

371.26

ณ 2187

จ. 2

การเปรียบเทียบความสอดคล้องของผลการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน
ระหว่าง วิธี Lord's χ^2 วิธี Raju's Area Measures
และ วิธี Closed Interval Area

ปริญญานิพนธ์

ของ

ญาณภัทร สีหะมงคล



- 8 ส.ค. 2541

เสนอต่อมหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร เพื่อเป็นส่วนหนึ่งของการศึกษา
ตามหลักสูตรปริญญาการศึกษาดุขุภักดิ์ สาขาการทดสอบและวัดผลการศึกษา

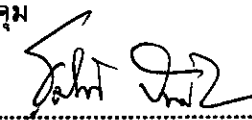
กันยายน 2540

ลิขสิทธิ์เป็นของมหาวิทยาลัยศรีนครินทรวิโรฒ

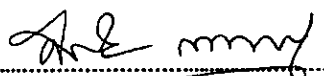
444307

คณะกรรมการควบคุมและคณะกรรมการสอบได้พิจารณาปริญญานิพนธ์ฉบับนี้แล้ว
เห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาการศึกษาดุขบัณฑิต
สาขาการทดสอบและวัดผลการศึกษา ของมหาวิทยาลัยศรีนครินทรวิโรฒได้

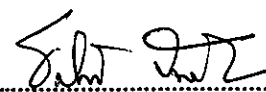
คณะกรรมการควบคุม

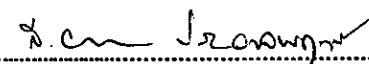

.....ประธาน
(ดร.ชูศักดิ์ ชัมภลชิต)

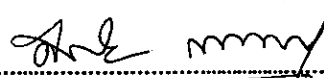

.....กรรมการ
(รศ.ดร.ส. วาสนา ประवालพฤษ)

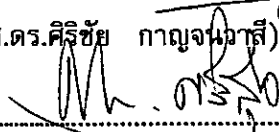

.....กรรมการ
(รศ.ดร.ศิริชัย กาญจนวาสี)

คณะกรรมการสอบ


.....ประธาน
(ดร.ชูศักดิ์ ชัมภลชิต)


.....กรรมการ
(รศ.ดร.ส. วาสนา ประवालพฤษ)


.....กรรมการ
(รศ.ดร.ศิริชัย กาญจนวาสี)


.....กรรมการที่แต่งตั้งเพิ่มเติม
(รศ.ดร.ดิเรก ศรีสุข)

บัณฑิตวิทยาลัยอนุมัติให้รับปริญญานิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตาม
หลักสูตรปริญญาการศึกษาดุขบัณฑิต สาขาการทดสอบและวัดผลการศึกษา ของมหาวิทยาลัย
ศรีนครินทรวิโรฒ


.....คณบดีบัณฑิตวิทยาลัย
(ดร.ศิริยุภา พูลสุวรรณ)

วันที่ 1๓ เดือน กันยายน พ.ศ. 2๕4๐

ประกาศคุณูปการ

ปริญญานิพนธ์ฉบับนี้สำเร็จลงได้ด้วยความกรุณาช่วยเหลือและเอาใจใส่ของคณะกรรมการที่ปรึกษาปริญญานิพนธ์ ทั้ง 3 ท่าน คือ อาจารย์ ดร. ชูศักดิ์ ชัมภลชิต ประธานกรรมการ รศ.ดร.ส. วาสนา ประवालพฤษ์ และ รศ.ดร. ศิริชัย กาญจนวาสี กรรมการที่ปรึกษา รวมทั้ง รศ.ดร.ดิเรก ศรีสุโข อาจารย์ภาควิชาวิจัยการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ที่ร่วมเป็นกรรมการสอบปริญญานิพนธ์ และให้คำแนะนำในการปรับปรุงแก้ไขปริญญานิพนธ์อย่างดี ผู้วิจัยขอกราบขอบพระคุณด้วยความซาบซึ้ง

ขอขอบคุณ Prof. Dr. Allan S. Cohen และ Assistant Prof. Dr. Seock-Ho Kim แห่งมหาวิทยาลัยวิสคอนซิน ที่ให้คำแนะนำอย่างดี พร้อมเอกสารและโปรแกรมหลักในการวิเคราะห์ข้อมูล (IRTDIF 1.1) Prof. Dr. Frank B. Baker แห่งมหาวิทยาลัยวิสคอนซิน ที่ให้โปรแกรมการเทียบคะแนน (EQUATE 2.1) ใช้ในการวิเคราะห์ข้อมูลบางส่วน และ Prof. Dr. William Stout แห่งมหาวิทยาลัยจอร์เจียที่ให้เอกสารบางส่วนประกอบการศึกษา

ขอขอบคุณ รศ.ดร.สมบูรณ์ ชิตพงษ์ รศ.นิภา ศรีไพโรจน์ ดร.กาญจนา วัฒนสุนทร อาจารย์บุญเรือง-ดร.รัตนา ศรีเทรัญ ดร.บุญชู ชลัษเฐียร และเพื่อน ๆ ชาววัดผล ที่ให้คำแนะนำและกำลังใจในการทำปริญญานิพนธ์

ท้ายสุดขอกราบขอบพระคุณบิดา-มารดา และครู-อาจารย์ที่ประสิทธิ์ประสาทวิชาความรู้แก่ผู้วิจัย และจะลืมนเสียมิได้ คือ กำลังใจจากอาจารย์ตรึงใจ สีหะมงคล และเด็กหญิงธีรดา สีหะมงคล ผู้ที่อยู่เบื้องหลังความสำเร็จของปริญญานิพนธ์ฉบับนี้

ญาณภัทร สีหะมงคล

สารบัญ

บทที่	หน้า
1 บทนำ.....	1
ภูมิหลัง.....	1
ความมุ่งหมายของการศึกษา.....	7
สมมติฐานในการศึกษา.....	8
ขอบเขตของการศึกษา.....	10
ข้อตกลงเบื้องต้น.....	11
ความสำคัญของการศึกษา.....	11
นิยามศัพท์เฉพาะ.....	12
2 เอกสารและงานวิจัยที่เกี่ยวข้อง.....	14
การทำหน้าที่ต่างกันของข้อสอบ.....	14
ความเป็นมา.....	14
ความหมาย.....	15
การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ.....	16
การใช้วิธีภายนอก.....	16
การใช้วิธีภายใน.....	17
การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามแนวคิดของทฤษฎี.....	
การตอบข้อสอบ.....	19
ความเป็นมาของทฤษฎี.....	19
หลักการของทฤษฎี.....	20
ข้อตกลงเบื้องต้นของทฤษฎี.....	20
แบบจำลองทฤษฎี.....	21
การตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันตามแนวคิดทฤษฎีการตอบข้อสอบ.....	24
งานวิจัยที่เกี่ยวข้อง.....	27
งานวิจัยของต่างประเทศ.....	27
งานวิจัยภายในประเทศ.....	32
3 วิธีดำเนินการศึกษา.....	39
แหล่งข้อมูลและกลุ่มตัวอย่าง.....	39
เครื่องมือที่ใช้ในการศึกษา.....	39
การเตรียมข้อมูลในการศึกษา.....	40

บทที่	หน้า
การวิเคราะห์ข้อมูล.....	44
4 ผลการวิเคราะห์ข้อมูล.....	52
ตอนที่ 1 ค่าสถิติพื้นฐานของข้อมูลผลการสอบ คุณภาพของแบบทดสอบ.....	53
การตรวจสอบความเป็นมิติเดียวของแบบทดสอบ และค่าพารามิเตอร์	
ของข้อสอบ.....	53
ตอนที่ 2 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ.....	58
2.1 จำนวนข้อสอบที่ทำหน้าที่ต่างกัน ความสอดคล้องในการตรวจสอบ	
และความสัมพันธ์ระหว่างวิธี ตรวจสอบ.....	58
2.2 อัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกัน	
ของข้อสอบ.....	97
2.3 อัตราความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกัน	
ของข้อสอบ.....	149
5 สรุป อภิปรายผลและข้อเสนอแนะ.....	164
สรุปผลการวิจัย.....	165
อภิปรายผลการวิจัย.....	169
ข้อเสนอแนะ.....	174
บรรณานุกรม.....	176
ภาคผนวก.....	185
ประวัติย่อของผู้วิจัย.....	198

บัญชีตาราง

ตาราง		หน้า
1	ค่าเฉลี่ย ค่าส่วนเบี่ยงเบนมาตรฐาน และค่าความเชื่อมั่นของแบบทดสอบ วิชาคณิตศาสตร์ จำนวน 75 ข้อ.....	53
2	ค่าไอเกนและร้อยละของความแปรปรวนขององค์ประกอบ.....	54
3	ค่าพารามิเตอร์ของข้อสอบ แยกตามเพศชาย-หญิง.....	55
4	จำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบ เมื่อกลุ่มตัวอย่าง 300 คน และ ความยาวแบบทดสอบ 20 ข้อ.....	59
5	ผลการทดสอบความแตกต่างของจำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบ ระหว่างวิธีตรวจสอบ เมื่อกลุ่มตัวอย่าง 300 คน และความยาวแบบทดสอบ 20 ข้อ.....	60
6	ร้อยละของความสอดคล้องในการตรวจสอบ ข้อสอบที่ทำหน้าที่ต่างกัน และ สัมประสิทธิ์สหสัมพันธ์แบบเรียงอันดับ ระหว่างวิธีตรวจสอบ.....	60
7	จำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบเมื่อกลุ่มตัวอย่าง 500 คน และ ความยาวแบบทดสอบ 20 ข้อ.....	61
8	ร้อยละของความสอดคล้องในการตรวจสอบ ข้อสอบที่ทำหน้าที่ต่างกัน และ สัมประสิทธิ์สหสัมพันธ์แบบเรียงอันดับ ระหว่างวิธีตรวจสอบ.....	62
9	จำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบเมื่อกลุ่มตัวอย่าง 700 คน และ ความยาวแบบทดสอบ 20 ข้อ.....	63
10	ผลการทดสอบความแตกต่างของจำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบ ระหว่างวิธีตรวจสอบ เมื่อกลุ่มตัวอย่าง 700 คน และความยาวแบบทดสอบ 20 ข้อ.....	64
11	ร้อยละของความสอดคล้องในการตรวจสอบ ข้อสอบที่ทำหน้าที่ต่างกัน และ สัมประสิทธิ์สหสัมพันธ์แบบเรียงอันดับ ระหว่างวิธีตรวจสอบ.....	64
12	จำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบเมื่อกลุ่มตัวอย่าง 1,000 คน และ ความยาวแบบทดสอบ 20 ข้อ.....	65
13	ผลการทดสอบความแตกต่างของจำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบ ระหว่างวิธีตรวจสอบ เมื่อกลุ่มตัวอย่าง 1,000 คน และความยาวแบบทดสอบ 20 ข้อ	66
14	ร้อยละของความสอดคล้องในการตรวจสอบ ข้อสอบที่ทำหน้าที่ต่างกัน และ สัมประสิทธิ์สหสัมพันธ์แบบเรียงอันดับ ระหว่างวิธีตรวจสอบ.....	66
15	จำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบเมื่อกลุ่มตัวอย่าง 300 คน และ ความยาวแบบทดสอบ 30 ข้อ.....	67

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

104	ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบ การทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 1,000 คน ความยาว แบบทดสอบ 60 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน ร้อยละ 20...	148
105	จำนวนความสอดคล้องในการตรวจสอบข้อสอบทำหน้าที่ต่างกันแต่ละวิธี เมื่อ กลุ่มตัวอย่าง 300 คน ความยาวแบบทดสอบ 20, 30, 40 และ 60 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน 0%, 10%, 20%.....	149
106	อัตราความสอดคล้องในการตรวจสอบข้อสอบทำหน้าที่ต่างกันแต่ละวิธี เมื่อ กลุ่มตัวอย่าง 300 คน ความยาวแบบทดสอบ 20, 30, 40 และ 60 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน 0%, 10%, 20%.....	151
107	จำนวนความสอดคล้องในการตรวจสอบข้อสอบทำหน้าที่ต่างกันแต่ละวิธี เมื่อ กลุ่มตัวอย่าง 500 คน ความยาวแบบทดสอบ 20, 30, 40 และ 60 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน 0%, 10%, 20%.....	153
108	อัตราความสอดคล้องในการตรวจสอบข้อสอบทำหน้าที่ต่างกันแต่ละวิธี เมื่อ กลุ่มตัวอย่าง 500 คน ความยาวแบบทดสอบ 20, 30, 40 และ 60 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน 0%, 10%, 20%.....	155
109	จำนวนความสอดคล้องในการตรวจสอบข้อสอบทำหน้าที่ต่างกันแต่ละวิธี เมื่อ กลุ่มตัวอย่าง 700 คน ความยาวแบบทดสอบ 20, 30, 40 และ 60 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน 0%, 10%, 20%.....	156
110	อัตราความสอดคล้องในการตรวจสอบข้อสอบทำหน้าที่ต่างกันแต่ละวิธี เมื่อ กลุ่มตัวอย่าง 700 คน ความยาวแบบทดสอบ 20, 30, 40 และ 60 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน 0%, 10%, 20%.....	158
111	จำนวนความสอดคล้องในการตรวจสอบข้อสอบทำหน้าที่ต่างกันแต่ละวิธี เมื่อ กลุ่มตัวอย่าง 1,000 คน ความยาวแบบทดสอบ 20, 30, 40 และ 60 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน 0%, 10%, 20%.....	160
112	อัตราความสอดคล้องในการตรวจสอบข้อสอบทำหน้าที่ต่างกันแต่ละวิธี เมื่อ กลุ่มตัวอย่าง 1,000 คน ความยาวแบบทดสอบ 20, 30, 40 และ 60 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน 0%, 10%, 20%.....	162
113	ค่าพารามิเตอร์ของข้อสอบ ที่ประมาณโดยใช้แบบจำลอง 3 พารามิเตอร์.....	186
114	ค่าความยาก (p) และค่าอำนาจจำแนก (r) ของแบบทดสอบ แยกตามเพศ ชาย - หญิง.....	189
115	การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ วิชาคณิตศาสตร์ ชั้นประถมศึกษา ปีที่ 4.....	192
116	ตัวอย่างการคำนวณการทดสอบความแตกต่างของจำนวนข้อสอบที่ทำหน้าที่ต่างกัน ที่ตรวจพบในแต่ละวิธีตรวจสอบ.....	195

บัญชีภาพประกอบ

ภาพประกอบ

หน้า

1	ข้อสอบที่ทำหน้าที่ต่างกันระหว่างกลุ่มย่อย 2 กลุ่ม.....	3
2	การทำหน้าที่ต่างกันของข้อสอบเมื่อใช้วิธีภายนอกตรวจสอบ.....	17
3	โค้งแสดงลักษณะข้อสอบแบบ 3 พารามิเตอร์.....	24
4	ผลการตรวจสอบความเป็นมิติเดียวของแบบทดสอบ.....	54
5	โค้งแสดงลักษณะข้อสอบโดยเฉลี่ยของแบบทดสอบ.....	57
6	โค้งสารสนเทศของแบบทดสอบ (Test Information Curve).....	58
7	จำนวนความไม่สอดคล้อง เมื่อ $N = 300$, $I = 20$, $0\%DIF$	98
8	จำนวนความไม่สอดคล้อง เมื่อ $N = 300$, $I = 20$, $10\%DIF$	98
9	จำนวนความไม่สอดคล้อง เมื่อ $N = 300$, $I = 20$, $20\%DIF$	98
10	จำนวนความไม่สอดคล้อง เมื่อ $N = 300$, $I = 30$, $0\%DIF$	99
11	จำนวนความไม่สอดคล้อง เมื่อ $N = 300$, $I = 30$, $10\%DIF$	99
12	จำนวนความไม่สอดคล้อง เมื่อ $N = 300$, $I = 30$, $20\%DIF$	99
13	จำนวนความไม่สอดคล้อง เมื่อ $N = 300$, $I = 40$, $0\%DIF$	100
14	จำนวนความไม่สอดคล้อง เมื่อ $N = 300$, $I = 40$, $10\%DIF$	100
15	จำนวนความไม่สอดคล้อง เมื่อ $N = 300$, $I = 40$, $20\%DIF$	100
16	จำนวนความไม่สอดคล้อง เมื่อ $N = 300$, $I = 60$, $0\%DIF$	101
17	จำนวนความไม่สอดคล้อง เมื่อ $N = 300$, $I = 60$, $10\%DIF$	101
18	จำนวนความไม่สอดคล้อง เมื่อ $N = 300$, $I = 60$, $20\%DIF$	101
19	จำนวนความไม่สอดคล้อง เมื่อ $N = 500$, $I = 20$, $0\%DIF$	111
20	จำนวนความไม่สอดคล้อง เมื่อ $N = 500$, $I = 20$, $10\%DIF$	111
21	จำนวนความไม่สอดคล้อง เมื่อ $N = 500$, $I = 20$, $20\%DIF$	111
22	จำนวนความไม่สอดคล้อง เมื่อ $N = 500$, $I = 30$, $0\%DIF$	112
23	จำนวนความไม่สอดคล้อง เมื่อ $N = 500$, $I = 30$, $10\%DIF$	112
24	จำนวนความไม่สอดคล้อง เมื่อ $N = 500$, $I = 30$, $20\%DIF$	112
25	จำนวนความไม่สอดคล้อง เมื่อ $N = 500$, $I = 40$, $0\%DIF$	113
26	จำนวนความไม่สอดคล้อง เมื่อ $N = 500$, $I = 40$, $10\%DIF$	113
27	จำนวนความไม่สอดคล้อง เมื่อ $N = 500$, $I = 40$, $20\%DIF$	113
28	จำนวนความไม่สอดคล้อง เมื่อ $N = 500$, $I = 60$, $0\%DIF$	114
29	จำนวนความไม่สอดคล้อง เมื่อ $N = 500$, $I = 60$, $10\%DIF$	114
30	จำนวนความไม่สอดคล้อง เมื่อ $N = 500$, $I = 60$, $20\%DIF$	114

31	จำนวนความไม่สอดคล้อง เมื่อ $N = 700, I = 20, 0\%DIF$	123
32	จำนวนความไม่สอดคล้อง เมื่อ $N = 700, I = 20, 10\%DIF$	123
33	จำนวนความไม่สอดคล้อง เมื่อ $N = 700, I = 20, 20\%DIF$	123
34	จำนวนความไม่สอดคล้อง เมื่อ $N = 700, I = 30, 0\%DIF$	124
35	จำนวนความไม่สอดคล้อง เมื่อ $N = 700, I = 30, 10\%DIF$	124
36	จำนวนความไม่สอดคล้อง เมื่อ $N = 700, I = 30, 20\%DIF$	124
37	จำนวนความไม่สอดคล้อง เมื่อ $N = 700, I = 40, 0\%DIF$	125
38	จำนวนความไม่สอดคล้อง เมื่อ $N = 700, I = 40, 10\%DIF$	125
39	จำนวนความไม่สอดคล้อง เมื่อ $N = 700, I = 40, 20\%DIF$	125
40	จำนวนความไม่สอดคล้อง เมื่อ $N = 700, I = 60, 0\%DIF$	126
41	จำนวนความไม่สอดคล้อง เมื่อ $N = 700, I = 60, 10\%DIF$	126
42	จำนวนความไม่สอดคล้อง เมื่อ $N = 700, I = 60, 20\%DIF$	126
43	จำนวนความไม่สอดคล้อง เมื่อ $N = 1,000, I = 20, 0\%DIF$	136
44	จำนวนความไม่สอดคล้อง เมื่อ $N = 1,000, I = 20, 10\%DIF$	136
45	จำนวนความไม่สอดคล้อง เมื่อ $N = 1,000, I = 20, 20\%DIF$	136
46	จำนวนความไม่สอดคล้อง เมื่อ $N = 1,000, I = 30, 0\%DIF$	137
47	จำนวนความไม่สอดคล้อง เมื่อ $N = 1,000, I = 30, 10\%DIF$	137
48	จำนวนความไม่สอดคล้อง เมื่อ $N = 1,000, I = 30, 20\%DIF$	137
49	จำนวนความไม่สอดคล้อง เมื่อ $N = 1,000, I = 40, 0\%DIF$	138
50	จำนวนความไม่สอดคล้อง เมื่อ $N = 1,000, I = 40, 10\%DIF$	138
51	จำนวนความไม่สอดคล้อง เมื่อ $N = 1,000, I = 40, 20\%DIF$	138
52	จำนวนความไม่สอดคล้อง เมื่อ $N = 1,000, I = 60, 0\%DIF$	139
53	จำนวนความไม่สอดคล้อง เมื่อ $N = 1,000, I = 60, 10\%DIF$	139
54	จำนวนความไม่สอดคล้อง เมื่อ $N = 1,000, I = 60, 20\%DIF$	139

บทที่ 1

บทนำ

ภูมิหลัง

แบบทดสอบเป็นเครื่องมือชนิดหนึ่ง ที่นิยมใช้ในการวัดผลการศึกษาในเกือบทุกระดับการศึกษา เนื่องจากมีความสะดวกในการนำมาใช้ ครูและนักการศึกษาส่วนใหญ่มีความเข้าใจและคุ้นเคยกับเครื่องมือประเภทนี้ รวมถึงขั้นตอนการสร้างก็ไม่ยุ่งยากจนเกินไป หากได้ทำการศึกษาถึงลักษณะเด่นเฉพาะของแบบทดสอบแต่ละชนิดและวิธีการสร้าง จะทำให้สามารถเลือกใช้ได้เหมาะสมกับจุดมุ่งหมายที่ต้องการวัด เช่น แบบทดสอบประเภทความเรียงเหมาะสำหรับใช้วัดความสามารถในการบูรณาการความรู้เขียนเป็นเรื่องราว เชิงวิเคราะห์วิจารณ์ แบบทดสอบแบบเติมคำเหมาะสำหรับการวัดความรู้เกี่ยวกับมโนทัศน์ ความหมายในเนื้อหาที่ต้องการวัดหรือคำสำคัญ และแบบทดสอบประเภทแบบเลือกตอบเป็นแบบทดสอบที่ใช้วัดความสามารถได้ตั้งแต่ขั้นความรู้ความจำไปจนถึงการสังเคราะห์และประเมินค่า ประกอบกับมีจุดเด่นที่สำคัญอีกอย่าง คือ มีความเป็นปรนัย (Objectivity) สูง หมายถึง การเข้าใจคำถามตรงกัน การตรวจให้คะแนนตรงกัน และการแปลผลคะแนนที่ได้ตรงกัน ด้วยเหตุนี้จึงได้รับความนิยมอย่างแพร่หลาย

คุณสมบัติของแบบทดสอบที่ดีประกอบด้วย ความเที่ยงตรง (Validity) ความเชื่อมั่น (Reliability) งามลึก (Searching) ยั่วให้อยากตอบ (Exemplary) มีความเป็นปรนัย (Objectivity) สูง และมีความยุติธรรม (Fairness) มีความจำเพาะเจาะจง (Definite) มีประสิทธิภาพ (Efficiency) มีความยากง่าย (Difficulty) พอเหมาะ และมีอำนาจจำแนก (Discrimination) (ชวาล แพริตกุล. 2509 : 10 - 125) สำหรับความเที่ยงตรงนั้นอาจกล่าวได้ว่าเป็นคุณสมบัติที่สำคัญที่สุดทั้งนี้เพราะความเที่ยงตรงของแบบทดสอบหรือเครื่องมือวัดจะเป็นตัวชี้ให้เห็นว่า แบบทดสอบหรือเครื่องมือวัดนั้น ๆ วัดได้ตรงตามเป้าหมายของคุณลักษณะ (Trait) หรือความสามารถ (Ability) ที่ต้องการจะวัด ถ้าหากแบบทดสอบขาดความเที่ยงตรงแล้วก็จะทำให้เกิดความคลาดเคลื่อนในการแปลผลคะแนนที่ได้จากการสอบวัด

โดยทั่วไปการสร้างแบบทดสอบจะมุ่งวัดความสามารถหรือคุณลักษณะด้านใดด้านหนึ่ง ผู้สร้างจะเขียนข้อคำถามวัดคุณลักษณะเฉพาะด้านนั้น ๆ นั่นคือ การพิจารณาหรือคำนึงถึงความเป็นมิติเดียวของแบบทดสอบ (Unidimensionality) ซึ่งทฤษฎีการทดสอบแนวใหม่ เรียกว่า “ ทฤษฎีการตอบข้อสอบ ” (Item Response Theory : IRT) ถือเป็นข้อตกลงเบื้องต้นที่สำคัญอันดับแรกสำหรับทฤษฎีการทดสอบแบบดั้งเดิม (Classical Test Theory : CTT) นั้นใช้ในความหมายของความเป็นเอกพันธ์ของแบบทดสอบ (Homogeneity) หมายถึง ข้อสอบแต่ละข้อวัดในองค์ประกอบร่วมเดียวกันเรียงตามลำดับความยากง่าย (Guilford. 1954 : 363 ; citing Loevinger. 1947) อย่างไรก็ตามเมื่อนำแบบทดสอบไปสอบวัดกับกลุ่มตัวอย่างที่มีความสามารถเท่ากัน แต่มีความแตกต่างกันเกี่ยวกับ เพศ เชื้อชาติ ศาสนา ภูมิสำเนา และระดับสติปัญญา ฯลฯ แล้ว โอกาสในการตอบข้อสอบ

ข้อนั้นได้ถูกต้องไม่เท่าเทียมกัน เรียกว่า ข้อสอบหรือแบบทดสอบนั้นทำหน้าที่ต่างกัน (Differential Item / Test Functioning : DIF, DTF) ทำให้เกิดการได้เปรียบเสียเปรียบกันขึ้นระหว่างกลุ่มย่อยต่าง ๆ นั่นคือ เกิดความลำเอียงหรือความไม่ยุติธรรมขึ้นในการสอบ ซึ่งได้รับการวิพากษ์วิจารณ์และเป็นเรื่องสำคัญอีกอย่างหนึ่งที่ผู้เกี่ยวข้องกับกระบวนการทดสอบต้องคำนึงถึงและให้ความสนใจ ทั้งนี้เพราะหากข้อสอบหรือแบบทดสอบมีความลำเอียงโดยเข้าข้างกลุ่มใดกลุ่มหนึ่ง หรือเอื้อประโยชน์ต่อคนบางกลุ่มแล้ว จะทำให้ผลการสอบขาดความน่าเชื่อถือ และเกิดความไม่เสมอภาคในการตัดสิน

กระบวนการในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสามารถทำได้ 2 ขั้นตอน (อุทุมพร จามรมาน. 2537 : 129) ดังนี้

1. การตรวจสอบในขั้นตอนการสร้างข้อสอบ โดยให้ผู้เชี่ยวชาญพิจารณาว่า ข้อสอบวัดตรงจุดมุ่งหมายหรือพฤติกรรมและเนื้อหาที่ต้องการวัดหรือไม่จากตารางวิเคราะห์หลักสูตร (Table of Specification) รวมทั้งพิจารณาภาษาที่ใช้สามารถสื่อความหมายได้ชัดเจนไม่คลุมเครือ เวลานำไปใช้ก็ไม่เกิดปัญหาในการตีความ หรือเกิดประโยชน์ต่อคนบางกลุ่ม ซึ่งเป็นการสนับสนุนการตรวจสอบความเที่ยงตรงเชิงโครงสร้าง (Construct Validity) และความเที่ยงตรงเชิงเนื้อหา (Content Validity) ไปพร้อมกัน

2. การตรวจสอบโดยใช้สถิติ หลังจากที่น่าแบบทดสอบไปใช้แล้ว สามารถทำได้ 2 แนวทาง ดังนี้

2.1 ใช้วิธีภายนอก (External Method) หรือเกณฑ์ภายนอก (External Criterion) โดยใช้แบบทดสอบที่สร้างขึ้นหาความสัมพันธ์กับเกณฑ์ที่เป็นมาตรฐาน เช่น เกณฑ์เฉลี่ย หรือ ผลการปฏิบัติงาน เป็นต้น วิธีนี้จะมีปัญหาก่อนอยู่ที่เกณฑ์ภายนอกที่นำมาหาความสัมพันธ์ เนื่องจากถ้าเกณฑ์ไม่มีมาตรฐานแล้วจะทำให้การตรวจสอบขาดความถูกต้อง ดังนั้นนักวัดผลส่วนใหญ่จึงไม่นิยมใช้วิธีนี้

2.2 ใช้วิธีภายใน (Internal Method) หรือ เกณฑ์ภายใน (Internal Criterion) เป็นวิธีที่ได้รับความนิยมอย่างมาก เนื่องจากไม่ต้องกังวลเรื่องเกณฑ์ สามารถศึกษาจากการตรวจสอบโครงสร้างภายในของข้อสอบ โดยพิจารณาคะแนนที่ได้จากผลการตอบของผู้เข้าสอบแต่ละกลุ่มว่าวัดในคุณลักษณะที่ต้องการวัดตามโครงสร้างเช่นเดียวกันหรือไม่ ถือเป็นกระบวนการของการสนับสนุนการตรวจสอบความเที่ยงตรงเชิงโครงสร้าง การตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันตามแนวนี้สามารถแบ่งออกเป็น 3 กลุ่มใหญ่ ๆ (Camilli and Shepard. 1994 : 22 - 131) ดังนี้

2.2.1 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามทฤษฎีการทดสอบดั้งเดิม (Classical Test Theory) การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามวิธีนี้ ได้แก่ วิธีการแปลงค่าความยาก (Transformation Item Difficulty) วิธี Golden Rule วิธีวิเคราะห์ความแปรปรวน (Analysis of Variance) และวิธีพอยต์ไบซีเรียล (Point Biserial) วิธีนี้มีปัญหาตรงที่ไม่สามารถแยกลักษณะประจำของข้อสอบระหว่างค่าความยากกับค่าอำนาจจำแนกออกจากกันได้ ทำให้เกิดความคลาดเคลื่อนในการตรวจสอบค่อนข้างสูง

2.2.2 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้ตารางการถัว (Contingency Table) การตรวจสอบในแนวนี้ ได้แก่ วิธีแมนเทล ฮาเซนเซล (Mantel Haenzel) และวิธีไคสแควร์ เป็นต้น วิธีนี้ค่อนข้างสะดวก สามารถคำนวณด้วยเครื่องคิดเลขได้ไม่ต้องอาศัยเครื่องคอมพิวเตอร์

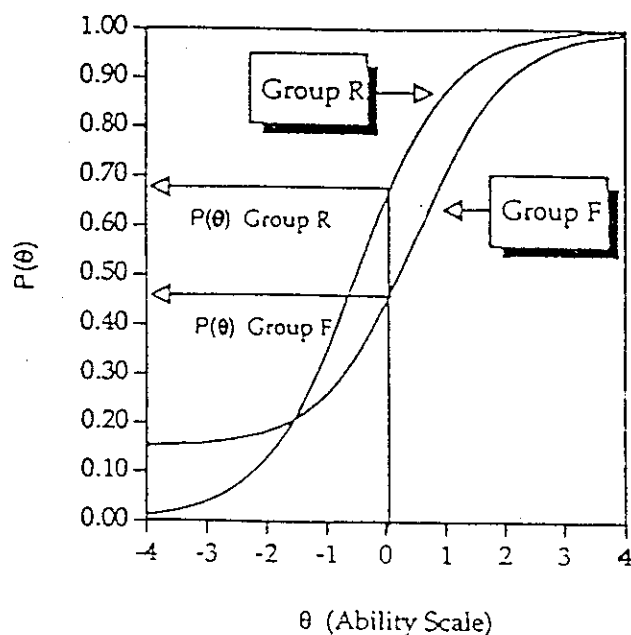
แต่มีจุดอ่อนที่ขาดประสิทธิภาพในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ (Nonuniform Differential Item Functioning) แต่ปัจจุบันได้ปรับปรุงการคำนวณให้มีประสิทธิภาพในการตรวจสอบ ได้ดีขึ้น (เสรี ชัดแจ้ง. 2540)

2.2.3 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามแนวคิดของทฤษฎีการตอบข้อสอบ เป็นแนวคิดที่พิจารณาผลการตอบข้อสอบกับโอกาสตอบข้อสอบของผู้ที่มีความสามารถระดับเดียวกันแต่ต่างกลุ่มกัน เช่น หญิง และชาย หากเส้นโค้งคุณลักษณะข้อสอบ (Item Characteristic Curve) มีความแตกต่างกัน นั่นคือ เกิดพื้นที่ขึ้นระหว่างเส้นโค้งคุณลักษณะข้อสอบ 2 เส้น แสดงว่าข้อสอบข้อนั้นทำหน้าที่ต่างไปจากปกติทำให้เกิดการได้เปรียบและเสียเปรียบกันขึ้น แนวคิดนี้มีความสมเหตุสมผลมากกว่าแนวคิดอื่น เนื่องจากมีจุดเด่นกว่าแนวคิดทฤษฎีการทดสอบแบบดั้งเดิม และวิธีตารางการณักร ดังนี้

1) ทฤษฎีการตอบข้อสอบสามารถประมาณค่าพารามิเตอร์ของข้อสอบแยกจากกันได้ชัดเจน เช่น ค่าความยาก (b) ค่าอำนาจจำแนก (a) และค่าการเดา (c) ทำให้ลดการปนกัน (Confounding) ของพารามิเตอร์ข้อสอบลงได้

2) คุณสมบัติทางสถิติของข้อสอบ (Statistical Properties of Items) สามารถอธิบายได้ละเอียดและชัดเจนกว่า เมื่อข้อสอบทำหน้าที่ต่างกันระหว่าง 2 กลุ่ม

3) คุณสมบัติทางสถิติของข้อสอบ สามารถแสดงด้วยแผนภาพที่เป็นกราฟ ทำให้เข้าใจง่ายและนำไปใช้ได้กว้างขวาง (ดังภาพประกอบ 1)



ภาพประกอบ 1 ข้อสอบที่ทำหน้าที่ต่างกันระหว่างกลุ่มย่อย 2 กลุ่ม (ค่าพารามิเตอร์ a,b,c แตกต่างกัน) (Camilli and Shepard. 1994 : 60)

4) สามารถตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอได้

การตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันตามแนวคิดของทฤษฎีการตอบข้อสอบ สามารถแบ่งได้เป็น 2 แนวทางใหญ่ ๆ (Hambleton, Swaminathan and Rogers. 1991 : 110 ; Kim and Cohen. 1995) ดังนี้

1) การเปรียบเทียบค่าพารามิเตอร์ของข้อสอบระหว่าง 2 กลุ่มย่อย (Subgroup) ว่ามีความแตกต่างกันหรือไม่ (Kim and Cohen. 1994 : 217 ; citing Draba. 1977 , Lord. 1977, 1980 ; Thissen, Steinberg and Wainer. 1988, 1993 ; Wright and Stone. 1979) ในกลุ่มนี้นิยมใช้วิธี Lord's χ^2 และล่าสุดใช้ Likelihood Ratio Test ในระยะแรกผู้วิจัยสนใจจะศึกษาทั้ง 2 วิธี แต่ในที่สุดก็เลือกเฉพาะวิธี Lord's χ^2 เนื่องจากวิธี Likelihood Ratio Test ค่อนข้างมีปัญหาในการคำนวณ แม้ว่าจะใช้โปรแกรม MULTOLOG ในการคำนวณก็ตาม จากการศึกษาของคิม และโคเฮน (Kim and Cohen. 1995 : 291 - 312) ต้องใช้เครื่องคอมพิวเตอร์ 5 เครื่อง (รุ่น 486 2 เครื่อง และรุ่น pentium 3 เครื่อง) วิเคราะห์ข้อมูลประมาณ 2 เดือนจึงเสร็จ ประกอบกับผลการวิจัยที่ได้ก็ค่อนข้างสอดคล้องกันมากกับวิธี Lord's χ^2 ทำให้ผู้วิจัยตัดออกจากกรอบของการศึกษาตามคำแนะนำของ คิม และโคเฮน

2) การวัดพื้นที่ความแตกต่างของฟังก์ชันการตอบข้อสอบระหว่าง 2 กลุ่มย่อย (Kim and Cohen. 1991 ; Linn, Levine, Hastings and Wardrop. 1981 ; Raju. 1988, 1990 ; Rudner. 1977 ; Wainer. 1993) ในกลุ่มนี้ระยะหลังนิยมใช้วิธี Raju's Area Measures และวิธี Closed Interval Area สำหรับการศึกษาครั้งนี้ ผู้วิจัยเลือกศึกษาทั้ง 2 วิธี เนื่องจากวิธี Raju's Area Measures ใช้ช่วงความสามารถ (Ability) ของผู้เข้าสอบเป็นช่วงเปิด บางครั้งเรียกว่า การวัดพื้นที่ชนิดสัมบูรณ์ (Exact Area Measures) ส่วนวิธี Closed Interval Area โคเฮน และคิม เป็นผู้เสนอให้ใช้ช่วงความสามารถของผู้เข้าสอบแบบช่วงปิด นั่นคือ กำหนดช่วงความสามารถสำหรับขอบเขตบนและขอบเขตล่าง

การศึกษาเกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยทฤษฎีการตอบข้อสอบนั้น ส่วนใหญ่จะศึกษาร่วมกับวิธีอื่น ๆ เช่น งานวิจัยของ ไอรอนสัน และซับโคเวียค (Ironson and Subkoviak. 1979) สงบ ลักษณะ (Laksana. 1979) เชปเพิร์ด และคณะ (Subkoviak et al. 1984 : 50 ; citing Shepard et al. 1981) ซับโคเวียค ไอรอนสัน และแครก (Subkoviak, Ironson and Craig. 1984) บาغيและเฟอร์รารา (Baghi and Ferrara. 1990) ชัชชัย เฝ้าพงศ์ (2527) ทศนีย์ พิรมนตรี (2529) สุศักดิ์ อมรัตน์ศักดิ์ (2531) พชร ปิยะภักดิ์ (2531) และเจน จันทรแสง (2537) ซึ่งส่วนใหญ่ศึกษาร่วมกับวิธีแปลงค่าความยาก วิธีไคสแควร์ วิธีวิเคราะห์ความแปรปรวน วิธีพอยท์ไบซีเรียล และวิธีแมนเทล แชนเซล พบว่า วิธีตรวจสอบการทำหน้าที่ต่างกันที่ใช้ทฤษฎีการตอบข้อสอบเป็นวิธีที่ดีที่สุด โดยเฉพาะวิธีทฤษฎีการตอบข้อสอบแบบ 3 พารามิเตอร์ เนื่องจากมีความครอบคลุมและเหมาะสมในเชิงทฤษฎีมากกว่าวิธีอื่น นอกจากนี้จากการศึกษาด้วย

ข้อมูลจำลองของรัตเนอร์ เกทสัน และไนท์ (Rudner, Getson and Knight. 1980) เพิร์ลแมน และคณะ (Perlman et al. 1985) เชปเพิร์ด คามิลลี และวิลเลียมส์ (Shepard, Camilli and Williams. 1985) รวมทั้งเบจาร์ ฮันเตอร์ ลอร์ด และปีเตอร์สัน (Berk. 1982 : 119 ; citing Bejar. 1980, Hunter. 1978, Lord. 1980 and Peterson. 1977) ยังพบอีกว่า วิธีตรวจสอบโดยใช้ทฤษฎีการตอบข้อสอบชนิด 2 และ 3 พารามิเตอร์ มีความถูกต้องและสอดคล้องกันมากที่สุดเมื่อเทียบกับวิธีอื่น ๆ แม้ว่าจะไม่สามารถตรวจสอบได้ถูกต้องถึงร้อยละ 100 ก็ตาม ทั้งนี้เพราะเกิดความคลาดเคลื่อนจากผลการตรวจสอบอยู่ 2 ประการ คือ การยอมรับผิดพลาด (False Positive : FP) (ตัดสินว่าข้อสอบทำหน้าที่ต่างกัน ทั้ง ๆ ที่ทำหน้าที่ไม่ต่างกัน) และการปฏิเสธผิดพลาด (False Negative : FN) (ตัดสินว่า ข้อสอบทำหน้าที่ไม่ต่างกัน ทั้ง ๆ ที่ข้อสอบทำหน้าที่ต่างกัน)

สำหรับงานวิจัยที่ศึกษาการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันตามแนวคิดของทฤษฎีการตอบข้อสอบเพียงอย่างเดียวก็พอมีบ้าง แต่ไม่มากนัก ได้แก่

คิม และโคเฮน (Kim and Cohen. 1991) ได้ศึกษาเปรียบเทียบความแตกต่างของพื้นที่แบบไม่มีเครื่องหมาย (Unsigned Area) และแบบมีเครื่องหมาย (Signed Area) พบว่า พื้นที่แบบมีเครื่องหมายมีความคลาดเคลื่อนต่ำกว่าแบบไม่มีเครื่องหมาย และการวัดพื้นที่ช่วงเปิดและแบบช่วงปิดจะให้ผลสอดคล้องกันค่อนข้างมาก

โคเฮน และคิม (Cohen and Kim. 1993) ได้ศึกษาเปรียบเทียบการใช้สถิติทดสอบ χ^2 ของลอร์ด และสถิติทดสอบพื้นที่ความแตกต่างระหว่างโค้งลักษณะข้อสอบชนิดมีเครื่องหมาย และไม่มีเครื่องหมาย สถิติ Z ของราชู โดยจำลองข้อมูลตามทฤษฎีการตอบข้อสอบชนิด 2 พารามิเตอร์ กำหนดให้แบบทดสอบ ยาว 20 และ 60 ข้อ จำนวนผู้สอบ 100 และ 500 และมีจำนวนข้อสอบที่ทำหน้าที่ต่างกันแบบทดสอบจำนวนร้อยละ 0 , 10 และ 20 พบว่า

1. แบบทดสอบที่มีความยาวมากขึ้นและมีจำนวนข้อสอบที่ทำหน้าที่ต่างกันแบบทดสอบน้อย ความคลาดเคลื่อนในการยอมรับผิดพลาด (False Positive : FP) จะมากขึ้น และการกำหนดระดับนัยสำคัญต่างกันก็มีผลต่อความคลาดเคลื่อนประเภทนี้ (ที่ระดับ .05 จะเกิด FP มากกว่าที่ระดับ .01) สำหรับสถิติทดสอบ Z จะเกิดขึ้นทุกเงื่อนไขของการศึกษา ส่วนสถิติทดสอบ χ^2 จะเกิดขึ้นบางเงื่อนไข

2. แบบทดสอบที่มีความยาวมากขึ้นและมีจำนวนข้อสอบที่ทำหน้าที่ต่างกันแบบทดสอบมาก ความคลาดเคลื่อนในการปฏิเสธผิดพลาด (False Negative : FN) จะมากขึ้น โดยสถิติทดสอบ Z จะเกิดขึ้นมากกว่าสถิติทดสอบ χ^2

3. ความแตกต่างของการเกิดความคลาดเคลื่อนในการปฏิเสธผิดพลาดระหว่างสถิติทดสอบ Z และ χ^2 เมื่อจำนวนผู้เข้าสอบ 100 คน จะน้อยกว่าเมื่อจำนวนผู้เข้าสอบ 500 คน และสถิติทดสอบ χ^2 เกิดความคลาดเคลื่อนในการปฏิเสธผิดพลาดน้อยกว่าสถิติทดสอบ Z โดยเฉพาะกรณีที่มีจำนวนข้อสอบที่ทำหน้าที่ต่างกันแบบทดสอบมากที่สุด คือ ร้อยละ 20

คิม และโคเฮน (Kim and Cohen. 1994) และโคเฮน คิม และวอลแลค (Cohen, Kim and Wollack. 1995) ได้ศึกษาอัตราการเกิดความคลาดเคลื่อนชนิดที่ 1 (Type I error rates) หรือ

การยอมรับผิดพลาด (False Positive) ของวิธีลอร์ด (Lord's χ^2) และวิธีการทดสอบอัตราส่วนไลเคิลฮูด (Likelihood Ratio Test) จำลองข้อมูลโดยกำหนดขนาดกลุ่มตัวอย่างเป็น 250 และ 1,000 คน จำนวนข้อสอบ 50 ข้อ และแบบจำลองที่ศึกษาเป็นแบบจำลอง 2, 3 และ 3 พารามิเตอร์ ที่กำหนดให้ค่าการเดาคงที่ พบว่า โดยภาพรวมอัตราการเกิดความคลาดเคลื่อนชนิดที่ 1 ของวิธีการทดสอบอัตราส่วนไลเคิลฮูด จะมีค่าใกล้เคียงค่าคาดหวังในทางทฤษฎีมากกว่าวิธีลอร์ด ณ ระดับนัยสำคัญ .01 และ .001 แต่ ณ ระดับนัยสำคัญ (α) .05 ทั้งสองวิธีให้ค่าใกล้เคียงกัน เมื่อใช้แบบจำลอง 2 และ 3 พารามิเตอร์ ที่กำหนดให้ค่าการเดาคงที่ อย่างไรก็ตาม สำหรับวิธีลอร์ดเมื่อใช้แบบจำลอง 3 พารามิเตอร์ จะเกิดความคลาดเคลื่อนชนิดที่ 1 สูงกว่าเมื่อใช้แบบจำลอง 2 พารามิเตอร์และ 3 พารามิเตอร์ ที่กำหนดให้ค่าการเดาคงที่

คิม และโคเฮน (Kim and Cohen. 1995) ได้ศึกษาเปรียบเทียบผลการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันระหว่างวิธีลอร์ด วิธีราชู และวิธีการทดสอบอัตราส่วนไลเคิลฮูด โดยศึกษาจากข้อมูลแบบทดสอบจัดชั้นเรียนของนักศึกษาระดับมหาวิทยาลัย วิชาคณิตศาสตร์ พบว่า ทั้ง 3 วิธี ให้ผลการตรวจสอบสอดคล้องกันมาก

จากการศึกษาและงานวิจัยที่กล่าวมาข้างต้น พอสรุปประเด็นที่น่าสนใจเกี่ยวกับการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันได้ดังนี้

1. วิธีที่เหมาะสมและดีที่สุดทางทฤษฎี รวมทั้งอาจกล่าวได้ว่าในทางปฏิบัติด้วย ได้แก่ วิธีการตรวจสอบตามแนวคิดของทฤษฎีการตอบข้อสอบ ชนิด 2 และ 3 พารามิเตอร์

เนื่องจากปัจจุบันมีการพัฒนาในด้านโปรแกรมคอมพิวเตอร์ไปมาก จึงไม่ค่อยมีปัญหาในการคิดคำนวณ เช่น โปรแกรม IRTDIF GRAPHDIF และ Lord_DIF เป็นต้น

2. ตัวแปรที่เกี่ยวข้องในการสอบ ได้แก่ จำนวนผู้เข้าสอบ จำนวนข้อสอบ และสัดส่วนจำนวนข้อสอบที่ลำเอียงในแบบทดสอบ ซึ่งแต่ละวิธีที่ใช้ตรวจสอบจะมีความเจ็บคมหรือไวในการตรวจสอบต่างกัน รวมทั้งมีความไม่คงที่ในการตัดสินข้อสอบที่ทำหน้าที่ต่างกับกับความยาวข้อสอบ จำนวนผู้เข้าสอบ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันแบบทดสอบแตกต่างกัน เช่น งานวิจัยของรูดเนอร์ เกทสัน และไนท์ (Rudner, Getson and Knight. 1980) บาเกียและเฟอร์รารา (Baghi and Ferrara. 1990) โคเฮน และคิม (Cohen and Kim. 1993) และงานวิจัยของกาญจนา วัฒนสุนทร (2537) เป็นต้น

3. เกณฑ์ที่บ่งชี้ประสิทธิภาพของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ หรือแบบทดสอบ ได้แก่ ความคลาดเคลื่อนในการยอมรับผิดพลาด หรือความคลาดเคลื่อนประเภทที่ 1 และการปฏิเสธผิดพลาด หรือความคลาดเคลื่อนชนิดที่ 2 ในกรณีศึกษาจากข้อมูลจำลอง เช่น งานวิจัยของ โคเฮน และคิม (Cohen and Kim. 1993) คิม และโคเฮน (Kim and Cohen. 1994) โคเฮน คิม และวอลแลค (Cohen, Kim and Wollack. 1995) เป็นต้น และจำนวนข้อสอบที่พบว่ามีค่าทำหน้าที่ต่างกันในแต่ละวิธี หรือความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เช่น งานวิจัยของสงบ ลักษณะ (Laksana. 1979) ไอรอนสัน และซับโคเวียค (Ironson and Subkoviak. 1979) ราชู และคณะ (Raju et al. 1993) คิม และโคเฮน (Kim and Cohen. 1995)

ทัศนีย์ พิรมนตรี (2529) สุรศักดิ์ อมรรัตนศักดิ์ (2531) และกาญจนา วัฒนสุนทร (2537) เป็นต้น

อนึ่ง วิธีที่นิยมใช้ศึกษาเกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามแนวคิด ทฤษฎีการตอบข้อสอบ ในระยะหลัง ได้แก่ วิธี Lord's χ^2 วิธี Raju's Area Measures และวิธี Closed Interval Area ซึ่งมีการศึกษาทั้งจากข้อมูลจริงและข้อมูลจำลอง แต่ยังมีประเด็นที่ไม่มีคำตอบ ชัดเจนว่าวิธีใดมีประสิทธิภาพดีกว่าในเงื่อนไขของการศึกษาที่แตกต่างกัน เช่น ขนาดของกลุ่ม ตัวอย่าง ความยาวของแบบทดสอบ และจำนวนข้อสอบที่ทำหน้าที่ต่างกันแบบทดสอบ แม้ว่าจะ มีผู้ศึกษาอยู่บ้าง อย่างไรก็ตาม หากมีการขยายขนาดของกลุ่มตัวอย่าง ความยาวของแบบทดสอบ และจำนวนข้อสอบที่ทำหน้าที่ต่างกันแบบทดสอบให้ละเอียดมากขึ้น จะทำให้ข้อสรุปจากการศึกษา มีความน่าเชื่อถือและสมบูรณ์ยิ่งขึ้น

ดังนั้น ผู้วิจัยจึงสนใจที่จะศึกษาผลการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันระหว่าง วิธี Lord's χ^2 วิธี Raju's Area Measures และวิธี Closed Interval Area ตามเงื่อนไขของการศึกษา ค้นคว้าดังกล่าว สำหรับการศึกษาครั้งนี้ได้เลือกข้อมูลผลการสอบประเมินคุณภาพและความก้าวหน้าทางการศึกษาระดับประเทศ ปีการศึกษา 2536 วิชาคณิตศาสตร์ ชั้นประถมศึกษาปีที่ 4 ของ สำนักงานการประถมศึกษาแห่งชาติ เนื่องจากมีขนาดกลุ่มตัวอย่างและความยาวข้อสอบมากเพียงพอ ต่อการสุ่ม ทำให้สามารถดำเนินการตามกรอบเงื่อนไขของการศึกษาได้ และกำหนดตัวแปรจำแนก เป็นเพศ ทั้งนี้เพราะจากการศึกษา พบว่า ความสามารถทางคณิตศาสตร์ระหว่างเพศมีความ แตกต่างกัน (Doolittle and Cleary. 1987 ; Gallagher and Lisi. 1992 ; Harris and Carlton. 1993 ; Neil and McPeck. 1993 ; Wang and Lane. 1996)

ความมุ่งหมายของการศึกษา

1. เพื่อเปรียบเทียบจำนวนข้อสอบที่ทำหน้าที่ต่างกันและศึกษาความสอดคล้องของผลการ ตรวจสอบระหว่าง วิธี Lord's χ^2 วิธี Raju's Area Measures และวิธี Closed Interval Area เมื่อ ขนาดของกลุ่มตัวอย่าง และความยาวของแบบทดสอบ ต่างกัน
2. เพื่อเปรียบเทียบอัตราความไม่สอดคล้องในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน ระหว่างวิธี Lord's χ^2 วิธี Raju's Area Measures และวิธี Closed Interval Area เมื่อขนาดของ กลุ่มตัวอย่าง ความยาวของแบบทดสอบ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันแบบทดสอบ ต่างกัน

สมมติฐานในการวิจัย

จากการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องกับการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน ตามแนวคิดของทฤษฎีการตอบข้อสอบ ที่ใช้วิธี Lord's χ^2 วิธี Raju's Area Measures และวิธี Closed Interval Area ได้แก่ การศึกษาของ กาญจนนา วันสุนทร (2537) พบว่า จำนวนข้อสอบ ในรูปของร้อยละของข้อสอบที่ทำหน้าที่ต่างกัน และความสอดคล้องของการตัดสินข้อสอบที่ทำหน้าที่ ต่างกันมีความไม่คงที่ข้ามความยาวแบบทดสอบ และขนาดของกลุ่มตัวอย่าง แม้ว่าจากการศึกษา ของคิม และโคเฮน (Kim and Cohen. 1991, 1995) จะพบว่ามีความความสอดคล้องกันค่อนข้าง มากก็ตาม แต่การศึกษาของคิม และโคเฮนใช้ความยาวของแบบทดสอบและขนาดของกลุ่มตัวอย่าง คงที่ครั้งเดียว ข้อค้นพบจากการศึกษาจึงยังไม่ครอบคลุม ดังนั้น ผู้วิจัยจึงตั้งสมมติฐาน ดังนี้

1. เมื่อขนาดของกลุ่มตัวอย่างต่างกัน จำนวนข้อสอบที่พบว่าทำหน้าที่ต่างกันจากการ ตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน ทั้ง 3 วิธี แตกต่างกัน

2. เมื่อความยาวของแบบทดสอบต่างกัน จำนวนข้อสอบที่ทำหน้าที่ต่างกันจากการ ตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน ทั้ง 3 วิธี แตกต่างกัน

ในการศึกษครั้งนี้ เป็นการศึกษาจากข้อมูลเชิงประจักษ์ (Empirical data) การจัดทำ กระทำ เพื่อกำหนดให้ข้อสอบแต่ละข้อทำหน้าที่ต่างกัน ไม่สามารถกระทำได้ตั้งแต่เริ่มต้น ด้วยเหตุนี้จึงใช้ การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจากข้อมูลทั้งหมดด้วยวิธีตรวจสอบทั้ง 3 วิธี ข้อสอบ ข้อใดตรวจพบว่าทำหน้าที่ต่างกันซ้ำกันหรือสอดคล้องกันทุกวิธี ถือว่าข้อสอบข้อนั้นเป็นข้อสอบที่ทำ หน้าที่ต่างกัน ดังนั้น ในการศึกษาครั้งนี้จึงใช้คำว่า “อัตราความไม่สอดคล้อง” แทนคำว่า “อัตรา ความคลาดเคลื่อน” สำหรับอัตราความไม่สอดคล้องในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน จาก การศึกษาของ โคเฮน และคิม (Cohen and Kim. 1993) พบว่า มีความไม่คงที่ของผลการศึกษา เมื่อขนาดของกลุ่มตัวอย่าง ความยาวของแบบทดสอบ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน ดังนั้น ผู้วิจัยจึงตั้งสมมติฐาน ดังนี้

3. เมื่อกลุ่มตัวอย่างขนาดเล็ก (300 คน) แบบทดสอบสั้น (20 - 30 ข้อ) และสัดส่วน จำนวนข้อสอบที่ทำหน้าที่ต่างกันแบบทดสอบน้อย (ร้อยละ 0 - 10) อัตราความไม่สอดคล้องใน การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ทั้ง 3 วิธี แตกต่างกัน

4. เมื่อกลุ่มตัวอย่างขนาดเล็ก แบบทดสอบสั้น และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ ต่างกันแบบทดสอบมาก (ร้อยละ 20) อัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่าง กันของข้อสอบ ทั้ง 3 วิธี แตกต่างกัน

5. เมื่อกลุ่มตัวอย่างขนาดเล็ก แบบทดสอบปานกลาง (40 ข้อ) และสัดส่วนจำนวนข้อสอบ ที่ทำหน้าที่ต่างกันแบบทดสอบน้อย อัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกัน ของข้อสอบ ทั้ง 3 วิธี แตกต่างกัน

6. เมื่อกลุ่มตัวอย่างขนาดเล็ก แบบทดสอบปานกลาง และสัดส่วนจำนวนข้อสอบที่ทำ หน้าที่ต่างกันแบบทดสอบมาก อัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของ ข้อสอบ ทั้ง 3 วิธี แตกต่างกัน

19. เมื่อกลุ่มตัวอย่างขนาดค่อนข้างใหญ่ แบบทดสอบยาว และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันแบบทดสอบน้อย อัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ทั้ง 3 วิธี ไม่แตกต่างกัน

20. เมื่อกลุ่มตัวอย่างขนาดค่อนข้างใหญ่ แบบทดสอบยาว และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันแบบทดสอบมาก อัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ทั้ง 3 วิธี ไม่แตกต่างกัน

21. เมื่อกลุ่มตัวอย่างขนาดใหญ่ (1,000 คน) แบบทดสอบสั้น และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันแบบทดสอบน้อย อัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ทั้ง 3 วิธี แตกต่างกัน

22. เมื่อกลุ่มตัวอย่างขนาดใหญ่ แบบทดสอบสั้น และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันแบบทดสอบมาก อัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ทั้ง 3 วิธี แตกต่างกัน

23. เมื่อกลุ่มตัวอย่างขนาดใหญ่ แบบทดสอบปานกลาง และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันแบบทดสอบน้อย อัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ทั้ง 3 วิธี แตกต่างกัน

24. เมื่อกลุ่มตัวอย่างขนาดใหญ่ แบบทดสอบปานกลาง และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันแบบทดสอบมาก อัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ทั้ง 3 วิธี แตกต่างกัน

25. เมื่อกลุ่มตัวอย่างขนาดใหญ่ แบบทดสอบยาว และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันแบบทดสอบน้อย อัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ทั้ง 3 วิธี ไม่แตกต่างกัน

26. เมื่อกลุ่มตัวอย่างขนาดใหญ่ แบบทดสอบยาว และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันแบบทดสอบมาก อัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ทั้ง 3 วิธี ไม่แตกต่างกัน

ขอบเขตของการศึกษา

1. ข้อมูลที่ใช้ในการศึกษาครั้งนี้ เป็นข้อมูลผลการสอบประเมินคุณภาพและความก้าวหน้าทางการศึกษาระดับประเทศ ในปีการศึกษา 2536 วิชาคณิตศาสตร์ชั้นประถมศึกษาปีที่ 4 ของสำนักงานการประถมศึกษาแห่งชาติ

2. ตัวแปรที่ศึกษา

2.1 ตัวแปรอิสระ ประกอบด้วย

2.1.1 วิธีการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน ได้แก่ วิธี Lord's χ^2 วิธี

Raju's Area Measures และวิธี Closed Interval Area

2.1.2 ขนาดของกลุ่มตัวอย่าง ได้แก่ 300, 500, 700 และ 1,000 คน

2.1.3 ความยาวของแบบทดสอบ ได้แก่ 20, 30, 40 และ 60 ข้อ

2.1.4 สัดส่วนของจำนวนข้อสอบที่ทำหน้าที่ต่างกัน ได้แก่ 0 %, 10 % และ 20 %

2.2 ตัวแปรตาม ประกอบด้วย

2.2.1 อัตราความไม่สอดคล้องในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน ในที่นี้พิจารณา 2 ลักษณะ คือ อัตราความไม่สอดคล้องชนิดที่ 1 และอัตราความไม่สอดคล้องชนิดที่ 2

2.2.2 อัตราความสอดคล้อง แบ่งเป็น 2 ลักษณะ ได้แก่ อัตราความสอดคล้องชนิดที่ 1 และอัตราความสอดคล้องชนิดที่ 2

2.2.3 จำนวนข้อสอบที่ทำหน้าที่ต่างกัน ได้จากการตรวจสอบแต่ละวิธี

2.2.4 ความสอดคล้องในการตรวจสอบ

2.2.5 ความสัมพันธ์ระหว่างวิธีตรวจสอบ

3. แบบจำลองที่ใช้ในการศึกษา ได้แก่ แบบจำลองการตอบข้อสอบแบบโลจิสติกที่มี 2 พารามิเตอร์

ข้อตกลงเบื้องต้น

เนื่องจากข้อมูลที่ใช้ในการศึกษาครั้งนี้ เป็นข้อมูลผลการสอบประเมินคุณภาพและความก้าวหน้าทางการศึกษา ระดับประเทศ ของสำนักงานการประถมศึกษาแห่งชาติ จากข้อสอบมาตรฐาน วิชาคณิตศาสตร์ ชั้นประถมศึกษาปีที่ 4 ปีการศึกษา 2536 ผู้วิจัยมุ่งที่จะใช้ข้อมูลดังกล่าวมาวิเคราะห์เพื่อประโยชน์ในการศึกษาและเปรียบเทียบความสอดคล้องระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกัน มิได้มีเจตนาศึกษาคุณภาพของข้อสอบดังกล่าว

ความสำคัญของการศึกษา

การศึกษาครั้งนี้ เป็นการเปรียบเทียบความสอดคล้องของผลการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันตามแนวคิดทฤษฎีการตอบข้อสอบ 3 วิธี ซึ่งผลที่ได้จากการศึกษามีประโยชน์และความสำคัญ ดังนี้

1. ทำให้ได้ข้อสรุปที่ชัดเจนเกี่ยวกับผลของวิธีการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันตามแนวคิดของทฤษฎีการตอบข้อสอบ ทั้ง 3 วิธี ว่าวิธีใดบ้างที่ให้ผลสอดคล้องกันหรือขัดแย้งกัน ในเงื่อนไขใดของการศึกษา ซึ่งจะมีประโยชน์ต่อการเลือกใช้ให้เหมาะสมและเกิดความผิดพลาดน้อยที่สุดในการตรวจสอบ

2. เป็นการพัฒนาทฤษฎีการทดสอบในด้านการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันตามแนวคิดของทฤษฎีการตอบข้อสอบ ซึ่งทำให้เกิดแง่มุมในการศึกษาครั้งต่อไปอีก ทั้งในกรณีศึกษาจากข้อมูลเชิงประจักษ์ (Empirical data) และข้อมูลจำลอง (Simulated data)

นัยามศัพท์เฉพาะ

ข้อสอบที่ทำหน้าที่ต่างกัน หมายถึง ข้อสอบที่ผู้สอบ ณ ระดับความสามารถเดียวกัน แต่อยู่ต่างกลุ่มกัน มีโอกาสของการตอบข้อสอบข้อนั้นได้ถูกต้องไม่เท่ากัน ในการศึกษาครั้งนี้ กลุ่มย่อย (Subgroup) ที่ใช้ คือ กลุ่มเพศชาย และกลุ่มเพศหญิง

วิธีตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน หมายถึง วิธีตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันตามแนวคิดทฤษฎีการตอบข้อสอบ ประกอบด้วย วิธี Lord's χ^2 วิธี Raju's Area Measures และวิธี Closed Interval Area

1. วิธี Lord's χ^2 หมายถึง วิธีการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันตามแนวคิดของ ลอร์ด ที่ใช้การเปรียบเทียบค่าพารามิเตอร์ของข้อสอบระหว่างกลุ่มย่อย 2 กลุ่ม

2. วิธี Raju's Area Measures หมายถึง วิธีการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันตามแนวคิดของราจู โดยการวัดความแตกต่างพื้นที่ของฟังก์ชันการตอบข้อสอบระหว่างกลุ่มย่อย 2 กลุ่ม เมื่อกำหนดช่วงของความสามารถ (Ability) ผู้เข้าสอบเป็นแบบเปิด (Open Interval)

3. วิธี Closed Interval Area หมายถึง วิธีการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันตามแนวคิดของโคเชน และคิม โดยการวัดความแตกต่างพื้นที่ของฟังก์ชันการตอบข้อสอบระหว่างกลุ่มย่อย 2 กลุ่ม เมื่อกำหนดช่วงของความสามารถ (Ability) ผู้เข้าสอบ เป็นแบบปิด (Closed Interval)

สำหรับวิธี Raju's Area Measures และ วิธี Closed Interval Area จะพิจารณาทั้งพื้นที่แบบมีเครื่องหมาย (Signed Area : SA) และพื้นที่แบบไม่มีเครื่องหมาย (Unsigned Area : UA)

ดัชนีการทำหน้าที่ต่างกันของข้อสอบ หมายถึง ค่าสถิติที่ใช้แสดงความแตกต่างของผลการตอบข้อสอบระหว่างกลุ่มย่อย ขึ้นอยู่กับว่าใช้วิธีใดในการตรวจสอบ ในที่นี้ ดัชนีการทำหน้าที่ต่างกันของข้อสอบ ได้แก่ ขนาดของพื้นที่ความแตกต่างของฟังก์ชันการตอบระหว่างกลุ่มย่อย 2 กลุ่ม สำหรับวิธี วิธี Raju's Area Measures และวิธี Closed Interval Area และค่าสถิติ ไคสแคว์สำหรับวิธี Lord's χ^2

เกณฑ์การตัดสินข้อสอบที่ทำหน้าที่ต่างกัน หมายถึง ค่าของดัชนีพื้นที่แบบมีเครื่องหมาย (Signed Area : SA) และพื้นที่แบบไม่มีเครื่องหมาย (Unsigned Area : UA) รวมทั้งค่าไคสแคว์ที่แสดงว่าข้อสอบทำหน้าที่ต่างกัน ได้แก่

- วิธี Lord's χ^2 ทดสอบนัยสำคัญทางสถิติที่ระดับ .05
- วิธี Raju's Area Measures ใช้ Z ทดสอบนัยสำคัญทางสถิติที่ระดับนัยสำคัญ .05
- วิธี Closed Interval Area ใช้เกณฑ์ ดังนี้
 - Unsigned Area ขนาดพื้นที่ตั้งแต่ .40 ขึ้นไป
 - Signed Area ขนาดพื้นที่ตั้งแต่ .30 ขึ้นไป

ผลการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน หมายถึง จำนวนข้อสอบที่ทำหน้าที่ต่างกัน ความสอดคล้องในการตรวจสอบ ความสัมพันธ์ระหว่างวิธีตรวจสอบ อัตราความไม่สอดคล้องในการตรวจสอบ และอัตราความสอดคล้องในการตรวจสอบ

อัตราความไม่สอดคล้องชนิดที่ 1 หมายถึง ความไม่สอดคล้องที่เกิดจากการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันผิดพลาด อันเนื่องจากการตัดสินใจว่าข้อสอบทำหน้าที่ต่างกัน ทั้ง ๆ ที่ความจริงข้อสอบข้อนั้นไม่ได้ทำหน้าที่ต่างกัน คำนวณได้ ดังนี้

$$\text{อัตราความไม่สอดคล้องชนิดที่ 1} = \frac{\text{จำนวนข้อสอบที่ยอมรับผิดพลาด}}{\text{จำนวนข้อสอบที่ทำหน้าที่ไม่ต่างกัน (Non DIF) ทั้งหมด}}$$

อัตราความไม่สอดคล้องชนิดที่ 2 หมายถึง ความไม่สอดคล้องที่เกิดจากการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันผิดพลาด อันเนื่องจากการตัดสินใจว่าข้อสอบไม่ทำหน้าที่ต่างกัน ทั้ง ๆ ที่ความจริงข้อสอบข้อนั้นทำหน้าที่ต่างกัน คำนวณได้ดังนี้

$$\text{อัตราความไม่สอดคล้องชนิดที่ 2} = \frac{\text{จำนวนข้อสอบที่ปฏิเสธผิดพลาด}}{\text{จำนวนข้อสอบที่ทำหน้าที่ต่างกัน (DIF) ทั้งหมด}}$$

อัตราความสอดคล้องชนิดที่ 1 หมายถึง ความสอดคล้องของผลการตรวจสอบที่ระบุได้ถูกต้องว่าข้อสอบข้อนั้น ๆ ไม่ได้ทำหน้าที่ต่างกัน คำนวณได้ดังนี้

$$\text{อัตราความสอดคล้องชนิดที่ 1} = \frac{\text{จำนวนข้อสอบที่ตรวจสอบได้ถูกต้องว่าไม่ทำหน้าที่ต่างกัน}}{\text{จำนวนข้อสอบที่ทำหน้าที่ไม่ต่างกัน (Non DIF) ทั้งหมด}}$$

อัตราความสอดคล้องชนิดที่ 2 หมายถึง ความสอดคล้องของผลการตรวจสอบที่ระบุได้ถูกต้องว่าข้อสอบข้อนั้น ๆ ทำหน้าที่ต่างกันจริง คำนวณได้ดังนี้

$$\text{อัตราความสอดคล้องชนิดที่ 2} = \frac{\text{จำนวนข้อสอบที่ตรวจสอบได้ถูกต้องว่าทำหน้าที่ต่างกัน}}{\text{จำนวนข้อสอบที่ทำหน้าที่ต่างกัน (DIF) ทั้งหมด}}$$

บทที่ 2

เอกสารและงานวิจัยที่เกี่ยวข้อง

ในการศึกษาครั้งนี้ ได้จัดเรียงเนื้อหาของเอกสารและงานวิจัยที่เกี่ยวข้อง ดังนี้

1. การทำหน้าที่ต่างกันของข้อสอบ
2. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ
3. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามแนวคิดของทฤษฎีการตอบ

ข้อสอบ

4. งานวิจัยที่เกี่ยวข้อง

1. การทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning)

1.1 ความเป็นมา

การทำหน้าที่ต่างกันของข้อสอบ เดิมใช้ในความหมายของความลำเอียงของข้อสอบ ซึ่งเริ่มต้นได้รับความสนใจตั้งแต่ปี 1910 (Holland and Wainer. 1993) โดยบินด์ และไซมอน (Binet and Simon) ได้ทดสอบเด็กที่มาจากฐานะทางเศรษฐกิจและสังคมต่างกัน พบว่า ข้อสอบบางข้อน่าจะวัดผลจากวัฒนธรรมการฝึก (เช่น ที่บ้าน หรือ ที่โรงเรียน) มากกว่าที่จะวัดความสามารถทางสมอง จากนั้นได้เอาข้อสอบเหล่านั้นออกไป เช่น ถ้าเป็นเด็กอายุ 7 ปี ข้อคำถามที่เกี่ยวข้องกับนิ้วมือ การคัดลอกประโยค และการเรียกชื่อเหรียญถูกเอาออกจากแบบทดสอบ โดยทั่วไปในการทดสอบจะพิจารณาเฉพาะข้อคำถามที่ไม่ขึ้นกับการฝึกฝนของทางบ้าน (Home Training) ความตั้งใจ (Attention) ภาษา (Language) นิสัยในการดูภาพ (Habit of looking at picture) และแบบฝึกหัดเชิงวิชาการ (Scholastic exercise)

ปี 1912 วิลเลียม สเตอร์น (William Stern) ได้ศึกษาความแตกต่างของชั้นเรียน (Class difference) ในประเทศเยอรมัน และพบว่ามีความคล้ายคลึงกับที่บินด์พบ โดยพยายามที่จะทำความเข้าใจเกี่ยวกับแหล่งที่ทำให้เกิดความแตกต่าง เพื่อที่จะเสนอแนะว่า แบบทดสอบนั้นเข้าข้างนักเรียนชั้นหนึ่งมากกว่าชั้นอื่น ๆ อย่างชัดเจน

ต่อมาปี 1951 มีงานวิจัยที่ถือเป็นการเริ่มต้นของการศึกษาเกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบ “ ข้อสอบลำเอียง ” ได้แก่ งานวิจัยของเอลส์ (Ellis) เดวิส (Davis) ฮาวิเกอร์สต์ (Havighurst) เฮร์ริค (Herrick) และไทเลอร์ (Tyler) โดยเอลส์ และคณะ สนใจความแตกต่างของการวัดบางอย่างที่พิจารณาถึงความไม่ถูกต้องของการตอบข้อสอบอันเกิดจากความสามารถ (Ability) เช่น ความแตกต่างของไอคิวเด็กนักเรียน จุดมุ่งหมายหลักเพื่อที่จะอธิบายความแตกต่างระหว่างกลุ่มในแบบทดสอบไอคิว โดยมุ่งไปที่ข้อคำถามแต่ละข้อในลักษณะเช่นเดียวกับ บินด์ และสเตอร์น

คือดูว่าข้อคำถามใดมีการทำหน้าที่ต่างกันหรือไม่ แม้ว่า เอสส์ และคณะใช้สัดส่วนการตอบถูกสำหรับกลุ่มที่แตกต่างกัน แต่ก็ไม่ได้รับการแนะนำให้ใช้ในเวลาไม่นานนัก ขณะเดียวกันก็เกิดคำถามเกี่ยวกับความคงที่ของความแตกต่างระหว่างกลุ่มว่าขึ้นกับเนื้อหา หรือรูปแบบคำถาม ความแตกต่างของวัฒนธรรมในการเลือกคำตอบที่ผิด มีความสัมพันธ์กับความยากของข้อสอบหรือไม่ เขาได้พยายามแก้ไขอิทธิพลของชั้นทางสังคมและภูมิหลังของเชื้อชาติที่มีผลต่อการตอบข้อสอบ โดยการเปรียบเทียบกลุ่มของเชื้อชาติที่อยู่ในชั้นทางสังคม

ผลจากการศึกษาของ เอสส์ ได้นำไปสู่การศึกษาอย่างกว้างขวางในช่วง 20 ปีที่ผ่านมา เมื่อความมุ่งหวังของรัฐในการที่จะทำให้เกิดความยุติธรรมในการใช้ผลการสอบ ในปี 1960 และ 1970 ความตื่นตัวเกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบ ได้มุ่งไปที่แบบทดสอบไอคิว เนื่องจากแบบทดสอบไอคิวได้ถูกนำไปใช้ในการคัดเลือก (Selection) และ จัดบุคคลเข้าสู่ตำแหน่ง (Placement) อย่างแพร่หลาย จนกระทั่งปี 1971 บริษัท Griggs v. Duke Power ซึ่งว่าผลการทดสอบไอคิวใช้ในการคัดเลือกอาจไม่เหมาะสม ยกเว้นกรณีที่ผลของการปฏิบัติงานมีความสอดคล้องกับการทดสอบ

ในช่วงเวลาเดียวกันมีการโต้แย้งเกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบมีทั้งบทความในวารสารต่าง ๆ เจนเซน ได้เสนอว่า ไอคิวเป็นผลมาจากการถ่ายทอดทางพันธุกรรมมากกว่ามาจากสิ่งแวดล้อม ความแตกต่างของไอคิวระหว่างคนผิวขาวกับผิวดำไม่ได้มาจากสิ่งแวดล้อมเพียงอย่างเดียว และได้รับการยืนยันจากนักวิทยาศาสตร์ค่อนข้างชัดเจน เขายังได้อธิบายต่อว่า พันธุกรรมอธิบายความแปรปรวนผลการปฏิบัติงานของแต่ละคนได้ประมาณ ร้อยละ 80 จากนั้นก็มีการตอบสนองงานของเจนเซน ในวารสาร Harvard Education Review (1969) โดยสมาคมนักจิตวิทยาคนผิวดำ เสนอว่า การทดสอบคนผิวดำทุกชนิดควรทำให้เท่าเทียมกันมากเท่าที่จะเป็นไปได้ (Camilli and Shepard. 1994 : 4 - 7)

การตรวจสอบและค้นหาการทำหน้าที่ต่างกันของข้อสอบได้มีมาอย่างต่อเนื่องควบคู่กับการพยายามคิดค้นวิธีการทางสถิติในการวิเคราะห์ การทำหน้าที่ต่างกันของข้อสอบดังปรากฏจากเอกสารการวัดผลในช่วงกลางปี 1970 และในปัจจุบันก็มีการพัฒนาเทคนิควิธีการต่าง ๆ อย่างมากมายเพื่อให้ผลการวิเคราะห์มีความถูกต้องที่สุด

1.2 ความหมายของการทำหน้าที่ต่างกันของข้อสอบ

สำหรับความหมายของการทำหน้าที่ต่างกันของข้อสอบมีผู้ให้ความเห็นไว้ต่าง ๆ ดังนี้

อลเลน และเยน (Allen and Yen. 1979 : 129) ให้ความหมายของการทำหน้าที่ต่างกันของข้อสอบว่า เป็นข้อสอบที่วัดความสามารถหรือคุณลักษณะทางจิตวิทยาของผู้สอบแต่ละกลุ่มไม่ตรงกัน

ลอร์ด (Lord. 1980 : 212) แมคคอลเลย์ และเมนโดซา (McCauley and Mendoza. 1985 : 389) แฮมเบิลตัน และคณะ (Hambleton et. al. 1991 : 110) ฮอลแลนด์และไวเนอร์ (Holland and Wainer. 1993 : 4 ; citing Angoff. 1993) ได้ให้ความหมายข้อสอบที่ทำหน้าที่ต่างกันว่าเป็นข้อสอบที่ผู้สอบมีความสามารถเท่ากัน แต่อยู่ต่างกลุ่มกัน มีโอกาสตอบข้อสอบข้อนั้นได้ถูกต้องไม่เท่ากัน

โปพแฟม (Popham. 1990 : 178) ได้ให้ความหมายของความลำเอียงว่าความชอบหรือความโน้มเอียงที่ทำให้เกิดการตัดสินอย่างไม่ยุติธรรม

มิลล์แซป และอีเวอร์สัน (Millsap and Everson. 1993 : 297) ได้ให้ความหมายของความลำเอียงว่าเป็นความไม่แม่นยำ (Inaccuracy) อย่างเป็นระบบของการวัด

คามิลลี และเชพเพิร์ด (Camilli and Shepard. 1994 : 8) ได้ให้ความหมายของแบบทดสอบที่ทำหน้าที่ต่างกันว่าเป็นความคลาดเคลื่อนอย่างเป็นระบบ (Systematic Error) หรือความไม่เที่ยงตรง (Invalidity) ของการวัดชนิดหนึ่ง ทำให้สมาชิกบางกลุ่มโดยเฉพาะที่ได้ประโยชน์หรือเสียประโยชน์ เช่น แบบทดสอบหรือข้อสอบ ที่เข้าข้างกลุ่มนักเรียนชายมากกว่านักเรียนหญิง หรือคนผิวขาวกับคนผิวดำ หรือระหว่างคนที่อยู่ต่างภูมิภาคกัน เป็นต้น

ฮอลแลนด์ และวายเนอร์ (Holland and Wainer. 1993) ได้ให้ความหมายของการทำหน้าที่ต่างกันของข้อสอบว่า ข้อสอบที่ผู้เข้าสอบ ณ ความสามารถเดียวกัน แต่อยู่ต่างกลุ่มกัน มีโอกาสในการตอบถูกไม่เท่ากัน กลุ่มที่ให้ความสนใจ หรือกลุ่มที่เสียเปรียบ เรียกว่า “ Focal group ” และกลุ่มที่ใช้เปรียบเทียบกับ ซึ่งโดยทั่วไปจะเป็นกลุ่มได้เปรียบ เรียกว่า “ Reference group ”

โดยสรุปข้อสอบหรือแบบทดสอบที่ทำหน้าที่ต่างกัน หมายถึง ข้อสอบ หรือแบบทดสอบที่ทำให้ผู้เข้าสอบที่มีความสามารถเท่ากัน แต่อยู่ต่างกลุ่มภายใต้ประชากรเดียวกันมีโอกาสในการที่จะตอบถูกไม่เท่ากัน หรืออาจกล่าวได้ว่า เป็นข้อสอบที่ขาดความเที่ยงตรงในคุณลักษณะที่ต้องการวัดโดยเอื้อประโยชน์ต่อผู้เข้าสอบบางกลุ่ม ซึ่งอาจใช้ตัวแปรจำแนกในการแบ่งกลุ่มเป็น เพศ เชื้อชาติ ศาสนา ภูมิภาค และระดับสติปัญญา

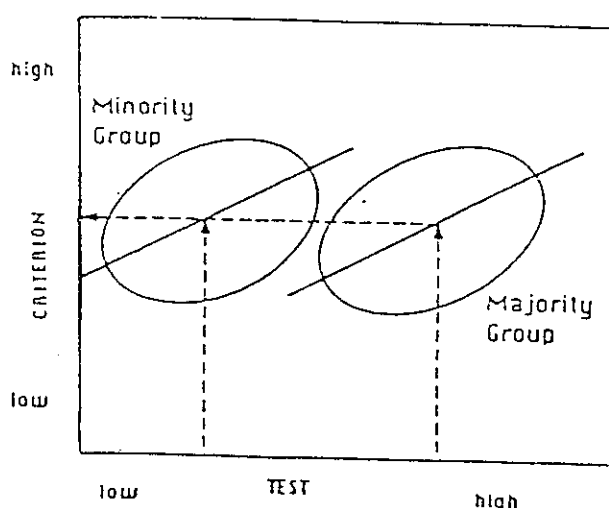
2. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

แบบทดสอบที่ดีควรมีความยุติธรรม (Fairness) กับผู้เข้าสอบทุกกลุ่ม เนื่องจากถ้าข้อสอบหรือแบบทดสอบทำหน้าที่ เบี่ยงเบนไป โดยเข้าข้างกลุ่มใดกลุ่มหนึ่ง ทำให้เกิดการได้เปรียบเสียเปรียบกันขึ้นในการทดสอบครั้งหนึ่ง ๆ ผลการสอบก็ขาดความน่าเชื่อถือ หรือเกิดความไม่เสมอภาคในการตัดสิน เช่น แบบทดสอบที่ใช้คัดเลือกบุคคล แบบทดสอบประเมินคุณภาพการศึกษา และแบบทดสอบเลื่อน ตำแหน่ง เป็นต้น

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสามารถทำได้ 2 ขั้นตอน ได้แก่ ขั้นตอนแรกเป็นขั้นของการสร้างแบบทดสอบ โดยให้ผู้เชี่ยวชาญตรวจสอบว่า ข้อสอบ หรือ แบบทดสอบทำหน้าที่ต่างกันหรือไม่ วัดในคุณลักษณะที่ต้องการวัดเดียวกันหรือไม่ นั่นคือ เป็นการตรวจสอบความเที่ยงตรงของแบบทดสอบ (Validity) ส่วนขั้นตอนต่อมา ใช้เทคนิคทางสถิติเข้ามาช่วย หลังจากนำแบบทดสอบไปใช้แล้ว นำผลมาวิเคราะห์ที่ดูลักษณะการตอบข้อสอบของกลุ่มคนที่อยู่ต่างกลุ่มกัน เช่น อาจเป็น เพศ เชื้อชาติ ศาสนา ท้องถิ่น และระดับสติปัญญา ว่าข้อสอบหรือแบบทดสอบยังคงทำหน้าที่ไม่ต่างกัน หรือเอื้อประโยชน์ต่อคนบางกลุ่มหรือไม่ ซึ่งการศึกษาการทำหน้าที่ต่างกันของข้อสอบที่มีพื้นฐานจากการใช้หลักสถิติสามารถทำได้ 2 แนวทาง (Camilli and Shepard. 1994 : 9) ได้แก่

1) การใช้วิธีภายนอก (External Method)

การศึกษาการทำหน้าที่ต่างกันของข้อสอบ โดยใช้วิธีภายนอกหรือการใช้เกณฑ์ภายนอก (External Criterion) ถือเป็นการหาความเที่ยงตรงเชิงพยากรณ์ (Predictive Validity) นั้นเอง ทำได้โดยการเขียนกราฟแสดงความสัมพันธ์ระหว่างตัวทำนาย (ข้อสอบ) กับเกณฑ์ (เกรดเฉลี่ย หรือ ผลการปฏิบัติงาน เป็นต้น) พิจารณาค่าความลาดชันของกราฟ (Slope) และค่าจุดตัดแกนตั้ง (Intercept) ระหว่าง 2 กลุ่ม ถ้าต่างกันแสดงว่าข้อสอบทำหน้าที่ต่างกันเข้าข้างกลุ่มใดกลุ่มหนึ่ง (ดังภาพประกอบ 2)



ภาพประกอบ 2 การทำหน้าที่ต่างกันของข้อสอบ เมื่อใช้วิธีภายนอกตรวจสอบ
(Camilli and Shepard. 1994 : 11)

สำหรับวิธีนี้จะมีปัญหามากตรงที่เกณฑ์มีความเป็นมาตรฐานหรือไม่ ดังนั้นจึงไม่ได้รับความนิยมในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

2) การใช้วิธีภายใน (Internal Method)

การศึกษาการทำหน้าที่ต่างกันของข้อสอบโดยวิธีภายในหรือการใช้เกณฑ์ภายใน (Internal Criterion) ได้รับความนิยมอย่างมาก เนื่องจากไม่ต้องกังวลเรื่องเกณฑ์สามารถศึกษาจากการตรวจสอบโครงสร้างภายในของข้อสอบ ซึ่งพิจารณาคะแนนที่ได้จากผลการตอบของผู้เข้าสอบต่างกลุ่มว่าวัดในคุณลักษณะที่ต้องการวัดตามโครงสร้างเช่นเดียวกันหรือไม่ ถือเป็นการกระบวนการของการสนับสนุนการตรวจสอบความเที่ยงตรงเชิงโครงสร้าง (Construct Validity) การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแนวนี้ สามารถแบ่งเป็น 3 กลุ่ม ใหญ่ ๆ (Camilli and Shepard. 1994) ดังนี้

2.1 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามแนวคิดของทฤษฎีการทดสอบ

แบบดั้งเดิม (Classical Test Theory)

เป็นวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ใช้กันมาตั้งแต่สมัยเริ่มต้นของการพัฒนาเทคนิควิธีการตรวจสอบ แม้ว่าปัจจุบันจะไม่ค่อยได้นำมาใช้แล้ว เนื่องจากนักวัดผลได้มีการพัฒนาเทคนิควิธีการใหม่ ๆ มาใช้แทน และได้ผล ดีกว่า ประกอบกับวิธีการตรวจสอบแนวนี้ก็มีจุดอ่อนตรงที่ไม่สามารถแยกค่าความยากของข้อสอบออกจากค่าอำนาจจำแนกได้ ทำให้เกิดความคลาดเคลื่อนในการตรวจสอบค่อนข้างสูง สำหรับวิธีการในกลุ่มนี้ประกอบด้วย วิธีแปลงค่าความยาก (Transformation Item Difficulty) วิธี Golden Rule วิธีวิเคราะห์ความแปรปรวน (Analysis of Variance) และ วิธีพอยต์ไบซีเรียล (Point Biserial) เป็นต้น

2.2 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามแนวคิดของทฤษฎีการตอบข้อสอบ

(Item Response Theory)

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแนวนี้ จะพิจารณาเปรียบเทียบฟังก์ชันการตอบข้อสอบ (Item Response Function) หรือ เส้นโค้งคุณลักษณะข้อสอบ (Item Characteristic Curves) ระหว่างกลุ่มตั้งแต่ 2 กลุ่ม ขึ้นไป ว่ามีความแตกต่างกันหรือไม่

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามแนวคิดของทฤษฎีการตอบข้อสอบ มีจุดเด่นกว่าวิธีตามแนวคิดทฤษฎีการทดสอบแบบดั้งเดิม ดังนี้

1) ทฤษฎีการตอบข้อสอบสามารถประมาณค่าพารามิเตอร์ของข้อสอบ เช่น ค่าความยาก (b) ค่าอำนาจจำแนก (a) และค่าการเดา (c) ทำให้ลดการปนกัน (Confounding) ของพารามิเตอร์ข้อสอบลงได้

2) คุณสมบัติทางสถิติของข้อสอบ (Statistical Properties of Items) สามารถอธิบายได้ละเอียดและชัดเจนกว่า เมื่อข้อสอบทำหน้าที่ต่างกันระหว่าง 2 กลุ่ม

3) คุณสมบัติทางสถิติของข้อสอบ สามารถแสดงด้วยแผนภาพที่เป็นกราฟ ทำให้เข้าใจง่ายและนำไปใช้ได้กว้างขวาง สำหรับรายละเอียดจะแยกกล่าวในตอนต่อไป

2.3 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้ตารางการณัจร (Contingency Table Approach)

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแนวนี้ มีลักษณะ ดังนี้

1) ใช้สถิติ непараметрических ในการทดสอบ (Non-Parametric Statistics)

2) ใช้คะแนนที่สังเกตได้ (Observed Score) ประมาณค่าความสามารถระหว่างกลุ่ม

3) มีความสอดคล้องมากกับการเปรียบเทียบระหว่างเส้นกราฟคุณลักษณะของข้อสอบ

(Item Characteristic Curve) ตามแนวทฤษฎีการตอบข้อสอบ

4) ไม่จำเป็นต้องใช้สถิติขั้นสูง

5) ค่อนข้างประหยัด

รวมทั้งมีมีนัยสำคัญกับการตรวจสอบตามแนวคิดทฤษฎีการตอบข้อสอบ ได้แก่

1) มีข้อกำหนดเบื้องต้น ในการเปรียบเทียบความแตกต่างของผลการตอบข้อสอบระหว่างผู้เข้าสอบ 2 กลุ่ม โดยใช้เกณฑ์ภายใน เช่น ใช้คะแนนจากการสอบเป็น 0 - 1 เหมือนกัน

2) ทั้ง 2 วิธีสามารถนำไปประยุกต์ในการตรวจสอบซ้ำได้ (Iterative) เช่น ตามแนวคิด Purification ของ ลอร์ด (Lord. 1980 : 220) หรือแนวคิดของ Candell and Drasgow (1988) เป็นต้น

3) เป็นแนวคิดการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning) ในลักษณะการวัดแบบหลายมิติ (Multidimensionality) ชนิดหนึ่ง เช่นเดียวกับทฤษฎีการตอบข้อสอบ

ตัวอย่างวิธีการตรวจสอบในกลุ่มนี้ ได้แก่ ไคสแควร์ (Chi-square) แมนเทิล แฮนเซล (Mantel-Haenzel) และ ลอกลินีเย (log-linear) เป็นต้น

อย่างไรก็ตามปัจจุบันนักวัดผลได้พัฒนาเทคนิควิธีใหม่ ๆ ขึ้นมาอย่างต่อเนื่อง เช่น วิธีการทำให้เป็นมาตรฐาน (Standardization Method) แบบจำลองการถดถอยโลจิสติก (Logistic Regression Models) และล่าสุดวิธี SIBTEST (Shealy and Stout. 1993)

3. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามแนวคิดของทฤษฎีการตอบข้อสอบ

3.1 ความเป็นมาของทฤษฎีการตอบข้อสอบ

ทฤษฎีการตอบข้อสอบ (Item Response Theory) หรือ ทฤษฎีคุณลักษณะแฝง (Latent Trait Theory) ได้รับการพัฒนามานานตั้งแต่บิเน็ต และ ซีมอน (Binet and Simon. 1916) และริชาร์ดสัน (1936) ได้แสดงความสัมพันธ์ระหว่างพารามิเตอร์ของแบบจำลองทฤษฎีการตอบข้อสอบและทฤษฎีแบบดั้งเดิม ต่อมาปี 1943 - 1944 ลอร์เลย์ (Hambleton and Swaminathan. 1985) ได้เสนอวิธีการประมาณค่าพารามิเตอร์แนวใหม่ ซึ่งมีหลักการว่าผลการสอบของผู้สอบจากแบบทดสอบใด ๆ ขึ้นอยู่กับความสามารถ (Ability or Skill) ของผู้สอบ จนกระทั่งปี 1952 ลอร์ด (Lord) ได้เสนอทฤษฎีใหม่ในรูปโค้งลักษณะข้อสอบ (Item Characteristic Curve : ICC) โดยลอร์ดได้เสนอว่าโค้งลักษณะข้อสอบมีลักษณะเป็นโค้งปกติสะสม ต่อมาเรียกว่า แบบจำลองนอร์มัลอจิว (Normal Ogive Model) ซึ่งแบบจำลองนี้จะกล่าวถึงพารามิเตอร์ 2 ตัว คือ ค่าความยาก และค่าอำนาจจำแนก แต่เนื่องจากแบบจำลองนี้มีความยากในการคำนวณ และขาดโปรแกรมคอมพิวเตอร์ที่มีประสิทธิภาพในการใช้วิเคราะห์ข้อมูลเพื่อพิสูจน์ทฤษฎี จึงทำให้ ลอร์ด ชะงักการทำงานเกี่ยวกับทฤษฎีนี้ไประยะหนึ่ง

ปี 1960 ราสช์ (Rasch) ได้ศึกษาเกี่ยวกับเรื่องนี้โดยไม่ทราบแนวคิดของ ลอร์ด และได้เสนอแนวคิดในรูปของพารามิเตอร์ตัวเดียว คือ ค่าความยากของข้อสอบเท่านั้น

ปี 1965 ลอร์ด ได้หันมาสนใจและพัฒนาทฤษฎีการตอบข้อสอบใหม่อีกครั้ง

ปี 1968 เบิร์นบอม (Birnbaum) ได้เสนอแบบจำลองโลจิสติก (Logistic Model) ที่ใช้พารามิเตอร์ 2 ตัว คือ ค่าความยาก และค่าอำนาจจำแนก ซึ่งเป็นแบบจำลองที่ง่ายกว่าของ ลอร์ด จึงทำให้แบบจำลองนี้เป็นที่นิยมแพร่หลาย และมีการพัฒนาขึ้นเรื่อย ๆ จนกระทั่งสามารถใช้ได้กับพารามิเตอร์ตัวเดียว และพารามิเตอร์สามตัว

3.2 หลักการของทฤษฎีการตอบข้อสอบ

ทฤษฎีการตอบข้อสอบ เป็นทฤษฎีที่อธิบายถึงความสัมพันธ์ระหว่างคุณลักษณะหรือความสามารถที่มีอยู่ในตัวบุคคลกับพฤติกรรมการตอบข้อสอบของบุคคลนั้น ๆ ในรูปของฟังก์ชันทางคณิตศาสตร์ กล่าวคือ พฤติกรรมการตอบข้อสอบของบุคคลจะถูกกำหนดโดยคุณลักษณะหรือความสามารถที่มีอยู่ในตัวบุคคล (Lord and Novick. 1968 : 358) ซึ่งไม่สามารถจะสังเกตได้โดยตรง เขียนในรูปฟังก์ชันทางคณิตศาสตร์ได้ ดังนี้

$$y = f(\theta)$$

เมื่อ y แทน คะแนนที่ได้จากการตอบข้อสอบ ซึ่งเป็นพฤติกรรม
การตอบข้อสอบ

θ แทน คุณลักษณะหรือความสามารถภายในตัวบุคคล

f แทน ฟังก์ชัน

ซึ่ง ลอร์ด เรียกฟังก์ชันนี้ว่า ฟังก์ชันการตอบข้อสอบ (Item Response Function) หรือ ฟังก์ชันคุณลักษณะของข้อสอบ (Item Characteristic Function) (Hambleton and Swaminathan. 1985 : 13)

3.3 ข้อตกลงเบื้องต้นของทฤษฎีการตอบข้อสอบ

1) ความเป็นมิติเดียว (Unidimensionality) ถ้าความสามารถของคน (Ability or Trait) มีอยู่หลายอย่าง ซึ่งความสามารถแต่ละอย่างนี้ ต่างก็ส่งผลต่อการตอบข้อสอบ (Test Performance) ต่าง ๆ ที่รวมกันเป็นแบบทดสอบ ถ้าผลการตอบข้อสอบ หรือคะแนนของผู้สอบนั้น สามารถอธิบายได้เพียงความสามารถเดียวก็เพียงพอแล้ว เรียกข้อตกลงนี้ว่ามีความเป็นมิติเดียว (Hambleton and Swaminathan. 1985 : 16) สำหรับในทางปฏิบัติ ก็คือ ข้อสอบแต่ละข้อในแบบทดสอบจะต้องวัดในคุณลักษณะ (Trait) เดียว แต่อย่างไรก็ตาม แฮมเบิลตัน และสวามินาธาน (Hambleton and Swaminathan. 1985 : 16 - 17) ได้กล่าวว่าข้อตกลงเบื้องต้นข้อนี้ไม่ได้เข้มงวดนักขอให้มีความเด่น (Dominant) ที่จะวัดในองค์ประกอบใดองค์ประกอบหนึ่งก็ใช้ได้ ซึ่งสอดคล้องกับผลการวิจัยของ ดรอสโกว์ และพาร์สันส์ และเรคเคส (McCauley and Mendoza. 1985 : 390 ; citing Drasgow and Parsons. 1983 ; Reckase. 1979) ที่พบว่า การประมาณค่าพารามิเตอร์ต่าง ๆ ยังสามารถทำได้เมื่อการวัดนั้นอาจจะมีการเบนไปในทางหลายมิติบ้างเล็กน้อย

วิธีการตรวจสอบว่าแบบทดสอบนั้นวัดในมิติเดียวหรือไม่นั้น ทำได้หลายวิธี เช่น โดยการวิเคราะห์องค์ประกอบ (Factor Analysis) แล้วสังเกตค่าไอเกน (Eigen Value) ค่าสูงสุดว่าแตกต่างจากค่าอื่น ๆ อย่างชัดเจนหรือไม่ สำหรับ ลอร์ด และโนวิก (Lord and Novick. 1968 : 382 ; Lord. 1980 : 20) ยังได้แนะนำว่าการวิเคราะห์องค์ประกอบควรจะใช้ค่าสหสัมพันธ์แบบเตตราคอร์ริก

ไม่ควรใช้การคำนวณค่าสหสัมพันธ์แบบเพียร์สัน เพราะว่าข้อมูลที่เหมาะ (Fit) กับแบบจำลองของ ทฤษฎีการตอบข้อสอบไม่ได้สัมพันธ์กันในเชิงเส้นตรง (Hambleton and Swaminathan. 1985 : 306)

(2) **ความเป็นอิสระในการตอบข้อสอบ (Local Independence)** หมายถึง โอกาสในการตอบข้อสอบแต่ละข้อได้ถูกต้องเป็นอิสระจากกัน นั่นคือ การตอบข้อสอบข้อใดข้อหนึ่งได้ถูกหรือผิดจะไม่มีผลต่อการตอบข้ออื่นๆ ด้วย (Lord. 1980 : 19 ; Hambleton and Swaminathan. 1985 : 23) หรือจะกล่าวในเชิงคณิตศาสตร์ได้ว่า ความเป็นอิสระในการตอบข้อสอบ หมายถึง ความน่าจะเป็นที่จะตอบข้อสอบถูกทั้งหมด จะเท่ากับผลคูณของความน่าจะเป็นของการตอบข้อสอบถูกเป็นรายข้อ (Lord. 1980 : 19 ; Hambleton and Swaminathan. 1985 : 23) แมกโดนัลด์ (McDonald. 1982 : 379 - 396) และแฮมเบิลตันกับสวามินาธาน (Hambleton and Swaminathan. 1985 : 22) มีความเห็นตรงกันว่าถ้ามีความเป็นมิติเดียวแล้วจะมีความเป็นอิสระในการตอบข้อสอบ ซึ่งมีความสำคัญต่อการประมาณค่าพารามิเตอร์ต่าง ๆ มาก ถ้าขาดความเป็นอิสระในการตอบข้อสอบแล้ว ค่าพารามิเตอร์ต่าง ๆ ที่ประมาณค่าได้ จะมีความคลาดเคลื่อนจากความเป็นจริง

3) **โค้งแสดงลักษณะข้อสอบ (Item Characteristic Curve : ICC)** โค้งแสดงลักษณะข้อสอบ เป็นฟังก์ชันทางคณิตศาสตร์ที่แสดงถึงความสัมพันธ์ระหว่างความน่าจะเป็นในการตอบข้อสอบข้อนั้นได้ถูกต้องกับระดับความสามารถที่วัดได้โดยใช้ชุดของข้อสอบ หรือแบบทดสอบนั้น (Hambleton and Swaminathan. 1985 : 25) ซึ่งจะเห็นว่าความน่าจะเป็นที่ผู้สอบตอบข้อสอบถูกจะขึ้นอยู่กับโค้งแสดงลักษณะข้อสอบในแต่ละแบบจำลองที่ใช้ โดยที่รูปร่างของโค้งแสดงลักษณะข้อสอบในแต่ละข้อมีคุณสมบัติไม่แปรเปลี่ยน (Invariant) ไปตามกลุ่มตัวอย่างหรือตัวผู้สอบ (Hambleton and Cook. 1977 : 80) ฉะนั้นจึงทำให้ความน่าจะเป็นในการตอบข้อสอบถูกในแต่ละข้อไม่แปรเปลี่ยนด้วย

เนื่องจากมีความเชื่อหลายความเชื่อเกี่ยวกับความสัมพันธ์ระหว่างความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องกับระดับความสามารถที่วัดได้โดยใช้ชุดของข้อสอบหรือแบบทดสอบนั้น จึงทำให้เกิดมีแบบจำลองของโค้งแสดงลักษณะข้อสอบ (Item Response Model) ขึ้นมาหลายแบบจำลอง ดังจะกล่าวต่อไปนี้

3.4 แบบจำลองในทฤษฎีการตอบข้อสอบ (Item Response Model)

แบบจำลองในทฤษฎีการตอบข้อสอบมีการพัฒนาต่อเนื่องกันมานานและมีหลายลักษณะ ประกอบด้วยแบบจำลองกิตต์แมนสมบูรณ์ (Guttman Perfect Scale) แบบจำลองเชิงเส้น (Linear Model) แบบจำลองนอร์มินัล (Nominal Model) แบบจำลองนอร์มัลออร์จิฟ (Normal Ogive Model) แบบจำลองเกรดเรสปอนส์ (Grad Response Model) และแบบจำลองโลจิสติก (Logistic Model) แต่ในที่นี้จะกล่าวถึงบางลักษณะที่น่าสนใจ และใช้กันแพร่หลาย ดังนี้

3.4.1 แบบจำลองนอร์มัลออร์จิฟ

ลอว์เลย์ (Lord and Novick. 1968 : 369 ; citing Lawley. 1943, 1944) เป็นผู้พัฒนาขึ้น และในปี 1952 ลอร์ด ได้พัฒนาวิธีการประมาณค่าพารามิเตอร์ และนำมาประยุกต์ใช้

โดยใช้พารามิเตอร์ 2 ตัว ได้แก่ ค่าความยาก และค่าอำนาจจำแนก เขียนเป็นรูปแบบจำลองทางคณิตศาสตร์ ดังนี้

$$P_i(\theta) = \int_{-a}^{a_j(\theta-b_i)} \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} dz$$

เมื่อ $P_i(\theta)$ แทน ความน่าจะเป็นที่ผู้มีความสามารถ (θ) ตอบข้อสอบข้อที่ i ได้ถูกต้อง

a_i แทน ค่าอำนาจจำแนกของข้อสอบข้อที่ i

b_i แทน ค่าความยากของข้อสอบข้อที่ i

θ แทน ระดับความสามารถของผู้สอบ

Z แทน ความเบี่ยงเบนปกติ จากการแจกแจงที่ค่าเฉลี่ย b
และความเบี่ยงเบนมาตรฐาน $1/a$

แบบจำลองนอร์มัลลอจีฟ (Normal Ogive Model) เป็นแบบจำลองที่ ลอร์ด ได้ใช้เวลาอันยาวนานในการพัฒนาแบบจำลองนี้ ในระยะแรก ๆ เป็นไปอย่างเชื่องช้ามาก เนื่องจากความสลับซับซ้อนทางคณิตศาสตร์ที่เป็นพื้นฐานของทฤษฎียากแก่การปฏิบัติ หรือนำมาใช้จริง คนทั่วไปจึงไม่นิยมนำแบบจำลองนี้มาใช้ แม้แต่ ลอร์ด เอง ในระยะหลังนี้ก็ได้หันมาใช้แบบจำลองโลจิสติก (Logistic Model) แทน เพราะเป็นแบบจำลองที่มีลักษณะคล้ายกันมากแต่มีวิธีการคำนวณค่าพารามิเตอร์ต่างๆ ได้ง่ายกว่า (Lord and Novick. 1968 : 366) แต่อย่างไรก็ตาม แบบจำลองนี้ก็นับว่ามีประโยชน์อย่างมาก สำหรับการพัฒนาแบบจำลองโลจิสติก (Logistic Model) ต่อมา

3.4.2 แบบจำลองโลจิสติก (Logistic Model)

ผู้มีบทบาทสำคัญในการพัฒนาแบบจำลองนี้ คือ เบิร์นบอม (Hambleton and Swaminathan. 1985 : 36 ; citing Birnbaum. 1957, 1958a, 1958b, 1968) โดยในครั้งแรก เบิร์นบอมได้พัฒนาแบบจำลองโลจิสติก แบบ 2 พารามิเตอร์ คือ พารามิเตอร์ค่าความยาก (b) และค่าอำนาจจำแนก (a) ซึ่งเป็นแบบจำลองที่มีโค้งแสดงลักษณะข้อสอบ (Item Characteristic Curve : ICC) คล้ายกับโค้งแสดงลักษณะข้อสอบ (ICC) ของแบบจำลองนอร์มัลลอจีฟ (Normal Ogive Model) แบบ 2 พารามิเตอร์ ที่ ลอร์ด ได้เสนอไว้เมื่อปี ค.ศ. 1952 มาก โดยโค้งแสดงลักษณะทั้งสองมีความแตกต่างกันไม่เกิน .01 สำหรับทุก ๆ ค่าของความสามารถ (Lord. 1980 : 14 ; Hambleton and Swaminathan. 1985 : 37 ; citing Haley. 1952)

เนื่องจากแบบจำลองโลจิสติก ได้ใช้แบบจำลองทางคณิตศาสตร์ที่สามารถนำไปประยุกต์ใช้ได้ง่ายกว่าแบบนอร์มัลลอจีฟ จึงทำให้มีคนนำแบบจำลองนี้ไปใช้กันมาก ต่อมา เบิร์นบอม ได้เพิ่มพารามิเตอร์ค่าการเดา (c) เข้าไปอีกตัวหนึ่ง จึงกลายเป็นแบบ 3 พารามิเตอร์ นอกจากนั้นยังได้ตัดพารามิเตอร์ค่าอำนาจจำแนก (a) ออกจากแบบ 2 พารามิเตอร์ จึงเหลือเพียงพารามิเตอร์เดียว

คือ ค่าความยาก (b) ซึ่งเรียกว่า 1 พารามิเตอร์ สำหรับแบบจำลองทางคณิตศาสตร์ของแต่ละแบบจำลองย่อย มีดังนี้

(1) แบบจำลองโลจิสติก แบบ 3 พารามิเตอร์ (Three Parameter Logistic Model) มีแบบจำลองทางคณิตศาสตร์ ดังนี้ (Hambleton Swaminathan and Rogers. 1991 : 17)

$$P_i(\theta) = c_i + (1 + c_i) \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$

เมื่อ $P_i(\theta)$ คือ ความน่าจะเป็นที่ผู้ตอบซึ่งมีระดับความสามารถ (θ) ตอบข้อสอบข้อที่ i ได้ถูกต้อง

a_i คือ ค่าอำนาจจำแนกของข้อสอบข้อที่ i โดยทฤษฎีแล้ว a จะมีค่าอยู่ระหว่าง $-\infty$ ถึง $+\infty$ ข้อสอบที่ใช้ได้ควรมีค่า a เป็นบวก โดยปกติจะมีค่าไม่เกิน 2.00

b_i คือ ค่าความยากของข้อสอบข้อที่ i โดยทฤษฎีแล้ว b จะมีค่าอยู่ระหว่าง $-\infty$ ถึง $+\infty$ แต่ในทางปฏิบัติ b จะมีค่าอยู่ระหว่าง -2.00 ถึง +2.00 ข้อที่มีค่า b ใกล้ -2.00 เป็นข้อที่ง่ายมาก และข้อที่มีค่า b ใกล้ +2.00 เป็นข้อที่ยากมาก (Hambleton and Cook. 1977 : 81)

c_i คือ ค่าการเดาของข้อสอบ หมายถึง ค่าความน่าจะเป็นที่ผู้สอบมีความสามารถต่ำสุด ทำข้อสอบนี้ได้ถูกต้อง ข้อสอบที่มีค่าการเดาสูงเป็นข้อสอบที่ไม่ดี ควรมีค่าการเดาไม่เกิน 0.30 (Urry. 1977 : 184)

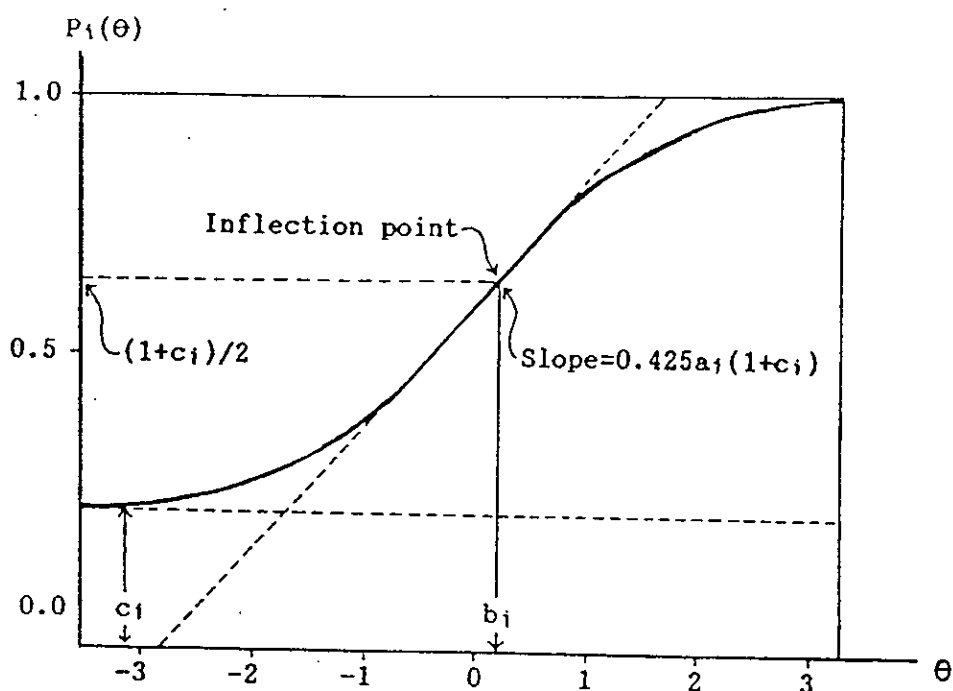
θ คือ ระดับความสามารถของผู้สอบ โดยทฤษฎีแล้วจะมีค่าอยู่ระหว่าง $-\infty$ ถึง $+\infty$ ในทางปฏิบัติจะมีค่าอยู่ระหว่าง -3.00 ถึง +3.00

D คือ ค่าคงที่ (Scalling Factor) เพื่อปรับโค้งแสดงลักษณะข้อสอบ (ICC) ให้มีลักษณะใกล้เคียงกับโค้งแสดงลักษณะข้อสอบของแบบจำลองนอร์มัลลอจีฟ (Normal Ogive Model)

e คือ ค่าเอ็กซ์โปเนนเชียล (Exponential) เป็นค่าคงที่ตัวหนึ่งซึ่งมีค่าเท่ากับ 2.7182818

จากแบบจำลองทางคณิตศาสตร์ของ 3 พารามิเตอร์ เมื่อนำมาเขียนเป็นกราฟแสดงลักษณะข้อสอบได้ ดังนี้ (Lord. 1980 : 12 - 14 ; Hambleton and Swaminathan. 1985 : 38 - 39)

ความน่าจะเป็นในการตอบถูก



ความสามารถ

ภาพประกอบ 3 โค้งแสดงลักษณะข้อสอบแบบ 3 พารามิเตอร์

จากภาพประกอบ 3 เป็นโค้งแสดงลักษณะข้อสอบ แบบ 3 พารามิเตอร์ จะเห็นว่า

- 1) ค่าอำนาจจำแนก (a_i) จะเป็นสัดส่วนกับความชัน (Slope) ของโค้ง ณ จุดเปลี่ยนโค้ง หรือจุดที่มีความชันมากที่สุด [ความชัน ณ จุดนี้มีค่าเท่ากับ $.425a_i(1-c_i)$]
- 2) ค่าความยาก (b_i) จะอยู่บนแกนความสามารถ (θ) ตรงจุดตัดกับเส้นที่ลากมาจากจุดเปลี่ยนโค้งมาตั้งฉากกับแกนความสามารถ
- 3) ค่าการเดา (c_i) เป็นส่วนตัดบนแกนตั้งของโค้งแสดงลักษณะข้อสอบ

(2) แบบจำลองโลจิสติก แบบ 2 พารามิเตอร์ (Two Parameter Logistic Model) มีแบบจำลองทางคณิตศาสตร์ ดังนี้ (Hambleton Swaminathan and Rogers. 1991 : 15)

$$P_i(\theta) = \frac{e^{Da_i(\theta-b_i)}}{1 + e^{Da_i(\theta-b_i)}}$$

(3) แบบจำลองโลจิสติก แบบ 1 พารามิเตอร์ (One- Parameter Logistic Model) เป็นแบบจำลองที่ถือว่าข้อสอบทุกข้อมีค่าอำนาจจำแนกเท่ากันหมด ซึ่งเท่ากับ a และมีค่าการเดา (c) เป็นศูนย์ โดยมีแบบจำลองทางคณิตศาสตร์ ดังนี้ (Hambleton Swaminathan and Rogers. 1991 : 12)

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}$$

3.5 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามแนวคิดทฤษฎีการตอบข้อสอบ

การทำหน้าที่ต่างกันของข้อสอบในแง่ของทฤษฎีการตอบข้อสอบ ก็คือ การที่ฟังก์ชันการตอบข้อสอบข้อนั้น ๆ ของประชากรย่อย 2 กลุ่ม มีความแตกต่างกันทำให้เกิดพื้นที่ระหว่างโค้งแสดงลักษณะข้อสอบ

จากนิยามดังกล่าว การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามแนวคิดทฤษฎีการตอบข้อสอบสามารถกระทำได้โดยเปรียบเทียบฟังก์ชันการตอบข้อสอบของผู้สอบกลุ่มย่อยที่ต้องการศึกษา ซึ่งสามารถแบ่งออกเป็น 2 แนวทางใหญ่ ๆ ดังนี้ (Hambleton Swaminathan and Rogers. 1991 : 110 ; Kim and Cohen. 1995 : 291)

1. การเปรียบเทียบค่าพารามิเตอร์ของข้อสอบระหว่าง 2 กลุ่ม ว่ามีความแตกต่างกันหรือไม่ มีผู้ศึกษาค้นคว้ามากพอสมควร ได้แก่ งานวิจัยของดราบา (Draba. 1977) คิม และโคเฮน (Kim and Cohen. 1994) ลอร์ด (Lord. 1977, 1980) อีสเซน สเติร์นเบิร์ก และวายเนอร์ (Thissen Steinberg and Wainer. 1988, 1993) และไรท์ และสโตน (Wright and Stone. 1979) อย่างไรก็ตาม ในกลุ่มนี้วิธีที่ได้รับความนิยม ก็คือ วิธีที่อาศัยแนวคิดลอร์ด (Lord's χ^2) ในระยะหลังมีวิธีที่น่าสนใจอีกวิธี คือ วิธีทดสอบอัตราส่วนไลค์ลิฮูด (Likelihood Ratio Test) ซึ่ง อีสเซน สเติร์นเบิร์ก และวายเนอร์ คิม และโคเฮน ก็ได้ศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีนี้ รวมทั้งเปรียบเทียบผลการตรวจสอบกับวิธีอื่นด้วย เช่น Lord's χ^2 และ Raju's Area Measures เป็นต้น

สำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามแนวคิดของลอร์ด สามารถใช้ได้กับแบบจำลองโลจิสติก ชนิด 1 พารามิเตอร์ 2 พารามิเตอร์ และ 3 พารามิเตอร์ มีขั้นตอนสรุปได้ ดังนี้

1. ทดสอบข้อมูล (Test Model) กับแบบจำลองที่ศึกษา
2. ประเมินค่าพารามิเตอร์ของข้อสอบ (a , b , c)
3. แปลงหน่วยของค่าพารามิเตอร์ให้อยู่ในหน่วยเดียวกัน หรือหน่วยร่วม (Common Scale)
4. นำค่าพารามิเตอร์ของข้อสอบมาเปรียบเทียบกัน โดยใช้สถิติทดสอบ χ^2

โดยมีสมมติฐานในการทดสอบ คือ $H_0 : b_1 = b_2 , a_1 = a_2 , c_1 = c_2$

สถิติทดสอบ (Hambleton Swaminathan and Rogers. 1991 : 11)

$$\chi^2 = (a_{diff} \ b_{diff} \ c_{diff})' \Sigma^{-1} (a_{diff} \ b_{diff} \ c_{diff})$$

เมื่อ

$$\begin{aligned} a_{diff} &= a_2 - a_1 \\ b_{diff} &= b_2 - b_1 \\ c_{diff} &= c_2 - c_1 \end{aligned}$$

และ Σ แทน เมตริกซ์ Variance และ Covariance ของความแตกต่างระหว่างพารามิเตอร์ข้อสอบ หากศึกษาเป็นแบบจำลอง 1 พารามิเตอร์ สถิติทดสอบ เป็นดังนี้

$$\chi^2 = \frac{b_{diff}^2}{\text{Var}(b_1) + \text{Var}(b_2)}$$

2) การวัดพื้นที่ความแตกต่างระหว่างฟังก์ชันการตอบข้อสอบจาก 2 กลุ่ม เพื่อพิจารณาขนาดของการทำหน้าที่ต่างกันของข้อสอบ มีผู้ศึกษาค้นคว้าไว้ เช่น งานวิจัยของ คิม และโคเฮน (Kim and Cohen. 1991) ลินน์ เลวิน ฮาสติงส์ และวาร์ลดรอป (Linn Levine Hastings and Wardrop. 1981) ราชู (Raju. 1988, 1990) รัดเนอร์ (Rudner. 1977) และวายนอร์ (Wainer. 1993) เป็นต้น ในกลุ่มนี้ระยะหลังนิยมใช้วิธีตามแนวคิดของราชู และแนวคิดของ คิม และโคเฮน ซึ่งราชูเสนอให้ใช้การวัดขนาดพื้นที่บนช่วงความสามารถแบบเปิด (Open Interval) บางครั้งเรียกว่า “ การวัดพื้นที่ที่สมบูรณ์ ” (Exact Area Measures) ส่วนคิม และโคเฮน เสนอให้วัดขนาดพื้นที่บนช่วงความสามารถแบบปิด (Closed Interval Area) ภายใต้การวัดขนาดพื้นที่ทั้ง 2 แนวคิดนี้ ก็สามารถวัดพื้นที่ได้ 2 ลักษณะอีกด้วย คือ พื้นที่แบบมีเครื่องหมาย และไม่มีเครื่องหมาย (Signed and Unsigned Area) โดยมีสูตรในการคำนวณ ดังนี้

ถ้าเป็นแบบจำลอง 3 พารามิเตอร์

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

โดยที่

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

a b และ c เป็นพารามิเตอร์ที่แสดงลักษณะของข้อสอบ และ D เป็นตัวคงที่ปกติกำหนดให้ = 1.7

การคำนวณขนาดของพื้นที่ระหว่างความสามารถที่กำหนดไว้ 2 จุด บนมาตรฐานความสามารถหาได้จาก

$$S(\theta_1 - \theta_2) = \int_{\theta_1}^{\theta_2} P(\theta) d(\theta) = c(\theta_1 - \theta_2) + (1-c)(Da)^{-1} \ln \frac{\{1 - \exp[Da(\theta_2 - b)]\}}{\{1 + \exp[Da(\theta_2 - b)]\}}$$

ในการศึกษาการทำหน้าที่ต่างกันของข้อสอบ จะมีพารามิเตอร์ของ 2 กลุ่ม คือ กลุ่มอ้างอิง (Reference group) ได้แก่ a_R , b_R และ c_R และกลุ่มเปรียบเทียบ (Focal group) ได้แก่ a_F , b_F และ c_F พื้นที่ชนิดมีเครื่องหมาย (Closed - Interval Signed Area - CSA) หาได้จาก

$$CSA = \int_{\theta_1}^{\theta_2} [P_R(\theta) - P_F(\theta)] d\theta = S_R(\theta_1, \theta_2) - S_F(\theta_1, \theta_2)$$

และพื้นที่ชนิดไม่มีเครื่องหมาย (Closed-Interval Unsigned Area - CUA) หาได้จาก

$$CUA = \int_{\theta_1}^{\theta_2} |P_R(\theta) - P_F(\theta)| d\theta$$

4. งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่ศึกษาเกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ที่ใช้แนวคิดทฤษฎีการตอบข้อสอบอย่างเดียวนั้นไม่ค่อยพบ ส่วนใหญ่จะศึกษาร่วมกับวิธีอื่น ๆ โดยมีผู้ศึกษาไว้มากพอสมควร ในที่นี้จะแยกกล่าวเป็น 2 ส่วน คือ งานวิจัยของต่างประเทศ กับงานวิจัยภายในประเทศ

4.1 งานวิจัยของต่างประเทศ

งานวิจัยของต่างประเทศที่ทำการศึกษามีทั้งที่ใช้ข้อมูลจำลอง (Simulated data) และข้อมูลจริง (Real data) ในระยะหลังมานี้ส่วนใหญ่ใช้ข้อมูลจำลอง ซึ่งจะนำเสนอ ดังนี้

4.1.1 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยข้อมูลจริง

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยข้อมูลจริงส่วนมากเป็นวิธีการศึกษาเปรียบเทียบเทคนิคการตรวจสอบต่าง ๆ ดังนี้

เพิร์ลแมน และคณะ (Perlman et al. 1988) ได้ศึกษาความคงที่ของวิธีประมาณค่าการทำหน้าที่ต่างกันของข้อสอบ 4 วิธี ได้แก่ วิธีแปลงค่าความยาก วิธีแปลงค่าความยากที่ดัดแปลงโดย Shepard วิธีวิเคราะห์ค่าที่เหลือแบบ 1 พารามิเตอร์ของราสส์ และวิธี แมนเทิล แชนเซล (MH)

กลุ่มตัวอย่างที่ใช้มี 30 กลุ่ม ตั้งแต่ 600 จนถึง 2,000 คน ซึ่งสุ่มมาจากนักเรียนระดับ 9 จำนวน 54,896 คน ประกอบด้วยคนผิวขาว ผิวดำ และชนชาติที่พูดภาษาสเปนให้ทำแบบทดสอบทักษะความสามารถขั้นต่ำของรัฐชิคาโก ซึ่งเป็นแบบทดสอบแบบเลือกตอบ 46 ข้อ การศึกษาพบว่าแต่ละวิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ไม่คงที่เมื่อขนาดของกลุ่มตัวอย่างต่างกัน และ

ความเชื่อมั่นของตัวบ่งชี้การทำหน้าที่ต่างกันของข้อสอบจะมีปัญหา ถ้าใช้ขนาดของกลุ่มตัวอย่างต่ำกว่า 666 คน

บากิ และเฟอรารา (Baghi and Ferrara. 1989) ได้เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วย 3 วิธี ได้แก่ การใช้ทฤษฎีการตอบข้อสอบ วิธีแปลงค่าความยาก และวิธี แมนเทล เฮนเซล โดยศึกษาในแง่ของผลกระทบจากขนาดของกลุ่มตัวอย่างในผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ และระดับของความสัมพันธ์ระหว่างค่าสถิติจากทั้ง 3 วิธีนี้

ข้อมูลที่ใช้ในการศึกษา เป็นข้อมูลที่สุ่มจากนักเรียนระดับ 9 จำนวน 50,000 คน ที่ตอบแบบทดสอบทักษะความเป็นพลเมืองดี แบบเลือกตอบ จำนวน 45 ข้อ ประเมินความรู้ และทักษะ รวม 3 ด้าน ได้แก่ ด้านรัฐบาลประชาธิปไตย ด้านการเมือง และพฤติกรรมทางการเมือง และด้านหลักการ สิทธิ และความรับผิดชอบ กลุ่มตัวอย่างมี 4 ขนาด ได้แก่ 200, 500, 750 และ 1,000 คน

ผลจากการศึกษา พบว่า ไม่มีข้อสอบข้อใดที่มีการทำหน้าที่ต่างกันทั้งในการเปรียบเทียบระหว่างผู้สอบผิวดำ และผิวขาว และผู้สอบหญิง และชาย การแสดงค่าประมาณความยากด้วยแผนภูมิเส้นแสดงความสัมพันธ์ระหว่างค่าความยากต่างกลุ่มอยู่ในรูปเส้นตรงเกือบสมบูรณ์ ความสอดคล้องกันระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบของ ราสซ์ และวิธีแปลงค่าความยากมีความสัมพันธ์กันสูงมากทุกขนาดของกลุ่มตัวอย่างโดยใช้ค่าสัมประสิทธิ์สหสัมพันธ์แบบลำดับที่

แฮมเบิลตัน และโรเจอร์ส (Hambleton and Rogers. 1989) ได้เปรียบเทียบวิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ 2 วิธี คือ วิธีทฤษฎีการตอบข้อสอบ และวิธีแมนเทล เฮนเซล (MH) ข้อมูลที่ใช้เป็นผลการสอบแบบทดสอบ New Mexico High School Proficiency Exam (NMHSPE) จำนวน 150 ข้อ ของนักเรียนผิวขาว 8,000 คน และนักเรียนพื้นเมือง 2,600 คน

NMHSPE เป็นแบบทดสอบชนิดเลือกตอบ 4 ตัวเลือก วัดทักษะการดำเนินชีวิต 5 ด้าน คือ ความรู้เกี่ยวกับแหล่งทรัพยากรของชุมชน ความรู้เกี่ยวกับผู้บริโภค ความรู้เกี่ยวกับรัฐบาลและกฎหมาย ความรู้เกี่ยวกับสุขภาพทางกายและจิต และความรู้เกี่ยวกับอาชีพ

กลุ่มตัวอย่างที่ศึกษา สุ่มมากลุ่มละ 2,000 คน และแบ่งเป็นกลุ่มย่อย ๆ ละ 1,000 คน การศึกษาการทำหน้าที่ต่างกันของข้อสอบ 2 คู่ นี้ เพื่อตรวจสอบความคงที่ในการระบุข้อสอบที่ทำหน้าที่ต่างกัน โดยพิจารณาร้อยละของข้อสอบที่ถูกระบุว่าทำหน้าที่ต่างกันหรือไม่ต่างกันซ้ำกันในการตรวจสอบ 2 ครั้ง

ผลการศึกษา พบว่า วิธีทฤษฎีการตอบข้อสอบมีความสอดคล้องของผลการตรวจสอบร้อยละ 73 ส่วนวิธี แมนเทล เฮนเซล มีความสอดคล้องของผลการตรวจสอบร้อยละ 80 และพบว่าร้อยละ 47 ของข้อสอบที่ถูกระบุว่าทำหน้าที่ต่างกันในการตรวจสอบคู่ที่ 1 เป็นข้อสอบทำหน้าที่ต่างกันในการตรวจสอบคู่ที่ 2 ด้วย สำหรับวิธีแมนเทล เฮนเซล และร้อยละ 61 สำหรับวิธีทฤษฎีการตอบข้อสอบ ในทางกลับกันพบว่า ร้อยละ 64 ของข้อสอบที่ทำหน้าที่ต่างกันในการตรวจสอบคู่ที่ 2 เป็นข้อสอบที่ทำหน้าที่ต่างกันในการตรวจสอบคู่ที่ 1 สำหรับวิธีแมนเทล เฮนเซล และร้อยละ 56 สำหรับวิธีทฤษฎีการตอบข้อสอบ จะเห็นได้ว่าความคงที่ของการตรวจสอบทั้ง 2 วิธี อยู่ในระดับปานกลาง

แม็คโกรกัน (McGrogan. 1989) ได้ศึกษาการทำหน้าที่ต่างกันของข้อสอบตามวัฒนธรรม จากแบบทดสอบ California Achievement Tests โดยใช้ทฤษฎีการตอบข้อสอบชนิด 3 พารามิเตอร์ กลุ่มตัวอย่างเป็นนักเรียนระดับ 2 จำนวน 1,500 คน แบ่งเป็นกลุ่มละ 500 คน ตามกลุ่มผิวขาว ผิวดำ และฮิสแปนิก การตรวจสอบใช้วิธี Likelihood Rate χ^2 (RT-LR) ด้วยโปรแกรมสำเร็จรูป MULTOLOG (Thissen. 1984) ผลการศึกษา พบว่า กลุ่มผิวขาวมีคะแนนเฉลี่ยสูงกว่าอีก 2 กลุ่ม ประมาณ 0.5 - 1.0 ของความเบี่ยงเบนมาตรฐาน และเมื่อพิจารณาด้านความยากของข้อสอบ พบว่า ไม่แตกต่างกัน

เรย์โนลด์ (Reynold. 1989) ได้เปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของ ข้อสอบ CTBS/S ฉบับภาษาสเปน ระหว่างวิธีการแปลงค่าความยาก ไคสแคว์ วิธีของราสส์ และ วิธีทฤษฎีการตอบข้อสอบชนิด 3 พารามิเตอร์

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยการพิจารณาค่าสหสัมพันธ์ระหว่างวิธี ตรวจสอบ ร้อยละของความสอดคล้องระหว่างวิธีตรวจสอบที่ใช้ และการวิเคราะห์องค์ประกอบ

ผลจากการศึกษา พบว่า วิธีแปลงค่าความยาก กับไคสแคว์ มีความสัมพันธ์เกือบจะ สมบูรณ์ วิธีที่ 1 และ 2 กับวิธีที่ 3 และ 4 มีความสัมพันธ์กันน้อย ซึ่งจากการวิเคราะห์ร้อยละของ ความสอดคล้อง และการวิเคราะห์องค์ประกอบ พบว่า วิธีแปลงค่าความยาก ไคสแคว์ และทฤษฎี การตอบข้อสอบ ชนิด 3 พารามิเตอร์ วัดโครงสร้าง (construct) เดียวกัน คือ ข้อสอบที่ทำหน้าที่ ต่างกันด้านภาษา

บากิ และเฟอร์รา (Baghi and Ferrara. 1990) ได้เปรียบเทียบผลการตรวจสอบการทำ หน้าที่ต่างกันของข้อสอบด้วยทฤษฎีการตอบข้อสอบชนิด 3 พารามิเตอร์ และสถิติไคสแคว์แบบ แมนเทิล แชนเซล โดยใช้ข้อมูลผลการสอบจากการแบบทดสอบทักษะความเป็นพลเมืองดีของรัฐ แเมริแลนด์ กลุ่มตัวอย่างเป็นนักเรียนระดับ 9 จำนวน 50,000 คน ขนาดของกลุ่มตัวอย่างที่ใช้ใน การตรวจสอบด้วย วิธี แมนเทิล แชนเซล มี 4 ขนาด ได้แก่ 200, 500, 750 และ 1,000 คน ใน แต่ละกลุ่มเปรียบเทียบ

ผลการศึกษา พบว่า วิธีทฤษฎีการตอบข้อสอบตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ระหว่างกลุ่มหญิงและชาย พบข้อสอบที่ทำหน้าที่ต่างกัน 4 ข้อ และพบข้อสอบที่ทำหน้าที่ต่างกัน 3 ข้อ ในการเปรียบเทียบระหว่างผู้สอบผิวขาวและผิวดำ และในขณะที่ผลจากการหาค่าสหสัมพันธ์ ระหว่างดัชนีทั้งสองตัว แสดงความไม่สอดคล้องกันในการบ่งชี้ข้อสอบที่ทำหน้าที่ต่างกันนั้น ความ สอดคล้องโดยรวมพบว่า ไคสแคว์ แบบแมนเทิล แชนเซล เป็นวิธีที่ใช้แทนทฤษฎีการตอบข้อสอบ แบบ 3 พารามิเตอร์ ได้ถ้าขนาดของกลุ่มตัวอย่าง 750 คน ขึ้นไป

คิม และโคเฮน (Kim and Cohen. 1991) ได้ศึกษาการทำหน้าที่ต่างกันของข้อสอบที่ สร้างขึ้นอย่างจงใจให้ทำหน้าที่ต่างกัน โดยการเปรียบเทียบผลการคำนวณพื้นที่ความแตกต่าง ระหว่างโครงสร้างข้อสอบของกลุ่มตัวอย่างคนผิวขาว และผิวดำ 2 วิธี ประกอบด้วย การคำนวณ พื้นที่ภายใต้ความสามารถที่มีขอบเขตจำกัด (ปิด) และที่ไม่มีขอบเขตจำกัด (เปิด)

ขนาดของกลุ่มตัวอย่างกลุ่มละ 1,000 คน สุ่มมาจากคนผิวขาว 1,021 คน และผิวดำ 1,008 คน เครื่องมือที่ใช้ในการศึกษาเป็นแบบทดสอบคำศัพท์แบบเลือกตอบชนิด 4 ตัวเลือก จำนวน 50 ข้อ เป็นคำศัพท์ภาษาอังกฤษมาตรฐาน 40 ข้อ อีก 10 ข้อ เป็นคำศัพท์ที่มีความง่าย

สำหรับกลุ่มผิวดำ มากกว่าผิวดำ

การคำนวณพื้นที่ระหว่างฟังก์ชันการตอบ ทั้งภายใต้ช่วงความสามารถแบบเปิด และแบบปิด จะพิจารณา 2 ลักษณะ คือ พื้นที่ชนิดมีเครื่องหมาย (Signed Area) และพื้นที่ชนิดไม่มีเครื่องหมาย (Unsigned Area) แบบจำลองที่ใช้ในการศึกษาเป็นแบบจำลอง 3 พารามิเตอร์ และ 3 พารามิเตอร์ ที่กำหนดให้ c คงที่ = .23

ผลการศึกษา พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างการคำนวณพื้นที่แบบช่วงเปิด และแบบช่วงปิด แทบไม่มีความแตกต่างกัน และพื้นที่แบบไม่มีเครื่องหมายไม่ว่าจะเป็นแบบช่วงเปิดหรือปิด มีความไวต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมากกว่าแบบมีเครื่องหมาย นอกจากนั้นยังพบอีกว่าพื้นที่แบบมีเครื่องหมาย และไม่มีเครื่องหมายของการคำนวณแบบช่วงเปิด มีความสัมพันธ์กันมากกว่าการคำนวณพื้นที่แบบช่วงปิด

สำหรับความคลาดเคลื่อนในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน พบว่าเมื่อใช้แบบจำลอง 3 พารามิเตอร์ พื้นที่แบบมีเครื่องหมาย มีอัตราความคลาดเคลื่อนต่ำกว่าวิธีอื่น ๆ และมีความคลาดเคลื่อนในการยอมรับผิดพลาด จำนวน 2 ข้อ ที่ระดับ .05 ส่วนพื้นที่แบบไม่มีเครื่องหมายมีอัตราความคลาดเคลื่อนสูงกว่าวิธีอื่น ๆ เล็กน้อย

ราชู และคณะ (Raju et al. 1993) ได้ศึกษาเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่าง วิธีลอร์ด (χ^2) วิธีราชู (Z) และวิธี แมนเทิล แชนเซล (MH) เครื่องมือที่ใช้ในการศึกษาเป็นแบบทดสอบวัดค่าศัพท์ จำนวน 45 ข้อ ชนิด 5 ตัวเลือก กลุ่มตัวอย่างเป็นนักเรียนระดับ 10 และ 12 จำนวน 839 คน แยกเป็นชาย 399 คน หญิง 440 คน ผิวดำ 436 คน และผิวดำ 245 คน (อีก 158 คน เป็นเชื้อชาติอื่น)

ผลการศึกษา พบว่า การตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันระหว่างนักเรียนผิวดำและผิวดำ มีข้อสอบที่ทำหน้าที่ต่างกัน 1 ข้อ สำหรับการทดสอบด้วยวิธีลอร์ด และวิธีราชู ทั้งพื้นที่ชนิดมีเครื่องหมายและไม่มีเครื่องหมาย ได้แก่ ข้อ 41 ทั้ง 3 วิธี ส่วนวิธีแมนเทิล แชนเซล พบข้อสอบที่ทำหน้าที่ต่างกัน 2 ข้อ ได้แก่ ข้อ 14 และ 27 ทั้ง 3 ข้อ เป็นข้อที่มีความยากต่างกันมากสำหรับนักเรียนผิวดำและผิวดำ

ส่วนผลการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันระหว่างชายและหญิง ด้วยวิธีลอร์ด และวิธีราชู พบข้อสอบทำหน้าที่ต่างกันซ้ำกัน 4 ข้อ และวิธีแมนเทิล แชนเซล พบ 5 ข้อ ซ้ำกับ 3 วิธีแรก 4 ข้อ (ร้อยละ 80) ได้แก่ ข้อ 2, 18, 23, 33 และข้อที่ไม่ซ้ำกับ 3 วิธีแรกคือ ข้อ 41 ทุกข้อมีค่าความยากสูงกว่าข้อที่ทำหน้าที่ไม่ต่างกัน

เป็นที่น่าสังเกตว่า วิธีแมนเทิล แชนเซล ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันได้มากกว่าวิธีทฤษฎีการตอบข้อสอบ ซึ่งยังกับการศึกษาของ สวามินาธาน และโรเจอร์ (1990) ที่พบว่าวิธีทฤษฎีการตอบข้อสอบตรวจพบข้อสอบที่ทำหน้าที่ต่างกันได้มากกว่าวิธีแมนเทิล แชนเซล

คิม และโคเฮน (Kim and Cohen. 1995) ได้ศึกษาเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยทฤษฎีการตอบข้อสอบ ด้วยวิธีลอร์ด (Lord's χ^2) วิธีราชู (Raju's Area Measures) และวิธีทดสอบอัตราส่วนไลลิสูด (Likelihood Ratio Test) โดยใช้ข้อมูลจากแบบทดสอบจัดชั้นเรียนของนักศึกษาระดับมหาวิทยาลัย วิชาคณิตศาสตร์ 2 รูปแบบ 7 ละ 28 ข้อ

ฟอร์มแรกแบ่งเป็น 2 ส่วน สำหรับฟอร์มหลังเป็นข้อสอบชุดเดียวกัน แต่เรียงสลับกันระหว่างส่วนที่ 1 และส่วนที่ 2 ผลการศึกษาพบว่า ทั้ง 3 วิธี ให้ผลการตรวจสอบสอดคล้องกันมาก

4.1.2 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยข้อมูลจำลอง

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยข้อมูลจำลอง ส่วนใหญ่ เป็นการ ศึกษาเพื่อเปรียบเทียบวิธีการตรวจสอบต่าง ๆ ที่มีผู้เสนอขึ้นมาใช้ ดังนี้

โคเฮน และคิม (Cohen and Kim, 1993) ได้ศึกษาเปรียบเทียบการใช้ทฤษฎีการตอบข้อสอบ 2 พารามิเตอร์ โดยใช้สถิติทดสอบ χ^2 ของ ลอร์ด และสถิติทดสอบพื้นที่ความแตกต่างระหว่างโค้งลักษณะข้อสอบชนิดมีเครื่องหมาย และไม่มีเครื่องหมาย ได้แก่ สถิติ Z ของราชู โดยใช้ข้อมูลจำลองตามทฤษฎีการตอบข้อสอบ 2 พารามิเตอร์ จำลองการตอบข้อสอบ ความยาว 20 และ 60 ข้อ จำนวนผู้สอบ 100 และ 500 และมีจำนวนข้อสอบที่ทำหน้าที่ต่างกันแบบทดสอบจำนวนร้อยละ 0, 10 และ 20

ผลการวิเคราะห์พบว่า

1. แบบทดสอบที่ยาวการระบุข้อสอบที่ทำหน้าที่ต่างกันผิดพลาด (FP) จะมากขึ้น และถ้ามีจำนวนข้อสอบที่ทำหน้าที่ต่างกันอยู่ในแบบทดสอบน้อย จำนวนข้อสอบที่ถูกระบุว่าทำหน้าที่ต่างกันผิดพลาดจะยิ่งมากขึ้น

องค์ประกอบที่มีอิทธิพลต่อการระบุว่าข้อสอบทำหน้าที่ต่างกันผิดพลาดอีกอย่าง คือ ระดับนัยสำคัญ ค่ายิ่งมาก FP จะเกิดมาก (ที่ .05 จะเกิด FP มากกว่า ที่ .01) ผลที่เกิดขึ้นนี้เกิดในทุกสถานการณ์ที่ศึกษา สำหรับ Z แต่เกิดในบางสถานการณ์ สำหรับ χ^2

2. แบบทดสอบที่ยาวกว่า และมีร้อยละของข้อสอบที่ทำหน้าที่ต่างกันมากกว่า จะมีข้อสอบที่ถูกระบุว่าทำหน้าที่ไม่ต่างกันผิดพลาด (FN) มากกว่า แต่จะเกิดขึ้นกับการทดสอบด้วย Z มากกว่า

กรณีที่จำนวนผู้เข้าสอบ 100 คน พบว่า ความแตกต่างในการระบุข้อสอบที่ทำหน้าที่ไม่ต่างกันผิดพลาด สำหรับ Z และ χ^2 ไม่แตกต่างกันมากเท่ากรณีใช้ตัวอย่าง 500 คน ซึ่งมีการระบุว่าข้อสอบทำหน้าที่ไม่ต่างกันผิดพลาด สำหรับ χ^2 น้อยกว่า Z โดยเฉพาะอย่างยิ่งกรณีที่ข้อสอบที่ทำหน้าที่ต่างกันมากที่สุด คือ ร้อยละ 20

คิม และโคเฮน (Kim and Cohen, 1994) ได้ศึกษาวิธีการของลอร์ด (Lord's χ^2 Test) ในการตรวจสอบฟังก์ชันข้อสอบที่ทำหน้าที่ต่างกัน โดยพิจารณาอัตราการเกิดความคลาดเคลื่อนชนิดที่ 1 (Type I error rates) ใช้การจำลองข้อมูล กลุ่มตัวอย่าง 250 และ 1,000 คน จำนวนข้อสอบ 50 ข้อ แบบจำลอง 2, 3 และ 3 พารามิเตอร์ ที่กำหนดให้ค่าการเดาคงที่ (Fixed Guessing Parameter) และประมาณค่าพารามิเตอร์ด้วยวิธี มาร์จินัล แมกซิมัม โลลิฮูด (Marginal Maximum Likelihood Estimation) และมาร์จินัล เบย์เซียน (Marginal Bayesian Estimation) พบว่า อัตราการเกิดความคลาดเคลื่อนชนิดที่ 1 (Type I error rates) ที่ใช้วิธีประมาณค่า 2 วิธี ดังกล่าว เมื่อเทียบกับวิธีประมาณค่า แบบ จอยท์ แมกซิมัม โลลิฮูด (Joint Maximum Likelihood Estimation)

สำหรับแบบจำลอง 3 พารามิเตอร์ มีค่าคงที่ และสูงกว่าค่าคาดหวัง ในทางทฤษฎี สำหรับแบบจำลอง 2 และ 3 พารามิเตอร์ ที่กำหนดให้ค่าการเดาคงที่ มีค่าคงที่ และต่ำกว่าค่าคาดหวัง ณ ระดับนัยสำคัญ (α) ที่ใช้ในการศึกษา

โคเฮน คิม และวอลแลค (Cohen, Kim and Wollack. 1995) ได้ศึกษาวิธีการทดสอบอัตราส่วนไลเคิลฮูด (Likelihood Ratio Test) ในการตรวจสอบฟังก์ชันข้อสอบที่ทำหน้าที่ต่างกัน โดยศึกษาจากข้อมูลจำลอง กลุ่มตัวอย่าง 250 และ 1,000 คน จำนวนข้อสอบ 50 ข้อ แบบจำลอง 2, 3 และ 3 พารามิเตอร์ ที่กำหนดให้ค่าการเดาคงที่ และประมาณค่าพารามิเตอร์ด้วยวิธีมาร์จินัลแมกซ์ซิมัม ไลเคิลฮูด พบว่า อัตราการเกิดความคลาดเคลื่อนชนิดที่ 1 สำหรับแบบจำลอง 2 พารามิเตอร์ เป็นไปตามความคาดหวังในทางทฤษฎี และเมื่อใช้แบบจำลอง 3 พารามิเตอร์ และ 3 พารามิเตอร์ ที่กำหนดให้ค่าการเดาคงที่ มีความแตกต่างจากค่าคาดหวังในทางทฤษฎี ณ ระดับนัยสำคัญที่ศึกษาหากเปรียบเทียบกับวิธีของลอร์ด โดยภาพรวมแล้วจะมีค่าใกล้เคียงกับค่าคาดหวังในทางทฤษฎีมากกว่า

เฟนสเทน (Feinstein. 1995) ได้ศึกษาผลของความแตกต่างของพารามิเตอร์ข้อสอบ สำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี Closed Interval Area ใช้การจำลองข้อมูลโดยกำหนดให้การกระจายของค่าพารามิเตอร์ระหว่างกลุ่มอ้างอิง (Reference group) และกลุ่มเปรียบเทียบ (Focal group) แตกต่างกัน ผลการศึกษาพบว่า เมื่อกำหนดให้การเดา (c) แตกต่างกัน วิธีคิม 1 (CSA) มีความสามารถในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันได้ดีกว่า วิธีคิม 2 (CUA) ในกรณีที่ความแตกต่างระหว่างค่าพารามิเตอร์ มีขนาดปานกลางถึงขนาดใหญ่

4.2 งานวิจัยในประเทศ

สำหรับงานวิจัยในประเทศที่ศึกษาเกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบ มีดังนี้

ชัชชัย เผ่าพงศ์ (2527) ได้วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบมาตรฐานวัดความถนัดทางการเรียนคณิตศาสตร์และภาษาระดับมัธยมศึกษาตอนต้น ซึ่งพัฒนาโดยสำนักทดสอบทางการศึกษาและจิตวิทยา มหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร ระหว่างกลุ่มผู้สอบเพศชายและเพศหญิงด้วยการเปรียบเทียบโค้งลักษณะข้อสอบ ชนิด 3 พารามิเตอร์ กลุ่มตัวอย่างเป็นนักเรียนชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2524 ทุกภาคภูมิศาสตร์ทั่วประเทศ คือ ภาคเหนือ ภาคกลาง ภาคใต้ ภาคตะวันออก และภาคตะวันออกเฉียงเหนือ ผลการวิเคราะห์พบว่าแบบทดสอบวัดความถนัดทางการเรียนคณิตศาสตร์จำนวน 30 ข้อ มีข้อสอบที่ทำหน้าที่ต่างกันต่อกลุ่มใดกลุ่มหนึ่งโดยเฉพาะ 9 ข้อ คือ ข้อสอบทำหน้าที่ต่างกันต่อกลุ่มนักเรียนชาย 7 ข้อ และทำหน้าที่ต่างกันต่อกลุ่มนักเรียนหญิง 2 ข้อ ข้อสอบที่ทำหน้าที่ต่างกันต่อกลุ่มนักเรียนชายและหญิงในระดับปานกลางขึ้นไป มีจำนวน 5 ข้อ ซึ่งวัดในเรื่องร้อยละ การหาปริมาตรและการหาความยาวของเส้นรอบรูปสามเหลี่ยม จำนวนเรื่องละ 1 ข้อ และโจทย์ปัญหา จำนวน 2 ข้อ ข้อสอบจำนวน 5 ข้อ นี้เป็นข้อสอบที่ทำหน้าที่ต่างกันต่อกลุ่มนักเรียนหญิงในช่วงความสามารถแรกและทำหน้าที่ต่างกันต่อกลุ่มนักเรียนชายในช่วงความสามารถต่อมา จำนวน 1 ข้อ ส่วนแบบทดสอบวัดความถนัดทางการเรียนด้านภาษาจำนวน 40 ข้อ มีข้อสอบที่ทำหน้าที่ต่างกันเฉพาะกลุ่มนักเรียนชาย 3 ข้อ และเฉพาะกลุ่มนักเรียนหญิงมี 8 ข้อ

ข้อสอบที่ทำหน้าที่ต่างกันในระดับปานกลางขึ้นไปมีจำนวน 9 ข้อ ซึ่งวัดความเข้าใจเกี่ยวกับการอ่าน คำประพันธ์ บทร้อยกรอง อย่างละ 1 ข้อ และวัดความเข้าใจเกี่ยวกับการอ่านข้อความ 7 ข้อ ข้อสอบ จำนวน 9 ข้อนี้ เป็นข้อที่ทำหน้าที่ต่างกันเฉพาะกลุ่มนักเรียนชายจำนวน 1 ข้อ และเป็นข้อสอบที่ทำหน้าที่ต่างกันเฉพาะกลุ่มนักเรียนหญิง 6 ข้อ และอีก 2 ข้อ ทำหน้าที่ต่างกันต่อกลุ่มนักเรียนชายในระดับความสามารถต้น ๆ และทำหน้าที่ต่างกันต่อกลุ่มนักเรียนหญิงในระดับความสามารถที่สูงขึ้น

X ทศนีย์ พิรมนตรี (2529) ได้วิเคราะห์การทำหน้าที่ต่างกันของแบบทดสอบวิชาคณิตศาสตร์ โครงการตรวจสอบคุณภาพการศึกษาของสำนักงานทดสอบทางการศึกษา กรมวิชาการ กระทรวงศึกษาธิการ กลุ่มตัวอย่างเป็นนักเรียนชั้นมัธยมศึกษาปีที่ 6 ปีการศึกษา 2526 จำนวน 7,036 คน โดยเปรียบเทียบจำนวนข้อสอบที่ทำหน้าที่ต่างกันระหว่างกลุ่มนักเรียนในกรุงเทพมหานครกับกลุ่มนักเรียนในภาคภูมิศาสตร์ทั้ง 5 ภาค คือ ภาคเหนือ ภาคกลาง ภาคใต้ ภาคตะวันออกเฉียงเหนือ และภาคตะวันออกเฉียงเหนือ ด้วยวิธีวิเคราะห์ 3 วิธี คือ วิธีกำหนดจุดค่าเฉลี่ย วิธีทดสอบความแตกต่างระหว่างกลุ่มด้วยสถิติ ไคสแคว์ ในโมเดลล็อกลิเนียร์ 2 โมเดล คือ โมเดลที่ไม่มีพารามิเตอร์ผลร่วมระหว่างระดับคะแนนกับกลุ่ม และโมเดลที่ไม่มีพารามิเตอร์ของผลหลักที่เกิดจากกลุ่ม และวิธีโค้งลักษณะข้อสอบที่ใช้พารามิเตอร์ 3 ตัว ผลการวิเคราะห์ พบว่า ข้อสอบที่ทำหน้าที่ต่างกันจากวิธีโค้งลักษณะข้อสอบที่ใช้พารามิเตอร์ 3 ตัว มีจำนวนมากที่สุด และในแต่ละวิธีวิเคราะห์ พบว่า มีข้อสอบที่ทำหน้าที่ต่างกันซ้ำกันระหว่างกลุ่มนักเรียนในกรุงเทพมหานครกับกลุ่มนักเรียนในทุกภาค

X สุรศักดิ์ อมรรัตนศักดิ์ (2531) ได้เปรียบเทียบผลของการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบที่แตกต่างกัน 4 วิธี คือ วิธีวิเคราะห์ความแปรปรวน วิธีแปลงค่าความยากของข้อสอบ วิธีโค้งลักษณะข้อสอบที่มีพารามิเตอร์ 1 ตัว และวิธีโค้งลักษณะข้อสอบที่มีพารามิเตอร์ 3 ตัว โดยหาสัมประสิทธิ์สหสัมพันธ์ของดัชนีการทำหน้าที่ต่างกันระหว่างวิธีวิเคราะห์การทำหน้าที่ต่างกัน 4 วิธี และเปรียบเทียบความแตกต่างของผลการคัดเลือกก่อนและหลังการศึกษาการทำหน้าที่ต่างกันของข้อสอบตามวิธีการคิดคะแนนรวมที่แตกต่างกัน 6 วิธี ในด้านจำนวนผู้ได้รับการคัดเลือก สัดส่วนของชาย : หญิง ที่ได้รับการคัดเลือก และความเชื่อมั่นของแบบทดสอบ โดยกลุ่มตัวอย่าง คือ ผู้เข้าสอบแข่งขันเพื่อบรรจุเข้ารับราชการเป็นข้าราชการครูในปี พ.ศ. 2529 จำนวน 2,340 คน เป็นชาย 1,170 คน หญิง 1,170 คน เครื่องมือที่ใช้ในการรวบรวมข้อมูลได้แก่แบบทดสอบที่ใช้สอบแข่งขันเพื่อบรรจุเข้ารับราชการเป็นข้าราชการครูในปี พ.ศ. 2529 ซึ่งมี 4 ฉบับ คือ ฉบับที่ 1 แบบทดสอบวิชาเอก (ความรู้เกี่ยวกับการประถมศึกษา) ฉบับที่ 2 แบบทดสอบวิชาการศึกษาและกฎหมาย ฉบับที่ 3 แบบทดสอบภาษาไทย ฉบับที่ 4 แบบทดสอบความสามารถทั่วไป ผลการศึกษา พบว่า วิธีวิเคราะห์การทำหน้าที่ต่างกันแต่ละวิธีพบจำนวนข้อสอบที่ทำหน้าที่ต่างกันแตกต่างกัน โดยวิธีโค้งลักษณะข้อสอบที่มีพารามิเตอร์ 3 ตัว พบจำนวนข้อสอบที่ทำหน้าที่ต่างกันมากที่สุด รองลงมาได้แก่ วิธีวิเคราะห์ความแปรปรวน และวิธีแปลงค่าความยากของข้อสอบพบจำนวนข้อที่ทำหน้าที่ต่างกันน้อยที่สุด ส่วนค่าสัมประสิทธิ์สหสัมพันธ์ของดัชนีข้อสอบที่ทำหน้าที่ต่างกันระหว่างวิธีวิเคราะห์

ข้อสอบที่ทำหน้าที่ต่างกัน 4 วิธี พบว่า มีค่าอยู่ระหว่าง .7535 ถึง .9921 ในด้านจำนวนผู้ได้รับการคัดเลือกเมื่อเปรียบเทียบความแตกต่างของผลการคัดเลือกก่อนและหลังการศึกษาการทำหน้าที่ต่างกันของข้อสอบตามวิธีคิดคะแนนดิบกับวิธีรวมคะแนนแบบอื่นๆ อีก 5 วิธี มีจำนวนผู้ได้รับการคัดเลือกแตกต่างกันประมาณ 4% ถึง 24% ส่วนวิธีแปลงคะแนนดิบเป็นคะแนนมาตรฐานที่ปกติกับวิธีรวมคะแนนแบบอื่นๆ อีก 4 วิธี มีจำนวนผู้ได้รับการคัดเลือกแตกต่างกันประมาณ 6 % ถึง 23 % ส่วนในเรื่องความเชื่อมั่นของแบบทดสอบ พบว่า ค่าความเชื่อมั่นของแบบทดสอบลดลงกว่าแบบทดสอบเดิมเล็กน้อย

X พัทธ์ ปิยะพันธ์ (2531) ได้ศึกษาการทำหน้าที่ต่างกันของข้อสอบจากแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนวิชาคณิตศาสตร์ เรื่องสมการ ที่มีต่อกลุ่มนักเรียนชายกับหญิงและกลุ่มนักเรียนในกรุงเทพมหานครกับจังหวัดสมุทรสาคร ด้วยวิธีวิเคราะห์ 3 วิธี คือ วิธีแปลงค่าความยากของข้อสอบ วิธีไคสแคว์ และวิธีโค้งลักษณะข้อสอบ ใช้เทคนิคการวิเคราะห์ตัวลงเพื่อตรวจทานแหล่งของการทำหน้าที่ต่างกัน กลุ่มตัวอย่างเป็นนักเรียนชั้นประถมศึกษาปีที่ 6 ปีการศึกษา 2530 ของโรงเรียนสังกัดสำนักงานคณะกรรมการการประถมศึกษาแห่งชาติในกรุงเทพมหานครกับจังหวัดสมุทรสาคร จำนวน 1,422 คน ซึ่งเลือกโดยวิธีสุ่มแบบแบ่งชั้น พบว่า ในแต่ละวิธีพบจำนวนข้อสอบที่ทำหน้าที่ต่างกันแตกต่างกัน วิธีโค้งลักษณะข้อสอบพบจำนวนข้อสอบที่ทำหน้าที่ต่างกันมากที่สุดจากแบบทดสอบ 45 ข้อ พบข้อสอบที่ทำหน้าที่ต่างกันระหว่างกลุ่มนักเรียนชายกับนักเรียนหญิงจำนวน 7 ข้อ และพบข้อสอบที่ทำหน้าที่ต่างกันระหว่างกลุ่มนักเรียนในกรุงเทพมหานครกับสมุทรสาคร จำนวน 18 ข้อ ส่วนวิธีแปลงค่าความยากของข้อสอบ และวิธี ไคสแคว์ พบข้อสอบที่ทำหน้าที่ต่างกันระหว่างกลุ่มนักเรียนชายกับหญิง จำนวน 2 ข้อ และ 3 ข้อ และพบข้อสอบที่ทำหน้าที่ต่างกันระหว่างกลุ่มนักเรียนในกรุงเทพมหานครกับสมุทรสาคร จำนวน 1 ข้อ และ 3 ข้อ ตามลำดับ

ค่าสหสัมพันธ์ของดัชนีการทำหน้าที่ต่างกันของข้อสอบจากการวิเคราะห์ 3 วิธี พบว่าดัชนีการทำหน้าที่ต่างกันของข้อสอบระหว่างกลุ่มนักเรียนชายกับหญิงจากการวิเคราะห์โดยวิธีแปลงค่าความยากของข้อสอบกับวิธีโค้งลักษณะข้อสอบมีค่า .5618 ($p < .001$) และดัชนี การทำหน้าที่ต่างกันของข้อสอบระหว่างกลุ่มนักเรียน ในกรุงเทพมหานครกับจังหวัดสมุทรสาครจากการวิเคราะห์โดยวิธีไคสแคว์กับวิธีโค้งลักษณะข้อสอบมีค่า .6009 ($p < .001$) ส่วนวิธีอื่นสัมพันธ์กันอย่างไม่มีความสำคัญทางสถิติ

ค่าความเชื่อมั่นของแบบทดสอบหลังการคัดเลือกข้อสอบที่ทำหน้าที่ต่างกันออกไปแล้ว แตกต่างกับค่าความเชื่อมั่นของแบบทดสอบก่อนการคัดเลือกข้อสอบที่ทำหน้าที่ต่างกันอย่างไม่มีความสำคัญทางสถิติ โดยความเชื่อมั่นของแบบทดสอบก่อนการคัดเลือกข้อสอบที่ทำหน้าที่ต่างกันมีค่า .9277 ค่าความเชื่อมั่นของแบบทดสอบหลังการคัดเลือกข้อสอบที่ทำหน้าที่ต่างกันระหว่างกลุ่มนักเรียนชายกับหญิง โดยวิธีแปลงค่าความยากของข้อสอบ วิธีไคสแคว์ และวิธีโค้งลักษณะข้อสอบมีค่า .9316, .9277 และ .9286 ตามลำดับ และค่าความเชื่อมั่นของแบบทดสอบหลังการคัดเลือกข้อสอบที่ทำหน้าที่ต่างกันระหว่างกลุ่มนักเรียนในกรุงเทพมหานครกับจังหวัดสมุทรสาครมีค่า .9339, .9311 และ .9392 ตามลำดับ เมื่อวิเคราะห์หาแหล่งการทำหน้าที่ต่างกันของข้อสอบ

ในข้อสอบที่ทำหน้าที่ต่างกันทุกข้อไม่พบว่าแหล่งการทำหน้าที่ต่างกันของข้อสอบแต่ละข้อเกิดจากการเขียนตัวเลือก

✕ สุพัฒน์ สุกมลสันต์ (2534) ได้วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบภาษาอังกฤษเพื่อคัดเลือกเข้าศึกษาในมหาวิทยาลัย โดยใช้ข้อมูลการตอบแบบทดสอบภาษาอังกฤษเข้ามหาวิทยาลัยชุด กข และชุด กขค ปี 2531 - 2533 ซึ่งมีข้อสอบชุดละ 100 ข้อ ตัวแปรที่ศึกษาการทำหน้าที่ต่างกันของข้อสอบ คือ เพศ และภาคภูมิศาสตร์ของผู้สอบ ซึ่งแยกออกเป็น 5 ภาค ตามภูมิลำเนาของผู้สอบ ได้แก่ ภาคกลาง ภาคตะวันออก ภาคตะวันออกเฉียงเหนือ ภาคเหนือ และภาคใต้ เทคนิควิธีที่นำมาวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกัน ได้แก่ วิธีกำหนดจุดค่าเดลต้า (Delta Plot Method) วิธีไคสแคว์ชนิดที่แบ่งความสามารถของผู้สอบเป็น 3 ระดับ ได้แก่ กลุ่มความสามารถระดับต่ำ (ผู้ได้คะแนนรวม 0 - 40 คะแนน) กลุ่มความสามารถระดับปานกลาง (ผู้ได้คะแนนรวม 41 - 70 คะแนน) และระดับสูง (ผู้ได้คะแนนรวม 71 - 100 คะแนน) วิธีที่ 3 ที่นำมาใช้วิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกัน ได้แก่ วิธีการวัดพื้นที่ความแตกต่างระหว่างโค้งลักษณะข้อสอบที่วิเคราะห์ตามทฤษฎีการตอบข้อสอบชนิด 3 พารามิเตอร์

เกณฑ์การตัดสินข้อสอบที่ทำหน้าที่ต่างกันสำหรับวิธีต่าง ๆ ที่ สุพัฒน์ สุกมลสันต์ กำหนดขึ้น คือ

1. วิธีกำหนดจุดค่าเดลต้า ข้อสอบที่ทำหน้าที่ต่างกัน คือ ข้อที่มีระยะตั้งฉากจากค่าคู่อันดับเดลต้าไปยังเส้นแกนหลัง (d) มากกว่า $\pm 1.964 SD$
2. วิธีไคสแคว์ ข้อสอบที่ทำหน้าที่ต่างกัน คือ ข้อสอบที่ผู้สอบต่างกลุ่มที่อยู่ในระดับคะแนนเดียวกันมีสัดส่วนการตอบถูกหรือผิดแตกต่างกันอย่างมีนัยสำคัญที่ระดับ $\alpha = 0.017$
3. วิธีการวัดพื้นที่ความแตกต่างระหว่างโค้งลักษณะข้อสอบ ข้อสอบที่ทำหน้าที่ต่างกัน คือ ข้อที่มีพื้นที่ความแตกต่างมากกว่า 0.40 โดยแบ่งเป็น 3 ระดับ
 - 3.4 พื้นที่ > 0.70 แสดงว่า ข้อสอบทำหน้าที่ต่างกันมาก
 - 3.5 พื้นที่ระหว่าง 0.40 - 0.70 แสดงว่า ข้อสอบทำหน้าที่ต่างกันปานกลาง
 - 3.6 พื้นที่มีค่ามากกว่า 0.00 แต่น้อยกว่า 0.40 แสดงว่า ข้อสอบทำหน้าที่ต่างกันน้อย

ผลการวิเคราะห์พบว่า แบบทดสอบภาษาอังกฤษ กข กขค มีจำนวนข้อสอบที่ทำหน้าที่ต่างกันต่อเพศ 7 - 28 ข้อ และ 4 - 41 ข้อ ตามลำดับ โดยมีแนวโน้มทำหน้าที่ต่างกันต่อเพศชายมากกว่าหญิง และมีข้อสอบที่ทำหน้าที่ต่างกันต่อภาคอื่นมากกว่าภาคกลาง วิธีตรวจสอบที่พบจำนวนข้อสอบที่ทำหน้าที่ต่างกันมากที่สุด คือ วิธีโค้งลักษณะข้อสอบ เมื่อกำหนดขนาดพื้นที่มากกว่า .40 ขึ้นไป รองลงมา คือ วิธีไคสแคว์ และวิธีกำหนดจุดเดลต้า

กาญจนา วัฒนสุนทร (2537) ได้พัฒนาเกณฑ์ตัดสินข้อสอบที่ทำหน้าที่ต่างกันทางเพศ โดยใช้ข้อมูลเชิงประจักษ์สำหรับดัชนี 4 ตัว คือ พื้นที่ระหว่างโค้งการตอบข้อสอบแบบมีเครื่องหมาย (SA) และไม่มีเครื่องหมาย (UA) จากวิธีทฤษฎีการตอบข้อสอบแบบจำลอง 2 พารามิเตอร์ วิธีแมนเทิล แชนเซล และวิธี SIBTEST

ข้อมูลที่ใช้เป็นข้อมูลจากการตอบข้อสอบคัดเลือกบุคคลเข้าศึกษาในสถาบันอุดมศึกษาของทบวงมหาวิทยาลัยปีการศึกษา 2535 โดยใช้ความยาวแบบทดสอบ 20, 30 และ 40 ข้อ สำหรับวิชาคณิตศาสตร์ และ 50, 60, 70 และ 80 ข้อ สำหรับวิชาภาษาอังกฤษ โดยใช้กลุ่มผู้สอบ 6 ขนาด คือ 100, 200, 400, 600, 800 และ 1,000 คน

การพัฒนาเกณฑ์กระทำโดยคำนวณค่าดัชนีทั้ง 4 ตัว จากข้อมูลการตอบข้อสอบของผู้สอบเพศเดียวกัน เพศละ 50 ค่าสำหรับแต่ละความยาวแบบทดสอบ และขนาดผู้สอบ จากนั้นนำค่าดัชนีที่ได้ทั้งหมดมาวิเคราะห์ค่าเฉลี่ยและกำหนดเกณฑ์จากค่าเฉลี่ย 2 ลักษณะ คือ เกณฑ์ที่กำหนดโดยพิจารณาค่าเฉลี่ยรวมจากค่าดัชนีทุกข้อโดยไม่พิจารณาความแตกต่างในด้านความยาวแบบทดสอบ และขนาดของกลุ่มผู้สอบ และเกณฑ์ที่พัฒนาจากค่าเฉลี่ย ซึ่งพิจารณาถึงความยาวแบบทดสอบ และขนาดผู้สอบด้วย

จากนั้นนำเกณฑ์ที่กำหนดไว้ไปตัดสินค่าดัชนีที่ได้จากการวิเคราะห์ระหว่างผู้สอบเพศหญิงและชาย พบว่า ไม่ว่าจะใช้เกณฑ์ลักษณะใด ผลด้านจำนวนข้อสอบในรูปของร้อยละของข้อสอบที่ทำหน้าที่ต่างกัน และความสอดคล้องของการตัดสินมีความไม่คงที่ข้ามความยาวแบบทดสอบและขนาดผู้สอบ โดยที่เกณฑ์ในลักษณะหลังจะลดความไม่คงที่ลงได้ระดับหนึ่ง ทั้งนี้เนื่องจากค่าดัชนีแต่ละตัวที่ได้จากผู้สอบต่างเพศกันมีความไม่คงที่แปรเปลี่ยนไปเมื่อความยาวแบบทดสอบที่ใช้และขนาดผู้สอบเปลี่ยนไปเช่นกัน

นอกจากนั้นพบว่าที่ความยาวแบบทดสอบ 40 ข้อ สำหรับวิชาคณิตศาสตร์ และ 80 ข้อ สำหรับวิชาภาษาอังกฤษมีจำนวนข้อสอบที่ได้รับการตัดสินว่าทำหน้าที่ต่างกันมากกว่าในความยาวอื่นมาก ทั้งยังพบว่าเกณฑ์ที่กำหนดได้จากการศึกษาค้างนี้มีค่าสูงกว่าเกณฑ์ ที่มีผู้ใช้น้อยอยู่ในกรณีของค่า UA ในขณะที่ค่า MH มีค่าใกล้เคียงกับเกณฑ์ขั้นต่ำที่ควรใช้ในการตัดสินข้อสอบที่ทำหน้าที่ต่างกันที่พัฒนาได้ครั้งนี้ คือ $ISAI > .40$ $UA > .50$ เมื่อใช้ความยาวแบบทดสอบ 50 ข้อขึ้นไป $ISAI > .80$ $UA > 1.20$ กรณีความยาวแบบทดสอบสั้นกว่า 50 ข้อ ค่า $.60 > MH > 1.40$ และ $ISIBI > .06$ และควรปรับใช้เกณฑ์สูงขึ้นโดยพิจารณาถึงวัตถุประสงค์ที่ต้องการใช้แบบทดสอบแต่ไม่ควรต่ำกว่านี้เพื่อลดปัญหาของการเกิดความคลาดเคลื่อนประเภทที่ 1 นอกจากนั้นยังต้องการการศึกษาเพิ่มเติมเพื่อหาความยาวแบบทดสอบและจำนวนผู้สอบที่ทำให้ค่าดัชนีที่มีความคงที่เพิ่มขึ้น

เจน จันทรแสง (2537) ได้ศึกษาดัชนีการทำหน้าที่ต่างกันของข้อสอบวัดผลสัมฤทธิ์ทางการเรียนในวิชาคณิตศาสตร์ เรื่อง จำนวนจริง ที่เกิดจากตัวแปรเพศ และที่ตั้งของโรงเรียน ด้วยวิธีวิเคราะห์ข้อสอบที่ทำหน้าที่ต่างกัน 4 วิธี คือ วิธีแปลงค่าความยากของข้อสอบ วิธีวิเคราะห์ความแปรปรวน วิธีไคสแคว์ และวิธีโค้งลักษณะข้อสอบที่มีพารามิเตอร์ 1 ตัว ศึกษาสัมประสิทธิ์สหสัมพันธ์ระหว่างดัชนีการทำหน้าที่ต่างกันของข้อสอบทั้ง 4 วิธี และเปรียบเทียบความเชื่อมั่นของแบบทดสอบก่อนและหลังการคัดเลือกข้อสอบที่ทำหน้าที่ต่างกัน เครื่องมือที่ใช้เป็นแบบทดสอบชนิด

เลือกตอบ จำนวน 40 ข้อ กลุ่มตัวอย่างเป็นนักเรียนชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2536 ของโรงเรียนสังกัดกรมสามัญศึกษา จังหวัดนครสวรรค์ จำนวน 1,000 คน พบว่า

1. วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบแต่ละวิธีตรวจพบจำนวนข้อสอบที่ทำหน้าที่ต่างกันแตกต่างกัน โดยวิธีโค้งลักษณะข้อสอบที่มีพารามิเตอร์ 1 ตัว ตรวจพบจำนวนข้อสอบที่ทำหน้าที่ต่างกันได้มากที่สุด โดยพบข้อสอบที่ทำหน้าที่ต่างกันตามตัวแปรเพศ 10 ข้อ และพบข้อสอบที่ทำหน้าที่ต่างกันตามตัวแปรที่ตั้งของโรงเรียน 15 ข้อ รองลงมา คือ วิเคราะห์ความแปรปรวน และวิธีไคสแคว์ พบข้อสอบที่ทำหน้าที่ต่างกันตามตัวแปรเพศ 7 ข้อ และพบข้อสอบที่ทำหน้าที่ต่างกันตามตัวแปรที่ตั้งของโรงเรียน 14 ข้อ ส่วนวิธีแปลงค่าความยากของข้อสอบเป็นวิธีที่พบจำนวนข้อสอบที่ทำหน้าที่ต่างกันน้อยที่สุด โดยพบข้อสอบที่ทำหน้าที่ต่างกันตามตัวแปรเพศ 1 ข้อ และพบข้อสอบที่ทำหน้าที่ต่างกันตามตัวแปรที่ตั้งของโรงเรียน 5 ข้อ ตามลำดับ ซึ่งแสดงว่าเพศ และที่ตั้งของโรงเรียนเป็นตัวแปรที่ทำให้ข้อสอบทำหน้าที่ต่างกัน

2. ความสัมพันธ์ระหว่างดัชนีการทำหน้าที่ต่างกันของข้อสอบของวิธีวิเคราะห์แต่ละวิธี มีค่าเป็นบวก และมีนัยสำคัญทางสถิติที่ระดับ .001 เมื่อวิเคราะห์ตามตัวแปรเพศ พบว่า สัมประสิทธิ์สหสัมพันธ์ของดัชนีแต่ละวิธีมีค่าระหว่าง .5284 ถึง .9970 เมื่อวิเคราะห์ตามตัวแปรที่ตั้งของโรงเรียน พบว่า สัมประสิทธิ์สหสัมพันธ์ของดัชนีแต่ละวิธีมีค่าระหว่าง .6008 ถึง .9475

3. แบบทดสอบหลังจากคัดเลือกข้อสอบที่ทำหน้าที่ต่างกันออกไปแล้ว มีค่าความเชื่อมั่นลดลงจากเดิม อย่างไม่มีนัยสำคัญทางสถิติ และเมื่อปรับค่าความเชื่อมั่นของแบบทดสอบหลังจากคัดเลือกข้อสอบที่ทำหน้าที่ต่างกันออกไปแล้วให้มีจำนวนเท่าเดิมโดยใช้สูตรของ สเปียร์แมน บรวน์ พบว่า ค่าความเชื่อมั่นแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

การศึกษาเกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ผ่านมาส่วนใหญ่จะศึกษาร่วมกับวิธีอื่นๆ เช่น งานวิจัยของสงบ ลักษณะ (Laksana. 1979) เชปเพิร์ด และ คณะ (Subkoviak et Al. 1984 : 50 ; citing Shepard et Al. 1984) ไอรอนสัน และ ซับโคเวียค (Ironson and Subkoviak. 1979) บาจิ และเฟอร์รารา (Baghi and Ferrara. 1990) สุรศักดิ์ อมรรัตนศักดิ์ (2531) และ กาญจนา วัฒนสุนทร (2537) เป็นต้น ได้ข้อสรุปค่อนข้างสอดคล้องกันว่า วิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ใช้แนวคิดทฤษฎีการตอบข้อสอบมีความเหมาะสมและครอบคลุมกว่า โดยเฉพาะชนิด 3 พารามิเตอร์ นอกจากนี้ยังพบอีกว่า ทฤษฎีการตอบข้อสอบชนิด 2 และ 3 พารามิเตอร์ มีความถูกต้องและสอดคล้องกันมากที่สุดเมื่อเทียบกับวิธีอื่น ๆ แม้ว่า จะไม่สามารถตรวจสอบได้ถูกต้องร้อยละ 100 ก็ตาม ทั้งนี้เนื่องจากเกิดความคลาดเคลื่อนจากผลการตรวจสอบอยู่ 2 ลักษณะ คือ การยอมรับผิดพลาด (False Positive : FP) และการปฏิเสธผิดพลาด (False Negative : FN)

ในระยะประมาณ 10 ปี มาแล้ว วิธีที่นิยมใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ตามแนวคิดทฤษฎีการตอบข้อสอบ ได้แก่ วิธี Lord's χ^2 วิธี Raju's Area Measures และ วิธี Closed Interval Area ซึ่งมีการศึกษาทั้งจากข้อมูลจำลองและข้อมูลจริง รวมทั้งกำหนดเงื่อนไขในการสอบแตกต่างกันไป เช่น ขนาดของกลุ่มตัวอย่างต่างกัน ความยาวของแบบทดสอบต่างกัน

และจำนวนข้อสอบที่ทำหน้าที่ต่างกันแบบทดสอบต่างกัน แต่ยังไม่มีย่อสรุปที่ชัดเจนว่าวิธีใดมีประสิทธิภาพดีกว่ากันในเรื่องของการศึกษาที่กำหนด แม้ว่าจะมีผู้ศึกษาอยู่บ้าง อย่างไรก็ตามหากมีการศึกษาโดยขยายขนาดของกลุ่มตัวอย่าง ความยาวของแบบทดสอบ และจำนวนข้อสอบที่ทำหน้าที่ต่างกันแบบทดสอบให้ละเอียดยิ่งขึ้น ข้อค้นพบที่ได้น่าจะมีคุณสมบัติและน่าเชื่อถือยิ่งขึ้น

ดังนั้น ผู้วิจัยจึงสนใจที่จะศึกษาผลการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันระหว่าง วิธี Lord's χ^2 วิธี Raju's Area Measures และ วิธี Closed Interval Area เมื่อกำหนดเงื่อนไขของการสอบแตกต่างกันดังกล่าว

บทที่ 3

วิธีดำเนินการศึกษา

การศึกษาค้นคว้าครั้งนี้ มีความมุ่งหมายเพื่อเปรียบเทียบความสอดคล้องของผลการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันระหว่าง วิธี Lord's χ^2 วิธี Raju's Area Measures และ วิธี Closed Interval Area มีขั้นตอนการศึกษา ดังนี้

แหล่งข้อมูลและกลุ่มตัวอย่าง

ข้อมูลที่ใช้ในการศึกษาเป็นผลการสอบประเมินคุณภาพและความก้าวหน้าทางการศึกษาระดับประเทศ วิชาคณิตศาสตร์ ของนักเรียนชั้นประถมศึกษาปีที่ 4 ปีการศึกษา 2536 สังกัดสำนักงานการประถมศึกษาแห่งชาติ จำนวน 11,404 คน แยกเป็นเพศชาย 5,715 คน และเพศหญิง 5,689 คน

เครื่องมือที่ใช้ในการศึกษา

เครื่องมือที่ใช้ในการศึกษา ได้แก่ แบบทดสอบประเมินคุณภาพและความก้าวหน้าทางการศึกษาวิชาคณิตศาสตร์ ชั้นประถมศึกษาปีที่ 4 มีลักษณะเป็นแบบเลือกตอบ (Multiple choice) ชนิด 4 ตัวเลือก แบ่งเป็น 3 ตอน คือ

ตอนที่ 1 ทักษะการคิดคำนวณ จำนวน 20 ข้อ

ตอนที่ 2 ความรู้ความเข้าใจ จำนวน 35 ข้อ

ตอนที่ 3 โจทย์ปัญหา จำนวน 25 ข้อ

โดยมีรายละเอียดของเนื้อหาในแต่ละตอน ดังนี้

ตอนที่ 1 ทักษะการคิดคำนวณ ประกอบด้วยเนื้อหาเกี่ยวกับ

- การบวกลบจำนวนที่ไม่เกิน 10,000 7 ข้อ
- การคูณ - ทหาร จำนวนเต็ม 4 ข้อ
- การบวก - ลบ เศษส่วน 3 ข้อ
- การคูณ - ทหาร เศษส่วน 3 ข้อ
- โจทย์ระคน 3 ข้อ

ตอนที่ 2 ความรู้ความเข้าใจ ประกอบด้วยเนื้อหาเกี่ยวกับ

- ค่าประจำหลักของจำนวน 3 ข้อ
- การบวก - ลบ จำนวน 3 ข้อ
- การคูณ - ทหาร จำนวน 4 ข้อ

- เศษส่วน	4 ข้อ
- ทศนิยม	4 ข้อ
- การวัดความยาว การชั่ง การตวง	4 ข้อ
- เงินและการบันทึกรายรับ - รายจ่าย	1 ข้อ
- เวลา	3 ข้อ
- แผนภูมิรูปภาพ แผนภูมิแท่ง และตาราง	2 ข้อ
- เส้นตรง มุม รูปเรขาคณิต เส้นขนาน รูปสมมาตร	3 ข้อ
- พื้นที่ของรูปเหลี่ยม	2 ข้อ
- บทประยุกต์	2 ข้อ

ตอนที่ 3 โจทย์ปัญหา ประกอบด้วยเนื้อหาเกี่ยวกับ

- การบวก - ลบ จำนวน	2 ข้อ
- การคูณ - ทหาร จำนวน	3 ข้อ
- การบวก ลบ คูณ และหารจำนวน	5 ข้อ
- การบวก ลบ คูณ และหารเศษส่วน	5 ข้อ
- เวลา	2 ข้อ
- การวัดความยาว การชั่ง การตวง	3 ข้อ
- บทประยุกต์	5 ข้อ

แบบทดสอบฉบับนี้สร้างโดยศึกษานิเทศก์ และครูผู้สอนวิชาคณิตศาสตร์ ชั้นประถมศึกษาปีที่ 4 ที่ผ่านการคัดเลือกจากสำนักงานการประถมศึกษาแห่งชาติ และนำไปทดลองใช้ 2 ครั้ง เพื่อหาข้อบกพร่อง แล้วนำมาปรับปรุงแก้ไข ก่อนนำไปใช้ในสถานการณ์จริง

การเตรียมข้อมูลในการศึกษา

1. คัดเลือกข้อสอบ ผู้วิจัยได้วิเคราะห์หาคุณภาพรายข้อ ด้วยโปรแกรม BILOG แล้วคัดเลือกเอาเฉพาะข้อที่มีค่าความยากระหว่าง .20 - .80 และค่าอำนาจจำแนกตั้งแต่ .10 ขึ้นไป รวมทั้งค่าพารามิเตอร์ของข้อสอบที่ประมาณค่าได้ตามทฤษฎีการตอบข้อสอบ ซึ่งเหมาะสม (fit) กับแบบจำลองที่ศึกษา ในการศึกษานี้ ใช้แบบจำลอง 2 พารามิเตอร์ เนื่องจากการประมาณค่าพารามิเตอร์การเดา (c) ในแบบจำลอง 3 พารามิเตอร์ มีค่าเฉลี่ยค่อนข้างต่ำ (.14) เมื่อนำมาคำนวณในการเปรียบเทียบผลการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันด้วย วิธี Raju's Area Measures จะต้องกำหนดให้ c เท่ากันทั้ง 2 กลุ่ม จึงไม่น่าจะให้ผลต่างกัน เมื่อใช้แบบจำลอง 2 พารามิเตอร์ ที่ $c = 0$ ประกอบกับจากผลการศึกษาของ คิม และโคเฮน (Kim and Cohen. 1994 : 222) พบว่า แบบจำลอง 3 พารามิเตอร์ จะทำให้เกิดอัตราความคลาดเคลื่อนชนิดที่ 1 (Type I error rate) สูงกว่าแบบจำลอง 2 พารามิเตอร์ และ 3 พารามิเตอร์ เมื่อกำหนดให้ c คงที่สำหรับวิธีตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันด้วย Lord's χ^2 ดังนั้น จึงเลือกใช้แบบจำลอง

2 พารามิเตอร์แทน ทำให้ได้ข้อสอบที่ใช้ในการศึกษาครั้งนี้ จำนวน 75 ข้อ จากทั้งหมด 80 ข้อ และหาค่าคุณภาพของแบบทดสอบด้านความเชื่อมั่นด้วยสูตรครุเดอร์-ริชาร์ดสัน (KR-20) เท่ากับ .906

2. การจัดเตรียมข้อมูลให้สอดคล้องกับจุดมุ่งหมายของการศึกษา แบ่งออกเป็น 2 ส่วน ดังนี้

ส่วนที่ 1 การตรวจสอบจำนวนข้อสอบที่ทำหน้าที่ต่างกัน ความสอดคล้องในการตรวจสอบ และสหสัมพันธ์ระหว่างวิธีตรวจสอบ 3 วิธี มีขั้นตอน ดังนี้

1. สุ่มกลุ่มตัวอย่างขนาด 300, 500, 700 และ 1,000 คน ตามลำดับ แยกเพศ ชาย - หญิง

2. สุ่มข้อสอบ จากข้อสอบ 75 ข้อ เพื่อจัดชุดแบบทดสอบใหม่ ขนาด 20, 30, 40 และ 60 ข้อ ตามลำดับ โดยใช้การสุ่มอย่างง่าย (จับสลากหมายเลขข้อสอบที่เขียนไว้ในตะกร้า) แบบใส่คืน แยกตามทักษะ 3 ตอน ดังกล่าวข้างต้น

ได้กรอบหรือแบบแผนของการศึกษาในส่วนที่ 1 ดังนี้

I	N	วิธีตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน				
		LC	ESA	EUA	CSA	CUA
20	300					
	500					
	700					
	1000					
30	300					
	500					
	700					
	1000					
40	300					
	500					
	700					
	1000					
60	300					
	500					
	700					
	1000					

* I = จำนวนข้อสอบ
N = ขนาดกลุ่มตัวอย่าง

- ส่วนที่ 2 การตรวจสอบความไม่สอดคล้องของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่าง 3 วิธี มีขั้นตอน ดังนี้
- 1) ประมาณค่าพารามิเตอร์ข้อสอบรายข้อ จากกลุ่มตัวอย่างนักเรียนทั้งหมด 11,404 คน เพื่อทดสอบข้อมูลกับแบบจำลอง
 - 2) ประมาณค่าพารามิเตอร์ข้อสอบรายข้อ แยกเพศชาย-หญิง
 - 3) ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบรายข้อ ด้วยวิธีทั้ง 3 เพื่อหาว่าข้อใดที่ทุกวิธีตรวจสอบได้เข้ากัน แสดงว่าข้อสอบข้อนั้นทำหน้าที่ต่างกันจริง แล้วแยกจำนวนข้อสอบที่ทำหน้าที่ต่างกันออกจากข้อสอบที่ทำหน้าที่ไม่ต่างกัน ซึ่งจะได้นำไปจัดชุดแบบทดสอบต่อไป
 - 4) จัดชุดแบบทดสอบ ให้มีจำนวนข้อสอบ ขนาด 20, 30, 40 และ 60 ข้อ โดยให้มีสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันอยู่ในแบบทดสอบ เป็นร้อยละ 0, 10 และ 20 ตามลำดับ
 - 5) สุ่มกลุ่มตัวอย่างขนาด 300, 500, 700 และ 1,000 คน แยกเพศ ชาย-หญิง
 - 6) ตรวจสอบความไม่สอดคล้องในการตัดสินข้อสอบที่ทำหน้าที่ต่างกัน ระหว่าง 3 วิธี

ได้กรอบหรือแบบแผนของการศึกษาในส่วนที่ 2 ดังนี้

N	I	% DIF	วิธีตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน				
			LC	ESA	EUA	CSA	CUA
300	20	0%					
		10%					
		20%					
	30	0%					
		10%					
		20%					
	40	0%					
		10%					
		20%					
	60	0%					
		10%					
		20%					

N	I	% DIF	วิธีตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน				
			LC	ESA	EUA	CSA	CUA
500	20	0%					
		10%					
		20%					
	30	0%					
		10%					
		20%					
	40	0%					
		10%					
		20%					
	60	0%					
		10%					
		20%					
700	20	0%					
		10%					
		20%					
	30	0%					
		10%					
		20%					
	40	0%					
		10%					
		20%					
	60	0%					
		10%					
		20%					

N	I	% DIF	วิธีตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน				
			LC	ESA	EUA	CSA	CUA
1,000	20	0%					
		10%					
		20%					
	30	0%					
		10%					
		20%					
	40	0%					
		10%					
		20%					
	60	0%					
		10%					
		20%					

* N = ขนาดกลุ่มตัวอย่าง

I = จำนวนข้อสอบ

% DIF = สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันแบบทดสอบ

การวิเคราะห์ข้อมูล

1) การตรวจสอบความเป็นมิติเดียวของแบบทดสอบ ใช้การวิเคราะห์องค์ประกอบ (Factor Analysis) ด้วยการวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis) หมุนแกนด้วยวิธีแวนแมกซ์ (Varimax) ตามข้อเสนอของลอร์ด สำหรับการตรวจสอบความเป็นมิติเดียวครั้งนี้ใช้โปรแกรมสำเร็จรูป SPSS-PC⁺

2) การประมาณค่าพารามิเตอร์ของข้อสอบ ใช้โปรแกรม BILOG 3 (Mislevy and Bock, 1990) ซึ่งเป็นโปรแกรมที่สามารถใช้กับไมโครคอมพิวเตอร์ เพื่อประมาณค่าพารามิเตอร์ของข้อสอบ ได้แก่ ค่าอำนาจจำแนก (a) ค่าความยาก (b) และค่าการเดา (c) ตามแนวคิดทฤษฎีการตอบข้อสอบ (Item Response Theory) วิธีประมาณค่ามีทั้งวิธีมาร์จินัล แมกซิมัม โลลิฮูด (Marginal Maximum Likelihood Method) และวิธีเบย์ (Bayesian Method) ข้อมูลที่ใช้เป็นแบบ 0 - 1 (Binary) แบ่งเป็นโปรแกรมย่อยที่สำคัญ ดังนี้

INPUT เป็นโปรแกรมย่อยที่ใช้ในขั้นตอนการกำหนดลักษณะของข้อมูล การแปรข้อมูลดิบ และแบบจำลองที่ใช้ในการวิเคราะห์

CALIBRATE เป็นโปรแกรมย่อยที่ใช้ในขั้นตอนการประมาณค่าพารามิเตอร์ของข้อสอบ เช่น การกำหนดการคำนวณซ้ำด้วยวิธี EM Newton - Gauss ที่รอบ เมื่อความยาวแบบทดสอบมากกว่าหรือน้อยกว่า 50 ข้อ การกำหนดวิธีการประมาณค่าพารามิเตอร์ เป็นต้น สำหรับผลของการวิเคราะห์ (Out put) ประกอบด้วย 3 ส่วน (Phase) คือ

ส่วนที่ 1 เป็นผลการวิเคราะห์ข้อมูลที่เกี่ยวข้องกับค่าอำนาจจำแนก (r) และค่าความยาก (p) ตามแนวคิดทฤษฎีดั้งเดิม (Classical Test Theory : CTT)

ส่วนที่ 2 เป็นการประมาณค่าพารามิเตอร์ของข้อสอบ ตามแนวคิดทฤษฎีการตอบข้อสอบ ได้แก่ ค่าอำนาจจำแนก (a) ค่าความยาก (b) ค่าการเดา (c) เมตริกซ์ความแปรปรวน และความแปรปรวนร่วม รวมทั้งจะมีผลการทดสอบความเหมาะสมระหว่างแบบจำลองกับข้อมูลในการศึกษาด้วย ซึ่งการคำนวณเพื่อเปรียบเทียบค่าพารามิเตอร์ของข้อสอบ จะต้องใช้ไฟล์พารามิเตอร์ของข้อสอบ และไฟล์เมตริกซ์ความแปรปรวนและความแปรปรวนร่วม ในโปรแกรม EQUATE และโปรแกรม IRTDIF ต่อไป

ส่วนที่ 3 เป็นผลการวิเคราะห์เกี่ยวกับฟังก์ชันสารสนเทศของข้อสอบหรือแบบทดสอบ (Test Information Function)

3) การแปลงหน่วยของพารามิเตอร์ข้อสอบ (Metric Transformation) ถือเป็นขั้นตอนที่สำคัญขั้นตอนหนึ่ง ก่อนที่จะนำมาเปรียบเทียบกัน นั่นคือ ต้องแปลงให้อยู่ในหน่วยเดียวกันหรือสเกลร่วม (Common Scale) จึงจะเปรียบเทียบกันได้ โดยใช้โปรแกรม EQUATE 2.1 (Baker, 1993)

ซึ่งโปรแกรมนี้ได้ใช้แนวคิดของสต็อกกิง และลอร์ด (Stocking and Lord, 1983) ในการแสดงความสัมพันธ์ระหว่างเมตริกซ์ของพารามิเตอร์ที่ประมาณค่าได้ระหว่าง 2 กลุ่ม เป็นเชิงเส้นตรง (Linear) เขียนในรูปของสมการได้ ดังนี้

$$a_F^* = \frac{a_F}{A}$$

$$\text{Var}(a_F^*) = \text{Var}\left(\frac{a_F}{A^2}\right)$$

$$b_F^* = Ab_F + K$$

$$\text{Var}(b_F^*) = A^2 \text{Var}(b_F)$$

เมื่อ	A	แทน	ค่าความลาดชันของเส้นถดถอย
	K	แทน	จุดตัดบนแกนตั้ง (Intercept)
	a_F	แทน	ค่าอำนาจจำแนกของข้อสอบก่อนการแปลงหน่วย
	a_F^*	แทน	ค่าอำนาจจำแนกของข้อสอบบนมาตราที่แปลงแล้ว
	b_F	แทน	ค่าความยากของข้อสอบก่อนการแปลงหน่วย
	b_F^*	แทน	ค่าความยากของข้อสอบบนมาตราที่แปลงแล้ว

ค่าที่ได้จากโปรแกรม EQUATE จะให้ค่า A และ K เพื่อที่จะนำไปใช้ในการแปลงหน่วย ของค่าพารามิเตอร์ข้อสอบในโปรแกรม IRTDIF ต่อไป

4) การตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน ใช้โปรแกรม IRTDIF 1.0 พัฒนาโดยคิม และโคเฮน (Kim and Cohen. 1991) ซึ่งสามารถตรวจวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ ได้ ทั้ง 3 วิธีใหญ่ ได้แก่ วิธีตามแนวคิดของลอร์ด (Lord's χ^2) วิธีตามแนวคิดของราชู (Raju's Area Measures) และวิธีตามแนวคิดของคิม และโคเฮน (Closed Interval Area)

ลักษณะของโปรแกรม เป็นแบบ Interactive mode คือ วิธีการทำงานในลักษณะที่มีการโต้ตอบกันระหว่างระบบกับผู้ใช้งาน ประกอบด้วย 7 ไฟล์หลัก ได้แก่

- IRTDIF.FOR
- IRTDIF.OBJ
- IRTDIF.EXE
- IRTDIF.TEX เป็นคู่มือการใช้โปรแกรม
- SAMPLE1.COV เป็นเมตริกซ์ความแปรปรวนร่วมกลุ่มที่ 1
- SAMPLE2.COV เป็นเมตริกซ์ความแปรปรวนร่วมกลุ่มที่ 2
- SAMPLE.OUT เป็นตัวอย่างผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ

ระหว่าง 5 วิธีย่อย (Lord's χ^2 ESA EUA CSA และ CUA)

สำหรับไฟล์ที่นำมาใช้ในการคำนวณประกอบด้วย ไฟล์เมตริกซ์ความแปรปรวนและความแปรปรวนร่วม (Variance - Covariance Matrices) ได้จากโปรแกรม BILOG ส่วนค่า A และ K ได้จากโปรแกรม EQUATE

การศึกษาข้อสอบที่ทำหน้าที่ต่างกัน ซึ่งหมายถึง ข้อสอบที่ผู้เข้าสอบความสามารถเท่ากัน แต่อยู่ต่างกลุ่มกันมีโอกาสในการตอบข้อสอบข้อนั้น ๆ ได้ถูกต้องไม่เท่ากัน สามารถเขียนฟังก์ชันในการตอบข้อสอบได้ ดังนี้

$$P(\theta) = \{1 + \exp[-Da(\theta - b)]\}^{-1} \quad \text{ในกรณีเป็นแบบจำลอง 2 พารามิเตอร์}$$

เมื่อ D แทน ค่าคงที่ เพื่อปรับโค้งลักษณะข้อสอบ มีค่าเท่ากับ 1.7

a แทน ค่าอำนาจจำแนก

b แทน ค่าความยาก

สำหรับการศึกษาการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันตามแนวคิดของลอร์ด (Lord. 1980) เป็นการเปรียบเทียบค่าพารามิเตอร์ของข้อสอบระหว่าง 2 กลุ่มย่อย ซึ่งเขียนสมมติฐานการวิจัย ดังนี้

$$\begin{aligned} H_0 &: a_R = a_F, \quad b_R = b_F \\ H_a &: H_0 \quad \text{ไม่เป็นจริง} \end{aligned}$$

เมื่อ	a_R แทน	ค่าอำนาจจำแนกของข้อสอบกลุ่ม Reference
	a_F แทน	ค่าอำนาจจำแนกของข้อสอบกลุ่ม Focal
	b_R แทน	ค่าความยากของข้อสอบกลุ่ม Reference
	b_F แทน	ค่าความยากของข้อสอบกลุ่ม Focal

สถิติทดสอบ

$$\chi^2 = (\xi_R - \xi_F)' \Sigma^{-1} (\xi_R - \xi_F)$$

$$\begin{aligned} \text{เมื่อ } \xi_R &= (\hat{a}_R, \hat{b}_R)' \\ \xi_F &= (\hat{a}_F, \hat{b}_F)' \end{aligned}$$

Σ อยู่ในรูปเมตริกซ์ 2×2 และ

$$\Sigma = \Sigma_{JR} + \Sigma_{JF}$$

เมื่อ Σ_{JR} และ Σ_{JF} เป็นเมตริกซ์ Variance และ Covariance ของ ξ_R และ ξ_F ตามลำดับ

การศึกษาการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันตามแนวคิดของราชู (Raju's Area Measures) เป็นการวัดพื้นที่ความแตกต่างของฟังก์ชันการตอบข้อสอบระหว่าง 2 กลุ่มย่อย ซึ่งราชูเสนอให้ใช้ช่วงความสามารถแบบเปิด บางครั้งเรียกว่า “ การวัดพื้นที่สมบูรณ์ ” (Exact Area Measures) โดยพิจารณาทั้ง 2 กรณี คือ แบบมีเครื่องหมาย (Signed Area) และแบบไม่มีเครื่องหมาย (Unsigned Area) ใช้สถิติทดสอบ Z เขียนในรูปสัญลักษณ์ได้ดังนี้

$$ESA = \hat{b}_F - \hat{b}_R$$

$$ESA = \begin{cases} |\hat{b}_F - \hat{b}_R| & \text{เมื่อ } \hat{a}_R = \hat{a}_F \\ H & \end{cases}$$

เมื่อ

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

สถิติทดสอบของ ESA คือ

$$Z(\text{ESA}) = \frac{\hat{b}_F - \hat{b}_R}{\left[\text{Var}(\hat{b}_F) + \text{Var}(\hat{b}_R) \right]^{\frac{1}{2}}}$$

ส่วนสถิติทดสอบ EUA คือ

$$Z(H) = \frac{H}{\left[\text{Var}(H) \right]^{\frac{1}{2}}}$$

ส่วนการศึกษการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันตามแนวคิดของ คิม และโคเฮน (Closed Interval Area) ก็เป็นการวัดพื้นที่ความแตกต่างของฟังก์ชันการตอบข้อสอบระหว่าง 2 กลุ่มย่อย เช่นเดียวกันแต่กำหนดช่วงความสามารถเป็นแบบปิด รวมทั้งพิจารณาพื้นที่แบบมีเครื่องหมาย และพื้นที่แบบไม่มีเครื่องหมาย ที่ต่างจากแนวคิดของราชู คือ ไม่ใช้สถิติทดสอบตัดสินข้อสอบที่ทำหน้าที่ต่างกัน หากแต่นำขนาดของพื้นที่ไปเทียบกับเกณฑ์ที่ตัดสินการทำหน้าที่ต่างกันของข้อสอบแทน

เขียนในรูปทั่วไปได้ ดังนี้

$$s(\theta_L, \theta_U) = \frac{1}{Da} \ln \left[\frac{1 + \exp\{Da(\theta_U - b)\}}{1 + \exp\{Da(\theta_L - b)\}} \right]$$

เมื่อ

$s(\theta_L, \theta_U)$ แทน ขนาดของพื้นที่อยู่ในช่วงความสามารถ θ_L, θ_U

และ

$$\text{CSA} = s_R(\theta_L, \theta_U) - s_F(\theta_L, \theta_U)$$

เมื่อ $a_R = a_F = a$ โค้งแสดงลักษณะของข้อสอบ (ICC) จะไม่ตัดกัน

จะได้

$$\text{CUA} = |s_R(\theta_L, \theta_U) - s_F(\theta_L, \theta_U)|$$

ถ้า $a_R \neq a_F$ โค้งแสดงลักษณะของข้อสอบจะตัดกัน

$$\theta_X = \frac{a_R b_R - a_F b_F}{a_R - a_F}$$

เมื่อ θ_x แทน ระดับความสามารถที่อยู่ข้างนอกช่วง (θ_L, θ_U) จะได้ว่า

$$CUA = |s_R(\theta_L, \theta_U) - s_F(\theta_L, \theta_U)|$$

แต่ถ้า θ_x อยู่ข้างในช่วงความสามารถ (θ_L, θ_U) จะได้ว่า

$$CUA = |s_R(\theta_L, \theta_x) - s_F(\theta_L, \theta_x)| + |s_R(\theta_x, \theta_U) - s_F(\theta_x, \theta_U)|$$

5) การหาความสัมพันธ์ระหว่างวิธีตรวจสอบ 3 วิธี ด้วยสัมประสิทธิ์สหสัมพันธ์แบบเรียงอันดับ (Rank Order Correlation Coefficient) ซึ่งหาได้จากสูตร

$$\rho = 1 - \frac{6 \sum D^2}{N(N^2 - 1)} \quad (\text{Guilford and Fruchter. 1978 : 295})$$

เมื่อ ρ แทน สัมประสิทธิ์สหสัมพันธ์แบบเรียงอันดับ

N แทน จำนวนคู่ของข้อมูล (ดัชนีการทำหน้าที่ต่างกันของข้อสอบ)

D แทน ผลต่างของอันดับที่ของดัชนีการทำหน้าที่ต่างกันของข้อสอบ

6) การหาความสอดคล้องในการตรวจสอบระหว่างวิธีตรวจสอบ เป็นการวิเคราะห์ความสอดคล้องของผลการตัดสินการทำหน้าที่ต่างกันของข้อสอบ คำนวณได้จาก

สูตร

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

เมื่อ R คือ จำนวนข้อสอบที่ซ้ำกันในการตัดสิน

T คือ จำนวนข้อสอบที่ทำหน้าที่ต่างกันทั้งหมดในการเปรียบเทียบแต่ละคู่

(กาญจนา วัฒนสุนทร. 2537 : 104)

7) การทดสอบความแตกต่างของผลการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันระหว่างวิธีตรวจสอบ

7.1 จำนวนข้อสอบที่พบว่าทำหน้าที่ต่างกันในแต่ละวิธี โดยใช้สถิติทดสอบ คือ Cochran Q Test

$$Q = \frac{(K-1) \left[K \sum_{j=1}^k G_j^2 - \left(\sum_{j=1}^k G_j \right)^2 \right]}{K \sum_{i=1}^N L_i - \sum_{i=1}^N L_i^2} \quad (\text{Siegel. 1956 : 162})$$

เมื่อ K แทน จำนวนกลุ่มตัวอย่าง (จำนวนวิธีตรวจสอบ)
 G_j แทน ผลรวมในแนวตั้ง (จำนวนข้อสอบที่ทำหน้าที่ต่างกันแต่ละวิธี)
 L_i แทน ผลรวมในแนวนอน

7.2 อัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ใช้สถิติทดสอบ Z - Test ดังนี้

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\bar{p} \cdot \bar{q} \cdot \left(\frac{n_1 + n_2}{n_1 n_2} \right)}} \quad (\text{Guilford and Fruchter. 1978 : 159})$$

7.2.1 อัตราความไม่สอดคล้องชนิดที่ 1

เมื่อ $\hat{p}_1$ แทน อัตราความไม่สอดคล้องชนิดที่ 1 คำนวณจากจำนวนข้อสอบที่ตรวจสอบผิดพลาดด้วยจำนวนข้อสอบที่ทำหน้าที่ไม่ต่างกัน (Non DIF) ทั้งหมด ซึ่งตรวจสอบด้วยวิธีที่ 1

$$\hat{p}_1 = \frac{f_1}{n_1}$$

เมื่อ $\hat{p}_2$ แทน อัตราความไม่สอดคล้องชนิดที่ 1 คำนวณจากจำนวนข้อสอบที่ตรวจสอบผิดพลาดด้วยจำนวนข้อสอบที่ทำหน้าที่ไม่ต่างกัน (Non DIF) ทั้งหมด ซึ่งตรวจสอบด้วยวิธีที่ 2

$$\hat{p}_2 = \frac{f_2}{n_2}$$

เมื่อ $\bar{p}$ แทน อัตราความไม่สอดคล้องชนิดที่ 1 คำนวณจากจำนวนข้อสอบที่ตรวจสอบผิดพลาด ทั้ง 2 วิธี ซึ่งคำนวณจากจำนวนข้อสอบที่ตรวจสอบผิดพลาดที่ตรวจสอบด้วยวิธีที่ 1 และวิธีที่ 2 ทารด้วยจำนวนข้อสอบที่ทำหน้าที่ไม่ต่างกันทั้งหมด ของทั้ง 2 วิธี

$$\bar{p}_e = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

$$\text{และ } \bar{q}_e = 1 - \bar{p}_e$$

7.2.2 อัตราความไม่สอดคล้องชนิดที่ 2

เมื่อ $\hat{p}_1$ แทน อัตราความไม่สอดคล้องชนิดที่ 2 คำนวณจากจำนวนข้อสอบที่ตรวจสอบ
ผิดพลาดด้วยจำนวนข้อสอบที่ทำหน้าที่ต่างกัน (DIF) ทั้งหมด ซึ่ง
ตรวจสอบด้วยวิธีที่ 1

เมื่อ $\hat{p}_2$ แทน อัตราความไม่สอดคล้องชนิดที่ 2 คำนวณจากจำนวนข้อสอบที่ตรวจสอบ
ผิดพลาดด้วยจำนวนข้อสอบที่ทำหน้าที่ต่างกัน (DIF) ทั้งหมด ซึ่ง
ตรวจสอบด้วยวิธีที่ 2

เมื่อ $\bar{p}_e$ แทน อัตราความไม่สอดคล้องชนิดที่ 2 คำนวณจากจำนวนข้อสอบที่ตรวจสอบ
ผิดพลาด ทั้ง 2 วิธี ซึ่งคำนวณจากจำนวนข้อสอบที่ตรวจสอบผิดพลาด
ที่ตรวจสอบด้วยวิธีที่ 1 และวิธีที่ 2 หารด้วยจำนวนข้อสอบที่ทำหน้าที่
ต่างกันทั้งหมด ของทั้ง 2 วิธี

$$\text{และ } \bar{q}_e = 1 - \bar{p}_e$$

บทที่ 4

ผลการวิเคราะห์ข้อมูล

การนำเสนอผลการวิเคราะห์ข้อมูล ได้เรียงตามลำดับ ดังนี้

ตอนที่ 1 ค่าสถิติพื้นฐานของข้อมูลผลการสอบ คุณภาพของแบบทดสอบ

การตรวจสอบความเป็นมิติเดียวของแบบทดสอบ และค่าพารามิเตอร์ของข้อสอบ

ตอนที่ 2 ผลการตรวจสอบ การทำหน้าที่ต่างกันของข้อสอบ ด้วย 3 วิธี ได้แก่

วิธี Lord's χ^2 วิธี Raju's Area Measures และวิธี Closed Interval Area โดยใช้เกณฑ์ในการพิจารณา 3 ส่วน คือ

2.1 จำนวนข้อสอบที่ทำหน้าที่ต่างกัน ความสอดคล้องในการตรวจสอบ และความสัมพันธ์ระหว่างวิธีตรวจสอบ

2.2 อัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

2.3 อัตราความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

เพื่อให้การนำเสนอผลการวิเคราะห์เข้าใจตรงกันและง่ายขึ้น จึงกำหนดสัญลักษณ์แทน

ดังนี้

$\bar{X}$	แทน	ค่าเฉลี่ยเลขคณิต
S.D.	แทน	ค่าส่วนเบี่ยงเบนมาตรฐาน
KR-20	แทน	ความเชื่อมั่นของแบบทดสอบ
N	แทน	จำนวนคน
I	แทน	จำนวนข้อสอบ
a_i	แทน	ค่าอำนาจจำแนกของข้อสอบรวม
b_i	แทน	ค่าความยากของข้อสอบรวม
a_1	แทน	ค่าอำนาจจำแนกของข้อสอบ กลุ่มเพศชาย
a_2	แทน	ค่าอำนาจจำแนกของข้อสอบ กลุ่มเพศหญิง
b_1	แทน	ค่าความยากของข้อสอบ กลุ่มเพศชาย
b_2	แทน	ค่าความยากของข้อสอบ กลุ่มเพศหญิง
DIF	แทน	การทำหน้าที่ต่างกันของข้อสอบ
LC	แทน	วิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ตามแนวคิดของลอว์ด
ESA	แทน	วิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ตามแนวคิดของราชู โดยคิดเครื่องหมาย เรียกว่า “วิธีราชู1”
EUA	แทน	วิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ตามแนวคิดของราชู โดยไม่คิดเครื่องหมาย เรียกว่า “วิธีราชู2”

CSA	แทน	วิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ตามแนวคิดของคิม โดยคิดเครื่องหมาย เรียกว่า “วิธีคิม 1”
CUA	แทน	วิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ตามแนวคิดของคิม โดยไม่คิดเครื่องหมาย เรียกว่า “วิธีคิม 2”
FP	แทน	การยอมรับผิดพลาด เป็นการตัดสินใจข้อสอบทำหน้าที่ต่างกัน ทั้ง ๆ ที่ข้อสอบทำหน้าที่ไม่ต่างกัน
FN	แทน	การปฏิเสธผิดพลาด เป็นการตัดสินใจข้อสอบทำหน้าที่ไม่ต่างกัน ทั้ง ๆ ที่ข้อสอบทำหน้าที่ต่างกัน
*	แทน	$p < .05$
**	แทน	$p < .01$
***	แทน	$p < .001$

ตอนที่ 1 ค่าสถิติพื้นฐานของข้อมูลผลการสอบ คุณภาพของแบบทดสอบ การตรวจสอบความเป็นมิติเดียวของแบบทดสอบ และค่าพารามิเตอร์ของข้อสอบ

1) ค่าสถิติพื้นฐานของข้อมูลผลการตอบข้อสอบ และผลการวิเคราะห์คุณภาพของแบบทดสอบ ได้แก่ ค่าเฉลี่ย ($\bar{X}$) ค่าส่วนเบี่ยงเบนมาตรฐาน (S.D.) และค่าความเชื่อมั่นของแบบทดสอบ (KR-20) ดังตาราง 1

ตาราง 1 ค่าเฉลี่ย ค่าส่วนเบี่ยงเบนมาตรฐาน และค่าความเชื่อมั่นของแบบทดสอบวิชาคณิตศาสตร์ จำนวน 75 ข้อ

กลุ่ม	N (คน)	$\bar{X}$	S.D.	KR-20
ชาย	5,715	35.15	12.95	.904
หญิง	5,689	37.49	12.70	.907
รวม	11,404	36.38	12.87	.906

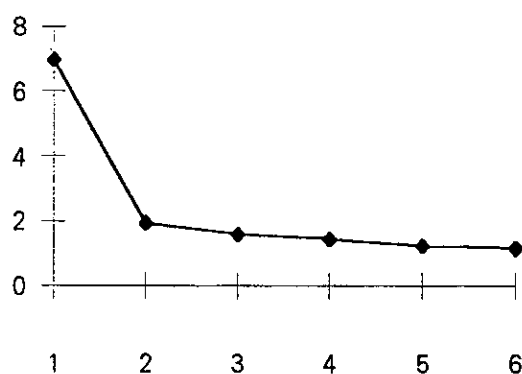
จากตาราง 1 พบว่า กลุ่มตัวอย่างเพศชาย 5,715 คน หญิง 5,689 คน ค่าเฉลี่ยของคะแนนสอบเพศชายเท่ากับ 35.15 คะแนน เพศหญิงเท่ากับ 37.49 คะแนน และค่าความเชื่อมั่นของแบบทดสอบ โดยใช้สูตรครุเดอร์ริชาร์ดสัน (KR-20) เท่ากับ .906

2) การตรวจสอบความเป็นมิติเดียวของแบบทดสอบ แสดงผลดังตาราง 2 และ
ภาพประกอบ 4

ตาราง 2 ค่าไอเกนและร้อยละของความแปรปรวนขององค์ประกอบ

องค์ประกอบ	ค่าไอเกน	ร้อยละความแปรปรวน
1	6.97	8.7
2	1.95	2.4
3	1.62	2.0
4	1.47	1.8
5	1.22	1.5
6	1.19	1.5

ค่าไอเกน



องค์ประกอบ

ภาพประกอบ 4 ผลการตรวจสอบความเป็นมิติเดียวของแบบทดสอบ

จากตาราง 2 และภาพประกอบ 4 พบว่า ค่าไอเกนขององค์ประกอบที่ 1 มีค่าสูงกว่าค่าไอเกนขององค์ประกอบที่ 2 ประมาณ 3.57 เท่า นับว่ามีค่าเด่นกว่าองค์ประกอบอื่นแม้ว่าค่าร้อยละของความแปรปรวนจะไม่สูงนัก (ร้อยละ 8.7) หากพิจารณาตามเกณฑ์ที่เรคเคส (Reckase. อ้างถึงใน Raju, 1993 : 304) เสนอว่าค่าร้อยละของความแปรปรวนควรมีค่าไม่น้อยกว่าร้อยละ 20 จึงจะถือว่าแบบทดสอบมีความเป็นมิติเดียว ซึ่งในทางปฏิบัติหาพบได้ยากเนื่องจากแบบทดสอบประกอบด้วยเนื้อหาหลายอย่างและวัดหลายสมรรถภาพ

3) ค่าพารามิเตอร์ของข้อสอบ (a, b) แยกตามเพศชาย-หญิง แสดงผลดังตาราง 3

ตาราง 3 ค่าพารามิเตอร์ของข้อสอบ แยกตามเพศชาย-หญิง

ข้อที่	พารามิเตอร์ข้อสอบ					
	a_1	b_1	a_2	b_2	a_t	b_t
1	.552	.047	.513	-.348	.542	-.144
2	.535	-.183	.514	-.553	.532	-.361
3	.525	.054	.474	-.225	.506	-.082
4	.704	.534	.679	.256	.698	.391
5	.726	.508	.782	.198	.763	.344
6	.441	-.668	.396	-1.224	.429	-.913
7	.514	.813	.537	.563	.528	.682
8	.517	.993	.497	.662	.513	.819
9	.635	1.249	.598	1.067	.623	1.148
10	.603	1.393	.603	1.128	.611	1.244
11	.647	.746	.658	.466	.659	.598
12	.804	.724	.805	.469	.816	.587
13	.804	1.022	.780	.757	.800	.880
14	.739	.426	.715	.165	.736	.292
15	.806	-1.152	.755	-1.518	.796	-1.312
16	.731	-.916	.696	-1.334	.727	-1.103
17	.734	-.936	.743	-1.299	.750	-1.103
18	.678	-.378	.666	-.717	.681	-.543
19	.576	.910	.532	.873	.554	.889
20	.481	1.010	.465	.911	.474	.955
21	.545	-.638	.534	-.674	.535	-.664
22	.352	.346	.333	.578	.336	.463
23	.347	-.208	.356	-.359	.353	-.287
24	.257	2.176	.287	2.036	.266	2.140
25	.279	1.680	.320	1.613	.295	1.666
26	.200	2.178	.231	1.968	.212	2.094
27	.312	-.004	.293	.134	.299	.059
28	.633	-.910	.664	-1.181	.655	-1.040
29	.512	.554	.470	.512	.491	.530
30	.426	-.156	.402	-.407	.416	-.281

ตาราง 3 (ต่อ)

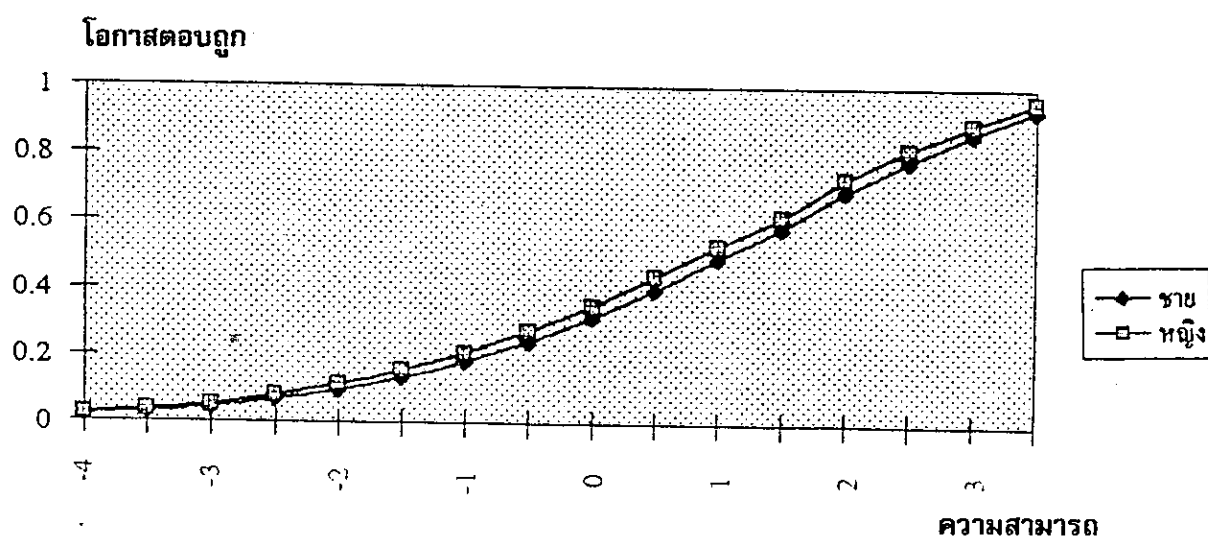
ข้อที่	พารามิเตอร์ข้อสอบ					
	a_1	b_1	a_2	b_2	a_t	b_t
31	.136	-.640	.101	-.857	.113	-.769
32	.331	.643	.349	.646	.337	.646
33	.126	3.922	.124	4.039	.120	4.156
34	.428	1.183	.423	1.281	.419	1.243
35	.457	-.307	.436	-.244	.443	-.281
36	.240	2.025	.265	1.522	.253	1.759
37	.226	1.100	.266	.548	.249	.791
38	.313	-.397	.276	-.628	.295	-.506
39	.443	.165	.411	.150	.426	.156
40	.469	-.496	.454	-.677	.463	-.587
41	.357	-1.420	.382	-1.679	.373	-1.540
42	.495	-.121	.517	-.217	.505	-.172
43	.287	1.275	.276	1.169	.278	1.338
44	.501	-1.468	.469	-2.026	.495	-1.706
45	.367	1.635	.416	1.295	.392	1.450
46	.432	-.310	.494	-.078	.452	-.194
47	.459	-1.307	.428	-1.612	.444	-1.455
48	.437	-3.204	.356	-4.148	.410	-3.681
49	.522	-1.496	.434	-2.185	.488	-1.780
50	.491	1.081	.566	1.049	.518	1.074
51	.417	.614	.456	.524	.435	.566
52	.494	1.071	.461	1.044	.478	1.055
53	.302	1.371	.349	1.090	.325	1.222
54	.219	1.389	.274	.764	.248	1.033
55	.134	3.705	.142	3.283	.133	3.625
56	.523	-.272	.547	-.179	.525	-.229
57	.609	-.559	.532	-.649	.565	-.607
58	.442	-.349	.398	-.669	.422	-.501
59	.487	.046	.486	-.169	.486	-.062
60	.518	-.229	.514	-.405	.517	-.318
61	.315	.040	.248	-.397	.284	-.154
62	.605	-.616	.584	-.911	.600	-.757
63	.448	-.177	.489	-.290	.469	-.238
64	.176	1.634	.236	1.148	.202	1.379
65	.274	.762	.236	.969	.250	.871

ตาราง 3 (ต่อ)

ข้อที่	พารามิเตอร์ข้อสอบ					
	a_1	b_1	a_2	b_2	a_t	b_t
66	.588	-.202	.585	-.505	.589	-.352
67	.473	.577	.439	.707	.450	.644
68	.544	-.280	.534	-.292	.532	-.290
69	.237	3.441	.174	4.945	.199	4.205
70	.448	.627	.485	.411	.466	.513
71	.377	.375	.374	.482	.369	.431
72	.219	1.390	.226	.897	.223	1.134
73	.539	.138	.549	-.135	.548	-.002
74	.309	.674	.309	.458	.309	.564
75	.396	.481	.399	.417	.396	.447
$\bar{X}$	.456	.383	.450	.192	.454	.309
S.D.	.168	1.178	.163	1.317	.168	1.250

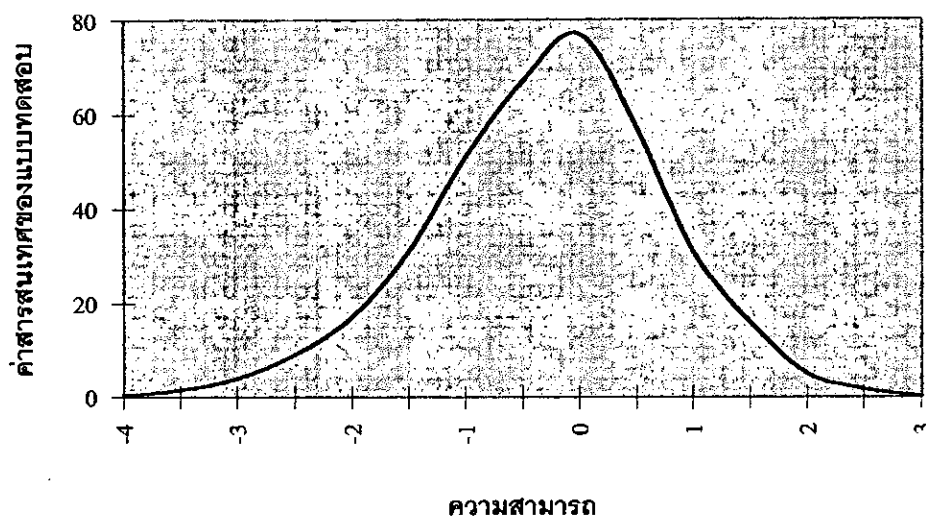
จากตาราง 3 ค่าความยากของข้อสอบเฉลี่ยเท่ากับ .309 และค่าอำนาจจำแนกเฉลี่ยเท่ากับ .454 ซึ่งค่าความยากของข้อสอบ อยู่ในระดับปานกลาง และค่าอำนาจจำแนกอยู่ในเกณฑ์ดี เป็นที่น่าสังเกต ค่าความยากของกลุ่มเพศชายจะมีค่าสูงกว่ากลุ่มเพศหญิง

เมื่อนำแบบทดสอบฉบับนี้ มาคำนวณหาความน่าจะเป็นโดยเฉลี่ยในการตอบข้อสอบได้ถูกต้อง ที่ระดับความสามารถต่างกัน 16 ระดับ แสดงได้ดังภาพประกอบ 5



ภาพประกอบ 5 ได้แสดงลักษณะข้อสอบโดยเฉลี่ยของแบบทดสอบ

สำหรับฟังก์ชันสารสนเทศของแบบทดสอบ วิชาคณิตศาสตร์ ชั้นประถมศึกษาปีที่ 4
จำนวน 75 ข้อ แสดงได้ดังภาพประกอบ 6



ภาพประกอบ 6 โค้งสารสนเทศของแบบทดสอบ (Test Information Curve)

จากภาพประกอบ 6 จะเห็นว่าโค้งสารสนเทศของแบบทดสอบกระจายไปทุกระดับความสามารถ และค่าสารสนเทศของแบบทดสอบสูงสุดที่ระดับความสามารถ (θ) เท่ากับ $-.155$

ตอนที่ 2 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วย 3 วิธี ได้แก่ วิธี Lord's χ^2 วิธี Raju's Area Measures และวิธี Closed Interval Area โดยใช้เกณฑ์ในการพิจารณา 3 ส่วน คือ

2.1 จำนวนข้อสอบที่ทำหน้าที่ต่างกัน ความสอดคล้องในการตรวจสอบ และความสัมพันธ์ระหว่างวิธีตรวจสอบ

ตาราง 4 จำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบเมื่อกลุ่มตัวอย่าง 300 คน และ
ความยาวแบบทดสอบ 20 ข้อ

ข้อที่	ค่าพารามิเตอร์				วิธีตรวจสอบ				
	a_1	b_1	a_2	b_2	LC	ESA	EUA	CSA	CUA
1	.767	.076	.564	-.407	6.22	-2.35	-1.97	<u>-0.353</u>	<u>0.517</u>
2	.911	.489	.846	.452	0.12	0.18	-0.35	0.016	0.110
3	.635	.918	.588	.743	0.81	-0.45	-0.47	-0.115	0.146
4	.799	.709	.857	.516	0.59	-0.76	0.47	-0.108	0.120
5	.693	.614	.733	.218	3.78	-1.90	1.85	-0.289	0.289
6	.699	-1.205	1.056	-.883	3.09	1.63	1.62	0.246	<u>0.538</u>
7	.851	-.237	.739	-.330	0.44	-0.24	-0.66	-0.024	0.203
8	.672	-.630	.560	-.530	2.18	0.72	-0.86	0.165	0.285
9	.269	1.944	.306	1.504	0.32	-0.54	0.38	-0.095	0.133
10	.418	-.205	.749	-.118	5.52	0.69	2.03	0.085	<u>0.758</u>
11	.338	-.398	.252	-.493	0.96	-0.11	-0.91	0.032	0.304
12	.288	1.823	.351	1.144	1.23	-1.04	0.67	-0.209	0.251
13	.529	-.524	.621	-.592	0.48	-0.07	0.59	-0.042	0.197
14	.368	1.528	.508	1.130	1.35	-0.76	1.09	-0.035	0.365
15	.476	.795	.785	.811	6.40	0.35	1.83	0.213	<u>0.647</u>
16	.195	2.866	.292	1.471	1.90	-1.37	1.09	-0.246	0.371
17	.399	-.040	.588	-.119	2.14	-0.09	1.34	-0.024	<u>0.483</u>
18	.614	-.859	.902	-.706	2.29	0.98	1.46	0.100	<u>0.493</u>
19	.474	-.376	.754	-.118	5.21	1.60	1.78	0.223	<u>0.631</u>
20	.643	-.039	.651	-.097	0.01	0.01	0.01	0.001	0.002
รวม					2	1	2	1	7

จากตาราง 4 จะเห็นว่าวิธีลอร์ด (LC) ตรวจพบจำนวนข้อสอบที่ทำหน้าที่ต่างกัน 2 ข้อ
วิธีราชู 1 (ESA) 1 ข้อ วิธีราชู 2 (EUA) 2 ข้อ วิธีคิม 1 (CSA) 1 ข้อ วิธีคิม 2 จำนวน 7 ข้อ
และมีเพียง 1 ข้อ ที่ตรวจพบว่าทำหน้าที่ต่างกันซ้ำกันในทุกวิธี

ตาราง 5 ผลการทดสอบความแตกต่างของจำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบระหว่างวิธี
ตรวจสอบ เมื่อกลุ่มตัวอย่าง 300 คน และความยาวแบบทดสอบ 20 ข้อ

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA	χ^2 รวม
LC		1.00	0.00	1.00	5.00	18.00**
ESA			1.00	0.00	6.00	
EUA				1.00	5.00	
CSA					6.00	
CUA						

$\chi^2_{.05,1} = 3.84$ $\chi^2_{.01,1} = 6.63$ $\chi^2_{.05,4} = 9.49$ $\chi^2_{.01,4} = 13.28$

จากตาราง 5 จะเห็นว่าจำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบระหว่างวิธีตรวจสอบ
แตกต่างกัน อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 โดยวิธีลอร์ด กับวิธีคิม2 วิธีราชู1 กับวิธีคิม2
วิธีราชู2 กับ วิธีคิม2 และวิธีคิม1 กับวิธีคิม2 พบจำนวนข้อสอบที่ทำหน้าที่ต่างกันแตกต่างกัน
อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นอกนั้นไม่แตกต่างกัน

ตาราง 6 ร้อยละของความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ และ
สัมประสิทธิ์สหสัมพันธ์แบบเรียงอันดับ ระหว่างวิธีตรวจสอบ

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	.608**	.954***	.689***	.911***
ESA	50.00	-	.578**	.854***	.454*
EUA	33.33	50.00	-	.583**	.929***
CSA	50.00	100.00	50.00	-	.479*
CUA	28.57	14.29	28.57	14.29	-

จากตาราง 6 ความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อยู่
ระหว่างร้อยละ 14.29 - 100 โดยที่วิธีราชู1 กับวิธีคิม1 มีความสอดคล้องกันสูงสุด (ร้อยละ
100) ส่วนวิธีราชู1 กับวิธีคิม2 และวิธีคิม1 กับวิธีคิม2 มีความสอดคล้องกันต่ำสุด (ร้อยละ
14.29)

สำหรับสัมประสิทธิ์สหสัมพันธ์ระหว่างวิธีตรวจสอบ พบว่า มีนัยสำคัญทางสถิติทุกค่า ที่มี
ค่าสูงสุดคือ ระหว่างวิธีลอร์ด กับวิธีราชู2 (.954) รองลงมาคือ ระหว่างวิธีราชู2 กับวิธีคิม2 (.929)
และวิธีลอร์ด กับวิธีคิม2 (.911)

ตาราง 7 จำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบ เมื่อกลุ่มตัวอย่าง 500 คน และ ความยาวแบบทดสอบ 20 ข้อ

ข้อที่	ค่าพารามิเตอร์				วิธีตรวจสอบ				
	a ₁	b ₁	a ₂	b ₂	LC	ESA	EUA	CSA	CUA
1	.521	.238	.473	-.295	6.15	-2.15	1.32	-0.250	0.358
2	.661	.639	.894	.316	10.54**	-2.22	2.98**	<u>-0.318</u>	<u>0.782</u>
3	.652	.932	.564	.651	4.27	-2.02	1.16	-0.239	0.306
4	.845	.623	.728	.579	0.79	-0.56	0.87	-0.041	0.211
5	.815	.493	.791	.271	3.90	-1.54	1.60	-0.128	0.370
6	.956	-.986	.772	-1.321	1.24	1.06	0.77	0.141	0.173
7	.848	-.155	.764	-.469	2.21	-0.44	1.21	-0.051	0.266
8	.567	-.456	.523	-.553	2.04	1.27	1.22	0.134	0.339
9	.320	1.365	.309	1.393	1.18	-0.57	1.02	0.026	0.314
10	.422	-.161	.424	-.049	3.87	1.58	1.46	0.195	<u>0.445</u>
11	.193	-.392	.155	-.744	0.10	-0.02	0.28	-0.018	0.076
12	.351	1.342	.279	1.671	0.33	0.03	0.30	0.058	0.103
13	.616	-.342	.464	-.524	0.48	0.68	0.36	0.080	0.083
14	.523	1.122	.391	1.381	0.08	0.08	0.13	0.028	0.039
15	.558	.911	.494	1.522	11.58**	1.46	1.53	<u>0.348</u>	0.373
16	.182	2.710	.203	2.459	1.36	-0.89	1.14	0.031	0.330
17	.637	-.258	.542	-.101	6.57	2.54	1.38	0.280	0.316
18	.639	-.740	.447	-1.022	1.32	0.73	-0.85	0.122	0.122
19	.535	-.296	.572	-.206	6.53	2.09	1.97	<u>0.304</u>	<u>0.546</u>
20	.742	-.230	.540	.125	0.01	0.01	0.01	0.000	0.000
รวม					5	5	2	3	3

จากตาราง 7 จะเห็นว่า วิธีลอร์ด ตรวจพบจำนวนข้อสอบที่ทำหน้าที่ต่างกัน 5 ข้อ วิธี ราชู1 5 ข้อ วิธีราชู2 2 ข้อ วิธีคิม1 3 ข้อ วิธี คิม2 3 ข้อ และมียู่ 2 ข้อ ที่ตรวจพบว่า ทำหน้าที่ต่างกันซ้ำกันในทุกวิธี

ตาราง 8 ร้อยละของความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ และ
สัมประสิทธิ์สหสัมพันธ์แบบเรียงอันดับ ระหว่างวิธีตรวจสอบ

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	.898 ^{***}	.934 ^{***}	.893 ^{***}	.877 ^{***}
ESA	66.67	-	.845 ^{***}	.887 ^{***}	.818 ^{***}
EUA	40.00	40.00	-	.794 ^{***}	.965 ^{***}
CSA	60.00	33.33	66.67	-	.749 ^{***}
CUA	33.33	33.33	66.67	50.00	-

จากตาราง 8 ความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบอยู่ระหว่าง
ร้อยละ 33.33 - 66.67 โดยที่วิธีลอร์ด กับวิธีราชู1 วิธีราชู2 กับวิธีคิม1 และวิธีราชู2 กับวิธีคิม2
มีความสอดคล้องกันสูงสุด (ร้อยละ 66.67) ส่วนวิธีลอร์ด กับวิธีคิม2 วิธีราชู1 กับวิธีคิม1 และวิธี
ราชู1 กับวิธีคิม2 มีความสอดคล้องกันต่ำสุด (ร้อยละ 33.33)

สำหรับสัมประสิทธิ์สหสัมพันธ์ระหว่างวิธีตรวจสอบ พบว่า มีนัยสำคัญทางสถิติทุกค่า ที่
มีค่าสูงสุดคือ ระหว่างวิธีราชู2 กับวิธีคิม2 (.965) รองลงมาคือ ระหว่างวิธีลอร์ด กับวิธีราชู2
(.934) และวิธีลอร์ด กับวิธีราชู1 (.898)

ตาราง 9 จำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบ เมื่อกลุ่มตัวอย่าง 700 คน และ
ความยาวแบบทดสอบ 20 ข้อ

ข้อที่	ค่าพารามิเตอร์				วิธีตรวจสอบ				
	a_1	b_1	a_2	b_2	LC	ESA	EUA	CSA	CUA
1	.619	.204	.517	-.406	17.92 ^{***}	-3.69 ^{***}	2.77 ^{**}	<u>-0.417</u>	<u>0.630</u>
2	.944	.498	.768	.300	7.09 [*]	0.27	-2.64 ^{**}	-0.033	<u>0.575</u>
3	.705	.958	.598	.749	5.39 [*]	0.80	-2.23 [*]	-0.042	<u>0.511</u>
4	.885	.683	.712	.453	8.06 [*]	0.31	-2.75 ^{**}	-0.060	<u>0.589</u>
5	.769	.717	.870	.119	16.36 ^{**}	-3.59 ^{***}	-3.62 ^{***}	<u>-0.387</u>	<u>0.463</u>
6	.878	-1.036	1.007	-1.117	2.46	-1.47 [*]	-1.06	-0.177	0.198
7	.790	-.186	.874	-.479	5.32	-2.30 [*]	-1.49	-0.230	0.263
8	.664	-.484	.591	-.476	5.85	0.21	-2.10 [*]	0.088	<u>0.465</u>
9	.265	1.708	.413	1.131	1.36	-0.31	0.87	0.076	0.206
10	.519	-.150	.504	-.202	2.65	0.28	-1.52	0.046	0.337
11	.293	-.268	.183	-.545	6.42 [*]	-0.71	-2.15 [*]	-0.010	<u>0.539</u>
12	.253	1.938	.320	1.612	0.81	0.46	0.46	0.110	0.110
13	.472	-.509	.492	-.551	1.19	-0.20	-1.05	0.018	0.235
14	.408	1.581	.451	1.219	0.60	0.33	-0.67	-0.028	0.147
15	.538	.872	.545	1.080	9.31 ^{**}	2.91 ^{**}	-1.81	<u>0.352</u>	<u>0.403</u>
16	.166	2.978	.230	1.828	0.54	-0.57	0.34	-0.093	0.095
17	.568	-.062	.501	-.234	4.33	-0.55	-2.02 [*]	-0.046	<u>0.463</u>
18	.732	-.537	.640	-.722	5.37	-1.53	-2.22 [*]	-0.100	<u>0.497</u>
19	.606	-.181	.573	-.313	3.04	-0.45	-1.70	-0.024	0.382
20	.589	-.065	.744	-.067	0.00	-0.00	0.00	-0.000	0.000
รวม					6	4	9	3	10

จากตาราง 9 จะเห็นว่า วิธีลอร์ดตรวจพบจำนวนข้อสอบที่ทำหน้าที่ต่างกัน 6 ข้อ วิธี
ราซ 1 4 ข้อ วิธีราซ 2 9 ข้อ วิธีคิม 1 3 ข้อ วิธีคิม 2 10 ข้อ และมีอยู่ 2 ข้อที่ตรวจพบว่า
ทำหน้าที่ต่างกัน ซ้ำกันในทุกวิธี

ตาราง 10 ผลการทดสอบความแตกต่างของจำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบระหว่างวิธี
ตรวจสอบ เมื่อกลุ่มตัวอย่าง 700 คน และความยาวแบบทดสอบ 20 ข้อ

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA	χ^2 รวม
LC		1.00	1.80	3.00	4.00	14.88**
ESA			2.78	1.00	4.50	
EUA				4.50	1.00	
CSA					12.25**	
CUA						

จากตาราง 10 จะเห็นว่าจำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบระหว่างวิธีตรวจสอบ
แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 โดยวิธีลอร์ด กับวิธีคิม2 วิธีราชู1 กับวิธีคิม2
และวิธีราชู2 กับวิธีคิม1 พบจำนวนข้อสอบที่ทำหน้าที่ต่างกันแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่
ระดับ .05 ส่วนวิธีคิม1 กับ วิธีคิม2 พบจำนวนข้อสอบที่ทำหน้าที่ต่างกันแตกต่างกันอย่างมีนัย
สำคัญทางสถิติที่ระดับ .01 นอกนั้นไม่แตกต่างกัน

ตาราง 11 ร้อยละของความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ และ
สัมประสิทธิ์สหสัมพันธ์แบบเรียงอันดับ ระหว่างวิธีตรวจสอบ

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	.491	.931***	.412	.888***
ESA	42.86	-	.395	.720**	.266
EUA	50.00	18.18	-	.259	.943***
CSA	50.00	75.00	20.00	-	.106
CUA	60.00	27.27	90.00	30.00	-

จากตาราง 11 ความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อยู่
ระหว่างร้อยละ 18.18 - 90.00 โดยที่วิธีราชู2 กับวิธีคิม2 มีความสอดคล้องกันสูงสุด (ร้อยละ
90.00) และวิธีราชู1 กับวิธีราชู2 มีความสอดคล้องกันต่ำสุด (ร้อยละ 18.18)

สำหรับสัมประสิทธิ์สหสัมพันธ์ระหว่างวิธีตรวจสอบ พบว่า ที่มีค่าสูงสุดคือ ระหว่างวิธี
ราชู2 กับวิธีคิม2 (.943) รองลงมาคือ ระหว่างวิธีลอร์ด กับวิธีราชู2 (.931) และวิธีลอร์ด กับ
วิธีคิม2 (.888)

ตาราง 12 จำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบ เมื่อกลุ่มตัวอย่าง 1000 คน และ
ความยาวแบบทดสอบ 20 ข้อ

ข้อที่	ค่าพารามิเตอร์				วิธีตรวจสอบ				
	a_1	b_1	a_2	b_2	LC	ESA	EUA	CSA	CUA
1	.589	.143	.530	-.309	3.68	-0.60	-1.84	-0.067	0.358
2	.845	.556	.791	.349	9.77 ^{**}	3.04 ^{**}	-2.14 [*]	<u>0.312</u>	<u>0.440</u>
3	.633	.909	.591	.741	7.50 [*]	2.73 ^{**}	-1.84	0.209	0.337
4	.700	.797	.711	.541	5.47	2.33	-1.47	0.181	0.239
5	.922	.575	.911	.218	2.82	1.32	-1.43	0.080	0.223
6	1.171	-.774	.941	-1.221	7.60 [*]	-1.84	-2.66 ^{**}	-0.151	<u>0.460</u>
7	.953	-.111	.848	-.493	5.24	-0.35	-2.17 [*]	-0.022	0.352
8	.614	-.367	.565	-.510	9.43 ^{**}	1.65	-1.89	0.186	0.345
9	.309	1.485	.364	1.300	6.43 [*]	1.45	1.45	0.248	0.248
10	.540	-.028	.516	-.262	4.14	1.25	-1.46	0.111	0.272
11	.131	-.490	.158	-.714	0.03	0.19	0.09	0.019	0.021
12	.296	1.741	.313	1.500	2.09	1.13	0.63	0.136	0.145
13	.540	-.339	.465	-.626	5.23	0.27	-1.94	0.070	0.388
14	.450	1.480	.464	1.238	3.30	1.61	-1.04	0.167	0.201
15	.633	.882	.531	1.197	30.56 ^{***}	5.27 ^{***}	-3.35 ^{***}	<u>0.536</u>	<u>0.618</u>
16	.207	2.308	.181	2.674	2.74	1.52	-1.08	0.152	0.219
17	.601	-.118	.537	-.270	8.93 [*]	2.02	-2.02 [*]	0.182	0.378
18	.783	-.388	.565	-.802	12.39 ^{**}	-0.93	-3.18 ^{**}	0.021	<u>0.640</u>
19	.731	-.134	.529	-.300	17.90 ^{***}	1.95	-3.28 ^{**}	0.179	<u>0.653</u>
20	.624	-.392	.728	.034	0.00	0.01	0.01	0.001	0.001
รวม					9	5	7	2	5

จากตาราง 12 จะเห็นว่าวิธีลอร์ดตรวจพบจ ๑ ข้อสอบที่ทำหน้าที่ต่างกัน 9 ข้อ วิธี
ราชู 1 5 ข้อ วิธีราชู 2 7 ข้อ วิธีคิม 1 2 ข้อ วิธีคิม 2 5 ข้อ และมีอยู่ 2 ข้อที่ตรวจพบว่า
ทำหน้าที่ต่างกันซ้ำกันในทุกวิธี

ตาราง 13 ผลการทดสอบความแตกต่างของจำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบระหว่างวิธีตรวจสอบ เมื่อกลุ่มตัวอย่าง 1,000 คน และความยาวแบบทดสอบ 20 ข้อ

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA	χ^2 รวม
LC		2.67	1.00	7.00	4.00	11.83
ESA			0.67	3.00	0.00	
EUA				5.00	2.00	
CSA					3.00	
CUA						

จากตาราง 13 จะเห็นว่าจำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบระหว่างวิธีตรวจสอบแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 โดยวิธีลอร์ด กับวิธีคิม2 และวิธีราชู2 กับวิธีคิม1 พบจำนวนข้อสอบที่ทำหน้าที่ต่างกันแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ส่วนวิธีลอร์ด กับวิธีคิม1 พบจำนวนข้อสอบที่ทำหน้าที่ต่างกันแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 นอกนั้นไม่แตกต่างกัน

ตาราง 14 ร้อยละของความสอดคล้องในการตรวจสอบ การทำหน้าที่ต่างกันของข้อสอบ และสัมประสิทธิ์สหสัมพันธ์แบบเรียงอันดับ ระหว่างวิธีตรวจสอบ

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	.675	.898	.612	.893
ESA	40.00	-	.460	.902	.417
EUA	60.00	33.33	-	.330	.980
CSA	22.22	40.00	28.57	-	.309
CUA	55.56	25.00	71.43	40.00	-

จากตาราง 14 ความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อยู่ระหว่างร้อยละ 22.22 - 71.43 โดยที่วิธีราชู2 กับวิธีคิม2 มีความสอดคล้องกันสูงสุด (ร้อยละ 71.43) และวิธีลอร์ด กับวิธีคิม1 มีความสอดคล้องกันต่ำสุด (ร้อยละ 22.22) สำหรับสัมประสิทธิ์สหสัมพันธ์ระหว่างวิธีตรวจสอบ พบว่า ที่มีค่าสูงสุดคือระหว่างวิธีราชู2 กับวิธีคิม2 (.980) รองลงมาคือ ระหว่างวิธีราชู1 กับวิธีคิม1 (.902) และวิธีลอร์ด กับวิธีราชู2 (.898)

ตาราง 15 จำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบ เมื่อกลุ่มตัวอย่าง 300 คน และ
ความยาวแบบทดสอบ 30 ข้อ

ข้อที่	ค่าพารามิเตอร์				วิธีตรวจสอบ				
	a_1	b_1	a_2	b_2	LC	ESA	EUA	CSA	CUA
1	.632	.279	.519	-.420	22.47 ^{***}	-4.54 ^{***}	-2.66 ^{**}	<u>-0.762</u>	<u>0.825</u>
2	.640	.016	.736	-.364	14.83 ^{***}	-3.84 ^{***}	-3.82 ^{***}	<u>-0.592</u>	<u>0.592</u>
3	.792	.157	.628	-.103	12.07 ^{**}	-2.86 ^{**}	-2.36 [*]	<u>-0.434</u>	<u>0.684</u>
4	.914	.567	.809	.339	10.63 ^{**}	-2.42 [*]	-2.08 [*]	<u>-0.393</u>	<u>0.532</u>
5	.688	.641	.593	.718	2.96 [*]	-0.03	-1.48	-0.091	<u>0.470</u>
6	.790	.875	.633	.737	7.49 [*]	-0.98	-1.88	-0.292	<u>0.594</u>
7	.851	1.368	.740	.954	15.33 ^{***}	-1.70	-2.21 [*]	<u>-0.496</u>	<u>0.578</u>
8	.894	-.859	.966	.604	7.20 [*]	-2.07 [*]	-2.47 [*]	<u>-0.356</u>	<u>0.408</u>
9	1.026	-.608	.844	-.936	15.97 ^{***}	-3.99 ^{***}	-2.85 ^{**}	<u>-0.622</u>	<u>0.738</u>
10	.741	-.002	.631	-.337	12.70 ^{**}	-3.43 ^{***}	-2.21 [*]	<u>-0.509</u>	<u>0.646</u>
11	.302	-.250	.539	-.322	4.59	-1.34	1.29	<u>-0.318</u>	<u>0.481</u>
12	.364	1.211	.449	1.016	0.50	-0.50	0.50	-0.127	0.127
13	.396	-.161	.471	.105	0.03	0.14	-0.14	0.027	0.045
14	.578	-.112	.448	-.030	3.90	-0.72	-1.73	0.086	<u>0.586</u>
15	.359	-.172	.186	-.748	7.74 [*]	-1.82	-2.10 [*]	-0.236	<u>0.750</u>
16	.514	-.226	.408	-.138	3.00	-0.73	-1.63	-0.069	<u>0.532</u>
17	.316	1.494	.343	1.642	0.19	0.43	-0.40	0.057	0.128
18	.446	-.674	.546	-.371	4.26	-0.12	-0.10	-0.021	0.021
19	.615	-1.386	.603	-1.498	4.32	-1.84	-1.27	<u>-0.370</u>	<u>0.423</u>
20	.467	1.435	.477	1.118	4.44	-0.76	-0.98	<u>-0.308</u>	0.324
21	.690	-1.384	.590	-1.568	5.09	-2.10 [*]	-1.67	<u>-0.357</u>	<u>0.537</u>
22	.594	.764	.589	1.415	6.53 [*]	2.13 [*]	-1.48	<u>0.456</u>	<u>0.482</u>
23	.216	1.962	.481	.502	7.14 [*]	-2.48 [*]	1.75	<u>-0.547</u>	<u>0.724</u>
24	.421	-.092	.338	-.538	2.41	-2.48 [*]	-1.58	<u>-0.390</u>	<u>0.556</u>
25	.318	2.572	.234	4.423	12.68 ^{**}	1.52	-1.37	<u>0.369</u>	<u>0.488</u>
26	.777	-.524	.490	-.968	4.47	-3.31 [*]	-2.54 [*]	<u>-0.460</u>	<u>0.939</u>
27	.603	-.109	.686	-.216	3.74	-2.11 [*]	-1.80	<u>-0.323</u>	0.323
28	.553	.587	.570	.425	4.55	-1.32	-1.04	-0.271	0.314
29	.652	-.166	.601	-.218	0.00	-1.79	-1.39	-0.245	<u>0.433</u>
30	.541	-.116	.665	.109	0.01	0.00	0.00	0.000	0.000
รวม					13	13	10	18	22

จากตาราง 15 จะเห็นว่าวิธีลอร์ดตรวจพบจำนวนข้อสอบที่ทำหน้าที่ต่างกัน 13 ข้อ วิธีราชา1 13 ข้อ วิธีราชา2 10 ข้อ วิธีคิม1 18 ข้อ วิธีคิม2 22 ข้อ และมีอยู่ 8 ข้อ ที่ตรวจพบว่าทำหน้าที่ต่างกัน ซ้ำกันในทุกวิธี

ตาราง 16 ผลการทดสอบความแตกต่างของจำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบระหว่างวิธีตรวจสอบ เมื่อกลุ่มตัวอย่าง 300 คน และความยาวแบบทดสอบ 30 ข้อ

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA	χ^2 รวม
LC		0.00	1.50	2.78	9.00**	22.70**
ESA			1.29	5.00*	7.36**	
EUA				6.40*	12.00**	
CSA					2.00	
CUA						

จากตาราง 16 จะเห็นว่าจำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบระหว่างวิธีตรวจสอบ แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 โดยวิธีลอร์ดกับวิธีคิม2 วิธีราชา1 กับวิธีคิม2 และวิธีราชา2 กับวิธีคิม2 พบจำนวนข้อสอบที่ทำหน้าที่ต่างกันแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ส่วนวิธีราชา1 กับวิธีคิม1 และวิธีราชา2 กับวิธีคิม1 พบจำนวนข้อสอบที่ทำหน้าที่ต่างกันแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นอกนั้นไม่แตกต่างกัน

ตาราง 17 ร้อยละของความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ และสัมประสิทธิ์สหสัมพันธ์แบบเรียงอันดับ ระหว่างวิธีตรวจสอบ

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	.855***	.906***	.845***	.816***
ESA	62.50	-	.796***	.919***	.724***
EUA	76.92	53.33	-	.745***	.832***
CSA	57.89	72.22	47.37	-	.698***
CUA	59.91	54.54	45.45	68.18	-

จากตาราง 17 ความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อยู่ระหว่างร้อยละ 45.45 - 76.92 โดยวิธีลอร์ด กับวิธีราชา2 มีความสอดคล้องกันสูงสุด (ร้อยละ 76.92) และวิธีราชา2 กับวิธีคิม2 มีความสอดคล้องกันต่ำสุด (ร้อยละ 45.45)

สำหรับสัมประสิทธิ์สหสัมพันธ์ระหว่างวิธีตรวจสอบ พบว่า มีนัยสำคัญทางสถิติทุกค่า ที่มีค่าสูงสุดคือระหว่างวิธีราชา1 กับวิธีคิม1 (.919) รองลงมาคือระหว่างวิธีลอร์ด กับวิธีราชา2 (.906) และวิธีลอร์ด กับวิธีราชา1 (.855)

ตาราง 18 จำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบ เมื่อกลุ่มตัวอย่าง 500 คน
และความยาวแบบทดสอบ 30 ข้อ

ข้อที่	ค่าพารามิเตอร์				วิธีตรวจสอบ				
	a_1	b_1	a_2	b_2	LC	ESA	EUA	CSA	CUA
1	.647	.180	.491	-.300	19.86 ^{***}	-4.21 ^{***}	-2.43 [*]	<u>-0.517</u>	<u>0.624</u>
2	.688	-.011	.618	-.437	19.57 ^{***}	-4.42 ^{***}	-2.84 ^{**}	<u>-0.519</u>	<u>0.524</u>
3	.644	.181	.536	-.032	8.72 [*]	-2.61 ^{**}	-1.72 [*]	<u>-0.317</u>	<u>0.428</u>
4	.914	.453	.748	.334	10.11 ^{**}	-2.30 [*]	-2.05 [*]	-0.272	<u>0.449</u>
5	.615	.662	.588	.632	2.42	-0.92	-0.81	-0.156	0.199
6	.729	.893	.596	.616	15.50 ^{***}	-2.46 [*]	-2.02 [*]	<u>-0.407</u>	<u>0.483</u>
7	.865	1.258	.680	.955	23.11 ^{***}	-2.16 [*]	-2.55 [*]	<u>-0.448</u>	<u>0.526</u>
8	.953	-.761	.917	.619	8.22 [*]	-2.18 [*]	-2.44 [*]	-0.269	0.274
9	.850	-.770	.792	-1.164	19.09 ^{***}	-3.89 ^{***}	-3.35 ^{***}	<u>-0.549</u>	<u>0.549</u>
10	.818	-.089	.670	-.526	22.51 ^{***}	-4.74 ^{***}	-2.92 ^{**}	<u>-0.545</u>	<u>0.595</u>
11	.353	-.071	.482	-.328	6.88 [*]	-2.25 [*]	1.25	<u>-0.335</u>	0.384
12	.275	1.559	.304	1.402	0.42	-0.46	0.25	-0.095	0.095
13	.344	.002	.398	.053	0.40	-0.56	0.35	0.078	0.102
14	.478	-.204	.440	-.061	0.60	-0.21	-0.75	0.012	0.199
15	.289	-.393	.160	-.737	4.77	-1.28	-1.81	-0.075	<u>0.513</u>
16	.444	-.216	.448	-.119	0.29	-0.48	-0.35	-0.056	0.092
17	.260	1.920	.306	1.526	1.11	-0.79	-0.43	-0.159	0.159
18	.504	-.516	.520	-.493	1.19	-1.06	-0.61	-0.134	0.134
19	.516	-1.574	.611	-1.609	6.34 [*]	-1.07	2.20 [*]	<u>-0.309</u>	0.309
20	.459	1.484	.457	1.204	5.28 [*]	-1.13	-1.89	-0.291	0.291
21	.611	-1.382	.524	-1.878	7.86 [*]	-2.42 [*]	-1.74	<u>-0.442</u>	<u>0.477</u>
22	.594	.730	.480	1.547	6.95 [*]	2.57 [*]	-1.72	<u>0.437</u>	<u>0.518</u>
23	.212	1.855	.406	.559	7.85 [*]	-2.75 ^{**}	1.89	<u>-0.446</u>	<u>0.659</u>
24	.361	-.197	.367	-.392	3.10	-1.75	-0.94	-0.233	0.233
25	.301	2.814	.193	4.799	2.56	1.44	-1.42	0.164	0.399
26	.760	-.531	.493	-.961	12.98 ^{**}	-3.46 ^{***}	-2.61 ^{**}	<u>-0.386</u>	<u>0.738</u>
27	.624	-.133	.673	-.359	9.73 ^{**}	-3.03 ^{***}	3.30 ^{**}	<u>-0.365</u>	0.365
28	.576	.527	.542	.498	2.48	-0.97	-0.86	-0.157	0.215
29	.617	-.209	.575	-.219	2.37	-1.39	-0.94	-0.150	0.229
30	.532	-.064	.573	.107	0.00	0.00	0.00	0.000	0.000
รวม					16	15	11	14	13

จากตาราง 18 จะเห็นว่าวิธีลอร์ดตรวจพบจำนวนข้อสอบที่ทำหน้าที่ต่างกัน 16 ข้อ วิธีราชู1 15 ข้อ วิธีราชู2 11 ข้อ วิธีคิม1 14 ข้อ วิธีคิม2 13 ข้อ และมีอยู่ 7 ข้อ ที่ตรวจพบว่าทำหน้าที่ต่างกันซ้ำกันในทุกวิธี

ตาราง 19 ร้อยละของความสอดคล้องในการตรวจสอบ การทำหน้าที่ต่างกันของข้อสอบ และสัมประสิทธิ์สหสัมพันธ์แบบเรียงอันดับ ระหว่างวิธีตรวจสอบ

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	.903 ^{***}	.923 ^{***}	.902 ^{***}	.883 ^{***}
ESA	93.75	-	.845 ^{***}	.903 ^{***}	.885 ^{***}
EUA	68.75	62.50	-	.819 ^{***}	.826 ^{***}
CSA	87.50	81.25	56.25	-	.856 ^{***}
CUA	70.59	75.00	50.00	68.75	-

จากตาราง 19 ความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อยู่ระหว่างร้อยละ 50.00 - 93.75 โดยที่วิธีลอร์ด กับวิธีราชู1 มีความสอดคล้องกันสูงสุด (ร้อยละ 93.75) และวิธีราชู2 กับวิธีคิม2 มีความสอดคล้องกันต่ำสุด (ร้อยละ 50.00)

สำหรับสัมประสิทธิ์สหสัมพันธ์ระหว่างวิธีตรวจสอบ พบว่า มีนัยสำคัญทางสถิติทุกค่า ที่มีค่าสูงสุดคือ ระหว่างวิธีลอร์ด กับวิธีราชู2 (.923) รองลงมาคือระหว่างและวิธีราชู1 กับวิธีคิม1 (.903) วิธีลอร์ด กับวิธีราชู1 (.903)

ตาราง 20 จำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบ เมื่อกลุ่มตัวอย่าง 700 คน และ
ความยาวแบบทดสอบ 30 ข้อ

ข้อที่	ค่าพารามิเตอร์				วิธีตรวจสอบ				
	a_1	b_1	a_2	b_2	LC	ESA	EUA	CSA	CUA
1	.568	.257	.572	-.302	0.43	-0.19	0.62	-0.010	0.136
2	.611	-.014	.653	-.427	1.50	0.92	1.08	0.095	0.230
3	.732	.225	.515	-.227	3.35	0.67	-1.60	0.042	0.350
4	.854	.651	.799	.135	0.06	-0.21	0.20	-0.015	0.036
5	.622	.758	.648	.435	3.56	1.14	1.06	0.161	0.215
6	.703	.909	.623	.487	0.11	0.33	-0.21	0.029	0.043
7	.856	1.235	.681	1.022	1.92	1.36	-0.99	0.104	0.203
8	.969	.829	.788	.555	3.31	1.81	-1.18	0.152	0.191
9	.785	-.625	.656	-1.399	0.79	-0.89	-0.70	-0.108	0.145
10	.642	-.091	.654	-.734	1.44	-0.55	0.80	-0.064	0.161
11	.279	-.379	.372	-.414	6.60	2.39	1.68	0.280	<u>0.469</u>
12	.412	1.082	.295	1.713	11.84	2.89	-1.58	<u>0.425</u>	<u>0.446</u>
13	.420	.054	.305	.236	14.08	3.70	-1.52	<u>0.385</u>	<u>0.410</u>
14	.532	-.089	.394	-.404	4.82	1.53	-1.25	0.168	0.296
15	.175	-.147	.141	-.847	0.16	-0.26	-0.37	-0.029	0.079
16	.521	-.197	.397	-.138	18.72	3.98	-1.99	<u>0.421</u>	<u>0.429</u>
17	.287	1.690	.271	1.395	0.35	0.21	0.19	0.070	0.070
18	.530	-.430	.505	-.626	7.39	2.55	2.25	0.288	0.288
19	.588	-1.224	.471	-2.049	0.53	-0.46	-0.66	-0.037	0.164
20	.455	1.546	.398	1.260	0.18	0.39	-0.25	0.045	0.061
21	.568	-1.409	.363	2.418	4.40	0.71	-1.46	0.026	<u>0.417</u>
22	.682	.883	.511	1.324	27.31	4.42	-3.11	<u>0.579</u>	<u>0.585</u>
23	.278	1.385	.252	.945	0.00	-0.04	-0.02	-0.009	0.009
24	.545	-.169	.337	-.759	4.45	-0.07	-1.78	0.019	<u>0.485</u>
25	.259	3.531	.205	3.849	0.20	0.44	-0.44	0.011	0.101
26	.700	-.401	.510	-.962	3.06	0.22	-1.30	0.058	0.302
27	.693	-.012	.625	-.495	0.44	0.62	-0.64	0.061	0.061
28	.497	.581	.343	.862	12.63	3.47	-1.77	<u>0.375</u>	<u>0.456</u>
29	.694	-.155	.464	-.237	20.02	3.65	-2.32	<u>0.362</u>	<u>0.506</u>
30	.590	.356	.537	.168	0.00	0.00	0.00	0.000	0.000
รวม					8	8	4	6	9

จากตาราง 20 จะเห็นว่าวิธีลอร์ดตรวจพบจำนวนข้อสอบที่ทำหน้าที่ต่างกัน 8 ข้อ วิธีราชู1 8 ข้อ วิธีราชู2 4 ข้อ วิธีคิม1 6 ข้อ วิธีคิม2 9 ข้อ และมีอยู่ 3 ข้อ ที่ตรวจพบว่าทำหน้าที่ต่างกันซ้ำกันในทุกวิธี

ตาราง 21 ผลการทดสอบความแตกต่างของจำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบระหว่างวิธีตรวจสอบ เมื่อกลุ่มตัวอย่าง 700 คน และความยาวแบบทดสอบ 30 ข้อ

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA	χ^2 รวม
LC		0.00	4.00	2.00	0.33	10.00
ESA			5.33	2.00	0.33	
EUA				1.00	3.57	
CSA					3.00	
CUA						

จากตาราง 21 จะเห็นว่าจำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบระหว่างวิธีตรวจสอบ แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 โดยวิธีลอร์ด กับวิธีราชู2 และวิธีราชู1 กับวิธีราชู2 พบจำนวนข้อสอบที่ทำหน้าที่ต่างกันแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นอกนั้นไม่แตกต่างกัน

ตาราง 22 ร้อยละของความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ และสัมประสิทธิ์สหสัมพันธ์แบบเรียงอันดับ ระหว่างวิธีตรวจสอบ

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	.856 ***	.957 ***	.831 ***	.944 ***
ESA	100.00	-	.768 ***	.922 ***	.710 ***
EUA	40.00	50.00	-	.714 ***	.954 ***
CSA	60.00	75.00	42.86	-	.675 ***
CUA	77.78	70.00	30.00	66.67	-

จากตาราง 22 ความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อยู่ระหว่างร้อยละ 30.00 - 100.00 โดยวิธีลอร์ด กับวิธีราชู1 มีความสอดคล้องกันสูงสุด (ร้อยละ 100.00) และวิธีราชู2 กับวิธีคิม2 มีความสอดคล้องกันต่ำสุด (ร้อยละ 30.00)

สำหรับสัมประสิทธิ์สหสัมพันธ์ระหว่างวิธีตรวจสอบ พบว่า มีนัยสำคัญทางสถิติทุกค่า ที่มีค่าสูงสุดคือ ระหว่างวิธีลอร์ด กับวิธีราชู2 (.957) รองลงมา คือ ระหว่างวิธีราชู2 กับวิธีคิม2 (.954) และวิธีลอร์ด กับวิธีคิม2 (.944)

ตาราง 23 จำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบ เมื่อกลุ่มตัวอย่าง 1,000 คน และ
ความยาวแบบทดสอบ 30 ข้อ

ข้อที่	ค่าพารามิเตอร์				วิธีตรวจสอบ				
	a_1	b_1	a_2	b_2	LC	ESA	EUA	CSA	CUA
1	.621	.205	.519	-.239	10.66**	-3.08**	-1.79	-0.268	0.304
2	.616	-.020	.626	-.318	3.77	-1.81	1.68	-0.152	0.152
3	.738	.191	.662	-.106	4.75	-2.07	-1.34	-0.164	0.181
4	.890	.564	.722	.379	4.52	-0.96	-1.61	-0.097	0.252
5	.607	.706	.622	.542	0.32	-0.56	0.40	-0.041	0.070
6	.676	.914	.614	.592	7.09	-1.91	-1.88	-0.208	0.208
7	.848	1.147	.793	.902	4.32	-1.36	-2.06	-0.149	0.149
8	1.029	.765	.909	.685	0.69	0.20	-0.81	0.003	0.122
9	.782	-.672	.917	-.853	4.78	-0.52	1.57	-0.073	0.234
10	.695	-.120	.699	-.390	3.17	-1.62	1.77	-0.133	0.133
11	.313	-.195	.412	-.409	3.87	-0.57	1.60	-0.089	0.344
12	.346	1.198	.310	1.466	2.09	1.20	-0.64	0.144	0.151
13	.388	.123	.320	.197	1.94	1.14	-1.01	0.095	0.207
14	.464	-.086	.601	-.055	5.97	1.30	2.06	0.111	0.383
15	.174	-.184	.142	-.887	2.21	-1.48	-0.65	-0.151	0.162
16	.466	-.167	.563	-.134	3.47	1.30	1.52	0.106	0.290
17	.293	1.700	.257	1.678	0.53	0.13	-0.48	0.046	0.101
18	.535	-.398	.582	-.435	0.91	0.77	0.80	0.054	0.148
19	.553	-1.318	.645	-1.435	3.14	0.14	1.22	-0.060	0.222
20	.398	1.613	.502	1.171	2.69	-1.47	1.54	-0.088	0.313
21	.529	-1.449	.539	-1.547	0.07	0.22	0.27	0.012	0.052
22	.620	.979	.696	.907	2.37	0.15	1.08	0.061	0.182
23	.237	1.510	.341	.832	3.39	-1.77	1.72	-0.104	0.382
24	.473	-.051	.440	-.662	13.22**	-3.55***	-3.23**	<u>-0.344</u>	<u>0.402</u>
25	.197	4.653	.278	2.923	2.40	-1.55	1.36	-0.182	0.305
26	.674	-.423	.666	-.609	0.47	-0.60	0.66	-0.053	0.053
27	.670	-.007	.698	-.332	6.63	-2.43	2.03	-0.195	0.195
28	.495	.471	.482	.526	1.68	1.17	-1.17	0.115	0.115
29	.635	-.224	.648	-.107	6.15	2.47	2.24	0.197	0.197
30	.597	.259	.584	.150	0.00	-0.00	-0.00	-0.000	0.000
รวม					6	5	5	1	1

จากตาราง 23 จะเห็นว่าวิธีลอร์ดตรวจพบจำนวนข้อสอบที่ทำหน้าที่ต่างกัน 6 ข้อ วิธีราชู 5 ข้อ วิธีราชู 2 5 ข้อ วิธีคิม 1 1 ข้อ วิธีคิม 2 1 ข้อ และมีอยู่ 1 ข้อ ที่ตรวจพบว่าทำหน้าที่ต่างกันซ้ำกันในทุกวิธี

ตาราง 24 ผลการทดสอบความแตกต่างของจำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบระหว่างวิธีตรวจสอบ เมื่อกลุ่มตัวอย่าง 1,000 คน และความยาวแบบทดสอบ 30 ข้อ

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA	χ^2 รวม
LC		0.33	0.33	5.00	5.00	10.67
ESA			0.00	4.00	4.00	
EUA				4.00	4.00	
CSA					0.00	
CUA						

จากตาราง 24 จะเห็นว่าจำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบระหว่างวิธีตรวจสอบแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 โดยวิธีลอร์ด กับวิธีคิม 1 วิธีลอร์ด กับวิธีคิม 2 วิธีราชู 1 กับวิธีคิม 1 วิธีราชู 1 กับวิธีคิม 2 วิธีราชู 2 กับวิธีคิม 1 และวิธีราชู 2 กับวิธีคิม 2 พบจำนวนข้อสอบที่ทำหน้าที่ต่างกันแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นอกนั้นไม่แตกต่างกัน

ตาราง 25 ร้อยละของความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ และสัมประสิทธิ์สหสัมพันธ์แบบเรียงอันดับ ระหว่างวิธีตรวจสอบ

วิธี	LC	ESA	EUA	CSA	CUA
LC	-	.726 ***	.915 ***	.776 ***	.729 ***
ESA	66.67	-	.721 ***	.918 ***	.461 **
EUA	42.86	42.86	-	.721 ***	.664 ***
CSA	20.00	20.00	20.00	-	.485 **
CUA	20.00	20.00	20.00	100.00	-

จากตาราง 25 ความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อยู่ระหว่างร้อยละ 20.00 - 100.00 โดยที่วิธีคิม 1 กับวิธีคิม 2 มีความสอดคล้องกันสูงสุด (ร้อยละ 100.00) ส่วนวิธีลอร์ด และวิธีราชู 1 วิธีราชู 2 เมื่อจับคู่กับวิธีคิม 1 วิธีคิม 2 พบว่า มีความสอดคล้องกันต่ำสุด (ร้อยละ 20.00)

สำหรับสัมประสิทธิ์สหสัมพันธ์ระหว่างวิธีตรวจสอบ พบว่า มีนัยสำคัญทางสถิติทุกค่า ที่สูงสุดคือระหว่างวิธีราชู 1 กับวิธีคิม 1 (.918) รองลงมา คือ ระหว่าง วิธีลอร์ด กับวิธีราชู 2 (.915) และวิธีลอร์ด กับวิธีคิม 1 (.776)

ตาราง 26 จำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบ เมื่อกลุ่มตัวอย่าง 300 คน และ
ความยาวแบบทดสอบ 40 ข้อ

ข้อที่	ค่าพารามิเตอร์				วิธีตรวจสอบ				
	a_1	b_1	a_2	b_2	LC	ESA	EUA	CSA	CUA
1	.558	.022	.529	-.313	2.69	-0.35	1.51	-0.052	<u>0.481</u>
2	.566	.220	.447	-.113	1.02	-0.71	0.77	-0.095	0.247
3	.849	.348	.692	.389	1.70	0.62	1.06	0.089	0.260
4	.526	.720	.523	.666	2.53	-0.52	1.54	-0.001	<u>0.532</u>
5	.760	1.241	.549	1.224	1.75	-1.15	0.81	-0.206	0.214
6	.894	.616	.878	.525	3.13	-0.73	1.72	-0.068	<u>0.446</u>
7	.905	.854	.788	.716	2.85	-1.51	1.43	-0.171	0.347
8	.745	-.740	.503	-1.826	3.41	-1.21	1.21	0.272	0.272
9	.690	-.665	.592	-.717	4.60	2.14	1.44	0.294	<u>0.417</u>
10	.408	.310	.286	.705	2.09	1.25	1.02	0.235	0.235
11	.431	.030	.335	-.328	0.50	-0.35	0.58	-0.057	0.206
12	.519	-.047	.388	.118	2.70	1.56	0.81	0.240	0.255
13	.796	-.712	.915	-.972	5.32	1.62	2.27	0.181	<u>0.650</u>
14	.622	.394	.516	.466	1.86	0.49	1.05	0.104	0.307
15	.220	2.116	.203	3.050	3.43	0.09	0.89	<u>0.304</u>	0.358
16	.509	1.015	.424	1.280	1.38	0.02	0.87	0.091	0.296
17	.664	-.235	.583	-.107	5.91	2.25	1.61	0.297	<u>0.441</u>
18	.332	1.420	.325	1.068	2.00	-1.31	1.25	-0.157	<u>0.472</u>
19	.213	.753	.274	.793	4.10	-0.13	1.73	0.167	<u>0.664</u>
20	.399	-.119	.365	-.690	2.55	-0.71	1.17	-0.151	<u>0.421</u>
21	.665	.376	.384	.485	0.55	0.62	-0.62	0.075	0.200
22	.695	-.258	.436	-.343	1.77	1.22	-0.97	0.177	0.178
23	.436	.902	.257	1.206	0.16	0.20	-0.38	-0.002	0.140
24	.529	-1.127	.518	-1.760	3.18	0.44	1.47	-0.032	<u>0.498</u>
25	.497	.636	.584	.498	6.03	-0.67	2.28	-0.010	<u>0.752</u>
26	.516	1.107	.505	1.104	2.32	-0.80	1.50	-0.033	<u>0.503</u>
27	.506	1.082	.554	.551	6.15	-2.21	2.12	<u>-0.322</u>	<u>0.711</u>
28	.574	-.628	.757	-.586	10.18	2.35	2.77	<u>0.306</u>	<u>0.911</u>
29	.362	-.676	.517	-.373	10.02	2.35	2.49	<u>0.345</u>	<u>1.027</u>
30	.503	-.065	.671	-.311	8.38	0.17	2.68	0.014	<u>0.919</u>

ตาราง 26 (ต่อ)

ข้อที่	ค่าพารามิเตอร์				วิธีตรวจสอบ				
	a ₁	b ₁	a ₂	b ₂	LC	ESA	EUA	CSA	CUA
31	.646	-.148	.710	-.391	5.26	0.40	2.19	0.039	<u>0.605</u>
32	.323	2.830	.229	3.718	0.04	-0.20	0.19	-0.028	0.065
33	.837	-.636	.891	-.581	10.84**	3.10	2.58	<u>0.376</u>	<u>0.630</u>
34	.595	-.252	.556	-.145	5.64	2.03	1.68	0.277	<u>0.508</u>
35	.594	.624	.640	.406	4.98	-1.05	2.14	-0.095	<u>0.639</u>
36	.509	.606	.678	.425	8.67	-0.79	2.68**	-0.028	<u>0.905</u>
37	.375	.379	.322	.691	2.81	0.94	0.98	0.228	0.358
38	.234	1.303	.302	.616	4.35	-1.54	1.82	-0.149	<u>0.709</u>
39	.509	.042	.611	-.282	6.39	-0.35	2.29	-0.048	<u>0.789</u>
40	.521	.486	.346	.476	0.00	0.00	-0.00	0.000	0.000
รวม					8	7	11	5	23

จากตาราง 26 จะเห็นว่าวิธีลอร์ด (LC) ตรวจพบจำนวนข้อสอบที่ทำหน้าที่ต่างกัน 8 ข้อ วิธีราชู1 (ESA) 7 ข้อ วิธีราชู2 (EUA) 11 ข้อ วิธีคิม1 (CSA) 5 ข้อ วิธีคิม2 (CUA) 23 ข้อ และมีอยู่ 5 ข้อ ที่ตรวจพบว่าทำหน้าที่ต่างกันซ้ำกันในทุกวิธี

ตาราง 27 ผลการทดสอบความแตกต่างของจำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบระหว่างวิธีตรวจสอบ เมื่อกลุ่มตัวอย่าง 300 คน และความยาวแบบทดสอบ 40 ข้อ

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA	χ^2 รวม
LC		0.14	3.00	1.80	13.07**	17.94**
ESA			1.60	1.00	14.06**	
EUA				4.50*	10.08**	
CSA					14.45**	
CUA						

จากตาราง 27 จะเห็นว่าจำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบระหว่างวิธีตรวจสอบแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 โดยวิธีลอร์ด กับวิธีคิม2 วิธีราชู1 กับวิธีคิม2 และวิธีราชู2 กับวิธีคิม2 และวิธีคิม1 กับวิธีคิม2 พบจำนวนข้อสอบที่ทำหน้าที่ต่างกันแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ส่วนวิธีราชู2 กับวิธีคิม1 พบจำนวนข้อสอบที่ทำหน้าที่ต่างกันแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นอกนั้นไม่แตกต่างกัน

ตาราง 28 ร้อยละของความสอดคล้องในการตรวจสอบ การทำหน้าที่ต่างกันของข้อสอบ และ
สัมประสิทธิ์สหสัมพันธ์แบบเรียงอันดับ ระหว่างวิธีตรวจสอบ

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	.481 ^{**}	.914 ^{***}	.388 [*]	.875 ^{***}
ESA	36.36	-	.377 [*]	.729 ^{***}	.269 ^{***}
EUA	72.73	28.57	-	.141	.952 ^{***}
CSA	44.44	50.00	33.33	-	.094
CUA	36.36	31.18	50.00	17.39	-

จากตาราง 28 ความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อยู่
ระหว่างร้อยละ 17.39 - 72.73 โดยที่วิธีลอร์ด กับวิธีราชู2 มีความสอดคล้องกันสูงสุด (ร้อยละ
72.73) ส่วนวิธีคิม1 กับวิธีคิม2 มีความสอดคล้องกันต่ำสุด (ร้อยละ 17.39)

สำหรับสัมประสิทธิ์สหสัมพันธ์ระหว่างวิธีตรวจสอบ พบว่า ที่มีค่าสูงสุด คือ ระหว่างวิธี
ราชู2 กับวิธีคิม2 (.952) รองลงมา คือ ระหว่างวิธีลอร์ด กับวิธีราชู2 (.914) และวิธีลอร์ด กับ
วิธีคิม2 (.875)

ตาราง 29 จำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบ เมื่อกลุ่มตัวอย่าง 500 คน และ
ความยาวแบบทดสอบ 40 ข้อ

ข้อที่	ค่าพารามิเตอร์				วิธีตรวจสอบ				
	a ₁	b ₁	a ₂	b ₂	LC	ESA	EUA	CSA	CUA
1	.532	.195	.568	.026	0.25	-0.22	-0.44	-0.033	0.107
2	.533	.195	.710	-.025	1.03	-0.63	0.84	-0.069	0.201
3	.764	.496	.684	.509	3.56	1.55	-1.60	0.140	0.360
4	.621	.989	.615	.749	0.90	0.03	-0.82	-0.052	0.201
5	.691	1.449	.715	.946	4.34	-1.11	-1.37	-0.234	0.238
6	.876	.791	.934	.764	2.64	1.57	-1.16	0.185	0.195
7	.807	1.075	.914	.799	0.08	-0.15	-0.18	-0.026	0.031
8	.674	-.903	.845	-.480	11.27 ^{**}	3.15 ^{**}	2.84 ^{**}	<u>0.401</u>	<u>0.401</u>
9	.669	-.500	.792	.113	31.04 ^{***}	5.49 ^{***}	5.49 ^{***}	<u>0.662</u>	<u>0.662</u>
10	.270	.438	.401	.392	1.99	0.51	0.98	0.132	0.272
11	.373	-.078	.513	-.513	4.93	-1.86	1.07	-0.279	0.307
12	.332	-.007	.298	.335	4.15	1.89	-1.14	0.248	0.315
13	.645	-.795	.778	-.789	0.05	0.07	0.23	-0.001	0.057
14	.633	.484	.701	.400	0.49	0.70	-0.43	0.082	0.089
15	.114	4.110	.154	3.157	0.36	-0.20	0.41	0.062	0.085

ตาราง 29 (ต่อ)

ข้อที่	ค่าพารามิเตอร์				วิธีตรวจสอบ				
	a_1	b_1	a_2	b_2	LC	ESA	EUA	CSA	CUA
16	.374	1.403	.521	1.261	4.59	0.54	0.98	0.236	0.275
17	.464	-.222	.585	.026	5.05	2.22	1.37	0.283	0.283
18	.244	2.038	.276	1.328	1.44	-0.62	-0.81	-0.175	0.175
19	.224	1.245	.386	.652	2.42	-0.89	1.37	-0.006	0.375
20	.345	-.454	.516	-.423	1.55	0.42	1.15	0.014	0.308
21	.440	.329	.383	.546	3.82	1.82	-1.43	0.206	0.369
22	.538	-.025	.590	.350	11.63 ^{**}	3.38 ^{***}	-3.33 ^{***}	<u>0.429</u>	<u>0.429</u>
23	.289	1.145	.323	1.468	4.84	1.57	-1.49	<u>0.318</u>	0.318
24	.528	-1.232	.728	-.896	2.31	1.52	1.13	0.187	0.278
25	.438	.617	.512	.773	4.72	1.78	1.78	0.285	0.285
26	.535	1.164	.488	1.204	1.95	1.33	-1.30	0.119	0.306
27	.345	1.384	.465	.958	0.49	-0.44	0.67	-0.003	0.168
28	.604	-.404	.674	-.568	0.65	-0.80	-0.50	-0.096	0.097
29	.408	-.190	.491	-.416	0.79	-0.81	0.44	-0.114	0.114
30	.584	.053	.505	.184	4.40	1.70	-1.57	0.196	<u>0.407</u>
31	.551	-.046	.571	-.163	0.35	-0.08	-0.59	-0.008	0.144
32	.371	2.344	.338	2.233	1.02	0.57	-0.86	-0.050	0.225
33	.828	-.530	.793	-.427	4.01	1.29	-1.34	0.171	0.279
34	.770	-.103	.707	-.068	3.27	1.18	-1.49	0.136	0.328
35	.549	.537	.533	.356	1.01	0.01	-0.94	-0.036	0.226
36	.449	.680	.612	.371	0.81	-0.61	0.85	-0.048	0.216
37	.385	.247	.449	.689	10.31 ^{**}	2.91 ^{**}	2.91 ^{**}	<u>0.438</u>	<u>0.438</u>
38	.191	1.497	.231	.956	0.28	-0.46	0.19	-0.076	0.076
39	.555	-.077	.710	.361	18.49 ^{***}	4.11 ^{***}	4.16 ^{***}	<u>0.520</u>	<u>0.520</u>
40	.322	.643	.372	.445	0.00	-0.00	0.00	-0.000	0.001
รวม					5	6	5	6	6

จากตาราง 29 จะเห็นว่าวิธีลอร์ด (LC) ตรวจพบจำนวนข้อสอบที่ทำหน้าที่ต่างกัน 5 ข้อ วิธีราชู1(ESA) 6 ข้อ วิธีราชู2 (EUA) 5 ข้อ วิธีคิม1 (CSA) 6 ข้อ วิธีคิม2 (CUA) 6 ข้อ และมีอยู่ 5 ข้อ ที่ตรวจพบว่าทำหน้าที่ต่างกันซ้ำกันในทุกวิธี

ตาราง 30 ร้อยละของความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ และ
สัมประสิทธิ์สหสัมพันธ์แบบเรียงอันดับ ระหว่างวิธีตรวจสอบ

วิธี	LC	ESA	EUA	CSA	CUA
LC	-	.879 ^{***}	.922 ^{***}	.904 ^{***}	.891 ^{***}
ESA	83.33	-	.832 ^{***}	.885 ^{***}	.793 ^{***}
EUA	100.00	83.33	-	.766 ^{***}	.950 ^{***}
CSA	83.33	71.43	83.33	-	.702 ^{***}
CUA	83.33	71.43	83.33	71.43	-

จากตาราง 30 ความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อยู่
ระหว่างร้อยละ 71.43 - 100.00 โดยที่วิธีลอร์ด กับวิธีราชู2 มีความสอดคล้องกันสูงสุด (ร้อยละ
100) ส่วนวิธีราชู1 กับวิธีคิม1 วิธีคิม2 และวิธีคิม1 กับวิธีคิม2 มีความสอดคล้องกันต่ำสุด
(ร้อยละ 71.43)

สำหรับสัมประสิทธิ์สหสัมพันธ์ระหว่างวิธีตรวจสอบ พบว่า มีนัยสำคัญทางสถิติทุกค่า ที่มี
ค่าสูงสุดคือ ระหว่างวิธีราชู2 กับวิธีคิม2 (.950) รองลงมา คือ ระหว่าง วิธีลอร์ด กับวิธีราชู2
(.922) และวิธีลอร์ด กับวิธีคิม1(.904)

ตาราง 31 จำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบ เมื่อกลุ่มตัวอย่าง 700 คน และ
ความยาวแบบทดสอบ 40 ข้อ

ข้อที่	ค่าพารามิเตอร์				วิธีตรวจสอบ				
	a ₁	b ₁	a ₂	b ₂	LC	ESA	EUA	CSA	CUA
1	.551	.147	.555	-.062	3.02	-0.69	-1.60	-0.086	0.327
2	.562	.245	.669	-.046	2.21	-1.32	-0.88	0.147	0.161
3	.653	.615	.727	.424	1.01	0.43	-1.02	0.013	0.197
4	.531	1.157	.633	.642	3.25	-1.07	-1.01	-0.187	0.188
5	.636	1.309	.759	.927	0.33	0.07	-0.48	-0.025	0.097
6	.866	.892	.946	.692	1.51	0.99	-1.20	0.074	0.222
7	.770	1.111	.998	.721	0.21	-0.40	0.40	-0.047	0.047
8	.701	-.888	.864	-.675	2.13	1.03	-1.23	0.133	0.134
9	.689	-.487	.707	.035	33.52 ^{***}	5.24 ^{***}	-3.71 ^{***}	<u>0.564</u>	<u>0.576</u>
10	.322	.438	.419	.471	2.06	1.30	1.30	0.168	0.168
11	.344	-.319	.514	-.443	1.47	-0.66	0.79	-0.104	0.185
12	.446	-.017	.307	.281	12.82 ^{**}	2.21 [*]	-2.88 ^{**}	<u>0.327</u>	<u>0.691</u>
13	.786	-.763	.820	-.845	2.63	-1.47	-1.43	-0.121	0.298
14	.694	.385	.690	.400	5.24	1.93	-1.98	0.164	0.375
15	.161	3.523	.160	3.009	1.24	0.37	-0.76	-0.113	0.154

ตาราง 31 (ต่อ)

ข้อที่	ค่าพารามิเตอร์				วิธีตรวจสอบ				
	a_1	b_1	a_2	b_2	LC	ESA	EUA	CSA	CUA
16	.485	1.109	.560	1.116	4.62	1.94	-1.27	0.252	0.260
17	.460	-.394	.577	-.054	8.74	2.86**	-2.88**	<u>0.324</u>	0.324
18	.299	1.403	.275	1.637	3.02	1.73	-1.47	0.142	0.312
19	.246	1.108	.334	.805	0.20	0.15	0.21	0.046	0.055
20	.404	-.302	.572	-.543	3.69	-1.58	0.91	-0.206	0.208
21	.572	.170	.438	.507	15.86***	3.30**	-3.12**	<u>0.319</u>	<u>0.694</u>
22	.516	-.111	.630	.074	5.64	2.35	-1.94	0.251	0.251
23	.292	1.207	.307	1.250	1.89	1.36	-0.96	0.132	0.209
24	.493	-1.234	.523	-1.150	1.38	-0.51	-1.10	0.021	0.233
25	.452	.539	.567	.602	4.82	1.98	-1.91	0.246	0.246
26	.559	.994	.466	1.270	9.99**	3.05**	-2.73**	<u>0.312</u>	<u>0.573</u>
27	.317	1.433	.402	1.266	1.79	0.91	-0.85	0.161	0.161
28	.676	-.480	.667	-.500	3.16	-0.30	-1.72	0.016	0.358
29	.488	-.250	.542	-.412	1.43	-0.98	-0.95	0.091	0.203
30	.472	.039	.583	.129	2.76	1.66	-1.12	0.180	0.180
31	.589	-.127	.621	-.208	1.70	-0.12	-1.31	-0.001	0.268
32	.184	5.195	.323	2.350	2.39	-1.47	1.16	<u>-0.312</u>	0.332
33	.850	-.508	.762	-.538	6.08	-0.52	-2.46	0.004	<u>0.483</u>
34	.661	-.089	.657	-.104	3.83	0.63	-1.88	0.070	0.356
35	.594	.461	.519	.499	6.83	1.82	-2.47	0.128	<u>0.511</u>
36	.481	.420	.587	.426	2.43	1.52	-0.92	0.172	0.172
37	.508	.315	.487	.657	12.35**	3.50***	-2.20	<u>0.375</u>	<u>0.476</u>
38	.158	1.807	.251	.705	1.88	-1.33	0.74	-0.169	0.188
39	.691	.156	.642	.285	9.42**	2.53	-2.54	0.236	<u>0.481</u>
40	.299	.813	.386	.533	0.00	-0.00	0.00	-0.000	0.000
รวม					9	9	10	7	8

จากตาราง 31 จะเห็นว่าวิธีลอร์ด (LC) ตรวจพบจำนวนข้อสอบที่ทำหน้าที่ต่างกัน 9 ข้อ วิธีราชู1 (ESA) 9 ข้อ วิธีราชู2 (EUA) 10 ข้อ วิธีคิม1 (CSA) 7 ข้อ วิธีคิม2 (CUA) 8 ข้อ และมีอยู่ 5 ข้อ ที่ตรวจพบว่าทำหน้าที่ต่างกันซ้ำกันในทุกวิธี

ตาราง 32 ร้อยละของความสอดคล้องในการตรวจสอบ การทำหน้าที่ต่างกันของข้อสอบ และสัมประสิทธิ์สหสัมพันธ์แบบเรียงอันดับ ระหว่างวิธีตรวจสอบ

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	.830 ^{***}	.885 ^{***}	.714 ^{***}	.833 ^{***}
ESA	63.64	-	.683 ^{***}	.908 ^{***}	.605 ^{***}
EUA	90.00	58.33	-	.489 ^{***}	.902 ^{***}
CSA	70.00	66.67	54.55	-	.411 ^{**}
CUA	88.89	54.55	80.00	50.00	-

จากตาราง 32 ความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อยู่ระหว่างร้อยละ 50.00 - 90.00 โดยที่วิธีลอร์ด กับวิธีราชู2 มีความสอดคล้องกันสูงสุด (ร้อยละ 90.00) และวิธีคิม1 กับวิธีคิม2 มีความสอดคล้องกันต่ำสุด (ร้อยละ 50.00)

สำหรับสัมประสิทธิ์สหสัมพันธ์ระหว่างวิธีตรวจสอบ พบว่า มีนัยสำคัญทางสถิติทุกค่า ที่มีค่าสูงสุดคือ ระหว่าง วิธีราชู1 กับวิธีคิม1 (.908) รองลงมาคือ ระหว่างวิธีราชู2 กับวิธีคิม2 (.902) และวิธีลอร์ด กับวิธีราชู2 (.885)

ตาราง 33 จำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบ เมื่อกลุ่มตัวอย่าง 1,000 คน และความยาวแบบทดสอบ 40 ข้อ

ข้อที่	ค่าพารามิเตอร์				วิธีตรวจสอบ				
	a ₁	b ₁	a ₂	b ₂	LC	ESA	EUA	CSA	CUA
1	.528	.231	.571	-.281	16.04 ^{***}	-3.99 ^{**}	3.99 ^{**}	<u>-0.355</u>	0.355
2	.558	.165	.489	-.144	5.84	-2.04 [*]	-1.52	-0.187	0.281
3	.679	.618	.698	.415	1.43	-1.04	-1.15	-0.094	0.094
4	.535	.969	.513	.794	1.34	-0.37	-0.69	-0.078	0.131
5	.648	1.207	.564	1.129	2.42	0.36	-1.31	-0.040	0.241
6	.724	.855	.787	.493	7.58 [*]	-2.57 [*]	2.51 [*]	-0.223	0.223
7	.728	1.040	.730	.879	0.63	-0.31	-0.46	-0.047	0.072
8	.620	-.898	.618	-1.476	21.35 ^{***}	-3.51 ^{***}	-3.42 ^{**}	<u>-0.442</u>	<u>0.442</u>
9	.770	-.566	.550	-.625	10.53 ^{**}	-0.14	-2.85 ^{**}	0.062	<u>0.515</u>
10	.394	.343	.385	.376	0.69	0.81	-0.51	0.068	0.101
11	.372	-.146	.417	-.468	3.86	-1.77	0.88	-0.181	0.181
12	.297	-.030	.325	.168	2.55	1.55	0.76	0.155	0.155
13	.729	-.752	.824	-.949	5.51	-1.61	1.76	-0.168	0.171
14	.567	.636	.468	.392	7.62 [*]	-1.10	-1.83	-0.168	0.338
15	.161	3.313	.119	4.301	1.20	0.95	-1.05	-0.007	0.173

ตาราง 33 (ต่อ)

ข้อที่	ค่าพารามิเตอร์				วิธีตรวจสอบ				
	a_1	b_1	a_2	b_2	LC	ESA	EUA	CSA	CUA
16	.455	1.110	.462	1.394	8.02	2.08	-1.95	0.276	0.276
17	.468	-.352	.465	-.266	1.80	1.13	-0.64	0.113	0.187
18	.223	1.947	.257	1.232	2.94	-1.29	0.61	-0.187	0.118
19	.253	.962	.332	.496	2.21	-1.44	1.10	-0.113	0.187
20	.357	-.339	.304	-.713	2.66	-1.62	-0.92	-0.151	0.228
21	.533	.250	.478	.098	2.08	-0.57	-1.19	-0.067	0.198
22	.491	-.072	.594	-.190	1.32	-0.46	1.02	-0.046	0.216
23	.388	1.182	.311	1.184	2.97	0.54	-1.45	-0.065	0.188
24	.490	-1.244	.514	-1.523	4.38	-1.34	-1.34	-0.199	0.289
25	.414	.627	.586	.411	4.55	-0.83	2.00	-0.015	0.199
26	.539	1.034	.515	1.048	0.75	0.86	-0.72	0.066	0.377
27	.367	1.145	.516	.709	3.91	-1.63	1.80	-0.093	0.134
28	.739	-.408	.610	-.603	4.53	-1.44	1.99	-0.090	0.356
29	.476	-.170	.527	.453	4.16	-1.87	1.30	-0.178	0.337
30	.589	.108	.548	-.154	4.17	-1.81	-1.20	-0.156	0.207
31	.613	-.233	.549	-.319	1.60	-0.25	-1.26	-0.006	0.216
32	.256	3.366	.183	4.442	2.15	1.21	-1.31	-0.001	0.257
33	.772	-.374	.746	-.766	13.85***	-3.64***	-2.91**	-0.305	0.491
34	.546	-.018	.543	-.254	2.56	-1.56	-0.85	-0.133	0.138
35	.530	.406	.571	.569	6.32	2.18	2.24	0.216	0.216
36	.493	.607	.564	.517	0.58	0.09	0.60	0.030	0.107
37	.453	.352	.332	.396	4.91	0.86	-2.06	0.021	0.423
38	.210	1.220	.257	.850	0.56	-0.72	0.64	-0.044	0.125
39	.560	.210	.603	-.131	6.58	-2.56	2.56	-0.219	0.219
40	.338	.622	.356	.518	0.00	-0.00	-0.00	-0.000	0.000
รวม					8	8	10	3	4

จากตาราง 33 จะเห็นว่าวิธีลอร์ด (LC) ตรวจพบจำนวนข้อสอบที่ทำหน้าที่ต่างกัน 8 ข้อ วิธีราชู1 (ESA) 8 ข้อ วิธีราชู2 (EUA) 10 ข้อ วิธีคิม1 (CSA) 3 ข้อ วิธีคิม2 (CUA) 4 ข้อ และมีอยู่ 2 ข้อ ที่ตรวจพบว่าทำหน้าที่ต่างกันซ้ำกันในทุกวิธี

ตาราง 34 ผลการทดสอบความแตกต่างของจำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบระหว่างวิธี
ตรวจสอบ เมื่อกลุ่มตัวอย่าง 1,000 คน และความยาวแบบทดสอบ 40 ข้อ

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA	χ^2 รวม
LC		0.00	0.67	5.00	2.67	14.08**
ESA			0.67	5.00	2.00	
EUA				7.00**	6.00	
CSA					0.33	
CUA						

จากตาราง 34 จะเห็นว่าจำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบระหว่างวิธีตรวจสอบ
แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 โดยวิธีลอร์ด กับวิธีคิม1 วิธีราชู1 กับวิธีคิม1
และวิธีราชู2 กับวิธีคิม2 พบจำนวนข้อสอบที่ทำหน้าที่ต่างกันแตกต่างกันอย่างมีนัยสำคัญทางสถิติ
ที่ระดับ .05 ส่วนวิธีราชู2 กับวิธีคิม1 พบจำนวนข้อสอบที่ทำหน้าที่ต่างกันแตกต่างกันอย่างมี
นัยสำคัญทางสถิติที่ระดับ .01 นอกนั้นไม่แตกต่างกัน

ตาราง 35 ร้อยละของความสอดคล้องในการตรวจสอบความทำหน้าที่ต่างกันของข้อสอบ และ
สัมประสิทธิ์สหสัมพันธ์แบบเรียงอันดับ ระหว่างวิธีตรวจสอบ

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	.703***	.891***	.701***	.783***
ESA	60.00	-	.542***	.862***	.358***
EUA	50.00	50.00	-	.440**	.866***
CSA	37.50	37.50	30.00	-	.241
CUA	22.22	22.22	20.00	50.00	-

จากตาราง 35 ความสอดคล้องในการตรวจสอบความทำหน้าที่ต่างกันของข้อสอบ อยู่
ระหว่างร้อยละ 20.00 - 60.00 โดยที่วิธีลอร์ด กับวิธีราชู1 มีความสอดคล้องกันสูงสุด (ร้อยละ
60.00) และวิธีราชู2 กับวิธีคิม2 มีความสอดคล้องกันต่ำสุด (ร้อยละ 20.00)

สำหรับสัมประสิทธิ์สหสัมพันธ์ระหว่างวิธีตรวจสอบ พบว่า ที่มีค่าสูงสุดคือ ระหว่างวิธี
ลอร์ด กับวิธีราชู2 (.891) รองลงมา คือ ระหว่างวิธีราชู2 กับวิธีคิม2 (.866) และวิธีราชู1 กับ
วิธีคิม1 (.862)

ตาราง 36 จำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบ เมื่อกลุ่มตัวอย่าง 300 คน และ
ความยาวแบบทดสอบ 60 ข้อ

ข้อที่	ค่าพารามิเตอร์				วิธีตรวจสอบ				
	a_1	b_1	a_2	b_2	LC	ESA	EUA	CSA	CUA
1	.576	.157	.618	-.192	8.43	-2.89**	-2.89**	<u>-0.451</u>	<u>0.451</u>
2	.591	-.089	.537	-.231	3.42	-1.73	-1.16	-0.258	0.365
3	.707	.132	.583	-.184	9.43**	-2.79**	-1.94	<u>-0.424</u>	<u>0.555</u>
4	.709	.366	1.064	.263	4.18	-1.70	1.54	-0.219	<u>0.400</u>
5	.289	-.891	.416	-1.054	5.58	-1.20	0.99	<u>-0.393</u>	<u>0.401</u>
6	.505	.679	.648	.469	2.20	-1.48	0.90	-0.237	0.253
7	.720	1.028	.559	1.163	3.08	0.40	-1.52	-0.062	<u>0.492</u>
8	.652	1.480	.678	1.128	5.31	-1.29	-2.27*	<u>-0.349</u>	0.349
9	.733	.596	.797	.595	0.57	-0.61	-0.52	-0.101	0.106
10	.766	.907	.727	.961	1.02	0.00	-0.82	-0.050	0.240
11	.804	.435	.677	.231	7.71*	-2.13*	-1.90	<u>-0.338</u>	<u>0.486</u>
12	.698	-1.182	.611	-1.246	2.04	-1.40	-1.25	-0.210	0.379
13	.658	-.870	.710	-1.125	8.55*	-2.35*	-2.32*	<u>-0.474</u>	<u>0.474</u>
14	.564	-.326	.597	-.212	0.31	-0.51	-0.40	-0.072	0.116
15	.681	.958	.657	.818	3.26	-0.95	-1.06	-0.227	0.280
16	.745	-.640	.680	-.679	2.78	-1.63	-1.20	-0.218	0.348
17	.435	.387	.382	.598	0.96	0.34	-0.96	0.016	0.304
18	.302	-.435	.534	-.301	3.22	-0.33	1.46	-0.155	<u>0.515</u>
19	.340	1.027	.556	1.229	5.88	0.48	1.37	0.295	<u>0.466</u>
20	.263	1.197	.411	1.396	3.98	0.40	1.03	0.281	0.373
21	.448	.018	.467	.125	0.23	-0.26	-0.39	-0.044	0.115
22	.638	.395	.535	.322	4.33	-1.01	-1.47	-0.205	<u>0.435</u>
23	.532	-.144	.403	-.077	3.02	-0.53	-1.59	-0.054	<u>0.508</u>
24	.177	-.564	.239	-.550	0.70	-0.46	0.48	-0.131	0.157
25	.176	3.311	.174	2.794	1.60	-0.22	-0.42	-0.221	0.221
26	.378	1.289	.530	1.127	0.68	-0.50	0.78	-0.020	0.245
27	.520	-.129	.473	-.198	2.24	-1.24	-1.03	-1.181	0.321
28	.248	.949	.377	.651	1.09	-0.90	0.92	-0.114	0.292
29	.428	-.393	.354	-.649	2.94	-1.62	-1.23	-0.237	<u>0.409</u>
30	.645	.236	.545	.406	1.88	0.17	-1.33	0.000	<u>0.401</u>
31	.319	-1.907	.519	-1.125	1.59	0.83	1.13	0.009	0.386
32	.483	-.059	.695	.040	1.18	-0.42	1.00	-0.075	0.314
33	.319	1.347	.432	.906	1.39	-1.15	0.70	-0.213	0.254

ตาราง 36 (ต่อ)

ข้อที่	ค่าพารามิเตอร์				วิธีตรวจสอบ				
	a_1	b_1	a_2	b_2	LC	ESA	EUA	CSA	CUA
34	.325	1.873	.405	1.070	4.58	-1.59	1.18	<u>-0.422</u>	<u>0.422</u>
35	.469	-.543	.563	-.105	1.05	1.02	0.63	0.164	0.164
36	.585	-1.029	.481	-1.567	6.53	-2.50	-1.73	<u>-0.477</u>	<u>0.571</u>
37	.622	-1.271	.556	-1.657	5.14	-2.08	-1.48	<u>-0.444</u>	<u>0.488</u>
38	.548	.937	.761	.856	0.81	-0.62	0.89	-0.061	0.265
39	.469	.720	.583	.370	4.15	-1.95	1.38	<u>-0.346</u>	0.346
40	.453	.932	.572	.825	0.39	-0.62	0.48	-0.094	0.149
41	.308	1.481	.448	.816	2.75	-1.63	1.02	<u>-0.317</u>	0.370
42	.224	2.591	.411	1.175	3.46	-1.83	1.46	<u>-0.359</u>	<u>0.528</u>
43	.609	-.120	.783	-.177	3.01	-1.57	1.09	-0.242	0.259
44	.786	-.410	.553	-.535	6.48	-1.89	-2.21	-0.218	<u>0.668</u>
45	.534	.029	.627	-.553	17.00 ^{***}	-4.01 ^{***}	4.01 ^{***}	<u>-0.674</u>	<u>0.674</u>
46	.566	.160	.513	-.025	4.63	-1.76	-1.28	-0.285	0.374
47	.574	-.206	.590	-.345	3.44	-1.85	-1.16	-0.281	0.284
48	.291	3.550	.230	3.478	3.89	0.16	-0.94	<u>-0.359</u>	0.380
49	.318	-.098	.348	-.238	1.23	-1.11	-0.69	-0.194	0.194
50	.773	-.371	.768	-.733	13.62 ^{***}	-3.64 ^{**}	-3.17 ^{**}	<u>-0.548</u>	<u>0.548</u>
51	.564	-.074	.659	-.348	7.31	-2.67 ^{**}	2.67 ^{**}	<u>-0.419</u>	<u>0.419</u>
52	.328	.736	.244	.774	2.77	-0.08	-1.30	-0.142	<u>0.405</u>
53	.694	-.245	.754	-.321	3.23	-1.78	-1.69	-0.263	0.263
54	.545	.356	.780	.447	1.19	-0.18	1.03	-0.001	0.305
55	.697	-.215	.797	-.122	0.48	-0.69	0.69	-0.099	0.099
56	.610	.506	.554	.663	1.15	0.25	-1.03	-0.000	0.300
57	.438	.385	.409	.610	0.58	0.41	-0.74	-0.040	0.240
58	.583	.255	.733	-.156	11.56 ^{**}	-3.39 ^{***}	3.20	<u>-0.524</u>	<u>0.524</u>
59	.405	.477	.358	.461	1.64	-0.44	-0.64	-0.132	0.303
60	.462	.427	.527	.540	0.00	0.00	0.00	0.000	0.000
รวม					10	10	7	17	23

จากตาราง 36 จะเห็นว่าวิธีลอร์ด (LC) ตรวจพบจำนวนข้อสอบที่ทำหน้าที่ต่างกัน 10 ข้อ วิธีราชู1 (ESA) 10 ข้อ วิธีราชู2 (EUA) 7 ข้อ วิธีคิม1 (CSA) 17 ข้อ วิธีคิม2 (CUA) 23 ข้อ และมีอยู่ 6 ข้อ ที่ตรวจพบว่าทำหน้าที่ต่างกันซ้ำกันในทุกวิธี

ตาราง 37 ผลการทดสอบความแตกต่างของจำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบระหว่างวิธีตรวจสอบ เมื่อกลุ่มตัวอย่าง 300 คน และความยาวแบบทดสอบ 60 ข้อ

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA	χ^2 รวม
LC		0.00	1.80	5.44	13.00**	35.25**
ESA			1.80	7.00**	13.00**	
EUA				10.00**	14.22**	
CSA					2.57	
CUA						

จากตาราง 37 จะเห็นว่าจำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบระหว่างวิธีตรวจสอบแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 โดยวิธีลอร์ด กับวิธีคิม2 วิธีราชู1 กับวิธีคิม1 วิธีราชู1 กับวิธีคิม2 วิธีราชู2 กับวิธีคิม1 และวิธีราชู2 กับวิธีคิม2 พบจำนวนข้อสอบที่ทำหน้าที่ต่างกันแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ส่วนวิธีลอร์ด กับวิธีคิม1 พบจำนวนข้อสอบที่ทำหน้าที่ต่างกันแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นอกนั้นไม่แตกต่างกัน

ตาราง 38 ร้อยละของความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ และสัมประสิทธิ์สหสัมพันธ์แบบเรียงอันดับ ระหว่างวิธีตรวจสอบ

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	.711***	.878***	.860***	.840***
ESA	81.82	-	.652***	.789***	.502***
EUA	54.55	54.55	-	.619***	.878***
CSA	50.00	58.82	41.18	-	.578***
CUA	43.48	43.48	25.00	48.15	-

จากตาราง 38 ความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อยู่ระหว่างร้อยละ 25.00 - 81.82 โดยวิธีลอร์ด กับวิธีราชู1 มีความสอดคล้องกันสูงสุด (ร้อยละ 81.82) ส่วนวิธีราชู2 กับวิธีคิม2 มีความสอดคล้องกันต่ำสุด (ร้อยละ 25.00)

สำหรับสัมประสิทธิ์สหสัมพันธ์ระหว่างวิธีตรวจสอบ พบว่า มีนัยสำคัญทางสถิติทุกค่า ที่มีค่าสูงสุด คือ ระหว่างวิธีราชู2 กับวิธีคิม2 และวิธีลอร์ด กับวิธีราชู2 (.878) รองลงมา คือ วิธีลอร์ด กับวิธีคิม1 (.860) และวิธีลอร์ด กับวิธีคิม2 (.840)

ตาราง 39 จำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบ เมื่อกลุ่มตัวอย่าง 500 คน และ
ความยาวแบบทดสอบ 60 ข้อ

ข้อที่	ค่าพารามิเตอร์				วิธีตรวจสอบ				
	a_1	b_1	a_2	b_2	LC	ESA	EUA	CSA	CUA
1	.462	.257	.501	-.348	15.67 ^{***}	-3.88 ^{***}	3.77 ^{***}	<u>-0.501</u>	<u>0.501</u>
2	.546	.136	.575	-.427	17.60 ^{***}	-4.08 ^{***}	4.12 ^{***}	<u>-0.509</u>	<u>0.509</u>
3	.622	.190	.421	-.259	12.66 ^{**}	-3.06 ^{**}	-2.19 [*]	<u>-0.383</u>	<u>0.588</u>
4	.922	.522	.857	.209	13.79 ^{**}	-3.53 ^{***}	-3.67 ^{***}	<u>-0.361</u>	0.361
5	.518	-.638	.354	-1.350	7.06 [*]	-2.65 [*]	-1.81	<u>-0.356</u>	<u>0.541</u>
6	.518	.883	.606	.446	8.29 [*]	-2.83 [*]	1.67	<u>-0.362</u>	0.378
7	.667	1.242	.641	1.171	1.30	-0.71	-1.12	-0.129	0.129
8	.640	1.665	.654	1.238	8.73 [*]	-2.04 [*]	2.04 [*]	<u>-0.391</u>	0.391
9	.773	.653	.739	.497	3.87	-1.70	-1.92	-0.205	0.205
10	.845	.992	1.012	.717	6.50 [*]	-2.52 [*]	1.83	-0.292	0.339
11	.710	.529	.741	.330	4.55 [*]	-2.13 [*]	2.04 [*]	-0.231	0.231
12	.900	-1.029	.758	-1.400	6.24 [*]	-2.35 [*]	-1.79	<u>-0.336</u>	0.366
13	.858	-.758	.580	-1.435	12.43 ^{**}	-3.52 ^{***}	-2.70 ^{**}	<u>-0.496</u>	<u>0.659</u>
14	.742	-.183	.624	-.669	14.88 ^{***}	-3.82 ^{***}	-2.57 [*]	<u>-0.453</u>	<u>0.462</u>
15	.697	.991	.637	.837	4.07	-1.31	-1.39	-0.213	0.215
16	.587	-.452	.608	-.540	1.40	-0.97	1.00	-0.132	0.132
17	.358	.725	.333	.699	0.43	-0.33	-0.32	-0.075	0.094
18	.376	-.386	.377	-.477	0.56	-0.68	0.69	-0.099	0.099
19	.346	1.468	.365	1.749	1.52	0.51	0.56	0.161	0.161
20	.331	1.333	.299	1.656	0.36	0.59	-0.43	0.076	0.116
21	.357	-.057	.315	.214	0.98	0.88	-0.58	0.118	0.161
22	.586	.558	.608	.445	1.42	-1.17	0.84	-0.141	0.141
23	.454	-.152	.380	-.433	3.12	-1.71	-0.99	-0.208	0.269
24	.191	-.390	.132	-.788	1.37	-0.86	-1.00	-0.077	0.229
25	.218	2.790	.159	2.810	4.20	-0.04	-0.89	-0.281	0.297
26	.464	1.090	.476	1.382	1.83	0.83	1.10	0.173	0.173
27	.449	-.297	.378	-.105	1.26	0.65	-0.80	0.103	0.218
28	.240	.631	.345	.785	3.20	0.27	1.31	0.147	0.371
29	.322	-.410	.296	-.814	2.76	-1.58	-0.68	0.231	0.231
30	.434	.210	.415	.387	0.31	0.56	-0.29	0.072	0.078
31	.363	-1.478	.439	-1.565	3.79	-0.40	0.88	-0.220	0.267
32	.569	-.067	.530	-.239	2.50	-1.56	-0.92	-0.184	0.186
33	.253	1.268	.299	1.428	1.51	0.20	0.63	0.135	0.187

ตาราง 39 (ต่อ)

ข้อที่	ค่าพารามิเตอร์				วิธีตรวจสอบ				
	a_1	b_1	a_2	b_2	LC	ESA	EUA	CSA	CUA
34	.560	1.076	.368	1.434	4.59	0.96	-1.84	-0.013	<u>0.503</u>
35	.470	-.321	.608	-.094	2.53	1.02	1.38	0.102	0.356
36	.489	-1.082	.435	-1.768	8.84	-2.25	-1.67	<u>-0.444</u>	<u>0.444</u>
37	.589	-1.372	.397	-2.146	4.57	-2.08	-1.85	-0.292	<u>0.526</u>
38	.513	.973	.578	1.135	1.67	0.44	0.70	0.122	0.178
39	.511	.741	.439	.468	6.24	-1.70	-1.25	-0.284	0.299
40	.241	1.862	.371	.928	4.59	-2.11	1.65	-0.280	<u>0.483</u>
41	.232	1.327	.275	.940	1.10	-1.04	0.65	-0.135	0.184
42	.166	2.974	.243	2.259	1.42	-0.82	1.08	0.036	0.270
43	.576	-.295	.433	-.203	2.16	0.16	-1.37	0.053	0.370
44	.685	-.527	.538	-.663	2.52	-1.27	-1.39	-0.111	0.333
45	.459	-.185	.348	-.510	3.99	-1.82	-1.30	-0.211	0.366
46	.487	-.099	.489	.070	0.39	0.62	0.62	0.077	0.077
47	.715	-.248	.576	-.416	3.52	-1.66	-1.41	-0.171	0.321
48	.244	3.333	.228	3.473	0.16	0.06	-0.18	-0.046	0.057
49	.307	.031	.296	-.315	2.49	-1.57	-1.18	-0.218	0.218
50	.591	-.790	.642	-.786	0.76	-0.34	0.55	-0.073	0.123
51	.459	-.022	.555	-.262	5.11	-1.86	1.28	-0.246	0.315
52	.335	.657	.287	-.852	0.40	0.40	-0.62	0.022	0.159
53	.571	-.107	.719	-.495	15.45 ^{***}	-3.24 ^{**}	2.33	<u>-0.410</u>	<u>0.457</u>
54	.520	.486	.453	.715	0.78	0.81	-0.73	0.088	0.188
55	.554	-.293	.589	-.261	0.20	-0.21	0.36	-0.033	0.087
56	.538	.485	.492	.451	0.97	-0.55	-0.57	-0.092	0.133
57	.366	.396	.400	.425	0.17	-0.16	0.41	-0.008	0.108
58	.668	.236	.544	-.162	12.54 ^{**}	-3.28 ^{**}	-2.02	<u>-0.386</u>	<u>0.421</u>
59	.341	.479	.309	.612	0.15	0.24	-0.38	0.016	0.102
60	.443	.437	.442	.503	0.00	-0.00	0.00	-0.000	0.001
รวม					15	17	10	13	12

จากตาราง 39 จะเห็นว่าวิธีลอร์ด (LC) ตรวจพบจำนวนข้อสอบที่ทำหน้าที่ต่างกัน 15 ข้อ
 วิธีราฐู1 (ESA) 17 ข้อ วิธีราฐู2 (EUA) 10 ข้อ วิธีคิม1 (CSA) 13 ข้อ วิธีคิม2 (CUA) 12 ข้อ
 และมีอยู่ 7 ข้อ ที่ตรวจพบว่าทำหน้าที่ต่างกันซ้ำกันในทุกวิธี

ตาราง 40 ผลการทดสอบความแตกต่างของจำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบระหว่างวิธี
ตรวจสอบ เมื่อกลุ่มตัวอย่าง 500 คน และความยาวแบบทดสอบ 60 ข้อ

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA	χ^2 รวม
LC		1.00	3.57	2.00	2.00	11.31
ESA			7.00**	4.00	4.50	
EUA				1.80	0.14	
CSA					0.57	
CUA						

จากตาราง 40 จะเห็นว่าจำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบระหว่างวิธีตรวจสอบ
แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 โดยวิธีราชู1 กับวิธีคิม1 และวิธีราชู1 กับวิธีคิม2
พบจำนวนข้อสอบที่ทำหน้าที่ต่างกันแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ส่วนวิธีราชู1
กับวิธีราชู2 พบจำนวนข้อสอบที่ทำหน้าที่ต่างกันแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01
นอกนั้นไม่แตกต่างกัน

ตาราง 41 ร้อยละของความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ และ
สัมประสิทธิ์สหสัมพันธ์แบบเรียงอันดับ ระหว่างวิธีตรวจสอบ

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	.845***	.922***	.905***	.907***
ESA	77.78	-	.847***	.838***	.743***
EUA	56.25	58.82	-	.783***	.884***
CSA	86.67	76.47	62.29	-	.742***
CUA	60.00	64.71	50.00	60.00	-

จากตาราง 41 ความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อยู่
ระหว่างร้อยละ 50.00 - 86.67 โดยที่วิธีลอร์ด กับวิธีคิม1 มีความสอดคล้องกันสูงสุด (ร้อยละ
86.67) และ วิธีราชู2 กับวิธีคิม2 มีความสอดคล้องกันต่ำสุด (ร้อยละ 50.00)

สำหรับสัมประสิทธิ์สหสัมพันธ์ระหว่างวิธีตรวจสอบ พบว่า มีนัยสำคัญทางสถิติทุกค่า ที่มี
ค่าสูงสุด คือ ระหว่างวิธีลอร์ด กับวิธีราชู2 (.922) รองลงมาคือ ระหว่างวิธีลอร์ด กับวิธีคิม2
(.907) และวิธีลอร์ด กับวิธีคิม1 (.905)

ตาราง 42 จำนวนข้อสอบทำหน้าที่ต่างกันที่ตรวจพบ เมื่อกลุ่มตัวอย่าง 700 คน และ
ความยาวแบบทดสอบ 60 ข้อ

ข้อที่	ค่าพารามิเตอร์				วิธีตรวจสอบ				
	a_1	b_1	a_2	b_2	LC	ESA	EUA	CSA	CUA
1	.488	.126	.587	-.271	15.43 ^{***}	-3.30 ^{***}	2.31 [*]	<u>-0.348</u>	<u>0.491</u>
2	.530	-.154	.525	-.494	9.29 ^{**}	-2.58 ^{**}	1.66	-0.287	0.298
3	.585	.152	.518	-.134	10.22 ^{**}	-3.19 ^{**}	-3.19 ^{**}	<u>-0.324</u>	0.324
4	.830	.475	.840	.376	4.69	-2.16 [*]	1.46	-0.183	0.211
5	.506	-.743	.447	-1.158	4.45	-2.07 [*]	-1.29	-0.233	0.299
6	.472	1.165	.580	.548	14.30 ^{***}	-3.75 ^{***}	-2.36 [*]	<u>-0.447</u>	<u>0.538</u>
7	.610	1.357	.581	1.110	6.03 [*]	-1.93	1.69	-0.268	0.268
8	.544	1.651	.612	1.201	4.99	-1.92	1.32	-0.284	0.311
9	.643	.861	.741	.460	10.37 ^{***}	-3.22 ^{**}	2.17 [*]	<u>-0.315</u>	0.396
10	.759	1.172	.764	.861	9.07 [*]	-2.74 ^{**}	2.05 [*]	<u>-0.319</u>	0.327
11	.719	.534	.666	.286	11.19 ^{**}	-3.31 ^{***}	3.26 ^{**}	<u>-0.309</u>	0.309
12	.707	-1.254	.699	-1.417	4.01	-0.45	1.15	-0.118	0.203
13	.816	-.843	.742	-1.313	11.99 ^{**}	-2.74 ^{**}	-2.73 ^{**}	<u>-0.332</u>	0.332
14	.710	-.267	.591	-.561	6.18 [*]	-2.48 [*]	-1.56	-0.234	0.246
15	.598	.923	.484	.981	0.79	-0.82	0.52	-0.093	0.096
16	.665	-.597	.551	-.621	0.67	-0.81	0.50	-0.077	0.094
17	.378	.353	.371	.392	0.92	0.76	0.42	-0.102	0.102
18	.344	-.133	.396	-.485	7.59 [*]	-2.11 [*]	1.46	-0.281	0.379
19	.355	1.544	.370	1.490	1.45	-1.09	1.16	-0.083	0.264
20	.378	1.103	.271	1.675	1.42	1.01	1.13	0.055	0.277
21	.350	.160	.324	.172	0.46	-0.67	0.29	-0.075	0.079
22	.601	.608	.445	.412	8.65 [*]	-2.30 [*]	-2.06 [*]	-0.277	0.277
23	.421	-.066	.457	-.331	3.62	1.45	1.20	-0.170	0.272
24	.219	-.258	.155	-.486	0.92	-0.91	0.91	-0.115	0.115
25	.164	3.239	.117	4.380	0.28	0.37	-0.48	-0.017	0.091
26	.482	1.127	.440	1.455	1.08	0.57	1.02	0.108	0.108
27	.487	-.393	.448	-.269	2.91	1.70	1.49	0.177	0.177
28	.270	.877	.356	.471	3.85	-1.60	1.71	-0.130	<u>0.429</u>
29	.344	-.281	.299	-.721	7.17 [*]	-2.14 [*]	1.11	-0.296	0.296
30	.496	.227	.447	.109	0.05	-0.19	-0.11	-0.022	0.022
31	.425	-1.212	.494	-1.396	7.33 [*]	0.15	1.91	-0.146	<u>0.449</u>
32	.506	-.181	.601	-.184	3.84	0.24	1.88	0.005	0.394
33	.376	.999	.319	1.162	0.79	0.79	-0.84	0.052	0.199

ตาราง 42 (ต่อ)

ข้อที่	ค่าพารามิเตอร์				วิธีตรวจสอบ				
	a_1	b_1	a_2	b_2	LC	ESA	EUA	CSA	CUA
34	.261	2.039	.420	1.211	6.69	-2.15	2.20	-0.177	<u>0.575</u>
35	.466	-.305	.468	-.069	2.73	1.40	1.09	0.133	0.235
36	.458	-1.221	.487	-1.584	10.19**	-1.14	1.43	-0.297	0.326
37	.567	-1.383	.464	-2.068	4.00	-1.36	-1.24	-0.234	0.234
38	.537	1.177	.583	1.062	3.05	-1.74	1.44	-0.182	0.309
39	.454	.495	.588	.417	5.36	-1.58	2.08	-0.131	<u>0.452</u>
40	.333	1.220	.506	.725	8.99	-2.54	2.54	-0.233	<u>0.640</u>
41	.263	1.098	.374	.622	5.23	-1.84	1.93	-0.141	<u>0.506</u>
42	.150	3.279	.218	2.291	4.73	-1.66	1.82	-0.054	<u>0.451</u>
43	.563	-.200	.631	-.197	2.00	-0.55	1.25	-0.068	0.248
44	.643	-.540	.554	-.639	0.94	-0.93	-0.69	-0.080	0.137
45	.416	-.238	.508	-.460	10.54**	-1.65	2.36	-0.238	<u>0.556</u>
46	.480	.007	.529	-.153	4.51	-1.58	1.52	-0.170	0.319
47	.516	-.141	.536	-.320	3.77	-1.33	1.33	-0.150	0.274
48	.201	4.670	.170	4.770	0.89	-0.58	0.43	-0.143	0.143
49	.396	-.020	.317	-.088	0.60	-0.77	-0.75	-0.085	0.085
50	.783	-.492	.668	-.814	7.99	-2.51	2.57	-0.244	0.244
51	.521	-.004	.529	-.253	7.62	-2.33	1.74	-0.244	0.354
52	.361	.545	.231	.951	4.06	0.60	-1.72	-0.024	<u>0.434</u>
53	.622	-.171	.674	-.367	4.93	-1.99	1.47	-0.201	0.205
54	.511	.440	.543	.596	2.08	0.28	1.14	0.063	0.237
55	.568	-.342	.668	-.249	3.89	1.08	1.74	0.080	0.350
56	.444	.426	.544	.536	4.32	0.00	1.76	0.051	0.382
57	.388	.287	.327	.404	0.32	0.56	-0.34	0.059	0.085
58	.621	.151	.612	-.124	13.94***	-3.64***	2.49	<u>-0.346</u>	<u>0.401</u>
59	.403	.655	.338	.545	4.04	-1.63	-1.34	-0.218	0.218
60	.420	.456	.390	.567	0.83	-0.07	0.82	0.018	0.177
รวม					21	20	14	8	12

จากตาราง 42 จะเห็นว่าวิธีลอร์ด (LC) ตรวจพบจำนวนข้อสอบที่ทำหน้าที่ต่างกัน 21 ข้อ วิธีราชู1(ESA) 20 ข้อ วิธีราชู2 (EUA) 14 ข้อ วิธีคิม1 (CSA) 8 ข้อ วิธีคิม2 (CUA) 12 ข้อ และมีอยู่ 3 ข้อ ที่ตรวจพบว่าทำหน้าที่ต่างกันซ้ำกันในทุกวิธี

ตาราง 43 ผลการทดสอบความแตกต่างของจำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบระหว่างวิธีตรวจสอบ เมื่อกลุ่มตัวอย่าง 700 คน และความยาวแบบทดสอบ 60 ข้อ

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA	χ^2 รวม
LC		0.14	5.44	13.00**	4.26	19.35**
ESA			3.60	12.00**	2.91	
EUA				6.00*	0.33	
CSA					1.14	
CUA						

จากตาราง 43 จะเห็นว่าจำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบระหว่างวิธีตรวจสอบแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 โดยวิธีลอร์ด กับวิธีคิม1 และ วิธีราชู1 กับวิธีคิม1 พบจำนวนข้อสอบที่ทำหน้าที่ต่างกันแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ส่วนวิธีลอร์ด กับวิธีราชู2 วิธีลอร์ด กับวิธีคิม2 และวิธีราชู2 กับวิธีคิม1 พบจำนวนข้อสอบที่ทำหน้าที่ต่างกันแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นอกนั้นไม่แตกต่างกัน

ตาราง 44 ร้อยละของความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ และสัมประสิทธิ์สหสัมพันธ์แบบเรียงอันดับ ระหว่างวิธีตรวจสอบ

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	.799***	.881***	.835***	.770***
ESA	70.83	-	.685***	.871***	.494***
EUA	59.09	54.55	-	.598***	.833***
CSA	38.09	40.00	57.14	-	.424**
CUA	30.77	44.44	36.84	25.00	-

จากตาราง 44 ความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อยู่ระหว่างร้อยละ 25.00 - 70.83 โดยที่วิธีลอร์ด กับวิธีราชู1 มีความสอดคล้องกันสูงสุด (ร้อยละ 70.83) ส่วนวิธีคิม1 กับวิธีคิม2 มีความสอดคล้องกันต่ำสุด (ร้อยละ 25.00)

สำหรับสัมประสิทธิ์สหสัมพันธ์ระหว่างวิธีตรวจสอบ พบว่า มีนัยสำคัญทางสถิติทุกค่า ที่มีค่าสูงสุด คือ ระหว่างวิธีลอร์ด และวิธีราชู2 (.881) รองลงมา คือ ระหว่างวิธีราชู1 กับวิธีคิม1 (.871) และวิธีลอร์ด กับวิธีคิม1 (.835)

ตาราง 45 จำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบ เมื่อกลุ่มตัวอย่าง 1,000 คน และ
ความยาวแบบทดสอบ 60 ข้อ

ข้อที่	ค่าพารามิเตอร์				วิธีตรวจสอบ				
	a_1	b_1	a_2	b_2	LC	ESA	EUA	CSA	CUA
1	.535	.137	.511	-.320	21.51 ^{***}	-4.56 ^{***}	-2.57 ^{***}	<u>-0.410</u>	<u>0.419</u>
2	.544	.001	.572	-.373	17.64 ^{***}	-4.20 ^{***}	-3.37 ^{***}	<u>-0.368</u>	0.368
3	.529	.181	.522	-.127	11.53 ^{**}	-3.20 ^{**}	1.85	-0.289	0.290
4	.800	.617	.822	.325	15.46 ^{***}	-3.40 ^{***}	-3.27 ^{***}	-0.288	<u>0.583</u>
5	.482	-.580	.362	-1.275	17.76 ^{***}	-4.20 ^{***}	-2.84 ^{**}	<u>-0.406</u>	0.231
6	.545	.966	.623	.639	6.67 [*]	-2.17 [*]	2.17 [*]	-0.231	0.259
7	.683	1.334	.663	1.061	9.15 [*]	-1.25	-1.66	-0.217	0.315
8	.661	1.532	.585	1.410	6.21 [*]	0.09	-1.70	-0.128	0.333
9	.785	.674	.702	.554	8.21 [*]	-0.99	-2.17 [*]	-0.133	0.295
10	.832	1.145	.846	.795	16.20 ^{***}	-2.72 [*]	-3.51 ^{***}	-0.294	0.291
11	.852	.478	.842	.227	14.78 ^{***}	-3.29 ^{***}	-2.68 ^{**}	-0.266	<u>0.416</u>
12	.808	-1.067	.834	-1.319	18.81 ^{***}	-3.67 ^{***}	-3.24 ^{**}	<u>-0.395</u>	<u>0.434</u>
13	.785	-.814	.855	-1.077	23.23 ^{***}	-4.02 ^{***}	-4.01 ^{***}	<u>-0.404</u>	0.397
14	.727	-.224	.710	-.557	21.44 ^{***}	-4.62 ^{***}	-3.18 ^{**}	<u>-0.384</u>	0.194
15	.559	1.120	.547	1.049	1.95	0.10	-1.11	-0.056	<u>0.440</u>
16	.656	-.506	.536	-.535	6.86 [*]	-1.32	-2.55 [*]	-0.053	0.128
17	.314	.553	.325	.763	1.41	1.18	-0.62	0.110	0.139
18	.341	-.125	.401	-.258	2.12	-1.40	0.85	-0.139	0.034
19	.285	1.846	.334	1.588	0.07	-0.25	0.18	-0.025	0.100
20	.306	1.398	.318	1.474	0.47	0.68	-0.52	0.054	0.099
21	.360	.137	.398	.030	1.13	-1.02	-0.54	-0.099	0.238
22	.594	.554	.580	.391	5.98	-1.38	-1.52	-0.161	0.208
23	.408	-.071	.463	-.284	4.97	-2.21 [*]	-2.21 [*]	-0.208	0.218
24	.157	-.048	.154	.725	4.36	-2.08 [*]	-0.63	-0.218	0.100
25	.139	3.785	.134	3.583	0.93	0.18	-0.50	-0.087	0.043
26	.453	1.306	.500	1.179	0.20	-0.10	-0.24	-0.032	0.236
27	.422	-.308	.395	-.162	2.29	0.43	-1.34	0.066	0.167
28	.250	1.150	.322	.697	2.34	-1.47	0.74	-0.152	0.343
29	.360	-.320	.295	-.462	4.07	-1.34	-1.81	-0.081	<u>0.465</u>
30	.486	.313	.375	.408	7.12 [*]	0.58	-2.50 [*]	-0.000	<u>0.465</u>
31	.435	-1.195	.433	-1.437	5.61	-2.10 [*]	-1.28	-0.237	0.254
32	.467	-.060	.587	-.059	0.83	-0.55	0.76	0.053	0.132
33	.291	1.211	.239	1.733	3.23	1.75	-1.55	0.120	0.296

ตาราง 45 (ต่อ)

ข้อที่	ค่าพารามิเตอร์				วิธีตรวจสอบ				
	a ₁	b ₁	a ₂	b ₂	LC	ESA	EUA	CSA	CUA
34	.279	1.997	.465	1.253	5.37	-1.84	2.05	0.083	<u>0.401</u>
35	.466	-.427	.467	-.090	5.35	1.98	-1.24	0.203	0.227
36	.519	-1.095	.471	-1.517	11.42 ^{**}	-3.16 ^{***}	-2.05 [*]	<u>0.342</u>	<u>0.409</u>
37	.595	-1.219	.463	-1.910	15.71 ^{***}	-3.93 ^{***}	-2.97 ^{**}	<u>0.445</u>	<u>0.593</u>
38	.451	1.220	.536	1.208	1.18	0.50	0.54	0.091	0.091
39	.492	.637	.444	.728	2.79	0.84	-1.67	0.022	0.288
40	.291	1.701	.449	1.026	4.25	-1.99 [*]	1.77	-0.153	0.345
41	.178	1.976	.325	.740	7.02 [*]	-2.59 ^{**}	1.89	-0.264	<u>0.431</u>
42	.160	3.234	.165	2.965	0.64	0.08	-0.38	-0.077	0.082
43	.558	-.126	.726	-.020	1.46	0.44	1.12	0.032	0.180
44	.635	-.506	.606	-.467	2.25	-0.77	-1.47	-0.033	0.238
45	.451	-.251	.445	-.433	4.70	-2.11 [*]	-1.22	-0.179	0.225
46	.460	.057	.535	-.062	2.25	-1.50	1.49	-0.135	0.135
47	.557	-.111	.527	-.275	6.42 [*]	-2.23 [*]	-1.65	-0.181	0.282
48	.184	4.812	.186	4.423	1.10	0.11	-0.38	-0.105	0.107
49	.315	.107	.359	-.203	5.08	-2.25	-2.25 [*]	-0.221	0.221
50	.727	-.520	.656	-.735	12.42 ^{**}	-3.46 ^{***}	-2.42 [*]	-0.261	0.381
51	.518	-.100	.490	-.149	3.06	-1.04	-1.48	-0.081	0.248
52	.339	.624	.230	1.029	7.08 [*]	1.65	-2.35 [*]	0.058	<u>0.474</u>
53	.632	-.109	.643	-.309	9.05 [*]	-2.94 ^{**}	-1.86	-0.238	0.254
54	.542	.511	.523	.593	1.74	0.79	-1.31	0.040	0.219
55	.575	-.229	.576	-.284	2.68	-1.39	-1.15	-0.105	0.190
56	.458	.521	.480	.462	0.89	-0.36	-0.66	-0.055	0.112
57	.456	.450	.341	.616	7.32 [*]	1.07	-2.52 [*]	0.020	<u>0.480</u>
58	.627	.153	.721	-.037	6.91 [*]	-2.62 ^{**}	2.62 ^{**}	-0.216	0.216
59	.348	.669	.348	.501	2.37	-0.80	-0.84	-0.123	0.166
60	.423	.485	.482	.477	0.00	0.00	-0.00	0.000	0.000
รวม					25	25	22	8	13

จากตาราง 45 จะเห็นว่าวิธีลอร์ด (LC) ตรวจพบจำนวนข้อสอบที่ทำหน้าที่ต่างกัน 25 ข้อ วิธีราชู1 (ESA) 25 ข้อ วิธีราชู2 (EUA) 22 ข้อ วิธีคิม1 (CSA) 8 ข้อ วิธีคิม2 (CUA) 13 ข้อ และมีอยู่ 6 ข้อ ที่ตรวจพบว่าทำหน้าที่ต่างกันซ้ำกันในทุกวิธี

ตาราง 46 ผลการทดสอบความแตกต่างของจำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบระหว่างวิธีตรวจสอบ เมื่อกลุ่มตัวอย่าง 1,000 คน และความยาวแบบทดสอบ 60 ข้อ

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA	χ^2 รวม
LC		0.00	1.00	17.00**	8.47**	31.21**
ESA			0.53	17.00**	6.26*	
EUA				14.00**	5.79*	
CSA					2.08	
CUA						

จากตาราง 46 จะเห็นว่าจำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบระหว่างวิธีตรวจสอบแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 โดยวิธีลอร์ด กับวิธีคิม1 วิธีลอร์ด กับวิธีคิม2 วิธีราชู1 กับวิธีคิม1 และวิธีราชู2 กับวิธีคิม1 พบจำนวนข้อสอบที่ทำหน้าที่ต่างกันแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ส่วนวิธีราชู1 กับวิธีคิม2 และวิธีราชู2 กับวิธีคิม2 พบจำนวนข้อสอบที่ทำหน้าที่ต่างกันแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นอกนั้นไม่แตกต่างกัน

ตาราง 47 ร้อยละของความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ และสัมประสิทธิ์สหสัมพันธ์แบบเรียงอันดับ ระหว่างวิธีตรวจสอบ

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	.832***	.918***	.757***	.858***
ESA	56.25	-	.766***	.877***	.626***
EUA	64.29	51.61	-	.599***	.854***
CSA	32.00	32.00	36.36	-	.449***
CUA	42.31	23.33	47.83	42.86	-

จากตาราง 47 ความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อยู่ระหว่างร้อยละ 23.33 - 64.29 โดยที่วิธีลอร์ด กับวิธีราชู2 มีความสอดคล้องกันสูงสุด (ร้อยละ 64.29) และวิธีราชู1 กับวิธีคิม2 มีความสอดคล้องกันต่ำสุด (ร้อยละ 23.33) สำหรับสัมประสิทธิ์สหสัมพันธ์ระหว่างวิธีตรวจสอบ พบว่า มีนัยสำคัญทางสถิติทุกค่า ที่มีค่าสูงสุด คือ ระหว่างวิธีลอร์ด กับวิธีราชู2 (.918) รองลงมาคือ ระหว่างวิธีราชู1 กับวิธีคิม1 (.877) และวิธีลอร์ด กับวิธีคิม2 (.858)

ตาราง 48 ผลการทดสอบความแตกต่างของจำนวนข้อสอบที่ทำหน้าที่ต่างกัน ที่ตรวจพบระหว่างวิธี
 ตรวจสอบ เมื่อกลุ่มตัวอย่าง 500 คน ความยาวแบบทดสอบ 20, 30 และ 40 ข้อ
 และกลุ่มตัวอย่าง 700 คน ความยาวแบบทดสอบ 40 ข้อ

กลุ่มตัวอย่าง/ ความยาวแบบทดสอบ	วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA	χ^2
N=500, l=20							5.54
N=500, l=30							7.40
N=500, l=40							2.00
N=700, l=40							2.89

จากตาราง 48 จำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบจากการตรวจสอบด้วยวิธี
 ลอร์ด วิธีราชู1 วิธีราชู2 วิธีคิม1 และวิธีคิม2 เมื่อกลุ่มตัวอย่าง 500 คน แบบทดสอบยาว 20,
 30 และ 40 ข้อ และกลุ่มตัวอย่าง 700 คน แบบทดสอบยาว 40 ข้อ พบว่า ไม่มีความแตกต่างกัน

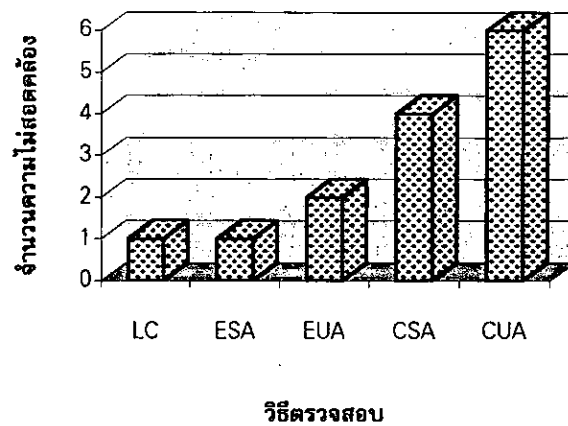
2.2 อัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ แบ่งเป็น 2 ลักษณะ คือ อัตราความไม่สอดคล้องชนิดที่ 1 และอัตราความไม่สอดคล้องชนิดที่ 2 ดังนี้

ตาราง 49 จำนวนความไม่สอดคล้องของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 300 คน ความยาวแบบทดสอบ 20, 30, 40 และ 60 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน 0%, 10%, 20%

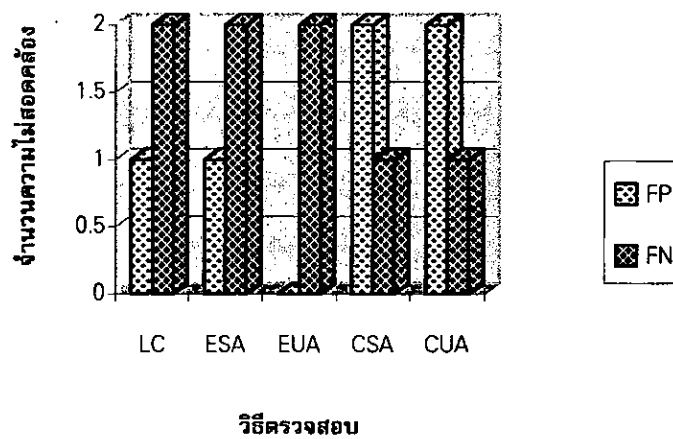
I % DIF	LC		ESA		EUA		CSA		CUA	
	I	II	I	II	I	II	I	II	I	II
I = 20										
0% DIF	1	-	1	-	2	-	4	-	6	-
10% DIF	1	2	1	2	0	2	2	1	2	1
20% DIF	8	3	9	4	5	3	10	3	10	3
I = 30										
0% DIF	1	-	1	-	2	-	5	-	6	-
10% DIF	2	3	2	3	2	3	3	3	7	2
20% DIF	3	6	4	5	3	5	5	4	7	3
I = 40										
0% DIF	2	-	1	-	1	-	3	-	6	-
10% DIF	1	4	2	3	2	4	2	3	8	3
20% DIF	7	7	4	6	6	7	5	6	11	4
I = 60										
0% DIF	1	-	2	-	1	-	5	-	16	-
10% DIF	19	4	14	5	13	5	18	3	20	3
20% DIF	20	9	16	7	15	8	16	6	22	5

I หมายถึง ความไม่สอดคล้องชนิดที่ 1

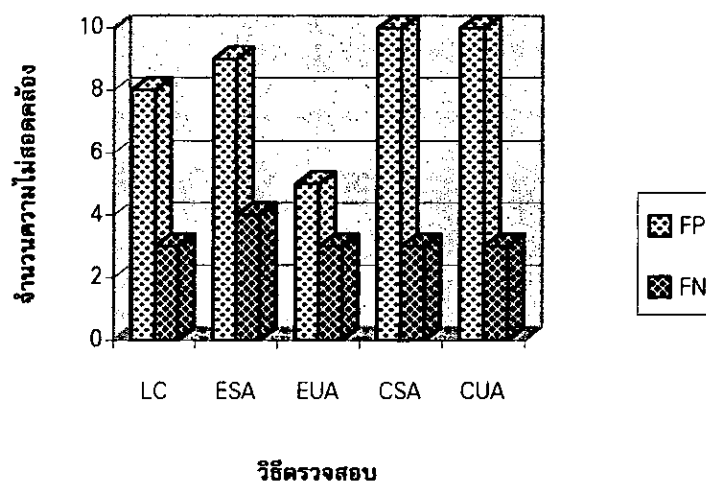
II หมายถึง ความไม่สอดคล้องชนิดที่ 2



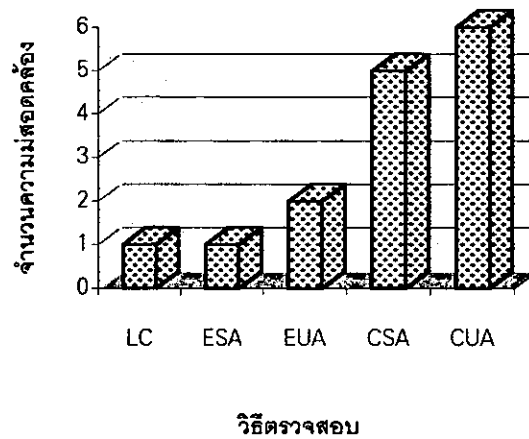
ภาพประกอบ 7 จำนวนความไม่สอดคล้อง เมื่อ $N = 300$, $I = 20$, 0%DIF



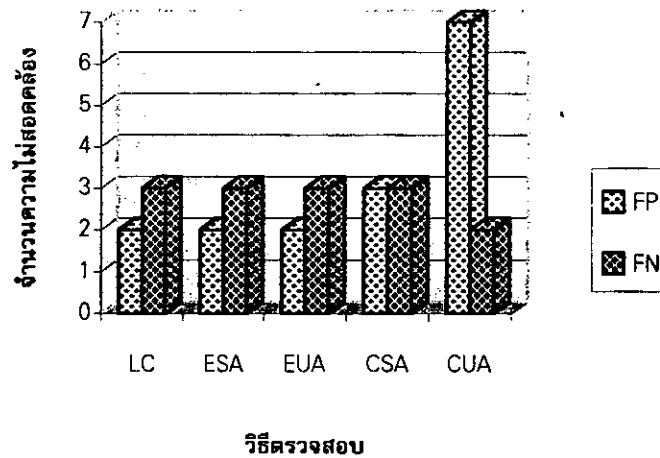
ภาพประกอบ 8 จำนวนความไม่สอดคล้อง เมื่อ $N = 300$, $I = 20$, 10%DIF



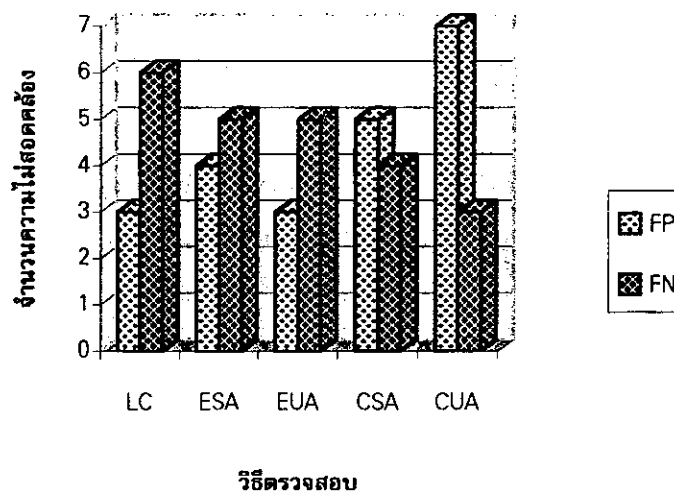
ภาพประกอบ 9 จำนวนความไม่สอดคล้อง เมื่อ $N = 300$, $I = 20$, 20%DIF



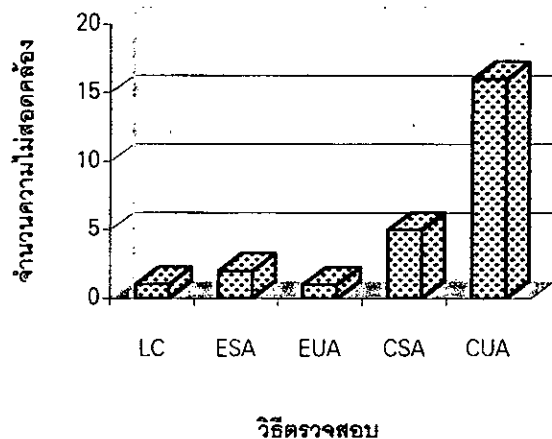
ภาพประกอบ 10 จำนวนความไม่สอดคล้อง เมื่อ $N = 300, I = 30, 0\%DIF$



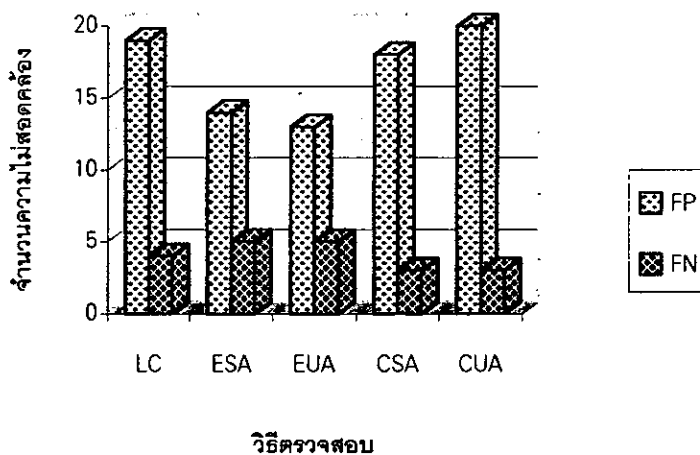
ภาพประกอบ 11 จำนวนความไม่สอดคล้อง เมื่อ $N = 300, I = 30, 10\%DIF$



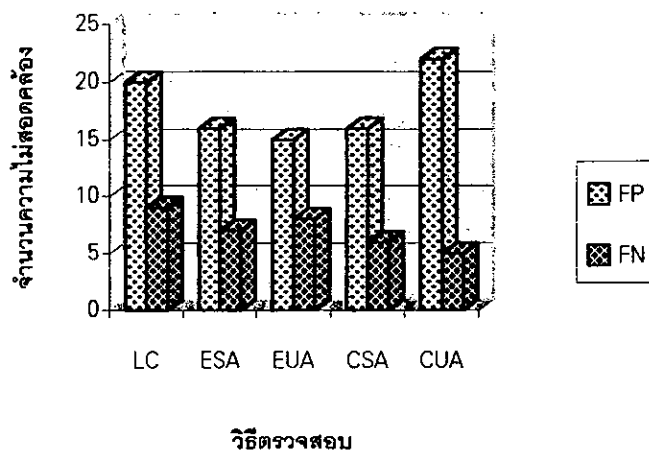
ภาพประกอบ 12 จำนวนความไม่สอดคล้อง เมื่อ $N = 300, I = 30, 20\%DIF$



ภาพประกอบ 16 จำนวนความไม่สอดคล้อง เมื่อ $N = 300$, $I = 60$, 0%DIF



ภาพประกอบ 17 จำนวนความไม่สอดคล้อง เมื่อ $N = 300$, $I = 60$, 10%DIF



ภาพประกอบ 18 จำนวนความไม่สอดคล้อง เมื่อ $N = 300$, $I = 60$, 20%DIF

วิธีที่มีความไม่สอดคล้องต่ำสุด คือ วิธีคิม2 (4) นอกนั้นมีความไม่สอดคล้องใกล้เคียงกัน (6 - 7)

เมื่อแบบทดสอบ 60 ข้อ สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน ร้อยละ 10 วิธีที่มีความไม่สอดคล้องต่ำสุด คือ วิธีคิม1 (3) และวิธีคิม2 (3) นอกนั้นมีความไม่สอดคล้องใกล้เคียงกัน (4 - 5) และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน ร้อยละ 20 วิธีที่มีความไม่สอดคล้องต่ำสุด คือ วิธีคิม2 (5) ส่วนวิธีที่มีความไม่สอดคล้องสูงสุด คือ วิธีลอร์ด (9)

ตาราง 50 อัตราความไม่สอดคล้องของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อ
กลุ่มตัวอย่าง 300 คน ความยาวแบบทดสอบ 20, 30, 40 และ 60 ข้อ และสัดส่วนจำนวน
ข้อสอบที่ทำหน้าที่ต่างกัน 0%, 10%, 20%

I % DIF	LC		ESA		EUA		CSA		CUA	
	I	II	I	II	I	II	I	II	I	II
I = 20										
0% DIF	.05	-	.05	-	.10	-	.20	-	.30	-
10% DIF	.06	1.00	.06	1.00	.00	1.00	.11	.50	.11	.50
20% DIF	.50	.75	.56	1.00	.31	.75	.63	.75	.63	.75
I = 30										
0% DIF	.03	-	.03	-	.07	-	.17	-	.20	-
10% DIF	.07	1.00	.07	1.00	.07	1.00	.11	1.00	.26	.67
20% DIF	.13	1.00	.17	.83	.13	.83	.21	.67	.29	.50
I = 40										
0% DIF	.05	-	.03	-	.03	-	.08	-	.15	-
10% DIF	.03	1.00	.06	.75	.06	1.00	.06	.75	.22	.75
20% DIF	.22	.88	.13	.75	.19	.88	.16	.75	.34	.50
I = 60										
0% DIF	.02	-	.03	-	.02	-	.08	-	.27	-
10% DIF	.35	.67	.26	.83	.24	.83	.33	.50	.37	.50
20% DIF	.42	.75	.33	.58	.31	.67	.33	.50	.46	.42

จากตาราง 50 อัตราความไม่สอดคล้องชนิดที่ 1 เมื่อแบบทดสอบ 20 ข้อ สัดส่วน
จำนวนข้อสอบที่ทำหน้าที่ต่างกัน ร้อยละ 0 วิธีที่มีอัตราความไม่สอดคล้องสูงสุด คือ วิธีคิม2 (.30)
รองลงมา คือ วิธีคิม1 (.20) และต่ำสุด คือ วิธีลอร์ด และวิธีราชู1 (.05) สัดส่วนจำนวนข้อสอบ
ที่ทำหน้าที่ต่างกัน ร้อยละ 10 วิธีราชู2 ไม่มีความไม่สอดคล้องเกิดขึ้น นอกนั้นม้อัตราความ
ไม่สอดคล้องใกล้เคียงกัน (.06 - .11) และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20
วิธีลอร์ด วิธีราชู1 วิธีคิม1 และวิธีคิม2 มีอัตราความไม่สอดคล้องใกล้เคียงกัน (.50 - .63)
ยกเว้น วิธีราชู2 มีอัตราความไม่สอดคล้องต่ำสุด (.31)

เมื่อแบบทดสอบ 30 ข้อ สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0 วิธีที่มีอัตราความไม่สอดคล้องสูงสุด คือ วิธีคิม2 (.20) รองลงมา คือ วิธีคิม1 (.17) และวิธีที่มีอัตราความไม่สอดคล้องต่ำสุด คือ วิธีลอร์ด และวิธีราชู1 (.03) สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน ร้อยละ 10 วิธีที่มีอัตราความไม่สอดคล้องสูงสุด คือ วิธีคิม2 (.26) นอกนั้นมีอัตราความไม่สอดคล้องใกล้เคียงกัน (.07 - .11) และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20 วิธีที่มีอัตราความไม่สอดคล้องสูงสุด คือ วิธีคิม2 (.29) นอกนั้นมีอัตราความไม่สอดคล้องใกล้เคียงกัน

เมื่อแบบทดสอบ 40 ข้อ สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0 วิธีที่มีอัตราความไม่สอดคล้องสูงสุด คือ วิธีคิม2 (.15) และวิธีที่มีอัตราความไม่สอดคล้องต่ำสุด คือ วิธีราชู1 และวิธีราชู2 (.03) สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน ร้อยละ 10 วิธีที่มีอัตราความไม่สอดคล้องสูงสุด คือ วิธีคิม2 (.22) และวิธีที่มีอัตราความไม่สอดคล้องต่ำสุด คือ วิธีลอร์ด (.03) และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20 วิธีที่มีอัตราความไม่สอดคล้องสูงสุด คือ วิธีคิม2 (.34) และวิธีที่มีอัตราความไม่สอดคล้องต่ำสุด คือ วิธีราชู1 (.13)

เมื่อแบบทดสอบ 60 ข้อ สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0 วิธีที่มีอัตราความไม่สอดคล้องสูงสุด คือ วิธีคิม2 (.27) และวิธีที่มีอัตราความไม่สอดคล้องต่ำสุด คือ วิธีลอร์ด และวิธีราชู2 (.02) สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 วิธีที่มีอัตราความไม่สอดคล้องสูงสุด คือ วิธีคิม2 (.37) รองลงมา คือ วิธีลอร์ด (.35) และวิธีคิม1 (.33) ตามลำดับ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20 วิธีที่มีอัตราความไม่สอดคล้องสูงสุด คือ วิธีคิม2 (.46) และวิธีที่มีอัตราความไม่สอดคล้องต่ำสุด คือ วิธีราชู2 (.31)

สำหรับอัตราความไม่สอดคล้องชนิดที่ 2 ส่วนใหญ่มีค่าลดลง เมื่อสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันเพิ่มขึ้น

ตาราง 51 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 300 คน ความยาวแบบทดสอบ 20 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC		0.00	0.60	1.43	2.08
ESA			0.60	1.43	2.08
EUA				0.89	1.58
CSA					0.73
CUA					

จากตาราง 51 จะเห็นว่า วิธีลอร์ด กับวิธีคิม2 และวิธีราชู1 กับวิธีคิม2 มีอัตราความไม่สอดคล้องชนิดที่ 1 แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นอกนั้นไม่แตกต่างกัน

ตาราง 52 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 300 คน ความยาวแบบทดสอบ 20 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	0.00	1.10	0.55	0.55
ESA	0.00	-	1.10	0.55	0.55
EUA	0.00	0.00	-	1.45	1.45
CSA	1.15	1.15	1.15	-	0.00
CUA	1.15	1.15	1.15	0.00	-

สามเหลี่ยมบน หมายถึง อัตราความไม่สอดคล้องชนิดที่ 1

สามเหลี่ยมล่าง หมายถึง อัตราความไม่สอดคล้องชนิดที่ 2

จากตาราง 52 จะเห็นว่า อัตราความไม่สอดคล้องชนิดที่ 1 และชนิดที่ 2 ทั้ง 2 กรณี ไม่พบความแตกต่าง

ตาราง 53 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 300 คน ความยาวแบบทดสอบ 20 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	0.34	1.09	0.74	0.74
ESA	1.07	-	1.43	0.40	0.40
EUA	0.00	1.07	-	1.81	1.81
CSA	0.00	1.07	0.00	-	0.00
CUA	0.00	1.07	0.00	0.00	-

จากตาราง 53 จะเห็นว่า อัตราความไม่สอดคล้องชนิดที่ 1 และชนิดที่ 2 ทั้ง 2 กรณี ไม่พบความแตกต่าง

ตาราง 54 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 300 คน ความยาวแบบทดสอบ 30 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC		0.00	0.71	1.81	2.05
ESA			0.71	1.81	2.05
EUA				1.21	1.48
CSA					0.30
CUA					

จากตาราง 54 จะเห็นว่า วิธีลอร์ด กับวิธีคิม2 และวิธีราซู1 กับวิธีคิม2 มีอัตราความไม่สอดคล้องชนิดที่ 1 แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05

ตาราง 55 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 300 คน ความยาวแบบทดสอบ 30 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	0.00	0.00	0.51	1.73
ESA	0.00	-	0.00	0.51	1.73
EUA	0.00	0.00	-	0.51	1.73
CSA	0.00	0.00	0.00	-	1.43
CUA	1.17	1.17	1.17	1.17	-

จากตาราง 55 จะเห็นว่าอัตราความไม่สอดคล้องชนิดที่ 1 และชนิดที่ 2 ทั้ง 2 กรณีไม่พบความแตกต่าง

ตาราง 56 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 300 คน ความยาวแบบทดสอบ 30 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	0.39	0.00	0.74	1.37
ESA	1.07	-	0.39	0.36	0.98
EUA	1.07	0.00	-	0.74	1.37
CSA	1.53	0.64	0.64	-	0.64
CUA	2.00	1.49	1.49	0.59	-

จากตาราง 56 จะเห็นว่า อัตราความไม่สอดคล้องชนิดที่ 1 ไม่แตกต่างกันสำหรับอัตราความไม่สอดคล้องชนิดที่ 2 พบว่า วิธีลอร์ด กับวิธีคิม2 มีอัตราความไม่สอดคล้องแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05

ตาราง 57 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 300 คน ความยาวแบบทดสอบ 40 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC		0.47	0.47	0.55	1.49
ESA			0.00	1.03	1.90
EUA				1.03	1.90
CSA					0.99
CUA					

จากตาราง 57 จะเห็นว่าอัตราความไม่สอดคล้องชนิดที่ 1 ไม่แตกต่างกัน

ตาราง 58 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 300 คน ความยาวแบบทดสอบ 40 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	0.64	0.64	0.64	2.44 ^{**}
ESA	1.07	-	0.00	0.00	1.96 [*]
EUA	0.00	1.07	-	0.00	1.96 [*]
CSA	1.07	0.00	1.07	-	1.96 [*]
CUA	1.07	0.00	1.07	0.00	-

จากตาราง 58 จะเห็นว่า วิธีลอร์ด กับวิธีคิม2 วิธีราซู1 กับวิธีคิม2 วิธีราซู2 กับวิธีคิม2 และวิธีคิม1 กับวิธีคิม2 มีอัตราความไม่สอดคล้องชนิดที่ 1 แตกต่างกันอย่างมีนัยสำคัญทางสถิติ ที่ระดับ .05 นอกนั้นไม่แตกต่างกัน

สำหรับอัตราความไม่สอดคล้องชนิดที่ 2 พบว่า ไม่แตกต่างกัน

ตาราง 59 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 300 คน ความยาวแบบทดสอบ 40 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	0.95	0.29	0.61	1.07
ESA	0.67	-	0.66	0.35	1.98 [*]
EUA	0.00	0.67	-	0.95	1.36
CSA	0.67	0.00	0.67	-	1.66
CUA	1.64	1.03	1.64	1.03	-

จากตาราง 59 จะเห็นว่า วิธีราซู1 กับวิธีคิม2 มีอัตราความไม่สอดคล้องชนิดที่ 1 แตกต่างกันอย่างมีนัยสำคัญทางสถิติ ที่ระดับ .05 นอกนั้นไม่แตกต่างกัน

สำหรับอัตราความไม่สอดคล้องชนิดที่ 2 พบว่า ไม่แตกต่างกัน

ตาราง 60 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 300 คน ความยาวแบบทดสอบ 60 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC		0.35	0.00	1.52	3.93 ^{***}
ESA			0.35	1.17	3.68 ^{***}
EUA				1.52	3.93 ^{***}
CSA					2.74 ^{**}
CUA					

จากตาราง 60 จะเห็นว่า วิธีลอร์ด กับวิธีคิม2 วิธีราซู1 กับวิธีคิม2 วิธีราซู2 กับวิธีคิม2 มีอัตราความไม่สอดคล้องชนิดที่ 1 แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .001 ส่วนวิธีคิม1 กับวิธีคิม2 พบอัตราความไม่สอดคล้องชนิดที่ 1 แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 นอกนั้นไม่แตกต่างกัน

ตาราง 61 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 300 คน ความยาวแบบทดสอบ 60 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	1.02	1.25	0.22	0.22
ESA	0.64	-	0.24	0.79	1.23
EUA	0.64	0.00	-	1.03	1.47
CSA	0.59	1.49	1.49	-	0.44
CUA	0.59	1.49	1.49	0.00	-

จากตาราง 61 จะเห็นว่าอัตราความไม่สอดคล้องชนิดที่ 1 และชนิดที่ 2 ทั้ง 2 กรณีไม่พบความแตกต่าง

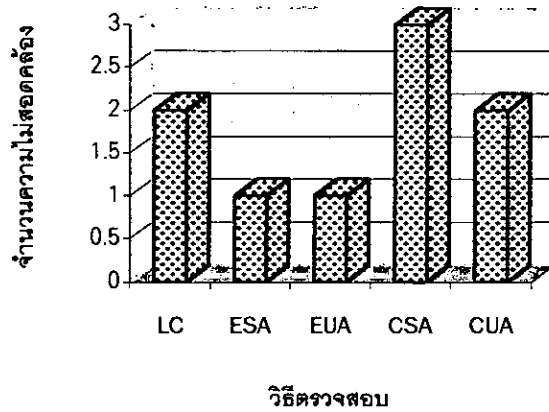
ตาราง 62 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 300 คน ความยาวแบบทดสอบ 60 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	0.91	1.12	0.91	0.39
ESA	1.08	-	0.21	0.00	1.30
EUA	0.97	0.46	-	0.21	1.51
CSA	1.26	0.39	0.84	-	1.30
CUA	1.64	0.78	1.03	0.39	-

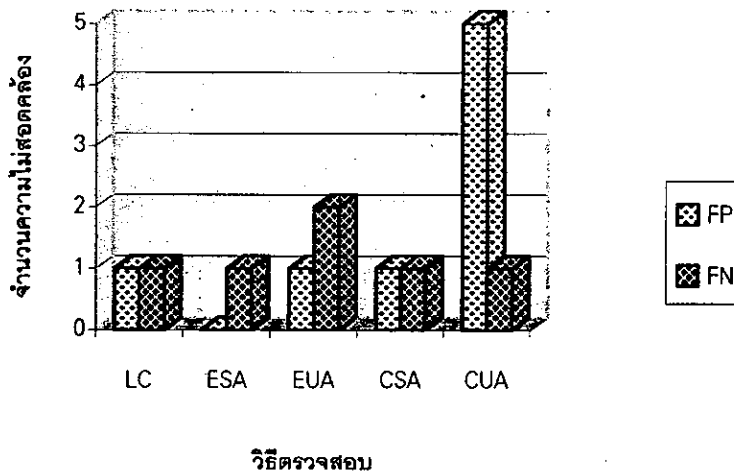
จากตาราง 62 จะเห็นว่าอัตราความไม่สอดคล้องชนิดที่ 1 และชนิดที่ 2 ทั้ง 2 กรณีไม่พบความแตกต่าง

ตาราง 63 จำนวนความไม่สอดคล้องของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 500 คน จำนวนข้อสอบ 20, 30, 40 และ 60 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน 0%, 10%, 20%

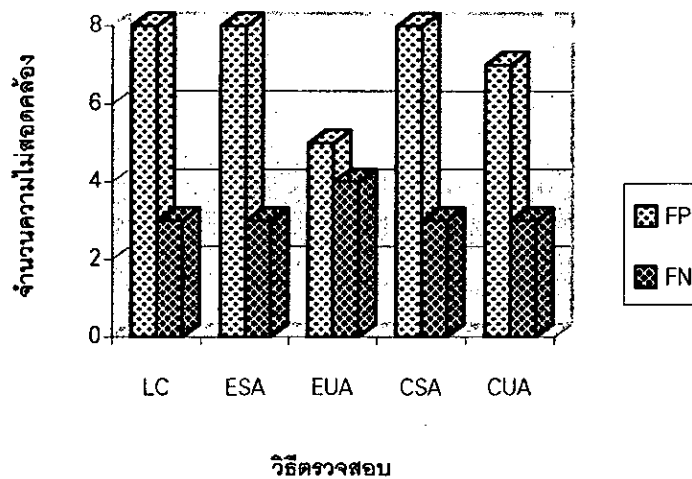
I % DIF	LC		ESA		EUA		CSA		CUA	
	I	II	I	II	I	II	I	II	I	II
I = 20										
0% DIF	2	-	1	-	1	-	3	-	2	-
10% DIF	1	1	0	1	1	2	1	1	5	1
20% DIF	8	3	8	3	5	4	8	3	7	3
I = 30										
0% DIF	2	-	2	-	1	-	2	-	2	-
10% DIF	8	1	9	2	8	2	6	1	4	1
20% DIF	11	3	11	4	8	3	10	3	9	3
I = 40										
0% DIF	2	-	3	-	3	-	3	-	3	-
10% DIF	11	2	11	3	9	3	8	2	9	3
20% DIF	14	4	13	5	11	4	12	4	13	5
I = 60										
0% DIF	8	-	8	-	7	-	7	-	8	-
10% DIF	21	4	20	4	17	3	19	4	20	4
20% DIF	24	5	21	5	17	6	22	5	23	6



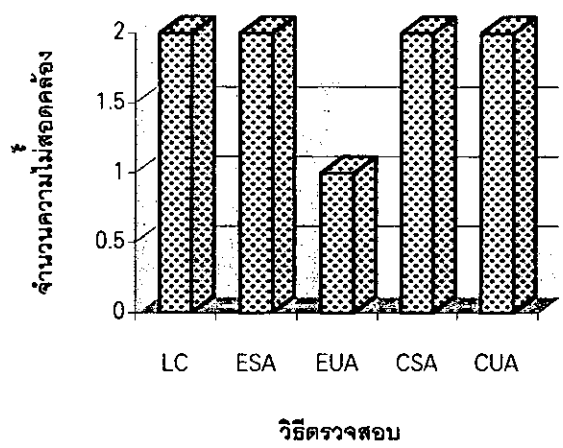
ภาพประกอบ 19 จำนวนความไม่สอดคล้อง เมื่อ $N = 500, I = 20, 0\%DIF$



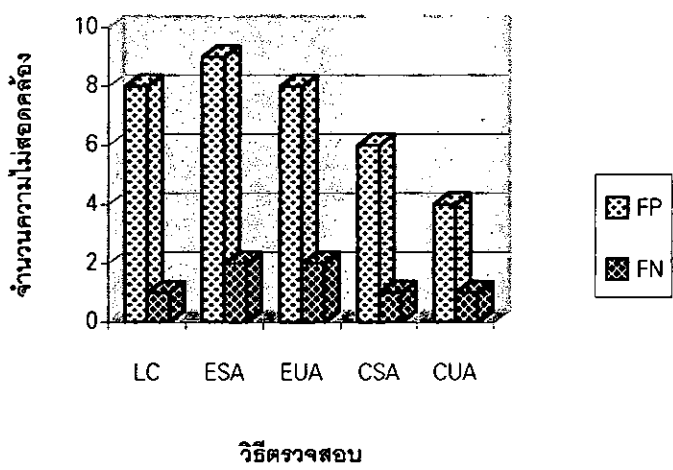
ภาพประกอบ 20 จำนวนความไม่สอดคล้อง เมื่อ $N = 500, I = 20, 10\%DIF$



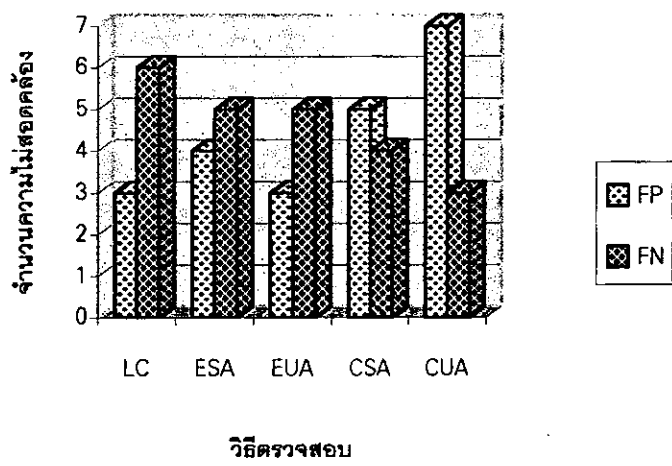
ภาพประกอบ 21 จำนวนความไม่สอดคล้อง เมื่อ $N = 500, I = 20, 20\%DIF$



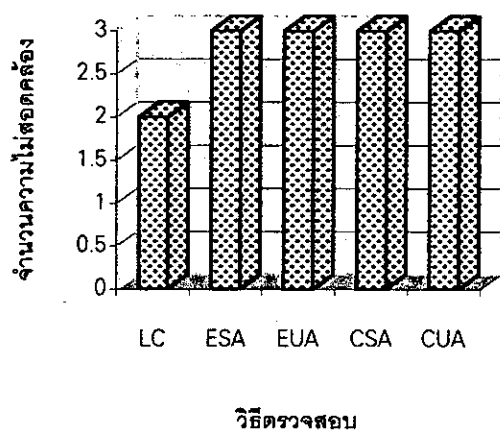
ภาพประกอบ 22 จำนวนความไม่สอดคล้อง เมื่อ $N = 500, I = 30, 0\%DIF$



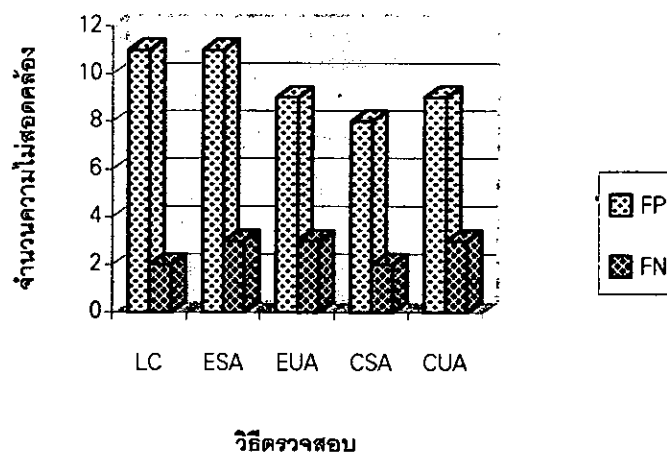
ภาพประกอบ 23 จำนวนความไม่สอดคล้อง เมื่อ $N = 500, I = 30, 10\%DIF$



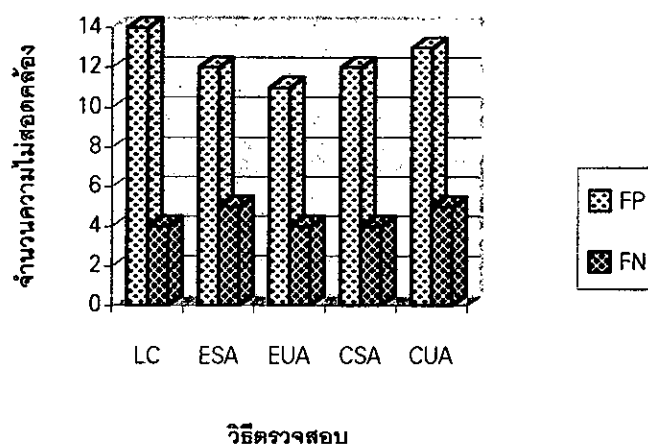
ภาพประกอบ 24 จำนวนความไม่สอดคล้อง เมื่อ $N = 500, I = 30, 20\%DIF$



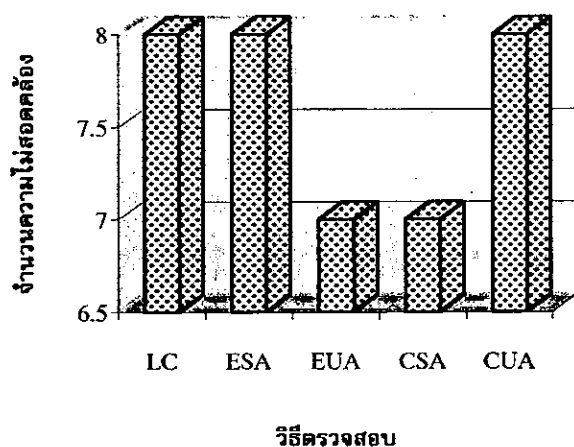
ภาพประกอบ 25 จำนวนความไม่สอดคล้อง เมื่อ $N = 500$, $I = 40$, 0%DIF



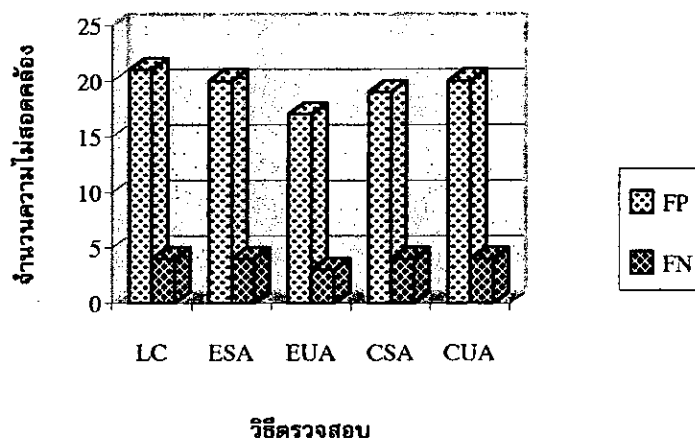
ภาพประกอบ 26 จำนวนความไม่สอดคล้อง เมื่อ $N = 500$, $I = 40$, 10%DIF



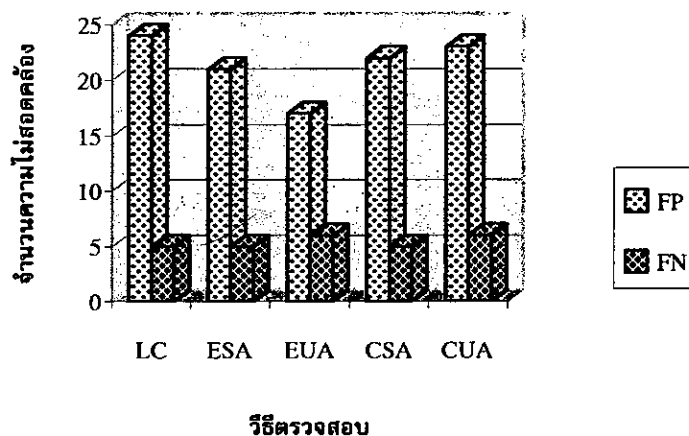
ภาพประกอบ 27 จำนวนความไม่สอดคล้อง เมื่อ $N = 500$, $I = 40$, 20%DIF



ภาพประกอบ 28 จำนวนความไม่สอดคล้อง เมื่อ $N = 500, I = 60, 0\%DIF$



ภาพประกอบ 29 จำนวนความไม่สอดคล้อง เมื่อ $N = 500, I = 60, 10\%DIF$



ภาพประกอบ 30 จำนวนความไม่สอดคล้อง เมื่อ $N = 500, I = 60, 20\%DIF$

เมื่อแบบทดสอบ 40 ข้อ สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0 ทุกวิธีมีความไม่สอดคล้องใกล้เคียงกัน (2 - 3) สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 วิธีที่มีความไม่สอดคล้องสูงสุด คือ วิธีลอร์ด และวิธีราชู 1 (11) นอกนั้นมีความไม่สอดคล้องใกล้เคียงกัน (8 - 9) และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20 ทุกวิธีมีความไม่สอดคล้องสูงใกล้เคียงกัน คือ วิธีลอร์ด (14) วิธีราชู 1 (13) วิธีคิม 2 (13) วิธีคิม 1 (12) และวิธีราชู 2 (11)

สำหรับจำนวนความไม่สอดคล้องชนิดที่ 2 เมื่อแบบทดสอบ 20 ข้อ สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 วิธีที่มีความไม่สอดคล้องสูงสุด คือ วิธีราชู2 (2) นอกนั้นมีความไม่สอดคล้องเท่ากัน (1) และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20 วิธีที่มีความไม่สอดคล้องสูงสุด คือ วิธีราชู2 (4) นอกนั้นมีความไม่สอดคล้องเท่ากัน (3)

เมื่อแบบทดสอบ 40 ข้อ สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน ร้อยละ 10 วิธีที่มีความไม่สอดคล้องสูงสุด คือ วิธีราชู1 วิธีราชู2 (3) และวิธีคิม2 (3) ส่วนวิธีลอร์ด และวิธีคิม1 มีความไม่สอดคล้องเท่ากัน (2) และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20 วิธีที่มีความไม่สอดคล้องสูงสุด คือ วิธีราชู1 (5) และวิธีคิม2 (5) นอกนั้นมี ความไม่สอดคล้องเท่ากัน (4)

เมื่อแบบทดสอบ 60 ข้อ สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 วิธีที่มีความไม่สอดคล้องต่ำสุด คือ วิธีราชู2 (3) นอกนั้นมีความไม่สอดคล้องเท่ากัน (4) และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20 ทุกวิธีมีความไม่สอดคล้องใกล้เคียงกัน (5 - 6)

ตาราง 64 อัตราความไม่สอดคล้องของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 500 คนความยาวแบบทดสอบ 20, 30, 40 และ 60 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน 0%, 10%, 20%

I % DIF	LC		ESA		EUA		CSA		CUA	
	I	II	I	II	I	II	I	II	I	II
I = 20										
0% DIF	.10	-	.05	-	.05	-	.15	-	.10	-
10% DIF	.06	.50	.00	.50	.06	1.00	.06	.50	.28	.50
20% DIF	.50	.75	.50	.75	.31	1.00	.50	.75	.44	.75
I = 30										
0% DIF	.07	-	.07	-	.03	-	.07	-	.07	-
10% DIF	.30	.33	.33	.67	.30	.67	.22	.33	.15	.33
20% DIF	.46	.50	.46	.67	.33	.50	.42	.50	.38	.50
I = 40										
0% DIF	.05	-	.08	-	.08	-	.08	-	.08	-
10% DIF	.31	.50	.31	.75	.25	.75	.22	.50	.25	.75
20% DIF	.44	.50	.41	.63	.34	.50	.38	.50	.41	.63
I = 60										
0% DIF	.13	-	.13	-	.12	-	.12	-	.13	-
10% DIF	.39	.67	.37	.67	.31	.50	.35	.67	.37	.67
20% DIF	.50	.42	.44	.42	.35	.50	.46	.42	.48	.50

จากตาราง 64 อัตราความไม่สอดคล้องชนิดที่ 1 เมื่อแบบทดสอบ 20 ข้อ สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน ร้อยละ 0 วิธีที่มีอัตราความไม่สอดคล้องสูงสุด คือ วิธีคิม1 (.15) และวิธีที่มีอัตราความไม่สอดคล้องต่ำสุด คือ วิธีราชู1 และวิธีราชู2 (.05) สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 วิธีที่มีอัตราความไม่สอดคล้องสูงสุด คือ วิธีคิม2 (.28) นอกนั้นมีความไม่สอดคล้องใกล้เคียงกัน (.00 - .06) และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20 วิธีที่มีอัตราความไม่สอดคล้องสูงสุด คือ วิธีลอร์ด วิธีราชู1 วิธีคิม1 (.50) รองลงมา คือ วิธีคิม2 (.44) ส่วนวิธีราชู2 มีอัตราความไม่สอดคล้องต่ำสุด (.31)

เมื่อแบบทดสอบ 30 ข้อ สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0 ทุกวิธีมีอัตราความไม่สอดคล้องเท่ากัน (.07) ยกเว้นวิธีราชู2 มีอัตราความไม่สอดคล้องต่ำสุด (.03) สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 วิธีที่มีอัตราความไม่สอดคล้องต่ำสุด คือ วิธีคิม2 (.15) และวิธีที่มีอัตราความไม่สอดคล้องสูงใกล้เคียงกัน คือ วิธีลอร์ด (.30) วิธีราชู1 (.33) และวิธีราชู2 (.30) และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20 วิธีที่มีอัตราความไม่สอดคล้องสูงสุด คือ วิธีลอร์ด และวิธีราชู1 (.46)

เมื่อแบบทดสอบ 40 ข้อ สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0 วิธีที่มีอัตราความไม่สอดคล้องต่ำสุด คือ วิธีลอร์ด (.05) นอกนั้นมีอัตราความไม่สอดคล้องเท่ากัน สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 วิธีที่มีอัตราความไม่สอดคล้องสูงสุด คือ วิธีลอร์ด และวิธีราชู1 (.31) นอกนั้นมีอัตราความไม่สอดคล้องใกล้เคียงกัน (.22 - .25) และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20 ทุกวิธีมีอัตราความไม่สอดคล้องสูงใกล้เคียงกัน คือ วิธีลอร์ด (.44) วิธีราชู1 (.41) วิธีคิม2 (.41) วิธีคิม1 (.38) และวิธีราชู2 (.34)

เมื่อแบบทดสอบ 60 ข้อ สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0 ทุกวิธีมีอัตราความไม่สอดคล้องสูงใกล้เคียงกัน (.12 - .13) สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 ทุกวิธีมีอัตราความไม่สอดคล้องสูงใกล้เคียงกัน (.37 - .39) ยกเว้นวิธีราชู2 มีอัตราความไม่สอดคล้องต่ำกว่าวิธีอื่น (.31) และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20 ทุกวิธีมีอัตราความไม่สอดคล้องสูงใกล้เคียงกัน (.44 - .50) ยกเว้น วิธีราชู2 มีอัตราความไม่สอดคล้องต่ำกว่าวิธีอื่น (.35)

สำหรับอัตราความไม่สอดคล้องชนิดที่ 2 จะเกิดมากขึ้น หรือน้อยลงในสัดส่วนใกล้เคียงกัน เมื่อสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันเพิ่มขึ้น

ตาราง 65 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 500 คน ความยาวแบบทดสอบ 20 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC		0.60	0.60	0.48	0.00
ESA			0.00	1.05	0.60
EUA				1.05	0.60
CSA					0.48
CUA					

จากตาราง 65 จะเห็นว่าอัตราความไม่สอดคล้องชนิดที่ 1 ไม่แตกต่างกัน

ตาราง 66 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 500 คน ความยาวแบบทดสอบ 20 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	0.65	0.00	0.00	1.78
ESA	0.00	-	0.65	0.65	2.44
EUA	1.15	1.15	-	0.00	1.78
CSA	0.00	0.00	1.15	-	1.78
CUA	0.00	0.00	1.15	0.00	-

จากตาราง 66 จะเห็นว่า วิธีราชู1 กับวิธีคิม2 มีอัตราความไม่สอดคล้องชนิดที่ 1 แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นอกนั้นไม่แตกต่างกัน
สำหรับอัตราความไม่สอดคล้องชนิดที่ 2 พบว่า ไม่แตกต่างกัน

ตาราง 67 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 500 คน ความยาวแบบทดสอบ 20 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	0.00	1.09	0.00	0.34
ESA	0.00	-	1.09	0.00	0.34
EUA	1.07	1.07	-	1.09	0.76
CSA	0.00	0.00	1.07	-	0.34
CUA	0.00	0.00	1.07	0.00	-

จากตาราง 67 จะเห็นว่าอัตราความไม่สอดคล้องชนิดที่ 1 และชนิดที่ 2 ทั้ง 2 กรณี ไม่พบความแตกต่าง

ตาราง 68 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 500 คน ความยาวแบบทดสอบ 30 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC		0.00	0.71	0.00	0.00
ESA			0.71	0.00	0.00
EUA				0.71	0.71
CSA					0.00
CUA					

จากตาราง 68 จะเห็นว่าอัตราความไม่สอดคล้องชนิดที่ 1 ไม่แตกต่างกัน

ตาราง 69 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 500 คน ความยาวแบบทดสอบ 30 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	0.24	0.00	0.67	1.33
ESA	0.83	-	0.24	0.90	0.69
EUA	0.83	0.00	-	0.67	1.33
CSA	0.00	0.83	0.83	-	0.66
CUA	0.00	0.83	0.83	0.00	-

จากตาราง 69 จะเห็นว่าอัตราความไม่สอดคล้องชนิดที่ 1 และชนิดที่ 2 ทั้ง 2 กรณีไม่พบความแตกต่าง

ตาราง 70 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 500 คน ความยาวแบบทดสอบ 30 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	0.00	0.92	0.28	0.56
ESA	0.84	-	0.92	0.28	0.56
EUA	0.00	0.84	-	0.65	0.36
CSA	0.00	0.84	0.00	-	0.28
CUA	0.00	0.84	0.00	0.00	-

จากตาราง 70 จะเห็นว่าอัตราความไม่สอดคล้องชนิดที่ 1 และชนิดที่ 2 ทั้ง 2 กรณี ไม่พบความแตกต่าง

ตาราง 71 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 500 คน ความยาวแบบทดสอบ 40 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC		0.55	0.55	0.55	0.55
ESA			0.00	0.00	0.00
EUA				0.00	0.00
CSA					0.00
CUA					

จากตาราง 71 จะเห็นว่าอัตราความไม่สอดคล้องชนิดที่ 1 ไม่แตกต่างกัน

ตาราง 72 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 500 คน ความยาวแบบทดสอบ 40 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	0.00	0.57	0.87	0.57
ESA	0.73	-	0.57	0.87	0.57
EUA	0.73	0.00	-	0.28	0.00
CSA	0.00	0.73	0.73	-	0.28
CUA	0.73	0.73	0.00	0.73	-

จากตาราง 72 จะเห็นว่าอัตราความไม่สอดคล้องชนิดที่ 1 และชนิดที่ 2 ทั้ง 2 กรณี ไม่พบความแตกต่าง

ตาราง 73 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่
ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 500 คน ความยาวแบบทดสอบ 40 ข้อ และสัดส่วน
จำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	0.24	0.82	0.49	0.24
ESA	0.52	-	0.58	0.25	0.00
EUA	0.00	0.52	-	0.33	0.58
CSA	0.00	0.52	0.00	-	0.25
CUA	0.52	0.00	0.52	0.52	-

จากตาราง 73 จะเห็นว่าอัตราความไม่สอดคล้องชนิดที่ 1 และชนิดที่ 2 ทั้ง 2 กรณี
ไม่พบความแตกต่าง

ตาราง 74 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่
ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 500 คน ความยาวแบบทดสอบ 60 ข้อ และสัดส่วน
จำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC		0.00	0.17	0.17	0.00
ESA			0.17	0.17	0.00
EUA				5.00	0.17
CSA					0.17
CUA					

จากตาราง 74 จะเห็นว่าอัตราความไม่สอดคล้องชนิดที่ 1 ไม่แตกต่างกัน

ตาราง 75 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่
ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 500 คน ความยาวแบบทดสอบ 60 ข้อ และสัดส่วน
จำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	0.21	0.87	0.43	0.21
ESA	0.00	-	0.66	0.22	0.00
EUA	0.59	0.59	-	0.44	0.66
CSA	0.00	0.00	0.59	-	0.22
CUA	0.00	0.00	0.59	0.00	-

จากตาราง 75 จะเห็นว่าอัตราความไม่สอดคล้องชนิดที่ 1 และชนิดที่ 2 ทั้ง 2 กรณี ไม่พบความแตกต่าง

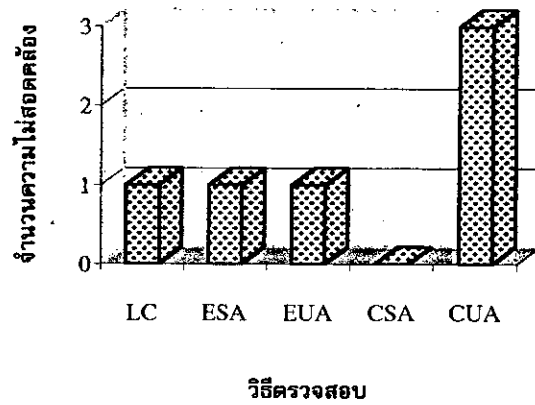
ตาราง 76 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 500 คน ความยาวแบบทดสอบ 60 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	0.59	1.49	0.39	0.20
ESA	0.00	-	0.50	0.20	0.39
EUA	0.39	0.39	-	1.09	1.29
CSA	0.00	0.00	0.39	-	0.20
CUA	0.39	0.39	0.00	0.39	-

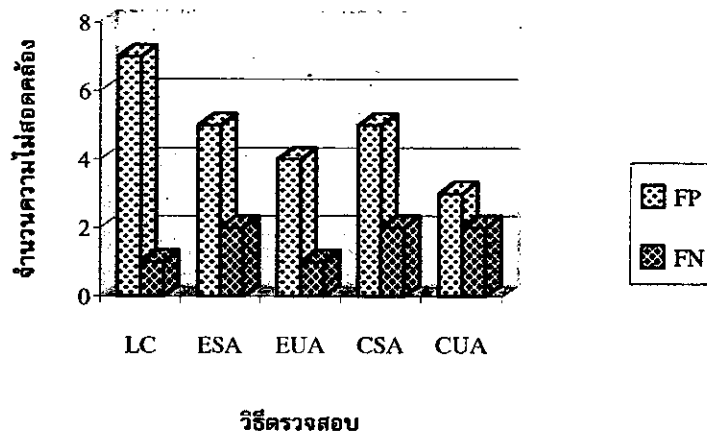
จากตาราง 76 จะเห็นว่าอัตราความไม่สอดคล้องชนิดที่ 1 และชนิดที่ 2 ทั้ง 2 กรณี ไม่พบความแตกต่าง

ตาราง 77 จำนวนความไม่สอดคล้องของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 700 คน ความยาวแบบทดสอบ 20, 30, 40 และ 60 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน 0%, 10%, 20%

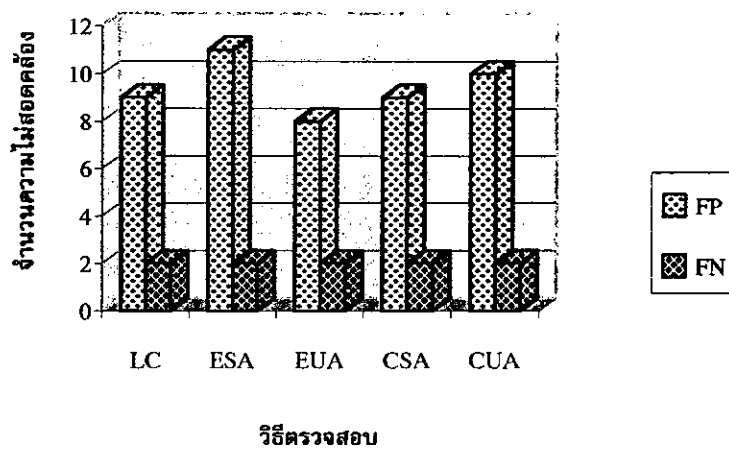
I % DIF	LC		ESA		EUA		CSA		CUA	
	I	II	I	II	I	II	I	II	I	II
I = 20										
0% DIF	1	-	1	-	1	-	0	-	3	-
10% DIF	7	1	5	2	4	1	5	2	3	2
20% DIF	9	2	11	2	8	2	9	2	10	2
I = 30										
0% DIF	2	-	3	-	3	-	1	-	2	-
10% DIF	10	1	10	1	9	1	8	1	7	1
20% DIF	13	2	13	3	12	3	13	2	11	2
I = 40										
0% DIF	3	-	3	-	2	-	1	-	3	-
10% DIF	9	3	10	4	5	4	4	4	3	3
20% DIF	11	4	12	5	8	6	7	5	6	6
I = 60										
0% DIF	5	-	7	-	6	-	1	-	6	-
10% DIF	15	3	17	2	18	4	8	4	9	5
20% DIF	24	5	20	5	21	7	18	5	19	6



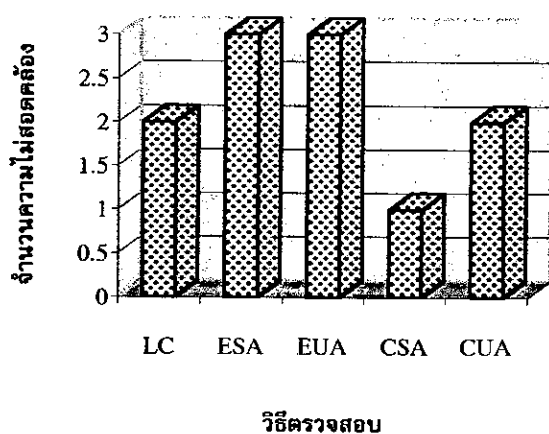
ภาพประกอบ 31 จำนวนความไม่สอดคล้อง เมื่อ $N = 700, I = 20, 0\%DIF$



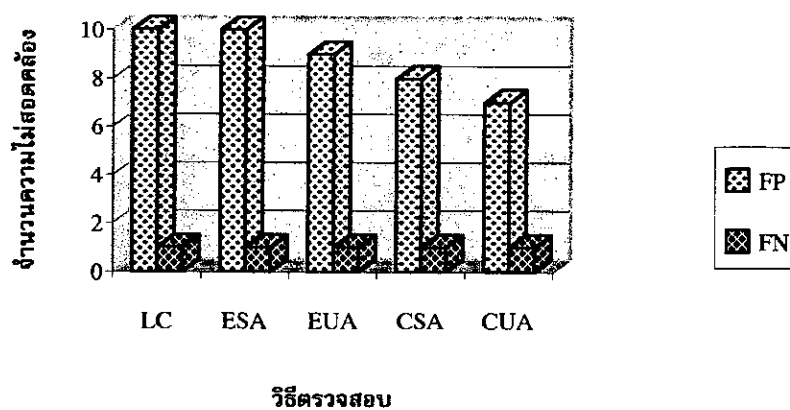
ภาพประกอบ 32 จำนวนความไม่สอดคล้อง เมื่อ $N = 700, I = 20, 10\%DIF$



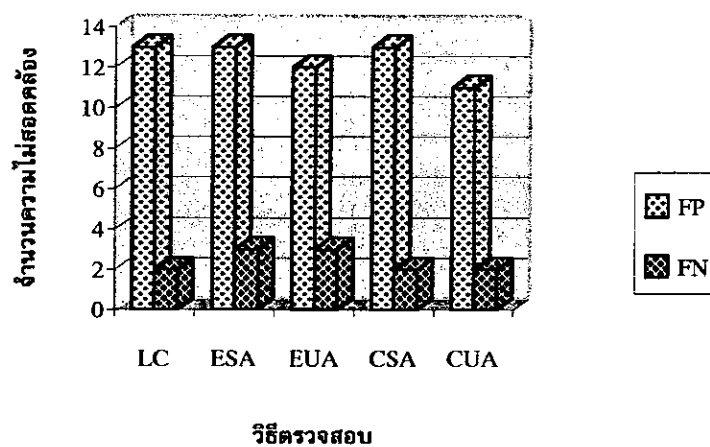
ภาพประกอบ 33 จำนวนความไม่สอดคล้อง เมื่อ $N = 700, I = 20, 20\%DIF$



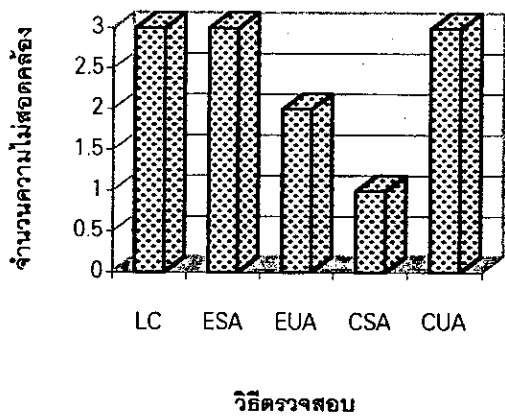
ภาพประกอบ 34 จำนวนความไม่สอดคล้อง เมื่อ $N = 700, I = 30, 0\%DIF$



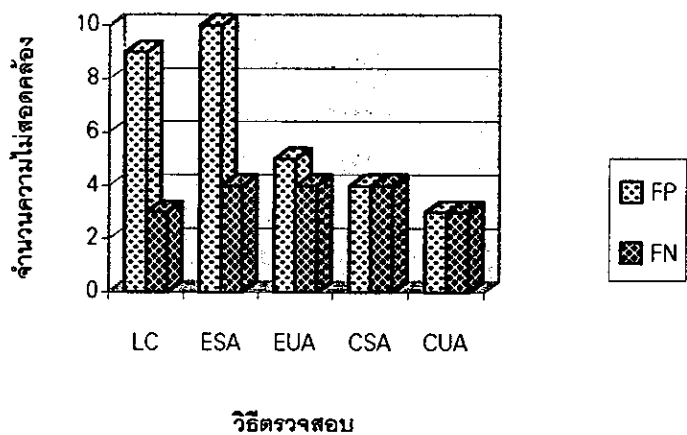
ภาพประกอบ 35 จำนวนความไม่สอดคล้อง เมื่อ $N = 700, I = 30, 10\%DIF$



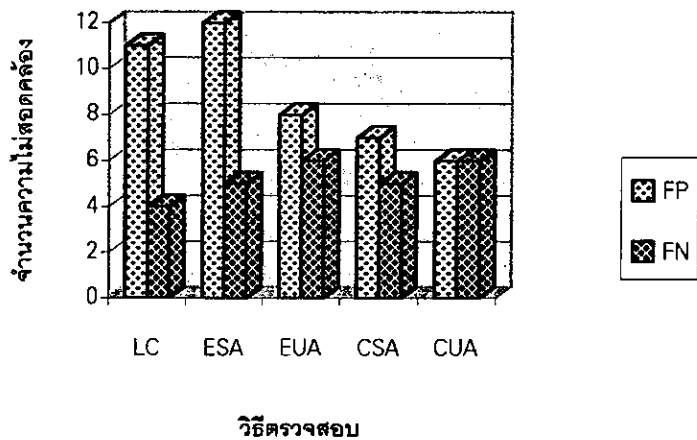
ภาพประกอบ 36 จำนวนความไม่สอดคล้อง เมื่อ $N = 700, I = 30, 20\%DIF$



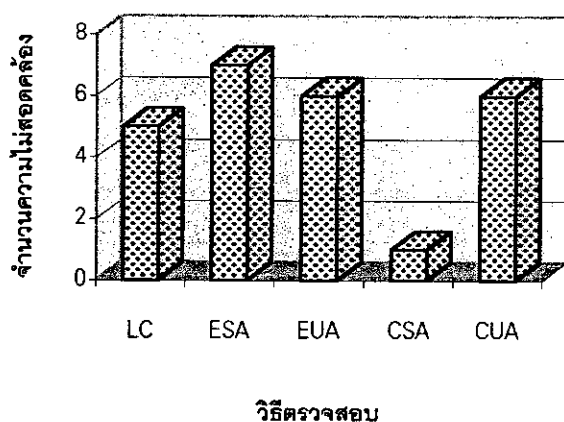
ภาพประกอบ 37 จำนวนความไม่สอดคล้อง เมื่อ $N = 700, I = 40, 0\%DIF$



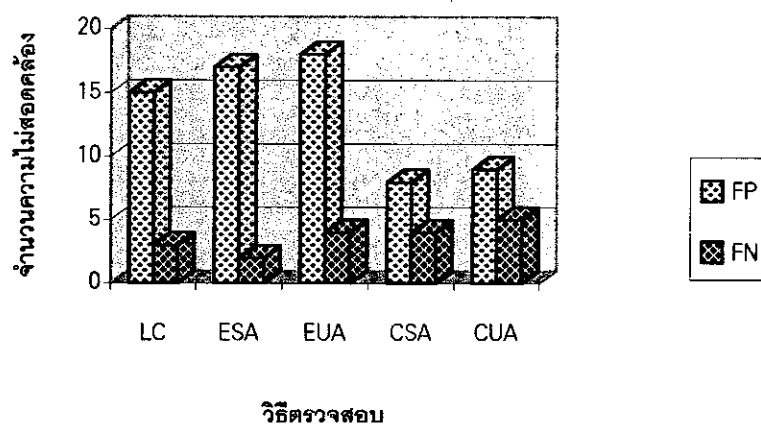
ภาพประกอบ 38 จำนวนความไม่สอดคล้อง เมื่อ $N = 700, I = 40, 10\%DIF$



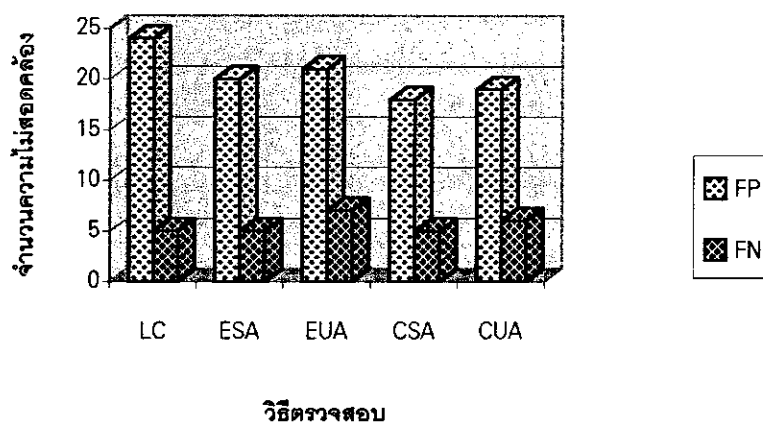
ภาพประกอบ 39 จำนวนความไม่สอดคล้อง เมื่อ $N = 700, I = 40, 20\%DIF$



ภาพประกอบ 40 จำนวนความไม่สอดคล้อง เมื่อ $N = 700, I = 60, 0\%DIF$



ภาพประกอบ 41 จำนวนความไม่สอดคล้อง เมื่อ $N = 700, I = 60, 10\%DIF$



ภาพประกอบ 42 จำนวนความไม่สอดคล้อง เมื่อ $N = 700, I = 60, 20\%DIF$

เมื่อแบบทดสอบ 60 ข้อ สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน ร้อยละ 10 วิธีที่มีความไม่สอดคล้องต่ำสุด คือ วิธีราชู1 (2) และวิธีที่มีความไม่สอดคล้องสูงสุด คือ วิธีคิม2 (5) นอกนั้นมีความไม่สอดคล้องใกล้เคียงกัน (3 - 4) และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน ร้อยละ 20 วิธีที่มีความไม่สอดคล้องสูงสุด คือ วิธีราชู2 (7) นอกนั้นมีความไม่สอดคล้องใกล้เคียงกัน (5 - 6)

ตาราง 78 อัตราความไม่สอดคล้องของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อ
กลุ่มตัวอย่าง 700 คน ความยาวแบบทดสอบ 20, 30, 40 และ 60 ข้อ และสัดส่วนจำนวน
ข้อสอบที่ทำหน้าที่ต่างกัน 0%, 10%, 20%

I % DIF	LC		ESA		EUA		CSA		CUA	
	I	II	I	II	I	II	I	II	I	II
I = 20										
0% DIF	.05	-	.05	-	.05	-	.00	-	.15	-
10% DIF	.39	.50	.28	1.00	.22	.50	.28	1.00	.17	1.00
20% DIF	.56	.50	.69	.50	.50	.50	.56	.50	.63	.50
I = 30										
0% DIF	.07	-	.10	-	.10	-	.03	-	.07	-
10% DIF	.37	.33	.37	.33	.33	.33	.30	.33	.26	.33
20% DIF	.54	.33	.54	.50	.50	.50	.54	.33	.46	.33
I = 40										
0% DIF	.08	-	.08	-	.05	-	.03	-	.08	-
10% DIF	.25	.75	.27	1.00	.14	1.00	.11	1.00	.08	.75
20% DIF	.34	.50	.38	.63	.25	.75	.22	.63	.19	.75
I = 60										
0% DIF	.08	-	.12	-	.10	-	.02	-	.10	-
10% DIF	.28	.50	.31	.33	.33	.67	.15	.67	.17	.83
20% DIF	.50	.42	.42	.42	.44	.58	.38	.42	.40	.50

จากตาราง 78 อัตราความไม่สอดคล้องชนิดที่ 1 เมื่อแบบทดสอบ 20 ข้อ สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0 วิธีที่มีอัตราความไม่สอดคล้องสูงสุด คือ วิธีคิม2 (.15) นอกนั้นไม่มีอัตราความไม่สอดคล้องต่ำมาก (.00) และไม่มีอัตราความไม่สอดคล้องเลย คือ วิธีคิม1 สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 วิธีที่มีอัตราความไม่สอดคล้องสูงสุดคือ วิธีลอร์ด (.39) รองลงมา คือ วิธีราชู1 และวิธีคิม1 (.28) ส่วนวิธีคิม2 (.17) และวิธีราชู2 (.22) มีอัตราความไม่สอดคล้องใกล้เคียงกัน และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20 วิธีที่

มีอัตราความไม่สอดคล้องสูงสุด คือ วิธีราชู1 (.69) รองลงมา คือ วิธีคิม2 (.63) วิธีลอร์ด (.56) และวิธีคิม1 (.56)

เมื่อแบบทดสอบ 30 ข้อ สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0 วิธีคิม1 (.03) มีอัตราความไม่สอดคล้องต่ำสุด นอกนั้นมีความไม่สอดคล้องใกล้เคียงกัน (.07 - .10) สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 วิธีที่มีอัตราความไม่สอดคล้องต่ำสุด คือ วิธีคิม2 (.26) นอกนั้นมีความไม่สอดคล้องใกล้เคียงกัน (.30 - .37) และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20 ทุกวิธีมีอัตราความไม่สอดคล้องใกล้เคียงกัน (.46 - .54)

เมื่อแบบทดสอบ 40 ข้อ สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0 วิธีที่มีอัตราความไม่สอดคล้องต่ำสุด คือ วิธีคิม1 (.03) นอกนั้นมีความไม่สอดคล้องใกล้เคียงกัน (.05 - .08) สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 วิธีที่มีอัตราความไม่สอดคล้องต่ำสุด คือ วิธีคิม2 (.08) วิธีที่มีอัตราความไม่สอดคล้องสูงใกล้เคียงกัน คือ วิธีลอร์ด (.25) และวิธีราชู1 (.27) และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20 วิธีที่มีอัตราความไม่สอดคล้องต่ำสุด คือ วิธีคิม1 (.19) วิธีที่มีอัตราความไม่สอดคล้องใกล้เคียงกัน คือ วิธีลอร์ด (.34) และวิธีราชู1 (.38)

เมื่อแบบทดสอบ 60 ข้อ สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0 วิธีที่มีอัตราความไม่สอดคล้องต่ำสุด คือ วิธีคิม1 (.02) นอกนั้นมีความไม่สอดคล้องใกล้เคียงกัน (.08 - .12) สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 วิธีที่มีอัตราความไม่สอดคล้องต่ำใกล้เคียงกัน คือ วิธีคิม1 (.15) และวิธีคิม2 (.17) นอกนั้นมีความไม่สอดคล้องสูงใกล้เคียงกัน คือ วิธีราชู2 (.33) วิธีราชู1 (.31) และวิธีลอร์ด (.28) และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20 วิธีที่มีอัตราความไม่สอดคล้องสูงสุด คือ วิธีลอร์ด (.50) นอกนั้นมีความไม่สอดคล้องใกล้เคียงกัน (.38 - .44)

สำหรับอัตราความไม่สอดคล้องชนิดที่ 2 ส่วนใหญ่จะลดลง เมื่อสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันเพิ่มขึ้น

ตาราง 79 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 700 คน ความยาวแบบทดสอบ 20 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC		0.00	0.00	1.01	1.05
ESA			0.00	1.01	1.05
EUA				1.01	1.05
CSA					1.80
CUA					

จากตาราง 79 จะเห็นว่าอัตราความไม่สอดคล้องชนิดที่ 1 ไม่แตกต่างกัน

ตาราง 80 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 700 คน ความยาวแบบทดสอบ 20 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	0.71	1.11	0.71	1.47
ESA	1.15	-	0.42	0.00	0.79
EUA	0.00	1.15	-	0.42	0.38
CSA	1.15	0.00	1.15	-	0.79
CUA	1.15	0.00	1.15	0.00	-

จากตาราง 80 จะเห็นว่าอัตราความไม่สอดคล้องชนิดที่ 1 และชนิดที่ 2 ทั้ง 2 กรณี ไม่พบความแตกต่าง

ตาราง 81 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 700 คน ความยาวแบบทดสอบ 20 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	0.76	0.34	0.00	0.40
ESA	0.00	-	1.09	0.76	0.36
EUA	0.00	0.00	-	0.34	0.74
CSA	0.00	0.00	0.00	-	0.40
CUA	0.00	0.00	0.00	0.00	-

จากตาราง 81 จะเห็นว่าอัตราความไม่สอดคล้องชนิดที่ 1 และชนิดที่ 2 ทั้ง 2 กรณี ไม่พบความแตกต่าง

ตาราง 82 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 700 คน ความยาวแบบทดสอบ 30 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	0.42	0.42	0.71	0.00
ESA		-	0.00	1.09	0.42
EUA			-	1.09	0.42
CSA				-	0.71
CUA					-

จากตาราง 82 จะเห็นว่าอัตราความไม่สอดคล้องชนิดที่ 1 ไม่แตกต่างกัน

ตาราง 83 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 700 คน ความยาวแบบทดสอบ 30 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	0.00	0.31	0.55	0.87
ESA	0.00	-	0.31	0.55	0.87
EUA	0.00	0.00	-	0.23	0.56
CSA	0.00	0.00	0.00	-	0.33
CUA	0.00	0.00	0.00	0.00	-

จากตาราง 83 จะเห็นว่าอัตราความไม่สอดคล้องชนิดที่ 1 และชนิดที่ 2 ทั้ง 2 กรณีไม่พบความแตกต่าง

ตาราง 84 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 700 คน ความยาวแบบทดสอบ 30 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	0.00	0.28	0.00	0.55
ESA	0.59	-	0.28	0.00	0.55
EUA	0.59	0.00	-	0.28	0.28
CSA	0.00	0.59	0.59	-	0.55
CUA	0.00	0.59	0.59	0.00	-

จากตาราง 84 จะเห็นว่าอัตราความไม่สอดคล้องชนิดที่ 1 และชนิดที่ 2 ทั้ง 2 กรณีไม่พบความแตกต่าง

ตาราง 85 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่
ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 700 คน ความยาวแบบทดสอบ 40 ข้อ และสัดส่วน
จำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC		0.00	0.55	1.03	0.00
ESA			0.55	1.03	0.00
EUA				0.47	0.55
CSA					1.03
CUA					

จากตาราง 85 จะเห็นว่าอัตราความไม่สอดคล้องชนิดที่ 1 ไม่แตกต่างกัน

ตาราง 86 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่
ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 700 คน ความยาวแบบทดสอบ 40 ข้อ และสัดส่วน
จำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	0.19	1.18	1.55	1.94
ESA	1.07	-	1.36	1.72	2.11
EUA	1.07	0.00	-	0.39	0.81
CSA	1.07	0.00	0.00	-	0.43
CUA	0.00	1.07	1.07	1.07	-

จากตาราง 86 จะเห็นว่าอัตราความไม่สอดคล้องชนิดที่ 1 ของวิธีราชู1 กับวิธีคิม2
แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นอกนั้นไม่แตกต่างกัน
สำหรับอัตราความไม่สอดคล้องชนิดที่ 2 พบว่าไม่แตกต่างกัน

ตาราง 87 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 700 คน ความยาวแบบทดสอบ 40 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	0.33	0.82	1.07	1.36
ESA	0.52	-	1.12	1.45	1.69
EUA	1.03	0.52	-	0.28	0.58
CSA	0.52	0.00	0.52	-	0.29
CUA	1.03	0.52	0.00	0.52	-

จากตาราง 87 จะเห็นว่าอัตราความไม่สอดคล้องชนิดที่ 1 และชนิดที่ 2 ทั้ง 2 กรณีไม่พบความแตกต่าง

ตาราง 88 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 700 คน ความยาวแบบทดสอบ 60 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC		0.73	0.38	1.52	0.38
ESA			0.35	2.22	0.35
EUA				1.88	0.00
CSA					1.88
CUA					

จากตาราง 88 จะเห็นว่า วิธีราชู1 กับวิธีคิม1 เกิดอัตราความไม่สอดคล้องชนิดที่ 1 แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นอกนั้นไม่แตกต่างกัน

ตาราง 89 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 700 คน ความยาวแบบทดสอบ 60 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	0.34	0.56	1.65	1.38
ESA	0.59	-	0.22	1.97	1.71
EUA	0.59	1.18	-	2.19	1.92
CSA	0.59	1.18	0.00	-	0.29
CUA	1.21	1.76	0.68	0.68	-

จากตาราง 89 จะเห็นว่า วิธีราชู1 กับวิธีคิม1 และวิธีราชู2 กับวิธีคิม1 มีอัตราความไม่สอดคล้องชนิดที่ 1 แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05

สำหรับอัตราความไม่สอดคล้องในชนิดที่ 2 พบว่าไม่แตกต่างกัน

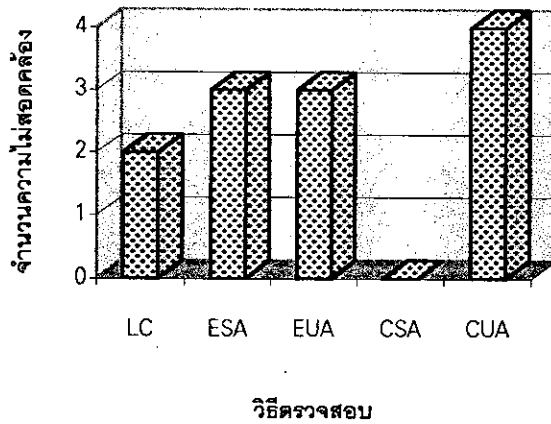
ตาราง 90 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 700 คน ความยาวแบบทดสอบ 60 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	0.75	0.59	1.17	0.98
ESA	0.00	-	0.20	0.40	0.20
EUA	0.78	0.78	-	0.60	0.39
CSA	0.00	0.00	0.78	-	0.20
CUA	0.39	0.39	0.39	0.39	-

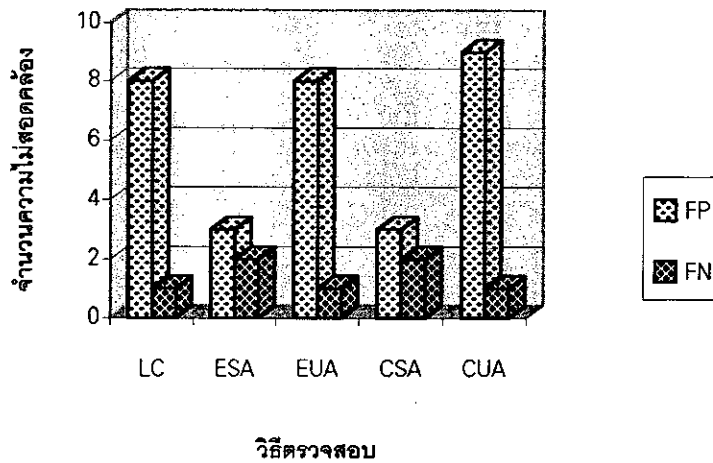
จากตาราง 90 จะเห็นว่า อัตราความไม่สอดคล้องชนิดที่ 1 และชนิดที่ 2 ทั้ง 2 กรณีไม่พบความแตกต่าง

ตาราง 91 จำนวนความไม่สอดคล้องของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อ
กลุ่มตัวอย่าง 1,000 คน ความยาวแบบทดสอบ 20, 30, 40 และ 60 ข้อ และ
สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน 0%, 10%, 20%

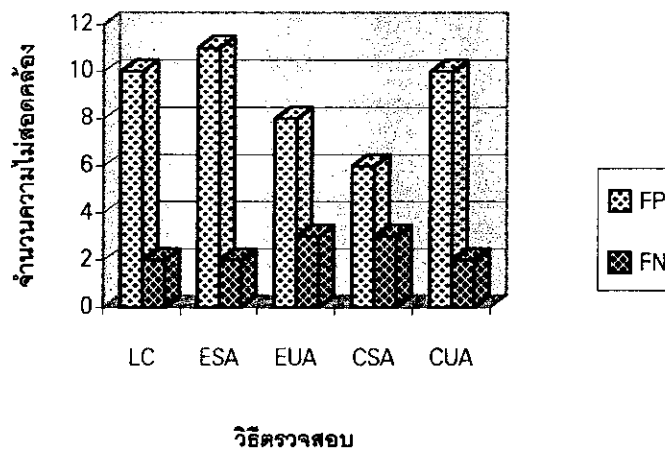
I % DIF	LC		ESA		EUA		CSA		CUA	
	I	II	I	II	I	II	I	II	I	II
I = 20										
0% DIF	2	-	3	-	3	-	0	-	4	-
10% DIF	8	1	3	2	8	1	3	2	9	1
20% DIF	10	2	11	2	8	3	6	3	10	2
I = 30										
0% DIF	3	-	1	-	3	-	0	-	2	-
10% DIF	12	0	10	2	11	2	8	3	6	1
20% DIF	14	1	14	5	13	2	8	5	9	3
I = 40										
0% DIF	4	-	3	-	2	-	0	-	2	-
10% DIF	12	2	12	3	9	2	7	3	9	2
20% DIF	16	3	15	4	13	3	10	5	12	5
I = 60										
0% DIF	14	-	13	-	10	-	2	-	3	-
10% DIF	16	3	15	4	11	4	3	5	3	4
20% DIF	20	7	19	6	20	6	8	7	19	6



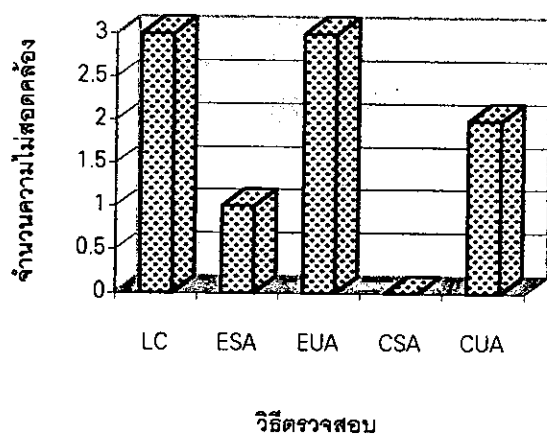
ภาพประกอบ 43 จำนวนความไม่สอดคล้อง เมื่อ $N = 1,000$, $I = 20$, $0\%DIF$



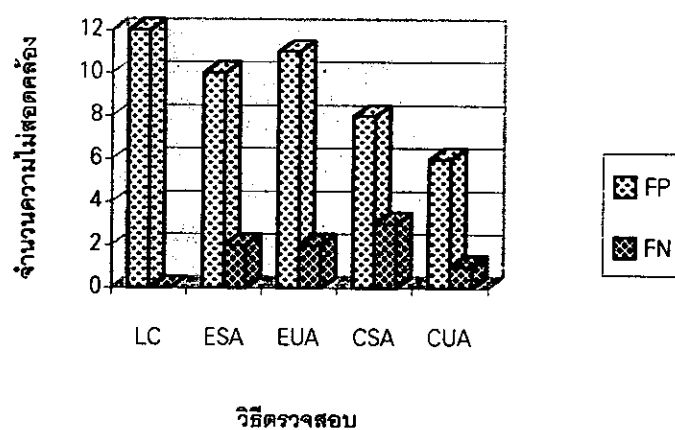
ภาพประกอบ 44 จำนวนความไม่สอดคล้อง เมื่อ $N = 1,000$, $I = 20$, $10\%DIF$



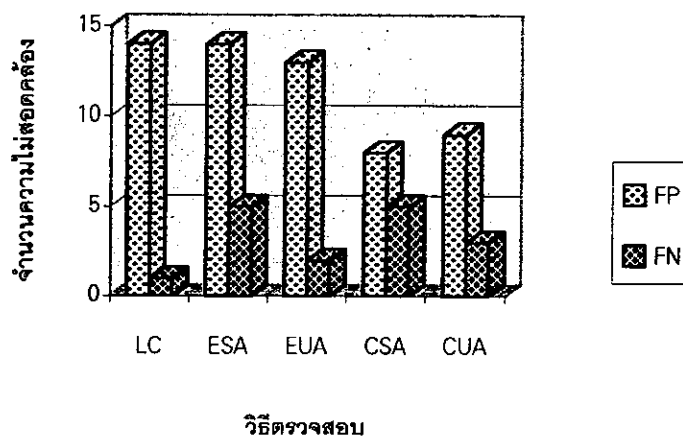
ภาพประกอบ 45 จำนวนความไม่สอดคล้อง เมื่อ $N = 1,000$, $I = 20$, $20\%DIF$



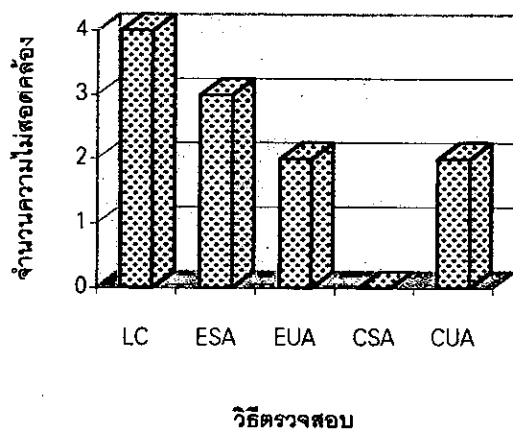
ภาพประกอบ 46 จำนวนความไม่สอดคล้อง เมื่อ $N = 1,000$, $I = 30$, 0%DIF



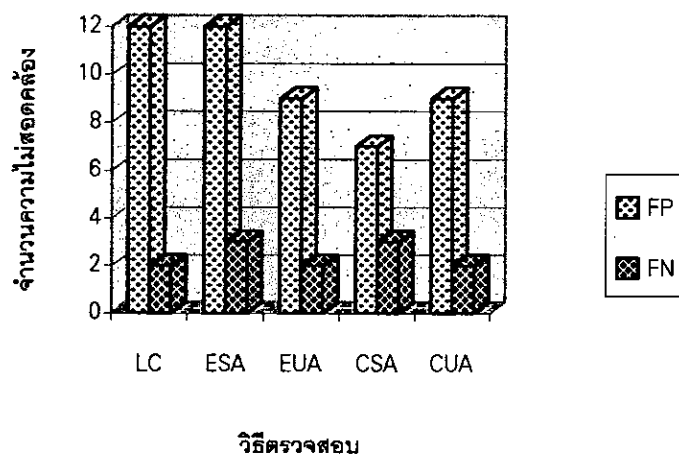
ภาพประกอบ 47 จำนวนความไม่สอดคล้อง เมื่อ $N = 1,000$, $I = 30$, 10%DIF



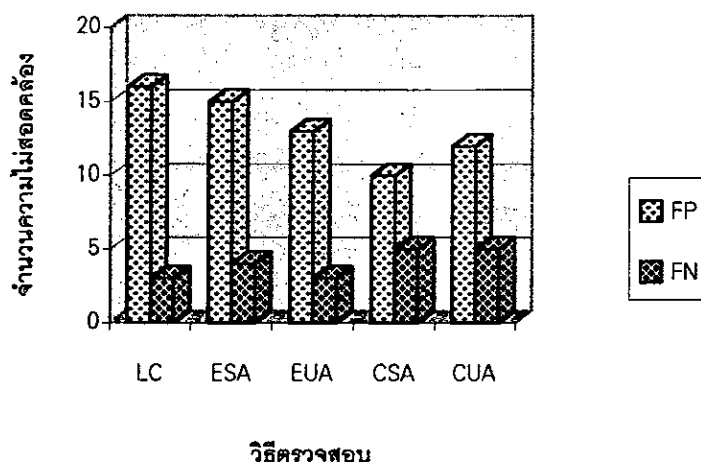
ภาพประกอบ 48 จำนวนความไม่สอดคล้อง เมื่อ $N = 1,000$, $I = 30$, 20%DIF



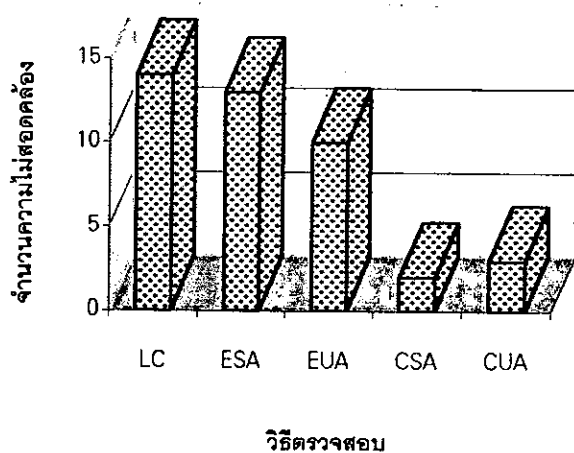
ภาพประกอบ 49 จำนวนความไม่สอดคล้อง เมื่อ $N = 1,000$, $I = 40$, 0%DIF



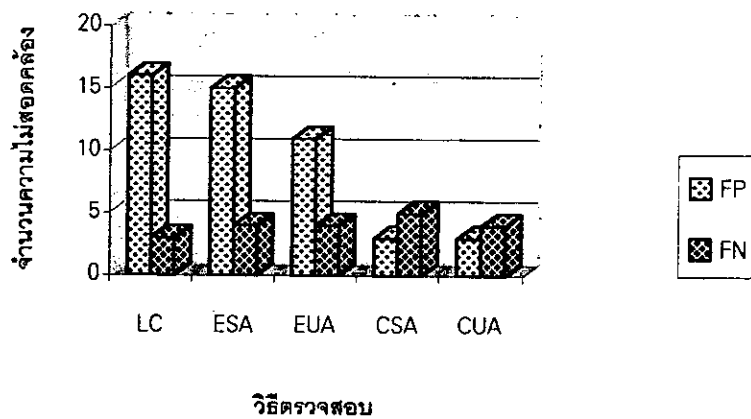
ภาพประกอบ 50 จำนวนความไม่สอดคล้อง เมื่อ $N = 1,000$, $I = 40$, 10%DIF



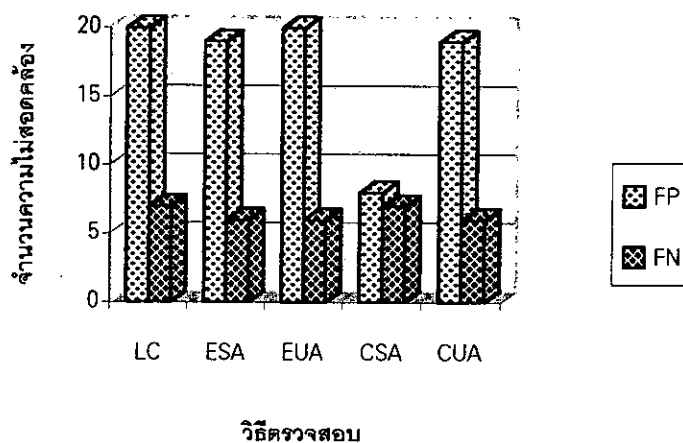
ภาพประกอบ 51 จำนวนความไม่สอดคล้อง เมื่อ $N = 1,000$, $I = 40$, 20%DIF



ภาพประกอบ 52 จำนวนความไม่สอดคล้อง เมื่อ $N = 1,000$, $I = 60$, 0%DIF



ภาพประกอบ 53 จำนวนความไม่สอดคล้อง เมื่อ $N = 1,000$, $I = 60$, 10%DIF



ภาพประกอบ 54 จำนวนความไม่สอดคล้อง เมื่อ $N = 1,000$, $I = 60$, 20%DIF

เมื่อแบบทดสอบ 40 ข้อ สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0 วิธีที่มีความไม่สอดคล้องสูงใกล้เคียงกัน คือ วิธีลอร์ด (4) และวิธีราชู1 (3) ส่วนวิธีที่ไม่มี ความไม่สอดคล้องเลย คือ วิธีคิม1 สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 วิธีที่มีความไม่สอดคล้องสูงใกล้เคียงกัน คือ วิธีลอร์ด (12) และวิธีราชู1 (12) ส่วนวิธีคิม1 มีความไม่สอดคล้องต่ำกว่าวิธีอื่น (7) และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20 วิธีที่มีความไม่สอดคล้องสูงใกล้เคียงกัน คือ วิธีลอร์ด (16) และวิธีราชู1 (15) ส่วนวิธีคิม1 มีความไม่สอดคล้องต่ำสุด (10)

ส่วนจำนวนความไม่สอดคล้องชนิดที่ 2 เมื่อแบบทดสอบ 20 ข้อ สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 วิธีที่มีความไม่สอดคล้องสูงสุด คือ วิธีราชู1 (2) และวิธีคิม1 (2) นอกนั้นมีความไม่สอดคล้องเท่ากัน (1) และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20 วิธีที่มีความไม่สอดคล้องสูงสุด คือ วิธีราชู2 (3) และวิธีคิม1 (3) นอกนั้นมีความไม่สอดคล้องเท่ากัน (2)

เมื่อแบบทดสอบ 30 ข้อ สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 วิธีที่ไม่มีความไม่สอดคล้อง คือ วิธีลอร์ด และวิธีที่มีความไม่สอดคล้องสูงสุด คือ วิธีคิม1 (3) และ สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20 วิธีที่มีความไม่สอดคล้องสูงสุด คือ วิธีราชู1 (5) และวิธีคิม1 (5) ส่วนวิธีที่มีความไม่สอดคล้องต่ำสุด คือ วิธีลอร์ด (1)

เมื่อแบบทดสอบ 40 ข้อ สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 วิธีที่มีความไม่สอดคล้องสูงสุด คือ วิธีราชู1 (3) และวิธีคิม1 (3) นอกนั้นมีความไม่สอดคล้องเท่ากัน (2) และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20 วิธีที่มีความไม่สอดคล้องสูงสุด คือ วิธีคิม1 (5) และวิธีคิม2 (5) นอกนั้นมีความไม่สอดคล้องใกล้เคียงกัน (3 - 4)

เมื่อแบบทดสอบ 60 ข้อ สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 วิธีที่มีความไม่สอดคล้องสูงสุด คือ วิธีคิม1 (5) นอกนั้นมีความไม่สอดคล้องใกล้เคียงกัน (3 - 4) และ สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20 ทุกวิธีมีความไม่สอดคล้องใกล้เคียงกัน (6 - 7)

ตาราง 92 อัตราความไม่สอดคล้องของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 1,000 คน ความยาวแบบทดสอบ 20, 30, 40 และ 60 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน 0%, 10%, 20%

I	LC		ESA		EUA		CSA		CUA	
	I	II	I	II	I	II	I	II	I	II
I = 20										
0% DIF	.10	-	.15	-	.15	-	.00	-	.20	-
10% DIF	.44	.50	.17	1.00	.44	.50	.17	1.00	.50	.50
20% DIF	.63	.50	.69	.50	.50	.75	.38	.75	.63	.50
I = 30										
0% DIF	.10	-	.03	-	.10	-	.00	-	.07	-
10% DIF	.44	.00	.37	.67	.41	.67	.30	1.00	.22	.33
20% DIF	.58	.17	.58	.83	.54	.33	.33	.83	.38	.50
I = 40										
0% DIF	.10	-	.08	-	.05	-	.00	-	.05	-
10% DIF	.33	.50	.33	.75	.25	.50	.19	.75	.25	.50
20% DIF	.50	.38	.47	.50	.41	.38	.31	.63	.38	.63
I = 60										
0% DIF	.23	-	.22	-	.17	-	.03	-	.05	-
10% DIF	.30	.50	.28	.67	.20	.67	.06	.83	.06	.67
20% DIF	.42	.58	.40	.50	.42	.50	.17	.58	.40	.50

ตาราง 93 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 1,000 คน ความยาวแบบทดสอบ 20 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC		0.48	0.48	1.45	0.89
ESA			0.00	1.80	0.42
EUA				1.80	0.42
CSA					2.11
CUA					

จากตาราง 93 จะเห็นว่า วิธีคิม1 กับวิธีคิม2 มีอัตราความไม่สอดคล้องชนิดที่ 1 แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นอกนั้นไม่แตกต่างกัน

ตาราง 94 ผลการทดสอบความแตกต่างของอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่าง 1,000 คน ความยาวแบบทดสอบ 20 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10

วิธีตรวจสอบ	LC	ESA	EUA	CSA	CUA
LC	-	0.46	0.00	0.46	0.36
ESA	1.15	-	0.46	0.00	2.11
EUA	0.00	1.15	-	0.46	0.36
CSA	1.15	0.00	1.15	-	2.11
CUA	0.00	1.15	0.00	1.15	-

จากตาราง 94 จะเห็นว่า วิธีราชู1 กับวิธีคิม2 และวิธีคิม1 กับวิธีคิม2 มีอัตราความไม่สอดคล้องชนิดที่ 1 แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นอกนั้นไม่แตกต่างกัน สำหรับอัตราความไม่สอดคล้องชนิดที่ 2 พบว่าไม่แตกต่างกัน

บทที่ 5

สรุป อภิปรายผล และข้อเสนอแนะ

การศึกษาครั้งนี้เป็นการศึกษาเพื่อตรวจสอบและพัฒนาทฤษฎีการทดสอบ ซึ่งใช้แนวคิดทฤษฎีการตอบข้อสอบเป็นพื้นฐานในการศึกษา โดยมีความมุ่งหมายของการศึกษา ดังนี้

1. เพื่อเปรียบเทียบจำนวนข้อสอบที่ทำหน้าที่ต่างกัน และความสอดคล้องของผลการตรวจสอบระหว่าง วิธี Lord's χ^2 วิธี Raju's Area Measures และวิธี Closed Interval Area เมื่อขนาดของกลุ่มตัวอย่างและความยาวของแบบทดสอบ ต่างกัน
2. เพื่อเปรียบเทียบอัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ระหว่าง วิธี Lord's χ^2 วิธี Raju's Area Measures และ วิธี Closed Interval Area เมื่อขนาดของกลุ่มตัวอย่าง ความยาวของแบบทดสอบ และสัดส่วนของจำนวนข้อสอบที่ทำหน้าที่ต่างกัน ในแบบทดสอบ ต่างกัน

ข้อมูลที่ใช้ในการศึกษาเป็นข้อมูลผลการสอบประเมินคุณภาพและความก้าวหน้าทางการศึกษา ปีการศึกษา 2536 วิชาคณิตศาสตร์ ชั้นประถมศึกษาปีที่ 4 ของนักเรียนจำนวน 11,404 คน แบบทดสอบเป็นแบบเลือกตอบ จำนวน 80 ข้อ แบ่งเป็น 3 ตอน ประกอบด้วย ความรู้ความเข้าใจ การคิดคำนวณ และการแก้โจทย์ปัญหา ทาคุณภาพโดยใช้สูตรคูเดอริชาร์ดสัน (KR-20) ได้ความเชื่อมั่นเท่ากับ .906 และวิเคราะห์องค์ประกอบ (Factor Analysis) เพื่อตรวจสอบความเป็นมิติเดียวของแบบทดสอบ พบว่า ค่าไอเกนขององค์ประกอบที่ 1 มากกว่าองค์ประกอบที่ 2, 3, 4, 5 ประมาณ 3.57 เท่าขึ้นไป ถือว่ามีความเป็นมิติเดียว แม้ว่าร้อยละของความแปรปรวนมีค่าค่อนข้างต่ำ (ร้อยละ 8.7)

การคัดเลือกข้อสอบ เพื่อใช้ในการศึกษาครั้งนี้ ได้วิเคราะห์ทาคุณภาพรายข้อ คัดเอาเฉพาะข้อที่มีค่าความยากระหว่าง .20 - .80 และค่าอำนาจจำแนกตั้งแต่ .10 ขึ้นไป รวมทั้งค่าพารามิเตอร์ของข้อสอบที่ประมาณค่าได้ตามทฤษฎีการตอบข้อสอบ ซึ่งเหมาะสม (fit) กับแบบจำลอง 2 พารามิเตอร์ได้จำนวน 75 ข้อ แล้วจัดเตรียมข้อมูลให้สอดคล้องกับจุดมุ่งหมายของการวิจัย แบ่งเป็น 2 ส่วน ดังนี้

ส่วนที่ 1 จัดชุดแบบทดสอบให้มีกลุ่มตัวอย่างขนาด 300, 500, 700 และ 1,000 คน และความยาวแบบทดสอบ 20, 30, 40 และ 60 ข้อ

ส่วนที่ 2 จัดชุดแบบทดสอบให้มีกลุ่มตัวอย่างขนาด 300, 500, 700 และ 1,000 คน ความยาวแบบทดสอบ 20, 30, 40 และ 60 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน ในแบบทดสอบ ร้อยละ 0, 10 และ 20

สำหรับการวิเคราะห์ข้อมูล ในส่วนที่ 1 วิเคราะห์เพื่อหาจำนวนข้อที่ทำหน้าที่ต่างกัน ความสอดคล้องในการตรวจสอบ และความสัมพันธ์ระหว่างวิธีตรวจสอบ และการวิเคราะห์ข้อมูล ในส่วนที่ 2 เป็นการหาอัตราความไม่สอดคล้อง และอัตราความสอดคล้องในการตรวจสอบ การทำหน้าที่ต่างกันของข้อสอบ

สรุปผลการวิจัย

1) จำนวนข้อสอบที่ทำหน้าที่ต่างกัน

จำนวนข้อสอบที่ทำหน้าที่ต่างกันได้จากการตรวจสอบ ระหว่าง 3 วิธี แตกต่างกัน เมื่อขนาดของกลุ่มตัวอย่าง และความยาวของแบบทดสอบต่างกัน ยกเว้น กรณีกลุ่มตัวอย่างขนาด 500 คน สำหรับแบบทดสอบ 20, 30 และ 40 ข้อ และกลุ่มตัวอย่างขนาด 700 คน สำหรับแบบทดสอบ 40 ข้อ สรุปได้ ดังนี้

1.1 เมื่อกลุ่มตัวอย่างขนาดใหญ่ (1,000 คน) วิธีลอร์ดตรวจพบจำนวนข้อสอบที่ทำหน้าที่ต่างกันมากกว่าวิธีอื่น

1.2 เมื่อกลุ่มตัวอย่างขนาดใหญ่ และแบบทดสอบปานกลาง (40 ข้อ) และแบบทดสอบยาว (60 ข้อ) วิธีลอร์ด วิธีราชู1 และวิธีราชู2 ตรวจพบจำนวนข้อสอบที่ทำหน้าที่ต่างกันได้มากกว่าวิธีอื่น

1.3 เมื่อกลุ่มตัวอย่างขนาดเล็ก (300 คน) วิธีคิม2 ตรวจพบจำนวนข้อสอบที่ทำหน้าที่ต่างกันได้มากกว่าวิธีอื่น

1.4 เมื่อกลุ่มตัวอย่างขนาดเล็ก และแบบทดสอบสั้น (30 ข้อ) และแบบทดสอบยาว (60 ข้อ) วิธีคิม1 และวิธีคิม2 ตรวจพบจำนวนข้อสอบที่ทำหน้าที่ต่างกันได้มากกว่าวิธีอื่น

2) ความสัมพันธ์ระหว่างวิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

สัมประสิทธิ์สหสัมพันธ์ระหว่าง 3 วิธี โดยภาพรวมมีขนาดค่อนข้างสูงมากและมีนัยสำคัญทางสถิติเกือบทุกเงื่อนไขของการศึกษา

2.1 สัมประสิทธิ์สหสัมพันธ์ ระหว่างวิธีลอร์ด กับวิธีราชู2 และวิธีลอร์ด กับวิธีคิม2 มีขนาดค่อนข้างสูงมาก ในเกือบทุกเงื่อนไขของการศึกษา ส่วนสัมประสิทธิ์สหสัมพันธ์ระหว่างวิธีลอร์ด กับวิธีราชู1 และวิธีลอร์ด กับวิธีคิม1 บางเงื่อนไขมีขนาดสูง และบางเงื่อนไขมีขนาดปานกลาง

2.2 สัมประสิทธิ์สหสัมพันธ์ ระหว่างวิธีราชู1 กับวิธีคิม1 และวิธีราชู2 กับวิธีคิม2 มีขนาดค่อนข้างสูงมากในทุกเงื่อนไขของการศึกษา

2.3 สัมประสิทธิ์สหสัมพันธ์ ระหว่างวิธีราชู1 กับวิธีคิม2 วิธีราชู2 กับวิธีคิม1 มีขนาดค่อนข้างสูง ในบางเงื่อนไขของการศึกษา และค่อนข้างต่ำในบางเงื่อนไขของการศึกษา

2.4 เมื่อกลุ่มตัวอย่างขนาด 500 คน สัมประสิทธิ์สหสัมพันธ์มีขนาดค่อนข้างสูงมาก และมีนัยสำคัญที่ระดับ .001 ทุกค่า

3) ความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

สำหรับความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ส่วนใหญ่จะมีค่าปานกลางถึงต่ำ ในเกือบทุกเงื่อนไขของการศึกษา ยกเว้น

เมื่อกลุ่มตัวอย่างขนาด 500 คน และความยาวแบบทดสอบ 30, 40 และ 60 ข้อ พบว่าแต่ละวิธีมีความสอดคล้องกันอยู่ในระดับปานกลางถึงสูงมาก (ร้อยละ 50 - 100) ซึ่งในกรณีนี้วิธีที่มีความสอดคล้องกันในการตรวจสอบสูง ได้แก่ วิธีลอร์ด กับวิธีราชู1 (ร้อยละ 77.78 - 93.75) วิธีลอร์ด กับวิธีคิม1 (ร้อยละ 83.33 - 87.50) วิธีราชู1 กับวิธีคิม1 (ร้อยละ 71.43 - 81.25) และวิธีลอร์ด กับวิธีราชู2 (ร้อยละ 100)

และเมื่อกลุ่มตัวอย่างขนาด 700 คน และความยาวแบบทดสอบ 40 ข้อ พบว่า แต่ละวิธีที่มีความสอดคล้องกันอยู่ในระดับปานกลางถึงสูงมาก และที่มีค่าสูงมาก ได้แก่ วิธีลอร์ด กับวิธีราชู2 (ร้อยละ 90) วิธีลอร์ด กับวิธีคิม2 (ร้อยละ 88.89) และวิธีราชู2 กับวิธีคิม2 (ร้อยละ 80.00)

4. อัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

เมื่อแบบทดสอบยาวมากขึ้น และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันมากขึ้น อัตราความไม่สอดคล้องชนิดที่ 1 ส่วนใหญ่เพิ่มขึ้น ส่วนอัตราความไม่สอดคล้องชนิดที่ 2 ส่วนใหญ่ลดลง มีรายละเอียด ดังนี้

4.1 กรณีกลุ่มตัวอย่างขนาดเล็ก (300 คน)

4.1.1 แบบทดสอบสั้น (20 - 30 ข้อ) และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันน้อย (ร้อยละ 0 - 10) พบว่า อัตราความไม่สอดคล้องชนิดที่ 1 แตกต่างกัน เมื่อแบบทดสอบ 20 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน ร้อยละ 0 ได้แก่ วิธีลอร์ด กับวิธีคิม2 และวิธีราชู1 กับวิธีคิม2 และเมื่อแบบทดสอบ 30 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน ร้อยละ 0 ได้แก่ วิธีลอร์ด กับวิธีคิม2 และวิธีราชู1 กับวิธีคิม2

4.1.2 แบบทดสอบสั้น และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันมาก (ร้อยละ 20) พบว่า อัตราความไม่สอดคล้องชนิดที่ 1 ไม่แตกต่างกัน

4.1.3 แบบทดสอบปานกลาง (40 ข้อ) และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันน้อย พบว่า อัตราความไม่สอดคล้องชนิดที่ 1 แตกต่างกัน เมื่อแบบทดสอบ 40 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 ได้แก่ วิธีลอร์ด กับวิธีคิม2 วิธีราชู1 กับวิธีคิม2 วิธีราชู2 กับวิธีคิม2 และวิธีคิม1 กับวิธีคิม2

4.1.4 แบบทดสอบปานกลาง และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันมาก พบว่า อัตราความไม่สอดคล้องชนิดที่ 1 แตกต่างกัน เมื่อแบบทดสอบ 40 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20 ได้แก่ วิธีราชู1 กับวิธีคิม2

4.1.5 แบบทดสอบยาว (60 ข้อ) และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันน้อย พบว่า อัตราความไม่สอดคล้องชนิดที่ 1 แตกต่างกัน เมื่อแบบทดสอบ 60 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0 ได้แก่ วิธีลอร์ด กับวิธีคิม2 วิธีราชู1 กับวิธีคิม2 วิธีราชู2 กับวิธีคิม2 และวิธีคิม1 กับวิธีคิม2

4.1.6 แบบทดสอบยาว และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันมาก พบว่า อัตราความไม่สอดคล้องชนิดที่ 1 ไม่แตกต่างกัน

* 4.2 กรณีกลุ่มตัวอย่างขนาดปานกลาง (500 คน)

ทุกเงื่อนไขของการศึกษา พบว่า มีอัตราความไม่สอดคล้องชนิดที่ 1 ไม่แตกต่างกัน ยกเว้น เมื่อแบบทดสอบ 20 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 เท่านั้น ที่ต่างกัน

4.3 กรณีกลุ่มตัวอย่างขนาดค่อนข้างใหญ่ (700 คน)

4.3.1 แบบทดสอบสั้น และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันน้อย พบว่า อัตราความไม่สอดคล้องชนิดที่ 1 ไม่แตกต่างกัน

4.3.2 แบบทดสอบสั้น และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันมาก พบว่า อัตราความไม่สอดคล้องชนิดที่ 1 ไม่แตกต่างกัน

4.3.3 แบบทดสอบปานกลาง และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันน้อย พบว่า อัตราความไม่สอดคล้องชนิดที่ 1 แตกต่างกัน เมื่อแบบทดสอบ 40 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 ได้แก่ วิธีราชู1 กับวิธีคิม2

4.3.4 แบบทดสอบปานกลาง และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันมาก พบว่า อัตราความไม่สอดคล้องชนิดที่ 1 ไม่แตกต่างกัน

4.3.5 แบบทดสอบยาว และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันน้อย พบว่า อัตราความไม่สอดคล้องชนิดที่ 1 แตกต่างกัน เมื่อแบบทดสอบ 60 ข้อ สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0 ได้แก่ วิธีราชู1 กับวิธีคิม1 และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 ได้แก่ วิธีราชู1 กับวิธีคิม1 และวิธีราชู2 กับวิธีคิม1

4.3.6 แบบทดสอบยาว และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันมาก พบว่า อัตราความไม่สอดคล้องชนิดที่ 1 ไม่แตกต่างกัน

4.4 กรณีกลุ่มตัวอย่างขนาดใหญ่ (1,000 คน)

4.4.1 แบบทดสอบสั้น และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันน้อย พบว่า อัตราความไม่สอดคล้องชนิดที่ 1 แตกต่างกัน เมื่อแบบทดสอบ 20 ข้อ สัดส่วนจำนวนข้อสอบ

ที่ทำหน้าที่ต่างกันร้อยละ 0 ได้แก่ วิธีคิม1 กับวิธีคิม2 และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 ได้แก่ วิธีราชู1 กับวิธีคิม2 และวิธีคิม1 กับวิธีคิม2

4.4.2 แบบทดสอบสั้น และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันมาก พบว่า อัตราความไม่สอดคล้องชนิดที่ 1 แตกต่างกัน เมื่อแบบทดสอบ 20 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20 ได้แก่ วิธีราชู1 กับวิธีคิม1

4.4.3 แบบทดสอบปานกลาง และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันน้อย พบว่า อัตราความไม่สอดคล้องชนิดที่ 1 แตกต่างกัน เมื่อแบบทดสอบ 40 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0 ได้แก่ วิธีลอร์ด กับวิธีคิม1

4.4.4 แบบทดสอบปานกลาง และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันมาก พบว่า อัตราความไม่สอดคล้องชนิดที่ 1 ไม่แตกต่างกัน

4.4.5 แบบทดสอบยาว และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันน้อย พบว่า อัตราความไม่สอดคล้องชนิดที่ 1 แตกต่างกัน เมื่อแบบทดสอบ 60 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0 ได้แก่ วิธีลอร์ด กับวิธีคิม1 วิธีลอร์ด กับวิธีคิม2 วิธีราชู1 กับวิธีคิม1 วิธีราชู1 กับวิธีคิม2 วิธีราชู2 กับวิธีคิม1 และวิธีราชู2 กับวิธีคิม2 และเมื่อสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 ได้แก่ วิธีลอร์ด กับวิธีคิม1 วิธีลอร์ด กับวิธีคิม2 วิธีราชู1 กับวิธีคิม1 และวิธีราชู1 กับวิธีคิม2

4.4.6 แบบทดสอบยาว และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันมาก (ร้อยละ 20) พบว่า อัตราความไม่สอดคล้องชนิดที่ 1 แตกต่างกัน ได้แก่ วิธีลอร์ด กับวิธีคิม1 วิธีราชู1 กับวิธีคิม1 วิธีราชู2 กับวิธีคิม1 และวิธีคิม1 กับวิธีคิม2

สำหรับอัตราความไม่สอดคล้องชนิดที่ 2 พบว่า มีเพียงแบบทดสอบ 30 ข้อ สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 และ 20 สำหรับกลุ่มตัวอย่างขนาด 1,000 คน และแบบทดสอบ 30 ข้อ สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20 สำหรับกลุ่มตัวอย่างขนาด 300 คน เท่านั้นที่แตกต่างกัน นอกนั้นไม่แตกต่างกัน

5. อัตราความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

เมื่อแบบทดสอบยาวมากขึ้น และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันมากขึ้น อัตราความสอดคล้องชนิดที่ 1 ส่วนใหญ่ลดลง ส่วนอัตราความสอดคล้องชนิดที่ 2 ส่วนใหญ่เพิ่มขึ้น มีรายละเอียด ดังนี้

5.1 กรณีกลุ่มตัวอย่างขนาดเล็ก

แบบทดสอบสั้น (20 - 30 ข้อ) วิธีลอร์ด วิธีราชู1 และวิธีราชู2 มีอัตราความสอดคล้องชนิดที่ 1 สูงใกล้เคียงกัน ส่วนวิธีคิม1 และวิธีคิม2 มีอัตราความสอดคล้องชนิดที่ 1

5.2 กรณีกลุ่มตัวอย่างขนาดปานกลาง

ทุกเงื่อนไขของการศึกษา มีอัตราความสอดคล้องชนิดที่ 1 ใกล้เคียงกัน ยกเว้น กรณีแบบทดสอบ 20 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน ร้อยละ 10 โดยวิธีคิม2 มีอัตราความสอดคล้องชนิดที่ 1 ต่ำกว่าวิธีอื่น (.72)

5.3 กรณีกลุ่มตัวอย่างขนาดค่อนข้างใหญ่

ส่วนใหญ่มีอัตราความสอดคล้องชนิดที่ 1 ใกล้เคียงกัน ยกเว้น แบบทดสอบ 40 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน ร้อยละ 10 วิธีคิม2 มีอัตราความสอดคล้องชนิดที่ 1 สูงสุด (.92) และแบบทดสอบ 60 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน ร้อยละ 0 และ 20 วิธีคิม1 มีอัตราความสอดคล้องชนิดที่ 1 สูงสุด (.98 และ .85) ส่วนวิธีราชู1 และวิธีราชู2 มีอัตราความสอดคล้องชนิดที่ 1 ต่ำกว่าวิธีอื่น

5.4 กรณีกลุ่มตัวอย่างขนาดใหญ่

วิธีคิม1 มีอัตราความสอดคล้องชนิดที่ 1 สูงกว่าวิธีอื่นในเกือบทุกเงื่อนไขของการศึกษา ยกเว้นกรณี แบบทดสอบ 30 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน ร้อยละ 10 ที่วิธีคิม2 มีอัตราความสอดคล้องชนิดที่ 1 สูงกว่า

สำหรับอัตราความสอดคล้องชนิดที่ 2 พบว่า ทุกเงื่อนไขมีอัตราความสอดคล้อง ใกล้เคียงกัน ยกเว้นกรณี แบบทดสอบ 30 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกัน ร้อยละ 10 และ 20 ที่แตกต่างกัน

อภิปรายผล

1) จำนวนข้อสอบที่ทำหน้าที่ต่างกัน จากการศึกษา พบว่า เมื่อขนาดกลุ่มตัวอย่างใหญ่ขึ้น สำหรับวิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามแนวคิดของลอร์ด ใช้การเปรียบเทียบค่า พารามิเตอร์ของข้อสอบแล้วทดสอบนัยสำคัญทางสถิติด้วยสถิติทดสอบไคสแควร์ (χ^2) สามารถตรวจพบได้มากขึ้น ซึ่งสอดคล้องกับการศึกษาของกาญจนา วัฒนสุนทร (2537) และข้อเสนอแนะของ มิลแซป และอีเวอร์สัน (Millsap and Everson. 1993 : 314) และ แฮมเบอร์ตัน สวามินาทาน และโรเจอร์ส (Hambleton, Swaminathan and Rogers. 1991: 113) ว่า วิธีนี้มีความไวต่อการตัดสินข้อสอบที่ทำหน้าที่ต่างกันมาก โดยเฉพาะเมื่อกลุ่มตัวอย่างขนาดใหญ่ อย่างไรก็ตาม วิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามแนวคิดของราชู (Raju's Area Measures) และวิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามแนวคิดของ คิม และโคเฮน (Closed Interval Area) ทั้ง 2 แนวคิดใช้การวัดขนาดความแตกต่างของพื้นที่ (Area Measures) สำหรับวิธีตรวจสอบตามแนวคิดราชู ใช้สถิติทดสอบ “ Z ” ในการทดสอบความมีนัยสำคัญทางสถิติ ซึ่งก็พบว่า เมื่อกลุ่มตัวอย่างขนาดใหญ่ และความยาวแบบทดสอบปานกลาง - ยาว จำนวนข้อสอบที่พบว่า

ทำหน้าที่ต่างกันมีจำนวนใกล้เคียงกับวิธีลอร์ด ทั้งนี้เพราะสถิติทดสอบ Z และ χ^2 จะไวต่อการทดสอบใกล้เคียงกันในเงื่อนไขการสอบดังกล่าว ส่วนวิธีตรวจสอบตามแนวคิดของ คิม และโคเฮน ไม่ใช้สถิติทดสอบ แต่พิจารณาขนาดของพื้นที่ความแตกต่างระหว่าง 2 กลุ่ม กับเกณฑ์ที่กำหนดว่าขนาดของพื้นที่เท่าไรจึงถือว่า ข้อสอบทำหน้าที่ต่างกัน ในการศึกษาครั้งนี้ใช้เกณฑ์ที่ปรับจากการศึกษาของ กาญจนา วัฒนสุนทร (2537) ซึ่งพัฒนาเกณฑ์ขั้นต่ำในการตัดสินข้อสอบที่ทำหน้าที่ต่างกัน สำหรับวิชาคณิตศาสตร์ คือ $|SA| > .30$ และ $|UA| > .40$ ด้วยเหตุที่การศึกษาของ กาญจนา หาพื้นที่โดยใช้การคำนวณตามแนวคิดของราชู ที่กำหนดให้ช่วงของความสามารถเป็นแบบเปิด ทำให้สามารถได้พื้นที่แน่นอนหรือสมบูรณ์ บางครั้งเรียกว่า “ การวัดพื้นที่สมบูรณ์ ” (Exact Area Measures) ซึ่งโดยทั่วไปจะมีค่ามากกว่าขนาดของพื้นที่ที่คำนวณได้จากวิธีที่เสนอโดยคิม และโคเฮน เนื่องจากกำหนดให้ช่วงของความสามารถผู้เข้าสอบเป็นแบบปิด เรียกวิธีนี้ว่า “ Closed Interval Area ” จึงทำให้ผู้วิจัยกำหนดเกณฑ์ขั้นต่ำในการตัดสินข้อสอบที่ทำหน้าที่ต่างกันลดลงจากการเสนอของกาญจนา เล็กน้อย ($|SA| > .40$ และ $|UA| > .50$)

2) ความสัมพันธ์ระหว่างวิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ จากการศึกษาพบว่า โดยภาพรวมมีความสัมพันธ์กันค่อนข้างสูงมาก ในเกือบทุกเงื่อนไขของการศึกษา ซึ่งสอดคล้องกับการศึกษาของ คิม และโคเฮน (Kim and Cohen. 1991, 1995) และราชูและคณะ (Raju et al. 1993) สำหรับการศึกษาของคิม และโคเฮน ในปี 1995 ได้ศึกษาผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธีลอร์ด (Lord's χ^2) วิธีราชู (Raju's Area Measures) และวิธีการทดสอบอัตราส่วนไลเคิลฮูด (Likelihood Ratio Test) ใช้ข้อมูลจากแบบทดสอบจัดชั้นเรียนของนักศึกษาระดับมหาวิทยาลัย จำนวน 28 ข้อ วิชาคณิตศาสตร์ ซึ่งสหสัมพันธ์ระหว่างวิธีลอร์ดกับวิธีราชู ทั้งแบบมีเครื่องหมาย (Signed Area) และแบบไม่มีเครื่องหมาย (Unsigned Area) พบว่ามีขนาดค่อนข้างสูงมาก (.851 - .996) ส่วนการศึกษาของคิม และโคเฮน ในปี 1991 ได้ศึกษาผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธีวัดขนาดพื้นที่สมบูรณ์ (Exact Area) กับการวัดพื้นที่แบบช่วงปิด (Closed Interval Area) ใช้ข้อมูลจากการศึกษาของ ซับโคเวียค และคณะ (Subkoviak et al. 1984) พบว่า ความสัมพันธ์ระหว่างวิธีวัดขนาดพื้นที่สมบูรณ์ กับวิธีการวัดพื้นที่แบบช่วงปิด ทั้งที่พื้นที่แบบมีเครื่องหมายและแบบไม่มีเครื่องหมายมีขนาดค่อนข้างสูงมาก (.842 - .985)

อย่างไรก็ตาม การศึกษาที่ผ่านมาในอดีต ส่วนใหญ่จะกำหนดความยาวของแบบทดสอบ และขนาดของกลุ่มตัวอย่างคงที่ ทำให้ไม่พบความสัมพันธ์ในเงื่อนไขของการศึกษา ที่ขนาดของกลุ่มตัวอย่าง และความยาวของแบบทดสอบแตกต่างกัน ผลการวิจัยจึงสรุปได้จำกัดมากกว่าการศึกษาครั้งนี้

สำหรับความสัมพันธ์ระหว่างวิธีราชู1 กับวิธีคิม1 และวิธีราชู2 กับวิธีคิม2 พบว่า มีขนาดค่อนข้างสูงมากในทุกเงื่อนไขของการศึกษา เนื่องจาก วิธีราชู1 กับวิธีคิม1 ต่างก็เป็นการคำนวณพื้นที่แบบมีเครื่องหมายเหมือนกัน และวิธีราชู2 กับวิธีคิม2 ก็เป็นการคำนวณพื้นที่แบบไม่มีเครื่องหมายเช่นเดียวกัน เพียงแต่การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามแนวคิด

สำหรับความสัมพันธ์ระหว่างวิธีราชู1 กับวิธีคิม1 และวิธีราชู2 กับวิธีคิม2 พบว่า มีขนาดค่อนข้างสูงมากในทุกเงื่อนไขของการศึกษา เนื่องจาก วิธีราชู1 กับวิธีคิม1 ต่างก็เป็นการคำนวณพื้นที่แบบมีเครื่องหมายเหมือนกัน และวิธีราชู2 กับวิธีคิม2 ก็เป็นการคำนวณพื้นที่แบบไม่มีเครื่องหมายเช่นเดียวกัน เพียงแต่การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามแนวคิดของราชู ใช้สถิติทดสอบ Z ส่วนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามแนวคิดของคิม และโคเฮน ใช้ขนาดพื้นที่เทียบกับเกณฑ์ที่กำหนดการทำหน้าที่ต่างกันของข้อสอบ

3) ความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ระหว่างวิธีลอร์ด (Lord's χ^2) วิธีราชู (Raju's Area Measures) และวิธีคิม (Closed Interval Area) พบว่า ส่วนใหญ่มีค่าปานกลางถึงต่ำ ในเกือบทุกเงื่อนไขของการศึกษา ซึ่งสอดคล้องกับการศึกษาของ กาญจนา วัฒนสุนทร (2537) ที่พบความสอดคล้องระหว่างวิธี Lord's χ^2 กับวิธีทฤษฎีการตอบข้อสอบแบบมีเครื่องหมาย (Signed Area : SA) และแบบไม่มีเครื่องหมาย (Unsigned Area : UA) มีค่าค่อนข้างต่ำ (ส่วนมากต่ำกว่าร้อยละ 50) และมีความไม่คงที่ ในแบบทดสอบที่มีความยาวต่างกัน และขนาดของกลุ่มตัวอย่างต่างกัน สาเหตุที่ทำให้การตัดสินข้อสอบที่ทำหน้าที่ต่างกัน ในเงื่อนไขของการศึกษาต่างกัน ไม่คงที่ เนื่องจากเกณฑ์ที่ใช้ตัดสินข้อสอบที่ทำหน้าที่ต่างกันยังไม่สามารถกำหนดแน่นอน และความไวต่อการตัดสินของสถิติทดสอบที่ใช้ในการศึกษาต่างกัน

อย่างไรก็ตาม เมื่อกลุ่มตัวอย่างขนาด 500 คน และความยาวแบบทดสอบ 30, 40 และ 60 ข้อ และเมื่อกลุ่มตัวอย่างขนาด 700 คน และความยาวแบบทดสอบ 40 ข้อ พบว่า แต่ละวิธีมีความสอดคล้องกันอยู่ในระดับปานกลางถึงสูงมาก ทั้งนี้อาจเป็นเพราะว่า สถิติทดสอบ χ^2 ของลอร์ด สถิติทดสอบ Z ตามแนวคิดของราชู และการคำนวณหาพื้นที่ที่ใช้ช่วงความสามารถแบบปิด ตามแนวคิดของคิม และโคเฮน ณ เงื่อนไขดังกล่าว มีความสอดคล้องในการตัดสินข้อสอบที่ทำหน้าที่ต่างกันตรงกันพอดี ซึ่งจากการศึกษาครั้งนี้ไม่ค่อยพบนัก

4) อัตราความไม่สอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ พบว่า เมื่อแบบทดสอบยาวมากขึ้น และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันมากขึ้น อัตราความไม่สอดคล้องชนิดที่ 1 ส่วนใหญ่มีแนวโน้มเพิ่มขึ้น ซึ่งขัดแย้งกับผลการศึกษาของโคเฮน และคิม (Cohen and Kim. 1993 : 48) ทั้งนี้เนื่องจากการศึกษาครั้งนี้คำนวณความไม่สอดคล้องในรูปของอัตราส่วนเทียบกับฐานจำนวนข้อสอบที่ทำหน้าที่ไม่ต่างกันทั้งหมดในแต่ละเงื่อนไข ส่วนการศึกษาของ โคเฮน และคิม คัดจำนวนข้อสอบที่เกิดความไม่สอดคล้องชนิดที่ 1 เฉลี่ยจากการทำซ้ำ 5 ครั้ง ไม่ได้คิดเทียบกับฐานของแต่ละเงื่อนไขของการศึกษา รวมทั้งอาจเกิดจากอิทธิพลของการสุ่ม (Sampling) และจำนวนครั้งในการศึกษา (Replication) น้อยเกินไปก็เป็นได้ สำหรับการอภิปรายในข้อ 4 จะพิจารณาตามกรอบใหญ่ของสมมติฐานการวิจัย ดังนี้

4.1 กรณีกลุ่มตัวอย่างขนาดเล็ก ผลการศึกษา พบว่า อัตราความไม่สอดคล้องชนิดที่ 1 หรือในเมโนทัศน์ (Concept) เดียวกับอัตราความคลาดเคลื่อนชนิดที่ 1 (Type I error rate) หรืออัตราความคลาดเคลื่อนในการยอมรับผิดพลาด (False Positive error rate) มีอยู่ 5 เงื่อนไขของการศึกษาที่แตกต่างกัน จากทั้งหมด 12 เงื่อนไข ได้แก่ เมื่อแบบทดสอบสั้น (20 และ 30 ข้อ) สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0 แบบทดสอบปานกลาง (40 ข้อ)

สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 และ 20 และแบบทดสอบยาว (60 ข้อ) สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0

และหากพิจารณาในรายละเอียดของแต่ละวิธีของการตรวจสอบทั้งหมด 3 วิธีใหญ่ หรือ 5 วิธีย่อย ได้แก่ วิธีลอร์ด วิธีราชู1 (ESA) วิธีราชู2 (EUA) วิธีคิม1 (CSA) และวิธีคิม2 (CUA) มีเพียงวิธีลอร์ด วิธีราชู1 และวิธีราชู2 กับวิธีคิม2 ที่จับคู่กันแล้วมีอัตราความไม่สอดคล้องชนิดที่ 1 ใน 5 เงื่อนไขดังกล่าวแล้ว แตกต่างกัน จะเห็นได้ว่า วิธีคิม2 เกิดความไม่สอดคล้องสูงกว่าวิธีอื่น เป็นที่น่าสังเกตว่าในการศึกษาเกี่ยวกับจำนวนข้อสอบที่ทำหน้าที่ต่างกันในตอนต้น ก็พบว่าเมื่อกลุ่มตัวอย่างขนาดเล็ก วิธีคิม2 สามารถตรวจสอบจำนวนข้อสอบที่ทำหน้าที่ต่างกันได้มากกว่าวิธีอื่น ความเป็นจริงในจำนวนข้อสอบเหล่านั้นอาจไม่ใช่ข้อสอบที่ทำหน้าที่ต่างกันทั้งหมดก็ได้ นั่นคือมีโอกาสที่จะนำไปสู่การเกิดความไม่สอดคล้องชนิดที่ 1 มากกว่าวิธีอื่น ในทางกลับกันวิธีตามแนวคิดของลอร์ด เกิดความไม่สอดคล้องต่ำกว่า ซึ่งสอดคล้องกับการศึกษาในตอนต้นเช่นเดียวกันที่พบว่าวิธีลอร์ดตรวจพบจำนวนข้อสอบที่ทำหน้าที่ต่างกันได้น้อย เมื่อกลุ่มตัวอย่างขนาดเล็ก และสามารถตรวจพบได้มากขึ้น เมื่อกลุ่มตัวอย่างขนาดใหญ่ขึ้น (Millsap and Everson. 1993 : 314 ; Hambleton, Swaminathan and Rogers. 1991 : 113)

4.2 กรณีกลุ่มตัวอย่างขนาดปานกลาง ผลการศึกษา พบว่า อัตราความไม่สอดคล้องชนิดที่ 1 ของเกือบทุกเงื่อนไขของการศึกษา ไม่แตกต่างกัน มีเพียงเงื่อนไขเดียวเท่านั้นที่ต่างกัน คือ เมื่อแบบทดสอบ 20 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 ซึ่งเป็นที่น่าสังเกตว่า สอดคล้องกับผลการศึกษาตอนต้นที่เกี่ยวกับจำนวนข้อสอบที่ทำหน้าที่ต่างกัน ที่ทุกวิธีตรวจพบได้จำนวนข้อไม่แตกต่างกัน สำหรับความสัมพันธ์ระหว่างวิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ทั้ง 5 วิธีย่อย ก็พบว่ามีความสัมพันธ์ที่สัมพันธ์กันค่อนข้างสูงมาก และมีนัยสำคัญทางสถิติที่ระดับ .001 ทุกค่า และเมื่อพิจารณาในด้านความสอดคล้องของผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบก็พบในลักษณะเช่นเดียวกัน คือ เมื่อกลุ่มตัวอย่างขนาดปานกลาง (500 คน) ทุกวิธีมีความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบอยู่ในระดับปานกลางถึงสูงมาก ขณะที่ในเงื่อนไขอื่น ๆ พบว่า มีความสอดคล้องกันอยู่ในระดับปานกลางถึงต่ำ ซึ่งอาจกล่าวได้ว่า ในกรณีกลุ่มตัวอย่างขนาดปานกลาง (500 คน) ไม่ว่าแบบทดสอบจะสั้นหรือยาว และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันแบบทดสอบ จะมากหรือน้อย ทุกวิธีให้ผลการตรวจสอบสอดคล้องกัน ทั้งนี้อาจเป็นเพราะว่า ในการศึกษาครั้งนี้ใช้แบบจำลอง 2 พารามิเตอร์ (พิจารณาเฉพาะค่าความยาก (b) และค่าอำนาจจำแนก (a)) ขนาดของกลุ่มตัวอย่างที่เหมาะสมและเพียงพอกับแบบจำลองที่ใช้ในการศึกษา มีขนาดประมาณ 500 คน ซึ่งสอดคล้องกับงานวิจัยของกาญจนา วัฒนสุนทร (2537) ที่ใช้แบบจำลอง 2 พารามิเตอร์ เช่นเดียวกัน และเสนอแนะว่าควรใช้กลุ่มตัวอย่างไม่น้อยกว่า 400 คน ขึ้นไป จึงจะทำให้ผลการตรวจสอบค่อนข้างคงที่และใกล้เคียงกัน

4.3 กรณีกลุ่มตัวอย่างขนาดค่อนข้างใหญ่ ผลการศึกษา พบว่า มีเพียง 3 เงื่อนไขของการศึกษาเท่านั้นที่มีอัตราความไม่สอดคล้องชนิดที่ 1 แตกต่างกัน (จากทั้งหมด 12 เงื่อนไข) ได้แก่ เมื่อแบบทดสอบปานกลาง (40 ข้อ) สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10

แบบทดสอบยาว (60 ข้อ) และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0 และ 10 อย่างไรก็ตาม นับว่ามีความสอดคล้องกันอยู่ในเกณฑ์ค่อนข้างสูง (ร้อยละ 75) ซึ่งโดยทั่วไปการศึกษาเปรียบเทียบวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อขนาดของกลุ่มตัวอย่างต่างกัน ความยาวแบบทดสอบต่างกัน และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันแบบทดสอบต่างกัน มีความไม่คงที่ของผลการศึกษา เช่น งานวิจัยของเพิร์ลแมน และคณะ (Perlman et al. 1988) งานวิจัยของบาเกิ และเฟอร์รารา (Baghi and Ferrara. 1990) งานวิจัยของโคเฮน และคิม (Cohen and Kim. 1993) และงานวิจัยของกาญจนา วัฒนสุนทร (2537) เป็นต้น

4.4 กรณีกลุ่มตัวอย่างขนาดใหญ่ ผลการศึกษา พบว่า อัตราความไม่สอดคล้องชนิดที่ 1 เกิดขึ้น 7 เงื่อนไขของการศึกษา ได้แก่ เมื่อแบบทดสอบ 20 ข้อ สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0, 10 และ 20 แบบทดสอบปานกลาง (40 ข้อ) สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0 แบบทดสอบยาว (60 ข้อ) สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 0, 10 และ 20 เป็นที่น่าสังเกตว่า เมื่อกลุ่มตัวอย่างขนาดใหญ่ขึ้น (700 - 1,000 คน) ความยาวแบบทดสอบมากขึ้น ทั้งกรณีที่มีสัดส่วนการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบน้อย และมาก พบว่า วิธีตามแนวคิดของลอว์รี และวิธีตามแนวคิดของราชู มีอัตราความไม่สอดคล้องชนิดที่ 1 สูงกว่าวิธีตามแนวคิดของคิม และโคเฮน ทั้งนี้เพราะสถิติทดสอบ χ^2 ของลอว์รี และสถิติทดสอบ Z ของราชู มีความไวต่อการตัดสินใจการทำหน้าที่ต่างกันของข้อสอบ ซึ่งวิธีราชู 2 (EUA) เป็นการคำนวณพื้นที่ความแตกต่างของฟังก์ชันการตอบข้อสอบ (Unsigned Area) ใช้ค่าความยาก (b) และค่าอำนาจจำแนก (a) มาคำนวณในสูตรเช่นเดียวกับ χ^2 ของลอว์รี ส่วนวิธีราชู 1 (ESA) เป็นการคำนวณพื้นที่ความแตกต่างของฟังก์ชันการตอบแบบมีเครื่องหมาย (Signed Area) ไม่ได้นำค่าอำนาจจำแนก (a) มาคำนวณด้วย และจากการศึกษาของโคเฮน และคิม (Cohen and Kim. 1993) พบว่า วิธีราชู 1 มีความไม่สอดคล้องสูงกว่า วิธีราชู 2 และวิธีลอว์รี

สำหรับอัตราความไม่สอดคล้องชนิดที่ 2 มีเพียง 3 เงื่อนไขของการศึกษาเท่านั้นที่แตกต่างกัน ได้แก่ เมื่อแบบทดสอบ 30 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 และ 20 ในกลุ่มตัวอย่าง 300 คน และ 1,000 คน โดยพบว่า แบบทดสอบที่มีสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันมากขึ้น อัตราความไม่สอดคล้องชนิดที่ 2 ส่วนใหญ่มีแนวโน้มลดลง ซึ่งขัดแย้งกับผลการศึกษาของโคเฮน และคิม (Cohen and Kim. 1993) ทั้งนี้เพราะการศึกษาค้างนี้คำนวณความไม่สอดคล้องในรูปของอัตราส่วนเทียบกับฐานจำนวนข้อสอบที่ทำหน้าที่ต่างกันทั้งหมดในแต่ละเงื่อนไข ส่วนการศึกษาของโคเฮน และคิม คิดจำนวนข้อสอบที่เกิดความคลาดเคลื่อนในการปฏิเสธผิดพลาดเลย ไม่ได้คิดเทียบกับฐานของแต่ละเงื่อนไขของการศึกษา

5) อัตราความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ในการศึกษาครั้งนี้ แบ่งเป็น 2 ลักษณะ คือ อัตราความสอดคล้องชนิดที่ 1 และอัตราความสอดคล้องชนิดที่ 2 สำหรับอัตราความสอดคล้องชนิดที่ 1 จะแปรผกผันกับอัตราความไม่สอดคล้องชนิดที่ 1 กล่าวคือ ถ้าอัตราความไม่สอดคล้องชนิดที่ 1 มีค่ามาก อัตราความสอดคล้องชนิดที่ 1 จะมีค่าน้อย

ถ้าอัตราความไม่สอดคล้องชนิดที่ 1 มีค่าน้อย อัตราความสอดคล้องชนิดที่ 1 จะมีค่ามาก ส่วนอัตราความสอดคล้องชนิดที่ 2 จะแปรผกผันกับอัตราความไม่สอดคล้องชนิดที่ 2 ซึ่งมีความหมายในทำนองเดียวกับอัตราความสอดคล้องชนิดที่ 1

เมื่อกลุ่มตัวอย่างขนาดเล็ก วิธีตามแนวคิดของลอร์ด และวิธีตามแนวคิดของราชู มีอัตราความสอดคล้องมากกว่าวิธีตามแนวคิดของคิม และโคเฮน ทั้งนี้เพราะวิธีการตรวจสอบทั้งสองมีความไวต่อขนาดของกลุ่มตัวอย่าง ถ้ากลุ่มตัวอย่างขนาดเล็ก ทำให้เกิดความไม่สอดคล้องต่ำกว่าเมื่อกลุ่มตัวอย่างขนาดใหญ่

เมื่อกลุ่มตัวอย่างขนาดปานกลาง ทุกวิธีมีอัตราความสอดคล้องชนิดที่ 1 และอัตราความสอดคล้องชนิดที่ 2 ไม่มีความแตกต่างกัน ยกเว้นกรณี แบบทดสอบ 20 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 และวิธีคิม2 ยังคงมีอัตราความสอดคล้องต่ำกว่าวิธีอื่น

และเมื่อกลุ่มตัวอย่างขนาดใหญ่มากขึ้น วิธีตามแนวคิดของคิม และโคเฮน มีอัตราความสอดคล้องสูงกว่าวิธีตามแนวคิดของลอร์ด และราชู โดยเฉพาะ วิธีคิม1 (CSA) มีอัตราความสอดคล้องสูงกว่าวิธีคิม2 (CUA) เกือบทุกเงื่อนไข ยกเว้นกรณี แบบทดสอบ 30 ข้อ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 ที่วิธีคิม2 มีอัตราความสอดคล้องสูงกว่าวิธีคิม1 ซึ่งสอดคล้องกับการศึกษาของเฟนสเทน (Freinstein, 1995) ที่พบว่า วิธีคิม1 มีความสามารถในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันได้ดีกว่าวิธีคิม2 ในกรณีที่ความแตกต่างระหว่างค่าพารามิเตอร์มีขนาดปานกลางถึงขนาดใหญ่

เนื่องจากในการศึกษาครั้งนี้ เป็นการศึกษาจากข้อมูลเชิงประจักษ์ (Empirical data) โดยใช้ข้อมูลผลการสอบประเมินความก้าวหน้าและคุณภาพทางการศึกษา วิชาคณิตศาสตร์ ชั้นประถมศึกษาปีที่ 4 ของสำนักงานการประถมศึกษาแห่งชาติ ปีการศึกษา 2536 ในการศึกษาส่วนที่ 2 ที่ต้องจัดชุดข้อสอบ มีตัวแปรสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันในความยาวแบบทดสอบต่าง ๆ กันนั้น ข้อสอบที่ถือว่า เป็นข้อสอบที่ทำหน้าที่ต่างกัน ได้จาก การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างเทคนิควิธีทั้ง 3 ที่ตรวจสอบได้ตรงกัน หรือสอดคล้องกัน ซึ่งไม่สามารถกำหนดให้เป็นข้อสอบที่ทำหน้าที่ต่างกันได้ตั้งแต่เริ่มต้นเช่นเดียวกับการสร้างจากข้อมูลจำลอง และจากการศึกษาครั้งนี้ ยังพบว่า ความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ มีความไม่คงที่ และมีขนาดปานกลางถึงต่ำ ดังนั้น ในการศึกษาครั้งนี้ จึงใช้คำว่า “อัตราความไม่สอดคล้อง” แทนคำว่า “อัตราความคลาดเคลื่อน” และทำให้ข้อสรุปจากการศึกษาไม่สามารถบอกได้ชัดเจนว่าวิธีใดมีอัตราความคลาดเคลื่อนต่ำกว่า หรือสูงกว่า ซึ่งอาจกล่าวได้ว่า เป็นเพียงแนวโน้มว่ามีทิศทางไปทางใดเท่านั้น

ข้อเสนอแนะในการนำไปใช้

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ควรใช้อย่างน้อย 3 วิธีขึ้นไป ในการตัดสินใจ ทั้งนี้เพราะจากการศึกษา พบว่า แม้ว่าทั้งวิธีลอร์ด วิธีตามแนวคิดของราชู และวิธีตามแนวคิดของคิม จะตั้งอยู่บนพื้นฐานทฤษฎีการตอบข้อสอบ (Item Response Theory) เหมือนกัน แต่ผล

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบยังแตกต่างกัน ไม่ว่าจะเป็นจำนวนข้อสอบที่พบว่าทำหน้าที่ต่างกัน ความสัมพันธ์ระหว่างวิธีตรวจสอบ ความสอดคล้องในการตรวจสอบ อัตราความไม่สอดคล้องในการตรวจสอบ และอัตราความสอดคล้อง ทั้งนี้ขึ้นอยู่กับขนาดของกลุ่มตัวอย่าง ความยาวของแบบทดสอบ เกณฑ์ที่ใช้ในการตัดสินการทำหน้าที่ต่างกันของข้อสอบ และความไวต่อการตัดสินข้อสอบที่ทำหน้าที่ต่างกันของสถิติทดสอบ ต่างกัน ขณะเดียวกันก็กับแนวโน้มในการเลือกใช้เฉพาะบางเงื่อนไข ดังนี้

1. เมื่อกลุ่มตัวอย่างขนาดเล็ก (300 คน) ควรเลือกใช้วิธี Lord's χ^2 และวิธีราชู
2. เมื่อกลุ่มตัวอย่างขนาดใหญ่ (1,000 คน) ควรเลือกใช้วิธีตามแนวคิดของ คิม และ โคเฮน
3. เมื่อกลุ่มตัวอย่างขนาดเล็ก (300 คน) และมีจำนวนข้อสอบที่ทำหน้าที่ต่างกันในรูปแบบทดสอบน้อย ควรเลือกใช้วิธีลอร์ด และวิธีราชู ถ้าแบบทดสอบยาวขึ้น และมีจำนวนข้อสอบที่ทำหน้าที่ต่างกันในรูปแบบทดสอบมาก ควรเลือกใช้วิธีราชู
4. เมื่อกลุ่มตัวอย่างขนาดปานกลาง (500 คน) โดยทั่วไปสามารถใช้ได้ทั้ง 3 วิธี แต่ถ้าแบบทดสอบยาว (60 ข้อ) และมีจำนวนข้อสอบที่ทำหน้าที่ต่างกันในรูปแบบทดสอบมาก ควรเลือกใช้ วิธีราชู2 และถ้าแบบทดสอบปานกลาง (40 ข้อ) และมีจำนวนข้อสอบที่ทำหน้าที่ต่างกันในรูปแบบทดสอบร้อยละ 10 และ 20 ควรเลือกใช้วิธีคิม
5. เมื่อกลุ่มตัวอย่างขนาดค่อนข้างใหญ่ (700 คน) แบบทดสอบปานกลาง และมีจำนวนข้อสอบที่ทำหน้าที่ต่างกันในรูปแบบทดสอบร้อยละ 10 และ 20 ควรเลือกใช้วิธีคิม
6. เมื่อกลุ่มตัวอย่างขนาดใหญ่ (1,000 คน) แบบทดสอบสั้น (20 ข้อ) และมีจำนวนข้อสอบที่ทำหน้าที่ต่างกันในรูปแบบทดสอบร้อยละ 10 ควรเลือกใช้ วิธีคิม1 และถ้าแบบทดสอบยาว (60 ข้อ) และมีจำนวนข้อสอบที่ทำหน้าที่ต่างกันในรูปแบบทดสอบน้อย (ร้อยละ 0 และ 10) ควรเลือกใช้วิธีคิม

ข้อเสนอแนะในการวิจัยครั้งต่อไป

1. ควรศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในลักษณะเดียวกับงานวิจัยครั้งนี้ แต่ใช้ข้อมูลจำลอง (Simulated data) เพื่อยืนยันผลการศึกษว่าได้ข้อสรุปสอดคล้องกันหรือไม่
2. ควรศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในลักษณะเดียวกับงานวิจัยครั้งนี้ แต่เลือกตัวแปรอื่น เช่น ขนาดของการทำหน้าที่ต่างกันของข้อสอบ แบบจำลองในการศึกษา และวิธีประมาณค่าพารามิเตอร์ เป็นต้น จะทำให้ได้ข้อสรุปในแง่มุมต่างๆ ที่สมบูรณ์ยิ่งขึ้น
3. ควรศึกษาร่วมกับวิธีอื่นด้วยการคัดเลือกวิธีที่กำลังเป็นที่นิยม เช่น วิธีแมนเทส แชนเซล และวิธีซิปเทส (SIBTEST)
4. ควรศึกษาถึงความสัมพันธ์ระหว่างการทำหน้าที่ต่างกันของข้อสอบ กับความเป็นหลายมิติของข้อสอบ

บรรณานุกรม

บรรณานุกรม

- กาญจนา วิธสุนทร. การพัฒนาเกณฑ์ตัดสินข้อสอบลำเอียงทางเพศ. วิทยานิพนธ์ ค.ด. กรุงเทพฯ : จุฬาลงกรณ์มหาวิทยาลัย, 2537. อัดสำเนา.
- เจน จันทรแสง. การวิเคราะห์ความลำเอียงของข้อสอบของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียน วิชาคณิตศาสตร์ ชั้นมัธยมศึกษาปีที่ 3. วิทยานิพนธ์ กศ.ม. พิษณุโลก : มหาวิทยาลัยนเรศวร, 2537. อัดสำเนา.
- ชัยชัย เฝ้าพงศ์. การวิเคราะห์ความลำเอียงของข้อสอบจากแบบทดสอบวัดความถนัดทางการเรียน ด้านคณิตศาสตร์และภาษา ระดับชั้นมัธยมศึกษาตอนต้น. วิทยานิพนธ์ ศศ.ม. ขอนแก่น : มหาวิทยาลัยขอนแก่น, 2527. อัดสำเนา.
- ชวาล แพรตกุล. เทคนิคการวัดผล. กรุงเทพฯ : วัฒนาพานิช, 2509.
- ทัศนีย์ พิรมนตรี. การวิเคราะห์ความลำเอียงของแบบสอบวิชาคณิตศาสตร์ โครงการตรวจสอบคุณภาพการศึกษาชั้นมัธยมศึกษาปีที่ 6 ปีการศึกษา 2526. วิทยานิพนธ์ ค.ม. กรุงเทพฯ : จุฬาลงกรณ์มหาวิทยาลัย, 2530. อัดสำเนา.
- พัชรี ปิยะภันต์. การวิเคราะห์ความลำเอียงของข้อสอบจากแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียน. วิทยานิพนธ์ กศ.ม. กรุงเทพฯ : มหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร, 2531. อัดสำเนา.
- สุพัฒน์ สุกมลสันต์. การวิเคราะห์ความลำเอียงของข้อสอบภาษาอังกฤษเข้ามหาวิทยาลัยปี 2531-2533. กรุงเทพฯ : สถาบันภาษา จุฬาลงกรณ์มหาวิทยาลัย, 2534. อัดสำเนา.
- เสรี ชัดแจ้ง. การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันแบบไม่สม่ำเสมอของข้อสอบ ระหว่างวิธีแมนเทิล แชนเซล แบบปกติ กับวิธีแมนเทิล แชนเซล แบบแบ่งกลุ่มตามความสามารถและความยาวของข้อสอบ. วิทยานิพนธ์ ค.ด. กรุงเทพฯ : จุฬาลงกรณ์มหาวิทยาลัย, 2540. อัดสำเนา.
- สุรศักดิ์ อมรรัตน์ศักดิ์. การศึกษาเปรียบเทียบผลของวิธีวิเคราะห์ความลำเอียงของข้อสอบที่ต่างกัน 4 วิธี. วิทยานิพนธ์ ค.ด. กรุงเทพฯ : จุฬาลงกรณ์มหาวิทยาลัย, 2531. อัดสำเนา.
- อุทุมพร จามรมาน. ทฤษฎีการวัดทางจิตวิทยา. กรุงเทพฯ : ฟีนีฟับลิชชิ่ง, 2537.

Allen, Mary J. and Wendy M. Yen. Introduction to Measurement Theory. California : Brooks/Cole Publishing Company, 1979.

Baghi, Heibatollah and Steven F. Ferrara. A Comparison of IRT, Delta Plot, and Mantel-Haenszel Techniques for Detecting Differential Item Functioning Across Subpopulations in the Maryland Test of Citizenship Skills. Paper presented at the Annual Meeting of the American Educational Research Association. San Francisco, California, March: 27 - 31 ; 1989.

Baghi, Heibatollah and Steven F. Ferrara. Detecting Differential Item Functioning Using IRT and Mantel-Haenszel Techniques: Implementing Procedures and Comparing Results. Paper Presented at the Annual Meeting of the Eastern Educational Research Association. Clearwater, Florida : 14 - 17 ; February, 1990.

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 : 78 ; 1991.

Berk, R.A. Handbook of Methods for Detecting item Bias. Baltimore, MD : John Hopkins University Press, 1982.

Camilli, Gregory. " A Conceptual Analysis of Differential Item Functioning in Terms of a Multidimensional Item Response Model, " Applied Psychological Measurement. 10(2) : 129 - 147 ; June, 1992.

✓ Camilli, Gregory and Lorrie A. Shepard. Methods for Identifying Biased Test Items. California : Sage Publications, 1994.

✓ Candell, G.L. and F. Drasgow. " An iterative procedure for linking metrics and assessing Item Bias in Item Response Theory, " Applied Psychological Measurement. 12 : 253 - 260 ; 1988.

✓ Cohen, Allan S. and Seock-Ho. Kim. " A Comparison of Lord's χ^2 and Raju's Area Measures in Detection of Differential Item Functioning, " Applied Psychological Measurement. 17(1) : 39 - 52 ; Spring, 1993.

✓ Cohen, Allan S., Seock-Ho. Kim and Michael J. Subkoviak. " Influence of Distributions on Detection of DIF, " Journal of Educational Measurement. 28(1) : 49 - 59 ; Spring, 1991.

✓ Cohen, Allan S., Seock-Ho. Kim and Wollack, James A. " An Investigation of the Likelihood Ratio Test for Detection of Differential Item Functioning, " Applied Psychological Measurement. 7 September, 1995.

- Cole, Nancy S. and Pamela A. Moss. Bias in Test Use. p 201 – 219 in Robert L. Linn (ed.) Educational Measurement (3 rd ed.) New York : Macmillan Publishing Company, 1989.
- Doolittle, Allen E. Gender Differences in Performance on Categories of Mathematics Achievement Test Items. Unpublished Doctoral Dissertation, The University of Iowa, 1987.
- Doolittle, Allen E. and T. Anne. Cleary. “ Gender-Based Differential Item Performance in Mathematics Achievement Items, ” Journal of Educational Measurement. 24(2) : 157 – 166 ; Summer, 1987.
- Gallagher, A.M. and R.D. Lisi. Gender Differences in Mathematics Problem Solving Strategies. Paper presented at the American Educational Research Association, San Francisco : February, 1992.
- Feinstein, Zachary S. “ Effects of Differing Item Parameters on Closed-Interval DIF Statistics,” Applied Psychological Measurement. 19(2) : 131 – 142 ; June, 1995.
- Guilford J.P. and B. Fruchter. Fundamental Statistics in Psychology and Education. McGraw-Hill Singapore, 1978.
- Hambleton, Ronald K. and L.L. Cook. “ Latent Trait Model and their Use in the Analysis of Educational Test Data, ” Journal of Educational Measurement. 14 (2) : 75 – 96 ; 1977.
- Hambleton, Ronald K. and H. Jane. Rogers. “ Detecting Potential Biased Test Items. Comparison of IRT Area and Mantel-Haenszel Methods, ” Applied Measurement in Education. 2(4) : 313 – 334 ; 1989.
- Hambleton, Ronald K. and Hariharan Swaminathan. Item Response Theory Principles and Applications. Boston, Kluwer-Nijhoff Publishing, 1985.
- Hambleton, Ronald K. , H. Swaminathan and H. Jane. Rogers. Fundamentals of Item Response Theory. Newbury Park, California : Sage Publications, 1991.

- Harris, Abigail M. and Sydel T. Carlton. "Patterns of Gender Differences on Mathematics Items on the Scholastic Aptitude Test," Applied Measurement in Education. 6(2) : 137 - 151 ; 1993.
- Holland, P.W. and H. Wainer. Differential Item Functioning. Hillsdale NJ : Erlbaum, 1993.
- Hulin, Charles L., Fritz Drasgow and Charles K. Parsons. Item Response Theory Application to Psychological Measurement. Illinois : Dow Jones Irwin, 1983.
- Ironson, G.H. and M.J. Subkoviak. "A Comparison of Seven Methods of Assessing Item Bias," Journal of Educational Measurement. 16 : 209 - 225, 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.
- Kim, Seock-Ho and Allan S. Cohen. "Effects of Linking Methods on Detecting of DIF," Journal of Educational Measurement. 29(1) : 51 - 66 ; Spring 1992.
- Kim, Seock-Ho and Allan S. Cohen and Hae-OK. Kim. "An Investigation of Lord's Procedure for the Detection of Differential Item Functioning," Applied Psychological Measurement. 18(3) : 217 - 228 ; 1994.
- Kim, Seock-Ho and Allan S. Cohen and Hae-OK. Kim. "A Comparison of Lord's Chi-Square, Raju's Area Measures and the Likelihood Ratio Test on Detection of Differential Item Functioning," Applied Measurement in Education. 8(4) : 291 - 312 ; 1995.
- Lautenschlager, Gary J. and Dong-Gun. Park. "IRT Item Bias Detection Procedure : Issues of Model Misspecification, Robutness, and Parameter Linking," Applied Psychological Measurement. 12(4) : 365 - 376 ; December, 1988.
- Lautenschlager, Gary J., Vicki L. Flaherty and Dong-Gun. Park. "IRT Differential Item Functioning : An Examination of Ability Scale Purifications," Educational and Psychological Measurement. 54(1) : 21 - 31 ; Spring, 1994.

- Linn, Robert L. Educational Measurement. 3rd ed.. New York : Mcmillan Publishing Company, 1989.
- Linn, Robert L. and Delwyn L. Harnisch. " Interactions Between Item context and Group Membership on Achievement Test Items, " Journal of Educational measurement. 18(2) : 109 - 118 ; Summer, 1981.
- Linn, Robert L. and others. " Item Bias in Test of Reading Comprehension, " Applied Psychological Measurement. 5(2) : 159 - 173 ; 1981.
- Lord, F.M. and M.R. Novick. Statistical Theories of Mental Test Score. Massachusetts : Addison Wesley Publishing Company, 1968.
- Lord, F.M. Application of Item Response Theory to Practical Testing Problems. Hillsdale, N.J. : Lawrence Erlbaum Associates, 1980.
- McCauley, Cynthia D. and Jorge. Mendoza. " A Simulation Study of Item Bias Using a Two-Parameter Item Response Model, " Applied Psychological Measurement. 9(4) : 389 - 400 ; 1985.
- McGrogan, Harold James Jr. The Determination of Cultural Item Bias in the California Achievement Test. Unpublished Doctoral Dissertation, The University of Arizona, 1989.
- McLanghlin, M.E. and F. Drasgow. " Lord's Chisquare Test of Item Bias with estimated and with known person parameters, " Applied Psychological Measurement. 11 : 161 - 173 ; 1987.
- Millsap, R.E. and H.T. Everson. " Methodology review : Statistical Approaches for Assessing Measurement Bias, " Applied Psychological Measurement. 17 : 297 - 334 ; 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 1989.

- Oshima, T.C. and M.D. Miller. "Multidimensionality and IRT Based Item Invariance Indexes : The Effect of Between Group variation in Trait Correlation, " Journal of Educational measurement. 27(3) : 273 - 283 ; Fall, 1990.
- Osterlind, Steven J. Test Item Bias. California : Sage Publications, Inc., 1983.
- Park, Dong-Gun and Gary J. Lautenschlager. "Improving IRT Item Bias Detection with Iterative Linking and ability Scale Purification, " Applied Psychological measurement. 14(2) : 167 - 173 ; June, 1990.
- Perلمان, C.L. et al. Investigating the Stability of Four Methods for Estimating Item Bias. Paper presented at the Annual Meeting of the National Council for Measurement in Education. New Orleans, LA, April 5 - 9, 1988.
- Popham, W. James. Modern Educational Measurement A Practitioner's Perspective. 2nd Ed. Englewood Cliffs New Jersey : Prentice Hall, 1990.
- Raju, Drasgow Fritz and Jeffrey A. Slinde. "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 : 301 - 314 ; 1993.
- Raju, Nambury S. "The Area Between Two Item Characteristic Curves, " Psychometrika. 53 : 495 - 502 ; 1988.
- 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, Nambury S. , Wim J. Van der Linden and Paul F. Fleer. "IRT-Based Internal Measures of Differential Functioning of Item and Tests, " Applied Psychological Measurement. 19(4) : 353 - 368 ; December ; 1995.
- Reynolds, Trudy Lynn. An Application of Several Differential Item Functioning Procedures to Assess the Quality of a Translated Test. Unpublished Doctoral Dissertation, Southern Illinois University at Carbondale, 1989.

- Rudner, L.M. An Evaluation of Select Approaches for Biased Item Identification. Unpublished Doctoral Dissertation Catholic University of America, 1977.
- Rudner, L.M., P.R. Getson and D.L. Knight. "A Monte Carlo Comparison of Seven Biased Item Detection Techniques," Journal of Educational Measurement. 17(1) : 1 - 10 ; 1980.
- Scheuneman, Janice Dowd. and Kalle. Gerritz. "Using Differential Item Functioning Procedure to Explore Source of Item Difficulty and Group Performance Characteristics," Journal of Educational Measurement. 27(2) : 109 - 131 ; 1990.
- Schmitt, Alicia P. and Neil J. Dorans. "Differential Item Functioning for Minority Examination on the SAT," Journal of Educational Measurement. 27(1) : 67 - 81 ; 1990.
- Shealy, R. and W. Stout. "A Model-Based Standardization Approach that Separates True Bias / DIF from Group Ability Differences and Detects Test Bias /DTF as well as Bias / DIF," Psychometrika. (58) : 159 - 194 ; 1993.
- Seigel, Sidney. Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill Kogakusha Tokyo, 1956.
- Shepard, L.A.,G. Camilli and D.M. Williams. "Validity of Approximation Techniques for Detecting Item Bias," Journal of Educational Measurement. 22(2) : 77 - 105 ; 1985.
- Shepard, L.A.,G. Camilli and D.M. Williams. "Accounting for Statistics Artifacts in Item Bias Research," Journal of Educational Measurement. 9(2) : 93 - 126 ; 1984.
- Stocking, M.L. and F.M. Lord. "Developing a Common metric in item response theory," Applied Psychological Measurement. 7 : 201 - 210 ; 1983.
- Subkoviak, Micheal J. and others. "Empirical Comparison of Deleted Item Bias Detection Prodecures with Bias Manipulation," Journal of Educational Measurement. 21(10) : 49 - 59 ; Spring, 1984.

Thissen, David, Lynne Steinberg and Howard. Wainer. Use of Item Response Theory in the Study of Group Differences in Trace Test Validity. p 147 - 170 in Wainer, Howard and Braun, Henry I. Test Validity. Hillsdale, New Jersey : Lawrence Erlbaum Associates , 1988.

_____. Detection of differential Item functioning using the parameters of item response models. p 67-113 In P.W. Holland and H. Wainer Differential Item Functioning. Hillsdale NJ: Erlbaum, 1993.

Wainer, Howard, Stephen G. Sireci and David. Thissen. " Differential Testlet Functioning : Definitions and Detection, " Journal of Educational Measurement. 28(3) : 197 - 219 ; Fall, 1991.

Wright, B.D. and M.H. Stone. Best Test Design. Chicago : Mesa Press, 1979.

ภาคผนวก

ตาราง 113 ค่าพารามิเตอร์ของข้อสอบ ที่ประมาณค่าโดยใช้แบบจำลอง 3 พารามิเตอร์

ข้อที่	ค่าพารามิเตอร์ข้อสอบ								
	a_i	b_i	c_i	a_2	b_2	c_2	a_i	b_i	c_i
1	.557	.165	.030	.523	-.241	.031	.540	-.074	.016
2	.533	-.098	.021	.521	-.436	.038	.526	-.303	.014
3	.530	.161	.026	.485	-.128	.026	.505	-.017	.013
4	.707	.608	.010	.687	.318	.011	.694	.446	.005
5	.730	.582	.010	.789	.257	.009	.761	.398	.005
6	.442	-.509	.051	.412	-.929	.095	.427	-.791	.043
7	.523	.905	.017	.545	.632	.014	.528	.737	.006
8	.520	1.076	.012	.509	.728	.014	.513	.869	.005
9	.650	1.299	.006	.619	1.109	.010	.633	1.184	.004
10	.612	1.456	.008	.621	1.164	.008	.618	1.283	.004
11	.657	.807	.008	.674	.528	.013	.663	.647	.005
12	.809	.787	.005	.824	.524	.008	.820	.637	.003
13	.817	1.076	.005	.795	.803	.006	.809	.923	.003
14	.743	.492	.007	.724	.220	.008	.738	.342	.004
15	.797	-1.113	.027	.759	-1.472	.033	.783	-1.300	.016
16	.724	-.870	.020	.700	-1.280	.028	.718	-1.080	.012
17	.728	-.892	.019	.746	-1.254	.025	.741	-1.082	.011
18	.677	-.321	.012	.670	.942	.019	.676	-.503	.007
19	.602	.984	.016	.559	.942	.019	.568	.941	.009
20	.498	1.081	.014	.508	1.006	.032	.491	1.006	.001
21	.549	-.475	.060	.590	-.326	.126	.550	-.476	.068
22	.496	.960	.171	.517	1.195	.184	.505	1.108	.184
23	.672	.952	.316	.589	.675	.282	.639	.818	.303
24	.799	2.050	.197	.866	1.856	.181	.832	1.955	.190
25	.048	1.743	.227	.932	1.664	.193	1.006	1.705	.213
26	.554	2.413	.227	.528	2.233	.199	.537	2.335	.214
27	.471	.925	.230	.476	1.058	.234	.469	1.002	.234
28	.708	.525	.162	.810	-.607	.251	.761	-.570	.204
29	.772	.933	.143	.735	.923	.158	.741	.917	.147
30	.503	.285	.128	.435	-.129	.078	.452	.008	.082
31	.165	.900	.172	.590	2.726	.489	.203	2.625	.342
32	.611	1.332	.219	.917	1.296	.262	.749	1.335	.247
33	.953	2.596	.269	1.157	2.329	.267	1.106	2.437	.271
34	.741	1.402	.132	.743	1.462	.131	.736	1.437	.132
35	.487	-.033	.081	.487	.062	.091	.472	-.041	.071

ตาราง 113 (ต่อ)

ข้อที่	ค่าพารามิเตอร์ข้อสอบ								
	a_1	b_1	c_1	a_2	b_2	c_2	a_t	b_t	c_t
36	.503	2.275	.180	.571	1.905	.200	.523	2.077	.186
37	.561	1.949	.269	.504	1.470	.248	.541	1.694	.261
38	.356	.201	.139	.302	-.167	.099	.315	-.113	.088
39	.820	.874	.240	.818	.894	.254	.800	.877	.244
40	.641	.230	.230	.610	.067	.234	.631	.158	.237
41	.369	-1.084	.102	.387	-1.420	.094	.374	-1.345	.069
42	.723	.505	.207	.687	.296	.175	.719	.416	.199
43	.524	1.777	.186	.598	1.781	.204	.558	1.799	.197
44	.496	-1.308	.083	.469	-1.847	.094	.484	-1.618	.060
45	.902	1.692	.162	.728	1.476	.131	.812	1.587	.150
46	.713	.579	.272	.685	.445	.176	.714	.556	.241
47	.537	-.601	.240	.497	-.826	.259	.529	-.647	.268
48	.440	2.986	.124	.359	4.253	.147	.412	-3.483	.112
49	.512	-1.338	.090	.443	-1.891	.128	.480	-1.649	.079
50	.676	1.293	.112	.851	1.185	.091	.800	1.247	.103
51	.695	1.104	.177	.818	.984	.185	.771	1.049	.186
52	.703	1.273	.096	.759	1.266	.126	.717	1.263	.107
53	.875	1.669	.223	.834	1.467	.209	.871	1.560	.219
54	.388	2.165	.204	.389	1.409	.162	.394	1.762	.187
55	.340	3.801	.212	.433	3.231	.233	.365	3.587	.220
56	.806	.418	.236	.768	.349	.188	.792	.408	.222
57	.682	-.206	.133	.721	.017	.233	.694	-.103	.184
58	.673	.472	.253	.648	.421	.313	.667	.440	.282
59	.669	.575	.173	.590	.261	.139	.635	.437	.163
60	.722	.570	.202	.658	.131	.182	.692	.253	.194
61	.125	2.969	.233	.106	2.915	.209	.116	2.843	.222
62	.114	2.575	.189	.091	2.424	.173	.103	2.474	.183
63	.074	2.717	.140	.069	2.582	.149	.072	2.648	.144
64	.518	1.029	.253	.437	1.088	.310	.469	1.008	.272
65	.060	2.288	.500	.048	1.702	.500	.019	1.770	.266
66	.775	-.072	.204	.726	-.326	.225	.765	-.187	.218
67	.659	.530	.221	.746	.404	.233	.714	.476	.232
68	.311	2.575	.207	.409	1.896	.202	.353	2.217	.205
69	1.045	1.538	.312	1.053	1.633	.318	1.054	1.594	.318
70	.717	.167	.130	.733	-.037	.172	.718	.046	.144

ตาราง 113 (ต่อ)

ข้อที่	ค่าพารามิเตอร์ข้อสอบ								
	a_1	b_1	c_1	a_2	b_2	c_2	a_t	b_t	c_t
71	.729	.998	.155	.901	1.134	.195	.815	1.080	.180
72	.822	.370	.227	.747	.285	.202	.786	.343	.221
73	.719	2.564	.141	.628	3.102	.143	.669	2.828	.143
74	.931	1.110	.203	.950	.927	.209	.956	1.013	.208
75	.689	1.094	.230	.742	1.149	.234	.732	1.139	.240
76	.750	1.955	.282	.669	1.776	.298	.697	1.862	.289
77	.617	4.12	.087	.731	.317	.160	.671	.348	.120
78	.783	1.342	.187	.493	1.217	.208	.495	1.277	.200
79	.817	1.125	.233	.793	1.054	.231	.810	1.087	.233
80	.919	2.698	.238	.819	3.308	.239	.908	2.932	.240
$\bar{X}$	.608	.966	.149	.623	.944	.158	.618	.773	.147

ตาราง 114 ค่าความยาก (p) และค่าอำนาจจำแนก (r) ของแบบทดสอบ แยกตามเพศชาย-หญิง

ข้อที่	p ₁	r ₁	p ₂	r ₂	p (รวม)	r (รวม)
1	.489	.488	.565	.463	.527	.480
2	.533	.477	.602	.457	.567	.470
3	.488	.469	.540	.433	.514	.453
4	.379	.581	.445	.549	.412	.568
5	.383	.590	.454	.616	.418	.606
6	.607	.410	.678	.630	.643	.392
7	.353	.458	.396	.469	.374	.465
8	.322	.456	.384	.444	.353	.454
9	.253	.529	.294	.505	.274	.519
10	.237	.503	.282	.507	.260	.508
11	.342	.546	.401	.540	.372	.546
12	.327	.631	.389	.618	.358	.627
13	.265	.621	.326	.605	.295	.615
14	.400	.594	.464	.567	.432	.584
15	.756	.577	.814	.542	.785	.565
16	.699	.552	.774	.524	.736	.544
17	.704	.555	.777	.544	.740	.554
18	.581	.546	.654	.533	.618	.544
19	.324	.514	.341	.479	.333	.497
20	.328	.439	.350	.434	.339	.437
21	.618	.490	.627	.487	.623	.487
22	.452	.363	.426	.328	.439	.342
23	.527	.358	.550	.360	.539	.360
24	.287	.255	.282	.293	.285	.272
25	.319	.294	.307	.330	.313	.310
26	.327	.209	.323	.238	.325	.222
27	.499	.328	.485	.307	.493	.315
28	.683	.535	.741	.541	.712	.541
29	.398	.490	.414	.442	.406	.466
30	.523	.414	.563	.387	.543	.403
31	.536	.141	.537	.095	.536	.118
32	.416	.335	.414	.350	.415	.341
33	.303	.122	.301	.114	.302	.117
34	.317	.426	.307	.410	.312	.415
35	.550	.443	.541	.421	.546	.430

ตาราง 114 (ต่อ)

ข้อที่	p_1	r_1	p_2	r_2	p (รวม)	r (รวม)
36	.311	.254	.343	.272	.327	.266
37	.399	.237	.442	.279	.421	.262
38	.548	.312	.570	.275	.559	.295
39	.471	.436	.478	.411	.475	.425
40	.583	.455	.613	.436	.598	.447
41	.687	.345	.730	.357	.708	.354
42	.520	.476	.541	.488	.531	.483
43	.357	.300	.353	.289	.355	.293
44	.744	.437	.806	.403	.775	.427
45	.282	.366	.307	.401	.295	.386
46	.548	.429	.515	.466	.532	.442
47	.708	.429	.740	.397	.724	.416
48	.783	.341	.810	.271	.796	.319
49	.754	.449	.809	.383	.782	.424
50	.314	.468	.304	.511	.309	.486
51	.403	.415	.414	.439	.408	.427
52	.315	.475	.330	.445	.322	.460
53	.340	.317	.357	.334	.348	.336
54	.377	.233	.417	.279	.397	.259
55	.303	.130	.314	.144	.309	.138
56	.549	.496	.536	.512	.542	.500
57	.612	.538	.622	.488	.617	.513
58	.556	.430	.601	.390	.578	.413
59	.490	.470	.531	.457	.510	.465
60	.541	.482	.575	.483	.558	.484
61	.240	.023	.218	.018	.229	.017
62	.216	.068	.212	.049	.214	.059
63	.197	.094	.216	.087	.207	.089
64	.494	.322	.541	.256	.517	.293
65	.609	-.106	.637	-.143	.623	-.120
66	.623	.538	.678	.514	.650	.530
67	.528	.437	.553	.462	.540	.450
68	.382	.195	.392	.245	.387	.219
69	.416	.293	.408	.261	.412	.275
70	.539	.530	.602	.518	.570	.527

ตาราง 114 (ต่อ)

ข้อที่	p_1	r_1	p_2	r_2	p (รวม)	r (รวม)
71	.400	.465	.387	.445	.394	.453
72	.552	.510	.556	.503	.554	.505
73	.208	.224	.193	.170	.220	.195
74	.396	.442	.429	.452	.412	.449
75	.445	.385	.432	.386	.438	.383
76	.377	.231	.419	.236	.398	.236
77	.472	.506	.527	.509	.500	.510
78	.417	.322	.445	.317	.431	.321
79	.427	.403	.438	.407	.433	.404
80	.271	.133	.258	.061	.265	.096

ตาราง 115 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ วิชาคณิตศาสตร์ ชั้นประถมศึกษาปีที่ 4

ข้อที่	ค่าพารามิเตอร์ข้อสอบ				วิธีตรวจสอบ				
	a ₁	b ₁	a ₂	b ₂	LC	ESA	EUA	CSA	CUA
1	.552	.047	.513	-.348	23.17***	-4.79***	-2.82**	-.281	.326
2	.535	-.183	.514	.553	16.51***	-3.97***	-3.20**	-.245	.297
3	.525	.054	.474	-.225	6.97*	-2.27*	-1.81	-.190	.241
4	.704	.534	.679	.256	10.40**	-2.99**	-3.10**	-.242	.307
5	.726	.508	.782	.198	17.86***	-4.09***	2.81**	-.293	.346
6	.441	-.668	.396	-1.224	26.15***	-4.72***	-2.57*	<u>-.406</u>	<u>.445</u>
7	.514	.813	.537	.563	2.88	-1.69	1.15	.092	.144
8	.517	.993	.497	.662	14.39***	-2.89**	-3.73***	-.296	.363
9	.635	1.249	.598	1.067	2.76	-0.46	-0.98	.028	.048
10	.603	1.393	.603	1.128	4.91	-1.69	1.66	-.113	.113
11	.647	.746	.658	.466	7.98*	-2.77**	2.10	-.238	.238
12	.804	.724	.805	.469	7.16*	-2.58**	2.56*	-.204	.250
13	.804	1.022	.780	.757	10.54**	-2.53*	-2.98**	-.253	.307
14	.739	.426	.715	.165	7.81*	-2.69**	-2.62**	-.214	.262
15	.806	-1.152	.755	-1.518	12.68**	-3.16**	-2.53	-.286	.348
16	.731	-.916	.696	-1.334	26.29***	-4.28***	-3.92***	<u>-.418</u>	<u>.457</u>
17	.734	-.936	.743	-1.229	25.63***	-3.43***	3.56***	-.347	.398
18	.678	-.378	.666	-.717	17.59***	-3.79***	-3.79***	-.283	.354
19	.576	.910	.532	.873	4.43	2.10*	-1.48	-.150	.192
20	.481	1.010	.465	.911	0.76	0.85	-0.51	.011	.062
21	.545	-.638	.534	-.674	9.04*	2.54*	-2.56*	-.242	.280
22	.352	.346	.333	.578	30.34***	5.28***	-3.15**	<u>-.437</u>	<u>.526</u>
23	.347	-.208	.356	-.359	0.29	0.27	0.52	.009	.048
24	.257	2.176	.287	2.036	8.51*	0.03	1.36	-.192	.192
25	.279	1.680	.320	1.613	16.81***	0.56	1.88	-.235	.294
26	.200	2.178	.231	1.968	6.01*	-0.27	1.36	-.263	.305
27	.312	-.004	.293	.134	16.12***	3.99***	-1.55	-.214	.214
28	.633	-.910	.664	-1.181	10.51**	1.73	1.87	-.120	.163
29	.512	.554	.470	.512	4.79	-2.13*	-1.56	-.168	.214
30	.426	-.156	.402	-.407	2.07	-1.38	-0.87	-.183	.233
31	.136	-.640	.101	-.857	3.30	-0.18	-1.49	.024	.081
32	.331	.643	.349	.646	9.28**	1.99*	1.22	-.243	.243
33	.126	3.922	.124	4.039	1.35	0.34	-0.32	.014	.052
34	.428	1.183	.423	1.281	17.79***	4.22***	-2.89**	-.265	.298
35	.457	-.307	.436	-.244	20.72***	4.22***	-3.53***	-.270	.341

ตาราง 115 (ต่อ)

ข้อที่	ค่าพารามิเตอร์ข้อสอบ				วิธีตรวจสอบ				
	a ₁	b ₁	a ₂	b ₂	LC	ESA	EUA	CSA	CUA
36	.240	2.025	.265	1.522	4.29	-1.89	1.24	-.153	.194
37	.226	1.100	.266	.548	10.01**	-3.16**	1.93	-.251	.294
38	.313	-.397	.276	-.628	2.04	-0.68	-1.41	.012	.012
39	.443	.165	.411	.150	7.37*	-2.66**	-1.40	-.224	.224
40	.469	-.496	.454	-.677	0.13	-0.14	-0.36	.001	.002
41	.357	-1.420	.382	-1.679	7.90*	-0.65	1.19	-.187	.230
42	.495	-.121	.517	-.217	2.89	1.48	1.12	.091	.143
43	.287	1.275	.276	1.169	6.47*	-1.84	-0.99	-.225	.225
44	.501	-1.468	.469	-2.026	23.76***	-3.40***	-2.79**	<u>-.361</u>	<u>.412</u>
45	.367	1.635	.416	1.295	4.09	-1.73	1.99*	-.178	.224
46	.432	-.310	.494	-.078	60.37***	7.61***	3.97***	<u>-.586</u>	<u>.724</u>
47	.459	-1.307	.428	-1.612	1.66	-1.28	-1.01	-.016	.154
48	.437	-.324	.356	-4.148	13.62**	-2.97**	-2.60**	<u>-.345</u>	<u>.407</u>
49	.522	-1.496	.434	-2.185	20.19***	-4.26***	-3.24**	-.357	.416
50	.491	1.081	.566	1.049	31.89***	1.99	3.13**	<u>-.419</u>	<u>.507</u>
51	.417	.614	.456	.524	7.51*	1.13	1.76	-.179	.214
52	.494	1.071	.461	1.044	3.22	1.77	-1.21	-.136	.172
53	.302	1.371	.349	1.090	4.40	-1.11	1.98*	-.149	.188
54	.219	1.389	.274	.764	10.93**	-3.29***	2.45*	<u>-.340</u>	<u>.401</u>
55	.134	3.705	.142	3.283	0.26	-0.49	0.44	.001	.001
56	.523	-.272	.547	-.179	32.14***	5.66***	4.05***	<u>-.424</u>	<u>.424</u>
57	.609	-.559	.532	-.649	14.43***	1.67	-2.56*	-.201	.243
58	.442	-.349	.398	-.669	6.03*	-2.37*	-1.65	-.175	.197
59	.487	.046	.486	-.169	0.97	-0.97	0.81	-.158	.180
60	.518	-.229	.514	-.405	0.02	-0.12	0.08	.001	.003
61	.315	.040	.248	-.397	14.86***	-3.13**	-2.68**	<u>-.348</u>	<u>.411</u>
62	.605	-.616	.584	-.911	6.37*	-2.38*	-1.78	-.189	.215
63	.448	-.177	.489	-.290	3.66	1.06	1.75	-.132	.132
64	.176	1.634	.236	1.148	7.54*	-1.97	2.51*	<u>-.330</u>	<u>.403</u>
65	.274	.762	.236	.969	9.59**	3.04**	-1.65	-.230	.267
66	.588	-.202	.585	-.505	9.77**	-2.94**	2.99**	-.244	.273
67	.473	.577	.439	.707	23.69***	4.65***	-2.60**	<u>-.402</u>	<u>.451</u>
68	.544	-.280	.534	-.292	12.15***	3.39***	-3.39***	-.271	.304
69	.237	3.441	.174	4.945	8.25*	2.79**	-2.32*	<u>-.359</u>	<u>.420</u>
70	.448	.627	.485	.411	2.52	-0.94	1.56	-.121	.152

ตาราง 115 (ต่อ)

ข้อที่	ค่าพารามิเตอร์ข้อสอบ				วิธีตรวจสอบ				
	a_1	b_1	a_2	b_2	LC	ESA	EUA	CSA	CUA
71	.377	.375	.374	.482	18.59***	3.95***	3.95***	-.279	.308
72	.219	1.390	.226	.897	7.74*	-2.18*	0.87	-.215	.253
73	.539	.138	.549	-.135	5.74*	-2.35*	1.56	-.143	.143
74	.309	.674	.309	.458	0.48	-0.66	0.28	.001	.001
75	.396	.481	.399	.417	3.27	1.51	1.07	.183	.232

จากตาราง 115 จะเห็นว่าข้อสอบที่ตรวจพบว่าทำหน้าที่ต่างกันซ้ำกันทุกวิธี ได้แก่ ข้อที่ 6, 16, 22, 44, 46, 48, 49, 50, 54, 56, 61, 64, 67 และ 69

ตาราง 116 ตัวอย่างการคำนวณการทดสอบความแตกต่างของจำนวนข้อสอบที่ทำหน้าที่ต่างกันที่
ตรวจพบในแต่ละวิธีตรวจสอบ

ข้อที่	ค่าพารามิเตอร์ข้อสอบ				วิธีตรวจสอบ						L _i	L _i ²
	a ₁	b ₁	a ₂	b ₂	LC	ESA	EUA	CSA	CUA			
1	.767	.076	.564	-.407	<u>6.22*</u>	<u>-2.35*</u>	<u>-1.97*</u>	<u>-0.353</u>	<u>0.517</u>	5	25	
2	.911	.489	.846	.452	0.12	0.18	-0.35	0.016	0.110			
3	.635	.918	.588	.743	0.81	-0.45	-0.47	-0.115	0.146			
4	.799	.709	.857	.516	0.59	-0.76	0.47	-0.108	0.120			
5	.693	.614	.733	.218	3.78	-1.90	1.85	-0.289	0.289			
6	.699	-1.205	1.056	-.883	3.09	1.63	1.62	0.246	<u>0.538</u>	1	1	
7	.851	-.237	.739	-.330	0.44	-0.24	-0.66	-0.024	0.203			
8	.672	-.630	.560	-.530	2.18	0.72	-0.86	0.165	0.285			
9	.269	1.944	.306	1.504	0.32	-0.54	0.38	-0.095	0.133			
10	.418	-.205	.749	-.118	5.52	0.69	<u>2.03*</u>	0.085	<u>0.758</u>	2	4	
11	.338	-.398	.252	-.493	0.96	-0.11	-0.91	0.032	0.304			
12	.288	1.823	.351	1.144	1.23	-1.04	0.67	-0.209	0.251			
13	.529	-.524	.621	-.592	0.48	-0.07	0.59	-0.042	0.197			
14	.368	1.528	.508	1.130	1.35	-0.76	1.09	-0.035	0.365			
15	.476	.795	.785	.811	<u>6.40*</u>	0.35	1.83	0.231	<u>0.647</u>	2	4	
16	.195	2.866	.292	1.471	1.90	-1.37	1.09	-0.246	0.371			
17	.399	-.040	.588	-.119	2.14	-0.09	1.34	-0.024	<u>0.483</u>	1	1	
18	.614	-.859	.902	-.706	2.29	0.98	1.46	0.100	<u>0.493</u>	1	1	
19	.474	-.376	.754	-.118	5.21	1.60	1.78	0.223	<u>0.631</u>	1	1	
20	.643	-.039	.651	-.097	0.01	0.01	0.01	0.001	0.002			
รวม				G _j	2	1	2	1	7	ΣL _i	ΣL _i ²	

สูตร

$$Q = \frac{(K-1) \left[\left(K \sum_{j=1}^K G_j^2 \right) - \left(\sum_{j=1}^K G_j \right)^2 \right]}{K \sum_{i=1}^N L_i - \sum_{i=1}^N L_i^2}$$

เมื่อ K แทน จำนวนกลุ่มตัวอย่าง (จำนวนวิธีตรวจสอบ)
 L_i แทน ผลรวมในแนวนอน
 G_j แทน ผลรวมในแนวตั้ง (จำนวนข้อสอบลำเอียงที่ตรวจพบในแต่ละวิธี)

แทนค่า

$$Q = \frac{(5-1) [5(2^2 + 1^2 + 2^2 + 1^2 + 7^2) - (2+1+2+1+7)^2]}{5(13) - 37}$$

$$= \frac{4(295 - 169)}{65 - 37}$$

$$= \frac{504}{28} = 18.00^{**}$$

$$\chi^2_{.05,4} = 9.49, \chi^2_{.01,4} = 13.28$$

ดังนั้น จำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบในแต่ละวิธีแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01

การทดสอบรายคู่

ให้ Q_1 แทน คู่ระหว่าง LC กับ ESA

$$Q_1 = \frac{(2-1)[2(2^2+1^2)-(2+1)^2]}{2(3)-5}$$

$$= \frac{(10-9)}{6-5} = \frac{1}{1}$$

$$= 1.00$$

$$\chi^2_{.05,1} = 3.84, \chi^2_{.01,1} = 6.63$$

ดังนั้น จำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ตรวจพบระหว่างวิธีลอร์ด กับวิธีราชู1
ไม่แตกต่างกัน

ประวัติผู้วิจัย

ชื่อ-สกุล	นายญาณภัทร สีหะมงคล
วัน เดือน ปีเกิด	1 มกราคม 2504
สถานที่เกิด	ตำบลบางทรายใหญ่ อำเภอเมือง จังหวัดมุกดาหาร
สถานที่อยู่ปัจจุบัน	399/13 หมู่ 14 ถนนอินทจินดา ตำบลในเมือง อำเภอพิมาย จังหวัดนครราชสีมา

ประวัติการศึกษา

2526	กศ.บ. (คณิตศาสตร์-ฟิสิกส์) มหาวิทยาลัยศรีนครินทรวิโรฒ มหาสารคาม
2534	ค.ม. (วิจัยการศึกษา) จุฬาลงกรณ์มหาวิทยาลัย
2540	กศ.ด. (การทดสอบและวัดผลการศึกษา) มหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร

ประวัติการทำงาน

2526	โรงเรียนธาตุพนม อำเภอธาตุพนม จังหวัดนครพนม
2527	วิทยาลัยเทคนิคกระนอง อำเภอเมือง จังหวัดระนอง
2529 - ปัจจุบัน	วิทยาลัยเทคนิคหนองคาย อำเภอเมือง จังหวัดหนองคาย 43000

การเปรียบเทียบความสอดคล้องของผลการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน
 ระหว่างวิธี Lord's χ^2 วิธี Raju's Area Measures
 และ วิธี Closed Interval Area

บทคัดย่อ
 ของ
 ญาณภัทร สีหะมงคล

เสนอต่อมหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร เพื่อเป็นส่วนหนึ่งของการศึกษา
 ตามหลักสูตรการศึกษาดุษฎีบัณฑิต สาขาการทดสอบและวัดผลการศึกษา
 กันยายน 2540

การศึกษาครั้งนี้มีความมุ่งหมาย เพื่อเปรียบเทียบความสอดคล้องของผลการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันระหว่างวิธี Lord's χ^2 วิธี Raju's Area Measures และ วิธี Closed Interval Area เมื่อขนาดของกลุ่มตัวอย่าง ความยาวของแบบทดสอบ และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันแบบทดสอบแตกต่างกัน ข้อมูลที่ใช้ในการศึกษาเป็นผลการสอบประเมินคุณภาพและความก้าวหน้าทางการศึกษา วิชาคณิตศาสตร์ ของนักเรียนชั้นประถมศึกษาปีที่ 4 ปีการศึกษา 2536 สังกัดสำนักงานการประถมศึกษาแห่งชาติ จำนวน 11,404 คน เครื่องมือที่ใช้เป็นแบบทดสอบแบบเลือกตอบจำนวน 80 ข้อ แบ่งเป็น 3 ตอน คือ การคิดคำนวณ ความรู้ความเข้าใจ และโจทย์ปัญหา การศึกษาแบ่งออกเป็น 2 ส่วน ได้แก่ ส่วนที่ 1 เป็นการตรวจสอบจำนวนข้อสอบที่ทำหน้าที่ต่างกัน ความสอดคล้องในการตรวจสอบ และความสัมพันธ์ระหว่างวิธีตรวจสอบ และ ส่วนที่ 2 เป็นการตรวจสอบอัตราความไม่สอดคล้องในการตรวจสอบ และอัตราความสอดคล้องในการตรวจสอบ ผลการศึกษาพบว่า

1. จำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ได้จากการตรวจสอบ ระหว่าง วิธี Lord's χ^2 วิธี Raju's Area Measures และ วิธี Closed Interval Area แตกต่างกันเมื่อขนาดของกลุ่มตัวอย่าง และความยาวของแบบทดสอบ ต่างกัน ยกเว้นกรณีกลุ่มตัวอย่าง 500 คน แบบทดสอบความยาว 20, 30 และ 40 ข้อ และกรณีกลุ่มตัวอย่าง 700 คน แบบทดสอบความยาว 40 ข้อ

2. ความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ส่วนมากมีค่าปานกลางถึงต่ำเกือบทุกเงื่อนไขของการศึกษา ยกเว้นกรณีกลุ่มตัวอย่าง 500 คน แบบทดสอบความยาว 30, 40 และ 60 ข้อ และกลุ่มตัวอย่าง 700 คน แบบทดสอบความยาว 40 ข้อ ที่มีความสอดคล้องกันในระดับปานกลางถึงสูงมาก

3. ความสัมพันธ์ระหว่างวิธีการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน ทั้ง 3 วิธี โดยภาพรวมมีสัมประสิทธิ์สหสัมพันธ์ค่อนข้างสูงมาก และมีนัยสำคัญทางสถิติเกือบทุกเงื่อนไขของการศึกษา

4. อัตราความไม่สอดคล้องในการตรวจสอบ เมื่อแบบทดสอบยาวมากขึ้น และมีสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันมากขึ้น อัตราความไม่สอดคล้องชนิดที่ 1 ส่วนใหญ่เพิ่มขึ้น แต่อัตราความไม่สอดคล้องชนิดที่ 2 ส่วนใหญ่ลดลง โดยที่อัตราความไม่สอดคล้องชนิดที่ 1 และอัตราความไม่สอดคล้องชนิดที่ 2 ส่วนมากมีค่าใกล้เคียงกัน

เมื่อกลุ่มตัวอย่างขนาดใหญ่ วิธี Lord's χ^2 และวิธี Raju's Area Measures มีอัตราความไม่สอดคล้องสูงกว่าวิธี Closed Interval Area และเมื่อกลุ่มตัวอย่างขนาดเล็ก วิธี Lord's χ^2 และวิธี Raju's Area Measures มีอัตราความไม่สอดคล้องต่ำกว่าวิธี Closed Interval Area

5. อัตราความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อแบบทดสอบยาวมากขึ้น และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันมากขึ้น อัตราความสอดคล้องชนิดที่ 1 ส่วนใหญ่ลดลง แต่อัตราความสอดคล้องชนิดที่ 2 ส่วนใหญ่เพิ่มขึ้น

A COMPARISON OF CONSISTENCY OF LORD'S χ^2 , RAJU'S AREA MEASURES
AND CLOSED INTERVAL AREA ON DETECTION
DIFFERENTIAL ITEM FUNCTIONING

ABSTRACT

BY

YANAPHAT SEEHAMONGKON

Presented in partial fulfillment of the requirements for the Doctor
of Education Degree in Testing and Education Measurement
at Srinakharinwirot University

September 1997

The purpose of this study was to compare the consistency of three different methods of differential item function detecting ; Lord's χ^2 , Raju's Area Measures and Closed Interval Area. The comparisons were investigated under various situations focusing on sample sizes, test lengths and percentages of DIF-items including in the test differences. The data were an 80 items mathematics test score, from the 1993 Nation-Wide Progressive Assessment. The test measured computational skill, mathematical concept and problem solving. The first part of the study was to investigate the amount of DIF-items, degree of consistency and correlation among the methods. The second part of the study investigated the inconsistency rate and the consistency rate among the three methods. The findings were as follows :-

1. The number of DIF-items were statistical significant different among Lord's χ^2 , Raju's Area Measures and Closed Interval Area with respect to the difference between sample sizes and test lengths. However, at the 20, 30, 40 items tests for 500 persons and 40 items tests for 700 persons, the amount of DIF-items were not different.

2. In most cases, the levels of agreement of detecting DIF among the methods were moderate to low except at the test length 30, 40, 60 items for 500 persons and 40 items for 700 persons the levels of agreement were moderate to very high.

3. All of the correlations between DIF indices were high.

4. In detecting inconsistency rates, the longer test with more DIF-items increased the first inconsistency rate and decreased the second inconsistency rate.

The inconsistency rates of Lord's χ^2 and Raju's Area Measures were larger than Closed Interval Area for the large sample, but were less with small sample.

However, most of the first inconsistency rates were about the same and almost the second inconsistency rates were closely.

5. In detecting consistency rate, the longer test with more DIF-items had lower in the first consistency rate and higher in the second consistency rate