บทที่ 2

วรรณคดีที่เกี่ยวข้อง

การวิจัยครั้งนี้ มีวัตถุประสงค์เพื่อเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันแบบไม่สม่ำเสมอของข้อสอบระหว่างวิธีแมนเทล-แฮนส์เซล แบบปกติ กับ วิธีแมนเทล-แฮนส์เซล แบบแบ่งกลุ่มความสามารถของผู้สอบและความยากของข้อสอบ ดังนั้น วรรณคดีที่เกี่ยวข้องจึงไม่กล่าวถึงวิธีการอื่น ๆ แต่จะกล่าวถึงวิธีแมนเทล-แฮนส์เซล เท่านั้น โดยแบ่งวรรณคดีที่เกี่ยวข้องเป็น 4 ตอน ดังนี้

ตอนที่ 1 แนวคิดและวิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธีแมนเทล-แฮนส์เซล

ตอนที่ 2 แนวคิดเกี่ยวกับหลักการพิจารณาข้อสอบสำเอียง

ตอนที่ 3 งานวิจัยที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

ตอนที่ 4 แนวทางการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ตามข้อเสนอแนะของ Hambleton, et al. (1993)

ตอนที่ 1 แนวคิดและวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธีแมนเทล-แฮนส์เซล

การทำหน้าที่ต่างกันของข้อสอบ (DIFferential Item Functioning) เป็นคุณลักษณะทางสถิติของข้อสอบ ที่แสดงให้เห็นความแตกต่างในการทำหน้าที่ของข้อสอบต่อกลุ่มผู้สอบ 2 กลุ่ม ที่นำมาเปรียบเทียบกัน ซึ่งกลุ่มผู้สอบส่วนใหญ่จำแนกตามสารสนเทศของประชากร เช่น เชื้อชาติ เผ่าพันธุ์ เพศ ศาสนา ภาษา เป็นต้น ดังนั้น หลักการเรื่อง การจับคู่กลุ่มผู้สอบตามภาวะสันนิษฐานที่แบบสอบมุ่งประสงค์จะวัด แล้วเปรียบเทียบกลุ่มผู้สอบ จึงเป็นหลักการสำคัญ เพราะเป็นการแยกแยะเรื่อง ความแตกต่างในการทำหน้าที่ของข้อสอบ ออกจากเรื่อง ความแตกต่างของผลการตอบข้อสอบระหว่างกลุ่มผู้สอบ (Dorans and Holland, 1993)

ความแตกต่างของผลการตอบข้อสอบระหว่างกลุ่มผู้สอบ 2 กลุ่ม หรือที่เรียกว่า impact เป็นเรื่องปกติทั่ว ๆ ไป เนื่องจากบุคคลแต่ละคนมีคุณลักษณะแตกต่างกันไป ตามหลักความแตกต่างระหว่างบุคคล เช่น กลุ่มผู้สอบชาวอเมริกันเอเชีย ได้คะแนนเฉลี่ยในวิชาคณิตศาสตร์

สูงกว่ากลุ่มผู้สอบชาวอเมริกันผิวขาว หรือ กลุ่มผู้สอบเพศชาย ได้คะแนนเฉลี่ยในวิชาคณิตศาสตร์สูงกว่ากลุ่มผู้สอบเพศหญิง เป็นต้น แต่เรื่องข้อสอบทำหน้าที่ต่างกัน เป็นความแตกต่างในการทำหน้าที่ของข้อสอบ หลังจาก กลุ่มผู้สอบทั้งสองถูกจับคู่กันตามความสามารถหรือคุณลักษณะที่ข้อสอบข้อนั้นมุ่งวัด ดังนั้นทิศทางของ impact จึงอาจไม่เป็นเช่นเดียวกับ ทิศทางของข้อสอบทำหน้าที่ต่างกัน ดังเช่น โดยภาพรวมแล้วข้อสอบวิชาคณิตศาสตร์ อาจยากกับกลุ่มผู้สอบเพศหญิง แต่เมื่อพิจารณาโดยการจับคู่กลุ่มผู้สอบตามความสามารถแล้ว ข้อสอบข้อนั้นอาจง่ายกับกลุ่มผู้สอบเพศหญิง ก็ได้ หลักการนี้จะเข้าใจได้ดียิ่งขึ้น ถ้าพิจารณาจาก Simpson's paradox (simpson, 1951 quoted in Dorans and Holland, 1993) ดังแสดงในตารางที่ 1

ตารางที่ 1 สรุปผลการตอบข้อสอบข้อหนึ่งของกลุ่มผู้สอบ 2 กลุ่ม

กลุ่ม ก			กลุ่ม ข		
N_m	N_{om}	N_{om} / N_m	N_m	N_{om}	N_{om} / N_m
400	40	.10	1000	200	.20
1000	500	.50	1000	600	.60
1000	900	.90	400	400	1.00
2400	1440	.60	2400	1200	.50

N_m แทน จำนวนผู้สอบ ในระดับความสามารถ m

N_{om} แทน จำนวนผู้ตอบข้อสอบถูก ในระดับความสามารถ m

N_{om} / N_m แทน สัดส่วนผู้ตอบข้อสอบถูก ในระดับความสามารถ m

จำนวนผู้สอบในแต่ละแถว จัดเรียงตามระดับความสามารถ (m) จากต่ำไปหาสูง

จากตารางที่ 1 ถ้าพิจารณาสัดส่วนโดยรวมของผู้ตอบข้อสอบถูก จะเห็นว่า กลุ่ม ก มีสัดส่วนผู้ตอบข้อสอบถูกเท่ากับ .60 หรือ 60 % แต่ในกลุ่ม ข มีสัดส่วนผู้ตอบข้อสอบถูก เท่ากับ .50 หรือ 50 % ดังนั้น กลุ่ม ก ได้เปรียบกลุ่ม ข เท่ากับ .10 (.60 - .50)

แสดงว่า ข้อสอบเข้าข้างกลุ่ม ก แต่ถ้าพิจารณาเปรียบเทียบสัดส่วนผู้ตอบข้อสอบถูกของกลุ่ม ก กับกลุ่ม ข ในแต่ละแถว จะเห็นว่า กลุ่ม ก มีสัดส่วนผู้ตอบข้อสอบถูกต่ำกว่ากลุ่ม ข อยู่ .10 ในทุกระดับความสามารถ ดังนั้น เมื่อเปรียบเทียบกลุ่มทั้งสองตามระดับความสามารถแล้ว จะพบว่า กลุ่ม ข ได้เปรียบกลุ่ม ก ข้อแตกต่างระหว่างการเปรียบเทียบสัดส่วนโดยรวม กับ การเปรียบเทียบสัดส่วนตามระดับความสามารถ เกิดจากการแจกแจงความสามารถของผู้สอบในกลุ่ม ก กับ กลุ่ม ข ไม่เท่ากัน ดังจะเห็นได้จากจำนวนคนในแต่ละแถวของสองกลุ่ม ไม่เท่ากัน จริง ๆ แล้ว ถ้าพิจารณากลุ่มผู้สอบตามระดับความสามารถ กลุ่ม ก เสียเปรียบกลุ่ม ข แต่เนื่องจากคนมีความสามารถสูงในกลุ่ม ก มีจำนวนมากกว่าคนมีความสามารถสูงในกลุ่ม ข จึงทำให้สัดส่วนผู้ตอบข้อสอบถูกโดยรวมแสดงออกมาว่า ข้อสอบข้อนี้เข้าข้างกลุ่ม ก

Shepard (1982) ได้อธิบายความหมายของคำว่า "ความลำเอียงของข้อสอบ" ว่า ถ้าข้อสอบไม่ลำเอียง ผู้สอบแต่ละคนที่ได้คะแนนเท่ากันจากแบบสอบที่เป็นเอกพันธ์ มีสัดส่วนการตอบข้อสอบถูกเช่นเดียวกับกลุ่มประชากรที่ศึกษา จากนิยามนี้ได้นำไปสู่วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยการใช้ตารางการณ์จร ภายใต้เงื่อนไขคะแนนที่สังเกตได้ในช่วงแรก ๆ ของ Scheuneman (1975, quoted in Dorans and Holland, 1993)

Mantel และ Haenszel (1959) ได้แนะนำ วิธีการจับคู่กลุ่มแบบใหม่ โดยการแบ่งชั้นคะแนน (categories) ให้แต่ละชั้นคะแนนเท่ากับหนึ่งหน่วย (unit interval) แทนการแบ่งชั้นคะแนนตามวิธีไค-สแควร์ แบบเดิม ทั้งนี้ เพื่อหลีกเลี่ยงปัญหาในเรื่องการรวมสเกลของแต่ละชั้นคะแนน และเขาได้นำวิธีการนี้ไปใช้กับการวิจัยเรื่อง การศึกษาย้อนหลังเกี่ยวกับสาเหตุของโรค

ต่อมา Holland และ Thayer (1986 ; 1988) ได้นำวิธีการจับคู่กลุ่มของ Mantel และ Haenszel มาใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยนำมาใช้ครั้งแรกกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ของหน่วยงานบริการทดสอบทางการศึกษาแห่งสหรัฐอเมริกา (Dorans and Holland, 1993)

วิธีแมนเทล-แฮนส์เซล (Mantel-Haenszel procedure : MH) เป็นวิธีตารางการณ์จรแบบไม่คำนวณซ้ำทวน (noniterative contingency table method) ที่ใช้ประมาณและทดสอบค่าพารามิเตอร์ขององค์ประกอบร่วม 2 องค์ประกอบที่สัมพันธ์กัน ในตารางการณ์จรแบบ 2 X 2 X M (M แทน จำนวนชั้นคะแนน) ดังนั้น จึงสามารถนำมาใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับกลุ่มผู้สอบสองกลุ่มได้ (Holland and Thayer, 1988)

การใช้วิธีแมนเทล-แฮนส์เซล ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับกลุ่มผู้สอบสองกลุ่มนั้น Holland และ Thayer (1988) ได้เรียกชื่อกลุ่มผู้สอบที่ใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ดังนี้

1. กลุ่มสนใจ (focal group : f) เป็นกลุ่มผู้สอบที่เชื่อว่าจะเสียเปรียบ ในกรณีที่ข้อสอบทำหน้าที่ต่างกัน

2. กลุ่มอ้างอิง (reference group : r) เป็นกลุ่มผู้สอบที่ใช้เป็นมาตรฐาน ในการเปรียบเทียบกับกลุ่มสนใจ เพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ตัวอย่าง เช่น การศึกษาการทำหน้าที่ต่างกันของข้อสอบ ระหว่างผู้สอบต่างเชื้อชาติ กลุ่มสนใจ ได้แก่ ผู้สอบผิวดำ ในขณะที่กลุ่มอ้างอิง ได้แก่ ผู้สอบผิวขาว เป็นต้น

ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ระหว่างกลุ่มสนใจกับกลุ่มอ้างอิงครั้งหนึ่ง ๆ จะเรียกข้อสอบแต่ละข้อที่ถูกตรวจสอบการทำหน้าที่ต่างกันว่า "ข้อสอบที่ศึกษา (studied item)"

หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามวิธีแมนเทล-แฮนส์เซล คือ การเปรียบเทียบผลการตอบแบบสอบของผู้สอบกลุ่มอ้างอิงกับกลุ่มสนใจ แล้วจำแนกข้อสอบที่แสดงว่า ทำหน้าที่ต่างกันต่อผู้สอบกลุ่มอ้างอิงหรือกลุ่มสนใจออกมา โดยมีการตรวจสอบ ในทุก ๆ ระดับความสามารถของกลุ่มผู้สอบทั้งสองกลุ่ม ข้อสอบข้อใดที่กลุ่มผู้สอบทั้งสองกลุ่มทำคะแนนได้เท่า ๆ กัน แสดงว่า ข้อสอบข้อนั้นทำหน้าที่ต่อกลุ่มผู้สอบทั้งสองกลุ่มไม่ต่างกัน ดังนั้น เกณฑ์ที่ใช้ในจับคู่กลุ่มผู้สอบเพื่อแทนระดับความสามารถของผู้สอบทั้งสองกลุ่ม จึงมีความสำคัญ ในทางปฏิบัติ ส่วนใหญ่ใช้คะแนนของแบบสอบทั้งฉบับซึ่งมีข้อสอบข้อนั้นรวมอยู่ด้วย เป็นเกณฑ์การจับคู่กลุ่มผู้สอบ เนื่องจากเห็นว่าแบบสอบฉบับนั้นวัดความสามารถเดียวกับข้อสอบที่ต้องการตรวจสอบการทำหน้าที่ต่างกัน (Holland and Thayer, 1988)

หลังจากเลือกเกณฑ์การจับคู่กลุ่มผู้สอบสองกลุ่มแล้ว ข้อมูลผลการตอบข้อสอบที่ศึกษาจะถูกจัดลงในตารางการณ์จรแบบ 2 X 2 X M หรือ ตาราง 3 มิติแบบ 2 X 2 X M ขนาดใหญ่ 1 ตาราง ตารางการณ์จร 1 ตาราง แทนผลการตอบข้อสอบที่ศึกษาตามเกณฑ์การจับคู่กลุ่มผู้สอบ 1 ชั้นคะแนน ดังนั้น ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ศึกษาข้อหนึ่ง จึงมีจำนวนตารางการณ์จรเท่ากับ M ตาราง (M แทน จำนวนชั้นคะแนน) ดังแสดงในตารางที่ 2

ตารางที่ 2 ผลการตอบข้อสอบที่ศึกษาในตารางการณ์จร 2 X 2 X M
(กลุ่มผู้สอบ X คะแนนผลการตอบ X จำนวนชั้นคะแนน)

กลุ่มผู้สอบ	คะแนนผลการตอบข้อสอบที่ศึกษา		
	ถูก (R)	ผิด (W)	รวม
กลุ่มสนใจ (f)	R_{fm}	W_{fm}	N_{fm}
กลุ่มอ้างอิง (r)	R_{rm}	W_{rm}	N_{rm}
รวม (t)	R_{tm}	W_{tm}	N_{tm}

เมื่อ R_{fm} และ W_{fm} แทน จำนวนผู้สอบกลุ่มสนใจที่ตอบข้อสอบถูก (1) และตอบข้อสอบผิด (0) ตามลำดับ ณ ระดับชั้นคะแนน m

R_{rm} และ W_{rm} แทน จำนวนผู้สอบกลุ่มอ้างอิงที่ตอบข้อสอบถูก (1) และตอบข้อสอบผิด (0) ตามลำดับ ณ ระดับชั้นคะแนน m

N_{fm} และ N_{rm} แทน จำนวนผู้สอบกลุ่มสนใจ และกลุ่มอ้างอิง ตามลำดับ ณ ระดับชั้นคะแนน m

R_{tm} และ W_{tm} แทน จำนวนผู้สอบที่ตอบข้อสอบถูก (1) และตอบข้อสอบผิด (0) ตามลำดับ ณ ระดับชั้นคะแนน m

N_{tm} แทน จำนวนผู้สอบทั้งหมดของสองกลุ่ม ณ ระดับชั้นคะแนน m ซึ่ง มีจำนวนเท่ากับ $N_{fm} + N_{rm}$

ตารางการณ์จร 2 X 2 X M

การจัดทำตารางการณ์จร 2 X 2 X M ทำได้ดังนี้

1. คะแนนผลการตอบของข้อสอบแต่ละข้อจะถูกแจกแจงเป็น คำตอบถูก (1) หรือ คำตอบผิด (0)

2. นับจำนวนคำตอบถูก (1) และ คำตอบผิด (0) ของแต่ละข้อ แล้วแจกแจงลงในตารางการณ์จรแบบ 2 X 2 X M

3. ชั้นคะแนนมีจำนวน M ชั้น ตามเกณฑ์การจับคู่กลุ่มผู้สอบ เช่น คะแนนของแบบสอบทั้งฉบับ เป็นต้น

ค่าความถี่ในตารางที่ 2 สามารถเปลี่ยนให้เป็นค่าสัดส่วนได้ ดังตารางที่ 3

ตารางที่ 3 สัดส่วนผลการตอบข้อสอบที่ศึกษาในตารางการณ์จร 2 X 2 X M
(กลุ่มผู้สอบ X คะแนนผลการตอบ X จำนวนชั้นคะแนน)

	คะแนนผลการตอบข้อสอบที่ศึกษา		
กลุ่มผู้สอบ	ถูก (R)	ผิด (W)	รวม
กลุ่มสนใจ (f)	p_{fm}	q_{fm}	1
กลุ่มอ้างอิง (r)	p_{rm}	q_{rm}	1

เมื่อ p_{fm} แทน สัดส่วนของผู้สอบกลุ่มสนใจที่ตอบข้อสอบถูก (1) ณ ระดับชั้นคะแนน m

q_{fm} แทน สัดส่วนของผู้สอบกลุ่มสนใจที่ตอบข้อสอบผิด (0) ณ ระดับชั้นคะแนน m ($q_{fm} = 1 - p_{fm}$)

p_{rm} แทน สัดส่วนของผู้สอบกลุ่มอ้างอิงที่ตอบข้อสอบถูก (1) ณ ระดับชั้นคะแนน m

q_{rm} แทน สัดส่วนของผู้สอบกลุ่มอ้างอิงที่ตอบข้อสอบผิด (0) ณ ระดับชั้นคะแนน m ($q_{rm} = 1 - p_{rm}$)

การทดสอบสมมุติฐาน

สมมุติฐานสูญการทำหน้าที่ต่างกันของข้อสอบ (null DIF hypothesis) ของวิธี แมนเทล-แฮนส์เซล แสดงได้ดังนี้

$$H_0 : [R_{rm}/W_{rm}] / [R_{fm}/W_{fm}] = 1 , \quad m = 1,...,M, \qquad (1)$$

หรือ

$$H_0 : [R_{rm}/W_{rm}] = [R_{fm}/W_{fm}] , \quad m = 1,...,M. \qquad (2)$$

นั่นก็คือ แต้มต่อของการตอบข้อสอบถูก ณ ระดับชั้นคะแนน m (ตามเกณฑ์การจับคู่กลุ่มผู้สอบ) เหมือนกันทั้งกลุ่มสนใจและกลุ่มอ้างอิง ในทุกระดับชั้นคะแนน M (ตามเกณฑ์การจับคู่กลุ่มผู้สอบ)

การทดสอบสมมุติฐานอัตราส่วนแต้มต่อคงที่

Mantel และ Haenszel (1959) ได้พัฒนาการทดสอบไค-สแควร์ของสมมุติฐานสูญการทำหน้าที่ต่างกันของข้อสอบ ต่อสมมุติฐานแย้ง เป็นการเฉพาะ ที่เรียกว่า "สมมุติฐานอัตราส่วนแต้มต่อคงที่ (constant odds ratio hypothesis)" ดังนี้

$$H_a : [R_{rm}/W_{rm}] = \alpha [R_{fm}/W_{fm}] , \quad m = 1,...,M \text{ และ } \alpha \neq 1 \qquad (3)$$

จะสังเกตว่าเมื่อ $\alpha = 1.0$ สมมุติฐานแย้งจะเท่ากับสมมุติฐานสูญการทำหน้าที่ต่างกันของข้อสอบ ค่าพารามิเตอร์ α นี้เรียกว่า อัตราส่วนแต้มต่อร่วม (common odds ratio)

ในตารางการณ์จร 2 X 2 X M และภายใต้สมมุติฐาน H_a ค่าของ α คือ อัตราส่วนแต้มต่อ ซึ่งมีค่าเหมือนกัน ในทุกระดับชั้นคะแนน m

$$\alpha_m = [R_{rm}/W_{rm}] / [R_{fm}/W_{fm}] = [R_{rm}W_{fm}] / [R_{fm}W_{rm}] \quad (4)$$

สถิติทดสอบไค-สแควร์

การทดสอบไค-สแควร์ ตามวิธีแมนเทล-แฮนส์เซล เรียกว่า การทดสอบสมมุติฐานสูญการทำหน้าที่ต่างกันของข้อสอบ คือ $H_0 : \alpha_m = 1$,

$$MH - \chi^2 = [\left|\sum_m R_{rm} - \sum_m E(R_{rm})\right| - .5]^2 \Big/ \sum_m Var(R_{rm}), \quad (5)$$

$$E(R_{rm}) = E(R_{rm} \mid \alpha = 1) = N_{rm}R_{tm}/N_{tm},$$

$$Var(R_{rm}) = Var(R_{rm} \mid \alpha = 1)$$

$$= [N_{rm}R_{tm}N_{fm}W_{tm}] / [N_{tm}^2(N_{tm} - 1)]. \quad (6)$$

เมื่อ R_{rm} แทน จำนวนผู้สอบกลุ่มอ้างอิงที่ตอบข้อสอบถูก (1) ณ ระดับชั้นคะแนน m

$E(R_{rm})$ แทน ความถี่ที่คาดว่าควรจะเป็น (expected frequency) ของจำนวนผู้สอบกลุ่มอ้างอิงที่ตอบข้อสอบถูก (1) ณ ระดับชั้นคะแนน m

$Var(R_{rm})$ แทน ความแปรปรวนของจำนวนผู้สอบกลุ่มอ้างอิงที่ตอบข้อสอบถูก (1) ณ ระดับชั้นคะแนน m

N_{rm} แทน จำนวนผู้สอบกลุ่มอ้างอิง ณ ระดับชั้นคะแนน m

R_{tm} แทน จำนวนผู้สอบที่ตอบข้อสอบถูก (1) ณ ระดับชั้นคะแนน m

N_{tm} แทน จำนวนผู้สอบทั้งหมดของสองกลุ่ม ณ ระดับชั้นคะแนน m

N_{fm} แทน จำนวนผู้สอบกลุ่มสนใจ ณ ระดับชั้นคะแนน m

W_{tm} แทน จำนวนผู้สอบที่ตอบข้อสอบผิด (0) ณ ระดับชั้นคะแนน m

เนื่องจากการแจกแจงของค่า MH-χ^2 ใกล้เคียงกับการแจกแจงค่าไค-สแควร์ที่มี degrees of freedom เท่ากับ 1 ทำให้ค่า MH-χ^2 ที่คำนวณได้สูงกว่าที่ควรจะเป็นจริง ดังนั้น จึงต้องปรับค่า MH-χ^2 ที่คำนวณได้ให้ถูกต้อง โดยนำค่า 0.5 ซึ่งเรียกว่า ค่าแก้ (continuity correction) ไปหักออกจากผลต่างที่ไม่คิดเครื่องหมายทั้งหมด ระหว่างความถี่ที่ได้จากการสังเกตกับความถี่ที่คาดว่าควรจะเป็น

Holland และ Thayer (1988) ได้รายงานว่า สถิติทดสอบ MH-χ^2 เป็นการทดสอบ H_0 กับ H_a ที่ไม่ลำเอียงแบบสม่ำเสมอที่มีอำนาจมากที่สุด ดังนั้น จึงไม่มีสถิติตัวใดที่มีอำนาจสูงกว่า ในกรณีที่ H_a มากกว่า 1 ตาม MH-χ^2 หรือ กล่าวอีกนัยหนึ่งได้ว่า วิธีแมนเทล-แฮนส์เซลเป็นการทดสอบทางสถิติที่มีอำนาจมากที่สุด ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจากสมมุติฐานสูญการทำหน้าที่ต่างกันของข้อสอบ ซึ่งคงเส้นคงวากับสมมุติฐานอัตราส่วนแต้มต่อคงที่

ตัวประมาณค่าอัตราส่วนแต้มต่อคงที่

Mantel และ Haenszel (1959) ได้เสนอตัวประมาณค่าอัตราส่วนแต้มต่อคงที่ (constant odds ratio) ดังนี้

$$\alpha_{MH} = [\sum_m R_{rm} W_{fm} / N_{tm}] / [\sum_m R_{fm} W_{rm} / N_{tm}] \quad (7)$$

α_{MH} เป็นตัวประมาณค่าขนาดอิทธิพลของข้อสอบทำหน้าที่ต่างกัน (DIF effect size) ในมาตราที่มีช่วงจาก 0 ถึง ∞

ดังนั้น ค่า $\alpha_{MH} = 1.0$ แสดงสมมุติฐานสูญการทำหน้าที่ต่างกันของข้อสอบ หรือข้อสอบทำหน้าที่ไม่ต่างกัน

ค่า $\alpha_{MH} < 1.0$ แสดงว่า ข้อสอบง่ายสำหรับผู้สอบกลุ่มสนใจ

ค่า $\alpha_{MH} > 1.0$ แสดงว่า ข้อสอบง่ายสำหรับผู้สอบกลุ่มอ้างอิง

เนื่องจากเมตริกอัตราส่วนแต้มต่อ เป็นสเกลที่คนส่วนใหญ่เข้าใจได้ยาก โดยทั่ว ๆ ไป นักพัฒนาแบบสอบจะแปลงแต้มต่อ ให้เป็น log odds เพราะการแจกแจงมีลักษณะสมมาตรรอบค่าเฉลี่ย ทำให้ง่ายต่อการตีความหมาย

MH DIF ในเมตริกค่าความยากของข้อสอบ

นักพัฒนาแบบสอบมักจะคุ้นเคยกับตัวประมาณค่าความยากของข้อสอบ ในเมตริกค่าความยากมาตรฐานของข้อสอบ (delta metric) ซึ่งมีค่าเฉลี่ยเท่ากับ 13 และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 4 การหาค่าความยากมาตรฐานของข้อสอบ (delta : Δ) ทำได้โดยการแปลงค่าสัดส่วนการตอบข้อสอบถูก (p) ให้เป็นคะแนน Z ด้วยการใช้ส่วนกลับของฟังค์ชั่นสะสมปกติ ตามวิธีการแปลงค่าเชิงเส้นตรง ให้เป็นเมตริกที่มีค่าเฉลี่ยเท่ากับ 13 และส่วนเบี่ยงเบนมาตรฐาน เท่ากับ 4 ดังนี้

$$\Delta = 13 + 4Z \qquad (8)$$

ดังนั้น ค่า Δ มาก แสดงว่า ข้อสอบยาก

ค่า Δ น้อย แสดงว่า ข้อสอบง่าย

Holland และ Thayer (1985) ได้แปลงค่า α_{MH} ไปเป็นความแตกต่างในค่าความยากมาตรฐานของข้อสอบ ดังนี้

$$\text{MH D-DIF} = -2.35 \ln [\alpha_{MH}]. \qquad (9)$$

ค่า MH D-DIF เป็นดัชนีที่ใช้วัดปริมาณการทำหน้าที่ต่างกันของข้อสอบ ดังนี้

1. ค่า MH D-DIF เป็นศูนย์ แสดงว่า ข้อสอบทำหน้าที่ไม่ต่างกัน ซึ่งหมายความว่า คำถามยากกับผู้สอบกลุ่มอ้างอิงเท่ากับผู้สอบกลุ่มสนใจ

2. ค่า MH D-DIF เป็นบวก แสดงว่า ข้อสอบทำหน้าที่เข้าข้างกลุ่มสนใจ ซึ่งหมายความว่า คำถามยากกับผู้สอบกลุ่มอ้างอิงมากกว่าผู้สอบกลุ่มสนใจ

3. ค่า MH D-DIF เป็นลบ แสดงว่า ข้อสอบทำหน้าที่เข้าข้างกลุ่มอ้างอิง ซึ่งหมายความว่า คำถามยากกับผู้สอบกลุ่มสนใจมากกว่าผู้สอบกลุ่มอ้างอิง

กฎการจำแนกข้อสอบทำหน้าที่ต่างกันของหน่วยงานบริการทดสอบทางการศึกษาแห่งสหรัฐอเมริกา

หน่วยงานบริการทดสอบทางการศึกษา แห่งสหรัฐอเมริกา ได้จำแนกข้อสอบที่ศึกษาออกเป็น 3 กลุ่ม โดยพิจารณาจาก 2 องค์ประกอบ คือ 1) ค่าสัมบูรณ์ของ MH D-DIF 2) ค่าสถิติ MH -χ^2 มีนัยสำคัญหรือไม่ ที่ต้องพิจารณาค่าสัมบูรณ์ประกอบค่าสถิติมีนัยสำคัญก็เนื่องจากว่าในกรณีที่ใช้กลุ่มผู้สอบขนาดใหญ่วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ การทำหน้าที่ต่างกันของข้อสอบเพียงเล็กน้อย ก็อาจแสดงนัยสำคัญได้ (Zieky, 1993)

การจำแนกกลุ่มข้อสอบ ออกเป็น 3 กลุ่ม ได้แก่

กลุ่ม A ค่า MH- χ^2 ไม่มีนัยสำคัญ หรือ MH- χ^2 มีนัยสำคัญ แต่ค่าสัมบูรณ์ของ MH D-DIF น้อยกว่า 1.0

กลุ่ม B ค่า MH- χ^2 มีนัยสำคัญ และค่าสัมบูรณ์ของ MH D-DIF ตั้งแต่ 1.0 ถึงน้อยกว่า 1.5

กลุ่ม C ค่า MH- χ^2 มีนัยสำคัญ และค่าสัมบูรณ์ของ MH D-DIF ตั้งแต่ 1.5 ขึ้นไป

ดังนั้น กลุ่ม A ประกอบด้วย ข้อสอบทำหน้าที่ต่างกันเล็กน้อยหรือไม่มีนัยสำคัญ นักพัฒนาแบบสอบจึงสามารถนำข้อสอบในกลุ่มนี้ไปใช้ได้อย่างมั่นใจ

กลุ่ม B ประกอบด้วย ข้อสอบทำหน้าที่ต่างกันที่มีนัยสำคัญ และเป็นข้อสอบทำหน้าที่ต่างกันระดับเล็กน้อยถึงระดับปานกลาง ซึ่งเป็นข้อสอบที่อาจนำไปใช้ได้

กลุ่ม C ประกอบด้วย ข้อสอบทำหน้าที่ต่างกันที่มีนัยสำคัญ และเป็นข้อสอบหน้าที่ต่างกันระดับปานกลางถึงระดับมาก การพิจารณานำข้อสอบในกลุ่มนี้ไปใช้ ในกรณีที่ต้องการให้ครบถ้วนตามคุณสมบัติเฉพาะของแบบสอบเท่านั้น

ตอนที่ 2 แนวคิดเกี่ยวกับหลักการพิจารณาข้อสอบลำเอียง

หลักการพิจารณาว่า ข้อสอบข้อใดข้อหนึ่งลำเอียงต่อผู้สอบกลุ่มใดของประชากร ต้องพิจารณาข้อสอบข้อนั้นตามจุดมุ่งหมายในการวัดของแบบสอบ เนื่องจากความไม่ยุติธรรม หรือความลำเอียงของข้อสอบขึ้นอยู่กับจุดมุ่งหมายในการวัดของแบบสอบ ข้อสอบที่ถามเรื่องเกี่ยวกับอุณหภูมิที่ใช้ในการอบขนมเค้ก ถ้าพิจารณากันอย่างผิน ๆ แล้ว อาจจะเห็นว่าเป็นข้อสอบที่ลำเอียงเข้าข้างผู้สอบเพศหญิง ทำให้ผู้สอบเพศชายเสียเปรียบ แต่ถ้าแบบสอบฉบับนี้มีจุดมุ่งหมาย

ในการคัดเลือกช่างทำอาหาร หรือช่างทำขนม ข้อสอบข้อนี้ก็อาจเป็นข้อสอบที่เหมาะสมได้ (Ironson, 1982) หรือ ในกรณีข้อสอบวัดความเข้าใจในการอ่านซึ่งมีเนื้อหาเกี่ยวกับกีฬาเบสบอลล์ อาจจะถูกพิจารณาว่า เป็นข้อสอบที่ลำเอียงเข้าข้างผู้สอบเพศชาย ทำให้ผู้สอบเพศหญิงเสียเปรียบ เนื่องจากว่า ผู้สอบเพศหญิงขาดความคุ้นเคยกับการใช้ถ้อยคำในเรื่องกีฬาเบสบอลล์ แต่ถ้าแบบสอบฉบับนี้มีจุดมุ่งหมายจะวัดความรู้ในเรื่องกีฬาของผู้สอบ ก็อาจเป็นข้อสอบที่เหมาะสมได้ เพียงแต่เป็นข้อสอบที่ไม่เหมาะสมสำหรับการวัดความเข้าใจในการอ่านเท่านั้น (Camilli and Shepard, 1994) ดังนั้น ในบางกรณีความไม่ยุติธรรมที่ปรากฏอยู่ในตัวข้อสอบหรือการที่ข้อสอบลำเอียงต่อผู้สอบกลุ่มใดของประชากร อาจจะพิจารณาว่าเป็นข้อสอบที่ยุติธรรมได้ ทั้งนี้ขึ้นอยู่กับจุดมุ่งหมายในการวัดของแบบสอบนั้น (Angoff, 1993)

Angoff (1993) ได้กล่าวว่า เนื้อหาในข้อสอบที่มีแนวโน้มลำเอียง ทำให้ผู้สอบกลุ่มใดกลุ่มหนึ่งได้เปรียบผู้สอบกลุ่มอื่นของประชากร มีดังนี้

1. ผู้สอบกลุ่มผิวดำ มีแนวโน้มได้เปรียบในเนื้อหาเรื่อง ประวัติศาสตร์และวรรณคดีของคนผิวดำ

2. ผู้สอบเพศชาย มีแนวโน้มได้เปรียบในเนื้อหาเกี่ยวกับเรื่อง กีฬาที่ผู้ชายนิยม สงคราม วิทยาศาสตร์ และคณิตศาสตร์

3. ผู้สอบเพศหญิง มีแนวโน้มได้เปรียบในเนื้อหาเกี่ยวกับเรื่อง บุคคลและความสัมพันธ์ระหว่างบุคคล ศิลปะ ดนตรี วรรณคดี และการละคร

Linn (1993) ได้กล่าวว่า ข้อสอบทำหน้าที่ต่างกัน มักจะพบในแบบสอบทางด้านภาษามากกว่าแบบสอบทางด้านคณิตศาสตร์ โดยเขาได้ยกตัวอย่าง การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบสอบ SAT และแบบสอบ preliminary SAT (PSAT) จำนวน 7 ฉบับ รวม 955 ข้อ ซึ่งสอบกับกลุ่มผู้สอบในรัฐนิวยอร์ค ปี ค.ศ. 1988 - 1989 พบว่ามีข้อสอบทำหน้าที่ต่างกัน จำนวน 29 ข้อ จำแนกเป็นข้อสอบทางด้านคณิตศาสตร์ จำนวน 4 ข้อ และเป็นข้อสอบทางด้านภาษา จำนวน 25 ข้อ ในจำนวนข้อสอบทำหน้าที่ต่างกันทางด้านภาษา เป็นข้อสอบเรื่องคำศัพท์ จำนวน 17 ข้อ

การพิจารณาลักษณะของข้อสอบแต่ละข้อว่าจะลำเอียงหรือไม่ อาจใช้การพิจารณาข้อสอบทางด้านตรรก ที่เรียกว่า "วิธีการตัดสินข้อสอบ (judgmental method)" หลักการสำคัญของวิธีการนี้ ก็คือ การใช้ผู้เชี่ยวชาญด้านการพิจารณาความลำเอียงของข้อสอบ พิจารณาตัวข้อสอบอย่างละเอียดถี่ถ้วนว่า ข้อสอบลำเอียงต่อผู้สอบกลุ่มใดของประชากรหรือไม่ ผู้เชี่ยวชาญเหล่านี้เรียกว่า "ผู้ตัดสินข้อสอบ (judge)" การพิจารณาตัดสินข้อสอบอาจกระทำโดยคน ๆ เดียว หรือ

เป็นองค์คณะบุคคล ก็ได้ ผู้ตัดสินข้อสอบจะพิจารณารูปแบบและเนื้อหาของข้อสอบ ให้มีความเป็นธรรมแก่ผู้สอบทั้งจากกลุ่มอ้างอิงและกลุ่มสนใจ โดยยึดหลักการว่า ผู้สอบกลุ่มอ้างอิงและกลุ่มสนใจ ต้องมีความคุ้นเคยและมีประสบการณ์ในเนื้อหาของข้อสอบในระดับที่เท่าเทียมกัน (Tittle, 1982)

Hambleton และ Jones (1993, quoted in Hambleton and others, 1993) ได้ศึกษาเปรียบเทียบความสอดคล้องของผลการตรวจพบข้อสอบทำหน้าที่ต่างกัน ระหว่างวิธีพิจารณาตัดสินข้อสอบ กับวิธี IRT area และวิธีแมนเทล-แฮนส์เซล ผลการวิจัยพบว่า

1. ผลการตรวจพบข้อสอบทำหน้าที่ต่างกันของทั้งวิธี IRT area และวิธีแมนเทล-แฮนส์เซล ไม่ค่อยคงเส้นคงวานัก และมีความสอดคล้องกันในระดับปานกลาง

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

และเขาได้ให้ข้อเสนอแนะเกี่ยวกับการศึกษาการทำหน้าที่ต่างกันของข้อสอบ ไว้ดังนี้

1. นักพัฒนาแบบสอบ ต้องคิดถึงความไม่คงเส้นคงวาของวิธีการทางสถิติเอาไว้ด้วย ดังนั้น จึงควรใช้กลุ่มตัวอย่างขนาดใหญ่ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ และต้องตีความผลการวิเคราะห์ทางสถิติด้วยความระมัดระวัง

2. วิธีแมนเทล-แฮนส์เซล สามารถนำไปใช้ทดแทนวิธี IRT area ได้

3. วิธีการพิจารณาตัดสินข้อสอบ สามารถนำไปใช้จำแนกข้อสอบทำหน้าที่ต่างกันในทางปฏิบัติได้

ในปี ค.ศ. 1988 Hambleton และ Rogers (1988, quoted in Hambleton and Others, 1993) ได้ศึกษาวรรณคดีที่เกี่ยวกับวิธีการตัดสินข้อสอบ เพื่อหาหลักเกณฑ์ในการพิจารณาข้อสอบที่สำเอียง โดยได้รวบรวมหลักเกณฑ์การพิจารณาตัดสินข้อสอบสำเอียงไว้ เป็น 2 กลุ่มใหญ่ ๆ ดังนี้

ลักษณะของข้อสอบสำเอียงทั่ว ๆ ไป

1. เนื้อหาหรือภาษาในข้อสอบยั่วยุให้ผู้สอบกลุ่มสนใจโกรธ เกิดการโต้แย้งหรือเกิดอารมณ์ไม่พอใจ

2. เนื้อหาหรือภาษาในข้อสอบมีความหมายไปในทางลบ ดูถูกเหยียดหยาม หรือก้าวร้าวต่อผู้สอบกลุ่มสนใจ

3. เนื้อหาหรือภาษาในข้อสอบ แสดงว่า ผู้สอบกลุ่มสนใจเป็นคนมีปมด้อย เกี่ยวกับอำนาจ หรือ ความเป็นผู้นำ

4. รูปแบบหรือเนื้อหาของข้อสอบ แสดงว่า ผู้สอบกลุ่มสนใจเป็นคนเจ้าอารมณ์

5. รูปแบบหรือเนื้อหาของข้อสอบ แสดงว่า ผู้สอบกลุ่มสนใจเป็นคนที่มีลักษณะไม่ดี

6. รูปแบบหรือเนื้อหาของข้อสอบ แสดงว่า ผู้สอบกลุ่มสนใจเป็นคนที่มีอาชีพไม่ดี

7. การจัดรูปแบบของข้อสอบลำเอียง หรือก้าวร้าวต่อผู้สอบกลุ่มสนใจ

ลักษณะของข้อสอบลำเอียงต่อเพศ เชื้อชาติ วัฒนธรรม ศาสนา และ ชนชั้นทางสังคม

8. ข้อสอบมีเนื้อหาที่ผู้สอบกลุ่มสนใจไม่คุ้นเคย

9. ข้อสอบวัดในสิ่งที่ไม่สอดคล้องกับสิ่งที่กำหนดไว้ในหลักสูตร (กรณีแบบสอบวัดผลสัมฤทธิ์ทางการเรียน)

10. ข้อสอบมีเนื้อหา ทักษะ หรือสารสนเทศที่ไม่เป็นการทั่วไป กับผู้สอบทุกคน

11. ข้อสอบมีสารสนเทศที่เป็นประโยชน์ กับผู้สอบกลุ่มอ้างอิง

12. ข้อสอบมีข้อความที่รู้ความหมายกันเฉพาะผู้สอบกลุ่มอ้างอิง หรือมีความหมายแตกต่างออกไป สำหรับผู้สอบกลุ่มสนใจ

13. ข้อสอบใช้ภาษา คำศัพท์ หรือสรรพนามเป็นการเฉพาะกับผู้สอบกลุ่มใดกลุ่มหนึ่ง

14. ข้อสอบมีตัวลวงที่ดึงดูดความสนใจในการเลือกตอบ เฉพาะแก่ผู้สอบกลุ่มสนใจด้วยเหตุผลทางด้านวัฒนธรรม

15. รูปแบบหรือโครงสร้างของข้อสอบ เป็นปัญหากับผู้สอบกลุ่มใดกลุ่มหนึ่งมากกว่ากลุ่มอื่น

ตอนที่ 3 งานวิจัยที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

งานวิจัยที่เกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ หรือความลำเอียงของข้อสอบเท่าที่มีผู้ทำไว้นั้น แบ่งได้เป็นงานวิจัยในประเทศไทยและงานวิจัยต่างประเทศ ดังนี้

3.1 งานวิจัยในประเทศ

งานวิจัยในประเทศนั้น ผู้วิจัยขอเสนองานวิจัยที่เกี่ยวกับความลำเอียงของข้อสอบเท่าที่ตรวจพบทั้งหมด ดังนี้

ชัชชัย เผ่าพงศ์ (2527) ได้ศึกษาความลำเอียงของข้อสอบระหว่างผู้สอบเพศชายและเพศหญิง ในแบบทดสอบวัดความถนัดทางการเรียนด้านคณิตศาสตร์และภาษา ระดับชั้นมัธยมศึกษาตอนต้น ซึ่งพัฒนาโดยสำนักทดสอบทางการศึกษาและจิตวิทยา มหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร จำนวน 2 ฉบับ โดยใช้วิธีวิเคราะห์โค้งลักษณะข้อสอบแบบ 3 พารามิเตอร์ กลุ่มตัวอย่างเป็นนักเรียนชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2524 จากทุกภาคภูมิศาสตร์ของประเทศ ได้แก่ ภาคเหนือ ภาคกลาง ภาคใต้ ภาคตะวันออก และ ภาคตะวันออกเฉียงเหนือ กลุ่มตัวอย่างที่ใช้วิเคราะห์แบบทดสอบวัดความถนัดทางการเรียนด้านคณิตศาสตร์ เป็นนักเรียนชาย 1,610 คนและนักเรียนหญิง 1,337 คน ส่วนกลุ่มตัวอย่างที่ใช้วิเคราะห์แบบทดสอบวัดความถนัดทางการเรียนด้านภาษา เป็นนักเรียนชาย 1,316 คน และนักเรียนหญิง 985 คน

ผลการศึกษา พบว่า

1. ในแบบทดสอบวัดความถนัดทางการเรียนด้านคณิตศาสตร์ จำนวน 30 ข้อ พบข้อสอบลำเอียงต่อกลุ่มนักเรียนชาย 7 ข้อ และลำเอียงต่อกลุ่มนักเรียนหญิง 2 ข้อ ข้อสอบลำเอียงต่อกลุ่มนักเรียนชายและกลุ่มนักเรียนหญิงในระดับปานกลางขึ้นไป มีจำนวน 5 ข้อ ซึ่งเนื้อหาของข้อสอบวัดในเรื่อง ร้อยละ การหาปริมาตร และการหาความยาวเส้นรอบรูปสามเหลี่ยม เรื่องละ 1 ข้อ ส่วนอีก 2 ข้อ เนื้อหาของข้อสอบวัดในเรื่องโจทย์ปัญหาทางคณิตศาสตร์ ในจำนวนข้อสอบ 5 ข้อนี้ เป็นข้อสอบลำเอียงต่อกลุ่มนักเรียนชาย 4 ข้อ ส่วนอีก 1 ข้อ ลำเอียงต่อกลุ่มนักเรียนหญิงในช่วงความสามารถต่ำ และลำเอียงต่อกลุ่มนักเรียนชายในช่วงความสามารถสูง

2. ในแบบทดสอบวัดความถนัดทางการเรียนด้านภาษาเกี่ยวกับความเข้าใจในการอ่าน จำนวน 40 ข้อ พบข้อสอบลำเอียงต่อกลุ่มนักเรียนชาย 3 ข้อ และลำเอียงต่อกลุ่มนักเรียนหญิง 8 ข้อ ข้อสอบลำเอียงต่อกลุ่มนักเรียนชายและกลุ่มนักเรียนหญิงในระดับปานกลางขึ้นไป มีจำนวน 9 ข้อ ซึ่งเนื้อหาของข้อสอบวัดในเรื่อง ความเข้าใจเกี่ยวกับการอ่านคำประพันธ์และการอ่านบทร้อยกรองเรื่องละ 1 ข้อ ส่วนอีก 7 ข้อเนื้อหาของข้อสอบวัดในเรื่อง ความเข้าใจเกี่ยวกับการอ่านข้อความ ในจำนวนข้อสอบ 9 ข้อนี้เป็นข้อสอบที่ลำเอียงต่อกลุ่ม

นักเรียนชาย 1 ข้อ และสำเอียงต่อกลุ่มนักเรียนหญิง 6 ข้อ ส่วนอีก 2 ข้อสำเอียงต่อกลุ่มนักเรียนชายในช่วงความสามารถต่ำ และสำเอียงต่อกลุ่มนักเรียนหญิงในช่วงความสามารถสูง

ทัศนีย์ พีรมนตรี (2530) ได้ศึกษาความลำเอียงของแบบสอบวิชาคณิตศาสตร์ของโครงการตรวจสอบคุณภาพการศึกษานักเรียนชั้นมัธยมศึกษาปีที่ 6 ปีการศึกษา 2526 โดยใช้วิธีวิเคราะห์ 3 วิธี ได้แก่ 1) วิธีกำหนดจุดค่าเคลต้า 2) วิธีทดสอบความแตกต่างระหว่างกลุ่มด้วยสถิติไค-สแควร์ในโมเดลล็อก-ลิเนียร์ 2 โมเดล คือ โมเดลที่ไม่มีพารามิเตอร์ผลร่วมระหว่างระดับคะแนนกับกลุ่ม และโมเดลที่ไม่มีพารามิเตอร์ของผลหลักที่เกิดจากกลุ่ม 3) วิธีตามทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ ตัวแปรที่ศึกษาความลำเอียงคือ ภาคภูมิศาสตร์ของผู้สอบ โดยเปรียบเทียบจำนวนข้อสอบสำเอียงระหว่างกลุ่มนักเรียนในกรุงเทพมหานครกับกลุ่มนักเรียนในภาคภูมิศาสตร์ 5 ภาค ได้แก่ ภาคกลาง ภาคใต้ ภาคเหนือ ภาคตะวันออกเฉียงเหนือ และภาคตะวันออก กลุ่มตัวอย่างเป็นนักเรียนชั้นมัธยมศึกษาปีที่ 6 ปีการศึกษา 2526 จำนวนทั้งหมด 7,036 คน แยกเป็น กรุงเทพมหานคร 1,410 คน ภาคกลาง 1,169 คน ภาคใต้ 972 คน ภาคเหนือ 1,185 คน ภาคตะวันออกเฉียงเหนือ 1,921 คน และภาคตะวันออก 379 คน

ผลการศึกษา พบว่า

1. เมื่อจำแนกข้อสอบในแบบสอบวิชาคณิตศาสตร์ จำนวน 60 ข้อ ออกตามค่าความยากซึ่งวิเคราะห์ด้วยทฤษฎีการวัดแบบดั้งเดิมเป็น 3 ระดับ คือ กลุ่มข้อที่ยาก กลุ่มข้อที่ยากง่ายปานกลาง และกลุ่มข้อที่ง่าย แล้ว พบข้อสอบไม่สำเอียง จำนวน 17 ข้อ และข้อสอบสำเอียงเข้าหาภาคภูมิศาสตร์ของผู้สอบ จำนวน 43 ข้อ เหมือนกันในแต่ละระดับความยากและทุก ๆ ภาคภูมิศาสตร์ รวมทั้งกรุงเทพมหานคร

2. เมื่อเปรียบเทียบจำนวนข้อสอบสำเอียงระหว่างกลุ่มนักเรียนในกรุงเทพมหานคร กับกลุ่มนักเรียนในแต่ละภาคภูมิศาสตร์ พบว่า วิธีที่ใช้ทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ พบจำนวนข้อสอบสำเอียงมากที่สุด แต่ละวิธีพบข้อสอบสำเอียงซ้ำกัน ระหว่างกลุ่มนักเรียนในกรุงเทพมหานครกับกลุ่มนักเรียน ในทุก ๆ ภาคภูมิศาสตร์ ในแต่ละภาคภูมิศาสตร์มีจำนวนข้อสอบสำเอียงกับกรุงเทพมหานครไม่เท่ากันในแต่ละวิธี วิธีที่ 1 และที่ 2 พบจำนวนข้อสอบสำเอียงมากที่สุดระหว่างกลุ่มนักเรียนในกรุงเทพมหานครกับกลุ่มนักเรียนในภาคตะวันออกเฉียงเหนือ ส่วนวิธีที่ 3 พบข้อสอบสำเอียงมากที่สุดระหว่างกลุ่มนักเรียนในกรุงเทพมหานครกับกลุ่มนักเรียนในภาคตะวันออก

3. เมื่อเปรียบเทียบจำนวนข้อสอบลำเอียงภายในภาคภูมิศาสตร์เดียวกัน ของวิธีวิเคราะห์ 3 วิธี พบข้อสอบลำเอียงซ่ำกัน ข้อสอบลำเอียงส่วนใหญ่เป็นข้อสอบที่ง่าย สำหรับกลุ่มนักเรียนในกรุงเทพมหานครมากกว่ากลุ่มนักเรียนในภาคภูมิศาสตร์อื่น ๆ ตามวิธีที่ 1 และเป็นข้อสอบลำเอียงอย่างสม่ำเสมอตามวิธีที่ 2 แต่เป็นข้อสอบลำเอียงในเกณฑ์ต่ำตามวิธีที่ 3 ทั้ง 3 วิธีไม่พบข้อสอบลำเอียงซ่ำกันระหว่างกลุ่มนักเรียนในกรุงเทพมหานครกับกลุ่มนักเรียนในทุก ๆ ภาคภูมิศาสตร์ของผู้สอบ

พัชรี ปิยภัณฑ์ (2531) ได้ศึกษาความลำเอียงของข้อสอบจากแบบสอบวัดผลสัมฤทธิ์ทางการเรียนวิชาคณิตศาสตร์ ชั้นประถมศึกษาปีที่ 6 จำนวน 45 ข้อซึ่งผู้วิจัยสร้างขึ้นเอง โดยใช้วิธีวิเคราะห์ 3 วิธี ได้แก่ 1) วิธีแปลงค่าความยากของข้อสอบ 2) วิธีไค-สแควร์ และ 3) วิธีโค้งลักษณะข้อสอบ แบบ 1 พารามิเตอร์ ตัวแปรที่ศึกษาความลำเอียง คือ เพศ และภาคภูมิศาสตร์ของผู้สอบ (กรุงเทพมหานครกับจังหวัดสมุทรสาคร) กลุ่มตัวอย่างที่ใช้เป็นนักเรียนชายและหญิงชั้นประถมศึกษาปีที่ 6 ปีการศึกษา 2530 ของโรงเรียนประถมศึกษาสังกัดสำนักงานคณะกรรมการการประถมศึกษาแห่งชาติในกรุงเทพมหานคร จำนวน 913 คน และจังหวัดสมุทรสาคร จำนวน 509 คน รวมจำนวน 1,422 คน โดยวิเคราะห์ความลำเอียงของข้อสอบ แล้วหาค่าสัมประสิทธิ์สหสัมพันธ์ของวิธีวิเคราะห์ทั้ง 3 วิธี เปรียบเทียบค่าความเที่ยงของแบบสอบก่อนและหลังการคัดเลือกข้อสอบลำเอียงออกจากแบบสอบ และวิเคราะห์ตัวลวงเพื่อตรวจทานแหล่งของความลำเอียง

ผลการศึกษา พบว่า

1. การวิเคราะห์ความลำเอียงของข้อสอบ โดยวิธีโค้งลักษณะข้อสอบ แบบ 1 พารามิเตอร์ พบจำนวนข้อสอบลำเอียงมากที่สุด รองลงมา คือวิธีไค-สแควร์ ส่วนวิธีแปลงค่าความยากของข้อสอบ พบจำนวนข้อสอบลำเอียงน้อยที่สุด

2. ค่าสัมประสิทธิ์สหสัมพันธ์ของดัชนีความลำเอียงระหว่างกลุ่มนักเรียนชายกับนักเรียนหญิงจากการวิเคราะห์โดยวิธีแปลงค่าความยากของข้อสอบ กับวิธีโค้งลักษณะข้อสอบ แบบ 1 พารามิเตอร์ มีค่า .5618 ($p < .001$) และค่าสัมประสิทธิ์สหสัมพันธ์ของดัชนีความลำเอียงระหว่างนักเรียนในกรุงเทพมหานครกับจังหวัดสมุทรสาคร จากการวิเคราะห์โดยวิธีไค-สแควร์กับวิธีโค้งลักษณะข้อสอบ แบบ 1 พารามิเตอร์ มีค่า .6009 ($p < .001$) ส่วนวิธีอื่นสัมพันธ์กันอย่างไม่มีนัยสำคัญ

3. ค่าความเที่ยงของแบบสอบก่อน และหลังการคัดเลือกข้อสอบลำเอียงออกจากแบบสอบแตกต่างกัน อย่างไม่มีนัยสำคัญ

4. แหล่งของความลำเอียงในข้อสอบลำเอียงทุกข้อ ไม่ได้เกิดจากตัวลวง

สุรศักดิ์ อมรรัตนศักดิ์ (2531) ได้ศึกษาผลของวิธีการวิเคราะห์ความลำเอียงของข้อสอบในแบบสอบแข่งขันเพื่อบรรจุเข้ารับราชการครู สังกัดสำนักงานคณะกรรมการการประถมศึกษาแห่งชาติ ปี พ.ศ. 2529 ในด้านสัมประสิทธิ์สหสัมพันธ์ระหว่างวิธีวิเคราะห์ความลำเอียงของข้อสอบ 4 วิธี ได้แก่ 1) วิธีวิเคราะห์ความแปรปรวน 2) วิธีแปลงค่าความยากของข้อสอบ 3) วิธีโค้งลักษณะข้อสอบ แบบ 1 พารามิเตอร์ 4) วิธีโค้งลักษณะข้อสอบ แบบ 3 พารามิเตอร์ และเปรียบเทียบความแตกต่างของผลการคัดเลือกก่อนและหลังการศึกษาความลำเอียงของข้อสอบตามวิธีการคิดคะแนนรวมที่แตกต่างกัน 6 วิธีในด้านจำนวนผู้ได้รับการคัดเลือก สัดส่วนของชาย : หญิงที่ได้รับการคัดเลือกและความเที่ยงของแบบสอบ กลุ่มตัวอย่างเป็นบุคคลที่สอบแข่งขันเพื่อบรรจุเข้าเป็นข้าราชการครู สังกัดสำนักงานคณะกรรมการการประถมศึกษาแห่งชาติ ปี พ.ศ. 2529

ผลการศึกษา พบว่า

1. วิธีโค้งลักษณะข้อสอบ แบบ 3 พารามิเตอร์ ค้นพบจำนวนข้อสอบลำเอียงได้มากที่สุด รองลงมาได้แก่ วิธีวิเคราะห์ความแปรปรวน ส่วนวิธีที่ค้นพบจำนวนข้อสอบลำเอียงได้น้อยที่สุด ได้แก่ วิธีแปลงค่าความยากของข้อสอบ

2. วิธีวิเคราะห์ความลำเอียงของข้อสอบทั้ง 4 วิธี มีความสัมพันธ์กันทางบวกอย่างมีนัยสำคัญที่ระดับ .001 โดยมีค่าอยู่ระหว่าง .7535 ถึง .9921

3. การใช้คะแนนดิบและคะแนนรวมแบบอื่น ๆ อีก 5 วิธี มีจำนวนผู้ได้รับการคัดเลือกแตกต่างกันประมาณร้อยละ 4 ถึง 24 ส่วนการใช้คะแนนมาตรฐานทีปกติรวมกับคะแนนแปลงแบบอื่น ๆ อีก 4 วิธี มีจำนวนผู้ได้รับการคัดเลือกแตกต่างกันร้อยละ 4 ถึง 23

4. เมื่อตัดข้อสอบลำเอียงออก พบว่า สัดส่วนเพศชายต่อเพศหญิงที่ได้รับการคัดเลือกใกล้เคียงกัน และค่าความเที่ยงของแบบสอบลดลงเล็กน้อย

สุพัฒน์ สุกมลสันต์ (2534) ได้วิเคราะห์ความลำเอียงของข้อสอบวิชาภาษาอังกฤษ เพื่อคัดเลือกบุคคลเข้าศึกษาในสถาบันอุดมศึกษาสังกัดทบวงมหาวิทยาลัย โดยใช้ข้อมูลการตอบแบบสอบวิชาภาษาอังกฤษชุด กข และ ชุด กขค ปี พ.ศ. 2531-2533 ซึ่งมีข้อสอบชุดละ 100 ข้อ ตัวแปรที่ศึกษาความลำเอียง คือ เพศ และภาคภูมิศาสตร์ของผู้สอบ ซึ่งแยกออกตามภูมิลำเนาของผู้สอบเป็น 5 ภาค ได้แก่ ภาคกลาง ภาคตะวันออก ภาคตะวันออกเฉียงเหนือ ภาคเหนือ และภาคใต้ วิธีวิเคราะห์ความลำเอียงของข้อสอบมี 3 วิธี ได้แก่ 1) วิธีกำหนดจุดค่าเดลต้า 2) วิธีไค-สแควร์ ชนิดที่แบ่งความสามารถของผู้สอบเป็น 3 ระดับ ได้แก่ กลุ่ม

ความสามารถระดับต่ำ (ผู้ได้คะแนนรวม 0 - 40 คะแนน) กลุ่มความสามารถระดับปานกลาง (ผู้ได้คะแนนรวม 41 - 70 คะแนน) กลุ่มความสามารถระดับสูง (ผู้ได้คะแนนรวม 71 - 100 คะแนน) 3) วิธีวัดพื้นที่ความแตกต่างระหว่างโค้งลักษณะข้อสอบซึ่งวิเคราะห์ตามทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์

ผลการศึกษา พบว่า

1. แบบสอบวิชาภาษาอังกฤษชุด กข และ ชุด กขค ปี พ.ศ. 2531-2533 พบ ข้อสอบลำเอียงต่อเพศ 7 ถึง 28 ข้อ และ 4 ถึง 41 ข้อ ตามลำดับ โดยมีแนวโน้มลำเอียงต่อเพศชายมากกว่าเพศหญิง และพบข้อสอบลำเอียงต่อภาคภูมิศาสตร์ 6 ถึง 45 ข้อ และ 5 ถึง 43 ข้อ ตามลำดับ โดยลำเอียงต่อผู้สอบภาคภูมิศาสตร์อื่น ๆ มากกว่าภาคกลาง 2 - 3 เท่า

2. การเปรียบเทียบผลการวิเคราะห์ความลำเอียงของข้อสอบ โดยใช้วิธีวิเคราะห์ที่ต่างกัน พบว่า มีจำนวนข้อสอบลำเอียงต่างกันอย่างมีนัยสำคัญ แต่ละวิธีให้ผลที่มีความสัมพันธ์กันอย่างไม่มีนัยสำคัญ ผลการวิเคราะห์ที่ใช้ทฤษฎีการตอบสนองข้อสอบ พบจำนวนข้อสอบลำเอียงมากที่สุด (เกณฑ์พื้นที่ความแตกต่างระหว่างโค้งลักษณะข้อสอบมากกว่า 0.40) รองลงมาได้แก่ วิธีไค-สแควร์ และวิธีกำหนดจุดค่าเคลต้า ตามลำดับ ซึ่งวิธีกำหนดจุดค่าเคลต้าพบจำนวนข้อสอบลำเอียง น้อยกว่าสองวิธีแรก ประมาณ 3 - 4 เท่า

นิรมล ชัยชวลิต (2537) ได้เปรียบเทียบผลการวิเคราะห์ความลำเอียงของข้อสอบในแบบสอบความเข้าใจในการอ่านภาษาไทย ชั้นมัธยมศึกษาปีที่ 1 จำนวน 50 ข้อ ชนิด 5 ตัวเลือก ซึ่งผู้วิจัยสร้างขึ้นเอง โดยใช้วิธีวิเคราะห์ต่างกัน 3 วิธี ได้แก่ 1) วิธีแปลงค่าความยากของข้อสอบ 2) วิธีไค-สแควร์ 3) วิธีวิเคราะห์ความแปรปรวน โดยเปรียบเทียบจำนวนข้อสอบลำเอียง ค่าความเที่ยงของแบบสอบหลังคัดเลือกข้อสอบลำเอียงออกแล้ว โดยสูตรคำนวณ แบบ KR-20 และสูตรคำนวณ แบบแบ่งครึ่งฉบับ กลุ่มตัวอย่างเป็นนักเรียนชายและนักเรียนหญิง ชั้นมัธยมศึกษาปีที่ 1 ปีการศึกษา 2536 ของโรงเรียนมัธยมศึกษา สังกัดกรมสามัญศึกษา เขตพระโขนง กรุงเทพมหานคร จำนวน 1,066 คน แบ่งเป็นกลุ่มตัวอย่างที่ใช้เปรียบเทียบผลการวิเคราะห์ความลำเอียงของข้อสอบ จำนวน 466 คน เป็นนักเรียนชาย 233 คน กับนักเรียนหญิง 233 คน และกลุ่มตัวอย่างที่ใช้เปรียบเทียบความเที่ยงของแบบสอบจำนวน 600 คน

ผลการศึกษา พบว่า

1. การวิเคราะห์ความลำเอียงของข้อสอบด้วยวิธีการแปลงค่าความยากของข้อสอบ กับ วิธีไค-สแควร์ และวิธีการแปลงค่าความยากของข้อสอบ กับ วิธีวิเคราะห์ความแปรปรวน พบจำนวนข้อสอบลำเอียงแตกต่างกัน อย่างมีนัยสำคัญที่ระดับ .05 ส่วนวิธีไค-สแควร์ กับ วิธีวิเคราะห์ความแปรปรวน พบจำนวนข้อสอบลำเอียงแตกต่างกันอย่างไม่มีนัยสำคัญ

2. ค่าความเที่ยงของแบบสอบที่ใช้สูตรคำนวณแบบ KR-20 และแบบแบ่งครึ่งฉบับ หลังจากคัดเลือกข้อสอบลำเอียงออกด้วยวิธีแปลงค่าความยากของข้อสอบ วิธีไค-สแควร์ และ วิธีวิเคราะห์ความแปรปรวน มีค่าแตกต่างกันอย่างไม่มีนัยสำคัญ

กาญจนา วัจนสุนทร (2538) ได้พัฒนาเกณฑ์ตัดสินข้อสอบลำเอียงทางเพศ ด้วยข้อมูลเชิงประจักษ์ สำหรับดัชนี 4 ตัว ได้แก่ 1) พื้นที่ระหว่างโค้งลักษณะข้อสอบ ชนิดคิดเครื่องหมาย (SA) จากทฤษฎีการตอบสนองข้อสอบ แบบ 2 พารามิเตอร์ 2) พื้นที่ระหว่างโค้งลักษณะข้อสอบ ชนิดไม่คิดเครื่องหมาย (US) จากทฤษฎีการตอบสนองข้อสอบ แบบ 2 พารามิเตอร์ 3) ดัชนีแอลฟา (α_{MH}) จากวิธีแมนเทล-แฮนส์เซล 4) ดัชนีเบต้า (β_{SIB}) จากวิธี SIBTEST โดยใช้ข้อมูลผลการตอบข้อสอบคัดเลือกบุคคลเข้าศึกษาในสถาบันอุดมศึกษาสังกัดทบวงมหาวิทยาลัย ปีการศึกษา 2535 ในความยาวแบบสอบ 20, 30 และ 40 ข้อ สำหรับวิชาคณิตศาสตร์ และความยาวแบบสอบ 50, 60, 70 และ 80 ข้อ สำหรับวิชาภาษาอังกฤษ และใช้กลุ่มผู้สอบ 6 ขนาด คือ 100, 200, 400, 600, 800 และ 1,000 คนต่อกลุ่ม การพัฒนาเกณฑ์ทำโดยการคำนวณค่าดัชนีทั้ง 4 ตัว จากข้อมูลผลการตอบของผู้สอบเพศเดียวกัน เพศละ 50 ค่า สำหรับแต่ละความยาวแบบสอบและขนาดกลุ่มผู้สอบ จากนั้นนำค่าดัชนีที่ได้ทั้งหมดมาวิเคราะห์ค่าเฉลี่ยและกำหนดเกณฑ์จากค่าเฉลี่ย 2 ลักษณะ คือ เกณฑ์ที่กำหนดจากค่าเฉลี่ยซึ่งรวมค่าดัชนีทุกข้อ โดยไม่พิจารณาความแตกต่างในด้านความยาวของแบบสอบและขนาดกลุ่มผู้สอบ และเกณฑ์ที่กำหนดจากค่าเฉลี่ยโดยพิจารณาถึงความยาวของแบบสอบและขนาดกลุ่มผู้สอบ จากนั้นนำเกณฑ์ที่กำหนดไปตัดสินค่าดัชนีที่ได้จากการวิเคราะห์ระหว่างผู้สอบเพศหญิงและชายพบว่า ความสอดคล้องของการตัดสินภายในดัชนีเดียวกัน มีความไม่คงที่ข้ามขนาดกลุ่มผู้สอบ แต่ความสอดคล้องมีแนวโน้มสูงขึ้น ที่ขนาดกลุ่มผู้สอบตั้งแต่ 600 คน ขึ้นไป

ผลการศึกษา พบว่า

1. เกณฑ์ที่พัฒนาจากข้อมูลเชิงประจักษ์เพื่อใช้ในการตัดสินความลำเอียงของข้อสอบระหว่างเพศชายและเพศหญิง คือ

(1) |SA| > .80 และ UA > .50 กรณีความยาวแบบสอบต่ำกว่า 50 ข้อ

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

(3) .60 > α_{MH} > 1.40 และ β_{SIB} > .06 ในทุกขนาดความยาวของแบบสอบและทุกขนาดกลุ่มผู้สอบ

ทั้งนี้ ควรใช้ขนาดกลุ่มผู้สอบอย่างน้อย 800 คน สำหรับดัชนี SA และ UA และขนาดกลุ่มผู้สอบ 600 คน สำหรับดัชนี α_{MH} และ β_{SIB}

2. การตรวจค้นข้อสอบลำเอียงทางเพศมีความไม่คงที่ข้ามขนาดกลุ่มผู้สอบและขนาดความยาวของแบบสอบ

3. ความสอดคล้องในการตรวจค้นข้อสอบลำเอียงภายในวิธีเดียวกันข้ามขนาดกลุ่มผู้สอบค่อนข้างต่ำ แต่จะสูงขึ้นที่ขนาดกลุ่มผู้สอบ ตั้งแต่ 600 คน ขึ้นไป

4. ข้อสอบลำเอียงวิชาคณิตศาสตร์ ส่วนใหญ่ลำเอียงเข้าข้างผู้สอบเพศชาย ส่วนข้อสอบวิชาภาษาอังกฤษลำเอียงเข้าข้างผู้สอบเพศหญิง เมื่อใช้ดัชนี SA และ α_{MH} แต่เมื่อใช้ดัชนี β_{SIB} ให้ผลตรงกันข้าม

เกษร หว่างจิตร (2539) ได้วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ สำหรับแบบสอบคัดเลือกระดับบัณฑิตศึกษาวิชาภาษาไทยและวิชาภาษาอังกฤษ ด้วยวิธีแมนเทล-แฮนส์เซล โดยใช้ข้อมูลผลการตอบข้อสอบวิชาสอบร่วมในส่วนที่เป็นข้อสอบแบบเลือกตอบ ของศูนย์ทดสอบทางการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย รหัสการสอบที่ 383 ตัวแปรที่ศึกษา ได้แก่ เพศ ภูมิลำเนา ประสบการณ์ในการสอบ และสังกัดของสถานศึกษา กลุ่มตัวอย่างที่ใช้เป็นผู้สอบวิชาภาษาไทย จำนวน 506 คน และ วิชาภาษาอังกฤษ จำนวน 501 คน วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ โดยใช้โปรแกรมคอมพิวเตอร์ MH-DIF ของ Angel M. Fidalgo ซึ่งเป็นโปรแกรมคอมพิวเตอร์ที่เขียนด้วยภาษาปาสคาน และวิเคราะห์ความเที่ยงและความตรงตามทฤษฎีของแบบสอบก่อนและหลังการตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบสอบ

ผลการวิจัย พบว่า

1. ข้อสอบทำหน้าที่ต่างกันที่พบ ส่วนใหญ่เป็นแบบอเนกรูป (nonuniform DIF) เมื่อวิเคราะห์กลุ่มผู้สอบจำแนกตามเพศ พบ ข้อสอบทำหน้าที่ต่างกัน ทั้งแบบเอกรูป (uniform DIF) และแบบอเนกรูป มากที่สุด รองลงมาได้แก่ จำแนกกลุ่มผู้สอบตามภูมิลำเนาและสังกัดของสถานศึกษา ส่วนการจำแนกกลุ่มผู้สอบตามประสบการณ์ในการสอบ พบข้อสอบทำหน้าที่ต่างกัน น้อยที่สุด

2. ผลการวิเคราะห์ลักษณะของข้อสอบทำหน้าที่ต่างกันที่พบ ส่วนใหญ่เป็นข้อสอบที่มีค่าจำแนกค่อนข้างต่ำ (0.20 - 0.39) สอดคล้องกันทั้งวิชาภาษาไทยและวิชาภาษาอังกฤษ สำหรับวิชาภาษาไทยข้อสอบทำหน้าที่ต่างกันที่พบ ส่วนมากเป็นข้อสอบที่ง่ายมาก (.80 - 1.00) แต่ในวิชาภาษาอังกฤษข้อสอบทำหน้าที่ต่างกันที่พบ ส่วนใหญ่เป็นข้อสอบค่อนข้างยาก (0.00 - 0.19) เมื่อพิจารณาลักษณะของข้อสอบทำหน้าที่ต่างกัน พบว่า ในวิชาภาษาไทยเป็นข้อสอบที่มีเนื้อหาด้านการใช้ภาษามากที่สุด และในวิชาภาษาอังกฤษเป็นข้อสอบที่มีเนื้อหาด้านการให้รายละเอียดมากที่สุด

3. ค่าความเที่ยงและความตรงตามทฤษฎีของแบบสอบก่อนและหลัง กรณีการตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบสอบ ส่วนใหญ่ไม่แตกต่างกันอย่างมีนัยสำคัญทางสถิติ ทั้งในวิชาภาษาไทยและวิชาภาษาอังกฤษ

จะเห็นว่า การศึกษาความลำเอียงของข้อสอบในประเทศไทยทั้ง 8 เรื่อง ใช้วิธี การตรวจสอบความลำเอียงของข้อสอบต่างกันหลายวิธี ได้แก่ วิธีแปลงค่าความยากของข้อสอบหรือวิธีกำหนดจุดค่าเดลต้า (ทัศนีย์ พีรมนตรี ; สุรศักดิ์ อมรรัตนศักดิ์ ; พัชรี ปิยภัณฑ์ ; สุพัฒน์ สุกมลสันต์ และนิรมล ชัยชวลิต) วิธีวิเคราะห์ความแปรปรวน (สุรศักดิ์ อมรรัตนศักดิ์ และ นิรมล ชัยชวลิต) วิธีที่ใช้ทฤษฎีการตอบสนองข้อสอบ แบบ 1 พารามิเตอร์ (สุรศักดิ์ อมรรัตนศักดิ์ และพัชรี ปิยภัณฑ์) แบบ 2 พารามิเตอร์ (area method) (กาญจนา วัจนสุนทร) แบบ 3 พารามิเตอร์ (ชัชชัย เผ่าพงศ์ ; ทัศนีย์ พีรมนตรี ; สุรศักดิ์ อมรรัตนศักดิ์ ; สุพัฒน์ สุกมลสันต์) วิธีไค-สแควร์ (ทัศนีย์ พีรมนตรี ; พัชรี ปิยภัณฑ์ ; สุพัฒน์ สุกมลสันต์ และนิรมล ชัยชวลิต) วิธีแมนเทล-แฮนส์เซล (กาญจนา วัจนสุนทร ; เกษร หว่างจิตร) และวิธี SIBTEST (กาญจนา วัจนสุนทร) ซึ่งเป็นวิธีที่ใช้แนวความคิดพหุมิติ (multidimentional conception of bias) (Millsap and Everson, 1993 ; Shealy and Stout, 1993) ส่วนตัวแปรจำแนกกลุ่มประชากรที่ศึกษา ได้แก่ เพศของผู้สอบ ประสบการณ์ในการสอบ สังกัดของสถานศึกษา และภาคภูมิศาสตร์ของผู้สอบ งานวิจัยเหล่านี้ส่วนใหญ่เป็นการตรวจสอบข้อสอบลำเอียงในแบบสอบต่าง ๆ และเปรียบเทียบจำนวนข้อสอบลำเอียงที่ตรวจพบเมื่อใช้วิธีวิเคราะห์ทางสถิติที่ต่างกัน แม้ว่างานวิจัยของกาญจนา วัจนสุนทร เป็นการพัฒนาเกณฑ์การตัดสินข้อสอบลำเอียงทางเพศ ซึ่งส่วนหนึ่งใช้แนวคิดของวิธีแมนเทล-แฮนส์เซล ก็เป็นการกำหนดเกณฑ์ตัดสินข้อสอบลำเอียงทางเพศ โดยใช้ข้อมูลจริงภายใต้เงื่อนไขขนาดของกลุ่มตัวอย่างและความยาวของแบบสอบเท่านั้น ดังนั้น ผู้วิจัยจึงมีแนวความคิดที่จะศึกษาวิจัยในเชิงวิธีการ การนำวิธีแมนเทล-แฮนส์เซล ไปใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในสถานการณ์สอบจริง โดย

ศึกษาในประเด็นการแบ่งกลุ่มความสามารถของผู้สอบและความยากของข้อสอบ ต่อการตรวจพบข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ นั่นคือการเพิ่มเงื่อนไขหรือตัวแปรความสามารถผู้สอบและความยากของข้อสอบ เข้าไปในโมเดลการวิเคราะห์ข้อมูลตามวิธีแมนเทล-แฮนส์เซล และเน้นเฉพาะการตรวจพบข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอเท่านั้น เนื่องจากวิธีแมนเทล-แฮนส์เซล ถูกมองว่ามีข้อด้อยในประเด็นการตรวจพบข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ผู้วิจัยเชื่อว่า งานวิจัยนี้จะมีส่วนช่วยให้การนำวิธีแมนเทล-แฮนส์เซลไปใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ มีประสิทธิผลยิ่งขึ้น และจะช่วยให้มีการศึกษาวิจัยในเรื่องนี้มากขึ้น

3.2 งานวิจัยต่างประเทศ

งานวิจัยเกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบในต่างประเทศนั้น ผู้วิจัยนำเสนองานวิจัยที่เกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธีแมนเทล-แฮนส์เซล ดังนี้

3.2.1 งานวิจัยที่เกี่ยวกับเกณฑ์การจับคู่กลุ่มผู้สอบตามความสามารถ

Hambleton, et al. (1990, quoted in Hambleton and Others, 1993) ได้เปรียบเทียบผลการตรวจพบข้อสอบทำหน้าที่ต่างกัน ในแบบสอบวัดความรอบรู้ระดับชั้นมัธยมศึกษาตอนปลาย ระหว่างการใช้คะแนนรวมของแบบสอบทั้งฉบับ (เกณฑ์ภายในแบบสอบ) กับคะแนนจากแบบสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชั้นมัธยมศึกษาตอนปลาย (เกณฑ์ภายนอกแบบสอบ) เป็นเกณฑ์การจับคู่กลุ่มผู้สอบสองกลุ่ม ส่วนตัวแปรจำแนกกลุ่มประชากรที่ศึกษา ได้แก่ ตัวแปรเพศ (กลุ่มเพศชายกับกลุ่มเพศหญิง)

ผลการวิจัย พบว่า

ผลการตรวจพบข้อสอบทำหน้าที่ต่างกันในแบบสอบวัดความรอบรู้ระดับชั้นมัธยมศึกษาตอนปลาย ระหว่างการใช้คะแนนของแบบสอบทั้งฉบับ กับคะแนนจากเกณฑ์ภายนอกแบบสอบเป็นเกณฑ์การจับคู่กลุ่มผู้สอบสองกลุ่ม มีความสอดคล้องกันสูง แม้ว่าสหสัมพันธ์ระหว่างคะแนนของแบบสอบทั้งฉบับ กับคะแนนจากแบบสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชั้นมัธยมศึกษาตอนปลาย จะมีค่าเพียงปานกลางก็ตาม ข้อค้นพบนี้สนับสนุนให้ใช้เกณฑ์ภายในแบบสอบเป็นเกณฑ์การจับคู่กลุ่มผู้สอบสองกลุ่ม ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตาม วิธีแมนเทล-แฮนส์เซล ต่อไป

Clauser, et al. (1991a) ได้เปรียบเทียบผลการตรวจพบข้อสอบทำหน้าที่ต่างกัน ในแบบสอบวัดความสามารถระดับชั้นมัธยมศึกษาตอนปลาย ระหว่างการใช้คะแนนของแบบสอบทั้งฉบับ กับ คะแนนของแบบสอบย่อย (subtest score) ในแบบสอบ เป็นเกณฑ์การจับคู่กลุ่มผู้สอบสองกลุ่ม ส่วนตัวแปรจำแนกกลุ่มประชากรที่ศึกษา ได้แก่ ตัวแปรเชื้อชาติ จำนวน 2 กลุ่ม (กลุ่มชาวอเมริกันผิวขาว 1,000 คน กับ กลุ่มชาวอเมริกันพื้นเมือง 1,000 คน) โดยการวิเคราะห์ข้อมูลผลการตอบข้อสอบ จำนวน 91 ข้อ ข้อสอบเหล่านี้สามารถจำแนกออกเป็นแบบสอบย่อยได้ 5 ฉบับ ซึ่งแต่ละฉบับผู้สอบ ต้องใช้ทักษะการตอบข้อสอบให้ถูกแตกต่างกัน ได้แก่ 1) การอ่าน 2) การคำนวณทางด้านคณิตศาสตร์ 3) การตีความหมายตาราง แผนภูมิ หรือแผนที่ 4) ความรู้เดิม (ข้อเท็จจริง) และ 5) ความรู้เดิม (คำตอบถูกไม่ชัดเจน) ในครั้งแรกวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยใช้คะแนนของแบบสอบทั้งฉบับเป็นเกณฑ์การจับคู่กลุ่มผู้สอบ ในครั้งต่อมาวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแต่ละแบบสอบย่อยโดยใช้คะแนนแบบสอบย่อยฉบับนั้น ๆ เป็นเกณฑ์การจับคู่กลุ่มผู้สอบสองกลุ่ม และเพื่อเป็นการควบคุมเนื้อหาของแต่ละแบบสอบย่อยไม่ให้มีผลกับการวิเคราะห์ข้อมูล จึงได้สร้างแบบสอบย่อยขึ้นใหม่ จำนวน 3 ฉบับ โดยการสุ่มข้อสอบจากแบบสอบ 91 ข้อ ซึ่งแบบสอบย่อยใหม่ทั้ง 3 ฉบับ กับ แบบสอบ 91 ข้อ แตกต่างกันในเรื่องขนาดความยาวของแบบสอบเท่านั้น จากนั้นวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบสอบย่อยใหม่ โดยใช้คะแนนของแบบสอบย่อยใหม่เป็นเกณฑ์การจับคู่กลุ่มผู้สอบสองกลุ่ม การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบใช้วิธีแมนเทล-แฮนส์เซล แบบ 2 ขั้นตอน (two steps procedure)

ผลการวิจัย พบว่า

1. ข้อสอบทำหน้าที่ต่างกันที่ตรวจพบ เมื่อวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ โดยใช้คะแนนของแบบสอบทั้งฉบับ เป็นเกณฑ์การจับคู่กลุ่มผู้สอบสองกลุ่ม เมื่อวิเคราะห์ใหม่โดยใช้คะแนนของแบบสอบย่อย เป็นเกณฑ์การจับคู่กลุ่มผู้สอบสองกลุ่ม พบจำนวนข้อสอบทำหน้าที่ต่างกันลดลง 32 % ดังนั้น ในกรณีที่แบบสอบประกอบด้วยกลุ่มเนื้อหาที่แตกต่างกันหลายกลุ่มเนื้อหา นักพัฒนาแบบสอบควรตรวจสอบข้อสอบทำหน้าที่ต่างกันในบริบทของแบบสอบย่อยที่อิงเนื้อหานั้น ทั้งนี้เพื่อหลีกเลี่ยงความคลาดเคลื่อน ประเภทที่ I

2. การตรวจพบข้อสอบทำหน้าที่ต่างกันข้ามแบบสอบย่อย (เนื้อหา) และแบบสอบย่อยใหม่ (สุ่ม) มีจำนวนมากกว่าการใช้คะแนนของแบบสอบทั้งฉบับ เป็นเกณฑ์การจับคู่กลุ่มผู้สอบสองกลุ่ม

3. แบบสอบย่อยซึ่งมีข้อสอบที่ต้องใช้ความรู้เดิมในการตอบ (คำตอบถูกไม่ชัดเจน) การใช้คะแนนของแบบสอบทั้งฉบับ เป็นเกณฑ์การจับคู่กลุ่มผู้สอบสองกลุ่ม พบข้อสอบทำหน้าที่ต่างกันจำนวน 50 % แต่เมื่อวิเคราะห์ใหม่โดยการใช้คะแนนของแบบสอบย่อย ฉบับนี้เป็นเกณฑ์การจับคู่กลุ่มผู้สอบสองกลุ่ม ข้อสอบทำหน้าที่ต่างกันที่พบทั้งหมดในการวิเคราะห์ในครั้งแรก ก็ยังพบว่าทำหน้าที่ต่างกันอีก ดังนั้น การใช้เกณฑ์การจับคู่กลุ่มผู้สอบสองกลุ่มอย่างระมัดระวัง อาจจะช่วยให้สามารถแยกข้อสอบที่เป็นปัญหาซึ่งเกิดจากสร้างข้อสอบ กับข้อสอบทำหน้าที่ต่างกัน ออกจากกันได้

Clauser, et al. (1993) ได้ศึกษาอิทธิพลของการทำให้เกณฑ์การจับคู่กลุ่มผู้สอบมีความบริสุทธิ์ (purification of the matching criterion) ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธีแมนเทล-แฮนส์เซล ระหว่าง แบบ 2 ขั้นตอน (two steps procedure) กับแบบขั้นตอนเดียว (single step procedure) โดยใช้สถานการณ์จำลองสร้างกลุ่มผู้สอบ 2 กลุ่ม ๆ ละ 1,000 คน และสร้างแบบสอบความยาว 3 ขนาด ได้แก่ 20 ข้อ 40 ข้อ และ 80 ข้อ รวมทั้งสร้างข้อสอบทำหน้าที่ต่างกันแบบสม่ำเสมอ ใส่ลงไปในแบบสอบแต่ละขนาดความยาว จำนวน 0 %, 3 %, 8 %, และ 20 %

ผลการวิจัย พบว่า

ผลการตรวจพบข้อสอบทำหน้าที่ต่างกันด้วยวิธีแมนเทล-แฮนส์เซล แบบ 2 ขั้นตอน เท่ากับหรือสูงกว่าแบบขั้นตอนเดียว ในทุกเงื่อนไขของการทดสอบ และไม่เพิ่มความคลาดเคลื่อนประเภทที่ I มากกว่าแบบขั้นตอนเดียว

Donoghue, et al. (1993) ได้ศึกษาองค์ประกอบที่มีผลต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธีแมนเทล-แฮนส์เซล โดยศึกษาจากสถานการณ์จำลอง (monte carlo study) ในประเด็นขนาดความยาวของแบบสอบ 4 ขนาด ได้แก่ 4 ข้อ 9 ข้อ 19 ข้อ และ 39 ข้อ

ผลการวิจัย พบว่า

ถ้าใช้แบบสอบขนาดความยาว 4 ข้อ และ 9 ข้อ เป็นเกณฑ์การจับคู่กลุ่มผู้สอบสองกลุ่มแล้ว ผลการตรวจพบข้อสอบทำหน้าที่ต่างกัน ไม่น่าพอใจนัก แต่ถ้าใช้แบบสอบขนาดความยาว 19 ข้อ และ 39 ข้อ เป็นเกณฑ์การจับคู่กลุ่มผู้สอบสองกลุ่ม ผลการตรวจพบข้อสอบทำหน้าที่ต่างกัน เชื่อถือได้ และองค์ประกอบอื่น ๆ ไม่มีอิทธิพลต่อผลการตรวจพบข้อสอบทำหน้าที่ต่างกัน

3.2.2 งานวิจัยที่เกี่ยวกับอำนาจของสถิติ MH

Swaminathan และ Rogers (1990) ได้เปรียบเทียบผลการตรวจพบข้อสอบทำหน้าที่ต่างกัน ระหว่างวิธีการถดถอยโลจิสติก กับวิธีแมนเทล-แฮนส์เซล โดยใช้สถานการณ์จำลองจำลองเงื่อนไข 6 เงื่อนไข ได้แก่ กลุ่มตัวอย่าง 2 ขนาด คือ ขนาด 250 คน และขนาด 500 คน จำลองแบบสอบที่มีความยาว 3 ขนาด คือ ขนาด 40 ข้อ ขนาด 60 ข้อ และ ขนาด 80 ข้อ ในแบบสอบแต่ละชุดจะมีข้อสอบทำหน้าที่ต่างกัน 20 % เป็นข้อสอบทำหน้าที่ต่างกันแบบสม่ำเสมอครึ่งหนึ่ง ส่วนอีกครึ่งหนึ่งเป็นข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ

ผลการวิจัย พบว่า

1. ในกรณีข้อสอบทำหน้าที่ต่างกันแบบสม่ำเสมอ พบว่า วิธีทั้งสองได้ผลการวิเคราะห์ใกล้เคียงกัน แต่ในกรณีข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ พบว่า วิธีการถดถอยโลจิสติก ได้ผลดีกว่าวิธีแมนเทล-แฮนส์เซล

2. ระดับความคลาดเคลื่อน ประเภท I พบว่า วิธีแมนเทล-แฮนส์เซล คลาดเคลื่อนประมาณร้อยละ 1 แต่วิธีการถดถอยโลจิสติก คลาดเคลื่อนประมาณร้อยละ 1 - 6

3. วิธีการถดถอยโลจิสติก เสียค่าใช้จ่ายในการวิเคราะห์มากกว่าวิธีแมนเทล-แฮนส์เซล ประมาณ 3 - 4 เท่า

Clauser, et al. (1991b) ได้ศึกษาว่าข้อสอบทำหน้าที่ต่างกันลักษณะใด ที่ตรวจไม่พบว่าทำหน้าที่ต่างกัน โดยการจำลองข้อมูลตามทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ จำนวน 75 ข้อ (ค่า c-parameter ของทุกข้อ เท่ากับ 0.20) ประกอบด้วย ข้อสอบทำหน้าที่ไม่ต่างกัน จำนวน 59 ข้อ ซึ่งจำลองค่าความยากและค่าอำนาจจำแนกของข้อสอบตามผลการสอบจากแบบสอบ GMAT ปี ค.ศ. 1985 และ จำลองข้อสอบทำหน้าที่ต่างกันแบบสม่ำเสมอ จำนวน 16 ข้อ ซึ่งมีค่าอำนาจจำแนก 4 ระดับ (0.25, 0.60, 0.90, 1.25) คูณกับระดับความแตกต่างของค่าความยากระหว่างกลุ่มสนใจกับกลุ่มอ้างอิง 4 ระดับ (0.25, 0.50, 1.0, 1.5) ทั้งในกรณีการแจกแจงความสามารถของผู้สอบสองกลุ่มเท่ากันและแตกต่างกัน และจำลองกลุ่มผู้สอบ 2 กลุ่ม ๆ ละ 1,000 คน การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบใช้วิธีแมนเทล-แฮนส์เซล แบบ 2 ขั้นตอน

ผลการวิจัย พบว่า

1. ค่าอำนาจจำแนกของข้อสอบ มีอิทธิพลต่อผลการตรวจพบข้อสอบทำหน้าที่ต่างกัน ข้อสอบที่มีค่าอำนาจจำแนกต่ำ อาจจะตรวจไม่พบการทำหน้าที่ต่างกันของข้อสอบ

2. ค่าความยากของข้อสอบ มีอิทธิพลต่อผลการตรวจพบข้อสอบทำหน้าที่ต่างกัน ข้อสอบที่ยากมาก ๆ อาจจะตรวจไม่พบการทำหน้าที่ต่างกันของข้อสอบ

3. ระดับความแตกต่างของค่าความยากระหว่างกลุ่มผู้สอบสองกลุ่ม มีอิทธิพลต่อผลการตรวจพบข้อสอบทำหน้าที่ต่างกัน

Mazor, et al. (1992) ได้ศึกษาอิทธิพลของขนาดกลุ่มตัวอย่าง ที่มีต่อผลการตรวจพบข้อสอบทำหน้าที่ต่างกัน โดยการจำลองข้อมูลตามทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ (ค่า c-parameter ของทุกข้อเท่ากับ 0.20) ประกอบด้วย ข้อสอบทำหน้าที่ไม่ต่างกัน จำนวน 59 ข้อ ซึ่งจำลองค่าความยากและค่าอำนาจจำแนกของข้อสอบ ตามผลการสอบจากแบบสอบ GMAT ปี ค.ศ. 1985 และจำลองข้อสอบทำหน้าที่ต่างกัน ซึ่งมีค่าอำนาจจำแนกของข้อสอบ 4 ระดับ (0.25, 0.60, 0.90, 1.25) คูณกับค่าความยากของข้อสอบกลุ่มอ้างอิง 5 ระดับ (-2.5, -1.0, 0, 1.0, 2.5) คูณกับระดับความแตกต่างของค่าความยากของข้อสอบระหว่างกลุ่มผู้สอบ 4 ระดับ (0.25, 0.50, 1.00, 1.50) รวมแล้วได้ข้อสอบทำหน้าที่ต่างกัน จำนวน 80 ข้อ (4 X 5 X 4) สุ่มข้อสอบ 80 ข้อ ครั้งละ 16 ข้อ ได้ข้อสอบทำหน้าที่ต่างกัน จำนวน 5 ชุด แล้วนำข้อสอบที่ได้แต่ละชุดไปรวมกับข้อสอบทำหน้าที่ไม่ต่างกัน จำนวน 59 ข้อ ทำให้ได้แบบสอบที่ต้องตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ จำนวน 5 ฉบับ ๆ ละ 75 ข้อ จำลองขนาดกลุ่มผู้สอบ 5 ขนาด ได้แก่ 100, 200, 500, 1,000 และ 2,000 คนต่อกลุ่ม ทั้งในกรณีการแจกแจงความสามารถของกลุ่มผู้สอบทั้งสอง เป็นโค้งปกติที่มีค่าเฉลี่ยเท่ากัน และในกรณีการแจกแจงความสามารถของกลุ่มผู้สอบทั้งสองกลุ่ม มีรูปร่างเหมือนกัน แต่กำหนดให้ค่าเฉลี่ยความสามารถของกลุ่มสนใจต่ำกว่ากลุ่มอ้างอิงหนึ่งส่วนเบี่ยงเบนมาตรฐาน การวิเคราะห์ข้อมูลใช้วิธีแมนเทล-แฮนส์เซล แบบ 2 ขั้นตอน

ผลการวิจัย พบว่า

การเพิ่มขึ้นของขนาดกลุ่มตัวอย่าง สัมพันธ์กับการเพิ่มขึ้นของอำนาจสถิติ MH การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยการใช้กลุ่มผู้สอบขนาดใหญ่ จะทำให้ได้ผลการตรวจพบข้อสอบทำหน้าที่ต่างกันถูกต้องมากกว่าการใช้กลุ่มผู้สอบขนาดเล็ก การใช้กลุ่มผู้สอบขนาด 200 ถึง 1,000 คนต่อกลุ่ม ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ มีความ

เพียงพอ ต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ แต่การใช้กลุ่มผู้สอบขนาด 100 คน ต่อกลุ่ม ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อาจไม่มีความเพียงพอต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

Clauser, et al. (1994) ได้ศึกษาผลของการลดจำนวนชั้นคะแนนของเกณฑ์การจับคู่กลุ่มผู้สอบสองกลุ่ม ที่มีต่ออำนาจสถิติ MH ตามที่ Holland และ Thayer (1988) ได้แนะนำไว้ว่า จำนวนชั้นคะแนนที่ใช้เป็นเกณฑ์การจับคู่กลุ่มผู้สอบสองกลุ่ม ตามวิธีแมนเทล-แฮนส์เซล มีจำนวนเท่ากับ k + 1 ชั้นคะแนน (k แทน จำนวนข้อสอบในแบบสอบ) ดังนั้น ถ้าลดจำนวนชั้นคะแนนมาเป็น 4 หรือ 5 ช่วง เหมือนกับวิธีไค-สแควร์ แบบอื่น ๆ จะเป็นการช่วยเพิ่มอำนาจสถิติ MH หรือไม่ โดยการจำลองข้อมูลตามทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ จำนวน 80 ข้อ ประกอบด้วย ข้อสอบทำหน้าที่ไม่ต่างกัน จำนวน 70 ข้อ และข้อสอบทำหน้าที่ต่างกัน จำนวน 10 ข้อ จำลองกลุ่มผู้สอบขนาด 100, 200, 500, 1,000 และ 2,000 คนต่อกลุ่ม ใช้เกณฑ์การจับคู่กลุ่มผู้สอบตามวิธีของ Holland และ Thayer (1988) และ วิธีแบ่งคะแนนของแบบสอบออกเป็น 20, 10, 5 และ 2 ชั้นคะแนน การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ ใช้วิธีแมนเทล-แฮนส์เซล แบบ 2 ขั้นตอน

ผลการวิจัย พบว่า

1. ในกรณีกลุ่มผู้สอบสองกลุ่มมีการแจกแจงความสามารถคล้ายคลึงกัน การลดจำนวนชั้นคะแนนที่ใช้เป็นเกณฑ์การจับคู่กลุ่มผู้สอบสองกลุ่ม ไม่มีผลต่ออำนาจสถิติ MH ในขณะที่ความคลาดเคลื่อน ประเภทที่ I เพิ่มขึ้นเล็กน้อย หรือไม่เพิ่มขึ้นเลย

2. ในกรณีกลุ่มผู้สอบสองกลุ่มมีการแจงแจงความสามารถแตกต่างกัน การลดจำนวนชั้นคะแนนที่ใช้เป็นเกณฑ์การจับคู่กลุ่มผู้สอบสองกลุ่ม ทำให้อำนาจสถิติ MH เพิ่มขึ้นเล็กน้อย ในขณะที่ความคลาดเคลื่อน ประเภทที่ I เพิ่มขึ้นอย่างมาก

Mazor, et al. (1994) ได้ศึกษาการใช้วิธีแมนเทล-แฮนส์เซล ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบไม่สม่ำเสมอ โดยการจำลองข้อสอบตามทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ (ค่า c-parameter ของทุกข้อ เท่ากับ .20) ประกอบด้วย ข้อสอบทำหน้าที่ไม่ต่างกัน จำนวน 59 ข้อ ซึ่งจำลองค่าความยากและค่าอำนาจจำแนกของข้อสอบตามผลการสอบจากแบบสอบ GMAT ปี ค.ศ 1985 และจำลองข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ตามเงื่อนไขค่าอำนาจจำแนกของข้อสอบ 4 ระดับ (.25, .60, .90, 1.25) คูณกับค่าความยากของข้อสอบกลุ่มอ้างอิง 5 ระดับ (-1.5, -1.0, 0, 1.0, 1.5) คูณกับระดับความแตกต่างของค่าอำนาจจำแนกของสองกลุ่ม 5 ระดับ (0, .25, .50, .75, 1.0) คูณกับระดับความ

แตกต่างของค่าความยากของสองกลุ่ม 4 ระดับ (0, .3, .6, 1.0) ได้ข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ จำนวน 400 ข้อ สุ่มข้อสอบครั้งละ 16 ข้อ ได้ข้อสอบ 25 ชุด นำข้อสอบแต่ละชุดไปรวมกับข้อสอบทำหน้าที่ไม่ต่างกัน ได้แบบสอบจำนวน 25 ฉบับ จำลองกลุ่มผู้สอบขนาด 1,000 คนต่อกลุ่ม การวิเคราะห์ข้อมูลใช้เทคนิคการวิเคราะห์ 3 ครั้ง คือ ครั้งที่ 1 วิเคราะห์การทำหน้าที่ต่างกัน ของข้อสอบในกลุ่มสนใจกับกลุ่มอ้างอิง ตามวิธีการปกติ จากนั้นแบ่งครึ่งกลุ่มผู้สอบแต่ละกลุ่ม ออกเป็น 2 กลุ่มย่อย โดยใช้ค่าเฉลี่ยของคะแนนผลการสอบของผู้สอบกลุ่มสนใจและกลุ่มอ้างอิงรวมกัน ครั้งที่ 2 วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในกลุ่มผู้สอบที่ได้คะแนนผลการสอบสูงกว่าค่าเฉลี่ย จำนวน 1 ครั้ง ครั้งที่ 3 วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ ในกลุ่มผู้สอบที่ได้คะแนนผลการสอบต่ำกว่าค่าเฉลี่ย 1 ครั้ง การวิเคราะห์ข้อมูลใช้วิธีแมนเทล-แฮนส์เซล แบบ 2 ขั้นตอน

ผลการวิจัย พบว่า

1. วิธีแมนเทล-แฮนส์เซล แบบปกติ สามารถตรวจพบข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ได้ถูกต้อง 68 % ส่วนใหญ่แล้วเป็นข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอชนิดที่โค้งลักษณะข้อสอบตัดกันห่างจุดกลางของการแจกแจงความสามารถ แต่ตรวจไม่ค่อยพบข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ชนิดที่โค้งลักษณะข้อสอบตัดกันใกล้จุดกลางของการแจกแจงความสามารถ แต่เมื่อใช้เทคนิคการวิเคราะห์ 3 ครั้ง แล้ว พบว่า สามารถตรวจพบข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ เพิ่มขึ้นอีก 14 % และความคลาดเคลื่อน ประเภทที่ I ไม่เพิ่มขึ้น ซึ่งผลการวิจัยประเด็นนี้เป็นไปในทิศทางเดียวกัน ทั้งกรณีการแจกแจงความสามารถของกลุ่มผู้สอบสองกลุ่มเท่ากันและไม่เท่ากัน

2. วิธีแมนเทล-แฮนส์เซล แบบการวิเคราะห์ 3 ครั้ง สามารถตรวจพบข้อสอบที่ยากง่ายปานกลางได้เพิ่มขึ้น

3. ข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอที่ตรวจพบ มักเป็นข้อสอบที่ง่าย ส่วนข้อสอบที่ยากมักตรวจไม่พบ แสดงให้เห็นว่าค่าความยากของข้อสอบมีผลต่อการตรวจพบข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ

เขาได้ให้ข้อเสนอแนะว่า ควรมีการศึกษาเพิ่มเติมเรื่องนี้กับข้อมูลจริง รวมทั้งมีการเปรียบเทียบผลการวิจัยที่ได้กับวิธีที่ใช้ทฤษฎีการตอบสนองข้อสอบ

จากผลงานวิจัยทั้ง 9 เรื่องที่กล่าวมาแล้ว ได้สนับสนุนให้นำวิธีแมนเทล-แฮนส์เซล ไปใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบต่อไป งานวิจัย 2 เรื่องแรก ได้แก่ งานวิจัยของ Hambleton, et al. (1990) และงานวิจัยของ Clauser, et al. (1991a) เน้นไปที่ เกณฑ์การ

จับคู่กลุ่มผู้สอบ ผลการวิจัยแนะนำว่า การใช้คะแนนรวมของแบบสอบเป็นเกณฑ์การจับคู่กลุ่มผู้สอบมีความเหมาะสม แต่อย่างไรก็ดี ความตรงของการใช้เกณฑ์นี้ขึ้นอยู่กับว่า คะแนนรวมของแบบสอบแทนความสามารถของผู้สอบได้มากน้อยเพียงใด ถ้าข้อสอบแต่ละข้อวัดความสามารถของผู้สอบมากกว่าหนึ่งความสามารถ หรือในกรณีกลุ่มของข้อสอบเหล่านี้วัดความสามารถแตกต่างจากแบบสอบทั้งฉบับ การนำเอาคะแนนของแบบสอบทั้งฉบับมาเป็นเกณฑ์การจับคู่กลุ่มผู้สอบ อาจทำให้ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบคลาดเคลื่อนไปได้ งานวิจัยโดยใช้สถานการณ์จำลองของ Clauser, et al. (1993) เน้นไปที่ การใช้วิธีแมนเทล-แฮนส์เซล แบบ 2 ขั้นตอน ผลวิจัยแนะนำว่า การใช้เทคนิคการวิเคราะห์ แบบ 2 ขั้นตอน มีความเหมาะสม งานวิจัยโดยใช้สถานการณ์จำลองของ Donoghue, et al. (1993) แนะนำว่า แบบสอบที่ใช้เป็นเกณฑ์การจับคู่กลุ่มผู้สอบสองกลุ่ม ควรมีความยาวอย่างน้อย 19 ข้อ ส่วนงานวิจัย 5 เรื่องสุดท้ายเป็นการวิจัยโดยใช้สถานการณ์จำลอง ซึ่งเน้นที่อำนาจสถิติ MH ภายใต้เงื่อนไขต่าง ๆ งานวิจัยของ Swaminathan และ Rogers (1990) ยืนยันว่า วิธีแมนเทล-แฮนส์เซล ไม่ไวต่อการตรวจพบข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ งานวิจัยของ Clauser, et al. (1991b) ศึกษาความสัมพันธ์ระหว่างอำนาจสถิติ MH กับ ความแตกต่างระหว่างค่าความยากของข้อสอบระหว่างกลุ่มผู้สอบ และ ค่าอำนาจจำแนกของข้อสอบ รวมทั้งปฏิสัมพันธ์ของค่าความยากของข้อสอบกับการแจกแจงความสามารถของผู้สอบ งานวิจัยของ Mazor, et al. (1992) ยืนยันว่า วิธีแมนเทล-แฮนส์เซล ใช้ได้กับกลุ่มตัวอย่างขนาดเล็ก แต่ไม่สนับสนุนให้ใช้กับกลุ่มผู้สอบที่น้อยกว่า 200 คนต่อกลุ่ม นอกจากในกรณีที่ต้องการผลวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบอย่างหยาบ ๆ งานวิจัยของ Clauser, et al. (1994) เป็นการศึกษาประโยชน์ของการลดจำนวนชั้นคะแนนในเกณฑ์การจับคู่กลุ่มผู้สอบ ผลการวิจัยแนะนำว่า ในกรณีที่อำนาจของสถิติเพิ่มขึ้น อาจจะมีผลดีแก่การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบอยู่บ้าง แต่ต้องระวังเกี่ยวกับเรื่อง การแจกแจงความสามารถของกลุ่มผู้สอบและความตรงของเกณฑ์การจับคู่กลุ่มผู้สอบ อาจจะลดลง เมื่อจำนวนชั้นของคะแนนน้อยลง โดยเฉพาะในกรณีที่ผู้สอบมาจากกลุ่มผู้สอบที่มีความสามารถแตกต่างกัน งานวิจัยชิ้นสุดท้าย ได้แก่ งานวิจัยโดยใช้สถานการณ์จำลองของ Mazor, et al. (1994) ได้แนะนำการใช้วิธีแมนเทล-แฮนส์เซล ตรวจสอบการทำหน้าที่ต่างกันแบบไม่สม่ำเสมอของข้อสอบ ผลการวิจัยในเบื้องต้น แสดงให้เห็นว่าน่าจะได้ผลดี แม้ว่าในเรื่องนี้ ยังต้องการงานวิจัยเพิ่มเติมอีก

จะเห็นว่าจนถึงขณะนี้ งานวิจัยเกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันแบบไม่สม่ำเสมอของข้อสอบ ด้วยวิธีแมนเทล-แฮนส์เซล มีน้อยมาก และเป็นงานวิจัยขั้นต้นจากสถานการณ์จำลองเท่านั้น ยังไม่มีการศึกษาเรื่องนี้จากสถานการณ์จริง

ตอนที่ 4 แนวทางการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ตามข้อเสนอแนะของ Hambleton, et al. (1993)

Hambleton, et al. (1993) ได้เขียนรายงานความก้าวหน้าเรื่อง "การศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ของมหาวิทยาลัยแมสสาชูเสทท์" สรุปว่า จากการที่คณะของเขาทำการวิจัยเรื่อง การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยวิธีที่ใช้ทฤษฎีการตอบสนองข้อสอบ กับ วิธีแมนเทล-แฮนส์เซล ตลอดช่วงเวลา 12 ปี (ค.ศ.1981 - 1993) ทำให้มั่นใจว่าสามารถออกแบบการศึกษาหรือวิจัยการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบให้มีความตรงสูงได้ จึงได้ให้ข้อเสนอแนะสำหรับใช้เป็นแนวทางในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ดังต่อไปนี้

ข้อเสนอแนะที่ 1 ไม่มีวิธีการใด ๆ ที่สามารถตรวจพบข้อสอบทำหน้าที่ต่างกันในแบบสอบได้ทั้งหมด แต่ละวิธีก็มีข้อบกพร่องในตัวเอง และผลการตรวจพบข้อสอบทำหน้าที่ต่างกัน ก็แตกต่างกันไป ดังนั้น นักพัฒนาแบบสอบที่ต้องตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ในโครงการทดสอบที่สำคัญ จึงควรใช้วิธีการตรวจสอบหลาย ๆ วิธี รวมทั้งการให้ผู้เชี่ยวชาญพิจารณาผลการตรวจพบข้อสอบทำหน้าที่ต่างกัน ที่ไม่สอดคล้องกันด้วย

ข้อเสนอแนะที่ 2 วิธีพิจารณาตัดสินข้อสอบ ก็เป็นวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่เป็นประโยชน์ และมีข้อได้เปรียบวิธีการอื่น ๆ หลายประการ เป็นต้นว่า

1. วิธีพิจารณาตัดสินข้อสอบ เสียค่าใช้จ่ายถูกกว่าวิธีการทางสถิติ เนื่องจากไม่ต้องเก็บรวบรวมข้อมูล

2. การใช้ผู้ตัดสินข้อสอบที่เหมาะสม จะช่วยให้สามารถพิจารณาความตรงเฉพาะหน้าได้ และถ้าใช้ผู้ตัดสินข้อสอบที่มาจากคนกลุ่มสนใจ จะเป็นประโยชน์ในด้านสังคม เชื้อชาติ และการเมือง

3. วิธีการพิจารณาตัดสินข้อสอบ สามารถดำเนินการได้ก่อนการนำแบบสอบไปใช้ จึงทำให้สามารถคัดเลือกข้อสอบที่ไม่เหมาะสม ออกจากแบบสอบก่อนนำข้อสอบไปใช้ในสถานการณ์สอบจริง

4. ในกรณีที่ผู้ตัดสินข้อสอบมีความเชี่ยวชาญทางด้านเนื้อหาวิชา จะช่วยพิจารณาความตรงตามเนื้อหาของแบบสอบ ได้อีกด้วย

ส่วนข้อเสียเปรียบของวิธีพิจารณาตัดสินข้อสอบ มีดังนี้

1. ผลการพิจารณาตัดสินข้อสอบทำหน้าที่ต่างกัน มักไม่สอดคล้องกับผลที่ได้จากวิธีการทางสถิติ

2. ในกรณีที่ต้องนำผู้ตัดสินข้อสอบมาพิจารณาข้อสอบร่วมกัน หรือต้องฝึกอบรมผู้ตัดสินข้อสอบ ทำให้ต้องเสียเวลาและค่าใช้จ่ายเพิ่มขึ้น

3. ผู้ตัดสินข้อสอบมักจะเกิดปัญหา เรื่อง ความเบื่อ ความเหนื่อยล้า ซึ่งอาจกระทบต่อความตรงของผลการพิจารณาตัดสินข้อสอบ

อย่างไรก็ตาม แม้ว่าวิธีพิจารณาตัดสินข้อสอบจะมีข้อเสียเปรียบอยู่บ้าง แต่ก็เป็นประโยชน์ในการพิจารณาข้อสอบทำหน้าที่ต่างกัน

ข้อเสนอแนะที่ 8 ในการพัฒนาแบบสอบ ควรนำวิธีแมนเทล-แฮนส์เซล ไปใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยดำเนินการตามแนวทาง ดังนี้

1. วิธีแมนเทล-แฮนส์เซล แบบ 2 ขั้นตอน (two steps procedure) ซึ่งเสนอโดย Holland และ Thayer (1988) เป็นวิธีการที่ไม่ยุ่งยากในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ การทำให้เกณฑ์การจับคู่กลุ่มผู้สอบมีความบริสุทธิ์ ด้วยวิธีการคัดเลือกข้อสอบทำหน้าที่ต่างกันซึ่งตรวจพบในการวิเคราะห์ขั้นตอนแรก ออกจากเกณฑ์การจับคู่กลุ่มผู้สอบในการวิเคราะห์ขั้นตอนที่สอง ชอบด้วยเหตุผลทางทฤษฎีและมีหลักฐานเชิงประจักษ์สนับสนุน (Clauser, 1993)

2. เกณฑ์ที่ใช้ในการจับคู่กลุ่มผู้สอบ ต้องคาดหมายได้ว่า มีความเป็นเอกมิติ (approximately unidimentional) การฝ่าฝืนข้อตกลงข้อนี้ อาจมีผลต่อความคลาดเคลื่อนประเภทที่ I (Clauser and Others, 1991a ; Ackerman, 1992) ถ้าแบบสอบที่ใช้เป็นเกณฑ์การจับคู่กลุ่มผู้สอบมีปัญหาในเรื่องนี้ ทางแก้ก็คือ การแยกแบบสอบฉบับนี้ออกเป็นแบบสอบย่อย ๆ ตามกลุ่มเนื้อหา แล้ววิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ โดยการใช้คะแนนรวมจากแบบสอบย่อยนั้น เป็นเกณฑ์การจับคู่กลุ่มผู้สอบสองกลุ่ม

3. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยการใช้กลุ่มผู้สอบขนาดใหญ่ดีกว่าการใช้กลุ่มผู้สอบขนาดเล็ก การใช้กลุ่มผู้สอบระหว่าง 200 ถึง 1,000 คนต่อกลุ่ม มีความเพียงพอสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ แต่การใช้กลุ่มผู้สอบที่น้อยกว่า 200 คนต่อกลุ่ม (กลุ่มอ้างอิงหรือกลุ่มสนใจ) อาจไม่มีความเพียงพอสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในบางกรณี

4. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ในกรณีผู้สอบกลุ่มสนใจมีจำนวนน้อย วิธีการเพิ่มอำนาจของสถิติ MH สามารถทำได้ โดยการเพิ่มจำนวนผู้สอบกลุ่มอ้างอิงขึ้น และคงจำนวนผู้สอบกลุ่มสนใจไว้ตามเดิม การใช้อัตราส่วนระหว่างผู้สอบกลุ่มอ้างอิงต่อผู้สอบกลุ่มสนใจเท่ากับ 9 : 1 จะทำให้อำนาจของสถิติ MH สูงขึ้น และไม่ทำให้ความคลาดเคลื่อนประเภทที่ I หรือ ความลำเอียงอื่น ๆ เพิ่มขึ้น (Clauser, 1993)

5. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยการใช้กลุ่มผู้สอบขนาดใหญ่ อาจต้องทดสอบนัยสำคัญ และวัดขนาดอิทธิพลของค่าสถิติที่ได้ เนื่องจากอำนาจของสถิติเพิ่มขึ้นตามขนาดของกลุ่มตัวอย่าง ในกรณีการใช้กลุ่มผู้สอบเกินกว่า 1,000 คนต่อกลุ่ม แม้การทำหน้าที่ต่างกันของข้อสอบเพียงเล็กน้อย ข้อสอบก็อาจจะมีนัยสำคัญทางสถิติได้

6. ผู้สอบที่ใช้เป็นกลุ่มตัวอย่างต้องแทนกลุ่มประชากรที่สนใจศึกษาได้ เนื่องจากค่าสถิติในแต่ละระดับความสามารถของผู้สอบถูกถ่วงน้ำหนัก โดยจำนวนผู้ตอบในแต่ละระดับความสามารถนั้น ๆ ในกรณีที่ข้อสอบยากมาก ๆ อาจตรวจไม่พบการทำหน้าที่ต่างกันของข้อสอบ เพราะข้อสอบเหล่านี้ผู้สอบมีโอกาสตอบถูกน้อย เมื่อแจกแจงจำนวนผู้ตอบตามระดับความสามารถแล้ว ทำให้สัดส่วนผู้ตอบข้อสอบถูกในระดับความสามารถสูง ๆ มีน้อย จึงทำให้อาจตรวจไม่พบข้อสอบทำหน้าที่ต่างกันชนิดนี้ ทางแก้ก็คือ การสุ่มจำนวนผู้สอบที่มีระดับความสามารถสูง ๆ ให้มากขึ้น

7. การรวมกลุ่มชั้นคะแนน ที่ใช้เป็นเกณฑ์การจับคู่กลุ่มผู้สอบตามความสามารถของผู้สอบอาจจะเป็นประโยชน์สำหรับการเพิ่มอำนาจของสถิติ MH อยู่บ้าง ในกรณีที่การแจกแจงความสามารถของผู้สอบกลุ่มอ้างอิงกับกลุ่มสนใจคล้ายคลึงกัน อาจจะทำให้อำนาจของสถิติ MH เพิ่มขึ้นบ้าง และความคลาดเคลื่อนเพิ่มขึ้นเล็กน้อย แต่ในกรณีที่การแจกแจงความสามารถของผู้สอบกลุ่มอ้างอิงกับกลุ่มสนใจแตกต่างกัน ความคลาดเคลื่อนที่เพิ่มขึ้น อาจมีผลกระทบต่อความตรงของเทคนิคนี้ได้ ดังนั้น ในทางปฏิบัติจึงควรหลีกเลี่ยงเทคนิคนี้ หันไปใช้การทำให้การแจกแจงความสามารถของผู้สอบของสองกลุ่มคล้ายคลึงกัน โดยการเก็บรวบรวมข้อมูลให้มากขึ้น เพื่อให้ได้กลุ่มผู้สอบขนาดใหญ่ ซึ่งสามารถทำได้ไม่ยากนัก โดยเฉพาะในกรณีที่ใช้กลุ่มผู้สอบเพศชาย กับเพศหญิง

8. ผู้ที่นำวิธีแมนเทล-แฮนส์เซล ไปใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในสถานการณ์สอบจริง ต้องระมัดระวังเรื่อง ข้อสอบที่มีค่าอำนาจจำแนกต่ำ ซึ่งอาจตรวจไม่พบการทำหน้าที่ต่างกันของข้อสอบ กรณีที่ในแบบสอบมีข้อสอบค่าอำนาจจำแนกต่ำกว่า จำนวนน้อย

ก็อาจคัดเลือกข้อสอบเหล่านี้ออกจากการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ แต่กรณีที่ในแบบสอบมีข้อสอบค่าอำนาจจำแนกต่ำ จำนวนมาก ก็อาจต้องใช้ระดับความมีนัยสำคัญต่ำ

9. วิธีแมนเทล-แฮนส์เซล สามารถใช้ตรวจสอบการทำหน้าที่ต่างกันแบบไม่สม่ำเสมอของข้อสอบได้ แต่อาจต้องใช้เทคนิคอื่น ๆ เพิ่มเติมอีก อย่างไรก็ตาม วิธีนี้ยังไม่ไวกับข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ในบางกรณี (Mazor and Others, 1994)

จากการพิจารณาทบทวนวรรณคดีที่เกี่ยวข้อง ผู้วิจัยได้นำแนวคิด ข้อเสนอแนะและผลการวิจัย มาเป็นกรอบงานวิจัยเรื่อง การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันแบบไม่สม่ำเสมอของข้อสอบ ระหว่างวิธีแมนเทล-แฮนส์เซล แบบปกติ กับ วิธีแมนเทล-แฮนส์เซล แบบแบ่งกลุ่มความสามารถของผู้สอบและความยากของข้อสอบ ดังต่อไปนี้

1. ศึกษากับข้อมูลจริง ตามข้อเสนอแนะของ Uttaro และ Millsap (1994) และ Mazor, et al. (1994)

2. ใช้วิธีแมนเทล-แฮนส์เซล แบบ 2 ขั้นตอน ตามผลงานวิจัยของ Clauser, et al. (1993) และข้อเสนอแนะ ของ Hambleton, et al. (1993)

3. ตัวแปรจำแนกกลุ่มประชากรที่ศึกษา ได้แก่ ตัวแปรเพศ เนื่องจากเป็นตัวแปรที่รู้ชัดและสามารถแบ่งได้โดยปราศจากความคลาดเคลื่อน ตามแนวคิดของ Millsap และ Everson (1993) , และ Zieky (1993)

4. ใช้กลุ่มตัวอย่าง จำนวน 1,200 คน แบ่งเป็นเพศชาย 600 คน และเพศหญิง 600 คน ตามผลงานวิจัยของ Mazor, et al. (1992) และข้อเสนอแนะของ Hambleton, et al. (1993) , กาญจนา วัจนสุนทร (2538)

5. สร้างแบบสอบวัดความสามารถในการอ่านภาษาไทย เพื่อเป็นเครื่องมือเก็บรวมรวมข้อมูล เนื่องจากเพศหญิงมีแนวโน้มได้เปรียบในข้อสอบที่มีเนื้อหาสาระเกี่ยวกับภาษา (สุพัฒน์ สุกมลสันต์, 2534 ; Angoff, 1993 ; กาญจนา วัจนสุนทร, 2538) นอกจากนี้แบบสอบวัดความสามารถในการอ่าน (reading ability) เป็นแบบสอบที่มีความเป็นเอกมิติ ตามแนวคิดของ Millsap และ Everson (1993)

6. ผู้วิจัยเพิ่มเงื่อนไขการแบ่งกลุ่มผู้สอบ ออกเป็นกลุ่มผู้สอบที่มีความสามารถสูง และกลุ่มผู้สอบที่มีความสามารถต่ำ ซึ่งเป็นแนวคิดที่เกิดจากการศึกษาแนวทางการวิจัยของ Mazor, et al. (1994) ในเรื่องการแบ่งครึ่งกลุ่มผู้สอบโดยใช้คะแนนรวมของแบบสอบ และการที่ใช้ค่าประมาณความสามารถของผู้สอบ ($\hat{\theta}$) ที่ได้จากทฤษฎีการตอบสนองข้อสอบ เป็น

เกณฑ์การแบ่งกลุ่มผู้สอบ แทนการใช้คะแนนรวมของแบบสอบ เนื่องจากตามทฤษฎีการตอบสนองข้อสอบ ค่าประมาณความสามารถของผู้สอบ มีความเชื่อถือได้มากกว่า คะแนนรวมของแบบสอบ (Holland and Thayer, 1988 ; Hambleton and Others, 1993)

7. ผู้วิจัยเพิ่มเงื่อนไขการแบ่งกลุ่มข้อสอบ ออกเป็นกลุ่มข้อสอบยาก กลุ่มข้อสอบยากง่ายปานกลาง และกลุ่มข้อสอบง่าย ซึ่งเป็นแนวคิดที่เกิดจากการศึกษางานวิจัยของ Mazor,et al. (1994) ในประเด็น โอกาสในการตรวจพบข้อสอบทำหน้าที่ต่างกันของวิธีแมนเทล-แฮนส์เซลขึ้นอยู่กับฟังก์ชั่นค่าความยากของข้อสอบและการแจกแจงความสามารถของผู้สอบ และการใช้ค่าประมาณความยากของข้อสอบ ($\hat{b}$) ที่ได้จากทฤษฎีการตอบสนองข้อสอบ เป็นเกณฑ์แบ่งกลุ่มข้อสอบ เนื่องจากค่าประมาณความยากของข้อสอบ มีความคงที่มากกว่าค่าความยากของข้อสอบที่ได้จากทฤษฎีการสอบแบบดั้งเดิม (Hambleton and Others, 1993)