บทที่ 3
วิธีดำเนินการวิจัย

ในการศึกษาครั้งนี้ ผู้วิจัยจำลองข้อมูลภายใต้ทฤษฎีการตอบสนองข้อสอบ โดยใช้โมเดลโลจิสติกแบบ 3 พารามิเตอร์ชนิดกำหนดค่า c คงที่ (3PLM-c) แล้วจัดกระทำข้อมูลตามปัจจัยที่ศึกษา 4 ตัว คือ (1) ลักษณะของข้อสอบ 9 ลักษณะ ประกอบด้วย a ต่ำกับ b ต่ำ, a ต่ำกับ b ปานกลาง, a ต่ำกับ b สูง, a ปานกลางกับ b ต่ำ, a ปานกลางกับ b ปานกลาง, a ปานกลางกับ b สูง, a สูงกับ b ต่ำ, a สูงกับ b ปานกลาง และ a สูงกับ b สูง (2) ความยาวของแบบสอบ 2 ระดับ ประกอบด้วย 30 ข้อ และ 60 ข้อ (3) สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน 3 ระดับ ประกอบด้วย 5%, 10% และ 20% (4) ขนาดกลุ่มตัวอย่าง 6 ระดับ ประกอบด้วย จำนวนผู้สอบกลุ่มอ้างอิงต่อจำนวนผู้สอบกลุ่มเปรียบเทียบเท่ากับ 250 คนต่อ 250 คน, 500 คนต่อ 250 คน, 500 คนต่อ 500 คน, 1000 คนต่อ 250 คน, 1000 คนต่อ 500 คน และ 1000 คนต่อ 1000 คน รวมข้อมูลที่จัดกระทำทั้งหมด 324 เงื่อนไข ($9 \times 2 \times 3 \times 6$) ต่อจากนั้นจึงนำข้อมูลดังกล่าวมาวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปด้วยวิธีซิปเทสท์ปรับใหม่ วิธีซิปเทสท์ วิธีแมนเทล-แฮนส์เซล และวิธีการถดถอยโลจิสติก เพื่อเปรียบเทียบความถูกต้องแม่นยำระหว่างวิธีการตรวจสอบ 4 วิธีดังกล่าว โดยพิจารณาจากค่าอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 สำหรับวิธีดำเนินการวิจัยแบ่งออกเป็น 5 ตอน ดังนี้

ตอนที่ 1 การจำลองข้อมูล

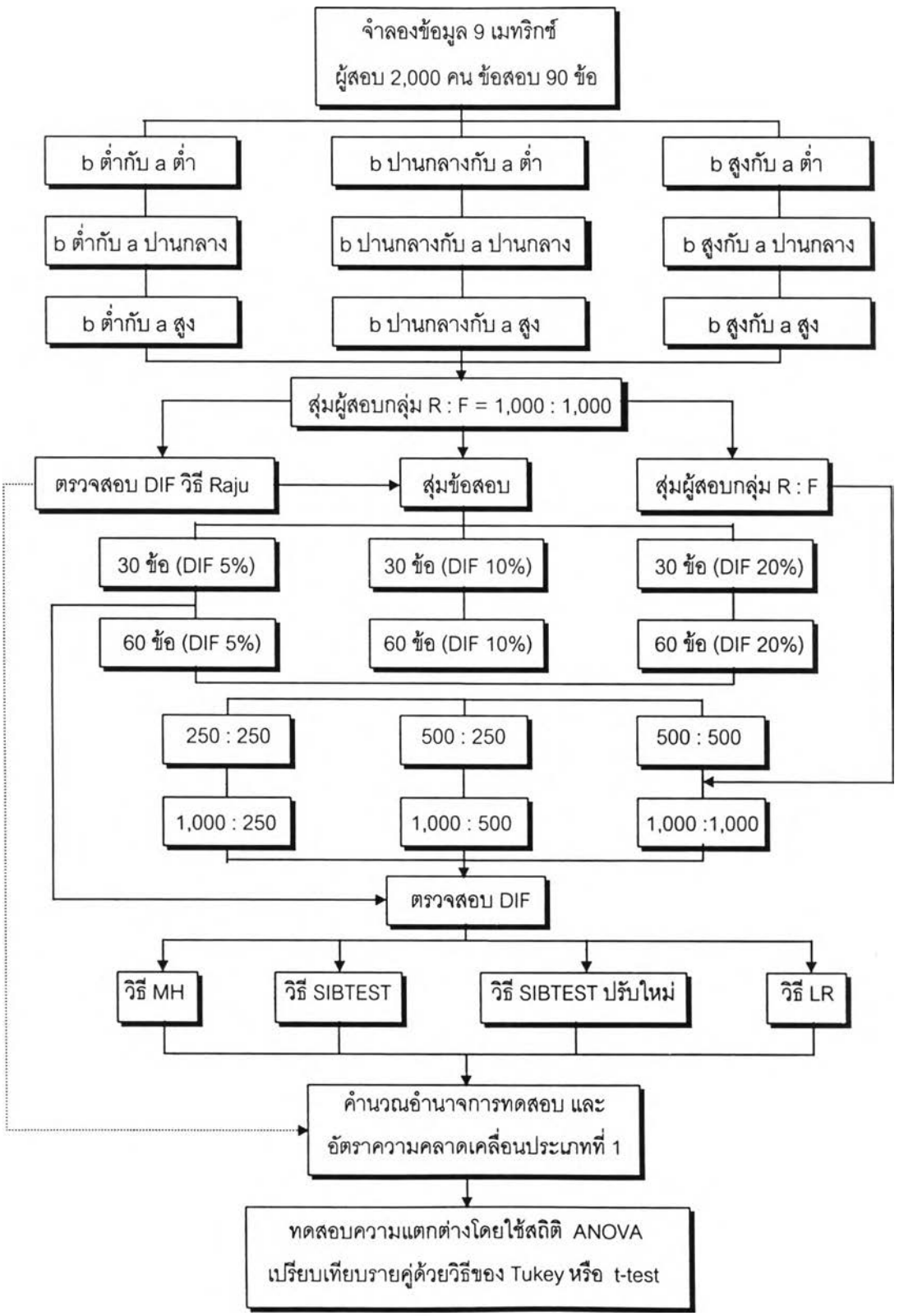
ตอนที่ 2 การจัดกระทำข้อมูลตามปัจจัยที่ศึกษา

ตอนที่ 3 การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธีซิปเทสท์ วิธีซิปเทสท์ปรับใหม่ วิธีแมนเทล-แฮนส์เซล และวิธีการถดถอยโลจิสติก

ตอนที่ 4 การวิเคราะห์อำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีซิปเทสท์ปรับใหม่ วิธีซิปเทสท์ วิธีแมนเทล-แฮนส์เซลและวิธีการถดถอยโลจิสติก

ตอนที่ 5 การทดสอบความแตกต่างของอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ระหว่างวิธีซิปเทสท์ปรับใหม่ วิธีซิปเทสท์ วิธีแมนเทล-แฮนส์เซล และวิธีการถดถอยโลจิสติก

ผู้วิจัยดำเนินการวิจัยตามแผนภาพดังต่อไปนี้



แผนภาพที่ 3 ขั้นตอนการดำเนินการวิจัย

การดำเนินการวิจัยทั้ง 5 ตอน มีรายละเอียดดังนี้

ตอนที่ 1 การจำลองข้อมูล

ผู้วิจัยศึกษาในสถานการณ์จำลอง โดยใช้โปรแกรม IRTDATA (Johanson,1992) จำลองข้อมูลภายใต้ทฤษฎี IRT โมเดลโลจิสติกแบบ 3 พารามิเตอร์ชนิดกำหนดค่า c คงที่ (3PLM-c) โมเดลดังกล่าวมีลักษณะนี้

$$P_i(\theta) = c_i + (1-c_i)\left\{1+\exp\left[-Da_i(\theta-b_i)\right]\right\}^{-1}$$

เมื่อ $P_i(\theta)$ แทน โอกาสของผู้สอบซึ่งมีระดับความสามารถ θ จะตอบข้อสอบข้อที่ i ได้ถูกต้อง

θ แทน ค่าความสามารถของผู้สอบ

a_i แทน ค่าอำนาจจำแนกของข้อสอบข้อที่ i

b_i แทน ค่าความยากของข้อสอบข้อที่ i

c_i แทน ค่าการเดาของข้อสอบข้อที่ i

D แทน ค่าองค์ประกอบของการปรับสเกล ซึ่งมีค่าเท่ากับ 1.7

exp แทน ค่าคงที่ของลอการิทึมธรรมชาติ ซึ่งมีค่าประมาณ 2.71818...

โปรแกรม IRTDATA ใช้ในการจำลองข้อมูลที่มีการให้คะแนนแบบทวิภาค ซึ่งสามารถจำลองข้อมูลโดยใช้โมเดลโลจิสติกแบบ 1, 2 หรือ 3 พารามิเตอร์ สมรรถนะสูงสุดในการจำลองผู้สอบไม่เกิน 3,000 คน และข้อสอบไม่เกิน 200 ข้อ โปรแกรมดังกล่าวประกอบด้วยโปรแกรมย่อย 2 โปรแกรม คือ โปรแกรมย่อย IRTDATA 1 ใช้ในการกำหนดคำสั่งเพื่อเตรียมข้อมูลในการจำลอง แล้วนำข้อมูลที่กำหนดป้อนเข้าสู่โปรแกรมย่อย IRTDATA 2 ผลการจำลองจะได้แฟ้มข้อมูล 4 แฟ้ม ดังนี้

(1) temp.irt เป็นแฟ้มคำสั่งที่ใช้กำหนดลักษณะของข้อมูล ได้แก่ การแจกแจงของข้อมูล (uniform หรือ normal) ค่าองค์ประกอบของการปรับสเกล (D) จำนวนรอบในการจำลองข้อมูล ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของค่าพารามิเตอร์ของข้อสอบ (a, b และ c) ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) จำนวนผู้สอบและจำนวนข้อสอบ

(2) temp.itm เป็นแฟ้มแสดงค่าพารามิเตอร์ของข้อสอบ ซึ่งประกอบด้วยค่าความยาก (b) ค่าอำนาจจำแนก (a) และค่าการเดา (c)

(3) temp.per เป็นแฟ้มแสดงค่าพารามิเตอร์ความสามารถของผู้สอบ ซึ่งประกอบด้วยคะแนนจริง (true score) ระดับความสามารถ (θ) และจำนวนข้อสอบที่ตอบถูก

(4) temp.mat เป็นแฟ้มแสดงเมทริกซ์คำตอบของผู้สอบทุกคน ซึ่งเป็นคะแนนดิบ (0-1) ของการตอบข้อสอบในแต่ละข้อ

ผู้วิจัยจำลองเมทริกซ์คำตอบจำนวน 9 เมทริกซ์ โมเดลโลจิสติกแบบ 3 พารามิเตอร์ซนิด กำหนดค่า c คงที่ โดยกำหนดการแจกแจงของค่าพารามิเตอร์แบบปกติ ค่า D เท่ากับ 1.7 จำนวนผู้สอบ 2,000 คน จำนวนข้อสอบ 90 ข้อ ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของค่าพารามิเตอร์ในแต่ละเมทริกซ์ ผู้วิจัยกำหนดดังตารางที่ 4

ตารางที่ 4 ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของค่าอำนาจจำแนก (a) ค่าความยาก (b) ค่าการเดา (c) และค่าความสามารถ (θ) ของเมทริกซ์ผลการตอบข้อสอบ

เมทริกซ์ที่	$\bar{a}$	$\bar{b}$	$\bar{c}$	$\bar{\theta}$
1	0.83	-1.34	0.20	0.00
2	0.83	0.00	0.20	0.00
3	0.83	+1.34	0.20	0.00
4	1.30	-1.34	0.20	0.00
5	1.30	0.00	0.20	0.00
6	1.30	+1.34	0.20	0.00
7	1.77	-1.34	0.20	0.00
8	1.77	0.00	0.20	0.00
9	1.77	+1.34	0.20	0.00
SD	0.077	0.222	0.000	1.000

ค่า SD ของค่าพารามิเตอร์ของข้อสอบและผู้สอบดังกล่าวจะมีผลทำให้ค่าพารามิเตอร์ของข้อสอบที่ได้จากการจำลองมีค่าอยู่ในช่วงที่ต้องการศึกษา ซึ่งผู้วิจัยคำนวณคร่าว ๆ โดยใช้สูตรดังนี้

$$SD = \frac{X_{\max} - X_{\min}}{6}$$

โดยที่ $X_{\max}$ แทน ค่าพารามิเตอร์ของข้อสอบที่มีค่าสูงสุด

$X_{\min}$ แทน ค่าพารามิเตอร์ของข้อสอบที่มีค่าต่ำสุด

ผลการจำลองข้อมูล 9 เมทริกซ์ ได้ค่าพารามิเตอร์ของข้อสอบและค่าพารามิเตอร์ของผู้สอบสอดคล้องกับลักษณะของข้อสอบและผู้สอบที่ต้องการศึกษา กล่าวคือ ค่าการเดา (c) ของข้อสอบทุกข้อมีค่าเท่ากับ .20 ค่าอำนาจจำแนก (a) มีค่าอยู่ในช่วงต่ำ ($0.60 \leq a \leq 1.06$) ช่วงปานกลาง ($1.11 \leq a \leq 1.50$) และช่วงสูง ($1.58 \leq a \leq 1.97$) ค่าความยาก (b) มีค่าอยู่ในช่วงต่ำ ($-1.93 \leq b \leq -0.67$) ช่วงปานกลาง ($-0.59 \leq b \leq 0.64$) และช่วงสูง ($0.76 \leq b \leq 1.91$) ค่าพารามิเตอร์ของผู้สอบ (θ) มีค่าอยู่ในช่วง $-2.98 \leq \theta \leq 3.05$ และคะแนนผลการตอบข้อสอบ (X) มีค่าอยู่ในช่วง $7 \leq X \leq 90$ สำหรับรายละเอียดของผลการจำลองค่าพารามิเตอร์ของข้อสอบและคะแนนผลการตอบข้อสอบ แสดงดังตารางที่ 65 – 82 ในภาคผนวก ก หน้า 299 – 312

ตอนที่ 2 การจัดกระทำข้อมูลตามปัจจัยที่ศึกษา

ในการศึกษาครั้งนี้ ผู้วิจัยจัดกระทำข้อมูลตามปัจจัยที่ศึกษา 4 ตัว ดังนี้

1. ลักษณะของข้อสอบ 9 ลักษณะ
   - 1.1 ค่า a ต่ำกับ b ต่ำ ($\bar{a}$ = 0.83 กับ $\bar{b}$ = –1.34)
   - 1.2 ค่า a ต่ำกับ b ปานกลาง ($\bar{a}$ = 0.83 กับ $\bar{b}$ = 0.00)
   - 1.3 ค่า a ต่ำกับ b สูง ($\bar{a}$ = 0.83 กับ $\bar{b}$ = +1.34)
   - 1.4 ค่า a ปานกลางกับ b ต่ำ ($\bar{a}$ = 1.30 กับ $\bar{b}$ = –1.34)
   - 1.5 ค่า a ปานกลางกับ b ปานกลาง ($\bar{a}$ = 1.30 กับ $\bar{b}$ = 0.00)
   - 1.6 ค่า a ปานกลางกับ b สูง ($\bar{a}$ = 1.30 กับ $\bar{b}$ = +1.34)
   - 1.7 ค่า a สูงกับ b ต่ำ ($\bar{a}$ = 1.77 กับ $\bar{b}$ = –1.34)
   - 1.8 ค่า a สูงกับ b ปานกลาง ($\bar{a}$ = 1.77 กับ $\bar{b}$ = 0.00)
   - 1.9 ค่า a สูงกับ b สูง ($\bar{a}$ = 1.77 กับ $\bar{b}$ = +1.34)
2. ความยาวของแบบสอบ 2 ระดับ
   - 2.1 แบบสอบที่มีจำนวน 30 ข้อ
   - 2.2 แบบสอบที่มีจำนวน 60 ข้อ

3. สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน 3 ระดับ
   3.1 มีข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบจำนวน 5 %
   3.2 มีข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบจำนวน 10 %
   3.3 มีข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบจำนวน 20 %
4. ขนาดกลุ่มตัวอย่าง 6 ระดับ
   4.1 กลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ เท่ากับ 250 คนต่อ 250 คน
   4.2 กลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ เท่ากับ 500 คนต่อ 250 คน
   4.3 กลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ เท่ากับ 500 คนต่อ 500 คน
   4.4 กลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ เท่ากับ 1,000 คนต่อ 250 คน
   4.5 กลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ เท่ากับ 1,000 คนต่อ 500 คน
   4.6 กลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ เท่ากับ 1,000 คนต่อ 1,000 คน

การจัดกระทำข้อมูลมีรายละเอียดดังนี้

1. ใช้โปรแกรม SPSS for windows version 7.52 สุ่มขนาดกลุ่มตัวอย่างในแต่ละเมทริกซ์คำตอบของแฟ้ม temp.mat ที่ได้จากการจำลองข้อมูล โดยสุ่มอย่างง่ายแบบไม่ใส่คืนเพื่อให้ได้ขนาดของผู้สอบกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบเท่ากับ 1,000 คนต่อ 1,000 คน ดังนั้นผลการสุ่มขนาดกลุ่มตัวอย่างจำนวน 9 เมทริกซ์ (ลักษณะของข้อสอบ 9 ลักษณะ) จะได้เมทริกซ์ข้อมูลทั้งหมด 18 แฟ้ม

2. นำเมทริกซ์ข้อมูลที่สุ่มได้ในข้อ 1 มาวิเคราะห์หาค่าพารามิเตอร์ของข้อสอบ (a, b และ c) ภายใต้ทฤษฎี IRT ด้วยโปรแกรม BILOG version 3.04 (Mislevy and Bock, 1990) โดยวิเคราะห์แยกกันระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ โปรแกรมดังกล่าวใช้วิเคราะห์ข้อมูลที่มีการให้คะแนนแบบทวิภาคซึ่งสามารถเลือกโมเดลโลจิสติกแบบ 1, 2, 3 พารามิเตอร์หรือแบบ 3 พารามิเตอร์ชนิดกำหนดค่า c คงที่ สำหรับวิธีการประมาณค่าพารามิเตอร์ของข้อสอบสามารถเลือกได้ 2 วิธี คือ วิธี marginal maximum likelihood estimation (MMLE) และวิธีของ Bayes ที่เรียกว่า marginal maximum a posteriori estimation (MMAPE) ส่วนวิธีการประมาณค่าความสามารถมี 3 วิธี ได้แก่ วิธี maximum likelihood estimation (MLE) วิธี expected a posteriori estimation (EAPE) และวิธี maximum likelihood posteriori estimation (MAPE) โปรแกรม BILOG ประกอบด้วยโปรแกรมย่อย 3 โปรแกรม ดังนี้

(1) INPUT เป็นโปรแกรมกำหนดคำสั่งเพื่ออ่านข้อมูล สร้างแฟ้มข้อมูล และคำนวณค่าสถิติรายข้อตามทฤษฎี CTT

(2) CALIBRATE เป็นโปรแกรมคำนวณค่าประมาณพารามิเตอร์ของข้อสอบ (a, b และ c) ค่าความคลาดเคลื่อนมาตรฐาน และทดสอบความเหมาะสมของข้อมูลกับโมเดล

(3) SCORE เป็นโปรแกรมคำนวณค่าประมาณความสามารถของผู้สอบ (θ) และค่าความคลาดเคลื่อนมาตรฐาน

ในการประมาณค่าพารามิเตอร์ของข้อสอบ ผู้วิจัยใช้โมเดลโลจิสติกแบบ 3 พารามิเตอร์ชนิดกำหนดค่า c คงที่ (3PLM-c) ซึ่งจะสอดคล้องกับการคำนวณดัชนีการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการวัดพื้นที่ของ Raju (1990) โดยใช้โมเดล 3 พารามิเตอร์ชนิดกำหนดค่า c คงที่เหมือนกัน ทั้งนี้เนื่องจากถ้าค่า c ของข้อสอบระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบมีค่าไม่เท่ากันแล้วจะส่งผลให้การคำนวณค่าประมาณพื้นที่ไม่มีที่สิ้นสุด (infinite) (Kim and Cohen, 1991) ในการประมาณค่าพารามิเตอร์ของข้อสอบดังกล่าวใช้วิธี MMLE สำหรับการประมาณค่าความสามารถผู้วิจัยไม่ได้วิเคราะห์เนื่องจากไม่ได้นำผลมาใช้ในการคำนวณ ดังนั้นผลการวิเคราะห์ค่าประมาณพารามิเตอร์ของข้อสอบจะได้แฟ้มข้อมูลนามสกุล .cov จำนวน 18 แฟ้ม ซึ่งเป็นเมทริกซ์ความแปรปรวน-ความแปรปรวนร่วมของค่าประมาณพารามิเตอร์ของข้อสอบ ซึ่งประกอบด้วยค่าอำนาจจำแนก (slopes) ค่าความยาก (thresholds) และค่าการเดา (asymptotes) สำหรับรายละเอียดของผลการประมาณค่าพารามิเตอร์ของข้อสอบในกลุ่มอ้างอิงและกลุ่มเปรียบเทียบทั้ง 9 เมทริกซ์ แสดงดังตารางที่ 83 – 91 ในภาคผนวก ข หน้า 313 – 321

3. นำแฟ้มข้อมูลเมทริกซ์ความแปรปรวน-ความแปรปรวนร่วมของค่าประมาณพารามิเตอร์ของข้อสอบในผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่ได้จากการวิเคราะห์ในข้อ 2 มาปรับเทียบสเกลพารามิเตอร์ของข้อสอบ ทั้งนี้เนื่องจากพารามิเตอร์ของข้อสอบดังกล่าวประมาณค่ามาจากการวิเคราะห์แยกกันในแต่กลุ่มผู้สอบ ดังนั้นก่อนที่จะนำข้อมูลทั้งสองกลุ่มมาวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบจะต้องปรับเทียบสเกลพารามิเตอร์ของข้อสอบให้อยู่บนเมทริกซ์เดียวกันก่อน ซึ่งผู้วิจัยปรับเทียบสเกลพารามิเตอร์ของข้อสอบด้วยวิธีโค้งลักษณะแบบสอบ (test characteristic curve; TCC) ของ Stocking และ Lord (1983 cited in Kim and Cohen, 1992a) ภายใต้โมเดลโลจิสติกแบบ 2 พารามิเตอร์ ซึ่งจะปรับเทียบเฉพาะค่าอำนาจจำแนกและค่าความยากของข้อสอบโดยใช้โปรแกรม EQUATE version 2.0 (Baker, 1993) ในการปรับเทียบใช้การแปลงค่าเชิงเส้น ดังนี้

$$a^{*}_{iF} = a_{iF} / A$$
$$b^{*}_{iF} = Ab_{iF} / A + K$$

เมื่อ a^{*}_{iF} แทน ค่าอำนาจจำแนกของข้อสอบข้อที่ i จากผู้สอบกลุ่มเปรียบเทียบ ซึ่งได้แปลงค่าแล้ว

b^{*}_{iF} แทน ค่าความยากของข้อสอบข้อที่ i จากผู้สอบกลุ่มเปรียบเทียบ ซึ่งได้แปลงค่าแล้ว

a_{iF} แทน ค่าอำนาจจำแนกของข้อสอบข้อที่ i จากผู้สอบกลุ่มเปรียบเทียบ ซึ่งยังไม่ได้แปลงค่า

b_{iF} แทน ค่าความยากของข้อสอบข้อที่ i จากผู้สอบกลุ่มเปรียบเทียบ ซึ่งยังไม่ได้แปลงค่า

A แทน ค่าความชัน (slope)

K แทน ค่าจุดตัด (intercept)

ค่าความชัน (A) และค่าจุดตัด (K) ที่คำนวณโดยวิธีโค้งลักษณะแบบสอบภายใต้โมเดลโลจิสติกแบบ 2 พารามิเตอร์ เป็นค่าที่ได้จากการแปลงค่า T^{*}_{jF} ไปยัง T_{jF} ซึ่งทำให้ฟังก์ชัน F มีค่าน้อยที่สุด ดังนี้

$$T_{jF} = \sum_{i=1}^{n} P(\theta_{jF}, a_{iR}, b_{iR})$$

$$T^{*}_{jF} = \sum_{i=1}^{n} P(\theta_{jF}, a^{*}_{iF}, b^{*}_{iF})$$

$$F = \frac{1}{N}\sum_{j=1}^{N}\left[\sum_{i=1}^{n} P(\theta_{jF}, a_{iR}, b_{iR}) - \sum_{i=1}^{n} P(\theta_{jF}, a^{*}_{iF}, b^{*}_{iF})\right]^{2}$$

เมื่อ θ_{jF} แทน ระดับความสามารถของผู้สอบคนที่ j ในกลุ่มเปรียบเทียบ

n แทน จำนวนข้อสอบ

N แทน จำนวนผู้สอบทั้งหมด

T_{jF} แทน คะแนนจริงของผู้สอบคนที่ j ในกลุ่มเปรียบเทียบ ซึ่งยังไม่ได้แปลงค่า

T^{*}_{jF} แทน คะแนนจริงของผู้สอบคนที่ j ในกลุ่มเปรียบเทียบ ซึ่งได้แปลงค่าแล้ว

a_{iR} แทน ค่าอำนาจจำแนกของข้อสอบข้อที่ i จากผู้สอบกลุ่มอ้างอิง ซึ่งยังไม่ได้แปลงค่า

b_{iR} แทน ค่าความยากของข้อสอบข้อที่ i จากผู้สอบกลุ่มอ้างอิง ซึ่งยังไม่ได้แปลงค่า

ผลจากการปรับเทียบสเกลพารามิเตอร์ของข้อสอบในแต่ละเมทริกซ์ได้ค่า A และ K ทั้งหมด 9 คู่ ซึ่งมีรายละเอียดดังตารางที่ 92 ในภาคผนวก ค หน้า 322 ค่าดังกล่าวจะนำไปใช้ในการวิเคราะห์ดัชนีการทำหน้าที่ต่างกันของข้อสอบต่อไป

4. นำแฟ้มข้อมูลเมทริกซ์ความแปรปรวน-ความแปรปรวนร่วมที่ได้จากการวิเคราะห์ด้วยโปรแกรม BILOG ในข้อ 2 กับค่า A และ K ที่ได้จากการวิเคราะห์ด้วยโปรแกรม EQUATE ในข้อ 3 มาวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป โดยใช้โปรแกรม IRTDIF version 1.0 (Kim and Cohen, 1992b) โปรแกรมดังกล่าวสามารถวิเคราะห์ดัชนี DIF 3 วิธี คือ วิธีการทดสอบไค-สแควร์ของ Lord (1980) วิธีการวัดพื้นที่ในช่วงเปิดของ Raju (1990) และวิธีการวัดพื้นที่ในช่วงปิดของ Kim และ Cohen (1991) โดยสามารถเลือกคำนวณภายใต้โมเดลโลจิสติกแบบ 1, 2, 3 พารามิเตอร์ และแบบ 3 พารามิเตอร์ชนิดกำหนดค่า c คงที่ ในการศึกษาครั้งนี้ผู้วิจัยใช้วิธีการวัดพื้นที่ในช่วงเปิดของ Raju ทั้งนี้เพราะว่าวิธีดังกล่าวมีสูตรคำนวณดัชนี DIF ทั้งในกรณีแบบเอกรูปและแบบอเนกรูป ซึ่งวิธีการทดสอบไค-สแควร์ของ Lord ไม่มี ทั้งยังสามารถทดสอบนัยสำคัญทางสถิติกับดัชนี DIF แต่วิธีการวัดพื้นที่ในช่วงปิดของ Kim และ Cohen ไม่สามารถกระทำได้ ดังนั้นผู้วิจัยจึงเลือกวิธีการวัดพื้นที่ในช่วงเปิดของ Raju คำนวณดัชนี DIF แบบอเนกรูป ($\hat{a}_{iR} \neq \hat{a}_{iF}$) ภายใต้โมเดลโลจิสติกแบบ 3 พารามิเตอร์ชนิดกำหนดค่า c คงที่ และใช้สูตรการคำนวณชนิดไม่คิดเครื่องหมาย (exact unsigned area; EUA) ทั้งนี้จะได้ไม่มีปัญหาในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปประเภทโค้งลักษณะข้อสอบระหว่างกลุ่มผู้สอบตัดกันในลักษณะสมมาตร (Feinstein, 1995) สำหรับสูตรที่ใช้ในการคำนวณมีดังนี้

$$EUA = (1-c)|H| \qquad \text{เมื่อ } \hat{a}_{iR} \neq \hat{a}_{iF}$$

โดยที่ $$H = \frac{2(\hat{a}_F - \hat{a}_R)}{D\hat{a}_R a_F} \ln\left\{1 + \exp\left[\frac{D\hat{a}_R\hat{a}_F(\hat{b}_F - \hat{b}_R)}{\hat{a}_F - \hat{a}_R}\right] - (\hat{b}_F - \hat{b}_R)\right\}$$

สมมติฐานของการทดสอบดัชนี DIF กำหนดดังนี้

$$H_0 : H_i = 0$$

$$H_1 : H_i \neq 0$$

สำหรับการทดสอบนัยสำคัญของดัชนี EUA จะใช้สถิติ Z_i โดยนำไปเปรียบเทียบกับ 0 ดังนี้

$$Z_i(H) = \frac{H_i}{\sqrt{Var(H_i)}} \qquad \text{เมื่อ } \bar{a}_{iR} \neq \bar{a}_{iF}$$

โดยที่

$$Var(H) = B_R^2 Var(\hat{b}_R) + B_F^2 Var(\hat{b}_F) + A_R^2 Var(\hat{a}_R) + A_F^2 Var(\hat{a}_F) + 2B_R A_R Cov(\hat{b}_R, \hat{a}_R) + 2B_F A_F Cov(\hat{b}_F, \hat{a}_F)$$

$$B_R = 1 - \frac{2\exp(Y)}{1+\exp(Y)}$$

$$B_F = -B_R$$

$$A_R = \frac{2}{a_R^2}\left\{\frac{a_R a_F (b_R - b_F)}{a_R - a_F} \cdot \frac{\exp(Y)}{1+\exp(Y)} - \frac{\ln[1+\exp(Y)]}{D}\right\}$$

$$A_F = -\frac{a_R^2}{a_F^2} A_R$$

$$Y = \frac{D a_R a_F (b_F - b_R)}{a_F - a_R}$$

$$Var(\hat{b}_i) = \frac{I_{a_i}}{I_{a_i} I_{b_i} - I_{a_i b_i}^2}$$

$$Var(\hat{a}_i) = \frac{I_{b_i}}{I_{a_i} I_{b_i} - I_{a_i b_i}^2}$$

$$Cov(\hat{a}_i, \hat{b}_i) = \frac{-I_{a_i b_i}}{I_{a_i} I_{b_i} - I_{a_i b_i}^2}$$

$$I_{a_i} = \frac{D^2}{(1-c)^2} \sum_{j=1}^{N} \left\{(\theta_j - b_i)^2 \left[P_i(\theta_j) - c\right]^2 \frac{1 - P_i(\theta_j)}{P_i(\theta_j)}\right\}$$

$$I_{b_i} = \frac{D^2 a_i^2}{(1-c)^2} \sum_{j=1}^{N} \left\{\left[P_i(\theta_j) - c\right]^2 \frac{1 - P_i(\theta_j)}{P_i(\theta_j)}\right\}$$

$$I_{a_i b_i} = \frac{D^2 a_i}{(1-c)^2} \sum_{j=1}^{N} \left\{ (\theta_j - b_i) \left| P_i(\theta_j) - c \right|^2 \frac{1 - P_i(\theta_j)}{P_i(\theta_j)} \right\}$$

โปรแกรม IRTDIF จะคำนวณพื้นที่พร้อมทั้งทดสอบนัยสำคัญทางสถิติที่ระดับ .001, .01 และ .05 ในการศึกษาครั้งนี้ผู้วิจัยใช้เกณฑ์ตัดสินที่ระดับ .05 สำหรับผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปในเมทริกซ์ข้อมูล 9 เมทริกซ์ ได้จำนวนข้อสอบที่ทำหน้าที่ต่างกันอยู่ในช่วง 12 – 29 ข้อ ซึ่งมีรายละเอียดดังตารางที่ 93 – 101 ในภาคผนวก ง หน้า 323 – 331

5. สุ่มข้อสอบที่ทำหน้าที่ต่างกันในเมทริกซ์ข้อมูลซึ่งตรวจสอบแล้วในข้อ 4 โดยสุ่มอย่างง่ายเมทริกซ์ละ 12 ข้อ จะได้ข้อสอบที่ทำหน้าที่ต่างกันทั้งหมด 108 ข้อ แสดงดังภาพที่ 12 – 20 ในภาคผนวก จ หน้า 332 – 349 ต่อจากนั้นจึงสุ่มข้อสอบที่ทำหน้าที่ต่างกันในแต่ละเมทริกซ์ จำนวน 6 ข้อจาก 12 ข้อ แล้วสุ่มอีก 3 ข้อจาก 6 ข้อ และสุ่มอีก 2 ข้อ จาก 3 ข้อ ตามลำดับ และในทำนองเดียวกัน สุ่มข้อสอบที่ทำหน้าที่ไม่ต่างกันในแต่ละเมทริกซ์ข้อมูล โดยสุ่มมาจำนวน 57 ข้อ, 54 ข้อ, 48 ข้อ, 28 ข้อ, 27 ข้อ และ 24 ข้อ ตามลำดับ แล้วนำข้อสอบที่สุ่มได้ในแต่ละเมทริกซ์ข้อมูลดังกล่าวมาจัดกระทำเป็นแบบสอบ 6 ฉบับ ซึ่งมีรูปแบบดังตารางที่ 5

ตารางที่ 5 รูปแบบของการจัดกระทำแบบสอบ 6 ฉบับ ในแต่ละเมทริกซ์ข้อมูล ภายใต้สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันจำนวน 5%, 10% และ 20% ในแบบสอบที่มีความยาว 30 ข้อ และ 60 ข้อ

แบบสอบฉบับที่	จำนวนข้อสอบ	สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน	จำนวนข้อสอบที่ทำหน้าที่ต่างกัน	จำนวนข้อสอบที่ทำหน้าที่ไม่ต่างกัน
1		5%	2	28
2	30	10%	3	27
3		20%	6	24
4		5%	3	57
5	60	10%	6	54
6		20%	12	48

6. สุ่มขนาดกลุ่มตัวอย่างในแต่ละเมทริกซ์ข้อมูลจากข้อ 1 โดยสุ่มอย่างง่ายแบบใส่คืน เพื่อให้ได้ขนาดกลุ่มผู้สอบจำนวน 6 ระดับ คือ จำนวนผู้สอบกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ เท่ากับ 250 คนต่อ 250 คน, 500 คนต่อ 250 คน, 500 คนต่อ 500 คน, 1000 คนต่อ 250 คน, 1000 คนต่อ 500 คน และ 1000 คนต่อ 1000 คน อนึ่งในการสุ่มขนาดกลุ่มตัวอย่างดังกล่าว จำนวนผู้สอบกลุ่มอ้างอิงต้องสุ่มมาจากผู้สอบกลุ่มอ้างอิงขนาด 1000 คน และในทำนองเดียวกัน จำนวนผู้สอบกลุ่มเปรียบเทียบต้องสุ่มมาจากผู้สอบกลุ่มเปรียบเทียบขนาด 1000 คน

ดังนั้นจะได้ผลการจัดกระทำข้อมูลทั้งหมด 324 เงื่อนไข (ลักษณะของข้อสอบ 9 ลักษณะ × ความยาวของแบบสอบ 2 ระดับ × สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน 3 ระดับ × ขนาดกลุ่มตัวอย่าง 6 ระดับ)

ตอนที่ 3 การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธีซิปเทสท์ วิธีซิปเทสท์ปรับใหม่ วิธีแมนเทล-แฮนส์เซล และวิธีการถดถอยโลจิสติก

1. การวิเคราะห์ DIF ด้วยวิธีซิปเทสท์

ในการตรวจสอบ DIF แบบอเนกรูปด้วยวิธีซิปเทสท์ ผู้วิจัยใช้โปรแกรม SIBTEST ของ Stout และ Roussos version 1.1 (1992) ซึ่งมีสูตรในการคำนวณดังนี้

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

เมื่อ $\hat{\beta}_{uni}$ แทน ดัชนีซิปเทสท์

$\bar{Y}_{Rk}$ แทน ค่าเฉลี่ยของคะแนนรวมจากชุดแบบสอบที่ต้องการศึกษาของผู้สอบกลุ่มอ้างอิง ซึ่งได้คะแนน $X = k$

$\bar{Y}_{Fk}$ แทน ค่าเฉลี่ยของคะแนนรวมจากชุดแบบสอบที่ต้องการศึกษาของผู้สอบกลุ่มเปรียบเทียบ ซึ่งได้คะแนน $X = k$

$\hat{P}_k$ แทน สัดส่วนของผู้สอบทั้งหมด (กลุ่มอ้างอิงและกลุ่มเปรียบเทียบ) ซึ่งตอบชุดแบบสอบที่มีความตรงโดยได้คะแนนรวม $X = k$

k แทน คะแนนรวมจากชุดแบบสอบที่มีความตรง

โดยที่ $\hat{P}_k$ คำนวณได้จากสูตรดังนี้

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

เมื่อ J_{Fk} แทน จำนวนผู้สอบกลุ่มเปรียบเทียบซึ่งตอบชุดแบบสอบที่มีความตรงแล้วได้คะแนนรวม $X = k$

J_{Rk} แทน จำนวนผู้สอบกลุ่มอ้างอิงซึ่งตอบชุดแบบสอบที่มีความตรงแล้วได้คะแนนรวม $X = k$

สมมติฐานของการทดสอบดัชนี DIF กำหนดดังนี้

$$H_0 : \beta_{uni} = 0$$
$$H_1 : \beta_{uni} > 0$$

นำค่า $\hat{\beta}_{uni}$ ดังกล่าวมาคำนวณสถิติ B_{uni} เพื่อนำไปทดสอบสมมติฐาน H_0 โดยนำไปเปรียบเทียบกับสถิติ Z ที่ระดับ .05 ดังนี้

$$B_{uni} = \frac{\hat{\beta}_{uni}}{\hat{\sigma}(\hat{\beta}_{uni})}$$

ขณะที่ $\hat{\sigma}(\hat{\beta}_{uni})$ คำนวณได้จากสูตร

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

เมื่อ $\hat{\sigma}(\hat{\beta}_{uni})$ แทน ค่าประมาณความคลาดเคลื่อนมาตรฐานของ β_{uni}

$\hat{\sigma}^2(Y|k,g)$ แทน ค่าประมาณความแปรปรวนของคะแนนจากชุดแบบสอบที่ต้องการศึกษาสำหรับผู้สอบกลุ่ม g (R หรือ F) ซึ่งมีคะแนนรวมเท่ากับ k

J_{gk} แทน จำนวนผู้สอบกลุ่ม g (R หรือ F) ซึ่งตอบชุดแบบสอบที่มีความตรงแล้วได้คะแนนรวม $X = k$

ผู้วิจัยใช้เกณฑ์ตัดสินข้อสอบทำหน้าที่ต่างกันที่ระดับนัยสำคัญ .05 ถ้าผลการทดสอบปรากฏว่า $B_{uni} > Z_{\alpha}$ อย่างมีนัยสำคัญที่ระดับ .05 แสดงว่าปฏิเสธ H_0 นั่นคือ ข้อสอบที่นำมาตรวจสอบทำหน้าที่ต่างกัน

2. การวิเคราะห์ DIF ด้วยวิธีซิปเทสท์ปรับใหม่

ผู้วิจัยนำวิธีซิปเทสท์ในข้อ 1 มาปรับปรุงขั้นตอนการวิเคราะห์ข้อมูล เพื่อใช้ตรวจสอบ DIF แบบอเนกรูป ในการวิเคราะห์จะนำจำนวนผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบมาแบ่งกลุ่มผู้สอบออกเป็น 2 กลุ่มตามระดับความสามารถ คือ กลุ่มผู้สอบที่มีความสามารถสูงและกลุ่มผู้สอบที่มีความสามารถต่ำ โดยใช้ค่าเฉลี่ยของคะแนนรวมเป็นเกณฑ์ในการแบ่งกลุ่มผู้สอบ แล้ววิเคราะห์ดัชนี DIF ด้วยวิธีซิปเทสท์ในกลุ่มผู้สอบทั้งสอง โดยวิเคราะห์แยกกันคนละกลุ่ม ในการศึกษาครั้งนี้ผู้วิจัยใช้โปรแกรม SIBTEST ของ Stout และ Roussos version 1.1 (1992) ซึ่งมีสูตรในการคำนวณทำนองเดียวกับข้อ 1 ดังนี้

$$\hat{\beta}_{uni} = \sum_{k=0}^{n} \hat{P}_k (\overline{Y}_{Rk} - \overline{Y}_{Fk})$$

สำหรับการทดสอบนัยสำคัญทางสถิติ ผู้วิจัยใช้เกณฑ์ตัดสินข้อสอบทำหน้าที่ต่างกันที่ระดับนัยสำคัญ .05

3. การวิเคราะห์ DIF ด้วยวิธีแมนเทล-แฮนส์เซล

ในการวิเคราะห์ดัชนี DIF แบบอเนกรูปด้วยวิธีแมนเทล-แฮนส์เซล ผู้วิจัยใช้โปรแกรม MHDIF version 1.0 (Fidalgo, 1995) ซึ่งพัฒนาตามข้อเสนอแนะของ Mazor และคณะ (1994) โปรแกรมดังกล่าวใช้กับข้อมูลที่มีการให้คะแนนแบบทวิภาค สามารถวิเคราะห์ข้อสอบได้ไม่เกิน 75 ข้อ และผู้สอบไม่เกิน 5,000 คน ทั้งยังสามารถเลือกวิเคราะห์ดัชนี DIF แบบเอกรูปและแบบอเนกรูป สำหรับการทดสอบนัยสำคัญทางสถิติสามารถเลือกได้ 7 ระดับ คือ 0.1, 0.05, 0.025, 0.02, 0.01, 0.005 และ 0.001 สูตรที่ใช้ในการคำนวณดัชนี DIF ด้วยวิธีแมนเทล-แฮนส์เซลมีดังนี้

$$\hat{\alpha}_{MH} = \frac{\sum_{j=1}^{K} A_j D_j / T_j}{\sum_{j=1}^{K} B_j C_j / T_j}$$

เมื่อ $\hat{\alpha}_{MH}$ แทน ดัชนีแมนเทล-แฮนส์เซล

A_j แทน จำนวนผู้สอบกลุ่มอ้างอิงที่ระดับคะแนน j ซึ่งตอบข้อสอบถูก

B_j แทน จำนวนผู้สอบกลุ่มอ้างอิงที่ระดับคะแนน j ซึ่งตอบข้อสอบผิด

C_j แทน จำนวนผู้สอบกลุ่มเปรียบเทียบที่ระดับคะแนน j ซึ่งตอบข้อสอบถูก

D_j แทน จำนวนผู้สอบกลุ่มเปรียบเทียบที่ระดับคะแนน j ซึ่งตอบข้อสอบผิด

m_{1j} แทน จำนวนผู้สอบทั้งหมดที่ระดับคะแนน j ซึ่งตอบข้อสอบถูก

m_{0j} แทน จำนวนผู้สอบทั้งหมดที่ระดับคะแนน j ซึ่งตอบข้อสอบผิด

nR_j แทน จำนวนผู้สอบกลุ่มอ้างอิงที่ระดับคะแนน j

nF_j แทน จำนวนผู้สอบกลุ่มเปรียบเทียบที่ระดับคะแนน j

T_j แทน จำนวนผู้สอบทั้งหมดที่ระดับคะแนน j

สมมติฐานของการทดสอบดัชนี DIF กำหนดดังนี้

$$H_0 : \left[\pi_{R_j} \Big/ (1-\pi_{R_j})\right] = \left[\pi_{F_j} \Big/ (1-\pi_{F_j})\right] \quad ; \; j = 1, 2, 3, \ldots, K$$

$$H_1 : \left[\pi_{R_j} \Big/ (1-\pi_{R_j})\right] = \alpha\left[\pi_{F_j} \Big/ (1-\pi_{F_j})\right] \quad ; \; j = 1, 2, 3, \ldots, K, \; \alpha \neq 1$$

เมื่อ π_{R_j} แทน โอกาสของผู้สอบกลุ่มอ้างอิงที่ระดับคะแนน j จะตอบข้อสอบได้ถูกต้อง

π_{F_j} แทน โอกาสของผู้สอบกลุ่มเปรียบเทียบที่ระดับคะแนน j จะตอบข้อสอบได้ถูกต้อง

นำดัชนี $\hat{\alpha}_{MH}$ ไปทดสอบนัยสำคัญกับสถิติแมนเทล-แฮนส์เซลไค-สแควร์ (χ^2_{MH}) ที่ระดับชั้นของความเป็นอิสระเท่ากับ 1 (df = 1) โดยนำไปเปรียบเทียบกับ 1 สำหรับสถิติ χ^2_{MH} มีสูตรคำนวณดังนี้

$$\chi^2_{MH} = \frac{\left[\left|\sum_{j=1}^{K} A_j - E(A_j)\right| - 0.5\right]^2}{\sum_{j=1}^{K} Var(A_j)}$$

โดยที่ $E(A_j) = \dfrac{nR_j m_{1j}}{T_j}$

$$Var(A_j) = \frac{nR_j nF_j m_{1j} m_{0j}}{T_j^2 (T_j - 1)}$$

เมื่อ $E(A_j)$ แทน ค่าคาดหวังของจำนวนผู้สอบกลุ่มอ้างอิงที่ระดับคะแนน j ซึ่งตอบข้อสอบถูก

$Var(A_j)$ แทน ค่าความแปรปรวนของจำนวนผู้สอบกลุ่มอ้างอิงที่ระดับคะแนน j ซึ่งตอบข้อสอบถูก

ในการศึกษาครั้งนี้ ผู้วิจัยเลือกการวิเคราะห์ DIF แบบอเนกรูป โดยใช้คะแนนรวมของผู้สอบซึ่งห่างกันช่วงละ 1 คะแนนเป็นเกณฑ์ในการจับคู่ของกลุ่มผู้สอบ โปรแกรม MHDIF จะคำนวณดัชนี DIF โดยแบ่งกลุ่มผู้สอบออกเป็น 2 กลุ่ม คือ กลุ่มผู้สอบที่มีความสามารถสูงและกลุ่มผู้สอบที่มีความสามารถต่ำ ซึ่งใช้ค่าเฉลี่ยของคะแนนรวมเป็นเกณฑ์ แล้วคำนวณดัชนี DIF ในแต่ละกลุ่มโดยใช้วิธีการวิเคราะห์แบบสองขั้นตอน ดังนี้

ขั้นตอนที่ 1 โปรแกรมจะตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ถ้าไม่พบข้อสอบที่ทำหน้าที่ต่างกันโปรแกรมจะหยุดการตรวจสอบ แต่ถ้าพบข้อสอบที่ทำหน้าที่ต่างกันโปรแกรมจะคำนวณคะแนนรวมของผู้สอบอีกครั้งหนึ่ง โดยไม่รวมคะแนนจากข้อสอบซึ่งตรวจพบว่าทำหน้าที่ต่างกัน

ขั้นตอนที่ 2 โปรแกรมจะตรวจสอบการทำหน้าที่ต่างกันของข้อสอบอีกครั้งหนึ่ง โดยใช้คะแนนรวมครั้งหลังจากขั้นตอนที่ 1 เป็นเกณฑ์ในการจับคู่ของกลุ่มผู้สอบ ซึ่งจะทำให้เกณฑ์ในการจับคู่ของกลุ่มผู้สอบมีความบริสุทธิ์ (purification of matching criterion)

สำหรับค่า $\hat{\alpha}_{MH}$ ที่ใช้ในการทดสอบกับสถิติแมนเทล-แฮนส์เซลไค-สแควร์ (χ^2_{MH}) จะใช้เกณฑ์ตัดสินข้อสอบทำหน้าที่ต่างกันที่ระดับนัยสำคัญ .05

4. *การวิเคราะห์ DIF ด้วยวิธีการถดถอยโลจิสติก*

ในการวิเคราะห์ดัชนี DIF แบบอเนกรูปด้วยวิธีการถดถอยโลจิสติก ผู้วิจัยใช้โปรแกรม SPSS/PC+ version 4.01 ซึ่งมีโมเดลในการคำนวณดังนี้

$$\text{logit}[P(U_{pj} = 1 \mid X)] = \tau_0 + \tau_1 X + \tau_2 G + \tau_3 (XG)$$

เมื่อ $\text{logit}[P(U_{pj} = 1 \mid X)]$ แทน ค่า log ของอัตราส่วนของโอกาสในการตอบข้อสอบถูกของผู้สอบคนที่ p ต่อโอกาสของการตอบข้อสอบผิดของผู้สอบในกลุ่ม j

X_{pj} แทน ระดับความสามารถของผู้สอบคนที่ p ในกลุ่ม j

τ_0 แทน พารามิเตอร์จุดตัด

τ_1 แทน สัมประสิทธิ์ของระดับความสามารถ

τ_2 แทน ความแตกต่างของกลุ่มผู้สอบ

τ_3 แทน ปฏิสัมพันธ์ระหว่างสมาชิกในกลุ่มและระดับความสามารถ

G_j แทน สมาชิกผู้สอบในกลุ่ม j

ในการศึกษาครั้งนี้ ผู้วิจัยใช้คะแนนรวมของแบบสอบ (total test score) แทนระดับความสามารถของผู้สอบ ส่วนสมาชิกของกลุ่มผู้สอบจะกำหนดให้ G_p = 1 เมื่อผู้สอบอยู่ในกลุ่มอ้างอิง และ G_p = 0 เมื่อผู้สอบอยู่ในกลุ่มเปรียบเทียบ สำหรับการตัดสินข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปจะพิจารณาพารามิเตอร์ τ_2 และ τ_3 จากโมเดลข้างบน กล่าวคือ ถ้า $\tau_2 = \tau_3 = 0$ แสดงว่าข้อสอบทำหน้าที่ไม่ต่างกัน ถ้า $\tau_2 \neq 0$ และ $\tau_3 = 0$ แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบมีทิศทางเดียว และถ้า $\tau_3 \neq 0$ ($\tau_2 = 0$ หรือไม่ก็ได้) แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบไม่มีทิศทาง (Li and Stout, 1996) สำหรับสมมติฐานของการทดสอบดัชนี DIF กำหนดดังนี้

$$H_0 : \tau_2 = \tau_3 = 0$$

$$H_1 : \tau_2 \neq 0 \text{ และ } \tau_3 \neq 0$$

ในการทดสอบสมมติฐานจะใช้สถิติดังนี้

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

โดยที่ $\hat{\tau}' = \left[\tau_0, \tau_1, \tau_2, \tau_3 \right]$

และ C แทน เมทริกซ์ขนาด 2 × 4 ดังนี้

$$C = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

สถิติไค-สแควร์ดังกล่าวมีการแจกแจงไค-สแควร์แบบเชิงเส้นกำกับ (asymptotic χ^2 distribution) ที่ระดับชั้นของความเป็นอิสระเท่ากับ 2 (df = 2) ในการศึกษาครั้งนี้ผู้วิจัยใช้เกณฑ์ตัดสินข้อสอบทำหน้าที่ต่างกันที่ระดับนัยสำคัญ .05

ตอนที่ 4 การวิเคราะห์อำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีซิปเทสท์ปรับใหม่ วิธีซิปเทสท์ วิธีแมนเทล-แฮนส์เซล และวิธีการถดถอยโลจิสติก

ในการวิเคราะห์อำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป ผู้วิจัยนำผลการวิเคราะห์ดัชนี DIF ด้วยวิธีซิปเทสท์ปรับใหม่ วิธีซิปเทสท์ วิธีแมนเทล-แฮนส์เซล และวิธีการถดถอยโลจิสติกมาเปรียบเทียบกับวิธีการวัดพื้นที่ของ Raju โดยถือว่าผลจากการวิเคราะห์ดัชนี DIF ด้วยวิธีการวัดพื้นที่ของ Raju เป็นเกณฑ์ที่ถูกต้อง กล่าวคือ ถ้าวิธีซิปเทสท์ปรับใหม่ วิธีซิปเทสท์ วิธีแมนเทล-แฮนส์เซล และวิธีการถดถอยโลจิสติกระบุข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปได้ตรงกับข้อสอบที่ถูกระบุด้วยวิธีการวัดพื้นที่ของ Raju แสดงว่าวิธีดังกล่าวสามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปได้ถูกต้อง และถ้าวิธีซิปเทสท์ปรับใหม่ วิธีซิปเทสท์ วิธีแมนเทล-แฮนส์เซล และวิธีการถดถอยโลจิสติกระบุข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปไม่ตรงกับข้อสอบที่ถูกระบุด้วยวิธีการวัดพื้นที่ของ Raju แสดงว่าวิธีดังกล่าวตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปไม่ถูกต้อง รายละเอียดของผลการวิเคราะห์แสดงดังตารางที่ 102 ในภาคผนวก ฉ หน้า 350 – 359 สำหรับสูตรที่ใช้ในการคำนวณมีดังนี้

$$P = \frac{n_1}{N_1} \times 100$$

$$E_1 = \frac{n_2}{N_2} \times 100$$

เมื่อ P แทน อำนาจการทดสอบ

E_1 แทน อัตราความคลาดเคลื่อนประเภทที่ 1

n_1 แทน จำนวนข้อสอบที่ระบุถูกต้องว่าทำหน้าที่ต่างกัน

n_2 แทน จำนวนข้อสอบที่ระบุผิดพลาดว่าทำหน้าที่ต่างกัน

N_1 แทน จำนวนข้อสอบที่ทำหน้าที่ต่างกันทั้งหมดในแบบสอบที่ตรวจสอบด้วยวิธีเกณฑ์

N_2 แทน จำนวนข้อสอบที่ทำหน้าที่ไม่ต่างกันทั้งหมดในแบบสอบที่ตรวจสอบด้วยวิธีเกณฑ์

ตอนที่ 5 การเปรียบเทียบอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ระหว่างวิธีซิปเทสท์ปรับใหม่ วิธีซิปเทสท์ วิธีแมนเทล-แฮนส์เซล และวิธีการถดถอยโลจิสติก

การเปรียบเทียบอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูประหว่างวิธีซิปเทสท์ปรับใหม่ วิธีซิปเทสท์ วิธีแมนเทล-แฮนส์เซล และวิธีการถดถอยโลจิสติก ภายใต้เงื่อนไขของปัจจัยที่ศึกษา 4 ตัว คือ ลักษณะของข้อสอบ 9 ลักษณะ ความยาวของแบบสอบ 2 ระดับ สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน 3 ระดับ และขนาดกลุ่มตัวอย่าง 6 ระดับ ผู้วิจัยใช้สถิติการวิเคราะห์ความแปรปรวนแบบ 2 ทาง (two-way ANOVA) เพื่อทดสอบผลของวิธีการตรวจสอบ ผลของปัจจัยที่ศึกษา และผลของปฏิสัมพันธ์ระหว่างวิธีการตรวจสอบและปัจจัยที่ศึกษาถ้าผลการทดสอบดังกล่าวมีนัยสำคัญทางสถิติจะใช้สถิติการวิเคราะห์ความแปรปรวนแบบทางเดียว (one-way ANOVA) ทดสอบในแต่ละระดับของตัวแปร ถ้าผลการทดสอบดังกล่าวมีนัยสำคัญทางสถิติจะเปรียบเทียบความแตกต่างเป็นรายคู่โดยใช้วิธีการทดสอบของ Tukey ยกเว้นการเปรียบเทียบภายใต้ความยาวของแบบสอบ 2 ระดับจะใช้การทดสอบ t-test กรณีกลุ่มตัวอย่างเป็นอิสระต่อกัน (independent sample) โดยจะทดสอบใน 3 กรณีต่อไปนี้ **กรณีที่** 1 ทดสอบอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ของแต่ละวิธีการตรวจสอบภายใต้ปัจจัยที่ศึกษาต่างระดับกัน **กรณีที่** 2 ทดสอบอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ระหว่างวิธีการตรวจสอบ 4 วิธีภายใต้ปัจจัยที่ศึกษาระดับเดียวกัน และ**กรณีที่** 3 ทดสอบอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ระหว่างวิธีการตรวจสอบ 4 วิธีภายใต้ปัจจัยที่ศึกษาต่างระดับกัน