บทที่ 3

วิธีดำเนินการวิจัยและการพัฒนาเกณฑ์ตัดสินความลำเอียงของข้อสอบ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อพัฒนาเกณฑ์ตัดสินข้อสอบลำเอียงทางเพศของดัชนีที่ได้จากวิธีการตรวจสอบข้อสอบลำเอียง 4 ตัว ได้แก่ ดัชนี SA และ UA จากวิธี IRT-2 พารามิเตอร์ ดัชนี α_{MH} จากวิธี MH และดัชนี β_{SIB} จากวิธี SIBTEST และในบทนี้ผู้วิจัยได้นำเสนอเนื้อหาเป็น 4 ตอนคือ

ตอนที่ 1 เครื่องมือและข้อมูลที่ใช้ในการวิจัย

ตอนที่ 2 การเตรียมข้อมูลเพื่อพัฒนาเกณฑ์และวิเคราะห์ความลำเอียงของข้อสอบ

ตอนที่ 3 การพัฒนาเกณฑ์ตัดสินข้อสอบลำเอียง

ตอนที่ 4 การทดลองใช้เกณฑ์ตัดสินที่ได้พัฒนาขึ้นจากตอนที่ 3

ตอนที่ 1 เครื่องมือและข้อมูลที่ใช้ในการวิจัย

1. เครื่องมือที่ใช้ในการวิจัย

เครื่องมือที่ใช้ในการวิจัยครั้งนี้ ได้แก่ แบบสอบคัดเลือกบุคคลเพื่อศึกษาต่อในสถาบันอุดมศึกษาของทบวงมหาวิทยาลัยประจำปีการศึกษา 2535 วิชาภาษาอังกฤษ กข และวิชาคณิตศาสตร์ กข

วิชาภาษาอังกฤษ กข ประกอบด้วยข้อสอบแบบเลือกตอบ 4 ตัวเลือก จำนวน 100 ข้อ ส่วนวิชาคณิตศาสตร์ กข ประกอบด้วยข้อสอบเลือกตอบ 4 ตัวเลือก จำนวน 41 ข้อ และข้อสอบเติมคำ 6 ข้อ การวิจัยครั้งนี้นำมาใช้เฉพาะข้อสอบเลือกตอบ

สำหรับแบบสอบคัดเลือกทั้ง 2 ฉบับนี้ ทบวงมหาวิทยาลัยได้วิเคราะห์คุณภาพของแบบทั้งฉบับและรายข้อไว้ดังนี้

แบบสอบวิชาภาษาอังกฤษมีค่าความเที่ยง (KR-20) เท่ากับ .85 ค่าความยาก (p) ของข้อสอบระหว่าง.13 และ .88 ค่าอำนาจจำแนก (r) ของข้อสอบระหว่าง -.07 และ .54 ส่วนแบบสอบวิชาคณิตศาสตร์มีค่าความเที่ยง .63 มีค่าความยากของข้อสอบระหว่าง .15-.52 และค่าอำนาจจำแนกรายข้อระหว่าง .08 และ -.40

2. การเก็บรวบรวมข้อมูล

ผู้วิจัยได้เก็บรวบรวมข้อมูลการตอบข้อสอบคัดเลือกวิชาภาษาอังกฤษและคณิตศาสตร์จากทบวงมหาวิทยาลัย ได้ข้อมูลซึ่งเก็บอยู่ในเทปคอมพิวเตอร์ ซึ่งสามารถแบ่งออกเป็น 2 ส่วน

ส่วนที่ 1 เป็นข้อมูลการตอบข้อสอบของผู้เข้าสอบทั้งหมด ซึ่งมีผู้เข้าสอบวิชาภาษาอังกฤษ 441,851 คน และวิชาคณิตศาสตร์ 43,094 คน ข้อมูลที่เก็บไว้ในเทปคอมพิวเตอร์ส่วนนี้ไม่มีข้อมูลเกี่ยวกับเพศของผู้สอบ ซึ่งเป็นตัวแปรที่ต้องการศึกษาในการวิจัยครั้งนี้

ส่วนที่ 2 เป็นข้อมูลการตอบแบบสอบถามที่ทางทบวงมหาวิทยาลัย ได้สุ่มแจกให้ผู้เข้าสอบคัดเลือกตอบ ข้อมูลที่เก็บไว้ในเทปส่วนนี้มีข้อมูลเกี่ยวกับเพศของผู้สอบอยู่ด้วย

ดังนั้น ในการนำข้อมูลทั้งในส่วนที่เป็นเพศ และส่วนการตอบข้อสอบมาใช้ ผู้วิจัยจึงต้องใช้คำสั่ง merge ในโปรแกรมสำเร็จรูป SAS เชื่อมข้อมูล 2 ส่วนเข้าด้วยกัน โดยใช้เลขที่นั่งสอบเป็นตัวเชื่อม ได้กลุ่มผู้สอบวิชาภาษาอังกฤษ จำนวน 16,026 คน เป็นชาย 7,144 คน (ร้อยละ 44.58) หญิง 8,878 คน (ร้อยละ 55.40) ไม่ระบุเพศ 4 คน (ร้อยละ 0.02) ส่วนกลุ่มผู้สอบวิชาคณิตศาสตร์มีจำนวน 25,459 คน เป็นชาย 11,977 คน (ร้อยละ 47.04) หญิง 13,480 คน (ร้อยละ 52.95) และไม่ระบุเพศ 2 คน (ร้อยละ 0.01)

3. สถิติที่ใช้ในการวิจัย

การวิจัยครั้งนี้ผู้วิจัยใช้โปรแกรมสำเร็จรูปเพื่อคำนวณค่าสถิติต่าง ๆ ดังนี้

1. ใช้โปรแกรมสำเร็จรูป SAS แปลงคะแนนการตอบข้อสอบเป็น 0 (ตอบผิด) และ 1 (ตอบถูก) และใช้ในการสุ่มกลุ่มตัวอย่างตามขนาดที่กำหนด

2. ใช้โปรแกรมสำเร็จรูป SPSS วิเคราะห์ตัวประกอบ (Factor Analysis) ของข้อมูลการตอบข้อสอบวิชาภาษาอังกฤษและคณิตศาสตร์ เพื่อตรวจสอบความเป็นเอกมิติของข้อสอบ

3. ใช้โปรแกรม BILOG ประมาณค่าพารามิเตอร์ของข้อสอบวิชาภาษาอังกฤษและวิชาคณิตศาสตร์

4. ใช้โปรแกรม EQUATE เทียบมาตรค่าพารามิเตอร์ที่ประมาณได้ของกลุ่มผู้สอบ 2 กลุ่ม ที่ต้องการหาพื้นที่ของความแตกต่างระหว่างโค้งข้อสอบ

5. ใช้โปรแกรม AREAB และ AREASEB คำนวณค่าพื้นที่ของความแตกต่างระหว่างโค้งข้อสอบของกลุ่มตัวอย่าง 2 กลุ่ม ได้แก่ พื้นที่ที่มีเครื่องหมาย (SA) และพื้นที่ที่ไม่มีเครื่องหมาย (UA) และค่าไคสแควร์ของลอร์ดซึ่งแสดงผลการทดสอบสมมุติฐานของความแตกต่างของค่าพารามิเตอร์ของข้อสอบระหว่างผู้สอบ 2 กลุ่ม

6. ใช้โปรแกรม SIBTEST คำนวณค่าเบต้าและค่าสถิติ Z ซึ่งทดสอบสมมุติฐานของความแตกต่างระหว่างผลการตอบข้อสอบของกลุ่มตัวอย่าง 2 กลุ่ม พร้อมทั้งคำนวณค่าเดลต้าและแอลฟา (Δ และ α) ของแมนเทล-แฮนส์เซล และค่าไคสแควร์ของ MH ซึ่งใช้ทดสอบสมมุติฐานของความแตกต่างในการตอบข้อสอบของกลุ่มสอบ 2 กลุ่ม

7. ใช้โปรแกรมสำเร็จรูป SPSS-PC วิเคราะห์ค่าเฉลี่ยและค่าความเบี่ยงเบนมาตรฐานของค่าสถิติ SA UA α_{MH} และ β_{SIB}

ตอนที่ 2 การเตรียมข้อมูลเพื่อพัฒนาเกณฑ์และวิเคราะห์ความลำเอียงของข้อสอบ

ในการวิจัยครั้งนี้ผู้วิจัยได้เตรียมข้อมูลเพื่อการพัฒนาเกณฑ์ และการวิเคราะห์ความลำเอียงของข้อสอบดังนี้

1. การสุ่มตัวอย่าง

ใช้โปรแกรมสำเร็จรูป SAS สุ่มตัวอย่างผู้สอบวิชาภาษาอังกฤษและวิชาคณิตศาสตร์แยกตามเพศโดยวิธีสุ่มแบบไม่ใส่คืน (random sampling without replacement) ได้กลุ่มตัวอย่างขนาด 100 200 400 600 800 และ 1,000 คน ขนาดละ 101 กลุ่ม ในแต่ละเพศและแต่ละวิชา เตรียมไว้ขนาดละ 100 กลุ่ม เพื่อการพัฒนาเกณฑ์และอีก 1 กลุ่มที่เหลือใช้ในการวิเคราะห์ข้อสอบลำเอียง

2. ในแต่ละขนาดของกลุ่มตัวอย่างเพศเดียวกันผู้วิจัยได้สุ่มจับคู่กลุ่มตัวอย่างทั้ง 101 กลุ่มให้เป็น 50 คู่ เพื่อใช้ในการพัฒนาเกณฑ์การตัดสินข้อสอบลำเอียง ที่เหลืออีกเพศละ 1 กลุ่ม (แต่ละเพศมี 101 กลุ่ม) จับคู่กลุ่มตัวอย่างต่างเพศชาย/หญิง ขนาดละ 1 คู่ เพื่อใช้ในการวิเคราะห์ความลำเอียง ผลการจับคู่ในแต่ละขนาด (รวม 6 ขนาด คือ 100, 200, 400, 600, 800 และ 1,000 คน) ดังตารางที่ 4

ตารางที่ 4 ข้อมูลเกี่ยวกับกลุ่มตัวอย่างที่ใช้ในการศึกษาแต่ละขนาด

วิชา	จำนวนข้อสอบ	ตัวอย่าง		จำนวนกลุ่มที่ใช้ในการพัฒนาเกณฑ์	จำนวนกลุ่มที่ใช้ในการตรวจสอบความลำเอียง
		เพศ	จำนวนกลุ่ม		
ภาษาอังกฤษ	40 ข้อ	ชาย	101	100 (50 คู่)	1 (กลุ่มเปรียบเทียบ = F)
		หญิง	101	100 (50 คู่)	1 (กลุ่มอ้างอิง = R)
คณิตศาสตร์	80 ข้อ	ชาย	101	100 (50 คู่)	1 (กลุ่มเปรียบเทียบ = R)
		หญิง	101	100 (50 คู่)	1 (กลุ่มอ้างอิง = F)

3. การเตรียมข้อมูลสำหรับการวิเคราะห์เมื่อใช้ความยาวแบบสอบต่างกัน

3.1 จัดเรียงข้อสอบเป็นกลุ่มตามเนื้อหาที่วัด โดยจัดให้ข้อสอบที่อยู่ในกลุ่มเดียวกันวัดเนื้อหาเดียวกัน หรือใกล้เคียงกันมากที่สุด ได้แก่

ก. วิชาคณิตศาสตร์ ข้อสอบวิชาคณิตศาสตร์สามารถแยกตามเนื้อหาที่วัดได้เป็น 7 กลุ่ม คือ

1) เซทและฟังชั่น ได้แก่ข้อสอบข้อที่ 1-6,15,16 และ 18-22 รวม 13 ข้อ
2) ระบบจำนวน ได้แก่ข้อสอบข้อที่ 7-10 รวม 4 ข้อ
3) สมการ ได้แก่ข้อสอบข้อที่ 11-14 17 28-30 และ 41 รวม 9 ข้อ
4) อนุกรม ได้แก่ข้อสอบข้อที่ 23 และ 24 รวม 2 ข้อ
5) แมตริกซ์ ได้แก่ข้อสอบข้อที่ 25-27 รวม 3 ข้อ
6) ความน่าจะเป็นและการสุ่ม ได้แก่ข้อสอบข้อที่ 31-35 รวม 5 ข้อ
7) การแจกแจงความถี่ ได้แก่ข้อสอบข้อที่ 36-40 รวม 5 ข้อ

ข. วิชาภาษาอังกฤษ ข้อสอบวิชาภาษาอังกฤษสามารถแยกตามเนื้อหาที่วัดเป็น 3 กลุ่มใหญ่ ๆ ได้แก่

1) การเติมคำหรือข้อความในบทสนทนา 30 ข้อ (ข้อ 1 - 30)

2) คำศัพท์และการใช้คำศัพท์ 15 ข้อ (ข้อ 31 - 45)

3) ความเข้าใจภาษา 55 ข้อ (ข้อ 46 - 100)

และเนื่องจากแต่ละกลุ่มมีเนื้อหาและสถานการณ์แตกต่างกันไป ผู้วิจัยจึงได้แยกกลุ่มข้อสอบออกเป็นกลุ่มย่อย ๆ ตามสถานการณ์ดังนี้

1) การเติมคำ/ข้อความในบทสนทนา

1.1 บทสนทนาทั่วไประหว่างเพื่อน ข้อ 1-4 รวม 4 ข้อ

1.2 บทสนทนาระหว่างแพทย์และคนไข้ข้อ 5-9 รวม 4 ข้อ

1.3 บทสนทนาระหว่างลูกค้าและเจ้าหน้าที่รับจองตั๋วรถไฟ ข้อ 10-14 รวม 5 ข้อ

1.4 บทสนทนาเกี่ยวกับการนัดหมายข้อ 15-20 รวม 6 ข้อ

1.5 บทสนทนาระหว่างลูกค้าและเจ้าหน้าที่วีดีทัศน์ ข้อ 21-30 รวม 10 ข้อ

2) คำศัพท์และการใช้ศัพท์

2.1 การขอบคุณสำหรับการต้อนรับ ข้อ 36-40 รวม 5 ข้อ

2.2 การวางแผนท่องเที่ยว ข้อ 41-45 รวม 5 ข้อ

2.3 การอนุรักษ์ต้นไม้ ข้อ 46-50 รวม 5 ข้อ

3) ความเข้าใจภาษาหรือการอ่าน

3.1 การใช้แรงงานในอุตสาหกรรมตัดไม้ ข้อ31 รวม 1 ข้อ

3.2 การท่องเที่ยว ข้อ 32-35 รวม 4 ข้อ

3.3 บรรยากาศชั้น BIOSPHERE ข้อ 51-62 รวม 12 ข้อ

3.4 การนำของที่ใช้แล้วกลับมาใช้อีก ข้อ63-68 รวม 6 ข้อ

3.5 การเลือกคู่ของนกแม่หม้าย ข้อ 69-74 รวม 6 ข้อ

3.6 คอมพิวเตอร์ ข้อ 75-81 รวม 7 ข้อ

3.7 ชีวิตและความโกรธ ข้อ 82-89 รวม 8 ข้อ

3.8 ความพยายามและความสำเร็จ ข้อ 90-100 รวม 11 ข้อ

3.2 จัดเรียงลำดับข้อสอบตามค่าอำนาจจำแนก และค่าความยากของข้อสอบ

ผู้วิจัยได้เรียงลำดับข้อสอบเพื่อความสะดวกในการจัดข้อสอบตามความยาวแบบสอบที่ใช้ในการศึกษาครั้งนี้

1. เลือกเฉพาะข้อสอบที่มีค่าอำนาจจำแนกมากกว่า .20 และค่าความยากระหว่าง .20 ถึง .80
2. เรียงลำดับข้อสอบที่เลือกไว้ใน 1 ตามค่าอำนาจจำแนกจากสูงสุดไปต่ำสุดกรณีที่มีข้อสอบตั้งแต่ 2 ข้อ ที่มีค่าอำนาจจำแนกเท่ากัน จะพิจารณาเลือกข้อที่มีค่าความยากอยู่ใกล้ .50 มากกว่าก่อน
3. ข้อสอบที่เหลือ นำมาเรียงลำดับต่อจากการจัดลำดับในข้อ 2 โดยเรียงตามค่าอำนาจจำแนกสูงสุดลงไปตามลำดับจนถึงต่ำสุด

เหตุผลในการคัดเลือกข้อสอบ โดยพิจารณาค่าอำนาจจำแนกเป็นเกณฑ์หลักเมื่อมีคุณภาพด้านอื่นอยู่ในเกณฑ์แล้วนั้น เพราะผู้วิจัยพิจารณาว่าแบบสอบที่นำมาใช้เป็นแบบสอบที่ใช้เพื่อการคัดเลือกจึงควรมีค่าอำนาจจำแนกสูงกว่าเมื่อมีคุณภาพด้านอื่นเท่า ๆ กัน

3.3 การคัดเลือกข้อสอบตามขนาดแบบสอบที่สนใจศึกษา

หลังจากเรียงลำดับข้อสอบแล้ว ผู้วิจัยได้คัดข้อสอบ 20 ข้อท้ายในวิชาภาษาอังกฤษและข้อสอบข้อสุดท้ายในวิชาคณิตศาสตร์ออกเหลือข้อสอบวิชาภาษาอังกฤษ 80 ข้อ และวิชาคณิตศาสตร์ 40 ข้อ โดยที่ข้อสอบวิชาภาษาอังกฤษมีค่าความยากระหว่าง .13-.77 ค่าเฉลี่ยเท่ากับ 0.43 ค่าอำนาจจำแนกระหว่าง .11-.54 ค่าเฉลี่ยเท่ากับ 0.31

และข้อสอบวิชาคณิตศาสตร์ความยาว 40 ข้อ มีค่าความยากระหว่าง .15-.50 ค่าเฉลี่ยเท่ากับ 0.31 ค่าอำนาจจำแนกระหว่าง .90-.40 ค่าเฉลี่ยเท่ากับ 0.26

หลังจากนั้นใช้การสุ่มแบบง่าย (simple random sampling) สุ่มข้อสอบตามกลุ่มเนื้อหาที่จัดไว้โดยพิจารณาจากสัดส่วนจำนวนข้อสอบในแต่ละกลุ่มเนื้อหา สุ่มข้อสอบเข้ารวมเป็นแบบสอบยาว 20 ข้อ 30 ข้อ และ 40 ข้อ สำหรับวิชาคณิตศาสตร์ 50 60 70 และ 80 ข้อ สำหรับวิชาภาษาอังกฤษ ทั้งนี้มีค่าเฉลี่ยความยากในแต่ละความยาวแบบสอบเป็น .31 และค่าอำนาจจำแนกเป็น .26 เท่ากันทุกความยาวแบบสอบสำหรับวิชาคณิตศาสตร์ ส่วนวิชาภาษาอังกฤษมีค่าเฉลี่ยความยากเป็น .42 .43 .43 และ .43 และค่าอำนาจจำแนกเป็น .30 .30 .30 และ .31 ตามลำดับ รายละเอียดดังตารางที่ 51 และ 52 ในภาคผนวก ก

ตอนที่ 3 การพัฒนาเกณฑ์ตัดสินข้อสอบลำเอียง

เนื้อหาในตอนนี้จะแบ่งเป็น 2 หัวข้อคือ

1. การตรวจสอบความเป็นเอกมิติของแบบสอบ
2. การพัฒนาเกณฑ์ตัดสินความลำเอียงของข้อสอบ มีรายละเอียดแต่ละตอนดังนี้

1. การตรวจสอบความเป็นเอกมิติของแบบสอบ

ตรวจสอบความเป็นเอกมิติของแบบสอบ โดยการวิเคราะห์ตัวประกอบ (Factor Analysis) ด้วยวิธีวิเคราะห์ตัวประกอบสำคัญ (Principal Component Analysis) และหมุนแกนด้วยวิธี Varimax ตามที่ Lord (1980) เสนอแนะ ใช้โปรแกรมสำเร็จรูป SPSS วิเคราะห์ โดยใช้กลุ่มตัวอย่างที่สุ่มมาจำนวน 2,000 คน (ชาย 1,000 คน และหญิง 1,000 คน) ผลการวิเคราะห์ดังตารางที่ 5

ตารางที่ 5 ค่าไอเกนและร้อยละของความแปรปรวนของตัวประกอบวิชาคณิตศาสตร์และวิชาภาษาอังกฤษ*

ตัวประกอบที่	คณิตศาสตร์			ภาษาอังกฤษ		
	ญ	ช	รวม	ญ	ช	รวม
1	2.85	3.87	3.39	8.80	8.67	8.70
	(7.1)	(9.7)	(8.5)	(11.0)	(10.8)	(10.9)
2	1.53	1.51	1.46	2.26	2.12	2.14
	(3.8)	(3.8)	(3.7)	(2.8)	(2.6)	(2.7)
3	1.44	1.35	1.33	1.50	1.58	1.4
	(3.6)	(3.4)	(3.3)	(1.9)	(2.0)	(1.8)
4	1.37	1.30	1.26	1.42	1.50	1.36
	(3.4)	(3.2)	(3.1)	(1.8)	(1.9)	(1.7)
5	1.28	1.28	1.21	1.40	1.41	1.29
	(3.2)	(3.2)	(3.0)	(1.8)	(1.8)	(1.6)

* ในวงเล็บคือค่าร้อยละของความแปรปรวน

พบจากตารางที่ 5 ว่าค่าไอเกนสูงสุดของวิชาคณิตศาสตร์ คือ 3.39 และของวิชาภาษาอังกฤษ คือ 8.70 โดยที่ผลการวิเคราะห์แยกกลุ่มตามเพศชายและหญิงมีผลใกล้เคียงกับผลที่ได้จากข้อมูลรวม และคิดเป็นความแปรปรวนร้อยละ 8.5 และ 10.9 ของความแปรปรวนทั้งหมดของแต่ละวิชา ซึ่งพิจารณาแล้วเห็นว่ามีค่าน้อยเกินกว่าที่จะสรุปถึงความเป็นเอกมิติชนิดมีตัวประกอบหลักตัวแรกเด่นกว่าตัวประกอบอื่น ๆ (essential unidimensionality) นอกจากนั้นยังต่ำกว่าค่าความแปรปรวน ซึ่ง Rackase (Rackase อ้างใน Raju, 1993: 21-22) เสนอแนะว่าการใช้ทฤษฎีการตอบข้อสอบวิเคราะห์ข้อสอบนั้น ค่าความแปรปรวนจากตัวประกอบหลักตัวแรกของแบบสอบควรมีค่าไม่ต่ำกว่าร้อยละ 20.00 และด้วยเหตุนี้จึงสรุปได้ว่าการวิเคราะห์ด้วยทฤษฎีการตอบข้อสอบอาจไม่เหมาะสมกับแบบสอบทั้ง 2 ชุดนี้

ผลการวิเคราะห์ข้อสอบด้วยโปรแกรม BILOG โดยใช้โมเดล 3 พารามิเตอร์ พบข้อยืนยันความไม่เหมาะสม โดยมีจำนวนข้อสอบที่ไม่เหมาะกับโมเดลจากการทดสอบนัยสำคัญของโปรแกรม ทั้งยังไม่สามารถประมาณค่าพารามิเตอร์ของข้อสอบบางข้อได้ โดยที่ข้อสอบและจำนวนข้อสอบที่ไม่เหมาะสม และไม่สามารถหาค่าพารามิเตอร์ได้มีความแตกต่างกันไปเมื่อใช้ขนาดกลุ่มตัวอย่างต่างกัน หรือแม้แต่ใช้กลุ่มตัวอย่างขนาดเดียวกันแต่มีขนาดความยาวของข้อสอบต่างกัน

ดังนั้น ผู้วิจัยจึงได้วิเคราะห์ข้อมูลอีกครั้งโดยใช้โมเดล 2 พารามิเตอร์ พบว่ายังมีข้อสอบที่ไม่เหมาะกับโมเดลอยู่บ้าง และมีความแตกต่างกันไปเมื่อใช้กลุ่มตัวอย่างหรือความยาวแบบสอบต่างขนาดกัน แต่ประมาณค่าพารามิเตอร์ได้ทุกข้อ จึงได้ตัดสินใจใช้โมเดล 2 พารามิเตอร์ในการวิเคราะห์ข้อมูลตามทฤษฎีการตอบข้อสอบเพื่อให้สามารถเปรียบเทียบผลการวิเคราะห์ได้ตามวัตถุประสงค์ที่เสนอไว้

2. <u>การพัฒนาเกณฑ์ตัดสินข้อสอบลำเอียง</u>

ดัชนีที่นำมาใช้ในการพัฒนาเกณฑ์ครั้งนี้คือ ดัชนี SA ดัชนี UA ดัชนี α_{MH} และดัชนี β_{SIB} เนื่องจากดัชนี SA และ UA เป็นดัชนีที่ได้จากวิธีการวิเคราะห์ความลำเอียงด้วยทฤษฎีการตอบข้อสอบเช่นเดียวกัน ผู้วิจัยจึงนำเสนอวิธีการพัฒนาเกณฑ์ควบคู่ไปด้วยกัน เช่นเดียวกับดัชนี α_{MH} และดัชนี β_{SIB} ที่แม้ว่าจะมาจากวิธีการต่างกัน แต่โปรแกรมการวิเคราะห์ที่ใช้ให้ผลทั้งค่า α_{MH} และ β_{SIB} ผู้วิจัยจึงขอนำเสนอวิธีการพัฒนาเกณฑ์ไปพร้อม ๆ กันเช่นกัน

2.1 การพัฒนาเกณฑ์ของดัชนี SA และ UA มีวิธีการดังนี้

1) คำนวณแมตริกซ์ค่าพารามิเตอร์และค่าความแปรปรวนรวมของข้อมูลผลการตอบข้อสอบแต่ละแฟ้ม (file) ด้วยโปรแกรม BILOG version 3.04

โปรแกรม BILOG ที่ผู้วิจัยนำมาใช้เป็นโปรแกรมที่ใช้กับเครื่องคอมพิวเตอร์ PC เพื่อวิเคราะห์แบบสอบตามทฤษฎีตอบข้อสอบ ข้อมูลการตอบข้อสอบต้องเป็นข้อมูลแบบ 2 ประเภท เช่น ถูก-ผิด วิธีการวิเคราะห์ที่ใช้ คือ วิธีการประมาณค่าแบบ Marginal maximum likelihood และวิธีของ Bayes โดยที่สามารถให้ผลการประมาณค่าที่ถูกต้องทั้งแบบสอบสั้นและยาวการวิเคราะห์ แบ่งออกเป็น 3 ขั้นตอน ดังนี้

ขั้นตอนที่ 1 เป็นขั้นตอนการแปลงข้อมูลดิบให้อยู่ในรูปข้อมูลสำหรับการวิเคราะห์ในขั้นตอนที่ 2 โดยใช้โปรแกรมย่อยชื่อ INPUT

ขั้นตอนที่ 2 ประมาณค่าพารามิเตอร์และการทดสอบความเหมาะสมของข้อมูลต่อโมเดลที่ใช้ด้วยโปรแกรมย่อย CALIBRATE

ในขั้นตอนครั้งแรกนี้ผู้วิจัยได้เลือกใช้โมเดลการตอบข้อสอบแบบ 3 พารามิเตอร์ ซึ่งโปรแกรม BILOG ใช้ค่า D=1.7 ในการประมาณค่าพารามิเตอร์ของข้อสอบโดยผู้วิจัยได้กำหนดค่าพารามิเตอร์การเดาเป็น 0.25 เท่ากันทุกข้อเพื่อให้สามารถคำนวณพื้นที่ความแตกต่างระหว่างโค้งข้อสอบได้ การประมาณค่าพารามิเตอร์ด้วยโปรแกรม BILOG ใช้วิธี Marginal maximum likelihood ซึ่งสามารถใช้กับแบบสอบขนาด 3 ข้อขึ้นไป การประมาณค่าพารามิเตอร์ของโปรแกรมใช้การทำซ้ำด้วยวิธี EM และ Newton Gauss ซึ่งผู้วิจัยได้ปล่อยให้โปรแกรมดำเนินการเองตามปกติ โดยใช้การประมาณด้วย EM 10 รอบ ตามด้วย Newton-Gauss อีก 2 รอบ สำหรับข้อสอบขนาด 40 และ 50 ข้อ และประมาณด้วย EM 5 รอบ Newton 2 รอบสำหรับข้อสอบขนาด 60-80 ข้อ

การวิเคราะห์สุดท้ายในขั้นตอนที่ 2 คือ การทดสอบความเหมาะสมของข้อสอบต่อโมเดลที่ใช้ ปกติโปรแกรม BILOG ทดสอบความเหมาะสมต่อโมเดลของข้อสอบรายข้อเมื่อมีข้อสอบตั้งแต่ 20 ข้อขึ้นไป โดยไม่ต้องใช้คำสั่งใน OPTIONS สถิติที่ทดสอบ คือ

$$G^2 = 2 \sum_{i=1}^{2^n} r_i \log_e \frac{r_i}{NP_i}$$

เมื่อ 2^n เป็นจำนวนรูปแบบที่เป็นไปได้ทั้งหมดของการตอบข้อสอบ โดยที่ r_i เป็นความถี่ของการตอบในรูปแบบ i N คือ จำนวนของผู้สอบ และ P_i คือ ค่าความน่าจะเป็นที่คำนวณได้ในรูปแบบ i และมีค่าชั้นแห่งความเป็นอิสระ = $2_n - k_n - 1$ เมื่อ k คือ จำนวนพารามิเตอร์ในโมเดลที่ใช้

อย่างไรก็ตามผลการวิเคราะห์ทั้งภาษาอังกฤษความยาว 80 ข้อ และคณิตศาสตร์ความยาว 40 ข้อ พบว่ามีข้อสอบที่ไม่เหมาะสมกับโมเดลการตอบข้อสอบ 3 พารามิเตอร์ เป็นจำนวนมาก ในกลุ่มตัวอย่างขนาดเดียวกัน แต่ละกลุ่มมีทั้งข้อสอบไม่เหมาะสมกับโมเดลที่เป็นข้อเดียวกันและต่างข้อกัน ข้อสอบเหล่านี้ไม่สามารถคำนวณค่าพารามิเตอร์ของข้อสอบได้ โดยเฉพาะอย่างยิ่งค่าพารามิเตอร์ความยาก ทำให้ต้องตัดข้อสอบข้อนั้น ๆ ออกจากการคำนวณค่าพื้นที่ความแตกต่างระหว่างโค้งข้อสอบ เนื่องจากไม่สามารถคำนวณพื้นที่ความแตกต่างได้เช่นกัน

ตัวอย่างข้อสอบที่ไม่เหมาะสมกับโมเดล และไม่สามารถคำนวณค่าพารามิเตอร์ได้ดังนำเสนอในตารางที่ 53 ในภาคผนวก ข ซึ่งสรุปจำนวนข้อสอบที่ต้องตัดออกทั้งหมดเนื่องจากไม่สามารถคำนวณค่าพารามิเตอร์ของข้อสอบได้

จากสาเหตุดังกล่าวทำให้ไม่สามารถเปรียบเทียบผลการตัดสินข้อสอบลำเอียงของแต่ละความยาวได้ตามจุดมุ่งหมายของการวิจัย ผู้วิจัยจึงได้เปลี่ยนใช้โมเดลการตอบข้อสอบ 2 พารามิเตอร์แทน โดยโปรแกรมจะกำหนดค่าพารามิเตอร์การเดาเป็น 0.00 โดยอัตโนมัติแทนการกำหนด ค่าการเดาเป็น 0.25 เท่ากันทุกข้อ ตามที่ผู้วิจัยกำหนดไว้ในตอนแรก

2) เทียบมาตรของค่าพารามิเตอร์สำหรับกลุ่มผู้สอบที่ต้องการตรวจสอบความลำเอียงของข้อสอบด้วยโปรแกรม EQUATE ของ Baker, Al-Karni และ Al-Dosary (1991)

เนื่องจากการเปรียบเทียบโค้งลักษณะข้อสอบของผู้สอบ 2 กลุ่ม ที่ได้มาจากการวิเคราะห์ตามทฤษฎีการตอบข้อสอบ ต้องการการแปลงมาตรของค่าพารามิเตอร์ที่แสดงลักษณะข้อสอบที่ได้จากผู้สอบ 2 กลุ่มให้อยู่บนมาตรเดียวกันก่อนการเปรียบเทียบ ในการวิจัยครั้งนี้ผู้วิจัยได้ใช้วิธีการประมาณค่าสัมประสิทธิ์การเทียบมาตรซึ่งพัฒนาขึ้นโดย Baker Al-Kani และ Al-Dosary (1991)ตามวิธีของ Stocking และ Lord (1983) โดยที่แสดงความสัมพันธ์ระหว่างแมตริกซ์ของค่าพารามิเตอร์ที่ประมาณได้จากกลุ่มตัวอย่าง 2 กลุ่ม ที่ทำแบบสอบเดียวกันด้วยความสัมพันธ์เชิงเส้นของค่าพารามิเตอร์ที่ประมาณได้จากทฤษฎีการตอบข้อสอบแบบ 3 พารามิเตอร์ ในการแปลงค่าความสามารถของผู้สอบในกลุ่มหนึ่งให้อยู่บนมาตรเดียวกันอีกกลุ่มดังนี้

$$\theta_{*} = A\theta_{*} + K \qquad (1)$$

เมื่อ A คือ ค่าความชันของเส้นถดถอย

K คือ intercept

θ_{*} คือ ความสามารถของผู้สอบก่อนการเทียบมาตร

θ_{*} คือ ความสามารถของผู้สอบบนมาตรที่เทียบแล้ว

และสามารถแปลงค่าพารามิเตอร์จากมาตรหนึ่งไปยังอีกมาตรดังนี้

$$a = a_{*} / A \qquad (2)$$

$$b = Ab_{*} + K \qquad (3)$$

ดังนั้น ในการเทียบมาตรตามทฤษฎีการตอบข้อสอบ หลักการประการแรก คือ การคำนวณค่าความชันและค่า intercept ของสมการเชิงเส้นที่จะใช้ในการแปลงแมตริกซ์ของค่าพารามิเตอร์ ซึ่ง Stocking และ Lord (1983) ได้พัฒนาวิธีการคำนวณค่าสัมประสิทธิ์ 2 ตัวนี้ โดยอาศัยโค้งลักษณะแบบสอบในรูปของฟังชันของผลบวกของความแตกต่างกำลังสอง ระหว่างโค้งลักษณะข้อสอบ 2 โค้งบนมาตรความสามารถ ความคิดของ Stocking และ Lord ก็คือ ถ้าทราบค่าพารามิเตอร์ของข้อสอบตามทฤษฎีการตอบข้อสอบแล้ว คะแนนจริงของผู้สอบคนที่ 1 ที่สอบแบบสอบฉบับเดียวกัน 2 ครั้งจะเท่ากัน ถ้าได้มีการเทียบมาตรอย่างถูกต้อง นั่นคือ ถ้าให้ T_i เป็นคะแนนจริงของผู้สอบคนที่ i ที่ยังไม่ได้รับการเทียบมาตรและ T_{i*} เป็นคะแนนจริงของผู้สอบคนที่ i ที่เป็นคะแนนที่เทียบมาตรแล้ว พบว่า

$$F = \frac{1}{N} \sum_{i=1}^{N} (T_i - T_{i*})^2 \qquad (4)$$

จะเป็นฟังชั่นที่ควรจะมีค่าน้อยที่สุด (ซึ่งก็คือ มีความแตกต่างระหว่าง $T_i - T_{i*}$ น้อยที่สุด) เมื่อ N คือ จำนวนของผู้สอบทั้งหมด และภายใต้ทฤษฎีการตอบข้อสอบจะได้ว่า

$$T_i = \sum_{j=1}^{n} P_j(\theta_i, a_j, b_j, c_j) \quad \text{และ}$$

$$T_{i*} = \sum_{j=1}^{n} P_j(\theta_i, a_j, b_j, c_j) \quad (5)$$

เมื่อ n คือ จำนวนข้อสอบที่ใช้ร่วมกัน θ คือ ค่าความสามารถบนมาตรที่ใช้เป็นหลักในการเทียบ a_j และ b_j เป็นผลที่ได้จากการคำนวณตามสมการ (2) และ (3) ส่วนแมตริกซ์ของค่า c_j ไม่ต้องนำมาเทียบมาตร และสามารถเขียนสมการ (4) ใหม่เป็น

$$F = \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{i=1}^{n} P_j(\theta_i, a_j, b_j, c_j) - \sum_{i=1}^{n} P_j(\theta_i, a_i, b_i, c_i)\right] \quad (6)$$

วิธีของ Stocking และ Lord ใช้คำนวณค่า A และ K ที่นำมาแปลงค่าคะแนนจริงที่ได้จากค่าพารามิเตอร์ข้อสอบในแมตริกซ์ที่ยังไม่ได้แปลงค่า (T_i) ให้อยู่ในมาตรเดียวกับแมตริกซ์หลักเป็น T_{i*} โดยที่ค่า A และ K นั้นเป็นค่าที่ทำให้ฟังชั่น F มีค่าน้อยที่สุด

Stocking และ Lord ใช้เทคนิค multivariate search technique ของ Davidson (1959 cited in Baker และ Al-Karni, 1991: 150) และFletcher และ Powell (1963 cited in Baker หน้าเดียวกัน)ซึ่งBaker Al-karni และ Al-Dosory ได้นำมาพัฒนาเป็นโปรแกรม EQUATE โดยใช้ภาษา FORTRAN เพื่อใช้กับคอมพิวเตอร์ขนาดเล็ก ดังที่ได้นำมาใช้ในการวิจัยครั้งนี้

การวิเคราะห์ด้วยโปรแกรม EQUATE ทำการวิเคราะห์โดยใช้แมตริกซ์ที่ดัดแปลงจากแมตริกซ์ค่าความแปรปรวนร่วมที่ได้จากการวิเคราะห์ด้วยโปรแกรม BILOG version 3.04 แต่การใช้แมตริกซ์ในโปรแกรม EQUATE ต้องใช้ครั้งละ 2 แมตริกซ์ โดยแมตริกซ์หนึ่งจะเป็นแมตริกซ์ของกลุ่มอ้างอิง (reference group) อีกแมตริกซ์เป็นแมตริกซ์ของกลุ่มที่ต้องการศึกษา (focal group) เนื่องจากต้องการเทียบมาตรของข้อมูลจากผู้สอบเพศเดียวกันในความยาวแบบสอบ และขนาดผู้สอบเดียวกัน ผู้วิจัยจึงสุ่มจับคู่แมตริกซ์ของผู้สอบแต่ละเพศแต่ละความยาวแบบสอบ และแต่ละขนาดผู้สอบออกเป็น 50 คู่ ผลการ EQUATE ค่าพารามิเตอร์แต่ละคู่ตามโปรแกรมของ Baker Al-karni และ Al-Dosary จะได้ค่า A และค่า K จำนวน 50x6x7 = 2100 คู่ สำหรับข้อมูลการตอบข้อสอบของผู้สอบแต่ละเพศในทุกบริบทของความยาวแบบสอบ และขนาดผู้สอบที่นำมาศึกษา

3) ใช้โปรแกรมคำนวณพื้นที่ของ Raju (1988) ได้แก่ โปรแกรม AREAB และ AREASEB วัดพื้นที่ความแตกต่างระหว่างโค้งลักษณะข้อสอบของผู้สอบ 2 กลุ่ม ที่ได้มีการเทียบมาตรพารามิเตอร์แล้ว โดยใช้แมตริกซ์ความแปรปรวนร่วมแต่ละคู่ที่ใช้ในโปรแกรม EQUATE และแฟ้มข้อมูลค่า A และ K ที่เป็นผลจากการเทียบมาตรของแมตริกซ์ความแปรปรวนร่วมคู่นั้น ในการวิเคราะห์แต่ละครั้ง

โปรแกรมของ Raju เป็นโปรแกรมคำนวณพื้นที่ความแตกต่างระหว่างโค้งลักษณะข้อสอบที่ใช้ได้ทั้งกับโค้งที่ได้จากการใช้ทฤษฎีการตอบข้อสอบ 1, 2 และ 3 พารามิเตอร์ เป็นโปรแกรมที่คำนวณพื้นที่ช่วงเปิด (open interval) คือ ตั้งแต่ความสามารถ $-\infty$ จนถึง $+\infty$ กรณีที่ใช้โมเดล 2 พารามิเตอร์ โปรแกรมจะกำหนดค่าพารามิเตอร์การเดา (c) = 0.00

ค่าสถิติที่ได้จากการวิเคราะห์ด้วยโปรแกรมวัดพื้นที่ของ Raju ประกอบด้วยค่า UA (พื้นที่ที่คำนวณโดยไม่มีเครื่องหมาย) SA (พื้นที่ที่คำนวณโดยมีเครื่องหมาย) ค่าสถิติ Z และค่าสถิติ χ^2 ของ Lord

การคำนวณตามโปรแกรมของราชูมี 2 ขั้นตอน คือ ขั้นแรกใช้โปรแกรม AREAB คำนวณค่า SA และ UA โดยใช้แมตริกซ์ความแปรปรวนร่วมซึ่งเป็นผลจาก BILOG คู่เดียวกับที่ใช้ในโปรแกรม EQUATE และค่า A และ K ของแมตริกซ์ของแต่ละคู่ที่ได้จากโปรแกรม EQUATE จากนั้นใช้โปรแกรม AREASEB เพื่อทดสอบนัยสำคัญตามวิธีของ Lord โดยใช้แมตริกซ์ความแปรปรวนร่วมและแมตริกซ์ที่เป็นผลลัพธ์จาก AREAB เป็นตัวป้อน

4) ทำการวิเคราะห์ตามข้อ 1) 2) และ 3) โดยมีจำนวนข้อมูลที่ต้องทำการวิเคราะห์ (100 กลุ่ม x 7 ขนาดความยาวแบบสอบ x 6 ขนาดกลุ่มตัวอย่าง x 2 เพศ) = 8,400 ครั้ง สำหรับโปรแกรม BILOG ทำการปรับผลจาก BILOG จำนวน 8,400 ไฟล์ เพื่อเตรียมสำหรับโปรแกรม EQUATE และ AREA จากนั้น วิเคราะห์ด้วย EQUATE 4,200 ครั้ง ด้วย AREAB จำนวน 4,200 ครั้ง และด้วย AREASEB อีก 4,200 ครั้ง รวมเป็นจำนวนการวิเคราะห์ทั้งหมด 25,400 ครั้ง ดังแสดงขั้นตอนในภาพที่ 8

5) วิเคราะห์ค่าสถิติเบื้องต้นของดัชนี SA และ UA ซึ่งจากผลการวิเคราะห์ที่ผ่านมา จะได้ค่าดัชนีแต่ละตัวจำนวน 4,200 ค่า เป็นค่าจากวิชาคณิตศาสตร์ 1,800 ค่า และวิชาภาษาอังกฤษ 2,400 ค่า โปรแกรมที่ใช้วิเคราะห์คือ โปรแกรม SPSS-PC$^+$ โดยมีการวิเคราะห์แยกเป็นรายวิชาดังนี้

- วิเคราะห์การกระจายของค่าดัชนี UA และ SA

- คำนวณค่ามัชฌิมเลขคณิต ค่าส่วนเบี่ยงเบนมาตรฐาน และค่าความคลาดเคลื่อนมาตรฐานของดัชนีแต่ละตัว โดยวิเคราะห์ทั้งรายข้อและรวมทุกข้อ วิเคราะห์ทั้งจำแนกตามความยาวแบบสอบและขนาดผู้สอบ และวิเคราะห์รวมทุกความยาวแบบสอบและขนาดผู้สอบ

ในกรณีของค่า SA ซึ่งมีค่าทั้งลบและบวก โดยที่เครื่องหมายลบและบวกแสดงถึงทิศทางที่เบี่ยงเบนออกจากค่าศูนย์ และเนื่องจากต้องการผลที่แสดงถึงปริมาณของความเบี่ยงเบนโดยไม่ต้องการพิจารณาเครื่องหมาย การวิเคราะห์ค่าเฉลี่ยจึงใช้ค่าสัมบูรณ์ของค่า SA ในการวิเคราะห์

6) วิเคราะห์ค่า $|SA|$ และค่า UA ที่แสดงถึงค่าสูงสุดที่เบี่ยงเบนออกจากศูนย์ เมื่อวิเคราะห์โดยใช้กลุ่มตัวอย่างเดียวกัน โดยใช้สูตร $\bar{X} \pm 2.58$ (SE) ทั้งนี้โดยใช้ช่วงความเชื่อมั่นร้อยละ 99.00

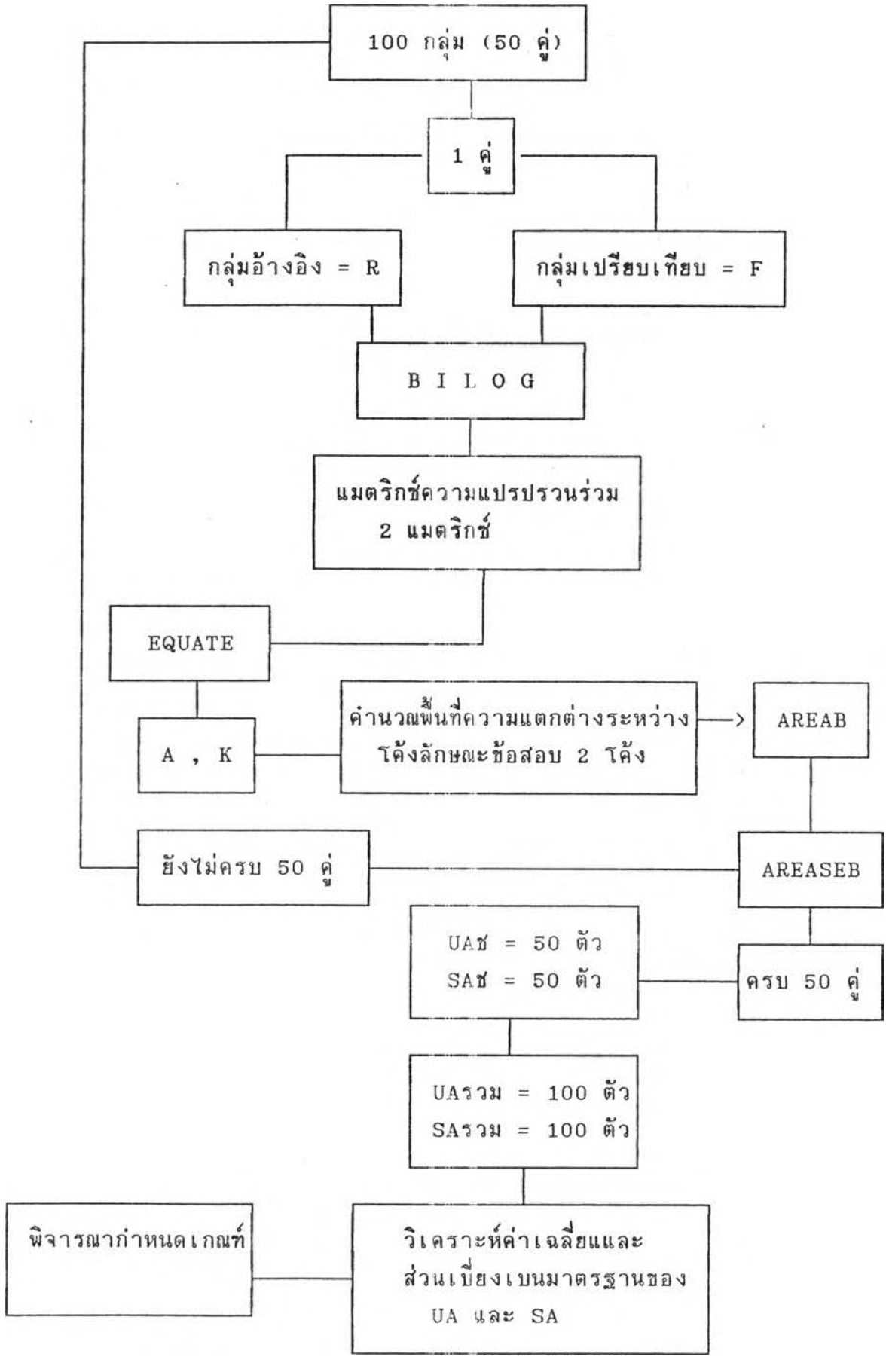
7) กำหนดเกณฑ์ตัดสินข้อสอบลำเอียงของดัชนี SA และ UA โดยใช้ค่าที่มากกว่าค่าสูงสุดจากที่คำนวณได้ในข้อ 5 ทั้งนี้ได้กำหนดเกณฑ์การตัดสินออกเป็น 2 ลักษณะ คือ

ก) กำหนดเกณฑ์เพียงค่าเดียวสำหรับดัชนีแต่ละตัวด้วยการพิจารณาจากค่าเฉลี่ยรวมทุกข้อ ทุกความยาวแบบสอบและทุกขนาดผู้สอบ

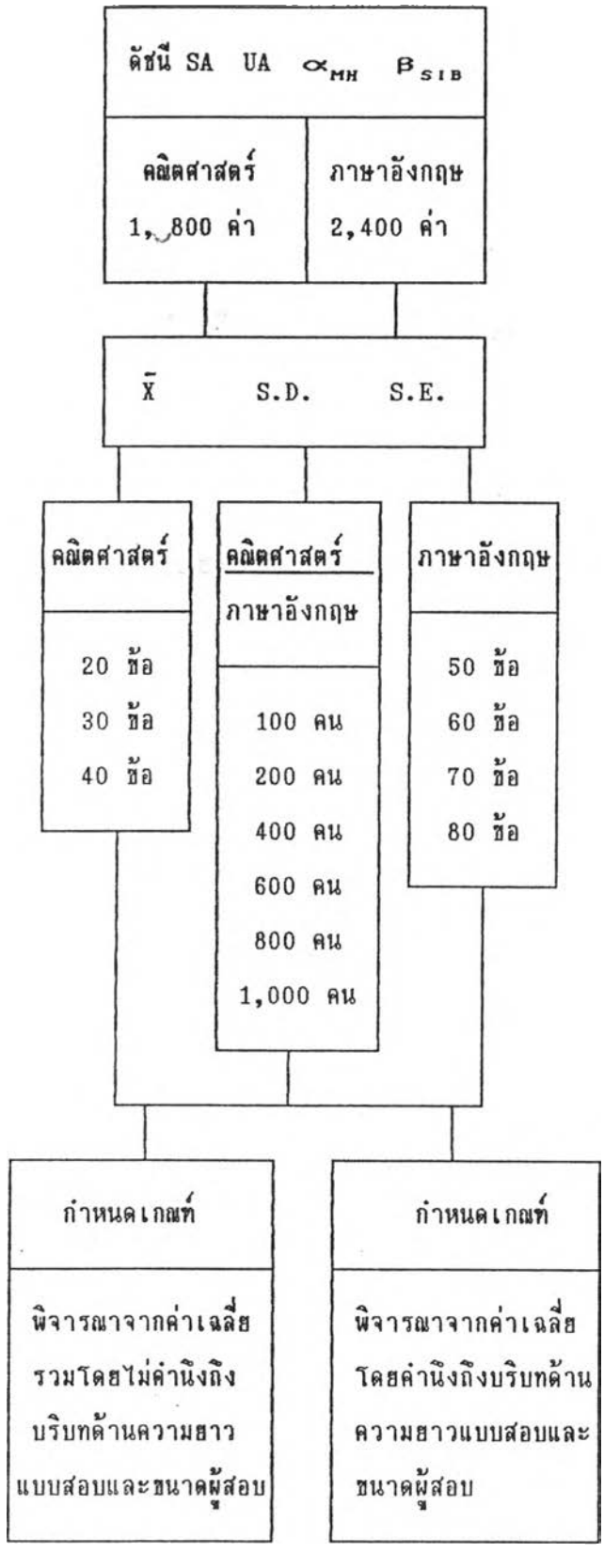
ข) กำหนดเกณฑ์โดยพิจารณาจากความแตกต่างของค่ามัชฌิมเลขคณิตระหว่างความยาวแบบสอบที่ต่างกันและ/หรือความยาวแบบสอบที่ต่างกัน

ดังแสดงขั้นตอนของการพัฒนาเกณฑ์ตั้งแต่ขั้นที่ 4 ไว้ในภาพที่ 9

ภาพที่ 8 ขั้นตอนการพัฒนาเกณฑ์ตัดสินความลำเอียงด้วย IRT-2 สำหรับแต่ละความยาวข้อสอบแต่ละขนาดผู้สอบและแต่ละเพศ



ภาพที่ 9 ขั้นตอนการวิเคราะห์เกณฑ์การตัดสินความลำเอียงของดัชนี SA UA α_{MH} และ β_{SIB} (ตัดตอนเฉพาะการวิเคราะห์ค่าเฉลี่ยของค่าดัชนีแต่ละตัว)



2.2 การพัฒนาเกณฑ์ของดัชนี α_{MH} และ β_{SIB}

ดัชนี α_{MH} และ β_{SIB} เป็นดัชนีที่คำนวณได้พร้อมกันด้วยโปรแกรม SIBTEST ของ Stout และ Rousos (1992)

วิธีคำนวณค่าดัชนี α_{MH} และ β_{SIB} มี 2 ขั้นตอนดังนี้

1) ใช้โปรแกรมย่อย SIBIN (=SIB INPUT) จัดเตรียมข้อมูลจากแฟ้มข้อมูลการตอบแต่ละคู่ (ใช้แฟ้มข้อมูลคู่เดียวกับที่ใช้ในการคำนวณค่าดัชนี SA และ UA) เนื่องจากค่า β_{SIB} และ α_{MH} มีทิศทางที่แสดงถึงความลำเอียงต่อเพศที่เป็นกลุ่มอ้างอิงหรือกลุ่มที่ ต้องการศึกษา (reference group หรือ focal group) ด้วย ในขั้นนี้ผู้วิจัยได้ระบุกลุ่มผู้สอบชายเป็นกลุ่มอ้างอิงในวิชาคณิตศาสตร์ และผู้สอบหญิงเป็นกลุ่มอ้างอิงในวิชาภาษาอังกฤษ

2) ใช้โปรแกรมย่อย SIBTEST วิเคราะห์ค่า β ค่า Z ค่า α_{MH} และ MH-χ^2 โดยใช้แฟ้มข้อมูลซึ่งเป็นผลจากการใช้โปรแกรมย่อย SIBIN เป็นตัวป้อน

รวมการวิเคราะห์ข้อมูลทั้งหมด (2 ครั้ง x 2 เพศ x 50 คู่ x 7 ความยาวแบบสอบ x 6 ขนาดผู้สอบ) = 8,400 ครั้ง

3) ขั้นตอนต่อจากนี้คือ การวิเคราะห์ค่าสถิติ ได้แก่ การวิเคราะห์การกระจายของค่าดัชนีทั้ง 2 ตัว ซึ่งมีทั้งหมดตัวละ 4,200 ค่า ได้จากวิชาคณิตศาสตร์ 1,800 ค่า และภาษาอังกฤษ 2,400 ค่า วิเคราะห์ค่าเฉลี่ยรายข้อ รวมทุกข้อ วิเคราะห์จำแนกตามความยาวแบบสอบ และตามขนาดผู้สอบเช่นเดียวกับดัชนี

ค่า β_{SIB} มีลักษณะเช่นเดียวกับดัชนี SA คือ มีทั้งค่าลบและบวก แล้วแต่จะเบี่ยงเบนไปทางกลุ่มอ้างอิง หรือกลุ่มที่ต้องการศึกษา ดังนั้นการวิเคราะห์ค่าเฉลี่ยจึงใช้ค่าสัมบูรณ์

สำหรับค่า α_{MH} มี 2 ทิศทางเช่นกัน แต่ไม่มีค่าลบ มีค่าเบี่ยงเบนจากค่า 1.00 ไปทางน้อยกว่าและมากกว่า การวิเคราะห์ค่าเฉลี่ยจึงแยกวิเคราะห์ค่าที่มากกว่า 1.00 และค่าที่น้อยกว่า 1.00

การวิเคราะห์ค่าเฉลี่ยมีขั้นตอนดังสรุปไว้ในภาพที่ 9 เช่นกัน

เหตุผลที่ใช้ค่ามัชฌิมเลขคณิตของดัชนีแต่ละตัวในการกำหนดเกณฑ์เป็นเพราะการพบว่า การกระจายของดัชนีทั้ง 4 ตัวที่ได้จากทุกขนาดแบบสอบและผู้สอบมีลักษณะค่อนข้างสมมาตร (Symmetry) ในกรณีที่มีทิศทางของการเบี่ยงเบนออกจากค่าที่แสดงถึงความปราศจากลำเอียงทั้งในทางลบและบวก และมีลักษณะเป็นฐานนิยมเดียว (unimodal) ด้วย นอกจากนั้น เมื่อพิจารณาถึงวัตถุประสงค์ของการวิเคราะห์ที่ต้องการกำหนดขอบเขตของความคลาดเคลื่อนในการสุ่มตัวอย่างอันอาจเกิดขึ้นในการเปรียบเทียบค่าที่ได้จากกลุ่มผู้สอบเพศเดียวกัน อีกทั้งข้อเขียนของ Glass and Hopkins (1984: 191-194) ที่ระบุว่าตัวประมาณค่า

พารามิเตอร์ที่เหมาะสมและไม่ลำเอียง สำหรับทุกลักษณะการกระจายคือค่ามัชฌิมเลขคณิตแล้ว ผู้วิจัยจึงได้ตัดสินใจเลือกใช้ค่ามัชฌิมเลขคณิต แสดงถึงความเบี่ยงเบนโดยเฉลี่ยของค่าดัชนีต่าง ๆ ออกจากค่าที่คาดหวังของดัชนีทั้ง 4 ตัว เมื่อไม่มีความลำเอียงระหว่างกลุ่มผู้สอบ สำหรับผลการวิเคราะห์การกระจายของค่าดัชนีทั้ง 4 ตัว ที่ได้จากทุกความยาวแบบสอบและทุกขนาดผู้สอบได้ผลการกระจายและโค้งการกระจายของค่าดัชนีแต่ละตัวเมื่อเทียบกับโค้งปกติดังภาพที่ 10-17 ในภาคผนวก ค

- ภาพที่ 10 และ 11 แสดงลักษณะการกระจายของค่าดัชนี SA วิชาคณิตศาสตร์ และวิชาภาษาอังกฤษ มีค่าความเบ้ .495 และ .193 ตามลำดับ มีค่าฐานนิยมและค่ามัชฌิมเลขคณิตใกล้เคียงกัน คือ .003 และ .006 สำหรับวิชาคณิตศาสตร์ และ .024 และ .0004 สำหรับวิชาภาษาอังกฤษนิจารณาแล้วเป็นโค้งที่ค่อนข้างจะสมมาตร และมีฐานนิยมเดียว (symmetry and unimodal)

ค่าของดัชนี SA เมื่อโค้งการตอบข้อสอบของกลุ่มผู้สอบ 2 กลุ่ม ไม่แตกต่างกันคือ 0.00 ดังนั้นการคำนวณค่าเฉลี่ยจึงได้ใช้เฉพาะค่าสัมบูรณ์ของดัชนี SA เนื่องจากเครื่องหมายที่ได้เป็นเพียงเครื่องแสดงทิศทางของความเบี่ยงเบนออกจากค่า 0.00 เท่านั้น

- ภาพที่ 12 และ 13 เป็นลักษณะการกระจายของค่าดัชนี UA ซึ่งแม้ค่าเมื่อไม่มีความลำเอียงระหว่างกลุ่มผู้สอบจะเท่ากับ 0.00 เช่นเดียวกับ SA แต่เนื่องจากค่าที่วิเคราะห์ได้จะมีค่าเป็นบวกเท่านั้น ลักษณะของโค้งจึงเบ้ทางขวา และการคำนวณค่าเฉลี่ยจึงใช้ค่าที่ปรากฏอยู่ตามปกติ

- ภาพที่ 14 และ 15 เป็นลักษณะการกระจายของค่าดัชนี α_{MH} ค่า α_{MH} จะแตกต่างจากดัชนี 2 ตัวที่กล่าวมาแล้วตรงที่ว่า เมื่อกลุ่มผู้สอบที่เปรียบเทียบไม่มีความแตกต่างกัน α_{MH} จะมีค่า 1.00 ดังนั้น การเบี่ยงเบนที่เกิดขึ้นต่างทิศกันจะทำให้ได้ค่า α_{MH} มากกว่าหรือน้อยกว่า 1.00 การคำนวณค่าเฉลี่ยจึงได้กระทำใน 2 ลักษณะ คือ ค่าเฉลี่ยเมื่อ $\alpha_{MH} < 1.00$ และค่าเฉลี่ย $\alpha > 1.00$

- ภาพที่ 16 และ 17 แสดงลักษณะการกระจายของดัชนี β_{STD} ซึ่งมีค่าที่แสดงว่าไม่มีความแตกต่างระหว่างการตอบข้อสอบของกลุ่มที่นำมาเปรียบเทียบกัน เป็น 0.00 ดังนั้น การคำนวณค่าเฉลี่ยจึงใช้วิธีเดียวกับ SA คือ คิดเฉพาะค่าสัมบูรณ์ของค่าที่ได้

ตอนที่ 4 การทดลองใช้เกณฑ์คัดสินที่ได้พัฒนาขึ้นจากตอนที่ 3

1. การตรวจสอบความลำเอียงของข้อสอบระหว่างผู้สอบชายและหญิง

สำหรับการวิเคราะห์เพื่อตรวจสอบความลำเอียงของข้อสอบ ระหว่างผู้สอบชายและหญิงนั้นมีวิธีการดังนี้

1.1 วิเคราะห์ความลำเอียงของข้อสอบระหว่างเพศชายและหญิงด้วย IRT-2 พารามิเตอร์ MH และ SIBTEST โดยใช้โปรแกรมและขั้นตอนในการวิเคราะห์เช่นเดียวกับการพัฒนาเกณฑ์แต่ใช้กลุ่มอ้างอิงและกลุ่มเปรียบเทียบต่างเพศกันดังนี้

วิชาภาษาอังกฤษ กลุ่มอ้างอิง ได้แก่ ผู้สอบหญิง กลุ่มเปรียบเทียบ ได้แก่ ผู้สอบชาย ส่วนวิชาคณิตศาสตร์ กลุ่มอ้างอิง ได้แก่ ผู้สอบชาย และกลุ่มเปรียบเทียบคือ ผู้สอบหญิง

แบ่งการวิเคราะห์ตามขนาดกลุ่มตัวอย่าง 6 ขนาด และความยาวข้อสอบ 7 ขนาด (4 ขนาด ในวิชาภาษาอังกฤษ และ 3 ขนาด ในวิชาคณิตศาสตร์) เช่นเดียวกับการวิเคราะห์เพื่อพัฒนาเกณฑ์

1.2. การระบุข้อสอบลำเอียงมีเกณฑ์ 2 ลักษณะ

1.2.1 ใช้เกณฑ์การตรวจสอบด้วยค่าสถิติซึ่งเป็นการทดสอบสมมุติฐานที่ว่าไม่มีความแตกต่างในการตอบข้อสอบระหว่างผู้สอบต่างเพศที่มีความสามารถเท่ากัน วิธี IRT-2 พารามิเตอร์ ทดสอบด้วยสถิติ X^2 ของ Lord โดยมีค่าชั้นแห่งความเป็นอิสระเท่ากับ 2 ใช้ค่า $\alpha \leq 0.05$ วิธี MH ทดสอบด้วย MH-X^2 และ SIBTEST ทดสอบด้วยค่า Z ทั้ง 2 วิธีหลังใช้ผลการทดสอบจากโปรแกรมที่ใช้

1.2.2 เปรียบเทียบค่าสถิติ SA UA α_{MH} และ β_{SIB} ที่คำนวณได้กับเกณฑ์ที่พัฒนาขึ้น

2. การวิเคราะห์ความสอดคล้องของผลการวิเคราะห์และการตัดสินความลำเอียงของข้อสอบ

ผู้วิจัยได้พิจารณาถึงความสอดคล้องของการตัดสินข้อสอบลำเอียงจากจำนวนข้อสอบที่ถูกระบุว่าลำเอียง ข้อสอบที่ลำเอียงและร้อยละของความสอดคล้องของข้อสอบที่ลำเอียงสำหรับแต่ละวิชาในสถานการณ์ต่าง ๆ ดังนี้

2.1 เมื่อใช้เกณฑ์การตัดสินเดียวกัน (การทดสอบสมมุติฐานหรือใช้เกณฑ์ ที่พัฒนาขึ้น) จำนวนข้อสอบและกลุ่มผู้สอบขนาดเดียวกัน แต่ใช้วิธีการวิเคราะห์ต่างกัน

2.2 เมื่อใช้เกณฑ์การตัดสินเดียวกัน วิธีวิเคราะห์เดียวกันแต่มีจำนวนข้อสอบหรือขนาดผู้สอบต่างกัน

2.3 เมื่อใช้เกณฑ์การตัดสินต่างกัน วิธีวิเคราะห์เดียวกันและมีจำนวนข้อสอบหรือขนาดผู้สอบเดียวกัน

สูตรการวิเคราะห์ความสอดคล้อง คือ

$$\text{ความสอดคล้อง} = \frac{R}{T} \times 100$$

เมื่อ R คือ จำนวนข้อสอบที่ซ้ำกันในการตัดสิน

T คือ จำนวนข้อสอบที่ลำเอียงทั้งหมดในการเปรียบเทียบแต่ละคู่