บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญ

แบบสอบเป็นเครื่องมือหลักที่ใช้ในการวัดและประเมินผลทางการศึกษา เพื่อตรวจสอบว่าผู้สอบมีความสามารถหรือมีคุณลักษณะแฝงภายในที่ต้องการวัด (latent trait) อยู่ในระดับใด ดังนั้นในการพัฒนาแบบสอบจึงต้องคำนึงถึงคุณภาพของแบบสอบเป็นสำคัญ ทั้งนี้ เพื่อให้ได้ข้อมูลที่ถูกต้องตรงตามความเป็นจริงมากที่สุด คุณสมบัติที่สำคัญที่สุดที่บ่งชี้ถึงคุณภาพของแบบสอบคือ ความตรง ซึ่งความตรงเป็นคุณสมบัติที่แสดงถึงสามารถในการวัดคุณลักษณะแฝงได้ตามที่ต้องการวัดหรือแบบสอบทำหน้าที่ได้ตามวัตถุประสงค์ที่กำหนดไว้ ในการตรวจสอบว่าข้อสอบมีความตรงหรือไม่นั้นมีหลายวิธี และวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบก็เป็นอีกวิธีหนึ่งที่ใช้เพื่อตรวจสอบความตรงของข้อสอบ

เมื่อนำแบบสอบไปทดสอบกับผู้สอบกลุ่มย่อยตั้งแต่ 2 กลุ่มขึ้นไป แล้วนำผลการสอบมาวิเคราะห์พบว่าผู้สอบที่มีความสามารถหรือมีคุณลักษณะแฝงที่ต้องการวัดเท่ากัน แต่อยู่ต่างกลุ่มกันแล้ว ผู้สอบมีโอกาสในการตอบข้อสอบได้ถูกต้องไม่เท่ากัน ก็กล่าวได้ว่าแบบสอบหรือข้อสอบทำหน้าที่ต่างกัน (Differential item/test functioning)(Green,B.F.,1994;Mazor,et al.,1995) เมื่อเป็นเช่นนี้แสดงว่าแบบสอบหรือข้อสอบนั้นขาดความตรงเพราะไม่ได้วัดเฉพาะคุณลักษณะแฝงเป้าหมายตามที่ต้องการวัดเท่านั้น แต่ยังวัดคุณลักษณะแฝงแทรกซ้อนที่ไม่ต้องการวัดของผู้สอบอีกด้วย หากผู้สอบกลุ่มย่อยกลุ่มใดมีคุณลักษณะแฝงแทรกซ้อนนั้นสูงกว่าย่อมมีโอกาสที่จะตอบข้อสอบได้ถูกต้องมากกว่า ทั้ง ๆ ที่มีคุณลักษณะแฝงเป้าหมายเท่ากันกับผู้สอบกลุ่มย่อยกลุ่มอื่น ซึ่งจะทำให้เกิดการได้เปรียบเสียเปรียบกันระหว่างกลุ่มผู้สอบย่อย ลักษณะเช่นนี้เดิมเรียกว่า ข้อสอบลำเอียง (Item bias) แต่ในระยะหลังเรียกว่า ข้อสอบทำหน้าที่ต่างกัน (Differential Item Functioning ; DIF) เพราะวิธีการใหม่ ๆ ที่ใช้ในการตรวจสอบความลำเอียงนั้นจะเน้นไปที่ความแตกต่างระหว่างกลุ่มผู้สอบที่ตอบสนองต่างกันต่อข้อสอบข้อเดียวกัน ความแตกต่างที่เกิดขึ้นนี้อาจมาจากข้อคำถาม ประสบการณ์หรือพื้นฐานเดิมที่แตกต่างกันของกลุ่มผู้สอบซึ่งในบางสถานการณ์ก็ไม่เหมาะสมที่จะใช้คำว่าลำเอียง ด้วยเหตุนี้จึงควรใช้คำว่าข้อสอบทำหน้าที่ต่างกันเพราะเป็นคำที่มีความเป็นกลางมากกว่าและเหมาะสมกว่า (Holland and Thayer,1988;Green,B.F.,1994)

Narayanan (1993) Shealy และStout (1993) เสนอแนวคิดว่า ความลำเอียงของข้อสอบจะพิจารณาที่ความตรงของข้อสอบ ถ้าข้อสอบมีความตรงสำหรับกลุ่มหนึ่งน้อยกว่าอีกกลุ่มหนึ่งจะส่งผลให้คะแนนสอบแตกต่างกันระหว่างกลุ่มทั้งที่มีความสามารถที่ต้องการวัดเท่ากัน ส่วนการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ จะพิจารณาที่ข้อสอบว่ามีข้อสอบบางข้อที่มีส่วนทำให้คะแนนสอบมีความแตกต่างกันระหว่างผู้สอบต่างกลุ่มที่จับคู่เปรียบเทียบกัน นั่นคือข้อสอบจะเข้าข้าง(favor)ผู้สอบกลุ่มหนึ่งมากกว่าอีกกลุ่มหนึ่ง โดยการตัดสินว่าข้อสอบทำหน้าที่ต่างกันหรือไม่จะสัมพันธ์กับเกณฑ์(criterion)ที่ใช้ในการจับคู่เปรียบเทียบผู้สอบระหว่างกลุ่ม ดังนั้นการวิเคราะห์ความลำเอียงของข้อสอบ จึงเป็นกรณีย่อยกรณีหนึ่งของการวิเคราะห์ DIF เพราะเกณฑ์ในการจับคู่เปรียบเทียบที่ใช้ตัดสินว่าข้อสอบมีความลำเอียงคือความตรง ทั้งนี้เพราะเกณฑ์ที่ใช้ในการจับคู่เปรียบเทียบผู้สอบระหว่างกลุ่ม ซึ่งจะใช้ในการตัดสินว่าข้อสอบทำหน้าที่ต่างกันหรือไม่มีหลายเกณฑ์ เช่น เพศ เชื้อชาติ ภูมิลำเนา เป็นต้น

โดยสรุปการทำหน้าที่ต่างกันของข้อสอบ จะเกิดขึ้นเมื่อข้อสอบวัดคุณลักษณะแฝงอื่นนอกเหนือจากคุณลักษณะแฝงที่ต้องการวัด ทำให้ผู้สอบแต่ละกลุ่มมีโอกาสในการตอบข้อสอบถูกแตกต่างกันทั้ง ๆ ที่มีความสามารถที่ต้องการวัดเท่ากัน

การตรวจสอบเกี่ยวกับการได้เปรียบเสียเปรียบในการสอบระหว่างผู้สอบกลุ่มย่อยตั้งแต่ 2 กลุ่มขึ้นไป เริ่มมีตั้งแต่ปี ค.ศ.1951 โดยเป็นการศึกษาเปรียบเทียบระหว่างผู้สอบที่มีความต่างกันทางด้านเศรษฐกิจ สังคม เพศ วัฒนธรรมและเชื้อชาติ ระดับสติปัญญา วิธีสอน (กาญจนา วัธนสุนทร,2537) รวมทั้งสถาบันการศึกษา ประสบการณ์ (Holland and Thayer,1988) นอกจากจะศึกษาปัจจัยอันเกิดจากผู้สอบซึ่งส่งผลให้เกิดการได้เปรียบเสียเปรียบระหว่างกลุ่มผู้สอบ ในระยะหลังได้มีการศึกษาเปรียบเทียบวิธีการต่าง ๆ ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ทั้งนี้ เพราะมีวิธีการตรวจสอบหลายวิธีที่ถูกคิดค้นและพัฒนา ปรับปรุงเพื่อให้สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้อย่างมีประสิทธิภาพมากที่สุด

วิธีการในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมีหลายวิธี วิธีที่เป็นที่รู้จักและนิยมใช้ในการเปรียบเทียบที่ผ่านมา กาญจนา วัธนสุนทร (2537)ได้สรุปว่าวิธีที่นิยมใช้มี 6 วิธีได้แก่ วิธีค่าความยากแปลง (Transformed item difficulties) วิธีวิเคราะห์ความแปรปรวน (Analysis of variance)

วิธีไคสแควร์ (Chi square) วิธีทฤษฎีการตอบสนองข้อสอบ (Item Response Theory) วิธีMantel-Haenszel(MH)และวิธี SIBTEST ซึ่งแต่ละวิธีมีขั้นตอนการวิเคราะห์ ข้อดี ข้อเสียต่างกันดังนี้

ตารางที่ 1 เปรียบเทียบวิธีต่าง ๆ ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

ประเด็น	TID	ANOVA	χ^2	IRT	MH	SIBTEST
1.ข้อตกลงเบื้องต้น	ผลร่วมระหว่างกลุ่มกับข้อกระทงเป็นตัวบ่งชี้ความลำเอียง	ความแปรปรวนและความแปรปรวนร่วม ของข้อสอบต้องเท่ากัน	คะแนนรวมจากแบบสอบเป็นตัวแทนความสามารถของผู้สอบ	แบบสอบเป็นเอกมิติและโค้งลักษณะข้อสอบสามารถแสดงฟังก์ชั่นของค่าความสามารถและโอกาสในการตอบข้อสอบถูก	คะแนนรวมจากแบบสอบเป็นตัวแทนความสามารถของผู้สอบ	คะแนนรวมจากแบบสอบเป็นตัวแทนความสามารถของผู้สอบและมีมิติการวัด 2 มิติ คือ คุณลักษณะแฝงเป้าหมายและคุณลักษณะแฝงแทรกซ้อน
2.สิ่งที่ทำการวิเคราะห์	ผลร่วมระหว่างการเป็นสมาชิกในกลุ่มกับการตอบถูก	ผลร่วมระหว่างการเป็นสมาชิกในกลุ่มกับการตอบถูก	ความแตกต่างของอัตราส่วนการตอบถูกตามระดับคะแนนรวม	ความแตกต่างของฟังก์ชั่นการตอบข้อสอบที่ระดับความสามารถเดียวกัน	ความแตกต่างของอัตราส่วนการตอบระหว่างผู้ที่มีความสามารถระดับเดียวกัน	ความแตกต่างระหว่างคะแนนเฉลี่ยและอัตราส่วนการตอบข้อสอบระหว่างผู้ที่มีความสามารถระดับเดียวกัน

ประเด็น	TID	ANOVA	χ^2	IRT	MH	SIBTEST
3.สิ่งที่พิจารณาในการตัดสิน DIF	ระยะห่างของจุดเดลต้าจากเส้นแกนหลัก	ความมีนัยสำคัญทางสถิติของ F-test	ความมีนัยสำคัญทางสถิติของ χ^2	พื้นที่ระหว่างโค้งลักษณะข้อสอบ	ค่าของดัชนี α_{MH} และความมีนัยสำคัญทางสถิติ	ค่าของดัชนี β_{SIB} และความมีนัยสำคัญทางสถิติ
4.ทฤษฎีพื้นฐาน	CTT	-	-	IRT	-	IRT ชนิดพหุมิติ
5.ข้อดี	คำนวณง่าย ใช้กลุ่มตัวอย่างน้อย	ใช้กลุ่มตัวอย่างน้อย	คำนวณง่าย มีเกณฑ์ตายตัวในการแปลผล	ให้รายละเอียดมาก และการไม่แปรเปลี่ยนของค่าพารามิเตอร์	คำนวณง่าย ใช้กลุ่มตัวอย่างน้อยประหยัด	คำนวณง่าย ใช้กลุ่มตัวอย่างน้อย ตรวจสอบ DIF ได้หลายข้อในคราวเดียวกัน
6.ข้อเสีย	มีความคลาดเคลื่อนเมื่อค่า a สูง และค่า b เปลี่ยนตามกลุ่มผู้สอบ	การคำนวณค่อนข้างยุ่งยากและไม่มีดัชนีบอกระดับความลำเอียง	ไม่มีเกณฑ์ตายตัวในการกำหนดช่วงคะแนนและค่า b เปลี่ยนตามกลุ่มผู้สอบ	มีการคำนวณซับซ้อนหลายรอบ แปลผลยาก ใช้กลุ่มตัวอย่างมาก ค่าใช้จ่ายสูง	ไม่มีความไวในการตรวจสอบ DIF แบบอเนกรูป (nonuniform DIF)	อัตราความคลาดเคลื่อนชนิดที่ 1 เพิ่มสูงเมื่อคะแนนเฉลี่ยแตกต่างกันมาก

จากตารางที่ 1 วิธีค่าความยากแปลง (TID) วิธีวิเคราะห์ความแปรปรวน (ANOVA) และวิธีไคสแควร์ (χ^2) เป็นการวิเคราะห์ข้อสอบโดยอาศัยทฤษฎีการวัดแบบดั้งเดิม ซึ่งมีจุดด้อยคือ ค่าพารามิเตอร์ของข้อสอบจะแปรเปลี่ยนไปตามกลุ่มผู้สอบ นอกจากนี้วิธีวิเคราะห์ความแปรปรวนและวิธีไคสแควร์ไม่มีดัชนีบอกระดับของการทำหน้าที่ต่างกันของข้อสอบ แต่เป็นวิธีที่ประหยัดและใช้กลุ่มตัวอย่างน้อย วิธีทฤษฎีการตอบสนองข้อสอบ (IRT) เป็นวิธีที่วิเคราะห์ความแตกต่างของฟังก์ชั่นการตอบข้อสอบระหว่างกลุ่มผู้สอบ มีดัชนีบอกระดับของการทำหน้าที่ต่างกันของข้อสอบและทดสอบนัยสำคัญทางสถิติ ซึ่งเป็นวิธีที่ให้รายละเอียดมากและมีข้อดีคือการไม่แปรเปลี่ยนของค่าพารามิเตอร์ แต่มีข้อเสียคือ ค่อนข้างสิ้นเปลือง วิธีแมนเทล-แฮนส์เซล(MH) คล้าย

กับวิธีไคสแควร์ คือ ใช้คะแนนรวมจากแบบสอบเป็นตัวแทนของความสามารถ แต่วิธีแมนเทล-แฮนส์เซลจะวิเคราะห์ทีละระดับความสามารถ และมีดัชนีบอกระดับการทำหน้าที่ต่างกันของข้อสอบรวมทั้งการทดสอบนัยสำคัญทางสถิติ ส่วนวิธีชิบเทสท์(SIBTEST) ใช้คะแนนรวมจากแบบสอบเป็นตัวแทนความสามารถเช่นกัน โดยมีข้อตกลงว่ามีมิติการวัด 2 มิติ ดังนั้นคะแนนจากแบบสอบจึงมี 2 ส่วนคือ คะแนนจากแบบสอบที่มีความตรง(valid subtest) ซึ่งวัดคุณลักษณะแฝงเป้าหมาย และคะแนนจากแบบสอบที่ศึกษา(studied subtest) ซึ่งวัดคุณลักษณะแฝงแทรกซ้อน มีดัชนีบอกระดับการทำหน้าที่ต่างกันของข้อสอบและทดสอบนัยสำคัญทางสถิติ ทั้งวิธีแมนเทล-แฮนส์เซลและวิธีชิบเทสท์เป็นวิธีที่ประหยัด ใช้กลุ่มตัวอย่างน้อย

ที่ผ่านมามีการศึกษาเปรียบเทียบวิธีการในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการต่าง ๆ เช่น ทัศนีย์ พีรมนตรี (2530) ได้ทำการวิเคราะห์ความลำเอียงของแบบสอบวิชาคณิตศาสตร์ด้วยวิธีวิเคราะห์ 3 วิธี คือ วิธีกำหนดจุดค่าเดลต้า วิธีทดสอบความแตกต่างระหว่างกลุ่มด้วยสถิติไคสแควร์ในโมเดลลอกลีเนียและวิธี IRT 3 พารามิเตอร์ พบว่าวิธี IRT 3 พารามิเตอร์พบจำนวนข้อสอบที่มีความลำเอียงมากที่สุด ซึ่งสอดคล้องกับสุรศักดิ์ อมรรัตนศักดิ์ (2531)ที่ได้ศึกษาเปรียบเทียบผลของวิธีวิเคราะห์ความลำเอียงของข้อสอบ 4 วิธี ได้แก่ วิธีวิเคราะห์ความแปรปรวน วิธีค่าความยากแปลง วิธี IRT 1 พารามิเตอร์ และวิธี IRT 3 พารามิเตอร์ พบว่าวิธี IRT 3 พารามิเตอร์ พบจำนวนข้อสอบที่ลำเอียงมากที่สุด รองลงมาคือวิธีวิเคราะห์ความแปรปรวน และวิธีค่าความยากแปลงพบจำนวนข้อสอบที่ลำเอียงน้อยที่สุด

เพ็ญพนา สุขสม (2539) ทำการเปรียบเทียบผลของวิธีวิเคราะห์ความลำเอียงของข้อสอบที่แตกต่างกัน 3 วิธี ได้แก่ วิธีค่าความยากแปลง วิธี IRT 3 พารามิเตอร์และวิธีแมนเทล-แฮนส์เซล พบว่าวิธีแมนเทล-แฮนส์เซลตรวจสอบพบข้อสอบลำเอียงมีจำนวนข้อมากที่สุด รองลงมาคือวิธี IRT 3 พารามิเตอร์และวิธีค่าความยากแปลงพบข้อสอบที่ลำเอียงมีจำนวนน้อยที่สุด ทั้งสามวิธีมีความสัมพันธ์กันทางบวก โดยวิธี IRT 3 พารามิเตอร์มีความสัมพันธ์กับวิธี MH สูงกว่าวิธีค่าความยากแปลง

Subkoviak และคณะ (1984)ได้ศึกษาเปรียบเทียบวิธีการตรวจสอบความลำเอียง พบว่าวิธีโค้งคุณลักษณะข้อสอบ 3 พารามิเตอร์ (3 parameter item charactevistic curve procedure) มีประสิทธิภาพมากที่สุด รองลงมาคือวิธีไคสแควร์และวิธีค่าความยากแปลง Shepard และคณะ (1985)

ศึกษาโดยใช้ข้อมูลจริงและข้อมูลจำลองพบว่าวิธีการตอบสนองข้อสอบเทียม (The pseudo-IRT approach)เป็นวิธีที่ดีที่สุด วิธีไคสแควร์เป็นวิธีที่สามารถตรวจสอบได้ถูกต้องใกล้เคียงกับวิธีการตอบสนองข้อสอบเทียม และวิธีกำหนดจุดค่าเดลต้าของแองกอฟฟ์ (The Angoff delta-plot method) มีประสิทธิภาพใกล้เคียงกับวิธีไคสแควร์

Rogers และ Swaminathan (1993) พบว่าวิธี MH มีประสิทธิภาพเท่าเทียมกับวิธีถดถอยโลจิสติในการตรวจสอบการทำหน้าที่ต่างกันแบบเอกรูป (Uniform DIF) และวิธี MH มีประสิทธิภาพน้อยกว่าในการตรวจสอบการทำหน้าที่ต่างกันแบบอเนกรูป (nonuniform DIF) ซึ่งสอดคล้องกับผลจากการศึกษาของ Mazor และคณะ (1994) ที่พบว่าวิธี MH สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปได้ดี แต่ไม่มีความไว (sensitive) ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป เกษร หว่างจิตร (2539) ใช้วิธี MH ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบพบว่าข้อสอบที่ทำหน้าที่ต่างกันส่วนใหญ่เป็นข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูป Shealy และ Stout (1993)กล่าวว่าในกรณีข้อสอบลำเอียงข้อเดียว (singlebiased item) ทั้งวิธี MH และวิธี SIBTEST มีประสิทธิภาพดีสำหรับใช้ในการตรวจสอบ และในกรณีข้อสอบลำเอียงหลายข้อ (several biased item) วิธี SIBTEST ก็สามารถใช้ตรวจสอบได้อย่างมีประสิทธิภาพ

Narayanan และ Swaminathan (1994) ได้ผลจากการศึกษาว่าวิธี MH และวิธีSIBTEST มีประสิทธิภาพเท่าเทียมกันในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปเมื่อการกระจายของความสามารถของผู้สอบเท่ากันระหว่าง 2 กลุ่ม ในกรณีที่การกระจายของความสามารถของผู้สอบไม่เท่า กันวิธี SIBTEST จะมีประสิทธิภาพมากกว่าวิธี MH ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

การศึกษาเปรียบเทียบวิธีการต่าง ๆ ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้ข้อมูลจำลอง พบว่าวิธี IRT สามารถตรวจสอบพบข้อสอบที่ทำหน้าที่ต่างกันได้ถูกต้องมากที่สุด ทั้งนี้ Subkoviak และคณะ (1984 cited in Shepard et al.,1985) กล่าวว่าข้อมูลจำลองสร้างขึ้นภายใต้ทฤษฎี IRT ดังนั้น จึงเอื้อต่อวิธี IRT ทำให้วิธีนี้ตรวจสอบการทำหน้าที่ต่างกันได้ดีที่สุด วิธี MH

มีความสอดคล้องสูงกับวิธี IRT โมเดล 2 และ 3 พารามิเตอร์ อีกทั้งสามารถใช้วิธี MH แทนวิธี IRT ได้อย่างประหยัดกว่า(Hambleton et al., 1986 ; Baghi and Ferrara, 1989 อ้างถึงในกาญจนา วัธนสุนทร, 2537) Miller และ Oshima (1992) กล่าวว่าวิธี MH จะระบุข้อสอบที่มีความลำเอียงปานกลางได้ดีพอ ๆ กับดัชนีของวิธี IRT

จากตารางและผลการศึกษาของนักวิจัยที่เปรียบเทียบวิธีการต่าง ๆ ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อาจกล่าวได้ว่าวิธี IRT เป็นวิธีที่ดีที่สุด แต่มีข้อจำกัดคือต้องใช้กลุ่มตัวอย่างขนาดใหญ่ ข้อมูลต้องเป็นไปตามข้อตกลงเบื้องต้น การคำนวณซับซ้อนและต้องคำนวณหลายรอบ ซึ่งเป็นการสิ้นเปลืองทั้งเวลาและค่าใช้จ่าย (Osterlind, 1993 ; Ryan, 1991) วิธีไคสแควร์จึงเป็นวิธีที่เป็นทางเลือกที่ดีแทนวิธี IRT (Shepard et al.,1985) ส่วนวิธี MH เป็นวิธีที่พัฒนาจากวิธีไคสแควร์แบบดั้งเดิม จึงเป็นอีกวิธีหนึ่งที่เป็นทางเลือกแทนวิธี IRT ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ(Holland, 1985 ; Holland and Thayer, 1986 ; cited in Ryan, 1991) วิธี SIBTEST เป็นวิธีที่มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ใกล้เคียงกับวิธี MH (Shealy and Stout,1993;Narayanan and Swaminathan,1994) ส่วนวิธีอื่นๆ คือ วิธีวิเคราะห์ความแปรปรวน เป็นวิธีที่คำนวณและตัดสินการทำหน้าที่ต่างกันโดยการทดสอบความมีนัยสำคัญทางสถิติ ไม่มีดัชนีบอกระดับความลำเอียง และวิธีค่าความยากแปลงเป็นวิธีที่คำนวณจากค่าความยากของข้อสอบ ซึ่งค่าความยากจะเปลี่ยนไปตามกลุ่มผู้สอบและไม่มีการทดสอบนัยสำคัญทางสถิติ

ผลจากการเปรียบเทียบวิธีการในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ พบว่าปัจจัยสำคัญที่ส่งผลต่อประสิทธิภาพการตรวจสอบ คือ ขนาดของกลุ่มตัวอย่าง เช่น จากการศึกษาของ Swaminathan และ Rogers (1990) ใช้กลุ่มตัวอย่างในแต่ละกลุ่มมีขนาด 250 และ500 คนตรวจสอบด้วยวิธีMHและวิธีถดถอยโลจิสต์ อีกทั้ง Mazor และคณะ(1991) ได้ศึกษาโดยใช้กลุ่มตัวอย่างขนาด 100, 200, 500, 1000 และ 2000 คนใช้วิธี MH ในการตรวจสอบ โดยศึกษาจากข้อมูลจำลองต่างก็ได้ข้อค้นพบที่สอดคล้องกัน คือเมื่อขนาดกลุ่มตัวอย่างใหญ่ขึ้นอัตราการตรวจสอบจะเพิ่มขึ้น ส่วน Narayanan และ Swaminathan(1993) ใช้ขนาดของกลุ่มตัวอย่างโดยมีอัตราส่วนของกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบในการวิเคราะห์แต่ละครั้ง ดังนี้ 300:100, 300:200, 300:300, 500:100, 500:200, 500:300 , 1000:100, 1000:200 และ 1000:300 ใช้วิธี MH กับวิธี SIBTEST อีกทั้ง Ackerman และ Evans (1994) ศึกษาโดยใช้กลุ่มตัวอย่างเป็นอัตราส่วนต่อกัน

เช่นเดียวกัน ซึ่งขนาดของกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบเป็น 1000:250, 1000:500, 500:250 โดยศึกษาจากข้อมูลจำลอง ข้อค้นพบที่ได้สอดคล้องกับเมื่อใช้ขนาดของกลุ่มตัวอย่างย่อยเป็นอัตราส่วนเท่ากัน คือ เมื่อกลุ่มตัวอย่างย่อยของทั้งสองกลุ่มมากขึ้น อัตราการตรวจสอบจะเพิ่มขึ้น

ปัจจัยที่ส่งผลต่อประสิทธิภาพการตรวจสอบอีกอย่างหนึ่ง คือ ความยาวของแบบสอบ ทั้งนี้เพราะความยาวของแบบสอบจะมีผลกระทบต่อความถูกต้องในการจับคู่เปรียบเทียบระหว่างผู้สอบ เนื่องจากทั้งวิธี MH และวิธี SIBTEST ใช้คะแนนรวมจากการสอบแทนคุณลักษณะภายในของผู้สอบที่วัดได้ แบบสอบที่มีความยาวมากกว่าย่อมส่งผลให้มีความน่าเชื่อถือ(reliable)มากกว่า ทำให้การจับคู่เปรียบเทียบระหว่างผู้สอบมีความถูกต้องมากขึ้น เช่น จากการศึกษาของ Swaminathan และ Rogers (1990) โดยใช้ข้อมูลจำลองและตรวจสอบด้วยวิธี ถดถอยโลจิสต์(Logistic regression) มีขนาดของแบบสอบเท่ากับ 40, 60 และ 80 ข้อ พบว่าเมื่อใช้แบบสอบยาวอัตราการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจะดีกว่าการใช้แบบสอบสั้น ต่อมาในปี 1993 ทั้งสองได้ศึกษาอีกครั้งโดยใช้ขนาดของแบบสอบเท่ากับ 40 และ 80 ข้อ ตรวจสอบด้วยวิธี MH กับ วิธีถดถอยโลจิสต์ พบว่าความยาวของแบบสอบไม่มีผลกระทบต่ออัตราการตรวจสอบ และกาญจนา วัธนสุนทร (2537) พบว่า อัตราการตรวจสอบไม่คงที่ข้ามขนาดของแบบสอบ

จากการศึกษางานวิจัยที่ผ่านมา มีประเด็นปัญหาคือ

1. วิธีการในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี MH และวิธีSIBTEST ยังไม่มีความชัดเจนว่าวิธีใดที่จะมีประสิทธิภาพดีกว่า เชื่อถือได้ และเหมาะสมในทางปฏิบัติมากกว่า

2. ขนาดของกลุ่มตัวอย่างส่งผลกระทบต่ออัตราการตรวจสอบ ดังนั้นควรจะใช้ขนาดกลุ่มตัวอย่างเท่าใด จึงจะเหมาะสมเพื่อให้การตรวจสอบมีประสิทธิภาพดีที่สุด

3. ที่ผ่านมา นักวิจัยศึกษาเปรียบเทียบโดยใช้ขนาดผู้สอบของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบเป็นอัตราส่วนต่อกัน ดังนั้นในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ ควรใช้ขนาดกลุ่มตัวอย่างเป็นอัตราส่วนต่อกันเท่าใด จึงจะเหมาะสมและดีกว่าในทางปฏิบัติ

4. ความยาวของแบบสอบจะส่งผลต่อความเที่ยงของคะแนนจากแบบสอบและความถูกต้องในการจับคู่เปรียบเทียบ ซึ่งอาจจะส่งผลต่อความคลาดเคลื่อนชนิดที่ 1 (Narayanan and Swaminathan, 1993) ดังนั้น ความยาวของแบบสอบจะส่งต่อประสิทธิภาพการตรวจสอบหรือไม่

จากปัญหาที่กล่าวข้างต้นนี้ ผู้วิจัยจึงสนใจที่จะศึกษาเปรียบเทียบการตรวจสอบ DIF ด้วยวิธี MH กับวิธี SIBTESTเพราะเป็นวิธีที่มีแนวคิดเดียวกัน คือ ใช้คะแนนรวมเป็นตัวแทนของความสามารถ ใช้กลุ่มตัวอย่างน้อย ประหยัดค่าใช้จ่าย การคำนวณและการแปลผลง่าย ตัวแปรที่ศึกษาคือ ขนาดของกลุ่มตัวอย่างโดยที่ขนาดของกลุ่มตัวอย่างย่อยระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบเป็นอัตราส่วนต่อกัน และความยาวของแบบสอบ ทั้งนี้เพราะนับว่าเป็นปัจจัยสำคัญที่มีผลกระทบต่ออัตราการตรวจสอบและอัตราความคลาดเคลื่อนในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยศึกษาจากข้อมูลจำลอง ทั้งนี้เพราะมีข้อดีคือผู้วิจัยจะทราบก่อนแล้วว่าข้อสอบข้อใดทำหน้าที่ต่างกัน ซึ่งจะทำให้ตัดสินได้ว่าวิธีการตรวจสอบที่ใช้นั้นระบุได้ถูกต้องหรือไม่(Shepard et al.,1985) จึงเหมาะสมที่จะใช้ศึกษาเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีต่าง ๆ

1.2 วัตถุประสงค์ในการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี MH และวิธี SIBTEST เมื่อ

1).ขนาดกลุ่มตัวอย่างย่อยระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบเป็นอัตราส่วนต่อกัน ภายใต้ขนาดกลุ่มตัวอย่างต่างกันที่ระดับความยาวแบบสอบเดียวกัน

2).ขนาดกลุ่มตัวอย่างย่อยระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบเป็นอัตราส่วนต่อกัน ภายใต้ขนาดกลุ่มตัวอย่างเดียวกันแต่มีระดับความยาวแบบสอบต่างกัน

1.3 สมมุติฐานทางการวิจัย

จากการศึกษาของ Narayanan และ Swaminathan (1994) พบว่าเมื่อการกระจายของความสามารถไม่เท่ากันระหว่างผู้สอบกลุ่มย่อย วิธี SIBTEST จะมีประสิทธิภาพมากกว่าวิธี MH ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อีกทั้ง Roussos และ Stout(1996) ศึกษาพบว่าวิธี MH จะมีอัตราความคลาดเคลื่อนชนิดที่ 1 (หมายถึงการระบุผิดพลาดว่าข้อสอบทำหน้าที่ต่างกัน ทั้งที่ความเป็นจริงข้อสอบทำหน้าที่ไม่ต่างกัน)มากกว่าวิธี SIBTEST เมื่อการกระจายของความสามารถเป้าหมายที่ต้องการวัด มีความแตกต่างกันระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ดังนั้นจึงมีสมมุติฐานดังนี้

1.เมื่อขนาดกลุ่มตัวอย่างย่อยระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบเป็นอัตราส่วนต่อกันเป็น 1:1, 1:0.9, 1:0.75, และ 1:0.5 ภายใต้ขนาดกลุ่มตัวอย่างต่างกันที่มีระดับความยาวแบบสอบเดียวกัน วิธี SIBTEST จะมีประสิทธิภาพมากกว่าวิธีMH ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

2.เมื่อขนาดกลุ่มตัวอย่างย่อยระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบเป็นอัตราส่วนต่อกันเป็น 1:1, 1:0.9, 1:0.75, และ 1:0.5 ภายใต้ขนาดกลุ่มตัวอย่างเดียวกันแต่มีระดับความยาวแบบสอบต่างกัน วิธี SIBTEST จะมีประสิทธิภาพมากกว่าวิธี MHในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

1.4 ขอบเขตของการวิจัย

1.การวิจัยครั้งนี้ใช้วิธีวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ 2 วิธีคือวิธี MH และวิธี SIBTEST

2.ตัวแปรที่ศึกษาในครั้งนี้

ตัวแปรอิสระ คือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ขนาดกลุ่มตัวอย่าง และความยาวของแบบสอบซึ่งมีรายละเอียดดังนี้

2.1 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ 2 วิธีคือ วิธี MH และวิธี SIBTEST

2.2 ขนาดกลุ่มตัวอย่างย่อยที่ศึกษามี 3 ขนาดคือ 200, 600, 1000 คนโดยมีอัตราส่วนของกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ 4 ระดับ คือ 1:1, 1:0.9, 1:0.75, 1:0.5 ดังนั้นจึงมีกลุ่มตัวอย่าง 12 ขนาดดังนี้

200:200	200:180	200:150	200:100
600:600	600:540	600:450	600:300
1000:1000	1000:900	1000:750	1000:500

2.3 ความยาวแบบสอบศึกษา 3 ขนาดคือ 30 60และ 90 ข้อ

ตัวแปรตาม คือ ประสิทธิภาพในการตรวจสอบ

3.ข้อมูลที่นำมาใช้ในการวิเคราะห์เป็นข้อมูลที่จำลองขึ้นด้วยโปรแกรม IRTDATA ของ George A. Johanson

4. เกณฑ์ที่ใช้ในการวิเคราะห์ว่าข้อสอบทำหน้าที่ต่างกันมีดังนี้

4.1 วิธี MH ได้แก่ค่า $\alpha_{MH} \neq 1.0$ และการทดสอบนัยสำคัญด้วย χ^2_{MH} ที่ระดับ .05

4.2 วิธี SIBTEST ได้แก่ค่า $\beta_U > 0$ และการทดสอบนัยสำคัญด้วยสถิติ Z ที่ระดับ .05

4.3 วิธี IRT ได้แก่ การทดสอบนัยสำคัญของพื้นที่ระหว่างโค้งคุณลักษณะข้อสอบด้วยสถิติ Z ที่ระดับ .05 (Raju et al.,1993,Budgell et al.,1995)

1.5 ข้อตกลงเบื้องต้น

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบครั้งนี้เป็นการศึกษาจากข้อมูลจำลอง ดังนั้นจึง ไม่อาจระบุแหล่งหรือสาเหตุของการทำหน้าที่ต่างกันของข้อสอบ

1.6 คำจำกัดความที่ใช้ในการวิจัย

ข้อสอบทำหน้าที่ต่างกัน หมายถึงข้อสอบที่ทำให้ผู้สอบที่มีความสามารถที่ต้องการวัดเท่ากัน มีโอกาสในการตอบข้อสอบได้ถูกต้องต่างกัน ทั้งนี้เนื่องจากผู้สอบอยู่ในกลุ่มย่อยต่างกัน ในที่นี้คะแนนของผู้สอบและข้อสอบจำลองขึ้นด้วยโปรแกรม IRTDATA ดังนั้นข้อสอบที่ทำหน้าที่ต่างกันหมายถึง ข้อสอบที่ทำให้ผู้สอบที่มีความสามารถเท่ากันแต่อยู่ต่างกลุ่มกันมีโอกาสตอบข้อสอบถูกไม่เท่ากัน ที่ตรวจสอบพบด้วยวิธี IRT

ความสามารถของผู้สอบ หมายถึงคะแนนรวมของผู้สอบแต่ละคนที่ได้จากการจำลองข้อมูลด้วยโปรแกรม IRTDATA

วิธี MH หมายถึง วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบของแมนเทล-แฮนส์เซล ซึ่งพิจารณาจากความแตกต่างของอัตราส่วนการตอบข้อสอบระหว่างผู้สอบที่มีความสามารถระดับเดียวกัน (Holland and Thayer, 1988)

วิธี SIBTEST หมายถึงวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่พิจารณาความแตกต่างของคะแนนจริงระหว่างผู้สอบที่มีความสามารถระดับเดียวกัน(Shealy and Stout,1993)

วิธี IRT หมายถึงวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ที่พิจารณาจากความแตกต่างของฟังก์ชั่นการตอบข้อสอบ ระหว่างผู้สอบต่างกลุ่มที่ได้จากการตอบข้อสอบชุดเดียวกัน (Lord,1980 cited in Green,1944)

เกณฑ์การตัดสินข้อสอบที่ทำหน้าที่ต่างกันตามวิธี MH หมายถึงข้อสอบที่มีค่า α_{MH}แตกต่าง

จาก 1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 หรือมีค่า Δ_{MH} แตกต่างจาก 0 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

เกณฑ์การตัดสินข้อสอบที่ทำหน้าที่ต่างกันตามวิธี SIBTEST หมายถึงข้อสอบที่มีค่าβ_U แตกต่างจาก 0 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

เกณฑ์การตัดสินข้อสอบที่ทำหน้าที่ต่างกันตามวิธี IRT หมายถึงข้อสอบที่มีพื้นที่อันเกิดจากความ

แตกต่างระหว่างโค้งคุณลักษณะข้อสอบของผู้สอบ 2 กลุ่ม ในที่นี้ใช้การทดสอบนัยสำคัญด้วยสถิติ Z ที่ระดับ .05

อัตราความถูกต้องของการตรวจสอบ หมายถึงอัตราส่วนหรือร้อยละของจำนวนข้อสอบที่ตรวจสอบพบว่าทำหน้าที่ต่างกันได้อย่างถูกต้อง โดยคำนวณจากจำนวนข้อสอบที่ตรวจสอบพบว่าทำหน้าที่ต่างกันได้ถูกต้องต่อจำนวนข้อสอบที่ทำหน้าที่ต่างกันทั้งหมดในแบบสอบ

อัตราความคลาดเคลื่อนของการตรวจสอบ หมายถึงอัตราส่วนหรือร้อยละของจำนวนข้อสอบที่ระบุผิดพลาดซึ่งมี 2 ประเภทคือ ความคลาดเคลื่อนประเภทที่ 1 และความคลาดเคลื่อนประเภทที่ 2

ความคลาดเคลื่อนประเภทที่1 (Type I error) คือการระบุผิดพลาดว่าข้อสอบทำหน้าที่ต่างกัน (false positive) ทั้งที่ความเป็นจริงข้อสอบทำหน้าที่ไม่ต่างกัน ซึ่งในที่นี้คำนวนได้จากค่าอัตราส่วนของจำนวนข้อสอบที่ระบุผิดพลาดว่าทำหน้าที่ต่างกันต่อจำนวนข้อสอบที่ทำหน้าที่ไม่ต่างกันทั้งหมดในแบบสอบ

ความคลาดเคลื่อนประเภทที่2 (Type II error) คือการระบุผิดพลาดว่าข้อสอบทำหน้าที่ไม่ต่างกัน (false negative) ทั้งที่ความเป็นจริงข้อสอบทำหน้าที่ต่างกัน ซึ่งในที่นี้คำนวนได้จากค่าอัตราส่วนของจำนวนข้อสอบที่ระบุผิดพลาดว่าทำหน้าที่ไม่ต่างกันต่อจำนวนข้อสอบที่ทำหน้าที่ต่างกันทั้งหมดในแบบสอบ

ประสิทธิภาพของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ หมายถึงความถูกต้องของการระบุการทำหน้าที่ต่างกันของข้อสอบ จากการตรวจสอบด้วยวิธี MH และวิธี SIBTEST ซึ่งพิจารณาได้จากอัตราความถูกต้องของการตรวจสอบและความคลาดเคลื่อนของการตรวจสอบ

กลุ่มอ้างอิง (reference group) หมายถึงกลุ่มผู้สอบที่คาดว่าจะได้ประโยชน์จากการตอบข้อสอบที่ทำหน้าที่ต่างกัน คือเป็นกลุ่มที่มีความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องสูงกว่าผู้สอบอีกกลุ่มหนึ่งทั้ง ๆ ที่มีความสามารถเท่ากัน ในที่นี้กลุ่มอ้างอิงได้จากการสุ่มผู้สอบจากผู้สอบที่จำลองขึ้นด้วยโปรแกรม IRTDATA

กลุ่มเปรียบเทียบ (focal group) หมายถึงกลุ่มของผู้สอบที่คาดว่าจะเสียประโยชน์จากการตอบข้อสอบที่ทำหน้าที่ต่างกัน คือเป็นกลุ่มทีมีความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องต่ำกว่าผู้สอบอีกกลุ่มหนึ่ง ทั้ง ๆ ที่มีความสามารถเท่ากัน ในที่นี้กลุ่มเปรียบเทียบได้จากการสุ่มผู้สอบจากผู้สอบที่จำลองขึ้นด้วยโปรแกรม IRTDATA

ขนาดกลุ่มตัวอย่าง หมายถึงจำนวนผู้สอบที่ได้จากการจำลองที่นำมาใช้ในการตรวจสอบ มี 12 ขนาดโดยมีจำนวนผู้สอบกลุ่มย่อยระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบดังนี้

200:200 200:180 200:150 200:100 600:600 600:540 600:450 600:300
1000:1000 1000:900 1000:750 1000:500

ความยาวแบบสอบ หมายถึงจำนวนข้อสอบในแบบสอบ ในการวิจัยครั้งนี้ศึกษาความยาวแบบสอบ 3 ขนาดคือ 30, 60 และ 90 ข้อ

1.7 ประโยชน์ที่จะได้รับ

1.เพื่อเป็นแนวทางในการเลือกวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่เหมาะสมต่อการปฏิบัติ โดยคำนึงถึงประสิทธิภาพ ความสะดวกและประหยัด

2.เพื่อเป็นแนวทางในการเลือกใช้ขนาดกลุ่มตัวอย่างและความยาวแบบสอบ ว่าควรใช้กลุ่มตัวอย่างขนาดเท่าใด และควรใช้กลุ่มผู้สอบย่อยระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบเป็นอัตราส่วนต่อกันเท่าใด จึงจะทำให้ประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบดีที่สุด

3.เพื่อเป็นแนวทางในการเลือกใช้ความยาวแบบสอบ ว่าควรใช้แบบสอบ ที่มีระดับความยาวเท่าใดจึงจะทำให้ประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบดีที่สุด