การเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบวัดพหุมิติให้คะแนนหลายค่า ด้วยวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์

ณัฐพร ภักดี

ดุษฎีนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปรัชญาดุษฎีบัณฑิต
สาขาวิชาวิจัย วัดผลและสถิติการศึกษา
คณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา
สิงหาคม 2560

คณะกรรมการควบคุมดุษฎีนิพนธ์และคณะกรรมการสอบดุษฎีนิพนธ์ ได้พิจารณาดุษฎีนิพนธ์ของ ณัฐพร ภักดี ฉบับนี้แล้ว เห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปรัชญาดุษฎีบัณฑิต สาขาวิชาวิจัย วัดผลและสถิติการศึกษา ของมหาวิทยาลัยบูรพาได้

คณะกรรมการควบคุมดุษฎีนิพนธ์

.....อาจารย์ที่ปรึกษาหลัก
(รองศาสตราจารย์ ดร.ไพรัตน์ วงษ์นาม)

..... อาจารย์ที่ปรึกษาร่วม
(ผู้ช่วยศาสตราจารย์ ดร.สุรีพร อนุศาสนนันท์)

คณะกรรมการสอบดุษฎีนิพนธ์

.....ประธาน
(ดร.อาวีพร ปานทอง)

.....กรรมการ
(รองศาสตราจารย์ ดร.ไพรัตน์ วงษ์นาม)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.สุรีพร อนุศาสนนันท์)

..... กรรมการ
(ดร.สมพงษ์ ปั้นหุ่น)

คณะศึกษาศาสตร์อนุมัติให้รับดุษฎีนิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปรัชญาดุษฎีบัณฑิต สาขาวิชาวิจัย วัดผลและสถิติการศึกษา ของมหาวิทยาลัยบูรพา

..... คณบดีคณะศึกษาศาสตร์
(รองศาสตราจารย์ ดร.วิชิต สุรัตน์เรืองชัย)
วันที่ 10 เดือน มีนาคม พ.ศ. 2560

กิตติกรรมประกาศ

ดุษฎีนิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยดี เป็นเพราะผู้วิจัยได้รับความกรุณาอย่างยิ่งจาก รองศาสตราจารย์ ดร.ไพรัตน์ วงษ์นาม ผู้ช่วยศาสตราจารย์ ดร.สุรีพร อนุศาสนนันท์ ดร.สมพงษ์ ปั้นหุ่น และ ดร.อาวีพร ปานทอง ที่กรุณาให้คำปรึกษาแนะนำแนวทางที่ถูกต้อง ตลอดจนแก้ไขข้อบกพร่องต่าง ๆ ไปได้ด้วยความละเอียดถี่ถ้วนและเอาใจใส่ด้วยดีเสมอมา ผู้วิจัยรู้สึกซาบซึ้งเป็นอย่างยิ่ง จึงขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ โอกาสนี้

ขอขอบพระคุณ อาจารย์ธวัชชัย เอี่ยมไพโรจน์ ที่กรุณาให้คำปรึกษาแนะแนวทางในการแก้ไขปัญหาด้านการวิเคราะห์ข้อมูล ขอบคุณ นายมาโนชญ์ ใจกว้าง นักวิชาการคอมพิวเตอร์ห้องปฏิบัติการวิจัยวิศวกรรมระบบสารสนเทศ มหาวิทยาลัยบูรพา ที่กรุณาให้ความรู้ ให้คำปรึกษาด้านการเขียนโปรแกรม ทำให้งานวิจัยมีความสมบูรณ์และสำเร็จลุล่วง นอกจากนี้ยังได้รับการสนับสนุนช่วยเหลือด้านโปรแกรมทางสถิติจากรุ่นพี่ สาขาวิชาวิจัย วัดผลและสถิติการศึกษา และน้อง ๆ ที่ให้ความช่วยเหลือด้านต่าง ๆ ทำให้ดุษฎีนิพนธ์ฉบับนี้สำเร็จได้ด้วยดี

กราบขอบพระคุณบิดา มารดา และบุคคลในครอบครัวทุกคนที่คอยให้กำลังใจ และเสียสละคอยสนับสนุนให้ข้าพเจ้าเพื่อให้เกิดความสำเร็จในการดำเนินการวิจัยอย่างยิ่ง คุณค่าและประโยชน์ของดุษฎีนิพนธ์ฉบับนี้ ผู้วิจัยขอมอบเป็นกตัญญูกตเวทิตาแด่บุพการี บูรพาจารย์ และผู้มีพระคุณทุกท่านทั้งในอดีตและปัจจุบัน ที่ทำให้ข้าพเจ้าเป็นผู้มีการศึกษาและประสบความสำเร็จมาจนตราบเท่าทุกวันนี้

ณัฐพร ภักดี

55810089: สาขาวิชา: วิจัย วัดผลและสถิติการศึกษา; ปร.ด. (วิจัย วัดผลและสถิติการศึกษา)
คำสำคัญ: การทำหน้าต่างกันของข้อสอบ/ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ/
วิธีการทดสอบวอลด์

ณัฐพร ภักดี: การเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบวัดพหุมิติให้คะแนนหลายค่า ด้วยวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ (COMPARING PERFORMANCE OF DIFFERENTIAL ITEM FUNCTION DETECTION FOR MULTIDIMENSIONAL POLYTOMOUS SCORED ITEMS USING POLY-SIBBTEST, MULTI GROUP CONFIRMATORY FACTOR ANALYSIS AND WALD TEST) คณะกรรมการควบคุมดุษฎีนิพนธ์: ไพรัตน์ วงษ์นาม, ค.ด.,สุรีพร อนุศาสนนันท์, ค.ด. 200 หน้า. ปี พ.ศ. 2560.

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบวัดพหุมิติที่ให้คะแนนหลายค่า ด้วยวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ จากอัตราความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบ ในการศึกษาครั้งนี้ใช้ข้อมูลจำลองโดยจำลองภายใต้โมเดลเกรดเรสพอนพหุมิติ ซึ่งแต่ละข้อคำถามจะมีรายการตอบ 5 รายการ โดยให้คะแนนเป็น 1, 2, 3, 4 หรือ 5 คะแนน ข้อมูลดังกล่าวจำลองผลการตอบข้อสอบภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ ความยาวของแบบวัด 2 ขนาด ขนาดของการทำหน้าที่ต่างกัน 3 ระดับ สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 2 ขนาด และขนาดของกลุ่มตัวอย่าง 5 รูปแบบ รวมข้อมูลทั้งหมดที่ต้องจัดกระทำเพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจำนวน 60 เงื่อนไข (2x3x2x5) ในแต่ละเงื่อนไขวนซ้ำ 100 รอบ ผลการวิจัยพบว่า

1. อัตราความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบวัดพหุมิติให้คะแนนหลายค่า วิธีวิเคราะห์องค์ประกอบกลุ่มพหุ และวิธีการทดสอบวอลด์ สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ดีกว่าวิธีโพลีโตมัสซิปเทสท์ โดยวิธีวอลด์มีค่าอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำกว่าและควบคุมได้ดีกว่าวิธีวิเคราะห์องค์ประกอบกลุ่มพหุ เมื่อความยาวของแบบวัดเพิ่มขึ้น

2. อำนาจการทดสอบในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบวัดพหุมิติให้คะแนนหลายค่า วิธีวิเคราะห์องค์ประกอบกลุ่มพหุมีอำนาจการทดสอบสูงกว่าวิธีการทดสอบวอลด์ และวิธีโพลีโตมัสซิปเทสท์ ทั้งสามวิธีมีอำนาจการทดสอบสูง และใกล้เคียงทุกเงื่อนไข

3. โดยภาพรวมวิธีวิเคราะห์องค์ประกอบกลุ่มพหุมีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามดีกว่าวิธีวอดล์ เมื่อความยาวของแบบวัดมากขึ้น วิธีวอดล์มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันดีกว่าวิธีวิเคราะห์องค์ประกอบกลุ่มพหุ โดยทั้งสองวิธีมีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันดีกว่าวิธีโพลีโตมัสซิปเทสท์ในทุกเงื่อนไข

55810089: MAJOR: EDUCATONAL RESEARCH, MEASUREMENT AND STATISTICS; Ph.D. (EDUCATIONAL RESEARCH, MEASUREMENT AND STATISTICS)
KEYWORDS: DIFFERENTAL ITEM FUNCTIONING IN MULTIDIMENSIONAL TESTS/ MULTI GROUP CONFIRMATORY FACTOR ANALYSIS/ WALD TEST

NUTTAPORN PHAKDEE: COMPARING PERFORMANCE OF DIFFERENTIAL ITEM FUNCTION DETECTION FOR MULTIDIMENSIONAL POLYTOMOUS SCORED ITEMS USING POLY-SIBTEST, MULTI GROUP CONFIRMATORY FACTOR ANALYSIS AND WALD TEST. DISSERTATION ADVISORS: PAIRAT WONGNAME, Ph.D., SUREEPORN ANUSANANUN, Ph.D. 200 P. 2017.

The purposes of this research were to; compare the performance of the power of the type I error rate of Poly-SIBTEST, perform multi group confirmatory factor analysis and Wald test procedures in detecting the differential item functioning (DIF) for multidimensional polytomous items. The data were simulated under the multidimensional graded response model. The type of all items were in five response categories scoring as 1, 2, 3, 4 or 5. This data was simulated under a variety of four factors: two differing levels of test length, three differing levels of magnitude of DIF, two differing levels of proportion of DIF items, and five differing levels of sample sizes. A total of 60 conditions were studied. The data were replicated 100 times for each condition. The major findings were as follows;

1. The type I error rate of multi group confirmatory factor analysis and Wald test procedures on detecting of DIF for multidimensional polytomous items has control type I error rate lower than Poly-SIBTEST, Wald test has type I error rate lowest and when test length increased, Wald test was control type I error rate better than other methods.

2. The power of testing for Poly-SIBTEST, multi group confirmatory factor analysis, and Wald test procedures on detecting of differential item functioning (DIF) for multidimensional polytomous item have high power and were similar in all conditions.

3. Multi group confirmatory factor analysis is more efficient than the Wald test when test length increased. Wald test is more efficient than multi group confirmatory factor analysis and multi group confirmatory factor analysis and Wald test is more efficient than Poly-SIBTEST procedure on detecting of differential item functioning (DIF) for multidimensional polytomous items under all conditions.

สารบัญ

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญภาพ

สารบัญภาพ (ต่อ)

บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

การวัดและการประเมินผล เป็นกระบวนการที่นักวัดต้องการวัดคุณลักษณะภายในของบุคคลที่ไม่สามารถสังเกตได้โดยตรง ซึ่งคุณลักษณะภายในเหล่านี้จะใช้ข้อคำถามเป็นสิ่งเร้าให้ผู้ตอบแสดงคุณลักษณะภายในที่ต้องการวัดออกมา ดังนั้น ข้อคำถามที่สร้างขึ้นจะต้องสามารถวัดได้ตรงตามสิ่งที่ต้องการวัด และมีความคงเส้นคงวา ก็ย่อมมั่นใจได้ระดับหนึ่งว่า ข้อคำถามที่สร้างขึ้นมีความเชื่อถือได้ อย่างไรก็ตาม คุณภาพของข้อคำถามที่สร้างขึ้นจะมีมากเพียงใดนั้น ผู้พัฒนาข้อคำถามต้องวางแผนการสร้างอย่างรอบคอบ ครอบคลุมเนื้อหาที่ต้องการวัด มีความรู้ถึงแก่นแท้ของเนื้อหาที่จะวัด มีทักษะการเขียนข้อคำถาม รวมทั้งมีการตรวจสอบคุณภาพของข้อคำถามโดยนำไปทดลองกับกลุ่มตัวอย่าง แล้วนำผลการตอบมาวิเคราะห์คุณภาพรายข้อ ผลการวิเคราะห์รายข้อนี้ จะทำให้ทราบว่าข้อคำถามแต่ละข้อสามารถทำหน้าที่ได้ตรงตามที่ผู้พัฒนาต้องการหรือไม่ เพื่อเป็นข้อมูลพื้นฐานสำหรับการจัดทำเป็นแบบทดสอบที่เหมาะสมต่อไป

แบบทดสอบเป็นเครื่องมือหลักสำหรับการวัดผลทางการศึกษาและจิตวิทยา แบบทดสอบที่มีคุณภาพจะต้องประกอบไปด้วยความเที่ยง (Reliability) ความตรง (Validity) และ คุณลักษณะของข้อคำถามที่ได้มาตรฐาน โดยเฉพาะความตรงนั้น นับว่าเป็นหัวใจที่สำคัญยิ่งของแบบทดสอบที่แสดงถึงความสามารถในการวัดได้ถูกต้องแม่นยำ นั่นคือ คะแนนที่วัดได้จากแบบทดสอบสามารถวัดค่าได้ใกล้เคียงกับค่าคุณลักษณะภายในที่แท้จริงมากเพียงใด ก็ถือว่า การวัดมีความตรงมากขึ้นเพียงนั้น (Ayala, 2008; DeMars, 2010; Embretson & Reise, 2000; Holland & Wainer, 1993; Kunnan, 2000; ศิริชัย กาญจนวาสี, 2550) โดยทั่วไป การตรวจสอบความตรงของแบบทดสอบแบ่งออกได้เป็นสามประเภทหลัก คือ ความตรงเชิงเนื้อหา (Content validity) ความตรงตามเกณฑ์สัมพันธ์ (Criterion-related validity) และความตรงเชิงโครงสร้าง (Construct validity) นอกจากนี้ยังมีคุณสมบัติอีกประการหนึ่งที่นักวิจัยให้ความสำคัญสำหรับการตรวจสอบคุณภาพของเครื่องมือ นั่นคือ การตรวจสอบความไม่ยุติธรรมของข้อคำถามและแบบทดสอบ (Item and test unfairness)

ความยุติธรรม (Fairness) ของข้อคำถามหรือแบบทดสอบเกิดขึ้นในกรณีที่ผู้ตอบกลุ่มย่อยต่างกลุ่มกันมีลักษณะเฉพาะบางอย่างแตกต่างกัน มีความได้เปรียบหรือเสียเปรียบกัน ทั้งที่มีความสามารถจริงเท่าเทียมกัน แต่เดิมใช้คำว่า การตรวจสอบ “ความลำเอียงของข้อคำถามและแบบทดสอบ” (Item/ Test bias) ต่อมาระยะหลังได้มีผู้เสนอให้ใช้คำว่า “การตรวจสอบการทำหน้าที่

ต่างกันของข้อคำถามหรือแบบทดสอบ" (Differential item/ Test function) หรือ เรียกว่า DIF/ DTF (Angoff, 1993; Hambleton, Swaminathan, & Rogers, 1991) คำดังกล่าวมีความเหมาะสมมากกว่าในการอธิบายสารสนเทศเชิงสถิติ นักวิจัยจึงนำมาใช้กันอย่างแพร่หลายจนถึงทุกวันนี้

การทำหน้าที่ต่างกันของข้อคำถาม ที่ผ่านมามีผู้ให้ความหมายไว้หลายความหมายด้วยกัน สำหรับผู้วิจัย ขอให้ความหมายของการทำหน้าที่ต่างกันของข้อคำถามว่า การที่บุคคลหรือผู้ตอบมีเชื้อชาติ เพศ วัฒนธรรม หรืออื่น ๆ แตกต่างกัน มีความสามารถในแบบทดสอบที่ต้องการวัดเท่ากัน จะมีความน่าจะเป็นในการตอบข้อคำถามได้ในแต่ละรายการคำตอบแตกต่างกัน นอกจากนี้ Park and Lautenschlager (1990) ได้ให้ข้อสังเกตเกี่ยวกับการทำหน้าที่ต่างกันของข้อคำถามอยู่ 2 ประเด็น คือ 1) การตอบข้อคำถามมีอิทธิพลจากแหล่งความแปรปรวนจากข้อคำถามและจากผู้ตอบ 2) แหล่งความแปรปรวนภายนอกมีผลต่อกลุ่มผู้ตอบบางกลุ่มที่แตกต่างจากกลุ่มอื่น ส่วนการวิจัยของ Adams (2003) และ Adams, Wu, and Carstensen (2007) กล่าวว่า เกิดจากรูปแบบของข้อคำถามหรือการจัดพิมพ์ผิดพลาดของข้อคำถามฉบับย่อย ส่วน Le (2009) พบว่า การทำหน้าที่ต่างกันของข้อคำถามเกิดจากรูปแบบของข้อคำถามและเนื้อหาที่ใช้ออกข้อคำถาม และการศึกษาของ Wetzel, Carstensen, and Böhnke (2013) พบว่า ความแตกต่างของคะแนนความสามารถในกลุ่มใดกลุ่มหนึ่ง เกิดจากความยากของเนื้อหาในข้อคำถาม

การวิเคราะห์การทำหน้าที่ต่างกันของข้อคำถาม สามารถแบ่งได้เป็น 2 ประเภท (Mellenbergh, 1982) ได้แก่ 1) การทำหน้าที่แตกต่างกันแบบเอกรูป หรือเรียกว่า ข้อคำถามที่ทำหน้าที่ต่างกันแบบรูปแบบเดียวกัน (Uniform DIF) ซึ่งจะเกิดขึ้นเมื่อความสามารถของผู้ตอบกับการเป็นสมาชิกกลุ่มย่อยนั้นไม่มีปฏิสัมพันธ์ (Non-interaction) นั่นคือ โอกาสของการตอบข้อคำถามได้ถูกต้องของผู้ตอบกลุ่มย่อยกลุ่มหนึ่งสูงกว่าผู้ตอบกลุ่มย่อยอีกกลุ่มหนึ่งตลอดช่วงความสามารถ 2) การทำหน้าที่แตกต่างกันแบบไม่เป็นรูปแบบเดียวกัน (Non uniform DIF) ซึ่งจะเกิดขึ้นเมื่อความสามารถของผู้ตอบกับการเป็นสมาชิกกลุ่มย่อยนั้นมีปฏิสัมพันธ์ (Interaction) นั่นคือ โอกาสของการตอบข้อคำถามได้ถูกต้องของผู้ตอบกลุ่มย่อยกลุ่มหนึ่งสูงกว่าผู้ตอบกลุ่มย่อยอีกกลุ่มหนึ่งไม่ตลอดช่วงความสามารถ ซึ่งถ้าพิจารณาตามทฤษฎีการตอบสนองข้อคำถาม สามารถพิจารณาปฏิสัมพันธ์ได้จากความแตกต่างของค่าพารามิเตอร์อำนาจจำแนกของข้อคำถามระหว่างผู้ตอบสองกลุ่ม ถ้าข้อคำถามระหว่างผู้ตอบสองกลุ่มย่อยมีค่าพารามิเตอร์อำนาจจำแนกเท่ากันแล้ว โค้งคุณลักษณะข้อคำถาม (Item characteristics curves: ICC) ของผู้ตอบสองกลุ่มย่อยจะขนานกัน แสดงว่า เป็นข้อคำถามที่ทำหน้าที่ต่างกันแบบรูปแบบเดียวกัน แต่ถ้าข้อคำถามระหว่างผู้ตอบสองกลุ่มย่อยมีค่าพารามิเตอร์อำนาจจำแนกไม่เท่ากันแล้ว โค้งคุณลักษณะข้อคำถามของกลุ่มผู้ตอบจะไม่ขนานกัน แสดงว่า เป็นข้อคำถามที่ทำหน้าที่ต่างกันแบบไม่เป็นรูปแบบเดียวกัน

หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม เป็นการเปรียบเทียบการตอบระหว่างผู้ตอบกลุ่มย่อยสองกลุ่มที่มีความสามารถระดับเดียวกัน โดยที่ผู้ตอบกลุ่มหนึ่งเป็นตัวแทนกลุ่มหลักในประชากร เรียกว่า “กลุ่มอ้างอิง” (Reference group: R) ส่วนอีกกลุ่มหนึ่งเป็นตัวแทนกลุ่มรองในประชากร เรียกว่า “กลุ่มสนใจ” (Focal group: F) ซึ่งเป็นกลุ่มเป้าหมายที่ต้องการศึกษาข้อคำถามที่ใช้ในการตรวจสอบเรียกว่า “ข้อคำถามศึกษา” (Studies item) เมื่อข้อคำถามเกิดการทำหน้าที่ต่างกัน กลุ่มอ้างอิงจะได้เปรียบในการตอบ ส่วนผู้ตอบในกลุ่มสนใจคาดว่าจะเสียเปรียบในการตอบ เกณฑ์ที่ใช้ในการเปรียบเทียบผลการตอบระหว่างกลุ่มอ้างอิงและกลุ่มสนใจจำเป็นต้องใช้การจับคู่ (Matching) ตามความสามารถ

การศึกษาการทำหน้าที่ต่างกันของข้อคำถามได้มีการพัฒนาอย่างจริงจัง ระยะแรกมุ่งเน้นการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามที่วัดความสามารถเอกมิติและให้คะแนนสองค่า (Unidimensional dichotomous) โดยวิเคราะห์ภายใต้ทฤษฎีการทดสอบแบบดั้งเดิม (Classical test theory methods) โดยใช้คะแนนรวมของผู้ตอบเป็นเกณฑ์การจับคู่ของกลุ่มผู้ตอบ ได้แก่ วิธีการวิเคราะห์ความแปรปรวน (Analysis of variance: ANOVA) (Cardall & Coffman, 1964) ซึ่งเป็นจุดเริ่มต้นของวิธีแปลงค่าความยาก (Transformed item difficulty: TID) (Angoff, 1993) ต่อมา Scheuneman (1979) ได้เสนอวิธีไคกำลังสอง และได้ปรับแก้วิธีการจากวิธีไคกำลังสอง เป็นไคกำลังสองแบบเต็มรูป (Full chi square) จากแนวคิดการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ทำให้มีนักวิจัยหลายท่านได้คิดค้นและพัฒนาวิธีการตรวจสอบให้มีความเหมาะสมมากยิ่งขึ้น เช่น แมนเทล-แฮนเซล (Mantel-Heanszel) (Holland & Thayer, 1988) วิธีการทำให้เป็นมาตรฐาน (Standardization: STD) (Dorans & Kulick, 1986) วิธีการถดถอยโลจิสติก (Logistics regression: LR) (Swaminathan & Rogers, 1990) โดยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามตามทฤษฎีการทดสอบแบบดั้งเดิม มีจุดอ่อนที่ข้อตกลงเบื้องต้น คือ คะแนนความคลาดเคลื่อน ค่าพารามิเตอร์ข้อคำถามและแบบทดสอบแปรเปลี่ยนไปตามกลุ่มผู้ตอบ ด้วยเหตุนี้ จึงมีนักวิจัยพัฒนาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามด้วยทฤษฎีการตอบสนองข้อคำถาม (Item response theory)

การตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามด้วยทฤษฎีการตอบสนองข้อคำถามเป็นการตรวจสอบโดยใช้ค่าประมาณคุณลักษณะภายในหรือตัวแปรแฝงเป็นเกณฑ์ในการจับคู่ นักวิจัยส่วนมากให้การยอมรับกันโดยทั่วไปว่ามีประสิทธิภาพสูงกว่าการใช้คะแนนสังเกตตามทฤษฎีการทดสอบมาตรฐานเดิม ทั้งนี้เนื่องจากการประมาณค่าพารามิเตอร์ตามทฤษฎีการตอบสนองข้อคำถามไม่แปรเปลี่ยนไปตามกลุ่มผู้ตอบ (Hambleton et al., 1991) ทำให้ความแตกต่างของความสามารถระหว่างกลุ่มผู้ตอบไม่เป็นปัญหาในการตรวจสอบ อีกทั้งค่าประมาณ

พารามิเตอร์ของข้อคำถาม ได้แก่ ความยาก ค่าอำนาจจำแนก และค่าการเดาของข้อคำถาม มีความคลาดเคลื่อนน้อยกว่าการใช้ทฤษฎีการทดสอบแบบดั้งเดิม วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามของกลุ่มทฤษฎีการตอบสนองข้อคำถามนี้ จำแนกออกเป็นสองกลุ่มใหญ่ คือ 1) ใช้การเปรียบเทียบค่าประมาณพารามิเตอร์ของข้อคำถามระหว่างกลุ่มผู้ตอบ วิธีที่สำคัญในกลุ่มนี้ ได้แก่ วิธีการทดสอบไค-สแควร์ของลอร์ด (Lord's chi-square test) และวิธีการทดสอบอัตราส่วนความน่าจะเป็นในทฤษฎีการตอบสนองข้อคำถาม (IRT likelihood ratio test: IRT-LR) (Thissen, Steinberg, & Wainer, 1993) 2) ใช้การเปรียบเทียบค่าประมาณฟังก์ชันการตอบข้อคำถามระหว่างกลุ่มผู้ตอบ โดยการวัดพื้นที่ระหว่างฟังก์ชันการตอบข้อคำถาม วิธีที่สำคัญในกลุ่มนี้ ได้แก่ วิธีการวัดพื้นที่ชนิดไม่คิดเครื่องหมาย (Unsigned area) ในรูปรากกำลังสองของความแตกต่างเฉลี่ย (Root mean square difference) ระหว่าง IRFs (Linn & Hamisch, 1981) วิธีการวัดพื้นที่ชนิดคิดเครื่องหมายและไม่คิดเครื่องหมายในรูปผลรวมของกำลังสอง (Sum of square: SOS) (Shepard, Camilli, & Williams, 1984) วิธีการวัดพื้นชนิดคิดเครื่องหมายในช่วงเปิด (Exact signed area: ESA) และพื้นที่ชนิดไม่คิดเครื่องหมายในช่วงเปิด (Exactunsigned area: EUA)(Raju, 1990) และวิธีการวัดพื้นที่ชนิดคิดเครื่องหมายในช่วงปิด (Closed-interval signed area: CSA) และพื้นที่ชนิดไม่คิดเครื่องหมายในช่วงปิด (Closed-interval unsigned area: CUA) (Kim & Cohen, 1991)

ปัจจุบันการวัดผลทางการศึกษาและจิตวิทยาได้ให้ความสนใจการตรวจสอบการทำหน้าที่ต่างกันของคำถามสำหรับการให้คะแนนหลายค่า (Polytomous) ซึ่งส่วนมากปรับขยายหรือพัฒนามาจากวิธีการตรวจสอบสำหรับการให้คะแนนแบบสองค่า สำหรับในกลุ่มของทฤษฎีการทดสอบแบบดั้งเดิม ได้แก่ วิธีดัชนีมาตรฐานพหุวิภาค (Polytomous standardization) (Dorans & Kulick, 1986) วิธีแมนเทล-แฮนส์เซลทั่วไป (General Mantal-Haenzel: GMH) (Holland & Thayer, 1988) และวิธีการวิเคราะห์การถดถอยโลจิสติกพหุวิภาค (Polytomous logistic regression) (Swaminathan & Rogers, 1990) ส่วนในกลุ่มของทฤษฎีการตอบสนองข้อคำถาม เช่น วิธีอัตราส่วนไลค์ลิฮูดในรูปทั่วไป (General IRT likelihood ratio) วิธีการให้คะแนนบางส่วน (Partial credit model: PCM) (Masters, 1982) วิธีซิปเทสท์พหุวิภาค (Polytomous SIBTEST) (Shealy & Stout, 1993) วิธีการให้คะแนนบางส่วนทั่วไป (Gereralized partial credit model: GPCM) (Muraki, 1992) และวิธีการทำหน้าที่ต่างกันเป็นขั้น (Differential step functioning: DSF) (Penfield, 2007; 2008; 2010)

การตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ทั้งแบบการให้คะแนนสองค่า หรือหลายค่า มักอยู่บนเงื่อนไขของโมเดลการวัดความสามารถแบบเอกมิติ ซึ่งนักวิจัยหลายท่านเห็นว่าการตรวจสอบในแนวคิดแบบนี้มีจุดด้อยและขาดความเหมาะสม (Ackerman, 1992) ทั้งนี้เพราะว่าแบบทดสอบทางด้านการศึกษาและจิตวิทยามักจะประกอบไปด้วยคุณลักษณะภายในหลายมิติ

ประกอบกัน หากนำโมเดลการตอบสนองข้อคำถามแบบเอกมิติไปใช้ในการวิเคราะห์ จะทำให้เกิดปัญหา คือ เกิดความลำเอียงในการประมาณค่าพารามิเตอร์ของข้อคำถาม และการประมาณค่าความสามารถของผู้ตอบได้ (Wilson & Hoskens, 2005) โดย Walker and Beretvas (2001) ได้ทำการศึกษาการทำหน้าที่ต่างกันของแบบทดสอบพหุมิติกับเอกมิติ กับข้อมูลเชิงประจักษ์ ระหว่างวิธี Poly-SIBTEST และ LISREL ซึ่งผลการวิเคราะห์พบว่า การตรวจสอบการทำหน้าที่ต่างกันของแบบทดสอบพหุมิติมีความสอดคล้องกับข้อมูลเชิงประจักษ์มากกว่าการวิเคราะห์การทำหน้าที่ต่างกันของแบบทดสอบแบบเอกมิติ จากผลการวิจัยดังกล่าว การใช้คะแนนรวมหรือค่าประมาณความสามารถบนฐานคิดแบบเอกมิติมาใช้จำแนกระดับความสามารถของผู้ตอบ เพื่อจัดอันดับผู้ตอบแต่ละคนจึงไม่ถูกต้อง ถ้านำผลที่ได้ไปวิเคราะห์โดยใช้ทฤษฎีการตอบสนองข้อคำถามภายใต้เอกมิติจะทำให้ไม่เหมาะสม และมีผลทำให้ฟังก์ชันการตอบสนองข้อคำถามระหว่างกลุ่มแตกต่างกัน แล้วสรุปว่าข้อคำถามหรือข้อคำถามทำหน้าที่แตกต่างกัน (Oshima & Miller, 1992) ข้อสรุปนี้จึงมีความผิดพลาด เนื่องจากการละเมิดข้อตกลงเอกมิติ ดังนั้น สำหรับแบบทดสอบที่มีความซับซ้อนควรใช้ความสามารถในทุกมิติเป็นเกณฑ์การจับคู่ โดยใช้โมเดลการตอบสนองแบบพหุมิติ (Multidimensional item response model) (Reckase, 2009)

จากการศึกษางานวิจัยที่ผ่านมา Bolt (2002); Bolt, Hare, Vitale, & Newman (2004); Boughton (2004); Cai (2008; 2012); Cao, Tay, & Liu (2017); Chang, Mazzeo, & Roussos (1996); Gierl, Gotzmann, & Kannan & Kim (2009); Kannan & Ye (2008); Langer (2008); Miller and Spray (1993); Oshima, Raju, & Flowers (1997); Raju, Laffitte, & Byne (2002); Stark, Chernyshenko, & Drasgow (2006); Swaminathan and Rogers (1990); Tian (1999); Woods, Cai, & Wang (2013); Woods & Grimm (2011); Wu & Lei (2009); ชัยยศ ชาวระนอง (2553); มิ่ง เทพครเมือง (2554); สิริรัตน์ วิภาสศิลป์ (2545); สุชาติ สิริมีนนันท์ (2554); อรินทร์ น่วมถนอม (2549); อาวีพร ปานทอง (2558) และอุทัยวรรณ สายพัฒนะ (2547) พบว่า วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามได้มีการคิดค้นและพัฒนามาอย่างต่อเนื่องเป็นลำดับ จากวิธีหนึ่งไปสู่อีกวิธีหนึ่ง จากวิธีที่ใช้คะแนนสังเกตได้ตามทฤษฎีการทดสอบมาตรฐานเดิมไปสู่วิธีที่ใช้คะแนนความสามารถแฝงตามทฤษฎีการตอบสนองข้อคำถาม จากการให้คะแนนแบบสองค่าไปสู่การให้คะแนนแบบหลายค่า และจากการมุ่งวัดความสามารถเอกมิติไปสู่ข้อคำถามที่วัดความสามารถหลายมิติ ซึ่งผู้วิจัยขอนำเสนอประเด็นที่น่าสนใจ 3 ประเด็น ดังนี้

ประเด็นที่หนึ่ง ปัจจุบันนักวัดผลได้ปรับเปลี่ยนวิธีการวัดและประเมินผลจากแนวทางเดิมไปสู่วิธีการวัดและประเมินผลแนวใหม่ ที่มีลักษณะการให้คะแนนหลายค่า ซึ่งเป็นการประเมินตามสภาพจริง ในอดีตเราใช้วิธีการวิเคราะห์การให้คะแนนสองค่ามาใช้วิเคราะห์แบบทดสอบที่มี

การให้คะแนนหลายค่า จึงทำให้สารสนเทศเกี่ยวกับข้อคำถามบางส่วนถูกตัดทิ้งไป จึงมีนักวิจัยได้คิดค้นและพัฒนาทฤษฎีการตอบสนองข้อคำถามแบบหลายค่ามาช่วยในการวิเคราะห์แบบทดสอบดังกล่าว นอกจากนี้ แบบทดสอบทางด้านการศึกษาและจิตวิทยามักเป็นแบบทดสอบที่มีลักษณะการวัดความสามารถหลายมิติและให้คะแนนหลายค่า แต่ยังคงถูกวิเคราะห์ภายใต้ข้อตกลงเบื้องต้นของความเป็นเอกมิติของข้อคำถาม จึงทำให้เกิดปัญหาตามมา คือ เกิดความลำเอียงในการประมาณค่าพารามิเตอร์ของข้อคำถาม และการประมาณค่าความสามารถของผู้ตอบได้ ดังนั้น การวิเคราะห์ข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า จึงเป็นประเด็นที่สนใจในการวิจัยครั้งนี้

ประเด็นที่สอง การประเมินการทำหน้าที่ต่างกันของข้อคำถาม ที่ใช้คุณลักษณะแฝงเป็นเกณฑ์ในการจับคู่จะมีความตรงมากเพียงใด ขึ้นอยู่กับวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามที่โดยวิธีการที่ใช้ในการตรวจสอบมีหลายวิธีด้วยกัน สำหรับงานวิจัยนี้ ผู้วิจัยเลือกใช้วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ได้แก่ 1) วิธีโพลีโตมัสซิปเทสท์ (Poly-SIBTEST) พัฒนาโดย Chang et al. (1996) ปรับขยายมาจากวิธีซิปเทสท์ (SIBTEST) (Shealy & Stout, 1993) เป็นวิธีการตรวจสอบการทำหน้าที่ต่างกันแบบนอนพาราเมทริกซ์ สามารถใช้เพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามที่ให้คะแนนหลายค่า มีจุดเด่นคือ สามารถตรวจสอบได้ทั้งข้อคำถามที่วัดความสามารถเพียงมิติเดียวและข้อคำถามวัดความสามารถหลายมิติ นอกจากนี้ยังมีความถูกต้องแม่นยำ เนื่องจากใช้คะแนนการจับคู่ด้วยคุณลักษณะแฝงเป็นเกณฑ์ สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ที่สูงเกินปกติได้ 2) วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ (Multi group confirmatory factor analysis: MG-CFA) พัฒนาโดย Reise, Widaman, and Pugh (1993); Vandenberg and Lance (2000) เป็นการวิเคราะห์โดยนำแนวคิดของการวิเคราะห์แบบสมการเชิงโครงสร้าง (Structural equation model) มาปรับใช้กับการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ที่มีความยืดหยุ่นสำหรับการวิเคราะห์ความสามารถหลายมิติ (Kannan & Kim, 2009; Kannan & Ye, 2008; Raju et al., 2002) นอกจากนี้ Gonzales-Roma, Hernandez, and Gomez-Benito (2006) พบว่า วิธีการนี้มีอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามได้ดีโดยเฉพาะโมเดล Graded response และพบว่ามีอำนาจการทดสอบตั้งแต่ 0.70 ขึ้นไป สำหรับกลุ่มตัวอย่างขนาดเล็กที่มีขนาดของการทำหน้าที่ต่างกันระดับปานกลาง และมีอำนาจการทดสอบเพิ่มขึ้นเมื่อขนาดตัวอย่างและขนาดของการทำหน้าที่ต่างกันของข้อคำถามเพิ่มขึ้น นอกจากนี้ วิธีการนี้สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ในระดับที่ต่ำกว่า 0.10 ส่วน Wu and Lei (2009) ได้นำวิธี MG-CFA ไปใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม สำหรับโมเดลแบบพหุมิติแบบให้คะแนนแบบสองค่า ผลการวิเคราะห์พบว่า อัตราความคลาดเคลื่อนประเภทที่ 1

ลดลงอย่างมีนัยสำคัญทางสถิติ และมีอำนาจการทดสอบเพิ่มขึ้นอย่างมีนัยสำคัญ เมื่อวิเคราะห์ MG-CFA แบบสองมิติ 3) วิธีการทดสอบวอลด์ (Wald test) ซึ่งพัฒนามาจาก Lord's Wald เนื่องจากวิธีการทดสอบวอลล์ดั้งเดิมให้อัตราความคลาดเคลื่อนประเภทที่ 1 สูง (Donoghue & Isham, 1998; Kim, Cohen, & Kim, 1994; Lim & Drasgow, 1990; McLaughlin & Drasgow, 1987) นอกจากนี้ยังพบว่า การประมาณค่าจากเมทริกซ์ความแปรปรวนร่วมมีปัญหา (Donoghue & Isham, 1998; Kim, Chosen and Kim, 1995; McLaughlin & Drasgow, 1987) ดังนั้น วิธีของ Wald (Cai, 2012; Cai, Thissen, & du Toit, 2011) จึงได้ถูกปรับปรุงขึ้น Woods et al. (2013) ได้ทำการปรับปรุง Lord's Wald และได้นำไปใช้ทดสอบความแตกต่างกันของข้อคำถามเอกมิติสำหรับประชากรหลายกลุ่มภายใต้โมเดลเกรดเรสพอน พบว่า Wald-1 มีอัตราความคลาดเคลื่อนประเภทที่ 1 น้อยกว่า Wald-2 และพบว่า มีอำนาจการทดสอบเท่ากับหรืออาจจะมากกว่ากับ IRT-LR เมื่อขนาดตัวอย่างมากขึ้น และ Yao and Li (2010) ได้นำ Wald test ไปศึกษาการทำหน้าที่ต่างกันกับข้อคำถามแบบพหุมิติ พบว่า Wald test มีความคลาดเคลื่อนประเภทที่ 2 น้อยกว่าการทดสอบของราจู แต่มีอำนาจการทดสอบมากกว่าการทดสอบของราจู ด้วยเหตุผลที่กล่าวมา ผู้วิจัยจึงสนใจศึกษาประสิทธิภาพของวิธีการวิเคราะห์การทำหน้าที่ต่างกันของข้อคำถาม อันได้แก่ วิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์

ประเด็นที่สาม การวัดประเมินผลแนวใหม่ใช้แบบทดสอบพหุมิติให้คะแนนหลายค่ามีความสำคัญทางการศึกษาและจิตวิทยา เพื่อให้เกิดความยุติธรรมกับผู้ตอบทุกคนในแต่ละกลุ่ม จึงจำเป็นต้องมีการตรวจสอบคุณภาพแบบทดสอบด้านความเที่ยงตรงและคุณสมบัติของข้อคำถาม ซึ่งรวมถึงการทำหน้าที่ต่างกันของข้อคำถามด้วย เพื่อให้ผลการตรวจสอบมีประสิทธิภาพ ควรคำนึงถึงปัจจัยร่วมที่นำมาใช้ในการตรวจสอบ เช่น ขนาดของการทำหน้าที่ต่างกัน (Magnitude of DIF) ความยาวของแบบทดสอบ (Test lenght) สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน (Proportion of DIF) และขนาดตัวอย่าง (Sample size) เป็นต้น เพื่อเป็นสารสนเทศส่วนหนึ่งในการตัดสินใจเลือกใช้วิธีการตรวจสอบวิธีใดวิธีหนึ่ง ดังนั้น ปัจจัยดังกล่าวจึงเป็นประเด็นที่สงสัยว่าน่าจะมีผลต่อประสิทธิภาพในการตรวจสอบของวิธีโพลีโตมัสซิปเทส์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ หรือไม่ ทั้งสามวิธีสามารถตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามได้ดีที่สุดในเงื่อนไขใด

จากประเด็นต่าง ๆ ที่กล่าวมา ผู้วิจัยมีความสนใจที่จะนำวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ มาใช้ทดสอบการทำหน้าที่ต่างกันของข้อคำถาม ภายใต้ปัจจัย 4 ปัจจัย ประกอบด้วย ความยาวของแบบทดสอบ ขนาดของการทำหน้าที่ต่างกัน สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง การศึกษาครั้งนี้ ผู้วิจัย

อาศัยการจำลองข้อมูลเพื่อให้ได้ข้อมูลตามข้อกำหนดดังกล่าว โดยหวังว่า การวิจัยครั้งนี้จะช่วยให้นักวิจัยนำแนวคิดที่ได้จากการวิจัยไปใช้ในการตรวจสอบความตรงสำหรับเครื่องมือทางการวัดด้านการศึกษาและจิตวิทยา ซึ่งเป็นแบบทดสอบพหุมิติและมีการให้คะแนนหลายค่าในเรื่องของการทำหน้าที่ต่างกันของข้อคำถาม เพื่อให้เครื่องมือในการวัดมีคุณภาพและมีประสิทธิภาพในการวัดและประเมินผล

คำถามการวิจัย

1. การตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามพหุมิติที่ให้คะแนนหลายค่า ระหว่างวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ ภายใต้ปัจจัย 4 ปัจจัย คือ ความยาวของแบบทดสอบ ขนาดของการทำหน้าที่ต่างกัน สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่างต่างกัน จะมีอัตราความคลาดเคลื่อนประเภทที่ 1 ด้วยวิธีการตรวจสอบทั้งสามวิธี ต่างกันหรือไม่

2. การตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามพหุมิติที่ให้คะแนนหลายค่าระหว่างวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ ความยาวของแบบทดสอบ ขนาดของการทำหน้าที่ต่างกัน สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่างต่างกัน จะมีอำนาจการทดสอบของวิธีการตรวจสอบทั้งสามวิธี ต่างกันหรือไม่

วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาอำนาจการทดสอบ และอัตราความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามพหุมิติที่ให้คะแนนหลายค่าด้วยวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ ความยาวของแบบทดสอบ ขนาดของการทำหน้าที่ต่างกัน สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง

2. เพื่อเปรียบเทียบอำนาจการทดสอบ และอัตราความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามพหุมิติที่ให้คะแนนหลายค่าระหว่างวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ ความยาวของแบบทดสอบ ขนาดของการทำหน้าที่ต่างกัน สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง

สมมติฐานของการวิจัย

1. การตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามพหุมิติที่ให้คะแนนหลายค่าด้วยวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ มีความคลาดเคลื่อนประเภทที่ 1 ได้ต่ำกว่าหรือเท่ากับ .05 ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ ความยาวของแบบทดสอบ ขนาดของการทำหน้าที่ต่างกัน สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง

2. การตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามพหุมิติที่ให้คะแนนหลายค่าด้วยวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ มีอำนาจการทดสอบมากกว่าหรือเท่ากับ .80 ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ ความยาวของแบบทดสอบ ขนาดของการทำหน้าที่ต่างกัน สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง

3. การตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามพหุมิติที่ให้คะแนนหลายค่าด้วยวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ ความยาวของแบบทดสอบ ขนาดของการทำหน้าที่ต่างกัน สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง ของวิธีการตรวจสอบทั้งสามวิธี แตกต่างกัน

กรอบแนวคิดในการวิจัย

การตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามที่ให้คะแนนสองค่าหรือหลายค่ามีข้อจำกัดอยู่บนเงื่อนไขของข้อตกลงเบื้องต้นของความเป็นเอกมิติของแบบทดสอบ แนวคิดดังกล่าว นักวิจัยหลายคนวิพากษ์วิจารณ์ว่า มีจุดด้อยขาดความเหมาะสม (Ackerman, 1992; Ackerman & Evan, 1994; Camilli, 1992; Mazor, Hambleton, & Clauser,1998; Oshima & Miller, 1992; Oshima et al, 1997; Stout, Li, Nandakumar, & Bold, 1997) เนื่องจากแบบทดสอบทั่วไปมักประกอบด้วยความสามารถหลายมิติ ซึ่งความสามารถแต่ละมิติมีอิทธิพลในการตอบข้อคำถาม ดังนั้น การประมาณค่าความสามารถตามทฤษฎีการตอบสนองข้อคำถาม ภายใต้เงื่อนไขของข้อตกลงของความเป็นเอกมิติของแบบทดสอบ จึงไม่เหมาะสมที่จะนำมาใช้เป็นเกณฑ์ในการจับคู่ความสามารถของผู้ตอบ ถ้าผู้ตอบมาจากกลุ่มที่แตกต่างกันภายใต้การแจกแจงความสามารถหลายมิติและข้อคำถามสามารถจำแนกระดับความสามารถของผู้ตอบในหลายมิติ นั่นคือ การใช้คะแนนความสามารถแบบเอกมิติจะทำให้ข้อคำถามทำหน้าที่ต่างกัน ซึ่งมีผลทำให้การตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามมีอัตราความคลาดเคลื่อนประเภทที่ 1 สูงเกินปกติ ดังนั้น

ในแบบทดสอบที่วัดคุณลักษณะที่ซับซ้อนน่าจะใช้ความสามารถในทุกมิติเป็นเกณฑ์ในการจับคู่ เพื่อให้อัตราความคลาดเคลื่อนประเภทที่ 1 ลดลง

ในการวิจัยครั้งนี้ ผู้วิจัยได้นำวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ มาใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามพหุมิติให้คะแนนหลายค่า ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ ความยาวของแบบทดสอบ ขนาดของการทำหน้าที่ต่างกัน สัดส่วนข้อคำถามทำหน้าที่ต่างกัน และขนาดตัวอย่าง

ความยาวของแบบทดสอบ

การตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามด้วยการจำลองข้อมูล ความยาวของแบบทดสอบ เป็นปัจจัยหนึ่งที่ถูกนำมาใช้ในการศึกษา French and Maller (2007); Jodoin and Gierl (2001); Narayanan and Swaminathan, (1996); Paek and Wilson, (2011); Rogers & Swaminathan, (1993); Woods (2009); Yao & Li (2010) พบว่า อำนาจการทดสอบมีค่าสูงขึ้นถ้าความยาวของแบบทดสอบมากขึ้น สำหรับการศึกษาข้อคำถามให้คะแนนหลายค่าด้วยวิธีการจำลองข้อมูลส่วนใหญ่จะศึกษาที่ความยาวของแบบทดสอบน้อยกว่า 50 ข้อ (Su & Wang, 2005) โดยศึกษาอยู่ในช่วง 10 ถึง 40 ข้อ (Flowers, Oshima, & Raju, 1999; Wang & Su, 2004; Williams & Beretvas, 2006; Woods, 2009) จากการศึกษาของ Kim and Chen (1998 อ้างถึงใน ปิยะทิพย์ ตินวร, 2549) พบว่าความยาวของแบบทดสอบมีผลกระทบต่ออำนาจการทดสอบในการตรวจสอบการทำหน้าที่ต่างกัน และงานวิจัยของ Narayanan and Swaminathan (1996) พบว่า แบบทดสอบที่มีความยาว 40 ข้อ เป็นแบบทดสอบที่มีความยาวเพียงพอในการตรวจสอบการทำหน้าที่ต่างกัน ซึ่งสอดคล้องกับ Snow and Oshima (2009) นอกจากนี้ Wu and Lei (2009) ได้ทำการศึกษาโดยใช้การจำลองข้อมูลเพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามด้วยวิธี MG-CFA ด้วยความยาวของแบบทดสอบจำนวน 40 ข้อ และพบว่า วิธี MG-CFA มีความแกร่งในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม นอกจากนี้ วริษยา ชะม้อย (2550) ได้ทำการเปรียบเทียบวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามแบบให้คะแนนหลายค่า ด้วยความยาวของแบบทดสอบ จำนวน 20 และ 40 ข้อ จากเหตุผลดังกล่าว ผู้วิจัยจึงนำความยาวมาเป็นปัจจัยในการศึกษาการทำหน้าที่ต่างกัน คือ 20 ข้อ 40 ข้อ เนื่องจากเป็นระดับความยาวที่เหมาะสม นำมาใช้กันอย่างแพร่หลาย

ขนาดของการทำหน้าที่ต่างกันของข้อคำถาม

จากการงานวิจัยที่ศึกษาการทำหน้าที่ต่างกันของข้อคำถามที่ผ่านมา (Flowers et al., 1999; Meade, Lautenschlager, & Johnson, 2006; Stark et al., 2006) พบว่า ข้อคำถามที่มีขนาดของการทำหน้าที่ต่างกันที่สูงกว่า สามารถตรวจพบการทำหน้าที่ต่างกันได้มากกว่าขนาดของการทำหน้าที่ต่างกันต่ำ นอกจากนี้ยังพบว่า การศึกษาการทำหน้าที่ต่างกันของข้อคำถามสำหรับ

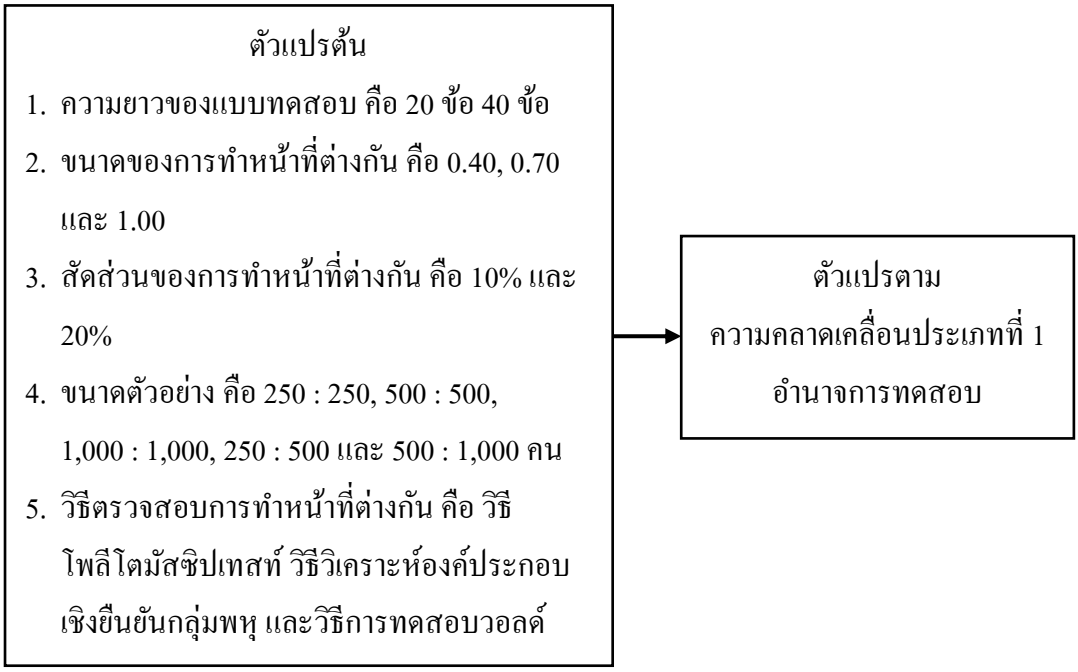
วิธีการทดสอบโดยใช้ทฤษฎีการตอบสนองข้อคำถาม (Flowers et al, 1999; Meade et al., 2006) ที่มีขนาดของการทำหน้าที่ต่างกันของข้อคำถามขนาดใหญ่ ($\geq$ .4) จะมีอำนาจการตรวจสอบการทำหน้าที่ต่างกันได้มากกว่า ดังนั้น งานวิจัยนี้ ผู้วิจัยจึงใช้ขนาดของการทำหน้าที่ต่างกัน คือ 0.40, 0.70 และ 1.00 เพื่อทดสอบผลการอ้างอิงดังกล่าว สำหรับการทดสอบในรูปแบบของการทำหน้าที่ต่างกันของข้อคำถามแบบไม่เป็นรูปแบบเดียวกัน (Non uniform DIF) คือ ค่าอำนาจจำแนกของข้อคำถามและค่าเทรสโฮลในแต่ละรายการคำตอบของกลุ่มสนใจ (FG) มีค่ามากกว่ากลุ่มอ้างอิง (RG) เท่ากับ 0.40, 0.70 และ 1.00

สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน

เนื่องจากการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามเป็นการเปรียบเทียบความสามารถของผู้สอบภายใต้เกณฑ์การจับคู่ที่มีความเชื่อมั่น (Clauser & Mazor,1988; Potenza & Dorans, 1995) ถ้าในแบบทดสอบที่มีข้อคำถามทำหน้าที่ต่างกันปะปนอยู่ จะทำให้ค่าประมาณความสามารถมีความเชื่อมั่นต่ำลง ซึ่งจะมีผลทำให้อำนาจการทดสอบลดลง หรืออัตราความคลาดเคลื่อนประเภทที่ 1 สูงกว่าปกติ การกำหนดสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกันได้มีการศึกษาอย่างต่อเนื่อง ซึ่งส่วนใหญ่จะทำการศึกษาที่ 10% ของจำนวนข้อคำถามทั้งหมด Raju, van der Linder, and Fleer (1995 cited in Flowers et al, 1999) พบว่า สัดส่วนของการทำหน้าที่ต่างกันของข้อคำถามที่มีสัดส่วนการทำหน้าที่ต่างกันมาก (มากกว่า 20%) จะมีผลทำให้พบขนาดของผลบวกลวง (False positive) และผลลบลวง (False negative) มีค่าสูงขึ้น เมื่อสัดส่วนของการทำหน้าที่ต่างกันมากขึ้น กล่าวคือ อัตราความคลาดเคลื่อนประเภทที่ 1 และความคลาดเคลื่อนประเภทที่ 2 มีค่ามากขึ้น ดังนั้น นักวิจัยรุ่นหลังจึงกำหนดสัดส่วนของการทำหน้าที่ต่างกันของข้อคำถามที่มีค่าอยู่ใกล้ 10% (Bolt, 2002; Gonzales-Roma et al., 2006; Stark et al., 2006; Wu & Lei, 2009)จากการศึกษาของอรินทร์ น่วมถนอม (2549) เมื่อสัดส่วนข้อคำถามที่ทำหน้าที่เบี่ยงเบนเพิ่มขึ้นในช่วง 10% ถึง 30% ไม่มีผลต่อวิธีโพลีโตมัสซิปเทสท์ และวิธีการถดถอยโลจิสติกแบบจัดอันดับหลายมิติ แต่มีผลต่อวิธีการถดถอยโลจิสติกแบบจัดอันดับ และสุชาติ สิริมีนนันท์ (2554) ได้ทำการศึกษาการทำหน้าที่ต่างกันของข้อคำถามด้วยวิธีการจำลอง โดยให้สัดส่วนการทำหน้าที่ต่างกันเป็น 10% 20% และ 30% พบว่า วิธีการวิเคราะห์ฟังก์ชันการจำแนกโลจิสติกมีอำนาจการทดสอบสูงสุด และมีอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำสุด ในเงื่อนไขสัดส่วนของการทำหน้าที่ต่างกัน 10% ในการศึกษาครั้งนี้ ผู้วิจัยจึงสนใจสัดส่วนของการทำหน้าที่ต่างกัน 2 ขนาด ได้แก่ 10% และ 20%

ขนาดตัวอย่าง

สำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ซึ่งเป็นการเปรียบเทียบผลการตอบระหว่างผู้สอบสองกลุ่ม การตรวจสอบการทำหน้าที่ต่างกันนี้ ขนาดตัวอย่างก็เป็นปัจจัยหนึ่งที่สำคัญสำหรับการตรวจสอบการทำหน้าที่ต่างกัน (MacCallum, Widaman, Zhang, & Hong, 1999) ในการศึกษาที่ผ่านมามีการกำหนดขนาดตัวอย่างสำหรับการศึกษาทั้งแบบที่มีสัดส่วนเท่ากัน (Lubke & Muthén, 2004; Meade & Lautenschlager, 2004; Gonzales-Roma et al., 2006) และสัดส่วนที่ต่างกัน (Finch & French, 2007; Maller & French, 2004) โดยการศึกษามีช่วงของการกำหนดตัวอย่างระหว่าง 250 ถึง 1,000 (Meade et al., 2006, Meade & Lautenschlager, 2004; Stark et al, 2006, Gonzales-Roma et al., 2006) ดังนั้น ผู้วิจัยจึงได้กำหนดตัวอย่างสำหรับการศึกษาในครั้งนี้ โดยอัตราส่วนของกลุ่มสนใจกับกลุ่มอ้างอิง เป็น 1 : 1 และ 1 : 2 จำนวน 5 ขนาด คือ 250 : 250, 500 : 500, 1,000 : 1,000, 250 : 500 และ 500 : 1000 คน



ภาพที่ 1 กรอบแนวคิดในการวิจัย

ประโยชน์ที่ได้รับจากการวิจัย

1. ทราบถึงประสิทธิภาพของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามที่เหมาะสมกับโมเดลการตอบสนองสองมิติให้คะแนนหลายค่า ด้วยวิธีการทดสอบโพลีโตมัส

ซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ ภายใต้เงื่อนไขปัจจัยความยาวของแบบทดสอบ ขนาดของการทำหน้าที่ต่างกันของข้อคำถาม สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง

2. เป็นแนวทางในการเลือกใช้วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามสำหรับโมเดลการตอบสนองข้อคำถามในแบบทดสอบสองมิติและให้คะแนน 5 ค่า สำหรับแบบทดสอบที่มีความยาวขนาด 20 ข้อ และ 40 ข้อ

3. เป็นแนวทางในการศึกษาเกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบสอบ แบบทดสอบ หรือเครื่องมือวัดอื่น ๆ ต่อไป

ขอบเขตของการวิจัย

1. ข้อมูลที่ใช้ในการศึกษาครั้งนี้เป็นข้อมูลจำลอง สำหรับข้อมูลจำลอง จำลองโดยใช้โมเดลเกรดเรสพอนพหุมิติ (Multidimensional graded response model) ที่มีโครงสร้างวัดความสามารถสองมิติ (Two-dimensional) และเป็นแบบทดสอบที่มีความเป็นพหุมิติระหว่างข้อคำถาม (Multidimensional between item test) โดยมีค่าความสัมพันธ์ของทั้งสองมิติเป็น .50 และข้อคำถามแต่ละข้อมีรายการคำตอบ 5 รายการ โดยให้คะแนนเป็น 1, 2, 3, 4 หรือ 5 ใช้รูปแบบของการทำหน้าที่ต่างกันไม่เป็นรูปแบบเดียวกัน คือ Non uniform DIF โดยกำหนดให้ค่าอำนาจจำแนกและค่าเทรสโฮลทุกรายการของข้อคำถามระหว่างกลุ่มอ้างอิงและกลุ่มสนใจมีค่าแตกต่างกัน และจำลองผลการตอบภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย ได้แก่ ความยาวของแบบทดสอบ 2 ขนาด ขนาดของการทำหน้าที่ต่างกัน 3 ขนาด สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 2 ขนาด และขนาดตัวอย่าง 5 รูปแบบ รวมข้อมูลที่ต้องกระทำเพื่อตรวจสอบการทำหน้าที่ต่างกันจำนวน 60 เงื่อนไข (3x2x2x5) ในแต่ละเงื่อนไขจำลองข้อมูลซ้ำ 100 ครั้ง รวมทั้งหมด 6,000 เงื่อนไข

2. ตัวแปรที่ใช้ในการศึกษา ประกอบด้วย

2.1 ตัวแปรอิสระมี 5 ตัวแปร ดังนี้

2.1.1 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามมี 3 วิธี ดังนี้

2.1.1.1 วิธีโพลีโตมัสซิปเทสท์

2.1.1.2 วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ

2.1.1.3 วิธีการทดสอบวอลด์

2.1.2 ความยาวของแบบทดสอบมี 2 ขนาด ดังนี้

2.1.2.1 จำนวน 20 ข้อ

2.1.2.2 จำนวน 40 ข้อ

2.1.3 ขนาดของการทำหน้าที่ต่างกัน 3 วิธี ดังนี้

2.1.3.1 ขนาดเล็ก 0.40

2.1.3.2 ขนาดกลาง 0.70

2.1.3.3 ขนาดใหญ่ 1.00

2.1.4 สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 2 ขนาด ดังนี้

2.1.4.1 จำนวน 10%

2.1.4.2 จำนวน 20%

2.1.5 ขนาดตัวอย่าง ประกอบด้วยกลุ่มสนใจและกลุ่มอ้างอิงที่มีจำนวนผู้ตอบ/ผู้สอบในแต่ละกลุ่มเท่ากันมีอัตราส่วนเป็น 1 : 1 และ 1 : 2 ดังนี้

2.1.5.1 จำนวน 250 : 250 คน

2.1.5.2 จำนวน 500 : 500 คน

2.1.5.3 จำนวน 1000 : 1000 คน

2.1.5.4 จำนวน 250 : 500 คน

2.1.5.5 จำนวน 500 : 1000 คน

2.2 ตัวแปรตาม มี 2 ตัวแปร ดังนี้

2.2.1 อัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม

2.2.2 อำนาจการทดสอบของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม

นิยามศัพท์เฉพาะ

1. แบบทดสอบหลายมิติ (Multidimensional test) หมายถึง แบบทดสอบที่มีโครงสร้างในการวัดความสามารถของผู้สอบที่มีลักษณะเด่นจำนวนสองมิติหรือมากกว่า ในการศึกษาครั้งนี้จำลองแบบทดสอบสองมิติ (Two-dimensional) ซึ่งเป็นแบบทดสอบความสามารถหลายมิติระหว่างข้อคำถาม (Multidimensional between-item test) โดยข้อคำถามแต่ละข้อวัดความสามารถมิติใดมิติหนึ่ง กำหนดให้ θ_1 และ θ_2 แทนความสามารถหลักในมิติที่หนึ่งและมิติที่สอง ตามลำดับ ข้อคำถามทุกข้อมีรายการคำตอบ 5 รายการ โดยให้คะแนนหลายค่า (Polytomous) ผลการตอบในรายการเป็น 1, 2, 3, 4 และ 5 และให้คะแนนเป็น 1, 2, 3, 4 หรือ 5 ตามลำดับ

2. ข้อคำถามที่ให้คะแนนหลายค่า (Polytomously scored item) หมายถึง ข้อคำถามที่มีรายการตอบหลายรายการ ในแต่ละรายการตอบจะมีการกำหนดขั้นของการให้คะแนนผลการตอบ

ข้อคำถามไว้เป็นจำนวนเต็ม ซึ่งในการวิจัยนี้ ข้อคำถามแต่ละข้อมีรายการตอบ 5 รายการ โดยให้คะแนนเป็น 1, 2, 3, 4 หรือ 5

3. การทำหน้าที่ต่างกันของข้อคำถาม (Differential item function: DIF) หมายถึง ผู้สอบที่มาจากกลุ่มที่แตกต่างกัน และมีความสามารถตามที่ข้อคำถามต้องการวัดเท่ากัน จะมีความน่าจะเป็นในการตอบข้อคำถามได้ถูกต้องไม่เท่ากัน

4. กลุ่มอ้างอิง (Reference group: R) หมายถึง กลุ่มผู้สอบที่คาดว่าจะได้เปรียบในการตอบข้อคำถามเมื่อข้อคำถามทำหน้าที่ต่างกัน โดยมีความน่าจะเป็นในการตอบข้อคำถามได้ถูกต้องมากกว่ากลุ่มสนใจ

5. กลุ่มสนใจ (Focal group: F) หมายถึง กลุ่มผู้สอบที่เป็นเป้าหมายของการศึกษา ซึ่งคาดว่าจะเสียเปรียบในการตอบข้อคำถามเมื่อข้อคำถามทำหน้าที่ต่างกัน โดยมีความน่าจะเป็นในการตอบข้อคำถามได้ถูกต้องน้อยกว่ากลุ่มอ้างอิง

6. สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน (Proportion of DIF) หมายถึง จำนวนข้อคำถามที่ทำหน้าที่ต่างกันต่อจำนวนข้อคำถามทั้งหมดในแบบทดสอบ ในการวิจัยครั้งนี้ ศึกษาสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 2 ขนาด 10% และ 20% คือ สำหรับกรณีความยาวแบบทดสอบเป็น 20 ข้อ จะมีข้อที่ทำหน้าที่ต่างกันจำนวน 2 ข้อ (10%) และ 4 ข้อ (20%) สำหรับกรณีความยาวแบบทดสอบเป็น 40 ข้อ จะมีข้อที่ทำหน้าที่ต่างกันจำนวน 4 ข้อ (10%) และ 8 ข้อ (20%)

7. ขนาดตัวอย่าง (Sample size) หมายถึง จำนวนผู้สอบของกลุ่มสนใจ และกลุ่มอ้างอิง ในการวิจัยครั้งนี้ใช้จำนวนผู้สอบของกลุ่มสนใจต่อกลุ่มอ้างอิง 5 รูปแบบ คือ 250 : 250 คน 500 : 500 คน 250 : 500 คน 500 : 1,000 คน และ 1,000 : 1,000 คน

8. ความยาวของแบบทดสอบ (Test length) หมายถึง จำนวนข้อคำถามทั้งหมดในแบบทดสอบ ในการวิจัยนี้ ศึกษาความยาวของแบบทดสอบ 2 ขนาด คือ 20 ข้อ และ 40 ข้อ

9. ขนาดของการทำหน้าที่ต่างกันของข้อคำถาม (Magnitude DIF) หมายถึง ขนาดของการทำหน้าที่ต่างกันของรูปแบบการทำหน้าที่ต่างกันไม่เป็นรูปแบบเดียวกัน ในการวิจัยครั้งนี้ ศึกษาขนาดของการทำหน้าที่ต่างกัน 3 ขนาด คือ .40, .70 และ 1.00 คือ

$$\left(a_{iR} = a_{iF} + 0.40, \tau_{i1R} = \tau_{i1F} - 0.40, \tau_{i2R} = \tau_{i2F} - 0.40, \tau_{i3R} = \tau_{i3F} - 0.40, \tau_{i4R} = \tau_{i4F} - 0.40\right)$$

$$\left(a_{iR} = a_{iF} + 0.70, \tau_{i1R} = \tau_{i1F} - 0.70, \tau_{i2R} = \tau_{i2F} - 0.70, \tau_{i3R} = \tau_{i3F} - 0.70, \tau_{i4R} = \tau_{i4F} - 0.70\right)$$

$$\left(a_{iR} = a_{iF} + 1.00, \tau_{i1R} = \tau_{i1F} - 1.00, \tau_{i2R} = \tau_{i2F} - 1.00, \tau_{i3R} = \tau_{i3F} - 1.00, \tau_{i4R} = \tau_{i4F} - 1.00\right)$$

10. การทำหน้าที่ต่างกันไม่เป็นรูปแบบเดียวกัน (Non uniform DIF) หมายถึง การที่พารามิเตอร์อำนาจจำแนกของข้อคำถามทุกข้อสำหรับกลุ่มสนใจมีค่าน้อยกว่ากลุ่มอ้างอิง และพารามิเตอร์เทรสโฮลในทุกรายการตอบข้อคำถามสำหรับกลุ่มสนใจมีค่าสูงกว่ากลุ่มอ้างอิง

11. วิธีโพลีโตมัสซิปเทสท์ (Polytomous SIBTEST: Poly-SIBTEST) หมายถึง วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในกลุ่มวิธีที่ใช้คุณลักษณะแฝง ซึ่งเป็นไปตามทฤษฎีการตอบสนองข้อสอบ รูปแบบนอนพาราเมตริก (Nonparametric) ซึ่งพัฒนาโดย Chang et al. (1996) เพื่อใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามที่ให้คะแนนหลายค่า ในการศึกษาครั้งนี้นำมาตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบสองมิติและให้คะแนน 5 ค่า

12. วิธีการวิเคราะห์องค์ประกอบเชิงยืนยันพหุ (Multi group confirmatory factor analysis) หมายถึง วิธีการวิเคราะห์การทำหน้าที่ต่างกันของข้อคำถาม โดยนำแนวคิดของสมการเชิงโครงสร้าง (Structural equation model) มาปรับใช้กับการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามสำหรับรูปแบบการวัดพหุมิติ โดยพิจารณาจากความไม่สอดคล้องกันของการวิเคราะห์ความแตกต่างของน้ำหนักองค์ประกอบ (ค่าอำนาจจำแนก) และค่าความยากแต่ละรายการคำตอบ (เทรสโฮล) ของข้อคำถาม เพื่อนำมาใช้ในการวิเคราะห์โครงสร้างของแบบทดสอบแยกตามกลุ่มผู้สอบ ในการวิจัยนี้ ผู้วิจัยได้นำมาใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบสองมิติและให้คะแนน 5 ค่า

13. วิธีการทดสอบแบบวอลด์ (Wald test) หมายถึง วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ที่ Cai (2012) พัฒนามาจากวิธีการของลอร์ดวอลด์ (Lord's Wald) โดยขจัดข้อบกพร่องจากการประมาณค่าพารามิเตอร์ความแปรปรวนร่วมของวิธีการของลอร์ดวอลด์ด้วยวิธีการประมาณ Supplemented expectation maximization หรือ SEM เป็นการทดสอบบนพื้นฐานของทฤษฎีการตอบสนองข้อสอบโดยใช้ตัวแปรแฝงแบบพาราเมตริก (Parametric) เพื่อใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ซึ่งการศึกษานี้ได้นำมาใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบสองมิติและให้คะแนน 5 ค่า

14. อำนาจการทดสอบ (Power of test) หมายถึง สัดส่วนของจำนวนข้อคำถามที่ตรวจสอบได้ถูกต้องว่าทำหน้าที่ต่างกันได้อย่างถูกต้อง โดยคำนวณจากจำนวนข้อสอบที่ตรวจสอบได้ถูกต้องว่าทำหน้าที่ต่างกัน ต่อจำนวนข้อคำถามที่ทำหน้าที่ต่างกันทั้งหมดในแบบทดสอบ

15. อัตราความคลาดเคลื่อนประเภทที่ 1 (Type I error) หมายถึง สัดส่วนของจำนวนข้อคำถามที่ตรวจสอบผิดพลาดว่าทำหน้าที่ต่างกัน ทั้งที่จริงข้อคำถามทำหน้าที่ไม่ต่างกัน ต่อจำนวนข้อคำถามที่ทำหน้าที่ไม่ต่างกันทั้งหมดในแบบทดสอบ

16. ประสิทธิภาพการทำหน้าที่ต่างกันของข้อคำถาม หมายถึง ความถูกต้องของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม จากการตรวจสอบด้วยวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ ซึ่งพิจารณาได้จากอัตราความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบ

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

ในการวิจัยครั้งนี้ มีจุดมุ่งหมายเพื่อเปรียบเทียบวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามพหุมิติแบบให้คะแนนหลายค่าด้วยวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ ขนาดของการทำหน้าที่ต่างกันของข้อคำถาม ความยาวของแบบทดสอบ สัดส่วนของการทำหน้าที่ต่างกัน และขนาดตัวอย่าง ผู้วิจัยได้ศึกษากรอบแนวคิด ทฤษฎี หลักการ รูปแบบ และวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ตลอดจนงานวิจัยต่าง ๆ ที่เกี่ยวข้อง เพื่อนำองค์ความรู้ที่ได้จากการศึกษามาประยุกต์ใช้กับการศึกษา จากการศึกษาสามารถกำหนดเป็นกรอบการศึกษาใน 5 ประเด็น ดังนี้

ตอนที่ 1 โมเดลการตอบสนองข้อสอบพหุมิติ

ตอนที่ 2 แนวคิดและหลักการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม

ตอนที่ 3 รูปแบบของข้อคำถามทำหน้าที่ต่างกัน

ตอนที่ 4 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม

ตอนที่ 5 งานวิจัยที่เกี่ยวข้อง

ตอนที่ 1 โมเดลการตอบสนองข้อสอบพหุมิติ (Multidimensional item response theory)

ความเป็นมาของทฤษฎีการตอบสนองข้อคำถามพหุมิติ

ในช่วงปลายปี ค.ศ. 1970s ถึงช่วงต้นปี ค.ศ. 1980s มีนักวิจัยหลายท่านพัฒนาแนวคิดเกี่ยวกับทฤษฎีการตอบสนองแบบพหุมิติ ดังจะเห็นได้จากการศึกษาของ Reckase (1972) ที่พัฒนาโมเดลของ Rasch แบบพหุมิติ นอกจากนี้ยังมี Mulaik (1972), Sympson (1978) และ Whitely (1980; 1981) ได้เสนอโมเดลพหุมิติที่มีปฏิสัมพันธ์ระหว่างผู้ตอบกับข้อคำถาม ดังนี้

$$P\left(U_{ij}\middle|\boldsymbol{\theta}_{\mathbf{j}},\boldsymbol{\eta}_{\mathbf{i}}\right)=\frac{\sum_{k=1}^{m} e^{\left(\theta_{jk}-\eta_{ik}\right)u_{ij}}}{1+\sum_{k=1}^{m} e^{\left(\theta_{jk}-\eta_{ik}\right)u_{ij}}} \qquad (1)$$

เมื่อ U_{ij} แทน คะแนน ที่มีค่าเป็น 0 หรือ 1

θ_j แทน เวกเตอร์ความสามารถของผู้ตอบคนที่ j

m แทน จำนวนมิติในโมเดล

โมเดลนี้มีคุณสมบัติที่น่าสนใจ คือ เมื่อกำหนดค่าคงที่ให้แก่ส่วนยกกำลังของเอ็กโพแนนเชียล (e) โอกาสในการตอบถูกเพิ่มขึ้น เมื่อจำนวนมิติ (m) เพิ่มขึ้น ถ้าค่าของส่วนยกกำลังของเอ็กโพแนนเชียลเป็นศูนย์ในทุกมิติ โอกาสในการตอบถูกจะมีค่าเท่ากับ $m/m+1$ และเมื่อกำหนดความน่าจะเป็นที่จะตอบถูกให้คงที่ ค่าของเอ็กโพแนนเชียลจะมีการเปลี่ยนแปลงถ้าจำนวนมิติเพิ่มขึ้น

Sympson (1978) และ Whitely (1980) ได้นำเสนอโมเดลความสัมพันธ์ระหว่างจำนวนมิติและขนาดของเอ็กโพแนนเชียลซึ่งตรงข้ามกับ Mulaik (1972) กล่าวคือ ถ้ามีการกำหนดค่าให้แก่ส่วนที่ยกกำลังของเอ็กโพแนนเชียล ความน่าจะเป็นในการตอบถูกจะมีค่าลดลง เมื่อจำนวนมิติเพิ่มขึ้น สามารถแสดงได้ดังสมการต่อไปนี้

$$P(U_{ij} \mid \theta_j, \alpha_j, b_i, c_i) = C_i + (1-C_i).\prod_{k=1}^{m} \frac{e^{a_{ik}\left(\theta_{jk} - b_{ik}\right)}}{1+e^{a_{ik}\left(\theta_{jk} - b_{ik}\right)}} \qquad (2)$$

เมื่อ C_i แทน เป็นพารามิเตอร์เสกลลาร์ (Scalar)

จากสมการที่ 2 พบว่า ถ้าส่วนยกกำลังของเอ็กโพแนนเชียลมีค่าเป็น 0 ความน่าจะเป็นในการตอบถูกจะมีค่าเท่ากับ $C_i + (1-C_i)(.5)^m$ และเมื่อ m มีค่ามากขึ้น ค่าของสมการจะมีค่าลู่เข้า C_i

McKinley and Reckase (1982) ได้ศึกษาความแปรเปลี่ยนของโมเดล Rasch และยังคงเป็นโมเดลแบบพหุตัวแปร (Multivariate) ได้นำเสนอในรูปของโมเดลโลจิสติก ซึ่งเป็นโมเดลที่นิยมใช้ในปัจจุบัน เหมาะสำหรับการวิเคราะห์เชิงสำรวจ (Exploratory) กับการให้คะแนนสองค่า ดังแสดงได้ตามสมการต่อไปนี้

$$P\left(u_{ij}=1 \mid \mathbf{\theta_j}, \mathbf{a_i}, d_i\right) = \frac{e^{\sum_{k=1}^{m} a_{ik}\theta_{ik} + d_i}}{1+e^{\sum_{k=1}^{m} a_{ik}\theta_{ik} + d_i}} \qquad (3)$$

เมื่อ $\mathbf{a_i}$ แทน เวกเตอร์พารามิเตอร์อำนาจจำแนกข้อคำถามข้อที่ i

$\boldsymbol{\theta}_\mathbf{j}$ แทน เวกเตอร์ความสามารถของผู้สอบ (Ability vector) คนที่ j

d_i แทน ค่าพารามิเตอร์จุดตัด (Intercept term) ของข้อสอบข้อที่ i

จากการศึกษาความเป็นมาของโมเดลทฤษฎีการตอบสนองพหุมิติ ถือว่าเป็นการเริ่มต้นแนวคิดที่สำคัญที่นำไปสู่การพัฒนาวิธีการศึกษาโมเดลทฤษฎีการตอบสนองพหุมิติ และนำไปสู่การนำไปสู่การประยุกต์ใช้มากขึ้น

ประเภทของโมเดลการตอบสนองข้อสอบแบบพหุมิติ

โมเดลทฤษฎีการตอบสนองพหุมิติ ประกอบด้วย พารามิเตอร์ความสามารถของผู้ตอบตั้งแต่สองค่าขึ้นไป ค่าอำนาจจำแนกจะมีผลกระทบต่อข้อคำถามทุกมิติ โดยจะเพิ่มความเหมาะสมสำหรับผลการตอบข้อคำถาม เมื่อบุคคลมีความสามารถแตกต่างกัน ในข้อคำถามที่มีความยากง่ายแตกต่างกัน

1. โมเดลการตอบสนองข้อสอบพหุมิติให้คะแนนสองค่า (Multidimensional dichotomous item response model)

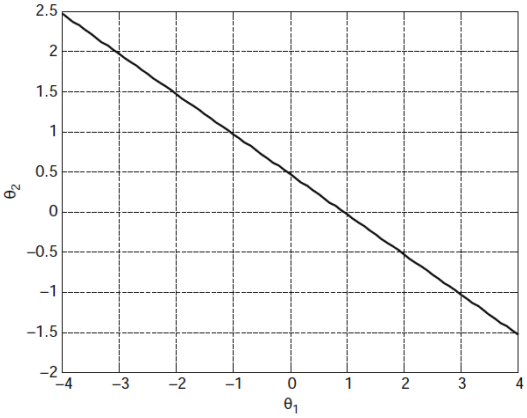
Reckase (1985) นำเสนอโมเดลการตอบสนองข้อสอบแบบพหุมิติ ที่นำมาใช้ในการวิจัย แบ่งออกเป็นสองประเภท ได้แก่ โมเดลชดเชยได้ (Compensatory model) และโมเดลชดเชยไม่ได้ (Noncompensatory model) ซึ่งมีรายละเอียด ดังนี้

1.1 โมเดลชดเชยได้ (Compensatory model)

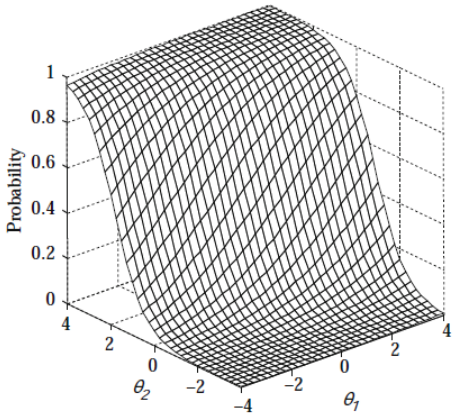
โมเดลนี้อธิบายได้ว่า ผู้สอบที่มีความสามารถในมิติใดมิติหนึ่งต่ำ จะมีความสามารถจากมิติอื่นที่มีความสามารถสูงมาชดเชย สามารถแสดงในรูปของสมการโลจิสติกสองพารามิเตอร์ ดังสมการที่ 3

การนำเสนอกราฟของทฤษฎีการตอบสนองข้อคำถามพหุมิติ มีข้อจำกัดในการนำเสนอเพียงสองมิติเท่านั้น เพราะข้อจำกัดทางด้านกราฟฟิก และ Reckase (1985) เรียกระนาบของข้อคำถามหลายมิติ ว่า ระนาบการตอบสนองข้อคำถาม (Item response surface: IRS)

ถ้า $\sum_{k=1}^{m} a_{ik}\theta_{ik} + d_i = 0$ จะทำให้ค่าความน่าจะเป็นในสมการที่ 3 มีค่าเท่ากับ 0.5 ถ้ากำหนดให้เวกเตอร์อำนาจจำแนกมีค่าเท่ากับ [0.75 1.5] และ ค่าพารามิเตอร์ $d = -0.7$ จะได้ว่า $0.75\theta_1 + 1.5\theta_2 - 0.7 = 0$ และ ฟังก์ชันเชิงเส้นตรง ที่แสดงความชันและจุดตัดของสมการ $\theta_2 = -0.5\theta_1 + \frac{7}{1.5}$ ดังภาพที่ 2 นอกจากนี้ ยังสามารถแสดงความชันและจุดตัดของสมการแสดงดังภาพที่ 3



ภาพที่ 2 เวกเตอร์เมื่อกำลังเอ็กโพแนนเชียลเป็น 0 สำหรับแบบทดสอบ เมื่อ $a_1 = 0.75, a_2 = 1.5$ $d = -0.7$ (Reckase, 2009)



ภาพที่ 3 ระนาบการตอบสนองข้อคำถาม สำหรับโมเดลแบบชดเชย ระหว่างความสามารถในมิติที่ 1 และ 2 (Reckase, 2009)

โมเดลแบบชดเชยของการตอบสนองข้อคำถามพหุมิติด้วยโมเดลโลจิสติกแบบสามพารามิเตอร์ (Multidimensional three-parameter logistic model) มีสมการ ดังนี้

$$P(u_{ij}=1|\theta_j, a_i, c_i, d_i) = c_i + (1-c_i)\frac{e^{\sum_{k=1}^{m} a_{ik}\theta_{jk} + d_i}}{1 + e^{\sum_{k=1}^{m} a_{ik}\theta_{jk} + d_i}} \quad (4)$$

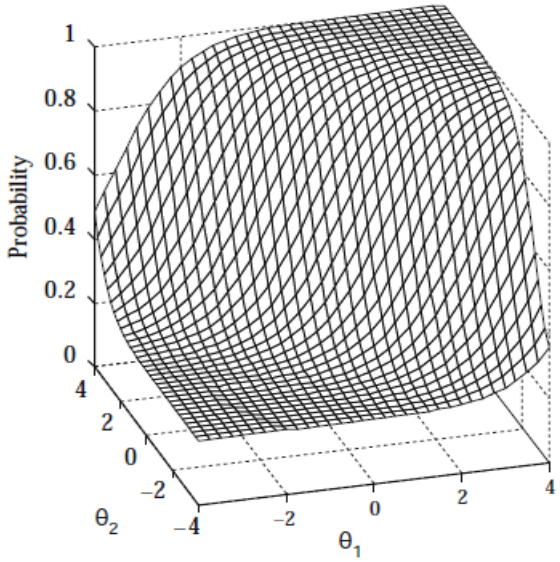
เมื่อ a_i แทน เวกเตอร์อำนาจจำแนกข้อคำถาม ข้อที่ i ในมิติที่ k

θ_j แทน เวกเตอร์ความสามารถของผู้สอบ คนที่ j ในมิติที่ k

c_i แทน ค่าพารามิเตอร์การเดา ข้อที่ i

d_i แทน ค่าพารามิเตอร์จุดตัด ข้อที่ i

ความสัมพันธ์ของความสามารถแต่ละมิติ จากการวิเคราะห์ด้วยโมเดลโลจิสติกแบบสามพารามิเตอร์ สามารถแสดงด้วยระนาบการตอบสนองข้อคำถามข้อที่ 1 ของความสามารถมิติที่ 1 และ 2 เมื่อ $a_{11}=1.3, a_{12}=1.4, d_1=-1, c_1=0.2$ ตามสมการที่ 4 ดังแสดงในภาพที่ 4



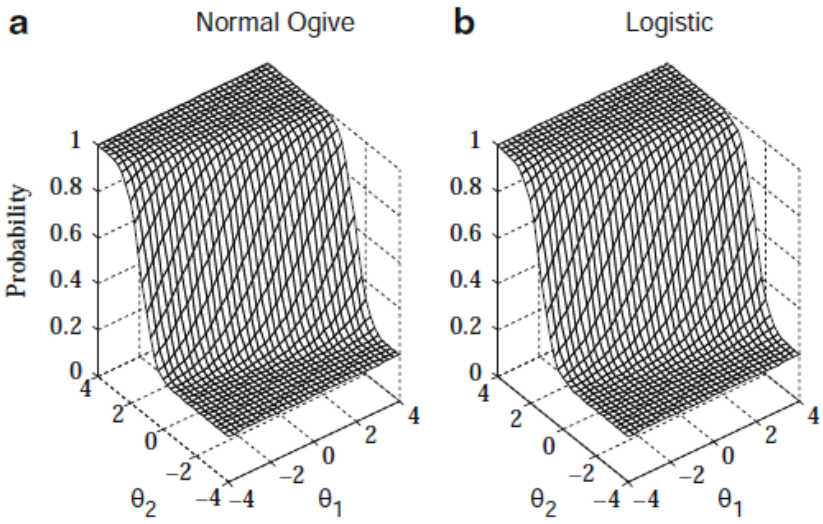
ภาพที่ 4 ระนาบการตอบสนองข้อคำถามของโมเดลโลจิสติกแบบสามพารามิเตอร์ (Reckase, 2009)

สามารถใช้โมเดลพหุมิติแบบปกติสะสมสามพารามิเตอร์ (Multidimensional three-parameter normal ogive model) นำเสนอโดย Bock and Schilling (2003), McDonald (1999) และ Samejima (1974) วิเคราะห์ความน่าจะเป็นในการตอบถูก ดังสมการต่อไปนี้

$$P\left(u_{ij}=1 \mid \theta_j, a_i, c_i, d_i\right)=c_i+\left(1-c_i\right)\int_{-z_i(\theta)}^{\alpha}\frac{1}{\sqrt{2\pi}}e^{\frac{-t^z}{z}}dt \qquad (5)$$

เมื่อ $z_i\left(\theta_j\right)=\sum_{k=1}^{m}a_{ik}\theta_{jk}+d_i$

ระนาบของความน่าจะเป็นของการวิเคราะห์ด้วยโมเดลโลจิสติกแบบสามพารามิเตอร์ที่ปรับค่าด้วย $1.702\sum_{k=1}^{m} a_{ik}\theta_{jk} + d_i = 0$ และโมเดลแบบปกติสะสม จะเป็นดังแสดงในภาพที่ 5 แสดงว่า การปรับค่าโมเดลโลจิสติกด้วย 1.702 ไม่มีผลกระทบกับระนาบการตอบสนองข้อคำถาม



ภาพที่ 5 ระนาบการตอบสนองข้อคำถามของโมเดลปกติสะสมและโลจิสติกโมเดล
เมื่อ $a_1 = .5, a_2 = 1.5, d = 0, c = .2$ ปรับค่าด้วย 1.702 (Reckase, 2009)

1.2 โมเดลไม่สามารถชดเชยได้ (Noncompensatory model)

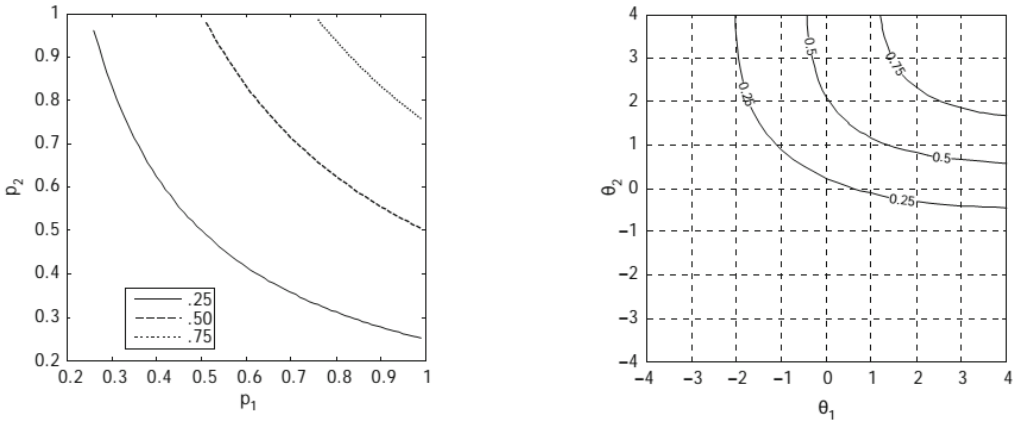
โมเดลนี้ อธิบายได้ว่า ผู้สอบที่มีความสามารถในมิติใดมิติหนึ่งต่ำ จะไม่สามารถนำความสามารถสูงในอีกมิติมาชดเชยได้หรือชดเชยได้น้อย หรืออาจเรียกว่า โมเดลชดเชยได้บางส่วน (Partially compensatory model) สมการที่ใช้ในการวิเคราะห์ด้วยโมเดลโลจิสติกแบบสามพารามิเตอร์ (Three parameter logistic model) แสดงดังนี้

$$P(u_{ij} = 1 | \theta_j, a_i, b_i, c_i) = c_i + (1 - c_i)\prod_{k=1}^{m} \frac{e^{1.70za_{ik}(\theta_{jk} - b_{ik})}}{1 + e^{1.70za_{ik}(\theta_{jk} - b_{ik})}} \quad (6)$$

เมื่อ a_i แทน เวกเตอร์อำนาจจำแนกของข้อคำถาม ข้อที่ i
θ_j แทน เวกเตอร์ความสามารถของผู้สอบ คนที่ j
b_i แทน เวกเตอร์ความยากของข้อคำถามข้อที่ i

c_i แทน ค่าพารามิเตอร์การเดา ข้อที่ i

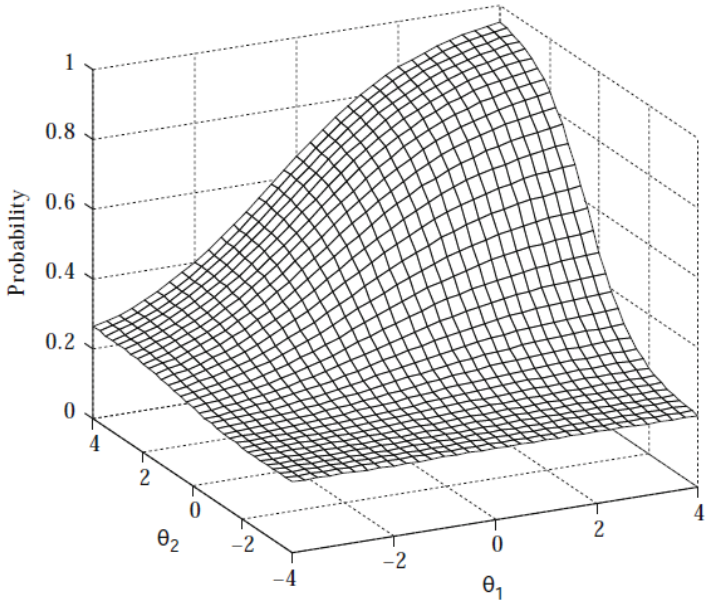
จากสมการที่ 6 เมื่อ $c_i = 0$ และความน่าจะเป็นของการตอบข้อคำถามได้ถูกต้องมีค่าเท่ากับ k และแต่ละเทอมของความน่าจะเป็นในแต่ละมิติที่สนใจ l จะได้ว่า $k = \prod_{l=1}^{m} P_i$ ถ้ามิติที่สนใจมีเพียงสองมิติ แล้ว $k = p_1 p_2$ สมการที่ได้ความน่าจะเป็นของการตอบข้อคำถามได้ถูกต้อง เช่น $k = .25,.50$ และ $.75$ ถ้า $k = .5$ และข้อคำถามข้อนั้นใช้ความสามารถจากสองมิติ ผลการตอบข้อคำถามจะแสดงในเส้นประแบบไฮเปอร์โบลา ในภาพที่ 6 (a) และถ้าพารามิเตอร์ข้อคำถามในโมเดลสองมิติ คือ $c_i = 0, a_{i1} = .7, a_{i2} = 1.1, b_{i1} = .5$ และ $b_{i2} = .5$ โค้งในระนาบของความสามารถ θ ซึ่งมีความสัมพันธ์กับไฮเปอร์โบลาทางซ้าย แสดงดังภาพที่ 6 (b)



(a) (b)

ภาพที่ 6 โค้งของความน่าจะเป็นที่เท่ากันของการตอบสนองข้อคำถามสำหรับโมเดลไม่ชดเชยของสองพิกัดและ $c_i = 0$ (Reckase, 2009)

เมื่อ $a_1 = .7, a_2 = 1.1, b_1 = -.5, b_2 = .5$ และ $c_i = .2$ สามารถแสดงระนาบการตอบสนองข้อคำถามได้ดังภาพที่ 7



ภาพที่ 7 ระนาบของการตอบสนองข้อคำถามของโมเดลแบบไม่ชดเชย (Reckase, 2009)

โมเดลการตอบสนองข้อสอบแบบพหุมิติสำหรับการให้คะแนนหลายค่า

(Multidimensional polytomous item response model)

1. Multidimensional generalized partial credit model: MGPM

Yao and Schwarz (2006) ได้นำเสนอโมเดลพหุมิติแบบ Generalized partial credit เป็นโมเดลที่อธิบายปฏิสัมพันธ์ระหว่างผู้ตอบกับข้อคำถามหรือข้อคำถาม ซึ่งมีการตรวจให้คะแนนแบบหลายค่า โดยค่าสูงสุดของข้อคำถามหรือข้อคำถามที่ i คือ K_i คะแนนต่ำสุดมีค่าเท่ากับ 0 และจำนวนรายการทั้งหมดมีค่าเท่ากับ K_i+1 ดังนั้น คะแนนของผู้ตอบในข้อคำถามหรือข้อคำถามคือ $k=0,1,\ldots,K_i$ โมเดลของ MGPM แสดงได้ดังนี้

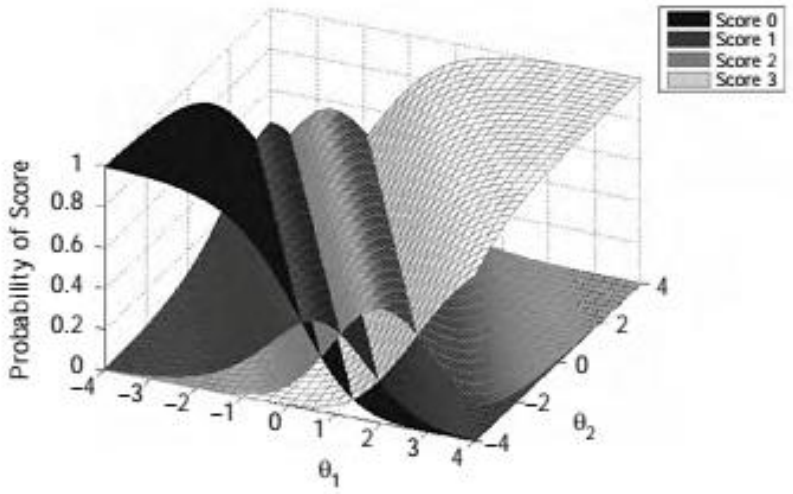
$$P\left(u_{ij}=k\mid\theta_j\right)=\frac{e^{ka_j\theta_j^i-\Sigma_{u=0}^{k}\beta_{iu}}}{\sum_{\nu=0}^{K_i}e^{\nu a_i\theta_j^i-\Sigma_{u=0}^{\nu}\beta_{iu}}} \tag{7}$$

เมื่อ β_{iu} แทนพารามิเตอร์เทรสโฮลในรายการคำตอบที่ $u_{ij}=k$ และ $\beta_{iu}=0$

เมื่อทำการเปรียบเทียบโมเดล MGPC กับ GPC จะพบความแตกต่างที่สำคัญสองประการ คือ ประการแรก โมเดลนี้ไม่ได้เสนอค่าพารามิเตอร์ความยาก และค่าเทรสโฮล ประการที่สอง θ เป็นเวกเตอร์ และ β_{iu} เป็นสเกลลาร์

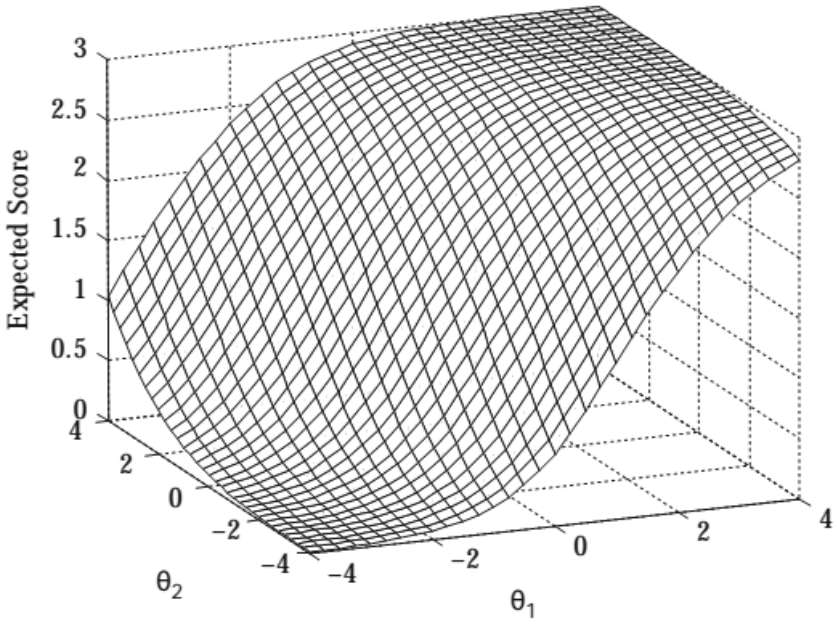
จากสมการที่ 7 สามารถนำเสนอความน่าจะเป็นที่อยู่ในระนาบการตอบสนองข้อคำถาม โดยกำหนดให้แบบทดสอบนี้มีการให้คะแนนการตอบตั้งแต่ 0 ถึง 3 และ $a_i = [1.2,.7]$ และ $\beta_{iu} = 0,-2.5,-1.5,.5$ แสดงดังภาพที่ 7 นอกจากนี้ยังสามารถนำเสนอการคำนวณค่าคาดหวังของรายการคำตอบของผู้ตอบในมิติความสามารถดังสมการที่ 7 และสามารถนำเสนอในรูปของระนาบคะแนนที่คาดหวัง ดังภาพที่ 8

$$\mathrm{E}(U_{ij} \mid \theta_j) = \sum_{k=0}^{K_i} kP(U_{ij} = k \mid \theta_j) \tag{8}$$



ภาพที่ 8 ระนาบการตอบสนองข้อคำถามของแต่ละรายการตอบ สำหรับโมเดล MGPC (Reckase, 2009)

ภาพที่ 8 แสดงระนาบของแต่ละรายการคำตอบ จำนวน 4 ระนาบ ซึ่งเป็นจำนวนรายการคำตอบที่เป็นไปได้ ระนาบสีดำที่สุดจะแทนคะแนนศูนย์ ซึ่งความน่าจะเป็นจะลดลงเมื่อความสามารถเพิ่มขึ้นทั้งสองมิติ ส่วนระนาบของรายการคะแนนหนึ่งและสอง ในช่วงแรกจะเพิ่มขึ้นและมีค่าลดลงเมื่อความสามารถทั้งสองมิติเพิ่มขึ้น และเมื่อรายการคะแนนคำตอบเป็นสาม จะมีค่าความน่าจะเป็นเพิ่มขึ้นเมื่อความสามารถในแต่ละมิติเพิ่มขึ้น



ภาพที่ 9 ระนาบคะแนนที่คาดหวังของข้อคำถาม สำหรับโมเดล MGPC (Reckase, 2009)

จากภาพที่ 9 จะแสดงระนาบคะแนนที่คาดหวังของโมเดลแบบชดเชยของ MGPC ซึ่งมีลักษณะเดียวกันกับการวิเคราะห์การตอบสนองข้อคำถามแบบเอกมิติ ผลของการวิเคราะห์จะแสดงในรูปที่มีพิกัด (4, -4) ซึ่งพบว่ามีคะแนนที่คาดหวังเข้าใกล้ 3

2. Multidimensional partial credite model: MPC

Kelderman and Rijkes (1994) ได้เสนอรูปแบบทั่วไปของโมเดลพหุมิติหนึ่งพารามิเตอร์ของ Rasch กรณีตรวจให้คะแนนหลายค่า ซึ่ง Adams, Wilson, and Wang (1997) ได้ปรับปรุงสมการใหม่เพื่อให้สามารถเปรียบเทียบกับโมเดลอื่น ๆ ได้ ดังสมการต่อไปนี้

$$P\left(u_{ij}=k \mid \theta_j\right)=\frac{e^{\sum_{l=1}^{m}\left(\theta_{jl}-b_{ilk}\right)w_{ilk}}}{\sum_{r=0}^{K_i} e^{\sum_{l=1}^{m}\left(\theta_{jl}-b_{ilk}\right)w_{ilk}}} \qquad (9)$$

เมื่อ b_{ilk} แทน พารามิเตอร์ความยาก ข้อที่ i บนมิติ l ในรายการที่ k

W_{ilk} แทน น้ำหนักคะแนนก่อนนิยาม ข้อที่ i บนมิติ l ในรายการที่ k

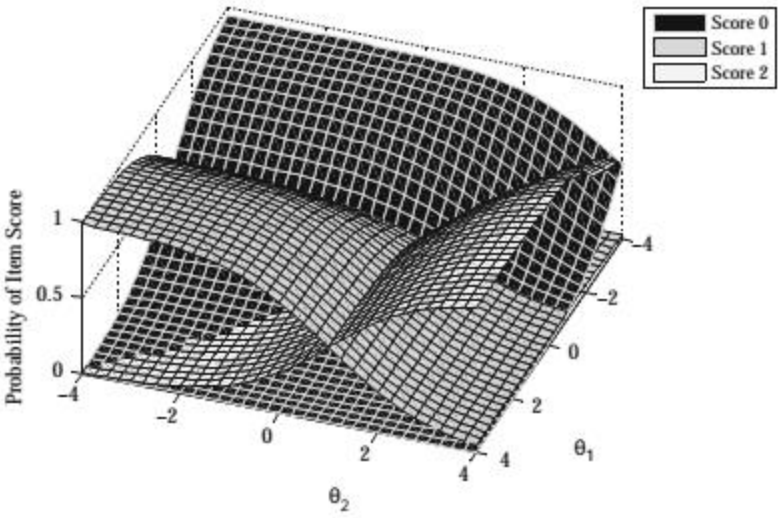
สำหรับฟังก์ชันที่สำคัญของโมเดลนี้คือ เมทริกซ์น้ำหนักเฉพาะ W_{ilk} เช่น ข้อคำถามมี $K_i = 3$ ระดับคะแนน ได้แก่ 0, 1 และ 2 ดังนั้น คำตอบจะมีความไวบนความแตกต่างของสองมิติ โดยที่เมทริกซ์ของน้ำหนักของข้อคำถามเป็น $\begin{bmatrix} 0\,0 \\ 1\,0 \\ 1\,1 \end{bmatrix}$ แถวของเมทริกซ์จะแทนด้วยระดับคะแนน คือ k และ คอลัมน์แทนด้วยมิติ คือ l ซึ่งเมทริกซ์นี้แสดงให้เห็นว่า ในแถวแรก ผู้ตอบ ตอบผิด (ได้คะแนน 0) ทั้งสองมิติ และแถวที่สอง หมายถึง มิติแรกตอบถูก มิติที่สอง ตอบผิด และแถวสุดท้าย หมายถึง ผู้ตอบ ตอบได้ถูกต้องทั้งสองมิติ

สำหรับกรณีเฉพาะนี้ เมื่อ $k = 0$ในสมการที่ 9 จะทำให้ $e^0 = 1$สำหรับค่าอื่น ๆ ของ k และทำให้ง่ายขึ้นดังสมการที่ 10

$$P(U_{ij} = k \mid \theta) = \frac{e^{\sum_{l=1}^{m}(\theta_{jl} - b_{ilk})}}{1 + \sum_{r=0}^{m} e^{\sum_{l=1}^{m}(\theta_{jl} - b_{ilk})}}, k = 1, \ldots, K_i \qquad (10)$$

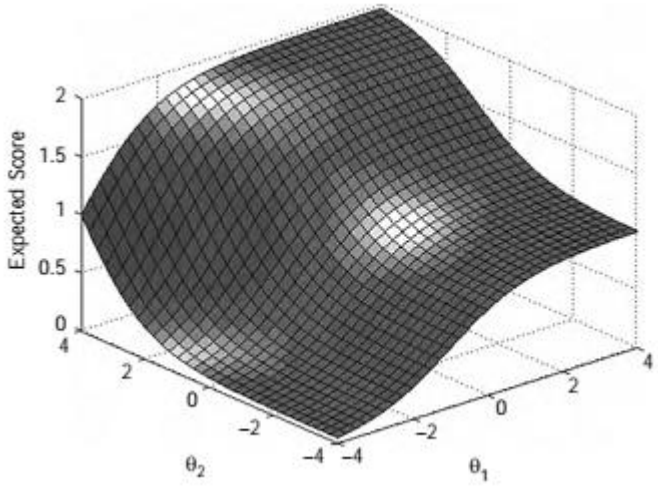
ผู้พัฒนาโมเดลนี้แสดงให้เห็นว่า การประมาณค่าพารามิเตอร์ b_{ilk} มีค่าไม่คงที่ จึงต้องกำหนดค่าให้เท่ากันในทุกระดับของรายการคำตอบ ได้แก่ $k = 1, \ldots, K_i$ ซึ่งข้อคำถามจะมีค่าพารามิเตอร์ความยากที่แตกต่างกันในแต่ละมิติ แต่จะมีค่าเท่ากันในรายการคำตอบภายใต้มิติเดียวกัน คือ ฟังก์ชันของข้อคำถาม (Item function) ของชุดการตอบของข้อคำถามที่มีการให้คะแนนแบบ 2 ค่า สำหรับในแต่ละมิติ ด้วยค่าพารามิเตอร์ความยาก

ระนาบการตอบสนองข้อคำถาม สำหรับรายการคำตอบเป็น 0, 1 หรือ 2 ในสองมิติ เมื่อค่าพารามิเตอร์ b_{ilk} มีค่าเป็น -1 ในมิติที่ 1 และเป็น +1 ในมิติที่ 2 ดังแสดงในภาพที่ 10 โดยระนาบของการตอบสำหรับรายการคำตอบที่มีค่าเป็น 0 จะมีค่าความน่าจะเป็นสูงสุดอยู่ที่ θ = (-4, -4) และค่าความน่าจะเป็นต่ำสุดอยู่ที่ θ = (4, 4) นอกจากนี้ยังพบว่า ระนาบของรายการคำตอบที่มีค่าเป็น 0 ยังตัดกับระนาบรายการคำตอบมีค่าเป็น 1 ที่ θ_1 = -1 นอกจากนี้ยังพบว่า ระนาบของรายการคำตอบที่มีคะแนนเป็น 1 จะมีค่าความน่าจะเป็นเข้าใกล้ 1 เมื่อ θ = (-4, -4) และเข้าใกล้ 0 เมื่อ θ = (4, 4) ยังพบอีกว่า ระนาบรายการคำตอบที่มีค่าเป็น 1 ตัดกับระนาบของรายการคำตอบที่มีค่าเป็น 2 ที่ θ_2 = 1 และระนาบรายการคำตอบที่มีค่าเป็น 2 พบว่า เมื่อ θ_1 มีค่ามากกว่า -1 และ θ_2 มากกว่า 1 ความน่าจะเป็นจะมีค่าเป็นเท่ากับ 1



ภาพที่ 10 ระนาบการตอบสนองข้อคำถาม ของโมเดล MPC (Reckase, 2009)

นอกจากนี้ยังสามารถแสดงระนาบคะแนนที่คาดหวัง จากโมเดล MPC ซึ่งเป็นผลรวมของระดับคะแนนการตอบคูณด้วยความน่าจะเป็นของคะแนนการตอบในแต่ละระดับรายการตอบ ซึ่งระนาบคะแนนที่คาดหวัง แสดงดังภาพที่ 11



ภาพที่ 11 ระนาบคะแนนคาดหวังของรายการคำตอบ ของโมเดล MPC (Reckase, 2009)

3. Multidimensional graded response model (MGRM)

Muraki and Carlson (1993) ได้พัฒนา MGRM มาจาก GRM ซึ่งเป็นโมเดลการตอบสนองข้อสอบแบบเอกมิติ และใช้ฟังก์ชันการตอบสนองข้อคำถามด้วยโมเดลปกติสะสมเหมือนกับโมเดล

การตอบสนองแบบเอกมิติ โดยโมเดลการตอบสนองข้อสอบแบบพหุมิติเชื่อว่า การที่จะบรรลุเป้าหมายจากการตอบในขั้นที่ k ได้ จะต้องสำเร็จในขั้นที่ $k-1$ เสียก่อน ซึ่งทำให้โมเดล MGRM มีความเหมาะสมกับเครื่องมือวัดแบบมาตรประมาณค่า ยกตัวอย่างเช่น ถ้าเราเลือกรายการของมาตรประมาณค่าในการทำงาน คือ 2 ชั่วโมง แสดงว่า เวลาในการทำงานดำเนินการผ่านเวลาที่น้อยกว่า 2 ชั่วโมงมาแล้ว เป็นต้น

ค่าพารามิเตอร์ของรายการคำตอบของข้อคำถามที่ i มีค่าต่ำสุดเป็น 0 และคะแนนสูงสุดเป็น m_i ซึ่งความน่าจะเป็นของรายการคำตอบระดับที่ k หรือมากกว่า อนุมานว่า จะมีค่าเพิ่มขึ้นทางเดียว (Monotonically) เมื่อความสามารถ (θ) ในแต่ละมิติมีค่าเพิ่มขึ้น

ซึ่งสมมูลกับโมเดลการตรวจให้คะแนนแบบ 2 ค่า คือ คะแนน k หรือมากกว่า k ให้มีค่าเป็น 1 และต่ำกว่า k มีค่าเป็น 0 โดยผลที่ได้จะสอดคล้องกับโมเดลการให้คะแนนแบบ 2 ค่า ซึ่งความน่าจะเป็นของการตอบระดับ k หรือมากกว่า สามารถคำนวณได้จากโมเดลปกติสะสมแบบสองพารามิเตอร์ (Two-parameter normal ogive model) กับพารามิเตอร์ของผู้ตอบที่แสดงด้วยสมการเชิงเส้นตรงระหว่างเวกเตอร์ความสามารถ (θ-vector) กับพารามิเตอร์อำนาจจำแนก

ความน่าจะเป็นของคะแนนที่ k มีค่าเท่ากับผลต่างระหว่างความน่าจะเป็นของความสำเร็จในขั้นที่ k หรือสูงกว่า กับความน่าจะเป็นของความสำเร็จในขั้นที่ $k+1$ หรือมากกว่าที่กำหนดโดยระดับความสามารถ (θ) คือ $P^*\left(u_{ij}=k\mid\theta_j\right)$ สามารถแสดงด้วยสมการที่ 11

$$P\left(u_{ij}=k\mid\theta_j\right)=P^*\left(u_{ij}=k\mid\theta_j\right)-P^*\left(u_{ij}=k+1\mid\theta_j\right) \tag{11}$$

โดยที่ $P^*\left(u_{ij}=k\mid\theta_j\right)=1$ เพราะการตอบในขั้นที่ 0 หรือมากกว่านั้น เป็นการตอบที่สมบูรณ์ที่สุดจึงมีค่าเป็น 1 และ $P^*\left(u_{ij}=m_i+1\mid\theta_j\right)=0$ เพราะว่าเป็นไปไม่ได้ที่จะตอบขั้นที่มากกว่าระดับคะแนน m_i สำหรับความน่าจะเป็นในการตอบ สามารถแสดงได้ในสมการที่ 11 ซึ่ง Samejima (1969) ได้กำหนดสัญลักษณ์ของเทอมด้านขวาแทนฟังก์ชันความน่าจะเป็นของรายการคำตอบแบบสะสม (Cumulative category) และฟังก์ชันด้านซ้ายแสดงฟังก์ชันการตอบสนองข้อคำถามแต่ละรายการคำตอบ

โมเดลปกติสะสม สำหรับ Graded response สามารถแสดงได้ในสมการที่ 12

$$P\left(u_{ij}=\theta_j\right)=\frac{1}{\sqrt{2\pi}}\int_{\mathbf{a_i'\theta_j}+d_{i,k+1}}^{\mathbf{a_i'\theta_j}+d_{ik}} e^{\frac{-t^2}{2}}dt \tag{12}$$

เมื่อ k คือ ผลการตอบข้อคำถามที่ i โดยที่ $k = 0,1,\ldots,m_i$

$\mathbf{a_i}$ คือ เวกเตอร์พารามิเตอร์อำนาจจำแนกของข้อคำถามที่ i

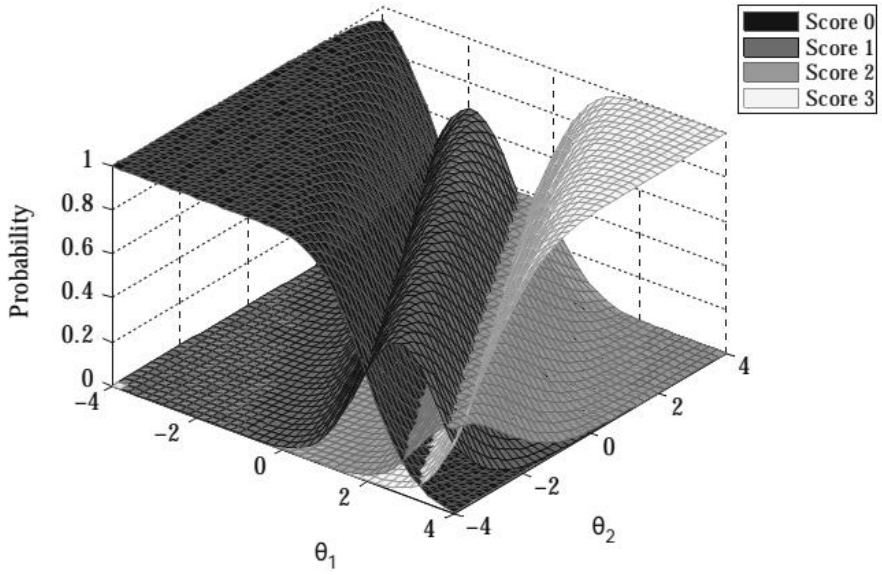
d_{ik} คือ ค่าพารามิเตอร์ของผู้ตอบกับรายการคำตอบที่ k ของข้อที่ i

จะเห็นได้ว่า ค่าพารามิเตอร์ d_{ik} มีค่าสูงมากทางบวก เมื่อคะแนนของรายการคำตอบที่มีคะแนนน้อย และมีค่ามากทางลบ เมื่อมีคะแนนรายการคำตอบมาก ซึ่งค่า d_{ik} จะมีความสัมพันธ์ผกผัน (Inverse relationship) กับคะแนนของรายการคำตอบของข้อคำถามที่ i เช่น รายการคำตอบ 0 หรือ $d_{i0} = \infty$ และเมื่อรายการคำตอบเป็น $m_i + 1$ ค่า $d_{i,m+1} = -\infty$ ส่วนค่า d_{ik} ที่มีค่า $k = 1$ ถึง m_i จึงจะใช้วิธีการประมาณค่าความน่าจะเป็นของรายการคำตอบที่ k ซึ่งสามารถคำนวณได้จากผลต่างของการอินทิเกรต (Integral) ดังสมการที่ 13 ดังนี้

$$P(u_{ij} = K \mid \boldsymbol{\theta}_\mathbf{j}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\mathbf{a_i'\theta_j} + d_{ik}} e^{\frac{-t^2}{z}} dt - \int_{-\infty}^{\mathbf{a_i'\theta_j} + d_{ik+1}} e^{\frac{-t^2}{2}} dt \qquad (13)$$

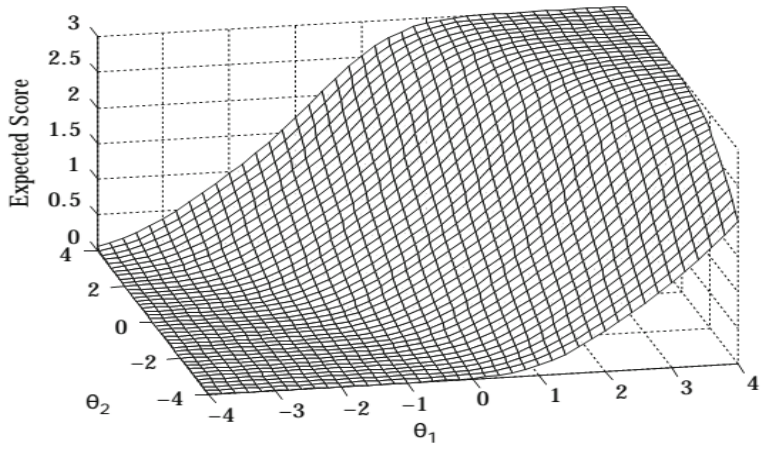
โมเดลที่นำเสนอนี้ทำให้เกิดความชัดเจนโดยใช้ค่าความแตกต่างระหว่างคะแนนการสอบสองค่า โดยใช้โมเดลปกติสะสม สำหรับการอธิบายโอกาสความน่าจะเป็นของการตอบข้อคำถามรายการคำตอบที่ k คำนวณจากผลต่างระหว่างความน่าจะเป็นของรายการคำตอบที่ k หรือมากกว่า k กับระดับ k+1 หรือมากกว่า k+1

กราฟแสดงระนาบของรายการคำตอบสำหรับแบทดสอบที่มีรายการ 4 ค่า ที่มีการให้คะแนน 0, 1, 2 และ 3 แสดง ดังภาพ ที่ 12 โดยข้อคำถามนี้มีค่าพารามิเตอร์ $a_{i1} = 1.2, a_{i2} = .7, d_{i1} = .5, d_{i2} = -1.5$ และ $d_{i3} = -2.5$



ภาพที่ 12 ระนาบของการตอบแต่ละรายการคำตอบ 4 รายการ โดยโมเดล MGRM (Reckase, 2009)

จากภาพแสดงให้เห็นว่า เมื่อ θ_s มีค่าเพิ่มขึ้น ความน่าจะเป็นของคะแนน 0 จะมีค่าลดลง แต่ถ้าคะแนนมีค่าเป็น 3 จะมีค่าเพิ่มขึ้น และความน่าจะเป็นของคะแนนที่ 1 และ 2 มีค่าเพิ่มขึ้น จากนั้นจะมีค่าลดลงเมื่อ ค่า θ_s ที่เพิ่มขึ้น นอกจากผลการตอบในภาพที่ 12 ยังสามารถแสดงระนาบของค่าคาดหวังคะแนนการตอบของโมเดล MGRM ได้ดังภาพที่ 13



ภาพที่ 13 ระนาบค่าคาดหวังคะแนนการตอบของข้อคำถาม ด้วยโมเดล MGRM (Reckase, 2009)

ตอนที่ 2 แนวคิดและหลักการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม

นักวิจัยตั้งแต่อดีตจนถึงปัจจุบันได้พยายามคิดค้น ดัดแปลง ปรับปรุง และพัฒนาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม เพื่อให้ผลการตรวจสอบมีความถูกต้องและแม่นยำมากที่สุด เริ่มต้นจากฐานคิดของการตรวจสอบในข้อคำถามที่วัดความสามารถเอกมิติ (Unidimensional) และเป็นข้อคำถามที่ให้คะแนนสองค่า (Dichotomous) จนถึงข้อคำถามที่ให้คะแนนหลายค่า (Polytomous) ต่อมามีการพัฒนาโมเดลการวัดความสามารถพหุมิติ (Multidimensional) ประกอบกับความเจริญก้าวหน้าของการพัฒนาโปรแกรมคอมพิวเตอร์ที่ใช้ในการคำนวณ ทำให้นักวิจัยสนใจพัฒนาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามวัดความสามารถพหุมิติ ซึ่งถือว่ามีความสมบูรณ์ในการวัดมากกว่า

แนวคิดในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม สำหรับข้อคำถามที่ให้คะแนนสองค่า แบ่งออกเป็นสองกลุ่มใหญ่ ๆ คือ กลุ่มที่ใช้ทฤษฎีการทดสอบแบบมาตรฐานเดิม (Classical test theory method) และกลุ่มที่ใช้แนวคิดทฤษฎีการตอบสนองข้อคำถาม (Item response theory method) กลุ่มวิธีที่ใช้ทฤษฎีการทดสอบแบบมาตรฐานเดิม (CTT method) จะใช้คะแนนรวมของแบบทดสอบเป็นตัวแทนของระดับความสามารถของผู้สอบเพื่อเป็นตัวแปรการจับคู่ (Matching variable) ระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มสนใจ วิธีที่สำคัญในกลุ่มนี้ เช่น วิธีการวิเคราะห์ความแปรปรวน (Analysis of variance: ANOVA) (Cardall & Coffman, 1964 cited in Angoff, 1993) วิธีแปลงค่าความยากของข้อคำถาม (Transformed item difficulty: TID) (Angoff, 1972 cited in Angoff, 1993) วิธีไค-กำลังสอง (Scheuneman, 1979; 1981) วิธีล็อก-ลิเนียร์ (Log-linear) (Mellenbergh, 1982) วิธีแมนเทล-แฮนส์เซล (Mantel-Heanszel: MH) (Holland & Thayer, 1988) วิธีการทำให้เป็นมาตรฐาน (Standardization: STD) (Dorans & Kulick, 1986) วิธีการถดถอยโลจิสติก (Logistic regression: LR) (Swaminathan & Rogers, 1990) เป็นต้น

สำหรับกลุ่มวิธีที่ใช้ทฤษฎีการตอบสนองข้อคำถาม (Item response methods) จะใช้ค่าประมาณความสามารถของผู้สอบซึ่งเป็นตัวแปรแฝง (Latent variable) เป็นเกณฑ์การจับคู่ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในกลุ่มนี้ สามารถจำแนกออกเป็นสองกลุ่ม คือ กลุ่มแรก ใช้ค่าประมาณพารามิเตอร์ของข้อคำถามระหว่างกลุ่มผู้สอบ ได้แก่ วิธีเปลี่ยนค่าความยากของ Hulin, Drasgow, & Parsons (1983) วิธีไค-กำลังสองของลอร์ด (Lord, 1980) และวิธีอัตราส่วนความน่าจะเป็นของ Thissen et al. (1993) ส่วนอีกกลุ่มหนึ่งใช้ค่าประมาณฟังก์ชันการตอบสนองข้อคำถามระหว่างกลุ่มผู้สอบ โดยการวัดพื้นที่ของฟังก์ชันการตอบสนองข้อคำถาม ได้แก่ วิธีการวัดพื้นที่ของ Rudner (1977) วิธีการวัดพื้นที่ของ Linn and Hamisch (1981) วิธีการวัดพื้นที่ของ

Shepard et al. (1984) วิธีการวัดพื้นที่ของ Raju (1990) วิธีการวัดพื้นที่ของ Kim and Cohen (1991) และวิธีดีเอฟไอที (Raju, van der Linden, & Fleer, 1995) เป็นต้น

การใช้ทฤษฎีการทดสอบแบบมาตรฐานเดิมมีข้อได้เปรียบหลายประการ เช่น สามารถนำไปใช้ในเชิงปฏิบัติ มีกระบวนการตรวจสอบที่ง่าย ไม่จำเป็นต้องใช้โปรแกรมคอมพิวเตอร์ที่ซับซ้อน แปลผลเข้าใจง่าย ใช้ตัวอย่างขนาดเล็ก ทำให้ประหยัดเวลา และเสียค่าใช้จ่ายไม่มาก แต่การตรวจสอบในกลุ่มนี้มีข้อจำกัดคือ ค่าความยากและอำนาจจำแนกของข้อคำถามมีค่าแปรเปลี่ยนไปตามกลุ่มผู้สอบ (Camilli & Shepard, 1994) ดังนั้น จึงทำให้การตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามเกิดความคลาดเคลื่อนประเภทที่ 1 และ 2 ปัญหาดังกล่าวสามารถแก้ไขได้โดยใช้ทฤษฎีการตอบสนองข้อคำถาม เนื่องจากคุณสมบัติความไม่แปรเปลี่ยน (Invariance) ของพารามิเตอร์ของข้อคำถาม (Hambleton et al., 1991) โดยที่ค่าประมาณพารามิเตอร์ของข้อคำถามจะไม่ขึ้นกับการแจกแจงความสามารถของกลุ่มผู้สอบ ดังนั้น ฟังก์ชันการตอบข้อคำถามจะมีค่าเท่ากันเมื่อผู้สอบมีระดับความสามารถเท่ากัน โดยไม่คำนึงผู้สอบมาจากกลุ่มใด ทำให้สามารถเปรียบเทียบผลการตอบข้อคำถามที่ระดับความสามารถเดียวกันได้ ซึ่งเป็นหลักการสำคัญของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม วิธีการตรวจสอบภายใต้ทฤษฎีการตอบสนองข้อคำถามในปัจจุบันที่เป็นที่นิยม คือ วิธีไคกำลังสองของลอร์ด วิธีอัตราส่วนความน่าจะเป็น วิธีการวัดพื้นที่ของราชู และวิธีดีเอฟไอที (Raju & Ellis, 2002) ทั้ง 4 วิธี สามารถประยุกต์ใช้ภายใต้โมเดลแบบหนึ่ง สอง หรือสามพารามิเตอร์ ส่วนข้อจำกัดของวิธีดังกล่าว คือ มีรูปแบบพาราเมตริก (Parametric form) ซึ่งจำเป็นต้องมีข้อตกลงเกี่ยวกับโมเดลที่ใช้ประมาณค่าพารามิเตอร์ ก่อนที่จะวิเคราะห์การทำหน้าที่ต่างกันของข้อคำถามต้องแน่ใจว่าค่าพารามิเตอร์ของข้อคำถามระหว่างกลุ่มผู้สอบอยู่บนมาตรเดียวกัน และใช้ตัวอย่างขนาดใหญ่ เพื่อให้การประมาณค่าพารามิเตอร์มีความแม่นยำ โดยเฉพาะโมเดลสอง หรือสามพารามิเตอร์ (Clauser & Mazor, 1998)

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามที่กล่าวมา เป็นวิธีการตรวจสอบภายใต้ข้อคำถามที่ให้คะแนนสองค่า ต่อมาได้มีการปรับขยายสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามที่ให้คะแนนหลายค่า เช่น วิธีแมนเทล (Mantel) วิธีแมนเทล-แฮนส์เซลทั่วไป (Generalized Mantel-Haenszel: GMH) (Zwick, Donoghue, & Grima, 1993) วิธีการทำให้เป็นมาตรฐานแบบให้คะแนนหลายค่า (Polytomous STND) (Potenza & Dorans, 1995) วิธีการฟังก์ชันการจำแนกโลจิสติก (Logistic discriminant function analysis) (Miller & Spray, 1993) วิธีการถดถอยโลจิสติกแบบจัดอันดับ (Ordinal logistic regression: OLR) (Zumbo, 1999) วิธีการเหล่านี้อยู่ในกลุ่มวิธีที่ใช้ทฤษฎีการทดสอบมาตรฐานเดิม สำหรับกลุ่มวิธีที่ใช้ทฤษฎีการตอบสนองข้อคำถาม เช่น วิธีการวัดพื้นที่ (Cohen, Kim, & Baker, 1993) วิธีโพลีซิปเทสท์ (Poly-SIBTEST)

(Chang et al., 1996) วิธีดีเอฟไอที (DFIT) (Flowers et al., 1999) เป็นต้น นอกจากนี้ยังมีการปรับขยายวิธีการตรวจสอบสำหรับข้อคำถามที่วัดความสามารถพหุมิติ เช่น วิธีมัลติซิป (MULTISIB) (Stout et al., 1997) วิธีดีเอฟไอที (DFIT) (Oshima et al., 1997) วิธีดถดถอยโลจิสติก (Mazor, Hambleton, & Clauser, 1998)และ วิธีการวิเคราะห์องค์ประกอบ (Factor analysis) เป็นต้น นอกจากนี้ยังมี นักวิจัยหลายท่านได้นำแนวคิดของโมเดลการวิเคราะห์สมการโครงสร้าง (Structure equation model: SEM) มาใช้สำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม เช่น วิธี Multiple-indicator multiple-cause: MIMIC (Muthén, 1989) Mean and covariance structure model: MACS (Sörbom, 1974) และ Multi group confirmatory factor analysis: MG-CFA (Wu & Lei, 2009) เป็นต้น

กรอบแนวคิดในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม

การทำหน้าที่ต่างกันของข้อคำถาม หมายถึง การที่ผู้สอบมาจากกลุ่มที่แตกต่างกันและมีการจับคู่ความสามารถตามที่แบบทดสอบหรือข้อคำถามต้องการวัดเท่ากัน มีความน่าจะเป็นในการตอบข้อคำถามได้ถูกต้องไม่เท่ากัน โดยทั่วไปแล้ว วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามจะอยู่บนฐานแนวคิดของทฤษฎีการทดสอบแบบมาตรฐานเดิม และทฤษฎีการตอบสนองข้อคำถาม ทั้งสองแนวคิดจะเปรียบเทียบผลการตอบข้อคำถามระหว่างกลุ่มผู้สอบสองกลุ่มหรือมากกว่าสองกลุ่ม บนเงื่อนไขของการจับคู่ความสามารถของผู้สอบ โดยกำหนดให้ผู้สอบกลุ่มหนึ่งเป็นกลุ่มสนใจ และอีกกลุ่มหนึ่งเป็นกลุ่มอ้างอิง ผู้สอบกลุ่มแรกจะเป็นกลุ่มเป้าหมายในการศึกษาการทำหน้าที่ต่างกันของข้อคำถาม ส่วนผู้สอบกลุ่มหลังเป็นเป็นกลุ่มเปรียบเทียบกับกลุ่มสนใจ ถ้าข้อคำถามทำหน้าที่ต่างกันแล้ว ความน่าจะเป็นในการตอบข้อคำถามถูกของผู้สอบทั้งสองกลุ่มไม่เท่ากัน โดยคาดว่าข้อคำถามเข้าข้างผู้สอบกลุ่มอ้างอิง ทำให้ได้เปรียบในการตอบข้อคำถาม สำหรับเกณฑ์ที่ใช้ในการจำแนกผู้สอบกลุ่มอ้างอิงและกลุ่มสนใจมีหลายลักษณะ เช่น เชื้อชาติ ศาสนา ภาษา เพศ วัฒนธรรม ประสบการณ์ เป็นต้น

ข้อคำถามที่ใช้ในการวิเคราะห์การทำหน้าที่ต่างกัน เรียกว่า “ข้อคำถามศึกษา (Studies items)” ส่วนการเปรียบเทียบกลุ่มผู้สอบที่มีความสามารถระดับเดียวกันเรียกว่า “ตัวแปรการจับคู่ (Matching variable)” ความสามารถดังกล่าวเป็นความสามารถเป้าหมายตามที่ต้องการวัด สาเหตุที่ต้องกำหนดเป็นความสามารถเป้าหมาย เนื่องจากในการพัฒนาแบบทดสอบที่ใช้วัดทักษะ ความรู้ความสามารถอย่างสมบูรณ์จะต้องจำแนกลักษณะของเนื้อหาเฉพาะที่จะวัด ถ้าความแตกต่างของผลการตอบข้อคำถามถูกขึ้นอยู่กับความสามารถอื่นมากกว่า จะเป็นสาเหตุที่ทำให้ข้อคำถามทำหน้าที่ต่างกัน (Clauser & Mazor, 1998) ตัวอย่างเช่น แบบทดสอบวิชาคณิตศาสตร์ การตอบข้อคำถามได้ถูกต้องขึ้นอยู่กับความสามารถในการคิดคำนวณ และความสามารถในการอ่าน

ถ้าให้ความสามารถในการคิดคำนวณเป็นความสามารถเป้าหมายที่ต้องการวัดของข้อคำถาม ซึ่งเป็นความสามารถที่ใช้ในการจับคู่ และถ้าผู้สอบกลุ่มหนึ่งมีความสามารถในการอ่านมากกว่าผู้สอบอีกกลุ่มหนึ่ง ความแตกต่างของผลการตอบข้อคำถามระหว่างกลุ่มจะแสดงว่า “ข้อคำถามทำหน้าที่ต่างกัน” แต่ถ้าความแตกต่างของผลการตอบข้อคำถามระหว่างกลุ่มผู้สอบเกิดขึ้นจากการแจกแจงความสามารถของผู้สอบทั้งหมด ซึ่งไม่ได้มีการควบคุมตัวแปรดังกล่าว จะเรียกว่าผลกระทบของข้อคำถาม (Camilli, 1992; Clauser & Mazor, 1998; Dorans & Holland, 1993; Zumbo, 1999)

จากแนวคิดดังกล่าว จึงทำให้นักวิจัยให้ความสำคัญกับมิติ (Dimension) ของการวัด โดยเชื่อว่าการทำหน้าที่ต่างกันของข้อคำถามมีอิทธิพลของมิติในการวัดเข้ามาเกี่ยวข้อง เช่น ในแบบทดสอบที่วัดความสามารถแบบสองมิติ ประกอบด้วยมิติหลัก (Primary dimension) และมิติรอง (Secondary dimension) ในมิติหลักเป็นความสามารถที่ข้อคำถามต้องการวัด เรียกว่าความสามารถเป้าหมาย (Target ability) แทนด้วยสัญลักษณ์ θ ส่วนมิติรองอาจมีหลายความสามารถ ถ้าความสามารถนั้นเป็นความสามารถที่ข้อคำถามต้องการวัด เรียกว่า “ความสามารถเสริม (Auxiliary ability)” แต่ถ้าเป็นความสามารถนั้น เป็นความสามารถที่ข้อคำถามไม่ต้องการวัด เรียกว่า “ความสามารถแทรกซ้อน (Nuisance ability)” แทนด้วยสัญลักษณ์ η (Ackerman, 1992; Nandakumar,1993; Roussos & Stout, 1996b; Shealy & Stout, 1993) ข้อคำถามที่วัดความสามารถนอกเหนือมิติหลักอย่างน้อยหนึ่งมิติเป็นสาเหตุที่ทำให้ข้อคำถามทำหน้าที่ต่างกัน (Cronbach, 1990; Dorans & Schmitt, 1989; Lord, 1980; Roussos & Stout, 1996) ดังนั้น ความสามารถในมิติรองเป็นสาเหตุที่ทำให้ข้อคำถามทำหน้าที่ต่างกัน นักวิจัยจึงคิดค้นและพัฒนาโมเดลการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม โดยสามารถจำแนกกรอบแนวคิดเป็นสองแนวใหญ่ ๆ ซึ่งมีรายละเอียด ดังนี้

1. การทำหน้าที่ต่างกันของข้อคำถามแบบมิติเดียว (Unidimensional DIF)

การตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามโดยใช้ทฤษฎีการตอบสนองข้อคำถามในยุคเริ่มต้นได้พัฒนาวิธีการตรวจสอบภายใต้โมเดลแบบเอกมิติ โดยพัฒนาภายใต้กรอบแนวคิดสองแนวทางใหญ่ ๆ คือ ตรวจสอบด้วยการเปรียบเทียบความแตกต่างของค่าพารามิเตอร์ข้อคำถามระหว่างกลุ่มผู้สอบที่มีความสามารถระดับเดียวกัน ส่วนอีกกรอบแนวคิดหนึ่งจะเปรียบเทียบความแตกต่างระหว่างพื้นที่โค้งลักษณะข้อคำถาม (Item characteristic curves: ICCs) หรือ ฟังก์ชันการตอบข้อคำถาม (Item response functions: IRFs) ระหว่างกลุ่มผู้สอบที่มีความสามารถระดับเดียวกัน ในการเปรียบเทียบจะใช้โมเดลโลจิสติกแบบหนึ่ง สอง หรือสามพารามิเตอร์ สำหรับโมเดลโลจิสติกแบบสามพารามิเตอร์ สามารถคำนวณได้ดังนี้

$$P(U_{ij}=k \mid \boldsymbol{\theta}_j)=c_i+(1-c_i)\frac{e^{1.702a_i(\boldsymbol{\theta}_j-b_i)}}{1+e^{1.702a_i(\boldsymbol{\theta}_j-b_i)}} \quad (14)$$

การเปรียบเทียบความแตกต่างของค่าพารามิเตอร์ข้อคำถามระหว่างกลุ่ม ในยุคเริ่มต้น Wright, Mead and Draba (1976 cited in Hulin et al., 1983) ได้เสนอใช้ สถิติดราบา (Draba statistic) ทดสอบความแตกต่างของค่าพารามิเตอร์ความยากของข้อคำถามระหว่างกลุ่มผู้สอบภายใต้โมเดลของราส์ช (Rasch model) โดยใช้สูตร ดังนี้

$$Z=\frac{\hat{b}_{iR}-\hat{b}_{iF}}{\sqrt{\left(SE\left(\hat{b}_{iR}\right)\right)^2+\left(SE\left(\hat{b}_{iF}\right)\right)^2}} \quad (15)$$

เมื่อ $\hat{b}_{iR}$ และ $\hat{b}_{iF}$ แทนค่าประมาณความยากของข้อคำถามที่ i จากผู้สอบกลุ่มอ้างอิงและกลุ่มสนใจ ตามลำดับ $SE\left(\hat{b}_{iR}\right)$ และ $SE\left(\hat{b}_{iF}\right)$ แทนค่าความคลาดเคลื่อนมาตรฐานของค่าประมาณความยากของข้อคำถาม จากผู้สอบกลุ่มอ้างอิงและกลุ่มสนใจ ตามลำดับ ต่อมา Lord (1980) ได้ใช้สถิติโฮลเทลลิง (Holtelling statistics) ทดสอบความแตกต่างของค่าพารามิเตอร์ของข้อคำถามภายใต้โมเดลสามพารามิเตอร์ ดังนี้

$$x^2=v_i'\Sigma_i^{-1}v_i \quad (16)$$

เมื่อ $v_i'=\left[\hat{b}_{iR}-\hat{b}_{iF},\hat{a}_{iR}-\hat{a}_{iF}\right]$ และ $\Sigma_i^{-1}v_i$ แทน เมทริกซ์ค่าประมาณความแปรปรวน-ความแปรปรวนร่วม (Variance-Covariance matrix) ของ $\hat{b}_{iR}-\hat{b}_{iF}$ และ $\hat{a}_{iR}-\hat{a}_{iF}$ วิธีนี้มักนิยมเรียกว่า “วิธีการทดสอบไค-กำลังสองของลอร์ด” ส่วนการตรวจสอบอีกวิธีหนึ่งจะใช้ “สถิติอัตราส่วนความน่าจะเป็น” (Likelihood ratio statistic) เปรียบเทียบความแตกต่างของค่าพารามิเตอร์ระหว่างกลุ่ม (Thissen et al., 1993) โดยทดสอบความเหมาะสมระหว่างโมเดลสองโมเดล คือ โมเดลออกเมนท์ (Augmented model: A) ประกอบด้วยพารามิเตอร์ของข้อคำถามระหว่างกลุ่มผู้สอบที่มีค่าแตกต่างกัน และโมเดลคอมแพคท์ (Compact model: C) ประกอบด้วยพารามิเตอร์ของข้อคำถามระหว่างกลุ่มผู้สอบที่มีค่าเท่ากัน สำหรับสถิติอัตราส่วนความน่าจะเป็นมีลักษณะ ดังนี้

$$G^2(df)=2\log\left[\frac{likelihood[A]}{likelihood[C]}\right] \quad (17)$$

ส่วนกรอบแนวคิดในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม โดยการวัดพื้นที่ของฟังก์ชันการตอบสนองข้อคำถามจะเปรียบเทียบผลการตอบข้อคำถามโดยตรง มีนักวิจัยหลายคนได้พัฒนาสูตรการวัดพื้นที่ เช่น วิธีการวัดพื้นที่ของ Rudner (1977) วิธีการวัดพื้นที่ของ Linn and Hamisch (1981) วิธีการวัดพื้นที่ของ Shepard et al. (1984) วิธีการวัดพื้นที่ของ Raju (1990) วิธีการวัดพื้นที่ของ Kim and Cohen (1991) เป็นต้น สำหรับวิธีการวัดพื้นที่แต่ละสูตรดังกล่าว วิธีการวัดพื้นที่ของ Raju มีความโดดเด่นมากที่สุด เนื่องจากสามารถคำนวณพื้นที่โดยการหาปริพันธ์แบบต่อเนื่อง (Continuous integration) ซึ่งเป็นวิธีการทางคณิตศาสตร์ พร้อมทั้งใช้สถิติทดสอบนัยสำคัญ ทำให้ผลการตรวจสอบมีความถูกต้องและแม่นยำสูง ราชูได้พัฒนาสูตรการวัดพื้นที่ในช่วงเปิดแบบคิดเครื่องหมาย (Exact signed area: ESA) และแบบไม่คิดเครื่องหมาย (Exact unsigned area: EUA) ภายใต้โมเดลหนึ่ง สอง หรือสามพารามิเตอร์ สูตรทั่วไปในการคำนวณพื้นที่ของ Raju เป็นดังนี้

$$ESA = \int_{-\infty}^{+\infty} [P_R(\theta) - P_F(\theta)] d\theta \tag{18}$$

$$EUA = \int_{-\infty}^{+\infty} [P_R(\theta) - P_F(\theta)] d\theta \tag{19}$$

เมื่อ $P_R(\theta)$และ $P_F(\theta)$ แทนฟังก์ชันการตอบสนองข้อคำถาม ณ ระดับความสามารถ θ จากกลุ่มอ้างอิงและกลุ่มสนใจ ตามลำดับ ซึ่งคำนวณจากโมเดลในสมการ (14) ส่วนสูตรการวัดพื้นที่ซึ่งนักวิจัยนิยมนำมาใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม หรือนำมาใช้คำนวณขนาดและพื้นที่ของข้อคำถามทำหน้าที่ต่างกันของข้อคำถามเพื่อใช้ในการจำลองข้อมูล คือ สูตรการคำนวณพื้นที่แบบไม่คิดเครื่องหมาย เนื่องจากข้อคำถามทำหน้าที่ต่างกันไม่เป็นรูปแบบเดียวกัน (Non uniform DIF) จึงทำให้ฟังก์ชันการตอบข้อคำถามแต่ละกลุ่มผู้สอบจะตัดกัน ถ้าตัดกันตรงช่วงกึ่งกลางของความสามารถแล้วคำนวณโดยใช้สูตรการวัดพื้นที่แบบคิดเครื่องหมายจะไม่สามารถตรวจพบข้อคำถามที่มีรูปแบบดังกล่าวได้ เพราะว่าทิศทางข้อคำถามทำหน้าที่ต่างกันเป็นเครื่องหมายบวกและเครื่องหมายลบจะหักล้างกัน (Cancellation) สำหรับสูตรการวัดพื้นที่แบบคิดเครื่องหมายภายใต้โมเดลโลจิสติกแบบสามพารามิเตอร์มี 2 กรณี ดังนี้ (Raju, 1990)

$$EUA = (1-C)|\hat{b}_R - \hat{b}_F| \text{ เมื่อ } \hat{a}_R - \hat{a}_F \tag{20}$$

$$EUA = (1-\hat{C})\left|\frac{2(\hat{a}_R - \hat{a}_F)}{D\hat{a}_R\hat{a}_F}\ln\left\{1+\exp\left[\frac{D\hat{a}_R\hat{a}_F\left(\hat{b}_R-\hat{b}_F\right)}{\hat{a}_R-\hat{a}_F}\right]\right\} - \left(\hat{b}_R-\hat{b}_F\right)\right| \quad (21)$$

เมื่อ $(\hat{a}_R \neq \hat{a}_F)$

สูตรการวัดพื้นที่ในกรณีที่ 1 เมื่อ $\hat{a}_R = \hat{a}_F$ ใช้คำนวณพื้นที่ข้อคำถามทำหน้าที่เบี่ยงเบนที่เป็นรูปแบบเดียวกัน โดยกำหนดให้ความแตกต่างของค่าพารามิเตอร์ความยากของข้อคำถามระหว่างกลุ่มอ้างอิงและกลุ่มสนใจมีค่าแปรเปลี่ยน $(\hat{b}_R \neq \hat{b}_F)$ ส่วนค่าพารามิเตอร์อำนาจจำแนกของข้อคำถามระหว่างกลุ่มมีค่าเท่ากัน $(\hat{a}_R = \hat{a}_F)$ และกรณีที่ 2 เมื่อ $\hat{a}_R \neq \hat{a}_F$ ใช้คำนวณพื้นที่ข้อคำถามทำหน้าที่เบี่ยงเบนที่ไม่เป็นรูปแบบเดียวกัน โดยกำหนดให้ความแตกต่างของค่าพารามิเตอร์อำนาจจำแนกของข้อคำถามระหว่างกลุ่มอ้างอิงและกลุ่มสนใจมีค่าแปรเปลี่ยน $(\hat{a}_{iR} \neq \hat{a}_{iF})$ ส่วนค่าพารามิเตอร์ความยากของข้อคำถามระหว่างกลุ่มจะมีค่าเท่ากันหรือแตกต่างกันได้ $(\hat{b}_{iR} = \hat{b}_{iF}$ หรือ $\hat{b}_{iR} \neq \hat{b}_{iF})$ สำหรับค่าพารามิเตอร์การเดาของข้อคำถามทั้งสองมักกำหนดให้มีค่าคงที่ สูตรดังกล่าวสามารถปรับเป็นโมเดลแบบหนึ่ง สอง หรือสามพารามิเตอร์ก็ได้ ขึ้นอยู่กับเงื่อนไขของพารามิเตอร์ที่กำหนด จากแนวคิดและหลักการดังกล่าวนี้ Cohen et al. (1993) ได้ปรับขยายสูตรการวัดพื้นที่ของ Raju เพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามให้คะแนนหลายค่า โดยใช้โมเดลเกรดเรสพอน (Graded response model; GRM)

2. การทำหน้าที่ต่างกันของข้อคำถามพหุมิติ (Multidimensional DIF)

การตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ในข้อคำถามที่ให้คะแนนแบบสองค่าหรือหลายค่า มักมีข้อตกลงเกี่ยวกับความเป็นเอกมิติของแบบทดสอบ เช่น ข้อคำถามวัดความสามารถพหุมิติแต่ให้คะแนนเสมือนวัดความสามารถมิติเดียว จะมีผลทำให้ข้อคำถามทำหน้าที่ต่างกันและผลการทำหน้าที่ต่างกันจะทำให้เกิดความคลาดเคลื่อนประเภทที่ 1 เนื่องจากการเลือกโมเดลผิดพลาด ไม่เหมาะสมกับข้อมูล ดังนั้น เมื่อแบบทดสอบวัดความสามารถหลายมิติก็ควรจะต้องเลือกใช้โมเดลการวัดความสามารถพหุมิติซึ่งจะทำให้จำแนกระดับความสามารถของผู้สอบได้อย่างถูกต้องและแม่นยำ ในปัจจุบันนักวิจัยได้ให้ความสำคัญในประเด็นดังกล่าวมากยิ่งขึ้น ดังจะเห็นได้จากมีการศึกษาเกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในข้อมูลหลายมิติของนักวิจัยหลายคน (Ackerman, 1992; Camilli, 1992; Mazor et al., 1998; Oshima & Miller, 1992; Oshima et al., 1997; Roussos & Stout, 1996; Shealy & Stout, 1993; Stout et al., 1997) ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ที่วัดความสามารถหลายมิติและให้คะแนนสองค่า ภายใต้ทฤษฎีการตอบสนองข้อคำถามแบบหลายมิติ นักวิจัยส่วนมากนิยมใช้โมเดลโลจิสติกแบบสองหรือสามพารามิเตอร์หลายมิติที่มีลักษณะ ดังนี้ (Reckase, 1997)

$$P(U_{ij}=1|\mathbf{a_i},d_i,c_i,\boldsymbol{\theta}_\mathbf{j})=c_i+(1-c_i)\frac{e^{(a_i'\theta_j+d_i)}}{1+e^{(a_i'\theta_j+d_i)}} \quad (22)$$

เมื่อ $P(U_{ij}=1|\mathbf{a_i},d_i,c_i,\boldsymbol{\theta}_\mathbf{j})$ แทน ความน่าจะเป็นของผู้สอบคนที่ j ที่มีเวกเตอร์ของความสามารถ θ จะตอบข้อคำถามข้อที่ i ได้ถูกต้อง

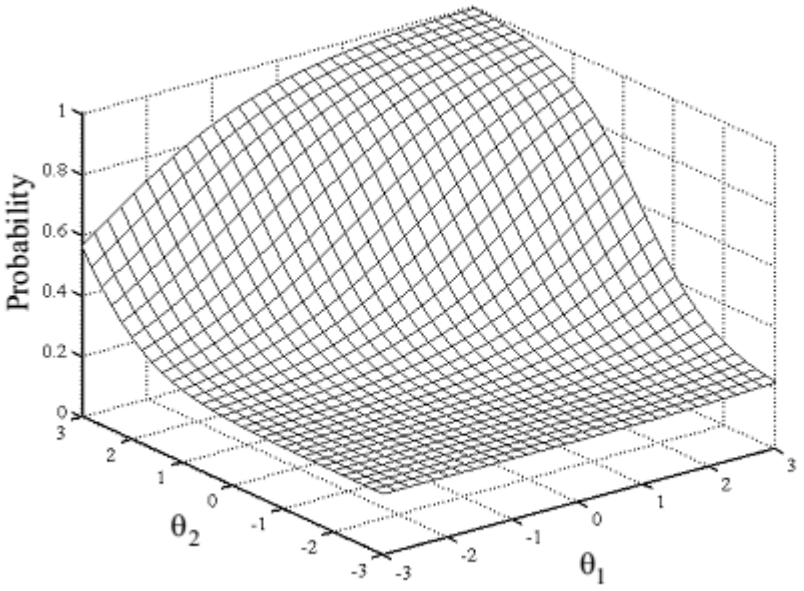
$\mathbf{a_i}$ แทน เวกเตอร์ของพารามิเตอร์อำนาจจำแนกของข้อคำถามข้อที่ i

d_i แทน พารามิเตอร์ความยากของข้อคำถามข้อที่ i

c_i แทน พารามิเตอร์การเดาของข้อคำถามข้อที่ i และ

$\boldsymbol{\theta}_\mathbf{j}$ แทน เวกเตอร์พารามิเตอร์ความสามารถของผู้สอบคนที่ j

สำหรับความน่าจะเป็นในการตอบข้อคำถามดังกล่าว สามารถแสดงด้วยระนาบการตอบสนองข้อคำถาม (Item response surface: IRS) ดังภาพที่ 14



ภาพที่ 14 ระนาบการตอบข้อคำถาม (IRS) ภายใต้โมเดลโลจิสติกแบบพหุมิติ 3 พารามิเตอร์ (M3PL) (Reckase, 1997)

พารามิเตอร์ความยากของข้อคำถามหลายมิติ (Multidimensional item difficulty: MDIFF) สามารถคำนวณได้ ดังนี้

$$MDIFF_i = \frac{d_i}{\sqrt{\sum_{k=1}^{m} a_{ik}^2}} \tag{23}$$

เมื่อ a_{ik} แทนสมาชิกตัวที่ k ของเวกเตอร์ d_i โดยที่พารามิเตอร์ความยากของข้อคำถามหลายมิติ ซึ่งแทนระยะห่างจากจุดกำเนิดของระนาบความสามารถไปยังระนาบการตอบสนองข้อคำถาม (Item response surface: IRS) ณ จุดที่มีความชันมากที่สุด จุดดังกล่าวจะให้สารสนเทศสูงสุดเกี่ยวกับผู้สอบซึ่งถูกวัดด้วยข้อคำถามข้อดังกล่าว เส้นที่เชื่อมระหว่างจุดบนระนาบการตอบสนองข้อคำถามไปยังจุดกำเนิดทำมุมกับมิติความสามารถที่ k เท่ากับ a_{ik} ซึ่งคำนวณขนาดของมุม ดังนี้

$$\alpha_{ik} = \arccos \frac{a_{ik}}{\sqrt{\sum_{k=1}^{m} a_{ik}^2}} \tag{24}$$

ค่าของ α_{ik} แสดงทิศทางข้อคำถาม (Item direction) ซึ่งใช้กำหนดน้ำหนักส่วนประกอบของความสามารถที่วัดโดยข้อคำถาม ทิศทางข้อคำถามแต่ละข้อจะทำมุมกับแกนความสามารถ θ_1 (แกนบวก) ในโมเดลแบบสองมิติ สมมติให้แกนความสามารถ θ_1 แทนข้อคำถามวัดความสามารถ θ_1 และแกนความสามารถ θ_2 แทนข้อคำถามวัดความสามารถ θ_2 แล้ว α_{i1} เท่ากับ 0 องศา ถ้าข้อคำถามวัดเฉพาะความสามารถ θ_2 แล้ว α_{i1} เท่ากับ 90 องศา แต่ถ้าข้อคำถามวัดความสามารถ θ_1 และ θ_2 เท่ากันแล้ว α_{i1} เท่ากับ 45 องศา ค่าของ α_{i1} มีค่าตั้งแต่ 0 ถึง 90 องศา ค่าจะมากหรือน้อยขึ้นอยู่กับข้อคำถามที่วัดความสามารถทั้งสอง สำหรับพารามิเตอร์อำนาจจำแนกของข้อคำถามหลายมิติ (Multidimensional item discrimination: MDISC) สามารถคำนวณ ดังนี้

$$MDISC_i = \sqrt{\sum_{k=1}^{m} a_{ik}^2} \tag{25}$$

พารามิเตอร์อำนาจจำแนกของข้อคำถามหลายมิติถูกกำหนดโดยสมาชิกของเวกเตอร์ a_i สำหรับสมาชิกของเวกเตอร์ a_i ก็คือ พารามิเตอร์อำนาจจำแนกในโมเดลมิติเดียว พารามิเตอร์อำนาจจำแนกของข้อคำถามหลายมิติมีความสัมพันธ์กับพารามิเตอร์ความยากของข้อคำถามหลายมิติ ดังนี้

$$MDISC_i = -\frac{d}{MDIFF_i} \tag{26}$$

Oshima et al. (1997) ได้ใช้กรอบแนวคิดในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ตามทฤษฎีการตอบสนองข้อคำถามเอกมิติมาประยุกต์ใช้ในการตรวจสอบแบบพหุมิติ โดยใช้โมเดลโลจิสติกสองพารามิเตอร์แบบหลายมิติ ที่มีการชดเชยค่าความสามารถ จำลองข้อคำถามโดยใช้โมเดลสองมิติ แล้วตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามด้วยวิธีดีเอฟไอที (DFIT) (Raju et al., 1995) โดยการคำนวณดัชนี 3 ประเภท คือ ดัชนีการทำหน้าที่ต่างกันของข้อคำถามที่มีการชดเชย ดัชนีการทำหน้าที่ต่างกันของข้อคำถามที่ไม่มีการชดเชย และ ดัชนีการทำหน้าที่ต่างกันของแบบทดสอบ (DTF) ดังนี้

$$\hat{CDIF} = \hat{Cov}(\hat{d}, \hat{D}) + \hat{\mu}_{\hat{d}_i}\hat{\mu}_{\hat{D}} \tag{27}$$

$$\hat{NCDIF} = \hat{\sigma}^2_{\hat{d}_i} + \hat{\mu}^2_{\hat{d}_i} \tag{28}$$

$$\hat{DTF} = \hat{\sigma}^2_{\hat{D}_i} + \hat{\mu}^2_{\hat{D}} \tag{29}$$

โดยที่ $\hat{D} = \sum_{i=1}^{k} \hat{d}_i$ และ $\hat{d}_i = \hat{p}_{iF} - \hat{p}_{iR}$

เมื่อ $\hat{p}_{iF}$ และ $\hat{p}_{iR}$ แทนความน่าจะเป็นในการตอบข้อคำถามข้อที่ *i* ถูกต้องสำหรับผู้สอบกลุ่มสนใจและกลุ่มอ้างอิง ตามลำดับ ซึ่งประมาณค่ามาจากสมการที่ 22 [$\hat{d}_i$ ในสมการที่ 27 และ 28 ไม่ใช่ $\hat{d}_i$ จากสมการที่ 22 หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามที่วัดความสามารถพหุมิติ จะเปรียบเทียบฟังก์ชันการตอบข้อคำถามระหว่างกลุ่ม ทำนองเดียวกับการตรวจสอบที่วัดความสามารถเอกมิติ สมมติให้แบบทดสอบมีข้อคำถาม *k* ข้อ และมีพารามิเตอร์ของข้อคำถามสองชุด สำหรับกลุ่มอ้างอิงและกลุ่มสนใจ โดยที่พารามิเตอร์ของข้อคำถามสองชุดดังกล่าว มีมาตรร่วมกัน เมื่อเวกเตอร์ของความสามารถเท่ากันแล้ว $\hat{p}_{iF}$ และ $\hat{p}_{iR}$ มีค่าแตกต่างกัน นั่นแสดงว่า ข้อคำถามทำหน้าที่ต่างกัน ส่วนการศึกษาเกี่ยวกับรูปแบบของข้อคำถามทำหน้าที่ต่างกันภายใต้การจำลองข้อมูลแบบพหุมิติก็มีแนวคิดเช่นเดียวกัน กล่าวคือ ในเงื่อนไขของการจำลองข้อคำถามทำหน้าที่ต่างกันที่เป็นรูปแบบเดียวกัน จะกำหนดให้สเกลาร์ของพารามิเตอร์ d_i ระหว่างกลุ่มผู้สอบมีค่าแตกต่างกัน ในขณะที่เวกเตอร์ของพารามิเตอร์ a_i มีค่าเท่ากันส่วนในเงื่อนไขของข้อคำถามทำหน้าที่ต่างกันที่ไม่เป็นรูปแบบเดียวกัน จะกำหนดให้เวกเตอร์ของพารามิเตอร์ a_i

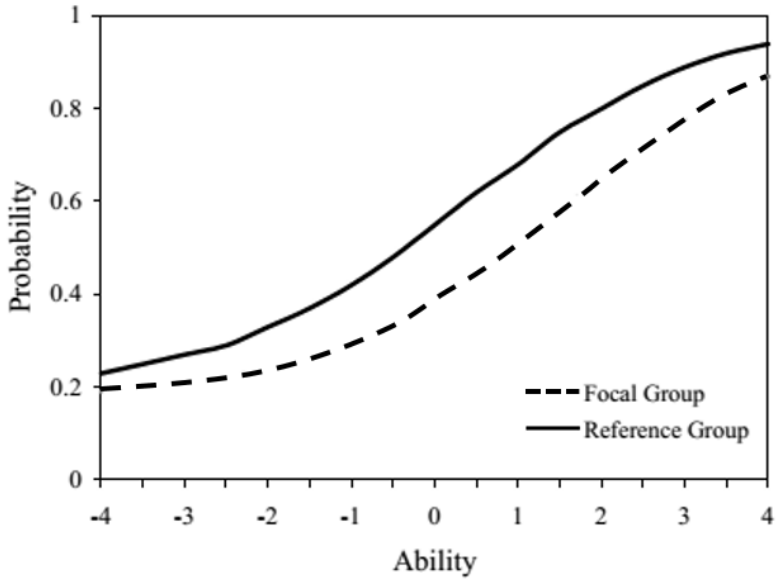
ระหว่างกลุ่มผู้สอบมีค่าแตกต่างกัน ในขณะที่สเกลาร์ของพารามิเตอร์ d_i มีค่าเท่ากันหรือแตกต่างกันก็ได้

จะเห็นได้ว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามโดยใช้ทฤษฎีการตอบสนองข้อคำถามแบบเอกมิติ หรือพหุมิติ ต่างก็อาศัยหลักการความไม่แปรเปลี่ยนของพารามิเตอร์ข้อคำถาม

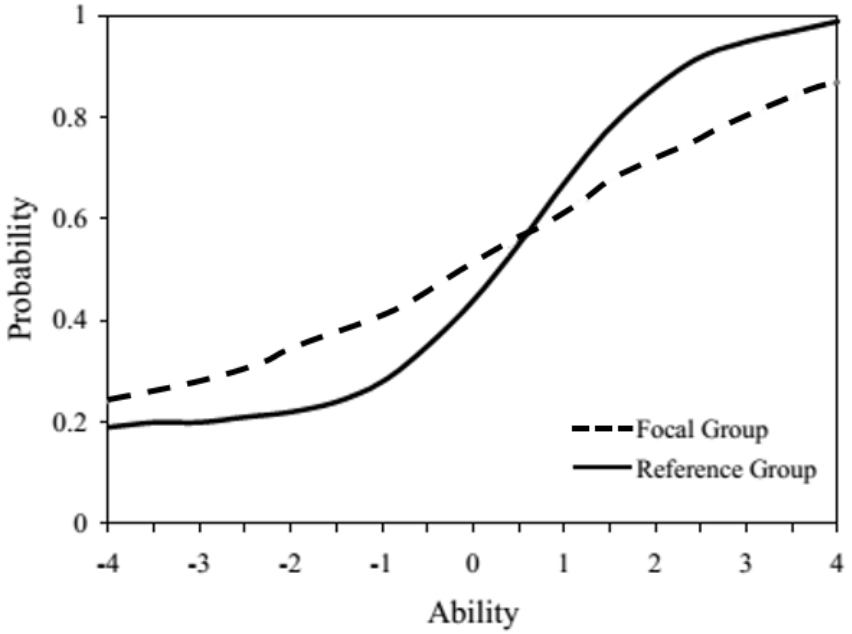
ตอนที่ 3 รูปแบบของการทำหน้าที่ต่างกันของข้อคำถาม

Mellenbergh (1982) ได้ทำการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามให้คะแนนสองค่า โดยใช้โมเดลล็อก-ลิเนียร์ และโมเดลโลจิสติก ทำนายผลการตอบข้อคำถาม ภายใต้ตารางการณ์จรแบบสามมิติ ภายใต้ปัจจัยสามประการ คือ การเป็นสมาชิกของกลุ่ม ระดับความสามารถ และปฏิสัมพันธ์ระหว่างการเป็นสมาชิกของกลุ่มกับระดับความสามารถ การศึกษาครั้งนี้ ได้จำแนกข้อคำถามทำหน้าที่ต่างกันสองรูปแบบ คือ ข้อคำถามทำหน้าที่ต่างกันรูปแบบเดียวกัน (Uniform DIF) และข้อคำถามทำหน้าที่ต่างกันไม่เป็นรูปแบบเดียวกัน (Non uniform DIF) ข้อคำถามทำหน้าที่ต่างกันรูปแบบแรก เกิดขึ้นเมื่อไม่มีปฏิสัมพันธ์ระหว่างระดับความสามารถกับการเป็นสมาชิกของกลุ่ม นั่นแสดงว่า ความน่าจะเป็นในการตอบข้อคำถามถูกของผู้สอบกลุ่มหนึ่งมากกว่าอีกกลุ่มหนึ่งอย่างคงที่ทุกระดับความสามารถ ส่วนข้อคำถามทำหน้าที่ต่างกันแบบหลังเกิดขึ้นเมื่อมีปฏิสัมพันธ์ระหว่างระดับความสามารถกับการเป็นสมาชิกของกลุ่ม แสดงว่า ความแตกต่างของความน่าจะเป็นในการตอบข้อคำถามถูกของกลุ่มไม่เหมือนกันทุกระดับความสามารถ

ทฤษฎีการตอบสนองข้อคำถาม (Item response theory: IRT) สามารถพิจารณาปฏิสัมพันธ์ระหว่างระดับความสามารถกับการเป็นสมาชิกของกลุ่มได้จากความแตกต่างของค่าพารามิเตอร์อำนาจจำแนกของข้อคำถามระหว่างกลุ่มผู้สอบ กล่าวคือ ถ้าค่าพารามิเตอร์อำนาจจำแนกของข้อคำถามระหว่างกลุ่มผู้สอบเท่ากัน แล้วฟังก์ชันการตอบสนองข้อคำถามระหว่างกลุ่มขนานกัน แสดงว่า ข้อคำถามทำหน้าที่ต่างกันรูปแบบเดียวกัน โดยรูปแบบของข้อคำถามทำหน้าที่ต่างกันทั้งสอง แสดงดังภาพที่ 15 และ 16



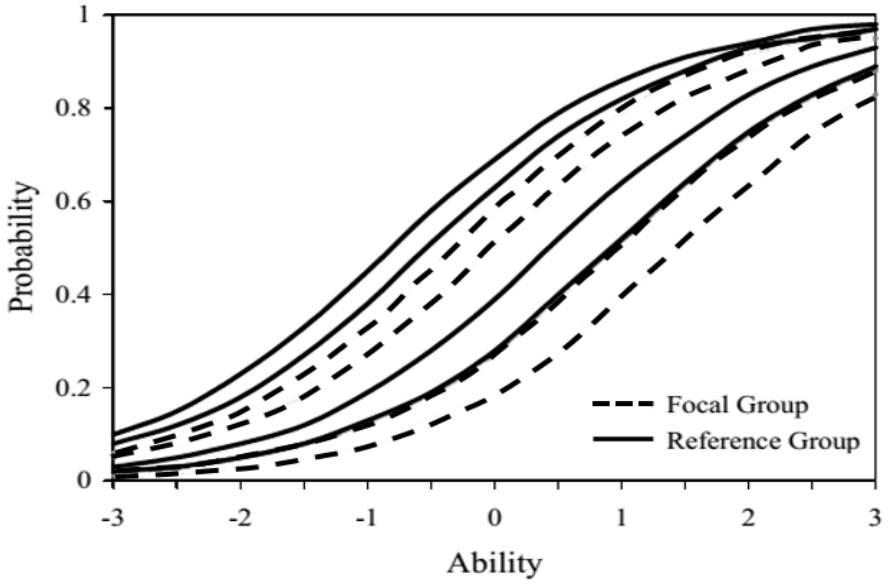
ภาพที่ 15 ข้อคำถามที่ทำหน้าที่ต่างกันที่เป็นรูปแบบเดียวกัน (Uniform DIF) ภายใต้โมเดลโลจิสติกแบบ 3 พารามิเตอร์ (3PL)



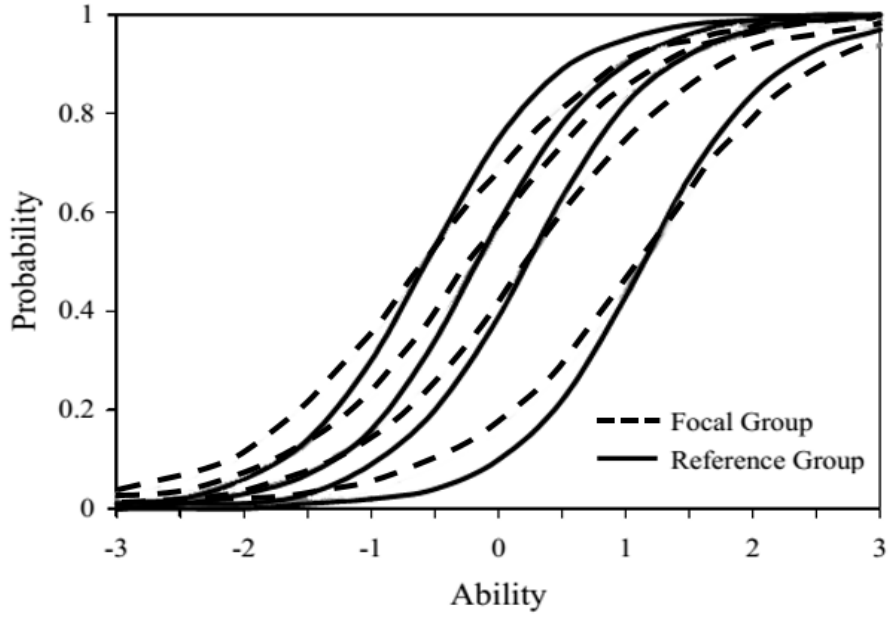
ภาพที่ 16 ข้อคำถามที่ทำหน้าที่ต่างกันไม่เป็นรูปแบบเดียวกัน (Non uniform DIF) ภายใต้โมเดลโลจิสติกแบบ 3 พารามิเตอร์ (3PL)

เมื่อพิจารณานิยามข้อคำถามทำหน้าที่ต่างกันของข้อคำถามไม่เป็นรูปแบบเดียวกัน จะเห็นว่า ฟังก์ชันการตอบข้อคำถามระหว่างกลุ่มผู้สอบซึ่งไม่ขนานกันอาจจะตัดกันหรือไม่ตัดกันก็ได้ จากผลการศึกษาของ Swaminathan and Rogers (1990) พบว่า รูปแบบของข้อคำถามทำหน้าที่ต่างกันไม่เป็นรูปแบบเดียวกัน สามารถจำแนกออกเป็นสองรูปแบบย่อย คือ ช่วงความสามารถตามทฤษฎีการตอบสนองข้อคำถาม เมื่อฟังก์ชันการตอบสนองข้อคำถามระหว่างกลุ่มผู้สอบตัดกันตรงกึ่งกลางของช่วงความสามารถดังกล่าว ข้อคำถามทำหน้าที่ต่างกันไม่เป็นรูปแบบเดียวกันเมื่อวิเคราะห์ด้วยโมเดล ANOVA จะลักษณะ “ปฏิสัมพันธ์ไม่เป็นลำดับ” (Disordinal interaction) เมื่อฟังก์ชันการตอบข้อคำถามระหว่างกลุ่มตัดกันนอกช่วงความสามารถหรือฟังก์ชันการตอบข้อคำถามระหว่างกลุ่มไม่ขนานกัน แต่ไม่ตัดกัน ข้อคำถามทำหน้าที่ต่างกันไม่เป็นรูปแบบเดียวกันจะมีลักษณะ “ปฏิสัมพันธ์เป็นลำดับ” (Ordinal interaction) Li and Stout (1993 cited in Narayanan and Swaminathan, 1996) ได้เรียกรูปแบบของข้อคำถามทั้งสองว่า “ข้อคำถามทำหน้าที่ต่างกันของข้อคำถามไม่มีทิศทางเดียวกัน (Nonunidirectional DIF) และข้อคำถามทำหน้าที่ต่างกันที่มีทิศทางเดียวกัน (Unidirectional DIF) ตามลำดับ เพื่อไม่ให้เกิดความสับสนในการเรียกชื่อ ดังกล่าว Li and Stout, 1996) ได้ใช้คำใหม่ว่า ข้อคำถามทำหน้าที่ต่างกันตัดกัน (Crossing DIF) และข้อคำถามทำหน้าที่ต่างกันแบบไม่ตัดกัน (Non-crossing DIF) แต่คำดังกล่าวไม่เป็นที่นิยมใช้กัน ส่วนมากจะใช้คำเดิมมากกว่า

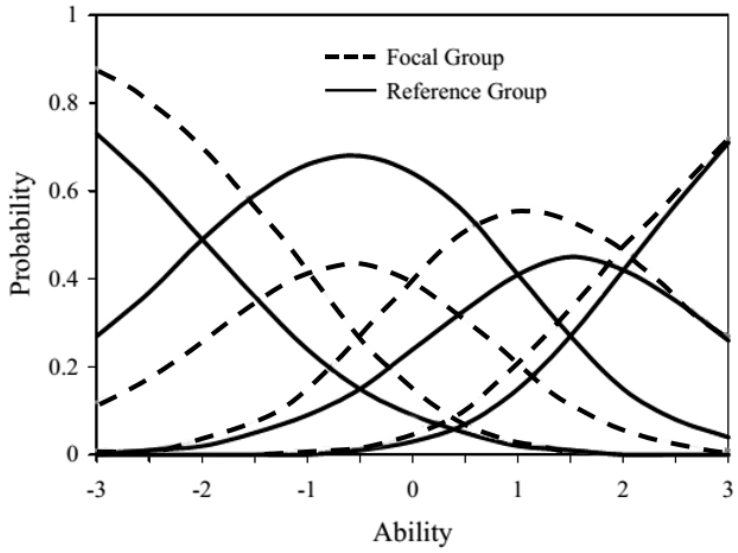
สำหรับเงื่อนไขของปฏิสัมพันธ์ดังกล่าวเป็นรูปแบบของการทำหน้าที่ต่างกันของข้อคำถามให้คะแนนสองค่า ส่วนข้อคำถามที่ให้คะแนนหลายค่าจะมีความสมบูรณ์มากกว่า ทั้งนี้เพราะว่าไม่เพียงแต่จะเกิดปฏิสัมพันธ์ระหว่างระดับความสามารถกับการเป็นสมาชิกของกลุ่มเท่านั้น แต่ยังมีตัวแปรที่สาม คือ ระดับคะแนนเข้ามาเกี่ยวข้องด้วย ดังนั้น การทำหน้าที่ต่างกันของข้อคำถามให้คะแนนหลายค่า จะเกิดขึ้นภายใต้รายการคะแนน (Score categories) (French & Miller, 1996) รูปแบบของการทำหน้าที่ต่างกันของข้อคำถามให้คะแนนหลายค่า จะมีลักษณะคล้ายกับรูปแบบของข้อคำถามทำหน้าที่ต่างกันที่ให้คะแนนสองค่า ซึ่งสามารถแสดงด้วยฟังก์ชันการตอบขอบรายการ (Boundary response functions: BRFs) โดยใช้โมเดลเกรดเรสพอน (Graded response model: GRM) ดังภาพที่ 17 และ 18 สำหรับฟังก์ชันรายการตอบสนองข้อคำถาม (Item-category response functions: ICRFs) โดยใช้โมเดลพาเชียลเครดิตทั่วไป (Generalized partial credit model: GPCM) แสดงดังภาพที่ 19 และ 20



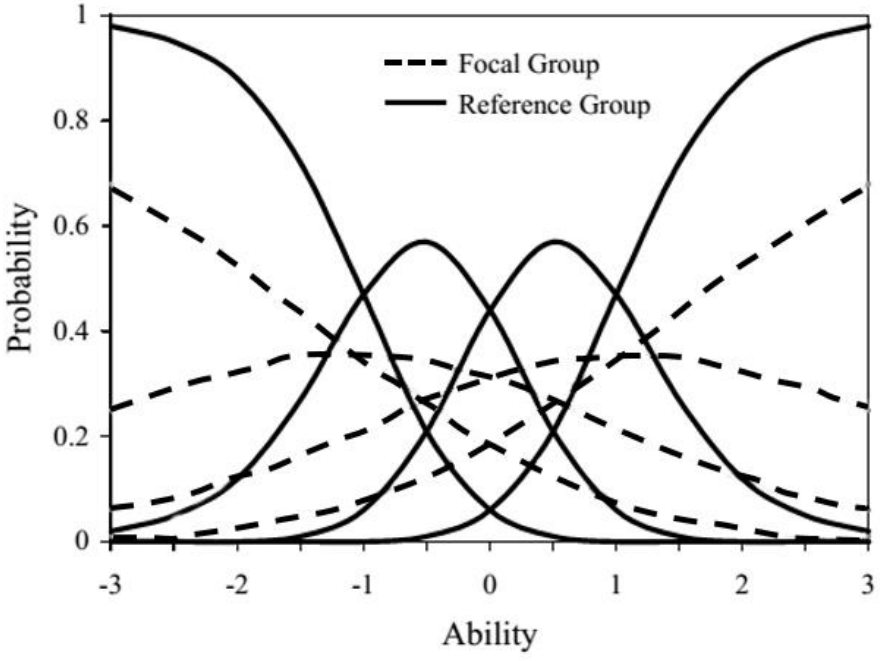
ภาพที่ 17 ข้อคำถามทำหน้าที่ต่างกันรูปแบบเดียวกัน (Uniform DIF)ภายใต้โมเดลเกรดเรสพอน (GRM)



ภาพที่ 18 ข้อคำถามทำหน้าที่ต่างกันที่ไม่เป็นรูปแบบเดียวกัน (Non uniform DIF) ภายใต้โมเดลเกรดเรสพอน (GRM)

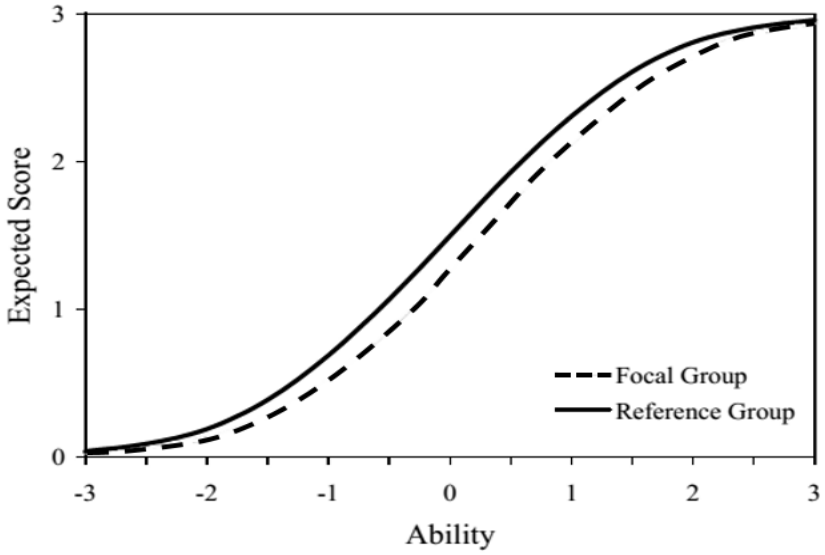


ภาพที่ 19 ข้อคำถามที่ทำหน้าที่ต่างกันที่เป็นรูปแบบเดียวกัน (Uniform DIF) ภายใต้โมเดลพาเชียลเครดิตทั่วไป (GPCM)

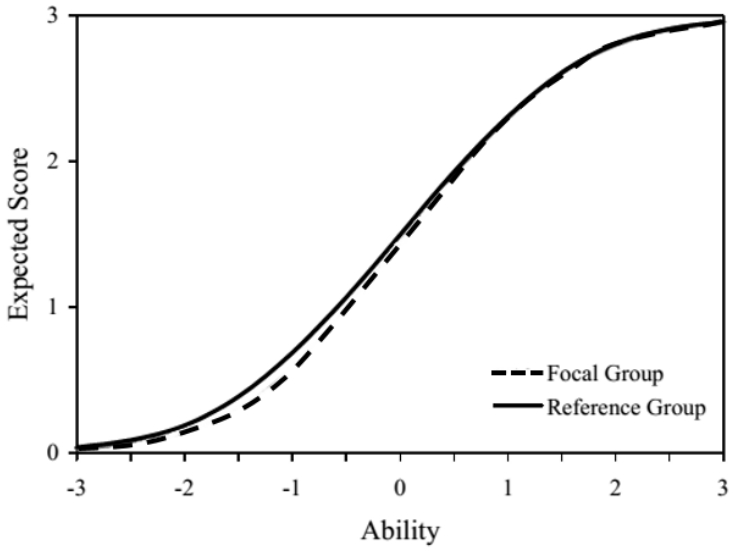


ภาพที่ 20 ข้อคำถามที่ทำหน้าที่ต่างกันที่ไม่เป็นรูปแบบเดียวกัน (Non uniform DIF) ภายใต้โมเดลพาเชียลเครดิตทั่วไป (GPCM)

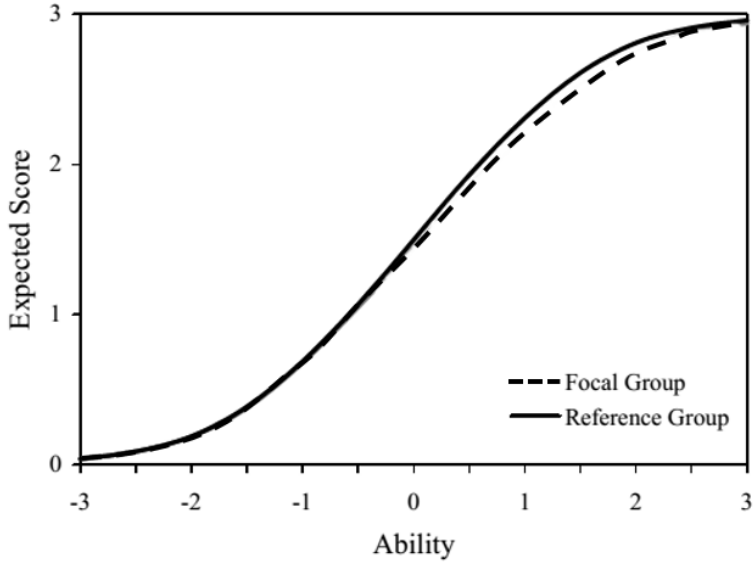
นอกจากนี้ Zwick et al. (1993) ได้ศึกษารูปแบบของข้อคำถามทำหน้าที่ต่างกันให้คะแนนหลายค่า และได้กำหนดนิยามโดยใช้ฟังก์ชันความน่าจะเป็นสะสมของการตอบในรายการคะแนน (Score categories) ซึ่งมีรูปแบบแตกต่างกัน 4 รูปแบบ คือ 1) ข้อคำถามทำหน้าที่ต่างกันคงที่ (Constant DIF) เกิดขึ้นเมื่อค่าพารามิเตอร์ความยากในทุกรายการของข้อคำถาม สำหรับกลุ่มสนใจมีค่าสูงกว่ากลุ่มอ้างอิงในระดับคงที่ 2) ข้อคำถามทำหน้าที่ต่างกันของข้อคำถามเปลี่ยนขนาดที่ระดับความสามารถต่ำ (Low-shift DIF) เกิดขึ้นเมื่อค่าพารามิเตอร์ความยากในรายการต่ำสุดของข้อคำถามสำหรับกลุ่มสนใจมีค่าสูงกว่ากลุ่มอ้างอิง ส่วนค่าพารามิเตอร์ความยากในรายการของข้อคำถามที่เหลือสำหรับสองกลุ่มมีค่าเท่ากัน 3) ข้อคำถามทำหน้าที่ต่างกันแบบเปลี่ยนขนาดที่ระดับความสามารถสูง (High-shift DIF) เกิดขึ้นเมื่อค่าพารามิเตอร์ความยากในรายการสูงสุดของข้อคำถามสำหรับกลุ่มสนใจมีค่าสูงกว่ากลุ่มอ้างอิง ส่วนค่าพารามิเตอร์ความยากในรายการของข้อคำถามที่เหลือสำหรับสองกลุ่มมีค่าเท่ากัน และ 4) ข้อคำถามทำหน้าที่ต่างกันของข้อคำถามแบบสมดุล (Balanced DIF) เกิดขึ้นเมื่อค่าพารามิเตอร์ความยากในรายการสูงของข้อคำถามสำหรับกลุ่มสนใจมีค่าต่ำกว่ากลุ่มอ้างอิง ส่วนค่าพารามิเตอร์ความยากในรายการของข้อคำถามที่เหลือสำหรับสองกลุ่มมีค่าเท่ากัน รูปแบบของข้อคำถามทั้ง 4 รูปแบบสามารถแสดงด้วยฟังก์ชันการตอบข้อคำถามสะสมในทุกรายการคำตอบ ดังภาพที่ 21-24



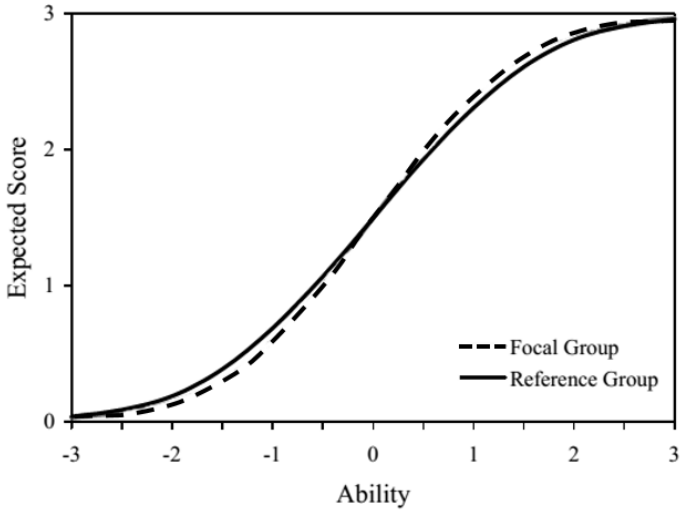
ภาพที่ 21 ข้อคำถามที่ทำหน้าที่ต่างกันแบบคงที่ (Constant DIF) ภายใต้โมเดลพาเชียลเครดิต (PCM)



ภาพที่ 22 ข้อคำถามที่ทำหน้าที่ต่างกันแบบเปลี่ยนขนาดที่ระดับความสามารถต่ำ (Low-shift DIF) ภายใต้โมเดลพาเชียลเครดิต (PCM)



ภาพที่ 23 ข้อคำถามทำหน้าที่ต่างกันแบบเปลี่ยนขนาดที่ระดับความสามารถสูง (High-shift DIF) ภายใต้โมเดลพาเชียลเครดิต (PCM)



ภาพที่ 24 ข้อคำถามทำหน้าที่ต่างกันแบบสมดุล (Balanced DIF) ภายใต้โมเดลพาเชียลเครดิต (PCM)

จากภาพที่ 24 เมื่อเปรียบเทียบกับรูปแบบของข้อคำถามทำหน้าที่ต่างกันให้คะแนนสองค่า และจากภาพที่ 21-23 จะเห็นว่า ข้อคำถามทำหน้าที่ต่างกันคงที่ (Constant DIF) ข้อคำถามทำหน้าที่ต่างกันแบบเปลี่ยนขนาดที่ระดับความสามารถต่ำ (Low-shift DIF) และข้อคำถามทำหน้าที่ต่างกันแบบเปลี่ยนขนาดที่ระดับความสามารถสูง (High-shift DIF) ก็คือ ข้อคำถามทำหน้าที่ต่างกันที่มีทิศทางเดียวกัน ส่วนข้อคำถามทำหน้าที่ต่างกันแบบสมดุล (Balance DIF) ก็คือ ข้อคำถามทำหน้าต่างกันที่ไม่มีทิศทางเดียวกัน (Chang et al., 1996)

ในการวิจัยครั้งนี้ ผู้วิจัยมีความสนใจศึกษารูปแบบการทำหน้าที่ต่างกันของข้อคำถาม ซึ่งจะนำมาจัดกระทำข้อมูลเพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามที่วัดความสามารถหลายมิติ และให้คะแนนหลายค่า ภายใต้การจำลองข้อมูลโดยใช้ทฤษฎีการตอบสนองข้อคำถามเกรดเรสพอนพหุมิติ โดยใช้รูปแบบการทำหน้าที่ต่างกันของข้อคำถามที่ไม่เป็นรูปแบบเดียวกัน (Nonuniform DIF) โดยพารามิเตอร์อำนาจจำแนกของกลุ่มสนใจและพารามิเตอร์เทรสโฮลของทุกรายการของกลุ่มสนใจมีค่ามากกว่ากลุ่มอ้างอิง

ตอนที่ 4 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม

ในการศึกษาครั้งนี้ ผู้วิจัยนำขอนำเสนอวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีผู้วิจัยได้นำเสนอ สำหรับแบบทดสอบแบบพหุมิติและให้คะแนนหลายค่า ได้แก่ วิธีโพลีโตมัส

ซิปเทสท์ วิธีวิเคราะห์องค์ประกอบยืนยันกลุ่มพหุ วิธีการถดถอยโลจิสติก วิธีการถดถอยแบบโลจิสติกแบบจัดอันดับ และวิธีตรวจสอบแบบวอลด์ ทั้งหมด 6 วิธี ดังนี้

1. วิธีซิปเทสท์ (SIBTEST)

Shealy and Stout (1993) ได้เสนอวิธี “Simultaneous item bias test”หรือเรียกสั้น ๆ ว่า “วิธีซิปเทสท์” (SIBTEST) เพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามและแบบทดสอบ (Differential item/test functioning: DIF/ DTF) ในข้อสอบที่ให้คะแนนสองค่า หลักการตรวจสอบด้วยวิธีซิปเทสท์มีแนวคิดคล้ายกับวิธีการทำให้เป็นมาตรฐาน (Standardization: STD) (Dorans & Kulick, 1986) โดยพัฒนามาจากโมเดลความลำเอียงของแบบทดสอบภายใต้ทฤษฎีการตอบข้อสอบแบบหลายมิติ (Multidimensional IRT) มีรูปแบบนันพาราเมทริก (Nonparametric form) ซึ่งไม่ต้องใช้ฟังก์ชันการตอบข้อสอบประมาณค่าความสามารถ วิธีซิปเทสท์เป็นวิธีที่ใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามรูปแบบเดียวกัน (Unidirectional DIF) โดยเฉพาะ ดังนั้น จึงมีข้อจำกัดในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามที่ไม่มีทิศทางเดียวกัน (Non unidirectional DIF) (Li & Stout, 1996; Shealy & Stout, 1993) ส่วนข้อได้เปรียบของวิธีซิปเทสท์ก็คือ สามารถคำนวณได้ง่าย เสียค่าใช้จ่ายไม่มาก และไม่จำเป็นต้องใช้ตัวอย่างขนาดใหญ่ สามารถใช้สถิติทดสอบนัยสำคัญเพื่อตัดสินการทำหน้าที่ต่างกันของข้อคำถามครั้งละหนึ่งข้อ หรือมากกว่าหนึ่งข้อพร้อมกัน (Simultaneous) ผลการวิเคราะห์ทำให้ทราบขนาดและทิศทางของการทำหน้าที่ต่างกันของข้อคำถาม (Nandakumar, 1993) นักวิจัยหลายคนได้นำวิธีซิบเทสท์มาศึกษาและปรับขยายเพื่อใช้ตรวจสอบในประเด็นต่าง ๆ เช่น Nandakumar (1993) ได้ศึกษาการทำหน้าที่ต่างกันของข้อคำถามในสองกรณี คือ การขยายข้อสอบทำหน้าที่เบี่ยงเบน (DIF amplification) และการหักล้างข้อสอบทำหน้าที่เบี่ยงเบน (DIF cancellation) Li and Stout (1996) ได้พัฒนาวิธีครอซซิงซิปเทสท์ (Crossing SIBTEST) เพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามแบบตัดกัน Roussos and Stout (1996) ได้ศึกษาอัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีซิปเทสท์ในกลุ่มตัวอย่างขนาดเล็ก Chang et al. (1996) ได้พัฒนาวิธีโพลีโตมัสซิปเทสท์ (Poly-SIBTEST) เพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามที่ให้คะแนนหลายค่า Stout et al. (1997) ได้พัฒนาวิธีมัลติซิป (MULTISIB) เพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามแบบสองมิติ เป็นต้น

แนวคิดและหลักการ

Shealy and Stout (1993) ได้อธิบายการทำหน้าที่ต่างกันของข้อคำถาม โดยใช้ขอบของฟังก์ชันการตอบข้อสอบ (Marginal IRFs) ของความสามารถเป้าหมายที่ต้องการวัด θ สำหรับกลุ่ม g (กลุ่มอ้างอิง หรือกลุ่มสนใจ) ดังนี้

$$Mig(\theta) = E\left[P_i(\Theta,\eta) \mid \Theta = \theta, G = g\right] \tag{30}$$

ถ้า $\eta \mid \Theta = \theta, G = g$ เป็นฟังก์ชันความหนาแน่นแบบมีเงื่อนไขของ η เมื่อกำหนดความสามารถ θ ของกลุ่ม g มีค่าคงที่ ซึ่งแทนด้วย $fg(\eta \mid \theta)$ ดังนั้น การกำหนดนิยามในสมการที่ 30 สามารถคำนวณ ได้ดังนี้

$$Mig(\theta) = \int_{-\infty}^{+\infty} P_i(\theta,\eta) fg(\eta \mid \theta) d\eta \tag{31}$$

จากสมการที่ 31 ถ้าการแจกแจงแบบมีเงื่อนไขของความสามารถ η มีค่าเท่ากันสำหรับผู้สอบสองกลุ่ม แล้วข้อสอบจะทำหน้าที่ไม่เบี่ยงเบน (No-DIF) เพราะว่ามีความสามารถ θ เท่ากันจะทำให้ความน่าจะเป็นในการตอบข้อสอบถูกเท่ากัน (Ackerman, 1992) จากแนวคิดดังกล่าวสามารถนิยามการทำหน้าที่ต่างกันของข้อคำถาม (DIF) โดยใช้ Marginal IRFs ได้ว่า ถ้าฟังก์ชันการตอบข้อสอบของความสามารถเป้าหมายสำหรับกลุ่มอ้างอิงมีค่ามากกว่าฟังก์ชันการตอบข้อสอบสำหรับกลุ่มสนใจ แล้วข้อสอบจะทำหน้าที่เบี่ยงเบน โดยข้อสอบเข้าข้างกลุ่มอ้างอิง ซึ่งแสดงในรูปสัญลักษณ์ทางคณิตศาสตร์ ดังนี้

$$M_{iR}(\theta) > M_{iF}(\theta) \tag{32}$$

ฟังก์ชันการตอบข้อสอบของแบบทดสอบที่ต้องการศึกษา สำหรับผู้สอบกลุ่มอ้างอิงและกลุ่มสนใจ กำหนดดังนี้

$$M_{SR}(\theta) = \sum_{i=n+1}^{N} M_{iR}(\theta) \tag{33}$$

$$M_{SR}(\theta) = \sum_{i=n+1}^{N} M_{iF}(\theta) \tag{34}$$

เมื่อ $M_{SR}(\theta)$ และ $M_{SF}(\theta)$ แทนผลรวม Marginal IRFs ของข้อสอบที่ต้องการศึกษา ณ ระดับความสามารถ θ จากผู้สอบกลุ่มอ้างอิงและกลุ่มสนใจ ตามลำดับ สำหรับปริมาณของการทำหน้าที่ต่างกันของข้อคำถาม (Amount of DIF) สามารถคำนวณจากความแตกต่างระหว่าง $M_{SR}(\theta)$ และ $M_{SF}(\theta)$ ดังนี้

$$B(\theta) = MSR(\theta) - MSF(\theta) \tag{35}$$

ขนาดของความแตกต่างดังกล่าว แสดงถึงปริมาณของการทำหน้าที่ต่างกันของข้อคำถามจากแบบทดสอบชุดย่อยที่ต้องการศึกษา ณ ระดับความสามารถ θ ซึ่งเข้าข้างกลุ่มอ้างอิง Shealy and Stout (1993) ได้คำนวณค่าเฉลี่ยของปริมาณการทำหน้าที่ต่างกันของข้อคำถามรูปแบบเดียวกัน (Unidirectional DIF) ดังนี้

$$\beta_{uni} = \int_{-\infty}^{+\infty} B(\theta) f_F(\theta) d\theta \tag{36}$$

เมื่อ β_{uni} แทนดัชนีการทำหน้าที่ต่างกันของข้อคำถามที่มีรูปแบบเดียวกัน และ $f_F(\theta)$ แทนฟังก์ชันความหนาแน่นความน่าจะเป็นของการแจกแจงความสามารถเป้าหมาย จากผู้สอบกลุ่มรวมทั้งหมด

กระบวนการตรวจสอบ

ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามรูปแบบเดียวกันตามแนวคิดของ Shealy and Stout (1993) จะเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มอ้างอิงและกลุ่มสนใจ โดยใช้แบบทดสอบจำนวน N ข้อ แล้วแบ่งแบบทดสอบดังกล่าวออกเป็นสองชุด คือ แบบทดสอบชุดย่อยที่มีความถูกต้อง (Valid subtests) และแบบทดสอบชุดย่อยที่ใช้ในการศึกษา (Studied subtests) กล่าวคือ แบบทดสอบชุดแรกใช้ในการจับคู่เปรียบเทียบ (Matching subtests) ประกอบด้วยข้อสอบข้อที่ 1 ถึง n ซึ่งเป็นข้อสอบที่ไม่สงสัยว่าทำหน้าที่เบี่ยงเบน โดยวัดความสามารถเป้าหมาย θ เพียงความสามารถเดียว ส่วนแบบทดสอบชุดหลังเป็นส่วนที่เหลือจากชุดแรกประกอบด้วยข้อสอบข้อที่ $n+1$ ถึง N ข้อสอบดังกล่าวสงสัยว่าทำหน้าที่เบี่ยงเบน โดยวัดทั้งความสามารถเป้าหมาย θ และความสามารถแทรกซ้อน η

การทดสอบสมมติฐาน

ในการทดสอบสมมติฐานของการทำหน้าที่ต่างกันของข้อคำถามรูปแบบเดียวกัน เมื่อข้อสอบเข้าข้างกลุ่มอ้างอิง นำดัชนี β_{uni} มากำหนดสมมติฐานศูนย์ (H_0) และสมมติฐานทางเลือก (H_1) ดังนี้

$$H_0 : \beta_{uni} = 0$$

$$H_1 : \beta_{uni} > 0 \tag{37}$$

การทดสอบสมมติฐานการทำหน้าที่ต่างกันของข้อคำถามรูปแบบเดียวกันจะประมาณค่าดัชนี β_{uni} โดยคำนวณจากคะแนนของแบบทดสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบและแบบทดสอบชุดย่อยที่ต้องการศึกษา ดังนี้

$$X = \sum_{i=n+1}^{N} U_i \tag{38}$$

$$Y = \sum_{i=n+1}^{N} U_i \tag{39}$$

เมื่อ X แทน คะแนนรวมจากแบบทดสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบ
Y แทน คะแนนรวมจากแบบทดสอบชุดย่อยที่ต้องการศึกษา
U_i แทน ผลการตอบข้อสอบข้อที่ i
(ตอบถูกได้ 1 คะแนน และตอบผิดได้ 0 คะแนน)

คำนวณคะแนนเฉลี่ยจากผลการตอบข้อสอบในแบบทดสอบชุดย่อยที่ต้องการศึกษาของผู้สอบกลุ่มอ้างอิงและกลุ่มสนใจที่มีความสามารถระดับเดียวกัน แล้วนำคะแนนเฉลี่ยดังกล่าวมาจับคู่เปรียบเทียบ โดยพิจารณาได้จากคะแนนรวมที่เท่ากันของแบบทดสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบ ($X=k$) ดังนี้

$$\bar{Y}_{Rk} - \bar{Y}_{Fk}; \cdots\cdots k = 0,1,2,\ldots,n \tag{40}$$

เมื่อ $\bar{Y}_{Rk}$ และ $\bar{Y}_{Fk}$ แทนค่าเฉลี่ยของคะแนน Y จากการตอบแบบทดสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบ แล้วได้คะแนนรวม $X = k$ สำหรับผู้สอบในกลุ่มอ้างอิงและกลุ่มสนใจตามลำดับ คะแนนเฉลี่ยที่ใช้ในการเปรียบเทียบดังกล่าวอาจทำให้การตรวจสอบผิดพลาดจากความเป็นจริง กล่าวคือ เมื่อเกิดความแตกต่างของการแจกแจงค่าความสามารถ (Ability distribution) ของกลุ่มอ้างอิงและกลุ่มสนใจจะมีผลทำให้ $\bar{Y}_{Rk} - \bar{Y}_{Fk}$ มีค่าแตกต่างจาก 0 อย่างเป็นระบบ ทำให้ตรวจพบว่าข้อสอบทำหน้าที่เบี่ยงเบน ซึ่งความเป็นจริงแล้วข้อสอบทำหน้าที่ไม่เบี่ยงเบน ดังนั้นความแตกต่างของการแจกแจงค่าความสามารถของกลุ่มอ้างอิงและกลุ่มสนใจที่เกิดขึ้นสามารถปรับแก้ค่าการถดถอย (Regression correction) เพื่อกำจัดค่าที่สูงเกินปกติ (Inflate) (Shealy and Stout, 1993) สำหรับค่าเฉลี่ย $\bar{Y}_{Rk}$ และ $\bar{Y}_{Fk}$ ที่ปรับแก้แล้วแทนด้วย $\bar{Y}^*_{Rk}$ และ $\bar{Y}^*_{Fk}$ ตามลำดับ

ค่า $\bar{Y}_{Rk}^{*} - \bar{Y}_{Fk}^{*}$ เป็นความแตกต่างของผลการตอบข้อสอบในแบบทดสอบชุดย่อยที่ศึกษาระหว่างกลุ่มผู้สอบที่มีความสามารถระดับเดียวกัน ถ้า $\bar{Y}_{Rk}^{*} - \bar{Y}_{Fk}^{*} = 0$ ทุกคะแนน k แสดงว่า ข้อสอบที่สงสัยในแบบทดสอบชุดย่อยทำหน้าที่ไม่แตกต่างกัน และ ถ้า $\bar{Y}_{Rk}^{*} - \bar{Y}_{Fk}^{*} > 0$ ทุกคะแนน k แสดงว่า ข้อสอบที่สงสัยในแบบทดสอบชุดย่อยทำหน้าที่เบี่ยงเบนที่มีทิศทางเดียวกัน โดยข้อสอบเข้าข้างกลุ่มอ้างอิง ถ้า $\bar{Y}_{Rk}^{*} - \bar{Y}_{Fk}^{*} < 0$ ทุกคะแนน k แสดงว่า ข้อสอบที่สงสัยในแบบทดสอบชุดย่อยทำหน้าที่เบี่ยงเบนที่มีทิศทางเดียวกัน โดยข้อสอบเข้าข้างกลุ่มสนใจ สำหรับค่าความแตกต่างของผลการตอบข้อสอบดังกล่าว สามารถนำมาประมาณค่าในรูป $\hat{\beta}_{uni}$ ดังนี้

$$\hat{\beta}_{uni} = \sum_{k=0}^{n} \hat{P}_k \left(\bar{Y}_{Rk}^{*} - \bar{Y}_{Fk}^{*} \right) \tag{41}$$

เมื่อ $\hat{P}_k$ แทนสัดส่วนของผู้สอบทั้งหมด (กลุ่มอ้างอิงและกลุ่มสนใจ) ตอบแบบทดสอบชุดย่อยที่ใช้จับคู่เปรียบเทียบ แล้วได้คะแนนรวม $X = k$ สัดส่วนของผู้สอบดังกล่าวสามารถเขียนในรูปสัญลักษณ์ ดังนี้

$$\hat{P}_k = \frac{\left(J_{Rk} + J_{Fk} \right)}{\sum_{k=0}^{n} \left(J_{Rk} + J_{Fk} \right)} \tag{42}$$

เมื่อ J_{Rk} และ J_{Fk} แทนจำนวนผู้สอบซึ่งตอบแบบทดสอบชุดย่อยที่ใช้จับคู่เปรียบเทียบแล้วได้คะแนนรวม $X = k$ สำหรับกลุ่มอ้างอิงและกลุ่มสนใจ ตามลำดับ จากนั้นจึงนำค่าประมาณ $\hat{\beta}_{uni}$ ที่คำนวณในสมการที่ (41) มาทดสอบสมมติฐานศูนย์ (No-DIF) โดยใช้สถิติ β_{uni} ดังนี้

$$B_{uni} = \frac{\hat{\beta}_{uni}}{\hat{\sigma}\left(\hat{\beta}_{uni} \right)} \tag{43}$$

$\hat{\sigma}\left(\hat{\beta}_{uni} \right)$ เป็นค่าประมาณความคลาดเคลื่อนมาตรฐานของ B_{uni} คำนวณจาก

$$\hat{\sigma}(\hat{\beta}_{uni}) = \sqrt{\Sigma_{k=0}^{n} \hat{P} \left[\frac{1}{J_{Rk}} \hat{\sigma}^2 (Y \mid k, R) + \frac{1}{J_{Fk}} \hat{\sigma}^2 (Y \mid k, F) \right]} \tag{44}$$

เมื่อ $\hat{\sigma}^2(Y|k,R)$ และ $\hat{\sigma}^2(Y|k,F)$ แทนค่าประมาณความแปรปรวนของคะแนนจากแบบทดสอบชุดย่อยที่ต้องการศึกษาในกลุ่มอ้างอิงและกลุ่มสนใจ ตามลำดับ สำหรับสถิติ B_{uni} มีการแจกแจงใกล้เคียงการแจกแจงแบบปกติมาตรฐาน เมื่อข้อสอบทำหน้าที่ไม่เบี่ยงเบนและถ้าผลการทดสอบพบว่า $B_{uni} > Z_a$ อย่างมีนัยสำคัญที่ระดับ a โดยที่ $P[N(0,1) > Z_a] = a$ แสดงว่าปฏิเสธ H_0 นั่นคือ ข้อสอบทำหน้าที่เบี่ยงเบนที่มีทิศทางเดียวกัน (Unidirectional DIF) เมื่อ $B_{uni} > 0$ แสดงว่า ข้อสอบเข้าข้างกลุ่มอ้างอิง และเมื่อ $B_{uni} < 0$ แสดงว่า ข้อสอบเข้าข้างกลุ่มสนใจ Roussos and Stout (1996) ได้เสนอเกณฑ์เพื่อใช้จำแนกขนาดของการทำหน้าที่ต่างกันของข้อคำถามไว้ดังนี้

DIF ขนาดเล็ก: ปฏิเสธสมมติฐานศูนย์ และ $|\hat{\beta}_{uni}| < 0.059$

DIF ขนาดปานกลาง: ปฏิเสธสมมติฐานศูนย์ และ $0.059 \leq |\hat{\beta}_{uni}| < 0.088$

DIF ขนาดใหญ่: ปฏิเสธสมมติฐานศูนย์ และ $|\hat{\beta}_{uni}| \geq 0.088$

การปรับแก้ค่าการถดถอย

การตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามจะมีปัญหาในเชิงสถิติ ซึ่งเกิดจากความแตกต่างของการแจกแจงความสามารถเป้าหมายระหว่างกลุ่มอ้างอิงและกลุ่มสนใจ กล่าวคือ ถ้าการแจกแจงความสามารถเป้าหมายของผู้สอบกลุ่มอ้างอิงสูงกว่ากลุ่มสนใจ จะเกิดเงื่อนไขที่เรียกว่า “ผลกระทบ” (Impact) ซึ่งทำให้สถิติ B_{uni} มีค่าเฟ้อ (Inflate) หรือมีค่าสูงเกินปกติแล้วจะส่งผลให้การตรวจสอบเกิดความคลาดเคลื่อนประเภทที่ 1 (Type I error) เพราะในความเป็นจริงข้อสอบทำหน้าที่ไม่เบี่ยงเบน แต่ตรวจพบว่าข้อสอบทำหน้าที่เบี่ยงเบน ดังนั้น จึงมีความจำเป็นที่จะต้องปรับแก้ความแตกต่างของการแจกแจงความสามารถเป้าหมายโดยใช้การปรับแก้การถดถอย (Regression correction) เพื่อกำจัดอิทธิพลค่าเฟ้อของผลกระทบ ซึ่งจะแปลงค่า $\bar{Y}_{Rk}$, $\bar{Y}_{Fk}$ เป็น $\bar{Y}^*_{Rk}$ $\bar{Y}^*_{Fk}$ ทีละคู่ ดังนั้น $\bar{Y}^*_{gk}$ เป็นตัวประมาณค่าอิทธิพลของค่าเฉลี่ยคะแนนจริงจากแบบทดสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบในกลุ่มย่อย k กลุ่ม ซึ่งแบ่งมาจากแต่ละกลุ่มของกลุ่มอ้างอิงและกลุ่มสนใจ โดยมีรายละเอียดดังนี้ (Shealy & Stout, 1993)

กำหนดให้ $X = k$ แทนคะแนนรวม ซึ่งเป็นคะแนนสังเกตจากแบบทดสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบ $V_g(k)$ แทนค่าการถดถอยของคะแนนจริงจากแบบทดสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบ ซึ่งได้คะแนนเท่ากับ k ในการประมาณค่า $V_g(k)$ สมมติว่าเป็นเส้นตรงดังนี้

$$V_g(k) = \alpha + \beta k \tag{45}$$

ในการประมาณค่า $V_g(k)$ จะใช้ข้อตกลงของโมเดลคะแนนจริง ดังนี้

$$X = T + e \tag{46}$$

$$E(e) = 0 \text{ และ } \quad cov(T, e) = 0 \tag{47}$$

เมื่อ X , T และ e แทนคะแนนสังเกต คะแนนจริง และคะแนนความคลาดเคลื่อน จากแบบทดสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบ ตามลำดับ แล้วใช้ทฤษฎีการถดถอยมาตรฐานประมาณค่า $V_g(k)$ ดังนี้

$$V_g(k) = ET + \left(\frac{\rho_{XT}\sigma_T}{\sigma_X} \right)(k - EX) \tag{48}$$

เมื่อ EX และ ET แทนค่าคาดหวังของ X และ T ในกลุ่ม g (R หรือ F) สำหรับ σ_x และ σ_T แทนส่วนเบี่ยงเบนมาตรฐานของ X และ T ในกลุ่ม g (R หรือ F) ส่วน ρ_{XT} แทนความสัมพันธ์ระหว่าง X และ T ในกลุ่ม g (R หรือ F) จากโมเดลคะแนนจริงสามารถคำนวณค่าความเชื่อมั่น ดังนี้

$$\frac{\rho_{XT}\sigma_T}{\sigma_X} = 1 - \frac{\sigma_e^2}{\sigma_X^2} \tag{49}$$

เมื่อ σ_e^2 และ σ_X^2 แทนความแปรปรวนของความคลาดเคลื่อนและความแปรปรวนของคะแนนสังเกตในกลุ่ม g (R หรือ F) ตามลำดับ จากสมการที่ 46 และ 47 แสดงว่า $ET = EX$ ดังนั้น สมการที่ 48 และ 49 สามารถเขียนใหม่ได้ ดังนี้

$$V_g(k) = ET + \left(1 - \frac{\sigma_e^2}{\sigma_X^2} \right)(k - EX) \tag{50}$$

ค่าประมาณของ $V_g(k)$ สามารถคำนวณ ดังนี้

$$\hat{V}_g(k) = \overline{X}_g + 1\left(1 - \frac{\hat{\sigma}^2(e|g)}{\hat{\sigma}^2(X|g)}\right)(k - \overline{X}_g) \tag{51}$$

โดยที่

$$\hat{\sigma}^2_{(X|g)} = \frac{1}{(J_g - 1)}\sum_{J=1}^{J_g}(X_{gj} - \overline{X}_g)^2 \tag{52}$$

และ

$$\hat{\sigma}^2_{(e|g)}\hat{\sigma}^2(e|g) = \sum_{i=1}^{n}\overline{U}_{ig}(1 - \overline{U}_{ig}) \tag{53}$$

เมื่อ X_{gj} แทน คะแนนจากแบบทดสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบของผู้สอบคนที่ j ในกลุ่ม g

$\overline{X}_g$ แทน คะแนนเฉลี่ยจากแบบทดสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบของผู้สอบคนที่ j ในกลุ่ม g

$\overline{U}_{ig}$ แทน สัดส่วนการตอบข้อสอบถูกของผู้สอบกลุ่ม g ซึ่งตอบข้อสอบที่ i จากแบบทดสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบ

ค่า $\hat{V}_g(k)$ จากสมการที่ 51 เป็นค่าประมาณคะแนนจริงของความสามารถเป้าหมายสำหรับผู้สอบกลุ่ม g (R หรือ F) ที่ได้คะแนน $X = k$ จากแบบทดสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบ ซึ่งสอดคล้องกับคะแนนจากแบบทดสอบชุดย่อยที่ต้องการศึกษา โดยใช้วิธีอนุกรมของเทเลอร์ (Taylor) ปรับแก้คะแนน ดังนี้

$$\overline{Y}^*_{gk} = \overline{Y}_{gk} + \hat{M}_{gk}\left[\hat{V}(k) - \hat{V}_g(k)\right] \tag{54}$$

เมื่อ $\overline{Y}_{gk}$ แทนค่าเฉลี่ยของคะแนนสังเกตจากแบบทดสอบชุดย่อยที่ต้องการศึกษาของผู้สอบกลุ่ม g (R หรือ F) ซึ่งได้คะแนน $X = k$ จากแบบทดสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบสำหรับ $\hat{V}(k)$ และ $\hat{M}_{gk}$ คำนวณ ดังนี้

$$\hat{V}(k) = \frac{\hat{V}_R(k) + \hat{V}_F(k)}{2} \tag{55}$$

$$\hat{M}_{gk} = \frac{\bar{Y}_{g,k+1} - \bar{Y}_{g,k-1}}{\hat{V}_g(k+1) - \hat{V}_g(k-1)} \tag{56}$$

เมื่อ $\bar{Y}_{g,k+1}$ และ $\bar{Y}_{g,k-1}$ แทนค่าเฉลี่ยของคะแนนสังเกตจากแบบทดสอบชุดย่อยที่ต้องการศึกษาของผู้สอบกลุ่มย่อย จำนวน $k+1$ กลุ่มและ $k-1$กลุ่ม ตามลำดับ $\hat{V}_g(k+1)$ และ $\hat{V}_g(k-1)$ แทนค่าประมาณของคะแนนจริงจากแบบสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบของผู้สอบกลุ่มย่อย จำนวน $k+1$ กลุ่มและ $k-1$ กลุ่ม ตามลำดับ สำหรับค่า $\bar{Y}^*_{gk}$ ในสมการที่ 54 เป็นค่าประมาณของคะแนนจริงจากแบบทดสอบชุดย่อยที่ต้องการศึกษาของผู้สอบกลุ่มย่อย k ในกลุ่ม g (R หรือ F) ซึ่งสมมติว่าเท่ากับค่าประมาณคะแนนจริงจากแบบทดสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบ $\hat{V}(k)$ ของผู้สอบทั้งสองกลุ่ม

2. วิธีโพลีโตมัสซิปเทสท์ (Poly-SIBTEST)

วิธีโพลีโตมัสซิปเทสท์ (Chang et al., 1996) เป็นวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามให้คะแนนหลายค่า ที่ปรับขยายมาจากวิธีซิปเทสท์ ซึ่งเป็นการตรวจสอบการทำหน้าที่ต่างกันข้อคำถามให้คะแนนสองค่า

แนวคิดและหลักการของวิธีโพลีโตมัสซิปเทสท์

การทำหน้าที่ต่างกันของข้อคำถามที่ให้คะแนนสองค่าของตัวแปรแฝง กำหนดไว้ว่า เมื่อกำหนดให้ $E_R[Y \mid \theta]$ และ $E_F[Y \mid \theta]$ แทนการถดถอยของ Y บนตัวแปรแฝง ของกลุ่มอ้างอิงและกลุ่มสนใจ ตามลำดับ ข้อคำถามจะไม่แสดงการทำหน้าที่ต่างกัน ถ้าทุกค่าของความสามารถที่มี

$$E_R[Y \mid \theta] = E_F[Y \mid \theta] \tag{57}$$

และเมื่อกำหนด $E_R[Y \mid \theta]$ และ $E_F[Y \mid \theta]$ แทนการถดถอยของ Y บนคะแนนสังเกตของกลุ่มอ้างอิงและกลุ่มสนใจ ตามลำดับ ข้อคำถามจะไม่แสดงการทำหน้าที่ต่างกันของข้อคำถามถ้าทุกค่าของคะแนน X ที่มี

$$E_R[Y \mid X] = E_F[Y \mid X] \tag{58}$$

การนำนิยามของตัวแปรแฝงในสมการที่ 57 มาใช้ทดสอบการทำหน้าที่ต่างกันของข้อคำถามที่ให้คะแนนหลายค่า สามารถกำหนดได้ว่า ข้อคำถามไม่แสดงการทำหน้าที่ต่างกันของข้อคำถาม ถ้าการถดถอยของคะแนนข้อคำถามบนตัวแปรแฝงเท่ากัน สำหรับกลุ่มผู้สอบภายใต้การศึกษา เมื่อกำหนด Y แทนคะแนนของข้อคำถามที่ต้องการศึกษา โดยเป็นคะแนนในรายการจัดอันดับ (Ordered categories) ซึ่งมี $m+1$ รายการ $(Y=k,0\le k\le m)$ และ $P_{k,g}(\theta)$ แทนฟังก์ชันการตอบในแต่รายการของข้อคำถาม (Item-category response function: ICRF) ที่ระดับคะแนน k ในกลุ่ม g ดังนั้น การถดถอยของคะแนนข้อคำถามบนความสามารถ k จะกำหนดในรูปผลรวมของฟังก์ชันการตอบรายการของข้อคำถามแบบถ่วงน้ำหนัก ดังนี้

$$E_g[Y\mid\theta]=\sum_{k=1}^{m}kP_{k,g}(\theta) \tag{59}$$

ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามที่ให้คะแนนสองค่าจะเปรียบเทียบความแตกต่างของ IRFs ระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มสนใจทุกระดับความสามารถ θ โดยอาศัยนิยามจากสมการที่ (57) ส่วนการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามที่ให้คะแนนหลายค่าจะเปรียบเทียบความแตกต่างของ ICRFs เช่นเดียวกัน โดยสามารถเขียนในรูปสัญลักษณ์ได้ดังนี้

$$P_{kR}(\theta)=P_{kF}(\theta),k=1,2,\ldots,m \tag{60}$$

เมื่อ $P_{kR}(\theta)$ และ $P_{kF}(\theta)$ แทน ICFs ที่ระดับคะแนน k ของผู้สอบกลุ่มอ้างอิงและกลุ่มสนใจ ตามลำดับ โดยที่ข้อคำถาม 1 ข้อ จะมี ICFs จำนวน m ฟังก์ชัน ดังนั้น นิยามจากสมการที่ 57 ของตัวแปรแฝงของข้อคำถามที่ให้คะแนนสองค่า สามารถนำมาใช้กับข้อคำถามที่ให้คะแนนหลายค่าแบบเรียงลำดับ

สำหรับวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามที่ให้คะแนนหลายค่า ซึ่งใช้คะแนนสังเกตในสมการที่ 58 ดังเช่น วิธีแมนเทล-แฮนส์เซล (Mantel-Haenszel: MH) (Holland & Thayer, 1988) และวิธีการทำให้เป็นมาตรฐาน (Standardization: STD) (Dorans & Kulick, 1986) สามารถกำหนดเป็นนิยามได้ว่า “ข้อคำถามไม่แสดงทำหน้าที่ต่างกัน ถ้าการถดถอยของคะแนนสอบที่ให้คะแนนหลายค่า บนคะแนนสังเกตได้ของแบบทดสอบที่ใช้ในการจับคู่มีค่าเท่ากัน สำหรับกลุ่มผู้สอบภายใต้การศึกษา

กระบวนการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม

การทำหน้าที่ต่างกันของข้อคำถามที่ให้คะแนนหลายค่า โดยใช้วิธีโพลีโตมัสซิปเทสท์ จะใช้กรอบแนวคิดเช่นเดียวกับวิธีซิปเทสท์ต้นฉบับเดิม (Original categories) ดังนี้

กำหนดให้

Y แทน คะแนนของข้อคำถามที่ต้องการศึกษา ซึ่งให้คะแนนตามรายการแบบจัดอันดับจำนวน $m+1$ รายการ $(Y = 0,1,2,\ldots,m)$

$X_1, X_2, \ldots, X_n$ แทน คะแนนของข้อคำถามที่ใช้ในการจับคู่เปรียบเทียบ จำนวน n ข้อ

$m_1, m_2, \ldots, m_n$ แทน คะแนนมากที่สุดที่เป็นไปได้ของ $X_1, X_2, \ldots, X_n$ ตามลำดับ

X แทน คะแนนที่ใช้ในการจับคู่เปรียบเทียบ คำนวณจากสูตร

$$X = \Sigma_{j=1}^{n} X_j \tag{61}$$

เมื่อ $X = 0,1,2,\ldots,n_H$ โดยที่ n_H เป็นคะแนนของข้อคำถามที่ใช้ในการจับคู่เปรียบเทียบซึ่งมีค่ามากที่สุดเป็นไปได้ สามารถคำนวณได้ ดังนี้

$$n_H = \Sigma_{j=1}^{n} m_j \tag{62}$$

$\bar{Y}_{gk}$ แทน คะแนนเฉลี่ยของข้อคำถามที่ต้องการศึกษาสำหรับผู้สอบทั้งหมดในกลุ่ม g (R หรือ F) ซึ่งได้คะแนน $X = k$ ถึงแม้ว่าวิธีซิปเทสท์พัฒนามาจากโมเดลของทฤษฎีการตอบสนองข้อคำถาม (IRT) แต่ในที่นี้จะนำทฤษฎีการทดสอบแบบมาตรฐานเดิม (CTT) มาอธิบายเพื่อทำความเข้าใจกระบวนการตรวจสอบของวิธีดังกล่าว ตามข้อตกลงของทฤษฎีการทดสอบมาตรฐานเดิมเกี่ยวกับคะแนน X กำหนดว่า $X = T + E$ เมื่อ T แทนคะแนนจริงของแบบทดสอบที่ใช้ในการจับคู่ ดังนั้น ตัวแปร $E = X - T$ แทนความคลาดเคลื่อนในการวัดและสมมติว่ามีค่าเฉลี่ยเป็นศูนย์ทั้งสองกลุ่ม เมื่อให้ $f_g(t)$ แทนความหนาแน่นของคะแนนจริงของแบบทดสอบที่ใช้ในการจับคู่ในกลุ่ม g (R หรือ F) ซึ่งการถดถอยดังกล่าวเขียนในรูปลักษณ์ได้ว่า

$$E_g[Y \mid \theta] = [Y \mid T = t, G = g] \tag{63}$$

นิยามของข้อคำถามที่ต้องการศึกษาไม่แสดงการทำหน้าที่ต่างกันของข้อคำถามตามทฤษฎีการทดสอบแบบมาตรฐานเดิม กำหนดได้ว่า ถ้า $E_R[Y \mid t] = E_F[Y \mid t]$ สำหรับทุกค่า

ของคะแนนจริง *t* จากแบบทดสอบที่ใช้ในการจับคู่ นิยามดังกล่าวสมมูลกับนิยามที่กำหนด โดยใช้ตัวแฝงตามทฤษฎีการตอบสนองข้อคำถาม (Chang & Mazzeo, 1994 cited in Chang et al., 1996) ดังนั้น สามารถนำนิยามดังกล่าวไปอธิบายการทำหน้าที่ต่างกันของข้อคำถามให้คะแนนหลายค่าด้วยวิธีโพลีโตมัสซิปเทสท์ ซึ่งมีรายละเอียด ดังนี้

ซิปเทสท์ พัฒนามาจากโมเดลในทฤษฎีการตอบสนองข้อคำถาม ดังนั้น การกำหนดนิยามการทำหน้าที่ต่างกันของข้อคำถามที่ให้คะแนนสองค่าจะใช้ตัวแปรแฝง ซึ่งสามารถกำหนดได้ว่า การทำหน้าที่ต่างกันของข้อคำถามเกิดขึ้นเมื่อ $E_R[Y \mid \theta] \neq E_F[Y \mid \theta]$ ที่ระดับความสามารถ θ และปริมาณของการทำหน้าที่ต่างกันของข้อคำถามสามารถวัดได้จาก $B_0 = E_R[Y \mid t] = E_F[Y \mid t]$ เมื่อ $f_F(\theta)$ แทนความหนาแน่นของความสามารถ θ ในกลุ่มสนใจ ดัชนีการทำหน้าที่ต่างกันของข้อคำถามสามารถคำนวณได้จาก $B_0 = \int B_0(\theta) f_f(\theta) d\theta$

จากนิยามการทำหน้าที่ต่างกันของข้อคำถามที่ให้คะแนนสองค่าโดยใช้ตัวแปรแฝง เมื่อเปลี่ยนตัวแฝงเป็นคะแนนจริงตามทฤษฎีการทดสอบแบบมาตรฐานเดิม สามารถกำหนดนิยามการทำหน้าที่ต่างกันของข้อคำถามที่ให้คะแนนหลายค่า จะได้ว่า $B(t) = E_R[Y \mid t] - E_F[Y \mid t]$ โดยที่ดัชนีการทำหน้าที่ต่างกันของข้อคำถามสามารถคำนวณจาก $f_F(t)\beta = \int B_0(t) f_F(t) dt$ ซึ่งสามารถประมาณค่าได้ ดังนี้

$$d_k = \bar{Y}_{RK} - \bar{Y}_{FK}, k = 1,2,\ldots,n_H \tag{64}$$

เมื่อ $\bar{Y}_{RK}$ และ $\bar{Y}_{FK}$ แทน ค่าเฉลี่ยของคะแนนข้อคำถามที่ต้องการศึกษาของผู้สอบกลุ่มอ้างอิงและกลุ่มสนใจ ตามลำดับ และได้คะแนน $X = k$ โดยที่ $\bar{Y}_{RK} - \bar{Y}_{FK}$ เป็นความแตกต่างของผลการตอบข้อคำถามที่ต้องการศึกษาระหว่างกลุ่มผู้สอบ ซึ่งมีคะแนนสังเกตจากแบบทดสอบที่ใช้ในการจับคู่เท่ากัน ถ้าสมมติว่าผู้สอบมีคะแนนสังเกตได้เท่ากับหรือใกล้เคียงคะแนนจริงจากแบบทดสอบที่ใช้ในการจับคู่ ซึ่งจะเป็นจริงได้ถ้าแบบทดสอบดังกล่าวมีความยาวมากพอเพื่อจะทำให้ความเชื่อมั่นสูง หรือกลุ่มผู้สอบที่ศึกษามีการแจกแจงความสามารถคล้ายคลึงกัน ดังนั้นสมการที่ 64 ถือได้ว่าเป็นความแตกต่างในคะแนนข้อคำถามที่ระดับคะแนนจริงเท่ากัน ถ้าข้อคำถามที่ศึกษาไม่มีคะแนนสังเกตที่ทำให้ข้อคำถามทำหน้าที่ต่างกันแล้ว คาดว่า $d_k \approx 0$ สำหรับการประมาณค่าดัชนีการทำหน้าที่ต่างกันของข้อคำถาม สามารถคำนวณได้ ดังนี้

$$\hat{\beta} = \Sigma_{k=0}^{nH} \sum_{k=0}^{n_H} p_k d_k \tag{65}$$

โดย

$$P_k = \frac{N_{Rk}+N_{Fk}}{N} \tag{66}$$

เมื่อ P_k แทน สัดส่วนของผู้สอบทั้งหมด (กลุ่มอ้างอิงและกลุ่มสนใจ) ซึ่งตอบแบบทดสอบที่ใช้ในการจับคู่ $X_1, X_2, \ldots, X_n$ แล้ว ได้คะแนน $X = k$ ต่อจากนั้นจะนำค่าประมาณ $\hat{\beta}$ มาทดสอบสมมติฐาน ดังนี้

การทดสอบสมมติฐาน

นำดัชนี $\hat{\beta}$ มาทดสอบสมมติฐานศูนย์ โดยใช้สถิติ B ดังนี้

$$B = \frac{\hat{\beta}}{\hat{\sigma}(\hat{\beta})} \tag{67}$$

โดยที่

$$\hat{\sigma}(\hat{\beta}) = \sqrt{\sum_{k=0}^{n_H} P_k^2 \left[\frac{\hat{\sigma}^2(Y \mid k, R)}{N_{RK}} + \frac{\hat{\sigma}^2(Y \mid k, F)}{N_{FK}} \right]} \tag{68}$$

เมื่อ $\hat{\sigma}(\hat{\beta})$ แทนค่าประมาณความคลาดเคลื่อนมาตรฐานของ β และ $\hat{\sigma}^2(Y \mid k, R)$ แทนค่าประมาณความแปรปรวนของคะแนนจากแบบทดสอบที่ต้องการศึกษาในกลุ่ม g (R หรือ F) ซึ่งมีคะแนนรวมเท่ากับ k ในกรณีที่ $\beta = 0$ สถิติ B มีการแจกแจงใกล้เคียงการแจกแจงแบบปกติมาตรฐาน [$N(0, 1)$] ข้อคำถามทำหน้าที่ต่างกันของข้อคำถาม ถ้าผลการสอบพบว่า $|B| > Z_{1-\frac{\alpha}{2}}$ อย่างมีนัยสำคัญที่ระดับ α แสดงว่า ปฏิเสธ H_0 นั่นคือ ข้อคำถามทำหน้าที่ต่างกันโดยเข้าข้างผู้สอบกลุ่มใดกลุ่มหนึ่ง

3. วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ (Multi-group confirmatory factor analysis: MG-CFA)

การวิเคราะห์องค์ประกอบเชิงยืนยัน (Confirmatory factor analysis)

สมการการวิเคราะห์องค์ประกอบเชิงยืนยัน สำหรับโมเดลการตอบสนองข้อสอบให้คะแนนเรียงลำดับ คือ

$$x_i^* = \lambda_{i1}\xi_1 + \lambda_{i2}\xi_2 + \cdots + \lambda_{iH}\xi_H + \delta_i, i = 1,2,\ldots, p \quad (69)$$

x_i^* แทน ตัวแปรแฝงที่มีค่าต่อเนื่อง เกิดจากข้อคำถาม (x_i^*) ที่ให้คะแนนหลายค่า

ξ_n แทน คะแนนองค์ประกอบคุณลักษณะแฝง

λ_{in} แทน ค่าน้ำหนักองค์ประกอบของข้อคำถามข้อที่ i บนองค์ประกอบ h หรืออาจเรียกว่าค่าสัมประสิทธิ์การถดถอยของ x_i^* บน ξ_n

และ δ_i แทนความคลาดเคลื่อนขององค์ประกอบและความคลาดเคลื่อนจากการวัด ซึ่งเป็นผลรวมจากองค์ประกอบ H ตัว ตัวแปรแบบไม่ต่อเนื่อง x_i จะได้รับการเปรียบเทียบโดยตัวแปร x_i^* กับค่าเทรสโฮล τ_{ij}

$$x_i = j \text{ ถ้า } \tau_{ij} < x_i^* < \tau_{i(j+1)} \text{ โดยที่ } j = 1,\ldots,k_i - 1 \quad (70)$$

เมื่อ k_i แทนจำนวนของรายการตอบแบบเรียงลำดับสำหรับข้อคำถามข้อที่ i และ $\tau_{i0} = -\infty$ และ $\tau_{ik} = +\infty$

ภายใต้การวิเคราะห์องค์ประกอบสำหรับโครงสร้างที่ซับซ้อนกับองค์ประกอบสององค์ประกอบที่นำมาวิเคราะห์ ซึ่งแต่ละองค์ประกอบมีค่าน้ำหนักสองค่า $\lambda_i S$ ต่อหนึ่งข้อ ซึ่งสามารถเขียนโมเดลได้ ดังนี้

$$x_i^* = \lambda_{i1}\xi_1 + \lambda_{i2}\xi_2 + \delta_i \quad \text{โดยที่ } i = 1,2,\ldots, p \quad (71)$$

อย่างไรก็ตาม สำหรับข้อคำถามที่มีน้ำหนักองค์ประกอบเฉพาะองค์ประกอบเดียว และ λ_{iH} สำหรับองค์ประกอบอื่น ๆ มีค่าเป็นศูนย์ จากสมการที่ 71 สามารถนำมาเขียนอยู่ในรูปเมทริกซ์ได้ ดังนี้

$$X^* = \Lambda\xi + \delta \quad (72)$$

เมื่อ X^* แทน เวกเตอร์ของข้อคำถาม ขนาด (p x 1)

Λ แทน เมทริกซ์น้ำหนักองค์ประกอบของตัวแปรที่วัดได้ P บนตัวแปรแฝง H

ξ แทน เวกเตอร์คะแนนองค์ประกอบ

และ δ แทน เวกเตอร์ของเศษเหลือจากการวัด และมีข้อสมมติว่า $(\xi, \delta) = 0$ และความแปรปรวนร่วมของข้อคำถาม คือ (Reise et al., 1993; Vandenburg & Lance, 2000)

$$\Sigma = \Lambda \Phi \Lambda' + \Theta_{\delta} \tag{73}$$

เมื่อ Σ แทน เมทริกซ์ความแปรปรวนร่วมของประชากร ระหว่างตัวแปรการวัด ในขนาด (p x p)

Φ แทน เมทริกซ์ความแปรปรวนร่วมระหว่างตัวแปรแฝง ขนาด (H x H)

และ Θ_{δ} แทน เมทริกซ์แนวทแยงของความแปรปรวนค่าเดียว (Unique variance)

ความสัมพันธ์ของพารามิเตอร์ ระหว่างการวิเคราะห์องค์ประกอบเชิงยืนยันกับการวิเคราะห์ตามทฤษฎีการตอบสนองข้อคำถาม

โดยปกติพารามิเตอร์ของการวิเคราะห์องค์ประกอบไม่ได้มีความสัมพันธ์โดยตรงกับพารามิเตอร์ของการวิเคราะห์ตามทฤษฎีการตอบสนองข้อคำถาม แต่เมื่อทำการแปลงน้ำหนักองค์ประกอบ (λ) และค่าเทรสโฮล (τ) ที่ใช้ในการประมาณพารามิเตอร์ของข้อคำถามสำหรับการวิเคราะห์พหุมิติระหว่างในข้อคำถาม (Between-item multidimensional) สำหรับโมเดลการวิเคราะห์ข้อคำถามพหุมิติแบบสองมิติ ในโมเดล MGRM สามารถคำนวณค่าพารามิเตอร์ α_{ih} จากน้ำหนักองค์ประกอบ ดังนี้ (McLeod, Swygert, & Thissen, 2001; Swygert, McLeod, & Thissen, 2001; Takane & de Leeuw, 1987)

$$\alpha_{ih} = \frac{(D)\lambda_{ih}}{\sqrt{1-(\Sigma_{h=1}^{H}\lambda_{ih}^{2} + 2\lambda_{i1}\Phi\lambda_{i2})}} \tag{74}$$

เมื่อ λ_{i1} คือ น้ำหนักองค์ประกอบสำหรับข้อที่ i บน ξ_1

λ_{i2} คือ น้ำหนังองค์ประกอบสำหรับข้อที่ i บน ξ_2

และ Φ คือ ความสัมพันธ์ระหว่าง ξ_1 และ ξ_2

ในแต่ละข้อคำถามของตัวแปรแฝงสองตัวจะมีพารามิเตอร์ความชัน α_{ih} หรือ λ_{ih} สองค่า

สำหรับโมเดลพหุมิติแบบสองมิติ เราสามารถแสดงพารามิเตอร์เทรสโฮลของรายการคำตอบในข้อคำถาม β_{ih} ในรูปของพารามิเตอร์องค์ประกอบเชิงยืนยัน

$$\beta_{ij} = \frac{\tau_{ij}}{\sqrt{1-(\Sigma_{h=1}^{H}\lambda_{ih}^{2} + 2\lambda_{i1}\Phi\lambda_{i2})}} \quad (75)$$

Kannan and Kim (2009) พบว่า การวิเคราะห์องค์ประกอบเชิงยืนยันด้วย Weighted least squares-means and variance adjustment (WLSMV) มีความแม่นยำในการประมาณค่าพารามิเตอร์สำหรับโมเดล MGRM ซึ่งมีค่า RMSD และมีค่าความลำเอียง (Bias) ต่ำ และยังพบว่า วิธี WLSMV ยังเป็นวิธีการประมาณค่าที่มีความคลาดเคลื่อนมาตรฐานมีขนาดเล็ก และใช้เวลาในการประมาณค่าพารามิเตอร์น้อย เมื่อเปรียบเทียบกับวิธีการประมาณค่าแบบภาวะความควรจะเป็น (Maximum likelihood) และประโยชน์ของการประมาณค่าด้วยวิธี WLSMV โดยเฉพาะอย่างยิ่งเมื่อมิติแฝงมีความสัมพันธ์กันสูง ($p \geq .5$) และ ค่าน้ำหนักองค์ประกอบในมิติที่สองมีค่ามาก ($\lambda = 0.3$)

วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ (Multi group confirmatory factor analysis: MG-CFA)

เพื่อแสดงความแตกต่างระหว่างค่าเฉลี่ยของกลุ่มจากคะแนนสังเกตได้ ซึ่งเป็นสิ่งที่สำคัญของความเท่าเทียมกันระหว่างกลุ่มในโมเดลการวัด เมื่อเราทำการเปรียบเทียบกลุ่มที่ต้องการศึกษา สิ่งที่สำคัญก็คือ ค่าตัวเลขที่ทำการวัดจะต้องอยู่ในมาตรวัดเดียวกัน หรืออีกนัยหนึ่งคือ ควรสร้างขอบเขตของคุณลักษณะที่ต้องการวัดของตัวแปรที่สามารถสังเกตได้เพื่อที่จะสามารถเปรียบเทียบในแต่ละกลุ่ม (g) ได้ ดังนั้นอาจจะกล่าวได้ว่า แบบทดสอบ/ แบบประเมินมี “ความเท่าเทียม/ ความไม่แปรเปลี่ยน” ระหว่างกลุ่ม (Measurement equivalence/ Invariance: ME/ I) (Reise et al., 1993; Vandenberg & Lance, 2000) และเมื่อค่าของตัวเลขที่กำหนดให้กับคุณลักษณะไม่เท่าเทียมกันข้ามกลุ่ม แสดงว่า การวัดต่างกลุ่มกันอาจจะเกิดการไม่เท่ากันแบบเทียม (Reise et al., 1993; Vandenberg & Lance, 2000)

ดังนั้น การเปรียบเทียบระหว่างกลุ่ม โดยปรับจากสมการที่ 74 และ 75 โดยใส่กลุ่มจะได้

$$x^{g} = \Lambda^{g}\xi^{g} + \delta^{g} \quad (76)$$

และ

$$\Sigma^{g} = \Lambda^{g}\Phi\Lambda^{g'} + \Theta_{\delta}^{g} \quad (77)$$

ข้อตกลงเบื้องต้นเกี่ยวกับ “ความเท่าเทียม/ ความไม่แปรเปลี่ยน” ที่ใช้ทดสอบสมมติฐานความแตกต่างกันระหว่างกลุ่ม (Horn & McArdle, 1992; Vandenburg & Lance, 2000)

1. ตัวแปรแฝงตามทฤษฎี (ξ^g) จะต้องมีค่าเท่ากันในแต่ละกลุ่ม

2. ความสัมพันธ์ λ^g ระหว่าง x^g และ ξ^g ต้องมีค่าเท่ากันระหว่างกลุ่ม

3. x^g ต้องได้รับอิทธิพลในระดับเดียวกันจากองค์ประกอบเฉพาะหนึ่งเดียว $\left(\delta^g\right)$ ข้ามกลุ่ม

การเปรียบเทียบข้ามกลุ่มสำหรับการดำเนินการทางการวัด ซึ่งเป็นการตรวจสอบความไม่แปรเปลี่ยนข้ามกลุ่ม การไม่ละเมิดข้อตกลงข้างต้น จะทำให้ข้อสรุปจากการตรวจสอบไม่แปรเปลี่ยนและเกิดข้อโต้แย้ง นอกจากนี้ ความน่าเชื่อถือของคะแนนและความเที่ยงตรงของคะแนนที่อ้างถึง จะเกิดข้อคำถาม (Horn & McArdle, 1992; Vandenburg & Lance, 2000) สำหรับข้อตกลงข้างต้น มีสมมติฐานบางอย่างเกี่ยวกับ ME/ I ระหว่างกลุ่มที่ต้องทำการทดสอบ โดยมีขั้นตอนของการทดสอบสมมติฐานที่เกี่ยวข้องกับ ME/ I ที่แสดงตามสมการที่ 76 และ 77

1. $\xi^g = \xi^{g'}$ นั่นคือ ชุดของข้อคำถาม *P* ข้อ ในกรอบแนวคิดเดียวกันของโครงสร้างคุณลักษณะแฝง ξ ในแต่ละกลุ่มเปรียบเทียบ '*g*' มีค่าเท่ากัน

2. $\Lambda^g = \Lambda^{g'}$ นั่นคือ น้ำหนักองค์ประกอบที่มีค่าเท่ากันระหว่างกลุ่ม

3. $\tau^g = \tau^{g'}$ นั่นคือ เทรสโฮลของรายการคำตอบมีค่าเท่ากันระหว่างกลุ่ม

4. รูปแบบของ CFA มีความเหมือนกันและถือว่ามีรูปแบบเดียวกันระหว่างกลุ่ม

5. $\Theta_\delta^g = \Theta_\delta^{g'}$ คือ ความแปรปรวนเท่ากันระหว่างกลุ่ม

6. $\Phi^g = \Phi^{g'}$ คือ ความแปรปรวนและความแปรปรวนร่วมของตัวแปรแฝงเท่ากันระหว่างกลุ่ม

การทดสอบความไม่แปรเปลี่ยนของการวัด

1. การทดสอบรวม (Omnibus test) ของเมทริกซ์ความแปรปรวนร่วม $\Sigma^g = \Sigma^{g'}$

เมทริกซ์ความแปรปรวนร่วมของกลุ่มตัวอย่าง $S^g = S^{g'}$ เป็นการเปรียบเทียบในลักษณะการดำเนินการของ MG-CFA โดยมีสมมติฐานศูนย์ คือ $\Sigma^g = \Sigma^{g'}$ โดยทดสอบด้วยสถิติไค-กำลังสองและสถิติทดสอบภาวะสารูปสนิทดี (Goodness of fit) และมีการทดสอบความแตกต่างระหว่างกลุ่มในแต่ละกลุ่มตัวอย่าง (Bollen & Long, 1993) ถ้าการทดสอบปรากฏว่าเป็นไปตามสมมติฐานศูนย์แล้วถือว่าการทดสอบสิ้นสุด แต่ถ้ามีการปฏิเสธสมมติฐานศูนย์ จะเริ่มกระบวนการทดสอบความเท่ากันของแต่ละคู่ของกลุ่มตัวอย่าง (Vandenburg & Lance, 2000) ซึ่งไม่มีข้อมูลในกรณีอื่น ๆ โดยเฉพาะอย่างยิ่ง การปฏิเสธ H_0 ไม่จำเป็นต้องแสดงให้เห็นถึงแหล่งที่มาของความไม่เท่ากันระหว่างกลุ่ม ดังนั้น ถ้าปฏิเสธสมมติฐานศูนย์แล้ว ME/ I กำลังทดสอบ ซึ่งเป็นการทดสอบที่มีความเข็มงวดของการทดสอบสมมติฐาน ซึ่งเป็นการรับประกันถึง ME/ I (Byrne, Shavelson, & Muthen, 1989) ท้ายสุด Vandenburg and Lance (2000) ในการวิเคราะห์อภิมาน (Meta-analysis) พบว่า มากกว่า

62% ที่บทความที่ใช้สถิติมีการทดสอบรวมเป็นอันดับแรก และน้อยกว่า 20% ของการงานด้านศึกษาประยุกต์ ใช้วิธีการทดสอบโดยตรง (Cheung & Rensvold, 2002)

2. ความไม่แปรเปลี่ยนของรูปแบบความสัมพันธ์ของตัวแปร (Configural invariance) $\Lambda^{g}_{form} = \Lambda^{g'}_{form}$.

การทดสอบความเท่าเทียมกันของรูปแบบความสัมพันธ์ของตัวแปร ซึ่งสมมติฐานศูนย์ที่มีรูปแบบตายตัวและน้ำหนักองค์ประกอบที่อิสระถูกกำหนดไว้ในข้อคำถามที่มีความเท่าเทียมกันระหว่างกลุ่ม ซึ่งเรียกว่า “Weak factorial invariance” (Horn & McArdle, 1992) ซึ่งการทดสอบนี้เป็นการทดสอบความสอดคล้องของโมเดลที่มีการเปรียบเทียบของเมทริกซ์ Λ สองกลุ่มหรือหลายกลุ่ม และถ้าความแตกต่างของกลุ่มคล้ายกันในแนวคิดของโครงสร้างคุณลักษณะแฝง โดยในแต่ละกลุ่มควรจะมีกลุ่มข้อคำถามย่อยเดียวกันในองค์ประกอบเดียวกัน (Cheung & Rensvold, 2002) โครงสร้างองค์ประกอบ มีข้อสมมติว่าเป็นเหตุผลที่แสดงถึงกรอบแนวคิดภายใต้การอ้างอิง (Reference) ควรมีการเปรียบเทียบระหว่างกลุ่ม (Vandenburg & Lance, 2000).ถ้ามีการปฏิเสธสมมติฐานศูนย์ ความเท่าเทียมกันของรูปแบบความสัมพันธ์ของตัวแปรที่ว่า $\Lambda^{g}_{form} = \Lambda^{g'}_{form}$ ในสองความหมาย คือ 1) แต่ละผลการตอบต่างใช้กรอบแนวคิดเดียวกัน 2) การทดสอบ ME/ I อาจจะยอมรับในการเปรียบเทียบของกลุ่ม (Cheung & Rensvold, 2002; Vandenburg & Lance, 2000)

อย่างไรก็ตาม ถ้าการทดสอบความเท่าเทียมกันของรูปแบบความสัมพันธ์ของตัวแปรปฏิเสธสมมติฐานศูนย์แล้ว สรุปได้ว่า การปฏิเสธนี้คือการรับรองทดสอบ ME/ I (Vandenburg & Lance, 2000) แต่การทดสอบความเท่าเทียมกันของรูปแบบความสัมพันธ์ของตัวแปรไม่เป็นจริง ยกตัวอย่างเช่น เมื่อผู้ทดสอบมาจากวัฒนธรรมต่างกัน (กลุ่มย่อย) แสดงถึงความแตกต่างของค่าเฉลี่ยและกรอบแนวคิดของกลุ่มอ้างอิงจากทฤษฎี/ โครงสร้าง (Cheung & Rensvold, 2002) นอกจากนี้ สมมติฐานที่ว่า $\Lambda^{g}_{form} = \Lambda^{g'}_{form}$ อาจจะทำให้เกิดการปฏิเสธ H_0 อันเนื่องมาจากมีปัญหาจากการเก็บข้อมูล (Data collection) มีความคลาดเคลื่อนจากการแปลงข้อมูลและแปลงข้อมูลกลับ (Translation and back-translation errors) ความคลาดเคลื่อนจากการบริการจัดการการสำรวจ (Survey administration) และความคลาดเคลื่อนเนื่องจากการสอน (Instructional errors) และอื่น ๆ อีก (Cheung & Rensvold, 2002)

ถ้าโครงสร้างในการทดสอบไม่แปรเปลี่ยนระหว่างกลุ่มแล้ว การเปรียบเทียบที่ได้จะไม่มีความหมาย นอกจากนี้ การทดสอบไม่สามารถแสดงให้เห็นเกี่ยวกับการทดสอบและเปรียบเทียบ ME/ I ถ้าข้อคำถามที่มีขนาดเท่ากัน ดังนั้น โครงสร้างจะไม่สามารถเปรียบเทียบระหว่างกลุ่มได้

และการทดสอบความเท่าเทียมกันของรูปแบบความสัมพันธ์ของตัวแปรจำเป็นต้องเป็นที่ยอมรับ เพื่อให้การทดสอบ ME/ I ที่เป็นไปตามสมมติฐานและมีความหมาย (Vandenburg & Lance, 2000)

3. ความไม่แปรเปลี่ยนของเมทริกซ์ $\Lambda^g = \Lambda^{g'}$.

ความไม่แปรเปลี่ยนของเมทริกซ์เป็นการทดสอบที่มีสมมติฐานศูนย์ คือ พารามิเตอร์น้ำหนักองค์ประกอบมีค่าเท่ากันระหว่างกลุ่ม (Cheung & Rensvold, 2002; Vandenburg & Lance, 2000) เป็นการทดสอบที่มีความแข็งแกร่งของความสัมพันธ์ระหว่างข้อคำถาม ภายใต้โครงสร้างทฤษฎี (Construct) แสดงให้เห็นว่า ถ้าโครงสร้างทางทฤษฎีมีความเหมือนกันข้ามกลุ่ม จะสรุปได้ว่า $\Lambda^g = \Lambda^{g'}$ มีค่าเท่ากันในสเกลเดียวกันข้ามกลุ่ม (Schmitt, 1982; Vandenburg & Lance, 2000) ความไม่แปรเปลี่ยนของเมทริกซ์ที่สมบูรณ์เป็นการแสดงถึงความไม่แปรเปลี่ยนของเมทริกซ์ในระดับโครงสร้างทางทฤษฎี (Construct-level metric invariance) (Cheung & Rensvold, 2002)

ข้อมูลที่ได้รับจากประชากรที่มีความแตกต่างกันอาจจะแสดงให้เห็นถึงข้อตกลงในเทอมของชนิดและตัวเลขที่อยู่ภายใต้โครงสร้างทฤษฎีและข้อคำถามในแต่ละโครงสร้างทฤษฎี (ความเท่าเทียมกันของรูปแบบความสัมพันธ์ของตัวแปร) ความแข็งแกร่งของความสัมพันธ์ระหว่างข้อคำถามที่มีสเกลเฉพาะกับโครงสร้างทฤษฎีอาจจะเปลี่ยนแปลงข้ามกลุ่ม (Cheung & Rensvold, 2002) การทดสอบความไม่แปรเปลี่ยนของเมทริกซ์เต็มรูปจะสำเร็จได้โดยการกำหนดให้น้ำหนักองค์ประกอบคงที่ (λ_{ih}) ทั้งหมดให้เหมือนกันข้ามกลุ่ม โดยเชื่อว่าโมเดลที่คงไว้ถูกต้องทั้งสองกลุ่ม และ $\Lambda^g = \Sigma^{g'}$ อย่างไม่มีเงื่อนไข ดังนั้น ความไม่แปรเปลี่ยนของเมทริกซ์เป็นการทดสอบที่เข้มงวดกว่าการทดสอบความไม่แปรเปลี่ยนของรูปแบบความสัมพันธ์ (Horn & McArdle, 1992) และมีความสำคัญที่จำเป็นต้องมีการทดสอบก่อนการเปรียบเทียบข้ามกลุ่ม (Bollen, 1989)

ความไม่แปรเปลี่ยนของเมทริกซ์บางส่วน (Partial metric invariance)

เมื่อมีการปฏิเสธสมมติฐานศูนย์ ที่ว่า $\Lambda^g = \Lambda^{g'}$ อาจจะเกิดจากความไม่ชัดเจนในทฤษฎี (Reise et al., 1993; Vandenburg & Lance, 2000) นักวิจัยบางท่านกล่าวว่า การปฏิเสธสมมติฐานศูนย์สำหรับความไม่แปรเปลี่ยนของเมทริกซ์แบบเต็มรูป ทำให้หมดข้อสงสัยเกี่ยวกับการทดสอบ ME/ I ซึ่งมีความหมายเหมือนกับการทดสอบความไม่แปรเปลี่ยนของรูปแบบตัวแปร อย่างไรก็ตาม นักวิจัยคนอื่น ๆ เช่น Byrne, Shavelson and Muthen (1989) เห็นด้วยกับ “ความไม่แปรเปลี่ยนของเมทริกซ์บางส่วน” ซึ่งควรมีการทดสอบเมื่อไม่พบความไม่แปรเปลี่ยนของเมทริกซ์แบบเต็มรูป ซึ่งทำด้วยการทดสอบบางส่วน ด้วยการไม่กำหนดค่าคงที่ใน Λ เป็นการไม่แปรเปลี่ยนข้ามกลุ่ม (Byrne et al., 1989)

Cheung and Rensvold (2002) ได้กล่าวถึงความแปรเปลี่ยนบางส่วน คือ ความไม่แปรเปลี่ยนของระดับข้อคำถามในเมทริกซ์ (Item-level metric invariance) หรือความไม่แปรเปลี่ยนของ

น้ำหนักองค์ประกอบ (Factor loading invariance) พวกเขาแสดงให้เห็นว่า ชุดของการทดสอบความไม่แปรเปลี่ยนของระดับข้อคำถามควรได้รับการยอมรับ ถ้าความไม่แปรเปลี่ยนของเมทริกซ์โครงสร้างแต่ละระดับไม่ตรวจพบ นอกจากนี้ยังสามารถเลือกหนึ่งรายการของการตอบสนองข้อคำถามสำหรับความแปรเปลี่ยนทั้งหมดของเมทริกซ์น้ำหนักองค์ประกอบ (Cheung & Rensvold, 2002) การทดสอบระดับของข้อคำถามสามารถใช้โปรแกรมในการวิเคราะห์ได้ โดย Modification indices (MIs) แสดงถึงค่าคงที่ของพารามิเตอร์เมื่อมีองศาอิสระเท่ากับ 1 โดยที่ MIs แสดงถึงค่าไค-กำลังสองที่มีการเปลี่ยนแปลงเมื่อมีตัวแปรมีการเพิ่มหรือตัดออก (Reise et al., 1993) ซึ่งมันเป็นวิธีที่ช่วยทำให้นักวิจัยค้นหาข้อคำถามย่อยที่มีการแปรเปลี่ยนของน้ำหนักองค์ประกอบที่ไม่มีความแปรเปลี่ยนระหว่างกลุ่ม (Cheung & Rensvold, 2002; Reise et al., 1993) นอกจากนี้ข้อคำถามอาจจะไม่แปรเปลี่ยนข้ามกลุ่มซึ่งเป็นการพิจารณาถึงความแตกต่างกันของข้อคำถาม DIF

การทดสอบความไม่แปรเปลี่ยนเมทริกซ์บางส่วนเป็นการสนับสนุนการวัดความไม่เท่ากันโดยเฉพาะตัวบ่งชี้ที่ไม่ตอบสนองกับค่าคงที่ไม่แปรเปลี่ยน และช่วยในการทดสอบ ME/ I ข้อนี้ถือว่าเป็นสิ่งที่สำคัญที่ทำให้เกิดการโต้แย้งกันในเรื่องของการทดสอบความไม่แปรเปลี่ยนบางส่วนซึ่งเกิดจากเหตุผล 2 ข้อ คือ 1) ไม่มีคงเส้นคงวาของค่าวิกฤตที่ใช้สำหรับความไม่แปรเปลี่ยนของค่าคงที่ในทฤษฎี 2) ค่าคงที่ของความแปรเปลี่ยนบางส่วนเกิดจากการวิเคราะห์เชิงสำรวจเป็นส่วนใหญ่และมีโอกาสทำให้เกิดประโยชน์ในอนาคตอย่างมหาศาล (Vandenburg & Lance, 2000) สำหรับการเปรียบเทียบข้ามกลุ่มนั้นมีความหมาย เพราะสิ่งที่เกิดขึ้นไม่ได้เกิดขึ้นโดยบังเอิญ และข้อคำถามส่วนใหญ่บนตัวแปรแฝงควรมีน้ำหนักที่ไม่แปรเปลี่ยนข้ามกลุ่ม (Reise et al., 1993) และข้อคำถามที่มีความแปรเปลี่ยนควรจะมีได้เล็กน้อยในโมเดล (Cheung & Rensvold, 2002; Vandenburg & Lance, 2000) ในการทบทวนวรรณกรรมสำหรับการทำหน้าที่ต่างกันของข้อคำถาม ซึ่งเป็นการเสนอแนะเกี่ยวกับข้อคำถามที่แปรเปลี่ยนที่ต้องตัดออกจากแบบทดสอบหรือแบบประเมิน นอกจากนี้ Vandenburg and Lance (2000) ได้ให้ข้อเสนอแนะเกี่ยวกับการตัดข้อคำถามว่าความไม่แปรเปลี่ยนบางส่วนที่เกิดขึ้นต้องมีการยืนยันภายใต้ทฤษฎีที่เข้มแข็ง

4. ความไม่แปรของความแปรปรวน $\Theta_{\delta}^{g} = \Theta_{\delta}^{g'}$.

สมมติฐานศูนย์ของการทดสอบ คือ ความแปรปรวนของเศษเหลือ (Residual variance) มีค่าเท่ากันข้ามกลุ่ม ถ้าสเกลการวัดข้อคำถามภายใต้โครงสร้างคุณลักษะแฝงมีความคลาดเคลื่อนของการวัดระดับเดียวกัน (Cheung & Rensvold, 2002; Vandenburg & Lance, 2000) การทดสอบนี้อยู่ภายใต้ข้อคำถามที่มีค่าคงที่ ที่มีค่าเท่ากันระหว่างกลุ่ม (Vandenburg & Lance, 2000) ถ้าผู้สอบอยู่ในหนึ่งหรือมากกว่าหนึ่งกลุ่ม ต่างมีค่าไม่เท่ากันในคะแนนแบบเดียวกัน และมีแนวโน้มจะตอบสนองอย่างต่อเนื่องในข้อคำถาม (Millsap & Everson, 1993) นอกจากนี้ ความแตกต่าง

ระหว่างกลุ่มในเรื่องของคำศัพท์ (Vocabulary) ไวยากรณ์ (Grammar) วากยสัมพันธ์ (Syntax) และประสบการณ์ร่วมกัน อาจจะทำให้ความแปรปรวนของเศษเหลือไม่มีความเท่าเทียมกัน (Millsap, 1995)

การทดสอบความไม่แปรเปลี่ยนของโครงสร้าง (Tests of structural invariance)

1. ความไม่แปรเปลี่ยนความแปรปรวนขององค์ประกอบ $\Phi_{pp}^{g} = \Phi_{pp}^{g'}$.

การทดสอบนี้มีสมมติฐานศูนย์ที่ว่า ความแปรปรวนขององค์ประกอบต่างไม่แปรเปลี่ยนข้ามกลุ่ม ความแปรปรวนขององค์ประกอบเป็นตัวแทนของการกระจายหรือความแปรผันได้ของตัวแปรแฝง และการทดสอบนี้พบบ่อยในการทดสอบความไม่แปรเปลี่ยนของเมทริกซ์ (Schmitt, 1982) ความแตกต่างของความแปรปรวนขององค์ประกอบที่มีความแตกต่างกันของคะแนนจริงที่เทียบกันข้ามกลุ่ม การปฏิเสธสมมติฐานศูนย์ แสดงว่า กลุ่มที่มีความแปรปรวนน้อยกว่าในองค์ประกอบ จะมีแนวโน้มที่มีระยะห่างของโครงสร้างของต่อเนื่องน้อยกว่า (Vandenburg & Lance, 2000)

2. ความไม่แปรเปลี่ยนของความแปรปรวนร่วมขององค์ประกอบ $\Phi_{pp'}^{g} = \Phi_{pp'}^{g'}$.

การทดสอบนี้มีสมมติฐานที่กล่าวว่า ความแปรปรวนร่วมขององค์ประกอบต่างไม่แปรเปลี่ยนข้ามกลุ่ม ซึ่งการทดสอบนี้ใช้บ่อยสำหรับการทดสอบความไม่แปรเปลี่ยนแบบ (Configural invariance) ความแตกต่างในความแปรปรวนร่วมขององค์ประกอบซึ่งเป็นความแตกต่างของกรอบแนวคิดของคะแนนจริง (Schmitt, 1982; Vandenburg & Lance, 2000) การทดสอบความไม่แปรเปลี่ยนของความแปรปรวนขององค์ประกอบ และเมทริกซ์ความแปรปรวนร่วม ซึ่งเป็นการรวมการทดสอบในภาพรวมของเมทริกซ์ความแปรปรวนร่วมและความแปรปรวนของตัวแปรแฝงข้ามกลุ่ม เช่น $\phi^{g} = \phi^{g'}$ (Byrne et al., 1989; Vandenburg & Lance, 2000) อย่างไรก็ตาม มีหลายอย่างที่ไม่เป็นเช่นนั้น เพราะการทดสอบของความแปรปรวนร่วมขององค์ประกอบไม่แปรเปลี่ยนไม่ได้ทำการทดสอบแยกต่างหาก ดังนั้น นักวิจัยส่วนใหญ่กล่าวว่าไม่เห็นด้วยความเชื่อนี้มากนักในเรื่องของการทดสอบความเท่ากันของเมทริกซ์ความแปรปรวนร่วมขององค์ประกอบ ซึ่งเป็นการทดสอบหนึ่งในการ Configural invariance

3. ความไม่แปรเปลี่ยนของค่าเฉลี่ยองค์ประกอบ $\kappa^{g} = \kappa^{g'}$.

การทดสอบสมมติฐานศูนย์ของความไม่แปรเปลี่ยนของค่าเฉลี่ยองค์ประกอบข้ามกลุ่ม ซึ่งเป็นการทดสอบคล้ายกับการทดสอบค่าเฉลี่ยแต่ละกลุ่มแบบ ANOVA และเป็นการเริ่มต้นการทดสอบภาพรวมของค่าเฉลี่ยทั้งหมดก่อนที่จะมีการทดสอบเฉพาะกรณีย่อย (คล้ายกับการทดสอบภายหลัง (Post-hoc) ซึ่งเป็นการอธิบายความแตกต่างระหว่างสองกลุ่ม (Cheung & Rensvold, 2002; Schmitt, 1982; Vandenburg & Lance, 2000) ซึ่งใช้ในการทดสอบความไม่แปรเปลี่ยนของค่าเฉลี่ย

องค์ประกอบ อย่างไรก็ตาม เป็นแนวคิดของการเปรียบเทียบค่าเฉลี่ยที่นิยมใช้ทดสอบ ซึ่งเป็นการทดสอบโดยตรงสำหรับเพื่อให้ความคลาดเคลื่อนในการวัดมีความเที่ยงตรงมากขึ้น (Schmitt, 1982) และเป็นการควบคุมการไม่เท่าเทียมกันของการวัดบางส่วน โดยการบังคับให้ความไม่แปรเปลี่ยนบางส่วนเกิดผล (Byrne et al., 1989)

การทดสอบที่แตกต่างกันทั้งหมดเจ็ดข้อสำหรับการไม่แปรเปลี่ยนที่กล่าวมาข้างต้นเป็นการทดสอบภายใต้กรอบแนวคิดของ MG-CFA การทดสอบเหล่านี้อาจจะไม่ได้ใช้ในการทดสอบทั้งหมด อยู่กับผู้วิจัย (Vandenburg & Lance, 2000) โดยส่วนใหญ่จะใช้การทดสอบเกี่ยวกับเมทริกซ์ความแปรปรวนร่วมก่อน (Bagozzi & Edwards, 1998; Byrne et al., 1989; Horn & McArdle, 1992) และงานวิจัยหลายงานก็มุ่งหวังว่าเมทริกซ์ความแปรปรวนร่วมจะมีความไม่แปรเปลี่ยน เพราะเป็นการทดสอบที่สำคัญสำหรับการทดสอบ ME/ I (Bagozzi & Edwards, 1998; Horn & McArdle, 1992; Jöreskog, 1971) นอกจากนี้ Vandenburg and Lance (2000) พบว่า เมทริกซ์และความไม่แปรเปลี่ยนของเมทริกซ์บางส่วนเป็นการทดสอบที่พบมากในวิธี MG-CFA

อย่างไรก็ตาม ในมุมมองของการทดสอบความแตกต่างกันของข้อคำถาม การนำการทดสอบ MG-CFA มาเพื่อทำการทดสอบความแตกต่างกัน โดยใช้การทดสอบความไม่แปรเปลี่ยนผู้วิจัยจะทำการทดสอบความไม่แปรเปลี่ยนของพารามิเตอร์ ได้แก่ พารามิเตอร์จุดตัด (Intercept) ซึ่งเปรียบเหมือนพารามิเตอร์ความยากของข้องสอบ เนื่องจากงานวิจัยนี้เป็นข้อคำถามให้คะแนนหลายค่า ดังนั้น พารามิเตอร์จุดตัดของตัวแปรสังเกตได้ (Indicators) ก็คือ พารามิเตอร์ความยากของแต่ละรายการตอบ หรือเรียกว่า เทรสโฮล (Threshold) อีกพารามิเตอร์หนึ่งที่ใช้สำหรับการทดสอบคือ พารามิเตอร์น้ำหนักองค์ประกอบ (Factor loading) ซึ่งก็คือ พารามิเตอร์อำนาจจำแนกของข้อคำถาม ซึ่งการทดสอบเหล่านี้สามารถทำการทดสอบด้วยการทดสอบความไม่แปรเปลี่ยน โดยใช้ซอฟต์แวร์ที่สามารถวิเคราะห์ SEM ได้ เช่น Mplus, EQS, LISREL และ อื่น ๆ การศึกษาครั้งนี้ผู้วิจัยจะทำการทดสอบความไม่แปรเปลี่ยนของพารามิเตอรอำนาจจำแนกและความไม่แปรเปลี่ยนของเทรสโฮล ซึ่งเป็นการทดสอบที่มีความสำคัญมากกว่าการทดสอบอื่น ๆ ภายใต้พื้นฐานของการทดสอบความแตกต่างกันของข้อคำถามภายใต้ทฤษฎีการตอบสนองข้อคำถาม

4. วิธีการถดถอยโลจิสติก (Logistic regression: LR)

Swaminathan and Rogers (1990) ได้นำโมเดลการถดถอยโลจิสติกไปประยุกต์ใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนสองค่า โดยการดัดแปลงมาจากวิธีล็อกลิเนียร์และโมเดลโลจิท ของ Mellenbergh รวมทั้งเชื่อมแนวคิดระหว่างวิธีแมนเทล-แฮนส์เซล ของ Holland and Thayer วิธีการทำให้เป็นมาตรฐาน ที่พัฒนาโดย Dorans and Kulick วิธีซิปเทสท์ ของ Shealy and Stout และวิธีที่ใช้ทฤษฎีการตอบสนองข้อสอบ หลักการตรวจสอบจะใช้โมเดล

การถดถอยโลจิสติกทำนายความน่าจะเป็นในการตอบข้อสอบถูก โดยใช้คะแนนรวมแทนระดับความสามารถ ซึ่งสมมติว่าเป็นตัวแปรความสามารถแบบต่อเนื่อง โมเดลดังกล่าวมีพจน์สำหรับทดสอบปฏิสัมพันธ์ระหว่างการเป็นสมาชิกของกลุ่มผู้สอบ กับระดับความสามารถ ดังนั้น จึงสามารถตรวจสอบการทำหน้าที่ต่างกันที่เป็นแบบรูปแบบเดียวกัน และข้อสอบทำหน้าที่ต่างกันในรูปแบบเดียวกัน โมเดลการถดถอยโลจิสติกมีความยืดหยุ่น สามารถนำไปปรับขยายเพื่อใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนหลายค่า (French & Miller, 1996; Zumbo, 1999) และตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่วัดความสามารถหลายมิติ (Multidimensional) (Mazor et al., 1998)

แนวคิดและหลักการ

ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจะใช้โมเดลการถดถอยโลจิสติกทำนายความน่าจะเป็นในการตอบถูก ดังนี้

$$P(u=1|\theta)=\frac{e^{(\beta_{0+}\beta_1\theta)}}{1+e^{(\beta_{0+}\beta_1\theta)}} \tag{78}$$

เมื่อ u แทน ผลการตอบข้อสอบ

θ แทน ความสามารถที่สังเกตได้ของผู้สอบ

β_0 แทน พารามิเตอร์ส่วนตัด (Intercept parameter)

β_1 แทน พารามิเตอร์ความชัน (Slope parameter)

สมการดังกล่าว เป็นโมเดลการถดถอยโลจิสติกแบบมาตรฐาน ซึ่งใช้ทำนายตัวแปรตามแบบสองค่าจากตัวแปรอิสระที่กำหนดให้ โมเดลการถดถอยโลจิสติกสามารถนำไปประยุกต์ใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม โดยสามารถแยกสมการกลุ่มผู้สอบสองกลุ่มที่สนใจ ได้ดังนี้

$$P(u_{ij}=1|\theta_{ij})=\frac{e^{(\beta_{0j+}\beta_{1j}\theta_{ij})}}{1+e^{(\beta_{0j+}\beta_{1j}\theta_{ij})}} \quad i=1,...,n,\ j=1,2,... \tag{79}$$

เมื่อ u_{ij} แทน ผลการตอบข้อสอบของผู้สอบคนที่ i ใน กลุ่ม j

θ_{ij} แทน ผลการตอบข้อสอบของผู้สอบคนที่ i ใน กลุ่ม j

β_{0j} แทน พารามิเตอร์ส่วนตัดในกลุ่ม j

β_{1j} แทน พารามิเตอร์ความชันในกลุ่ม j

จากนิยามของการทำหน้าที่ต่างกันของข้อสอบ (Differential item functioning: DIF) กำหนดไว้ว่า ข้อสอบที่ทำหน้าที่ต่างกัน ถ้าผู้สอบมีความสามารถเท่ากัน แต่มาจากกลุ่มผู้สอบต่างกัน มีความน่าจะเป็นในการตอบข้อสอบถูกไม่เท่ากัน (Hambleton & Swaminathan, 1985 cited in Swaminathan & Rogers, 1990) จากนิยามดังกล่าว เมื่อพิจารณาโมเดลการถดถอยโลจิสติก สามารถสรุปได้ว่า ถ้า $\beta_{01} = \beta_{02}$ และ $\beta_{11} = \beta_{12}$ แล้ว โค้งการถดถอยโลจิสติกระหว่างกลุ่มผู้สอบเท่ากัน แสดงว่า ข้อคำถามทำหน้าที่ต่างกัน ถ้า $\beta_{11} = \beta_{12}$ และ $\beta_{01} \neq \beta_{02}$ แล้ว โค้งการถดถอยโลจิสติกระหว่างกลุ่มผู้สอบขนานกันแต่ไม่ทับกัน แสดงว่า ข้อสอบทำหน้าที่ต่างกันเป็นรูปแบบเดียวกัน และถ้า $\beta_{01} = \beta_{02}$ และ $\beta_{11} \neq \beta_{12}$ แล้ว โค้งการถดถอยโลจิสติกระหว่างกลุ่มผู้สอบไม่ขนาดกัน แสดงว่า ข้อสอบทำหน้าที่ต่างกันไม่เป็นรูปแบบเดียวกัน

กระบวนการตรวจสอบ

ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจะใช้โมเดลการถดถอยโลจิสติกทดสอบความแตกต่างของผลการตอบข้อสอบ ระหว่างกลุ่มผู้สอบที่ระดับความสามารถเดียวกัน โดยสามารถทดสอบการทำหน้าที่ต่างกันของข้อสอบที่เป็นรูปแบบเดียวกัน และที่ไม่เป็นรูปแบบเดียวกันได้พร้อมกัน ดังนี้

$$P(u_{ij} = 1) = \frac{e^{Z_{ij}}}{1 + e^{Z_{ij}}} \tag{80}$$

เมื่อ $$Z_{ij} = \tau_0 + \tau_1\theta + \tau_2 g + \tau_3(\theta g) \tag{81}$$

โมเดลในสมการที่ 81 สมมูลกับโมเดลในสมการที่ 78 โดยรวมพารามิเตอร์ของข้อสอบทำหน้าที่ต่างกันที่เป็นรูปแบบเดียวกัน และที่ไม่เป็นรูปแบบเดียวกันไว้ในโมเดลเดียวกัน สำหรับ τ_0 เป็นพารามิเตอร์ส่วนตัด τ_1 τ_2 และ τ_3 เป็นค่าสัมประสิทธิ์ของระดับความสามารถของผู้สอบ (θ) การเป็นสมาชิกของกลุ่มผู้สอบ (g) และปฏิสัมพันธ์ระหว่างระดับความสามารถของผู้สอบกับการเป็นสมาชิกของกลุ่มผู้สอบ (θ_g) ตามลำดับ ส่วนตัวแปร g ที่ใช้แทนสมาชิกของกลุ่มผู้สอบจะกำหนดรหัสเป็นตัวดัมมี่ โดยที่ g = 1 หมายถึง ถ้าผู้สอบเป็นสมาชิกในกลุ่มอ้างอิง และ g = 0 หมายถึง ถ้าผู้สอบเป็นสมาชิกในกลุ่มสนใจ

จากโมเดลการถดถอยโลจิสติกในสมการที่ 80 พจน์ θ_g เป็นผลคูณของตัวแปรอิสระ 2 ตัว คือ θ และ g ซึ่งกำหนดรหัสตัวแปรดังข้างบน สำหรับพารามิเตอร์ τ_2 และ τ_3 มีความเกี่ยวข้องกับพารามิเตอร์ของโมเดลในสมการที่ 79 ดังนี้

$$\tau_2 = \beta_{01} - \beta_{02} \tag{82}$$

$$\tau_3 = \beta_{11} - \beta_{12} \tag{83}$$

สำหรับสมการที่ 80 สามารถเขียนให้อยู่ในรูปล๊อกธรรมชาติของอัตราส่วนแต้มต่อ (Odd ratio) ได้ดังนี้ (Zumbo, 1999, p. 23)

$$\ln[\frac{P}{1-P}] = \tau_0 + \tau_1\theta + \tau_2 g + \tau_3(\theta g) \tag{84}$$

การประมาณค่าพารามิเตอร์

ในการประมาณค่าพารามิเตอร์ของข้อสอบจะใช้การประมาณค่าความน่าจะเป็นมากที่สุด (Maximum likelihood estimation: MLE) โดยใช้ฟังก์ชันในการคำนวณดังนี้ (Narayanan & Swaminathan, 1996)

$$L(u_{ij} \mid \theta) = \prod_{i=1}^{N}\prod_{j=1}^{n} P(u_{ij})^{u_{ij}}[1 - P(u_{ij})]^{1-u_{ij}} \tag{85}$$

เมื่อ N แทนขนาดตัวอย่าง n แทนความยาวของแบบทดสอบ u_{ij} และ $P(u_{ij})$ เป็นค่าที่คำนวณจากสมการที่ 80 ในการประมาณค่าพารามิเตอร์โดยใช้วิธีการประมาณค่าความน่าจะเป็นมากที่สุดมีการแจกแจงปกติหลายตัวแปรแบบเชิงเส้นกำกับ (Asymptotically multivariate normal) โดยมีเวกเตอร์เฉลี่ย τ และเมทริกซ์ความแปรปรวน-ความแปรปรวนร่วม Σ ดังนี้

$$\hat{\tau} \sim N(\tau, \Sigma) \tag{86}$$

เมื่อ $\hat{\tau}' = [\tau_0, \tau_1, \tau_2, \tau_3]$ สำหรับค่าความคลาดเคลื่อนมาตรฐานแบบเชิงเส้นกำกับของค่าประมาณ $\tau_s (s = 0,...,3)$ สามารถคำนวณจากรากที่สองของสมาชิกในแนวเส้นทแยงมุมสี่ค่าของเมทริกซ์ Σ ดังนี้

$$SE(\hat{\tau} = \sqrt{\Sigma^{ss}}) \tag{87}$$

การทดสอบสมมติฐาน

การทดสอบสมมติฐานของข้อสอบทำหน้าที่เบี่ยงเบนที่เป็นรูปแบบเดียวกัน (Uniform DIF) และข้อสอบทำหน้าที่เบี่ยงเบนที่ไม่เป็นรูปแบบเดียวกัน (Nonuniform DIF) จะทดสอบพร้อมกัน โดยกำหนดสมมติฐานศูนย์ $H_0 : \tau_2 = 0$ และ $H_0 : \tau_3 = 0$ และสมมติฐานทางเลือก $H_1 : \tau_2 \neq 0$ หรือ $H_1 : \tau_3 \neq 0$ พารามิเตอร์ τ_2 บ่งชี้ถึงความแตกต่างของค่าเฉลี่ยในผลการตอบข้อสอบของกลุ่มผู้สอบ และ τ_3 บ่งชี้ถึงปฏิสัมพันธ์ระหว่างระดับความสามารถกับการเป็นสมาชิกของกลุ่มผู้สอบ ถ้า $\tau_2 \neq 0$ และ $\tau_3 > 0$ แสดงว่า ข้อคำถามทำหน้าที่ต่างกันเป็นรูปแบบเดียวกัน โดยเข้าข้างกลุ่มอ้างอิง เมื่อ $\tau_2 > 0$ หรือเข้าข้างกลุ่มสนใจเมื่อ $\tau_2 < 0$ แต่ถ้า $\tau_3 \neq 0$ ($\tau_2 = 0$ หรือ $\tau_2 \neq 0$ ก็ได้) แสดงว่า ข้อสอบทำหน้าที่เบี่ยงเบนที่ไม่เป็นรูปแบบเดียวกัน เมื่อ $\tau_3 > 0$ ข้อสอบจะเข้าข้างกลุ่มอ้างอิงที่ระดับความสามารถสูงกว่า และเข้าข้างกลุ่มสนใจที่ระดับความสามารถต่ำกว่า ในทางตรงกันข้าม เมื่อ $\tau_3 < 0$ ข้อสอบจะเข้าข้างกลุ่มสนใจที่ระดับความสามารถสูงกว่า และเข้าข้างกลุ่มอ้างอิงที่ระดับความสามารถต่ำกว่า สำหรับสมมติฐานที่ใช้ทดสอบข้อสอบทำหน้าที่เบี่ยงเบนสองรูปแบบพร้อมกัน เป็นดังนี้

$$H_0 : C_\tau = 0$$
$$H_0 : C_\tau \neq 0 \qquad (88)$$

เมื่อ
$$C = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad (89)$$

สถิติที่ใช้ทดสอบสมมติฐานมีลักษณะ ดังนี้

$$\chi^2 = \hat{\tau} C' \left(C \Sigma C' \right)^{-1} C \hat{\tau}' \qquad (90)$$

สถิติการทดสอบดังกล่าวมีการแจกแจงแบบไค-กำลังสอง $\left(\chi^2\right)$ ที่ระดับชั้นของความเป็นอิสระเท่ากับ 2 ($df = 2$) เมื่อทดสอบนัยสำคัญทางสถิติโดยใช้สมการที่ 90 แล้ว มีค่ามากกว่า 2 $\chi^2_{\alpha/2}$ แสดงว่า ปฏิเสธสมมติฐานของการทำหน้าที่ไม่ต่างกันของข้อคำถาม (No-DIF) Jodoin and Gierl (2001) ได้เสนอดัชนี $R^2\Delta$ เพื่อใช้จำแนกขนาดของการทำหน้าที่ต่างกันของข้อคำถาม โดยมีพื้นฐานจากค่าประมาณกำลังสองน้อยที่สุดแบบถ่วงน้ำหนัก (Weighted least squares estimate) ดัชนี $R^2\Delta$ มี 3 ระดับ ดังนี้

DIF ระดับ A ขนาดเล็ก: $R^2\Delta < 0.035$

DIF ระดับ B ขนาดปานกลาง: ปฏิเสธสมมติฐานศูนย์ และ $0.035 \le R^2\Delta < 0.070$

DIF ระดับ C ขนาดใหญ่: ปฏิเสธสมมติฐานศูนย์ และ $R^2\Delta \ge 0.070$

5. วิธีการถดถอยโลจิสติกแบบจัดอันดับ (Ordinal logistic regression: OLR)

นักวิจัยหลายคนได้นำโมเดลการถดถอยโลจิสติกไปประยุกต์ใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ในข้อสอบที่มีการให้คะแนนหลายค่า (Polytomous) ดังเช่น Miller and Spray (1993) ได้เสนอการวิเคราะห์ฟังก์ชันการจำแนกโลจิสติก (Logistic discriminant function analysis) ส่วน French and Miller (1996) ได้ปรับขยายโมเดลการถดถอยโลจิสติก 3 โมเดล คือ โมเดลโลจิตของอัตราส่วนแบบต่อเนื่อง (Continuation ratio logits) โมเดลโลจิตแบบสะสม (Cumulative logits) และโมเดลโลจิตของรายการคะแนนที่อยู่ติดกัน (Adjacent categories logits) ต่อมา Zumbo (1999) ได้ใช้โมเดลการถดถอยโลจิสติกแบบจัดอันดับ (Ordinal logistic regression model) พร้อมกับเสนอแนวทางการทดสอบนัยสำคัญทางสถิติในการตัดสินข้อสอบทำหน้าที่เบี่ยงเบนที่เป็นรูปแบบเดียวกัน และไม่เป็นรูปแบบเดียวกัน โดยใช้การทดสอบโมเดล 3 ขั้นตอน นอกจากนี้ยังได้พัฒนาดัชนีการวัดขนาดอิทธิพลของการทำหน้าที่ต่างกันของข้อคำถาม (DIF effect size) โดยคำนวณค่า R^2 ที่ได้จากการทดสอบในแต่ละโมเดล

แนวคิดและหลักการ

การวิเคราะห์โดยใช้โมเดลการถดถอยโลจิสติกมีข้อได้เปรียบเหนือกว่าการถดถอยเชิงเส้น คือ ตัวแปรตามไม่เป็นตัวแปรต่อเนื่อง และตัวแปรอิสระกับตัวแปรตามไม่จำเป็นต้องมีความสัมพันธ์ เชิงเส้น โมเดลการถดถอยโลจิสติกสามารถวิเคราะห์อิทธิพลของการเป็นสมาชิกของกลุ่มบนตัวแปรตาม โดยที่ความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามจะเป็นเส้นตรงหรือไม่เป็นเส้นตรงก็ได้ ดังนั้น โมเดลการถดถอยโลจิสติกเหมาะสมที่จะนำมาประยุกต์ใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ไม่ว่าจะเป็นข้อมูลแบบสองค่าหรือหลายค่า ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม โดยใช้โมเดลการถดถอยโลจิสติกแบบจัดอันดับตามกรอบแนวคิดของ Zumbo (1999) จะประมาณค่าความน่าจะเป็นสะสม (Cumulative probability) ในรายการคำตอบที่ 1 ถึง *J-1* เมื่อ *J* เป็นจำนวนรายการคำตอบในมาตรแบบจัดอันดับ โดยที่ความน่าจะเป็นสะสมในรายการคำตอบสุดท้ายมีค่าเท่ากับ 1 ตัวอย่าง เช่น ข้อสอบมีรายการคำตอบ 3 รายการ โมเดลการถดถอยจะประมาณค่าความน่าจะเป็นสะสมในรายการที่ 1 ถึง 2 คือ ความน่าจะเป็นสะสมสำหรับการตอบในรายการที่ 1 หรือต่ำกว่า และความน่าจะเป็นสะสมสำหรับการตอบในรายการที่ 2 หรือต่ำกว่า โดยที่และความน่าจะเป็นสะสมในรายการที่ 3 ซึ่งเป็นรายการ

สุดท้ายมีค่าเท่ากับ 1 ดังนั้นจะคำนวณความน่าจะเป็นสะสมจำนวน *J-1* รายการ สำหรับโมเดลการถดถอยโลจิสติกแบบจัดอันดับ สามารถเขียนในรูปสัญลักษณ์ ดังนี้

$$P(Y \le j) = \frac{e^{\alpha_j}}{1+e^{\alpha_j}} \tag{91}$$

เมื่อ $P(Y \le j)$ แทนความน่าจะเป็นสะสมสำหรับการตอบในรายการที่ j หรือต่ำกว่าและ α_j เป็นพารามิเตอร์ส่วนตัดของความน่าจะเป็นสะสมในแต่ละรายการ โมเดลในสมการที่ 91 สามารถเขียนในรูปของล็อกได้ ดังนี้

$$log\left[\frac{P(Y \le j)}{P(Y > j)}\right] = \alpha_j \tag{92}$$

หรือเขียนในรูปโลจิต (logit) ดังนี้

$$logit[(Y \le j)] = \alpha_j \tag{93}$$

โมเดลในสมการที่ 93 อยู่ในรูปโลจิต เมื่อโลจิตเป็นล็อกธรรมชาติของอัตราส่วนแต้มต่อ (Odds ratio) สมการดังกล่าวมีเฉพาะพารามิเตอร์ส่วนตัดเท่านั้น เมื่อนำตัวแปรทำนายเพิ่มเข้าไปในโมเดล ได้แก่ คะแนนรวม (Total score) ซึ่งเป็นระดับความสามารถ โดยสมมติว่าเป็นตัวแปรต่อเนื่อง ตัวแปรกลุ่มผู้สอบ (Group variable) และตัวแปรปฏิสัมพันธ์ระหว่างคะแนนรวมกับกลุ่มผู้สอบ (Interaction variable) จะได้โมเดลใหม่ ดังนี้

$$logit[(Y \le j)] = \alpha_j + b_1X + b_2G + b_3(X*G) \tag{94}$$

เมื่อ X แทน ระดับความสามารถของผู้สอบ

G แทน กลุ่มผู้สอบ

b_1 แทน พารามิเตอร์ที่เกี่ยวข้องกับระดับความสามารถของผู้สอบ (X)

b_2 แทน พารามิเตอร์ที่เกี่ยวข้องกับกลุ่มผู้สอบ (G)

b_3 แทน พารามิเตอร์ที่เกี่ยวข้องกับปฏิสัมพันธ์ระหว่างระดับความสามารถของผู้สอบกับกลุ่มผู้สอบ ($X*G$)

กระบวนการตรวจสอบ

ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม โดยใช้วิธีการถดถอยโลจิสติกแบบจัดอันดับของ Zumbo (1999) จะใช้กระบวนการตรวจสอบความเหมาะสมของข้อมูลกับโมเดล โดยจะทดสอบโมเดล 3 ขั้นตอน ดังนี้

ขั้นตอนที่หนึ่ง ทดสอบตัวแปรที่ใช้ในการจับคู่ความสามารถ (Matching variable) หรือตัวแปรเงื่อนไข (Conditioning variable) ในที่นี้คือ คะแนนรวม ซึ่งเป็นระดับความสามารถของผู้สอบ (X) ในขั้นตอนนี้ถือว่าอิทธิพลของตัวแปรคะแนนรวมมีค่าเท่ากันในทุกรายการตอบ และสมมติให้ค่าประมาณของพารามิเตอร์ความชันมีเพียงค่าเดียว การนำตัวแปรดังกล่าวเข้าไปในสมการถดถอยมีลักษณะ ดังนี้

$$logit\left[\left(Y \le j\right)\right] = \alpha_j + b_1 X \tag{95}$$

ขั้นตอนที่สอง ทดสอบตัวแปรสมาชิกของกลุ่มผู้สอบ (G) ในที่นี้จะกำหนดรหัสเป็นตัวแปรดัมมี่ (Dummy variable) โดยให้กลุ่มอ้างอิงแทนด้วย 0 และกลุ่มสนใจแทนด้วย 1 แล้วเพิ่มตัวแปรดังกล่าวในสมการที่ 95 ของขั้นตอนที่ 1 ดังนี้

$$logit\left[\left(Y \le j\right)\right] = \alpha_j + b_1 X + b_2 G \tag{96}$$

ขั้นตอนที่สาม ทดสอบตัวแปรสมาชิกของกลุ่มผู้สอบ (G) และปฏิสัมพันธ์ระหว่างระดับความสามารถกับสมาชิกของกลุ่มผู้สอบ ($X*G$) โดยเพิ่มพจน์ทั้งสองดังกล่าวในโมเดลสมการที่ 95 ของขั้นตอนที่ 1 ดังนี้

$$logit\left[\left(Y \le j\right)\right] = \alpha_j + b_1 X + b_2 G + b_3\left(X * G\right) \tag{97}$$

การทดสอบสมมติฐาน

การทดสอบการทำหน้าที่ต่างกันของข้อคำถามที่เป็นรูปแบบเดียวกัน และที่ไม่เป็นรูปแบบเดียวกัน สามารถทดสอบได้พร้อมกัน โดยกำหนดสมมติฐานศูนย์ (H_0) และสมมติฐานทางเลือก (H_1) ดังนี้

$$\begin{aligned} &H_0 : b_2 = 0 \text{ และ } b_3 = 0 \\ &H_0 : b_2 \neq 0 \text{ หรือ } b_3 = 0 \end{aligned} \tag{98}$$

การทดสอบนัยสำคัญทางสถิติของการทำหน้าที่ต่างกันของข้อคำถาม ตามกระบวนทดสอบความเหมาะสมของข้อมูลกับโมเดล 3 ขั้นตอน โดยการนำตัวแปรเข้าทดสอบกับโมเดลตามลำดับขั้น ซึ่งได้กล่าวมาแล้วนั้น ในการตัดสินว่าข้อสอบข้อใดทำหน้าที่เบี่ยงเบนที่เป็นรูปแบบเดียวกันหรือที่ไม่เป็นรูปแบบเดียวกัน Zumbo (1999) ได้เสนอแนะให้เปรียบเทียบความแตกต่างระหว่างค่าไค-กำลังสองที่คำนวณได้ในขั้นตอนที่ 1 และ 3 ที่ระดับชั้นของความเป็นอิสระเท่ากับ 2 ($df = 2$) โดยมีค่า $p \geq .05$ สำหรับค่าของ *df* ดังกล่าวได้มาจากการเปรียบเทียบความแตกต่างระหว่าง *df* ของขั้นตอนทั้งสอง ในขั้นตอนที่ 1 มี $df = 1$ และในขั้นตอนที่ 3 มี $df = 3$ ดังนั้น จะทดสอบค่าไค-กำลังสองที่ระดับ $df = 2$ ซึ่งเป็นการทดสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสองรูปแบบพร้อมกัน (Swaminathan & Rogers, 1990) ถ้าผลการทดสอบมีนัยสำคัญทางสถิติโดยปฏิเสธสมมติฐานศูนย์ ($b_3 = 0$) นั่นแสดงว่า ข้อสอบทำหน้าที่เบี่ยงเบน ซึ่งเป็นผลมาจากอิทธิพลของการเป็นสมาชิกของกลุ่มผู้สอบ/ อิทธิพลของปฏิสัมพันธ์ระหว่างการเป็นสมาชิกของกลุ่มกับระดับความสามารถของผู้สอบ ส่วนการตัดสินว่าข้อสอบทำหน้าที่เบี่ยงเบนที่เป็นรูปแบบเดียวกันหรือที่ไม่เป็นรูปแบบเดียวกัน สามารถพิจารณาได้จากความแตกต่างของค่า R^2 ที่วิเคราะห์ได้ระหว่างช่วงแรก (ขั้นตอนที่ 1 กับ 2) และช่วงหลัง (ขั้นตอนที่ 2 กับ 3) ถ้าความแตกต่างในช่วงแรกมากกว่าช่วงหลัง สามารถสรุปได้ว่า ข้อสอบทำหน้าที่เบี่ยงเบนที่เป็นรูปแบบเดียวกัน แต่ถ้าความแตกต่างในช่วงแรกน้อยกว่าช่วงหลัง สรุปได้ว่า ข้อสอบทำหน้าที่เบี่ยงเบนที่ไม่เป็นรูปแบบเดียวกัน (Zumbo, 1999)

6. การทดสอบวอลด์ (Wald test)

การทดสอบวอลด์ (Lord, 1977; 1980) เป็นเทคนิคหนึ่งที่มีการนำไปใช้สำหรับการวิเคราะห์การทำหน้าที่ต่างกันของข้อคำถาม บนพื้นฐานแนวคิดของการวิเคราะห์การตอบสนองข้อคำถามแบบเอกมิติ ถือว่าเป็นวิธีการวิเคราะห์การตอบสนองข้อคำถามแบบเอกมิติที่ใช้ในการเปรียบเทียบการประมาณค่าพารามิเตอร์ของข้อคำถามระหว่างกลุ่ม เริ่มต้น วิธีการนี้ Kim, Cohen, & Park (1995) ได้นำมาใช้ในโมเดลพหุมิติสองพารามิเตอร์ (M2PL) สูตรในการคำนวณ คือ

$$\chi^2 = \hat{\nu}\Sigma^{-1}\hat{\nu} \tag{99}$$

เมื่อ $\hat{\nu} = \left[\hat{a}_{f1} - \hat{a}_{r1}, \hat{a}_{f2} - \hat{a}_{r2}, \hat{d}_f - \hat{d}_r\right]$ เป็นเวกเตอร์ผลต่างระหว่างพารามิเตอร์ข้อคำถามทุกข้อ ระหว่างกลุ่มอ้างอิงและกลุ่มสนใจ และ Σ^{-1} เป็นเมทริกซ์ความคลาดเคลื่อนของความแปรปรวน-ความแปรปรวนร่วม

$$\Sigma^{-1} = \left(\Sigma_F + \Sigma_R\right)^{-1} \tag{100}$$

เมื่อ Σ_F แทน ความคลาดเคลื่อนของความแปรปรวน-ความแปรปรวนร่วม ของกลุ่มสนใจ

Σ_R แทน ความคลาดเคลื่อนของความแปรปรวน-ความแปรปรวนร่วม ของกลุ่มอ้างอิง

ยกตัวอย่างเช่น เมทริกซ์ความแปรปรวน-ความแปรปรวนร่วมของโมเดลพหุมิติสองพารามิเตอร์ แสดงดังสมการ

$$\Sigma_F = \begin{pmatrix} \text{var}\left(a_{f1}\right) & \text{cov}\left(a_{f1}, a_{f2}\right) & \text{cov}\left(a_{f1}, d_f\right) \\ \text{cov}\left(a_{f1}, a_{f2}\right) & \text{var}\left(a_{f2}\right) & \text{cov}\left(a_{f2}, d_f\right) \\ \text{cov}\left(a_{f1}, d_f\right) & \text{cov}\left(a_{f2}, d_f\right) & \text{var}\left(d_f\right) \end{pmatrix}_{3x3}$$

สถิติทดสอบ χ^2 มีลักษณะการแจกแจงแบบ χ^2 ที่ df = จำนวนพารามิเตอร์ของข้อคำถาม โดยมีสมมติฐานหลักสำหรับการทดสอบการทำหน้าที่ต่างกันของโมเดลพหุมิติแบบสองพารามิเตอร์

$$H_0 : a_{f1} = a_{r1}; a_{f2} = a_{r2}; d_f = d_r \tag{101}$$

การทดสอบลอร์ดวอลด์ สำหรับการทดสอบโมเดลพหุมิติ สามพารามิเตอร์ ได้แก่ พารามิเตอร์อำนาจจำแนก ความยาก และโอกาสการเดา โดยมีสมการ ดังนี้

$$\chi^2 = \hat{\nu}\Sigma^{-1}\hat{\nu} \tag{102}$$

เมื่อ $\hat{\nu} = \left[\hat{a}_{f1} - \hat{a}_{r1}, \hat{a}_{f2} - \hat{a}_{r2}, \hat{d}_f - \hat{d}_r, \hat{g}_f - \hat{g}_r\right]$ เป็นเวกเตอร์ผลต่างระหว่างพารามิเตอร์ ข้อคำถามทุกข้อ ระหว่างกลุ่มอ้างอิงและกลุ่มสนใจ และ Σ^{-1} เป็นเมทริกซ์ความคลาดเคลื่อน ความแปรปรวน-ความแปรปรวนร่วม

$$\Sigma_F = \begin{pmatrix} \text{var}(a_{f1}) & \text{cov}(a_{f1}, a_{f2}) & \text{cov}(a_{f1}, d_f) & \text{cov}(a_{f1}, g_f) \\ \text{cov}(a_{f1}, a_{f2}) & \text{var}(a_{f2}) & \text{cov}(a_{f2}, d_f) & \text{cov}(a_{f2}, g_f) \\ \text{cov}(a_{f1}, d_f) & \text{cov}(a_{f2}, d_f) & \text{var}(d_f) & \text{cov}(d_f, g_f) \\ \text{cov}(a_{f1}, g_f) & \text{cov}(a_{f2}, g_f) & \text{cov}(d_f, g_f) & \text{var}(g_f) \end{pmatrix}_{4x4}$$

สถิติทดสอบ χ^2 มีลักษณะการแจกแจงแบบ χ^2 ที่ df = จำนวนพารามิเตอร์ของข้อคำถาม โดยมีสมมติฐานหลักสำหรับการทดสอบการทำหน้าที่ต่างกันของโมเดลพหุมิติแบบสองพารามิเตอร์

$$H_0 : a_{f1} = a_{r1}; a_{f2} = a_{r2}; d_f = d_r; g_f = g_r \tag{103}$$

วิธีการประมาณค่าของ Wald แบบดั้งเดิมจะใช้วิธีการประมาณค่าพารามิเตอร์แบบ Joint maximum likelihood (JML) และปัจจุบันใช้การประมาณค่าพารามิเตอร์ด้วยวิธี Expectation Maximization (EM) และ Maginal Maximium likelihood (MML)

ต่อมาได้มีนักวิจัยพัฒนาการทดสอบ Wald (Cai, 2012; Cai et al., 2011; Langer, 2008) จาก Lord (1980) เพราะประสิทธิภาพสำหรับการทดสอบการทำหน้าที่ต่างกันของข้อคำถามมีค่าความคลาดเคลื่อนประเภทที่ 1 ขนาดใหญ่ (Donoghue & Isham, 1998; Kim et al., 1994; Lim & Drasgow, 1990; McLaughlin & Drasgow, 1987) นักวิจัยบางคนก็พบว่า การประมาณค่าเมทริกซ์ความแปรปรวนร่วมมีปัญหา (Donoghue & Isham,1998; Kim et al., 1994; McLaughlin & Drasgow, 1987) ดังนั้น วิธีการทดสอบ Wald จึงประมาณค่าเมทริกซ์ความแปรปรวนร่วมได้ถูกต้องมากขึ้น และค่าความสามารถจะถูกประมาณค่าพร้อมกันกับการประมาณค่าพารามิเตอร์ข้อคำถามระหว่างกลุ่มซึ่งเป็นวิธีที่ดีกว่าแบบเดิม ซึ่งมีการควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ (Langer, 2008) ภายใต้การจำลองข้อมูล

Langer (2008) ได้เสนอขั้นตอนการปรับเทียบแบบสองขั้นโดยเรียกว่า Wald-2 และ Cai et al. (2011) ได้นำเสนอขั้นตอนการปรับเทียบแบบขั้นตอนเดียว โดยเรียกว่า Wald-1 โดยทั้งสองวิธีเป็นการเชื่อมโยงเมทริกซ์ระหว่างกลุ่มด้วยการประมาณค่าพารามิเตอร์และการทดสอบการทำหน้าที่ต่างกันของข้อคำถาม โดยทั้งสองวิธีสามารถวิเคราะห์ได้ในโปแกรม IRTPRO และ flexMIRT และกำหนดการประมาณค่าเมทริกซ์ความแปรปรวนร่วม (Covariance matrix) ด้วยวิธี Supplemented expectation maximization (SEM)

6.1 การวิเคราะห์ด้วย Wald-2

วิธีการวิเคราะห์ด้วย Wald 2 (Langer, 2008) เป็นการวิเคราะห์ที่ไม่ต้องมีข้อคำถามร่วม (Anchors) โดยเริ่มวิเคราะห์จาก ขั้นตอนที่ 1 ทำการประมาณค่าพารามิเตอร์ความสามารถของกลุ่มสนใจ จากนั้นกำหนดให้ค่าพารามิเตอร์ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของพารามิเตอร์ความสามารถของกลุ่มสนใจให้มีค่าเป็น 0 และ 1 ตามลำดับ โมเดลที่มีความกลมกลืนแล้ว และทำการกำหนดค่าพารามิเตอร์ของข้อคำถามของทั้งสองกลุ่มให้มีค่าเท่ากัน ขั้นตอนที่ 2 กำหนดค่าให้กับการแจกแจงของความสามารถของกลุ่มสนใจด้วยค่าที่เกิดจากขั้นตอนที่ 1 และทำการประมาณค่าพารามิเตอร์ของทั้งสองกลุ่ม จากนั้นใช้ค่าพารามิเตอร์ของข้อคำถามจากการประมาณได้ในขั้นตอนที่ 2 ในการคำนวณค่า Chi-square (สมการที่ 55)

ข้อดีของการวิเคราะห์การทำหน้าที่ต่างกันของข้อคำถามแบบ Wald-II คือ ไม่ต้องมีการข้อคำถามร่วม (Anchors) สำหรับการวิเคราะห์ ส่วนข้อเสีย คือ การประมาณค่าพารามิเตอร์ความสามารถที่เกิดขึ้นในขั้นตอนแรกเป็นการประมาณภายใต้ข้อตกลงที่ว่า ทุกข้อจะต้องไม่มีการทำหน้าที่ต่างกันของข้อคำถาม และถ้าข้อตกลงนี้ไม่เป็นจริง การประมาณค่าพารามิเตอร์ความสามารถในกลุ่มสนใจจะมีความลำเอียง ซึ่งนำไปสู่การตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามไม่ถูกต้อง (Tay, Meade, & Cao, 2015) ซึ่งสอดคล้องกับการศึกษาภายใต้การจำลองข้อมูลและทดสอบด้วย Wald-2 พบว่า มีความคลาดเคลื่อนประเภทที่ 1 สูง ในทุกเงื่อนไขของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม (Woods et al., 2013)

6.2 การวิเคราะห์ด้วย Wald-I

Cai et al. (2011) ได้นำเสนอการทดสอบแบบ Wald-I มีวิธีการเช่นเดียวกันกับการทดสอบการทำหน้าที่ต่างกันของข้อคำถามทั่วไป คือ ต้องมีการกำหนดข้อคำถามร่วม ซึ่งข้อคำถามร่วมควรกำหนดเบื้องต้นก่อนการวิจัยหรือก่อนการทดสอบ วิธีการเลือกข้อคำถามร่วมเป็นวิธีการที่กล่าวถึงมากที่สุด เพราะทำให้เกิดความผิดพลาดประเภทที่ 1 ที่เหมาะสมสำหรับการตรวจสอบการทำหน้าที่ต่างกัน (Studies items) (Kim & Cohen, 1991; Wang, 2004; Woods, 2009)

Wald-1 มีขั้นตอนการวิเคราะห์เช่นเดียวกับการวิเคราะห์ด้วย Wald-2 ในขั้นตอนที่ 1 โดยมีข้อสมมติว่าทราบข้อคำถามร่วม และพารามิเตอร์ของข้อคำถามมีเสกลเดียวกันในเมทริกซ์เดียวกัน โดยที่พารามิเตอร์ของข้อคำถามร่วมมีค่าเท่ากันระหว่างกลุ่มและข้อคำถามข้ออื่น ๆ มีค่าอิสระระหว่างกลุ่ม และการทดสอบการทำหน้าที่ต่างกันของข้อคำถามทุกข้อทดสอบด้วยสมการที่ 55 อย่างไรก็ตาม การวิเคราะห์ด้วย Wald-1 มีประสิทธิภาพที่ดีกว่าเนื่องจากทราบข้อคำถามร่วมมาก่อน (Anchor) สำหรับการใช้โปรแกรมการทดสอบการทำหน้าที่ต่างกันของวิธี

Wald ถ้าเป็นโปรแกรม IRTPRO และ flexMIRT จะเรียกใช้คำสั่ง “test candidate item, estimate group difference with anchor items” ซึ่ง Candidate item คืออีกชื่อหนึ่งสำหรับข้อคำถามศึกษา (Studied items)

สำหรับงานวิจัยนี้ ผู้วิจัยทำการทดสอบการทำหน้าที่ต่างกันของข้อคำถามด้วยวิธี Wald I โดยการวิเคราะห์ครั้งนี้ใช้โปรแกรม IRTPRO

ฟังก์ชันการทำหน้าที่ต่างกันของข้อคำถามและแบบทดสอบ (Differential functional functioning of items and test: DFIT)

กระบวนการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามและแบบทดสอบ (DFIT) เริ่มจากการประมาณค่าฟังก์ชันการทำหน้าที่ต่างกันของแบบทดสอบ (DTF) และประมาณระดับของการทำหน้าที่ต่างกันของข้อคำถามในแต่ละข้อคำถาม (DIF) จากความแปรปรวนร่วมระหว่างฟังก์ชันการทำหน้าที่ต่างกันของข้อคำถามและแบบทดสอบ ในลำดับแรก ผู้วิจัยขอนำเสนอวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามและแบบทดสอบสำหรับโมเดลเอกมิติให้คะแนนสองค่า ภายใต้โมเดลแบบสองพารามิเตอร์ ซึ่งมีฟังก์ชันดังต่อไปนี้

$$P(\theta_j)=\frac{e^{D\alpha_i(\theta_j-\beta_i)}}{1+e^{D\alpha_i(\theta_j-\beta_i)}} \tag{104}$$

เมื่อ $P_i(\theta_j)$ แทน ความน่าจะเป็นที่ ผู้ตอบคนที่ j ที่มีความสามารถ θ ในการตอบข้อคำถามข้อที่ i ถูกต้องของผู้ตอบคนที่ j

α_i แทน พารามิเตอร์อำนาจจำแนกของข้อคำถามข้อที่ i

β_i แทน พารามิเตอร์ความยากของข้อคำถามข้อที่ i

กระบวนการของ DFIT ต้องการค่าประมาณพารามิเตอร์ของกลุ่มสนใจและกลุ่มอ้างอิง เมื่อทำแบบทดสอบแล้ว $P_{iR}(\theta)$ คือความน่าจะเป็นของผู้สอบในกลุ่มอ้างอิงที่ตอบข้อคำถามถูกด้วยความสามารถ θ และ $P_{iF}(\theta)$ คือความน่าจะเป็นของผู้สอบกลุ่มสนใจที่ตอบข้อคำถามถูกด้วยความสามารถ θ (Raju et al., 1995) โดยค่าคาดหมายของการตอบข้อคำถามได้ถูกต้องของผู้สอบใด (Examinee’s expected proportion correct: EPC) คำนวณภายใต้ทฤษฎีการตอบสนองข้อคำถาม ดังสมการ

$$T_S=\sum_{i=1}^{P}P_i(\theta) \tag{105}$$

เมื่อ P แทน จำนวนข้อคำถามทั้งหมดในแบบทดสอบ ซึ่งคำนวณได้จากสมการที่ 104 ภายใต้ DFIT เกิดจากการประมาณค่าพารามิเตอร์ของข้อคำถามแต่ละข้อสำหรับผู้ตอบแต่ละคน ณ ความสามารถ θ

ผู้สอบแต่ละคนมีค่ามีค่า EPC สองค่า คือ T_{SF} และ T_{SR} ถ้า $T_{SF} = T_{SR}$ แล้ว ค่า EPC ของผู้ตอบจะมีความเป็นอิสระจากการเป็นสมาชิกในกลุ่ม ค่าความต่างระหว่าง T_{SF} และ T_{SR} ขนาดใหญ่ จะทำให้เกิดการทำหน้าที่ต่างกันของแบบทดสอบขนาดใหญ่ ซึ่งค่าของ DTF ของผู้สอบในแต่ละระดับความสามารถสามารถคำนวณได้จาก

$$D^2 = \left(T_{SF} - T_{SR}\right)^2 \tag{106}$$

นอกจากนี้ยังสามารถคำนวณ DTF จากค่าคาดหมาย (Expectation: E) ของกลุ่มใด (FG หรือ RG) เช่น กลุ่มสนใจ (FG) ได้จาก

$$DTF = E_F\left(D^2\right) = E_F\left(T_{SF} - T_{SR}\right)^2 \tag{107}$$

Raju et al. (1995) แสดงการคำนวณค่า DTF ด้วยการอินทิเกรตด้วยฟังก์ชันหนาแน่นน่าจะเป็นของ $\theta[f_F(\theta)]$ ทั้งหมดในกลุ่มสนใจ

$$DTF = \int D^2 f_F(\theta) d\theta \tag{108}$$

หรือ

$$DTF = \sigma_D^2 + \left(\mu_{TF} - \mu_{TR}\right)^2 = \sigma_D^2 + \mu_D^2 \tag{109}$$

เมื่อ μ_{TF} แทน ค่าเฉลี่ยของคะแนนคาดหมายของผู้ตอบกลุ่ม FG

μ_{TR} แทน ค่าเฉลี่ยของคะแนนคาดหมายของผู้ตอบกลุ่ม RG

σ_D^2 แทน ความแปรปรวนของ D

โดยที่ $DIF\left(d_{is}\right) = P_{iF}\left(\theta_S\right) - P_{iR}\left(\theta_S\right)$ แล้วจะได้ว่า

$$DTF = E\left(\sum_{i=1}^{P} d_i\right)^2 \tag{110}$$

และ *DTF* ยังสามารถคำนวณได้จากความแปรปรวนร่วมของผลต่างคะแนนคาดหวังของข้อคำถามกับความแตกต่างระหว่างคะแนนคาดหมายของแบบทดสอบ คือ

$$DTF = \sum_{i=1}^{P}\left[Cov(d_i, D) + \mu_{di}\mu_D\right] \tag{111}$$

Flowers et al. (1999); Oshima et al. (1997); Raju et al. (1995) ได้นำสมการที่ 111 มาใช้สำหรับโมเดลแบบชดเชยในการคำนวณ DIF (Compensatory DIF: CDIF) โดยมีสมการ ดังนี้

$$CDIF_i = Cov(d_i, D) + \mu_{di}\mu_D \tag{112}$$

เมื่อ $Cov(d_i, D) = \sigma_{di}^2 + Cov(d_i, d_j), i \neq j$

Raju et al. (1995) ได้นำเสนอโมเดลแบบไม่ชดเชยสำหรับการคำนวณ DIF (Non-compensatory DIF: NCDIF) เมื่อ DIF คำนวณได้จากข้อคำถามข้อที่ *i* และข้ออื่น ๆ ไม่มี DIF จะได้ว่า

$$NCDIF_i = \sigma_{di}^2 + \mu_{di}^2 \tag{113}$$

การทดสอบนัยสำคัญของ DFIT

ถ้า *D* เป็นค่าความแตกต่างระหว่างคะแนนคาดหมายที่มีการแจกแจงแบบปกติด้วยค่าเฉลี่ย μ_D และส่วนเบี่ยงเบนมาตรฐาน σ_D^2 แล้ว *Z* สามารถคำนวณได้จาก

$$Z = \frac{D_S - \mu_D}{\sigma_D} \tag{114}$$

เมื่อ $Z_S{}^2$ มีการแจกแจงแบบ χ^2 ด้วย องศาอิสระเท่ากับ 1 และผลรวมของ $Z_S{}^2$ ของผู้สอบทุกคน (*N*) มีการแจงแจงแบบ χ^2 ด้วย องศาอิสระเท่ากับ *N*

$$\chi^2_N = \sum_{S=1}^{N} Z_S^2 = \frac{\sum_{S=1}^{N} (D_S - \mu_D)^2}{\sigma_D^2} \quad (115)$$

ถ้าลดค่าคาดหมายของ DTF ให้ต่ำสุด ด้วยค่าเฉลี่ย μ_D^2 ที่มีค่าเท่ากับ 0 แล้ว สมการที่ 115 จะอยู่ในรูป

$$\chi^2_N = \frac{\sum_{S=1}^{N} D^2}{\sigma_D^2} = \frac{N(DTF)}{\sigma_D^2} \quad (116)$$

เมื่อ χ^2 ที่มีนัยสำคัญจะหมายถึงมีข้อคำถามหนึ่งข้อหรือมากกว่าทำหน้าที่ต่างกัน Raju et al. (1995) ให้ข้อเสนอแนะว่า ผู้วิจัยควรเริ่มจากการนำข้อคำถามแบบ CDIF ที่มีนัยสำคัญออก จนกว่าค่า χ^2 ไม่มีนัยสำคัญ

การวิเคราะห์อำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1

ในการวิเคราะห์อำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ถ้าเป็นการทดสอบในสถานการณ์ที่ใช้ข้อมูลจริง จะสามารถดำเนินการได้โดยกำหนดวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามที่มีประสิทธิภาพขึ้นมาหนึ่งวิธี โดยถือว่า ผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อคำถามด้วยวิธีดังกล่าวมีความถูกต้อง และจะใช้เป็นเกณฑ์ในการวิเคราะห์อำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม เช่น ถ้าผู้วิจัยต้องการวิเคราะห์อำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามด้วยวิธีซิปเทสท์ และวิธีการถดถอยโลจิสติก ถ้าให้วิธีการวัดพื้นที่ของราชู เป็นเกณฑ์ในการตรวจสอบ ดังนั้น ผู้วิจัยต้องนำผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อคำถามวิธีซิปเทสท์ และวิธีการถดถอยโลจิสติก มาเทียบกับวิธีการวัดพื้นที่ราชู โดยถือว่า ผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อคำถามด้วยวิธีการวัดพื้นที่ของราชูเป็นเกณฑ์ที่ถูกต้อง กล่าวคือ ถ้าวิธีซิปเทสท์และวิธีการถดถอยโลจิสติกระบุข้อคำถามที่ทำหน้าที่ต่างกันได้ตรงกับข้อคำถามที่ถูกระบุด้วยวิธีการวัดพื้นที่ของราชู แสดงว่า วิธีดังกล่าวสามารถตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามได้ถูกต้อง และถ้าวิธีซิปเทสท์และวิธีการถดถอยโลจิสติกระบุข้อคำถามที่ทำหน้าที่ต่างกันของข้อคำถามไม่ตรงกับข้อคำถามที่ระบุด้วยวิธีการวัดของราชู แสดงว่า วิธีดังกล่าวตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามไม่ถูกต้อง สำหรับ

การศึกษาในสถานการณ์ที่ใช้ข้อมูลจำลอง เราสามารถกำหนดว่าจะให้ข้อคำถามข้อใดบ้างทำหน้าที่ต่างกันของข้อคำถาม จึงทำให้ผู้วิจัยทราบล่วงหน้าว่าข้อคำถามข้อใดบ้างเป็นข้อคำถามที่ลำเอียง จึงไม่จำเป็นต้องนำวิธีตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามวิธีใดวิธีหนึ่งมาเป็นเกณฑ์ในการตรวจสอบ ฉะนั้น ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามวิธีใดวิธีหนึ่งมาเป็นเกณฑ์ในการตรวจสอบ ฉะนั้น ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามด้วยวิธีใดก็ตาม ถ้าวิธีดังกล่าวระบุข้อคำถามที่ทำหน้าที่ต่างกันของข้อคำถามที่ทราบว่าทำหน้าที่ต่างกันอยู่ก่อนแล้ว (True positive: TP) แสดงว่า วิธีดังกล่าวตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามได้ถูกต้อง เราสามารถคำนวณหาอำนาจการทดสอบ (Power of test) ของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ได้จากสูตร

$$P = \frac{n_1}{N_1} \tag{117}$$

เมื่อ P แทน อำนาจการทดสอบ

n_1 แทน จำนวนข้อคำถามที่ตรวจสอบได้ถูกต้องว่าทำหน้าที่ต่างกัน

N_1 แทน จำนวนข้อคำถามทำหน้าที่ต่างกันทั้งหมดในแบบทดสอบ

ถ้าระบุข้อคำถามที่ต่างกันของข้อคำถามที่ทราบว่าทำหน้าที่ต่างกันของข้อคำถามอยู่ก่อนแล้ว โดยระบุว่าข้อคำถามทำหน้าที่ต่างกัน ทั้งที่ความเป็นจริงข้อคำถามทำหน้าที่ต่างกัน (False positive: FP) แสดงว่า วิธีดังกล่าวตรวจสอบทำหน้าที่ต่างกันของข้อสบไม่ถูกต้อง ทำให้เกิดความคลาดเคลื่อนประเภทที่ 1 (Type I error) สามารถคำนวณหาอัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ได้จากสูตร

$$E = \frac{n_2}{N_2} \tag{118}$$

เมื่อ E แทน อัตราความคลาดเคลื่อนประเภทที่ 1

n_2 แทน จำนวนข้อคำถามที่ตรวจสอบได้ผิดพลาดว่าทำหน้าที่ต่างกัน ทั้งที่เป็นข้อคำถามที่ไม่ทำหน้าที่ต่างกัน

N_2 แทน จำนวนข้อคำถามทำหน้าที่ทำหน้าที่ไม่ต่างกันทั้งหมดในแบบทดสอบ

ตอนที่ 5 งานวิจัยที่เกี่ยวข้อง

งานวิจัยในประเทศ

อรินทร์ น่วมถนอม (2549) ศึกษาเปรียบเทียบวิธีโพลีซิปเทสต์ วิธีการถดถอยโลจิสติกแบบจัดอันดับ และวิธีการถดถอยโลจิสติกแบบจัดอันดับหลายมิติ ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามที่วัดความสามารถหลายมิติ และข้อคำถามเป็นการให้คะแนนแบบหลายค่า เป็นการจำลองข้อมูลโดยใช้โมเดลพาเชียลเครดิตทั่วไปแบบหลายมิติจำลองผลการตอบจากแบบทดสอบที่วัดความสามารถสองมิติ จำนวน 40 ข้อ โดยให้คะแนน 5 ระดับ เป็น 1, 2, 3, 4 และ 5 คะแนน ในทุกข้อคำถาม และได้ทำการเปรียบเทียบวิธีการตรวจสอบภายใต้ปัจจัย 4 ปัจจัย ได้แก่ รูปแบบข้อคำถามทำหน้าที่ต่างกัน 2 รูปแบบ คือ รูปแบบเดียวกัน และที่ไม่เป็นรูปเดียวกัน สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกันของข้อคำถาม 3 ขนาด คือ 10%, 20% และ 30% ความแตกต่างของการแจกแจงความสามารถ 3 ระดับ คือ แตกต่างกัน 0SD, 0.5SD และ1.0SD และขนาดของกลุ่มตัวอย่าง 4 ขนาด คือ 250 คน, 500 คน, 1,000 คน และ 2,000 คน รวมเงื่อนไขทั้งสิ้น 72 เงื่อนไข ทำการจำลองซ้ำ 50 ครั้ง ผลการวิจัยพบว่า วิธีการถดถอยโลจิสติกแบบจัดอันดับ และวิธีการถดถอยโลจิสติกแบบจัดอันดับหลายมิติ มีอำนาจการทดสอบใกล้เคียงกัน ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งที่เป็นรูปแบบเดียวกัน และไม่เป็นรูปแบบเดียวกัน ซึ่งทั้งสองวิธีมีอำนาจการทดสอบสูงกว่าวิธีโพลีซิปเทสท์ ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามที่ไม่เป็นรูปแบบเดียวกัน สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบ ไม่มีผลต่อวิธีโพลีซิปเทสท์และวิธีการถดถอยโลจิสติกแบบจัดอันดับหลายมิติ แต่มีผลต่อวิธีการถดถอยโลจิสติกแบบจัดอันดับ เมื่อมีความแตกต่างของการแจกแจงความสามารถผู้สอบเพิ่มขึ้น วิธีโพลีซิปเทสท์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้สูงกว่าวิธีอื่น ๆ เมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้นทำให้ทุกวิธีในการตรวจสอบมีอำนาจการทดสอบเพิ่มขึ้นด้วยในทุก ๆ เงื่อนไข

อิทธิฤทธิ์ พงษ์ปิยะรัตน์ (2551) ได้ศึกษาค่าพารามิเตอร์ข้อคำถาม พารามิเตอร์ความสามารถของผู้สอบ ตรวจสอบและเปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อคำถาม (DIF) โดยการประยุกต์ใช้โปรแกรมโมเดลเชิงเส้นตรงระดับลดหลั่น (HLM) และโปรแกรม BILOG-MG ผลการวิจัยพบว่า ผลการประมาณค่าพารามิเตอร์ข้อคำถามโดยโมเดล HGLM-2L และ HGLM-3L ด้วยสถิติ Empirical Bayesian มีความสัมพันธ์อย่างสัมบูรณ์กับผลการประมาณค่าด้วยโปรแกรม BILOG-MG ส่วนผลการประมาณค่าพารามิเตอร์ผู้สอบด้วยโมเดล HGLM-2L มีความสัมพันธ์อย่างสัมบูรณ์กับผลการประมาณค่าด้วยโปรแกรม BILOG-MG ส่วนโมเดล HGLM-3L มีระดับของค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0.793 ซึ่งน้อยกว่าค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างโมเดล HGLM-2L กับโปรแกรม BILOG-MG ให้ผล

การตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามจากโมเดล HGLM และจากโปรแกรม BILOG-MG ให้ผลการตรวจสอบที่เหมือนกันในทุกข้อ ทั้งนี้เป็นเพราะหลักการวิเคราะห์การทำหน้าที่ต่างกันของข้อคำถามจากสองวิธีมีความคล้ายคลึงกัน นั่นคือ การควบคุมให้ค่าพารามิเตอร์ความสามารถผู้สอบมีค่าคงที่ โดยโปรแกรมมีหลักการประมาณดังนี้ โปรแกรม HLM ค่าความยากรายข้อถือเป็นค่าคงที่ (Fixed effect) ซึ่งเป็นคุณสมบัติที่เหมาะสมตามหลักการสร้างข้อคำถามเพราะตรงตามคุณสมบัติความไม่ผันแปรไปตามกลุ่มผู้สอบ (Item invariance) ส่วนค่าพารามิเตอร์ความสามารถของผู้สอบที่ได้จากการประมาณค่าด้วยวิธีเบส์ ก็ถูกปรับให้มีค่าเฉลี่ยเท่ากับศูนย์ ส่วนเบี่ยงเบนมาตรฐานเท่ากับหนึ่ง ดังนั้น โปรแกรมจะทดสอบอิทธิพลของตัวแปรดัมมี่เพศที่สร้างขึ้นในสมการได้จากการทดสอบด้วยสถิติทดสอบที (*t*-test) ส่วนโปรแกรม BILOG-MG จะทำการคำนวณค่าพารามิเตอร์ความสามารถของผู้สอบทั้งสองกลุ่มร่วมกัน เพื่อเป็นการปรับฐานให้อยู่บนมาตรเดียวกัน ควบคุมค่าความคลาดเคลื่อนมาตรฐานและปรับค่าเฉลี่ยค่าพารามิเตอร์เทรสโฮล (Threshold) ของข้อคำถามกลุ่มอ้างอิงให้เท่ากับศูนย์ ควบคุมค่าพารามิเตอร์ความยากของข้อคำถามกลุ่มอ้างอิงให้เท่ากับศูนย์ ควบคุมค่าพารามิเตอร์การเดาให้เท่ากับศูนย์ และค่าพารามิเตอร์อำนาจจำแนกให้เท่ากันทั้งสองกลุ่ม จากนั้นจึงทำการวิเคราะห์ค่าพารามิเตอร์ความยากอีกครั้งโดยแยกวิเคราะห์ตามแต่ละกลุ่มเพื่อนำค่า -2lnL ratio ของแต่ละกลุ่มมาเปรียบเทียบกัน ด้วยหลักการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามที่ใช้หลักการคำนวณและวิธีการทางสถิติวิเคราะห์ที่คล้ายคลึงกัน จึงทำให้ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามเหมือนกัน

ชัยยศ ชาวระนอง (2553) ทำการเปรียบเทียบประสิทธิภาพของโมเดลการวิเคราะห์องค์ประกอบเชิงยืนยันอันดับหนึ่ง องค์ประกอบเชิงยืนยันอันดับสอง องค์ประกอบแฝงยืนยันอันดับหนึ่งแฝงภายใน และองค์ประกอบเชิงยืนยันอันดับสองแฝงภายใน ภายใต้เงื่อนไขกลุ่มตัวอย่างขนาด 100, 200, 400, 800, 1,200 1,600 และ 2,000 คน จำนวนข้อคำถามขนาด 3, 5, 8, 10 และ 15 ต่อองค์ประกอบ และขนาดตัวอย่างในการทดสอบการทำหน้าที่ต่างกันขนาด 1,600, 2,000, 2,400, 2,800, 3,200, 3,600 และ 4,000 เป็นนักเรียนระดับชั้นมัธยมศึกษาปีที่ 3 ในเขตบุรีรัมย์เขต 2 จากการทดสอบแห่งชาติ ตรวจสอบการทำหน้าที่ต่างกันด้วยโปรแกรม AMOS ผลการวิจัยพบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามโดยวิธีการวิเคราะห์องค์ประกอบเชิงยืนยันอันดับสองแฝงภายในจำแนกตามเพศ โดยใช้เทคนิค Multi-group พบว่า มีข้อคำถามทำหน้าที่ต่างกัน 5 ข้อ และเมื่อตัดข้อคำถามที่ทำหน้าที่ต่างกันออก พบว่า ค่าสัมประสิทธิ์ความเชื่อมั่นของข้อคำถามสูงขึ้น

มิ่ง เทพครเมือง (2554) ได้ตรวจสอบความเท่าเทียมกันของการวัดบนพื้นฐานทฤษฎีการทดสอบแบบคะแนนจริงสัมพันธ์และทฤษฎีการตอบสนองข้อคำถาม ข้อมูลที่ใช้ในการศึกษา

เป็นการจำลองข้อมูลรายการคำตอบในการตรวจให้คะแนนสองค่า และรายการคำตอบ เป็นการตรวจให้คะแนนหลายค่า ฉบับละ 25 ข้อ ภายใต้เงื่อนไขขนาดของการทำหน้าที่ต่างกันของข้อคำถาม 3 ขนาด คือ ไม่มีการทำหน้าที่ต่างกัน การทำหน้าที่ต่างกันขนาดเล็ก และการทำหน้าที่ต่างกันขนาดใหญ่ กลุ่มตัวอย่างแบ่งตามสัดส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบเป็น 2 สัดส่วน คือ 1 : 1 และ 1 : 2 ทำการจำลองข้อมูลซ้ำ 100 ครั้ง ผลการวิจัยพบว่า 1) อำนาจการทดสอบในการตรวจสอบความเท่าเทียมกันของการวัดกรณีรายการคำตอบของผู้สอบที่เป็นการตรวจให้คะแนนหลายค่า การทำหน้าที่ต่างกันขนาดใหญ่ที่กลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบเป็น 200 : 400 คน วิธีการตรวจสอบบนพื้นฐานทฤษฎีการทดสอบแบบคะแนนจริงสัมพันธ์ มีอำนาจการทดสอบในการตรวจสอบต่ำกว่าวิธีการตรวจสอบบนพื้นฐานทฤษฎีการตอบข้อคำถาม อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 2) อัตราความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบความเท่าเทียมกันของการวัดกรณีรายการคำตอบของผู้สอบที่เป็นการตรวจให้คะแนนหลายค่า ทั้งวิธีตรวจสอบความเท่าเทียมกันของการวัดบนพื้นฐานทฤษฎีการทดสอบแบบคะแนนจริงสัมพันธ์ และวิธีบนพื้นฐานทฤษฎีการตอบข้อคำถามมีความคลาดเคลื่อนประเภทที่ 1 ไม่แตกต่างกัน

สุชาติ สิริมีนนันท์ (2554) การเปรียบเทียบวิธีโพลีโตมัสซิบเทสต์ วิธีการวิเคราะห์ฟังก์ชันการจำแนกโลจิสติก และวิธีการถดถอยโลจิสติกแบบจัดอันดับ ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบที่มีการให้คะแนนแบบหลายค่า โดยใช้การจำลองข้อมูลภายใต้โมเดลพาร์เชียลเครดิตทั่วไปแบบมิติเดียว มีรายการคำตอบ 5 รายการ จำลองข้อมูลภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ รูปแบบของข้อคำถามทำหน้าที่ต่างกัน 2 รูปแบบ ความยาวของแบบทดสอบ 3 ขนาด สัดส่วนของข้อคำถามทำหน้าที่ต่างกัน 3 ขนาด และขนาดตัวอย่าง 3 ขนาด พบว่า 1) อำนาจการทดสอบในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบที่มีการให้คะแนนหลายค่าของวิธีการวิเคราะห์ฟังก์ชันการจำแนกโลจิสติกสูงกว่าวิธีโพลีโตมัสซิบเทสต์ และวิธีการถดถอยโลจิสติกแบบจัดอันดับในทุกเงื่อนไข 2) อัตราความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามวิธีการวิเคราะห์ฟังก์ชันการจำแนกโลจิสติกต่ำกว่าวิธีโพลีโตมัสซิบเทสต์และวิธีการถดถอยโลจิสติกจัดอันดับในทุกเงื่อนไข

ธเกียรติกมล ทองงอก (2554) เปรียบเทียบอัตราความถูกต้อง และอัตราความคลาดเคลื่อนประเภทที่ 1 ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ที่มีรูปแบบการให้คะแนนแบบทวิภาค โดยการจำลองข้อมูลและข้อมูลเชิงประจักษ์ในวิธีการถดถอยโลจิสติก ระหว่างการวัดขนาดอิทธิพลตามเกณฑ์ Jodoin and Gierl กับเกณฑ์ Zumbo and Thomas การศึกษาครั้งนี้จำลองภายใต้ทฤษฎีการตอบสนองข้อคำถามแบบสองพารามิเตอร์ จำลองผลการตอบภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย รวมทั้งหมด 24 เงื่อนไข คือ รูปแบบข้อคำถามที่ทำหน้าที่ต่างกัน (ทั้งฉบับ

คิดเป็นร้อยละ 10 และ 20) และความยาวของแบบทดสอบทั้งฉบับ ในทุกเงื่อนไขจำลองข้อมูลซ้ำ 25 ครั้ง วิเคราะห์ข้อมูลในแต่ละเงื่อนไขด้วยวิธีถอดถอยโลจิสติก ระหว่างการวัดขนาดอิทธิพลตามเกณฑ์ Jodoin and Gierl กับเกณฑ์ Zumbo and Thomas ผลการวิจัยสรุปได้ว่า วิธีถดถอยโลจิสติก โดยการวัดขนาดอิทธิพลตาม Jodoin and Gierl มีอัตราการถูกต้อง ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามสูงกว่าเกณฑ์ Zumbo and Thomas ภายใต้เงื่อนไขเกือบทุกเงื่อนไข ส่วนข้อคำถามที่ทำหน้าที่ต่างกันแบบเอกรูปมีอัตราความถูกต้องจากการวัดขนาดอิทธิพลทั้ง 2 เกณฑ์สูงกว่าแบบเอกรูป แบบทดสอบที่มีจำนวนข้อคำถามที่ทำหน้าที่ต่างกัน ทั้งฉบับคิดเป็นร้อยละ 10 และเมื่อขนาดอิทธิพลทั้ง 2 เกณฑ์สูงกว่าในแบบทดสอบที่มีจำนวนข้อคำถามที่ทำหน้าที่ต่างกันทั้งฉบับคิดเป็นร้อยละ 10 และเมื่อขนาดอิทธิพลของข้อคำถามที่ทำหน้าที่ต่างกันเพิ่มขึ้น มีผลทำให้อัตราความถูกต้องจากการวัดขนาดอิทธิพลทั้ง 2 เกณฑ์ เพิ่มขึ้นภายใต้เกือบทุกเงื่อนไข

สุพัฒนา หอมบุปผา (2556) ศึกษาเปรียบเทียบผลการประมาณค่าพารามิเตอร์ความยากของข้อคำถาม พารามิเตอร์ความสามารถของผู้สอบ จำแนกตามเพศ สถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน ระหว่างวิธี HGLM, MIMIC และวิธี BAYESIAN และเปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อคำถามสำหรับผู้สอบ จำแนกตามเพศ สถานที่ตั้งของโรงเรียนระหว่างวิธี HGLM, MIMIC และวิธี BAYESIAN และศึกษาลักษณะการทำหน้าที่ต่างกันของข้อคำถามที่เกิดจากการทำหน้าที่ต่างกันที่ได้จากการวิเคราะห์การทำหน้าที่ต่างกัน ด้วยวิธี HGLM MIMIC และวิธี BAYESIAN โดยใช้ข้อมูลเป็นคะแนนสอบวัดผลสัมฤทธิ์ทางการเรียนเพื่อประเมินคุณภาพการศึกษาระดับชาติ ของนักเรียนชั้นประถมศึกษาปีที่ 3 ในรายวิชาภาษาไทย คณิตศาสตร์ และวิทยาศาสตร์ ของสำนักทดสอบทางการศึกษา กระทรวงศึกษาธิการ ปีการศึกษา 2553 จำนวน 1,000 คน จำแนกเพศชายและหญิง ที่อยู่ในเขตกรุงเทพมหานครและปริมณฑล และนอกเขตกรุงเทพมหานครและปริมณฑล พบว่า ผลการวิเคราะห์ค่าพารามิเตอร์ความยากและพารามิเตอร์ความสามารถผู้สอบและผลการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในวิชาภาษาไทย คณิตศาสตร์ และวิทยาศาสตร์ ด้วยวิธี HGLM, MIMIC และวิธี BAYESIAN มีความสัมพันธ์กันสูงมาก ที่ระดับนัยสำคัญทางสถิติ .01 นอกจากนี้ยังพบว่า วิธี HGLM ตรวจสอบพบการทำหน้าที่ต่างกันของข้อคำถามมากที่สุด ส่วนตรวจพบน้อยที่สุดคือ วิธี MIMIC

อาวีพร ปานทอง (2558) ได้ศึกษาเพื่อเปรียบเทียบอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามที่ให้คะแนนหลายค่า โดยวิธีทดสอบอัตราส่วนความควรจะเป็น วิธีเบส์เซียน และวิธีโพลีโตมัสซิปเทสท์ ในการประมาณค่าพารามิเตอร์ ภายใต้ปัจจัยที่แตกต่างกัน 4 ปัจจัย ดังนี้ ความยาวของแบบทดสอบ ขนาดของการทำหน้าที่ต่างกันของข้อคำถาม ความแตกต่างของการแจกแจงความสามารถ และ

ขนาดตัวอย่าง โดยใช้ข้อมูลจำลองภายใต้โมเดลพาเชียลเครดิตทั่วไป ภายใต้การตอบสนองข้อคำถามแบบมิติเดียว โดยข้อคำถามแต่ละข้อมีตัวเลือกให้ตอบ จำนวน 5 ตัวเลือก โดยให้คะแนนเป็น 0,1, 2, 3 หรือ 4 คะแนน รวมทั้งหมด 54 เงื่อนไข วนซ้ำ 500 รอบ ผลการวิจัยพบว่า วิธีทดสอบอัตราส่วนความควรจะเป็น และวิธีเบส์เซียน มีอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ใกล้เคียงกัน ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามที่ให้คะแนนหลายค่าที่เป็นรูปแบบเดียวกัน ทั้งอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 สูงกว่าวิธีโพลีโตมัสซิปเทสท์ ภายใต้เงื่อนไขของปัจจัยที่แปรเปลี่ยน เมื่อความยาวของข้อคำถามเพิ่มขึ้น วิธีการทดสอบอัตราส่วนความควรจะเป็น และเบส์เซียน สามารถควบคุมคลาดคลาดเคลื่อนประเภทที่ 1 ได้ดีกว่าวิธีโพลีโตมัสซิปเทสท์ นอกจากนี้ เมื่อขนาดตัวอย่างเพิ่มขึ้น มีผลทำให้ทุกวิธีมีอำนาจการทดสอบเพิ่มขึ้นภายใต้ทุกปัจจัย

งานวิจัยต่างประเทศ

Flowers et al. (1999) ได้ปรับขยายวิธี DFIT (Differential item/test functioning) ที่ใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ในข้อคำถามที่ให้คะแนนแบบสองค่า ซึ่งพัฒนาโดย Raju et al. (1995) เพื่อนำมาใช้ตรวจสอบในข้อคำถามที่ให้คะแนนหลายค่า ข้อมูลที่ใช้ในการศึกษาเป็นข้อมูลจำลองภายใต้โมเดลเกรดเรสพอน (Graded Response Model: GRM) โดยนำค่าพารามิเตอร์จากการศึกษาของ Cohen and Kim (1993) และ Fleer (1993) มาปรับใหม่เพื่อให้เหมาะสมกับโมเดล GRM แล้วจำลองผลการตอบข้อคำถามที่มี 5 รายการ ข้อคำถามทุกข้อให้คะแนน 5 ระดับ (0, 1, 2, 3 และ 4) และขนาดตัวอย่างกลุ่มละ 1,000 คน ภายใต้เงื่อนไขที่แปรเปลี่ยน 4 ปัจจัย คือ 1) ความยาวของแบบทดสอบ 2 ขนาด คือ 20 ข้อ และ 40 ข้อ 2) ความแตกต่างของการแจกแจงความสามารถระหว่างกลุ่มอ้างอิงและกลุ่มสนใจ 2 ระดับ คือ แบบเท่ากันและแบบไม่เท่ากัน โดยแบบเท่ากันทั้ง 2 กลุ่มมีการแจกแจง $N(0, 1)$ ส่วนแบบไม่เท่ากันกลุ่มอ้างอิงมีการแจกแจง $N(0, 1)$แต่กลุ่มสนใจมีการแจกแจง $N(-1, 1)$(3) สัดส่วนของข้อคำถามทำหน้าที่ต่างกัน 4 ขนาด คือ 0% 5% 10% และ 20% 4) ทิศทางของข้อคำถามทำหน้าที่ต่างกัน 2 ลักษณะ คือ ข้อคำถามทำหน้าที่ต่างกันมีทิศทางเดียวกัน ซึ่งเข้าข้างกลุ่มอ้างอิงทุกข้อ และข้อคำถามทำหน้าที่ต่างกันมีสองทิศทาง ซึ่งเข้าข้างกลุ่มอ้างอิงและกลุ่มสนใจจำนวนเท่ากัน ยกเว้นเงื่อนไข 5% ของแบบทดสอบขนาด 10 ข้อ ไม่มีการจำลองข้อคำถามดังกล่าว และ 5) รูปแบบของข้อคำถามทำหน้าที่ต่างกัน 2 รูปแบบ คือ ข้อคำถามทำหน้าที่ต่างกัน และไม่เป็นรูปแบบเดียวกัน การจำลองรูปแบบของข้อคำถามทำหน้าที่ต่างกันในเงื่อนไขสัดส่วนข้อคำถามทำหน้าที่ต่างกัน 20% จำลองข้อคำถาม Uniform DIF และ Nonuniform DIF แต่เงื่อนไขอื่น ๆ จำลองเฉพาะ Uniform DIF เท่านั้น รวมข้อมูลที่จำลองทั้งหมด 26 เงื่อนไข ในแต่ละเงื่อนไขจำลองซ้ำ 5 ครั้ง การประมาณค่าพารามิเตอร์

ของข้อคำถามและผู้สอบใช้โปรแกรม PASCALE ภายใต้โมเดล GRM แล้วใช้โปรแกรม EQUATE ปรับเทียบมาตรประมาณพารามิเตอร์ของกลุ่มอ้างอิงให้อยู่บนเมทริกซ์ของกลุ่มสนใจ การคำนวณดัชนี DFIT ใช้โปรแกรม FORTRAN สำหรับดัชนี CDIF และ NCDIF จะคำนวณความถูกต้องเชิงบวก (True positive: TP) และความผิดพลาดเชิงบวก (False positive: FP) แล้วเปรียบเทียบภายใต้ 2 เงื่อนไข คือ เงื่อนไขจริง ซึ่งเป็นค่าพารามิเตอร์จริง และเงื่อนไขค่าประมาณ ซึ่งเป็นค่าพารามิเตอร์จากการประมาณ

ผลการศึกษาพบว่า วิธี DFIT มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม และแบบทดสอบในข้อคำถามที่ให้คะแนนหลายค่า ซึ่งเป็นข้อมูลจากการจำลอง สำหรับความยาวของแบบทดสอบความแตกต่างของการแจกแจงความสามารถ สัดส่วนของข้อคำถามทำหน้าที่ต่างกัน รูปแบบของข้อคำถามทำหน้าที่ต่างกันและทิศทางของข้อคำถามทำหน้าที่ต่างกันมีอิทธิพลต่อความถูกต้องเชิงบวก และความผิดพลาดเชิงบวกน้อยมาก ในทุกเงื่อนไขของการตรวจสอบ และเป็นไปตามที่คาดไว้เกี่ยวกับข้อคำถามที่มี DIF ไม่คงที่เหมือนกันกับดัชนี NCDF ในทุกเงื่อนไข ซึ่งเหมือนกับการตรวจสอบในกรณีมิติเดียว (Fleer, 1993) และในกรณีหลายมิติ สำหรับการศึกษาครั้งนี้ CDIF ที่ตรวจสอบใน 2 เงื่อนไขมีความผันแปรไปจากที่คาดไว้ คือ เงื่อนไขความยาวของแบบทดสอบ 20 ข้อ และการแจกแจงความสามารถแบบเท่ากัน มีความผิดพลาดในการตรวจสอบ 18% และในเงื่อนไขความยาวของแบบทดสอบ 50 ข้อ และการแจกแจงความสามารถแบบไม่เท่ากันสามารถตรวจสอบได้ถูกต้องเพียง 50%

Gierl et al. (2004) ได้ศึกษาผลของวิธีซิปเทสท์ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม เมื่อร้อยละของข้อคำถามทำหน้าที่ต่างกันมีขนาดใหญ่ ข้อมูลที่ใช้ในการศึกษาเป็นข้อมูลจำลอง โดยจำลองผลการตอบข้อคำถาม จำนวน 40 ข้อ ภายใต้การจัดกระทำ 4 ตัวแปร ได้แก่ ตัวแปร 1 สัดส่วนของข้อคำถามทำหน้าที่ต่างกัน 3 ขนาด คือ 20% 30% และ 40% (8, 16 และ 24 ข้อ ตามลำดับ) ในแต่ละระดับดังกล่าวแบ่งออกเป็น 2 เงื่อนไข คือ 75/ 25 และ 50/ 50 ในเงื่อนไขแรก 75% ของข้อคำถามมี DIF ขนาดปานกลาง และ 25% ของข้อคำถามมี DIF ขนาดใหญ่ ในเงื่อนไขสอง 50% ของข้อคำถามมี DIF ขนาดปานกลาง และ 50% ของข้อคำถามมี DIF ขนาดใหญ่ ซึ่ง DIF ขนาดปานกลางมีค่าอยู่ในช่วง $.05 \leq \left|\hat{\beta}_{uni}\right| < .01$ และ DIF ขนาดใหญ่มีค่าอยู่ในช่วง $\left|\hat{\beta}_{uni}\right| \geq .10$ ตัวแปร 2 ทิศทางของข้อคำถามทำหน้าที่ต่างกัน 2 ลักษณะ คือ แบบสมดุล และแบบไม่สมดุล โดยแบบสมดุลมีข้อคำถาม DIF เข้าข้างกลุ่มอ้างอิงและกลุ่มสนใจจำนวนเท่ากัน ส่วนแบบไม่สมดุลมีข้อคำถาม DIF เข้าข้างกลุ่มอ้างอิงและกลุ่มสนใจจำนวนไม่เท่ากัน กล่าวคือ ในเงื่อนไข DIF 20% 40% และ 60% เข้าข้างกลุ่มอ้างอิงจำนวน 6, 12 และ 18 ข้อ ตามลำดับ และเข้าข้างกลุ่มสนใจจำนวน 2, 4 และ 6 ตามลำดับ ตัวแปร 3 ขนาดตัวอย่าง 4 ขนาด คือ 500, 1,000 1,500 และ 2,000 คน

ต่อกลุ่มอ้างอิงและกลุ่มสนใจ และตัวแปร 4 ความแตกต่างของการแจกแจงความสามารถระหว่างกลุ่ม 2 ระดับ คือ แบบเท่ากันและแบบไม่เท่ากัน ในกรณีเท่ากันทั้งสองกลุ่มมีการแจกแจงความสามารถปกติ N(0, 1) ส่วนในกรณีไม่เท่ากันกลุ่มอ้างอิงและกลุ่มสนใจมีการแจกแจงความสามารถ N(0, 1) และ N(0, 1) ตามลำดับ รวมข้อมูล ที่จำลองทั้งหมด 96 เงื่อนไข ในแต่ละเงื่อนไขจำลองซ้ำ 100 ครั้ง โดยใช้โปรแกรม DIFSIM จำลองผลการตอบภายใต้ 2 เงื่อนไข คือ เงื่อนไข N0-DIF ใช้พารามิเตอร์ของข้อคำถามจากการศึกษาของดราสโกว์ จำนวน 32 ข้อ จำลองข้อคำถามวัดเฉพาะความสามารถในมิติหลัก และให้ค่าพารามิเตอร์ของกลุ่มอ้างอิงกับกลุ่มสนใจเท่ากัน โดยใช้โมเดลโลจิสติก 3 พารามิเตอร์แบบมิติเดียว (Unidimensional 3-parameter logistic) ส่วนในเงื่อนไข DIF ใช้พารามิเตอร์จากการศึกษาของ Nandakumar (1993) จำนวน 24 ข้อ จำลองข้อคำถามวัดทั้งความสามารถในมิติหลักและมิติรอง ภายใต้โมเดลโลจิสติก 3 พารามิเตอร์แบบหลายมิติที่มีการชดเชย (Compensatory multidimensional) และกำหนดให้ค่าสหสัมพันธ์ระหว่างมิติในแต่ละกลุ่มเท่ากับ .5 ส่วนความแปรปรวนภายในมิติของแต่ละกลุ่มเท่ากับ 1 แล้วคำนวณอัตราการตรวจสอบ DIF 2 ชนิด คือ อัตราการตัดสินใจไม่ถูกต้อง (Incorrectdecisions) และอัตราการตัดสินใจถูกต้อง (Correct decisions)

ผลการศึกษาพบว่า ในเงื่อนไข DIF 75/ 25 (Unbalanced DIF) เมื่อข้อคำถามทำหน้าที่ต่างกันจำนวนเล็กน้อย คือ 20% ประกอบด้วยข้อคำถามเข้าข้างกลุ่มอ้างอิง 6 ข้อ และเข้าข้างกลุ่มสนใจ 2 ข้อ ไม่มีผลต่ออัตราการตรวจสอบ เมื่อสัดส่วนของการทำหน้าที่ต่างกันของข้อคำถามขนาดใหญ่ คือ 40% และ 60% อัตราการตัดสินใจไม่ถูกต้องลดลง แต่อัตราการตัดสินใจถูกต้องเพิ่มขึ้น โดยเพิ่มขึ้นสูงสุดในเงื่อนไขของขนาดตัวอย่างมากที่สุด สำหรับเงื่อนไขสัดส่วน DIF 50/ 50 เมื่อสัดส่วนของการทำหน้าที่ต่างกันขนาดใหญ่ คือ 40% และ 60% อัตราการตัดสินใจไม่ถูกต้องลดลงในขณะที่ขนาดตัวอย่างเพิ่มขึ้น อย่างก็ตาม ในเงื่อนไข Balanced DIF แต่ในเงื่อนไข Balanced DIF และ Unbalanced DIF พบว่า อัตราการตัดสินใจไม่ถูกต้องในเงื่อนไข Unbalanced DIF

Meade and Lautenschlager (2004) ได้ทำการเปรียบเทียบทฤษฎีการตอบสนองข้อคำถามและการวิเคราะห์องค์ประกอบเชิงยืนยัน ในการแสดงความเท่าเทียมกันของการวัด โดยการจำลองข้อมูลจากข้อคำถามแบบ Likert 5 ระดับ กับขนาดกลุ่มตัวอย่าง 2 กลุ่ม ที่มีการแจกแจงปกติ 3 ขนาด คือ 150, 500, 1,000 คน แต่ละเงื่อนไขมีการทำซ้ำ 100 ครั้ง และใช้ค่าพารามิเตอร์ความยากและค่าพารามิเตอร์อำนาจจำแนกในการประเมินความเหมาะสม ความเหมือน และความแตกต่างระหว่างการวิเคราะห์องค์ประกอบเชิงยืนยัน และทฤษฎีการตอบข้อคำถาม พบว่า กรณีที่ค่าพารามิเตอร์ความยากในการตรวจสอบความเท่าเทียมกันของการวัด วิธีการวิเคราะห์องค์ประกอบเชิงยืนยันจะตรวจสอบได้ดีกับขนาดกลุ่มตัวอย่าง 500, 1,000 คน ส่วนขนาดกลุ่มตัวอย่าง 150 คน และเมื่อใช้

อัตราส่วนความเป็นไปได้ตรวจสอบความเท่าเทียมกันของการวัด พบว่า ขนาดกลุ่มตัวอย่าง 150 คน จะมีความถูกต้องในการตรวจสอบต่ำ การตรวจสอบความเท่าเทียมกันของการวัดด้วยวิธีการวิเคราะห์องค์ประกอบเชิงยืนยัน พบว่า ตรวจสอบได้ดีในทุกขนาดกลุ่มตัวอย่าง ส่วนการตรวจสอบด้วยวิธีอัตราส่วนความควรจะเป็น พบว่า ตรวจสอบได้ดีในกลุ่มตัวอย่าง 1,000 คน ส่วนขนาดกลุ่มตัวอย่าง 150 คน ให้ผลการตรวจสอบไม่คงที่ และกลุ่มตัวอย่างขนาด 500 คน ให้ผลการตรวจสอบไม่แตกต่างจากขนาดกลุ่มตัวอย่าง 150 คน

Finch and French (2007) ได้ตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ที่มีการชดเชยคำตอบใน 4 วิธี เป็นการศึกษารูปแบบการทำหน้าที่ต่างกันของข้อคำถามแบบอเนกรูป ได้แก่ วิธีสมการถดถอยโลจิสติก (LR) ซิปเทสท์ (SIBTEST) การทดสอบอัตราส่วนความควรจะเป็น และการวิเคราะห์องค์ประกอบเชิงยืนยัน (CFA) จากการจำลองข้อมูลภายใต้เงื่อนไขต่าง ๆ ได้แก่ ขนาดกลุ่มตัวอย่าง ใช้สัดส่วนของกลุ่มสนใจกับกลุ่มอ้างอิงแบบ 1 : 1 และ แบบ 1 : 2 จำนวน 3 ขนาด คือ 250 : 250, 250 : 500 และ 500 : 1,000 คน ระดับความสามารค่าเฉลี่ย 2 ระดับ คือ ระดับความสามารถเฉลี่ยเท่ากัน ($M = 0$, $SD = 1$) และระดับความสามารถเฉลี่ยกลุ่มอ้างอิงสูงกว่ากลุ่มสนใจ ($MR = 0.5$, $MF = 0$) ขนาดของการทำหน้าที่ต่างกันของข้อคำถามที่มีการชดเชยการตอบ 3 ขนาด คือ 0% 10% หรือ 20% กระทำภายใต้โมเดลโลจิสติก 2 และ 3 พารามิเตอร์ แต่ละเงื่อนไขทำซ้ำ 1,000 ครั้ง ผลการศึกษาพบว่า การตรวจสอบอัตราความคลาดเคลื่อนประเภทที่ 1 ข้ามกลุ่ม ไม่มีปฏิสัมพันธ์ร่วมใดที่มีนัยสำคัญทางสถิติ แสดงถึงระดับของเงื่อนไขวิธีการไม่มีนัยสำคัญทางสถิติ และไม่ใช่ความแตกต่างที่เด่นชัด อัตราความคลาดเคลื่อนประเภทที่ 1 ไม่มีผลกระทบต่อโมเดลการประมาณค่าพารามิเตอร์ภายใต้ขนาดกลุ่มตัวอย่างและความสามารถของกลุ่ม ส่วนการตรวจสอบอัตราความถูกต้อง พบว่า ความแปรปรวนของขนาดกลุ่มตัวอย่างส่งผลต่ออัตราความถูกต้องในวิธีการถดถอยโลจิสติก วิธีการทดสอบอัตราส่วนความควรจะเป็น ทฤษฎีการตอบสนองข้อคำถาม และวิธีซิปเทสท์ ซึ่งอธิบายสัดส่วนของความแปรปรวนมากที่สุด (มากกว่า 70%) ส่วนวิธีการวิเคราะห์องค์ประกอบไม่มีองค์ประกอบที่มีนัยสำคัญทางสถิติ วิธีอัตราส่วนความเป็นไปได้มีปฏิสัมพันธ์ภายใต้โมเดลการประมาณค่าพารามิเตอร์และความแตกต่างในความสามารถอย่างมีนัยสำคัญทางสถิติ และอธิบายความแปรปรวนได้ 4.5% โดยที่ความแตกต่างในความสามรถของกลุ่มไม่กระทบต่ออัตราความถูกต้อง และเมื่อกลุ่มมีความสามารถไม่เท่ากันของข้อมูลแบบ 3 พารามิเตอร์ ในการเปรียบเทียบวิธีการถดถอยโลจิสติกจะกระทำได้ยาก นอกจากนี้ยังพบว่า วิธีซิปเทสท์มีอัตราความถูกต้องที่สุดในทุกเงื่อนไข โดยทั่วไป ทั้งวิธีการทดสอบอัตราส่วนความเป็นไปได้ตามทฤษฎีการตอบข้อคำถาม และวิธีวิเคราะห์องค์ประกอบเชิงยืนยันจะมีอัตราความถูกต้องต่ำ ซึ่งนำไปสู่การเปรียบเทียบอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ดีกว่า สำหรับ

องค์ประกอบอื่น ๆ พบว่า โมเดลโลจิสติก 3 พารามิเตอร์มีส่วนทำให้อัตราความถูกต้องของวิธีซิปเทสท์ มีนัยสำคัญทางสถิติ และวิธีการถดถอยโลจิสติก โมเดลโลจิสติก 3 พารามิเตอร์ มีปฏิสัมพันธ์ร่วมกับความแตกต่างในความสามารถ และเมื่อมีการนำเสนอค่าการเดา จะทำให้วิธีการถดถอยโลจิสติกและวิธีซิปเทสท์ให้อัตราความถูกต้องต่ำ ส่วนในวิธีอัตราความควรจะเป็นของทฤษฎีการตอบข้อคำถามภายใต้โมเดลการประมาณค่าพารามิเตอร์ไม่มีผลต่ออัตราความถูกต้อง ในกรณีของการวิเคราะห์องค์ประกอบเชิงยืนยันอัตราความถูกต้องมีค่าต่ำ

Oishi (2006) ได้ตรวจสอบความเท่าเทียมของการวัดความพึงพอใจในชีวิตระหว่างกลุ่มตัวอย่างชาวอเมริกันและชาวจีน โดยใช้ Multi-group structure equation modeling (SEM) Multiple indicator multiple case model (MIMIC) และทฤษฎีการตอบสนองข้อคำถาม (IRT) กลุ่มตัวอย่างเป็นนักเรียน 422 คน ใน University of Illinois ที่ลงทะเบียนวิชาจิตวิทยาเบื้องต้น ใช้แบบทดสอบถามให้เด็กทำในชั้นเรียน การวัดผลทำโดยใช้แบบทดสอบ SWLS ที่ใช้ในการประเมินความพึงพอใจในชีวิตของคนทั่วโลก แบบทดสอบนี้ประกอบด้วย 5 ข้อคำถาม ผู้เข้าร่วมตอบสนองแต่ละข้อโดยใช้สเกล 7 ระดับ เพื่อจัดลำดับจาก 1 (ไม่เห็นด้วยมากที่สุด) ถึง 7 (เห็นด้วยมากที่สุด) ผลการวิจัยพบว่า การวิเคราะห์การทำหน้าที่ต่างกันของข้อคำถามจัดให้เห็นแง่มุมความแตกต่างของวิธีแบบดั้งเดิมในการวัดประเด็นการวิจัยทางวัฒนธรรมและความเป็นอยู่ที่ดี การวิเคราะห์ IRT แสดงให้เห็นความแตกต่างของค่าเฉลี่ยระหว่างกลุ่มชาวจีนและชาวอเมริกัน ข้อคำถามที่มีความลำเอียงจะให้คะแนนโดยมีน้ำหนักน้อยกว่า ดังนั้น ความแตกต่างของค่าเฉลี่ยที่ค้นพบก่อนหน้าระหว่างกลุ่มชาวจีนและชาวอเมริกันอาจจะไม่ค้นพบได้ง่ายในข้อคำถามที่มีความลำเอียง สุดท้าย การวิเคราะห์ IRT จัดหาข้อมูลที่เกี่ยวข้องกับแนวคิดของความพึงพอใจในชีวิต การวิจัยครั้งนี้แสดงให้เห็นความสำคัญและประโยชน์ของการวิเคราะห์การทำหน้าที่ต่างกันของข้อคำถามในโครงสร้างอื่น ๆ

Stark et al. (2006) ได้พัฒนาและทดสอบแผนการร่วมที่ใช้ในการระบุการทำหน้าที่ต่างกันของข้อคำถามรวมถึงการทดสอบอัตราส่วนไลค์ลิฮูด ซึ่งสามารถใช้ได้ทั้งการวิเคราะห์องค์ประกอบเชิงยืนยันและทฤษฎีการตอบสนองข้อคำถาม โดยใช้ข้อมูลจำลองในการตรวจสอบความตรงของทั้งสองวิธี IRT ตั้งบนพื้นฐานของวิธี Likelihood ratio test และ CFA ตั้งบนพื้นฐานของวิธี Mean and covariance structures (MACS) โดยใช้แบบทดสอบมิติเดียวจำนวน 15 ข้อคำถาม มีตัวแปรที่เกี่ยวข้อง 8 ตัว จำนวนเงื่อนไขที่ใช้ในการจำลองข้อมูลคือ 320 ($2^6 + 2^8$) แต่ละเงื่อนไขจะมีการทำซ้ำ 50 ครั้ง วิเคราะห์วิธี MACS โดยใช้โปรแกรมลิสเรล 8 และการทดสอบด้วยทฤษฎีการตอบสนองข้อคำถามด้วยวิธี Likelihood ratio test ใช้พื้นฐานของโมเดล Graded response โดยใช้โปรแกรมคอมพิวเตอร์ MULTILOG ผลการวิจัยพบว่า IRT วิธี Likelihood ratio test ให้ผล

ดีกว่าวิธี MACS ในขณะที่กลุ่มตัวอย่างขนาดเล็ก การวิเคราะห์ MACS ให้ผลดีกว่า หากในกรณีกลุ่มตัวอย่างมีขนาดใหญ่ และข้อมูลเป็นแบบสองค่า มิติเดียว การตรวจสอบการทำหน้าที่ต่างกันของแบบทดสอบโดยใช้วิธี IRT ให้ผลดีกว่า อย่างไรก็ตามวิธี IRT หลาย ๆ วิธีจะมีความแกร่งในการทดสอบ หากมีการฝ่าฝืนข้อตกลงเบื้องต้นในเรื่องความเป็นเอกมิติ

Kim, Cohen, Alagoz, and Kim (2007) ศึกษาการทำหน้าที่ต่างกันของขนาดอิทธิพลของข้อคำถามที่มีการตรวจให้คะแนนแบบหลายค่า (Polytomous) กลุ่มตัวอย่างขนาดใหญ่ (N = 105,731) เพื่อเปรียบเทียบถึงความสอดคล้องตามวิธีการตรวจสอบทั้ง 5 วิธี ได้แก่ วิธีการทดสอบอัตราส่วนไลด์ลิฮู้ดแบบ IRT วิธีการถดถอยโลจิสติก วิธีการทดสอบอัตราส่วนไลด์ลิฮู้ด วิธีแมนเทล และวิธีแมนเทล-แฮนส์เซลแบบทั่วไป (GMH) โดยใช้โปรแกรม Multilog และโปรแกรม IRTLRDIF วิเคราะห์ด้วยวิธีการทดสอบอัตราส่วนไลด์ลิฮู้ด ส่วนวิธีแมนเทล และวิธีแมนเทล-แฮนส์เซลแบบทั่วไปเขียนโปรแกรมด้วยภาษาฟอร์แทน ผลการวิจัยพบว่า สามารถตรวจพบข้อคำถามที่ทำหน้าที่ต่างกันจากทั้ง 5 วิธี ข้อค้นพบที่สำคัญ คือ การใช้กลุ่มตัวอย่างขนาดใหญ่เกินไปจะไม่มีประโยชน์ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม

Elosua and López-Jáuregui (2007) ได้ศึกษาแหล่งของการทำหน้าที่ต่างกันของข้อคำถามที่ส่งผลต่อการแปลแบบทดสอบการศึกษาในครั้งนี้ มีวัตถุประสงค์เพื่อหาแหล่งของการทำหน้าที่ต่างกันของข้อคำถามที่ส่งผลต่อการแปลแบบทดสอบ โดยจำแนกบนพื้นฐานของเกณฑ์ทางภาษาและวัฒนธรรม 4 แบบ คือ ความเกี่ยวข้องทางวัฒนธรรม ปัญหาการแปล ไวยากรณ์ และการตีความหมายคำ การศึกษาในครั้งนี้ใช้ข้อคำถามที่ตรวจให้คะแนนแบบ 2 ค่า เกี่ยวกับภาษาจำนวน 53 ข้อ ซึ่งเป็นข้อคำถามที่สร้างเป็นภาษาสเปน จากนั้นแปลเป็นภาษาบาสต์ และมีการแปลย้อนกลับอีกครั้ง (Black tanslation) กลุ่มตัวอย่างที่ศึกษาเป็นนักเรียนอายุระหว่าง 9-11 ปี 1,048 คน แบ่งเป็นกลุ่มอ้างอิง 498 คน และกลุ่มเปรียบเทียบ 550 คน วิเคราะห์ข้อมูลด้วยวิธีแมนเทล-แฮนส์เซล และจากความเห็นของผู้เชี่ยวชาญ พบว่า เกณฑ์ทั้ง 4 แบบ คือ ความเกี่ยวข้องทางวัฒนธรรม ปัญหาการแปล ไวยากรณ์ และการตีความหมายคำ ส่งผลต่อการทำหน้าที่ต่างกันของข้อคำถามได้ทั้งสิ้น 32 ข้อ ผู้เชี่ยวชาญตรวจสอบพบว่า การทำหน้าที่ต่างกันของข้อคำถามได้ 28 ข้อ และมีข้อคำถามที่ทั้งผู้เชี่ยวชาญ และวิธีแมนเทล-แฮนเซล ตรวจสอบพบว่า เกิดการทำหน้าที่ต่างกันของข้อคำถามได้ตรงกัน จำนวน 22 ข้อ และมีแหล่งของการทำหน้าที่ต่างกันของข้อคำถามทั้งสิ้น 29 แหล่ง

Walker, Zhang, and Surber (2008) ศึกษาการใช้กรอบแนวคิดกระบวนทัศน์ของการวิเคราะห์การทำหน้าที่ต่างกันของข้อคำถามแบบพหุมิติ ในการตัดสินความสามารถในการอ่านที่ส่งผลต่อความสามารถทางคณิตศาสตร์ วิเคราะห์ข้อมูลด้วยโปรแกรม NOHARM ผลการวิจัยพบว่า ความสามารถในการอ่านส่งผลต่อความสามารถทางคณิตศาสตร์ในทางบวก นั่นคือ นักเรียน

ที่มีความสามารถในการอ่านสูง จะสามารถทำคะแนนในส่วนของคณิตศาสตร์ได้สูงด้วย และมีนักเรียนเพียงส่วนหนึ่งเท่านั้น ที่มีความสามารถในการอ่านสูง แต่ทำคะแนนในส่วนของคณิตศาสตร์ได้ไม่ค่อยดี

Wiberg (2009) ศึกษาการทำหน้าที่ต่างกันของข้อคำถามของแบบทดสอบความสามารถระดับสูง เปรียบเทียบสามวิธี โดยใช้ข้อมูลจริง คือ วิธีลอกลิเนียร์ โมเดลโลจิสติกรีเกรสชัน และวิธีแมนเทล แฮนส์เซล ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบความสามารถระดับสูง ข้อมูลที่ใช้ในการวิเคราะห์มาจากผลการสอบเครื่องมือ Swedish theory driving license test (SDLT) และ Mastery test ประกอบด้วยข้อคำถามจำนวน 65 ข้อ ในระดับยาก ผู้เข้าร่วมต้องทำข้อคำถามได้อย่างน้อย 52 ข้อขึ้นไป ถึงผ่านการทดสอบ และจากผู้เข้าสอบ จำนวน 5,404 คน และสุ่มคัดเลือกข้อคำถามมา 15 ข้อ ที่ครอบคลุมหลักสูตรเพื่อนำมาตรวจสอบ DIF สถิติในการตรวจสอบ DIF คือ วิธีลอกลิเนียร์ โมเดลโลจิสติกรีเกรสชัน และวิธีแมนเทล แฮนส์เซล โดยใช้โปรแกรม R ผลการวิจัยพบว่า มีความสัมพันธ์กันค่อนข้างสูงเกี่ยวกับขนาดของการทำหน้าที่ต่างกันของข้อคำถาม การใช้วิธีแมนเทล-แฮนน์เซลให้ผลที่เหมือนกันกับทั้งสองวิธี วิธีลอกลิเนียร์ โมเดลโลจิสติกรีเกรสชัน ให้ผลสอดคล้องกันพอสมควร ส่วนวิธีลอกลิเนียร์โมเดลจะมีประโยชน์ในการให้ค่าช่วงคะแนนในการสอบที่แน่นอน ซึ่งถือเป็นสิ่งที่น่าสนใจเป็นพิเศษในแบบทดสอบความสามารถระดับสูง ซึ่งในการทดสอบคะแนนส่วนนี้ วิธีโลจิสติกรีเกรสชันและวิธีแมนเทล-แฮนส์เซล ให้ผลลัพธ์ที่แตกต่างกัน

Kahraman, De Boeck, and Janssen (2009) ศึกษารูปแบบ DIF ของข้อคำถามที่มีผลการตอบซับซ้อน โดยใช้ยุทธวิธีในการออกแบบแบบทดสอบ จุดมุ่งหมายของการศึกษาเพื่อนำเสนอวิธีการสร้างรูปแบบของการตอบสนองข้อมูลพหุมิติ กับกลุ่มโครงสร้างที่เกี่ยวข้องและปัจจัยหลักของกระบวนการประมาณค่าพารามิเตอร์ของข้อคำถาม ถูกขยายเพื่อรวมผลกระทบของมิติของแบบทดสอบและปัจจัยจากกลุ่มแตกต่างในสมรรถนะของการทำหน้าที่ต่างกันของข้อคำถามใน 2 ระดับ ข้อมูลที่ใช้ในการวิเคราะห์เป็นข้อมูลจริง จากการสุ่มนักเรียนประถมศึกษาเกรด 3, 4 ระดับละ 269 คน แบบทดสอบที่ใช้เป็นแบบเขียนตอบเกี่ยวกับคำศัพท์ที่กำหนดให้ สถิติที่ใช้ในการทดสอบ DIF ใช้ประมาณการปรับเหมาะของวิธีถดถอยโลจิสติก ภายใต้วิธีการของทฤษฎีการทดสอบแนวใหม่ ผลการวิจัยพบว่า การให้ตัวอย่างประกอบนี้เป็นการนำเสนอการใช้มาตราวัดความเชี่ยวชาญหรือชำนาญในการสะกดคำของชาวดัสช์ โดยดำเนินการจากกลุ่มย่อยคือปฏิสัมพันธ์ระหว่างกลุ่มกับข้อคำถามในแต่ละโมเดลหลัก โดยเฉพาะของข้อคำถามแต่ละข้อ

Gómez-Benito, Hidalgo, and Padilla (2009) ศึกษาประสิทธิภาพของขนาดอิทธิพลในการพัฒนาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามด้วยวิธีถดถอยโลจิสติก

โดยการจำลองข้อมูลที่แปรเปลี่ยน 5 ปัจจัย คือ รูปแบบของการทำหน้าที่ต่างกันของข้อคำถาม ขนาดอิทธิพลของการทำหน้าที่ต่างกันของข้อคำถาม จำนวนข้อคำถามที่เกิดการทำหน้าที่ต่างกันในแบบทดสอบแต่ละฉบับ ขนาดกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบและความยาวของข้อคำถามทั้งฉบับ ศึกษาภายใต้เงื่อนไข 225 เงื่อนไข วิธีการตรวจสอบการทำหน้าที่ต่างกัน เลือกใช้วิธีการถดถอยโลจิสติกภายใต้โมเดลทฤษฎีการตอบสนองข้อคำถามแบบ 2 พารามิเตอร์ ผลการวิจัยพบว่า ขนาดอิทธิพลที่เหมาะสมโดยพิจารณาเปรียบเทียบค่า R^2 จากเกณฑ์ Jodoin and Gierl (2001) ได้ศึกษาประสิทธิภาพของการวัดขนาดอิทธิพลในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม พบว่า การวัดขนาดอิทธิพลโดยสถิติ R^2 ร่วมกับการทดสอบนัยสำคัญจะได้ค่าที่ลดลงจนเกือบจะเป็นศูนย์ของการสรุปผิดว่าข้อคำถามทำหน้าที่ต่างกัน (DIF) ทั้งที่ความเป็นจริง ข้อคำถามไม่ได้ทำหน้าที่ต่างกัน (No DIF)(False positive: FP) โดยเมื่อข้อคำถามยิ่งมีความยาวมากขึ้น FP ยิ่งใกล้ศูนย์ และในทางกลับกัน การทดสอบนัยสำคัญของสถิติเพียงอย่างเดียวจะทำให้ได้ FP สูงกว่าเล็กน้อยหรือใกล้เคียงจากค่าปกติทั่วไป อย่างไรก็ตาม การวัดขนาดอิทธิพลโดยสถิติ R^2 ให้ผลของอำนาจการทดสอบ (1-β) ที่ต่ำลงจากาการทดสอบนัยสำคัญ ซึ่งผลการวิจัยสนับสนุนให้ศึกษาการวัดขนาดอิทธิพลโดยสถิติ R^2 ร่วมกับการทดสอบนัยสำคัญทางสถิติ จะทำให้ได้สารสนเทศมากยิ่งขึ้น

Kannan (2011) ได้ทำการศึกษาความแกร่งของการตรวจสอบการตอบสนองข้อคำถามแบบพหุมิติแบบหลายค่าและสองค่า โดยใช้วิธีการตรวจสอบการทำหน้าที่ต่างกันสองวิธี คือ MG-CFA และ MGRM-DFIT ใช้การจำลองตามเงื่อนไขที่แปรเปลี่ยน ได้แก่ ขนาดกลุ่มตัวอย่าง อัตราส่วนระหว่างกลุ่มเปรียบเทียบและกลุ่มอ้างอิง ชนิดของการทำหน้าที่ต่างกัน การแจกแจงของค่าความสามารถจริง โดยข้อมูลที่จำลองเป็นข้อคำถามให้คะแนนแบบสองค่าจำนวน 26 ข้อ และข้อคำถามแบบให้คะแนนหลายค่า 14 ข้อ 5 ระดับ และกำหนดความสัมพันธ์ระหว่างมิติเป็น .6 โดยตรวจสอบความมีประสิทธิภาพด้วยความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกัน ผลการทดสอบพบว่า วิธี MGRM-DFIT มีการควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดี มีอำนาจการทดสอบสูง และเมื่อเป็นการตรวจสอบการทำหน้าที่ต่างกันสำหรับแบบทดสอบผสมจะมีอำนาจในการทดสอบ DIF สูง แต่มีอำนาจการทดสอบต่ำสำหรับการตรวจสอบแบบแยกการตรวจให้คะแนน และ MGRM-DFIT มีอำนาจการทดสอบ DIF สูง เมื่อมีการแจกแจงของความสามารถจริงที่แตกต่างกัน และเมื่อมีการเปรียบเทียบข้อคำถามที่ไม่มี DIF และเป็นโมเดลแบบไม่ชดเชย มีอำนาจการทดสอบมากกว่า และค่าของโมเดลแบบชดเชยมีค่าน้อยกว่าเมื่อข้อคำถามมี DIF 4 ข้อ ซึ่งโมเดลแบบชดเชยเมื่อมี DIF ไม่สอดคล้องกับโมเดลแบบไม่ชดเชยที่มี DIF สำหรับวิธี MG-CFA มีอำนาจการทดสอบเพิ่มขึ้นเล็กน้อยในแต่ละเงื่อนไข ซึ่งทั้งสองวิธีสามารถตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามที่มีประสิทธิภาพคงเส้นคงวา

ทั้งการทดสอบแบบตัวแปรเดียวและหลายตัวแปร อย่างไรก็ตาม MGRM-DIF มีอำนาจการทดสอบสูงสำหรับการตรวจสอบการทำหน้าที่ต่างกันในแบบทดสอบ แต่ใช้ได้ไม่ดีสำหรับการตรวจสอบการทำหน้าที่ต่างกันของแบบรายข้อ ส่วน MG-CFA ง่ายต่อการวิเคราะห์เมื่อใช้โปรแกรมสำเร็จรูป

Woods et al. (2013) ศึกษาเปรียบเทียบวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามด้วยวิธี IRT-LR-DIF, Wald-I และ Wald-2 ภายใต้ปัจจัยที่แตกต่างกัน 3 ปัจจัย ได้แก่ จำนวนกลุ่ม (2 และ 3 กลุ่ม) ขนาดของกลุ่มตัวอย่าง (1,000/ 1,000, 500/ 500, 1,500/ 500, 1,500/ 500/ 500, 750/ 250 และ 750/ 250/ 250 จำนวนข้อที่ทำหน้าที่ต่างกันในแบบทดสอบ (25% และ 50%) ทำทั้งหมด 500 ซ้ำ ในแต่ละเงื่อนไข โดยการศึกษาครั้งนี้ใช้การจำลองข้อมูลภายใต้โมเดลเกรดเรสพอน ข้อคำถามจำนวน 24 ข้อ แต่ละข้อมี 5 ระดับ ตรวจสอบการทำหน้าที่ต่างกันด้วยโปรแกรม IRTLRDIF สำหรับวิธี IRT-LR-DIF และใช้โปรแกรม flexMIRT สำหรับการทดสอบ Wald-I และ Wald-II ผลการศึกษาพบว่า Wald-I มีประสิทธิภาพในการทดสอบสูงกว่าในทุกกรณี ส่วนวิธี Wald-2 มีความคลาดเคลื่อนประเภทที่ 1 สูงกว่าวิธีอื่น ๆ และ IRTLRDIF และ Wald-I มีประสิทธิภาพเท่า ๆ กัน สำหรับกรณีทดสอบแบบสามกลุ่ม

Cao et al. (2017) ศึกษาประสิทธิภาพของวิธีการทดสอบ Wald แบบวนซ้ำสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามระหว่างกลุ่มสองกลุ่ม เมื่อทราบว่าไม่มีข้อคำถามร่วม ซึ่งวิธีการทำซ้ำนี้จะใช้วิธีการของ Wald-2 เพื่อค้นหาข้อคำถามร่วม (Anchor) และเมื่อทราบแล้วจะใช้วิธีการทดสอบ Wald-1 (Woods et al., 2013) ซึ่งการทำซ้ำนี้จะใช้การจำลองข้อมูลภายใต้เงื่อนไขความแตกต่างกันของจำนวนรายการตอบของข้อคำถาม ความยาวข้อคำถาม ขนาดตัวอย่าง จำนวนข้อคำถามที่มี DIF ขนาดของ DIF และชนิดของ DIF (Compenasory, noncompensatory) ผลการศึกษาพบว่า ประสิทธิภาพของการทำซ้ำสูงเมื่อข้อมูลเป็นแบบให้คะแนนหลายค่าในทุกเงื่อนไข และสามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดีและมีอำนาจการทดสอบสูง สำหรับรายการตอบให้คะแนนสองค่าพบว่า วิธีการทำซ้ำมีการนำเสนอค่าของความคลาดเคลื่อนประเภทที่ 1 ได้ดีกว่าวิธี Wald-2 โดยไม่เสียอำนาจของการทดสอบ DIF อย่างไรก็ตาม ก็ยังคงพบความคลาดเคลื่อนประเภทที่ 1 ขนาดใหญ่ ในกรณีที่เป็น Noncompensatory จำนวนข้อคำถาม DIF มาก และในขนาดของ DIF ขนาดใหญ่ อย่างไรก็ตาม ขนาดของความคลาดเคลื่อนประเภทที่ 1 ยังมีค่าน้อยกว่าเมื่อเทียบกับวิธีทดสอบ Wald-2

จากการศึกษางานวิจัยภายในประเทศและต่างประเทศที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ด้วยวิธีโพลีโตมัสซิปเทสท์ วิธีการวิเคราะห์องค์ประกอบเชิงยืนยันจำแนกพหุ และวิธีการทดสอบวอลด์ ภายใต้เงื่อนไขปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ ขนาด

ของการทำหน้าที่ต่างกัน ความยาวของแบบทดสอบ สัดส่วนของการทำหน้าที่ต่างกัน และขนาดตัวอย่าง สามารถสรุปได้ ดังนี้

ขนาดของการทำหน้าที่ต่างกัน (Magnitude of DIF)

จากการงานวิจัยที่ศึกษาการทำหน้าที่ต่างกันของข้อคำถามที่ผ่านมา (Flowers et al., 1999; Meade et al., 2006; Stark et al., 2006) พบว่า ข้อคำถามที่มีขนาดของการทำหน้าที่ต่างกันที่สูงกว่าสามารถตรวจพบการทำหน้าที่ต่างกันได้มากกว่า ขนาดของการทำหน้าที่ต่างกันต่ำ นอกจากนี้ยังพบว่า การศึกษาการทำหน้าที่ต่างกันของข้อคำถามสำหรับวิธีการทดสอบโดยใช้ทฤษฎีการตอบสนองข้อคำถาม (Flowers et al., 1999; Meade et al., 2006) ที่มีขนาดของการทำหน้าที่ต่างกันของข้อคำถามขนาดใหญ่ ($\geq$ .4) จะมีอำนาจการทดสอบการทำหน้าที่ต่างกันได้มากกว่า

ความยาวของแบบทดสอบ

การตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามด้วยการจำลองข้อมูล ความยาวของแบบทดสอบ เป็นปัจจัยหนึ่งที่ถูกนำมาใช้ในการศึกษา French and Maller (2007); Jodoin and Gierl (2001); Narayanan and Swaminathan (1996); Paek and Wilson (2011); Rogers and Swaminathan (1993); Woods (2009); Yao and Li (2010) พบว่า อำนาจการทดสอบมีค่าสูงขึ้นถ้าความยาวของแบบทดสอบมากขึ้น สำหรับการศึกษาข้อคำถามให้คะแนนหลายค่าด้วยวิธีการจำลองข้อมูลส่วนใหญ่จะศึกษาที่ความยาวของแบบทดสอบน้อยกว่า 50 ข้อ (Su & Wang, 2005) โดยศึกษาอยู่ในช่วง 10 ถึง 40 ข้อ (Flowers et al., 1999; Wang & Su, 2004; Williams & Beretvas, 2006; Woods, 2009) จากการศึกษาของ Kim and Chen (1998 อ้างถึงใน ปิยะทิพย์ ตินวร, 2549) พบว่า ความยาวของแบบทดสอบมีผลกระทบต่ออำนาจการทดสอบในการตรวจสอบการทำหน้าที่ต่างกัน และงานวิจัยของ Narayanan and Swaminathan (1996) พบว่า แบบทดสอบที่มีความยาว 40 ข้อ เป็นแบบทดสอบที่มีความยาวเพียงพอในการตรวจสอบการทำหน้าที่ต่างกัน ซึ่งสอดคล้องกับ Snow and Oshima (2009) นอกจากนี้ Wu and Lei (2009) ได้ทำการศึกษาโดยใช้การจำลองข้อมูลเพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามด้วยวิธี MG-CFA ด้วยความยาวของแบบทดสอบจำนวน 40 ข้อ และพบว่า วิธี MG-CFA มีความแกร่งในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม

สัดส่วนของการทำหน้าที่ต่างกันของข้อคำถาม

เนื่องจากการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามจำเป็นต้องเปรียบเทียบความสามารถของผู้สอบภายใต้คะแนนเกณฑ์การจับคู่ที่มีความเชื่อมั่น (Clauser & Mazor, 1988; Potenza & Dorans, 1995) ถ้าในแบบทดสอบที่มีข้อคำถามทำหน้าที่ต่างกันปะปนอยู่ จะทำให้ค่าประมาณความสามารถมีความเชื่อมั่นต่ำลง ซึ่งจะมีผลทำให้อำนาจการทดสอบลดลง หรืออัตราความคลาดเคลื่อนประเภทที่ 1 สูงกว่าปกติ สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกันได้มีการศึกษา

อย่างต่อเนื่อง ซึ่งส่วนใหญ่จะทำการศึกษาที่ 10% ของจำนวนข้อคำถามทั้งหมด Raju et al. (1995); Flowers et al. (1999) พบว่า สัดส่วนของการทำหน้าที่ต่างกันของข้อคำถามที่มีสัดส่วนการทำหน้าที่ต่างกันมาก (มากกว่า 20%) จะมีผลทำให้พบขนาดของผลบวกลวง (False positive) และผลลบลวง (False negative) มีค่าสูงขึ้น เมื่อสัดส่วนของการทำหน้าที่ต่างกันมากขึ้น กล่าวคือ อัตราความคลาดเคลื่อนประเภทที่ 1 และความคลาดเคลื่อนประเภทที่ 2 มีค่ามากขึ้น ดังนั้น นักวิจัยรุ่นหลังจึงกำหนดสัดส่วนของการทำหน้าที่ต่างกันของข้อคำถามที่มีค่าอยู่ใกล้ 10% (Bolt, 2002; Gonzales-Roma et al., 2006; Stark et al., 2006; Wu & Lei, 2009)

ขนาดตัวอย่าง

สำหรับการทดสอบการทำหน้าที่ต่างกันของข้อคำถาม ซึ่งเป็นการเปรียบเทียบผลการตอบระหว่างผู้สอบสองกลุ่ม การทดสอบการทำหน้าที่ต่างกันนี้ ขนาดตัวอย่างก็เป็นปัจจัยหนึ่งที่สำคัญสำหรับการทดสอบการทำหน้าที่ต่างกัน (MacCallum et al., 1999) ในการศึกษาที่ผ่านมามีการกำหนดขนาดตัวอย่างสำหรับการศึกษาทั้งแบบที่มีสัดส่วนเท่ากัน (Gonzales-Roma et al., 2006; Lubke & Muthén, 2004; Meade & Lautenschlager, 2004) และสัดส่วนที่ต่างกัน (Finch & French, 2007; Maller & French, 2004) โดยการศึกษามีช่วงของการกำหนดตัวอย่างระหว่าง 250 ถึง 1,000 (Gonzalez-Roma et al., 2006; Meade & Lautenschlager, 2004; Meade et al., 2006; Stark et al., 2006)

บทที่ 3
วิธีดำเนินการวิจัย

การวิจัยครั้งนี้มีจุดมุ่งหมาย เพื่อศึกษาและเปรียบเทียบวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ได้แก่ วิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ ภายใต้ 4 ปัจจัยที่แปรเปลี่ยน คือ ความยาวของแบบทดสอบ ขนาดของการทำหน้าที่ต่างกัน สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง โดยใช้วิธีการจำลองข้อมูลมีการดำเนินการวิจัย ตามขั้นตอนดังต่อไปนี้

1. การจัดกระทำตัวแปร
2. การจำลองข้อมูล
3. การวิเคราะห์ข้อมูล

การจัดกระทำตัวแปร

การศึกษาวิจัยครั้งนี้ ศึกษาในสถานการณ์จำลองโดยใช้โมเดลเกรดเรสพอนพหุมิติ (Multidimensional graded response model: MGRM) จำลองข้อมูลภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย ได้แก่ ความยาวของแบบทดสอบ 2 ขนาด ขนาดของการทำหน้าที่ต่างกัน 3 วิธี สัดส่วนของการทำหน้าที่ต่างกัน 2 ขนาด และขนาดตัวอย่างที่แตกต่างกัน 5 รูปแบบ โดยจำลองผลการตอบแบบทดสอบที่มีโครงสร้างแบบพหุมิติ 2 มิติ ข้อคำถามแต่ละข้อมีรายการคำตอบ 5 รายการ ให้คะแนน 1, 2, 3, 4 หรือ 5 สำหรับการจัดกระทำดังกล่าว ผู้วิจัยได้ออกแบบการวิจัยโดยมีรายละเอียด ดังนี้

ความยาวของแบบทดสอบ

ความยาวของแบบทดสอบระดับปานกลางขึ้นไป มีผลต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามมากที่สุด แบบทดสอบที่มีความยาว 40 ข้อ เป็นแบบทดสอบที่มีความยาวเพียงพอในการตรวจสอบการทำหน้าที่ต่างกัน (Narayanan & Swaminathan, 1996; Wu & Lei, 2009) ส่วน Klockars and Lee (2008) พบว่า ความยาวของแบบทดสอบควรมากขึ้นถ้าขนาดของกลุ่มตัวอย่างมากขึ้น ($\geq$ 1,000) เพราะจะส่งผลให้ค่าอำนาจการทดสอบมีค่ามากขึ้น โดยเฉพาะการตรวจสอบด้วย SIBTEST แต่จะไม่มีผลถ้าขนาดตัวอย่างน้อยกว่า 500 สำหรับการทดสอบด้วยวิธี MG-CFA พบว่า แบบทดสอบ จำนวน 40 ข้อ มีความแกร่งสำหรับการทดสอบการทำหน้าที่ต่างกัน (Wu and Lei, 2009) สำหรับการศึกษาแบบทดสอบให้คะแนนหลายค่าด้วยการจำลองส่วนใหญ่

จะศึกษาในช่วง 10 ถึง 40 ข้อ (Flowers et al., 1999; Wang & Su, 2004; Williams & Beretvas, 2006; Woods, 2009) วริษยา ชะม้อย (2550) ได้ทำการเปรียบเทียบวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบให้คะแนนหลายค่า ด้วยความยาวของแบบทดสอบ จำนวน 20 และ 40 ข้อ วิจัยครั้งนี้สนใจศึกษาความยาวของแบบทดสอบ 2 ขนาด คือ 20 ข้อ และ 40 ข้อ

ขนาดของการทำหน้าที่ต่างกัน

จากการศึกษาของ Flowers et al. (1999); Meade et al. (2006); Stark et al. (2006) พบว่า ข้อคำถามที่มีขนาดของการทำหน้าที่ต่างกันสูงจะทำให้การตรวจพบการทำหน้าที่ต่างกันได้มากกว่า โดยพบว่า ขนาดของการทำหน้าที่ต่างกันของข้อคำถามขนาดใหญ่ ($\geq$ .4) จะมีอำนาจในการทดสอบสูง (Flowers et al., 1999; Meade et al., 2006) ผู้วิจัยจึงกำหนดขนาดของการทำหน้าที่ต่างกัน 3 ขนาด คือ 0.4, 0.7 และ 1.0

สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน

สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกันได้มีการศึกษาอย่างต่อเนื่อง ซึ่งส่วนใหญ่จะทำการศึกษาที่ 10% ของจำนวนข้อคำถามทั้งหมด ที่เป็นเช่นนี้เพราะว่า Raju et al. 1995; Flowers et al., 1999 พบว่า สัดส่วนของการทำหน้าที่ต่างกันของข้อคำถามที่มากกว่า 20% จะทำให้เกิด False positives และ False negatives ขนาดใหญ่ ดังนั้น นักวิจัยรุ่นหลังจึงกำหนดสัดส่วนของการทำหน้าที่ต่างกันของข้อคำถามที่มีค่าอยู่ใกล้ 10% (Bolt, 2002; Gonzales-Roma et al., 2006; Stark et al., 2006; Wu & Lei, 2009) ในการศึกษาครั้งนี้ ผู้วิจัยจึงสนใจสัดส่วนของการทำหน้าที่ต่างกัน 2 ขนาด ได้แก่ 10% และ 20%

ขนาดตัวอย่าง

การศึกษาที่ผ่านมามีการกำหนดขนาดตัวอย่างสำหรับการศึกษาทั้งแบบที่มีสัดส่วนเท่ากัน (Lubke & Muthén, 2004; Meade & Lautenschlager, 2004; Gonzales-Roma et al., 2006) และสัดส่วนที่ต่างกัน (Finch & French, 2007; Maller & French, 2004) โดยการศึกษามีช่วงของการกำหนดตัวอย่างระหว่าง 250 ถึง 1,000 (Meade et al., 2006; Meade & Lautenschlager, 2004; Stark et al., 2006, Gonzales-Roma et al., 2006) ดังนั้น ผู้วิจัยจึงได้กำหนดตัวอย่างสำหรับการศึกษาในครั้งนี้ โดยใช้สัดส่วนของกลุ่มสนใจต่อกลุ่มอ้างอิง โดยใช้อัตราส่วน 1 : 1 และ 1 : 2 จำนวน 5 รูปแบบ คือ 250 : 250, 500 : 500, 1,000 : 1,000, 250 : 500 และ 500 : 1,000 คน

การจำลองข้อมูล

การศึกษาครั้งนี้ ผู้วิจัยใช้ข้อมูลจำลอง ภายใต้ความยาวของแบบทดสอบจำนวน 20 โดยข้อที่ 1 ถึง ข้อที่ 10 วัดความสามารถในมิติที่ 1 และข้อที่ 11 ถึง 20 และ 40 ข้อ โดย ข้อที่ 1

ถึงข้อที่ 20 วัดความสามารถในมิติที่ 1 และข้อที่ 21 ถึง 40 วัดความสามารถในมิติที่ 2 โดยทั้งสองมิติมีความสัมพันธ์ที่ระดับ .50 แต่ละข้อมีรายการคำตอบ 5 รายการให้คะแนนเป็น 1, 2, 3, 4, หรือ 5 ขนาดของการทำหน้าที่ต่างกันของข้อคำถาม 3 ขนาด คือ 0.4, 0.7 และ 1.0 สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 2 ขนาด คือ 10% และ 20% และขนาดตัวอย่าง 5 รูปแบบ คือ 250 : 250, 500 : 500, 1,000 : 1,000, 250 : 500 และ 500 : 1,000 คน จำลองภายใต้โมเดลเกรดเรสพอนพหุมิติ คือ

$$P\left(u_{vij}=k\middle|\boldsymbol{\theta}_{\mathbf{vj}}\right)=\frac{1}{\sqrt{2\pi}}\int_{a'_{vi}\theta_{vj}+d_{vik+1}}^{a'_{vi}\theta_{vj}+d_{vik}} e^{\frac{-t^2}{2}}dt$$

เมื่อ $P\left(u_{vij}\right)$ แทน ความน่าจะเป็นของผู้สอบคนที่ j ที่มีความสามารถ θ ในมิติ v ตอบข้อคำถามที่ i จากรายการคำตอบ k

$\boldsymbol{\theta}_{\mathbf{vj}}$ แทน พารามิเตอร์ความสามารถของผู้สอบคนที่ j ในมิติที่ v

$a_{\mathbf{v}i}$ แทน พารามิเตอร์อำนาจจำแนกของข้อคำถามข้อที่ i ในมิติที่ v

d_{vik} แทน พารามิเตอร์ข้อคำถามเทรสโฮล ของข้อคำถามข้อที่ i ในรายการที่ k

สำหรับการจำลองความสามารถผู้สอบ และพารามิเตอร์ของข้อคำถาม จำลองโดยโปรแกรม R version 3.3.2 จำลองภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย ซึ่งแบ่งการนำเสนอเป็นขั้นตอนหลัก 3 ขั้นตอน ดังนี้

ขั้นที่ 1 การสร้างข้อมูลผู้สอบ

ระบุจำนวนผู้สอบทั้งกลุ่มสนใจต่อกลุ่มอ้างอิงตามสัดส่วนของขนาดตัวอย่าง (250 : 250, 500 : 500, 1,000 : 1,000, 250 : 500 และ 500 : 1,000 คน) โดยจำลองข้อมูลความสามารถของผู้สอบในแต่ละกลุ่มทั้งสองมิติด้วยการแจกแจงแบบปกติหลายตัวแปร (Multivariate normal) ค่าเฉลี่ย 0 และส่วนเบี่ยงเบนมาตรฐานเป็น 1 และกำหนดให้ความสามารถทั้งสองมิติให้มีความสัมพันธ์กันที่ระดับ .50

ขั้นที่ 2 การสร้างข้อมูลข้อคำถาม

1. ระบุจำนวนของข้อคำถามในแบบทดสอบตามความยาวของข้อคำถามที่กำหนด (20 และ 40 ข้อ) พร้อมทั้งกำหนดจำนวนรายการคำตอบในแต่ละข้อเป็นแบบ 5 รายการตอบ ภายใต้โมเดลเกรดเรสพอนพหุมิติ โดยผู้วิจัยกำหนดสัดส่วนจำนวนข้อคำถามในแต่ละมิติเป็น 1 : 1 ถ้าความยาวข้อคำถามที่กำหนดเป็น 20 และ 40 ข้อ จะทำให้ข้อคำถามที่วัดในมิติที่ 1 และ 2 เป็นมิติละ 10 และ 20 ข้อ ตามลำดับ

2. เนื่องจากการทดสอบครั้งนี้เป็นการทดสอบภายใต้การทำหน้าที่ต่างกันของข้อคำถามแบบ Non uniform โดยให้ค่าพารามิเตอร์อำนาจจำแนกและพารามิเตอร์เทรสโฮลมีค่าแตกต่างกันทั้งกลุ่มสนใจและกลุ่มอ้างอิง จำลองการแจกแจงพารามิเตอร์ของข้อสอบ ได้แก่ ค่าอำนาจจำแนกของข้อคำถาม เป็นการแจกแจงแบบปกติ ค่าเฉลี่ย 0 และความแปรปรวน 1 โดยค่าที่จำลองได้มีค่าอยู่ตั้งแต่ 0 ถึง 1 โดยลักษณะค่าพารามิเตอร์อำนาจจำแนกของกลุ่มสนใจจะมีค่าแตกต่างจากกลุ่มอ้างอิงสำหรับข้อคำถามที่ทำหน้าที่ต่างกัน ดังนี้

2.1 กรณีความยาวของแบบทดสอบ 20 ข้อ และสัดส่วนข้อคำถามทำหน้าที่ต่างกัน 10% ในเงื่อนไขนี้จะมีข้อคำถามที่ทำหน้าที่ต่างกันสองข้อคือ ข้อที่ 10 และ ข้อที่ 20 ถ้าขนาดของการทำหน้าที่ต่างกัน .40 ดังนั้น ข้อคำถามสองข้อนี้จะมีค่าอำนาจจำแนกแตกต่างกัน โดยอำนาจจำแนกของกลุ่มสนใจจะมีค่าน้อยกว่ากลุ่มอ้างอิงเท่ากับ .40 ($a_{iR} = a_{iF} + .40$) เมื่อ $i = 10, 20$ ถ้าขนาดของการทำหน้าที่ต่างกัน .70 ค่าอำนาจจำแนกของกลุ่มสนใจจะมีค่าน้อยกว่ากลุ่มอ้างอิงเท่ากับ .70 ($a_{iR} = a_{iF} + .70$) เมื่อ $i = 10, 20$ และถ้าขนาดของการทำหน้าที่ต่างกัน 1.00 ค่าอำนาจจำแนกของกลุ่มสนใจจะมีค่าน้อยกว่ากลุ่มอ้างอิงเท่ากับ 1.00 ($a_{iR} = a_{iF} + 1.00$) เมื่อ $i = 10, 20$ ข้ออื่น ๆ จะมีค่าอำนาจจำแนกเท่ากันทั้งสองกลุ่ม

2.2 กรณีความยาวของแบบทดสอบ 20 ข้อ และสัดส่วนข้อคำถามทำหน้าที่ต่างกัน 20% ในเงื่อนไขนี้จะมีข้อคำถามที่ทำหน้าที่ต่างกันสองข้อคือ ข้อที่ 9, 10 และ ข้อที่ 19, 20 ถ้าขนาดของการทำหน้าที่ต่างกัน .40 ดังนั้น ข้อคำถามสี่ข้อนี้จะมีค่าอำนาจจำแนกแตกต่างกัน โดยอำนาจจำแนกของกลุ่มสนใจจะมีค่าน้อยกว่ากลุ่มอ้างอิงเท่ากับ .40 ($a_{iR} = a_{iF} + .40$) เมื่อ $i = 9, 10, 19, 20$ ถ้าขนาดของการทำหน้าที่ต่างกัน .70 ค่าอำนาจจำแนกของกลุ่มสนใจจะมีค่าน้อยกว่ากลุ่มอ้างอิงเท่ากับ .70 ($a_{iR} = a_{iF} + .70$) เมื่อ $i = 9, 10, 19, 20$ และถ้าขนาดของการทำหน้าที่ต่างกัน 1.00 ค่าอำนาจจำแนกของกลุ่มสนใจจะมีค่าน้อยกว่ากลุ่มอ้างอิงเท่ากับ 1.00 ($a_{iR} = a_{iF} + 1.00$) เมื่อ $i = 9, 10, 19, 20$ ส่วนข้ออื่น ๆ จะมีค่าอำนาจจำแนกเท่ากันทั้งสองกลุ่ม

2.3 กรณีความยาวของแบบทดสอบ 40 ข้อ และสัดส่วนข้อคำถามทำหน้าที่ต่างกัน 10% ในเงื่อนไขนี้จะมีข้อคำถามที่ทำหน้าที่ต่างกันสองข้อคือ ข้อที่ 9, 10 และข้อที่ 19, 20 ถ้าขนาดของการทำหน้าที่ต่างกัน .40 ดังนั้น ข้อคำถามสี่ข้อนี้จะมีค่าอำนาจจำแนกแตกต่างกัน โดยอำนาจจำแนกของกลุ่มสนใจจะมีค่าน้อยกว่ากลุ่มอ้างอิงเท่ากับ .40 ($a_{iR} = a_{iF} + .40$) เมื่อ $i = 9, 10, 19, 20$ ถ้าขนาดของการทำหน้าที่ต่างกัน .70 อำนาจจำแนกของกลุ่มสนใจจะมีค่าน้อยกว่ากลุ่มอ้างอิงเท่ากับ .70 ($a_{iR} = a_{iF} + .70$) เมื่อ $i = 9, 10, 19, 20$ และถ้าขนาดของการทำหน้าที่ต่างกัน 1.00 ค่าอำนาจจำแนกของกลุ่มสนใจจะมีค่าน้อยกว่ากลุ่มอ้างอิงเท่ากับ 1.00 ($a_{iR} = a_{iF} + 1.00$) เมื่อ $i = 9, 10, 19, 20$ ส่วนข้ออื่น ๆ จะมีค่าอำนาจจำแนกเท่ากันทั้งสองกลุ่ม

2.4 กรณีความยาวของแบบทดสอบ 40 ข้อ และสัดส่วนข้อคำถามทำหน้าที่ต่างกัน 20% ในเงื่อนไขนี้จะมีข้อคำถามที่ทำหน้าที่ต่างกันสองข้อคือ ข้อที่ 7, 8, 9, 10 และข้อที่ 17, 18, 19, 20 ถ้าขนาดของการทำหน้าที่ต่างกัน .40 ดังนั้น ข้อคำถามแปดข้อนี้จะมีค่าอำนาจจำแนกแตกต่างกัน โดยอำนาจจำแนกของกลุ่มสนใจจะมีค่าน้อยกว่ากลุ่มอ้างอิงเท่ากับ .40 ($a_{iR} = a_{iF} + .40$) เมื่อ i = 7, 8, 9, 10, 17, 18, 19, 20 ถ้าขนาดของการทำหน้าที่ต่างกัน .70 ค่าอำนาจจำแนกของกลุ่มสนใจจะมีค่าน้อยกว่ากลุ่มอ้างอิงเท่ากับ .70 ($a_{iR} = a_{iF} + .70$) เมื่อ i = 7, 8, 9, 10, 17, 18, 19, 20 และถ้าขนาดของการทำหน้าที่ต่างกัน 1.00 ค่าอำนาจจำแนกของกลุ่มสนใจจะมีค่าน้อยกว่ากลุ่มอ้างอิงเท่ากับ 1.00 ($a_{iR} = a_{iF} + 1.00$) เมื่อ i = 7, 8, 9, 10, 17, 18, 19, 20 ส่วนข้ออื่น ๆ จะมีค่าอำนาจจำแนกเท่ากันทั้งสองกลุ่ม

3. จำลองพารามิเตอร์เทรสโฮลของรายการคำตอบเป็นการแจกแจงแบบปกติ มีค่าเฉลี่ย 0 ความแปรปรวน 1 โดยลักษณะค่าพารามิเตอร์เทรสโฮลของกลุ่มสนใจจะมีค่าแตกต่างจากกลุ่มอ้างอิงสำหรับข้อคำถามที่ทำหน้าที่ต่างกัน ดังนี้

3.1 กรณีความยาวของแบบทดสอบ 20 ข้อ และสัดส่วนข้อคำถามทำหน้าที่ต่างกัน 10% ในเงื่อนไขนี้จะมีข้อคำถามที่ทำหน้าที่ต่างกันสองข้อคือ ข้อที่ 10 และข้อที่ 20 ถ้าขนาดของการทำหน้าที่ต่างกัน .40 ดังนั้น ข้อคำถามสองข้อนี้จะมีค่าเทรสโฮลแตกต่างกัน

โดยค่าเทรสโฮล ของกลุ่มสนใจจะมีค่ามากกว่ากลุ่มอ้างอิงเท่ากับ .40

$$(\tau_{i1R} = \tau_{i1F} - 0.40, \tau_{i2R} = \tau_{i2F} - 0.40, \tau_{i3R} = \tau_{i3F} - 0.40, \tau_{i4R} = \tau_{i4F} - 0.40)$$

เมื่อ i = 10, 20

ถ้าขนาดของการทำหน้าที่ต่างกัน .70 ค่าเทรสโฮลของกลุ่มสนใจ จะมีค่ามากกว่ากลุ่มอ้างอิงเท่ากับ .70

$$(\tau_{i1R} = \tau_{i1F} - 0.70, \tau_{i2R} = \tau_{i2F} - 0.70, \tau_{i3R} = \tau_{i3F} - 0.70, \tau_{i4R} = \tau_{i4F} - 0.70)$$

เมื่อ i = 10, 20

และถ้าขนาดของการทำหน้าที่ต่างกัน 1.00 ค่าเทรสโฮลของกลุ่มสนใจจะมีค่ามากกว่ากลุ่มอ้างอิงเท่ากับ 1.00

$$(\tau_{i1R} = \tau_{i1F} - 1.00, \tau_{i2R} = \tau_{i2F} - 1.00, \tau_{i3R} = \tau_{i3F} - 1.00, \tau_{i4R} = \tau_{i4F} - 1.00)$$

เมื่อ i = 10, 20

ส่วนข้ออื่น ๆ จะมีค่าเทรสโฮลเท่ากันทั้งสองกลุ่ม

3.2 กรณีความยาวของแบบทดสอบ 20 ข้อ และสัดส่วนข้อคำถามทำหน้าที่ต่างกัน 20% ในเงื่อนไขนี้จะมีข้อคำถามที่ทำหน้าที่ต่างกันสองข้อคือ ข้อที่ 9, 10 และข้อที่ 19, 20 ถ้าขนาดของการทำหน้าที่ต่างกัน .40 ดังนั้น ข้อคำถามสี่ข้อนี้จะมีค่าเทรสโฮลแตกต่างกัน

โดยเทรสโฮล ของกลุ่มสนใจจะมีค่ามากกว่ากลุ่มอ้างอิงเท่ากับ .40

$(\tau_{i1R} = \tau_{i1F} - 0.40, \tau_{i2R} = \tau_{i2F} - 0.40, \tau_{i3R} = \tau_{i3F} - 0.40, \tau_{i4R} = \tau_{i4F} - 0.40)$

เมื่อ i = 9, 10, 19, 20

ถ้าขนาดของการทำหน้าที่ต่างกัน .70 ค่าเทรสโฮลของกลุ่มสนใจจะมีค่ามากกว่ากลุ่มอ้างอิงเท่ากับ .70

$(\tau_{i1R} = \tau_{i1F} - 0.70, \tau_{i2R} = \tau_{i2F} - 0.70, \tau_{i3R} = \tau_{i3F} - 0.70, \tau_{i4R} = \tau_{i4F} - 0.70)$

เมื่อ i = 9, 10, 19, 20

และถ้าขนาดของการทำหน้าที่ต่างกัน 1.00 ค่าเทสโฮลของกลุ่มสนใจจะมีค่ามากกว่ากลุ่มอ้างอิงเท่ากับ 1.00

$(\tau_{i1R} = \tau_{i1F} - 1.00, \tau_{i2R} = \tau_{i2F} - 1.00, \tau_{i3R} = \tau_{i3F} - 1.00, \tau_{i4R} = \tau_{i4F} - 1.00)$

เมื่อ i = 9, 10, 19, 20

ส่วนข้ออื่น ๆ จะมีค่าเทรสโฮลเท่ากันทั้งสองกลุ่ม

3.3 กรณีความยาวของแบบทดสอบ 40 ข้อ และสัดส่วนข้อคำถามทำหน้าที่ต่างกัน 10% ในเงื่อนไขนี้จะมีข้อคำถามที่ทำหน้าที่ต่างกันสองข้อคือ ข้อที่ 9, 10 และข้อที่ 19, 20 ถ้าขนาดของการทำหน้าที่ต่างกัน .40 ดังนั้น ข้อคำถามสี่ข้อนี้จะมีค่าเทรสโฮลแตกต่างกัน

โดยเทรสโฮลของกลุ่มสนใจจะมีค่าน้อยกว่ากลุ่มอ้างอิงเท่ากับ .40

$(\tau_{i1R} = \tau_{i1F} - 0.40, \tau_{i2R} = \tau_{i2F} - 0.40, \tau_{i3R} = \tau_{i3F} - 0.40, \tau_{i4R} = \tau_{i4F} - 0.40)$

เมื่อ i = 9, 10, 19, 20

ถ้าขนาดของการทำหน้าที่ต่างกัน .70 ค่าเทรสโฮลของกลุ่มสนใจจะมีค่ามากกว่ากลุ่มอ้างอิงเท่ากับ 0.70

$(\tau_{i1R} = \tau_{i1F} - 0.70, \tau_{i2R} = \tau_{i2F} - 0.70, \tau_{i3R} = \tau_{i3F} - 0.70, \tau_{i4R} = \tau_{i4F} - 0.70)$

เมื่อ i = 9, 10, 19, 20

และถ้าขนาดของการทำหน้าที่ต่างกัน 1.00 ค่าเทรสโฮลของกลุ่มสนใจจะมีค่ามากกว่ากลุ่มอ้างอิงเท่ากับ 1.00

$(\tau_{i1R} = \tau_{i1F} - 1.00, \tau_{i2R} = \tau_{i2F} - 1.00, \tau_{i3R} = \tau_{i3F} - 1.00, \tau_{i4R} = \tau_{i4F} - 1.00)$

เมื่อ i = 9, 10, 19, 20

ส่วนข้ออื่น ๆ จะมีเทรสโฮลเท่ากันทั้งสองกลุ่ม

3.4 กรณีความยาวของแบบทดสอบ 40 ข้อ และสัดส่วนข้อคำถามทำหน้าที่ต่างกัน 20% ในเงื่อนไขนี้จะมีข้อคำถามที่ทำหน้าที่ต่างกันสองข้อคือ ข้อที่ 7, 8, 9, 10 และข้อที่ 17, 18, 19, 20 ถ้าขนาดของการทำหน้าที่ต่างกัน .40 ดังนั้น ข้อคำถามแปดข้อนี้จะมีค่าเทรสโฮลแตกต่างกัน

โดยค่าเทรสโฮลของกลุ่มสนใจจะมีค่ามากกว่ากลุ่มอ้างอิงเท่ากับ .40

$(\tau_{i1R} = \tau_{i1F} - 0.40, \tau_{i2R} = \tau_{i2F} - 0.40, \tau_{i3R} = \tau_{i3F} - 0.40, \tau_{i4R} = \tau_{i4F} - 0.40)$

เมื่อ i = 7, 8, 9, 10, 17, 18, 19, 20

ถ้าขนาดของการทำหน้าที่ต่างกัน .70 ค่าเทรสโฮลกลุ่มสนใจจะมีค่ามากกว่ากลุ่มอ้างอิงเท่ากับ .70

$(\tau_{i1R} = \tau_{i1F} - 0.70, \tau_{i2R} = \tau_{i2F} - 0.70, \tau_{i3R} = \tau_{i3F} - 0.70, \tau_{i4R} = \tau_{i4F} - 0.70)$

เมื่อ i = 7, 8, 9, 10, 17, 18, 19, 20

และถ้าขนาดของการทำหน้าที่ต่างกัน 1.00 ค่าเทรสโฮลของกลุ่มสนใจจะมีค่ามากกว่ากลุ่มอ้างอิงเท่ากับ 1.00

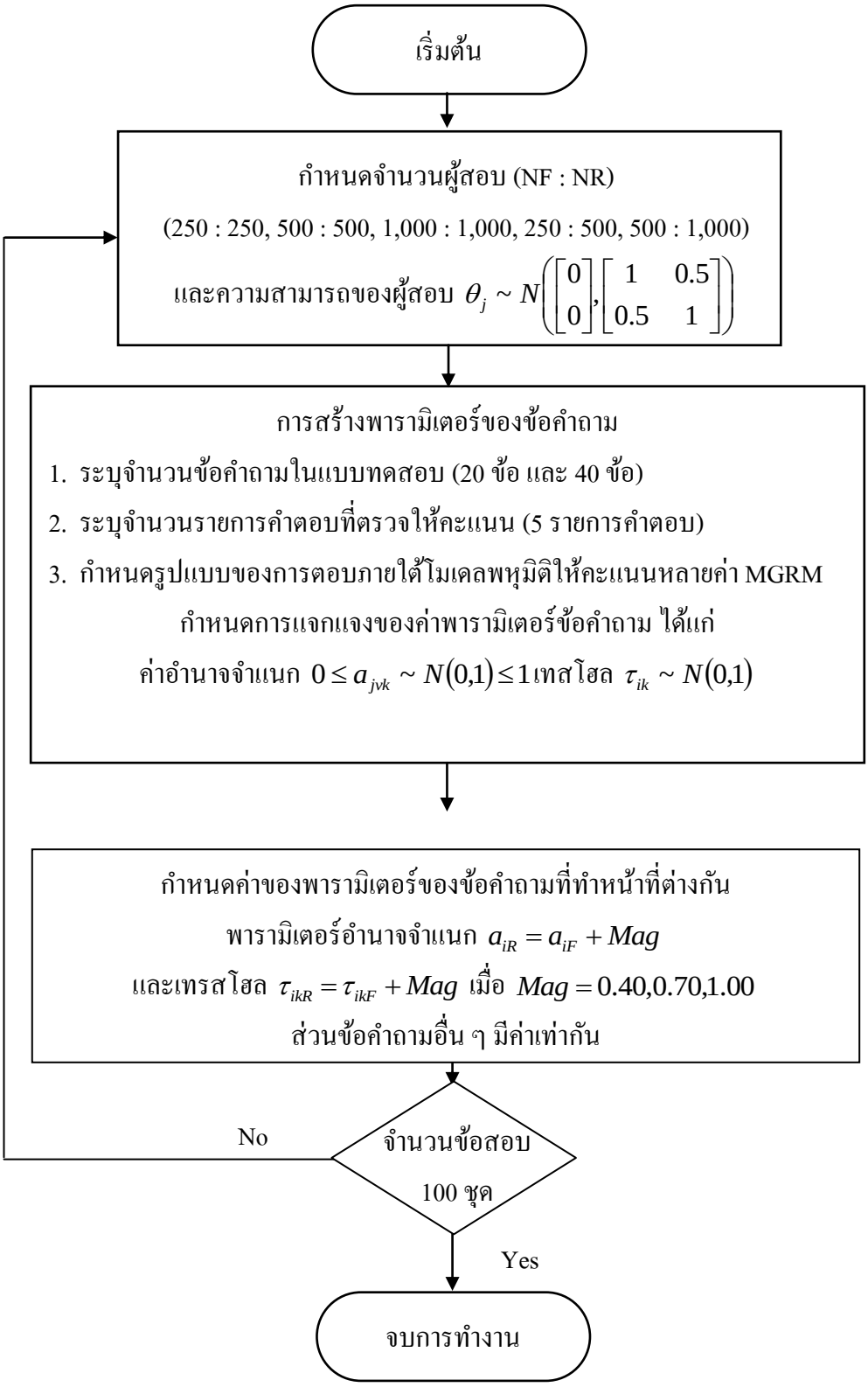
$(\tau_{i1R} = \tau_{i1F} - 1.00, \tau_{i2R} = \tau_{i2F} - 1.00, \tau_{i3R} = \tau_{i3F} - 1.00, \tau_{i4R} = \tau_{i4F} - 1.00)$

เมื่อ i = 7, 8, 9, 10, 17, 18, 19, 20

ส่วนข้ออื่น ๆ จะมีค่าเทรสโฮลเท่ากันทั้งสองกลุ่ม

ขั้นที่ 3 การสร้างข้อมูลการตอบข้อคำถาม

ในขั้นตอนนี้เป็นการนำเซตข้อมูลที่เกิดจากการจำลองในขั้นตอนที่ 1 และ 2 มาจัดทำข้อมูลรายการคำตอบของกลุ่มตัวอย่างที่จะทำการศึกษาโดยใช้โมเดลเกรตเรสพอนแบบพหุมิติในแต่ละเงื่อนไขกระทำซ้ำ 100 ชุด ซึ่งจะได้ข้อมูลเป็นไปตามเงื่อนไขทั้งสิ้น 6,000 ชุด เงื่อนไขทั้งหมดแสดงภาพที่ 25



ภาพที่ 25 แผนผังของการจำลองข้อมูล

การวิเคราะห์ข้อมูล

1. การวิเคราะห์การทำหน้าที่ต่างกันของข้อคำถาม

สำหรับงานวิจัยนี้ ในแต่ละวิธีของการทดสอบการทำหน้าที่ต่างกันของข้อคำถามจะใช้สถิติที่แตกต่างกันในแต่ละวิธีการตรวจสอบ โดยวิธีการตรวจสอบด้วยวิธีโพลีโตมัสซิปเทสท์ใช้สถิติ B ซึ่งวิเคราะห์ด้วยโปรแกรม DIFPACK วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุใช้สถิติค่าความแตกต่างของไค-สแควร์แต่ละโมเดลวิเคราะห์ด้วยโปรแกรม Mplus และวิธีการทดสอบวอลด์ใช้ Chi-square ประมาณค่าเมทริกซ์ความแปรปรวนร่วม (Covariance matrix) ด้วยวิธี Supplemented expectation maximization (SEM) ด้วยโปรแกรม IRTPRO ดังนี้

1.1 วิธีโพลีโตมัสซิปเทสท์

ผู้วิจัยใช้โปรแกรม DIFPACK ซึ่งมีสูตรในการคำนวณ ดังนี้

$$\hat{\beta} = \sum_{k=0}^{n_H} P_k d_k$$

$$P_k = \frac{N_{Rk} + N_{Fk}}{N}$$

เมื่อ P_k แทนสัดส่วนของผู้สอบทั้งหมด (กลุ่มอ้างอิงและกลุ่มสนใจ) ซึ่งตอบแบบทดสอบเดียวกันที่ใช้จับคู่ $X_1, X_2, ..., X_n$ แล้วได้คะแนน $x = k \quad X_1, X_2, ..., X_n$

$$d_k = \bar{Y}_{Rk} - \bar{Y}_{FK}, \ k = 0, ..., n_H$$

เมื่อ $\bar{Y}_{Rk}$ และ $\bar{Y}_{FK}$ แทนค่าเฉลี่ยคะแนนสอบของกลุ่มอ้างอิงและกลุ่มสนใจภายใต้ข้อคำถามศึกษา ณ คะแนนรวมในแบบทดสอบ $x = k$ (ณ ที่คะแนนรวมเท่ากัน อนุมานว่ามีความสามารถเท่ากัน) ถ้าข้อคำถามศึกษาใดไม่มีคะแนนสังเกตได้แสดงว่า $d_k \approx 0$

สมมติฐานของการทดสอบดัชนี DIF กำหนด ดังนี้

$$H_0 : \beta = 0$$

$$H_0 : \beta > 0$$

นำดัชนี $\hat{\beta}$ มาทดสอบสมมติฐาน โดยใช้สถิติ B ดังนี้

$$B\frac{\hat{\beta}}{\hat{\sigma}\left(\hat{\beta}\right)}$$

โดยที่ $\hat{\sigma}\left(\hat{\beta}\right)=\sqrt{\sum_{k=0}^{n_H}P_k^2\left[\frac{\hat{\sigma}^2\left(Y|k,R\right)}{N_{Rk}}+\frac{\hat{\sigma}^2\left(Y|k,F\right)}{N_{Fk}}\right]}$

เมื่อ $\hat{\sigma}\left(\hat{\beta}\right)$ แทน ค่าประมาณความคลาดเคลื่อนมาตรฐานของ β

$\hat{\sigma}^2\left(Y|k,g\right)$ แทน ค่าประมาณความแปรปรวนของคะแนนจากข้อคำถามศึกษากลุ่ม g (R หรือ F) ซึ่งมีคะแนนรวมในแบบทดสอบเท่ากับ k

ผู้วิจัยใช้เกณฑ์ตัดสินข้อคำถามทำหน้าที่ต่างกันที่ระดับนัยสำคัญ .05 ถ้าผลการทดสอบพบว่า ง $|B|>Z_{1-\frac{\alpha}{2}}$ อย่างมีนัยสำคัญที่ระดับ .05 แสดงว่า ปฏิเสธ H_0 นั่นคือ ข้อคำถามที่นำมาตรวจทำหน้าที่ต่างกัน โดยเข้าข้างผู้สอบกลุ่มใดกลุ่มหนึ่ง

1.2 วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ

การใช้วิธีการวิเคราะห์แบบ MG-CFA ผู้วิจัยใช้โปรแกรม Mplus และใช้การประมาณค่าแบบ WLSMV เพื่อใช้สำหรับการประเมินความไม่แปรเปลี่ยนของพารามิเตอร์ของข้อคำถามระหว่างกลุ่ม ซึ่งมีขั้นตอนในการดำเนินการวิจัย โดยแบ่งกรณีของการวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ สำหรับการคำนวณค่าอำนาจการทดสอบ และความคลาดเคลื่อนประเภทที่ 1 ดังนี้

2.1.1 วิธีวิเคราะห์เพื่อหาอำนาจการทดสอบ

2.1.1.1 Baselinemodel: กำหนดให้ค่าพารามิเตอร์ Loading (λ) และ Threshold (τ) ของข้อคำถามทั้งหมดในกลุ่มสนใจและกลุ่มอ้างอิงให้มีค่าเท่ากัน

2.1.1.2 Model I กำหนดให้ค่าพารามิเตอร์ Loading (λ)และ Threshold (τ) ของข้อคำถามที่ไม่ทำหน้าที่ต่างกันทุกข้อ และข้อคำถามที่ทำหน้าที่ต่างกันหนึ่งข้อ ให้มีค่าเท่ากันระหว่างกลุ่ม ส่วนข้ออื่น ๆ อิสระกันระหว่างกลุ่ม

2.1.1.3 Model II กำหนดให้ค่าพารามิเตอร์ Loading (λ) และ Threshold (τ) ของข้อคำถามที่ไม่ทำหน้าที่ต่างกันทุกข้อ และข้อคำถามที่ทำหน้าที่ต่างกันหนึ่งข้อ (ต้องเป็นข้อที่ทำหน้าที่ต่างกันไม่ซ้ำกับ Model I) ให้มีค่าเท่ากันระหว่างกลุ่ม ส่วนข้ออื่น ๆ อิสระกันระหว่างกลุ่ม

2.1.1.4 Model III กำหนดให้ค่าพารามิเตอร์ Loading (λ) และ Threshold (τ) ของข้อคำถามที่ไม่ทำหน้าที่ต่างกันทุกข้อ และข้อคำถามที่ทำหน้าที่ต่างกันหนึ่งข้อ (ต้องเป็นข้อคำถามที่ทำหน้าที่ต่างกันไม่ซ้ำกับ Model I และ Model II) ให้มีค่าเท่ากันระหว่างกลุ่ม ส่วนข้ออื่น ๆ อิสระกันระหว่างกลุ่ม

1.2.1.5 บังคับค่าข้อที่ทำหน้าที่ต่างกันที่เหลือไปเรื่อย ๆ จนครบทุกข้อคำถามที่ทำหน้าที่ต่างกัน

1.2.2 วิธีวิเคราะห์เพื่อหาความคลาดเคลื่อนประเภทที่ 1

1.2.2.1 Baseline model: กำหนดให้ค่าพารามิเตอร์ Loading (λ) และ Threshold (τ) ของข้อคำถามที่ไม่ทำหน้าที่ต่างกันทุกข้อให้เท่ากันระหว่างกลุ่ม และข้อที่ทำหน้าที่ต่างกันอิสระระหว่างกลุ่ม

1.2.2.2 Model I กำหนดให้ค่าพารามิเตอร์ Loading (λ) และ Threshold (τ) ของข้อคำถามที่ไม่ทำหน้าที่ต่างกันหนึ่งข้อ และข้อที่ทำหน้าที่ต่างกันทุกข้ออิสระกันระหว่างกลุ่ม ส่วนข้อที่ไม่ทำหน้าที่ต่างกันกำหนดให้เท่ากันระหว่างกลุ่ม

1.2.2.3 Model II กำหนดให้ค่าพารามิเตอร์ Loading (λ) และ Threshold (τ) ของข้อคำถามที่ไม่ทำหน้าที่ต่างกันหนึ่งข้อ (ต้องเป็นข้อคำถามที่ไม่ซ้ำกับ Model I) และข้อที่ทำหน้าที่ต่างกันทุกข้ออิสระกันระหว่างกลุ่ม ส่วนข้อไม่ทำหน้าที่ต่างกันกำหนดให้เท่ากันระหว่างกลุ่ม

1.2.2.4 model III กำหนดให้ค่าพารามิเตอร์ Loading (λ) และ Threshold (τ) ของข้อคำถามที่ไม่ทำหน้าที่ต่างกันหนึ่งข้อ (ข้อคำถามไม่ซ้ำกับ Model I และ Model II) และข้อที่ทำหน้าที่ต่างกันทุกข้ออิสระกันระหว่างกลุ่ม ส่วนข้อไม่ทำหน้าที่ต่างกันกำหนดให้เท่ากันระหว่างกลุ่ม

1.2.2.5 สร้าง Model ต่อไปเรื่อย ๆ จนครบทุกข้อตามจำนวนข้อคำถามที่ไม่ทำหน้าที่ต่างกัน

การทดสอบการทำหน้าที่ต่างกันของข้อคำถาม ทดสอบด้วยค่าความแตกต่างกันของ Chi-square เป็นการทดสอบภายใต้การนำค่า Chi-square ที่คำนวณจากผลต่างระหว่าง Model i กับ Baseline model ถ้าความแตกต่างของค่า Chi-square พบนัยสำคัญ แสดงว่า พารามิเตอร์ของข้อคำถามมีค่าแปรเปลี่ยนระหว่างกลุ่ม หรือกล่าวว่าข้อคำถามทำหน้าที่ต่างกัน และในทางกลับกัน ถ้าความแตกต่างของค่า Chi-square ไม่พบนัยสำคัญ แสดงว่า พารามิเตอร์ของข้อคำถามมีค่าไม่แปรเปลี่ยนระหว่างกลุ่ม หรือกล่าวว่าข้อคำถามไม่ทำหน้าที่ต่างกัน

1.3 วิธีการทดสอบวอลด์

การใช้วิธีการวิเคราะห์แบบ Wald ผู้วิจัยใช้โปรแกรม IRTPRO และวิธีการประมาณค่าเมทริกซ์ความแปรปรวนร่วมด้วยวิธี Supplement expectation maximization (SEM) โดยมีขั้นตอนในการทดสอบ ดังนี้

สมมติฐานสำหรับทดสอบ

$$H_0 : \alpha_{f1} = \alpha_{r1} : \beta_{f1} = \beta_{r1} : \beta_{f2} = \beta_{f2} : \beta_{f3} = \beta_{r3} : \beta_{f4} = \beta_{r4}$$

สถิติทดสอบ

$$\chi^2 = \hat{\nu}\Sigma^{-1}\hat{\nu} \qquad \text{ที่ } df = 5$$

เมื่อ $\nu = \alpha_{f1} - \alpha_{r1} : \beta_{f1} - \beta_{r1} : \beta_{f2} - \beta_{f2} : \beta_{f3} - \beta_{r3} : \beta_{f4} - \beta_{r4}$ เป็นเวกเตอร์ผลต่างระหว่างพารามิเตอร์ ข้อคำถามทุกข้อ ระหว่างกลุ่มอ้างอิงและกลุ่มสนใจ และ Σ^{-1} เป็นเมทริกซ์ความแปรปรวน-ความแปรปรวนร่วมของความคลาดเคลื่อน

$$\Sigma^{-1} = \left(\Sigma_F + \Sigma_R\right)^{-1} \tag{54}$$

เมื่อ Σ_F แทน เมทริกซ์ความแปรปรวน-ความแปรปรวนร่วมของความคลาดเคลื่อนของกลุ่มสนใจ

Σ_R แทน เมทริกซ์ความแปรปรวน-ความแปรปรวนร่วมของความคลาดเคลื่อนของกลุ่มอ้างอิง

เมทริกซ์ความแปรปรวน-ความแปรปรวนร่วมของความคลาดเคลื่อนของกลุ่มสนใจเป็นดังสมการ

$$\Sigma_F = \begin{pmatrix} \text{var}(a_f) & \text{cov}(a_r, b_{f1}) & \text{cov}(a_f, b_{f2}) & \text{cov}(a_f, b_{f3}) & \text{cov}(a_f, b_{f4}) \\ \text{cov}(a_f, b_{f1}) & \text{var}(b_{f1}) & \text{cov}(b_{f1}, b_{f2}) & \text{cov}(b_{f1}, b_{f3}) & \text{cov}(b_{f1}, b_{f4}) \\ \text{cov}(a_f, b_{f2}) & \text{cov}(b_{f1}, b_{f2}) & \text{var}(b_{f2}) & \text{cov}(b_{f2}, b_{f3}) & \text{cov}(b_{f2}, b_{f4}) \\ \text{cov}(a_f, b_{f3}) & \text{cov}(b_{f1}, b_{f3}) & \text{cov}(b_{f2}, b_{f3}) & \text{var}(b_{f3}) & \text{cov}(b_{f3}, b_{f4}) \\ \text{cov}(a_f, b_{f4}) & \text{cov}(b_{f1}, b_{f4}) & \text{cov}(b_{f2}, b_{f4}) & \text{cov}(b_{f3}, b_{f4}) & \text{var}(b_{f4}) \end{pmatrix}_{5\times5}$$

เมทริกซ์ความแปรปรวน-ความแปรปรวนร่วมของความคลาดเคลื่อนของกลุ่มอ้างอิง เป็นดังสมการ

$$\Sigma_F = \begin{pmatrix} \text{var}(a_r) & \text{cov}(a_r, b_{r1}) & \text{cov}(a_r, b_{r2}) & \text{cov}(a_r, b_{r3}) & \text{cov}(a_r, b_{r4}) \\ \text{cov}(a_r, b_{r1}) & \text{var}(b_{r1}) & \text{cov}(b_{r1}, b_{r2}) & \text{cov}(b_{r1}, b_{r1}) & \text{cov}(b_{r1}, b_{r4}) \\ \text{cov}(a_r, b_{r2}) & \text{cov}(b_{r1}, b_{r2}) & \text{var}(b_{r2}) & \text{cov}(b_{r2}, b_{r3}) & \text{cov}(b_{r2}, b_{r4}) \\ \text{cov}(a_r, b_{r3}) & \text{cov}(b_{r1}, b_{r3}) & \text{cov}(b_{r2}, b_{r3}) & \text{var}(b_{r3}) & \text{cov}(b_{r3}, b_{r4}) \\ \text{cov}(a_r, b_{r4}) & \text{cov}(b_{r1}, b_{r4}) & \text{cov}(b_{r2}, b_{r4}) & \text{cov}(b_{r3}, b_{r4}) & \text{var}(b_{r4}) \end{pmatrix}_{5\times5}$$

2. วิธีวิเคราะห์ประสิทธิภาพของการทำหน้าที่ต่างกันของข้อคำถาม

2.1 การคำนวณอำนาจการทดสอบ

การคำนวณอำนาจการทดสอบ (Power of the test) ของวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ ใช้สูตรในการคำนวณ ดังนี้

$$P = \frac{n_1}{N_1}$$

เมื่อ P แทน อำนาจการทดสอบ

n_1 แทน จำนวนข้อคำถามที่ตรวจสอบได้ถูกต้องว่าทำหน้าที่ต่างกัน

N_1 แทน จำนวนข้อคำถามที่ทำหน้าที่ต่างกัน

2.2 อัตราความคลาดเคลื่อนประเภทที่ 1

อัตราความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ ใช้สูตรในการคำนวณ ดังนี้

$$E = \frac{n_2}{N_2}$$

เมื่อ E แทน อัตราความคลาดเคลื่อนประเภทที่ 1

n_2 แทน จำนวนข้อคำถามที่ตรวจสอบผิดพลาดว่าทำหน้าที่ต่างกัน

N_2 แทน จำนวนข้อคำถามที่ไม่ทำหน้าที่ต่างกัน

3. สถิติสำหรับทดสอบสมมติฐาน

ผู้วิจัยได้นำข้อมูลที่ได้จากการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามด้วยวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ ความยาวของแบบทดสอบ ขนาดของการทำหน้าที่ต่างกัน สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และ ขนาดตัวอย่าง คำนวณอัตราความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบ และนำผลที่ได้มาพิจารณาตามเกณฑ์ ต่อไปนี้

3.1 การทดสอบสมมติฐานอัตราความคลาดเคลื่อนประเภทที่ 1

เกณฑ์การพิจารณาหากมีค่าความคลาดเคลื่อนประเภทที่ 1 ต่ำกว่าหรือเท่ากับ .05 ถือว่าควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดี (Atar & Kamata, 2011) นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามตรวจไม่พบข้อคำถามที่ทำหน้าที่ต่างกันในข้อที่ไม่ทำหน้าที่ต่างกันได้จริง การทดสอบสมมติฐานความคลาดเคลื่อนประเภทที่ 1 ดังนี้

ตั้งสมมติฐานการทดสอบ

$$H_0 : P \leq .05$$
$$H_1 : P > .05$$

สถิติทดสอบ $$Z = \frac{\hat{P} - P}{\sqrt{\frac{P(1-P)}{n}}}$$

โดยที่ $\hat{P}$ แทน สัดส่วนการเกิดความคลาดเคลื่อนประเภทที่ 1 ของกลุ่มตัวอย่าง

P แทน สัดส่วนการเกิดความคลาดเคลื่อนประเภทที่ 1 ของประชากร

n แทน จำนวนของการทำซ้ำในการจำลองข้อมูล

การกำหนดขอบเขตวิกฤต

เนื่องจากเป็นการทดสอบสมมติฐานทางเดียว จึงนำ Z ที่คำนวณได้จากสูตรเทียบกับ $Z_{.05} = 1.645$ ถ้า Z ที่คำนวณได้น้อยกว่า $Z_{.05} = 1.645$ จะไม่สามารถปฏิเสธสมมติฐาน H_0 แสดงว่าวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามนั้นสามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 อยู่ในระดับที่ยอมรับได้

3.2 การทดสอบสมมติฐานอำนาจการทดสอบ

เกณฑ์ที่ใช้พิจารณาอำนาจการทดสอบ จะพิจารณาเมื่อวิธีการตรวจสอบการทำหน้าที่ต่างกันสามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ และมีอำนาจการทดสอบต้องมีค่าตั้งแต่

.80 ขึ้นไป จึงถือว่ามีอำนาจการทดสอบเพียงพอ (Sufficient power) หากต่ำกว่า .80 ถือว่าวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามนั้น ๆ เป็นวิธีการตรวจสอบที่ไม่ดี (Atar & Kamata, 2011) การทดสอบสมมติฐานอำนาจการทดสอบ ดังนี้

ตั้งสมมติฐานการทดสอบ

$$H_0 : P \geq .80$$
$$H_1 : P < .80$$

สถิติทดสอบ $$Z = \frac{\hat{P} - P}{\sqrt{\frac{P(1-P)}{n}}}$$

เมื่อ $\hat{P}$ แทน สัดส่วนอำนาจการทดสอบของกลุ่มตัวอย่าง

P แทน สัดส่วนอำนาจการทดสอบของประชากร

n แทน จำนวนของการทำซ้ำในการจำลองข้อมูล

การกำหนดขอบเขตวิกฤต

เนื่องจากเป็นการทดสอบสมมติฐานทางเดียว จึงนำ Z ที่คำนวณได้จากสูตรเทียบกับ $Z_{.05} = 1.645$ ถ้า Z ที่คำนวณได้มากกว่า $Z_{.05} = 1.645$ จะไม่สามารถปฏิเสธสมมติฐาน H_0 แสดงว่าวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามนั้นมีอำนาจการทดสอบอย่างมีนัยสำคัญทางสถิติที่ .05

3.3 การทดสอบประสิทธิภาพของวิธีการตรวจสอบการทำหน้าที่ต่างกัน

สถิติสำหรับการทดสอบความแตกต่างของประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามพหุมิติให้คะแนนหลายค่าทั้งสามวิธี ผู้วิจัยใช้การวิเคราะห์ความแปรปรวนแบบวัดซ้ำ ประเภทสี่ตัวแปรระหว่างกลุ่ม และหนึ่งตัวแปรภายในกลุ่มแบบวัดซ้ำ (Repeated-measures analysis: four between-subjects variables and one within-subjects variable) (Howell, 2010)

บทที่ 4
ผลการวิเคราะห์ข้อมูล

ในการวิจัยครั้งนี้ มีวัตถุประสงค์เพื่อศึกษาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามที่เหมาะสมกับเครื่องมือวัดที่มีการวัดความสามารถสองมิติและให้คะแนน 5 ค่า ด้วยวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันพหุ และวิธีการทดสอบแบบวอลด์ ภายใต้ปัจจัยที่แตกต่างกัน 4 ปัจจัย ได้แก่ ความยาวของแบบทดสอบ ขนาดของการทำหน้าที่ต่างกัน สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง โดยพิจารณาจากอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1

ข้อมูลที่ใช้ในการศึกษาครั้งนี้เป็นข้อมูลจำลองภายใต้โมเดลเกรตเรสพอนพหุมิติแบบสองมิติและให้คะแนน 5 ค่า ตามปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ ความยาวของแบบทดสอบ ขนาดของการทำหน้าที่ต่างกัน สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง จำนวน 60 เงื่อนไข (3x2x2x5) ในแต่ละเงื่อนไขจำลองข้อมูลซ้ำ 100 ครั้ง แล้วนำข้อมูลดังกล่าวมาตรวจสอบการทำหน้าที่ต่างกัน 3 วิธี สำหรับผลการวิเคราะห์ข้อมูลจำแนกออกเป็น 3 ตอน ดังนี้

ตอนที่ 1 ผลการวิเคราะห์อัตราความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในโมเดลพหุมิติให้คะแนนหลายค่า ด้วยวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ ขนาดของการทำหน้าที่ต่างกัน ความยาวของแบบทดสอบ สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง

ตอนที่ 2 ผลการวิเคราะห์ประสิทธิภาพการควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในโมเดลพหุมิติให้คะแนนหลายค่า ด้วยวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ ความยาวของแบบทดสอบ ขนาดของการทำหน้าที่ต่างกัน สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง

ตอนที่ 3 ผลการวิเคราะห์เปรียบเทียบความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในโมเดลพหุมิติให้คะแนนหลายค่าระหว่างวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ ความยาวของแบบทดสอบ ขนาดของการทำหน้าที่ต่างกัน สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง

ในการนำเสนอผลการวิเคราะห์ข้อมูล ได้กำหนดสัญลักษณ์ที่ใช้ในการวิจัย ดังนี้

MGRM แทน โมเดลเกรตเรสพอน (Multidimensional graded response model)

DIF แทน การทำหน้าที่ต่างกันของข้อคำถาม (Differential item function)

POLYS แทน วิธีการตรวจสอบการทำหน้าที่ต่างกันด้วยวิธีโพลีโตมัสซิปเทสท์

MG-CFA แทน วิธีการตรวจสอบการทำหน้าที่ต่างกันด้วยการวิเคราห์องค์ประกอบเชิงยืนยันกลุ่มพหุ

WALD แทน วิธีการตรวจสอบการทำหน้าที่ต่างกันด้วยวิธีการทดสอบวอลด์

MAGDIF แทน ขนาดของการทำหน้าที่ต่างกัน (Magnitude of DIF)

TESTLG แทน ความยาวของแบบทดสอบ (Test length)

PROPDIF แทน สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน (Proportion DIF)

SAMPSIZE แทน ขนาดของกลุ่มตัวอย่าง (Sample size; NF : NR)

ตอนที่ 1 ผลการวิเคราะห์อัตราความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในโมเดลพหุมิติให้คะแนนหลายค่า ด้วยวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ ขนาดของการทำหน้าที่ต่างกัน ความยาวของแบบทดสอบ สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง

ผู้วิจัยนำข้อมูลที่ได้จากการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามของวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันพหุ และวิธีการทดสอบแบบวอลด์ โดยนำผลการคำนวณแสดงในรูปอัตราความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบ ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ ความยาวของแบบทดสอบ ขนาดของการทำหน้าที่ต่างกัน สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง แสดงดังตารางที่ 1

ตารางที่ 1 อัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีการตรวจสอบ DIF 3 วิธี สำหรับความยาวของแบบทดสอบจำนวน 20 ข้อ จำแนกตามปัจจัยที่แปรเปลี่ยน 3 ปัจจัย

TESTLG	MAGDIF	PROPDIF	SAMPSIZE (NF : NR)	TYPE I EROR		
				POLYS	MG-CFA	WALD
20 ข้อ	.40	10%	250 : 250	0.095	0.064*	0.037*
			500 : 500	0.143	0.079*	0.037*
			1,000 : 1,000	0.166	0.084*	0.052*
			250 : 500	0.092	0.077*	0.043*
			500 : 1,000	0.132	0.093	0.057*
		20%	250 : 250	0.208	0.067*	0.045*
			500 : 500	0.325	0.081*	0.080*
			1,000 : 1,000	0.496	0.091	0.125
			250 : 500	0.250	0.074*	0.053*
			500 : 1,000	0.420	0.098	0.105
	.70	10%	250 : 250	0.186	0.077*	0.033*
			500 : 500	0.245	0.079*	0.063*
			1,000 : 1,000	0.351	0.079*	0.079*
			250 : 500	0.202	0.078*	0.046*
			500 : 1,000	0.289	0.082*	0.086
		20%	250 : 250	0.526	0.062*	0.089
			500 : 500	0.578	0.084*	0.153
			1,000 : 1,000	0.728	0.087	0.264
			250 : 500	0.528	0.069*	0.116
			500 : 1,000	0.673	0.081*	0.244
	1.00	10%	250 : 250	0.291	0.065*	0.053*
			500 : 500	0.373	0.067*	0.076*
			1,000 : 1,000	0.521	0.081*	0.119
			250 : 500	0.261	0.069*	0.073*
			500 : 1,000	0.435	0.082*	0.086

ตารางที่ 1 (ต่อ)

TESTLG	MAGDIF	PROPDIF	SAMPSIZE (NF : NR)	TYPE I EROR POLYS	MG-CFA	WALD
			250 : 250	0.623	0.100	0.161
			500 : 500	0.694	0.082*	0.284
		20%	1,000 : 1,000	0.814	0.093	0.336
			250 : 500	0.686	0.090	0.210
			500 : 1,000	0.782	0.084*	0.332

หมายเหตุ ตัวเลข* หมายถึง สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ณ ระดับนัยสำคัญ .05

จากตารางที่ 1 ภายใต้ปัจจัยความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขของการทำหน้าที่ต่างกันของข้อคำถาม 0.40, 0.70 และ 1.00 สัดส่วนของจำนวนข้อคำถามที่ทำหน้าที่ต่างกัน 10% และ 20% และขนาดตัวอย่าง สัดส่วน 1 : 1 คือ 250 : 250, 500 : 500 และ 1,000 : 1,000 และสัดส่วน 1 : 2 คือ 250 : 500 และ 500 : 1,000 พบว่า วิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 อยู่ในช่วง 0.095-0.814, 0.064-0.100 และ 0.037-0.336 หรือ คิดเป็นร้อยละ 9.5%-81.4%, 6.2%-10.0% และ 3.7%-33.6% ตามลำดับ

1. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 250 : 250 คน วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีการทดสอบวอลด์มีอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำกว่าวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ

2. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 500 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์

สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีการทดสอบวอลด์มีอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำกว่าวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ

3. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 1,000 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีการทดสอบวอลด์มีอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำกว่าวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ

4. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 250 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีการทดสอบวอลด์มีอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำกว่าวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ

5. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 500 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

6. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 250 : 250 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีการทดสอบวอลด์มีอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำกว่าวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ

7. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 500 :500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีการทดสอบวอลด์ มีอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำกว่าวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุเพียงเล็กน้อย

8. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 1,000 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีไม่สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

9. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 250 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีการทดสอบวอลด์มีอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำกว่าวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ

10. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 500 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีไม่สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

11. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 250 : 250 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

โดยวิธีการทดสอบวอลด์มีอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำกว่าวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ

12. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 500 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีการทดสอบวอลด์มีอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำกว่าวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ

13. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 1,000 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05 และมีอัตราความคลาดเคลื่อนประเภทที่ 1 เท่ากัน

14. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 250 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีการทดสอบวอลด์มีอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำกว่าวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ

15. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 500 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

16. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 250 : 250 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติ

ให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

17. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 500 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

18. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 1,000 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีไม่สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

19. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 250 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

20. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 500 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

21. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 250 : 250 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีการทดสอบวอลด์มีอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำกว่าวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ

22. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 500 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำกว่าวิธีการทดสอบวอลด์

23. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 1,000 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

24. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 250 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำกว่าวิธีการทดสอบวอลด์

25. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 500 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

26. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 250 : 250 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีไม่สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

27. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 500 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

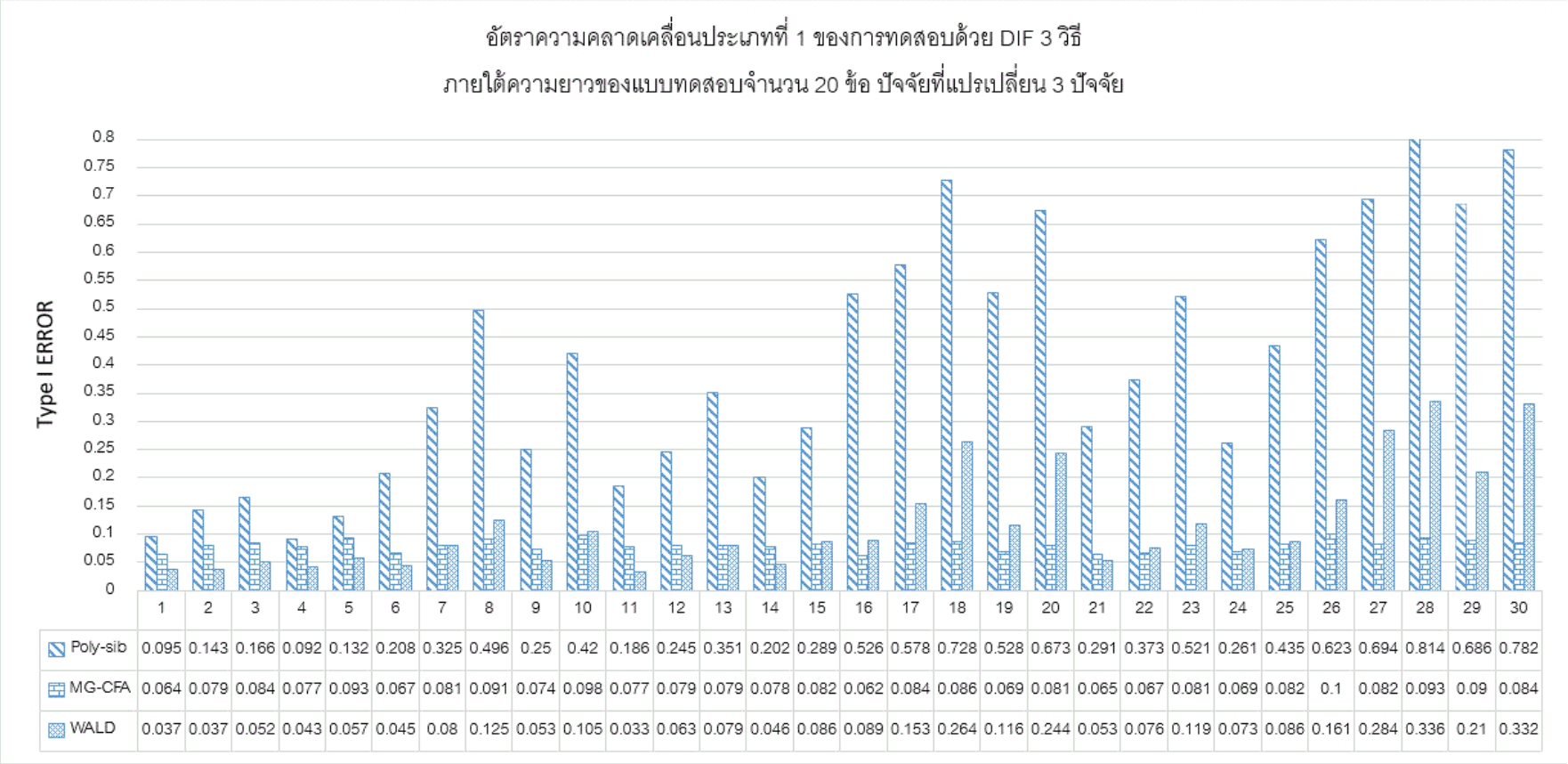
28. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 1,000 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีไม่สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

29. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 250 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธี ไม่สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

30. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 500 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

สรุปผลของการทดสอบอัตราความคลาดเคลื่อนประเภทที่ 1 ภายใต้ความยาวของแบบทดสอบ จำนวน 20 ข้อ พบว่า วิธีโพลีโตมัสซิปเทสท์ไม่สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ในทุกเงื่อนไข สำหรับกรณีที่ขนาดของการทำหน้าที่ต่างกันของข้อคำถามเท่ากับ 0.40 สัดส่วนของการทำหน้าที่ต่างกัน 10% วิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ดีกว่าวิธีการวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ ในทุกเงื่อนไขของขนาดตัวอย่าง แต่เมื่อขนาดของการทำหน้าที่ต่างกันมากขึ้น วิธีการวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุสามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ดีกว่าทั้งสองวิธี และเมื่อสัดส่วนของการทำหน้าที่ต่างกันเพิ่มขึ้นเป็น 20% วิธีการวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุสามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ดีกว่าวิธีการทดสอบวอลด์ และเมื่อขนาดของการทำหน้าที่ต่างกันเพิ่มขึ้น วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุสามารถควบคุมอัตรา

ความคลาดเคลื่อนประเภทที่ 1 ได้ดีกว่าวิธีการทดสอบวอลด์เกือบทุกเงื่อนไข ยกเว้นกรณีที่ขนาดของการทำหน้าที่ต่างกัน 0.40 สำหรับขนาดตัวอย่าง 1,000 : 1,000, 500 : 1,000 ขนาดของการทำหน้าที่ต่างกัน 0.70 สำหรับขนาดตัวอย่าง 1,000 : 1,000 และขนาดของการทำหน้าที่ต่างกัน 1.00 สำหรับขนาดตัวอย่าง 250 : 250, 1,000 : 1,000 และ 250 : 500 ทั้งสองวิธีไม่สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Poly-sib	0.095	0.143	0.166	0.092	0.132	0.208	0.325	0.496	0.25	0.42	0.186	0.245	0.351	0.202	0.289	0.526	0.578	0.728	0.528	0.673	0.291	0.373	0.521	0.261	0.435	0.623	0.694	0.814	0.686	0.782
MG-CFA	0.064	0.079	0.084	0.077	0.093	0.067	0.081	0.091	0.074	0.098	0.077	0.079	0.079	0.078	0.082	0.062	0.084	0.086	0.069	0.081	0.065	0.067	0.081	0.069	0.082	0.1	0.082	0.093	0.09	0.084
WALD	0.037	0.037	0.052	0.043	0.057	0.045	0.08	0.125	0.053	0.105	0.033	0.063	0.079	0.046	0.086	0.089	0.153	0.264	0.116	0.244	0.053	0.076	0.119	0.073	0.086	0.161	0.284	0.336	0.21	0.332

ภาพที่ 26 กราฟแสดงอัตราความคลาดเคลื่อนประเภทที่ 1 ด้วยวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ เมื่อความยาวของแบบทดสอบ จำนวน 20 ข้อ ภายใต้ปัจจัยที่แปรเปลี่ยน 3 ปัจจัย

ตารางที่ 2 อำนาจการทดสอบของวิธีการตรวจสอบ DIF 3 วิธี สำหรับความยาวของแบบทดสอบ จำนวน 20 ข้อ จำแนกตามปัจจัยที่แปรเปลี่ยน 3 ปัจจัย

TESTLG	MAGDIF	PROPDIF	SAMPSIZE (NF : NR)	POWER OF TEST		
				POLYS	MG-CFA	WALD
20 ข้อ	0.40	10%	250 : 250	0.865*	0.995*	0.905*
			500 : 500	0.990*	1.000*	1.000*
			1,000 : 1,000	1.000*	1.000*	1.000*
			250 : 500	0.980*	1.000*	0.975*
			500 : 1,000	1.000*	1.000*	1.000*
		20%	250 : 250	0.825*	0.980*	0.895*
			500 : 500	0.963*	1.000*	0.975*
			1,000 : 1,000	0.995*	1.000*	0.992*
			250 : 500	0.938*	0.998*	0.967*
			500 : 1,000	0.993*	1.000*	0.992*
	0.70	10%	250 : 250	1.000*	1.000*	1.000*
			500 : 500	1.000*	1.000*	1.000*
			1,000 : 1,000	1.000*	1.000*	1.000*
			250 : 500	1.000*	1.000*	1.000*
			500 : 1,000	1.000*	1.000*	1.000*
		20%	250 : 250	0.965*	0.988*	0.967*
			500 : 500	1.000*	1.000*	0.998*
			1,000 : 1,000	1.000*	1.000*	0.995*
			250 : 500	0.988*	1.000*	0.992*
			500 : 1,000	0.998*	1.000*	0.995*
	1.00	10%	250 : 250	1.000*	1.000*	1.000*
			500 : 500	1.000*	1.000*	1.000*
			1,000 : 1,000	1.000*	1.000*	1.000*
			250 : 500	0.995*	1.000*	1.000*
			500 : 1,000	1.000*	1.000*	1.000*

ตารางที่ 2 (ต่อ)

TESTLG	MAGDIF	PROPDIF	SAMPSIZE (NF : NR)	POWER OF TEST		
				POLYS	MG-CFA	WALD
			250 : 250	0.940*	0.983*	0.985*
			500 : 500	0.993*	1.000*	0.990*
		20%	1,000 : 1,000	0.997*	1.000*	0.990*
			250 : 500	0.998*	1.000*	0.995*
			500 : 1,000	0.995*	0.998*	0.997*

หมายเหตุ ตัวเลข* หมายถึง มีอำนาจการทดสอบตามเกณฑ์ที่กำหนด ณ ระดับนัยสำคัญ .05

จากตารางที่ 2 ภายใต้ปัจจัยความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขของการทำหน้าที่ต่างกันของข้อคำถาม 0.40, 0.70 และ 1.00 สัดส่วนของจำนวนข้อคำถามที่ทำหน้าที่ต่างกัน 10% และ 20% และขนาดตัวอย่าง สัดส่วน 1 : 1 คือ 250 : 250, 500 : 500 และ 1,000 : 1,000 และสัดส่วน 1 : 2 คือ 250 : 500 และ 500 : 1,000 พบว่า วิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ มีค่าเฉลี่ยของอำนาจการทดสอบอยู่ในช่วง 0.825-1.000, 0.980-1.000 และ 0.895-1.000 หรือคิดเป็นร้อยละ 82.5%-100.0%, 98.0%-100.0% และ 89.5%-100.0% ตามลำดับ

1. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 250 : 250 คน วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ มีอำนาจการทดสอบสูงที่สุด

2. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 500 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญที่ .05 โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์มีอำนาจการทดสอบสูงที่สุดถึง 1.00

3. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 1,000 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยทั้งสามวิธีมีอำนาจการทดสอบสูงถึง 1.00

4. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 250 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดย วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ มีอำนาจการทดสอบสูงสุด

5. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 500 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยทั้งสามมีอำนาจการทดสอบสูงถึง 1.00

6. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 250 : 250 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดย วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ มีอำนาจการทดสอบสูงสุด

7. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 500 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ มีอำนาจการทดสอบสูงสุด

8. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 1,000 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ มีอำนาจการทดสอบสูงสุด

9. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 250 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดย วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ มีอำนาจการทดสอบสูงสุด

10. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 500 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดย วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ มีอำนาจการทดสอบสูงสุด

11. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 250 : 250 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยทั้งสามวิธีมีอำนาจการทดสอบสูงถึง 1.00

12. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 500 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยทั้งสามวิธีมีอำนาจการทดสอบสูงถึง 1.00

13. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 1,000 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยทั้งสามวิธีมีอำนาจการทดสอบสูงถึง 1.00

14. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 250 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยทั้งสามวิธีมีอำนาจการทดสอบสูงถึง 1.00

15. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 500 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยทั้งสามวิธีมีอำนาจการทดสอบสูงถึง 1.00

16. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 250 : 250 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธี มีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ มีอำนาจการทดสอบสูงสุด

17. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 500 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธี มีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดย วิธีโพลีโตมัสซิปเทสท์ และวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ มีอำนาจการทดสอบสูงถึง 1.00

18. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 1,000 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีโพลีโตมัสซิปเทสท์ และวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ มีอำนาจการทดสอบสูงถึง 1.00

19. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 250 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ มีอำนาจการทดสอบสูงสุด

20. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 500 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติ

ให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ มีอำนาจการทดสอบสูงสุด

21. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 250 : 250 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธี มีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยทั้งสามวิธีมีอำนาจการทดสอบสูงถึง 1.00

22. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 500 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธี มีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยทั้งสามวิธีมีอำนาจการทดสอบสูงถึง 1.00

23. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 1,000 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยทั้งสามวิธีมีอำนาจการทดสอบสูงถึง 1.00

24. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 250 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธี มีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05

25. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 500 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธี มีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยทั้งสามวิธีมีอำนาจการทดสอบสูงถึง 1.00

26. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 250 : 250 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติ

ให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีการทดสอบวอลด์มีอำนาจการทดสอบสูงสุด

27. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 500 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ มีอำนาจการทดสอบสูงสุด

28. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 1,000 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ มีอำนาจการทดสอบสูงสุด

29. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 250 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ มีอำนาจการทดสอบสูงสุด

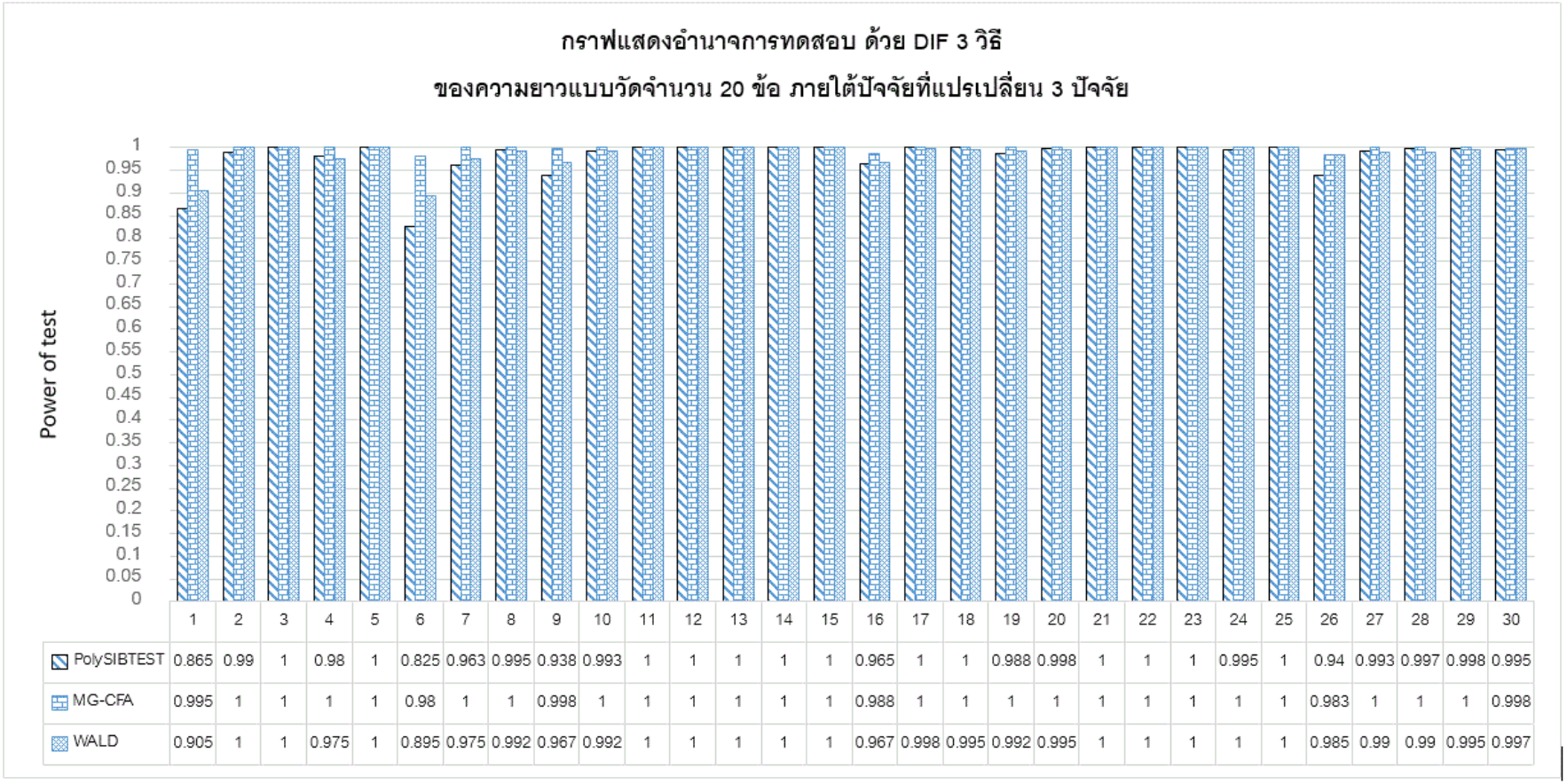
30. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 20 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 500 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธี มีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05

สรุปผลภายใต้ความยาวของแบบทดสอบจำนวน 20 ข้อ พบว่า วิธีการตรวจสอบการทำหน้าที่ต่างกันทั้งสามวิธีมีอำนาจการทดสอบเป็นไปตามเกณฑ์ที่กำหนด ณ ระดับนัยสำคัญที่ .05 ในทุกเงื่อนไข โดยวิธีการวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุมีอำนาจการทดสอบสุงสุด รองลงมาคือ วิธีการทดสอบวอลด์ และวิธีโพลีโตมัสซิปเทสท์

สำหรับกรณีที่ขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 10% ขนาดตัวอย่างเท่ากับ 1,000 : 1,000 และ 500 : 1,000 ทั้งสามวิธีมีอำนาจการทดสอบในระดับที่เท่ากัน ส่วนกรณีขนาดตัวอย่าง 500 : 500 ที่สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดตัวอย่าง 1,000 : 1,000 ที่สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 20% ที่พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์มีอำนาจการทดสอบในระดับที่เท่ากัน

สำหรับกรณีขนาดของการทำหน้าที่ต่างกัน 0.70 ที่สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 10% ทุกขนาดตัวอย่าง พบว่า ทั้งสามวิธีมีอำนาจการทดสอบในระดับที่เท่ากัน ส่วนกรณีสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 20% ที่ขนาดตัวอย่าง 500 : 500 และ 1,000 : 1,000 พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ มีอำนาจการทดสอบในระดับที่เท่ากัน ส่วนกรณีอื่น ๆ วิธีการวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุมีอำนาจการทดสอบมากที่สุด รองลงมาคือ วิธีการทดสอบวอลด์ และวิธีโพลีโตมัสซิปเทสท์

สำหรับกรณีขนาดของการทำหน้าที่ต่างกัน 1.00 พบว่า เมื่อสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 10% ขนาดตัวอย่าง 250 : 250, 500 : 500, 1,000 : 1,000 และ 500 : 1,000 พบว่า วิธีการตรวจสอบการทำหน้าที่ต่างกันทั้งสามวิธีมีอำนาจการทดสอบในระดับที่เท่ากัน ส่วนในกรณีอื่น ๆ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ มีอำนาจการทดสอบสูงสุด รองลงมาวิธีการทดสอบวอลด์ และวิธีโพลีโตมัสซิปเทสท์ โดยทั้งสามวิธีมีค่าอำนาจการทดสอบใกล้เคียงกัน



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
PolySIBTEST	0.865	0.99	1	0.98	1	0.825	0.963	0.995	0.938	0.993	1	1	1	1	1	0.965	1	1	0.988	0.998	1	1	1	0.995	1	0.94	0.993	0.997	0.998	0.995
MG-CFA	0.995	1	1	1	1	0.98	1	1	0.998	1	1	1	1	1	1	0.988	1	1	1	1	1	1	1	1	1	0.983	1	1	1	0.998
WALD	0.905	1	1	0.975	1	0.895	0.975	0.992	0.967	0.992	1	1	1	1	1	0.967	0.998	0.995	0.992	0.995	1	1	1	1	1	0.985	0.99	0.99	0.995	0.997

ภาพที่ 27 กราฟแสดงอำนาจการทดสอบ ด้วยวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ เมื่อความยาวของแบบทดสอบ จำนวน 20 ข้อ ภายใต้ปัจจัยที่แปรเปลี่ยน 3 ปัจจัย

ตารางที่ 3 อัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีการตรวจสอบ DIF 3 วิธี สำหรับความยาวของแบบทดสอบจำนวน 40 ข้อ จำแนกตามปัจจัยที่แปรเปลี่ยน 3 ปัจจัย

TESTLG	MAGDIF	PERDIF	SAMPSIZE (NF : NR)	TYPE I ERROR		
				POLYS	MG-CFA	WALD
40 ข้อ	0.40	10%	250 : 250	0.083*	0.089	0.038*
			500 : 500	0.116	0.075*	0.044*
			1,000 : 1,000	0.165	0.096	0.052*
			250 : 500	0.106	0.071*	0.044*
			500 : 1,000	0.123	0.076*	0.055*
		20%	250 : 250	0.183	0.079*	0.058*
			500 : 500	0.257	0.116	0.077*
			1,000 : 1,000	0.423	0.116	0.121
			250 : 500	0.185	0.090	0.061*
			500 : 1,000	0.368	0.114	0.112
	0.70	10%	250 : 250	0.128	0.072*	0.043*
			500 : 500	0.193	0.108	0.044*
			1,000 : 1,000	0.329	0.087	0.076*
			250 : 500	0.150	0.074*	0.049*
			500 : 1,000	0.253	0.083*	0.055*
		20%	250 : 250	0.329	0.091	0.057*
			500 : 500	0.497	0.118	0.144
			1,000 : 1,000	0.667	0.133	0.190
			250 : 500	0.416	0.092	0.117
			500 : 1,000	0.595	0.126	0.195
	1.00	10%	250 : 250	0.173	0.061*	0.044*
			500 : 500	0.309	0.075*	0.048*
			1,000 : 1,000	0.477	0.106	0.078*
			250 : 500	0.226	0.071*	0.054*
			500 : 1,000	0.369	0.089	0.074*

ตารางที่ 3 (ต่อ)

TESTLG	MAGDIF	PERDIF	SAMPSIZE (NF : NR)	TYPE I ERROR		
				POLYS	MG-CFA	WALD
			250 : 250	0.473	0.092	0.087
			500 : 500	0.650	0.103	0.160
		20%	1,000 : 1,000	0.766	0.139	0.207
			250 : 500	0.562	0.093	0.124
			500 : 1,000	0.733	0.116	0.244

หมายเหตุ ตัวเลข* หมายถึง สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ณ ระดับนัยสำคัญ .05

จากตารางที่ 3 ภายใต้ปัจจัยความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขของการทำหน้าที่ต่างกันของข้อคำถาม 0.40, 0.70 และ 1.00 สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 10% และ 20% และขนาดตัวอย่าง สัดส่วน 1 : 1 คือ 250 : 250, 500 : 500 และ 1,000 : 1,000 และสัดส่วน 1 : 2 คือ 250 : 500 และ 500 : 1,000 พบว่า วิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 อยู่ในช่วง 0.083-0.766, 0.061-0.139 และ 0.038-0.244 หรือคิดเป็นร้อยละ 8.3%-76.6%, 6.1%-13.9% และ 3.8%-24.4% ตามลำดับ

1. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 250 : 250 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีโพลีโตมัสซิปเทสท์ และวิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีการทดสอบวอลด์มีค่าอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำสุด

2. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 500 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีการทดสอบวอลด์มีอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำสุด

3. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 1,000 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

4. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 250 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

5. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 500 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีการทดสอบวอลด์มีอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำสุด

6. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 250 : 250 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีการทดสอบวอลด์มีอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำสุด

7. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 500 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

8. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 1,000 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติ

ให้คะแนนหลายค่า พบว่า ทั้งสามวิธี ไม่สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

9. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และ ขนาดกลุ่มตัวอย่างขนาด 250 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

10. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 500 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีไม่สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

11. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 250 : 250 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีการทดสอบวอลด์มีอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำสุด

12. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 500 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

13. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 1,000 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

14. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่าง

ขนาด 250 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีการทดสอบวอลด์มีอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำสุด

15. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 500 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีการทดสอบวอลด์มีอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำสุด

16. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 250 : 250 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

17. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 500 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีไม่สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

18. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 1,000 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีไม่สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

19. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 250 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีไม่สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

20. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 500 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีไม่สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

21. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 250 : 250 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีการทดสอบวอลด์มีอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำสุด

22. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 500 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีการทดสอบวอลด์มีอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำสุด

23. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 1,000 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

24. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 250 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีการทดสอบวอดล์มีอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำสุด

25. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่าง

ขนาด 500 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

26. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 250 : 250 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีไม่สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

27. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 500 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีไม่สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

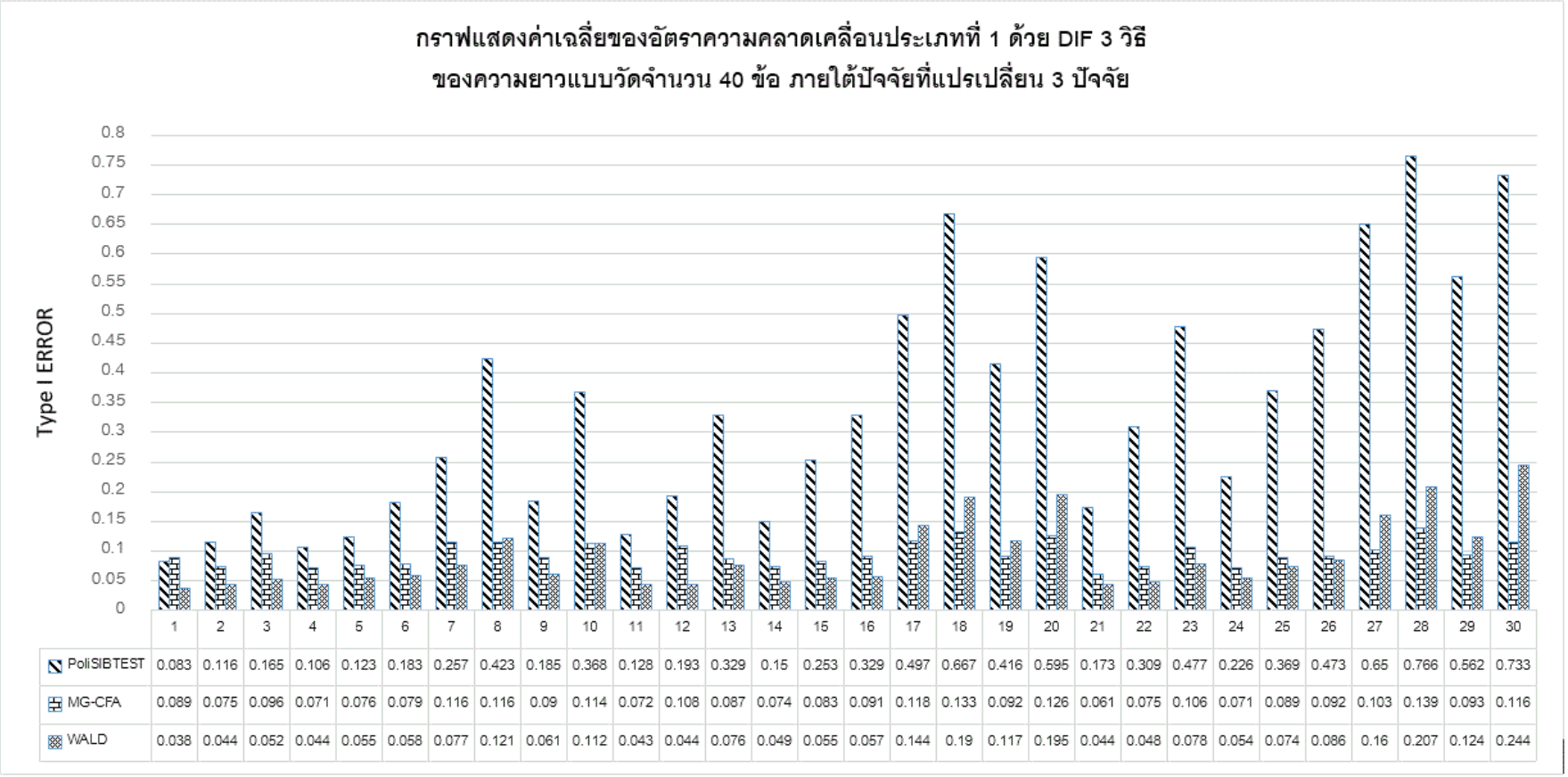
28. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 1,000 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธี ไม่สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

29. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 250 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีไม่สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

30. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 500 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีไม่สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05

สรุปผลของการทดสอบอัตราความคลาดเคลื่อนประเภทที่ 1 ภายใต้ความยาวของแบบทดสอบ จำนวน 40 ข้อ พบว่า วิธีโพลีโตมัสซิปเทสท์ไม่สามารถควบคุมอัตราความคลาดเคลื่อน

ประเภทที่ 1 ได้เกือบทุกเงื่อนไข ยกเว้นกรณีขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 10% ขนาดตัวอย่าง 250 : 250 สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05 สำหรับวิธีการทดสอบวอลด์ทุกขนาดของการทำหน้าที่ต่างกัน สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 10% ทุกขนาดตัวอย่าง สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ ณ ระดับนัยสำคัญ .05 แต่เมื่อสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกันเป็น 20% เมื่อขนาดของการทำหน้าที่ต่างกันเพิ่มขึ้น วิธีการทดสอบวอลด์ไม่สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ ส่วนวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ พบว่า ที่ขนาดการทำหน้าที่ต่างกันต่ำ สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 10% ขนาดตัวอย่างไม่เกิน 500 ของทั้งกลุ่มอ้างอิงและกลุ่มสนใจ สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ ส่วนกรณีอื่น ๆ ไม่สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
PoliSIBTEST	0.083	0.116	0.165	0.106	0.123	0.183	0.257	0.423	0.185	0.368	0.128	0.193	0.329	0.15	0.253	0.329	0.497	0.667	0.416	0.595	0.173	0.309	0.477	0.226	0.369	0.473	0.65	0.766	0.562	0.733
MG-CFA	0.089	0.075	0.096	0.071	0.076	0.079	0.116	0.116	0.09	0.114	0.072	0.108	0.087	0.074	0.083	0.091	0.118	0.133	0.092	0.126	0.061	0.075	0.106	0.071	0.089	0.092	0.103	0.139	0.093	0.116
WALD	0.038	0.044	0.052	0.044	0.055	0.058	0.077	0.121	0.061	0.112	0.043	0.044	0.076	0.049	0.055	0.057	0.144	0.19	0.117	0.195	0.044	0.048	0.078	0.054	0.074	0.086	0.16	0.207	0.124	0.244

ภาพที่ 28 กราฟแสดงอัตราความคลาดเคลื่อนประเภทที่ 1 ด้วยวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ เมื่อความยาวของแบบทดสอบ จำนวน 40 ข้อ ภายใต้ปัจจัยที่แปรเปลี่ยน 3 ปัจจัย

ตารางที่ 4 อำนาจการทดสอบของวิธีการตรวจสอบ DIF 3 วิธี สำหรับความยาวของแบบทดสอบ จำนวน 40 ข้อ จำแนกตามปัจจัยที่แปรเปลี่ยน 3 ปัจจัย

TESTLG	MAGDIF	PERDIF	SAMPSIZE (NF : NR)	POWERTEST		
				POLYS	MG-CFA	WALD
40 ข้อ	0.40	10%	250 : 250	0.815*	0.995*	0.958*
			500 : 500	0.998*	1.000*	1.000*
			1,000 : 1,000	1.000*	1.000*	1.000*
			250 : 500	0.965*	0.998*	0.995*
			500 : 1,000	1.000*	1.000*	1.000*
		20%	250 : 250	0.885*	0.985*	0.946*
			500 : 500	0.975*	1.000*	0.998*
			1,000 : 1,000	0.996*	1.000*	1.000*
			250 : 500	0.883*	0.998*	0.975*
			500 : 1,000	0.990*	1.000*	0.998*
	0.70	10%	250 : 250	0.998*	1.000*	1.000*
			500 : 500	0.993*	1.000*	1.000*
			1,000 : 1,000	1.000*	1.000*	1.000*
			250 : 500	0.998*	1.000*	1.000*
			500 : 1,000	0.995*	1.000*	1.000*
		20%	250 : 250	0.984*	1.000*	0.998*
			500 : 500	0.999*	1.000*	0.997*
			1,000 : 1,000	1.000*	1.000*	1.000*
			250 : 500	1.000*	1.000*	1.000*
			500 : 1,000	0.999*	1.000*	0.997*
	1.00	10%	250 : 250	1.000*	1.000*	1.000*
			500 : 500	0.998*	1.000*	1.000*
			1,000 : 1,000	1.000*	1.000*	1.000*
			250 : 500	0.993*	1.000*	1.000*
			500 : 1,000	0.998*	1.000*	1.000*

ตารางที่ 4 (ต่อ)

TESTLG	MAGDIF	PERDIF	SAMPSIZE (NF : NR)	POWERTEST POLYS	MG-CFA	WALD
			250 : 250	0.999*	1.000*	1.000*
			500 : 500	0.990*	1.000*	1.000*
		20%	1,000 : 1,000	1.000*	1.000*	1.000*
			250 : 500	0.998*	1.000*	1.000*
			500 : 1,000	1.000*	1.000*	1.000*

หมายเหตุ ตัวเลข* หมายถึง มีอำนาจการทดสอบตามเกณฑ์ที่กำหนด ณ ระดับนัยสำคัญ .05

จากตารางที่ 4 ภายใต้ปัจจัยความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขของการทำหน้าที่ต่างกันของข้อคำถาม 0.40, 0.70 และ 1.00 สัดส่วนของจำนวนข้อคำถามที่ทำหน้าที่ต่างกัน 10% และ 20% และขนาดตัวอย่าง สัดส่วน 1 : 1 คือ 250 : 250, 500 : 500 และ 1,000 : 1,000 และสัดส่วน 1 : 2 คือ 250 : 500 และ 500 : 1,000 พบว่า วิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ มีค่าเฉลี่ยของอำนาจการทดสอบ อยู่ในช่วง 0.815-1.000, 0.985-1.000 และ 0.946-1.000หรือคิดเป็นร้อยละ 81.5%-100.0%, 98.5.0%-100.0% และ 94.6%-100.0% ตามลำดับ

1. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 250 : 250 คน วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ มีอำนาจการทดสอบสูงสุด

2. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 500 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์มีอำนาจการทดสอบสูงสุดในระดับที่เท่ากัน

3. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 1,000 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยทุกวิธีมีอำนาจการทดสอบสูงในระดับที่เท่ากัน

4. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 250 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุมีอำนาจการทดสอบสูงสุด

5. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 500 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยทุกวิธีมีอำนาจการทดสอบสูงสุดในระดับที่เท่ากัน

6. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 250 : 250 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธี มีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ มีอำนาจการทดสอบสูงสุด

7. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 500 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธี มีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุสูงสุด

8. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 1,000 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05

โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุสูงสุดและวิธีการทดสอบวอลด์ มีอำนาจการทดสอบสูงสุดในระดับที่เท่ากัน

9. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 250 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุมีอำนาจการทดสอบสูงสุด

10. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 500 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธี มีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุมีอำนาจการทดสอบสูงสุด

11. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 250 : 250 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธี มีอำนาจการทดสอบเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ มีอำนาจการทดสอบสูงในระดับที่เท่ากัน

12. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 500 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธี มีอำนาจการทดสอบเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ มีอำนาจการทดสอบสูงในระดับที่เท่ากัน

13. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 1,000 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธี มีอำนาจการทดสอบเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ มีอำนาจการทดสอบสูงในระดับที่เท่ากัน

14. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 250 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธี มีอำนาจการทดสอบเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์มีอำนาจการทดสอบสูงในระดับที่เท่ากัน

15. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 500 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธี มีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์มีอำนาจการทดสอบสูงในระดับที่เท่ากัน

16. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 250 : 250 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธี มีอำนาจการทดสอบเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุมีอำนาจการทดสอบสูงสุด

17. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 500 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธี มีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ มีอำนาจการทดสอบสูงสุด

18. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 1,000 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธี มีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยทั้งสามวิธี มีอำนาจการทดสอบสูงสุดในระดับที่เท่ากัน

19. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 250 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติ

ให้คะแนนหลายค่า พบว่า ทั้งสามวิธี มีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยทั้งสามวิธีมีอำนาจในการทดสอบสูงในระดับที่เท่ากัน

20. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 0.70 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 500 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธี มีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุมีอำนาจการทดสอบสูงสุด

21. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 250 : 250 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธี มีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยทั้งสามวิธีมีอำนาจการทดสอบสูงในระดับที่เท่ากัน

22. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 500 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธี มีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ มีอำนาจการทดสอบสูงในระดับที่เท่ากัน

23. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 1,000 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธี มีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยทั้งสามวิธีมีอำนาจการทดสอบสูงในระดับที่เท่ากัน

24. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 250 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธี มีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ มีอำนาจการทดสอบสูงในระดับที่เท่ากัน

25. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 10% และขนาดกลุ่มตัวอย่างขนาด 500 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธี มีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ มีอำนาจการทดสอบสูงในระดับที่เท่ากัน

26. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 250 : 250 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธี มีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์มีอำนาจการทดสอบสูงสุดในระดับที่เท่ากัน

27. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 500 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธี มีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ มีอำนาจการทดสอบสูงในระดับที่เท่ากัน

28. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 1,000 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธีมีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยทั้งสามวิธีมีอำนาจการทดสอบสูงในระดับที่เท่ากัน

29. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 250 : 500 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธี มีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ มีอำนาจการทดสอบสูงในระดับที่เท่ากัน

30. เมื่อพิจารณาข้อมูลที่มีความยาวของแบบทดสอบจำนวน 40 ข้อ ในเงื่อนไขขนาดของการทำหน้าที่ต่างกัน 1.00 สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน 20% และขนาดกลุ่มตัวอย่างขนาด 500 : 1,000 คน ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า ทั้งสามวิธี มีอำนาจการทดสอบสูงเป็นไปตามเกณฑ์ ณ ระดับนัยสำคัญ .05 โดยทั้งสามวิธีมีอำนาจการทดสอบในระดับที่เท่ากัน

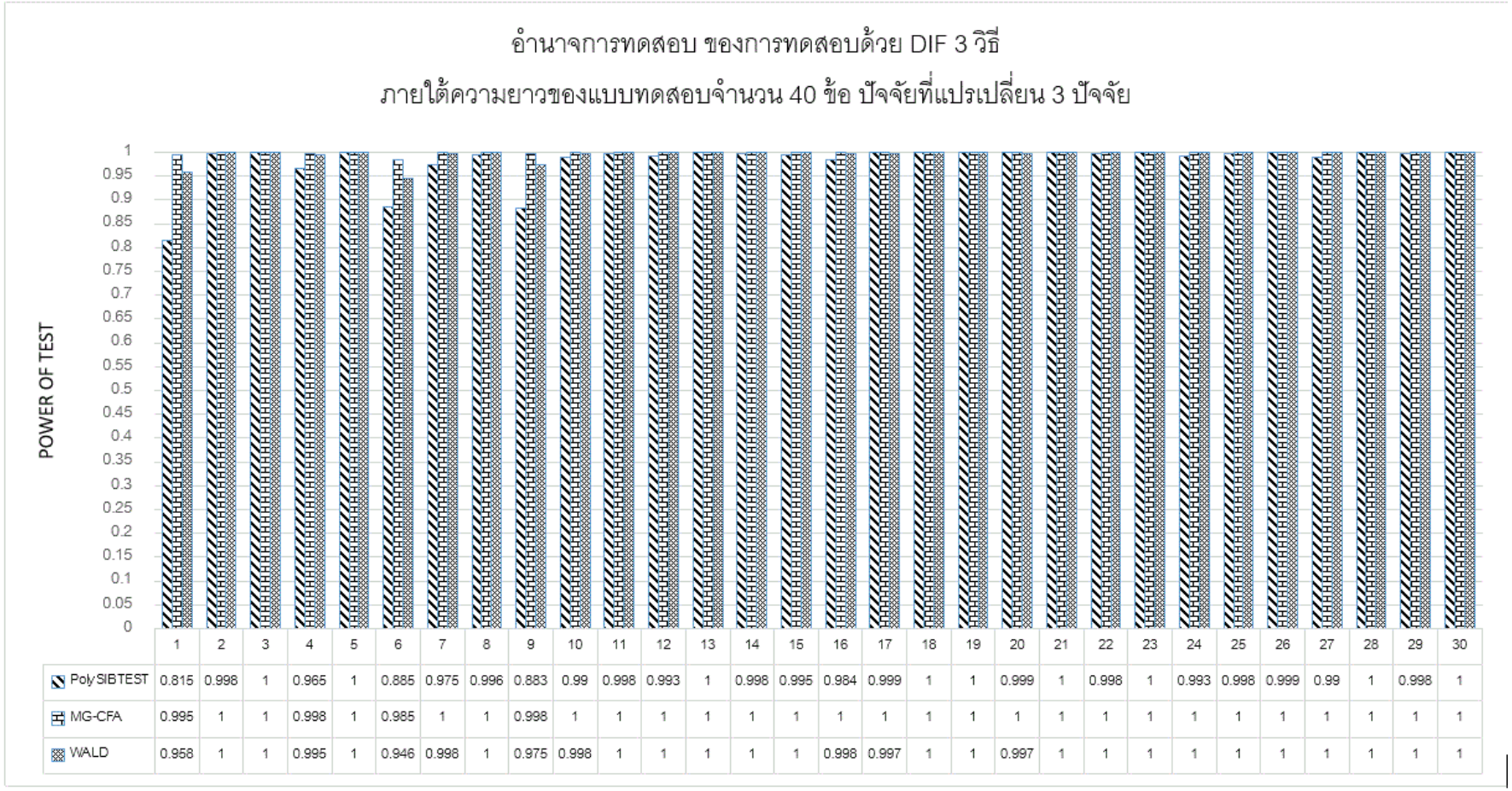
สรุปผลของการวิเคราะห์อำนาจการทดสอบ ภายใต้ความยาวของแบบทดสอบ จำนวน 40 ข้อ พบว่า วิธีการตรวจสอบการทำหน้าที่ต่างกันทั้งสามวิธี มีอำนาจการทดสอบเป็นไปตามเกณฑ์ที่กำหนด ณ ระดับนัยสำคัญที่ .05 ในทุกเงื่อนไข

สำหรับกรณีที่ขนาดของการทำหน้าที่ต่างกัน 0.40 สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 10% ขนาดตัวอย่าง 1,000 : 1,000 และ 500 : 1,000 พบว่า วิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ มีอำนาจการทดสอบในระดับที่เท่ากัน กรณีขนาดตัวอย่าง 500 : 500 ของสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 10% และกรณีขนาดตัวอย่าง 1,000 : 1,000 ของสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 20% วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ มีอำนาจการทดสอบในระดับที่เท่ากัน และมากกว่าวิธีโพลีโตมัสซิปเทสท์ กรณีอื่น ๆ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ มีอำนาจการทดสอบสูงกว่าวิธีโพลีโตมัสซิปเทสท์ และวิธีการทดสอบวอลด์

สำหรับกรณีขนาดของการทำหน้าที่ต่างกัน 0.70 พบว่า ทั้งสามวิธีมีอำนาจการทดสอบในระดับที่เท่ากัน เมื่อสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 10% ขนาดตัวอย่าง 1,000 : 1,000 สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 20% ขนาดตัวอย่าง 1,000 : 1,000 และ 250 : 500 ส่วนกรณีอื่น ๆ เมื่อสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 10% วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์มีอำนาจการทดสอบในระดับที่เท่ากัน ส่วนกรณีสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 20% กรณีอื่น ๆ พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุมีอำนาจการทดสอบสูงกว่าวิธีอื่น

สำหรับกรณีขนาดของการทำหน้าที่ต่างกัน 1.00 พบว่า เมื่อสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 10% พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ มีอำนาจการทดสอบในระดับที่เท่ากัน และมากกว่าวิธีโพลีโตมัสซิปเทสท์ ยกเว้น กรณีขนาดตัวอย่าง 250 : 250 และ 1,000 : 1,000 ทั้งสามวิธีมีอำนาจการทดสอบในระดับที่เท่ากัน และเมื่อสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 20% พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ มีอำนาจการทดสอบในระดับที่เท่ากัน และสูงกว่าวิธีโพลีโตมัสซิปเทสท์

เกือบทุกเงื่อนไข ยกเว้นเมื่อขนาดตัวอย่างเป็น 1,000 : 1,000 และ 500 : 1,000 ทั้งสามวิธีมีอำนาจการทดสอบในระดับที่เท่ากัน



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
PolySIBTEST	0.815	0.998	1	0.965	1	0.885	0.975	0.996	0.883	0.99	0.998	0.993	1	0.998	0.995	0.984	0.999	1	1	0.999	1	0.998	1	0.993	0.998	0.999	0.99	1	0.998	1
MG-CFA	0.995	1	1	0.998	1	0.985	1	1	0.998	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
WALD	0.958	1	1	0.995	1	0.946	0.998	1	0.975	0.998	1	1	1	1	1	0.998	0.997	1	1	0.997	1	1	1	1	1	1	1	1	1	1

ภาพที่ 29 กราฟแสดงอำนาจการทดสอบ ด้วยวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ เมื่อความยาวของแบบทดสอบจำนวน 40 ข้อ ภายใต้ปัจจัยที่แปรเปลี่ยน 3 ปัจจัย

ตอนที่ 2 ผลการวิเคราะห์ประสิทธิภาพการควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในโมเดลพหุมิติให้คะแนนหลายค่าด้วยวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ ความยาวของแบบทดสอบ ขนาดของการทำหน้าที่ต่างกัน สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง

ตารางที่ 5 ผลการวิเคราะห์ประสิทธิภาพการควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามด้วยวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัยหลัก

TEST LG	MAG DIF	PER DIF	SAMPSIZE (NF : NR)														
			250 : 250			500 : 500			1,000 : 1,000			250 : 500			500 : 1,000		
			P	M	W	P	M	W	P	M	W	P	M	W	P	M	W
20 ข้อ	0.40	10%	-	✓	✓	-	✓	✓	-	✓	✓	-	✓	✓	-	-	✓
		20%	-	✓	✓	-	✓	✓	-	-	-	-	✓	✓	-	-	-
	0.70	10%	-	✓	✓	-	✓	✓	-	✓		-	✓		-	✓	-
		20%	-	✓	-	-	✓	-	-	-	-	-	✓		-	✓	-
	1.00	10%	-	✓	✓	-	✓	✓	-	✓	-	-	✓	✓	-	✓	-
		20%	-	-	-	-	✓		-	-	-	-	-	-	-	✓	-
40 ข้อ	0.40	10%	✓	-	✓	-	✓	✓	-	-	✓	-	✓	✓	-	✓	✓
		20%	-	✓	✓	-	-	✓	-	-	-	-	-	✓	-	-	-
	0.70	10%	-	✓	✓	-		✓	-	-	✓	-	✓		-	✓	✓
		20%	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-
	1.00	10%	-	✓	✓	-	✓	✓	-	-	✓	-	✓	✓	-	-	✓
		20%	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-

หมายเหตุ P หมายถึง วิธีโพลีโตมัสซิปเทสท์
M หมายถึง วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ
และ W หมายถึง วิธีการทดสอบวอลด์

จากตารางที่ 5 ผลการวิเคราะห์ประสิทธิภาพการควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัยหลัก ได้แก่ ความยาวของแบบทดสอบ ขนาดของการทำหน้าที่ต่างกันของข้อคำถาม สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง ผลการวิเคราะห์พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ภายใต้ปัจจัยที่แปรเปลี่ยนสำหรับข้อคำถาม จำนวน 20 ข้อ ส่วนใหญ่วิธีการวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ รองลงมาคือ วิธีการทดสอบวอลด์ แต่สำหรับข้อคำถาม จำนวน 40 ข้อ ส่วนใหญ่วิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ดี รองลงมาคือ วิธีการทดสอบวอลด์ ส่วนวิธีโพล-ซิปเทสท์ ไม่สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ทุกปัจจัยที่แปรเปลี่ยน

ตารางที่ 6 ผลการวิเคราะห์ประสิทธิภาพอำนาจการทดสอบของวิธีการการตรวจสอบของวิธีทดสอบการทำหน้าที่ต่างกันของข้อคำถาม ด้วยวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัยหลัก

TEST LG	MAG DIF	PER DIF	SAMPSIZE (NF : NR)														
			250 : 250			500 : 500			1,000 : 1,000			250 : 500			500 : 1,000		
			P	M	W	P	M	W	P	M	W	P	M	W	P	M	W
20 ข้อ	0.40	10%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		20%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.70	10%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		20%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	1.00	10%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		20%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
40 ข้อ	0.40	10%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		20%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.70	10%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		20%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	1.00	10%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		20%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

หมายเหตุ P หมายถึง วิธีโพลีโตมัสซิปเทสท์
M หมายถึง วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ
และ W หมายถึง วิธีการทดสอบวอลด์

จากตารางที่ 6 ผลการวิเคราะห์ประสิทธิภาพอำนาจการทดสอบด้วยวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัยหลัก ได้แก่ ความยาวของแบบทดสอบ ขนาดของการทำหน้าที่ต่างกันของข้อคำถาม สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง ผลการวิเคราะห์พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ภายใต้ทุกปัจจัยที่แปรเปลี่ยน ทั้งสามวิธีมีอำนาจการทดสอบตามเกณฑ์ที่กำหนด

ตอนที่ 3 ผลการวิเคราะห์เปรียบเทียบความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในโมเดลพหุมิติให้คะแนนหลายค่า ระหว่างวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ ความยาวของแบบทดสอบ ขนาดของการทำหน้าที่ต่างกัน สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง

1. การเปรียบเทียบความแตกต่างของความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามของวิธีการตรวจสอบการทำหน้าที่ต่างกันสามวิธีภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย ผู้วิจัยใช้การวิเคราะห์ความแปรปรวนแบบวัดซ้ำประเภทสี่ตัวแปรระหว่างกลุ่มและหนึ่งตัวแปรภายในกลุ่มแบบวัดซ้ำ (Repeated-measures analysis: four between-subjects variables and one within-subjects variable) โดยตัวแปรระหว่างกลุ่มสี่ตัวแปร ได้แก่ ความยาวของแบบทดสอบ ขนาดของการทำหน้าที่ต่างกัน สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง และตัวแปรภายในกลุ่ม ได้แก่ วิธีการตรวจสอบการทำหน้าที่ต่างกัน ผลการวิเคราะห์แสดงดังต่อไปนี้

ตารางที่ 7 ผลการวิเคราะห์ความแปรปรวนของวิธีการตรวจสอบการทำหน้าที่ต่างกัน (แหล่งความแปรปรวนภายใน) ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย

Source	*SS*	*df*	*MS*	*F*
METHOD	308.022	2	154.011	21846.329***
METHOD X MAGDIF	46.400	4	11.600	1645.445***
METHOD X TESTLG	4.207	2	2.104	298.396***
METHOD X PROPDIF	55.740	2	27.870	3953.344***
METHOD X SAMPSIZE	15.865	8	1.983	281.297***

ตารางที่ 7 (ต่อ)

Source	*SS*	*df*	*MS*	*F*
METHOD X MAGDIF X TESTLG	.758	4	.189	26.875***
METHOD X MAGDIF X PROPDIF	3.327	4	.832	117.990***
METHOD X MAGDIF X SAMPSIZE	.729	16	.046	6.463***
METHOD X TESTLG X PROPDIF	.944	2	.472	66.988***
METHOD X TESTLG X SAMPSIZE	.622	8	.078	11.035***
METHOD X PROPDIF X SAMPSIZE	1.184	8	.148	20.986***
METHOD X MAGDIF X TESTLG X PROPDIF	.484	4	.121	17.171***
METHOD X MAGDIF X TESTLG X SAMPSIZE	.515	16	.032	4.564***
METHOD X MAGDIF X PROPDIF X SAMPSIZE	1.944	16	.121	17.234***
METHOD X TESTLG X PROPDIF X SAMPSIZE	.205	8	.026	3.644***
METHOD X MAGDIF X TESTLG X PROPDIF X SAMPSIZE	.396	16	.025	3.508***
Error	83.582	11856	.007	

***$p < .001$

จากตารางที่ 7 พบว่า วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามแตกต่างกัน มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกัน นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธี มีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 1 ของการตรวจสอบการทำหน้าที่ต่างกันแตกต่างกันที่ระดับนัยสำคัญ .001 ($F = 21846.329$)

เมื่อพิจารณาปฏิสัมพันธ์กันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามกับขนาดของการทำหน้าที่ต่างกันของข้อคำถาม พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธี มีผลให้อัตราความคลาดเคลื่อน

ประเภทที่ 1 แตกต่างกันที่ระดับนัยสำคัญ .001(F = 1645.445) ภายใต้ขนาดของการทำหน้าที่ต่างกันของข้อคำถามที่แตกต่างกัน

เมื่อพิจารณาปฏิสัมพันธ์กันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามกับความยาวของแบบทดสอบ พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธีมีอัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกันที่ระดับนัยสำคัญ .001(F = 298.396) ภายใต้เงื่อนไขความยาวของแบบทดสอบที่แตกต่างกัน

เมื่อพิจารณาปฏิสัมพันธ์กันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามกับสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธีมีอัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกันที่ระดับนัยสำคัญ .001(F = 3953.344) ภายใต้เงื่อนไขสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกันที่แตกต่างกัน

เมื่อพิจารณาปฏิสัมพันธ์กันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามกับขนาดตัวอย่าง พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธีมีอัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกันที่ระดับนัยสำคัญ .001 (F = 281.297) ภายใต้เงื่อนไขขนาดตัวอย่างที่แตกต่างกัน

เมื่อพิจารณาปฏิสัมพันธ์กันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามขนาดของการทำหน้าที่ต่างกัน และความยาวของแบบทดสอบ พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธีมีอัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกันที่ระดับนัยสำคัญ .001 (F = 26.875) ภายใต้เงื่อนไขขนาดของการทำหน้าที่ต่างกัน และความยาวของแบบทดสอบที่แตกต่างกัน

เมื่อพิจารณาปฏิสัมพันธ์กันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามขนาดของการทำหน้าที่ต่างกัน และสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธีมีอัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกันที่ระดับนัยสำคัญ .001 (F = 117.990) ภายใต้เงื่อนไขขนาดและสัดส่วนข้อคำถามที่แตกต่างกัน

เมื่อพิจารณาปฏิสัมพันธ์กันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามขนาดของการทำหน้าที่ต่างกัน และขนาดตัวอย่าง พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธีมีอัตราความคลาดเคลื่อน

ประเภทที่ 1 แตกต่างกันที่ระดับนัยสำคัญ .001 (F = 6.463) ภายใต้เงื่อนไขขนาดและขนาดตัวอย่างที่แตกต่างกัน

เมื่อพิจารณาปฏิสัมพันธ์กันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ความยาวของแบบทดสอบ และสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธีมีอัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกันที่ระดับนัยสำคัญ .001 (F = 66.988) ภายใต้เงื่อนไขความยาวของแบบทดสอบและสัดส่วนข้อคำถามที่แตกต่างกัน

เมื่อพิจารณาปฏิสัมพันธ์กันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ความยาวของแบบทดสอบ และขนาดตัวอย่าง พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธีมีอัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกันที่ระดับนัยสำคัญ .001 (F = 11.035) ภายใต้เงื่อนไขความยาวของแบบทดสอบและขนาดตัวอย่างที่แตกต่างกัน

เมื่อพิจารณาปฏิสัมพันธ์กันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม สัดส่วนข้อคำถาม และขนาดตัวอย่าง พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธีมีอัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกันที่ระดับนัยสำคัญ .001 (F = 20.986) ภายใต้เงื่อนไขสัดส่วนข้อคำถามและขนาดตัวอย่างที่แตกต่างกัน

เมื่อพิจารณาปฏิสัมพันธ์กันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ขนาดของการทำหน้าที่ต่างกัน ความยาวของแบบทดสอบ และสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธีมีอัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกันที่ระดับนัยสำคัญ .001 (F = 17.171) ภายใต้เงื่อนไขขนาดของการทำหน้าที่ต่างกัน ความยาวของแบบทดสอบ และสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน

เมื่อพิจารณาปฏิสัมพันธ์กันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ขนาดของการทำหน้าที่ต่างกัน ความยาวของแบบทดสอบ และขนาดตัวอย่าง พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธีมีอัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกันที่ระดับนัยสำคัญ .001 (F = 4.564) ภายใต้เงื่อนไขขนาดของการทำหน้าที่ต่างกัน ความยาวของแบบทดสอบ และขนาดตัวอย่าง

เมื่อพิจารณาปฏิสัมพันธ์กันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ขนาดของการทำหน้าที่ต่างกัน สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง พบว่า

มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธีมีอัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกันที่ระดับนัยสำคัญ .001 (F = 17.234) ภายใต้เงื่อนไขขนาดของการทำหน้าที่ต่างกัน สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง

เมื่อพิจารณาปฏิสัมพันธ์กันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ความยาวของแบบทดสอบ สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธีมีอัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกันที่ระดับนัยสำคัญ .001 (F = 3.644) ภายใต้เงื่อนไขความยาวของแบบทดสอบ สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง

เมื่อพิจารณาปฏิสัมพันธ์กันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ขนาดของการทำหน้าที่ต่างกัน ความยาวของแบบทดสอบ สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธีมีอัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกันที่ระดับนัยสำคัญ .001 (F = 3.508) ภายใต้เงื่อนไขขนาดของการทำหน้าที่ต่างกัน ความยาวของแบบทดสอบ สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง

ตารางที่ 8 ผลการวิเคราะห์ความแตกต่างของอัตราความคลาดเคลื่อนประเภทที่ 1 ภายใต้ขนาดของการทำหน้าที่ต่างกัน ความยาวของแบบทดสอบ สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง

Source	*SS*	*df*	*MS*	*F*
MAGDIF	47.221	2	23.611	2958.780***
TESTLG	2.420	1	2.420	303.274***
PROPDIF	76.122	1	76.122	9539.268***
SAMPSIZE	28.709	4	7.177	899.438***
MAGDIF * TESTLG	.918	2	.459	57.549***
MAGDIF * PROPDIF	6.099	2	3.050	382.176***
MAGDIF * SAMPSIZE	.997	8	.125	15.625***
TESTLG * PROPDIF	.444	1	.444	55.633***

ตารางที่ 8 (ต่อ)

Source	*SS*	*df*	*MS*	*F*
TESTLG * SAMPSIZE	.129	4	.032	4.054**
PROPDIF * SAMPSIZE	2.771	4	.693	86.808***
MAGDIF * TESTLG * PROPDIF	.056	2	.028	3.521*
MAGDIF * TESTLG* SAMPSIZE	.215	8	.027	3.360**
MAGDIF * PROPDIF * SAMPSIZE	.669	8	.084	10.478***
TESTLG * PROPDIF * SAMPSIZE	.099	4	.025	3.100*
MAGDIF * TESTLG * PROPDIF * SAMPSIZE	.109	8	.014	1.705
Error	47.305	5928	.008	

***$p < .001$, **$p < .01$, *$p < .05$

จากตารางที่ 8 พบว่า ขนาดของการทำหน้าที่ต่างกันของข้อคำถามแตกต่างกันมีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกัน นั่นคือ ขนาดของการทำหน้าที่ต่างกันของข้อคำถามที่แตกต่างกันสามขนาด มีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 1 ของการตรวจสอบการทำหน้าที่ต่างกันแตกต่างกันที่ระดับนัยสำคัญ .001 ($F = 2958.780$)

ความยาวของแบบทดสอบแตกต่างกันมีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกัน นั่นคือ ความยาวของแบบทดสอบที่แตกต่างกันสองขนาด มีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 1 ของการตรวจสอบการทำหน้าที่ต่างกันแตกต่างกันที่ระดับนัยสำคัญ .001 ($F = 303.274$)

สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน แตกต่างกันมีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกัน นั่นคือ สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกันที่แตกต่างกันสองขนาด มีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 1 ของการตรวจสอบการทำหน้าที่ต่างกันแตกต่างกันที่ระดับนัยสำคัญ .001 ($F = 9539.268$)

ขนาดตัวอย่างต่างกันมีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกัน นั่นคือ ขนาดตัวอย่างที่แตกต่างกันห้าขนาด มีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 1 ของการตรวจสอบการทำหน้าที่ต่างกันแตกต่างกันที่ระดับนัยสำคัญ .001 ($F = 899.438$)

เมื่อพิจารณาปฏิสัมพันธ์ระหว่างขนาดของการทำหน้าที่ต่างกัน และความยาวของแบบทดสอบ พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ ขนาดของการทำหน้าที่

ต่างกันที่แตกต่างกัน ความยาวของแบบทดสอบที่แตกต่างกัน มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกันที่ระดับนัยสำคัญ .001 (F = 57.549)

เมื่อพิจารณาปฏิสัมพันธ์ระหว่างขนาดของการทำหน้าที่ต่างกัน และสัดส่วนของการทำหน้าที่ต่างกัน พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ ขนาดของการทำหน้าที่ต่างกันที่แตกต่างกัน สัดส่วนของการทำหน้าที่ต่างกันที่แตกต่างกัน มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกันที่ระดับนัยสำคัญ .001 (F = 382.176)

เมื่อพิจารณาปฏิสัมพันธ์ระหว่างขนาดของการทำหน้าที่ต่างกัน และขนาดตัวอย่างที่แตกต่างกัน พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ ขนาดของการทำหน้าที่ต่างกันที่แตกต่างกัน ขนาดตัวอย่างที่แตกต่างกัน มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกันที่ระดับนัยสำคัญ .001 (F = 15.625)

เมื่อพิจารณาปฏิสัมพันธ์ระหว่างความยาวของแบบทดสอบ และสัดส่วนข้อคำถามที่แตกต่างกัน พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ ความยาวของแบบทดสอบ สัดส่วนข้อคำถามที่แตกต่างกัน มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกันที่ระดับนัยสำคัญ .001 (F = 55.633)

เมื่อพิจารณาปฏิสัมพันธ์ระหว่างความยาวของแบบทดสอบ และขนาดตัวอย่างที่แตกต่างกัน พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .01 นั่นคือ ความยาวของแบบทดสอบ ขนาดตัวอย่างที่แตกต่างกัน มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกันที่ระดับนัยสำคัญ .001 (F = 4.054)

เมื่อพิจารณาปฏิสัมพันธ์ระหว่างสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่างที่แตกต่างกัน พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ ความยาวของแบบทดสอบ ขนาดตัวอย่างที่แตกต่างกัน มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกันที่ระดับนัยสำคัญ .001 (F = 86.808)

เมื่อพิจารณาปฏิสัมพันธ์ระหว่างขนาดของการทำหน้าที่ต่างกัน ความยาวของแบบทดสอบ และสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .05 นั่นคือ ขนาดของการทำหน้าที่ต่างกัน ความยาวของแบบทดสอบ และสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกันที่แตกต่างกัน มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกันที่ระดับนัยสำคัญ .05 (F = 3.521)

เมื่อพิจารณาปฏิสัมพันธ์ระหว่างขนาดของการทำหน้าที่ต่างกัน ความยาวของแบบทดสอบ และขนาดตัวอย่างที่แตกต่างกัน พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .05 นั่นคือ ขนาด

ของการทำหน้าที่ต่างกัน ความยาวของแบบทดสอบ และขนาดตัวอย่างที่แตกต่างกัน มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกันที่ระดับนัยสำคัญ .01 (F = 3.360)

เมื่อพิจารณาปฏิสัมพันธ์ระหว่างขนาดของการทำหน้าที่ต่างกัน สัดส่วนของการทำหน้าที่ต่างกัน และขนาดตัวอย่างที่แตกต่างกัน พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ ขนาดของการทำหน้าที่ต่างกัน สัดส่วนของการทำหน้าที่ต่างกัน และขนาดตัวอย่างที่แตกต่างกัน มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกันที่ระดับนัยสำคัญ .001 (F = 10.478)

เมื่อพิจารณาปฏิสัมพันธ์ระหว่างความยาวของแบบทดสอบ สัดส่วนของการทำหน้าที่ต่างกัน และขนาดตัวอย่างที่แตกต่างกัน พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ ความยาวของแบบทดสอบ สัดส่วนของการทำหน้าที่ต่างกัน และขนาดตัวอย่างที่แตกต่างกัน มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกันที่ระดับนัยสำคัญ .001 (F = 3.100)

เมื่อพิจารณาปฏิสัมพันธ์ระหว่างขนาดของการทำหน้าที่ต่างกัน ความยาวของแบบทดสอบ สัดส่วนของการทำหน้าที่ต่างกัน และขนาดตัวอย่างที่แตกต่างกัน พบว่า ไม่มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญ นั่นคือ ขนาดของการทำหน้าที่ต่างกัน ความยาวของแบบทดสอบ สัดส่วนของการทำหน้าที่ต่างกัน และขนาดตัวอย่างที่แตกต่างกัน ไม่มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 ไม่แตกต่างกัน (F = 1.705)

2. การเปรียบเทียบอำนาจการทดสอบในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามของวิธีการตรวจสอบการทำหน้าที่ต่างกันสามวิธี ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย

ตารางที่ 9 ผลการวิเคราะห์ความแตกต่างของอำนาจการทดสอบด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันสามวิธี ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย

Source	*SS*	*df*	*MS*	*F*
METHOD	1.900	2	.950	23.296***
METHOD X MAGDIF	2.787	4	.697	17.084***
METHOD X TESTLG	.256	2	.128	3.141*
METHOD X PROPDIF	.870	2	.435	10.664***
METHOD X SAMPSIZE	1.880	8	.235	5.761***
METHOD X MAGDIF X TESTLG	.087	4	.022	.532
METHOD X MAGDIF X PROPDIF	1.729	4	.432	10.596***
METHOD X MAGDIF X SAMPSIZE	2.867	16	.179	4.393***

ตารางที่ 9 (ต่อ)

Source	*SS*	*df*	*MS*	*F*
METHOD X TESTLG X PROPDIF	.080	2	.040	.984
METHOD X TESTLG X SAMPSIZE	.839	8	.105	2.572**
METHOD X PROPDIF X SAMPSIZE	.595	8	.074	1.824
METHOD X MAGDIF X TESTLG X PROPDIF	.081	4	.020	.497
METHOD X MAGDIF X TESTLG X SAMPSIZE	2.350	16	.147	3.601***
METHOD X MAGDIF X PROPDIF X SAMPSIZE	1.238	16	.077	1.897*
METHOD X TESTLG X PROPDIF X SAMPSIZE	1.163	8	.145	3.564***
METHOD X MAGDIF X TESTLG X PROPDIF X SAMPSIZE	1.979	16	.124	3.033***
Error	484.529	11880	.041	

***$p < .001$, **$p < .01$, *$p < .05$

จากตารางที่ 9 พบว่า วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามที่แตกต่างกัน มีผลให้อำนาจการทดสอบแตกต่างกัน นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธี มีผลต่ออำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันที่แตกต่างกัน ณ ระดับนัยสำคัญ .001 ($F = 23.296$)

เมื่อพิจารณาปฏิสัมพันธ์กันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามกับขนาดของการทำหน้าที่ต่างกันของข้อคำถาม พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธี มีผลให้อำนาจการทดสอบแตกต่างกันที่ระดับนัยสำคัญ .001 ($F = 17.084$) ภายใต้ขนาดของการทำหน้าที่ต่างกันของข้อคำถามที่แตกต่างกัน

เมื่อพิจารณาปฏิสัมพันธ์กันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามกับความยาวของแบบทดสอบ พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .05 นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธี มีผลให้อำนาจการทดสอบแตกต่างกัน ณ ระดับนัยสำคัญ .05 ($F = 3.141$) ภายใต้เงื่อนไขความยาวของแบบทดสอบที่แตกต่างกัน

เมื่อพิจารณาปฏิสัมพันธ์กันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามกับสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธี มีผลทำให้อำนาจการทดสอบแตกต่างกันที่ระดับนัยสำคัญ .001 (F = 10.664) ภายใต้เงื่อนไขสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกันที่แตกต่างกัน

เมื่อพิจารณาปฏิสัมพันธ์กันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามกับขนาดตัวอย่าง พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธีมีผลให้อำนาจการทดสอบ แตกต่างกันที่ระดับนัยสำคัญ .001 (F = 5.761) ภายใต้เงื่อนไขขนาดตัวอย่างที่แตกต่างกัน

เมื่อพิจารณาปฏิสัมพันธ์กันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามขนาดของการทำหน้าที่ต่างกัน และความยาวของแบบทดสอบ พบว่า ไม่มีปฏิสัมพันธ์กัน นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธีไม่มีผลทำให้อำนาจการทดสอบแตกต่างกัน (F = .984) ภายใต้เงื่อนไขขนาดของการทำหน้าที่ต่างกัน และความยาวของแบบทดสอบที่แตกต่างกัน

เมื่อพิจารณาปฏิสัมพันธ์กันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามขนาดของการทำหน้าที่ต่างกัน และสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธีมีผลทำให้อำนาจการทดสอบแตกต่างกันที่ระดับนัยสำคัญ .001 (F = 10.596) ภายใต้เงื่อนไขขนาดและสัดส่วนข้อคำถามที่แตกต่างกัน

เมื่อพิจารณาปฏิสัมพันธ์กันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามขนาดของการทำหน้าที่ต่างกัน และขนาดตัวอย่าง พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธีมีผลให้อำนาจการทดสอบแตกต่างกันที่ระดับนัยสำคัญ .001 (F = 4.393) ภายใต้เงื่อนไขขนาดและขนาดตัวอย่างที่แตกต่างกัน

เมื่อพิจารณาปฏิสัมพันธ์กันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามความยาวของแบบทดสอบ และสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน พบว่า ไม่มีปฏิสัมพันธ์กัน นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธีไม่มีผลทำให้อำนาจการทดสอบแตกต่างกัน (F = .984) ภายใต้เงื่อนไขความยาวของแบบทดสอบและสัดส่วนข้อคำถามที่แตกต่างกัน

เมื่อพิจารณาปฏิสัมพันธ์กันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามความยาวของแบบทดสอบ และขนาดตัวอย่าง พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .01 นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธีมีผลให้อำนาจการทดสอบแตกต่างกันที่ระดับนัยสำคัญ .01 (F = 2.572) ภายใต้เงื่อนไขความยาวของแบบทดสอบและขนาดตัวอย่างที่แตกต่างกัน

เมื่อพิจารณาปฏิสัมพันธ์กันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม สัดส่วนข้อคำถาม และขนาดตัวอย่าง พบว่า ไม่มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญ นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธีไม่มีผลให้อำนาจการทดสอบแตกต่างกัน (F = 1.824) ภายใต้เงื่อนไขสัดส่วนข้อคำถามและขนาดตัวอย่างที่แตกต่างกัน

เมื่อพิจารณาปฏิสัมพันธ์กันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ขนาดของการทำหน้าที่ต่างกัน ความยาวของแบบทดสอบ และสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน พบว่า ไม่มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญ นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธีไม่มีผลให้อำนาจการทดสอบแตกต่างกัน (F = .497) ภายใต้เงื่อนไขขนาดของการทำหน้าที่ต่างกัน ความยาวของแบบทดสอบ และสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน

เมื่อพิจารณาปฏิสัมพันธ์กันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ขนาดของการทำหน้าที่ต่างกัน ความยาวของแบบทดสอบ และขนาดตัวอย่าง พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธีมีผลให้อำนาจการทดสอบแตกต่างกันที่ระดับนัยสำคัญ .001 (F = 3.601) ภายใต้เงื่อนไขขนาดของการทำหน้าที่ต่างกัน ความยาวของแบบทดสอบ และขนาดตัวอย่าง

เมื่อพิจารณาปฏิสัมพันธ์กันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ขนาดของการทำหน้าที่ต่างกัน สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .05 นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธีมีผลให้อำนาจการทดสอบแตกต่างกันที่ระดับนัยสำคัญ .05 (F = 1.897) ภายใต้เงื่อนไขขนาดของการทำหน้าที่ต่างกัน สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง

เมื่อพิจารณาปฏิสัมพันธ์กันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ความยาวของแบบทดสอบ สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธีมีผลให้อำนาจการทดสอบแตกต่างกันที่ระดับนัยสำคัญ .001 (F = 3.564) ภายใต้เงื่อนไขความยาวของแบบทดสอบ สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง

เมื่อพิจารณาปฏิสัมพันธ์กันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ขนาดของการทำหน้าที่ต่างกัน ความยาวของแบบทดสอบ สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธีมีผลให้อำนาจการทดสอบแตกต่างกันที่ระดับนัยสำคัญ .001 (F = 3.033) ภายใต้เงื่อนไขขนาดของการทำหน้าที่ต่างกัน ความยาวของแบบทดสอบ สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง

ตารางที่ 10 ผลการวิเคราะห์ความแตกต่างของอำนาจการทดสอบ ภายใต้ขนาดของการทำหน้าที่ต่างกัน ความยาวของแบบทดสอบ สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง

Source	*SS*	*df*	*MS*	*F*
MAGDIF	3.861	2	1.930	45.599***
TESTLG	.150	1	.150	3.549
PROPDIF	.047	1	.047	1.104
SAMPSIZE	3.278	4	.819	19.356***
MAGDIF * TESTLG	.070	2	.035	.824
MAGDIF * PROPDIF	1.065	2	.532	12.576***
MAGDIF * SAMPSIZE	3.237	8	.405	9.559***
TESTLG * PROPDIF	.165	1	.165	3.898*
TESTLG * SAMPSIZE	.502	4	.126	2.966*
PROPDIF * SAMPSIZE	.342	4	.086	2.021
MAGDIF * TESTLG * PROPDIF	.062	2	.031	.730
MAGDIF * TESTLG * SAMPSIZE	.830	8	.104	2.450*
MAGDIF * PROPDIF * SAMPSIZE	.820	8	.103	2.422*
testlengh * PROPDIF * SAMPSIZE	.514	4	.129	3.037*
MAGDIF * TESTLG * PROPDIF * SAMPSIZE	.859	8	.107	2.537**
Error	251.462	5940	.042	

***$p < .001$, ** $p < .01$, *$p < .05$

จากตารางที่ 10 พบว่า ขนาดของการทำหน้าที่ต่างกันของข้อคำถามแตกต่างกันมีผลให้อำนาจการทดสอบแตกต่างกัน นั่นคือ ขนาดของการทำหน้าที่ต่างกันของข้อคำถามที่แตกต่างกันสามขนาด มีผลต่ออำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันแตกต่างกันที่ระดับนัยสำคัญ .001 ($F = 45.599$)

ความยาวของแบบทดสอบแตกต่างกันไม่มีผลต่ออำนาจการทดสอบแตกต่างกัน นั่นคือ ความยาวของแบบทดสอบที่แตกต่างกันสองขนาด ไม่มีผลต่ออำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกัน ($F = 3.549$)

สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกันแตกต่างกันไม่มีผลต่ออำนาจการทดสอบแตกต่างกัน นั่นคือ สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกันที่แตกต่างกันสองขนาด ไม่มีผลต่ออำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันแตกต่างกันที่ระดับนัยสำคัญ .001 (F = 1.104)

ขนาดตัวอย่างต่างกันมีผลให้อำนาจการทดสอบแตกต่างกัน นั่นคือ ขนาดตัวอย่างที่แตกต่างกันห้าขนาด มีผลต่ออำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันแตกต่างกันที่ระดับนัยสำคัญ .001 (F = 19.356)

เมื่อพิจารณาปฏิสัมพันธ์ระหว่างขนาดของการทำหน้าที่ต่างกัน และความยาวของแบบทดสอบ พบว่า ไม่มีปฏิสัมพันธ์กัน นั่นคือ ขนาดของการทำหน้าที่ต่างกันที่แตกต่างกัน ความยาวของแบบทดสอบที่แตกต่างกัน ไม่มีผลต่ออำนาจการทดสอบที่แตกต่างกัน (F = .824)

เมื่อพิจารณาปฏิสัมพันธ์ระหว่างขนาดของการทำหน้าที่ต่างกัน และสัดส่วนของการทำหน้าที่ต่างกัน พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ ขนาดของการทำหน้าที่ต่างกันที่แตกต่างกัน สัดส่วนของการทำหน้าที่ต่างกันที่แตกต่างกัน มีผลให้อำนาจการทดสอบแตกต่างกันที่ระดับนัยสำคัญ .001 (F = 12.576)

เมื่อพิจารณาปฏิสัมพันธ์ระหว่างขนาดของการทำหน้าที่ต่างกัน และขนาดตัวอย่างที่แตกต่างกัน พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .001 นั่นคือ ขนาดของการทำหน้าที่ต่างกันที่แตกต่างกัน ขนาดตัวอย่างที่แตกต่างกัน มีผลให้อำนาจการทดสอบแตกต่างกันที่ระดับนัยสำคัญ .001 (F = 9.559)

เมื่อพิจารณาปฏิสัมพันธ์ระหว่างความยาวของแบบทดสอบ และสัดส่วนข้อคำถามที่แตกต่างกัน พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .05 นั่นคือ ความยาวของแบบทดสอบ สัดส่วนข้อคำถามที่แตกต่างกัน มีผลให้อำนาจการทดสอบแตกต่างกันที่ระดับนัยสำคัญ .05 (F = 3.898)

เมื่อพิจารณาปฏิสัมพันธ์ระหว่างความยาวของแบบทดสอบ และขนาดตัวอย่างที่แตกต่างกัน พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .05 นั่นคือ ความยาวของแบบทดสอบ ขนาดตัวอย่างที่แตกต่างกัน มีผลให้อำนาจการทดสอบแตกต่างกันที่ระดับนัยสำคัญ .001 (F = 2.966)

เมื่อพิจารณาปฏิสัมพันธ์ระหว่างสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่างที่แตกต่างกัน พบว่า ไม่มีปฏิสัมพันธ์กัน นั่นคือ ความยาวของแบบทดสอบ ขนาดตัวอย่างที่แตกต่างกัน ไม่มีผลทำให้อำนาจการทดสอบแตกต่างกัน (F = 2.021)

เมื่อพิจารณาปฏิสัมพันธ์ระหว่างขนาดของการทำหน้าที่ต่างกัน ความยาวของแบบทดสอบ และสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน พบว่า ไม่มีปฏิสัมพันธ์กัน นั่นคือ ขนาดของการทำหน้าที่

ต่างกัน ความยาวของแบบทดสอบ และสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน ที่แตกต่างกัน ไม่มีผลให้อำนาจการทดสอบแตกต่างกัน (F = .730)

เมื่อพิจารณาปฏิสัมพันธ์ระหว่างขนาดของการทำหน้าที่ต่างกัน ความยาวของแบบทดสอบ และขนาดตัวอย่างที่แตกต่างกัน พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .05 นั่นคือ ขนาดของการทำหน้าที่ต่างกัน ความยาวของแบบทดสอบ และขนาดตัวอย่างที่แตกต่างกัน มีผลให้อำนาจการทดสอบแตกต่างกันที่ระดับนัยสำคัญ .05 (F = 2.450)

เมื่อพิจารณาปฏิสัมพันธ์ระหว่างขนาดของการทำหน้าที่ต่างกัน สัดส่วนของการทำหน้าที่ต่างกัน และขนาดตัวอย่างที่แตกต่างกัน พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .05 นั่นคือ ขนาดของการทำหน้าที่ต่างกัน สัดส่วนของการทำหน้าที่ต่างกัน และขนาดตัวอย่างที่แตกต่างกัน มีผลให้อำนาจการทดสอบแตกต่างกันที่ระดับนัยสำคัญ .05 (F = 2.422)

เมื่อพิจารณาปฏิสัมพันธ์ระหว่างความยาวของแบบทดสอบ สัดส่วนของการทำหน้าที่ต่างกัน และขนาดตัวอย่างที่แตกต่างกัน พบว่า มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ .05 นั่นคือ ความยาวของแบบทดสอบ สัดส่วนของการทำหน้าที่ต่างกัน และขนาดตัวอย่างที่แตกต่างกัน มีผลให้อำนาจการทดสอบแตกต่างกันที่ระดับนัยสำคัญ .05 (F = 3.037)

เมื่อพิจารณาปฏิสัมพันธ์ระหว่างขนาดของการทำหน้าที่ต่างกัน ความยาวของแบบทดสอบ สัดส่วนของการทำหน้าที่ต่างกัน และขนาดตัวอย่างที่แตกต่างกัน พบว่า ไม่มีปฏิสัมพันธ์กันอย่างมีนัยสำคัญ นั่นคือ ขนาดของการทำหน้าที่ต่างกัน ความยาวของแบบทดสอบ สัดส่วนของการทำหน้าที่ต่างกัน และขนาดตัวอย่างที่แตกต่างกัน ไม่มีผลให้อำนาจการทดสอบแตกต่างกันที่ระดับนัยสำคัญ .05 (F = 2.537)

บทที่ 5
สรุปผล อภิปรายผล และข้อเสนอแนะ

ในการวิจัยครั้งนี้ มีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า ด้วยวิธีโพลีโตมัสซิปเทสท์ วิเคราะห์โดยใช้โปรแกรม DIFPACK วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกล่มพหุ วิเคราะห์ด้วยโปรแกรม Mplus และวิธีการทดสอบวอลด์ วิเคราะห์ด้วยโปรแกรม IRTPRO ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย ได้แก่ ความยาวของแบบทดสอบ 2 ขนาด ขนาดของการทำหน้าที่ต่างกันของข้อคำถาม 3 ขนาด สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 2 ขนาด และขนาดตัวอย่าง 5 ขนาด ข้อมูลที่ใช้ในการศึกษาครั้งนี้เป็นข้อมูลจำลอง จำนวน 60 เงื่อนไข (2x3x2x5) ในแต่ละเงื่อนไขจำลองข้อมูลซ้ำ 100 ครั้ง

สรุปผลการวิจัย

การสรุปผลการวิจัยแบ่งออกเป็น 2 ตอน คือ 1) ผลการทดสอบปัจจัยที่ศึกษาที่มีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 1 ของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามพหุมิติที่ให้คะแนนหลายค่า และ 2) ผลการทดสอบปัจจัยที่ศึกษาที่มีผลต่ออำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามพหุมิติที่ให้คะแนนหลายค่า ระหว่างวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย ดังรายละเอียดต่อไปนี้

1. ผลการวิเคราะห์ประสิทธิภาพการควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ด้วยวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย ได้แก่ ความยาวของแบบทดสอบ ขนาดของการทำหน้าที่ต่างกันของข้อคำถาม สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง ผลการวิเคราะห์พบว่า เมื่อความยาวของแบบทดสอบ จำนวน 20 ข้อ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ดี รองลงมาคือ วิธีการทดสอบวอลด์ ส่วนวิธีโพลีโตมัสซิปเทสท์ ไม่สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ทุกเงื่อนไขภายใต้ปัจจัยที่แปรเปลี่ยน และเมื่อขนาดของการทำหน้าที่ต่างกันเพิ่มขึ้น สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกันเพิ่มขึ้น วิธีการวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ มีความสามารถในการควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ลดลง และเมื่อพิจารณา

ความยาวของแบบทดสอบ จำนวน 40 ข้อ พบว่า วิธีการทดสอบวอลด์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ดีที่สุด รองลงมาคือ วิธีการวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และเมื่อขนาดของการทำหน้าที่ต่างกันเพิ่มขึ้น สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกันเพิ่มขึ้น ทั้งสองวิธีจะมีความสามารถในการควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ลดลง และไม่มีวิธีใดสามารถควบคุมอัตราความคลาดเคลื่อนได้เมื่อขนาดของการทำหน้าที่ต่างกันขนาดใหญ่และสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกันมากขึ้น

2. ผลการวิเคราะห์อำนาจการทดสอบด้วยวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย ได้แก่ ความยาวของแบบทดสอบ ขนาดของการทำหน้าที่ต่างกันของข้อคำถาม สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง ผลการวิเคราะห์พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามภายใต้ทุกปัจจัยที่แปรเปลี่ยน ทั้งสามวิธีมีอำนาจการทดสอบตามเกณฑ์ที่กำหนด โดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุมีอำนาจการทดสอบสูงสุดเป็นส่วนใหญ่ รองลงมาคือ วิธีการทดสอบวอลด์ และโพลีโตมัสซิปเทสท์ และเมื่อความยาวของแบบทดสอบ ขนาดของการทำหน้าที่ต่างกัน และขนาดตัวอย่างเพิ่มขึ้น ทั้งสามวิธีมีอำนาจการทดสอบเพิ่มขึ้น

สรุปได้ว่า ประสิทธิภาพเมื่อพิจารณาทั้งอัตราความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของวิธีตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ดี และมีค่าใกล้เคียงกัน โดยทั้งสองวิธีควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ลดลงเมื่อสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกันเพิ่มขึ้น ส่วนอำนาจการทดสอบพบว่า ทั้งสามวิธีมีอำนาจการทดสอบเป็นไปตามเกณฑ์ที่กำหนด และมีค่าใกล้เคียงกันในทุกปัจจัยที่แปรเปลี่ยน

อภิปรายผลการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อตอบคำถามในสองประเด็น คือ 1) ศึกษาและเปรียบเทียบอัตราความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติที่ให้คะแนนหลายค่า ด้วยวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย และ 2) ศึกษาและเปรียบเทียบอำนาจการทดสอบในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติที่ให้คะแนนหลายค่าของวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ ความยาวของแบบทดสอบ 2 ขนาด

ขนาดของการทำหน้าที่ต่างกันของข้อคำถาม 3 ขนาด สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 2 ขนาด และขนาดตัวอย่าง 5 รูปแบบ จากผลการวิจัยที่สรุปมาข้างต้นมีประเด็นสำคัญที่นำมาอภิปรายผลดังนี้

1. ปัจจัยความยาวของแบบทดสอบ

ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกัน ที่ระดับนัยสำคัญ .001 ภายใต้ปัจจัยความยาวของแบบทดสอบ โดยที่วิธีโพลีโตมัสซิปเทสท์ไม่สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ที่กำหนด โดยมีอัตราความคลาดเคลื่อนประเภทที่ 1 เฉลี่ยประมาณ 0.404 และ 0.344 ตามลำดับ ซึ่งไม่สอดคล้องกับการศึกษาของสิริรัตน์ วิภาสศิลป์ (2545) สุชาติ สิริมีนนันท์ (2554) และ Cohen and Kim (1993) พบว่า สำหรับวิธีการตรวจสอบการทำหน้าที่ต่างกันของวิธีโพลีโตมัสซิปเทสท์ เมื่อความยาวของแบบทดสอบเพิ่มขึ้น ส่งผลต่อการระบุผิดพลาดในการตรวจสอบสูง โดยมีอัตราความคลาดเคลื่อนประเภทที่ 1 เพิ่มขึ้น ส่วนวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุและวิธีการทดสอบวอลด์ สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดีกว่าวิธีโพลีโตมัสซิปเทสท์ อย่างมีนัยสำคัญที่ .05 ภายใต้ความยาวของแบบทดสอบทั้งสองขนาด และเมื่อพิจารณาเฉพาะวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ พบว่า สำหรับความยาวของแบบทดสอบ 20 ข้อ และ 40 ข้อ สามารถในการควบคุมอัตราความคลาดเคลื่อนที่ประเภทที่ 1 ได้ 38% และ 15% ตามลำดับ โดยมีอัตราความคลาดเคลื่อนประเภทที่ 1 เฉลี่ยประมาณ 0.079 และ 0.095 ตามลำดับจะเห็นว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุมีอัตราความคลาดเคลื่อนประเภทที่ 1 สูงขึ้น เมื่อความยาวของแบบทดสอบเพิ่มขึ้น ซึ่งไม่สอดคล้องกับการศึกษาของ French and Finch (2008) ที่พบว่า เมื่อความยาวของแบบทดสอบมีค่าเพิ่มขึ้น อัตราความคลาดเคลื่อนประเภทที่ 1 มีค่าลดลง เมื่อพิจารณาเฉพาะวิธีการทดสอบวอลด์ พบว่า วิธีนี้สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ 25% และ 31.67% ตามลำดับ โดยมีอัตราความคลาดเคลื่อนประเภทที่ 1 เฉลี่ยประมาณ 0.118 และ 0.092 ตามลำดับ จะเห็นว่า เมื่อความยาวของแบบทดสอบเพิ่มขึ้น วิธีการทดสอบวอลด์มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 ลดลง ซึ่งไม่สอดคล้องกับ Cao et al. (2017) ที่พบว่า อัตราความคลาดเคลื่อนประเภทที่ 1 มีค่าคงที่ หรือกล่าวว่า ความยาวของแบบทดสอบไม่มีผลกับอัตราความคลาดเคลื่อนประเภทที่ 1

เมื่อพิจารณาอำนาจการทดสอบ พบว่า ทั้งสามวิธีมีผลให้อำนาจการทดสอบในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามแตกต่างกันที่ระดับนัยสำคัญ .05 ภายใต้ความยาวของแบบทดสอบที่แตกต่างกัน โดยวิธีโพลีโตมัสซิปเทสท์มีอำนาจการทดสอบเฉลี่ยคงที่

เมื่อความยาวของแบบทดสอบแปรเปลี่ยน ซึ่งไม่สอดคล้องกับการศึกษาของสุชาติ สิริมีนนันท์ (2554) ที่พบว่า เมื่อเพิ่มความยาวแบบทดสอบจาก 20 ข้อ เป็น 40 ข้อ อัตราความถูกต้องในการตรวจสอบการทำหน้าที่ต่างกันของวิธีโพลีโตมัสซิปเทสท์จะเพิ่มขึ้น และเมื่อพิจารณาวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ พบว่า เมื่อความยาวของแบบทดสอบเพิ่มขึ้น ทั้งสองวิธีจะมีอำนาจการทดสอบเพิ่มขึ้น โดยมีอำนาจการทดสอบเฉลี่ย 0.998, 0.999 และ 0.986, 0.995 ตามลำดับ ซึ่งสอดคล้องกับ Cao et al. (2017) Langer (2008) ที่พบว่า วิธีการทดสอบวอลด์ มีอำนาจการทดสอบเพิ่มขึ้นเมื่อความยาวของแบบทดสอบเพิ่มขึ้น แต่ไม่สอดคล้องกับ French and Finch (2008) ที่พบว่า วิธีการวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ มีอำนาจการทดสอบลดลงเมื่อความยาวของแบบทดสอบมากขึ้น

2. ปัจจัยขนาดของการทำหน้าที่ต่างกันของข้อคำถาม

ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติให้คะแนนหลายค่า พบว่า วิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกัน ที่ระดับนัยสำคัญ .001 ภายใต้ปัจจัยขนาดของการทำหน้าที่ต่างกัน 0.40, 0.70 และ 1.00

เมื่อพิจารณาวิธีโพลีโตมัสซิปเทสท์ พบว่า ไม่สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ โดยมีอัตราความคลาดเคลื่อนประเภทที่ 1 เฉลี่ยประมาณ 0.217, 0.393 และ 0.519 ตามลำดับ ซึ่งมีค่าเพิ่มขึ้นเมื่อขนาดของการทำหน้าที่ต่างกันเพิ่มขึ้น ซึ่งสอดคล้องกับการศึกษาของ Kilmen (2016) ที่พบว่า วิธีโพลีโตมัสซิปเทสท์มีอัตราความคลาดเคลื่อนประเภทที่ 1 เพิ่มขึ้นเมื่อขนาดของการทำหน้าที่ต่างกันเพิ่มขึ้น เมื่อพิจารณาวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ พบว่า สามารถในการควบคุมอัตราความคลาดเคลื่อนที่ประเภทที่ 1 ได้ 18.33%, 20.00% และ 16.67% ตามลำดับ โดยมีอัตราความคลาดเคลื่อนประเภทที่ 1 เฉลี่ย 0.086, 0.087 และ 0.088 ตามลำดับ ซึ่งไม่สอดคล้องกับการศึกษาของ Chang, Huang, and Tsai (2015) ที่พบว่า ขนาดของการทำหน้าที่ต่างกันไม่มีผลต่อความคลาดเคลื่อนประเภทที่ 1 ของวิธีการตรวจสอบการทำหน้าที่ต่างกันด้วยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ Kim and Yoon (2011) และ Stark, Chernyshenko, and Drasgow (2006) ที่พบว่า ขนาดของการทำหน้าที่ต่างกันเพิ่มขึ้น อัตราความคลาดเคลื่อนประเภทที่ 1 เพิ่มขึ้น และวิธีการทดสอบวอลด์ พบว่า สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ 26.67%, 16.67% และ 13.33% ตามลำดับ โดยมีอัตราความคลาดเคลื่อนประเภทที่ 1 เฉลี่ยประมาณ 0.065, 0.101 และ 0.142 ตามลำดับ ซึ่งสอดคล้องกับ Cao et al. (2017) พบว่า ขนาดของการทำหน้าที่ต่างกันเพิ่มขึ้น อัตราความคลาดเคลื่อนประเภทที่ 1 มีค่าเพิ่มขึ้น

เมื่อพิจารณาอำนาจการทดสอบในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามที่ให้คะแนนหลายค่า ภายใต้ปัจจัยขนาดของการทำหน้าที่ต่างกันของข้อคำถาม ทั้งสามวิธี พบว่า วิธีการตรวจสอบการทำหน้าที่ต่างกันทั้งสามวิธีมีผลทำให้ อำนาจการทดสอบมีค่าแตกต่างกันที่ระดับนัยสำคัญ .001 โดยทั้งสามวิธีมีอำนาจการทดสอบในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามเป็นไปตามเกณฑ์ที่กำหนด ที่ระดับนัยสำคัญที่ .05 เมื่อพิจารณาวิธีโพลีโตมัสซิปเทสท์ พบว่า มีอำนาจการทดสอบเฉลี่ย 0.953, 0.995 และ 0.995 ตามลำดับ โดยมีค่าเพิ่มขึ้นเมื่อขนาดของการทำหน้าที่ต่างกันเพิ่มขึ้น ซึ่งสอดคล้องกับการศึกษาของ Awour (2008); Finch (2005); Gierl et al. (2004); Gonzales-Roma et al. (2006); Kilmen (2016); Lei and Li (2013); Narayanan and Swaminathan (1994) ที่พบว่า ขนาดของการทำหน้าที่ต่างกันเพิ่มขึ้น มีผลต่อประสิทธิภาพของวิธีโพลีโตมัสซิปเทสท์ และมีอำนาจการทดสอบเพิ่มขึ้น สำหรับวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ พบว่า มีอำนาจการทดสอบเพิ่มขึ้น เมื่อขนาดของการทำหน้าที่ต่างกันเพิ่มขึ้น โดยมีค่าอำนาจการทดสอบเฉลี่ย 0.997, 0.999 และ 0.999 ตามลำดับ สอดคล้องกับการศึกษาของ Flowers et al. (1999); Meade et al. (2006); Stark et al. (2006); Chang et al. (2015); Kim and Yoon (2011) พบว่า อำนาจการทดสอบของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามด้วยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุมีความสัมพันธ์กับขนาดของการทำหน้าที่ต่างกันอย่างมีนัยสำคัญ นั่นคือ เมื่อขนาดของการทำหน้าที่ต่างกันมากขึ้น อำนาจการทดสอบมีค่าเพิ่มขึ้น และสอดคล้องกับการศึกษาของ Gonzales-Roma et al. (2006) ที่พบว่า ขนาดของการทำหน้าที่ต่างกันขนาดเล็กจะมีอำนาจในการทดสอบสำหรับการทำหน้าที่ต่างกันของข้อสอบต่ำ แต่ไม่สอดคล้องกับการศึกษาของ French and Finch (2008) และมิ่ง เทพครเมือง (2554) ที่พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ มีอำนาจการทดสอบในการตรวจสอบการทำหน้าที่ต่างกันลดลง เมื่อขนาดของการทำหน้าที่ต่างกันเพิ่มขึ้น และเมื่อพิจารณาวิธีการทดสอบวอลด์ พบว่า เมื่อขนาดของการทำหน้าที่ต่างกันเพิ่มขึ้น จะมีค่าอำนาจการทดสอบเพิ่มขึ้น ซึ่งมีค่าเฉลี่ย 0.978, 0.997 และ 0.998 ตามลำดับ จะเห็นว่าทั้งสามวิธีมีอำนาจการทดสอบเพิ่มขึ้นเมื่อขนาดของการทำหน้าที่ต่างกันของข้อคำถามมากขึ้น ซึ่งสอดคล้องกับการศึกษาของ Cao et al. (2017) และ Langer (2008) พบว่า เมื่อขนาดของการทำหน้าที่ต่างกันเพิ่มขึ้น จะมีอำนาจการทดสอบการทำหน้าที่ต่างกันเพิ่มขึ้น

3. ปัจจัยสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน

ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามที่ให้คะแนนหลายค่า ภายใต้ปัจจัยสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 10% และ 20% พบว่า ทั้งสามวิธีมีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกันที่ระดับนัยสำคัญ .001 โดยวิธีโพลีโตมัสซิปเทสท์มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 เฉลี่ย 0.233 และ 0.515 ตามลำดับ ผลการทดสอบสอดคล้องกับ

Kilmen (2016); Narayanan and Swaminathan (1994) และสุชาติ สิริมีนนันท์ (2554) พบว่า วิธีโพลีโตมัสซิปเทสท์มีอัตราความคลาดเคลื่อนประเภทที่ 1 มากขึ้น เมื่อสัดส่วนข้อสอบที่ทำหน้าที่ต่างกันมากขึ้น ซึ่งไม่สอดคล้องกับ Cohen and Kim (1993); Shih, Liu, and Wang (2014) พบว่า เมื่อสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกันเพิ่มขึ้นแล้ว อัตราความคลาดเคลื่อนประเภทที่ 1 จะมีค่าลดลง ส่วนวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ ที่ระดับนัยสำคัญ .05 สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ 38.33% และ 18.33% ตามลำดับ และมีอัตราความคลาดเคลื่อนประเภทที่ 1 เฉลี่ย 0.079 และ 0.095 ตามลำดับ โดยอัตราความคลาดเคลื่อนจะมีค่าเพิ่มขึ้น เมื่อสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกันเพิ่มขึ้น ซึ่งผลที่ได้ไม่สอดคล้องกับการศึกษาของ French and Finch (2008) และ Chang et al. (2015) ที่พบว่า สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกันเพิ่มขึ้น ไม่มีผลต่อความคลาดเคลื่อนประเภทที่ 1 เมื่อพิจารณาวิธีการทดสอบวอลด์ พบว่า สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ ที่ระดับนัยสำคัญที่ .05 โดยสามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ 45% และ 11.67% ตามลำดับของเงื่อนไขทั้งหมด และมีอัตราความคลาดเคลื่อนประเภทที่ 1 เฉลี่ย 0.058 และ 0.152 ตามลำดับ ซึ่งจะเห็นว่า วิธีการทดสอบวอลด์มีอัตราความคลาดเคลื่อนประเภทที่ 1 เพิ่มขึ้น เมื่อสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกันเพิ่มขึ้น ซึ่งสอดคล้องกับการศึกษาของ Cao et al. (2017); Woods et al. (2013) และ Carroll (2015)

เมื่อพิจารณาอำนาจการทดสอบ ภายใต้สัดส่วนของการทำหน้าที่ต่างกันของข้อคำถาม 10% และ 20% พบว่า วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งสามวิธีมีผลให้อำนาจการทดสอบมีค่าแตกต่างกัน ที่ระดับนัยสำคัญ .001 โดยทั้งสามวิธีมีอำนาจการทดสอบในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามเป็นไปตามเกณฑ์ที่กำหนด ที่ระดับนัยสำคัญ .05 เมื่อพิจารณาวิธีโพลีโตมัสซิปเทสท์ พบว่า มีอำนาจการทดสอบเฉลี่ย 0.986 และ 0.976 ตามลำดับ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ มีอำนาจการทดสอบเฉลี่ย 0.999 และ 0.997 ตามลำดับ และวิธีการทดสอบวอลด์ มีอำนาจการทดสอบเฉลี่ย 0.992 และ 0.987 ตามลำดับ จะเห็นว่า เมื่อสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกันมากขึ้น ทั้งสามวิธีมีอำนาจการทดสอบลดลง ซึ่งสอดคล้องกับการศึกษาของ Kilmen (2016); Oshima and Miller (1992); Shih et al. (2014) พบว่า วิธีโพลีโตมัสซิปเทสท์มีอัตราความถูกต้องลดลง แต่ไม่สอดคล้องกับการศึกษาของสุชาติ สิริมีนนันท์ (2554); Gonzales-Roma et al. (2006) ที่พบว่า วิธีโพลีโตมัสซิปเทสท์มีอำนาจการทดสอบเพิ่มขึ้น เมื่อสัดส่วนของการทำหน้าที่ต่างกันเพิ่มขึ้น และ Kabasakal, Arsan, Gök, and Kelecioglu (2014) ที่พบว่า สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกันไม่มีผลต่ออำนาจการทดสอบด้วยวิธีโพลีโตมัสซิปเทสท์ ส่วนวิธีการวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ ผลการศึกษาสอดคล้องกับการศึกษาของ

French and Finch (2008) ที่พบว่า เมื่อสัดส่วนของการทำหน้าที่ต่างกันเพิ่มขึ้น อำนาจการทดสอบการทำหน้าที่ต่างกันมีค่าลดลง และไม่สอดคล้องกับการศึกษาของ Chang et al. (2015) ที่พบว่า เมื่อสัดส่วนของการทำหน้าที่ต่างกันเพิ่มขึ้น อำนาจการทดสอบการทำหน้าที่ต่างกันของข้อสอบมีค่าเพิ่มขึ้น และเมื่อพิจารณาวิธีการทดสอบวอลด์ พบว่า มีอำนาจการทดสอบลดลง เมื่อสัดส่วนของการทำหน้าที่ต่างกันมีค่ามากขึ้น ซึ่งไม่สอดคล้องกับการศึกษาของ Cao et al. (2017) ที่พบว่า สัดส่วนของการทำหน้าที่ต่างกันไม่มีผลต่ออำนาจการทดสอบด้วยวิธีการทดสอบวอลด์ และไม่สอดคล้องกับ Carroll (2015) ที่พบว่า เมื่อสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกันเพิ่มขึ้น อำนาจการทดสอบมีค่าเพิ่มขึ้น แต่สอดคล้องกับการศึกษาของ Woods et al. (2012) ที่พบว่า วิธีการทดสอบวอลด์มีอำนาจการทดสอบลดลง

4. ปัจจัยขนาดตัวอย่าง

ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามที่ให้คะแนนหลายค่า ภายใต้ปัจจัยขนาดตัวอย่าง 5 ขนาดได้แก่ 250 : 250, 500 : 500, 1,000 : 1,000, 250 : 500 และ 500 : 1,000 พบว่า ทั้งสามวิธีมีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกัน ที่ระดับนัยสำคัญ .001 โดยที่วิธีโพลีโตมัสซิปเทสท์สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ 1.67% สำหรับขนาดตัวอย่างที่มีอัตราส่วน 1 : 1 และขนาดตัวอย่างเล็ก และมีอัตราความคลาดเคลื่อนประเภทที่ 1 เพิ่มขึ้นเมื่อขนาดตัวอย่างมีขนาดใหญ่ขึ้น ซึ่งสอดคล้องกับการศึกษาของ Gonzales-Roma et al. (2006); Kilmen (2016) พบว่า วิธีการวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุมีขนาดของกลุ่มสนใจหรือกลุ่มอ้างอิงมีขนาดใหญ่ จะทำให้อัตราความคลาดเคลื่อนประเภทที่ 1 มีค่ามากขึ้น และสอดคล้องกับสุชาติ สิริมีนนันท์ (2554) ที่พบว่า วิธีโพลีโตมัสซิปเทสท์มีแนวโน้มว่า เมื่อขนาดตัวอย่างเพิ่มขึ้น อัตราความคลาดเคลื่อนประเภทที่ 1 จะเพิ่มสูงขึ้นด้วย เช่นเดียวกับอุทัยวรรณ สายพัฒนะ (2547) ที่พบว่า ความผิดพลาดของการตรวจสอบการทำหน้าที่ต่างกันด้วยวิธีโพลีโตมัสซิปเทสท์เมื่อกลุ่มตัวอย่างขนาดใหญ่ขึ้น ความผิดพลาดในการตรวจสอบจะมีค่ามากขึ้นด้วย ซึ่งขัดแย้งกับการศึกษาของ Narayanan and Swaminathan (1996) และทองอยู่ สาระ (2543) ที่พบว่า ขนาดกลุ่มตัวอย่างไม่มีผลกระทบต่ออัตราความคลาดเคลื่อนประเภทที่ 1 เมื่อพิจารณาวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ที่กำหนด ที่ระดับนัยสำคัญ .05 โดยสามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ 13.33%, 13.33%, 5.00%, 13.33% และ 10.00% ตามลำดับ ซึ่งผลการวิเคราะห์พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ มีอัตราความคลาดเคลื่อนประเภทที่ 1 เพิ่มขึ้น เมื่อขนาดตัวอย่างเพิ่มขึ้น ซึ่งไม่สอดคล้องกับการศึกษาของ French and Finch (2008) ที่พบว่า เมื่อขนาดตัวอย่างเพิ่มขึ้น ไม่มีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 1 และการศึกษาของ Chang et al. (2015) และ

มิ่ง เทพครเมือง (2554) ที่พบว่า วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ มีอัตราความคลาดเคลื่อนประเภทที่ 1 มีแนวโน้มลดลง เมื่อขนาดตัวอย่างเพิ่มขึ้น แต่สอดคล้องกับการศึกษาของ French and Finch (2008); Kim and Yoon (2011); Stark et al. (2006) และ ที่พบว่า เมื่อขนาดตัวอย่างเพิ่มขึ้น อัตราความคลาดเคลื่อนประเภทที่ 1 มีค่าเพิ่มขึ้น สำหรับวิธีการทดสอบวอลด์ สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ที่กำหนด ที่ระดับนัยสำคัญ .05 โดยสามารถอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ 15.00%, 13.33%, 8.33%, 13.33% และ 6.67% ตามลำดับ ซึ่งพบว่า เมื่อขนาดตัวอย่างมีขนาดใหญ่มากขึ้น อัตราความคลาดเคลื่อนประเภทที่ 1 จากการทดสอบด้วยวิธีการทดสอบวอลด์จะมีค่าเพิ่มขึ้น ซึ่งสอดคล้องกับการศึกษาของ Cao et al. (2016); Langer (2008); Woods et al. (2012) ที่พบว่า วิธีการทดสอบวอลด์มีอัตราความคลาดเคลื่อนประเภทที่ 1 เพิ่มขึ้นเมื่อขนาดตัวอย่างเพิ่มขึ้น

เมื่อพิจารณาอำนาจการทดสอบ ภายใต้ปัจจัยขนาดตัวอย่าง 5 ขนาดได้แก่ 250 : 250, 500 : 500, 1,000 : 1,000, 250 : 500 และ 500 : 1,000 พบว่า ทั้งสามวิธีมีผลให้อำนาจการทดสอบแตกต่างกันที่ระดับนัยสำคัญ .001 โดยทั้งสามวิธีมีอำนาจการทดสอบในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามเป็นไปตามเกณฑ์ที่กำหนด ที่ระดับนัยสำคัญ .05 เมื่อปัจจัยปัจจัยขนาดตัวอย่างมากขึ้น ทั้งสามวิธีมีอำนาจการทดสอบเพิ่มขึ้น ซึ่งสอดคล้องกับการศึกษาของมิ่ง เทพครเมือง (2554); Chang et al. (2015); French and Finch (2008); Gonzales-Roma et al. (2006); Kim and Yoon (2011); Stark et al. (2006) สำหรับการตรวจสอบการทำหน้าที่ต่างกันของวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ สำหรับการตรวจสอบการทำหน้าที่ต่างกันแบบไม่เป็นเอกรูป พบว่า อำนาจการทดสอบมีความสัมพันธ์กันอย่างมีนัยสำคัญกับขนาดของกลุ่มตัวอย่าง เมื่อขนาดตัวอย่างเพิ่มขึ้นอำนาจในการทดสอบเพิ่มขึ้น ซึ่งสอดคล้องกับการศึกษาของสิริรัตน์ วิภาสศิลป์ (2545); สุชาติ สิริมีนนันท์ (2554); Kilmen (2016); Narayanan and Swaminathan (1996); Rogers and Swaminathan (1993) ที่พบว่า การตรวจสอบการทำหน้าที่ต่างกันด้วยวิธีโพลีโตมัสซิปเทสท์ มีอำนาจการทดสอบเพิ่มขึ้นเมื่อขนาดตัวอย่างเพิ่มขึ้น แต่ไม่สอดคล้องกับการศึกษาของ Kabasakal et al. (2014) พบว่า ขนาดตัวอย่างไม่มีผลต่ออำนาจการทดสอบด้วยวิธีโพลีโตมัสซิปเทสท์ สำหรับวิธีการทดสอบวอลด์ สอดคล้องกับ Cao et al. (2016); Langer (2008); Woods et al. (2012) ที่พบว่า วิธีโพลีโตมัสซิปเทสท์มีอำนาจการทดสอบเพิ่มขึ้น เมื่อขนาดตัวอย่างเพิ่มมากขึ้น

ข้อเสนอแนะ

ข้อเสนอแนะในการนำผลการวิจัยไปใช้

1. การตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามพหุมิติให้คะแนนหลายค่า ภายใต้เงื่อนไขของปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ ความยาวของแบบทดสอบ 2 ขนาด ขนาดของการทำหน้าที่ต่างกัน 3 ขนาด สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 2 ขนาด และขนาดตัวอย่าง ต่างกัน 5 รูปแบบ การเลือกใช้วิธีการตรวจสอบการทำหน้าที่ต่างกันทั้งสามวิธี ผู้นำไปใช้ควรพิจารณานำไปใช้ในแต่ละเงื่อนไข เพื่อให้เกิดประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกัน

2. ถ้าผู้ใช้จะนำวิธีการตรวจสอบการทำหน้าที่ต่างกันไปใช้ในงานวิจัย ถ้าแบบทดสอบมีความยาว จำนวน 20 ข้อ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุจะมีความเหมาะสมในการตรวจสอบการทำหน้าที่ต่างกันมากกว่าอีกสองวิธี และเมื่อความยาวแบบทดสอบ จำนวน 40 ข้อ วิธีการทดสอบวอลด์จะมีความเหมาะสมในการตรวจสอบการทำหน้าที่ต่างกันมากกว่าอีกสองวิธี เพราะมีความสามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ และมีอำนาจการทดสอบสูง

3. การตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามพหุมิติที่ให้คะแนนหลายค่า ด้วยวิธีโพลีโตมัสซิปเทสท์ อาจไม่เหมาะสมกับความยาวข้อคำถาม 20 ข้อ และ 40 ข้อ ด้วยขนาดของการทำหน้าที่ต่างกันตั้งแต่ 0.40 ขึ้นไป โดยมีสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 10% และ 20% ทุกขนาดตัวอย่าง 250, 500 และ 1,000 เนื่องจากมีอัตราความคลาดเคลื่อนประเภทที่ 1 สูงเกินปกติ (Inflate)

4. สำหรับวิธีการตรวจสอบการทำหน้าที่ต่างกันทั้งสามวิธี ถ้าผู้ใช้คำนึงถึงปัจจัยสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกันสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามพหุมิติที่ให้คะแนนหลายค่า ทั้งสามวิธี คือ วิธีโพลีโตมัสซิปเทสท์ วิเคราะห์องค์ประกอบกลุ่มพหุ และวิธีการทดสอบวอลด์ สามารถตรวจพบข้อคำถามที่ทำหน้าที่ต่างกันของข้อคำถามได้ทุกเงื่อนไข แต่เมื่อพิจารณาความคลาดเคลื่อนประเภทที่ 1 พบว่า วิธีโพลีโตมัสซิปเทสท์มีความผิดพลาดสูงในทุกเงื่อนไข ส่วนวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุสามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดีเมื่อสัดส่วนข้อคำถาม 10% สำหรับความยาวข้อคำถาม 20 ข้อ แต่สำหรับความยาวข้อคำถาม 40 ข้อ วิธีการทดสอบวอลด์สามารถควบคุมความคลาดเคลื่อนได้ดี

5. ถ้าผู้นำไปใช้คำนึงถึงปัจจัยขนาดของการทำหน้าที่ต่างกันและสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน เมื่อมีขนาดตัวอย่างมากขึ้น ทำให้การตรวจสอบการทำหน้าที่ต่างกันด้วยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ มีอำนาจการทดสอบเพิ่มขึ้นเล็กน้อย และมีความคลาดเคลื่อนประเภทที่ 1 สูงขึ้น ดังนั้น ถ้าคำนึงถึงปัจจัยดังกล่าว วิธีวิเคราะห์

องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ เป็นทางเลือกหนึ่งที่จะนำไปใช้ในการตรวจสอบการทำหน้าที่ต่างกันโดยต้องมีขนาดของการทำหน้าที่ต่างกัน 0.40 และสัดส่วนข้อคำถาม 10%

6. การตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามพหุมิติที่ให้คะแนนหลายค่า ด้วยวิธีโพลีโตมัสซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ มีข้อดีข้อจำกัดแตกต่างกัน ข้อดีของวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ คือ โปรแกรมที่ใช้ในการวิเคราะห์ไม่ยุ่งยาก เวลาที่ใช้ในการวิเคราะห์ข้อมูลได้รวดเร็ว และสามารถใช้งานร่วมกับโปรแกรม R ได้ ข้อจำกัดของวิธีการวิเคราะห์องค์ประกอบกลุ่มพหุ คือ ผลการตอบข้อคำถามในทุกข้อของทั้งสองกลุ่มต้องมีรายการคำตอบทุกรายการคำตอบ ถ้าไม่เช่นนั้นแล้ว จะทำให้ไม่สามารถวิเคราะห์ผลได้ สำหรับวิธีการทดสอบวอลด์ ข้อดีคือ สามารถวิเคราะห์ได้รวดเร็ว และสามารถกำหนดค่าของกระบวนการวิเคราะห์ในคอมพิวเตอร์ได้ ข้อจำกัดคือ ใช้เวลาในการประมวลผลมากกว่าอีกสองวิธี ส่วนวิธีโพลีโตมัสซิปเทสท์ คือ โปรแกรมที่ใช้ในการวิเคราะห์ใช้งานง่าย ข้อจำกัด ไม่เหมาะกับขนาดตัวอย่างขนาดใหญ่ ความยาวแบบทดสอบจำนวนมาก และขนาดของการทำหน้าที่ต่างกันของข้อคำถามสูง

ข้อเสนอแนะในการวิจัยครั้งต่อไป

1. ควรทำการศึกษาโดยกำหนดตัวแปรอื่น ๆ เพิ่มเติม เช่น การแจกแจงความสามารถของผู้สอบ รูปแบบของการทำหน้าที่ต่างกันที่เป็นแบบเอกรูป เป็นต้น

2. ควรทำการศึกษาโดยศึกษากับเครื่องมือวัดประเภทหลายมิติ ที่มีมิติของการศึกษามากกว่า 2 มิติ หรือ แบบทดสอบที่ประกอบด้วยข้อคำถามที่ใช้โจทย์หรือสถานการณ์ร่วมกัน (Testlets)

3. ควรศึกษากลุ่มตัวอย่างที่มีมากกว่าสองกลุ่ม เช่น ใช้เกณฑ์เชื้อชาติ ภูมิลำเนา แล้วพิจารณาผลการตรวจสอบว่าแตกต่างหรือสอดคล้องกับผลการศึกษาครั้งนี้

บรรณานุกรม

ชัยยศ ชาวระนอง. (2553). *ประสิทธิภาพของโมเดลและการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบพหุมิติโดยวิธีการวิเคราะห์องค์ประกอบเชิงยืนยันแฝงภายใน*. ดุษฎีนิพนธ์ปรัชญาดุษฎีบัณฑิต, สาขาวิชาวิจัย วัดผลและสถิติการศึกษา, คณะศึกษาศาสตร์, มหาวิทยาลัยบูรพา.

ทองอยู่ สาระ. (2543). *การเปรียบเทียบอำนาจการทดสอบและการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอระหว่างวิธีแมนเทล-แฮนส์เซล และวิธีถดถอยโลจิสติก*. วิทยานิพนธ์การศึกษามหาบัณฑิต, สาขาวิชาการวัดผลการศึกษา, บัณฑิตวิทยาลัย, มหาวิทยาลัยศรีนครินทรวิโรฒ.

ธเกียรติกมล ทองงอก. (2554). *ประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในวิธีถดถอยโลจิสติก โดยใช้เกณฑ์ขนาดอิทธิพล 2 วิธี สำหรับข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบทวิภาค: ข้อมูลจำลองและข้อมูลเชิงประจักษ์*. ดุษฎีนิพนธ์ครุศาสตรดุษฎีบัณฑิต, สาขาวิชาการวัดและประเมินผลการศึกษา, บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย.

ปิยะทิพย์ ตินวร. (2549). *การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ: การเปรียบเทียบประสิทธิภาพระหว่างวิธีการวิเคราะห์องค์ประกอบจำกัดกับวิธีถดถอยโลจิสติก*. วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต, สาขาวิชาเทคโนโลยีการวัดผลทางการศึกษา, บัณฑิตวิทยาลัย, มหาวิทยาลัยบูรพา.

มิ่ง เทพครเมือง. (2554). *การตรวจสอบความเท่าเทียมกันของการวัดบนพื้นฐานทฤษฎีการทดสอบแบบคะแนนจริงสัมพันธ์และทฤษฎีการตอบข้อสอบ*. ปริญญานิพนธ์การศึกษาดุษฎีบัณฑิต, สาขาวิชาการทดสอบและวัดผลการศึกษา, บัณฑิตวิทยาลัย, มหาวิทยาลัยศรีนครินทรวิโรฒ.

วริษยา ชะม้อย. (2550). *การเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ตรวจให้คะแนนแบบพหุวิภาคระหว่างวิธีโครงสร้างความแปรปรวนร่วมและค่าเฉลี่ยกับวิธีการวิเคราะห์ฟังก์ชันเชิงจำแนกแบบโลจิสติก*. วิทยานิพนธ์ครุศาสตรมหาบัณฑิต, สาขาวิชาการวัดและประเมินผลการศึกษา, บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย.

ศิริชัย กาญจนวาสี. (2550). *ทฤษฎีการทดสอบแนวใหม่* (พิมพ์ครั้งที่ 3) กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

สิริรัตน์ วิภาสศิลป์. (2545). *การเปรียบเทียบวิธีซิบเทสท์และดีเอฟไอทีในการตรวจสอบการทำหน้าที่เบี่ยงเบน ของข้อสอบ หมวดข้อสอบและแบบทดสอบจากข้อมูลการตอบข้อสอบที่ใช้ความสามารถหลายมิติ*. ปริญญานิพนธ์การศึกษาดุษฎีบัณฑิต, สาขาวิชาการทดสอบและวัดผลการศึกษา, บัณฑิตวิทยาลัย, มหาวิทยาลัยศรีนครินทรวิโรฒ.

สุชาติ สิริมีนนันท์. (2554). *การเปรียบเทียบวิธีโพลีโตมัสซิบเทสท์ วิธีการวิเคราะห์ฟังก์ชันการจำแนกโลจิสติก และวิธีการถดถอยโลจิสติกแบบจัดอันดับ ในการตรวจสอบการทำหน้าที่เบี่ยงเบนของข้อสอบในแบบทดสอบที่มีการให้คะแนนแบบหลายค่า*. ปริญญานิพนธ์การศึกษาดุษฎีบัณฑิต, สาขาวิชาการทดสอบและวัดผลการศึกษา, บัณฑิตวิทยาลัย, มหาวิทยาลัยศรีนครินทรวิโรฒ.

สุพัฒนา หอมบุปผา. (2556). *การเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN*. ดุษฎีนิพนธ์ปรัชญาดุษฎีบัณฑิต, สาขาวิชาวิจัย วัดผลและสถิติการศึกษา, คณะศึกษาศาสตร์, มหาวิทยาลัยบูรพา

อรินทร์ น่วมถนอม. (2549). *การเปรียบเทียบวิธีโพลี-ซิปเทสท์ วิธีการถดถอยโลจิสติกแบบจัดอันดับและวิธีการถดถอยโลจิสติกแบบจัดอันดับหลายมิติ ในการตรวจสอบการทำหน้าที่เบี่ยงเบนของข้อสอบที่วัดความสามารถหลายมิติและให้คะแนนหลายค่า*. ปริญญานิพนธ์การศึกษาดุษฎีบัณฑิต, สาขาวิชาการทดสอบและวัดผลการศึกษา, บัณฑิตวิทยาลัย, มหาวิทยาลัยศรีนครินทรวิโรฒ.

อาวีพร ปานทอง. (2558). *การเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบให้คะแนนหลายค่าโดยวิธีทดสอบอัตราส่วนความควรจะเป็นวิธีเบส์เซียน และวิธีโพลี-ซิปเทสท์*. ดุษฎีนิพนธ์ปรัชญาดุษฎีบัณฑิต, สาขาวิชาวิจัย วัดผลและสถิติการศึกษา, คณะศึกษาศาสตร์, มหาวิทยาลัยบูรพา.

อิทธิฤทธิ์ พงษ์ปิยะรัตน์. (2551). *การวิเคราะห์ข้อสอบและการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ: การวิเคราะห์พหุระดับ*. ดุษฎีนิพนธ์ครุศาสตรดุษฎีบัณฑิต, สาขาวิชาการวัดและประเมินผลการศึกษา, บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย.

อุทัยวรรณ สายพัฒนะ. (2547). *การเปรียบเทียบประสิทธิภาพของผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ในแบบทดสอบที่มีการให้คะแนนแบบหลายค่าระหว่างวิธี GHM และวิธี Polytomous-SIBTEST*. ปริญญานิพนธ์การศึกษาดุษฎีบัณฑิต. สาขาวิชาการทดสอบและวัดผลการศึกษา, บัณฑิตวิทยาลัย, มหาวิทยาลัยศรีนครินทรวิโรฒ.

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*(1), 67-91.

Adams, R. J. (2003). Response to "Cautions on OECD's recent educational survey (PISA)". *Oxford Review of Education, 29*(3), 379-389.

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1-24.

Adams, R. J., Wu, M. L., & Carstensen, C. H. (2007). *Application of multivariate Rash models in international large-scale educational assessments: Statistics for social and behavioral sciences*. New York: Springer.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Angoff, W. H. (1972). *A technique for the investigation of cultural differences*. Paper presented at the annual meeting of the American Psychological Association. Honolulu, HI.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Hillside: Lawrence Erlbaum Associates.

Atar, B., & Kamata, A. (2011). Comparison of IRT likelihood ratio test and logistic regression DIF detect procedures. *Hacettepe University Journal of Education, 41*, 36-47.

Ayala, R. J. (2008). *The theory and practice of item response theory*. New York: Guilford Press.

Bagozzi, R. P., & Edwards, J. R. (1998). A general approach for representing constructs in organizational research. *Organizational Research Methods, 1*, 45-87.

Bock, R. D., & Schilling, S. G. (2003). IRT based item factor analysis. In M. du Toit (Ed.), *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT* (pp. 584-591). Lincolnwood, IL: Scientific Software International.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bollen, K. A., & Long, J. S. (Eds.) (1993). *Testing structural equation models*. Newbury Park, CA: Sage.

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 2*, 113-141.

Bolt, D. M., & Gierl, M. J. (2006). Testing features of graphical DIF: Application of a regression correlation three nonparametric statistical tests. *Journal of Educational Measurement*, *43*(4), 313-333.

Bolt, D. M., Hare, R. D., Vitale, J. E., & Newman, J. P. (2004). A multigroup item response theory analysis of the psychopathy checklist-revised. *Psychological Assessment, 16*, 155-168.

Byrne, B. M., Shavelson, R. J., & Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456-466.

Cai, L. (2008). SEM of another flavor: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology, 61*, 309-329.

Cai, L. (2012). *flexMIRT*: *Flexible multilevel item factor analysis and test scoring [Computer software]*. Seattle, WA: Vector Psychometric Group, LLC.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]*. Lincolnwood, IL: Scientific Software International.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Cao, M., Tay, L., & Liu, Y. (2017). A Monte Carlo Study of an iterative Wald test procedure for DIF analysis. *Educational and Psychological measurement, 77*(1), 104-118.

Cardall, C., & Coffman, W. E. (1964). *A method for comparing the performance of different groups on the same items of a test*. Princeton, NJ: Educational Testing Service.

Carroll, H. F. C. (2015). *An examination of the improved Wald test for differential item functioning detection with multiple groups*. Doctoral dissertation, Educational Psychology, University of Kansas.

Chang, Y., Huang, W., & Tsai, R. (2015). DIF detection using multiple-group categorical CFA with minimum free baseline approach. *Journal of Educational Measurement, 52*(2), 181-199.

Chang, H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33*(3), 333-353.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233-255.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*(1), 31-44.

Cohen, A. S., Kim, S. H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17*(4), 335-350.

Cohen, A.S., & Kim, S. H. (1993). A comparison of Lord's chi-square and Raju's area measures in detection of DIF. *Applied Psychological Measurement, 17*, 39-52.

Cronbach, L. J. (1990). *Essentials of psychology testing* (5th ed.). New York: Harper Collins.

DeMars, C. (2010). *Item response theory: Understanding statistics measurement*. Oxford: Oxford University Press.

Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement, 22*, 33-51.

Dorans, N, J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N, J., & Kulick, E. (1986). Demonstrating the unity of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. *Journal of Educational Measurement, 23*(4), 355-368.

Dorans, N. J., & Schmitt, A. J. (1991). *Constructed response and differential item functioning: A pragmatic approach*. Princeton, NJ: Educational Testing Service.

Elosua, P., & López-Jáuregui, A. (2007). Application of four procedures for detecting differential item functioning in polytomous items. *Psicothema, 19*(2), 329-336.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. New Jersey: Lawrence Erlbaum Associates.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST and the IRT likelihood ratio test. *Applied Psychological Measurement, 29*, 278-295.

Finch, H., & French, B. F. (2007). Detection of crossing differential item functioning item: A comparison of four methods. *Educational and Psychological Measurement, 67*, 565-582.

Fleer, P. F. (1993). A Monte Carlo assessment of a new measure of item and test bias. Illinois Institute of Technology. *Dissertation Abstracts International, 54*(04B), 2266.

Flowers, C. P., Oshima, T. C., & Raju, N. S. (1999). A description and demonstration of the polytomous-DFIT framework. *Applied Psychological Measurement, 23*, 309-326.

French, B. F., & Finch, W. H. (2008). Multi-group confirmatory factor analysis: Locating the invariant referent. *Structural Equation Modeling, 15*, 96-113.

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33*(3), 315-332.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement, 67*, 373-393.

Gierl, M. J., Gotzmann, A., & Boughton, K. A. (2004). Performance of SIBTEST when the percentage of DIF Items is large. *Applied Measurement in Education, 17*(3), 241-264.

Gómez-Benito, J., Hidalgo, M. D., & Padilla, J. L. (2009). Efficacy of effect size measures in logistic regression: An application for detecting DIF. *Methodology: European Journal of Research Methods for The Behavior and Social Sciences, 5*(1), 18-25.

Gonzales-Roma, V., Hernandez, A., & Gomez-Benito, J. (2006). Power and type-I error of the mean and covariance structure analysis model for detecting differential item functioning in graded response items. *Multivariate Behavioral Research, 41*(1), 29-53.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel Haenszel procedure. In H. Wainer, & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*, 117-144.

Howell, D. C. (2010). *Statistical methods for psychology* (7th ed.). Belmont, CA: Wadsworth Press.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Applications to psychological measurement*. Homewood, IL: Dow Jones-Irwin.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*, 329-349.

Jöreskog, K. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*, 409-426.

Kabasakal, K. A., Arsan, N., Gök, B., & Kelecioglu, H. (2014). Comparing Performances (Type I Error and Power) of IRT Likelihood Ratio SIBTEST and Mantel-Haenszel Methods in the Determination of Differential Item Functioning. *Educational Sciences: Theory and Practice, 14*(6), 2186-2193.

Kahraman, N., De Boeck, P., & Janssen, R. (2009). Modeling DIF in complex response data using test design strategies. *International Journal of Testing, 9*, 151-166.

Kannan, P. (2011). *Comparing DIF detection for multidimensional polytomous models using multi group confirmatory factor analysis and the differential functioning of items and tests*. Doctoral dissertation, University of Pittsburgh. (Unpublished).

Kannan, P., & Kim, K. H. (2009). *Item parameter recovery for a within-item multidimensional graded response model: A SEM-CFA perspective*. Paper presented at the Annual Meeting of the National Council on Measurement in Education. San Diego, CA: The National Council on Measurement in Education

Kannan, P., & Ye, F. (2008). *Item parameter recovery for a between-item multidimensional graded response model*. Paper presented at the annual meeting of the National Council on Measurement in Education. New York.

Kelderman, H., & Rijkes, C. P. M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika, 59*, 149-176.

Kilmen, S. (2016). Effect of DIF magnitudes, focal group sample size, and DIF ratio on the *performance of SIBTEST. International Journal of Social Sciences and Education, 6*(1), 91-97.

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling, 18*, 212-228.

Kim, S. H., & Cohen, A. S. (1991). A comparison of two area measures for detecting differential item functioning. *Applied Psychological Measurement, 15*(3), 269-278.

Kim, S. H., Cohen, A. S., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Education Measurement, 44*(2), 93-116.

Kim, S. H., Cohen, A. S., & Park, T. H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement, 32*(3), 261-276.

Klockars, A. J., & Lee, Y. (2008). Simulated tests of differential item functioning using SIBTEST with and without impact. *Journal of Educational Measurement, 45*, 271-285.

Kunnan, A. J. (2000). *Fairness and validation in language assessment: Selected papers from the 19th language testing research colloquium*. Orlando, FL: The University of Cambridge.

Langer, M. (2008). *A reexamination of Lord's Wald test for differential item functioning using item response theory and modern error estimation*. North Carolina: University of North Carolina, Chapel Hill.

Le, L. T. (2009). Investigating gender differential item functioning across countries and test languages for PISA Science items. *International Journal of Testing, 9*(2), 122-133. doi: 10.1080/15305050902880769

Li, H., & Stout, W. (1996). A new procedure for detecting crossing DIF. *Psychometrika, 61*(4), 647-677.

Lim, R. G., & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. *Journal of Applied Psychology, 75*, 164-174.

Linn, R. L., & Hamisch, D. L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18*, 109-118.

Linn, R. L., Levin, M. V., Hastings, C. N., & Wardrop, J. L. (1981). An investigation of item bias in a test of reading comprehension. *Applied Psychological Measurement, 5*, 159-173.

Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19-29). Amsterdam: Swets and Zeitlinger.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003). On the relationship between sources of within-group and between-group differences and measurement invariance in the context of the common factor model. *Intelligence, 31*, 543-566.

Lubke, G. H., & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling, 11*, 514-534.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*, 84-99. doi: 10.1037/1082-989X.4.1.84.

Maller, S. J., & French, B. F. (2004). Factor invariance of the UNIT across deaf and standardization samples. *Educational and Psychological Measurement, 64*, 647-660.

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Mazor, K. M., Hambleton, R. K., & Clauser, B. E. (1998). Multidimensional DIF analyses: The effects of matching on unidimensional subtest score. *Applied Psychological Measurement, 22*, 357-367.

McDonald, R. P. (1999) *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

McKinley, R. L., & Reckase, M. D. (1982). *The use of the general Rasch model with multidimensional item response data (Research Report ONR 82-1)*. Iowa City, IA: American College Testing.

McLaughlin, M. E., & Drasgow, F. (1987). Lord's chi-square test of item bias with estimated and with known person parameters. *Applied Psychological Measurement, 11*, 161-173.

McLeod, L. D., Swygert, K. A., & Thissen, D. (2001). Factor analysis for items scored in two categories. In D. Thissen, & H. Wainer (Eds.), *Test scoring* (pp. 189-216). Hillsdale, NJ: Lawrence Erlbaum Associates.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/ invariance. *Organizational Research Methods, 7*(4), 361-388.

Meade, A. W., & Lautenschlager, G. J. (2005). *Sensitivity of DFIT tests of measurement invariance for Likert data*. Paper presented at the 20th annual meeting of the Society for Industrial/ Organizational Psychology. Los Angeles, CA.

Meade, A. W., Lautenschlager, G. J., & Johnson, E. C. (2006). *Alternative cutoff values and DFIT tests of measurement invariance*. Paper presented at the 21st annual conference of the Society for Industrial and Organizational Psychology. Dallas, TX.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*(2), 105-118.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement, 30*(2), 107-122.

Millsap, R. E. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivariate Behavioral Research, 30*, 577-605.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*, 297-334.

Mulaik, S. A. (1972). *A mathematical investigation of some multidimensional Rasch models for psychological tests*. Paper presented at the annual meeting of the Psychometric Society. Princeton, NJ.

Muraki, E. (1992). A generalized partial credit model: Application of an EM Algorithm. *Applied Psychological Measurement, 16*(2), 159-176.

Muraki. E, & Carlson, J. E. (1993). *Full-information factor analysis for polytomous item responses*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*, 557-585.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement, 30*(4), 293-311.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item biasprocedures for detecting differential item functioning. *Applied Psychological Measurement, 20*, 315-338.

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20*, 257-274.

Oishi, S. (2006). The concept of life satisfaction across cultures: An IRT analysis. *Journal of Research in Personality, 40*(4), 411-423.

Oshima, T. C., & Miller, M. D. (1992). Multidimensionality and item bias in item response theory. *Applied Psychological Measurement, 16*(3), 237-248.

Oshima, T. C., Raju, N. S., & Flowers, C. P. (1997). Development and demonstration of multidimensional IRT-based internal measures of differential functioning of items and tests. *Journal of Educational Measurement, 34*(3), 253-272.

Paek, I., & Wilson, M. (2011). Formulating the Rasch differential item functioning model under the marginal maximum likelihood estimation context and its comparison with Mantel-Haenszel procedure in short test and small sample conditions. *Educational and Psychological Measurement, 71*, 1023-1046.

Park, D. G., & Lautenschlager, G. J. (1990). Improving IRT item bias detection with iterative linking and ability scale purification. *Applied Psychological Measurement, 14*, 163-173.

Penfield, R. D. (2007). Assessing differential step functioning in polytomous items using a common odds ratio estimator. *Journal of Educational Measurement, 44*(3), 187-210.

Penfield, R. D. (2008). Three Classes of Nonparametric Differential Step Functioning Effect Estimators. *Applied Psychological Measurement, 32*(6), 480-591. doi: 10.1177/0146621607305399

Penfield, R. D. (2010). Distinguishing between net and global DIF in polytomous items. *Journal of Educational Measurement, 47*(2), 129-149.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19*, 23-37.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, 53, 495-502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*(2), 197-207.

Raju, N. S., & Ellis, B. B. (2002). Differential item and test functioning. In F. Drasgow, & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations* (pp. 156-188). San Francisco: Jossey-Bass.

Raju, N. S., Laffitte, L. J., & Byne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*(3), 517-529.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measure of differential functioning of items and tests. *Applied Psychological Measurement, 19*, 353-368.

Reckase, M. D. (1972). *Development and application of a multivariate logistic latent trait model*. Unpublished doctoral dissertation, Syracuse University, Syracuse, NY.

Reckase, M. D. (1977). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*(1), 25-36.

Reckase, M. D. (1985).The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*, 401-412.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552-566.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of the logistic regression and Mantel-Haenszel procedures for detecting dierential item functioning. *Applied Psychological Measurement, 17*(2), 105-116.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*(4), 355-371.

Rudner, L. M. (1977). *An evaluation of select approaches for biased item identification*. Unpublished doctoral dissertation, Catholic University of America, Washing DC.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*(17).

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika, 39*, 111-121.

Scheuneman, J. D. (1979). A method of assessing bias in test items. *Journal of Educational Measurement, 16*(3), 143-152.

Scheuneman, J. D. (1981). A response to Baker's criticism. *Journal of Education Measurement, 18*(1), 63-66.

Schmitt, N. (1982). The use of analysis of covariance structures to assess beta and gamma change. *Multivariate Behavioral Research, 17*, 343-358.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/ DIF from group ability differences and detects test bias/ DIF as well as item bias/ DIF. *Psychometrika*, 58, 159-194.

Shepard, L. A., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics, 9*(2), 93-128.

Shih, C. L., Liu, T. H., & Wang, W. C. (2014). Controlling type I error rates in qssessing DIF for logistic regression method combined with SIBTEST regression correction procedure and DIF-Free-Then-DIF strategy. *Educational and Psychological Measurement*, *74*(6), 1018–1048. doi: org/10.1177/0013164413520545

Snow, T. K., & Oshima, T. C. (2009). A comparison of unidimensional and three-dimensional differential item functioning analysis using two-dimensional data. *Educational and Psychological Measurement, 69*, 732-747.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structures between groups. *British Journal of Mathematical and Statistical Psychology*, *27*, 229-239.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292-1306.

Stout, W., Li, H., Nandakumar, R., & Bold, D. (1997). MULTISIB: A procedure to investigate DIF when a test is intentionally two-dimensional. *Applied Psychological Measurement, 21*(3), 195-213.

Su, Y., & Wang, W. (2005). Efficiency of the Mantel, generalized Mantel-Haenszel, and logistic discriminant function analysis methods in detecting differential item functioning for polytomous items. *Applied Measurement in Education, 18*, 313-350.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370.

Swygert, K. A.,, McLeod, L. D., & Thissen, D. (2001). *Factor analysis for items or testlets scored in more than two categories*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 computerized adaptive testing conference* (pp. 82-98). Minneapolis: University of Minnesota.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*(3), 393-408.

Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods, 18*, 3-46. doi: 10.1177/1094428114553062.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In W. P. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tian, F. (1999). *Detecting DIF in polytomous items*. Unpublished doctoral dissertation, Faculty of Education, University of Ottawa.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-70.

Walker, C. M., & Beretvas, S. N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: A cognitive explanation for DIF. *Journal of Educational Measurement, 38*, 147-163.

Walker, C. M., Zhang, B., & Surber, J. (2008). Using a multidimensional differential item functioning framework to determine if reading ability affects student performance in mathematics. *Applied Measurement in Education, 21*(2), 162-181.

Wang, W. C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *The Journal of Experimental Education, 72*(3), 221-261.

Wang, W. C., & Su, Y. Y. (2004). Factors influencing the Mantel and generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement, 28*, 450-480.

Wetzel, E., Carstensen, C. H., & Böhnke, J. R. (2013). Consistency of extreme response style and non-extreme response style across traits. *Journal of Research in Personality, 47*(2), 178-189.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45*, 479-494.

Whitely, S. E. (1981). *Measuring aptitude processes with multicomponent latent trait models*. *Journal of Educational Measurement, 18*, 67-84.

Wiberg, M. (2009). Differential item functioning in mastery tests: A comparison of three methods using real data. *International Journal of Testing, 9*(1), 41-59.

Williams, N. J., & Beretvas, S. N. (2006). DIF identification using HGLM for polytomous items. *Applied Psychological Measurement, 30*, 22-42.

Wilson, M., & Hoskens, M. (2005). Multidimensional item response: Multimethod/ multitrait perspective. In S. Alagumalai, D. D. Curtis, & N. Hungi (Eds.), *Applied Rasch measurement: A Book exemplars* (pp. 287-307). Netherlands: Springer.

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement, 33*(1), 42-57.

Woods, C. M., Cai, L., & Wang, M. (2013). The langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement, 73*(3), 532-547.

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement, 35*, 339-361.

Wu, Q., & Lei, P. W. (2009). *Using multi-group confirmatory factor analysis to detect differential item functioning when tests are multidimensional*. Paper presented at the Annual Meeting of the National Council for Measurement in Education, San Diego: CA.

Yao, L., & Li, F. (2010). *A DIF detection procedure in multidimensional item response theory framework and its applications*. Paper presented at the 2010, Annual Meeting of the National Council on Measurement in Education, Colorado, Denver.

Yao, L., & Schwarz, R. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed format tests. *Applied Psychological Measurement, 30*, 469-492.

Zumbo, B. D. (1999). *A handbook of the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item score*. Ottawa, Ontario, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measure*ment*, 30*(3), 233-251.

ประวัติย่อของผู้วิจัย

ชื่อ-สกุล	นางสาวณัฐพร ภักดี
วัน เดือน ปีเกิด	15 พฤศจิกายน พ.ศ. 2524
สถานที่เกิด	จังหวัดสุโขทัย
สถานที่อยู่ปัจจุบัน	บ้านเลขที่ 67 ถนนเทศบาลพัฒนา 1 ตำบลเหมือง อำเภอเมือง จังหวัดชลบุรี
ตำแหน่งและประวัติการทำงาน	
พ.ศ. 2552-2557	อาจารย์ประจำหลักสูตรสาขาวิชาสถิติประยุกต์ คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏจันทรเกษม
พ.ศ. 2558-ปัจจุบัน	นักวิจัย เขตอุตสาหกรรมซอฟต์แวร์ภาคตะวันออก คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา
ประวัติการศึกษา	
พ.ศ. 2547	ศึกษาศาสตรบัณฑิต (คณิตศาสตร์) มหาวิทยาลัยบูรพา
พ.ศ. 2552	วิทยาศาสตรมหาบัณฑิต (สถิติประยุกต์) สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ. 2560	ปรัชญาดุษฎีบัณฑิต (วิจัย วัดผล และสถิติการศึกษา) มหาวิทยาลัยบูรพา