

371.261

ท 3-๒๗

๙ 3

การเปรียบเทียบอำนาจการตรวจสอบและการจำแนกชนิดผลาด
ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอและแบบไม่สม่ำเสมอ
ระหว่างวิธีแมนเทล-แฮนส์เซลและวิธีการถดถอยโลจิสติก
โดยใช้ความยาวแบบทดสอบและขนาดกลุ่มตัวอย่างต่างกัน

ปริญญานิพนธ์

ของ

ทองอยู่ สาระ

เสนอต่อมหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร เพื่อเป็นส่วนหนึ่งของการศึกษา
ตามหลักสูตรปริญญาการศึกษามหาบัณฑิต วิชาเอกการวัดผลการศึกษา

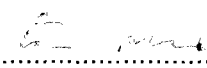
พฤษภาคม 2543

ลิขสิทธิ์เป็นของมหาวิทยาลัยศรีนครินทรวิโรฒ

คณะกรรมการควบคุมและคณะกรรมการสอบได้พิจารณาปริญญานิพนธ์ฉบับนี้แล้ว เห็นสมควร
รับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาการศึกษามหาบัณฑิต วิชาเอก การวัดผลการศึกษา
ของมหาวิทยาลัยศรีนครินทรวิโรฒได้

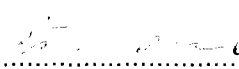
คณะกรรมการควบคุม


.....ประธาน
(รองศาสตราจารย์ ดร.บุญเชิด ภิญโญนนตพงษ์)

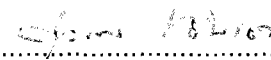

.....กรรมการ
(รองศาสตราจารย์ นิภา ศรีไพโรจน์)

คณะกรรมการสอบ

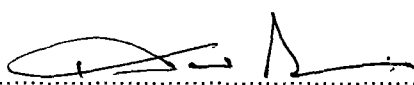

.....ประธาน
(รองศาสตราจารย์ ดร.บุญเชิด ภิญโญนนตพงษ์)


.....กรรมการ
(รองศาสตราจารย์ นิภา ศรีไพโรจน์)


.....กรรมการที่แต่งตั้งเพิ่มเติม
(รองศาสตราจารย์ วัณญา วิศาลาภรณ์)


.....กรรมการที่แต่งตั้งเพิ่มเติม
(อาจารย์ ดร.สุพร เข้มเฮง)

บัณฑิตวิทยาลัยอนุมัติให้รับปริญญานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาการศึกษามหาบัณฑิต วิชาเอกการวัดผลการศึกษา ของมหาวิทยาลัยศรีนครินทรวิโรฒ


.....คณบดีบัณฑิตวิทยาลัย
(ศาสตราจารย์ ดร.เสริมศักดิ์ วิศาลาภรณ์)

วันที่.....เดือน พฤษภาคม พ.ศ 2543

ประกาศคุณูปการ

ปริญญานิพนธ์ฉบับนี้สำเร็จลงได้ด้วยความช่วยเหลือและเอาใจใส่เป็นอย่างดียิ่งของ คณะกรรมการที่ปรึกษาปริญญานิพนธ์ ได้แก่ รองศาสตราจารย์ ดร.บุญเชิด ภิญโญอนันตพงษ์ ในฐานะ ประธานที่ปรึกษาปริญญานิพนธ์ รองศาสตราจารย์นภา ศรีไพโรจน์ ในฐานะกรรมการที่ปรึกษาปริญญานิพนธ์ รองศาสตราจารย์วิญญา วิศาลาภรณ์ และ ดร.สุพร เข้มแข็ง ในฐานะกรรมการสอบปริญญานิพนธ์ และ รองศาสตราจารย์อังคณา สายยศ ในฐานะอาจารย์ที่ปรึกษา ที่ได้กรุณาช่วยชี้แนะแนวทางในการแก้ปัญหา พร้อมทั้งให้ความรู้และข้อคิดเห็น เพื่อนำไปแก้ไขและปรับปรุงข้อบกพร่องต่าง ๆ ของปริญญานิพนธ์ และให้ กำลังใจแก่ผู้วิจัยด้วยความเอาใจใส่อย่างดียิ่งตลอดมา

อาจารย์อัมพร วิชัยศรี อาจารย์ลัดดา เสือสืบพันธุ์ อาจารย์สุทธิ กระจะจ่าง อาจารย์จิราพร แดงก่อ และอาจารย์อรุณศรี แสงแก้ว ที่กรุณาตรวจสอบเครื่องมือในการวิจัย ตลอดจนให้คำแนะนำและ ข้อคิดเห็นต่าง ๆ ในการแก้ไขและปรับปรุงเครื่องมือที่ใช้ในการวิจัยครั้งนี้

คณาจารย์ภาควิชาวัดผลและวิจัยทางการศึกษาทุกท่านที่ได้อบรมสั่งสอนวิชาความรู้ในศาสตร์ เกี่ยวกับการวัดผลการศึกษาแก่ผู้วิจัย

อาจารย์มานพ นพศิริกุล ผู้อำนวยการโรงเรียนศรีอยุธยา อาจารย์ประวิทย์ พุทธิพิทูล ผู้อำนวยการ โรงเรียนสตรีมหาพฤฒาราม อาจารย์กาญจนา สังวรวงศา ผู้ช่วยผู้อำนวยการฝ่ายวิชาการโรงเรียนศรีอยุธยา อาจารย์ลำเจียก เตราชูสงฆ์ อาจารย์ศิริพรรณ อรรถวิน อาจารย์ภาวิตา หาญวัฒนกุล และพี่ ๆ เพื่อน ๆ น้อง ๆ ชาวศรีอยุธยาทุกท่าน ตลอดจนนิสิตสาขาการวัดผลการศึกษาทุกคน ที่คอยช่วยเหลือและให้กำลังใจ ในการเรียนและการทำวิจัยในครั้งนี้อย่างยิ่งตลอดมา

อาจารย์อรินทร์ น่วมถนอม และ อาจารย์วัลลิมาต แซ่ฮ้อ ที่กรุณาให้ความอนุเคราะห์โปรแกรม คอมพิวเตอร์และช่วยเหลือในการวิเคราะห์ข้อมูลด้วยระบบคอมพิวเตอร์ ตลอดจนให้คำแนะนำต่าง ๆ ในการ ปรับปรุงและแก้ไขปริญญานิพนธ์เป็นอย่างดี

อาจารย์สุกัญญา ภูริยะพันธ์ อาจารย์ธีรา พงศ์ศาสตร์ อาจารย์ทิวดี นกบิน อาจารย์สุภาวค์ เรืองศรี อาจารย์ประสิทธิ์ เมืองหมุด และอาจารย์วิรัตน์ กิจสมพงษ์ ที่ให้ความช่วยเหลือในการเก็บรวบรวม ข้อมูลที่ใช้ในการวิจัยครั้งนี้เป็นอย่างดี

ผู้บริหารโรงเรียน คณะครูโรงเรียนที่เป็นกลุ่มตัวอย่าง และขอขอบใจนักเรียนทุกคนที่ให้ความร่วมมือ ในการเก็บข้อมูลเป็นอย่างดี

สุดท้ายนี้ ผู้วิจัยขอกราบขอบพระคุณ คุณพ่อ คุณแม่และขอขอบคุณ พี่ ๆ น้อง ๆ ทุกคนโดยเฉพาะ คุณณฤชณา แก้วประเสริฐ ที่สนับสนุนทั้งด้านกำลังใจกำลังใจทรัพย์และคอยดูแลช่วยเหลือในทุกสิ่งทุกอย่างแก่ ผู้วิจัยอย่างดีเสมอมา

ทองอยู่ สาระ

สารบัญ

บทที่	หน้า
1 บทนำ.....	1
ภูมิหลัง.....	1
ความมุ่งหมายของการศึกษาค้นคว้า.....	4
ความสำคัญของการศึกษาค้นคว้า.....	4
ขอบเขตของการศึกษาค้นคว้า.....	4
นิยามศัพท์เฉพาะ.....	6
2 เอกสารและงานวิจัยที่เกี่ยวข้อง.....	9
ความหมายของการทำหน้าที่ต่างกันของข้อสอบ.....	9
ประเภทของการทำหน้าที่ต่างกันของข้อสอบ.....	9
หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ.....	10
วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ.....	10
งานวิจัยที่เกี่ยวข้องกับการทำหน้าที่ต่างกันของข้อสอบ.....	26
สมมติฐานการวิจัย.....	37
3 วิธีดำเนินการศึกษาค้นคว้า.....	38
ประชากร.....	38
กลุ่มตัวอย่าง.....	38
เครื่องมือที่ใช้ในการศึกษาค้นคว้า.....	39
วิธีดำเนินการสร้างเครื่องมือที่ใช้ในการศึกษา.....	40
ลักษณะเนื้อหาของข้อสอบ.....	41
การเก็บรวบรวมข้อมูล.....	45
การวิเคราะห์ข้อมูล.....	45
4 ผลการวิเคราะห์ข้อมูล.....	50
ตอนที่ 1 การวิเคราะห์ข้อมูลเบื้องต้น.....	51
ตอนที่ 2 การวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันตามทฤษฎีการตอบข้อสอบ	
ด้วยวิธีการวัดพื้นที่ของราชู(EUA).....	53
ตอนที่ 3 การวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันด้วยวิธีแมนเทล-แฮนส์เซล.....	71
ตอนที่ 4 การวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันด้วยวิธีการถดถอยโลจิสติก.....	74
ตอนที่ 5 การวิเคราะห์อำนาจการตรวจสอบและการจำแนกผิดพลาด.....	77
ตอนที่ 6 การเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาด	
ระหว่างวิธีแมนเทล-แฮนส์เซล(MH)กับวิธีการถดถอยโลจิสติก(LR).....	80

บทที่	หน้า
5 สรุป อภิปรายผล ข้อเสนอแนะ.....	100
สรุปผลการวิจัย.....	101
อภิปรายผล.....	107
ข้อเสนอแนะ.....	113
บรรณานุกรม.....	114
ภาคผนวก.....	119
ประวัติย่อของผู้วิจัย.....	121

บัญชีตาราง

ตาราง		หน้า
1	วิธีตรวจสอบ DIF จำแนกตามการให้คะแนนของข้อมูล ทฤษฎีที่ใช้ในการวิเคราะห์ และรูปแบบของโมเดล.....	11
2	ผลการตอบข้อสอบระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ณ ระดับคะแนน j.....	18
3	สัดส่วนของผลการตอบข้อสอบระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ณ ระดับคะแนน j	19
4	แสดงรายชื่อโรงเรียน จำนวนห้องเรียนและจำนวนนักเรียนที่เป็นกลุ่มตัวอย่าง.....	39
5	ความยาวของแบบทดสอบและขนาดกลุ่มตัวอย่างที่ศึกษา.....	47
6	ค่าสถิติพื้นฐานของแบบทดสอบความสามารถทางสมองตามแนวโครงสร้างของ โอติส-เลนนอน ความยาว 20 ข้อ, 40 ข้อ, 60 ข้อ และ 80 ข้อ.....	51
7	ค่าความยากง่าย ค่าอำนาจจำแนก และค่าความเชื่อมั่น (KR-20) ของแบบทดสอบ ความสามารถทางสมองตามแนวโครงสร้างของโอติส-เลนนอน.....	52
8	การวิเคราะห์องค์ประกอบของแบบทดสอบความยาว 20 ข้อ.....	53
9	การวิเคราะห์องค์ประกอบของแบบทดสอบความยาว 40 ข้อ.....	54
10	การวิเคราะห์องค์ประกอบของแบบทดสอบความยาว 60 ข้อ.....	55
11	ค่าพารามิเตอร์อำนาจจำแนกและความยากของแบบทดสอบความยาว 60 ข้อ.....	56
12	ค่าพารามิเตอร์อำนาจจำแนกและความยากของแบบทดสอบความยาว 40 ข้อ.....	58
13	ค่าพารามิเตอร์อำนาจจำแนกและความยากของแบบทดสอบความยาว 20 ข้อ.....	59
14	ค่า A และค่า K ที่ได้จากการปรับเทียบค่าพารามิเตอร์ข้อสอบของแบบทดสอบ ความยาว 20 ข้อ, 40 ข้อ และ 60 ข้อ ระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ.....	60
15	การตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน ของแบบทดสอบความยาว 60 ข้อ.....	61
16	การตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน ของแบบทดสอบความยาว 40 ข้อ.....	66
17	การตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน ของแบบทดสอบความยาว 20 ข้อ.....	69
18	การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบแบบสมำเสมอและแบบไม่สมำเสมอด้วยวิธี MH แต่ละความยาวของแบบทดสอบและกลุ่มตัวอย่างแต่ละขนาด.....	71
19	การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบแบบสมำเสมอและแบบไม่สมำเสมอด้วยวิธี LR แต่ละความยาวของแบบทดสอบและกลุ่มตัวอย่างแต่ละขนาด.....	74
20	การวิเคราะห์อำนาจการตรวจสอบและการจำแนกผิดพลาดในการตรวจสอบการทำหน้าที่ ต่างกันของข้อสอบแบบสมำเสมอและแบบไม่สมำเสมอด้วยวิธี MH และวิธี LR.....	78
21	การเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาดในการตรวจสอบข้อสอบ ที่ทำหน้าที่ต่างกันแบบสมำเสมอ ระหว่างวิธีวิธี MH และวิธี LR เมื่อใช้แบบทดสอบความยาวเท่ากัน.....	80
22	การเปรียบเทียบอำนาจการตรวจสอบในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน แบบสมำเสมอ ระหว่างแบบทดสอบความยาวต่างกัน เมื่อตรวจสอบด้วยวิธี MH.....	81

ตาราง

หน้า

38	การเปรียบเทียบการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน แบบไม่สม่ำเสมอ ระหว่างกลุ่มตัวอย่างขนาดต่างกัน เมื่อตรวจสอบด้วยวิธี MH...	95
39	การเปรียบเทียบอำนาจการตรวจสอบในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน แบบไม่สม่ำเสมอ ระหว่างกลุ่มตัวอย่างขนาดต่างกัน เมื่อตรวจสอบด้วยวิธี LR....	96
40	การเปรียบเทียบการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน แบบไม่สม่ำเสมอ ระหว่างกลุ่มตัวอย่างขนาดต่างกัน เมื่อตรวจสอบด้วยวิธี LR....	98

บัญชีภาพประกอบ

ภาพประกอบ	หน้า
1 ข้อสอบทำหน้าที่ต่างกันแบบสม่ำเสมอ (Uniform DIF)	10
2 ข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ (Nonuniform DIF)	10
3 ลำดับขั้นในการดำเนินการสร้างแบบทดสอบวัดความสามารถทางสมอง.....	40
4 แสดงขั้นตอนการวิเคราะห์ข้อมูล.....	46
5 การตรวจสอบความเป็นมิติเดียวของแบบทดสอบความยาว 20 ข้อ ด้วยการทดสอบสกรี	53
6 การตรวจสอบความเป็นมิติเดียวของแบบทดสอบความยาว 40 ข้อ ด้วยการทดสอบสกรี	54
7 การตรวจสอบความเป็นมิติเดียวของแบบทดสอบความยาว 60 ข้อ ด้วยการทดสอบสกรี	55
8 เส้นภาพข้อสอบที่ทำหน้าที่ต่างกัน ของแบบทดสอบความยาว 60 ข้อ.....	63
9 เส้นภาพข้อสอบที่ทำหน้าที่ต่างกัน ของแบบทดสอบความยาว 40 ข้อ.....	67
10 เส้นภาพข้อสอบที่ทำหน้าที่ต่างกัน ของแบบทดสอบความยาว 20 ข้อ.....	70

บทที่ 1

บทนำ

ภูมิหลัง

คุณสมบัติที่สำคัญที่สุดที่บ่งชี้ถึงคุณภาพของแบบทดสอบ คือ ความเที่ยงตรง (Validity) ซึ่งความเที่ยงตรงเป็นคุณสมบัติที่แสดงถึงความสามารถในการวัดคุณลักษณะแฝงได้ตามที่ต้องการวัดหรือแบบทดสอบทำหน้าที่ได้ตามวัตถุประสงค์ที่กำหนดไว้ การตรวจสอบว่าข้อสอบมีความเที่ยงตรงหรือไม่นั้นมีหลายวิธี และวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบก็เป็นอีกวิธีหนึ่งที่ใช้เพื่อตรวจสอบความเที่ยงตรงของข้อสอบ แบบทดสอบที่สอบกับกลุ่มผู้สอบย่อยตั้งแต่สองกลุ่มขึ้นไป แล้วนำผลการสอบมาวิเคราะห์พบว่าผู้สอบที่มีความสามารถหรือมีคุณลักษณะแฝงที่ต้องการวัดเท่ากัน แต่อยู่ต่างกลุ่มกันแล้วมีโอกาสในการตอบข้อสอบได้ถูกไม่เท่ากัน แสดงว่าแบบทดสอบขาดความเที่ยงตรง เพราะไม่ได้วัดคุณลักษณะแฝงเป้าหมายตามที่ต้องการวัด แต่อาจวัดคุณลักษณะแฝงแทรกซ้อนอื่นที่ไม่ต้องการวัดอีกด้วย หากผู้สอบกลุ่มย่อยกลุ่มใดมีคุณลักษณะแฝงแทรกซ้อนนั้นสูงกว่าย่อมมีโอกาสที่จะตอบข้อสอบได้ถูกมากกว่าทั้ง ๆ ที่มีคุณลักษณะแฝงเป้าหมายเท่ากันกับผู้สอบกลุ่มย่อยกลุ่มอื่น ซึ่งทำให้เกิดการได้เปรียบเสียเปรียบกันระหว่างกลุ่มผู้สอบย่อย ลักษณะเช่นนี้เดิมใช้คำว่าความลำเอียงของข้อสอบ (Item Bias) หรือความลำเอียงของแบบทดสอบ (Test Bias) ซึ่งเป็นภาษาที่ใช้กันในทางสังคมและมีความหมายในเชิงลบ ต่อมาในระยะหลังนักวิจัยได้เปลี่ยนมาใช้คำใหม่ว่า การทำหน้าที่ต่างกันของข้อสอบ (Differential Item Function : DIF) หรือการทำหน้าที่ต่างกันของแบบทดสอบ (Differential Test Function : DTF) ซึ่งเป็นคำที่มีความเป็นกลางและเหมาะสมกว่า (วลิมาต แซ็ง 2543 : 26 ; อ้างอิงจาก Holland and Wainer. 1993)

เมลเลนเบอร์ก (วลิมาต แซ็ง 2543 : 28 ; อ้างอิงจาก Mellenbergh. 1982) ได้จำแนกรูปแบบของการทำหน้าที่ต่างกันของข้อสอบออกเป็นสองประเภท คือ ข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอ (Uniform DIF) ซึ่งจะเกิดขึ้นเมื่อไม่มีปฏิสัมพันธ์ (Interaction) ระหว่างระดับความสามารถของผู้สอบและการเป็นสมาชิกของกลุ่มผู้สอบและข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ (Nonuniform DIF) ซึ่งจะเกิดขึ้นเมื่อมีปฏิสัมพันธ์ระหว่างระดับความสามารถของผู้สอบและการเป็นสมาชิกของกลุ่ม ในทฤษฎีการตอบข้อสอบ (Item Response Theory : IRT) สามารถพิจารณา “ปฏิสัมพันธ์” ได้จากความแตกต่างของค่าพารามิเตอร์อำนาจจำแนก (a-Parameter) ของข้อสอบระหว่างผู้สอบกลุ่มย่อยสองกลุ่ม ถ้าข้อสอบทำหน้าที่ต่างกันแบบสม่ำเสมอแล้วโค้งลักษณะข้อสอบ (Item Characteristic Curves : ICCs) ระหว่างผู้สอบกลุ่มย่อยสองกลุ่มจะขนานกัน แต่ถ้าข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอแล้วโค้งลักษณะข้อสอบระหว่างผู้สอบกลุ่มย่อยสองกลุ่มจะไม่ขนานกัน

สำหรับวิธีวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบนั้นในปัจจุบันมีหลายวิธี ซึ่งโปเทนซาและโดแรนส์ (Potenza and Dorans. 1995) ได้จำแนกประเภทของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบออกเป็นหลายลักษณะ กล่าวคือ ถ้าจำแนกตามการให้คะแนนของข้อมูลจะจำแนกได้สองวิธี คือ การทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนแบบทวิพันธ์ (Dichotomous Score : คือการให้คะแนนเป็น 1

เมื่อตอบถูก และให้คะแนนเป็น 0 เมื่อตอบผิด) และการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนแบบพหุพจน์ (Polytomous Score) และถ้าจำแนกตามทฤษฎีการวิเคราะห์สามารถแบ่งออกเป็นสองกลุ่ม คือ กลุ่มวิธี nonIRT (Non Item Response Theory Methods) และกลุ่มวิธี IRT กลุ่มวิธี nonIRT จะวิเคราะห์ดัชนีการทำหน้าที่ต่างกันของข้อสอบโดยใช้คะแนนที่สอบได้ (Observe Score) วิธีการในกลุ่มนี้ได้แก่ วิธีแมนเทล-แฮนส์เซล (Holland and Thayer. 1988) วิธี Standardization (Dorans and Holland. 1993) และวิธีการถดถอยโลจิสติก (Swaminathan and Rogers. 1990) กลุ่มวิธี IRT จะวิเคราะห์ดัชนีการทำหน้าที่ต่างกันของข้อสอบโดยใช้คะแนนที่สอบวัดโดยตรงไม่ได้หรือตัวแปรแฝง (Latent Variable) วิธีการในกลุ่มนี้ได้แก่วิธีที่มีพื้นฐานจากทฤษฎีการตอบข้อสอบ (IRT) และวิธีซิปเทสต์ (SIBTEST) (Shealy and Stout. 1993)

วิธีที่ใช้ทฤษฎีการตอบข้อสอบ เป็นวิธีที่มีข้อดีในทางทฤษฎี (Holland and Thayer. 1988 ; Hambleton and others. 1991) เนื่องจากใช้การประมาณค่าความน่าจะเป็นสูงสุด (Marginal Maximum Likelihood Estimation : MMLE) หรือ การประมาณค่ารวมแบบเบย์ (Marginal Bayes Estimation : MBE) ในการทดสอบพารามิเตอร์ของโมเดล IRT จึงมีอำนาจทดสอบสูง แต่ข้อจำกัดที่สำคัญของวิธีนี้ก็คือกระบวนการวิเคราะห์ค่อนข้างยุ่งยาก ซับซ้อน เสียค่าใช้จ่ายสูง กลุ่มตัวอย่างต้องมีขนาดใหญ่ ทำให้เกิดความยุ่งยากในการปฏิบัติ ด้วยเหตุนี้จึงได้มีการศึกษาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยหลีกเลี่ยงการใช้ทฤษฎีการตอบข้อสอบ คือวิธีที่ใช้สถิติแบบพาราเมตริก (Parametric Form) เช่นวิธีการถดถอยโลจิสติก (Logistic Regression : LR) และวิธีที่ใช้สถิติแบบนอปปาราเมตริก (Non-parametric Form) เช่น วิธีแมนเทล-แฮนส์เซล (Mantel-Haenszel : MH) วิธีซิปเทสต์ (SIBTEST) และวิธี Standardization (STND) ซึ่งวิธี MH เป็นวิธีที่มีข้อดีในทางปฏิบัติ (Hambleton and others. 1989 ; Holland and Wainer. 1993) เนื่องจากสามารถคำนวณได้ง่าย เพราะใช้ตารางการถ่วง (Contingency Table) และสถิติไค-กำลังสอง แต่ข้อจำกัดที่สำคัญของวิธีนี้ก็คือการใช้คะแนนที่สอบวัดโดยตรงแทนความสามารถของกลุ่มผู้สอบทำให้ผลการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันอาจเปลี่ยนแปลงไป เมื่อใช้กลุ่มตัวอย่างแตกต่างกัน ในปัจจุบันนี้วิธีวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบที่ได้รับความสนใจและเป็นที่ยอมรับมากคือวิธี MH (Clauser and others. 1991) สำนักงานบริการการทดสอบทางการศึกษา (Education Testing Service : ETS) ของสหรัฐอเมริกา ได้แนะนำให้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี MH และวิธีนี้ได้ถูกนำไปใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในการทดสอบของสำนักงานบริการการทดสอบทางการศึกษาแห่งสหรัฐอเมริกามากที่สุด (Zieky.1993 ; Dorans and Holland. 1993)

ลักษณะสำคัญที่ทำให้วิธี MH ได้รับความนิยมนำมาใช้ในปัจจุบันนี้คือ วิธี MH มีหลักการที่ดีสามารถเข้าใจได้ง่าย กระบวนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบไม่ยุ่งยากซับซ้อน และสามารถปฏิบัติได้ง่ายกว่าวิธีที่ใช้ทฤษฎีการตอบข้อสอบ มีเทคนิคทางสถิติเชื่อถือได้ มีการทดสอบนัยสำคัญทางสถิติ ใช้ได้ดีเมื่อกลุ่มตัวอย่างมีขนาดเล็ก จึงไม่มีข้อจำกัดในเรื่องขนาดกลุ่มตัวอย่างเหมือนวิธี IRT นอกจากนั้นยังสามารถใช้กับกลุ่มตัวอย่างที่กลุ่มประชากรย่อยมีขนาดไม่เท่ากันได้ เสียค่าใช้จ่ายในการดำเนินการไม่สูงเหมือนวิธี IRT และผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยวิธี MH สอดคล้องกับวิธีอื่นๆ เช่นวิธี IRT วิธี LR และวิธี Standardization (Holland and Thayer. 1988; Hambleton and Rogers.1989; Swaminathan and Rogers. 1990 ; Millsap and Everson. 1993 ; Dorans and Holland. 1993 ; Clauser and others.1991) วิธี LR เป็นวิธีการทดสอบสถิติแบบพารามิเตอร์มีการทดสอบโมเดลทางสถิติ วิธีนี้สามารถ

ใช้วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบได้ดี ทั้งแบบสม่ำเสมอ (Uniform DIF) และแบบไม่สม่ำเสมอ (Nonuniform DIF) แต่จะเสียค่าใช้จ่ายมากกว่าวิธี MH ประมาณ 3-4 เท่า (Swaminathan and Rogers. 1990)

จากผลการศึกษาของนักวิจัยที่ใช้วิธี MH และ LR ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันพบว่า ในกรณีข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอทั้งสองวิธีมีอำนาจการตรวจสอบใกล้เคียงกัน (Swaminathan and Rogers. 1990 ; Rogers and Swaminathan. 1993) ส่วนในกรณีข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ วิธี LR มีอำนาจการตรวจสอบมากกว่าวิธี MH (Swaminathan and Rogers. 1990 ; Rogers and Swaminathan. 1993 ; Narayanan and Swaminathan. 1996)

นอกจากนี้ผลจากการเปรียบเทียบวิธีการในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ พบว่า ปัจจัยสำคัญที่ส่งผลต่ออำนาจการตรวจสอบ คือ ขนาดของกลุ่มตัวอย่าง เช่นจากการศึกษาของโรเจอร์สและสวามินาธาน (Rogers and Swaminathan. 1993) เมเซอร์และคณะ (Mazor and others. 1992) นารายานาน และสวามินาธาน (Narayanan and Swaminathan. 1996) ต่างก็ได้ข้อค้นพบที่สอดคล้องกัน คือ เมื่อขนาดของกลุ่มตัวอย่างใหญ่ขึ้น อำนาจการตรวจสอบจะเพิ่มขึ้น

ปัจจัยที่ส่งผลต่ออำนาจการตรวจสอบอีกอย่างหนึ่งคือ ความยาวของแบบทดสอบ ทั้งนี้เพราะว่าแบบทดสอบที่มีความยาวมากกว่าย่อมส่งผลให้มีความน่าเชื่อถือมากกว่า ทำให้การจับคู่เปรียบเทียบระหว่างผู้สอบมีความถูกต้องมากขึ้น เช่น จากการศึกษาของสวามินาธานและโรเจอร์ส (Swaminathan and Rogers. 1990) โดยใช้ข้อมูลจำลองและตรวจสอบด้วยวิธี MH กับ LR ใช้แบบทดสอบความยาว 40 ข้อ, 60 ข้อ และ 80 ข้อ พบว่าเมื่อใช้แบบทดสอบยาว อำนาจการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจะดีกว่าการใช้แบบทดสอบสั้น ต่อมาในปี 1993 ทั้งสองได้ศึกษาอีกครั้งหนึ่ง โดยใช้แบบทดสอบความยาว 40 และ 80 ข้อ ตรวจสอบด้วยวิธี MH และ วิธี LR พบว่า ความยาวของแบบทดสอบไม่มีผลกระทบต่ออำนาจการตรวจสอบ

จากการศึกษางานวิจัยที่ผ่านมา มีประเด็นที่น่าสนใจคือ

1.วิธีการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอและแบบไม่สม่ำเสมอด้วยวิธี MH และวิธี LR ยังไม่มีความชัดเจนว่าวิธีใดมีประสิทธิภาพดีกว่า และมีความเหมาะสมในทางปฏิบัติมากกว่ากัน

2.ขนาดของกลุ่มตัวอย่าง มีผลกระทบต่ออำนาจการตรวจสอบและการจำแนกผิดพลาด ดังนั้นควรจะใช้ขนาดกลุ่มตัวอย่างของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบเท่าใด จึงจะเหมาะสมเพื่อให้การตรวจสอบมีประสิทธิภาพมากที่สุด

3.ความยาวของแบบทดสอบจะส่งผลกระทบต่ออำนาจการตรวจสอบและการจำแนกผิดพลาดหรือไม่

จากประเด็นที่กล่าวมาข้างต้นนี้ ผู้วิจัยจึงสนใจศึกษาเปรียบเทียบการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอและแบบไม่สม่ำเสมอด้วยวิธี MH และวิธี LR โดยมีตัวแปรที่ศึกษาคือขนาดกลุ่มตัวอย่าง ซึ่งใช้ขนาดของกลุ่มตัวอย่างย่อยระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบเป็นสัดส่วนต่อกัน จำนวน 10 กลุ่ม และความยาวแบบทดสอบสามขนาด เพราะว่าตัวแปรดังกล่าวเป็นปัจจัยสำคัญที่มีผลกระทบต่ออำนาจการตรวจสอบและการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน ในการศึกษาครั้งนี้ผู้วิจัยศึกษาจากข้อมูลจริง เพื่อจะได้เป็นแนวทางในการพัฒนาวิธีการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน และนำมาใช้ในการตรวจสอบคุณภาพของแบบทดสอบให้มีความยุติธรรม และเป็นที่น่าเชื่อถือได้มากยิ่งขึ้นต่อไป

ความมุ่งหมายของการศึกษาค้นคว้า

1. เพื่อเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอ ระหว่างวิธี MH และวิธี LR เมื่อใช้แบบทดสอบความยาวเท่ากัน
2. เพื่อเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอ ระหว่างแบบทดสอบความยาวต่างกัน เมื่อใช้วิธีตรวจสอบวิธีเดียวกัน
3. เพื่อเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอ ระหว่างวิธี MH และวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาดเท่ากัน
4. เพื่อเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอ ระหว่างกลุ่มตัวอย่างขนาดต่างกัน เมื่อใช้วิธีตรวจสอบวิธีเดียวกัน
5. เพื่อเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สมำเสมอ ระหว่างวิธี MH และวิธี LR เมื่อใช้แบบทดสอบความยาวเท่ากัน
6. เพื่อเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สมำเสมอ ระหว่างแบบทดสอบความยาวต่างกัน เมื่อใช้วิธีตรวจสอบวิธีเดียวกัน
7. เพื่อเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สมำเสมอ ระหว่างวิธี MH และวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาดเท่ากัน
8. เพื่อเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สมำเสมอ ระหว่างกลุ่มตัวอย่างขนาดต่างกัน เมื่อใช้วิธีตรวจสอบวิธีเดียวกัน

ความสำคัญของการศึกษาค้นคว้า

ผลจากการศึกษาในครั้งนี้จะเป็นข้อมูลเบื้องต้นสำหรับใช้เป็นแนวทางในการเลือกวิธีการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอและแบบไม่สมำเสมอระหว่างวิธี MH และวิธี LR ภายใต้เงื่อนไขของปัจจัยด้านความยาวแบบทดสอบและขนาดกลุ่มตัวอย่าง ตลอดจนใช้เป็นแนวทางในการศึกษาค้นคว้าวิจัยเกี่ยวกับการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอและแบบไม่สมำเสมอ ต่อไป

ขอบเขตของการศึกษาค้นคว้า

1. ประชากรในการศึกษาในครั้งนี้เป็นนักเรียนชายและหญิงชั้นมัธยมศึกษาปีที่ 3 ภาคเรียนที่ 2 ปีการศึกษา 2542 ของโรงเรียนในกรุงเทพมหานคร สังกัดกรมสามัญศึกษา กลุ่มที่ 3 จำนวน 12 โรงเรียน จำนวน 164 ห้องเรียน จำนวนนักเรียน 7,058 คน จำแนกเป็นนักเรียนชาย 2,823 คน และนักเรียนหญิง 4,235 คน
2. กลุ่มตัวอย่างที่ใช้ในการศึกษาค้นคว้าครั้งนี้ เป็นนักเรียนชายและนักเรียนหญิง ชั้นมัธยมศึกษาปีที่ 3 ภาคเรียนที่ 2 ปีการศึกษา 2542 ของโรงเรียนในกรุงเทพมหานคร สังกัดกรมสามัญศึกษา กลุ่มที่ 3 ซึ่งได้มาโดยการสุ่มแบบแบ่งชั้น (Stratified Random Sampling) มีขนาดโรงเรียนเป็นชั้น (Strata) โรงเรียนเป็น

หน่วยการสุ่ม (Sampling Unit) ได้จำนวนโรงเรียน 8 โรงเรียน จำนวนห้องเรียน 71 ห้องเรียน นักเรียนทั้งหมด 3,242 คน จำแนกเป็นนักเรียนชาย 1,282 คน และนักเรียนหญิง 1,960 คน

3. ตัวแปรที่ใช้ในการศึกษาค้นคว้าครั้งนี้ ประกอบด้วย

3.1 ตัวแปรอิสระ ได้แก่

3.1.1 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมีสองวิธี คือ

(1) วิธี MH

(2) วิธี LR

3.1.2 ความยาวของแบบทดสอบมีสามขนาด คือ

(1) แบบทดสอบความยาว 20 ข้อ

(2) แบบทดสอบความยาว 40 ข้อ

(3) แบบทดสอบความยาว 60 ข้อ

3.1.3 ขนาดกลุ่มตัวอย่างมี 10 กลุ่ม

(1) กลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ เท่ากับ 300 คน ต่อ 100 คน

(2) กลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ เท่ากับ 300 คน ต่อ 200 คน

(3) กลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ เท่ากับ 300 คน ต่อ 300 คน

(4) กลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ เท่ากับ 500 คน ต่อ 100 คน

(5) กลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ เท่ากับ 500 คน ต่อ 200 คน

(6) กลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ เท่ากับ 500 คน ต่อ 300 คน

(7) กลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ เท่ากับ 1,000 คน ต่อ 100 คน

(8) กลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ เท่ากับ 1,000 คน ต่อ 200 คน

(9) กลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ เท่ากับ 1,000 คน ต่อ 300 คน

(10) กลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ เท่ากับ 1,000 คน ต่อ 1,000 คน

3.2 ตัวแปรตาม ได้แก่

3.2.1 อำนาจการตรวจสอบ

3.2.2 การจำแนกผิดพลาด

4. ตัวแปรควบคุม ได้แก่รูปแบบของการทำหน้าที่ต่างกันของข้อสอบมีสองรูปแบบ

4.1 แบบสม่ำเสมอ

4.2 แบบไม่สม่ำเสมอ

5. ตัวแปรที่ใช้จำแนกกลุ่มตัวอย่างในการวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันคือ เพศ ซึ่งในการวิจัยครั้งนี้ กลุ่มอ้างอิง ได้แก่เพศหญิง และกลุ่มเปรียบเทียบ ได้แก่ เพศชาย

6. แบบจำลองที่ใช้ในการศึกษาครั้งนี้ได้แก่แบบจำลองการตอบข้อสอบแบบโลจิสติกที่มีสองพารามิเตอร์

7. ข้อตกลงเบื้องต้น

เนื่องจากเครื่องมือที่ใช้ในการศึกษาครั้งนี้เป็นแบบทดสอบความสามารถทางสมองซึ่งผู้วิจัยสร้างขึ้นตามแนวโครงสร้างของแบบทดสอบความสามารถทางสมองของโอติส-เลนนอน ฟอร์ม เจ ระดับสูง แต่ผู้วิจัยได้พัฒนาและปรับปรุงแบบทดสอบให้สอดคล้องกับพื้นฐานทางภาษา สภาพสังคม วัฒนธรรมไทย และความมุ่งหมายในการวิจัยครั้งนี้ โดยตัดเนื้อหาด้านปริมาณซึ่งวัดความสามารถทางสมองด้านคณิตศาสตร์ออก เพื่อให้เหลือแบบทดสอบที่วัดความสามารถทางด้านภาษา เหตุผลด้านภาษาและเหตุผลด้านภาพเท่านั้น เนื่องจากตัวแปรที่จำแนกกลุ่มตัวอย่างที่ใช้ในการวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกัน คือตัวแปรเพศ ซึ่งผู้วิจัยได้กำหนดให้กลุ่มอ้างอิงได้แก่นักเรียนหญิง เนื่องจากนักเรียนหญิงมีความถนัดทางด้านภาษามากกว่านักเรียนชาย แล้วเพิ่มจำนวนข้อสอบในเนื้อหาด้านความเข้าใจด้านภาษา เหตุผลด้านภาษาและเหตุผลด้านภาพแทนข้อสอบที่ตัดออก ซึ่งการศึกษาในครั้งนี้ผู้วิจัยมุ่งที่จะใช้แบบทดสอบดังกล่าวมาพัฒนาและปรับปรุงเพื่อศึกษาเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาดในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธี MH และวิธี LR เท่านั้น มิได้มีเจตนาศึกษาคุณภาพของแบบทดสอบดังกล่าวแต่อย่างใด

นิยามศัพท์เฉพาะ

1. การทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning : DIF) หมายถึงโอกาสของการตอบข้อสอบได้ถูกไม่เท่ากัน เมื่อผู้สอบที่มีความสามารถในระดับเดียวกันแต่มาจากกลุ่มผู้สอบที่แตกต่างกัน ซึ่งในการวิจัยครั้งนี้การทำหน้าที่ต่างกันของข้อสอบหมายถึงนักเรียนหญิงและนักเรียนชายที่มีความสามารถในระดับเดียวกัน มีโอกาสในการตอบข้อสอบได้ถูกไม่เท่ากัน
2. ข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอ (Uniform DIF) หมายถึง โอกาสของการตอบข้อสอบได้ถูกต้อง สำหรับผู้สอบกลุ่มหนึ่งสูงกว่าผู้สอบอีกกลุ่มหนึ่งตลอดช่วงความสามารถ ซึ่งในทฤษฎี IRT สามารถพิจารณาได้จากโค้งลักษณะข้อสอบของผู้สอบสองกลุ่มขนานกัน ในการวิจัยครั้งนี้ ข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอหมายถึงโอกาสของการตอบข้อสอบได้ถูกต้องของนักเรียนหญิงสูงกว่านักเรียนชายคงเส้นคงวาตลอดช่วงความสามารถ
3. ข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ (Nonuniform DIF) หมายถึงความแตกต่างในโอกาสของการตอบข้อสอบได้ถูก สำหรับผู้สอบสองกลุ่มเปลี่ยนเครื่องหมายตลอดช่วงความสามารถ ซึ่งในทฤษฎี IRT สามารถพิจารณาได้จากโค้งลักษณะข้อสอบของผู้สอบสองกลุ่มไม่ขนานกัน ในการวิจัยครั้งนี้ ข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ หมายถึง โอกาสของการตอบข้อสอบได้ถูกของนักเรียนหญิงและนักเรียนชาย ไม่คงเส้นคงวาตลอดช่วงความสามารถ
4. ขนาดกลุ่มตัวอย่าง หมายถึงจำนวนผู้สอบที่ใช้ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในการศึกษาครั้งนี้ผู้วิจัยกำหนดขนาดกลุ่มตัวอย่าง 10 กลุ่ม คือจำนวนผู้สอบกลุ่มอ้างอิงต่อจำนวนผู้สอบกลุ่มเปรียบเทียบ เท่ากับ 300 : 100, 300 : 200, 300 : 300, 500 : 100, 500 : 200, 500 : 300, 1000 : 100, 1000 : 200, 1000 : 300 และ 1,000 : 1,000

5. ความยาวของแบบทดสอบ หมายถึง จำนวนข้อสอบของแบบทดสอบที่ใช้ในการวิเคราะห์ การทำหน้าที่ต่างกันของข้อสอบ ในการศึกษาครั้งนี้ผู้วิจัยกำหนดความยาวของแบบทดสอบสามขนาด คือ 20, 40 และ 60 ข้อ

6. กลุ่มอ้างอิง (Reference Group) หมายถึง กลุ่มผู้สอบที่คาดว่าจะได้เปรียบในการตอบข้อสอบ เมื่อข้อสอบทำหน้าที่ต่างกัน โดยมีโอกาสของการตอบข้อสอบได้ถูกมากกว่าผู้สอบกลุ่มเปรียบเทียบ ในการ วิจัยครั้งนี้ กลุ่มอ้างอิง ได้แก่กลุ่มผู้สอบเพศหญิง เนื่องจากมีความถนัดทางด้านภาษามากกว่าเพศชาย

7. กลุ่มเปรียบเทียบ (Focal Group) หมายถึงกลุ่มผู้สอบที่คาดว่าจะเสียเปรียบในการตอบข้อสอบ เมื่อข้อสอบทำหน้าที่ต่างกัน โดยมีโอกาสของการตอบข้อสอบได้ถูกน้อยกว่าผู้สอบกลุ่มอ้างอิง ในการวิจัย ครั้งนี้ กลุ่มเปรียบเทียบ ได้แก่กลุ่มผู้สอบเพศชาย เนื่องจากมีความถนัดทางด้านภาษาน้อยกว่าเพศหญิง

8. วิธีการวัดพื้นที่ของราชู (Raju's Area Measurement) หมายถึง วิธีตรวจสอบการทำหน้าที่ ต่างกันของข้อสอบที่พัฒนาโดยราชู (Raju, 1990) ซึ่งจะคำนวณดัชนีการทำหน้าที่ต่างกันของข้อสอบจาก พื้นที่ระหว่างโค้งลักษณะข้อสอบจากผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ โดยใช้การอินทิเกรตพื้นที่ในช่วง เปิด แล้วใช้สถิติ Z ทดสอบนัยสำคัญกับดัชนี DIF ที่ระดับ .05 ซึ่งผู้วิจัยใช้วิธีนี้เป็นเกณฑ์มาตรฐานในการ ตัดสินข้อสอบที่ทำหน้าที่ต่างกัน และในการศึกษาครั้งนี้ผู้วิจัยคำนวณพื้นที่ภายใต้โมเดลโลจิสติกแบบสอง พารามิเตอร์ กรณีข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอและแบบไม่สม่ำเสมอ โดยใช้โปรแกรม IRTDIF Version 1.0 ซึ่งพัฒนาโดยคิมและโคเฮน (Kim and Cohen, 1992)

9. วิธีแมนเทล-แฮนส์เซล (Mantel-Haenszel : MH) หมายถึง วิธีตรวจสอบการทำหน้าที่ต่างกัน ของข้อสอบที่พัฒนาโดยฮอลแลนด์และเทเยอร์ (Holland and Thayer, 1988) ซึ่งคำนวณดัชนีการทำหน้าที่ ต่างกันของข้อสอบจากสัดส่วนการตอบข้อสอบระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ แล้วใช้สถิติ χ^2_{MH} ทดสอบนัยสำคัญกับดัชนี DIF ที่ระดับ .05 ในการศึกษาครั้งนี้ผู้วิจัยใช้โปรแกรม MHDIF Version 1.0 ของฟิดัลโก (Fidalgo, 1995)

10. วิธีการถดถอยโลจิสติก (Logistic Regression : LR) หมายถึง วิธีตรวจสอบการทำหน้าที่ ต่างกันของข้อสอบ ในกลุ่มวิธี nonIRT ที่พัฒนาโดยสวามินาธานและโรเจอร์ส (Swaminathan and Rogers, 1990)ซึ่งวิเคราะห์ดัชนีการทำหน้าที่ต่างกันของข้อสอบจากความแตกต่างของฟังก์ชันการถดถอยโลจิสติก ระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบแล้วใช้สถิติ χ^2 ทดสอบนัยสำคัญกับดัชนี DIF ที่ระดับ .05 ในการศึกษาครั้งนี้ผู้วิจัยใช้โปรแกรม SPSS/PC⁺ Version 4.01

11. อำนาจการตรวจสอบ (Power) หมายถึง ร้อยละของจำนวนข้อสอบที่ระบุว่าทำหน้าที่ต่างกัน ได้ถูก ซึ่งคำนวณได้จากจำนวนข้อสอบที่ระบุว่าทำหน้าที่ต่างกันได้ถูกต้องจำนวนข้อสอบที่ทำหน้าที่ต่างกัน ทั้งหมดในแบบทดสอบ ซึ่งในการวิจัยครั้งนี้จำนวนข้อสอบที่ทำหน้าที่ต่างกันทั้งหมดในแบบทดสอบ ผู้วิจัยใช้ วิธีการวัดพื้นที่ของราชูเป็นเกณฑ์มาตรฐานตัดสิน

12. การจำแนกผิดพลาด (Type I Error Rates) หมายถึง ร้อยละของจำนวนข้อสอบที่ระบุว่า ทำหน้าที่ต่างกันผิดพลาด ทั้งที่ความเป็นจริงข้อสอบทำหน้าที่ไม่ต่างกัน ซึ่งคำนวณได้จากจำนวนข้อสอบที่ ระบุว่าทำหน้าที่ต่างกันผิดพลาดต่อจำนวนข้อสอบที่ทำหน้าที่ไม่ต่างกันแบบทดสอบ ซึ่งในการวิจัยครั้งนี้ จำนวนข้อสอบที่ทำหน้าที่ไม่ต่างกันแบบทดสอบ ผู้วิจัยใช้วิธีการวัดพื้นที่ของราชูเป็นเกณฑ์มาตรฐานตัดสิน

13. แบบทดสอบวัดความสามารถทางสมองหมายถึงแบบทดสอบที่ผู้วิจัยได้สร้างขึ้นโดยพัฒนาและปรับปรุงตามแนวโครงสร้างของแบบทดสอบทางสมองของโอตีส-เลนนอน ฟอรัม เจ ระดับสูง (Otis-Lennon Mental Ability Test Advanced Level Form J) จำนวน 1 ฉบับ เป็นแบบทดสอบวัดความสามารถทั่วไป ซึ่งแบ่งออกเป็นสามด้าน คือ ความเข้าใจด้านภาษา 30 ข้อ เหตุผลด้านภาษา 30 ข้อ และเหตุผลด้านภาพ 20 ข้อ รวม 80 ข้อ เป็นข้อสอบแบบปรนัยเลือกตอบห้าตัวเลือก ใช้เวลาในการทำ 40 นาที

14. ผู้เชี่ยวชาญ หมายถึงผู้ที่มีคุณวุฒิตั้งแต่ระดับปริญญาโทสาขาการวัดผลการศึกษา และมีประสบการณ์ด้านการสร้างข้อสอบวัดเชาวปัญญาหรือความถนัด จำนวนห้าท่าน

บทที่ 2

เอกสารและงานวิจัยที่เกี่ยวข้อง

ในการศึกษาครั้งนี้ ผู้วิจัยนำเสนอแนวคิดและทฤษฎีที่เกี่ยวข้อง โดยแบ่งออกเป็น 5 ตอน ดังนี้

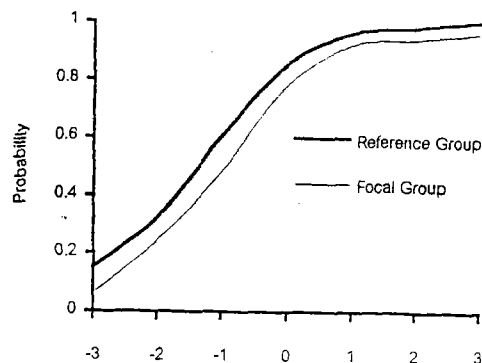
- ตอนที่ 1 ความหมายของการทำหน้าที่ต่างกันของข้อสอบ
- ตอนที่ 2 ประเภทของการทำหน้าที่ต่างกันของข้อสอบ
- ตอนที่ 3 หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ
- ตอนที่ 4 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ
- ตอนที่ 5 งานวิจัยที่เกี่ยวข้องกับการทำหน้าที่ต่างกันของข้อสอบ

ตอนที่ 1 ความหมายของการทำหน้าที่ต่างกันของข้อสอบ

การศึกษาเรื่องความยุติธรรมของข้อสอบหรือแบบทดสอบในกรณีที่จะทำให้ผู้สอบระหว่างกลุ่มย่อยมีความได้เปรียบหรือเสียเปรียบกันเดิมใช้คำว่า **“ความลำเอียงของข้อสอบ” (Item Bias)** หรือ **“ความลำเอียงของแบบทดสอบ” (Test Bias)** ซึ่งเป็นภาษาที่ใช้กันในทางสังคมและมีความหมายในเชิงลบ ส่วนการตัดสินว่าข้อสอบมีความลำเอียงหรือไม่นั้น มักจะพิจารณาอิทธิพลที่สังเกตได้ของกลุ่มผู้สอบย่อยที่นำมาศึกษาโดยไม่คำนึงถึงวิธีทางสถิติ จึงทำให้เกิดความคลุมเครือเกี่ยวกับเกณฑ์ที่ใช้ในการตัดสินความลำเอียง ต่อมาในระยะหลังนักวิจัยได้นำสารสนเทศทางสถิติมาใช้เป็นเกณฑ์ในการตัดสินความลำเอียงของข้อสอบ และได้เปลี่ยนไปใช้คำใหม่ว่า **“การทำหน้าที่ต่างกันของข้อสอบ” (Differential Item Functioning : DIF)** และ **“การทำหน้าที่ต่างกันของแบบทดสอบ” (Differential Test Functioning : DTF)** ซึ่งเป็นคำที่มีความเป็นกลางและเหมาะสมมากกว่า (วลิมาศ แซฮึง, 2543 : 26 - 27 ; อ้างอิงจาก Holland and Wainer, 1993) การทำหน้าที่ต่างกันของข้อสอบ หมายถึง โอกาสของการตอบข้อสอบได้ถูกต้องไม่เท่ากันเมื่อผู้สอบทั้งหมดที่มีความสามารถระดับเดียวกันแต่มาจากกลุ่มผู้สอบที่แตกต่างกัน (S.-H. Kim, H.-O. Kim and Cohen, 1994 ; citing Pine, 1977)

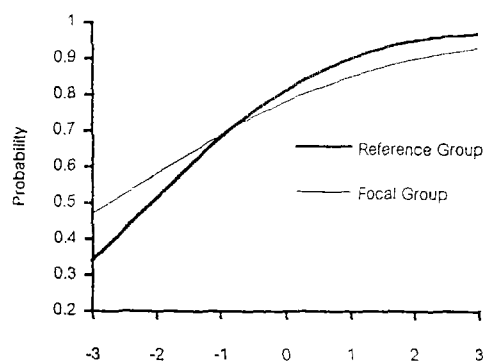
ตอนที่ 2 ประเภทของการทำหน้าที่ต่างกันของข้อสอบ

เมลเลนเบอร์ก (วลิมาศ แซฮึง, 2543 : 28 ; อ้างอิงจาก Mellenbergh, 1982) ได้จำแนกรูปแบบของการทำหน้าที่ต่างกันของข้อสอบออกเป็นสองประเภท คือ **การทำหน้าที่ต่างกันของข้อสอบแบบสม่ำเสมอ (Uniform DIF)** และ **การทำหน้าที่ต่างกันของข้อสอบแบบไม่สม่ำเสมอ (Nonuniform DIF)** การทำหน้าที่ต่างกันของข้อสอบประเภทแรกจะเกิดขึ้นเมื่อไม่มีปฏิสัมพันธ์ระหว่างระดับความสามารถของผู้สอบและการเป็นสมาชิกของกลุ่มผู้สอบ ส่วนการทำหน้าที่ต่างกันของข้อสอบประเภทหลังจะเกิดขึ้นเมื่อมีปฏิสัมพันธ์ระหว่างระดับความสามารถของผู้สอบและการเป็นสมาชิกของกลุ่มผู้สอบ ในทฤษฎีการตอบข้อสอบ (Item Response Theory : IRT) สามารถพิจารณา“ปฏิสัมพันธ์”ดังกล่าวได้จากความแตกต่างของค่าพารามิเตอร์อำนาจจำแนกของข้อสอบระหว่างผู้สอบกลุ่มย่อยสองกลุ่ม กล่าวคือถ้าข้อสอบทำหน้าที่ต่างกันแบบสม่ำเสมอแล้วโค้งลักษณะข้อสอบ (Item Characteristic Curves : ICCs) ระหว่างผู้สอบกลุ่มย่อยสองกลุ่มจะขนานกัน แต่ถ้าข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอแล้วโค้งลักษณะข้อสอบระหว่างผู้สอบกลุ่มย่อยสองกลุ่มจะไม่ขนานกัน ข้อสอบที่ทำหน้าที่ต่างกันทั้งสองประเภทแสดงดังภาพประกอบ 1 และ 2



Ability Scale

ภาพประกอบ 1 ข้อสอบทำหน้าที่ต่างกันแบบสม่ำเสมอ (Uniform DIF)



Ability Scale

ภาพประกอบ 2 ข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ (Nonuniform DIF)

ตอนที่ 3 หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ สามารถทำได้โดยการเปรียบเทียบผลการตอบข้อสอบระหว่างผู้สอบสองกลุ่ม ได้แก่กลุ่มเปรียบเทียบซึ่งเป็นกลุ่มที่สนใจศึกษาและเป็นกลุ่มที่คาดว่าจะเสียประโยชน์ในการตอบข้อสอบ คือมีโอกาสในการตอบข้อสอบถูกน้อยกว่าผู้สอบอีกกลุ่มหนึ่ง และกลุ่มอ้างอิงซึ่งเป็นกลุ่มที่คาดว่าจะได้ประโยชน์ในการตอบข้อสอบคือมีโอกาสในการตอบข้อสอบถูกมากกว่าผู้สอบอีกกลุ่มหนึ่ง การจำแนกผู้สอบเป็นกลุ่มอ้างอิงและกลุ่มเปรียบเทียบนั้นมีหลายลักษณะ เช่น จำแนกตามเพศ สีมัว เชื้อชาติ ภาษา สถาบันการศึกษา เป็นต้น

ตอนที่ 4 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ สามารถจำแนกตามการให้คะแนนของข้อมูลได้สองลักษณะ คือ การทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนแบบทวิพันธ์ (Dichotomous) และการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนแบบพหุพันธ์ (Polytomous DIF) และถ้าจะจำแนกตามทฤษฎีการ

วิเคราะห์ สามารถแบ่งออกเป็นสองกลุ่มใหญ่ ๆ คือ **กลุ่มวิธี nonIRT (Non Item Response Theory Methods)** และ **กลุ่มวิธี IRT (Item Response Theory Methods)** (Feinstein. 1995) กลุ่มวิธี nonIRT จะวิเคราะห์ดัชนีการทำหน้าที่ต่างกันของข้อสอบโดยใช้คะแนนที่สอบวัดได้โดยตรงตามทฤษฎีมาตรฐานเดิม ส่วนกลุ่มวิธี IRT จะวิเคราะห์โดยใช้คะแนนที่สอบวัดไม่ได้โดยตรงหรือตัวแปรแฝงตามทฤษฎีการตอบข้อสอบ หรือถ้าจะจำแนกวิธีการตรวจสอบตามข้อตกลงเบื้องต้นของโมเดล สามารถแบ่งออกได้เป็นสองรูปแบบ คือ **รูปแบบพาราเมตริก (Parametric Form)** และ **รูปแบบนัยพาราเมตริก (Nonparametric Form)** รูปแบบพาราเมตริก จะวิเคราะห์ดัชนีการทำหน้าที่ต่างกันของข้อสอบตามข้อตกลงเบื้องต้นของโมเดลที่อธิบายความสัมพันธ์ระหว่างคะแนนของข้อสอบและการจับคู่ตัวแปร ส่วนรูปแบบนัยพาราเมตริกจะไม่มีข้อตกลงดังกล่าวข้างต้น การจำแนกดังกล่าวสรุปรวมเป็นตารางได้ดังนี้ (Potenza and Dorans. 1995)

ตาราง 1 วิธีการตรวจสอบ DIF จำแนกตามการให้คะแนนของข้อมูล ทฤษฎีที่ใช้ในการวิเคราะห์ และรูปแบบของโมเดล

Type of Procedure	Parametric	Nonparametric
And Matching Variable		
Dichotomous DIF		
Observed Score	LRDIF	MH STND
Latent Variable	General IRTLR Limited Information IRTLR Loglinear IRTLR IRT-D ² Lord's χ^2	SIBTEST
Polytomous DIF		
Observed Score	Polytomous LRDIF	MNTLP-DIF Polytomous STND HW 1 HW 3 GMH
Latent Variable	Geneal IRTLR PCM	Polytomous SIBTEST GPCM

สำหรับรายละเอียดของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ จะกล่าวเฉพาะวิธีการวัดพื้นที่ของราชู(EUA) วิธีแมนเทล-ฮันส์เซล (MH) และวิธีการถดถอยโลจิสติก (LR) ดังนี้

วิธีการวัดพื้นที่ของราชู (Raju's area Measurement)

ราชู (วลิมาต แซฮ์อิง. 2543 : 58-63 ; อ้างอิงจาก Raju. 1990) ได้เสนอสูตรการคำนวณพื้นที่ชนิดมีเครื่องหมายในช่วงเปิด (Open-interval Signed Area or Exact Signed Area : ESA) และพื้นที่ชนิดไม่มีเครื่องหมายในช่วงเปิด (Open-interval Unsigned Area or Exact Unsigned Area : EUA) เพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนแบบทวิพันธ์ โมเดลโลจิสติกแบบหนึ่ง, สอง และสามพารามิเตอร์ สูตรทั่วไปที่ใช้ในการคำนวณพื้นที่ทั้งสองชนิดมีลักษณะดังนี้

$$ESA_{kl} = \int_{-\infty}^{+\infty} [P_R(\theta) - P_F(\theta)] d\theta$$

$$EUA_{kl} = \int_{-\infty}^{+\infty} |P_R(\theta) - P_F(\theta)| d\theta$$

โดยที่ k แทนโมเดลการตอบข้อสอบ

ถ้า $k = 1$ แสดงว่าเป็นโมเดลแบบหนึ่งพารามิเตอร์

ถ้า $k = 2$ แสดงว่าเป็นโมเดลแบบสองพารามิเตอร์

ถ้า $k = 3$ แสดงว่าเป็นโมเดลแบบสามพารามิเตอร์

และ i แทนพารามิเตอร์อำนาจจำแนกของข้อสอบข้อที่ i

ถ้า $i = 0$ แสดงว่าเป็นดัชนีชนิดคิดเครื่องหมาย โดยมี a_i แบบเท่ากันหรือไม่เท่ากัน

ถ้า $i = 1$ แสดงว่าเป็นดัชนีชนิดไม่คิดเครื่องหมาย โดยมี a_i แบบเท่ากัน

ถ้า $i = 2$ แสดงว่าเป็นดัชนีชนิดไม่คิดเครื่องหมาย โดยมี a_i แบบไม่เท่ากัน

ค่า IRFs ระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ในโมเดลโลจิสติกแบบสามพารามิเตอร์ สามารถคำนวณได้จากสูตร

$$P(\theta) = c + (1 - c)P^*(\theta)$$

โดยที่ $P^*(\theta) = \{1 + \exp[-Da(\theta - b)]\}^{-1}$

a , b และ c เป็นพารามิเตอร์ที่แสดงคุณลักษณะของข้อสอบ และ D เป็นค่าคงที่ โดยปกติกำหนดให้เท่ากับ 1.7 ถ้าพารามิเตอร์ c เท่ากับ 0 แล้ว IRF จะมีเพียงสองพารามิเตอร์ คือ a และ b และ ถ้าพารามิเตอร์ a เท่ากับ 1 แล้วจะเหลือเพียงพารามิเตอร์ b เท่านั้น การคำนวณพื้นที่ชนิด ESA และ EUA ในโมเดลโลจิสติกแบบหนึ่ง, สอง และสามพารามิเตอร์ มีรายละเอียดดังนี้

(1) โมเดลโลจิสติกแบบหนึ่งพารามิเตอร์ (1PLM)

การคำนวณพื้นที่ชนิด ESA และ EUA ระหว่าง IRFs ในโมเดลโลจิสติกแบบหนึ่งพารามิเตอร์ หรือโมเดล Rasch สามารถคำนวณได้จากสูตร

$$ESA_{10} = \hat{b}_F - \hat{b}_R$$

$$EUA_{11} = |\hat{b}_F - \hat{b}_R|$$

เนื่องจากค่าประมาณความยากของข้อสอบมีการแจกแจงลักษณะเชิงเส้นกำกับ(Asymptotically) ดังนั้นค่าเฉลี่ย และค่าความแปรปรวนของ ESA_{10} สามารถคำนวณได้จากสูตร

$$\mu(ESA_{10}) = b_F - b_R$$

$$\sigma^2(ESA_{10}) = Var(\hat{b}_F) + Var(\hat{b}_R)$$

$$\text{โดยที่ } Var(\hat{b}_i) = \left[\sum_{j=1}^N P_i(\theta_j) Q_i(\theta_j) \right]^{-1}$$

$$\text{และ } Q_i(\theta_j) = 1 - P_i(\theta_j)$$

เมื่อ $Var(\hat{b})$ แทน ความแปรปรวนของค่าประมาณความยากของข้อสอบข้อที่ i

N แทน จำนวนผู้สอบในแต่ละกลุ่ม

$P_i(\theta_j)$ แทน โอกาสของผู้สอบคนที่ j ที่มีระดับความสามารถ θ จะตอบข้อสอบข้อที่ i ถูก

$Q_i(\theta_j)$ แทน โอกาสของผู้สอบคนที่ j ที่มีระดับความสามารถ θ จะตอบข้อสอบข้อที่ i ผิด

ถ้ากำหนดให้ $X = \hat{b}_F - \hat{b}_R$ และ X มีข้อตกลงการแจกแจงแบบปกติ ดังนั้นค่าเฉลี่ยและค่าความแปรปรวนของ EUA_{11} สามารถคำนวณได้จากสูตร

$$\mu(EUA_{11}) = \mu(ESA_{10})[1 - 2\Phi(z_0)] + (2/\pi)^{1/2} \sigma(ESA_{10}) \exp(-z_0^2/2)$$

$$\sigma^2(EUA_{11}) = \sigma^2(ESA_{10}) + \mu^2(ESA_{10}) - \mu^2(EUA_{11})$$

$$\text{โดยที่ } \Phi(z_0) = \int_{-\infty}^{z_0} g(z) dz$$

$$z_0 = [0 - \mu(ESA_{10})]/\sigma(ESA_{10})$$

เมื่อ $g(z)$ แทน ฟังก์ชันความหนาแน่นของความน่าจะเป็นแบบปกติ

(2) โมเดลโลจิสติกแบบสองพารามิเตอร์ (2PLM)

การคำนวณพื้นที่ชนิด ESA และ EUA ระหว่าง IRFs ในโมเดลโลจิสติกแบบสองพารามิเตอร์ สามารถคำนวณได้จากสูตร

$$\begin{aligned}
 ESA_{20} &= \hat{b}_F - \hat{b}_R \\
 EUA_{21} &= |\hat{b}_F - \hat{b}_R| && \text{เมื่อ } \hat{a}_R = \hat{a}_F \\
 EUA_{22} &= \left| \frac{2(\hat{a}_F - \hat{a}_R)}{D\hat{a}_R\hat{a}_F} \ln \left\{ 1 + \exp \left[\frac{D\hat{a}_R\hat{a}_F(\hat{b}_F - \hat{b}_R)}{\hat{a}_F - \hat{a}_R} \right] \right\} - (\hat{b}_F - \hat{b}_R) \right| && \text{เมื่อ } \hat{a}_R \neq \hat{a}_F
 \end{aligned}$$

เนื่องจากค่าประมาณความยากของข้อสอบมีการแจกแจงลักษณะเชิงเส้นกำกับ(Asymptotically) ดังนั้นค่าเฉลี่ยและความแปรปรวนของ ESA_{20} สามารถคำนวณได้จากสูตร

$$\begin{aligned}
 \mu(ESA_{20}) &= b_F - b_R \\
 \sigma^2(ESA_{20}) &= Var(\hat{b}_F) + Var(\hat{b}_R)
 \end{aligned}$$

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

และ $I_{a_i} = D^2 \sum_{j=1}^N (\theta_j - b_i)^2 P_i(\theta_j) Q_i(\theta_j)$

$$I_{b_i} = D^2 a_i^2 \sum_{j=1}^N P_i(\theta_j) Q_i(\theta_j)$$

$$I_{a_i b_i} = D^2 a_i \sum_{j=1}^N (\theta_j - b_i)^2 P_i(\theta_j) Q_i(\theta_j)$$

เมื่อ I_{a_i} แทน ฟังก์ชันสารสนเทศของ a_i
 I_{b_i} แทน ฟังก์ชันสารสนเทศของ b_i
 $I_{a_i b_i}$ แทน ฟังก์ชันสารสนเทศของ $a_i b_i$

ถ้ากำหนดให้ $X = \hat{b}_F - \hat{b}_R$ และ X มีข้อตกลงการแจกแจงแบบปกติ ดังนั้นค่าเฉลี่ยและค่าความแปรปรวนของ EUA_{21} เมื่อ $\hat{a}_R = \hat{a}_F$ สามารถคำนวณได้จากสูตร

$$\mu(EUA_{21}) = \mu(ESA_{20})[1 - 2\Phi(z_0)] + (2/\pi)^{1/2} \sigma(ESA_{20}) \exp(-z_0^2/2)$$

$$\sigma^2(EUA_{21}) = \sigma^2(ESA_{20}) + \mu^2(ESA_{20}) - \mu^2(EUA_{21})$$

โดยที่ $z_0 = [0 - \mu(ESA_{20})] / \sigma(ESA_{20})$

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

และ X มีข้อตกลงการแจกแจงแบบปกติ ดังนั้นค่าเฉลี่ยและค่าความแปรปรวนของ EUA_{22} เมื่อ $\hat{a}_R \neq \hat{a}_F$ สามารถคำนวณได้จากสูตร

$$\mu(EUA_{22}) = \mu(H)[1 - 2\Phi(z_0)] + (2/\pi)^{1/2} \sigma(H) \exp(-z_0^2/2)$$

$$\sigma^2(EUA_{22}) = \sigma^2(H) + \mu^2(H) - \mu^2(H)$$

โดยที่ $z_0 = [0 - \mu(H)] / \sigma(H)$

$$\mu(H) = \frac{2(a_F - a_R)}{Da_R a_F} \ln \left\{ 1 + \exp \left[\frac{Da_R a_F (b_F - b_R)}{a_F - a_R} \right] - (b_F - b_R) \right\}$$

$$\sigma^2(H) = Var(H)$$

$$\begin{aligned} &= B_R^2 Var(\hat{b}_R) + B_F^2 Var(\hat{b}_F) + A_R^2 Var(\hat{a}_R) + A_F^2 Var(\hat{a}_F) \\ &\quad + 2B_R A_R Cov(\hat{b}_R, \hat{a}_R) + 2B_F A_F Cov(\hat{b}_F, \hat{a}_F) \end{aligned}$$

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

$$B_F = -B_R$$

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

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

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

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

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

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

เมื่อ	$Var(\hat{b})$	แทน ความแปรปรวนของ $(\hat{b})$
	$Var(\hat{a})$	แทน ความแปรปรวนของ $(\hat{a})$
	$Cor(\hat{a}, \hat{b})$	แทน ความแปรปรวนร่วมระหว่าง $(\hat{a})$ และ $(\hat{b})$

(3) โมเดลโลจิสติกแบบสามพารามิเตอร์ (3PLM)

ราชู (Kim and Cohen. 1991 ; citing Raju. 1988) ได้เสนอให้กำหนดพารามิเตอร์ c เท่ากัน ($c_R = c_F = c$) ในการคำนวณพื้นที่ชนิด ESA และ EUA ระหว่าง IRFs โมเดลโลจิสติกแบบสามพารามิเตอร์ ทั้งนี้เพราะว่าถ้าพารามิเตอร์ c ระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบมีค่าไม่เท่ากันแล้ว ค่าที่คำนวณได้จะไม่มีที่สิ้นสุด (Infinite) ดังนั้นค่า ESA และ EUA โมเดลโลจิสติกแบบสามพารามิเตอร์เมื่อกำหนด c คงที่สามารถคำนวณได้จากสูตร

$$\begin{aligned} ESA_{30} &= (1-c)ESA_{20} \\ EUA_{31} &= (1-c)EUA_{21} && \text{เมื่อ } \hat{a}_R = \hat{a}_F \\ EUA_{32} &= (1-c)EUA_{22} && \text{เมื่อ } \hat{a}_R \neq \hat{a}_F \end{aligned}$$

ค่าเฉลี่ยและค่าความแปรปรวนของ ESA_{30} EUA_{31} และ EUA_{32} เมื่อกำหนดค่า c คงที่ สามารถคำนวณโดยใช้สูตรจากโมเดลแบบ 2 พารามิเตอร์ ดังนี้

$$\begin{aligned} \mu(ESA_{30}) &= (1-c)\mu(ESA_{20}) \\ \sigma^2(ESA_{30}) &= (1-c)^2 \sigma^2(ESA_{20}) \\ \mu(EUA_{31}) &= (1-c)\mu(EUA_{21}) \\ \sigma^2(EUA_{31}) &= (1-c)^2 \sigma^2(EUA_{21}) \\ \mu(EUA_{32}) &= (1-c)\mu(EUA_{22}) \\ \sigma^2(EUA_{32}) &= (1-c)^2 \sigma^2(EUA_{22}) \end{aligned}$$

สำหรับการคำนวณค่าฟังก์ชันสารสนเทศ (I) ใช้สูตรดังนี้

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

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

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

ในการตัดสินใจว่าข้อสอบที่นำมาตรวจสอบทำหน้าที่ต่างกันหรือไม่ ราชู (Raju, 1990) ได้เสนอให้นำดัชนีการวัดพื้นที่มาทดสอบนัยสำคัญโดยใช้สถิติ Z ภายใต้สมมติฐานการแจกแจงแบบปกติ ในการทดสอบดังกล่าวแบ่งออกเป็นสองชนิด ดังนี้

ชนิดที่ 1 การทดสอบนัยสำคัญของ ESA

นำดัชนี ESA ของข้อสอบข้อที่ i ไปทดสอบความแตกต่างกับ 0 โดยใช้สถิติทดสอบ Z ดังนี้

$$Z_i(ESA) = \frac{\hat{b}_{iF} - \hat{b}_{iR}}{\sqrt{\text{Var}(\hat{b}_{iF}) + \text{Var}(\hat{b}_{iR})}}$$

ชนิดที่ 2 การทดสอบนัยสำคัญของ EUA

ในการทดสอบนัยสำคัญของดัชนี EUA สามารถแบ่งออกเป็น 2 กรณีคือ **กรณีที่ 1** เมื่อ $\hat{a}_{iR} = \hat{a}_{iF}$ จะทดสอบเหมือนกับดัชนี ESA สำหรับ **กรณีที่ 2** $\hat{a}_{iR} \neq \hat{a}_{iF}$ จะนำดัชนี EUA ของข้อสอบข้อที่ i ไปทดสอบความแตกต่างกับ 0 โดยใช้สถิติทดสอบ Z ดังนี้

$$Z_i(H) = \frac{H_i}{\sqrt{\text{Var}(H_i)}}$$

วิธีแมนเทล-แฮนส์เซล (Mantel-Haenszel : MH)

วิธี MH พัฒนาโดยแมนเทลและแฮนส์เซล (วลิมาต แซ็ง. 2543 : 38 – 44 ; อ้างอิงจาก Camilli and Shepard, 1994) ซึ่งประกอบด้วยอัตราส่วนแต้มต่อร่วม (Common odds ratio) และสถิติไค-กำลังสอง (Chi-square statistic) ต่อมาฮอลแลนด์ (Holland and Thayer, 1988 ; citing Holland, 1985) ได้นำวิธี MH ไปประยุกต์ใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในสำนักงานบริการทดสอบทางการศึกษา (Education Testing Service : ETS) ของประเทศสหรัฐอเมริกา จึงทำให้วิธี MH เป็นที่ยอมรับจากนักวิจัยอย่างกว้างขวาง วิธี MH เป็นวิธีตารางการณั้จรแบบไม่คำนวณทวนซ้ำ (Noniterative Contingency Table)

พัฒนามาจากวิธีไค-กำลังสองแบบประเพณีนิยม (Traditional) สามารถนำไปใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ง่ายมีขั้นตอนการคำนวณที่ไม่สลับซับซ้อนมีการทดสอบทางสถิติแบบนั้พาราเมตริก (Nonparametric) ซึ่งไม่จำเป็นต้องใช้โมเดลประมาณค่า

หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยวิธี MH จะเปรียบเทียบผลการตอบข้อสอบระหว่างผู้สอบสองกลุ่ม กลุ่มหนึ่งเรียกว่า“**กลุ่มเปรียบเทียบ**”ซึ่งเป็นกลุ่มผู้สอบที่ผู้วิจัยสนใจศึกษา ถ้าข้อสอบที่นำมาศึกษาทำหน้าที่ต่างกัน คาดว่าผู้สอบกลุ่มนี้จะเสียเปรียบในการตอบข้อสอบ ส่วนผู้สอบอีกกลุ่มหนึ่งเรียกว่า“**กลุ่มอ้างอิง**”ซึ่งเป็นกลุ่มผู้สอบที่ใช้เป็นมาตรฐานในการเปรียบเทียบผลการตอบข้อสอบกับผู้สอบกลุ่มแรก ถ้าข้อสอบทำหน้าที่ต่างกันคาดว่าผู้สอบกลุ่มหลังนี้จะได้เปรียบในการตอบข้อสอบ ส่วนการเปรียบเทียบผลการตอบข้อสอบจะเปรียบเทียบทุกระดับความสามารถของผู้สอบกลุ่มย่อยสองกลุ่มที่มีระดับความสามารถเท่ากัน ในทางปฏิบัติมักใช้คะแนนรวมของแบบทดสอบเป็นเกณฑ์การจับคู่กลุ่มผู้สอบ เกณฑ์ในลักษณะดังกล่าวมีจุดอ่อนกล่าวคือถ้าในแบบทดสอบมีข้อสอบที่ทำหน้าที่ต่างกันการใช้คะแนนรวมของแบบทดสอบที่ใช้เป็นเกณฑ์ในการจับคู่กลุ่มผู้สอบจะรวมคะแนนจากข้อสอบที่ทำหน้าที่ต่างกันดังกล่าวด้วยซึ่งจะมีผลทำให้ได้เกณฑ์ที่มีความบิดเบือน ผลการตรวจสอบจึงไม่ถูกต้อง ในการแก้ไขจุดอ่อนดังกล่าวฮอลแลนด์และ เทเยอร์ (Holland and Thayer, 1988) ได้เสนอให้ใช้วิธีการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบแบบสองขั้นตอน เพื่อจะทำให้เกณฑ์ในการจับคู่กลุ่มผู้สอบมีความบริสุทธิ์ (Purification of Matching Criterion) สำหรับขั้นตอนการวิเคราะห์มีดังนี้

ขั้นตอนแรก ใช้คะแนนรวมของแบบทดสอบทั้งฉบับเป็นเกณฑ์การจับคู่กลุ่มผู้สอบย่อยสองกลุ่มแล้ววิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ เมื่อพบว่าข้อสอบข้อใดที่ทำหน้าที่ต่างกันให้นำคะแนนของข้อสอบข้อนั้น ออกจากคะแนนรวมของผู้สอบแต่ละคน

ขั้นตอนที่สอง ใช้คะแนนรวมของแบบทดสอบที่นำเอาข้อสอบที่ทำหน้าที่ต่างกัน ซึ่งตรวจพบในขั้นตอนแรกออกไปแล้ว เพื่อวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบซ้ำอีกครั้งหนึ่ง

เมื่อจับคู่กลุ่มผู้สอบแล้วจะนำข้อมูลผลการตอบข้อสอบระหว่างผู้สอบย่อยสองกลุ่มมาจัดลงในตารางการณ้จร (Contingency Table) แบบ 2×2 (กลุ่มผู้สอบ 2 กลุ่ม $\times$ ผลการตอบ 2 แบบ) โดยที่ตารางการณ้จร 1 ตารางแทนคะแนนรวม 1 ระดับ ดังนั้นถ้ามีคะแนนรวม K ระดับ จะต้องสร้างตารางการณ้จรแบบ 2×2 ทั้งหมด K ตาราง สำหรับตารางการณ้จรแบบ 2×2 ของข้อสอบแต่ละข้อที่มีคะแนนรวมระดับ j แสดงดังตาราง 2

ตาราง 2 ผลการตอบข้อสอบระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ณ ระดับคะแนน j

กลุ่ม	คะแนนของผลการตอบข้อสอบที่ศึกษา		รวม
	ตอบถูกได้ 1 คะแนน	ตอบผิดได้ 0 คะแนน	
R	A_j	B_j	NR_j
F	C_j	D_j	NF_j
รวม	m_{1j}	m_{0j}	T_j

เมื่อ	A_j	แทน จำนวนผู้สอบกลุ่มอ้างอิงที่ตอบข้อสอบถูก ณ ระดับคะแนน j
	B_j	แทน จำนวนผู้สอบกลุ่มอ้างอิงที่ตอบข้อสอบผิด ณ ระดับคะแนน j
	C_j	แทน จำนวนผู้สอบกลุ่มเปรียบเทียบที่ตอบข้อสอบถูก ณ ระดับคะแนน j
	D_j	แทน จำนวนผู้สอบกลุ่มเปรียบเทียบที่ตอบข้อสอบผิด ณ ระดับคะแนน j
	m_{1j}	แทน จำนวนผู้สอบที่ตอบถูก ณ ระดับคะแนน j
	m_{0j}	แทน จำนวนผู้สอบที่ตอบผิด ณ ระดับคะแนน j
	nR_j	แทน จำนวนผู้สอบกลุ่มอ้างอิง ณ ระดับคะแนน j
	nF_j	แทน จำนวนผู้สอบกลุ่มเปรียบเทียบ ณ ระดับคะแนน j
	T_j	แทน จำนวนผู้สอบทั้งหมด ณ ระดับคะแนน j

ต่อจากนั้นจึงนำผลการตอบข้อสอบจากตารางที่ 2 มาคำนวณสัดส่วนของผลการตอบข้อสอบถูกและผิด ดังตาราง 3

ตาราง 3 สัดส่วนของผลการตอบข้อสอบระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ณ ระดับคะแนน j

กลุ่ม	คะแนนของผลการตอบข้อสอบที่ศึกษา		รวม
	ตอบถูกได้ 1 คะแนน	ตอบผิดได้ 0 คะแนน	
R	pR_j	qR_j	1
F	pF_j	qF_j	1

เมื่อ	pR_j	แทน สัดส่วนของผู้สอบกลุ่มอ้างอิงที่ตอบข้อสอบถูก ณ ระดับคะแนน j
	qR_j	แทน สัดส่วนของผู้สอบกลุ่มอ้างอิงที่ตอบข้อสอบผิด ณ ระดับคะแนน j
	pF_j	แทน สัดส่วนของผู้สอบกลุ่มเปรียบเทียบที่ตอบข้อสอบถูก ณ ระดับคะแนน j
	qF_j	แทน สัดส่วนของผู้สอบกลุ่มเปรียบเทียบที่ตอบข้อสอบผิด ณ ระดับคะแนน j

สำหรับการทดสอบสมมติฐานของการทำหน้าที่ต่างกันของข้อสอบ จะกำหนดสมมติฐานกลาง (H_0) และสมมติฐานเลือก (H_1) ดังนี้

$$H_0: \frac{pR_j}{qR_j} = \frac{pF_j}{qF_j} \quad ; \quad j = 1, 2, 3, \dots, K$$

$$H_1: \frac{pR_j}{qR_j} = \alpha \frac{pF_j}{qF_j} \quad ; \quad j = 1, 2, 3, \dots, K \text{ เมื่อ } \alpha \neq 1$$

สมมติฐานกลางเป็นสมมติฐานที่เป็นอิสระอย่างมีเงื่อนไขของสมาชิกในกลุ่มผู้สอบและคะแนนจากข้อสอบที่ศึกษา ดังนั้นคะแนนที่ได้จากตารางที่ 3 และภายใต้สมมติฐานกลางสามารถสรุปเป็นค่าคาดหวัง (Expected Values) ในแต่ละเซลล์ได้ดังนี้

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

$$E(B_j) = \frac{nR_j m_{0j}}{T_j}$$

$$E(C_j) = \frac{nF_j m_{1j}}{T_j}$$

$$E(D_j) = \frac{nF_j m_{0j}}{T_j}$$

พารามิเตอร์ α ภายใต้สมมติฐานเลือกเรียกว่าอัตราส่วนแอดมิตต่อร่วม (Common Odds Ratio) ซึ่งคำนวณได้จาก

$$\alpha = \frac{\frac{pR_j}{qR_j}}{\frac{pF_j}{qF_j}} = \frac{pR_j qF_j}{qR_j pF_j}$$

ถ้า $\alpha = 1$ แสดงว่าโอกาสของการตอบข้อสอบถูกต้องระหว่างผู้สอบทั้งสองกลุ่มมีค่าเท่ากัน

ถ้า $\alpha > 1$ แสดงว่าผู้สอบกลุ่มอ้างอิงมีโอกาสของการตอบข้อสอบถูกมากกว่าผู้สอบกลุ่มเปรียบเทียบ และ

ถ้า $\alpha < 1$ แสดงว่าผู้สอบกลุ่มเปรียบเทียบมีโอกาสของการตอบข้อสอบถูกมากกว่าผู้สอบกลุ่มอ้างอิง แมนเทิลและแฮนส์เซล (Holland and Thayer, 1988 ; citing Mantel and Haenszel, 1959) ได้ประมาณค่า α จากตารางไขว้แบบ 2×2 ดังนี้

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

$\hat{\alpha}_{MH}$ เป็นค่าประมาณขนาดอิทธิพลของการทำหน้าที่ต่างกันของข้อสอบ (DIF Effect Size) ซึ่งมีค่าอยู่ระหว่าง 0 ถึง ∞ ฮอลแลนด์และเทเยอร์ (Holland and Thayer. 1988 ; citing Holland and Thayer. 1985) ได้เสนอให้แปลงค่า $\hat{\alpha}_{MH}$ เป็นสเกลเดลตา (Delta Scale : MH_{D-DIF}) ซึ่งเป็นค่าที่ใช้ในสำนักงานบริการทดสอบทางการศึกษาของสหรัฐอเมริกา ดังนี้

$$MH_{D-DIF} = -2.35 \ln(\hat{\alpha}_{MH})$$

ค่า MH_{D-DIF} สามารถนำไปใช้วัดค่าความยากง่ายของข้อสอบ กล่าวคือ ถ้า MH_{D-DIF} มีค่าเป็นศูนย์ แสดงว่าข้อสอบยากง่ายเท่ากันระหว่างผู้สอบทั้งสองกลุ่ม ถ้า MH_{D-DIF} มีค่าเป็นลบ แสดงว่าข้อสอบยากสำหรับผู้สอบกลุ่มเปรียบเทียบมากกว่ากลุ่มอ้างอิง และถ้า MH_{D-DIF} มีค่าเป็นบวก แสดงว่าข้อสอบยากสำหรับผู้สอบกลุ่มกลุ่มอ้างอิงมากกว่ากลุ่มเปรียบเทียบ สำหรับการประมาณค่าความคลาดเคลื่อนมาตรฐาน (Standard Error : SE) ของ MH_{D-DIF} สามารถคำนวณได้จากสูตรดังนี้

$$SE(MH_{D-DIF}) = 2.35 \sqrt{Var[\ln(\hat{\alpha}_{MH})]}$$

$$\text{โดยที่ } Var[\ln(\hat{\alpha}_{MH})] = \frac{\sum_{j=1}^K U_j V_j / T_j^2}{2 \left[\sum_{j=1}^K A_j D_j / T_j \right]^2}$$

$$\text{ขณะที่ } U_i = A_j D_j + \hat{\alpha}_{MH} (B_j C_j)$$

$$\text{และ } V_i = (A_j + D_j) + \hat{\alpha}_{MH} (B_j + C_j)$$

สำหรับการทดสอบนัยสำคัญของสมมติฐาน จะนำค่า $\hat{\alpha}_{MH}$ และค่า MH_{D-DIF} ไปทดสอบกับสถิติแมนเทิล-แฮนส์เซลไค-กำลังสอง (χ^2_{MH}) ที่ระดับชั้นความเป็นอิสระเท่ากับ 1 (df = 1) โดยนำค่า $\hat{\alpha}_{MH}$ ไปเปรียบเทียบกับ 1 ส่วนค่า MH_{D-DIF} จะเปรียบเทียบกับ 0 สถิติทดสอบ χ^2_{MH} มีสูตรในการคำนวณดังนี้

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

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

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

เมื่อ $E(A_j)$ แทน ค่าคาดหวังของจำนวนผู้สอบกลุ่มอ้างอิงที่ตอบข้อสอบถูก ณ ระดับคะแนน j

$Var(A_j)$ แทน ค่าความแปรปรวนของจำนวนผู้สอบกลุ่มอ้างอิงที่ตอบข้อสอบถูก ณ ระดับคะแนน j

ถ้าค่า $\hat{\alpha}_{MH}$ ไม่แตกต่างจาก 1 อย่างมีนัยสำคัญ แสดงว่าข้อสอบทำหน้าที่ไม่ต่างกัน (no DIF) แต่ถ้าค่า $\hat{\alpha}_{MH}$ แตกต่างจาก 1 อย่างมีนัยสำคัญ แสดงว่าข้อสอบทำหน้าที่ต่างกัน (DIF) โดยจะเข้าข้างกลุ่มเปรียบเทียบกับเมื่อ $\hat{\alpha}_{MH}$ มีค่าเป็นบวก และจะเข้าข้างกลุ่มอ้างอิงเมื่อ $\hat{\alpha}_{MH}$ มีค่าเป็นลบ ส่วนการพิจารณาค่า MH_{D-DIF} ว่าแตกต่างจาก 0 อย่างมีนัยสำคัญทางสถิติหรือไม่ ก็มีลักษณะทำนองเดียวกัน

วิธีการถดถอยโลจิสติก (logistic regression : LR)

สวามินาธานและโรเจอร์ส (วลีมาต แซ็ง. 2543 : 45 – 48 ; อ้างอิงจาก Swaminathan and Rogers. 1990) ได้เสนอวิธี LR เพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนแบบทวิพันธ์ (Dichotomous) วิธีดังกล่าวมีแนวคิดมาจากวิธีตารางการถ่วง โดยดัดแปลงมาจากวิธีล็อกลิเนียร์ของเมลเลนเบิร์ก (Mellenbergh. 1982) และวิธีแมนเทิล-แฮนส์เซลของฮอลแลนด์และเทเยอร์ (Holland and Thayer. 1988) จุดเด่นของวิธี LR ก็คือ สามารถคำนวณได้ง่ายโดยใช้โมเดลการถดถอยโลจิสติก ซึ่งโมเดลนี้ใช้ตัวแปรความสามารถแบบต่อเนื่อง ซึ่งมีความไวในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้งแบบสม่ำเสมอ (Uniform DIF) และแบบไม่สม่ำเสมอ (Nonuniform DIF) นอกจากนี้ยังเป็นวิธีที่มีความยืดหยุ่น ทั้งนี้เนื่องจากสามารถนำไปประยุกต์กับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีผู้สอบหลายกลุ่มและการให้คะแนนข้อสอบแบบพหุพันธ์ (polytomous) (Miller and Spray. 1993)

ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธี LR จะใช้สมการมาตรฐานของโมเดลการถดถอยโลจิสติกคำนวณผลการตอบข้อสอบถูก ดังนี้ (Swaminathan and Rogers. 1990 ; citing Bock. 1975)

$$P(U_{pj} = 1 | \theta_{pj}) = \frac{\exp(\beta_{0j} + \beta_{1j} \theta_{pj})}{1 + \exp(\beta_{0j} + \beta_{1j} \theta_{pj})} ; p = 1, \dots, n_j, j = 1, 2.$$

เมื่อ $P(U_{pj} = 1 | \theta_{pj})$ แทน ผลการตอบข้อสอบของผู้สอบคนที่ p ในกลุ่ม j

β_{0j} แทน พารามิเตอร์จุดตัด

β_{1j} แทน พารามิเตอร์ความชันสำหรับกลุ่ม j

θ_{pj} แทน ค่าความสามารถที่สังเกตได้ของผู้สอบคนที่ p ในกลุ่ม j

จากโมเดลดังกล่าว ถ้า $\beta_{01} = \beta_{02}$ และ $\beta_{11} = \beta_{12}$ แล้วฟังก์ชันการถดถอยโลจิสติกของผู้สอบสองกลุ่มเหมือนกัน แสดงว่าข้อสอบทำหน้าที่ไม่ต่างกัน (no DIF) ถ้า $\beta_{11} = \beta_{12}$ และ $\beta_{01} \neq \beta_{02}$ แล้วฟังก์ชันการถดถอยโลจิสติกของผู้สอบสองกลุ่มขนานกันแต่ไม่ทับกัน แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบสม่ำเสมอ (uniform DIF) และถ้า $\beta_{01} = \beta_{02}$ แต่ $\beta_{11} \neq \beta_{12}$ แล้วฟังก์ชันการถดถอยโลจิสติกของผู้สอบไม่ขนานกัน แสดงว่า ข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ (nonuniform DIF) จากโมเดลการถดถอยโลจิสติกดังกล่าว สามารถเปลี่ยนเป็นโมเดลในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบแบบสม่ำเสมอและแบบไม่สม่ำเสมอดังนี้

$$P(U_{pj} = 1) = \frac{\exp^{Z_{pj}}}{1 + \exp^{Z_{pj}}}$$

โดยที่ $Z_{pj} = \tau_0 + \tau_1\theta_{pj} + \tau_2g_j + \tau_3(\theta_{pj}g_j)$

เมื่อ $P(U_{pj} = 1)$ แทน โอกาสในการตอบข้อสอบถูกของผู้สอบคนที่ p ในกลุ่ม j

θ_{pj} แทน ระดับความสามารถของผู้สอบ (คะแนนรวม) คนที่ p ในกลุ่ม j

τ_0 แทน พารามิเตอร์จุดตัด

τ_1 แทน สัมประสิทธิ์ของความสามารถ

τ_2 แทน ความแตกต่างในกลุ่มของผลการตอบข้อสอบ

τ_3 แทน ปฏิสัมพันธ์ระหว่างกลุ่มและระดับความสามารถ

g_j แทน สมาชิกผู้สอบในกลุ่ม j (โดยกำหนดให้ $g_p = 1$ ถ้าผู้สอบอยู่ในกลุ่ม 1 และ $g_p = 0$ ถ้าผู้สอบอยู่ในกลุ่ม 2)

โมเดลการถดถอยโลจิสติกข้างต้นสามารถเปลี่ยนเป็นโมเดลเชิงเส้นในเมทริกโลจิสติก(logit metric) ซึ่งจะอยู่ในรูป $\log$ ของอัตราส่วนของโอกาสในการตอบข้อสอบถูกต้องต่อโอกาสในการตอบข้อสอบผิด ดังนี้

$$\log \left[\frac{P}{1-P} \right] = \tau_0 + \tau_1\theta_{pj} + \tau_2g_j + \tau_3(\theta_{pj}g_j)$$

จากโมเดลดังกล่าว θ_{pj} และ g เป็นตัวแปรอิสระ ส่วนพารามิเตอร์ τ_2 และ τ_3 สอดคล้องกับพารามิเตอร์ในสมการมาตรฐานของโมเดลการถดถอยโลจิสติก ดังนี้

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

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

ในการตัดสินใจว่าข้อสอบทำหน้าที่ต่างกันแบบสม่ำเสมอหรือไม่สม่ำเสมอสามารถพิจารณาได้จากพารามิเตอร์ τ_2 และ τ_3 กล่าวคือ ถ้า $\tau_2 \neq 0$ และ $\tau_3 = 0$ แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบสม่ำเสมอ และถ้า $\tau_3 \neq 0$ ซึ่ง $\tau_2 = 0$ หรือไม่ก็ได้แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ สำหรับค่าพารามิเตอร์ของข้อสอบแต่ละข้อในโมเดล Z_{pj} สามารถประมาณค่าโดยใช้วิธีความน่าจะเป็นสูงสุด (Maximum Likelihood Estimation : MLE) ซึ่งกำหนดในรูปฟังก์ชันดังนี้

$$L(U_{pj}|\theta) = \prod_{p=1}^N \prod_{j=1}^n P(U_{pj})^{u_{pj}} [1 - P(U_{pj})]^{1-u_{pj}}$$

เมื่อ N แทน ขนาดกลุ่มตัวอย่าง
 n แทน ความยาวของแบบสอบ

ค่าประมาณของพารามิเตอร์โดยใช้วิธีการประมาณค่าความน่าจะเป็นสูงสุดมีการแจกแจงแบบปกติของตัวแปรพหุในรูปเชิงเส้นกำกับ (asymptotically multivariate normal) ซึ่งมีค่าเฉลี่ยของเวกเตอร์ τ และเมทริกซ์ความแปรปรวน และความแปรปรวนร่วมในรูป Σ ในขณะที่ Σ^{-1} เป็นเมทริกซ์สารสนเทศกำหนดดังนี้

$$\Sigma^{-1} = -E \left[\frac{\partial^2}{\partial \tau_r \partial \tau_s} \ln L \right] ; \quad r, s = 0, 1, 2, 3.$$

เมื่อ E แทน ค่าคาดหวังของเมทริกซ์
 $\ln L$ แทน ลอการิทึมของฟังก์ชันไลค์ลิฮูด

ดังนั้นการแจกแจงของการประมาณค่าพารามิเตอร์ด้วยวิธี MLE จะอยู่ในรูป

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

โดยที่ $\hat{\tau}' = [\tau_0, \tau_1, \tau_2, \tau_3]$ ส่วนความคลาดเคลื่อนมาตรฐานเชิงเส้นกำกับของค่าประมาณของ τ_s ($s = 0, 1, 2, 3$) เมื่อ s เป็นสมาชิกแนวเส้นทแยงมุมของ Σ สามารถคำนวณได้จากสูตร

$$SE \hat{\tau}_s = \sqrt{\Sigma_{ss}}$$

สำหรับการทดสอบสมมติฐานของการทำหน้าที่ต่างกันของข้อสอบ จะทดสอบสมาชิกของ τ_s ซึ่งสมมติฐานที่สนใจคือ $H_0 : \tau_2 = 0$ และ $H_0 : \tau_3 = 0$ สมมติฐานทั้งสองสามารถทดสอบพร้อม ๆ กัน (simultaneously) ดังนี้

$$H_0: C_\tau = 0$$

$$H_1: C_\tau \neq 0$$

โดยที่ C เป็นเมทริกซ์ขนาด 2×4 ดังนี้

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

ส่วนการทดสอบนัยสำคัญของสมมติฐาน จะใช้สถิติไค-กำลังสองที่ระดับชั้นความเป็นอิสระเท่ากับ 2 ($df = 2$) ดังนี้

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

ถ้า χ^2 มีค่ามากกว่า $\chi^2_{\alpha;2}$ แสดงว่าปฏิเสธสมมติฐานของข้อสอบที่ทำหน้าที่ไม่ต่างกัน (no DIF) นั่นคือ ข้อสอบทำหน้าที่ต่างกัน (DIF) นั่นเอง

ตอนที่ 5 งานวิจัยที่เกี่ยวข้องกับการทำหน้าที่ต่างกันของข้อสอบ

งานวิจัยที่เกี่ยวข้องกับการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบนั้น แบ่งเป็นงานวิจัยภายในประเทศ และงานวิจัยต่างประเทศ ดังนี้

งานวิจัยในประเทศ

จากการศึกษารายงานการวิจัยภายในประเทศที่ผ่านมา มีผู้ทำการวิจัยเกี่ยวกับการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบไว้หลายท่าน โดยศึกษาในการเปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธีต่างๆ กับกลุ่มย่อยที่มีลักษณะแตกต่างกันในด้านภาษา เพศ และสภาพภูมิศาสตร์ รวมทั้งการเปรียบเทียบคุณภาพของแบบทดสอบหลังคัดข้อสอบที่ทำหน้าที่ต่างกันออกแล้ว ซึ่งผู้วิจัยจะได้นำเสนอ ดังนี้

กาญจนา วัฒนสุนทร (2538 : 195 - 207) ได้ศึกษาการพัฒนาเกณฑ์ตัดสินข้อสอบลำเอียงทางเพศ ด้วยข้อมูลเชิงประจักษ์ โดยใช้ดัชนีความลำเอียง 4 ตัว คือ พื้นที่ระหว่างโค้งการตอบข้อสอบชนิดคิดเครื่องหมาย (SA) และไม่คิดเครื่องหมาย (UA) จากทฤษฎีการตอบข้อสอบโมเดลสองพารามิเตอร์ ดัชนีแอลฟา (α_{MH}) จากวิธี MH และดัชนีเบต้า (β_{SIB}) จากวิธี SIBTEST ใช้ข้อมูลการตอบข้อสอบคัดเลือกบุคคลเข้าศึกษาในสถาบันอุดมศึกษาของทบวงมหาวิทยาลัย ปีการศึกษา 2535 ความยาวของแบบทดสอบเป็น 20, 30 และ 40 ข้อ สำหรับวิชาคณิตศาสตร์ และ 50, 60, 70 และ 80 ข้อ สำหรับวิชาภาษาอังกฤษ ขนาดกลุ่มผู้สอบมี 6 ขนาด คือ 100, 200, 400, 600, 800 และ 1,000 คน การพัฒนาเกณฑ์กระทำโดยคำนวณค่าดัชนีทั้ง 4 ตัว จากข้อมูลการตอบข้อสอบของผู้สอบเพศเดียวกัน เพศละ 50 ค่า สำหรับแต่ละความยาวแบบทดสอบ และขนาดกลุ่มผู้สอบ จากนั้นนำค่าดัชนีที่ได้ทั้งหมดมาวิเคราะห์หาค่าเฉลี่ย และกำหนดเกณฑ์จากค่าเฉลี่ยสองลักษณะ คือ เกณฑ์ที่กำหนดจากค่าเฉลี่ยรวมจากค่าดัชนีทุกข้อ โดยไม่พิจารณาความแตกต่างในด้านความยาวของแบบทดสอบและขนาดกลุ่มผู้สอบด้วย และเกณฑ์ที่พัฒนาจากค่าเฉลี่ยซึ่งพิจารณาถึงความยาวของแบบทดสอบและขนาดของกลุ่มผู้สอบ จากนั้นนำเกณฑ์ที่กำหนดไว้ไปตัดสินค่าดัชนีที่ได้จากการวิเคราะห์ระหว่างผู้สอบเพศหญิงและชาย ข้อค้นพบที่สำคัญมีดังต่อไปนี้

1. เกณฑ์ที่พัฒนาจากข้อมูลเชิงประจักษ์เพื่อใช้ในการตัดสินความลำเอียงของข้อสอบระหว่างผู้สอบเพศหญิงและชาย คือ

$$(1) |SA| > .80 \text{ และ } UA > .50 \text{ กรณีความยาวแบบทดสอบที่สั้นกว่า 50}$$

(2) $|SA| > .40$ และ $UA > 1.20$ กรณีความยาวแบบทดสอบที่มีความยาวตั้งแต่ 50 ข้อขึ้นไป

(3) $.06 > \alpha_{MH} > 1.40$ และ $\beta_{SIB} > .06$ สำหรับทุกความยาวแบบทดสอบและทุกขนาดกลุ่มผู้สอบ ทั้งนี้ควรใช้ขนาดผู้สอบอย่างน้อย 800 คน สำหรับดัชนี SA และ UA และ 600 คน สำหรับดัชนี α_{MH} และดัชนี β_{SIB}

2. การตรวจค้นข้อสอบลำเอียงทางเพศมีความไม่คงที่ข้ามขนาดกลุ่มผู้สอบและความยาวของแบบทดสอบ

3. ความสอดคล้องในการตรวจค้นข้อสอบที่ลำเอียงภายในวิธีเดียวกันข้ามขนาดผู้สอบก่อนข้างต่ำ แต่จะสูงขึ้นที่ขนาดผู้สอบตั้งแต่ 600 คนขึ้นไป

4. ข้อสอบที่ลำเอียงในวิชาคณิตศาสตร์ส่วนใหญ่จะลำเอียงเข้าสู่เพศชายแต่วิชาภาษาอังกฤษจะลำเอียงเข้าสู่เพศหญิง เมื่อใช้ดัชนี SA และ α_{MH} แต่ดัชนี β_{SIB} ให้ผลตรงข้าม

เรวดี อินทสระ (2539 : 248 - 252) ได้ศึกษาความเที่ยงตรงเชิงพยากรณ์ของแบบทดสอบคัดเลือกที่วิเคราะห์ความลำเอียงต่อเพศด้วยวิธี IRT วิธี MH และวิธี SIBTEST การตัดสินผลการสอบที่คิดคะแนนมาตรฐานที่-ปกติ คะแนนน้ำหนักความสามารถและสาเหตุความลำเอียงของข้อสอบ โดยศึกษาความลำเอียงของข้อสอบคัดเลือกเข้าศึกษาในชั้นปีที่ 1 ประเภทรับตรง ปีการศึกษา 2538 ของมหาวิทยาลัยสงขลานครินทร์ ในวิชาภาษาไทย ก วิชาสังคม ก วิชาภาษาอังกฤษ กข วิชาละ 8,127 คน (ชาย 2,722 คน หญิง 5,405 คน) วิชาภาษาไทย กข วิชาสังคม กข วิชาภาษาอังกฤษ กขค วิชาละ 5,415 คน (ชาย 1,454 คน หญิง 3,961 คน) ความเที่ยงตรงเชิงพยากรณ์ศึกษาจากคะแนนสอบคัดเลือกกับระดับผลการเรียนภาคเรียนที่ 1 ปีการศึกษา 2538 ของนักศึกษาที่ได้รับการคัดเลือกจากประเภทรับตรง สายวิทยาศาสตร์ 763 คน และสายศิลปศาสตร์ 281 คน และสาเหตุความลำเอียงของข้อสอบจากการระบุสาเหตุของนักวัดผลการศึกษาหรืออาจารย์ผู้สอนจำนวน 50 คน และนิสิตนักศึกษาที่เรียนในสาขาวิชานั้นๆ วิชาละ 50 คน

ผลการศึกษาพบว่าวิธีการตรวจสอบความลำเอียงทั้งสามวิธีตัดสินจำนวนข้อสอบที่ลำเอียงแตกต่างกันในวิชาภาษาไทย ฉบับ ก ฉบับที่ 2 และวิชาสังคม ก ฉบับที่ 1 ที่ระดับนัยสำคัญทางสถิติ .05 นอกนั้นแตกต่างกันที่ระดับนัยสำคัญทางสถิติที่ระดับ .01 โดยวิธี IRT ตัดสินจำนวนข้อที่ลำเอียงได้มากที่สุด ความสัมพันธ์ของลำดับที่การสอบไม่ว่าจะคิดคะแนนมาตรฐานที่-ปกติ หรือคิดคะแนนน้ำหนักความสามารถ และใช้จำนวนข้อทั้งหมดหรือใช้ข้อสอบที่ปราศจากความลำเอียงต่างมีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ประสิทธิภาพในการทำนายผลสัมฤทธิ์ทางการเรียนในสายวิทยาศาสตร์ การคิดคะแนนน้ำหนักความสามารถทั้งใช้ข้อสอบจำนวนทั้งหมดและใช้เฉพาะข้อสอบที่ปราศจากความลำเอียงมีประสิทธิภาพในการทำนายสูงกว่าการคิดคะแนนมาตรฐานที่-ปกติ ส่วนสายศิลปศาสตร์ การคิดคะแนนมาตรฐานที่-ปกติทั้งใช้ข้อสอบจำนวนทั้งหมด และใช้เฉพาะข้อสอบที่ปราศจากความลำเอียง มีประสิทธิภาพในการทำนายสูงกว่าการคิดคะแนนน้ำหนักความสามารถ และสาเหตุของความลำเอียงของข้อสอบต่อเพศทั้งชายและหญิง เกิดจากข้อสอบเป็นข้อคำถามที่ผู้สอบได้รับการฝึกฝนเฉพาะจะมีโอกาสตอบถูกมากกว่าเป็นเรื่องราวที่กลุ่มนั้นๆสนใจ และเป็นข้อสอบที่ตามความจำ

คมศักดิ์ ชื่นชม (2539 : 80) ได้ทำการวิเคราะห์ความลำเอียงของแบบทดสอบวัดจริยธรรมด้านความซื่อสัตย์ ชั้นประถมศึกษาปีที่ 6 จำนวน 30 ข้อ ที่ผู้วิจัยสร้างขึ้น ระหว่างกลุ่มผู้สอบที่ใช้ภาษาในชีวิตประจำวันในครอบครัวต่างกัน คือภาษาไทย และไม่ใช่ภาษาไทยและนับถือศาสนาต่างกันคือศาสนาพุทธและศาสนาอิสลาม ด้วยวิธีวิเคราะห์สามวิธี คือ วิธีค่าอำนาจจำแนก วิธีวิเคราะห์องค์ประกอบ วิธีวิเคราะห์การถดถอยโดยทำการเปรียบเทียบจำนวนข้อสอบที่ลำเอียงและศึกษาเปรียบเทียบความเชื่อมั่นของแบบทดสอบ หลังจากคัดเลือกข้อที่มีความลำเอียงออกแล้ว กลุ่มตัวอย่างที่ใช้เป็นนักเรียนชั้นประถมศึกษาปีที่ 6 ภาคเรียนที่ 2 ปีการศึกษา 2538 ของโรงเรียนในจังหวัดยะลา ซึ่งได้มาโดยวิธีการสุ่มแบบแบ่งชั้น โดยใช้ภาษาที่นักเรียนใช้ในชีวิตประจำวันเป็นชั้น จำนวน 1,081 คนแบ่งออกเป็น 2 กลุ่ม กลุ่มที่ 1 ใช้ในการศึกษาเปรียบเทียบผลการวิเคราะห์ความลำเอียงจำนวน 800 คน กลุ่มที่ 2 ใช้เปรียบเทียบความเชื่อมั่นของแบบทดสอบ หลังคัดเลือกข้อที่มีความลำเอียงออกแล้ว จำนวน 281 คน

ผลการศึกษาพบว่า การวิเคราะห์ความลำเอียงของแบบทดสอบทั้ง 3 วิธี มีจำนวนข้อที่มีความลำเอียงน้อย และแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ ส่วนค่าความเชื่อมั่นของแบบทดสอบวัดจริยธรรม ด้านความซื่อสัตย์ที่คำนวณด้วยสูตรแอลฟา (α - Coefficient) หลังคัดเลือกข้อที่ลำเอียงออกแล้ว พบว่ามีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

จิตสุตา ธารพร (2539 : 131 - 134) ได้ทำการวิเคราะห์ความลำเอียงของแบบทดสอบโจทย์ปัญหาทางคณิตศาสตร์และแบบทดสอบวัดความเข้าใจทางภาษาของนักเรียนชั้นมัธยมศึกษาปีที่ 2 ระหว่างกลุ่มนักเรียนที่นับถือศาสนาต่างกัน ด้วยวิธีวิเคราะห์สามวิธีคือ วิธีแปลงค่าความยากของข้อสอบ วิธีโค้งลักษณะข้อสอบแบบสามพารามิเตอร์ และวิธี MH โดยทำการเปรียบเทียบจำนวนข้อสอบที่ลำเอียง และศึกษาเปรียบเทียบความเชื่อมั่นของแบบทดสอบหลังคัดเลือกข้อสอบที่มีความลำเอียงออกแล้ว กลุ่มตัวอย่างที่ใช้เป็นนักเรียนชายหญิงชั้นมัธยมศึกษาปีที่ 2 ภาคเรียนที่ 2 ปีการศึกษา 2538 ของโรงเรียนสังกัดกรมสามัญศึกษาในจังหวัดกระบี่ ซึ่งเลือกโดยการสุ่มอย่างง่าย จำนวน 2,689 คน เครื่องมือที่ใช้มี 2 ฉบับ ฉบับที่ 1 คือแบบทดสอบโจทย์ปัญหาทางคณิตศาสตร์เรื่องสมการและอสมการ อัตราส่วนและร้อยละ ปริมาตรและพื้นที่ผิว ชนิด 5 ตัวเลือก จำนวน 40 ข้อ ฉบับที่ 2 เป็นแบบทดสอบความเข้าใจทางภาษา วิชาภาษาไทย ชนิด 5 ตัวเลือก จำนวน 40 ข้อ ซึ่งแบบทดสอบทั้ง 2 ฉบับ ผู้วิจัยเป็นผู้สร้างขึ้นเอง

ผลการศึกษาพบว่าวิธีวิเคราะห์ความลำเอียงของข้อสอบ จากแบบทดสอบทั้ง 2 ฉบับ แต่ละวิธีพบจำนวนข้อสอบที่มีความลำเอียงแตกต่างกัน ซึ่งวิธี MH พบจำนวนข้อสอบที่มีความลำเอียงมากที่สุด และวิธีแปลงค่าความยากของข้อสอบพบจำนวนข้อสอบที่มีความลำเอียงน้อยที่สุด โดยที่การวิเคราะห์ความลำเอียงด้วยวิธีแปลงค่าความยากของข้อสอบกับวิธีโค้งลักษณะข้อสอบ และวิธีแปลงค่าความยากของข้อสอบกับวิธี MH มีจำนวนข้อสอบที่ลำเอียงแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .001 ส่วนวิธีโค้งลักษณะข้อสอบกับวิธี MH มีจำนวนข้อสอบที่ลำเอียงแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ สำหรับค่าความเชื่อมั่นของแบบทดสอบโจทย์ปัญหาทางคณิตศาสตร์ที่คำนวณโดยใช้สูตร KR-20 หลังจากคัดเลือกข้อสอบที่มีความลำเอียงออกแล้ว ด้วยวิธีแปลงค่าความยากของข้อสอบกับวิธีโค้งลักษณะข้อสอบและวิธีแปลงค่าความยากของข้อสอบกับวิธี MH มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 และวิธีโค้งลักษณะข้อสอบกับวิธี MH มีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ ส่วนค่าความเชื่อมั่นที่คำนวณโดยใช้สูตรแบบแบ่งครึ่งฉบับ หลังจากคัดเลือกข้อสอบที่มีความลำเอียงออกแล้ว พบว่าทั้ง 3 วิธี มีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ สำหรับค่าความเชื่อมั่นของแบบทดสอบความเข้าใจทางภาษาซึ่งคำนวณด้วยสูตร KR-20 และแบบแบ่งครึ่งฉบับ หลังจากคัดเลือกข้อสอบที่มีความลำเอียงออกแล้วทั้ง 3 วิธี พบว่า มีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

อังคณา สายยศ (2539 : 24 - 26) ได้ศึกษาผลการวิเคราะห์ความลำเอียงของแบบทดสอบความเข้าใจในการอ่านตามลักษณะการใช้ภาษาพูดในครอบครัวเป็นภาษาลาว เขมร และภาษาส่วย โดยศึกษาเปรียบเทียบจำนวนข้อที่ลำเอียง ความเชื่อมั่นของแบบทดสอบหลังคัดข้อที่ลำเอียงออกแล้ว และคะแนนที่ได้จากการสอบหลังคัดข้อที่ลำเอียงออกแล้วกลุ่มตัวอย่างเป็นนักเรียนชั้นประถมที่ 6 ในจังหวัดสุรินทร์ ได้มาโดยการสุ่มแบบแบ่งชั้น อำเภอเป็นชั้น และโรงเรียนเป็นหน่วยการสุ่ม ได้จำนวนโรงเรียนทั้งสิ้น 29 โรงเรียน นักเรียน 860 คน ทำการวิเคราะห์ความลำเอียงด้วยวิธี MH ผลการวิจัยพบว่าจำนวนข้อสอบที่ลำเอียงของแบบทดสอบความเข้าใจในการอ่านระหว่างกลุ่มผู้สอบที่ใช้ภาษาพูดเป็นภาษาลาว ภาษาเขมร ภาษาส่วยแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ สำหรับค่าความเชื่อมั่นของแบบทดสอบหลังคัดข้อที่ลำเอียงออกแล้ว

พบว่า ค่าความเชื่อมั่นระหว่างกลุ่มผู้สอบที่ใช้ภาษาพูดเป็นภาษาลาว ภาษาเขมร และภาษาส่วย แตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ ส่วนคะแนนที่ได้จากการสอบด้วยแบบทดสอบที่คัดข้อที่ลำเอียงออกแล้วมีคะแนนระหว่างกลุ่มผู้สอบที่พูดภาษาลาวกับภาษาส่วย และภาษาเขมรกับภาษาส่วยมีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติ ส่วนภาษาลาวกับภาษาเขมร มีคะแนนแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

เกษร ห่วงจิตร (2539 : 122 - 128) ได้วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธี MH โดยกลุ่มอ้างอิงและกลุ่มเปรียบเทียบจำแนกตามเพศ ภูมิสำเนา ประสบการณ์ในการสอบ และสังกัดของสถานศึกษา ข้อมูลที่ใช้เป็นผลการตอบข้อสอบวิชาภาษาไทยของผู้สอบจำนวน 506 คน และผลการตอบข้อสอบวิชาภาษาอังกฤษของผู้สอบจำนวน 501 คน ในส่วนที่เป็นข้อสอบแบบเลือกตอบของศูนย์ทดสอบทางการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ผลการศึกษาพบว่า ข้อสอบที่ทำหน้าที่ต่างกันส่วนมากจะเป็นข้อสอบที่มีคำอ่านจำแนกก่อนข้างต่ำทั้งสองวิชา เมื่อพิจารณาในด้านค่าความยากพบว่าข้อสอบที่ทำหน้าที่ต่างกันส่วนมากเป็นข้อสอบที่ง่ายมากสำหรับวิชาภาษาไทย ส่วนวิชาภาษาอังกฤษข้อสอบที่ทำหน้าที่ต่างกันส่วนมากเป็นข้อที่ยากมาก อีกทั้งพบว่าส่วนมากเป็นข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ เมื่อจำแนกกลุ่มอ้างอิงและกลุ่มเปรียบเทียบตามเพศ จะพบข้อสอบที่ทำหน้าที่ต่างกันมีจำนวนมากที่สุด รองลงมาคือการจำแนกตามภูมิสำเนา สังกัดของสถานศึกษาและประสบการณ์ในการสอบ ตามลำดับ

เสรี ชัดแฉ่ม (2539 : 84 - 126) ได้เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบไม่สม่ำเสมอระหว่างวิธี MH แบบปกติ กับวิธี MH แบบแบ่งกลุ่มความสามารถของผู้สอบและความยากของแบบทดสอบ โดยกลุ่มอ้างอิงและกลุ่มเปรียบเทียบจำแนกตามเพศ คือเพศชาย 600 คน และเพศหญิง 600 คน เครื่องมือที่ใช้เก็บรวบรวมข้อมูลคือแบบทดสอบวัดความสามารถในการอ่านภาษาไทย ซึ่งผู้วิจัยสร้างขึ้น

ผลการศึกษาพบว่า วิธี MH แบบปกติ ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอสอดคล้องกับวิธี IRT Area คิดเป็นร้อยละ 33.33 วิธี MH แบบแบ่งกลุ่มความสามารถของผู้สอบและความยากของข้อสอบ ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอสอดคล้องกับวิธี IRT Area คิดเป็นร้อยละ 61.11 และเพิ่มขึ้นจากวิธี MH แบบปกติ คิดเป็นร้อยละ 27.33

พรณี จินตมาศ (2540 : 90 - 92) ได้ทำการวิเคราะห์ความลำเอียงต่อเพศของแบบทดสอบคณิตศาสตร์โจทย์ปัญหาชั้นมัธยมศึกษาปีที่ 1 ด้วยวิธีวิเคราะห์ความลำเอียง สามวิธี คือวิธีแปลงค่าความยาก วิธี MH วิธี SIBTEST ในแต่ละขนาดของกลุ่มผู้สอบ 500 และ 1000 คน โดยเปรียบเทียบจำนวนข้อที่มีความลำเอียงและเปรียบเทียบค่าความเชื่อมั่นแบบครึ่งฉบับของแบบทดสอบหลังคัดเลือกข้อสอบที่มีความลำเอียงออกแล้ว กลุ่มตัวอย่างเป็นนักเรียนชั้นมัธยมศึกษาปีที่ 1 ภาคเรียนที่ 2 ปีการศึกษา 2539 ของโรงเรียนสังกัดกรมสามัญศึกษา ส่วนกลาง จำนวน 2,200 คน ซึ่งเลือกมาโดยการสุ่มแบบแบ่งชั้น มีขนาดของโรงเรียนเป็นชั้นและโรงเรียนเป็นหน่วยการสุ่ม เครื่องมือที่ใช้คือแบบทดสอบคณิตศาสตร์โจทย์ปัญหาเรื่อง สมการ อัตราส่วน และร้อยละ แบบเลือกตอบชนิด 5 ตัวเลือก จำนวน 40 ข้อ ที่ผู้วิจัยสร้างขึ้นเอง จำนวน 1 ฉบับ

ผลการวิจัยพบว่า เมื่อวิเคราะห์จากกลุ่มผู้สอบขนาด 500 คน วิธี SIBTEST พบข้อสอบที่มีความลำเอียงมากที่สุดและวิธีแปลงค่าความยากพบข้อสอบที่มีความลำเอียงน้อยที่สุด โดยจำนวนข้อสอบที่มีความลำเอียงแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ ทุกวิธีวิเคราะห์ และเมื่อวิเคราะห์จากกลุ่มผู้สอบขนาด 1,000

คน วิธี MH พบข้อสอบมีความลำเอียงมากที่สุด วิธีแปลงค่าความยากไม่พบข้อสอบที่ลำเอียง โดยจำนวนข้อสอบที่ลำเอียงจากการวิเคราะห์ด้วยวิธีแปลงค่าความยาก กับ วิธี MH และวิธีแปลงค่าความยาก กับ วิธี SIBTEST แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นอกนั้นมีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ สำหรับจำนวนข้อสอบที่ลำเอียงจากการวิเคราะห์ด้วยวิธีแปลงค่าความยากระหว่างกลุ่มผู้สอบขนาด 500 คนและขนาดกลุ่มผู้สอบขนาด 1,000 คน จะมีจำนวนข้อแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ.05 นอกนั้นมีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ สำหรับค่าความเชื่อมั่นของแบบทดสอบหลังคัดเลือกข้อสอบที่มีความลำเอียงออกแล้ว เมื่อวิเคราะห์จากกลุ่มผู้สอบขนาด 500 คน จากการวิเคราะห์ความลำเอียงด้วยวิธีแปลงค่าความยากกับวิธี SIBTEST มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ.05 นอกนั้นมีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ และเมื่อวิเคราะห์จากกลุ่มตัวอย่างขนาด 1,000 คน ค่าความเชื่อมั่นของแบบทดสอบหลังคัดเลือกข้อสอบที่ลำเอียงออกแล้วจากการวิเคราะห์ความลำเอียง 3 วิธี มีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ สำหรับค่าความเชื่อมั่นของแบบทดสอบหลังคัดเลือกข้อสอบที่ลำเอียงออกแล้วจากการวิเคราะห์ด้วยวิธี SIBTEST ระหว่างกลุ่มผู้สอบขนาด 500 คน และกลุ่มผู้สอบ 1,000 คน มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นอกนั้นมีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

จิตติมา วรณศรี (2540 : 90 – 99) ได้ศึกษาเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธี MH และวิธี SIBTEST โดยใช้ข้อมูลจำลองจากโปรแกรม IRTDATA เงื่อนไขที่ศึกษาได้แก่ (1) ความยาวของแบบทดสอบ 3 ขนาด คือ 30 ข้อ 60 ข้อ และ 90 ข้อ (2) ขนาดกลุ่มตัวอย่าง 3 ขนาด คือ 200 คน 600 คน และ 1,000 คน โดยแต่ละขนาดมีอัตราส่วนระหว่างผู้สอบกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบเป็น 4 อัตราส่วน คือ 1 : 1, 1 : 0.9, 1 : 0.75, 1 : 0.5 รวมเงื่อนไขที่ศึกษาทั้งหมด 36 เงื่อนไข

ผลการศึกษาพบว่า วิธี MH กับวิธี SIBTEST มีประสิทธิภาพเท่าเทียมกันในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ทุกขนาดกลุ่มตัวอย่างและทุกอัตราส่วน ภายใต้ความยาวแบบทดสอบเดียวกัน เมื่อขนาดกลุ่มตัวอย่าง 200 และ 600 คน สามารถตรวจสอบพบข้อสอบที่ทำหน้าที่ต่างกันได้ถูกต้องร้อยละ 50 และเมื่อขนาดกลุ่มตัวอย่าง 1,000 คน สามารถตรวจสอบพบข้อสอบที่ทำหน้าที่ต่างกันได้ถูกต้องร้อยละ 100 โดยส่วนมากวิธี SIBTEST มีอัตราความคลาดเคลื่อนประเภทที่ 1 มากกว่าวิธี MH เล็กน้อย และพบว่าวิธี MH และวิธี SIBTEST มีประสิทธิภาพเท่าเทียมกันที่ทุกระดับความยาวของแบบทดสอบโดยพบว่าเมื่อใช้แบบทดสอบที่มีความยาวปานกลาง (60 ข้อ) มีประสิทธิภาพในการตรวจสอบมากที่สุด

งานวิจัยต่างประเทศ

สำหรับงานวิจัยที่เกี่ยวข้องกับการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในต่างประเทศนั้นมีผู้วิจัยได้ทำการวิจัยไว้หลายท่าน ทั้งที่เป็นการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบจากข้อมูลจำลองและการวิเคราะห์จากข้อมูลจริง ซึ่งผู้วิจัยรวบรวมและสรุปได้ดังนี้

ผจงจิต อินทสุวรรณ (Intasuwan, 1979 : 92) ศึกษาวิธีวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบสามวิธี คือ วิธีโค้งลักษณะข้อสอบแบบสามพารามิเตอร์ วิธีราสช์ และวิธีไค-กำลังสอง ของแบบทดสอบ IEA (The International Association of Education Achievement) กลุ่มตัวอย่างที่ใช้ศึกษาเป็นนักเรียนอายุ 14 ปี ชาวอังกฤษ อเมริกัน และนิวซีแลนด์ จำนวน 8,601 คน

ผลการวิจัยพบว่าสหสัมพันธ์ระหว่างวิธีทั้ง 3 มีค่าอยู่ระหว่าง .51 - .98 และวิธีไค-กำลังสองกับวิธีราสช์ มีสหสัมพันธ์กันสูงถึง .98

สงบ ลักษณะ (Laksana, 1979 : 168) ได้วิเคราะห์การทำหน้าที่ต่างกันต่อเพศและผิว ในแบบทดสอบ ITBS (IOWA Test of Basic Skills) ฉบับภาษาต้นคำศัพท์และฉบับคณิตศาสตร์ ด้วยวิธีโค้งลักษณะข้อสอบแบบสามพารามิเตอร์ และวิธีวิเคราะห์ความแปรปรวน

ผลการวิจัยพบว่าวิธีวิเคราะห์ความแปรปรวนเป็นวิธีที่ใช้ได้ดีในทางปฏิบัติแต่เป็นวิธีที่ไม่ครอบคลุม ส่วนวิธีโค้งลักษณะข้อสอบแบบสามพารามิเตอร์ เป็นวิธีที่ครอบคลุมกว่า แต่ต้องใช้กลุ่มตัวอย่างขนาดใหญ่และการคำนวณยุ่งยากซับซ้อน

สับโคเวียต และคณะ (Subkoviak and others, 1984 : 49-59) ได้ศึกษาวิธีวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบสามวิธี คือวิธีโค้งลักษณะข้อสอบแบบสามพารามิเตอร์ วิธีไค-กำลังสอง และวิธีแปลงค่าความยาก ของแบบทดสอบวัดความรู้เกี่ยวกับคำศัพท์ 50 ข้อ เป็นคำศัพท์ภาษาอังกฤษมาตรฐาน จำนวน 40 ข้อ ศัพท์แสงของชาวผิวดำ จำนวน 10 ข้อ กลุ่มตัวอย่างที่ใช้ในการศึกษาเป็นนักเรียนผิวขาว 1,021 คน นักเรียนผิวดำ จำนวน 1,008 คน

ผลการวิจัยพบว่าวิธีโค้งลักษณะข้อสอบแบบสามพารามิเตอร์เป็นวิธีที่มีประสิทธิภาพมากที่สุด รองลงมาคือวิธีไค-กำลังสอง ส่วนวิธีแปลงค่าความยากเป็นวิธีที่นำไปใช้ในทางปฏิบัติได้ ในกรณีไม่มีเครื่องคอมพิวเตอร์

ทิสเซน สเตนเบิร์ก และ ไวน์เนอร์ (Thissen, Steinberg and Wainer, 1988) ได้ศึกษาความแตกต่างของโค้งการตอบข้อสอบระหว่างกลุ่มด้วยทฤษฎีการตอบข้อสอบเพื่อศึกษาถึงสภาพการณ์ในการทดสอบสมมติฐานความเท่ากันของค่าพารามิเตอร์ด้วยค่าไค-กำลังสองแบบอัตราส่วน Likelihood (IRT-LR) และวิธี MH โดยการจำลองข้อมูล

ผลการศึกษาพบว่า วิธีไค-กำลังสองแบบอัตราส่วน Likelihood (IRT-LR) สามารถใช้วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบได้เมื่อมีจำนวนกลุ่มตัวอย่างน้อยๆ คือประมาณ 500 คน ความแตกต่างของค่าความยากของข้อสอบระหว่างกลุ่มมีค่าเท่ากับ 0.3 และความแตกต่างในด้านค่าสัมประสิทธิ์การเดามีค่าเท่ากับ 0.1 โดยจำนวนข้อสอบรวมจะมากหรือน้อยก็ไม่มีผลกระทบถ้าความแตกต่างมีค่าน้อยกว่านี้ไม่สามารถที่จะใช้วิธีไค-กำลังสองแบบอัตราส่วนไลค์ลิฮูด (Likelihood : IRT-LR) วิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันได้ แต่ถ้าใช้

กลุ่มตัวอย่างมากกว่านี้ก็อาจจะทำการวิเคราะห์ได้ การระบุถึงการทำหน้าที่ต่างกันของข้อสอบใช้วิธีเปรียบเทียบความแตกต่างระหว่างกลุ่มในข้อสอบที่ต้องการวิเคราะห์การทำหน้าที่ต่างกับข้อสอบร่วมซึ่งมีข้อตกลงว่าข้อสอบร่วมนั้นทำหน้าที่ไม่ต่างกัน สำหรับผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธีไค-กำลังสองแบบอัตราส่วน Likelihood กับวิธี MH ทั้งสองวิธีให้ผลการวิเคราะห์คล้ายคลึงกัน ข้อเสนอแนะจากการวิจัยครั้งนี้คือ วิธี MH อาจจะใช้ในการวิเคราะห์การทำหน้าที่ต่างกันของแบบทดสอบยาวๆหรืออาจใช้เป็นเครื่องมือในการกลั่นกรองข้อสอบก่อนที่จะใช้วิธีไค-กำลังสองแบบอัตราส่วน Likelihood (IRT-LR) ต่อไป

สวามินาธานและโรเจอร์ส (Swaminathan and Rogers. 1990) ได้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี LR กับวิธี MH โดยใช้ขนาดตัวอย่างสองขนาด คือ 200 และ 500 คน และความยาวของแบบทดสอบสามขนาด คือ 40, 60 และ 80 ข้อ โดยจำลองข้อมูลการตอบข้อสอบด้วยโปรแกรม DATAGEN แบบสองพารามิเตอร์ โดยจำลองข้อมูลที่ทำหน้าที่ต่างกันสองแบบ คือ การทำหน้าที่ต่างกันของข้อสอบแบบสม่ำเสมอ (Uniform DIF) และการทำหน้าที่ต่างกันของข้อสอบแบบไม่สม่ำเสมอ (Nonuniform DIF) โดยให้ค่าจำแนกของทั้งสองกลุ่มเท่ากันส่วนค่าความยากของข้อสอบจะมีค่าแปรเปลี่ยนไปตามกลุ่มผู้สอบ เพื่อให้เกิดระดับการทำหน้าที่ต่างกันต่าง ๆ กันตามต้องการแต่ในลักษณะการทำหน้าที่ต่างกันของข้อสอบแบบไม่สม่ำเสมอ (Nonuniform DIF) นั้นให้ค่าความยากของข้อสอบของสองกลุ่มเท่ากัน แต่ค่าอำนาจจำแนกแปรเปลี่ยนไปตามกลุ่ม

ผลการศึกษาพบว่าตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้งสองวิธี ให้ผลใกล้เคียงกัน ในกรณีการทำหน้าที่ต่างกันของข้อสอบแบบสม่ำเสมอ (Uniform DIF) วิธี MH ตรวจพบการทำหน้าที่ต่างกันของข้อสอบได้ดีกว่าวิธี LR เล็กน้อย โดยมีการตรวจพบการทำหน้าที่ต่างกันได้ถูกต้องร้อยละ 75 กรณีใช้กลุ่มตัวอย่าง 200 คน และตรวจพบได้ถูกต้องร้อยละ 100 ในกรณีใช้กลุ่มตัวอย่าง 500 คน ในทุกความยาวของแบบทดสอบ แต่ในกรณีของการทำหน้าที่ต่างกันของข้อสอบแบบไม่สม่ำเสมอ (Nonuniform DIF) นั้นวิธี MH ไม่สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ ส่วนวิธี LR ตรวจพบการทำหน้าที่ต่างกันของข้อสอบได้ถูกต้องร้อยละ 50 ในกรณีกลุ่มตัวอย่างน้อยและแบบทดสอบสั้น และตรวจสอบการทำหน้าที่ต่างกันได้ถูกต้องร้อยละ 75 ในกรณีกลุ่มตัวอย่างขนาดใหญ่และแบบทดสอบยาว สำหรับการตรวจสอบการจำแนกผิดพลาด (Type I error rate) พบว่าวิธี MH มีประสิทธิภาพดีกว่า โดยตรวจสอบผิดพลาดประมาณร้อยละ 1 ส่วนวิธี LR ตรวจสอบผิดพลาดประมาณร้อยละ 1 - 6 นอกจากนี้ยังพบว่าวิธี LR จะเสียค่าใช้จ่ายมากกว่าประมาณ 3 - 4 เท่า

คลอสเซอร์ และคณะ (Clauser and others. 1991) ได้ศึกษาอิทธิพลต่างๆที่มีผลต่อการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี MH ข้อมูลที่ใช้เป็นข้อมูลจำลองขึ้นเพื่อใช้ในการวิเคราะห์การทำหน้าที่ต่างกันแบบสม่ำเสมอ (Uniform DIF) โดยจำลองข้อมูลเป็นแบบทดสอบจำนวน 75 ข้อ เป็นข้อสอบที่ลำเอียง 16 ข้อและมีระดับความแตกต่างของค่าความยากหลายระดับระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ กลุ่มตัวอย่างที่ใช้ศึกษามีจำนวน 3,000 คน เป็นกลุ่มอ้างอิง 1,500 คน และกลุ่มเปรียบเทียบ 1,500 คน มีการวิเคราะห์ระดับความยากที่ต่างกันหลายระดับ กับระดับความยากของแบบทดสอบ 5 ระดับ ค่าอำนาจจำแนก 4 ระดับ และการกระจายความสามารถ 2 ลักษณะ

ผลการศึกษาพบว่าสิ่งที่มีอิทธิพลต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี MH คือความแตกต่างในระดับความยากระหว่างผู้สอบทั้งสองกลุ่ม รวมทั้งค่าอำนาจจำแนกและค่าความยากของ

ข้อสอบด้วย จากการวิจัยครั้งนี้ทำให้ได้ข้อเสนอแนะว่าวิธี MH นี้ใช้ได้สำหรับข้อสอบที่มีค่าอำนาจจำแนกสูง และมีแนวโน้มว่าจะไม่สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ในข้อสอบที่มีค่าความยากสูงได้

เมเซอร์และคณะ (Mazor and others. 1992) ได้ใช้วิธี MH ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยศึกษาข้อมูลที่จำลองขึ้นตามทฤษฎีการตอบข้อสอบ ซึ่งใช้กลุ่มตัวอย่างแตกต่างกัน 5 ขนาด ได้แก่ 2,000, 1,000, 500, 200 และ 100 คน เพื่อเปรียบเทียบอัตราการตรวจค้นโดยใช้ข้อสอบที่จำลองขึ้น 5 ชุด ชุดละ 75 ข้อ ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยการทดสอบ $MH-\chi^2$ ที่ระดับ $df = 1$

ผลการศึกษาพบว่า เมื่อใช้กลุ่มตัวอย่างขนาด 2,000 คน วิธีนี้วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบผิดพลาดร้อยละ 30 และเมื่อใช้จำนวนผู้สอบขนาด 500 คนลงมา พบว่าวิธีนี้จะวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบได้ผิดพลาดร้อยละ 50 ส่วนข้อสอบที่ไม่สามารถตรวจพบการทำหน้าที่ต่างกันของข้อสอบได้แก่ข้อสอบที่ยากมากๆ และเป็นข้อสอบที่มีค่าอำนาจจำแนกต่ำ

คลอสเซอร์และคณะ (Clauser and others. 1993) ได้ศึกษาอิทธิพลของการทำให้เกณฑ์การจับคู่ผู้สอบที่มีความบริสุทธิ์ (purification of the matching criterion) ระหว่างเทคนิคหนึ่งขั้นตอน (one step procedure) และเทคนิคสองขั้นตอน (two-step procedure) ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธี MH โดยการใช้สถานการณ์จำลอง สร้างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบกลุ่มละ 1,000 คน จำลองสถานการณ์ตามเงื่อนไขต่างๆ กัน 24 เงื่อนไข คือ สร้างแบบทดสอบยาว 3 ขนาด คือ 20 ข้อ 40 ข้อ และ 80 ข้อ สร้างข้อสอบที่ทำหน้าที่ต่างกันลงในแบบทดสอบแต่ละขนาดจำนวน 0% , 3% , 8% และ 20% รวมทั้งสร้างเงื่อนไขระดับความสามารถ 2 ระดับ

ผลการศึกษาปรากฏว่า ผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยเทคนิคสองขั้นตอน เท่ากับหรือเหนือกว่าเทคนิคหนึ่งขั้นตอนในทุกเงื่อนไขของการทดสอบ และเทคนิคสองขั้นตอน ไม่เพิ่มการจำแนกผิดพลาด (Type I error rate) มากกว่าเทคนิคหนึ่งขั้นตอน

โรเจอร์ส และสวามินาธาน (Rogers and Swaminathan. 1993) ได้เปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธี LR กับ วิธี MH โดยใช้ข้อมูลที่จำลองขึ้น เงื่อนไขการศึกษาเป็น 4 แบบ คือ ขนาดกลุ่มตัวอย่าง 2 กลุ่ม ขนาด 250 คน และ 500 คน ต่อกลุ่ม ระดับความเหมาะสมของโมเดล ข้อมูลเป็น 2 ระดับ คือระดับที่เหมาะสมใช้รูปแบบสองพารามิเตอร์ ระดับที่ไม่เหมาะสมใช้รูปแบบสามพารามิเตอร์ โดยค่าการเดา (c) จะถูกกำหนดที่ .2 เครื่องมือที่ใช้เป็น ข้อสอบจำนวน 40 ข้อ ที่จำลองขึ้นมา 7 แบบ ดังนี้ แบบที่ 1 ค่าความยากและค่าอำนาจจำแนกต่ำ แบบที่ 2 ค่าความยากสูงค่าอำนาจจำแนกต่ำ แบบที่ 3 ค่าความยากปานกลางค่าอำนาจจำแนกต่ำ แบบที่ 4 ค่า ความยากปานกลางค่าอำนาจจำแนกสูง แบบที่ 5 ค่าความยากสูงค่าอำนาจจำแนกต่ำ แบบที่ 6 ค่าความยากและค่าอำนาจจำแนกสูง แบบที่ 7 เป็นแบบผสมคือค่าอำนาจจำแนกและค่าความยากต่างกันสำหรับ 2 กลุ่ม

ผลการศึกษาพบว่าทั้งสองวิธีมีประสิทธิภาพเท่าเทียมกันในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบสม่ำเสมอ (Uniform DIF) แต่วิธี LR มีประสิทธิภาพมากกว่าในการตรวจสอบการทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ (Nonuniform DIF) และค้นพบว่าขนาดของกลุ่มตัวอย่างเป็นปัจจัยที่มีผลกระทบต่ออำนาจการตรวจสอบด้วยสองวิธีนี้ กล่าวคือเมื่อเพิ่มขนาดกลุ่มตัวอย่าง อำนาจการตรวจสอบจะเพิ่มขึ้น ส่วนความยาวของแบบทดสอบและการกระจายของคะแนนไม่มีผลกระทบต่ออำนาจการตรวจสอบ

ราชู และคณะ (Raju and others. 1993) ได้ศึกษาเปรียบเทียบผลการทดสอบไค-กำลังสองของลอว์รี การทดสอบ Z ของราชู (1990) ในการวัดพื้นที่ความแตกต่างระหว่างโค้งลักษณะข้อสอบทั้งพื้นที่ชนิดที่คิดเครื่องหมายและไม่คิดเครื่องหมาย โดยใช้โมเดลสองพารามิเตอร์ และวิธีไค-กำลังสองของ MH ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบใช้เกณฑ์การตัดสินความลำเอียงดังนี้คือ ค่า χ^2 ของลอว์รี มีค่าเท่ากับ 13.80 Z มีค่าอยู่ระหว่าง ± 8.30 และไค-กำลังสองของ MH มีค่าเท่ากับ 10.83 โดยใช้ข้อมูลจากการตอบแบบทดสอบ Gater-Mac Ginitie Reading Tests (GMRT) เฉพาะแบบทดสอบย่อยที่วัดคำศัพท์ 45 ข้อ เป็นแบบเลือกตอบชนิด 5 ตัวเลือก กลุ่มตัวอย่างที่ใช้เป็นนักเรียนระดับ 10 และ 12 ในปี 1987 จำนวน 839 คน เป็นนักเรียนผิวดำ 245 คน นักเรียนผิวขาว 436 คน นักเรียนหญิง 440 คน และนักเรียนชาย 399 คน

ผลการศึกษาพบว่าในการวิเคราะห์การทำหน้าที่ต่างกันระหว่างนักเรียนผิวดำและนักเรียนผิวขาว จากการทดสอบด้วย χ^2 และ Z ทั้งพื้นที่ชนิดคิดเครื่องหมายและไม่คิดเครื่องหมาย ทั้ง 3 วิธี พบข้อสอบที่มีทำหน้าที่ต่างกัน 1 ข้อ คือ ข้อ 41 ส่วนวิธีไค-กำลังสองของ MH พบข้อสอบที่ทำหน้าที่ต่างกัน 2 ข้อ คือ ข้อ 14 และ 27 ซึ่งข้อสอบทั้ง 3 ข้อเป็นข้อที่มีค่าความยากต่างกันมากระหว่างกลุ่มผู้สอบผิวดำและผิวขาว ส่วนผลการวิเคราะห์การทำหน้าที่ต่างกันระหว่างนักเรียนชายและนักเรียนหญิง 3 วิธีแรก พบข้อสอบที่ทำหน้าที่ต่างกันซ้ำกัน 4 ข้อ ส่วนวิธีไค-กำลังสองของ MH พบข้อสอบที่ทำหน้าที่ต่างกัน 5 ข้อ ซ้ำกับ 3 วิธีแรก 4 ข้อ คิดเป็นร้อยละ 80 ได้แก่ข้อ 2, 18, 23, 33 และข้อสอบที่ทำหน้าที่ต่างกันไม่ซ้ำกับวิธีแรก คือ ข้อ 41 และข้อสอบทุกข้อมีค่าความยากสูงกว่าข้อสอบที่ทำหน้าที่ไม่ต่างกัน นอกจากนี้ยังพบว่า วิธีไค-กำลังสองของ MH ค้นพบข้อสอบที่มีทำหน้าที่ต่างกันได้มากกว่าวิธีที่ใช้ทฤษฎีการตอบข้อสอบ ซึ่งค้านกับการค้นพบของสวามินาธานและโรเจอร์ส(1990)ที่พบว่าวิธีที่ใช้ทฤษฎีการตอบข้อสอบ สามารถค้นพบข้อสอบที่ทำหน้าที่ต่างกันได้มากกว่าวิธี MH

โคเฮนและคิม (Cohen and Kim. 1993) ได้ศึกษาเปรียบเทียบการใช้ทฤษฎีการตอบข้อสอบแบบสองพารามิเตอร์ โดยใช้การทดสอบ χ^2 ของลอว์รี และการทดสอบความแตกต่างของพื้นที่ระหว่างโค้งลักษณะข้อสอบชนิดคิดเครื่องหมายและไม่คิดเครื่องหมายโดยใช้สถิติ Z ของราชู จากข้อมูลจำลองขึ้นตามทฤษฎีการตอบข้อสอบแบบสองพารามิเตอร์ซึ่งจำลองข้อสอบ 2 ขนาด คือ 20 และ 60 ข้อ จำนวนผู้สอบ 2 ขนาด คือ 100 และ 500 คน และมีจำนวนข้อสอบที่ทำหน้าที่ต่างกันแบบทดสอบที่จำลอง 3 แบบ คือ ข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0, 10 และ 20

ผลการศึกษาพบว่า

1. แบบทดสอบที่ยาว การระบุข้อสอบทำหน้าที่ต่างกันผิดพลาด (Fault Positive) จะมากขึ้นและถ้ามีจำนวนข้อสอบที่ทำหน้าที่ต่างกันอยู่ในแบบทดสอบน้อย จำนวนข้อสอบที่ถูกระบุว่าทำหน้าที่ต่างกันผิดพลาดจะยิ่งมากขึ้นด้วย องค์ประกอบที่มีอิทธิพลต่อการระบุว่าข้อสอบทำหน้าที่ต่างกันผิดพลาด (Fault Positive) อีกประการหนึ่งคือ ระดับนัยสำคัญ (α) ถ้าระดับนัยสำคัญยิ่งมากจะเกิด Fault Positive มาก นั่นคือที่ $\alpha = .05$ จะเกิด Fault Positive มากกว่าที่ $\alpha = .01$ สำหรับการทดสอบ Z ของราชูเกิดในทุกสถานการณ์ที่ศึกษา แต่การทดสอบ χ^2 ของลอว์รีจะเกิดในบางสถานการณ์เท่านั้น

2. แบบทดสอบที่ยาวกว่า มีร้อยละของข้อสอบที่ทำหน้าที่ต่างกันมากกว่า จะมีข้อสอบที่ถูกระบุว่าทำหน้าที่ไม่ต่างกันผิดพลาด (Fault Negative) มากกว่า แต่จะเกิดขึ้นกับการทดสอบด้วย Z ของราชูมากกว่าการทดสอบด้วย χ^2 ของลอว์รี กรณีที่จำนวนผู้สอบเท่ากับ 100 คน พบว่าความแตกต่างในการระบุว่าข้อสอบทำหน้าที่ไม่ต่างกันผิดพลาด สำหรับ Z ของราชู และ χ^2 ของลอว์รี ไม่แตกต่างกันมากเท่ากับกรณีที่ใช้

กลุ่มตัวอย่าง 500 คน ซึ่งมีการระบุว่าทำหน้าที่ไม่ต่างกันผิดพลาด สำหรับ χ^2 ของลอร์ด น้อยกว่า Z ของราชู โดยเฉพาะอย่างยิ่งในกรณีที่มีข้อสอบที่ทำหน้าที่ต่างกันแบบทดสอบมากที่สุดคือร้อยละ 20

โคเฮนและคิม ได้ ให้ความเห็นว่า การทดสอบด้วย χ^2 ของลอร์ด ให้ผลการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันได้ดีกว่าการทดสอบ Z ของพื้นที่ที่คิดเครื่องหมาย อาจเป็นเพราะ Z(ESA) ใช้การเปรียบเทียบเฉพาะค่าความยาก (b) ในขณะที่ χ^2 ใช้ทั้งค่าความยากและค่าอำนาจจำแนก (b และ a) แต่ไม่สามารถอธิบายความแตกต่างได้ในกรณีใช้ Z(H) ซึ่งทดสอบพื้นที่ที่ไม่คิดเครื่องหมาย เพราะ Z(H) ก็ใช้การทดสอบค่าพารามิเตอร์ทั้ง 2 ตัวเหมือนกัน

นารายานานและสวามินาธาน (Narayanan and Swaminathan. 1994) ได้ศึกษาผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี MH กับ วิธี SIBTEST โดยใช้ข้อมูลจำลองความยาวของแบบทดสอบ 40 ข้อ ตัวแปรที่ศึกษาคือ (1) ขนาดของกลุ่มตัวอย่าง โดยกลุ่มอ้างอิงมี 3 ขนาด ได้แก่ 300, 500 และ 1,000 คน กลุ่มเปรียบเทียบ 3 ขนาด ได้แก่ 100, 200 และ 300 คน ซึ่งจะจับคู่ศึกษาได้ 9 เงื่อนไข (2) การกระจายของความสามารถ 2 แบบ (3) สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันที่มีภายในแบบทดสอบ 2 ขนาด (4) ขนาดของพื้นที่ระหว่างโค้งคุณลักษณะข้อสอบของผู้สอบสองกลุ่ม 4 ขนาด (5) ค่าความยากและค่าอำนาจจำแนกของแบบทดสอบ 6 ระดับ

ผลการศึกษาพบว่า ขนาดกลุ่มตัวอย่าง สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน ขนาดของพื้นที่ระหว่างโค้งคุณลักษณะข้อสอบ ค่าความยากและค่าอำนาจจำแนกเป็นตัวแปรที่มีผลกระทบต่ออำนาจการตรวจสอบของทั้งสองวิธีอย่างมีนัยสำคัญทางสถิติ วิธี MH และ วิธี SIBTEST มีประสิทธิภาพเท่ากันในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมมาตร เมื่อการกระจายความสามารถเท่ากันระหว่างกลุ่ม แต่เมื่อการกระจายความสามารถไม่เท่ากันระหว่างกลุ่ม วิธี SIBTEST จะมีประสิทธิภาพในการตรวจสอบมากกว่าวิธี MH จึงสรุปได้ว่าการกระจายความสามารถไม่มีผลกระทบต่อตรวจสอบด้วยวิธี SIBTEST แต่มีผลกระทบต่อวิธี MH ส่วนการจำแนกผิดพลาด วิธี MH จะมีระดับการเกิดอยู่ในระดับปกติ แต่วิธี SIBTEST จะเพิ่มขึ้นเล็กน้อยในกลุ่มที่มีความสามารถเท่ากัน ในกรณีที่มีการกระจายความสามารถต่างกันเพิ่มขึ้นระหว่างผู้สอบสองกลุ่มจะทำให้การจำแนกผิดพลาดเพิ่มขึ้น

เมเซอร์และคณะ (Mazor and others. 1994) ใช้วิธี MH ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สมมาตร กลุ่มตัวอย่างได้จากการจำลองกลุ่มละ 1,000 คน ใช้แบบทดสอบ 25 ฉบับ ๗๕ 75 ข้อ ในแต่ละฉบับมีข้อสอบที่ทำหน้าที่ไม่ต่างกัน (no-DIF) จำนวน 59 ข้อ และข้อสอบที่ศึกษาซึ่งทำหน้าที่ต่างกัน(DIF) จำนวน 16 ข้อ โดยข้อสอบที่ศึกษาจะแปรเปลี่ยนค่าพารามิเตอร์ดังนี้ ค่าอำนาจจำแนก 4 ระดับ ความแตกต่างของค่าอำนาจจำแนกระหว่างกลุ่ม 5 ระดับ ค่าความยาก 5 ระดับ ความแตกต่างของค่าความยากระหว่างกลุ่ม 4 ระดับ ค่าการเดากำหนดเป็น 0.2 และการกระจายของความสามารถระหว่างกลุ่ม 2 แบบ คือการกระจายความสามารถเท่ากันและไม่เท่ากัน ในการตรวจสอบจะวิเคราะห์สองแบบ คือ แบบแรกจะใช้คะแนนรวมของแต่ละคนเป็นเกณฑ์ในการจับคู่เปรียบเทียบระหว่างกลุ่ม ส่วนแบบที่สองจะแยกวิเคราะห์เฉพาะกลุ่มตัวอย่างที่เป็นกลุ่มสูงหรือกลุ่มต่ำ โดยใช้ค่าเฉลี่ยจากคะแนนของกลุ่มตัวอย่างทุกคนเป็นเกณฑ์ในการแยกเป็นกลุ่มสูงและกลุ่มต่ำ จากนั้นจึงนำเฉพาะกลุ่มสูงหรือกลุ่มต่ำที่ได้มาแบ่งเป็นกลุ่มอ้างอิงและกลุ่มเปรียบเทียบแล้ววิเคราะห์

ผลการศึกษาพบว่า การวิเคราะห์โดยการแบ่งกลุ่มตัวอย่างเป็นกลุ่มสูงกลุ่มต่ำ จะทำให้อำนาจการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบไม่สม่ำเสมอสูงกว่าการวิเคราะห์แบบแรก (รวมผู้สอบกลุ่มสูงและกลุ่มต่ำไว้ด้วยกัน) อีกทั้งไม่ทำให้การจำแนกผิดพลาดเพิ่มขึ้น และพบว่าเมื่อมีความแตกต่างของค่าอำนาจจำแนกและค่าความยากระหว่างกลุ่มเพิ่มขึ้นจะทำให้อำนาจการตรวจสอบพบข้อสอบที่ทำหน้าที่ต่างกันได้มากขึ้น

นารายานานและสวามินาธาน (Narayanan and Swaminathan, 1996) ได้ศึกษาเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบไม่สม่ำเสมอ ด้วยวิธี MH วิธี LR และวิธี SIBTEST โดยศึกษาอำนาจการตรวจสอบและการจำแนกผิดพลาด โดยการจำลองข้อมูล ตัวแปรที่ศึกษาได้แก่ขนาดกลุ่มตัวอย่าง ความแตกต่างของการกระจายความสามารถระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันที่มีภายในแบบทดสอบ ขนาดของพื้นที่ระหว่างโค้งคุณลักษณะข้อสอบของผู้สอบสองกลุ่ม ค่าความยากและค่าอำนาจจำแนกของแบบทดสอบ

ผลการศึกษาพบว่า วิธี SIBTEST และวิธี LR มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบไม่สม่ำเสมอได้เท่าเทียมกันในทุกเงื่อนไขที่ศึกษา ส่วนวิธี MH ไม่สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ และการจำแนกผิดพลาด วิธี MH จะสูงกว่า วิธี SIBTEST และวิธี LR

จากรายงานการวิจัยทั้งภายในประเทศและต่างประเทศข้างต้น จะพบว่าการเปรียบเทียบวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบนิยมใช้วิธี MH วิธี SIBTEST วิธี LR เป็นส่วนใหญ่ทั้งการศึกษาจากข้อมูลจริงและข้อมูลจำลอง ซึ่งจะพบว่าวิธี MH และวิธี SIBTEST จะมีประสิทธิภาพเท่าเทียมกัน แต่วิธี MH กับวิธี LR ยังไม่พบข้อสรุปที่แน่นอนทั้งการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอและแบบไม่สม่ำเสมอ นอกจากนี้รายงานการวิจัยข้างต้นยังพบว่าปัจจัยที่มีผลกระทบต่ออำนาจการตรวจสอบคือขนาดกลุ่มตัวอย่าง ซึ่งข้อค้นพบที่ได้จะสอดคล้องกัน กล่าวคือเมื่อขนาดกลุ่มตัวอย่างมากขึ้น อำนาจการตรวจสอบพบ DIF จะเพิ่มขึ้น ซึ่งนารายานานและสวามินาธาน (Narayanan and Swaminathan, 1994) แนะนำว่าโดยทั่วไปใช้ขนาดกลุ่มตัวอย่างกลุ่มละ 300 คน ก็เพียงพอที่จะตรวจสอบอย่างมีประสิทธิภาพและพบว่าอำนาจการตรวจสอบของ MH จะได้รับผลกระทบจากกลุ่มตัวอย่างขนาดเล็กของกลุ่มเปรียบเทียบมากกว่ากลุ่มตัวอย่างขนาดใหญ่ของกลุ่มอ้างอิง จึงอาจกล่าวได้ว่าสัดส่วนระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ น่าจะมีผลกระทบต่ออำนาจการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ดังนั้นจึงทำให้ผู้วิจัยต้องการศึกษาเกี่ยวกับขนาดกลุ่มตัวอย่างที่เป็นสัดส่วนต่อกันระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ทั้งนี้เพราะในทางปฏิบัติจะพบว่าขนาดกลุ่มตัวอย่างทั้งสองกลุ่มมักจะไม่เท่ากัน อีกทั้งยังไม่มีผลงานการวิจัยที่ระบุได้ชัดเจนว่าควรจะใช้กลุ่มตัวอย่างที่เป็นสัดส่วนต่อกันเท่าใด นอกจากนี้โรเจอร์สและสวามินาธาน (Rogers and Swaminathan, 1993) พบว่าความยาวของแบบทดสอบไม่มีผลกระทบต่ออำนาจการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ในขณะที่โคเฮนและคิม (Cohen and Kim, 1993) พบว่าความยาวของแบบทดสอบมีผลกระทบต่ออำนาจการตรวจสอบ อีกทั้งจิตติมา วรรณศรี (2540 : 90 – 99) พบว่า วิธี MH มีประสิทธิภาพในการตรวจสอบได้ดีที่สุดที่ขนาดความยาวของแบบทดสอบปานกลาง ดังนั้นเพื่อเป็นการยืนยันข้อค้นพบให้มีความชัดเจนมากยิ่งขึ้น จึงควรศึกษาเกี่ยวกับความยาวของแบบทดสอบว่าจะมีผลต่ออำนาจการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบอย่างไร หรือไม่ จากการศึกษารายงานการวิจัยดังกล่าว ผู้วิจัยจึงกำหนดสมมติฐานในการวิจัยดังนี้

บทที่ 3

วิธีดำเนินการศึกษาค้นคว้า

การศึกษาค้นคว้าครั้งนี้ มีความมุ่งหมายเพื่อเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอและแบบไม่สมำเสมอ ระหว่างวิธี MH และวิธี LR โดยใช้ความยาวแบบทดสอบและขนาดกลุ่มตัวอย่างต่างกัน ซึ่งผู้วิจัยได้ดำเนินการศึกษาค้นคว้าดังต่อไปนี้

ประชากร

ประชากรในการศึกษาค้นคว้าครั้งนี้ เป็นนักเรียนชายและหญิง ชั้นมัธยมศึกษาปีที่ 3 ภาคเรียนที่ 2 ปีการศึกษา 2542 ของโรงเรียนมัธยมศึกษาสังกัดกรมสามัญศึกษา ส่วนกลาง กลุ่มที่ 3 จำนวน 12 โรงเรียน จำนวน 164 ห้องเรียน จำนวนนักเรียนทั้งหมด 7,058 คน จำแนกเป็นนักเรียนชาย 2,823 คน นักเรียนหญิง 4,235 คน

กลุ่มตัวอย่าง

กลุ่มตัวอย่างที่ใช้ในการศึกษาค้นคว้าครั้งนี้ เป็นนักเรียนชายและหญิงชั้นมัธยมศึกษาปีที่ 3 ภาคเรียนที่ 2 ปีการศึกษา 2542 ของโรงเรียนมัธยมศึกษาสังกัดกรมสามัญศึกษา ส่วนกลาง กลุ่มที่ 3 จำนวน 8 โรงเรียน จำนวนนักเรียน 3,242 คน จำแนกเป็นนักเรียนชาย 1,282 คน นักเรียนหญิง 1,960 คน โดยแบ่งเป็นกลุ่มตัวอย่างที่ใช้ในการทดลองเครื่องมือ และกลุ่มตัวอย่างที่ใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ซึ่งได้มาโดยวิธีการสุ่มแบบแบ่งชั้น (Stratified random sampling) มีขนาดโรงเรียนเป็นชั้น (strata) โรงเรียนเป็นหน่วยการสุ่ม (Sampling unit) ซึ่งมีลำดับขั้นตอนในการสุ่มดังนี้

1. กำหนดขนาดกลุ่มตัวอย่างทั้งหมดด้วยความเชื่อมั่น 95% ($\alpha = .05$) ซึ่งเทียบจากตารางขนาดกลุ่มตัวอย่างของ Yamane (ล้วน สายยศ และ อังคณา สายยศ, 2536 : 260 อ้างอิงจาก Yamane, 1967) ปรากฏว่าต้องใช้ขนาดกลุ่มตัวอย่างอย่างน้อย 364 คน แต่เพื่อสะดวกในการศึกษาค้นคว้าครั้งนี้ ผู้วิจัยใช้ขนาดกลุ่มตัวอย่างอย่างน้อย 400 คน สำหรับขนาดกลุ่มตัวอย่างที่ใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี MH แฮมเบิลตันและคณะ (Hambleton and others, 1989) ได้แนะนำว่าควรใช้กลุ่มผู้สอบระหว่าง 200 ถึง 1,000 คน และกลุ่มตัวอย่างที่ตรวจสอบด้วยวิธีตามทฤษฎีการตอบข้อสอบ (IRT) คิม และคณะ (Kim and others, 1994) กล่าวว่าถ้าวิเคราะห์โดยใช้โมเดล 2 พารามิเตอร์ กลุ่มตัวอย่างที่ให้ผลการประมาณค่าดีที่สุดคือ 1,000 คนขึ้นไป ดังนั้นเพื่อให้กลุ่มตัวอย่างทั้ง 2 กลุ่ม มีความเหมาะสมกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี MH วิธี LR และวิธีตามทฤษฎีการตอบข้อสอบ (IRT) ผู้วิจัยจึงดำเนินการสุ่มตัวอย่างดังนี้

ขั้นที่ 1 สุ่มโรงเรียนจากแต่ละขนาดของโรงเรียนด้วยวิธีการสุ่มแบบแบ่งชั้น (Stratified random sampling) มีขนาดของโรงเรียนเป็นชั้น (Strata) คือโรงเรียนขนาดใหญ่พิเศษ (พิจารณาจากจำนวนนักเรียนทั้งหมดมากกว่า 2,500 คน) โรงเรียนขนาดใหญ่ (พิจารณาจากจำนวนนักเรียนทั้งหมดมากกว่า 1,500 แต่ไม่น้อยกว่า 2,500 คน) โรงเรียนขนาดกลาง (พิจารณาจากจำนวนนักเรียนทั้งหมดมากกว่า 500 แต่ไม่น้อยกว่า 1,500 คน) และมีโรงเรียนเป็นหน่วยการสุ่ม (Sampling unit) ได้โรงเรียนขนาดใหญ่พิเศษ 6 โรงเรียน โรงเรียนขนาดใหญ่ 1 โรงเรียน และโรงเรียนขนาดกลาง 1 โรงเรียน

ขั้นที่ 2 สุ่มห้องเรียนจากโรงเรียนที่สุ่มได้ในแต่ละขนาด โดยวิธีสุ่มอย่างง่าย (Simple random sampling) มีห้องเรียนเป็นหน่วยการสุ่ม ได้จำนวนห้องเรียน 71 ห้องเรียน แบ่งเป็นกลุ่มตัวอย่างที่ใช้ในการทดลองเครื่องมือ จำนวน 12 ห้องเรียน จำนวนนักเรียน 528 คน และกลุ่มตัวอย่างที่ใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจำนวน 59 ห้องเรียน จำนวนนักเรียน 2,714 คน จำแนกเป็นนักเรียนชาย 1,070 คน นักเรียนหญิง 1,644 คน ดังตาราง 4

ตาราง 4 แสดงรายชื่อโรงเรียน จำนวนห้องเรียนและจำนวนนักเรียนที่เป็นกลุ่มตัวอย่าง

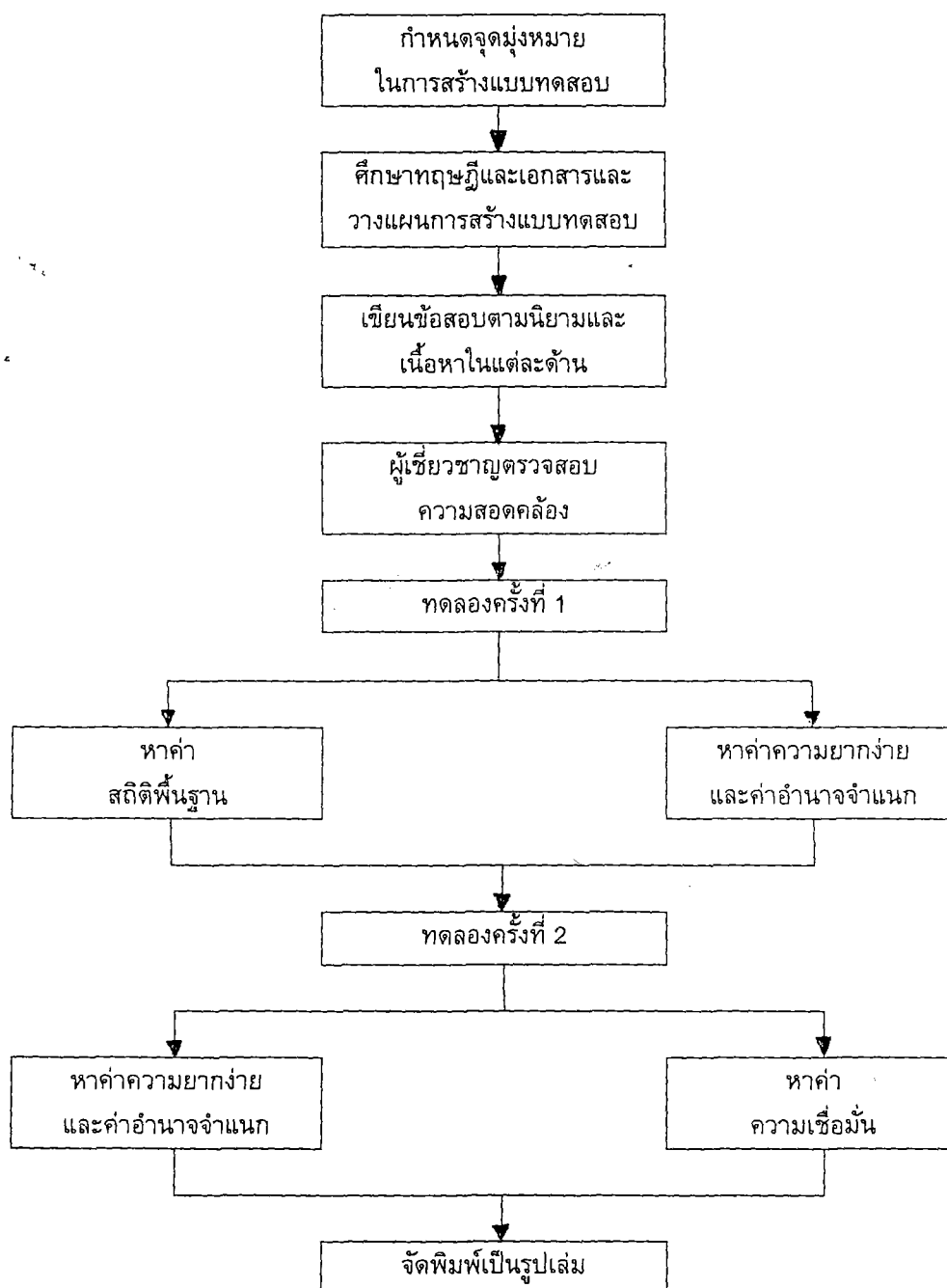
ขนาด	โรงเรียน	จำนวนนักเรียนที่ใช้ ในการทดลองเครื่องมือ			จำนวนนักเรียนที่ใช้ ตรวจสอบการทำหน้าที่ ต่างกันของข้อสอบ		
		รวม(ห้อง)	ชาย	หญิง	รวม(ห้อง)	ชาย	หญิง
ใหญ่พิเศษ	ศรีอยุธยา	528(12)	212	316			
	สามเสนวิทยาลัย				392(8)	157	235
	สันติราษฎร์วิทยาลัย				376(8)	151	225
	หอวัง				577(10)	231	346
	สารวิทยา				368(8)	147	221
	ฤทธิยะวรรณาลัย				516(12)	206	310
ใหญ่	ราชวินิตบางเขน				325(8)	114	211
กลาง	มักกะสันพิทยา				160(5)	64	96
	รวม	528(12)	212	316	2,714(59)	1,070	1,644

เครื่องมือที่ใช้ในการศึกษาค้นคว้า

เครื่องมือที่ใช้ในการศึกษาครั้งนี้ได้แก่แบบทดสอบความสามารถทางสมองที่ผู้วิจัยสร้างขึ้นโดยพัฒนาและปรับปรุงตามแนวโครงสร้างของแบบทดสอบความสามารถทางสมองของโอติส-เลนนอน ฟอรัม เจ ระดับสูง (Otis-Lennon Mental Ability Test Advanced Level Form J) จำนวน 1 ฉบับ เป็นแบบทดสอบที่วัดความสามารถทั่วไปซึ่งแบ่งออกเป็นสามด้าน คือความเข้าใจด้านภาษา 30 ข้อ เหตุผลด้านภาษา 30 ข้อ และเหตุผลด้านภาพ 20 ข้อ รวม 80 ข้อ เป็นแบบทดสอบปรนัยแบบเลือกตอบ 5 ตัวเลือก ใช้เวลาในการทำ 40 นาที

วิธีดำเนินการสร้างเครื่องมือที่ใช้ในการศึกษา

วิธีดำเนินการสร้างแบบทดสอบความสามารถทางสมองสำหรับนักเรียนระดับชั้นมัธยมศึกษาปีที่ 3
ผู้วิจัยดำเนินการตามขั้นตอน ดังแสดงในภาพประกอบ 3



ภาพประกอบ 3 ลำดับขั้นตอนในการดำเนินการสร้างแบบทดสอบวัดความสามารถทางสมอง

ผู้วิจัยได้ดำเนินการสร้างเครื่องมือสำหรับใช้ในงานวิจัยตามภาพประกอบข้างต้น ดังมี รายละเอียด ดังนี้

1. กำหนดจุดมุ่งหมายในการสร้างแบบทดสอบ
2. ศึกษาทฤษฎีเอกสารและงานวิจัยที่เกี่ยวข้องเพื่อเป็นแนวทางในการสร้างแบบทดสอบ
3. สร้างแบบทดสอบวัดความสามารถทางสมองตามแนวโครงสร้างของแบบทดสอบทางสมองของโอติส-เลนนอน ฟอรั่ม เจ ระดับสูง 1 ฉบับ โดยผู้วิจัยได้ปรับปรุงโดยตัดเนื้อหาด้านเหตุผลด้านปริมาณซึ่งวัดความสามารถทางสมองด้านคณิตศาสตร์ออก เนื่องจากตัวแปรที่จำแนกกลุ่มตัวอย่างที่ใช้ในการวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกัน คือ ตัวแปรเพศ ซึ่งผู้วิจัยได้กำหนดให้กลุ่มอ้างอิงได้แก่นักเรียนหญิงเพราะมีความถนัดด้านภาษามากกว่านักเรียนชาย ดังนั้นแบบทดสอบที่สร้างขึ้นจึงประกอบด้วยเนื้อหาต่าง ๆ รวมสามด้านคือความเข้าใจทางด้านภาษา 40 ข้อ เหตุผลทางภาษา 40 ข้อ และเหตุผลด้านภาพ 30 ข้อ รวมจำนวนข้อสอบทั้งหมด 110 ข้อ
4. นำแบบทดสอบที่สร้างในข้อ 3 ไปให้ผู้เชี่ยวชาญจำนวน 5 ท่าน ตรวจสอบความสอดคล้องของข้อคำถามของแบบทดสอบ
5. ปรับปรุงและแก้ไขข้อสอบแล้วจัดพิมพ์โดยเรียงข้อสอบตามเนื้อหาเป็นด้าน ๆ ดังนี้คือ ความเข้าใจด้านภาษา (ข้อ 1 – 40) เหตุผลด้านภาษา (ข้อ 41 – 80) และเหตุผลด้านภาพ (ข้อ 81 – 110)
6. นำแบบทดสอบที่สร้างขึ้นไปทดสอบครั้งที่ 1 กับนักเรียนกลุ่มตัวอย่าง ระดับชั้นมัธยมศึกษาปีที่ 3 โรงเรียนศรีอยุธยา จำนวน 6 ห้องเรียน จำนวนนักเรียน 264 คน
7. ทำการวิเคราะห์ข้อสอบรายข้อเพื่อหาค่าความยากง่ายและค่าอำนาจจำแนก โดยใช้โปรแกรม Item Analysis แล้วคัดเลือกข้อสอบที่มีค่าอำนาจจำแนกตั้งแต่ .20 ขึ้นไป และมีค่าความยากง่ายในช่วง .20 ถึง .80 ไว้จำนวน 80 ข้อ ได้ข้อสอบความเข้าใจด้านภาษา 30 ข้อ เหตุผลด้านภาษา 30 ข้อ และเหตุผลด้านภาพ 20 ข้อ
8. นำข้อสอบที่ได้ในข้อที่ 7 จำนวน 80 ข้อไปทดสอบครั้งที่ 2 กับนักเรียนกลุ่มตัวอย่าง ระดับชั้นมัธยมศึกษาปีที่ 3 โรงเรียนศรีอยุธยา จำนวน 6 ห้องเรียน จำนวนนักเรียน 264 คน เพื่อหาค่าความยากง่าย ค่าอำนาจจำแนกและค่าความเชื่อมั่น โดยใช้โปรแกรม Item Analysis
9. จัดพิมพ์แบบทดสอบฉบับที่สมบูรณ์ เพื่อนำไปใช้ในการวิจัยต่อไป

ลักษณะเนื้อหาของข้อสอบ

แบบทดสอบความสามารถทางสมองฉบับที่ผู้วิจัยสร้างขึ้นเป็นแบบทดสอบวัดความสามารถทางสมองตามแนวโครงสร้างของโอติส-เลนนอน ฟอรั่ม เจ ซึ่งแบ่งลักษณะเนื้อหาของแบบทดสอบออกเป็นสามด้านคือ

1. การวัดความเข้าใจด้านภาษา (Verbal Comprehension) ประกอบด้วยแบบทดสอบชนิดต่างๆ ดังนี้คือ ชนิดความหมายเหมือนกัน ชนิดคำตรงข้าม ชนิดประโยคสมมูล และชนิดกระจายประโยค
 - 1.1 ชนิดความหมายเหมือนกัน ข้อสอบแต่ละข้อจะกำหนดคำมาให้แล้วให้นักเรียนหาคำที่มีความหมายเหมือนกับคำที่กำหนดหรือคำที่ขีดเส้นใต้ในประโยค ตัวอย่างเช่น

ข้อ 1. ซี้ฮ่าย มีความหมายเหมือนคำใด

- | | | |
|------------|------------|------------|
| ก. กระดาก | ข. ขวยเขิน | ค. ประหม่า |
| ง. ซี้ซลัด | จ. อ่อนแอ | |

เฉลย (ข)

1.2 ชนิดคำตรงข้าม ข้อสอบแต่ละข้อจะกำหนดคำมาให้แล้วให้นักเรียนหาคำที่มีความหมายตรงข้ามกับคำที่กำหนด หรือคำที่ขีดเส้นใต้ในประโยค ตัวอย่างเช่น

ข้อ 2. “ปากหนัก” มีความหมายตรงข้ามกับคำใด

- | | | |
|-------------|-------------|-----------|
| ก. ปากจัด | ข. ปากโป้ง | ค. ปากบอน |
| ง. ปากสว่าง | จ. ปากเปราะ | |

เฉลย (จ)

1.3 ชนิดประโยคสมบูรณ์ ข้อสอบแต่ละข้อจะกำหนดข้อความมาให้ โดยข้อความนั้นจะเว้นช่องว่างไว้ให้แล้วให้นักเรียนเติมคำที่เหมาะสมที่สุดลงในช่องว่าง เพื่อให้ประโยคสมบูรณ์ที่สุด ตัวอย่างเช่น

ข้อ 3. การทำมาหากินในระยะนี้ดูช่าง.....ไปเสียทุกอย่าง

- | | | |
|-----------|-------------|-------------|
| ก. ลำบาก | ข. กันดาร | ค. ผิดเคือง |
| ง. อัดคัด | จ. แห้งแล้ง | |

เฉลย (ค)

1.4 ชนิดกระจายประโยค ข้อสอบแต่ละข้อจะกำหนดข้อความมาให้ แล้วให้จัดเรียงเป็นประโยคที่สมบูรณ์ และสละสลวยที่สุด ตัวอย่างเช่น

ข้อ 4. ให้พิจารณาบทกลอนที่กำหนดให้ในแต่ละข้อต่อไปนี้แล้วเรียงเป็นคำกลอนที่สมบูรณ์ที่สุด

- 1) การอบรมบ่มนิสัยให้ศิษย์นั้น
- 2) เพียงแค่วันละนาทีกี่ดีถม
- 3) อาจเป็นเพราะอบรมหนึ่งนาทีกี่
- 4) ศิษย์จะดีมีชื่อหรือลุ่มจม

- | | | |
|------------|------------|------------|
| ก. 1 3 2 4 | ข. 1 2 4 3 | ค. 4 1 2 3 |
| ง. 3 1 2 4 | จ. 4 3 2 1 | |

เฉลย (ข)

2. การวัดเหตุผลด้านภาษา (Verbal Reasoning) ประกอบด้วยแบบทดสอบชนิดต่าง ๆ ดังนี้ คือ ชนิดอักษรไขว้ในตาราง ชนิดอุปมาอุปไมยด้านภาษา ชนิดสรุปความ และชนิดการเลือกแบบตรรกวิทยา

2.1 ชนิดอักษรไขว้ในตาราง ข้อสอบแต่ละข้อจะกำหนดอนุกรมตัวอักษรให้และให้ดูความหมายที่สอดคล้องกับอนุกรมที่กำหนด และเติมชุดตัวอักษรที่เป็นไปได้ ตัวอย่างเช่น

ข้อ 5. ตัวอักษรที่อยู่ในกรอบสี่เหลี่ยม จะเรียงตามกฎเกณฑ์อย่างใดอย่างหนึ่งให้พิจารณาว่าเครื่องหมายคำถาม(?) คือคำใด

สาบ	?
บอน	นอบ
กาบ	บาก
ดาน	นาด

ก. ปาส

ข. นาก

ค. ดอน

ง. สาก

จ. บาน

เฉลย (ก)

2.2 ชนิดอุปมาอุปไมยด้านภาษา ข้อสอบแต่ละข้อจะเป็นการหาสิ่งที่คู่กันคู่หนึ่ง แล้วนำไปเปรียบเทียบกับคู่อื่นๆที่มีลักษณะความสัมพันธ์เป็นแนวเดียวกัน ตัวอย่างเช่น

ข้อ 6. กันยายน : ตุลาคม → อังคาร : ?

ก. อาทิตย์

ข. จันทร์

ค. พุธ

ด. พฤหัสบดี

จ. ศุกร์

เฉลย (ค)

2.3 ชนิดสรุปความ ข้อสอบแต่ละข้อจะมีข้อความทั้งหมดมาให้แล้วให้สรุปความจากข้อความนั้นๆ ตัวอย่างเช่น

ข้อ 7. สมชายนั่งอยู่ระหว่างสมศรีกับสมศักดิ์

สมศรีนั่งอยู่ระหว่างสมชายกับสมหญิง

สมหญิงนั่งอยู่ระหว่างสมศรีกับสมปอง

ถ้าทั้ง 5 คนนั่งอยู่ในแถวเดียวกัน

ใครนั่งอยู่ตรงกลาง

ก. สมชาย

ข. สมศรี

ค. สมศักดิ์

ง. สมหญิง

จ. สมปอง

เฉลย (ข)

2.4 ชนิดการเลือกแบบตรรกวิทยา ข้อสอบแต่ละข้อจะกำหนดข้อความมาให้แล้วให้หาความสมเหตุสมผลจากข้อความนั้น ตัวอย่างเช่น

ข้อ 8. ดำและแดงเป็นพ่อลูกกัน

แดงเป็นพ่อ

ฉะนั้นสรุปได้ว่าอย่างไร

ก. แดงมีลูกชาย

ข. แดงมีลูกสาว

ค. แดงมีลูกคนเดียว

ง. ดำเป็นลูกของแดง

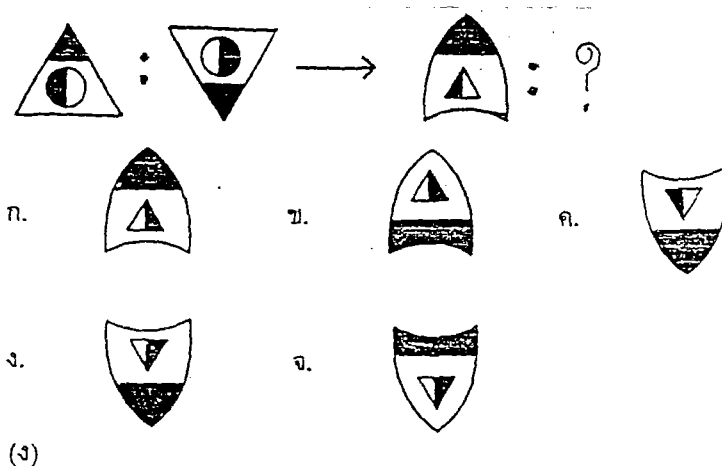
จ. ยังสรุปแน่นอนไม่ได้

เฉลย (ง)

3. การวัดเหตุผลด้านภาพ (Figural Reasoning) ประกอบด้วยแบบทดสอบชนิดต่าง ๆ ดังนี้ คือ ชนิดอุปมาอุปไมยภาพ ชนิดอนุกรมภาพ และชนิดตารางสัมพันธ์

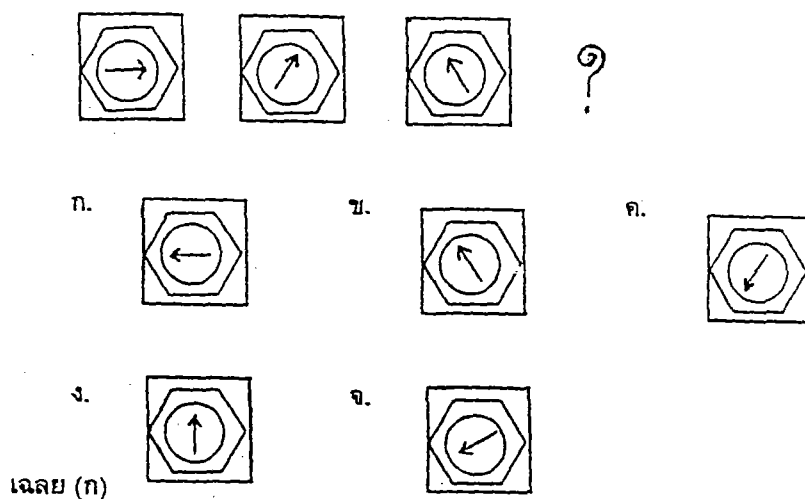
3.1 ชนิดอุปมาอุปไมยภาพ ข้อสอบแต่ละข้อจะประกอบไปด้วยภาพมาให้ 2 หรือ 3 ภาพ โดยภาพแรกจะมีความสัมพันธ์อย่างใดอย่างหนึ่งกับภาพที่สอง แล้วกำหนดภาพที่สามให้หาภาพที่สี่ ที่มีความสัมพันธ์เหมือนกับความสัมพันธ์ของภาพแรกกับภาพที่สอง ตัวอย่างเช่น

ข้อ 9. จากภาพ 2 ภาพซึ่งมีความสัมพันธ์กันอย่างใดอย่างหนึ่ง ให้พิจารณาว่าภาพที่กำหนดให้จะมีความสัมพันธ์กับภาพใด?



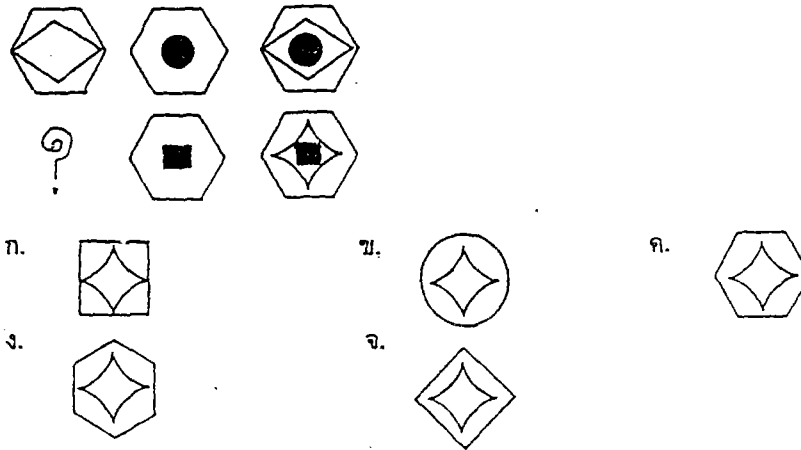
3.2 ชนิดอนุกรมภาพ ข้อสอบแต่ละข้อหาลำดับความเปลี่ยนแปลงอย่างมีระบบของภาพที่กำหนดให้ แล้วดูความสัมพันธ์ที่มีลักษณะทิศทางเดียวกัน ตัวอย่างเช่น

ข้อ 10. จากภาพ 3-4 ภาพ ให้ศึกษาระบบการเปลี่ยนแปลงของภาพที่กำหนดให้ แล้วพิจารณาว่า ภาพต่อไป ควรเป็นภาพใด?



3.3 แบบทดสอบภาพตารางสัมพันธ์ ข้อสอบแต่ละข้อหาลำดับความเปลี่ยนแปลงอย่างมีระบบของภาพทั้งแนวนอนและแนวตั้งในตารางที่กำหนดให้และให้หาภาพที่ขาดหายไป ตัวอย่างเช่น

ข้อ 11. รูปภาพในกรอบสี่เหลี่ยมที่กำหนดให้มีความสัมพันธ์กันอย่างไรอย่างหนึ่งทั้งแนวนอนและแนวตั้ง ให้นักเรียนศึกษาและพิจารณาว่ารูปที่หายไปควรเป็นรูปใด?



เฉลย (ค)

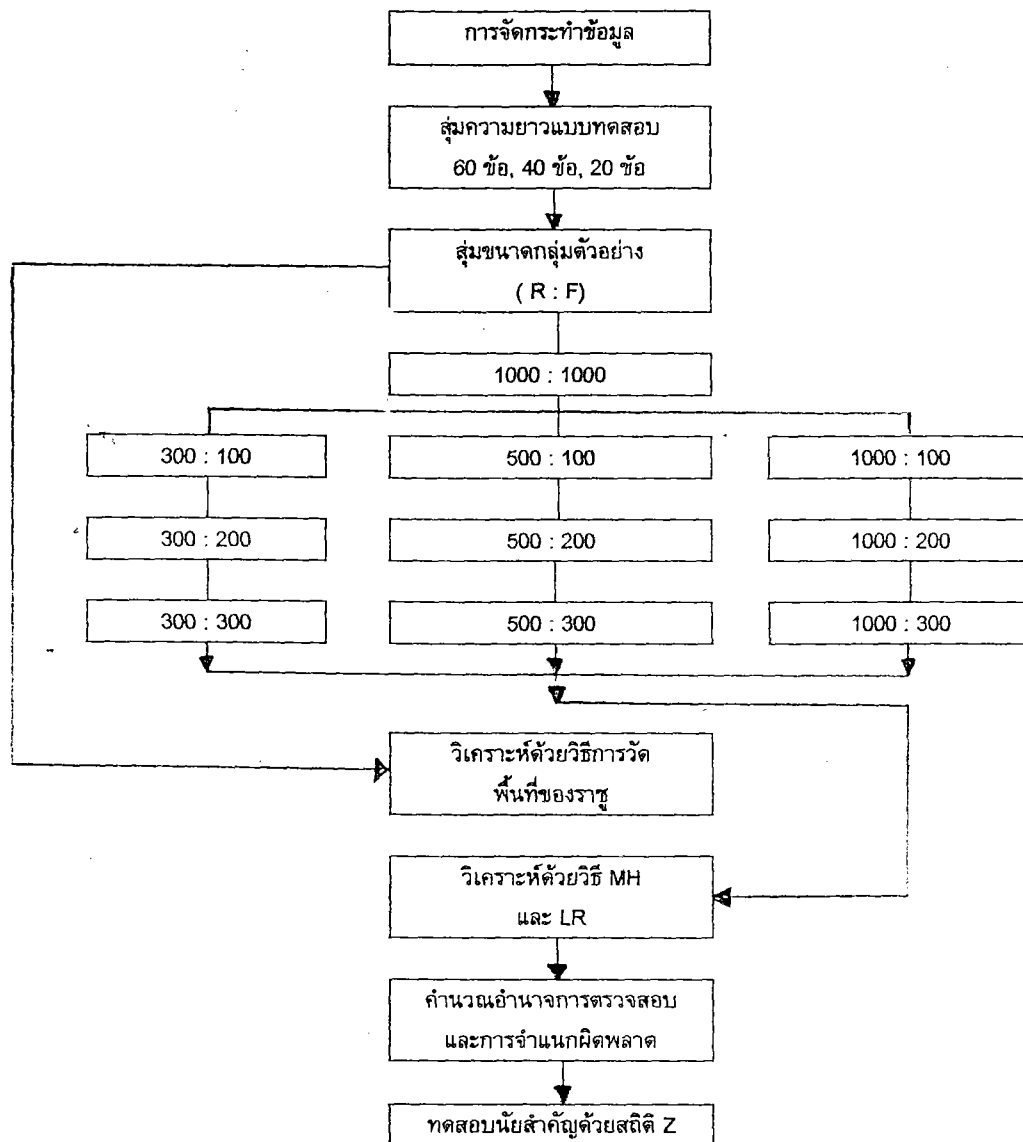
การเก็บรวบรวมข้อมูล

ผู้วิจัยได้ดำเนินการ ดังนี้

1. ขออนุญาตจากบัณฑิตวิทยาลัยมหาวิทยาลัยศรีนครินทรวิโรฒเพื่อขอความอนุเคราะห์ในการเก็บรวบรวมข้อมูล โดยทำหนังสือถึงอธิบดีกรมสามัญศึกษา เพื่อขอความร่วมมือจากโรงเรียนในสังกัด
2. ติดต่อประสานงานกับโรงเรียนที่เป็นกลุ่มตัวอย่าง เพื่อบรรยายและกำหนดวัน เวลา ในการนำแบบทดสอบ ไปทดสอบกับนักเรียนที่เป็นกลุ่มตัวอย่าง
3. ดำเนินการสอบ โดยนำแบบทดสอบ ไปทำการทดสอบกับนักเรียนที่เป็นกลุ่มตัวอย่าง ในช่วงวันที่ 1 มกราคม 2543 ถึง 31 มกราคม 2543
4. นำผลการตอบข้อสอบในแบบทดสอบ มาตรวจดูความเรียบร้อยแล้วลงรหัสผลการตอบข้อสอบทุกข้อ (ตอบถูกได้ 1 และตอบผิดได้ 0) ของนักเรียนที่เป็นกลุ่มตัวอย่างทุกคน
5. นำข้อมูลที่ได้มาวิเคราะห์ทางสถิติและทดสอบสมมติฐานที่ตั้งไว้

การวิเคราะห์ข้อมูล

ในการดำเนินการวิเคราะห์ข้อมูล ผู้วิจัยได้ดำเนินการตามขั้นตอน ดังแสดงในภาพประกอบ 4



ภาพประกอบ 4 แสดงขั้นตอนการวิเคราะห์ข้อมูล

1. การจัดกระทำข้อมูล

นำผลการตอบข้อสอบจากผู้สอบจำนวน 2,214 คน (เพศชาย 1,070 คน และเพศหญิง 1,144) ซึ่งตอบข้อสอบจำนวน 80 ข้อ มาแยกเป็นกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ โดยสุ่มกระดานคำตอบมากลุ่มละ 1,000 คน เพื่อจัดทำแฟ้มข้อมูล 2 แฟ้ม

2. การสุ่มความยาวของแบบทดสอบ

เป็นการเตรียมข้อมูลเพื่อให้ได้ตามเงื่อนไขที่กำหนดด้วยวิธีสุ่มอย่างง่าย จากข้อมูลที่ได้เตรียมไว้ในข้อ 1 โดยทำการสุ่ม 3 ครั้ง ดังนี้

สุ่มครั้งที่ 1 เป็นการสุ่มข้อสอบ จำนวน 60 ข้อ จากข้อสอบทั้งหมด 80 ข้อ

สุ่มครั้งที่ 2 เป็นการสุ่มข้อสอบ จำนวน 40 ข้อ จากข้อสอบทั้งหมด 60 ข้อ

สุ่มครั้งที่ 3 เป็นการสุ่มข้อสอบ จำนวน 20 ข้อ จากข้อสอบทั้งหมด 40 ข้อ

ดังนั้นจะได้ข้อมูล 3 เงื่อนไข ดังนี้

2.1 ความยาวของแบบทดสอบ 60 ข้อ ขนาดกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบเท่ากับ 1000 : 1000

2.2 ความยาวของแบบทดสอบ 40 ข้อ ขนาดกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบเท่ากับ 1000 : 1000

2.3 ความยาวของแบบทดสอบ 20 ข้อ ขนาดกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบเท่ากับ 1000 : 1000

3. การสุ่มขนาดกลุ่มตัวอย่าง

เป็นการสุ่มขนาดกลุ่มตัวอย่าง ซึ่งผู้วิจัยใช้วิธีการสุ่มอย่างง่ายโดยสุ่มจากข้อมูลที่ได้ในข้อ 2 เพื่อให้ได้จำนวนผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบตามเงื่อนไขที่จะศึกษาทั้งหมด 30 เงื่อนไข (3 ขนาดความยาวแบบทดสอบ $\times$ 10 ขนาดกลุ่มตัวอย่าง) ดังรายละเอียด ตามตาราง 5

ตาราง 5 ความยาวของแบบทดสอบและขนาดกลุ่มตัวอย่างที่ศึกษา

แบบทดสอบความยาว	กลุ่มตัวอย่างขนาด		
	กลุ่มอ้างอิง : กลุ่มเปรียบเทียบ		
20	300 : 100	300 : 200	300 : 300
	500 : 100	500 : 200	500 : 300
	1000 : 100	1000 : 200	1000 : 300
	1000 : 1000		
40	300 : 100	300 : 200	300 : 300
	500 : 100	500 : 200	500 : 300
	1000 : 100	1000 : 200	1000 : 300
	1000 : 1000		
60	300 : 100	300 : 200	300 : 300
	500 : 100	500 : 200	500 : 300
	1000 : 100	1000 : 200	1000 : 300
	1000 : 1000		

4. การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการวัดพื้นที่ของราชู

ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ผู้วิจัยใช้วิธีวิเคราะห์ทั้งหมด 3 วิธี คือ วิธี การวัดพื้นที่ของราชู วิธี MH และวิธี LR โดยจะใช้วิธี MH และวิธี LR วิเคราะห์ในทุกเงื่อนไขเพื่อเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาด แต่วิธีการวัดพื้นที่ของราชูจะใช้เป็นวิธีมาตรฐานที่จะระบุว่าข้อสอบข้อใดทำหน้าที่ต่างกันเพื่อจะใช้เป็นเกณฑ์ในการประเมินประสิทธิภาพในการตรวจสอบด้วยวิธี MH กับวิธี LR โดยวิธีการวัดพื้นที่ของราชูจะวิเคราะห์ในเงื่อนไขที่กลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ มีขนาด 1,000 คน : 1,000 คน ในแบบทดสอบทั้ง 3 ขนาด คือ 20 ข้อ, 40 ข้อ และ 60 ข้อ ทั้งนี้เพราะว่าวิธีการวัดพื้นที่ของ

ราชู ตามทฤษฎีการตอบข้อสอบจะตรวจสอบได้อย่างมีประสิทธิภาพเมื่อใช้กลุ่มตัวอย่างที่มีขนาดใหญ่ ดังนั้นจึงวิเคราะห์ด้วยวิธีนี้เพียง 3 เงื่อนไข คือ

เงื่อนไขที่ 1 แบบทดสอบยาว 60 ข้อ ขนาดกลุ่มตัวอย่างระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ เป็น 1,000 : 1,000

เงื่อนไขที่ 2 แบบทดสอบยาว 40 ข้อ ขนาดกลุ่มตัวอย่างระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ เป็น 1,000 : 1,000

เงื่อนไขที่ 3 แบบทดสอบยาว 20 ข้อ ขนาดกลุ่มตัวอย่างระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ เป็น 1,000 : 1,000

การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยวิธีการวัดพื้นที่ของราชู มีรายละเอียดดังนี้

4.1 การตรวจสอบความเป็นมิติเดียวของแบบทดสอบ

ผู้วิจัยตรวจสอบความเป็นมิติเดียวของแบบทดสอบก่อนที่จะนำไปวิเคราะห์ตามทฤษฎีการตอบข้อสอบ โดยการวิเคราะห์องค์ประกอบ (Factor analysis) ด้วยการวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis) แล้วหมุนแกนด้วยวิธีแวนแมกซ์ (Varimax) สำหรับการตรวจสอบความเป็นมิติเดียวครั้งนี้ใช้โปรแกรมสำเร็จรูป SPSS for windows version 7.5

4.2 การประมาณค่าพารามิเตอร์ของข้อสอบ

ผู้วิจัยนำข้อมูลที่สุ่มไว้ในข้อ 2 มาประมาณค่าพารามิเตอร์ของข้อสอบ ได้แก่ ค่าอำนาจจำแนก (a) และค่าความยาก โดยใช้โปรแกรม BILOG version 3.04 (Mislevy and Bock. 1990) ตามทฤษฎีการตอบข้อสอบ ผลจากการวิเคราะห์จะได้แฟ้มข้อมูลความแปรปรวน-ความแปรปรวนร่วมของค่าพารามิเตอร์ซึ่งแฟ้มข้อมูลดังกล่าวจะนำไปวิเคราะห์การปรับเทียบค่าพารามิเตอร์ระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ แล้วนำไปวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันโดยใช้วิธีการวัดพื้นที่ของราชูต่อไป

4.3 การแปลงหน่วยของพารามิเตอร์ข้อสอบ

หลังจากประมาณค่าพารามิเตอร์ระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบแล้วผู้วิจัยนำแฟ้มข้อมูลความแปรปรวน-ความแปรปรวนร่วมของค่าพารามิเตอร์มาเทียบมาตรฐานเพื่อให้อยู่บนสเกลเดียวกัน โดยใช้โปรแกรม EQUATE 2.0 (Baker and others. 1991) ทั้งนี้เพราะการประมาณค่าพารามิเตอร์ของข้อสอบที่แยกกันวิเคราะห์แต่ละกลุ่มนั้น ไม่สามารถนำค่าพารามิเตอร์มาเปรียบเทียบกันได้โดยตรง ดังนั้นจึงจำเป็นต้องเทียบมาตรฐานของค่าพารามิเตอร์ของทั้งสองกลุ่มให้อยู่บนมาตรฐานเดียวกันก่อน ผลจากการวิเคราะห์จะได้ค่า A (slope) และค่า K (intercept) ซึ่งจะนำไปวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันด้วยวิธีการวัดพื้นที่ของราชูต่อไป

4.4 หลังจากเทียบมาตรฐานของค่าพารามิเตอร์ของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบให้อยู่บนมาตรฐานเดียวกันแล้ว ต่อจากนั้นเป็นการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้วิธีการวัดพื้นที่ของราชู โดยใช้โปรแกรม IRTDIF version 1.0 ซึ่งพัฒนาโดย คิมและโคเฮน (Kim and Cohen. 1992 ; citing Raju. 1990) ซึ่งผู้วิจัยนำแฟ้มข้อมูลความแปรปรวน-ความแปรปรวนร่วมจากข้อ 4.2 และค่า A และค่า K จากข้อ 4.3 มาวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอและแบบไม่สมำเสมอ โดยใช้โมเดลแบบสองพารามิเตอร์ คำนวณพื้นที่ในช่วงเปิดชนิดไม่คิดเครื่องหมาย ผลที่ได้จากการวิเคราะห์จะนำไปเป็นเกณฑ์มาตรฐานตัดสินข้อสอบที่ทำหน้าที่ต่างกันด้วยวิธี MH และวิธี LR ต่อไป

5. การวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันด้วยวิธี MH

ในการวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอและแบบไม่สมำเสมอด้วยวิธี MH นั้น ผู้วิจัยได้ดำเนินการวิเคราะห์ข้อมูลโดยใช้โปรแกรม MHDIF version 1.0 (Fidalgo. 1995)

6. การวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันด้วยวิธี LR

ในการวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอและแบบไม่สมำเสมอด้วยวิธี LR นั้น ผู้วิจัยได้ดำเนินการวิเคราะห์ข้อมูลโดยใช้โปรแกรม SPSS/PC⁺ version 4.01

7. การคำนวณค่าอำนาจการตรวจสอบและการจำแนกผิดพลาด

หลังจากวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอและแบบไม่สมำเสมอด้วยวิธี MH และวิธี LR แล้ว ผู้วิจัยนำผลการวิเคราะห์ที่ได้จากข้อ 5 และ 6 มาคำนวณค่าอำนาจการตรวจสอบและการจำแนกผิดพลาด ด้วยการเปรียบเทียบกับผลที่ได้จากการตรวจสอบด้วยวิธีการวัดพื้นที่ของราชูในข้อ 4 เป็นเกณฑ์มาตรฐาน หากวิธี MH หรือวิธี LR ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันซึ่งเป็นข้อเดียวกันกับที่ตรวจสอบด้วยวิธีวัดพื้นที่ของราชู ถือว่าวิธีนั้นมีความถูกต้องในการตรวจสอบ แต่ถ้าตรวจพบข้อสอบข้ออื่นที่ทำหน้าที่ต่างกันนอกเหนือจากที่ตรวจพบด้วยวิธีการวัดพื้นที่ของราชู ถือว่ามีการจำแนกผิดพลาดในการตรวจสอบ ในการคำนวณค่าอำนาจการตรวจสอบและการจำแนกผิดพลาด ผู้วิจัยใช้สูตรดังนี้(Narayanan and Swaminathan. 1994)

$$\text{อำนาจการตรวจสอบ} = \frac{\text{จำนวนข้อสอบที่ระบุว่าทำหน้าที่ต่างกันได้ถูกต้อง}}{\text{จำนวนข้อสอบที่ทำหน้าที่ต่างกันทั้งหมดในแบบทดสอบ}} \times 100$$

$$\text{การจำแนกผิดพลาด} = \frac{\text{จำนวนข้อสอบที่ระบุผิดพลาดว่าทำหน้าที่ต่างกัน}}{\text{จำนวนข้อสอบที่ทำหน้าที่ไม่ต่างกันทั้งหมดในแบบทดสอบ}} \times 100$$

8. การเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาด ระหว่างวิธี MH กับวิธี LR เป็นการเปรียบเทียบความแตกต่างของอำนาจการตรวจสอบและการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอและแบบไม่สมำเสมอ ระหว่างวิธี MH กับวิธี LR โดยการทดสอบด้วยค่าสถิติ Z ที่ระดับนัยสำคัญ .05 (ชูศรี วงศ์รัตน์, 2530 : 209)

บทที่ 4

ผลการวิเคราะห์ข้อมูล

การวิจัยครั้งนี้ มีความมุ่งหมายเพื่อเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอและแบบไม่สม่ำเสมอระหว่างวิธี MH และวิธี LR เมื่อใช้ความยาวแบบทดสอบและขนาดกลุ่มตัวอย่างต่างกัน ซึ่งผลการวิเคราะห์ข้อมูลครั้งนี้ผู้วิจัยได้นำเสนอเป็น 6 ตอนดังนี้

- ตอนที่ 1 การวิเคราะห์ข้อมูลเบื้องต้น
- ตอนที่ 2 การวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันตามทฤษฎีการตอบข้อสอบด้วยวิธี EUA
- ตอนที่ 3 การวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันด้วยวิธี MH
- ตอนที่ 4 การวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันด้วยวิธี LR
- ตอนที่ 5 การวิเคราะห์อำนาจการตรวจสอบและการจำแนกผิดพลาด
- ตอนที่ 6 การเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาดระหว่างวิธี MH กับวิธี LR

เพื่อให้การนำเสนอผลการวิเคราะห์ข้อมูลมีความเข้าใจตรงกัน ผู้วิจัยจึงกำหนดสัญลักษณ์ดังนี้

$\bar{X}$	แทน ค่าเฉลี่ยเลขคณิต
S.D.	แทน ค่าส่วนเบี่ยงเบนมาตรฐาน
$N_R : N_F$	แทน ขนาดผู้สอบกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ
MH	แทน วิธีแมนเทิล-แฮนส์เซล
LR	แทน วิธีการถดถอยโลจิสติก
EUA	แทน วิธีการวัดพื้นที่แบบไม่คิดเครื่องหมายของราชู
p	แทน ค่าความยากง่ายตามทฤษฎีมาตรฐานเดิม
r	แทน ค่าอำนาจจำแนกตามทฤษฎีมาตรฐานเดิม
b	แทน ค่าความยากง่ายตามทฤษฎีการตอบข้อสอบ
a	แทน ค่าอำนาจจำแนกตามทฤษฎีการตอบข้อสอบ
P	แทน อำนาจการตรวจสอบ
E1	แทน การจำแนกผิดพลาด
*	แทน $p < .05$
**	แทน $p < .01$
***	แทน $p < .001$
****	แทน ไม่สามารถคำนวณค่าได้
A	แทน ค่าความลาดชันของเส้นถดถอย
K	แทน ค่าจุดตัดบนแกนตั้ง
H	แทน ค่าสถิติ H ของราชู

ผลการวิเคราะห์ข้อมูล

ตอนที่ 1 การวิเคราะห์ข้อมูลเบื้องต้น

1.1 ค่าสถิติพื้นฐาน

การวิเคราะห์ส่วนนี้ผู้วิจัยได้นำคะแนนของแบบทดสอบความสามารถทางสมองตามแนวโครงสร้างของโอตีส-เลนนอน จำนวน 80 ข้อ และข้อสอบที่สุ่มมาจำนวน 20 ข้อ, 40 ข้อ และ 60 ข้อ คำนวณค่าสถิติพื้นฐาน ปรากฏผลดังตาราง 6

ตาราง 6 ค่าสถิติพื้นฐานของแบบทดสอบความสามารถทางสมองตามแนวโครงสร้างของโอตีส-เลนนอนตามระดับความยาว 20 ข้อ, 40 ข้อ, 60 ข้อ และ 80 ข้อ

ค่าสถิติ	ความยาวของแบบทดสอบ			
	20 ข้อ	40 ข้อ	60 ข้อ	80 ข้อ
คะแนนเต็ม	20	40	60	80
ค่าเฉลี่ยเลขคณิต ($\bar{X}$)	14.31	27.55	40.92	56.47
ส่วนเบี่ยงเบนมาตรฐาน (S.D.)	3.43	6.05	9.03	12.54
ค่าสูงสุด (Maximum)	20	38	56	77
ค่าต่ำสุด (Minimum)	1	5	7	12
ความเชื่อมั่น (Reliability)	.73	.83	.88	.92
จำนวนผู้สอบ	2000	2000	2000	2000

ผลการวิเคราะห์ตามตาราง 6 พบว่า แบบทดสอบความสามารถทางสมองตามแนวโครงสร้างของโอตีส-เลนนอนความยาว 20 ข้อ มีคะแนนต่ำสุดเท่ากับ 1 คะแนน คะแนนสูงสุดเท่ากับ 20 คะแนน มีค่าเฉลี่ยของคะแนนเท่ากับ 14.31 ค่าส่วนเบี่ยงเบนมาตรฐานเท่ากับ 3.43 และมีความเชื่อมั่นเท่ากับ .73

แบบทดสอบความสามารถทางสมองตามแนวโครงสร้างของโอตีส-เลนนอน ความยาว 40 ข้อ มีคะแนนต่ำสุดเท่ากับ 5 คะแนน คะแนนสูงสุดเท่ากับ 38 คะแนน มีค่าเฉลี่ยของคะแนนเท่ากับ 27.55 ส่วนเบี่ยงเบนมาตรฐานเท่ากับ 6.05 และมีความเชื่อมั่นเท่ากับ .83

แบบทดสอบความสามารถทางสมองตามแนวโครงสร้างของโอตีส-เลนนอน ความยาว 60 ข้อ มีคะแนนต่ำสุดเท่ากับ 7 คะแนน คะแนนสูงสุดเท่ากับ 56 คะแนน มีค่าเฉลี่ยของคะแนนเท่ากับ 40.92 ส่วนเบี่ยงเบนมาตรฐานเท่ากับ 9.03 และมีความเชื่อมั่นเท่ากับ .88

แบบทดสอบความสามารถทางสมองตามแนวโครงสร้างของโอตีส-เลนนอน ความยาว 80 ข้อ มีคะแนนต่ำสุดเท่ากับ 12 คะแนน คะแนนสูงสุดเท่ากับ 77 คะแนน มีค่าเฉลี่ยของคะแนนเท่ากับ 56.47 ส่วนเบี่ยงเบนมาตรฐานเท่ากับ 12.54 และมีความเชื่อมั่นเท่ากับ .92

1.2 การวิเคราะห์คุณภาพของแบบทดสอบ

การวิเคราะห์ส่วนนี้ผู้วิจัยได้นำคะแนนของแบบทดสอบความสามารถทางสมองตามแนวโครงสร้างของโอทิส-เลนนอนจำนวน 80 ข้อมาคำนวณค่าความยากง่าย ค่าอำนาจจำแนกและค่าความเชื่อมั่น (KR-20) โดยวิเคราะห์ตามทฤษฎีมาตรฐานเดิม (Classical Test Theory : CTT) ปรากฏผลดังตาราง 7

ตาราง 7 ค่าความยากง่าย ค่าอำนาจจำแนก และค่าความเชื่อมั่น (KR-20) ของแบบทดสอบความสามารถทางสมองตามแนวโครงสร้างของโอทิส-เลนนอน

ข้อ	ค่า p	ค่า r	ข้อ	ค่า p	ค่า r	ข้อ	ค่า p	ค่า r	ข้อ	ค่า p	ค่า r
1	0.70	0.66	21	0.64	0.25	41	0.48	0.73	61	0.63	0.63
2	0.73	0.58	22	0.66	0.56	42	0.64	0.70	62	0.58	0.69
3	0.79	0.43	23	0.77	0.45	43	0.63	0.63	63	0.57	0.82
4	0.77	0.55	24	0.72	0.50	44	0.60	0.66	64	0.55	0.79
5	0.79	0.43	25	0.66	0.58	45	0.65	0.65	65	0.69	0.69
6	0.71	0.60	26	0.75	0.58	46	0.65	0.72	66	0.42	0.69
7	0.74	0.56	27	0.75	0.47	47	0.70	0.64	67	0.68	0.68
8	0.72	0.59	28	0.69	0.68	48	0.70	0.65	68	0.53	0.47
9	0.68	0.65	29	0.78	0.32	49	0.46	0.24	69	0.56	0.60
10	0.79	0.51	30	0.58	0.43	50	0.49	0.82	70	0.61	0.63
11	0.60	0.72	31	0.60	0.59	51	0.65	0.28	71	0.59	0.77
12	0.75	0.52	32	0.64	0.73	52	0.69	0.66	72	0.69	0.52
13	0.61	0.81	33	0.60	0.66	53	0.71	0.53	73	0.41	0.46
14	0.71	0.62	34	0.69	0.58	54	0.71	0.61	74	0.43	0.71
15	0.79	0.50	35	0.59	0.38	55	0.65	0.46	75	0.42	0.22
16	0.58	0.81	36	0.65	0.54	56	0.66	0.70	76	0.38	0.49
17	0.46	0.72	37	0.36	0.64	57	0.59	0.57	77	0.37	0.57
18	0.71	0.50	38	0.70	0.45	58	0.59	0.72	78	0.42	0.25
19	0.44	0.71	39	0.66	0.72	59	0.62	0.59	79	0.36	0.45
20	0.68	0.50	40	0.69	0.63	60	0.58	0.77	80	0.45	0.29
KR-20		.83									

ผลการวิเคราะห์ตามตาราง 7 พบว่าแบบทดสอบความสามารถทางสมองตามแนวโครงสร้างของโอทิส-เลนนอน จำนวน 80 ข้อ มีค่าความยากง่ายอยู่ในช่วง 0.36 – 0.39 จำนวน 4 ข้อ ความยากง่ายอยู่ในช่วง 0.40 – 0.59 จำนวน 23 ข้อ ความยากง่ายอยู่ในช่วง 0.60 – 0.79 จำนวน 53 ข้อ มีค่าอำนาจจำแนกอยู่ในช่วง 0.22 – 0.82 และค่าความเชื่อมั่น (KR-20) เท่ากับ .83

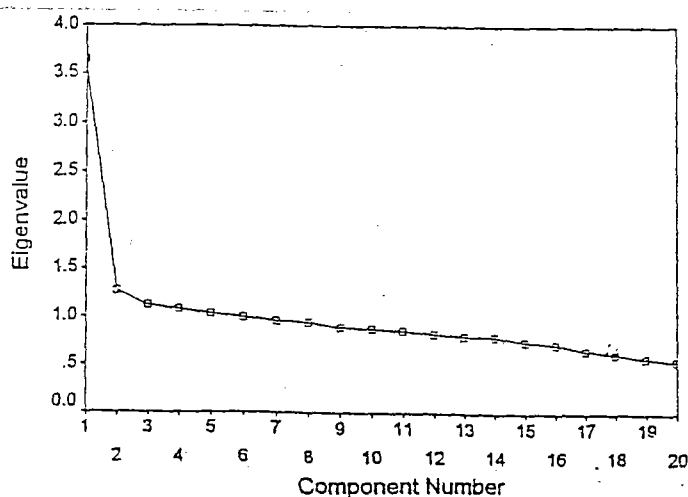
ตอนที่ 2 การวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันตามทฤษฎีการตอบข้อสอบด้วยวิธี การวัดพื้นที่ของราชู (EUA)

2.1 การตรวจสอบความเป็นมิติเดียว

การวิเคราะห์ส่วนนี้ผู้วิจัยได้นำคะแนนของแบบทดสอบความสามารถทางสมองตามแนวโครงสร้างของโอติส-เลนนอน ที่สุ่มมาจำนวน 20 ข้อ, 40 ข้อ และ 60 ข้อ มาตรวจสอบความเป็นมิติเดียว โดยใช้การวิเคราะห์องค์ประกอบแบบสำรวจ หาค่าไอเกน และร้อยละความแปรปรวน พร้อมทั้งทำการทดสอบสกรี (Scree-test) ปรากฏผลดังตาราง 8 – 10 และภาพประกอบ 5 – 7

ตาราง 8 การวิเคราะห์องค์ประกอบของแบบทดสอบความยาว 20 ข้อ

องค์ประกอบ	ค่าไอเกน (Eigen value)	ร้อยละความแปรปรวน
1	3.64	18.19
2	1.27	6.36
3	1.12	5.61
4	1.08	5.39
5	1.03	5.17

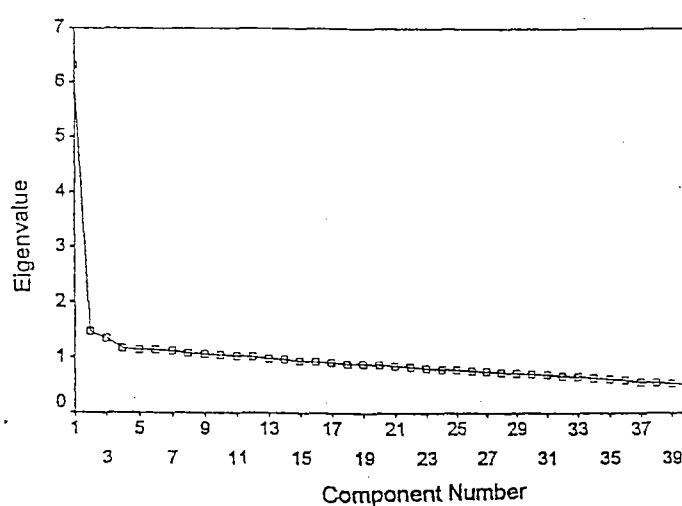


ภาพประกอบ 5 การตรวจสอบความเป็นมิติเดียวของแบบทดสอบความยาว 20 ข้อ ด้วยการทดสอบสกรี (Scree-test)

ผลการวิเคราะห์ตามตาราง 8 พบว่า แบบทดสอบความยาว 20 ข้อ มีจำนวนองค์ประกอบ 5 องค์ประกอบ ค่าไอเกนขององค์ประกอบที่ 1 มีค่าสูงสุด (3.64) ซึ่งมีค่าแตกต่างจากค่าไอเกนขององค์ประกอบอื่นๆที่เหลือน้อยอย่างชัดเจนและค่าไอเกนขององค์ประกอบที่เหลือทั้งหมดมีค่าใกล้เคียงกัน (Lord, 1980) ประกอบกับเส้นภาพขององค์ประกอบในภาพประกอบ 5 แสดงว่าแบบทดสอบมีความเป็นมิติเดียว ดังนั้นจึงสามารถนำไปวิเคราะห์ตามทฤษฎีการตอบข้อสอบได้

ตาราง 9 การวิเคราะห์องค์ประกอบของแบบทดสอบความยาว 40 ข้อ

องค์ประกอบ	ค่าไอเกน (Eigen value)	ร้อยละความแปรปรวน
1	6.32	15.79
2	1.48	3.69
3	1.35	3.38
4	1.17	2.92
5	1.15	2.87
6	1.14	2.85
7	1.13	2.82
8	1.08	2.69
9	1.06	2.65
10	1.05	2.62
11	1.03	2.57
12	1.01	2.54

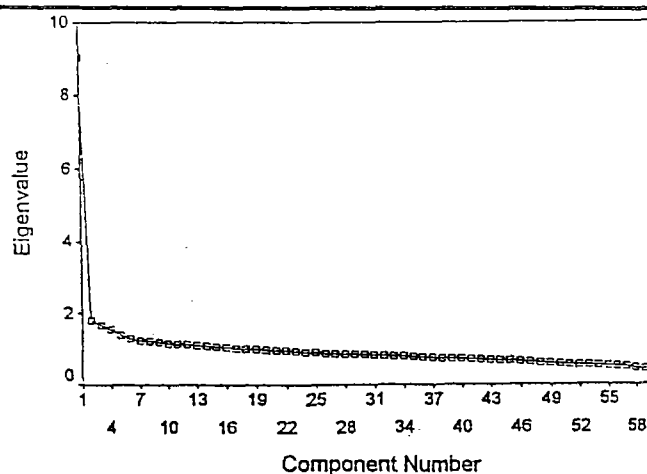


ภาพประกอบ 6 การตรวจสอบความเป็นมิติเดียวของแบบทดสอบความยาว 40 ข้อ ด้วยการทดสอบสกรี้ (Scree-test)

ผลการวิเคราะห์ตามตาราง 9 พบว่า แบบทดสอบความยาว 40 ข้อ มีจำนวนองค์ประกอบ 12 องค์ประกอบ ค่าไอเกนขององค์ประกอบที่ 1 มีค่าสูงสุด (6.32) ซึ่งมีค่าแตกต่างจากค่าไอเกนขององค์ประกอบอื่นๆที่เหลือน้อยอย่างชัดเจนและค่าไอเกนขององค์ประกอบที่เหลือทั้งหมดมีค่าใกล้เคียงกัน (Lord, 1980) ประกอบกับเส้นภาพขององค์ประกอบในภาพประกอบ 6 แสดงว่าแบบทดสอบมีความเป็นมิติเดียว ดังนั้นจึงสามารถนำไปวิเคราะห์ตามทฤษฎีการตอบข้อสอบได้

ตาราง 10 การวิเคราะห์องค์ประกอบของแบบทดสอบความยาว 60 ข้อ

องค์ประกอบ	ค่าไอเกน (Eigen value)	ร้อยละความแปรปรวน
1	9.05	15.09
2	1.80	3.00
3	1.66	2.77
4	1.54	2.57
5	1.38	2.29
6	1.29	2.15
7	1.22	2.04
8	1.21	2.01
9	1.18	1.97
10	1.15	1.92
11	1.14	1.89
12	1.12	1.86
13	1.09	1.82
14	1.08	1.80
15	1.06	1.76
16	1.05	1.75
17	1.02	1.70



ภาพประกอบ 7 การตรวจสอบความเป็นมิติเดียวของแบบทดสอบความยาว 60 ข้อ ด้วยการทดสอบสกรี้ (Scree-test)

ผลการวิเคราะห์ตามตาราง 10 พบว่า แบบทดสอบความยาว 60 ข้อ มีจำนวนองค์ประกอบ 17 องค์ประกอบ ค่าไอเกนขององค์ประกอบที่ 1 มีค่าสูงสุด (9.05) ซึ่งมีค่าแตกต่างจากค่าไอเกนขององค์ประกอบอื่นๆที่เหลือน้อยอย่างชัดเจนและค่าไอเกนขององค์ประกอบที่เหลือทั้งหมดมีค่าใกล้เคียงกัน (Lord, 1980)ประกอบ

กับเส้นภาพขององค์ประกอบในภาพประกอบ 7 แสดงว่าแบบทดสอบมีความเป็นมิติเดียว ดังนั้นจึงสามารถนำไปวิเคราะห์ตามทฤษฎีการตอบข้อสอบได้

2.2 การวิเคราะห์ค่าพารามิเตอร์ของข้อสอบ

การวิเคราะห์ส่วนนี้ผู้วิจัยได้ใช้โปรแกรม BILOG version 3.04 วิเคราะห์ค่าอำนาจจำแนกและค่าความยากของแบบทดสอบความสามารถทางสมองตามแนวโครงสร้างของโอตีส-เลนนอนของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ปรากฏผล ดังตาราง 11 – 13

ตาราง 11 ค่าพารามิเตอร์อำนาจจำแนกและความยาก ของแบบทดสอบความยาว 60 ข้อ

ข้อ	กลุ่มอ้างอิง		กลุ่มเปรียบเทียบ	
	ค่าอำนาจจำแนก (a)	ค่าความยาก (b)	ค่าอำนาจจำแนก (a)	ค่าความยาก (b)
1	.79	-1.51	.75	-1.56
2	.68	-2.52	.68	-2.11
3	.84	-2.17	.75	-2.08
4	.53	-1.89	.51	-2.14
5	.60	-1.78	.71	-1.59
6	.43	-1.52	.38	-1.85
7	.91	-1.78	.68	-2.11
8	.60	-1.03	.66	-.82
9	.89	-1.23	.89	-1.03
10	.58	-1.99	.81	-1.38
11	.81	-1.95	1.02	-1.65
12	.86	-1.11	.81	-1.10
13	.32	-2.30	.35	-2.35
14	.58	-2.08	.67	-1.58
15	.53	-1.69	.51	-1.66
16	.42	-2.52	.43	-2.48
17	.48	-2.34	.55	-1.94
18	.55	-1.56	.70	-1.17
19	.79	-1.17	.95	-.95
20	.34	-3.47	.40	-2.90
21	.67	-1.45	.63	-1.40
22	.68	-1.62	.70	-1.56
23	.44	-.94	.45	-.89
24	1.03	-1.45	1.03	-1.21

ตาราง 11 (ต่อ)

ข้อ	กลุ่มอ้างอิง		กลุ่มเปรียบเทียบ	
	ค่าอำนาจจำแนก (a)	ค่าความยาก (b)	ค่าอำนาจจำแนก (a)	ค่าความยาก (b)
25	1.11	-1.71	1.11	-1.45
26	.36	-1.60	.39	-1.57
27	.29	-1.94	.34	-1.37
28	.90	-1.35	.71	-1.46
29	.80	-1.70	.81	-1.56
30	.49	-1.17	.49	-.95
31	.63	-1.20	.61	-1.27
32	.50	-1.05	.58	-1.07
33	.73	-1.33	.70	-1.31
34	.96	-1.79	.86	-1.81
35	.17	-.33	.20	-.48
36	.65	.33	.67	.49
37	.25	-2.23	.17	-3.00
38	.23	-.84	.28	-1.06
39	.27	-1.12	.38	-.70
40	.42	-.85	.38	-1.01
41	.10	1.55	.08	1.71
42	.23	.20	.29	.29
43	.55	.09	.56	-.95
44	.28	1.23	.26	1.54
45	.53	-1.14	.57	-.94
46	.74	-.64	.83	-.42
47	.08	1.69	.09	1.66
48	.46	.96	.43	1.01
49	.90	-1.43	1.04	-1.25
50	.30	-.82	.39	-.53
51	.57	-1.54	.85	-1.07
52	.32	1.45	.41	.97
53	1.08	-1.62	1.08	-1.38
54	.65	-1.46	.74	-1.15
55	.49	-1.42	.69	-.98
56	.32	.72	.34	.88

ตาราง 11 (ต่อ)

ข้อ	กลุ่มอ้างอิง		กลุ่มเปรียบเทียบ	
	ค่าอำนาจจำแนก	ค่าความยาก	ค่าอำนาจจำแนก	ค่าความยาก
	(a)	(b)	(a)	(b)
57	.45	1.12	.54	.95
58	.19	1.35	.28	.87
59	.54	-.61	.50	-.37
60	.29	.34	.26	.33

ตาราง 12 ค่าพารามิเตอร์อำนาจจำแนกและความยากของแบบทดสอบความยาว 40 ข้อ

ข้อ	กลุ่มอ้างอิง		กลุ่มเปรียบเทียบ	
	ค่าอำนาจจำแนก	ค่าความยาก	ค่าอำนาจจำแนก	ค่าความยาก
	(a)	(b)	(a)	(b)
1	.75	-1.57	.74	-1.57
2	.64	-2.67	.64	-2.17
3	.79	-1.99	1.04	-1.64
4	.81	-2.22	.74	-2.12
5	.60	-1.76	.67	-1.68
6	1.01	-1.47	1.01	-1.27
7	.70	-1.25	.72	-1.20
8	.56	-1.55	.83	-1.13
9	.44	-1.50	.39	-1.80
10	.52	-1.69	.53	-1.62
11	.40	-2.63	.39	-2.73
12	.50	-1.99	.48	-2.21
13	.40	-1.01	.42	-.99
14	.64	-1.69	.72	-1.58
15	.23	-.85	.30	-1.02
16	.38	-1.53	.37	-1.61
17	.30	-1.90	.34	-1.38
18	.75	-1.30	.72	-1.36
19	.95	-1.80	.85	-1.85
20	.17	-.31	.22	-.38
21	.60	.36	.61	.50
22	.25	-2.24	.17	-2.94

ตาราง 12 (ต่อ)

ข้อ	กลุ่มอ้างอิง		กลุ่มเปรียบเทียบ	
	ค่าอำนาจจำแนก	ค่าความยาก	ค่าอำนาจจำแนก	ค่าความยาก
	(a)	(b)	(a)	(b)
23	.23	.21	.29	.23
24	.65	-1.47	.59	-1.50
25	.42	-.84	.39	-1.01
26	.29	-2.53	.36	-2.29
27	.70	-1.39	.76	-1.14
28	.95	-1.31	.76	-1.40
29	.49	-1.40	.49	-.94
30	.55	-1.09	.62	-.90
31	.74	-.62	.85	-.44
32	.08	1.67	.09	1.88
33	.43	1.03	.42	1.06
34	.51	-1.13	.63	-.97
35	.33	-.74	.41	-.52
36	.10	1.48	.08	1.65
37	.52	-.62	.54	-.38
38	1.10	-1.60	1.10	-1.38
39	1.20	-1.66	1.20	-1.46
40	.28	1.25	.23	1.66

ตาราง 13 ค่าพารามิเตอร์อำนาจจำแนกและความยากของแบบทดสอบความยาว 20 ข้อ

ข้อ	กลุ่มอ้างอิง		กลุ่มเปรียบเทียบ	
	ค่าอำนาจจำแนก	ค่าความยาก	ค่าอำนาจจำแนก	ค่าความยาก
	(a)	(b)	(a)	(b)
1	.50	-2.01	.47	-2.27
2	.66	-1.33	.67	-1.26
3	.27	-2.74	.42	-2.03
4	.43	-1.56	.38	-1.85
5	.56	-1.63	.52	-1.67
6	.38	-2.76	.36	-2.73
7	.31	-1.86	.34	-1.37
8	.77	-1.31	.72	-1.34

ตาราง 13 (ต่อ)

ข้อ	กลุ่มอ้างอิง		กลุ่มเปรียบเทียบ	
	ค่าอำนาจจำแนก (a)	ค่าความยาก (b)	ค่าอำนาจจำแนก (a)	ค่าความยาก (b)
9	.96	-1.81	.90	-1.77
10	.20	-.29	.24	-.40
11	.66	.32	.62	.51
12	.27	-2.10	.19	-2.77
13	1.35	-1.60	1.24	-1.43
14	.67	-1.46	.67	-1.53
15	.45	-.83	.45	-.97
16	.76	-.65	.79	-.46
17	1.16	-1.58	1.16	-1.34
18	.42	1.02	.40	1.09
19	1.02	-1.28	.81	-1.37
20	.36	-.71	.47	-.47

ผลการวิเคราะห์ตามตาราง 11 – 13 แสดงค่าพารามิเตอร์ค่าอำนาจจำแนกและค่าความยากของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ของแบบทดสอบความยาว 60 ข้อ, 40 ข้อ และ 20 ข้อ ตามลำดับ ซึ่งค่าพารามิเตอร์ดังกล่าว นำไปใช้ในการปรับเทียบมาตรฐานของค่าพารามิเตอร์ข้อสอบต่อไป

2.3 การปรับเทียบค่าพารามิเตอร์

การวิเคราะห์ส่วนนี้ผู้วิจัยนำค่าพารามิเตอร์ข้อสอบมาวิเคราะห์ด้วยโปรแกรม EQUATE version 2.0 เพื่อปรับเทียบค่าพารามิเตอร์ข้อสอบ ของแบบทดสอบความยาว 20 ข้อ, 40 ข้อ และ 60 ข้อ ระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ปรากฏผลดังตาราง 14

ตาราง 14 ค่า A และค่า K ที่ได้จากการปรับเทียบค่าพารามิเตอร์ข้อสอบ ของแบบทดสอบความยาว 20 ข้อ, 40 ข้อ และ 60 ข้อ ระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ

แบบทดสอบความยาว	A	K
20 ข้อ	.99	-.07
40 ข้อ	1.06	-.04
60 ข้อ	1.07	-.07

ผลการวิเคราะห์ตามตาราง 14 แสดงค่าความลาดชันของเส้นถดถอย(A)และค่าจุดตัดบนแกนตั้ง (K) ของแบบทดสอบความยาว 20 ข้อ, 40 ข้อ และ 60 ข้อ ตามลำดับ

2.4 การวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกัน

การวิเคราะห์ส่วนนี้ผู้วิจัยนำค่า A และ K มาวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันตามทฤษฎีการตอบข้อสอบด้วยวิธี EUA โดยใช้โปรแกรม IRTDIF version 1.0 ปรากฏผลดังตาราง 15 – 17

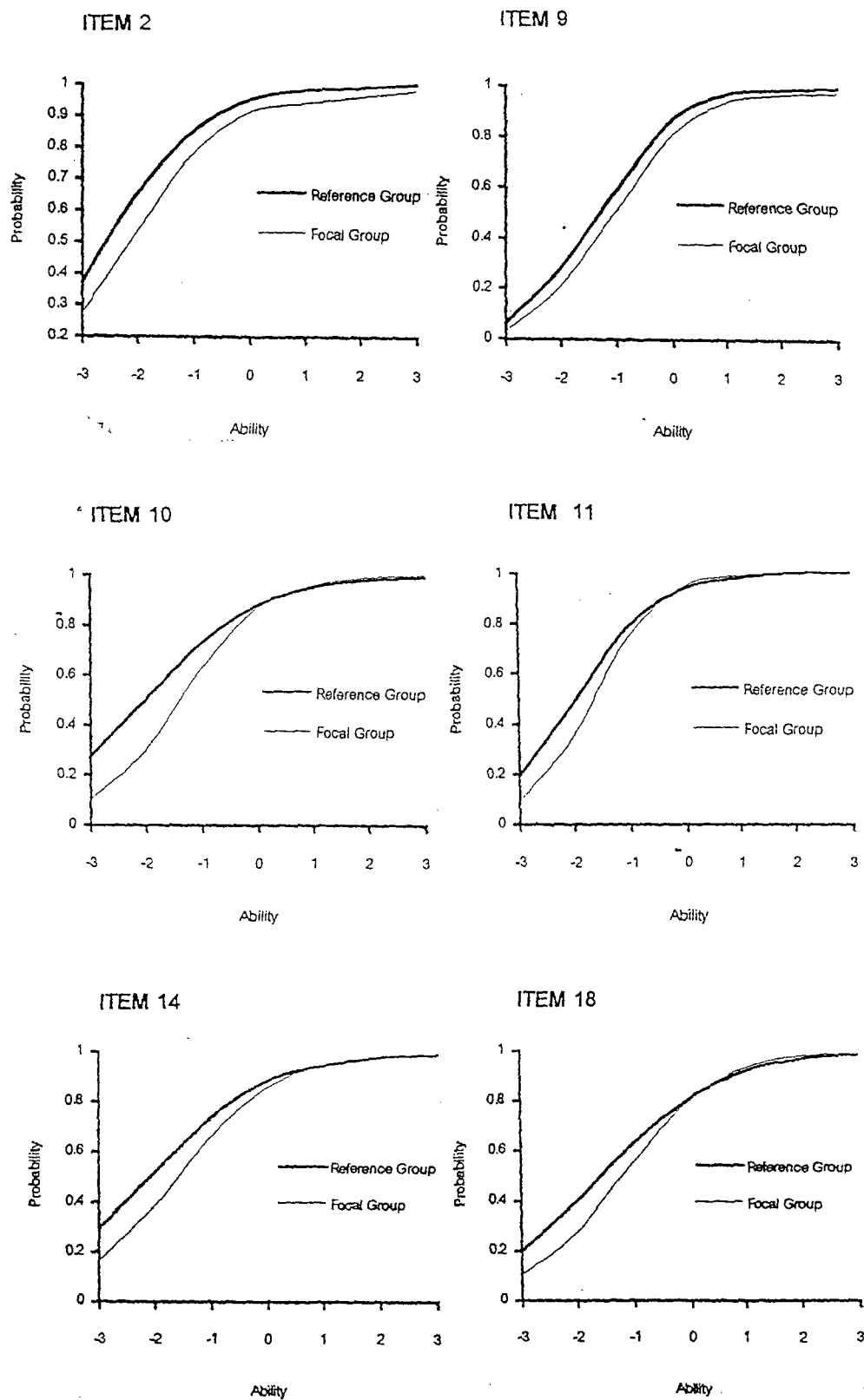
ตาราง 15 การตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน ของแบบทดสอบความยาว 60 ข้อ

ข้อ	EUA	H	Z-test
1	.13	-.13	-.79
2	.66	-.66	-2.05*
3	.32	.32	1.40
4	.07	-.07	-.23
5	.64	-.64	-1.75
6	.27	.27	.42
7	.42	.42	1.41
8	.52	-.52	-1.83
9	.36	.36	2.88**
10	1.11	-1.11	-3.11**
11	.66	-.66	-2.45*
12	.16	-.16	-1.33
13	.65	-.65	-.73
14	.82	-.82	-2.33*
15	.25	-.25	-.66
16	.36	-.36	-.53
17	.78	-.78	-1.54
18	.84	-.84	-2.45*
19	.53	-.53	-2.51*
20	1.18	-1.18	-1.17
21	.22	-.22	-1.31
22	.31	-.31	-1.06
23	.38	-.38	-.75
24	.44	-.44	-2.97**
25	.47	-.47	-3.32***
26	.63	-.63	-.87
27	1.20	-1.20	-1.31
28	.28	.28	1.33
29	.35	-.35	-1.76

ตาราง 15 (ต่อ)

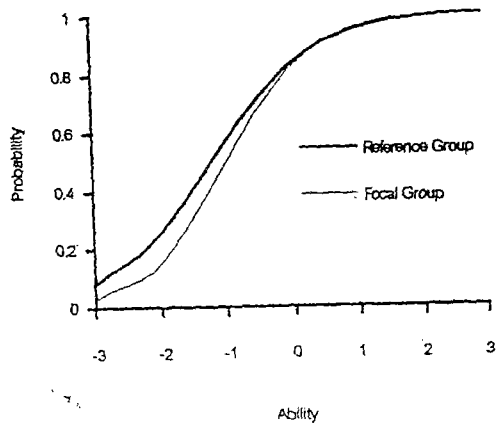
ข้อ	EUA	H	Z-test
30	.89	-.89	-2.35*
31	.10	-.10	-.37
32	.60	-.60	-1.55
33	.22	-.22	-.92
34	.18	.18	1.16
35	1.87	-1.87	-.84
36	.27	-.27	-1.45
37	2.03	2.03	1.02
38	1.45	-1.45	-1.14
39	2.00	-2.00	-2.09*
40	.09	.09	.15
41	2.08	2.08	.28
42	1.75	-1.75	-1.35
43	.35	-.35	-1.41
44	.29	-.29	-1.26
45	.51	-.51	-1.51
46	.44	-.44	-2.46*
47	4.88	-4.88	.59
48	.17	-.17	-.39
49	.44	-.44	-2.30*
50	1.53	-1.53	-1.86
51	1.12	-1.12	-3.66***
52	1.40	-1.40	-1.80
53	.44	-.44	-3.03**
54	.60	-.60	-2.19*
55	1.15	-1.15	-3.01**
56	.56	-.56	-.74
57	.82	-.82	-1.76
58	2.91	-2.91	-1.73
59	.36	-.36	-2.77**
60	.17	-.17	-.17

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

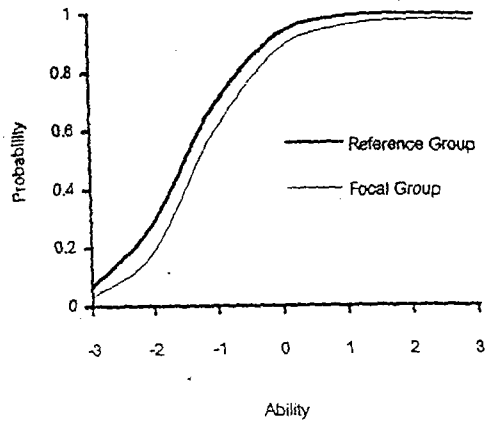


ภาพประกอบ 8 เส้นภาพข้อสอบที่ทำหน้าที่ต่างกัน ของแบบทดสอบความยาว 60 ข้อ

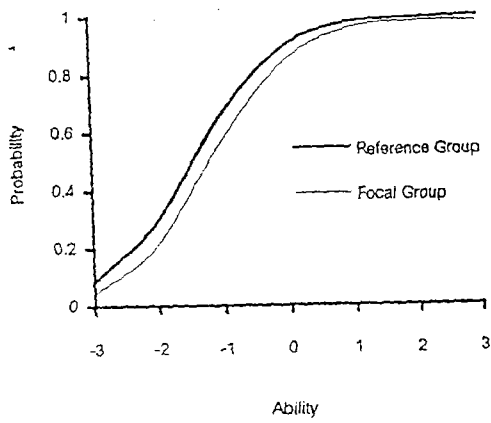
ITEM 19



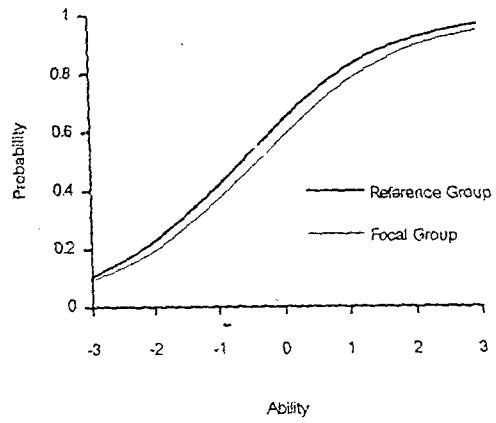
ITEM 24



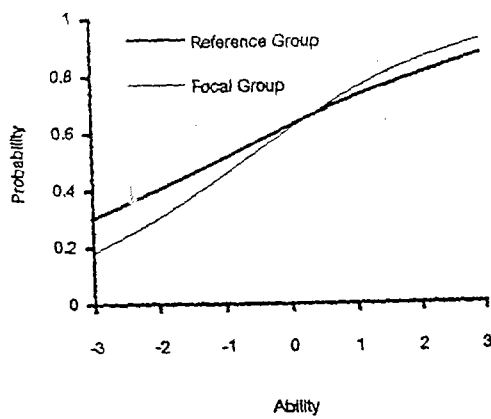
ITEM 25



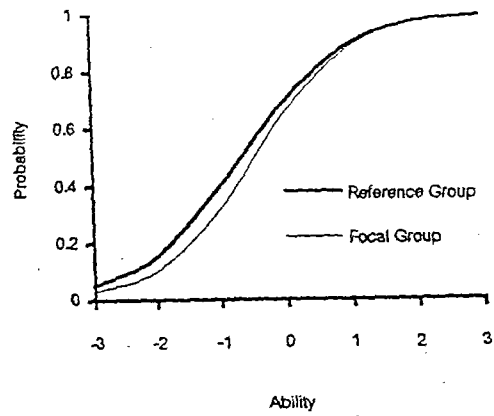
ITEM 30

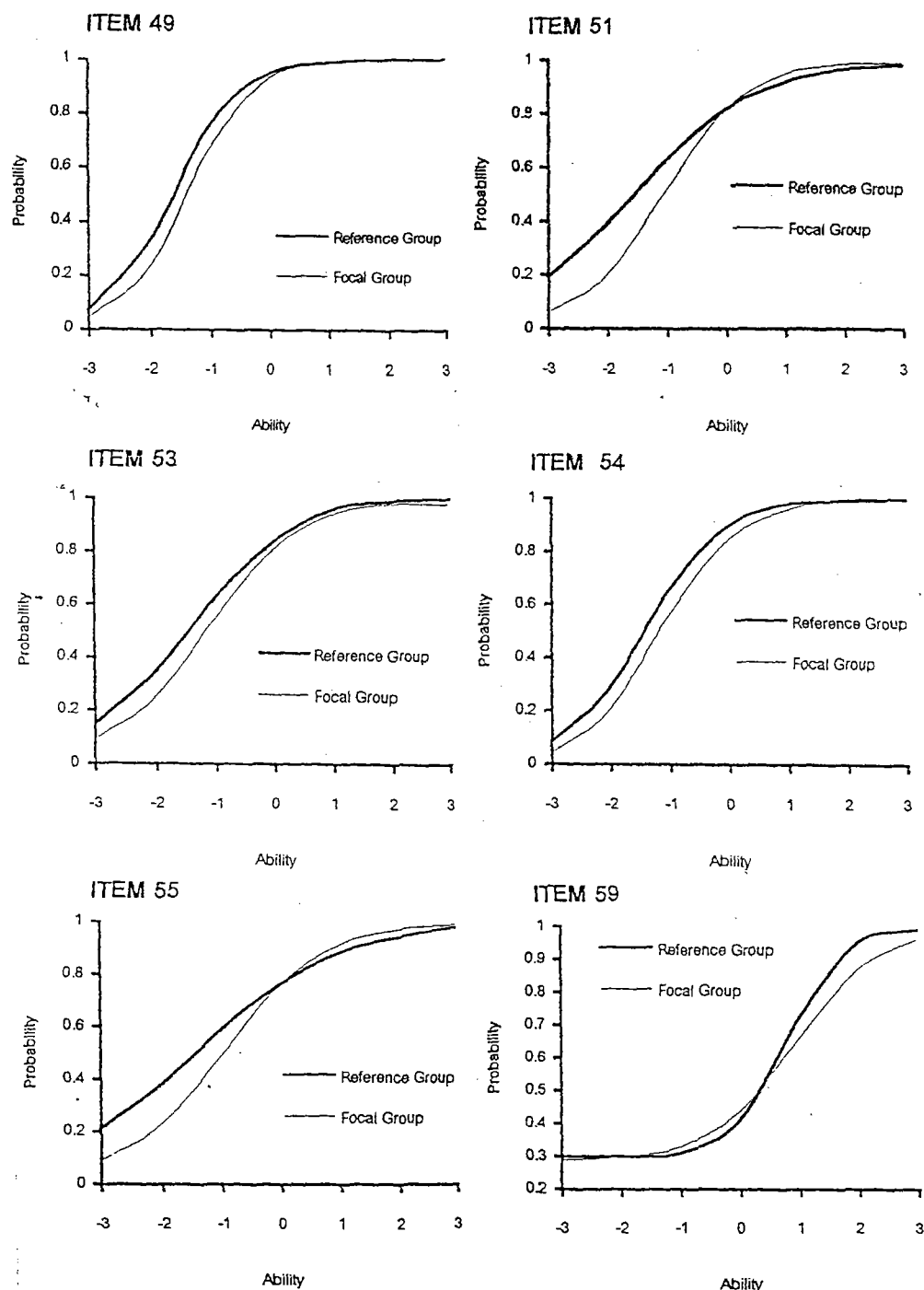


ITEM 39



ITEM 46





ภาพประกอบ 8 (ต่อ)

ผลการวิเคราะห์ตามตาราง 15 และภาพประกอบ 8 พบว่าวิธีการวัดพื้นที่โดยไม่คิดเครื่องหมายของ
 ราชู(EUA) ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน 18 ข้อ คือ ข้อ 2, 9, 10, 11, 14, 18, 19, 24, 25, 30, 39, 46,
 49, 51, 53, 54, 55, 59 โดยเป็นข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอ 6 ข้อ คือข้อ 2, 9, 24, 25, 30, 53
 และเป็นข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ 12 ข้อ คือข้อ 10, 11, 14, 18, 19, 39, 46, 49, 51, 54,
 55, 59

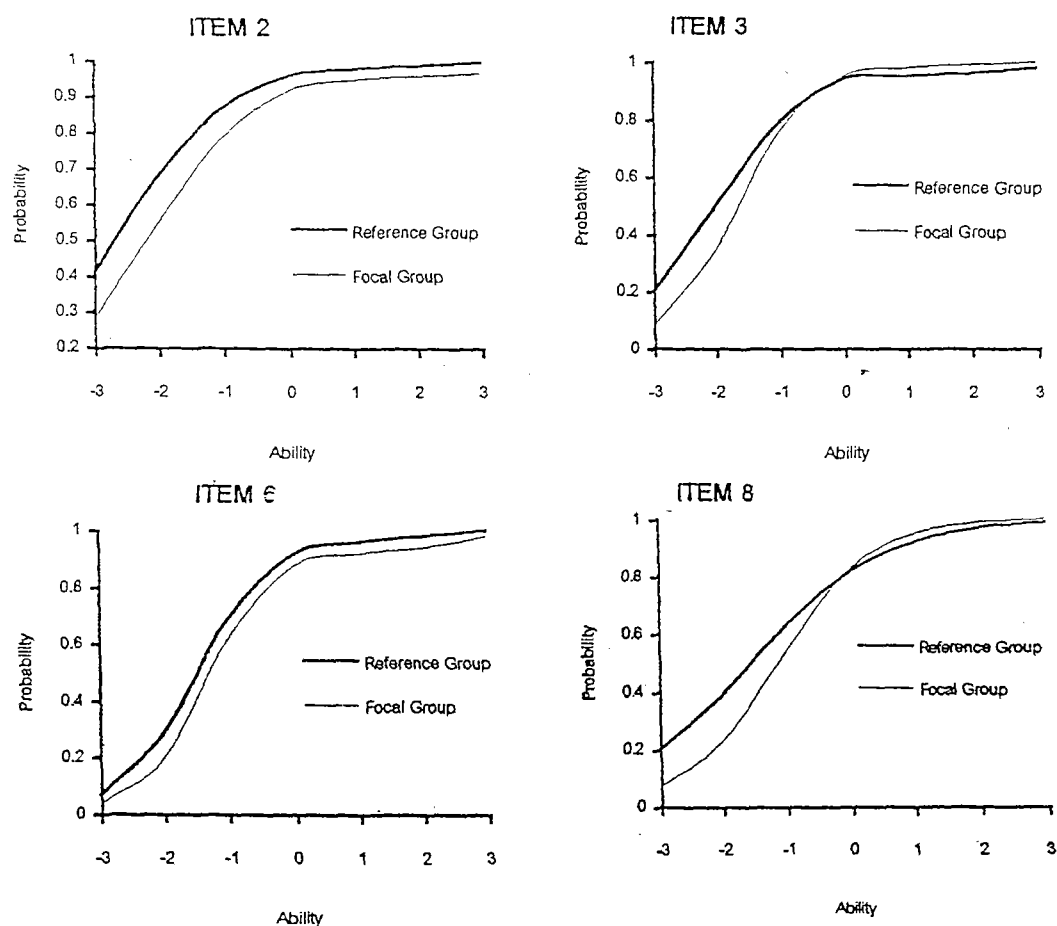
ตาราง 16 การตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน ของแบบทดสอบความยาว 40 ข้อ

ข้อ	EUA	H	Z-test
1	.16	-.16	-.59
2	.72	-.72	-2.89**
3	.70	-.70	-2.42*
4	.29	.29	1.13
5	.44	-.44	-1.12
6	.37	-.37	-2.17*
7	.22	-.22	-.82
8	1.05	-1.05	-3.24**
9	.28	.28	.44
10	.27	-.27	-.62
11	.16	-.16	-.20
12	.14	-.14	-.35
13	.40	-.40	-.64
14	.44	-.44	-1.33
15	1.84	-1.84	-1.40
16	.26	-.26	-.36
17	1.07	-1.07	-1.16
18	.06	-.06	-.30
19	.13	.13	1.49
20	2.46	-2.46	-1.18
21	.21	-.21	-.92
22	1.80	1.80	.91
23	1.64	-1.64	-1.25
24	.13	.13	.94
25	.08	-.08	-.37
26	1.33	-1.33	-1.33
27	.44	-.44	-1.80
28	.27	.27	1.25
29	1.22	-1.22	-3.24**
30	.50	-.50	-1.52

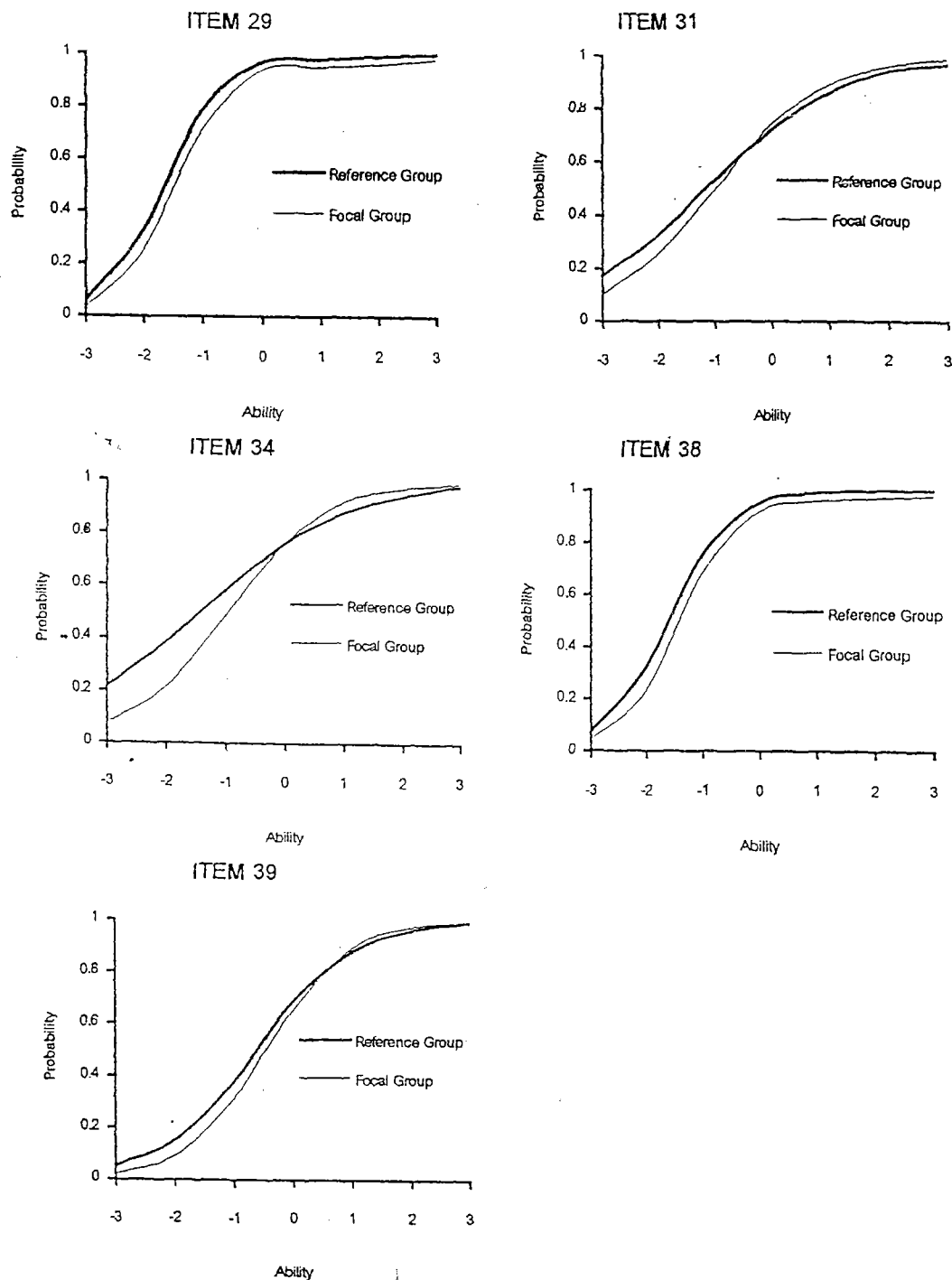
ตาราง 16 (ต่อ)

ข้อ	EUA	H	Z-test
31	.43	-.43	-2.16*
32	4.04	-4.04	-.51
33	.15	-.15	-.27
34	.74	-.74	-1.96*
35	1.12	-1.12	-1.51
36	1.96	1.96	.27
37	.40	-.40	-1.38
38	.37	-.37	-2.66**
39	.34	-.34	-2.75**
40	.64	.64	.52

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



ภาพประกอบ 9 เส้นภาพข้อสอบที่ทำหน้าที่ต่างกัน ของแบบทดสอบความยาว 40 ข้อ



ภาพประกอบ 9 (ต่อ)

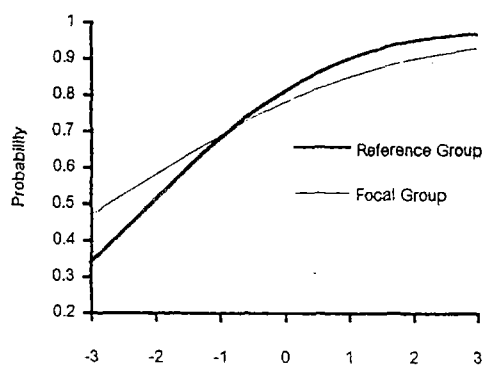
ผลการวิเคราะห์ตามตาราง 16 และภาพประกอบ 9 พบว่าวิธีการวัดพื้นที่โดยไม่คิดเครื่องหมายของ ราชู(EUA) ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน 9 ข้อ คือ ข้อ 2, 3, 6, 8, 29, 31, 34, 38, 39 โดยเป็นข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอ 4 ข้อ คือข้อ 2, 6, 29, 38 และเป็นข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ 5 ข้อ คือข้อ 3, 8, 31, 34, 39

ตาราง 17 การตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน ของแบบทดสอบความยาว 20 ข้อ

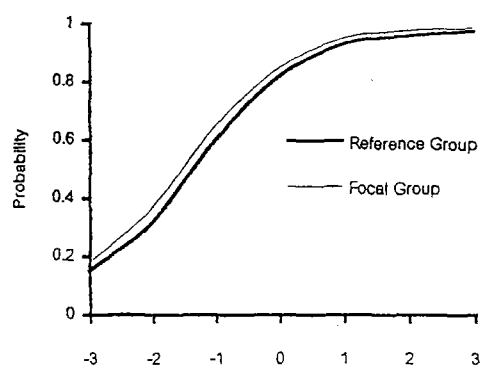
ข้อ	EUA	H	Z-test
1	.26	.26	.48
2	.12	-.12	-.72
3	1.86	-1.86	-2.75**
4	.47	.47	.67
5	.25	.25	.60
6	.08	.08	.28
7	.63	-.63	-.84
8	.14	.14	.54
9	.13	.13	1.18
10	1.13	-1.13	-.69
11	.29	.29	1.52
12	2.23	2.23	1.23
13	.23	.23	1.71
14	.40	.40	2.31*
15	.82	.82	2.23*
16	.25	-.25	-2.60**
17	.30	-.30	-2.39**
18	.23	.23	.41
19	.96	.96	2.69**
20	.94	-.94	-1.52

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

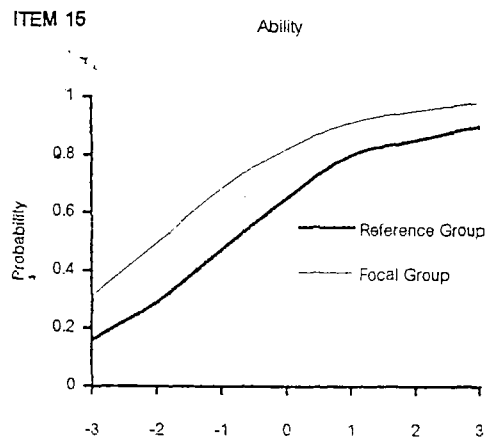
ITEM 3



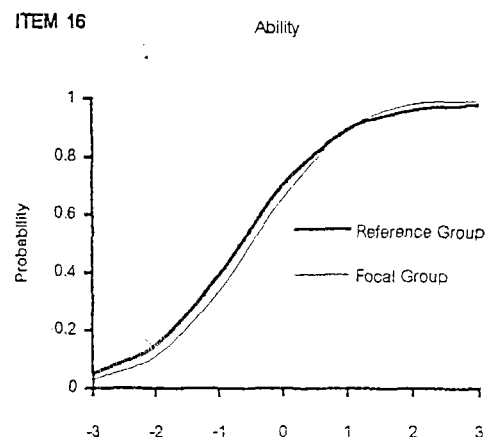
ITEM 14



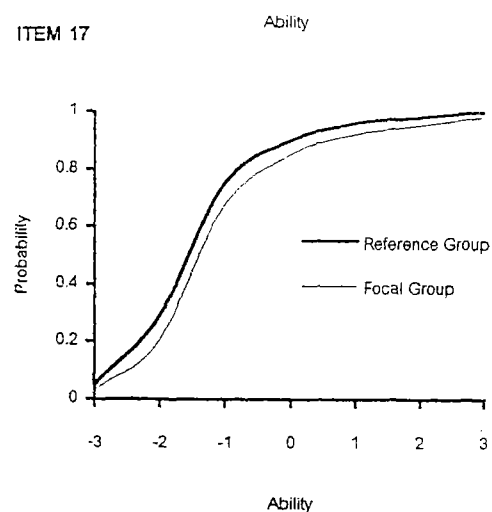
ITEM 15



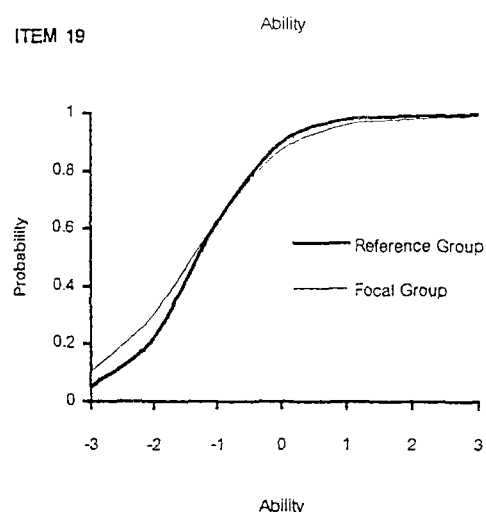
ITEM 16



ITEM 17



ITEM 19



ภาพประกอบ 10 เส้นภาพข้อสอบที่ทำหน้าที่ต่างกัน ของแบบทดสอบความยาว 20 ข้อ

ผลการวิเคราะห์ตามตาราง 17 และภาพประกอบ 10 พบว่าวิธีการวัดพื้นที่โดยไม่คิดเครื่องหมายของราชู(EUA) ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน 6 ข้อ คือข้อ 3, 14, 15, 16, 17, 19 โดยเป็นข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอ 3 ข้อ คือข้อ 14, 15, 17 และเป็นข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สมำเสมอ 3 ข้อ คือ ข้อ 3, 16, 19

ตอนที่ 3 การวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันด้วยวิธีแมนเทิล-แฮนส์เซล

การวิเคราะห์ส่วนนี้ผู้วิจัยใช้โปรแกรม MHDIF version 1.0 วิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอและแบบไม่สม่ำเสมอ ด้วยวิธี MH ปรากฏผลดังตาราง 18

ตาราง 18 การวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอและแบบไม่สม่ำเสมอด้วยวิธี MH ของแบบทดสอบแต่ละความยาวและกลุ่มตัวอย่างแต่ละขนาด

แบบทดสอบ	กลุ่มตัวอย่างขนาด ($N_R : N_F$)	MH	
		แบบสม่ำเสมอ	แบบไม่สม่ำเสมอ
20 ข้อ	300 : 100	-	-
	300 : 200	-	-
	300 : 300	15	-
	500 : 100	14	11, 20
	500 : 200	17	16, 18
	500 : 300	-	-
	1000 : 100	14	11
	1000 : 200	-	16, 19
	1000 : 300	-	13, 16, 19
	1000 : 1000	14	3, 16
40 ข้อ	300 : 100	-	-
	300 : 200	-	2, 39
	300 : 300	19, 38	11, 12, 15, 19, 20, 21 23, 26, 28, 29, 31, 33, 37
	500 : 100	17, 38	4, 38
	500 : 200	2	39
	500 : 300	-	8

ตาราง 18 (ต่อ)

แบบทดสอบ ความยาว	กลุ่มตัวอย่างขนาด ($N_R : N_F$)	MH	
		แบบสม่ำเสมอ	แบบไม่สม่ำเสมอ
	1000 : 100	38	-
	1000 : 200	6, 38	2, 39
	1000 : 300	6, 8, 38	8, 13, 20, 27, 39
	1000 : 1000	3, 24, 29, 38	3, 24, 31, 39
60 ข้อ	300 : 100	2, 59	29, 33, 36, 43, 46
	300 : 200	-	7, 18
	300 : 300	1, 46, 53	1, 18, 39
	500 : 100	59	46, 48, 59
	500 : 200	24	2, 24, 46, 54
	500 : 300	3, 30, 53	33, 46, 51
	1000 : 100	2, 36	2, 10, 11, 14, 36, 48, 59
	1000 : 200	-	18, 54, 55, 59
	1000 : 300	30, 53	30, 39, 46, 51
	1000 : 1000	2, 9, 25, 52, 53	5, 10, 14, 18, 19, 37, 51, 52, 59

ผลการวิเคราะห์ตามตาราง 18 การวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอด้วยวิธี MH พบว่า แบบทดสอบความยาว 20 ข้อ กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 100, 1000 : 100, 1000 : 1000$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 1 ข้อ คือข้อ 14 กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 300$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 1 ข้อ คือข้อ 15 และ กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 200$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 1 ข้อ คือข้อ 17

แบบทดสอบ 40 ข้อ กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 300$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 2 ข้อ คือข้อ 19, 38 กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 100$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันจำนวน 2 ข้อ คือข้อ 17, 38 กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 200$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 1 ข้อ คือข้อ 2 กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 100$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันจำนวน 1 ข้อ คือข้อ 38 กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 200$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 2 ข้อ คือข้อ 6, 38 กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 300$ สามารถตรวจพบ ข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 3 ข้อ คือข้อ 6, 8, 38 และ กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 1000$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันจำนวน 4 ข้อ คือข้อ 3, 24, 29 และ 38

แบบทดสอบความยาว 60 ข้อ กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 5 ข้อ คือข้อ 29, 33, 36, 43, 46 กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 200$ จำนวน 2 ข้อ คือข้อ 7, 18 กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 300$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันจำนวน 3 ข้อ คือข้อ 1, 18, 39 กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 100$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันจำนวน 3 ข้อ คือ ข้อ 46, 48, 59 กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 200$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันจำนวน 4 ข้อ คือข้อ 2, 24, 46, 54 กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 300$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันจำนวน 3 ข้อ คือ ข้อ 33, 46, 51 กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 100$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันจำนวน 7 ข้อ คือ ข้อ 2, 10, 11, 14, 36, 48, 59 กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 200$ สามารถตรวจพบข้อสอบที่ทำ

หน้าที่ต่างกันจำนวน 4 ข้อ คือ 18, 54, 55, 59 กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 300$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันจำนวน 4 ข้อ คือข้อ 30, 39, 46, 51 และกลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 1000$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันจำนวน 9 ข้อ คือข้อ 5, 10, 14, 18, 19, 37, 51, 52, 59

ตอนที่ 4 การวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันด้วยวิธีการถดถอยโลจิสติก

การวิเคราะห์ส่วนนี้ผู้วิจัยใช้โปรแกรม *SPSS/PC*⁺ version 4.01 วิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอและแบบไม่สม่ำเสมอ ด้วยวิธี LR ปรากฏผลดังตาราง 19

ตาราง 19 การวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอและแบบไม่สม่ำเสมอด้วยวิธี LR ของแบบทดสอบแต่ละความยาวและกลุ่มตัวอย่างแต่ละขนาด

แบบทดสอบ	กลุ่มตัวอย่างขนาด ($N_R : N_F$)	LR	
		แบบสม่ำเสมอ	แบบไม่สม่ำเสมอ
20 ข้อ	300 : 100	-	-
	300 : 200	-	16
	300 : 300	15	14
	500 : 100	11, 14	-
	500 : 200	6, 17	19
	500 : 300	-	-
	1000 : 100	14	-
	1000 : 200	-	16, 19
	1000 : 300	19	-
	1000 : 1000	13, 17	3, 19, 20
40 ข้อ	300 : 100	39	23, 31
	300 : 200	2, 24, 39	9, 23, 33, 37
	300 : 300	19, 29	3, 4, 30
	500 : 100	17, 25, 29	31
	500 : 200	2	3, 9, 37
	500 : 300	29	27, 34

ตาราง 19 (ต่อ)

แบบทดสอบ ความยาว	กลุ่มตัวอย่างขนาด ($N_R : N_F$)	LR	
		แบบสม่ำเสมอ	แบบไม่สม่ำเสมอ
	1000 : 100	25, 29, 38	31, 40
	1000 : 200	2, 14	3, 9, 19, 34, 37
	1000 : 300	6, 38	34, 39
	1000 : 1000	2, 15, 38	26, 29, 31, 34, 39
60 ข้อ	300 : 100	2, 26, 33, 46, 51	3, 12, 18, 36
	300 : 200	10, 26	14, 18, 34, 49
	300 : 300	1, 46, 53	11, 12, 39
	500 : 100	2, 19, 27	51, 55, 59
	500 : 200	24	14, 54, 60
	500 : 300	3, 33, 53	10, 23, 39, 46, 55
	1000 : 100	2	3, 49, 51, 55, 59
	1000 : 200	2, 24	3, 14, 18, 34, 37, 49, 54
	1000 : 300	3, 30, 53	11, 23, 39, 46, 49, 50, 55
	1000 : 1000	2, 11, 25, 32, 53	19, 41, 44, 46, 49

ผลการวิเคราะห์ตามตาราง 19 การวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอด้วยวิธี LR พบว่า แบบทดสอบความยาว 20 ข้อ เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 300$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 1 ข้อ คือข้อ 15 เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 100$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันจำนวน 2 ข้อคือข้อ 11, 14 เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 200$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันจำนวน 2 ข้อ คือข้อ 6, 17 เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 100$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 1 ข้อ คือข้อ 14 เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 300$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 1 ข้อ คือข้อ 19 และเมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 1000$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 2 ข้อ คือข้อ 13, 17

แบบทดสอบความยาว 40 ข้อ เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 1 ข้อ คือข้อ 39 เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 200$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันจำนวน 3 ข้อ คือข้อ 2, 24, 39 เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 300$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 2 ข้อ คือข้อ 19, 29 เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 100$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 3 ข้อ คือข้อ 17, 25, 29 เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 200$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันจำนวน 1 ข้อ คือข้อ 2 เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 300$ สามารถตรวจพบข้อสอบที่

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 300$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันจำนวน 2 ข้อ คือข้อ 34, 39 และเมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 1000$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันจำนวน 5 ข้อ คือข้อ 26, 29, 31, 34, 39

แบบทดสอบความยาว 60 ข้อ เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 4 ข้อ คือข้อ 3, 12, 18, 36 เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 200$ จำนวน 4 ข้อ คือข้อ 14, 18, 34, 49 เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 300$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันจำนวน 3 ข้อ คือข้อ 11, 12, 39 เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 100$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันจำนวน 3 ข้อ คือข้อ 51, 55, 59 เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 200$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันจำนวน 3 ข้อ คือข้อ 14, 54, 60 เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 300$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันจำนวน 5 ข้อ คือข้อ 10, 23, 39, 46, 55 เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 100$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันจำนวน 5 ข้อ คือข้อ 3, 49, 51, 55, 59 เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 200$ จำนวน 7 ข้อ คือ 3, 14, 18, 34, 37, 49, 54 เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 300$ จำนวน 7 ข้อ คือข้อ 11, 23, 39, 46, 49, 50, 55 และเมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 1000$ สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันจำนวน 5 ข้อ คือข้อ 19, 41, 44, 46, 49

ตอนที่ 5 การวิเคราะห์อำนาจการตรวจสอบและการจำแนกผิดพลาด

การวิเคราะห์ส่วนนี้ผู้วิจัยคำนวณอำนาจการตรวจสอบและการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอและแบบไม่สมำเสมอ ระหว่างวิธี MH และวิธี LR ปรากฏผลดังตาราง 20

ตาราง 20 การวิเคราะห์อำนาจการตรวจสอบและการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอและแบบไม่สม่ำเสมอ ระหว่างวิธี MH และวิธี LR

แบบทดสอบ ความยาว	กลุ่มตัวอย่าง ขนาด ($N_R : N_F$)	MH				LR			
		แบบ สม่ำเสมอ		แบบ ไม่สม่ำเสมอ		แบบ สม่ำเสมอ		แบบ ไม่สม่ำเสมอ	
		P	E1	P	E1	P	E1	P	E1
20 ข้อ	300 : 100	.00	.00	.00	.00	.00	.00	.00	.00
	300 : 200	.00	.00	.00	.00	.00	.00	.33	.00
	300 : 300	.33	.00	.00	.00	.33	.00	.00	.06
	500 : 100	.33	.00	.00	.12	.33	.06	.00	.00
	500 : 200	.33	.00	.33	.06	.33	.06	.33	.00
	500 : 300	.00	.00	.00	.00	.00	.00	.00	.00
	1000 : 100	.33	.00	.00	.06	.33	.06	.00	.00
	1000 : 200	.00	.00	.67	.00	.00	.00	.67	.00
	1000 : 300	.00	.00	.00	.06	.00	.06	.00	.00
	1000 : 1000	.33	.00	.67	.06	.33	.06	.67	.06
40 ข้อ	300 : 100	.00	.00	.00	.00	.00	.00	.20	.03
	300 : 200	.00	.00	.20	.03	.25	.06	.00	.11
	300 : 300	.25	.03	.20	.34	.25	.03	.20	.06
	500 : 100	.25	.03	.00	.06	.25	.06	.20	.00
	500 : 200	.25	.00	.20	.00	.25	.00	.20	.06
	500 : 300	.00	.00	.20	.00	.25	.00	.20	.03
	1000 : 100	.25	.00	.00	.00	.50	.03	.20	.03
	1000 : 200	.50	.00	.20	.03	.25	.03	.20	.14
	1000 : 300	.50	.03	.40	.09	.50	.00	.60	.00
	1000 : 1000	.50	.06	.60	.03	.50	.03	.60	.06

ตาราง 20 (ต่อ)

แบบทดสอบ ความยาว	กลุ่มตัวอย่าง ขนาด ($N_R : N_F$)	MH				LR			
		แบบ สม่ำเสมอ		แบบ ไม่สม่ำเสมอ		แบบ สม่ำเสมอ		แบบ ไม่สม่ำเสมอ	
		P	E1	P	E1	P	E1	P	E1
60 ข้อ	300 : 100	.17	.02	.08	.08	.17	.07	.08	.06
	300 : 200	.00	.00	.08	.02	.00	.04	.17	.04
	300 : 300	.17	.04	.17	.02	.17	.04	.17	.02
	500 : 100	.00	.02	.17	.02	.17	.04	.25	.00
	500 : 200	.17	.00	.17	.04	.17	.00	.17	.02
	500 : 300	.33	.02	.17	.02	.17	.04	.33	.02
	1000 : 100	.33	.02	.33	.06	.17	.00	.33	.02
	1000 : 200	.00	.00	.33	.00	.33	.02	.33	.06
	1000 : 300	.33	.00	.25	.02	.33	.02	.42	.04
	1000 : 1000	.67	.02	.50	.06	.50	.04	.25	.04

ผลการวิเคราะห์ตามตาราง 20 พบว่าอำนาจการตรวจสอบในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอด้วยวิธี MH เมื่อใช้แบบทดสอบความยาว 20 ข้อ และกลุ่มตัวอย่างทุกขนาด มีค่าอยู่ในช่วง .00 - .33 เมื่อใช้แบบทดสอบความยาว 40 ข้อ และกลุ่มตัวอย่างทุกขนาด มีค่าอยู่ในช่วง .00 - .50 และเมื่อใช้แบบทดสอบความยาว 60 ข้อ และกลุ่มตัวอย่างทุกขนาด มีค่าอยู่ในช่วง .00 - .67

การจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอด้วยวิธี MH เมื่อใช้แบบทดสอบความยาว 20 ข้อ และกลุ่มตัวอย่างทุกขนาด มีค่าเท่ากับ .00 เมื่อใช้แบบทดสอบความยาว 40 ข้อ และกลุ่มตัวอย่างทุกขนาด มีค่าอยู่ในช่วง .00 - .06 และเมื่อใช้แบบทดสอบความยาว 60 ข้อ และกลุ่มตัวอย่างทุกขนาด มีค่าอยู่ในช่วง .00 - .04

อำนาจการตรวจสอบในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอด้วยวิธี LR เมื่อใช้แบบทดสอบความยาว 20 ข้อ และกลุ่มตัวอย่างทุกขนาด มีค่าอยู่ในช่วง .00 - .33 เมื่อใช้แบบทดสอบความยาว 40 ข้อ และกลุ่มตัวอย่างทุกขนาด มีค่าอยู่ในช่วง .00 - .50 และเมื่อใช้แบบทดสอบความยาว 60 ข้อ และกลุ่มตัวอย่างทุกขนาด มีค่าอยู่ในช่วง .00 - .50

การจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอด้วยวิธี LR เมื่อใช้แบบทดสอบความยาว 20 ข้อ และกลุ่มตัวอย่างทุกขนาด มีค่าอยู่ในช่วง .00 - .06 เมื่อใช้แบบทดสอบความยาว 40 ข้อ และกลุ่มตัวอย่างทุกขนาด มีค่าอยู่ในช่วง .00 - .06 และเมื่อใช้แบบทดสอบความยาว 60 ข้อ และกลุ่มตัวอย่างทุกขนาด มีค่าอยู่ในช่วง .00 - .07

อำนาจการตรวจสอบในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอด้วยวิธี MH เมื่อใช้แบบทดสอบความยาว 20 ข้อ และกลุ่มตัวอย่างทุกขนาด มีค่าอยู่ในช่วง .00 - .67 เมื่อใช้แบบทดสอบความยาว 40 ข้อ และกลุ่มตัวอย่างทุกขนาด มีค่าอยู่ในช่วง .00 - .60 และเมื่อใช้แบบทดสอบความยาว 60 ข้อ และกลุ่มตัวอย่างทุกขนาด มีค่าอยู่ในช่วง .00 - .50

การจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สมมาตรด้วยวิธี MH โดยใช้แบบทดสอบความยาว 20 ข้อ และกลุ่มตัวอย่างทุกขนาด มีค่าอยู่ในช่วง .00 - .12 เมื่อใช้แบบทดสอบความยาว 40 ข้อ และกลุ่มตัวอย่างทุกขนาด มีค่าอยู่ในช่วง .00 - .34 และเมื่อใช้แบบทดสอบความยาว 60 ข้อ และกลุ่มตัวอย่างทุกขนาด มีค่าอยู่ในช่วง .00 - .08

อำนาจการตรวจสอบในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สมมาตรด้วยวิธี LR โดยใช้แบบทดสอบความยาว 20 ข้อ และกลุ่มตัวอย่างทุกขนาด มีค่าอยู่ในช่วง .00 - .67 เมื่อใช้แบบทดสอบความยาว 40 ข้อ และกลุ่มตัวอย่างทุกขนาด มีค่าอยู่ในช่วง .00 - .60 และเมื่อใช้แบบทดสอบความยาว 60 ข้อ และกลุ่มตัวอย่างทุกขนาด มีค่าอยู่ในช่วง .00 - .42

การจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สมมาตรด้วยวิธี LR โดยใช้แบบทดสอบความยาว 20 ข้อ และกลุ่มตัวอย่างทุกขนาด มีค่าอยู่ในช่วง .00 - .06 เมื่อใช้แบบทดสอบความยาว 40 ข้อ และกลุ่มตัวอย่างทุกขนาด มีค่าอยู่ในช่วง .00 - .14 และเมื่อใช้แบบทดสอบความยาว 60 ข้อ และกลุ่มตัวอย่างทุกขนาด มีค่าอยู่ในช่วง .00 - .06

ตอนที่ 6 การเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาด ระหว่างวิธี

แมนเทิล – เฮนส์เชล (MH) กับวิธีการถดถอยโลจิสติก (LR)

การวิเคราะห์ส่วนนี้ผู้วิจัยเปรียบเทียบความแตกต่างของอำนาจการตรวจสอบและการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน ระหว่างวิธี MH กับวิธี LR โดยการทดสอบด้วยสถิติ Z ที่ระดับนัยสำคัญ .05 โดยมีรายละเอียดดังนี้

6.1 การเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมมาตร ระหว่างวิธี MH และวิธี LR เมื่อใช้แบบทดสอบความยาวเท่ากัน ดังตาราง 21

ตาราง 21 การเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมมาตร ระหว่างวิธี MH และวิธี LR เมื่อใช้แบบทดสอบความยาวเท่ากัน

แบบทดสอบ	P			E1		
	MH	LR	Z - test	MH	LR	Z - test
ความยาว						
20 ข้อ	.17	.17	.00	.00	.03	1.74
40 ข้อ	.25	.30	.79	.01	.02	.46
60 ข้อ	.22	.22	.02	.01	.03	.81

* $p < .05$

ผลการวิเคราะห์ตามตาราง 21 พบว่า อำนาจการตรวจสอบและการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมมาตร ระหว่างวิธี MH และวิธี LR เมื่อใช้แบบทดสอบความยาว 20 ข้อ, 40 ข้อ และ 60 ข้อ มีค่าไม่แตกต่างกัน

6.2 การเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอ ระหว่างแบบทดสอบความยาวต่างกัน เมื่อใช้วิธีตรวจสอบวิธีเดียวกัน ดังแสดงในตาราง 22 – 25

ตาราง 22 การเปรียบเทียบอำนาจการตรวจสอบในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอ ระหว่างแบบทดสอบความยาวต่างกัน เมื่อตรวจสอบด้วยวิธี MH

แบบทดสอบความยาว	20 ข้อ	40 ข้อ	60 ข้อ
(P)	(.17)	(.25)	(.22)
20 ข้อ	-	1.48	.93
40 ข้อ		-	.55
60 ข้อ			-

* $p < .05$

ผลการวิเคราะห์ตามตาราง 22 พบว่า อำนาจการตรวจสอบในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอด้วยวิธี MH เมื่อใช้แบบทดสอบความยาว 20 ข้อ กับ 40 ข้อ, 20 ข้อ กับ 60 ข้อ และ 40 ข้อ กับ 60 ข้อ มีค่าไม่แตกต่างกัน

ตาราง 23 การเปรียบเทียบการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอ ระหว่างแบบทดสอบความยาวต่างกัน เมื่อตรวจสอบด้วยวิธี MH

แบบทดสอบความยาว	20 ข้อ	40 ข้อ	60 ข้อ
(E1)	(.00)	(.01)	(.14)
20 ข้อ	-	1.23	1.19
40 ข้อ		-	.06
60 ข้อ			-

* $p < .05$

ผลการวิเคราะห์ตามตาราง 23 พบว่า การจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอ ด้วยวิธี MH เมื่อใช้แบบทดสอบความยาว 20 ข้อ กับ 40 ข้อ, 20 ข้อ กับ 60 ข้อ และ 40 ข้อ กับ 60 ข้อ มีค่าไม่แตกต่างกัน

ตาราง 24 การเปรียบเทียบอำนาจการตรวจสอบในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอ
ระหว่างแบบทดสอบความยาวต่างกัน เมื่อตรวจสอบด้วยวิธี LR

แบบทดสอบความยาว (P)	20 ข้อ (.17)	40 ข้อ (.30)	60 ข้อ (.02)
20 ข้อ	-	2.26*	.95
40 ข้อ		-	1.32
60 ข้อ			-

* $p < .05$

ผลการวิเคราะห์ตามตาราง 24 พบว่า อำนาจการตรวจสอบในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอ ด้วยวิธี LR เมื่อใช้แบบทดสอบความยาว 20 ข้อ กับ 40 ข้อ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่เมื่อใช้ความยาวของแบบทดสอบ 20 ข้อ กับ 60 ข้อ และ 40 ข้อ กับ 60 ข้อ มีค่าไม่แตกต่างกัน

ตาราง 25 การเปรียบเทียบการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอ
ระหว่างแบบทดสอบความยาวต่างกัน เมื่อตรวจสอบด้วยวิธี LR

แบบทดสอบความยาว (E1)	20 ข้อ (.03)	40 ข้อ (.02)	60 ข้อ (.03)
20 ข้อ	-	.26	.04
40 ข้อ		-	.30
60 ข้อ			-

* $p < .05$

ผลการวิเคราะห์ตามตาราง 25 พบว่า การจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอ ด้วยวิธี LR เมื่อใช้แบบทดสอบความยาว 20 ข้อ กับ 40 ข้อ, 20 ข้อ กับ 60 ข้อ และ 40 ข้อ กับ 60 ข้อ มีค่าไม่แตกต่างกัน

6.3 การเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอ ระหว่างวิธี MH กับวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาดเท่ากัน ปรากฏผลดังตาราง 26

ตาราง 26 การเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอ ระหว่างวิธี MH และวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาดเท่ากัน

กลุ่มตัวอย่างขนาด ($N_R : N_F$)	P			E1		
	MH	LR	Z – test	MH	LR	Z – test
300 : 100	.06	.06	.00	.02	.03	.75
300 : 200	.00	.08	2.95*	.00	.03	1.84
300 : 300	.25	.25	.00	.03	.02	.00
500 : 100	.19	.25	.97	.02	.06	1.59
500 : 200	.25	.25	.00	.00	.02	1.42
500 : 300	.11	.14	.64	.01	.01	.47
1000 : 100	.30	.33	.46	.01	.03	1.23
1000 : 200	.17	.19	.49	.00	.02	1.30
1000 : 300	.28	.28	.00	.01	.03	.88
1000 : 1000	.50	.44	.80	.03	.04	.64

* $p < .05$

ผลการวิเคราะห์ตามตาราง 26 พบว่า อำนาจการตรวจสอบ ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอ ระหว่างวิธี MH และวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 200$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100$, $300 : 300$, $500 : 100$, $500 : 200$, $500 : 300$, $1000 : 100$, $1000 : 200$, $1000 : 300$ มีค่าไม่แตกต่างกัน

การจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอระหว่างวิธี MH และวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100$, $300 : 200$, $300 : 300$, $500 : 100$, $500 : 200$, $500 : 300$, $1000 : 100$, $1000 : 200$, $1000 : 300$ และ $1000 : 1000$ มีค่าไม่แตกต่างกัน

6.4 การเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอ ระหว่างกลุ่มตัวอย่างขนาดต่างกัน เมื่อใช้วิธีตรวจสอบวิธีเดียวกัน ปรากฏผลดังตาราง 27 – 30

ตาราง 27 การเปรียบเทียบอำนาจการตรวจสอบในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอ ระหว่างกลุ่มตัวอย่างขนาดต่างกัน เมื่อตรวจสอบด้วยวิธี MH

กลุ่มตัวอย่าง ขนาด ($N_R : N_F$) (P)	300 : 100 (.00)	300 : 200 (.25)	300 : 300 (.19)	500 : 100 (.25)	500 : 200 (.11)	500 : 300 (.30)	1000 : 100 (.17)	1000 : 200 (.28)	1000 : 300 (.28)	1000 : 1000 (.44)
300 : 100	-	2.42*	3.79*	2.92*	3.79*	1.36	4.54*	2.47*	4.17*	6.99*
300 : 200		-	5.35*	4.63*	5.35*	3.41*	5.98*	4.26*	5.57*	8.16*
300 : 300			-	.97	.00	2.58*	.84	1.45	.43	3.65*
500 : 100				-	.97	1.64	1.80	.49	1.39	4.56*
500 : 200					-	2.58*	.84	1.45	.43	3.65*
500 : 300						-	3.36*	1.16	2.98*	5.99*
1000 : 100							-	2.28*	.41	2.84*
1000 : 200								-	1.87	4.99*
1000 : 300									-	3.24*
1000 : 1000										-

* $p < .05$

ผลการวิเคราะห์ตามตาราง 27 อำนาจการตรวจสอบ ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอด้วยวิธี MH พบว่า

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100$ กับ $N_R : N_F = 300 : 200, 300 : 300, 500 : 100, 500 : 200, 1000 : 100, 1000 : 200, 1000 : 300$ และ $1000 : 1000$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 300$

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 200$ กับ $N_R : N_F = 300 : 100, 300 : 300, 500 : 100, 500 : 200, 500 : 300, 1000 : 100, 1000 : 200, 1000 : 300$ และ $1000 : 1000$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 300$ กับ $N_R : N_F = 300 : 100, 300 : 200, 500 : 300$ และ $1000 : 1000$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 100$ กับ $N_R : N_F = 300 : 100, 300 : 200$ และ $1000 : 1000$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 200$ กับ $N_R : N_F = 300 : 100$, $300 : 200$, $500 : 300$ และ $1000 : 1000$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 300$ กับ $N_R : N_F = 300 : 200$, $300 : 300$, $500 : 200$, $1000 : 100$, $1000 : 300$ และ $1000 : 1000$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 100$ กับ $N_R : N_F = 300 : 100$, $300 : 200$, $500 : 300$, $1000 : 200$ และ $1000 : 1000$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 200$ กับ $N_R : N_F = 300 : 100$, $300 : 200$, $1000 : 100$ และ $1000 : 1000$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 300$ กับ $N_R : N_F = 300 : 100$, $300 : 200$, $500 : 300$ และ $1000 : 1000$ พบว่า มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 1000$ กับกลุ่มตัวอย่างขนาดอื่น ๆ ทั้งหมดพบว่า มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05

ตาราง 28 การเปรียบเทียบการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอ
ระหว่างกลุ่มตัวอย่างขนาดต่างกัน เมื่อตรวจสอบด้วยวิธี MH

กลุ่มตัวอย่าง ขนาด ($N_R : N_F$) (E1)	300 : 100	300 : 200	300 : 300	500 : 100	500 : 200	500 : 300	1000 : 100	1000 : 200	1000 : 300	1000 : 1000
	(.02)	(.00)	(.03)	(.02)	(.00)	(.01)	(.01)	(.00)	(.01)	(.03)
300 : 100	-	1.30	.75	.00	1.30	.66	.66	1.30	.41	.49
300 : 200		-	1.84	1.30	****	.82	.82	****	1.00	1.65
300 : 300			-	.75	1.84	1.34	1.34	1.84	1.13	.27
500 : 100				-	1.30	.66	.66	1.30	.41	.49
500 : 200					-	.82	.82	****	1.00	1.65
500 : 300						-	.00	.82	.26	1.10
1000 : 100							-	.82	.26	1.10
1000 : 200								-	1.00	1.22
1000 : 300									-	.88
1000 : 1000										-

* $p < .05$

**** หมายถึง ไม่สามารถคำนวณค่าได้

ผลการวิเคราะห์ตามตาราง 28 การจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน
แบบสมำเสมอ ด้วยวิธี MH พบว่า

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100$ กับกลุ่มตัวอย่างขนาดอื่น ๆ ทั้งหมด มีค่า
ไม่แตกต่างกัน

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 200$ กับกลุ่มตัวอย่างขนาดอื่น ๆ ทั้งหมด มีค่า
ไม่แตกต่างกัน

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 300$ กับกลุ่มตัวอย่างขนาดอื่น ๆ ทั้งหมด มีค่า
ไม่แตกต่างกัน

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 100$ กับกลุ่มตัวอย่างขนาดอื่น ๆ ทั้งหมด มีค่า
ไม่แตกต่างกัน

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 200$ กับกลุ่มตัวอย่างขนาดอื่น ๆ ทั้งหมด มีค่า
ไม่แตกต่างกัน

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 300$ กับกลุ่มตัวอย่างขนาดอื่น ๆ ทั้งหมด มีค่า
ไม่แตกต่างกัน

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 100$ กับกลุ่มตัวอย่างขนาดอื่น ๆ ทั้งหมด มีค่า
ไม่แตกต่างกัน

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 200$ กับกลุ่มตัวอย่างขนาดอื่น ๆ ทั้งหมด มีค่าไม่แตกต่างกัน

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 300$ กับกลุ่มตัวอย่างขนาดอื่น ๆ ทั้งหมด มีค่าไม่แตกต่างกัน

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 1000$ กับกลุ่มตัวอย่างขนาดอื่น ๆ ทั้งหมด มีค่าไม่แตกต่างกัน

ตาราง 29 การเปรียบเทียบอำนาจการตรวจสอบในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมมาตรระหว่างกลุ่มตัวอย่างขนาดต่างกัน เมื่อตรวจสอบด้วยวิธี LR

กลุ่มตัวอย่าง ขนาด ($N_R : N_F$)	300 : 100 (.06)	300 : 200 (.08)	300 : 300 (.25)	500 : 100 (.25)	500 : 200 (.25)	500 : 300 (.14)	1000 : 100 (.33)	1000 : 200 (.19)	1000 : 300 (.28)	1000 : 1000 (.44)
300 : 100	-	.74	3.79*	3.79*	3.79*	1.98*	4.94*	2.92*	4.17*	6.31*
300 : 200		-	3.16	3.16*	3.16*	1.27	4.35*	2.25*	3.56*	5.78*
300 : 300			-	.00	.00	1.96	1.30	.96	.43	2.87*
500 : 100				-	.00	1.96	1.30	.96	.43	2.87*
500 : 200					-	1.96	1.30	.96	.43	2.87*
500 : 300						-	3.21*	1.01	2.38*	4.72*
1000 : 100							-	2.25*	.87	1.60
1000 : 200								-	1.39	3.79*
1000 : 300									-	2.45*
1000 : 1000										-

* $p < .05$

ผลการวิเคราะห์ตามตาราง 29 อำนาจการตรวจสอบ ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมมาตร ด้วยวิธี LR พบว่า

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 300, 500 : 100, 500 : 200, 500 : 300, 1000 : 100, 1000 : 200, 1000 : 300$ และ $1000 : 1000$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 200$

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 200$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 300, 500 : 100, 500 : 200, 1000 : 100, 1000 : 200, 1000 : 300$ และ $1000 : 1000$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 300$ กับกลุ่มตัวอย่างขนาด $N_R : N_F =$

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 1000$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100, 300 : 200, 300 : 300, 500 : 100, 500 : 200, 500 : 300, 1000 : 200$ และ $1000 : 300$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 100$

ตาราง 30 การเปรียบเทียบการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอ
ระหว่างกลุ่มตัวอย่างขนาดต่างกัน เมื่อตรวจสอบด้วยวิธี LR

กลุ่มตัวอย่าง ขนาด ($N_R : N_F$) (E1)	300 : 100 (.03)	300 : 200 (.03)	300 : 300 (.02)	500 : 100 (.06)	500 : 200 (.02)	500 : 300 (.01)	1000 : 100 (.03)	1000 : 200 (.02)	1000 : 300 (.03)	1000 : 1000 (.04)
300 : 100	-	.00	.43	.89	.58	.94	.13	.75	.27	.37
300 : 200		-	.43	.89	.58	.94	.13	.75	.27	.37
300 : 300			-	1.30	.16	.53	.29	.33	.15	.79
500 : 100				-	1.44	1.76	1.02	1.59	1.16	.53
500 : 200					-	.37	.45	.17	.31	.94
500 : 300						-	.81	.19	.68	1.28
1000 : 100							-	.62	.14	.50
1000 : 200								-	.49	1.10
1000 : 300									-	.64
1000 : 1000										-

*p < .05

ผลการวิเคราะห์ตามตาราง 30 พบว่า การจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอ ด้วยวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100, 300 : 200, 300 : 300, 500 : 100, 500 : 200, 500 : 300, 1000 : 100, 1000 : 200, 1000 : 300$ และ $1000 : 1000$ กับกลุ่มตัวอย่างขนาดอื่น ๆ ทั้งหมด มีค่าไม่แตกต่างกัน

6.5 การเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ระหว่างวิธี MH กับวิธี LR เมื่อใช้แบบทดสอบความยาวเท่ากัน ปรากฏผล ดังตาราง 31

ตาราง 31 การเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ระหว่างวิธี MH กับวิธี LR เมื่อใช้แบบทดสอบความยาวเท่ากัน

แบบทดสอบ ความยาว	P			E1		
	MH	LR	Z - test	MH	LR	Z - test
20 ข้อ	.17	.20	.60	.04	.01	1.11
40 ข้อ	.20	.26	1.01	.06	.05	.19
60 ข้อ	.23	.25	.42	.03	.03	.08

*p < .05

ผลการวิเคราะห์ตามตาราง 31 พบว่า อำนาจการตรวจสอบและการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอระหว่างวิธี MH กับวิธี LR เมื่อใช้แบบทดสอบความยาว 20 ข้อ, 40 ข้อ และ 60 ข้อ มีค่าไม่แตกต่างกัน

6.6 การเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ระหว่างแบบทดสอบความยาวต่างกัน เมื่อใช้วิธีตรวจสอบวิธีเดียวกัน ดังแสดง ในตาราง 32 – 35

ตาราง 32 การเปรียบเทียบอำนาจการตรวจสอบในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ระหว่างแบบทดสอบความยาวต่างกัน เมื่อตรวจสอบด้วยวิธี MH

แบบทดสอบความยาว (P)	20 ข้อ (.17)	40 ข้อ (.20)	60 ข้อ (.23)
20 ข้อ	-	.60	1.03
40 ข้อ		-	.43
60 ข้อ			-

*p < .05

ผลการวิเคราะห์ตามตาราง 32 พบว่า อำนาจการตรวจสอบ ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ด้วยวิธี MH เมื่อใช้แบบทดสอบความยาว 20 ข้อ กับ 40 ข้อ, 20 ข้อ กับ 60 ข้อ และ 40 ข้อ กับ 60 ข้อ มีค่าไม่แตกต่างกัน

ตาราง 33 การเปรียบเทียบการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ
ระหว่างแบบทดสอบความยาวต่างกัน เมื่อตรวจสอบด้วยวิธี MH

แบบทดสอบความยาว (E1)	20 ข้อ (.04)	40 ข้อ (.05)	60 ข้อ (.03)
20 ข้อ	-	.74	.08
40 ข้อ		-	.06
60 ข้อ			-

*p < .05

ผลการวิเคราะห์ตามตาราง 33 พบว่า การจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ด้วยวิธี MH เมื่อใช้แบบทดสอบความยาว 20 ข้อ กับ 40 ข้อ, 20 ข้อ กับ 60 ข้อ และ 40 ข้อ กับ 60 ข้อ มีค่าไม่แตกต่างกัน

ตาราง 34 การเปรียบเทียบอำนาจการตรวจสอบในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ
ระหว่างแบบทดสอบความยาวต่างกัน เมื่อตรวจสอบด้วยวิธี LR

แบบทดสอบความยาว (P)	20 ข้อ (.20)	40 ข้อ (.26)	60 ข้อ (.25)
20 ข้อ	-	1.01	.85
40 ข้อ		-	.16
60 ข้อ			-

*p < .05

ผลการวิเคราะห์ตามตาราง 34 พบว่า อำนาจการตรวจสอบ ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ด้วยวิธี LR เมื่อใช้แบบทดสอบความยาว 20 ข้อ กับ 40 ข้อ, 20 ข้อ กับ 60 ข้อ และ 40 ข้อ กับ 60 ข้อ มีค่าไม่แตกต่างกัน

ตาราง 35 การเปรียบเทียบการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ
ระหว่างแบบทดสอบความยาวต่างกัน เมื่อตรวจสอบด้วยวิธี LR

แบบทดสอบความยาว (E1)	20 ข้อ (.01)	40 ข้อ (.05)	60 ข้อ (.03)
20 ข้อ	-	1.61	.96
40 ข้อ		-	.71
60 ข้อ			-

*p < .05

ผลการวิเคราะห์ตามตาราง 35 พบว่า การจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ด้วยวิธี LR เมื่อใช้แบบทดสอบความยาว 20 ข้อ กับ 40 ข้อ, 20 ข้อ กับ 60 ข้อ และ 40 ข้อ กับ 60 ข้อ มีค่าไม่แตกต่างกัน

6.7 การเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ระหว่างวิธี MH กับวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาดเท่ากัน ดังแสดงในตาราง 36

ตาราง 36 การเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ระหว่างวิธี MH และวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาดเท่ากัน

กลุ่มตัวอย่างขนาด ($N_R : N_F$)	P			E1		
	MH	LR	Z - test	MH	LR	Z - test
300 : 100	.03	.09	1.98*	.03	.04	.00
300 : 200	.09	.17	1.54	.02	.07	1.48
300 : 300	.12	.12	.00	.13	.05	2.01*
500 : 100	.06	.15	2.16*	.07	.00	2.63*
500 : 200	.23	.23	.00	.03	.06	.99
500 : 300	.12	.18	1.06	.02	.02	.33
1000 : 100	.11	.18	1.35	.04	.02	.99
1000 : 200	.40	.40	.00	.01	.08	2.31*
1000 : 300	.22	.34	1.94	.06	.01	1.67
1000 : 1000	.59	.51	1.18	.05	.05	.11

*p < .05

ผลการวิเคราะห์ตามตาราง 36 พบว่า อำนาจการตรวจสอบในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ระหว่างวิธี MH และวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100$ และ $500 : 100$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 200, 300 : 300, 500 : 200, 500 : 300, 1000 : 100, 1000 : 200, 1000 : 300$ และ $1000 : 1000$ มีค่าไม่แตกต่างกัน

การจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอระหว่างวิธี MH และวิธี LR พบว่า เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 300, 500 : 100$ และ $1000 : 200$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100, 300 : 200, 500 : 200, 500 : 300, 1000 : 100, 1000 : 300$ และ $1000 : 1000$ พบว่ามีค่าไม่แตกต่างกัน

6.8 การเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ระหว่างกลุ่มตัวอย่างขนาดต่างกัน เมื่อใช้วิธีตรวจสอบวิธีเดียวกัน ดังแสดงในตาราง 37 – 40

ตาราง 37 การเปรียบเทียบอำนาจการตรวจสอบในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ระหว่างกลุ่มตัวอย่างขนาดต่างกัน เมื่อตรวจสอบด้วยวิธี MH

กลุ่มตัวอย่าง ขนาด ($N_R : N_F$)	300 : 100	300 : 200	300 : 300	500 : 100	500 : 200	500 : 300	1000 : 100	1000 : 200	1000 : 300	1000 : 1000
(P)	(.03)	(.09)	(.12)	(.06)	(.23)	(.12)	(.11)	(.40)	(.22)	(.59)
300 : 100	-	1.98*	2.59*	1.06	4.34*	2.59*	2.33*	6.44*	4.11*	8.62*
300 : 200		-	.68	.98	2.68*	.68	.39	5.03*	2.41*	7.40*
300 : 300			-	1.65	2.03*	.00	.29	4.45*	1.76	6.89*
500 : 100				-	3.55*	1.65	1.36	5.78*	3.29*	8.06*
500 : 200					-	2.03*	2.31*	2.53*	.28	5.13*
500 : 300						-	.29	4.45*	1.76	6.89*
1000 : 100							-	4.70*	2.04*	7.12*
1000 : 200								-	2.81*	2.69*
1000 : 300									-	5.38*
1000 : 1000										-

*p < .05

ผลการวิเคราะห์ตามตาราง 37 อำนาจการตรวจสอบ ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน แบบไม่สมมาตร ด้วยวิธี MH พบว่า

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100$ กับ $N_R : N_F = 300 : 200, 300 : 300, 500 : 200, 500 : 300, 1000 : 100, 1000 : 200, 1000 : 300$ และ $1000 : 1000$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 100$

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 200$ กับ $N_R : N_F = 300 : 100, 500 : 200, 1000 : 200, 1000 : 300$ และ $1000 : 1000$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 300$ กับ $N_R : N_F = 300 : 100, 500 : 200, 1000 : 200$ และ $1000 : 1000$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 100$ กับ $N_R : N_F = 500 : 200, 1000 : 200, 1000 : 300$ และ $1000 : 1000$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 200$ กับ $N_R : N_F = 300 : 100, 300 : 200, 300 : 300, 500 : 100, 500 : 300, 1000 : 100, 1000 : 200$ และ $1000 : 1000$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 300$

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 300$ กับ $N_R : N_F = 300 : 100, 500 : 200, 1000 : 200$ และ $1000 : 1000$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 100$ กับ $N_R : N_F = 300 : 100, 500 : 200, 1000 : 200, 1000 : 300$ และ $1000 : 1000$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 200$ กับกลุ่มตัวอย่างขนาดอื่นๆ ทั้งหมด มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 300$ กับ $N_R : N_F = 300 : 100, 300 : 200, 500 : 100, 1000 : 100, 1000 : 200$ และ $1000 : 1000$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 1000$ กับกลุ่มตัวอย่างขนาดอื่นๆ ทั้งหมด มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05

ตาราง 38 การเปรียบเทียบการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ
ระหว่างกลุ่มตัวอย่างขนาดต่างกัน เมื่อตรวจสอบด้วยวิธี MH

กลุ่มตัวอย่าง ขนาด ($N_R : N_F$)	300 : 100	300 : 200	300 : 300	500 : 100	500 : 200	500 : 300	1000 : 100	1000 : 200	1000 : 300	1000 : 1000
(E1)	(.03)	(.02)	(.13)	(.07)	(.03)	(.02)	(.04)	(.01)	(.06)	(.05)
300 : 100	-	.43	2.43*	1.08	.00	.43	.25	1.13	.80	.59
300 : 200		-	2.78*	1.48	.43	.00	.67	.73	1.21	1.00
300 : 300			-	1.44	2.43*	2.78*	2.22*	3.27*	1.71	1.91
500 : 100				-	1.08	1.48	.84	2.09*	.29	.50
500 : 200					-	.43	.25	1.13	.79	.59
500 : 300						-	.67	.73	1.21	1.00
1000 : 100							-	1.36	.55	.34
1000 : 200								-	1.84	1.66
1000 : 300									-	.21
1000 : 1000										-

* $p < .05$

ผลการวิเคราะห์ตามตาราง 38 การจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ด้วยวิธี MH พบว่า

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 300$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 200$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 300$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 300$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100, 300 : 200, 500 : 200, 500 : 300, 1000 : 100$ และ $1000 : 200$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 100$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 200$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 200$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 300$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 300$ กับกลุ่มตัวอย่างตัวอย่าง $N_R : N_F = 300 : 300$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 100$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 300$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 200$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 300$ และ $500 : 100$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 300$ และ $1000 : 1000$ กับกลุ่มตัวอย่างขนาดอื่น ๆ ทั้งหมด มีค่าไม่แตกต่างกัน

ตาราง 39 การเปรียบเทียบอำนาจการตรวจสอบในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ระหว่างกลุ่มตัวอย่างขนาดต่างกัน เมื่อตรวจสอบด้วยวิธี LR

กลุ่มตัวอย่าง ขนาด ($N_R : N_F$) (P)	300 : 100 (.09)	300 : 200 (.17)	300 : 300 (.12)	500 : 100 (.15)	500 : 200 (.23)	500 : 300 (.18)	1000 : 100 (.18)	1000 : 200 (.40)	1000 : 300 (.34)	1000 : 1000 (.51)
300 : 100	-	1.54	.68	1.23	2.68*	1.73	1.73	5.03*	4.23*	6.38*
300 : 200		-	.87	.32	1.18	.19	.19	3.56*	2.82*	5.09*
300 : 300			-	.55	2.03*	1.06	1.06	4.45*	3.63*	5.84*
500 : 100				-	1.50	.51	.51	3.96*	3.12*	5.37*
500 : 200					-	.99	.99	2.53*	1.67	4.00*
500 : 300						-	.00	3.49*	2.64*	4.92*
1000 : 100							-	3.49*	2.64*	4.92*
1000 : 200								-	.88	1.52
1000 : 300									-	2.39*
1000 : 1000										-

*p < .05

ผลการวิเคราะห์ตามตาราง 39 อำนาจการตรวจสอบ ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน แบบไม่สม่ำเสมอ ด้วยวิธี LR พบว่า

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 200$, $1000 : 200$, $1000 : 300$ และ $1000 : 1000$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ

ตาราง 40 การเปรียบเทียบการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอระหว่างกลุ่มตัวอย่างขนาดต่างกัน เมื่อตรวจสอบด้วยวิธี LR

กลุ่มตัวอย่าง ขนาด ($N_R : N_F$) (E1)	300 : 100	300 : 200	300 : 300	500 : 100	500 : 200	500 : 300	1000 : 100	1000 : 200	1000 : 300	1000 : 1000
	(.04)	(.07)	(.05)	(.00)	(.06)	(.02)	(.02)	(.08)	(.01)	(.05)
300 : 100	-	.73	.12	2.10*	.63	1.10	1.10	.99	1.28	.33
300 : 200		-	.61	2.63*	.09	1.77	1.77	.27	1.93	.39
300 : 300			-	2.19*	.51	1.21	1.21	.88	1.38	.21
500 : 100				-	2.56*	1.30	1.30	2.82*	1.16	2.34*
500 : 200					-	1.68	1.68	.37	1.84	.30
500 : 300						-	.00	2.01*	.19	1.41
1000 : 100							-	2.01*	.19	1.41
1000 : 200								-	2.16	.67
1000 : 300									-	1.58
1000 : 1000										-

* $p < .05$

ผลการวิเคราะห์ตามตาราง 40 การจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ด้วยวิธี LR พบว่า

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 100$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 200$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 100$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 300$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 100$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 100$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100, 300 : 200, 300 : 300, 500 : 200, 1000 : 200$ และ $1000 : 1000$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 200$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 100$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 300$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 200$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่นๆ ที่เหลือ

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 100$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 200$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่นๆ ที่เหลือ

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 200$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 100$, $500 : 300$ และ $1000 : 100$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่นๆ ที่เหลือ

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 300$ กับกลุ่มตัวอย่างขนาดอื่นๆ มีค่าไม่แตกต่างกัน

เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 1000$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 100$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่นๆ ที่เหลือ

บทที่ 5

สรุป อภิปรายผล และข้อเสนอแนะ

การศึกษาในครั้งนี้ มีความมุ่งหมาย เพื่อเปรียบเทียบอำนาจการตรวจสอบ และการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอและแบบไม่สม่ำเสมอ ระหว่างวิธี MH และวิธี LR โดยใช้ความยาวแบบทดสอบและขนาดกลุ่มตัวอย่างต่างกัน ซึ่งมีความมุ่งหมายเฉพาะในการศึกษาดังนี้

1. เพื่อเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอ ระหว่างวิธี MH และวิธี LR เมื่อใช้แบบทดสอบความยาวเท่ากัน
2. เพื่อเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอ ระหว่างแบบทดสอบความยาวต่างกัน เมื่อใช้วิธีตรวจสอบวิธีเดียวกัน
3. เพื่อเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอระหว่างวิธี MH และวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาดเท่ากัน
4. เปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอ ระหว่างกลุ่มตัวอย่างขนาดต่างกัน เมื่อใช้วิธีตรวจสอบวิธีเดียวกัน
5. เพื่อเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอระหว่างวิธี MH และวิธี LR เมื่อใช้แบบทดสอบความยาวเท่ากัน
6. เพื่อเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ระหว่างแบบทดสอบความยาวต่างกัน เมื่อใช้วิธีตรวจสอบวิธีเดียวกัน
7. เพื่อเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอระหว่างวิธี MH และวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาดเท่ากัน
8. เพื่อเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ระหว่างกลุ่มตัวอย่างขนาดต่างกัน เมื่อใช้วิธีตรวจสอบวิธีเดียวกัน

กลุ่มตัวอย่างที่ใช้ในการศึกษานี้ เป็นนักเรียนชายและหญิง ชั้นมัธยมศึกษาปีที่ 3 ภาคเรียนที่ 2 ปีการศึกษา 2542 ของโรงเรียนมัธยมศึกษาสังกัดกรมสามัญศึกษา ส่วนกลาง กลุ่มที่ 3 ซึ่งได้มาโดยการสุ่มแบบแบ่งชั้น (Stratified Random Sampling) มีขนาดโรงเรียนเป็นชั้น (Strata) โรงเรียนเป็นหน่วยการสุ่ม (Sampling Unit) ได้จำนวนโรงเรียน 8 โรงเรียน จำนวนห้องเรียน 71 ห้องเรียน นักเรียนทั้งหมด 3,242 คน จำแนกเป็นนักเรียนชาย 1,282 คน และนักเรียนหญิง 1,960 คน

เครื่องมือที่ใช้ในการศึกษาค้นคว้าในครั้งนี้เป็นแบบทดสอบวัดความสามารถทางสมองที่ผู้วิจัยได้สร้างขึ้นโดยพัฒนาและปรับปรุงตามแนวโครงสร้างของแบบทดสอบทางสมองของโอตีส-เลนนอน พอร์ม เจ ระดับสูง (Otis-Lennon Mental Ability Test Advanced Level Form J) จำนวน 1 ฉบับ เป็นแบบทดสอบวัดความสามารถทั่วไป ซึ่งแบ่งออกเป็นสามด้าน คือ วัดความเข้าใจด้านภาษา 30 ข้อ เหตุผลด้านภาษา 30 ข้อ และเหตุผลด้านภาพ 20 ข้อ รวม 80 ข้อ ซึ่งเป็นแบบทดสอบแบบเลือกตอบ 5 ตัวเลือก

ในการดำเนินการศึกษาค้นคว้า ผู้วิจัยได้ดำเนินการดังนี้

1. ดำเนินการสร้างแบบทดสอบความสามารถทางสมองตามแนวโครงสร้างของโอดิส-เลนนอน และคัดเลือกข้อสอบที่มีคุณภาพโดยคัดเฉพาะข้อที่มีค่าความยากง่ายระหว่าง .20 - .80 และค่าอำนาจจำแนกตั้งแต่ .20 ขึ้นไป
2. ดำเนินการเก็บรวบรวมข้อมูลโดยการนำแบบทดสอบไปทดสอบกับนักเรียนที่เป็นกลุ่มตัวอย่าง
3. นำกระดาษคำตอบของนักเรียนทั้งหมดมาตรวจความเรียบร้อยแล้วตรวจให้คะแนน(ตอบถูกได้ 1 และตอบผิดได้ 0) แล้วสร้างแฟ้มคำตอบของนักเรียน โดยแยกเป็นกลุ่มอ้างอิง(เพศหญิง) 1 แฟ้ม จำนวน 1000 คน และกลุ่มเปรียบเทียบ(เพศชาย) 1 แฟ้ม จำนวน 1000 คน
4. การจัดกระทำข้อมูลเป็นการเตรียมข้อมูลเพื่อให้ได้ตามเงื่อนไขที่ต้องการศึกษาโดยสุ่มความยาวแบบทดสอบจาก 80 ข้อ ให้ได้แบบทดสอบความยาว 3 ขนาด คือ 60 ข้อ 40 ข้อ และ 20 ข้อ แล้วสุ่มกลุ่มตัวอย่างขนาดกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ 10 กลุ่ม คือ 300 : 100 , 300 : 200, 300 : 300, 500 : 100, 500 : 200, 500 : 300, 1000 : 100, 1000 : 200, 1000 : 300 และ 1000 : 1000 ดังนั้นจะได้ข้อมูลที่จะนำไปใช้ในการศึกษาครั้งนี้ทั้งหมด 30 เงื่อนไข
5. การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธีทฤษฎีการตอบข้อสอบ โดยมีขั้นตอนดังนี้
 - 5.1 วิเคราะห์องค์ประกอบของแบบทดสอบความยาว 20 ข้อ, 40 ข้อ และ 60 ข้อ โดยใช้โปรแกรม SPSS for windows version 7.5
 - 5.2 ประเมินค่าพารามิเตอร์ของข้อสอบ โดยใช้โปรแกรม BILOG version 3.04
 - 5.3 เทียบมาตรของค่าพารามิเตอร์ของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบโดยใช้โปรแกรม EQUATE version 2.0
 - 5.4 วิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันโดยการวัดพื้นที่ของราชู โดยใช้โปรแกรม IRTDIF version 1.0
6. การวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันด้วยวิธี MH โดยใช้โปรแกรม MHDIF version 1.0
7. การวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันด้วยวิธี LR โดยใช้โปรแกรม SPSS/PC⁺ version 4.01
8. คำนวณค่าอำนาจการตรวจสอบและการจำแนกผิดพลาด
9. เปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาด ระหว่างวิธี MH และวิธี LR โดยการทดสอบด้วยค่าสถิติ Z ที่ระดับนัยสำคัญ .05

สรุปผลการวิจัย

1. ผลการเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอ ระหว่างวิธี MH และวิธี LR เมื่อใช้แบบทดสอบความยาวเท่ากัน ปรากฏผลดังนี้

1.1 อำนาจการตรวจสอบระหว่างวิธี MH และวิธี LR เมื่อใช้แบบทดสอบความยาว 20 ข้อ, 40 ข้อ และ 60 ข้อ พบว่า มีค่าไม่แตกต่างกัน

1.2 การจำแนกผิดพลาดระหว่างวิธี MH และวิธี LR เมื่อใช้แบบทดสอบความยาว 20 ข้อ, 40 ข้อ และ 60 ข้อ พบว่า มีค่าไม่แตกต่างกัน

2. ผลการเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอ ระหว่างแบบทดสอบความยาวต่างกันเมื่อใช้วิธีตรวจสอบวิธีเดียวกัน ปรากฏผลดังนี้

2.1 อำนาจการตรวจสอบของวิธี MH เมื่อใช้แบบทดสอบความยาว 20 ข้อ กับ 40 ข้อ, 20 ข้อ กับ 60 ข้อ และ 40 ข้อ กับ 60 ข้อ พบว่ามีค่าไม่แตกต่างกัน ส่วนอำนาจการตรวจสอบของวิธี LR เมื่อใช้แบบทดสอบความยาว 20 ข้อ กับ 40 ข้อ พบว่ามีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่เมื่อใช้แบบทดสอบความยาว 20 กับ 60 ข้อ และ 40 ข้อ กับ 60 ข้อ พบว่ามีค่าไม่แตกต่างกัน

2.2 การจำแนกผิดพลาดของวิธี MH เมื่อใช้แบบทดสอบความยาว 20 ข้อ กับ 40 ข้อ, 20 ข้อ กับ 60 ข้อ และ 40 ข้อ กับ 60 ข้อ พบว่ามีค่าไม่แตกต่างกัน ส่วนการจำแนกผิดพลาดของวิธี LR เมื่อใช้แบบทดสอบความยาว 20 ข้อ กับ 40 ข้อ, 20 ข้อ กับ 60 ข้อ และ 40 ข้อ กับ 60 ข้อ พบว่ามีค่าไม่แตกต่างกัน

3. ผลการเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอระหว่างวิธี MH และวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาดเท่ากัน ปรากฏผลดังนี้

3.1 อำนาจการตรวจสอบระหว่างวิธี MH และวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 200$ พบว่ามีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100, 300 : 300, 500 : 100, 500 : 200, 500 : 300, 1000 : 100, 1000 : 200, 1000 : 300$ และ $1000 : 1000$ พบว่ามีค่าไม่แตกต่างกัน

3.2 การจำแนกผิดพลาดระหว่างวิธี MH และวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100, 300 : 200, 300 : 300, 500 : 100, 500 : 200, 500 : 300, 1000 : 100, 1000 : 200, 1000 : 300$ และ $1000 : 1000$ พบว่ามีค่าไม่แตกต่างกัน

4. ผลการเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอ ระหว่างกลุ่มตัวอย่างขนาดต่างกันเมื่อใช้วิธีตรวจสอบวิธีเดียวกัน ปรากฏผลดังนี้

4.1 อำนาจการตรวจสอบของวิธี MH เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100$ กับกลุ่มตัวอย่างขนาดอื่น ๆ ยกเว้น $N_R : N_F = 500 : 300$ พบว่ามีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 200$ กับกลุ่มตัวอย่างขนาดอื่น ๆ ทั้งหมด พบว่ามีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 300$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100, 300 : 200, 500 : 300$ และ $1000 : 1000$ พบว่ามีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 100$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100, 300 : 200$ และ $1000 : 1000$ พบว่ามีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 200$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100, 300 : 200, 500 : 300$ และ $1000 : 1000$ พบว่ามีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับ

[illegible]

กลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 1000$ กับกลุ่มตัวอย่างขนาดอื่น ๆ ยกเว้นกลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 100$ พบว่ามีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05

4.4 การจำแนกผิดพลาดของวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100$, $300 : 200$, $300 : 300$, $500 : 100$, $500 : 200$, $500 : 300$, $1000 : 100$, $1000 : 200$, $1000 : 300$ และ $1000 : 1000$ กับกลุ่มตัวอย่างขนาดอื่น ๆ ทั้งหมด พบว่ามีค่าไม่แตกต่างกัน

5. ผลการเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ระหว่างวิธี MH และวิธี LR เมื่อใช้แบบทดสอบความยาวเท่ากัน ปรากฏผลดังนี้

5.1 อำนาจการตรวจสอบระหว่างวิธี MH และวิธี LR เมื่อใช้แบบทดสอบความยาว 20 ข้อ, 40 ข้อ และ 60 ข้อ พบว่ามีค่าไม่แตกต่างกัน

5.2 การจำแนกผิดพลาดระหว่างวิธี MH และวิธี LR เมื่อใช้แบบทดสอบความยาว 20 ข้อ, 40 ข้อ และ 60 ข้อ พบว่ามีค่าไม่แตกต่างกัน

6. ผลการเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ระหว่างแบบทดสอบความยาวต่างกัน เมื่อใช้วิธีตรวจสอบวิธีเดียวกัน ปรากฏผลดังนี้

6.1 อำนาจการตรวจสอบของวิธี MH เมื่อใช้แบบทดสอบความยาว 20 ข้อ กับ 40 ข้อ, 20 กับ 60 ข้อ และ 40 ข้อ กับ 60 ข้อ พบว่ามีค่าไม่แตกต่างกัน ส่วนอำนาจการตรวจสอบของวิธี LR เมื่อใช้แบบทดสอบความยาว 20 ข้อ กับ 40 ข้อ, 20 กับ 60 ข้อ และ 40 ข้อ กับ 60 ข้อ พบว่ามีค่าไม่แตกต่างกัน

6.2 การจำแนกผิดพลาดของวิธี MH เมื่อใช้แบบทดสอบความยาว 20 ข้อ กับ 40 ข้อ, 20 กับ 60 ข้อ และ 40 ข้อ กับ 60 ข้อ พบว่ามีค่าไม่แตกต่างกัน ส่วนการจำแนกผิดพลาดของวิธี LR เมื่อใช้แบบทดสอบความยาว 20 ข้อ กับ 40 ข้อ, 20 กับ 60 ข้อ และ 40 ข้อ กับ 60 ข้อ พบว่ามีค่าไม่แตกต่างกัน

7. ผลการเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ระหว่างวิธี MH และวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาดเท่ากัน ปรากฏผลดังนี้

7.1 อำนาจการตรวจสอบระหว่างวิธี MH และวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100$ และ $500 : 100$ พบว่ามีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 200$, $300 : 300$, $500 : 200$, $500 : 300$, $1000 : 100$, $1000 : 200$, $1000 : 300$ และ $1000 : 1000$ พบว่ามีค่าไม่แตกต่างกัน

7.2 การจำแนกผิดพลาดระหว่างวิธี MH และวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 300$, $500 : 100$ และ $1000 : 200$ พบว่ามีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100$, $300 : 200$, $500 : 100$, $500 : 200$, $500 : 300$, $1000 : 100$, $1000 : 300$ และ $1000 : 1000$ พบว่ามีค่าไม่แตกต่างกัน

8.2 การจำแนกผิดพลาดของวิธี MH เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 300$ พบว่า มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่นๆ ที่เหลือ เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 200$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 300$ พบว่ามีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่นๆ ที่เหลือ เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 300$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100, 300 : 200, 500 : 200, 500 : 300, 1000 : 100$ และ $1000 : 200$ พบว่ามีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่นๆ ที่เหลือ เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 100$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 200$ พบว่ามีค่า

[illegible]

8.4 การจำแนกผิดพลาดของวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 100$ พบว่ามีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 200$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 100$ พบว่ามีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 300$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 100$ พบว่ามีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 100$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100, 300 : 200, 300 : 300, 500 : 200, 1000 : 200$ และ $1000 : 1000$ พบว่ามีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 200$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 100$ พบว่ามีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 300$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 200$ พบว่ามีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 100$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 200$ พบว่ามีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 200$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 100, 500 : 300$ และ $1000 : 100$ พบว่ามีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 1000$ กับกลุ่มตัวอย่างขนาด $N_R : N_F = 500 : 100$ พบว่ามีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างกับกลุ่มตัวอย่างขนาดอื่น ๆ ที่เหลือ

อภิปรายผลการวิจัย

1. การเปรียบเทียบอำนาจการตรวจสอบ และการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอ ระหว่างวิธี MH และวิธี LR เมื่อใช้แบบทดสอบความยาวเท่ากัน

1.1 อำนาจการตรวจสอบระหว่างวิธี MH และวิธี LR เมื่อใช้แบบทดสอบความยาว 20 ข้อ, 40 ข้อ และ 60 ข้อ มีค่าไม่แตกต่างกัน ซึ่งขัดแย้งกับสมมติฐานในข้อที่ 1 แต่ผลการศึกษาดังกล่าวสอดคล้องกับผลการศึกษาของโรเจอร์สและสวามินาธาน (Rogers and Swaminathan. 1993) ทั้งนี้เนื่องจากวิธี MH ถูกออกแบบมาเพื่อวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอ ชนิดที่มีโครงสร้างข้อสอบ (Item Characteristic Curves : ICCs) ของผู้สอบสองกลุ่มขนานกันโดยเฉพาะ (Li and Stout. 1996) ซึ่งมีผลทำให้การตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันมีลักษณะดังกล่าวมีความแม่นยำสูง ส่วนวิธี LR สามารถใช้โมเดลคำนวณข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอได้อย่างมีประสิทธิภาพ (Swaminathan and Rogers. 1990) ดังนั้นผลการวิเคราะห์อำนาจการตรวจสอบระหว่างวิธี MH กับวิธี LR เมื่อใช้แบบทดสอบความยาวเท่ากัน จึงมีค่าใกล้เคียงกัน

1.2 การจำแนกผิดพลาดระหว่างวิธี MH และวิธี LR เมื่อใช้แบบทดสอบความยาว 20 ข้อ, 40 ข้อ และ 60 ข้อ มีค่าไม่แตกต่างกัน ซึ่งขัดแย้งกับสมมติฐานในข้อที่ 1 ทั้งนี้เนื่องจากในการวิเคราะห์ด้วยวิธี MH ในครั้งนี้ผู้วิจัยได้ใช้การวิเคราะห์แบบสองขั้นตอน (Two-stage) โดยขั้นตอนแรกใช้คะแนนรวมของแบบทดสอบทั้งฉบับเป็นเกณฑ์การจับคู่กลุ่มผู้สอบย่อยสองกลุ่ม แล้ววิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ เมื่อพบว่าข้อสอบข้อใดที่ทำหน้าที่ต่างกัน จะนำคะแนนของข้อสอบข้อนั้นออกจากคะแนนรวมของผู้สอบแต่ละคน และขั้นตอนที่สองใช้คะแนนรวมของแบบทดสอบที่นำเอาข้อสอบที่ทำหน้าที่ต่างกันซึ่งตรวจพบในขั้นตอนแรกออกไปแล้วเพื่อวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบซ้ำอีกครั้งหนึ่ง การวิเคราะห์ในลักษณะดังกล่าวจะทำให้เกณฑ์การจับคู่ของกลุ่มผู้สอบมีความบริสุทธิ์ (Holland and Thayer, 1988) ซึ่งจะมีผลทำให้การตรวจสอบมีความถูกต้องแม่นยำสูง ส่วนวิธี LR สามารถใช้โมเดลวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอได้อย่างมีประสิทธิภาพ (Swaminathan and Rogers, 1990) ดังนั้นผลการวิเคราะห์การจำแนกผิดพลาดระหว่างวิธี MH และวิธี LR เมื่อใช้แบบทดสอบความยาวเท่ากันจึงมีค่าใกล้เคียงกัน

2. การเปรียบเทียบอำนาจการตรวจสอบ และ การจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอ ระหว่างแบบทดสอบความยาวเท่ากัน เมื่อใช้วิธีตรวจสอบวิธีเดียวกัน

2.1 อำนาจการตรวจสอบของวิธี MH เมื่อใช้แบบทดสอบความยาว 20 ข้อ กับ 40 ข้อ, 20 ข้อ กับ 60 ข้อ และ 40 ข้อ กับ 60 ข้อ มีค่าไม่แตกต่างกัน ส่วนอำนาจการตรวจสอบของวิธี LR เมื่อใช้แบบทดสอบความยาว 20 ข้อ กับ 40 ข้อ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่เมื่อใช้แบบทดสอบความยาว 20 ข้อ กับ 60 ข้อ และ 40 ข้อ กับ 60 ข้อ มีค่าไม่แตกต่างกัน ผลการศึกษาดังกล่าวมีเพียงกรณีเดียว คือ อำนาจการตรวจสอบของวิธี LR เมื่อใช้แบบทดสอบความยาว 20 ข้อ กับ 40 ข้อ ซึ่งสอดคล้องกับสมมติฐานที่กำหนดไว้ในข้อที่ 2 แสดงว่าความยาวของแบบทดสอบมีผลต่ออำนาจการตรวจสอบของวิธี LR กล่าวคือเมื่อความยาวของแบบทดสอบเพิ่มขึ้นจาก 20 ข้อ เป็น 40 ข้อ จะส่งผลให้อำนาจการตรวจสอบของวิธี LR มีค่าเพิ่มขึ้นประมาณร้อยละ 14 สาเหตุดังกล่าวอาจเนื่องมาจาก ถ้าแบบทดสอบมีจำนวนข้อเพิ่มมากขึ้นแล้วความเชื่อมั่นจะมีค่าเพิ่มมากขึ้นและมีผลให้อำนาจการตรวจสอบมีค่าเพิ่มมากขึ้นด้วย (Narayanan and Swaminathan, 1996) แต่เป็นที่น่าสังเกตว่าเมื่อเพิ่มความยาวจาก 20 ข้อ เป็น 60 ข้อ หรือเพิ่มมากขึ้นจาก 40 ข้อ เป็น 60 ข้อ จะไม่มีผลต่ออำนาจการตรวจสอบ ทั้งนี้อาจเป็นเพราะว่าอำนาจการตรวจสอบมีค่าเพิ่มมากขึ้นในแบบทดสอบที่มีความเชื่อมั่นระดับปานกลาง ข้อค้นพบดังกล่าวสอดคล้องกับผลการศึกษาของแอกเกอร์แมนและอีแวนส์ (Narayanan and Swaminathan, 1994 ; citing Ackerman and Evans, 1992)

2.2 การจำแนกผิดพลาดของวิธี MH เมื่อใช้แบบทดสอบความยาว 20 ข้อ กับ 40 ข้อ, 20 ข้อ กับ 40 ข้อ และ 40 ข้อ กับ 60 ข้อ มีค่าไม่แตกต่างกัน และการจำแนกผิดพลาดของวิธี LR เมื่อใช้แบบทดสอบความยาว 20 ข้อ กับ 40 ข้อ, 20 ข้อ กับ 40 ข้อ และ 40 ข้อ กับ 60 ข้อ มีค่าไม่แตกต่างกัน ซึ่งไม่สอดคล้องกับสมมติฐานที่กำหนดไว้ในข้อที่ 2 และไม่สอดคล้องกับการศึกษาของสวามินาธานและโรเจอร์ส (Swaminathan and Rogers, 1990) ที่ใช้แบบทดสอบความยาว 40 ข้อ กับ 60 ข้อ, 40 ข้อ กับ 80 ข้อ และ 60 ข้อ กับ 80 ข้อ แล้วพบว่าการจำแนกผิดพลาดของวิธี MH เกิดขึ้นประมาณร้อยละ 1 แต่การจำแนกผิดพลาดของวิธี LR เกิดขึ้นประมาณร้อยละ 1 – 6 ที่เป็นเช่นนี้อาจเนื่องมาจากในการวิเคราะห์ด้วยวิธี MH ครั้งนี้ ผู้วิจัยใช้เทคนิคสองขั้นตอน ซึ่งทำให้ผลการวิเคราะห์มีความถูกต้องแม่นยำสูง และไม่เพิ่มการจำแนก

ผิดพลาด ข้อค้นพบดังกล่าวสอดคล้องกับผลการศึกษาของคลอเซอร์และคณะ (Clauser and others, 1993) ส่วนวิธี LR สามารถวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอ ได้อย่างถูกต้องแม่นยำ (Swaminathan and Rogers, 1990) ทำให้การจำแนกผิดพลาดเกิดขึ้นค่อนข้างต่ำ ดังนั้นความยาวของแบบทดสอบจึงไม่มีผลกระทบต่อการจำแนกผิดพลาดของวิธี MH และวิธี LR

3. การเปรียบเทียบอำนาจการตรวจสอบ และ การจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอ ระหว่างวิธี MH และวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาดเท่ากัน

3.1 อำนาจการตรวจสอบระหว่างวิธี MH และวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 200$ พบว่ามีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่เมื่อใช้กลุ่มตัวอย่างขนาดอื่น ๆ พบว่ามีค่าไม่แตกต่างกัน ผลการศึกษาดังกล่าวมีเพียงกรณีเดียวที่สอดคล้องกับสมมติฐานที่กำหนดไว้ในข้อ 3 นอกนั้นในกรณีอื่น ๆ ไม่สอดคล้องกับสมมติฐานดังกล่าว แสดงว่าอำนาจการตรวจสอบระหว่างวิธี MH และวิธี LR เกือบทุกขนาดกลุ่มตัวอย่างมีค่าใกล้เคียงกัน ซึ่งข้อค้นพบดังกล่าวสอดคล้องกับผลการศึกษาของโรเจอร์สและสวามินาธาน (Rogers and Swaminathan, 1993) ซึ่งพบว่าอำนาจการตรวจสอบของวิธี MH กับวิธี LR มีค่าใกล้เคียงกันในเกือบทุกเงื่อนไข ทั้งนี้อาจเนื่องมาจากวิธี MH ถูกออกแบบมาเพื่อวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอชนิดที่มีโครงสร้างข้อสอบของผู้สอบสองกลุ่มขนานกันโดยเฉพาะ (Li and Stout, 1996) ซึ่งจะมีความถูกต้องและแม่นยำในการตรวจสอบข้อสอบที่มีลักษณะดังกล่าว ส่วนวิธี LR สามารถใช้โมเดลคำนวณข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอได้อย่างมีประสิทธิภาพ (Swaminathan and Rogers, 1990) ดังนั้นผลการวิเคราะห์อำนาจการตรวจสอบระหว่างวิธี MH และวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาดเท่ากัน จึงมีค่าใกล้เคียงกัน

3.2 การจำแนกผิดพลาดระหว่างวิธี MH และวิธี LR เมื่อใช้กลุ่มตัวอย่างทุกขนาด พบว่ามีค่าไม่แตกต่างกัน ซึ่งไม่สอดคล้องกับสมมติฐานในข้อ 3 และไม่สอดคล้องกับผลการศึกษาของสวามินาธานและโรเจอร์ส (Swaminathan and Rogers, 1990) ทั้งนี้อาจเนื่องมาจากในการวิเคราะห์ด้วยวิธี MH ผู้วิจัยได้ใช้การวิเคราะห์แบบสองขั้นตอน (Two-stage) ดังที่กล่าวมาแล้วข้างต้น ในขณะที่วิธี LR สามารถใช้โมเดลวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอได้อย่างมีประสิทธิภาพ ผลการวิเคราะห์ที่ได้จากทั้งสองวิธีดังกล่าวจึงมีความถูกต้องแม่นยำสูง ดังนั้นผลการวิเคราะห์การจำแนกผิดพลาดระหว่างวิธี MH และวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาดเท่ากัน จึงมีค่าใกล้เคียงกัน

4. การเปรียบเทียบอำนาจการตรวจสอบ และ การจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอ ระหว่างกลุ่มตัวอย่างขนาดต่างกัน เมื่อใช้วิธีตรวจสอบวิธีเดียวกัน

4.1 อำนาจการตรวจสอบของวิธี MH เมื่อใช้กลุ่มตัวอย่างขนาดต่างกัน เกือบทุกกลุ่ม พบว่ามีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ซึ่งสอดคล้องกับสมมติฐานที่กำหนดไว้ในข้อ 4 เกือบทุกกรณีของการตรวจสอบ แสดงว่าขนาดกลุ่มตัวอย่างมีผลต่ออำนาจการตรวจสอบของวิธี MH กล่าวคือ เมื่อขนาดกลุ่มตัวอย่างเพิ่มมากขึ้นจะมีผลทำให้อำนาจการตรวจสอบของวิธีดังกล่าวมีค่าเพิ่มมากขึ้น โดยเฉพาะถ้าเพิ่มขนาดกลุ่มตัวอย่างจาก $N_R : N_F = 300 : 200$ เป็น $N_R : N_F = 1000 : 1000$ จะมีผลทำให้อำนาจการตรวจสอบเพิ่มขึ้นสูงที่สุด ประมาณร้อยละ 50 ผลการศึกษานี้สอดคล้องกับผลการศึกษาของโรเจอร์สและสวามินาธาน (Rogers and Swaminathan, 1993)

4.2 การจำแนกผิดพลาดของวิธี MH เมื่อใช้กลุ่มตัวอย่างขนาดต่างกันทุกขนาด พบว่ามีค่าไม่แตกต่างกัน ซึ่งไม่สอดคล้องกับสมมติฐานที่กำหนดไว้ในข้อ 4 และไม่สอดคล้องกับผลการศึกษาของสวามินาธานและโรเจอร์ส (Swaminathan and Rogers. 1990) ทั้งนี้อาจเนื่องมาจากการวิเคราะห์ด้วยวิธี MH ผู้วิจัยได้ใช้การวิเคราะห์แบบสองขั้นตอน (Two-stage) ดังที่กล่าวมาแล้วข้างต้น ทำให้ผลการวิเคราะห์ที่ได้จากวิธี MH มีความถูกต้องแม่นยำสูง ดังนั้นการจำแนกผิดพลาดของวิธี MH เมื่อใช้กลุ่มตัวอย่างขนาดต่างกัน จึงมีค่าใกล้เคียงกัน นั่นคือขนาดกลุ่มตัวอย่างไม่มีผลต่อการจำแนกผิดพลาดของวิธี MH

4.3 อำนาจการตรวจสอบของวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาดต่างกัน เกือบทุกกลุ่ม พบว่ามีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ซึ่งสอดคล้องกับสมมติฐานที่กำหนดไว้ในข้อ 4 เกือบทุกกรณีของการตรวจสอบ แสดงว่าขนาดกลุ่มตัวอย่างมีผลต่ออำนาจการตรวจสอบของวิธี LR กล่าวคือ เมื่อขนาดกลุ่มตัวอย่างเพิ่มมากขึ้นจะมีผลทำให้อำนาจการตรวจสอบของวิธีดังกล่าวมีค่าเพิ่มมากขึ้น โดยเฉพาะถ้าเพิ่มกลุ่มตัวอย่างขนาดจาก $N_R : N_F = 300 : 200$ เป็นกลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 1000$ จะมีผลทำให้อำนาจการตรวจสอบเพิ่มขึ้นสูงที่สุด ประมาณร้อยละ 36 ผลการศึกษาในครั้งนี้สอดคล้องกับผลการศึกษาของโรเจอร์สและสวามินาธาน (Rogers and Swaminathan. 1993)

4.4 การจำแนกผิดพลาดของวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาดต่างกันทุกขนาด พบว่ามีค่าไม่แตกต่างกัน ซึ่งไม่สอดคล้องกับสมมติฐานที่กำหนดไว้ในข้อ 4 และไม่สอดคล้องกับผลการศึกษาของสวามินาธานและโรเจอร์ส (Swaminathan and Rogers. 1990) ทั้งนี้อาจเนื่องมาจากวิธี LR สามารถใช้โมเดลวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันแบบสมมาตรได้อย่างมีประสิทธิภาพ ผลการวิเคราะห์ที่ได้จากวิธี LR จึงมีความถูกต้องแม่นยำสูง ดังนั้นการจำแนกผิดพลาดของวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาดต่างกัน จึงมีค่าใกล้เคียงกัน นั่นคือขนาดกลุ่มตัวอย่างไม่มีผลต่อการจำแนกผิดพลาดของวิธี LR

5. การเปรียบเทียบอำนาจการตรวจสอบ และ การจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สมมาตร ระหว่างวิธี MH และวิธี LR เมื่อใช้แบบทดสอบความยาวเท่ากัน

5.1 อำนาจการตรวจสอบระหว่างวิธี MH และวิธี LR เมื่อใช้แบบทดสอบความยาว 20 ข้อ, 40 ข้อ และ 60 ข้อ มีค่าไม่แตกต่างกัน ซึ่งไม่สอดคล้องกับสมมติฐานในข้อ 5 และไม่สอดคล้องกับผลการศึกษาของโรเจอร์สและสวามินาธาน (Rogers and Swaminathan. 1993) ซึ่งพบว่าอำนาจการตรวจสอบของวิธี LR มีค่าสูงกว่าอำนาจการตรวจสอบของวิธี MH เกือบทุกเงื่อนไขในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สมมาตร ทั้งนี้อาจเนื่องมาจากโรเจอร์สและสวามินาธาน วิเคราะห์โดยใช้วิธี MH แบบเดิมของฮอลแลนด์และเทเยอร์ (Holland and Thayer. 1988) ซึ่งไม่สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สมมาตรชนิดที่มีโค้งลักษณะข้อสอบ (Item Characteristic Curves : ICCs) ของกลุ่มผู้สอบสองกลุ่มตัดกันตรงจุดกึ่งกลางของช่วงความสามารถของผู้สอบ (Swaminathan and Rogers. 1990) เพราะว่าวิธี MH ดังกล่าวคำนวณโดยใช้ค่าสถิติที่มีเครื่องหมาย (Signed statistic) ทำให้ความแตกต่างทางลบระหว่างกลุ่มผู้สอบสองกลุ่มหักล้างกับความแตกต่างทางบวกระหว่างกลุ่มผู้สอบสองกลุ่ม ดังนั้นจึงไม่สามารถตรวจสอบข้อสอบที่มีลักษณะดังกล่าวได้ แต่ในการศึกษาครั้งนี้ผู้วิจัยใช้วิธี MH แบบปรับปรุงการวิเคราะห์ของเมเซอร์และคณะ (Mazor and others. 1994) โดยจะแบ่งกลุ่มผู้สอบออกเป็นสองกลุ่มตามระดับความสามารถ คือกลุ่มผู้สอบที่มีความสามารถสูงและกลุ่มผู้สอบที่มีความสามารถต่ำ แล้ววิเคราะห์แยกกลุ่มกัน ดังนั้นวิธีนี้จึงสามารถแก้ปัญหาข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สมมาตรชนิดที่มีโค้งลักษณะข้อสอบของกลุ่มผู้สอบสองกลุ่มตัดกันตรงจุดกึ่งกลางของช่วงความสามารถของผู้สอบได้ ส่วนวิธี LR ไม่มีปัญหาในการวิเคราะห์

ข้อสอบที่มีปัญหาดังกล่าว เพราะว่าวิธี LR นอกจากจะใช้โมเดลทางคณิตศาสตร์คำนวณข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอแล้วยังสามารถคำนวณข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอได้อีกด้วย ดังนั้นผลการวิเคราะห์อำนาจการตรวจสอบระหว่างวิธี MH แบบปรับปรุงการวิเคราะห์โดยการแบ่งกลุ่มผู้สอบและวิธี LR เมื่อใช้แบบทดสอบความยาวเท่ากัน จึงมีค่าใกล้เคียงกัน

5.2 การจำแนกผิดพลาดระหว่างวิธี MH และวิธี LR เมื่อใช้แบบทดสอบความยาว 20 ข้อ, 40 ข้อ และ 60 ข้อ มีค่าไม่แตกต่างกัน ซึ่งไม่สอดคล้องกับสมมติฐานในข้อ 5 ทั้งนี้อาจเนื่องมาจากการวิเคราะห์ด้วยวิธี MH ผู้วิจัยใช้การวิเคราะห์แบบสองขั้นตอน(Two-stage) โดยขั้นตอนแรกใช้คะแนนรวมของแบบสอบทั้งฉบับเป็นเกณฑ์การจับคู่กลุ่มผู้สอบย่อยสองกลุ่ม แล้ววิเคราะห์การทำหน้าที่ต่างกันของข้อสอบเมื่อพบว่าข้อสอบข้อใดที่ทำหน้าที่ต่างกันให้นำคะแนนของข้อสอบข้อนั้นออกจากคะแนนรวมของผู้สอบแต่ละคน และขั้นตอนที่สองใช้คะแนนรวมของแบบสอบที่นำเอาข้อสอบที่ทำหน้าที่ต่างกันซึ่งตรวจพบในขั้นตอนแรกออกไปแล้ว เพื่อวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบซ้ำอีกครั้งหนึ่ง การวิเคราะห์ดังกล่าวจะทำให้เกณฑ์การจับคู่ของกลุ่มผู้สอบมีความบริสุทธิ์(Holland and Thayer. 1988) ในขณะเดียวกันวิธี LR สามารถใช้โมเดลวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอได้อย่างมีประสิทธิภาพผลการวิเคราะห์จากสองวิธีดังกล่าวจึงมีความถูกต้องแม่นยำสูง ดังนั้นผลการวิเคราะห์การจำแนกผิดพลาดระหว่างวิธี MH และวิธี LR เมื่อใช้แบบทดสอบความยาวเท่ากัน จึงมีค่าใกล้เคียงกัน

6. การเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ระหว่างแบบทดสอบความยาวต่างกัน เมื่อใช้วิธีตรวจสอบวิธีเดียวกัน

6.1 อำนาจการตรวจสอบของวิธี MH เมื่อใช้แบบทดสอบความยาว 20 ข้อ กับ 40 ข้อ, 20 กับ 60 ข้อ และ 40 ข้อ กับ 60 ข้อ มีค่าไม่แตกต่างกัน และอำนาจการตรวจสอบของวิธี LR เมื่อใช้แบบทดสอบความยาว 20 ข้อ กับ 40 ข้อ, 20 ข้อ กับ 60 ข้อ และ 40 ข้อ กับ 60 ข้อ มีค่าไม่แตกต่างกันซึ่งไม่สอดคล้องกับสมมติฐานที่กำหนดในข้อ 6 และไม่สอดคล้องกับผลการศึกษาของสวามินาทานและโรเจอร์ส (Swaminathan and Rogers. 1990) ที่พบว่า ความยาวของแบบทดสอบมีผลกระทบต่ออำนาจการตรวจสอบของวิธี MH และวิธี LR แต่สอดคล้องกับผลการศึกษาของโรเจอร์สและสวามินาทาน (Rogers and Swaminathan. 1993) ที่พบว่าความยาวของแบบทดสอบไม่มีผลกระทบต่ออำนาจการตรวจสอบของวิธี MH และวิธี LR

6.2 การจำแนกผิดพลาดของวิธี MH เมื่อใช้แบบทดสอบความยาว 20 ข้อ กับ 40 ข้อ, 20 ข้อ กับ 60 ข้อ และ 40 ข้อ กับ 60 ข้อ มีค่าไม่แตกต่างกัน และการจำแนกผิดพลาดของวิธี LR เมื่อใช้แบบทดสอบความยาว 20 ข้อ กับ 40 ข้อ, 20 ข้อ กับ 60 ข้อ และ 40 ข้อ กับ 60 ข้อ มีค่าไม่แตกต่างกัน ซึ่งไม่สอดคล้องกับสมมติฐานที่กำหนดในข้อ 6 และไม่สอดคล้องกับผลการศึกษาของสวามินาทานและโรเจอร์ส (Swaminathan and Rogers. 1990) ที่พบว่า การจำแนกผิดพลาดของวิธี MH เกิดขึ้นน้อยกว่าการจำแนกผิดพลาดของวิธี LR

7. การเปรียบเทียบอำนาจการตรวจสอบ และการจำแนกผิดพลาด ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ระหว่างวิธี MH และวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาดเท่ากัน

7.1 อำนาจการตรวจสอบระหว่างวิธี MH และวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100$ และ $500 : 100$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 จากผลการศึกษาดังกล่าวพบว่า มีเพียง 2 กรณี คือ เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 100$ และ $500 : 100$ สอดคล้องกับ

สมมติฐานที่กำหนดไว้ในข้อ 7 นอกนั้นในกรณีอื่น ๆ ไม่สอดคล้องกับสมมติฐานดังกล่าว แสดงว่าอำนาจการตรวจสอบ ระหว่างวิธี MH และวิธี LR เกือบทุกขนาดกลุ่มตัวอย่างมีค่าใกล้เคียงกัน ผลการศึกษาดังกล่าว สอดคล้องกับผลการศึกษาของโรเจอร์สและสวามินาธาน (Rogers and Swaminathan, 1993) ซึ่งพบว่าอำนาจการตรวจสอบของวิธี MH กับวิธี LR มีค่าใกล้เคียงกันเกือบทุกเงื่อนไข ทั้งนี้อาจเนื่องมาจากวิธี MH ถูกออกแบบมาเพื่อวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอโดยเฉพาะ (Li and Stout, 1996) ซึ่งจะสามารถตรวจสอบข้อสอบที่มีลักษณะดังกล่าวได้อย่างถูกต้องและแม่นยำ ส่วนวิธี LR สามารถใช้โมเดลคำนวณข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอได้อย่างมีประสิทธิภาพ (Swaminathan and Rogers, 1990) ดังนั้นผลการวิเคราะห์อำนาจการตรวจสอบระหว่างวิธี MH และวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาดเท่ากัน จึงมีค่าใกล้เคียงกัน

7.2 การจำแนกผิดพลาดระหว่างวิธี MH และวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 300, 500 : 100$ และ $1000 : 200$ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 จากผลการศึกษาดังกล่าวมีเพียง 3 กรณี คือ เมื่อใช้กลุ่มตัวอย่างขนาด $N_R : N_F = 300 : 300, 500 : 100$ และ $1000 : 200$ สอดคล้องกับสมมติฐานที่กำหนดไว้ในข้อ 7 นอกนั้นในกรณีอื่น ๆ ไม่สอดคล้องกับสมมติฐานดังกล่าว แสดงว่าการจำแนกผิดพลาดระหว่างวิธี MH และวิธี LR เกือบทุกขนาดกลุ่มตัวอย่างมีค่าใกล้เคียงกัน ทั้งนี้ อาจเนื่องมาจากในการวิเคราะห์ด้วยวิธี MH ผู้วิจัยใช้การวิเคราะห์แบบสองขั้นตอน (Two-stage) ดังที่กล่าวมาแล้วข้างต้น ส่วนวิธี LR สามารถใช้โมเดลวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอได้อย่างมีประสิทธิภาพ ผลการวิเคราะห์จากทั้งสองวิธีดังกล่าวจึงมีความถูกต้องแม่นยำสูง ดังนั้นผลการวิเคราะห์การจำแนกผิดพลาดระหว่างวิธี MH และวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาดเท่ากัน จึงมีค่าใกล้เคียงกัน

8. ผลการเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ ระหว่างกลุ่มตัวอย่างขนาดต่างกัน เมื่อใช้วิธีตรวจสอบวิธีเดียวกัน

8.1 อำนาจการตรวจสอบของวิธี MH เมื่อใช้กลุ่มตัวอย่างขนาดต่างกันพบว่ามีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ซึ่งสอดคล้องกับสมมติฐานที่กำหนดไว้ในข้อ 8 เกือบทุกกรณีของการตรวจสอบ แสดงว่าขนาดกลุ่มตัวอย่างมีผลต่ออำนาจการตรวจสอบของวิธี MH กล่าวคือเมื่อขนาดกลุ่มตัวอย่างเพิ่มมากขึ้น จะมีผลให้อำนาจการตรวจสอบของวิธีดังกล่าวมีค่าเพิ่มมากขึ้น โดยเฉพาะถ้าเพิ่มกลุ่มตัวอย่างจากขนาด $N_R : N_F = 300 : 100$ เป็นกลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 1000$ จะทำให้อำนาจการตรวจสอบเพิ่มขึ้นสูงที่สุด ประมาณร้อยละ 56 ผลการศึกษาในครั้งนี้สอดคล้องกับการศึกษาของโรเจอร์สและสวามินาธาน (Rogers and Swaminathan, 1993) และนารายานานและสวามินาธาน (Narayanan and Swaminathan, 1996)

8.2 การจำแนกผิดพลาดของวิธี MH เมื่อใช้กลุ่มตัวอย่างขนาดต่างกัน พบว่ามีค่าไม่แตกต่างกันในเกือบทุกกรณีที่ศึกษา ซึ่งสอดคล้องกับสมมติฐานที่กำหนดไว้ในข้อ 8 ในบางกรณีของการตรวจสอบ แสดงว่าขนาดกลุ่มตัวอย่างมีแนวโน้มที่จะไม่มีผลต่อการจำแนกผิดพลาดของวิธี MH

8.3 อำนาจการตรวจสอบของวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาดต่างกัน พบว่ามีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ซึ่งสอดคล้องกับสมมติฐานที่กำหนดไว้ในข้อ 8 เกือบทุกกรณีของการตรวจสอบ แสดงว่าขนาดกลุ่มตัวอย่างมีผลต่ออำนาจการตรวจสอบของวิธี LR กล่าวคือเมื่อขนาดกลุ่มตัวอย่างเพิ่มมากขึ้น จะมีผลให้อำนาจการตรวจสอบของวิธีดังกล่าวมีค่าเพิ่มมากขึ้น โดยเฉพาะถ้าเพิ่มกลุ่มตัวอย่างจากขนาด $N_R : N_F = 300 : 300$ เป็นกลุ่มตัวอย่างขนาด $N_R : N_F = 1000 : 1000$ จะทำให้

อำนาจการตรวจสอบเพิ่มขึ้นสูงที่สุด ประมาณร้อยละ 38 ผลการศึกษาในครั้งนี้สอดคล้องกับผลการศึกษาของ โรเจอร์สและสวามินาธาน(Rogers and Swaminathan, 1993)และนารายานานและสวามินาธาน (Narayanan and Swaminathan, 1996)

8.4 การจำแนกผิดพลาดของวิธี LR เมื่อใช้กลุ่มตัวอย่างขนาดต่างกัน พบว่ามีค่าไม่แตกต่างกันในเกือบทุกกรณีที่ศึกษา ซึ่งสอดคล้องกับสมมติฐานที่กำหนดไว้ในข้อ 8 ในบางกรณีของการตรวจสอบ แสดงว่าขนาดกลุ่มตัวอย่างมีแนวโน้มที่จะไม่มีผลต่อการจำแนกผิดพลาดของวิธี LR

ข้อเสนอแนะ

1. ข้อเสนอแนะในการนำไปใช้

1.1 การศึกษาในครั้งนี้ ได้ข้อค้นพบว่าในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้งแบบ สม่่าเสมอและแบบไม่สม่่าเสมอ วิธี MH และวิธี LR มีอำนาจการตรวจสอบ (Power) และการจำแนกผิดพลาด (Type I error) ใกล้เคียงกัน แสดงว่าทั้งสองวิธีมีประสิทธิภาพเท่าเทียมกัน จึงสามารถจะเลือกใช้วิธีใดวิธีหนึ่ง แทนกันได้ แต่กระบวนการในการวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกันโดยใช้วิธี MH สามารถกระทำได้ง่ายและไม่ยุ่งยากซับซ้อนเหมือนกับวิธี LR นอกจากนี้วิธี MH ยังสามารถแปลผลได้ง่าย ประหยัดเวลาและค่าใช้จ่ายมากกว่าวิธี LR

1.2 การศึกษาในครั้งนี้ได้ข้อค้นพบว่าในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันทั้งแบบสม่่าเสมอและแบบไม่สม่่าเสมอ ขนาดกลุ่มตัวอย่างมีผลต่ออำนาจการตรวจสอบของวิธี MH และวิธี LR เกือบทุกเงื่อนไขของการตรวจสอบ กล่าวคือเมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้น จะมีผลให้อำนาจการตรวจสอบมีค่าเพิ่มขึ้น ในขณะที่การจำแนกผิดพลาดมีผลต่อวิธีการตรวจสอบทั้งสอง เพียงบางเงื่อนไขของการตรวจสอบ นั้นแสดงว่า ถึงแม้ว่าจะใช้ขนาดกลุ่มตัวอย่างที่มีขนาดเล็กก็จะมีผลต่อการจำแนกผิดพลาด

1.3 การศึกษาในครั้งนี้ ได้ข้อค้นพบว่าในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสม่่าเสมอ ความยาวของแบบทดสอบมีผลต่ออำนาจการตรวจสอบของวิธี LR แต่ไม่มีผลต่ออำนาจการตรวจสอบของวิธี MH กล่าวคือเมื่อเพิ่มความยาวของแบบทดสอบจาก 20 ข้อ เป็น 40 ข้อ จะทำให้อำนาจการตรวจสอบของวิธี LR มีค่าเพิ่มขึ้น แต่เมื่อเพิ่มความยาวของแบบทดสอบจาก 20 ข้อ เป็น 60 ข้อ จะไม่มีผลต่ออำนาจการตรวจสอบของวิธีดังกล่าว แสดงว่าอำนาจการตรวจสอบของวิธี LR จะมีค่าสูงในแบบทดสอบขนาดปานกลาง ส่วนการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่่าเสมอ พบว่าความยาวของแบบทดสอบจะไม่มีผลต่ออำนาจการตรวจสอบของทั้งสองวิธี นอกจากนี้ยังได้ข้อค้นพบว่า ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสม่่าเสมอและแบบไม่สม่่าเสมอ ความยาวของแบบทดสอบไม่มีผลต่อการจำแนกผิดพลาดของทั้งสองวิธี

2. ข้อเสนอแนะในการทำวิจัยในครั้งต่อไป

2.1 ในการศึกษาวิจัยในครั้งต่อไป ควรศึกษาลักษณะเดียวกับงานวิจัยครั้งนี้ แต่ควรใช้ตัวแปรอื่นๆ ในการแบ่งกลุ่ม เช่นกลุ่มโรงเรียนในส่วนกลางกับส่วนภูมิภาค เพื่อเปรียบเทียบกับผลที่ได้จากการวิจัยในครั้งนี้

2.2 ในการศึกษาวิจัยในครั้งต่อไป ควรใช้วิธีตรวจสอบอื่น ๆ เช่น วิธีอัตราส่วนความน่าจะเป็นสูงสุด วิธีทดสอบไค-กาวงสองของลอว์รี วิธีการวัดพื้นที่ของคิมและโคเฮน เป็นต้น และควรศึกษาร่วมกับตัวแปรอื่น ๆ เช่น ระดับความแตกต่างของค่าความยาก ค่าอำนาจจำแนกของข้อสอบ และสัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน เป็นต้น

บรรณานุกรม

บรรณานุกรม

- กาญจนา วัชรสุนทร. การพัฒนาเกณฑ์ตัดสินข้อสอบลำเอียงทางเพศ. วิทยานิพนธ์ ค.ด. กรุงเทพฯ : จุฬาลงกรณ์มหาวิทยาลัย, 2538. อัดสำเนา.
- เกษร ห่วงจิตร์. การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบสำหรับแบบสอบคัดเลือกระดับบัณฑิตศึกษา วิชาภาษาไทยและภาษาอังกฤษด้วยวิธีแมนเทิล-แฮนส์เซล. วิทยานิพนธ์ ค.ม. กรุงเทพฯ : จุฬาลงกรณ์มหาวิทยาลัย, 2539. อัดสำเนา.
- คมศักดิ์ ชื่นชม. การศึกษาผลการวิเคราะห์ความลำเอียงที่ใช้วิธีต่างกันของแบบทดสอบวัดจริยธรรมด้าน ความซื่อสัตย์. วิทยานิพนธ์ กศ.ม. กรุงเทพฯ : มหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร, 2539. อัดสำเนา.
- จิตติมา วรณศรี. การเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี แมนเทิล-แฮนส์เซลกับวิธีชิปเทสต์ เมื่อขนาดกลุ่มตัวอย่างและความยาวของแบบสอบต่างกัน. วิทยานิพนธ์ ค.ม. กรุงเทพฯ : จุฬาลงกรณ์มหาวิทยาลัย, 2540. อัดสำเนา.
- จิตสุตา ธรพร. การเปรียบเทียบผลการวิเคราะห์ความลำเอียงของข้อสอบจากแบบทดสอบโจทย์ปัญหา ทางคณิตศาสตร์และแบบทดสอบวัดความเข้าใจทางภาษา. วิทยานิพนธ์ กศ.ม. กรุงเทพฯ : มหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร, 2539. อัดสำเนา.
- ชูศรี วงศ์รัตน์. เทคนิคการใช้สถิติเพื่อการวิจัย. พิมพ์ครั้งที่ 4. กรุงเทพฯ : ศูนย์หนังสือจุฬาลงกรณ์ มหาวิทยาลัย, 2530.
- พรรณี จินตมาศ. การเปรียบเทียบผลการวิเคราะห์ความลำเอียงของข้อสอบ โดยใช้ขนาดกลุ่มผู้สอบและวิธี วิเคราะห์ต่างกัน. วิทยานิพนธ์ กศ.ม. กรุงเทพฯ : มหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร, 2540. อัดสำเนา.
- เรวดี อินทสระ. ผลการตรวจสอบความลำเอียงของข้อสอบต่อการศึกษาคำความที่ตรงเชิงพยากรณ์ของ แบบทดสอบคัดเลือกที่คิดคะแนนต่างกัน. วิทยานิพนธ์ กศ.ด. กรุงเทพฯ : มหาวิทยาลัย ศรีนครินทรวิโรฒ ประสานมิตร, 2539. อัดสำเนา.
- ล้วน สายยศ และ อังคณา สายยศ. เทคนิคการวิจัยทางการศึกษา. พิมพ์ครั้งที่ 3. กรุงเทพฯ : ศูนย์ส่งเสริม วิชาการ, 2536.
- วลีมาศ แซ่อึ้ง. การเปรียบเทียบอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบ การทำหน้าที่ต่างกันของข้อสอบแบบอนุกรมระหว่างวิธีชิปเทสต์ปรับใหม่ วิธีชิปเทสต์ วิธี แมนเทิล-แฮนส์เซล และวิธีการถดถอยโลจิสติก. วิทยานิพนธ์ ค.ด. กรุงเทพฯ : จุฬาลงกรณ์ มหาวิทยาลัย, 2543. อัดสำเนา.
- เสรี ชัดเข้ม. การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันแบบไม่สมมาตรของข้อสอบระหว่างวิธี แมนเทิล-แฮนส์เซล แบบปกติ กับวิธีแมนเทิล-แฮนส์เซลแบบแบ่งกลุ่มตามความสามารถและความ ยาวของข้อสอบ. วิทยานิพนธ์ ค.ด. กรุงเทพฯ : จุฬาลงกรณ์มหาวิทยาลัย, 2539. อัดสำเนา.
- อังคณา สายยศ. การศึกษาผลการวิเคราะห์ความลำเอียงของแบบทดสอบความเข้าใจในการอ่านตาม ลักษณะ การใช้ภาษาพูดในครอบครัว. กรุงเทพฯ : ภาควิชาการวัดผลและวิจัยทางการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร, 2539.

- Baker, F.B., A Al-Karni and I.M.Al-Dosary. "EQUATE : A computer program for the test characteristic method of IRT equating." Applied Psychological Measurement. 15(1) : 78 ; March, 1991.
- Camilli, Gregory and Lorrie A. Shepard. "Methods for Identifying Biased Test Items," California : Sage Publications, 1994.
- Clauser, Brian E. et al. Examination of various Influence on the Mantel-Haenszel Statistic. Paper Presented at the Annual Meeting of the American Educational Research Association. Chicago Illinois, April 3 – 7 ; 1991.
- Clauser, B., Mazor, K. and Hambleton, R.K. "The influence of the criterion variable on the identification of differentially functioning test item using the Mantel-Haenszel Statistic," Applied Psychological Measurement. 15(4) : 353-359 ; December, 1991.
- Clauser, B., Mazor, K. and Hambleton, R.K. "The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel," Applied Measurement in Education. 6(4) : 269-279 ; 1993.
- Cohen, Allan S. and Seock-Ho. Kim. " A comparison of Lord's chi-square and Raju's area measures in detection of Differential Item Functioning," Applied Psychological Measurement. 17(1) : 39-52 ; March, 1993.
- Dorans, N.J. and Holland, P.W. DIF Detection and Description : Mantel-Haenszel and Standardization. In P.W. Holland and Wainer (Eds.) Differential Item Functioning. Hillsdale, NJ. Erlbaum Associates, Inc. : 35 – 66 ; 1993.
- Feinstein, Zachary S. "Effects of differing item parameters on closed-interval DIF statistics," Applied Psychological Measurement. 19(2) : 131 – 142 ; June, 1995.
- Fidalgo, A.M. "MHDIF : A computer program for detecting uniform and nonuniform Differential Item Functioning with the Mantel-Haenszel procedure," Applied Psychological Measurement. 18(3) : 1994.
- _____. MHDIF Version 1.0. University of Oviedo ; Spain : 1995.
- Hambleton, Ronald K. and Roger H. Jane "Detecting potentially biased test item : Comparison of IRT area and Mantel-Haenszel methods," Applied Measurement in Education. 2(4) : 313 – 334 ; 1989.
- Hambleton, Ronald K., H. Swaminathan and Roger H. Jame. Foundamentals of Item Response Theory. Newbury Park, Calofonia, Sage Publications, 1991.
- Holland, P.W., and Thayer, D.T. Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H.I. Braun (Eds.) Test validity. Hillsdale, NJ . lawrence Erlbaum : 129 – 145 ; 1988.
- Holland, Paul W. and Wainer, Haward (Eds.) Differential item functioning. Hillsdale, New Jersey : Lawrence Erlbaum Associates, 1993.

- Intasuwan, P. A Comparison of Three Approaches for Determining Item Bias in Cross-National Testing. Doctor's Thesis The University of Pittsburgh, 1979.
- Kim, Seock-Ho and Allan S. Cohen. "A comparison of two area measures for detecting Differential Item Functioning," Applied Psychological Measurement. 15(3) : 269 – 278 ; September, 1991.
- _____. IRTDIF : A computer program for IRT Differential Item Functioning analysis. Applied Psychological Measurement. 16 : 158 ; 1992.
- Kim, Seock-Ho and Allan S. Cohen and Hae-OK. Kim. "Effects of linking methods on detecting of DIF," Journal of Educational Measurement. 29(1) : 51 – 66 ; Spring, 1992.
- _____. "An investigation of Lord's procedure for the detection of Differential Item Functioning," Applied Psychological Measurement. 18(3) : 217 - 228 ; September, 1994.
- Laksana, S. Application of Analysis of Variance Approach and ITBS Form 7. The University of Iowa, 1979.
- Li, H., Stout, W. "A new procedure for detection of crossing DIF," Psychometrika. 61 : 647 – 677 ; December, 1996.
- Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale NJ:Erlbaum, 1980.
- Mazor, Kathleen M., Clauser, Brian E. and Hambleton, Ronald K. "The effect of sample size on the functioning of the Mantel-Haenszel statistic," Educational and Psychological Measurement. 52(2) : 443-451 ; Summer, 1992.
- Mazor, Kathleen M., Clauser, Brian E. and Hambleton, Ronald K. "Identification of nonuniform Differential Item Functioning using A variation of the Mantel-Haenszel Procedure," Educational and Psychological Measurement. 54(2) : 284-291 ; Summer, 1994.
- Mellenbergh, G.J. "Contingency table models for assessing item bias," Journal of Educational Statistics. 7(2) : 105 – 118 ; Summer, 1982.
- Miller, M.D. and T.C. Oshima. "Effect of sample size, number of biased items, and magnitude of Bias on a two-stage item bias estimation method," Applied Psychological Measurement. 16(4) : 381 – 388 ; December, 1992.
- Miller, T.R. and Spray, J. A. "Logistic discriminant function analysis for DIF identification of polytomously scored items," Journal of Educational Measurement. 30(2) : 107 –122 ; Summer, 1993.
- Millsap, Roger E. and Howard T. Everson. "Methodology Review : Statistical fpproaches for assessing measurement bias," Applied Measurement in Education. 17(4) : 297 – 334 ; December, 1993.

- Mislevy, R.J. and R.D. Bock. "BILOG 3 : Item analysis and test scoring with binary logistic model (Computer Program). Mooresville IN : Scientific Software. Mislevy, Robert J. and Stocking, Martha L. A Consumer's Guide to LOGIST and BILOG," "Applied Psychological Measurement. 13(1) : 57 - 75 ; March, 1990
- Narayanan, Pankaja and H.Swaminathan. "Performance of the mantel-Haenszel and Simultaneous Item Bias procedures for detecting Differential Item Function," Applied Psychological Measurement. 18(4) : 315 – 328 ; December, 1994.
- _____. "Identification of Items that show nonuniform DIF," Applied Psychological Measurement. 20(3) : 257 – 274 ; September, 1996.
- Potenza M.T. and Dorans M.T. "DIF Assessment for polytomously scored items : A framework for classification and evaluation," Applied Psychological Measurement. 19(1) : 23 – 37 ; March, 1995.
- Raju, Nambury S. "Determining the significance of signed and unsigned areas between two item response function," Applied Psychological Measurement. 14(2) : 197 - 207 ; June, 1990.
- Raju, Drasgow Fritz and Jeffrey A. Slind. "An empirical comparison of the area methods, Lord's chi-square test and the Mantel-Haenszel technique for assessing Differential Item Functioning," Educational and Psychological Measurement. 53(2) : 301-314 ; Summer, 1993.
- Rogers, H. Jane and H. Swaminathan. "A comparison of Logistic Regression and Mantel-Haenszel Procedures for detecting Differential Item Functioning," Applied Psychological Measurement. 17(2) : 105 – 116 ; June, 1993.
- Shealy, Robin and William Stout. "A Model – Based Standardization Approach that Separates True Bias/ DIF from Group Ability Difference and Detects Test Bias / DIF as Well as Item Bias / DIF," Psychometrika. 58(2) : 159 – 194 ; June, 1993.
- Subkoviak, Micheal J. et al. "Empirical Comparison of Detected Item Bias Detection Procedure with Bias Manipulation," Journal of Education Measurement. 21(10) : 49 – 59 ; Spring, 1984.
- Swaminathan, H. and H.J. Rogers. "Detecting differential item functioning using logistic regression procedures," Journal of Educational Measurement. 27(4) : 361 – 370 ; Winter, 1990.
- Thissen David, Lynne Steinberg and Wainer Howard. " Use of item Response Theory in the Study of Group Difference in Trace Lines." In Test Validity, P. 147-170. Hillsdale, New Jersey : Lawrence Erlbaum Associates, 1988.
- Zieky M. Practical questions in the use of DIF Statistics in Development. In P.W. Holland and Wainer (Eds.) Differential item functioning. Hillsdale NJ. Erlbaum : 123 – 135 ; 1993.

ภาคผนวก

รายชื่อผู้เชี่ยวชาญ

- | | | |
|------------------|---------------|---|
| 1. นางสาวอรุณศรี | แสงแก้ว | อาจารย์ 3 ระดับ 8
โรงเรียนศรีอยุธยา ในพระอุปถัมภ์ฯ |
| 2. นางสาวจิราพร | แดงก่อ | อาจารย์ 3 ระดับ 8
โรงเรียนศรีอยุธยา ในพระอุปถัมภ์ฯ |
| 3. นางสาวอัมพร | วิชัยศรี | อาจารย์ 2 ระดับ 6
โรงเรียนอนุวัชรเทพวรวิหาร |
| 4. นางลัดดา | เสือสืบพันธุ์ | อาจารย์ 2 ระดับ 6
โรงเรียนสุวรรณารามวิทยาคม |
| 5. นายสุทธิ | กระจะจ่าง | อาจารย์ 2 ระดับ 5
โรงเรียนบางช้างวิทยา |

ประวัติย่อของผู้วิจัย

ชื่อ-สกุล	นายทองอยู่ สาระ
วัน เดือน ปีเกิด	24 พฤษภาคม 2506
สถานที่เกิด	ตำบลดอนลาน อำเภอดงหลวง จังหวัดแพร่
สถานที่อยู่ปัจจุบัน	497 ถนนศรีอยุธยา แขวงถนนพญาไท เขตราชเทวี กรุงเทพมหานคร 10400

ประวัติการศึกษา

2528	ค.บ. (คณิตศาสตร์-ภาษาอังกฤษ) วิทยาลัยครูพระนครศรีอยุธยา
2543	กศ.ม. (การวัดผลการศึกษา) มหาวิทยาลัยศรีนครินทรวิโรฒ

ประวัติการทำงาน

2530-ปัจจุบัน	อาจารย์ 2 ระดับ 6 โรงเรียนศรีอยุธยา ในพระอุปถัมภ์ฯ เขตราชเทวี กรุงเทพมหานคร
---------------	---

การเปรียบเทียบอำนาจการตรวจสอบและการจำแนกชนิดพลาด
ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอและแบบไม่สม่ำเสมอ
ระหว่างวิธีแมนเทล-แฮนส์เซลและวิธีการถดถอยโลจิสติก
โดยใช้ความยาวแบบทดสอบและขนาดกลุ่มตัวอย่างต่างกัน

บทคัดย่อ
ของ
ทองอยู่ สาระ

เสนอต่อมหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร เพื่อเป็นส่วนหนึ่งของการศึกษา
ตามหลักสูตรปริญญาการศึกษามหาบัณฑิต วิชาเอกการวัดผลการศึกษา
พฤษภาคม 2543

การศึกษาค้างนี้มีความมุ่งหมายเพื่อเปรียบเทียบอำนาจการตรวจสอบและการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอและแบบไม่สมำเสมอ ระหว่างวิธี MH วิธี LR โดยใช้ความยาวแบบทดสอบและขนาดกลุ่มตัวอย่างต่างกัน กลุ่มตัวอย่างที่ใช้ในการศึกษาค้างนี้เป็นนักเรียนระดับชั้นมัธยมศึกษาปีที่ 3 ภาคเรียนที่ 2 ปีการศึกษา 2542 ในโรงเรียนสังกัดกรมสามัญศึกษา ส่วนกลาง กลุ่มที่ 3 จำนวน 3,242 คน เครื่องมือที่ใช้เป็นแบบทดสอบความสามารถทางสมองที่ผู้วิจัยสร้างขึ้นตามแนวโครงสร้างของโอติส-เลนนอน ซึ่งเป็นแบบทดสอบเลือกตอบห้าตัวเลือก จำนวน 80 ข้อ วัดความสามารถทั่วไปสามด้าน คือความเข้าใจด้านภาษา เหตุผลด้านภาษา และเหตุผลด้านภาพ ในการศึกษาค้างนี้ผู้วิจัยได้สุ่มแบบทดสอบความยาวสามขนาด คือ 20 ข้อ, 40 ข้อ, 60 ข้อ และสุ่มกลุ่มตัวอย่างขนาดกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบจำนวน 10 กลุ่ม คือ 300 : 100, 300 : 200, 300 : 300, 500 : 100, 500 : 200, 500 : 300, 1000 : 100, 1000 : 200, 1000 : 300 และ 1000 : 1000 รวมเงื่อนไขที่ศึกษาทั้งหมด 30 เงื่อนไข (3 ขนาดความยาวแบบทดสอบ $\times$ 10 ขนาดกลุ่มตัวอย่าง) ผลการศึกษา พบว่า

1.อำนาจการตรวจสอบและการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันทั้งแบบสมำเสมอและแบบไม่สมำเสมอ ระหว่างวิธี MH และวิธี LR ภายใต้ความยาวแบบทดสอบและขนาดกลุ่มตัวอย่างที่ศึกษา เกือบทุกเงื่อนไข มีค่าไม่แตกต่างกัน

2.ความยาวของแบบทดสอบไม่มีผลต่ออำนาจการตรวจสอบและการจำแนกผิดพลาดในการตรวจสอบด้วยวิธี MH และวิธี LR ทั้งการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอและแบบไม่สมำเสมอ

3.ขนาดของกลุ่มตัวอย่างมีผลต่ออำนาจการตรวจสอบในการตรวจสอบด้วยวิธี MH และวิธี LR เกือบทุกเงื่อนไขของการศึกษาทั้งการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอและแบบไม่สมำเสมอ กล่าวคือเมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้น อำนาจการตรวจสอบ จะมีค่าเพิ่มขึ้น แต่พบว่าขนาดกลุ่มตัวอย่างไม่มีผลต่อการจำแนกผิดพลาด ในเกือบทุกเงื่อนไขที่ศึกษา ทั้งการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสมำเสมอและแบบไม่สมำเสมอ

A COMPARISON OF THE POWER AND TYPE I ERROR RATE
IN DETECTION OF UNIFORM AND NONUNIFORM DIFFERENTIAL ITEM FUNCTIONING
BETWEEN THE MANTEL-HAENSZEL AND THE LOGISTIC REGRESSION PROCEDURES
BY USING DIFFERENT TEST LENGTHS AND SAMPLE SIZES

AN ABSTRACT
BY
THONGYOO SARA

Presented in partial fulfillment of the requirements for the
Master of Education degree in Educational Measurement
At Srinakharinwirot University
May 2000

The purpose of this study was to compare the power and Type I error rate of two different procedures of uniform and nonuniform differential item function detecting, Mantel-Haenszel(MH) and Logistic Regression(LR). The comparison was investigated under different test lengths and sample sizes. The samples in this study were 3,242 Mathayom Suksa III students in the second semester of the academic year 1999. These students were in the school of the Department of General Education in Bangkok Group III. The instrument used in this study was 80 items of The Otis-Lennon Mental Ability Test Advance Level Form J. The test measured verbal comprehension, verbal reasoning and figural reasoning. Examinees response data were manipulated (1) 3 sizes of test lengths; 20, 40 and 60 items and (2) 10 sample sizes of examinees reference and focal groups ; $N_R : N_F = 300 : 100$, $300 : 200$, $300 : 300$, $500 : 100$, $500 : 200$, $500 : 300$, $1000 : 100$, $1000 : 200$, $1000 : 300$ and $1000 : 1000$. Thirty conditions were investigated in this study.

Results of this research were as follows :

1. Both the MH and LR procedures had equally power and Type I error rate in detecting uniform DIF and nonuniform DIF under most conditions, test lengths and sample sizes.
2. The test lengths had no effected to power and Type I error rate in detecting uniform DIF and nonuniform DIF with MH and LR procedures under most conditions.
3. The sample sizes had effected to power in detecting uniform DIF and nonuniform DIF with MH and LR procedures under most conditions, but had no effected to Type I error rate in detecting uniform DIF and nonuniform DIF with MH and LR procedures under most conditions.