

บทที่ 1

บทนำ



ความเป็นมาและความสำคัญของปัญหา

การศึกษาความลำเอียงหรือความไม่ยุติธรรมของแบบสอบเป็นเรื่องที่ได้รับการวิพากษ์วิจารณ์และให้ความสนใจเป็นอย่างมาก เนื่องจากแบบสอบเป็นเครื่องมือสำคัญในการวัดผลการเรียนรู้ เครื่องมือวัดใด ๆ ก็ตามจะต้องมีความตรง ซึ่งเป็นคุณสมบัติที่สำคัญอย่างยิ่งในการสร้างและตรวจสอบคุณภาพของเครื่องมือ ดังนั้นการนำเครื่องมือที่ขาดความตรงหรือมีความลำเอียงในการวัดมาใช้จึงเป็นการไม่ยุติธรรมแก่ผู้สอบ อีกทั้งผลที่ได้จากการวัดไม่สามารถสะท้อนถึงความสามารถที่แท้จริงของผู้สอบทั้งนี้เนื่องจากผู้สอบแต่ละกลุ่มมีการได้เปรียบเสียเปรียบกันในการตอบข้อสอบข้อเดียวกัน ดังนั้นผู้สร้างแบบสอบจึงจำเป็นต้องคำนึงถึงคุณภาพของแบบสอบ โดยต้องสร้างแบบสอบที่เชื่อถือได้ วัดได้ตามคุณลักษณะที่ต้องการวัด เกิดความคลาดเคลื่อนในการวัดน้อยที่สุด และจะต้องระมัดระวังไม่ให้เกิดการได้ประโยชน์หรือเสียประโยชน์ของผู้สอบในแต่ละกลุ่มที่มีความสามารถเท่ากัน (Allen and Yen, 1979; Popham, 1981)

โดยทั่วไปเครื่องมือที่ใช้ในการวัดความสามารถ ส่วนใหญ่นิยมใช้แบบสอบซึ่งถูกสร้างขึ้นเพื่อวัดคุณลักษณะแฝง (Latent trait) ที่ต้องการวัด แต่เมื่อนำไปใช้พบผลการวิเคราะห์พบว่าผู้สอบนอกจากต้องใช้ความรู้ในลักษณะแฝงเป้าหมายของการวัดแล้วยังต้องใช้ความรู้ในลักษณะแฝงอื่น (ลักษณะแฝงแทรกซ้อน) เข้ามาช่วยในการทำข้อสอบหรือแบบสอบนั้น ๆ ด้วย เรียกว่าข้อสอบนั้นหรือแบบสอบนั้นทำหน้าที่ต่างกัน (Differential Item Functioning : DIF) และเมื่อใดที่นำข้อสอบหรือแบบสอบที่ทำหน้าที่ต่างกันนี้ไปวัดประชากรที่ประกอบด้วยประชากรกลุ่มย่อยตั้งแต่ 2 กลุ่มขึ้นไป โดยที่ประชากรย่อยกลุ่มหนึ่งมีความรู้ในลักษณะแฝงแทรกซ้อนสูงกว่าประชากรกลุ่มย่อยกลุ่มอื่น ๆ ทำให้สามารถตอบข้อสอบหรือแบบสอบได้ถูกต้องมากกว่าประชากรย่อยอื่น ๆ ทั้งที่มีระดับความรู้ในลักษณะแฝงเป้าหมายเท่ากัน ลักษณะการเช่นนี้จะทำให้เกิดความลำเอียงของข้อสอบหรือแบบสอบได้หากแปลความหมายของผลการวัดโดยอ้างอิงที่ลักษณะแฝงเป้าหมายเพียงลักษณะเดียว ซึ่งความจริงแล้วผลการตอบข้อสอบที่แตกต่างกันระหว่างกลุ่มประชากรย่อยไม่ได้

เกิดจากความแตกต่างในลักษณะแง่เป้าหมายเพียงอย่างเดียว แต่มีความแตกต่างในลักษณะแง่แทรกซ้อนที่มีได้ตั้งใจวัดเข้ามาด้วย (กาญจนา วัฒนสุนทร, 2538)

เพื่อแก้ปัญหาการได้เปรียบเสียเปรียบกันระหว่างผู้สอบ ได้มีผู้เสนอวิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบไว้หลายวิธี โดยแบ่งเป็น 2 แนวทางใหญ่ ๆ (Ruder Getson and Knight, 1980) คือ แนวทางแรก เป็นการตรวจสอบโดยใช้เกณฑ์ภายนอก (External Criterion) โดยการนำคะแนนจากการตอบข้อสอบไปสัมพันธ์กับเกณฑ์ที่ได้จากภายนอก เช่น เกณฑ์เฉลี่ยหรือผลสัมฤทธิ์ของงานที่ทำ เป็นต้น แต่จะมีจุดอ่อนเพราะเป็นการยากที่จะหาเกณฑ์ภายนอกที่มีความตรงและยุติธรรม ซึ่งถ้าหากขาดคุณสมบัติข้อนี้จะทำให้ผลการตรวจสอบเกิดความคลาดเคลื่อนได้ ส่วนแนวทางที่สอง เป็นแนวทางที่มีผู้นิยมทำการศึกษามากในปัจจุบัน คือ การตรวจสอบเมื่อไม่มีเกณฑ์ภายนอก (Absence of Criterion) เป็นการนำวิธีการทางสถิติมาตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยพิจารณาโครงสร้างภายในของแบบสอบ(ทัศนีย์ พิรมนตรี, 2530) ซึ่งในปัจจุบันมีหลายวิธี โดยจิตติมา วรณศรี (2540) ได้สรุปวิธีที่นิยมใช้ได้เป็นตารางเพื่อเปรียบเทียบข้อดีข้อเสียของวิธีต่าง ๆ ได้แก่ วิธีวิเคราะห์ความแปรปรวน (ANOVA) วิธีไคสแควร์ (Chi square) วิธีทฤษฎีการตอบสนองข้อสอบ(IRT) วิธีแมนเทิล-แฮนส์เซล(MH) และวิธีซิบเทสต์(SIBTEST) ทั้งนี้ได้เพิ่มวิธีถดถอยโลจิสติก(LR) เพื่อเปรียบเทียบอีกหนึ่งวิธี ซึ่งมีรายละเอียดต่าง ๆ ดังตารางที่ 1

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

ตารางที่ 1 การเปรียบเทียบวิธีต่าง ๆ ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

| ประเด็น | ANOVA | χ^2 | IRT | MH | SUBTEST | LR |
|--------------------------------|---------------------------------------------------|------------------------------------------------|-------------------------------------------------------------------------------------------------|-------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------|
| 1. ข้อตกลงเบื้องต้น | ความแปรปรวนและความแปรปรวนร่วมของข้อสอบต้องเท่ากัน | คะแนนรวมจากแบบสอบเป็นตัวแทนความสามารถของผู้สอบ | แบบสอบเป็นเอกมิติและโด่งลักษณะข้อสอบสามารถแสดงฟังก์ชันของค่าความสามารถและโอกาสในการตอบข้อสอบถูก | คะแนนรวมจากแบบสอบเป็นตัวแทนความสามารถของผู้สอบ | คะแนนรวมจากแบบสอบเป็นตัวแทนความสามารถของผู้สอบ และมีมิติการวัด 2 มิติ คือคุณลักษณะแฝงเป้าหมายและคุณลักษณะแฝงแทรกซ้อน | คะแนนจากแบบสอบเป็นตัวแทนความสามารถที่ต่อเนื่องของผู้สอบ |
| 2. สิ่งที่ทำ การวิเคราะห์ | ผลรวมระหว่างการเป็นสมาชิกในกลุ่มกับการตอบถูก | ความแตกต่างของสัดส่วนการตอบถูกตามระดับคะแนนรวม | ความแตกต่างของฟังก์ชันการตอบข้อสอบ | ความแตกต่างของสัดส่วนการตอบระหว่างผู้ที่มีความสามารถระดับเดียวกัน | ความแตกต่างระหว่างคะแนนเฉลี่ยและสัดส่วนการตอบข้อสอบระหว่างผู้ที่มีความสามารถระดับเดียวกัน | ความแตกต่างระหว่างจุดตัดและค่าพารามิเตอร์ในแต่ละกลุ่ม |
| 3. สิ่งที่ใช้ในการตัดสินใจ DIF | ความมีนัยสำคัญทางสถิติของ F-test | ความมีนัยสำคัญทางสถิติของ χ^2 | พื้นที่ระหว่างโด่งลักษณะข้อสอบ | ค่าของดัชนี α_{MH} และความมีนัยสำคัญทางสถิติ | ค่าของดัชนี β_{SB} และความมีนัยสำคัญทางสถิติ | จุดตัด (β_0) และค่าความชัน (β_1) ของพารามิเตอร์และความมีนัยสำคัญทางสถิติ |

ตารางที่ 1 (ต่อ)

| ประเด็น | ANOVA | χ^2 | IRT | MH | SUBTEST | LR |
|-----------------|---------------------------------------------------------|-------------------------------------------------------------------|---------------------------------------------------------------------|---------------------------------------------------------|---------------------------------------------------------------------|-------------------------------------------------------------------------|
| 4. ทฤษฎีพื้นฐาน | - | - | IRT | - | IRT ชนิดพหุมิติ | สถิติ Logistic Regression |
| 5. ข้อดี | ใช้กลุ่มตัวอย่างน้อย | คำนวณง่าย มีเกณฑ์ตายตัวในการแปลผล | ให้รายละเอียดและการไม่แปรเปลี่ยนของพารามิเตอร์ | คำนวณง่าย ใช้กลุ่มตัวอย่างประหยัด | คำนวณง่าย ใช้กลุ่มตัวอย่างน้อย ตรวจสอบ DIF ได้หลายข้อในคราวเดียวกัน | คำนวณง่าย ใช้กลุ่มตัวอย่างน้อย ตรวจสอบ DIF ได้ดีทั้งแบบเอกรูปและอเนกรูป |
| 6. ข้อเสีย | การคำนวณค่อนข้างยุ่งยากและไม่มีดัชนีบอกระดับความลำเอียง | ไม่มีเกณฑ์ตายตัวในการกำหนดช่วงคะแนนและค่า b เปลี่ยนตามกลุ่มผู้สอบ | มีการคำนวณซับซ้อนหลายรอบ แปลผลยาก ใช้กลุ่มตัวอย่างมาก ค่าใช้จ่ายสูง | ไม่มีความไวในการตรวจสอบ DIF แบบอเนกรูป (nonuniform DIF) | อัตราความคลาดเคลื่อนชนิดที่ 1 เพิ่มสูงเมื่อคะแนนเฉลี่ยแตกต่างกันมาก | ประสิทธิภาพในการตรวจสอบต่ำเมื่อข้อสอบมีค่าความยากและค่าอำนาจจำแนกสูง |

จากตารางวิธีวิเคราะห์ความแปรปรวน (ANOVA) และวิธีไคสแควร์ (χ^2) มีจุดอ่อนเนื่องจากทั้งสองวิธีนี้เป็นวิธีวิเคราะห์ข้อสอบโดยอาศัยทฤษฎีการวัดแบบดั้งเดิมจึงทำให้ค่าพารามิเตอร์ของข้อสอบจะแปรเปลี่ยนไปตามกลุ่มผู้สอบ ทั้งยังไม่มีดัชนีบอกระดับการทำหน้าที่ต่างกันของข้อสอบ แต่ข้อดี คือเป็นวิธีที่ประหยัดและใช้กลุ่มตัวอย่างน้อย ส่วนวิธีทฤษฎีการตอบสนองของข้อสอบ (IRT) เป็นวิธีที่พิจารณาฟังก์ชันการตอบข้อสอบระหว่างกลุ่มผู้สอบ ซึ่งมีข้อดีคือ มีดัชนีในการตรวจสอบและทดสอบนัยสำคัญทางสถิติ ค่าพารามิเตอร์จะไม่แปรเปลี่ยนไปตามกลุ่มผู้สอบ แต่ข้อเสียคือ ค่าใช้จ่ายสูงเนื่องจากการคำนวณซับซ้อนและต้องใช้กลุ่มตัวอย่างจำนวนมาก วิธีแมนเทิล-แฮนสเฟล (MH) เป็นวิธีที่วิเคราะห์ที่ระดับความสามารถ มีดัชนีบอกระดับการทำหน้าที่ต่างกันของข้อสอบและการทดสอบนัยสำคัญทางสถิติ ซึ่งมีข้อดีคือ สะดวก ประหยัดและคำนวณง่าย แต่มีข้อเสียคือ ประสิทธิภาพในการตรวจสอบ DIF แบบอเนกรูป (nonuniform DIF) ด้วยวิธีนี้ยังไม่ดีนัก ส่วนวิธีซิบเทสต์ (SIBTEST) เป็นวิธีที่มีข้อตกลงว่ามีมิติการวัด 2 มิติ โดย

คะแนนจากแบบสอบจะแบ่งเป็น 2 ส่วน ในส่วนแรกเป็นคะแนนจากแบบสอบที่มีความตรง (valid subtest) ซึ่งมุ่งวัดคุณลักษณะแฝงเป้าหมายและอีกส่วนหนึ่งเป็นคะแนนจากแบบสอบที่ศึกษา (studied subtest) โดยมุ่งวัดคุณลักษณะแฝงแทรกซ้อน วิธีในการตรวจสอบวิธีนี้มีข้อดีเช่นเดียวกับวิธี MH ทั้งยังสามารถตรวจสอบ nonuniform DIF ได้ดีกว่าด้วย สำหรับวิธีถดถอยโลจิสติก (LR) เป็นวิธีการทดสอบสถิติแบบพารามิเตอร์มีการทดสอบโมเดลทางสถิติ Logistic Regression วิธีนี้ใช้วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบได้ดีทั้งแบบเอกกรุปและอนเนกรุป แต่จะเสียค่าใช้จ่ายสูงกว่าวิธี MH

ที่ผ่านมาได้มีการศึกษาเปรียบเทียบวิธีการในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการต่าง ๆ เช่น Rogers และ Swaminathan (1993) พบว่า วิธีแมนเทิล-แฮนส์เซล (MH) มีประสิทธิภาพเท่าเทียมกับวิธีถดถอยโลจิสติกในการตรวจสอบการทำหน้าที่ต่างกันแบบเอกกรุป (Uniform DIF) และวิธี MH มีประสิทธิภาพดีกว่าในการตรวจสอบการทำหน้าที่ต่างกันแบบอนเนกรุป (nonuniform DIF) ซึ่งสอดคล้องกับผลจากการศึกษาของ Mazor และคณะ(1994) ที่พบว่าวิธี MH สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกกรุปได้ดีแต่ไม่มีความไว (sensitive) ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอนเนกรุป เกษร ห่วงจิตร(2539) ใช้วิธี MH ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ พบว่าข้อสอบที่ทำหน้าที่ต่างกันส่วนใหญ่เป็นข้อสอบที่ทำหน้าที่ต่างกันแบบอนเนกรุป Shealy และ Stout (1993) ทำการศึกษาพบว่าในกรณีข้อสอบลำเอียงข้อเดียว (single biased item) ทั้งวิธี MH และวิธี SIBTEST มีประสิทธิภาพดีสำหรับใช้ในการตรวจสอบและในกรณีข้อสอบลำเอียงหลายข้อ (several biased item) วิธี SIBTEST ก็สามารถใช้ตรวจสอบได้อย่างมีประสิทธิภาพ

สำหรับการวิจัยในครั้งนี้ได้เลือกวิธีที่กำลังได้รับความสนใจ คือ วิธีแมนเทิล-แฮนส์เซล(MH) และวิธีถดถอยโลจิสติก (LR) เนื่องจากการศึกษางานวิจัยที่ผ่านมาพบว่า ทั้งสองวิธีนี้เป็นวิธีที่มีประสิทธิภาพในการตรวจสอบ ประหยัดและสะดวกเพราะมีโปรแกรมคอมพิวเตอร์สำเร็จรูปซึ่งง่ายในการคำนวณ (Logers and Swaminathan,1993 ; Mazor,1994)

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เป็นการเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มผู้สอบ 2 กลุ่มหลังจากการจับคู่ (matching) ผู้สอบตามความรู้หรือความสามารถของผู้สอบ การจับคู่ของผู้สอบสองกลุ่มเป็นเงื่อนไขที่สำคัญเนื่องจากเกณฑ์การจับคู่(matching criteria) เป็นเกณฑ์ที่ใช้แทนความสามารถที่แท้จริงของผู้สอบสองกลุ่ม ในทางปฏิบัติมักใช้คะแนนรวมของ

แบบสอบ (total test score) เป็นเกณฑ์ในการจับคู่เปรียบเทียบ เนื่องจากเกี่ยวข้องกับ ความรู้หรือความสามารถที่วัดได้ โดยแบบสอบนั้นสามารถตรวจสอบความตรงและความเที่ยงได้ อีกทั้งผู้สอบทุกคนสอบภายใต้สถานการณ์เดียวกัน แต่จุดอ่อนของการใช้คะแนนรวมของแบบ สอบเป็นเกณฑ์จับคู่เปรียบเทียบผู้สอบ คือ มีการรวมเอาคะแนนจากข้อสอบที่ทำหน้าที่ต่างกันมา เป็นเกณฑ์ในการจับคู่ผู้สอบด้วย (Holland and Thayer, 1988)

งานวิจัยที่ศึกษาเกี่ยวกับเกณฑ์จับคู่เปรียบเทียบในการตรวจสอบการทำหน้าที่ต่างกัน ของข้อสอบเหิงจะได้รับความสนใจในระยะหลังมานี้ โดยสนใจที่จะหาเกณฑ์การจับคู่ที่เหมาะสม ในการตรวจสอบ DIF ที่ผ่านมามีการศึกษาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ใช้ ทั้งเกณฑ์ภายนอกและเกณฑ์ภายใน ดังเช่น Baeza (1989) ได้ศึกษาพฤติกรรมการตอบข้อสอบ ชาวอเมริกันอินเดียและอเมริกันคอเคเซียน วิธีการที่ใช้ในการตรวจสอบ ได้แก่ วิธี MH และวิธี การจับคู่กลุ่มตัวอย่างของ McNemar เกณฑ์การจับคู่สำหรับวิธี MH คือ คะแนนรวมจากแบบสอบ ส่วนเกณฑ์การจับคู่สำหรับ McNemar คือ ตัวแปรด้านเศรษฐกิจสังคมและระดับคะแนนเฉลี่ย มัธยมปลาย ซึ่งผลการวิเคราะห์พบว่า วิธี MH ซึ่งใช้เกณฑ์ภายในตรวจพบข้อสอบที่ลำเอียงเพียง เล็กน้อย ส่วน McNemar ซึ่งใช้เกณฑ์ภายนอกนั้นพบว่า เมื่อใช้เกณฑ์การจับคู่เป็นสถานภาพ ทางเศรษฐกิจสังคมพบว่า ข้อสอบมีระดับความลำเอียงต่ำกว่าการจับคู่ด้วยระดับคะแนนเฉลี่ย

ส่วน Clauser และคณะ (1991) ได้ศึกษาเกี่ยวกับความสำคัญของการจับคู่เกณฑ์ในการ ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี MH โดยใช้แบบสอบที่มีเนื้อหาในการวัดหลากหลายซึ่งใช้เกณฑ์ในการจับคู่ทั้งคะแนนรวม(total test score) และคะแนนแบบสอบย่อย(subtest score) มีสมมติฐานว่า การใช้คะแนนรวมเป็นเกณฑ์ในการจับคู่ในแบบสอบที่มีเนื้อหาหลากหลาย จะเพิ่มอัตราการเกิดความคลาดเคลื่อนประเภทที่ 1 ให้สูงขึ้น ซึ่งสมมติฐานนี้สอดคล้องกับผลการ ศึกษาจากข้อมูลจำลองโดย Ackerman(1992) ที่ชี้ให้เห็นว่า เมื่อแบบสอบที่ใช้ในการตรวจสอบ DIF มีเนื้อหาที่มีติดกันข้างซับซ้อน การใช้คะแนนรวมเป็นเกณฑ์ในการจับคู่จะส่งผลให้เกิดความคลาด เคลื่อนประเภทที่ 1 สูงขึ้น และClauser (1995) ยังได้ศึกษาเพิ่มเติมโดยการนำเกณฑ์ภายนอกซึ่งเป็น ความสามารถหลากหลาย (multiple ability) เข้ามาเป็นเกณฑ์จับคู่ร่วมกับคะแนนรวมจากแบบ สอบในการตรวจสอบ DIF ผลปรากฏว่า ข้อสอบที่ถูกตรวจพบว่าทำหน้าที่ต่างกันมีจำนวนลดลง และการนำเกณฑ์ความสามารถมาวิเคราะห์ด้วยจะทำให้ผลการตรวจสอบมีความถูกต้องมากขึ้น ซึ่งวิธีที่เหมาะสมกับการตรวจสอบ DIF ในลักษณะนี้ คือ วิธีถดถอยโลจิสติก

นอกจากนี้ Clauser และคณะ(1996) ได้ศึกษาการจับคู่เกณฑ์ในการตรวจสอบการทำหน้าที่

ที่ต่างกันของข้อสอบชนิดพหุมิติ โดยเปรียบเทียบผลการตรวจสอบโดยใช้เกณฑ์การจับคู่ 3 เกณฑ์ ได้แก่ คะแนนรวม (total score) คะแนนแบบสอบย่อย (subtest score) และคะแนนหลายแบบสอบย่อย (multiple subtest scores) โดยวิธีที่ใช้ในการตรวจสอบคือ วิธีแมนเทิล-แฮนส์เซล (MH) และวิธีถดถอยโลจิสติก (Logistic Regression) ทั้งในข้อมูลจริงและข้อมูลจำลองพบว่า ในแบบสอบที่มีมิติซับซ้อน การใช้คะแนนรวมเป็นเกณฑ์ในการจับคู่ นั้นไม่เหมาะสม แต่เมื่อใช้คะแนนหลายแบบสอบย่อย (multiple subtest scores) เป็นเกณฑ์ในการจับคู่ จะมีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ดีกว่าซึ่งพิจารณาจากจำนวนข้อสอบที่ถูกตรวจสอบว่าทำหน้าที่ต่างกัันมีจำนวนน้อยที่สุดเมื่อเปรียบเทียบกับ การใช้เกณฑ์การจับคู่อื่น

จากการศึกษางานวิจัยที่ผ่านมาพบว่า ในการตรวจสอบ DIF ส่วนใหญ่จะศึกษาผลการตอบข้อสอบจากแบบสอบเอกมิติ (unidimensional test) ซึ่งเป็นแบบสอบที่มีเป้าหมายในการวัดเพียงมิติเดียวหรือมีลักษณะที่แฝงเด่นอยู่ลักษณะเดียว โดยการตรวจสอบ DIF จะพิจารณาจากค่าสถิติที่ได้จากการตรวจสอบความเป็นพหุมิติของแบบสอบซึ่งแสดงว่าแบบสอบนั้นมีการทำหน้าที่ต่างกันระหว่างกลุ่มผู้สอบ ส่วนในแบบสอบชนิดพหุมิติ (multidimensional test) หรือแบบสอบที่วัดลักษณะแฝงเด่นมากกว่า 1 ลักษณะพบว่า วิธีการตรวจสอบ DIF ยังมีไม่มากนัก ทั้งที่มีงานวิจัยสนับสนุนว่า ในการสร้างข้อสอบนั้นจุดประสงค์ในการวัดของข้อสอบแต่ละข้อสามารถที่จะวัดคุณลักษณะที่ต้องการวัดได้หลาย ๆ คุณลักษณะ (Clauser, 1996) และโดยทั่วไปแล้วการสร้างข้อสอบวิชาหนึ่ง ๆ นั้นมีเนื้อหาในการวัดหลายองค์ประกอบ แต่ก็พบว่าการตรวจสอบ DIF ในแบบสอบชนิดนี้มีไม่มากนัก ที่เป็นเช่นนี้ก็เนื่องมาจากวิธีการตรวจสอบ DIF ส่วนใหญ่ตั้งอยู่บนพื้นฐานของการวัดเพียงมิติเดียวซึ่งเป็นข้อตกลงพื้นฐานในทฤษฎีการตอบสนองข้อสอบ (IRT) ดังนั้นผู้วิจัยจึงสนใจที่จะนำวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ได้รับการยอมรับว่ามีประสิทธิภาพมาใช้ตรวจสอบ DIF ในแบบสอบพหุมิติ ภายใต้เกณฑ์การจับคู่เปรียบเทียบอื่นนอกจากคะแนนรวม (total test score) เนื่องจากมีงานวิจัยพบว่าการตรวจสอบ DIF ในแบบสอบที่มีมิติซับซ้อนการนำคะแนนรวมมาใช้เป็นเกณฑ์ในการจับคู่เปรียบเทียบนั้นไม่เหมาะสมเพราะจะทำให้อัตราความคลาดเคลื่อนประเภทที่ 1 สูงขึ้น (Clauser, 1993) ดังนั้นในการวิจัยนี้ผู้วิจัยจึงได้เลือกเกณฑ์การจับคู่เปรียบเทียบมาใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบสอบชนิดพหุมิติ ได้แก่ คะแนนรวม (total test score) คะแนนแบบสอบย่อย (subtest score) และคะแนนหลายแบบสอบย่อย (multiple subtest scores) เพื่อให้ได้แนวทางในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบสอบชนิดพหุมิติที่มีประสิทธิภาพที่สุด

ดังนั้นคำถามวิจัยที่น่าสนใจ คือ ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบสอบชนิดพหุมิติ วิธีการตรวจสอบใดที่เป็นวิธีที่มีประสิทธิภาพมากที่สุดและการใช้เกณฑ์จับคู่เปรียบเทียบแตกต่างกันจะทำให้ผลการตรวจสอบ DIF มีประสิทธิภาพแตกต่างกันมากน้อยเพียงใด ซึ่งจากการวิจัยของ Clauser และคณะ (1996) พบว่า การนำคะแนนหลายแบบสอบย่อย (multiple subtest scores) มาเป็นเกณฑ์จับคู่ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบชนิดพหุมิติจะมีประสิทธิภาพมากกว่าคะแนนรวม (total test score) และคะแนนแบบสอบย่อย (subtest score) โดยพิจารณาจากจำนวนข้อสอบที่ถูกตรวจสอบว่าทำหน้าที่ต่างกันของข้อสอบที่มีจำนวนน้อยที่สุด ซึ่งหากพิจารณาถึงอัตราความถูกต้องของการตรวจสอบและอัตราความคลาดเคลื่อนควบคู่กันไปด้วยแล้ว คะแนนหลายแบบสอบย่อยยังเป็นเกณฑ์จับคู่เปรียบเทียบที่มีประสิทธิภาพสูงสุดอยู่หรือไม่

วัตถุประสงค์ของการวิจัย

1. เพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบชนิดพหุมิติด้วยวิธีแมนเทล - แฮนส์เชลกับวิธีถดถอยโลจิสติก เมื่อใช้เกณฑ์การจับคู่เปรียบเทียบ 3 เกณฑ์ คือ คะแนนรวม คะแนนแบบสอบย่อย และคะแนนหลายแบบสอบย่อย
2. เพื่อเปรียบเทียบประสิทธิภาพด้านอัตราความถูกต้องและอัตราความคลาดเคลื่อนของผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบชนิดพหุมิติ ด้วยวิธีแมนเทล - แฮนส์เชลกับวิธีถดถอยโลจิสติกเมื่อใช้เกณฑ์การจับคู่เปรียบเทียบ 3 เกณฑ์ คือ คะแนนรวม คะแนนแบบสอบย่อย และคะแนนหลายแบบสอบย่อย

สมมติฐานของการวิจัย

จากการวิจัยของ Clauser และคณะ (1996) เปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบสอบชนิดพหุมิติ โดยเปรียบเทียบผลการตรวจสอบโดยใช้เกณฑ์การจับคู่ 3 เกณฑ์ ได้แก่ คะแนนรวม (total test score) คะแนนแบบสอบย่อย (subtest score) และคะแนนหลายแบบสอบย่อย (multiple subtest scores) พบว่า การใช้คะแนนรวมเป็นเกณฑ์ในการจับคู่นั้นไม่เหมาะสม แต่เมื่อใช้คะแนนหลายแบบย่อยเป็นเกณฑ์จะส่งผลให้การตรวจสอบมีประสิทธิภาพมากขึ้น และงานวิจัยของ Clauser และคณะ (1996) ยังได้ศึกษาเพิ่มเติมโดยการนำเกณฑ์ภายนอกอื่นซึ่งเป็นความสามารถหลากหลายเข้ามาเป็นเกณฑ์จับคู่ร่วมกับคะแนนรวมจากแบบสอบ

ในการตรวจสอบ DIF ผลปรากฏว่า ข้อสอบที่ถูกตรวจพบว่าทำหน้าที่ต่างกันมีจำนวนลดลง และการนำเกณฑ์ความสามารถมาวิเคราะห์ด้วยจะทำให้ผลการตรวจสอบมีความถูกต้องมากขึ้นซึ่งวิธีที่เหมาะสมกับการตรวจสอบ DIF ในลักษณะนี้ คือ วิธีถดถอยโลจิสติก นอกจากนี้ผลการศึกษาของ Mazor และ Clauser (1996) สนับสนุนว่าวิธีถดถอยโลจิสติกสามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ดีกว่าวิธีแมนเทิล - แฮนส์เซล ในแบบสอบที่มีตัวแปรความสามารถสองตัวหรือมากกว่า ทั้งนี้เป็นเพราะความยืดหยุ่นของวิธีถดถอยโลจิสติกที่มีพื้นฐานอยู่บนสมการถดถอยโลจิสติก ซึ่งสามารถเพิ่มตัวแปรความสามารถหลาย ๆ ตัวเข้าไปในการวิเคราะห์ได้ จึงทำให้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ดียิ่งขึ้น จากการศึกษาที่ผ่านมาทำให้ผู้วิจัยตั้งสมมติฐานของการวิจัยดังนี้ คือ

1. การเปรียบเทียบผลตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบสอบชนิดพหุมิติ โดยแยกตามการใช้เกณฑ์จับคู่เปรียบเทียบ วิธีถดถอยโลจิสติกเมื่อใช้คะแนนหลายแบบสอบย่อย (multiple subtest scores) เป็นเกณฑ์จับคู่จะเป็นวิธีที่มีประสิทธิภาพที่สุด เมื่อพิจารณาทั้งอัตราความถูกต้องและอัตราความคลาดเคลื่อนของผลการตรวจสอบ
2. การเปรียบเทียบผลตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบสอบชนิดพหุมิติ ระหว่างวิธีแมนเทิล-แฮนส์เซลกับวิธีถดถอยโลจิสติก เมื่อพิจารณาตามเกณฑ์จับคู่เปรียบเทียบ วิธีถดถอยโลจิสติกจะมีอัตราความถูกต้องของผลการตรวจสอบมากกว่าวิธีแมนเทิล-แฮนส์เซล
3. การเปรียบเทียบผลตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบสอบชนิดพหุมิติ ระหว่างวิธีแมนเทิล-แฮนส์เซลกับวิธีถดถอยโลจิสติก เมื่อพิจารณาตามเกณฑ์จับคู่เปรียบเทียบ วิธีถดถอยโลจิสติกจะมีอัตราความคลาดเคลื่อนน้อยกว่าวิธีแมนเทิล-แฮนส์เซล

ขอบเขตของการวิจัย

1. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในการศึกษาครั้งนี้ใช้วิธีตรวจสอบ 2 วิธี คือ วิธีแมนเทิล-แฮนส์เซล (MH) และวิธีถดถอยโลจิสติก (LR)
2. ตัวแปรที่ใช้ศึกษาในครั้งนี้
 - ตัวแปรอิสระ คือ วิธีในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบภายใต้เกณฑ์ที่ใช้ในการจับคู่เปรียบเทียบ ซึ่งมีรายละเอียดดังนี้
 - 2.1 วิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ 2 วิธี คือวิธีแมนเทิล-แฮนส์เซล (MH) และวิธีถดถอยโลจิสติก (LR)

2.2 เกณฑ์ที่ใช้ในการจับคู่เปรียบเทียบ มี 3 เกณฑ์ ได้แก่ คะแนนรวม (total test score) คะแนนแบบสอบย่อย (subtest score) และคะแนนหลายแบบสอบย่อย (multiple subtest scores)

โดยเกณฑ์จับคู่คะแนนรวมและคะแนนแบบสอบย่อยใช้เป็นเกณฑ์จับคู่เพื่อเปรียบเทียบประสิทธิภาพของวิธีในการตรวจสอบทั้งสองวิธี และเนื่องจากข้อดีของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีถดถอยโลจิสติกมีความยืดหยุ่นในการนำตัวแทนความสามารถอื่นมาใช้เป็นเกณฑ์จับคู่ได้ จึงเพิ่มเกณฑ์จับคู่คะแนนหลายแบบสอบย่อยเข้ามาใช้ในการตรวจสอบด้วย ดังนั้น ตัวแปรอิสระในการวิจัยครั้งนี้จึงมีระดับของการศึกษา 5 ลักษณะ ได้แก่

2.2.1 ตรวจสอบด้วยวิธีแมนเทล-แฮนส์เซล โดยใช้คะแนนรวมเป็นเกณฑ์ (MH_{TOT})

2.2.2 ตรวจสอบด้วยวิธีแมนเทล-แฮนส์เซลโดยใช้คะแนนแบบสอบย่อยเป็นเกณฑ์ (MH_{SUB})

2.2.3 ตรวจสอบด้วยวิธีถดถอยโลจิสติก โดยใช้คะแนนรวมเป็นเกณฑ์ (LR_{TOT})

2.2.4 ตรวจสอบด้วยวิธีถดถอยโลจิสติกโดยใช้คะแนนแบบสอบย่อยเป็นเกณฑ์ (LR_{SUB})

2.2.5 ตรวจสอบด้วยวิธีถดถอยโลจิสติก โดยใช้คะแนนหลายแบบสอบย่อยเป็นเกณฑ์

($LR_{MULTIPLE}$)

ตัวแปรตาม คือ ประสิทธิภาพในการตรวจสอบ โดยพิจารณาจากอัตราความถูกต้อง และอัตราความคลาดเคลื่อนของผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

3. ประชากรและกลุ่มตัวอย่างที่ใช้ศึกษาในงานวิจัยครั้งนี้ คือ นักเรียนชั้นประถมศึกษาปีที่ 6 ปีการศึกษา 2540 ในโรงเรียนสังกัดสำนักงานการประถมศึกษากรุงเทพมหานคร

4. เกณฑ์ที่ใช้ในการวิเคราะห์ว่าข้อสอบทำหน้าที่ต่างกัน มีดังนี้

4.1 วิธีแมนเทล-แฮนส์เซล ได้แก่ ค่า $\alpha_{MH} \neq 1.0$ และการทดสอบนัยสำคัญด้วย χ^2_{MH} ที่ระดับ .05

4.2 วิธีถดถอยโลจิสติก ได้แก่ ค่าอิทธิพลจากปฏิสัมพันธ์ระหว่างกลุ่มผู้สอบและความสามารถของผู้สอบหรือค่าอิทธิพลจากกลุ่มผู้สอบ และการทดสอบนัยสำคัญด้วย χ^2 ที่ระดับ .05

4.3 วิธี SIBTEST ได้แก่ ค่า $\beta_n > 0$ และการทดสอบนัยสำคัญด้วย Z ที่ระดับ .05

ข้อตกลงเบื้องต้น

1. เนื่องจากแบบทดสอบความสามารถทางคณิตศาสตร์ชั้นประถมศึกษาปีที่ 6 มีจำนวน 80 ข้อ เพื่อสะดวกแก่การเก็บรวบรวมข้อมูลและหลีกเลี่ยงความคลาดเคลื่อนอันเกิดจากความเมื่อยล้าในการทำแบบสอบ ผู้วิจัยจึงแบ่งแบบทดสอบออกเป็น 2 ฉบับ ๆ ละ 40 ข้อ ใช้เวลาใน

การทดสอบฉบับละ 1 คาบ (50 นาที) รวมเป็นเวลา 2 คาบ ทั้งนี้ได้ทิ้งช่วงเวลาระหว่างการทำแบบสอบทั้ง 2 ชุด แบ่งสอบเป็นช่วงเช้าและช่วงบ่าย โดยเวลาที่แตกต่างกันนั้นถือว่าไม่มีผลต่อการเปลี่ยนแปลงความสามารถของกลุ่มตัวอย่าง

2. วิธีชิบเทสท์ เป็นวิธีที่นำวิธีการทางสถิติมาใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยพิจารณาโครงสร้างภายในของแบบสอบที่เป็นแบบสอบชนิดพหุมิติ (multidimensional test) โดยมีพื้นฐานบนทฤษฎีการตอบสนองข้อสอบแบบพหุมิติ ในการวิเคราะห์มีข้อตกลงว่ามีมิติการวัด 2 มิติ คือ มิติลักษณะแง่เป้าหมายที่ต้องการวัดและลักษณะแง่แทรกซ้อนที่ไม่ต้องการวัด (Shealy and Stout, 1993) จึงเป็นวิธีที่ใช้ระบุว่าข้อสอบข้อใดทำหน้าที่ต่างกันแบบสอบชนิดพหุมิติที่ถือว่าถูกต้องใกล้เคียงความจริง เพื่อนำมาใช้เป็นเกณฑ์ในการประเมินอัตราความถูกต้องและอัตราความคลาดเคลื่อนในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมนเทส-แฮนส์เทสกับวิธีตัดรอยโลจิสติก

คำจำกัดความที่ใช้ในการวิจัย

แบบสอบชนิดพหุมิติ (multidimensional test) หมายถึง แบบสอบที่ผู้วิจัยสร้างขึ้น ใช้วัดความสามารถเกี่ยวกับทักษะวิชาคณิตศาสตร์ในระดับชั้นประถมศึกษาปีที่ 6 ซึ่งมีองค์ประกอบในการวัด 4 ด้าน ได้แก่ ความรู้ความเข้าใจทางคณิตศาสตร์ การคิดคำนวณ การพิจารณาผลลัพธ์ที่สมเหตุสมผลและการแก้โจทย์ปัญหา แบบสอบชนิดพหุมิติเป็นข้อสอบปรนัยแบบเลือกตอบ 4 ตัวเลือก โดยแบ่งออกเป็นแบบสอบย่อยตามองค์ประกอบในการวัด จำนวน 2 ฉบับ ๆ ละ 40 ข้อ

ข้อสอบทำหน้าที่ต่างกัน (DIF) หมายถึง ข้อสอบที่ทำให้ผลการตอบของผู้สอบที่มีความสามารถในสิ่งที่ต้องการวัดเท่ากัน มีโอกาสตอบข้อสอบข้อนั้นได้ถูกต้องไม่เท่ากันเนื่องจากอยู่ในกลุ่มย่อยต่างกัน ซึ่งในการวิจัยนี้ศึกษาการทำหน้าที่ต่างกันของข้อสอบสำหรับกลุ่มผู้สอบจำแนกตามเพศ

เกณฑ์จับคู่เปรียบเทียบ (matching criteria) หมายถึง เกณฑ์ที่ใช้แทนความสามารถที่แท้จริงของผู้สอบสองกลุ่มซึ่งใช้ในการจับคู่เปรียบเทียบเพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ในการวิจัยครั้งนี้เกณฑ์จับคู่เปรียบเทียบมี 3 เกณฑ์ คือ คะแนนรวม คะแนนแบบสอบย่อย และคะแนนหลายแบบสอบย่อย

- **เกณฑ์จับคู่คะแนนรวม** หมายถึง เกณฑ์ที่ใช้แทนความสามารถที่แท้จริงซึ่งเป็นผลรวมของคะแนนที่ได้จากผลการสอบด้วยแบบทดสอบความสามารถทางคณิตศาสตร์ชั้นประถมศึกษาปีที่ 6 ทั้งฉบับ ซึ่งในการวิเคราะห์จะใช้ตัวแปรความสามารถของผู้สอบ คือ คะแนนรวมของผลการสอบทั้งฉบับ เป็นเกณฑ์จับคู่ระหว่างนักเรียนชายและหญิงเพื่อตรวจสอบ DIF
- **เกณฑ์จับคู่คะแนนแบบสอบย่อย** หมายถึง เกณฑ์ที่ใช้แทนความสามารถที่แท้จริงซึ่งเป็นคะแนนที่ได้จากผลการสอบด้วยแบบทดสอบความสามารถทางคณิตศาสตร์ชั้นประถมศึกษาปีที่ 6 โดยแยกพิจารณาในแต่ละด้านขององค์ประกอบของการวัดทั้ง 4 ด้าน การวิเคราะห์ใช้ตัวแปรความสามารถของผู้สอบ คือ คะแนนผลการตอบข้อสอบรวมในแต่ละองค์ประกอบเป็นเกณฑ์จับคู่ระหว่างนักเรียนชายและหญิงเพื่อตรวจสอบ DIF โดยแยกวิเคราะห์ทีละองค์ประกอบ
- **เกณฑ์จับคู่คะแนนหลายแบบสอบย่อย** หมายถึง เกณฑ์ที่ใช้แทนความสามารถที่แท้จริงซึ่งเป็นคะแนนที่ได้จากผลการสอบด้วยแบบทดสอบความสามารถทางคณิตศาสตร์ชั้นประถมศึกษาปีที่ 6 การวิเคราะห์ใช้ตัวแปรความสามารถของผู้สอบ คือ คะแนนผลการตอบข้อสอบรวมในแต่ละองค์ประกอบทุกองค์ประกอบ เป็นตัวทำนายโอกาสในการตอบข้อสอบได้ถูกต้องร่วมกันระหว่างนักเรียนชายและหญิงเพื่อตรวจสอบ DIF

วิธี MH หมายถึง วิธีการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบของแมนเทล-เฮนส์เซล (Mantel - Haenszel) ซึ่งเป็นการวิเคราะห์ความแตกต่างของสัดส่วนการตอบข้อสอบระหว่างผู้สอบที่มีความสามารถในระดับเดียวกัน

วิธี LR หมายถึง วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ทดสอบด้วยโมเดลทางสถิติ Logistic Regression เพื่อทำนายโอกาสในการตอบข้อสอบได้ถูกต้อง ซึ่งพิจารณาจากจุดตัดและค่าความชันของพารามิเตอร์ รวมทั้งความสามารถของผู้สอบแต่ละคนในกลุ่มผู้สอบ

วิธี SIBTEST หมายถึง วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่พิจารณาจากความแตกต่างของคะแนนจริงระหว่างผู้สอบที่มีความสามารถระดับเดียวกัน

ประสิทธิภาพของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ หมายถึง ความถูกต้องของการระบุการทำหน้าที่ต่างกันของข้อสอบ จากการตรวจสอบด้วยวิธี MH และวิธี LR ซึ่งพิจารณาได้จากอัตราความถูกต้องของการตรวจสอบและความคลาดเคลื่อน

อัตราความถูกต้องของการตรวจสอบ หมายถึง สัดส่วนหรือร้อยละของจำนวนข้อสอบที่ตรวจพบว่าถูกต้องตรงกับวิธี SIBTEST ซึ่งมี 2 ประเภท คือ อัตราความถูกต้องประเภทที่ 1 และอัตราความถูกต้องประเภทที่ 2

- อัตราความถูกต้องประเภทที่ 1 หมายถึง สัดส่วนหรือร้อยละของจำนวนข้อสอบที่ตรวจพบว่าทำหน้าที่ไม่ต่างกันได้ถูกต้องตรงกับวิธี SIBTEST ในการวิจัยนี้คำนวณจากสัดส่วนของจำนวนข้อสอบที่ทำหน้าที่ไม่ต่างกันได้ถูกต้องต่อจำนวนข้อสอบที่ทำหน้าที่ไม่ต่างกันทั้งหมดในแบบสอบ
- อัตราความถูกต้องประเภทที่ 2 หมายถึง สัดส่วนหรือร้อยละของจำนวนข้อสอบที่ตรวจพบว่าทำหน้าที่ต่างกันได้ถูกต้องตรงกับวิธี SIBTEST ในการวิจัยนี้คำนวณจากสัดส่วนของจำนวนข้อสอบที่ทำหน้าที่ต่างกันได้ถูกต้องต่อจำนวนข้อสอบที่ทำหน้าที่ต่างกันทั้งหมดในแบบสอบ

อัตราความคลาดเคลื่อนของการตรวจสอบ หมายถึง สัดส่วนหรือร้อยละของจำนวนข้อสอบที่ระบุผิดพลาดซึ่งมี 2 ประเภท คือ ความคลาดเคลื่อนประเภทที่ 1 และความคลาดเคลื่อนประเภทที่ 2

- ความคลาดเคลื่อนประเภทที่ 1 (type I error) หมายถึง การระบุผิดพลาดว่าข้อสอบทำหน้าที่ต่างกัน (fault positive) ทั้งที่ความเป็นจริงข้อสอบทำหน้าที่ไม่ต่างกัน ซึ่งในการวิจัยนี้คำนวณได้จากค่าสัดส่วนของจำนวนข้อสอบที่ระบุผิดพลาดว่าทำหน้าที่ต่างกันต่อจำนวนข้อสอบที่ทำหน้าที่ไม่ต่างกันทั้งหมดในแบบสอบ
- ความคลาดเคลื่อนประเภทที่ 2 (type II error) หมายถึง การระบุผิดพลาดว่าข้อสอบทำหน้าที่ไม่ต่างกัน (fault negative) ทั้งที่ความเป็นจริงข้อสอบทำหน้าที่ต่างกัน ซึ่งในการวิจัยนี้คำนวณได้จากค่าสัดส่วนของจำนวนข้อสอบที่ระบุผิดพลาดว่าทำหน้าที่ไม่ต่างกันต่อจำนวนข้อสอบที่ทำหน้าที่ต่างกันทั้งหมดในแบบสอบ

กลุ่มอ้างอิง (reference group) หมายถึง กลุ่มผู้สอบที่คาดว่าจะได้ประโยชน์จากการตอบข้อสอบที่ทำหน้าที่ต่างกัน คือเป็นกลุ่มที่มีความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องสูงกว่าผู้สอบอีกกลุ่มหนึ่งทั้ง ๆ ที่มีความสามารถเท่ากัน ในการวิจัยนี้กลุ่มอ้างอิง คือ เพศชาย

กลุ่มเปรียบเทียบ (focal group) หมายถึง กลุ่มของผู้สอบที่คาดว่าจะเสียประโยชน์จากการตอบข้อสอบที่ทำหน้าที่ต่างกัน คือเป็นกลุ่มที่มีความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องต่ำกว่าผู้สอบอีกกลุ่มหนึ่งทั้ง ๆ ที่มีความสามารถเท่ากันในการวิจัยนี้กลุ่มเปรียบเทียบ คือ เพศหญิง

ข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูป (uniform DIF) หมายถึง ข้อสอบที่ทำหน้าที่ต่างกันสำหรับผู้สอบใน กลุ่มอ้างอิงและกลุ่มเปรียบเทียบในลักษณะเดียวกันทุกระดับความสามารถของผู้สอบ

ข้อสอบที่ทำหน้าที่ต่างกันแบบอนเอกรูป (nonuniform DIF) หมายถึง ข้อสอบที่ทำหน้าที่ต่างกันสำหรับผู้สอบในกลุ่มอ้างอิงและกลุ่มเปรียบเทียบในลักษณะที่ไม่สม่ำเสมอทุกระดับความสามารถของผู้สอบ

ความเที่ยง (Reliability) หมายถึง คุณภาพของแบบสอบซึ่งพิจารณาได้จากความสอดคล้องภายใน (Internal Consistency) ซึ่งคำนวณค่าความเที่ยงของแบบสอบ โดยใช้สูตร KR-20 ของคูเดอร์-ริชาร์ดสัน

ความตรง (Validity) หมายถึง คุณภาพของแบบสอบซึ่งพิจารณาจากความตรงเชิงโครงสร้าง (Construct Validity) โดยการตรวจสอบด้วยวิธีวิเคราะห์องค์ประกอบเชิงยืนยัน (Confirmatory Factor Analysis: CFA)

ประโยชน์ที่คาดว่าจะได้รับ

1. ได้วิธีและเกณฑ์การจับคู่เปรียบเทียบที่เหมาะสมในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบสอบชนิดพหุมิติ
2. เพื่อเป็นแนวทางในการศึกษาและพัฒนาเกณฑ์ในการจับคู่เปรียบเทียบที่เหมาะสมในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ
3. เพื่อเป็นแนวทางในการศึกษาเกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบสอบชนิดพหุมิติต่อไป