บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

"ความตรง" เป็นหัวใจสำคัญของแบบสอบ ในการสร้างและการตรวจสอบคุณภาพของแบบสอบจะต้องคำนึงถึงคุณภาพด้านความตรงเป็นสำคัญ ทั้งนี้เพราะว่าความตรงเป็นคุณสมบัติของแบบสอบที่แสดงถึงความสามารถในการวัดได้ถูกต้องแม่นยำ ถ้าผลการวัดได้ค่าที่ใกล้เคียงกับค่าคุณลักษณะที่แท้จริงเพียงใด ก็ถือว่าการวัดมีความตรงมากขึ้นเพียงนั้น (ศิริชัย กาญจนวาสี, 2535) นักวิธีวิทยาการวิจัยมักจะนิยมตรวจสอบความตรงของแบบสอบ 3 ประเภทหลัก คือ (1) ความตรงตามเนื้อหา (content validity) (2) ความตรงตามเกณฑ์สัมพันธ์ (criterion-related validity) และ (3) ความตรงตามภาวะสันนิษฐาน (construct validity) ส่วนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (differential item functioning; DIF) ก็เป็นอีกประเภทหนึ่งที่ใช้ตรวจสอบคุณภาพด้านความตรงของแบบสอบ (Mazor, Clauser and Hambleton, 1992; S.-H. Kim, H.-O. Kim and Cohen, 1994) โดยเป็นการตรวจสอบในประเด็นความอยุติธรรมของข้อสอบ (item unfairness) ข้อสอบข้อใดที่ทำหน้าที่ต่างกันจะถูกคัดออกจากแบบสอบ โดยทั่วไปแล้วในแบบสอบมาตรฐานวัดผลสัมฤทธิ์ทางการเรียนถ้ามีสัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 ถึง 15 ถือว่าไม่ผิดปกติ แต่ถ้ามีสัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20 ถือว่าเป็นเรื่องผิดพลาดอย่างมาก (Clauser, 1993 cited in Narayanan and Swaminathan, 1994) ในปัจจุบันนักวิธีวิทยาการวิจัยได้ให้ความสนใจการตรวจสอบในประเด็นดังกล่าวมากยิ่งขึ้น ทั้งนี้เนื่องจากแบบสอบมาตรฐานไม่เพียงแต่นำไปใช้ในการสอบวัดผลสัมฤทธิ์ทางการเรียนเท่านั้น แต่ยังนำไปใช้ในด้านอื่น ๆ อีก เช่น การสอบคัดเลือกเพื่อศึกษาต่อ การสอบบรรจุบุคคลเข้าทำงาน และการสอบเลื่อนตำแหน่ง เป็นต้น

การศึกษาเรื่องความอยุติธรรมของข้อสอบในกรณีที่ทำให้ผู้สอบระหว่างกลุ่มย่อยมีความได้เปรียบหรือเสียเปรียบกัน เดิมใช้คำว่าการตรวจสอบ ***"ความลำเอียงของข้อสอบหรือแบบสอบ"*** (*item* / test *bias*) ซึ่งเป็นภาษาที่ใช้กันในทางสังคมและมีความหมายในเชิงลบ สำหรับการตัดสินว่าข้อสอบมีความลำเอียงหรือไม่นั้น มักจะพิจารณาอิทธิพลที่สังเกตได้ของผู้สอบกลุ่มย่อยที่นำมาศึกษา โดยไม่ได้คำนึงถึงวิธีทางสถิติ จึงทำให้เกิดความคลุมเครือเกี่ยวกับเกณฑ์ที่ใช้ในการตัดสินความลำเอียง ต่อมาในระยะหลังนักวิธีวิทยาการวิจัยได้พัฒนาวิธีใหม่ ๆ เพื่อใช้วิเคราะห์

ดัชนีความลำเอียงของข้อสอบ โดยมุ่งเน้นความแตกต่างระหว่างกลุ่มผู้สอบที่ตอบข้อสอบข้อเดียวกัน และเลือกเกณฑ์ที่ใช้ในการจับคู่กลุ่มผู้สอบ (matching criterion) ทั้งยังนำสารสนเทศทางสถิติมาใช้เป็นเกณฑ์ในการตัดสินความลำเอียงของข้อสอบ ดังนั้นเพื่อความเหมาะสมจึงเปลี่ยนมาใช้คำใหม่ว่าการตรวจสอบ ***"การทำหน้าที่ต่างกันของข้อสอบหรือแบบสอบ"*** (*differential item / test function*; *DIF / DTF*) ซึ่งเป็นคำที่มีความหมายเป็นกลางมากกว่า (Holland and Thayer, 1988; Holland and Wainer, 1993; Shealy and Stout, 1993)

ในการทดสอบแต่ละครั้ง ผู้สอบระหว่างกลุ่มย่อยอาจมีลักษณะแตกต่างกัน เช่น เชื้อชาติ ศาสนา วัฒนธรรม ภูมิลำเนา สังคม เพศ ภาษา อายุ และประสบการณ์ เป็นต้น ผู้สอบกลุ่มย่อยดังกล่าวอาจไม่ได้รับความยุติธรรมในการทำข้อสอบ โดยข้อสอบบางข้ออาจมีความลำเอียงเข้าข้างผู้สอบกลุ่มย่อยบางกลุ่มของผู้เข้าสอบทั้งหมด ซึ่งทำให้เกิดการได้เปรียบหรือเสียเปรียบระหว่างผู้สอบกลุ่มย่อยด้วยกัน ทั้ง ๆ ที่สอบด้วยข้อสอบข้อเดียวกันหรือแบบสอบฉบับเดียวกัน แสดงว่าแบบสอบหรือข้อสอบดังกล่าวขาดความตรง สาเหตุดังกล่าวอาจเนื่องมาจากแบบสอบไม่ได้วัดความสามารถเป้าหมายที่ต้องการวัด (target ability; θ) เพียงอย่างเดียว แต่ยังวัดความสามารถแทรกซ้อนที่ไม่ต้องการวัด (nuisance ability; η) อีกด้วย ตัวอย่างเช่น แบบสอบวัดคำศัพท์ในวิชาภาษาอังกฤษฉบับหนึ่ง ข้อสอบบางข้ออาจถามความรู้สำหรับผู้ชายเป็นพิเศษ เช่น ความรู้เรื่องกีฬา ในขณะที่ข้อสอบบางข้ออาจถามความรู้สำหรับผู้หญิงโดยเฉพาะ เช่น ความรู้เกี่ยวกับงานในบ้าน จากสถานการณ์ดังกล่าว ทักษะวัดคำศัพท์ในวิชาภาษาอังกฤษเป็นความสามารถเป้าหมาย (θ) ส่วนทักษะวัดความรู้ทางด้านกีฬา (η_1) และงานในบ้าน (η_2) เป็นความสามารถแทรกซ้อน ข้อสอบทุกข้อในแบบสอบจะวัดความสามารถเป้าหมาย ส่วนข้อสอบบางข้อที่ทำหน้าที่ต่างกันจะวัดทั้งความสามารถเป้าหมายและความสามารถแทรกซ้อน (Nandakumar, 1993) นั่นคือ ถ้าผู้สอบกลุ่มย่อยกลุ่มใดมีความสามารถแทรกซ้อนสูงกว่าก็มีโอกาสในการตอบข้อสอบได้ถูกต้องมากกว่า ทั้ง ๆ ที่ระดับความสามารถเป้าหมายที่ต้องการวัดเท่ากัน จึงมีผลทำให้ข้อสอบทำหน้าที่ต่างกัน

ข้อสอบทำหน้าที่ต่างกันเมื่อ***ผู้สอบที่มีความสามารถระดับเดียวกันแต่อยู่ต่างกลุ่มกันมีโอกาสของการตอบข้อสอบได้ถูกต้องไม่เท่ากัน*** (Li and Stout, 1996) ซึ่งขนาดและทิศทางของข้อสอบที่ทำหน้าที่ต่างกันจะแปรเปลี่ยนไปตามระดับความสามารถที่แตกต่างกัน โดยข้อสอบที่ทำหน้าที่ต่างกันแบ่งออกเป็น 2 ประเภท (Mellenbergh, 1982) คือ ***ข้อสอบทำหน้าที่ต่างกันแบบเอกรูป*** (*uniform DIF*) และ***ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป*** (*nonuniform DIF*) ข้อสอบที่ทำหน้าที่ต่างกันประเภทแรกเกิดขึ้นเมื่อไม่มีปฏิสัมพันธ์ (interaction) ระหว่างระดับ

ความสามารถของผู้สอบและการเป็นสมาชิกของกลุ่ม (group membership) ส่วนข้อสอบที่ทำหน้าที่ต่างกันประเภทหลังเกิดขึ้นเมื่อมีปฏิสัมพันธ์ระหว่างระดับความสามารถของผู้สอบและการเป็นสมาชิกของกลุ่ม ซึ่งตามทฤษฎีการตอบสนองข้อสอบ (item response theory; IRT) สามารถพิจารณาปฏิสัมพันธ์ดังกล่าวได้จากความแตกต่างของค่าพารามิเตอร์อำนาจจำแนกของข้อสอบระหว่างผู้สอบกลุ่มย่อย 2 กลุ่ม กล่าวคือ ถ้าข้อสอบระหว่างผู้สอบกลุ่มย่อย 2 กลุ่มมีค่าอำนาจจำแนกเท่ากันแล้วโค้งลักษณะข้อสอบ (item characteristic curves; ICCs) ระหว่างผู้สอบดังกล่าวจะขนานกัน แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบเอกรูป แต่ถ้าข้อสอบระหว่างผู้สอบกลุ่มย่อย 2 กลุ่มมีค่าอำนาจจำแนกไม่เท่ากันแล้วโค้งลักษณะข้อสอบระหว่างผู้สอบดังกล่าวจะไม่ขนานกัน แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป (Camilli and Shepard, 1994) โดยปกติแล้วในแบบสอบมาตรฐานจะมีข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูปมากกว่าข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูป แต่ในข้อมูลจริงจะมีข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปมากกว่า จากการศึกษาของ Swaminathan และ Rogers (1990) ภายใต้ทฤษฎี IRT เมื่อกำหนดระดับความสามารถในช่วง –3 ถึง +3 พบว่า ข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปแบ่งออกเป็น 2 ลักษณะ คือ (1) ***ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปโดยมีปฏิสัมพันธ์ไม่เป็นลำดับ*** (*disordinal interaction*) เกิดขึ้นเมื่อโค้งลักษณะข้อสอบตัดกันตรงจุดกึ่งกลางของช่วงความสามารถ และ (2) ***ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปโดยมีปฏิสัมพันธ์เป็นลำดับ*** (*ordinal interaction*) เกิดขึ้นเมื่อโค้งลักษณะข้อสอบตัดกันนอกช่วงความสามารถ ซึ่งอาจตัดกันตรงปลายสุดของช่วงความสามารถต่ำหรือสูง นอกจากนี้ยังพบว่า โค้งลักษณะข้อสอบที่ไม่ขนานกันอาจไม่ตัดกันก็ได้ ดังเช่น กรณีโมเดลแบบ 3 พารามิเตอร์ ดังนั้นโค้งลักษณะข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปซึ่งไม่ขนานกันอาจจะตัดกันหรือไม่ตัดกันก็ได้ ต่อมา Li และ Stout (1993 cited in Narayanan and Swaminathan, 1996) เรียกข้อสอบที่ทำหน้าที่ต่างกันโดยมีปฏิสัมพันธ์ไม่เป็นลำดับว่า ***"ข้อสอบทำหน้าที่ต่างกันแบบไม่มีทิศทาง"*** (*nondirectional DIF*) และเรียกข้อสอบที่ทำหน้าที่ต่างกันโดยมีปฏิสัมพันธ์เป็นลำดับว่า ***"ข้อสอบทำหน้าที่ต่างกันแบบมีทิศทางเดียว"*** (*unidirectional DIF*)

หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจะเปรียบเทียบผลการตอบข้อสอบระหว่างผู้สอบกลุ่มย่อยสองกลุ่มที่มีความสามารถระดับเดียวกัน โดยที่ผู้สอบกลุ่มหนึ่งเป็นตัวแทนกลุ่มหลักในประชากรเรียกว่า ***"กลุ่มอ้างอิง"*** (*reference group*; *R*) ซึ่งเป็นกลุ่มพื้นฐาน ส่วนอีกกลุ่มหนึ่งเป็นตัวแทนกลุ่มรองในประชากรเรียกว่า ***"กลุ่มเปรียบเทียบ"*** (*focal group*; *F*) ซึ่งตามปกติเป็นกลุ่มผู้สอบที่สนใจจะทำการศึกษาการทำหน้าที่ต่างกันของข้อสอบ (Angoff, 1993) ถ้าข้อสอบทำหน้าที่ต่างกันแล้วโอกาสในการตอบข้อสอบถูกของผู้สอบแต่ละกลุ่มจะไม่เท่ากัน โดย

คาดว่าผู้สอบกลุ่มแรกจะได้เปรียบในการตอบข้อสอบ ส่วนผู้สอบกลุ่มหลังคาดว่าจะเสียเปรียบในการตอบข้อสอบ สำหรับวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนแบบทวิภาค (dichotomous) จำแนกออกเป็น 2 กลุ่มวิธีใหญ่ ๆ ดังนี้ (Feinstein, 1995; Potenza and Dorans, 1995)

1. **กลุ่มวิธี** non-IRT เป็นวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ใช้คะแนนสังเกตได้ (observe score) ซึ่งวิเคราะห์ตามทฤษฎีแบบดั้งเดิม (classical test theory; CTT) วิธีในกลุ่มนี้มักใช้คะแนนรวมของผู้สอบเป็นเกณฑ์การจับคู่ของกลุ่มผู้สอบ (matching criterion) ซึ่งเป็นเกณฑ์ภายใน วิธีการตรวจสอบที่สำคัญในกลุ่มนี้ ได้แก่ วิธีการวิเคราะห์ความแปรปรวน (analysis of variance; ANOVA) (Cleary and Hillton, 1968 cited in Osterlind, 1983) วิธีแปลงค่าความยากของข้อสอบ (transformed item difficulty; TID) (Angoff, 1972 cited in Osterlind, 1983) วิธีตารางการณ์จร (contingency table; CT) (Camilli and Shepard, 1994) วิธีการทำให้เป็นมาตรฐาน (standardization; STND) (Dorans and Kulick, 1986) และวิธีการถดถอยโลจิสติก (logistic regression; LR) (Swaminathan and Rogers, 1990) สำหรับวิธีตารางการณ์จรยังแบ่งเป็นวิธีย่อย ๆ อีก เช่น วิธีไค-สแควร์ (chi-square; χ^2) (Scheuneman, 1979) วิธีลอก-ลิเนียร์ (log-linear; LL) (Mellenbergh, 1982) และวิธีแมนเทล-แฮนส์เซล (Mantel-Haenszel; MH) (Holland and Thayer, 1988) เป็นต้น

2. **กลุ่มวิธี** IRT เป็นวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ใช้คะแนนสังเกตไม่ได้ (unobserved score) หรือใช้ตัวแปรแฝง (latent variable) ซึ่งวิเคราะห์ตามทฤษฎีการตอบสนองข้อสอบ (item response theory; IRT) ดังนั้นวิธีในกลุ่มนี้จะใช้ค่าประมาณความสามารถของผู้สอบเป็นเกณฑ์การจับคู่กลุ่มผู้สอบ ซึ่งแตกต่างจากวิธี non-IRT ที่ใช้คะแนนรวมของผู้สอบ วิธีการตรวจสอบในกลุ่มนี้แบ่งออกเป็น 3 วิธีใหญ่ ๆ คือ วิธีการวัดพื้นที่ วิธีการเปรียบเทียบค่าพารามิเตอร์ และวิธีซิปเทสท์ (SIBTEST) (Shealy and Stout, 1993) วิธีการวัดพื้นที่แบ่งออกเป็นวิธีย่อย ๆ อีกหลายวิธี เช่น วิธีการวัดพื้นที่ของ Rudner (1977) วิธีการวัดพื้นที่ของ Linn, Levine, Hastings และ Wardrup (1981) วิธีการวัดพื้นที่ของ Shepard, Camilli และ Williams (1984) วิธีการวัดพื้นที่ของ Raju (1990) วิธีการวัดพื้นที่ของ Kim และ Cohen (1991) เป็นต้น สำหรับวิธีการเปรียบเทียบค่าพารามิเตอร์แบ่งออกเป็นวิธีย่อย ๆ เช่น วิธีเปลี่ยนค่าความยาก (difficulty shift) ของ Wright, Mead และ Komocar (1976 cited in Hulin, Drasgow and Parson, 1983) วิธีการทดสอบ F ของ Hulin, Drasgow และ Komocar (1982) วิธีการตอบสนองข้อสอบแบบเทียม (pseudo-IRT) ของ Linn และ Harnisch (1981) วิธีการทดสอบไค-สแควร์ของ Lord (Lord's

chi-square test) (1980) และวิธีการทดสอบอัตราส่วนไลค์ลิฮู้ด (likelihood ratio test; LR) (Thissen, Steinberge and Wainer, 1993)

เทคนิคการวิเคราะห์ความแปรปรวนและวิธีแปลงค่าความยากของข้อสอบเป็นวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในยุคเริ่มต้น ทั้งสองวิธีมีแนวความคิดในการตรวจสอบที่คล้ายกัน หลักการตรวจสอบจะพิจารณาค่าความยากสัมพัทธ์ของข้อสอบจากผลของปฏิสัมพันธ์ (interaction effect) ระหว่างกลุ่มผู้สอบและข้อสอบ จุดเด่นของวิธีการวิเคราะห์ความแปรปรวนก็คือใช้สถิติ F ทดสอบนัยสำคัญเพื่อตัดสินการทำหน้าที่ต่างกันของข้อสอบ สถิติดังกล่าวมีอำนาจการทดสอบสูง (powerful) ส่วนวิธีแปลงค่าความยากของข้อสอบไม่ได้ใช้สถิติทดสอบนัยสำคัญแต่จะใช้เกณฑ์ที่เชื่อถือได้ซึ่งมีผู้เสนอขึ้นมา วิธีแปลงค่าความยากของข้อสอบมีจุดเด่นตรงที่มีดัชนีวัดขนาดของการทำหน้าที่ต่างกันของข้อสอบ (size of DIF) ซึ่งวิธีการวิเคราะห์ความแปรปรวนไม่มี ทั้ง 2 วิธีใช้กลุ่มตัวอย่างที่มีขนาดเล็ก โดยเฉพาะวิธีแปลงค่าความยากของข้อสอบใช้กลุ่มตัวอย่างเพียง 30 – 50 คนต่อกลุ่ม (Osterlind, 1983) ในปัจจุบันวิธีทั้งสองไม่นิยมใช้แล้ว เพราะว่าดัชนีการทำหน้าที่ต่างกันของข้อสอบที่วิเคราะห์ตามทฤษฎี CTT จะประกอบด้วยค่าความยากและค่าอำนาจจำแนกของข้อสอบที่ไม่สามารถแยกออกจากกัน (compound) จึงทำให้มีปัญหา 2 ประการ คือ ***ประการแรก*** ข้อสอบที่มีค่าอำนาจจำแนกสูงมีแนวโน้มทำให้ความแตกต่างของความยากของข้อสอบ (ค่า p) มีค่ามากขึ้น และ***ประการที่สอง*** ข้อสอบที่มีค่าความยากปลายสุด (extreme difficulty) เช่น ข้อสอบที่ยากมากหรือง่ายมาก มีแนวโน้มทำให้ความแตกต่างของค่า p น้อยกว่าข้อสอบที่มีความยากปานกลาง (Camilli and Shepard, 1994) จากปัญหาดังกล่าวจะส่งผลให้การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบบิดเบือน (distort) ไปจากความเป็นจริง กล่าวคือเกิดความคลาดเคลื่อนประเภทที่ 1 (type I error) โดยอาจระบุผิดพลาดว่าข้อสอบทำหน้าที่ต่างกันทั้งที่ความเป็นจริงข้อสอบทำหน้าที่ไม่ต่างกัน หรือเกิดความคลาดเคลื่อนประเภทที่ 2 (type II error) โดยไม่สามารถระบุข้อสอบที่ทำหน้าที่ต่างกันจริงได้

จากข้อจำกัดดังกล่าวจึงทำให้นักวิธีวิทยาการวิจัยหันมาพัฒนาวิธีตารางการณ์จรและวิธีในกลุ่ม IRT ซึ่งทั้งสองวิธีมีแนวคิดในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบคล้ายกัน โดยใช้หลักการวิเคราะห์ความแตกต่างของผลการตอบข้อสอบจากผู้สอบสองกลุ่มที่มีความสามารถระดับเดียวกัน ในวิธีตารางการณ์จรจะใช้คะแนนรวมที่สังเกตได้แทนระดับความสามารถ ส่วนวิธีในกลุ่ม IRT จะประมาณค่าความสามารถโดยใช้ฟังก์ชันทางคณิตศาสตร์ตามลักษณะโมเดล IRT จุดเด่นของวิธีตารางการณ์จรก็คือ สามารถนำไปประยุกต์ใช้กับกลุ่มตัวอย่างที่มีขนาดเล็ก ทั้งยังมีวิธีดำเนินการตรวจสอบง่าย และไม่จำเป็นต้องใช้โปรแกรมคอมพิวเตอร์ที่ซับซ้อนเหมือนกับวิธี IRT

ทั่ว ๆ ไป (Camilli and Shepard, 1994) ในการวิเคราะห์ด้วยวิธีนี้จะนำข้อมูลมาสร้างตารางการณ์จรแบบ 3 ทิศทาง คือ กลุ่มผู้สอบ (กลุ่มอ้างอิง / กลุ่มเปรียบเทียบ) ผลการตอบข้อสอบ (ถูก / ผิด) และคะแนนรวม (K ระดับ) วิธีตารางการณ์จรแบ่งออกเป็นวิธีย่อย ๆ หลายวิธี แต่วิธีที่นิยมใช้ในปัจจุบันก็คือ วิธีแมนเทล-แฮนส์เซล ซึ่งพัฒนาโดย Holland และ Thayer (1988) จุดเด่นของวิธีแมนเทล-แฮนส์เซลก็คือ เป็นวิธีนันพาราเมทริก (nonparametric) ไม่จำเป็นต้องใช้โมเดลประมาณค่าพารามิเตอร์ สามารถคำนวณได้ง่าย ไม่จำเป็นต้องใช้กลุ่มตัวอย่างที่มีขนาดใหญ่ และเสียค่าใช้จ่ายไม่แพง นอกจากนี้ยังมีดัชนีวัดขนาดของการทำหน้าที่ต่างกันของข้อสอบ และใช้สถิติทดสอบนัยสำคัญ (Narayanan and Swaminathan, 1994, 1996) เทคนิคการตรวจสอบด้วยวิธีแมนเทล-แฮนส์เซลมีความสัมพันธ์สูงกับวิธีการทำให้เป็นมาตรฐานที่พัฒนาโดย Doran และ Kulick (1986) จุดเด่นของวิธีการทำให้เป็นมาตรฐานก็คือ มีดัชนีวัดขนาดของการทำหน้าที่ต่างกันของข้อสอบ แต่ไม่มีรูปแบบสถิติที่ใช้ทดสอบสมมติฐาน (Potenza and Doran, 1995) สำหรับวิธีไค-สแควร์ของ Scheuneman (1979) เป็นวิธีที่ใช้ตารางการณ์จรในยุคต้น ๆ ซึ่งไม่นิยมนำมาใช้ในปัจจุบันนี้แล้ว ทั้งนี้เนื่องจากใช้สถิติที่ไม่มีอำนาจการทดสอบ และไม่มีดัชนีวัดขนาดของการทำหน้าที่ต่างกันของข้อสอบ (Green, 1994) ส่วนวิธีลอก-ลิเนียร์ของ Mellenbergh (1982) จะใช้โมเดลลอก-ลิเนียร์วิเคราะห์ข้อมูลแล้วพิจารณาผลของปฏิสัมพันธ์ระหว่างสมาชิกกลุ่มผู้สอบและระดับความสามารถซึ่งคล้ายกับวิธีการวิเคราะห์ความแปรปรวน จุดอ่อนของวิธีลอก-ลิเนียร์ก็คือ ในการวิเคราะห์จะใช้ระดับความสามารถของผู้สอบแยกกันเป็นกลุ่ม ๆ ไม่เป็นลำดับ ซึ่งวิธีที่ใช้ระดับความสามารถแบบต่อเนื่องจะมีอำนาจการทดสอบมากกว่า จากข้อจำกัดดังกล่าวจึงทำให้ Swaminathan และ Rogers (1990) นำวิธีลอก-ลิเนียร์และวิธีแมนเทล-แฮนส์เซลมาดัดแปลงเป็นวิธีการถดถอยโลจิสติกโดยใช้ตัวแปรความสามารถแบบต่อเนื่อง วิธีนี้จะวิเคราะห์โดยใช้โมเดลการถดถอยโลจิสติก ซึ่งมีจุดเด่นตรงที่สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ทั้งแบบเอกรูปและแบบอเนกรูป ทั้งยังเป็นวิธีที่มีความยืดหยุ่นสูง โดยสามารถนำไปปรับขยายเพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีผู้สอบหลายกลุ่มและข้อสอบที่มีการให้คะแนนแบบพหุวิภาค (polytomous) (Miller and Spray, 1993)

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่กล่าวมาแล้วข้างต้นเป็นวิธีในกลุ่ม non-IRT ที่วิเคราะห์ข้อมูลด้วยทฤษฎี CTT ทฤษฎีดังกล่าวมีข้อจำกัดเกี่ยวกับความคงที่ของค่าสถิติโดยจะมีค่าแปรเปลี่ยนไปตามกลุ่มตัวอย่าง ซึ่งจะส่งผลให้การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบไม่น่าเชื่อถือ นอกจากนี้วิธีในกลุ่มดังกล่าวยังใช้เกณฑ์การจับคู่กลุ่มผู้สอบภายใน (internal matching criterion) โดยใช้คะแนนรวมของแบบสอบแทนระดับความสามารถของผู้สอบ ซึ่งอาจ

มีผลทำให้แบบสอบเกิดความลำเอียง เพราะว่าการใช้เกณฑ์ดังกล่าวจะใช้คะแนนรวมที่ได้มาจากข้อสอบที่ทำหน้าที่ต่างกันรวมอยู่ด้วย (Zieky, 1993) จากจุดอ่อนของทฤษฎี CTT จึงทำให้นักวิธีวิทยาการวิจัยนำทฤษฎี IRT ไปประยุกต์ใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยเชื่อกันว่าเป็นทฤษฎีที่มีความแกร่ง สามารถแก้ปัญหาของทฤษฎี CTT ได้ จุดเด่นของทฤษฎี IRT ก็คือ ความไม่แปรเปลี่ยนของค่าพารามิเตอร์ของข้อสอบ โดยสามารถนำคุณสมบัติสถิติของข้อสอบไปอธิบายลักษณะของข้อสอบที่ทำหน้าที่ต่างกันได้แม่นยำกว่าทฤษฎี CTT (Camilli and Shepard, 1994) ส่วนเกณฑ์การจับคู่กลุ่มผู้สอบจะใช้ค่าประมาณระดับคุณลักษณะของบุคคล ซึ่งต่างจากวิธีในกลุ่ม CTT ที่ใช้คะแนนรวม ในการพัฒนาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบภายใต้ทฤษฎี IRT ตามโมเดลแบบทวิภาคแบ่งออกเป็น 2 วิธีใหญ่ ๆ คือ วิธีการวัดพื้นที่และวิธีการเปรียบเทียบค่าพารามิเตอร์ของข้อสอบ ทั้ง 2 วิธียังแบ่งออกเป็นวิธีย่อย ๆ อีกหลายวิธี แต่วิธีที่นิยมนำมาใช้ในปัจจุบันมี 4 วิธี คือ วิธีการวัดพื้นที่ของ Raju (1990) วิธีการวัดพื้นที่ของ Kim และ Cohen (1991) วิธีการทดสอบไค-สแควร์ของ Lord (1980) และวิธีการทดสอบอัตราส่วน-ไลค์ลิฮู้ด (Thissen, Steinberge and Wainer, 1993) จากการศึกษาของ Kim และ Cohen (1991) Cohen และ Kim (1993) Raju, Drasgow และ Slinde (1993) Kim และ Cohen (1995) พบว่า ทั้ง 4 วิธีให้ผลการตรวจสอบที่คล้ายกัน โดยเฉพาะต้องใช้กลุ่มตัวอย่างที่มีขนาดใหญ่และแบบสอบที่ยาวมาก

วิธีการวัดพื้นที่ของ Raju จะคำนวณพื้นที่ในช่วงเปิด (open-interval or exact area) ส่วนวิธีการวัดพื้นที่ของ Kim และ Cohen จะคำนวณพื้นที่ในช่วงปิด (closed-interval area) จุดเด่นของวิธีการวัดพื้นที่ทั้งสองลักษณะก็คือ สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปและแบบอเนกรูปได้ถูกต้องแม่นยำ โดยจะใช้วิธีการคำนวณทางคณิตศาสตร์ที่เรียกว่า "การอินทิเกรทแบบต่อเนื่อง" (continuous integration) คำนวณพื้นที่ระหว่างโค้งลักษณะข้อสอบจากผู้สอบสองกลุ่ม ภายใต้โมเดลโลจิสติกแบบ 1, 2 และ 3 พารามิเตอร์ ทั้งยังสามารถคำนวณพื้นที่ชนิดคิดเครื่องหมาย (signed) และชนิดไม่คิดเครื่องหมาย (unsigned) ในการตัดสินข้อสอบที่ทำหน้าที่ต่างกันของวิธีการวัดพื้นที่ของ Raju จะใช้สถิติทดสอบนัยสำคัญ ส่วนวิธีการวัดพื้นที่ของ Kim และ Cohen จะไม่ใช้การทดสอบทางสถิติ แต่จะนำขนาดของพื้นที่ไปเปรียบเทียบกับเกณฑ์ที่กำหนดไว้ สำหรับวิธีการวัดพื้นที่ของ Rudner (1977) วิธีการวัดพื้นที่ของ Linn และคณะ (1981) วิธีการวัดพื้นที่ของ Shepard และคณะ (1984) ไม่นิยมนำมาใช้ในปัจจุบัน เพราะว่าวิธีดังกล่าวจะคำนวณพื้นที่แบบหยาบ ๆ โดยใช้การประมาณค่าแบบไม่ต่อเนื่อง (discrete approximation) สำหรับวิธีการทดสอบไค-สแควร์ของ Lord (1980) จะใช้สถิติ Wald ทดสอบการเท่ากันของค่า

พารามิเตอร์ของข้อสอบจากผู้สอบสองกลุ่ม ภายใต้โมเดลโลจิสติกแบบ 2 หรือ 3 พารามิเตอร์ ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีดังกล่าว เมื่อประมาณค่าความสามารถด้วยวิธี marginal maximum likelihood หรือวิธี Bayes modal จะทำให้ผลการตรวจสอบมีความถูกต้องแม่นยำมากกว่าวิธี joint maximum likelihood (McLaughlin and Drasgow, 1987 cited in Raju and others, 1993) จากข้อได้เปรียบของวิธีการทดสอบไค-สแควร์ของ Lord จึงทำให้วิธีเปลี่ยนค่าความยากของ Wright และคณะ (1976 cited in Hulin and others, 1983) วิธีการทดสอบ F ของ Hulin และคณะ (1982) วิธี IRT แบบเทียมของ Linn และ Harnisch (1981) ไม่นิยมนำมาใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ทั้งนี้เพราะว่าวิธีเปลี่ยนค่าความยากจะทดสอบความแตกต่างของค่าพารามิเตอร์โดยใช้สถิติ Draba ภายใต้โมเดลของ Rasch ซึ่งดัชนีการทำหน้าที่ต่างกันของข้อสอบที่คำนวณจากโมเดลแบบ 2 หรือ 3 พารามิเตอร์จะมีความแม่นยำมากกว่า (Camilli and Shepard, 1994) สำหรับวิธีการทดสอบ F จะทดสอบการเท่ากันของฟังก์ชันการถดถอยด้วยสถิติการทดสอบอัตราส่วน F ภายใต้โมเดลโลจิสติกแบบ 2 พารามิเตอร์ ส่วนวิธี IRT แบบเทียมเป็นวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบภายใต้โมเดลโลจิสติกแบบ 3 พารามิเตอร์ ซึ่งนำมาประยุกต์ใช้กับกลุ่มตัวอย่างที่มีขนาดเล็กมากเพียง 300 คน ทั้งวิธีการทดสอบ F และวิธี IRT แบบเทียมเป็นวิธีการประมาณค่าแบบคร่าว ๆ ทำให้ขาดความแม่นยำ นอกจากนี้ยังเป็นวิธีที่ล้าสมัย เพราะในปัจจุบันมีโปรแกรมคอมพิวเตอร์ที่ใช้กับเครื่องคอมพิวเตอร์ส่วนบุคคล สามารถประมาณค่าพารามิเตอร์ได้สะดวกและรวดเร็ว เช่น โปรแกรม BILOG (Mislevy and Bock, 1990) ดังนั้นการประมาณค่าพารามิเตอร์โดยใช้โมเดลโลจิสติกแบบ 2 พารามิเตอร์ หรือ 3 พารามิเตอร์จึงไม่เป็นปัญหาอีกต่อไป สำหรับวิธีการทดสอบอัตราส่วนไลค์ลิฮู้ดจะทดสอบการเท่ากันของค่าพารามิเตอร์ของข้อสอบระหว่างผู้สอบสองกลุ่มโดยใช้สถิติอัตราส่วนไลค์ลิฮู้ด จุดเด่นของวิธีนี้ก็คือไม่จำเป็นต้องปรับเทียบ (equating) สเกลพารามิเตอร์ของข้อสอบจากเมทริกซ์หนึ่งไปยังอีกเมทริกซ์หนึ่ง เนื่องจากพารามิเตอร์ของข้อสอบระหว่างผู้สอบสองกลุ่มจะถูกประมาณค่าพร้อม ๆ กัน (simultaneous calibration) (Kim and Cohen, 1995)

นอกจากนี้แล้วยังมีวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในกลุ่ม IRT อีกวิธีหนึ่ง คือ วิธีซิปเทสท์ (simultaneous item bias test; SIBTEST) วิธีนี้ Shealy และ Stout (1993) พัฒนามาจากโมเดล IRT ชนิดพหุมิติ (multidimensional) มีรูปแบบนันพาราเมทริกซึ่งแตกต่างจากวิธี IRT ทั่ว ๆ ไป เนื่องจากไม่ต้องใช้ค่าประมาณความสามารถแฝง สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้งเป็นรายข้อ (single-item) และเป็นกลุ่มข้อสอบ (item-bundle) ในแบบสอบเอกมิติ (unidimensional test) และแบบสอบพหุมิติ (multidimensional test) (Stout, Li

and Nandakumar, 1997) จุดเด่นของวิธีซิบเทสท์ก็คือ สามารถคำนวณได้ง่าย เสียค่าใช้จ่ายไม่แพง และไม่จำเป็นต้องใช้กลุ่มตัวอย่างที่มีขนาดใหญ่ นอกจากนี้ยังใช้สถิติทดสอบนัยสำคัญ (Narayanan and Swaminathan, 1996)

จากเหตุผลที่กล่าวมาข้างต้นจะเห็นว่าวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในกลุ่ม IRT ที่ยอมรับกันโดยทั่วไปว่ามีประสิทธิภาพสูงมี 4 วิธี คือ วิธีการวัดพื้นที่ของ Raju วิธีการวัดพื้นที่ของ Kim และ Cohen (1991) วิธีการทดสอบไค-สแควร์ของ Lord (1980) และวิธีการทดสอบอัตราส่วนไลค์ลิฮู้ด อย่างไรก็ตาม ถึงแม้ว่าทั้ง 4 วิธีจะมีประสิทธิภาพสูงแต่ก็มีข้อจำกัดตรงที่ต้องใช้กลุ่มตัวอย่างที่มีขนาดใหญ่และแบบสอบที่มีความยาวมาก ทำให้ต้องเสียค่าใช้จ่ายสูง นอกจากนี้ยังมีวิธีดำเนินการตรวจสอบที่ยุ่งยาก เช่น ข้อมูลต้องเป็นไปตามข้อตกลงเบื้องต้นของทฤษฎี IRT ค่าประมาณพารามิเตอร์ของข้อสอบระหว่างผู้สอบ 2 กลุ่มต้องปรับเทียบให้อยู่บนสเกลเดียวกัน การวิเคราะห์ข้อมูลค่อนข้างซับซ้อนต้องใช้โปรแกรมคอมพิวเตอร์ที่สามารถคำนวณข้อมูลทวนซ้ำหลายรอบทำให้เสียเวลามาก และการแปลผลก็ค่อนข้างยาก เป็นต้น ดังนั้นจึงทำให้นักวิธีวิทยาการวิจัยพยายามคิดค้น พัฒนา และปรับปรุงวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เพื่อให้ได้วิธีการตรวจสอบที่มีทั้งประสิทธิภาพและประสิทธิผล สามารถนำมาใช้แทนวิธีการตรวจสอบในกลุ่ม IRT ได้

วิธีการตรวจสอบที่นักวิธีวิทยาการวิจัยให้ความสนใจในปัจจุบันมี 3 วิธี คือ วิธีแมนเทล-แฮนส์เซล วิธีซิบเทสท์ และวิธีการถดถอยโลจิสติก โดยวิธีแมนเทล-แฮนส์เซลจัดเป็นวิธีมาตรฐานที่หน่วยงานการทดสอบทางการศึกษาของประเทศสหรัฐอเมริกา (Educational Testing Service; ETS) นำมาใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในโครงการทดสอบ (Zieky, 1993) ส่วนวิธีแมนเทล-แฮนส์เซลมีจุดเด่นคล้ายกับวิธีซิบเทสท์ เช่น เป็นวิธีนันพาราเมทริกเหมือนกัน ซึ่งถูกออกแบบเพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปโดยเฉพาะ สามารถคำนวณได้ง่ายไม่ยุ่งยากซับซ้อน การแปลผลไม่ยาก ทั้งยังใช้กลุ่มตัวอย่างขนาดเล็ก ทำให้เสียค่าใช้จ่ายไม่มาก นอกจากนี้ยังมีดัชนีวัดขนาดของการทำหน้าที่ต่างกันของข้อสอบและใช้สถิติทดสอบนัยสำคัญ สำหรับวิธีการถดถอยโลจิสติกมีจุดเด่นตรงที่ใช้โมเดลการถดถอยโลจิสติกวิเคราะห์ดัชนีการทำหน้าที่ต่างกันของข้อสอบ โมเดลนี้มีเทอมที่สามารถคำนวณปฏิสัมพันธ์ระหว่างสมาชิกกลุ่มผู้สอบและระดับความสามารถ จึงทำให้วิธีการถดถอยโลจิสติกสามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ทั้งแบบเอกรูปและแบบอเนกรูป นักวิธีวิทยาการวิจัยมักนำวิธีการตรวจสอบ 3 วิธีดังกล่าวมาศึกษาเปรียบเทียบเพื่อหาข้อสรุปว่าวิธีใดสามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ถูกต้องแม่นยำมากที่สุด ความถูกต้องแม่นยำสามารถพิจารณาได้จากค่าอำนาจ

การทดสอบหรืออัตราการตรวจสอบ (power of test or detection rates) อัตราความคลาดเคลื่อนประเภทที่ 1 (type I error rates) และอัตราความคลาดเคลื่อนประเภทที่ 2 (type II error rates) จากการศึกษางานวิจัยที่ผ่านมาพบว่ามีปัจจัยบางตัวมีผลต่อความถูกต้องแม่นยำในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เช่น ขนาดกลุ่มตัวอย่าง ความยาวของแบบสอบ ลักษณะของข้อสอบ การแจกแจงค่าความสามารถ สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบ และขนาดอิทธิพลของข้อสอบที่ทำหน้าที่ต่างกัน เป็นต้น ผลการศึกษาดังกล่าวสรุปได้ดังนี้

(1) ขนาดกลุ่มตัวอย่าง (sample size)

ผลการศึกษาของ Swaminathan และ Rogers (1990) พบว่า ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปและแบบอเนกรูป เมื่อเพิ่มขนาดกลุ่มตัวอย่างจะมีผลทำให้อำนาจการทดสอบของวิธีการถดถอยโลจิสติกมีค่าเพิ่มมากขึ้นเกือบทุกเงื่อนไข ส่วนอำนาจการทดสอบของวิธีแมนเทล-แฮนส์เซลจะมีค่าเพิ่มมากขึ้นเฉพาะกรณีแบบเอกรูปเท่านั้น ในขณะที่อัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีแมนเทล-แฮนส์เซลมีค่าลดลงเกือบทุกเงื่อนไข ต่อมาทั้ง 2 คนได้ศึกษาใหม่อีกครั้งหนึ่ง (Rogers and Swaminathan, 1993) ผลปรากฏว่า ขนาดกลุ่มตัวอย่างมีผลต่ออำนาจการทดสอบของวิธีการถดถอยโลจิสติกและวิธีแมนเทล-แฮนส์เซล เมื่อขนาดกลุ่มตัวอย่างเพิ่มมากขึ้นจะทำให้อำนาจการทดสอบของทั้ง 2 วิธีมีค่าเพิ่มมากขึ้น สำหรับ Narayanan และ Swaminathan (1994) ได้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป ผลการศึกษาพบว่า ขนาดกลุ่มตัวอย่างและอัตราส่วนของกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบมีผลต่ออำนาจการทดสอบของวิธีการตรวจสอบ กล่าวคือ เมื่อเพิ่มขนาดกลุ่มตัวอย่างจะทำให้อำนาจการทดสอบของวิธีแมนเทล-แฮนส์เซลและวิธีซิปเทสท์มีค่าเพิ่มมากขึ้น โดยเฉพาะเมื่อเพิ่มขนาดกลุ่มเปรียบเทียบจะมีผลทำให้อำนาจการทดสอบของทั้ง 2 วิธีมีค่าเพิ่มขึ้นมากกว่าเพิ่มขนาดกลุ่มอ้างอิง ต่อมาทั้ง 2 คนได้ศึกษาในกรณีแบบอเนกรูปอีกครั้งหนึ่ง (Narayanan and Swaminathan, 1996) พบว่า ขนาดกลุ่มตัวอย่างมีผลต่ออำนาจการทดสอบของวิธีโคร-ซิป วิธีแมนเทล-แฮนส์เซล และวิธีการถดถอยโลจิสติก ผลการศึกษาดังกล่าวสอดคล้องกับผลการศึกษาในครั้งแรก สำหรับอัตราความคลาดเคลื่อนประเภทที่ 1 พบว่า เมื่อใช้กลุ่มตัวอย่างขนาดเล็ก อัตราความคลาดเคลื่อนประเภทที่ 1 ของทั้ง 3 วิธีมีค่าต่ำกว่าใช้กลุ่มตัวอย่างขนาดใหญ่ ภายใต้เกือบทุกเงื่อนไขของการตรวจสอบ ต่อมาจิติมา วรรณศรี (2539) ได้เปรียบเทียบประสิทธิภาพของวิธีแมนเทล-แฮนส์เซลและวิธีซิปเทสท์ พบว่า เมื่อขนาดกลุ่มตัวอย่าง 200 และ 600 คน ทั้งสองวิธีสามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ถูกต้อง 50% แต่ถ้าขนาดกลุ่มตัวอย่าง 1,000 คน สามารถตรวจสอบได้ถูกต้อง 100%

*(2) **ความยาวของแบบสอบ** (test length)*

ผลการศึกษาของ Swaminathan และ Rogers (1990) พบว่า ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปและแบบอเนกรูป เมื่อใช้แบบสอบที่มีความยาวมากขึ้นจะมีผลทำให้อำนาจการทดสอบของวิธีแมนเทล-แฮนส์เซลและวิธีการถดถอยโลจิสติกมีค่ามากขึ้น ยกเว้นในกรณีแบบอเนกรูปของวิธีแมนเทล-แฮนส์เซล ต่อมาทั้ง 2 คนได้ทำการศึกษาใหม่อีกครั้งหนึ่ง (Rogers and Swaminathan, 1993) พบว่า ให้ผลขัดแย้งกับครั้งแรก กล่าวคือ ความยาวของแบบสอบไม่มีผลต่ออำนาจการทดสอบของวิธีแมนเทล-แฮนส์เซลและวิธีการถดถอยโลจิสติก ยกเว้นในกรณีแบบอเนกรูปของวิธีการถดถอยโลจิสติก ส่วนผลการศึกษาของกาญจนา วัธนสุนทร (2537) ปรากฏว่า ความยาวของแบบสอบไม่มีผลกระทบต่ออัตราการตรวจสอบของวิธีแมนเทล-แฮนส์เซลและวิธีซิปเทสท์ ซึ่งขัดแย้งกับผลการศึกษาของจิติมา วรรณศรี (2539) ที่พบว่า เมื่อใช้แบบสอบขนาด 60 ข้อ จะมีผลทำให้วิธีแมนเทล-แฮนส์เซลและวิธีซิปเทสท์มีประสิทธิภาพในการตรวจสอบดีที่สุด ส่วนผลการศึกษาของ Uttaro และ Millsap (1994) พบว่า เมื่อใช้แบบสอบที่มีความยาวมากขึ้นแล้วอัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีแมนเทล-แฮนส์เซลจะมีค่าลดลง ในขณะที่ผลการศึกษาของ Cohen และ Kim (1993) พบว่า เมื่อใช้แบบสอบที่มีความยาวมากขึ้นแล้วอัตราความคลาดเคลื่อนประเภทที่ 1 และ 2 ที่ตรวจสอบด้วยวิธีการวัดพื้นที่ของ Raju และวิธีการทดสอบไค-สแควร์ของ Lord จะมีค่ามากขึ้นด้วย

*(3) **ลักษณะของข้อสอบ** (type of item)*

ผลการศึกษาของ Rogers และ Swaminathan (1993) พบว่า ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปและแบบอเนกรูป ลักษณะของข้อสอบมีผลต่ออำนาจการทดสอบของวิธีการถดถอยโลจิสติกและวิธีแมนเทล-แฮนส์เซล ในกรณีแบบเอกรูปเมื่อข้อสอบมีค่า b ปานกลางและ a สูง จะทำให้อำนาจการทดสอบของทั้ง 2 วิธีมีค่าสูงสุด แต่ในกรณีแบบอเนกรูป พบว่า อำนาจการทดสอบของทั้ง 2 วิธีมีค่าสูงสุด เมื่อข้อสอบมีค่า b ปานกลางและ a สูงชนิดผสม (ข้อสอบมีค่า a และ b แตกต่างกันระหว่างผู้สอบ 2 กลุ่ม) สำหรับ Narayanan และ Swaminathan (1994) ได้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป พบว่า ภายใต้เกือบทุกเงื่อนไขลักษณะของข้อสอบที่มีค่า b ปานกลางและ a สูง จะทำให้อำนาจการทดสอบของวิธีแมนเทล-แฮนส์เซลและวิธีซิปเทสท์มีค่าสูงใกล้เคียงกัน ส่วนอัตราความคลาดเคลื่อนประเภทที่ 1 พบว่า ลักษณะของข้อสอบไม่มีผลต่อวิธีการตรวจสอบทั้งสอง ต่อมาทั้ง 2 คนได้ตรวจสอบในกรณีแบบอเนกรูป (Narayanan and Swaminathan, 1996) ผลการศึกษาพบว่า เมื่อข้อสอบมีค่า a เพิ่มขึ้นจะมีผลทำให้อำนาจการทดสอบของวิธีโคร-ซิปและวิธีการถดถอยโลจิสติกมีค่าเพิ่มขึ้น และเมื่อข้อสอบมีค่า b เพิ่มขึ้นหรือลดลงจะมีผลทำให้อำนาจการทดสอบของวิธีแมนเทล-

แฮนส์เซลมีค่าเพิ่มขึ้น ส่วนอัตราความคลาดเคลื่อนประเภทที่ 1 พบว่า เมื่อข้อสอบมีค่า a เพิ่มขึ้น จะมีผลทำให้อัตราความคลาดเคลื่อนประเภทที่ 1 ของทั้ง 3 วิธีมีค่าเพิ่มขึ้น โดยที่วิธีแมนเทล-แฮนส์เซลมีค่าต่ำสุด ต่อมาเกษร หว่างจิตร (2539) ได้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมนเทล-แฮนส์เซล ผลการศึกษาพบว่า ข้อสอบที่ตรวจพบว่าทำหน้าที่ต่างกันแบบเอกรูปและแบบอเนกรูปส่วนมากเป็นข้อสอบที่มีค่า a ค่อนข้างต่ำ ทั้งวิชาภาษาไทยและวิชาภาษาอังกฤษ ซึ่งไม่สอดคล้องกับผลการศึกษาของรัชรินทร์ มุคดา (2540) ที่พบว่าข้อสอบที่ตรวจพบว่าทำหน้าที่ต่างกันแบบอเนกรูปส่วนมากเป็นข้อสอบที่มีค่า a สูง ภายใต้ลักษณะของข้อสอบ a สูงกับ b ต่ำ, a สูงกับ b ปานกลาง, a สูงกับ b สูง และกลุ่มผู้สอบที่มีความสามารถสูง ปานกลาง และต่ำ

(4) การแจกแจงค่าความสามารถ (ability distribution)

ผลการศึกษาของ Rogers และ Swaminathan (1993) พบว่า ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปและแบบอเนกรูป ความแตกต่างของการแจกแจงคะแนนระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบไม่มีผลต่ออำนาจการทดสอบของวิธีแมนเทล-แฮนส์เซลและวิธีการถดถอยโลจิสติก สำหรับ Narayanan และ Swaminathan (1994) ได้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป ปรากฏว่า ความแตกต่างของการแจกแจงค่าความสามารถระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบไม่มีผลต่ออำนาจการทดสอบของวิธีซิปเทสท์ แต่มีผลต่ออำนาจการทดสอบของวิธีแมนเทล-แฮนส์เซล เมื่อความแตกต่างของการแจกแจงค่าความสามารถมีค่าเพิ่มขึ้นจะมีผลให้อำนาจการทดสอบมีค่าลดลง สำหรับอัตราความคลาดเคลื่อนประเภทที่ 1 พบว่า เมื่อความแตกต่างของการแจกแจงค่าความสามารถมีค่าเพิ่มขึ้นจะมีผลทำให้อัตราความคลาดเคลื่อนประเภทที่ 1 ของทั้ง 2 วิธีมีค่าเพิ่มขึ้น ต่อมา Narayanan และ Swaminathan (1996) ได้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป พบว่า ความแตกต่างของการแจกแจงค่าความสามารถไม่มีผลต่ออำนาจการทดสอบของวิธีแมนเทล-แฮนส์เซล แต่จะมีผลต่ออำนาจการทดสอบของวิธีโคร-ซิปและวิธีการถดถอยโลจิสติก กล่าวคือ เมื่อความแตกต่างของการแจกแจงค่าความสามารถมีค่าเพิ่มขึ้นจะมีผลให้อำนาจการทดสอบของทั้ง 2 วิธีดังกล่าวมีค่าลดลง ส่วนอัตราความคลาดเคลื่อนประเภทที่ 1 ปรากฏว่า ภายใต้เงื่อนไขความแตกต่างของการแจกแจงค่าความสามารถแบบไม่เท่ากัน อัตราความคลาดเคลื่อนประเภทที่ 1 ของทั้ง 3 วิธีมีค่าสูงกว่าภายใต้เงื่อนไขแบบเท่ากัน โดยที่วิธีโคร-ซิปมีค่าสูงสุดทั้งสองเงื่อนไข

(5) สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน (proportion of DIF items)

ผลการศึกษาของ Rogers และ Swaminathan (1993) พบว่า ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปและแบบอเนกรูป สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันไม่มีผลต่ออำนาจการทดสอบของวิธีแมนเทล-แฮนส์เซลและวิธีการถดถอยโลจิสติก ยกเว้นใน

กรณีแบบเอกรูปของวิธีการถดถอยโลจิสติก เมื่อสัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบมีจำนวนลดลงจะมีผลทำให้อำนาจการทดสอบของวิธีการถดถอยโลจิสติกมีค่าเพิ่มมากขึ้น ต่อมา Narayanan และ Swaminathan (1994) ได้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกรณีแบบเอกรูป ผลการศึกษาพบว่า เมื่อสัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบมีจำนวนลดลงจะมีผลทำให้อำนาจการทดสอบของวิธีซิปเทสท์และวิธีการถดถอยโลจิสติกมีค่าเพิ่มขึ้น ทั้งยังจะส่งผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 ของทั้ง 2 วิธีมีค่าเพิ่มขึ้นด้วย ต่อมา Narayanan และ Swaminathan (1996) ได้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบอีกครั้งหนึ่ง โดยตรวจสอบในกรณีแบบอเนกรูป ผลการศึกษาพบว่า เมื่อสัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบมีจำนวนลดลงจะมีผลทำให้อำนาจการทดสอบของวิธีการถดถอยโลจิสติกมีค่าเพิ่มขึ้น แต่จะไม่มีผลต่ออำนาจการทดสอบของวิธีแมนเทล-แฮนส์เซลและวิธีโคร-ซิป สำหรับอัตราความคลาดเคลื่อนประเภทที่ 1 พบว่า เมื่อสัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบมีจำนวนเพิ่มมากขึ้นจะส่งผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 ของทั้ง 3 วิธีจะมีค่าเพิ่มขึ้นด้วย โดยที่วิธีแมนเทล-แฮนส์เซลมีค่าต่ำสุด

(6) ***ขนาดอิทธิพลของข้อสอบที่ทำหน้าที่ต่างกัน*** (*DIF effect size*)

ผลการศึกษาของ Rogers และ Swaminathan (1993) พบว่า ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปและแบบอเนกรูป ขนาดอิทธิพลของข้อสอบที่ทำหน้าที่ต่างกัน (ค่าพื้นที่ระหว่างโค้งลักษณะข้อสอบของผู้สอบ 2 กลุ่ม) มีผลต่ออำนาจการทดสอบของวิธีการถดถอยโลจิสติกและวิธีแมนเทล-แฮนส์เซล เมื่อขนาดอิทธิพลของข้อสอบที่ทำหน้าที่ต่างกันมีค่าเพิ่มขึ้นจะส่งผลให้อำนาจการทดสอบของทั้ง 2 วิธีมีค่าเพิ่มมากขึ้น สำหรับ Narayanan และ Swaminathan (1994) ได้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป ปรากฏว่า ขนาดอิทธิพลของข้อสอบที่ทำหน้าที่ต่างกันมีผลต่ออำนาจการทดสอบของวิธีซิปเทสท์และวิธีแมนเทล-แฮนส์เซล เมื่อขนาดอิทธิพลของข้อสอบที่ทำหน้าที่ต่างกันมีค่าเพิ่มขึ้น อำนาจการทดสอบของทั้ง 2 วิธีมีค่าเพิ่มขึ้น ต่อมา Narayanan และ Swaminathan (1996) ได้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบอีกครั้งหนึ่ง โดยศึกษาในกรณีแบบอเนกรูป ปรากฏว่า ขนาดอิทธิพลของข้อสอบที่ทำหน้าที่ต่างกันมีผลต่ออำนาจการทดสอบของวิธีการถดถอยโลจิสติก วิธีแมนเทล-แฮนส์เซล และวิธีโคร-ซิป เมื่อขนาดอิทธิพลของข้อสอบที่ทำหน้าที่ต่างกันมีค่าเพิ่มขึ้น อำนาจการทดสอบของทั้ง 3 วิธีจะมีค่าเพิ่มขึ้น ซึ่งสอดคล้องกับผลการศึกษาในครั้งแรก

จากการศึกษางานวิจัยของ Swaminathan และ Rogers (1990) พบว่า วิธีแมนเทล-แฮนส์เซลและวิธีการถดถอยโลจิสติกมีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปได้เท่าเทียมกัน แต่การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป พบว่า วิธีการถดถอยโลจิสติกมีประสิทธิภาพสูงกว่าวิธีแมนเทล-แฮนส์เซล ภายใต้เกือบทุกเงื่อนไขของการตรวจสอบ ต่อมาทั้ง 2 คน ได้ทำการศึกษาอีกครั้งหนึ่ง (1993) ปรากฏว่า ให้ผลสอดคล้องกับครั้งแรก สำหรับ Narayanan และ Swaminathan (1994) ได้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป ผลการศึกษาพบว่า เมื่อการแจกแจงค่าความสามารถระหว่างกลุ่มผู้สอบมีค่าเท่ากัน วิธีซิปเทส์ทและวิธีแมนเทล-แฮนส์เซลมีประสิทธิภาพเท่าเทียมกัน แต่เมื่อการแจกแจงค่าความสามารถระหว่างกลุ่มผู้สอบมีค่าไม่เท่ากัน วิธีซิปเทสท์มีประสิทธิภาพสูงกว่าวิธีแมนเทล-แฮนส์เซล ต่อมาทั้ง 2 คน ได้ทำการศึกษาอีกครั้งหนึ่ง (1996) โดยตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป ผลปรากฏว่า วิธีการถดถอยโลจิสติกมีประสิทธิภาพสูงกว่าวิธีแมนเทล-แฮนส์เซล ภายใต้เกือบทุกเงื่อนไขของการตรวจสอบ จากผลดังกล่าว พบว่า วิธีแมนเทล-แฮนส์เซลไม่สามารถระบุข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูป โดยเฉพาะข้อสอบที่ทำหน้าที่ต่างกันแบบไม่มีทิศทาง (nondirectional DIF) ซึ่งเกิดจากโค้งลักษณะข้อสอบตัดกันตรงจุดกึ่งกลางของช่วงความสามารถ เนื่องจากวิธีแมนเทล-แฮนส์เซลใช้สถิติชนิดคิดเครื่องหมาย เมื่อขนาดของข้อสอบที่ทำหน้าที่ต่างกันเปลี่ยนทิศทางตรงจุดกึ่งกลางของช่วงความสามารถจะมีผลทำให้ความแตกต่างที่มีเครื่องหมายลบของช่วงคะแนนในส่วนหนึ่งหักล้างกับความแตกต่างที่มีเครื่องหมายบวกของช่วงคะแนนในอีกส่วนหนึ่ง ดังนั้นวิธีแมนเทล-แฮนส์เซลจึงไม่สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่มีทิศทาง

จากปัญหาดังกล่าว จึงทำให้ Mazor, Clauser และ Hambleton (1994) นำวิธีแมนเทล-แฮนส์เซลมาปรับปรุงขั้นตอนการวิเคราะห์ข้อมูล โดยแบ่งกลุ่มผู้สอบออกเป็นสองกลุ่มตามระดับคะแนนผลการสอบ คือ กลุ่มผู้สอบที่มีความสามารถสูงและกลุ่มผู้สอบที่มีความสามารถต่ำ แล้ววิเคราะห์เหมือนเดิมแต่แยกกันในแต่ละกลุ่มผู้สอบ ผลปรากฏว่า วิธีแมนเทล-แฮนส์เซลแบบใหม่มีอัตราการตรวจสอบสูงกว่าวิธีแมนเทล-แฮนส์เซลแบบเดิม ทั้งยังไม่ทำให้อัตราความคลาดเคลื่อนประเภทที่ 1 เพิ่มมากขึ้น ต่อมาเสรี ชัดแช้ม (2539) ได้ศึกษาทำนองเดียวกับ Mazor และคณะ (1994) ปรากฏว่า ได้ผลสอดคล้องกัน แสดงว่าการวิเคราะห์ด้วยวิธีแมนเทล-แฮนส์เซลถ้าแบ่งกลุ่มผู้สอบออกเป็นสองกลุ่มตามข้อเสนอของ Mazor และคณะ จะมีผลทำให้สามารถตรวจสอบข้อสอบแบบอเนกรูปได้ดีขึ้น ส่วนวิธีการถดถอยโลจิสติกไม่มีปัญหาดังกล่าว ทั้งนี้เนื่องจากในการวิเคราะห์จะใช้โมเดลการถดถอยโลจิสติก โมเดลดังกล่าวมีพารามิเตอร์สำหรับการคำนวณข้อสอบที่ทำหน้า

ที่ต่างกันแบบมีทิศทางเดียวและไม่มีทิศทาง ดังนั้นจึงสามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ดีทั้ง 2 แบบ อย่างไรก็ตาม ถึงแม้ว่าวิธีการถดถอยโลจิสติกจะมีข้อได้เปรียบดังกล่าว แต่ก็มีข้อเสียเปรียบในเรื่องความยุ่งยากในการประมาณค่าพารามิเตอร์ โดยจะต้องมีการคำนวณทวนซ้ำหลายรอบ (iterative) ทำให้เสียเวลามาก ทั้งยังเสียค่าใช้จ่ายสูง สำหรับวิธีซิปเทสท์เป็นวิธีการตรวจสอบที่ออกแบบมาเพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบมีทิศทางเดียว จึงไม่สามารถระบุข้อสอบที่ทำหน้าที่ต่างกันแบบไม่มีทิศทาง สาเหตุดังกล่าวอาจเนื่องมาจากวิธีซิปเทสท์ใช้สถิติชนิดคิดเครื่องหมาย ซึ่งน่าจะมีปัญหาในการวิเคราะห์ทำนองเดียวกับวิธีแมนเทล-แฮนส์เซล ดังนั้นถ้าปรับปรุงขั้นตอนในการวิเคราะห์ข้อมูลโดยแบ่งกลุ่มผู้สอบออกเป็นสองกลุ่ม คือ กลุ่มผู้สอบที่มีความสามารถสูงกับกลุ่มผู้สอบที่มีความสามารถต่ำ แล้ววิเคราะห์เหมือนเดิมแต่แยกกันคนละกลุ่ม ผู้วิจัยคาดว่าน่าจะสามารถระบุข้อสอบที่ทำหน้าที่ต่างกันแบบไม่มีทิศทางได้ ทั้งนี้เพราะว่าเมื่อแบ่งกลุ่มผู้สอบแล้ววิเคราะห์จะมีผลทำให้โค้งลักษณะข้อสอบที่ตัดกันตรงจุดกึ่งกลางของช่วงความสามารถเปลี่ยนเป็นตัดกันตรงปลายสุดของช่วงความสามารถสูงหรือต่ำ ดังนั้นความแตกต่างของพื้นที่ระหว่างเครื่องหมายบวกและลบจึงไม่หักล้างกัน

จากประเด็นปัญหาของวิธีซิปเทสท์ ผู้วิจัยจึงมีความสนใจที่จะเปรียบเทียบความถูกต้องแม่นยำในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูประหว่างวิธีซิปเทสท์ปรับใหม่ (ปรับปรุงขั้นตอนการวิเคราะห์ดังที่กล่าวมาข้างต้น) วิธีซิปเทสท์แบบเดิม วิธีแมนเทล-แฮนส์เซล ที่พัฒนาโดย Mazor และคณะ และวิธีการถดถอยโลจิสติก โดยจะพิจารณาความถูกต้องแม่นยำจากอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ระหว่างวิธีการตรวจสอบ ภายใต้การจำลองข้อมูลด้วยการจัดกระทำปัจจัยที่แปรเปลี่ยน 4 ตัว คือ ***(1) ลักษณะของข้อสอบที่มีค่า a และ b แตกต่างกัน 9 ลักษณะ*** ประกอบด้วย a ต่ำกับ b ต่ำ, a ต่ำกับ b ปานกลาง, a ต่ำกับ b สูง, a ปานกลางกับ b ต่ำ, a ปานกลางกับ b ปานกลาง, a ปานกลางกับ b สูง, a สูง กับ b ต่ำ, a สูงกับ b ปานกลาง และ a สูงกับ b สูง ***(2) ความยาวของแบบสอบแตกต่างกัน 2 ระดับ*** ประกอบด้วย 30 ข้อ และ 60 ข้อ ***(3) สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันมีจำนวนแตกต่างกัน 3 ระดับ*** ประกอบด้วย 5%, 10% และ 20% ***(4) ขนาดกลุ่มตัวอย่างแตกต่างกัน 6 ระดับ*** ประกอบด้วย จำนวนผู้สอบกลุ่มอ้างอิงต่อจำนวนผู้สอบกลุ่มเปรียบเทียบเท่ากับ 250 คนต่อ 250 คน, 500 คนต่อ 250 คน, 500 คนต่อ 500 คน, 1000 คนต่อ 250 คน, 1000 คนต่อ 500 คน และ 1000 คนต่อ 1000 คน ดังนั้นในการศึกษาครั้งนี้จะต้องจัดกระทำข้อมูลทั้งหมด 324 เงื่อนไข ($9 \times 2 \times 3 \times 6$) สำหรับเกณฑ์ที่ใช้เปรียบเทียบผลการตรวจสอบของวิธีการตรวจสอบ 4 วิธีดังกล่าว ผู้วิจัยเลือกวิธีการวัดพื้นที่ในช่วงเปิดของ Raju (1990) กรณีแบบอเนกรูป ($a_F \neq a_R$) ภายใต้โมเดล

โลจิสติกแบบ 3 พารามิเตอร์ชนิดกำหนดค่า c คงที่ (3PLM-c) ทั้งนี้เนื่องจากถ้าไม่กำหนดค่า c คงที่แล้วผลการคำนวณจะไม่มีที่สิ้นสุด (infinite) ส่วนการตัดสินการทำหน้าที่ต่างกันของข้อสอบจะพิจารณาจากดัชนีพื้นที่ชนิดไม่คิดเครื่องหมายซึ่งมีนัยสำคัญกับสถิติ Z ที่ระดับ .05 ผู้วิจัยใช้ดัชนีพื้นที่ชนิดไม่คิดเครื่องหมายเพราะว่าในกรณีที่ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปประเภทโค้งลักษณะข้อสอบระหว่างผู้สอบ 2 กลุ่มตัดกันในลักษณะสมมาตร (symmetrically) ถ้าใช้ดัชนีพื้นที่ชนิดคิดเครื่องหมายจะมีผลทำให้ความแตกต่างของพื้นที่ระหว่างเครื่องหมายบวกและลบหักล้างกัน จึงไม่สามารถตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันในลักษณะดังกล่าวได้ แต่ถ้าใช้ดัชนีพื้นที่ชนิดไม่คิดเครื่องหมายจะไม่มีปัญหาดังกล่าว ดังนั้นจึงสามารถตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันได้มากกว่า (Feinstein, 1995)

วัตถุประสงค์ในการวิจัย

1. เพื่อเปรียบเทียบอำนาจการทดสอบในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูประหว่างวิธีซิบเทสท์ปรับใหม่ วิธีซิบเทสท์ วิธีแมนเทล-แฮนส์เซล และวิธีการถดถอยโลจิสติก ภายใต้เงื่อนไขของปัจจัยที่แตกต่างกันทางด้าน

   1.1 ลักษณะของข้อสอบ

   1.2 ความยาวของแบบสอบ

   1.3 สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน

   1.4 ขนาดกลุ่มตัวอย่าง

2. เพื่อเปรียบเทียบอัตราความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูประหว่างวิธีซิบเทสท์ปรับใหม่ วิธีซิบเทสท์ วิธีแมนเทล-แฮนส์เซล และวิธีการถดถอยโลจิสติก ภายใต้เงื่อนไขของปัจจัยที่แตกต่างกันทางด้าน

   2.1 ลักษณะของข้อสอบ

   2.2 ความยาวของแบบสอบ

   2.3 สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน

   2.4 ขนาดกลุ่มตัวอย่าง

สมมติฐานการวิจัย

จากการศึกษาของ Swaminathan และ Rogers (1990) Rogers และ Swaminathan (1993) Narayanan และ Swaminathan (1996) พบว่า ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป อำนาจการทดสอบของวิธีการถดถอยโลจิสติกมีค่าสูงกว่าอำนาจการทดสอบของวิธีแมนเทล-แฮนส์เซล ภายใต้เกือบทุกเงื่อนไขของปัจจัยลักษณะของข้อสอบ ความยาวของแบบสอบ สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบและขนาดกลุ่มตัวอย่าง ทั้งนี้อาจเนื่องมาจากข้อสอบที่หน้าที่ต่างกันแบบอเนกรูปมี 2 ลักษณะ คือ ***ข้อสอบทำหน้าที่ต่างกันแบบไม่มีทิศทาง*** (*nondirectional DIF*) และ***ข้อสอบทำหน้าที่ต่างกันแบบมีทิศทางเดียว*** (*unidirectional DIF*) ข้อสอบที่ทำหน้าที่ต่างกันแบบไม่มีทิศทางเกิดขึ้นเมื่อมีปฏิสัมพันธ์ไม่เป็นลำดับ (disordinal interaction) ระหว่างระดับความสามารถและการเป็นสมาชิกของกลุ่ม (group membership) ซึ่งในทฤษฎีการตอบสนองข้อสอบ (item response theory) สามารถพิจารณาได้จากโค้งลักษณะข้อสอบ (item characteristic curves) ระหว่างกลุ่มผู้สอบ 2 กลุ่มตัดกันตรงจุดกึ่งกลางของช่วงความสามารถ ข้อสอบลักษณะดังกล่าวจะมีค่าความยากปานกลางซึ่งไม่สามารถตรวจสอบด้วยวิธีแมนเทล-แฮนส์เซลของ Holland and Thayer (1988) เพราะว่าวิธีแมนเทล-แฮนส์เซลใช้สถิติชนิดคิดเครื่องหมายซึ่งมีความไวต่อทิศทางของข้อสอบที่ทำหน้าที่ต่างกัน เมื่อขนาดของข้อสอบที่ทำหน้าที่ต่างกันเปลี่ยนทิศทางตรงจุดกึ่งกลางของช่วงความสามารถจะมีผลทำให้ความแตกต่างที่มีเครื่องหมายลบของช่วงคะแนนส่วนหนึ่งหักล้างกับความแตกต่างที่มีเครื่องหมายบวกของช่วงคะแนนอีกส่วนหนึ่ง ข้อสอบลักษณะดังกล่าวจึงไม่สามารถตรวจสอบด้วยวิธีแมนเทล-แฮนส์เซล สำหรับข้อสอบที่ทำหน้าที่ต่างกันแบบมีทิศทางเดียวเกิดขึ้นเมื่อมีปฏิสัมพันธ์เป็นลำดับ (ordinal interaction) ระหว่างระดับความสามารถและการเป็นสมาชิกของกลุ่ม ซึ่งจะทำให้โค้งลักษณะข้อสอบระหว่างกลุ่มผู้สอบ 2 กลุ่มตัดกันตรงปลายสุดของช่วงความสามารถต่ำหรือสูง ข้อสอบลักษณะดังกล่าวเป็นข้อสอบที่ยากหรือง่าย ซึ่งไม่มีปัญหาเมื่อตรวจสอบด้วยวิธีแมนเทล-แฮนส์เซลแบบเดิม

จากข้อบกพร่องดังกล่าวจึงทำให้ Mazor และคณะ (1994) ปรับปรุงขั้นตอนการวิเคราะห์ของวิธีแมนเทล-แฮนส์เซล โดยแบ่งกลุ่มผู้สอบออกเป็น 2 กลุ่มตามระดับคะแนนรวม คือ กลุ่มที่มีคะแนนสูงกับกลุ่มที่มีคะแนนต่ำ แล้ววิเคราะห์แยกกันในแต่ละกลุ่ม ผลการวิเคราะห์ ปรากฏว่าวิธีแมนเทล-แฮนส์เซลที่พัฒนาขึ้นใหม่มีอำนาจการทดสอบสูงกว่าวิธีแมนเทล-แฮนส์เซลแบบเดิม ต่อมาเสรี ชัดแช้ม (2539) ได้ศึกษาทำนองเดียวกับ Mazor และคณะ (1994) ปรากฏว่า ได้ผล

สอดคล้องกัน นอกจากนี้ รัชรินทร์ มุคคา (2540) ได้นำวิธีแมนเทล-แฮนส์เซลไปเปรียบเทียบกับวิธีการถดถอยโลจิสติก ผลปรากฏว่า ทั้ง 2 วิธีมีประสิทธิภาพใกล้เคียงกัน สำหรับวิธีซิปเทสท์เป็นวิธีที่ออกแบบเพื่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบมีทิศทางเดียวโดยเฉพาะ (Shealy and Stout, 1993) จึงไม่สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบไม่มีทิศทาง แต่ถ้าจะนำวิธีซิปเทสท์มาตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปที่ไม่มีทิศทางโดยการแบ่งกลุ่มผู้สอบออกเป็น 2 กลุ่ม เช่นเดียวกับวิธีแมนเทล-แฮนส์เซลของ Mazor และคณะ (1994) ผู้วิจัยคาดว่าน่าจะสามารถระบุข้อสอบลักษณะดังกล่าวได้ ทั้งนี้เนื่องจากวิธีซิปเทสท์ใช้สถิติชนิดคิดเครื่องหมายเช่นเดียวกับวิธีแมนเทล-แฮนส์เซล ดังนั้นจึงน่าจะมีปัญหาในทำนองเดียวกัน นั่นคือ วิธีซิปเทสท์ปรับใหม่น่าจะสามารถระบุข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปได้ดีกว่าวิธีซิปเทสท์แบบเดิม สำหรับวิธีการถดถอยโลจิสติกไม่น่าจะมีปัญหาในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปทั้ง 2 ลักษณะ เนื่องจากวิธีดังกล่าวมีพื้นฐานมาจากโมเดลการถดถอยโลจิสติก ซึ่งโมเดลนี้มีเทอมที่สามารถคำนวณปฏิสัมพันธ์ระหว่างระดับความสามารถและการเป็นสมาชิกของกลุ่ม ดังนั้นวิธีการถดถอยโลจิสติกน่าจะสามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ทั้งแบบมีทิศทางเดียวและแบบไม่มีทิศทาง

สำหรับผลการศึกษาเกี่ยวกับอัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ มีข้อสังเกต 2 ประการ คือ ***ประการที่ 1*** เมื่อวิธีการตรวจสอบมีอำนาจการทดสอบสูงมักจะมีอัตราความคลาดเคลื่อนประเภทที่ 1 สูงด้วย แต่เมื่อวิธีการตรวจสอบมีอำนาจการทดสอบต่ำมักจะมีอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำด้วย แสดงว่า ค่าทั้งสองแปรผันตามกัน ซึ่งเป็นไปตามทฤษฎีของค่าอำนาจการทดสอบและความคลาดเคลื่อนประเภทที่ 1 ดังเช่น ผลการศึกษาของ Swaminathan และ Rogers (1990) Narayanan และ Swaminathan (1996) ที่พบว่า เมื่ออำนาจการทดสอบของวิธีโคร-ซิปและวิธีการถดถอยโลจิสติกมีค่าสูงแล้วอัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีดังกล่าวจะมีค่าสูงด้วย แต่เมื่ออำนาจการทดสอบของวิธีแมนเทล-แฮนส์เซลมีค่าต่ำแล้วอัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีดังกล่าวจะมีค่าต่ำด้วยภายใต้เกือบทุกเงื่อนไขของการตรวจสอบ และ***ประการที่ 2*** ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ระดับ α .05 อัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีแมนเทล-แฮนส์เซล วิธีซิปเทสท์ และวิธีการถดถอยโลจิสติกมีค่าเฉลี่ยต่ำกว่า 10% เช่น ผลการศึกษาของ Narayanan และ Swaminathan (1994) ที่ตรวจสอบในกรณีแบบเอกรูป พบว่า อัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีแมนเทล-แฮนส์เซลและวิธีซิปเทสท์มีค่าเฉลี่ย 4.75% และ 6.65% ตามลำดับ และผลการศึกษาของ Narayanan และ Swaminathan (1996) ที่ตรวจสอบในกรณีแบบอเนกรูป

พบว่า อัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีแมนเทล-แฮนส์เซล วิธีโคร-ซิป และวิธีการ-ถดถอยโลจิสติกมีค่าเฉลี่ย 4.81%, 8.45% และ 8.15% ตามลำดับ

ในการศึกษาครั้งนี้ ผู้วิจัยนำวิธีการตรวจสอบ 4 วิธี คือ วิธีซิปเทสท์ปรับใหม่ วิธีซิปเทสท์ วิธีแมนเทล-แฮนส์เซลของ Mazor และคณะ (1994) และวิธีการถดถอยโลจิสติก มาเปรียบเทียบความถูกต้องแม่นยำในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป ความถูกต้องแม่นยำดังกล่าวพิจารณาจากอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ภายใต้การจัดกระทำปัจจัยที่แปรเปลี่ยน 4 ตัว คือ ลักษณะของข้อสอบ ความยาวของแบบสอบ สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน และขนาดกลุ่มตัวอย่าง โดยผู้วิจัยตั้งสมมติฐาน ดังนี้

1. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปภายใต้เงื่อนไขของปัจจัยที่แตกต่างกันทางด้านลักษณะของข้อสอบ ความยาวของแบบสอบ สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน และขนาดกลุ่มตัวอย่าง อำนาจการทดสอบของวิธีซิปเทสท์ปรับใหม่ วิธีแมนเทล-แฮนส์เซล และวิธีการถดถอยโลจิสติกน่าจะมีค่าสูงกว่าวิธีซิปเทสท์

2. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปภายใต้เงื่อนไขของปัจจัยที่แตกต่างกันทางด้านลักษณะของข้อสอบ ความยาวของแบบสอบ สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน และขนาดกลุ่มตัวอย่าง อัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีซิปเทสท์ปรับใหม่ วิธีแมนเทล-แฮนส์เซล และวิธีการถดถอยโลจิสติกน่าจะมีค่าสูงกว่าวิธีซิปเทสท์ เมื่อเปรียบเทียบอัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีการตรวจสอบ 4 วิธีดังกล่าวกับเกณฑ์ที่ระดับ 10% อัตราความคลาดเคลื่อนประเภทที่ 1 ของทั้ง 4 วิธีน่าจะมีค่าอยู่ภายในเกณฑ์ที่กำหนด

ขอบเขตของการวิจัย

1. การวิจัยครั้งนี้ศึกษาในสถานการณ์จำลอง โดยจำลองข้อมูลด้วยโปรแกรม IRTDATA (Johanson, 1992) ภายใต้ทฤษฎีการตอบสนองข้อสอบ โมเดลโลจิสติกแบบ 3 พารามิเตอร์ชนิดกำหนดค่า c คงที่ (3PLM-c) โดยกำหนดค่า c เท่ากับ 0.20 แล้วจัดกระทำข้อมูลทั้งหมด 324 เงื่อนไข (9 ลักษณะของข้อสอบ × 2 ความยาวของแบบสอบ × 3 สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน × 6 ขนาดกลุ่มตัวอย่าง)

2. ในการจำลองข้อมูล Swaminathan และ Gifford (1985 cited in Lautenschlager and Park, 1988) ได้เสนอแนะให้ใช้ค่าอำนาจจำแนก (a) ในช่วงตั้งแต่ .60 ถึง 2 ค่าความยาก (b) ในช่วงตั้งแต่ –2 ถึง 2 และค่าการเดา (c) มีค่าคงที่เท่ากับ .20 ในการศึกษาครั้งนี้ผู้วิจัยใช้ค่า

พารามิเตอร์ a และ b ในช่วงดังกล่าวแล้วแบ่งค่า a และ b ออกเป็น 3 ช่วง คือ ต่ำ ปานกลาง และสูง โดยใช้สเกลในแต่ละช่วงเท่ากัน ผลการแบ่งช่วงของค่า a ปรากฏว่า ช่วงต่ำมีค่าตั้งแต่ 0.60 ถึง 1.06 ($\bar{a}$ = 0.83) ช่วงปานกลางมีค่าระหว่าง 1.06 และ 1.54 ($\bar{a}$ = 1.30) ช่วงสูงมีค่าตั้งแต่ 1.54 ถึง 2.00 ($\bar{a}$ = 1.77) ผลการแบ่งช่วงของค่า b ปรากฏว่า ช่วงต่ำมีค่าตั้งแต่ –2.00 ถึง –0.67 ($\bar{b}$ = –1.34) ช่วงปานกลางมีค่าระหว่าง –0.67 และ 0.67 ($\bar{b}$ = 0.00) และช่วงสูงมีค่าตั้งแต่ 0.67 ถึง 2.00 ($\bar{b}$ = 1.34) สำหรับพารามิเตอร์ c กำหนดให้มีค่าคงที่เท่ากับ 0.20 ส่วนพารามิเตอร์ของผู้สอบ (θ) กำหนดให้มีค่าตั้งแต่ –3.00 ถึง 3.00 ($\bar{\theta}$ = 0.00)

3. ใช้วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ 4 วิธี คือ

   3.1 วิธีซิปเทสท์ปรับใหม่

   3.2 วิธีซิปเทสท์

   3.3 วิธีแมนเทล-แฮนส์เซล

   3.4 วิธีการถดถอยโลจิสติก

4. ตัวแปรที่ใช้ในการวิจัยประกอบด้วย

4.1 ตัวแปรอิสระ มี 4 ตัว คือ ลักษณะของข้อสอบ ความยาวของแบบสอบ สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน และขนาดกลุ่มตัวอย่าง ในแต่ละตัวแปรดังกล่าวยังแบ่งออกเป็นระดับต่าง ๆ ดังนี้

4.1.1 ลักษณะของข้อสอบมี 9 ลักษณะ

(1) ค่า a ต่ำกับ b ต่ำ ($\bar{a}$ = 0.83 กับ $\bar{b}$ = –1.34)

(2) ค่า a ต่ำกับ b ปานกลาง ($\bar{a}$ = 0.83 กับ $\bar{b}$ = 0.00)

(3) ค่า a ต่ำกับ b สูง ($\bar{a}$ = 0.83 กับ $\bar{b}$ = 1.34)

(4) ค่า a ปานกลางกับ b ต่ำ ($\bar{a}$ = 1.30 กับ $\bar{b}$ = –1.34)

(5) ค่า a ปานกลางกับ b ปานกลาง ($\bar{a}$ = 1.30 กับ $\bar{b}$ = 0.00)

(6) ค่า a ปานกลางกับ b สูง ($\bar{a}$ = 1.30 กับ $\bar{b}$ = 1.34)

(7) ค่า a สูงกับ b ต่ำ ($\bar{a}$ = 1.77 กับ $\bar{b}$ = –1.34)

(8) ค่า a สูงกับ b ปานกลาง ($\bar{a}$ = 1.77 กับ $\bar{b}$ = 0.00)

(9) ค่า a สูงกับ b สูง ($\bar{a}$ = 1.77 กับ $\bar{b}$ = 1.34)

4.1.2 ความยาวของแบบสอบมี 2 ระดับ

(1) แบบสอบที่มีจำนวน 30 ข้อ

(2) แบบสอบที่มีจำนวน 60 ข้อ

4.1.3 สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันมี 3 ระดับ

(1) มีข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบจำนวน 5%

(2) มีข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบจำนวน 10%

(3) มีข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบจำนวน 20%

4.1.4 ขนาดกลุ่มตัวอย่างมี 6 ระดับ

(1) กลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบเท่ากับ 250 คนต่อ 250 คน

(2) กลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบเท่ากับ 500 คนต่อ 250 คน

(3) กลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบเท่ากับ 500 คนต่อ 500 คน

(4) กลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบเท่ากับ 1,000 คนต่อ 250 คน

(5) กลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบเท่ากับ 1,000 คนต่อ 500 คน

(6) กลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบเท่ากับ 1,000 คนต่อ 1,000 คน

4.2 ตัวแปรตาม มี 2 ตัว ดังนี้

4.2.1 อำนาจการทดสอบ

4.2.2 อัตราความคลาดเคลื่อนประเภทที่ 1

5. ในการศึกษาครั้งนี้ ผู้วิจัยใช้วิธีการวัดพื้นที่ชนิดไม่คิดเครื่องหมายของ Raju (1990) กรณีแบบอเนกรูปภายใต้โมเดลโลจิสติกแบบ 3 พารามิเตอร์ชนิดกำหนดค่า c คงที่ (3PLM-c) เป็นวิธีเกณฑ์ เพื่อใช้เป็นเกณฑ์สำหรับการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป ซึ่งตรวจสอบด้วยวิธีที่ศึกษา 4 วิธี ได้แก่ วิธีซิปเทสท์ปรับใหม่ วิธีซิปเทสท์ วิธีแมนเทล-แฮนส์เซล และวิธีการถดถอยโลจิสติก กล่าวคือ ถ้าวิธีที่ศึกษาระบุข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปได้ตรงกับข้อสอบที่ถูกระบุด้วยวิธีการวัดพื้นที่ของ Raju แสดงว่า วิธีที่ศึกษาตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปได้ถูกต้อง แต่ถ้าวิธีที่ศึกษาระบุข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปได้ไม่ตรงกับข้อสอบที่ถูกระบุด้วยวิธีการวัดพื้นที่ของ Raju แสดงว่า วิธีที่ศึกษาตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปไม่ถูกต้อง

คำจำกัดความที่ใช้ในการวิจัย

การทำหน้าที่ต่างกันของข้อสอบ หมายถึง โอกาสของการตอบข้อสอบได้ถูกต้องไม่เท่ากัน เมื่อผู้สอบที่มีความสามารถในระดับเดียวกัน แต่มาจากกลุ่มผู้สอบที่แตกต่างกัน

การทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป หมายถึง ความแตกต่างระหว่างโอกาสของการตอบข้อสอบได้ถูกต้องสำหรับผู้สอบ 2 กลุ่มมีค่าไม่คงที่ตลอดช่วงความสามารถ ซึ่งในทฤษฎี IRT สามารถพิจารณาได้จากโค้งลักษณะข้อสอบระหว่างผู้สอบ 2 กลุ่มไม่ขนานกัน

ลักษณะของข้อสอบ หมายถึง ข้อสอบในแบบสอบที่มีค่าพารามิเตอร์อำนาจจำแนก (a) และความยาก (b) แตกต่างกัน ส่วนพารามิเตอร์การเดา (c) มีค่าคงที่ ในการศึกษาครั้งนี้ผู้วิจัยจัดกระทำค่าอำนาจจำแนกและความยาก 9 ลักษณะ คือ a ต่ำกับ b ต่ำ, a ต่ำกับ b ปานกลาง, a ต่ำกับ b สูง, a ปานกลางกับ b ต่ำ, a ปานกลางกับ b ปานกลาง, a ปานกลางกับ b สูง, a สูงกับ b ต่ำ, a สูงกับ b ปานกลาง และ a สูงกับ b สูง ข้อสอบทั้ง 9 ลักษณะได้มาจากการจำลองข้อมูล 9 เมทริกซ์ ในแต่ละเมทริกซ์มีจำนวนข้อสอบ 90 ข้อ และจำนวนผู้สอบ 2,000 คน

กลุ่มข้อสอบที่มีค่าอำนาจจำแนกต่ำ หมายถึง กลุ่มข้อสอบที่มีค่าอำนาจจำแนกเฉลี่ย 0.83 โดยมีค่าอำนาจจำแนกตั้งแต่ 0.60 ถึง 1.06

กลุ่มข้อสอบที่มีค่าอำนาจจำแนกปานกลาง หมายถึง กลุ่มข้อสอบที่มีค่าอำนาจจำแนกเฉลี่ย 1.30 โดยมีค่าอำนาจจำแนกระหว่าง 1.06 และ 1.54

กลุ่มข้อสอบที่มีค่าอำนาจจำแนกสูง หมายถึง กลุ่มข้อสอบที่มีค่าอำนาจจำแนกเฉลี่ย 1.77 โดยมีค่าอำนาจจำแนกตั้งแต่ 1.54 ถึง 2.00

กลุ่มข้อสอบที่มีค่าความยากต่ำ หมายถึง กลุ่มข้อสอบที่มีค่าความยากเฉลี่ย –1.34 โดยมีค่าความยากตั้งแต่ –2.00 ถึง –0.67

กลุ่มข้อสอบที่มีค่าความยากปานกลาง หมายถึง กลุ่มข้อสอบที่มีค่าความยากเฉลี่ย 0.00 โดยมีค่าความยากระหว่าง –0.67 และ 0.67

กลุ่มข้อสอบที่มีค่าความยากสูง หมายถึง กลุ่มข้อสอบที่มีค่าความยากเฉลี่ย 1.34 โดยมีค่าความยากตั้งแต่ 0.67 ถึง 2.00

ความยาวของแบบสอบ หมายถึง จำนวนข้อสอบในแบบสอบ ในการศึกษาครั้งนี้ผู้วิจัยจัดกระทำความยาวของแบบสอบ 2 ระดับ คือ 30 ข้อ และ 60 ข้อ โดยสุ่มมาจากข้อสอบ 90 ข้อ ที่ได้จากการจำลองข้อมูล ดังนั้นข้อมูลทั้ง 9 เมทริกซ์จะจัดกระทำเป็นแบบสอบ 30 ข้อ และ 60 ข้อ ทั้งหมด 18 ฉบับ

สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน หมายถึง ร้อยละของจำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบ ในการศึกษาครั้งนี้ผู้วิจัยจัดกระทำสัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน 3 ระดับคือ 5%, 10% และ 20% โดยสุ่มจากข้อสอบที่ตรวจสอบแล้วว่าทำหน้าที่ต่างกันแบบอเนกรูปด้วยวิธีการวัดพื้นที่ของ Raju ซึ่งเป็นวิธีเกณฑ์ ดังนั้นสัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน 5%, 10% และ 20% ในแบบสอบที่มีความยาว 30 ข้อ จะมีข้อสอบที่ทำหน้าที่ต่างกันกับข้อสอบที่ทำหน้าที่ไม่ต่างกันจำนวน 2 ข้อกับ 28 ข้อ, 3 ข้อกับ 27 ข้อ และ 6 ข้อกับ 24 ข้อ ตามลำดับ และสัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน 5%, 10% และ 20% ในแบบสอบที่มีความยาว 60 ข้อ จะมีข้อสอบที่ทำหน้าที่ต่างกันกับข้อสอบที่ทำหน้าที่ไม่ต่างกันจำนวน 3 ข้อกับ 57 ข้อ, 6 ข้อกับ 54 ข้อ และ 12 ข้อกับ 48 ข้อ ตามลำดับ

ขนาดกลุ่มตัวอย่าง หมายถึง จำนวนผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่ใช้ในการวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกัน ในการศึกษาครั้งนี้ผู้วิจัยจัดกระทำขนาดกลุ่มตัวอย่าง 6 ระดับ คือ จำนวนผู้สอบกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบเท่ากับ 250 คน : 250 คน, 500 คน : 250 คน, 500 คน : 500 คน, 1000 คน : 250 คน, 1000 คน : 500 คน และ 1000 คน : 1000 คน ขนาดกลุ่มตัวอย่างทั้ง 6 ระดับสุ่มมาจากจำนวนผู้สอบ 2,000 คน ที่ได้มาจากการจำลองข้อมูล ดังนั้นข้อมูลทั้ง 9 เมทริกซ์จะจัดกระทำเป็นขนาดกลุ่มตัวอย่างทั้งหมด 54 กลุ่ม

กลุ่มอ้างอิง หมายถึง กลุ่มผู้สอบที่คาดว่าจะได้เปรียบในการตอบข้อสอบเมื่อข้อสอบทำหน้าที่ต่างกัน โดยมีโอกาสของการตอบข้อสอบได้ถูกต้องมากกว่าผู้สอบกลุ่มเปรียบเทียบ

กลุ่มเปรียบเทียบ หมายถึง กลุ่มผู้สอบที่คาดว่าจะเสียเปรียบในการตอบข้อสอบเมื่อข้อสอบทำหน้าที่ต่างกัน โดยมีโอกาสของการตอบข้อสอบได้ถูกต้องน้อยกว่าผู้สอบกลุ่มอ้างอิง

วิธีการวัดพื้นที่ของ Raju หมายถึง วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในกลุ่มวิธี IRT ที่พัฒนาโดย Raju (1990) ซึ่งจะคำนวณดัชนีการทำหน้าที่ต่างกันของข้อสอบจากพื้นที่ระหว่างโค้งลักษณะข้อสอบจากผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบโดยใช้การอินทิเกรทพื้นที่ในช่วงเปิด แล้วทดสอบนัยสำคัญด้วยสถิติ Z ในการศึกษาครั้งนี้ผู้วิจัยตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกรณีแบบอเนกรูป ($\hat{a}_{iR} \neq \hat{a}_{iF}$) โดยคำนวณพื้นที่ซนิดไม่คิดเครื่องหมายภายใต้โมเดลโลจิสติกแบบ 3 พารามิเตอร์ซนิดกำหนดค่า c คงที่ (3PLM-c) ด้วยโปรแกรม IRTDIF version 1.0 (Kim and Cohen, 1992b)

วิธีซิบเทสท์ หมายถึง วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในกลุ่มวิธี IRT ที่พัฒนาโดย Shealy และ Stout (1993) ซึ่งจะคำนวณดัชนีการทำหน้าที่ต่างกันของข้อสอบจากค่าเฉลี่ยสัดส่วนการตอบข้อสอบถูกระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบในชุดแบบสอบ

ที่ใช้ในการศึกษา (studied subtest) โดยใช้คะแนนการจับคู่เปรียบเทียบระหว่างกลุ่มผู้สอบจากชุดแบบสอบที่มีความตรง (valid subtest) แล้วทดสอบนัยสำคัญด้วยสถิติ Z ในการศึกษาครั้งนี้ผู้วิจัยวิเคราะห์โดยใช้โปรแกรม SIBTEST version 1.1 (Stout and Roussos, 1992)

วิธีซิปเทสท์ปรับใหม่ หมายถึง วิธีซิปเทสท์ที่ผู้วิจัยนำมาปรับปรุงขั้นตอนการวิเคราะห์โดยแบ่งกลุ่มผู้สอบออกเป็น 2 กลุ่ม ตามระดับความสามารถ คือ กลุ่มผู้สอบที่มีความสามารถต่ำและกลุ่มผู้สอบที่มีความสามารถสูง โดยใช้คะแนนเฉลี่ยของผู้สอบทั้งหมดเป็นเกณฑ์ในการแบ่งกลุ่ม แล้วคำนวณดัชนีการทำหน้าที่ต่างกันของข้อสอบแยกกันในแต่ละกลุ่ม ในการศึกษาครั้งนี้ผู้วิจัยวิเคราะห์โดยใช้โปรแกรม SIBTEST version 1.1 (Stout and Roussos, 1992)

วิธีแมนเทล-แฮนส์เซล หมายถึง วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในกลุ่มวิธี non-IRT ที่พัฒนาโดย Holland และ Thayer (1988) ซึ่งจะคำนวณดัชนีการทำหน้าที่ต่างกันของข้อสอบจากสัดส่วนการตอบข้อสอบถูกระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบแล้วจึงทดสอบนัยสำคัญด้วยสถิติ χ^2_{MH} ในการศึกษาครั้งนี้ผู้วิจัยใช้โปรแกรม MHDIF version 1.0 ของ Fidalgo (1995) วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปตามข้อเสนอแนะของ Mazor และคณะ (1994)

วิธีการถดถอยโลจิสติก หมายถึง วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในกลุ่มวิธี non-IRT ที่พัฒนาโดย Swaminathan และ Rogers (1990) ซึ่งจะคำนวณดัชนีการทำหน้าที่ต่างกันของข้อสอบจากผลการตอบข้อสอบถูกระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบโดยใช้โมเดลการถดถอยโลจิสติก แล้วทดสอบนัยสำคัญด้วยสถิติ χ^2 ในการศึกษาครั้งนี้ผู้วิจัยวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปโดยใช้โปรแกรม SPSS/PC$^+$ version 4.01

อำนาจการทดสอบ หมายถึง ร้อยละของจำนวนข้อสอบที่ระบุถูกต้องว่าทำหน้าที่ต่างกัน โดยคำนวณได้จากจำนวนข้อสอบที่ระบุถูกต้องว่าทำหน้าที่ต่างกันซึ่งตรวจสอบด้วยวิธีที่ศึกษา (วิธีซิปเทสท์ปรับใหม่ วิธีซิปเทสท์ วิธีแมนเทล-แฮนส์เซล และวิธีการถดถอยโลจิสติก) ต่อจำนวนข้อสอบที่ทำหน้าที่ต่างกันทั้งหมดในแบบสอบซึ่งตรวจสอบด้วยวิธีเกณฑ์ (วิธีการวัดพื้นที่ของ Raju)

อัตราความคลาดเคลื่อนประเภทที่ 1 หมายถึง ร้อยละของจำนวนข้อสอบที่ระบุผิดพลาดว่าทำหน้าที่ต่างกัน ทั้งที่ความเป็นจริงข้อสอบทำหน้าที่ไม่ต่างกัน โดยคำนวณได้จากจำนวนข้อสอบที่ระบุผิดพลาดว่าทำหน้าที่ต่างกันซึ่งตรวจสอบด้วยวิธีที่ศึกษา (วิธีซิปเทสท์ปรับใหม่ วิธีซิปเทสท์ วิธีแมนเทล-แฮนส์เซล และวิธีการถดถอยโลจิสติก) ต่อจำนวนข้อสอบที่ทำหน้าที่ไม่ต่างกันทั้งหมดในแบบสอบที่ตรวจสอบด้วยวิธีเกณฑ์ (วิธีการวัดพื้นที่ของ Raju)

เกณฑ์ของอัตราความคลาดเคลื่อนประเภทที่ 1 หมายถึง เกณฑ์ที่ใช้เปรียบเทียบกับอัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีการตรวจสอบที่ศึกษา 4 วิธี ในการศึกษาครั้งนี้ผู้วิจัยใช้เกณฑ์ที่ระดับ 10%

ประโยชน์ที่คาดว่าจะได้รับ

1. เพื่อเป็นแนวทางในการปรับปรุงข้อสอบให้มีความยุติธรรมต่อกลุ่มผู้สอบ ซึ่งเป็นอีกทางเลือกหนึ่งในการพัฒนาแบบสอบให้มีคุณภาพ

2. เพื่อเป็นแนวทางในการเลือกวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปที่มีประสิทธิภาพสูง ระหว่างวิธีซิปเทสท์ปรับใหม่ วิธีซิปเทสท์ วิธีแมนเทล-แฮนส์เซล และวิธีการถดถอยโลจิสติก ภายใต้เงื่อนไขของปัจจัยที่ศึกษา 4 ตัว คือ ลักษณะของข้อสอบ ความยาวของแบบสอบ สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน และขนาดกลุ่มตัวอย่าง

3. เพื่อเป็นแนวทางในการศึกษา ค้นคว้า วิจัย เกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปภายใต้การจำลองข้อมูล โดยใช้วิธีซิปเทสท์ วิธีซิปเทสท์ปรับใหม่ วิธีแมนเทล-แฮนส์เซล วิธีการถดถอยโลจิสติก และวิธีการวัดพื้นที่ของ Raju