บทที่ 2

วรรณคดีที่เกี่ยวข้อง

ในบทนี้จะนำเสนอวรรณคดีที่เกี่ยวข้องกับการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบสำหรับแบบสอบคัดเลือกวิชาภาษาไทยและภาษาอังกฤษด้วยวิธีแมนเทล-เฮนส์เซล โดยจะนำเสนอเป็น 4 ตอน ดังนี้

ตอนที่ 1 ความเป็นมาและความสำคัญของการสอบคัดเลือกวิชาสอบร่วมของศูนย์ทดสอบทางการศึกษา

ตอนที่ 2 ความเป็นมาและวิธีการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ

ตอนที่ 3 วิธีการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมนเทล - เฮนส์เซล

ตอนที่ 4 งานวิจัยที่เกี่ยวข้องกับการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ

ตอนที่ 1 ความเป็นมาและความสำคัญของการสอบคัดเลือกวิชาสอบร่วมของศูนย์ทดสอบทางการศึกษา

คณะครุศาสตร์ซึ่งเป็นคณะวิชาในจุฬาลงกรณ์มหาวิทยาลัย มีหน้าที่ร่วมกับบัณฑิตวิทยาลัยในการผลิตมหาบัณฑิตและดุษฎีบัณฑิตทางการศึกษา โดยมีส่วนร่วมกับภาควิชาในการบริหารและกำหนดโครงสร้างของหลักสูตร ซึ่งปัจจุบัน (ปีการศึกษา 2538) บัณฑิตศึกษา คณะครุศาสตร์เปิดสอนหลักสูตรปริญญาครุศาสตรมหาบัณฑิต รวม 11 ภาควิชา รวม 23 สาขาวิชา หลักสูตรปริญญาครุศาสตรดุษฎีบัณฑิต 8 สาขาวิชา ดังนี้ (สำนักงานบัณฑิตศึกษา, 2538)

หลักสูตรปริญญาครุศาสตรมหาบัณฑิต			
ที่	ภาควิชา	ที่	สาขาวิชา
1.	สารัตถศึกษา	1.	พื้นฐานการศึกษา
2.	วิจัยการศึกษา	1.	วิจัยการศึกษา
		2.	สถิติการศึกษา
		3.	การวัดและประเมินผลการศึกษา
3.	ประถมศึกษา	1.	ประถมศึกษา
		2.	การศึกษาปฐมวัย
4.	มัธยมศึกษา	1.	การสอนภาษาไทย
		2.	การสอนภาษาอังกฤษ
		3.	การสอนสังคมศึกษา
		4.	การศึกษาคณิตศาสตร์
		5.	การศึกษาวิทยาศาสตร์
5.	พลศึกษา	1.	พลศึกษา
		2.	สุขศึกษา
6.	บริหารการศึกษา	1.	บริหารการศึกษา
		2.	นิเทศการศึกษาและพัฒนาหลักสูตร
7.	จิตวิทยา	1.	จิตวิทยาการศึกษา
		2.	จิตวิทยาการปรึกษา
		3.	จิตวิทยาสังคม
		4.	จิตวิทยาพัฒนาการ
8.	โสตทัศนศึกษา	1.	โสตทัศนศึกษา
9.	ศิลปศึกษา	1.	ศิลปศึกษา
10.	อุดมศึกษา	1.	อุดมศึกษา
11.	การศึกษานอกระบบโรงเรียน	1.	การศึกษานอกระบบโรงเรียน

สำหรับหลักสูตรปริญญาครุศาสตร์ดุษฎีบัณฑิต มีดังนี้

หลักสูตรปริญญาครุศาสตร์ดุษฎีบัณฑิต			
ที่	ภาควิชา	ที่	สาขาวิชา
1.	สารัตถศึกษา	1.	พัฒนศึกษา
2.	วิจัยการศึกษา	2.	การวัดและประเมินผลการศึกษา
3.	บริหารการศึกษา	3.	บริหารการศึกษา
4.	จิตวิทยา	4.	จิตวิทยาการศึกษา
5.	พลศึกษา	5.	พลศึกษา
6.	อุดมศึกษา	6.	อุดมศึกษา
7.	โสตทัศนศึกษา	7.	เทคโนโลยีและการสื่อสารการศึกษา
8.	คณะกรรมการบริหารหลักสูตรดุษฎีบัณฑิต	8.	หลักสูตรและการสอน

ศูนย์ทดสอบทางการศึกษาเป็นหน่วยงานในคณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัยจัดตั้งขึ้นเมื่อปี พ.ศ. 2531 เดิมใช้ชื่อว่า 'ศูนย์ทดสอบ' ซึ่งอยู่ฝ่ายวิจัยของคณะครุศาสตร์ โดยมีวัตถุประสงค์เพื่อพัฒนาแบบสอบ และให้บริการทดสอบเพื่อคัดเลือกผู้เรียนเข้าศึกษาตามหลักสูตรของคณะครุศาสตร์ เป็นศูนย์กลางในการส่งเสริมการสร้างและพัฒนาแบบสอบมาตรฐานด้านสัมฤทธิผลของการศึกษาในวิชาและระดับต่างๆ ส่งเสริมพัฒนาและให้บริการแบบสอบสำหรับการคัดเลือกบุคคลเข้าทำงานหรือคัดเลือกบุคคล เพื่อเลื่อนตำแหน่งในวงงานต่างๆ นอกจากนี้ยังเป็นแหล่งศึกษาค้นคว้าแบบสอบสำหรับนิสิตนักศึกษาและผู้สนใจทั่วไป (จุฬาลงกรณ์มหาวิทยาลัย, 2535) ปัจจุบันศูนย์ทดสอบทางการศึกษากำหนดวัตถุประสงค์ดังนี้ (ศูนย์ทดสอบทางการศึกษา, 2538)

1. สร้างและพัฒนาแบบสอบทางการศึกษาและวิชาชีพ
2. บริการทดสอบทางการศึกษาและวิชาชีพ
3. วิจัยและเผยแพร่องค์ความรู้ทางการทดสอบและการประเมินผล
4. งานอื่นๆ ที่ได้รับมอบหมายโดยมุ่งให้ศูนย์ทดสอบทางการศึกษา สนับสนุนงานวิชาการด้านการทดสอบและประเมินผลการสอน

ภาระกิจหลักที่สำคัญอย่างหนึ่งของศูนย์ทดสอบทางการศึกษาคือการให้บริการทดสอบความสามารถพื้นฐานวิชาภาษาไทยและภาษาอังกฤษ ซึ่งเป็นแบบสอบร่วมของหลักสูตรบัณฑิตศึกษา คณะครุศาสตร์ โดยผู้ประสงค์จะสอบคัดเลือกเข้าศึกษาในคณะครุศาสตร์ทุกสาขาวิชา จะต้องผ่านการทดสอบวิชาร่วมภาษาไทยและภาษาอังกฤษของศูนย์ทดสอบ

ทางการศึกษา ให้ได้ตามเกณฑ์ที่แต่ละสาขากำหนดไว้ (จุฬาลงกรณ์มหาวิทยาลัย, 2538) แล้วจึงจะสามารถสมัครเข้าสอบในสาขาวิชาเฉพาะต่างๆ ต่อไป ทั้งนี้เพื่อให้ผู้สอบทดสอบความรู้ความสามารถที่พึงประสงค์และจำเป็นต่อการศึกษาต่อในระดับบัณฑิตเพื่อให้คะแนนมีความเป็นมาตรฐานและสามารถเปรียบเทียบกันได้ ตลอดจนสามารถทำนายได้ว่าผู้ที่สอบผ่านตามเกณฑ์ที่กำหนดจะสามารถเรียนในระดับบัณฑิตวิทยาลัยได้จนสำเร็จสูง (วรรณา ปูรณโชติ, 2538 : สัมภาษณ์)

จะเห็นได้ว่าแบบสอบวิชาร่วมภาษาไทยและภาษาอังกฤษดังกล่าว มีความสำคัญต่อการสอบคัดเลือกเข้าศึกษาต่อในระดับบัณฑิต คณะครุศาสตร์ ดังนั้นแบบสอบที่ใช้จึงต้องเป็นแบบสอบที่มีคุณภาพทั้งด้านความเที่ยง ความตรง และข้อสอบควรทำหน้าที่คัดเลือกผู้สอบได้อย่างยุติธรรม การศึกษาเรื่องการทำหน้าที่ต่างกันของข้อสอบ/แบบสอบ เป็นวิธีการวิเคราะห์คุณภาพของข้อสอบหรือแบบสอบที่สำคัญอย่างหนึ่งในปัจจุบัน

ตอนที่ 2 ความเป็นมาและวิธีการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ

การศึกษาผลการสอบของผู้สอบกลุ่มต่าง ๆ ของประชากรมีมานานแล้ว แต่ประเด็นเรื่องความยุติธรรมในการสอบระหว่างผู้สอบกลุ่มต่าง ๆ เริ่มได้รับการศึกษาอย่างจริงจังในช่วงปลายทศวรรษ 1960 โดยมีการเสนอวิธีการต่าง ๆ เพื่อตรวจสอบความลำเอียงของแบบสอบ (Test Bias) หรือความลำเอียงในการคัดเลือกผู้สอบ (Selection-Bias) ในขณะเดียวกันผู้สร้างแบบสอบก็สนใจวิธีการจำแนกข้อสอบที่ไม่เหมาะสมในแบบสอบ เพื่อนำข้อสอบไปปรับปรุงหรือคัดข้อสอบออกจากแบบสอบ เป็นการขจัดข้อสอบที่ทำให้เกิดปัญหาเรื่องความไม่ยุติธรรมระหว่างกลุ่มประชากรย่อยที่ต่างกัน ก่อนที่จะพัฒนาเป็นแบบสอบฉบับสมบูรณ์ต่อไป

ปัจจุบันนักวิจัยส่วนใหญ่ใช้คำว่า 'การทำหน้าที่ต่างกันของข้อสอบ' (Differential Item Functioning/DIF) แทนคำว่า ความลำเอียงของข้อสอบ/แบบสอบ (Item Bias) ความหมายของคำสองคำนี้ไม่เหมือนกัน แต่มีความเกี่ยวข้องกัน กล่าวคือ

Popham (1981) ให้ความหมายของความลำเอียงของข้อสอบว่าหมายถึง ความชอบหรือความโอนเอียงที่ทำให้การพิจารณาตัดสินเป็นไปอย่างไม่ยุติธรรม

Osterlind (1983) ให้ความหมายว่า ความลำเอียงเป็นความคลาดเคลื่อนอย่างเป็นระบบของการวัด

Holland และ Wainer (1993) กล่าวว่า การศึกษาในเรื่องของความยุติธรรมของข้อสอบหรือแบบสอบเป็นประเด็นหลักที่สำคัญ เดิมใช้คำว่า 'ความลำเอียงของข้อสอบ' (Item Bias) ซึ่งถือเป็นคำหนึ่งในการศึกษา หรือทดสอบความยุติธรรมของข้อสอบ ต่อมาในระยะหลังเกิดความคลุมเครือในการที่จะใช้เกณฑ์ในการตัดสินใจเรื่องความลำเอียง จึงใช้สารสนเทศทางสถิติมาเป็นเกณฑ์ในการตัดสินและใช้คำ 'การทำหน้าที่ต่างกันของข้อสอบ' (Differential

Item Functioning/DIF) ดังนั้น Holland และ Wainer จึงเสนอแนะว่าการใช้คำ 'การทำหน้าที่ต่างกันของข้อสอบ' หรือ DIF น่าจะเหมาะสมและมีความเป็นกลางมากกว่าการที่จะใช้คำว่า 'ความลำเอียง' หรือ 'Item Bias' สำหรับสารสนเทศทางสถิติที่ได้จากผลการตอบของผู้สอบต่างกลุ่มที่มีความสามารถเท่ากัน แต่มีโอกาสในการตอบข้อสอบได้ถูกต้องไม่เท่ากัน

Dorans และ Holland (1993) กล่าวว่า การทำหน้าที่ต่างกันของข้อสอบ หมายถึง ความแตกต่างในการทำหน้าที่ของข้อสอบหลังจากกลุ่มของผู้สอบได้ถูกจับคู่ตามความสามารถหรือคุณลักษณะที่ข้อสอบนั้นวัด

Millsap และ Everson (1993) กล่าวว่า การทำหน้าที่ต่างกันของข้อสอบ หมายถึง ความแตกต่างในการทำหน้าที่ของแบบสอบหรือข้อสอบ ระหว่างกลุ่มผู้สอบซึ่งถูกจับคู่ตามคุณลักษณะที่วัดโดยแบบสอบหรือข้อสอบนั้น

Shepard และ Camilli (1994) ให้ข้อสรุปเกี่ยวกับ DIF ตามหลักในการวิเคราะห์ของ DIF นั้นมิได้ตรวจสอบหาความลำเอียงโดยตรง แต่เป็นการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างกลุ่มผู้สอบตั้งแต่ 2 กลุ่มขึ้นไปที่มีความสามารถหลัก (primary abilities) ที่มุ่งวัดเท่ากัน สถิติ DIF ได้มาจากการตรวจสอบความเป็นพหุมิติของข้อสอบโดยจะแสดงว่าข้อสอบมีการทำหน้าที่ต่างกันระหว่างกลุ่ม เมื่อมีการแจกแจงของความสามารถรอง (secondary abilities) แตกต่างกันในกลุ่มผู้สอบที่มีความสามารถหลักเท่ากัน

กล่าวโดยสรุป การทำหน้าที่ต่างกันของข้อสอบ(DIF) เป็นกรณีที่เกิดขึ้นเมื่อข้อสอบหรือแบบสอบ ทำให้ผลการตอบของผู้สอบที่มีความสามารถหลักที่ต้องการวัดหรือคุณลักษณะแฝงที่เป็นเป้าหมายของการวัดเท่ากันแต่มีคุณลักษณะแฝงอื่นต่างกัน และเมื่อทำข้อสอบนั้นๆ แล้วข้อสอบนั้นทำให้ผู้ที่มาจากต่างกลุ่มกัน มีโอกาสในการตอบข้อสอบได้ถูกต้องไม่เท่ากัน

สำหรับวิธีการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ (Analysis DIF) หรือที่นักวิจัยบางท่านเรียกว่า การตรวจค้นการทำหน้าที่ต่างกันของข้อสอบ (Detect DIF) ในปัจจุบันมีอยู่หลายวิธี Dorans and Potenza (1995) ได้จำแนกประเภทของวิธีการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในลักษณะที่มีการให้คะแนนแบบ 2 ค่า (dichotomously score : โดยการให้คะแนนเป็น 1 เมื่อตอบถูก และให้คะแนนเป็น 0 เมื่อตอบผิด) ออกเป็น 2 กลุ่ม คือ

กลุ่มที่ 1 เป็นกลุ่มที่ใช้คะแนนที่สังเกตไม่ได้หรือตัวแปรแฝง (Latent Variable) ได้แก่วิธีที่มีพื้นฐานจากทฤษฎีการตอบสนองข้อสอบ (IRT) และวิธี SIBTEST (Shealy and Stout, 1993)

กลุ่มที่ 2 เป็นกลุ่มที่ใช้คะแนนที่สังเกตได้ (Observe Score) ได้แก่วิธี Mantel-Haenszel หรือวิธี MH (Holland and Thayer, 1988) วิธี Standardization (Dorans and Holland, 1993) และวิธี Logistic Regression (Swaminatan and Rojer,1990)

วิธีที่ใช้ทฤษฎีการตอบสนองข้อสอบ (IRT) เป็นวิธีที่ได้รับการพัฒนาและแตกย่อยออกไปหลายวิธี เช่น วิธี General IRTLR วิธี Loglinear IRTLR วิธี IRT-D^2 และวิธี Lord 's χ^2 กระบวนการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในกลุ่มนี้ใช้สถิติพารามิเตอร์ (Parameter) และสามารถใช้ได้กับตัวแปรแฝงหรือคะแนนที่สังเกตไม่ได้ (Latent Score) วิธีการในกลุ่มนี้ใช้ได้ดีถ้าคะแนนของข้อสอบ/แบบสอบเป็นไปตามโมเดล IRT แต่กระบวนการวิเคราะห์ DIF ค่อนข้างยุ่งยากซับซ้อน เสียค่าใช้จ่ายสูง กลุ่มตัวอย่างต้องมีขนาดใหญ่ และจะเกิดปัญหาในการนำไปใช้มากเพราะข้อมูลจริงมักจะไม่เป็นไปตามโมเดล IRT ซึ่งทำให้การประมาณค่าของพารามิเตอร์ต่าง ๆ ทำได้ไม่ดี (Sheapard and Camilli, 1994)

วิธี SIBTEST เป็นวิธีการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ/แบบสอบชนิดพหุมิติ (multidimentional) เป็นวิธีการทดสอบทางสถิติที่นันพารามิเตอร์ (non parameter) สามารถคำนวนได้ง่ายไม่ซับซ้อน ใช้ได้ดีกับการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป (unifrom DIF) และมีทิศทางเดียว (unidirectional) ข้อจำกัดของวิธีนี้คือ กลุ่มตัวอย่างต้องมีขนาดใหญ่พอและเสียค่าใช้จ่ายค่อนข้างสูง (Shealy and Stout, 1993)

วิธี Logistic Regression (LRDIF) เป็นวิธีการทดสอบสถิติแบบพารามิเตอร์ มีการทดสอบโมเดลทางสถิติ Logistic Regression วิธีนี้สามารถใช้วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบได้ดีทั้งแบบเอกรูป (unifrom DIF) และแบบอเนกรูป (nonunifrom DIF) แต่จะเสียค่าใช้จ่ายมากกว่าวิธี MH ประมาณ 3-4 เท่า (Swaminatan and Rojer, 1990)

วิธี Standardization (STND) และ วิธี Mantel-Haenszel (MH) ทั้ง 2 วิธีนี้ เป็นการทดสอบสถิติแบบนันพารามิเตอร์ (non parameter) มีหลักการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบคล้ายกัน ผลการทดสอบทางสถิติในสถานการณ์จำลองระหว่างสองวิธีนี้ มีความสัมพันธ์กัน และได้ผลการวิเคราะห์ DIF คล้ายกัน (Dorans and Holland, 1993 ; Millsap and Everson, 1993)

การวิจัยครั้งนี้ผู้วิจัยเห็นสมควรที่จะนำวิธีการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบตามหลักของแมนเทล-เฮนส์เซล (MH) มาใช้กับข้อมูลเชิงประจักษ์ ซึ่งจะได้กล่าวถึงรายละเอียดในตอนต่อไป

ตอนที่ 3 วิธีการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบตามหลักของแมนเทล - เฮนส์เซล

การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบทำได้หลายวิธี Holland and Thayer (1988) ; Mazor และคณะ (1994) กล่าวว่าวิธี MH เป็นวิธีที่ใช้ง่าย สะดวก และประหยัดได้พัฒนาขึ้นใช้ตั้งแต่ ปี 1959 วิธี MH เป็นวิธีการที่พัฒนามาจากวิธีไค-สแควร์ หลักการของวิธี MH เป็นการเปรียบเทียบผลการสอบของผู้สอบ 2 กลุ่ม หรือที่ Holland เรียกว่า กลุ่มอ้างอิงและกลุ่มเปรียบเทียบ (reference and focal groups) กลุ่มที่ใช้อ้างอิงคือกลุ่มที่คาดว่าจะได้ประโยชน์จากข้อสอบ และกลุ่มเปรียบเทียบเป็นกลุ่มที่เสียประโยชน์จากข้อสอบที่ทำหน้าที่ต่างกัน ซึ่งจะมีการตรวจสอบทุกๆ ระดับคะแนนรวมจากแบบสอบ ข้อสอบใดที่ผู้สอบในกลุ่มที่มีความสามารถเท่ากัน ทำได้ถูกต้องเท่ากันถือว่าเป็นข้อสอบที่ทำหน้าที่ไม่ต่างกัน (no DIF) ระหว่างกลุ่มผู้สอบ

การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธี MH แยกวิเคราะห์ข้อสอบเป็นรายข้อ โดยที่การวิเคราะห์แต่ละข้อจะต้องสร้างตารางไขวัขนาด 2x2 แสดงความถี่ของผู้สอบที่ตอบถูก/ผิด ในกลุ่มอ้างอิง(R)และกลุ่มเปรียบเทียบ(F) ตามช่วงคะแนนของผู้สอบสำหรับช่วงคะแนน j ดังตารางที่ 1 ถ้ามีการแบ่งช่วงคะแนนการสอบเป็น k ช่วง จะได้ตารางไขวัขนาด 2x2 รวม k ตาราง

ตารางที่ 1 แสดงความถี่ของผู้สอบที่ตอบถูกและผิดในกลุ่มอ้างอิงและกลุ่มเปรียบเทียบสำหรับช่วงคะแนน j

กลุ่มผู้สอบ	คะแนนที่ได้จากข้อสอบที่ต้องการวิเคราะห์DIF		
	ตอบถูก (1)	ตอบผิด (0)	รวม
R	A_j	B_j	N_{Rj}
F	C_j	D_j	N_{Fj}
รวม	m_{1j}	m_{0j}	N_j

เมื่อ A_j = เป็นความถี่ที่สังเกตได้ในการตอบถูกในช่วงคะแนน j ของกลุ่ม R

B_j = เป็นความถี่ที่สังเกตได้ในการตอบผิดในช่วงคะแนน j ของกลุ่ม R

C_j = เป็นความถี่ที่สังเกตได้ในการตอบถูกในช่วงคะแนน j ของกลุ่ม F

D_j = เป็นความถี่ที่สังเกตได้ในการตอบผิดในช่วงคะแนน j ของกลุ่ม F

m_{1j} = จำนวนผู้สอบที่ตอบถูกทั้งหมดในช่วงคะแนน j

m_{0j} = จำนวนผู้สอบที่ตอบผิดทั้งหมดในช่วงคะแนน j

N_j = จำนวนผู้สอบทั้งหมด

จากนั้นแสดงเป็นสัดส่วนการตอบข้อสอบของกลุ่มประชากร 2 กลุ่มได้ดังตารางที่ 2

ตารางที่ 2 สัดส่วนการตอบข้อสอบของกลุ่มประชากร 2 กลุ่ม

ผลการตอบของกลุ่ม	คะแนน		
	1	0	รวม
R	P_{Rj}	Q_{Rj}	1
F	P_{Fj}	Q_{Fj}	1

โดยที่ P_{Rj} คือ สัดส่วนของกลุ่มอ้างอิงที่อยู่ในช่วงความสามารถ j ที่ตอบข้อสอบถูก

P_{Fj} คือ สัดส่วนของกลุ่มเปรียบเทียบที่อยู่ในช่วงความสามารถ j ที่ตอบข้อสอบถูก

Q_{Rj} คือ $1 - P_{Rj}$

Q_{Fj} คือ $1 - P_{Fj}$

หลักการวิเคราะห์ตามวิธี MH เป็นการนำข้อมูลจากตารางไขว้ k ตารางมาดำเนินการตามขั้นตอนดังนี้

1. คำนวนค่าความน่าจะเป็นในรูปของสัดส่วนของการตอบข้อสอบถูกและผิดระหว่างกลุ่ม ในทุกช่วงคะแนน จากสูตร

$$\alpha MH = \frac{\sum A_j D_j / N_j}{\sum B_j C_j / N_j}$$

เมื่อ αMH = สัดส่วนของการตอบข้อสอบถูกและผิดระหว่างกลุ่มในแต่ละข้อในทุกช่วงคะแนน j

A_j = เป็นความถี่ที่สังเกตได้ในการตอบถูกในช่วงคะแนน j ของกลุ่ม R

B_j = เป็นความถี่ที่สังเกตได้ในการตอบผิดในช่วงคะแนน j ของกลุ่ม R

C_j = เป็นความถี่ที่สังเกตได้ในการตอบถูกในช่วงคะแนน j ของกลุ่ม F

D_j = เป็นความถี่ที่สังเกตได้ในการตอบผิดในช่วงคะแนน j ของกลุ่ม F

N_j = จำนวนผู้สอบทั้งหมด

2. ทดสอบนัยสำคัญของค่าสถิติไค-สแควร์ เพื่อทดสอบค่า αMH ที่คำนวนได้ว่าแตกต่างจาก 1 อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05 หรือไม่ ที่ระดับชั้นความเป็นอิสระเท่ากับ 1 ตามสูตร ดังนี้

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

เมื่อ $E(A_j) = (N_{Rj})\,(m_{1j}) / N_j$

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

และสมมุติฐานศูนย์แสดงว่าข้อสอบที่ไม่ DIF ได้แก่

$H_0 : P_{Rj} = P_{Fj}$ สำหรับทุกชั้นคะแนน j

สมมุติฐานศูนย์นี้เป็นสมมุติฐานของความเป็นอิสระอย่างมีเงื่อนไขของสมาชิกกลุ่มและคะแนนที่ได้จากการตอบข้อสอบที่ต้องการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ และภายใต้สมมุติฐานศูนย์จะได้ค่าคาดหวังของการตอบ ดังนี้

$$E(A_j) = (N_{Rj}\ m_{1j}) / N_j$$
$$E(B_j) = (N_{Rj}\ m_{0j}) / N_j$$
$$E(C_j) = (N_{Fj}\ m_{1j}) / N_j$$
$$E(D_j) = (N_{Fj}\ m_{0j}) / N_j$$

สมมุติฐานอื่นของแมนเทลและเฮนส์เซล ได้แก่

$H_1 : \frac{P_{Rj}}{q_{Rj}} = \alpha \frac{P_{Fj}}{q_{Fj}}$ $\quad j = 1, \ldots, K$ เมื่อ α ไม่เท่ากับ 1

เมื่อ α มีค่า = 1 ซึ่งสอดคล้องกับสมมุติฐานศูนย์ จะได้ว่า

$$H_0 : \frac{P_{Rj}}{q_{Rj}} = \frac{P_{Fj}}{q_{Fj}}$$

และค่าประมาณของ α MH $= \frac{P_{Rj}}{q_{Rj}} \cdot \frac{P_{Fj}}{q_{Fj}} = \frac{P_{Rj}\, q_{Fj}}{P_{Fj}\, q_{Rj}}$ สำหรับทุก j=1, . . ,K

3. Holland และ Thayer เสนอแนะให้แปลงค่า α MH ให้เป็นค่าเดลต้า (Δ MH หรือ MH_{DIF}) ตามสูตร

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

สำหรับเกณฑ์ในการตัดสินข้อสอบที่ทำหน้าที่ต่างกัน คือ ข้อสอบที่มีค่า α MH แตกต่าง จาก 1 อย่างมีนัยสำคัญทางสถิติ หรือค่า MH_{DIF} แตกต่างจาก 0 อย่างมีนัยสำคัญทางสถิติ และมีเกณฑ์ในการพิจารณาค่าของ MH_{DIF} ดังนี้

1). ค่า MH_{DIF} เท่ากับ 0 หรือไม่แตกต่างจาก 0 อย่างมีนัยสำคัญทางสถิติแสดงว่าข้อสอบนั้นทำหน้าที่ไม่แตกต่างกันระหว่างกลุ่ม (no DIF)

2). ค่า MH_{DIF} แตกต่างจาก 0 อย่างมีนัยสำคัญทางสถิติ และมีค่าเป็นบวก (positive) แสดงว่าข้อสอบนั้นทำหน้าที่แตกต่างกันระหว่างกลุ่ม โดยจะเข้าข้าง (favor) กลุ่มเปรียบเทียบ (F)

3). ค่า MH_{DIF} แตกต่างจาก 0 อย่างมีนัยสำคัญทางสถิติ และมีค่าเป็นลบ (negative)แสดงว่าข้อสอบนั้นทำหน้าที่แตกต่างกันระหว่างกลุ่ม โดยจะเข้าข้าง (favor) กลุ่มอ้างอิง (R)

การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยวิธี MH นี้ เป็นการเปรียบเทียบผลการตอบข้อสอบของผู้สอบสองกลุ่ม หลังจากการจับคู่ผู้สอบตามความรู้ หรือความสามารถของผู้สอบการจับคู่ผู้สอบของผู้สอบสองกลุ่มเป็นเงื่อนไขที่สำคัญเพราะเป็นเกณฑ์ที่ใช้แทนความสามารถที่แท้จริงของผู้สอบสองกลุ่ม ในทางปฏิบัติมักใช้คะแนนรวมของแบบสอบ (Total Test Score) เป็นเกณฑ์การจับคู่ เพราะเกี่ยวข้องโดยตรงกับความรู้หรือความสามารถที่วัดได้โดยแบบสอบนั้นสามารถตรวจสอบความตรงและความเที่ยงของแบบสอบได้ และผู้สอบทุกคนสอบภายใต้สถานการณ์เดียวกัน แต่จุดอ่อนของการใช้คะแนนรวมของแบบสอบเป็นเกณฑ์การ

จับคู่ผู้สอบก็คือ มีการรวมเอาคะแนนจากข้อสอบที่ทำหน้าที่ต่างกันมาเป็นเกณฑ์ในการจับคู่ผู้สอบด้วย ในการแก้ไขจุดอ่อนนี้ Holland และ Thayer (1988) ได้เสนอให้ใช้วิธีการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบแบบ 2 ขั้นตอน ดังนี้

ขั้นตอนที่ 1 ใช้คะแนนรวมของแบบสอบทั้งฉบับเป็นเกณฑ์การจับคู่ผู้สอบ แล้ววิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ

ขั้นตอนที่ 2 นำคะแนนของข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบในขั้นตอนที่ 1 ออกจากเกณฑ์การจับคู่ผู้สอบแล้วใช้คะแนนรวมของข้อสอบที่ทำหน้าที่ไม่แตกต่างกัน(no DIF) หรือข้อสอบที่เหลือ เป็นเกณฑ์การจับคู่ผู้สอบในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบของแบบสอบทั้งฉบับในขั้นตอนที่สอง

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้เทคนิค 2 ขั้นตอนนี้ เรียกว่า การทำให้เกณฑ์การจับคู่ผู้สอบมีความบริสุทธิ์ (purification of matching criterion)

ในระยะเวลาที่ผ่านมามีนักวิจัยได้นำหลักการของแมนเทลและเฮนส์เซล (MH) มาใช้ศึกษาเกี่ยวกับการวิเคราะห์การการทำหน้าที่ต่างกันของข้อสอบหรือความลำเอียงของข้อสอบหลายท่านตัวอย่างเช่น Tissen Steinberg และ Wainer (1988) พบว่าวิธี MH ให้ผลการวิเคราะห์ DIF คล้าย IRT-LR และจะเป็นประโยชน์ในการประเมินแบบสอบยาว ๆ และอาจใช้เป็นเครื่องมือกลั่นกรองข้อสอบก่อนที่จะใช้ IRT-LR ต่อไป Hambleton และคณะ (1989) พบว่าวิธี MH เป็นวิธีที่สามารถเลือกใช้แทนวิธีทฤษฎีการตอบสนองข้อสอบได้อย่างประหยัดทั้งเวลาและค่าใช้จ่าย โดยวิธี MH มีความสอดคล้องของการตรวจสอบเท่ากับ 80 % ในขณะที่ IRT เท่ากับ 73 % สำหรับ Swaminatan และ Rojer (1990) พบว่าวิธี MH วิเคราะห์ได้ดีกว่า LR-DIF เล็กน้อยโดยตรวจค้นได้ถูกต้องร้อยละ 75 ในกลุ่มตัวอย่างขนาด 250 คน และร้อยละ 100 ในกลุ่มตัวอย่าง 500 คน สำหรับการวิเคราะห์ DIF แบบเอกรูป ส่วนการวิเคราะห์ DIF แบบอเนกรูปนั้น MH ทำได้ไม่ดีนักแต่วิธี LR-DIF จะเสียค่าใช้จ่ายมากกว่าวิธี MH ประมาณ 3-4 เท่า ต่อมา Clauser และคณะ (1991) ใช้ MH วิเคราะห์ DIF แบบเอกรูป พบว่า วิธี MH ใช้ได้ดีในกรณีที่ข้อสอบมีค่าอำนาจจำแนกสูงแต่กรณีที่ข้อสอบมีค่าความยากสูงมากจะไม่สามารถตรวจค้นได้ นอกจากนี้ Mazor และคณะ (1994) พบว่า วิธี MH สามารถวิเคราะห์ DIF ได้ดีทั้งแบบเอกรูปและแบบอเนกรูป

การวิจัยครั้งนี้ผู้วิจัยเลือกใช้วิธี MH มาทำการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบกับข้อมูลจริงคือแบบสอบวิชาภาษาไทยและภาษาอังกฤษของศูนย์ทดสอบทางการศึกษา ทั้งนี้จะศึกษาการวิเคราะห์ DIF ทั้งแบบเอกรูปและแบบอเนกรูป โดยใช้เทคนิคการวิเคราะห์แบบ 2 ขั้นตอนเนื่องจาก Clauser และคณะ (1993) พบว่าผลการตรวจค้น DIF จากสถานการณ์จำลอง โดยใช้เทคนิคแบบ 2 ขั้นตอน พบข้อสอบที่ DIF เท่ากับหรือเหนือกว่า

เทคนิคแบบ 1 ขั้นตอน ในทุกเงื่อนไขและเทคนิคแบบ 2 ขั้นตอน ไม่เพิ่มความคลาดเคลื่อนชนิดที่ 1 (Type 1 error rate) มากกว่าเทคนิค 1 ขั้นตอน

สำหรับเกณฑ์ในการจับคู่ผู้สอบ ผู้วิจัยใช้คะแนนรวมที่ผู้สอบทำได้เป็นเกณฑ์และใช้จำนวนชั้นคะแนน (Score categores) ที่ผู้สอบทำได้สูงสุดเป็นเกณฑ์ในการจับคู่ผู้สอบ ตามที่ Clauser และคณะ (1994) เสนอแนะว่าการลดจำนวนชั้นคะแนนถึงแม้จะเพิ่มอำนาจการทดสอบ (Power of test) แต่ก็จะเพิ่ม Type 1 error rate ด้วย ซึ่งนักวิจัยควรระมัดระวังไม่ให้เกิดการเพิ่ม Type 1 error rate ดังนั้นการใช้จำนวนชั้นคะแนนที่เป็นไปได้มากที่สุดจึงควรจะนำไปเลือกใช้ในทางปฏิบัติ

ในการศึกษาครั้งนี้ผู้วิจัยใช้โปรแกรม MH_{DIF} ที่พัฒนาโดย Fidalgo A. (1995) ทำการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบตามวิธีของแมนเทล-เฮนส์เซล

ตอนที่ 4 งานวิจัยที่เกี่ยวข้องกับการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ

4.1 งานวิจัยในประเทศ

ในช่วงระยะเวลาที่ผ่านมายังไม่มีงานวิจัยที่ศึกษาการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยตรง แต่มีการศึกษาเกี่ยวกับการวิเคราะห์ความลำเอียงของข้อสอบ สำหรับงานวิจัยในประเทศไทยที่เกี่ยวข้องกับการวิเคราะห์ความลำเอียงของข้อสอบมีดังนี้

ชัชชัย เผ่าพงศ์ (2527) ศึกษาความลำเอียงของข้อสอบด้วยวิธีโค้งลักษณะข้อสอบ 3 พารามิเตอร์ ตัวแปรที่ศึกษาคือเพศ แบบทดสอบที่ใช้คือแบบทดสอบวัดความถนัดทางการเรียนด้านคณิตศาสตร์และภาษาในระดับมัธยมต้น ซึ่งแบบทดสอบฉบับนี้พัฒนาโดยสำนักทดสอบทางการศึกษาและจิตวิทยา มหาวิทยาลัยศรีนครินทรวิโรฒประสานมิตร กลุ่มตัวอย่างที่ใช้เป็นนักเรียนชายและหญิงชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2524 ทั่วทุกภาคของประเทศ (ภาคเหนือ ภาคกลาง ภาคใต้ ภาคตะวันออกเฉียงเหนือ) กลุ่มที่ทำการวิเคราะห์แบบทดสอบด้านคณิตศาสตร์ เป็นนักเรียนชายจำนวน 1,610 คน นักเรียนหญิง 1,337 คน อีกกลุ่มหนึ่งทำการวิเคราะห์แบบทดสอบด้านภาษาเป็นนักเรียนชาย 1,316 คน นักเรียนหญิง 985 คน ผลการศึกษา พบว่า แบบทดสอบด้านคณิตศาสตร์ จำนวน 30 ข้อ มีความลำเอียงต่อกลุ่มใดกลุ่มหนึ่ง โดยเฉพาะ 9 ข้อ คือลำเอียงต่อกลุ่มนักเรียน 7 ข้อ และลำเอียงต่อกลุ่มนักเรียนหญิง 2 ข้อ ข้อสอบที่ลำเอียงต่อ กลุ่มนักเรียนชายและหญิงในระดับปานกลางขึ้นไปมีจำนวน 5 ข้อ ซึ่งวัดในเรื่องร้อยละ การหาปริมาตร และการหาความยาวของด้านรูปสามเหลี่ยม จำนวนเรื่องละ 1 ข้อ อีกจำนวน 2 ข้อ เป็นเรื่องเกี่ยวกับโจทย์ปัญหา และจากข้อสอบจำนวน 5 ข้อนี้เป็นข้อสอบที่มีความลำเอียงต่อกลุ่มนักเรียนหญิงในช่วงความสามารถแรก และลำเอียงต่อกลุ่มนักเรียนชายในช่วงความสามารถต่อมาจำนวน 1 ข้อ ส่วนแบบทดสอบด้านภาษาเกี่ยวกับความเข้าใจในการอ่านจำนวน 40 ข้อ พบว่ามีความลำเอียงต่อ

กลุ่มนักเรียนชายโดยเฉพาะ 3 ข้อ และลำเอียงต่อกลุ่มนักเรียนหญิงโดยเฉพาะมี 8 ข้อ ข้อสอบที่มีความลำเอียงในระดับปานกลางขึ้นไปมีจำนวน 9 ข้อ ซึ่งวัดความเข้าใจเกี่ยวกับการอ่านคำประพันธ์ บทร้อยกรองอย่างละ 1 ข้อ และวัดความเข้าใจเกี่ยวกับการอ่านข้อความจำนวน 7 ข้อ และจากข้อสอบจำนวน 9 ข้อนี้ เป็นข้อที่มีความลำเอียงต่อกลุ่มนักเรียนชายโดยเฉพาะ 1 ข้อ เป็นข้อที่มีความลำเอียงต่อกลุ่มนักเรียนหญิงโดยเฉพาะ 6 ข้อ และมีข้อสอบ 2 ข้อ ที่มีความลำเอียงต่อกลุ่มนักเรียนชายในช่วงความสามารถแรกช่วงความสามารถต่อมาลำเอียงต่อกลุ่มนักเรียนหญิง

ทัศนีย์ พีรมนตรี (2530) ศึกษาวิเคราะห์ความลำเอียงของข้อสอบ 3 วิธี คือ วิธีกำหนดจุดค่าเคลดา วิธีทดสอบความแตกต่างระหว่างกลุ่มด้วยสถิติไค-สแควร์ ในโมเดลลอกลิเนียร์ คือ โมเดลที่ไม่มีพารามิเตอร์ผลร่วมระหว่างระดับคะแนนกับกลุ่มและโมเดลที่ไม่มีพารามิเตอร์ของผลหลักที่เกิดจากกลุ่ม และวิธีโค้งลักษณะข้อสอบ 3 พารามิเตอร์ แบบทดสอบที่ใช้ศึกษาคือ แบบทดสอบวิชาคณิตศาสตร์โครงการตรวจสอบคุณภาพการศึกษาของสำนักงานทดสอบทางการศึกษา กรมวิชาการ กระทรวงศึกษาธิการ กลุ่มตัวอย่างเป็นนักเรียนชั้นมัธยมศึกษาปีที่ 6 ปีการศึกษา 2526 ทั่วทุกภาคของประเทศ และกรุงเทพมหานคร จำนวนทั้งหมด 7,036 คน ผู้วิจัยเปรียบเทียบจำนวนข้อสอบที่มีความลำเอียงระหว่างนักเรียนกรุงเทพมหานครกับกลุ่มนักเรียนภาค ภูมิศาสตร์ทั้ง 5 ภาค ดังกล่าว

ผลการศึกษา พบว่า เมื่อแยกข้อสอบออกตามค่าความยากที่วิเคราะห์ด้วยวิธีดั้งเดิม พบว่า ข้อที่ลำเอียงมีจำนวน 43 ข้อ และไม่ลำเอียง 17 ข้อ เหมือนกันทุกคู่ทุกภาค และเมื่อเปรียบเทียบจำนวนข้อที่ลำเอียงระหว่ากลุ่มนักเรียนในกรุงเทพมหานครและกลุ่มนักเรียนภาคอื่น ๆ พบว่า ในการใช้ IRT-3 พารามิเตอร์ พบข้อกระทงที่มีความลำเอียงจำนวนมากที่สุด มีข้อกระทงที่ลำเอียงซ้ำกันระหว่ากลุ่มนักเรียนในกรุงเทพมหานครกับทุก ๆ ภาค แต่จำนวนไม่เท่ากันในแต่ละวิธี วิธีที่ 1 และ 3 มีข้อลำเอียงซ้ำกันมากสุดระหว่างกรุงเทพมหานครและภาคตะวันออกเฉียงเหนือ ส่วนวิธีที่ 2 มีจำนวนข้อที่ลำเอียงซ้ำกันมากที่สุดระหว่างกรุงเทพมหานครและภาคตะวันออก

การเปรียบเทียบผลการวิเคราะห์ทั้ง 3 วิธี กับความลำเอียงที่เกิดขึ้นในภาคเดียวกัน พบข้อกระทงที่ลำเอียงซ้ำกัน ข้อกระทงที่ลำเอียงส่วนใหญ่เป็นข้อที่ง่ายสำหรับนักเรียนในกรุงเทพมหานครมากกว่ากลุ่มนักเรียนในภาคอื่น ๆ สำหรับวิธีวิเคราะห์ที่ 1 แต่เป็นข้อที่มีความลำเอียงในเกณฑ์ต่ำสำหรับวิธีที่ 2 และเป็นข้อที่ลำเอียงอย่างสม่ำเสมอในการวิเคราะห์ด้วยวิธีที่ 3

สุรศักดิ์ อมรรัตนศักดิ์ (2531) ได้ทำการศึกษาเปรียบเทียบผลของวิธีวิเคราะห์หาความลำเอียงของข้อสอบ 4 วิธี คือ 1) วิธีวิเคราะห์ความแปรปรวน 2) วิธีแปลงค่าความยากง่ายของข้อสอบ 3) วิธีโค้งลักษณะของข้อทดสอบที่มีพารามิเตอร์ 1 ตัว และ 4) วิธีโค้งลักษณะของข้อสอบที่มีพารามิเตอร์ 3 ตัว เมื่อใช้วิเคราะห์หาความลำเอียงต่อเพศของข้อสอบ ที่ใช้สอบแข่งขันเพื่อบรรจุเป็นข้าราชการ 4 ฉบับ และใช้ผลวิจัยเพศชาย 1,170 คน และเพศหญิงอีก 1,170 คน ผู้วิจัยวิเคราะห์หาดัชนีความลำเอียงของข้อสอบแล้วหาสัมประสิทธิ์สหสัมพันธ์ระหว่างวิธีการวิเคราะห์ทั้ง 4 วิธี และเปรียบเทียบความแตกต่างของผลการคัดเลือกก่อนและหลังการศึกษาความลำเอียงของข้อสอบตามวิธีการคิดคะแนนรวมที่แตกต่างกัน 6 วิธี ในด้านจำนวนผู้ที่ได้รับการคัดเลือกสัดส่วนของชายต่อหญิงที่ได้รับการคัดเลือกและความเที่ยงของแบบสอบ ผลการวิจัยพบว่าวิธีวิเคราะห์หาความลำเอียงแต่ละวิธีพบข้อทดสอบที่มีความลำเอียงต่างกัน โดยวิธีโค้งลักษณะของข้อทดสอบที่มีพารามิเตอร์ 3 ตัว พบข้อทดสอบที่มีความลำเอียงจำนวนมากที่สุด รองลงมา ได้แก่ วิธีวิเคราะห์ความแปรปรวน ส่วนวิธีแปลงค่าความยากง่ายของข้อทดสอบพบข้อสอบที่มีความลำเอียงจำนวนน้อยที่สุดและยังพบว่าค่าสัมประสิทธิ์สหสัมพันธ์ของดัชนีที่บ่งบอกความลำเอียงของข้อทดสอบของทั้ง 4 วิธี สูงมากคือมีค่า r_{xy} ระหว่าง 0.754 - 0.992 สำหรับการใช้คะแนนดิบและคะแนนรวมแบบต่าง ๆ กัน 5 วิธี มีจำนวนผู้ได้รับการคัดเลือกแตกต่างกันประมาณร้อยละ 4-24 ส่วนการให้คะแนนมาตรฐานทีปกติ รวมกับคะแนนแปลงแบบอื่น ๆ 4 วิธี มีจำนวนผู้ได้รับการคัดเลือกแตกต่างกันร้อยละ 4-23 และเมื่อตัดข้อสอบที่มีความลำเอียงออกแล้ว พบว่า สัดส่วนหญิงและชายที่ได้รับการคัดเลือกมีความใกล้เคียงกัน และค่าความเที่ยงของแบบสอบลดลงเล็กน้อย

พัชรี ปิยภัณฑ์ (2531) ได้ทำการวิเคราะห์ความลำเอียงของข้อสอบด้วยวิธีแปลงค่าความยาก วิธีไค-สแควร์ และวิธีโค้งลักษณะข้อสอบ 1 พารามิเตอร์ จากแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนวิชาคณิตศาสตร์ จำนวน 45 ข้อ ซึ่งผู้วิจัยสร้างขึ้นเอง ตัวแปรที่ศึกษา คือ เพศและภาคภูมิศาสตร์ กลุ่มตัวอย่างที่ใช้ในการศึกษาเป็นนักเรียนชายหญิงชั้นประถมศึกษาปีที่ 6 ปีการศึกษา 2530 ของโรงเรียนสังกัด สำนักงานคณะกรรมการการประถมศึกษาแห่งชาติ ในเขตกรุงเทพมหานคร และจังหวัดสมุทรสาคร จำนวน 1,948 คน ผู้วิจัยวิเคราะห์หาดัชนีความลำเอียงของข้อสอบแล้วหาสัมประสิทธิ์สหสัมพันธ์ของวิธีวิเคราะห์ 3 วิธีและศึกษาเปรียบเทียบค่าความเที่ยงของแบบทดสอบก่อน และหลังการคัดเลือกข้อสอบที่ลำเอียงออก ผลการวิจัย พบว่า การวิเคราะห์ความลำเอียงของข้อสอบวิธีโค้งลักษณะข้อสอบพบจำนวนข้อสอบที่ลำเอียงมากที่สุด รองลงมา คือ วิธีไค-สแควร์ และวิธีแปลงค่าความยากพบจำนวนข้อสอบที่ลำเอียงน้อยที่สุด ค่าสัมประสิทธิ์สหสัมพันธ์ของดัชนีความลำเอียงต่อเพศระหว่างวิธีวิเคราะห์ความลำเอียง 3 วิธี มีค่าระหว่าง .1713 ถึง .5618 และค่าสัมประสิทธิ์สหสัมพันธ์ของดัชนีความลำเอียงต่อกลุ่มนักเรียนกลุ่มนักเรียนกรุงเทพมหานครกับกลุ่ม

นักเรียนสมุทรสาครมีค่าระหว่าง .1868 ถึง .6009 ส่วนค่าความเที่ยงของแบบสอบก่อนและหลังคัดข้อสอบที่ลำเอียงแล้วแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

สุพัฒน์ สุกมลสันต์ (2534) ได้วิเคราะห์ความลำเอียงของข้อสอบภาษาอังกฤษเพื่อคัดเลือกเข้าศึกษาในมหาวิทยาลัยโดยใช้ข้อมูลการตอบแบบสอบภาษาอังกฤษเข้ามหาวิทยาลัย ชุด กข และ ชุด กขค ปี 2531 - 2533 ซึ่งมีข้อสอบชุดละ 100 ข้อ ตัวแปรที่ศึกษา คือเพศ และภาคภูมิศาสตร์ของผู้สอน ซึ่งแยกตามภูมิลำเนาของข้อสอบ เทคนิควิธีที่นำมาวิเคราะห์ความลำเอียงมี 3 วิธีที่นำมาวิเคราะห์ความลำเอียงมี 3 วิธี ได้แก่ วิธีกำหนดจุดค่าเดลต้า (Delta Plot) วิธีไค-สแควร์ชนิดที่แบ่งความสามารถของผู้สอบเป็น 3 ระดับ ได้แก่กลุ่มความสามารถระดับต่ำ กลุ่มความสามารถระดับกลาง และกลุ่มความสามารถระดับสูง วิธีที่ 3 ที่นำมาใช้ คือ วิธีการวัดพื้นที่ความแตกต่างระหว่างโค้งลักษณะข้อสอบที่วิเคราะห์ตามทฤษฎีการตอบข้อสอบแบบ 3 พารามิเตอร์ ประชากรของการวิจัยได้แก่ผู้สอบแบบสอบดังกล่าวจำนวน 6 กลุ่มๆ ละ ประมาณ 30,000 - 80,000 คน และแต่ละกลุ่มแบ่งตามเพศและภาคภูมิศาสตร์ของผู้สอบ ผลการวิจัยได้จากการสุ่มอย่างง่ายจากประชากรแต่ละกลุ่ม โดยกำหนดให้ได้กลุ่มละ 424 - 3,000 คน

ผลการวิเคราะห์ พบว่า แบบสอบภาษาอังกฤษ ฉบับ กข และ กขค ปี 2531-2533 มีความลำเอียงต่อเพศ 7-28 ข้อ และ 4-41 ข้อ ตามลำดับ โดยมีแนวโน้มที่ลำเอียงต่อเพศชายมากกว่าเพศหญิงและลำเอียงต่อภาคภูมิศาสตร์ 6-45 ข้อ และ 5-43 ข้อตามลำดับ โดยมีความลำเอียงต่อผู้สอบจากภาคอื่นมากกว่าภาคกลางประมาณ 2-3 เท่า

การวิเคราะห์เปรียบเทียบผลการวิเคราะห์ความลำเอียงของข้อสอบทั้ง 3 วิธี พบข้อสอบที่ลำเอียงต่อเพศและต่อภูมิภาคศาสตร์ของผู้สอบจำนวนแตกต่างกันอย่างมีนัยสำคัญ แต่จำนวนข้อสอบที่ลำเอียงของแต่ละวิธีความสัมพันธ์กันอย่างไม่มีนัยสำคัญ ผลการวิเคราะห์ด้วยวิธีทฤษฎีการตอบสนองข้อสอบเมื่อไม่ได้วิเคราะห์ความลำเอียงระดับต่ำพบจำนวนข้อสอบที่ลำเอียงมากที่สุด รองลงมา ได้แก่ วิธีไค-สแควร์ และวิธีกำหนดจุดเดลต้าพบข้อสอบที่ลำเอียงจำนวนน้อยที่สุด

กาญจนา วัชนสุนทร (2537) ได้พัฒนาเกณฑ์ตัดสินข้อสอบลำเอียงทางเพศด้วยข้อมูลเชิงประจักษ์ สำหรับดัชนี 4 ตัว คือ พื้นที่ระหว่างโค้งการตอบข้อสอบชนิดคิดเครื่องหมาย (SA) และไม่คิดเครื่องหมาย (UA) จากวิธีทฤษฎีการตอบข้อสอบโมเดล 2 พารามิเตอร์ ดัชนีแอลฟา (α MH) จากวิธีแมนเทล-แฮนส์เซล และเบต้า (β SIB) จากวิธี SIBTESTโดยใช้ข้อมูลการตอบข้อสอบคัดเลือกเข้าศึกษาในสถาบันอุดมศึกษาของทบวงมหาวิทยาลัย ปีการศึกษา 2535 ในความยาวแบบสอบ 20, 30 และ 40 ข้อ สำหรับวิชา

คณิตศาสตร์ และ 50, 60 และ 80 ข้อ สำหรับวิชาภาษาอังกฤษ ใช้กลุ่มผู้สอบ 6 ขนาด คือ 100, 200, 400, 600, 800 และ 1,000 คน

เกณฑ์ที่พัฒนาจากข้อมูลชิงประจักษ์ เพื่อใช้ในการตัดสินใจความลำเอียงของข้อสอบระหว่างผู้สอบหญิงและชาย คือ

1. SA > .80 และ UA > .50 กรณีความยาวแบบสอบต่ำกว่า 50 ข้อ
2. SA > .80 และ UA > 1.20 กรณีความยาวแบบสอบ 50 ข้อขึ้นไป
3. .60 > α MH > 1.40 และ β SIB > .06 ทุกความยาวแบบสอบ และทุกขนาดผู้สอบ

ผลการศึกษา พบว่า ค่าเฉลี่ยของดัชนี SA UA αMH และ β SIB ที่ได้จากการเปรียบเทียบระหว่างผู้สอบเพศเดียวกันภายในความยาวแบบสอบ และขนาดผู้สอบต่างกัน มีค่าใกล้เคียงกันในแต่ละวิชา ค่าเฉลี่ยที่ได้จากวิชาภาษาอังกฤษค่อนข้างจะต่ำกว่าค่าเฉลี่ยที่ได้จากวิชาคณิตศาสตร์ สำหรับดัชนี SA และ UA ในขณะที่ค่าเฉลี่ยของ α MH และ β SIB มีค่าใกล้เคียงกันทั้ง 2 วิชา ผลการตรวจค้นข้อสอบลำเอียงทางเพศ เมื่อใช้ดัชนีตามที่กำหนด พบว่า มีความไม่คงที่ข้ามขนาดผู้สอบและความยาวแบบสอบ ความสอดคล้องในการตรวจค้นข้อสอบลำเอียงภายในวิธีเดียวกันข้ามขนาดผู้สอบต่ำแต่จะสูงขึ้นที่ขนาดผู้สอบ 600 คน สำหรับผลการวิเคราะห์ความลำเอียงของข้อสอบที่มีต่อเพศผู้สอบพบว่า ข้อสอบลำเอียงในวิชาภาษาอังกฤษเข้าข้างเพศหญิงมากกว่าชายเป็นส่วนใหญ่ ในกรณีของการใช้ดัชนี SA และ αMH ในขณะที่ผลการใช้ดัชนี β SIB จะให้ผลตรงข้าม ส่วนวิชาคณิตศาสตร์ข้อสอบลำเอียงที่พบจะลำเอียงเข้าข้างเพศชายมากกว่าเพศหญิง

4.2 งานวิจัยในต่างประเทศ

4.2.1 การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบจากข้อมูลจำลอง

Thissen, Steinberg และ Wainer (1988) ได้ศึกษาความแตกต่างของโค้งการตอบข้อสอบระหว่างกลุ่มด้วยทฤษฎีการตอบข้อสอบเพื่อศึกษาถึงสภาพการณ์ในการทดสอบสมมุติฐาน ความเท่ากันของค่าพารามิเตอร์ด้วยค่าไค-สแควร์แบบอัตราส่วน Likelihood (IRT-LR) โดยใช้ข้อมูลจำลอง และได้ทำการเปรียบเทียบการทดสอบสมมุติฐานความเท่ากันของค่าพารามิเตอร์ ด้วยวิธี IRT-LR และวิธี MH พบว่าวิธี IRT-LR สามารถตรวจค้นข้อสอบที่ DIF ได้ในเมื่อค่าความแตกต่างของค่าความยากระหว่างกลุ่มมีค่า 0.3 และความแตกต่างของค่าการเดามีค่า 0.1 จำนวนตัวอย่างในกลุ่มขนาด 500 คน โดยจำนวนข้อสอบร่วมจะไม่มีผลกระทบต่อขนาดความยาวของแบบสอบ สำหรับค่าความแตกต่างที่น้อยกว่านี้ไม่สามารถใช้ IRT-LR ตรวจค้น DIF ได้ แต่อาจทำได้ถ้าจำนวนตัวอย่างมากกว่านี้ การระบุถึงข้อสอบที่ทำหน้าที่ต่างกัน (DIF) ใช้วิธีเปรียบเทียบความแตกต่างระหว่างกลุ่มใน

ข้อสอบที่ต้องการศึกษาสำหรับผลการตรวจค้นการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี IRT-LR และ วิธี MH ให้ผลคล้ายกัน นอกจากนี้ วิธี MH อาจมีประโยชน์ในการประเมินแบบสอบยาวๆ และอาจใช้เป็นเครื่องมือกลั่นกรองก่อนที่ใช้ IRT-LR

Swaminatan และ Rojer (1990) ได้ตรวจค้นการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการถดถอยแบบโลจิส (LR-DIF) และวิธี MH โดยใช้ขนาดตัวอย่าง 2 ขนาด คือ 200 และ 500 คน และความยาวแบบสอบ 3 ขนาด คือ 40, 60 และ 80 ข้อ จำลองข้อมูลการตอบตามทฤษฎีการตอบข้อสอบด้วยโปรแกรม DATAGEN แบบ 3 พารามิเตอร์ ข้อมูลการจำลองข้อสอบที่ DIF สร้างแบบ 2 พารามิเตอร์ โดยให้ค่าอำนาจจำแนกของสองกลุ่มเท่ากัน ส่วนค่าความยากผันแปรไปเพื่อให้เกิดระดับของ DIF ที่ต้องการ แต่ในลักษณะของข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปนั้นให้ค่าความยากของ 2 กลุ่มเท่ากัน แต่ค่าอำนาจจำแนกเปลี่ยนไป

ผลการวิเคราะห์พบว่าการตรวจค้นการทำหน้าที่ต่างกันของข้อสอบ 2 วิธี โดยใช้การทดสอบ χ^2 ที่ชั้นความเป็นอิสระเท่ากับ 2 และการทดสอบ M H- χ^2 ให้ผลใกล้เคียงกันโดยวิธี MH ดีกว่าเล็กน้อยคือมีการตรวจค้นได้ถูกต้องร้อยละ 75 กรณีที่ใช้กลุ่มตัวอย่าง 250 คน และสามารถตรวจค้นได้ร้อยละ 100 ในกลุ่มตัวอย่างขนาด 500 คน ในทุกความยาวแบบสอบและในกรณีของการทำหน้าที่ต่างกันแบบเอกรูป สำหรับกรณีของการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปนั้นวิธี MH ไม่สามารถตรวจค้นได้ ส่วนวิธี LR-DIF สามารถตรวจค้นได้ถูกต้องประมาณร้อยละ 50 ในกรณีตัวอย่างน้อยและแบบสอบสั้น และถูกต้องประมาณร้อยละ 75 ในแบบสอบยาวและกลุ่มตัวอย่างขนาดใหญ่ แต่กรณี DIF แบบอเนกรูปไม่ว่าจะเป็นด้านความสามารถระดับสูงหรือระดับต่ำก็ตาม วิธี MH จะดีกว่า

สำหรับการตรวจค้นระดับความคลาดเคลื่อนชนิดที่ 1 (Type 1 error rate) พบว่าวิธี MH ตรวจค้นผิดพลาดประมาณ ร้อยละ 1 ส่วนวิธี LR-DIF ตรวจค้นผิดประมาณร้อยละ 1-6 นอกจากนี้พบว่าวิธี LR-DIF จะเสียค่าใช้จ่ายมากกว่าวิธี MH ประมาณ 3-4 เท่า

Clauser และคณะ (1991) ได้ศึกษาอิทธิพลต่างๆ ที่มีผลต่อวิธีการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธี MH ข้อมูลที่ใช้เป็นข้อมูลจำลองขึ้นเพื่อใช้ในการวิเคราะห์ DIF แบบเอกรูป โดยจำลองข้อมูลเป็นแบบสอบจำนวน 75 ข้อ เป็นข้อสอบที่ทำหน้าที่ต่างกันจำนวน16 ข้อ มีระดับของความแตกต่างในค่าความยากหลายระดับระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ซึ่งมีจำนวนกลุ่มละ 1,500 คน มีการวิเคราะห์ระดับความยากที่ต่างกันหลายระดับกับระดับความยากของแบบสอบ 5 ระดับ ส่วนอำนาจจำแนก 4 ระดับ และการกระจายของความสามารถ 2 ลักษณะ ผลการวิเคราะห์แสดงถึงอิทธิพลจากความแตกต่างในระดับความยากและค่าอำนาจจำแนกซึ่งผู้วิจัย ได้ให้ข้อเสนอแนะว่า ค่าสถิติจากวิธี MH นี้

จะใช้ได้ดีสำหรับข้อกระทงที่มีค่าอำนาจจำแนกสูงและมีแนวโน้มว่า จะไม่สามารถวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบได้ ในกรณีที่ข้อสอบมีค่าความยากสูง

Mazor และคณะ (1991) ได้ใช้วิธี MH เพื่อวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ โดยศึกษากับข้อมูลที่จำลองขึ้นตามทฤษฎีการตอบข้อสอบซึ่งใช้กลุ่มตัวอย่างแตกต่างกัน 5 ขนาดได้แก่ 2,000, 1,000, 500, 200 และ 100 คน เพื่อเปรียบเทียบอัตราการตรวจค้นโดยใช้ข้อสอบที่จำลองขึ้น 5 ชุด ชุดละ 75 ข้อ ตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันด้วยการทดสอบ MH-χ^2 ที่ระดับ df=1 ผลการศึกษาพบว่าเมื่อใช้กลุ่มตัวอย่างขนาด 2,000 คน วิธีนี้ตรวจค้นได้ผิดพลาดร้อยละ 50 และเมื่อใช้จำนวนผู้สอบขนาด 500 คน พบว่าตรวจค้นผิดพลาดร้อยละ 50 ข้อกระทงที่ไม่สามารถตรวจค้น DIF ได้ ได้แก่ข้อกระทงที่มีค่าความยากต่างกันเล็กน้อยสำหรับ 2 กลุ่ม และเป็นข้อกระทงที่มีค่าอำนาจจำแนกต่ำ

Clauser และคณะ (1993) ได้ศึกษาอิทธิพลของการทำให้เกณฑ์การจับคู่ผู้สอบที่มีความบริสุทธิ์ (purification of the matching criterion) ระหว่างเทคนิค 1 ขั้นตอน (one step procedure) และเทคนิค 2 ขั้นตอน (two step procedure) ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธี MH โดยการใช้สถานการณ์จำลองสร้างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบกลุ่มละ 1,000 คน จำลองสถานการณ์ตามเงื่อนไขต่างๆ กัน 24 เงื่อนไข คือ สร้างแบบสอบความยาว 3 ขนาด ได้แก่ 20 ข้อ 40 ข้อ และ 80 ข้อ สร้างข้อสอบที่ทำหน้าที่ต่างกันลงในแบบสอบแต่ละขนาดจำนวน 0%, 3%, 8% และ 20% รวมทั้งสร้างเงื่อนไขของระดับความสามารถ 2 ระดับ ผลการศึกษาปรากฏว่า ผลการตรวจพบข้อสอบที่ทำหน้าที่ต่างกันด้วยเทคนิค 2 ขั้นตอน เท่ากับหรือเหนือกว่าเทคนิค 1 ขั้นตอน ในทุกเงื่อนไขของการทดสอบ และเทคนิค 2 ขั้นตอน ไม่เพิ่มระดับความคลาดเคลื่อนประเภทที่ 1 (type 1 error rate) มากกว่าเทคนิค 1 ขั้นตอน

Clauser และคณะ (1994) ได้ศึกษาผลกระทบจากความกว้างของชั้นคะแนน (score group) ที่มีต่ออำนาจการทดสอบและระดับความคลาดเคลื่อนประเภทที่ 1 ของวิธี MH จากการใช้ข้อมูลจำลองซึ่งมีทั้งข้อสอบที่ทำหน้าที่ต่างกัน (DIF) และข้อสอบที่ทำหน้าที่ไม่แตกต่างกัน (no DIF) ร่วมกันจำนวน 80 ข้อ สุ่มขนาดผู้สอบออกเป็น 5 ขนาด ได้แก่ขนาด 2,000, 1,000, 500, 200 และ 100 จากนั้นทำการตรวจค้นการทำหน้าที่ต่างกันของข้อสอบโดยใช้จำนวนชั้นคะแนนเป็น 2, 5, 10, 20 และ 80 ตามลำดับ ผลการทดสอบทางสถิติพบว่าการเพิ่มขนาดของผู้สอบจะช่วยให้เพิ่มอำนาจการทดสอบและไม่เพิ่มระดับความคลาดเคลื่อนประเภทที่ 1 สำหรับการลดจำนวนชั้นคะแนนในการจับคู่เกณฑ์ของคะแนน (matching score) นั้น ถึงแม้ว่าจะเพิ่มอำนาจการทดสอบแต่จะเพิ่มระดับความคลาดเคลื่อน

ประเภทที่ 1 ด้วย Clauser และคณะ จึงได้เสนอแนะว่าในการใช้วิธี MH วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ นักวิจัยควรใช้จำนวนชั้นคะแนนที่เป็นเกณฑ์จำนวนคะแนนที่เป็นไปได้มากที่สุดจากคะแนนรวมของผู้สอบ(total score) เป็นเกณฑ์ในการจับคู่ โดยเฉพาะอย่างยิ่งเมื่อการกระจายของคะแนนของความสามารถของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบแตกต่างกัน

Mazor และคณะ (1994) ได้ศึกษาถึงการระบุการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปด้วยวิธี MH โดยการจำลองผู้สอบขึ้นจำนวน 1,000 คน และสร้างข้อสอบที่จำลองขึ้นตามทฤษฎีการตอบข้อสอบ 3 พารามิเตอร์ โดยการจำลองลักษณะของข้อสอบตามเงื่อนไขต่างๆ ทั้งหมด 400 ข้อ ได้แก่ ข้อสอบที่มีค่าความยากต่างกัน 5 ระดับ ข้อสอบที่มีค่าอำนาจจำแนกต่างกัน 5 ระดับ ข้อสอบที่ค่าความยากระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบมีความต่างกัน 4 ระดับ และข้อสอบที่ค่าอำนาจจำแนกระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบมีความต่างกัน 4 ระดับ จากนั้นทำการสุ่มข้อสอบที่ทำหน้าที่ไม่ความต่างกัน (no DIF) จำนวน 59 ข้อ และข้อสอบที่ทำหน้าที่ต่างกัน (DIF) จำนวน 16 ข้อ รวมเป็นฉบับละ 75 ข้อ ทำการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยใช้โปรแกรม MH ของ Rojer และ Hambleton (1989)

ผลการวิเคราะห์พบว่าวิธี MH สามารถระบุลักษณะของข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปได้ถูกต้องมากกว่าร้อยละ 60 ในทุกเงื่อนไข โดยสามารถวิเคราะห์ได้ถูกต้องร้อยละ 82 เมื่อการกระจายของความสามารถของผู้สอบสองกลุ่มเท่ากัน และถูกต้องร้อยละ 75 เมื่อการกระจายของความสามารถของผู้สอบสองกลุ่มแตกต่างกัน นอกจากนั้นยังพบว่าไม่มีการเพิ่มระดับของความคลาดเคลื่อนประเภทที่ 1 ในทุกเงื่อนไข ดังนั้นผลจากการศึกษานี้จึงชี้ให้เห็นว่าวิธี MH สามารถระบุถึงลักษณะของข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปได้เช่นเดียวกับวิธีที่มีพื้นฐานจากทฤษฎี IRT หรือวิธี LR-DIF

4.2.2 งานวิจัยในต่างประเทศที่วิเคราะห์ DIF โดยใช้ข้อมูลจริง

Shoener (1984) ได้เปรียบเทียบการตรวจค้นความลำเอียงของข้อสอบด้วยวิธีการทางสถิติ ได้แก่ IRT 3 พารามิเตอร์ ซึ่งให้การทดสอบ χ^2 และการตัดสินด้วยผู้ตัดสินจำนวน 24 คน โดยใช้แบบสอบอิงเกณฑ์กับนักเรียนชั้นมัธยมศึกษาชายและหญิง จำนวน 1,064 คน ที่มีภูมิหลังทางวัฒนธรรมต่าง ๆ กัน การตัดสินใช้ผู้ตัดสินที่มีความรู้ทางด้านคณิตศาสตร์ และหลักสูตรคณิตศาสตร์เป็นอย่างดีและใช้แบบฟอร์มการให้คะแนนเป็นช่วง (rating form) ผู้ตัดสินเป็นผิวดำ 8 คน ผิวขาว 8 คน และเชื้อสายสเปน 8 คน เป็นชายและหญิงเท่า ๆ กัน

การวิเคราะห์ได้พิจารณาถึงความสอดคล้องของการตัดสินใจของกลุ่มผู้ตัดสิน 3 กลุ่ม และความสอดคล้องระหว่างวิธีการทางสถิติและการตัดสินของผู้ตัดสิน โดยใช้ค่าสถิติของKappa ผลจากการวิเคราะห์ พบว่า ความสอดคล้องระหว่างการตัดสินโดยผู้เชี่ยวชาญต่างกลุ่มไม่มีนัยสำคัญสำหรับทั้งความลำเอียงด้านวัฒนธรรมและทางเพศ พบความสอดคล้องกันอย่างมีนัยสำคัญในตัวบ่งชี้บางตัวที่คำนวณขึ้น เพื่อรวมการให้คะแนนของผู้ตัดสิน แต่ไม่พบความสอดคล้องระหว่างวิธีการตรวจค้นความลำเอียงทั้ง 2 วิธี การตัดข้อกระทงที่พบทางสถิติว่าลำเอียงออกไปไม่ทำให้ลำดับของคนในกลุ่มย่อยและรวมทั้งหมดเปลี่ยนไปและทั้งแบบสอบฉบับเดิม และแบบสอบที่ตัดข้อกระทงที่ลำเอียงออกแล้วมีความสัมพันธ์อย่างมีนัยสำคัญกับแบบสอบวิธีคณิตศาสตร์อิงกลุ่มที่เป็นแบบสอบมาตรฐาน

Hambleton และคณะ (1986) ได้เปรียบเทียบวิธีการตรวจค้น DIF ของข้อสอบ 4 วิธี ได้แก่ วิธีของ Mantel-Haenszel วิธีการลงจุดค่าความยาก วิธีค่าความแตกต่างของค่าเฉลี่ยกำลังสอง (the root meansquared difference) และวิธีของพื้นที่รวม ซึ่ง 2 วิธีหลังนี้เป็นวิธีการที่ใช้ทฤษฎีการตอบข้อสอบแบบสอบถามที่ใช้ในการศึกษา เป็นแบบสอบความสามารถในการอ่านของคลีฟแลนด์ มีข้อสอบจำนวน 75 ข้อ โดยใช้กลุ่มตัวอย่างชาย 451 คน หญิง 486 คน คะแนนจุดตัดในการแปลความหมายค่าสถิติที่แสดงถึงข้อสอบที่ทำหน้าที่ต่างกัน (DIF) ได้จากการจำลองข้อมูลขึ้น ผลปรากฏว่าวิธีทั้ง 4 วิธี ให้ผลการตรวจค้นข้อสอบที่ DIFใกล้เคียงกัน และพบว่า วิธี MH เป็นวิธีเลือกใช้แทนวิธีทฤษฎีการตอบข้อสอบได้อย่างประหยัดกว่าทั้งเงินและเวลา

Doolitte และ Cleary (1987) ได้ตรวจค้นการทำหน้าที่ต่างกันระหว่างเพศของข้อสอบในแบบวัดผลสัมฤทธิ์ทางคณิตศาสตร์ โดยใช้ตัวอย่างจำนวน 8 กลุ่ม ที่ทำแบบสอบ ACT Assessment Mathematic Test รวม 8 ฟอร์ม แต่ละกลุ่มมีความเท่าเทียมกันและมีจำนวนประมาณ 1,300-1,400 คน ตัวอย่างเป็นหญิงประมาณร้อยละ 55 วิเคราะห์ด้วยดัชนีของ ลินน์ และ ฮาร์นิสช์ (Linn and Harnisch, 1981) ซึ่งเป็นโมเดลโลจิสติคแบบ 3 พารามิเตอร์ ถ้าดัชนี Z ที่ได้เป็นบวกแสดงว่าง่าย สำหรับกลุ่มเปรียบเทียบถ้าเป็นลบแสดงว่ายากสำหรับกลุ่มเปรียบเทียบ ผลการวิเคราะห์พบว่า ดัชนี Z มีค่าเป็นลบ สำหรับข้อกระทงที่วัดด้านเรขาคณิต และการใช้เหตุผลเชิงพีชคณิต และเลขคณิตในทุกฟอร์มแสดงว่าง่ายสำหรับชายมากกว่าหญิงที่เหลือส่วนมากเป็นบวก สำหรับผลจากการวิเคราะห์ด้วยการวิเคราะห์ความแปรปรวน พบผลเช่นเดียวกับการคำนวณค่า Z

Perlman และคณะ (1988) ได้ศึกษาความคงที่ของวิธีการประมาณค่าความลำเอียงของข้อสอบ 4 วิธี ได้แก่ วิธีค่าความยากแปลงวิธีค่าความยากแปลงที่นำมาดัดแปลงโดย Shepard วิธีการวิเคราะห์ค่าที่เหลือแบบ 1 พารามิเตอร์ของ Rasch และวิธี MH

กลุ่มตัวอย่างที่ใช้มี 30 กลุ่ม โดยมีขนาด 600 จนถึง 2,000 คน ซึ่งสุ่มมาจากนักเรียนชั้น 9 จำนวน 54,896 คน ซึ่งมีทั้งผิวขาว ผิวดำ และชนชาติที่พูดภาษาสเปนที่ทำแบบทดสอบทักษะความสามารถขั้นต่ำของรัฐซิคาโก ซึ่งประกอบด้วยข้อสอบเลือกตอบ 46 ข้อ จากธนาคารข้อสอบจำนวน 1,000 ข้อ ผลจากการศึกษา พบว่า ความเที่ยงของตัวบ่งชี้ความลำเอียงจะมีปัญหา ถ้าใช้จำนวนตัวอย่างน้อยกว่า 666 คน ไม่มีวิธีใดที่ให้ผลการตรวจค้นความลำเอียงได้มากกว่า หรือน้อยกว่าวิธีอื่น ๆ อย่างคงที่ข้ามขนาดตัวอย่าง

Baeza (1989) ได้ศึกษาพฤติกรรมการตอบข้อสอบของชาวอเมริกันอินเดียนและอเมริกันคอเคเซี่ยน เพื่อพิจารณาระดับของความลำเอียงทั้งภายในและภายนอกข้อสอบ โดยใช้ตัวแปรด้านเศรษฐกิจสังคม และระดับคะแนนเฉลี่ยในชั้นมัธยมปลายเป็นตัวแปร จับคู่กลุ่มตัวอย่างนอกเหนือไปจากเพศ โรงเรียนที่เรียนอยู่ และปีการศึกษาที่เรียน

วิธีการที่ใช้ในการตรวจค้นความลำเอียง ได้แก่ วิธี MH และวิธีการจับคู่กลุ่มตัวอย่างของ McNemar ตัวแปรจับคู่สำหรับ MH ได้แก่ คะแนนรวมจากแบบสอบ ส่วนตัวแปรจับคู่สำหรับ McNemar ได้แก่ ตัวแปรด้านเศรษฐกิจสังคมหรือระดับคะแนนเฉลี่ย

ผลการวิเคราะห์ พบว่า วิธี MH ซึ่งใช้เกณฑ์ภายในตรวจพบข้อสอบที่ลำเอียงเพียงเล็กน้อย ส่วน McNemar ซึ่งใช้เกณฑ์ภายนอกนั้น พบว่าเมื่อใช้ตัวแปรจับคู่เป็นสถานภาพทางเศรษฐกิจสังคม พบว่า ข้อสอบมีระดับความลำเอียงต่ำกว่าการจับคู่ด้วยระดับคะแนนเฉลี่ยแสดงว่า ระดับคะแนนเฉลี่ยของคน 2 กลุ่มนี้ มีระดับของความหมายไม่เท่ากัน

Baghi และ Ferrara (1989) ได้เปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธีการแตกต่างกัน 3 วิธี ได้แก่ การใช้ทฤษฎีการตอบข้อสอบ วิธีการใช้ค่าเดลต้า (ค่าความยากแปลง) และวิธี MH โดยศึกษาในแง่ของผลกระทบจากขนาดของกลุ่มตัวอย่างในผลการตรวจค้น DIF และระดับของความสัมพันธ์ระหว่างค่าสถิติจากทั้ง 3 วิธีนี้ ข้อมูลที่ใช้ในการวิเคราะห์ เป็นข้อมูลที่สุ่มจากนักเรียนระดับ 9 จำนวน 50,000 คน ที่ตอบแบบสอบถามทักษะความเป็นพลเมืองดี ประกอบด้วยข้อสอบแบบเลือกตอบ 45 ข้อ ประเมินความรู้และทักษะ รวม 3 ด้าน ได้แก่ ด้านรัฐบาลประชาธิปไตย ด้านการเมือง และพฤติกรรมทางการเมือง และด้านหลักการ สิทธิ และความรับผิดชอบ จำนวนกลุ่มตัวอย่างที่สุ่มมาใช้มี 4 ขนาด ได้แก่ ขนาด 1,000 คน 750 คน 500 คน และ 200 คน

ผลจากการวิเคราะห์ไม่พบว่ามีข้อสอบข้อใดที่ DIF ทั้งในการเปรียบเทียบระหว่างผู้สอบผิวดำและผิวขาว และผู้สอบหญิงและชาย การแสดงค่าประมาณความยากด้วย

แผนภูมิเส้นแสดงความสัมพันธ์ระหว่างค่าความยากต่างกลุ่มอยู่ในรูปเส้นตรงเกือบสมบูรณ์ ความสอดคล้องกันระหว่างวิธีการของ Rasch และค่าความยากแปลงมีความสัมพันธ์กันสูงมาก ทุกขนาดของกลุ่มตัวอย่างจากค่าสัมประสิทธิ์สหสัมพันธ์แบบลำดับ และทั้ง 2 วิธี มีความสอดคล้องสูงสุดกับวิธีการตอบข้อสอบ 3 พารามิเตอร์ ในการตรวจค้นความลำเอียงและไม่ลำเอียงของข้อสอบ

Hambleton และ Rogers (1989) ได้ศึกษาเปรียบเทียบวิธีการวิเคราะห์ DIF ของข้อสอบ 2 วิธี ได้แก่ MH และ IRT โดยใช้ข้อมูลการตอบแบบสอบ New Mexico High School Proficiency Exam (NMHSPE) จำนวน 150 ข้อ ของนักเรียนชาวอเมริกันผิวขาว 8,000 คน และชาวอเมริกันพื้นเมือง 2,600 คน จากผู้สอบทั้งหมด 23,000 คน

NMHSPE เป็นแบบสอบชนิดเลือกตอบ 4 ตัวเลือก วัดทักษะการดำเนินชีวิต 5 ด้าน คือ (1) ความรู้เกี่ยวกับแหล่งทรัพยากรของชุมชน (2) ความรู้ด้านผู้บริโภค (3) ความรู้เกี่ยวกับรัฐบาลและกฎหมาย (4) ความรู้ด้านสุขภาพทางกายและจิต และ(5)ความรู้ด้านอาชีพ

กลุ่มตัวอย่างที่นำมาศึกษา ซึ่งได้แก่ ชาวอเมริกันผิวขาวและอเมริกันพื้นเมือง นั้นได้สุ่มมาศึกษาอย่างละ 2,000 คน และแบ่งเป็น 2 กลุ่ม ๆ ละ 1,000 คน ตามลำดับคู่-คี่ การศึกษาความลำเอียง 2 คู่ เช่นนี้ เพื่อพิจารณาถึงความคงที่ในการระบุข้อสอบลำเอียง

ในการศึกษาการทำหน้าหน้าที่ต่างกันของข้อสอบ ผู้วิจัยได้ใช้ข้อสอบเพียง 75 ข้อ โดยตัดข้อสอบที่ง่ายมาก (p >.70) และค่าอำนาจจำแนกต่ำ (r <.10) วิธี MH วิเคราะห์โดยใช้ 76 ระดับคะแนน ส่วนวิธีทฤษฎีการตอบข้อสอบใช้การวัดพื้นที่ระหว่างโค้งลักษณะข้อสอบ (ICC) พิสัยของความสามารถที่นำมาใช้การคำนวณพื้นที่กำหนดใช้ 2 ค่าเบี่ยงเบนมาตรฐานสูงกว่าความสามารถเฉลี่ยของชาวอเมริกันผิวขาว และ 2 ค่าเบี่ยงเบนมาตรฐานต่ำกว่าความสามารถเฉลี่ยของชาวอเมริกันพื้นเมือง ความคงที่ของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ได้แก่ ร้อยละของข้อสอบที่ถูกระบุว่า DIF หรือ no DIF ซ้ำกันในการวิเคราะห์ 2 ครั้ง

ผลการวิเคราะห์ พบว่า วิธี MH มีความสอดคล้องของการตรวจสอบร้อยละ 80 ส่วนวิธี ICC ร้อยละ 73 และพบว่า ร้อยละ 47 ของข้อสอบที่ถูกระบุว่า DIF ในการ วิเคราะห์คู่ที่ 1 เป็นข้อสอบที่ DIF ในการวิเคราะห์คู่ที่ 2 ด้วย สำหรับ MH และร้อยละ 61 สำหรับวิธี ICC ในทางกลับกัน พบว่า ร้อยละ 64 ของข้อสอบที่ DIF ในการวิเคราะห์คู่ที่ 2 เป็น ข้อสอบที่ DIF ในการวิเคราะห์คู่ที่ 1 สำหรับวิธี MH และร้อยละ 56 สำหรับวิธี ICC ผลดังกล่าวแสดงให้เห็นว่าความคงที่ของการวิเคราะห์ทั้งวิธี MH และ ICC อยู่ในระดับปานกลางซึ่งอาจเป็นเพราะจำนวนตัวอย่างค่อนข้างน้อย (วิธี MH มีข้อสอบที่ระบุว่า DIF ในการวิเคราะห์ทั้ง 2 ครั้ง 7 ข้อ วิธี ICC 14 ข้อ และพื้นที่แสดงความแตกต่างของ ICC มีค่าระหว่าง -2.7 และ 1.5)

สรุปได้ว่า ทั้งวิธี MH และ ICC มีความไม่เที่ยงในระดับหนึ่งในการใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยที่ วิธี MH มีความคงที่ประมาณร้อยละ 80 และ วิธี ICC มีความคงที่ประมาณร้อยละ 73

Sudweeks และ Tolman (1990) ได้เปรียบเทียบผลจากการใช้วิธี MH และการตัดสินด้วยผู้เชี่ยวชาญในการตรวจค้นความลำเอียงของข้อสอบวิชาวิทยาศาสตร์ระหว่างนักเรียนชายและหญิงระดับ 5 จำนวน 926 คน เครื่องมือที่เก็บรวบรวมข้อมูลเป็นแบบสอบอิงเกณฑ์ชนิดเลือกตอบที่สร้างขึ้นตามมาตรฐานและวัตถุประสงค์ของหลักสูตรวิทยาศาสตร์ร่วมในรัฐยูท่าห์ ผลจากการวิเคราะห์ พบว่าวิธีการทั้งสองให้ผลต่างกัน และข้อสอบหลายข้อที่ถูกระบุด้วยวิธีการ 2 วิธีนี้ว่า ลำเอียง จะเป็นข้อสอบที่ยาก

Baghi และ Ferrara (1990) ได้เปรียบเทียบผลจากการวิเคราะห์ DIF ของข้อสอบด้วยทฤษฎีการตอบข้อสอบแบบ 3 พารามิเตอร์ (IRT) และสถิติไค-สแควร์แบบ MH (MH-χ^2) โดยใช้ข้อมูลจากการบริหารแบบสอบทักษะความเป็นพลเมืองดีของรัฐแมรี่แลนด์ ให้แก่นักเรียนระดับ 9 จำนวน 50,000 คน ในเดือนมกราคม - กุมภาพันธ์ 1988 จำนวนตัวอย่างที่ใช้ในการวิเคราะห์ด้วย MHCS มี 4 ขนาด ได้แก่ 1,000 คน 750 คน 500 คน และ 200 คน ส่วนจำนวนตัวอย่างในการวิเคราะห์ด้วย IRT ใช้กลุ่มละ 1,000 คน ในแต่ละกลุ่มเปรียบเทียบการวิเคราะห์ประกอบด้วย (1) การพิจารณาความคงที่ของค่าสถิติ MH-χ^2 เมื่อใช้กลุ่มตัวอย่างขนาดต่าง ๆ (2) ความคงที่ของดัชนี αMH (MH Alpha) ในวิธี MH-χ^2 ข้ามช่วงคะแนน (3) ความสัมพันธ์ระหว่างดัชนีแบบ IRT และค่า MH Alpha และ (4) ความสอดคล้องระหว่างวิธีทั้งหมดในการวิเคราะห์ข้อกระทงที่ทำหน้าที่ต่างกัน

ผลปรากฏว่า เมื่อใช้วิธี IRT วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบระหว่างกลุ่มหญิงและชายพบข้อสอบที่ DIF 4 ข้อ และพบข้อสอบ DIF 3 ข้อในการเปรียบเทียบระหว่างผู้สอบผิวขาวและผิวดำ ซึ่งมีความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องไม่เท่ากันอย่างมีนัยสำคัญ และในขณะที่ผลจากการหาค่าสหสัมพันธ์ระหว่างดัชนีทั้งสองตัวแสดงความไม่สอดคล้องกันในการบ่งชี้ข้อสอบที่ DIF นั้น ความสอดคล้องในผลรวมแสดงว่าเทคนิค MH-χ^2 เป็นวิธีที่ใช้แทนIRT แบบ 3 พารามิเตอร์ได้อย่างเพียงพอถ้าจำนวนตัวอย่างมีขนาดอย่างน้อย 750 คน

Ryan (1991) ได้วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการ MH โดยใช้ตัวอย่างกลุ่มอ้างอิง 500 คน และกลุ่มเปรียบเทียบ 100 คน เพื่อตรวจสอบความคงที่ของค่าประมาณด้วยวิธี MH ข้ามกลุ่มตัวอย่างและขนาดตัวอย่างต่าง ๆ กัน และเพื่อศึกษาว่าวิธี MH มีความแกร่งต่อบริบทของข้อสอบเพียงไร โดยมีข้อสอบร่วมกันจำนวน 40 ข้อ และข้อสอบที่ใช้สลับกันอีกจำนวน 35 ข้อ กลุ่มตัวอย่างเป็นชาวผิวขาวและผิวดำ จำนวน 5,015

คน และ 670 คน ตามลำดับโดยมีเกณฑ์การแยกประเภทข้อสอบจากผลการวิเคราะห์เป็น 3 ประเภท คือ

ประเภท A ข้อสอบที่ทำทำหน้าที่ได้ตรง (no DIF) มีค่า α MH < 1.00 และมีค่าไม่ต่างจาก 0 อย่างมีนัยสำคัญ

ประเภท B ข้อสอบที่พอจะนำมาใช้ได้ แต่ควรเลือกข้อที่มีค่า α MH ต่ำ คือ 1 = α MH < 1.5 และไม่ต่างจาก 1 อย่างมีนัยสำคัญ

ประเภท C เลือกมาใช้ถ้าตรงตามวัตถุประสงค์เฉพาะที่ระบุไว้อย่างมีนัยสำคัญ ถ้าต้องการข้อสอบลักษณะนี้ ได้แก่ ข้อที่มีค่า α MH > 1.5 และมากกว่า 1 อย่างมีนัยสำคัญ

ขั้นตอนในการวิเคราะห์ขั้นแรกมีการวิเคราะห์ 5 แบบ ในการสร้างฐานเรียกว่า ชุด 1 ทำโดยการสุ่มตัวอย่างผิวขาว 670 คน เป็นกลุ่มเป้าหมาย ที่เหลือเป็นกลุ่มอ้างอิงแล้วสุ่มกลุ่มผิวขาวมาอีก 4 กลุ่ม เป็นกลุ่มอ้างอิง ได้แก่ W_1 W_2 W_3 W_4 และวิเคราะห์ข้อสอบร่วม 40 ข้อ ด้วย MH จับคู่การวิเคราะห์เป็น W/W W/W_1 W/W_2 W/W_3 W/W_4 คะแนนที่ใช้แบ่งระดับความสามารถ คือ คะแนนรวมจากข้อสอบ 40 ข้อ

การวิเคราะห์ชุด 2 ทำให้ลักษณะเดียวกับ ชุด 1 แต่ใช้กลุ่มเป้าหมายเป็นผิวดำวิเคราะห์ด้วย MH 5 คู่ ได้แก่ BW BW_1 BW_2 BW_3 BW_4 เป็นการวิเคราะห์เพื่อดูความคงที่ของค่า α MH ข้ามกลุ่มและขนาดตัวอย่าง

ชุดที่ 3 เป็นการวิเคราะห์ผลของบริบทของข้อสอบ BW_1 BW_2 BW_3 และ BW_4 โดยใช้ข้อสอบ 75 ข้อจาก ฟอร์ม 1 ฟอร์ม 2 ฟอร์ม 3 ฟอร์ม 4 ใช้คะแนนรวมเป็นเกณฑ์แบ่งช่วงความสามารถ (แต่ละฟอร์มมีข้อสอบร่วม 40 ข้อ)

และชุดที่ 4 เป็นการวิเคราะห์เช่นเดียวกับชุด 3 แต่ใช้ข้อสอบ 4 ฟอร์มที่ไม่มีข้อสอบร่วม (มีข้อสอบ 35 ข้อ) และใช้คะแนนรวมจาก 35 ข้อ เป็นเกณฑ์แบ่งระดับความสามารถ

ผลการวิเคราะห์ พบว่า ข้อสอบเข้าข้างสำหรับกลุ่มผิวขาวด้วยกัน 5 คู่นั้นมีความสัมพันธ์กันต่ำมาก และไม่มีความแตกต่างอันเนื่องมาจากข้อสอบที่DIFแต่มีความแตกต่างจากความคลาดเคลื่อนในการสุ่มตัวอย่างที่มีความซ้ำซ้อนกันมากกว่า

เกณฑ์การแบ่งช่วงระดับความสามารถ ซึ่งเป็นบริบทที่ศึกษาไม่มีผลต่อค่า α MH ไม่ว่าจะใช้เกณฑ์คะแนนรวม 35 ข้อ หรือ 75 ข้อ ก็ได้ค่าสหสัมพันธ์ของการเปรียบเทียบแต่ละคู่เท่า ๆ กัน

ผลจากการจำแนกประเภทข้อสอบมีเพียง 1 ข้อ ที่มีความสอดคล้องกับทุกการวิเคราะห์ คือ เป็นประเภท B ทุกครั้ง นอกนั้นมีความแปรปรวนไปตามการวิเคราะห์

Harris และ Carlton (1993) ได้ศึกษาลักษณะกว้าง ๆ ของข้อสอบเพื่อชี้ให้เห็นจุดเด่นและจุดด้อย สำหรับผู้สอนหญิงและชายโดยแบ่งลักษณะที่ศึกษาเป็นประเด็นที่วัดรูปแบบและเนื้อหาของข้อสอบ

ข้อสอบที่นำมาศึกษาเป็นข้อมูลการตอบข้อสอบในแบบสอบ SAT-M จำนวน 6 ฟอร์ม ซึ่งเป็นตอนที่ทดสอบเฉพาะคณิตศาสตร์ ประกอบด้วยข้อสอบ 60 ข้อ เป็นการแก้ปัญหาทางคณิตศาสตร์ทั่วไป 40 ข้อ และการเปรียบเทียบปริมาณตัวเลข 20 ข้อ ผู้สอบเป็นนักเรียนชั้นมัธยมทั้งต้นและปลาย เป็นชาย 181,228 คน หญิง 98,668 คน ทุกคนรายงานด้วยตนเองว่า ภาษาอังกฤษเป็นภาษาที่แต่ละคนใช้ได้ดีที่สุด จำนวนกลุ่มตัวอย่างที่ใช้ใน SAT-M ฟอร์ม 1-6 เป็นชาย 6,329-74,283 คน หญิง 6,712-83,945 คน ตรวจสอบด้วย วิธี MH และพิจารณาความแตกต่างของค่าเดลต้า ซึ่งใช้แสดงถึงความยากของข้อสอบใน SAT อยู่แล้ว โดยที่ค่าเดลต้า = 1.00 แสดงว่า ข้อสอบมีความยากสำหรับกลุ่มหนึ่งมากกว่าอีกกลุ่มหนึ่งเท่ากับ 1 เดลต้า หรือเท่ากับความแตกต่างร้อยละ 10 โดยประมาณ

ค่าเดลต้าที่เป็นลบแสดงว่า ยากสำหรับหญิงมากกว่าชาย ค่าบวกแสดงว่ายากสำหรับชายมากกว่าหญิง

ผลการวิเคราะห์โดยจับคู่ชายและหญิงในด้านความสามารถ ซึ่งได้แก่คะแนนรวมจากแบบสอบแล้ว ได้ค่าเดลต้ามีค่า-1.86 ถึง 1.27 ค่าเฉลี่ย -.01 และค่า ความเบี่ยงเบนมาตรฐาน = .51 และใช้การวิเคราะห์ความแปรปรวนทดสอบนัยสำคัญของความแตกต่างเฉลี่ยระหว่างข้อที่มีและไม่มีลักษณะดังที่ระบุไว้ พบว่า โดยทั่ว ๆ ไป ถ้าเนื้อหาหลักเป็นเรขาคณิตหรือเลขคณิตแล้วชายจะทำได้ดีกว่าหญิง และกรณีเป็นเนื้อหาทั่ว ๆ ไป (เช่น โครงสร้างของระบบจำนวน เซท ฯลฯ) หญิงจะทำได้ดีกว่าชาย ถ้าเป็นเลขคณิตและพีชคณิต หญิงจะทำได้ดีกว่าชายแต่ถ้าเป็นเลขคณิตและเรขาคณิตแล้ว ชายได้ดีกว่าหญิง

นอกจากนั้น พบว่า ชายจะทำได้ดีกว่าในข้อสอบที่ต้องการกระบวนการทางสมองในระดับสูงกว่า ในข้อที่เป็นการประยุกต์นำมาใช้ในชีวิตประจำวัน ส่วนหญิงทำได้ดีกว่ากรณีที่มีตัวแปร (เช่น x, a, b) ในปัญหาและในตัวเลือก กรณีเป็นปัญหาตรงไปตรงมาจากหลักสูตรหรือหนังสือเรียนโดยไม่มีการนำมาประยุกต์และในกรณีมีการอ้างอิงถึงบุคคลโดยไม่มีการแสดงถึงเพศหรือสถานภาพทางเชื้อชาติ

Raju และคณะ (1993) ได้ศึกษาเปรียบเทียบผลการทดสอบค่า Z ของ Raju ทั้งที่ได้จากพื้นที่ชนิดมีและไม่คิดเครื่องหมาย การทดสอบค่า χ^2 ของ Lord (ใช้โมเดล 2 พารามิเตอร์) และการทดสอบ MH- χ^2 ข้อมูลที่ใช้ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบคือผลการตอบแบบสอบ Gater-Mac Ginitie Reading Tests (GMRT) เฉพาะแบบสอบย่อยที่วัดคำศัพท์ 45 ข้อ เป็นแบบเลือกตอบ 5 ตัวเลือก กับนักเรียนระดับ 10 และ 12 ใน

ปี 1987 จำนวน 839 คน ในจำนวนนี้มีนักเรียนผิวดำ 245 คน ผิวขาว 436 คน เป็นหญิง 440 คน และชาย 399 คน

ผลจากการศึกษาในกรณีการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบระหว่างนักเรียนผิวดำและผิวขาว พบว่า มีข้อสอบที่ DIF 1 ข้อ สำหรับการทดสอบด้วย χ^2 และ Z ทั้งพื้นที่ชนิดคิดเครื่องหมายและไม่คิดเครื่องหมายได้แก่ ข้อ 41 ส่วนวิธี MH พบข้อสอบที่ DIF 2 ข้อ ได้แก่ ข้อ 14 และ 27 ทั้ง 3 ข้อ เป็นข้อที่มีความยากต่างกันมากสำหรับผู้สอบผิวดำและผิวขาว ส่วนผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบระหว่างชายและหญิง 3 วิธีแรกพบข้อสอบที่ DIF ซ้ำกัน 4 ข้อ ส่วนวิธี MH พบ 5 ข้อ ซ้ำกับ 3 วิธีแรก 4 ข้อ (ร้อยละ 80)ได้แก่ ข้อ 2, 18, 23, 33 และข้อที่ไม่ซ้ำกับ 3 วิธีแรก คือ ข้อ 41 ทุกข้อที่พบ DIF มีค่าความยากสูงกว่าข้อที่ไม่มีความแตกต่างกันระหว่างกลุ่มหรือ no DIF

จากการศึกษาเรื่องความยุติธรรมของข้อสอบหรือแบบสอบที่กล่าวมานี้ พอสรุปได้ว่าเดิมนักวิจัยใช้คำว่าความลำเอียงของข้อสอบ/แบบสอบ แต่ในปัจจุบันนักวิจัยใช้คำว่า การทำหน้าที่ต่างกันของข้อสอบ (DIF) สำหรับการศึกษาเรื่อง DIF ในประเทศไทยยังมีน้อยมาก เท่าที่ผ่านมามีการศึกษาใน 2 ลักษณะ คือ การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบระหว่างกลุ่มผู้สอบเพศชายและหญิง (ชัชชัย เผ่าพงศ์, 2526 ; สุรศักดิ์ อมรรัตนศักดิ์, 2531 ; พัชรี ปิยภัณฑ์, 2531 ; สุพัฒน์ สุกมลสันต์, 2534 ; กาญจนา วัชนสุนทร, 2537) และการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบระหว่างกลุ่มผู้สอบที่มีภูมิลำเนาหรือสภาพสังคมต่างกัน (ทัศนีย์ พีรมนตรี, 2530 ; พัชรี ปิยภัณฑ์, 2531) ผลการศึกษาส่วนใหญ่พบว่าข้อสอบวัดความถนัดทางคณิตศาสตร์ที่ลำเอียงส่วนใหญ่จะเข้าข้างเพศชาย และข้อสอบวัดความถนัดทางภาษามักจะเข้าข้างเพศหญิง สำหรับผลการศึกษาการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบระหว่างกลุ่มผู้สอบที่มีภูมิลำเนาหรือสภาพสังคมต่างกันนั้นยังไม่ได้ข้อสรุปที่ชัดเจน กล่าวคือ ทัศนีย์ พีรมนตรี พบว่าข้อสอบที่ลำเอียงมักเข้าข้างผู้สอบในกรุงเทพมหานครในขณะที่ สุพัฒน์ สุกมลสันต์ พบว่า ข้อสอบที่ลำเอียงส่วนใหญ่จะเข้าข้างผู้สอบภาคอื่น ๆ มากกว่าภาคกลาง นอกจากนี้ผลการศึกษาพบข้อความรู้สำคัญอีกประการหนึ่งคือ วิธีการที่ใช้ในการวิเคราะห์ที่แตกต่างกัน จะให้ผลการระบุลักษณะของข้อสอบที่ DIF และจำนวนข้อสอบที่ DIF ได้ไม่ตรงกัน ตลอดจนไม่มีความสัมพันธ์กัน (สุพัฒน์ สุกมลสันต์, 2534 ; กาญจนา วัชนสุนทร, 2537)

ส่วนการศึกษาการเคราะห์การทำหน้าที่ต่างกันของข้อสอบในต่างประเทศ มีการศึกษาทั้งกรณีใช้ข้อมูลจริงและข้อมูลจำลอง สำหรับการศึกษาการเคราะห์ DIF เมื่อใช้ข้อมูลจริงจะวิเคราะห์ลักษณะของข้อสอบที่ DIF ระหว่างกลุ่มผู้สอบจำแนกตามเพศ สีผิว เชื้อชาติ ภาษา ภูมิลำเนา สภาพเศรษฐกิจและสังคม ศาสนา และระดับสติปัญญา เป็นต้น อีกกรณีหนึ่งเป็นการศึกษาโดยใช้ข้อมูลจำลองที่จำลองขึ้นตามทฤษฎี IRT ซึ่งส่วนมากจะมุ่งศึกษา

เปรียบเทียบผลการระบุลักษณะของข้อสอบที่ DIF ในขนาดกลุ่มตัวอย่าง ความยาวแบบสอบ ค่าความยากและค่าอำนาจจำแนกที่แตกต่างกัน และหาความสัมพันธ์ของแต่ละวิธี ผลการศึกษาส่วนใหญ่พบว่าวิธี IRT มีความถูกต้องมากกว่าวิธีอื่น ๆ แต่ก็ยังคงมีความคลาดเคลื่อนอยู่พอสมควร และยังมีข้อจำกัดในเรื่องขนาดกลุ่มตัวอย่าง ทำให้เสียค่าใช้จ่ายมาก นักวิจัยหลายท่านจึงนิยมใช้วิธีที่ได้ผลการวิเคราะห์ที่สอดคล้องกับ วิธี IRT แต่สะดวกกว่า เช่น วิธี MH (Hambleton และคณะ, 1989 ; Swaminatan และ Rojer, 1990) และจากการศึกษาของ Sudweeks และ Tolman (1990) พบว่าข้อสอบที่ทำหน้าที่ต่างกันระหว่างกลุ่มหลายข้อเป็นข้อสอบที่ยาก ต่อมา Clauser และคณะ (1991) ใช้วิธี MH วิเคราะห์ DIF แบบเอกรูปพบว่า วิธี MH ใช้ได้ดีในกรณีที่ข้อสอบมีค่าอำนาจจำแนกสูงและกรณีที่ข้อสอบมีค่าความยากสูงมากจะไม่สามารถตรวจค้นได้นอกจากนี้ Mazor และคณะ (1994) พบว่า วิธี MH สามารถวิเคราะห์ DIF ได้ดีทั้งแบบเอกรูปและแบบอเนกรูป

การวิจัยครั้งนี้ผู้วิจัยเลือกใช้วิธี MH มาทำการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบกับข้อมูลจริงคือแบบสอบวิชาภาษาไทยและภาษาอังกฤษของศูนย์ทดสอบทางการศึกษา เมื่อจำแนกผู้สอบตามเพศ ภูมิลำเนา ประสบการณ์ในการสอบ และสังกัดของสถานศึกษา ทั้งนี้จะศึกษาการวิเคราะห์ DIF ทั้งแบบเอกรูปและแบบอเนกรูป โดยใช้เทคนิคการวิเคราะห์แบบ 2 ขั้นตอนเนื่องจาก Clauser และคณะ (1993) พบว่าผลการตรวจค้น DIF จากสถานการณ์จำลอง โดยใช้เทคนิคแบบ 2 ขั้นตอน พบข้อสอบที่ DIF เท่ากับหรือเหนือกว่าเทคนิคแบบ 1 ขั้นตอน ในทุกเงื่อนไขและเทคนิคแบบ 2 ขั้นตอน ไม่เพิ่มความคลาดเคลื่อนชนิดที่ 1 (Type 1 error rate) มากกว่าเทคนิค 1 ขั้นตอน

สำหรับเกณฑ์ในการจับคู่ผู้สอบ ผู้วิจัยใช้คะแนนรวมที่ผู้สอบทำได้เป็นเกณฑ์และใช้จำนวนชั้นคะแนน (Score Categores) ที่ผู้สอบทำได้สูงสุดเป็นเกณฑ์ในการจับคู่ผู้สอบ ตามข้อเสนอแนะของ Clauser และคณะ (1994)

ในการศึกษาครั้งนี้ผู้วิจัยใช้โปรแกรม MH_{DIF} ที่พัฒนาโดย Fidalgo A. (1995) ทำการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบตามวิธีของแมนเทล-เฮนส์เซล