บทที่ 2

วรรณคดีที่เกี่ยวข้อง

เพื่อนำไปสู่การเปรียบเทียบประสิทธิภาพระหว่างวิธีแมนเทล-แฮนส์เซล(MH)กับวิธีถดถอยโลจิสติก(LR) ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อใช้เกณฑ์จับคู่เปรียบเทียบแตกต่างกันในแบบสอบพหุมิติ ผู้วิจัยได้เสนอสาระสำคัญโดยแบ่งออกเป็น 5 ตอน ดังนี้

ตอนที่ 1 แนวคิดในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

ตอนที่ 2 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมนเทล-แฮนส์เซล

ตอนที่ 3 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีถดถอยโลจิสติก

ตอนที่ 4 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีซิบเทสท์

ตอนที่ 5 เอกสารและงานวิจัยที่เกี่ยวข้อง

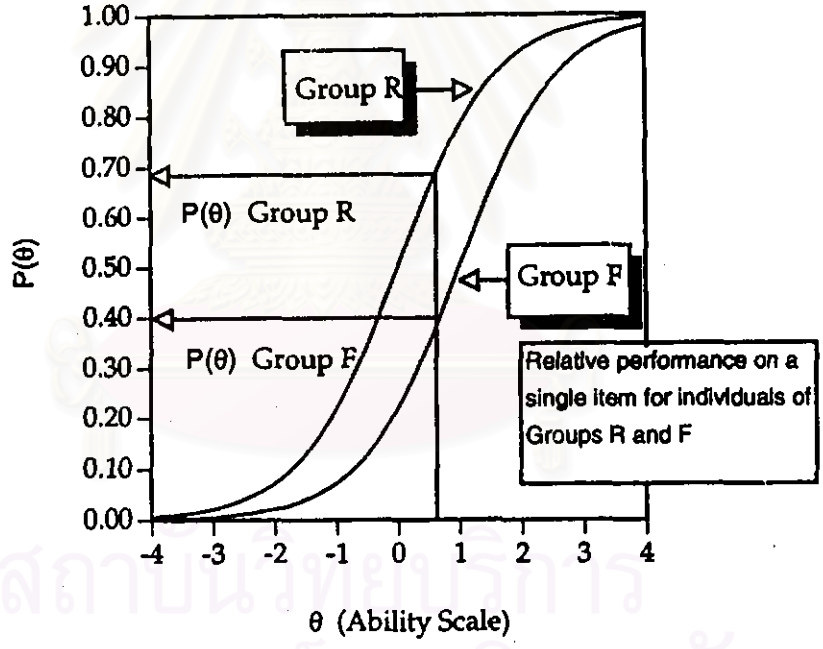
ตอนที่ 1 แนวคิดในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

การที่นำข้อสอบหรือแบบสอบไปใช้กับผู้สอบที่มีความสามารถหลัก (primary abilities) ที่ต้องการวัดหรือคุณลักษณะแฝงเป้าหมายของการวัดเท่ากัน แต่มีคุณลักษณะแฝงอื่นต่างกันแล้วทำให้ผู้ที่มาจากต่างกลุ่มกันมีโอกาสในการตอบข้อสอบได้ถูกต้องไม่เท่ากัน แสดงว่าข้อสอบหรือแบบสอบนั้นขาดความตรงในการแปลคะแนน ทั้งนี้เนื่องจากแบบสอบไม่ได้วัดในมิติหรือลักษณะแฝงเป้าหมายเพียงอย่างเดียวทำให้ผลการวัดเบี่ยงเบนไปจากวัตถุประสงค์ที่กำหนดไว้ ซึ่งการที่ข้อสอบทำหน้าที่ได้ไม่ตรงเช่นนี้มีคำที่ใช้เรียกต่าง ๆ กัน เช่น ความลำเอียงของข้อสอบ(Item Bias) การทำหน้าที่ต่างกันของข้อสอบ หรือ การทำหน้าที่เบี่ยงเบนของข้อสอบ (Differential Item Functioning) แต่เดิมคำว่า "ความลำเอียงของข้อสอบ" เป็นคำที่ใช้ในการศึกษาหรือทดสอบความยุติธรรมของข้อสอบ ต่อมาเกิดความคลุมเครือในการใช้เกณฑ์ในการตัดสินใจเรื่องความลำเอียง จึงสนใจที่จะใช้สารสนเทศทางสถิติมาเป็นเกณฑ์ในการตัดสินแทน โดยพิจารณาจากค่าสถิติ DIF ที่ได้มาจากการตรวจสอบความเป็นพหุมิติของข้อสอบ ซึ่งแสดงว่าข้อสอบมีการทำหน้าที่ต่างกันระหว่างกลุ่มเมื่อการแจกแจงของความสามารถรอง (secondary abilities) แตกต่างกันในกลุ่มผู้สอบที่มีความสามารถหลักเท่ากัน อีกทั้งวิธีที่ใช้ตรวจสอบความลำเอียงในระยะหลังเน้น

ไปที่ความแตกต่างระหว่างกลุ่มผู้สอบที่ตอบสนองต่อข้อสอบข้อเดียวกันในลักษณะที่แตกต่างกัน จึงเปลี่ยนมาใช้คำว่า " การทำหน้าที่ต่างกันของข้อสอบ " ซึ่งเหมาะสมและมีความเป็นกลางมากกว่า (Holland and Thayer, 1988 ; เกษร หว่างจิตร, 2539)

ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ จะพบลักษณะของข้อสอบที่ทำหน้าที่ต่างกัน 2 รูปแบบ คือ

1. **การทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป** (uniform DIF) หมายถึง การที่ข้อสอบทำให้ผู้สอบกลุ่มหนึ่งมีโอกาสในการตอบข้อสอบได้ถูกต้องมากกว่าอีกกลุ่มหนึ่งสม่ำเสมอในทุกระดับความสามารถ โดยเมื่อพิจารณาโค้งคุณลักษณะข้อสอบของผู้สอบทั้งสองกลุ่มจะพบว่าไม่มีปฏิสัมพันธ์ระหว่างโค้งคุณลักษณะข้อสอบในทุกระดับความสามารถ

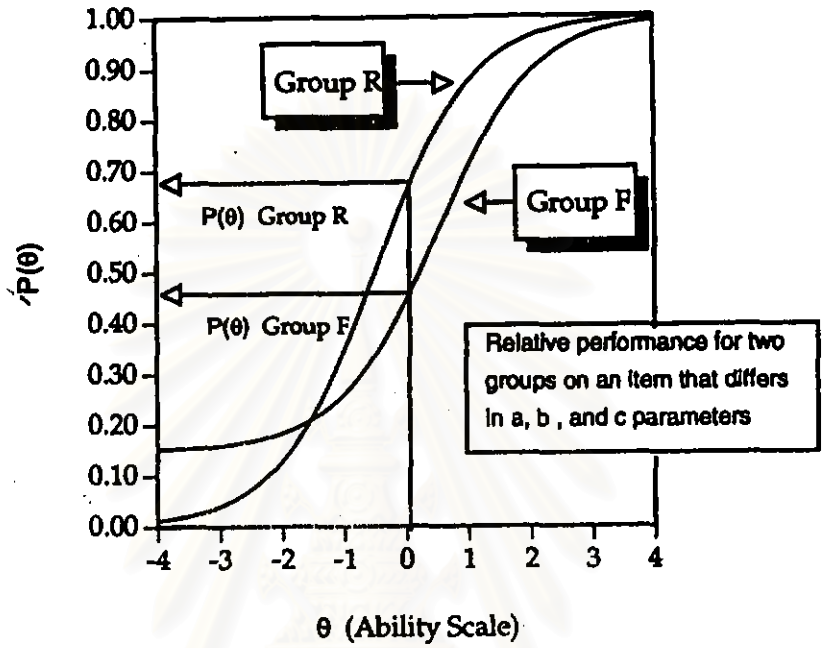


แผนภาพที่ 1 โค้งคุณลักษณะของข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูป

ที่มา : Camilli, G. & Shepard, L. A., 1994: 59

2. **การทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป** (nonuniform DIF) หมายถึง การที่ข้อสอบทำให้โอกาสในการตอบข้อสอบได้ถูกต้องของผู้สอบระหว่างกลุ่มไม่สม่ำเสมอกันในทุกระดับความสามารถ เมื่อพิจารณาโค้งคุณลักษณะข้อสอบของทั้งสองกลุ่มจะพบว่ามีปฏิสัมพันธ์ระหว่างโค้งคุณลักษณะข้อสอบ เช่น ที่ความสามารถระดับหนึ่ง กลุ่ม A มีโอกาสในการตอบ

ข้อสอบถูกมากกว่ากลุ่ม B แต่ที่ความสามารถอีกระดับหนึ่ง กลุ่ม B กลับมีโอกาสในการตอบข้อสอบได้ถูกมากกว่ากลุ่ม A เป็นต้น



แผนภาพที่ 2 โค้งคุณลักษณะของข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูป

ที่มา : Camilli, G. & Shepard, L. A., 1994: 60

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF detection) เป็นการเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มผู้สอบ 2 กลุ่มที่มีความสามารถหลัก(primary abilities) ที่มุ่งวัดเท่ากัน โดยกลุ่มแรกคือ กลุ่มอ้างอิง (reference group) ซึ่งเป็นกลุ่มที่คาดว่าจะได้ประโยชน์จากการตอบข้อสอบหรือเป็นกลุ่มที่มีโอกาสตอบข้อสอบได้ถูกมากกว่า และกลุ่มที่สองคือ กลุ่มเปรียบเทียบ (focal group)เป็นกลุ่มที่เราสนใจศึกษาและคาดว่าจะเสียประโยชน์ในการตอบข้อสอบ ในการตรวจสอบ DIF จะเป็นการเปรียบเทียบผลการตอบข้อสอบหลังจากการจับคู่ (matching) ผู้สอบตามความรู้หรือความสามารถของผู้สอบ การจับคู่ของผู้สอบสองกลุ่มเป็นเงื่อนไขที่สำคัญเนื่องจากเกณฑ์การจับคู่ (matching criteria) เป็นเกณฑ์ที่ใช้แทนความสามารถที่แท้จริงของผู้สอบสองกลุ่ม โดยทั่วไปแล้วมักใช้คะแนนรวมของแบบสอบ (total test score) เป็นเกณฑ์ในการจับคู่เนื่องจากเกี่ยวข้องโดยตรงกับความรู้หรือความสามารถที่วัดได้ และแบบสอบนั้นสามารถตรวจสอบความตรงและความเที่ยงของแบบสอบได้อีกทั้งผู้สอบทุกคนสอบภายใต้สถานการณ์เดียวกัน แต่จุดอ่อนของการใช้คะแนนรวมของแบบสอบเป็นเกณฑ์การจับคู่ผู้สอบ คือ มีการรวมเอาคะแนน

จากข้อสอบที่ทำหน้าที่ต่างกันมาเป็นเกณฑ์ในการจับคู่ผู้สอบด้วย ในการแก้ไขจุดอ่อนนี้ Holland และThayer (1988) ได้เสนอให้ใช้วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้เทคนิค 2 ขั้นตอน ซึ่งเรียกว่า การทำให้เกณฑ์การจับคู่ผู้สอบมีความบริสุทธิ์ (purification of matching criterion)

ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ ได้มีการศึกษาใน 2 แนวทางใหญ่ ๆ โดยแปลตามปัจจัยหรือเกณฑ์ที่ใช้จับคู่เปรียบเทียบ (Rudner Getson and Knight, 1980) คือ

1. **การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยใช้เกณฑ์ภายนอก** (External Criterion) การวิเคราะห์ความลำเอียงของข้อสอบโดยวิธีนี้สามารถวิเคราะห์ได้ทั้งรายข้อและรายฉบับ ซึ่งวิธีการวิเคราะห์จะเขียนกราฟแสดงความสัมพันธ์ระหว่างคะแนนที่ใช้เป็นตัวแปรเกณฑ์ภายนอกกับตัวแปรทำนาย วิธีนี้มีจุดมุ่งหมายเพื่อวิเคราะห์ความถดถอยของตัวแปรทั้งสอง แล้วเปรียบเทียบค่าความชัน (slope) และค่าการตัดแกน (intercept) ของเส้นกราฟระหว่างกลุ่มผู้สอบ 2 กลุ่ม ในการวิเคราะห์ความลำเอียงของข้อสอบทั้งฉบับจะใช้คะแนนรวมของข้อสอบฉบับนั้นจากผู้สอบแต่ละคนเป็นตัวแปรทำนาย แต่ถ้าเป็นการวิเคราะห์ความลำเอียงของข้อสอบเป็นรายข้อจะใช้ค่าความยาก(p)ของข้อสอบแต่ละข้อเป็นตัวแปรทำนาย ส่วนตัวแปรที่ใช้เป็นเกณฑ์ภายนอกของการวิเคราะห์ความลำเอียงของข้อสอบรายข้อหรือรายฉบับจะใช้คะแนนรวมหรือเกรดเฉลี่ยหรือผลสัมฤทธิ์ของงานบางอย่างที่ให้ทำ (Cronbach, 1970)

ดังนั้นการวิเคราะห์ความลำเอียงของข้อสอบด้วยวิธีนี้ จึงเป็นการเปรียบเทียบเส้นกราฟที่แสดงความสัมพันธ์ระหว่างตัวแปรเกณฑ์กับตัวแปรทำนายของกลุ่มผู้สอบที่ต้องการวิเคราะห์ความลำเอียงของข้อสอบ ถ้าเส้นกราฟดังกล่าวมีค่าความชันและค่าการตัดแกนแตกต่างกันในแต่ละกลุ่มแล้วข้อสอบข้อนั้น หรือแบบสอบฉบับนั้นก็จะมีความลำเอียงต่อกลุ่มผู้สอบที่มีค่าการตัดแกน หรือค่าความชันมากกว่า

จะเห็นว่าการวิเคราะห์ความลำเอียงของข้อสอบตามวิธีการที่กล่าวมา มีจุดอ่อนตรงที่ในทางปฏิบัติแล้วเป็นการยากมากที่จะหาตัวแปรเกณฑ์ภายนอกที่มาจากแบบทดสอบที่มีความเที่ยงตรงเชิงพยากรณ์และมีความยุติธรรมสำหรับกลุ่มที่นำมาวิเคราะห์ความลำเอียง ถ้าหากตัวแปรเกณฑ์ภายนอกขาดคุณสมบัติดังกล่าว จะทำให้ผลการวิเคราะห์ความลำเอียงของข้อสอบตามวิธีนี้ขาดความสมบูรณ์และถูกต้องเท่าที่ควร

2. **การวิเคราะห์ความลำเอียงของข้อสอบเมื่อไม่มีเกณฑ์ภายนอก** (Absence of Criterion) หรือการวิเคราะห์ความลำเอียงของข้อสอบโดยใช้เกณฑ์ภายใน เป็นการนำวิธีการทางสถิติมาตรวจสอบความลำเอียงโดยพิจารณาโครงสร้างภายในของแบบสอบโดยยึดหลักการที่ว่า

ข้อสอบจะมีความยุติธรรมไม่ลำเอียงต่อกลุ่มผู้สอบ ก็ต่อเมื่อมีผู้สอบที่อยู่ต่างกลุ่มกันมีความสามารถเท่ากันแล้วจะมีคะแนนจริงของผลการสอบเท่ากัน แต่ถ้าข้อสอบใดที่ผู้สอบมีระดับความสามารถเท่ากัน แต่อยู่ต่างกลุ่มกันมีคะแนนจริงของผลการสอบแตกต่างกันแสดงว่าข้อสอบนั้นมีความลำเอียง การวิเคราะห์ความลำเอียงโดยวิธีการนี้อาจแบ่งออกได้เป็น 3 วิธีใหญ่ ๆ (Laksana and Coffman, 1980) ดังนี้

1. ใช้หลักการวัดความเบี่ยงเบนสัมพันธ์ (Relative Deviation) ของข้อสอบแต่ละข้อ จากแนวโน้มเข้าสู่ส่วนกลาง วิธีนี้มีข้อตกลงเบื้องต้นว่า ข้อสอบส่วนใหญ่ในแบบทดสอบมีความเป็นเอกพันธ์ในการวัดความสามารถใดความสามารถหนึ่งในผู้สอบต่างกลุ่มกัน ข้อสอบข้อใดที่เบี่ยงเบนไปจากส่วนกลางมากกว่าที่คาดหวังไว้ ก็อาจตั้งข้อสงสัยได้ว่าเป็นข้อสอบที่มีความลำเอียง วิธีการหนึ่งที่ใช้กันบ่อย ๆ ก็คือการวิเคราะห์ความแปรปรวน (ANOVA) ข้อสอบที่ลำเอียงคือข้อสอบที่มีปฏิสัมพันธ์ระหว่างข้อกับกลุ่มผู้สอบมีนัยสำคัญทางสถิติ ส่วนอีกวิธีหนึ่ง ได้แก่ การเขียนกราฟ วิธีการนี้ค่าพารามิเตอร์เกี่ยวกับค่าความยาก เช่น ค่า P ค่าเดลตา หรือค่า b จากกลุ่มหนึ่งจะลงจุดคู่อันดับ คู่กับค่าพารามิเตอร์จากอีกกลุ่มหนึ่ง ข้อสอบข้อใดที่เบี่ยงเบนไปจากเส้นแกนหลักมากถือว่าเป็นข้อสอบที่มีความลำเอียง

2. ใช้หลักการประเมินความเที่ยงตรงเชิงโครงสร้าง (Construct Validity) ของแบบทดสอบ วิธีการหนึ่งที่ใช้กันมาก คือ การวิเคราะห์องค์ประกอบ ดัชนีความลำเอียงของข้อสอบตามวิธีนี้ก็คือความแตกต่าง (discrepancy) ของน้ำหนักองค์ประกอบ (factor loading) จากผู้สอบต่างกลุ่มกัน หรือความแตกต่างอย่างมีนัยสำคัญระหว่างค่าเฉลี่ยของคะแนนองค์ประกอบ (factor scores) จากแต่ละกลุ่มที่นำมาเปรียบเทียบกัน ดัชนีความลำเอียงที่มาเป็นตัวชี้ว่าข้อสอบไม่ได้วัดสิ่งเดียวกันในผู้สอบต่างกลุ่มกัน

3. ใช้หลักการประมาณค่าโอกาสในการตอบข้อสอบแต่ละข้อถูก วิธีการที่ใช้กันทั่ว ๆ ไป คือ วิธีไคสแควร์ (Chi-square) และวิธีโค้งลักษณะข้อสอบ (Item Characteristic Curve : ICC) วิธีการทั้งสองนี้มีความคล้ายคลึงกันในแง่การใช้โอกาสในการตอบถูกที่แตกต่างกันจากผู้สอบ 2 กลุ่มหรือมากกว่าเป็นดัชนีความลำเอียงของข้อสอบ สำหรับข้อแตกต่างระหว่างวิธีการทั้งสองก็คือวิธีไคสแควร์จะประมาณความสามารถของผู้สอบโดยใช้คะแนนดิบ ส่วนวิธีโค้งลักษณะข้อสอบจะประมาณความสามารถของผู้สอบจากคุณลักษณะแฝง

วิธีในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ส่วนใหญ่จะนิยมใช้แนวทางในการวิเคราะห์โดยไม่ใช้เกณฑ์ภายนอก ทั้งนี้เพราะในการหาตัวแปรเกณฑ์ภายนอกให้มีความตรงเชิง

พยากรณ์และยุติธรรมนั้นเป็นสิ่งที่ทำได้ยากมาก ถ้าตัวแปรเกณฑ์ขาดคุณสมบัติดังกล่าวจะทำให้ผลการวิเคราะห์ความลำเอียงมีความคลาดเคลื่อนเกิดขึ้น ในปัจจุบันวิธีที่ใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) มีหลายวิธี Dorans and Potenza(1995) ได้จำแนกประเภทของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในลักษณะที่มีการให้คะแนนแบบ 2 ค่า (dichotomously score : โดยการให้คะแนนเป็น 1 เมื่อตอบถูก และให้คะแนนเป็น 0 เมื่อตอบผิด) ออกเป็น 2 กลุ่ม คือ

กลุ่มที่ 1 เป็นกลุ่มที่ใช้คะแนนที่สังเกตไม่ได้หรือตัวแปรแฝง (Latent Variable) ได้แก่ วิธีที่มีพื้นฐานจากทฤษฎีการตอบสนองข้อสอบ(IRT) และวิธี SIBTEST

กลุ่มที่ 2 เป็นกลุ่มที่ใช้คะแนนที่สังเกตได้ (Observe Score) ได้แก่ วิธี Mantel-Haenzel, วิธี Standardization และวิธี Logistic Regression

วิธีที่ใช้ทฤษฎีการตอบสนองข้อสอบ (IRT) สามารถวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบได้โดยการเปรียบเทียบฟังก์ชั่นการตอบข้อสอบระหว่างกลุ่มผู้สอบที่ต้องการศึกษา ถ้าฟังก์ชั่นการตอบสนองข้อนั้นของกลุ่มผู้สอบ 2 กลุ่มไม่เหมือนกันแสดงว่าข้อสอบนั้นทำหน้าที่แตกต่างกัน (Kim et al.,1944) หรือโดยการเปรียบเทียบความแตกต่างระหว่างโค้งคุณลักษณะข้อสอบระหว่างกลุ่ม ซึ่งพื้นที่ระหว่างโค้งคุณลักษณะข้อสอบจะเป็นดัชนีบอกระดับของการทำหน้าที่ต่างกันของข้อสอบ(Osterlind, 1992) ถ้าพบว่าโค้งคุณลักษณะของข้อสอบมีช่องว่างหรือมีพื้นที่ระหว่างโค้ง 2 โค้ง แสดงว่าข้อสอบข้อนั้นทำหน้าที่แตกต่างกัน กระบวนการตรวจสอบ DIF ด้วยวิธีนี้ค่อนข้างยุ่งยากซับซ้อน เสียค่าใช้จ่ายสูง กลุ่มตัวอย่างต้องมีขนาดใหญ่ การนำไปใช้อาจเกิดปัญหา เพราะข้อมูลจริงมักจะไม่เป็นไปตามโมเดล IRT ซึ่งทำให้การประมาณค่าของพารามิเตอร์ต่าง ๆ ทำได้ไม่ดี (Shepard and Camilli, 1994)

วิธีแมนเทล-แฮนส์เซล (MH) เป็นวิธีที่มีพื้นฐานบนวิธีไคสแควร์ โดยเปรียบเทียบความน่าจะเป็นที่จะตอบข้อสอบได้ถูกต้องระหว่างผู้สอบ 2 กลุ่มที่มีความสามารถเท่าเทียมกัน วิธีการเปรียบเทียบจะใช้คะแนนรวมเป็นเกณฑ์ในการเปรียบเทียบระหว่างผู้สอบ 2 กลุ่ม คือ กลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ซึ่งจะมีการตรวจสอบทุกๆ ระดับคะแนนรวมจากแบบสอบ นำสัดส่วนของผู้ที่ตอบผิดมาเขียนในรูปตาราง 2 ทาง คำนวณหาค่าอัลฟา (α_{MH}) แล้วทดสอบสมมติฐานด้วยค่าสถิติไคสแควร์ (χ^2_{MH}) วิธี MH ใช้ได้เหมาะสมกับข้อมูลจริงหรือข้อมูลเชิงประจักษ์เนื่องจากสามารถคำนวณได้ง่าย ประหยัดเวลาและค่าใช้จ่าย สามารถใช้กับกลุ่มตัวอย่างขนาดเล็กได้และสามารถคำนวณได้ถูกต้อง แม้ว่าจำนวนประชากรในกลุ่มย่อยที่ศึกษาจะมีขนาดต่างกัน ซึ่งในทาง

ปฏิบัติหรือข้อมูลจริงจะพบว่าจำนวนประชากรในกลุ่มย่อยย่อมไม่เท่ากัน (Holland and Thayer, 1988; Mazor และคณะ, 1994)

วิธีการทดสอบความแปรปรวน (Analysis of variance : ANOVA) เป็นวิธีที่พิจารณาผลจากปฏิสัมพันธ์ระหว่างกลุ่มผู้สอบย่อยและข้อสอบ ในการวิเคราะห์จะแยกความแปรปรวนเป็นส่วนๆ คือ ความแปรปรวนจากข้อสอบจากกลุ่มผู้สอบจากปฏิสัมพันธ์ระหว่างกลุ่มผู้สอบกับผู้สอบและจากความแปรปรวนภายในกลุ่ม แล้วทำการทดสอบสมมติฐาน ถ้าปฏิสัมพันธ์ระหว่างกลุ่มผู้สอบและข้อสอบมีนัยสำคัญทางสถิติ แสดงว่าแบบสอบฉบับนั้นทำหน้าที่แตกต่างกันต้องวิเคราะห์รายข้อต่อข้อที่แตกต่างกัน (กาญจนา วัชนสุนทร, 2538)

วิธีซิบเทสท์ (SIBTEST) เป็นวิธีการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ / แบบสอบชนิดพหุมิติ(multidimensional test) โดยมีพื้นฐานบนทฤษฎีการตอบข้อสอบแบบพหุมิติ ซึ่งรูปแบบการวิเคราะห์ข้อมูลมีข้อตกลงว่ามีมิติการวัด 2 มิติ ได้แก่ ความสามารถหรือลักษณะแฝงเป้าหมายที่ต้องการวัด (θ) และลักษณะแฝงแทรกซ้อนที่ไม่ต้องการวัด (η) ดังนั้นเวคเตอร์ความสามารถคือ(θ,η) มีฟังก์ชั่นการตอบข้อสอบคือ $P(\theta,\eta)$ โดยที่ข้อสอบทุกข้อจะวัดความสามารถเป้าหมาย(θ) และบางข้อ(ซึ่งมีความลำเอียง)จะวัดทั้งความสามารถเป้าหมายและความสามารถซ้อนโดยใช้คะแนนรวมจากแบบสอบเป็นเกณฑ์ในการจับคู่เปรียบเทียบ แล้วคำนวณค่าเบต้า(β) ทดสอบสมมติฐานศูญด้วย Z-test วิธีนี้คำนวณได้ง่ายไม่ซับซ้อนใช้ได้ดีกันการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป (unifrom DIF) และมีทิศทางเดียว(unidirectional) ข้อจำกัดของวิธีนี้คือกลุ่มตัวอย่างต้องมีขนาดใหญ่พอและเสียค่าใช้จ่ายค่อนข้างสูง (Shealy and Stout, 1993)

วิธีลอจิสติก (LR) เป็นวิธีการทดสอบสถิติแบบพารามิเตอร์มีการทดสอบโมเดลทางสถิติ Logistic Regression วิธีนี้สามารถใช้วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบได้ดีทั้งแบบเอกรูปและอเนกรูป แต่จะเสียค่าใช้จ่ายมากกว่าวิธี MH ประมาณ 3-4 เท่า (Swaminathan and Rojer, 1990)

ตอนที่ 2 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมนเทล-แฮนส์เซล (Mantel-Haenszel)

วิธีแมนเทล-แฮนส์เซล (MH) เป็นวิธีที่ Mantel และ Haenszel ได้เสนอขึ้นใช้ในปี 1959 โดยพัฒนามาจากวิธีไคสแควร์ซึ่งเป็นวิธีที่อาศัยทฤษฎีการวัดแบบดั้งเดิมเป็นพื้นฐานและ Holland and Thayer ได้เริ่มนำมาใช้เพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ซึ่งต่อมาถูกนำมาใช้อย่างแพร่หลายเนื่องจากเป็นวิธีที่ใช้ง่าย สะดวก และประหยัด (Holland and Thayer, 1988)

หลักการของวิธี MH เป็นการเปรียบเทียบผลการตอบข้อสอบของผู้สอบ 2 กลุ่ม ได้แก่ กลุ่มอ้างอิง (reference group) ซึ่งเป็นกลุ่มที่คาดว่าจะได้ประโยชน์จากการตอบข้อสอบ และกลุ่มเปรียบเทียบ (focal group) เป็นกลุ่มที่เสียประโยชน์จากการตอบข้อสอบ ในการตรวจสอบ DIF แต่ละครั้งจะเรียกข้อสอบที่ถูกตรวจว่า ข้อสอบที่ศึกษา (studied item) โดยจะเปรียบเทียบในทุก ๆ ระดับคะแนนรวมที่ได้จากผลการตอบข้อสอบของผู้สอบ ข้อสอบใดที่ผู้สอบในกลุ่มที่มีความสามารถเท่ากันทำได้ถูกต้องเท่ากัน แสดงว่า เป็นข้อสอบที่ทำหน้าที่ไม่ต่างกัน (no DIF) ระหว่างกลุ่มผู้สอบ การตรวจสอบด้วยวิธีนี้ทำได้โดยแยกวิเคราะห์ข้อสอบเป็นรายข้อ ซึ่งในแต่ละข้อจะต้องสร้างตารางไขว้ขนาด 2x2 แสดงความถี่ของผู้สอบที่ตอบถูก/ผิด ในกลุ่มอ้างอิง(R) และกลุ่มเปรียบเทียบ(F) ตามช่วงคะแนนของผู้สอบสำหรับช่วงคะแนน j ถ้ามีการแบ่งช่วงคะแนนการสอบเป็น k ช่วง จะได้ตาราง 2x2 จำนวน k ตาราง

ตารางที่ 2 แสดงความถี่ของผู้สอบที่ตอบข้อสอบถูกและผิดในกลุ่มอ้างอิงและกลุ่มเปรียบเทียบสำหรับช่วงคะแนน j

กลุ่มผู้สอบ	คะแนนที่ได้จากข้อสอบที่ต้องการวิเคราะห์ DIF		
	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} จำนวนผู้สอบที่ตอบถูกทั้งหมดในช่วงคะแนน j

N_j จำนวนผู้สอบทั้งหมด

โดยแสดงเป็นสัดส่วนการตอบข้อสอบของผู้สอบ 2 กลุ่มได้ดังตารางที่ 3

ตารางที่ 3 แสดงสัดส่วนการตอบข้อสอบของกลุ่มประชากร

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

เมื่อ 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. คำนวณค่าความน่าจะเป็นในรูปของสัดส่วนของการตอบข้อสอบถูกและผิดระหว่างกลุ่ม ในทุกช่วงคะแนน จากสูตร

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

เมื่อ αMH สัดส่วนของการตอบข้อสอบถูกและผิดระหว่างกลุ่มในแต่ละข้อในทุกช่วงคะแนน j

2. ทดสอบนัยสำคัญของค่าสถิติไคสแควร์ เพื่อทดสอบค่า αMH ที่คำนวณได้ว่าแตกต่างจาก 1 อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05 หรือไม่ ที่ระดับชั้นความเป็นอิสระเท่ากับ 1 ตามสูตร ดังนี้

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

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

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

3. Holland และ Thayer เสนอแนะให้แปลงค่า αMH ให้เป็นค่าเดลต้า (ΔMH หรือ MH_{DIF}) ตามสูตร

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

เกณฑ์ในการตัดสินข้อสอบที่ทำหน้าที่ต่างกัน พิจารณาจากข้อสอบที่มีค่าαMH แตกต่างจาก 1 อย่างมีนัยสำคัญทางสถิติ หรือค่า MH_{DIF} แตกต่างจาก 0 อย่างมีนัยสำคัญทางสถิติและมีเกณฑ์ในการพิจารณาค่าของ MH_{DIF} ดังนี้

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

2. ค่า MH_{DIF} แตกต่างจาก 0 อย่างมีนัยสำคัญทางสถิติ และมีค่าเป็นบวก แสดงว่าข้อสอบนั้นทำหน้าที่ต่างกัน โดยจะลำเอียงเข้าหากลุ่มเปรียบเทียบ

3. ค่า MH_{DIF} แตกต่างจาก 0 อย่างมีนัยสำคัญทางสถิติ และมีค่าเป็นลบ แสดงว่าข้อสอบนั้นทำหน้าที่ต่างกัน โดยจะลำเอียงเข้าหากลุ่มอ้างอิง

ตอนที่ 3 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีถดถอยโลจิสติก (Logistic Regression)

วิธีถดถอยโลจิสติกถูกนำมาประยุกต์ใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดย Swaminathan และ Rogers ในปี 1990 ซึ่งพัฒนามาจากสมการ logliear ของ Mellenberg (1982) เป็นวิธีที่มีสมการทางคณิตศาสตร์ที่ใช้กลุ่มผู้สอบ (group) ความสามารถ (ability) และปฏิสัมพันธ์ระหว่างกลุ่มผู้สอบและความสามารถ (group by ability) ทำนายความน่าจะเป็นของการตอบข้อสอบว่าถูก (1) หรือผิด (0) (Camili, G. and Shepard, L.A., 1994) จึงทำให้วิธีนี้มีประสิทธิภาพในการตรวจสอบ DIF ได้ดีทั้งแบบเอกรูปและอเนกรูป

วิธีถดถอยโลจิสติกเป็นวิธีที่ตั้งอยู่บนพื้นฐานของโมเดลทางสถิติ Logisic Regression ซึ่งเป็นการวิเคราะห์สมการทำนายเมื่อต้องการศึกษาผลของตัวแปรทำนาย (predictor variable) ที่มีต่อตัวแปรเกณฑ์ ซึ่งเป็นทวิวิภาค (dichotomous variable) โดยใช้ฟังก์ชันโลจิสติก (logistic

function) ในการแสดงความสัมพันธ์ระหว่างค่าของตัวแปรทำนายกับค่าความน่าจะเป็นของการเกิดเหตุการณ์ตามตัวแปรเกณฑ์ (ศิริชัย กาญจนวาสี, 2541)

การตรวจสอบการทำหน้าที่ต่างกันด้วยวิธีถดถอยโลจิสติก (LR) ใช้ผู้สอบเป็นหน่วยของการวิเคราะห์โมเดลทางสถิติ Logistic Regression เพื่อทำนายโอกาสในการตอบข้อสอบของผู้สอบ โดยสมการนี้ คือ

$$P(u_i = 1) = \frac{\exp \psi_i}{1 + \exp \psi_i}$$

$$\psi_i = \delta + \tau_1 G_i + \tau_2 X_i + \tau_3 (G_i X_i)$$

โดยที่ ψ_i คือ ค่าคาดหวังของ log odds ratio

u_i คือ ผลการตอบข้อสอบของผู้สอบคนที่ i (u_i =1 / 0)

X_i คือ คะแนนสอบทั้งหมดของผู้สอบคนที่ i

G_i คือ กลุ่มของผู้สอบคนที่ i (G_i = 1 / 2)

จากสมการถดถอยโลจิสติก ค่า τ_1 เป็นค่าที่ชี้ให้เห็นถึงโอกาสในการตอบข้อสอบได้ถูกระหว่างกลุ่มผู้สอบ ถ้าพบว่า ค่า τ_1 แตกต่างจากศูนย์อย่างมีนัยสำคัญ แสดงว่า มีความน่าจะเป็นในการตอบข้อสอบได้ถูกแตกต่างกันระหว่างผู้สอบทั้งสองกลุ่ม เมื่อควบคุมความสามารถของผู้สอบ ส่วนค่า τ_2 จะชี้ให้เห็นถึงความแตกต่างของความสามารถระหว่างกลุ่มผู้สอบมีผลต่อโอกาสในการตอบข้อสอบได้ถูก โดยปกติพบว่า ค่า τ_2 มักจะแตกต่างจากศูนย์อย่างมีนัยสำคัญ ทั้งนี้เนื่องจากผู้สอบที่มีคะแนนสอบสูงย่อมมีโอกาสตอบข้อสอบได้ถูกต้องมากกว่าผู้สอบที่มีคะแนนสอบต่ำกว่า และค่า τ_3 จะแสดงถึงปฏิสัมพันธ์ระหว่างกลุ่มผู้สอบและความสามารถของผู้สอบ ถ้าพบว่าค่า τ_3 แตกต่างจากศูนย์อย่างมีนัยสำคัญ แสดงว่า ข้อสอบนั้นทำให้คะแนนระหว่างกลุ่มผู้สอบในระดับความสามารถหนึ่งมีความแตกต่างกันมากกว่าระดับความสามารถอื่น นั่นคือ การเกิด DIF แบบอเนกรูป (nonuniform DIF)

เนื่องจากการตรวจ DIF ด้วยวิธี LR เป็นวิธีที่ใช้สมการทางคณิตศาสตร์จึงมีความยืดหยุ่นสูงในการนำตัวแทนความสามารถอื่นมาใช้แทนคะแนนรวม (total test score) ซึ่งโดยทั่วไปมักใช้เป็นตัวแทนความสามารถในการจับคู่เปรียบเทียบ เช่น ใช้ค่าประมาณความสามารถในลักษณะอื่นนอกจากคะแนนสอบ หรืออาจเป็นตัวแปรเชิงคุณภาพก็ได้

การประมาณค่าพารามิเตอร์ของสมการถดถอยโลจิสติกประมาณค่าโดยใช้วิธีแมกซิมัม-ไลลิฮูด (maximum likelihood) ดังนั้น การทดสอบนัยสำคัญทางสถิติซึ่งมีสมมติฐานที่เปรียบเทียบสมการทางคณิตศาสตร์หลายสมการ จะถูกทดสอบด้วยสถิติไลลิฮูด(likelihood ratio statistics) โดยมีสมการเต็มรูป ดังนี้

$$\psi_i = \delta + \tau_1 G_i + \tau_2 X_i + \tau_3(G_i X_i) \quad \text{.....(1)}$$

ส่วนสมการที่ 2 เป็นสมการที่แสดงการเกิด DIF แบบเอกรูป (uniform DIF) ซึ่งไม่มีส่วนที่เป็นปฏิสัมพันธ์เข้ามาเกี่ยวข้อง เขียนสมการได้ดังนี้

$$\psi_i = \delta + \tau_1 G_i + \tau_2 X_i \quad \text{.....(2)}$$

สำหรับสมการสุดท้าย เป็นสมการที่แสดงการไม่เกิด DIF จึงพิจารณาเพียงระดับความสามารถเท่านั้น เขียนสมการได้ดังนี้

$$\psi_i = \delta + \tau_2 X_i \quad \text{.....(3)}$$

จากทั้งสามสมการข้างต้นจะถูกนำไปเป็นพื้นฐานในการทดสอบสมมติฐาน มีขั้นตอนในการพิจารณา ดังนี้

ขั้นที่ 1 สมการที่ 1 จะถูกทดสอบเพื่อคัดค้านกับสมการที่ 2 เมื่อเปรียบเทียบทั้งสองสมการพบว่า สมการที่ 1 มีพารามิเตอร์ที่ต่างไปจากสมการที่ 2 คือ เทอมของ $\tau_3(G_i X_i)$ นั่นคือ ในขั้นตอนนี้ต้องการพิสูจน์ว่า สมการนี้จะเป็นจริงได้ย่อมเกิดจากอิทธิพลของปฏิสัมพันธ์

การทดสอบสมมติฐานในขั้นที่ 1 หากพบว่า เกิดนัยสำคัญทางสถิติ แสดงว่าปฏิสัมพันธ์ระหว่างกลุ่มผู้สอบและความสามารถ (group by ability) มีผลต่อโอกาสในการตอบข้อสอบได้ถูกระหว่างกลุ่มผู้สอบที่มีความสามารถเท่ากัน นั่นคือ การเกิด DIF แบบอเนกรูป (nonuniform DIF) แต่ถ้าไม่เกิดนัยสำคัญทางสถิติ จึงดำเนินการต่อในขั้นที่ 2

ขั้นที่ 2 สมการที่ 2 จะถูกทดสอบเพื่อคัดค้านกับสมการที่ 3 เมื่อเปรียบเทียบทั้งสองสมการพบว่า สมการที่ 2 มีพารามิเตอร์ที่ต่างไปจากสมการที่ 3 คือ เทอมของ $\tau_1 G_i$ นั่นคือ ในขั้นตอนนี้ต้องการพิสูจน์ว่า กลุ่มผู้สอบ (group) มีอิทธิพลต่อโอกาสในการตอบข้อสอบได้ถูกในทุกระดับความสามารถของผู้สอบ

การทดสอบสมมติฐานในขั้นที่ 2 หากพบว่า เกิดนัยสำคัญทางสถิติ แสดงว่า กลุ่มผู้สอบ (group)มีผลต่อโอกาสในการตอบข้อสอบถูกระหว่างกลุ่มผู้สอบที่มีความสามารถเท่ากัน นั่นคือ

การเกิด DIF แบบเอกรูป(nonuniform DIF) แต่ถ้าไม่เกิดนัยสำคัญทางสถิติ แสดงว่าไม่เกิด DIF อธิบายได้ว่า ปฏิสัมพันธ์ระหว่างกลุ่มผู้สอบและความสามารถ (group by ability) และ กลุ่มผู้สอบ (group) ไม่มีผลทำให้เกิดความแตกต่างของโอกาสในการตอบข้อสอบถูกของกลุ่มผู้สอบ ดังนั้น หากพบว่ามีความแตกต่างนี้ขึ้น ย่อมเกิดจากความแตกต่างของระดับความสามารถของผู้สอบซึ่งในการทดสอบโดยทั่วไปมักพบว่าความสามารถมีผลต่อโอกาสในการตอบข้อสอบถูกของผู้สอบแตกต่างกัน

ตอนที่ 4 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีซิบเทสท์ (SIBTEST)

Shealy และStoutได้เสนอวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในปี ค.ศ. 1991 ซึ่งมีแนวคิดในการตรวจสอบความลำเอียงของแบบสอบชนิดพหุมิติโดยมีพื้นฐานอยู่บนทฤษฎีการตอบข้อสอบแบบพหุมิติ และมีข้อตกลงว่ามีมิติการวัด 2 มิติ คือ มิติลักษณะแฝงเป้าหมายที่ต้องการวัด (θ) และลักษณะแฝงแทรกซ้อนที่ไม่ต้องการวัด (η) มีฟังก์ชั่นการตอบข้อสอบ คือ $P(\theta,\eta)$ โดยที่ข้อสอบทุกข้อจะวัดความสามารถเป้าหมาย(θ) และบางข้อ (ซึ่งมีความลำเอียง) จะวัดทั้งคุณลักษณะแฝงเป้าหมายและคุณลักษณะแฝงแทรกซ้อน

การตรวจสอบเป็นการเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ซึ่งผลการตอบข้อสอบนั้นมีอยู่ 2 ค่า โดยได้จากแบบสอบย่อย 2 ชุด นั่นคือ ในการตรวจสอบด้วยวิธี SIBTEST จะแบ่งแบบสอบจากเดิมที่มี 1 ชุด ให้เป็น 2 ชุดย่อย คือ แบบสอบย่อยที่มีความตรง เป็นแบบสอบที่ประกอบด้วยข้อสอบที่มีความตรง วัดได้ตามที่ต้องการ และแบบสอบที่ต้องการศึกษา เป็นแบบสอบที่ประกอบด้วยข้อสอบที่สงสัยว่าจะทำหน้าที่ต่างกันโดยที่

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

U_i คือผลการตอบข้อที่ i ได้ 1 ถ้าตอบถูก ได้ 0 ถ้าตอบผิด

$Y = \sum_{i=1}^{n} U_i$ โดยที่ Y คือคะแนนรวมจากแบบที่ศึกษา

ในการวิเคราะห์จะพิจารณาจากผลการตอบข้อสอบจากแบบสอบทั้ง 2 ชุดย่อย โดยผู้ได้คะแนนรวมเท่ากันจากแบบสอบที่มีความตรงของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบมาจับคู่เปรียบเทียบและใช้คะแนนรวมจากแบบสอบที่ศึกษาของผู้สอบเหล่านี้ในการคำนวณ

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

โดย k คือระดับคะแนนรวมจากแบบสอบที่มีความตรงของผู้สอบ

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

$\overline{Y}_{Fk}$ คือคะแนนเฉลี่ยจากแบบสอบที่ศึกษาของผู้สอบกลุ่มเปรียบเทียบทุกคนที่มีคะแนนจากแบบสอบที่มีความตรงที่ระดับ k

ดังนั้น $(\overline{Y}_{Rk} - \overline{Y}_{Fk})$ คือ ความแตกต่างของผลการตอบที่ได้จากแบบสอบที่ศึกษาระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่มีความสามารถเท่ากัน $(x = k)$

ถ้า $\overline{Y}_{Rk} - \overline{Y}_{Fk} = 0$ แสดงว่าข้อสอบในแบบสอบที่ศึกษาทำหน้าที่ไม่แตกต่างกัน

โดยมีสมมติฐานว่า $H_0 : \beta u = 0$

$H_1 : \beta u > 0$

ขั้นตอนในการวิเคราะห์มีดังนี้

1. ประมาณค่า βu ซึ่งเป็นดัชนีที่บ่งชี้การทำหน้าที่ต่างกันของข้อสอบ จากสูตร

$$\beta u = \Sigma P_k (\overline{Y}_{Rk} - \overline{Y}_{Fk})$$

โดย P_k คือสัดส่วนของผู้สอบกลุ่มเปรียบเทียบที่ตอบสอบที่มีความตรงได้ถูกต้อง k ข้อ

2. ทดสอบนัยสำคัญทางสถิติ

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

โดยที่ $\sigma(\beta u) = \left(\Sigma P_k^{\,2} \left[1/J_{Rk}\, \sigma^2(Y/k,R) + 1/J_{Fk}\, \sigma^2(Y/k,F) \right] \right)^{1/2}$

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

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

J_{Rk} , J_{Fk} คือจำนวนผู้สอบของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่มีคะแนนจากแบบสอบที่มีความตรงระดับ k

ตอนที่ 5 เอกสารและงานวิจัยที่เกี่ยวข้อง

5.1 งานวิจัยต่างประเทศ

Swaminathan และ Rogers (1990) ได้เปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันแบบเอกรูปและแบบอเนกรูประหว่างวิธีถดถอยโลจิสติกกับวิธีแมนเทล-แฮนส์เซล โดยจำลองข้อสอบที่ทำหน้าที่ต่างกัน 2 ชุด ชุดหนึ่งเป็นข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูป ซึ่งกำหนดให้ค่าอำนาจจำแนกรายข้อระหว่างกลุ่มผู้สอบเท่ากัน แต่ค่าความยากผันแปรไปตามระดับของข้อสอบที่ต้องการให้เกิดการทำหน้าที่ต่างกัน ส่วนชุดที่สองเป็นการจำลองข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูป ซึ่งกำหนดให้ค่าความยากของข้อสอบระหว่างกลุ่มผู้สอบเท่ากัน แต่ค่าอำนาจจำแนกรายข้อผันแปรไป โดยกลุ่มตัวอย่างที่ใช้มี 2 ขนาด คือ 250 และ 500 คน ความยาวของแบบสอบมี 3 ขนาด คือ 40, 60 และ 80 ข้อ

ผลการศึกษาพบว่าในการตรวจสอบการทำหน้าที่ต่างกันแบบเอกรูปทั้งสองวิธีให้ผลที่ใกล้เคียงกันในทุกความยาวของแบบสอบ โดยวิธี MH ดีกว่าเล็กน้อย คือ มีการตรวจพบถูกต้องร้อยละ 70 ในกลุ่มตัวอย่าง 250 คนและในกลุ่มตัวอย่าง 500 คน มีการตรวจพบถูกต้องร้อยละ 100 ส่วนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปปรากฏว่าวิธีถดถอยโลจิสติกตรวจพบในกลุ่มตัวอย่างน้อยและแบบสอบสั้นได้ถูกต้องร้อยละ 50 ในกลุ่มตัวอย่างใหญ่และแบบสอบยาวตรวจพบได้ถูกต้องร้อยละ 75 สำหรับวิธีแมนเทล - แฮนส์เซลพบว่าไม่สามารถตรวจค้นได้ สำหรับการตรวจสอบการทำหน้าที่ต่างกันแบบเอกรูป พบว่า วิธี MH จะตรวจค้นได้ดีกว่าทั้งในระดับความสามารถของผู้สอบสูงและต่ำ

Rogers และ Swaminathan (1993) ได้เปรียบเทียบวิธีถดถอยโลจิสติกกับวิธีแมนเทล-แฮนส์เซลในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เป็นการศึกษาเกี่ยวกับการกระจายของสถิติทดสอบและประสิทธิภาพของสถิติทดสอบของแต่ละวิธีโดยการใช้ข้อมูลจำลอง ในการศึกษาด้านการกระจายของสถิติทดสอบ มีปัจจัยที่แปรเปลี่ยน 5 ตัว คือ ขนาดกลุ่มตัวอย่าง(250, 500 คน) ความเหมาะสมของข้อมูลกับโมเดล ค่าความยากของข้อสอบ ค่าอำนาจจำแนกของข้อสอบ และความยาวของแบบสอบ(40 ข้อ) สำหรับการศึกษาด้านประสิทธิภาพของสถิติทดสอบของแต่ละวิธีมีปัจจัยที่แปรเปลี่ยน 8 ตัว คือ ขนาดกลุ่มตัวอย่าง (250, 500 คน) ความเหมาะสมของข้อมูลกับโมเดล ความยาวของแบบสอบ (40, 80 ข้อ) การกระจายของข้อสอบ สัดส่วนของข้อ

สอบที่ทำหน้าที่ต่างกัน ค่าความยากของข้อสอบ ค่าอำนาจจำแนกของข้อสอบ และพื้นที่ระหว่างโค้งคุณลักษณะข้อสอบระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ

ผลการศึกษาพบว่าการกระจายของสถิติของทั้งสองวิธีเป็นไปตามที่คาดหมายไว้ทั้งหมด ยกเว้นวิธีถดถอยโลจิสติกเนื่องจากข้อสอบมีค่าความยากและค่าอำนาจจำแนกของข้อสอบสูง ในด้านประสิทธิภาพของสถิติทดสอบของแต่ละวิธีพบว่าในการตรวจสอบการทำหน้าที่ต่างกันแบบเอกรูปทั้งสองวิธีได้ผลเท่าเทียมกัน ส่วนแบบอเนกรูปพบว่าวิธีถดถอยโลจิสติกมีประสิทธิภาพมากกว่า นอกจากนี้ขนาดกลุ่มตัวอย่างยังมีผลกระทบต่ออัตราการตรวจสอบของทั้งสองวิธีโดยแปรผันตรงกัน สำหรับความยาวของแบบสอบและการกระจายของคะแนนไม่มีผลกระทบต่ออัตราการตรวจสอบ

Narayanan และ Swaminathan (1996) เปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูประหว่างวิธีแมนเทล - แฮนส์เซล วิธีถดถอยโลจิสติกและวิธีโครชิบ(CRO-SIB) โดยใช้ข้อมูลจำลอง ความยาวแบบสอบ 40 ข้อและกำหนดเงื่อนไขที่แปรเปลี่ยนได้ คือ กลุ่มตัวอย่าง(กลุ่มอ้างอิง : 500, 1000 คน กลุ่มเปรียบเทียบ : 200, 500 คน) การกระจายความสามารถของทั้งสองกลุ่มตัวอย่าง(เท่ากัน : ไม่เท่ากัน) สัดส่วนของจำนวนข้อสอบที่ DIFในแบบสอบ(0%, 10%, 20%) พื้นที่ความแตกต่างระหว่างโค้งคุณลักษณะข้อสอบทั้งสองกลุ่มหรือขนาดของข้อสอบที่ DIF (0.4, 0.6, 0.8, 1.0) ค่าความยากและค่าอำนาจจำแนกของข้อสอบ สำหรับค่าการเดาคงที่ เท่ากับ 0.2

ผลการศึกษา พบว่า วิธีถดถอยโลจิสติกและวิธีโครชิบตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปได้ดีกว่าวิธีแมนเทล-แฮนส์เซลโดยให้ผลการตรวจสอบใกล้เคียงกัน เงื่อนไขที่ส่งผลให้อัตราในการตรวจสอบ DIF ของทั้งสามวิธีเพิ่มขึ้น ได้แก่ การเพิ่มขนาดตัวอย่าง การกระจายความสามารถของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบอย่างเท่ากัน และพื้นที่ความแตกต่างระหว่างโค้งคุณลักษณะข้อสอบสองกลุ่มเพิ่มจาก 0.4 เป็น1.0 วิธีแมนเทล-แฮนส์เซลสามารถตรวจ DIF ได้ดีเมื่อข้อสอบมีค่าความยากสูงและต่ำ โดยที่โค้งคุณลักษณะข้อสอบสองกลุ่มจะตัดกันที่ระดับความสามารถสูงหรือต่ำ ส่วนวิธีถดถอยโลจิสติกและวิธีโครชิบจะตรวจพบข้อสอบที่ DIF ในกรณีที่ข้อสอบมีค่าความยากต่ำ เมื่อพิจาณาอัตราความคลาดเคลื่อนประเภทที่ 1 พบว่าวิธีที่มีอัตราความคลาดเคลื่อนในการตรวจค้นสูงสุด ได้แก่ วิธีโครชิบ รองลงมา คือ วิธีถดถอยโลจิสติก และวิธีที่มีอัตราความคลาดเคลื่อนในการตรวจค้นต่ำสุด คือ วิธีแมนเทล-แฮนส์เซล

สำหรับงานวิจัยที่ศึกษาเกี่ยวกับเกณฑ์การจับคู่เปรียบเทียบในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเพิ่งจะได้รับความสนใจในระยะหลังมานี้ โดยสนใจที่จะหาเกณฑ์การจับคู่ที่เหมาะสมในการตรวจสอบ DIF ที่ผ่านมามีการศึกษาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ใช้ทั้งเกณฑ์ภายนอกและเกณฑ์ภายใน ดังเช่น Baeza (1989) ได้ศึกษาพฤติกรรมการตอบข้อสอบชาวอเมริกันอินเดียนและอเมริกันคอเคเซี่ยน วิธีการที่ใช้ในการตรวจสอบ ได้แก่ วิธี MH และวิธีการจับคู่กลุ่มตัวอย่างของMcNemar เกณฑ์การจับคู่สำหรับวิธี MH คือ คะแนนรวมจากแบบสอบ ส่วนเกณฑ์การจับคู่สำหรับ McNemar คือ ตัวแปรด้านเศรษฐกิจสังคมและระดับคะแนนเฉลี่ยมัธยมปลาย

ผลการวิเคราะห์พบว่า วิธี MH ซึ่งใช้เกณฑ์ภายในตรวจพบข้อสอบที่ลำเอียงเพียงเล็กน้อย ส่วน McNemar ซึ่งใช้เกณฑ์ภายนอกนั้นพบว่า เมื่อใช้เกณฑ์การจับคู่เป็นสถานภาพทางเศรษฐกิจสังคมพบว่า ข้อสอบมีระดับความลำเอียงต่ำกว่าการจับคู่ด้วยระดับคะแนนเฉลี่ย

Clauser และคณะ(1991) ได้ศึกษาเกี่ยวกับความสำคัญของการจับคู่เกณฑ์ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี MH โดยใช้แบบสอบที่มีเนื้อหาในการวัดหลากหลาย ซึ่งใช้เกณฑ์ในการจับคู่ทั้งคะแนนรวม(total test score) และคะแนนแบบสอบย่อย(subtest score) มีสมมติฐานว่า การใช้คะแนนรวมเป็นเกณฑ์ในการจับคู่ในแบบสอบที่มีเนื้อหาหลากหลายจะเพิ่มอัตราการเกิดความคลาดเคลื่อนประเภทที่ 1ให้สูงขึ้น ซึ่งสมมติฐานนี้สอดคล้องกับผลการศึกษาจากข้อมูลจำลอง โดย Ackerman(1992) ที่ชี้ให้เห็นว่า เมื่อแบบสอบที่ใช้ในการตรวจสอบ DIF มีเนื้อหาที่มิติค่อนข้างซับซ้อน การใช้คะแนนรวมเป็นเกณฑ์ในการจับคู่จะส่งผลให้เกิดความคลาดเคลื่อนประเภทที่ 1 สูงขึ้น

Ryan (1991) ศึกษาความคงที่(stability) ของวิธีแมนเทล-แฮนส์เซลโดยการใช้กลุ่มตัวอย่างและเกณฑ์จับคู่เปรียบเทียบที่แตกต่างกัน กลุ่มตัวอย่างจำแนกตามสีผิว ได้แก่ กลุ่มนักเรียนผิวขาวและกลุ่มนักเรียนผิวดำ สุ่มแบ่งกลุ่มออกเป็นกลุ่มละ 4 กลุ่มย่อย แบบสอบที่ใช้เป็นวิชาคณิตศาสตร์ประกอบด้วยเนื้อหาที่หลากหลาย โดยแบบสอบที่ใช้วัดความสามารถเพื่อจับคู่เปรียบเทียบแบ่งเป็น 3 กรณี ได้แก่ 1) ข้อสอบร่วม 40 ข้อ 2) ข้อสอบสุ่มจากเนื้อหาต่าง ๆ ฉบับละ 35 ข้อ จำนวน 4 ฉบับ 3) ข้อสอบจากการรวมข้อสอบร่วมกับข้อสอบที่สุ่มมาเป็นฉบับละ 75 ข้อ จำนวน 4 ฉบับ การศึกษาจะเปลี่ยนไปตามกลุ่มตัวอย่าง ขนาดกลุ่มตัวอย่าง สำหรับเกณฑ์จับคู่เปรียบเทียบใช้ 2 เกณฑ์ คือ 1) คะแนนจากแบบสอบร่วม (40 ข้อ) 2) คะแนนจากแบบสอบกรณีที่ 3 (75 ข้อ)

ผลการศึกษาพบว่า เมื่อนำแต่ละเงื่อนไขมาหาความสัมพันธ์กัน แต่ใช้เกณฑ์จับคู่ต่างกัน ค่าสหสัมพันธ์ที่ได้ไม่ต่างกัน นั่นคือเกณฑ์จับคู่ไม่ส่งผลกระทบต่อการตรวจสอบไม่ว่าใช้ขนาดของกลุ่มตัวอย่างขนาดใหญ่หรือย่อย แสดงว่าดัชนี MH มีความแกร่งและจะมีความคงที่ในการประมาณค่าเมื่อใช้กลุ่มตัวอย่างที่มีขนาดใหญ่

Clauser (1993) ยังได้ศึกษาอิทธิพลของการทำให้เกณฑ์การจับคู่ผู้สอบที่มีความบริสุทธิ์ (purification of the matching criterion)ระหว่างเทคนิค 1 ขั้นตอน (one step procedure) และเทคนิค 2 ขั้นตอน (two step procedure) ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธี MH โดยการใช้สถานการณ์จำลองสร้างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบกลุ่มละ 1,000 คน จำลองสถานการณ์ตามเงื่อนไขต่าง ๆ กัน 24 เงื่อนไข คือ สร้างแบบสอบความยาว 3 ขนาด ได้แก่ 20 ข้อ 40 ข้อและ 80 ข้อ สร้างข้อสอบที่ทำหน้าที่ต่างกันลงในแบบสอบแต่ละขนาดจำนวน 0%, 3%, 8% และ 20% รวมทั้งสร้างเงื่อนไขของระดับความสามารถ 2 ระดับ

ผลการศึกษาปรากฏว่า ผลการตรวจพบข้อสอบที่ทำหน้าที่ต่างกันด้วยเทคนิค 2 ขั้นตอนเท่ากับหรือเหนือกว่าเทคนิค 1 ขั้นตอนในทุกเงื่อนไขของการทดสอบและเทคนิค 2 ขั้นตอนไม่เพิ่มระดับความคลาดเคลื่อนประเภทที่ 1 (type 1 eror rate)

Clauser และคณะ(1994) ได้ศึกษาผลกระทบจากความกว้างของชั้นคะแนน (score group) ที่มีต่ออำนาจการทดสอบและระดับความคลาดเคลื่อนประเภทที่ 1 ของวิธี MH จากการใช้ข้อมูลจำลองซึ่งมีทั้งข้อสอบที่ทำหน้าที่ต่างกัน (DIF) และข้อสอบที่ทำหน้าที่ไม่แตกต่างกัน (no DIF) ผลการทดสอบทางสถิติพบว่า การเพิ่มขนาดของผู้สอบจะช่วยให้เพิ่มอำนาจการทดสอบและไม่เพิ่มระดับความคลาดเคลื่อนประเภทที่1 สำหรับการลดชั้นคะแนนในการจับคู่เกณฑ์ของคะแนน (matching score)นั้น ถึงแม้ว่าจะเพิ่มอำนาจการทดสอบแต่จะทำให้ระดับความคลาดเคลื่อนประเภทที่ 1 เพิ่มขึ้นด้วย Clauser และคณะ จึงได้เสนอแนะว่าในการใช้วิธี MH วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบนักวิจัยควรใช้จำนวนชั้นคะแนนที่เป็นไปได้มากที่สุดจากคะแนนรวมของผู้สอบ(total score)เป็นเกณฑ์ในการจับคู่ โดยเฉพาะอย่างยิ่งเมื่อการกระจายของคะแนนของความสามารถของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบแตกต่างกัน

Mezor และ Clauser (1995) ได้เปรียบเทียบวิธีถดถอยโลจิสติกกับวิธีแมนเทล-แฮนส์เซลในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อใช้เกณฑ์ภายนอกหรือความสามารถหลากหลาย(multiple ability)เข้ามาร่วมเป็นเกณฑ์เปรียบเทียบ โดยศึกษาจากข้อมูลจริงซึ่งเป็นผลการตอบข้อสอบวัดผลสัมฤทธิ์ทางการเรียนวิชาประวัติศาสตร์และวิชาเคมีของนักเรียนระดับมัธยม

ศึกษา ความยาวแบบสอบ 75 ข้อ จำแนกกลุ่มผู้สอบตามเพศและความสามารถทางภาษา สำหรับเกณฑ์ที่ใช้จับคู่ ได้แก่ ใช้คะแนนรวมของแบบสอบ ใช้คะแนนรวมของแบบสอบร่วมกับเกณฑ์ภายนอก นั่นคือ ตัวแปรความถนัดทางภาษาและความถนัดทางคณิตศาสตร์

ผลการศึกษาพบว่า วิธีถดถอยโลจิสติกตรวจพบข้อสอบทำหน้าที่ต่างกันมากกว่าวิธีแมนเทล-แฮนส์เซลในทุกลักษณะ ส่วนใหญ่เมื่อใช้เกณฑ์ภายนอกทั้งสองตัวแปรเข้ามาเป็นเกณฑ์จับคู่ร่วมกับคะแนนรวมจากแบบสอบจะทำให้พบข้อสอบที่ถูกระบุว่าทำหน้าที่ต่างกันน้อยลง โดยพบว่าเมื่อใช้คะแนนรวมเป็นเกณฑ์ร่วมกับตัวแปรความถนัดทางภาษาจะตรวจพบน้อยกว่าเมื่อใช้คะแนนรวมเป็นเกณฑ์ร่วมกับตัวแปรความถนัดทางคณิตศาสตร์

Clauser และคณะ(1996) ได้ศึกษาการจับคู่เกณฑ์ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบชนิดพหุมิติ โดยเปรียบเทียบผลการตรวจสอบโดยใช้เกณฑ์การจับคู่ 3 เกณฑ์ ได้แก่ คะแนนรวม (total score) คะแนนแบบสอบย่อย (subtest score) และคะแนนหลายแบบสอบย่อย (multiple subtest scores) วิธีที่ใช้ในการตรวจสอบคือ วิธีแมนเทล-แฮนส์เซล (MH) และวิธีถดถอยโลจิสติก (Logistic Regression) ทั้งในข้อมูลจริงและข้อมูลจำลอง พบว่าในแบบสอบที่มีมิติซับซ้อน การใช้คะแนนรวมเป็นเกณฑ์ในการจับคู่นั้นไม่เหมาะสม การใช้คะแนนหลายแบบย่อย (multiple subtest scores) เป็นเกณฑ์ในการจับคู่จะมีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ดีกว่า ทั้งนี้จำนวนข้อสอบที่ถูกตรวจพบว่าทำหน้าที่ต่างกันมีจำนวนน้อยที่สุดเมื่อเปรียบเทียบกับการใช้เกณฑ์การจับคู่อื่น

5.2 งานวิจัยในประเทศ

กาญจนา วัชนสุนทร (2537) ได้พัฒนาเกณฑ์ตัดสินข้อสอบลำเอียงทางเพศด้วยข้อมูลเชิงประจักษ์ โดยใช้ดัชนี 4 ตัว คือ พื้นที่ระหว่างโค้งการตอบข้อสอบชนิดคิดเครื่องหมาย (SA) และไม่คิดเครื่องหมาย(UA)จากทฤษฎีการตอบข้อสอบโมเดล 2 พารามิเตอร์ ดัชนีแอลฟา (α_{MH}) จากวิธีแมนเทล-แฮนส์เซล และดัชนีเบต้า (β_{SIB}) จากวิธี SIBTEST โดยใช้ข้อมูลการตอบข้อสอบคัดเลือกเข้าศึกษาในสถาบันอุดมศึกษาของทบวงมหาวิทยาลัย ปีการศึกษา 2535 ในวิชาคณิตศาสตร์ มีความยาวแบบสอบ 20, 30 และ 40 ข้อ และวิชาภาษาอังกฤษ มีความยาวข้อสอบ 50, 60 และ 80 ข้อ ใช้กลุ่มผู้สอบ 6 ขนาด ได้แก่ 100, 200, 400, 600, 800 และ 1,000 คน

เกณฑ์ที่พัฒนาขึ้นเพื่อใช้ในการตัดสินความลำเอียงของข้อสอบระหว่างผู้สอบหญิงและชาย ได้แก่

1. $|SA| > .80$ และ $UA > .50$ กรณีความยาวแบบสอบต่ำกว่า 50 ข้อ

2. $|SA| > .80$ และ $UA > 1.20$ กรณีความยาวแบบสอบ 50 ข้อขึ้นไป

3. $.60 < \alpha_{MH} > 1.40$ และ $\beta_{SIB} > .06$ ทุกขนาดผู้สอบและความยาวแบบสอบ

ผลการศึกษาพบว่าค่าเฉลี่ยที่ได้จากดัชนีทั้ง 4 ค่าที่ได้จากการเปรียบเทียบระหว่างผู้สอบเพศเดียวกัน ในความยาวแบบสอบและขนาดผู้สอบต่างกันมีค่าใกล้เคียงกันในแต่ละวิชา ผลการตรวจค้นข้อสอบลำเอียงทางเพศ เมื่อใช้ดัชนีตามที่กำหนด พบว่า มีความไม่คงที่ข้ามขนาดผู้สอบและความยาวแบบสอบ ความสอดคล้องในการตรวจค้นข้อสอบลำเอียงภายในวิธีเดียวกันข้ามขนาดผู้สอบต่ำ แต่จะสูงขึ้นที่ขนาดผู้สอบ 600 คน สำหรับการวิเคราะห์ความลำเอียงของข้อสอบที่มีต่อเพศผู้สอบพบว่า ในวิชาภาษาอังกฤษข้อสอบส่วนใหญ่เข้าข้างเพศหญิงมากกว่าเพศชายในกรณีที่ใช้ดัชนี SA และ α_{MH} โดยที่ผลการใช้ดัชนี β_{SIB} จะให้ผลตรงข้าม ส่วนวิชาภาษาอังกฤษข้อสอบที่ลำเอียงจะเข้าข้างเพศชายมากกว่าเพศหญิง

เกษร หว่างจิตร (2539) ศึกษาการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมนเทล-แฮนส์เซล โดยใช้ข้อมูลจากผลการตอบข้อสอบวิชาภาษาไทยของผู้สอบ จำนวน 506 คน และผลการตอบข้อสอบวิชาภาษาอังกฤษของผู้สอบ จำนวน 501 คน ในส่วนที่เป็นข้อสอบแบบเลือกตอบของศูนย์การทดสอบทางการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ซึ่งกลุ่มอ้างอิงและกลุ่มเปรียบเทียบจำแนกตามเพศ ภูมิลำเนา ประสบการณ์ในการสอบและสังกัดสถานศึกษา

ผลการศึกษาพบว่าข้อสอบที่ทำหน้าที่ต่างกันส่วนใหญ่เป็นข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูป มีค่าอำนาจจำแนกค่อนข้างต่ำทั้งสองวิชา โดยจะมีค่าต่ำมากในวิชาภาษาไทย และข้อสอบที่ DIF ในวิชาภาษาไทยเป็นข้อสอบง่าย ส่วนในวิชาภาษาอังกฤษเป็นข้อสอบที่ยาก เมื่อจำแนกกลุ่มอ้างอิงและกลุ่มเปรียบเทียบตามเพศ จะพบข้อสอบที่ DIF มีจำนวนมากที่สุด รองลงมา คือการจำแนกตามภูมิลำเนา สังกัดของสถานศึกษา และประสบการณ์ในการสอบ ตามลำดับ

จิติมา วรรณศรี (2540) ได้เปรียบเทียบประสิทธิภาพระหว่างวิธีแมนเทล - แฮนส์เซลกับวิธีชิบเทสท์ โดยใช้ข้อมูลจำลองจากโปรแกรม IRTDATA มีเงื่อนไขที่ศึกษา ได้แก่ ความยาวแบบสอบ 3 ขนาด (30, 60, 90 ข้อ) ขนาดกลุ่มตัวอย่าง 3 ขนาด (200, 600, 1,000 คน) โดยแต่ละขนาดมีอัตราส่วนระหว่างผู้สอบกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ 4 อัตราส่วน ได้แก่ 1:1, 1:0.9, 1:0.75 และ 1:0.50 รวมเงื่อนไขทั้งสิ้น 36 เงื่อนไข

ผลการศึกษา พบว่า วิธีแมนเทล-แฮนส์เซลกับวิธีชิบเทสท์มีประสิทธิภาพในการตรวจสอบเท่าเทียมกันทุกขนาดกลุ่มตัวอย่าง และทุกอัตราส่วนเมื่อใช้ความยาวแบบสอบเดียวกัน โดยมี

ประสิทธิภาพในการตรวจสอบสูงสุด ในกรณีที่ใช้แบบสอบที่มีความยาวปานกลาง(60ข้อ) เมื่อขนาดกลุ่มตัวอย่าง 200 และ 600 คน สามารถตรวจพบข้อสอบที่ DIF ได้ถูกต้องร้อยละ 50 และเมื่อเพิ่มขนาดเป็น 1,000 คน สามารถตรวจพบข้อสอบที่ DIF ได้ถูกต้องร้อยละ 100 โดยมากวิธีซิบเทสท์มีอัตราความคลาดเคลื่อนประเภทที่ 1 มากกว่าวิธีแมนเทล-แฮนส์เซลเล็กน้อย

สำหรับการวิจัยในครั้งนี้ได้เลือกวิธีที่กำลังได้รับความสนใจ คือ วิธีแมนเทล-เฮนส์เซล(MH) และวิธีถดถอยโลจิสติก (LR) เนื่องจากการศึกษางานวิจัยที่ผ่านมา พบว่า ทั้งสองวิธีนี้เป็นวิธีที่มีประสิทธิภาพในการตรวจสอบ ประหยัดและสะดวกเพราะมีโปรแกรมคอมพิวเตอร์สำเร็จรูปซึ่งง่ายในการคำนวณ (Logers and Swaminathan, 1993 ; Mazor, 1994) และจากการศึกษางานวิจัยที่ผ่านมาพบว่า ในการตรวจสอบ DIF ส่วนใหญ่จะศึกษาผลการตอบข้อสอบจากแบบสอบเอกมิติ (unidimensional test) ซึ่งเป็นแบบสอบที่มีเป้าหมายในการวัดเพียงมิติเดียวหรือมีลักษณะที่แฝงเด่นอยู่ลักษณะเดียว ส่วนในแบบสอบชนิดพหุมิติ (multidimensional test) หรือแบบสอบที่วัดลักษณะแฝงเด่นมากกว่า 1 ลักษณะ พบว่า วิธีการตรวจสอบ DIF ยังมีไม่มากนัก ซึ่งโดยทั่วไปแล้วการสร้างข้อสอบวิชาหนึ่งๆ นั้นมีเนื้อหาในการวัดหลายองค์ประกอบแต่ก็พบว่าการตรวจสอบ DIF ในแบบสอบชนิดนี้มีไม่มากนัก ที่เป็นเช่นนี้ก็เนื่องจากวิธีการตรวจสอบ DIF ส่วนใหญ่ตั้งอยู่บนพื้นฐานของการวัดเพียงมิติเดียวซึ่งเป็นข้อตกลงพื้นฐานในทฤษฎีการตอบสนองข้อสอบ (IRT) ดังนั้นผู้วิจัยจึงสนใจที่จะนำวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ได้รับการยอมรับว่ามีประสิทธิภาพมาใช้ตรวจสอบ DIF ในแบบสอบพหุมิติ เมื่อใช้เกณฑ์การจับคู่เปรียบเทียบอื่นนอกจากคะแนนรวม (total test score) เนื่องจากงานมีวิจัยพบว่าการตรวจสอบ DIF ในแบบสอบที่มีมิติซับซ้อนการนำคะแนนรวมมาใช้เป็นเกณฑ์ในการจับคู่เปรียบเทียบนั้นไม่เหมาะสมเพราะจะทำให้อัตราความคลาดเคลื่อนประเภทที่ 1 สูงขึ้น (Clauser,1993) ดังนั้นในการวิจัยนี้ผู้วิจัยจึงได้เลือกเกณฑ์การจับคู่เปรียบเทียบมาใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบสอบชนิดพหุมิติ ได้แก่ คะแนนรวม (total test score) คะแนนแบบสอบย่อย (subtest score) และคะแนนหลายแบบสอบย่อย (multiple subtest scores) เพื่อให้ได้แนวทางในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบสอบชนิดพหุมิติที่มีประสิทธิภาพที่สุด