การเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ
ด้วยวิธี IRT-likelihood ratio วิธี Bayesian และวิธี Multiple group CFA

เพ็ญศรี เทียมสุข

ดุษฎีนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปรัชญาดุษฎีบัณฑิต
สาขาวิชาวิจัย วัดผลและสถิติการศึกษา
คณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา
สิงหาคม 2560
ลิขสิทธิ์เป็นของมหาวิทยาลัยบูรพา

คณะกรรมการควบคุมดุษฎีนิพนธ์และคณะกรรมการสอบดุษฎีนิพนธ์ ได้พิจารณาดุษฎีนิพนธ์ของ เพ็ญศรี เทียมสุข ฉบับนี้แล้ว เห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปรัชญาดุษฎีบัณฑิต สาขาวิชาวิจัย วัดผลและสถิติการศึกษา ของมหาวิทยาลัยบูรพาได้

คณะกรรมการควบคุมดุษฎีนิพนธ์

……..อาจารย์ที่ปรึกษาหลัก
(ผู้ช่วยศาสตราจารย์ ดร.เชษฐ์ ศิริสวัสดิ์)

…….. อาจารย์ที่ปรึกษาร่วม
(รองศาสตราจารย์ ดร.ไพรัตน์ วงษ์นาม)

คณะกรรมการสอบดุษฎีนิพนธ์

……..ประธาน
(ดร.อาวีพร ปานทอง)

……..กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.เชษฐ์ ศิริสวัสดิ์)

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

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

คณะศึกษาศาสตร์อนุมัติให้รับดุษฎีนิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปรัชญาดุษฎีบัณฑิต สาขาวิชาวิจัย วัดผลและสถิติการศึกษา ของมหาวิทยาลัยบูรพา

…….. คณบดีคณะศึกษาศาสตร์
(รองศาสตราจารย์ ดร.วิชิต สุรัตน์เรืองชัย)
วันที่ 4 เดือน สิงหาคม พ.ศ. 2560

กิตติกรรมประกาศ

ดุษฎีนิพนธ์ฉบับนี้สำเร็จลงได้ด้วยความกรุณาจาก ผู้ช่วยศาสตราจารย์ ดร.เชษฐ์ ศิริสวัสดิ์ อาจารย์ที่ปรึกษาหลัก และรองศาสตราจารย์ ดร. ไพรรัตน์ วงษ์นาม อาจารย์ที่ปรึกษาร่วม ที่ได้กรุณาให้คำปรึกษา แนะนำ ถ่ายทอดความรู้ และตรวจสอบแก้ไขข้อบกพร่องต่าง ๆ ด้วยความเอาใจใส่ยิ่ง ผู้วิจัยขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ โอกาสนี้

ขอขอบคุณ ดร.อาวีพร ปานทอง ประธานสอบปากเปล่า และผู้ช่วยศาสตราจารย์ ดร.สุรีพร อนุศาสนนันท์ กรรมการ ที่ได้กรุณาให้ข้อเสนอแนะที่เป็นประโยชน์ทำให้ดุษฎีนิพนธ์ฉบับนี้มีความสมบูรณ์ยิ่งขึ้น

ขอกราบขอบพระคุณผู้อำนวยการสถาบันทดสอบการศึกษาแห่งชาติ (องค์การมหาชน) กระทรวงศึกษาธิการ ที่กรุณาอนุญาตให้ใช้ข้อมูลในการทำวิจัย และเจ้าหน้าที่ของสถาบันทุกท่านที่ให้ความช่วยเหลือเป็นอย่างดี

ขอกราบขอบพระคุณบุพการีอันเป็นที่เคารพรักยิ่ง ตลอดจนญาติพี่น้องทุกท่าน รวมทั้งเพื่อน ๆ พี่ ๆ ทุกท่านที่เป็นกำลังใจ ห่วงใยและให้ความช่วยเหลือ และสนับสนุนอย่างดียิ่ง จนทำให้งานวิจัยนี้สำเร็จลุล่วงไปได้ด้วยดี

สุดท้ายนี้ขอกราบขอบพระคุณครู อาจารย์ทุกท่านทั้งในอดีตและปัจจุบันที่ได้อบรมสั่งสอนและประสิทธิ์ประสาทวิชาความรู้ จนผู้วิจัยประสบความสำเร็จในการศึกษาและหน้าที่การงานด้วยดี

คุณค่าและประโยชน์อันพึงมีจากของดุษฎีนิพนธ์ฉบับนี้ ผู้วิจัยขอมอบเป็นกตัญญูกตเวทิตาแด่บุพการี คณาจารย์ทุกท่าน ตลอดจนผู้มีพระคุณต่อผู้วิจัยทุกท่าน ที่ทำให้ข้าพเจ้าประสบความสำเร็จ ในการทำดุษฎีนิพนธ์ฉบับนี้ซึ่งถือว่าเป็นความสำเร็จอีกก้าวหนึ่งของชีวิต และเป็นความภาคภูมิใจของครอบครัว ญาติพี่น้อง เป็นเกียรติ และศักดิ์ศรีแห่งวงศ์ตระกูล และเป็นแบบอย่างที่ดีให้แก่บุตรหลานต่อไปในภายภาคหน้า

เพ็ญศรี เทียมสุข

52810150: สาขาวิชา: วิจัย วัดผลและสถิติการศึกษา; ปร.ด. (วิจัย วัดผลและสถิติการศึกษา)
คำสำคัญ: การทำหน้าที่ต่างกันของข้อสอบ/ วิธี IRT-LIKELIHOOD RATIO/ วิธี BAYESIAN/ วิธี MULTIPLE GROUP CFA

เพ็ญศรี เทียมสุข: การเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี IRT-LIKELIHOOD RATIO วิธี BAYESIAN และวิธี MULTIPLE-GROUP CFA (A COMPARISON OF DIFFERENTIAL ITEM FUNCTIONING BY IRT-LIKELIHOOD RATIO, BAYESIAN AND MULTIPLE-GROUP CFA METHOD) คณะกรรมการควบคุมดุษฎีนิพนธ์: เชษฐ์ ศิริสวัสดิ์, ค.ด., ไพรัตน์ วงษ์นาม, ค.ด., สุรีพร อนุศาสนนันท์, ค.ด. 273 หน้า. ปี พ.ศ. 2560.

การวิจัยครั้งนี้มีวัตถุประสงค์ 1) เพื่อศึกษาผลการประมาณค่าพารามิเตอร์อำนาจจำแนก (a) ของข้อสอบ, ค่าพารามิเตอร์ความยาก (b) ของข้อสอบ, ค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) ด้วยวิธี Maximum likelihood,วิธี Bayesian และวิธี Confirmatory factor analysis 2) เพื่อเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ (DIF) จำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน ได้แก่ เขตกรุงเทพมหานคร และปริมณฑล และนอกเขตกรุงเทพมหานคร และปริมณฑล ด้วยวิธี IRT likelihood ratio,วิธี Bayesian และวิธี Multiple group CFA

ข้อมูลที่ใช้ในการศึกษาครั้งนี้ เป็นคะแนนการสอบวัดผลสัมฤทธิ์ทางการเรียนเพื่อประเมินคุณภาพการศึกษาระดับชาติของนักเรียนชั้นมัธยมศึกษาปีที่ 6 ในรายวิชาคณิตศาสตร์ และภาษาไทย ของสำนักทดสอบทางการศึกษา กระทรวงศึกษาธิการ ปีการศึกษา 2558 ซึ่งได้มาโดยการสุ่มแบบหลายขั้นตอน (Multi-stage random sampling technique) จำนวน 2,400 คน จำแนกเป็นเพศชาย และเพศหญิง ที่อยู่ในเขตกรุงเทพมหานคร และปริมณฑล และนอกเขตกรุงเทพมหานคร และปริมณฑล

วิเคราะห์ค่าพารามิเตอร์ข้อสอบตามโมเดลพารามิเตอร์ 2 ตัว (2PL) ได้แก่ ค่าพารามิเตอร์อำนาจจำแนกของข้อสอบ ค่าพารามิเตอร์ความยากของข้อสอบ และค่าพารามิเตอร์ความสามารถของผู้สอบ ด้วยวิธี Maximum likelihood, วิธี Bayesian และวิธี Confirmatory factor analysis วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี IRT-likelihood ratio,วิธี Bayesian และวิธี Multiple group CFA โดยใช้โปรแกรมสำเร็จรูป ได้แก่ โปรแกรม IRT PRO, โปรแกรม Open BUGS และโปรแกรม Mplus ตามลำดับ

ผลการวิจัย พบว่า

1. ผลการประมาณค่าพารามิเตอร์อำนาจจำแนกของข้อสอบ ค่าพารามิเตอร์ความยากของข้อสอบ และค่าพารามิเตอร์ความสามารถของผู้สอบ (DIF) ในวิชาคณิตศาสตร์ และ

ภาษาไทย ด้วยวิธี Maximum likelihood, วิธี Bayesian และวิธี Confirmatory factor analysis พบว่า ผลการประมาณค่าพารามิเตอร์อำนาจจำแนกด้วยวิธี Maximum likelihood,วิธี Bayesian และวิธี Confirmatory factor analysis มีความสัมพันธ์ทางบวกในระดับสูงมาก อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 การประมาณค่าพารามิเตอร์ความยากของข้อสอบวิชาคณิตศาสตร์ พบว่า วิธี Maximum likelihood และวิธี Confirmatory factor analysis มีความสัมพันธ์ทางบวกในระดับสูงมากอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ส่วนวิธี Bayesian มีความสัมพันธ์ทางบวกในระดับปานกลางค่อนข้างน้อยกับวิธี Maximum likelihood และวิธี Confirmatory factor analysis ผลการประมาณค่าพารามิเตอร์ความยากของข้อสอบวิชาภาษาไทยด้วยวิธี Maximum likelihood, วิธี Bayesian และวิธี Confirmatory factor analysis มีความสัมพันธ์ทางบวกในระดับสูงมาก อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ผลการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบวิชาคณิตศาสตร์ และภาษาไทยด้วยวิธี Maximum likelihood, วิธี Bayesian และวิธี Confirmatory factor analysis พบว่า มีความสัมพันธ์ทางบวกในระดับสูงมาก และมีนัยสำคัญทางสถิติที่ระดับ .05

2. วิธีตรวจสอบที่พบการทำหน้าที่ต่างกันของข้อสอบมากที่สุด วิชาคณิตศาสตร์ จำแนกตามเพศ คือ วิธี Bayesian รองลงมา คือ วิธี Multiple group CFA และวิธี IRT likelihood ratio ตามลำดับ เมื่อจำแนกตามเขตที่ตั้งทางภูมิศาสตร์ของโรงเรียน พบว่า วิธี Multiple group CFA พบการทำหน้าที่ต่างกันของข้อสอบมากที่สุด รองลงมา คือ วิธี Bayesian และวิธี IRT likelihood ratio

สำหรับวิชาภาษาไทย จำแนกตามเพศ และที่ตั้งทางภูมิศาสตร์ของโรงเรียน พบว่า วิธี Bayesian พบการทำหน้าที่ต่างกันของข้อสอบมากที่สุด รองลงมา คือ วิธี IRT likelihood ratio ส่วนวิธี Multiple group CF ตรวจไม่พบการทำหน้าที่ต่างกันของข้อสอบ

52810150 MAJOR: EDUCATIONAL RESEARCH, MEASUREMENT AND STATISTICS; Ph.D. (EDUCATIONAL RESEARCH, MEASUREMENTAND STATISTICS)
KEYWORDS: DIFERENTIAL ITEM FUNCTIONING/ IRT-LIKELIHOOD RATIO METHOD/ BAYESIAN METHOD/ MULTIPLE GROUP CONFIRMATORY FACTOR ANALYSIS METHOD

PENSRI THIAMSUK: A COMPARISON OF DIFFERENTIAL ITEM FUNCTIONING BY IRT-LIKELIHOOD RATIO, BAYESIAN AND MULTIPLE GROUP CONFIRMATORY FACTOR ANALYSIS METHOD. DISSERTATION ADVISORS: CHADE SIRISAWAT, Ph.D., PAIRAT WONGNAM, Ed.D., SUREEPORN NUSANSANANAN, Ph.D. 273 P. 2017.

This study aimed to: 1) compare the discriminant item parameter (a), difficulty item parameter (b) and ability person parameter (θ) of Mattayomsuksa 6 students' national test scores of Mathematics and Thai Language among Maximum likelihood, Bayesian and Confirmatory factor analysis methods: 2) compare the differential item functioning (DIF) of Mathematics and Thai Language test items on gender and school geographical locations variables with which estimate by IRT likelihood ratio, Bayesian and Multiple group CFA methods.

The national test scores of Mattayomsuksa 6 students in 2015 academic year were drawn from National test database, Ministry of Education by using the multi-stage random sampling techniques 2,400 cases of test scores were divided into two groups: gender (male & female) and school geographical location (Bangkok and Metropolitan areas & non Bangkok and Metropolitan areas) variables.

The analysis procedures to estimates the item parameter by IRT model (2 PL) include discrimination item parameters, difficulty item parameter and ability person parameters, The DIF by Maximum likelihood, Bayesian and Confirmatory factor analysis methods were used along with IRT PRO program software, Open BUGS program software and Mplus program software for data analysis, All parameter estimaters and DIF detection results were compared in terms of correlation and congruence when met DIF.

Results of the study were as follows: The discrimination parameters, difficulty parameters and ability person parameter of Mathematics and Thai Language test scores which were estimated by Maximum likelihood, Bayesian and Confirmatory factor analysis methods

revealed that the correlation of discriminant parameter estimate by Maximum likelihood, Bayesian and Confirmatory factor methods were very highly positive correlated with .05 statistical significant level. The correlation of difficulty parameters of Mathematics test scores estimated by Maximum likelihood and Confirmatory factor analysis methods were very highly positive correlated with .05 statistical significant level. While Bayesian method was positive correlated with slightly less to moderate correlated with Maximum likelihood and Confirmatory factor analysis methods with .05 statistical significant level. The difficulty parameters of Thai Language test scores estimated by Maximum likelihood, Bayesian and Confirmatory factor methods were very highly positive correlation with .05 statistical significant level. The correlation among ability person parameter which estimating by Maximum likelihood, Bayesian and Confirmatory factor methods were very high with .05 statistical significant level in all subject test score.

According to DIF detection. Bayesian method was very sensitive to detect DIF in Mathematics test scores option to gender variables, Multiple group CFA methods, and IRT likelihood ratio, respectively. Multiple group CFA methods was very sensitive to detect DIF in Mathematics test scores option to school geographical location variables, Bayesian method, and IRT likelihood ratio, respectively. Bayesian method was very sensitive to detect DIF in Thai Language test scores option to gender and school geographical location variables and IRT likelihood ratio methods. While the Multiple group CFA methods did not meet DIF.

สารบัญ

หน้า

บทคัดย่อภาษาไทย..... ง

บทคัดย่อภาษาอังกฤษ..... ฉ

สารบัญ..... ซ

สารบัญตาราง..... ญ

สารบัญภาพ..... ฒ

บทที่

1 บทนำ..... 1

ความเป็นมา..... 1

คำถามการวิจัย..... 11

วัตถุประสงค์ของการวิจัย..... 11

สมมติฐานในการวิจัย..... 11

ประโยชน์ที่คาดว่าจะได้รับ..... 12

ขอบเขตการวิจัย..... 12

นิยามศัพท์เฉพาะ..... 13

2 เอกสารและงานวิจัยที่เกี่ยวข้อง..... 16

ตอนที่ 1 แนวคิดทฤษฎีการตอบสนองข้อสอบ..... 16

ตอนที่ 2 การวิเคราะห์ข้อสอบตามแนวทฤษฎีการตอบสนองข้อสอบ IRT..... 24

ตอนที่ 3 การทำหน้าที่ต่างกันของข้อสอบ..... 33

ตอนที่ 4 ทฤษฎีการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ..... 47

ตอนที่ 5 แนวคิดการวิเคราะห์อัตราส่วนความควรจะเป็น..... 55

ตอนที่ 6 แนวคิดการวิเคราะห์ด้วยวิธี BAYESIAN..... 60

ตอนที่ 7 แนวคิดการวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ..... 71

ตอนที่ 8 การทดสอบวัดผลสัมฤทธิ์ระดับชาติของนักเรียน O-NET..... 81

ตอนที่ 9 งานวิจัยที่เกี่ยวข้องกับทำหน้าที่ต่างกันของข้อสอบ..... 93

3 วิธีดำเนินการวิจัย..... 106

ประชากรและกลุ่มตัวอย่าง..... 106

เครื่องมือที่ใช้ในการวิจัย..... 112

สารบัญ (ต่อ)

บทที่		หน้า
4	ผลการวิเคราะห์ข้อมูล	123
	สัญลักษณ์ที่ใช้ในการวิเคราะห์ข้อมูล	123
	สัญลักษณ์ที่ใช้แทนตัวแปรสังเกตได้	124
	ตอนที่ 1 การวิเคราะห์ค่าสถิติพื้นฐาน	124
	ตอนที่ 2 เปรียบเทียบผลการประมาณค่าพารามิเตอร์อำนาจจำแนกของข้อสอบ	129
	ตอนที่ 3 เปรียบเทียบผลการประมาณค่าพารามิเตอร์ความยากของข้อสอบ	138
	ตอนที่ 4 การเปรียบเทียบผลการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบ	149
	ตอนที่ 5 ผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบจำแนกตามเพศและสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน	154
	ตอนที่ 6 เปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ด้วยวิธี IRT-LR วิธี BAYESIAN และวิธี MGCFA จำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน	189
5	สรุป อภิปรายผล และข้อเสนอแนะ	200
	สรุปผลการวิจัย	201
	อภิปรายผลการวิจัย	208
	ข้อเสนอแนะจากการวิจัย	211
บรรณานุกรม		213
ภาคผนวก		227
	ภาคผนวก ก	228
	ภาคผนวก ข	230
	ภาคผนวก ค	235
	ภาคผนวก ง	238
	ภาคผนวก จ	241
	ภาคผนวก ฉ	244
	ภาคผนวก ช	248
ประวัติย่อของผู้วิจัย		273

สารบัญตาราง

ตารางที่		หน้า
2-1	การเปรียบเทียบวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ	48
2-2	แสดงเนื้อหาการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET)	82
2-3	จำนวนข้อสอบ O-NET มัธยมศึกษาปีที่ 6 แยกตามรูปแบบของข้อสอบ ประจำปีการศึกษา 2558	85
2-4	ค่าคะแนนเฉลี่ยของผลการทดสอบ O-NET ชั้นมัธยมศึกษาปีที่ 6 ปีการศึกษา 2558 จำแนกตามสังกัด	87
2-5	แสดงค่าคะแนนเฉลี่ยของผลการทดสอบ O-NET ชั้นมัธยมศึกษาปีที่ 6 ปีการศึกษา 2558 จำแนกตามขนาดโรงเรียน ที่ตั้ง และภูมิภาค	90
3-1	จำนวนโรงเรียน และจำนวนการสุ่มตัวอย่างโรงเรียนในเขตกรุงเทพมหานครและปริมณฑล	108
3-2	ประชากรและกลุ่มตัวอย่างนักเรียนในเขตกรุงเทพมหานครและปริมณฑล จำแนกตามจังหวัด และโรงเรียน	109
3-3	จำนวนจังหวัด จำนวนโรงเรียน และจำนวนการสุ่มตัวอย่างโรงเรียนนอกเขตกรุงเทพมหานคร และปริมณฑล	110
3-4	เขตที่ตั้งทางภูมิศาสตร์ จำนวนจังหวัด และจำนวนการสุ่มตัวอย่างโรงเรียนนอกเขตกรุงเทพมหานคร และปริมณฑล	110
3-5	การทดสอบการทำหน้าที่ต่างกันด้วย No-DIF model และ DIF model	121
4-1	แสดงจำนวนผู้สอบ ค่าต่ำสุด ค่าสูงที่สุด ค่าเฉลี่ย ค่าส่วนเบี่ยงเบนมาตรฐาน ค่าความเบ้ และค่าความโด่ง จากคะแนนการตอบแบบทดสอบวิชาคณิตศาสตร์ จำแนกตามเพศผู้สอบ	125
4-2	แสดงจำนวนผู้สอบ ค่าต่ำสุด ค่าสูงสุด ค่าเฉลี่ย ค่าส่วนเบี่ยงเบนมาตรฐาน ค่าความเบ้ และค่าความโด่ง จากคะแนนการตอบแบบทดสอบวิชาภาษาไทย จำแนกตามเพศผู้สอบ	126
4-3	แสดงจำนวนผู้สอบ ค่าต่ำสุด ค่าสูงสุด ค่าเฉลี่ย ค่าส่วนเบี่ยงเบนมาตรฐาน ค่าความเบ้ และค่าความโด่ง จากคะแนนการตอบแบบทดสอบวิชาคณิตศาสตร์ จำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียนผู้สอบ	127

สารบัญตาราง (ต่อ)

ตารางที่ หน้า

4-4 แสดงจำนวนผู้สอบ ค่าต่ำสุด ค่าสูงสุด ค่าเฉลี่ย ค่าส่วนเบี่ยงเบนมาตรฐาน ค่าความเบ้ และค่าความโด่ง จากคะแนนการตอบแบบทดสอบวิชาภาษาไทย จำแนกตาม สถานที่ตั้งทางภูมิศาสตร์ของโรงเรียนผู้สอบ..... 128

4-5 ผลการประมาณค่าพารามิเตอร์อำนาจจำแนกของข้อสอบวิชาคณิตศาสตร์ ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA..... 129

4-6 ผลการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าพารามิเตอร์อำนาจจำแนก ของข้อสอบวิชาคณิตศาสตร์ ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA..... 131

4-7 ผลการประมาณค่าพารามิเตอร์อำนาจจำแนกของข้อสอบวิชาภาษาไทย ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA..... 133

4-8 ผลการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าพารามิเตอร์อำนาจจำแนก ของข้อสอบวิชาภาษาไทย ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA..... 137

4-9 ผลการประมาณค่าพารามิเตอร์ความยากของข้อสอบวิชาคณิตศาสตร์ ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA..... 139

4-10 ผลการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าพารามิเตอร์ความยากของ ข้อสอบวิชาคณิตศาสตร์ ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA..... 141

4-11 ผลการประมาณค่าพารามิเตอร์ความยากของข้อสอบวิชาภาษาไทย ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA..... 143

4-12 ผลการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าพารามิเตอร์ความยากของ ข้อสอบวิชาภาษาไทย ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA..... 147

4-13 สถิติพื้นฐานของผลการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบวิชา คณิตศาสตร์ ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA..... 149

4-14 ผลการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าพารามิเตอร์ความสามารถ ของผู้สอบวิชาคณิตศาสตร์ ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA..... 150

4-15 สถิติพื้นฐานของผลการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบวิชา ภาษาไทย ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA..... 152

4-16 ผลการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าพารามิเตอร์ความสามารถ ของผู้สอบวิชาภาษาไทย ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA..... 153

สารบัญตาราง (ต่อ)

ตารางที่ หน้า

4-17 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์ ด้วยวิธี IRT-LR ประยุกต์ใช้โปรแกรม IRT-PRO จำแนกตามเพศ 155

4-18 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทย ด้วยวิธี IRT-LR ประยุกต์ใช้โปรแกรม IRT-PRO จำแนกตามเพศ 157

4-19 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์ด้วยวิธี IRT-LR ประยุกต์ใช้โปรแกรม IRT-PRO จำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน..... 160

4-20 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทยด้วยวิธี IRT-LR ประยุกต์ใช้โปรแกรม IRT-PRO จำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน..... 162

4-21 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์ ด้วยวิธี BAYSIAN จำแนกตามเพศ 166

4-22 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทย ด้วยวิธี BAYSIAN จำแนกตามเพศ 171

4-23 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์ ด้วยวิธี BAYSIAN จำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน 177

4-24 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทย ด้วยวิธี BAYSIAN จำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน 181

4-25 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์ ด้วยวิธี MGCFA จำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน 187

4-26 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทย ด้วยวิธี MGCFA จำแนกตามตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน 188

4-27 เปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์ด้วยวิธี IRT-LR วิธี BAYESIAN และวิธี MGCFA จำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน 189

4-28 แสดงความสอดคล้องของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์ ด้วยวิธี IRT-LR วิธี BAYSIAN และวิธี MGCFA 192

สารบัญตาราง (ต่อ)

ตารางที่ หน้า

4-29 เปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทย ด้วยวิธี IRT-LR วิธี BAYESIAN และวิธี MGCFA จำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน 194

4-30 แสดงความสอดคล้องของการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทย ด้วยวิธี IRT-LR, BAYSIAN และ MGCFA 198

5-1 แสดงความสอดคล้องของการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์ ด้วยวิธี IRT-LR วิธี BAYSIAN และวิธี MGCFA 203

5-2 แสดงความสอดคล้องของการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทย ด้วยวิธี IRT-LR วิธี BAYSIAN และวิธี MGCFA 205

5-3 วิเคราะห์ความเป็นเอกมิติของโครงสร้างคุณลักษณะแฝงวิชาคณิตศาสตร์ จำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน 209

5-4 วิเคราะห์ความเป็นเอกมิติของโครงสร้างคุณลักษณะแฝงวิชาภาษาไทย จำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน 211

ข-1 ผลการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบวิชาคณิตศาสตร์ ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA 231

ข-2 ผลการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบวิชาคณิตศาสตร์ ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA 233

ค-1 ผลการประมาณค่าพารามิเตอร์อำนาจจำแนกของข้อสอบวิชาคณิตศาสตร์ จำแนกตามเพศ ด้วยวิธี ML 236

ง-1 ผลการประมาณค่าพารามิเตอร์อำนาจจำแนกของแบบทดสอบ วิชาคณิตศาสตร์ จำแนกตามเพศ ด้วยวิธี BAYESIAN 239

จ-1 ผลการประมาณค่าพารามิเตอร์อำนาจจำแนกของแบบทดสอบ วิชาคณิตศาสตร์ จำแนกตามเพศ ด้วยวิธี CFA 242

ช-1 การสกัดองค์ประกอบหลักวิชาคณิตศาสตร์ จำแนกเป็นเพศชาย 249

ช-2 การสกัดองค์ประกอบหลักวิชาคณิตศาสตร์ จำแนกเป็นเพศหญิง 251

ช-3 การสกัดองค์ประกอบหลักวิชาคณิตศาสตร์ จำแนกเป็นเขตกรุงเทพมหานครและปริมณฑล 253

สารบัญตาราง (ต่อ)

ตารางที่ หน้า

ช-4 การสกัดองค์ประกอบหลักวิชาคณิตศาสตร์ จำแนกเป็นนอกเขตกรุงเทพมหานครและปริมณฑล..... 255

ช-5 การสกัดองค์ประกอบหลักวิชาภาษาไทย จำแนกเป็นเพศชาย..... 257

ช-6 การสกัดองค์ประกอบหลักวิชาภาษาไทย จำแนกเป็นเพศหญิง..... 261

ช-7 การสกัดองค์ประกอบหลักวิชาภาษาไทย จำแนกเป็นเขตกรุงเทพมหานครและปริมณฑล..... 265

ช-8 การสกัดองค์ประกอบหลักวิชาภาษาไทย จำแนกเป็นนอกเขตกรุงเทพมหานครและปริมณฑล..... 269

สารบัญภาพ

ภาพที่		หน้า
1-1	กรอบแนวคิดในการวิจัย	15
2-1	ฟังก์ชันโมเดลการตอบสนองข้อสอบแบบโลจิสติกส์ 1 พารามิเตอร์	26
2-2	ฟังก์ชันโมเดลการตอบสนองข้อสอบแบบโลจิสติกส์ 2 พารามิเตอร์	27
2-3	ฟังก์ชันโมเดลการตอบสนองข้อสอบแบบโลจิสติกส์ 3 พารามิเตอร์	28
2-4	โค้งลักษณะข้อสอบโลจิสติก	36
2-5	เส้นโค้งลักษณะข้อสอบที่มีความแตกต่างของตำแหน่งที่เป็นตัวแปรต่อเนื่อง	37
2-6	ตัวอย่างของข้อสอบที่ไม่แสดง DIF	38
2-7	ตัวอย่างของข้อสอบที่แสดง DIF	39
2-8	ตัวอย่างของข้อสอบที่แสดง DIF แบบ Nonunifrom DIF	40
2-9	ลักษณะของโค้งคุณลักษณะข้อสอบปกติ	43
2-10	การทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป	43
2-11	การทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป	44
2-12	แสดงแผนภาพการแจกแจงความหนาแน่นของฟังก์ชั่นก่อนหน้าและฟังก์ชั่นภายหลังที่ไม่ทราบค่าเฉลี่ย (_) ของการแจกแจงปกติ	66
2-13	แสดงการแจกแจงก่อนของห่วงโซ่มาร์คอฟที่แสดงการลู่เข้า (a) และการไม่ลู่เข้า (b)	69
4-1	กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ค่าอำนาจจำแนกของข้อสอบวิชาคณิตศาสตร์ที่ประมาณค่าด้วยวิธี ML และวิธี BAYESIAN	132
4-2	กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ค่าอำนาจจำแนกของข้อสอบวิชาคณิตศาสตร์ที่ประมาณค่าด้วยวิธี ML และวิธี CFA	132
4-3	กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ค่าอำนาจจำแนกของข้อสอบวิชาคณิตศาสตร์ที่ประมาณค่าด้วยวิธี BAYESIAN และวิธี CFA	133
4-4	กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ค่าอำนาจจำแนกข้อสอบวิชาภาษาไทยที่ประมาณค่าด้วยวิธี ML และวิธี BAYESIAN	137
4-5	กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ค่าอำนาจจำแนกข้อสอบวิชาภาษาไทยที่ประมาณค่าด้วยวิธี ML และวิธี CFA	138
4-6	กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ค่าอำนาจจำแนกข้อสอบวิชาภาษาไทยที่ประมาณค่าด้วยวิธี BAYESIAN และวิธี CFA	138

สารบัญภาพ (ต่อ)

ภาพที่ หน้า

4-7 กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ความยากของข้อสอบวิชาคณิตศาสตร์ที่ประมาณค่าด้วยวิธี ML และวิธี BAYESIAN..... 142

4-8 กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ความยากของข้อสอบวิชาคณิตศาสตร์ที่ประมาณค่าด้วยวิธี ML และวิธี CFA..... 142

4-9 กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ความยากของข้อสอบวิชาคณิตศาสตร์ที่ประมาณค่าด้วยวิธี BAYESIAN และวิธี CFA..... 142

4-10 กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ค่าความยากของข้อสอบวิชาภาษาไทยที่ประมาณค่าด้วยวิธี ML และวิธี BAYESIAN..... 147

4-11 กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ค่าความยากข้อสอบวิชาภาษาไทยที่ประมาณค่าด้วยวิธี ML และวิธี CFA..... 148

4-12 กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ค่าความยากของข้อสอบวิชาภาษาไทยที่ประมาณค่าด้วยวิธี BAYESIAN และวิธี CFA..... 148

4-13 กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ค่าความสามารถของผู้สอบวิชาคณิตศาสตร์ที่ประมาณค่าด้วยวิธี ML และวิธี BAYESIAN..... 151

4-14 กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ค่าความสามารถของผู้สอบวิชาคณิตศาสตร์ที่ประมาณค่าด้วยวิธี ML และวิธี CFA..... 151

4-15 กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ค่าความสามารถของผู้สอบวิชาคณิตศาสตร์ที่ประมาณค่าด้วยวิธี BAYESIAN และวิธี CFA..... 151

4-16 กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ค่าความสามารถของผู้สอบวิชาภาษาไทยที่ประมาณค่าด้วยวิธี ML และวิธี BAYESIAN..... 153

4-17 กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ค่าความสามารถของผู้สอบวิชาภาษาไทยที่ประมาณค่าด้วยวิธี ML และวิธี CFA..... 154

4-18 กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ค่าความสามารถของผู้สอบวิชาภาษาไทยที่ประมาณค่าด้วยวิธี BAYESIAN และวิธี CFA..... 154

4-19 คลื่นความถี่ (History) การเกิดการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์จำแนกตามเพศ ด้วยวิธี BAYESIAN..... 169

สารบัญภาพ (ต่อ)

ภาพที่ หน้า

4-20 คลื่นความถี่ (Density) ของการเกิดการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์ จำแนกตามเพศ ด้วยวิธี BAYESIAN..... 170

4-21 คลื่นความถี่ (History) การเกิดการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทย จำแนกตามเพศ ด้วยวิธี BAYESIAN..... 175

4-22 คลื่นความถี่ (Density) ของการเกิดการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทย จำแนกตามเพศ ด้วยวิธี BAYESIAN..... 176

4-23 คลื่นความถี่ (History) การเกิดการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์ จำแนกตามเพศ ด้วยวิธี BAYESIAN..... 179

4-24 คลื่นความถี่ (Density) ของการเกิดการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์ จำแนกตามเพศ ด้วยวิธี BAYESIAN..... 180

4-25 คลื่นความถี่ (History) การเกิดการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทย จำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน ด้วยวิธี BAYESIAN..... 185

4-26 คลื่นความถี่ (Density) ของการเกิดการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทย จำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน ด้วยวิธี BAYESIAN..... 186

ฉ-1 ตัวอย่าง History plot ของพารามิเตอร์ที่ใช้ในการตัดสิน DIF ที่ประมาณค่าด้วยวิธี BAYESIAN..... 245

ฉ-2 ตัวอย่าง Density plot ของพารามิเตอร์ที่ใช้ในการตัดสิน DIF ที่ประมาณค่าด้วยวิธี BAYESIAN..... 247

บทที่ 1
บทนำ

ความเป็นมา

ตามพระราชบัญญัติการศึกษาแห่งชาติ พุทธศักราช 2542 มาตรา 47 ได้กำหนดให้มีระบบการประกันคุณภาพการศึกษา เพื่อพัฒนาคุณภาพ และมาตรฐานการศึกษาในทุกระดับ การประเมินคุณภาพการศึกษาขั้นพื้นฐาน เป็นส่วนหนึ่งของการประกันคุณภาพภายในสถานศึกษา หลักสูตรการศึกษาขั้นพื้นฐาน พุทธศักราช 2544 ได้กำหนดแนวทางการวัดและประเมินผลการเรียนรู้เพื่อให้ได้ข้อมูลสารสนเทศที่แสดงพัฒนาการ ความก้าวหน้า และความสำเร็จของการเรียนรู้ของผู้เรียน ซึ่งสถานศึกษาต้องจัดให้มีการประเมินผลสัมฤทธิ์ทางการเรียนให้เป็นมาตรฐานเดียวกัน ทั้งในระดับสถานศึกษา ระดับเขตพื้นที่การศึกษา และระดับชาติ ต่อมากระทรวงศึกษาธิการได้ออกประกาศเรื่องการใช้ผลการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐานเป็นองค์ประกอบหนึ่งในการตัดสินผลการเรียนของผู้เรียนที่จบการศึกษาตามหลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน พุทธศักราช 2551 และให้ใช้หลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน พุทธศักราช 2551ซึ่งเป็นหลักสูตรอิงมาตรฐาน (Standard-base curriculum) เพื่อให้สถานศึกษาใช้เป็นมาตรฐานการเรียนรู้ และเป็นกรอบแนวทางในการจัดทำหลักสูตรสถานศึกษา ผลสัมฤทธิ์ทางการเรียนจึงเป็นสิ่งที่สะท้อนถึงคุณภาพผู้เรียนที่สถานศึกษาต้องรับผิดชอบ

การทดสอบทางการศึกษาระดับชาติ (O-NET: Ordinary national educational test) เป็นการทดสอบเพื่อวัดผลทางการศึกษาระดับประเทศ โดยการทดสอบความรู้รวบยอดปลายช่วงชั้น (6 ภาคเรียน) ดำเนินการโดยสถาบันทดสอบทางการศึกษาระดับชาติ (องค์การมหาชน) โดยใช้กรอบการประเมินเป็นไปตามมาตรฐานการเรียนรู้ของหลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน พุทธศักราช 2551 เพื่อให้การจัดการเรียนการสอนที่มีคุณภาพสอดคล้องกับเป้าหมายของหลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน พุทธศักราช 2551 ดังนั้นผลการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET) จึงเป็นเครื่องบ่งบอกคุณภาพของผู้เรียนที่เกิดจากการจัดการเรียนรู้ และการพัฒนาคุณภาพของสถาบันการศึกษา

ในปัจจุบันนี้คุณภาพการศึกษาของไทยก็ยังคงเป็นปัญหา เนื่องจากผลสัมฤทธิ์ทางการเรียนเป็นสิ่งที่สะท้อนถึงคุณภาพผู้เรียนที่สถานศึกษาต้องรับผิดชอบจากการประเมินผลนักเรียนนานาชาติ ตามโครงการ PISA จัดโดยกลุ่ม OECD วัดการเรียนรู้เรื่องการอ่าน คณิตศาสตร์ วิทยาศาสตร์ ของนักเรียนวัย 15 ปี (วัยจบการศึกษาพื้นฐาน) นักเรียนไทยประเมินการรู้เรื่อง

การอ่านได้คะแนน 409 คะแนน จากคะแนนมาตรฐานที่ 493 คะแนน อยู่ในช่วงลำดับที่ 56-60 ซึ่งต่ำกว่าค่าเฉลี่ย OECD มากกว่าหนึ่งระดับ ประเทศในเอเชียที่ร่วมการประเมินและมีคะแนนเฉลี่ยต่ำกว่าไทยมีเพียงอินโดนีเซียประเทศเดียวเท่านั้น (PISA Thailand สสวท. ผลการประเมิน PISA2015 การอ่านคณิตศาสตร์และวิทยาศาสตร์)

ปัญหาเรื่องการเรียนวิชาคณิตศาสตร์ที่ผ่านมา นักเรียนจำนวนไม่น้อยยังด้อยความสามารถเกี่ยวกับทักษะและกระบวนการต่าง ๆ ซึ่งปัญหาหนึ่งคือ นักเรียนไม่สามารถแสดงหรืออ้างอิงเหตุผลได้ ทำให้นักเรียนไม่สามารถนำความรู้ทางคณิตศาสตร์ไปประยุกต์ใช้ได้ (สถาบันส่งเสริมการสอนวิทยาศาสตร์และเทคโนโลยี, 2555, หน้า 1) ซึ่งพิจารณาได้จากคะแนนสอบวัดความถนัดทางคณิตศาสตร์ ซึ่งเป็นข้อสอบที่เน้นการคิดวิเคราะห์ และการให้เหตุผล พบว่า นักเรียนยังมีคะแนนวิชาคณิตศาสตร์อยู่ในระดับต่ำ (ชินภัทร ภูมิรัตน์, 2556) และอัมพร ม้าคะนอง (2553, หน้า 50) กล่าวว่า การที่นักเรียนได้คำตอบที่ถูกต้องแต่ใช้เหตุผลผิด เป็นอันตรายอย่างยิ่งต่อการเรียนรู้วิชาคณิตศาสตร์ เนื่องจากเมื่อนักเรียนได้คำตอบที่ถูกต้องแล้ว ผู้สอนอาจไม่ได้ให้โอกาสนักเรียนแสดงเหตุผล ซึ่งทำให้นักเรียนไม่ทราบเหตุผลว่าผิดนั้นผิดอย่างไร นอกจากนี้การสอนคณิตศาสตร์ยังไม่ประสบผลสำเร็จเท่าที่ควร ดังจะเห็นได้จากผลการทดสอบการศึกษาแห่งชาติขั้นพื้นฐาน (O-NET) ซึ่งเป็นการทดสอบวัดผลสัมฤทธิ์ทางการเรียนของนักเรียนชั้นมัธยมศึกษาปีที่ 6 ปีการศึกษา 2554-2556 พบว่า ไม่ผ่านเกณฑ์ร้อยละ 50 โดยเนื้อหาส่วนหนึ่งที่ใช้ในการทดสอบ คือ เนื้อหาเรื่องความสัมพันธ์ (ณัฏฐปิญชาน์ พิชญาชมชื่น, 2558, หน้า 78) และผลการทดสอบภาพรวมของประเทศได้คะแนนอยู่ในระดับต่ำ จึงทำให้เกิดเสียงวิพากษ์วิจารณ์ และสะท้อนปัญหาในแง่ต่าง ๆ ของประเทศ ได้แก่ ปัญหาคุณภาพครูไม่ตรงสาย, ปัญหาการออกข้อสอบที่ไม่ตรงกับหลักสูตร ความเบื่อหน่ายของเด็กที่ต้องสอบติดต่อกัน รวมถึงการขาดแรงจูงใจในการสอบ นักวิชาการทางการศึกษาบางท่านให้ข้อคิดเห็นว่าการสอบ O-NET เป็นการสอบแข่งขันของเด็กในเมืองมากกว่าเด็กในพื้นที่ต่างจังหวัด ปัญหาการสอบตกต่ำ เพราะว่าเด็กไม่คุ้นเคยกับรูปแบบของข้อสอบที่ต้องเป็นแบบสัมพันธ์กัน ข้อสอบบางข้อมีความลำเอียงเชิงวัฒนธรรม บางคำถามสำหรับเด็กฐานะเศรษฐกิจดี หรือเด็กในเมืองตอบได้ แต่เด็กชนบทหรือต่างจังหวัดไม่ทราบคำตอบ และตอบข้อสอบไม่ได้ (ชินภัทร ภูมิรัตน์, 2554)

การวัดและการประเมินผลการเรียนรู้เป็นกระบวนการหนึ่งที่ให้ผู้สอนใช้พัฒนาคุณภาพผู้เรียนเพราะจะได้ข้อมูลสารสนเทศที่แสดงพัฒนาการความก้าวหน้า และความสำเร็จทางการเรียนของผู้เรียนและเป็นประโยชน์ในการส่งเสริมการเรียนรู้อย่างเต็มศักยภาพเพื่อให้การจัดการศึกษามีคุณภาพที่แท้จริง การจัดระบบการศึกษาที่มีประสิทธิภาพเป็นสิ่งจำเป็น ดังนั้น ผู้ที่มีบทบาทสำคัญต่อประสิทธิภาพการศึกษา ได้แก่ ผู้บริหารโรงเรียน และครูผู้สอน ปัจจัยที่ส่งผลต่อประสิทธิผล

ทางวิชาการของนักเรียน ได้แก่ การจัดกระบวนการเรียนรู้ บรรยากาศทางวิชาการ ภาวะผู้นำทางวิชาการ และพฤติกรรมการสอนของครู (นวลแสง สูคีรี, 2560) นักวิชาการทางการศึกษาหลายท่าน มุ่งให้ความสนใจในเรื่องการคิดอย่างมีวิจารณญาณ (Critical thinking) ซึ่งเชื่อว่าทักษะการคิดเป็นพื้นฐานที่จำเป็นในการเปลี่ยนแปลงทางด้านสติปัญญาของคน การนำเทคนิคการสอนแบบ PMI (The PMI teaching technique) มาใช้ในการสอนนักเรียนระดับมัธยมศึกษาปีที่ 3 วิชาภาษาไทย พบว่า นักเรียนมีคะแนนความสามารถในการคิดอย่างมีวิจารณญาณหลังใช้เทคนิค PMI สูงกว่าคะแนนก่อนเรียน (สุวภัทร์ ทัพซ้าย, 2557, หน้า 63) ส่วนปัจจัยที่สัมพันธ์กับพฤติกรรมการบริหารงานของผู้บริหารโรงเรียนที่มีอิทธิพลต่อผลสัมฤทธิ์ทางการเรียนของนักเรียน สังกัดสำนักงานคณะกรรมการขั้นพื้นฐาน คือ ภาวะความเป็นผู้นำทางวิชาการ บรรยากาศโรงเรียนที่เน้นด้านวิชาการ การเสริมพลังอำนาจให้กับครู การรับรู้ความสามารถของครูในภาพรวม โดยปัจจัยด้านภาวะผู้นำทางวิชาการมีอิทธิพลโดยรวมต่อผลสัมฤทธิ์ทางการเรียนของนักเรียนมากที่สุด (ทศพร จันทนราช, 2556, หน้า 13)

การวัดผลทางการศึกษามีความต่อเนื่องมาอย่างยาวนานเนื่องจากแบบทดสอบไม่สามารถวัดความรู้ ความสามารถซึ่งเป็นลักษณะแฝงของบุคคลได้โดยตรง จึงเริ่มมีการคิดค้นทฤษฎีเกี่ยวกับการวัดขึ้นมา โดยในระยะแรก เรียกว่า ทฤษฎีการทดสอบแบบดั้งเดิม (Classical test theory: CTT) ซึ่งมีการวิเคราะห์คุณภาพข้อสอบ ได้แก่ ค่าความยาก (Item difficulty) ค่าอำนาจจำแนกของข้อสอบ (Item discrimination) และประสิทธิภาพตัวลวง (Item distractor) โดยมีข้อตกลงเบื้องต้นที่เป็นจุดอ่อน (Weak assumption) ซึ่งนักวัดผลเห็นว่า เป็นข้อตกลงเบื้องต้นที่ไม่สมเหตุสมผลและไม่สอดคล้องกับสภาพของการทดสอบในปัจจุบัน ไม่ว่าจะเป็นการกำหนดให้ค่าความคลาดเคลื่อนมีค่าเท่ากันในทุกผู้สอบ หรือข้อตกลงเกี่ยวกับความเป็นคู่ขนานที่เป็นไปได้ยากในทางปฏิบัติ ต่อมาจึงเกิดทฤษฎีการวัดแนวใหม่ที่เป็นที่ยอมรับกันอย่างกว้างขวางในปัจจุบัน คือ ทฤษฎีการตอบสนองข้อสอบ (Item response theory: IRT) เป็นทฤษฎีที่มีการอธิบายความสัมพันธ์ระหว่างความน่าจะเป็นในการตอบข้อสอบถูกต้องตรง กับ ลักษณะหรือความสามารถที่แท้จริงของผู้สอบด้วยการอธิบายโค้งคุณลักษณะข้อสอบ (Item characteristic curve: ICC) ซึ่งมีคุณลักษณะเป็นฟังก์ชั่น (Logistic function) ทำให้ค่าพารามิเตอร์ของข้อสอบที่ได้จากการวิเคราะห์คุณภาพของข้อสอบ ได้แก่ ค่าความยาก (Difficulty item parameter), ค่าอำนาจจำแนกของข้อสอบ (Discrimination item parameter) และค่าการเดา (Guessing parameter หรือ Pseudo-chance score level) แต่ละข้อมีลักษณะที่คงที่ ไม่แปรเปลี่ยนไปตามลักษณะกลุ่มผู้สอบ และพารามิเตอร์ของผู้สอบไม่เปลี่ยนแปลงไปตามคุณลักษณะของข้อสอบ ซึ่งเป็นจุดแข็งของทฤษฎีการตอบสนองข้อสอบ ข้อตกลงเบื้องต้นที่สำคัญ คือ ความเป็นอิสระระหว่างข้อสอบ และผู้สอบ ทำให้ผู้ตอบข้อสอบมีความเป็นอิสระในการตอบข้อคำถามแต่ละข้อ เพราะหากไม่มีความอิสระในการตอบข้อคำถาม

จะเกิดความคลาดเคลื่อนมาตรฐานของการวัด และค่าความเที่ยง (Reliability) จะได้ค่าพารามิเตอร์ที่ไม่ถูกต้อง

การสร้างแบบทดสอบเพื่อนำมาทดสอบสิ่งที่ต้องให้ความคำนึงถึง คือ ความยุติธรรมกับกลุ่มผู้สอบทุกกลุ่มที่มีระดับความสามารถเดียวกัน ถ้าหากข้อสอบทำให้ผู้ตอบกลุ่มใดกลุ่มหนึ่งได้ประโยชน์ เรียกว่า ข้อสอบนั้นมีความลำเอียง (Bias) เกิดขึ้น และเกิดความไม่ยุติธรรมต่อผู้ทดสอบเกิดขึ้น เรียกกว่า “ความลำเอียงของข้อสอบ” Zumbo (2007) ได้ศึกษาถึงพัฒนาการของวิธีการตรวจสอบความลำเอียงของข้อสอบโดยในสมัยแรก ๆ ใช้คำว่า ความลำเอียงของข้อสอบ โดยให้ความสำคัญของผลสัมฤทธิ์ทางการเรียนที่แตกต่างกันในกลุ่มเพศ และศาสนา เช่น เพศชาย และเพศหญิง ผิวดำและผิวขาว ซึ่งการตรวจสอบความลำเอียงของข้อสอบ โดยการเปรียบเทียบ เช่น (ก) การเปรียบเทียบเพียงสองกลุ่มของผู้สอบ (ข) คำศัพท์ที่นำมาใช้ ในการเปรียบเทียบกลุ่มผู้สอบสองกลุ่ม โดยกลุ่มที่สนใจ (Focal group) เป็นกลุ่มที่ต้องการศึกษา และกลุ่มอ้างอิง (Reference group) เป็นกลุ่มที่ใช้เปรียบเทียบ และ (ค) คะแนนข้อสอบเป็นการจับคู่ และต่อมาคำว่าข้อสอบมีความลำเอียงถูกแทนที่ด้วยคำศัพท์ว่า “การทำหน้าที่ต่างกันของข้อสอบ (DIF)” และได้มีการสร้างเงื่อนไขใหม่ และกรอบแนวคิดสำหรับการตรวจสอบ DIF ด้วยข้อมูลเชิงประจักษ์ ส่วนใหญ่มีการวิจัยที่มุ่งเน้นไปที่การพัฒนาวิธีการทางสถิติที่ซับซ้อนสำหรับการตรวจจับ DIF และให้คำอธิบายว่าเหตุใด DIF จึงเกิดขึ้นแนวคิด ดังนี้ 1) โมเดลการตอบสนองข้อสอบ, ตารางการณ์จร และ/ หรือโมเดลการถดถอย (Regression) 2) ทฤษฎีการตอบสนองข้อสอบ (IRT) และ 3) โมเดลพหุมิติ

การทำหน้าที่ต่างกันของข้อสอบ (DIF) กับความลำเอียงของข้อสอบ (Item bias) มีแนวคิดที่แตกต่างกัน สำหรับการทำหน้าที่ต่างกันของข้อสอบ เป็นกระบวนการที่เน้นการใช้วิธีการทางสถิติสำหรับตรวจสอบ เพื่อให้ได้สารสนเทศเกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบสำหรับกลุ่มผู้สอบกลุ่มย่อยที่มีลักษณะเฉพาะบางอย่างแตกต่างกัน ส่วนความลำเอียงของข้อสอบเป็นกระบวนการการตัดสินความยุติธรรมของข้อสอบ โดยนำสารสนเทศการทำหน้าที่ต่างกันของข้อสอบมาวิเคราะห์เชิงตรรกะ (Logical analysis) โดยผู้เชี่ยวชาญพิจารณาถึงการเขียนข้อสอบ เนื้อหา สาระของข้อสอบ และจุดมุ่งหมายของการวัด เพื่อระบุว่าข้อสอบนั้นลำเอียงเข้าข้างกลุ่มใดหรือไม่ เพราะเหตุใดจึงเป็นการตัดสินความลำเอียงของข้อสอบ

โมเดลตอบสนองข้อสอบ (IRT Models) หรือทฤษฎีคุณลักษณะแฝง (Latent trait theory) ได้รับการพัฒนามาอย่างต่อเนื่อง บิเนต และไซมอน (Binet & Simon, 1916) ได้แสดงความสัมพันธ์ระหว่างพารามิเตอร์ของแบบจำลองทฤษฎีการตอบสนองข้อสอบและทฤษฎีแบบดั้งเดิม ต่อมาในปี ค.ศ. 1943-1944 ลอร์เลย์ (Hambleton & Swaminathan, 1985) ได้เสนอวิธีการประมาณ

ค่าพารามิเตอร์แนวใหม่ ซึ่งมีหลักการว่าผลการสอบของผู้สอบจากแบบทดสอบใด ๆ ขึ้นอยู่กับความสามารถ (Ability or skill) ของผู้สอบ จนกระทั่งปี ค.ศ. 1952 ลอร์ด (Lord) ได้เสนอทฤษฎีใหม่ในรูปโค้งลักษณะข้อสอบ (Item characteristic curve: ICC) โดยลอร์ดได้เสนอว่าโค้งลักษณะข้อสอบเป็นโค้งปกติสะสม ต่อมา เรียกว่า แบบจำลอง Normal ogive ซึ่งแบบจำลองนี้มีความยากในการคำนวณ และขาดโปรแกรมคอมพิวเตอร์ที่มีประสิทธิภาพในการใช้วิเคราะห์ข้อมูลเพื่อพิสูจน์ทฤษฎี จึงทำให้ลอร์ด ชะงักการทำงานเกี่ยวกับทฤษฎีนี้ไประยะหนึ่ง

ทฤษฎีการตอบสนองข้อสอบ (IRT) เป็นทฤษฎีการวัดที่อธิบายความสัมพันธ์ระหว่างความสามารถที่มีอยู่ในตัวบุคคล (Latent trait or ability) กับ ผลการตอบข้อสอบ หรือข้อคำถาม โดยใช้โค้งลักษณะข้อสอบ (Item characteristic curve: ICC) ซึ่งมีการกำหนดลักษณะข้อสอบด้วยพารามิเตอร์อำนาจจำแนก (*a*) ความยาก (*b*) และโอกาสในการเดาข้อสอบ (*c*) IRT จึงอยู่บนพื้นฐานแนวคิดที่สำคัญ 2 ประการ คือ 1) ผลการตอบข้อสอบหรือข้อคำถามของผู้ตอบสามารถอธิบายได้ด้วยความสามารถที่มีอยู่ภายในของผู้ตอบ และ 2) ความสัมพันธ์ระหว่างผลการตอบข้อสอบกับความสามารถที่มีอยู่ภายใน สามารถอธิบายได้ด้วยฟังก์ชันลักษณะข้อสอบ หรือโค้งลักษณะข้อสอบ (ICC) อันมีลักษณะเป็นฟังก์ชันทางคณิตศาสตร์ เรียกว่า ฟังก์ชันโลจิสติก (Logistic function) ซึ่งมีลักษณะใกล้เคียงกับฟังก์ชันปกติสะสม (Normal ogive function) โดยพื้นฐานของทฤษฎีการตอบสนองข้อสอบนั้น มุ่งอธิบายผลการตอบข้อสอบว่าผู้สอบจะมีโอกาสในการตอบได้ถูกต้องเพียงใด

การตรวจสอบความลำเอียงของข้อสอบตามแนวคิดทฤษฎีการตอบสนองข้อสอบมีจุดเด่นกว่าวิธีตามแนวคิดทฤษฎีการตอบสนองข้อสอบแบบดั้งเดิม คือ สามารถประมาณค่าพารามิเตอร์ของข้อสอบ เช่น ค่าความยาก (b) ค่าอำนาจจำแนก (a) และค่าการเดา (c) สามารถแยกค่าพารามิเตอร์ข้อสอบได้ มีคุณสมบัติทางสถิติของข้อสอบ สามารถอธิบายได้ละเอียดและชัดเจนเมื่อข้อสอบมีความลำอียงระหว่างกลุ่ม และคุณสมบัติของข้อสอบสามารถแสดงด้วยแผนภาพทำให้เข้าใจง่ายและนำไปใช้ได้อย่างกว้างขวาง

Angoff (1993) ได้ทำการศึกษาเกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบมาอย่างยาวนาน แต่ก็ไม่มีคำอธิบายว่า ทำไมข้อสอบที่มีความสมบูรณ์เหมาะสมจึงเกิดการทำหน้าที่ต่างกันของข้อสอบขนาดใหญ่มาก Muthen และเพื่อนร่วมงานของเขา ได้ตั้งคำถามถึงแหล่งที่มาของการเกิดการทำหน้าที่ต่างกันของข้อสอบ (Muthen, 1985, 1988, 1989; Muthen, Kao & Burstein, 1991; Muthen & Lehman, 1985) และคิดวิธีการตรวจสอบ DIF โดยการใช้โครงสร้างโมเดลสมการโครงสร้างหลายตัวชี้วัดหลายสาเหตุ หนึ่งในกรอบการทำงานของ Muthen คือ ความกลมกลืนของ “การสร้างโมเดลตอบสนองข้อสอบผ่านตารางการณ์จร และ/ หรือโมเดลการถดถอย” และ

“โมเดลทฤษฎีการตอบสนองข้อสอบ” ซึ่งกรอบการทำงานอธิบายเพียงความเป็นไปได้ของความเป็นพหุมิติ (Multidimensionality) วิธีการของ Muthen มีความชัดเจน และง่ายต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

ต่อมาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมีความชัดเจน และโดดเด่นอย่างมากการศึกษามุ่งเน้นไปที่การทำหน้าที่ต่างกันของข้อสอบระหว่างเพศ ซึ่งจะทำการตรวจสอบลักษณะของข้อสอบ และเนื้อหาของข้อสอบ ที่อาจมีผลต่อผลสัมฤทธิ์ทางการเรียนของนักเรียน รวมทั้งตัวแปรตามบริบท เช่น ขนาดของห้องเรียน, สถานะทางเศรษฐกิจสังคม, การปฏิบัติการเรียนการสอน และรูปแบบของผู้ปกครอง ในปี ค.ศ. 1951 มีงานวิจัยที่ถือว่าเป็นการเริ่มต้นของการศึกษาเกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบ ได้แก่ งานวิจัยของเอลส์ (Eells) เดวิส (Davis) ฮาวิคเฮิร์ต (Havighurst) เฮอร์ริค (Herrick) และไทเลอร์ (Tyler) โดยเอลส์ และคณะ มีความสนใจในเรื่องความไม่ถูกต้องของการตอบสนองข้อสอบอันเกิดจากความสามารถ (Ability) เช่น ความแตกต่างของไอคิวเด็กนักเรียน ในแบบทดสอบไอคิว โดยมุ่งไปที่ข้อคำถามแต่ละข้อ เพื่อดูว่ามีข้อคำถามใดทำหน้าที่ต่างกันหรือไม่ เช่น มีข้อคำถามที่แสดงถึงความแตกต่างของวัฒนธรรม ซึ่งอาจมีความสัมพันธ์กับความยากของข้อสอบ และทำให้มีการเลือกคำตอบที่ผิด และเขาได้พยายามแก้ไขอิทธิพลทางสังคม และภูมิหลังของเชื้อชาติที่มีผลต่อการตอบข้อสอบ โดยการเปรียบเทียบกลุ่มของเชื้อชาติที่อยู่ในชั้นทางสังคม ผลการศึกษาของเอลส์ ได้นำไปสู่การศึกษาอย่างกว้างขวางในช่วง 20 ปีที่ผ่านมา

จากแนวคิดการตอบสนองข้อสอบได้มีการพัฒนาโมเดลหรือแบบจำลองขึ้นมาหลายรูปแบบด้วยกัน โดยแต่ละโมเดลจะมีฟังก์ชันทางคณิตศาสตร์และจำนวนพารามิเตอร์ในฟังก์ชันที่แตกต่างกัน ซึ่งสามารถจำแนกเป็น 3 ประเภทใหญ่ ๆ (Hambleton & Swiminathan, 1985) ดังนี้

1. โมเดลการตอบสนองข้อสอบที่ใช้กับคะแนนรายข้อแบบ 2 ค่า (Dichotomous) เป็นข้อสอบที่ตรวจให้คะแนน 0, 1 (ตอบผิดได้ 0 คะแนน, ตอบถูกได้ 1 คะแนน) โมเดลประเภทนี้ในระยะเริ่มแรก (ค.ศ. 1943-1968) เช่น Guttman perfect scale, Latent distance model, Linear model เป็นต้น ในระยะต่อมา (ค.ศ. 1952-1982) ได้มีการพัฒนาโมเดลประเภทนี้เพิ่มขึ้นมา เช่น One parameter normal ogive model, Two parameter normal ogive model, Three parameter normal ogive model, Four parameter normal ogive model เป็นต้น

2. โมเดลการตอบสนองที่ใช้กับคะแนนรายข้อแบบข้อแบบมากกว่า 2 ค่า (Multichotomous) เช่น Norminal response model, Grade response model, Partical credit model (Semejima, 1972) เป็นต้น

3. โมเดลการตอบสนองข้อสอบที่ใช้กับคะแนนรายข้อแบบต่อเนื่อง (Continuous) เช่น Continuous response model (Semejim, 1972) เป็นต้น

ซึ่งต่อมาได้มีการจำแนกวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ ดังนี้ (ศิริชัย กาญจนวาสี, 2545, หน้า 112)

1. วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนแบบทวิภาค

1.1 กลุ่มวิธีที่ใช้คะแนนสังเกตได้

วิธีในกลุ่มนี้มักวิเคราะห์ตามทฤษฎีการทดสอบแบบดั้งเดิม (CTT) หรือกลุ่มที่ไม่ใช้ทฤษฎีการตอบสนองข้อสอบ (Non-IRT approach) โดยใช้คะแนนรวมของผู้สอบเป็นเกณฑ์การจับคู่ของกลุ่มผู้สอบ วิธีการตรวจสอบที่สำคัญในกลุ่มนี้ ได้แก่

1.1.2 วิธีการวิเคราะห์การถดถอยโลจิสติก (Logistic regression: LR)

1.1.3 วิธีการแปลงค่ายากข้อสอบ (Transformed item difficulty: TID)

1.1.4 วิธีแมนเทล-แฮนส์เซล (Mantel-haenszel: MH)

1.1.5 วิธีดัชนีมาตรฐาน (Standardization: STND) การปรับปรุงให้เป็นมาตรฐานด้วยน้ำหนักองค์ประกอบ

1.2 กลุ่มวิธีที่ใช้คุณลักษณะแฝง

วิธีในกลุ่มนี้ใช้คุณลักษณะหรือตัวแปรแฝง ซึ่งวิเคราะห์บนพื้นฐานของทฤษฎีการตอบสนองข้อสอบ (IRT) สำหรับใช้เกณฑ์การจับคู่กลุ่มผู้สอบ วิธีการตรวจสอบที่สำคัญ ได้แก่

1.2.1 วิธีวัดพื้นที่ความแตกต่างระหว่างโค้งการตอบสนองข้อสอบ (IRT-D^2)

1.2.2 วิธีไคสแควส์ของลอร์ด (Lord's χ^2)

1.2.3 วิธีอัตราส่วนไลค์ลิฮู้ดทั่วไป (General IRT likelihood ratio)

1.2.4 วิธีอัตราส่วนไลค์ลิฮู้ด ลอกลิเนียร์ (Loglinear IRT likelihood ratio)

1.2.5 วิธีซิปเทสท์ (SIBTEST)

2. วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนแบบพหุวิภาค

2.1 กลุ่มวิธีที่ใช้คะแนนที่สังเกตได้

2.1.1 วิธีการวิเคราะห์การถดถอยโลจิสติกพหุวิภาค (Polytomous logistic regression)

2.1.2 วิธีดัชนีมาตรฐานพหุวิภาค (Polytomous standardization)

2.1.3 วิธีแมนเทล-แฮนส์เซลทั่วไป (General mantel-haenszel: GMH)

2.2 กลุ่มวิธีที่ใช้คุณลักษณะแฝง

2.2.1 วิธีอัตราส่วนไลค์ลิฮู้ดในรูปทั่วไป (General IRT likelihood ratio)

2.2.2 วิธีการให้คะแนนบางส่วน (Partial credit model: PCM)

2.2.3 วิธีซิปเทสท์พหุวิภาค (Polytomous SIBTEST)

2.2.4 วิธีการให้คะแนนบางส่วนทั่วไป (Generalized partial credit model: GPCM)

จะเห็นว่ามีวิธีการที่หลากหลายสำหรับตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีข้อดี และข้อจำกัดที่แตกต่างกันออกไป ผู้วิจัยมีความสนใจที่จะทำการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีอัตราส่วนความควรจะเป็น (Likelihood ratio test) วิธีเบส์เซียน (Bayesian) และวิธี Multiple group CFA และเปรียบเทียบวิธีที่เหมาะสมในการประมาณค่าพารามิเตอร์ข้อสอบ (ค่าพารามิเตอร์อำนาจจำแนก และค่าพารามิเตอร์ความยาก) และค่าพารามิเตอร์ข้อสอบ รวมทั้งต้องการเปรียบเทียบวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบว่า วิธีใดสามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ดี และมีประสิทธิภาพ และได้สรุปวิธีการตรวจสอบได้ดังต่อไปนี้

การตรวจสอบความเท่าเทียมกันของการวัด (Detection of measurement equivalence) หรืออาจกล่าวอีกนัยหนึ่ง คือ ความไม่แปรเปลี่ยนของโมเดลการวัด มีความใกล้เคียงกับแนวคิดของความลำเอียงของข้อสอบ (Item bias) และการทำหน้าที่ต่างกันของข้อสอบ (DIF) ที่มาจากกรอบทฤษฎีการตอบสนองข้อสอบ (DIF) การศึกษาเรื่องความเท่าเทียมกันของการวัด การทำหน้าที่ต่างกันของข้อสอบที่มีความสนใจกันในปัจจุบันนี้ คือ การวิเคราะห์องค์ประกอบเชิงยืนยันพหุกลุ่ม (Multiple group confirmatory factor analysis: MGCFA)

วิธีการ MGCFA Jöreskog & Sörbom (1993) ได้สร้างรากฐานของการใช้ MGCFA เพื่อศึกษาความไม่แปรเปลี่ยนของโมเดลการวัด (MI) และตรวจสอบความแตกต่างของค่าเฉลี่ยของตัวแปรแฝงในประชากรต่างกลุ่ม ขยายมาจากพื้นฐานของโมเดลองค์ประกอบเชิงยืนยัน, MGCFA เป็นวิธีการที่อยู่บนพื้นฐานของทฤษฎีการทดสอบความไม่แปรเปลี่ยนของโมเดลการวัด (Bollen, 1989; Byrne, 1998; Vandenberg & Lance, 2000) ซึ่งได้รับความสนใจจากนักวิจัย และมีการใช้กันอย่างแพร่หลายในการวิจัยความไม่แปรเปลี่ยนของโมเดลการวัด (French & Finch, 2008) ยกตัวอย่าง เช่น Muthén and Lehman (1985) และ Glockner-Rist and Hoijtnik (2003) กล่าวถึงการใช้งานของ MGCFA หลักการของ MGCFA คือ การทดสอบแบบลำดับขั้นเพื่อตรวจสอบความไม่แปรเปลี่ยนของโมเดลการวัดระหว่างกลุ่มประชากร

หลักการวิเคราะห์โครงสร้างค่าเฉลี่ยของตัวแปรแฝง คือ การเปรียบเทียบค่าเฉลี่ยตัวแปรแฝงระหว่างกลุ่มประชากรด้วยการพิจารณาจุดตัด (Threshold) ของตัวชี้วัดที่สังเกตได้ การวิเคราะห์ค่าเฉลี่ยโครงสร้างตัวแปรแฝง เพื่อตรวจสอบความแตกต่างของค่าเฉลี่ยตัวแปรแฝงนั้น ว่ามีการกระจายความสามารถแฝงของกลุ่มที่แตกต่างกันหรือไม่ การวัดความไม่แปรเปลี่ยนของโมเดล

จะต้องประเมินความสอดคล้องของโมเดลกับข้อมูลของโมเดลสอดแทรก (Nested models) เริ่มด้วยการไม่มีการกำหนดข้อบังคับกับโมเดลพื้นฐาน แล้วค่อย ๆ เพิ่มข้อบังคับของพารามิเตอร์ในโมเดล (เช่น ค่าน้ำหนักองค์ประกอบ (Factor loading) จุดตัด (Threshold) เป็นต้น จุดเด่นของวิธีการนี้ คือ การวิเคราะห์ที่สามารถกำหนดข้อบังคับของโมเดล และมีการผ่อนคลายข้อบังคับ เพื่อการตรวจสอบว่าโมเดลทั้งสองกลุ่มนั้นมีความไม่แปรเปลี่ยนของโมเดลการวัดเกิดขึ้นหรือไม่ และสามารถทำการวิเคราะห์แยกกลุ่ม เพื่อหลีกเลี่ยงปัญหา และข้อจำกัดที่อาจเกิดจากการวิเคราะห์รวมกลุ่มได้

วิธีทดสอบอัตราส่วนความควรจะเป็น (Likelihood ratio test: LRT) สำหรับใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยการทดสอบความแตกต่างของผลการตอบข้อสอบระหว่างผู้สอบกลุ่มย่อย 2 กลุ่ม เช่น เพศ ภูมิลำเนา เชื้อชาติ ศาสนา เป็นต้น วิธีทดสอบอัตราส่วนความควรจะเป็นจะใช้วิธีการตรวจสอบอัตราส่วนความควรจะเป็นสูงสุด (Maximum likelihood; ML) ในทฤษฎีการตอบสนองข้อสอบในรูปเข้าสู่ระบบเชิงเส้น (Loglinear; LR) เป็นการประมาณค่าพารามิเตอร์ด้วยวิธีความน่าจะเป็นสูงสุด วิธีดังกล่าวใช้ทดสอบอัตราส่วนความควรจะเป็นเพื่อทดสอบนัยสำคัญของการทำหน้าที่ต่างกันของข้อสอบ (Bolt, 2002)

ข้อดีของวิธีการทดสอบอัตราส่วนความควรจะเป็น คือ สามารถวัดความคลาดเคลื่อนของโมเดล จำแนกพารามิเตอร์ของข้อสอบในการประมาณค่าแต่ละกลุ่มผู้สอบ และขนาดของอิทธิพลพารามิเตอร์ของข้อสอบ เข้าใจได้ง่าย ข้อเสียของวิธีการนี้ คือ อยู่บนพื้นฐานของข้อตกลงเบื้องต้นมากมาย เช่น ฟังก์ชั่นการตอบสนองข้อสอบต้องเป็นสมการโลจิสติก และการแจกแจงของคุณลักษณะแฝง และคุณลักษณะหลักของผู้สอบ (θ) ต้องเป็นการแจกแจงปกติ ดังนั้นก่อนการทดสอบด้วยวิธีทดสอบอัตราส่วนความควรจะเป็น จะต้องมีการตรวจสอบให้เป็นไปตามข้อตกลงเบื้องต้นอย่างชัดเจน ซึ่งเมื่อเกิดการละเมิดข้อตกลงเบื้องต้นของวิธีการทดสอบอัตราส่วนความควรจะเป็น จึงมีทางเลือกอื่น ๆ ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่สามารถตรวจสอบกับข้อมูลที่มีการแจกแจงอื่น ๆ เช่น การประมาณค่าพารามิเตอร์ด้วยวิธีเบส์เซียน

วิธีเบส์เซียน (Bayesian method) มีข้อดีที่นำวิธีการของเบส์มาใช้สามารถทำการประมาณค่าพารามิเตอร์ข้อสอบและค่าความสามารถของผู้เข้าสอบพร้อม ๆ กัน วิธีของเบส์จึงอาจเป็นวิธีที่เหมาะสมกว่าในการประมาณค่าพารามิเตอร์ (วิชุดา บัวคง, 2533, หน้า 47) ค่าพารามิเตอร์ของแบบทดสอบที่ประมาณค่าด้วยวิธีแมกซิมัมไลค์ลิฮู้ด วิธีฮิวริสติก และวิธีของเบส์ พบว่าค่าความยากเฉลี่ยของแบบทดสอบและค่าส่วนเบี่ยงเบนมาตรฐานของค่าความยากของแบบทดสอบที่ประมาณค่าด้วยวิธีของเบส์ มีค่าสูงกว่าค่าความยากเฉลี่ยของแบบทดสอบและค่าเบี่ยงเบนมาตรฐานของ

ค่าความยากของแบบทดสอบที่ประมาณค่าด้วยวิธีฮิวริสติก และวิธีแมกซิมัมไลค์ลิฮู้ด (รัตนา ศรีหิรัญ, 2539) ซึ่งแนวคิดของวิธีเบส์มีแนวคิดบางประการต่างออกไปจากแนวคิดวิธีการอื่น ๆ คือ สามารถประมาณค่า θ และ a, b, c ร่วมกันได้อย่างมีประสิทธิภาพ (Swaminathan & Gifford, 1985) นอกจากนี้แล้วยังสามารถประมาณค่าพารามิเตอร์ได้ทุกรูปแบบการแจกแจง เหมาะสำหรับการวิเคราะห์กลุ่มตัวอย่างที่มีขนาดเล็ก และจำนวนพารามิเตอร์ที่มีจำนวนมาก มีงานวิจัยน้อยมากที่ศึกษาวิธีการเบส์เซียนในการประมาณค่าพารามิเตอร์ที่ใช้ฟังก์ชัน Cumulative logit (Lahiri & Gao, 2002) โดยวิธีการวิเคราะห์จะมีการกำหนดการแจกแจงก่อนหน้า (Posterior distribution) โดยผู้วิจัยจะอาศัยจากความรู้ ความเชื่อ และประสบการณ์ของผู้วิจัยเองที่มีต่อพารามิเตอร์

ทั้งนี้ผู้วิจัยสามารถเลือกการแจกแจงก่อนหน้าซึ่งเป็นคอนจูเกต (Congugate) กับการแจกแจงของค่าสารสนเทศซึ่งการแจกแจงคอนจูเกตจะทำให้การคำนวณเพื่อหาฟังก์ชันการแจกแจงภายหลังทำได้ง่ายขึ้น และจากการแจกแจงภายหลัง (Posterior distribution) ที่ได้จะถูกนำไปวิเคราะห์เพื่อหาข้อสารสนเทศที่เกี่ยวข้องกับการแจกแจงภายหลังต่อไป เช่น การหาค่าความคาดหวัง ค่าความแปรปรวน เป็นต้น ข้อเสียหรือส่วนที่ยากที่สุดของการประมาณค่าพารามิเตอร์ของข้อสอบด้วยวิธีเบส์ คือ ความซับซ้อนของการพิจารณาการแจกแจงภายหลัง (Posterior distribution) ที่ต้องอาศัยข้อมูลที่ได้จากตัวอย่างที่อยู่ในรูปแบบของฟังก์ชันความน่าจะเป็น (Likekihood function) และการแจกแจงก่อนหน้า (Prior distribution) และการประมาณค่าพารามิเตอร์จะต้องทำการประมาณค่าซ้ำ (Iterative) จนกว่าค่าประมาณจะลู่เข้าสู่ค่าคงที่ ซึ่งต้องทำซ้ำหลาย ๆ รอบ และใช้เวลานานในการวิเคราะห์ข้อมูล

จากเหตุผลที่กล่าวมาแต่ต้นข้างต้น ทำให้ผู้วิจัยสนใจที่จะใช้วิธีทั้ง 3 วิธีดังกล่าว เพื่อทำการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยใช้แบบทดสอบวัดผลสัมฤทธิ์ระดับชาติ (O-NET: Ordinary national educational test) ซึ่งเป็นการทดสอบเพื่อวัดผลทางการศึกษาระดับประเทศในระดับชั้นมัธยมศึกษาปีที่ 6 ประจำปีการศึกษา 2558 เนื่องจากคะแนน O-NET มีความสำคัญต่อการนำไปประกอบการสอบเข้ามหาวิทยาลัย และมีผลต่อการประเมินคุณภาพของการจัดการศึกษาของแต่ละโรงเรียน รวมทั้ง พบว่า ยังไม่มีงานวิจัยใดที่ใช้คะแนน O-NET ระดับชั้นมัธยมศึกษาปีที่ 6 มาทำการศึกษา งานวิจัยที่ผ่านมา พบว่า มีเพียงการทำวิจัยการทำหน้าที่ต่างกันในข้อสอบ Pre-O-NET เท่านั้น และในการวิจัยครั้งนี้ ผู้วิจัยมีความสนใจที่จะทำการศึกษาการทำหน้าที่ต่างกันของข้อสอบ (DIF) จำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน ได้แก่ เขตกรุงเทพมหานคร และปริมณฑล และนอกเขตกรุงเทพมหานคร และปริมณฑล เพื่อตรวจสอบดูว่าข้อสอบที่ใช้ทดสอบนั้นมีความลำเอียง หรือทำหน้าที่ต่างกันของข้อสอบเข้าข้างกลุ่มใดกลุ่มหนึ่งหรือไม่ ทั้งนี้ผู้วิจัยหวังเป็นอย่างยิ่งว่าข้อค้นพบที่ได้จากงานวิจัย จะเป็นประโยชน์แก่ผู้เกี่ยวข้อง

ทุกภาคส่วน สามารถนำไปปรับปรุง และพัฒนาการจัดทำแบบทดสอบ การจัดระบบการศึกษา เพื่อยกระดับคุณภาพผู้เรียนต่อไปในอนาคต

คำถามการวิจัย

1. ผลการประมาณค่าพารามิเตอร์อำนาจจำแนก (a) ค่าพารามิเตอร์ความยาก (b) ของข้อสอบ และค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) ด้วยวิธี Maximum likelihood, วิธี BAYESIAN และวิธี Confirmatory factor analysis มีความสอดคล้องกันหรือไม่

2. การทำหน้าที่ต่างกันของข้อสอบ (DIF) จำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียนได้แก่ เขตกรุงเทพมหานคร และปริมณฑล และนอกเขตกรุงเทพมหานคร และปริมณฑล ด้วยวิธี IRT-likelihood ratio, วิธี BAYESIAN และวิธี Multiple group confirmatory factor analysis มีความสอดคล้องกันหรือไม่

วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาผลการประมาณค่าพารามิเตอร์อำนาจจำแนก (a) ค่าพารามิเตอร์ความยาก (b) ของข้อสอบ ค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) ด้วยวิธี Maximum likelihood วิธี BAYESIAN และวิธี Confirmatory factor analysis

2. เพื่อเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ (DIF) จำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียนได้แก่ เขตกรุงเทพมหานคร และปริมณฑล และนอกเขตกรุงเทพมหานคร และปริมณฑล ด้วยวิธี IRT-likelihood ratio, วิธี BAYESIAN และวิธี Multiple group confirmatory factor analysis

สมมติฐานในการวิจัย

วิธีการประมาณค่าพารามิเตอร์อำนาจจำแนก (a) ค่าพารามิเตอร์ความยาก (b) ของข้อสอบ และค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) ด้วยวิธี Maximum likelihood วิธี BAYESIAN และวิธี Confirmatory factor analysis ทำให้ค่าพารามิเตอร์ที่ได้จากการประมาณค่าทั้งสามวิธี แตกต่างกัน และมีการทำหน้าที่ต่างกันของข้อสอบ (DIF) เป็นอย่างไร ดังนั้นผู้วิจัยจึงได้ตั้งสมมติฐานในการวิจัยไว้ ดังนี้

1. ค่าพารามิเตอร์อำนาจจำแนก (a) ค่าพารามิเตอร์ความยาก (b) ของข้อสอบ และค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) ด้วยวิธี Maximum likelihood วิธี BAYESIAN และ

วิธี Confirmatory factor analysis มีความสอดคล้องกัน

2. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ระหว่างวิธี IRT-likelihood ratio, วิธี BAYESIAN และวิธี Multiple group confirmatory factor analysis เมื่อจำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ ได้แก่ เขตกรุงเทพมหานคร และปริมณฑล และนอกเขตกรุงเทพมหานครและปริมณฑลของโรงเรียนต่างกัน มีความสอดคล้องกัน

ประโยชน์ที่คาดว่าจะได้รับ

1. ผลการวิจัยครั้งนี้ สามารถใช้เป็นแนวทางในการวางแผนปรับปรุงและพัฒนาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบให้มีความถูกต้อง และเกิดความยุติธรรมแก่ผู้เข้าสอบทุกกลุ่มที่มีความสามารถเท่ากัน

2. ผลการวิจัยจะเป็นประโยชน์ในเชิงวิชาการในด้านการประยุกต์ใช้วิธีการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ (DIF) ด้วยวิธี IRT-likelihood ratio, วิธี BAYESIAN และวิธี Multiple group confirmatory factor analysis ซึ่งจะเป็นแนวทางในการนำไปใช้วิเคราะห์ข้อสอบระดับชาติอื่น ๆ ต่อไป

3. ผลการวิจัยสามารถใช้เป็นแนวทางในการศึกษาเรื่องวิธีการประมาณค่าพารามิเตอร์อำนาจจำแนก (a), ค่าพารามิเตอร์ความยาก (b) ของข้อสอบ และค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) เพื่อตรวจสอบดูว่าวิธีการใดที่มีประสิทธิภาพมากกว่ากันระหว่างวิธี Maximum likelihood วิธี BAYESIAN และวิธี Confirmatory factor analysis เมื่อใช้แนวทางการวิเคราะห์ที่แตกต่างกัน

ขอบเขตการวิจัย

1. ประชากรที่ใช้ในการวิจัยครั้งนี้เป็นนักเรียนชั้นมัธยมศึกษาปีที่ 6 ที่อยู่ในโรงเรียนระดับมัธยมศึกษาขนาดใหญ่พิเศษ เฉพาะโรงเรียนสหศึกษา ที่เข้าสอบวัดผลทางการเรียนเพื่อประเมินคุณภาพการศึกษาระดับชาติ ปีการศึกษา 2558 ซึ่งเป็นนักเรียนที่เข้าสอบวิชาคณิตศาสตร์และภาษาไทย จำนวน 358,512 คน

2. การวิจัยครั้งนี้เป็นการศึกษาในบริบทของประเทศไทยโดยมุ่งศึกษาในกลุ่มตัวแปรเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน

3. การวิจัยครั้งนี้เป็นการศึกษาโมเดลตอบสนองข้อสอบ (Item response model) แบบ 2 พารามิเตอร์ ได้แก่ ค่าพารามิเตอร์อำนาจจำแนก (a) และค่าพารามิเตอร์ความยาก (b) ของข้อสอบ รวมทั้งประมาณค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) ด้วยวิธี Maximum likelihood

วิธี BAYESIAN และวิธี Confirmatory factor analysis และการทำหน้าที่ต่างกันของข้อสอบเพื่อเปรียบเทียบผลการวิเคราะห์ด้วยวิธีการที่แตกต่างกัน

นิยามศัพท์เฉพาะ

1. การทำหน้าที่ต่างกันของข้อสอบ (Differential item functioning: DIF) หมายถึง การที่ข้อสอบทำให้ผู้เข้าสอบซึ่งเป็นสมาชิกของกลุ่มอ้างอิง และผู้สอบที่เป็นสมาชิกของกลุ่มสนใจ ที่มีความสามารถในสิ่งที่ต้องการวัดเท่ากัน มีโอกาสตอบข้อสอบนั้นได้ถูกต้องไม่เท่ากัน ซึ่งในการวิจัยครั้งนี้ศึกษาการทำหน้าที่ต่างกันของข้อสอบและแบบสอบ สำหรับกลุ่มผู้สอบ จำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน

เพศ หมายถึง เพศของกลุ่มตัวอย่าง คือ เพศชาย และเพศหญิง

สถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน หมายถึง เขตพื้นที่ที่โรงเรียนตั้งอยู่ โดยจำแนกเป็น 2 กลุ่ม ได้แก่ เขตกรุงเทพมหานครและปริมณฑล และนอกเขตกรุงเทพมหานครและปริมณฑล

2. วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (Detecting DIF procedures) หมายถึง วิธีการวิเคราะห์ทางสถิติที่ใช้เพื่อบ่งบอกว่ามีการทำหน้าที่ต่างกันของข้อสอบระหว่างกลุ่มอ้างอิง และกลุ่มสนใจ ซึ่งวิธีการในงานวิจัยนี้เป็นการตรวจสอบที่ใช้กับข้อสอบที่มีการให้คะแนนแบบสองค่า

3. กลุ่มอ้างอิง (Reference group) หมายถึง กลุ่มผู้สอบที่คาดว่าจะได้ประโยชน์จากการตอบข้อสอบที่ทำหน้าที่ต่างกัน คือ เป็นกลุ่มที่มีความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องสูงกว่าผู้สอบอีกกลุ่มหนึ่งทั้ง ๆ ที่มีความสามารถเท่ากัน

ในการวิจัยครั้งนี้เมื่อจำแนกตามเพศ กลุ่มอ้างอิงในวิชาคณิตศาสตร์ คือ เพศชายและกลุ่มอ้างอิงในวิชาภาษาไทย คือ เพศหญิง จำแนกตามเขตที่ตั้งทางภูมิศาสตร์ กลุ่มอ้างอิงในวิชาคณิตศาสตร์ และวิชาภาษาไทย คือ โรงเรียนที่ตั้งอยู่ในเขตกรุงเทพมหานครและปริมณฑล

4. กลุ่มสนใจ (Focal group) หมายถึง กลุ่มผู้สอบที่คาดว่าเสียประโยชน์จากการตอบข้อสอบที่ทำหน้าที่ต่างกัน คือ เป็นกลุ่มที่มีความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องต่ำกว่าผู้สอบอีกกลุ่มหนึ่งทั้ง ๆ ที่มีความสามารถเท่ากัน

ในการวิจัยครั้งนี้เมื่อจำแนกตามเพศ กลุ่มสนใจวิชาคณิตศาสตร์ คือ เพศหญิงและกลุ่มสนใจวิชาภาษาไทย คือ เพศชาย จำแนกตามเขตที่ตั้งทางภูมิศาสตร์ กลุ่มสนใจในวิชาคณิตศาสตร์และวิชาภาษาไทย คือ โรงเรียนที่ตั้งอยู่ในเขตกรุงเทพมหานครและปริมณฑล

5. ตัวแปรต้น ได้แก่ วิธี Maximum likelihood วิธี BAYESIAN วิธี IRT-likelihood ratio วิธี CFA และวิธี Multiple group CFA

6. วิธี IRT-likelihood ratio (IRT-LR) หมายถึง วิธีการทดสอบอัตราส่วนความควรจะเป็นในทฤษฎีการตอบสนองข้อสอบในรูปของล็อกเชิงเส้น (Loglinear: LR) เป็นการประมาณค่าพารามิเตอร์ด้วยวิธีความเป็นไปได้สูงสุด (Maximum likelihood: ML) ใช้ทดสอบอัตราส่วนความควรจะเป็นเพื่อทดสอบนัยสำคัญของการทำหน้าที่ต่างกันของข้อสอบ

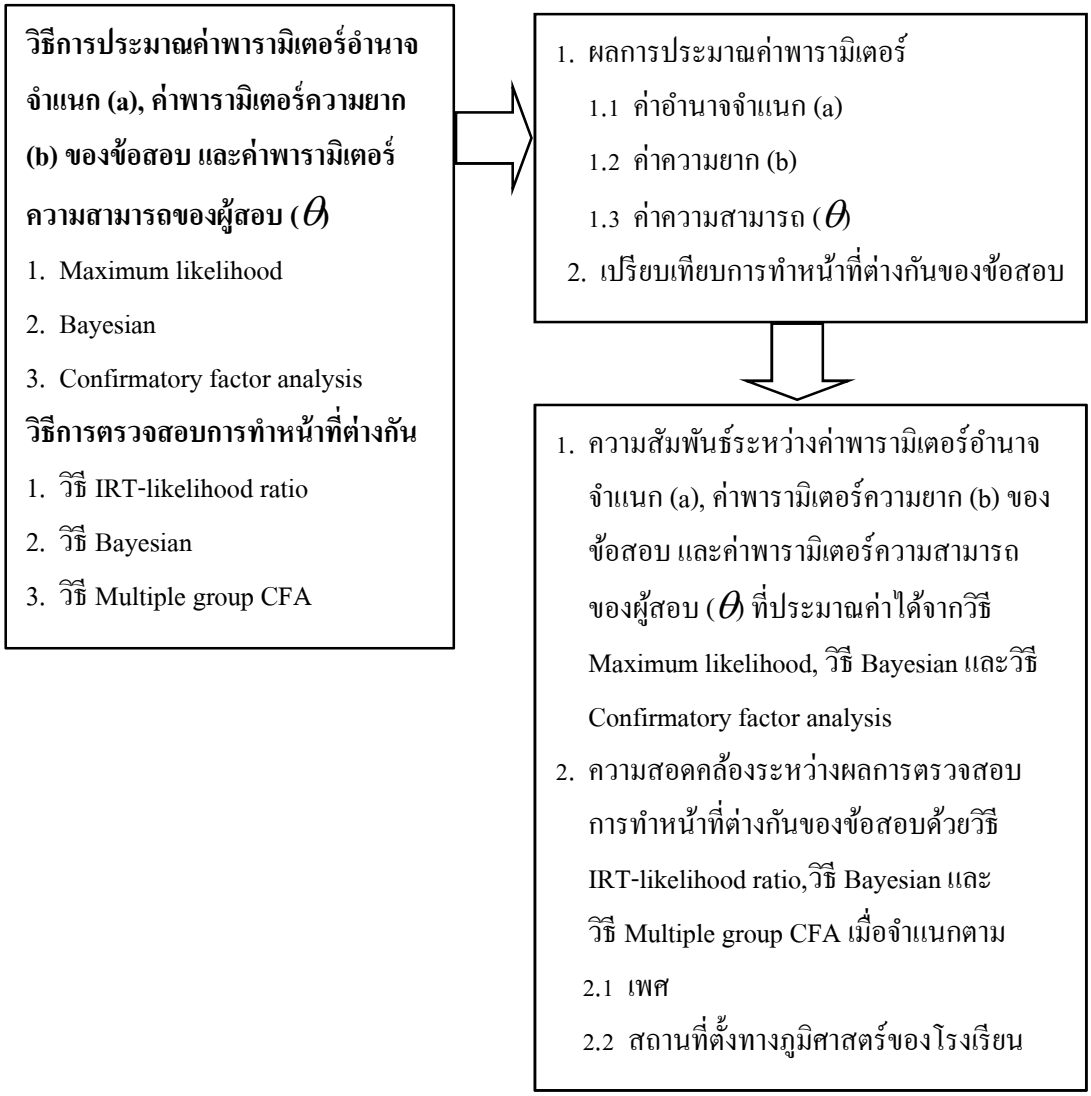
7. วิธี BAYESIAN หมายถึง วิธีการประมาณค่าโดยใช้ฟังก์ชันความน่าจะเป็นที่แสดงความสัมพันธ์ของข้อมูลที่สามารถหาค่าได้กับเซตของพารามิเตอร์ที่ไม่ทราบค่า กำหนดให้เป็น θ และประมาณค่า θ หรือการแจกแจงภายหลังนั้นขึ้นอยู่กับฟังก์ชันความน่าจะเป็นและการแจกแจงก่อนหน้า ซึ่งฟังก์ชันความน่าจะเป็นนั้นเกี่ยวข้องกับค่าสังเกต และฟังก์ชันการแจกแจงก่อนหน้าเป็นการแจกแจงที่ผู้วิจัยกำหนดขึ้นเอง โดยอาศัยความรู้ ความเชื่อ และประสบการณ์ของผู้วิจัยที่มีต่อค่า θ

8. วิธีการ Multiple group confirmatory factor analysis: MGCFA เป็นการศึกษาความไม่แปรเปลี่ยนของโมเดลการวัด (MI) ที่สัมพันธ์กับพารามิเตอร์ของข้อสอบ ได้แก่ โมเดลการตอบสนองข้อสอบแบบ 2 PL ได้แก่ ค่าน้ำหนักองค์ประกอบ (Factor loading) และค่าจุดตัด (Threshold) ของข้อสอบระหว่างประชากรต่างกลุ่ม

9. ตัวแปรตาม ได้แก่ ความสัมพันธ์ระหว่างค่าพารามิเตอร์อำนาจจำแนก (a) ค่าพารามิเตอร์ความยาก (b) ของข้อสอบ และค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) ที่ประมาณค่าได้ จากวิธี Maximum likelihood, วิธี Bayesian และวิธี Confirmatory factor analysis และความสอดคล้องระหว่างผลการตรวจสอบการทาหน้าที่ต่างกันของข้อสอบด้วยวิธี IRT-likelihood ratio,วิธี Bayesian และ วิธี Multiple group CFA เมื่อจำแนกตามเพศและสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน

10. O-NET (Ordinary national educational test) หมายถึง การทดสอบทางการศึกษาระดับชาติขั้นพื้นฐานเป็นการสอบความรู้รวบยอดปลายช่วงชั้น (6 ภาคเรียน) ของนักเรียนที่กำลังศึกษาในระดับชั้นประถมศึกษาปีที่ 6 ระดับชั้นมัธยมศึกษาปีที่ 3 และระดับชั้นมัธยมศึกษาปีที่ 6 ตามมาตรฐานการเรียนรู้หลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน พุทธศักราช 2551

11. คะแนนการทดสอบระดับชาติหรือ O-NET คือ คะแนนที่ได้มาจากการเข้าทดสอบซึ่งจัดสอบโดยสถาบันทดสอบแห่งชาติ หรือ สทศ ในทุกรายวิชารวม 5 กลุ่มสาระ ได้แก่ ภาษาไทย สังคมศึกษา ภาษาอังกฤษ คณิตศาสตร์ และวิทยาศาสตร์



ภาพที่ 1-1 กรอบแนวคิดในการวิจัย

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

ผู้วิจัยได้ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบและแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียน โดยนำเสนอ ตามลำดับดังนี้

ตอนที่ 1 แนวคิดทฤษฎีการตอบสนองข้อสอบ (Item response theory)

ตอนที่ 2 การวิเคราะห์ข้อสอบตามแนวทฤษฎีการตอบสนองข้อสอบ IRT

ตอนที่ 3 การทำหน้าที่ต่างกันของข้อสอบ (Differential item functioning)

ตอนที่ 4 ทฤษฎีการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ (Theoretic comparison of method)

ตอนที่ 5 แนวคิดการวิเคราะห์อัตราส่วนความควรจะเป็น (IRT-likelihood ratio test: LRT)

ตอนที่ 6 แนวคิดการวิเคราะห์ด้วยวิธี BAYESIAN

ตอนที่ 7 แนวคิดการวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ (Multiple group confirmatory factor analysis: MG-CFA)

ตอนที่ 8 การทดสอบวัดผลสัมฤทธิ์ระดับชาติของนักเรียน O-NET

ตอนที่ 9 งานวิจัยที่เกี่ยวข้องกับทำหน้าที่ต่างกันของข้อสอบ

ตอนที่ 1 แนวคิดทฤษฎีการตอบสนองข้อสอบ (Item response theory)

1. ทฤษฎีการทดสอบแบบดั้งเดิม (Classical test theory: CTT) เป็นทฤษฎีที่มุ่งศึกษาความสัมพันธ์ระหว่างคะแนนสังเกตได้กับความสามารถที่แท้จริงของผู้สอบ จากแนวคิดดังกล่าวได้นำมาสู่การสร้างข้อตกลงเบื้องต้นที่เกี่ยวข้องกับโมเดลการวัดและแบบสอบคู่ขนานอยู่หลายประการ และเป็นที่มาของการวิเคราะห์ข้อสอบตามแนวทฤษฎีการทดสอบแบบดั้งเดิม ซึ่งประกอบด้วย ตัวบ่งชี้คุณภาพข้อสอบ 3 ประการ ได้แก่ ค่าความยากของข้อสอบ (Item difficulty) ค่าอำนาจจำแนกของข้อสอบ (Item discrimination) และประสิทธิภาพตัวลวง (Item distractor) หลักการวิเคราะห์คุณภาพข้อสอบจะอยู่บนพื้นฐานของคุณลักษณะ และจำนวนผู้สอบในการสอบแต่ละครั้ง การวิเคราะห์ค่าความยากง่ายของข้อสอบคำนวณจากอัตราส่วนของจำนวนผู้สอบที่ตอบข้อสอบถูกต่อจำนวนผู้เข้าสอบทั้งหมด ยกตัวอย่าง เช่น ในการสอบครั้งหนึ่งมีจำนวน 100 คน ในข้อสอบ K ผู้สอบที่ตอบข้อสอบถูกต้องมี 65 คน ข้อสอบข้อนี้มีความยากง่ายของข้อสอบ เท่ากับ

0.65 จากวิธีการวิเคราะห์ที่อิงอยู่กับจำนวนผู้เข้าสอบ และคุณลักษณะของผู้สอบ ทำให้มีข้อพิพากษ์เกี่ยวกับความถูกต้อง ความคงเส้นคงวาของผลการวิเคราะห์ที่มีความผันแปรไปตามกลุ่มผู้สอบ ถึงแม้ว่าผู้สอบทุกกลุ่มจะทำข้อสอบชุดเดียวกัน การวิเคราะห์คุณภาพ

ทฤษฎีการทดสอบแบบดั้งเดิมมีข้อตกลงเบื้องต้นหลายข้อ ซึ่งนักวัดผลเห็นว่าเป็นข้อตกลงเบื้องต้นที่ไม่สมเหตุสมผล (Weak assumption) และไม่สอดคล้องกับสภาพของการทดสอบในปัจจุบัน ไม่ว่าจะเป็นการกำหนดให้ค่าความคลาดเคลื่อนมีความเท่ากันในทุกผู้สอบ หรือข้อตกลงเกี่ยวกับความเป็นคู่ขนานที่เป็นไปได้ยากในทางปฏิบัติ

Hambleton and Swaminathan (1985) ได้กล่าวถึงข้อด้อยของการวิเคราะห์คุณภาพข้อสอบตามทฤษฎีการทดสอบแบบดั้งเดิมไว้ 5 ประการ ดังนี้

1. ค่าสถิติที่ใช้ในการวิเคราะห์ข้อสอบ เช่น ค่าความยาก และค่าอำนาจจำแนกของข้อสอบ ขึ้นอยู่กับกลุ่มผู้สอบที่เข้าสอบในแต่ละครั้ง เช่น หากกลุ่มผู้สอบมีความสามารถแตกต่างกันมาก (Heterogeneoud in ability) ดัชนีค่าความยาก และค่าอำนาจจำแนกของข้อสอบจะมีแนวโน้มสูง เป็นต้น

2. วิธีการและเทคนิคการออกแบบทดสอบและการวิเคราะห์เพื่อการเปรียบเทียบค่าความสามารถของผู้สอบที่วัดได้ขึ้นอยู่กับชุดข้อสอบที่เป็นคู่ขนานกัน หรือสมมูลกัน หรือแบบสอบชุดเดียวกัน ภายใต้สถานการณ์การทดสอบเดียวกัน หรือใกล้เคียงกัน นอกจากนี้ยังไม่สามารถเปรียบเทียบกันได้ ภายใต้สถานการณ์การทดสอบเดียวกัน หรือใกล้เคียงกัน เช่น ผู้ที่ได้คะแนน 60% ในแบบสอบชุดที่ค่อนข้างง่าย มีความสามารถที่สูงหรือต่ำกว่าผู้สอบที่ได้คะแนน 40% จากแบบทดสอบที่มีความยากกว่า

3. การคำนวณค่าความเที่ยงของแบบสอบ กำหนดจากส่วนของความเป็นแบบสอบคู่ขนานซึ่งมีความเป็นไปได้ยากที่จะปฏิบัติจริง

4. ทฤษฎีการทดสอบแบบดั้งเดิมไม่มีสารสนเทศที่จะบ่งชี้ว่าผู้สอบแต่ละคนมีโอกาสที่จะตอบข้อสอบได้ถูกต้องในระดับใด ซึ่งสารสนเทศนี้มีความจำเป็นอย่างยิ่งต่อการออกแบบแบบทดสอบปรับเหมาะที่มีระดับความยากของข้อสอบเหมาะสมกับระดับความสามารถของผู้สอบ (Tailor test) หรือแบบทดสอบที่สามารถคัดเลือกข้อสอบที่เหมาะสมกับความสามารถของผู้สอบได้ด้วยระบบคอมพิวเตอร์ (Computer adaptive testing)

5. ทฤษฎีทดสอบแบบดั้งเดิมสมมติว่าความแปรปรวนของความคลาดเคลื่อนของการวัด (Error of measurement) มีค่าเท่ากันในผู้สอบทุกคน

ประมาณปี ค.ศ. 1936 มีผู้ริเริ่มเสนอแนวคิดเกี่ยวกับแบบจำลองการวัดคุณลักษณะของบุคคลเพื่อแก้ไขจุดอ่อนของทฤษฎีการทดสอบแบบมาตรฐานเดิม การพัฒนาเชิงคณิตศาสตร์ของ

ทฤษฎีการตอบสนองข้อสอบ เริ่มต้นโดย ริชาร์ดสัน (Richardson, 1936) ทัคเคอร์ (Tucker, 1936) ลอว์เลย์ (Lowley, 1944) ลอร์ด (Lord, 1952) แต่ผู้ที่ทำให้แนวคิดนี้สามารถนำไปปฏิบัติได้ คือ (Bimbuam, 1957; อุทุมพร จามรมาน, 2540, หน้า 155)

ทฤษฎีการการตอบสนองข้อสอบมีลักษณะที่สำคัญ ดังนี้

1. คุณลักษณะของพารามิเตอร์ข้อสอบไม่ขึ้นอยู่กับกลุ่มผู้ตอบข้อสอบเฉพาะกลุ่มใดกลุ่มหนึ่ง

2. มีการอธิบายความสามารถของผู้ตอบข้อสอบแยกออกจากพารามิเตอร์ข้อสอบ

3. อธิบายความหมายการตอบข้อสอบของผู้ตอบข้อสอบในระดับข้อสอบมากกว่าระดับแบบทดสอบ

4. ในการคำนวณค่าสารสนเทศของแบบทดสอบไม่ใช้แบบทดสอบคู่ขนาน

5. อธิบายความหมายการตอบสนองข้อสอบของผู้ตอบข้อสอบทุกคนอย่างถูกต้องที่ทุกระดับความสามารถและประมาณค่าความคลาดเคลื่อนมาตรฐานในการวัดเป็นรายบุคคล

ทฤษฎีการตอบสนองข้อสอบเป็นทฤษฎีที่อธิบายความสัมพันธ์ระหว่างคุณลักษณะภายในหรือความสามารถของผู้ตอบข้อสอบด้วยฟังก์ชันทางคณิตศาสตร์ที่เขียนในรูปแบบจำลองการตอบข้อสอบ และแสดงการตอบข้อสอบรายข้อด้วยโค้งคุณลักษณะข้อสอบ (Item characteristic curve: ICC) ที่แสดงความสัมพันธ์ระหว่างระดับความสามารถของบุคคลที่วัดด้วยข้อสอบข้อนั้นกับโอกาสที่บุคคลที่มีความสามารถระดับต่าง ๆ จะตอบข้อสอบได้ถูกต้อง

ทฤษฎีการตอบสนองข้อสอบ (Item response theory) เป็นทฤษฎีการวัดที่อธิบายความสัมพันธ์ระหว่างคุณลักษณะภายในหรือความสามารถที่มีอยู่ในตัวบุคคลกับพฤติกรรมการตอบสนองข้อสอบของบุคคลนั้น ว่ามีโอกาสตอบข้อสอบถูกมากน้อยเพียงใด ทฤษฎีนี้มีพื้นฐานความเชื่อว่าพฤติกรรมการตอบข้อสอบของผู้สอบถูกกำหนดโดยคุณลักษณะภายในหรือความสามารถที่มีอยู่ในตัวบุคคล (ศิริชัย กาญจนวาสี, 2545) ทฤษฎีนี้มีแนวความคิดพื้นฐานที่สำคัญ 2 ประการ คือ (Hambleton, Swaminatan & Roger, 1991)

1. พฤติกรรมการแสดงออกของบุคคลในการตอบข้อสอบสามารถทำนายได้ด้วยกลุ่มขององค์ประกอบที่เรียกว่าคุณลักษณะ (Trait) หรือคุณลักษณะแฝง (Latent traits) หรือความสามารถ (Abilities) ของบุคคลนั้น

2. ความสัมพันธ์ระหว่างการแสดงพฤติกรรมการตอบสนองของข้อสอบกับชุดของคุณลักษณะแฝงที่อยู่ภายใต้คุณลักษณะที่แสดงออกของข้อสอบสามารถอธิบายได้โดยฟังก์ชั่นที่เพิ่มขึ้นในลักษณะทิศทางเดียว (Monotonically) ที่รียกว่าฟังก์ชั่นคุณลักษณะข้อสอบหรือโค้งลักษณะข้อสอบ (Item characteristic curve)

หลักการของทฤษฎีการตอบสนองข้อสอบ

ทฤษฎีการตอบสนองของข้อสอบเป็นทฤษฎีที่อธิบายถึงความสัมพันธ์ระหว่างลักษณะหรือความสามารถที่มีอยู่ในตัวบุคคลกับพฤติกรรมการตอบสนองข้อสอบของบุคคลนั้น โดยทฤษฎีนี้มีความเชื่อว่าพฤติกรรมการตอบสนองข้อสอบของบุคคลจะถูกกำหนดโดยลักษณะหรือความสามารถที่มีอยู่ในตัวบุคคล (Lord & Novick, 1980, p. 353) ซึ่งไม่สามารถสังเกตได้ทฤษฎีนี้พยายามอธิบายความสัมพันธ์ระหว่างลักษณะภายในตัวบุคคลกับพฤติกรรมที่บุคคลตอบสนองต่อข้อสอบ การอธิบายความสัมพันธ์ระหว่างลักษณะภายในตัวบุคคลกับพฤติกรรมที่บุคคลตอบสนองต่อข้อสอบ โดยให้คะแนนที่ได้รับจากการตอบข้อสอบ (y) แทนพฤติกรรมการตอบสนองข้อสอบ ให้ θ แทนลักษณะหรือความสามารถในตัวบุคคล และ r_i เป็นเกณฑ์ที่บอกว่า y ระดับใดจึงจะทำข้อสอบข้อที่ i ได้ถูก ดังนั้นถ้า $y > r_i$ แสดงว่า ทำข้อสอบข้อที่ i ได้ถูก และถ้า $y < r_i$ แสดงว่าทำข้อสอบนั้นผิด

ฟังก์ชันทางคณิตศาสตร์ที่แสดงความสัมพันธ์ระหว่างระดับความสามารถกับพฤติกรรมการตอบสนองเรียกว่าฟังก์ชันการตอบสนองข้อสอบ หรือฟังก์ชันลักษณะข้อสอบ (Item characteristic function) ซึ่งสามารถเขียนฟังก์ชันได้ ดังนี้

$P(\theta)$ = Prob (U-1/ θ) เมื่อ U = 0, 1

จากฟังก์ชันข้างต้น หมายถึงโอกาสที่ผู้สอบซึ่งมีความสามารถ θ จะตอบคำถามข้อที่ i ได้ถูกต้อง

ข้อตกลงเบื้องต้นทฤษฎีการตอบสนองข้อสอบ

ฟังก์ชันทางคณิตศาสตร์ที่ใช้ในทฤษฎีการตอบข้อสอบ กำหนดไว้ว่าความน่าจะเป็นของการตอบข้อสอบต้องขึ้นอยู่กับความสามารถของผู้สอบและขึ้นอยู่กับคุณลักษณะของข้อสอบโมเดลการตอบสนองข้อสอบจึงตั้งอยู่บนพื้นฐานความเชื่อหรือข้อตกลงเบื้องต้นหลายประการเกี่ยวกับลักษณะข้อมูลที่จะทำให้โมเดลสามารถนำไปใช้ได้อย่างเหมาะสม ถึงแม้ข้อตกลงเบื้องต้นบางประการจะตรวจสอบไม่ได้โดยตรง แต่เราสามารถเก็บรวบรวมข้อมูลหรือหลักฐานทางอ้อมมาช่วยสนับสนุนยืนยันได้ ข้อตกลงเบื้องต้นที่สำคัญของทฤษฎีการตอบสนองข้อสอบมีดังนี้ (Lord & Novick, 1968: Hambleton & Swaminathan, 1985)

1. ความเป็นมิติเดียว (Unidimensionality: One trait)

ข้อตกลงเบื้องต้นที่ใช้กันทั่วไปสำหรับทฤษฎีการตอบสนองข้อสอบ คือ ข้อคำถามจะมุ่งวัดเพียงคุณลักษณะเดียว หรือความสามารถเดียว (One ability) ถ้าความสามารถของคนคนหนึ่งมี k ด้าน ซึ่งความสามารถแต่ละด้านต่างกันก็ส่งผลต่อการตอบข้อสอบข้อต่าง ๆ ที่รวมกันเป็นแบบทดสอบ ถ้าผลที่ได้จากการตอบแบบทดสอบ (Test perforance) หรือคะแนนของผู้ตอบ

สามารถอธิบายได้เพียงความสามารถเดียว จะเรียกว่าความเป็นมิติเดียว (Unidimensionality) อย่างไรก็ตามข้อตกลงเบื้องต้นนี้ไม่ได้เข้มงวดนัก ถ้าแบบทดสอบมีลักษณะเด่น (Dominant) ที่จะวัดองค์ประกอบใดองค์ประกอบหนึ่งก็ถือเสมือนว่ามีความเป็นมิติเดียว ซึ่งการทดสอบความเป็นมิติเดียวของแบบทดสอบสามารถทำได้หลายวิธี วิธีการหนึ่งที่นิยมใช้ในการตรวจสอบ คือ การวิเคราะห์องค์ประกอบ (Factor analysis) โดยพิจารณาจากค่า Eigen valua ว่าให้ค่าที่แตกต่างไปจากข้ออื่น ๆ หรือไม่

2. ความเป็นอิสระในการตอบข้อสอบ (Independence: Local independent)

แนวคิดเกี่ยวกับความเป็นอิสระระหว่างข้อสอบและผู้สอบ มีความเกี่ยวข้องมาจาก “ความเป็นมิติเดียวของแบบทดสอบ ความเป็นอิสระของข้อสอบและผู้สอบ หมายถึง ความน่าจะเป็นหรือโอกาส (Probability) ในการตอบข้อสอบได้ถูกต้องเป็นอิสระจากกัน นั่นคือ เมื่อมีการควบคุมความสามารถ (θ) ที่ส่งผลต่อการตอบข้อสอบ หรือให้ความสามารถของผู้สอบ (θ) คงที่แล้ว ผลการตอบข้อสอบแต่ละข้อจะต้องเป็นอิสระจากกัน หรือกล่าวได้อีกนัยหนึ่งว่าเมื่อควบคุมอิทธิพลของความสามารถ (θ) แล้วผลการตอบข้อสอบรายข้อไม่มีความสัมพันธ์กัน นั่นคือ โมเดลการตอบสนองข้อสอบมีเพียงความสามารถ (θ) ปัจจัยเดียวเท่านั้นที่มีผลต่อการตอบสนองรายข้อ ความเป็นอิสระสามารถจำแนกพิจารณาเป็นความอิสระระหว่างข้อสอบและความเป็นอิสระระหว่างผู้สอบ ดังนี้

2.1 ความเป็นอิสระระหว่างข้อสอบ (Local item independence) นั่นคือ เมื่อสุ่มผู้สอบซึ่งมีความสามารถ θ ขึ้นมา 1 คน ในการตอบข้อสอบ k ข้อให้ u_i เป็นผลการตอบหรือคะแนนข้อที่ i หลังจากควบคุมความสามารถ (θ) ของผู้สอบแล้ว คะแนนผลการตอบของผู้สอบคนนั้น ๆ ในแต่ละข้อไม่สัมพันธ์กัน ถ้าผลการตอบรายข้อของผู้สอบคนเดียวกันเป็นอิสระจากกัน ความน่าจะเป็นของแบบแผนการตอบข้อที่ k ของผู้สอบที่มีความสามารถ θ จะเท่ากับ ผลคูณระหว่างความน่าจะเป็นของผลการตอบข้อสอบแต่ละข้อ ซึ่งเขียนในรูปสมการคณิตศาสตร์ ดังต่อไปนี้

$$P(U_1,U_2,....,U_i \mid \theta) = \prod_{i=1}^{k} P(U_i \mid \theta) \qquad (1)$$

2.2 ความเป็นอิสระระหว่างผู้สอบ (Local person independence) นั่นคือ เมื่อสุ่มข้อสอบขึ้นมา 1 ข้อ ในการตอบข้อสอบของผู้สอบ n คน ให้เป็นผลการตอบหรือคะแนนข้อสอบของผู้สอบคนที่ i หลังจากควบคุมความสามารถ (θ) ของผู้สอบแต่ละคนแล้ว คะแนนผลการตอบข้อนั้นของผู้สอบแต่ละคนไม่สัมพันธ์กัน ถ้าผลการตอบข้อสอบข้อเดียวกันของผู้สอบแต่ละคนเป็นอิสระจากกัน ความน่าจะเป็นของแบบแผนการตอบข้อสอบของผู้สอบ n คน จะ เท่ากับ ผลคูณ

ระหว่างความน่าจะเป็นผลการตอบข้อนั้นของผู้สอบแต่ละคนซึ่งเขียนในรูปสมการทางคณิตศาสตร์ได้ดังต่อไปนี้

$$P(U_{1,},U_2,....,U_n \mid \theta) = \prod_{j=1}^{n} P(U_j \mid \theta) \qquad (2)$$

3. โค้งลักษณะข้อสอบ (Item characteristic curve: Item response model)

โค้งคุณลักษณะของข้อสอบเป็นฟังก์ชันทางคณิตศาสตร์สามารถใช้อธิบายความสัมพันธ์ระหว่างความน่าจะเป็น หรือโอกาสที่ผู้ตอบจะตอบข้อสอบได้ถูกต้องกับระดับความสามารถที่วัดได้โดยชุดของข้อสอบ หรือแบบทดสอบฉบับนั้น ทั้งนี้ ความน่าจะเป็นหรือโอกาสในการตอบข้อสอบถูกจะขึ้นอยู่กับโค้งลักษณะข้อสอบในแต่ละโมเดลที่เลือกใช้ โดยที่รูปร่าง (Shape) ของโค้งลักษณะข้อสอบในแต่ละข้อมีคุณสมบัติไม่แปรเปลี่ยน (Invariant) ไปตามกลุ่มตัวอย่างที่ใช้ ดังนั้นจึงทำให้ความน่าจะเป็นหรือโอกาสในการตอบข้อสอบถูกในแต่ละข้อไม่แปรเปลี่ยนด้วย คุณสมบัตินี้ถือว่าเป็นลักษณะเด่นของโมเดลต่าง ๆ ในทฤษฎีการตอบสนองข้อสอบ โค้งลักษณะข้อสอบมีหลายรูปแบบขึ้นอยู่กับว่าเลือกใช้โมเดลพารามิเตอร์ข้อสอบกี่ตัว

4. ข้อสอบที่ใช้ต้องไม่เป็นข้อสอบประเภทความเร็ว

ทฤษฎีการตอบสนองข้อสอบ (IRT) ถือว่าความสามารถ (θ) เป็นปัจจัยสำคัญต่อผลการสอบ ความเร็วในการตอบจะต้องไม่มีอิทธิพลต่อผลการตอบ การจัดการสอบจึงต้องไม่อยู่ในสถานการณ์ที่สอบแข่งขันกันด้วยเวลา การสอบจะต้องอยู่ในลักษณะที่ผู้สอบซึ่งมีความสามารถมีเวลาเพียงพอในการทำข้อสอบ

ข้อตกลงเบื้องต้นของ IRT มีประโยชน์ที่จะนำ IRT มาใช้ในการตรวจสอบ DIF เพราะนอกจากคุณสมบัติความเป็นเอกมิติแล้วทำให้สามารถทดสอบ DIF ได้จากหลักของความเป็นพหุมิติของข้อสอบ (Multidimemsional) ถ้าข้อสอบใดที่ผู้สอบอยู่ในกลุ่มใดกลุ่มหนึ่งจะมีโอกาสในการตอบข้อสอบข้อนั้นได้ถูกต้องกว่าผู้สอบอีกกลุ่มหนึ่งที่มีความสามารถเท่ากัน จำเป็นต้องอาศัยความสามารถอื่น ๆ มาประกอบ แล้วจะถือว่าข้อสอบข้อนั้นมี DIF เกิดขึ้น (Shealy & Stout, 1993) นอกจากนั้นการตกลงความเป็นอิสระของข้อสอบ ทำให้วิธีการตรวจสอบ DIF สามารถทำได้ทีละข้อ แต่อย่างไรก็ตามวิธีการตรวจสอบ DIF โดยใช้ IRT นั้นแตกต่างกันตามตัวแบบ IRT ที่ใช้

พารามิเตอร์ในทฤษฎีการตอบสนองข้อสอบ

พารามิเตอร์ในทฤษฎีการตอบสนองข้อสอบ แบ่งออกเป็น 2 ชนิด คือ พารามิเตอร์ข้อสอบ (Item parameter) ได้แก่ ค่าความยาก (b_i) ค่าอำนาจจำแนก (a_i) ค่าการเดา (c_i) และความรอบคอบ (y_i) ส่วนพารามิเตอร์ของผู้สอบ (Person parameter) ได้แก่ ระดับความสามารถหรือคุณลักษณะของผู้สอบ (θ) ซึ่งพิสัยของค่าพารามิเตอร์ต่าง ๆ มีดังนี้

1. พารามิเตอร์ความยาก (b) หมายถึง สัดส่วนของคนที่ทำข้อสอบข้อนั้นถูก หรือหมายถึงค่าที่แสดงถึงระดับความสามารถของผู้สอบ (θ) ที่จุดโค้งลักษณะข้อสอบมีความชันมากที่สุดมีค่าตั้งแต่ -∞ ถึง ∞ แต่ในทางปฏิบัติจะอยู่ระหว่าง -3 ถึง +3 ค่า -3 บ่งบอกว่าข้อสอบข้อนั้นง่ายมากและค่า +3 แสดงว่าข้อสอบข้อนั้นยากมาก

2. พารามิเตอร์อำนาจจำแนก (a) หมายถึง ความสามารถของผู้สอบที่แยกผู้สอบออกเป็น 2 กลุ่ม คือ กลุ่มตอบถูกกับกลุ่มตอบผิดหรือเป็นค่าที่เป็นสัดส่วนโดยตรงกับความชันของโค้งคุณลักษณะของข้อสอบ ณ จุดเปลี่ยนโค้งมีค่าตั้งแต่ -∞ ถึง ∞ แต่ในทางปฏิบัติมีค่าตั้งแต่ 0 ถึง 2 เพราะค่า a ที่เป็นลบแสดงว่าข้อสอบไม่ดี ใช้ไม่ได้ต้องตัดทิ้ง ค่า 0 แสดงว่าข้อสอบไม่มีค่าอำนาจจำแนก ค่า+2 แสดงว่าข้อสอบมีค่าอำนาจจำแนกสูงในการคัดเลือกข้อสอบ ข้อที่คัดไว้จะมีค่า a ตั้งแต่ 0.3 ขึ้นไป

3. พารามิเตอร์การเดา (c_i) เป็นค่าแสดงความน่าจะเป็นหรือโอกาสของการตอบข้อสอบได้ถูกต้อง โดยไม่มีความรอบรู้ หรือคุณลักษณะในเรื่องนั้น ๆ ในทางทฤษฎีพารามิเตอร์การเดามีค่าระหว่าง 0 ถึง 1 โดยทั่วไปนิยมใช้ข้อสอบที่มีค่าพารามิเตอร์การเดาไม่เกิน .30

4. ความรอบครอบ (y_i) McDoald (1967) และ Barton and Lord (1981) ได้เสนอพารามิเตอร์ที่แสดงถึงความรอบครอบของผู้สอบ เป็นค่าพารามิเตอร์ที่บ่งชี้ว่าผู้สอบที่มีความสามารถสูงอาจจะตอบข้อสอบได้ไม่ถูกต้องเสมอไป ซึ่งอาจเกิดจากความรอบครอบในการพิจารณาคำตอบ หรือผู้สอบอาจจะมีสารสนเทศอื่น ๆ เกี่ยวกับผู้ออกข้อสอบทำให้เลือกคำตอบในตัวเลือกที่ไม่ใช่คำตอบที่ถูกต้อง Barton and Lord (1981) กล่าวว่า พารามิเตอร์ตัวนี้จะเหมาะสมในการศึกษาทางทฤษฎีเท่านั้นในทางปฏิบัติแล้วไม่สามารถพบพารามิเตอร์นี้ได้ (Hambleton & Swaminathan, 1985)

5. พารามิเตอร์ผู้สอบ เป็นระดับความสามารถของผู้สอบ (θ) ที่ประมาณได้จากโมเดลตามทฤษฎีการตอบสนองข้อสอบ นิยมปรับให้เป็นคะแนนมาตรฐานที่มีค่าเฉลี่ยเป็น 0 และส่วนเบี่ยงเบนมาตรฐานเป็น 1 ซึ่งพารามิเตอร์ผู้สอบมีค่าระหว่าง -∞ ถึง ∞ แต่ส่วนใหญ่จะมีค่าอยู่ในช่วง -3.0 ถึง +3.0 ค่าที่เป็นลบแสดงว่าผู้สอบมีความสามารถต่ำ และค่าที่เป็นบวกแสดงว่าผู้สอบมีความสามารถสูง

ข้อตกลงเบื้องต้นและพารามิเตอร์ที่กล่าวมานี้ มีความหมายเด่นชัดในกรณีที่ข้อสอบนั้นคะแนนแบบทวิภาค ในการประยุกต์ทฤษฎีเพื่อใช้กับข้อสอบที่ให้คะแนนแบบหลายค่า ข้อตกลงเบื้องต้นทั้งหมดก็เทียบเคียงในทำนองเดียวกัน แตกต่างกันเพียงรายละเอียดปลีกย่อยเกี่ยวกับเงื่อนไขเฉพาะของแต่ละโมเดลเท่านั้น

การเปรียบเทียบประสิทธิภาพของวิธีประมาณค่าพารามิเตอร์

เนื่องจากวิธีประมาณค่าพารามิเตอร์มีหลายวิธี จึงจำเป็นต้องมีการเปรียบเทียบประสิทธิภาพในการประมาณค่าของวิธีประมาณ 2 วิธีใด ๆ ซึ่งทำได้โดยหาอัตราส่วนของค่าความแปรปรวนของวิธีประมาณ วิธีหนึ่ง ต่อค่าความแปรปรวนของวิธีประมาณอีกวิธีหนึ่ง เช่น สมมติว่า มีวิธีประมาณ 2 วิธี M_1 และ M_2 เป็นฟังก์ชัน 2 ฟังก์ชันของตัวแปรหุ่นในตัวอย่าง โดยมีค่าความแปรปรวนของวิธีประมาณเป็น $V(M_1)$ และ $V(M_2)$ ตามลำดับ ก็อาจวัดประสิทธิภาพของวิธีประมาณ M_1 เทียบกับ M_1 ได้ด้วยค่า $V(M_1)/ V(M_2)$ (สุชาดา กีระนันทน์, 2525, หน้า 42-43) แต่จากการศึกษาเรื่องดัชนีกำหนดคุณภาพของแบบทดสอบ พบว่า ความแปรปรวนที่ได้จากการประมาณค่า เป็นสัดส่วนกลับกันกับฟังก์ชันสารสนเทศแบบทดสอบ เมื่อพิจารณาความสัมพันธ์ดังกล่าว โดยใช้หลักการทางตรรกวิทยา สรุปได้ว่า

$$\frac{V(M_1)}{V(M_2)} = \frac{\mathrm{I}(\theta, \mathrm{M}_2)}{\mathrm{I}(\theta, \mathrm{M}_1)}$$

ดังนั้นจึงสามารถเปรียบเทียบค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับเดียวกันที่เป็นผลจากการประมาณค่าด้วยวิธีประมาณ 2 วิธีใด ๆ แทนการเปรียบเทียบประสิทธิภาพของวิธีประมาณค่าพารามิเตอร์ 2 วิธีใด ๆ ในแบบทดสอบฉบับเดียวกันได้ กล่าวคือ ประสิทธิภาพของวิธีประมาณค่าพารามิเตอร์ x เทียบกับวิธีประมาณค่าพารามิเตอร์ y ในแบบทดสอบฉบับหนึ่ง (EFF (x, y)) จะเท่ากับค่าประสิทธิภาพสัมพันธ์ของแบบทดสอบฉบับนั้น ที่เป็นผลจากการประมาณค่าพารามิเตอร์วิธี y เทียบกับวิธี x ณ ระดับความสามารถใด ๆ (RE (y, x)) ดังสมการ

$$EFF(\mathrm{x},\mathrm{y}) = \frac{v(\theta,\mathrm{x})}{\mathrm{v}(\theta,\mathrm{y})} = \frac{\mathrm{I}(\theta,\mathrm{y})}{\mathrm{I}(\theta,\mathrm{y})} = \mathrm{RE}(\mathrm{y},\mathrm{x}) \qquad (3)$$

ในการแปลความหมายสามารถพิจารณาจากเกณฑ์ ดังนี้

EFF (X, Y) = 1 แสดงว่าการประมาณค่าด้วยวิธี Y มีประสิทธภาพ เท่ากับการประมาณค่าด้วยวิธี X เนื่องจากการประมาณค่าพารามิเตอร์ด้วยวิธี X มีความแปรปรวนเท่ากับวิธี Y

EFF (X, Y) > 1 แสดงว่า การประมาณค่าด้วยวิธี Y มีประสิทธิภาพสูงกว่าการประมาณค่าด้วยวิธี X เนื่องจากการประมาณค่าพารามิเตอร์ด้วยวิธี X มีความแปรปรวนสูงกว่าวิธี Y

EFF (X, Y) < 1 แสดงว่าการประมาณค่าด้วยวิธี Y มีประสิทธิภาพต่ำกว่าการประมาณค่าด้วยวิธี X เนื่องจากการประมาณค่าพารามิเตอร์ด้วยวิธี X มีความแปรปรวนต่ำกว่าวิธี Y

ตอนที่ 2 การวิเคราะห์ข้อสอบตามแนวทฤษฎีการตอบสนองข้อสอบ (IRT)

ความเป็นมาของทฤษฎีการตอบสนองข้อสอบ

ทฤษฎีการตอบสนองข้อสอบ (Item response theory) หรือทฤษฎีคุณลักษณะแฝง (Latent trait theory) ได้รับการพัฒนามานานตั้งแต่บิเนต และไซมอน (Binet & Simon, 1916) และริชาร์ดสัน (Richardson, 1936) ได้แสดงความสัมพันธ์ระหว่างพารามิเตอร์ของแบบจำลองทฤษฎีการตอบสนองข้อสอบและทฤษฎีแบบดั้งเดิม ต่อมาในปี ค.ศ. 1943-1944 ลอร์เลย์ (Hambleton & Swaminathan, 1985) ได้เสนอวิธีการประมาณค่าพารามิเตอร์แนวใหม่ ซึ่งมีหลักการว่าผลการสอบของผู้สอบจากแบบทดสอบใด ๆ ขึ้นอยู่กับความสามารถ (Ability or skill) ของผู้สอบ จนกระทั่งปี ค.ศ. 1952 ลอร์ด (Lord) ได้เสนอทฤษฎีใหม่ในรูปโค้งลักษณะข้อสอบ (Item characteristic curve: ICC) โดยลอร์ดได้เสนอว่าโค้งลักษณะข้อสอบเป็นโค้งปกติสะสม ต่อมาเรียกว่า แบบจำลองนอร์มัลออไจฟ์ (Normal ogive model) ซึ่งแบบจำลองนี้มีความยากในการคำนวณ และขาดโปรแกรมคอมพิวเตอร์ที่มีประสิทธิภาพในการใช้วิเคราะห์ข้อมูลเพื่อพิสูจน์ทฤษฎี จึงทำให้ลอร์ด ชะงักการทำงานเกี่ยวกับทฤษฎีนี้ไประยะหนึ่ง

ปี ค.ศ. 1960 ราสซ์ (Rasch) ได้ศึกษาเกี่ยวกับเรื่องนี้ โดยไม่ทราบแนวคิดของลอร์ด และได้เสนอแนวคิดในรูปของพารามิเตอร์ตัวเดียว คือค่าความยากของข้อสอบเท่านั้น

ปี ค.ศ. 1965 ลอร์ด ได้หันมาสนใจและพัฒนาทฤษฎีการตอบข้อสอบใหม่อีกครั้ง

ปี ค.ศ. 1968 เบิอร์นบอร์ม (Bimbaum) ได้เสนอแบบจำลองโลจิสติก (Logistic model) ที่ใช้พารามิเตอร์ 2 ตัว คือ ค่าความยาก และค่าอำนาจจำแนก ซึ่งเป็นแบบจำลองที่ง่ายกว่าของลอร์ด จึงทำให้แบบจำลองนี้เป็นที่นิยมแพร่หลาย และมีการพัฒนาขึ้นเรื่อย ๆ จนกระทั่งสามารถใช้ได้กับพารามิเตอร์ตัวเดียว และพารามิเตอร์สามตัว

ทฤษฎีการตอบสนองข้อสอบ (IRT) เป็นทฤษฎีการวัดที่อธิบายความสัมพันธ์ระหว่างความสามารถที่มีอยู่ในตัวบุคคล (Latent trait or ability) กับผลการตอบข้อสอบหรือข้อคำถาม โดยใช้โค้งลักษณะข้อสอบ (Item characteristic curve: ICC) ซึ่งมีการกำหนดลักษณะข้อสอบด้วยพารามิเตอร์อำนาจจำแนก (*a*) ความยาก (*b*) และโอกาสในการเดาข้อสอบ (*c*) IRT จึงอยู่บนพื้นฐานแนวคิดที่สำคัญ 2 ประการ คือ 1) ผลการตอบข้อสอบหรือข้อคำถามของผู้ตอบสามารถอธิบายได้ด้วยความสามารถที่มีอยู่ภายในของผู้ตอบ และ 2) ความสัมพันธ์ระหว่างผลการตอบข้อสอบกับความสามารถที่มีอยู่ภายใน สามารถอธิบายได้ด้วยฟังก์ชันลักษณะข้อสอบหรือโค้งลักษณะข้อสอบ (ICC) อันมีลักษณะเป็นฟังก์ชันทางคณิตศาสตร์ เรียกว่า ฟังก์ชันโลจิสติก (Logistic function) หรือใกล้เคียงกับฟังก์ชันปกติสะสม (Normal ogive function) โดยพื้นฐาน

ของทฤษฎีการตอบสนองข้อสอบนั้น มุ่งอธิบายการตอบข้อสอบที่นำมากระตุ้นว่า ผู้สอบจะมีโอกาสในการตอบได้ถูกต้องเพียงใด

โมเดลของทฤษฎีการตอบข้อสอบ (Item response model)

ลักษณะทั่วไปของทฤษฎีการตอบสนองข้อสอบ คือ โมเดลการตอบสนองข้อสอบซึ่งในทางคณิตศาสตร์แสดงถึงความน่าจะเป็นในการตอบข้อสอบถูกในรูปฟังก์ชันของความสามารถของผู้สอบการตอบสนองข้อสอบ คือ ตอบถูกมีค่า เท่ากับ 1 ตอบผิดมีค่า เท่ากับ 0 ความสามารถของผู้สอบใช้แทนด้วย θ และความน่าจะเป็นของการตอบข้อสอบข้อที่ j ถูกคือ

$$P(Xj = 1 / \theta) = P_j(\theta)$$

ดังนั้นความน่าจะเป็นในการตอบผิดจึงเป็น

$$P(Xj = 0/ \theta) = 1\text{-}P_j(\theta)$$

โมเดลทางทฤษฎีตอบสนองข้อสอบที่เป็นที่นิยมใช้ในปัจจุบันมีความแตกต่างที่สำคัญ คือ จำนวนพารามิเตอร์ที่ใช้ในการบรรยายถึงข้อสอบซึ่งมีอยู่ 3 โมเดล คือ โมเดล 1, 2 และ 3 พารามิเตอร์ ดังนี้

1. โมเดลโลจิสติก 1 พารามิเตอร์ (One-parameter logistic model) โมเดลนี้เบิร์นบอร์ม (Birnbaum) ได้พัฒนาขึ้นในปี ค.ศ. 1968 เป็นโมเดลที่อธิบายข้อสอบด้วยค่าพารามิเตอร์ค่าเดียว คือ ค่าความยากซึ่งโอกาสที่ผู้สอบจะทำข้อสอบได้ถูกต้องหรือไม่นั้นขึ้นอยู่กับระดับความสามารถของผู้สอบกับระดับความยากของข้อสอบ ดังนั้นจึงถือว่าค่าการเดาเป็นศูนย์ (c = 0) และค่าอำนาจจำแนก (a = 1) ของข้อสอบนั้นคงที่ทั้งฉบับเขียนเป็นฟังก์ชั่นได้ดังนี้ (Birnbaurn, 1968, p. 402)

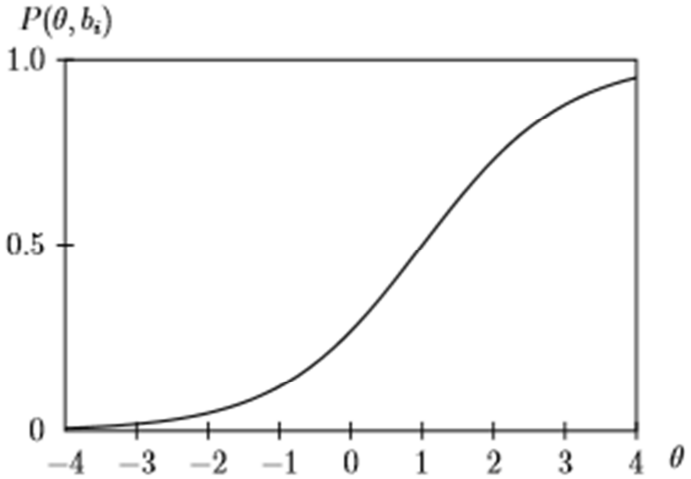
$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{\theta - b_i}}, i = 1,2,3,....n \qquad (4)$$

เมื่อ $P_i(\theta)$ แทน โอกาสที่ผู้มีความสามารถ θ จะทำข้อสอบข้อที่ i ได้ถูกต้อง

θ แทน ระดับความสามารถที่แท้จริงของผู้สอบ

b แทน ค่าความยากของข้อสอบข้อที่ i

e แทน ค่าคงที่ที่มีค่า เท่ากับ 2.7182818



ภาพที่ 2-1 ฟังก์ชันโมเดลการตอบสนองข้อสอบแบบโลจิสติกส์ 1 พารามิเตอร์ (1PL)

2. โมเดลโลจิสติกแบบ 2 พารามิเตอร์ (Two-parameter logistic (2Pl) model)

เบิอร์นบอร์ม (Birnbaum) ได้พัฒนาโมเดลนี้ขึ้นเพื่อมา และกำหนดให้ทุกข้อไม่มีการเดาเกิดขึ้น คือ ค่า c มีค่าเป็น 0 ทุกข้อ กล่าวคือ ผู้สอบที่มีความสามารถต่ำ ไม่มีโอกาสที่จะทำข้อสอบถูกในข้อสอบที่มีค่าความยากสูง ซึ่งเบิอร์นบอร์มได้เสนอรูปแบบของสมการดังนี้ (Birnbaurn, 1968, p. 400)

$$\mathrm{P_i}(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i\theta - b_i}}, i = 1,2,3,....n \qquad (5)$$

เมื่อ $\mathrm{P_i}(\theta)$ แทน โอกาสที่ผู้มีความสามารถ θ จะทำข้อสอบข้อที่ I ได้ถูกต้อง

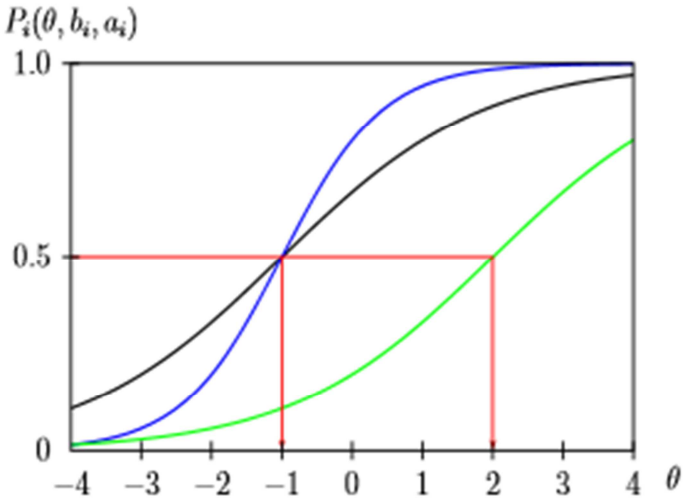
θ แทน ระดับความสามารถที่แท้จริงของผู้สอบ

a_i แทน ค่าอำนาจจำแนกข้อสอบข้อที่ i

b_i แทน ค่าความยากของข้อสอบข้อที่ i

e แทน ค่าคงที่มีค่า เท่ากับ 2.7182818

D แทน ค่าคงที่ซึ่งมีค่า 1.7



ภาพที่ 2-2 ฟังก์ชันโมเดลการตอบสนองข้อสอบแบบโลจิสติกส์ 2 พารามิเตอร์ (2PL)

3. โมเดลโลจิสติกแบบ 3 พารามิเตอร์ (Three-parameter logistic (3PL) model) เป็นโมเดลที่พัฒนามาจาก Two-parameter logistic model เพื่อให้เหมาะสมกับแบบทดสอบที่มีอิทธิพลจากการเดาเข้ามาแฝงอยู่ด้วย และเป็นโค้งลักษณะข้อสอบที่แสดงถึงลักษณะข้อสอบที่มีค่าพารามิเตอร์ของข้อสอบ 3 ตัว ที่เบิร์นบอร์ม ได้เสนอรูปแบบของสมการ ดังนี้

$$P_i(\theta) = c_i(1-c_i)\frac{e^{Da_i(\theta-b_i)}}{1+e^{Da_i\theta-b_i}}, i = 1,2,3,....,n \qquad (6)$$

เมื่อ $P_i(\theta)$ แทน โอกาสที่ผู้มีความสามารถ θ จะทำข้อสอบข้อที่ I ได้ถูกต้อง

θ แทน ระดับความสามารถที่แท้จริงของผู้สอบจะมีค่าอยู่ระหว่าง -3 ถึง +3 และค่า -3 แสดงว่ามีค่าความสามารถต่ำ +3 แสดงว่าค่าความสามารถสูง

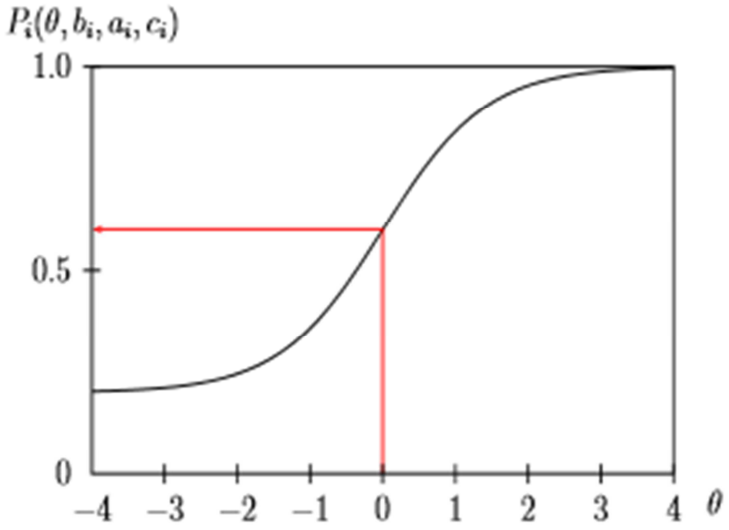
a_i แทน ค่าอำนาจจำแนกข้อสอบข้อที่ i ที่มีค่าสัดส่วนโดยตรงกับค่าความชันของโค้ง ณ จุดเปลี่ยนโค้งที่จุด $\theta = b_i$ โดยทั่วไปมักจะเลือกใช้ช่วงของค่าอำนาจจำแนกของข้อสอบอยู่ระหว่าง 0 ถึง 2

b_i แทน ค่าความยากของข้อสอบข้อที่ i เป็นระดับความสามารถของผู้สอบ ณ.จุดเปลี่ยนโค้งและมีค่าอยู่ระหว่าง -∞ ถึง +∞ แต่ในทางปฏิบัติจะมีค่าอยู่ระหว่าง -2 ถึง +2 ค่า -2 แสดงว่า ข้อสอบง่ายมากและค่า +2 แสดงว่าข้อสอบยาก

c_i แทน ค่าการเดาของข้อสอบข้อที่ i เป็นความน่าจะเป็นหรือโอกาสของคนที่มีความสามารถต่ำจะตอบข้อสอบถูก มีค่าอยู่ระหว่าง 0 ถึง 1 โดยทั่วไปแล้วข้อสอบที่ดีจะต้องมีค่าการเดาต่ำกว่า 0.30

e แทน ค่าคงที่มีค่า เท่ากับ 2.7182818

D แทน ค่าคงที่ซึ่งมีค่า 1.7



ภาพที่ 2-3 ฟังก์ชันโมเดลการตอบสนองข้อสอบแบบโลจิสติกส์ 3 พารามิเตอร์ (3PL)

ลักษณะเฉพาะของโค้งคุณลักษณะของข้อสอบเมื่อระดับความสามารถของผู้สอบ (θ) เปลี่ยนแปลงไปจะมีรูปแบบที่แตกต่างกัน รูปแบบเหล่านี้เปลี่ยนแปลงไปตามฟังก์ชันการตอบสนองรายข้อ (Item response function) หรือที่นิยมเรียกกันทั่วไปว่าโค้งลักษณะเฉพาะของข้อสอบซึ่งโมเดลที่ 3 ที่จะกล่าวต่อไปนี้มีลักษณะเฉพาะดังต่อไปนี้ (สุพัฒน์ สุกมลสันต์, 2538, หน้า 65)

1. เส้นกำกับบน (Upper asymptote) จะมีค่าสูงสุดได้เกือบ 1.0 แต่ไม่ เท่ากับ 1.0 เพราะโอกาสในการตอบข้อสอบได้ถูกต้องไม่เท่ากับ 1.0 หรือ 100% เพราะไม่มีอะไรแน่นอนที่สุด

2. เมื่อผู้สอบมีความสามารถมากขึ้น โอกาสในการตอบข้อสอบได้ถูกต้องมีมากขึ้น เส้นโค้งลักษณะของข้อสอบจะสูงขึ้นเรื่อย ๆ เมื่อเคลื่อนจากซ้ายไปขวาและจุดโค้งมีความชันมากที่สุด คือ จุดที่ดีที่สุดในการแบ่งผู้สอบที่มีความสามารถแตกต่างกัน (จุดนี้เรียกว่าจุด a ซึ่งเป็นค่าอำนาจจำแนก)

3. เส้นกำกับล่าง (Lower asymptote) จะมีค่าได้น้อยที่สุดเกือบ 0.0 แต่ไม่เท่า 0.0 เพราะการทดสอบย่อมมีการเดาบ้างการที่ไม่มีการเดาเลยย่อมเป็นไปไม่ได้

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามแนวคิดทฤษฎีการตอบข้อสอบ

จากนิยามดังกล่าว การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามแนวทฤษฎีการตอบข้อสอบสามารถกระทำได้โดยเปรียบเทียบฟังก์ชันการตอบสนองของผู้สอบกลุ่มย่อยที่ต้องการศึกษาซึ่งสามารถแบ่งออกเป็น 2 แนวทาง ดังนี้ (Hambleton, Swaminathan & Roger, 1991, p. 110)

1. การเปรียบเทียบพารามิเตอร์ของข้อสอบระหว่าง 2 กลุ่ม ว่ามีความแตกต่างกันหรือไม่ มีผู้ศึกษาค้นคว้ามากพอสมควร ได้แก่ งานวิจัยของคิม และโคเฮน (Kim & Cohen, 1994) ลอร์ด (Lord, 1977, p. 1980) ธิสเซน, สเติร์นเบอร์ก และวายเนอร์ (Thissen, Steinberg & Wainer, 1988-1993) และไรท์ และสโตน (Wright & Syone, p. 1979) อย่างไรก็ตาม ในกลุ่มนี้วิธีที่ได้รับความนิยมก็คือ วิธีที่อาศัยแนวคิดลอร์ด (Lord's χ^2) ในระยะหลังมีวิธีที่น่าสนใจอีกวิธี คือ วิธีทดสอบอัตราส่วนไลลิฮู้ด (Likelihood ratio test) ซึ่ง ธิสเซน สเติร์นเบอร์ก, วายเนอร์ และโคเฮน ก็ได้ศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีนี้ รวมทั้งเปรียบเทียบผลการตรวจสอบกับวิธีอื่นด้วย เช่น Lord's χ^2 และ Raju's Area Measures เป็นต้น

สำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามแนวคิดของลอร์ด สามารถใช้ได้กับแบบจำลองโลจิสติก ชนิด 1 พารามิเตอร์ 2 พารามิเตอร์ และ 3 พารามิเตอร์ มีขั้นตอนสรุปได้ดังนี้

1. ทดสอบข้อมูล (Test model) กับแบบจำลองที่ศึกษา
2. ประมาณค่าพารามิเตอร์ของข้อสอบ (a, b, c)
3. แปลงหน่วยของค่าพารามิเตอร์ให้อยู่ในหน่วยเดียวกัน หรือหน่วยร่วม (Common scale)
4. นำค่าพารามิเตอร์ของข้อสอบมาเปรียบเทียบกัน โดยใช้สถิติทดสอบ χ^2 โดยมีสมมติฐานในการทดสอบ คือ H: $b_1 = b_2$, $a_1 = a_2$, $c_1 = c_2$

สถิติทดสอบ (Hambleton, Swaminathan & Roger, 1991, p. 111)

$$\chi^2 = (a_{diff}\, b_{diff}\, c_{diff})' \sum^{-1} (a_{diff}\, b_{diff}\, c_{diff}) \qquad (7)$$

เมื่อ $a_{diff} = a_2 - a_1$

$b_{diff} = b_2 - b_1$

$c_{diff} = c_2 - c_1$

และ $\sum$ แทน เมทริกซ์ Variance และ Covariance ของความแตกต่างระหว่างพารามิเตอร์ข้อสอบ หากศึกษาเป็นแบบจำลอง 1 พารามิเตอร์ สถิติทดสอบ เป็นดังนี้

$$x^2 = \frac{b_{diff}{}^2}{\text{var}(b_i) + \text{var}(b_2)} \quad (8)$$

2.1 ความแตกต่างระหว่างฟังก์ชันการตอบสนองข้อสอบจาก 2 กลุ่ม เพื่อพิจารณาขนาดของการทำหน้าที่ต่างกันของข้อสอบ มีผู้ศึกษาค้นคว้าไว้ เช่น งานวิจัยของ คิม และโคเฮน (Kim & Cohen, 1991) ลินน์ เลวิน ฮาสติงส์ และวาร์ลดรอบ (Linn, Levien, Hasting &Wardrop, 1981) รัดเนอร์ (Rudner, 1977) และวายเนอร์ (Wainer, 1993) เป็นต้น ในกลุ่มนี้ระยะหลังนิยมใช้แนวคิดของราชู และแนวคิดของ คิม และโคเฮน ซึ่งราชูเสนอให้ใช้การวัดขนาดพื้นที่บนช่วงความสามารถแบบเปิด (Open interval) บางครั้งเรียกว่า “การวัดพื้นที่สมบูรณ์” (Exact area measures) ส่วนคิม และโคเฮน เสนอให้วัดขนาดพื้นที่บนช่วงความสามารถแบบปิด (Close 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 $$P^*(\theta) = 1 + \exp\left[Da(\theta - b)\right]^{-1} \quad (9)$$

a, b และ c เป็นพารามิเตอร์ที่แสดงลักษณะข้อสอบ และ D เป็นตัวคงที่กำหนดให้เท่ากับ 1.7 การคำนวณขนาดพื้นที่ระหว่างความสามารถที่กำหนดไว้ 2 จุด บนมาตราความสามารถหาได้จาก

$$S(\theta_1 - \theta_2) = \int_{\theta_1}^{\theta_2} P(\theta)d(\theta) = c(\theta_1 - \theta_2) + (1-c)(Da)^{-1}\text{In}\frac{\{1-\exp\left[Da(\theta_2 - b\right]}{\{1+\exp\left[Da(\theta_2 - b)\right]} \quad (10)$$

ในการศึกษาการทำหน้าที่ต่างกันของข้อสอบ จะมีพารามิเตอร์ของ 2 กลุ่ม คือ กลุ่มอ้างอิง (Reference group) ได้แก่ a_g b_g และ c_g และกลุ่มเปรียบเทียบ (Focal group) ได้แก่ a_R b_R และ C_R และกลุ่มเปรียบเทียบ (Focal group) ได้แก่ a_F b_F และ c_F พื้นที่ชนิดมีเครื่องหมาย (Closed-intervak signed area-CSA) หาได้จาก

$$CSA = \int_{\theta_1}^{\theta_2}\left[P_R(\theta) - P_F(\theta)\right]d\theta = S_R(\theta_1, \theta_2) - S_F(\theta_{1,}\theta_2) \quad (11)$$

และพื้นที่ชนิดไม่มีเครื่องหมาย (Closed-interval unsigned area-CUA) หาได้จาก

$$CUA = \int_{\theta_1}^{\theta_2}|P_R(\theta) - P_F(\theta)|\,d\theta \quad (12)$$

จากการศึกษา คิม และโคเฮน (Kim & Cohen, 1991) ได้ศึกษาเปรียบเทียบความแตกต่างของพื้นที่แบบไม่มีเครื่องหมาย (Unsigned area) และแบบมีเครื่องหมาย (Sign area)

พบว่า พื้นที่แบบมีเครื่องหมายมีความคลาดเคลื่อนต่ำกว่าแบบไม่มีเครื่องหมาย และการวัดพื้นที่ช่วงเปิดและแบบช่วงปิดจะให้ผลสอดคล้องกันค่อนข้างมาก

คุณลักษณะของทฤษฎีตอบสนองข้อสอบ

ตามหลักทฤษฎีการตอบสนองข้อสอบนั้น มีหลักคิดที่ว่าความสามารถของผู้ตอบสนองต่อข้อสอบสามารถอธิบาย หรือทำนายได้ด้วยคุณลักษณะของผู้สอบ (Examinee characteristics) ซึ่งอาจเป็นคุณลักษณะแฝงภายในตัวบุคคล (Trait) หรือความสามารถ (Abilities) ซึ่งการประมาณค่าคะแนน สำหรับผู้สอบในความสามารถแฝงอยู่ภายในจะเรียกว่า คะแนนความสามารถ (Ability score) และใช้คะแนนเหล่านี้ มาอธิบายความสามารถในการทำข้อสอบ ดังนั้นทฤษฎีการตอบสนองข้อสอบจึงเป็นทฤษฎีที่มุ่งอธิบายพฤติกรรมการตอบสนองข้อสอบของผู้สอบที่ถูกกำหนดโดยคุณลักษณะภายในหรือความสามารถที่แฝงอยู่ภายในตัวบุคคล และศึกษาความสัมพันธ์ระหว่างการตอบสนองข้อสอบของผู้สอบกับระดับความสามารถที่มีอยู่ด้วยโมเดลที่เป็นฟังก์ชันทางคณิตศาสตร์ที่เรียกว่า โค้งลักษณะข้อสอบ ซึ่งมีคุณสมบัติสรุปได้ ดังนี้

1. ความสามารถของผู้สอบในการทำแบบสอบถามสามารถอธิบายหรือทำนายในรูปแบบของคุณลักษณะที่แฝงอยู่ภายในตัวผู้สอบ

2. โมเดลการตอบสนองข้อสอบอธิบายควาสัมพันธ์ระหว่างความสามารถในการตอบข้อสอบที่สามารถสังเกตได้โดยตรงว่าผิดหรือถูกกับความสามารถที่แฝงอยู่ภายในตัวผู้สอบ

3. ความสำเร็จของโมเดลการตอบสนองข้อสอบจะให้ค่าเฉลี่ยของการประมาณค่าความสามารถของผู้สอบ

4. การประมาณค่าความสามารถที่แท้จริงของผู้สอบจะประมาณค่าจากความสามารถของผู้สอบที่ตอบสนองข้อสอบ

คุณสมบัติของความไม่แปรเปลี่ยนของค่าพารามิเตอร์

เมื่อโมเดลการตอบสนองข้อสอบมีความสอดคล้อง (Fit) กับข้อมูลเชิงประจักษ์จะทำให้เกิดคุณสมบัติความไม่แปรเปลี่ยน (Invariance) ของค่าพารามิเตอร์ข้อสอบ และพารามิเตอร์ความสามารถของผู้สอบในสองลักษณะดังนี้ (ศิริชัย กาญจนวาสี, 2550)

1. ความไม่แปรเปลี่ยนของพารามิเตอร์ข้อสอบ (Item invariance)

เป็นคุณสมบัติที่ว่าค่าพารามิเตอร์ของข้อสอบจะมีค่าคงที่ไม่เปลี่ยนแปลงถึงแม้จะเปลี่ยนกลุ่มผู้สอบนั้น คือ พารามิเตอร์ a, b และ c ในโค้งลักษณะข้อสอบเดียวกันจะคงที่สำหรับทุกกลุ่มความสามารถผู้สอบ แสดงว่าโค้งลักษณะข้อสอบมีความคงที่ข้ามกลุ่มผู้สอบ

2. ความไม่แปรเปลี่ยนของพารามิเตอร์ความสามารถของผู้สอบ (Item invariance)

ความไม่แปรเปลี่ยนมีคุณสมบัติ คือ ค่าพารามิเตอร์ความสามารถของผู้สอบจะมีค่าคงที่ไม่เปลี่ยนแปลงถึงแม้จะมีการเปลี่ยนแปลงชุดของแบบสอบ นั่นคือ หากนำแบบสอบที่มุ่งวัดคุณลักษณะเดียวกัน จำนวน 2 ชุด ค่าความสามารถของผู้สอบที่ประมาณค่าได้จากแบบทดสอบทั้งสองชุดจะมีค่าที่แตกต่างกันไม่เกินค่าความคาดเคลื่อนมาตรฐานของการประมาณค่า (SE_E) แสดงว่าการประมาณค่าความสามารถมีความคงที่ข้ามชุดของข้อสอบ

ในประเด็นของความไม่แปรเปลี่ยนของค่าพารามิเตอร์ เบเกอร์ (Baker, 1997, p. 170) กล่าวว่าต้องขึ้นอยู่กับเงื่อนไข 2 ประการ ประการแรก คือ ความสามารถ (Ability) ที่กล่าวถึงต้องสามารถนิยามได้ชัดเจนและวัดได้ด้วยข้อสอบ อีกประการหนึ่งความสามารถที่วัดนั้นจะต้องมีความคงที่ภายในช่วงระยะเวลาหนึ่ง นอกจากนี้ ลอร์ด (Lord, 1980, p. 36) ยังได้อธิบายเพิ่มเติมว่า การไม่แปรเปลี่ยนของค่าพารามิเตอร์นั้น มิได้หมายความว่า ค่าพารามิเตอร์ที่ได้จากการประมาณค่าโดยใช้กลุ่มตัวอย่างที่แตกต่างกันจะมีค่าเท่ากันเสมอ ค่าพารามิเตอร์ของข้อสอบจะเท่ากันหรือไม่ขึ้นอยู่กับเงื่อนไขในการทดสอบบางประการ เช่น ถ้าเลือกมาตรวัดที่มีจุดเริ่มต้นเดียวกัน และหน่วยในการวัดเดียวกัน ค่าพารามิเตอร์ที่ประมาณได้จากกลุ่มตัวอย่างที่แตกต่างกัน จะมีค่าเท่ากัน ในทางตรงข้าม ถ้าหากเลือกมาตรวัดที่มีจุดเริ่มต้น และมีหน่วยในการวัดแตกต่างกันแล้ว การไม่แปรเปลี่ยนของค่าพารามิเตอร์จะหมายความว่า ค่าพารามิเตอร์ที่ประมาณได้จากกลุ่มตัวอย่างที่ต่างกัน ของข้อสอบชุดหนึ่งจะมีความสัมพันธ์กันในเชิงเส้นตรง

จุดเด่นของการวิเคราะห์ด้วยทฤษฎีการตอบสนองข้อสอบ

การวิเคราะห์ข้อสอบด้วยทฤษฎีการตอบสนองข้อสอบให้ผลการวิเคราะห์ทั้งค่าพารามิเตอร์ของผู้สอบ ซึ่งได้แก่ ค่าความสามารถที่แท้จริงของผู้สอบ (θ) และค่าพารามิเตอร์ของข้อสอบซึ่ง ได้แก่ ค่าความยาก (b) ค่าอำนาจจำแนก (a) และค่าโอกาสในการเดาคำตอบที่ถูกต้อง (c) การวิเคราะห์ด้วยทฤษฎีนี้มีจุดเด่นหลายประการสรุปเป็นประเด็นสำคัญได้ ดังนี้

1. สารสนเทศของการวิเคราะห์

การวิเคราะห์ด้วยทฤษฎีการตอบข้อสอบสามารถวิเคราะห์ค่าพารามิเตอร์ของผู้สอบเป็นรายบุคคลโดยการประมาณค่าความสามารถที่แท้จริงของผู้สอบ (θ) นอกจากนี้ยังสามารถวิเคราะห์ค่าพารามิเตอร์ของการสอบเป็นรายข้อได้อีกด้วยตามโมเดลการทดสอบ โดยโมเดลการตอบสนองข้อสอบแบบ 1 พารามิเตอร์ (1 PL) จะเป็นโมเดลที่มีพารามิเตอร์ของข้อสอบเพียงค่าเดียว คือ ค่าความยากของข้อสอบ (b_i) ถ้าโมเดลการตอบสนองข้อสอบแบบ 2 พารามิเตอร์ (2PL) จะเป็นโมเดลที่มีค่าพารามิเตอร์ของข้อสอบสองค่า คือ ค่าความยากของข้อสอบ (b_i) และค่าอำนาจจำแนก (a) และถ้าโมเดลการตอบสนองข้อสอบแบบ 3 พารามิเตอร์ (3 PL) จะเป็นโมเดลที่มีค่าพารามิเตอร์

ของข้อสอบสามค่า คือ ค่าความยากของข้อสอบ (b_i) และค่าอำนาจจำแนก (a) และค่าโอกาสในการเดาคำตอบที่ถูกต้อง (c)

2. ความไม่แปรเปลี่ยนของค่าพารามิเตอร์ (Invariance)

ผลการวิเคราะห์ตามทฤษฎีการตอบสนองข้อสอบมีจุดเด่นที่สำคัญ เมื่อเปรียบเทียบกับทฤษฎีการทดสอบแบบดั้งเดิม (CTT) คือ ค่าพารามิเตอร์ของผู้สอบและค่าพารามิเตอร์ของข้อสอบมีความคงที่ไม่แปรเปลี่ยน โดยค่าพารามิเตอร์ของข้อสอบจะเป็นอิสระจากกลุ่มผู้สอบที่ใช้ในการประมาณค่านั้น คือ ทั้งค่าความยาก (b) ค่าอำนาจจำแนก (a) และค่าโอกาสในการเดาคำตอบที่ถูกต้อง (c) จะมีค่าไม่แปรเปลี่ยนแปลงไปตามกลุ่มผู้สอบไม่ว่าจะนำแบบสอบไปใช้กับกลุ่มผู้สอบกลุ่มใดก็ตาม ค่าพารามิเตอร์ก็จะไม่แปรเปลี่ยน สำหรับค่าพารามิเตอร์ของผู้สอบก็เช่นเดียวกันจะเป็นอิสระจากข้อสอบที่ใช้ในการประมาณค่า โดยจะไม่เปลี่ยนแปลงไปตามค่าพารามิเตอร์ของข้อสอบ ดังนั้นจึงทำให้สามารถคำนวณค่าความสามารถที่แท้จริงของผู้สอบได้โดยไม่ขึ้นกับสถานการณ์การทดสอบและกลุ่มของผู้สอบ

3. ความเป็นอิสระของข้อสอบ (Item-free)

การเปรียบเทียบความสามารถของผู้สอบจะไม่ขึ้นอยู่กับข้อคำถามในแบบสอบ โดยในการประมาณค่าความสามารถของผู้สอบจะใช้ชุดข้อคำถามที่ต่างกันก็สามารถนำค่าความสามารถมาเปรียบเทียบกันได้ ทั้งนี้ข้อสอบที่นำมาใช้ต้องได้รับการคัดเลือกจากคลังข้อสอบขนาดใหญ่ ที่ข้อสอบแต่ละข้อได้ระบุคุณลักษณะที่วัดเอาไว้ เนื่องจากค่าความสามารถที่แท้จริงที่ประมาณค่าได้เป็นคะแนนโลจิส (Logit) ซึ่งอยู่ในมาตรวัดอันเดียวกัน

4. การวัดคุณภาพตามแนวทฤษฎีการตอบสนองข้อสอบนั้น สามารถรายงานได้ทั้งรายข้อและทั้งฉบับ เนื่องจากสามารถวิเคราะห์คุณภาพของข้อสอบได้ทั้งค่าฟังก์ชันสารสนเทศของข้อสอบและค่าฟังก์ชันสารสนเทศของแบบสอบ ซึ่งค่าทั้งสองเป็นดัชนีที่แสดงถึงความถูกต้องแม่นยำในการประมาณค่าความสามารถนำมาใช้แปลความหมายได้เหมือนค่าความเที่ยง และค่าความคลาดเคลื่อนมาตรฐานตามแนวทฤษฎีการทดสอบดั้งเดิมได้

ตอนที่ 3 การทำหน้าที่ต่างกันของข้อสอบ (Differential item functioning)

ความหมายของการทำหน้าที่ต่างกันของข้อสอบ

การทำหน้าต่างกันของข้อสอบ หมายถึง โอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกัน (การวัดความสามารถ) หรือโอกาสในการตอบข้อสอบในทางบวกแตกต่างกัน (การวัดเจตคติ) เมื่อผู้สอบที่มีคุณลักษณะของการวัดในปริมาณเท่ากัน แต่มาจากกลุ่มประชากรย่อยที่แตกต่างกัน (Hulin, Drasgow & Komocar, 1982)

การทำหน้าที่ต่างกันของข้อสอบ หมายถึง ความคลาดเคลื่อนอย่างเป็นระบบในการวัด (Systematics error) หรือความไม่ตรง (Invalidity) ชนิดหนึ่งของการวัด ซึ่งทำให้ผลของการวัดบิดเบือนไปสำหรับสมาชิกของกลุ่มผู้สอบบางกลุ่มโดยเฉพาะ ทำให้สมาชิกบางกลุ่มได้ประโยชน์หรือเสียประโยชน์ เช่น เข้าข้างกลุ่มเพศชายมากกว่าเพศหญิง เข้าข้างคนผิวขาวมากกว่าคนผิวดำ หรือเข้าข้างภูมิภาคใดมากกว่าภูมิภาคหนึ่ง (Camilli & Shepard, 1994)

การทำหน้าต่างกันของข้อสอบ หมายถึง การที่ผู้เข้าสอบซึ่งเป็นกลุ่มย่อย (Subgroup) และมีความสามารถเท่าเทียมกับกลุ่มประชากรหลักได้รับการตอบสนองในการตอบถูกจากข้อสอบหรือแบบสอบแตกต่างจากผู้เข้าสอบในกลุ่มประชากรหลัก การทำหน้าที่ต่างกันของข้อสอบจะเกิดขึ้น เมื่อประชากรมีความแตกต่างกันด้านเพศ เชื้อชาติหรือวัฒนธรรม (Ahderson & Demar, 2002)

การทำหน้าต่างกันของข้อสอบ หมายถึง การที่ข้อสอบทำให้ผู้สอบจากกลุ่มต่างกันที่มีความสามารถ หรือคุณลักษณะที่มุ่งวัดเท่ากัน มีโอกาสในการตอบข้อสอบได้ถูกต้องต่างกัน หรือมีฟังก์ชันการตอบสนองข้อสอบแตกต่างกัน การทำหน้าที่ต่างกันของข้อสอบเกิดขึ้นเมื่อนำข้อสอบไปทดสอบกับกลุ่มผู้สอบย่อยต่างกัน ที่มีความสามารถหลัก (Primary ability) ระดับเดียวกันหรือมีคุณลักษณะแฝง (Latent trait) ที่ต้องการวัดเท่ากันแต่มีความสามารถรอง (Secondary ability) แตกต่างกัน (ศิริชัย กาญจนวาสี, 2545)

มีผู้ให้ความหมายของคำว่า “การทำหน้าที่ต่างกันของข้อสอบ (Differential Item functioning: DIF” ไว้หลายความหมาย แต่ความหมายที่เป็นที่ยอมรับกันอย่างกว้างขวาง คือ การทำหน้าที่ต่างกันของข้อสอบอยู่ภายใต้เงื่อนไขผู้เข้าสอบมีความสามารถเท่ากันแต่มาจากกลุ่มผู้เข้าสอบย่อยต่างกัน มีความน่าจะเป็นในการตอบข้อสอบข้อนั้นถูกต้องไม่เท่ากัน (Hamble & Others, 1993)

Zumbo and Hubley (1998) และอ้างอิงตามของ Camilli and Shepard (1994) และ Clauser and Mazor (1998) ได้ให้คำนิยามของความลำเอียงของข้อสอบ (DIF) ไว้ดังนี้

การวิเคราะห์ข้อสอบ (Item analysis) หมายถึง เทคนิคทางสถิติในการตรวจสอบประสิทธิภาพของข้อสอบแต่ละข้อ ความสำคัญ คือทำให้เกิดการพัฒนาแบบทดสอบแต่ละข้อเมื่อเราทราบกระบวนการวัด

ผลกระทบของข้อสอบ (Item impact) หมายถึง ผลกระทบของข้อสอบเกิดขึ้นเมื่อผู้สอบมาจากกลุ่มที่แตกต่างกันและมีโอกาสในการตอบข้อสอบ (หรือการรับรอง) ถูกต้องได้แตกต่างกันเนื่องจากมีความแตกต่างที่แท้จริงระหว่างกลุ่มภายใต้ความสามารถพื้นฐานที่วัดโดยข้อสอบ

การทำหน้าที่ต่างกันของข้อสอบ (DIF) เกิดขึ้นเมื่อผู้สอบมาจากกลุ่มที่แตกต่างกันที่มีความน่าจะเป็นของการมีประสิทธิผลที่ต่างกันในการทำข้อสอบ (หรือการรับรอง) ภายหลังทำการจับคู่ภายใต้ความสามารถพื้นฐานตามวัตถุประสงค์ในการวัดข้อสอบ

ความลำเอียงของข้อสอบ (Item bias) ความลำเอียงของข้อสอบเกิดขึ้นเมื่อผู้สอบของกลุ่มหนึ่งมีโอกาสน้อยกว่ากลุ่มผู้สอบกลุ่มอื่นในการตอบข้อสอบได้ถูกต้อง เนื่องจากมีลักษณะเฉพาะของข้อสอบที่ทดสอบหรือสถานการณ์การทดสอบที่ไม่เกี่ยวข้องกับวัตถุประสงค์ของการทดสอบ สามารถระบุการเกิด DIF แต่ไม่เพียงพอสำหรับการระบุความลำเอียงของข้อสอบ (Item bias)

มีสองวิธีในการตรวจสอบความลำเอียงของการวัด คือ (ก) การตัดสินใจ และ (ข) วิธีการทางสถิติ วิธีการตัดสินใจขึ้นอยู่กับความเห็นของผู้เชี่ยวชาญหนึ่งคนหรือหลายคนในการเลือกข้อสอบที่มีความลำเอียง แทนที่จะต้องพึ่งพาการตัดสินของผู้เชี่ยวชาญ (เช่น วิเคราะห์เนื้อหา) แต่เพียงผู้เดียว ควรมีการใช้เทคนิคทางสถิติเพื่อตรวจสอบความลำเอียง และอธิบายข้อสอบที่มีความลำเอียงได้

ทฤษฎีการทดสอบสมัยใหม่โดยทั่วไปตาม Hambleton, Swaminathan and Rogers (1991) มีพื้นฐานอยู่บนสมมติฐานสองข้อ ดังนี้

1. ประสิทธิภาพของผู้สอบในการทดสอบสามารถอธิบายหรือทำนายได้จากปัจจัยที่เรียกว่า “คุณลักษณะเฉพาะ คุณลักษณะแฝง หรือความสามารถ” ซึ่งเรียกว่าความแปรปรวนต่อเนื่อง

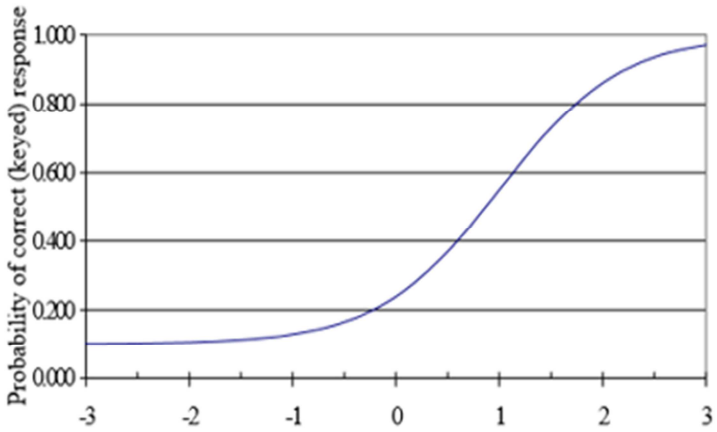
2. ความสัมพันธ์ระหว่างประสิทธิภาพของผู้สอบและความแปรปรวนต่อเนื่องจะอยู่ภายใต้ประสิทธิภาพของข้อสอบซึ่งสามารถอธิบายได้ว่าเป็นโค้งลักษณะข้อสอบ ICC หรือฟังก์ชันการตอบสนองต่อข้อสอบ (IRF)

การศึกษาความเป็นไปได้ในการวัดโดยการตรวจสอบความสัมพันธ์ภายนอกหรือความสัมพันธ์ภายใน ก่อนหน้านี้มักมุ่งเน้นไปที่การศึกษาเชิงทำนายเพื่อตรวจสอบความสัมพันธ์ระดับเกณฑ์ เช่น การคัดเลือกบุคลากร หลักฐานภายนอกของการวัดความลำเอียง (Bias) ถูกนำเสนอเมื่อความสัมพันธ์ระหว่างคะแนนและเกณฑ์แตกต่างกันไปสำหรับกลุ่มต่าง ๆ หลักฐานมาจากความสัมพันธ์ภายในของข้อสอบแต่ละข้อ DIF เป็นเรื่องของความสัมพันธ์ภายในของข้อสอบอย่างชัดเจน นอกจากนี้ยังมุ่งเน้นไปที่การใช้คะแนนรวม (Composite score) เป็นตัวบ่งชี้ของตัวแปรแฝง (ไม่สามารถสังเกตได้)

การแสดงของตัวแปรแฝง (Latent variable) และตัวแปรที่ไม่ได้รับการสังเกต (Unobservable) คือ ตัวแปรที่เรากำลังพยายามหาตัวบ่งชี้ (เช่น ข้อสอบ) เช่น สติปัญญา ความสามารถ ความสามารถทางการพูด สำหรับตัวแปรแฝง (Latent variable) ตัวนี้ (เรียกว่าความแปรปรวนต่อเนื่องซึ่งขึ้นอยู่กับความแตกต่างระหว่างบุคคล

ส่วนของรายละเอียดเกี่ยวกับโค้งลักษณะข้อสอบจะแสดงด้วยภาพต่าง ๆ เหล่านี้ เช่น ภาพที่ 2-4 แสดงตัวอย่างของพารามิเตอร์ของโค้งลักษณะข้อสอบ ICC โดยแกนนอนเป็นความแปรฝันต่อเนื่องของตัวแปรแฝง ระดับของตัวแปรแฝงอยู่ใน Z-Sores ข้อสอบแสดงไว้ใน ภาพที่ 2-4 จะมีค่าพารามิเตอร์อำนาจจำแนก (Discrimination parameter: ความลาดชัน (Slope) มีค่าเท่ากับ 2.0 ค่าพารามิเตอร์ความยาก (Difficulty parameter): Threshold มีค่าเท่ากับ 0.60 และค่าพารามิเตอร์การเดา (Guess parameter) มีค่าเท่ากับ 0.10

สังเกตว่าพารามิเตอร์อำนาจจำแนกข้อสอบ (Discrimaination parameter) เส้นโค้งจะเพิ่มขึ้นจากค่าต่ำสุดของค่า C จาก 0.10 ถึง 1.0 ถ้าเส้นโค้งมีลักษณะแบนราบแสดงว่าข้อสอบไม่สามารถจำแนกระหว่างคะแนนจากการวัดออกเป็นกลุ่มสูง, ปานกลาง และต่ำ ค่าอำนาจจำแนกของข้อสอบที่มีค่าตั้งแต่ 1.0 ขึ้นไป ถือว่าดีมาก ระดับของค่าพารามิเตอร์ คือ ค่าของตัวแปรแฝงที่มีความแปรผันต่อเนื่อง ซึ่งจะเห็นว่าเส้นโค้งจะอยู่กึ่งกลางระหว่างค่าต่ำสุด c และ 1.0 ดังนั้น การวัดผลสัมฤทธิ์หรือวัดทัศนคติ คือ สัญญลักษณ์หรือเครื่องหมายของความยากของข้อสอบ (Item difficulty) ข้อสอบที่มีค่าความยากน้อยกว่า -1.0 ชี้ให้เห็นว่าข้อสอบค่อนข้างง่าย ในขณะที่ข้อสอบที่มีค่าความยากของข้อสอบมากกว่า 1.0 แสดงให้เห็นว่าข้อสอบค่อนข้างยาก



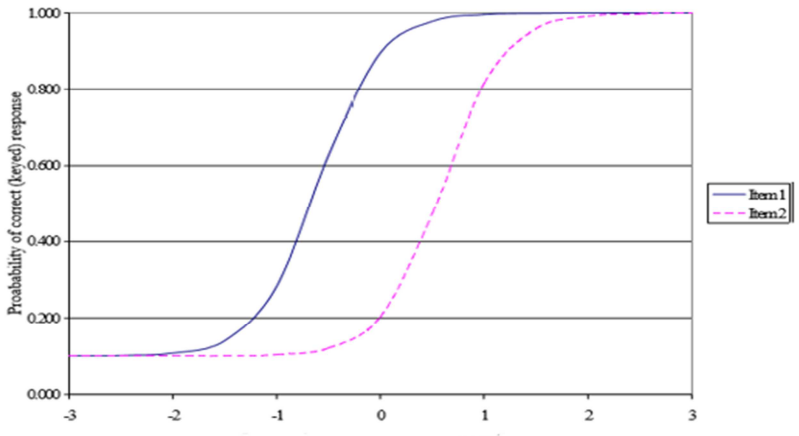
ภาพที่ 2-4 โค้งลักษณะข้อสอบโลจิสติก (Logistic item characterlistic curve)

โค้งลักษณะข้อสอบ ICCs ที่แสดงในภาพที่ 2-4 แสดงให้เห็นถึงข้อสอบสองข้อที่มีความลาดชัน (Slope) เท่ากัน นั่นคือ อำนาจจำแนกมีค่าเท่ากันระหว่างผู้ตอบข้อสอบ แต่มีตำแหน่ง

ของความแปรผันต่อเนื่องที่แตกต่างกัน ซึ่งจะต้องมีตัวแปรแฝงจำนวนมากซึ่งแสดงโดยเส้นประมากกว่าเส้นทึบ โดยเส้นประจะอยู่ด้านขวา ของเส้นทึบ

โดยสรุปแล้วโค้งลักษณะข้อสอบ ICC แสดงเห็นถึงความน่าจะเป็นของการถดถอยเชิงเส้น (Non-linear regression of the probability) ของความน่าจะเป็นของการตอบข้อสอบถูกต้องตามกรอบแนวคิดของความแปรผันต่อเนื่องของตัวแปรแฝงที่สนใจ ถ้าเราทำการทดสอบทัศนคติ ยกตัวอย่าง เช่น จุดตัดของ y คือ ความควรจะเป็น (Likelihood) ของบางคนที่จะตอบข้อสอบถูกต้อง ความลาดชัน (Slope) จะเป็นเส้นโค้งที่บ่งชี้ว่าข้อสอบเหล่านี้มีความแตกต่างกันอย่างไร ระหว่างระดับความสามารถ (โค้งแบน หมายถึง การไม่คำนึงถึงระดับความสามารถ) แต่ละบุคคลจะมีความควรจะเป็นที่เหมือนกัน (Likelihood) ในการตอบข้อสอบถูกต้อง

ตัวบ่งชี้ของจุดตัด (Threshold) ที่มีความแปรผันต่อเนื่อง (แกน X) มีความน่าจะเป็นที่จะตอบข้อสอบได้ถูกต้องจะอยู่กึ่งกลางระหว่างค่าต่ำสุด c และ 1.0 สำหรับข้อสอบที่พารามิเตอร์การเดา (Guessing parameter) เป็น 0 ค่าที่กึ่งกลางนี้จะแสดงถึงค่าของตัวแปรแฝง (Latent variable) ที่ความควรจะเป็น (likelihood) ของการตอบข้อสอบถูกต้องมากกว่า 0.50

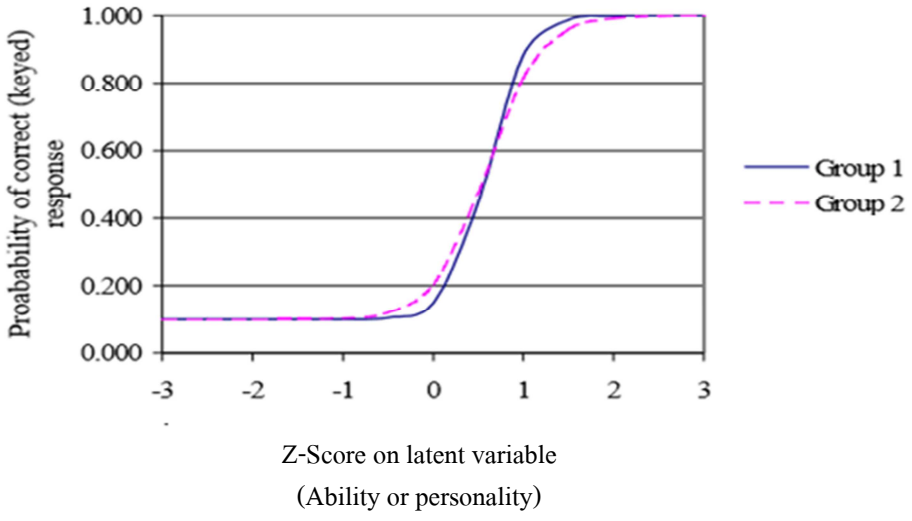


ภาพที่ 2-5 เส้นโค้งลักษณะข้อสอบที่มีความแตกต่างของตำแหน่งที่เป็นตัวแปรต่อเนื่อง

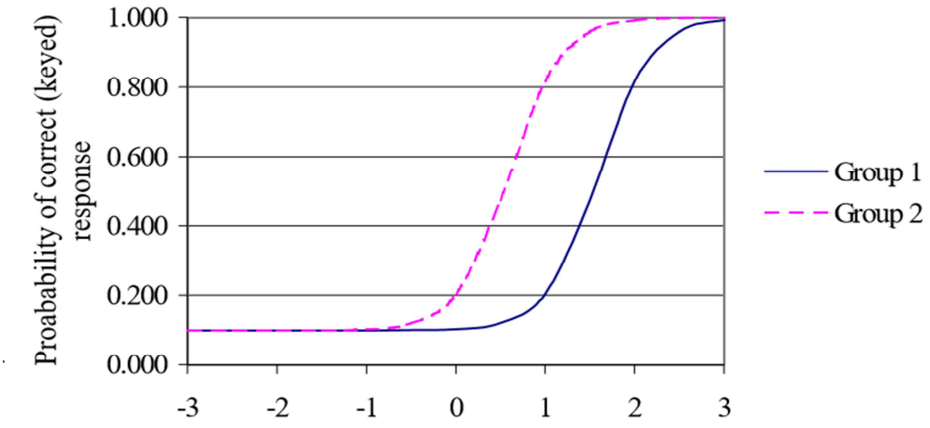
เนื่องจากวิธีการทดสอบทฤษฎีคลาสสิกแบบดั้งเดิม (Traditional classical test theory) ได้มีการสรุปสาระสำคัญทางสถิติว่า ค่าเฉลี่ยเป็นความแปรผันต่อเนื่อง ยกตัวอย่าง เช่น ค่าสหสัมพันธ์รวมของข้อสอบ Item total correlations (สถิติอำนาจจำแนกแบบดั้งเดิม) หรือค่าสัมประสิทธิ์ความน่าเชื่อถือเป็นตัวเลขหนึ่งส่วนโดยไม่คำนึงถึงระดับของแต่ละรูปแบบ นั่นคือ

ค่าสัมประสิทธิ์ (Reliability coefficients) จะมีจำนวนเท่ากันของระดับความแปรผันของผู้ตอบแต่ละคนค่าสัมประสิทธิ์อัลฟา (Coefficient alpha) จะมีจำนวนที่เท่ากันในผู้ตอบข้อสอบที่มีค่าเบี่ยงเบนมาตรฐาน (Standard deviations) เท่ากับ 3 หรือต่ำกว่า 3 หรือมากกว่า 3

ดังนั้นข้อสรุปของการวัดอธิบายด้วยตัวอย่างทั้งหมดโดยไม่สนใจว่าคุณสมบัติทางจิตวิทยาซึ่งการวัดอาจแปรผันภายในกลุ่มตัวอย่าง ทฤษฎีการทดสอบสมัยใหม่ (Modern test theory) สร้างขึ้นจากวิธีการทดสอบแบบคลาสสิก และคำนึงถึงรูปแบบตามตัวอย่างนี้และความต่อเนื่องของความแปรผัน เทคนิคทางด้านจิตวิทยามีมากขึ้นในจิตวิทยาสมัยใหม่ที่เกี่ยวกับ DIF ยังคำนึงถึงความแปรผันต่อเนื่อง แนวคิด DIF ถูกประเมินโดยการเปรียบเทียบโค้งลักษณะข้อสอบของกลุ่มที่แตกต่างกันในการทำข้อสอบ ซึ่งสามารถจิตนาการได้ว่ามีข้อสอบเดียวกันแต่แยกออกเป็นแต่ละกลุ่มที่นักวิจัยต้องการประเมิน (เช่น เพศ) ถ้าโค้งลักษณะข้อสอบ ICCs เหมือนกันในแต่ละกลุ่มหรือใกล้เคียงกันมาก อาจกล่าวได้ว่าข้อสอบนี้ไม่แสดง DIF อย่างไรก็ตามหากโค้งลักษณะข้อสอบ ICCs มีความแตกต่างกันอย่างมีนัยสำคัญในแต่ละกลุ่มแล้วข้อสอบจะแสดงว่ามี DIF เกิดขึ้น ในบริบทส่วนใหญ่ DIF ถือว่าเป็นความแตกต่างในการจัดตำแหน่ง (เช่น ความยาก (Difficulty) และจุดตัด (Threshold) ของเส้นโค้งลักษณะข้อสอบ ICCS สองเส้น แต่มีบางตัวอย่างของโค้งลักษณะข้อสอบ ICCS ที่แสดงให้เห็นถึง DIF และตัวอย่างของข้อสอบที่ไม่แสดง DIF ดูได้จากพื้นที่ระหว่างเส้นโค้งที่มีขนาดแคบมากและพารามิเตอร์สำหรับแต่ละเส้นโค้งเกือบจะเท่ากันตามภาพที่ 2-6 ดังนี้

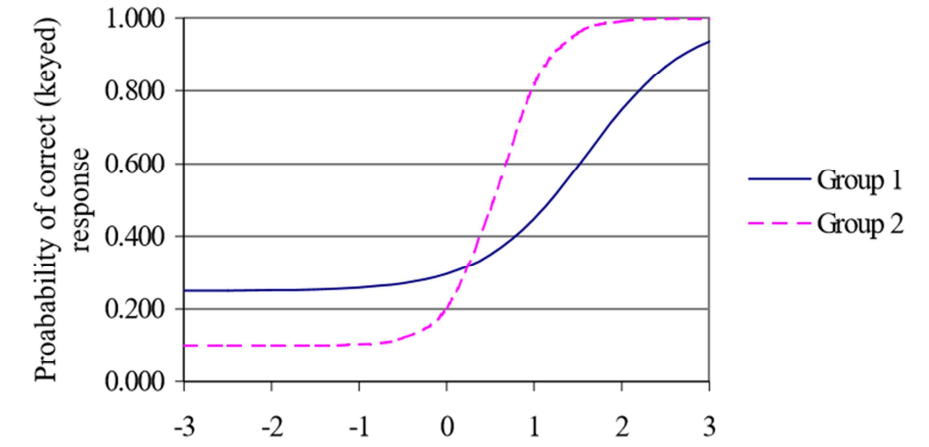


ภาพที่ 2-6 ตัวอย่างของข้อสอบที่ไม่แสดง DIF



ภาพที่ 2-7 ตัวอย่างของข้อสอบที่แสดง DIF

จากภาพที่ 2-7 แสดงให้เห็นข้อสอบที่แสดง DIF ที่มีขนาดใหญ่มากแสดงด้วยพื้นที่ขนาดใหญ่ระหว่างโค้งลักษณะข้อสอบสองเส้น ชนิดของ DIF เป็น uniform DIF เพราะมากระหว่างโค้งลักษณะข้อสอบ ICCS ไม่ไขว้กัน ข้อสอบที่แสดงในภาพที่ 2-5 อาจจะเป็นการวัดความเท่าเทียมกันของตัวแปรแฝงของทั้งสองกลุ่ม



ภาพที่ 2-8 ตัวอย่างของข้อสอบที่แสดง DIF แบบ Nonunifrom DIF

จากภาพที่ 2-8 จะแสดงข้อสอบที่เป็น DIF แบบ Nonunifrom DIF (ได้แก่ โค้ง ICCs ที่ไขว้ข้ามไปยังอันอื่น) แสดงให้เห็น DIF ที่ไม่สม่ำเสมอ (Non-uniform DIF) เนื่องจากบุคคลเหล่านั้นมีคะแนนต่ำกว่าค่าเฉลี่ย (นั่นคือ $z \leq 0$), กลุ่ม 1 เป็นกลุ่มที่มีคะแนนมากกว่าค่าเฉลี่ย (นั่นคือ $z > 0$) อาจทำให้เข้าใจได้ว่าภาพที่ 2-6 จะเป็น DIF ที่ซับซ้อน (และเกินความเป็นจริง)

วิธีการทางสถิติสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF): การถดถอยโลจิสติก (Logistic regression)

การให้คะแนนมีสองรูปแบบที่ใช้กันมากที่สุดสำหรับการทดสอบและการวัด ได้แก่ การให้คะแนนแบบสองค่า (Binry) และการจัดอันดับ (Ordinary) ยกตัวอย่างข้อสอบที่ให้คะแนนเป็นเลขฐานสองอย่างง่าย คือ (1) ข้อสอบที่ให้คะแนนเป็นถูกต้อง/ ไม่ถูกต้อง ในการทดสอบความถนัด หรือผลสัมฤทธิ์ทางการเรียน หรือ (2) ข้อสอบที่ตรวจให้คะแนนแบบสองค่า (Dichotomously scored) ตามเกณฑ์การวัดของบุคลิกภาพหรือมาตรการทางด้านสังคม ข้อสอบที่มีคะแนนเป็นมาตรการจัดอันดับ (Ordinal scale) อาจจะรวมถึงการให้สเกลคำตอบ (Likert-type scales)

ประเภทของคะแนนมีความสำคัญในงานวิจัย DIF เพราะวิธีทางสถิติของ DIF มีการให้ความสำคัญกับการให้คะแนนแบบสองค่า (Binary scored) ส่วนใหญ่แล้ว DIF จะใช้ข้อสอบที่ตรวจให้คะแนนแบบสองค่า สำหรับกรณีของการจัดอันดับ จะประยุกต์ใช้กับวิธีการทางสถิติที่เป็นส่วนขยายของข้อสอบที่ให้คะแนนสองค่า ซึ่งจะมีการแนะนำวิธีใหม่ในการวัดขนาดของ DIF สำหรับรูปแบบการตอบข้อสอบแบบจัดอันดับ วิธีการใหม่สำหรับการตรวจสอบหา DIF จะได้รับการแนะนำโดย Zumbo and Thomas (1997) สำหรับการตอบข้อสอบแบบให้คะแนนสองค่า

วัตถุประสงค์เพื่อแนะนำวิธีการถดถอยโลจิสติก (Logistic regression) พร้อมด้วยรายละเอียดและยกตัวอย่าง สิ่งแรกที่ต้องพิจารณา คือ แบบจำลองทางสถิติ ต่อมาก็เป็นการทดสอบนัยสำคัญทางสถิติ (Significance) สำหรับ DIF และสุดท้ายเป็นการวัดขนาดของ DIF (เช่น อิทธิพลของขนาด (Effect sizes) เป็นต้น

แบบจำลองทางสถิติ (The statistical models)

การตรวจให้คะแนนแบบสองค่า (Binary scored items)

สำหรับข้อสอบแบบให้คะแนนสองค่า (Binary scored items) เพื่อตรวจสอบหา DIF สามารถใช้วิธีการที่แตกต่างกัน (Clauser & Mazor, 1998, Green, 1994; Langenfeld, 1997) ปัจจุบันนี้วิธีการที่มีประสิทธิภาพ และแนะนำสำหรับใช้ในการตรวจสอบ DIF คือ การใช้วิธีการถดถอยโลจิสติก (Clauser & Mazor, 1998; Swaminathan & Rogers, 1990) การถดถอยโลจิสติกขึ้นอยู่กับการสร้างแบบจำลองทางสถิติของความน่าจะเป็นของการตอบข้อสอบถูกต้องโดยสมาชิกกลุ่ม เช่น กลุ่มอ้างอิง และกลุ่มเปรียบเทียบ (ตัวอย่าง เช่น กลุ่มที่ไม่ใช่ชนกลุ่มน้อย และชนกลุ่มน้อยที่มองเห็นได้ ตามลำดับ) และมีหลักเกณฑ์ หรือตัวแปรควบคุม เงื่อนไขหรือหลักเกณฑ์ของตัวแปรโดยปกติมักเป็นคะแนนรวมหรือมาตราส่วน แต่บางครั้งก็เป็นตัวแปรที่เหมือนกันแต่ถูกวัดความแตกต่าง

ขั้นตอนการถดถอยโลจิสติกจะใช้การตอบข้อสอบ (0 หรือ 1) และตัวแปรตาม, กับตัวแปรจัดกลุ่ม (ตัวแปรหุ่น (Dummy coded) โดยให้ 1 = กลุ่มอ้างอิง และ 2 = กลุ่มเปรียบเทียบ) คะแนนรวมทั้งหมดสำหรับแต่ละหัวข้อ (ซึ่งมีลักษณะเป็นตัวแปร TOT) และกลุ่มปฏิสัมพันธ์ของ TOT จะเป็นตัวแปรอิสระ วิธีนี้จะเป็นการทดสอบ DIF แบบมีเงื่อนไขเกี่ยวกับความสัมพันธ์ระหว่างการตอบข้อสอบและคะแนนรวมทั้งหมด โดยการทดสอบผลกระทบของกลุ่มสำหรับ Unifrom DIF และการโต้ตอบของกลุ่มและ TOT เพื่อประเมิน Non-unifrom DIF

สมการถดถอยโลจิสติก (The logistic regression equation)

$$Y = b_0 + b_1 TOT + b_2 GENDER + b_3 TOT * GENDER \qquad (13)$$

โดยที่ Y เป็น log ธรรมชาติของ odds ratio นั่น คือ สมการ

$$In\left[\frac{p_i}{(1-p_i)}\right] = b_0 + b_1 tot + b_2 group + b_3 (tot * group), \qquad (14)$$

โดยที่ p คือ สัดส่วนของประชากรที่เป็นตัวแปรแฝง จากนั้นเราสามารถทดสอบที่ระดับ 2 องศาอิสระในการทดสอบหาค่าไคสแควร์ (Chi-square) สำหรับ Unifrom DIF และ Non uniform DIF ประโยชน์ของการใช้การถดถอยโลจิสติก (มากกว่าวิธีการ DIF วิธีอื่น เช่น วิธี Mantel haenszel) ได้แก่ ไม่จำเป็นต้องเป็นตัวแปรกลุ่ม หรือตัวแปรต่อเนื่อง, สามารถใช้ได้ทั้ง Unifrom และ Nonunifrom DIF และสามารถใช้โมเดลถดถอยโลจิสติกแบบให้คะแนนสองค่ากับข้อสอบที่มีคะแนนแบบจัดอันดับได้

การทดสอบนัยสำคัญสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) (Tests of Significance for DIF)

การทดสอบความสำคัญทางสถิติของ DIF ตามคำนิยาม คือ การสร้างแบบจำลอง DIF มีลำดับชั้นสำหรับการป้อนตัวแปรลงในแบบจำลอง นั่นคือ

ขั้นตอนที่ 1 การป้อนเงื่อนไขของตัวแปร (เช่น คะแนนรวม)

ขั้นตอนที่ 2 ป้อนตัวแปรกลุ่ม (Group variable) และสุดท้าย

ขั้นตอนที่ 3 หาค่าปฏิสัมพันธ์ตามสมการ

ทำการทดสอบค่าไคสแควร์ (Chi-squared test) สำหรับการถดถอยโลจิสติก สำหรับ DIF นั่นคือจะได้รับค่าไคร์สแควร์สำหรับขั้นตอนที่ 3 และลบออกจากค่าไคร์สแควร์สำหรับขั้นตอนที่ 1

ค่าไคสแควร์ที่เป็นผลลัพธ์จะนำมาเปรียบเทียบกับฟังก์ชันการแจกแจงด้วย 2 องศาอิสระ ค่าของ 2 องศาอิสระจะเพิ่มขึ้น ตามข้อเท็จจริง ของแบบจำลองสถิติไคสแควร์ในขั้นตอนที่ 3 มีค่าเป็นสาม และแบบจำลองสถิติไคสแควร์ในขั้นตอนที่ 1 เป็นหนึ่ง (ความแตกต่างในองศาอิสระ

คือ สอง) ผลการทดสอบที่สององศาอิสระของการทดสอบไคสแควร์ คือ การทดสอบแบบจำลองของ Uniform และ Non-uniform DIF (Swaminathan & Rogers, 1990)

Swaminathan and Rogers (1990) ให้ความสำคัญกับการทดสอบที่ระดับ 2 องศาอิสระของการทดสอบ ไคสแควร์ ดังนั้นขนาดของอิทธิพลความสอดคล้อง คือ R-squared ที่เป็นของทั้งสองกลุ่มและเงื่อนไขการเกิดปฏิสัมพันธ์ (นั่นคือ R-squared ในขั้นตอนที่ 3 ลบ R-aquare ในขั้นตอนที่ 1) การถดถอยโลจิสติก DIF จะวัดโดยการทดสอบแบบจำลอง (Simultaneous test) ของ Uniform and non-uniform DIF อย่างไรก็ตามกลยุทธ์การสร้างแบบจำลองตามลำดับที่อธิบายข้างต้นช่วยให้สามารถเปรียบเทียบค่า R-squared ในขั้นตอนที่ 2 กับค่า R-squared ในขั้นตอนที่ 1 เพื่อวัดรูปแบบเฉพาะที่เกี่ยวข้องกับเป็นเอกลักษณ์ของตัวแปรในกลุ่มที่แตกต่างกัน (Unique variation attributable)

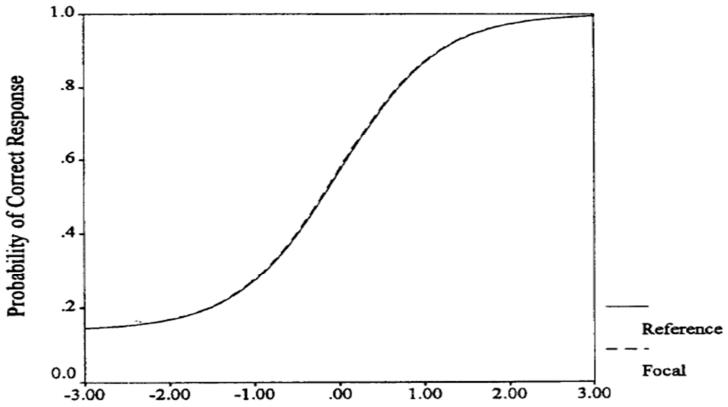
นอกจากนี้การเปรียบเทียบค่า R-squared สำหรับขั้นตอนที่ 3 และขั้นตอนที่ 2 จะวัดเอกลักษณ์ของตัวแปรซึ่งเป็นผลมาจากการปฏิสัมพันธ์เพื่อตรวจสอบว่าเป็น Non-uniform DIF กลยุทธ์นี้จำเป็นต้องได้รับการทดสอบเพิ่มเติมกับข้อมูลจริง ซึ่งแสดงให้เห็นประโยชน์ในการวิเคราะห์ข้อมูล กลยุทธ์การสร้างแบบจำลองในสามขั้นตอนจะใช้ทั้งการถดถอยโลจิสติกแบบให้คะแนนสองค่าหรือการถดถอยโลจิสติกแบบจัดอันดับ ในสาระสำคัญของกลยุทธ์การสร้างแบบจำลองนี้คล้ายกับการทดสอบว่ากลุ่มและตัวแปรมีปฏิสัมพันธ์อย่างมีนัยสำคัญทางสถิติมากกว่าเงื่อนไขตัวแปร (เช่น การจับคู่) สังเกตว่าการทดสอบ Uniform DIF จะมีส่วนเกี่ยวข้องกับความแตกต่างระหว่างขั้นตอนที่ 2 กับขั้นตอนที่ 1 ประโยชน์ของการถดถอยโลจิสติก คือ สามารถทดสอบทั้งแบบจำลองที่เป็นทั้ง Uniform and non-uniform DIF ได้ในเวลาเดียวกัน

ประเภทของการทำหน้าที่ต่างกันของข้อสอบ

การทำหน้าที่ต่างกันของข้อสอบ เป็นการเปรียบเทียบผลการตอบข้อสอบของผู้สอบอย่างน้อย 2 กลุ่มขึ้นไป โดยเรียกกลุ่มผู้เข้าสอบย่อยที่นำมาใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยกลุ่มแรก เรียกว่า กลุ่มเปรียบเทียบ (Focal group) เป็นกลุ่มผู้เข้าสอบย่อยที่เชื่อว่าจะเสียเปรียบในกรณีที่ข้อสอบทำหน้าที่ต่างกัน และกลุ่มที่สอง เรียกว่า กลุ่มอ้างอิง (Reference group) เป็นกลุ่มผู้เข้าสอบย่อยที่ใช้เป็นมาตรฐานในการเปรียบเทียบกับกลุ่มสนใจ เพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เช่น กลุ่มสนใจ ได้แก่ ผู้เข้าสอบผิวดำ ในขณะที่กลุ่มอ้างอิง ได้แก่ ผู้เข้าสอบผิวขาว เป็นต้น (Holland & Thayer, 1988)

ในการทดสอบตามปกติทั่วไปแล้วจะสร้างเครื่องมือให้มีคุณภาพสูงสุดโดยเฉพาะอย่างยิ่งคุณภาพด้านความตรง สังเกตได้จากการรายงานค่าสถิติของคุณภาพเครื่องมือ การใช้เครื่องมือวัดผลที่มีความตรงจะช่วยให้มั่นใจได้ว่าการวัดครั้งนั้นอาจไม่มีความลำเอียง ซึ่งตามปกติทั่วไปถ้า

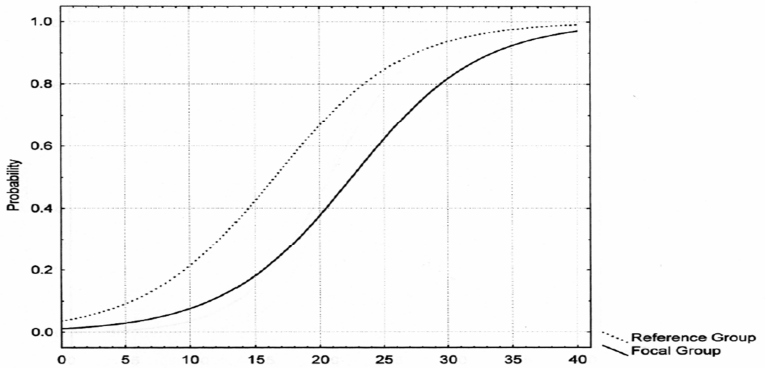
แบบสอบที่ใช้สอบไม่มีความลำเอียงโค้งคุณลักษณะข้อสอบของกลุ่มอ้างอิง และกลุ่มเปรียบเทียบจะเป็นเส้นเดียวกันหรือซ้อนทับกันสนิท และ P (θ) เมื่อเทียบกับกลุ่มอ้างอิง และกลุ่มเปรียบเทียบของผู้สอบคนเดียวกัน หรือมีความสามารถที่แท้จริงเท่ากัน อยู่ในตำแหน่งเดียวกัน ดังภาพที่ 2-10



ภาพที่ 2-9 ลักษณะของโค้งคุณลักษณะข้อสอบปกติ

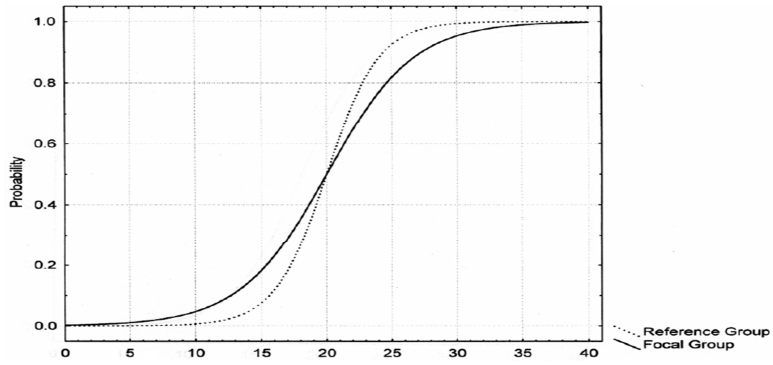
Camilli and Larried (1994) แบ่งการทำหน้าที่ต่างกันของข้อสอบออกเป็น 2 ประเภท คือ

1. การทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป (Unifrom differential item function: Unifrom DIF) หมายถึง คุณลักษณะของข้อสอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่ไม่ซ้อนทับกันและไม่ตัดกันในที่นี้คือกลุ่มอ้างอิง (R) มีโอกาสตอบถูกได้มากกว่ากลุ่มเปรียบเทียบ (F) ตลอดทุกช่วงความสามารถแท้จริงของผู้สอบ พื้นที่ระหว่างโค้งคุณลักษณะข้อสอบทั้งสองเส้นแสดงถึงขนาดของการทำหน้าที่ต่างกันของข้อสอบ ดังแผนภาพที่ 2-11



ภาพที่ 2-10 การทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป (Unifrom difference item functing: Unifrom DIF)

2. การทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป (Non-unifrom differential item functining: Non-unifrom DIF) หมายถึง คุณลักษณะของข้อสอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่ตัดไขว้มากกว่า 1 จุดขึ้นไป แสดงว่าข้อสอบทำให้ผู้สอบกลุ่มหนึ่งตอบถูกมากกว่าอีกกลุ่มในบางช่วงความสามารถที่แท้จริงของผู้สอบแต่ในช่วงอื่นข้อสอบนั้นกลับทำให้ผู้สอบกลุ่มที่มีโอกาสตอบถูกมากกว่ากลายเป็นกลุ่มที่มีโอกาสตอบถูกน้อยกว่า ดังแผนภาพที่ 2-11



ภาพที่ 2-11 การทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป (Non-unifrom difference item functing: Non-unifrom DIF)

ตามทฤษฎีการตอบสนองข้อสอบ (Item response theory: IRT) สามารถพิจารณาปฏิสัมพันธ์ ดังกล่าวได้จากความแตกต่างของค่าพารามิเตอร์อำนาจจำแนกของข้อสอบ ระหว่างผู้สอบกลุ่มย่อยสองกลุ่ม กล่าวคือ ถ้าข้อสอบทำหน้าที่ต่างกันแบบเอกรูป แล้วโค้งลักษณะข้อสอบ (Item characteristic curve: ICCs) ระหว่างผู้สอบกลุ่มย่อยสองกลุ่มจะขนานกัน หรือมีฟังก์ชันการตอบสนองข้อสอบ (Item response functions: IRF) เหมือนกัน แต่ถ้าข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปแล้วโค้งลักษณะข้อสอบระหว่างผู้สอบกลุ่มย่อยสองกลุ่มจะไม่ขนานกันหรือมีฟังก์ชันการตอบสนองข้อสอบต่างกัน ดังนั้นความแตกต่างระหว่างโค้งคุณลักษณะข้อสอบทั้งสองแบบจะบ่งบอกถึงขนาดและทิศทางของข้อสอบที่ทำหน้าที่ต่างกัน ซึ่งสามารถคำนวณได้โดยใช้สูตรการคำนวณพื้นที่ของ Raju (1990)

ข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูป สามารถจำแนกได้เป็น 2 ลักษณะ (Swaminathan & Rogers, 1990) ดังนี้

1. ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปโดยมีปฏิสัมพันธ์ไม่เป็นลำดับ (Disordinal interaction) เป็นการทำหน้าที่ต่างกันสำหรับกลุ่มผู้สอบซึ่งเกิดขึ้น เมื่อโค้งลักษณะข้อสอบตัดกันระหว่างช่วงความสามารถของผู้สอบ หรือเรียกว่า ข้อสอบทำหน้าที่ต่างกันแบบไม่มีทิศทาง (Nondirectional DIF)

2. ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป โดยมีปฏิสัมพันธ์เป็นลำดับ (Ordinal interaction) เป็นการทำหน้าที่ต่างกันสำหรับผู้สอบซึ่งเกิดขึ้นเมื่อโค้งลักษณะข้อสอบต่างกันอย่างไม่สม่ำเสมอ แต่ไม่ตัดกัน หรืออาจตัดกันนอกช่วง ความสามารถของผู้สอบตรงปลายสุดของช่วงความสามารถต่ำหรือสูง อาจเรียกข้อสอบลักษณะนี้ว่า ข้อสอบทำหน้าที่ต่างกันแบบมีทิศทางเดียว (Unidirectional DIF)

โดยทั่ว ๆ ไปในแบบสอบมาตรฐานมักจะมีข้อสอบที่ต่างกันแบบเอกรูปมากกว่าข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปได้มากกว่า แต่ในข้อมูลจริงจะมีข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปได้มากกว่าจะเห็นได้ว่าประเภทของการทำหน้าที่ต่างกันของข้อสอบแบ่งเป็น 2 ประเภท ได้แก่ การทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป (Unifrom DIF) เกิดขึ้นเมื่อผู้สอบกลุ่มหนึ่งมีโอกาสในการตอบข้อสอบถูกมากกว่าผู้สอบอีกกลุ่มหนึ่งในทุกระดับความสามารถ และการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป (Nonunifrom DIF) เกิดขึ้นเมื่อโอกาสในการตอบข้อสอบถูกของผู้สอบความไม่แปรเปลี่ยนของโมเดลการวัดย่อย 2 กลุ่มไม่สม่ำเสมอ สิ่งที่สำคัญมากในเรื่องการทำหน้าที่ต่างกันของข้อสอบ นั่นคือ หลักการและวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบซึ่งเกี่ยวกับวิธีการทางสถิติ

เกณฑ์การตรวจการทำหน้าที่ต่างกันของข้อสอบ

เกณฑ์การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF detection) จำแนกได้ดังนี้

1. ใช้เกณฑ์การให้คะแนนของข้อสอบ จะแบ่งออกเป็น 2 กลุ่ม คือ กลุ่มวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนแบบทวิภาค (Dichotomous DIF method) คือการให้คะแนนแบบ 0, 1 และกลุ่มวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบพหุภาค (Polytomous DIF method) คือ การให้คะแนนแบบหลายค่า และวิธีการตรวจสอบการทำหน้าที่ต่างกันทั้งสองแบบนี้ยังสามารถจำแนกได้ 2 มิติ ได้แก่ มิติลักษณะของตัวแปรเกณฑ์ซึ่งแบ่งเป็นกลุ่มที่ใช้คะแนนสังเกตได้ (Observed score) และกลุ่มวิธีที่ใช้คะแนนที่สังเกตไม่ได้หรือคะแนนของตัวแปรแฝง (Latent variable) และมิติลักษณะของสถิติวิเคราะห์ ซึ่งแบ่งเป็นกลุ่มที่ใช้สถิติพาราเมตริก (Parametric approach) และกลุ่มวิธีที่ใช้สถิตินันพาราเมตริก (Non-parametric approach) ซึ่งมีรายละเอียดดังต่อไปนี้

1.1 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ใช้คะแนนแบบทวิภาค ได้แก่

1.1.1 กลุ่มที่ใช้คะแนนสังเกตได้ วิธีการนี้มักวิเคราะห์ตามทฤษฎีการทดสอบแบบดั้งเดิม (CTT) หรือกลุ่มที่ไม่ใช้ทฤษฎีการตอบสนองข้อสอบ (Non-IRT approach) โดยใช้คะแนนรวมของผู้สอบเป็นเกณฑ์การจับคู่กลุ่มผู้สอบ วิธีการตรวจสอบที่สำคัญในกลุ่มนี้ ได้แก่

การวิเคราะห์ความแปรปรวน (ANOVA) การวิเคราห์ถดถอยโลจิสติก (Logistics การถดถอย) วิธีแมนเทล-แฮนส์เซล (Mantel-Heanszel: MH) และวิธีดัชนีมาตรฐาน (Standardization: STND) 2) กลุ่มวิธีที่ใช้คุณลักษณะหรือตัวแปรแฝง วิธีในกลุ่มนี้ใช้คุณลักษณะหรือตัวแปรแฝงซึ่งวิเคราะห์บนพื้นฐานของทฤษฎีการตอบสนองข้อสอบ (IRT) สำหรับใช้เกณฑ์การจัดคู่กลุ่มผู้สอบ วิธีการตรวจสอบที่สำคัญในกลุ่มนี้ ได้แก่ วิธีวัดพื้นที่ความแตกต่างระหว่างโค้งการตอบสนองข้อสอบ (*IRT*-D^2) วิธีไค-สแควร์ของลอร์ด (Lord's χ^2) วิธีอัตราส่วนไลค์ลิฮูดทั่วไป (General IRT-likelihood ratio) วิธีอัตราส่วนไลค์ลิฮูด ลอกลิเนียร์ (Loglinear IRT-likelihood ratio) และวิธีซิปเทสท์ (SIBTEST)

1.2 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบพหุวิภาค ได้แก่ 1) กลุ่มวิธีที่ใช้คะแนนสังเกตได้ ประกอบด้วย วิธีการวิเคราะห์ความแปรปรวน (ANOVA) วิธีการวิเคราะห์การถดถอยโลจิสติกพหุวิภาค (Polytomous logistic การถดถอย) วิธีดัชนีมาตรฐานพหุวิภาค (Polttomous satandardization) วิธีแมนเทล-แฮนส์เซลทั่วไป (General mantel-haenzel: GMH)

1.3 กลุ่มวิธีที่ใช้คุณลักษณะแฝง ได้แก่ วิธีอัตราส่วนไลค์ลิฮูดในรูปทั่วไป (General IRT-likelihood ratio) วิธีการให้คะแนนบางส่วน (Partial credit model: PCM) วิธีซิปเทสท์ พหุวิภาค (Polytomous SIBTEST) และวิธีการให้คะแนนบางส่วนทั่วไป (General patial credit model: GPMC)

2. ใช้เกณฑ์ตามทฤษฎีการวิเคราะห์ข้อมูล จะแบ่งออกเป็น 2 กลุ่ม คือ 1) กลุ่มที่ใช้วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ไม่ใช้ทฤษฎี IRT (Non-IRT: Non-item response theory-base DIF method) จะวิเคราะห์ดัชนีการทำหน้าที่ต่างกันของข้อสอบโดยใช้คะแนนที่สังเกตได้ภายใต้ทฤษฎีแบบดั้งเดิม 2) กลุ่มที่ใช้วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ใช้ทฤษฎี IRT (IRT: Item response theory-base DIF) ซึ่งจะวิเคราะห์โดยใช้คะแนนสังเกตไม่ได้หรือตัวแปรแฝงภายใต้ทฤษฎีการตอบสนองข้อสอบ

3. ใช้ข้อตกลงเบื้องต้นของโมเดลเป็นเกณฑ์ สามารถแบ่งออกเป็น 2 รูปแบบ คือ 1) รูปแบบพารามเตริก (Parametric from) ซึ่งจะวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยมีข้อตกลงเบื้องต้นของโมเดลสำหรับความสัมพันธ์ระหว่างคะแนนของข้อสอบและการจับคู่ตัวแปร 2) รูปแบบนันพาราเมตริก (Non-parametric from) ซึ่งจะวิเคราะห์ดัชนีการทำหน้าที่ต่างกันของข้อสอบโดยไม่มีข้อตกลงเบื้องต้นของโมเดลและการจับคู่ตัวแปรดังกล่าว

ตอนที่ 4 ทฤษฎีการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ (Theoretic comparison of method)

ความถูกต้องของเครื่องมือวัดขึ้นอยู่กับคุณภาพของข้อสอบ จุดมุ่งหมายโดยรวมเพื่อเปรียบเทียบวิธีการสำหรับการตรวจสอบ และตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (Differential item functioning, DIF) เพื่อหาวิธีการที่เหมาะสม ในข้อสอบที่ให้คะแนนแบบทวิภาค (Dichotomously) มีการอภิปรายถึง คือ ค่าพารามิเตอร์ โดยการจับคู่คะแนน ถ้าสามารถจัดการกับข้อสอบที่ตรวจให้คะแนนแบบทวิภาค และข้อสอบให้คะแนนแบบพหุวิภาคได้ก็สามารถที่จะทดสอบ หรือวัด DIF ได้ และตรวจสอบได้ทั้งข้อสอบแบบเอกรูป (Uniform DIF) และข้อสอบแบบอเนกรูป (Non-uniform DIF) นอกจากนี้ยังมีการอภิปรายถึงการตัดคะแนน (Cut-off score) และขนาดของตัวอย่างที่ต้องการ ผลการศึกษา พบว่า ไม่มีวิธีการใดเลยที่ได้รับการยอมรับ เพราะวิธีการต่าง ๆ เหล่านั้นมีสมมุติฐานที่เข้มงวดที่ในการตรวจสอบ เคยมีคำแนะนำที่เกิดจากการศึกษาเปรียบเทียบด้วยวิธีการ Mantel-Haenszel, Logistic regression, Log linear models and an IRT สุดท้ายแล้วก็ได้ข้อสรุปเพียงแนวทางสำหรับการดำเนินการถ้าพบว่าข้อสอบเกิด DIF (Wiberg, 2007)

การเปรียบเทียบวิธีการ DIF (Comparisons of methods)

เพื่อที่จะเลือกวิธีการ DIF จะต้องใช้เกณฑ์ที่กำหนดของวิธีการตรวจสอบซึ่งจะต้องมีการอภิปรายร่วมกัน และมีการตั้งข้อสังเกตเกี่ยวกับขนาดของกลุ่มตัวอย่าง ที่ผ่านมามีเพียงวิธีการทั่วไป และไม่มีวิธีการที่เฉพาะเจาะจงที่นำมาตรวจสอบหา DIF ยังไม่พบวิธีการใดที่มีอำนาจการทดสอบทางสถิติที่สูงในการตรวจสอบหา DIF เช่น โอกาสความน่าจะเป็นสูงในการระบุข้อสอบที่เกิด DIF แม้จะมีการควบคุมข้อผิดพลาดประเภทที่ 1 การสรุปผลของการเปรียบเทียบตารางที่ 2-1

ตารางที่ 2-1 การเปรียบเทียบวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

วิธีการตรวจสอบ DIF	พาราเมทริก (P)/ นันพาเมทริก (Non-p)	ตัวแปรสังเกต (Obs)/ ตัวแปรแฝง (Lat)	คะแนนสองค่า (D)/ พหุภาค (P)	ทดสอบ (T)/ วัด (M) DIF	Unifrom (U)/ Nonunifrom (N)
Mantel-Haenzel	*Non-p*	Obs	D/ P	T/ M	U
Standardization	*Non-p*	Obs	D	M	U
Chi-square methods	*Non-p*	Obs	D	T	U
SIBTEST	*Non-p*	Lat	D/ P	T/ M	U/ N
Logistic regression	Par	Obs	D/ P	T/ M	U/ N
Likelihood ratio test	Par	Obs/ Lat	D/ P	T/ M	U/ N
Prob.diff.indices	Par	Lat	D	M	U/ N
b.parameter indices	Par	Lat	D	M	U/ N
General IRT-LR	Par	Lat	D/ P	T/ M	U/ N
IRT LRT	Par	Lat	D/ P	T	U/ N
IRT methods	Par	Lat	D/ P	T/ M	U/ N
Lord’s chi-squared test	Par	Lat	D	T	U/ N
Log linear models	Par	Obs	D/ P	T	U/ N
Mixed effect models	Par	Lat	D/ P	T	U/ N

1. การเปรียบเทียบวิธีพาราเมตริก (Parametric) และวิธีนันพาราเมตริก (Non-parametric)

ขั้นแรกถ้าเราใช้วิธีพาราเมตริก สิ่งที่สำคัญมาก คือโมเดลจะต้องมีสมมติฐานที่เป็นจริง หากเลือกสมมติฐานของโมเดลที่มีการละเมิดโมเดลอื่น ก็ควรที่จะต้องเลือกวิธีการนันพาราเมตริกแทน อาจจะเลือกวิธีการนันพาราเมตริกเมื่อเริ่มต้น แต่โดยปกติแล้วจะควบคุมตัวแปรได้ยากมาก ประโยชน์ของการทดสอบด้วยวิธีนันพาราเมตริก เช่น การทดสอบไค-สแควร์, ไม่มีข้อตกลงเกี่ยวกับการกระจายความสามารถของประชากรที่สนใจ (Ironson, 1982) มีเพียงวิธีการ SIBTEST เท่านั้นที่เป็นตัวอย่างของ Monotonicity

วิธีพาราเมตริกทั้งหมดมีทั้งข้อสมมติฐานที่เข้มงวดน้อย และเข้มงวดมาก ซึ่งต้องมีการปฏิบัติตามมิฉะนั้นจะไม่ถูกนำมาใช้ เช่น วิธี LR มีสมมติฐานที่เข้มงวดของความสัมพันธ์ระหว่างความเป็นไปได้ของการตอบข้อสอบถูกต้อง และคะแนนสังเกตได้ต้องเป็นเส้นตรง Embretson & Reise (2000) และ Lord (1980) แสดงให้เห็นว่าการที่คะแนนข้อสอบสังเกตได้ไม่เป็นเส้นตรงนั้นเกี่ยวข้องกับความสามารถแฝงของผู้สอบ MH (Camilli, 2006) นอกจากนี้การใช้วิธีการ IRT วิธีการใดวิธีการหนึ่ง มีความหมายถึง สมมติฐานที่เข้มงวดในหนึ่งมิติความสามารถที่ต้องการทดสอบ โดยสรุปหากโมเดลมีความพอดีกับข้อมูลและมีสมมติฐานที่เป็นจริง วิธี DIF แบบพาราเมตริกจะถูกนำมาใช้ แต่ถ้าไม่สามารถหาโมเดลที่มีสมมุติฐานที่เป็นจริง หรือโมเดลไม่มีความพอดีกับข้อมูลก็สามารถใช้วิธีการ DIF แบบนันพาราเมตริกมาใช้แทนได้ วิธีการใช้ตารางการณ์จร (Contingency) (MH, LR, LLM, Chisquare tests) รวมทั้งวิธีนันพาราเมตริก และวิธีพาราเมตริก มีข้อเสียเปรียบถ้าเพราะใช้ไม่ได้กับพารามิเตอร์การเดา และถ้าเลือกวิธีเหล่านี้สำหรับการเปรียบเทียบกับวิธีการตอบสนองข้อสอบ (IRT) มีข้อเสียตรงที่ไม่มีพารามิเตอร์การเดา และการเลือกวิธีการเหล่านี้เพื่อเปรียบเทียบกับวิธีการ IRT การเลือกข้อสอบโดยปกติจะมีสมมติฐานว่าข้อสอบมีความเท่าเทียมกันทั้งในกลุ่มเปรียบเทียบ และกลุ่มอ้างอิง (Camilli, 2006) วิธี MH, LR และ LLM ทำงานดีที่สุดเมื่อข้อมูลเป็นโมเดลแบบ 1 PL แต่จะมีปัญหามากถ้าข้อมูลเป็นโมเดลแบบ 2PL หรือ 3PL (Millsap & Everson, 1993) ภายใต้สมมติฐานของ LLM สามารถให้ผลลัพธ์เหมือนกับ LR

2. การจับคู่ตัวแปร (Matching variable)

เกณฑ์ข้อที่สองเกี่ยวข้องกับธรรมชาติของตัวแปรที่จับคู่กัน เช่น ตัวแปรสังเกตหรือตัวแปรแฝง ให้เลือกระหว่างตัวแปรทั้งสองตัวนี้ว่าต้องใช้การทดสอบด้วยทฤษฎีแบบดั้งเดิม (คะแนนสังเกต) หรือทฤษฎีการตอบสนองข้อสอบ (คะแนนแฝง) ซึ่งอาจมีผลต่อวิธีการทดสอบทั้งหมดที่จะถูกตรวจสอบ และ/ หรือวิธีการที่รายงานผลลัพธ์ วิธีการที่ใช้คะแนนสังเกตมาจับคู่ตัวแปร ได้แก่ วิธี MH วิธีมาตรฐาน (Standardization) วิธีไค-สแควร์ LR และ LLM วิธีการทั้งหมดให้ผลลัพธ์ที่ไม่ดี ถ้าคะแนนรวมที่เป็นตัวแทนของความสามารถแฝงมีคะแนนต่ำ รวมทั้งความสามารถอื่น ๆ ด้วย ข้อเสียทั่วไปของวิธีการทั้งหมดที่ใช้คะแนนสังเกต คือ สมมติฐานจะแข็งแกร่งเพียงพอขึ้นอยู่กับความสามารถในการเป็นตัวแทนโดยใช้คะแนนทดสอบทั้งหมด ซึ่งคะแนนการทดสอบทั้งหมดไม่ได้เป็นตัวชี้วัดที่สมบูรณ์แบบของการทดสอบความสามารถของผู้ทำการทดสอบ ซึ่งแบบทดสอบยังต้องการทดสอบที่ถูกต้องด้วย (Ironson, 1982; Millsap & Everson, 1993)

มีการตั้งข้อสังเกตว่า แม้ว่าทั้งวิธี LR และ MH จะใช้การจับคู่คะแนนสังเกต ส่วนวิธี LR ใช้ทดสอบกับคะแนนที่เป็นตัวแปรต่อเนื่องที่ไม่จำเป็นต้องจัดแบ่งประเภท ดังนั้นมันจึงมี

ความคลาดเคลื่อนได้น้อย โมเดล LR ที่เป็นคู่สามารถใช้กับคะแนนที่ถูกจัดเรียงลำดับ (Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990) นอกจากนี้วิธี LR ยังมีประสิทธิภาพมากกว่าการจับคู่ที่ต้องการวิเคราะห์ด้วยวิธี MH (Mazor, Kanjee & Clauser, 1995) วิธี MH มีการประมาณค่า DIF ทั้งต่ำและสูงตามองค์ประกอบ เช่น การจับคู่ตัวแปร (Holland & Thayer, 1988) การคาดเดา (Camilli & Penfield, 1997) และเมื่อไม่มีสถิติที่เพียงพอในการจับคู่ (Zwick, 1990)

วิธี IRT วิธีนันพาราเมตริก SIBTEST และโมเดล Mixed effect ที่ไม่ใช้คะแนนสังเกตในการจับคู่ตัวแปรแทนการใช้ตัวแปรแฝง มันมีข้อโต้เถียงว่าคะแนนที่เป็นตัวแปรแฝงมีความแม่นยำในการวัดความสามารถของผู้ทำการทดสอบ ปัญหาของการใช้คะแนนที่เป็นตัวแปรสังเกตเป็นตัวแปรจับคู่จะสามารถแก้ไขได้โดยการเอาข้อสอบที่ทำให้เกิดความลำเอียงออก และทำการวิเคราะห์ซ้ำ (Van der Flier, Mellenbergh, Ader &Wijn, 1984) ดังนั้นไม่ว่าจะเป็นวิธีใดที่ถูกเลือก ทุกวิธีจะต้องมีการทำซ้ำ (Camilli & Shephard, 1994)

3. เปรียบเทียบคะแบบแบบทวิภาค (Dichotomously) และคะแนนแบบพหุวิภาค (Polytomously)

เกณฑ์ข้อที่สามไม่ว่าจะเป็นวิธีการที่สามารถจัดการคะแนนข้อสอบทั้งคะแนนแบบทวิภาค (Dichotomously) และคะแนนแบบพหุวิภาค (Polytomously) มีความสำคัญถ้าเป็นคะแนนข้อสอบแบบทวิภาค เกณฑ์นี้จะถูกใช้เพียงขึ้นอยู่กับวิธีการมีความหยืดหยุ่นถ้าข้อสอบบางข้อมีความแตกต่างกันของรูปแบบข้อสอบ มีการตั้งข้อสังเกตว่าวิธี MH, SIBTEST, LR, LRT, Generation IRT-LR, LLM และ Mixed effect models ทั้งหมดมีความยืดหยุ่น แต่เดิม MH ไม่ได้ถูกออกแบบเพื่อใช้กับคะแนนข้อสอบแบบพหุวิภาค แต่ได้รับการขยายเพื่อตอบสนองวัตถุประสงค์นี้ (Zwick, Donoghue & Grimo, 1993) วิธี LR ยังได้รับการปรับให้เข้ากับคะแนนสอบแบบพหุวิภาค (Camilli & Congdon, 1999) แม้จะมีการรายงานว่ามีความยากในการประยุกต์ใช้วิธี LR กับการตรวจสอบ DIF (Wonsuk, 2003)

4. เปรียบเทียบการวัด (Measure) และ/ หรือการทดสอบ DIF

เกณฑ์ข้อที่สี่ไม่ว่าจะเป็นวิธีการที่สามารถใช้ได้ทั้งตรวจสอบ และวัด DIF แล้วจะต้องมีการพิจารณาอีกว่า มันมีความเป็นไปได้ที่จะเลือกใช้เป็นส่วนเติมเต็มให้วิธีการอื่น ๆ ถ้าวิธีการที่เลือกนั้นสามารถช่วยเสริมกับวิธีการอื่น ๆ ได้ด้วย นอกจากนี้ยังต้องขึ้นอยู่กับวัตถุประสงค์ของการศึกษา DIF เป้าหมายหลักเพื่อระบุข้อสอบที่มีปัญหา หรือต้องการที่จะวัดขนาดของ DIF เพื่อเลือกข้อสอบที่จะมองข้าม วิธี MH วิธี SIBTEST, LR, General LRT, IRT-LR และวิธี IRT วิธีการทั้งหมดสามารถทดสอบ และการวัดขนาดของ DIF วิธีการอื่น ๆ มีข้อเสียอย่างหนึ่ง คือ วัดเพียงขนาดของ DIF (ความน่าจะเป็นของความแตกต่างของตัวบ่งชี้ และตัวบ่งชี้พารามิเตอร์ b) หรือเพียง

ได้รับการทดสอบอย่างมีนัยสำคัญของ DIF (Chisquare methods, Lord's chi-square test, LLM and mixed effect models)

5. เปรียบเทียบข้อสอบแบบเอกรูป (Uniform) และข้อสอบแบบอเนกรูป (Non-uniform DIF)

เกณฑ์ข้อที่ห้าไม่ว่าจะเป็นวิธีการที่สามารถจัดการกับข้อสอบแบบเอกรูป และข้อสอบแบบอเนกรูป เป็นสิ่งที่สำคัญมากเนื่องจากเราไม่สามารถสรุปได้ว่าพฤติกรรมของ DIF เป็นเส้นตรงหรือไม่ วิธีการนันพาราเมตริกสามารถจัดการกับข้อสอบแบบเอกรูป (Uniform DIF) ได้อย่างเป็นที่น่าพอใจ ส่วนวิธี MH มีข้อบกพร่องตรงที่มันถูกออกแบบมาเพื่อวัดข้อสอบแบบเอกรูป ซึ่งหมายความว่ามันไม่ไวต่อข้อสอบแบบอเนกรูป (Non-uniform DIF) มีคำแนะนำว่าวิธี MH สามารถประยุกต์ใช้ในการตรวจข้อสอบแบบอเนกรูป (Mazor, Clauser & Hambleton, 1994) แต่มันยังคงใช้กับการตรวจสอบข้อสอบแบบเอกรูปได้ด้วย ตารางที่ 2 เป็นการจัดสำหรับการตรวจสอบข้อสอบแบบเอกรูป ก่อนที่วิธีการตรวจสอบข้อสอบแบบอเนกรูป จะได้รับการยอมรับอย่างกว้างขวาง Hidalgo & Lopéz (2004) ในสหรัฐอเมริกา วิธี MH ค่อนข้างมีประสิทธิภาพมากกว่าวิธี LR ในการตรวจสอบข้อสอบแบบเอกรูป อย่างไรก็ตามวิธี LR มีระดับการทดสอบทางสถิติสูงกว่าวิธี MH ในการตรวจสอบข้อสอบแบบอเนกรูป (Hidalgo & López-Pina, 2004; Miller & Spray, 1993; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990) Hidalgo & Lopéz (2004) Camilli and Shephard (1994) แนะนำให้ใช้วิธี LR สำหรับตรวจสอบ DIF สรุปดังนั้นจึงมีการใช้วิธี LR แทน MH ถ้าวิธี LR โมเดลมีสมมติฐานที่เป็นจริง มิฉะนั้นต้องใช้วิธี MH ถ้าเพียงแต่ว่าวิธี LR เป็นที่สนใจในการนำมาตรวจสอบข้อสอบแบบเอกรูป แทนวิธี MH ส่วนวิธีการมาตรฐาน และวิธี MH ได้รับผลที่สลับกัน โดยวิธีมาตรฐาน Standardization (Standardized p-difference) มีปัญหาในการตรวจสอบข้อสอบแบบอเนกรูป (Penfield & Camilli, 2007) แต่วิธี SIBTEST สามารถทำได้การตรวจสอบข้อสอบแบบอเนกรูป แม้ว่ามันจะไม่ใช่การทดสอบที่เป็นนันพาราเมตริก ทุกวิธีที่เป็นพาราเมตริกสามารถจัดการกับทุกชนิดของ DIF ได้ (Wonsuk, 2003) ที่ซึ่งสรุปได้ว่าสามารถนำมาใช้ได้

6. เปรียบเทียบวิธีควบคุมคะแนน (Handle) และวิธีตัดคะแนนออก (The cut-off score)

เกณฑ์ข้อที่หกจะมีการพิจารณาคะแนนที่ถูกตัดออกไป มีความเป็นไปได้ที่จะมีการแบ่งคะแนนของผู้ทดสอบในเครื่องมือที่เป็นตารางการณ์จร (Contingency table) ออกเป็นคะแนนที่อยู่ต่ำกว่าคะแนนตัด (Cut-off) และคะแนนที่อยู่เหนือคะแนนตัดทั้งวิธี MH และวิธี LR มันมีความเป็นไปได้ในการแบ่งตัวแปรจับคู่ และควบคุมระดับความสามารถ เช่น ผู้ทดสอบที่สอบไม่ผ่าน หรือทดสอบระหว่างผู้ที่ทำการทดสอบผ่านเท่านั้น ความเป็นไปได้อื่น ๆ คือ การใช้วิธีที่คะแนนสังเกตมีการแบ่งประเภท และรวมถึงโมเดลดังต่อไปนี้ เช่น LLM และ Mixed effect models

มีการตั้งข้อสังเกตถึงการใช้ LRT กับหลาย ๆ วิธีการเหล่านี้ การแบ่งคะแนนรวมทั้งหมดเป็นชั้น ๆ อาจจะเกิดปัญหาถ้าตัวแปรแฝงถูกใช้ในการจับคู่ แม้ว่ามันอาจจะเป็นไปได้ที่โมเดลของผู้ทดสอบยอมรับในการอยู่สูงกว่าหรือต่ำกว่าคะแนนตัด เช่น โมเดลผสม (Mix model) ความเป็นไปได้อื่น ๆ คือ แต่ละโมเดลของผู้ทดสอบถูกแยกออกเป็นชั้น ๆ เช่น โมเดล IRT และแล้วก็นำมาเปรียบเทียบกัน ความคิดเหล่านี้ยังไม่ได้รับการวิจัยเพิ่มเติม

7. ขนาดกลุ่มตัวอย่าง (Sample size)

วิธีตารางการณ์จร (Contingency) (รวมทั้งวิธี MH, LR และ LLM) มีประโยชน์กับกลุ่มตัวอย่างขนาดเล็ก โดยเฉพาะอย่างยิ่งในการเปรียบเทียบกับวิธี IRT (Ironson 1982; Penfield Camilli, 2007) ถ้าขนาดตัวอย่างของกลุ่มเปรียบเทียบมีขนาดเล็กมันจะมีปัญหาต่อการใช้วิธี IRT-LR (Camilli 2006) Camilli and Shephard (1994) แนะนำให้ใช้วิธีโมเดลที่ซับซ้อน (เช่น 3PL model) ถ้ามีตัวอย่างขนาดใหญ่มากพอให้ใช้วิธีตารางการณ์จร เป็นเครื่องมือ วิธี IRT ยังมีการคำนวณอย่างเข้มข้น (Clauser & Mazor, 1998; Millsap & Everson, 1993) Swaminathan และ Gifford (1983) ใช้ผู้ทดสอบจำนวน 1,000 คน เมื่อใช้ข้อสอบจำนวน 20 ข้อ Hulin, Lissak, and Drasgow (1982) ตั้งข้อสังเกตระหว่างความยาวของข้อสอบ และขนาดของกลุ่มตัวอย่าง เมื่อใช้จำนวนผู้ทดสอบน้อยกว่า 1,000 คน และข้อสอบจำนวน 60 ข้อ หรือผู้สอบจำนวน 2,000 คน และข้อสอบ จำนวน 30 ข้อ จะมีความเป็นไปได้ในการประมาณค่าพารามิเตอร์ได้ถูกต้องเมื่อมีการใช้โมเดล 3PL มีการตั้งข้อสังเกตในเรื่องความถูกต้องของการวิเคราะห์ DIF ขึ้นอยู่กับความถูกต้องของโมเดล IRT (Camilli & Shephard, 1994) วิธี SIBTEST จะทำงานได้ดีกับกลุ่มตัวอย่างขนาดเล็ก เช่น จำนวนผู้ทดสอบ 250 คน แม้ว่าข้อสอบจะมีจำนวน 20 ข้อหรือมากกว่า

Millsap and Everson (1993) ยังไม่มีตัวบ่งชี้ที่ระบุความต้องการกลุ่มตัวอย่างขนาดใหญ่ที่ให้ผลลัพธ์ที่คงที่ และไม่ได้ตรวจสอบการกระจายความสามารถของผู้สอบ ซึ่งหมายความว่าตัวบ่งชี้เหล่านี้เกินจริงของในการนับการเกิด DIF

การเลือกกลุ่มตัวอย่างก็เป็นอีกหนึ่งสิ่งที่ต้องเลือกเพื่อดำเนินการวิเคราะห์ DIF เมื่อวิธีการที่ได้รับเลือกในการตรวจสอบ DIF ขนาดของกลุ่มตัวอย่างต้องขึ้นอยู่กับวิธีการที่เลือกเมื่อจะทำการวิเคราะห์ DIF ถ้าการวิเคราะห์ DIF ถูกทำการทดสอบ Clauser and Mazor (1998) เสนอแนะว่ากลุ่มตัวอย่างขนาดใหญ่จะทำให้การทดสอบมีความถูกต้องมากขึ้น เมื่อใช้วิธี IRT

นอกจากนี้แล้วการทำหน้าที่ต่างกันของข้อสอบ (DIF) จาก Good review by Teresi (2006) Medical care vol.44 ยังมีการสรุปไว้ ดังนี้

1. พาราเมตริก ได้แก่ วิธีแมนเทล-แฮนส์เซล (Mantel-kaenszel: MH) (Holland & Thayer, 1988)

2. นันพาราเมตริก ได้แก่

2.1 วิธีถดถอยโลจิสติก (Polytomous logistic regression) (Zumbo, 1999)

2.2 Ordinal Logistic regression (Crane et al., 2004)

2.3 วิธี MIMIC Model (Muthen, 2004)

2.4 โมเดลกลุ่มพหุ (Multiple group models)

2.5 IRT based methods (Thissen, 1991)

2. กลยุทธ์ในการวิเคราะห์ DIF (DIF analyses strategies)

จุดมุ่งหมายโดยรวม คือ การหาวิธีการที่เหมาะสมสำหรับการตรวจสอบ และการวัด DIF วิธีการที่แตกต่างกันสำหรับการตรวจสอบ DIF ได้รับการอภิปราย และเปรียบเทียบกัน วิธีการที่ถูกจัดเป็นวิธีการนันพาราเมตริก หรือวิธีพารามเตริก เป็นไปได้ที่การจับคู่ตัวแปรจะเป็นตัวแปรสังเกตหรือตัวแปรแฝง ไม่ว่าข้อสอบจะเป็นคะแนนแบบทวิภาค หรือคะแนนแบบพหุวิภาค หรือจะเป็นการทดสอบหรือการวัด DIF นอกจากนี้มันยังตรวจสอบ DIFได้ทั้งข้อสอบแบบเอกรูปและอเนกรูป เป็นที่น่าพอใจ และยังสามารถกระทำกับ (หรือโมเดล) กลุ่มสูงกว่า หรือต่ำกว่าคะแนนตัดโดยเฉพาะ สุดท้ายแล้วต้องมีการกล่าวถึงขนาดกลุ่มตัวอย่างที่ต้องการ การเปรียบเทียบวิธีการไม่สามารถทำเพียงวิธีการเดียว มีคำแนะนำว่าควรตรวจสอบอย่างน้อยสี่วิธี เช่น วิธี MH, LR, HLM และวิธี IRT ในการตัดสินใจระหว่างวิธีการเหล่านี้วิธีด้วยการศึกษาเชิงประจักษ์เป็นสิ่งจำเป็น มีข้อสังเกตว่าเพื่อให้มั่นใจว่าเรามีการทดสอบที่ถูกต้องก็ยังไม่เพียงพอแค่การตรวจสอบ DIF ในการทดสอบ

ขั้นตอนการตรวจสอบเป็นสิ่งจำเป็น เช่น Haladyna (2006) ได้ให้คำแนะนำเกี่ยวกับวิธีการดำเนินการ ที่เป็นสิ่งสำคัญของการตรวจสอบ (Mind, 2006) เช่น บริบทของการทดสอบที่เพื่อใช้ให้ถูกต้อง การอภิปรายเกี่ยวกับวิธีการดำเนินการตรวจสอบ DIF ที่จะมุ่งเน้นในครั้งแรกของการดำเนินการทั้งหมด คือ จะต้องมีการระบุข้อสอบที่ต้องการจะตรวจสอบก่อนที่จะทำการทดสอบ เพื่อจะไม่เป็นการล่วงละเมิดสำหรับสมาชิกกลุ่มอื่น ๆ Berk (1982) ได้สรุป และเผยแพร่กระบวนการทดสอบเหล่านี้จำนวน 6 ข้อ รุ่นที่มีการปรับปรุงดูจากของ Ramsey (1993) ที่มีการระบุความไวของการวิเคราะห์ การวิเคราะห์นี้จะดำเนินการเพื่อ 1) ความสมดุลของการทดสอบ 2) ไม่เป็นแบบแผนที่ตายตัว 3) ไม่นำเพศ และวัฒนธรรมมาใช้กำหนดสมมุติฐาน 4) หลีกเลี่ยงการทดสอบที่เป็นการล่วงละเมิดใด ๆ ต่อผู้ทำการทดสอบ 5) ไม่มีเนื้อหาขัดแย้งหรือไม่เป็นที่ต้องการ 6) หลีกเลี่ยงอภิสิทธิ์หรือประเพณี การวิเคราะห์ความไวถูกดำเนินการก่อนนำข้อสอบมาทดสอบกับกลุ่มที่มีขนาดเล็กลงเพื่อให้ข้อสอบทั้งหมดสามารถตรวจสอบ DIF (Longford, Holland & Thayer, 1993) Burton and Burton (1993) ตั้งข้อสังเกตว่าการคัดกรอง

ข้อสอบที่จะตรวจสอบ DIF โดยการทำ Pretested ก่อนเป็นการลดจำนวนข้อสอบที่ทำให้เกิด DIF ในการทดสอบปกติ การคัดกรอง DIF ไม่ได้มีผลกระทบต่อความยากของข้อสอบ, การคัดเลือกข้อสอบ หรือคะแนนเฉลี่ยของการทดสอบสำหรับกลุ่มที่เปรียบเทียบที่ถูกเปรียบเทียบกับกลุ่มอ้างอิง มันเป็นสิ่งสำคัญที่จะเน้นย้ำว่าข้อสอบมีความง่ายมากกว่า หรือข้อสอบที่มีความยากมากกว่าที่จะทำให้เกิด DIF (Linn, 1993)

ดังนั้นจึงควรที่จะเปรียบเทียบสถิติ DIF ที่ได้รับกับสถิติข้อสอบอื่น ๆ เนื่องจากมันมีความเกี่ยวข้องโดยรวมกับความยากของข้อสอบ และการคัดเลือกข้อสอบ (Burton & Burton, 1993; Linn, 1993) การคัดเลือกข้อสอบถูกนิยามโดยความสัมพันธ์ Biserial ระหว่างข้อสอบ และคะแนนทดสอบรวมทั้งหมด และความแตกต่างในความยากความไม่แปรเปลี่ยนของโมเดลการวัดเปรียบเทียบกับกลุ่มอ้างอิง (Linn, 1993) ความยากขอข้อสอบถูกนิยามเป็นเปอเซ็นความถูกต้องของข้อสอบ ในการศึกษาของ Burton and Burton (1993) วิธี MHD-DIF ถูกใช้ในการวัด DIF มีการตั้งข้อสังเกตว่าการศึกษา DIF ไม่สามารถที่จะดำเนินการโดยปราศจากการสะท้อนกลับ เขาต้องมีการติดตามตรวจสอบโดยเฉพาะอย่างยิ่งข้อสอบที่แสดง DIF การทดสอบสถิติอนุมานที่ให้ผลการตรวจสอบ DIF อย่างมีนัยสำคัญทางสถิติ ไม่ได้บ่งบอกถึงความสำคัญในทางปฏิบัติ และความต้องการที่จะได้รับการวัดขนาด ถ้าข้อสอบมีการแสดง DIF สิ่งหนึ่งที่ต้องพยายามหาคือ แหล่งของ DIF เพราะมันไม่จำเป็นว่าต้องเป็นข้อสอบที่ไม่ดี

ข้อสอบอาจจะแสดง DIF ถ้ามันมีรูปแบบที่มีความแตกต่างกันมากกว่าข้อสอบส่วนที่เหลือในการทดสอบ (Longford et al., 1993) ความเป็นไปได้อื่น ๆ ก็คือ การวัดข้อสอบที่มีความแตกต่างของความสามารถจากหนึ่งการวัดในแบบทดสอบ หรือการสะท้อนให้เห็นว่าทั้งสองกลุ่มได้เรียนรู้บางสิ่งบางอย่างด้วยวิธีการสอนที่แตกต่างกัน ดังนั้น มันจึงทำให้ข้อสอบง่ายสำหรับบางกลุ่ม (Camilli, 2006) มีกลยุทธ์ที่สามารถนำไปประยุกต์ใช้ได้ ที่พบมากที่สุดหลังตรวจสอบพบ DIF คือ 1) การเขียนข้อสอบใหม่ 2) การลบข้อสอบ 3) การควบคุมความแตกต่างพื้นฐาน โดยใช้รูปแบบการ IRT สำหรับผู้ตอบแบบสอบถามแต่ถ้าข้อสอบจะถูกเก็บไว้ในแบบทดสอบ ตัวสร้างแบบทดสอบควรจะมีเหตุผลในการตัดสินใจ ตัวสร้างแบบทดสอบควรจะมีเหตุผลในการตัดสินใจ ถ้าข้อสอบแสดง DIF ดังนั้นจึงมีการตัดสินว่าข้อสอบไม่มีความยุติธรรมภายในกลุ่ม

ข้อสอบที่มีความยุติธรรมขึ้นอยู่กับวัตถุประสงค์ของการทดสอบ มันเป็นไปได้ว่าความแตกต่างในคะแนนรวมทั้งหมดจะสะท้อนให้เห็นความแตกต่างความไม่แปรเปลี่ยนของโมเดลการวัด ดังนั้นจึงไม่มี DIF อยู่ สถิติ DIF สามารถตอบเพียงคำถามเล็กน้อย เช่น “ข้อสอบวัดกลุ่มที่แตกต่างกันหรือไม่?” อย่างไรก็ตามจะไม่มีคำถามเพิ่มเติม เช่น ผู้ตรวจสอบพบ DIF มักจะ

ไม่ถามถึงผลที่ตามมาจากการทดสอบ หรือข้อสอบนั้นมีความยุติธรรมหรือไม่ (Camilli & Shepard, 1994; Zieky, 1993) ดังนั้นมันจึงเป็นสิ่งสำคัญสิ่งหนึ่งในการหาเหตุผลเพื่อตัดสินใจลบข้อสอบออกหลังจากพบ DIF. นอกจากนี้ยังมีแนวทางเกี่ยวกับการส่งข้อสอบเพื่อทบทวน ซึ่งควรจะส่งเฉพาะในกรณีที่เป็นกลุ่มอ้างอิง (นั่นคือ ส่วนใหญ่) หรือเป็นกลุ่มเปรียบเทียบ อะไรที่จะกระทำถ้าพบว่าข้อสอบแสดง DIF ควรจะเป็นลบออกไปหรือเขียนข้อสอบใหม่ โดยปกติแล้วมักจะขึ้นอยู่กับการตัดสินใจ หรือข้อผิดพลาดร้ายแรงของการวัด (Zumbo, 1999) ในทางปฏิบัติมันเป็นเรื่องยากที่จะลบข้อสอบที่แสดง DIF ออกจากแบบทดสอบ อย่างน้อยมันควรพบก่อนที่ข้อสอบจะถูกบริหารจัดการ อย่างไรก็ตามการตรวจสอบ DIF สามารถช่วยพัฒนาแบบทดสอบให้มีความยุติธรรมมากขึ้น (Penfield & Camilli, 2007)

การทดสอบที่ข้อสอบไม่แสดง DIF ไม่ได้หมายความว่าการทดสอบนั้นมีความยุติธรรม ในขณะที่การวิเคราะห์ DIF ยังต้องพึ่งพาเกณฑ์ภายใน เช่น การใช้คะแนนรวมทั้งหมดแทนความสามารถ ในการศึกษา DIF ไม่สามารถตรวจสอบความลำเอียงได้อย่างต่อเนื่อง หากข้อสอบทั้งหมดที่ทดสอบแสดง DIF เราไม่สามารถตรวจสอบสิ่งนี้ได้เพราะคะแนนที่สังเกตโดยทั่วไปถูกประมาณค่าต่ำ และการประมาณค่าความสามารถของผู้ทดสอบจะเป็นตัวควบคุม ปัญหาอื่น ๆ อาจเกิดขึ้นเมื่อมีการวิเคราะห์ DIF ที่มันอาจจะเป็นปัญหาในการกำหนดตัวแปรจัดกลุ่ม เช่น วุฒิการศึกษา (Penfield & Camilli, 2007) วิธีการหนึ่งของการปรับปรุงการวิเคราะห์ คือ การใช้วิธีการ DIF ซ้ำ นั่นคือ การเอาข้อสอบที่ทำให้เกิดความลำเอียงออกไป และทำการทดสอบใหม่ DIF เพื่อตรวจสอบข้อสอบที่ทำให้เกิดความลำเอียง (Camilli & Shephard, 1994) มีการตั้งข้อสังเกตว่าศักยภาพของข้อสอบที่ DIF ควรจะรวมอยู่ในตัวแปรจับคู่ของคะแนนทดสอบ มิฉะนั้นการวิเคราะห์ DIF จะให้ผลที่แปลก และกระบวนการ MH ไม่สามารถทำงานได้ถูกต้องถ้าข้อสอบที่เกิด DIF ถูกลบออกไป (Dorans & Holland, 1993; Holland & Thayer, 1988; Lewis, 1993; Zwick, 1990) นอกจากนี้แล้วอาจจะเป็นไปไม่ได้ถ้าลบข้อสอบที่เกิด DIF ออกทั้งหมด เพราะทั้งกลุ่มเปรียบเทียบและกลุ่มอ้างอิงมีประสบการณ์ชีวิตที่ไม่เหมือนกัน

ตอนที่ 5 แนวคิดการวิเคราะห์ด้วยวิธีอัตราส่วนความควรจะเป็น (IRT-likelihood ratio: LR)

การทดสอบอัตราส่วน (IRT Likelihood ratio test) การทดสอบโดยใช้ทฤษฎีการตอบสนองข้อสอบด้วยการทดสอบอัตราส่วนสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (Differential item functioning: DIF) (IRT-LR. Thissen et al., 1988) สามารถใช้ได้ทั้งกับข้อมูลที่ให้คะแนนแบบทวิภาค และคะแนนแบบพหุวิภาค และได้รับการพิสูจน์ว่าเป็นวิธี

การหนึ่งที่มีประสิทธิภาพมากที่สุดสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป (Uniform) และข้อสอบแบบอเนกรูป (Nonuniform) และ DTF

การเปรียบเทียบโมเดลโดยวิธี IRT-LR ได้ถูกอภิปรายในเวลาต่อมา และได้มีการใช้กันอย่างแพร่หลาย โดยมีการใช้การทดสอบอัตราส่วนความควรจะเป็น (Likelihood ratio test) สำหรับตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โมเดลพื้นฐาน (Compact model) คือ การทดสอบโมเดลการตอบสนองข้อสอบ (IRT) และโมเดลเปรียบเทียบ (Augmented model) ซึ่งเป็นโมเดลพหุนามทั่วไป (Multinomial model) ที่รวมความเป็นไปได้ทั้งหมด พารามิเตอร์สามารถเพิ่มเติมตัวแปรที่สนใจไว้ในโมเดล ให้สังเกตว่าความถี่ของแต่ละเซลจะต้องเท่ากัน (Thissen et al., 1993) เมื่อมีการใช้ LRT สำหรับทดสอบ DIF ในโมเดลการตอบสนองข้อสอบที่เป็นโมเดลพื้นฐาน (Compact model) จะมีการกำหนดข้อบังคับ ว่าพารามิเตอร์ของทั้งสองกลุ่มมีความน่าสนใจเหมือนกันซึ่งเป็นเพียงข้อบังคับข้อหนึ่งเท่านั้นในการทดสอบโมเดล แต่จะมีอีกหลายโมเดลที่มีการเพิ่มเติมตัวแปรที่สนใจในข้อสอบ ยกตัวอย่างเช่น ถ้าเราใช้โมเดล 3PL โมเดลเปรียบเทียบ (Augmented model) รูปแบบจะถูกบังคับว่าต้องมีหกตัวแปร (สองตัวแปรสำหรับพารามิเตอร์ของข้อสอบตั้งแต่มีสองกลุ่มขึ้นไป) และโมเดลที่มีรูปแบบจำกัด (Restricted model) จะบังคับว่าต้องมีสามตัวแปร (สันนิษฐานว่าพารามิเตอร์ข้อสอบเหมือนกันในทุกกลุ่ม) (Camilli & Shephard, 1994)

ข้อสอบที่มี DIF จะต้องมีความแตกต่างของความน่าจะเป็นระหว่างโมเดลพื้นฐาน (Compact model) กับค่าพารามิเตอร์จำนวนหนึ่ง (เช่น พารามิเตอร์ที่มีข้อบังคับที่เหมือนกัน) และโมเดลที่มีการเพิ่มเติมตัวแปรทั้งหมดที่น่าสนใจ (เช่น พารามิเตอร์ที่กำหนดให้มีความแตกต่างกัน) โดยไม่คำนึงถึงรูปแบบ หรือข้อสอบที่ถูกใช้ในการกำหนดตัวชี้วัดที่แฝงอยู่เทียบกับการตรวจสอบข้อสอบที่สงสัยว่าจะเกิด DIF ข้อสอบ (Anchor items) จะถูกสันนิษฐานว่ามีความไม่แปรเปลี่ยนระหว่างกลุ่ม โดยการเปรียบเทียบ (Likelihood) ของทั้งสองโมเดล และเลือกโมเดลที่มีความน่าจะเป็นมากที่สุด สถิติทดสอบอัตราส่วน (Likelihood ratio test: LRT) ถูกนิยามไว้ดังนี้

วิธีการทดสอบอัตราส่วนความควรจะเป็น (Likelihood ratio test) จะวิเคราะห์โดยใช้โปรแกรม IRT Pro (Cai, Toit & Thissen, 2011) ทดสอบสมมติฐานระหว่าง 2 โมเดล คือ โมเดลพื้นฐาน (Compact) และโมเดลเปรียบเทียบ (Augmented) ในโมเดลแรกสมมติให้มีกลุ่มผู้สอบที่แตกต่างกัน ดังนั้นจึงบังคับให้พารามิเตอร์ของข้อสอบระหว่างผู้สอบกลุ่มอ้างอิง และกลุ่มเปรียบเทียบมีค่าเท่ากัน (Group 1-Item และ Group 2-Item 1) (Thissen, Steinberg, & Wainer, 1988; 1993 อ้างถึงใน อาวีพร ปานทอง, 2558, หน้า 39)

โดยใช้สมมติฐานในการทดสอบตามสูตร ดังนี้

$$H_0 : a_{\mathrm{iF}} = a_{iR} \text{ และ } b_{\mathrm{ijF}=} b_{\mathrm{ijR}} \text{ สำหรับทุก j}$$

H_A : พารามิเตอร์ของข้อสอบข้อที่ j ของทั้งสองกลุ่มไม่เทากัน อย่างน้อย 1 พารามิเตอร์

สูตรการทดสอบอัตราส่วนความน่าจะเป็นของสองโมเดลแสดงได้ดังนี้

$$LR = \frac{L^*(Model_C)}{L^*(Model_A)} \qquad (15)$$

เมื่อ $L^*(Model_C)$ คือ ฟังก์ชันความน่าจะเป็นของโมเดลพื้นฐาน (ค่าพารามิเตอร์น้อยกว่า)

$L^*(Model_A)$ คือ ฟังก์ชันความน่าจะเป็นของโมเดลเปรียบเทียบที่ยอมให้พารามิเตอร์ข้อสอบของข้อที่ j ระหว่างกลุ่มผู้สอบมีความหลากหลาย

กำหนดฟังก์ชันไลค์ลิฮูด (Likelihood function) ในการประมาณค่าพารามิเตอร์จากค่าที่มากที่สุดในฟังก์ชัน (Maximum likelihood function) ซึ่งสามารถแสดงได้ดังสูตรทางคณิตศาสตร์ดังนี้

$$L(X \mid \theta, a, b) = \prod_{i=1}^{N} \prod_{j=1}^{J} \prod_{h=1}^{k} P(\theta_i, a_j, b_j)^{x_{ij}} \left[1 - P(\theta_i, a_j, b_j)^{1-x_{ij}} \right] \qquad (16)$$

กำหนดให้ i แทน ผู้ตอบคำถามคนที่ 1, 2, …….N

j แทน ข้อคำถาม ข้อที่ 1, 2……...J

h แทน ตัวเลือกในข้อคำถาม ตัวเลือกที่ 1, 2…….k

นั่นคือโมเดลพื้นฐาน (Compact; ($Model_C$) จะประกอบด้วย ข้อสอบที่ทำหน้าที่ไม่ต่างกัน สำหรับในโมเดลเปรียบเทียบจะประกอบด้วย ข้อสอบที่มีค่าพารามิเตอร์ระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่มีค่าแปรเปลี่ยนไปตามกลุ่มโมเดลเปรียบเทียบ (Augmented; $Model_A$) อาจมีข้อสอบ จำนวน 1 ข้อ หรือมากกว่าที่ทำหน้าที่ต่างกัน ทำการเปรียบเทียบระหว่าง 2 โมเดล ด้วยสถิติอัตราส่วนความน่าจะเป็นได้ ดังนี้

$$G_i^2 = -2\log(\mathrm{LR})$$

$$G_i^2 = -2\ln(L^*(Model_C)) - \left[-2\ln\left[L^*(Model_A)\right]\right] \qquad (17)$$

เมื่อ G_i^2 แทน สถิติอัตราส่วนความควรจะเป็นของข้อสอบข้อที่ i

$L^*(Model_C)$ แทน ฟังก์ชันความควรจะเป็นของโมเดลพื้นฐาน (Compact)

$L^*(Model_A)$ แทน ฟังก์ชันความควรจะเป็นของโมเดลเปรียบเทียบ (Augmented)

โดยทั่วไปแล้ว $L^*(Model_C) < L^*(Model_A)$ และสถิติ $G_i^2 > 0$ มีการแจกแจงแบบไค-สแควร์ ซึ่งระดับของความเป็นอิสระ เท่ากับผลต่างของจำนวนพารามิเตอร์ในโมเดล $L^*(Model_C)$ และ $L^* Model_A$ (Atar, 2011) ดัชนีความไม่กลมกลืนในระดับนัยสำคัญทางสถิติของผลการทดสอบจะชี้บอกว่าโมเดลพื้นฐานกลมกลืนน้อยกว่าโมเดลเปรียบเทียบอย่างมีนัยสำคัญทางสถิติ ในการใช้ทดสอบอัตราส่วนความควรจะเป็นค่าไค-สแควร์ ที่องศาความเป็นอิสระ เท่ากับจำนวนพารามิเตอร์ของข้อสอบในการประมาณค่าของข้อสอบเหล่านั้นมีนัยสำคัญทางสถิติ แสดงว่า มีการทำหน้าที่ต่างกันของข้อสอบ

Stark, Chernyshenko and Drasgow (2006) กล่าวว่า การใช้การทดสอบ IRT-LR มีความหลากหลายและใช้อย่างกว้างขวางในงานวิจัย ทำให้การใช้เครื่องมือมีประสิทธิภาพมากขึ้นสำหรับการตรวจสอบ DIF และสามารถควบคุมข้อผิดพลาดประเภทที่ 1 มากกว่าเครื่องมืออื่น ๆ ที่มีอัตราข้อผิดพลาดประเภทที่ 1 ที่มีความแตกต่างกันจากระดับที่คาดหวังซึ่งสังเกตได้จากระดับค่าที่คาดหวังที่ 0.5 ไปจนถึงค่าที่สูงกว่า .90 ขึ้นไปในบางกรณี โดยเฉพาะอย่างยิ่งวิธีการที่แนะนำแต่เริ่มแรก โดย Thissen et al (1988) อ้างถึงหลักของการกำหนดค่าคงที่ (หรือข้อสอบพื้นฐานมีความเป็นอิสระ (Free-baseline) แสดงถึงประสิทธิภาพที่สูงกว่าวิธีทางเลือกอื่น ๆ ที่ทราบค่าหลักของการคงที่อื่น ๆ ทั้งหมด (หรือมีข้อบังคับของข้อสอบพื้นฐาน) เมื่อข้อสอบมีการทำหน้าที่ต่างกันจำนวนมาก (Stark et al., 2006; Wang & Yeh, 2003)

วิธีการของการกำหนดค่าคงที่ เริ่มด้วยการกำหนดโมเดลพื้นฐานที่มีโอกาสเกิดการเปลี่ยนแปลงความสอดคล้องของข้อมูลของผู้ตอบข้อสอบในกลุ่มอ้างอิง และกลุ่มเปรียบเทียบได้มากที่สุด กล่าวคือ พารามิเตอร์ของข้อสอบทั้งหมดมีความเป็นอิสระ ยกเว้นข้อสันนิษฐานการเกิด DIF จะต้องมีความเป็นอิสระของค่าคงที่ หรือแบบทดสอบย่อยของข้อสอบที่จำเป็นในการระบุตัวชี้วัดที่แฝงอยู่ การเปรียบเทียบโมเดลสำหรับการวิเคราะห์ DIF จะเกิดขึ้นโดยการกำหนดข้อบังคับตามเงื่อนไขของวิธีการซึ่งจะมีค่าคงที่ที่ใช้ในการเปรียบเทียบระหว่างกลุ่ม ในทางตรงกันข้าม วิธีการอื่น ๆ ทั้งหมดจะเริ่มด้วยโมเดลพื้นฐานที่ค่าของพารามิเตอร์ของข้อสอบทั้งหมดได้รับการกำหนดข้อบังคับระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบและทำการเปรียบเทียบโมเดลสำหรับการวิเคราะห์ DIF ที่เกิดขึ้นโดยมีความเป็นอิสระของข้อสอบหนึ่งข้อในเวลาที่ต่อเนื่องกัน โมเดลพื้นฐานจึงมีการเปลี่ยนแปลงโดยมีความจำเป็นที่จะต้องมีการทำซ้ำ และความสอดคล้องของข้อมูลก็จะได้รับผลกระทบหากการทดสอบมีการกำหนดข้อบังคับใด ๆ กับข้อสอบที่เกิด DIF ทั้งหมด งานวิจัยแสดงให้เห็นว่าข้อสอบจำนวนมากมีการเกิด DIF หรือกล่าวว่าโมเดลพื้นฐานเกิดข้อผิดพลาด ซึ่งมีความน่าจะเป็นของการเกิดข้อผิดพลาดประเภทที่ 1 ในปริมาณที่สูง (e.g., Finch, 2005; Wang & Yeh, 2003)

Stark et al. (2006) ได้เสนอข้อบังคับของข้อสอบหนึ่งข้อในช่วงเวลาหนึ่ง เงื่อนไขของการกำหนดค่าคงที่ต่อเนื่อง และการทดสอบ DIF จากข้อสอบทั้งหมดในเวลาเดียวกัน (การทดสอบกับแบบทดสอบ) เพื่อหลีกเลี่ยงปัญหา เช่น ความไม่แปรเปลี่ยนที่บางครั้งจะเกิดขึ้นกับ CFA ผลลัพธ์ที่ได้จากการโมเดลชี้ให้เห็นว่า การใช้ทั้ง IRT และ CFA โดยการดำเนินการให้มีค่าคงที่ต่อเนื่องในการทดสอบกับแบบทดสอบ จะมีประสิทธิภาพมากสำหรับการตรวจสอบข้อสอบที่เป็นเอกรูป และแบบอเนกรูป และสามารถควบคุมข้อผิดพลาดประเภทที่ 1 มากกว่าเครื่องมือแบบดั้งเดิมวิธีอื่น ๆ ผลลัพธ์เหล่านี้ได้รับการสนับสนุน และขยายผลการค้นพบของ Wang and Yeh (2003) ที่แสดงถึงการใช้ค่าคงที่ต่อเนื่องที่มีประสิทธิภาพมากกว่าวิธีการอื่น ๆ ภายใต้ความหลากหลายของเงื่อนไขการทดสอบจริง

นอกจากนี้การศึกษาในปัจจุบัน โดย Lopez Rivas, Stark and Chernyshenko (2009) แสดงให้เห็นอำนาจการทดสอบจะสูงขึ้นสำหรับการตรวจสอบหา DIF สามารถทำให้บรรลุผลด้วยวิธีการกำหนดค่าคงที่ต่อเนื่องโดยการใช้เพียงการอ้างอิงการจำแนกความไม่ลำเอียง แม้ว่าขั้นตอน IRT-LR สำหรับการตรวจสอบ DIF อาจจะเห็นบางอย่างที่ยุ่งยากที่ได้รับ แต่โดยภาพรวมแสดงถึงประสิทธิภาพ และความหลากหลายที่มีความสำคัญต่อการศึกษา การตรวจสอบประสิทธิภาพของการใช้ IRT-LR อื่น ๆ ทั้งหมด และพบว่า IRT-LR สามารถทำงานได้ดีภายใต้เงื่อนไขจำนวนมาก ในปัจจุบันนี้การนำเสนอการใช้ค่าคงที่ต่อเนื่องเป็นจุดเริ่มต้นของความสนใจในงานวิจัยมากมาย เพราะมีอำนาจการตรวจสอบ DIF สูงกว่าวิธีการอื่น ๆ รวมทั้งมีอัตราข้อผิดพลาดประเภทที่ 1 น้อย ในสถานการณ์การทดสอบจริง ทดสอบอัตราส่วนความน่าจะเป็น IRT ขั้นตอนแรก คือ การกำหนดค่าคงที่ต่อเนื่องของการทดสอบ IRT-LR (Lopez Rivas et al., 2009; Stark et al., 2006) เพื่อสร้างรูปแบบของโมเดลพื้นฐานตามพารามิเตอร์ข้อสอบที่แตกต่างกันระหว่างกลุ่มเปรียบเทียบและกลุ่มอ้างอิง ด้วยการการกำหนดให้ไม่มีความลำเอียงของข้อสอบทั้งสองกลุ่ม

การทดสอบข้อสอบ เริ่มด้วยการเปรียบเทียบโมเดลพื้นฐาน (Compact model) สำหรับข้อสอบแต่ละข้อภายใต้การวิเคราะห์ที่เกิดขึ้นจากการกำหนดข้อบังคับพารามิเตอร์ของข้อสอบในศึกษาสถานการณ์จริง (นั่นคือการกำหนดให้พารามิเตอร์เท่ากัน) ระหว่างกลุ่มอ้างอิง และกลุ่มเปรียบเทียบ สำหรับการใช้ IRT-LR โมเดลพื้นฐานจะถูกสร้างข้อกำหนดให้กับข้อสอบทุกข้อระหว่างกลุ่ม และทำการเปรียบเทียบโมเดลที่ถูกสร้างขึ้นมาโดยกำหนดพารามิเตอร์อิสระสำหรับข้อสอบหนึ่งข้อ การเปลี่ยนแปลงความสอดคล้องของโมเดล (Goodness of fit) ระหว่างโมเดลพื้นฐาน และโมเดลที่ทำการเปรียบเทียบจะดำเนินการด้วยการใช้ระดับของค่าไคสแควส์ที่มีองศาอิสระเท่ากันเพื่อเปรียบเทียบพารามิเตอร์ที่แตกต่างกัน ดังนั้นเมื่อทำการทดสอบ DIF ระหว่างพารามิเตอร์ a และ b พร้อมกัน ให้สังเกตค่าไคสแควส์ที่ถูกเปรียบเทียบจะต้องมีค่าวิกฤต

ไคสแควส์ที่ระดับ 2 องศาอิสระ และสังเกตว่าค่าไคสแควส์มีค่ามากกว่าค่าวิกฤตนั่นแสดงว่าข้อสอบนั้นทำหน้าที่ต่างกัน (DIF)

ตอนที่ 6 แนวคิดการวิเคราะห์ด้วยวิธี Bayesian

การวิเคราะห์ทางสถิติแบบเบส์มีเป้าหมายหลัก คือ การแจกแจงภายหลัง (Posterior distribution) ของพารามิเตอร์ของโมเดล ซึ่งเป็นค่าถ่วงน้ำหนักเฉลี่ย (Weighted average) ระหว่างพารามิเตอร์ที่ไม่ทราบค่าและข้อมูลที่สังเกตได้ โดยมีการระบุการแจกแจงก่อนหน้า (Prior distributions) ของพารามิเตอร์ที่ไม่ทราบค่า การรายงานผลของพารามิเตอร์แสดงโดยฟังก์ชันความควรจะเป็น (Likelihood function) การอนุมานต่าง ๆ สามารถอธิบายได้ด้วยผลการวิเคราะห์การแจกแจงภายหลัง (Posterior distribution) เช่น การประมาณค่าพารามิเตอร์แบบจุด และการประมาณค่าพารามิเตอร์แบบช่วง ซึ่งข้อตกลงเบื้องต้นของวิธีเบส์ คือ พารามิเตอร์ทั้งหมดเป็นตัวแปรสุ่ม (Random variables) ในขณะที่ในมุมมองของ IRT แบบดั้งเดิมมีเพียงพารามิเตอร์ความสามารถ (Ability parameters) เท่านั้นที่เป็นตัวแปรสุ่มและพารามิเตอร์ข้อสอบ (Item parameters) ถือว่าเป็นตัวแปรคงที่ (Fixed variables)

ทฤษฎีของเบส์ (Bayesian theorem) (Congdon, 2001)

ทฤษฎีของเบส์ เป็นฟังก์ชันความน่าจะเป็นที่แสดงความสัมพันธ์ของข้อมูลที่วัดค่าได้ Y เซตของพารามิเตอร์ที่ไม่ทราบค่า θ และต้องการประมาณค่า โดยทั่วไปแล้วในการประมาณค่าพารามิเตอร์ θ จะถือวาพารามิเตอร์ θ เป็นค่าคงที่แต่ไม่ทราบค่าและการประมาณค่า θ จะทำได้โดยการสุ่มตัวอย่าง (Random sample) จากการแจกแจงของประชากรนั้น ๆ ผู้ที่ประมาณค่าอาจทราบข้อเท็จจริงบางอย่างเกี่ยวกับ θ ก่อนที่จะทำการสุ่มตัวอย่างซึ่งหากนำข้อเท็จนั้นมาใช้ให้เกิดประโยชน์ก็จะช่วยให้การประมาณค่าได้ผลยิ่งขึ้น (ประชุม สุวัตถี, 2527)

ให้เซตของข้อมูล y คือ i = 1, 2, ……., n: $y = \{y_1, y_2, \ldots\ldots..y_n\}$

เซตของพารามิเตอร์ คือ $\theta = \{\theta_1, \theta_2, \ldots\ldots..\theta_k\}$

เซตของตัวแปรร่วมคือ $x = \{x_1, x_2, \ldots\ldots x_n\}$

ให้ $y_1, y_2, \ldots\ldots y_n$ เป็นตัวอย่างสุ่ม (Random sample) จากประชากรที่มีความน่าจะเป็นของฟังก์ชันความหนาแน่น (Probability density function หรือ PDF), $f(y|\theta)$ ในการประมาณค่าด้วยวิธีของเบส์จะถือว่าพารามิเตอร์ที่ไม่ทราบค่า θ จะถูกมองว่าเป็นตัวแปรสุ่ม ขณะที่ y ถูกมองว่าเป็นปริมาณที่ทราบค่า การแจกแจงของ θ เมื่อทราบค่า y ที่มีการแจกแจงความน่าจะเป็นที่ทราบค่าล่วงหน้าโดยแสดงได้ด้วยฟังก์ชั่นความหนาแน่น $f(\theta)$ และเรียก $f(\theta)$ นี้ว่า ความน่าจะเป็นของฟังก์ชันความหนาแน่นก่อนหน้า (Prior probability density function) และให้ฟังก์ชันความน่าจะ

เป็นสามารถเขียนในรูป $f(y|\theta, x)$ หรือ $f(\theta|y)$ และให้ $f(\theta|y)$ เป็นการแจกแจงของ θ เมื่อกำหนดค่า

$y_1 = y_1, y_2 = y_2, ...y_n = y_n$ เรียกว่า ความน่าจะเป็นของฟังก์ชันความหนาแน่นภายหลัง (Posterior probability density function) ซึ่งสามารถคำนวณค่าได้จากทฤษฎีของเบส์ ดังนี้

$$f(\theta \mid y) = \frac{f(y \mid \theta) f(\theta)}{f(y)} \tag{18}$$

เมื่อ $f(\theta|y)$ แทน การแจกแจงภายหลัง (Posterior distribution) ของพารามิเตอร์ θ ภายใต้ข้อมูล y

$f(y|\theta)$ แทน ฟังก์ชันไลค์ลิฮู้ด (Likelihood function)

$f(\theta)$ แทน การแจกแจงก่อนหน้า (Prior distribution)

$f(y)$ แทน การทำนายการแจกแจงก่อนหน้า (Prior predictive distribution)

โดยที่ $f(y) = \int_{\theta} f(y \mid \theta) f(\theta) d(\theta)$ ซึ่งเป็นค่าคงที่

ดังนั้น $f(\theta \mid y) \ \alpha \ f(y \mid \theta) f(\theta)$ เรียกว่า สมการของเบส์ (19)

การอนุมานแบบเบส์ (Bayesian Inference)

กรอบแนวคิดของวิธี Bayesian ได้รวบรวมพารามิเตอร์ที่สนใจ เรียกว่า การแจกแจงก่อนหน้าของพารามิเตอร์ (Prior distributions of parameters) ไปเป็นความน่าจะเป็นของพารามิเตอร์ สมมติให้ θ และ y เป็นเวกเตอร์ของพารามิเตอร์และข้อมูลสังเกต ตามลำดับ วัตถุประสงค์ของวิธี Bayesian เพื่อให้ข้อสรุปเชิงสถิติเกี่ยวกับ θ จากเงื่อนไขความน่าจะเป็นของ θ โดย y ซึ่งเรียกว่า ความหนาแน่นภายหลัง (posterior density), $p(\theta|y)$ (Gelman, Carlin, Stern & Rubin, 2004) ความหนาแน่นภายหลังเป็นสัดส่วนกับการแจกแจงก่อนหน้า (Prior distribution), $P(\theta)$ และ Likelihood function, $P(y|\theta)$ ดังนี้

$$P(\theta \mid y) \ \alpha \ P(\theta) P(y \mid \theta). \tag{20}$$

ตามแนวคิดทางสถิติของ Bayesian เกี่ยวกับพารามิเตอร์ θ หรือค่าสังเกตของ $\tilde{y}$ จะถูกกำหนดให้อยู่ในรูปของความน่าจะเป็น (Probability) ซึ่งความน่าจะเป็นจะเป็นไปตามเงื่อนไขของค่าสังเกตของ y และซึ่งจะอยู่ในรูปของ $p(\theta \mid y)$ หรือ $p(\tilde{y} \mid y)$ นอกจากนี้ยังมีเงื่อนไขเกี่ยวกับตัวแปรร่วม (Covariates) อื่น ๆ ที่ทราบค่า x ซึ่งเป็นหลักการพื้นฐานของตัวแปรสังเกตตามการอ้างอิงด้วยวิธี Bayesian ซึ่งมีการอธิบายไว้ในตำราหลายเล่ม ตามขั้นตอนที่ใช้ในการประเมิน θ หรือ $\tilde{y}$ สำหรับการแจกแจงความน่าจะเป็นของค่า y ของการไม่ทราบค่าที่แท้จริงของ θ จะเห็น

ได้ว่ามีผลการวิเคราะห์ที่ให้ข้อสรุปที่คล้ายคลึงกันด้วยวิธีในการอนุมานทางสถิติที่แตกต่างกัน แต่อย่างไรก็ตามการวิเคราะห์ด้วยวิธี Bayesian สามารถจัดการกับปัญหาที่ซับซ้อนได้ง่ายกว่าวิธีอื่น ๆ

กฎของเบส์ (Bayes's rule)

การหาความน่าจะเป็นของ θ โดย y จะเริ่มต้นด้วยโมเดลการแจกแจงความน่าจะเป็นร่วม (Joint probability distribution) สำหรับ θ และ y ซึ่งความน่าจะเป็นร่วม (Joint probability) หรือฟังก์ชั่นความหนาแน่น (Densities function) จะเรียกว่าเป็นการแจกแจงก่อนหน้า (Prior distribution) p (θ) และการแจกแจงกลุ่มตัวอย่าง (หรือการแจกแจงข้อมูล) p (y|θ) ตามลำดับ

$$p(\theta, Y) = p(\theta)\,p(y \mid \theta)$$

ตามเงื่อนไขโดยทั่วไปของการกำหนดค่าที่ทราบของ y จะใช้คุณสมบัติพื้นฐานของเงื่อนไขความน่าจะเป็นที่เรียกว่ากฎของเบส์ โดยความหนาแน่นภายหลัง (Posterior density) ดังนี้

$$p(\theta/y) = \frac{p\,(\theta,y)}{p\,(y)} = \frac{p(\theta)p\,(y/\theta)}{p\,(y)}, \qquad (21)$$

ขณะที่ p (y) = $\sum\theta$ p (θ) p (y|θ), และผลรวมของค่าที่เป็นไปได้ทั้งหมดของ θ กรณีที่ θ เป็นตัวแปรไม่ต่อเนื่อง (หรือ p (y) =$\int$p (θ) p (y|θ) dθ ในกรณีที่ θ เป็นตัวแปรต่อเนื่อง) เทียบ เท่ากับรูปแบบของ (1.1) ที่ไม่มีองค์ประกอบของ p (y) ซึ่งไม่ขึ้นอยู่กับ θ และค่าคงที่ของ y ดังนั้นสามารถพิจารณาให้เป็นค่าคงที่ ดังนั้นความหนาแน่นภายหลัง (posterior density)

$$p(\theta \mid y) \;\; \alpha \;\; p(\theta)\,p(y \mid \theta) \qquad (22)$$

ดังนั้น p (y | θ) ถูกนำมาใช้เป็นฟังก์ชันของ θ ซึ่งเป็นสูตรหลักทางเทคนิคการอนุมานแบบเบส์: ซึ่งเป็นการทำในขั้นตอนแรกสำหรับการพัฒนาโมเดล p (θ, y) และทำการคำนวณเพื่อให้ได้ข้อสรุปของ p (θ | y) ด้วยวิธีการที่เหมาะสม

หากพิจารณาเหตุการณ์ A และเหตุการณ์ B สมมติให้ A = $A_1 \cup A_2 \cup \ldots\ldots\ldots \cup A_n$เมื่อ $A_i \cap A_j = \varnothing$ สำหรับ I $\neq$ j แล้วใช้ทฤษฎีของเบส์ในการหาความน่าจะเป็นของเหตุการณ์ A ภายใต้เหตุการณ์ B จะได้สมการที่เรียกว่า กฎของเบส์ (Bayes'rule) ซึ่งถูกค้นพบโดย Piere-simon de Laplace (Ntzofras, 2009, p. 3) ดังนี้

$$\mathrm{P(A \mid B)} = \frac{\mathrm{P(B \mid A)P(A)}}{\mathrm{P(B)}} \qquad (23)$$

$$P(A_j) \mid B = \frac{P(B \mid A_j)P(A_j)}{\sum_{j=1}^{n} P(B \mid A_j)P(A_j)} \qquad (24)$$

จัดให้อยู่ในรูปแบบที่ง่ายขึ้นจะได้ว่า

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{25}$$

$P(A|B) \quad \alpha \quad P(B|A)P(A)$ เรียกว่า กฎของเบส์ (Bayes'rule)

การทำนาย (Prediction)

เพื่อให้การอนุมานเกี่ยวกับตัวแปรที่ไม่ทราบค่า มักเรียกว่า การอนุมานแบบการคาดการณ์หรือการทำนาย ตามตรรกะที่คล้ายกัน ก่อนที่ข้อมูล y จะได้รับการพิจารณา การแจกแจงของ y ที่ไม่ทราบค่าแต่สามารถสังเกตจาก

$$p(y) = \int p(y,\theta)d\theta = \int(\theta)\, p(y \setminus \theta)d\theta \tag{26}$$

บ่อยครั้งจะเรียกว่าการกระจายหน่วยสุดท้าย (Marginal distribution) ของ y หรือการทำนายการแจกแจงก่อนหน้า เนื่องจาก ก่อนหน้านี้ไม่มีเงื่อนไขของกระบวนการสังเกตและการทำนายเนื่องจากเป็นการแจกแจงในปริมาณที่สังเกตได้ $\tilde{y}$ จากการะบวนการเดียวกัน ยกตัวอย่าง เช่น $y = (y_1, ..., y_n)$ อาจเป็นเวกเตอร์ของค่าน้ำหนักที่บันทึกได้ในจำนวน n ครั้ง ในสเกลของ $\theta = (\mu, \sigma 2)$ ที่อาจเป็นน้ำหนักที่แท้จริงที่ไม่ทราบค่า และระดับของความแปรปรวน และ $\tilde{y}$ อาจจะยังไม่ได้รับการบันทึกค่าน้ำหนักใหม่ที่ได้รับ การแจกแจงของ $\tilde{y}$ เรียกว่า การทำนายการแจกแจงภายหลัง เนื่องจากมีเงื่อนไขของการสังเกต y และการทำนาย เพราะเป็นการทำนายสำหรับ y ที่สังเกตได้

$$\begin{aligned} p(\tilde{y} \setminus y) &= \int p(\tilde{y}, \theta | y)d\theta \\ &= \int p(\tilde{y} | \theta, y)\, p(\theta | y)d\theta \\ &= \int p(\tilde{y} | \theta)\, p(\theta | y)d\theta \end{aligned} \tag{27}$$

บรรทัดที่สองและสามแสดงการการทำนายการแจกแจงภายหลังซึ่งเป็นค่าเฉลี่ยการแจกแจงภายหลังของ θ ขั้นตอนสุดท้ายมาจากเงื่อนไขอิสระของ y และ $\tilde{y}$ ที่ให้ค่า θ

การทำนายการแจกแจงภายหลัง (Posterior predictive distribution)

การทำนายการแจกแจงภายหลังของตัวแปรสังเกตในอนาคต $\tilde{y}$, $p(\tilde{y} | y)$ สามารถคำนวณได้โดยตรงจากสมการอินทิเกรต (Integration) ที่ 28

$$p(\tilde{y} \setminus y) = \int p(\tilde{y} \setminus \theta)\, p(\theta | y)d\theta$$

$$\alpha \int \exp\left(-\frac{1}{2\sigma^2}(\tilde{y}-\theta)^2\right)\exp\left(-\frac{1}{2\tau_1^{\,2}}(\theta-\mu_1)^2\right)d\theta \qquad (28)$$

บรรทัดแรกเป็นการแจกแจงของตัวแปรในอนาคต โดย $\tilde{y}$ จะให้ค่า θ ที่ไม่ขึ้นอยู่กับข้อมูลที่ผ่านมา, y จะสามารถตรวจสอบการกระจายของ $\tilde{y}$ ได้จากคุณสมบัติของการแจกแจงแบบปกติของตัวแปรคู่ (Bivariate normal distribution) ภายในสมการอินทิแกรน (integrand) จะเป็นฟังก์ชั่นเอกซ์โพเนนเชียลกำลังสอง (Exponential of a quadratic) ของ ($\tilde{y}$, θ) ดังนั้น $\tilde{y}$ และ θ จึงมีการแจกแจงภายหลังร่วม (Joint normal posterior distribution) ดังนั้นการแจกแจงหน่วยสุดท้ายของ $\tilde{y}$ จึงมีการแจกแจงแบบปกติ

เราสามารถตรวจสอบค่าเฉลี่ยและความแปรปรวนของการทำนายการแจกแจงภายหลังได้จากการแจกแจงภายหลังซึ่ง $E(\tilde{y}\mid\theta)=\theta$ และ $\text{var}(\tilde{y}\mid\theta)=\sigma^2$ ได้ดังนี้

$$E(\tilde{y}\mid y)=E(E(\tilde{y}\mid\theta,y)\mid y)=E(\theta\mid y)=\mu_{1,} \qquad (29)$$

และ $\text{var}(\tilde{y}\mid y) = E(\text{var}(\tilde{y}\mid\theta, y)\mid y) + \text{var}(E(\tilde{y}\mid\theta, y)\mid y)$

$$\text{var}(\tilde{y}\mid y) = E(\text{var}(\tilde{y}\mid\theta, y\mid y) + \text{var}(E(\tilde{y}\mid\theta, y)\mid y)$$

$$= E(\sigma^2\mid y) + \text{var}(\theta\mid y)$$

$$= \sigma^2 + \tau_1^{\,2}$$

ดังนั้นการทำนายการแจกแจงภายหลังของ $\tilde{y}$ จึงมีค่าเฉลี่ย เท่ากับค่าเฉลี่ยภายหลังของ θ และความแปรปรวนของทั้ง $\tilde{y}$ และ θ จะทำนายความแปรปรวน σ^2 จากโมเดลและความแปรปรวน $\tau_1^{\,2}$ ขึ้นอยู่กับความไม่แน่นอนภายหลังของ θ

ห่วงโซ่มาร์คอฟ (Markov chain Monte Carlo: MCMC)

ความแตกต่างพื้นฐานระหว่าง MCMC และ เทคนิคการประมาณอื่น ๆ ที่มีชื่อเสียง เช่น การประมาณค่าความเป็นไปได้สูงสุด (Maximum likelihood: ML) คือ การประมาณค่าด้วยวิธีการอ้างอิงแบบเบส์จะเน้นที่การแจกแจง (Distribution) เมื่อเทียบกับการประมาณค่าแบบจุด เมื่อมีการอธิบายพารามิเตอร์ของโมเดล และยังให้คำอธิบายการประมาณค่าของการแจกแจงค่าพารามิเตอร์ที่ได้เพิ่มขึ้นด้วยการประมาณค่า ML

ในการวัดทางการศึกษาเพื่อใช้เทคนิคห่วงโซ่มาร์คอฟ (MCMC) สำหรับการประมาณค่าโมเดลการตอบสนองข้อสอบ (เช่น Beguin & Glas, 2001; Bolt & Lall, 2003; Bradlow, Wainer & Wang, 1999; De la Torre, Stark, & Chernyshenko, 2006; Fox & Glas, 2001; Johnson & Sinharay, 2005; Patz & Junker, 1999 a) MCMC มีข้อดีหลายประการ รวมถึงความสะดวก และง่ายต่อการใช้งาน และความพร้อมของฟรีซอฟต์แวร์ นักวิจัยหลายคนได้ใช้ MCMC กับการทดสอบ

โมเดลการวัดก่อนที่จะหาความน่าจะเป็นสูงสุด (Maximum likelihood procedures) เช่น การใช้ MCMC ในการประมาณค่าการอนุมานแบบเบส์ (Bayesian inference) สำหรับโมเดลการตอบสนองข้อสอบ (IRT) (เช่น Glas & Meijer, 2003; McLeod, Lewis & Thissen, 2003; Zwick, Thayer & Lewis, 2000)

ขั้นตอนวิธีการ MCMC มีความซับซ้อนมาก และต้องใช้อย่างระมัดระวัง ข้อบกพร่องหลักของ MCMC คือ ต้องใช้การคำนวณอย่างมาก ขั้นตอนการสุ่มตัวอย่างของวิธีการ MCMC จะต้องมีการทำซ้ำจำนวนมากก่อนที่พารามิเตอร์ของโมเดลจะสามารถประมาณค่าได้อย่างน่าเชื่อถือ (Reliability) ซึ่งการประมาณค่าในหนึ่งครั้งต้องใช้เวลาหลายชั่วโมง หรือใช้เวลาเป็นวันสำหรับโมเดลที่มีความซับซ้อนมากขึ้น หรือเมื่อข้อมูลมีจำนวนมาก

การประยุกต์ใช้ MCMC ในการประมาณค่าโมเดลตอบสนองข้อสอบ จะอธิบายโดยซอฟแวร์ของ WINBUGS (Spiegelhalter, Thomas, Best & Lunn, 2003) ซึ่งเป็นโปรแกรมคอมพิวเตอร์ที่ง่ายต่อการจัดการกับกระบวนที่หลากหลายของ MCMC, WINBUGS สามารถใช้หาความกระชับพอดี (Fit) ของกลุ่มที่มีโมเดลทางสถิติที่แตกต่างกัน และมีการสร้างกราฟิก เพื่อช่วยให้ผู้ใช้สามารถควบคุมกระบวนการประมาณค่า และประเมินผลลัพธ์จนเป็นที่ยอมรับ

การใช้ห่วงโซ่มาร์คอฟ (MCMC) กับโมเดลการตอบสนองข้อสอบ (IRT Models)

วิธีการต่าง ๆ ที่สามารถนำมาประยุกต์ใช้ในการประมาณค่าด้วยวิธี MCMC เช่น การสุ่มตัวอย่าง ข้อวินิจฉัยสำหรับการประเมินการลู่เข้าของห่วงโซ่ (Chain convergence) และการเปรียบเทียบโมเดล และเนื้อหาเกี่ยวกับความสอดคล้องพอดีภายใต้วิธีการของ MCMC

ความน่าจะเป็นของฟังก์ชันความหนาแน่น (Probability density functions)

การใช้ทฤษฎีของเบส์ ความน่าจะเป็นของฟังก์ชันความหนาแน่น (Probability density function) จะเป็นการแสดงความสัมพันธ์ของความควรจะเป็น (Likelihood) ในแต่ละผลลัพธ์ยกตัวอย่าง ซึ่งความน่าจะเป็นของฟังก์ชั่นความหนาแน่น (Probability density functions) รวมถึงฟังก์ชันความหนาแน่นปกติ (Normal density function) จะแสดงในรูปแบบที่เป็นเส้นโค้งคล้ายรูประฆัง (Bell-shap curve) ในบริบทของ IRT ทฤษฎีของเบส์ สามารถเขียนได้ด้วยความน่าจะเป็นของฟังก์ชันความหนาแน่น ได้ดังนี้

$$f(\Omega \mid X) = (f(X \mid \Omega / \left[\int_{\Omega} f(X \mid \Omega) f(\Omega) d\Omega\right]_{(2)} \qquad (30)$$

โดยที่ X = การตอบข้อสอบ

Ω = ค่าพารามิเตอร์ของข้อสอบและพารามิเตอร์ของผู้สอบ

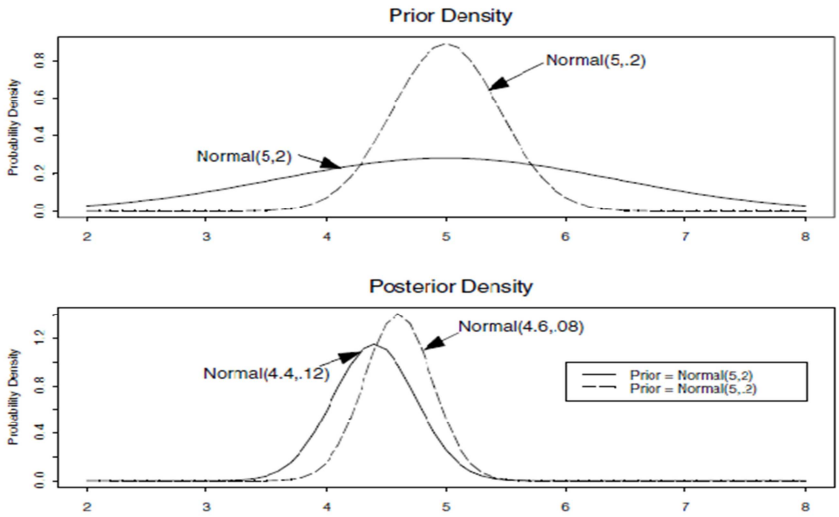
$f(X \mid \Omega)$ = likelihood function

$F(\Omega)$ = ความหนาแน่นก่อนหน้า (Prior density) ของพารามิเตอร์

ส่วนทางด้านซ้ายมือของ (2) เรียกว่า การแจกแจงความหนาแน่นร่วม (Joint posterior density) โดยทั่วไปให้ $f(\Omega \mid X)$ เป็นความหนาแน่นที่เหมาะสม (Proper density) เพราะยังไม่ทราบค่า อาจเขียนได้ ดังนี้

$$f(\Omega \mid X) \ \alpha \ f(X \mid \Omega) * f(\Omega) \qquad (31)$$

เพื่อแสดงให้เห็นว่า ความหนาแน่นภายหลังร่วม (Joint posterior density) คือ สัดส่วนที่แสดงอยู่ด้านขวามือ สัดส่วนของความสัมพันธ์จะอยู่บนพื้นฐานของกระบวนการสุ่มตัวอย่างภายใต้ MCMC โดยมีเป้าหมาย คือ การทำซ้ำของการแจกแจง $f(\Omega \mid X)$ ปฏิสัมพันธ์ของความหนาแน่นก่อนหน้า (Prior density) กับการระบุโมเดลจะถูกตรวจสอบจากการแจกแจงภายหลัง



ภาพที่ 2-12 แสดงแผนภาพการแจกแจงความหนาแน่นของฟังก์ชั่นก่อนหน้าและฟังก์ชั่นภายหลังที่ไม่ทราบค่าเฉลี่ย (_) ของการแจกแจงปกติ

กระบวนการของ MCMC ภายหลังการเก็บรวบรวมข้อมูล โมเดลการตอบสนองข้อสอบก็ถูกนำมาใช้ จะมีการระบุพารามิเตอร์ของโมเดลทั้งหมดมาก่อนจากกลุ่มตัวอย่างที่มีความน่าจะเป็น วัตถุประสงค์ของการ MCMC คือ การกำหนดกลไกการสังเกตกลุ่มตัวอย่างจากความหนาแน่นของการแจกแจงภายหลังร่วม (Joint posterior density) ของพารามิเตอร์ของโมเดลโดยจะมีกระบวนการทำซ้ำ (Iterative process) ที่ดำเนินการภายใต้ MCMC เพื่อพิจารณาความแตกต่างใน ML ซึ่งจะแสดง

ให้เห็นจากการแจกแจงภายหลัง (Posterior distribution) ของพารามิเตอร์ของโมเดล ในกระบวนการของ MCMC จะใช้การสุ่มตัวอย่างแบบกิ๊บส์ (Gibbs sampler)

Gibbs sampler

Gibbs sample หรือ Gibbs sampling เป็นส่วนหนึ่งของ MCMC Class โดยจะดำเนินการจำลองตัวแปรสุ่มจากการแจกแจงแบบมีเงื่อนไข ซึ่งอธิบาย ได้ดังนี้

กำหนดให้ $\theta = [\theta_1 \theta_2]$ ซึ่งมี Posterior Density $p(\theta) = p(\theta_1, \theta_2)$ โดยที่ไม่ต้องใช้ตัวแปรตามของ y ได้ ถ้าทราบเงื่อนไข (Conditional Density) ซึ่งไม่ต้องรับรองเงื่อนไขมากนัก แต่ทราบได้จากเงื่อนไข $p(\theta_1|\theta_2)$ ในการจำกัดจากการเลือกจาก $p(\theta_1|\theta_2)$ เช่น ถ้าพิจารณากรณีอย่างง่าย คือ มีค่าพารามิเตอร์ 3 ค่า $(\theta_1, \theta_2, \theta_3)$ แทนเป็น 3 เงื่อนไขของ Posterior Density เมื่อทำการหา Iterative 1 รอบ จะได้จาก

$f_1(\theta_1|\theta_2\theta_{3,}\, y)$

$f_2(\theta_2|\theta_3\theta_{1,}\, y)$

$f_3(\theta_3|\theta_1\theta_{2,}\, y)$ โดยที่ $y = (y_1, y_2 \ldots . y_a)^T$

ซึ่งกระบวนการของ Gibbs sampling เป็นดังนี้

1. พิจารณา Arbitrary set ของ Starting parameter value ซึ่งจะได้ค่า $\theta_{1,0}, \theta_{2,0}, \theta_{3,0}$
2. สร้าง M + N set ของ Random number โดยทำการ Iterative จาก Full conditional posteripr distribution เราจะได้อันดับแบบ General from แบบ *i-th* เป็น $\{\theta_{1,i}, \theta_{2,I}, \theta_{3,I}\}$ ดังนั้นเราจะได้

2.1 $\theta_{1,i+1}$ จาก $f_1(\theta_1|\theta_{2,i}\theta_{3,i}\, y)$

2.2 $\theta_{2,i+1}$ จาก $f_2(\theta_2|\theta_{3,i}\theta_{1,i+1}\, y)$ และ

2.3 $\theta_{3,i+1}$ จาก $f_3(\theta_3|\theta_{1,i+1}\theta_{2,i+1}\, y)$ จากอันดับ (i+1) -the Realization

3. ทิ้งตัวแรก M ที่ได้จากการรวบรวมในขั้นที่ 2 แล้วใช้ค่า N ที่เป็นตัวสุดท้ายเพื่อทำการสร้างแบบของ และทำการประมาณค่า Posterior marginal โดยใช้ Random sample ภายใต้เงื่อนไขอย่างง่าย German and Geman (1984) ได้แสดงว่า Joint distribution ของ Above random sample จะ Convergence expoentially สู่ Joint posterior distribution ของ $(\theta_1, \theta_2, \theta_3)$ ดังนั้นก่อนที่จะ Realization from ของ Random sample จาก Joint distribution $f_1(\theta_1, \theta_2, \theta_3|\, y)$ เราสามารถทำการอนุมานได้จาก

$$\hat{\theta}_i = \frac{1}{N \quad M} = \sum_{i=M+1}^{N} \theta_{i,j} \qquad (32)$$

$$\hat{\sigma}_i^2 = \frac{1}{N \quad M \quad 1} \Sigma_{i=M+1}^{N} (\theta_{i,j}, \hat{\theta}_i)^2 \qquad (33)$$

โดยค่า Credible interval $\{\iota, \mu\}$ ของ θ_i อยู่ภายใต้เงื่อนไข p (1 θ_i μ) = 0.95 เป็นต้น เป็น Algorithm อีกแบบซึ่งเป็นการแทนที่ค่าที่เป็นไปได้ในแต่ละ State เพื่อทำการหาค่าที่ดีที่สุด ในที่นี้เป็นค่าความเสี่ยง และ Probability function (PDF) เพื่อเป็นการตัดสินว่าค่าที่ได้เป็นค่าที่ดีที่สุด นั่นคือ เป็นหนึ่งในวิธีการหาค่าที่ดีที่สุดของข้อมูล (Finding best solution)

การตรวจสอบห่วงโซ่มาร์คอฟ (Monitoring the Markov chain)

เมื่อเริ่มกระบวนการสุ่มตัวอย่าง จำเป็นที่ต้องมีการระบุค่าเริ่มต้นของพารามิเตอร์ของโมเดล ค่าเหล่านี้สามารถสร้างได้โดยการสุ่มอย่างอิสระ ในขณะที่ผลลัพธ์โดยทั่วไปจะมีความสัมพันธ์เชิงบวกระหว่างค่าพารามิเตอร์ของกลุ่มตัวอย่าง แต่มักจะพบปัญหาในตำแหน่งการเริ่มต้นกลุ่มตัวอย่าง ซึ่งจะมีอิทธิพลต่อการเริ่มต้นตำแหน่งของห่วงโซ่ ดังนั้นจึงเป็นจะทำการตัดทิ้งจำนวนของตำแหน่งเริ่มต้น (เรียกว่าตำแหน่ง “Brun-in”) อีกเรื่องที่สำคัญในการตรวจสอบตำแหน่งของแบบจำลองห่วงโซ่มาร์คอฟ คือ การประเมินผลการลู่เข้ามาบรรจบ (Convergence) กันของห่วงโซ่ ซึ่งในทางทฤษฎีแล้วควรจะมาบรรจบ เมื่อการแจกแจงมีค่าคงที่ ดังนั้นตำแหน่งที่เกิดเหตุการณ์การลู่เข้าจะขึ้นอยู่กับหลายองค์ประกอบ ได้แก่ 1) ความสัมพันธ์ระดับสูงระหว่างตำแหน่งที่อยู่ติดกัน (Adjacent states) จะทำให้มีอัตราการลู่เข้าอย่างช้า ๆ ดังนั้นจะต้องมีการทำซ้ำจำนวนมากก่อนที่จะดูกลุ่มตัวอย่างจากการแจกแจงภายหลังได้ และ 2) อัลกอริทึมของกลุ่มตัวอย่างมีอิทธิพลต่ออัตราการลู่เข้า สาเหตุการไม่เกิดการลู่เข้าอาจจะเกี่ยวข้องกับการปัญหาการระบุโมเดล

การประมาณค่าพารามิเตอร์ด้วยวิธี MCMC การลู่เข้า (Convergence) ของการประมาณค่าพารามิเตอร์จะถูกตรวจสอบ ถ้าการประมาณค่าพารามิเตอร์ไม่พบการลู่เข้า จะแสดงให้เห็นถึงผลลัพธ์ในการอ้างอิงพารามิเตอร์ที่ทำการศึกษาไม่ถูกต้อง (Incorrect) ดังนั้นเราจึงต้องทำการตรวจสอบจำนวนของการ Burn-in ในระหว่างที่การประมาณค่าพารามิเตอร์มีความคงที่ หรือมีการลู่เข้า และต้องกำหนดจำนวนการทำซ้ำ (Iteration) หลังจากที่มีการ Burn-in ข้อมูลแล้ว เพื่อยืนยันว่าเราได้สุ่มตัวอย่างจากการแจกแจงภายหลังทั้งหมดแล้ว Sinhary (2004) ได้สรุปแนวทางการวินิจฉัยการลู่เข้า (Convergence) โดยการประมาณค่าพารามิเตอร์ด้วยกราฟิก เริ่มด้วย

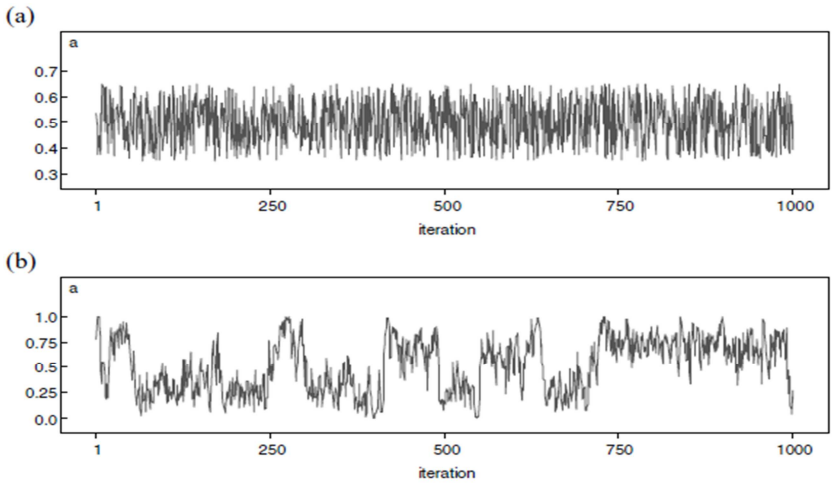
History plots ของการแจกแจงภายหลัง (Posterior distributions) จะได้รับการตรวจสอบเมื่อการประมาณค่าพารามิเตอร์มีความคงที่ (Stable)

Density plots จะถูกตรวจสอบว่าการแจกแจงภายหลังมีการแจกแจงแบบปกติ (Normal distribution) หรือไม่ จะแสดงรูปแบบเป็นเส้นโค้งคล้ายรูประฆัง (Bell-shap curve)

นอกจากนี้แล้ว Autocorrelation plots ก็จะแสดงความสัมพันธ์ระหว่างกลุ่มตัวอย่างจากการแจกแจงภายหลังในสองช่วงเวลาที่แตกต่างกัน

การตรวจสอบการลู่เข้ามาบรรจบ (Convergence) เป็นส่วนสำคัญในการระบุว่าการดำเนินการ MCMC ได้รับความสำเร็จ ในภาพที่ 2-13 (a) และ (b) แสดงภาพตัวอย่างของห่วงโซ่ MCMC ซึ่งแสดงให้เห็นถึงโอกาสสูงของการลู่เข้ามาบรรจบกัน (Convergence) ในภาพที่ 2-13 (a), และแสดงให้เห็นถึงไม่ลู่เข้ามาบรรจบกัน (Nonconvergence) ในภาพที่ 2-13 (b) ตัวอย่างหนึ่งของการวินิจฉัยการลู่เข้า คือ เกณฑ์ของ Geweke's (1992) ภายใต้วิธีการของ Geweke ค่าคะแนนของ Z สำหรับค่าพารามิเตอร์ที่ถูกกำหนด โดยการใช้ความแตกต่างระหว่างค่าเฉลี่ยของตำแหน่งแรกเริ่ม ร้อยละ 10 และค่าเฉลี่ยของตำแหน่งสุดท้าย ร้อยละ 50 และหารด้วยผลรวมของค่าส่วนเบี่ยงเบนมาตรฐาน ค่า Z ภายในช่วงของการไม่มีนัยสำคัญ (ได้แก่ $-1.96 \leq z \leq 1.96$) สามารถนำมาเป็นหลักฐานของการลู่เข้า (Convergence) สู่ค่าคงที่ และเกณฑ์ของ Gelman และ Rubin (1992) การทดสอบของ Gelman และ Rubin อยู่บนพื้นฐานของการเปรียบเทียบความแปรปรวนภายในและระหว่างห่วงโซ่ของแต่ละพารามิเตอร์ จากปริมาณเหล่านี้สถิติอัตราความแปรปรวน $\sqrt{\hat{R}}$ ถูกคำนวณในแต่ละพารามิเตอร์ ในขณะที่ $\sqrt{\hat{R} \approx 1}$ จะบ่งบอกถึงการลู่เข้า (Convergence) สู่ค่าคงที่

ยังคงมีกลยุทธ์อื่น ๆ สำหรับการประเมินการลู่เข้า (Convergence) สู่ค่าคงที่ ในแบบจำลองที่มีห่วงโซ่จำนวนมาก แต่ละอันก็จะต้องอยู่บนพื้นฐานของการมีตำแหน่งเริ่มต้นที่แตกต่างกัน ถ้าการลู่เข้าของห่วงโซ่อยู่ในตำแหน่งการแจกแจงที่เหมือนกัน



ภาพที่ 2-13 แสดงการแจกแจงก่อนของห่วงโซ่มาร์คอฟที่แสดงการลู่เข้า (a) และการไม่ลู่เข้า (b)

การสร้างการแจกแจงภายหลัง (Constructing posterior distributions)

เมื่อมีหลักฐานเพียงพอของการลู่เข้า (Convergence) สู่ค่าคงที่ ห่วงโซ่จำลองสามารถนำมาใช้ในการสร้างการแจกแจงภายหลังสูงสุด (Marginal posterior distributions) ที่เป็นพื้นฐานของการประมาณค่าพารามิเตอร์ของโมเดล ในการกำหนดวิธีการเก็บตัวอย่างในตำแหน่งของห่วงโซ่มีความจำเป็นต้องยอมรับอิทธิพลของความคลาดเคลื่อนมาตรฐาน Standard errors ในผลลัพธ์ของ Monte Carlo, Patz and Junker (1999b) เพราะการแจกแจงภายหลัง (Posterior distributions) อาจจะถูกสร้างจากกลุ่มตัวอย่างที่มีการสุ่มตัวอย่างผิดพลาด ดังนั้นเมื่อคำนวณการแจกแจงภายหลัง (Posterior distributions) เช่น ค่าเฉลี่ย ความผิดพลาดในการประมาณค่า จึงไม่สามารถอ้างได้ว่ามีเพียงความคลาดเคลื่อนมาตรฐานในการประมาณค่าแบบจุด แต่ยังมีความคลาดเคลื่อนในการสุ่มร่วมด้วย เรียกว่า ความคลาดเคลื่อนของ Monte Carlo ตามกฎของ Thumb แบบจำลองจะต้องมีการดำเนินการ (Run) จนกระทั่งความคลาดเคลื่อนของ Monte Carlo ในแต่ละพารามิเตอร์ที่สนใจน้อยกว่าร้อยละ 5 ของส่วนเบี่ยงเบนมาตรฐานของกลุ่มตัวอย่าง (Spiegelhalter et al., 2003)

วิธีการประมาณค่าพารามิเตอร์ด้วยวิธีเบส์ (Bayesian Estrimation)

ในการวิเคราะห์ทางสถิติแบบเบเซียน (Bayesian statistics analysis) จะนำทฤษฎีของเบส์เซียน เข้ามาประยุกต์ใช้ กล่าวคือ ในการพิจารณาการแจกแจงภายหลัง (Posterior distribution) จะอาศัยข้อมูลสารสนเทศจากตัวอย่างที่อยู่ในรูปแบบของฟังก์ชันควรจะเป็น (Likelihood function) และการแจกแจงก่อนหน้า (Posterior distribution) ดังสมการต่อไปนี้

$$f(\theta,\alpha,\beta \mid y) \propto f(\alpha,\beta) f(\theta \mid \alpha,\beta) f(y \mid \theta,\alpha,\beta)$$
$$= f(\alpha) f(\beta) f(\theta \mid \alpha,\beta) \ell(y \mid \theta,\alpha,\beta)$$

เมื่อ α และ β เป็นอิสระต่อกัน ขั้นต่อไป คือ การดำเนินการหาค่าการแจกแจงภายหลังแบบมีเงื่อนไข สำหรับค่าพารามิเตอร์ θ, α และ β จากการแจกแจงร่วมภายหลังเริ่มจาก $f(\theta \mid y,\alpha,\beta) = f(\theta, y,\alpha,\beta) \mid f(y,\alpha,\beta)$ โดยกฎของความน่าจะเป็นแบบมีเงื่อนไข อย่างไรก็ตาม $f(\theta, y,\alpha,\beta) = f(y \mid \theta,\alpha,\beta) f(\theta,\alpha,\beta)$ (Li & Baser, 2012) ดังนั้น $f(\theta \mid y,\alpha,\beta) = f(y \mid \theta,\alpha,\beta) f(\theta,\alpha,\beta) / f(y,\alpha,\beta)$ ถ้ากำหนดค่าพารามิเตอร์ θ, α และ β เป็นอิสระจะได้ว่า

$$f(\theta \mid y,\alpha,\beta) \propto f(y \mid \theta,\alpha,\beta) f(\theta) f(\alpha) f(\beta)$$
$$\propto \ell(y \mid \theta,\alpha,\beta) f(\theta)$$

เพราะการแจกแจงของ $f(\alpha)$ และ $f(\beta)$ เป็นการแจกแจงของค่าคงที่ α และ β ซึ่งการแจกแจงภายหลังเป็นสัดส่วนของการแจกแจงก่อนหน้ากับอัตราส่วนความควรจะเป็น ดังสมการต่อไปนี้

$$f(\theta \mid y,\alpha,\beta) \propto f(\theta)\ell(y \mid \theta,\alpha,\beta),$$
$$f(\alpha \mid y,\theta,\beta) \propto f(\alpha)\ell(y \mid \theta,\alpha,\beta),$$
$$f(\beta \mid y,\theta,\alpha) \propto f(\beta)\ell(y \mid \theta,\alpha,\beta),$$

MCMC เป็นกระบวนการที่มีความซับซ้อนในการประมาณค่า โดยการทดสอบจะใช้กระบวนการ Gibbs sampling ในการสุ่มค่าพารามิเตอร์ที่ทำการศึกษา ซึ่งโปรแกรมที่ใช้ในการประมาณค่าพารามิเตอร์นี้ คือ OpenBUGS ภายใต้ขั้นตอนของ Gibbs algorithm

ตอนที่ 7 การวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ (Multi-group confirmatory factor analysis (MGCFA)

การทดสอบความเที่ยงข้ามกลุ่มประชากร (Cross validation) ที่อิสระต่อกัน เป็นวิธีการทดสอบความเที่ยงตรงของโมเดล (Jöreskog & Sörbom, 1993, pp. 128-129) วิธีหนึ่งระหว่างประชากรที่มีการวัดซ้ำหลาย ๆ ครั้ง หรือการวิเคราะห์กลุ่มพหุ (Multi-group analysis) (นงลักษณ์ วิรัชชัย, 2542, หน้า 217) เพื่อตรวจสอบค่าพารามิเตอร์ของโมเดลของประชากรกลุ่มหนึ่งว่าจะเหมือนกับประชากรอีกกลุ่มหนึ่งหรือไม่ โดยการเปรียบเทียบโครงสร้างและค่าพารามิเตอร์ของโมเดลที่ เรียกว่า การศึกษาความไม่แปรเปลี่ยนของโมเดล การตรวจสอบความเท่าเทียมกันของการวัด โดยการวิเคราะห์องค์ประกอบเชิงยืนยัน เริ่มด้วยการกำหนดรูปแบบในการวิเคราะห์องค์ประกอบเชิงยืนยัน ซึ่งเป็นการระบุรูปแบบของตัวแปรทางทฤษฎีที่ผู้วิจัยต้องการยืนยันกับข้อมูลเชิงประจักษ์ ได้มาจากการกำหนดจำนวนตัวแปรแฝง (ξ) จำนวนตัวแปรสังเกต (x) ความแปรปรวนร่วมระหว่างตัวแปรแฝง ความสัมพันธ์ระหว่างตัวแปรสังเกตและตัวแปรแฝง ความสัมพันธ์ระหว่างตัวแปรสังเกตกับความคลาดเคลื่อนและความแปรปรวนและความแปรปรวนร่วมระหว่างความคลาดเคลื่อน (δ) ซึ่งข้อได้เปรียบของการวิเคราะห์องค์ประกอบเชิงยืนยัน คือ สามารถกำหนดส่วนประกอบต่าง ๆ ที่กล่าวมาตามความต้องการของนักวิจัย

Jöreskog and Sörbom (1993) ได้สร้างรากฐานของวิธีการ MGCFA เพื่อศึกษาความไม่แปรเปลี่ยนของโมเดลการวัด (MI) และตรวจสอบความแตกต่างของค่าเฉลี่ยของตัวแปรแฝงในประชากรข้ามกลุ่มซึ่ง ขยายมาจากโมเดลพื้นฐานของการวิเคราะห์องค์ประกอบเชิงยืนยัน, MGCFA เป็นวิธีการที่อยู่บนพื้นฐานของทฤษฎีการทดสอบความไม่แปรเปลี่ยนของโมเดลการวัด (Bollen, 1989; Byrne, 1998; Vandenberg & Lance, 2000) ซึ่งได้รับความสนใจจากนักวิจัย และมีการใช้กันอย่างแพร่หลายในการวิจัยเพื่อทดสอบความไม่แปรเปลี่ยนของโมเดลการวัด (French & Finch, 2008) ยกตัวอย่าง เช่น Muthén and Lehman (1985) และ Glockner-Rist and Hoijtnik (2003)

กล่าวถึงการใช้งานของ MGCFA เพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้งแบบเอกรูป (Uniform DIF) และอเนกรูป (No uniform DIF)

หลักการของ MGCFA คือ การประมาณค่าพารามิเตอร์ของประชากรในแต่ละกลุ่ม และการทดสอบแบบตามลำดับขั้น เพื่อตรวจสอบความไม่แปรเปลี่ยนของโมเดลการวัดระหว่างกลุ่มประชากร หลักการโดยการวิเคราะห์โครงสร้างค่าเฉลี่ยของตัวแปรแฝง คือ การเปรียบเทียบค่าเฉลี่ยตัวแปรแฝงระหว่างกลุ่มประชากรด้วยการพิจารณาจุดตัด (Thresholds) ของตัวชี้วัดสังเกต การวิเคราะห์ค่าเฉลี่ยโครงสร้างตัวแปรแฝง เพื่อตรวจสอบว่าค่าเฉลี่ยของตัวแปรแฝงเหล่านั้น มีการกระจายความสามารถแฝงแตกต่างกันระหว่างกลุ่มประชากร ความไม่แปรเปลี่ยนของโมเดลการวัด (MI) ถูกตรวจสอบด้วยการทดสอบ และประเมินความสอดคล้องของโมเดล กับ ข้อมูล ของโมเดลสอดแทรก (Nested models) เริ่มด้วยการไม่มีการกำหนดข้อบังคับกับโมเดลพื้นฐาน ในกรอบแนวคิดทฤษฎีที่ผ่านมา และต่อมาให้ดำเนินการใส่ข้อบังคับให้มีความเท่าเทียมกันของค่าพารามิเตอร์ในโมเดล (เช่น ค่าน้ำหนักองค์ประกอบ (Factor loading) จุดตัด (Threshole) เป็นต้น ของประชากรต่างกลุ่ม และดำเนินการทีละขั้นตอน

ในการศึกษาจากกลุ่มตัวอย่างหลายกลุ่ม เวกเตอร์ของจุดตัด (Thresholds) สำหรับตัวชี้วัดสังเกต เป็นสิ่งสำคัญเมื่อทำการทดสอบความไม่แปรเปลี่ยนของโมเดลการวัด (MI) และใช้ในการประมาณค่า ความแตกต่างของค่าเฉลี่ยของตัวแปรแฝง (Bollen, 1989; Byrne, 1998; Byrne, et al., 1989; Hancock, 1997; Ployhart & Oswald, 2004)

การทดสอบความไม่แปรเปลี่ยนของโมเดลการวัด เพื่อให้บรรลุความไม่แปรเปลี่ยนของโมเดลการวัดแบบเข้มงวด/ มีความสมบูรณ์ครบถ้วน ตามคำแนะนำในการวิจัยก่อนหน้า (Byrne, 1998; Chan & Schmitt, 1997; Lytle, Brett, Barsness, Tinsley & Janssens, 1995; Raju et al., 2002; Rensvold & Cheung, 1998; Vandenberg & Lance, 2000) Following Cheung & Rensvold (1999) and Vandenberg (2002) แนะนำให้ดำเนินการเป็นขั้นตอนตามลำดับชั้นของการทดสอบ MI (Hierarchical MI) โดยมีขั้นตอนดังนี้

1. การทดสอบความเท่าเทียมกันของความแปรปรวนร่วมของเมตริก
2. การทดสอบความไม่แปรเปลี่ยนของโมเดลการวัด (Measurement invariance)
   - 2.1 การทดสอบโมเดลพื้นฐาน (Baseline models) วิเคราะห์แยกกลุ่มตัวอย่าง
   - 2.2 การทดสอบความเท่าเทียมกันของค่าพารามิเตอร์ซึ่งประกอบด้วย Factor loading
   - 2.3 ความไม่แปรเปลี่ยนของจุดตัด (Thresholds)
   - 2.4 ความไม่แปรเปลี่ยนด้านความเฉพาะตัว (Uniqueness invariance)

3. การทดสอบความไม่แปรเปลี่ยนเชิงโครงสร้าง (Structure invariance)

3.1 การทดสอบความเท่าเทียมกันของโครงสร้างองค์ประกอบ (Marsh, Hau, Chung & SIU, 1998, pp. 149-150) ดังนี้

3.1.1 การทดสอบความไม่แปรเปลี่ยนของความแปรปรวน-ความแปรปรวนร่วมขององค์ประกอบ (Factor variance-covariance (Φ) Invariance test)

3.1.2 การทดสอบความเท่าเทียมกันของความคลาดเคลื่อนของความแปรปรวน (Measurement error)

3.1.3 การทดสอบความไม่แปรเปลี่ยนของค่าเฉลี่ยองค์ประกอบ (Invariance of factor Means)

1. การทดสอบความเท่าเทียมกันของความแปรปรวนร่วมของเมตริกของประชากรต่างกลุ่ม ตามคำแนะนำของ Jöreskog and Sörbom (1989) เป็นการดำเนินการก่อนการทดสอบชนิดของความไม่แปรเปลี่ยนของโมเดลการวัด ขั้นตอนนี้เป็นการทดสอบสมมติฐานว่าความแปรปรวนร่วมของเมตริกของประชากรต่างกลุ่มมีค่าเท่ากัน (Vandenberg, 2002) และเมตริกความแปรปรวนจากประชากรต่างกลุ่มกันมีค่าเท่ากัน (Vandenberg, 2002) ซึ่งเป็นการทดสอบสมมติฐานข้อแรกว่า:

$$H_0: \sum yy^{(1)} = \sum yy^{(2)} = \ldots\ldots = \sum yy^{(g)}$$

ถ้าค่าไค-สแควส์แสดงการปฏิเสธสมมติฐานก็จะต้องดำเนินการทดสอบ ความไม่แปรเปลี่ยนของโมเดลการวัด (MI) ส่วนที่เหลือ เพราะประชากรที่ศึกษามีความน่าจะเป็นว่าจะมีค่าพารามิเตอร์ของประชากรแตกต่างกัน (ค่าน้ำหนักองค์ประกอบ (Factor loadings), ความคลาดเคลื่อนของความแปรปรวนและความแปรปรวนร่วม (Error variance and Covariance) มิฉะนั้นมันจะบ่งชี้ว่าความเป็นไปได้ของความเท่าเทียมกันของพารามิเตอร์ประชากร และการทดสอบความไม่แปรเปลี่ยนของโมเดลการวัดส่วนที่เหลือไม่จำเป็นต้องดำเนินการต่อ การศึกษาความไม่แปรเปลี่ยนของการวัด (Measurement invariance) (Cheung & Rensvold, 1999) กล่าวว่า เป็นเรื่องยากที่จะจินตนาการว่าทั้งสองกลุ่มตัวอย่างมาจากประชากรที่แตกต่างกัน (ซึ่งโดยปกติในกรณี MI) เมตริกภาพรวมจะต้องเหมือนกัน ถ้าค่าความแปรปรวนของค่าพารามิเตอร์ของประชากรได้รับการปฏิเสธ ก็ยังสามารถดำเนินการทดสอบความเท่าเทียมกันของเมตริกความแปรปรวนขึ้นอยู่คำถามของงานวิจัย, เวลา, ระดับของทักษะ และเทคนิคอื่น ๆ, ระดับความยากลำบากในการดำเนินการทดสอบ เช่น การทดสอบภาพรวมของความเท่าเทียมกันของความแปรปรวนร่วมไม่มีค่าสูงพอ ถ้าหากสมมติฐานของการทดสอบภาพรวมได้รับการปฏิเสธ ก็สามารถทดสอบความไม่แปรเปลี่ยนของส่วนที่เหลือของพารามิเตอร์ประชากรต่อไปได้

2. การทดสอบโมเดลพื้นฐาน การทดสอบโมเดลพื้นฐาน เป็นการทดสอบความไม่แปรเปลี่ยนด้านโครงแบบ (Configural MI) (Thurstone, 1947) หรือที่เรียกว่า “สมมติฐานความไม่แปรเปลี่ยนของรูปแบบองค์ประกอบ” (Horn & McArdle, 1992; Vandenberg, 2002) เพื่อประเมินความสอดคล้องของโมเดลกับข้อมูลเชิงประจักษ์ในแต่ละกลุ่มย่อย โดยแยกวิเคราะห์แต่ละกลุ่มเพื่อตรวจสอบว่ามีโครงสร้างองค์ประกอบที่เหมือนหรือแตกต่างกันอย่างไร เมื่อมีการปรับโมเดลให้มีความสอดคล้องกับข้อมูลเชิงประจักษ์ ซึ่งเป็นแนวคิดที่จะนำไปสู่การทดสอบความไม่แปรเปลี่ยนของโครงสร้างองค์ประกอบ และค่าพารามิเตอร์ โดยทดสอบรูปร่างของโมเดลว่าไม่แปรเปลี่ยนจริง ซึ่งจะเกิดขึ้นเมื่อข้อมูลที่เก็บมาในแต่ละกลุ่มมีจำนวนองค์ประกอบ และจำนวนความสอดคล้องกันในแต่ละองค์ประกอบ (Meredith, 1993 cited in Cheung & Rensvold, 2002, pp. 235-237) กล่าวอีกนัยหนึ่งคือ เป็นการทดสอบโมเดลตามทฤษฎีที่สอดคล้องกับข้อมูลเชิงประจักษ์ของกลุ่มประชากรแต่ละกลุ่ม ซึ่งประกอบด้วย ตัวแปรในโมเดล ลักษณะโครงสร้างความสัมพันธ์ระหว่างตัวแปรเป็นแบบเดียวกันหมดทุกกลุ่ม เมตริกพารามิเตอร์ในโมเดลการวัดมีขนาดเท่ากัน มีรูปแบบเมตริก (Matrix from) และสถานะเมตริก (Matrix mode) เป็นแบบเดียวกันทุกกลุ่ม แต่ขนาดพารามิเตอร์ไม่เท่ากัน จะปรากฏเมื่อรูปแบบของเมตริกมี 0 เป็นองค์ประกอบอยู่ในตำแหน่งเดียวกันของประชากรทุกกลุ่ม นั่นคือเมื่อการทดสอบภาพรวมความเท่าเทียมกันของความแปรปรวนระหว่างกลุ่มมีการปฏิเสธสมมติฐานข้อที่หนึ่งอย่างชัดเจน ดังนี้

$$H_0 : \Lambda_{form}^{(1)} = \Lambda_{form}^{(2)} = ...\Lambda_{form}^{(g)}$$

เมื่อ $\sum_{yy}^{(g)} = \Lambda_y^{(g)}\Phi^{(g)}\Lambda_y^{'(g)} + \Psi_\varepsilon^{(g)}$, และ $\mu_y^{(g)} = \tau_y^{(g)} + \Lambda_y^{(g)}\kappa^{(g)}$

จากสมมติฐาน โดยธรรมชาติแล้วจะต้องดำเนินการทดสอบขั้นตอนนี้ก่อนที่จะทำการตรวจสอบความไม่แปรเปลี่ยนของค่าพารามิเตอร์ประชากร ในการทดสอบจะไม่มีการกำหนดจำนวนข้อบังคับให้มีความเท่าเทียมระหว่างกลุ่มในขั้นตอนนี้ มีเพียงดัชนีความสอดคล้องของข้อมูลที่จะได้รับการตรวจสอบสำหรับประชากรแต่ละคน โดยการพิจารณาค่า χ^2, RMSEA, CFI, NNFI และดัชนีความสอดคล้อง (Fit indexes) เพื่อตรวจสอบว่าโมเดลที่ทำงานร่วมกัน (Combined model) มีความสอดคล้องพอดี (Good model fit) จะเป็นการบ่งชี้ถึงความไม่แปรเปลี่ยนของโมเดลการวัดด้านโครงแบบ ซึ่งค่าที่ได้จะเป็นตัวบ่งชี้ว่ามีความไม่แปรเปลี่ยนด้านโครงแบบเกิดขึ้นหรือไม่ แต่ถ้ามีมีความแปรเปลี่ยนของโครงแบบ (Configural invariance) เกิดขึ้น ซึ่งหมายถึงองค์ประกอบไม่มีมีสอดคล้องกับข้อมูลเชิงประจักษ์ภายในประชากรบางกลุ่ม หรือหลายกลุ่มของประชากร นักวิจัยได้ให้คำแนะนำว่าควรจะกลับไปค้นหาโครงสร้างองค์ประกอบแฝงจริงที่มีในกลุ่มประชากรโดยการใช้วิธีการสำรวจ (Exploratory method) (เช่น วิธี EFA)

3. การทดสอบความเท่าเทียมกันของค่าพารามิเตอร์ ซึ่งประกอบด้วย Factor loading

3.1 ความไม่แปรเปลี่ยนของค่าน้ำหนักองค์ประกอบ (Factor loading invariance) (Horn & McArdle, 1992; Thurston, 1947) หรือเรียกว่า "สมมติฐานความไม่แปรเปลี่ยนขององค์ประกอบแบบอ่อน (Weak factorial invariance)" หรือ "ความไม่แปรเปลี่ยนของเมตริก (Metric Invariance)" ผลการทดสอบถ้าพบว่าจำนวนตัวบ่งชี้สังเกตมีค่าน้ำหนักองค์ประกอบ เท่ากับองค์ประกอบตัวแปรแฝง จะหมายถึงมีความไม่แปรเปลี่ยนของโครงแบบเกิดขึ้น หรืออย่างน้อยอาจจะเกิดขึ้นเพียงบางส่วน (Byrne et al., 1989)

นักวิจัยได้มีการทดสอบสมมติฐานข้อสาม ไว้ดังนี้

$$H_0 : \Lambda_y^{(1)} = \Lambda_y^{(2)} = ... = \Lambda_y^{(g)}$$

และ $\sum_{yy}^{(g)} = \Lambda_y \Phi^{(g)} \Lambda_y^{'} + \Psi_{\varepsilon}^{(g)}$, และ $\mu_y^{(g)} = \tau_y^{(g)} + \Lambda_y \kappa^{(g)}$

ความไม่แปรเปลี่ยนของค่าน้ำหนักองค์ประกอบ (Factor loading) หมายถึง ค่าน้ำหนักองค์ประกอบ ซึ่งเป็นส่วนที่สำคัญที่สุดในการกำหนดข้อบังคับของความเท่าเทียมกัน เพราะค่าน้ำหนักองค์ประกอบ เป็นตัวแทนของความสัมพันธ์ระหว่างตัวบ่งชี้สังเกตขององค์ประกอบแฝง ของความไม่แปรเปลี่ยนของประชากรต่างกลุ่ม ในขั้นตอนนี้ จะมีข้อบังคับกับพารามิเตอร์ประชากร ค่าสถิติไค-สแควร์ ของโมเดลที่ถูกข้อบังคับจะนำมาเปรียบเทียบกับค่า ไค-สแควร์ จากขั้นตอนที่ 2 ที่โมเดลมีความสอดคล้องของข้อมูลในแต่ละคนของกลุ่มประชากร ความแตกต่างของค่านัยสำคัญจะบ่งบอกถึงว่าสมมติฐานที่ถูกปฏิเสธนั้น มีค่าน้ำหนักองค์ประกอบไม่เท่ากันในประชากรต่างกลุ่ม นั่นคือ เกิดความไม่แปรเปลี่ยนของค่าน้ำหนักองค์ประกอบเกิดขึ้น

3.2 ความไม่แปรเปลี่ยนของจุดตัด (Thresholds) (Meredith, 1993;Steencamp & Baumgartner, 1998) หรือความไม่แปรเปลี่ยนด้านสเกล่าร์ (Scalar invariance) หรือความไม่แปรเปลี่ยนขององค์ประกอบแบบแกร่ง (Scalar or strong invariance) หมายถึง การถดถอยของจุดตัดบนตัวบ่งชี้สังเกตที่แสดงความไม่แปรเปลี่ยนของประชากรต่างกลุ่มนอกเหนือไปจากรูปแบบ และความไม่แปรเปลี่ยนของค่าน้ำหนักองค์ประกอบ (Factor loading) เรียกความไม่แปรเปลี่ยนของจุดตัด (Thresholds) เมื่อดำเนินการเปรียบเทียบค่าเฉลี่ยตัวแปรแฝง นำมาใช้เป็นรูปแบบทั่วไปที่มีเงื่อนไขของจุดตัด (Thresholds) ที่รวมถึงตัวชี้วัดที่สังเกต นั่นคือ สมมติฐานข้อที่สี่ จะทำการทดสอบความไม่แปรเปลี่ยนของจุดตัด (Thresholds):

$$H_0 : \tau_y^{(1)} = \tau_y^{(2)} = ... = \tau_y^{(g)}, \text{ เมื่อ } \sum_{yy}^{(g)} = \Lambda_y \Phi^{(g)} \Lambda_y^{'} + \Psi_{\varepsilon}^{(g)}, \text{ และ}$$

$$\mu_y^{(g)} = \tau_y + \Lambda y \kappa^{(g)}$$

ในการทดสอบความไม่แปรเปลี่ยนของจุดตัด (Thresholds) นั้น โดยปกติจะใช้การวิเคราะห์ความแตกต่างของค่าเฉลี่ยของตัวแปรแฝงจากสมการ $\mu_y^{(g)} = \tau_y + \Lambda y \kappa^{(g)}$ สามารถเห็นความแตกต่างของค่าเฉลี่ยของตัวบ่งชี้สังเกต ที่สามารถตรวจสอบกลับไปที่ความแตกต่างขององค์ประกอบตัวแปรแฝง วิธีการโดยการกำหนดความหมายของตัวแปรแฝงในประชากรกลุ่มหนึ่ง (เช่น กำหนดให้มีค่า = 0) และปล่อยให้ค่าเฉลี่ยตัวแปรแฝงของประชากรส่วนที่เหลือถูกประมาณค่าอิสระ ผลการประมาณค่าของค่าเฉลี่ยตัวแปรแฝงในประชากรส่วนที่เหลือ แทนค่าด้วยความแตกต่างของค่าเฉลี่ยระหว่างประชากรกลุ่มเปรียบเทียบ และประชากรกลุ่มอ้างอิง (Hancock, 1997)

3.3 ความไม่แปรเปลี่ยนด้านความเฉพาะตัว (Uniqueness invariance or equal residuals) เป็นการทดสอบความไม่แปรเปลี่ยนระหว่างกลุ่มประชากร เป็นการทดสอบความไม่แปรเปลี่ยนระหว่างกลุ่มประชากรที่เข้มงวด (Strict invariance) เพราะมีโอกาสน้อยมากในการไม่ปฏิเสธสมมติฐาน มีชื่อเรียกอีกชื่อหนึ่งในกรณีที่ความแปรปรวนขององค์ประกอบไม่แปรเปลี่ยนว่า การทดสอบความไม่แปรเปลี่ยนด้านความเที่ยง (Reliability invariance) การทดสอบในขั้นตอนนี้เป็นการทดสอบว่า เทอมความคลาดเคลื่อนของตัวบ่งชี้ในโมเดลการวัดความไม่แปรเปลี่ยนระหว่างกลุ่มประชากร เมื่อ θ คือ เมตริกความแปรปรวนร่วมของความคลาดเคลื่อน สมมติฐานทางสถิติ (นงลักษณ์ วิรัชชัย, 2554) คือ

$$H_0 = \theta_i = \theta_j$$

แต่สมมติฐานข้อ 3.3 ในชุดสมมติฐานนี้ เป็นสมมติฐานข้อสี่ข้อที่ผ่านมาด้วย นั่นคือ

$$H_0: (\sum, \text{Form}, \Lambda, \tau, \theta)_i = (\sum, \text{Form}, \Lambda, \tau, \theta)_j$$

ดังนั้นถ้าไม่ปฏิเสธสมมติฐานข้อที่ 3.3 แสดงว่าไม่ปฏิเสธสมมติฐานทั้งสี่ข้อที่ผ่านมา จึงสรุปได้ว่าโมเดลการวัดไม่แปรเปลี่ยนระหว่างกลุ่มประชากร ทั้งด้านเมตริกความแปรปรวน-ความแปรปรวนร่วม ด้านรูปแบบ น้ำหนักองค์ประกอบ สเกล่าร์ และความคลาดเคลื่อน หรือสรุปว่าโมเดลการวัดไม่แปรเปลี่ยนด้านเมตริก

4. การทดสอบความเท่าเทียมกันของโครงสร้างองค์ประกอบ (Marsh; Hau; Chung; & SIU, 1998, pp. 149-150) หรือความไม่แปรเปลี่ยนขององค์ประกอบแบบแกร่ง (Strong factorial invariance)

4.1 การทดสอบความแปรปรวน-ความแปรปรวนร่วมขององค์ประกอบ (Factor variance-Covariance (Φ) invariance test) ซึ่งความแปรปรวนของความเท่าเทียมกันของเมตริกขององค์ประกอบแฝงในประชากรต่างกลุ่มจะถูกทดสอบ โดยความไม่แปรเปลี่ยนขององค์ประกอบแบบแกร่ง (Strong factorial invariance) ซึ่งสมมติฐานของการทดสอบข้อที่ 5 มีดังนี้

$$H_0 : \Phi^{(1)} = \Phi^{(2)} = ... = \Phi^{(g)}$$

$$\sum_{yy}^{(g)} = \Lambda_y \Phi \Lambda_y^{'} + \Psi_\varepsilon, \text{ และ } \mu_y^{(g)} = \tau_y + \Lambda_y \kappa^{(g)}$$

นักวิจัยให้ความสนใจกับความสัมพันธ์ภายในความเท่าเทียมกันของโครงสร้างองค์ประกอบของประชากรต่างกลุ่ม เมื่อมีมากกว่าหนึ่งองค์ประกอบภายใต้โครงสร้างตัวแปรแฝงที่เสนอโดยทฤษฎี (Byrne, 1998; Cheung & Rensvold, 1999; Raju et al., 2002; Vandenberg & Lance, 2000) ในความสัมพันธ์ภายในโครงสร้าง เมื่อมีมากกว่าหนึ่งปัจจัยพื้นฐานที่แฝงอยู่โครงสร้างที่เสนอโดยทฤษฎี (Byrne, 1998; & Cheung Rensvold, 1999; Raju et al., 2002;Vandenberg & Lance, 2000) ความเท่ากันของโครงสร้างองค์ประกอบแฝง จะเป็นข้อบ่งชี้ว่าองค์ประกอบมีโครงสร้างทฤษฎีเดียวกัน จึงสามารถใช้ตรวจสอบระหว่างประชากรที่แตกต่างกันได้ ซึ่งจะต้องมีการกำหนดข้อสรุปของโครงสร้างที่ถูกต้องโดยคะแนน หรือเครื่องมือที่ใช้ตรวจสอบกับประชากรต่างกลุ่ม

4.2 การทดสอบความเท่าเทียมกันของความคลาดเคลื่อนของความแปรปรวน (Measurement error) ของประชากรต่างกลุ่มที่ถูกทดสอบหลังจากการทดสอบความไม่แปรเปลี่ยนทั้งห้าขั้นตอนแล้ว นั่นคือสมมติฐานข้อที่หก ที่จะทำการทดสอบ

$$H_0 : \Psi_g^{(1)} = \Psi_\varepsilon^{(2)} = ... = \Psi_\varepsilon^{(g)}, \text{เมื่อ } \sum_{yy}^{(g)} \Lambda_y \Phi \Lambda_y^{'} + \Psi_\varepsilon,$$

และ $\mu_y^{(g)} = \tau_y + \Lambda_y \kappa^{(g)}$

เพราะสันนิษฐานในกรอบของ SEM แบบดั้งเดิมที่ความคลาดเคลื่อนของการวัดมีความเป็นอิสระ ความคลาดเคลื่อนในการวัดของเมตริกความแปรปรวน-ความแปรปรวนร่วมโดยปกติแล้วจะเป็นเมตริกทแยงมุม (Diagonal matrix) ดังนั้นในขั้นตอนนี้ องค์ประกอบของเส้นทแยงมุมของความคลาดเคลื่อนในการวัดของเมตริกความแปรปรวน-ความแปรปรวนร่วม จะเป็นตัวทดสอบความไม่แปรเปลี่ยนในประชากรต่างกลุ่ม การทดสอบนี้เป็นจริงตามที่นักวิจัยกล่าวไว้ เพราะมีความจำเป็นที่จะต้องทำการทดสอบความคลาดเคลื่อนของการวัดความแปรปรวน สามารถดูได้ที่การวัดความเชื่อมั่น (Reliability) ของตัวแปรสังเกตที่เป็นตัวบ่งชี้ โดยการให้ Rxx = (1-Var (Error))/ Var (Total) โดยเฉพาะความไม่แปรเปลี่ยนของโมเดลการวัดความคลาดเคลื่อนของความแปรปรวนในประชากรต่างกลุ่มเป็นตัวบ่งชี้ของความเชื่อมั่นของตัวบ่งชี้สังเกต ที่จะแสดงถึงความเท่าเทียมกันประชากรต่างกลุ่ม (ความเท่าเทียมกันของความเชื่อมั่น; Byrne, 1998; Raju et al., 2002)

4.3 การทดสอบความไม่แปรเปลี่ยนของค่าเฉลี่ยองค์ประกอบ (Invariance of factor means) เป็นการทดสอบว่าความแตกต่างของค่าเฉลี่ยองค์ประกอบระหว่างกลุ่มประชากร ซึ่งเป็น

วัตถุประสงค์สำคัญของการวิจัยเชิงทดลอง สมมติฐานหลักทางสถิติที่ทดสอบ คือ

$$H_0: \alpha_i = \alpha_j ; \kappa_i = \kappa_j$$

เมื่อ α (Alpha) และ κ (Kappa) เป็นค่าเฉลี่ยของตัวแปรแฝงภายนอกและภายในตามลำดับ แต่สมมติฐานในข้อนี้เป็นสมมติฐานรวมสมมติฐานก่อนหน้านี้ทั้งหมดด้วยกัน ดังนั้น ถ้าไม่ปฏิเสธสมมติฐานทั้งชุด และสามารถสรุปได้ว่า โมเดลการวัดไม่แปรเปลี่ยนระหว่างกลุ่มประชากร ทั้งด้านเมตริก ความแปรปรวน-ความแปรปรวนร่วม ด้านรูปแบบ ด้านน้ำหนักองค์ประกอบ ด้านสเกล่าร์ ด้านความคลาดเคลื่อน ด้านความแปรปรวนและความแปรปรวนร่วมขององค์ประกอบ และด้านค่าเฉลี่ยองค์ประกอบในโมเดลการวัด (นงลักษณ์ วิรัชชัย, 2554)

ความเข้มงวดในทั้งสี่ขั้นตอน ควรจะดำเนินการอย่างใดอย่างหนึ่งก่อนที่จะเริ่มทำการเปรียบเทียบค่าเฉลี่ยตัวแปรแฝง ทั้งนี้ขึ้นอยู่กับคำถามการวิจัยเฉพาะ บางครั้งนักวิจัยยังให้ความสนใจในการทดสอบ MI อื่น ๆ เพิ่มเติม เช่น ค่าเฉลี่ยองค์ประกอบ (K) หรือการทดสอบความไม่แปรเปลี่ยนขององค์ประกอบความแปรปรวนร่วมระหว่างกลุ่มประชาชากรที่แตกต่างกันที่ไม่จำเป็นต้องใช้การวัดความไม่แปรเปลี่ยน (Millsap & Tein, 2004) ในแต่ละขั้นตอนของการทดสอบ MI ข้อมูลตัวบ่งชี้ความสอดคล้องของโมเดลได้จากซอฟต์แวร์ทั่วไป (เช่นโปรแกรม LISREL, Mplus, EQS ฯลฯ) จะรวบรวมทั้งโมเดลรุ่นปัจจุบัน และโมเดลที่มีการบังคับพารามิเตอร์ให้เท่ากัน ระดับของการลดลงของความสอดคล้องจะได้จากความแตกต่างของความสอดคล้องของข้อมูลในโมเดล (Chi-square) โดยองศาอิสระที่มีความแตกต่างกัน (ซึ่งนอกจากนี้ยังมีความแตกต่างของจำนวนการประมาณค่าพารามิเตอร์อิสระระหว่างสองโมเดล)

Timmons (2010) ได้กล่าวถึงความไม่แปรเปลี่ยนขององค์ประกอบ (Factorial invariance) ว่าเมื่อใดที่มีการวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ (Multiple group confirmatory factor analysis) สิ่งที่ต้องดำเนินการ คือ การสร้างความไม่แปรเปลี่ยนขององค์ประกอบ เพื่อที่จะทำการวัดโครงสร้างเดียวกันในประชากรต่างกลุ่ม เช่น ในเพศหญิง และเพศชาย หรือในช่วงเวลาที่แตกต่างกัน ความไม่แปรเปลี่ยนขององค์ประกอบ เป็นสมมติฐานที่สร้างขึ้นมาเพื่อทำการเปรียบเทียบระหว่างกลุ่ม หรือระหว่างช่วงเวลา เมื่อมีการวิเคราะห์ MGCFA ข้อสมมติฐานนี้สามารถทำการทดสอบได้โดยตรงด้วย การทดสอบความไม่แปรเปลี่ยนขององค์ประกอบ หรือความเท่าเทียมกัน และความไม่แปรเปลี่ยนของเมตริกซึ่ง Timmons (2010) ได้สรุปความไม่แปรเปลี่ยนขององค์ประกอบไว้ ดังนี้

1. ความไม่แปรเปลี่ยนองค์ประกอบด้านโครงแบบ (Configural invariance) ใช้ตรวจสอบกับประชากรต่างกลุ่ม (Across groups) มีทั้งรูปแบบของการกำหนดค่าคงที่ (Fixed) และค่าอิสระของพารามิเตอร์ ให้มีความเท่าเทียมกัน ความไม่เปลี่ยนองค์ประกอบด้านโครงแบบ

(Configural invariance) (Timmons, 2010) ต้องมีการตรวจสอบค่า X^2, RMSEA, CFI, NNFI และดัชนีความสอดคล้อง (Fit indexes) เพื่อตรวจสอบว่าโมเดลที่ทำงานร่วมกัน (Combined model) มีความสอดคล้องพอดี (Good model fit) จะเป็นการบ่งชี้ถึงความไม่แปรเปลี่ยนขององค์ประกอบด้านโครงแบบ

2. ความไม่แปรเปลี่ยนขององค์ประกอบแบบอ่อน (Weak factorial invariance) ใช้ตรวจสอบกับประชากรต่างกลุ่ม (Across groups) จะต้องมีความสอดคล้องของค่าน้ำหนักองค์ประกอบ (Factor loadings) และมีความเท่าเทียมกัน ให้ทำการเปรียบเทียบค่า RMSEA และช่วงความเชื่อมั่นของ RMSEA ในโมเดลความไม่แปรเปลี่ยนด้านโครงแบบอ่อน ถ้าพบว่าค่า RMSA ตกอยู่ในช่วงของความเชื่อมั่นช่วงอื่น จะเป็นการบ่งชี้ถึงความไม่แปรเปลี่ยนขององค์ประกอบแบบอ่อน นอกจากนี้ยังตรวจสอบการเปลี่ยนแปลงของค่า CFI สำหรับโมเดลความไม่แปรเปลี่ยนขององค์ประกอบด้านโครงแบบ และความไม่แปรเปลี่ยนขององค์ประกอบแบบอ่อน ถ้าพบว่าค่า CFI มีค่าน้อยกว่า .01 จะเป็นการบ่งชี้ว่ามีความไม่แปรเปลี่ยนขององค์ประกอบแบบอ่อน (Weak factorial invariance) เกิดขึ้น (Cheung & Rensvold, 2002)

3. ความไม่แปรเปลี่ยนขององค์ประกอบแบบแกร่ง (Strong factorial invariance) ใช้ตรวจสอบกับประชากรต่างกลุ่ม (Across groups) จะต้องมีความสอดคล้องของค่าเฉลี่ยตัวบ่งชี้ (Indicator means) และมีความเท่าเทียมกัน ทำการเปรียบเทียบค่า RMSEA และช่วงความเชื่อมั่น RMSEA ของโมเดลความไม่แปรเปลี่ยนขององค์ประกอบแบบอ่อน (Weak) และแบบแกร่ง (Strong) ถ้าค่า RMSEA ตกอยู่ในช่วงของความเชื่อมั่นช่วงอื่น จะบ่งชี้ว่ามีความไม่แปรเปลี่ยนขององค์ประกอบแบบแกร่ง (Strong factorial invariance) และตรวจสอบค่า CFI สำหรับโมเดลความไม่แปรเปลี่ยนขององค์ประกอบแบบอ่อน (Weak) และแบบแกร่ง (Strong) ถ้าพบว่า ค่า CFI มีค่าน้อยกว่า .01 จะเป็นการบ่งชี้ว่ามีความไม่แปรเปลี่ยนขององค์ประกอบแบบแกร่ง (Strong factorial invariance) เกิดขึ้น (Cheung & Rensvold, 2002)

4. ความไม่แปรเปลี่ยนขององค์ประกอบแบบเข้มงวด (Strict factorial invariance**)** ใช้ตรวจสอบกับประชากรต่างกลุ่ม (Across groups) จะต้องมีความสอดคล้องของตัวบ่งชี้เศษเหลือ (indicator residuals) ที่มีความเท่าเทียมกัน (ระดับนี้ไม่ค่อยแนะนำเพราะหลักเกณฑ์มีความเข้มงวดมากเกินไปยากต่อการปฏิบัติ)

การวัดความไม่แปรเปลี่ยนเป็นบางส่วน (Partial measurement invariance: PMI)

อ้างอิงถึงกรอบแนวคิดของ PMI ที่ถูกเสนอครั้งแรกโดย Byrne et al. (1989) ที่มีความพยายามที่จะละเมิดโดยการกำหนดแหล่งของการเกิดความไม่แปรเปลี่ยนของโมเดลการวัด (Measurement invariance) (เช่น การกำหนดเฉพาะค่าน้ำหนักองค์ประกอบ (Factor loading)

และจุดตัด (Thresholds) โดยการปล่อยให้มีการประมาณค่าอิสระของประชากรต่างกลุ่ม เก็บส่วนที่เหลือของความไม่แปรเปลี่ยนของพารามิเตอร์ และดำเนินการขั้นตอนต่อไป ผลกระทบต่อประโยชน์ และความเหมาะสมของ PMI (Cheung & Rensvold, 1999) มีข้อโต้เถียงว่ามันไม่มีความหมายถ้าดำเนินการเปรียบเทียบค่าเฉลี่ยของตัวแปรแฝงเมื่อทำ PMI (Bollen, 1989; Meredith, 1993; Ployhart & Oswald, 2004) มีข้อโต้เถียงว่าควรจะเก็บข้อสอบที่มีพารามิเตอร์ที่ไม่คงที่ไว้ และประมาณค่าอิสระก่อนที่จะทำการศึกษาความแตกต่างของค่าเฉลี่ยตัวแปรแฝง ปัญหาหนึ่งที่น่าสนใจเกี่ยวกับ PMI คือ สิ่งที่ต้องทำเมื่อเกิดความแปรเปลี่ยน (Noninvariance) ที่พบหลังจากการทบทวนวรรณกรรมเกี่ยวกับ MI ที่ผ่านมา

Byrne et al. (1989) ได้ระบุว่าไม่มีหลักฐานงานวิจัยที่พบว่า มีการดำเนินการทดสอบ PMI เมื่อสมมติฐานได้รับการปฏิเสธ พบว่า บางการศึกษา MI เพียงแต่ละเว้นข้อสอบที่มีความแปรเปลี่ยนไว้ (Noninvariance) และใช้ข้อสอบทั้งหมดแม้จะมีความจริงที่ว่าไม่ใช่ทั้งหมดของข้อสอบที่แสดงความไม่แปรเปลี่ยนของโมเดลการวัด แน่นอนว่ามันจะมีผลกระทบต่อการวัดคุณสมบัติด้านจิตวิทยา ในทางกลับกัน Cheung and Rensvold (1999) ระบุว่าจะได้รับข้อมูลที่ประโยชน์เกี่ยวกับความแตกต่างของค่าเฉลี่ยระหว่างกลุ่มประชากรที่แตกต่างกันจากการปฏิเสธสมมติฐานของ MI ยกตัวอย่าง เช่น Ellis (1989) ได้ใช้วิธีการตอบสนองข้อสอบ (IRT) เพื่อสำรวจว่าภาษา หรือปัญหาทางวัฒนธรรมที่เกี่ยวข้องกับการพบข้อสอบที่มีความแปรเปลี่ยน การลบข้อคำถามของข้อสอบเป็นวิธีที่จะจัดการกับปัญหานี้ แต่ได้รับคำแนะนำว่าข้อสอบทุกข้อถูกสันนิษฐานว่ามีพื้นฐานการพัฒนาบนพื้นฐานของทฤษฎี และควรจะมีส่วนร่วมในการวัดของโครงสร้างของตัวแปรแฝงทั้งหมด

การตัดข้อสอบออกไปอาจจะเป็นการทำลายความเที่ยงตรงเชิงโครงสร้างอย่างรุนแรง และทฤษฎีที่อยู่เบื้องหลังความเป็นไปได้ของโครงสร้าง (Cheung & Rensvold, 1999) นอกจากนี้ การพัฒนาข้อสอบต้องสูญเสียค่าใช้จ่ายที่แพง สูญเสียทั้งเวลา และเงิน ดังนี้การทดสอบ PMI ตามคำแนะนำของนักวิจัยจะกระทำได้ก็ต่อเมื่อการทดสอบ MI เต็มรูปแบบไม่สามารถกระทำได้เท่านั้น (Byrne, 1993; Jackson, Wall, Martin & Davids, 1993; Reise et al., 1993; Riordan & Vanderberg, 1994) เพื่อที่จะดำเนินการศึกษา และทำการทดสอบ MGCFA อย่างถูกต้อง ความเป็นไปได้ของการทดสอบแบบ PMI จะทำให้ได้รับผลการค้นหาที่มีความหมาย สามารถระบุรายการอ้างอิงของความไม่แปรเปลี่ยนของข้อสอบ ดังนั้นจึงต้องมีการกำหนดความเท่าเทียมกันของประชากรต่างกลุ่ม Cheung and Rensvold (1999) ได้เสนอคำถามที่ถูกใช้ในการสำหรับการระบุเพื่ออ้างอิงข้อสอบในการศึกษา MGCFA นั่นคือนักวิจัย ตามปกติแล้วจะเลือกข้อสอบที่มีการปรับวัตถุประสงค์ตามกระบวนการที่เป็นมาตรฐานในกลุ่มประชากรที่ต่างกัน หมายถึงการนำข้อสอบที่ไม่มีความแปรเปลี่ยนมาใช้ในการอ้างอิง

ตอนที่ 8 การทดสอบวัดผลสัมฤทธิ์ระดับชาติของนักเรียน O-NET

O-NET ย่อมาจาก Ordinary national education test ซึ่งหมายถึง การวัดผลสัมฤทธิ์รวบยอดของผู้เรียนที่ต้องการวัดผ่านการจัดสอบการทดสอบในระดับชาติ โดยมีสถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน) เป็นผู้ดำเนินการออกข้อสอบ และจัดสอบทั่วประเทศ มีวัตถุประสงค์เพื่อทดสอบความรู้ และความคิดของนักเรียนชั้นประถมศึกษาปีที่ 6, ชั้นมัธยมศึกษาปีที่ 3 และชั้นมัธยมศึกษาปีที่ 6 นำผลการทดสอบไปใช้เป็นองค์ประกอบหนึ่งในการจบการศึกษาตามหลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน พุทธศักราช 2551 เพื่อนำผลการทดสอบไปใช้ในการปรับปรุงคุณภาพการเรียน การสอนของโรงเรียน เพื่อนำผลการทดสอบไปใช้ในการประเมินผลการเรียนรู้ของนักเรียนระดับชาติ และนำไปใช้ในวัตถุประสงค์อื่น ๆ

ในปัจจุบันมีการจัดสอบแบบทดสอบทางการศึกษาระดับชาติสามัญ, แบบทดสอบทางการศึกษาระดับชาติสำหรับการศึกษานอกระบบโรงเรียน แบบทดสอบทางการศึกษาระดับชาติด้านอาชีวศึกษา, แบบทดสอบทางการศึกษาระดับชาติด้านอุดมศึกษา, แบบทดสอบทางการศึกษาระดับชาติด้านพุทธศาสนา และแบบทดสอบทางการศึกษาระดับชาติด้านอิสลาม โดยในอดีตมีการจัดการทดสอบทางการศึกษาระดับชาติขั้นสูงอีกด้วย ซึ่งในปัจจุบันมีการใช้แบบทดสอบความถนัดทั่วไป (GAT) และแบบทดสอบความถนัดทางวิชาการและวิชาชีพ (PAT) มาใช้แทน การทดสอบทางการศึกษาระดับสามัญ (Ordinary national education test: O-NET) เป็นการสอบความรู้รวบยอดปลายช่วงชั้น 6 ภาคเรียน โดยสอบชั้นประถมศึกษาปีที่ 3, 6 มัธยมศึกษาปีที่ 3 และ 6 ตามมาตรฐานการเรียนรู้ของหลักสูตรการศึกษาขั้นพื้นฐาน พ.ศ. 2544 ทำการทดสอบความรู้ในกลุ่มสาระต่าง ๆ รวม 8 กลุ่มสาระ ได้แก่ ภาษาไทย คณิตศาสตร์ วิทยาศาสตร์ สังคมศึกษา ศาสนาและวัฒนธรรม สุขศึกษา และพลศึกษา ศิลปะการงานอาชีพและเทคโนโลยี และภาษาต่างประเทศ มีช่วงเวลาสอบในเดือนกุมภาพันธ์ของทุกปี เป็นการสอบประจำปีเพียงครั้งเดียว และผลการทดสอบจะเป็นส่วนหนึ่งของคะแนนที่ใช้ในการสมัครศึกษาต่อในระดับมหาวิทยาลัย

การคัดเลือกบุคคลเข้าศึกษาในสถาบันอุดมศึกษาในระบบกลาง (Admissions) ที่ประชุมอธิการบดีแห่งประเทศไทย กำหนดให้พิจารณาคัดเลือกผู้สมัครโดยใช้

1. ผลการเรียนเฉลี่ยสะสมตลอดหลักสูตรมัธยมศึกษาตอนปลายหรือเทียบเท่า มีค่าน้ำหนัก ร้อยละ 20
2. ผลการสอบทางการศึกษาขั้นพื้นฐาน (O-NET) ให้ค่าน้ำหนักร้อยละ 30
3. ผลการสอบความถนัดทั่วไป (GAT) ให้ค่าน้ำหนักร้อยละ 10-50
4. ผลการทดสอบความถนัดทางวิชาชีพ และวิชา (PAT) ให้ค่าน้ำหนักร้อยละ 0 -40

5. ผลการสอบสัมภาษณ์และตรวจร่างกาย ใช้เป็นข้อมูลประกอบการพิจาณาความพร้อมและความเหมาะสมก่อนรับเข้าศึกษา ไม่คิดค่าน้ำหนัก

ตารางที่ 2-2 แสดงเนื้อหาการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET)

รหัสวิชา 01 วิชาภาษาไทย	**รหัสวิชา 02 วิชา สังคมศึกษา ศาสนาและวัฒนธรรม**
สาระที่ 1: การอ่าน	สาระที่ 1: ศาสนา ศีลธรรม จริยธรรม
สาระที่ 2: การเขียน	สาระที่ 2: หน้าที่พลเมือง วัฒนธรรม และการดำเนินชีวิตในสังคม
สาระที่ 3: การฟัง การดู และการพูด	สาระที่ 3: เศรษฐศาสตร์
สาระที่ 4: หลักการใช้ภาษา	สาระที่ 4: ประวัติศาสตร์
สาระที่ 5: วรรณคดิ และวรรณกรรม	สาระที่ 5: ภูมิศาสตร์
รหัสวิชา 03 วิชาภาษาอังกฤษ	**รหัสวิชา 04 วิชาคณิตศาสตร์**
สาระที่ 1: ภาษาเพื่อการสื่อสาร	สาระที่ 1: การอ่าน
สาระที่ 2: ภาษาและวัฒนธรรม	สาระที่ 2: การเขียน
สาระที่ 3: ภาษากับความสัมพันธ์กับกลุ่มสาระการเรียนรู้อื่น	สาระที่ 3: การฟัง การดู และการพูด
สาระที่ 4: ภาษากับความสัมพันธ์กับชุมชนและโลก	สาระที่ 4: หลักการใช้ภาษา
	สาระที่ 5: วรรณคดิ และวรรณกรรม
รหัสวิชา 05 วิชาวิทยาศาสตร์	**รหัสวิชา 06 สุขศึกษาและพลศึกษา**
สาระที่ 1: สิ่งมีชีวิตกับกระบวนการดำรงชีวิต	สาระที่ 1: การเจริญเติบโตและพัฒนาการของมนุษย์
สาระที่ 2: ชีวิตกับสิ่งแวดล้อม	สาระที่ 2: ชีวิตและครอบครัว
สาระที่ 3: สารและสมบัติของสาร	สาระที่ 3: การเคลื่อนไหว การออกกำลังกาย เกม กีฬาไทย และกีฬาสากล
สาระที่ 4: แรงและการเคลื่อนที่	สาระที่ 4: การเสริมสร้างสุขภาพ สมรรถภาพ และการป้องกันโรค
สาระที่ 5: พลังงาน	สาระที่ 5: ความปลอดภัยในชีวิต
สาระที่ 6: กระบวนการเปลี่ยนแปลงของโลก	
สาระที่ 7: ดาราศาสตร์และอวกาศ	
สาระที่ 8: ธรรมชาติของวิทยาศาสตร์และ	

ตารางที่ 2-2 (ต่อ)

รหัสวิชา 01 วิชาภาษาไทย	รหัสวิชา 02 วิชา สังคมศึกษา ศาสนาและวัฒนธรรม
รหัสวิชา 07 วิชาศิลปะ	**รหัสวิชา 08 วิชาการงานอาชีพและเทคโนโลยี**
สาระที่ 1: ทัศนศิลป์	สาระที่ 1: การดำรงชีวิตและครอบครัว
สาระที่ 2: ดนตรี	สาระที่ 2: การอาชีพ
สาระที่ 3: นาฏศิลป์	สาระที่ 3: การออกแบบและเทคโนโลยี
	สาระที่ 4: เทคโนโลยีสารสนเทศ
	สาระที่ 5: เทคโนโลยีเพื่อการทำงานและอาชีพ

และในวันที่ 26 มิถุนายน พ.ศ. 2558 ได้มีประกาศกระทรวงศึกษาธิการ เรื่องการกำหนดวิชาในการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน O-NET มีมติให้สถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน) ปรับลดการทดสอบการศึกษาระดับชาติขั้นพื้นฐาน O-NET จาก 8 กลุ่มสาระการเรียนรู้ เหลือ 5 กลุ่มสาระการเรียนรู้ ได้แก่ วิชาคณิตศาสตร์ วิชาวิทยาศาตร์ วิชาภาษาอังกฤษ วิชาภาษาไทย และวิชาสังคมศึกษา ศาสนาและวัฒนธรรม สำหรับอีก 3 กลุ่มสาระการเรียนรู้ ได้แก่ วิชาสุขศึกษาและพลศึกษา วิชาศิลปะและวิชาการงานอาชีพและเทคโนโลยี ให้โรงเรียนเป็นผู้ออกข้อสอบ และจัดสอบเอง โดยเริ่มตั้งแต่ปีการศึกษา 2558 การสอบเป็นบริการของรัฐให้แก่นักเรียนทุกคนโดยไม่ต้องเสียค่าใช้จ่ายใด ๆ ลักษณะของข้อสอบ และการประเมินผล O-NET ประกอบด้วย

1. แบบทดสอบจะมีทั้งปรนัย และอัตนัย ในอัตาส่วนระหว่าง 80%-90%: 10%-20% ข้อสอบแบบปรนัยจะเป็นข้อสอบแบบ 4 ตัวเลือก สำหรับข้อสอบอัตนัยจะเป็นข้อสอบแบบเขียนคำตอบสั้น ๆ (Short answer)
2. เวลาในการทำข้อสอบวิชาละ 2 ชั่วโมง
3. ข้อสอบแต่ละข้อ คะแนนอาจจะไม่เท่ากันขึ้นอยู่กับความยากง่ายของข้อสอบ
4. ข้อสอบครอบคลุมสาระและทักษะสำคัญของ 5 กลุ่มสาระการเรียนรู้

ในการวิจัยครั้งนี้ ได้ใช้คะแนนวัดผลสัมฤทธิ์ระดับชาติของนักเรียน O-NET ชั้นมัธยมศึกษาปีที่ 6 ปีการศึกษา 2558 มาใช้ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ โดยผลการทดสอบ O-NET ปีการศึกษา 2558 มีรายละเอียด ดังต่อไปนี้

การทดสอบวัดผลสัมฤทธิ์ระดับชาติของนักเรียน O-NET ปีการศึกษา 2558

สถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน) หรือเรียกโดยย่อ “สทศ” จัดการทดสอบระดับชาติขั้นพื้นฐาน O-NET ชั้นมัธยมศึกษาปีที่ 6 เมื่อวันที่ 6-7 กุมภาพันธ์ พ.ศ. 2559 และ สทศ. ได้รายงานผลค่าสถิติพื้นฐานเพิ่มเติม โดยมีการจำแนกตัวแปร คือ ค่าเฉลี่ย ($\overline{X}$) ตามสังกัดขนาดโรงเรียน ที่ตั้งและภูมิภาค สรุปได้ดังนี้

ตารางที่ 2-3 จำนวนข้อสอบ O-NET มัธยมศึกษาปีที่ 6 แยกตามรูปแบบของข้อสอบ ประจำปีการศึกษา 2558

ที่	รูปแบบ	วิชา									
		ภาษาไทย		สังคมศึกษา		อังกฤษ		คณิตศาสตร์		วิทยาศาสตร์	
		จำนวนข้อ	คะแนน	จำนวนข้อ	คะแนน	จำนวนข้อ	คะแนน	จำนวนข้อ	คะแนน	จำนวนข้อ	คะแนน
1	ปรนัย										
	1.1 4 ตัวเลือก 1 คำตอบ	-	-	-	-	-	-	-	-	-	-
	1.2 4 ตัวเลือก 2 คำตอบ	-	-	-	-	-	-	-	-	-	-
	1.3 5 ตัวเลือก 1 คำตอบ	0	90	80	80	80	80	32	80	80	80
	1.4 5 ตัวเลือก 2 คำตอบ	-	-	10	20	-	-	-	-	-	-
2	ปรนัย หลายตัวเลือก 1 คำตอบ	-	-	-	-	10	20	-	-	-	-
3	ปรนัย หลายตัวเลือก มากกกว่า 1 คำตอบ	-	-	-	-	-	-	-	-	10	20
4	แบบเลือกคำตอบจากแต่ละหมวดที่สัมพันธ์กัน	10	10	-	-	-	-	-	-	-	-
	แบบระบายคำตอบที่เป็นค่า/ ตัวเลข	-	-	-	-	-	-	8	20	-	-
	รวม	80	90	90	100	90	100	40	100	90	100
	จำนวนเวลาที่ใช้สอบ	2 ชั่วโมง		2 ชั่วโมง		2 ชั่วโมง		2 ชั่วโมง		2 ชั่วโมง	

ข้อมูล ณ วันที่ 21 สิงหาคม พ.ศ. 2558

จากตารางที่ 2-3 พบว่า ในแต่ละวิชามีคะแนนวิชาละ 100 คะแนน แต่จำนวนข้อสอบจะมีความแตกต่างกัน ลักษณะข้อสอบโดยส่วนใหญ่เป็นแบบปรนัย 5 ตัวเลือก 1 คำตอบ ใช้เวลาในการสอบแต่ละวิชา วิชาละ 2 ชั่วโมง เท่ากันทุกวิชา

ตารางที่ 2-4 ค่าคะแนนเฉลี่ยของผลการทดสอบ O-NET ชั้นมัธยมศึกษาปีที่ 6 ปีการศึกษา 2558 จำแนกตามสังกัด

สังกัด	ตัวแปร	วิชา				
		ภาษาไทย	สังคมศึกษา	ภาษาอังกฤษ	คณิตศาสตร์	วิทยาศาสตร์
สำนักงานคณะกรรมการการศึกษา	จำนวนผู้เข้าสอบ (N)	336,246	336,907	336,869	337,049	336,316
ขั้นพื้นฐาน	คะแนนเฉลี่ย (Mean)	49.95	40.00	24.68	26.65	33.55
สำนักบริหารคณะกรรมการส่งเสริม	จำนวนผู้เข้าสอบ (N)	52,536	52,688	52,644	52,685	52,552
การศึกษาเอกชน	คะแนนเฉลี่ย (Mean)	47.70	38.65	27.49	26.68	32.91
สำนักงานคณะกรรมการการอุดมศึกษา	จำนวนผู้เข้าสอบ (N)	4,270	4,273	4,278	4,275	4, 270
(โรงเรียนสาธิต)	คะแนนเฉลี่ย (Mean)	64.87	46.33	46.19	45.85	42.57
ในกำกับกระทรวงศึกษาธิการ	จำนวนผู้เข้าสอบ (N)	242	242	242	242	242
(โรงเรียนมหิดลวิทยานุสรณ์)	คะแนนเฉลี่ย (Mean)	81.49	58.32	75.53	92.21	67.22
สำนักการศึกษา กทม.	จำนวนผู้เข้าสอบ (N)	859	862	862	862	860
	คะแนนเฉลี่ย (Mean)	45.25	36.57	21.86	22.66	30.29
สำนักการศึกษาพัทยา	จำนวนผู้เข้าสอบ (N)	213	214	214	214	213
	คะแนนเฉลี่ย (Mean)	51.85	39.46	25.73	26.34	32.37
การส่งเสริมการปกครองท้องถิ่น	จำนวนผู้เข้าสอบ (N)	23, 237	23, 274	23, 274	23, 276	23, 240
	คะแนนเฉลี่ย (Mean)	43.42	37.03	20.25	21.98	31.03
สำนักการศึกษาพิเศษ	จำนวนผู้เข้าสอบ (N)	74	74	74	74	74
	คะแนนเฉลี่ย (Mean)	31.04	30.32	18.70	17.11	27.04

วิชาภาษาไทย มีผู้เข้าสอบจำนวน 422,625 คน มีคะแนนเฉลี่ยระดับประเทศ 49.36 ปรากฏว่าคะแนนเฉลี่ยในสังกัด ในกำกับของกระทรวงศึกษาธิการ (โรงเรียนมหิดลวิทยานุสรณ์) มีผู้เข้าสอบ 242 คน คะแนนเฉลี่ย 81.49 สังกัดสำนักงานคณะกรรมการการอุดมศึกษา (โรงเรียนสาธิต) มีผู้เข้าสอบ 4,270 คน คะแนนเฉลี่ย 64.87 สังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน มีผู้เข้าสอบ 336,246 คน คะแนนเฉลี่ย 49.95 และสังกัดสำนักบริหารงานคณะกรรมการส่งเสริมการศึกษาเอกชน มีผู้เข้าสอบ 52,536 คน คะแนนเฉลี่ย 47.70 ตามลำดับ

วิชาสังคมศึกษา ศาสนา และวัฒนธรรม มีผู้เข้าสอบ 423,519 คน **มี**คะแนนเฉลี่ยระดับประเทศ 39.70 ปรากฏว่าคะแนนเฉลี่ยในสังกัด ในกำกับกระทรวงศึกษาธิการ (โรงเรียนมหิดลวิทยานุสรณ์) ผู้เข้าสอบ 242 คน คะแนนเฉลี่ย 58.32 สังกัดสำนักงานคณะกรรมการการอุดมศึกษา (โรงเรียนสาธิต) มีผู้เข้าสอบ 4,273 คน คะแนนเฉลี่ย 46.33 สังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน มีผู้เข้าสอบ 336,907 คน คะแนนเฉลี่ย 40.00 และสังกัดสำนักบริหารงานคณะกรรมการส่งเสริมการศึกษาเอกชน มีผู้เข้าสอบ 52,688 คน คะแนนเฉลี่ย 38.65 ตามลำดับ

วิชาภาษาอังกฤษ มีผู้เข้าสอบ 423,417 คน มีคะแนนเฉลี่ยระดับประเทศ 24.98 ปรากฏว่าคะแนนเฉลี่ยในสังกัด ในกำกับกระทรวงศึกษาธิการ (โรงเรียนมหิดลวิทยานุสรณ์) ผู้เข้าสอบ 242 คน คะแนนเฉลี่ย 75.53 สังกัดสำนักงานคณะกรรมการการอุดมศึกษา (โรงเรียนสาธิต) มีผู้เข้าสอบ 4,278 คน คะแนนเฉลี่ย 46.19 สังกัดสำนักงานบริหารงานคณะกรรมการส่งเสริมการศึกษาเอกชน มีผู้เข้าสอบ 52,644 คน คะแนนเฉลี่ย 27.49 และสังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน มีผู้เข้าสอบ มีผู้เข้าสอบ 336,869 คน คะแนนเฉลี่ย 24.68 ตามลำดับ

วิชาคณิตศาสตร์ มีผู้เข้าสอบ 423,654 คน มีคะแนนเฉลี่ยระดับประเทศ 26.59 ปรากฏว่าคะแนนเฉลี่ยในสังกัด ในกำกับกระทรวงศึกษา (โรงเรียนมหิดลวิทยานุสรณ์) ผู้เข้าสอบ 242 คน คะแนนเฉลี่ย 92.21 สังกัดสำนักงานคณะกรรมการการอุดมศึกษา (โรงเรียนสาธิต) มีผู้เข้าสอบ 4,275 คน คะแนนเฉลี่ย 45.85 สังกัดสำนักบริหารงานคณะกรรมการส่งเสริมการศึกษาเอกชน มีผู้เข้าสอบ 52,685 คน คะแนนเฉลี่ย 26.68 และสังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน มีผู้เข้าสอบ 337,049 คน คะแนนเฉลี่ย 26.65 ตามลำดับ

วิชาวิทยาศาสตร์ มีผู้เข้าสอบ 422,718 คน มีคะแนนเฉลี่ยระดับประเทศ 33.40 ปรากฏว่าคะแนนเฉลี่ยในสังกัด ในกำกับกระทรวงศึกษาธิการ (โรงเรียนมหิดลวิทยานุสรณ์) ผู้เข้าสอบ 242 คน คะแนนเฉลี่ย 67.22 สังกัดสำนักงานคณะกรรมการการอุดมศึกษา (โรงเรียนสาธิต) มีผู้เข้าสอบ 4,270 คน คะแนนเฉลี่ย 42.57 สังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน

มีผู้เข้าสอบ 336,316 คน คะแนนเฉลี่ย 33.55 ตามลำดับ และสังกัดสำนักบริหารงานคณะกรรมการส่งเสริมการศึกษาเอกชน มีผู้เข้าสอบ 52,552 คน คะแนนเฉลี่ย 32.91 ตามลำดับ

จำแนกตามขนาดโรงเรียน ที่ตั้ง และภูมิภาค

ตารางที่ 2-5 แสดงค่าคะแนนเฉลี่ยของผลการทดสอบ O-NET ชั้นมัธยมศึกษาปีที่ 6 ปีการศึกษา 2558 จำแนกตามขนาดโรงเรียน ที่ตั้ง และภูมิภาค

วิชา	ขนาดโรงเรียน				ที่ตั้ง		กรุงเทพฯ และภูมิภาค							ระดับประเทศ
	เล็ก	กลาง	ใหญ่	ใหญ่พิเศษ	เมือง	นอกเมือง	กรุงเทพฯ	กลาง	ตะวันออก	เหนือ	ตะวันออกเฉียงเหนือ	ใต้	ตะวันตก	
ภาษาไทย	40.55	43.58	46.91	54.22	53.84	46.00	56.42	50.10	50.55	50.73	46.37	47.54	49.75	49.36
ผู้เข้าสอบ (N)	21,685	98,770	96,413	205,747	181,465	241,160	50,364	28,731	28,732	41, 818	136,773	60,062	18, 180	422,625
สังคมศึกษาฯ	36.09	37.27	38.59	41.76	41.47	38.36	42.23	39.83	40.12	40.63	38.64	38.94	39.66	39.70
ผู้เข้าสอบ (N)	21,724	98,990	96,590	206,215	182,091	241,428	50,095	86,939	28,798	41,839	136,889	60,119	18, 240	423,519
ภาษาอังกฤษ	18.95	20.65	22.70	28.76	29.11	21.87	34.42	25.15	25.47	25.61	21.94	23.44	26.68	24.98
ผู้เข้าสอบ (N)	21,705	98,926	96,592	206,194	181,944	241,473	50,560	86,899	28,827	41,854	136,948	60,134	18,222	423,417
คณิตศาสตร์	20.40	22.37	24.17	30.40	30.48	23.66	33.66	27.04	27.35	27.70	23.69	25.54	26.30	26.59
ผู้เข้าสอบ (N)	21,724	98,999	96,624	206,307	182,102	241,552	50,663	86,933	28,829	41,865	136,948	60,172	18,224	423,654
วิทยาศาสตร์	30.11	31.31	32.20	35.32	35.18	32.06	36.08	33.52	33.75	34.48	32.27	32.70	33.26	33.40
ผู้เข้าสอบ (N)	21,675	98,768	96,454	205,801	181,504	241,214	50,332	86,747	28,745	41,837	136,777	60,092	18,188	422,718

ข้อมูล ณวันที่ 21 สิงหาคม พ.ศ. 2558

จำแนกตามขนาดโรงเรียน

วิชาภาษาไทย มีผู้เข้าสอบ 422,625 คน มีคะแนนระดับประเทศ 49.36 และเมื่อพิจาณาตามขนาดโรงเรียน พบว่า คะแนนเฉลี่ยในโรงเรียนขนาดใหญ่พิเศษ มีผู้เข้าสอบ 205,747 คน คะแนนเฉลี่ย 54.22 สูงกว่าโรงเรียนขนาดใหญ่ มีผู้เข้าสอบ 96,413 คน คะแนนเฉลี่ย 46.91 ขนาดกลาง มีผู้เข้าสอบ 98,770 คน คะแนนเฉลี่ย 43.58 และขนาดเล็ก มีผู้เข้าสอบ 21,695 คน คะแนนเฉลี่ย 40.55 ตามลำดับ

วิชาสังคมศึกษา ศาสนาและวัฒนธรรม มีผู้เข้าสอบ 423,519 คน มีคะแนนเฉลี่ยระดับประเทศ 39.70 และเมื่อพิจารณาตามขนาดโรงเรียน พบว่า คะแนนเฉลี่ยนในโรงเรียนขนาดใหญ่พิเศษ มีผู้เข้าสอบ 206,215 คน คะแนนเฉลี่ย 41.76 สูงกว่าโรงเรียนขนาดใหญ่มีผู้เข้าสอบ 96,590 คน คะแนนเฉลี่ย 38.59 ขนาดกลาง มีผู้เข้าสอบ 98,990 คน คะแนนเฉลี่ย 37.27 และขนาดเล็ก มีผู้เข้าสอบ 21,724 คน คะแนนเฉลี่ย 36.09 ตามลำดับ

วิชาภาษาอังกฤษ มีผู้เข้าสอบ 423,417 คน มีคะแนนเฉลี่ยระดับประเทศ 24.98 และเมื่อพิจารณาตามขนาดโรงเรียน พบว่า คะแนนเฉลี่ยในโรงเรียนขนาดใหญ่พิเศษ มีผู้เข้าสอบ 206,194 คน คะแนนเฉลี่ย 28.76 สูงกว่าโรงเรียนขนาดใหญ่ มีผู้เข้าสอบ 96,592 คน คะแนนเฉลี่ย 22.70 ขนาดกลาง มีผู้เข้าสอบ 98,926 คน คะแนนเฉลี่ย 20.65 และขนาดเล็ก มีผู้เข้าสอบ 21,705 คน คะแนนเฉลี่ย 18.95 ตามลำดับ

วิชาคณิตศาสตร์ มีผู้เข้าสอบ 423,654 คน มีคะแนนเฉลี่ยระดับประเทศ 26.59 และเมื่อพิจาณาตามขนาดโรงเรียน พบว่า คะแนนเฉลี่ยในโรงเรียนขนาดใหญ่พิเศษ มีผู้เข้าสอบ 206,307 คน คะแนนเฉลี่ย 30.40 สูงกว่าโรงเรียนขนาดใหญ่ มีผู้เข้าสอบ 96,624 คน คะแนนเฉลี่ย 24.17 ขนาดกลาง มีผู้เข้าสอบ 98,999 คน คะแนนเฉลี่ย 22.37 และขนาดเล็ก มีผู้เข้าสอบ 21,724 คน คะแนนเฉลี่ย 20.40 ตามลำดับ

วิชาวิทยาศาสตร์ มีผู้เข้าสอบ 422,718 คน มีคะแนนเฉลี่ยระดับประเทศ 33.40 และเมื่อพิจาณาตามขนาดโรงเรียน พบว่า คะแนนเฉลี่ยในโรงเรียนขนาดใหญ่พิเศษ มีผู้เข้าสอบ 205,801 คน คะแนนเฉลี่ย 35.32 สูงกว่าโรงเรียนขนาดใหญ่ มีผู้เข้าสอบ 96,454 คน คะแนนเฉลี่ย 32.20 ขนาดกลาง มีผู้เข้าสอบ 98,768 คน คะแนนเฉลี่ย 31.31 และขนาดเล็ก มีผู้เข้าสอบ 21,695 คน คะแนนเฉลี่ย 30.11 ตามลำดับ

จำแนกตามที่ตั้ง และภูมิภาค

วิชาภาษาไทย มีผู้เข้าสอบ 422,625 คน มีคะแนนเฉลี่ยระดับประเทศ 49.36 เมื่อจำแนกตามที่ตั้งในเมือง ผู้เข้าสอบ 181,465 คน คะแนนเฉลี่ย 53.84 มีคะแนนเฉลี่ยสูงกว่าที่ตั้งนอกเมือง ผู้เข้าสอบ 241,160 คน คะแนนเฉลี่ยสูงกว่าที่ตั้งนอกเมือง ผู้เข้าสอบ 241,160 คน คะแนนเฉลี่ย

46.00 เมื่อจำแนกตามภูมิภาค พบว่า ในกรุงเทพฯ มีผู้เข้าสอบ 50,364 คน คะแนนเฉลี่ย 56.42 ภาคเหนือ มีผู้เข้าสอบ 41,818 คน คะแนนเฉลี่ย 50.73 ภาคตะวันออก มีผู้เข้าสอบ 28,731 คน คะแนนเฉลี่ย 50.55 ภาคกลาง ไม่รวมกรุงเทพฯ มีผู้เข้าสอบ 86,747 คน คะแนนเฉลี่ย 50.10 ภาคตะวันตก มีผู้เข้าสอบ 18,180 คน คะแนนเฉลี่ย 49.75 ภาคใต้ มีผู้เข้าสอบ 60,062 คน คะแนนเฉลี่ย 47.54 ภาคตะวันออกเฉียงเหนือ มีผู้เข้าสอบ 136,723 คน คะแนนเฉลี่ย 46.37 ตามลำดับ

วิชาสังคม ศาสนาและวัฒนธรรม มีผู้เข้าสอบ 423,519 คน คะแนนเฉลี่ยระดับประเทศ 39.70 เมื่อจำแนกตามที่ตั้งในเมือง มีผู้เข้าสอบ 182,091 คน คะแนนเฉลี่ย 41.47 มีคะแนนเฉลี่ยสูงกว่าที่ตั้งนอกเมือง มีผู้เข้าสอบ 241,428 คน คะแนนเฉลี่ย 38.36 เมื่อจำแนกตามภูมิภาค พบว่า ในกรุงเทพฯ มีผู้เข้าสอบ 50,695 คน คะแนนเฉลี่ย 42.23 ภาคเหนือมีผู้เข้าสอบ 41,839 คน คะแนนเฉลี่ย 40.63 ภาคตะวันออก มีผู้เข้าสอบ 28,798 คน คะแนนเฉลี่ย 40.12 ภาคกลางไม่รวมกรุงเทพฯ มีผู้เข้าสอบ 86,939 คน คะแนนเฉลี่ย 39.83 ภาคตะวันตก มีผู้เข้าสอบ 18,240 คน คะแนนเฉลี่ย 39.66 ภาคใต้ มีผู้เข้าสอบ 60,119 คน คะแนนเฉลี่ย 38.94 ภาคตะวันออกเฉียงเหนือ มีผู้เข้าสอบ 136,889 คน คะแนนเฉลี่ย 38.64 ตามลำดับ

ภาษาอังกฤษ มีผู้เข้าสอบ 423,417 คน มีคะแนนเฉลี่ยระดับประเทศ 24.98 เมื่อจำแนกตามที่ตั้งในเมือง มีผู้เข้าสอบ 181,944 คน คะแนนเฉลี่ย 29.11 มีคะแนนเฉลี่ยสูงกว่าที่ตั้งนอกเมือง มีผู้เข้าสอบ 241,473 คน คะแนนเฉลี่ย 21.87 เมื่อจำแนกตามภูมิภาค พบว่า ในกรุงเทพฯ มีผู้เข้าสอบ 50,560 คน คะแนนเฉลี่ย 34.42 ภาคเหนือมีผู้เข้าสอบ 41,854 คน คะแนนเฉลี่ย 25.61 ภาคตะวันออก มีผู้เข้าสอบ 28,827 คน คะแนนเฉลี่ย 25.47 ภาคกลางไม่รวมกรุงเทพ มีผู้เข้าสอบ 86,899 คน ค่าเฉลี่ย 25.15 ภาคตะวันตก มีผู้เข้าสอบ 18,222 คน คะแนนเฉลี่ย 23.68 ภาคใต้ มีผู้เข้าสอบ 60,134 คน คะแนนเฉลี่ย 23.44 ภาคตะวันออกเฉียงเหนือ มีผู้เข้าสอบ 136,921 คน คะแนนเฉลี่ย 21.94 ตามลำดับ

วิชาคณิตศาสตร์ มีผู้เข้าสอบ 423,654 คน มีคะแนนเฉลี่ยระดับประเทศ 26.59 เมื่อจำแนกตามที่ตั้งในเมือง มีผู้เข้าสอบ 182,102 คน คะแนนเฉลี่ย 30.48 มีคะแนนเฉลี่ยสูงกว่าที่ตั้งนอกเมือง มีผู้เข้าสอบ 241,552 คน คะแนนเฉลี่ย 23.66 เมื่อจำแนกตามภูมิภาค พบว่า ในกรุงเทพฯมีผู้เข้าสอบ 50,663 คน คะแนนเฉลี่ย 33.66 ภาคเหนือ มีผู้เข้าสอบ 41,865 คน คะแนนเฉลี่ย 27.70 ภาคตะวันออก มีผู้เข้าสอบ 28,829 คน คะแนนเฉลี่ย 27.35 ภาคกลางไม่รวมกรุงเทพฯ มีผู้เข้าสอบ 86,953 คน คะแนนเฉลี่ย 27.04 ภาคตะวันตก มีผู้เข้าสอบ 18,224 คน คะแนนเฉลี่ย 26.30 ภาคใต้ มีผู้เข้าสอบ 60,172 คน คะแนนเฉลี่ย 25.54 ภาคตะวันออกเฉียงเหนือ มีผู้เข้าสอบ 136,948 คน คะแนนเฉลี่ย 23.69 ตามลำดับ

วิชาวิทยาศาสตร์ มีผู้เข้าสอบ 422,718 คน มีคะแนนเฉลี่ยระดับประเทศ 33.40 เมื่อจำแนกตามที่ตั้งในเมือง มีผู้เข้าสอบ 181,504 คน คะแนนเฉลี่ย 35.18 มีคะแนนเฉลี่ยสูงกว่าที่ตั้งนอกเมือง มีผู้เข้าสอบ 241,214 คน คะแนนเฉลี่ย 32.06 เมื่อจำแนกตามภูมิภาค พบว่า ในกรุงเทพฯ มีผู้เข้าสอบ 50,332 คน คะแนนเฉลี่ย 36.08 ภาคเหนือ มีผู้เข้าสอบ 41,837 คน คะแนนเฉลี่ย 34.48 ภาคตะวันออก มีผู้เข้าสอบ 28,745 คน คะแนนเฉลี่ย 33.75 ภาคกลางไม่รวมกรุงเทพ มีผู้เข้าสอบ 86,747 คน คะแนนเฉลี่ย 33.52 ภาคตะวันตก มีผู้เข้าสอบ 18,188 คน คะแนนเฉลี่ย 33.26 ภาคใต้ มีผู้เข้าสอบ 60,092 คน คะแนนเฉลี่ย 32.70 ภาคตะวันออกเฉียงเหนือ มีผู้เข้าสอบ 136,777 คน คะแนนเฉลี่ย 32.27 ตามลำดับ

ตอนที่ ๑ งานวิจัยที่เกี่ยวข้องกับการทำหน้าที่ต่างกันของข้อสอบ

งานวิจัยต่างประเทศ

Finch and French (2007) ได้ศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการตรวจสอบ 4 วิธี เป็นการศึกษารูแบบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป ได้แก่ วิธี ซิปเทส (SIBTEST) วิธีถดถอยโลจิสติก (Logistic regression) วิธีทดสอบอัตราส่วนความน่าจะเป็น (Item response theory theory likelihood ratio test) และการวิเคราะห์องค์ประกอบเชิงยืนยัน (Confirmatory factor analysis) โดยมีปัจจัยที่เป็นเงื่อนไขในการจำลองข้อมูล คือ ขนาดของกลุ่มตัวอย่าง, ความแตกต่างของความสามารถระหว่างกลุ่มผู้สอบย่อยขนาดของการทำหน้าที่ต่างกันของข้อสอบและรูปแบบของการจำลองข้อมูล โดยใช้รูปแบบ 2 รูปแบบ คือ 3 พารามิเตอร์ และ 2 พารามิเตอร์ ผลการศึกษา พบว่า วิธีการตรวจสอบทั้งสี่วิธี สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้แก่ วิธี SIBTEST มีอำนาจการทดสอบของกระบวนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบดีที่สุด IRTLR และ CFA มีอำนาจการทดสอบต่ำกว่า LR และ SIBTEST ทุก ๆ เงื่อนไข

Glickman, Seal and Eisen (2009) ทำการศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีนอนพาราเมตริกซ์ โดยเปรียบเทียบค่านัยสำคัญของการแจกแจงภายหลัง (Posterior p-value) กับค่านัยสำคัญของค่าความน่าจะเป็น (Likelihood p-value) โดยใช้โปรแกรม R และ OpenBUGS ในการจำลองค่าพารามิเตอร์ความสามารถของผู้สอบจากการแจกแจงภายหลัง ค่าประมาณมอนติคาร์โลของ p-value คำนวณโดยการเปรียบเทียบความเหมาะสมของสมการถดถอยนอนพาราเมตริกซ์ของการตอบสนองข้อสอบในการจำลองค่าพารามิเตอร์ความสามารถของผู้สอบ และกลุ่มย่อยของผู้สอบ คุณสมบัติบางอย่างของวิธีการตรวจสอบการจำลองข้อมูลกับแบบสำรวจการประเมินสุขภาพของตนเอง นำมาตรวจสอบการทำหน้าที่ต่างกันของกลุ่มเชื้อชาติ

อังกฤษกับกลุ่มเชื้อชาติสเปน พบว่า มีข้อคำถามของการทำหน้าที่ต่างกันของข้อสอบ ผลการศึกษาพบว่า วิธีเบส์เซียนมีความเสถียรมากกว่าวิธีความน่าจะเป็น และวิธีเบส์เซียน ได้ค่านัยสำคัญที่น้อยกว่าวิธีความน่าจะเป็น

Tulin ACAR (2010) ได้ศึกษากำหนดให้มีการเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของ 3 วิธี ได้แก่ HGLM, LR และ IRT-LRโดยมีจุดมุ่งหมายของงานวิจัย คือ การตรวจสอบการทำหน้าที่ต่างกัน โดยวิธีโมเดลเชิงเส้นตรงทั่วไปพหุระดับ (HGLM) ของข้อสอบ และทำการเปรียบเทียบผลลัพธ์กับวิธี Logistic regression และ IRT-likelihood ratio (IRT-LR) วิเคราะห์ความแตกต่างระหว่างเทคนิคที่ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เขาพบว่า มีความสอดคล้องกันระหว่างเทคนิคที่ใช้ทดสอบ DIF ในข้อสอบย่อยวิชาวิทยาศาสตร์ระหว่างเทคนิค LR และ IRT-LR ส่วนวิชาสังคมศาสตร์ พบความสอดคล้องระหว่างเทคนิค HGLM กับ LR, HGLM กับเทคนิค IRT-LR, LR จำนวนข้อสอบที่พบการทำหน้าที่ต่างกันของข้อสอบระหว่างเพศทั้งสามวิธีตรวจพบเกือบครึ่งหนึ่งของจำนวนข้อสอบ เมื่อเปรียบเทียบทั้งสามวิธี พบว่า เทคนิค LR พบน้อยกว่าเทคนิคอื่นเพียงเล็กน้อย ในการเปรียบเทียบการทดสอบข้อสอบย่อยของวิชาสังคมศาสตร์และวิทยาศาสตร์ของตุรกีโดยการใช้เทคนิค HGLM พบ DIF จำนวนมากกว่าครึ่งของข้อสอบ สรุปได้ว่าการทดสอบข้อสอบย่อยของตุรกี พบว่า มีการทำหน้าที่ต่างกันของข้อสอบในจำนวนมากที่สุด

Burcu ATAR (2011) ได้เปรียบเทียบกระบวนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี IRT likelihood ratio test และวิธี Logistic regression พบว่า อัตราข้อผิดพลาดประเภทที่ 1 และอำนาจการทดสอบของ IRT likelihood ratio test และกระบวนการถดถอยโลจิสติกฟังก์ชั่นสะสมเพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบของคะแนนพหุภาค ถูกวิเคราะห์ด้วยการจำลองข้อมูล สำหรับโดยทดสอบเงื่อนไข 53 ข้อ (รวม 3 ขนาดตัวอย่าง, 2 อัตราส่วนขนาดตัวอย่าง, 3 ขนาดของการทำหน้าที่ต่างกันของข้อสอบ และ 3 เงื่อนไขการทำหน้าที่ต่างกันของข้อสอบ) ที่ถูกสร้างขึ้นในแต่ละเงื่อนไขและถูกจำลองแบบ จำนวน 200 ครั้ง) โดยทั่วไปอัตราข้อผิดพลาดชนิดที่ 1 ของการทดสอบ IRT likelihood ratio และ Ordinal logistic regression จะถูกควบคุมเป็นอย่างดีตามเงื่อนไขการจำลอง อำนาจการทดสอบ Likelihood-ratio จะสูงในขนาดของกลุ่มตัวอย่างขนาดเล็ก และขนาดกลาง หรือการเพิ่มขนาดของ DIF อำนาจการทดสอบจะเพิ่มขึ้นถ้าขนาดกลุ่มตัวอย่างใหญ่ขึ้น ในทางตรงกันข้ามอำนาจการทดสอบของวิธีถดถอยโลจิสติกจะอยู่ในระดับต่ำสำหรับทุกเงื่อนไขของ DIF ยกเว้นกลุ่มตัวอย่างขนาดใหญ่ หรือเงื่อนไขของ DIF มีขนาดใหญ่

Le (2009) ได้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างเพศข้ามประเทศและทดสอบภาษาของข้อสอบ PISA วิชาวิทยาศาสตร์ พบว่า การศึกษาครั้งนี้ใช้ข้อมูลการสอบ PISA รอบที่ 3 มาใช้ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบระหว่างเพศข้ามประเทศ และทดสอบภาษาสำหรับข้อสอบวิชาวิทยาศาสตร์ และรูปแบบในสี่มิติของ PISA ได้แก่ จุดเน้น, เนื้อหา, ความสามารถ และความรู้ด้านวิทยาศาสตร์ ข้อมูลรวบรวมมาจากความแตกต่างด้านภาษา จำนวน 60 กลุ่ม โดยมีผู้เข้าร่วม 50 ประเทศ เป็นนักเรียนที่มีอายุ 15 ปี ทั้งหมด 83,000 คน วิธีการ IRT ถูกนำมาใช้ในการตรวจสอบความแตกต่างทางเพศแบบอเนกรูปในกลุ่มภาษาและตัวอย่างระหว่างประเทศ ความแตกต่างทางเพศมีการมุ่งเน้นใน 5 มิติของข้อสอบ ยกตัวอย่าง เช่น ข้อสอบแบบหลายตัวเลือก และข้อสอบที่ตอบสนองต่อเพศชาย การศึกษาครั้งนี้ส่งผลกระทบของประเทศและการทดสอบภาษาที่มีการทำหน้าที่ต่างกันของข้อสอบระหว่างเพศ การค้นพบจากการศึกษานี้มีส่วนสนับสนุนให้มีการพัฒนาแบบทดสอบสำหรับการใช้งานระหว่างประเทศ

Weziak-Bialowolska (2015) ได้ศึกษาเรื่อง ความแตกต่างของบรรทัดฐานทางเพศระหว่างประเทศ: ถูกต้องหรือไม่? ปัญหาของความไม่แปรเปลี่ยนของโมเดลการวัด พบว่า ค่านิยมและทัศนคติต่อบทบาททางเพศจะถูกตรวจสอบบ่อยครั้ง และมีการเปรียบเทียบจากมุมมองแต่ละประเทศ โดยไม่ได้รับการปฏิบัติที่เหมาะสมทางสถิติของการประเมินความไม่แปรเปลี่ยน (MI) โดยพื้นฐานข้อสรุปขึ้นอยู่กับคะแนนรวมของบรรทัดฐานทางเพศ ทัศนคติบทบาททางเพศ หรือกิจวัตรประจำวันทางเพศ เขา พบว่า ไม่เคยมีการตรวจสอบระหว่างประเทศ ในความไม่แปรเปลี่ยนของโมเดลการวัด (MI) ด้านพื้นฐานความเท่าเทียมกันทางเพศ (Gender equality scale: GES) ตามข้อมูลการสำรวจทั่วโลก โดยการวิจัยได้ใช้การวิเคราะห์องค์ประกอบเชิงยืนยันหลายกลุ่ม (Multi-group confirmatory factor analysis) โดยวิธีการจัดตำแหน่ง และไม่มีการจัดตำแหน่ง (With and without alignment) เพื่อตรวจสอบความไม่แปรเปลี่ยนแบบ Configural, ความไม่แปรเปลี่ยนของโมเดลการวัดแบบอ่อน, ความไม่แปรเปลี่ยนของโมเดลการวัดแบบแข็งแกร่ง และความไม่แปรเปลี่ยนของโมเดลการวัดแบบเข้มงวด

ผลการศึกษา พบว่า แนวคิดของความเท่าเทียมทางเพศยังไม่มีการเปรียบเทียบในทุกประเทศที่มีส่วนร่วมในการสำรวจ ความแตกต่างพบในระหว่างยุโรปตะวันตกและยุโรปกลางและยุโรปตะวันออก ซึ่งได้เลือกเฉพาะประเทศแถบภาคกลางและประเทศแถบยุโรปตะวันออกประเทศมาแสดงความไม่แปรเปลี่ยน Configural MI แต่มีความล้มเหลวในการแสดงความไม่แปรเปลี่ยนแบบแข็งแกร่งเต็มรูปแบบ และความไม่แปรเปลี่ยนแบบเข้มงวดเต็มรูปแบบ ภายใต้กรอบการวัดการแสดงให้เห็นความแตกต่างในประเทศเหล่านี้ มีการเปรียบเทียบโดยการจัดอันดับประเทศซึ่งถูกยอมรับต่อ GES ว่ามีความถูกต้องโดยมีเงื่อนไขของความไม่แปรเปลี่ยน (Non-invariance)

ด้านความสัมพันธ์ของค่าน้ำหนักองค์ประกอบ และ/ หรือจุดตัด (Thresholdss) ที่ถูกนำมาใช้ การศึกษาแสดงให้เห็นว่า การวัดทัศนคติบทบาททางเพศโดย GES พบว่า ในสาธารณรัฐเช็ก ฮังการี ลิทัวเนีย โครเอเชีย มีทัศคติต่อความเท่าเทียมกันทางเพศสูงอย่างมีนัยสำคัญทางสถิติ และมีคะแนนต่ำที่สุดในประเทศโปแลนด์โปแลนด์ สโลวาเกีย แอลเบเนีย และโรมาเนีย

Fukuhara and Kamata (2011) ศึกษาวิธีการตรวจสอบ DIF ของข้อมูลที่มีลักษณะเป็นเทสต์เลท โดยขยายแนวคิด Bifactor Multidimensional Item Response Theory Model (Bi-factor MIRT) เปรียบเทียบกับ IRT DIF Model ประมาณค่าด้วยวิธีของเบส์ โดยใช้โปรแกรม WINBUGS ข้อมูลที่นำมาศึกษาใช้ทั้งข้อมูลจริงและข้อมูลจำลอง มีเงื่อนไขการจำลองข้อมูล 24 เงื่อนไข ได้แก่ อิทธิพล (0.5, 1, 2) ขนาด DIF (0.5, 0.7) ระดับค่าอำนาจจำแนกของข้อสอบ (0.8, 2.0) และจำนวนผู้สอบของกลุ่มสนใจ (Focal group) 500, 250 กำหนดให้มีความยาวแบบสอบในทุกสถานการณ์ 42 ข้อ (7 ข้อ ต่อ 1 Testlet แบบสอบมี 6 Testlets) และจำนวนผู้สอบ 1,000 คน เท่ากันทุกเงื่อนไข ทำการคำนวณซ้ำ 100 ครั้ง ในแต่ละเงื่อนไข ผลการศึกษา พบว่า

1. ทั้งสองโมเดล สามารถประมาณค่าขนาดของ DIF ในข้อที่ DIF ต่ำกว่าจริง (Understimatesd) สำหรับเงื่อนไขที่มีค่าอำนาจจำแนกของข้อสอบสูง (a = 2.0) และการประมาณค่าความยากของทั้งสองโมเดลมีค่าเข้าใกล้ 0 เหมือนกัน

2. Bifactor MIRT DIF model มีอัตราการตรวจสอบ DIF สูงกว่าและประมาณค่าขนาดของ DIF ได้ถูกต้องกว่า IRT DIF model มีอัตราการตรวจสอบ DIF สูงกว่าและประมาณค่าขนาดของ DIF ได้ถูกต้องกว่า IRT DIF model นอกจากนี้ Bifactor MIRT DIF model ยังมีอัตราความผิดพลาดในการตรวจ DIF ที่ต่ำมาก และมาณค่าพารามิเตอร์อื่นได้ดีกว่า IRT DIF model

3. Bifactor MIRT DIF model มีอคต หรือลำเอียง ในการประมาณค่าขนาด DIF น้อยกว่า 2 PL IRT DIF model

Huggins (2012) ได้ศึกษาผลกระทบของการทำหน้าที่ต่างกันของข้อสอบในความไม่เท่าเทียมกันของประชากรโดยทฤษฎีการตอบสนองข้อสอบโดยการเทียบจากคะแนนที่แท้จริง ความไม่แปรเปลี่ยนของประชากรในการเปรียบเทียบเกิดขึ้นเมื่อความสัมพันธ์ระหว่างสองระดับ (Scale) ของความสัมพันธ์ของผู้สอบสองกลุ่มหรือกลุ่มผู่สอบกลุ่มย่อย และด้วยเหตุนี้ฟังก์ชั่นที่ใช้ในการวัดระดับของความสัมพันธ์ไม่ได้ขึ้นกับกลุ่มผู้สอบย่อย การขาดความเท่าเทียมกันของความไม่แปรเปลี่ยน (เช่น การทำให้เท่ากัน) นำสู่การจำลองเหตุการณ์ของการตรวจสอบที่กำหนดให้มีคะแนนเท่ากันในระดับเดียวกัน แต่ขึ้นอยู่กับความแตกต่างกันของผู้สอบกลุ่มย่อย มีคะแนนการทดสอบที่แตกต่างกันในระดับของความสอดคล้องที่เท่ากัน การจำลองสถานการณ์ส่งผลให้เกิดสถานการณ์ที่ให้ประโยชน์สำหรับกลุ่มผู้สอบหนึ่งกลุ่มหรือมากกว่าหนึ่งกลุ่ม ด้วยเหตุ

นี้ความตระหนักเกี่ยวกับความยุติธรรมในการประเมินและความรับผิดชอบ มีความรู้เพียงเล็กน้อยเกี่ยวกับสาเหตุที่ต้องมีความเท่าเทียมกัน จุดประสงค์ของการวิจัย คือการตรวจสอบหาสาเหตุของปัญหานี้ มันเป็นข้อสมมุติฐานเพื่อทำการแสดงของการทำหน้าที่ต่างกันของข้อสอบในข้อสอบร่วมของการประเมินว่าอาจจะมีผลกระทบของความเท่าเทียมกันของความไม่แปรเปลี่ยนในประชากร

ผลการวิจัยแสดงให้เห็นว่าการทำหน้าที่ต่างกันของข้อสอบมีความแตกต่างกันของรูปแบบข้อสอบของกลุ่มผู้สอบย่อย ความไม่แปรเปลี่ยนของกลุ่มประชากรถูกทำให้มีความเท่าเทียมกันภายใต้เงื่อนไข การเพิ่มขึ้นของสมการความเท่าเทียมกันมีความสัมพันธ์ในระดับต่ำ กับการเพิ่มขนาดของการทำหน้าที่ต่างกันของข้อสอบ หรือการเพิ่มขึ้นของจำนวนข้อสอบร่วม ผลกระทบเหล่านี้อาจจะเป็นปัญหาในเงื่อนไขของการทำหน้าที่ต่างกันทั้งแบบทิศทางเดียว และแบบสองทิศทาง

งานวิจัยต่างในประเทศ

กาญจนา วัธนสุนทร (2538) ได้พัฒนาเกณฑ์การตัดสินข้อสอบลำเอียงทางเพศ โดยใช้ข้อมูลเชิงประจักษ์ โดยวิธีการตรวจสอบ 3 วิธี คือ IRT แบบ 2 พารามิเตอร์ วิธีซิปเทสท์ และวิธี MH ดัชนีที่พัฒนาเพื่อเป็นเกณฑ์ในการตัดสินข้อสอบลำเอียงคือ SA, UA, α_{MH}, β_{SIB} ตามลำดับ ซึ่งในการวิจัยครั้งยี้ได้จากการวิเคราะห์ค่าเฉลี่ยดัชนีแต่ละตัว ปัจจัยที่แปรเปลี่ยนในการศึกษา ได้แก่ ความยาวของแบบสอบ 20, 30 และ 40 ข้อในวิชาคณิตศาสตร์ และ 50, 60, 70 และ 80 ข้อ วิชาภาษาอังกฤษ ขนาดผู้สอบ 100, 200, 400, 600, 800 และ 1,000 คน

ผลการวิจัยพบว่า ขนาดของผู้เข้าสอบมีอิทธิพลต่อค่าเฉลี่ยของดัชนีทุกตัว ความยาวแบบสอบมีอิทธิพลต่อค่าเฉลี่ยของดัชนี SA และ UA แต่ไม่มีอิทธิพลต่าค่าเฉลี่ย α_{MH} และ β_{SIB} ซึ่งเกณฑ์ที่พัฒนาขึ้นเพื่อใช้ตัดสินความลำเอียงระหว่างผู้เข้าสอบเพศชาย และเพศหญิง เป็นดังนี้

1. $|SA| > 0.80$ และ $UA > .50$ เมื่อมีข้อสอบน้อยกว่า 50 ข้อ
2. $|SA| > 0.40$ และ $UA > 1.20$ เมื่อมีข้อสอบ 50 ข้อขึ้นไป
3. $\alpha_{MH} < 0.60$, $\alpha_{MH} > 1.40$ สำหรับทุกขนาดของผู้สอบและความยาวแบบสอบ
4. $\beta_{SIB} > 0.60$ สำหรับทุกขนาดของผู้สอบและความยาวแบบสอบ

นอกจากนี้ ยังพบว่าการใช้ดัชนี SA หรือ UA ควรใช้ผู้สอบขนาด 800 คนขึ้นไป ส่วนดัชนี α_{MH} และ β_{SIB} ควรใช้ขนาดผู้สอบอย่างน้อย 600 คน

เกสร หว่างจิตร (2539) ได้วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมนเทล-แฮนส์เซล (MH) โดยใช้เพศ ภูมิลำเนา ประสบการณ์ในการสอบ และสังกัดของสถานศึกษา เป็นเกณฑ์จำแนกผู้สอบเป็นกลุ่มอ้างอิง และกลุ่มเปรียบเทียบ ข้อมูลที่ใช้ศึกษาเป็นผลการสอบในวิชาสอบร่วมของศูนย์ทดสอบทางการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในส่วนเฉพาะ

ที่เป็นข้อสอบชนิดเลือกตอบ 2 วิชา คือ วิชาภาษาไทย ซึ่งมีผู้สอบจำนวน 506 คน และวิชาภาษาอังกฤษ ซึ่งมีผู้สอบจำนวน 501 คน แล้วนำผลการสอบมาวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปและแบบอเนกรูปโดยโปรแกรม MH_{DIF} สำหรับการตรวจสอบความเที่ยงและความตรงของแบบสอบใช้โปรแกรม CTIA และ LISREL ตามลำดับ

ผลการศึกษา พบว่า ข้อสอบที่ถูกระบุว่าทำหน้าที่ต่างกันส่วนมากมีลักษณะเป็นแบบอเนกรูป โดยพบในผู้สอบกลุ่มย่อยที่จำแนกตามเพศมากที่สุด รองลงมา คือ จำแนกตามภูมิลำเนาสังกัดสถานศึกษา และประสบการณ์ในการสอบตามลำดับ ข้อสอบที่ทำหน้าที่ต่างกันส่วนมากเป็นข้อสอบที่มีค่าอำนาจจำแนกค่อนข้างต่ำทั้งสองวิชา เมื่อพิจารณาลักษณะลักษณะของข้อสอบ พบว่า ในแบบสอบวิชาภาษาไทยที่ถูกระบุว่าทำหน้าที่ต่างกันส่วนมากจะเป็นข้อสอบที่ง่ายมากในแบบสอบวิชาภาษาอังกฤษข้อสอบที่ถูกระบุว่าทำหน้าที่ต่างกันส่วนมากจะเป็นข้อสอบที่ยากมาก นอกจากนี้ยังพบว่า ค่าความเที่ยงและความตรงของแบบสอบก่อนและหลังการตัดข้อสอบที่ทำหน้าที่ต่างกันออกจากแบบสอบไม่แตกต่างกันทั้งสองวิชา

เรวดี อินทะสะระ (2539) ได้ศึกษาความเที่ยงตรงเชิงพยากรณ์ของแบบสอบคัดเลือกที่วิเคราะห์ความลำเอียงต่อเพศด้วยวิธีทฤษฎีการตอบสนองข้อสอบ (IRT) วิธีแมนเทล-แฮนส์เซล (MH) และวิธีซิปเทสท์ (SIBTEST) พร้อมทั้งศึกษาการตัดสินผลการสอบที่คิดคะแนนมาตรฐานที่ปกติ และคะแนนน้ำหนักความสามารถและสาเหตุของความลำเอียงของข้อสอบ โดยศึกษาความลำเอียงของข้อสอบคัดเลือกเข้าศึกษาในชั้นปีที่ 1 ประเภทรับตรง ปีการศึกษา 2538 ของมหาวิทยาลัยสงขลานครินทร์ ในวิชาเอกภาษาไทย ก วิชาสังคม ก และวิชาภาษาอังกฤษ กข วิชาละ 8,127 คน (เป็นชาย 2,722 คน หญิง 5,405 คน) วิชาภาษาไทย กข วิชาสังศึกษา กข และวิชาภาษาอังกฤษ กขค วิชาละ 5,415 คน (เป็นชาย 1,451 คน หญิง 3,691 คน)

ผลการศึกษา พบว่า วิธีการตรวจสอบความลำเอียงทั้ง 3 วิธี ตัดสินจำนวนข้อสอบที่ลำเอียงแตกต่างกันในวิชาภาษาไทย ก ฉบับที่ 2 ที่ระดับนัยสำคัญทางสถิติที่ .05 นอกนั้นแตกต่างกันที่ระดับนัยสำคัญ .01 โดยวิธีทฤษฎีการตอบสนองข้อสอบตัดสินจำนวนข้อสอบที่ลำเอียงได้มากที่สุดความสัมพันธ์ของลำดับที่ของข้อสอบ ไม่ว่าจะคิดคะแนนมาตรฐานปกติที หรือคิดคะแนนน้ำหนักความสามารถ และใช้ข้อสอบทั้งหมด หรือใช้เฉพาะข้อสอบที่ปราศจากความลำเอียง ต่างมีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01

ทิพย์รัตน์ ผลบุญ (2540) ได้ศึกษาดัชนีความลำเอียงของข้อสอบวัดผลสัมฤทธิ์ทางการเรียนในวิชาคณิตศาสตร์ ซึ่งเกิดจากตัวแปรเพศและประเภทโรงเรียน ด้วยวิธีวิเคราะห์ 3 วิธี คือ วิธีไค-สแควร์ วิธีทฤษฎีการตอบสนองข้อสอบที่มี 3 พารามิเตอร์และวิธีแมนเทล-แฮนส์เซล เปรียบเทียบความลำเอียงของข้อสอบแต่ละข้อและจำนวนข้อที่มีต่อตัวแปรเพศ และประเภทของ

โรงเรียน ด้วยวิธีวิเคราะห์ความลำเอียงทั้ง 3 วิธี และศึกษาสัมประสิทธิ์สหสัมพันธ์ระหว่างดัชนีความลำเอียงทั้ง 3 วิธี เครื่องมือที่ใช้เป็นแบบวัดผลสัมฤทธิ์สหสัมพันธ์ทางการเรียนในวิชาคณิตศาสตร์ ค 102 จำนวน 30 ข้อ กลุ่มตัวอย่างเป็นนักเรียนชั้นมัธยมศึกษาปีที่ 1 ของกลุ่มโรงเรียนมัธยมศึกษาตอนต้น จังหวัดกำแพงเพชร จำนวน 1,600 คน

ผลการวิจัย พบว่า วิธีวิเคราะห์ความลำเอียงโดยวิธีทฤษฎีการตอบสนองข้อสอบที่มี 3 พารามิเตอร์ ตรวจพบจำนวนข้อสอบที่ลำเอียงมากที่สุด โดยพบข้อสอบที่ลำเอียงตามตัวแปรเพศ จำนวน 12 ข้อ ตามตัวแปรโรงเรียน จำนวน 12 ข้อ รองลงมา คือ วิธีแมนเทล-แฮนส์เซล โดยพบข้อสอบที่ลำเอียงตามตัวแปรเพศจำนวน 10 ข้อ ตามตัวแปรโรงเรียนจำนวน 11 ข้อ ส่วนวิธีไค-สแควร์ เป็นวิธีที่พบข้อสอบที่มีความลำเอียงน้อยที่สุด โดยพบตามตัวแปรเพศ จำนวน 1 ข้อ ตามตัวแปรโรงเรียน จำนวน 2 ข้อ ตามลำดับ ในการเปรียบเทียบความลำเอียง พบว่า ลำเอียงเข้าข้างเพศชายมากกว่าเพศหญิง และลำเอียงเข้าข้างโรงเรียนสามัญมากกว่าโรงเรียนขยายโอกาสทางการศึกษา ส่วนความสัมพันธ์ของความลำเอียงของวิธีวิเคราะห์แต่ละวิธี เมื่อวิเคราะห์ตามตัวแปรเพศมีความสัมพันธ์เป็นบวก และเมื่อวิเคราะห์ตามตัวแปรโรงเรียน พบว่า วิธีไค-สแควร์ มีค่า -0.3910 วิธีทฤษฎีการตอบสนองข้อสอบที่มี 3 พารามิเตอร์ มีค่า 0.2065 และวิธีแมนเทล-แฮนส์เซล มีค่า 0.2532

พรรณี จินตมาศ (2540) ได้ศึกษาผลการวิเคราะห์ความลำเอียงต่อเพศของข้อสอบจากแบบทดสอบคณิตศาสตร์โจทย์ปัญหา โดยจำแนกในแต่ละวิธีวิเคราะห์ความลำเอียง 3 วิธี คือ วิธีแปลงค่าความยากวิธี MH และวิธี SIBTEST ในแต่ละขนาดของกลุ่มผู้สอบ คือ ขนาดกลุ่มผู้สอบ 500 คน และขนาดกลุ่มผู้สอบ 1,000 คน โดยเปรียบเทียบจำนวนข้อสอบที่มีความลำเอียงและเปรียบเทียบค่าความเชื่อมั่นแบบครึ่งฉบับของแบบทดสอบหลังคัดเลือกข้อสอบที่มีความลำเอียงออกแล้ว ในการศึกษาครั้งนี้มีกลุ่มตัวอย่างเป็นนักเรียนชั้นมัธยมศึกษาปีที่ 1 ภาคเรียนที่ 2 ปีการศึกษา 2539 ของโรงเรียนสังกัดกรมสามัญศึกษาส่วนกลางจำนวน 2,200 คน ซึ่งเลือกมาโดยการสุ่มแบบแบ่งชั้น มีขนาดโรงเรียนเป็นชั้น และโรงเรียนเป็นหน่วยการสุ่ม

ผลการวิจัย พบว่า เมื่อวิเคราะห์จากกลุ่มตัวอย่างขนาด 500 คน ด้วยวิธี SIBTEST พบข้อสอบที่มีความลำเอียงมากที่สุด และวิธีแปลงค่าความยากพบข้อสอบที่มีความลำเอียงน้อยที่สุด โดยจำนวนข้อสอบที่มีความลำเอียงแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ ทุกวิธีวิเคราะห์และเมื่อวิเคราะห์จากกลุ่มผู้สอบขนาด 1,000 คน วิธี MH พบข้อสอบมีความลำเอียงมากที่สุด วิธีการแปลงค่าความยากไม่พบข้อสอบที่มีความลำเอียง โดยจำนวนข้อสอบที่ลำเอียงจากการวิเคราะห์ด้วยวิธีแปลงค่าความยากกับวิธี MH และวิธีแปลงค่าความยากกับวิธีซิปเทสท์แตกต่างกันอย่างมีนัยสำคัญที่ระดับ .05 นอกนั้นมีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติสำหรับจำนวนข้อสอบที่ลำเอียง

จากการวิเคราะห์ด้วยวิธีแปลงค่าความยากระหว่างกลุ่มผู้สอบขนาด 500 คน และกลุ่มผู้สอบขนาด 1,000 คน จะมีจำนวนข้อสอบแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นอกนั้นมีค่าความแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

ค่าความเที่ยงของแบบทดสอบหลังคัดเลือกข้อสอบที่มีความลำเอียงออกเมื่อวิเคราะห์จากกลุ่มผู้สอบขนาด 500 คน จากการวิเคราะห์ความลำเอียงด้วยวิธีแปลงค่ายากกับวิธี SIBTEST มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นอกนั้นมีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติและเมื่อวิเคราะห์จากกลุ่มผ้สอบขนาด 1,000 คน ค่าความเที่ยงของแบบสอบหลังคัดเลือกข้อสอบที่มีความลำเอียงออกแล้วจากการวิเคราะห์ความลำเอียง 3 วิธี มีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ สำหรับค่าความเที่ยงของแบบทดสอบหลังคัดเลือกข้อสอบที่ลำเอียงออกจากการวิเคราะห์ด้วยวิธี SIBTEST ระหว่างกลุ่มผู้สอบ 500 คน และกลุ่มผู้สอบ 1,000 คน มีค่าแตกต่างกันbอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นอกนั้นมีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

เพ็ญพนา สุขสม (2540) ได้ทำการเปรียบเทียบผลของวิธีวิเคราะห์ความลำเอียงของข้อสอบที่แตกต่างกัน 3 วิธี ได้แก่ วิธีแปลงค่าความยาก วิธีโค้งลักษณะข้อสอบที่มี 3 พารามิเตอร์ และวิธีแมนเทล-แฮนส์เซล โดยกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ จำแนกตามเพศ และที่ตั้งของโรงเรียน เครื่องมือที่ใช้ คือ แบบประเมินคุณภาพและวัดผลปลายปี วิชาภาษาไทย จำนวน 50 ข้อ โดยข้อสอบเป็นแบบหลายตัวเลือก กลุ่มตัวอย่าง คือ นักเรียนชั้นประถมศึกษาปีที่ 6 จำนวน 2,400 คน ผลการศึกษา พบว่า วิธีแมนเทล-แฮนเซล ตรวจพบข้อสอบลำเอียงมีจำนวนมากที่สุด รองลงมา คือวิธีโค้งลักษณะข้อสอบที่มี 3 พารามิเตอร์ และวิธีแปลงความยากน้อยที่สุด ทั้งสามวิธีมีความสัมพันธ์ทางบวก ค่าความสัมพันธ์ทางบวก ค่าความสัมพันธ์มีค่าระหว่าง 0.411-0.6662 เมื่อจำแนกตามเพศ และความสัมพันธ์ มีค่าระหว่าง 0.5676-0.7847 เมื่อจำแนกตามที่ตั้งของโรงเรียน โดยวิธีโค้งลักษณะข้อสอบที่มี 3 พารามิเตอร์ มีความสัมพันธ์กับวิธีแมนเทล-แฮนเซล สูงกว่าวิธีแปลงค่าความยาก

สมศักดิ์ จั่นผอง (2542) ศึกษาการเปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยใช้วิธีวิเคราะห์และขนาดกลุ่มผู้สอบต่างกัน การศึกษา พบว่า วิธีชิปเสท์ มีความคลาดเคลื่อนสูงกว่าแมนเทล-แฮนส์เซล เล็กน้อยแต่สรุปว่าเมื่อใช้กลุ่มตัวอย่างไม่เกิน 1,000 คน ทั้งสองวิธีให้ผลการวิเคราะห์ใกล้เคียงกันแต่ถ้าใช้กลุ่มต้วอย่างมากกว่า 1,000 คน อัตราความคลาดเคลื่อนจะสูงขึ้นด้วย ตัวแปรที่นำมาศึกษา ตัวแปรต้น ได้แก่ วิธีวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ แบ่งเป็นวิธี IRT-3 พารามิเตอร์ วิธี MH และวิธี SIBTEST, กลุ่มผู้สอบ 2 ขนาด คือ กลุ่มผู้สอบ 600 คน แยกเป็นหญิง 600 คน ชาย 600 คน และกลุ่มผู้สอบ 1,000 คน แยกเป็นหญิง 1,000 คน ชาย

1,000 คน และตัวแปรตาม ได้แก่ ดัชนีการทำหน้าที่ต่างกันของข้อสอบ, จำนวนข้อสอบที่ทำหน้าที่ต่างกัน (DIF) ที่ตรวจพบ

รัชนก ยี่สุ่นศรี (2544) ทำการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบและแบบสอบสำหรับกลุ่มผู้สอบเมื่อจำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์และโรงเรียนที่จบการศึกษาเพื่อเปรียบเทียบความเที่ยง ความตรง และฟังก์ชันสารสนเทศของแบบสอบระหว่างแบบสอบฉบับก่อนและหลังตัดข้อสอบที่พบ DIF โดยใช้ข้อมูลจากการตอบแบบสอบคัดเลือกบุคคลเข้าศึกษาในสถาบันอุดมศึกษา วิชาภาษาอังกฤษ และคณิตศาสตร์ ปีการศึกษา 2543 ครั้งที่ 1 และเลือกศึกษาในส่วนที่เป็นข้อสอบแบบหลายตัวเลือกจากกลุ่มตัวอย่างที่เป็นผู้เข้าสอบ จำนวน 4,000 คน และ 3,600 คน ตามลำดับ

ผลการวิจัย พบว่า การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ โดยใช้ดัชนี NCDIF พบว่าแบบสอบวิชาภาษาอังกฤษ เมื่อจำแนกกลุ่มผู้สอบตามเพศ มีข้อสอบที่พบ DIF จำนวน 30 ข้อ คิดเป็นร้อยละ 30.00 ของจำนวนข้อสอบทั้งหมด และเมื่อจำแนกกลุ่มผู้สอบตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียนที่จบการศึกษา มีข้อสอบที่พบ DIF จำนวน 16 ข้อ คิดเป็นร้อยละ 16.00 ของจำนวนข้อสอบทั้งหมด นอกจากนี้ในแบบสอบวิชาคณิตศาสตร์ เมื่อจำแนกกลุ่มผู้สอบตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียนที่จบการศึกษา มีข้อสอบที่พบ DIF จำนวน 16 ข้อ คิดเป็นร้อยละ 21.43 ของจำนวนข้อสอบทั้งหมด

นพดล มีชั้นช่วง (2544) ได้ศึกษาการเปรียบเทียบผลการประมาณค่าพารามิเตอร์ตามทฤษฎีการตอบสนองข้อสอบ ระหว่างวิธีแมกซิมัมไลค์ลิฮูด วิธีฮิวริสติก และวิธีของเบส์ของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนคณิตศาสตร์ ชั้นมัธยมศึกษาปีที่ 1 ผลการวิจัย พบว่า ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าความยากที่ประมาณค่าด้วยวิธีแมกซิมัมไลค์ลิฮูด วิธีฮิวริสติก และวิธีของเบส์ พบว่า ค่าสัมประสิทธิ์สหสัมพันธ์มีค่าตั้งแต่ .263 ถึง .446 วิธีแมกซิมัมไลค์ลิฮูด และวิธีของเบส์ มีนัยสำคัญทางสถิติที่ระดับ .05 ได้แก่ วิธีแมกซิมัมไลค์ลิฮูด และวิธีของเบส์ ส่วนวิธีฮิวริสติก กับวิธีของเบส์ มีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติ

ชัยยศ ชาวระนอง (2553) ได้ศึกษาเปรียบเทียบประสิทธิภาพของโมเดลการวิเคราะห์องค์ประกอบเชิงยืนยันอันดับหนึ่ง องค์ประกอบเชิงยืนยันอันดับสอง องค์ประกอบเชิงยืนยันอันดับหนึ่งแฝงภายใน และองค์ประกอบเชิงยืนยันอันดับสองแฝงภายใน และการทดสอบการทำหน้าที่ต่างกันโดยใช้วิธีการวิเคราะห์องค์ประกอบเชิงยืนยันอันดับสองแฝงภายในภายใต้เงื่อนไขกลุ่มตัวอย่างขนาด 100, 200, 400, 800, 1,200, 1,600 และ 2,000 คน จำนวนข้อสอบขนาด 3, 5, 8, 10 และ 15 ต่อองค์ประกอบ และขนาดกลุ่มตัวอย่างในการทดสอบการทำหน้าที่ต่างกันขนาด 1,600, 2,000, 2,400, 2,800, 3,200, 3,600 และ 4,000 ผลการวิจัยพบว่า

1. โมเดลการวิเคราะห์องค์ประกอบเชิงยืนยันอันดับสองแฝงภายใน มีประสิทธิภาพมากที่สุดโมเดลมีความสอดคล้องกับข้อมูลเชิงประจักษ์เมื่อไม่มีการปรับโมเดลอยู่ในระดับดี เมื่อพิจารณาค่าดัชนีวัดความสอดคล้องของโมเดล พบว่า ค่า χ^2= 1658.181, *df* = 1130, p = .000, GFI = .919, AGFI = .909, CFI = .643 และ RMSEA = .024 เมื่อใช้กลุ่มตัวอย่างตั้งแต่ 800 ตัวอย่างขึ้นไป และเมื่อพิจารณาตามขนาดของกลุ่มตัวอย่าง พบว่า ค่าดัชนีวัดความสอดคล้องไคสแควร์ไม่แปรเปลี่ยนตามขนาดกลุ่มตัวอย่างแต่แปรเปลี่ยนตามจำนวนข้อสอบต่อองค์ประกอบ

2. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยวีการวิเคราะห์องค์ประกอบเชิงยืนยันอันดับแฝงภายในจำแนกตามเพศโดยใช้เทคนิค Multi-group พบว่า มีข้อสอบทำหน้าที่ต่างกัน 5 ข้อ และเมื่อตัดข้อสอบที่ทำหน้าที่ต่างกันออก พบว่า ค่าสัมประสิทธิ์ความเชื่อมั่นของข้อสอบสูงขึ้น

มิ่ง เทพครเมือง (2554) ศึกษาการตรวจสอบความเท่าเทียมกันของการวัดบนพื้นฐานทฤษฎีการทดสอบแบบคะแนนจริงสัมพัทธ์และทฤษฎีการตอบสนองข้อสอบ โดยมีจุดมุ่งหมายเพื่อศึกษาการตรวจสอบความเท่าเทียมกันของการวัดบนพื้นฐานทฤษฎีการทดสอบแบบคะแนนจริงสัมพันธ์และทฤษฎีการตอบข้อสอบ และ 2) เพื่อเปรียบเทียบผลการตรวจสอบความเท่าเทียมกันของการวัดบนพื้นฐานทฤษฎีการทดสอบแบบคะแนนจริงสัมพันธ์และทฤษฎีการตอบข้อสอบ ข้อมูลที่ใช้ในการศึกษาเป็นข้อมูลจำลองที่รายการคำตอบเป็นการตรวจให้คะแนนแบบสองค่า (Dichotomous scoring) และรายการคำตอบเป็นการตรวจให้คะแนนสองค่า (Polytomous scoring) ฉบับละ 25 ข้อ ภายใต้เงื่อนไขขนาดของการทำหน้าที่เบี่ยงเบนของข้อสอบ 3 ขนาด คือ ไม่มีการทำหน้าที่เบี่ยงเบน การทำหน้าที่เบี่ยงเบน การทำหน้าที่เบี่ยงเบนเล็กน้อย และการทำหน้าที่เบี่ยงเบนขนาดใหญ่ กลุ่มตัวอย่างแบ่งตามสัดส่วนกลุ่มอ้างอิงต่อกลุ่มสนใจเป็น 2 สัดส่วน คือ 1: 1 และ 2: 1 แล้วทำการจำลองข้อมูลซ้ำในแต่ละเงื่อนไขที่ศึกษาจำนวน 100 ครั้ง

ผลการวิจัยสรุปได้ดังนี้

1. อัตราความถูกต้องในการตรวจสอบความเท่าเทียมกันของการวัด กรณีรายการคำตอบของผู้สอบที่เป็นการตรวจให้คะแนนแบบสองค่า การทำหน้าที่เบี่ยงเบนของข้อสอบขนาดเล็กและการทำหน้าที่เบี่ยงเบนของข้อสอบขนาดใหญ่ ที่กลุ่มอ้างอิงต่อกลุ่มสนใจ 100: 100 คน วิธีการตรวจสอบบนพื้นฐานทฤษฎีการตอบสนองข้อสอบแบบคะแนนจริงสัมพันธ์ มีอัตราความถูกต้องของการตรวจสอบสูงกว่าวิธีการตรวจสอบบนพื้นฐานทฤษฎีการตอบข้อสอบ อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ส่วนรายการคำตอบของผู้สอบที่เป็นการตรวจให้คะแนนแบบหลายค่า การทำหน้าที่เบี่ยงเบนขนาดใหญ่ ที่กลุ่มอ้างอิงต่อกลุ่มสนใจเป็น 200: 400 คน วิธีการตรวจสอบบนพื้นฐานทฤษฎีการทดสอบแบบคะแนนจริงสัมพันธ์ มีอัตราความถูกต้องในการตรวจสอบต่ำกว่า

วิธีการตรวจสอบบนพื้นฐานทฤษฎีการตอบข้อสอบ อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ส่วนเงื่อนไขอื่น ๆ มีอัตราความถูกต้องในการตรวจสอบไม่แตกต่างกัน

2. อัตราความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบความเท่าเทียมกันของการวัดกรณีรายการคำตอบของผู้สอบที่เป็นการตรวจให้คะแนนแบบสองค่า ที่กลุ่มอ้างอิงต่อกลุ่มสนใจเป็น 200: 400 คน วิธีตรวจสอบบนพื้นฐานทฤษฎีการทดสอบแบบคะแนนจริงสัมพันธ์มีความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบสูงกว่าวิธีบนพื้นฐานทฤษฎีการตอบข้อสอบอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ส่วนเงื่อนไขอื่น ๆ มีอัตราความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบไม่แตกต่างกัน

สุพัฒนา หอมบุปผา (2555) ได้ศึกษาเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN โดยมีวัตถุประสงค์เพื่อ 1) เปรียบเทียบผลการประมาณค่าพารามิเตอร์ความยากของข้อสอบ (δ_i) พารามิเตอร์ความสามารถของผู้ตอบ (θ_j) สำหรับผู้สอบจะแนกตามเพศ สถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน ระหว่างวิธี HGLM วิธี MIMIC และวิธี BAYESIAN 2) เปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ (DIF) สำหรับผู้สอบจำแนกตามเพศ สถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน ระหว่างวิธี HGLM วิธี MIMIC และวิธี BAYESIAN 3) ศึกษาลักษณะของข้อสอบที่เกิดการทำหน้าที่ต่างกันของข้อสอบ (DIF) ที่ได้จากการวิเคราะห์การทำหน้าที่ต่างกัน โดยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN ด้วยการวิเคราะห์ลักษณะและเนื้อหาของคำหรือข้อความที่ใช้ในการเขียนข้อสอบ

ข้อมูลที่ใช้ในการวิเคราะห์ครั้งนี้ เป็นคะแนนการสอบวัดผลสัมฤทธิ์ทางการเรียนเพื่อประเมินคุณภาพการศึกษาระดับชาติ ของนักเรียนชั้นประถมศึกษาปีที่ 3 ในรายวิชาภาษาไทย คณิตศาสตร์ และวิทยาศาสตร์ ของสำนักทดสอบทางการศึกษา กระทรวงศึกษาธิการ ปีการศึกษา 2553 ซึ่งได้มาโดยการสุ่มแบบหลายขั้นตอน (Multi-stage random sampling technique) จำนวน 1,000 คน จำแนกเป็นเพศชาย และเพศหญิง ที่อยู่ในเขตกรุงเทพมหานครและปริมณฑล และนอกเขตกรุงเทพมหานครและปริมณฑล

การวิเคราะห์ค่าพารามิเตอร์ความยากของข้อสอบ รามิเตอร์ความสามารถของผู้สอบ และการทำหน้าที่ต่างกันของข้อสอบ (DIF) ด้วยวิธี HGLM-2L วิธี MIMIC และวิธี BAYSESIAN โดยใช้โปรแกรมสำเร็จรูป 3 โปรแกรม ได้แก่ โปรแกรม Mplus และโปรแกรม WINBUGS ตามลำดับ เมื่อพบการทำหน้าที่ต่างกันของข้อสอบ (DIF) และทำการวิเคราะห์เนื้อหาคำ ประโยค หรือข้อความที่ใช้ในการเขียนข้อสอบของแต่ละวิชา ผลการวิจัย พบว่า

1. ผลการวิเคราะห์ค่าพารามิเตอร์ความยากของข้อสอบ ค่าพารามิเตอร์ความสามารถของผู้ตอบ และผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ในวิชาภาษาไทย

คณิตศาสตร์ และวิทยาศาสตร์ ด้วยวิธี HGLM-2L วิธี MIMIC และวิธี BAYESIAN พบว่ามีความสัมพันธ์กันในระดับสูงมาก อย่างมีนัยสำคัญทางสถิติที่ระดับ .01

2. วิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมากที่สุด คือ วิธี HGLM-2L ส่วนวิธีที่ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่น้อยที่สุด คือ วิธี MIMIC

3. ผลการศึกษาลักษณะของข้อสอบที่ตรวจพบการทำหน้าที่ต่างกันของข้อสอบ (DIF) วิชาภาษาไทย คณิตศาสตร์ และวิทยาศาสตร์ ลักษณะของข้อสอบที่เกิดการทำหน้าที่ต่างกันเมื่อจำแนกตามเพศ ส่วนใหญ่ข้อสอบที่เกิดการทำหน้าที่ต่างกันของข้อสอบ จะมีคำศัพท์ที่เกี่ยวข้องกับเพศนั้น จึงทำให้ข้อสอบเข้าข้างเพศนั้น และอาจเป็นเพราะความสามารถที่แตกต่างระหว่างเพศชายและเพศหญิงที่มีลักษณะความสามารถ ความถนัด และความสนใจในเรื่องนั้น ๆ ต่างกัน เมื่อจำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน ส่วนใหญ่สาเหตุที่ทำให้ข้อสอบเกิดการทำหน้าที่ต่างกันอาจเป็นเพาะประสบการณ์ ความคุ้นเคยเกี่ยวกับเรื่องนั้น สภาพแวดล้อมและการฝึกปฏิบัติที่แตกต่างกันระหว่างนักเรียนในเขตกรุงเทพมหานครและปริมณฑล นักเรียนนอกเขตกรุงเทพและปริมณฑล

อริสพา เตหลิ่ม (2556) ได้ศึกษาประสิทธิผลการประมาณค่าพารามิเตอร์และการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมกซิมัมไลค์ลิฮู้ด วิธีเบสี และวิธีเบส์แบบมีอิทธิพลเทสต์เลท โดยมีวัตถุประสงค์เพื่อ 1) ศึกษาประสิทธิผลในการประมาณค่าพารามิเตอร์ข้อสอบ (ค่าอำนาจจำแนกและความยาก) กับพารามิเตอร์ความสามารถของผู้สอบ และ 2) เพื่อศึกษาผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ระหว่างวิธีแมดซิมัมไลค์ลิฮู้ด (ML) วิธีของเบส์ (Bayes) และวิธีเบส์แบบมีอิทธิพลเทสต์เลท ($Bayes_{\gamma}$) ข้อมูลที่ศึกษาเป็นข้อมูลจำลองภายใต้ปัจจัยแปรเปลี่ยน 4 ปัจจัย คือ อิทธิพลเทสต์เลท (เท่ากันทุกเทสต์เลท แต่ละเทสต์เลทไม่เท่ากัน และข้อสอบที่เป็นอิสระผสมกับเทสต์เลท) การแจกแจงของความสามารถ (ปกติ เบ้ซ้าย เบ้ขวา) จำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบ (0, 5, 8 จากแบบสอบถามจำนวน 40 ข้อ) และอัตราส่วนของกลุ่มเปรียบเทียบต่อกลุ่มอ้างอิง (1000: 1000, 1000: 100) รวมจำนวนเงื่อนไขทั้งหมด 5 เงื่อนไข (3x3x3x2) กำหนดจำนวนรอบในการประมาณค่าพารามิเตอร์ ในแต่ละเงื่อนไข 100 รอบ

ผลการวิจัย พบว่า

1. ผลการประมาณค่าพารามิเตอร์ จำแนกเป็น (1) พารามิเตอร์ความยาก พบว่า วิธีเบส์แบบมีอิทธิพลเทสต์เลท ($Bayes_{\gamma}$) ประมาณค่าได้ดีเมื่อข้อมูลมีการแจกแจงความสามารถเป็นแบบปกติ ส่วนวิธีของเบส์ (Bayes) ยังไม่มีแนวโน้มแน่นอน แต่จะประมาณค่าพารามิเตอร์ความยากได้ดีเป็นส่วนใหญ่เมื่อข้อมูลมีการแจกแจงความสามารถเป็นแบบเบ้ซ้าย และวิธีแมกซิมัมไลค์ลิฮูด (ML)

จะประมาณค่าพารามิเตอร์ความยากได้ดี เมื่อข้อมูลมีการแจกแจงความสามารถเป็นแบบเบ้ขวา (2) พารมิเตอร์อำนาจจำแนก พบว่า ส่วนใหญ่วิธีการประมาณค่าด้วยวิธีเบส์แบบมีอิทธิพลเทสต์เลท ($Bayes_{\gamma}$) จะประมาณค่าพารามิเตอร์อำนาจจำแนกดีกว่าเมื่อข้อมูลความสามารถมีการแจกแจงแบบแกติ (3) พารามิเตอร์ความสามารถ พบว่า วิธีของเบส์แบบมีอิทธิพลเทสต์เลท ($Bayes_{\gamma}$) ประมาณค่าได้ดีที่สุด

1. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ พบว่า วิธีของเบส์แบบมีอิทธิพลเทสต์เลท ($Bayes_{\gamma}$) และวิธีของเบส์ (Bayes) สามารถควบคุมอัตราความคาดเคลื่อนประเภทที่ 1 ได้ดี (ยกเว้นกรณีที่อิทธิพลเทสต์เลทมีค่าเป็น 0.25, 0.5, 1, 2 ร่วมกับการแจกแจงความสามารถที่เป็นแบบเบ้ซ้ายและเบ้ขวา การประมาณค่าด้วยวิธีของเบส์ความคาดเคลื่อนประเภทที่ 1 สูง และมีอำนาจการตรวจสอบสูงเมื่อมีการแจกแจงความสามารถแบบเบ้ซ้ายและจำนวนตัวอย่างมาก แต่ไม่มากถึงเกณฑ์ที่กำหนด ตรงข้ามกับวิธีแมกซิมัมไลค์ลิฮู้ด (ML) ซึ่งไม่สามารถควบคุมอัตราความคาดเคลื่อนประเภททที่ 1 แต่มีอำนาจการตรวจสอบสูง

บทที่ 3
วิธีดำเนินการวิจัย

ประชากรและกลุ่มตัวอย่าง

ประชากร

ประชากรที่ใช้ในการวิจัยครั้งนี้ ได้แก่ นักเรียนชั้นมัธยมศึกษาปีที่ 6 ของโรงเรียนในสังกัดสำนักงานคณะกรรมการการศึกษาแห่งชาติขั้นพื้นฐาน การทรวงศึกษาธิการ โรงเรียนขนาดใหญ่พิเศษ (นักเรียน 2,501 คนขึ้นไป) ที่เข้าสอบวัดผลสัมฤทธิ์ทางการเรียนโดยแบบสอบทางการศึกษาแห่งชาติขั้นพื้นฐาน O-NET เพื่อประเมินคุณภาพการศึกษาระดับชาติประจำปีการศึกษา 2558 ที่เข้าสอบวิชาคณิตศาสตร์ และภาษาไทย จำนวนทั้งสิ้น 358,512 คน จำแนกตามเพศชาย จำนวน 304,751 คน และเพศหญิง จำนวน 285,412 คน

กลุ่มตัวอย่าง

กลุ่มตัวอย่างที่ใช้ในการวิจัยครั้งนี้ เป็นนักเรียนชั้นมัธยมศึกษาปีที่ 6 ที่เข้าทดสอบวัดผลสัมฤทธิ์ทางการเรียนโดยแบบสอบทางการศึกษาแห่งชาติขั้นพื้นฐาน O-NET เพื่อประเมินคุณภาพการศึกษาระดับชาติ ประจำปีการศึกษา 2558 วิชาคณิตศาสตร์ และภาษาไทย จำแนกตามตัวแปรเพศ และที่ตั้งทางภูมิศาสตร์ของโรงเรียนเป็นเขตกรุงเทพมหานครและปริมณฑล และนอกเขตกรุงเทพมหานครและปริมณฑล

การประมาณกลุ่มตัวอย่าง ผู้วิจัยได้ศึกษาการกำหนดขนาดกลุ่มตัวอย่างจากประชากรตามการวิเคราะห์คุณภาพตามทฤษฎีการทดสอบตอบสนองข้อสอบ (Item response theory: IRT) และตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (Differential item functioning: DIF) ซึ่งตามเงื่อนไขการกำหนดกลุ่มตัวอย่างเพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจะต้องไม่ต่ำกว่า 10-20 เท่า ของจำนวนข้อคำถาม (นงลักษณ์ วิรัชชัย, 2542, หน้า 311; สุวิมล ติรกานันท์, 2553, หน้า 168; สุภมาศ อังศุโชติ, สมถวิล วิจิตรวรรณา และรัชนีกูล ภิญโญภานุวัฒน์, 2554 หน้า 31 อ้างถึงใน สุพัฒนา หอมบุปผา, 2556, หน้า 109)

จากการศึกษาของเมเซอร์และคณะ (Mazor et al., 1992, pp. 443-451 อ้างถึงใน ปิยะทิพย์ ตินวร, 2549, หน้า 49) พบว่า เมื่อใช้กลุ่มตัวอย่างขนาดใหญ่ทำให้ตรวจพบข้อสอบทำหน้าที่ต่างกันได้มากกว่าการใช้กลุ่มตัวอย่างขนาดเล็ก คือ เมื่อใช้กลุ่มตัวอย่าง 2,000 คน มีความถูกต้องในการตรวจสอบ ร้อยละ 70 ถึง 75 และผลการศึกษาของจิติมา วรรณศรี (2539) พบว่า เมื่อใช้กลุ่มอ้างอิง 1,000 คน และกลุ่มเปรียบเทียบ 1,000 คน สามารถตรวจพบข้อสอบทำหน้าที่ต่างกัน

ได้ถูกต้อง ร้อยละ 100 และอัตราความคลาดเคลื่อนอยู่ในเกณฑ์ที่เหมาะสม จำนวนข้อสอบ O-NET มัธยมศึกษาปีที่ 6 ประจำปีการศึกษา 2558 มีจำนวน ดังนี้

วิชาคณิตศาสตร์	จำนวน	40	ข้อ
วิชาภาษาไทย	จำนวน	80	ข้อ
	รวม	120	ข้อ

ผู้วิจัยใช้กลุ่มตัวอย่าง 20 เท่า ของจำนวนข้อคำถาม ดังนั้นจะได้จำนวนกลุ่มตัวอย่างทั้งสิ้น 2,400 คน โดยมีขั้นตอนการสุ่มตัวอย่างแบบหลายขั้นตอน (Multi-stage random sampling technique) โดยมีขั้นตอนดังนี้

ขั้นตอนการสุ่มตัวอย่าง

1. ผู้วิจัยได้แบ่งประชากรในแต่ละจังหวัดออกเป็นประชากรย่อย (Subpopulation) โดยในแต่ละจังหวัดจะมีโรงเรียนเป็นสมาชิก กำหนดให้แต่ละโรงเรียนเป็น 1 กลุ่ม (Cluster) โดยทุกโรงเรียนที่อยู่ภายในชั้นภูมิเดียวกันหรือในเขตพื้นที่เดียวกันจะมีความแตกต่างกันมากที่สุด และระหว่างโรงเรียนภายในชั้นภูมิเดียวกันหรือในเขตพื้นที่เดียวกันจะมีความคล้ายคลึงกันมากที่สุด การแบ่งลักษณะนี้ใช้วิธีการสุ่มแบบกลุ่ม (Cluster sampling) ดำเนินการสุ่มโดยใช้วิธีการสุ่มตัวอย่างแบบกลุ่ม (Cluster random sampling) ซึ่งกำหนดให้โรงเรียนเป็นหน่วยการสุ่ม (Sampling unit) ผู้วิจัยจึงสุ่มโรงเรียนจากจังหวัด

2. เมื่อได้โรงเรียนจากการสุ่มข้อ 1 แล้ว ใช้ สถานที่ตั้งทางภูมิศาสตร์ของโรงเรียนเป็นตัวแปรจำแนกกลุ่ม (Classify variable) พบว่า แต่ละโรงเรียนจะถูกแบ่งออกเป็น 2 ชั้นภูมิ (Stratum) ได้แก่ เขตกรุงเทพมหานครและปริมณฑล และนอกเขตกรุงเทพมหานครและปริมณฑล จากนั้นทำการสุ่มตัวอย่างแบบชั้นภูมิ (Stratified random sampling) เมื่อได้กลุ่มตัวอย่างแล้วทำการแบ่งกลุ่มอีกครั้งโดยใช้เพศเป็นตัวแปรจำแนกกลุ่ม จะได้ชั้นภูมิย่อยอีก 2 ชั้นภูมิ คือ เพศชาย และเพศหญิง ผู้วิจัยได้ดำเนินการสุ่มโดยใช้วิธีการสุ่มตัวอย่างแบบชั้นภูมิ (Stratified random sampling) ได้กลุ่มตัวอย่าง จำนวน 2,400 คน รายละเอียดของจำนวนกลุ่มตัวอย่างที่ใช้ในการวิจัยแสดงได้ดังนี้

2.1 เขตกรุงเทพมหานคร และปริมณฑล มีจำนวนทั้งสิ้น 6 จังหวัด ได้แก่ กรุงเทพมหานคร, นครปฐม, นนทบุรี, สมุทรปราการ, ปทุมธานี และสมุทรสาคร

2.2 นอกเขตกรุงเทพมหานคร แบ่งได้เป็น 6 ภาค ได้แก่ ภาคตะวันออกเฉียงเหนือ ภาคกลาง ภาคใต้ ภาคตะวันออก ภาคเหนือ และภาคตะวันตก และจำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน ดังนี้

กรุงเทพมหานคร และปริมณฑล	จำนวน	1,200	คน
นอกเขตกรุงเทพมหานครและปริมณฑล	จำนวน	1,200	คน
รวมทั้งสิ้น		2,400	คน

ดำเนินการสุ่มโดยใช้วิธีการสุ่มตัวอย่างแบบกลุ่ม (Cluster random sampling) โดยมีโรงเรียนเป็นหน่วยสุ่มใน 6 จังหวัด ทำการสุ่มร้อยละ 20 ของจำนวนโรงเรียน ซึ่งมีรายละเอียดของจำนวนกลุ่มตัวอย่างที่ใช้ในการวิจัยแสดงได้ ดังนี้

ตารางที่ 3-1 จำนวนโรงเรียน และจำนวนการสุ่มตัวอย่างโรงเรียนในเขตกรุงเทพมหานครและปริมณฑล

จังหวัด	**จำนวนโรงเรียนทั้งหมด (แห่ง)**	**จำนวนสุ่มโรงเรียน (แห่ง)**	**กลุ่มตัวอย่างที่ต้องการ (คน)**
กรุงเทพมหานคร	50	10	632
สมุทรปราการ	14	3	189
นนทบุรี	11	2	126
ปทุมธานี	10	2	126
นครปฐม	6	1	63
สมุทรสาคร	2	1	63
รวมทั้งสิ้น	93	19	1,200

เมื่อทำการแบ่งอีกครั้ง โดยใช้เพศเป็นตัวแปรจำแนกกลุ่ม เราจะได้ชั้นภูมิย่อย ๆ อีก 2 ชั้นภูมิ คือ เพศชาย และเพศหญิง ได้กลุ่มตัวอย่างจำนวน 2,400 คน รายละเอียดของจำนวนกลุ่มตัวอย่างที่ใช้ในการวิจัยแสดงได้ดังนี้

ตารางที่ 3-2 ประชากรและกลุ่มตัวอย่างนักเรียนในเขตกรุงเทพมหานครและปริมณฑล จำแนกตามจังหวัด และโรงเรียน

จังหวัด	โรงเรียน	จำนวนนักเรียน					
		ประชากร			กลุ่มตัวอย่าง		
		ชาย	หญิง	รวม	ชาย	หญิง	รวม
กรุงเทพมหานคร	โรงเรียน 1	217	301	518	26	29	55
	โรงเรียน 2	524	866	1390	64	83	147
	โรงเรียน 3	268	432	700	32	41	73
	โรงเรียน 4	217	246	463	26	24	50
	โรงเรียน 5	285	383	668	35	37	72
	โรงเรียน 6	245	263	508	30	25	55
	โรงเรียน 7	196	205	401	24	20	44
	โรงเรียน 8	360	381	741	44	36	80
	โรงเรียน 9	199	313	512	24	30	54
	โรงเรียน 10	280	335	615	34	32	66
สมุทรปราการ	โรงเรียน 11	203	249	452	25	24	49
	โรงเรียน 12	134	256	390	16	25	41
	โรงเรียน 13	282	350	632	34	34	68
นนทบุรี	โรงเรียน 14	259	241	500	32	23	55
	โรงเรียน 15	432	438	870	52	42	94
ปทุมธานี	โรงเรียน 16	234	243	477	28	23	51
	โรงเรียน 17	145	184	329	17	18	35
นครปฐม	โรงเรียน 18	294	282	576	36	27	63
สมุทรสาคร	โรงเรียน 19	174	280	454	21	27	48
	รวมทั้งหมด	4,948	6,248	8,125	600	600	1,200

นอกเขตกรุงเทพมหานครและปริมณฑล ใช้เทคนิควิธีการสุ่มแบบหลายขั้นตอน (Multi-stage random sampling technique) ใช้วิธีการสุ่มตัวอย่างแบบกลุ่ม (Cluster random sampling) โดยมีจังหวัดเป็นหน่วยการสุ่ม (Sampling unit) ดังนี้

ตารางที่ 3-3 จำนวนจังหวัด จำนวนโรงเรียน และจำนวนการสุ่มตัวอย่างโรงเรียนนอกเขตกรุงเทพมหานคร และปริมณฑล

ภาค	จำนวนจังหวัด	จำนวนสุ่มจังหวัด	จำนวนโรงเรียน	จำนวนสุ่มโรงเรียน (แห่ง)	กลุ่มตัวอย่างที่ต้องการ (คน)
ตะวันออกเฉียงเหนือ	20	4	117	23	531
ภาคกลาง	16	3	49	10	231
ภาคใต้	14	3	42	8	185
ตะวันออก	7	1	26	5	115
ภาคเหนือ	8	2	18	4	92
ตะวันตก	5	1	8	2	46
รวมทั้งสิ้น	70	14	260	52	1,200

ตารางที่ 3-4 เขตที่ตั้งทางภูมิศาสตร์ จำนวนจังหวัด และจำนวนการสุ่มตัวอย่างโรงเรียนนอกเขตกรุงเทพมหานคร และปริมณฑล

ภูมิภาค	จังหวัด	โรงเรียน	จำนวนนักเรียน					
			ประชากร			กลุ่มตัวอย่าง		
			ชาย	หญิง	รวม	ชาย	หญิง	รวม
ตะวันออกเฉียงเหนือ	ขอนแก่น	โรงเรียน 1	349	547	896	20	21	41
		โรงเรียน 2	202	418	620	12	16	28
		โรงเรียน 3	199	454	563	12	17	29
		โรงเรียน 4	210	340	550	12	13	25
		โรงเรียน 5	141	246	387	8	10	18
		โรงเรียน 6	138	246	384	8	10	18
	นครราชสีมา	โรงเรียน 7	283	381	664	16	15	31
		โรงเรียน 8	173	282	455	10	11	21
		โรงเรียน 9	124	306	430	7	12	19
		โรงเรียน 10	200	319	519	12	12	24
		โรงเรียน 11	142	190	332	8	7	15

ตารางที่ 3-4 (ต่อ)

ภูมิภาค	จังหวัด	โรงเรียน	จำนวนนักเรียน					
			ประชากร			กลุ่มตัวอย่าง		
			ชาย	หญิง	รวม	ชาย	หญิง	รวม
	นครราชสีมา	โรงเรียน 12	115	246	361	7	10	17
		โรงเรียน 13	177	359	536	10	14	24
		โรงเรียน 14	135	323	458	8	12	20
	ร้อยเอ็ด	โรงเรียน 15	131	251	382	8	10	18
		โรงเรียน 16	193	330	523	11	13	24
		โรงเรียน 17	154	275	429	9	11	20
		โรงเรียน 18	412	289	701	24	11	35
		โรงเรียน 19	130	267	397	8	10	18
	อุดรธานี	โรงเรียน 20	147	256	403	9	10	19
		โรงเรียน 21	171	281	452	10	11	21
		โรงเรียน 22	504	469	973	29	18	47
		โรงเรียน 23	170	232	402	10	9	19
ภาคกลาง	สุพรรณบุรี	โรงเรียน 24	242	200	442	14	8	22
		โรงเรียน 25	146	214	360	8	8	16
		โรงเรียน 26	104	152	256	6	6	12
		โรงเรียน 27	120	209	329	7	8	15
	นครสวรรค์	โรงเรียน 28	143	187	330	8	7	15
		โรงเรียน 29	131	277	408	8	11	19
		โรงเรียน 30	285	394	679	17	15	32
	เพชรบูรณ์	โรงเรียน 31	252	271	523	15	10	25
		โรงเรียน 32	146	217	363	8	8	16
		โรงเรียน 33	201	267	468	12	10	22
ภาคใต้	นครศรีธรรมราช	โรงเรียน 34	161	380	541	9	15	24
		โรงเรียน 35	123	286	409	7	11	18
		โรงเรียน 36	107	177	284	6	7	13
	สงขลา	โรงเรียน 37	262	348	610	15	13	28
		โรงเรียน 38	290	509	799	17	20	37
		โรงเรียน 39	114	269	383	7	10	17

ตารางที่ 3-4 (ต่อ)

ภูมิภาค	จังหวัด	โรงเรียน	จำนวนนักเรียน					
			ประชากร			กลุ่มตัวอย่าง		
			ชาย	หญิง	รวม	ชาย	หญิง	รวม
	สุราษฎร์ธานี	โรงเรียน 40	201	287	488	12	11	23
		โรงเรียน 41	128	270	398	7	10	17
ตะวันตก	ชลบุรี	โรงเรียน 42	510	212	722	30	8	38
		โรงเรียน 43	132	296	428	8	11	19
		โรงเรียน 44	172	324	496	10	12	22
		โรงเรียน 45	198	306	504	12	12	24
		โรงเรียน 46	289	368	657	17	14	31
ภาคเหนือ	เชียงราย	โรงเรียน 47	300	320	620	17	12	29
		โรงเรียน 48	175	239	414	10	9	19
	เชียงใหม่	โรงเรียน 49	263	350	613	15	13	28
		โรงเรียน 50	146	282	428	8	11	19
ตะวันออก	ราชบุรี	โรงเรียน 51	153	377	530	9	14	23
		โรงเรียน 52	231	338	569	13	13	26
	รวม		10,325	15,633	25,868	600	600	1,200

เครื่องมือที่ใช้ในการวิจัย

เครื่องมือที่ใช้ในการวิจัยเป็นข้อมูลทุติยภูมิ ได้แก่ แบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนเพื่อประเมินคุณภาพการศึกษาระดับชาติของนักเรียนมัธยมศึกษาปีที่ 6 วิชาคณิตศาสตร์และภาษาไทย ที่สร้างโดยสถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน) ผู้สร้างแบบทดสอบประกอบด้วย ครูผู้สอน ศึกษานิเทศก์ นักวิชาการ และผู้เชี่ยวชาญที่เกี่ยวข้อง ซึ่งการสร้างแบบทดสอบเป็นไปตามวัตถุประสงค์ที่กำหนดในหลักสูตรนักเรียนมัธยมศึกษาปีที่ 6 ข้อสอบเป็นแบบปรนัย เลือกตอบ 5 ตัวเลือก ซึ่งประกอบด้วยข้อสอบ ดังต่อไปนี้

1. คณิตศาสตร์ จำนวน 40 ข้อ
2. ภาษาไทย จำนวน 80 ข้อ

รวมทั้งหมด 120 ข้อ

การเก็บรวบรวมข้อมูล

ผู้วิจัยรวบรวมข้อมูลผลการตอบของนักเรียนที่เป็นกลุ่มตัวอย่าง ซึ่งเป็นข้อมูลทุติยภูมิที่ได้จากสถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน) โดยมีขั้นตอนการเก็บรวบรวมข้อมูล ดังนี้

1. ขอหนังสือจากภาควิชาวิจัยและจิตวิทยาประยุกต์ คณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา เพื่อขอข้อมูลผลการตอบข้อสอบของนักเรียนที่เป็นกลุ่มตัวอย่าง จากสถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน)

2. ขอความร่วมมือจากเจ้าหน้าที่ในการสุ่มข้อมูลผลการตอบของนักเรียน แยกตามสามารถ เพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียนทั้งในเขตกรุงเทพมหานครและปริมณฑล และนอกเขตกรุงเทพมหานครและปริมณฑล ที่เป็นโรงเรียนสหศึกษา และเป็นโรงเรียนขนาดใหญ่พิเศษ เพื่อให้ได้กลุ่มตัวอย่างตามจำนวนที่ต้องการ

3. ตรวจสอบความเรียบร้อยของข้อมูล ได้แก่ โรงเรียนสหศึกษาและเป็นโรงเรียนขนาดใหญ่พิเศษ, สถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน, ความครบถ้วนของจำนวนข้อสอบและคำตอบของผู้ตอบแบบทดสอบที่ตรวจให้คะแนนแบบสองค่า (0 = ตอบผิด, 1 = ตอบถูกต้อง) ซึ่งเป็นนักเรียนที่สอบวัดผลสัมฤทธิ์ทางการเรียนเพื่อประเมินคุณภาพการศึกษาระดับชาติ ชั้นมัธยมศึกษาปีที่ 6 ปีการศึกษา 2558 แล้วนำข้อมูลมาวิเคราะห์ตามวัตถุประสงค์ของการวิจัย

การวิเคราะห์ข้อมูล

1. คำนวณค่าสถิติพื้นฐาน ได้แก่ ค่าเฉลี่ยเลขคณิต (Mean) ค่าความเบ้ (SK) ค่าความโด่ง (KU) ค่าเบี่ยงเบนมาตรฐาน (Standard deviation) เพื่อบรรยายการแจกแจงของคะแนนจากแบบทดสอบสอบวิชาคณิตศาสตร์ และภาษาไทย วิเคราะห์ความสัมพันธ์ด้วยค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน (Pearson's correlation) และ การหาความสอดคล้องด้วยการสร้างตารางไขว้ (Cross tabulation table) โดยใช้โปรแกรมสำเร็จรูป

2. การวิเคราะห์การประมาณค่าพารามิเตอร์อำนาจจำแนก (a) ค่าพารามิเตอร์ความยากของข้อสอบ (b) และค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) จากข้อมูลผลการตอบข้อสอบที่มีการให้คะแนนแบบ 2 ค่า (Dichotomous scores) ด้วยวิธี Maximum likelihood โดยโปรแกรม IPTPRO version 2.1, วิธี Bayesian โดยโปรแกรม Open BUGS Version 3.2.3 และวิธี CFA โดยโปรแกรม Mplus version 7.4

3. การเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบทั้ง 3 วิธีการ ได้แก่ วิธี IRT-LR วิธี BAYESIAN และวิธี MGCFA

3.1 การวิเคราะห์ด้วยวิธี IRT-LR

วิธีการทดสอบอัตราส่วนความควรจะเป็น (Likelihood ratio test) จะวิเคราะห์โดยใช้โปรแกรม IRT Pro (Cai, Toit & Thissen, 2011) ทดสอบสมมติฐานระหว่าง 2 โมเดล คือ โมเดลพื้นฐาน (Compact) และโมเดลเปรียบเทียบ (Augmented) ในโมเดลแรกสมมติให้มีกลุ่มผู้สอบที่แตกต่างกัน ดังนั้นจึงบังคับให้พารามิเตอร์ของข้อสอบระหว่างผู้สอบกลุ่มอ้างอิง และกลุ่มเปรียบเทียบมีค่าเท่ากัน (Group 1-item และ Group 2-item 1) (Thissen, Steinberg & Wainer, 1988, 1993 อ้างถึงใน อาวีพร ปานทอง, 2558, หน้า 39) โดยใช้สมมติฐานในการทดสอบตามสูตร ดังนี้

$H_0 : a_{\mathrm{iF}} = a_{iR}$ และ $b_{\mathrm{ijF}} = b_{\mathrm{ijR}}$ สำหรับทุก j

H_A : พารามิเตอร์ของข้อสอบข้อที่ j ของทั้งสองกลุ่มไม่เท่ากัน อย่างน้อย 1 พารามิเตอร์

สูตรการทดสอบอัตราส่วนความน่าจะเป็นของสองโมเดล แสดงได้ดังนี้

$$LR = \frac{L^*(Model_C)}{L^*(Model_A)} \tag{34}$$

เมื่อ $L^*(Model_C)$ คือ ฟังก์ชันความน่าจะเป็นของโมเดลพื้นฐาน (ค่าพารามิเตอร์น้อยกว่า)

$L^*(Model_A)$ คือ ฟังก์ชันความน่าจะเป็นของโมเดลเปรียบเทียบที่ยอมให้พารามิเตอร์ข้อสอบของข้อที่ j ระหว่างกลุ่มผู้สอบมีความหลากหลาย

กำหนดฟังก์ชันไลค์ลิฮูด (Likelihood function) ในการประมาณค่าพารามิเตอร์จากค่าที่มากที่สุดในฟังก์ชัน (Maximum likelihood function) ซึ่งสามารถแสดงได้ดังสูตรทางคณิตศาสตร์ดังนี้

$$L(X \mid \theta, a, b) = \prod_{i=1}^{N} \prod_{j=1}^{J} \prod_{h=1}^{k} P(\theta_i, a_j, b_j)^{x_{ij}} \left[1 - P(\theta_i, a_j, b_j)^{1-x_{ij}} \right] \tag{35}$$

กำหนดให้ i แทน ผู้ตอบคำถามคนที่ 1, 2, …….N

j แทน ข้อคำถาม ข้อที่ 1, 2……...J

h แทน ตัวเลือกในข้อคำถาม ตัวเลือกที่ 1, 2…….k

นั่นคือโมเดลพื้นฐาน (Compact; $\mathrm{Model_C}$) จะประกอบด้วยข้อสอบที่ทำหน้าที่ไม่ต่างกัน สำหรับในโมเดลเปรียบเทียบจะประกอบด้วยข้อสอบที่มีค่าพารามิเตอร์ระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่มีค่าแปรเปลี่ยนไปตามกลุ่มโมเดลเปรียบเทียบ (Augmented; $\mathrm{Model_A}$) อาจมีข้อสอบจำนวน 1 ข้อ หรือมากกว่าที่ทำหน้าที่ต่างกัน ทำการเปรียบเทียบระหว่าง

2 โมเดล ด้วยสถิติอัตราส่วนความน่าจะเป็นได้ดังนี้

$$G_i^2 = -2\log(\mathrm{LR})$$

$$G_i^2 = -2\ln(L^*(Model_C)) - \left[-2\ln\left[L^*(Model_A)\right]\right] \quad (36)$$

เมื่อ G_i^2 แทน สถิติอัตราส่วนความควรจะเป็นของข้อสอบข้อที่ i

$L^*(Model_C)$ แทน ฟังก์ชันความควรจะเป็นของโมเดลพื้นฐาน (Compact)

$L^*(Model_A)$ แทน ฟังก์ชันความควรจะเป็นของโมเดลเปรียบเทียบ (Augmented)

โดยทั่วไปแล้ว $L^*(Model_C) < L^*(Model_A)$ และสถิติ $G_i^2 > 0$ มีการแจกแจงแบบไค-สแควร์ ซึ่งระดับของความเป็นอิสระ เท่ากับผลต่างของจำนวนพารามิเตอร์ในโมเดล $L^*(Model_C)$ และ $L^*(Model_A)$ (ATAR, 2011) ดัชนีความไม่กลมกลืนในระดับนัยสำคัญทางสถิติของผลการทดสอบจะชี้บอกว่าโมเดลพื้นฐานกลมกลืนน้อยกว่าโมเดลเปรียบเทียบอย่างมีนัยสำคัญทางสถิติ ในการใช้ทดสอบอัตราส่วนความควรจะเป็นค่าไค-สแควร์ ที่องศาความเป็นอิสระ เท่ากับจำนวนพารามิเตอร์ของข้อสอบในการประมาณค่าของข้อสอบเหล่านั้นมีนัยสำคัญทางสถิติ แสดงว่ามีการทำหน้าที่ต่างกันของข้อสอบ

2. การวิเคราะห์ด้วยวิธี Bayesian

ในการวิจัยครั้งนี้ผู้วิจัยต้องการศึกษาผลการประมาณค่าอำนาจจำแนก (a), ค่าความยากของข้อสอบ และค่าความสามารถของผู้สอบ (θ) ตามกระบวนการและแนวคิดของเบส์ ซึ่งมีแนวการวิเคราะห์ ดังนี้

การประมาณค่าพารามิเตอร์ของข้อสอบ และความสามารถของผู้เข้าสอบกระบวนการประมาณค่าพารามิเตอร์ด้วยวิธีของเบส์ คือ การหาค่าประมาณ θ_i โดยที่ i = 1, 2, ……N

และ b_j, a_j c_j โดยที่ j = 1, 2, ……n ที่ทำให้ฟังก์ชันการแจกแจงภายหลัง f (θ, b, a, c|u) มีค่าสูงสุด

ถ้า In f (θ, b, a, c|u) เป็นฟังก์ชันที่หาอนุพันธ์ได้ การหาค่า θ_i, $b_j a_j$ และ c_j จะหาได้จากอนุพันธ์ของ

In f (θ, b, a, c|u) ซึ่ง และกำหนดให้อนุพันธ์ของ In f (θ, b, a, c/ u) มีค่า เท่ากับศูนย์ แล้วจึงหาค่ารากของอนุพันธ์ของ In f (θ, b, a, c|u) ได้จากสมการต่อไปนี้

$$f(\theta, b, a, c|u) = L(u \mid \theta, a, b, c).f(\theta).f(b).f(a).f(u)$$

$$\text{In } f(\theta, b, a, c|u) = \text{In } L(u \mid \theta, b, a, c).\text{In } f(\theta) + \text{In } f(b) + \text{In } f(a) + \text{constant}$$

$$d\,\text{In } f(\theta, b, a, c \mid u)/d\theta_i = 0$$

$$d\,\text{In } f(\theta, b, a, c \mid u)/db_j = 0$$

$$d\,In\,f(\theta,b,a,c\,|\,u)/da_j = 0$$

$$d\,In\,f(\theta,b,a,c\,|\,u)/dc_j = 0$$

สมการอนุพันธุ์ของ In f (θ, b, a, c / u) = 0 เรียกว่า สมการโมเดล (Model equation) ค่ารากของสมการโมเดล คือ ค่าความสามารถ θ และค่าพารามิเตอร์ของข้อสอบ b, a และ c ที่ทำให้ฟังก์ชันการแจกแจงภายหลังมีค่าสูงสุด อาจทำได้โดยใช้เทคนิค Newton-raphson ซึ่งเป็นการหาค่าประมาณโดยการหาค่าซ้ำ (Iterative) และมีขั้นตอนดังต่อไปนี้

ขั้นที่ 1 กำหนดค่าเริ่มต้นสำหรับใช้ในการประมาณค่าความสามารถ θ_i และค่าพารามิเตอร์ของข้อสอบ, a_j และ b_j มีค่าเริ่มต้นสำหรับใช้ในการประมาณค่าความสามารถ θ_i

$$\theta_i^{(0)} = \ln\left(\frac{x}{n - x_i}\right)$$

$$a_i^{(0)} = R_j/(1 - R_j^2)^{1/2}$$

$$b_i^{(0)} = Z_j / R_j$$

$$c_i^{(0)} = 1/m_j$$

เมื่อ In = Natural Logarithm

x_i = คะแนนสอบของผู้เข้าสอบคนที่ i

n = จำนวนข้อสอบ

R_j = Point-Biserial Correlation

Z_j = ค่า Z ของการแจกแจงปกติมาตรฐานที่พื้นที่ใต้โค้งปกติมาตรฐานด้านขวามือ มีค่า เท่ากับ P_j (U_j/ N)

N = จำนวนผู้สอบทั้งหมด

m_j = จำนวนตัวเลือกในข้อที่ j

ขั้นที่ 2 ทำการประมาณค่าความสามารถของผู้เข้าสอบแต่ละคน θ_i โดยกำหนดให้ค่าพารามิเตอร์ของข้อสอบที่ประมาณค่าได้ในครั้งก่อนเป็นค่าคงที่

$$\theta_i^{m+1} = \theta_i^{(m)} - g(\theta_i^{(m)} / h(\theta_i^{(m)}) \qquad (37)$$

เมื่อ θ_i^m, $\theta^{(m+1)}$ = ค่าประมาณความสามารถของคนที่ i ครั้งที่ m และ $m+1$

$g(\theta_i^m)$ = d In f (θ, b, a, c|u) / dθ_i

$h(\theta_i^{(m)})$ = d^2 In f (θ, b, a, c|u) / dθ_i^2

การประมาณค่าซ้ำ (Iterative) จะกระทำจนกว่าค่าประมาณความสามารถ θ_i จะข้าสู่ค่าคงที่ค่าใดค่าหนึ่ง (Convergence) คือ ค่าประมาณครั้งที่ m และครั้งที่ m+1 มีค่าแตกต่างกันน้อยกว่าค่าคงที่ที่กำหนดไว้ เช่น 0.001

ขั้นที่ 3 ประมาณค่าพารามิเตอร์ของข้อสอบแต่ละข้อ b_j, a_j และc_j โดยกำหนดให้ค่าความสามารถของผู้เข้าสอบที่ประมาณค่าได้ในครั้งก่อนเป็นค่าคงที่

$$a_j^{(m+1)} = a_j^{(m)} - g(a_j^{(m)}) / h(a_j^{(m)})$$

$$b_j^{(m+1)} = b_j^{(m)} - g(b_j^{(m)}) / h(b_j^{(m)})$$

$$c_j^{(m+1)} = c_j^{(m)} - g(c_j^{(m)}) / h(c_j^{(m)})$$

เมื่อ $a_j^m a_j^{m+1}$ = ค่าประมาณอำนาจจำแนกของข้อสอบข้อที่ j ครั้งที่ m และ m+1

b_j^m, b_j^{m+1} = ค่าประมาณความยากของข้อสอบข้อที่ j ครั้งที่ m และ m+1

c_j^m, c_j^{m+1} = ค่าประมาณโอกาสการเดาของข้อที่ j ครั้งที่ m และ m+1

$g(b_j^m)$ = $d \ln f(\theta, b, a, c|u) / db_j$

$h(b_j^m)$ = $d^2 \ln f(\theta, b, a, c|u) / db_j^2$

$g(a_j^m)$ = $d \ln f(\theta, b, a, c|u) / da_j$

$h(a_j^m)$ = $d^2 \ln f(\theta, b, a, c|u) / db_j^2$

$g(c_j^m)$ = $d \ln f(\theta, b, a, c|u) / dc_j$

$h(c_j^m)$ = $d^2 \ln f(\theta, b, a, c|u) / dc_j^2$

การประมาณค่าซ้ำ (Iterative) จะกระทำจนกว่าค่าประมาณจะเข้าสู่ค่าคงที่ (Convergence) ค่าใดค่าหนึ่ง สำหรับค่าอนุพันธุ์อันดับที่ 1 g (x) และอนุพันธุ์อันดับที่ 2 h (x) ประกอบ 2 ส่วน คือ ส่วนที่ได้จากฟังก์ชันไลค์ลิฮูด และส่วนที่ได้จากการแจกแจงเริ่มแรก

ขั้นที่ 4 ประมาณค่าซ้ำ ขั้นที่ 2 และขั้นที่ 3 จนกว่าค่าประมาณ θ_i, b_j, a_j และ c_j จะมีค่าคงที่และมีความถูกต้องเพียงพอ หรือทำให้ $f(\theta, b, a/ u)$ มีค่าสูงสุด

สำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี BAYESIAN โดยประยุกต์ใช้โปรแกรม Open BUGS สามารถคำนวณได้ดังนี้

สำหรับการวิเคราะห์ค่าพารามิเตอร์ และการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี BAYESIAN โดยประยุกต์ใช้โปรแกรม Open BUGS ผู้วิจัยได้ทำการวิเคราะห์ค่าพารามิเตอร์และการทำหน้าที่ต่างกันของข้อสอบ (DIF) (Fukuhara & Kamata (2011) แสดงสมการได้ ดังนี้

$$\text{In}\left[\frac{P(y_{ij}=1}{p(y_{ij}=0}\right]=a_j(\beta_\theta G_i+\xi_i-b_j-\beta_j G_i) \quad (38)$$

เมื่อ β_θ หมายถึง อิทธิพลของกลุ่ม G_i ต่อค่าความสามารถ θ_i

G_j คือ กลุ่มของผู้สอบ โดย Focal group = 0 และ Reference = 1

β_i คือ ค่าสัมประสิทธิ์ที่ใช้พิจารณาการทำหน้าที่ต่างกันของข้อสอบ

$$\beta_\theta^{adj}=\beta_\theta-\bar{\beta}, \text{and} \quad (39)$$

$$\beta_i^{adj}=\beta_i-\bar{\beta}, \quad (40)$$

โดยที่ β_θ^{adj} และ β_i^{adj} คือ การปรับพารามิเตอร์สำหรับค่าเฉลี่ยของความสามารถที่แตกต่างกัน และขนาดของ DIF ตามลำดับ และค่า $\bar{\beta}$ แสดงถึงค่าเฉลี่ยขนาดของ DIF ของข้อสอบทุกข้อ การประมาณค่าพารามิเตอร์จะบังคับให้ค่าเฉลี่ยขนาดของ DIF ของข้อสอบทุกข้อมีค่าเป็น 0 ซึ่งเป็นมาตรฐานของการทดสอบ และเป็นแนวทางที่คล้ายกันกับโมเดลการตอบสนองข้อสอบพื้นฐาน (IRT-based DIF) ที่ใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

สำหรับการพิจารณาการทำหน้าที่ต่างกันของข้อสอบ จะพิจารณาจากค่า β_j ซึ่งถ้าค่า β_j มีค่าขอบล่างของช่วงความเชื่อมั่น 95% ซึ่งตรงกับตำแหน่งเปอร์เซ็นไทล์ที่ 2.5 (val 2.5 pc) และค่าขอบบนของช่วงความเชื่อมั่น 95 % ซึ่งตรงกับตำแหน่งเปอร์เซ็นไทล์ที่ 97.5 (val 97.5 pc) ไม่คลุม 0 แสดงว่าข้อสอบข้อนั้นเกิดการทำหน้าที่ต่างกันของข้อสอบ (DIF)

ผู้วิจัยได้แสดงผลการวิเคราะห์ข้อมูลในช่วง 2001-10000 เนื่องจากความถี่เริ่มคงที่ตั้งแต่ 2001 ขึ้นไป โดยการ BURN ข้อมูลก่อนหน้า 2001

3. การวิเคราะห์ด้วยวิธี Multiple group confirmatory factor analysis: MGCFA

การวิเคราะห์ครั้งนี้ใช้หลักการของการวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ (MGCFA) ในการตรวจสอบหาการทำหน้าที่ต่างกันของข้อสอบ (DIF) กรอบการสร้างโมเดล (Modeling framework) ในการวิจัยครั้งนี้ คือ ต้องการให้มีความเท่าเทียมกันของค่าน้ำหนักองค์ประกอบ (Factor loadings) และจุดตัด (Thresholds) โดยการบังคับให้เท่ากันในทั้งสองกลุ่มที่ศึกษา ซึ่งโมเดลที่จะทำการศึกษาจะประกอบด้วย 2 โมเดล ได้แก่ โมเดลที่ไม่มีการทำหน้าที่ต่างกันของข้อสอบ คือ No-DIF model และโมเดลที่มีการทำหน้าที่ต่างกันของข้อสอบ คือ DIF model ที่จะใช้เพื่อตรวจสอบโมเดลการตอบสนองข้อสอบแบบ 2 พารามิเตอร์ (IRT-model 2 PL) และการทำหน้าที่ต่างกันของข้อสอบ หลักการ คือ การควบคุมให้ข้อสอบที่นำมาตรวจสอบหา DIF ไม่ให้มีข้อสอบทำหน้าที่ต่างกัน (No-DIF model) ในขั้นตอนแรก คือ ก่อนปรับโมดล

แล้วจึงหาข้อสอบที่ทำหน้าที่ต่างกันโดยการปรับโมเดล (DIF model) ในขั้นตอนสุดท้ายภายหลังปรับโมเดล ในกรณีของโมเดล No-DIF ก็จะไม่มีผลกระทบโดยตรง (Direct effects) ของตัวแปรเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน หรือผลกระทบของความไม่แปรเปลี่ยนของโมเดลการวัด (Measurement invariance) ตัวแปรแฝงจะถูกตรวจสอบด้วยการเปรียบเทียบด้วยการวิเคราะห์โมเดล No-DIF และโมเดล DIF การวิเคราะห์องค์ประกอบเชิงยืนยัน (Confirmatory factor analysis: CFA ก็จะถูกใช้ในการตรวจสอบความเป็นโมเดลการวัดแบบหนึ่งมิติ หรือโมเดลการวัดแบบสองมิติ (Mathew Niti et al., 2007, p. 366)

ขั้นตอนการวิเคราะห์ MGCFA ได้ดำเนินการวิเคราะห์ด้วยโปรแกรม Mplus ด้วยการวิเคราะห์โมเดลตอบสนองข้อสอบ (IRT-2PL) ด้วยการเขียนโปรแกรม สำหรับการวิเคราะห์ได้ดำเนินการตามลำดับ

ขั้นตอนของการทดสอบ No-DIF model และ DIF model

1. เขียนคำสั่ง Syntax ในโปรแกรมสร้างให้เป็นโมเดล No-DIF โดยการบังคับให้ Factor loading และ Thresholds มีค่าเท่ากัน เพื่อให้โมเดล No-DIF ไม่มีข้อสอบที่ทำหน้าที่ต่างกัน (DIF) และทำการทดสอบโดยการเลือกข้อสอบที่จะดำเนินการกับโมเดล DIF หลังปรับโดยดูที่ค่าโมเดลดัชนีปรับแก้ (Modification indices) เลือกข้อสอบที่มีค่า M.I. ที่มีค่าสูงในลำดับต้น ๆ ในส่วนของ Factor loading ประมาณ 4-5 ค่า เพื่อนำเข้าทดสอบในคำสั่ง DIF and NonDIF

2. จัดเตรียมข้อมูลนำเข้าสำหรับการวิเคราะห์ ด้วยโปรกรม Mplus สามารถจัดเตรียมข้อมูลนำเข้าด้วย โปรแกรม SPSS โดยสามารถจัดเตรียมข้อมูลทั้งหมดให้ อยู่ในไฟล์เดียวกัน เสร็จแล้ว SAVE ข้อมูลเป็น CSV หรือแปลงไฟล์ข้อมูลให้เป็นภาษา ASCII โดยการบันทึก Save as type: Fixed ASCII (*.dat) หรือ Save as type: Tab-delimited (*.dat) เมื่อบันทึกเสร็จแล้ว ข้อมูลจะอยู่ในไฟล์นามสกุล .dat ซึ่งสามารถเปิดดูได้ ด้วยโปรแกรม Notepad ข้อมูลที่นำเข้าวิเคราะห์ ได้แก่ MATH 40.csv และ THAI 80 .csv ทำการ Run โปรแกรม ตามการจำแนกกลุ่มที่ต้องการตรวจสอบหา DIF ได้แก่ กลุ่มเพศ (ชาย, หญิง) ตามกลุ่มที่กำหนดให้ ดังนี้

วิชาคณิตศาสตร์ จำแนกตามเพศ กำหนดให้เพศ ชาย เป็น Reference group (เพศชาย = Men) และเพศหญิงเป็น Focal group (เพศหญิง = Women) จำแนกตามที่ตั้งทางภูมิศาสตร์ของโรงเรียน กำหนดให้เขตกรุงเทพมหานครและปริมณฑลเป็น Reference group (เขตกรุงเทพมหานคและปริมณฑล = IN) และนอกเขตกรุงเทพมหานคและปริมณฑลเป็น Focal group (นอกเขตกรุงเทพมหานครและปริมณฑล = OUT)

วิชาภาษาไทย จำแนกตามเพศ กำหนดให้เพศหญิง (เพศหญิง = Women) เป็น Reference group และเพศชาย (เพศชาย = Men) เป็น Focal group จำแนกตามที่ตั้งทางภูมิศาสตร์ของโรงเรียน กำหนดให้เขตกรุงเทพมหานครและปริมณฑลเป็น Reference group (เขตกรุงเทพ

มหานครและปริมณฑล = IN) และนอกเขตกรุงเทพมหานครและปริมณฑลเป็น Focal group (นอกเขตกรุงเทพมหานครและปริมณฑล = OUT)

3. Run ข้อมูลในโปรแกรม Mplus โดยการนำข้อมูลวิเคราะห์ทีละกลุ่มทั้งกลุ่มเพศ และกลุ่มที่ตั้งทางภูมิศาสตร์ของโรงเรียน โดยในขั้นแรกทำการวิเคราะห์หาค่าพารามิเตอร์อำนาจจำแนกของข้อสอบ, พารามิเตอร์ความยากของข้อสอบ และพารามิเตอร์ค่าความสามารถของผู้หญิงของทุกกลุ่ม

4. ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ จำแนกตามกลุ่มเพศ และกลุ่มที่ตั้งทางภูมิศาสตร์ของโรงเรียน เพื่อตรวจสอบค่า χ^2 และ ค่า *df* เพื่อเป็นค่าตั้งต้นสำหรับใช้เป็นค่าเปรียบเทียบเพื่อหา DIF ของข้อสอบรายข้อ ตรวจสอบค่า χ^2 ของข้อสอบในกลุ่มที่ต้องการนำมาเปรียบเทียบ (กลุ่ม Focal) เพื่อที่จะนำเข้าตรวจสอบหา DIF ในคำสั่งที่เตรียมไว้ คือ DIFandNonDIF เลือกข้อสอบที่มีค่า M.I. ที่มีค่าสูงในลำดับต้น ๆ ในส่วนของ Factor loading ประมาณ 4-5 ค่า เพื่อนำเข้าทดสอบในคำสั่ง DIFandNonDIF

5. หลังจาก Run โปรแกรมแล้ว ให้ดูที่ค่า M.I ภายหลังจากนำข้อสอบที่เลือกได้ Run ในคำสั่ง DIFandNonDIF แล้วให้ทำการตรวจสอบดูค่า χ^2, *df* ซึ่งค่า χ^2 ที่ได้ควรจะมีค่าลดลงจากค่า χ^2 ของโมเดล No-DIF และค่า *df* ควรจะมีค่าลดลง หากพบว่าข้อสอบข้อนั้นมีค่า χ^2 และ *df* ลดลง นำเข้าโปรแกรม Excel เพื่อตรวจสอบนัยสำคัญทางสถิติที่ .05 โดยหาค่าสหสัมพันธ์ระหว่างองค์ประกอบได้แก่ ค่า ϕ, χ^2, *df* และค่า p-value ต่อไป ถ้าพบว่าข้อสอบข้อนั้นมีนัยสำคัญทางสถิติที่ .05 แล้ว ให้ทำการนำข้อสอบข้อที่มี ค่า χ^2 สูงข้อถัดมาเข้าทำการ RUN ด้วยวิธีเดิม ถ้าพบว่า χ^2 และ *df* เพิ่มขึ้นให้หยุด Run ในส่วนของ Factor loading ทันที

6. นำเข้าข้อสอบที่พบในส่วนของ Threshold มา Run ใน DIF and Non-DIF เพื่อตรวจสอบ DIF หากพบ χ^2 และ *df* ลดลง และมีนัยสำคัญทางสถิติให้บันทึกว่าข้อสอบข้อนั้นเกิดการทำหน้าที่ต่างกัน (DIF) ในการ Run ข้อสอบจะทำการ Run ต่อเนื่อง โดยไม่นำข้อสอบข้อที่พบ DIF ออกไป แต่จะนำข้อที่ต้องการตรวจสอบนำเข้าไป Run ต่อ แล้วบันทึกข้อสอบที่พบว่า มีนัยสำคัญทางสถิติแสดงว่าข้อสอบข้อนั้นทำหน้าที่ต่างกันของข้อสอบ

3. เปรียบเทียบวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจาก 3 วิธีการ

ได้แก่ วิธี IRT-LR, BAYESIAN และวิธี MGCFA จำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน ด้วยการสร้างตารางไขว้ (Cross tab) ทำการเปรียบเทียบครั้งละ 2 วิธีการ โดยใช้โปรแกรมสำเร็จรูป เพื่อตรวจสอบความสอดคล้องของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยการตรวจสอบจำนวนข้อสอบที่ DIF และจำนวนข้อสอบที่ Non-DIF ที่สอดคล้องกัน นับจำนวนข้อและร้อยละ, ตรวจสอบค่าสหสัมพันธ์องค์ประกอบ (Φ), χ^2, *df* และ p-value

พร้อมทั้งแปลผลความสอดคล้องของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้ง 3 วิธี ดังแสดงในตารางที่ 3-5 ดังต่อไปนี้

ตารางที่ 3-5 การทดสอบการทำหน้าที่ต่างกันด้วย No-DIF model และ DIF model

No-DIF Model	DIF Model
การทดสอบความเท่าเทียมกันของค่าพารามิเตอร์ Factor loading และ Thresholds **การทดสอบ No-DIF Model** สร้างคำสั่ง SYNTAX ตั้งชื่อว่า NonDIF กำหนดให้ค่า Factor loading และ Thresholds มีค่าเท่ากันทั้งสองกลุ่มโดยการเขียนคำสั่ง หลัง RUN โปรแกรมจะได้ค่า χ^2 และ *df* เพื่อใช้เป็นค่าตั้งต้นในการเปรียบเทียบเพื่อหา DIF ของข้อสอบรายข้อ เขียนคำสั่งสำหรับที่สำคัญของ No-DIF Model ยกตัวอย่างวิชาคณิตศาสตร์ดังนี้ กรณีผู้ชายเป็นกลุ่มอ้างอิง No-DIF Model for men reference group **Model men** Factor loadings all constraint equal to focal group Item thresholds all free Factor mean = 0 and Variance = 1 for Identification **Model women** Factor loadings all constraint equal to Focal group **Item thresholds all set equal to reference group**	การทดสอบความเท่าเทียมกันของค่าพารามิเตอร์ Factor loading และ thresholds **การทดสอบ DIF Model** สร้างคำสั่ง SYNTAX ตั้งชื่อว่า DIFandNonDIF 1. นำข้อสอบที่มีค่า χ^2 ของข้อสอบที่มีค่าสูงมาก 4-5 ลำดับต้นที่พบในคำสั่ง No-DIF ดูที่ค่าโมเดลดัชนีปรับแก้ (Modification indices) เพื่อที่จะนำเข้าตรวจสอบหา DIF ในคำสั่งที่ เขียนคำสั่งที่สำคัญเกี่ยวกับ DIF-Model ยกตัวอย่างวิชาคณิตศาสตร์ ดังนี้ กรณีผู้ชายเป็นกลุ่มอ้างอิง DIF Model for men reference group **Model men** Factor loadings all constraint equal to focal group math by q1-q40* (a1-a40); Item thresholds all free Item residual variance all fixed = 1 Factor mean = 0 and Variance = 1 for Identification

ตารางที่ 3-5 (ต่อ)

No-DIF Model	DIF Model
Item residual variance all fixed = 1 Factor mean = 0 and Variance = 1 for Identification	**Model women** Factor loadings all constraint equal to reference group **Item thresholds constrains some item free for focal group** Item residual variances all fixed = 1 Factor mean = 0 and variance = 1 for identification

บทที่ 4
ผลการวิเคราะห์ข้อมูล

การวิเคราะห์ครั้งนี้มีจุดมุ่งหมาย 1) เพื่อศึกษาผลการประมาณค่าอำนาจจำแนก (a) ของข้อสอบ, ค่าพารามิเตอร์ความยาก (b) ของข้อสอบ, ค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) ด้วยวิธี Maximum likelihood, วิธี BAYESIAN และวิธี Confirmatory factor analysis และ 2) เพื่อเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบที่มีการให้คะแนนแบบสองค่า (Dichotomous scores) ระหว่างวิธี IRT-LR, วิธี BAYESIAN และวิธี MGCFA ซึ่งผลการวิเคราะห์ข้อมูลครั้งนี้ผู้วิจัยได้เรียงตามลำดับดังนี้

ตอนที่ 1 การวิเคราะห์ค่าสถิติพื้นฐาน

ตอนที่ 2 เปรียบเทียบผลการประมาณค่าพารามิเตอร์อำนาจจำแนก, ค่าพารามิเตอร์ความยากของข้อสอบ และค่าพารามิเตอร์ความสามารถของผู้สอบจำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน

ตอนที่ 3 เปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบจำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน

สัญลักษณ์ที่ใช้ในการวิเคราะห์ข้อมูล

เพื่อให้การนำเสนอการวิเคราะห์ข้อมูลมีความเข้าใจตรงกัน ผู้วิจัยจึงได้กำหนดสัญลักษณ์และความหมายแทน ดังนี้

ML	หมายถึง วิธี Maximum likelihood
CFA	หมายถึง วิธี Confirmatory factor analysis
IRT-LR	หมายถึง วิธีการวิเคราะห์ข้อมูลด้วยวิธีการทดสอบอัตราส่วนความควรจะเป็น (Item response theory-Likelihood ratio)
BAYESIAN	หมายถึง วิธีการวิเคราะห์ข้อมูลด้วยทฤษฎีของเบส์ (Bayesian estimation)
MGCFA	หมายถึง วิธีการวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ (Multiple group confirmatory factor analysis: MGCFA)
$\overline{X}$	หมายถึง ค่าเฉลี่ย (Mean)
SD	หมายถึง ค่าส่วนเบี่ยงเบนมาตรฐาน (Standard deviation)

Max	หมายถึง ค่าสูงที่สุด (Maximum)
Min	หมายถึง ค่าต่ำที่สุด (Minimum)
SK	หมายถึง ค่าความเบ้ (Skewness)
KU	หมายถึง ค่าความโด่ง (Kurtosis)
SE	หมายถึง ค่าความคลาดเคลื่อนมาตรฐานในการประมาณค่า (Standard error)
a	หมายถึง ค่าพารามิเตอร์อำนาจจำแนก (Discrimination parameter)
b	หมายถึง ค่าพารามิเตอร์ความยาก (Difficulty parameter)
θ	หมายถึง ค่าพารามิเตอร์ความสามารถของผู้สอบ (Ability parameter)
χ^2	หมายถึง ค่าสถิติไค-สแควร์ (Chi-square)
df	หมายถึง องศาความเป็นอิสระ (Degree of freedom)

สัญลักษณ์ที่ใช้แทนตัวแปรสังเกตได้

Q 01-Q 40	หมายถึง ข้อสอบวิชาคณิตศาสตร์ ตั้งแต่ข้อที่ 1 ถึงข้อที่ 40
Q 01-Q 80	หมายถึง ข้อสอบวิชาภาษาไทย ตั้งแต่ข้อที่ 1 ถึงข้อที่ 80
*Val*2.5*pc*	หมายถึง ค่าขอบล่างของช่วงความเชื่อมั่น 95% ซึ่งตรงกับตำแหน่งเปอร์เซ็นไทล์ที่ 2.5
*Val*97.5*pc*	หมายถึง ค่าขอบบนของช่วงความเชื่อมั่น 95% ซึ่งตรงกับตำแหน่งเปอร์เซ็นไทล์ที่ 97.5

ตอนที่ 1 การวิเคราะห์ค่าสถิติพื้นฐาน

การวิเคราะห์ข้อมูลส่วนนี้ผู้วิจัยได้นำคะแนนที่ได้จากการตอบข้อสอบของแบบทดสอบวิชาคณิตศาสตร์ และภาษาไทย มีจำนวน 40 ข้อ และ 80 ข้อ ตามลำดับ ซึ่งจำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน

ตารางที่ 4-1 แสดงจำนวนผู้สอบ ค่าต่ำสุด ค่าสูงที่สุด ค่าเฉลี่ย ค่าส่วนเบี่ยงเบนมาตรฐาน ค่าความเบ้ และค่าความโด่ง จากคะแนนการตอบแบบทดสอบวิชาคณิตศาสตร์ จำแนกตามเพศผู้สอบ

	n	***Min***	***Max***	***Mean***	***SD***	***SK***	***KU***
เพศชาย	1,200	1.00	40.00	14.69	9.27	1.03 (.07)	.027 (.14)
เพศหญิง	1,200	1.00	39.00	14.26	8.30	1.11 (.07)	.46 (.14)
รวม	2,400	1.00	40.00	14.48	8.80	1.07 (.05)	.24 (.10)

หมายเหตุ ค่าความเบ้/ ค่าความคลาดเคลื่อนมาตรฐานของความเบ้ และค่าความโด่ง/ ค่าความคลาดเคลื่อนมาตรฐานของความโด่งมีค่าอยู่ระหว่าง -1.96 ถึง 1.96 ลักษณะของการแจกแจงเป็นโค้งปกติ

จากตารางที่ 4-1 พบว่า ผู้สอบจำนวนทั้งหมด 2,400 คน มีคะแนนต่ำที่สุด เท่ากับ 1 คะแนน และสูงที่สุด เท่ากับ 40 คะแนน มีค่าเฉลี่ย เท่ากับ 14.48 คะแนน และมีส่วนเบี่ยงเบนมาตรฐาน เท่ากับ 8.80 โค้งการแจกแจงมีลักษณะเบ้ขวา และมีความโด่งน้อย โดยมีความเบ้ เท่ากับ 1.07 ค่าความคลาดเคลื่อนมาตรฐานของความเบ้ เท่ากับ .05 และค่าความโด่ง เท่ากับ .24 ค่าความคลาดเคลื่อนมาตรฐานของความโด่ง เท่ากับ .10 เมื่อพิจารณาแล้ว พบว่า เพศชาย จำนวน 1,200 คน มีคะแนนต่ำที่สุด เท่ากับ 1 คะแนน และสูงที่สุด เท่ากับ 40 คะแนน มีคะแนนเฉลี่ยต่ำที่สุด เท่ากับ 14.69 ซึ่งมากกว่าค่าเฉลี่ยของนักเรียนทั้งหมด และมีส่วนเบี่ยงเบนมาตรฐาน เท่ากับ 9.27 โค้งการแจกแจงมีลักษณะเบ้ขวา และมีความโด่งมาก โดยมีความเบ้ เท่ากับ 1.03 ค่าความคลาดเคลื่อนมาตรฐานของความเบ้ เท่ากับ .07 และค่าความโด่ง เท่ากับ .02 ค่าความคลาดเคลื่อนมาตรฐานของความโด่ง เท่ากับ .14 ส่วนเพศหญิง มีจำนวน 1,200 คน มีคะแนนต่ำที่สุด เท่ากับ 1 คะแนน และคะแนนสูงที่สุด เท่ากับ 39 คะแนน ซึ่งน้อยกว่าเพศชาย มีคะแนนเฉลี่ย 14.26 คะแนน ซึ่งน้อยกว่าค่าเฉลี่ยของคะแนนการสอบจากนักเรียนทั้งหมด และมีส่วนเบี่ยงเบนมาตรฐาน เท่ากับ 8.30 โค้งการแจกแจงมีลักษณะเบ้ขวา และมีความโด่งมาก โดยมีความเบ้ เท่ากับ 1.11 ค่าความคลาดเคลื่อนมาตรฐานของความเบ้ เท่ากับ .07 และค่าความโด่ง เท่ากับ .46 ค่าความคลาดเคลื่อนมาตรฐานของความโด่ง เท่ากับ .14 ซึ่งเมื่อพิจารณาค่าความเบ้และค่าความโด่ง จากลักษณะการแจกแจงข้อมูลดังกล่าว แสดงให้เห็นว่านักเรียนส่วนใหญ่ได้คะแนนจากการสอบวิชาคณิตศาสตร์ในระดับน้อย (เบ้ขวา)

ตารางที่ 4-2 แสดงจำนวนผู้สอบ ค่าต่ำสุด ค่าสูงสุด ค่าเฉลี่ย ค่าส่วนเบี่ยงเบนมาตรฐาน ค่าความเบ้ และค่าความโด่ง จากคะแนนการตอบแบบทดสอบวิชาภาษาไทย จำแนกตามเพศผู้สอบ

	n	***Min***	***Max***	***Mean***	***SD***	***SK***	***KU***
เพศชาย	1,200	11.00	76.00	43.20	13.94	-.05 (.07)	-.73 (.14)
เพศหญิง	1,200	8.00	73.00	42.66	13.14	-.19 (.07)	-.54 (.14)
รวม	2,400	8.00	76.00	42.93	13.54	-.11 (.05)	-.63 (.10)

หมายเหตุ ค่าความเบ้/ ค่าความคลาดเคลื่อนมาตรฐานของความเบ้ และค่าความโด่ง/ ค่าความคลาดเคลื่อนมาตรฐานของความโด่งมีค่าอยู่ระหว่าง -1.96 ถึง 1.96 ลักษณะของการแจกแจงเป็นโค้งปกติ

จากตารางที่ 4-2 พบว่า ผู้สอบจำนวนทั้งหมด 2,400 คน มีคะแนนต่ำที่สุด เท่ากับ 8 คะแนน และสูงที่สุด เท่ากับ 76 คะแนน มีค่าเฉลี่ย เท่ากับ 42.93 คะแนน และมีส่วนเบี่ยงเบนมาตรฐาน เท่ากับ 13.54 โค้งการแจกแจงมีลักษณะเบ้ซ้าย และมีความโด่งน้อย โดยมีความเบ้ เท่ากับ -.11 ค่าความคลาดเคลื่อนมาตรฐานของความเบ้ เท่ากับ .05 และค่าความโด่ง เท่ากับ -.63 ค่าความคลาดเคลื่อนมาตรฐานของความโด่ง เท่ากับ .10 เมื่อพิจารณาแล้ว พบว่า เพศชาย จำนวน 1,200 คน มีคะแนนต่ำที่สุด เท่ากับ 11 คะแนน และสูงที่สุด เท่ากับ 76 คะแนน มีคะแนนเฉลี่ยต่ำที่สุด เท่ากับ 43.20 ซึ่งมากกว่าค่าเฉลี่ยของนักเรียนทั้งหมด และมีส่วนเบี่ยงเบนมาตรฐาน เท่ากับ 13.94 โค้งการแจกแจงมีลักษณะเบ้ขวาและมีความโด่งน้อย โดยมีความเบ้ เท่ากับ -.05 ค่าความคลาดเคลื่อนมาตรฐานของความเบ้ เท่ากับ .07 และค่าความโด่ง เท่ากับ -.73 ค่าความคลาดเคลื่อนมาตรฐานของความโด่ง เท่ากับ .14 ส่วนเพศหญิงมีจำนวน 1,200 คน มีคะแนนต่ำที่สุด เท่ากับ 8 คะแนน และคะแนนสูงที่สุด เท่ากับ 73 คะแนน ซึ่งน้อยกว่าเพศชาย มีคะแนนเฉลี่ย 42.66 คะแนน ซึ่งน้อยกว่าค่าเฉลี่ยของคะแนนการสอบจากนักเรียนทั้งหมด และมีส่วนเบี่ยงเบนมาตรฐาน เท่ากับ 13.14 โค้งการแจกแจงมีลักษณะเบ้ซ้าย และมีความโด่งน้อย โดยมีความเบ้ เท่ากับ -.19 ค่าความคลาดเคลื่อนมาตรฐานของความเบ้ เท่ากับ .07 และค่าความโด่ง เท่ากับ -.54 ค่าความคลาดเคลื่อนมาตรฐานของความโด่ง เท่ากับ .14 ซึ่งจากลักษณะการแจกแจงข้อมูลดังกล่าว แสดงให้เห็นว่านักเรียนส่วนใหญ่ได้คะแนนจากการสอบวิชาคณิตศาสตร์ในระดับมาก (เบ้ซ้าย)

ตารางที่ 4-3 แสดงจำนวนผู้สอบ ค่าต่ำสุด ค่าสูงสุด ค่าเฉลี่ย ค่าส่วนเบี่ยงเบนมาตรฐาน ค่าความเบ้ และค่าความโด่ง จากคะแนนการตอบแบบทดสอบวิชาคณิตศาสตร์ จำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียนผู้สอบ

	n	***Min***	***Max***	***Mean***	***SD***	***SK***	***KU***
กรุงเทพฯ และปริมณฑล	1,200	2.00	40.00	16.69	9.76	.73 (.07)	-.62 (.14)
นอกเขตกรุงเทพฯ และปริมณฑล	1,200	1.00	39.00	12.26	7.05	1.37 (.07)	1.54 (.14)
รวม	2,400	1.00	40.00	14.48	8.80	1.07 (.05)	.24 (.10)

หมายเหตุ ค่าความเบ้/ ค่าความคลาดเคลื่อนมาตรฐานของความเบ้ และค่าความโด่ง/ ค่าความคลาดเคลื่อนมาตรฐานของความโด่งมีค่าอยู่ระหว่าง -1.96 ถึง 1.96 ลักษณะของการแจกแจงเป็นโค้งปกติ

จากตารางที่ 4-3 พบว่า ผู้สอบทั้งหมด 2,400 คน มีคะแนนต่ำที่สุด เท่ากับ 1 คะแนน และสูงที่สุด เท่ากับ 40 คะแนน มีค่าเฉลี่ย เท่ากับ 14.48 คะแนน และมีส่วนเบี่ยงเบนมาตรฐาน เท่ากับ 8.80 โค้งการแจกแจงมีลักษณะเบ้ขวา และมีความโด่งน้อย โดยมีความเบ้ เท่ากับ 1.07 ค่าความคลาดเคลื่อนมาตรฐานของความเบ้ เท่ากับ .05 และค่าความโด่ง เท่ากับ .24 ค่าความคลาดเคลื่อนมาตรฐานของความโด่ง เท่ากับ .10 เมื่อพิจารณาตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียนผู้สอบแล้วพบว่าเขตกรุงเทพมหานครและปริมณฑล จำนวน 1,200 คน มีคะแนนต่ำที่สุด เท่ากับ 2 คะแนน และสูงที่สุด เท่ากับ 40 คะแนน มีคะแนนเฉลี่ยต่ำที่สุด เท่ากับ 16.69 ซึ่งมากกว่าค่าเฉลี่ยของนักเรียนทั้งหมด และมีส่วนเบี่ยงเบนมาตรฐาน เท่ากับ 9.76 โค้งการแจกแจงมีลักษณะเบ้ขวา และมีความโด่งน้อย โดยมีความเบ้ เท่ากับ .73 ค่าความคลาดเคลื่อนมาตรฐานของความเบ้ เท่ากับ .07 และค่าความโด่ง เท่ากับ -.62 ค่าความคลาดเคลื่อนมาตรฐานของความโด่ง เท่ากับ .14 เมื่อพิจารณาค่าความเบ้และค่าความโด่งแล้วยังเป็นโค้งปกติ ส่วนนอกเขตกรุงเทพมหานครและปริมณฑลจำนวน 1,200 คน มีคะแนนต่ำที่สุด เท่ากับ 1 คะแนน และคะแนนสูงที่สุด เท่ากับ 39 คะแนน ซึ่งน้อยกว่าเขตกรุงเทพมหานครและปริมณฑล มีคะแนนเฉลี่ยน้อยกว่าเขตกรุงเทพมหานครและปริมณฑล เท่ากับ 12.26 คะแนน และน้อยกว่าค่าเฉลี่ยของคะแนนการสอบจากนักเรียนทั้งหมด และมีส่วนเบี่ยงเบนมาตรฐาน เท่ากับ 7.05 โค้งการแจกแจงมีลักษณะเบ้ขวา และมีความโด่งมาก โดยมีความเบ้ เท่ากับ 1.37 ค่าความคลาดเคลื่อนมาตรฐานของความเบ้ เท่ากับ

.07 และค่าความโด่ง เท่ากับ 1.54 ค่าความคลาดเคลื่อนมาตรฐานของความโด่ง เท่ากับ .14 จากลักษณะการแจกแจงข้อมูลดังกล่าว แสดงให้เห็นว่านักเรียนส่วนใหญ่ได้คะแนนจากการสอบวิชาคณิตศาสตร์ในระดับน้อย (เบ้ขวา)

ตารางที่ 4-4 แสดงจำนวนผู้สอบ ค่าต่ำสุด ค่าสูงสุด ค่าเฉลี่ย ค่าส่วนเบี่ยงเบนมาตรฐาน ค่าความเบ้ และค่าความโด่ง จากคะแนนการตอบแบบทดสอบวิชาภาษาไทย จำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียนผู้สอบ

	n	***Min***	***Max***	***Mean***	***SD***	***SK***	***KU***
กรุงเทพฯ และปริมณฑล	1,200	8.00	74.00	45.46	14.21	-.29 (.07)	-.62 (.14)
นอกเขตกรุงเทพฯ และปริมณฑล	1,200	10.00	76.00	40.39	12.34	-.05 (.07)	-.56 (.14)
รวม	2,400	8.00	76.00	42.93	13.54	-.11 (.05)	-.63 (.10)

หมายเหตุ ค่าความเบ้/ ค่าความคลาดเคลื่อนมาตรฐานของความเบ้ และค่าความโด่ง/ ค่าความคลาดเคลื่อนมาตรฐานของความโด่งมีค่าอยู่ระหว่าง -1.96 ถึง 1.96 ลักษณะของการแจกแจงเป็นโค้งปกติ

จากตารางที่ 4-4 พบว่า ผู้สอบทั้งหมด 2,400 คน มีคะแนนต่ำที่สุด เท่ากับ 8 คะแนน และสูงที่สุด เท่ากับ 76 คะแนน มีค่าเฉลี่ย เท่ากับ 42.93 คะแนน และมีส่วนเบี่ยงเบนมาตรฐาน เท่ากับ 13.54 โค้งการแจกแจงมีลักษณะเบ้ซ้าย และมีความโด่งน้อย โดยมีความเบ้ เท่ากับ -.11 ค่าความคลาดเคลื่อนมาตรฐานของความเบ้ เท่ากับ .05 และค่าความโด่ง เท่ากับ -.63 ค่าความคลาดเคลื่อนมาตรฐานของความโด่ง เท่ากับ .10 เมื่อพิจารณาตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียนผู้สอบแล้ว พบว่า เขตกรุงเทพมหานครและปริมณฑล จำนวน 1, 200 คน มีคะแนนต่ำที่สุด เท่ากับ 8 คะแนน และสูงที่สุด เท่ากับ 74 คะแนน มีคะแนนเฉลี่ยต่ำที่สุด เท่ากับ 45.46 ซึ่งมากกว่าค่าเฉลี่ยของนักเรียนทั้งหมด และมีส่วนเบี่ยงเบนมาตรฐาน เท่ากับ 14.21 โค้งการแจกแจงมีลักษณะเบ้ซ้าย และมีความโด่งน้อย โดยมีความเบ้ เท่ากับ -.29 ค่าความคลาดเคลื่อนมาตรฐานของความเบ้ เท่ากับ .07 และค่าความโด่ง เท่ากับ -.62 ค่าความคลาดเคลื่อนมาตรฐานของความโด่ง เท่ากับ .14 ส่วนนอกเขตกรุงเทพมหานครและปริมณฑลจำนวน 1, 200 คน มีคะแนนต่ำที่สุด เท่ากับ 10 คะแนน และคะแนนสูงที่สุด เท่ากับ 76 คะแนนซึ่งมากกว่าเขตกรุงเทพมหานครและปริมณฑล มีคะแนนเฉลี่ย

น้อยกว่าเขตกรุงเทพมหานครและปริมณฑล เท่ากับ 40.39 คะแนน และน้อยกว่าค่าเฉลี่ยของคะแนนการสอบจากนักเรียนทั้งหมด และมีส่วนเบี่ยงเบนมาตรฐาน เท่ากับ 12.34 โค้งการแจกแจงมีลักษณะเบ้ซ้าย และมีความโด่งน้อย โดยมีความเบ้ เท่ากับ -.56 ค่าความคลาดเคลื่อนมาตรฐานของความเบ้ เท่ากับ .07 และค่าความโด่ง เท่ากับ -.56 ค่าความคลาดเคลื่อนมาตรฐานของความโด่ง เท่ากับ .14 จากลักษณะการแจกแจงข้อมูลดังกล่าว แสดงให้เห็นว่านักเรียนส่วนใหญ่ได้คะแนนจากการสอบวิชาภาษาไทยในระดับมาก (เบ้ซ้าย)

ตอนที่ 2 เปรียบเทียบผลการประมาณค่าพารามิเตอร์อำนาจจำแนกของข้อสอบ

การวิเคราะห์ค่าพารามิเตอร์อำนาจจำแนกของข้อสอบครั้งนี้ ประกอบด้วย วิชาคณิตศาสตร์ และภาษาไทย ซึ่งผู้วิจัยได้ทำการวิเคราะห์ด้วยวิธี Maximum likelihood: ML, วิธี BAYESIAN และวิธี Confirmatory factor analysis: CFA โดยการใช้โปรแกรม IRTPRO version 2.1, Open BUGS version 3.23 และ Mplus version 7.4 ตามลำดับ แล้วนำผลการประมาณค่าพารามิเตอร์อำนาจจำแนกที่ได้จากทั้ง 3 วิธี มาทำการเปรียบเทียบกัน ซึ่งได้ผลการวิเคราะห์ดังนี้

ตารางที่ 4-5 ผลการประมาณค่าพารามิเตอร์อำนาจจำแนกของข้อสอบวิชาคณิตศาสตร์ ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA

ข้อสอบ	ผลการประมาณค่าพารามิเตอร์อำนาจจำแนกของข้อสอบ (a)					
	ML		BAYESIAN		CFA	
	a	*SE*	*a*	*SE*	*a*	*SE*
1	1.07	.06	1.06	.06	1.07	.06
2	1.18	.07	1.17	.06	1.18	.07
3	1.07	.06	1.06	.06	1.07	.06
4	1.12	.06	1.11	.06	1.12	.06
5	1.37	.07	1.36	.07	1.37	.07
6	1.40	.07	1.38	.07	1.40	.07
7	.92	.06	.91	.06	.92	.06
8	1.23	.07	1.22	.07	1.23	.07
9	.86	.06	.85	.06	.86	.06
10	.95	.06	.94	.06	.95	.06

ตารางที่ 4-5 (ต่อ)

ข้อสอบ	ผลการประมาณค่าพารามิเตอร์อำนาจจำแนกของข้อสอบ (a)					
	ML		BAYESIAN		CFA	
	a	*SE*	*a*	*SE*	*a*	*SE*
11	.67	.05	.66	.05	.67	.05
12	1.43	.08	1.42	.08	1.43	.08
13	1.25	.07	1.24	.07	1.25	.07
14	.71	.05	.70	.05	.71	.05
15	.84	.06	.84	.05	.85	.06
16	1.19	.07	1.18	.06	1.19	.07
17	1.03	.06	1.02	.06	1.03	.06
18	.97	.06	.96	.06	.97	.06
19	.96	.06	.95	.06	.96	.06
20	1.38	.09	1.37	.08	1.38	.08
21	1.41	.09	1.40	.09	1.41	.09
22	.75	.06	.74	.05	.75	.06
23	.95	.06	.95	.06	.95	.06
24	.51	.05	.50	.05	.51	.05
25	.95	.06	.94	.06	.95	.06
26	1.60	.11	1.59	.10	1.60	.10
27	1.02	.06	1.00	.06	1.02	.06
28	.99	.06	.99	.06	.99	.06
29	.87	.06	.86	.06	.87	.06
30	-0.25	.06	.13	.02	-.25	.06
31	1.49	.09	1.48	.08	1.49	.08
32	1.48	.09	1.47	.08	1.48	.08
33	1.90	.10	1.89	.09	1.90	.09
34	1.85	.10	1.84	.09	1.85	.09
35	2.24	.12	2.23	.11	2.24	.11
36	2.44	.14	2.43	.12	2.44	.12

ตารางที่ 4-5 (ต่อ)

ข้อสอบ	ผลการประมาณค่าพารามิเตอร์อำนาจจำแนกของข้อสอบ (a)					
	ML		BAYESIAN		CFA	
	a	*SE*	*a*	*SE*	*a*	*SE*
37	2.51	.15	2.51	.13	2.51	.13
38	2.30	.13	2.28	.11	2.30	.11
39	1.39	.08	1.38	.08	1.39	.08
40	1.48	.10	1.47	.09	1.48	.09
เฉลี่ย	1.24	.08	1.24	.07	1.24	.07

จากตารางที่ 4-5 พบว่า ค่าพารามิเตอร์อำนาจจำแนกของข้อสอบวิชาคณิตศาสตร์ที่ประมาณค่าด้วยวิธี ML มีค่าอำนาจจำแนกต่ำที่สุด เท่ากับ -.25 และสูงที่สุด เท่ากับ 2.51 ค่าเฉลี่ย เท่ากับ 1.24 ค่าความคลาดเคลื่อนมาตรฐาน เท่ากับ .08 ค่าพารามิเตอร์อำนาจจำแนกของข้อสอบที่ประมาณค่าด้วยวิธี BAYESIAN มีค่าอำนาจจำแนกต่ำที่สุด เท่ากับ .13 และสูงที่สุด เท่ากับ 2.51 ค่าเฉลี่ย เท่ากับ 1.24 ค่าความคลาดเคลื่อนมาตรฐาน เท่ากับ .07 และค่าพารามิเตอร์อำนาจจำแนกของข้อสอบที่ประมาณค่าด้วยวิธี CFA มีค่าอำนาจจำแนกต่ำที่สุด เท่ากับ -.25 และสูงที่สุด เท่ากับ 2.51 ค่าเฉลี่ย เท่ากับ 1.24 ค่าความคลาดเคลื่อนมาตรฐาน เท่ากับ .07

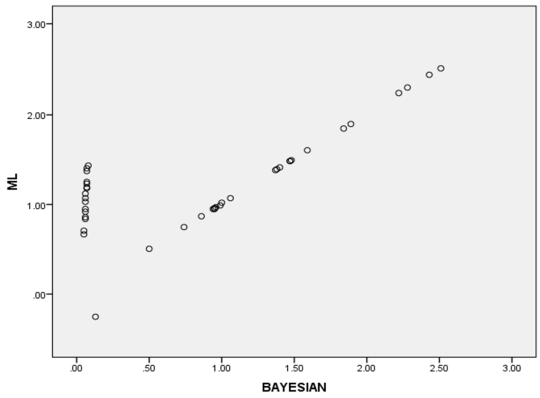
เมื่อนำผลการประมาณค่าพารามิเตอร์อำนาจจำแนกของข้อสอบวิชาคณิตศาสตร์ที่ได้จากการวิเคราะห์ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA มาหาความสัมพันธ์ระหว่างกันด้วยการหาค่าสหสัมพันธ์เพียร์สัน ได้ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน (Pearson's correlation coefficient) ดังตารางที่ 4-6

ตารางที่ 4-6 ผลการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าพารามิเตอร์อำนาจจำแนกของข้อสอบวิชาคณิตศาสตร์ ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA

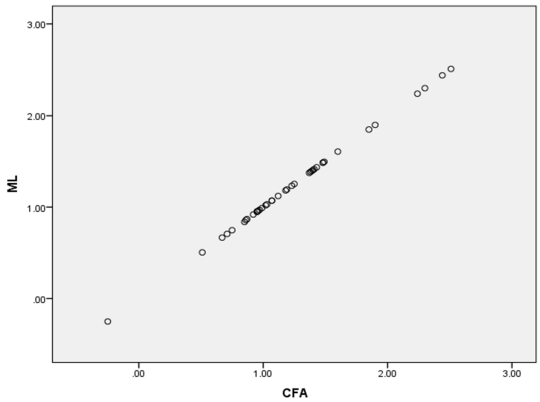
ตัวแปร	ML	BAYESIAN	CFA
ML	1.00		
BAYESIAN	1.00*	1.00	
CFA	1.00*	1.00*	1.00

* $p < .05$

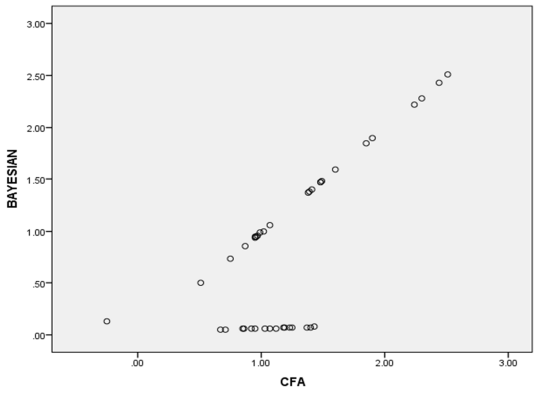
จากตารางที่ 4-6 พบว่า ผลการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าพารามิเตอร์อำนาจจำแนกของข้อสอบวิชาคณิตศาสตร์ทั้ง 3 วิธี ซึ่งได้แก่ วิธี ML วิธี BAYESIAN และวิธี CFA มีความสัมพันธ์ทางบวกในระดับสูงมาก และมีนัยสำคัญทางสถิติที่ .05 ซึ่งสามารถแสดงแผนภาพกราฟความสัมพันธ์ระหว่างค่าพารามิเตอร์อำนาจจำแนกของข้อสอบวิชาคณิตศาสตร์ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA แสดงได้ดังภาพต่อไปนี้



ภาพที่ 4-1 กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ค่าอำนาจจำแนกของข้อสอบวิชาคณิตศาสตร์ที่ประมาณค่าด้วยวิธี ML และวิธี BAYESIAN



ภาพที่ 4-2 กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ค่าอำนาจจำแนกของข้อสอบวิชาคณิตศาสตร์ที่ประมาณค่าด้วยวิธี ML และวิธี CFA



ภาพที่ 4-3 กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ค่าอำนาจจำแนกของข้อสอบวิชาคณิตศาสตร์ที่ประมาณค่าด้วยวิธี BAYESIAN และวิธี CFA

ตารางที่ 4-7 ผลการประมาณค่าพารามิเตอร์อำนาจจำแนกของข้อสอบวิชาภาษาไทย ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA

ข้อสอบ	ผลการประมาณค่าพารามิเตอร์อำนาจจำแนกของข้อสอบ (a)					
	ML		BAYESIAN		CFA	
	a	***SE***	***a***	***SE***	***a***	***SE***
1	.77	.06	.75	.06	.77	.06
2	1.15	.06	1.13	.06	1.15	.07
3	1.07	.07	1.06	.07	1.07	.07
4	.49	.05	.48	.05	.49	.05
5	1.15	.07	1.14	.07	1.15	.07
6	.42	.05	.41	.05	.42	.05
7	1.04	.06	1.03	.06	1.05	.06
8	1.06	.07	1.04	.07	1.06	.07
9	.20	.05	.19	.05	.20	.05
10	1.41	.08	1.40	.08	1.41	.08
11	.90	.06	.88	.06	.90	.06
12	.66	.06	.65	.06	.66	.06
13	.14	.05	.14	.04	.14	.05

ตารางที่ 4-7 (ต่อ)

ข้อสอบ	ผลการประมาณค่าพารามิเตอร์อำนาจจำแนกของข้อสอบ (a)					
	ML		BAYESIAN		CFA	
	a	*SE*	*a*	*SE*	*a*	*SE*
14	.62	.06	.60	.05	.62	.06
15	.54	.06	.54	.05	.54	.06
16	.71	.05	.69	.05	.71	.06
17	1.08	.06	1.07	.06	1.08	.06
18	1.06	.07	1.04	.07	1.06	.07
19	.64	.05	.63	.05	.64	.05
20	.51	.05	.50	.05	.51	.05
21	1.11	.07	1.10	.07	1.11	.07
22	1.01	.07	.99	.07	1.01	.07
23	.53	.05	.52	.05	.53	.05
24	.90	.06	.88	.07	.90	.06
25	.69	.06	.67	.06	.69	.06
26	.96	.06	.94	.06	.96	.06
27	1.04	.06	1.03	.06	1.04	.06
28	1.10	.07	1.09	.07	1.10	.07
29	.56	.05	.55	.05	.56	.05
30	.65	.05	.64	.05	.65	.05
31	.56	.05	.55	.05	.56	.05
32	.59	.05	.57	.05	.59	.05
33	.41	.05	.40	.05	.41	.05
34	.79	.05	.77	.05	.79	.05
35	.41	.05	.40	.05	.41	.05
36	.65	.05	.64	.05	.65	.05
37	1.16	.07	1.15	.07	1.17	.07
38	.28	.04	.27	.04	.29	.04

ตารางที่ 4-7 (ต่อ)

ข้อสอบ	ผลการประมาณค่าพารามิเตอร์อำนาจจำแนกของข้อสอบ (a)					
	ML		BAYESIAN		CFA	
	a	*SE*	*a*	*SE*	*a*	*SE*
39	.86	.05	.85	.05	.86	.06
40	1.00	.07	.98	.07	1.00	.07
41	.86	.06	.85	.06	.86	.06
42	.85	.06	.83	.05	.85	.06
43	.84	.06	.83	.06	.84	.06
44	.68	.05	.67	.05	.68	.05
45	.61	.05	.59	.05	.61	.05
46	1.15	.06	1.13	.06	1.15	.06
47	1.15	.07	1.13	.07	1.15	.07
48	.64	.05	.62	.05	.64	.05
49	.89	.06	.88	.06	.89	.06
50	.71	.05	.69	.05	.71	.05
51	.96	.06	.95	.06	.96	.06
52	.89	.06	.88	.06	.89	.06
53	.86	.06	.85	.06	.86	.06
54	.28	.04	.27	.04	.28	.04
55	.67	.05	.66	.54	.68	.05
56	.73	.05	.71	.05	.73	.05
57	.92	.06	.91	.06	.92	.06
58	1.08	.06	1.07	.06	1.08	.06
59	.91	.06	.90	.60	.91	.06
60	1.09	.06	1.08	.06	1.10	.06
61	1.43	.08	1.41	.08	1.43	.08
62	1.46	.08	1.45	.07	1.47	.08
63	.75	.05	.74	.05	.75	.05
64	.74	.06	.73	.06	.74	.06

ตารางที่ 4-7 (ต่อ)

ข้อสอบ	ผลการประมาณค่าพารามิเตอร์อำนาจจำแนกของข้อสอบ (a)					
	ML		BAYESIAN		CFA	
	a	*SE*	*a*	*SE*	*a*	*SE*
65	.74	.06	.73	.06	.74	.06
66	1.06	.07	1.04	.06	1.06	.07
67	1.06	.06	1.05	.06	.06	.06
68	1.58	.08	1.56	.08	1.58	.08
69	1.27	.07	1.25	.07	1.27	.07
70	1.51	.08	1.49	.08	1.51	.08
71	.95	.06	.93	.06	.95	.06
72	.87	.06	.85	.05	.87	.06
73	.43	.05	.42	.05	.44	.05
74	1.19	.07	1.17	.07	1.19	.07
75	.74	.05	.72	.05	.74	.05
76	2.04	.11	2.01	.10	2.04	.10
77	.80	.06	.79	.05	.81	.06
78	1.10	.06	1.09	.06	1.10	.06
79	1.18	.07	1.16	.06	1.18	.06
80	.64	.05	.63	.05	.64	.05
เฉลี่ย	.86	.06	.85	.07	.87	.06

จากตารางที่ 4-7 พบว่า ค่าพารามิเตอร์อำนาจจำแนกของข้อสอบวิชาภาษาไทยที่ประมาณค่าด้วยวิธี ML มีค่าอำนาจจำแนกต่ำที่สุด เท่ากับ .14 และสูงที่สุด เท่ากับ 2.04 ค่าเฉลี่ย เท่ากับ 0.86 ค่าความคลาดเคลื่อนมาตรฐาน เท่ากับ .06 ค่าพารามิเตอร์อำนาจจำแนกของข้อสอบที่ประมาณค่าด้วยวิธี BAYESIAN มีค่าอำนาจจำแนกต่ำที่สุด เท่ากับ .14 และสูงที่สุด เท่ากับ 2.01 ค่าเฉลี่ย เท่ากับ .85 ค่าความคลาดเคลื่อนมาตรฐาน เท่ากับ .07 และค่าพารามิเตอร์อำนาจจำแนกของข้อสอบที่ประมาณค่าด้วยวิธี CFA มีค่าอำนาจต่ำที่สุด เท่ากับ .14 และสูงที่สุด เท่ากับ 2.04 ค่าเฉลี่ย เท่ากับ .87 ค่าความคลาดเคลื่อนมาตรฐาน เท่ากับ .06

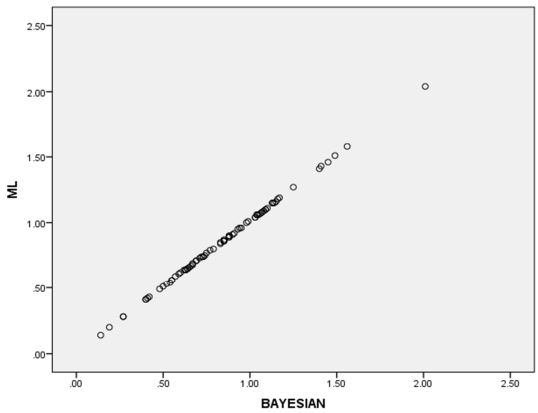
เมื่อนำผลการประมาณค่าพารามิเตอร์อำนาจจำแนกของข้อสอบวิชาภาษาไทยที่ได้จากการวิเคราะห์ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA มาหาความสัมพันธ์ระหว่างกัน ด้วยการหาค่าสหสัมพันธ์เพียร์สัน ได้ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน (Pearson's correlation coefficient) ดังตารางที่ 4-8

ตารางที่ 4-8 ผลการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าพารามิเตอร์อำนาจจำแนกของข้อสอบวิชาภาษาไทย ด้วยวิธี ML วิธี BAYESIAN และ วิธี CFA

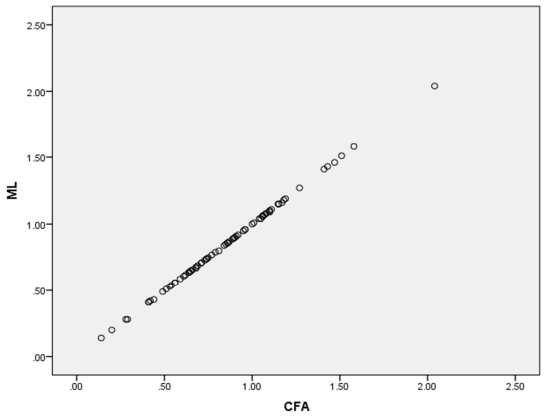
ตัวแปร	ML	BAYESIAN	CFA
ML	1.00		
BAYESIAN	1.00*	1.00	
CFA	1.00*	1.00*	1.00

* $p < .05$

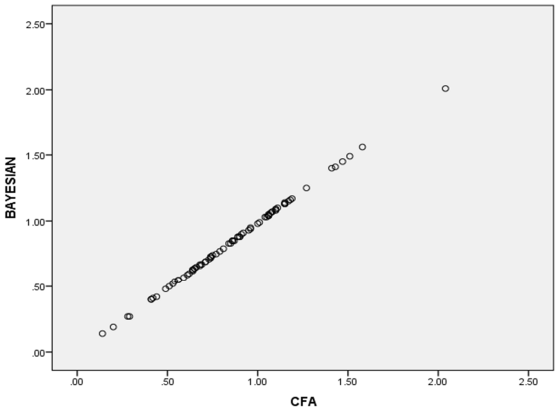
จากตารางที่ 4-8 พบว่า ผลการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าพารามิเตอร์อำนาจจำแนกของข้อสอบวิชาภาษาไทยทั้ง 3 วิธี ซึ่งได้แก่ ML วิธี BAYESIAN และวิธี CFA มีความสัมพันธ์ทางบวกในระดับสูงมาก และมีนัยสำคัญทางสถิติที่ระดับ .05 ซึ่งสามารถแสดงแผนภาพกราฟความสัมพันธ์ระหว่างค่าพารามิเตอร์ค่าอำนาจจำแนกของข้อสอบวิชาภาษาไทย ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA แสดงดังภาพต่อไปนี้



ภาพที่ 4-4 กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ค่าอำนาจจำแนกข้อสอบวิชาภาษาไทยที่ประมาณค่าด้วยวิธี ML และวิธี BAYESIAN



ภาพที่ 4-5 กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ค่าอำนาจจำแนกข้อสอบวิชาภาษาไทย ที่ประมาณค่าด้วยวิธี ML และวิธี CFA



ภาพที่ 4-6 กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ค่าอำนาจจำแนกข้อสอบวิชาภาษาไทย ที่ประมาณค่าด้วยวิธี BAYESIAN และวิธี CFA

ตอนที่ 3 เปรียบเทียบผลการประมาณค่าพารามิเตอร์ความยากของข้อสอบ

การวิเคราะห์ค่าพารามิเตอร์ความยากของข้อสอบครั้งนี้ประกอบด้วยวิชาคณิตศาสตร์และภาษาไทย ซึ่งผู้วิจัยได้ทำการวิเคราะห์ด้วยวิธี ML แล้วนำผลการวิเคราะห์มาเปรียบเทียบกับการประมาณค่าด้วยวิธี BAYSIAN และวิธี CFA โดยการใช้โปรแกรม IRTPRO version 2.1, Open BUGS version 3.23 และ Mplus version 7.4 ตามลำดับ ได้ผลการวิเคราะห์ ดังนี้

ตารางที่ 4-9 ผลการประมาณค่าพารามิเตอร์ความยากของข้อสอบวิชาคณิตศาสตร์ ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA

ข้อสอบ	ผลการประมาณค่าพารามิเตอร์ความยากของข้อสอบ (b)					
	ML		BAYESIAN		CFA	
	b	***SE***	***b***	***SE***	***b***	***SE***
1	.33	.05	.33	.05	.32	.05
2	.55	.05	.56	.05	.55	.05
3	.76	.06	.77	.06	.76	.06
4	.88	.06	.90	.06	.88	.06
5	.56	.05	.56	.05	.56	.05
6	.58	.05	.59	.05	.58	.05
7	.96	.07	.97	.08	.96	.08
8	.50	.05	.51	.05	.50	.05
9	1.43	.10	1.45	.10	1.43	.10
10	1.31	.08	1.32	.09	1.31	.09
11	.61	.08	.61	.09	.61	.08
12	.31	.05	.31	.04	.31	.04
13	.62	.05	.63	.05	.61	.05
14	1.23	.10	1.26	.11	1.23	.10
15	.80	.07	.82	.08	.80	.07
16	.53	.05	.54	.05	.53	.05
17	1.07	.07	1.08	.07	1.07	.07
18	.56	.06	.57	.06	.56	.06
19	1.18	.08	1.20	.08	1.18	.08
20	-.43	.05	-.43	.04	-.43	.04
21	-.71	.06	-.72	.05	-.71	.05
22	-.15	.07	-.15	.06	-.15	.06
23	.77	.07	.77	.07	.77	.07

ตารางที่ 4-9 (ต่อ)

ข้อสอบ	ผลการประมาณค่าพารามิเตอร์ความยากของข้อสอบ (b)					
	ML		BAYESIAN		CFA	
	b	***SE***	***b***	***SE***	***b***	***SE***
24	-.70	.11	-.72	.11	-.71	.11
25	1.08	.07	1.09	.08	1.08	.08
26	-.78	.06	-.78	.05	-.78	.05
27	.47	.06	.48	.06	.47	.06
28	.72	.06	.73	.06	.72	.06
29	.88	.07	.89	.08	.88	.08
30	-5.35	1.24	1.03	.14	-5.35	1.23
31	-.13	.05	-.13	.04	-.13	.04
32	-.16	.05	-.16	.04	-.16	.04
33	.56	.04	.56	.04	.56	.04
34	.55	.04	.55	.04	.55	.04
35	.44	.04	.45	.04	.44	.04
36	.46	.04	.47	.04	.46	.04
37	.77	.04	.78	.04	.77	.04
38	.81	.04	.81	.04	.80	.04
39	1.55	.07	1.57	.07	1.55	.07
40	1.90	.08	1.92	.09	1.90	.09
เฉลี่ย	.43	.09	.60	.06	.43	.09

จากตารางที่ 4-9 พบว่า ค่าพารามิเตอร์ความยากของข้อสอบวิชาคณิตศาสตร์ที่ประมาณค่าด้วยวิธี ML มีค่าความยากต่ำที่สุด เท่ากับ -5.35 และสูงที่สุด เท่ากับ 1.90 ค่าเฉลี่ย เท่ากับ .43 ค่าความคลาดเคลื่อนมาตรฐาน เท่ากับ .09 ค่าพารามิเตอร์ความยากของข้อสอบที่ประมาณค่าด้วยวิธี BAYESIAN มีค่าความยากต่ำที่สุด เท่ากับ -.78 และสูงที่สุด เท่ากับ 1.92 ค่าเฉลี่ย เท่ากับ .60 ค่าความคลาดเคลื่อนมาตรฐาน เท่ากับ .06 และค่าพารามิเตอร์ค่าความยากของข้อสอบที่ประมาณค่า

ด้วยวิธี CFA มีค่าความยากต่ำที่สุด เท่ากับ -5.35 และสูงที่สุด เท่ากับ 1.90 ค่าเฉลี่ย เท่ากับ .43 ค่าความคลาดเคลื่อนมาตรฐาน เท่ากับ .09

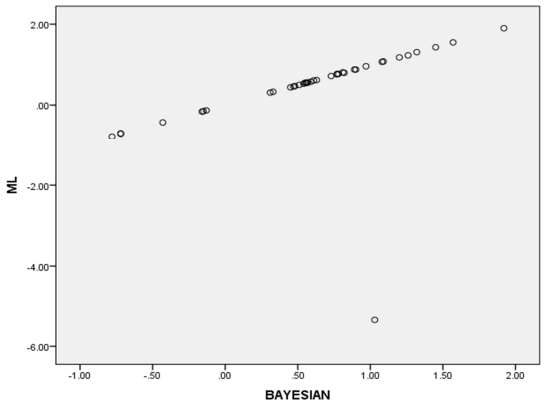
เมื่อนำผลการประมาณค่าพารามิเตอร์ความยากของข้อสอบวิชาคณิตศาสตร์ที่ได้จากการวิเคราะห์ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA มาหาความสัมพันธ์ระหว่างกัน ด้วยการหาค่าสหสัมพันธ์เพียร์สัน ได้ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน (Pearson's correlation coefficient) ดังตารางที่ 4-10

ตารางที่ 4-10 ผลการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าพารามิเตอร์ความยากของข้อสอบวิชาคณิตศาสตร์ ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA

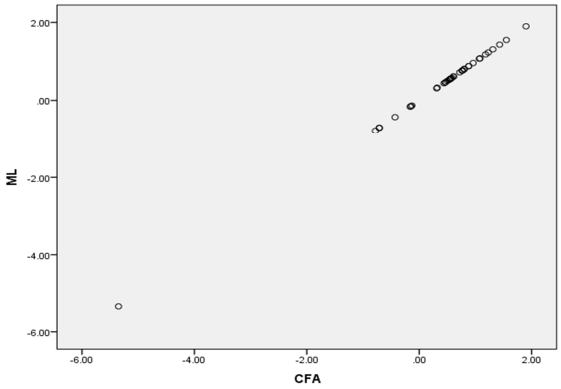
ตัวแปร	ML	BAYESIAN	CFA
ML	1.00		
BAYESIAN	.43*	1.00	
CFA	1.00*	.43*	1.00

* $p < .05$

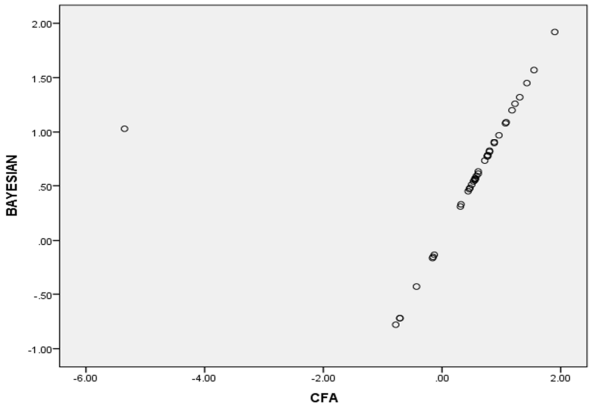
จากตารางที่ 4-10 พบว่า ผลการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าพารามิเตอร์ความยากของข้อสอบวิชาคณิตศาสตร์ ด้วยวิธี ML, BAYESIAN และวิธี CFA เมื่อพิจารณาผลการประมาณค่าด้วยวิธี ML เทียบกับผลการประมาณค่าด้วยวิธี CFA พบว่า วิธี ML กับวิธี CFA มีความสัมพันธ์ทางบวกในระดับสูงมาก และมีนัยสำคัญทางสถิติที่ระดับ .05 เมื่อพิจารณาผลการประมาณค่าด้วยวิธี ML เทียบกับผลการประมาณค่าด้วยวิธี BAYESIAN พบว่า วิธี ML มีความสัมพันธ์ทางบวกในระดับปานกลางค่อนข้างน้อย และมีนัยสำคัญทางสถิติที่ระดับ .05 และเมื่อพิจารณาผลการประมาณค่าด้วยวิธี BAYESIAN เทียบกับผลการประมาณค่าด้วยวิธี CFA พบว่า วิธี BAYESIAN มีความสัมพันธ์ทางบวกในระดับปานกลางค่อนข้างน้อย และมีนัยสำคัญทางสถิติที่ระดับ .05 เช่นกัน ซึ่งสามารถแสดงแผนภาพกราฟความสัมพันธ์ระหว่างค่าพารามิเตอร์ค่าอำนาจจำแนกของข้อสอบวิชาคณิตศาสตร์ ด้วยวิธี ML, วิธี BAYESIAN และวิธี CFA แสดงดังภาพต่อไปนี้



ภาพที่ 4-7 กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ความยากของข้อสอบวิชาคณิตศาสตร์ที่ประมาณค่าด้วยวิธี ML และวิธี BAYESIAN



ภาพที่ 4-8 กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ความยากของข้อสอบวิชาคณิตศาสตร์ที่ประมาณค่าด้วยวิธี ML และวิธี CFA



ภาพที่ 4-9 กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ความยากของข้อสอบวิชาคณิตศาสตร์ที่ประมาณค่าด้วยวิธี BAYESIAN และวิธี CFA

ตารางที่ 4-11 ผลการประมาณค่าพารามิเตอร์ความยากของข้อสอบวิชาภาษาไทย ด้วยวิธี ML, วิธี BAYESIAN และวิธี CFA

ข้อสอบ	ผลการประมาณค่าพารามิเตอร์ความยากของข้อสอบ (b)					
	ML		BAYESIAN		CFA	
	b	***SE***	***b***	***SE***	***b***	***SE***
1	1.67	.13	1.70	.13	1.67	.13
2	.77	.06	.79	.06	.77	.06
3	1.84	.11	1.87	.11	1.84	.11
4	.23	.09	.24	.09	.23	.09
5	1.26	.07	1.29	.07	1.26	.07
6	2.93	.37	3.06	.38	2.93	.37
7	.74	.06	.75	.06	.74	.06
8	1.95	.12	1.99	.12	1.95	.12
9	.50	.24	.55	.29	.51	.24
10	1.14	.06	1.16	.06	1.14	.06
11	-1.24	.08	-1.26	.09	-1.24	.09
12	-1.87	.16	-1.92	.16	-1.87	.16
13	-.70	.36	-.81	.47	-.70	.37
14	-1.83	.16	-1.89	.16	-1.83	.17
15	-2.24	.22	-2.28	.24	-2.24	.22
16	-.97	.09	-.99	.09	-.97	.09
17	-.51	.05	-.52	.05	-.51	.05
18	-1.29	.07	-1.31	.08	-1.29	.08
19	.83	.09	.85	.10	.84	.09
20	-1.53	.16	-1.57	.17	-1.53	.16
21	-.83	.05	-.85	.06	-.83	.06
22	-1.36	.08	-1.39	.09	-1.36	.09
23	-.08	.08	-.08	.09	-.08	.08
24	-1.24	.08	-1.27	.10	-1.24	.09

ตารางที่ 4-11 (ต่อ)

ข้อสอบ	ผลการประมาณค่าพารามิเตอร์ความยากของข้อสอบ (b)					
	ML		BAYESIAN		CFA	
	b	*SE*	*b*	*SE*	*b*	*SE*
25	-1.52	.12	-1.56	.13	-1.52	.13
26	-.70	.06	-.71	.06	-.69	.06
27	-.54	.05	-.55	.06	-.54	.05
28	-1.14	.07	-1.16	.07	-1.14	.07
29	1.10	.12	1.13	.13	1.10	.12
30	-.99	.09	-1.01	.10	-.99	.10
31	-1.58	.15	-1.64	.16	-1.58	.16
32	.05	.07	.05	.08	.05	.08
33	-1.68	.21	-1.72	.24	-1.67	.22
34	.15	.06	.16	.06	.16	.06
35	-.48	.11	-.50	.12	-.48	.12
36	-1.03	.10	-1.05	.10	-1.03	.10
37	-1.17	.06	-1.19	.17	-1.17	.07
38	-.28	.15	-.29	.16	-.27	.15
39	.16	.05	.17	.06	.16	.06
40	-1.38	.08	-1.41	.09	-1.38	.09
41	-.36	.06	-.37	.06	-.36	.06
42	-.21	.05	-.21	.06	-.21	.06
43	-1.16	.08	-1.18	.09	-1.16	.09
44	-1.18	.10	-1.20	.11	-1.18	.10
45	-1.39	.13	-1.44	.14	-1.39	.13
46	.22	.05	.23	.05	.22	.05
47	-1.10	.06	-1.12	.07	-1.10	.07
48	.54	.08	.56	.08	.55	.08
49	-.95	.07	-.96	.08	-.95	.07

ตารางที่ 4-11 (ต่อ)

ข้อสอบ	ผลการประมาณค่าพารามิเตอร์ความยากของข้อสอบ (b)					
	ML		BAYESIAN		CFA	
	b	*SE*	*b*	*SE*	*b*	*SE*
50	.13	.06	.14	.07	.14	.07
51	.52	.06	.53	.06	.52	.06
52	.84	.07	.86	.07	.85	.07
53	.75	.07	.76	.07	.75	.07
54	.21	.15	.22	.17	.21	.15
55	1.25	.11	1.28	.11	1.25	.11
56	.34	.07	.35	.07	.34	.07
57	-.01	.05	-.01	.05	-.01	.05
58	-.03	.05	-.02	.05	-.03	.05
59	-.15	.05	-.15	.05	-.14	.05
60	.14	.05	.15	.05	.15	.05
61	-.18	.04	-.17	.04	-.17	.04
62	-.49	.04	-.49	.04	-.49	.04
63	.63	.07	.64	.07	.63	.07
64	1.69	.13	1.74	.14	1.70	.13
65	2.27	.18	2.32	.17	2.27	.18
66	-.95	.06	-.96	.07	-.95	.06
67	.05	.05	.05	.05	.05	.05
68	-.76	.04	-.77	.04	-.76	.05
69	-.30	.04	-.30	.04	-.30	.04
70	-.82	.04	-.83	.05	-.81	.05
71	-.35	.05	-.36	.05	-.35	.05
72	-.41	.06	-.41	.06	-.41	.06
73	-.49	.11	-.51	.11	-.49	.11
74	-.76	.05	-.77	.06	-.76	.05

ตารางที่ 4-11 (ต่อ)

ข้อสอบ	ผลการประมาณค่าพารามิเตอร์ความยากของข้อสอบ (b)					
	ML		BAYESIAN		CFA	
	b	***SE***	***b***	***SE***	***b***	***SE***
75	.36	.06	.37	.07	.36	.07
76	-.69	.03	-.70	.04	-.69	.04
77	-.52	.06	-.53	.06	-.52	.07
78	-.20	.04	-.20	.05	-.20	.05
79	-.25	.04	-.26	.04	-.25	.05
80	-.07	.07	-.07	.07	-.07	.07
เฉลี่ย	-.21	.09	-.21	.10	-.21	.09

จากตารางที่ 4-11 พบว่า ค่าพารามิเตอร์ความยากของข้อสอบวิชาภาษาไทยที่ประมาณค่าด้วยวิธี ML มีค่าความยากต่ำที่สุด -2.24 และสูงที่สุด 2.93 ค่าเฉลี่ย เท่ากับ -.21 ค่าความคลาดเคลื่อนมาตรฐาน เท่ากับ .09 ค่าพารามิเตอร์ความยากของข้อสอบที่ประมาณค่าด้วยวิธี BAYESIAN มีค่าความยากต่ำที่สุด เท่ากับ -2.28 และสูงที่สุด เท่ากับ 3.06 ค่าเฉลี่ย เท่ากับ -.21 ค่าความคลาดเคลื่อนมาตรฐาน เท่ากับ .10 และค่าพารามิเตอร์ค่าความยากของข้อสอบที่ประมาณค่าด้วยวิธี CFA มีค่าความยากต่ำที่สุด เท่ากับ -2.24 และสูงที่สุด เท่ากับ 2.93 ค่าเฉลี่ย เท่ากับ -.21 ค่าความคลาดเคลื่อนมาตรฐาน เท่ากับ .09

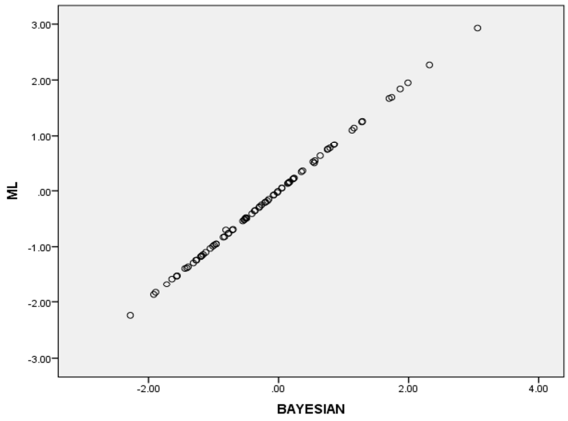
เมื่อนำผลการประมาณค่าพารามิเตอร์ความยากของข้อสอบวิชาภาษาไทยที่ได้จากการวิเคราะห์ด้วยวิธี ML, วิธี BAYESIAN และวิธี CFA มาหาความสัมพันธ์ระหว่างกัน ด้วยการหาค่าสหสัมพันธ์เพียร์สัน ได้ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน (Pearson’s correlation coefficient) ดังตารางที่ 4-12

ตารางที่ 4-12 ผลการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าพารามิเตอร์ความยากของข้อสอบวิชาภาษาไทย ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA

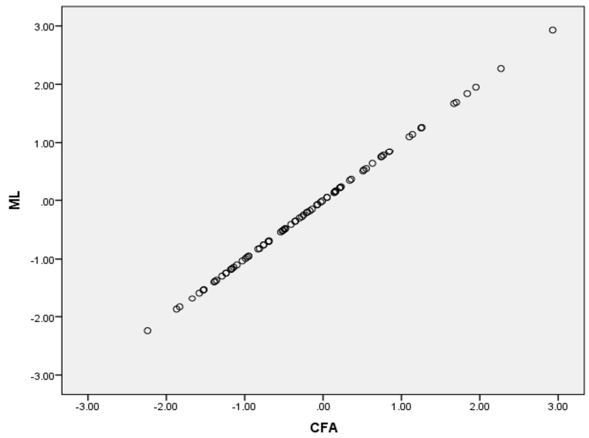
ตัวแปร	ML	BAYESIAN	CFA
ML	1.00		
BAYESIAN	1.00*	1.00	
CFA	1.00*	1.00*	1.00

* $p < .05$

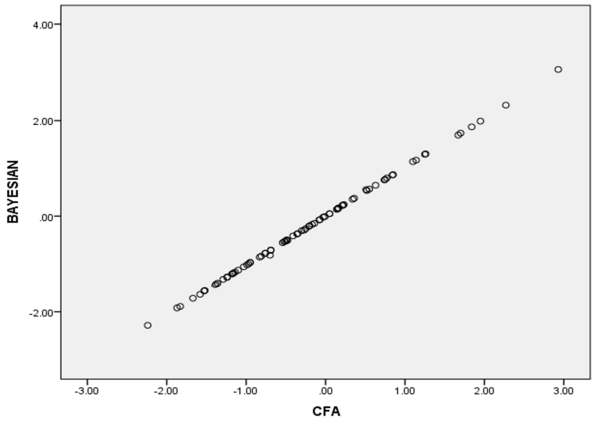
จากตารางที่ 4-12 พบว่า ผลการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าพารามิเตอร์ความยากของข้อสอบวิชาภาษาไทยทั้ง 3 วิธี ซึ่งได้แก่ ML, วิธี BAYESIAN และวิธี CFA พบว่า มีความสัมพันธ์ทางบวกในระดับสูงมาก และมีนัยสำคัญทางสถิติที่ระดับ .05 ซึ่งสามารถแสดงแผนภาพกราฟความสัมพันธ์ระหว่างค่าพารามิเตอร์ความยากของข้อสอบวิชาภาษาไทย ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA แสดงดังภาพต่อไปนี้



ภาพที่ 4-10 กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ค่าความยากของข้อสอบวิชาภาษาไทย ที่ประมาณค่าด้วยวิธี ML และวิธี BAYESIAN



ภาพที่ 4-11 กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ค่าความยากข้อสอบวิชาภาษาไทย ที่ประมาณค่าด้วยวิธี ML และวิธี CFA



ภาพที่ 4-12 กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ค่าความยากของข้อสอบวิชาภาษาไทย ที่ประมาณค่าด้วยวิธี BAYESIAN และวิธี CFA

ตอนที่ 4 การเปรียบเทียบผลการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบ

การเปรียบเทียบผลการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบในการวิจัยครั้งนี้ ประกอบด้วย วิชาคณิตศาสตร์ และภาษาไทย ใช้วิธีการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบ ได้แก่ วิธี ML, วิธี BAYSIAN และวิธี CFA โดยประยุกต์ใช้โปรแกรม IRT PRO, Open BUGS และ Mplus ตามลำดับ แสดงตามตารางที่ 4-13 ถึง 4-15 ตามลำดับ

ตารางที่ 4-13 สถิติพื้นฐานของผลการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบวิชาคณิตศาสตร์ ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA

ผลการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบ (θ)	*n*	*Min*	*Max*	*SD*	$\overline{X}$	*SE*
ML	2,400	-2.00	3.02	.95	-.00	.02
BAYESIAN	2,400	-2.00	3.12	.95	.00	.02
CFA	2,400	-2.00	3.09	.95	-.00	.02

จากตารางที่ 4-13 พบว่า ค่าพารามิเตอร์ความสามารถของผู้สอบวิชาคณิตศาสตร์จากการวิเคราะห์ด้วยวิธี ML มีค่าพารามิเตอร์ความสามารถของผู้สอบสูงที่สุด เท่ากับ 3.02 และต่ำที่สุด เท่ากับ -2.00 ส่วนเบี่ยงเบนมาตรฐานค่าพารามิเตอร์ความสามารถของผู้สอบ เท่ากับ .95 มีค่าเฉลี่ยค่าพารามิเตอร์ความสามารถ เท่ากับ -.00 และความคลาดเคลื่อนของค่าเฉลี่ย เท่ากับ .02

สำหรับการวิเคราะห์ค่าพารามิเตอร์ความสามารถของผู้สอบวิชาคณิตศาสตร์ จากการวิเคราะห์ด้วยวิธี BAYESIAN มีค่าพารามิเตอร์ความสามารถของผู้สอบสูงที่สุด เท่ากับ 3.12 และต่ำที่สุด เท่ากับ -2.00 ส่วนเบี่ยงเบนมาตรฐานค่าพารามิเตอร์ความสามารถของผู้สอบ เท่ากับ .95 มีค่าเฉลี่ยค่าพารามิเตอร์ความสามารถ เท่ากับ .00 และความคลาดเคลื่อนของค่าเฉลี่ย เท่ากับ .02 และการวิเคราะห์ค่าพารามิเตอร์ความสามารถของผู้สอบวิชาคณิตศาสตร์ ด้วยวิธี CFA มีค่าพารามิเตอร์ความสามารถของผู้สอบสูงที่สุด เท่ากับ 3.09 และต่ำที่สุด เท่ากับ -2.00 ส่วนเบี่ยงเบนมาตรฐานค่าพารามิเตอร์ความสามารถของผู้สอบ เท่ากับ .95 มีค่าเฉลี่ยค่าพารามิเตอร์ความสามารถของผู้สอบ เท่ากับ -.00 และความคลาดเคลื่อนของค่าเฉลี่ย เท่ากับ .02

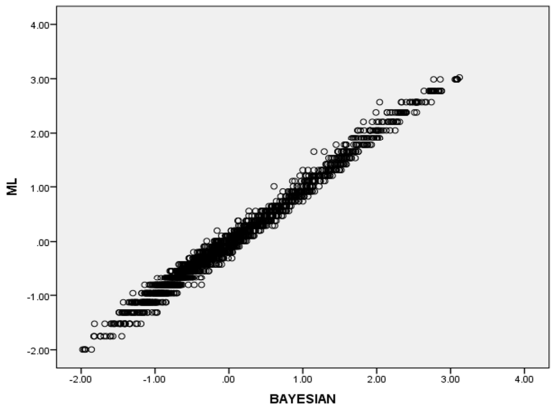
เมื่อนำผลการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบวิชาคณิตศาสตร์ที่ได้จากการวิเคราะห์ด้วยวิธี ML, วิธี BAYESIAN และวิธี CFA มาหาความสัมพันธ์ระหว่างกัน ด้วยการหาค่าสหสัมพันธ์เพียร์สัน ได้ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน (Pearson's correlation coefficient) ดังตารางที่ 4-14

ตารางที่ 4-14 ผลการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าพารามิเตอร์ความสามารถของผู้สอบวิชาคณิตศาสตร์ ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA

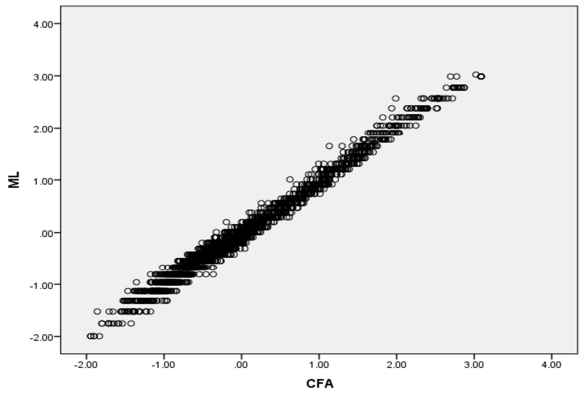
ตัวแปร	θ_{ML}	$\theta_{BAYESIAN}$	θ_{CFA}
θ_{ML}	1.00		
$\theta_{BAYESIAN}$	.99*	1.00	
θ_{CFA}	.99*	1.00*	1.00

* $p < .05$

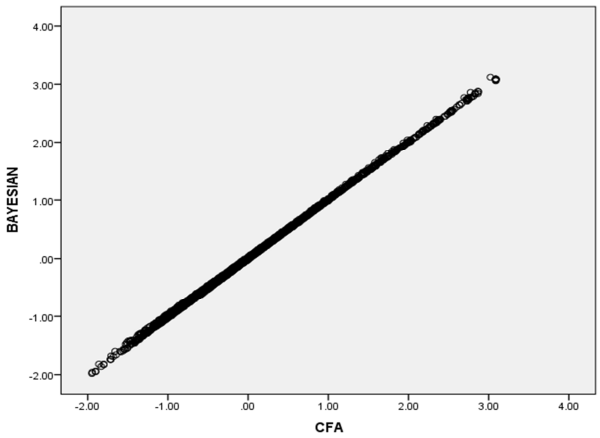
จากตารางที่ 4-14 ผลการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ของค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) วิชาคณิตศาสตร์ของทั้ง 3 วิธี ซึ่งได้แก่ วิธี ML, วิธี BAYESIAN และวิธี CFA พบว่า มีความสัมพันธ์ทางบวกในระดับสูงมาก เมื่อพิจารณาผลการประมาณค่าด้วยวิธี BAYESIAN เทียบกับผลการประมาณค่า CFA พบว่า $\theta_{BAYESIAN}$ กับ θ_{CFA} มีความสัมพันธ์ทางบวกในระดับสูงมาก เท่ากับ 1.00 และมีนัยสำคัญทางสถิติที่ระดับ .05 รองลงมา คือ ผลการประมาณค่าด้วยวิธี ML เทียบกับผลการประมาณค่าด้วยวิธี BAYESIAN พบว่า θ_{ML} กับ $\theta_{BAYESIAN}$ มีความสัมพันธ์ทางบวกในระดับสูงมาก เท่ากับ .99 และมีนัยสำคัญทางสถิติที่ระดับ .05 และผลการประมาณค่าด้วยวิธี ML เทียบกับผลการประมาณค่าด้วยวิธี CFA พบว่า θ_{ML} กับ θ_{CFA} มีความสัมพันธ์ทางบวกในระดับสูงมาก เท่ากับ .99 และมีนัยสำคัญทางสถิติที่ระดับ .05 เช่นกัน ซึ่งสามารถแสดงแผนภาพความสัมพันธ์ระหว่างค่าพารามิเตอร์ความสามารถของผู้สอบ (θ_j) วิชาคณิตศาสตร์ด้วยวิธี ML วิธี BAYESIAN และ วิธี CFA แสดงดังภาพต่อไปนี้



ภาพที่ 4-13 กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ค่าความสามารถของผู้สอบวิชาคณิตศาสตร์ที่ประมาณค่าด้วยวิธี ML และวิธี BAYESIAN



ภาพที่ 4-14 กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ค่าความสามารถของผู้สอบวิชาคณิตศาสตร์ที่ประมาณค่าด้วยวิธี ML และวิธี CFA



ภาพที่ 4-15 กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ค่าความสามารถของผู้สอบวิชาคณิตศาสตร์ที่ประมาณค่าด้วยวิธี BAYESIAN และวิธี CFA

ตารางที่ 4-15 สถิติพื้นฐานของผลการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบวิชาภาษาไทย ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA

ผลการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบ (θ)	*n*	*Min*	*Max*	*SD*	$\overline{X}$	SE
ML	2,400	-2.74	3.01	.96	-.00	.02
BAYESIAN	2,400	-2.69	3.12	.97	.00	.02
CFA	2,400	-2.67	3.09	.96	.00	.02

จากตารางที่ 4-15 พบว่า ค่าพารามิเตอร์ความสามารถของผู้สอบวิชาภาษาไทยจากการวิเคราะห์ด้วยวิธี ML มีค่าพารามิเตอร์ความสามารถของผู้สอบสูงที่สุด เท่ากับ 3.01 และต่ำที่สุด เท่ากับ -2.74 ส่วนเบี่ยงเบนมาตรฐานค่าพารามิเตอร์ความสามารถของผู้สอบ เท่ากับ .96 มีค่าเฉลี่ยพารามิเตอร์ความสามารถ เท่ากับ -.00 และความคลาดเคลื่อนของค่าเฉลี่ย เท่ากับ .02 สำหรับการวิเคราะห์ค่าพารามิเตอร์ความสามารถของผู้สอบวิชาภาษาไทยจากการวิเคราะห์ด้วยวิธี BAYESIAN มีค่าพารามิเตอร์ความสามารถของผู้สอบสูงที่สุด เท่ากับ 3.12 และต่ำที่สุด เท่ากับ -2.69 ส่วนเบี่ยงเบนมาตรฐานค่าพารามิเตอร์ความสามารถของผู้สอบ เท่ากับ .97 มีค่าเฉลี่ยพารามิเตอร์ความสามารถของผู้สอบ เท่ากับ .00 และความคลาดเคลื่อนของค่าเฉลี่ย เท่ากับ .02 และวิเคราะห์ค่าพารามิเตอร์ความสามารถของผู้สอบวิชาภาษาไทย จากการวิเคราะห์ด้วยวิธี CFA มีค่าพารามิเตอร์ความสามารถของผู้สอบสูงที่สุด เท่ากับ 3.09 และต่ำที่สุด เท่ากับ -2.67 ส่วนเบี่ยงเบนมาตรฐานค่าพารามิเตอร์ความสามารถของผู้สอบ เท่ากับ .96 มีค่าเฉลี่ยพารามิเตอร์ความสามารถของผู้สอบ เท่ากับ .00 และความคลาดเคลื่อนของค่าเฉลี่ย เท่ากับ .02

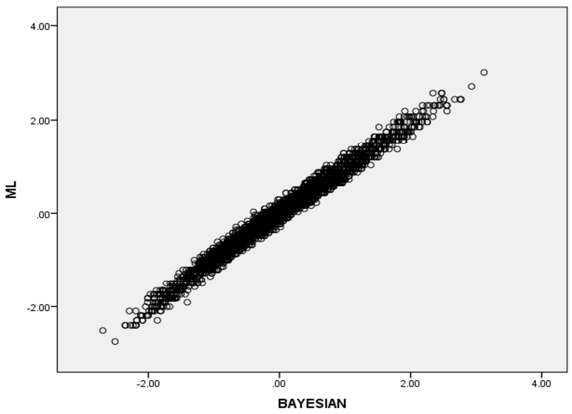
เมื่อนำผลการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบวิชาคณิตศาสตร์ที่ได้จากการวิเคราะห์ด้วยวิธี ML, วิธี BAYESIAN และวิธี CFA มาหาความสัมพันธ์ระหว่างกันด้วยการหาค่าสหสัมพันธ์เพียร์สัน ได้ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน (Pearson's correlation coefficient) ดังตารางที่ 4-16

ตารางที่ 4-16 ผลการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าพารามิเตอร์ความสามารถของผู้สอบวิชาภาษาไทย ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA

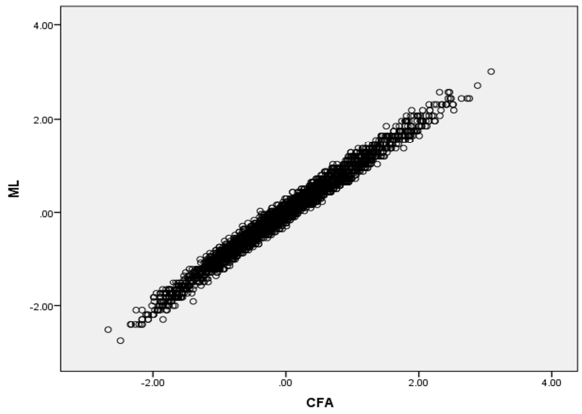
ตัวแปร	θ_{ML}	$\theta_{BAYESIAN}$	θ_{CFA}
θ_{ML}	1.00		
$\theta_{BAYESIAN}$	.99*	1.00	
θ_{CFA}	.99*	1.00*	1.00

* $p < .05$

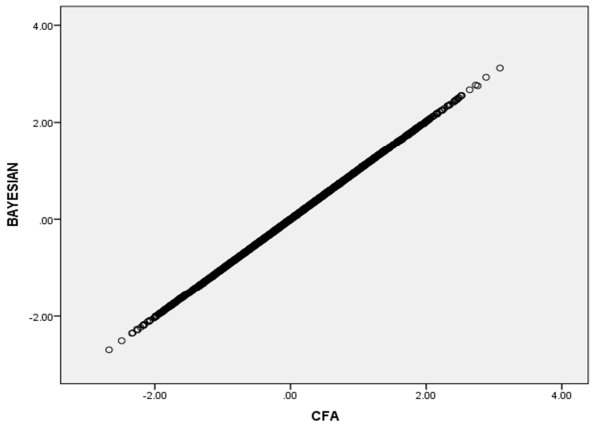
จากตารางที่ 4-16 ผลการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ของค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) วิชาภาษาไทยของทั้ง 3 วิธี ซึ่งได้แก่ วิธี ML, BAYESIAN และวิธี CFA พบว่า มีความสัมพันธ์ทางบวกในระดับสูงมาก เมื่อพิจารณาผลการประมาณค่าด้วยวิธี BAYESIAN เทียบกับวิธี CFA พบว่า $\theta_{BAYESIAN}$ กับ θ_{CFA} มีความสัมพันธ์ทางบวกในระดับสูงมาก เท่ากับ 1.00 และมีนัยสำคัญทางสถิติที่ระดับ .05 รองลงมา คือ ผลการประมาณค่าด้วยวิธี ML เทียบกับวิธี BAYESIAN พบว่าθ_{ML} กับ$\theta_{BAYESIAN}$ มีความสัมพันธ์ในระดับสูงมาก เท่ากับ .99 และมีนัยสำคัญทางสถิติที่ระดับ .05 และผลการประมาณค่าด้วยวิธี ML เทียบกับวิธี CFA พบว่า θ_{ML} กับ θ_{CFA} มีความสัมพันธ์ทางบวกในระดับสูงมาก เท่ากับ .99 และมีนัยสำคัญทางสถิติที่ระดับ .05 เช่นกัน ซึ่งสามารถแสดงแผนภาพความสัมพันธ์ระหว่างค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) วิชาภาษาไทยด้วยวิธี ML, วิธี BAYESIAN และวิธี CFA แสดงดังภาพต่อไปนี้



ภาพที่ 4-16 กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ค่าความสามารถของผู้สอบวิชาภาษาไทยที่ประมาณค่าด้วยวิธี ML และวิธี BAYESIAN



ภาพที่ 4-17 กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ค่าความสามารถของผู้สอบวิชาภาษาไทยที่ประมาณค่าด้วยวิธี ML และวิธี CFA



ภาพที่ 4-18 กราฟแสดงความสัมพันธ์ของค่าพารามิเตอร์ค่าความสามารถของผู้สอบวิชาภาษาไทยที่ประมาณค่าด้วยวิธี BAYESIAN และวิธี CFA

ตอนที่ 5 ผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบจำแนกตามเพศและสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน

การศึกษาการทำหน้าที่ต่างกันของข้อสอบในครั้งนี้ประกอบด้วยวิชาคณิตศาสตร์ และภาษาไทย มีจำนวนข้อสอบ 40 ข้อ และ 80 ข้อ ตามลำดับ ผู้วิจัยได้ใช้วิธี IRT-LR, BAYSIAN และ MGCFA โดยประยุกต์ใช้โปรแกรม IRT-PRO, Open BUGS และMplus ตามลำดับ โดยจำแนกตามเพศและสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียนซึ่งมีการวิเคราะห์ด้วยวิธีต่าง ๆ ดังนี้

1. การวิเคราะห์ IRT-LR ประยุกต์ใช้โปรแกรม IRT-PRO

การวิจัยครั้งนี้ผู้วิจัยได้ประยุกต์ใช้โปรแกรม IRT-PRO ในการวิเคราะห์ IRT-LR เพื่อศึกษาการทำหน้าที่ต่างกันของข้อสอบ ซึ่งประกอบด้วยวิชาคณิตศาสตร์ และภาษาไทย มีจำนวนข้อสอบ 40 ข้อ และ 80 ข้อ ตามลำดับ สามารถวิเคราะห์ได้ ดังนี้

ตารางที่ 4-17 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์ ด้วยวิธี IRT-LR ประยุกต์ใช้โปรแกรม IRT-PRO จำแนกตามเพศ

ข้อสอบข้อที่	Total χ^2	*df*	*p*
1	6.40*	2	.04
2	2.60	2	.28
3	.90	2	.65
4	2.10	2	.36
5	.00	2	1.00
6	4.70	2	.94
7	.00	2	.99
8	4.20	2	.12
9	7.40*	2	.02
10	.10	2	.97
11	6.50*	2	.04
12	18.20*	2	.00
13	2.60	2	.28
14	1.80	2	.40
15	1.40	2	.51
16	6.90*	2	.03
17	9.70*	2	.01
18	.2	2	.89
19	.2	2	.92
20	2.90	2	.23
21	.90	2	.63

ตารางที่ 4-17 (ต่อ)

ข้อสอบข้อที่	Total χ^2	*df*	*p*
22	2.00	2	.38
23	11.70*	2	.00
24	6.50*	2	.04
25	2.70	2	.26
26	4.40	2	.11
27	.60	2	.73
28	4.00	2	.14
29	3.70	2	.16
30	1.20	2	.55
31	.70	2	.71
32	4.60	2	.10
33	1.70	2	.43
34	4.70	2	.09
35	2.00	2	.36
36	1.50	2	.48
37	2.00	2	.37
38	.50	2	.77
39	9.60*	2	.01
40	.40	2	.81

* $p < .05$

จากตารางที่ 4-17 เป็นการวิเคราะห์ด้วยวิธี IRT-LR เพื่อศึกษาการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์ จำแนกตามเพศ ผลการวิเคราะห์ข้อสอบจำนวน 40 ข้อ พบว่า ข้อสอบข้อที่ 1, 9, 11, 12, 16, 17, 23, 24 และข้อที่ 39 ที่อิทธิพลของตัวแปรเพศชายมีผลทำให้ค่า χ^2 มากกว่า 3.84 และมีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ซึ่งหมายความว่าตัวแปรความเป็นเพศชายส่งผลให้เกิดโอกาสในการตอบข้อสอบข้อที่ 1, 9, 11, 12, 16, 17, 23, 24 และ

ข้อที่ 39 ได้แตกต่างกันกับเพศหญิง สามารถบ่งชี้ว่าข้อสอบข้อนั้นทำหน้าที่ต่างกัน จากผลการศึกษาพบว่า เพศชายจะได้เปรียบในการตอบข้อสอบข้อที่ 1, 9, 11, 12, 16, 17, 23, 24 และข้อที่ 39

ตารางที่ 4-18 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทย ด้วยวิธี IRT-LR ประยุกต์ใช้โปรแกรม IRT-PRO จำแนกตามเพศ

ข้อสอบข้อที่	Total χ^2	*df*	*p*
1	1.8	2	.41
2	3.2	2	.20
3	.30	2	.87
4	2.7	2	.26
5	.50	2	.77
6	2.4	2	.31
7	4.6	2	.10
8	.10	2	.96
9	.00	2	.98
10	1.2	2	.55
11	5.0	2	.08
12	.40	2	.83
13	1.9	2	.38
14	.50	2	.78
15	.50	2	.78
16	.90	2	.63
17	3.9	2	.14
18	.70	2	.70
19	3.60	2	.16
20	1.00	2	.62
21	.80	2	.68
22	1.30	2	.52
23	2.50	2	.29

ตารางที่ 4-18 (ต่อ)

ข้อสอบข้อที่	Total χ^2	*df*	*p*
24	3.10	2	.21
25	.10	2	.97
26	3.40	2	.18
27	1.40	2	.49
28	1.60	2	.46
29	4.50	2	.10
30	.10	2	.96
31	1.50	2	.48
32	4.50	2	.11
33	1.40	2	.49
34	1.90	2	.39
35	.60	2	.74
36	4.40	2	.11
37	.70	2	.70
38	9.30*	2	.01
39	1.90	2	.38
40	.60	2	.75
41	.00	2	.99
42	1.90	2	.40
43	4.20	2	.12
44	3.80	2	.15
45	2.00	2	.36
46	3.40	2	.18
47	.30	2	.86
48	.40	2	.81
49	3.30	2	.19
50	2.80	2	.25

ตารางที่ 4-18 (ต่อ)

ข้อสอบข้อที่	Total χ^2	*df*	*p*
51	5.30	2	.07
52	4.00	2	.14
53	3.80	2	.15
54	4.10	2	.13
55	.20	2	.91
56	3.30	2	.19
57	3.80	2	.15
58	.40	2	.82
59	2.50	2	.28
60	6.60*	2	.04
61	3.10	2	.22
62	.50	2	.76
63	2.30	2	.31
64	3.60	2	.16
65	1.10	2	.57
66	1.40	2	.49
67	1.60	2	.44
68	1.00	2	.61
69	3.10	2	.22
70	.30	2	.84
71	.00	2	.98
72	6.90*	2	.03
73	5.10	2	.79
74	2.70	2	.26
75	1.20	2	.55
76	.10	2	.96
77	1.10	2	.57

ตารางที่ 4-18 (ต่อ)

ข้อสอบข้อที่	Total χ^2	*df*	*p*
78	5.70	2	.06
79	1.70	2	.43
80	.90	2	.64

* $p < .05$

จากตารางที่ 4-18 เป็นการวิเคราะห์ด้วยวิธี IRT-LR เพื่อศึกษาการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทย จำแนกตามเพศ ผลการวิเคราะห์ข้อสอบจำนวน 40 ข้อ พบว่า ข้อสอบข้อที่ 38, 60 และข้อที่ 72 ที่อิทธิพลของตัวแปรเพศหญิง มีผลทำให้ค่า χ^2 มากกว่า 3.84 และมีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ซึ่งหมายความว่าตัวแปรความเป็นเพศหญิงส่งผลให้เกิดโอกาสในการตอบข้อสอบข้อที่ 38, 60 และข้อที่ 72 ได้แตกต่างกันระหว่างเพศ สามารถบ่งชี้ว่าข้อสอบนั้นทำหน้าที่ต่างกัน จากผลการศึกษา พบว่า เพศหญิงจะได้เปรียบในการตอบข้อสอบข้อที่ 38, 60 และข้อที่ 72

ตารางที่ 4-19 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์ด้วยวิธี IRT-LR ประยุกต์ใช้โปรแกรม IRT-PRO จำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน

ข้อสอบข้อที่	Total χ^2	*df*	*p*
1	.60	2	.74
2	1.10	2	.58
3	7.10*	2	.03
4	.50	2	.79
5	6.60*	2	.04
6	2.80	2	.25
7	2.90	2	.24
8	4.10	2	.13
9	6.00	2	.05

ตารางที่ 4-19 (ต่อ)

ข้อสอบข้อที่	Total χ^2	*df*	*p*
10	17.40*	2	.00
11	.30	2	.88
12	3.50	2	.18
13	.50	2	.76
14	2.50	2	.29
15	.10	2	.93
16	1.00	2	.61
17	6.80*	2	.03
18	3.60	2	.17
19	10.80*	2	.00
20	.10	2	.93
21	.60	2	.75
22	.10	2	.94
23	2.60	2	.27
24	1.30	2	.51
25	.30	2	.86
26	.20	2	.89
27	.30	2	.86
28	.80	2	.66
29	.70	2	.72
30	.80	2	.65
31	3.70	2	.16
32	.90	2	.64
33	1.00	2	.61
34	3.70	2	.16
35	.70	2	.72
36	5.60	2	.06

ตารางที่ 4-19 (ต่อ)

ข้อสอบข้อที่	Total χ^2	*df*	*p*
37	4.30	2	.12
38	4.80	2	.09
39	17.30*	2	.00
40	11.00*	2	.00

* $p < .05$

จากตารางที่ 4-19 เป็นการวิเคราะห์ด้วยวิธี IRT-LR เพื่อศึกษาการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์ จำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน ผลการวิเคราะห์ข้อสอบจำนวน 40 ข้อ พบว่า ข้อสอบข้อที่ 3, 5, 10, 17, 19, 39 และข้อที่ 40 ที่อิทธิพลของตัวแปรความเป็นเขตกรุงเทพมหานครและปริมณฑล มีผลทำให้ค่า χ^2 มากกว่า 3.84 และมีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ซึ่งหมายความว่าตัวแปรความเป็นเขตกรุงเทพมหานครและปริมณฑลส่งผลให้เกิดโอกาสในการตอบข้อสอบข้อที่ 3, 5, 10, 17, 19, 39 และข้อที่ 40 ได้แตกต่างกันระหว่างสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน สามารถบ่งชี้ว่าข้อสอบนั้นทำหน้าที่ต่างกัน จากผลการศึกษา พบว่า เขตกรุงเทพมหานครและปริมณฑล จะได้เปรียบในการตอบข้อที่ 3, 5, 10, 17, 19, 39 และข้อที่ 40

ตารางที่ 4-20 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทยด้วยวิธี IRT-LR ประยุกต์ใช้โปรแกรม IRT-PRO จำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน

ข้อสอบข้อที่	Total χ^2	*df*	*P*
1	1.40	2	.49
2	3.10	2	.21
3	1.70	2	.42
4	0.80	2	.68
5	0.40	2	.84
6	1.50	2	.48

ตารางที่ 4-20 (ต่อ)

ข้อสอบข้อที่	Total χ^2	*df*	*p*
7	1.40	2	.49
8	1.20	2	.54
9	1.80	2	.40
10	1.80	2	.40
11	1.00	2	.60
12	1.70	2	.42
13	4.20	2	.12
14	4.70	2	.10
15	5.90	2	.05
16	0.80	2	.67
17	0.10	2	.95
18	2.00	2	.37
19	1.50	2	.46
20	3.80	2	.15
21	0.40	2	.80
22	1.00	2	.59
23	4.30	2	.12
24	5.10	2	.08
25	6.20*	2	.04
26	.50	2	.78
27	1.60	2	.44
28	.30	2	.87
29	3.10	2	.21
30	4.90	2	.08
31	1.20	2	.55
32	.30	2	.88
33	2.60	2	.27

ตารางที่ 4-20 (ต่อ)

ข้อสอบข้อที่	Total χ^2	*df*	*p*
34	3.70	2	.16
35	9.70*	2	.01
36	7.20*	2	.03
37	5.40	2	.07
38	5.60	2	.06
39	1.80	2	.40
40	.40	2	.80
41	1.10	2	.56
42	1.80	2	.40
43	.00	2	1.00
44	2.80	2	.25
45	2.60	2	.28
46	4.10	2	.13
47	1.20	2	.55
48	11.10*	2	.00
49	.60	2	.74
50	5.20	2	.07
51	1.70	2	.43
52	4.40	2	.11
53	.30	2	.86
54	4.90	2	.08
55	6.90*	2	.03
56	1.50	2	.48
57	14.80*	2	.00
58	2.90	2	.23
59	4.50	2	.10
60	5.30	2	.07

ตารางที่ 4-20 (ต่อ)

ข้อสอบข้อที่	Total χ^2	*df*	*p*
61	1.50	2	.47
62	2.90	2	.24
63	9.50*	2	.01
64	3.10	2	.21
65	5.20	2	.07
66	5.80	2	.05
67	.70	2	.70
68	1.30	2	.53
69	4.30	2	.12
70	3.60	2	.16
71	.10	2	.97
72	3.10	2	.22
73	1.00	2	.61
74	2.30	2	.31
75	1.30	2	.51
76	.00	2	1.00
77	1.60	2	.45
78	2.50	2	.28
79	2.60	2	.27
80	1.70	2	.43

* $p < .05$

จากตารางที่ 4-20 เป็นการวิเคราะห์ด้วยวิธี IRT-LR เพื่อศึกษาการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทย จำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน ผลการวิเคราะห์ข้อสอบจำนวน 80 ข้อ พบว่า ข้อสอบข้อที่ 25, 35, 36, 48, 55, 57 และข้อที่ 63 ที่อิทธิพลของตัวแปรความเป็นเขตกรุงเทพมหานคร และปริมณฑล มีผลทำให้ค่า χ^2 มากกว่า 3.84 และมีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ซึ่งหมายความว่าตัวแปรความเป็นเขต

กรุงเทพมหานครและปริมณฑลส่งผลให้เกิดโอกาสในการตอบข้อสอบข้อที่ 25, 35, 36, 48, 55, 57 และข้อที่ 63 แตกต่างกันระหว่างสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน สามารถบ่งชี้ว่าข้อสอบนั้นทำหน้าที่ต่างกัน จากผลการศึกษา พบว่า เขตกรุงเทพมหานคร และปริมณฑล จะได้เปรียบในการตอบข้อที่ 25, 35, 36, 48, 55, 57 และข้อที่ 63

2. การวิเคราะห์ด้วยวิธี BAYESIAN

การวิจัยครั้งนี้ผู้วิจัยได้ประยุกต์ใช้โปรแกรม Open BUGS ในการศึกษาการทำหน้าที่ต่างกันของข้อสอบ ซึ่งประกอบด้วย วิชาคณิตศาสตร์ และภาษาไทย มีจำนวนข้อสอบ 40 ข้อ และ 80 ข้อ ตามลำดับ สามารถวิเคราะห์ได้ ดังนี้

ตารางที่ 4-21 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์ ด้วยวิธี BAYSIAN จำแนกตามเพศ

Parameter*	**Mean**	***SE***	***Val2.5pc***	***Val97.5pc***
β_1^{adj}	.13	.09	-.05	.33
β_2^{adj}	.05	.10	-.13	.24
β_3^{adj}	-.17	.09	-.35	.01
β_4^{adj}	-.12	.11	-.33	.09
β_5^{adj}	-.06	.08	-.24	.09
β_6^{adj}	-.27*	.08	-.42	-.11
β_7^{adj}	-.04	.11	-.27	.16
β_8^{adj}	-.04	.08	-.19	.12
β_9^{adj}	-.09	.12	-.33	.16
β_{10}^{adj}	-.06	.11	-.25	.16
β_{11}^{adj}	-.00	.14	-.27	.27
β_{12}^{adj}	.20*	.08	.05	.37
β_{13}^{adj}	-.19*	.08	-.34	-.01
β_{14}^{adj}	-.17	.14	-.46	.11
β_{15}^{dj}	-.04	.11	-.25	.17
β_{16}^{adj}	-.31*	.08	-.47	-.15

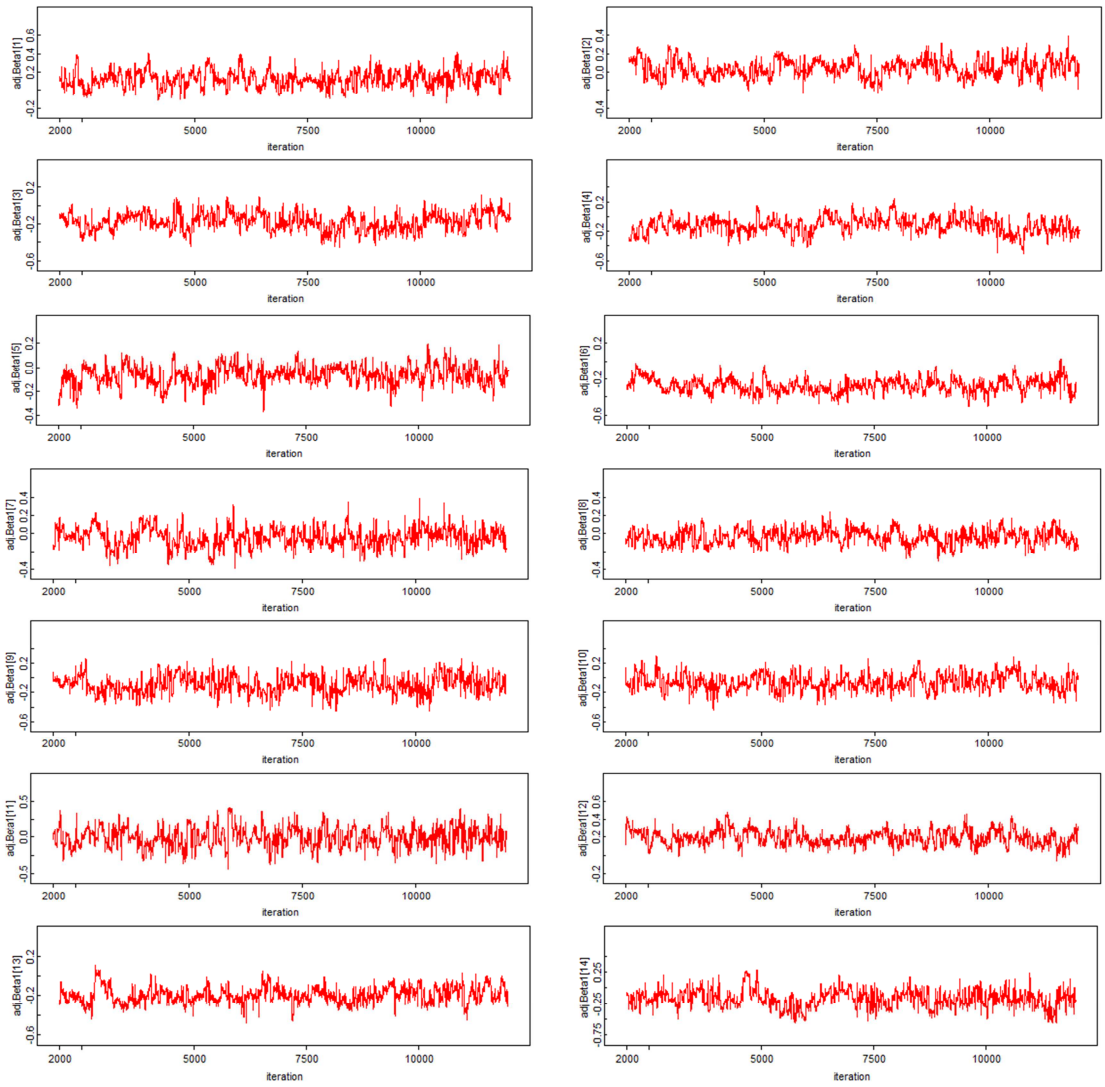
ตารางที่ 4-21 (ต่อ)

Parameter*	Mean	*SE*	*Val2.5pc*	*Val97.5pc*
β_{17}^{adj}	-.40*	.09	-.58	-.24
β_{18}^{adj}	-.09	.10	-.29	.09
β_{19}^{adj}	-.09	.10	-.30	.10
β_{20}^{adj}	.02	.07	-.12	.16
β_{21}^{adj}	-.04	.07	-.18	.10
β_{22}^{adj}	-.12	.12	-.34	.13
β_{23}^{adj}	-.41*	.11	-.64	-.21
β_{24}^{adj}	.31	.19	-.03	.73
β_{25}^{adj}	-.25*	.11	-.46	-.03
β_{26}^{adj}	-.17*	.07	-.31	-.04
β_{27}^{adj}	-.12	.09	-.31	.05
β_{28}^{adj}	.08	.10	-.11	.28
β_{29}^{adj}	.02	.11	-.20	.23
β_{30}^{adj}	2.95*	.77	1.51	4.56
β_{31}^{adj}	.00	.07	-.12	.13
β_{32}^{adj}	.11	.08	-.03	.27
β_{33}^{adj}	-.08	.07	-.22	.44
β_{34}^{adj}	-.08	.07	-.20	.05
β_{35}^{adj}	-.08	.05	-.18	.02
β_{36}^{adj}	-.16*	.06	-.28	-.06
β_{37}^{adj}	-.13*	.06	-.26	-.02
β_{38}^{adj}	-.10	.07	-.23	.02
β_{39}^{adj}	.20*	.09	.03	.37
β_{40}^{adj}	-.15	.11	-.35	.03

* $p < .05$

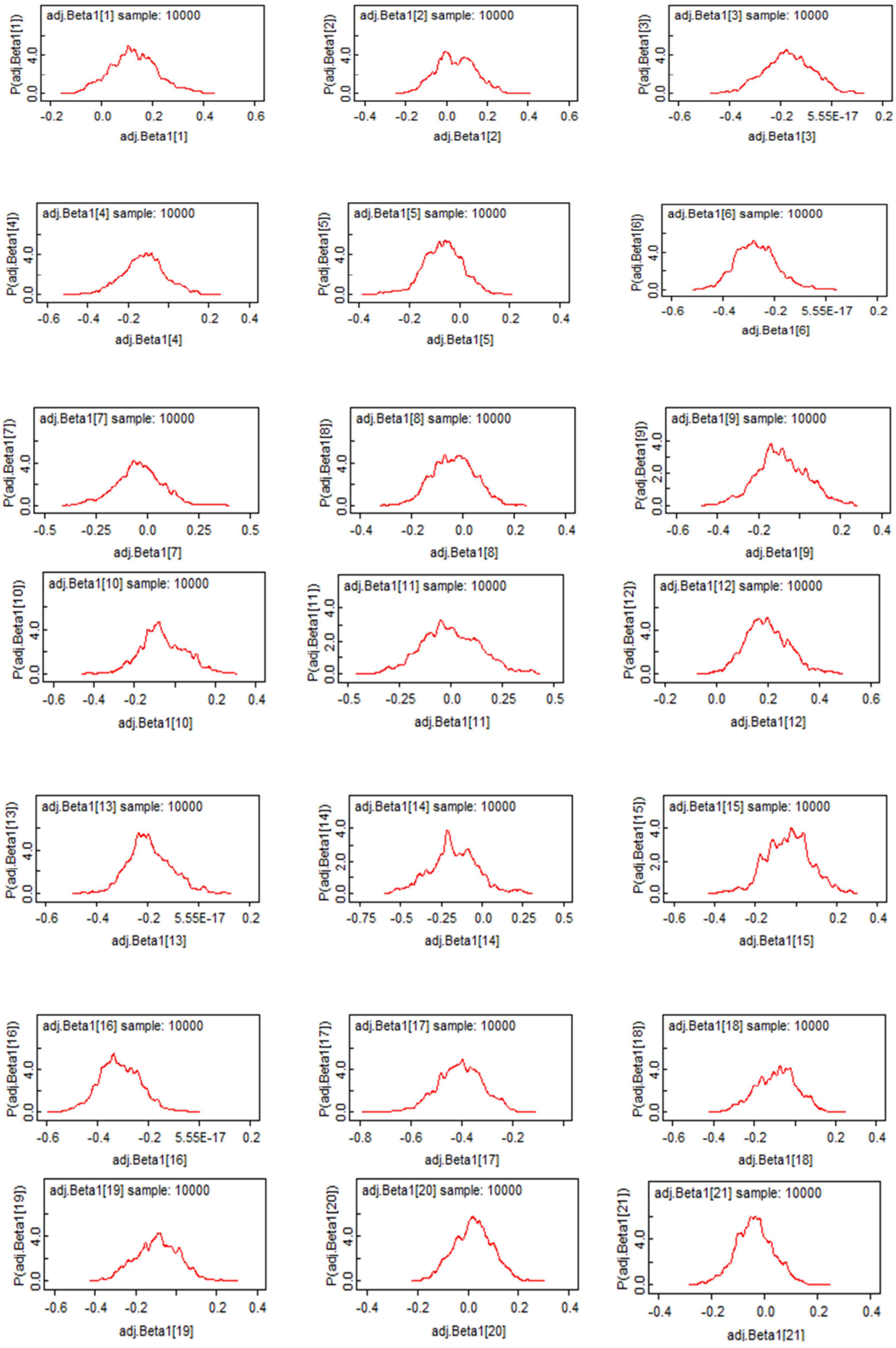
จากตารางที่ 4-21 เป็นการวิเคราะห์ด้วยวิธี BAYESIAN เพื่อศึกษาการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์ จำแนกตามเพศ ผลการวิเคราะห์ข้อสอบจำนวน 40 ข้อ พบว่ามีข้อสอบทำหน้าที่ต่างกันจำนวน 3 ข้อ ได้แก่ ข้อที่ 12, 30 และข้อที่ 39 โดยพิจารณาจากค่าที่อยู่ระหว่างขอบล่าง (val.2.5pc) และขอบบน (Val97.5pc) ของความเชื่อมั่นที่ 95% มีค่าไม่คลุม 0 ซึ่งหมายความว่า ตัวแปรความเป็นเพศชายส่งผลให้เกิดโอกาสในการตอบข้อสอบข้อที่ 12, 30 และข้อที่ 39 สามารถบ่งชี้ได้ว่าข้อสอบนั้นทำหน้าที่ต่างกันกับเพศหญิง จากผลการศึกษา พบว่าเพศชายจะได้เปรียบในการตอบข้อสอบข้อที่ 12, 30 และข้อที่ 39 และเพศหญิงจะได้เปรียบในการตอบข้อสอบข้อที่ 6, 13, 16, 17, 23, 25, 26, 36 และข้อที่ 37

ซึ่งสามารถแสดงแผนภาพคลื่นความถี่ (History) ของการเกิดการทำหน้าที่ต่างกันของข้อสอบ (DIF) ของวิชาคณิตศาสตร์ จำแนกตามเพศ ด้วยวิธี BAYESIAN ที่วิเคราะห์ด้วยโปรแกรม Open BUGS แสดงดังภาพต่อไปนี้



ภาพที่ 4-19 คลื่นความถี่ (History) การเกิดการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์ จำแนกตามเพศ ด้วยวิธี BAYESIAN

และสามารถแสดงแผนภาพคลื่นความถี่ (Density) ของการเกิดการทำหน้าที่ต่างกันของข้อสอบ (DIF) ของวิชาคณิตศาสตร์ จำแนกตามเพศ ด้วยวิธี BAYESIAN ที่วิเคราะห์ด้วยโปรแกรม Open BUGS ดังภาพต่อไปนี้



ภาพที่ 4-20 คลื่นความถี่ (Density) ของการเกิดการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์จำแนกตามเพศ ด้วยวิธี BAYESIAN

ตารางที่ 4-22 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทย ด้วยวิธี BAYSIAN จำแนกตามเพศ

Parameter*	Mean	*SE*	*Val2.5pc*	*Val97.5pc*
β_1^{adj}	-.26	.14	-.52	.04
β_2^{adj}	-.16	.09	-.34	.02
β_3^{adj}	-.18	.12	-.41	.06
β_4^{adj}	.17	.17	-.14	.52
β_5^{adj}	-.24*	.10	-.44	-.05
β_6^{adj}	-.14	.27	-.65	.44
β_7^{adj}	-.31*	.10	-.50	-.10
β_8^{adj}	-.15	.12	-.41	.10
β_9^{adj}	5.02*	1.09	2.88	7.07
β_{10}^{adj}	-.24*	.10	-.42	-.04
β_{11}^{adj}	-.38*	.12	-.60	-.13
β_{12}^{adj}	-.06	.18	-.40	.32
β_{13}^{adj}	4.64*	1.06	2.53	6.71
β_{14}^{adj}	-.08	.22	-.49	.34
β_{15}^{dj}	-.14	.20	-.53	.27
β_{16}^{adj}	-.06	.14	-.36	.21
β_{17}^{adj}	-.11	.09	-.27	.07
β_{18}^{adj}	-.12	.11	-.34	.10
β_{19}^{adj}	-.27	.15	-.54	.03
β_{20}^{adj}	.06	.20	-.36	.44
β_{21}^{adj}	-.19	.10	-.39	.01
β_{22}^{adj}	-.27*	.11	-.48	-.04
β_{23}^{adj}	-.03	.15	-.33	.27
β_{24}^{adj}	-.03	.11	-.23	.22
β_{25}^{adj}	-.15	.14	-.41	.14

ตารางที่ 4-22 (ต่อ)

Parameter*	Mean	*SE*	*Val2.5pc*	*Val97.5pc*
β_{26}^{adj}	-.29*	.11	-.51	-.09
β_{27}^{adj}	-.10	.10	-.31	.08
β_{28}^{adj}	-.24*	.10	-.45	-.05
β_{29}^{adj}	.27	.20	-.13	.65
β_{30}^{adj}	-.10	.16	-.41	.20
β_{31}^{adj}	-.15	.19	-.53	.18
β_{32}^{adj}	-.32*	.16	-.64	-.01
β_{33}^{adj}	.25	.25	-.21	.77
β_{34}^{adj}	-.27*	.13	-.52	-.02
β_{35}^{adj}	-.03	.25	-.51	.52
β_{36}^{adj}	-.07	.16	-.40	.27
β_{37}^{adj}	-.28	.10	-.47	.09
β_{38}^{adj}	.08	.62	-.63	1.90
β_{39}^{adj}	-.13	.12	-.34	.11
β_{40}^{adj}	-.26*	.12	-.50	-.04
β_{41}^{adj}	-.16	.11	-.36	.05
β_{42}^{adj}	-.03	.11	-.24	.18
β_{43}^{adj}	-.04	.12	-.27	.19
β_{44}^{adj}	-.30*	.15	-.60	-9.52
β_{45}^{adj}	-.28	.17	-.61	.05
β_{46}^{adj}	-.37*	.08	-.53	-.19
β_{47}^{adj}	-.22*	.10	-.43	-.03
β_{48}^{adj}	-.10	.15	-.38	.20
β_{49}^{adj}	-.10	.13	-.34	.16
β_{50}^{adj}	.07	.16	-.22	.40
β_{51}^{adj}	-.37*	.11	-.57	-.16

ตารางที่ 4-22 (ต่อ)

Parameter*	Mean	*SE*	*Val2.5pc*	*Val97.5pc*
β_{52}^{adj}	.02	.11	-.19	.24
β_{53}^{adj}	-.08	.12	-.30	.14
β_{54}^{adj}	1.70	2.07	-.41	6.28
β_{55}^{adj}	-.15	.16	-.46	.16
β_{56}^{adj}	-.06	.14	-.33	.20
β_{57}^{adj}	.01	.10	-.19	.22
β_{58}^{adj}	-.20*	.10	-.38	-.01
β_{59}^{adj}	-.11	.11	-.31	.12
β_{60}^{adj}	-.37*	.09	-.57	-.19
β_{61}^{adj}	-.11	.08	-.26	.04
β_{62}^{adj}	-.24*	.08	-.41	-.09
β_{63}^{adj}	.04	.13	-.21	.31
β_{64}^{adj}	.11	.15	-.21	.40
β_{65}^{adj}	-.16	.16	-.47	.19
β_{66}^{adj}	-.22*	.09	-.40	-.03
β_{67}^{adj}	-.18*	.09	-.37	-.01
β_{68}^{adj}	-.28*	.08	-.43	-.12
β_{69}^{adj}	-.18*	.09	-.35	-.01
β_{70}^{adj}	-.18*	.08	-.33	-.03
β_{71}^{adj}	-.21*	.11	-.41	.00
β_{72}^{adj}	-.47*	.12	-.69	-.24
β_{73}^{adj}	-.15	.23	-.59	.31
β_{74}^{adj}	-.06	.09	-.27	.11
β_{75}^{adj}	-0.16	0.13	-0.43	0.11
β_{76}^{adj}	-.21	.07	-.36	.08
β_{77}^{adj}	-.28*	.13	-.53	-.03

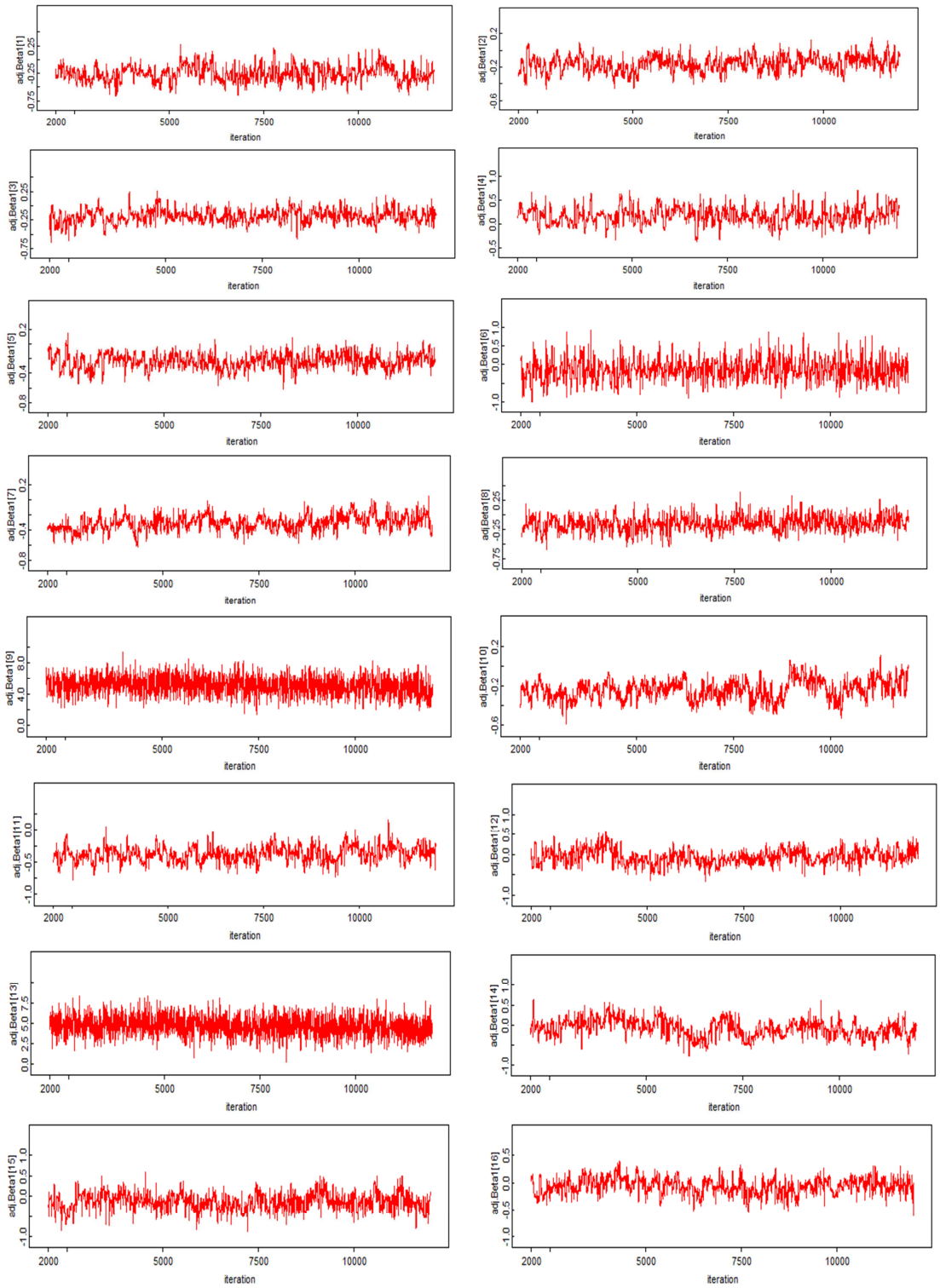
ตารางที่ 4-22 (ต่อ)

Parameter*	Mean	*SE*	*Val2.5pc*	*Val97.5pc*
β_{78}^{adj}	-.27*	.09	-.44	-.07
β_{79}^{adj}	-.20*	.09	-.38	-.03
β_{80}^{adj}	-.16	.14	-.44	.11

* $p < .05$

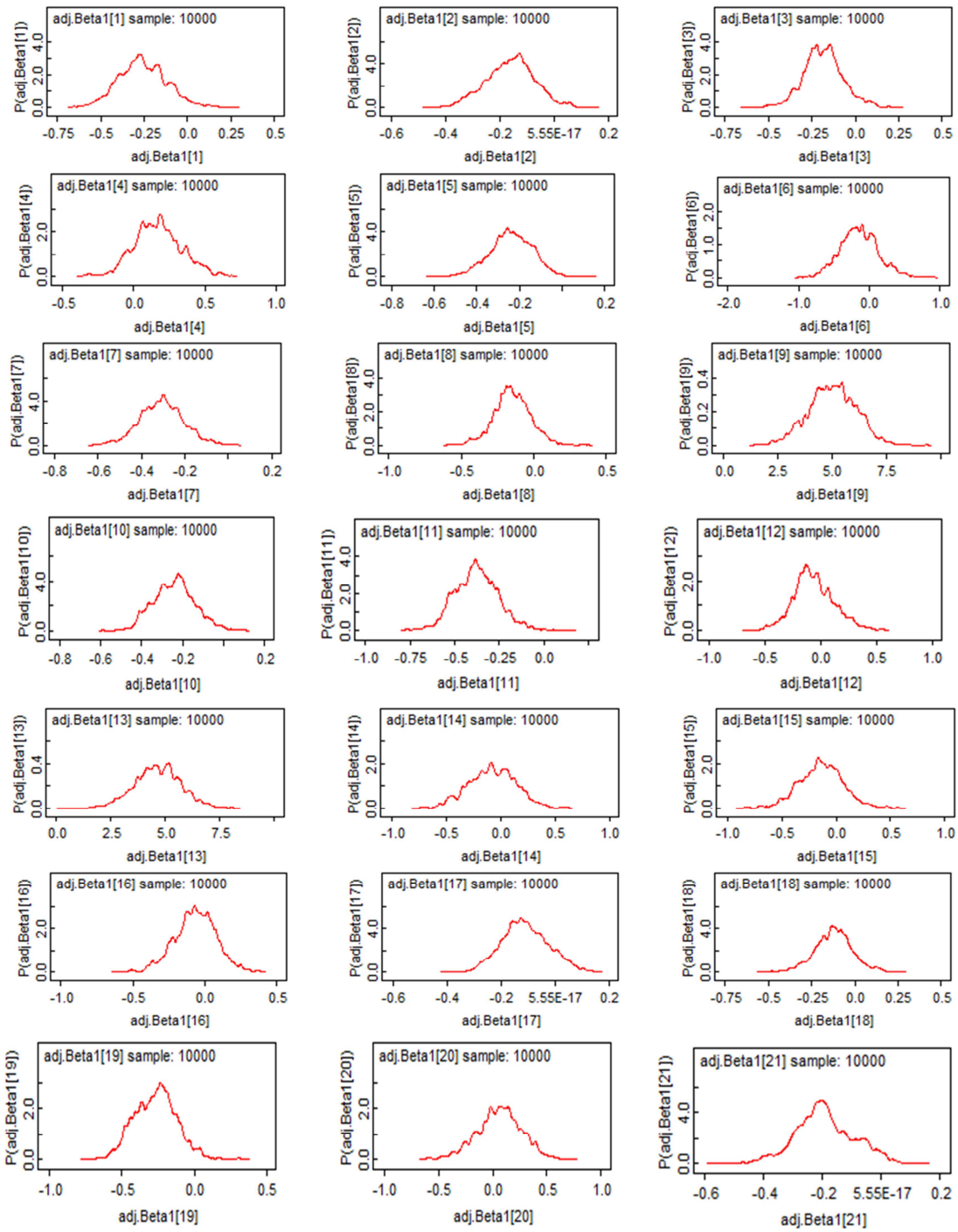
จากตารางที่ 4-22 เป็นการวิเคราะห์ด้วยวิธี BAYESIAN เพื่อศึกษาการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทย จำแนกตามเพศ ผลการวิเคราะห์ข้อสอบจำนวน 80 ข้อ พบว่า มีข้อสอบทำหน้าที่ต่างกัน จำนวน 2 ข้อ ได้แก่ ข้อที่ 9 และข้อที่ 13 โดยพิจารณาจากค่าที่อยู่ระหว่างขอบล่าง (val.2.5pc) และขอบบน (Val97.5pc) ของความเชื่อมั่นที่ 95% มีค่าไม่คลุม 0 ซึ่งหมายความว่าตัวแปรความเป็นเพศหญิงส่งผลให้เกิดโอกาสในการตอบข้อสอบข้อที่ 9 และข้อที่ 13 สามารถบ่งชี้ได้ว่าข้อสอบนั้นทำหน้าที่ต่างกันกับเพศชาย จากผลการศึกษา พบว่า เพศหญิงจะได้เปรียบในการตอบข้อสอบข้อที่ 9 และข้อที่ 13 และเพศชายจะได้เปรียบในการตอบข้อสอบข้อที่ 5, 7, 10, 11, 22, 26, 28, 32, 34, 40, 44, 46, 47, 51, 58, 60, 62, 66, 67, 68, 69, 70, 71, 72, 77, 78 และข้อที่ 79

ซึ่งสามารถแสดงแผนภาพคลื่นความถี่ (History) ของการเกิดการทำหน้าที่ต่างกันของข้อสอบ (DIF) ของวิชาภาษาไทย จำแนกตามเพศ ด้วยวิธี BAYESIAN ที่วิเคราะห์ด้วยโปรแกรม Open BUGS แสดงดังภาพต่อไปนี้



ภาพที่ 4-21 คลื่นความถี่ (History) การเกิดการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทย จำแนกตามเพศ ด้วยวิธี BAYESIAN

และสามารถแสดงแผนภาพคลื่นความถี่ (Density) ของการเกิดการทำหน้าที่ต่างกันของข้อสอบ (DIF) ของวิชาภาษาไทย จำแนกตามเพศ ด้วยวิธี BAYESIAN ที่วิเคราะห์ด้วยโปรแกรม Open BUGS ดังภาพต่อไปนี้



ภาพที่ 4-22 คลื่นความถี่ (Density) ของการเกิดการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทย จำแนกตามเพศ ด้วยวิธี BAYESIAN

ตารางที่ 4-23 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์ ด้วยวิธี BAYSIAN จำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน

Parameter*	Mean	*SE*	*Val2.5pc*	*Val97.5pc*
β_1^{adj}	-.13	.10	-.33	.04
β_2^{adj}	-.17	.10	-.33	.02
β_3^{adj}	-.11	.11	-.31	.10
β_4^{adj}	.02	.10	-.16	.23
β_5^{adj}	.02	.07	-.13	.16
β_6^{adj}	.11	.08	-.04	.28
β_7^{adj}	-.05	.12	-.27	.17
β_8^{adj}	-.01	.09	-.19	.16
β_9^{adj}	-.31*	.13	-.57	-.06
β_{10}^{adj}	.25*	.11	.03	.47
β_{11}^{adj}	.04	.14	-.29	.33
β_{12}^{adj}	-.14	.08	-.29	.02
β_{13}^{adj}	-.13	.09	-.30	.05
β_{14}^{adj}	-.07	.13	-.35	.19
β_{15}^{dj}	9.82	.13	-.24	.26
β_{16}^{adj}	-.03	.10	-.24	.15
β_{17}^{adj}	.08	.11	-.15	.28
β_{18}^{adj}	-.22*	.09	-.38	-.03
β_{19}^{adj}	-.06	.12	-.30	.17
β_{20}^{adj}	-.04	.08	-.20	.10
β_{21}^{adj}	.05	.08	-.09	.21
β_{22}^{adj}	-.04	.12	-.28	.19
β_{23}^{adj}	-.15	.11	-.38	.07
β_{24}^{adj}	.06	.18	-.30	.40
β_{25}^{adj}	-.07	.12	-.30	.17

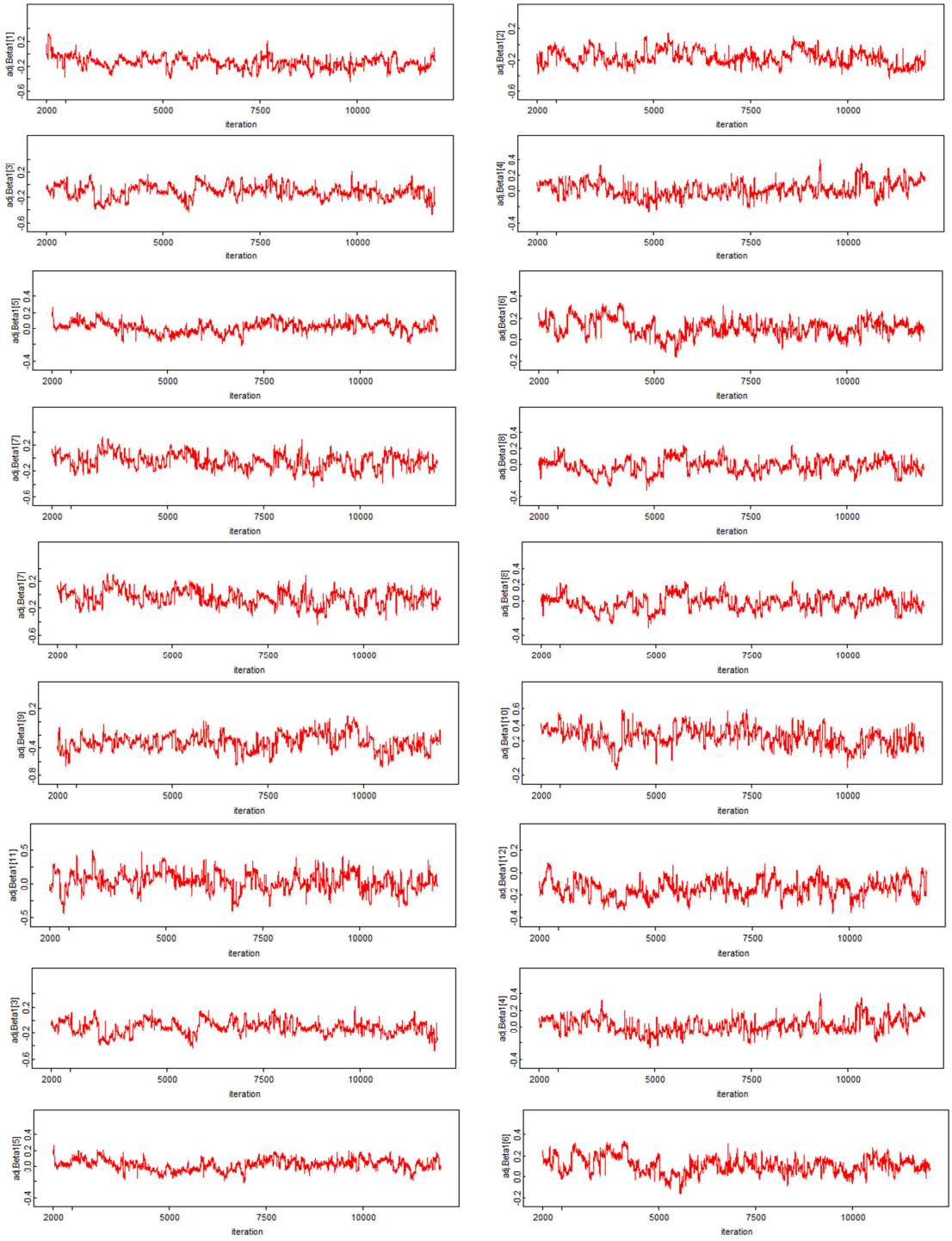
ตารางที่ 4-23 (ต่อ)

Parameter*	Mean	*SE*	*Val2.5pc*	*Val97.5pc*
β_{26}^{adj}	.04	.07	-.10	.17
β_{27}^{adj}	-.10	.11	-.35	.11
β_{28}^{adj}	-.01	.11	-.24	.20
β_{29}^{adj}	-.07	.12	-.33	.14
β_{30}^{adj}	2.73*	.69	1.42	4.09
β_{31}^{adj}	-.23*	.07	-.37	-.10
β_{32}^{adj}	-.14*	.07	-.27	-.01
β_{33}^{adj}	-.17*	.08	-.30	-.01
β_{34}^{adj}	-.03	.07	-.16	.10
β_{35}^{adj}	-.09	.07	-.24	.05
β_{36}^{adj}	-.02	.06	-.13	.09
β_{37}^{adj}	-.08	.07	-.22	.05
β_{38}^{adj}	-.22	.06	-.35	-.10
β_{39}^{adj}	-.12	.10	-.33	.10
β_{40}^{adj}	-.40*	.27	-1.26	-.10

* $p < .05$

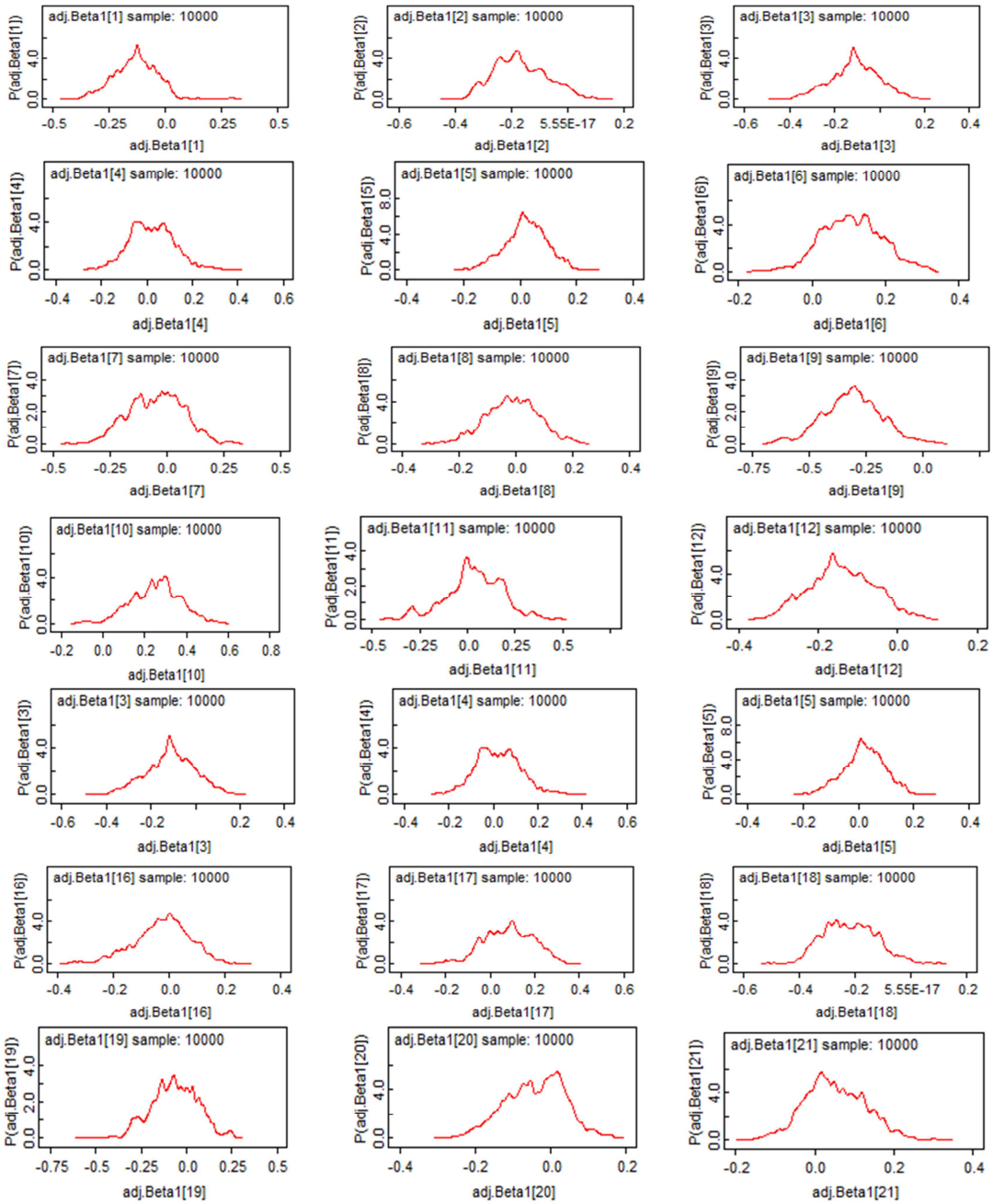
จากตารางที่ 4-23 เป็นการวิเคราะห์ด้วยวิธี BAYESIAN เพื่อศึกษาการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์ จำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน ผลการวิเคราะห์ข้อสอบ จำนวน 40 ข้อ พบว่า มีข้อสอบทำหน้าที่ต่างกันจำนวน 2 ข้อ ได้แก่ ข้อที่ 10 และข้อที่ 30 โดยพิจารณาจากค่าที่อยู่ระหว่างขอบล่าง (Val.2.5pc) และขอบบน (Val97.5pc) ของความเชื่อมั่นที่ 95% มีค่าไม่คลุม 0 ซึ่งหมายความว่า ตัวแปรความเป็นเขตกรุงเทพมหานครและปริมณฑลส่งผลให้เกิดโอกาสในการตอบข้อสอบข้อที่ 10 และข้อที่ 30 สามารถบ่งชี้ได้ว่าข้อสอบนั้นทำหน้าที่ต่างกันกับนอกเขตกรุงเทพมหานครและปริมณฑล จากผลการศึกษา พบว่า ในเขตกรุงเทพมหานคร และปริมณฑลจะได้เปรียบในการตอบข้อสอบข้อที่ 10 และข้อที่ 30 และนอกเขตกรุงเทพมหานครและปริมณฑลจะได้เปรียบในการตอบข้อสอบข้อที่ 9, 18, 31, 32, 33, 38 และข้อที่ 40

ซึ่งสามารถแสดงแผนภาพคลื่นความถี่ (History) ของการเกิดการทำหน้าที่ต่างกันของข้อสอบ (DIF) ของวิชาคณิตศาสตร์ จำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน ด้วยวิธี BAYESIAN ที่วิเคราะห์ด้วยโปรแกรม Open BUGS แสดงดังภาพต่อไปนี้



ภาพที่ 4-23 คลื่นความถี่ (History) การเกิดการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์ จำแนกตามเพศ ด้วยวิธี BAYESIAN

และสามารถแสดงแผนภาพคลื่นความถี่ (Density) ของการเกิดการทำหน้าที่ต่างกันของข้อสอบ (DIF) ของวิชาคณิตศาสตร์ จำแนกตามเพศ ด้วยวิธี BAYESIAN ที่วิเคราะห์ด้วยโปรแกรม Open BUGS ดังภาพต่อไปนี้



ภาพที่ 4-24 คลื่นความถี่ (Density) ของการเกิดการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์จำแนกตามเพศ ด้วยวิธี BAYESIAN

ตารางที่ 4-24 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทย ด้วยวิธี BAYSIAN จำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน

Parameter*	Mean	*SE*	*Val2.5pc*	*Val97.5pc*
β_1^{adj}	.24	.15	-.05	.53
β_2^{adj}	.09	.09	-.09	.28
β_3^{adj}	.24	.14	-.05	.51
β_4^{adj}	.24	.16	-.10	.55
β_5^{adj}	.06	.12	-.20	.30
β_6^{adj}	-.47	.33	-1.16	.10
β_7^{adj}	.02	.12	-.22	.24
β_8^{adj}	-.06	.16	-.39	.21
β_9^{adj}	-2.40*	.94	-4.31	-.69
β_{10}^{adj}	.26*	.09	.08	.45
β_{11}^{adj}	.00	.13	-.26	.25
β_{12}^{adj}	-.09	.19	-.49	.26
β_{13}^{adj}	-3.29*	.94	-5.22	-1.54
β_{14}^{adj}	-.03	.19	-.44	.33
β_{15}^{dj}	.29	.20	-.13	.66
β_{16}^{adj}	.13	.16	-.19	.43
β_{17}^{adj}	.14	.10	-.05	.32
β_{18}^{adj}	.05	.12	-.18	.27
β_{19}^{adj}	-.08	.15	-.40	.21
β_{20}^{adj}	.34	.19	-.04	.72
β_{21}^{adj}	.19	.10	-.00	.38
β_{22}^{adj}	.11	.12	-.13	.34
β_{23}^{adj}	.40*	.15	.10	.69
β_{24}^{adj}	.20	.12	-.02	.43
β_{25}^{adj}	.43*	.14	.16	.71

ตารางที่ 4-24 (ต่อ)

Parameter*	Mean	*SE*	*Val2.5pc*	*Val97.5pc*
β_{26}^{adj}	.21*	.10	.02	.43
β_{27}^{adj}	.05	.10	-.15	.25
β_{28}^{adj}	.13	.10	-.07	.33
β_{29}^{adj}	-.32	.22	-.79	.04
β_{30}^{adj}	.39*	.15	.10	.68
β_{31}^{adj}	.15	.21	-.30	.54
β_{32}^{adj}	-.03	.18	-.40	.31
β_{33}^{adj}	-.35	.27	-.94	.11
β_{34}^{adj}	-.17	.14	-.44	.10
β_{35}^{adj}	-.94*	.43	-1.94	-.18
β_{36}^{adj}	.43*	.15	.13	.75
β_{37}^{adj}	.02	.10	-.18	.21
β_{38}^{adj}	.11	.38	-.63	.77
β_{39}^{adj}	-.02	.11	-.26	.20
β_{40}^{adj}	.10	.11	-.09	.33
β_{41}^{adj}	.27*	.10	.06	.46
β_{42}^{adj}	.12	.11	-.09	.34
β_{43}^{adj}	.15	.12	-.10	.38
β_{44}^{adj}	-.08	.17	-.44	.25
β_{45}^{adj}	-.04	.19	-.40	.32
β_{46}^{adj}	-.05	.94	-.24	.13
β_{47}^{adj}	.23*	.11	.02	.44
β_{48}^{adj}	.43*	.16	.13	.73
β_{49}^{adj}	.05	.12	-.20	.27
β_{50}^{adj}	-.23	.16	-.57	.08
β_{51}^{adj}	-.03	.11	-.26	.20

ตารางที่ 4-24 (ต่อ)

Parameter*	Mean	*SE*	*Val2.5pc*	*Val97.5pc*
β_{52}^{adj}	-.13	.13	-.42	.08
β_{53}^{adj}	.18	.10	-.05	.37
β_{54}^{adj}	-.78	.64	-2.42	.14
β_{55}^{adj}	-.35	.18	-.72	.15
β_{56}^{adj}	.11	.15	-.19	.40
β_{57}^{adj}	.52*	.10	.32	.72
β_{58}^{adj}	-.03	.10	-.23	.16
β_{59}^{adj}	-.03	.11	-.25	.19
β_{60}^{adj}	.20*	.09	.03	.38
β_{61}^{adj}	.23*	.07	.08	.37
β_{62}^{adj}	.31*	.07	.17	.45
β_{63}^{adj}	.04	.14	-.24	.32
β_{64}^{adj}	.11	.15	-.17	.41
β_{65}^{adj}	-.17	.19	-.58	.17
β_{66}^{adj}	.36*	.09	.18	.56
β_{67}^{adj}	.23*	.10	.04	.43
β_{68}^{adj}	.09	.08	-.06	.25
β_{69}^{adj}	.28*	.09	.11	.45
β_{70}^{adj}	.25*	.08	.09	.41
β_{71}^{adj}	.15	.10	-.06	.36
β_{72}^{adj}	.20	.12	-.04	.43
β_{73}^{adj}	-.33	.31	-1.00	.19
β_{74}^{adj}	.30*	.09	.12	.47
β_{75}^{adj}	-.08	.14	-.34	.19
β_{76}^{adj}	.18*	.06	.06	.31
β_{77}^{adj}	.05	.14	-.23	.31

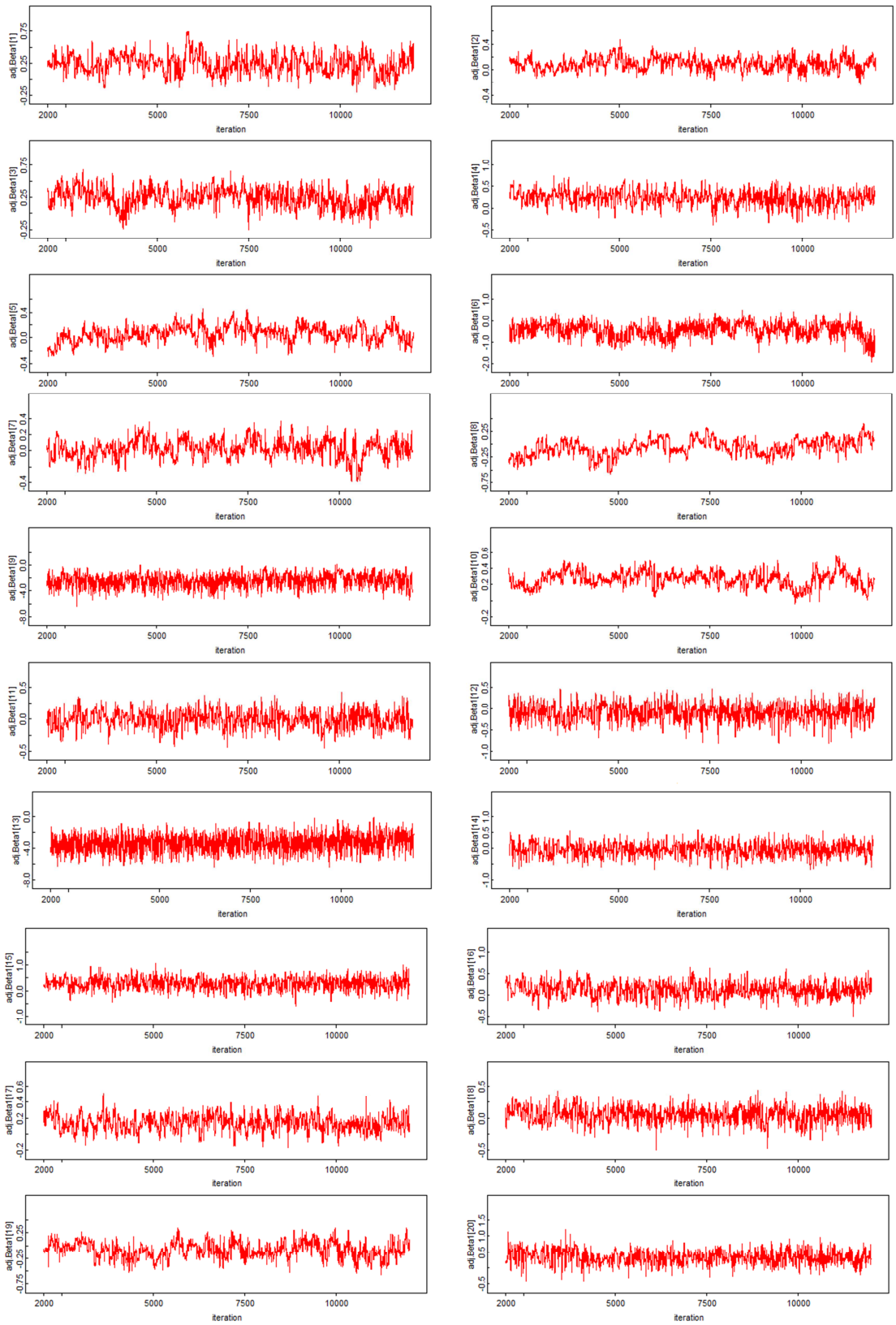
ตารางที่ 4-24 (ต่อ)

Parameter*	Mean	*SE*	*Val2.5pc*	*Val97.5pc*
β_{78}^{adj}	.29*	.09	.10	.47
β_{79}^{adj}	.30*	.09	.13	.47
β_{80}^{adj}	-.08	.16	-.38	.24

* $p < .05$

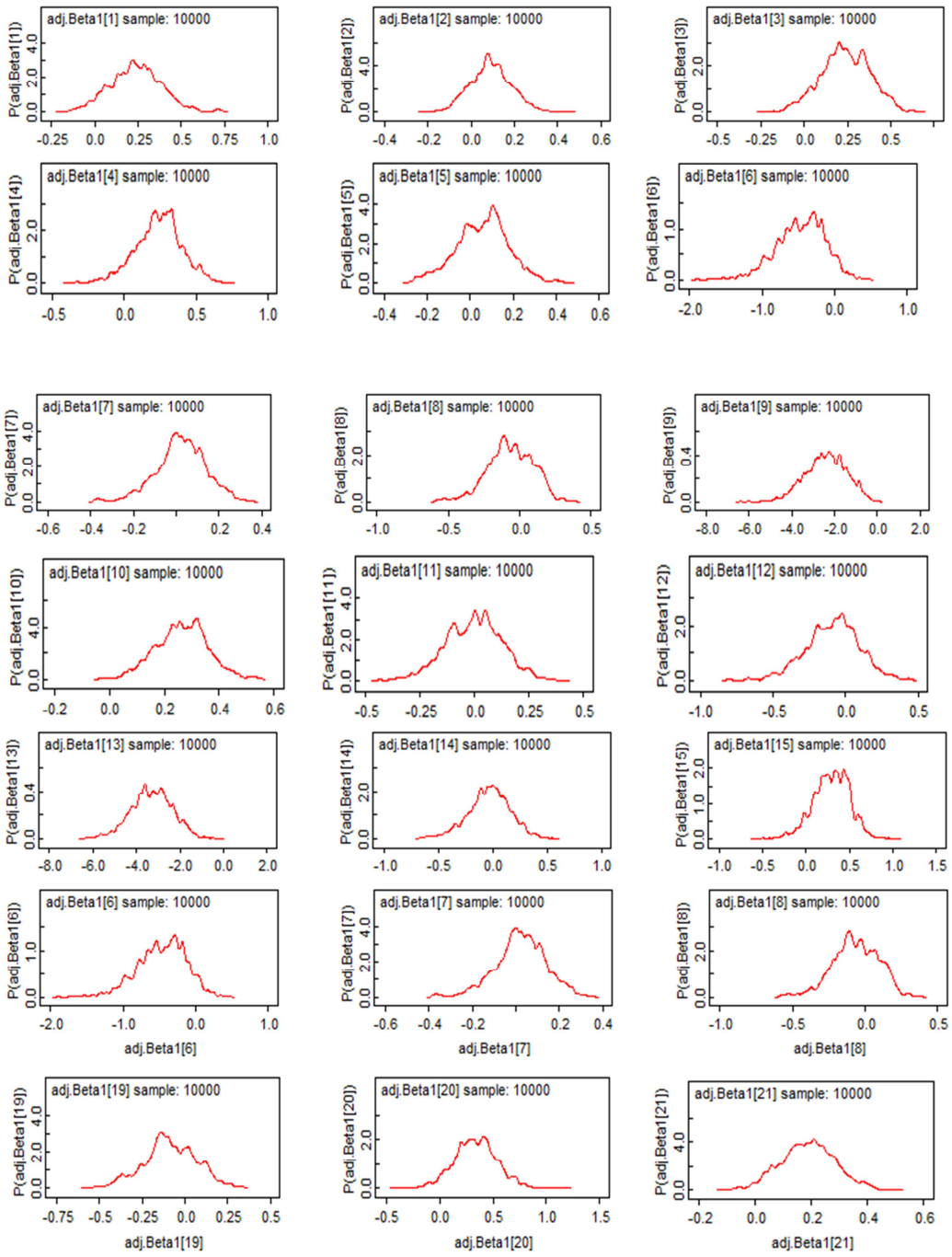
จากตารางที่ 4-24 เป็นการวิเคราะห์ด้วยวิธี BAYESIAN เพื่อศึกษาการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทย จำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน ผลการวิเคราะห์ข้อสอบจำนวน 80 ข้อ พบว่า มีข้อสอบทำหน้าที่ต่างกันจำนวน 21 ข้อ ได้แก่ ข้อที่ 10, 23, 25, 26, 30, 36, 41, 47, 48, 57, 60, 61, 62, 66, 67, 69, 70, 74, 76, 78 และข้อที่ 79 โดยพิจารณาจากค่าที่อยู่ระหว่างขอบล่าง (Val.2.5pc) และขอบบน (Val97.5pc) ของความเชื่อมั่นที่ 95% มีค่าไม่คลุม 0 ซึ่งหมายความว่า ตัวแปรความเป็นเขตกรุงเทพมหานครและปริมณฑลส่งผลให้เกิดโอกาสในการตอบข้อสอบข้อที่ 10, 23, 25, 26, 30, 36, 41, 47, 48, 57, 60, 61, 62, 66, 67, 69, 70,, 74, 76, 78 และข้อ79 สามารถบ่งชี้ได้ว่าข้อสอบข้อนั้นทำหน้าที่ต่างกันกับนอกเขตกรุงเทพมหานครและปริมณฑล จากผลการศึกษา พบว่า ในเขตกรุงเทพมหานครและปริมณฑล จะได้เปรียบในการตอบข้อสอบข้อที่ 10, 23, 25, 26, 30, 36, 41, 47, 48, 57, 60, 61, 62, 66, 67, 69, 70, 74, 76, 78 และข้อ 79 และนอกเขตกรุงเทพและปริมณฑลจะได้เปรียบในการตอบข้อสอบข้อที่ 9, 13 และข้อที่ 35

ซึ่งสามารถแสดงแผนภาพคลื่นความถี่ (History) ของการเกิดการทำหน้าที่ต่างกันของข้อสอบ (DIF) ของวิชาภาษาไทย จำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียนด้วยวิธี BAYESIAN ที่วิเคราะห์ด้วยโปรแกรม Open BUGS แสดงดังภาพต่อไปนี้



ภาพที่ 4-25 คลื่นความถี่ (History) การเกิดการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทย จำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน ด้วยวิธี BAYESIAN

และสามารถแสดงแผนภาพคลื่นความถี่ (Density) ของการเกิดการทำหน้าที่ต่างกันของข้อสอบ (DIF) ของวิชาภาษาไทย จำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน ด้วยวิธี BAYESIAN ที่วิเคราะห์ด้วยโปรแกรม Open BUGS ดังภาพต่อไปนี้



ภาพที่ 4-26 คลื่นความถี่ (Density) ของการเกิดการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทย จำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน ด้วยวิธี BAYESIAN

3. การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธี Multiple group CFA โดยประยุกต์ใช้โปรแกรม Mplus

การวิจัยครั้งนี้ผู้วิจัยได้ประยุกต์ใช้โปรแกรม Mplus ในการศึกษาการทำหน้าที่ต่างกันของข้อสอบ ซึ่งประกอบด้วย วิชาคณิตศาสตร์ และภาษาไทย มีจำนวนข้อสอบ 40 ข้อ และ 80 ข้อ ตามลำดับ สามารถวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธี MGCFA โดย ประยุกต์ใช้โปรแกรม Mplus ดังนี้

ตารางที่ 4-25 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์ ด้วยวิธี MGCFA จำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน

กลุ่ม	Chi-square	*df*	$\Delta\chi^2$	ΔDF	p-value
จำแนกตามเพศ					
No-DIF model	3564.05	1560	37.87	11	0.00*
DIF model	3526.18	1549			
จำแนกตามที่ตั้งทางภูมิศาสตร์ของโรงเรียน					
No-DIF	6014.92	1560	1487.56	52	0.00*
DIF model	4527.36	1508			

* $p < .05$ ** No-DIF model-ก่อนปรับโมเดล, DIF model-หลังปรับโมเดล

จากตารางที่ 4-25 พบว่า ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์ด้วยวิธี MGCFA จำแนกตามเพศ พบว่า จำนวนข้อสอบที่ทำหน้าที่ต่างกัน (DIF) จำนวน 11 ข้อ ได้แก่ ข้อ 1, 4, 6, 9, 11, 12, 16, 17, 20, 23 และข้อ 24 โดยมีค่า χ^2 ของ No-DIF model เท่ากับ 3564.05 ที่องศาอิสระ 1560 ต่อมาได้นำข้อมูลเข้าทดสอบใน DIF model (ปรับโมเดล) เพื่อทำการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันระหว่างเพศ ภายหลังปรับโมเดลแล้ว พบว่า มีค่า χ^2 หลังปรับใน DIF model เป็น 3526.18 ที่องศาอิสระ 1549 และมีค่า $\Delta\chi^2$ เท่ากับ 37.87 ค่า Δ DF เท่ากับ 11 และค่า p-value เท่ากับ 0.00 แสดงว่า No-DIF model และ DIF model มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ .05

ส่วนการตรวจสอบการทำหน้าที่ต่างกันจำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียนพบจำนวนข้อที่มีความแตกต่างกัน (DIF) จำนวน 39 ข้อ ได้แก่ ข้อ 1 ถึงข้อ 29 และข้อ 31 ถึงข้อ 40 โดยมีค่า χ^2 ก่อนปรับโมเดลใน No-DIF model เท่ากับ 6014.92 ที่องศาอิสระ 1560 ต่อมา

ได้ทำการปรับโมเดล ใน DIF model เพื่อทำการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันระหว่างสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน หลังปรับโมเดลแล้ว พบว่า มีค่า χ^2 หลังปรับโมเดลใน DIF model เป็น 4527.361 ที่องศาอิสระ 1508 และมีค่า $\Delta\chi^2$ เท่ากับ 1487.56 ค่า ΔDF เท่ากับ 52 และค่า p-value เท่ากับ .00 แสดงว่า No-DIF model และ DIF model มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ .05

ตารางที่ 4-26 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทย ด้วยวิธี MGCFA จำแนกตามตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน

กลุ่ม	Chi-square	*df*	$\Delta\chi^2$	ΔDF	p-value
จำแนกตามเพศ					
No-DIF model	39831.40	6318	-68.63	1	-
DIF model	39900.03	6317			
จำแนกตามที่ตั้งทางภูมิศาสตร์ของโรงเรียน					
No-DIF model	39472.85	6320	-21.26	1	-
DIF model	39494.11	6319			

* $p < .05$ ** No-DIF model-ก่อนปรับโมเดล, DIF model-หลังปรับโมเดล

จากตารางที่ 4-26 พบว่า ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทยด้วยวิธี MGCFA จำแนกตามเพศ ไม่พบข้อสอบที่ทำหน้าที่ต่างกันเมื่อจำแนกตามเพศ

โดยมีค่า χ^2 ของ No-DIF model (ก่อนปรับโมเดล) เท่ากับ 39831.40 ที่องศาอิสระ 6318 ต่อมานำข้อมูลเข้าทดสอบใน DIF model (ปรับโมเดล) เพื่อทำการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันระหว่างเพศ ภายหลังปรับโมเดลแล้ว พบว่า มีค่า χ^2 หลังปรับใน DIF model เป็น 39900.03 ที่องศาอิสระ 6317 และมีค่า $\Delta\chi^2$ เท่ากับ -68.63 ค่า ΔDF เท่ากับ 1 และค่า p-value ตรวจสอบไม่ได้ แสดงว่า No-DIF model และ DIF model ไม่มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ .05

ส่วนการตรวจสอบการทำหน้าที่ต่างกันจำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน ไม่พบข้อสอบที่ทำหน้าที่ต่างกันจำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของ โดยมีค่า χ^2 ก่อนปรับโมเดลใน No-DIF เท่ากับ 39472.84 ที่องศาอิสระ 6320 ต่อมาได้ทำการปรับโมเดล

ใน DIF model เพื่อทำการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันระหว่างสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน หลังปรับโมเดลแล้ว พบว่า มีค่า χ^2 หลังปรับโมเดล ใน DIF model เป็น 39494.11 ที่องศา 6319, $\Delta\chi^2$ = -21.26 ค่า ΔDF เท่ากับ 1 และค่า p-value ตรวจสอบไม่ได้ แสดงว่า No-DIF model และ DIF Model ไม่มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ .05

ตอนที่ 6 เปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ด้วยวิธี IRT-LR วิธี BAYESIAN และวิธี MGCFA จำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน

การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ (DIF) ครั้งนี้ประกอบด้วยวิชาคณิตศาสตร์และภาษาไทย มีข้อสอบจำนวน 40 ข้อ และ 80 ข้อ ตามลำดับ ซึ่งผู้วิจัยได้ทำการวิเคราะห์ด้วยวิธี IRT-LR วิธี BAYESIAN และวิธี MGCFA จำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน ดังนี้

ตารางที่ 4-27 เปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์ด้วยวิธี IRT-LR วิธี BAYESIAN และวิธี MGCFA จำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน

ข้อสอบข้อที่	เปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ (DIF)					
	จำแนกตามเพศ			จำแนกสถานที่ตั้งทางภูมิศาสตร์ฯ		
	IRT-LR	BAYESIAN	MGCFA	IRT-LR	BAYESIAN	MGCFA
1	1	0	1	0	0	1
2	0	0	0	0	0	1
3	0	0	0	1	0	1
4	0	0	1	0	0	1
5	0	0	0	1	0	1
6	0	1	1	0	0	1
7	0	0	0	0	0	1
8	0	0	0	0	0	1
9	1	0	1	0	1	1

ตารางที่ 4-27 (ต่อ)

ข้อสอบข้อที่	เปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ (DIF)					
	จำแนกตามเพศ			จำแนกสถานที่ตั้งทางภูมิศาสตร์ฯ		
	IRT-LR	BAYESIAN	MGCFA	IRT-LR	BAYESIAN	MGCFA
10	0	0	0	1	1	1
11	1	0	1	0	0	1
12	1	1	1	0	0	1
13	0	1	0	0	0	1
14	0	0	0	0	0	1
15	0	0	0	0	0	1
16	1	1	1	0	0	1
17	1	1	1	1	0	1
18	0	0	0	0	1	1
19	0	0	0	1	0	1
20	0	0	1	0	0	1
21	0	0	0	0	0	1
22	0	0	0	0	0	1
23	1	1	1	0	0	1
24	1	0	1	0	0	1
25	0	1	0	0	0	1
26	0	1	0	0	0	1
27	0	0	0	0	0	1
28	0	0	0	0	0	1
29	0	0	0	0	0	1
30	0	1	0	0	1	0
31	0	0	0	0	1	1
32	0	0	0	0	1	1
33	0	0	0	0	1	1
34	0	0	0	0	0	1

ตารางที่ 4-27 (ต่อ)

ข้อสอบ ข้อที่	เปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ (DIF)					
	จำแนกตามเพศ			จำแนกสถานที่ตั้งทางภูมิศาสตร์ฯ		
	IRT-LR	BAYESIAN	MGCFA	IRT-LR	BAYESIAN	MGCFA
35	0	0	0	0	0	1
36	0	1	0	0	0	1
37	0	1	0	0	0	1
38	0	0	0	0	1	1
39	1	1	0	1	0	1
40	0	0	0	1	1	1
รวม	9	12	11	7	9	39
ร้อยละ	22.50	30.00	27.50	17.50	22.50	97.50

*1 = DIF; 0 = No-DIF

จากตารางที่ 4-27 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) วิชาคณิตศาสตร์ จำแนกตามเพศ ด้วยวิธี IRT-LR, วิธี BAYESIAN และวิธี MGCFA พบว่า วิธี BAYESIAN พบจำนวนข้อสอบที่มีการทำหน้าที่ต่างกันมากที่สุด จำนวน 12 ข้อ ได้แก่ ข้อ 6, 12, 13, 16, 17, 23, 25, 26, 30, 36, 37 และข้อ 39 คิดเป็นร้อยละ 30 รองลงมา คือ วิธี MGCFA พบจำนวน 11 ข้อ ได้แก่ ข้อ 1, 4, 6, 9, 11, 12, 16, 17, 20, 23 และข้อ 24 คิดเป็นร้อยละ 27.50 และวิธี IRT-LR พบจำนวน 9 ข้อ ได้แก่ ข้อ 1, 9, 11, 12, 16, 17, 23, 24 และข้อ 39 คิดเป็นร้อยละ 22.50

ส่วนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) วิชาคณิตศาสตร์ จำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน พบว่า วิธี MGCFA พบจำนวนข้อสอบที่มีการทำหน้าที่ต่างกันมากที่สุด จำนวน 39 ข้อ ได้แก่ ข้อ 1 ถึงข้อ 29 และข้อ 31 ถึงข้อ40 คิดเป็นร้อยละ 97.50 รองลงมา คือ วิธี BAYESIAN พบจำนวน 9 ข้อ ได้แก่ ข้อ 9, 10, 18, 30, 31, 32, 33, 38 และข้อ 40 คิดเป็นร้อยละ 22.50 และวิธี IRT-LR พบจำนวน 7 ข้อ ได้แก่ ข้อ 3, 5, 10, 17, 19, 39 และข้อ 40 คิดเป็นร้อยละ 17.50

ตารางที่ 4-28 แสดงความสอดคล้องของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์ ด้วยวิธี IRT-LR วิธี BAYSIAN และวิธี MGCFA

วิธีการตรวจสอบ	จำนวนข้อ	จำนวนข้อที่สอดคล้อง			ข้อที่พบความสอดคล้องกัน	Φ	χ^2	*df*	*p*
		DIF	NO-DIF	รวม					
จำแนกตามเพศ									
IRT-LR & BAYESIAN	40	5 (12.5%)	24 (60.0%)	29 (72.5%)	12, 16, 17, 23, 39	.30	3.61	1	.06
IRT-LR & MGCFA	40	8 (20.0%)	28 (70.0%)	36 (90.0%)	1, 9, 11, 12, 16, 17, 23, 24	.74*	21.95*	1	.00
BAYESIAN & MGCFA	40	5 (12.5%)	22 (55.0%)	27 (67.5%)	6, 12, 16, 17, 23	.21	1.73	1	.19
จำแนกตามที่ตั้งทางภูมิศาสตร์									
IRT-LR & BAYESIAN	40	2 (5.0%)	26 (65.0%)	28 (70.0%)	10, 40	.08	.18	1	.67
IRT-LR & MGCFA	40	7 (17.5%)	1 (2.5%)	8 (20.0%)	3, 5, 10, 17, 19, 39, 40	.07	.22	1	.64
BAYESIAN & MGCFA	40	8 (20.0%)	0 (0.0%)	8 (20.0%)	9, 10, 18, 31, 32, 33, 38, 40	-.30	3.53	1	.60

* $p < .05$ หมายเหตุ: ตัวเลขในวงเล็บ หมายถึง ร้อยละของจำนวนข้อสอบที่ตรวจพบการทำหน้าที่ต่างกันของข้อสอบ (DIF) และข้อสอบที่ไม่ทำหน้าที่ต่างกัน (NO-DIF)

จากตาราง 4-28 พบว่า ความสอดคล้องของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์ จำนวน 40 ข้อ ได้ผลการวิเคราะห์ ดังนี้

เมื่อจำแนกตามเพศ วิธี IRT-LR เปรียบเทียบกับวิธี MGCFA พบว่า ทั้งสองวิธีตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน (DIF) ได้ตรงกันจำนวน 8 ข้อ ได้แก่ ข้อที่ 1, 9, 11, 12, 16, 17, 23 และข้อ 24 คิดเป็นร้อยละ 20 โดยมีค่า χ^2 เท่ากับ 21.95, Φ = .74 *df* = 1 และ p-value = .00 สรุปว่าทั้งสองวิธีความสอดคล้องกันอย่างมีนัยสำคัญทางสถิติที่ .05, วิธี IRT-LR เปรียบเทียบกับวิธี BAYESIAN พบว่า ทั้งสองวิธีตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน (DIF) ได้ตรงกันจำนวน 5 ข้อ ได้แก่ ข้อที่ 12, 16, 17, 23 และข้อ 39 คิดเป็นร้อยละ 12.5 โดยมีค่า χ^2 เท่ากับ 3.61, Φ = .30, *df* = 1 และ p-value = .06 สรุปว่าทั้งสองวิธีไม่มีความสอดคล้องกัน และวิธี BAYESIAN เปรียบเทียบกับวิธี MGCFA พบว่า ทั้งสองวิธีตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน (DIF) ได้ตรงกันจำนวน 5 ข้อ ได้แก่ ข้อ 6, 12, 16, 17 และข้อ 23 คิดเป็นร้อยละ 12.5 โดยมีค่า χ^2 เท่ากับ 1.73, Φ = .21, *df* = 1 และ p-value = .19 สรุปว่าทั้งสองวิธีไม่มีความสอดคล้องกัน

เมื่อจำแนกตามเขตที่ตั้งทางภูมิศาสตร์ของโรงเรียน พบว่า วิธี IRT-LR เปรียบเทียบกับวิธี BAYESIAN พบว่า ทั้งสองวิธีตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน (DIF) ได้ตรงกันจำนวน 2 ข้อ ได้แก่ ข้อ 10 และข้อ 40 คิดเป็นร้อยละ 5.0 โดยมีค่า χ^2 เท่ากับ .18, Φ = .08 *df* =1 และ p-value = .67 สรุปว่าทั้งสองวิธีไม่มีความสอดคล้องกัน, วิธี IRT-LR เปรียบเทียบกับวิธี MGCFA พบว่า ทั้งสองวิธีตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน (DIF) ได้ตรงกันจำนวน 7 ข้อ ได้แก่ ข้อที่ 3, 5, 10, 17, 19, 39 และข้อ 40 คิดเป็นร้อยละ 17.5 โดยมีค่า χ^2 เท่ากับ .22, Φ = .07, *df* = 1 และ p-value = .64 สรุปว่าทั้งสองวิธีไม่มีความสอดคล้องกัน และ วิธี BAYESIAN เปรียบเทียบกับวิธี MGCFA พบว่า ทั้งสองวิธีตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน (DIF) ได้ตรงกันจำนวน 8 ข้อ ได้แก่ ข้อ 9, 10, 18, 31, 32, 33, 38 และข้อ 40 คิดเป็นร้อยละ 20 โดยมีค่า χ^2 เท่ากับ 3.53, Φ = -.30 *df* = 1 และ p-value = .60 สรุปว่าทั้งสองวิธีไม่มีความสอดคล้องกัน

ตารางที่ 4-29 เปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทย ด้วยวิธี IRT-LR วิธี BAYESIAN และวิธี MGCFA จำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน

ข้อสอบข้อที่	เปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ (DIF)					
	จำแนกตามเพศ			จำแนกตามภูมิศาสตร์ฯ		
	IRT-LR	BAYESIAN	MGCFA	IRT-LR	BAYESIAN	MGCFA
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	1	0	0	0	0
6	0	0	0	0	0	0
7	0	1	0	0	0	0
8	0	0	0	0	0	0
9	0	1	0	0	1	0
10	0	1	0	0	1	0
11	0	1	0	0	0	0
12	0	0	0	0	0	0
13	0	1	0	0	1	0
14	0	0	0	0	0	0
15	0	0	0	0	0	0
16	0	0	0	0	0	0
17	0	0	0	0	0	0
18	0	0	0	0	0	0
19	0	0	0	0	0	0
20	0	0	0	0	0	0
21	0	0	0	0	0	0
22	0	1	0	0	0	0
23	0	0	0	0	1	0

ตารางที่ 4-29 (ต่อ)

ข้อสอบข้อที่	เปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ (DIF)					
	จำแนกตามเพศ			จำแนกตามภูมิศาสตร์ฯ		
	IRT-LR	BAYESIAN	MGCFA	IRT-LR	BAYESIAN	MGCFA
24	0	0	0	0	0	0
25	0	0	0	1	1	0
26	0	1	0	0	1	0
27	0	0	0	0	0	0
28	0	1	0	0	0	0
29	0	0	0	0	0	0
30	0	0	0	0	0	0
31	0	0	0	0	1	0
32	0	1	0	0	0	0
33	0	0	0	0	0	0
34	0	1	0	0	0	0
35	0	0	0	1	1	0
36	0	0	0	1	1	0
37	0	0	0	0	0	0
38	1	0	0	0	0	0
39	0	0	0	0	0	0
40	0	1	0	0	0	0
41	0	0	0	0	1	0
42	0	0	0	0	0	0
43	0	0	0	0	0	0
44	0	1	0	0	0	0
45	0	0	0	0	0	0
46	0	1	0	0	0	0
47	0	1	0	0	1	0
48	0	0	0	1	1	0

ตารางที่ 4-29 (ต่อ)

ข้อสอบข้อที่	เปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ (DIF)					
	จำแนกตามเพศ			จำแนกตามภูมิศาสตร์ฯ		
	IRT-LR	BAYESIAN	MGCFA	IRT-LR	BAYESIAN	MGCFA
49	0	0	0	0	0	0
50	0	0	0	0	0	0
51	0	1	0	0	0	0
52	0	0	0	0	0	0
53	0	0	0	0	0	0
54	0	0	0	0	0	0
55	0	0	0	1	0	0
56	0	0	0	0	0	0
57	0	0	0	1	1	0
58	0	1	0	0	0	0
59	0	0	0	0	0	0
60	1	1	0	0	1	0
61	0	0	0	0	1	0
62	0	1	0	0	1	0
63	0	0	0	1	0	0
64	0	0	0	0	0	0
65	0	0	0	0	0	0
66	0	1	0	0	1	0
67	0	1	0	0	1	0
68	0	1	0	0	0	0
69	0	1	0	0	1	0
70	0	1	0	0	1	0
71	0	1	0	0	0	0
72	1	1	0	0	0	0
73	0	0	0	0	0	0

ตารางที่ 4-29 (ต่อ)

ข้อสอบ ข้อที่	เปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ (DIF)					
	จำแนกตามเพศ			จำแนกตามภูมิศาสตร์ฯ		
	IRT-LR	BAYESIAN	MGCFA	IRT-LR	BAYESIAN	MGCFA
74	0	0	0	0	1	0
75	0	0	0	0	0	0
76	0	0	0	0	1	0
77	0	1	0	0	0	0
78	0	1	0	0	1	0
79	0	1	0	0	1	0
80	0	0	0	0	0	0
รวม	3	29	0	7	24	0
ร้อยละ	3.75	36.25	0	8.75	30	0

*1 = DIF; 0 = No-DIF

จากตารางที่ 4-29 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) วิชาภาษาไทย จำแนกตามเพศของโรงเรียน ด้วยวิธี IRT-LR วิธี BAYESIAN และวิธี MGCFA พบว่า วิธี BAYESIAN พบจำนวนข้อสอบที่มีการทำหน้าที่ต่างกันมากที่สุด จำนวน 29 ข้อ ได้แก่ ข้อ 5, 7, 9, 10, 11, 13, 22, 26, 28, 32, 34, 40, 44, 46, 47, 51, 58, 60, 62, 66, 67, 68, 69, 70, 71, 72, 77, 78, และข้อ 79 คิดเป็นร้อยละ 36.25 รองลงมา คือ วิธี IRT-LR พบจำนวน 3 ข้อ ได้แก่ ข้อ 38, 60 และข้อ 72 คิดเป็นร้อยละ 3.75 และวิธี MGCFA ไม่พบการทำหน้าที่ต่างกันของข้อสอบ คิดเป็นร้อยละ 0

ส่วนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) วิชาภาษาไทย จำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน พบว่า วิธี BAYESIAN พบจำนวนข้อสอบที่มีการทำหน้าที่ต่างกันมากที่สุด จำนวน 24 ข้อ ได้แก่ ข้อ 9, 10, 13, 23, 25, 26, 30, 35, 36, 41, 47, 48, 57, 60, 61, 62, 66, 67, 69, 70, 74, 76, 78 และข้อ 79 คิดเป็นร้อยละ 30 รองลงมา คือ วิธี IRT-LR พบจำนวน 7 ข้อ ได้แก่ ข้อ 25, 35, 36, 48, 55, 57 และข้อ 63 คิดเป็นร้อยละ 8.75 และวิธี MGCFA ไม่พบการทำหน้าที่ต่างกันของข้อสอบ คิดเป็นร้อยละ 0

ตารางที่ 4-30 แสดงความสอดคล้องของการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทย ด้วยวิธี IRT-LR, BAYSIAN และ MGCFA

วิธีการตรวจสอบ	จำนวนข้อ	จำนวนข้อที่สอดคล้อง			ข้อที่พบความสอดคล้องกัน	Φ	χ^2	*df*	*p*
		DIF	NO-DIF	รวม					
จำแนกตามเพศ									
IRT-LR & BAYESIAN	40	2 (2.50%)	50 (62.50%)	52 (65.00%)	60, 72	1.25	.13	1	.26
IRT-LR & MGCFA	40	0 (0%)	77 (96.30%)	77 (96.30%)	-	-	-	-	-
BAYESIAN & MGCFA	40	0 (0%)	51 (63.70%)	51 (63.70%)	-	-	-	-	-
จำแนกตามที่ตั้งทางภูมิศาสตร์									
IRT-LR & BAYESIAN	40	5 (6.30%)	54 (67.50%)	59 (73.80%)	25, 35, 36, 48, 57	.28*	6.27*	1	.01
IRT-LR & MGCFA	40	0 (0%)	73 (91.30%)	73 (91.30%)	-	-	-		-
BAYESIAN & MGCFA	40	0 (0%)	56 (70.00%)	56 (70.00%)	-	-			-

* $p < .05$ <u>หมายเหตุ</u>: ตัวเลขในวงเล็บ หมายถึง ร้อยละของจำนวนข้อสอบที่ตรวจพบการทำหน้าที่ต่างกันของข้อสอบ (DIF) และข้อสอบที่ไม่ทำหน้าที่ต่างกัน (NO-DIF)

จากตาราง 4-30 พบว่า ความสอดคล้องของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทย จำนวน 80 ข้อ ได้ผลการวิเคราะห์ ดังนี้

เมื่อจำแนกตามเพศ วิธี IRT-LR เปรียบเทียบกับวิธี BAYESIAN พบว่า ทั้งสองวิธีตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน (DIF) ได้ตรงกัน จำนวน 2 ข้อ ได้แก่ ข้อที่ 60 และข้อ 72 คิดเป็นร้อยละ 2.5 โดยมีค่า χ^2 เท่ากับ 1.25, Φ = .13 *df* = 1 และ p-value = .26 สรุปว่าทั้งสองวิธีไม่มีความสอดคล้องกัน ส่วนวิธี MGCFA ไม่สามารถหาความสอดคล้องได้ เนื่องจากวิธี MGCFA ตรวจไม่พบข้อสอบที่ทำหน้าที่ต่างกัน

เมื่อจำแนกตามเขตที่ตั้งทางภูมิศาสตร์ของโรงเรียน พบว่า วิธี IRT-LR เปรียบเทียบกับวิธี BAYESIAN พบว่า ทั้งสองวิธีตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน (DIF) ได้ตรงกันจำนวน 5 ข้อ ได้แก่ ข้อ 25, 35, 36, 48 และข้อ 57 คิดเป็นร้อยละ 6.3 โดยมีค่า χ^2 เท่ากับ 6.27, Φ = .28 *df* = 1 และ p-value = .01 สรุปว่าทั้งสองวิธีมีความสอดคล้องกัน, ส่วนวิธี MGCFA ไม่สามารถหาความสอดคล้องได้ เนื่องจากวิธี MGCFA ตรวจไม่พบข้อสอบที่ทำหน้าที่ต่างกัน

บทที่ 5
สรุป อภิปรายผล และข้อเสนอแนะ

การวิจัยครั้งนี้มีวัตถุประสงค์ คือ 1) เพื่อศึกษาผลการประมาณค่าอำนาจจำแนก (a) ของข้อสอบ, ค่าพารามิเตอร์ความยาก (b) ของข้อสอบ, ค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) ด้วยวิธี Maximum likelihood วิธี BAYESIAN และวิธี Confirmatory factor analysis 2) เพื่อเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ (DIF) จำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน ได้แก่ เขตกรุงเทพมหานครและปริมณฑล และนอกเขตกรุงเทพมหานครและปริมณฑล ด้วยวิธี IRT-LR วิธี BAYESIAN และวิธี MGCFA

ประชากรที่ใช้ในการวิจัยครั้งนี้ เป็นนักเรียนชั้นมัธยมศึกษาปีที่ 6 ในสังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน กระทรวงศึกษาธิการ ที่เข้าสอบวัดผลสัมฤทธิ์ทางการเรียนเพื่อประเมินคุณภาพการศึกษาระดับชาติ ปีการศึกษา 2558 ซึ่งนักเรียนที่เข้าสอบวิชาคณิตศาสตร์และภาษาไทย จำนวนทั้งสิ้น 358,512 คน จำแนกตามเพศชาย จำนวน 304,751 คน และเพศหญิง 285,412 คน ในการกำหนดกลุ่มตัวอย่างสำหรับการวิจัย ผู้วิจัยได้ศึกษาการกำหนดขนาดกลุ่มตัวอย่างจากประชากรตามการวิเคราะห์คุณภาพตามทฤษฎีการทดสอบการตอบสนองข้อสอบ (Item response theory: IRT) และการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (Differential item function: DIF) ซึ่งตามเงื่อนไขแล้วกลุ่มตัวอย่างไม่ต่ำกว่า 10-20 เท่า ของจำนวนข้อคำถาม (นงลักษณ์ วิรัชชัย, 2542, หน้า 311; สุวิมล ติรกานันท์, 2553, หน้า 168; สุภมาศ อังศุโชติ และคณะ, 2554, หน้า 31, หน้า 109 อ้างถึงใน สุพัฒนา หอมบุบผา, 2556, หน้า 109)

จากการศึกษาของ เมเซอร์และคณะ (Mazor et al., 1992, pp. 443-451 อ้างถึงใน ปิยะทิพย์ ตินวร, 2549, หน้า 49) พบว่า เมื่อใช้กลุ่มตัวอย่างขนาดใหญ่ทำให้ตรวจพบข้อสอบทำหน้าที่ต่างกันได้มากกว่าการใช้กลุ่มตัวอย่างขนาดเล็ก คือ เมื่อใช้กลุ่มตัวอย่าง 2,000 คน มีความถูกต้องในการตรวจสอบ ร้อยละ 70 ถึง 75 และผลการศึกษาของจิติมา วรรณศรี (2539) พบว่า เมื่อใช้กลุ่มอ้างอิง 1,000 คน และกลุ่มเปรียบเทียบ 1,000 คน สามารถตรวจพบข้อสอบทำหน้าที่ต่างกันได้ถูกต้องร้อยละ 100 และอัตราความคลาดเคลื่อนอยู่ในเกณฑ์ที่เหมาะสม

จากจำนวนข้อสอบ O-NET มัธยมศึกษาปีที่ 6 ประจำปีการศึกษา 2558 มีจำนวนดังนี้ คือ วิชาคณิตศาสตร์ จำนวน 40 ข้อ และภาษาไทย จำนวน 80 ข้อ รวมทั้งสิ้น 120 ข้อ ดังนั้นในการศึกษาครั้งนี้ผู้วิจัยจึงเลือกใช้จำนวนกลุ่มตัวอย่างเป็น 20 เท่าของจำนวนข้อสอบ โดยจะใช้กลุ่มตัวอย่างจำนวนทั้งสิ้น 2,400 คน

การวิเคราะห์ข้อมูล เริ่มจากการสร้างไฟล์ข้อมูลสำหรับการวิเคราะห์ด้วยโปรแกรม IRT PRO โปรแกรม Open BUGS และโปรแกรม Mplus เพื่อให้ได้ข้อมูลที่มีความเหมาะสมสำหรับการวิเคราะห์ในแต่ละโปรแกรม แล้วดำเนินการวิเคราะห์ข้อมูลเพื่อตอบคำถามการวิจัย ดังนี้

1. ผลการประมาณค่าพารามิเตอร์อำนาจจำแนก (a), พารามิเตอร์ความยาก (b) ของข้อสอบ และพารามิเตอร์ความสามารถของผู้สอบ (θ) ด้วยวิธี Maximum likelihood, วิธี BAYESIAN และวิธี CFA

2. การทำหน้าที่ต่างกันของข้อสอบ (DIF) จำแนกตามเพศ สถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน ได้แก่ เขตกรุงเทพมหานครและปริมณฑล และนอกเขตกรุงเทพมหานครและปริมณฑล ด้วยวิธี IRT-LR วิธี BAYESIAN และวิธี MGCFA มีความสอดคล้องกันหรือไม่

สรุปผลการวิจัย

จากการวิเคราะห์ค่าพารามิเตอร์อำนาจจำแนก (a), ค่าพารามิเตอร์ความยาก(b) ของข้อสอบ และค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) ครั้งนี้ ประกอบด้วย วิชาคณิตศาสตร์ และภาษาไทย ซึ่งผู้วิจัยได้ทำการวิเคราะห์ด้วยวิธี ML ด้วยโปรแกรม IRT PRO version 2.1 วิธี BAYESIAN ด้วยโปรแกรม Open BUGS version 3.2.1 และวิธี CFA ด้วยโปรแกรม Mplus version 7.4 แล้วนำผลการวิเคราะห์จากทั้งสามวิธีมาทำการเปรียบเทียบกันเพื่อตอบคำถามการวิจัย ได้ผล ดังนี้

1. ผลการประมาณค่าพารามิเตอร์อำนาจจำแนก (a), ค่าพารามิเตอร์ความยาก (b) ของข้อสอบ และค่าพารามิเตอร์ความสามารถของผู้สอบ (θ)

1.1 ผลการประมาณค่าพารามิเตอร์อำนาจจำแนกของข้อสอบ

ผลประมาณค่าพารามิเตอร์อำนาจจำแนก (a) ของข้อสอบ วิชาคณิตศาสตร์ และภาษาไทย ด้วยวิธี ML วิธี Bayesian และวิธี CFA วิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน พบว่า มีความสัมพันธ์ทางบวกในระดับสูงมาก (r = 1.00) และมีนัยสำคัญทางสถิติที่ระดับ .05

1.2 ผลการประมาณค่าพารามิเตอร์ความยากของข้อสอบ

ผลการประมาณค่าพารามิเตอร์ความยาก (b) ของข้อสอบ วิชาคณิตศาสตร์ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA วิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน พบว่า วิธี ML และวิธี CFA มีความสัมพันธ์ทางบวกในระดับสูงมาก (r = 1.00) และมีนัยสำคัญทางสถิติที่ระดับ .05 ผลการประมาณค่าพารามิเตอร์ความยากของข้อสอบระหว่างวิธี BAYESIAN กับวิธี ML และวิธี BAYESIAN กับวิธี CFA มีความสัมพันธ์ทางบวกในระดับปานกลางค่อนข้างน้อย

(r = .43) และมีนัยสำคัญทางสถิติที่ระดับ .05 ส่วนผลการประมาณค่าพารามิเตอร์ความยากของข้อสอบวิชาภาษาไทยระหว่างวิธี ML วิธี BAYESIAN และวิธี CFA พบว่า มีความสัมพันธ์ทางบวกในระดับสูงมาก (r = 1.00) และมีนัยสำคัญทางสถิติที่ระดับ .05

1.3 ผลการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบ

ผลการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) วิชาคณิตศาสตร์ และภาษาไทย ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA วิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน พบว่า วิธี BAYESIAN และวิธี CFA มีความสัมพันธ์ทางบวกในระดับสูงมาก (r = 1.00) และมีนัยสำคัญทางสถิติที่ระดับ .05 ผลการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบด้วยวิธี ML กับวิธี BAYESIAN และวิธี ML กับวิธี CFA มีความสัมพันธ์ทางบวกในระดับสูง (r = .99) และมีนัยสำคัญทางสถิติที่ระดับ .05

2. เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) จำแนกตามเพศและสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน

การศึกษาการทำหน้าที่ต่างกันของข้อสอบในครั้งนี้ ประกอบด้วย วิชาคณิตศาสตร์ จำนวน 40 ข้อ และภาษาไทย จำนวน 80 ข้อ รวมทั้งหมด 120 ข้อ ผู้วิจัยวิเคราะห์ข้อสอบด้วยวิธี IRT-LR วิธี BAYESIAN และวิธี MGCFA จำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน โดยประยุกต์ใช้โปรแกรม IRT-PRO โปรแกรม Open BUGS และโปรแกรม Mplus ตามลำดับ แล้วนำผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ในแต่ละวิธีมาวิเคราะห์หาความสอดคล้องของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยการวิเคราะห์ตารางไขว้ (Cross tabulation table) ได้ผลการวิเคราะห์ ดังนี้

2.1 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) จำแนกตามเพศและสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน วิชาคณิตศาสตร์

ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) วิชาคณิตศาสตร์ จากการวิเคราะห์ข้อสอบจำนวน 40 ข้อ ด้วยวิธี IRT-LR, BAYESIAN และวิธี MGCFA พบว่า วิธี BAYESIAN พบจำนวนข้อสอบที่มีการทำหน้าที่ต่างกันมากที่สุด จำนวน 12 ข้อ ได้แก่ ข้อ 6, 12, 13, 16, 17, 23, 25, 26, 30, 36, 37 และข้อ 39 คิดเป็นร้อยละ 30 รองลงมา คือ วิธี MGCFA พบจำนวน 11 ข้อ ได้แก่ ข้อ 1, 4, 6, 9, 11, 12, 16, 17, 20, 23 และข้อ 24 คิดเป็นร้อยละ 27.50 และวิธี IRT-LR พบจำนวน 9 ข้อ ได้แก่ ข้อ 1 9, 11, 12, 16, 17, 23, 24 และข้อ 39 คิดเป็นร้อยละ 22.50 และเมื่อตรวจสอบความสอดคล้องด้วยการสร้างตารางไขว้ (Cross tabulation table) พบว่า

ตารางที่ 5-1 แสดงความสอดคล้องของการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์ ด้วยวิธี IRT-LR วิธี BAYSIAN และวิธี MGCFA

วิธีตรวจสอบ DIF	**จำนวนข้อที่พบ DIF**	**ข้อที่พบ DIF ตรงกัน**	χ^2	**Φ**	*df*	*p*
จำแนกตามเพศ						
วิธี IRT-LR เปรียบเทียบ วิธี BAYESIAN	5 (12.5%)	12, 16, 17, 23, 39	3.61	.30	1	.06
วิธี IRT-LR เปรียบเทียบ วิธี MGCFA	8 (20%)	1, 9, 11, 12, 16, 17, 23, 24	21.95*	.74*	1	.00
วิธี BAYESIAN เปรียบเทียบ วิธี MGCFA	5 (12.5%)	6, 12, 16, 17, 23	1.73	.21	1	.19
จำแนกตามสถานที่ตั้งทางภูมิศาสตร์						
วิธี IRT-LR เปรียบเทียบ วิธี BAYESIAN	2 (5.0%)	10, 40	.18	.08	1	.67
วิธี IRT-LR เปรียบเทียบ วิธี MGCFA	7 (17.5%)	3, 5, 10, 17, 19, 39, 40	.22	.07	1	.64
วิธี BAYESIAN เปรียบเทียบ วิธี MGCFA	8 (20.0%)	9, 10, 18, 31, 32, 33, 38, 40	3.53	-.30	1	.60

* $p < .05$

สรุปได้ว่า เมื่อจำแนกตามเพศ วิธี IRT-LR เปรียบเทียบ วิธี MGCFA พบว่า มีความสอดคล้องกันอย่างมีนัยทางสถิติที่ .05 โดยมีค่า χ^2 เท่ากับ 21.95 Φ = .74 *df* =1, p-value = .00 ส่วนวิธี IRT-LR เปรียบเทียบกับ วิธี BAYESIAN และวิธี BAYESIAN เปรียบเทียบกับวิธี MGCFA พบว่า ไม่มีความสอดคล้องกัน เมื่อจำแนกตามเขตที่ตั้งทางภูมิศาสตร์ของโรงเรียน พบว่า ผลการเปรียบเทียบระหว่างวิธี IRT-LR วิธี BAYESIAN และวิธี MGCFA ไม่มีความสอดคล้องกัน

ส่วนผลตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) วิชาภาษาไทยจากการวิเคราะห์ข้อสอบจำนวน 80 ข้อ ด้วยวิธี IRT-LR, วิธี BAYESIAN และวิธี MGCFA พบว่า วิธี BAYESIAN พบจำนวนข้อสอบที่มีการทำหน้าที่ต่างกันมากที่สุด จำนวน 29 ข้อ ได้แก่ ข้อ 5, 7, 9, 10, 11, 13, 22, 26, 28, 32, 34, 40, 44, 46, 47, 51, 58, 60, 62, 66, 67, 68, 69, 70, 71, 72, 77, 78, และข้อ 79 คิดเป็นร้อยละ 36.25 รองลงมา คือวิธี IRT-LR พบจำนวน 3 ข้อ ได้แก่ ข้อ 38, 60 และข้อ 72 คิดเป็นร้อยละ 3.75 และวิธี MGCFA ไม่พบการทำหน้าที่ต่างกันของข้อสอบ คิดเป็นร้อยละ 0 และเมื่อตรวจสอบความสอดคล้องด้วยการสร้างตารางไขว้ (Cross tabulation table) พบว่า

ตารางที่ 5-2 แสดงความสอดคล้องของการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทย ด้วยวิธี IRT-LR วิธี BAYSIAN และวิธี MGCFA

วิธีตรวจสอบ DIF	**จำนวนข้อที่พบ DIF**	**ข้อที่พบ DIF ตรงกัน**	χ^2	Φ	*df*	*p*
จำแนกตามเพศ						
วิธี IRT-LR เปรียบเทียบวิธี BAYESIAN	2 (2.5%)	60, 72	1.25	.13	1	.26
วิธี IRT-LR เปรียบเทียบวิธี MGCFA	0 (0%)					
วิธี BAYESIAN เปรียบเทียบวิธี MGCFA	0 (0%)	-	-	-	-	-
จำแนกตามสถานที่ตั้งทางภูมิศาสตร์						
วิธี IRT-LR เปรียบเทียบวิธี MGCFA	5 (6.3%)	25, 35, 36, 48, 57	6.27*	.28*	1	.01
วิธี IRT-LR เปรียบเทียบวิธี BAYESIAN	0 (0%)	-	-	-	-	-
วิธี BAYESIAN เปรียบเทียบวิธี MGCFA	0 (0%)	-	-	-	-	-

* $p < .05$

สรุปได้ว่าเมื่อจำแนกตามเพศ พบว่า ผลการเปรียบเทียบระหว่างวิธี IRT-LR, วิธี BAYESIAN และวิธี MGCFA ไม่มีความสอดคล้องกัน และพบว่า วิธี MGCFA ตรวจไม่พบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ และเมื่อจำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน พบว่า ผลการเปรียบเทียบระหว่างวิธี IRT-LR กับวิธี BAYESIAN มีความสอดคล้องกันอย่างมีนัยทางสถิติที่ .05 โดยมีค่า χ^2 เท่ากับ 6.27 Φ = .28 *df* = 1, p-value = .01

โดยสรุป ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์ ระหว่างวิธี IRT-LR, วิธี BAYESIAN และวิธี MGCFA จำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน พบว่า เมื่อจำแนกตามเพศ วิธี BAYESIAN จะพบข้อสอบที่ทำหน้าที่ต่างกันได้มากที่สุด เมื่อจำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน วิธี MGCFA พบข้อสอบที่ทำหน้าที่ต่างกันได้มากที่สุด ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทย จำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน พบว่า เมื่อจำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน วิธี BAYESIAN จะพบข้อสอบที่ทำหน้าที่ต่างกันได้มากที่สุด และวิธี MGCFA ไม่พบการทำหน้าที่ต่างกันของข้อสอบ

โดยสรุปผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์ และวิชาภาษาไทย ด้วยวิธี IRT-LR, BAYESIAN และวิธี MGCFA พบการทำหน้าที่ต่างกันของข้อสอบดังนี้

วิชาคณิตศาสตร์

จำแนกตามเพศ

วิธีการตรวจสอบ	จำนวนข้อที่พบ DIF	ข้อสอบที่พบ DIF	ข้อสอบที่พบ DIF ตรงกันทั้ง 3 วิธี
IRT-LR	9	1, 9, 11, 12, 16, 17, 23, 24 และข้อ 39	12, 16, 17, 23
BAYESIAN	12	6, 12, 13, 16, 17, 23, 25, 26, 30, 36, 37 และข้อ 39	
MGCFA	11	1, 4, 6, 9, 11, 12, 16, 17, 20, 23 และ 24	

จำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน

วิธีการตรวจสอบ	จำนวนข้อที่พบ DIF	ข้อสอบที่พบ DIF	ข้อสอบที่พบ DIF ตรงกันทั้ง 3 วิธี
IRT-LR	7	3, 5, 10, 17, 19, 39 และข้อ 40	10, 40
BAYESIAN	9	9, 10, 18, 30, 31, 32, 33, 38 และข้อ 40	
MGCFA	39	1-29 และข้อ 31-40	

วิชาภาษาไทย

จำแนกตามเพศ

วิธีการตรวจสอบ	จำนวนข้อที่พบ DIF	ข้อสอบที่พบ DIF	ข้อสอบที่พบ DIF ตรงกันทั้ง 3 วิธี
IRT-LR	3	38, 60 และข้อ 72	0 ข้อ
BAYESIAN	29	5, 7, 9, 10, 11, 13, 22, 26, 28, 32, 34, 40, 46, 47, 51, 58, 60, 62, 66, 67, 68, 69, 70, 71, 72, 77, 78 และข้อ 79	
MGCFA	0	-	

จำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน

วิธีการตรวจสอบ	จำนวนข้อที่พบ DIF	ข้อสอบที่พบ DIF	ข้อสอบที่พบ DIF ตรงกันทั้ง 3 วิธี
IRT-LR	7	25, 35, 36, 48, 55, 57 และข้อ 63	0 ข้อ
BAYESIAN	24	9, 10, 13, 23, 25, 26, 30, 35, 36, 41, 47, 48, 57, 60, 61, 62, 66, 67, 69, 70, 74, 76, 78 และ 79	
MGCFA	0	-	

อภิปรายผลการวิจัย

1. ผลการประมาณค่าอำนาจจำแนก ค่าพารามิเตอร์ความยากของข้อสอบ และค่าพารามิเตอร์ความสามารถของผู้สอบ

ผลการประมาณค่าพารามิเตอร์อำนาจจำแนกของข้อสอบ และค่าพารามิเตอร์ความสามารถของผู้สอบวิชาคณิตศาสตร์ และวิชาภาษาไทย พบว่า วิธี Maximum likelihood, วิธี Bayesian และวิธี CFA สามารถประมาณค่าพารามิเตอร์ได้ใกล้เคียงกันมาก และมีความสัมพันธ์ทางบวกในระดับสูง และมีนัยสำคัญทางสถิติที่ .05 ทั้งนี้อาจเป็นเพราะวิธี Maximum likelihood, วิธี Bayesian และวิธี CFA วิเคราะห์ค่าพารามิเตอร์บนพื้นฐานของการวิเคราะห์ข้อสอบตามทฤษฎีการตอบสนองข้อสอบ (IRT) 2 PL parameter logistic measurement model ด้วยสมการเดียวกัน จึงทำให้ได้ผลการประมาณค่าพารามิเตอร์ได้ใกล้เคียงกันมาก และมีความสัมพันธ์กันในระดับสูง

ส่วนการประมาณค่าพารามิเตอร์ความยากของข้อสอบ พบว่า วิธี Maximum likelihood และวิธี CFA สามารถประมาณค่าพารามิเตอร์ได้ใกล้เคียงกันมาก และมีความสัมพันธ์ทางบวกในระดับสูง (r = 1.00) และมีนัยสำคัญทางสถิติที่ 0.05 ส่วนวิธี Bayesian พบว่า สามารถประมาณค่าพารามิเตอร์ความยากของข้อสอบได้สูงกว่าวิธี Maximum likelihood และวิธี CFA และมีความสัมพันธ์ทางบวกในระดับปานกลางค่อนข้างน้อยกับวิธี Maximum likelihood (r = 0.43) และวิธี CFA (r = 0.43) ซึ่งสอดคล้องกับผลการศึกษาของอริสพา เตหลิ่ม (2559) พบว่า วิธี Bayesian สามารถประมาณค่าพารามิเตอร์ความยากของข้อสอบได้ดี เมื่อข้อมูลมีการแจกแจงความสามารถเป็นแบบเบ้ซ้าย และวิธี Maximum likelihood (ML) จะประมาณค่าพารามิเตอร์ความยากของข้อสอบได้ดี เมื่อข้อมูลมีการแจกแจงความสามารถเป็นแบบเบ้ขวา ซึ่งผลการแจกแจงความสามารถด้วยวิธี Bayesian พบว่า มีการแจกแจงแบบเบ้ซ้าย (SK = -0.546) ส่วนวิธี Maximum likelihood (SK = -3.77) และ CFA (SK = -3.77) มีการแจกแจงแบบเบ้ซ้าย ซึ่งอาจเป็นสาเหตุที่ทำให้วิธี Bayesian สามารถประมาณค่าพารามิเตอร์ความยากได้สูงกว่าวิธีอื่น และสอดคล้องกับศาสตริยา ฉ่ำสมัย (2548, หน้า 78) ได้ศึกษาเปรียบเทียบการอนุมานค่าเฉลี่ยประชากรระหว่างวิธีภาวะน่าจะเป็น และวิธีการแบบเบส์เซียน พบว่า ในการอนุมานค่าเฉลี่ยกรณีประชากรมีการแจกแจงแบบปกติ และแบบไม่ปกติควรเลือกใช้วิธีเบส์เซียน ทุกขนาดตัวอย่าง เนื่องจากการประมาณค่าเฉลี่ยใกล้เคียงกับค่าเฉลี่ยที่แท้จริง การประมาณค่าแบบช่วงที่ระดับความเชื่อมั่น 95% และ 99% ได้ขนาดความกว้างช่วงเฉลี่ยแคบ และการทดสอบสมมติฐานที่ระดับนัยสำคัญ 0.01 และ 0.05 สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดี และอำนาจการทดสอบสูงกว่าวิธีการภาวะน่าจะเป็น

2. เปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ (DIF) จำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน ระหว่างวิธี IRT-likelihood ratio วิธี Bayesian และวิธี Multiple group CFA

2.1 เปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ (DIF) วิชาคณิตศาสตร์ จำแนกตามเพศ พบว่า วิธี Bayesian พบการทำหน้าที่ต่างกันของข้อสอบมากที่สุด และพบว่า วิธี

IRT-likelihood ratio เปรียบเทียบ วิธี Multiple group CFA มีความสอดคล้องกันอย่างมีนัยทางสถิติที่ 0.05 โดยมีค่า χ^2 เท่ากับ 21.95, Φ = 0.74, *df* = 1 และ p-value = 0.00 ส่วนวิธี IRT likelihood ratio เปรียบเทียบ วิธี Bayesian และวิธี Bayesian เปรียบเทียบวิธี Multiple group CFA พบว่า ไม่มีความสอดคล้องกัน เมื่อจำแนกตามสถานที่ตั้งทางภูมิศาสตร์ วิธี Multiple group CFA พบการทำหน้าที่ต่างกันของข้อสอบมากที่สุด และทั้งสามวิธีไม่มีความสอดคล้องกัน

ทั้งนี้เนื่องจากการวิเคราะห์ด้วยทฤษฎีการตอบสนองข้อสอบ (IRT) โมเดลจะต้องมีความเป็นเอกมิติ (Unidimensional IRT model) ซึ่งตามทฤษฎีการตอบสนองข้อสอบ IRT โมเดลสามารถวิเคราะห์คุณลักษณะของบุคคลได้ครั้งละหนึ่งองค์ประกอบเท่านั้น ดังนั้น ผู้วิจัยจึงได้ทำการตรวจสอบเพิ่มเติมว่าโครงสร้างของคุณลักษณะที่ต้องการวัดมีลักษณะความเป็นเอกมิติหรือไม่ โดยทำการสกัดองค์ประกอบ (Extraction factor analysis) ด้วยการหมุนแกนแบบออโธกอนอล (Orthogonal) วิธีวาริแมกซ์ (Varimax) แล้วตรวจสอบโครงสร้างของข้อมูลเชิงประจักษ์ของวิชาคณิตศาสตร์ จำแนกตามเพศและสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน โดยใช้เกณฑ์การตรวจแบบแนวตั้ง Morizot, Ainsworth and Reise (2007 อ้างถึงใน ชัยวิชิต เชียรชัยชนะ, 2552, หน้า 14) ได้เสนอให้พิจารณาความเป็นเอกมิติ (Unidimensionality) จากค่าอัตราส่วนระหว่างค่า Eigen องค์ประกอบแรกต่อค่า Eigen องค์ประกอบที่สอง หากพบว่ามีค่ามากกว่าหรือเท่ากับ 3.00 จะบ่งบอกถึงความเป็นเอกมิติของโครงสร้างของคุณลักษณะที่ต้องการวัด ดังนี้

ตารางที่ 5-3 วิเคราะห์ความเป็นเอกมิติของโครงสร้างคุณลักษณะแฝงวิชาคณิตศาสตร์ จำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน

กลุ่ม	Eigen value		Eigen 1: Eigen 2
	องค์ประกอบที่ 1	องค์ประกอบที่ 2	
เพศ			
เพศชาย	10.60	1.99	5.33
เพศหญิง	8.94	2.38	3.75
สถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน			
เขตกรุงเทพฯ	11.14	2.27	4.90
นอกเขตกรุงเทพฯ	7.24	2.13	3.39

จากการวิเคราะห์ พบว่า ข้อมูลเชิงประจักษ์ มีลักษณะเป็นเอกมิติ (Unidimensional) ซึ่งสอดคล้องกับโมเดลตอบสนองข้อสอบ (IRT) ซึ่งอาจเป็นสาเหตุทำให้วิธี Multiple group CFA สามารถตรวจสอบพบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ได้จำนวนมาก

2.2 เปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ (DIF) วิชาภาษาไทย จำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน พบว่า วิธี Bayesian พบการทำหน้าที่ต่างกันของข้อสอบ (DIF) มากที่สุด และวิธี Multiple group CFA ไม่พบการทำหน้าที่ต่างกันของข้อสอบ ทั้งนี้อาจเป็นเพราะวิธี Bayesian ประมาณค่าพารามิเตอร์มีเป้าหมายการแจกแจงภายหลัง (Posterior distribution) โดยใช้เทคนิค Markov chain monte carlo (MCMC) ทำการแจกแจงเริ่มต้น (Prior distribution) และมีการทำซ้ำจำนวนมากจนทำให้ประมวลผลออกมาได้ใกล้เคียงกับความเป็นจริงมาก ซึ่งสอดคล้องกับเกียรติเทพ ตั้งสันติถาวร (2548, หน้า 117) พบว่า วิธีการของเบส์จะมีประสิทธิภาพดีเมื่อขนาดตัวอย่างที่ใช้มีขนาดเล็กจนถึงขนาดปานกลาง เมื่อตัวอย่างมีขนาดใหญ่ขึ้น ตัวประมาณค่าจะลู่เข้าสู่พารามิเตอร์จริง และสอดคล้องกับ ปรางค์ทิพย์ รัชตะปิติ (2550, หน้า 243) พบว่า วิธีเบส์แบบการแจกแจงก่อนหน้าที่ให้ข้อมูลมีประสิทธิภาพดีที่สุด ยกเว้นกรณีที่ขนาดตัวอย่างมีขนาดเล็ก และค่าความแปรปรวนของความคลาดเคลื่อนจากการสุ่มในสมการหลักมีค่ามากกว่าค่าความแปรปรวนของความคลาดเคลื่อนในการสุ่มในสมการที่เกี่ยวข้องกับสมการหลัก จะพบว่าวิธีการแจกแจงก่อนที่ไม่ให้ข้อมูลจะมีประสิทธิภาพดีที่สุด

จากการที่วิธี Multiple group CFA ตรวจสอบไม่พบการทำหน้าที่ต่างกันของข้อสอบ (DIF) เนื่องจากพบว่าเมื่อทำการสกัดองค์ประกอบโดยการหมุนแกนแบบออโธกอนอล (Orthogonal) ด้วยวิธีวาริแมกซ์ (Varimax) แล้วพบว่า ข้อมูลเชิงประจักษ์ มีแนวโน้มเป็นสององค์ประกอบ (Two factor) ดังนั้นจึงอาจเป็นสาเหตุให้การวิเคราะห์ด้วยวิธี Multiple group CFA ตรวจสอบไม่พบการทำหน้าที่ต่างกันของข้อสอบ (DIF) เหมือนกับวิธีอื่น เพราะข้อมูลเชิงประจักษ์มีแนวโน้มไม่มีความเป็นเอกมิติซึ่งทำให้ไม่สอดคล้องกับโมเดลตอบสนองข้อสอบ (IRT) ซึ่งได้ผลการวิเคราะห์ดังนี้

ตารางที่ 5-4 วิเคราะห์ความเป็นเอกมิติของโครงสร้างคุณลักษณะแฝงวิชาภาษาไทย จำแนกตามเพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน

กลุ่ม	Eigen value		Eigen 1: Eigen 2
	องค์ประกอบที่ 1	องค์ประกอบที่ 2	
เพศ			
เพศชาย	11.67	7.80	1.50
เพศหญิง	10.45	8.08	1.30
จำแนกตามที่ตั้งทางภูมิศาสตร์ของโรงเรียน			
เขตกรุงเทพฯ	12.27	7.95	1.54
นอกเขตกรุงเทพฯ	9.29	8.04	1.15

ข้อเสนอแนะจากการวิจัย

1. ผลการศึกษาครั้งนี้ พบว่า การประมาณค่าพารามิเตอร์อำนาจจำแนกของข้อสอบ ค่าพารามิเตอร์ความสามารถของผู้สอบ ที่วิเคราะห์โดยทั้ง 3 วิธี ได้แก่ ML, วิธี BAYESIAN และวิธี CFA พบว่า ให้ผลการประมาณค่าที่ใกล้เคียงกันมาก ดังนั้นควรเลือกวิธี ML ประยุกต์ใช้โปรแกรม IRT-PRO ในการวิเคราะห์ เนื่องจากเป็นโปรแกรมสำเร็จรูป มีวิธีการใช้ง่ายไม่ต้องเขียนคำสั่ง ใช้เวลาในการวิเคราะห์น้อย ส่วนโปรแกรม Open BUGS และโปรแกรม Mplus ไม่แนะนำให้ใช้ เนื่องจากจะต้องมีการเขียนคำสั่งที่มีความซับซ้อน และใช้เวลาในการวิเคราะห์นานหลายชั่วโมง หรือหลายวัน โดยเฉพาะเมื่อมีการใช้กลุ่มตัวอย่างจำนวนมาก หรือข้อสอบจำนวนหลายข้อ

2. ในการประมาณค่าพารามิเตอร์ความยากของข้อสอบ ควรมีการตรวจสอบลักษณะของการแจกแจงความสามารถของผู้สอบก่อนว่าเป็นการแจกแจงแบบใด เช่น การแจกแจงแบบปกติ แจกแจงแบบเบ้ซ้าย หรือเบ้ขวา ทั้งนี้เพื่อใช้เป็นข้อมูลในการเลือกวิธีการประมาณค่าความยากที่เหมาะสม เช่น วิธี BAYESIAN จะประมาณค่าพารามิเตอร์ความยากของข้อสอบได้ดีเมื่อมีการแจกแจงความสามารถของผู้สอบเป็นแบบเบ้ซ้าย และวิธี Maximum likelihood (ML) จะประมาณค่าพารามิเตอร์ความยากของข้อสอบได้ดี เมื่อข้อมูลมีการแจกแจงความสามารถเป็นแบบเบ้ขวา

3. จากการศึกษาครั้งนี้ผู้วิจัยไม่ได้ทำการตรวจสอบลักษณะของข้อสอบก่อนว่าเป็นข้อสอบปกติทั่วไป หรือเรียกว่าข้อสอบอิสระ หรือเป็นข้อสอบที่ผสมกันระหว่างข้อสอบที่มี

ความเป็นอิสระและข้อสอบที่มีอิทธิพลเทสต์เลทผสมอยู่ แต่จากผลการวิจัย พบว่า วิธี ML วิธี BAYESIAN และวิธี CFA สามารถประมาณค่าพารามิเตอร์อำนาจจำแนกของข้อสอบ และพารามิเตอร์ความสามารถของผู้สอบได้ใกล้เคียงกัน ทั้งในวิชาคณิตศาสตร์ และภาษาไทย และมีความสัมพันธ์กันในระดับสูงอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ยกเว้นการประมาณค่าพารามิเตอร์ความยากของข้อสอบ ที่พบว่า วิธี BAYESIAN จะประมาณค่าได้ดีกว่าวิธี ML และวิธี CFA และมีความสัมพันธ์กับวิธี ML และวิธี CFA ปานกลางค่อนข้างต่ำ ทั้งนี้อาจเป็นเพราะอิทธิพลของข้อสอบเทสต์เลท (Testlet) มีไม่มากพอที่จะทำให้เกิดผลกระทบต่อการประมาณค่าพารามิเตอร์พารามิเตอร์อำนาจจำแนกและพารามิเตอร์ความสามารถของผู้สอบ จึงทำให้ผลการประมาณค่าดังกล่าวจากทั้งสามวิธีไม่แตกต่างกัน และเมื่อทำการตรวจสอบลักษณะของข้อสอบวิชาคณิตศาสตร์ และภาษาไทยแล้ว พบว่า ข้อสอบวิชาคณิตศาสตร์มีลักษณะข้อสอบเป็นแบบทั่วไปมีความเป็นอิสระของข้อสอบ แต่วิชาภาษาไทยมีลักษณะที่เป็นข้อสอบอิสระและข้อสอบที่มีอิทธิพลเทสต์เลทผสมอยู่ในแบบทดสอบ ตอนที่ 1 ตั้งแต่ข้อ 1-10 และตอนที่ 3 ข้อ 69-70

บรรณานุกรม

กาญจนา วัธนสุนทร. (2538). *การพัฒนาเกณฑ์ตัดสินข้อสอบลำเอียงทางเพศ*. วิทยานิพนธ์ครุศาสตรดุษฎีบัณฑิต, สาขาวิชาวิจัยการศึกษา, บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย.

เกสร หว่างจิตร. (2539). *การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบสำหรับแบบสอบคัดเลือกระดับบัณฑิตศึกษาวิชาภาษาไทยและภาษาอังกฤษด้วยวิธีแมนเทล เฮลว์เซล*. วิทยานิพนธ์ครุศาสตรมหาบัณฑิต, สาขาวิชาวิจัยการศึกษา, บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย.

เกียรติเทพ ตั้งสันติถาวร. (2548). *การเปรียบเทียบการประมาณค่าพารามิเตอร์ด้วยวิธีเบส์กับวิธีวิเคราะห์ความถดถอยแบบสองขั้นกำลังสองน้อยสุด*. วิทยานิพนธ์สถิติศาสตรมหาบัณฑิต, สาขาวิชาสถิติ, บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย.

จิติมา วรรณศรี. (2539). *การเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมนเทล-แฮนส์เซลกับวิธีซิปเทสท์ เมื่อความยาวแบบสอบ ขนาดกลุ่มตัวอย่างและอัตราส่วน ของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบต่างกัน*. วิทยานิพนธ์ครุศาสตรมหาบัณฑิต, สาขาวิชาการวัดและประเมินผลการศึกษา, คณะครุศาสตร์, จุฬาลงกรณ์มหาวิทยาลัย.

ชัยยศ ชาวระนอง. (2553). *ประสิทธิภาพของโมเดลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบพหุมิติโดยวิธีการวิเคราะห์องค์ประกอบเชิงยืนยันแฝงภายใน*. ดุษฎีนิพนธ์ปรัชญาดุษฎีบัณฑิต, สาขาวิชาวิจัย วัดผลและสถิติการศึกษา, บัณฑิตวิทยาลัย, มหาวิทยาลัยบูรพา.

ชัยวิชิต เชียรชัยชนะ. (2552). การวิเคราะห์พหุมิติ (Multidimensional analysis). *วารสารศึกษาศาสตร์ มหาวิทยาลัยขอนแก่น*, *32*(4), 13-22.

ชินภัทร ภูมิรัตน์. (2554). *สพฐ. ตีปี๊บใช้ O-NET รับ นร.ปีหน้าโยนข้อสอบมีจุดอ่อนทำเด็กมึน*. เข้าถึงได้จาก http://www.manager.co.th/QOL/ViewNews.aspx? NewsID= 9540000045848

ชินภัทร ภูมิรัตน์. (2556). *กพฐ. ชี้สอบ GAT-PAT ไม่ได้วัดคุณภาพเด็ก*. เข้าถึงได้จาก http://www.13nr.org/posts/514806

ณัฏฐปิญชาน์ พิชญาชมชื่น. (2558). ผลการจัดกิจกรรมการเรียนรู้แบบอุปนัยที่มีต่อมโนทัศน์ทางคณิตศาสตร์และความสามารถในการให้เหตุผลทางคณิตศาสตร์ของนักเรียนชั้นมัธยมศึกษาปีที่ 4. *วารสารมหาวิทยาลัยศิลปากร ฉบับภาษไทย สาขามนุษยศาสตร์ สังคมศาสตร์ และศิลปะ*, *8*(2), 76-91.

ทศพร จันทนราช. (2556). *โมเดลความสัมพันธ์เชิงสาเหตุของปัจจัยที่สัมพันธ์กับพฤติกรรมการบริหารงานของผู้บริหาร โรงเรียนที่มีอิทธิพลต่อผลสัมฤทธิ์ทางการเรียนของนักเรียน*. วิทยานิพนธ์ดุษฎีบัณฑิต, สาขาวิชาการบริหารการศึกษา, คณะศึกษาศาสตร์, มหาวิทยาลัยบูรพา.

ทิพย์รัตน์ ผลบุญ. (2540). *การวิเคราะห์ความลำเอียงของข้อสอบปลายภาควิชาคณิตศาสตร์ชั้นมัธยมศึกษาปีที่ 1 ของกลุ่มโรงเรียนมัธยมศึกษาตอนต้นในจังหวัดกำแพงเพชร*. วิทยานิพนธ์ศึกษาศาสตรมหาบัณฑิต, สาขาวิชาการวัดผลการศึกษา, บัณฑิตวิทยาลัย, มหาวิทยาลัยนเรศวร.

นพดล มีชั้นช่วง. (2544). *การเปรียบเทียบผลของการประมาณค่าพารามิเตอร์ ตามทฤษฎีการตอบข้อสอบ ระหว่างวิธีแมกซิมัมไลค์ลิฮูด วิธีฮิวริสติก และวิธีเบส์ ของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนคณิตศาสตร์ ชั้นมัธยมศึกษาปีที่ 1*. วิทยานิพนธ์การศึกษามหาบัณฑิต, สาขาวิชาวัดผลการศึกษา, บัณฑิตวิทยาลัย, วิทยาลัยมหาสารคาม.

นวลแสง สูคีรี. (2560). โมเดลความสัมพันธ์เชิงสาเหตุของปัจจัยที่ส่งผลต่อประสิทธิภาพทางวิชาการของโรงเรียนมัธยมศึกษาในจังหวัดภาคใต้. *วารสารมหาวิทยาลัยราชภัฏธนบุรี*, *11*(1), 75-90.

ประชุม สุวัตถี. (2527). *ทฤษฎีการอนุมานเชิงสถิติ*. กรุงเทพฯ: โอเดียนสโตร์.

ปรางค์ทิพย์ รัชตะปิติ. (2550). *การเปรียบเทียบการประมาณค่าพารามิเตอร์ด้วยวิธีเบส์กับวิธีวิเคราะห์ความถดถอยพหุคูณแบบกำลังสองน้อยสุดสองขั้น*. วิทยานิพนธ์สถิติศาสตรมหาบัณฑิต, สาขาวิชาสถิติ, บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย.

พรรณี จินตมาศ. (2540). *การเปรียบเทียบผลการวิเคราะห์ความลำเอียงข้อสอบ โดยใช้ขนาดกลุ่มผู้สอบและวิธีวิเคราะห์ต่างกัน*. ปริญญานิพนธ์การศึกษามหาบัณฑิต, สาขาวิชาการวัดผลการศึกษา, บัณฑิตวิทยาลัย, มหาวิทยาลัยศรีนครินทรวิโรฒ.

เพ็ญพนา สุขสม. (2540). *การเปรียบเทียบผลของวิธีวิเคราะห์ความลำเอียงของข้อสอบที่แตกต่างกัน 3 วิธี*. วิทยานิพนธ์ศึกษาศาสตรมหาบัณฑิต, สาขาวิชาการวัดผลการศึกษา, บัณฑิตวิทยาลัย, มหาวิทยาลัยนเรศวร.

มิ่ง เทพครเมือง. (2554). *การตรวจสอบความเท่าเทียมกันของการวัดบนพื้นฐานทฤษฎีการตอบสนองแบบคะแนนจริงสัมพันธ์และทฤษฎีการตอบข้อสอบ*. ปริญญานิพนธ์การศึกษาดุษฎีนิพนธ์, สาขาวิชาการทดสอบและวัดผลการศึกษา, บัณฑิตวิทยาลัย, มหาวิทยาลัยศรีนครินทรวิโรฒ.

รัชนก ยี่สุ่นศรี. (2544). *การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบและแบบสอบด้วยกระบวนการ ดี เอฟ ไอ ที สำหรับแบบสอบคัดเลือกบุคคลเข้าศึกษาในสถาบันอุดมศึกษาวิชาภาษาอังกฤษและวิชาคณิตศาสตร์*. วิทยานิพนธ์ครุศาสตรมหาบัณฑิต, สาขาวิชาวิจัยการศึกษา, บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย.

เรวดี อินทะสะระ. (2539). *ผลการตรวจสอบความลำเอียงของข้อสอบต่อการศึกษาความเที่ยงตรงเชิงพยากรณ์ของแบบทดสอบคัดเลือกที่คิดคะแนนต่างกัน*. ดุษฎีนิพนธ์การศึกษาดุษฎีบัณฑิต, สาขาวิชาการทดสอบและวัดผลการศึกษา, บัณฑิตวิทยาลัย, มหาวิทยาลัยศรีนครินทรวิโรฒ.

วิชุดา บัวคง. (2533). *การเปรียบเทียบประสิทธิผลของวิธีประมาณค่าพารามิเตอร์ของแบบจำลองโลจิสติค 3 พารามิเตอร์ ระหว่างวิธีแมกซิมัมไลค์ลิฮูด วิธีฮิวริสติค และวิธีของเบส์ในแบบสอบวัดผลสัมฤทธิ์และแบบสอบความถนัด*. วิทยานิพนธ์ครุศาสตรมหาบัณฑิต, สาขาวิชาวิจัยการศึกษา, บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย.

วิทยากร เชียงกูล. (2555). การรณรงค์ให้เด็กและเยาวชนไทยรักการอ่าน-กลยุทธ์สำคัญในการปฏิรูปการศึกษาและการปฏิรูปประเทศ. *วารสารสังคมศาสตร์และมนุษยศาสตร์*, *38*(2), 1-15.

ศาสตริยา ฉ่ำสมัย. (2548). *การศึกษาเปรียบเทียบการอนุมานค่าเฉลี่ยของประชากรระหว่างวิธีการแบบภาวะน่าจะเป็นและวิธีการแบบเบส์เซียน*.วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต, สาขาวิชาสถิติ, บัณฑิตวิทยาลัย, มหาวิทยาลัยเกษตรศาสตร์.

ศิริชัย กาญจนวาสี. (2545). *ทฤษฎีการทดสอบแนวใหม่* (พิมพ์ครั้งที่ 2). กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

ศิริชัย กาญจนวาสี. (2550). *ทฤษฎีการทดสอบแนวใหม่* (พิมพ์ครั้งที่ 3). กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

สถาบันส่งเสริมการสอนวิทยาศาสตร์และเทคโนโลยี. (2555). *การวัดผลประเมินผลคณิตศาสตร์*. กรุงเทพฯ: ซีเอ็ดยูเคชั่น.

สมศักดิ์ จันผ่อง. (2542). *การเปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยใช้วิธีวิเคราะห์และขนาดกลุ่มผู้สอบต่างกัน*. ปริญญานิพนธ์การศึกษามหาบัณฑิต, สาขาวิชาวัดผลการศึกษา, บัณฑิตวิทยาลัย, มหาวิทยาลัยนเรศวร.

สุชาดา กีระนันทน์. (2525). *การสำรวจตัวอย่าง*. กรุงเทพฯ: จุฬาลงกรณ์มหาวิทยาลัย.

สุพัฒนา หอมบุปผา. (2556). *การเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN*. ดุษฎีนิพนธ์ปรัชญาดุษฎีบัณฑิต, สาขาวิชาวิจัยวัดผลและสถิติการศึกษา, ศึกษาศาสตร์, มหาวิทยาลัยบูรพา.

สุภมาศ อังศุโชติ, สมถวิล วิจิตรวรรณา และรัชนีกูล ภิญโญภานุวัฒน์. (2554). *สถิติวิเคราะห์สำหรับการวิจัยทางสังคมศาสตร์และพฤติกรรมศาสตร์: เทคนิคการใช้โปรแกรม LISREL* (พิมพ์ครั้งที่ 3). กรุงเทพฯ: เจริญดีมั่นคงการพิมพ์.

สุวภัทร์ ทัพซ้าย. (2557). ผลของการใช้เทคนิคการสอน PMI ต่อความสามารถในการคิดอย่างมีวิจารณญาณวิชาภาษาไทยของนักเรียนชั้นมัธยมศึกษาปีที่ 3 โรงเรียนบ้านเหล่าจั่นหนองทุ่ม จังหวัดร้อยเอ็ด. *วารสารมหาวิทยาลัยราชภัฏธนบุรี*, *8*(2), 56-68.

สุวิมล ติรกานันท์. (2553). *การวิเคราะห์ตัวแปรพหุในงานวิจัยทางสังคมศาสตร์*. กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

อริสพา เตหลิ่ม. (2559). *การเปรียบเทียบประสิทธิผลการประมาณค่าพารามิเตอร์และการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมกซิมัมไลค์ลิฮู้ด วิธีของเบส์และวิธีของเบส์แบบมีอิทธิเทสต์เลท*. ดุษฎีนิพนธ์ปรัชญาดุษฎีบัณฑิต, สาขาวิชาวิจัย วัดผลและสถิติการศึกษา, คณะศึกษาศาสตร์, มหาวิทยาลัยบูรพา.

อัมพร ม้าคนอง. (2553). *ทักษะกระบวนการทางคณิตศาสตร์: การพัฒนาเพื่อพัฒนาการ*. กรุงเทพฯ: ศูนย์ตำราและเอกสารทางวิชาการ คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.

อาวีพร ปานทอง. (2558). *การเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบให้คะแนนหลายค่าโดย วิธีทดสอบอัตราส่วนความควรจะเป็นวิธีเบส์เซียนและวิธีโพลี-ซิปเทสท์*. ดุษฎีนิพนธ์ปรัชญาดุษฎีบัณฑิต, สาขาวิชาวิจัยวัดผลและสถิติการศึกษา, คณะศึกษาศาสตร์, มหาวิทยาลัยบูรพา.

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal Educational Measurement*, *10*, 95-106.

Atar, B., & Kamata, A. (2011). Comparison of IRT likelihood ratio test and logistic regression DIF detection procedures. *Hacettepe Universitesi Egitim Fakultesi Dergisi*, *41*(41).

Baker, F. B. (1997, June). Empirical sampling distributions of equating coefficients for graded and nominal response instruments. *Applied Psychological Measurement*, *21*(2), 157-172.

Barton, M. A., & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. *ETS Research Report Series*, *1981*(1), i-8.

Béguin, A. A., & Glas, C. A. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*(4), 541-561.

Berk, R. A. (1982). *Handbook of methods for detecting test bias*: Johns Hopkins University.

Binet, A., & Simon, T. (1916 a). *The development of intelligence in children: The binet-simon scale*: New York: Williams & Wilkins.

Binet, A., & Simon, T. (1916 b). *The intelligence of the feeble-minded*. New York: Williams & Wilkins.

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research*, *17*(3), 303-316.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using markov chain monte carlo. *Applied Psychological Measurement*, *27*(6), 395-414.

Burton, I. (1993). *The environment as hazard*. New York: Guilford.

Byrne, D. S. (1998). *Complexity theory and the social sciences: An introduction*. New York: Psychology.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological bulletin*, *105*(3), 456.

Cai, L., Du Toit, S., & Thissen, D. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]*. Chicago, IL: Scientific Software International.

Camilli, G. (2006). Test fairness. *Educational measurement*, *4*, 221-256.

Camilli, G., & Penfield, D. A. (1997). Variance estimation for differential test functioning based on mantel haenszel statistics. *Journal of Educational Measurement*, *34*(2), 123-139.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. London: Sage.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: subgroup differences in test performance and face validity perceptions. *Journal of applied psychology*, *82*(1), 143.

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of management*, *25*(1), 1-27.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233-255.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*(1), 31-44.

De La Torre, J., Stark, S., & Chernyshenko, O. S. (2006). Markov chain monte carlo estimation of item parameters for the generalized graded unfolding model. *Applied Psychological Measurement*, *30*(3), 216-232.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with mantel-haenszel, SIBTEST and the IRT likelihood ratio. *Applied Psychological Measurement*, *29*(5), 278-295.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with mantel-haenszel, sibtest, and the IRT likelihood ratio. *Applied Psychological Measurement*, *29*(4), 278-295.

Flier, H. Van der, Mellenbergh, G. J., Ader, H., & Wijn, M. (1984). An iterative bias detection procedure. *Journal of Educational Measurement*, *21*, 131-145.

French, B. F., & Finch, W. H. (2008). Multigroup confirmatory factor analysis: Locating the invariant referent sets. *Structural Equation Modeling*, *15*(1), 96-113.

Fukahara, H., & Kamata, A. (2007). *DIF detection in presence of locally dependent items.* Florida: Tampa.

Fukuhara, H., & Kamata, A. (2011). A bifactor multidimensional item response theory model for differential item functioning analysis on testlet-based items. *Applied Psychological Measurement*, *35*(8), 604-622.

Geman, T., & Geman, D. (1984). Stochastic relaxation gibbs distribution and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *3*(6), 721-741.

Glickman, M. E., Seal, P., & Eisen, S. V. (2009). A non-parametric bayesian diagnostic for detecting differential item functioning in IRT models. *Health Services and Outcomes Research Methodology*, *9*(3), 145-161.

Goodwin, A. P., Huggins, A. C., Carlo, M., Malabonga, V., Kenyon, D., Louguit, M., & August, D. (2012). Development and validation of extract the base: An english derivational morphology test for third through fifth grade monolingual students and spanish-speaking english language learners. *Language Testing, 29*(2), 265-289.

Green, T. H. (1994). Experimental studies of trace-element partitioning applicable to igneous pertro-genesis-Sedona 16 years later. *Chemical Geology*, 117, 1-36.

Haladyna, T. M. (2006). Perils of standardized achievement testing. *Educational Horizons*, *12*(33), 30-43.

Hambleton. R. K., & Swaminathan, H. (1985). *Item response theory: Principles and application*. Boston: Kluner-Nijhoff.

Hambleton, R. K, Swaminatan, H., & Rogers, H. J. (1991). *Fundamentals of item reaponse theory*. Newbury Park, California: Sage.

Hancock, B. C., & Zografi, G. (1997). Characteristics and significance of the amorphous state in pharmaceutical systems. *Journal of pharmaceutical sciences*, *86*(1), 1-12.

Hidalgo, M. D., & LÓPez-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement*, *64*(6), 903-915.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the mantel-haenszel procedure. *Test validity*, *5*(2), 129-145.

Holmes, F. W., & French, B. F. (2007). Detection of crossing differential item functioning: A comparison of four methods. *Educational and Psychological Measurement*, *67*(4), 565-582.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental aging research*, *18*(3), 117-144.

Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to analysis of attitude scale translations. *Journal of applied psychology*, *67*(6), 818.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two-and three-parameter logistic item characteristic curves: A monte carlo study. *Applied Psychological Measurement*, *6*(3), 249-260.

Ironson, G. H. (1982). Use of chi-square and latent trait approaches for detecting item bias. *Handbook of methods for detecting test bias*, *15*(38), 117-160.

Jackson, P. R., Wall, T. D., Martin, R., & Davids, K. (1993). New measures of job control, cognitive demand, and production responsibility. *Journal of applied psychology*, *78*(5), 753.

Johnson, M. S., & Sinharay, S. (2005). Calibration of polytomous item families using bayesian hierarchical modeling. *Applied Psychological Measurement*, *29*(5), 369-400.

Jöreskog, K. G. (1993). Testing structural equation models. *Sage focus editions, 154*, 294-294.

Kim, S. H., & Allen, S. C. (1991, September). A comparison of two area measures for detecting differential item functioning. *Applied Psychological Measurement*, *15*(3), 269-278.

Kim, S. H., & Cohen, A. S. (1994). An investigation of Lord's for the detection of differential item functioning. *Applied Psychological Measurement*, *18*(3), 217-228.

Lahiri, K., & Gao, J. (2002). *Bayesian analysis of nested logit model*. Available form https://papers.ssrn.com/sol3/papers.cfm?abstract_id=317779

Langenfeld, T. E. (1997). Test fairness: Internal and external investigations of gender bias in mathematics testing. *Educational Measurement: Issues and practice*, *16*(1), 20-26.

Le, L. T. (2009). Investigating gender differential item functioning across countries and test languages for PISA science items. *International Journal of Testing*, *9*(2), 122-133.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational evaluation and policy analysis*, *15*(1), 1-16.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement*, *5*(2), 159-173.

Longford, N. T. (1993). *VARCL: Software for variance component analysis of data with nested random effects (maximum likelihood): Manual*: iec ProGAMMA.

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). *Stability of the MH D-DIF statistics across populations*. In P. W. Holland & H. Wainer (Eds.). Differential item functioning (pp. 171-196). Hillsdale, NJ: Erlbaum.

Lopez Rivas, G. E., Stark, S., & Chernyshenko, O. S. (2009). The effects of referent item parameters on differential item functioning detection using the free baseline likelihood ratio test. *Applied Psychological Measurement*, *33*(4), 251-265.

Lord, F. M. (1980). *Application of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1980). *Application of item response theory to practical testing problem*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lytle, A., Brett, J., Barsness, Z., Tinsley, C., & Janssens, M. (1995). A paradigm for confirmatory cross-cultural research in organizational-behavior. *Research in organizational behavior: An annual series of analytical essays and critical reviews*, *17*(17), 167-214.

Main, I. G., Sammonds, P. R., & Meredith, P. G. (1993). Application of a modified griffith criterion to the evolution of fractal damage during compressional rock failure. *Geophysical Journal International*, *115*(2), 367-380.

Marsh, H. W., Hau, K. T., Chung, C. M., & Siu, T. L. (1998). Confirmatory factor analyses of chinese students' evaluations of university teaching. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(2), 143-164.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the mantel-haenszel statistic. *Educational and Psychological Measurement*, *52*(2), 443-451.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1994). Identification of nonuniform differential item functioning using a variation of the mantel-haenszel procedure. *Educational and Psychological Measurement*, *54*(2), 284-291.

Mazor, K. M., Kanjee, A., & Clauser, B. E. (1995). Using logistic regression and the mantel-haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement*, *32*(2), 131-144.

McLeod, L., Lewis, C., & Thissen, D. (2003). A bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement*, *27*(2), 121-137.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525-543.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement*, *30*(2), 107-122.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*(4), 297-334.

Morizot, J., Ainsworth, A. T., & Reise, S. (2007). Toward modern psychometrics: Application of item response theory models. In R. W. Robins, R. C. Fraley, and R. F. Krueger (eds.), pp. 407-423. *Handbook of research methods in personality psychology*. New York: Guilford.

Muthen, B. (1988). *Some uses of structural equation modeling in validity studies: Extending IRT to external variables*. In H. Wainer & H. I. Braun (Eds.), *Test validiry* (pp. 213-238). Hillsdale, NJ: Lawrence Erlbaum Associates.

Muthén, B., & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics*, *10*(2), 133-142.

Muthén, B. O., Kao, C. F., & Burstein, L. (1991). Instructionally sensitive psychometrics: application of a new IRT based detection technique to mathematics achievement test items. *Journal of Educational Measurement*, *28*(1), 1-22.

Muthén, B., & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics*, *10*(2), 133-142.

Narayanan, P., & Swaminathan, H. (1994). Performance of the mantel-haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement*, *18*(4), 315-328.

Niti, M., Ng, T. P., Chiam, P. C., & Kua, E. H. (2007). Item response bias was present in instrumental activity of daily living scale in Asian older adults. *Journal of Clinical Epidemiology*, *60*(4), 366-374.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to markov chain monte carlo methods for item response models. *Journal of educational and behavioral statistics*, *24*(2), 146-178.

Penfield, R., & Camilli, G. (2007). Psychometrics. *Handbook of statistics*, *26*(2), 112-130.

Ployhart, R. E., & Oswald, F. L. (2004). Applications of mean and covariance structure analysis: ntegrating correlational and experimental approaches. *Organizational research methods*, *7*(1), 27-65.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*(2), 197-207.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of applied psychology*, *87*(3), 517.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological bulletin*, *114*(3), 552.

Rensvold, R. B., & Cheung, G. W. (1998). Testing measurement models for factorial invariance: A systematic approach. *Educational and Psychological Measurement*, *58*(6), 1017-1034.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*(2), 105-116.

Rudner, L. M. (1977 a). Bias item detecting technique. *Journal of Educational Statistics*, *5,* 213-233.

Rudner, L. M. (1977 b). *An approach to biased item identification using latent trait measurement theory*. Paper presented at the Annual Meeting of the Amercian Educational Research Association, New York.

Shealy, R. T., & Stout, F. (1993). A model-based standardization approach that separates true bias/DIF from ability group differences and detectes test bias/DTF as well as Item bias/DIF. *Psychometrika*, *58*(2), 159-194.

Spiegelhalter, D. J., & Best, N. G. (2003). Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. *Statistics in medicine*, *22*(23), 3687-3709.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of applied psychology*, *91*(6), 1292.

Steenkamp, J. B. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of consumer research*, *25*(1), 78-90.

Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika*, *50*(3), 349-364.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361-370.

Teresi, J. A. (2006). Different approaches to differential item functioning in health applications: Advantages, disadvantages and some neglected topics. *Medical care*, *44*(11), S152-S170.

Thissen, D., Steinberg, L., & Gerrand, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin*, *99*, 118-128.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer and H. I. Braun, (Eds.), *Test validity*. (pp. 147-169). Hillsdale, NJ: Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Erlbaum.

Thurston, E. S. (1947). Recent developments in restitution: 1940-1947. *Michigan Law Review*, *45*(8), 935-976.

Thurstone, L. L. (1947). *Multiple factor analysis*: *A development and expansion of vectors of the mind*. Chicago: University of Chicago.

Tülin, A., & Keleecioglu, H. (2010). Comparison of differential item functioning determination techniques: HGLM, LR and IRT-LR.

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational research methods*, *5*(2), 139-158.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational research methods*, *3*(1), 4-70.

Van den Berg, R. (2002). Teachers' meanings regarding educational practice. *Review of Educational Research*, *72*(4), 577-625.

Van der Flier, Mellenbergh, G. J., Ader, H. J., &Wijn, M. (1984). The iterative item bias selection method. *Journal of Educational Measurement*, *21*, 131-145.

Weziak-Bialowolska, D. (2015). Differences in gender norms between countries: Are they valid? The issue of measurement invariance. *European Journal of Population*, *31*(1), 51-76.

Wiberg, M. (2007). *Measuring and detecting differential item functioning in criterion-referenced licensing test: A theoretic comparison of methods*. Sweden: Umea Universitet.

Wonsuk, K. (2003). *Develmopment of a differential item functioning (DIF) procedure using the hierarchical generalized liner model: A comparison study with logistic regression procedure*. A Thesis in Educational Psychology, Doctor of Philosophy, The Graduate School College of Education, The Pennsylvania State University.

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago, IL: Mesa.

Zieky, M. (1993). *Practical questions in the use of DIF statistics in test development*. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Lawrence Erlbaum Associates.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF)*. Ottawa: National Defense Headquarters.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language assessment quarterly*, *4*(2), 223-233.

Zumbo, B., & Thomas, D. (1997). A measure of effect size for a model-based approach for studying DIF. *Prince george, Canada: University of Northern British Columbia, Edgeworth Laboratory for Quantitative Behavioral Science*.

Zwick, R. (1990). When do item response function and mantel-haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics*, *15*(3), 185-197.

Zwick, R., Thayer, D. T., & Lewis, C. (2000). Using loss functions for DIF detection: An empirical bayes approach. *Journal of educational and behavioral statistics*, *25*(2), 225-247.

ภาคผนวก

ภาคผนวก ก

หนังสือขอข้อมูลคะแนนทดสอบระดับชาติของนักเรียน
ชั้นมัธยมศึกษาปีที่ 6 (O-NET)

(สำเนา)

ที่ ศธ ๖๒๑๘/๐๕๘ คณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา
๑๖๙ ถ.ลงหาดบางแสน ต.แสนสุข
อ.เมือง จ.ชลบุรี ๒๐๑๓๑

๑๑ มกราคม ๒๕๖๐

เรื่อง ขอความอนุเคราะห์ผลสอบรายข้อ

เรียน ผู้อำนวยการสถาบันการทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน)

สิ่งที่ส่งมาด้วย โครงร่างงานวิจัย จำนวน ๑ ฉบับ

ข้าพเจ้านางสาวเพ็ญศรี เทียมสุข นิสิตปริญญาเอก หลักสูตรปรัชญาดุษฎีบัณฑิต สาขาวิชาวิจัย วัดผล และสถิติการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา ได้รับอนุมัติให้ทำดุษฎีนิพนธ์ เรื่อง การเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี IRT-likelihood ratio วิธี Bayesian และวิธี Multiple-group CFA โดยมีผู้ช่วยศาสตราจารย์ ดร.เชษฐ์ ศิริสวัสดิ์ เป็นอาจารย์ที่ปรึกษาหลัก และรองศาสตราจารย์ ดร.ไพรัตน์ วงษ์นาม และผู้ช่วยศาสตราจารย์ ดร.สุรีพร อนุศาสนนันท์ เป็นอาจารย์ที่ปรึกษาร่วม ในการทำวิจัยครั้งนี้จำเป็นต้องใช้ข้อมูลผลการสอบรายข้อแบบทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET) ระดับชั้นมัธยมศึกษาปีที่ ๖ ประจำปีการศึกษา ๒๕๕๘ เพื่อนำข้อมูลมาวิเคราะห์หาข้อสรุปตามวัตถุประสงค์ของการวิจัย ข้อมูลที่ผู้วิจัยต้องการขอความอนุเคราะห์จากสถาบันการทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน) ได้แก่ ข้อมูลเพศผู้สอบ สถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน และคำตอบของนักเรียนรายบุคคลที่ผ่านการตรวจให้คะแนน แบบ ๑-๐ วิชาภาษาไทย, คณิตศาสตร์ และวิทยาศาสตร์ ประจำปีการศึกษา ๒๕๕๘ โดยต้องการขอข้อมูลเฉพาะโรงเรียนมัธยมศึกษาขนาดใหญ่พิเศษ ที่เป็นโรงเรียนสหศึกษา สังกัด สพฐ. กระทรวงศึกษาธิการ และข้อสอบฉบับจริงทั้ง ๓ วิชาดังกล่าว ซึ่งผลการวิจัยครั้งนี้จะเป็นประโยชน์ในเชิงวิชาการในด้านการประยุกต์ใช้วิธีการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ (DIF) ตามทฤษฎีการตอบสนองข้อสอบ ระหว่างการวิเคราะห์ข้อมูลผู้วิจัยจะเก็บรักษาข้อมูลดังกล่าวไว้เป็นความลับ และเข้าถึงข้อมูลได้เฉพาะข้าพเจ้าและผู้ร่วมวิเคราะห์ข้อมูลเท่านั้น เมื่อผู้วิจัยทำการวิจัย และสร้างข้อสรุปในภาพรวมเรียบร้อยแล้วจะดำเนินการลบข้อมูลดังกล่าวทิ้งทันที

จึงเรียนมาเพื่อโปรดพิจารณา คณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา หวังเป็นอย่างยิ่งว่าคงจะได้รับความอนุเคราะห์จากท่านด้วยดี และขอขอบคุณอย่างสูงมา ณ โอกาสนี้

ขอแสดงความนับถือ

(ลงชื่อ) เชษฐ์ ศิริสวัสดิ์

(ผู้ช่วยศาสตราจารย์ ดร.เชษฐ์ ศิริสวัสดิ์)
รองคณบดีฝ่ายบัณฑิตศึกษา ปฏิบัติการแทน
คณบดีคณะศึกษาศาสตร์ ปฏิบัติการแทน
ผู้รักษาการแทนอธิการบดีมหาวิทยาลัยบูรพา

ภาควิชาวิจัยและจิตวิทยาประยุกต์
โทรศัพท์ ๐-๓๘๑๐-๒๐๗๖
โทรสาร ๐-๓๘๓๙-๓๒๕๑
ผู้วิจัยโทร. ๐๘-๖๔๐๔-๐๗๖๓

ภาคผนวก ข

ตัวอย่างผลการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบด้วยวิธี ML
จากโปรแกรม IRT PRO, วิธี BAYESIAN จากโปรแกรม Open BUGS
และวิธี CFA จากโปรแกรม Mplus

ตารางที่ ข-1 ผลการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบวิชาคณิตศาสตร์ ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA

ผู้สอบคนที่	ผลการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบวิชาคณิตศาสตร์		
	ค่า θ ของ ML	ค่า θ ของ BAYESIAN	ค่า θ ของ CFA
1	-0.42	-0.52	-0.49
2	-0.42	-0.32	-0.31
3	0.20	0.33	0.33
4	-0.09	-0.07	-0.07
5	-0.54	-0.63	-0.61
6	-0.19	-0.39	-0.41
7	0.11	-0.02	0.00
8	-0.19	-0.11	-0.10
9	-0.19	-0.14	-0.16
10	-0.30	-0.45	-0.44
11	1.11	1.13	1.11
12	0.20	0.25	0.25
13	-0.30	-0.30	-0.33
14	-0.54	-0.64	-0.63
15	-0.67	-0.77	-0.79
16	0.64	0.70	0.71
17	-0.96	-1.00	-0.98
18	-0.67	-0.82	-0.80
19	-0.67	-0.59	-0.62
20	0.01	0.08	0.10
21	-0.81	-0.74	-0.72
22	-0.67	-0.71	-0.71
23	-0.09	-0.09	-0.08
24	1.01	0.99	1.00
25	-0.30	-0.42	-0.46

ตารางที่ ข-1 (ต่อ)

ผู้สอบคนที่	ผลการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบวิชาคณิตศาสตร์		
	ค่า θ ของ ML	ค่า θ ของ BAYESIAN	ค่า θ ของ CFA
26	-1.32	-0.98	-0.96
27	-0.30	-0.34	-0.34
28	-0.81	-1.04	-1.08
29	-0.67	-0.61	-0.61
30	3.02	3.12	3.02
.	.	.	.
.	.	.	.
.	.	.	.
2394	0.92	0.84	0.82
2395	-0.67	-0.78	-0.80
2396	0.20	0.19	0.19
2397	0.38	0.50	0.51
2398	0.20	0.23	0.23
2399	0.65	0.62	0.64
2400	1.11	0.98	1.00

ตารางที่ ข-2 ผลการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบวิชาคณิตศาสตร์ ด้วยวิธี ML วิธี BAYESIAN และวิธี CFA

ผู้สอบคนที่	ผลการประมาณค่าพารามิเตอร์ผู้สอบวิชาคณิตศาสตร์		
	ค่า θ ของ ML	ค่า θ ของ BAYESIAN	ค่า θ ของ CFA
1	-0.61	-0.80	-0.78
2	-0.03	-0.18	-0.17
3	-0.29	-0.27	-0.27
4	0.30	0.24	0.24
5	-0.29	-0.32	-0.31
6	-0.48	-0.45	-0.45
7	-2.40	-2.36	-2.34
8	-1.07	-0.87	-0.88
9	-1.43	-1.32	-1.31
10	1.55	1.42	1.40
11	-0.23	0.01	0.00
12	0.03	0.17	0.17
13	-0.93	-0.90	-0.89
14	2.19	1.91	1.89
15	-0.67	-0.62	-0.62
16	0.10	0.23	0.23
17	-1.66	-1.53	-1.51
18	-0.16	0.03	0.03
19	-0.48	-0.45	-0.44
20	-1.35	-1.32	-1.30
21	-0.35	-0.33	-0.34
22	-0.10	0.08	0.08
23	0.16	-0.02	-0.02
24	0.16	0.38	0.37
25	-0.42	-0.46	-0.46

ตารางที่ ข-2 (ต่อ)

ผู้สอบคนที่	ผลการประมาณค่าพารามิเตอร์ผู้สอบวิชาคณิตศาสตร์		
	ค่า θ ของ ML	ค่า θ ของ BAYESIAN	ค่า θ ของ CFA
26	-1.14	-1.02	-1.00
27	-2.19	-2.09	-2.07
28	-1.43	-1.31	-1.29
29	0.10	-0.01	-0.01
30	1.65	1.52	1.51
.	.	.	.
.	.	.	.
.	.	.	.
2394	0.43	0.54	0.54
2395	0.79	0.66	0.64
2396	-0.42	-0.43	-0.43
2397	0.03	-0.02	-0.03
2398	-0.42	-0.17	-0.17
2399	-0.03	0.04	0.04
2400	1.11	1.31	1.28

ภาคผนวก ค

ตัวอย่างผลการประมาณค่าพารามิเตอร์อำนาจจำแนกด้วยวิธี ML

จากโปรแกรม IRT PRO

ตารางที่ ค-1 ผลการประมาณค่าพารามิเตอร์อำนาจจำแนกของข้อสอบวิชาคณิตศาสตร์ จำแนกตามเพศ ด้วยวิธี ML

ข้อสอบข้อที่		Total X^2	*df*	*p*	X^2_a	*df*	*p*	$X^2_{c\|a}$	*df*	*p*
เพศชาย	เพศหญิง									
1	1	6.4	2	0.0405	3.6	1	0.0594	2.9	1	0.0919
2	2	2.6	2	0.2759	1. 5	1	0.2279	1.1	1	0.2899
3	3	0.9	2	0.6470	0.1	1	0.8054	0.8	1	0.3683
4	4	2.1	2	0.3565	1.7	1	0.1962	0.4	1	0.5323
5	5	0.0	2	0.9985	0.0	1	0.9821	0.0	1	0.9609
6	6	4.7	2	0.0940	0.4	1	0.5523	4.4	1	0.0365
7	7	0.0	2	0.9851	0.0	1	0.8714	0.0	1	0.9512
8	8	4.2	2	0.1238	3.9	1	0.0478	0.3	1	0.6108
9	9	7.4	2	0.0249	7.2	1	0.0072	0.2	1	0.6811
10	10	0.1	2	0.9679	0.1	1	0.8212	0.0	1	0.9054
11	11	6.5	2	0.0381	6.5	1	0.0106	0.0	1	0.9625
12	12	18.2	2	0.0001	6.9	1	0.0086	11.2	1	0.0008
13	13	2.6	2	0.2789	1.2	1	0.2701	1.3	1	0.2477
14	14	1.8	2	0.4013	0.9	1	0.3435	0.9	1	0.3356
15	15	1.4	2	0.5082	1.3	1	0.2494	0.0	1	0.8738
16	16	6.9	2	0.0319	1.0	1	0.3250	5.9	1	0.0150
17	17	9.7	2	0.0079	1.1	1	0.2957	8.6	1	0.0034
18	18	0.2	2	0.8934	0.1	1	0.7507	0.1	1	0.7245
19	19	0.2	2	0.9171	0.1	1	0.7167	0.0	1	0.8390
20	20	2.9	2	0.2329	2.5	1	0.1127	0.4	1	0.5305
21	21	0.9	2	0.6266	0.9	1	0.3357	0.0	1	0.9336
22	22	2.0	2	0.3774	1.4	1	0.2326	0.5	1	0.4698

ตารางที่ ค-1 (ต่อ)

ข้อสอบข้อที่		Total X^2	*df*	*p*	X^2_a	*df*	*p*	$X^2_{c\|a}$	*df*	*p*
เพศชาย	เพศหญิง									
23	23	11.7	2	0.0029	0.0	1	0.8513	11.6	1	0.0006
24	24	6.5	2	0.0397	4.7	1	0.0293	1.7	1	0.1930
25	25	2.7	2	0.2617	0.3	1	0.6151	2.4	1	0.1195
26	26	4.4	2	0.1095	2.2	1	0.1367	2.2	1	0.1394
27	27	0.6	2	0.7334	0.2	1	0.6758	0.4	1	0.5048
28	28	4.0	2	0.1378	2.4	1	0.1216	1.6	1	0.2107
29	29	3.7	2	0.1612	3.1	1	0.0792	0.6	1	0.4446
30	30	1.2	2	0.5511	0.1	1	0.7288	1.1	1	0.3009
31	31	0.7	2	0.7065	0.3	1	0.6105	0.4	1	0.5096
32	32	4.6	2	0.0993	0.0	1	0.9638	4.6	1	0.0317
33	33	1.7	2	0.4337	1.5	1	0.2222	0.2	1	0.6726
34	34	4.7	2	0.0946	4.4	1	0.0354	0.3	1	0.5902
35	35	2.0	2	0.3602	2.0	1	0.1541	0.0	1	0.9333
36	36	1.5	2	0.4783	0.4	1	0.5329	1.1	1	0.2976
37	37	2.0	2	0.3677	2.0	1	0.1576	0.0	1	0.9935
38	38	0.5	2	0.7680	0.5	1	0.4680	0.0	1	0.9770
39	39	9.6	2	0.0084	0.4	1	0.5307	9.2	1	0.0025
40	40	0.4	2	0.8067	0.4	1	0.5278	0.0	1	0.8607

ภาคผนวก ง

ตัวอย่างผลการประมาณค่าพารามิเตอร์อำนาจจำแนกด้วยวิธี BAYSIAN

จากโปรแกรม Open BUGS

ตารางที่ ง-1 ผลการประมาณค่าพารามิเตอร์อำนาจจำแนกของแบบทดสอบ วิชาคณิตศาสตร์ จำแนกตามเพศด้วยวิธี BAYESIAN

ค่าอำนาจจำแนกข้อที่	Mean	*SD*	MC_error	val2.5pc	median	Val97.5pc	start	Sample
a[1]	1.03	0.06	0.00	0.91	1.038	1.16	2001	10000
a[2]	1.14	0.06	0.00	1.14	1.14	1.28	2001	10000
a[3]	1.03	0.06	0.00	0.92	1.03	1.51	2001	10000
a[4]	1.09	0.07	0.00	0.10	1.09	1.22	2001	10000
a[5]	1.33	0.07	0.00	1.19	1.33	1.50	2001	10000
a[6]	1.36	0.07	0.00	1.22	1.36	1.51	2001	10000
a[7]	0.89	0.56	0.00	0.78	0.89	1.00	2001	10000
a[8]	1.19	0.07	0.00	1.06	1.19	1.32	2001	10000
a[9]	0.82	0.05	0.00	0.72	0.82	0.94	2001	10000
a[10]	0.91	0.05	0.00	0.81	0.91	1.02	2001	10000
a[11]	0.65	0.05	0.00	0.54	0.64	0.75	2001	10000
a[12]	1.40	0.07	0.00	1.26	1.40	1.54	2001	10000
a[13]	1.21	0.06	0.00	1.09	1.21	1.34	2001	10000
a[14]	0.68	0.05	0.00	0.58	0.38	0.78	2001	10000
a[15]	0.81	0.054	0.00	0.71	0.81	0.92	2001	10000
a[16]	1.16	0.065	0.00	1.03	1.15	1.29	2001	10000
a[17]	1.00	0.05	0.00	0.89	1.00	1.11	2001	10000
a[18]	0.93	0.05	0.00	0.82	0.93	1.05	2001	10000
a[19]	0.91	0.57	0.00	0.81	0.91	1.03	2001	10000
a[20]	1.33	0.08	0.00	1.19	1.33	1.50	2001	10000
a[21]	1.36	0.08	0.00	1.20	1.36	1.53	2001	10000
a[22]	0.72	0.05	0.00	0.62	0.72	0.83	2001	10000
a[23]	0.94	0.58	0.00	0.82	0.94	1.05	2001	10000
a[24]	0.47	0.05	0.00	0.37	0.48	0.58	2001	10000
a[25]	0.91	0.60	0.00	0.80	0.91	1.03	2001	10000
a[26]	1.56	0.10	0.00	1.37	1.51	1.75	2001	10000

ตารางที่ ง-1 (ต่อ)

ค่าอำนาจจำแนกข้อที่	Mean	*SD*	MC_error	val2.5pc	median	Val97.5pc	start	Sample
a[27]	0.99	0.06	0.00	0.87	0.99	1.10	2001	10000
a[28]	0.96	0.57	0.00	0.85	0.96	1.08	2001	10000
a[29]	0.83	0.05	0.00	0.73	0.83	0.94	2001	10000
a[30]	0.11	0.01	0.00	0.09	0.10	0.14	2001	10000
a[31]	1.42	0.08	0.00	1.28	1.42	1588	2001	10000
a[32]	1.83	0.40	0.00	1.65	1.83	2.02	2001	10000
a[33]	1.83	0.010	0.00	1.65	1.83	2.02	2001	10000
a[34]	1.80	0.09	0.00	1.62	1.80	1.99	2001	10000
a[35]	2.17	0.11	0.00	1.96	2.16	2.38	2001	10000
a[36]	2.35	0.12	0.00	2.11	2.35	2.60	2001	10000
a[37]	2.43	0.129	0.00	2.11	2.43	2.70	2001	10000
a[38]	2.21	0.12	0.00	1.10	2.21	2.4	2001	10000
a[39]	1.40	0.08	0.00	1.22	1.38	1.53	2001	10000
a[40]	1.42	0.09	0.00	1.25	1.42	1.60	2001	10000

ภาคผนวก จ

ตัวอย่างผลการประมาณค่าพารามิเตอร์อำนาจจำแนกด้วยวิธี CFA

จากโปรแกรม Mplus

ตารางที่ จ-1 ผลการประมาณค่าพารามิเตอร์อำนาจจำแนกของแบบทดสอบ วิชาคณิตศาสตร์ จำแนกตามเพศ ด้วยวิธี CFA

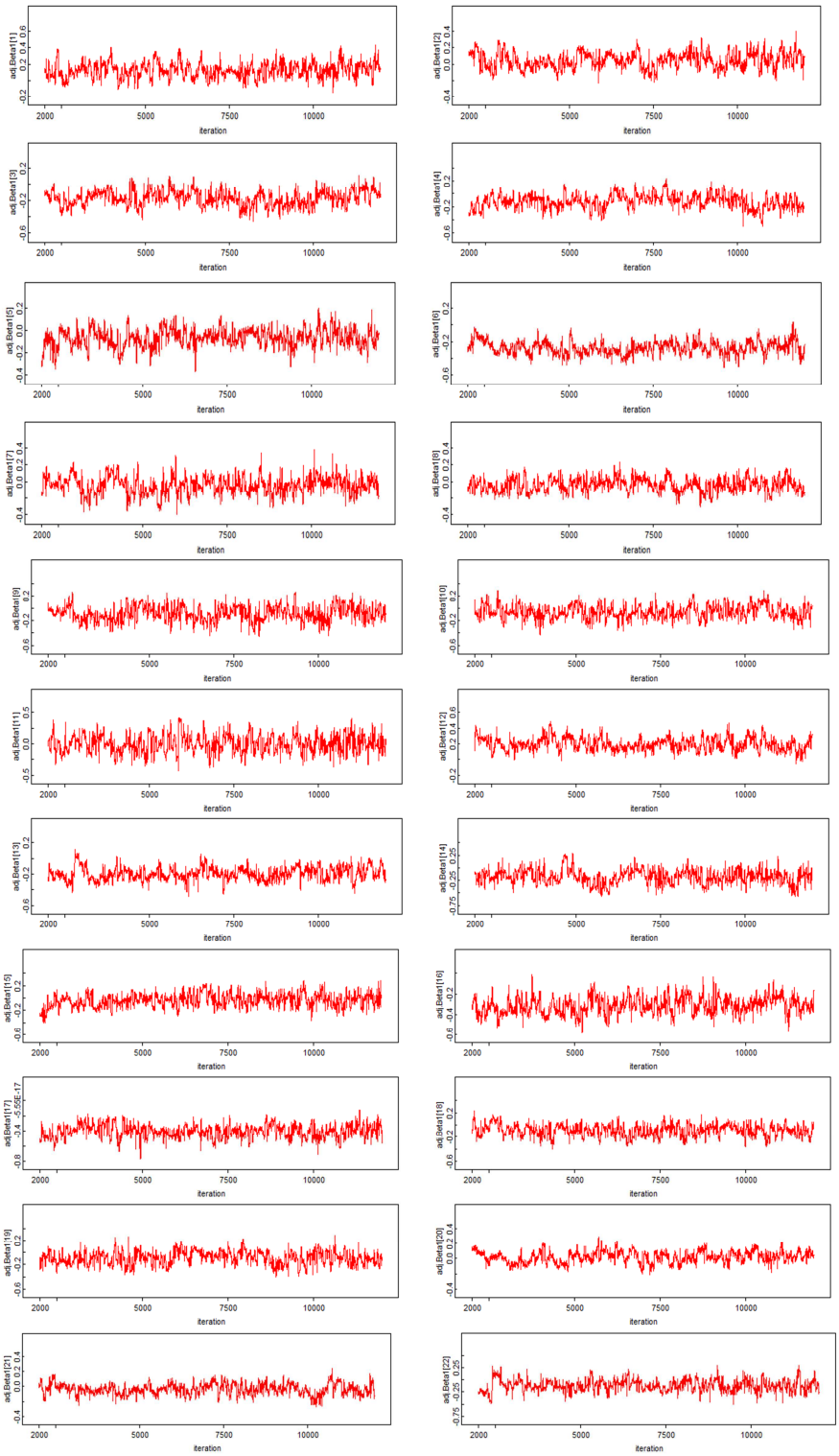
ข้อสอบข้อที่	Estimate	*SE*	*Est./ SE*	Two-Tailed
Q1	1.07	.06	17.11	0.00
Q2	1.18	.07	18.12	0.00
Q3	1.07	.06	17.43	0.00
Q4	1.12	.06	17.82	0.00
Q5	1.37	.07	19.09	0.00
Q6	1.40	.07	19.21	0.00
Q7	.92	.06	16.13	0.00
Q8	1.23	.07	18.35	0.00
Q9	.86	.06	15.30	0.00
Q10	.95	.06	16.25	0.00
Q11	.67	.05	13.15	0.00
Q12	1.43	.08	19.03	0.00
Q13	1.25	.07	18.57	0.00
Q14	.71	.05	13.67	0.00
Q15	.85	.06	15.37	0.00
Q16	1.19	.07	18.11	0.00
Q17	1.03	.06	17.06	0.00
Q18	.97	.06	16.49	0.00
Q19	.96	.06	16.46	0.00
Q20	1.38	.08	16.83	0.00
Q21	1.41	.09	15.79	0.00
Q22	.75	.06	13.61	0.00
Q23	.95	.06	16.47	0.00
Q24	.51	.05	19.97	0.00
Q25	.95	.06	16.40	0.00
Q26	1.60	.10	15.76	0.00

ตารางที่ จ-1 (ต่อ)

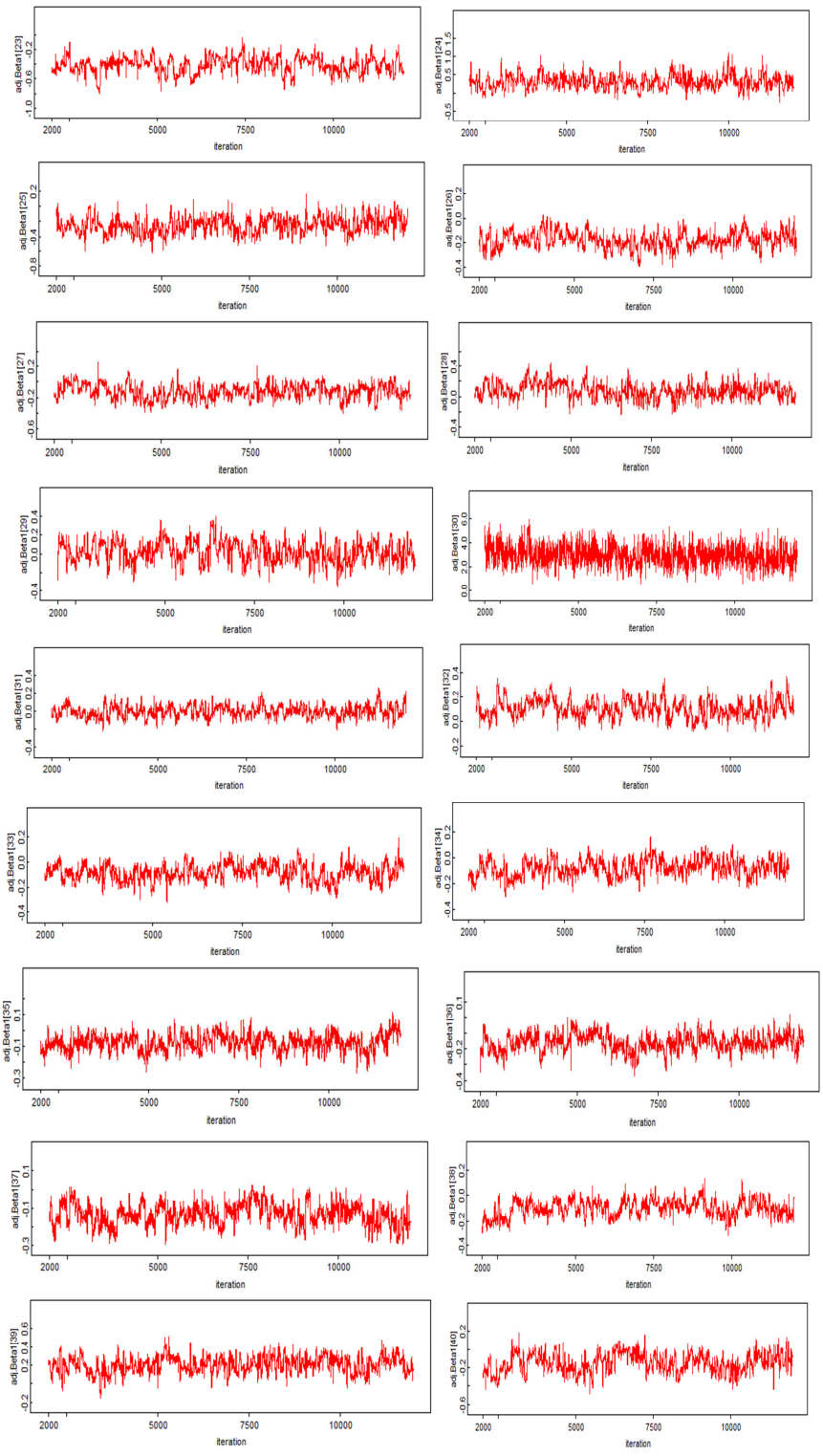
ข้อสอบข้อที่	Estimate	*SE*	*Est./ SE*	Two-Tailed
Q27	1.02	.06	16.86	0.00
Q28	.99	.06	16.78	0.00
Q29	.87	.06	15.66	0.00
Q30	-.25	.06	-4.318	0.00
Q31	1.49	.08	18.14	0.00
Q32	1.48	.08	17.99	0.00
Q33	1.90	.09	20.30	0.00
Q34	1.85	.09	20.25	0.00
Q35	2.24	.11	20.24	0.00
Q36	2.44	.12	20.18	0.00
Q37	2.51	.13	19.89	0.00
Q38	2.30	.11	20.06	0.00
Q39	1.39	.08	18.04	0.00
Q40	1.48	.09	16.97	0.00

ภาคผนวก ฉ

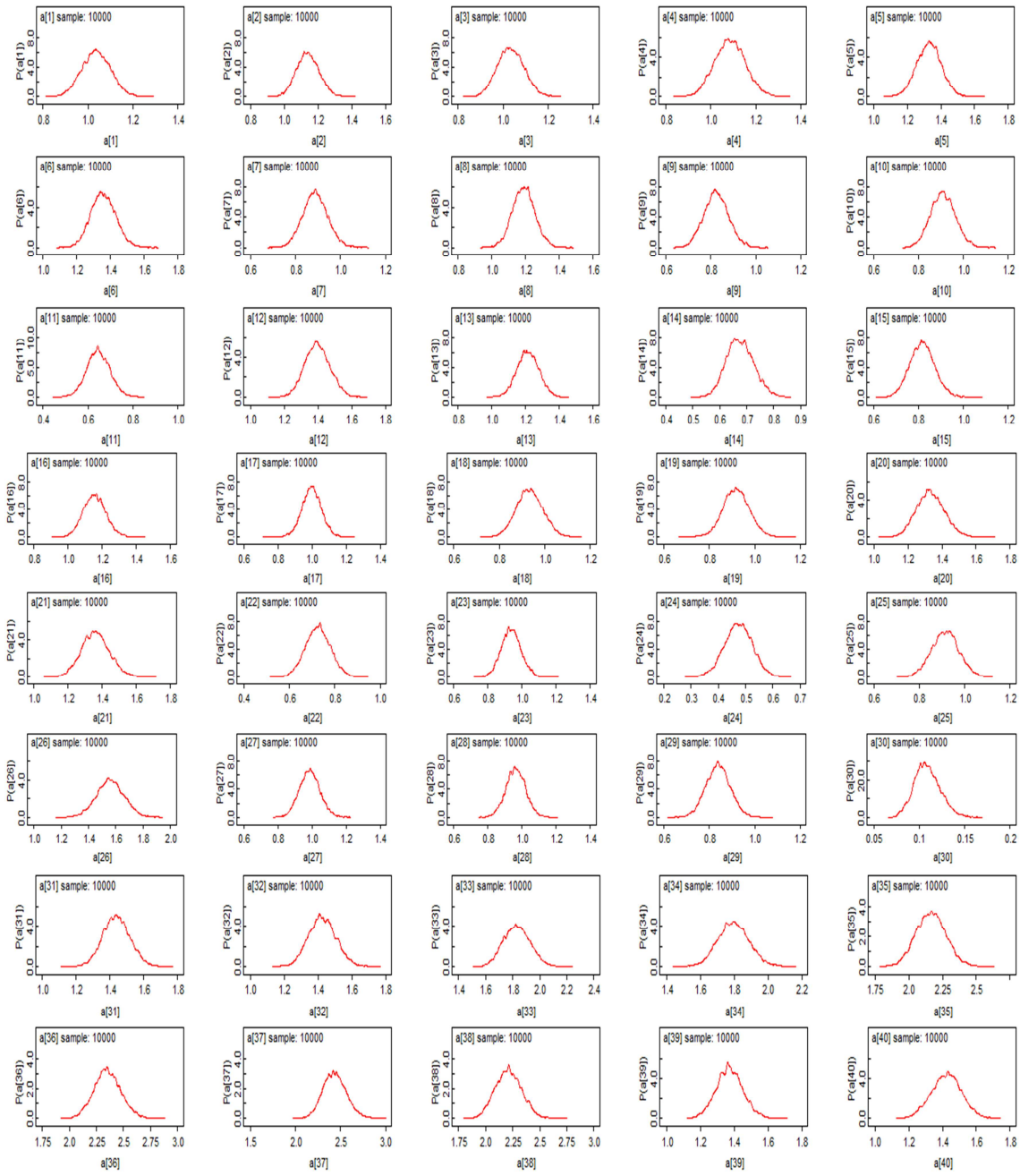
ตัวอย่าง History plot, Density plot ของพารามิเตอร์ที่ใช้ในการตัดสินการทำหน้าที่ต่างกัน (DIF) ของข้อสอบด้วยวิธี BAYSIAN



ภาพที่ ฉ-1 ตัวอย่าง History plot ของพารามิเตอร์ที่ใช้ในการตัดสิน DIF ที่ประมาณค่าด้วยวิธี BAYESIAN


adj.Beta1[23]
adj.Beta1[24]
adj.Beta1[25]
adj.Beta1[26]
adj.Beta1[27]
adj.Beta1[28]
adj.Beta1[29]
adj.Beta1[30]
adj.Beta1[31]
adj.Beta1[32]
adj.Beta1[33]
adj.Beta1[34]
adj.Beta1[35]
adj.Beta1[36]
adj.Beta1[37]
adj.Beta1[38]
adj.Beta1[39]
adj.Beta1[40]
iteration
2000
5000
7500
10000

ภาพที่ ฉ-1 (ต่อ)



ภาพที่ ฉ-2 ตัวอย่าง Density plot ของพารามิเตอร์ที่ใช้ในการตัดสิน DIF ที่ประมาณค่าด้วยวิธี BAYESIAN

ภาคผนวก ช

ผลการสกัดองค์ประกอบวิชาคณิตศาสตร์และวิทยาศาสตร์จำแนกตามเพศและเขตที่ตั้งทางภูมิศาสตร์ของโรงเรียน

ตารางที่ ช-1 การสกัดองค์ประกอบหลักวิชาคณิตศาสตร์ จำแนกเป็นเพศชาย

Component	Total variance explained[a]					
	Initial eigenvalues			Extraction sums of squared loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	10.596	26.490	26.490	10.596	26.490	26.490
2	1.986	4.965	31.455	1.986	4.965	31.455
3	1.372	3.430	34.885	1.372	3.430	34.885
4	1.152	2.880	37.765	1.152	2.880	37.765
5	1.080	2.701	40.466	1.080	2.701	40.466
6	.990	2.476	42.942			
7	.964	2.409	45.350			
8	.928	2.321	47.672			
9	.910	2.275	49.947			
10	.900	2.250	52.197			
11	.860	2.149	54.346			
12	.848	2.119	56.465			
13	.821	2.052	58.517			
14	.817	2.043	60.559			
15	.792	1.980	62.539			
16	.779	1.947	64.486			
17	.770	1.926	66.412			
18	.746	1.865	68.277			
19	.737	1.843	70.120			
20	.736	1.840	71.960			
21	.701	1.753	73.714			
22	.688	1.719	75.433			
23	.682	1.706	77.139			
24	.680	1.700	78.840			

ตารางที่ ช-1 (ต่อ)

Component	Total variance explained[a]					
	Initial eigenvalues			Extraction sums of squared loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
25	.656	1.640	80.480			
26	.640	1.601	82.081			
27	.635	1.587	83.667			
28	.617	1.544	85.211			
29	.594	1.486	86.697			
30	.579	1.448	88.145			
31	.566	1.415	89.559			
32	.544	1.360	90.919			
33	.534	1.336	92.255			
34	.503	1.259	93.514			
35	.492	1.229	94.743			
36	.479	1.198	95.941			
37	.429	1.071	97.012			
38	.425	1.063	98.076			
39	.393	.983	99.059			
40	.376	.941	100.000			

Extraction method: Principal component analysis.[a]

a. sex = 1

ตารางที่ ช-2 การสกัดองค์ประกอบหลักวิชาคณิตศาสตร์ จำแนกเป็นเพศหญิง

Component	Total variance explained[a]					
	Initial eigenvalues			Extraction sums of squared loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	8.935	22.336	22.336	8.935	22.336	22.336
2	2.380	5.950	28.286	2.380	5.950	28.286
3	1.367	3.418	31.704	1.367	3.418	31.704
4	1.204	3.009	34.713	1.204	3.009	34.713
5	1.094	2.736	37.449	1.094	2.736	37.449
6	1.007	2.516	39.965	1.007	2.516	39.965
7	.978	2.444	42.409			
8	.956	2.391	44.800			
9	.914	2.284	47.084			
10	.910	2.274	49.358			
11	.893	2.234	51.591			
12	.873	2.183	53.775			
13	.869	2.172	55.947			
14	.849	2.124	58.071			
15	.834	2.084	60.155			
16	.824	2.059	62.214			
17	.817	2.042	64.256			
18	.808	2.021	66.277			
19	.780	1.951	68.228			
20	.754	1.886	70.114			
21	.744	1.861	71.975			
22	.717	1.794	73.768			
23	.708	1.769	75.537			
24	.703	1.758	77.295			

ตารางที่ ช-2 (ต่อ)

Component	Total variance explained[a]					
	Initial eigenvalues			Extraction sums of squared loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
25	.691	1.728	79.023			
26	.683	1.707	80.730			
27	.662	1.656	82.386			
28	.650	1.624	84.010			
29	.630	1.576	85.586			
30	.619	1.547	87.133			
31	.614	1.535	88.668			
32	.585	1.463	90.131			
33	.567	1.417	91.549			
34	.541	1.352	92.901			
35	.527	1.317	94.218			
36	.512	1.280	95.498			
37	.489	1.222	96.720			
38	.462	1.156	97.876			
39	.442	1.105	98.981			
40	.408	1.019	100.000			

Extraction method: Principal component analysis.[a]

a. sex = 0

ตารางที่ ช-3 การสกัดองค์ประกอบหลักวิชาคณิตศาสตร์ จำแนกเป็นเขตกรุงเทพมหานครและปริมณฑล

Component	Total variance explained[a]						
	Initial eigenvalues			Extraction sums of squared loadings			Rotation sums of squared loadings
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	11.139	27.848	27.848	1.139	27.848	27.848	5.181
2	2.267	5.668	33.515	..267	5.668	33.515	5.061
3	1.335	3.338	36.854	..335	3.338	36.854	3.115
4	1.089	2.722	39.576	..089	2.722	39.576	2.447
5	1.027	2.567	42.143	..027	2.567	42.143	1.054
6	.959	2.397	44.540				
7	.933	2.333	46.873				
8	.901	2.252	49.125				
9	.874	2.186	51.311				
10	.859	2.149	53.460				
11	.855	2.138	55.598				
12	.829	2.072	57.670				
13	.806	2.016	59.686				
14	.800	1.999	61.685				
15	.782	1.956	63.642				
16	.759	1.897	65.539				
17	.751	1.878	67.416				
18	.732	1.830	69.246				
19	.723	1.806	71.052				
20	.709	1.772	72.824				
21	.688	1.719	74.544				
22	.662	1.655	76.198				

ตารางที่ ช-3 (ต่อ)

Component	Total variance explained[a]						
	Initial eigenvalues			Extraction sums of squared loadings			Rotation sums of squared loadings
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
23	.649	1.622	77.820				
24	.642	1.606	79.426				
25	.635	1.588	81.014				
26	.613	1.534	82.547				
27	.605	1.513	84.061				
28	.593	1.483	85.544				
29	.579	1.447	86.991				
30	.555	1.387	88.377				
31	.545	1.363	89.741				
32	.526	1.316	91.057				
33	.510	1.276	92.332				
34	.507	1.268	93.600				
35	.477	1.193	94.793				
36	.456	1.141	95.934				
37	.433	1.082	97.016				
38	.406	1.014	98.030				
39	.399	.998	99.028				
40	.389	.972	100.000				

ตารางที่ ช-4 การสกัดองค์ประกอบหลักวิชาคณิตศาสตร์ จำแนกเป็นนอกเขตกรุงเทพมหานครและปริมณฑล

Component	Total variance explained[a]						
	Initial eigenvalues			Extraction sums of squared loadings			Rotation sums of squared loadings
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	.241	18.104	18.104	7.241	18.104	18.104	4.227
2	2.132	5.330	23.434	2.132	5.330	23.434	2.588
3	1.355	3.386	26.820	1.355	3.386	26.820	2.124
4	1.284	3.211	30.031	1.284	3.211	30.031	1.737
5	1.123	2.807	32.838	1.123	2.807	32.838	1.500
6	1.077	2.694	35.531	1.077	2.694	35.531	1.414
7	1.052	2.629	38.160	1.052	2.629	38.160	1.330
8	1.043	2.608	40.768	1.043	2.608	40.768	1.265
9	1.009	2.524	43.292	1.009	2.524	43.292	1.132
10	.999	2.498	45.790				
11	.969	2.422	48.211				
12	.933	2.334	50.545				
13	.918	2.295	52.840				
14	.895	2.238	55.077				
15	.886	2.215	57.292				
16	.877	2.193	59.486				
17	.867	2.168	61.653				
18	.856	2.139	63.792				
19	.833	2.082	65.874				
20	.821	2.051	67.926				
21	.803	2.007	69.933				
22	.796	1.991	71.924				

ตารางที่ ช-4 (ต่อ)

Component	Total variance explained[a]						
	Initial eigenvalues			Extraction sums of squared loadings			Rotation sums of squared loadings
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
23	.773	1.932	73.856				
24	.765	1.913	75.769				
25	.748	1.870	77.639				
26	.738	1.845	79.484				
27	.701	1.754	81.237				
28	.678	1.694	82.932				
29	.667	1.668	84.600				
30	.657	1.642	86.243				
31	.645	1.612	87.854				
32	.627	1.567	89.421				
33	.618	1.544	90.965				
34	.588	1.470	92.435				
35	.586	1.466	93.901				
36	.550	1.376	95.277				
37	.504	1.260	96.537				
38	.472	1.181	97.718				
39	.461	1.151	98.869				
40	.452	1.131	100.000				

ตารางที่ ช-5 การสกัดองค์ประกอบหลักวิชาภาษาไทย จำแนกเป็นเพศชาย

Component	Total variance explained[a]								
	Initial eigenvalues			Extraction sums of squared loadings			Rotation sums of squared loadings		
	Total	% of Variance	Cumu lative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	11.673	14.592	14.592	11.673	4.592	14.592	8.670	10.837	10.837
2	7.794	9.743	24.334	7.794	9.743	24.334	6.733	8.416	19.253
3	1.837	2.296	26.631	1.837	2.296	26.631	2.482	3.102	22.355
4	1.554	1.943	28.573	1.554	1.943	28.573	2.034	2.542	24.897
5	1.267	1.584	30.157	1.267	1.584	30.157	1.871	2.338	27.236
6	1.208	1.510	31.666	1.208	1.510	31.666	1.506	1.883	29.118
7	1.181	1.476	33.142	1.181	1.476	33.142	1.435	1.793	30.912
8	1.170	1.462	34.604	1.170	1.462	34.604	1.401	1.751	32.663
9	1.139	1.424	36.029	1.139	1.424	36.029	1.386	1.732	34.395
10	1.129	1.411	37.440	1.129	1.411	37.440	1.321	1.651	36.046
11	1.088	1.361	38.801	1.088	1.361	38.801	1.295	1.619	37.665
12	1.083	1.354	40.155	1.083	1.354	40.155	1.291	1.614	39.279
13	1.057	1.321	41.475	1.057	1.321	41.475	1.284	1.605	40.884
14	1.038	1.297	42.773	1.038	1.297	42.773	1.214	1.517	42.401
15	1.019	1.274	44.046	1.019	1.274	44.046	1.213	1.516	43.917
16	1.007	1.258	45.304	1.007	1.258	45.304	1.110	1.387	45.304
17	.992	1.240	46.545						
18	.965	1.206	47.751						
19	.961	1.201	48.952						
20	.938	1.173	50.125						
21	.928	1.160	51.284						
22	.922	1.153	52.437						
23	.912	1.140	53.576						
24	.904	1.130	54.707						

ตารางที่ ช-5 (ต่อ)

Component	Total variance explained[a]								
	Initial eigenvalues			Extraction sums of squared loadings			Rotation sums of squared loadings		
	Total	% of Variance	Cumu lative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
25	.885	1.106	55.813						
26	.881	1.101	56.913						
27	.879	1.099	58.012						
28	.872	1.090	59.102						
29	.865	1.081	60.183						
30	.855	1.069	61.252						
31	.846	1.057	62.309						
32	.836	1.045	63.353						
33	.823	1.029	64.383						
34	.823	1.028	65.411						
35	.812	1.016	66.427						
36	.798	.998	67.424						
37	.789	.986	68.410						
38	.778	.973	69.383						
39	.772	.965	70.348						
40	.771	.964	71.312						
41	.749	.936	72.248						
42	.739	.924	73.172						
43	.736	.920	74.092						
44	.726	.908	75.000						
45	.710	.888	75.888						
46	.700	.875	76.763						
47	.698	.872	77.635						
48	.686	.857	78.492						

ตารางที่ ช-5 (ต่อ)

Component	Total variance explained[a]								
	Initial eigenvalues			Extraction sums of squared loadings			Rotation sums of squared loadings		
	Total	% of Variance	Cumu lative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
49	.682	.852	79.345						
50	.667	.834	80.179						
51	.654	.818	80.997						
52	.651	.814	81.811						
53	.649	.811	82.622						
54	.635	.794	83.416						
55	.626	.783	84.199						
56	.618	.773	84.972						
57	.605	.756	85.728						
58	.598	.748	86.476						
59	.594	.743	87.219						
60	.583	.729	87.948						
61	.575	.718	88.666						
62	.564	.705	89.371						
63	.556	.694	90.065						
64	.546	.683	90.748						
65	.537	.671	91.419						
66	.527	.659	92.079						
67	.517	.646	92.725						
68	.508	.635	93.360						
69	.497	.621	93.981						
70	.495	.618	94.600						
71	.491	.614	95.214						
72	.481	.601	95.815						

ตารางที่ ช-5 (ต่อ)

Component	Total variance explained[a]								
	Initial eigenvalues			Extraction sums of squared loadings			Rotation sums of squared loadings		
	Total	% of Variance	Cumu lative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
73	.474	.592	96.407						
74	.454	.568	96.975						
75	.431	.539	97.514						
76	.427	.534	98.049						
77	.416	.520	98.569						
78	.404	.505	99.074						
79	.377	.472	99.546						
80	.363	.454	100.000						

ตารางที่ ช-6 การสกัดองค์ประกอบหลักวิชาภาษาไทย จำแนกเป็นเพศหญิง

Component	Total variance explained[a]					
	Initial eigenvalues			Extraction sums of squared loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	10.450	13.063	13.063	10.450	13.063	13.063
2	8.079	10.099	23.162	8.079	10.099	23.162
3	1.866	2.332	25.494	1.866	2.332	25.494
4	1.496	1.869	27.363	1.496	1.869	27.363
5	1.326	1.657	29.020	1.326	1.657	29.020
6	1.277	1.596	30.617	1.277	1.596	30.617
7	1.198	1.497	32.114	1.198	1.497	32.114
8	1.170	1.463	33.577	1.170	1.463	33.577
9	1.155	1.443	35.020	1.155	1.443	35.020
10	1.128	1.410	36.429	1.128	1.410	36.429
11	1.103	1.379	37.808	1.103	1.379	37.808
12	1.081	1.352	39.160	1.081	1.352	39.160
13	1.063	1.328	40.489	1.063	1.328	40.489
14	1.053	1.316	41.804	1.053	1.316	41.804
15	1.042	1.302	43.106	1.042	1.302	43.106
16	1.026	1.282	44.389	1.026	1.282	44.389
17	1.015	1.269	45.657	1.015	1.269	45.657
18	1.000	1.250	46.908	1.000	1.250	46.908
19	.983	1.228	48.136			
20	.972	1.215	49.351			
21	.964	1.205	50.557			
22	.950	1.188	51.744			
23	.940	1.175	52.920			
24	.922	1.153	54.072			
25	.920	1.149	55.222			
26	.900	1.125	56.347			

ตารางที่ ช-6 (ต่อ)

Component	Total variance explained[a]					
	Initial eigenvalues			Extraction sums of squared loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
27	.898	1.122	57.469			
28	.883	1.104	58.573			
29	.872	1.090	59.662			
30	.866	1.083	60.745			
31	.858	1.073	61.818			
32	.840	1.051	62.868			
33	.832	1.041	63.909			
34	.819	1.024	64.933			
35	.812	1.015	65.948			
36	.809	1.011	66.960			
37	.798	.998	67.957			
38	.790	.987	68.945			
39	.775	.969	69.913			
40	.771	.964	70.877			
41	.754	.942	71.819			
42	.748	.935	72.755			
43	.743	.929	73.684			
44	.736	.921	74.604			
45	.725	.906	75.510			
46	.715	.893	76.403			
47	.702	.877	77.280			
48	.693	.866	78.147			
49	.685	.856	79.002			

ตารางที่ ช-6 (ต่อ)

Component	Total variance explained[a]					
	Initial eigenvalues			Extraction sums of squared loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
50	.674	.842	79.845			
51	.663	.829	80.673			
52	.662	.827	81.500			
53	.658	.823	82.323			
54	.645	.806	83.130			
55	.636	.795	83.924			
56	.623	.778	84.703			
57	.620	.775	85.478			
58	.613	.767	86.244			
59	.602	.753	86.997			
60	.596	.746	87.743			
61	.583	.729	88.471			
62	.573	.717	89.188			
63	.553	.691	89.879			
64	.547	.684	90.563			
65	.540	.675	91.238			
66	.538	.672	91.910			
67	.532	.665	92.575			
68	.522	.652	93.228			
69	.521	.651	93.879			
70	.514	.642	94.521			
71	.503	.629	95.150			
72	.489	.611	95.762			

ตารางที่ ช-6 (ต่อ)

Component	Total variance explained[a]					
	Initial eigenvalues			Extraction sums of squared loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
73	.483	.604	96.365			
74	.465	.582	96.947			
75	.462	.577	97.524			
76	.442	.552	98.077			
77	.420	.525	98.602			
78	.404	.505	99.107			
79	.365	.457	99.564			
80	.349	.436	100.000			

ตารางที่ ช-7 การสกัดองค์ประกอบหลักวิชาภาษาไทย จำแนกเป็นเขตกรุงเทพมหานครและปริมณฑล

Component	Total variance explained[a]					
	Initial eigenvalues			Extraction sums of squared loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	12.268	15.335	15.335	12.268	15.335	15.335
2	7.949	9.936	25.271	7.949	9.936	25.271
3	1.898	2.372	27.643	1.898	2.372	27.643
4	1.562	1.953	29.596	1.562	1.953	29.596
5	1.265	1.582	31.178	1.265	1.582	31.178
6	1.206	1.508	32.686	1.206	1.508	32.686
7	1.181	1.476	34.162	1.181	1.476	34.162
8	1.136	1.421	35.582	1.136	1.421	35.582
9	1.108	1.385	36.967	1.108	1.385	36.967
10	1.100	1.375	38.342	1.100	1.375	38.342
11	1.071	1.338	39.681	1.071	1.338	39.681
12	1.054	1.318	40.998	1.054	1.318	40.998
13	1.042	1.303	42.301	1.042	1.303	42.301
14	1.023	1.279	43.580	1.023	1.279	43.580
15	1.011	1.263	44.843	1.011	1.263	44.843
16	.992	1.240	46.083			
17	.985	1.231	47.314			
18	.978	1.222	48.536			
19	.967	1.209	49.745			
20	.944	1.180	50.925			
21	.927	1.159	52.084			
22	.920	1.150	53.234			
23	.911	1.139	54.373			
24	.897	1.121	55.494			

ตารางที่ ช-7 (ต่อ)

Component	Total variance explained[a]					
	Initial eigenvalues			Extraction sums of squared loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
25	.879	1.099	56.593			
26	.871	1.089	57.682			
27	.863	1.078	58.760			
28	.857	1.072	59.832			
29	.847	1.059	60.890			
30	.832	1.040	61.930			
31	.827	1.033	62.964			
32	.808	1.010	63.974			
33	.801	1.002	64.975			
34	.798	.998	65.973			
35	.796	.995	66.968			
36	.780	.975	67.943			
37	.773	.966	68.909			
38	.766	.958	69.867			
39	.756	.945	70.812			
40	.753	.941	71.753			
41	.737	.921	72.673			
42	.728	.910	73.584			
43	.717	.896	74.480			
44	.713	.891	75.371			
45	.700	.875	76.245			
46	.689	.861	77.106			
47	.684	.855	77.961			

ตารางที่ ช-7 (ต่อ)

Component	Total variance explained[a]					
	Initial eigenvalues			Extraction sums of squared loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
48	.670	.837	78.798			
49	.666	.832	79.630			
50	.660	.825	80.456			
51	.654	.817	81.273			
52	.647	.809	82.082			
53	.637	.797	82.879			
54	.631	.789	83.668			
55	.619	.774	84.441			
56	.610	.762	85.203			
57	.597	.746	85.949			
58	.588	.735	86.684			
59	.582	.728	87.412			
60	.580	.725	88.137			
61	.568	.709	88.846			
62	.563	.703	89.549			
63	.548	.685	90.235			
64	.541	.676	90.911			
65	.537	.671	91.582			
66	.528	.660	92.242			
67	.516	.645	92.887			
68	.508	.635	93.522			
69	.495	.619	94.141			
70	.482	.602	94.743			
71	.475	.594	95.337			

ตารางที่ ช-7 (ต่อ)

Component	Total variance explained[a]					
	Initial eigenvalues			Extraction sums of squared loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
72	.468	.585	95.922			
73	.464	.579	96.501			
74	.448	.560	97.061			
75	.434	.543	97.604			
76	.417	.521	98.125			
77	.414	.518	98.643			
78	.387	.484	99.127			
79	.361	.451	99.578			
80	.337	.422	100.000			

ตารางที่ ช-8 การสกัดองค์ประกอบหลักวิชาภาษาไทย จำแนกเป็นนอกเขตกรุงเทพมหานครและปริมณฑล

Component	Total variance explained[a]					
	Initial eigenvalues			Extraction sums of squared loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	9.289	11.612	11.612	9.289	11.612	11.612
2	8.043	10.054	21.666	8.043	10.054	21.666
3	1.794	2.242	23.908	1.794	2.242	23.908
4	1.463	1.829	25.738	1.463	1.829	25.738
5	1.274	1.592	27.330	1.274	1.592	27.330
6	1.255	1.568	28.898	1.255	1.568	28.898
7	1.206	1.508	30.406	1.206	1.508	30.406
8	1.186	1.483	31.889	1.186	1.483	31.889
9	1.163	1.453	33.342	1.163	1.453	33.342
10	1.143	1.429	34.770	1.143	1.429	34.770
11	1.098	1.373	36.143	1.098	1.373	36.143
12	1.091	1.363	37.507	1.091	1.363	37.507
13	1.088	1.360	38.867	1.088	1.360	38.867
14	1.070	1.338	40.204	1.070	1.338	40.204
15	1.056	1.320	41.524	1.056	1.320	41.524
16	1.041	1.302	42.826	1.041	1.302	42.826
17	1.026	1.282	44.108	1.026	1.282	44.108
18	1.014	1.267	45.376	1.014	1.267	45.376
19	1.010	1.263	46.638	1.010	1.263	46.638
20	1.000	1.250	47.888			
21	.985	1.232	49.119			
22	.968	1.210	50.330			
23	.958	1.198	51.527			
24	.954	1.193	52.720			

ตารางที่ ช-8 (ต่อ)

Component	Total variance explained[a]					
	Initial eigenvalues			Extraction sums of squared loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
25	.942	1.178	53.898			
26	.920	1.150	55.049			
27	.907	1.133	56.182			
28	.900	1.125	57.307			
29	.898	1.122	58.429			
30	.885	1.107	59.536			
31	.875	1.094	60.629			
32	.859	1.074	61.703			
33	.857	1.072	62.775			
34	.850	1.062	63.837			
35	.841	1.051	64.888			
36	.836	1.044	65.932			
37	.823	1.029	66.962			
38	.812	1.016	67.977			
39	.810	1.013	68.990			
40	.797	.996	69.985			
41	.785	.981	70.966			
42	.777	.971	71.937			
43	.764	.955	72.892			
44	.751	.939	73.831			
45	.747	.934	74.765			
46	.734	.918	75.683			
47	.728	.909	76.592			
48	.719	.899	77.491			

ตารางที่ ช-8 (ต่อ)

Component	Total variance explained[a]					
	Initial eigenvalues			Extraction sums of squared loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
49	.701	.876	78.367			
50	.696	.870	79.237			
51	.685	.856	80.093			
52	.682	.853	80.946			
53	.674	.843	81.788			
54	.658	.823	82.611			
55	.651	.814	83.425			
56	.645	.806	84.232			
57	.630	.788	85.019			
58	.628	.785	85.805			
59	.623	.779	86.584			
60	.617	.771	87.355			
61	.604	.754	88.109			
62	.592	.740	88.849			
63	.588	.735	89.584			
64	.574	.717	90.301			
65	.564	.705	91.006			
66	.559	.699	91.705			
67	.546	.683	92.388			
68	.533	.666	93.054			
69	.522	.652	93.706			
70	.515	.644	94.350			
71	.508	.636	94.986			
72	.501	.626	95.612			

ตารางที่ ช-8 (ต่อ)

Component	Total variance explained[a]					
	Initial eigenvalues			Extraction sums of squared loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
73	.490	.613	96.225			
74	.478	.597	96.822			
75	.466	.582	97.404			
76	.456	.570	97.974			
77	.436	.544	98.519			
78	.409	.511	99.029			
79	.392	.490	99.519			
80	.385	.481	100.000			

ประวัติย่อของผู้วิจัย

ชื่อ-สกุล	นางสาวเพ็ญศรี เทียมสุข
วัน เดือน ปี เกิด	11 กรกฎาคม พ.ศ. 2507
สถานที่เกิด	ตำบลเจ้าปลุก อำเภอมหาราช จังหวัดพระนครศรีอยุธยา
สถานที่อยู่ปัจจุบัน	บ้านเลขที่ 159 ซอย 4 ถนนพหลโยธิน
	อำเภอเมือง จังหวัดสระบุรี 18000
ประวัติการทำงาน	
พ.ศ. 2531-2537	พยาบาลวิชาชีพระดับปฏิบัติการ
	โรงพยาบาลวิหารแดง จังหวัดสระบุรี
พ.ศ. 2537-2547	พยาบาลวิชาชีพระดับชำนาญการ
	โรงพยาบาลหนองแค จังหวัดสระบุรี
พ.ศ. 2547-2556	พยาบาลวิชาชีพระดับชำนาญการ
	โรงพยาบาลเสาไห้เฉลิมพระเกียรติ 80 พรรษา
	จังหวัดสระบุรี
พ.ศ. 2556-ปัจจุบัน	พยาบาลวิชาชีพระดับชำนาญการ
	โรงพยาบาลสระบุรี จังหวัดสระบุรี
ประวัติการศึกษา	
พ.ศ. 2531	พยาบาลศาสตรบัณฑิต
	วิทยาลัยพยาบาลบรมราชชนนีสระบุรี
พ.ศ. 2548	วิทยาศาสตรมหาบัณฑิต (เวชศาสตร์ชุมชน)
	จุฬาลงกรณ์มหาวิทยาลัย
พ.ศ. 2560	ปรัชญาดุษฎีบัณฑิต (วิจัย วัดผล และสถิติการศึกษา)
	มหาวิทยาลัยบูรพา