียกแบบจำลองนี้ว่า Rasch Model ปี ค.ศ. 1968 Bimbaum ได้เสนอแนวคิดใหม่เกี่ยวกับ Logistic Model ที่ใช้พารามิเตอร์ 2 ตัวคือ ค่าความยากและค่าอำนาจจำแนก ซึ่งเป็นแบบจำลองที่ง่ายกว่าของ Lord จึงทำให้ Logistic Model เป็นที่นิยมแพร่หลายและมีการพัฒนาขึ้นเรื่อย ๆ จนกระทั่งใช้ได้กับพารามิเตอร์ตัวเดียวและพารามิเตอร์ 3 ตัว (Warm, 1979 อ้างถึงใน สุรศักดิ์ อมรรัตนศักดิ์, 2531)

วิธีการตอบสนองข้อสอบเป็นวิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยการพิจารณาเปรียบเทียบฟังก์ชั่นการตอบข้อสอบ ระหว่างกลุ่มผู้สอบที่มีต่อแบบสอบชุดเดียวกัน (Lord, 1980 cited in Green, 1994) หรือโดยการเปรียบเทียบความแตกต่างระหว่างโค้งคุณลักษณะข้อสอบ (ICC) ระหว่างกลุ่ม ซึ่งพื้นที่ระหว่างโค้งคุณลักษณะข้อสอบจะเป็นดัชนีบอกระดับของการทำหน้าที่ต่างกันของข้อสอบ

ในการวัดพื้นที่ระหว่างโค้งคุณลักษณะข้อสอบจะมี 2 ลักษณะคือ การวัดช่วงเปิด (open interval) จะเป็นการวัดในช่วงความสามารถทั้งหมดระหว่างโค้งทั้งสอง ซึ่งจะทำให้ได้พื้นที่ที่แน่นอนและการวัดช่วงปิด (closed interval) จะวัดในช่วงความสามารถตามที่กำหนดไว้ ซึ่งมีดัชนีที่ใช้บอกระดับการทำหน้าที่ต่างกันของข้อสอบได้แก่ พื้นที่ชนิดไม่มีเครื่องหมาย (unsigned areas) เป็นค่าสัมบูรณ์ของพื้นที่ระหว่างโค้งคุณลักษณะข้อสอบ พื้นที่ชนิดมีเครื่องหมาย (signed areas) เหมือนกับพื้นที่ชนิดไม่มีเครื่องหมาย แต่จะมีเครื่องหมายแสดงให้ทราบว่ากลุ่มใดได้ประโยชน์กลุ่มใดเสียประโยชน์ การทดสอบด้วยค่าสถิติ Z และไคสแควร์ (χ^2) เป็นการทดสอบนัยสำคัญของความแตกต่างของค่าพารามิเตอร์ a และ b ระหว่างกลุ่ม ในคราวเดียวกัน ซึ่งเป็นวิธีของ Lord (Shepard et al., 1994)

สูตรที่ใช้ในการคำนวณค่าฟังก์ชั่นการตอบข้อสอบแบบ 2 พารามิเตอร์ มีสูตรในการคำนวณดังนี้ คือ (Raju, 1988 cited in Cohen et al., 1991)

$$P_i(\theta) = [1+\exp\{-1.7a_i(\theta-b_i)\}]^{-1} \quad \text{.....(1)}$$

$$= [1+\exp\{-L_i(\theta)\}]^{-1} \quad \text{.....(2)}$$

$$= \frac{\exp\{-L_i(\theta)\}}{1+\exp\{-L_i(\theta)\}} \quad \text{.....(3)}$$

โดยที่ $L_i(\theta) = 1.7a_i(\theta-b_i)$ .....(4)

พื้นที่ (S_i) ภายใต้โค้งคุณลักษณะข้อสอบสำหรับข้อสอบข้อที่ i ระหว่างจุดสองจุดซึ่งเป็นช่วงที่สนใจ (θ_1, θ_2) สามารถเขียนได้ดังนี้

$$S_i(\theta_1, \theta_2) = \int_{\theta_1}^{\theta_2} P_i(\theta)d\theta \quad \text{.....(5)}$$

$$= (1.7a_i)^{-1}\ln[1+\exp\{L_i(\theta)\}]_{\theta_1}^{\theta_2} \quad \text{.....(6)}$$

$$= (1.7a_i)^{-1}[\ \ln[1+\exp\{L_i(\theta_2)\}] - \ln[1+\exp\{L_i(\theta_1)\}]\ ] \quad \text{....(7)}$$

โมเดลการตอบสนองข้อสอบแบบ 2 พารามิเตอร์ ของข้อสอบข้อที่ i มีค่าพารามิเตอร์ข้อสอบอยู่สองค่าคือ (a_{iR}, b_{iR}) สำหรับกลุ่มอ้างอิง และ (a_{iF}, b_{iF}) สำหรับกลุ่มเปรียบเทียบ

การคำนวณพื้นที่แบบมีเครื่องหมาย (SA) และพื้นที่แบบไม่มีเครื่องหมาย (UN) ระหว่างโค้งคุณลักษณะข้อสอบในช่วง θ_1, θ_2 สามารถคำนวณได้ดังนี้

$$SA = \int_{\theta_1}^{\theta_2}\{P_{iR}(\theta) - P_{iF}(\theta)\}d\theta \quad \text{..... (8)}$$

$$= \int_{\theta_1}^{\theta_2} P_{iR}(\theta)d\theta - \int_{\theta_1}^{\theta_2} P_{iF}(\theta)d\theta \quad \text{.....(9)}$$

$$= S_{iR}(\theta_1, \theta_2) - S_{iF}(\theta_1, \theta_2) \quad \text{.....(10)}$$

$$UN = \int_{\theta_1}^{\theta_2}|P_{iR}(\theta) - P_{iF}(\theta)|d\theta \quad \text{.....(11)}$$

เมื่อ θ_1, θ_2 คือ ระดับความสามารถที่ต่ำกว่าและสูงกว่าตามลำดับ

$P_{iR}(\theta)$ คือ ความน่าจะเป็นในการตอบข้อสอบข้อที่ i ได้ถูกต้องของผู้สอบที่ระดับความสามารถนั้นของกลุ่มอ้างอิง

$P_{iF}(\theta)$ คือ ความน่าจะเป็นในการตอบข้อสอบข้อที่ i ได้ถูกต้องของผู้สอบที่ระดับความสามารถนั้นของกลุ่มเปรียบเทียบ

S_{iR} คือ พื้นที่ของกลุ่มอ้างอิงสำหรับข้อสอบข้อที่ i

S_{iF} คือ พื้นที่ของกลุ่มเปรียบเทียบสำหรับข้อสอบข้อที่ i

ถ้าโค้งคุณลักษณะข้อสอบทั้งสองไม่ตัดกัน $(a_{iR} = a_{iF})$ สามารถคำนวณหาพื้นที่ระหว่างโค้งคุณลักษณะข้อสอบข้อที่ i ซึ่งเป็นพื้นที่ไม่คิดเครื่องหมายได้โดยพื้นที่ที่คำนวณได้จะเท่ากับพื้นที่แบบมีเครื่องหมาย

$$UN = \left|S_{iR}(\theta_1,\theta_2) - S_{iF}(\theta_1,\theta_2)\right| \quad \text{.....(12)}$$

$$= |SA| \quad \text{.....(13)}$$

ถ้าโค้งคุณลักษณะข้อสอบทั้งสองตัดกัน $(a_{iR} \neq a_{iF})$ ที่ θ_X

$$\theta_X = \frac{(a_{iF}b_{iF} - a_{iR}b_{iR})}{(a_{if} - a_{iR})} \quad \text{.....(14)}$$

สามารถคำนวณหาพื้นที่ระหว่างโค้งลักษณะข้อสอบข้อที่ i ซึ่งเป็นพื้นที่ไม่คิดเครื่องหมายได้ดังนี้

$$UN = \left|S_{iR}(\theta_1,\theta_X) - S_{iF}(\theta_1,\theta_X)\right| + \left|S_{iR}(\theta_X,\theta_2) - S_{iF}(\theta_X,\theta_2)\right| \quad \text{.....(15)}$$

$$= \left|\int_{\theta_1}^{\theta_X}\{P_{iR}(\theta) - P_{iF}(\theta)\}d\theta\right| + \left|\int_{\theta_X}^{\theta_2}\{P_{iR}(\theta) - P_{iF}(\theta)\}d\theta\right| \quad \text{.....(16)}$$

$$= \left|\int_{\theta_1}^{\theta_X} P_{iR}(\theta)d\theta - \int_{\theta_1}^{\theta_X} P_{iF}(\theta)d\theta\right| + \left|\int_{\theta_{X1}}^{\theta_2} P_{iR}(\theta)d\theta - \int_{\theta_X}^{\theta_2} P_{iF}(\theta)d\theta\right| \quad \text{.....(17)}$$

ตอนที่ 3 วิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมนเทล-แฮนส์เซล (Mantel-Haenszel procedure : MH)

วิธี MH เป็นวิธีที่พัฒนามาจากวิธีไคสแควร์แบบดั้งเดิม (traditional χ^2 approach) ซึ่งเสนอโดย Mantel และ Haenzel ในปี ค.ศ. 1959 เพื่อนำมาใช้ในงานวิจัยทางการแพทย์ แต่ผู้ที่วิเริ่มนำมาใช้เพื่อตรวจสอบการทำหน้าที่ต่างกันของแบบสอบ คือ Holland (Holland และ Thayer, 1988) หลังจากนั้นวิธี MH ถูกนำมาใช้อย่างกว้างขวางเพราะเป็นวิธีที่คำนวณได้ง่าย ประหยัดใช้กลุ่มตัวอย่างน้อย และการแปลผลไม่ยุ่งยาก

หลักการตรวจสอบการทำหน้าที่ต่างกันของวิธี MH เป็นการเปรียบเทียบผลการสอบของผู้สอบ 2 กลุ่ม คือ กลุ่มอ้างอิง (reference group) และกลุ่มเปรียบเทียบ (focal group) ในกรณีที่ข้อสอบทำหน้าที่ต่างกัน กลุ่มอ้างอิงคือ กลุ่มที่คาดว่าจะได้รับประโยชน์ในการตอบข้อสอบ ส่วนกลุ่มเปรียบเทียบเป็นกลุ่มที่คาดว่าจะเสียประโยชน์จากการตอบข้อสอบ ซึ่งจะมีการตรวจสอบกลุ่มผู้สอบทุกๆ ระดับคะแนนรวมจากแบบสอบ ข้อสอบข้อใดที่ผู้สอบในกลุ่มที่มีความสามารถเท่ากันทั้งสองกลุ่มทำได้ถูกต้องเท่ากันถือว่าเป็นข้อสอบทำหน้าที่ไม่ต่างกัน (no DIF)

การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธี MH จะแยกข้อสอบที่วิเคราะห์ออกเป็นรายข้อ ในการวิเคราะห์แต่ละข้อจะต้องสร้างตารางไขว้ขนาด 2x2 เพื่อแสดงความถี่ของผู้สอบที่ตอบข้อสอบข้อนั้นถูก (1) และตอบข้อสอบข้อนั้นผิด (0) ทั้งในกลุ่มอ้างอิง (R) และกลุ่มเปรียบเทียบ (F) ในช่วงคะแนน j ของผู้สอบทั้งสองกลุ่ม ดังตารางที่ 1 ถ้ามีการแบ่งช่วงคะแนนในการวิเคราะห์ออกเป็น k ช่วง จะได้ตารางไขว้ขนาด 2x2 ทั้งหมด k ตาราง

ตารางที่ 1 ความถี่ของกลุ่มผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่ระดับคะแนน j

กลุ่ม	คะแนนจากข้อสอบที่ต้องการตรวจสอบ DIF		
	ตอบถูก (1)	ตอบผิด (0)	รวม
อ้างอิง (R)	A_j	B_j	N_{Rj}
เปรียบเทียบ (F)	C_j	D_j	N_{Fj}
รวม	m_{1j}	m_{0j}	N_j

เมื่อ N_j เป็นความถี่ของผู้สอบทั้งหมดที่ได้ระดับคะแนน j

N_{Rj} , N_{Fj} เป็นความถี่ของกลุ่มผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่ได้ระดับคะแนน j ตามลำดับ

m_{1j} , m_{0j} เป็นความถี่ของผู้สอบที่ตอบข้อสอบถูกและผิดที่ระดับคะแนน j ตามลำดับ

A_j , B_j , C_j , D_j เป็นความถี่ของผู้สอบที่ตอบถูก (1) และตอบผิด (0) ของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่ระดับคะแนน j

จากนั้นนำผลการตอบข้อสอบมาแสดงเป็นสัดส่วนการตอบข้อสอบของกลุ่มตัวอย่างทั้งสองกลุ่มได้ดังตารางที่ 2

ตารางที่ 2 สัดส่วนการตอบข้อสอบของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่ระดับคะแนน j

กลุ่ม	คะแนนจากข้อสอบที่ต้องการตรวจสอบ DIF		
	1	0	รวม
อ้างอิง	P_{Rj}	Q_{Rj}	1
เปรียบเทียบ	P_{Fj}	Q_{Fj}	1

เมื่อ P_{Rj}, Q_{Rj} คือ สัดส่วนการตอบข้อสอบของกลุ่มอ้างอิงที่อยู่ในช่วงความสามารถ j ซึ่งตอบข้อสอบถูกและผิดตามลำดับ

P_{Fj}, Q_{Fj} คือ สัดส่วนการตอบข้อสอบของกลุ่มเปรียบเทียบที่อยู่ในช่วงความสามารถ j ซึ่งตอบข้อสอบถูกและผิดตามลำดับ

หลักการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธี MH เป็นการนำข้อมูลจากตารางไขว้ k ตารางมาดำเนินการวิเคราะห์ตามขั้นตอนดังนี้

1. คำนวณค่าความน่าจะเป็นในรูปสัดส่วนการตอบข้อสอบถูกและผิดระหว่างกลุ่มในแต่ละข้อในทุกช่วงคะแนน j โดยใช้สูตร

$$\alpha_{MH} = \frac{\sum A_j D_j / N_j}{\sum B_j C_j / N_j} \quad \text{.....(18)}$$

เมื่อ α_{MH} คือ สัดส่วนของการตอบข้อสอบถูกและผิดระหว่างกลุ่มในแต่ละข้อในทุกช่วงคะแนน j

ค่า α_{MH} มีค่าระหว่าง 0 ถึง α เกณฑ์ในการพิจารณาข้อสอบทำหน้าที่ต่างกัน ถ้าค่า α_{MH} ที่คำนวณได้มีค่าเท่ากับ 1 แสดงว่าข้อสอบทำหน้าที่ไม่แตกต่างกัน ถ้าค่า $\alpha_{MH} > 1$ แสดงว่าข้อสอบข้อนั้นทำหน้าที่ต่างกันโดยเข้าข้างกลุ่มอ้างอิง

2. ทดสอบนัยสำคัญของค่าไคสแควร์ เพื่อทดสอบค่า α_{MH} ที่คำนวณได้ว่ามีความแตกต่างจาก 1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 หรือไม่ ที่ระดับชั้นความเป็นอิสระเท่ากับ 1 ตามสูตรดังนี้

$$\chi^2_{MH} = \frac{\left\{\left|\sum A_j - \sum E(A_j)\right| - 0.5\right\}^2}{\sum Var(A_j)} \qquad (19)$$

เมื่อ $\sum(A_j) = (N_{Rj})(m_{1j})/N_j$ .....(20)

$$Var(A_j) = \frac{N_{Rj}N_{Fj}m_{1j}m_{0j}}{N_j^2(N_j - 1)} \qquad .....(21)$$

ดังนั้นการตรวจสอบด้วยวิธี MH จึงมีสมมติฐานหลักที่แสดงว่าข้อสอบข้อนั้นทำหน้าที่ไม่ต่างกันดังนี้ คือ $H_0 : P_{Rj} = P_{Fj}$ สำหรับทุกชั้นคะแนน j .....(22)

หรือ $H_0 : \frac{A_jD_j}{N_j} = \frac{B_jC_j}{N_j}$ สำหรับทุกชั้นคะแนน j ..... (23)

สมมุติฐานหลักนี้เป็นสมมุติฐานของความเป็นอิสระอย่างมีเงื่อนไขของสมาชิกกลุ่มและคะแนนที่ได้จากการตอบข้อสอบที่ต้องการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ และภายใต้สมมุติฐานหลักจะได้ค่าคาดหวังของการตอบ ดังนี้

$$\sum(A_j) = (N_{Rj}m_{1j})/N_j \qquad .....(24)$$
$$\sum(B_j) = (N_{Rj}m_{0j})/N_j \qquad .....(25)$$
$$\sum(C_j) = (N_{Fj}m_{1j})/N_j \qquad .....(26)$$
$$\sum(D_j) = (N_{Fj}m_{0j})/N_j \qquad .....(27)$$

สมมุติฐานอื่นของวิธีแมนเทล-แฮนส์เซล คือ

$$H_1 : \frac{P_{Rj}}{Q_{Rj}} = \alpha \frac{P_{Fj}}{Q_{Fj}} \quad j=1,...,k \text{ เมื่อ } \alpha \text{ ไม่เท่ากับ } 1 \quad \text{.....(28)}$$

แต่เมื่อ α เท่ากับ 1 ซึ่งสอดคล้องกับสมมุติฐานศูนย์ จะได้ว่า

$$H_0 : \frac{P_{Rj}}{Q_{Rj}} = \frac{P_{Fj}}{Q_{Fj}} \quad \text{.....(29)}$$

และค่าประมาณของ $\alpha_{MH} = \frac{P_{Rj} P_{Fj}}{Q_{Rj} Q_{Fj}} = \frac{P_{Rj} Q_{Fj}}{P_{Fj} Q_{Rj}}$ สำหรับทุก j= 1,...,k .....(30)

Holland และ Thayer (1988) ได้เสนอให้แปลงค่า α_{MH} ให้เป็นค่าเดลต้า (Δ_{MH}) หรือ MH_{DIF} ดังนี้

$$MH_{DIF} = -2.35 \ln (\alpha_{MH}) \quad \text{.....(31)}$$

ค่า MH_{DIF} มีค่าระหว่าง -2.6 ถึง 2.6 สำหรับเกณฑ์ในการตัดสินว่าข้อสอบทำหน้าที่ต่างกันคือ ข้อสอบที่มีค่า α_{MH} แตกต่างจาก 1 อย่างมีนัยสำคัญทางสถิติ หรือมีค่า MH_{DIF} แตกต่างจาก 0 อย่างมีนัยสำคัญทางสถิติ โดยมีเกณฑ์ในการพิจารณาค่า MH_{DIF} ดังนี้ 1) ถ้าค่า $MH_{DIF} = 0$ หรือไม่แตกต่างจาก 0 อย่างมีนัยสำคัญทางสถิติ แสดงว่าข้อสอบข้อนั้นทำหน้าที่ไม่ต่างกันระหว่างกลุ่ม 2) ถ้าค่า MH_{DIF} แตกต่างจาก 0 อย่างมีนัยสำคัญทางสถิติ และมีค่าเป็นบวก แสดงว่าข้อสอบข้อนั้นทำหน้าที่ต่างกันโดยเข้าข้างกลุ่มเปรียบเทียบ 3) ถ้าค่า MH_{DIF} แตกต่างจาก 0 อย่างมีนัยสำคัญทางสถิติ และมีค่าเป็นลบ แสดงว่าข้อสอบข้อนั้นทำหน้าที่ต่างกัน โดยเข้าข้างกลุ่มอ้างอิง

การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธี MH นี้ เป็นการเปรียบเทียบผลการตอบข้อสอบของผู้สอบสองกลุ่ม หลังจากการจับคู่ผู้สอบตามความรู้ หรือความสามารถของผู้สอบ การจับคู่ผู้สอบของผู้สอบสองกลุ่มเป็นเงื่อนไขที่สำคัญเพราะเป็นเกณฑ์ที่ใช้แทนความสามารถที่แท้จริงของผู้สอบสองกลุ่ม ในทางปฏิบัติมักใช้คะแนนรวมของแบบสอบ (total test score) เป็นเกณฑ์ในการจับคู่ เพราะเกี่ยวข้องโดยตรงกับความรู้หรือความสามารถที่วัดได้จากแบบสอบนั้น และจากผู้สอบทุกคนภายใต้สถานการณ์เดียวกัน แต่การใช้คะแนนรวมของแบบสอบเป็นเกณฑ์ในการจับคู่ผู้สอบมีจุดอ่อนคือ มีการรวมเอาคะแนนจากข้อสอบที่ทำหน้าที่ต่างกัน มาเป็นเกณฑ์ใน

การจับคู่ผู้สอบด้วย จุดอ่อนนี้สามารถแก้ไขได้โดยใช้การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบแบบ 2 ขั้นตอนซึ่ง Holland และ Thayer (1988) ได้เสนอขั้นตอนการวิเคราะห์ ดังนี้

ขั้นตอนที่ 1 ใช้คะแนนรวมของแบบสอบทั้งฉบับเป็นเกณฑ์ในการจับคู่ผู้สอบ แล้ววิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ

ขั้นตอนที่ 2 นำคะแนนของข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบในขั้นตอนที่ 1 ออกจากเกณฑ์การจับคู่ผู้สอบแล้วใช้คะแนนรวมของข้อสอบที่ทำหน้าที่ไม่แตกต่างกัน หรือข้อสอบที่เหลือเป็นเกณฑ์การจับคู่ผู้สอบในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบของแบบสอบทั้งฉบับในขั้นตอนที่ 2

ตอนที่ 4 วิธีการตรวจสอบการทำหน้าต่างกันของข้อสอบด้วยวิธีวิเคราะห์องค์ประกอบจำกัด (Restricted Factor Analysis procedure : RFA)

หลักการวิเคราะห์การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยวิธีวิเคราะห์องค์ประกอบจำกัด ใช้โมเดลองค์ประกอบร่วมเชิงเส้น โดยสมมุติว่าแบบสอบชุดหนึ่งสร้างขึ้นเพื่อวัดคุณลักษณะแฝงคุณลักษณะหนึ่งซึ่งมีข้อสอบทั้งหมด p ข้อ และการทำหน้าที่ต่างกันของข้อสอบจะศึกษาได้จากตัวฝ่าฝืน (potential violator) r ตัว สำหรับการเลือกข้อสอบแบบสุ่ม คะแนนของข้อสอบข้อที่ i แทนได้ด้วยค่า X_i ซึ่งค่า X_i แสดงได้ดังสมการที่ 32

$$X_i = m_i + a_i T + \sum_{k=1}^{r} (b_{ik} V_k) + D_i \quad \text{..... (32)}$$

เมื่อ T เป็นคุณลักษณะที่แสดงถึงคะแนนที่แท้จริง (true score)ของคุณลักษณะที่ศึกษา

V_k เป็นคุณลักษณะแฝงที่เป็นคะแนนจริง (true score) ของตัวฝ่าฝืนตัวที่ k

D_i เป็นองค์ประกอบของคะแนนความคลาดเคลื่อนของการทำข้อสอบข้อที่ i

m_i เป็นค่าจุดตัดแกน

a_i เป็นสัมประสิทธิ์การถดถอยของข้อสอบข้อที่ i ของคุณลักษณะ T_i

b_{ik} เป็นสัมประสิทธิ์การถดถอยของข้อสอบข้อที่ i ของคุณลักษณะ V_k

จากสมการที่ 32 ข้อสอบข้อที่ i จะทำหน้าที่ต่างกันตามตัวฝ่าฝืนที่ k ถ้ามีอิทธิพลทางตรงจากตัวฝ่าฝืนตัวที่ k มายังข้อสอบข้อที่ i แล้วส่งผลให้ค่าน้ำหนักองค์ประกอบไม่เท่ากับศูนย์ ($b_{ik} \neq 0$) ลักษณะเช่นนี้เรียกว่า การทำหน้าที่ต่างกันจากตัวฝ่าฝืนตัวที่ k บนข้อสอบข้อที่ i เพราะฉะนั้นการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการวิเคราะห์องค์ประกอบจำกัดจะตั้งสมมุติฐานหลัก (H_0) ว่า $b_{ik} = 0$ แล้วทำการตรวจสอบสมมุติฐานหลักสำหรับข้อสอบทุกข้อและสำหรับทุกตัวฝ่าฝืน

จากโมเดลการวิเคราะห์องค์ประกอบ ความสัมพันธ์ระหว่างข้อสอบ คุณลักษณะแฝง T และตัวฝ่าฝืน r ตัวที่ถูกวัดด้วยข้อสอบหรือจากตัวปงชี้ q ตัว ($q \geq r$) คะแนน Y_j จากตัวปงชี้ตัวที่ j สามารถคำนวณได้จากสมการที่ 33

$$Y_j = n_j + \sum_{k=1}^{r}(c_{jk}V_k) + E_j \quad \text{.....(33)}$$

เมื่อ Y_j เป็นคะแนนจากตัวปงชี้ตัวที่ j

n_j เป็นค่าเฉลี่ยของตัวปงชี้ที่ j

c_{jk} เป็นสัมประสิทธิ์การถดถอยของตัวปงชี้ตัวที่ j บนตัวฝ่าฝืนตัวที่ k

E_j เป็นองค์ประกอบของความคลาดเคลื่อน

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

จากสมการที่ 32 และสมการที่ 33 สามารถนำมาใช้ในโมเดลการวิเคราะห์องค์ประกอบความแปรปรวนร่วมในสมการที่ 34

$$S = LFL' + U^2 \quad \text{.....(34)}$$

เมื่อ S เป็นเมตริกซ์ความแปรปรวน-ความแปรปรวนร่วมของคะแนนสังเกตได้ X_i

Y_i, L เป็นเมตริกซ์ของน้ำหนักองค์ประกอบของตัวแปรสังเกตได้จากตัวแปรแฝงหรือองค์ประกอบร่วม T

V_k, F เป็นเมตริกซ์ความแปรปรวน-ความแปรปรวนร่วมของตัวแปรแฝง

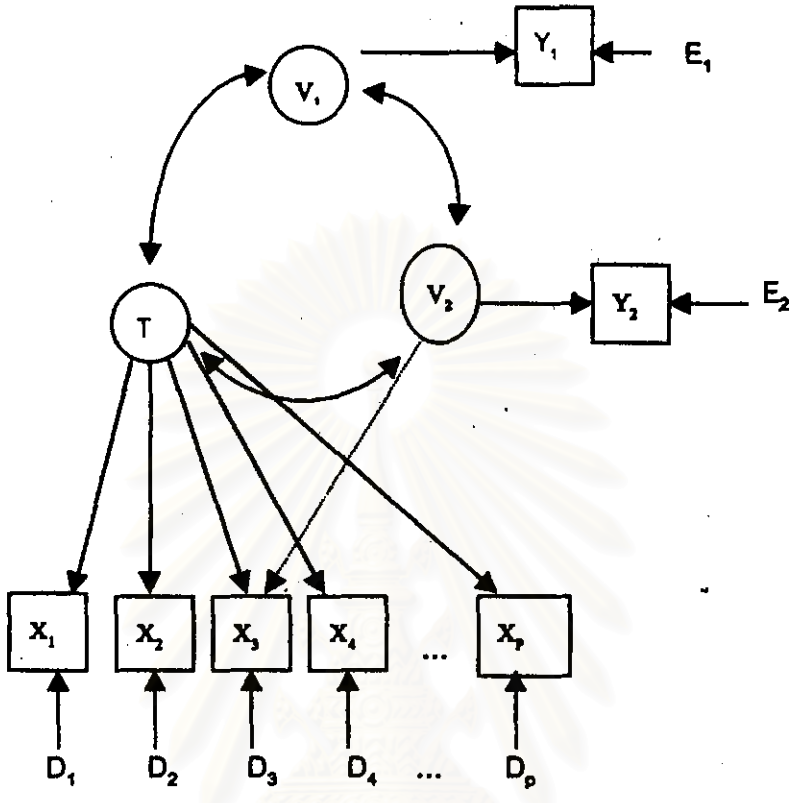
U^2 เป็นเมตริกซ์ไดแอกโกนอล (diagonal matrix) ของค่าความแปรปรวนขององค์ประกอบที่คลาดเคลื่อน

การวิเคราะห์องค์ประกอบ ตัวแปรทุกตัวจะวัดได้จากค่าเฉลี่ย จุดตัดแกนในสมการที่ 32 และสมการที่ 34

ในกรณีที่มีตัวบ่งชี้ 1 ตัวของแต่ละตัวฝำฝืน ($q = r$) ตัวฝำฝืน V_k ถูกแทนที่ด้วยตัวแปรสังเกตได้ Y_j ซึ่งหมายความว่าในโมเดลองค์ประกอบนี้ตัวฝำฝืนสันนิษฐานว่าสามารถวัดได้โดยปราศจากความคลาดเคลื่อนในการวัด และความแปรปรวนที่อธิบายไม่ได้ ($E_j = 0$) ของตัวฝำฝืนใน U^2 จะถูกกำหนดให้เป็น 0 เพื่อความสะดวกมากขึ้น ตัวแปรสังเกตได้เหล่านี้อาจจะเป็นคะแนนมาตรฐานก็ได้ และถ้าใช้เมตริกซ์สหสัมพันธ์แทนเมตริกซ์ความแปรปรวนร่วมจะทำให้มีประสิทธิภาพมากขึ้นเพราะค่าสัมประสิทธิ์คือ คะแนนมาตรฐานของค่าความแปรปรวนร่วมนั่นเอง และตัวแปรแฝงอาจจะทำให้เป็นคะแนนมาตรฐานได้โดยการกำหนดสมาชิกในแนวทะแยงของ F ให้เหมือนกัน นอกจากนั้นน้ำหนักองค์ประกอบของตัวแปร Y_j จาก V_k อาจจะกำหนดให้เหมือน ๆ กัน ($c_{jk} = 0$ ของทุก ๆ j และทุก ๆ k เมื่อ $j = k$) และกำหนดให้ตัวอื่น ๆ เป็น 0 ซึ่งตัวที่จะถูกประมาณค่าคือ น้ำหนักองค์ประกอบของข้อสอบจากคุณลักษณะแฝง T ความแปรปรวนของความคลาดเคลื่อนของข้อสอบ และสหสัมพันธ์ระหว่างตัวฝำฝืนและคุณลักษณะแฝง T

การทำหน้าที่ต่างกันของข้อสอบสามารถตรวจสอบได้จากค่าสัมประสิทธิ์การถดถอยของข้อสอบจากตัวฝำฝืนต่าง ๆ โมเดลการวิเคราะห์องค์ประกอบจำกัด ค่าสัมประสิทธิ์การถดถอยจะถูกนำเสนอในรูปน้ำหนักองค์ประกอบของตัวแปรสังเกตได้ X_i จากตัวฝำฝืน V_k โมเดลองค์ประกอบที่มีค่าน้ำหนักองค์ประกอบเป็น 0 ถูกกำหนดให้เป็นโมเดลหลัก (H_0) ซึ่งเป็นโมเดลที่ไม่มีความลำเอียง คือค่า $b_{ik} = 0$ ของทุก ๆ i และทุก ๆ k จากภาพที่ 1 โมเดลการทดสอบความลำเอียงของข้อสอบข้อที่ 3 ตามตัวฝำฝืนตัวที่ 2 โมเดลองค์ประกอบอื่นก็ต้องถูกตรวจสอบด้วย ซึ่งโมเดลอื่นๆ นี้จะมีลักษณะเหมือนกับโมเดลหลัก (H_0) ยกเว้นน้ำหนักองค์ประกอบของข้อที่ 3 จากตัวฝำฝืนตัวที่ 2 (b_{32}) ซึ่งจะถูกกำหนดให้เป็นพารามิเตอร์อิสระ ถ้าโมเดลนี้มีค่านัยสำคัญมากกว่า โมเดลหลัก (H_0) ข้อสอบข้อที่ 3 ก็จะเป็นข้อสอบข้อที่มีความลำเอียงตามตัวฝำฝืนตัวที่ 2 และจะต้องถูกขจัดออกจากแบบสอบ

ภาพที่ 1 โมเดลองค์ประกอบความสัมพันธ์ระหว่างข้อสอบแต่ละข้อ, คุณลักษณะ T และตัวฝาฝืน (Oort,1996:16)



ภาพที่ 1 แสดงรูปแบบที่อธิบายได้จากสมการที่ (33) และ (34) เมื่อมีตัวฝาฝืน 2 ตัว (V_1,V_2) ซึ่งแต่ละตัววัดจากตัวบ่งชี้เพียงตัวเดียวและคุณลักษณะ T มีความสัมพันธ์กับคะแนน Y_1,Y_2 และ V_1,V_2 ซึ่งคุณลักษณะ T วัดได้จากคะแนนข้อสอบ X_i นอกจากนี้ คุณลักษณะ T , V_k และ X_i อาจมีความสัมพันธ์กันด้วย จากภาพที่ 1 เส้นประจากตัวฝาฝืนตัวที่ 2 ไปยังข้อสอบข้อที่ 3 (X_3) แสดงถึงอิทธิพลทางตรงจากตัวฝาฝืนตัวที่ 2 ไปยังข้อสอบข้อที่ 3 ซึ่งอาจเป็นอิทธิพลทางอ้อมผ่านตัวแปร T ไปยัง X_3 ก็ได้ ค่าอิทธิพลทางตรงนี้บอกเป็นนัยว่าคะแนนที่วัดจาก T ที่มีค่าสูง ๆ ไม่จำเป็นต้องสัมพันธ์กับคุณลักษณะแฝง T ซึ่งมีความเป็นไปได้ที่ผู้สอบที่มีคะแนน T เท่ากันอาจจะได้คะแนนในข้อสอบข้อที่ 3 (X_3) ไม่เท่ากัน ทั้งนี้เพราะคะแนนดังกล่าวได้รับอิทธิพลจากตัวฝาฝืนตัวที่ 2 (V_2) แตกต่างกัน ดังนั้น ข้อสอบข้อที่ 3 (X_3) จึงไม่จำเป็นที่จะต้องวัดเพียงคุณลักษณะแฝง T อย่างเดียวแต่วัดตัวฝาฝืนตัวที่ 2 (V_2) ด้วย กล่าวได้ว่าข้อสอบข้อที่ 3 (X_3) จะมีความลำเอียงตามตัวฝาฝืนตัวที่ 2 (V_2)

โมเดลหลัก (H_0) สามารถตรวจสอบด้วยโปรแกรม LISREL โดยพิจารณาจากค่า Modification indices (MI) ซึ่งปกติแล้วการกระจายของค่า MI นี้จะมีการแจกแจงเป็นไคสแควร์ที่มีองศาอิสระเท่ากับ 1 ถ้าค่า MI ของข้อสอบข้อใดมีค่ามากและมีนัยสำคัญแสดงว่าข้อสอบข้อนั้นทำ

หน้าที่ต่างกันต้องตัดออกจากแบบสอบ นอกจากนี้ยังสามารถตรวจสอบขนาดและทิศทางของการทำหน้าที่ต่างกันของข้อสอบได้โดยการพิจารณาที่ค่าพารามิเตอร์ที่คาดหวัง (Expected Parameter Change : EPC) ซึ่งค่า EPC นี้เป็นการประมาณค่าการเปลี่ยนแปลงที่คาดหวังของค่าพารามิเตอร์กำหนด เมื่อในโมเดลสมมุติฐาน (H_0) กำหนดให้เป็นค่าพารามิเตอร์จริงระ ถ้าข้อสอบข้อที่ i เป็นข้อสอบที่ทำหน้าที่ต่างกันและค่า EPC เป็นค่าสัมประสิทธิ์การถดถอยของข้อสอบข้อที่ i จากตัวฝ่าฝืน k เป็นบวกแสดงว่าข้อสอบข้อที่ i ทำหน้าที่ต่างกันต่อกลุ่มผู้ตอบที่มีคะแนนของตัวฝ่าฝืน k มาก Kaplan (1990 อ้างถึงใน Oort, 1998) เสนอแนะว่า การที่จะขจัดข้อสอบที่ทำหน้าที่ต่างกันออกจากแบบสอบ สามารถพิจารณาโดยใช้ค่า MI และค่า EPC ที่แตกต่างจาก 0 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 โดยไม่ต้องพิจารณาขนาดและทิศทางของค่า EPC ก็ได้

การใช้วิธี RFA ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมีข้อดีอยู่หลายประการ (Oort, 1996)

1. สามารถตรวจสอบการทำหน้าที่ต่างกันของแบบสอบที่มีรูปแบบการตอบที่หลากหลาย เช่น ถ้าระดับการตอบข้อสอบเป็นแบบต่อเนื่อง จะสามารถวิเคราะห์โดยใช้สหสัมพันธ์แบบเพียร์สันและความแปรปรวนร่วมได้ ถ้าระดับการตอบข้อสอบแบ่งเป็นสองหรือแบ่งแบบพหุ จะสามารถวิเคราะห์โดยใช้สหสัมพันธ์เตตระคอริกหรือสหสัมพันธ์โพลีคอริกได้

2. จากคำอธิบายดังที่กล่าวมา ตัวฝ่าฝืนเป็นตัวแปรที่แสดงความเป็นสมาชิกของกลุ่ม คือเป็นตัวแปรนามบัญญัติ แต่ในวิธี RFA สามารถตรวจสอบตัวฝ่าฝืนที่มีลักษณะใดก็ได้ โดยไม่คำนึงถึงระดับของการวัด

3. วิธี RFA สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีตัวฝ่าฝืนหลายตัวได้ ถึงแม้ว่าจะเป็นคนละชนิดกันก็ตาม

4. วิธี RFA ไม่จำเป็นต้องแบ่งกลุ่มตัวอย่างออกเป็นกลุ่มย่อยๆ ก็ได้

5. หลีกเลี่ยงปัญหาการประมาณค่าพารามิเตอร์ของกลุ่มประชากรที่แตกต่างกันให้อยู่ในระดับเดียวกันได้

6. กลุ่มตัวอย่างที่ใช้ในการวิเคราะห์ด้วยวิธี RFA นี้ไม่จำเป็นต้องมีขนาดใหญ่เหมือนกับวิธี IRT

ตอนที่ 5 งานวิจัยที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

Ryan (1991) ได้ศึกษาความคงที่ (stability) ของวิธีแมนเทล-แฮนส์เซลด้วยการเปลี่ยนกลุ่มตัวอย่างและการจับคู่เปรียบเทียบ กลุ่มตัวอย่างเป็นนักเรียนเกรด 8 เป็นกลุ่มนักเรียนผิวขาว 5,015 คน และกลุ่มนักเรียนผิวดำ 670 คน ในแต่ละกลุ่มจะแบ่งเป็นกลุ่มย่อย 4 กลุ่มโดยวิธีสุ่ม แบบสอบที่ใช้เป็นวิชาคณิตศาสตร์ ประกอบด้วยเนื้อหาด้านพีชคณิต เรขาคณิต เลขคณิต การวัด และสถิติ ซึ่งแบบสอบที่จะใช้วัดความสามารถเพื่อจับคู่เปรียบเทียบนั้นศึกษา 3 แบบได้แก่ (1) ข้อสอบร่วม 40 ข้อ (2) ข้อสอบเวียนที่สุ่มจากเนื้อหาต่าง ๆ ฉบับละ 35 ข้อจำนวน 4 ฉบับ (3) ข้อสอบที่ได้จากการรวมข้อสอบร่วมกับข้อสอบเวียน เป็นฉบับละ 75 ข้อจำนวน 4 ฉบับ เงื่อนไขที่ศึกษาจะแปรเปลี่ยนกลุ่มตัวอย่าง (ระหว่างผิวขาวกับผิวขาว,ระหว่างผิวขาวกับผิวดำ) ขนาดกลุ่มตัวอย่าง (กลุ่มใหญ่และกลุ่มย่อย) เกณฑ์ในการจับคู่เปรียบเทียบใช้ 2 เกณฑ์ เกณฑ์ที่ 1 พิจารณาจากคะแนนจากแบบสอบร่วม 40 ข้อ และเกณฑ์ที่ 2 พิจารณาจากคะแนนแบบสอบแบบที่ 3 ที่มีข้อสอบ 75 ข้อ

ผลการศึกษาพบว่า เมื่อนำค่า MH ที่ได้จากการวิเคราะห์แต่ละเงื่อนไขมาหาความสัมพันธ์ในเงื่อนไขที่ใช้กลุ่มตัวอย่างใหญ่กับกลุ่มตัวอย่างย่อยระหว่างผิวขาวกับผิวดำ เมื่อจับคู่เปรียบเทียบโดยใช้เกณฑ์ที่ 1 มีค่าสหสัมพันธ์อยู่ระหว่าง 0.74 ถึง 0.88 เมื่อใช้เกณฑ์ที่ 2 มีค่าสหสัมพันธ์ระหว่าง 0.75 ถึง 0.88 ซึ่งแสดงให้เห็นว่าเกณฑ์ที่ใช้ในการจับคู่เปรียบเทียบไม่มีผลกระทบต่อค่า MH ไม่ว่าจะใช้กลุ่มตัวอย่างใหญ่หรือกลุ่มย่อย ซึ่งสรุปได้ว่าดัชนี MH มีความแกร่งต่อผลกระทบของบริบทข้อสอบและหากต้องการให้มีความคงที่ในการประมาณค่าจากวิธี MH ควรใช้กลุ่มตัวอย่างที่มีขนาดใหญ่ขึ้น

Mazor และคณะ (1992) ได้ศึกษาผลกระทบของขนาดกลุ่มตัวอย่างที่มีต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมนเทล-แฮนส์เซล โดยศึกษาจากข้อมูลจำลอง กลุ่มตัวอย่างที่ใช้มี 5 ขนาด คือ 100, 200, 500, 1,000 และ 2,000 คน ความยาวของแบบสอบ 75 ข้อ พบว่าเมื่อขนาดกลุ่มตัวอย่างเท่ากับ 100, 200 และ 500 คนสามารถระบุข้อสอบที่ทำหน้าที่ต่างกันได้ถูกต้องน้อยกว่าร้อยละ 50 และเมื่อขนาดกลุ่มตัวอย่างเท่ากับ 2,000 คน จะสามารถระบุข้อสอบที่ทำหน้าที่ต่างกันได้ถูกต้องร้อยละ 70 ถึงร้อยละ 75 นอกจากนี้ข้อสอบที่ไปสามารถตรวจสอบพบการทำหน้าที่ต่างกันได้ เนื่องมาจากข้อสอบเหล่านี้มีความยากมากหรือเป็นข้อสอบที่มีความยากต่างกันเพียงเล็กน้อยระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ อีกทั้งเป็นข้อที่มีค่าอำนาจจำแนกต่ำ

Roger และ Swaminathan (1993) ได้เปรียบเทียบวิธีถดถอยโลจิสติกกับวิธี MH ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ซึ่งทดสอบเกี่ยวกับการกระจายของสถิติทดสอบและประสิทธิภาพของสถิติทดสอบแต่ละวิธี โดยใช้ข้อมูลจำลอง การศึกษาด้านการกระจายของสถิติทดสอบ ค่าพารามิเตอร์ที่แปรเปลี่ยนได้แก่ ขนาดกลุ่มตัวอย่าง ความเหมาะสมของข้อมูลกับโมเดล ค่าความยาก ค่าอำนาจจำแนก ความยาวแบบสอบ 40 ข้อ ส่วนการศึกษาด้านประสิทธิภาพของแต่ละวิธีมีค่าพารามิเตอร์ที่แปรเปลี่ยนได้แก่ ขนาดกลุ่มตัวอย่าง ความเหมาะสมของโมเดลกับข้อมูล ขนาดของแบบสอบ การกระจายของคะแนนสอบ อัตราส่วนของข้อสอบที่ทำหน้าที่ต่างกัน ค่าความยาก ค่าอำนาจจำแนกและพื้นที่ระหว่างโค้งคุณลักษณะข้อสอบระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ

ผลการศึกษาพบว่าการกระจายของค่าสถิติเป็นไปตามสมมุติฐานที่ตั้งไว้เกือบทั้งหมดในทั้งสองวิธี กรณีที่การกระจายของค่าสถิติของวิธีถดถอยโลจิสติไม่เป็นไปตามที่คาดไว้ เนื่องมาจากข้อสอบมีค่าความยากสูงและมีค่าอำนาจจำแนกสูง ด้านประสิทธิภาพพบว่าทั้งสองวิธีมีประสิทธิภาพไม่ต่างกันในการตรวจสอบ DIF แบบเอกรูป (uniform DIF) แต่วิธีถดถอยโลจิสติมีประสิทธิภาพมากกว่าในการตรวจสอบ DIF แบบอเนกรูป (non-uniform DIF) ขนาดกลุ่มตัวอย่างเป็นปัจจัยที่มีผลกระทบอย่างมากต่ออัตราการตรวจสอบการทำหน้าที่ต่างกันของทั้งสองวิธีนี้กล่าวคือ เมื่อเพิ่มขนาดกลุ่มตัวอย่าง อัตราการตรวจสอบจะเพิ่มขึ้น ส่วนขนาดของแบบสอบและการกระจายของคะแนนไม่มีผลกระทบต่ออัตราการตรวจสอบ

Mazor และคณะ (1994) ใช้วิธี MH ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป (non-uniform DIF) กลุ่มตัวอย่างได้จากการจำลองขึ้นกลุ่มละ 1,000 คน ใช้ แบบสอบ 25 ฉบับ ๆ ละ 75 ข้อ ในแต่ละฉบับมีข้อสอบที่ทำหน้าที่ไม่แตกต่างกัน (no-DIF) จำนวน 59 ข้อ และข้อสอบที่ทำหน้าที่ต่างกัน (DIF) จำนวน 16 ข้อ โดยข้อสอบที่ทำหน้าที่ต่างกันนั้นจะแปรเปลี่ยนค่าพารามิเตอร์ดังนี้ ค่าอำนาจจำแนก 4 ระดับ ค่าการเดากำหนดเป็น 0.2 และการกระจายของความสามารถระหว่างกลุ่ม 2 แบบคือการกระจายความสามารถเท่ากันและไม่เท่ากัน

ในการตรวจสอบด้วยวิธี MH จะวิเคราะห์ 2 แบบคือ แบบที่ 1 จะใช้คะแนนรวมของแต่ละคนเป็นเกณฑ์ในการจับคู่เปรียบเทียบระหว่างกลุ่ม ส่วนแบบที่ 2 จะแยกวิเคราะห์เฉพาะกลุ่มตัวอย่างที่เป็นกลุ่มสูงหรือกลุ่มต่ำ โดยใช้ค่าเฉลี่ยจากคะแนนของกลุ่มตัวอย่างทุกคนเป็นเกณฑ์ในการแยกเป็นกลุ่มสูงกลุ่มต่ำ จากนั้นจึงนำเฉพาะกลุ่มสูงหรือกลุ่มต่ำที่ได้มาแบ่งเป็นกลุ่มอ้างอิงและกลุ่มเปรียบเทียบแล้ววิเคราะห์

ผลการศึกษาพบว่าการวิเคราะห์โดยแบ่งกลุ่มตัวอย่างเป็นกลุ่มสูงและกลุ่มต่ำจะทำให้อัตราการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปสูงกว่าการวิเคราะห์ด้วยแบบที่ 1 (รวมผู้สอบกลุ่มสูงและกลุ่มต่ำไว้ด้วยกัน) อีกทั้งไม่ทำให้อัตราความคลาดเคลื่อนชนิดที่ 1 เพิ่มขึ้น และพบว่าเมื่อค่าอำนาจจำแนกและค่าความยากระหว่างกลุ่มเพิ่มขึ้น อัตราการตรวจพบข้อที่ทำหน้าที่ต่างกันมากขึ้นด้วย

Narayanan และ Swaminathan (1994) ได้ศึกษาผลของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี MH กับวิธีซิบเทสท์ โดยใช้ข้อมูลจำลอง ความยาวแบบสอบ 40 ข้อ ตัวแปรที่ศึกษา (1) ขนาดกลุ่มตัวอย่าง โดยกลุ่มอ้างอิงมี 3 ขนาดได้แก่ 300, 500 และ1,000 คน กลุ่มเปรียบเทียบ 3 ขนาดได้แก่ 100, 200 และ 300 คน ซึ่งจะจับคู่ศึกษาได้ 9 เงื่อนไข (2) การกระจายความสามารถ 2 แบบ (3) อัตราส่วนของข้อสอบที่ทำหน้าที่ต่างกันที่มีภายในแบบสอบ 2 ขนาด (4) ขนาดพื้นที่ระหว่างโค้งคุณลักษณะข้อสอบของผู้สอบสองกลุ่ม 4 ขนาด (5) ค่าความยากและค่าอำนาจจำแนกของแบบสอบ 6 ระดับ

ผลการศึกษาพบว่าขนาดกลุ่มตัวอย่าง อัตราส่วนของข้อสอบที่ทำหน้าที่ต่างกัน ขนาดของพื้นที่ระหว่างโค้งคุณลักษณะ ค่าความยากและค่าอำนาจจำแนกเป็นตัวแปรที่มีผลกระทบต่ออัตราการตรวจสอบของทั้งสองวิธีอย่างมีนัยสำคัญ วิธี MH และวิธีซิบเทสท์มีประสิทธิภาพไม่ต่างกันในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูปเมื่อการกระจายความสามารถระหว่างกลุ่มไม่ต่างกัน แต่เมื่อกระจายความสามารถระหว่างกลุ่มต่างกันวิธีซิบเทสท์จะมีประสิทธิภาพในการตรวจสอบมากกว่าวิธี MH จึงสรุปได้ว่าการกระจายความสามารถไม่มีผลกระทบต่อการตรวจสอบด้วยวิธีซิบเทสท์ แต่มีผลกระทบต่อวิธี MH อย่างมีนัยสำคัญ ส่วนอัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธี MH เป็นไปตามที่คาดไว้ แต่อัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธี SIBTESTS สูงกว่าที่คาดไว้เล็กน้อย ในกรณีที่มีการกระจายของความสามารถต่างกันเพิ่มขึ้นระหว่างผู้สอบสองกลุ่มจะทำให้อัตราความคลาดเคลื่อนประเภทที่ 1 เพิ่มขึ้น

Roussos และ Stout (1996) ได้ศึกษาผลกระทบของกลุ่มตัวอย่างขนาดเล็กและค่าพารามิเตอร์ของข้อสอบที่มีต่ออัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีซิบเทสท์กับวิธี MHโดยใช้ข้อมูลจำลองทำการศึกษา 2 ครั้ง ครั้งแรกใช้ขนาดกลุ่มตัวอย่าง 100, 200, 500 และ 1,000 คน และความแตกต่างของค่าเฉลี่ยของการกระจายความสามารถระหว่างกลุ่มเป็น 0 และ 1.0 ใช้แบบสอบจำนวน 25 ข้อ การศึกษาครั้งที่สอง ใช้ขนาดกลุ่มตัวอย่าง 500 1,000 และ 3,000 คน ความแตกต่างของค่าเฉลี่ยของการกระจายความสามารถระหว่างกลุ่มเป็น 0 และ 1.0 ค่าอำนาจจำแนก 3 ระดับ ค่าความยาก 5 ระดับ ค่าการเดา 3 ระดับ

ผลการศึกษาครั้งที่ 1 พบว่าอัตราความคลาดเคลื่อนประเภทที่ 1 เพิ่มขึ้นอย่างไม่มีนัยสำคัญทั้งสองวิธี ผลการศึกษาครั้งที่ 2 พบว่าเมื่อความแตกต่างของค่าเฉลี่ยของการกระจายความสามารถระหว่างกลุ่มเป็น 1.0 จะทำให้อัตราความคลาดเคลื่อนประเภทที่ 1 เพิ่มขึ้นมากทั้งสองวิธี โดยวิธี MH จะมีความคลาดเคลื่อนมากกว่าวิธีซิบเทสท์ และเมื่อไม่มีความแตกต่างของค่าเฉลี่ยของการกระจายความสามารถทั้งวิธีซิบเทสท์และวิธี MH ให้ผลที่น่าพอใจทุกเงื่อนไข

Oort (1998) ใช้วิธีวิเคราะห์องค์ประกอบจำกัด (RFA) ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยการจำลองข้อมูล ใช้คะแนนรวมจากการสอบเป็นตัวแทนความสามารถ และเปรียบเทียบประสิทธิภาพในการตรวจสอบด้วยวิธี RFA กับวิธี IRT แบบ 1 พารามิเตอร์ พบว่าในข้อสอบแบบแบ่งเป็นสอง เมื่อกำหนดขนาดของความลำเอียงของข้อสอบออกเป็น 3 ระดับคือ ข้อสอบที่มีความลำเอียงสูง (0.8 SD) ปานกลาง (0.5 SD) และต่ำ (0.2 SD) ขนาดกลุ่มตัวอย่างแบ่งเป็น 2 ขนาดคือ ขนาดใหญ่ (2,000 คน) และขนาดเล็ก (200 คน) และค่าเฉลี่ยของการกระจายความสามารถไม่เท่ากัน พบว่าเมื่อข้อสอบมีความลำเอียงปานกลางและข้อสอบที่มีความลำเอียงสูง วิธี RFA ให้ผลการตรวจสอบได้ดีกว่าวิธี IRT และในกลุ่มตัวอย่างขนาดเล็กการตรวจสอบด้วยวิธี RFA ให้ผลการตรวจสอบที่ดีกว่าและทั้งสองวิธีจะให้ผลการตรวจสอบสมบูรณ์ที่สุดเมื่อกลุ่มตัวอย่างมีขนาดใหญ่

กาญจนา วัชนศุนธร (2537) ได้พัฒนาเกณฑ์ในการตัดสินข้อสอบลำเอียงทางเพศ โดยใช้ข้อมูลเชิงประจักษ์ ใช้วิธีการตรวจสอบ 3 วิธี คือ วิธี IRT วิธีซิบเทสท์ และวิธี MH ดัชนีที่พัฒนาเพื่อเป็นเกณฑ์ในการตัดสินข้อสอบลำเอียงคือ SA, UA, α_{MH} , β_{SIB} ตามลำดับ ซึ่งในการวิจัยครั้งนี้ได้จากการวิเคราะห์ค่าเฉลี่ยดัชนีของแต่ละตัว ปัจจัยที่แปรเปลี่ยนในการศึกษาได้แก่ ความยาวแบบสอบ 20, 30 และ 40 ข้อในวิชาคณิตศาสตร์ และ 50, 60, 70 และ 80 ข้อในวิชาภาษาอังกฤษ ขนาดผู้สอบ 100, 200,400, 600, 800 และ 1,000 คน

ผลการวิจัยพบว่าขนาดของผู้สอบมีอิทธิพลต่อค่าเฉลี่ยของดัชนีทุกตัว ความยาวแบบสอบมีอิทธิพลต่อค่าเฉลี่ยของดัชนี SA และ UA แต่ไม่มีอิทธิพลต่อค่าเฉลี่ย α_{MH} และ β_{SIB} ซึ่งเกณฑ์ที่พัฒนาขึ้นเพื่อใช้ตัดสินความลำเอียงระหว่างผู้สอบเพศชายและเพศหญิง เป็นดังนี้

1. SA > 0.80 และ UA > 0.50 เมื่อความยาวแบบสอบน้อยกว่า 50 ข้อ
2. SA > 0.40 และ UA > 1.20 เมื่อความยาวแบบสอบ 50 ข้อขึ้นไป
3. $\alpha_{MH} > 0.60$ และ $\alpha_{MH} > 1.40$ สำหรับทุกขนาดของผู้สอบและความยาวแบบสอบ

4. $\beta_{SIB} > 0.06$ สำหรับทุกขนาดของผู้สอบและความยาวแบบสอบ

นอกจากนี้ กาญจนา วัชนสุนธร ยังพบว่าการใช้ดัชนี SA หรือ UA ควรใช้ผู้สอบขนาด 800 คนขึ้นไป ส่วนดัชนี α_{MH} และ β_{SIB} ควรใช้ขนาดผู้สอบอย่างน้อย 600 คน

จิติมา วรรณศรี (2539) ได้เปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันระหว่างวิธีแมนเทล-แฮนส์เซลกับวิธีชิบเทสท์ โดยศึกษาจากข้อมูลจำลอง ปัจจัยที่ศึกษาได้แก่ความยาวแบบสอบ 3 ขนาด คือ 30, 60 และ 90 ข้อ ขนาดกลุ่มตัวอย่าง 3 ขนาด คือ 200, 600 และ 1000 คน โดยแต่ละขนาดมีอัตราส่วนระหว่างผู้สอบกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบต่างกัน คือ 1:1 1:0.9 , 1:0.75 และ 1:0.5

ผลการศึกษาพบว่า วิธี MH กับวิธีชิบเทสท์ มีประสิทธิภาพเท่าเทียมกันในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ทุกขนาดกลุ่มตัวอย่างและทุกอัตราส่วนภายใต้ความยาวแบบสอบเดียวกัน และเมื่อใช้แบบสอบที่มีความยาวปานกลาง (60 ข้อ) ทั้งสองวิธีสามารถตรวจสอบได้อย่างมีประสิทธิภาพที่สุด นอกจากนี้เมื่อใช้ขนาดกลุ่มตัวอย่างมากขึ้นจะสามารถตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันได้ถูกต้องมากขึ้น โดยส่วนมากวิธีชิบเทสท์มีอัตราความคลาดเคลื่อนประเภทที่ 1 มากกว่าวิธี MH เล็กน้อย

เสรี ชัดแช้ม (2540) ศึกษาเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูประหว่างวิธี MH แบบปกติกับวิธี MH แบบแบ่งกลุ่มความสามารถของผู้สอบและความยากของข้อสอบโดยใช้วิธี IRT เป็นเกณฑ์โดยศึกษาจากข้อมูลผลการตอบแบบสอบวัดความสามารถในการอ่านภาษาไทยของนักเรียนชั้นมัธยมศึกษาปีที่ 1 สังกัดกรมสามัญศึกษาจังหวัดชลบุรี จำนวน 1200 คน โดยกลุ่มผู้สอบจำแนกตามเพศ

ผลการศึกษาพบว่า วิธี MH แบบแบ่งกลุ่มความสามารถของผู้สอบและความยากของข้อสอบ สามารถตรวจพบข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปได้สอดคล้องกับวิธี IRT และตรวจพบข้อสอบทำหน้าที่ต่างกันมากกว่าวิธี MH แบบปกติ ข้อสอบที่ตรวจพบส่วนใหญ่เป็นข้อสอบยากปานกลางและข้อสอบง่าย ซึ่งมีโค้งลักษณะข้อสอบของกลุ่มผู้สอบสองกลุ่มตัดกันบริเวณใกล้ ๆ จุดกลางของช่วงความสามารถ

Hambleton และคณะ (1993 อ้างถึงใน เสรี ชัดแช้ม, 2539) ได้ให้ข้อเสนอแนะเกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ มีรายละเอียดดังต่อไปนี้

ข้อเสนอแนะที่ 1 ไม่มีวิธีการใด ๆ ที่สามารถตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบได้ทั้งหมด แต่ละวิธีก็มีข้อบกพร่องในตัวเองและผลการตรวจพบข้อสอบทำหน้าที่ต่างกันก็แตก

ต่างกันออกไป ดังนั้น นักวัดผลที่ต้องการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในโครงการทดสอบที่สำคัญ ควรเลือกใช้วิธีการตรวจสอบที่มีประสิทธิภาพในแต่ละเงื่อนไข รวมทั้งการให้ผู้เชี่ยวชาญพิจารณาผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ไม่สอดคล้องกันด้วย

ข้อเสนอแนะที่ 2 นอกจากจะใช้วิธีทางสถิติในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแล้ว วิธีพิจารณาตัดสินข้อสอบก็เป็นวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่เป็นประโยชน์และมีข้อดีหลายประการ เป็นต้นว่า

1. วิธีพิจารณาตัดสินข้อสอบ เสียค่าใช้จ่ายถูกกว่าวิธีการทางสถิติ เนื่องจากไม่ต้องเก็บรวบรวมข้อมูล

2. การใช้ผู้ตัดสินข้อสอบที่เหมาะสม จะช่วยให้สามารถพิจารณาความตรงเฉพาะหน้าได้ และถ้าใช้ผู้ตัดสินข้อสอบที่มาจากคนกลุ่มสนใจจะเป็นประโยชน์ในด้านสังคม เชื้อชาติ และการเมือง

3. วิธีการพิจารณาตัดสินข้อสอบ สามารถดำเนินการได้ก่อนนำแบบสอบไปใช้ จึงทำให้สามารถคัดเลือกข้อสอบที่ไม่เหมาะสม ออกจากแบบสอบก่อนนำข้อสอบไปใช้ในสถานการณ์สอบจริง

4. ในกรณีที่ผู้ตัดสินข้อสอบมีความเชี่ยวชาญทางด้านเนื้อหาวิชา จะช่วยพิจารณาความตรงตามเนื้อหาของแบบสอบได้อีกด้วย

ส่วนข้อเสียของวิธีพิจารณาตัดสินข้อสอบ มีดังนี้

1. ผลการพิจารณาตัดสินข้อสอบทำหน้าที่ต่างกัน มักไม่สอดคล้องกับผลที่ได้จากวิธีการทางสถิติ

2. ในกรณีที่ต้องนำผู้ตัดสินข้อสอบมาพิจารณาข้อสอบร่วมกัน หรือต้องฝึกอบรมผู้ตัดสินข้อสอบทำให้ต้องเสียเวลาและค่าใช้จ่ายเพิ่มขึ้น

3. ผู้ตัดสินข้อสอบมักจะเกิดปัญหาเรื่องความเบื่อหน่าย ความเหนื่อยล้า ซึ่งอาจกระทบต่อความตรงของผลการพิจารณาตัดสินข้อสอบ

อย่างไรก็ตาม แม้ว่าวิธีพิจารณาตัดสินข้อสอบจะมีข้อเสียอยู่บ้าง แต่ก็เป็นประโยชน์ในการพิจารณาข้อสอบที่ทำหน้าที่ต่างกัน

จากการศึกษางานวิจัยที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ สามารถสรุปวิธีที่ได้มีการศึกษามาแล้ว ดังตารางที่ 3

ตารางที่ 3 สรุปงานวิจัยที่ได้ศึกษาเกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

ผู้ที่ศึกษา	ประเด็นที่ศึกษา	ผลการศึกษา
Ryan (1991)	ศึกษาความคงที่ของวิธี MH เมื่อเปลี่ยนขนาดกลุ่มตัวอย่าง	เมื่อเพิ่มขนาดกลุ่มตัวอย่างให้มากขึ้นวิธี MH จะมีประสิทธิภาพสูงขึ้น
Mazor,et al (1992)	ผลกระทบของขนาดกลุ่มตัวอย่างด้วยวิธี MH	เมื่อเพิ่มขนาดกลุ่มตัวอย่างให้มากขึ้นวิธี MH จะมีประสิทธิภาพสูงขึ้น
Roger and Swaminathan (1993)	เปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธีโลจิสติก กับวิธี MH	ทั้งสองวิธีมีประสิทธิภาพเท่าเทียมกัน
Mazor,et al (1994),	เปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปด้วยวิธี MH เมื่อแบ่งกลุ่มตัวอย่างออกเป็นกลุ่มสูงและกลุ่มต่ำ และเมื่อไม่มีการแบ่งกลุ่มตัวอย่างออกเป็นกลุ่มสูงและกลุ่มต่ำ	เมื่อแบ่งกลุ่มตัวอย่างออกเป็นกลุ่มสูงและกลุ่มต่ำ จะทำให้อำนาจการตรวจสอบสูงกว่าเมื่อไม่มีการแบ่งกลุ่มตัวอย่างออกเป็นกลุ่มสูงและกลุ่มต่ำ
Narayanan and Swaminathan (1994)	เปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปด้วยวิธี MH และวิธีSIBTEST	ทั้งสองวิธีมีอำนาจการตรวจสอบเท่าเทียมกัน แต่วิธี SIBTEST มีอัตราความคลาดเคลื่อนประเภทที่ 1 สูงกว่า
Roussos and Stout (1996)	ผลกระทบของกลุ่มตัวอย่างขนาดเล็กที่มีต่ออัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธี MH และวิธีSIBTEST	แต่วิธี MH มีอัตราความคลาดเคลื่อนประเภทที่ 1 สูงกว่า

ตารางที่ 3 (ต่อ)

ผู้ที่ศึกษา	ประเด็นที่ศึกษา	ผลการศึกษา
จิตมา วรรณศรี (2539)	เปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปด้วยวิธี MH และวิธี SIBTEST	ทั้งสองวิธีมีอำนาจการตรวจสอบเท่าเทียมกัน แต่วิธี SIBTEST มีอัตราความคลาดเคลื่อนประเภทที่ 1 สูงกว่า
เสรี ชัดแช้ม (2540)	เปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปด้วยวิธี MH แบบปกติกับแบบแบ่งกลุ่มความสามารถ	วิธี MH แบบแบ่งกลุ่มความสามารถจะตรวจพบข้อสอบทำหน้าที่ต่างกันได้มากกว่าวิธี MH แบบปกติ
Oort (1998)	เปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี RFA กับวิธี IRT แบบ 1 พารามิเตอร์	วิธี RFA มีประสิทธิภาพสูงกว่าวิธี IRT แบบ 1 พารามิเตอร์ ในกลุ่มตัวอย่างขนาดเล็ก เมื่อเพิ่มขนาดกลุ่มตัวอย่างมากขึ้นทั้งสองมีประสิทธิภาพสูง พอ ๆ กัน