

บทที่ 1

บทนำ



ความเป็นมาและความสำคัญของปัญหา

แบบสอบเป็นเครื่องมือที่ใช้สำหรับวัดและประเมินผลการศึกษา เพื่อให้ได้ข้อมูลหรือสารสนเทศเกี่ยวกับการจัดกิจกรรมทางการศึกษา ผลที่ได้จากการสอบมีความสำคัญต่อการพัฒนาคุณภาพการศึกษา ตลอดจนใช้เป็นข้อมูลพื้นฐานในการตัดสินใจของครู ผู้บริหาร และผู้เกี่ยวข้องกับการจัดการศึกษาทุกระดับ ดังนั้นในการพัฒนาแบบสอบจึงต้องคำนึงถึงคุณภาพของแบบสอบ กล่าวคือแบบสอบนั้นสามารถวัดได้ตรงกับคุณลักษณะแ่งที่ต้องการวัดหรือแบบสอบนั้นทำหน้าที่ได้ตามวัตถุประสงค์ที่กำหนด ไม่มีการได้ประโยชน์หรือเสียประโยชน์จากผู้สอบแต่ละกลุ่ม ทั้งนี้ในการสอบแต่ละครั้งอาจจำแนกผู้สอบออกเป็นกลุ่มย่อยตามลักษณะที่แตกต่างกันในด้านต่างๆ เช่น เชื้อชาติ เพศ ศาสนา ประสบการณ์ เป็นต้น ข้อสอบหรือแบบสอบที่เอื้อต่อผู้สอบกลุ่มย่อยบางกลุ่ม อาจทำให้ผู้สอบกลุ่มย่อยนั้นได้รับประโยชน์มากกว่าผู้สอบกลุ่มย่อยอื่นๆ ทั้งๆที่มีความสามารถเท่ากันและสอบด้วยข้อสอบข้อเดียวกันหรือแบบสอบฉบับเดียวกัน ดังนั้นในการตรวจสอบว่าข้อสอบมีความตรงและมีการได้ประโยชน์หรือเสียประโยชน์ระหว่างผู้สอบกลุ่มย่อยหรือไม่นั้น วิธีการในการตรวจสอบการทำหน้าที่ต่างกันเป็นวิธีหนึ่งที่น่าสนใจนำมาใช้ตรวจสอบเพื่อช่วยพัฒนาแบบสอบให้มีคุณภาพต่อไป

วิธีการในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมีหลายวิธี ทั้งนี้เพราะมีการศึกษาและคิดค้นวิธีการต่างๆ เพื่อให้สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้อย่างมีประสิทธิภาพมากที่สุด ซึ่งสามารถแบ่งตามประเภทการวิเคราะห์ ได้เป็น 2 กลุ่ม ดังนี้ (Potenza and Dorans, 1995)

1. กลุ่มที่ใช้คะแนนที่สังเกตได้ (observed score) เป็นกลุ่มวิธีที่ใช้คะแนนรวมของแบบสอบเป็นเกณฑ์การจับคู่ผู้สอบสองกลุ่มตามความรู้หรือความสามารถที่แท้จริงของผู้สอบ วิธีการในกลุ่มนี้ได้แก่ วิธีแมนเทล-แฮนส์เชล (Holland and Thayer, 1988) วิธีถดถอยโลจิสติก (Swaminathan and Rogers, 1990) และวิธีทำให้เป็นมาตรฐาน (Dorans and Holland, 1993) จุดเด่นของวิธีการในกลุ่มนี้คือใช้กลุ่มตัวอย่างขนาดเล็ก การวิเคราะห์ไม่ยุ่งยากซับซ้อน ส่วนจุดด้อยของวิธีการในกลุ่มนี้คือค่าสถิติเปลี่ยนแปลงไปตามขนาดกลุ่มตัวอย่าง เมื่อกลุ่มตัวอย่างที่ศึกษาเปลี่ยนแปลงการตรวจสอบก็อาจเปลี่ยนแปลงด้วย

2. กลุ่มที่ใช้คะแนนที่สังเกตไม่ได้หรือใช้ตัวแปรแฝง (latent variable) เป็นกลุ่มวิธีที่มีทฤษฎีการทดสอบเป็นพื้นฐาน เช่นทฤษฎีการทดสอบแบบดั้งเดิม และทฤษฎีการตอบสนองข้อสอบ ใช้การประมาณค่าคุณลักษณะแฝง (latent trait) หรือใช้คะแนนจริงของผู้สอบเป็นเกณฑ์ในการจับคู่เปรียบเทียบผู้สอบ วิธีการในกลุ่มนี้ได้แก่ วิธีทฤษฎีการสนองข้อสอบ (IRT) และวิธีชิบเทสต์ (Shealy and Stout, 1993)

สำหรับแต่ละวิธี มีวิธีการพอสังเขปดังนี้

วิธีแมนเทล-แฮนส์เซล (Mantel-Haenszel : MH) เป็นวิธีที่พัฒนามาจากไควสแควร์แบบดั้งเดิม ใช้คะแนนรวมจากแบบสอบเป็นตัวแทนความสามารถโดยวิเคราะห์ที่ระดับความสามารถ มีดัชนีบอกระดับการทำหน้าที่ต่างกันของข้อสอบและการทดสอบนัยสำคัญทางสถิติ เป็นวิธีที่มีความสอดคล้องกับวิธี IRT อีกทั้งสามารถใช้วิธี MH แทนวิธี IRT ได้อย่างประหยัดกว่า (Hambleton et al., 1986 อ้างถึงใน กาญจนา วัชรสุนทร, 2537) ข้อดีของวิธีนี้คือคำนวณง่าย ใช้ง่ายกับกลุ่มตัวอย่างขนาดเล็ก

วิธีถดถอยโลจิสติก (logistic regression : LR) เป็นวิธีการที่มีโมเดลพื้นฐานทางสถิติจึงสามารถตรวจสอบการทำหน้าที่ต่างกันของแบบสอบได้ถูกต้องใกล้เคียงกับวิธี IRT และเข้าใจธรรมชาติการทำหน้าที่ต่างกันภายในแบบสอบได้ดี วิธีนี้มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของแบบสอบแบบเอกรูป (uniform DIF) ได้เท่าเทียมกับวิธีแมนเทล-แฮนส์เซล และวิธีชิบเทสต์ (Swaminatan and Roger, 1990 ; Ackerman, 1992) นอกจากนี้ยังสามารถตรวจสอบการทำหน้าที่ต่างกันของแบบสอบแบบอนเอกรูป (nonuniform DIF) ได้ดีเช่นเดียวกัน

วิธีทำให้เป็นมาตรฐาน (standardization : STND) โดยพื้นฐานแล้ววิธีนี้เป็นวิธีเชิงบรรยาย ไม่มีการทดสอบนัยสำคัญทางสถิติ หลักการสำคัญคือเป็นการเปรียบเทียบการถดถอยระหว่างคะแนนข้อสอบกับคะแนนแบบสอบของผลการตอบข้อสอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ มีหลักการวิเคราะห์การทำหน้าที่ต่างกันคล้ายกับวิธีแมนเทล-แฮนส์เซล ตรงที่ใช้หลักความแตกต่างระหว่างสัดส่วนการตอบข้อสอบที่ถูกที่ควรจะเป็นกับที่สังเกตได้ระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ให้ดัชนีการทำหน้าที่ต่างกัน 2 ค่า คือ ดัชนีคิดเครื่องหมาย (signed index) และดัชนีไม่คิดเครื่องหมาย (unsigned index) ข้อดีของวิธีนี้คือคำนวณง่าย เสียค่าใช้จ่ายน้อย นำไปใช้อธิบายธรรมชาติของข้อสอบทำหน้าที่ต่างกันได้ ส่วนข้อด้อยคือต้องใช้กลุ่มตัวอย่างขนาดใหญ่

วิธีทฤษฎีการตอบสนองข้อสอบ (IRT) เป็นวิธีที่วิเคราะห์ความแตกต่างของฟังก์ชันการตอบสนองข้อสอบ โดยเปรียบเทียบโค้งคุณลักษณะข้อสอบ (item characteristic curves : ICCs) ของกลุ่มผู้สอบตามระดับความสามารถผู้สอบ ถ้าโค้งคุณลักษณะข้อสอบของกลุ่มผู้สอบสองกลุ่ม

มีรูปร่างเหมือนกันแสดงว่าข้อสอบข้อนั้นทำหน้าที่ไม่ต่างกัน แต่ถ้าโค้งคุณลักษณะข้อสอบของผู้สอบสองกลุ่มมีรูปร่างต่างกันแสดงว่าข้อสอบข้อนั้นทำหน้าที่ต่างกัน จุดเด่นของวิธีนี้คือเป็นวิธีที่มีความถูกต้องสูง ตรวจพบจำนวนข้อสอบที่ทำหน้าที่ต่างกันมากที่สุด (Holland and Thayer, 1988) ค่าสถิติของข้อสอบไม่เปลี่ยนไปตามกลุ่มตัวอย่าง แต่มีข้อจำกัดคือต้องใช้กลุ่มตัวอย่างขนาดใหญ่มีการคำนวณซับซ้อนหลายรอบ แปรผลยาก เสียค่าใช้จ่ายสูง (Ryan, 1991 ; Osterlind, 1993 ; Narayanan and Swaminatan, 1994 ; สุรศักดิ์ อมรรัตนศักดิ์, 2531)

วิธีซิบเทสต์ (SIBTEST) เป็นวิธีที่มีพื้นฐานอยู่บนทฤษฎีการตอบข้อสอบแบบพหุมิติ ใช้คะแนนรวมจากแบบสอบเป็นตัวแทนความสามารถโดยมีข้อตกลงเบื้องต้นว่ามีมิติการวัด 2 มิติ ดังนั้นคะแนนรวมจากแบบสอบจึงมี 2 ส่วนคือคะแนนแบบสอบที่มีความตรง (valid subtest) ซึ่งวัดคุณลักษณะแฝงเป้าหมาย และคะแนนรวมจากแบบสอบที่ศึกษา (studied subtest) ซึ่งวัดคุณลักษณะแฝงแทรกซ้อน มีดัชนีบอกระดับการทำหน้าที่ต่างกันของข้อสอบและการทดสอบนัยสำคัญทางสถิติ วิธีนี้มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปได้ใกล้เคียงกับวิธีแมนเทิล-แฮนส์เชล (Shealy and Stout, 1993 ; Narayanan and Swaminatan, 1994 ; จิตินา วรณศรี, 2540)

ผลจากการเปรียบเทียบวิธีการต่างๆที่นำมาใช้ในการตรวจสอบการทำหน้าที่ต่างกันควรคำนึงถึงอัตราความคลาดเคลื่อนในการตรวจสอบของแต่ละวิธี ถ้าวิธีการในการตรวจสอบการทำหน้าที่ต่างกันมีอัตราความคลาดเคลื่อนประเภทที่ 1 สูง จะเกิดผลกระทบ (effect) ที่รุนแรงภายในแบบสอบดังนี้ (Roussos and Stout, 1996)

1. มีการตัดข้อสอบที่ดีออกจากแบบสอบ ทำให้การสร้างและการพัฒนาแบบสอบมีความเสี่ยงสูง หมดเปลืองทรัพยากรไปโดยไม่จำเป็น

2. เป็นอุปสรรคของความพยายามที่จะเข้าใจสาเหตุของการทำหน้าที่ต่างกัน ทั้งที่ความจริงเมื่อนำข้อสอบปกติมาตรวจสอบการทำหน้าที่ต่างกันแล้วพบว่ามีค่าต่ำ อัตราความคลาดเคลื่อนประเภทที่ 1 สูง สามารถให้เหตุผลได้ว่าการระบุผิดพลาด

3. ข้อสอบที่มีค่าอำนาจจำแนกสูงซึ่งเป็นคุณสมบัติประการหนึ่งของแบบสอบที่ดีเมื่อมีการระบุว่ามีความคลาดเคลื่อนประเภทที่ 1 สูง จะนำไปสู่การย้ายข้อสอบที่ไม่เหมาะสม กล่าวคือย้ายข้อสอบที่ดีออกจากแบบสอบ ดังนั้นการแสดงอัตราความคลาดเคลื่อนที่ 1 ของสถิติที่ตรวจสอบการทำหน้าที่ต่างกันจึงเป็นสิ่งสำคัญในการปฏิบัติและเป็นหลักฐานของความเชื่อถือได้ในประสิทธิภาพของการทำหน้าที่ต่างกัน

ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ พบว่ามีการระบุลักษณะของข้อสอบที่ทำหน้าที่ต่างกัน ใน 2 ประเภท (Mellenbergh, 1982) คือ

1. การทำหน้าที่ต่างกันของข้อสอบแบบเอกกรุป (uniform DIF) หมายถึงข้อสอบที่ทำให้ผู้สอบกลุ่มหนึ่งมีโอกาสในการตอบข้อสอบถูกมากกว่าผู้สอบอีกกลุ่มหนึ่งสม่ำเสมอในทุกระดับความสามารถ เมื่อพิจารณาโค้งคุณลักษณะข้อสอบของผู้สอบ 2 กลุ่มจะพบว่าไม่มีปฏิสัมพันธ์ระหว่างโค้งคุณลักษณะข้อสอบในทุกระดับความสามารถ

2. การทำหน้าที่ต่างกันของข้อสอบแบบอนเอกกรุป (nonuniform DIF) หมายถึงข้อสอบที่ทำให้โอกาสในการตอบข้อสอบถูกของผู้สอบระหว่างกลุ่มไม่สม่ำเสมอในทุกระดับความสามารถ เมื่อพิจารณาโค้งคุณลักษณะข้อสอบของผู้สอบ 2 กลุ่ม พบว่ามีปฏิสัมพันธ์ร่วมกันระหว่างโค้งคุณลักษณะ เช่น ที่ระดับความสามารถหนึ่ง กลุ่มผู้สอบกลุ่ม R มีโอกาสในการตอบข้อสอบถูกมากกว่ากลุ่มผู้สอบกลุ่ม F แต่ที่ระดับความสามารถอีกระดับหนึ่งกลุ่มผู้สอบกลุ่ม F มีโอกาสในการตอบข้อสอบถูกมากกว่ากลุ่มผู้สอบกลุ่ม R

จากการศึกษาการทำหน้าที่ต่างกันของข้อสอบแบบเอกกรุป (uniform DIF) ที่ผ่านมาพบว่าวิธีแมนเทิล-แฮนส์เชล เป็นวิธีที่นิยมและใช้กันอย่างกว้างขวางในการตรวจสอบการทำหน้าที่ต่างกันของแบบสอบ (French and Miller, 1996 ; Roussos and Stout, 1996) เพราะเป็นวิธีที่ไม่ยุ่งยาก ไม่ซับซ้อน ใช้กลุ่มตัวอย่างขนาดเล็ก คำนวณง่าย ประหยัดเวลาและค่าใช้จ่าย อีกทั้งมีสถิติที่เชื่อถือได้ ผลการวิเคราะห์สอดคล้องกับวิธี IRT วิธีถดถอยโลจิสติก และวิธีซิปเทสท์ เช่น ผลการศึกษาของ Hambleton และคณะ (1989) พบว่าวิธีแมนเทิล-แฮนส์เชล วิธีลงจุดค่าความยาก วิธีค่าความแตกต่างของค่าเฉลี่ยกำลังสอง และวิธีพื้นที่รวม ให้ผลการตรวจสอบการทำหน้าที่ต่างกันได้ใกล้เคียงกัน และวิธี MH สามารถเลือกใช้แทนวิธี IRT ได้อย่างประหยัดกว่า ข้อค้นพบของ Rogers และ Swaminatan (1993) พบว่าวิธี MH มีประสิทธิภาพเท่าเทียมกับวิธีถดถอยโลจิสติกในการตรวจสอบการทำหน้าที่ต่างกันแบบเอกกรุป ผลการศึกษาของ Narayanan และ Swaminatan (1994) ได้ศึกษาเปรียบเทียบประสิทธิภาพระหว่างวิธี MH กับวิธี SIBTEST พบว่าเมื่อการกระจายความสามารถของผู้สอบสองกลุ่มเท่ากันทั้งสองวิธีมีประสิทธิภาพเท่าเทียมกัน

เพ็ญพนา สุขสม (2539) เปรียบเทียบวิธีวิเคราะห์ความลำเอียงข้อสอบ 3 วิธี ได้แก่ วิธีค่าความยากแปลง วิธี IRT สามพารามิเตอร์และวิธี MH พบว่าวิธี MH ตรวจพบข้อสอบที่ลำเอียงมากที่สุด และจิตติมา วรณศรี (2540) เปรียบเทียบประสิทธิภาพของวิธี MH และวิธี SIBTEST พบว่าทั้งสองวิธีมีอัตราการตรวจสอบเท่าเทียมกัน และส่วนมากวิธี SIBTEST มีอัตราความคลาดเคลื่อนประเภทที่ 1 มากกว่าวิธี MH เล็กน้อย

สำหรับการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป (nonuniform DIF) โดยทั่วไปพบได้บ่อยในข้อสอบจริง (Mellenbergh, 1982 ; Hambleton and Rogers, 1989) ซึ่งเป็นลักษณะของข้อสอบที่ทำหน้าที่ต่างกันสำหรับผู้สอบสองกลุ่มในลักษณะที่มีปฏิสัมพันธ์ร่วมกันระหว่างกลุ่มกับระดับความสามารถ ในทฤษฎีการตอบสนองข้อสอบการมีปฏิสัมพันธ์ร่วมกันเป็นความแตกต่างของค่าอำนาจจำแนก (a) เมื่อพิจารณาจากโค้งลักษณะข้อสอบของผู้สอบกลุ่มย่อยสองกลุ่มพบว่าไม่ขนานกัน (Mazor., 1994) วิธีที่สามารถนำมาใช้ตรวจสอบการทำหน้าที่ต่างกันแบบอเนกรูปได้คือวิธีแมนเทิล-แฮนส์เชลและวิธีถดถอยโลจิสติก เนื่องจากทั้งสองวิธีมีความสะดวกและง่ายต่อการปฏิบัติและมีโปรแกรมสำเร็จรูปในการคิดคำนวณ (Narayanan and Swaminatan, 1996) จากการศึกษาเปรียบเทียบวิธีการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป ผลการศึกษาของ Rogers และ Swaminatan (1990, 1993) พบว่าวิธีแมนเทิล-แฮนส์เชลมีอัตราการตรวจสอบต่ำกว่าวิธีถดถอยโลจิสติก แต่อัตราความคลาดเคลื่อนชนิดที่ 1 ของวิธีถดถอยโลจิสติก สูงกว่าเล็กน้อย ซึ่งสอดคล้องกับผลการศึกษาของ Narayanan และ Swaminatan (1996) ที่พบว่าวิธีแมนเทิล-แฮนส์เชล มีอัตราการตรวจสอบต่ำกว่าวิธีถดถอยโลจิสติกและวิธี CRO-SIB แต่อัตราความคลาดเคลื่อนชนิดที่ 1 ของวิธีถดถอยโลจิสติกกับวิธี CRO-SIB สูงกว่าวิธีแมนเทิล-แฮนส์เชล ข้อค้นพบดังกล่าวขัดแย้งกับการศึกษาของ Mazor และคณะ (1994) ที่ใช้วิธีแมนเทิล-แฮนส์เชลตรวจสอบการทำหน้าที่ต่างกันแบบอเนกรูป โดยแบ่งกลุ่มตัวอย่างเป็นกลุ่มสูงกลุ่มต่ำ และใช้ค่าเฉลี่ยจากคะแนนของกลุ่มตัวอย่างทุกคนเป็นเกณฑ์ในการแยกกลุ่มสูงกลุ่มต่ำ กลุ่มที่ได้คะแนนต่ำกว่าค่าเฉลี่ยจัดให้เป็นกลุ่มที่มีความสามารถต่ำ และกลุ่มที่มีความสามารถสูงกว่าค่าเฉลี่ยจัดให้เป็นกลุ่มที่มีความสามารถสูง จากนั้นจึงนำเฉพาะกลุ่มสูงหรือกลุ่มต่ำที่ได้มาแบ่งเป็นกลุ่มอ้างอิงและกลุ่มเปรียบเทียบแล้ววิเคราะห์ พบว่าสามารถตรวจสอบการทำหน้าที่ต่างกันแบบอเนกรูปได้ดีโดยไม่ทำให้อัตราความคลาดเคลื่อนประเภทที่ 1 สูงขึ้น ซึ่งข้อค้นพบของ Mazor และคณะ (1994) สอดคล้องกับการศึกษาของเสรี ชัดรัมย์ (2540) ที่พบว่าวิธีแมนเทิล-แฮนส์เชลแบบแบ่งกลุ่มความสามารถสามารถตรวจพบข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปได้ดีกว่าวิธีแมนเทิล-แฮนส์เชลแบบปกติ

ปัจจัยที่มีผลต่อประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปคือค่าความยากและค่าอำนาจจำแนก เช่นผลการศึกษาของ Rogers และ Swaminatan (1993) พบว่าข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปส่วนใหญ่เป็นข้อสอบที่มีค่าความยากง่ายปานกลาง ค่าอำนาจจำแนกสูง ($b = 0, 1.0 \leq a \leq 1.20$ สำหรับกลุ่มอ้างอิง และ $1.40 \leq a \leq 1.80$ สำหรับกลุ่มเปรียบเทียบ) แต่ผลการศึกษาของ Mazor และคณะ (1994) พบว่าข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปที่ตรวจพบส่วนใหญ่เป็นข้อสอบง่าย และพบว่าเมื่อมีความแตกต่างของค่าอำนาจ

จำแนก (a) > 0.5 และค่าความยาก (b) > 0.3 ระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบเพิ่มขึ้นจะทำให้อัตราการตรวจสอบเพิ่มขึ้น ในขณะที่ผลการศึกษาของ Narayanan และ Swaminatan (1996) พบว่าข้อสอบที่ทำหน้าที่ต่างกันส่วนใหญ่เป็นข้อสอบที่มีค่าความยากต่ำและค่าอำนาจจำแนกสูง ($b = -1.50$, $0.47 \leq a \leq 0.90$ สำหรับกลุ่มอ้างอิงและ $1.68 \leq a \leq 2.01$ สำหรับกลุ่มเปรียบเทียบ) ส่วนผลการศึกษาในประเทศไทยของเกษร หว่างจิตร์ (2539) ใช้วิธีแมนเทิล-แฮนส์เชลตรวจสอบการทำหน้าที่ต่างกันของแบบสอบคัดเลือกบัณฑิตศึกษาวิชาภาษาไทยและภาษาอังกฤษ ผลพบว่าข้อสอบที่สามารถตรวจค้นได้ส่วนใหญ่เป็นข้อสอบแบบอเนกรูป มีค่าอำนาจจำแนกค่อนข้างต่ำ (0.20-0.39) ในวิชาภาษาไทยและวิชาภาษาอังกฤษ และเป็นข้อสอบที่ง่ายมาก (0.80-1.00) ในวิชาภาษาไทย แต่ในวิชาภาษาอังกฤษเป็นข้อสอบที่ยาก (0-0.19) และผลการศึกษาของเสรี ชัดแหม่ม (2540) พบว่าข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปที่ตรวจพบส่วนใหญ่เป็นข้อสอบที่มีค่าความยากปานกลาง และข้อสอบง่าย

จากการศึกษางานวิจัยที่ผ่านมา มีประเด็นปัญหาคือ

1. ข้อสอบลักษณะใดที่ทำหน้าที่ต่างกันแบบอเนกรูป
2. เมื่อจัดกลุ่มความสามารถของผู้สอบแตกต่างกัน ข้อสอบลักษณะใดที่ทำหน้าที่ต่างกันแบบอเนกรูปสำหรับแต่ละกลุ่มความสามารถผู้สอบ
3. เมื่อข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป วิธีแมนเทิล-แฮนส์เชล และวิธีถดถอยโลจิสติกวิธีใดที่มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันแบบอเนกรูปได้ดี โดยเกิดอัตราความคลาดเคลื่อนต่ำ และมีความเหมาะสมในทางปฏิบัติมากกว่า

จากปัญหาข้างต้นนี้ ผู้วิจัยจึงสนใจศึกษาเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันแบบอเนกรูปด้วยวิธีแมนเทิล-แฮนส์เชลกับวิธีถดถอยโลจิสติก เพราะทั้งสองวิธีมีความสะดวก ใช้กลุ่มตัวอย่างน้อย การคำนวณและการแปลผลง่าย ตัวแปรที่ศึกษาคือ กลุ่มความสามารถของผู้สอบ ค่าความยากของข้อสอบและค่าอำนาจจำแนกของข้อสอบ ทั้งนี้เพราะเป็นปัจจัยสำคัญที่มีผลต่ออัตราความถูกต้องและอัตราความคลาดเคลื่อนในการตรวจสอบการทำหน้าที่ต่างกันแบบอเนกรูป โดยศึกษาจากข้อมูลจำลองเพราะมีข้อดีคือ ผู้วิจัยทราบก่อนแล้วว่าข้อสอบข้อใดทำหน้าที่ต่างกัน ซึ่งทำให้ตัดสินใจได้ว่าวิธีการใดสามารถตรวจสอบได้ถูกต้อง

วัตถุประสงค์ในการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์หลักคือเพื่อเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกกรุประหว่างวิธีแมนเทิล-แฮนส์เชลกับวิธีถดถอยโลจิสติก โดยมีวัตถุประสงค์เฉพาะดังนี้

1. เพื่อศึกษาข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกกรุป เมื่อมีการจัดกลุ่มความสามารถผู้สอบออกเป็น กลุ่มผู้สอบที่มีความสามารถสูง กลุ่มผู้สอบที่มีความสามารถปานกลาง และกลุ่มผู้สอบที่มีความสามารถต่ำ โดยใช้วิธีทฤษฎีการตอบสนองข้อสอบตรวจสอบ
2. เพื่อเปรียบเทียบอัตราความถูกต้องและอัตราความคลาดเคลื่อนระหว่างวิธีแมนเทิล-แฮนส์เชลกับวิธีถดถอยโลจิสติกในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกกรุป

สมมติฐานการวิจัย

จากการศึกษางานวิจัยของ Mazor และคณะ (1994) ตรวจสอบการทำหน้าที่ต่างกันแบบอเนกกรุปด้วยวิธีแมนเทิล-แฮนส์เชล(MH) โดยการแบ่งกลุ่มผู้สอบออกเป็นสองกลุ่ม คือกลุ่มสูง และกลุ่มต่ำตามคะแนนผลการสอบ แล้ววิเคราะห์แต่ละกลุ่มแยกจากกัน พบว่าวิธีแมนเทิล-แฮนส์เชลสามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันได้เพิ่มขึ้น งานวิจัยของ Narayanan และSwaminatan (1996) พบว่าวิธีแมนเทิล-แฮนส์เชล (MH) มีอัตราการตรวจสอบต่ำกว่าวิธีถดถอยโลจิสติก(LR) ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกกรุป แต่อัตราความคลาดเคลื่อนของวิธีถดถอยโลจิสติก (LR) สูงกว่าวิธีแมนเทิล-แฮนส์เชล (MH) เล็กน้อย และพบว่าข้อสอบที่ทำหน้าที่ต่างกันส่วนใหญ่เป็นข้อสอบที่มีค่าความยากต่ำและค่าอำนาจจำแนกสูง ดังนั้นจึงมีสมมติฐานการวิจัยดังนี้

1. เมื่อจัดกลุ่มความสามารถผู้สอบออกเป็น กลุ่มผู้สอบที่มีความสามารถสูง กลุ่มผู้สอบที่มีความสามารถปานกลาง และกลุ่มผู้สอบที่มีความสามารถต่ำ ข้อสอบที่ตรวจพบการทำหน้าที่ต่างกันแบบอเนกกรุปส่วนใหญ่เป็นข้อสอบที่มีค่าความยากต่ำ และค่าอำนาจจำแนกสูง
2. เมื่อจัดกลุ่มความสามารถผู้สอบ ค่าความยากของข้อสอบและค่าอำนาจจำแนกของข้อสอบต่างกัน วิธีถดถอยโลจิสติกจะมีอัตราความถูกต้องในการตรวจสอบมากกว่าวิธีแมนเทิล-แฮนส์เชล และมีอัตราความคลาดเคลื่อนต่ำกว่าวิธีแมนเทิล-แฮนส์เชลในการตรวจสอบการทำหน้าที่ต่างกันของแบบสอบแบบอเนกกรุป

ขอบเขตการวิจัย

1. ตัวแปรที่ศึกษา

1.1 **ตัวแปรอิสระ** คือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของแบบสอบ การจัดกลุ่มความสามารถผู้สอบ ค่าความยากของข้อสอบและค่าอำนาจจำแนกของข้อสอบ ดังนี้

1.1.1 วิธีการตรวจสอบการทำหน้าที่ต่างกันของแบบสอบมี 2 วิธี คือ วิธีแมนเทล - แชนส์เรล กับวิธีถดถอยโลจิสติก

1.1.2 การจัดกลุ่มความสามารถผู้สอบมี 3 ระดับ คือกลุ่มผู้สอบที่มีความสามารถสูง กลุ่มผู้สอบที่มีความสามารถปานกลาง และกลุ่มผู้สอบที่มีความสามารถต่ำ

1.1.2.1 กลุ่มผู้สอบที่มีความสามารถสูง ($\bar{\theta} = 1.5$)

1.1.2.2 กลุ่มผู้สอบที่มีความสามารถปานกลาง ($\bar{\theta} = 0$)

1.1.2.3 กลุ่มผู้สอบที่มีความสามารถต่ำ ($\bar{\theta} = -1.5$)

1.1.3 ค่าความยากของข้อสอบแบ่งเป็น 3 ระดับคือ กลุ่มข้อสอบที่มีค่าความยากสูง กลุ่มข้อสอบที่มีค่าความยากปานกลาง และกลุ่มข้อสอบที่มีค่าความยากต่ำ

1.1.3.1 กลุ่มข้อสอบที่มีค่าความยากสูง ($\bar{b} = 1.835$)

1.1.3.2 กลุ่มข้อสอบที่มีค่าความยากปานกลาง ($\bar{b} = 0$)

1.1.3.3 กลุ่มข้อสอบที่มีค่าความยากต่ำ ($\bar{b} = -1.835$)

1.1.4 ค่าอำนาจจำแนกของข้อสอบแบ่งเป็น 3 ระดับ คือกลุ่มข้อสอบที่มีค่าอำนาจจำแนกสูง กลุ่มข้อสอบที่มีค่าอำนาจจำแนกปานกลาง และกลุ่มข้อสอบที่มีค่าอำนาจจำแนกต่ำ

1.1.4.1 กลุ่มข้อสอบที่มีค่าอำนาจจำแนกสูง ($\bar{a} = 1.675$)

1.1.4.2 กลุ่มข้อสอบที่มีค่าอำนาจจำแนกปานกลาง ($\bar{a} = 0.7$)

1.1.4.3 กลุ่มข้อสอบที่มีค่า \bar{a} -

1.2. **ตัวแปรตาม** คือ ประสิทธิภาพในการตรวจสอบ โดยพิจารณาจาก

1.2.1 อัตราความถูกต้อง ซึ่งมี 2 ประเภทคืออัตราความถูกต้องประเภทที่ 1 และอัตราความถูกต้องประเภทที่ 2

1.2.2 อัตราความคลาดเคลื่อน ซึ่งมี 2 ประเภท คืออัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราความคลาดเคลื่อนประเภทที่ 2

2. ข้อมูลที่นำมาใช้ในงานวิจัยครั้งนี้เป็นข้อมูลที่จำลองขึ้นด้วยโปรแกรม IRTDATA ของ George A. Johanson (1992)

3. เกณฑ์ที่ใช้ในการวิเคราะห์ว่าข้อสอบทำหน้าที่ต่างกันแบบอนุกรม มีดังนี้

3.1 วิธีแมนเทิล-แฮนส์เซล ได้แก่ค่า α_{MH} หรือค่า Δ_{MH} และการทดสอบนัยสำคัญทางสถิติ χ^2_{MH} ที่ระดับ .05

3.2 วิธีถดถอยโลจิสติก ได้แก่ค่าปฏิสัมพันธ์ระหว่างกลุ่มผู้สอบกับความสามารถผู้สอบ และการทดสอบนัยสำคัญทางสถิติ χ^2 ที่ระดับ .05

3.3 วิธี IRT ได้แก่การทดสอบนัยสำคัญของพื้นที่ระหว่างโค้งคุณลักษณะข้อสอบด้วยสถิติทดสอบ Z ที่ระดับ .05

ข้อตกลงเบื้องต้น

วิธีทฤษฎีการตอบสนองข้อสอบ (IRT) เป็นวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีความถูกต้องสูง และได้รับการยอมรับมากที่สุด (Halland and Thayer, 1988 ; Hambleton et al., 1993) จึงเป็นวิธีมาตรฐานที่ใช้ระบุว่าข้อสอบข้อใดทำหน้าที่ต่างกันที่ถือว่าถูกต้อง เพื่อใช้เป็นเกณฑ์ในการประเมินอัตราความถูกต้องในการตรวจสอบ และอัตราความคลาดเคลื่อนในการตรวจสอบด้วยวิธีแมนเทิล-แฮนส์เซลกับวิธีถดถอยโลจิสติก

คำจำกัดความที่ใช้ในการวิจัย

ข้อสอบทำหน้าที่ต่างกัน หมายถึง ข้อสอบที่ทำให้ผู้สอบที่มีความสามารถเท่ากันในสิ่งที่ต้องการวัด มีโอกาสตอบข้อสอบได้ถูกต้องแตกต่างกัน เนื่องจากผู้สอบอยู่ในกลุ่มย่อยต่างกัน

ข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูป หมายถึง ข้อสอบที่ทำให้ผู้สอบกลุ่มหนึ่งมีโอกาสในการตอบข้อสอบถูกมากกว่าผู้สอบอีกกลุ่มหนึ่งสม่ำเสมอในทุกระดับความสามารถ

ข้อสอบที่ทำหน้าที่ต่างกันแบบอนุกรม หมายถึง ข้อสอบที่ทำให้โอกาสในการตอบข้อสอบถูกของผู้สอบระหว่างกลุ่มไม่สม่ำเสมอในทุกระดับความสามารถ ในการวิจัยนี้จะเน้นความสามารถของผู้สอบและคะแนนรายข้อจำลองขึ้นด้วยโปรแกรม TRTDATA ข้อสอบที่ทำหน้าที่ต่างกันแบบอนุกรมในที่นี้จึงหมายถึงข้อสอบข้อที่จำลองคะแนนขึ้นด้วยโปรแกรม IRTDATA และตรวจพบว่าทำหน้าที่ต่างกันแบบอนุกรมด้วยวิธีตามทฤษฎีการตอบสนองข้อสอบ

กลุ่มความสามารถผู้สอบ (θ) หมายถึง ระดับความสามารถของผู้สอบซึ่งประมาณได้จากโมเดลตามทฤษฎีการตอบสนองข้อสอบ ในที่นี้ได้แบ่งกลุ่มผู้สอบออกเป็นกลุ่มผู้สอบที่มีความสามารถสูง กลุ่มผู้สอบที่มีความสามารถปานกลางและกลุ่มผู้สอบที่มีความสามารถต่ำ

กลุ่มผู้สอบที่มีความสามารถสูง หมายถึงกลุ่มผู้สอบที่มีระดับความสามารถอยู่ระหว่าง 0.5 ถึง 3.0

กลุ่มผู้สอบที่มีความสามารถปานกลาง หมายถึง กลุ่มผู้สอบที่มีระดับความสามารถอยู่ระหว่าง -1.0 ถึง 1.0

กลุ่มผู้สอบที่มีความสามารถต่ำ หมายถึงกลุ่มผู้สอบที่มีระดับความสามารถอยู่ระหว่าง -3.0 ถึง 0

ค่าความยากของข้อสอบ (b) หมายถึง ระดับความยากของข้อสอบซึ่งมีค่าเท่ากับระดับความสามารถของผู้สอบที่มีโอกาสทำข้อสอบข้อนั้นถูก $(1+c) / 2$ ในทางปฏิบัตินิยมใช้ข้อสอบที่มีค่าความยากอยู่ระหว่าง -2.5 ถึง 2.5 ในที่นี้ได้แบ่งค่าความยากของข้อสอบเป็น 3 ระดับคือ กลุ่มข้อสอบที่มีค่าความยากสูง กลุ่มข้อสอบที่มีค่าความยากปานกลาง และกลุ่มข้อสอบที่มีค่าความยากต่ำ

กลุ่มข้อสอบที่มีค่าความยากสูง หมายถึง ข้อสอบที่มีค่าความยากอยู่ระหว่าง 0.67 ถึง 2.5 ($0.67 < b < 2.5$)

กลุ่มข้อสอบที่มีค่าความยากปานกลาง หมายถึง ข้อสอบที่มีค่าความยากตั้งแต่ -0.67 ถึง 0.67 ($-0.67 \leq b \leq 0.67$)

กลุ่มข้อสอบที่มีค่าความยากต่ำ หมายถึง ข้อสอบที่มีค่าความยากอยู่ระหว่าง -2.5 ถึง -0.67 ($-2.5 < b < -0.67$)

ค่าอำนาจจำแนกของข้อสอบ (a) หมายถึง ค่าความชันของโค้งคุณลักษณะข้อสอบที่จุดเปลี่ยนโค้ง ในทางปฏิบัตินิยมใช้ข้อสอบที่มีค่าอำนาจจำแนกอยู่ระหว่าง 0.5 ถึง 2.5 ในที่นี้ได้แบ่งค่าอำนาจจำแนกของข้อสอบเป็น 3 ระดับ คือ กลุ่มข้อสอบที่มีค่าอำนาจจำแนกสูง กลุ่มข้อสอบที่มีค่าอำนาจจำแนกปานกลาง และกลุ่มข้อสอบที่มีค่าอำนาจจำแนกต่ำ

กลุ่มข้อสอบที่มีค่าอำนาจจำแนกสูง หมายถึง ข้อสอบที่มีค่าอำนาจจำแนกอยู่ระหว่าง 0.85 ถึง 2.5 ($0.85 < a < 2.5$)

กลุ่มข้อสอบที่มีค่าอำนาจจำแนกปานกลาง หมายถึง ข้อสอบที่มีค่าอำนาจจำแนกตั้งแต่ 0.55 ถึง 0.85 ($0.55 \leq a \leq 0.85$)

กลุ่มข้อสอบที่มีค่าอำนาจจำแนกต่ำ หมายถึง ข้อสอบที่มีค่าอำนาจจำแนกอยู่ระหว่าง 0.50 ถึง 0.55 ($0.50 < a < 0.55$)

วิธี IRT หมายถึง วิธีทฤษฎีการตอบสนองข้อสอบ (IRT) ซึ่งพิจารณาการทำหน้าที่ต่างกันของข้อสอบจากฟังก์ชันการตอบสนองข้อสอบระหว่างผู้สอบต่างกลุ่มที่ได้จากการตอบข้อสอบข้อเดียวกัน

วิธี MH หมายถึง วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบของแมนเทิล - แชนส์เซล(MH) ซึ่งวิเคราะห์จากความแตกต่างของสัดส่วนการตอบข้อสอบระหว่างผู้สอบที่มีความสามารถระดับเดียวกัน

วิธี LR หมายถึง วิธีถดถอยโลจิสติก(LR) ซึ่งใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยพิจารณาจากกลุ่มผู้สอบ ความสามารถผู้สอบและปฏิสัมพันธ์ระหว่างกลุ่มผู้สอบกับความสามารถผู้สอบในการทำนายโอกาสของการตอบข้อสอบถูกหรือผิด

เกณฑ์การตัดสินข้อสอบทำหน้าที่ต่างกันแบบอนุกรม หมายถึง สิ่งที่ใช้เป็นหลักในการพิจารณาตัดสินข้อสอบทำหน้าที่ต่างกันแบบอนุกรม ในการวิจัยนี้ใช้วิธี IRT วิธี MH และวิธี LR เป็นหลักในการพิจารณาตัดสินข้อสอบทำหน้าที่ต่างกันแบบอนุกรม

เกณฑ์การตัดสินข้อสอบที่ทำหน้าที่ต่างกันแบบอนุกรมตามวิธี IRT หมายถึง ความแตกต่างอย่างมีนัยสำคัญของพื้นที่ระหว่างโค้งคุณลักษณะข้อสอบ เมื่อทำการทดสอบนัยสำคัญด้วยสถิติทดสอบ Z ที่ระดับ .05

เกณฑ์การตัดสินข้อสอบที่ทำหน้าที่ต่างกันแบบอนุกรมตามวิธี MH หมายถึง ความแตกต่างของค่า α_{MH} แตกต่างจาก 1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 หรือ Δ_{MH} แตกต่างจาก 0 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

เกณฑ์การตัดสินข้อสอบที่ทำหน้าที่ต่างกันแบบอนุกรมตามวิธี LR หมายถึง ค่าปฏิสัมพันธ์ระหว่างกลุ่มผู้สอบกับระดับความสามารถผู้สอบที่คำนวณได้มีนัยสำคัญทางสถิติเมื่อทำการทดสอบด้วย χ^2 ที่ระดับ .05

กลุ่มอ้างอิง(reference group) หมายถึง กลุ่มผู้สอบที่คาดว่าจะมีโอกาสในการตอบข้อสอบที่ทำหน้าที่ต่างกันได้ถูกต้องมากกว่าผู้สอบอีกกลุ่มหนึ่ง ทั้งๆที่มีความสามารถเท่ากันในที่นี้กลุ่มอ้างอิงได้จากการสุ่มผู้สอบจากข้อมูลที่จำลองขึ้นด้วยโปรแกรม IRTDATA

กลุ่มเปรียบเทียบ (focal group) หมายถึง กลุ่มผู้สอบที่คาดว่าจะมีโอกาสในการตอบข้อสอบที่ทำหน้าที่ต่างกันได้ถูกต้องต่ำกว่าผู้สอบอีกกลุ่มหนึ่ง ทั้งๆที่มีความสามารถเท่ากัน ในที่นี้กลุ่มเปรียบเทียบได้จากการสุ่มผู้สอบจากข้อมูลที่จำลองขึ้นด้วยโปรแกรม IRTDATA

ประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป หมายถึง ความถูกต้องของการระบุการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปจากการตรวจสอบด้วยวิธี MH และวิธี LR ซึ่งพิจารณาจากอัตราความถูกต้องในการตรวจสอบและอัตราความคลาดเคลื่อนในการตรวจสอบ

อัตราความถูกต้องในการตรวจสอบ หมายถึง สัดส่วนหรือร้อยละของจำนวนข้อสอบที่ตรวจพบว่าถูกต้องตรงกับวิธี IRT ซึ่งมี 2 ประเภทคือ อัตราความถูกต้องประเภทที่ 1 และอัตราความถูกต้องประเภทที่ 2

อัตราความถูกต้องประเภทที่ 1 หมายถึง สัดส่วนหรือร้อยละของจำนวนข้อสอบที่ตรวจสอบพบว่าทำหน้าที่ไม่ต่างกันได้ถูกต้องตรงกับวิธี IRT ในการวิจัยนี้คำนวณจากสัดส่วนของจำนวนข้อสอบที่ทำหน้าที่ไม่ต่างกันได้ถูกต้องต่อจำนวนข้อสอบที่ทำหน้าที่ไม่ต่างกันทั้งหมดในแบบสอบ

อัตราความถูกต้องประเภทที่ 2 หมายถึง สัดส่วนหรือร้อยละของจำนวนข้อสอบที่ตรวจสอบพบว่าทำหน้าที่ต่างกันได้ถูกต้องตรงกับวิธี IRT ในการวิจัยนี้คำนวณจากสัดส่วนของจำนวนข้อสอบที่ทำหน้าที่ต่างกันได้ถูกต้องต่อจำนวนข้อสอบที่ทำหน้าที่ต่างกันทั้งหมดในแบบสอบ

อัตราความคลาดเคลื่อนของการตรวจสอบ หมายถึง สัดส่วนหรือร้อยละของจำนวนข้อสอบที่ระบุผิดพลาดซึ่งมี 2 ประเภท คือ อัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราความคลาดเคลื่อนประเภทที่ 2

อัตราความคลาดเคลื่อนประเภทที่ 1 (Type I error) หมายถึง การระบุผิดพลาดว่าข้อสอบทำหน้าที่ต่างกัน (fault positive) ทั้งที่ความจริงข้อสอบทำหน้าที่ไม่ต่างกัน ในการวิจัยนี้คำนวณได้จากสัดส่วนของจำนวนข้อสอบที่ระบุผิดพลาดว่าทำหน้าที่ต่างกันต่อจำนวนข้อสอบที่ทำหน้าที่ไม่ต่างกันทั้งหมดในแบบสอบ

อัตราความคลาดเคลื่อนประเภทที่ 2 (Type II error) หมายถึง การระบุผิดพลาดว่าข้อสอบทำหน้าที่ไม่ต่างกัน (fault negative) ทั้งที่ความจริงข้อสอบทำหน้าที่ต่างกันในการวิจัยนี้คำนวณได้จากสัดส่วนของจำนวนข้อสอบที่ระบุผิดพลาดว่าทำหน้าที่ไม่ต่างกันต่อจำนวนข้อสอบที่ทำหน้าที่ต่างกันทั้งหมดในแบบสอบ

ประโยชน์ที่คาดว่าจะได้รับ

1. ทำให้ทราบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบ
อเนกประหว่างวิธีแมนเทิล-แฮนส์เชลกับวิธีถดถอยโลจิสติก
2. เพื่อเป็นแนวทางในการเลือกวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบ
อเนกประเภทที่เหมาะสมต่อการปฏิบัติ
3. ทำให้ทราบว่าข้อสอบลักษณะใดทำหน้าที่ต่างกันแบบอเนกประเภท จะได้นำไปใช้เป็นแนว
ทางในการพัฒนาแบบสอบให้มีคุณภาพต่อไป



สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย