บทที่ 2
วรรณคดีที่เกี่ยวข้อง

เนื้อหาที่จะเสนอในบทนี้แบ่งเป็น 5 ตอนดังนี้

2.1.แนวคิดเกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบ

2.2.การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมนเทล-แฮนส์เซล

2.3.การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีซิบเทสท์

2.4.การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีทฤษฎีการตอบสนองข้อสอบ

2.5.งานวิจัยที่เกี่ยวข้อง

2.1 แนวคิดเกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบ

การทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning) เดิมเรียกว่าความลำเอียงของข้อสอบ (item bias) แต่ในระยะหลังเปลี่ยนมาใช้คำว่าการทำหน้าที่ต่างกันของข้อสอบ เพราะเป็นคำที่เป็นกลางและมีความเหมาะสมมากกว่า (Holland and Thayer,1988) มีผู้ให้ความหมายของคำว่าการทำหน้าที่ต่างกันของข้อสอบไว้ดังนี้

Shealy และ Stout (1993) กล่าวว่า ข้อสอบที่ทำหน้าที่ต่างกันหมายถึง ข้อสอบที่เข้าข้าง (favor) ผู้สอบกลุ่มหนึ่งมากกว่าผู้สอบอีกกลุ่มหนึ่งที่นำมาจับคู่เปรียบเทียบกัน ซึ่งทำให้ผู้สอบกลุ่มหนึ่งได้ประโยชน์แต่ผู้สอบอีกกลุ่มหนึ่งเสียประโยชน์

Potenza และ Dorans (1995) กล่าวว่า ถ้าข้อสอบทำหน้าที่ต่างกัน จะทำให้ผลการตอบข้อสอบระหว่างกลุ่มผู้สอบที่นำมาเปรียบเทียบกันแตกต่างกัน

Kim และคณะ (1994) สรุปว่า ข้อสอบจะทำหน้าที่ต่างกัน เมื่อฟังก์ชั่นการตอบข้อสอบจากผู้สอบต่างกลุ่ม มีค่าต่างกัน

Mazor และ Clauser (1995) กล่าวว่า ข้อสอบทำหน้าที่ต่างกัน เมื่อผู้สอบต่างกลุ่มที่มีความสามารถระดับเดียวกัน มีโอกาสในการตอบข้อสอบถูกแตกต่างกัน

ดังนั้นสรุปได้ว่าเมื่อข้อสอบวัดคุณลักษณะแฝงอื่น นอกเหนือจากคุณลักษณะแฝงที่ต้องการวัด จะส่งผลให้ผู้สอบต่างกลุ่มที่นำมาจับคู่เปรียบเทียบกัน มีโอกาสในการตอบข้อสอบถูกแตกต่างกัน ทั้ง ๆ ที่มีความสามารถที่ต้องการวัดเท่ากัน แสดงว่าข้อสอบข้อนั้นทำหน้าที่ต่างกัน

การทำหน้าที่ต่างกันของข้อสอบมี 2 แบบ คือ

2.1.1.การทำหน้าที่ต่างกันแบบเอกรูป (uniform DIF) หมายถึงผู้สอบกลุ่มหนึ่งมีโอกาสในการตอบข้อสอบถูกมากกว่าอีกกลุ่มหนึ่งในทุกระดับความสามารถ หรือเมื่อพิจารณาโค้งคุณลักษณะข้อสอบระหว่างกลุ่ม จะพบว่าไม่มีปฏิสัมพันธ์ระหว่างโค้งคุณลักษณะข้อสอบในทุกระดับความสามารถ

2.1.2.การทำหน้าที่ต่างกันแบบอเนกรูป (nonuniform DIF) หมายถึง โอกาสในการตอบข้อสอบถูกของผู้สอบระหว่างกลุ่มไม่สม่ำเสมอ เมื่อพิจารณาแต่ละระดับความสามารถ หรือเมื่อพิจารณาโค้งคุณลักษณะข้อสอบระหว่างกลุ่ม จะพบว่ามีปฏิสัมพันธ์ระหว่างโค้งคุณลักษณะข้อสอบ เช่น ที่ความสามารถระดับหนึ่ง กลุ่ม A มีโอกาสในการตอบข้อสอบถูกมากกว่ากลุ่ม B แต่ที่ความสามารถอีกระดับหนึ่ง กลุ่ม B มีโอกาสในการตอบข้อสอบถูกมากกว่ากลุ่ม A เป็นต้น

ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ดำเนินการโดยการเปรียบเทียบผลการตอบข้อสอบระหว่างผู้สอบ 2 กลุ่ม ได้แก่กลุ่มเปรียบเทียบซึ่งเป็นกลุ่มที่สนใจศึกษา และเป็นกลุ่มที่คาดว่าจะเสียประโยชน์ในการตอบข้อสอบ คือมีโอกาสในการตอบข้อสอบถูกน้อยกว่าผู้สอบอีกกลุ่มหนึ่ง และกลุ่มอ้างอิงซึ่งเป็นกลุ่มที่คาดว่าจะได้ประโยชน์ในการตอบข้อสอบ คือมีโอกาสในการตอบข้อสอบถูกมากกว่าผู้สอบอีกกลุ่มหนึ่ง การจำแนกผู้สอบเป็นกลุ่มอ้างอิงและกลุ่มเปรียบเทียบนั้นมีหลายลักษณะ เช่น จำแนกตามเพศ สีผิว เชื้อชาติ ภาษา สถาบันการศึกษา เป็นต้น

วิธีการในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมีหลายวิธี โดยแบ่งตามประเภทการวิเคราะห์ได้เป็น 2 กลุ่ม (Millsap and Everson,1993;Potenza and Dorans,1995) ได้แก่

กลุ่มที่ 1 ใช้คะแนนที่สังเกตได้ (observed score) ได้แก่ วิธีโคสแควร์ วิธี Loglinear models วิธี MH วิธี Standardization และวิธี Logistic regression

กลุ่มที่ 2 ใช้คะแนนที่สังเกตไม่ได้หรือตัวแปรแฝง (Latent Variable) ได้แก่ วิธีทฤษฎีการตอบสนองข้อสอบ (IRT) และวิธี SIBTEST ซึ่งใช้สถิตินอนพาราเมตริก (nonparametric)

2.2 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมนเทล-แฮนส์เซล(MH)

วิธี MH เป็นวิธีที่ใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่เสนอโดย Mantel และ Haenszel ในปี ค.ศ.1959 ผู้ที่เริ่มนำมาใช้ในการตรวจสอบคือ Holland (1985) Holland และ Thayer (1988) หลังจากนั้นวิธี MH ก็นิยมนำไปใช้กันอย่างกว้างขวาง วิธี MHเป็นวิธีที่พัฒนาจากวิธีไคสแควร์แบบดั้งเดิม (traditional χ^2 approach)

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีนี้เป็นการเปรียบเทียบผลการตอบของผู้สอบย่อย 2 กลุ่ม คือ กลุ่มเปรียบเทียบ (focal group)ซึ่งเป็นกลุ่มที่คาดว่าจะเสียประโยชน์ในการตอบข้อสอบในกรณีที่ข้อสอบทำหน้าที่ต่างกัน และกลุ่มอ้างอิง (reference group)เป็นกลุ่มที่คาดว่าจะได้ประโยชน์ในการตอบข้อสอบ โดยจะเปรียบเทียบตามระดับคะแนนรวมจากการสอบ มีวิธีการตรวจสอบดังนี้

พิจารณาผลการตอบข้อสอบของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่มีคะแนนรวมเท่ากัน แล้วแสดงความถี่ของผู้สอบทั้งสองกลุ่มที่ตอบข้อสอบข้อนั้นถูก (1) หรือผิด (0) ลงในตาราง 2 ทาง

ตารางที่ 2 ความถี่ของผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่ระดับคะแนน j

กลุ่ม	คะแนนของข้อสอบที่ต้องการตรวจสอบ		
	1	0	รวม
อ้างอิง	Aj	Bj	n_{Rj}
เปรียบเทียบ	Cj	Dj	n_{Fj}
รวม	m_{1j}	m_{0j}	Tj

โดยที่ Tj เป็นความถี่ของผู้สอบทั้งหมดที่ได้ระดับคะแนน j

n_{Rj} , n_{Fj} เป็นความถี่ของผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบตามลำดับ

m_{1j} , m_{0j} เป็นความถี่ของผู้สอบที่ตอบข้อสอบถูกและผิดตามลำดับ

Aj , Bj , Cj , Dj เป็นความถี่ของผู้สอบที่ตอบถูก(1)และผิด(0)ของแต่ละกลุ่มในระดับคะแนนที่ j

ตารางที่ 3 อัตราส่วนการตอบข้อสอบของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่ระดับคะแนน j

กลุ่ม	คะแนนของข้อสอบที่ต้องการตรวจสอบ		
	1	0	รวม
อ้างอิง	P_{Rj}	q_{Rj}	1
เปรียบเทียบ	P_{Fj}	q_{Fj}	1

โดยที่ P_{Rj}, q_{Rj} คืออัตราส่วนการตอบข้อสอบของกลุ่มอ้างอิงที่ตอบถูกและตอบผิดตามลำดับ

P_{Fj}, q_{Fj} คืออัตราส่วนการตอบข้อสอบของกลุ่มเปรียบเทียบที่ตอบถูกและตอบผิดตามลำดับ

ดังนั้นจึงมีสมมติฐานสูญดังนี้ $H_0 : P_{Rj} = P_{Fj}$ สำหรับทุกj

หรือ $H_0 : \frac{AjDj}{Tj} = \frac{BjCj}{Tj}$ สำหรับทุก j

ขั้นตอนในการวิเคราะห์ของวิธี MH มีดังนี้

1.คำนวณความน่าจะเป็นในการตอบข้อสอบระหว่างกลุ่มของแต่ละข้อในช่วงคะแนน j โดยใช้สูตร

$$\alpha_{MH} = \frac{\Sigma AjDj / Tj}{\Sigma BjCj / Tj}$$

ค่า α_{MH} มีค่าระหว่าง 0 ถึง ∞ ถ้าค่า α_{MH} ที่คำนวณได้มีค่าเท่ากับ 1 แสดงว่ากลุ่มตัวอย่างทั้งสองกลุ่มมีความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องเท่ากัน นั่นคือข้อสอบทำหน้าที่ไม่ต่างกัน ถ้า $\alpha_{MH} > 1$ แสดงว่าข้อสอบเข้าข้างกลุ่มอ้างอิง ถ้า $\alpha_{MH} < 1$ แสดงว่าข้อสอบเข้าข้างกลุ่มเปรียบเทียบ

2.ทดสอบนัยสำคัญด้วยค่าสถิติไคสแควร์ เพื่อทดสอบว่าค่าที่ได้จะมีความแตกต่างจาก 1 อย่างมีนัยสำคัญหรือไม่ที่ระดับชั้นความเป็นอิสระเท่ากับ 1

$$\chi^2_{MH} = \frac{[\,|\,\Sigma A_j - \Sigma E(A_j)\,|\, - .5\,]}{\Sigma Var(A_j)}$$

เมื่อ $E(A_j) = \frac{n_{Rj}\, m_{1j}}{T_j}$

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

Holland และ Thayer ได้เสนอให้แปลงค่า α_{MH} ให้เป็น Δ_{MH} (ค่าเดลต้า) ดังนี้

$$\Delta_{MH} = -2.35\ \mathrm{In}\ (\alpha_{MH})$$

Δ_{MH} มีค่าระหว่าง -2.6 ถึง 2.6 ถ้า $\Delta_{MH} = 0$ แสดงว่าข้อสอบทำหน้าที่ไม่ต่างกัน ถ้ามีค่าเป็นลบแสดงว่าข้อสอบเข้าข้างกลุ่มอ้างอิง

2.3 <u>การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี SIBTEST</u>

เป็นวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่เสนอโดย Shealy และStout ในปี ค.ศ. 1991 ซึ่งมีแนวคิดในการตรวจสอบความลำเอียงของแบบสอบชนิดพหุมิติโดยมีพื้นฐานอยู่บนทฤษฎีการตอบข้อสอบแบบพหุมิติ และมีข้อตกลงว่ามีมิติการวัด 2 มิติ คือมิติลักษณะแฝงเป้าหมายที่ต้องการวัด(θ)และมิติลักษณะแฝงแทรกซ้อนที่ไม่ต้องการวัด (η) มีฟังก์ชั่นการตอบข้อสอบเป็น P (θ,η)โดยที่ข้อสอบทุกข้อจะวัดคุณลักษณะแฝงเป้าหมาย (θ) และบางข้อ (ซึ่งทำหน้าที่ต่างกัน)จะวัดทั้งคุณลักษณะแฝงเป้าหมายและคุณลักษณะแฝงแทรกซ้อน

การตรวจสอบจะเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ซึ่งผลการตอบข้อสอบนั้นมีอยู่ 2 ค่า โดยได้จากแบบสอบย่อย 2 ชุด กล่าวคือในการตรวจสอบด้วยวิธี SIBTEST นี้จะแบ่งแบบสอบจากเดิมที่มี 1 ชุด ให้เป็น 2 ชุดย่อย คือแบบสอบย่อยที่มีความตรง (valid subtest) เป็นแบบสอบที่ประกอบด้วยข้อสอบที่มีความตรง วัดได้ตามที่ต้องการวัด และแบบสอบที่ศึกษา(studied subtest)เป็นแบบสอบที่ประกอบด้วยข้อสอบที่สงสัยว่าจะทำหน้าที่ต่างกันโดยที่

$X = \sum_{i=1}^{n} Ui$ โดยที่ X คือคะแนนรวมจากแบบสอบที่มีความตรง

Ui คือผลการตอบข้อที่ i ได้ 1 ถ้าตอบถูกได้ 0 ถ้าตอบผิด

$Y = \sum_{i=n+1}^{n} Ui$ โดยที่ Y คือคะแนนรวมจากแบบสอบที่ศึกษา

ในการวิเคราะห์จะพิจารณาจากผลการตอบข้อสอบจากแบบสอบทั้ง 2 ชุดย่อย โดยผู้ที่ได้คะแนนรวมเท่ากันจากแบบสอบที่มีความตรงของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบมาจับคู่เปรียบเทียบและใช้คะแนนรวมจากแบบสอบที่ศึกษาของผู้สอบเหล่านี้ในการคำนวณ

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

โดย k คือระดับคะแนนรวมจากแบบสอบที่มีความตรงของผู้สอบ

$\bar{Y}_{Rk}$ คือคะแนนเฉลี่ยจากแบบสอบที่ศึกษาของผู้สอบกลุ่มอ้างอิงทุกคนที่มีคะแนนจากแบบสอบที่มีความตรงที่ระดับ k

$\bar{Y}_{fk}$ คือคะแนนเฉลี่ยจากแบบสอบที่ศึกษาของผู้สอบกลุ่มเปรียบเทียบทุกคนที่มีคะแนนจากแบบสอบที่มีความตรงที่ระดับ k

ดังนั้น $(\bar{Y}_{Rk} - \bar{Y}_{Fk})$ คือความแตกต่างของผลการตอบที่ได้จากแบบสอบที่ศึกษาระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่มีความสามารถเท่ากัน (X=k) ถ้า $\bar{Y}_{Rk} - \bar{Y}_{fk} = 0$ แสดงว่าข้อสอบในแบบสอบที่ศึกษาทำหน้าที่ไม่ต่างกัน

โดยมีสมมติฐานว่า $H_0 : \beta u = 0$

$H_1 : \beta u > 0$

ขั้นตอนในการวิเคราะห์มีดังนี้

1.ประมาณค่า βu ซึ่งเป็นดัชนีที่บ่งชี้การทำหน้าที่ต่างกันของข้อสอบ จากสูตร

$$\beta u = \sum_{k=0}^{n} P_K(\bar{Y}_{RK} - \bar{Y}_{FK})$$

โดย Pk คืออัตราส่วนของผู้สอบกลุ่มเปรียบเทียบที่ตอบแบบสอบที่มีความตรงได้ถูกต้อง k ข้อ

2.ทดสอบนัยสำคัญทางสถิติ

$$B = \frac{\beta u}{\sigma(\beta u)} \quad ; N(0,1) \quad \text{เมื่อ} \quad \beta u = 0$$

โดยที่

$$\sigma(\beta u) = \left(\sum_{k=0}^{n} Pk^2 \left[1/J_{Rk}\ \sigma^2(Y \mid k,R) + 1/J_{Fk}\ \sigma^2(Y \mid k,F)\right]\right)^{1/2}$$

เมื่อ $\sigma(\beta u)$ คือความคลาดเคลื่อนในการประมาณค่าของ βu

$\sigma^2(Y \mid k,R)$, $\sigma^2(Y \mid k,F)$ เป็นความแปรปรวนของคะแนนกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่ระดับคะแนน k

J_{RK}, J_{FK} คือจำนวนผู้สอบของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่มีคะแนนจากแบบสอบที่มีความตรงระดับ k

2.4 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีทฤษฎีตอบสนองข้อสอบ

วิธีทฤษฎีตอบสนองข้อสอบเป็นวิธีที่ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยการพิจารณาเปรียบเทียบฟังก์ชั่นการตอบข้อสอบ (IRFs)ระหว่างกลุ่มผู้สอบที่มีต่อแบบสอบชุดเดียวกัน (Lord,1980 cited in Green,1994) ถ้าฟังก์ชั่นการตอบข้อสอบต่างกันระหว่างกลุ่มผู้สอบ แสดงว่าข้อสอบทำหน้าที่ต่างกัน (Kim et al.,1944) หรือโดยการเปรียบเทียบความแตกต่างระหว่างโค้งคุณลักษณะข้อสอบ (ICCs) ระหว่างกลุ่ม ซึ่งพื้นที่ระหว่างโค้งคุณลักษณะข้อสอบจะเป็นดัชนีบอกระดับของการทำหน้าที่ต่างกันของข้อสอบ (Osterlind,1992)

ในการวัดพื้นที่ระหว่างโค้งคุณลักษณะข้อสอบจะมี 2 ลักษณะคือ การวัดช่วงเปิด (open interval) จะเป็นการวัดในช่วงความสามารถทั้งหมดระหว่างโค้งทั้งสอง ซึ่งจะทำให้ได้พื้นที่แน่นอน และการวัดช่วงปิด (closed interval) จะวัดในช่วงความสามารถตามที่กำหนดไว้ ซึ่งมีดัชนีที่ใช้บอกระดับการทำหน้าที่ต่างกันของข้อสอบ ได้แก่ พื้นที่ขนิดไม่มีเครื่องหมาย (unsigned areas) เป็นค่าสัมบูรณ์ของพื้นที่ระหว่างโค้งคุณลักษณะข้อสอบ พื้นที่ขนิดมีเครื่องหมาย (signed areas) เหมือนกับ

พื้นที่ชนิดไม่มีเครื่องหมาย แต่จะมีเครื่องหมายแสดงให้ทราบว่ากลุ่มใดได้ประโยชน์กลุ่มใดเสียประโยชน์ ไคสแควร์ (IRT-χ^2) เป็นการทดสอบนัยสำคัญของความแตกต่างของค่าพารามิเตอร์ a และ b ระหว่างกลุ่ม ในคราวเดียวกัน ซึ่งเป็นวิธีของ Lord (Shepard et al.,1985)

การตรวจสอบด้วยวิธีนี้มีข้อตกลงเบื้องต้นเช่นเดียวกับทฤษฎีการตอบสนองข้อสอบ วิธีทฤษฎีการตอบสนองข้อสอบโมเดล 3 พารามิเตอร์นี้จะปล่อยให้ค่าความยาก ค่าอำนาจจำแนก และค่าการเดาแปรเปลี่ยนไปตามกลุ่ม การคำนวณจะประมาณค่าพารามิเตอร์ของข้อสอบและของผู้สอบโดยรวมทั้งสองกลุ่มก่อน แล้วแปลงค่าความยากให้เป็นค่ามาตรฐาน จากนั้นประมาณค่าพารามิเตอร์ใหม่โดยแยกกันระหว่างกลุ่มผู้สอบ ค่าที่ต้องประมาณใหม่คือค่าความยากและค่าอำนาจจำแนก ส่วนค่าการเดาใช้ค่าเดิมที่ได้จากการประมาณครั้งแรก แล้วแปลงค่าความยากที่ได้ใหม่ให้เป็นค่ามาตรฐานอีกครั้ง นำค่าเหล่านี้ไปใช้ในการคำนวณฟังก์ชั่นการตอบข้อสอบและพื้นที่ระหว่างโค้งคุณลักษณะข้อสอบ

สูตรที่ใช้ในการคำนวณค่าฟังก์ชั่นการตอบข้อสอบ (Hambleton and swaminathan,1985 cited in Feinstein,1995) คือ

$$Pi(\theta) = Ci + (1-Ci)\ \{\ 1+ \exp\ [\ -Dai\ (\ \theta - bi)\ ]\ \}^{-1}$$

โดย $Pi(\theta)$ คือโอกาสในการตอบข้อสอบถูกของผู้สอบที่มีระดับความสามารถ θ

i คือข้อสอบข้อที่ i,i=1,...,n

a , b และ c คือค่าพารามิเตอร์ของข้อสอบ

D คือค่าคงที่มีค่าเท่ากับ 1 .7

สูตรการคำนวณพื้นที่ระหว่างจุดสองจุดบนมาตรความสามารถ (Kim and Cohen,1991 cited in Feinstein,1995) คือ

$$Si\ (\theta_L - \theta_U) = Ci\ (\theta_U - \theta_L)\ +\ \frac{1-Ci}{Dai}\ \ In\left[\frac{1+\exp[Dai(\theta u - bi)]}{1+\exp[Dai(\theta_L - bi)]}\right]$$

โดย Si คือพื้นที่ระหว่างจุดสองจุด

θ_L คือระดับความสามารถที่ต่ำกว่า

θ_U คือระดับความสามารถที่สูงกว่า

พื้นที่ชนิดมีเครื่องหมายคำนวณจากสูตร

$$CSA = S_R(\theta_L, \theta_U) - S_F(\theta_L, \theta_U)$$

โดย S_R คือพื้นที่ของกลุ่มอ้างอิง

S_F คือพื้นที่ของกลุ่มเปรียบเทียบ

พื้นที่ชนิดไม่มีเครื่องหมายคำนวณจากสูตร

$$CUA = \left| S_R(\theta_L, \theta_U) - S_F(\theta_L, \theta_U) \right|$$

เมื่อฟังก์ชั่นการตอบข้อสอบตัดกันหนึ่งจุดหรือมากกว่าคำนวณโดย

$$CUA = \int \left| P_R(\theta) - P_F(\theta) \right| d\theta \quad \text{หรือ}$$

$$CUA = \sum_{j=1}^{n} \left| P_R(\theta) - P_F(\theta) \right| \Delta\theta + \frac{1}{2} \left[\left| P_R(\theta_L) - P_F(\theta_L) \right| - \left| P_R(\theta_U) - P_F(\theta_U) \right| \right] \Delta\theta$$

2.5. งานวิจัยที่เกี่ยวข้อง

Ryan (1991) ได้ศึกษาความคงที่(stability)ของวิธีแมนเทล-แฮนส์เซลด้วยการแปรเปลี่ยนกลุ่มตัวอย่างและการจับคู่เปรียบเทียบ กลุ่มตัวอย่างเป็นนักเรียนเกรด 8 เป็นกลุ่มนักเรียนผิวขาว 5,015 คนและกลุ่มนักเรียนผิวดำ 670 คน ในแต่ละกลุ่มจะแบ่งเป็นกลุ่มย่อย 4 กลุ่มโดยวิธีสุ่ม แบบสอบที่ใช้เป็นวิชาคณิตศาสตร์ประกอบด้วยเนื้อหาด้านพีชคณิต เรขาคณิต เลขคณิต การวัด และสถิติ ซึ่งแบบสอบที่จะใช้วัดความสามารถเพื่อจับคู่เปรียบเทียบนั้นศึกษา 3 แบบได้แก่ (1) ข้อสอบร่วม 40 ข้อ (2) ข้อสอบเวียนที่สุ่มจากเนื้อหาต่าง ๆ ฉบับละ 35 ข้อจำนวน 4 ฉบับ (3) ข้อสอบที่ได้จากการรวมข้อสอบร่วมกับข้อสอบเวียนเป็นฉบับละ 75 ข้อจำนวน 4 ฉบับ เงื่อนไขที่ศึกษาจะแปรเปลี่ยนกลุ่มตัวอย่าง(ระหว่างผิวขาวกับผิวขาว,ระหว่างผิวขาวกับผิวดำ) ขนาดกลุ่มตัวอย่าง(กลุ่มใหญ่และกลุ่มย่อย)เกณฑ์ ในการจับคู่เปรียบเทียบระหว่างกลุ่มใช้ 2 เกณฑ์ เกณฑ์

ที่ 1 พิจารณาจากคะแนนจากแบบสอบร่วม 40 ข้อและเกณฑ์ที่ 2 พิจารณาจากคะแนนแบบสอบแบบที่ 3 ที่มีข้อสอบ 75 ข้อ

ผลการศึกษาพบว่า เมื่อนำค่า MH D-DIF ที่ได้จากการวิเคราะห์แต่ละเงื่อนไขมาหาความสัมพันธ์ ในเงื่อนไขที่ใช้กลุ่มตัวอย่างใหญ่กับกลุ่มตัวอย่างย่อยระหว่างผิวขาวกับผิวดำ เมื่อจับคู่เปรียบเทียบโดยใช้เกณฑ์ที่ 1 มีค่าสหสัมพันธ์ระหว่าง .74 ถึง .88 เมื่อใช้เกณฑ์ที่ 2 มีค่าสหสัมพันธ์ระหว่าง .75 ถึง .88 ซึ่งแสดงให้เห็นว่าเกณฑ์ที่ใช้ในการจับคู่เปรียบเทียบไม่มีผลกระทบต่อค่า MH D-DIF ไม่ว่าจะใช้กลุ่มตัวอย่างกลุ่มใหญ่หรือกลุ่มย่อย สรุปได้ว่าดัชนี MH มีความแกร่งต่อผลกระทบของบริบทข้อสอบและหากต้องการให้มีความคงที่ในการประมาณค่าจากวิธีแมนเทล-แฮนส์เซลควรใช้กลุ่มตัวอย่างที่มีขนาดใหญ่ขึ้น

Mazor และคณะ (1992) ได้ศึกษาผลกระทบของขนาดกลุ่มตัวอย่างที่มีต่อการตรวจสอบการทำหน้าที่ต่างกันด้วยวิธีแมนเทล-แฮนส์เซล โดยศึกษาจากข้อมูลจำลอง กลุ่มตัวอย่างที่ใช้มี 5 ขนาด คือ 100 200 500 1000และ2000 คน ความยาวของแบบสอบ 75 ข้อ พบว่าเมื่อขนาดกลุ่มตัวอย่างเท่ากับ 500 คนหรือน้อยกว่า จะสามารถระบุข้อสอบที่ทำหน้าที่ต่างกันได้ถูกต้องน้อยกว่าร้อยละ 50 และเมื่อขนาดกลุ่มตัวอย่างเท่ากับ 2000 คน จะสามารถระบุข้อสอบที่ทำหน้าที่ต่างกันได้ถูกต้องร้อยละ 70ถึงร้อยละ 75 และได้กล่าวว่าข้อสอบที่ไม่สามารถตรวจสอบพบหรือระบุว่าทำหน้าที่ต่างกันได้เนื่องมาจากข้อสอบเหล่านั้นมีความยากมากหรือมีความยากต่างกันเพียงเล็กน้อยระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ อีกทั้งเป็นข้อที่มีค่าอำนาจจำแนกต่ำ

Raju และคณะ (1993) ได้ศึกษาเปรียบเทียบการประเมินการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีวัดขนาดพื้นที่(the area mothods) วิธีทดสอบไคสแควร์ของลอร์ด(Lord's chi-square test)และวิธีแมนเทล-แฮนส์เซล(Mantel-Haenszel) ซึ่งสองวิธีแรกนั้นเป็นวิธีที่มีพื้นฐานบนทฤษฎีการตอบสนองข้อสอบ(IRT) ใช้กลุ่มตัวอย่างทั้งหมด 839 คน ในการแบ่งกลุ่มตัวอย่างเป็นกลุ่มอ้างอิงและกลุ่มเปรียบเทียบนั้น ในกรณีแรกแบ่งตามเพศและกรณีที่สองแบ่งตามสีผิว ส่วนเครื่องมือที่ใช้เป็นแบบสอบวัดความรู้เกี่ยวกับศัพท์จำนวน 45 ข้อแต่ละข้อมี 5 ตัวเลือก

ผลการศึกษาพบว่าวิธีการวัดขนาดพื้นที่และวิธีทดสอบไคสแควร์ของลอร์ดให้ผลสอดคล้องกันอย่างมีนัยสำคัญในการระบุข้อสอบที่ทำหน้าที่ต่างกัน และวิธีแมนเทล-แฮนส์เซลให้ผล

สอดคล้องสูงมากกับวิธีวัดขนาดพื้นที่และวิธีทดสอบไคสแควร์ของลอร์ดในกรณีที่แบ่งกลุ่มตามเพศ ส่วนกรณีที่แบ่งกลุ่มตามสีผิวทั้งสามวิธีให้ผลแตกต่างกัน

Roger และ Swaminathan (1993) ได้เปรียบเทียบวิธีถดถอยโลจิสต์กับวิธีแมนเทล-แฮนส์เซล ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ซึ่งทดสอบเกี่ยวกับการกระจายของสถิติทดสอบ และประสิทธิภาพของสถิติทดสอบของแต่ละวิธี โดยใช้ข้อมูลจำลอง การศึกษาด้านการกระจาย ของสถิติทดสอบ ปัจจัย(factor)ที่แปรเปลี่ยนได้แก่ ขนาดกลุ่มตัวอย่าง ความเหมาะสมของข้อมูลกับ โมเดล ค่าความยาก ค่าอำนาจจำแนก ความยาวแบบสอบ 40 ข้อ ส่วนการศึกษาด้านประสิทธิ ภาพของแต่ละวิธี ปัจจัย(factor)ที่แปรเปลี่ยนได้แก่ ขนาดกลุ่มตัวอย่าง ความเหมาะสมของโมเดล กับข้อมูล ขนาดของแบบสอบ การกระจายของคะแนนสอบ อัตราส่วนของข้อสอบที่ทำหน้าที่ต่าง กัน ค่าความยาก ค่าอำนาจจำแนกและพื้นที่ระหว่างโค้งคุณลักษณะข้อสอบระหว่างกลุ่มอ้างอิงกับ กลุ่มเปรียบเทียบ

ผลการศึกษาพบว่าการกระจายของสถิติเป็นไปตามที่คาดไว้เกือบทั้งหมดทั้งสองวิธี กรณีที่ การกระจายของสถิติของวิธีถดถอยโลจิสต์ไม่เป็นตามที่คาดไว้เนื่องมาจากข้อสอบยากมากและค่า อำนาจจำแนกสูง ด้านประสิทธิภาพพบว่าทั้งสองวิธีมีประสิทธิภาพเท่ากันในการตรวจสอบ DIF แบบเอกรูป(uniform DIF) แต่วิธีถดถอยโลจิสต์มีประสิทธิภาพมากกว่าในการตรวจสอบ DIF แบบ อเนกรูป(non-uniform DIF) ขนาดกลุ่มตัวอย่างเป็นปัจจัยที่มีผลกระทบอย่างมากต่ออัตราการตรวจ สอบด้วยสองวิธีนี้คือเมื่อเพิ่มขนาดกลุ่มตัวอย่าง อัตราการตรวจสอบจะเพิ่มขึ้น ส่วนขนาดของแบบ สอบและการกระจายของคะแนนไม่มีผลกระทบต่ออัตราการตรวจสอบ

Mazor และคณะ (1994) ใช้วิธีแมนเทล-แฮนส์เซลในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน แบบอเนกรูป กลุ่มตัวอย่างได้จากการจำลองกลุ่มละ 1000 คน ใช้แบบสอบ 25 ฉบับ ๆ ละ 75 ข้อ ในแต่ละฉบับมีข้อสอบที่ทำหน้าที่ไม่ต่างกัน(no-DIF)จำนวน 59 ข้อและข้อสอบที่ศึกษาซึ่งทำหน้าที่ ต่างกัน(DIF)จำนวน 16 ข้อ โดยข้อสอบที่ศึกษาจะแปรเปลี่ยนค่าพารามิเตอร์ดังนี้ ค่าอำนาจจำแนก 4 ระดับ ความแตกต่างของค่าอำนาจจำแนกระหว่างกลุ่ม 5 ระดับ ค่าความยาก 5 ระดับ ความ แตกต่างของค่าความยากระหว่างกลุ่ม 4 ระดับ ค่าการเดากำหนดเป็น 0.2 และการกระจายของ ความสามารถระหว่างกลุ่ม 2 แบบคือการกระจายความสามารถเท่ากันและไม่เท่ากัน

ในการตรวจสอบด้วยวิธีแมนเทล-แฮนส์เซลจะวิเคราะห์ 2 แบบคือ แบบที่ 1 จะใช้คะแนนรวมของแต่ละคนเป็นเกณฑ์ในการจับคู่เปรียบเทียบระหว่างกลุ่ม ส่วนแบบที่ 2 จะแยกวิเคราะห์เฉพาะกลุ่มตัวอย่างที่เป็นกลุ่มสูงหรือกลุ่มต่ำ โดยใช้ค่าเฉลี่ยจากคะแนนของกลุ่มตัวอย่างทุกคนเป็นเกณฑ์ในการแยกเป็นกลุ่มสูงกลุ่มต่ำ จากนั้นจึงนำเฉพาะกลุ่มสูงหรือกลุ่มต่ำที่ได้มาแบ่งเป็นกลุ่มอ้างอิงและกลุ่มเปรียบเทียบแล้ววิเคราะห์

ผลการศึกษาพบว่าการวิเคราะห์โดยแบ่งกลุ่มตัวอย่างเป็นกลุ่มสูงกลุ่มต่ำจะทำให้อัตราการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปสูงกว่าการวิเคราะห์แบบที่ 1 (รวมผู้สอบกลุ่มสูงและกลุ่มต่ำไว้ด้วยกัน) อีกทั้งไม่ทำให้อัตราความคลาดเคลื่อนชนิดที่ 1 เพิ่มขึ้น และพบว่าเมื่อมีความแตกต่างของค่าอำนาจจำแนกและค่าความยากระหว่างกลุ่มเพิ่มขึ้นจะทำให้อัตราการตรวจสอบพบข้อที่ทำหน้าที่ต่างกันได้มากขึ้น

Uttaro และ Millsap (1994) ศึกษาปัจจัยที่มีผลต่อวิธีแมนเทล-แฮนส์เซลในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยใช้ข้อมูลจำลอง ขนาดกลุ่มตัวอย่างกลุ่มละ 500 คน ตัวแปรที่ศึกษาได้แก่ความยาวแบบสอบ 20 และ 40 ข้อ ค่าอำนาจจำแนก 3 ระดับ ค่าความยาก 3 ระดับ ค่าการเดา 2 ระดับ การกระจายความสามารถ 2 แบบ ซึ่งตัวแปรเหล่านี้จะศึกษาภายใต้เงื่อนไขที่ข้อสอบทำหน้าที่ต่างกัน (DIF conditions ซึ่งเป็นผลเนื่องมาจากฟังก์ชั่นการตอบข้อสอบของกลุ่มเปรียบเทียบจะแปรเปลี่ยนไปตามที่ศึกษา ส่วนฟังก์ชั่นการตอบของกลุ่มอ้างอิงจะกำหนดที่ a=1.0 , b=0.0 , c=0.2) และเงื่อนไขที่ข้อสอบทำหน้าที่ไม่ต่างกัน (no-DIF conditions ซึ่งเป็นผลเนื่องมาจากฟังก์ชั่นการตอบข้อสอบเหมือนกันและค่าพารามิเตอร์ที่ศึกษาเท่ากันทั้งสองกลุ่ม) ผลการศึกษาพบว่า

1) ภายใต้เงื่อนไข no-DIF พบว่าความยาวของแบบสอบ การกระจายของความสามารถ ค่าอำนาจจำแนกและค่าการเดามีผลกระทบต่ออัตราความคลาดเคลื่อนชนิดที่ 1 และการประมาณค่า α_{MH} แบบสอบ 20 ข้อกับแบบสอบ 40 ข้อให้ค่า α_{MH} แตกต่างกันอย่างมีนัยสำคัญ โดยข้อสอบขนาด 20 ข้อมีผลให้อัตราความคลาดเคลื่อนชนิดที่ 1 สูงและการประมาณค่า α_{MH} ผิดพลาด(ค่าที่ได้เบี่ยงเบนจาก 1.00 ไปมาก) อันเป็นผลเนื่องมาจากค่าอำนาจจำแนกและค่าการเดา ส่วนข้อสอบขนาด 40 ข้อไม่พบความคลาดเคลื่อนชนิดที่ 1 แต่ยังคงมีการประมาณค่าผิดพลาด ส่วนค่า χ^2_{MH} ระหว่างแบบสอบทั้งสองขนาดไม่แตกต่างกันอย่างมีนัยสำคัญ จึงได้แนะนำว่าในการประเมิน DIF ต้องพิจารณาทั้ง α_{MH} และ χ^2_{MH}

2) ภายใต้เงื่อนไข DIF พบว่าค่าพารามิเตอร์ (a,b,c) การกระจายของความสามารถและปฏิสัมพันธ์ระหว่างค่าพารามิเตอร์และการกระจายของความสามารถมีผลกระทบต่อการประมาณค่า α_{MH} แต่ไม่มีผลต่ออัตราความคลาดเคลื่อนชนิดที่ 2 โดยพบว่าทำให้การประมาณค่า α_{MH} ผิดพลาดในแบบสอบทั้งสองขนาดและค่า α_{MH} ที่ได้จากแบบสอบทั้งสองขนาดแตกต่างกันอย่างมีนัยสำคัญ

Narayanan และ Swaminathan (1994) ได้ศึกษาผลของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมนเทล-แฮนส์เซลกับวิธีซิบเทสท์ ใช้ข้อมูลจำลอง ความยาวแบบสอบ 40 ข้อ ตัวแปรที่ศึกษาคือ (1) ขนาดกลุ่มตัวอย่าง โดยกลุ่มอ้างอิงมี 3 ขนาดได้แก่ 300 500 และ 1000 คน กลุ่มเปรียบเทียบ 3 ขนาดได้แก่ 100 200 และ 300 คนซึ่งจะจับคู่ศึกษาได้ 9 เงื่อนไข (2) การกระจายของความสามารถ 2 แบบ (3) อัตราส่วนของข้อสอบที่ทำหน้าที่ต่างกันที่มีภายในแบบสอบ 2 ขนาด (4) ขนาดของพื้นที่ระหว่างโค้งคุณลักษณะข้อสอบของผู้สอบสองกลุ่ม 4 ขนาด (5) ค่าความยากและค่าอำนาจจำแนกของแบบสอบ 6 ระดับ

ผลการศึกษาพบว่าขนาดกลุ่มตัวอย่าง อัตราส่วนของข้อสอบที่ทำหน้าที่ต่างกัน ขนาดของพื้นที่ระหว่างโค้งคุณลักษณะ ค่าความยากและค่าอำนาจจำแนกเป็นตัวแปรที่มีผลกระทบต่ออัตราการตรวจสอบของทั้งสองวิธีอย่างมีนัยสำคัญ วิธีแมนเทล-แฮนส์เซลและวิธีซิบเทสท์มีประสิทธิภาพเท่ากันในการตรวจสอบข้อสอบที่ทำหน้าต่างกันแบบเอกรูปเมื่อการกระจายความสามารถเท่ากันระหว่างกลุ่ม แต่เมื่อการกระจายความสามารถไม่เท่ากันระหว่างกลุ่มวิธีซิบเทสท์จะมีประสิทธิภาพในการตรวจสอบมากกว่าวิธีแมนเทล-แฮนส์เซล จึงสรุปได้ว่าการกระจายความสามารถไม่มีผลกระทบต่อการตรวจสอบด้วยวิธีซิบเทสท์ แต่มีผลกระทบต่อวิธีแมนเทล-แฮนส์เซลอย่างมีนัยสำคัญ ส่วนอัตราความคลาดเคลื่อนชนิดที่ 1 ของสถิติ MH เป็นไปตามที่คาดไว้ แต่อัตราความคลาดเคลื่อนชนิดที่ 1 ของสถิติ SIBTEST สูงกว่าที่คาดไว้เล็กน้อย ในกรณีที่มีการกระจายของความสามารถต่างกันเพิ่มขึ้นระหว่างผู้สอบสองกลุ่มจะทำให้อัตราความคลาดเคลื่อนชนิดที่ 1 เพิ่มขึ้น

Roussos และ Stout (1996) ได้ศึกษาผลกระทบของกลุ่มตัวอย่างขนาดเล็กและค่าพารามิเตอร์ของข้อสอบที่มีต่ออัตราความคลาดเคลื่อนชนิดที่ 1 ของวิธีซิบเทสท์กับวิธีแมนเทล-แฮนส์เซล ใช้ข้อมูลจำลอง โดยศึกษา 2 ครั้ง ครั้งแรกใช้ขนาดกลุ่มตัวอย่าง 100 200 500 และ 1000 คน และความแตกต่างของค่าเฉลี่ยของการกระจายความสามารถระหว่างกลุ่มเป็น 0.0 0.5 และ 1.0

ใช้แบบสอบจำนวน 25 ข้อ การศึกษาครั้งที่สอง ใช้ขนาดกลุ่มตัวอย่าง 500 1000 และ 3000 คน ความแตกต่าง ของค่าเฉลี่ยของการกระจายความสามารถระหว่างกลุ่มเป็น 0.0 และ 1.0 ค่าอำนาจจำแนก 3 ระดับ ค่าความยาก 5 ระดับ ค่าการเดา 3 ระดับ

ผลการศึกษาครั้งที่ 1 พบว่าอัตราความคลาดเคลื่อนชนิดที่ 1 เพิ่มขึ้นอย่างไม่มีนัยสำคัญทั้งสองวิธี ผลการศึกษาครั้งที่ 2 เมื่อความแตกต่างของค่าเฉลี่ยของการกระจายความสามารถระหว่างกลุ่มเป็น 1.0 จะทำให้อัตราความคลาดเคลื่อนชนิดที่ 1 เพิ่มขึ้นมาก (inflated) ทั้งสองวิธี โดยวิธีแมนเทล-แฮนส์เซลจะมีความคลาดเคลื่อนมากกว่า และเมื่อไม่มีความแตกต่างของค่าเฉลี่ยของการกระจายความสามารถทั้งวิธีซิบเทสท์และวิธีแมนเทล-แฮนส์เซลให้ผลที่น่าพอใจทุกเงื่อนไข

Kim และคณะ (1994) ได้ทดสอบอัตราความคลาดเคลื่อนชนิดที่ 1 จากการตรวจสอบ DIF ด้วยวิธีทดสอบไคสแควร์ของลอร์ด(Lord's chi-square test)เมื่อใช้การประมาณค่าแบบ marginal maximum likelilood estimation (MMLE) และการประมาณค่าแบบ marginal Bayesian estimation (MBE) อีกทั้งยังเปรียบเทียบการประมาณค่าเมื่อใช้โมเดลต่างกันคือ โมเดล 3 พารามิเตอร์ โมเดล 3 พารามิเตอร์โดยกำหนดค่าการเดา(fixed c) และโมเดล 2 พารามิเตอร์ ศึกษาโดยใช้ข้อมูลจำลองขนาดกลุ่มตัวอย่าง 250 และ 1000 คน ข้อสอบ 50 ข้อ ในการตรวจสอบจะใช้โปรแกรม BILOG ในการประมาณค่าพารามิเตอร์ของข้อสอบ ซึ่งสามารถประมาณค่าได้ 2 แบบคือ MMLE และ MBE แล้วเทียบมาตรของค่าพารามิเตอร์ของข้อสอบให้อยู่บนเมตริกซ์เดียวกันด้วยโปรแกรม EQUATE แล้วทดสอบความแตกต่างของค่าพารามิเตอร์ด้วย Lord's chi-square test

ผลการศึกษาพบว่า การประมาณค่าแบบ MMLE และ MBE ให้ผลใกล้เคียงกัน เมื่อประมาณค่าโดยใช้โมเดล 3 พารามิเตอร์จะทำให้อัตราความคลาดเคลื่อนชนิดที่ 1 สูงกว่าที่คาดไว้ ส่วนโมเดล 3 พารามิเตอร์ที่กำหนดค่าการเดาและโมเดล 2 พารามิเตอร์ทำให้อัตราความคลาดเคลื่อนชนิดที่ 1 ต่ำกว่าที่ระดับนัยสำคัญที่ตั้งไว้ และยังได้แนะนำว่าการใช้โมเดล 2 พารามิเตอร์และขนาดกลุ่มตัวอย่าง 1000 คนเป็นเงื่อนไขที่ให้ผลดีที่สุด โดยสามารถใช้ได้ทั้งการประมาณค่าแบบ MMLE หรือแบบ MBE

Budgell Raju และ Quartetti (1995) ได้วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในเครื่องมือการประเมินที่ถูกแปลเป็น 2 ภาษา ทั้งนี้เพื่อประเมินความเท่าเทียมกันในการวัดด้วยเครื่องมือที่ถูกแปลเป็นภาษาอื่นด้วยวิธีการตรวจสอบ DIF 4 วิธีคือ วิธีการวัดพื้นที่แบบมีเครื่องหมาย (signed area method ; SA) ของRaju วิธีการวัดพื้นที่แบบไม่มีเครื่องหมาย (unsigned area method ; UA) ของRaju วิธีทดสอบไคสแควร์ของลอร์ด (Lord's χ^2) และวิธีแมนเทล-แฮนส์เซล

เครื่องมือที่ใช้เป็นชุดของแบบวัดความสามารถทั่วไปทางสมองที่ใช้ในประเทศแคนาดา ซึ่งจัดทำเป็น 2 ชุดคือชุดที่เป็นภาษาอังกฤษและชุดที่เป็นภาษาฝรั่งเศสโดยผู้เชี่ยวชาญทางภาษา ที่ใช้ในการศึกษาครั้งนี้เป็นแบบสอบที่เกี่ยวกับตัวเลข 15 ข้อ และแบบสอบที่เกี่ยวกับความเป็นเหตุเป็นผล 18 ข้อ ทุกข้อเป็นแบบหลายตัวเลือก มีเงื่อนไขในการสอบคือ (1) ผู้สอบต้องสอบโดยใช้แบบสอบชุดที่ตรงกับภาษาที่ใช้มาแต่กำเนิด (2) ผู้สอบต้องอาศัยอยู่ในประเทศแคนาดาและสอบโดยใช้แบบสอบที่ตรงกับภาษาหลักที่ใช้ในชุมชนที่อาศัยอยู่ ได้ผู้สอบทั้งหมด 16,362 คน แล้วสุ่มมา 4 กลุ่มคือกลุ่ม E1 และ E2 ซึ่งเป็นผู้ที่สอบแบบสอบชุดที่เป็นภาษาอังกฤษกลุ่มละ 1,000 คน และกลุ่ม F1 และ F2 ซึ่งเป็นผู้ที่สอบแบบสอบชุดภาษาฝรั่งเศสกลุ่มละ 1,000 คน การจับคู่เปรียบเทียบเพื่อวิเคราะห์มี 4 เงื่อนไขดังนี้ (1) E1 กับ F1 (2) E2 กับ F2 (3) E1 กับ E2 (4) F1 กับ F2 โดยจะแยกวิเคราะห์ระหว่างแบบสอบเกี่ยวกับตัวเลขและแบบสอบเกี่ยวกับความเป็นเหตุเป็นผล ดัชนีที่บ่งชี้การทำหน้าที่ต่างกันของวิธี SA และ UA คือการทดสอบด้วยสถิติ Z วิธีของลอร์ดและวิธีแมนเทล-แฮนส์เซลใช้การทดสอบด้วย χ^2

ผลการศึกษาพบว่าการจับคู่เปรียบเทียบเงื่อนไขที่ 3 และ 4 ไม่พบข้อสอบที่ทำหน้าที่ต่างกัน ซึ่งเป็นไปตามที่คาดหวังไว้ ส่วนเงื่อนไขที่ 1 และ 2 วิธีตรวจสอบทั้ง 4 วิธีให้ผลสอดคล้องกันสูง ยกเว้นกรณีที่ใช้แบบสอบเกียวกับความเป็นเหตุเป็นผลตรวจสอบด้วยวิธีการวัดพื้นที่แบบไม่มีเครื่องหมาย มีค่าสอดคล้องต่ำกับวิธีอื่น ๆ ในการระบุข้อสอบที่ทำหน้าที่ต่างกัน

กาญจนา วัธนสุนทร (2537) ได้พัฒนาเกณฑ์ในการตัดสินข้อสอบลำเอียงทางเพศ โดยใช้ข้อมูลเชิงประจักษ์ ใช้วิธีการตรวจสอบ 3 วิธีคือ วิธีทฤษฎีการตอบสนองข้อสอบ วิธีแมนเทล-แฮนส์เซลและวิธีชิบเทสท์ ดัชนีที่พัฒนาเพื่อเป็นเกณฑ์ในการตัดสินข้อสอบลำเอียงคือ SA , UA , α_{MH} , β_{SIB} ตามลำดับ ซึ่งในการวิจัยครั้งนี้ได้จากการวิเคราะห์ค่าเฉลี่ยของค่าดัชนีแต่ละตัว

ปัจจัยที่แปรเปลี่ยนในการศึกษาได้แก่ ความยาวแบบสอบ 20 30 40 ข้อสำหรับวิชาคณิตศาสตร์ และ 50 60 70 และ 80 ข้อสำหรับวิชาภาษาอังกฤษ ขนาดผู้สอบ 100 200 400 600 800 และ 1000 คน

ผลการวิจัยพบว่าขนาดผู้สอบมีอิทธิพลต่อค่าเฉลี่ยของดัชนีทุกตัว ความยาวแบบสอบมีอิทธิพลต่อค่าเฉลี่ยของดัชนี SA และ UA แต่ไม่มีอิทธิพลต่อค่าเฉลี่ยของดัชนี α_{MH} และ β_{SIB} ซึ่งเกณฑ์ที่พัฒนาขึ้นเพื่อใช้ตัดสินความลำเอียงระหว่างผู้สอบเพศชายและเพศหญิงเป็นดังนี้

1) SA > .80 และ UA > .50 เมื่อความยาวแบบสอบน้อยกว่า 50 ข้อ

2) SA > .40 UA > 1.20 เมื่อความยาวแบบสอบ 50 ข้อขึ้นไป

3) α_{MH} < .60 และ α_{MH} > 1.40 สำหรับทุกขนาดของผู้สอบและความยาวแบบสอบ

4) β_{SIB} > .06 สำหรับทุกขนาดของผู้สอบและความยาวแบบสอบ

ทั้งนี้ในการใช้ดัชนี SA หรือ UA ควรใช้ผู้สอบขนาด 800 คนขึ้นไป ส่วนดัชนี α_{MH} และ β_{SIB} ควรใช้ขนาดผู้สอบอย่างน้อย 600 คน

เกษร หว่างจิตร (2539) ได้วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมนเทล-แฮนส์เซล โดยกลุ่มอ้างอิงและกลุ่มเปรียบเทียบจำแนกตามเพศ ภูมิลำเนา ประสบการณ์ในการสอบและสังกัดของสถานศึกษา ข้อมูลที่ใช้เป็นผลการตอบข้อสอบวิชาภาษาไทยของผู้สอบจำนวน 506 คนและผลการตอบข้อสอบวิชาภาษาอังกฤษของผู้สอบจำนวน 501 คน ในส่วนที่เป็นข้อสอบแบบหลายตัวเลือกของศูนย์ทดสอบทางการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ผลการศึกษาพบว่าข้อสอบที่ทำหน้าที่ต่างกันส่วนมากจะเป็นข้อสอบที่มีค่าอำนาจจำแนก(a) ค่อนข้างต่ำทั้งสองวิชา เมื่อพิจารณาในด้านค่าความยากพบว่าข้อสอบที่ทำหน้าที่ต่างกันส่วนมากเป็นข้อที่ง่ายมากสำหรับวิชาภาษาไทย ส่วนวิชาภาษาอังกฤษข้อสอบที่ทำหน้าที่ต่างกัน

ส่วนมากเป็นข้อที่ยากมาก อีกทั้งพบว่าส่วนมากเป็นข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูป เมื่อจำแนกกลุ่มอ้างอิงและกลุ่มเปรียบเทียบตามเพศ จะพบข้อสอบที่ทำหน้าที่ต่างกันมีจำนวนมากที่สุด รองลงมาคือการจำแนกตามภูมิลำเนา สังกัดของสถานศึกษาและประสบการณ์ในการสอบตามลำดับ

เพ็ญพนา สุขสม (2539) ได้ศึกษาเปรียบเทียบผลการวิเคราะห์ความลำเอียงของข้อสอบด้วยวิธีวิเคราะห์ 3 วิธี คือวิธีค่าความยากแปลง วิธี MH และวิธี IRT 3 พารามิเตอร์ โดยกลุ่มอ้างอิงและกลุ่มเปรียบเทียบจำแนกตามเพศและที่ตั้งของโรงเรียน เครื่องมือที่ใช้คือแบบทดสอบประเมินคุณภาพและวัดผลปลายปีวิชาภาษาไทย จำนวน 50 ข้อโดยข้อสอบเป็นแบบหลายตัวเลือก กลุ่มตัวอย่างคือนักเรียนชั้นประถมศึกษาปีที่ 6 จำนวน 2400 คน ผลการศึกษาพบว่าวิธี MH ตรวจพบข้อสอบลำเอียงมีจำนวนมากที่สุด รองลงมาคือวิธี IRT 3 พารามิเตอร์ และวิธีค่าความยากแปลงตรวจพบข้อสอบที่ลำเอียงมีจำนวนน้อยที่สุด ทั้งสามวิธีมีความสัมพันธ์กันทางบวก ค่าความสัมพันธ์มีค่าระหว่าง.4011 - .6662 เมื่อจำแนกตามเพศ และความสัมพันธ์มีค่าระหว่าง .5676 - .7847 เมื่อจำแนกตามที่ตั้งของโรงเรียน โดยวิธี IRT มีความสัมพันธ์กับวิธี MH สูงกว่าวิธีค่าความยากแปลง

จากรายงานการวิจัยข้างต้น จะพบว่าปัจจัยที่มีผลกระทบต่ออัตราการตรวจสอบด้วยวิธี MH และวิธี SIBTEST คือ ขนาดกลุ่มตัวอย่าง ซึ่งข้อค้นพบที่ได้จะสอดคล้องกัน กล่าวคือเมื่อขนาดกลุ่มตัวอย่างมากขึ้นอัตราการตรวจสอบพบ DIF จะเพิ่มขึ้น เกี่ยวกับขนาดของกลุ่มตัวอย่าง Hill (1990 cited in Mazor et al.,1992) กล่าวไว้ว่าขนาดกลุ่มตัวอย่างที่พอเหมาะสำหรับวิธี MH ควรใช้ระหว่าง 100 ถึง 300 คนสำหรับกลุ่มใดกลุ่มหนึ่งหรือทั้งสองกลุ่ม ส่วน Mazor และคณะ(1992) แนะว่าการใช้กลุ่มตัวอย่างขนาด 200 คนก็เพียงพอแล้วและไม่ควรน้อยกว่านี้ Narayanan และ Swaminathan (1994)แนะนำว่าโดยทั่วไปใช้ขนาดกลุ่มตัวอย่าง กลุ่มละ 300 คนก็เพียงพอที่จะตรวจสอบอย่างมีประสิทธิภาพและพบว่าอัตราการตรวจสอบของวิธี MH และวิธี SIBTEST จะได้รับผลกระทบจากกลุ่มตัวอย่างขนาดเล็กของกลุ่มเปรียบเทียบมากกว่ากลุ่มตัวอย่างขนาดใหญ่ของกลุ่มอ้างอิง จึงกล่าวว่าอัตราส่วนระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบน่าจะมีผลกระทบ

ต่ออัตราการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ดังนั้นจึงทำให้ผู้วิจัยต้องการศึกษาเกี่ยวกับขนาดกลุ่มตัวอย่างที่เป็นอัตราส่วนต่อกันระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ทั้งนี้เพราะในทางปฏิบัติจะพบว่าขนาดกลุ่มตัวอย่างทั้งสองกลุ่มมักไม่เท่ากัน อีกทั้งยังไม่มีข้อความรู้ที่ชัดเจนว่าควรใช้กลุ่มตัวอย่างที่เป็นอัตราส่วนต่อกันเท่าใด

นอกจากนี้ Roger Swaminathan (1993) พบว่าความยาวแบบสอบไม่มีผลกระทบต่ออัตราการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ในขณะที่ Kim และ Cohen (1993) Uttaro และ Miller (1994) พบว่าความยาวของแบบสอบมีผลกระทบต่ออัตราการตรวจสอบ อีกทั้งกาญจนา วัธนสุนทร (2537) พบว่าการตรวจสอบ DIF มีความไม่คงที่ข้ามขนาดผู้สอบและความยาวแบบสอบ ดังนั้นเพื่อเป็นการยืนยันข้อค้นพบให้มีความชัดเจนมากยิ่งขึ้นจึงต้องศึกษาเกี่ยวกับความยาวของแบบสอบว่าจะมีผลต่ออัตราการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบอย่างไรบ้าง