การเปรียบเทียบประสิทธิผลการประมาณค่าพารามิเตอร์และการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมกซิมัมไลค์ลิฮู้ด วิธีของเบส์และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท

อริสฟา เตหลิ่ม

ดุษฎีนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปรัชญาดุษฎีบัณฑิต
สาขาวิชาวิจัย วัดผลและสถิติการศึกษา
คณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา
สิงหาคม 2559

คณะกรรมการควบคุมดุษฎีนิพนธ์และคณะกรรมการสอบดุษฎีนิพนธ์ ได้พิจารณาดุษฎีนิพนธ์ของ อริสพา เตหลิ่ม ฉบับนี้แล้ว เห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปรัชญาดุษฎีบัณฑิต สาขาวิชาวิจัย วัดผลและสถิติการศึกษา ของมหาวิทยาลัยบูรพาได้

คณะกรรมการควบคุมดุษฎีนิพนธ์

.....อาจารย์ที่ปรึกษาหลัก
(รองศาสตราจารย์ ดร.ไพรัตน์ วงษ์นาม)

..... อาจารย์ที่ปรึกษาร่วม
(ดร.สมพงษ์ ปั้นหุ่น)

คณะกรรมการสอบดุษฎีนิพนธ์

.....ประธาน
(ผู้ช่วยศาสตราจารย์ ดร.สังวรณ์ งัดกระโทก)

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

..... กรรมการ
(ดร.สมพงษ์ ปั้นหุ่น)

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

คณะศึกษาศาสตร์อนุมัติให้รับดุษฎีนิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปรัชญาดุษฎีบัณฑิต สาขาวิชาวิจัย วัดผลและสถิติการศึกษา ของมหาวิทยาลัยบูรพา

..... คณบดีคณะศึกษาศาสตร์
(รองศาสตราจารย์ ดร.วิชิต สุรัตน์เรืองชัย)
วันที่ 17 เดือน สิงหาคม พ.ศ. 2559

กิตติกรรมประกาศ

ดุษฎีนิพนธ์ฉบับนี้สำเร็จลงได้ด้วยความกรุณาจากรองศาสตราจารย์ ดร.ไพรัตน์ วงษ์นาม อาจารย์ที่ปรึกษาหลัก ดร.สมพงษ์ ปั้นหุ่น อาจารย์ที่ปรึกษาร่วม ที่กรุณาให้คำปรึกษาแนะนำ แนวทางตลอดจนแก้ไขข้อบกพร่องต่าง ๆ ด้วยความละเอียดถี่ถ้วนและเอาใจใส่เสมอมา ผู้วิจัยรู้สึกซาบซึ้งเป็นอย่างยิ่ง จึงขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ โอกาสนี้

ขอขอบคุณ ผู้ช่วยศาสตราจารย์ ดร.สังวรณ์ งัดกระโทก ประธานสอบปากเปล่า และผู้ช่วยศาสตราจารย์ ดร.สุรีพร อนุศาสนนันท์ กรรมการสอบปากเปล่า ที่ได้กรุณาให้ข้อเสนอแนะในการปรับปรุงแก้ไข ทำให้ดุษฎีนิพนธ์ฉบับนี้มีความสมบูรณ์ยิ่งขึ้น

ขอขอบคุณ Dr. Akihito Kamata และ Dr. Hirotaka Fukuhara ที่กรุณาให้คำแนะนำเกี่ยวกับการจำลองข้อมูล ขอขอบคุณ Dr. Alexander Robitzsch และ Dr. Margaret Wu ที่กรุณาให้คำแนะนำเกี่ยวกับการใช้งาน Packages TAM และให้ข้อแนะนำเกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบ และขอขอบคุณ Dr. Howard Wainer ที่กรุณามอบโปรแกรม SCORIGHT เพื่อใช้ในการศึกษาการประมาณค่าพารามิเตอร์ของข้อมูลที่มีอิทธิพลของเทสเลท

ขอขอบคุณ ผู้ช่วยศาสตราจารย์ ดร.สหัทยา รัตนะมงคลกุล ดร.อรทัย เจริญสิทธิ์ ดร.อำพล ชูสนุก Dr. Vir W., Mr. Ama L. พี่สิริกุล รัตนมณี พี่เสาวณีย์ สำราญสุข เพื่อน ๆ ภาควิชาวิจัยและจิตวิทยาประยุกต์ และเพื่อน ๆ กองทะเบียนฯ สำหรับการช่วยเหลือในขั้นตอนต่าง ๆ ระหว่างการวิจัย ผู้วิจัยพบความรักและน้ำใจมากมายจากดุษฎีนิพนธ์ฉบับนี้

สุดท้ายขอกราบขอบพระคุณ คุณแม่พัชรี คุณพ่อสุชาติ คุณตาแนบ คุณยายสุชิน คุณลุงอมรและครอบครัว ที่เป็นกำลังใจและสนับสนุนผู้วิจัยเสมอมา

คุณค่าและประโยชน์ของดุษฎีนิพนธ์ฉบับนี้ ผู้วิจัยขอมอบเป็นกตัญญูกตเวทิตาแด่บุพการี บูรพาจารย์ และผู้มีพระคุณทุกท่านทั้งในอดีตและปัจจุบัน ที่ทำให้ข้าพเจ้าเป็นผู้มีการศึกษาและประสบความสำเร็จมาจนตราบเท่าทุกวันนี้

อริสพา เตหลิ่ม

54810176: สาขาวิชา: วิจัย วัดผลและสถิติการศึกษา; ปร.ด. (วิจัย วัดผลและสถิติการศึกษา)
คำสำคัญ: การทำหน้าที่ต่างกันของข้อสอบ/ เทสต์เลท/ วิธีของเบส์

อริสพา เตหลิ่ม: การเปรียบเทียบประสิทธิผลการประมาณค่าพารามิเตอร์และการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมกซิมัมไลค์ลิฮู้ด วิธีของเบส์และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (COMPARING THE EFFECTIVENESS OF PARAMETER ESTIMATION AND DIFFERENTIAL ITEM FUNCTIONING AMONG MAXIMUM LIKELIHOOD, BAYESIAN, AND BAYESIAN TESTLET MODEL) กรรมการควบคุมดุษฎีนิพนธ์: ไพรัตน์ วงษ์นาม, ค.ด., สมพงษ์ ปั้นหุ่น, ค.ด. 230 หน้า. ปี พ.ศ. 2559.

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อ 1) ศึกษาประสิทธิผลในการประมาณค่าพารามิเตอร์ข้อสอบ (อำนาจจำแนกและความยาก) กับพารามิเตอร์ความสามารถของผู้สอบและ 2) เพื่อศึกษา ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ระหว่างวิธีแมกซิมัมไลค์ลิฮู้ด (ML) วิธีของเบส์ (Bayes) และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayes$_\gamma$) ข้อมูลที่ศึกษาเป็นข้อมูลจำลองภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ อิทธิพลเทสต์เลท (เท่ากันทุกเทสต์เลท แต่ละเทสต์เลทไม่เท่ากัน และข้อสอบที่เป็นอิสระผสมกับเทสต์เลท) การแจกแจงของความสามารถ (ปกติ เบ้ซ้าย เบ้ขวา) จำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบ (0, 5, 8 จากแบบสอบจำนวน 40 ข้อ) และอัตราส่วนของกลุ่มเปรียบเทียบต่อกลุ่มอ้างอิง (1000: 1000, 1000: 100) รวมจำนวนเงื่อนไขทั้งหมด 54 เงื่อนไข (3 x 3 x 3 x 2) กำหนดจำนวนรอบในการประมาณค่าพารามิเตอร์ในแต่ละเงื่อนไข 100 รอบ

ผลการวิจัยสรุปได้ ดังนี้

1. ผลการประมาณค่าพารามิเตอร์ จำแนกเป็น (1) พารามิเตอร์ความยาก พบว่า วิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayes$_\gamma$) ประมาณค่าได้ดีเมื่อข้อมูลมีการแจกแจงความสามารถเป็นแบบปกติ ส่วนวิธีของเบส์ (Bayes) ยังไม่มีแนวโน้มแน่นอน แต่จะประมาณค่าพารามิเตอร์ความยากได้ดีเป็นส่วนใหญ่เมื่อข้อมูลมีการแจกแจงความสามารถเป็นแบบเบ้ซ้าย และวิธีแมกซิมัมไลค์ลิฮู้ด (ML) จะประมาณค่าพารามิเตอร์ความยากได้ดี เมื่อข้อมูลมีการแจกแจงความสามารถเป็นแบบเบ้ขวา (2) พารามิเตอร์อำนาจจำแนก พบว่า ส่วนใหญ่วิธีการประมาณค่าด้วยวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayes$_\gamma$) จะประมาณค่าพารามิเตอร์อำนาจจำแนกดีกว่าเมื่อข้อมูลความสามารถมีการแจกแจงแบบปกติ (3) พารามิเตอร์ความสามารถ พบว่า วิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayes$_\gamma$) ประมาณค่าได้ดีที่สุด

2. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ พบว่า วิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayes$_\gamma$) และวิธีของเบส์ (Bayes) สามารถควบคุมอัตราความคาดเคลื่อนประเภทที่ 1 ได้ดี (ยกเว้นกรณีที่อิทธิพลเทสต์เลทมีค่าเป็น 0.25, 0.5, 1, 2 ร่วมกับการแจกแจงความสามารถที่เป็นแบบเบ้ซ้ายและเบ้ขวา การประมาณค่าด้วยวิธีของเบส์มีความคาดเคลื่อนประเภทที่ 1 สูง) และมีอำนาจการตรวจสอบสูงเมื่อมีการแจกแจงความสามารถแบบเบ้ซ้ายและจำนวนตัวอย่างมาก แต่ไม่มากถึงเกณฑ์ที่กำหนด ตรงข้ามกับวิธีแมกซิมัมไลค์ลิฮู้ด (ML) ซึ่งไม่สามารถควบคุมอัตราความคาดเคลื่อนประเภทที่ 1 แต่มีอำนาจการตรวจสอบสูง

54810176: MAJOR: EDUCATIONAL RESEARCH, MEASUREMENT AND STATISTICS; Ph.D. (EDUCATIONAL RESEARCH, MEASUREMENT AND STATISTICS)
KEYWORDS: DETECTING DIFFERENTIAL ITEM FUNCTIONING/ TESTLET/ BAYESIAN

ALISALA TAYLIM: COMPARING THE EFFECTIVENESS OF PARAMETER ESTIMATION AND DIFFERENTIAL ITEM FUNCTIONING AMONG MAXIMUM LIKELIHOOD, BAYESIAN, AND BAYESIAN TESTLET MODEL. DISSERTATION ADVISORY COMMITTEE: PAIRATTANA WONGNAM, Ph.D., SOMPONG PANHOON, Ph.D. 230 P. 2016.

The objectives of this research were: (1) to study the effectiveness of item parameter estimation (item difficulty and item discrimination) and person parameter estimation, and (2) to study the differential item functioning by using Maximum Likelihood (ML), Bayesian(Bayes), and Bayesian Testlet Model (Bayes$_\gamma$). In this study, the data were simulated under four variable conditions, they were: Testlet effects (equal effect, unequal effect, independent & testlet), Distributions ability (normal, negative, and positive skewness distributions), Levels of DIF (0, 5, and 8 items in the 40-item test length), and Levels of ratio of a reference and focal group (1000: 1000, 1000: 100). The entire total of testing conditions was 54 conditions (3 x 3 x 3 x 2). The total of 100 replications were performed to estimate the item parameters and statistical testing in each condition.

The research results were as follows:

1. Results of the estimation parameters were classified as (1) the difficulty parameters showed that the Bayesian Testlet Model (Bayes$_\gamma$) had the best estimator when there was normal distributions, the Bayesian (Bayes) estimates showed no definite trend but it will be a good estimation if the distributions of ability are negatively skewed, whereas Maximum Likelihood (ML) will be the best estimator if the distributions of ability are positively skewed. (2) the discrimination parameters showed that the Bayesian Testlet Model (Bayes$_\gamma$) had the best estimator when there is normal distributions. (3) the person parameters showed that the Bayesian Testlet Model (Bayes$_\gamma$) was the best estimator.

2. From the study of the detection of Differential Item Functioning, it was found that the Bayesian (Bayes), and Bayesian Testlet Model (Bayes$_\gamma$) estimate procedures had well-control of Type I error rate (except for magnitude of Testlet effect was set at level 0.25, 0.5, 1, 2 with skewed distributions, the Bayes$_\gamma$ had high Type I error rate) and had high power rate when negatively skewed ability distributions and sample sizes increased but it was not adequate for criteria, whereas Maximum Likelihood (ML) estimate procedures did not have control of Type I error rate but had high power of DIF detection.

สารบัญ

หน้า

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญตาราง (ต่อ)

ตารางที่ หน้า

สารบัญรูปภาพ

ภาพที่ หน้า

สารบัญรูปภาพ (ต่อ)

ภาพที่ หน้า

สารบัญรูปภาพ (ต่อ)

ภาพที่ หน้า

บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

จุดมุ่งหมายในการวัดทางด้านการศึกษาและจิตวิทยาที่สำคัญ คือ การวัดตัวแปรแฝง (Latent variable) เช่น ความสามารถในการอ่าน ความวิตกกังวล แรงจูงใจใฝ่สัมฤทธิ์ เป็นต้น จึงเกิดเป็นโมเดลการวัดขึ้นโดยมีสิ่งเร้าเป็นแบบสอบ โดยที่แบบสอบและการแปลความหมายของคะแนนเป็นที่รู้จักกันอย่างแพร่หลายและยาวนาน เริ่มตั้งแต่ปี ค.ศ. 1904 E.L. Thorndike ตีพิมพ์เกี่ยวกับทฤษฎีการวัดเป็นครั้งแรก ต่อมามีผู้ศึกษาต่อยอดจนเป็นที่รู้จักกันดีในชื่อทฤษฎีการทดสอบแบบดั้งเดิม (Classical test theory: CTT) ซึ่งแนวคิดนี้ มีลักษณะเป็นโมเดลเชิงเส้น (Linear model) ที่เชื่อว่าคะแนนที่สังเกตได้ (X) เป็นผลรวมของโมเดลเชิงเส้นระหว่างคะแนนจริงหรือความสามารถที่แท้จริงของผู้สอบ (T) กับความคลาดเคลื่อนที่เกิดขึ้นทั้งที่เป็นระบบและไม่เป็นระบบ (ε) โดยความคลาดเคลื่อนที่เกิดขึ้นนี้มีลักษณะเป็นหนึ่งเดียว ไม่สามารถแบ่งแยกได้ (Unique error) จากแนวคิดดังกล่าว เกิดเป็นข้อตกลงเบื้องต้นที่เกี่ยวข้องกับโมเดลการวัดและแบบสอบคู่ขนาน และมีวิธีการวิเคราะห์คุณภาพข้อสอบ ได้แก่ ค่าความยากของข้อสอบ (Item difficulty) ค่าอำนาจจำแนกของข้อสอบ (Item discrimination) และประสิทธิภาพตัวลวง (Item distractor) โดยที่หลักการวิเคราะห์คุณภาพของข้อสอบจะขึ้นอยู่กับจำนวนผู้เข้าสอบ และคุณลักษณะของผู้สอบ ทำให้มีข้อวิพากษ์เกี่ยวกับความถูกต้อง ความคงเส้นคงวาของผลการวิเคราะห์ข้อสอบที่มีความผันแปรไปตามกลุ่มผู้สอบ ถึงแม้ว่าผู้สอบทุกกลุ่มจะทำข้อสอบชุดเดียวกัน ซึ่งนักวัดผลเห็นว่าเป็นข้อตกลงเบื้องต้นที่ไม่สมเหตุสมผล (Weak assumption) และไม่สอดคล้องกับสภาพของการทดสอบในปัจจุบัน ไม่ว่าจะเป็นการกำหนดให้ค่าความคลาดเคลื่อนมีความเท่ากันในทุกผู้สอบ หรือข้อตกลงเกี่ยวกับความเป็นคู่ขนานของการทดสอบที่เป็นไปได้ยากในทางปฏิบัติ

ปัญหาการวัดความสามารถผู้สอบนี้ ได้รับการแก้ไขโดยใช้ทฤษฎีการวัดแนวใหม่ ซึ่งเป็นที่ยอมรับอย่างกว้างขวางในปัจจุบัน คือ ทฤษฎีการตอบสนองข้อสอบ (Item response theory: IRT) เป็นทฤษฎีที่พัฒนาขึ้นมาบนพื้นฐานของข้อตกลงเบื้องต้นที่มีความแกร่ง (Strong assumption) และมีความสอดคล้องกับการทดสอบจริงมากกว่าทฤษฎีการทดสอบแบบดั้งเดิม โดยทฤษฎีการตอบสนองข้อสอบ เป็นทฤษฎีที่อธิบายความสัมพันธ์ระหว่างความน่าจะเป็น

ในการตอบข้อสอบถูกกับคุณลักษณะหรือความสามารถที่แท้จริงของผู้สอบ อธิบายโดยใช้โค้งคุณลักษณะของข้อสอบ (Item characteristic curve: ICC) ซึ่งมีลักษณะเป็นฟังก์ชันทางคณิตศาสตร์ที่เรียกว่า ฟังก์ชันโลจิส (Logistic function) หรือใกล้เคียงกับฟังก์ชันปกติสะสม (Normal ogive function) ทำให้ค่าพารามิเตอร์ของข้อสอบที่ได้จากการวิเคราะห์คุณภาพของข้อสอบ ทั้งค่าความยาก (b) ค่าอำนาจจำแนก (a) และค่าโอกาสในการเดา (c) แต่ละข้อเป็นคุณลักษณะคงที่ ไม่แปรเปลี่ยนไปตามลักษณะกลุ่มผู้สอบ นอกจากนี้ค่าพารามิเตอร์ของผู้สอบหรือค่าความสามารถที่แท้จริงของผู้สอบยังเป็นคุณลักษณะที่มีอยู่ในตัวผู้สอบแต่ละคนที่ไม่เปลี่ยนแปลงไปตามคุณลักษณะของข้อสอบ ผลการวิเคราะห์จึงสามารถแก้ข้อจำกัดที่สำคัญของทฤษฎีการทดสอบแบบดั้งเดิมได้

ทฤษฎีการตอบสนองข้อสอบ มีการตกลงเกี่ยวกับการยอมรับความจริงเบื้องต้นข้อหนึ่ง นั่นคือ ความเป็นอิสระระหว่างข้อสอบและผู้สอบ (Item local independent) หมายถึง เมื่อควบคุมความสามารถ (Latent trait หรือ Ability) ที่ส่งผลต่อข้อสอบให้คงที่แล้ว ผลการตอบข้อสอบแต่ละข้อต้องเป็นอิสระกัน ซึ่งหากไม่คำนึงถึงข้อตกลงเบื้องต้นนี้แล้ว จะนำไปสู่ข้อผิดพลาดต่าง ๆ บ่อยครั้งข้อตกลงเบื้องต้นดังกล่าวมักขัดแย้งกับการนำไปใช้ โดยที่ผลการตอบข้อสอบข้อหนึ่งถูกจะมีผลต่อการตอบข้อสอบข้ออื่นถูกด้วย หรือเมื่อแบบสอบมีลักษณะเป็นเทสต์เลท (Testlet) เช่น ข้อสอบหลายข้อที่ต้องใช้ข้อมูลในการตอบจากกราฟ หรือแผนภูมิ หรือบทความเดียวกัน หรือแบบสอบที่เป็นการสอบทักษะการอ่าน (Reading comprehension test) ซึ่งแบ่งแบบสอบเป็นตอน ๆ ในตอนหนึ่งอาจประกอบด้วยข้อคำถาม 4 - 12 ข้อ ดังนั้น ผลการตอบสนองของข้อคำถามเหล่านี้ขึ้นอยู่กับผู้สอบเข้าใจบทความที่ใช้เป็นคำถามมากน้อยเพียงใด หรือ เมื่อมีการสุ่มเลือกข้อสอบในกรณีของการสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (Computerized adaptive testing: CAT) อาจเกิดสถานการณ์ที่ผลการตอบของข้อสอบข้อหนึ่งมาจากโจทย์ของอีกข้อหนึ่ง (Cross - information) ซึ่งแสดงถึงความไม่เป็นอิสระของการตอบคำถามในแต่ละข้อ เช่น ในมุมของทฤษฎีการทดสอบแบบดั้งเดิม หากเกิดความไม่เป็นอิสระระหว่างข้อสอบและผู้สอบ เนื่องจากเทสต์เลท (Testlet) แล้วทำให้ประมาณค่าความคาดเคลื่อนมาตรฐานของการวัด (Standard error of measurement) น้อยเกินจริง ซึ่งทำให้การประมาณค่าความเที่ยง (Reliability) สูงเกินจริง ส่วนในมุมของทฤษฎีการตอบสนองข้อสอบ (IRT) จะส่งผลให้ประมาณค่าสารสนเทศของแบบสอบสูงขึ้น นั่นหมายถึงมีการประมาณค่าความคลาดเคลื่อนมาตรฐานต่ำเกินจริง นอกจากนี้ ยังเกิดอคติในการประมาณค่าความยากและค่าอำนาจจำแนกของข้อสอบด้วย Sireci, Thissen & Wainer (1991) อธิบายว่าการละเลยความเป็นอิสระระหว่างข้อสอบและผู้สอบ ทำให้การประมาณค่าความเที่ยงเกินจริงประมาณ 10-15% และเพื่อให้โมเดล IRT ประมาณค่าได้ถูกต้องจะต้องเพิ่มความยาวของแบบสอบ

เป็น 2 เท่าเพื่อชดเชยค่าความเที่ยงที่ประมาณเกินจริงไป ซึ่งข้อมูลที่เกิดเหล่านี้มีโอกาสที่จะสรุปผลผิดพลาด หากการตัดสินใจของการสรุปผลนั้นขึ้นอยู่กับค่าที่ประมาณได้ (Sireci et al.,1991; Yen, 1993; Sedivy, 2009; Christine, 2012; สุนทร เทียนงาม, ศิริชัย กาญจนาวสี และดิเรก ศรีสุข, 2553) ดังนั้น จึงไม่เหมาะสมหากใช้โมเดล IRT ในการอธิบายแบบสอบลักษณะนี้ (Sedivy, 2009; Wang, Bradlow & Wainer, 2002)

สุนทร เทียนงาม และคณะ (2553) ให้ข้อเสนอแนะเกี่ยวกับการสร้างแบบสอบให้เป็นตามข้อตกลงเบื้องต้นของทฤษฎีการตอบสนองข้อสอบนั้น เป็นสิ่งสำคัญ เนื่องจากถ้าแบบสอบมีคุณภาพจะส่งผลทำให้การประมาณค่าพารามิเตอร์ข้อสอบ ค่าความสามารถเป็นค่าที่ถูกต้องแท้จริง ดังนั้น จึงจำเป็นต้องเลือกใช้โมเดลการวิเคราะห์ที่เหมาะกับจุดมุ่งหมายของการทดสอบ เช่น แบบสอบทั่วไป ที่มีลักษณะข้อสอบที่โอกาสเกิดความไม่เป็นอิสระของข้อสอบน้อย อาจเลือกใช้โมเดลการวิเคราะห์แบบ 2 หรือ 3 พารามิเตอร์ได้ แต่ในสถานการณ์ของการวิเคราะห์แบบสอบที่มีโอกาสเสี่ยงต่อการเกิดความไม่เป็นอิสระของข้อสอบสูง เช่น แบบสอบที่ถามคำถามในเนื้อหาเดียวกัน การใช้แผนภูมิ กราฟ เพื่อถามคำถามหลายข้อ หรือที่เรียกว่าแบบสอบมีลักษณะเป็นเทสต์เลท (Testlet) อาจเลือกใช้โมเดลการวิเคราะห์แบบ 1 พารามิเตอร์ได้ ซึ่งจะเหมาะสมกว่าจะเกิดความคลาดเคลื่อนในการประมาณค่าน้อยกว่า อย่างไรก็ตามการวิเคราะห์ด้วยโมเดลการวิเคราะห์ 1 พารามิเตอร์นั้น ไม่มีค่าอำนาจจำแนกและค่าโอกาสการเดาเข้ามาใช้ในการคำนวณค่าความสามารถ จึงทำให้ค่าความสามารถไม่ตรงกับความเป็นจริง

เนื่องจากข้อสอบที่มีลักษณะเป็นเทสต์เลท (Testlet) หรือบางครั้งเรียกว่ากลุ่มข้อสอบ (Item Bundle) เป็นการให้คะแนนโดยที่ใน 1 เทสต์เลทจะมีค่าน้อยกว่าคะแนนเต็มในแบบสอบ เช่น แบบสอบที่มีคะแนนเต็ม 10 คะแนน คะแนนใน 1 เทสต์เลท จะได้ไม่ถึง 10 คะแนน ข้อสอบที่อยู่ภายในเทสต์เลท มีความไม่เป็นอิสระในการตอบในแต่ละข้อ เนื่องจากถูกกระตุ้นจากสิ่งเร้า หรือโจทย์เดียวกัน นอกจากนี้ ความไม่เป็นอิสระในการตอบยังถือเป็นมิติที่เพิ่มขึ้นในโครงสร้างของแบบสอบด้วย ดังนั้น ความท้าทายของผู้สร้างแบบสอบไม่ได้อยู่ที่การกำจัดความไม่เป็นอิสระในการตอบออกไป แต่เป็นการค้นหาคำตอบ เช่น พารามิเตอร์ผู้สอบ พารามิเตอร์ข้อสอบ ที่เหมาะสม โดยที่ความไม่เป็นอิสระในการตอบไม่ไปมีผลกระทบต่อความตรง (Validity) และความเที่ยงของแบบสอบ (Reliability)

ลักษณะแบบสอบที่จะเกิดความไม่เป็นอิสระมักเป็นแบบสอบที่มีลักษณะเป็นเทสต์เลท ซึ่งมีแหล่งความแปรปรวนและปัจจัยอันเนื่องจากเทสต์เลทเพิ่มขึ้น ดังนั้น Wainer, Bradlow, & Du (2000) จึงเสนอทฤษฎีการตอบสนองข้อสอบที่มีลักษณะของเทสต์เลท (Testlet response theory: TRT) ซึ่งเป็น Parametric Bayesian Model สำหรับแบบสอบที่มีการให้คะแนนแบบสองค่าต่อเนื่อง

กับการให้คะแนนผสมกันด้วยการให้คะแนนแบบสองค่าระหว่างรายข้อที่เป็นอิสระกัน (Independent) กับข้อสอบแบบเทสต์เลท พัฒนาจากโมเดลตอบสนองข้อสอบ ทั้งแบบ 2 และ 3 พารามิเตอร์ โดยการเพิ่มส่วนของปฏิสัมพันธ์ของการตอบกับเทสต์เลทเข้าไป หรือเรียกว่า อิทธิพลของเทสต์เลท (Testlet effect) ซึ่งหมายถึง ถ้าละเลยความเป็นอิสระของข้อสอบจะทำให้เกิดอคติ (Bias) ในการประมาณค่าทั้งพารามิเตอร์ของผู้สอบและพารามิเตอร์ของข้อสอบ

เพื่อให้สอดคล้องกับสถานการณ์และพอดี (Fit) กับข้อมูล นักวิจัยหลายท่านจึงพัฒนาโมเดลและวิธีการต่าง ๆ ในการวิเคราะห์ข้อสอบที่มีลักษณะเทสต์เลทหลายวิธี เช่น การนำวิธีวิเคราะห์ข้อสอบแบบ Polytomous มาประยุกต์ใช้ (Sireci et al., 1991; Thissen, Steinberg, & Mooney, 1989; Wainer, 1995 cited in Zhang, 2010) โดยมีแนวคิด คือ กำจัดความไม่เป็นอิสระในการตอบออกไป โดยรวมคะแนนข้อที่อยู่ในเทสต์เลท (Testlet) เดียวกันให้เป็นข้อสอบแบบให้คะแนนหลายค่า (Polytomous) 1 ข้อ เช่น Graded Response Model (Samejima, 1968) Partial Credit Model (Masters, 1982) Rating Scale Model (Andrich, 1978) Nominal Response Model (Bock, 1972)

การใช้ Polytomous IRT Model ในการวิเคราะห์ข้อมูลแบบสอบที่มีลักษณะของเทสต์เลทนั้น จะกระทำได้โดยที่กำจัดความไม่เป็นอิสระในการตอบออกไปได้ ซึ่งวิธีนี้ทำให้หลีกเลี่ยงการประมาณค่าความเที่ยงของแบบสอบเกินจริง และสารสนเทศทางสถิติของ Polytomous IRT Model ยังมีความคงเส้นคงว่าดีกว่าการใช้ Standard Rasch Model อีกด้วย สอดคล้องกับงานวิจัยของ Zhang (2010) ได้ทำการเปรียบเทียบประสิทธิภาพการประมาณค่าพารามิเตอร์ผู้สอบ (ความสามารถ) กรณีที่กลุ่มตัวอย่างมีขนาดเล็ก 3 โมเดล ได้แก่ โมเดล Standard Rasch โมเดล Partial Credit และโมเดล Rasch Testlet โดยใช้การจำลองข้อมูลด้วยโปรแกรม R ประมาณค่าพารามิเตอร์ด้วยวิธีแมกซิมัมไลค์ลิฮู้ด (Maximum likelihood: ML) ด้วยโปรแกรม ConQuest ผลการวิจัยพบว่า ทั้งโมเดล Partial Credit และโมเดล Rasch Testlet มีประสิทธิภาพดีกว่าโมเดล Standard Rasch ผลการวิจัยยังชี้ให้เห็นว่า เมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้น ทำให้ความต่างของค่าพารามิเตอร์ความสามารถจากการประมาณค่ากับค่าจริงต่างกันมากขึ้นด้วย และเมื่อเปรียบเทียบการประมาณค่าพารามิเตอร์ของ Polytomous IRT Model กับ Standard Rasch Model พบว่า เมื่อในแบบสอบมีจำนวนเทสต์เลทมากขึ้น การประมาณค่าพารามิเตอร์โดยใช้ Polytomous IRT Model มีความคงเส้นคงว่าดีกว่าการใช้ Standard Rasch Model

แต่อย่างไรก็ตาม การใช้ Polytomous IRT Model มีจุดอ่อน คือ 1) สารสนเทศการตอบของผู้สอบที่เป็นรายข้อจะหายไป 2) พารามิเตอร์ของข้อสอบบางข้อจะหายไปเมื่อเทียบกับข้อสอบ

แบบ Dichotomous 3) ไม่มีข้อมูลเพื่อนำไปปรับปรุงแบบสอบ 4) อาจจะประมาณค่าความเที่ยงของแบบสอบต่ำกว่าจริง (Yen, 1993)

นอกจากนี้ การตอบสนองของข้อสอบแบบเทสต์เลท สามารถนำ Multidimensional Model มาประยุกต์ใช้ โดย Li, Bolt & Fu (2006) ได้ประยุกต์โดยนำ Bi-factor Multidimensional Item Response Theory Model (Bi-factor MIRT model) มาใช้วิเคราะห์ ซึ่งโมเดลนี้ประกอบด้วย 2 คุณลักษณะ ได้แก่ คุณลักษณะหลัก (Primary trait) และคุณลักษณะที่ 2 เป็นคุณลักษณะของเทสต์เลท (Testlet trait) หรืออธิบายได้ว่าโมเดลนี้เป็นกรณีพิเศษ (Special case) ของ Bi-factor MIRT model ซึ่งสำหรับข้อมูลที่มีความไม่เป็นอิสระในการตอบโดยมีสาเหตุจากเทสต์เลท (Testlet) หากทำการประมาณค่าพารามิเตอร์โดยใช้ Bi-factor MIRT model จะมีความพอดี (Fit) กับข้อมูลมากกว่าการประมาณค่าพารามิเตอร์โดยใช้ทฤษฎีการตอบสนองข้อสอบที่มีลักษณะของเทสต์เลท (Testlet Response Theory: TRT) (Li et al., 2006; DeMars, 2006 cited in Fukuhara & Kamata, 2011)

นอกจากนี้ ยังมีการประยุกต์ใช้โมเดลต่าง ๆ เช่น การวิเคราะห์พหุระดับ (Hierarchical generalized linear model) หรือ HGLM (Fukuhara & Kamata, 2011) เป็นการวิเคราะห์ตามโมเดลเชิงเส้นทั่วไป (Generalized linear model: GLM) แล้วใช้ฟังก์ชั่นโยง (Link function) ที่เป็นฟังก์ชั่นแบบโลจิท (Logit link function) ในการปรับค่าเฉลี่ยของการวิเคราะห์ระดับที่ 1 นำมาสู่การวิเคราะห์ในระดับต่อไปได้ โดยการวิเคราะห์ระดับที่ 1 ตัวแปรตามจึงเป็น log-odds ของความน่าจะเป็นในการตอบข้อสอบได้ถูก การวิเคราะห์นี้จัดโมเดลการวิเคราะห์เป็น 3 ระดับ โดยการวิเคราะห์ในระดับที่ 1 ระดับข้อสอบเป็นการจัดให้ข้อสอบสอดแทรกในเทสต์เลท (between item within Testlet) การวิเคราะห์ระดับที่ 2 ระดับเทสต์เลท เป็นการจัดให้เทสต์เลทสอดแทรกในผู้สอบ (between Testlet within person) และการวิเคราะห์ระดับที่ 3 ระดับผู้สอบ เป็นการวิเคราะห์ระหว่างผู้สอบ (between person) จากการศึกษาของ Kamata (2001) พบว่า ค่าความน่าจะเป็นที่บุคคลจะตอบข้อสอบได้ถูกต้องของโมเดลการวิเคราะห์ข้อสอบแบบพหุระดับ จะเป็นสมการคู่ขนาน (Equivalent) กับค่าความน่าจะเป็นที่บุคคลจะตอบข้อสอบได้ถูกต้องของโมเดลราสช์ (Rasch model) ดังนั้น จึงสามารถประยุกต์นำแนวคิดการวิเคราะห์แบบพหุระดับมาใช้ในการวิเคราะห์ข้อสอบที่แบ่งการวิเคราะห์ออกเป็นระดับต่าง ๆ ในลักษณะเดียวกันได้ แต่สามารถวิเคราะห์ข้อสอบได้เพียง 1 พารามิเตอร์

เนื่องจากที่ผ่านมามีการศึกษาพารามิเตอร์และศึกษาการทำหน้าที่ต่างกันของข้อสอบจำนวนมาก ซึ่งสาระสำคัญของการศึกษาการทำหน้าที่ต่างกันของข้อสอบ คือ พารามิเตอร์ของข้อสอบที่ประมาณค่าจะเป็นตัวที่ถูกนำมาใช้ในการตัดสินการทำหน้าที่ต่างกันของข้อสอบ เช่น

วิธีการของ Lord ที่ใช้การเปรียบเทียบค่าพารามิเตอร์ของข้อสอบระหว่างกลุ่มย่อย 2 กลุ่ม ดังนั้น การประมาณค่าพารามิเตอร์โดยไม่คำนึงว่ามีอิทธิพลของเทสต์เลท อาจจะทำให้การตัดสินการทำหน้าที่ต่างกันของข้อสอบคาดเคลื่อน ประกอบกับ ความเข้าใจเกี่ยวกับประสิทธิผลของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเทสต์เลทยังไม่กว้างขวาง เนื่องจากยังไม่มีข้อมูลการเปรียบเทียบระหว่างวิธีต่าง ๆ ส่งผลต่อความถูกต้องของการตัดสินใจหรือไม่ ดังนั้น ผู้วิจัยจึงสนใจประเด็นของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเทสต์เลท โดยหากวิเคราะห์แบบดั้งเดิมแล้วจะให้ผลการตรวจสอบแตกต่างกับการตรวจสอบแบบการใช้โมเดลที่มีอิทธิพลของเทสต์เลทหรือไม่

ซึ่งสำหรับการหาคุณภาพของแบบสอบที่วิเคราะห์ค่าพารามิเตอร์ของข้อสอบด้วยโมเดลต่าง ๆ สิ่งที่ต้องคำนึงถึงในการพัฒนาแบบสอบ คือ แบบสอบต้องมีความยุติธรรมกับผู้สอบทุกกลุ่ม ไม่เข้าข้างกลุ่มใดกลุ่มหนึ่งหรือทำให้กลุ่มใดกลุ่มหนึ่งเสียประโยชน์ ข้อสอบที่เอื้อต่อผู้สอบกลุ่มใดกลุ่มหนึ่งอาจทำให้ผู้สอบกลุ่มดังกล่าวมีความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องกว่ากลุ่มอื่นที่ระดับความสามารถเดียวกัน ลักษณะการทำหน้าที่ของข้อสอบประเภทนี้เรียกว่า **การทำหน้าที่ต่างกันของข้อสอบ** (Differential item functioning: DIF) ซึ่งเดิมเรียกว่าความลำเอียงของข้อสอบ (Item bias) เพราะวิธีการที่ใช้ในการตรวจสอบ ความลำเอียงนั้น เน้นความแตกต่างระหว่างกลุ่มของผู้สอบที่มีการตอบสนองต่อข้อสอบในข้อเดียวกัน ซึ่งความแตกต่างนี้ อาจเกิดจากข้อสอบประสบการณ์ พื้นฐานเดิมที่ต่างกันของกลุ่มผู้สอบและความคลุมเครือในการใช้เกณฑ์เพื่อตัดสินความลำเอียงของข้อสอบ จึงนิยมใช้สารสนเทศทางสถิติมาเป็นเกณฑ์ในการตัดสินใจ ด้วยเหตุนี้จึงนิยมใช้คำว่า การทำหน้าที่ต่างกันของข้อสอบเพราะเป็นคำที่มีความเป็นกลางและเหมาะสมกว่า (Holland & Wainer, 1993)

การตรวจสอบเกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบระหว่างผู้สอบกลุ่มย่อยตั้งแต่ 2 กลุ่มขึ้นไป เริ่มตั้งแต่ปี ค.ศ. 1951 โดยเป็นการศึกษาเปรียบเทียบระหว่างผู้สอบที่มีความต่างกันทางด้านเศรษฐกิจ สังคม เพศ วัฒนธรรมและเชื้อชาติ ระดับสติปัญญา วิธีการสอน (กาญจนา วัธนสุนทร, 2537) นอกจากจะศึกษาปัจจัยจากลักษณะของผู้สอบที่ทำให้เกิดการทำหน้าที่ต่างกันของข้อสอบแล้ว ระยะหลังได้มีการศึกษาเปรียบเทียบวิธีการในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เนื่องจากมีการพัฒนาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบหลายวิธี และแต่ละวิธีมีแนวคิด ข้อดี ข้อเสียต่างกัน โดยที่คมศักดิ์ ชื่นชม (2539) จำแนกวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามทฤษฎีพื้นฐาน ได้เป็น 2 กลุ่ม ดังนี้

กลุ่มที่ 1 กลุ่มที่ใช้หลักการของทฤษฎีการตอบสนองข้อสอบ (IRT) ได้แก่ วิธีโค้งคุณลักษณะข้อสอบแบบ 3 พารามิเตอร์ (Item characteristic curve - 3 parameter: ICC - 3) วิธีโค้ง

คุณลักษณะข้อสอบแบบ 2 พารามิเตอร์ (Item characteristic curve - 2 parameter: ICC - 2) วิธีโค้งคุณลักษณะข้อสอบแบบ 1 พารามิเตอร์ (Item characteristic curve - 1 parameter: ICC - 1) วิธี Likelihood Ratio Test วิธี SIBTEST และ วิธี Lord's χ^2 test

กลุ่มที่ 2 กลุ่มที่ใช้หลักการของทฤษฎีการทดสอบแบบดั้งเดิม (Classical test theory: CTT) ได้แก่ วิธีแปลงค่าความยากของข้อสอบ (Transformed item difficulty: TID) วิธีวิเคราะห์ความแปรปรวน (Analysis of variance: ANOVA) วิธีวิเคราะห์ด้วยไคสแคว์ (Chi - square: χ^2) วิธีวิเคราะห์องค์ประกอบ (Factor analysis) วิธีวิเคราะห์การถดถอย (Regression analysis) วิธีค่าอำนาจจำแนกของข้อสอบ (Item discrimination indices) วิธีลอกลิเนียร์ (Log-linear) วิธีแมนเทล-เฮนส์เซล (Mantel - Haenszel: MH) วิธีทำให้เป็นมาตรฐาน (Standardization: STND) และวิธีถดถอยโลจิสติก (Logistic regression: LR)

จากการศึกษาเปรียบเทียบวิธีการตรวจสอบการทำหน้าที่ต่างของข้อสอบที่ผ่านมา อาจกล่าวได้ว่าวิธีที่ใช้หลักการของทฤษฎีการตอบสนองข้อสอบ (IRT) เป็นวิธีที่ดีที่สุด แต่มีข้อจำกัด คือ ต้องใช้กลุ่มตัวอย่างขนาดใหญ่ การคำนวณซับซ้อนและต้องคำนวณหลายรอบ ข้อมูลต้องเป็นตามข้อตกลงเบื้องต้น ซึ่งข้อสอบแบบเทสต์เลทขัดแย้งกับข้อตกลงเบื้องต้นของทฤษฎีการตอบสนองข้อสอบ (จิติมา วรรณศรี, 2539; Lee, Cohen & Toro, 2009)

ซึ่งจากการศึกษาที่ผ่านมา พบว่า มีการพัฒนาวิธีการและโมเดลต่าง ๆ เพื่อนำมาประยุกต์ใช้สำหรับวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบที่มีลักษณะเทสต์เลท เช่น การจัดกระทำกับข้อมูล หรือปรับการให้คะแนน (Scoring) เนื่องจากวิธีนี้มีแนวคิด คือ กำจัดความไม่เป็นอิสระในการตอบออกไป โดยรวมคะแนนข้อที่อยู่ในเทสต์เลทเดียวกัน ให้เป็นข้อสอบแบบให้คะแนนหลายค่า (Polytomous) 1 ข้อ ดังนั้น เมื่อจัดกระทำกับข้อมูลแล้ว จึงสามารถใช้โปรแกรมการวิเคราะห์ข้อสอบที่รองรับการวิเคราะห์ข้อมูลแบบให้คะแนนหลายค่าได้ เช่น Polytomous - SIBTEST MULTILOG เป็นต้น อย่างไรก็ตาม การประยุกต์ใช้วิธีการนี้ไม่สามารถตรวจสอบการทำหน้าที่ต่างกันระดับข้อสอบได้ แต่จะตรวจสอบได้ในระดับเทสต์เลทเท่านั้น

นอกจากนี้ ยังพบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเทสต์เลท โดยการพัฒนาโมเดลและการประยุกต์ใช้โมเดล ได้แก่

การประยุกต์วิเคราะห์ข้อสอบแบบพหุระดับด้วยโมเดลเชิงเส้นตรงระดับลดหลั่น 3 ระดับ หรือ HGLM-3L (Jiao, Wang & Kamata, 2005) โดยกำหนดให้การวิเคาะห์ระดับที่ 1 เป็นระดับข้อสอบ การวิเคาะห์ระดับที่ 2 เป็นระดับเทสต์เลท (Testlet) และการวิเคาะห์ระดับที่ 3 เป็นระดับผู้สอบ ซึ่งในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) จะใส่ตัวแปรกลุ่มเข้าไปในระดับที่ 2 โดยวิธีนี้สามารถกำจัดปัญหาความไม่เป็นอิสระในการตอบข้อสอบออกได้

แต่สามารถวิเคราะห์ได้เทียบเท่ากับ 1 พารามิเตอร์ เท่านั้น และความแปรปรวนของอิทธิพลของเทสต์เลท (Testlet effect) ในแต่ละเทสต์เลทมีค่าคงที่

อีกวิธีหนึ่ง คือ การเพิ่มพารามิเตอร์ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นอิทธิพลสุ่ม (Random effect : $\gamma_{id(j)}$) เข้าไปในสมการของโมเดลแบบ IRT หรือ เรียกว่า Rasch Testlet Model (Wang & Wilson, 2005a) วิธีนี้ แม้จะผ่อนคลายข้อตกลงที่เกี่ยวกับความเป็นอิสระระหว่างข้อสอบและผู้สอบ แต่ก็สามารถวิเคราะห์ได้เทียบเท่ากับ 1 พารามิเตอร์ เช่นเดียวกับ HGLM - 3L

วิธีสุดท้าย คือ การประยุกต์ใช้ Bi-factor Multidimensional Item Response Theory Model for Testlets (Bi - factor MIRT) พัฒนาโดย Fukuhara & Kamata (2011) วิธีนี้เริ่มจากกำหนดให้โมเดลมี 2 คุณลักษณะ (หรือ 2 มิติ) ได้แก่ คุณลักษณะหลัก (Primary trait) หรือมิติที่ 1 เป็นการตอบสนองของข้อสอบแต่ละข้อหรือความสามารถ และคุณลักษณะที่ 2 หรือมิติที่ 2 เป็นคุณลักษณะของเทสต์เลท (Testlet trait) หรืออิทธิพลของเทสต์เลท (Testlet effect) วิธีนี้สามารถผ่อนคลายข้อตกลงที่เกี่ยวกับความเป็นอิสระระหว่างข้อสอบและผู้สอบ ไม่ทำให้สารสนเทศรายข้อหายไป มีความพอดีกับข้อมูลที่มีความไม่เป็นอิสระในการตอบโดยมีสาเหตุจากเทสต์เลทมากกว่าโมเดลทฤษฎีการตอบสนองข้อสอบที่มีลักษณะของเทสต์เลท (TRT) และสามารถวิเคราะห์ได้เทียบเท่ากับ 2 พารามิเตอร์

สำหรับการนำโมเดลที่มีแนวคิดมาจากทฤษฎีการตอบสนองข้อสอบเพื่อการวิเคราะห์สำหรับเทสต์เลทมาใช้ประมาณค่าพารามิเตอร์และการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อพิจารณาแต่ละโมเดลซึ่งมีข้อดีและข้อจำกัดต่างกัน ผู้วิจัยเลือกโมเดลที่มีความเหมาะสมที่สุด นั่นคือ Bi - factor MIRT มาศึกษาในครั้งนี้ ซึ่งพัฒนาโดย Fukuhara & Kamata (2011) เนื่องจากรองรับการวิเคราะห์ได้กับข้อสอบที่มีการให้คะแนนแบบสองค่าแบบ 2 พารามิเตอร์ ไม่ทำให้สารสนเทศรายข้อหายไป สามารถวิเคราะห์หาสารสนเทศและตรวจสอบการทำหน้าที่ต่างกันในระดับข้อสอบได้ และเป็นโมเดลที่สอดคล้องกับสถานการณ์จริงมากกว่า เช่น ไม่มีข้อจำกัดในเรื่องความแปรปรวนของอิทธิพลของเทสต์เลท (Testlet effect)

สำหรับวิธีที่ใช้ในการประมาณค่าพารามิเตอร์ในการวิเคราะห์ข้อสอบด้วยทฤษฎีการตอบสนองข้อสอบ ที่ผู้วิจัยสนใจศึกษา ได้แก่ วิธีแมกซิมัมไลค์ลิฮู้ด วิธีของเบส์ (Bayes) และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) เนื่องจากวิธีเหล่านี้ ได้ถูกนำมาใช้ในการวิเคราะห์เป็นจำนวนมาก โดยในช่วงปี ค.ศ. 1994 - 1995 ประมาณหนึ่งในสามของบทความในวารสาร Applied Psychological Measurement (APM) Psychometrika และ Journal of Educational Measurement (JEM) ใช้เทคนิคมาร์คอฟเชนมอนติคาร์โล (Markov chain monte carlo: MCMC)

ในการวิเคราะห์ข้อมูลโดยเฉพาะการประเมินประสิทธิภาพในการประมาณค่าพารามิเตอร์ (Harwell, Stone, Hsu & Kirisci, 1996) ประกอบกับ Zhang (2010) ได้สรุปข้อมูลบทความเกี่ยวกับเทสต์เลทระหว่างปี ค.ศ. 1989 - 2009 ในฐานข้อมูลของ EBSCO และ PsychInfo พบว่า ร้อยละ 82.76 ประมาณค่าพารามิเตอร์ด้วยวิธีแมกซิมัมไลค์ลิฮู้ดและร้อยละ 17.24 ประมาณค่าพารามิเตอร์ด้วยวิธีของเบส์

จะเห็นว่าในการวิเคราะห์ข้อสอบด้วยทฤษฎีการตอบสนองข้อสอบ มักนำวิธีแมกซิมัมไลค์ลิฮู้ดและวิธีของเบส์ มาใช้ในการประมาณค่าพารามิเตอร์ โดยที่ข้อดีของการประมาณค่าด้วยวิธีแมกซิมั่มไลค์ลิฮู้ด คือ ให้สารสนเทศของค่าพารามิเตอร์ที่ต้องการได้ทั้งหมด ไม่ว่าจะเป็นพารามิเตอร์ของผู้สอบและพารามิเตอร์ของข้อสอบ แต่การประมาณค่าพารามิเตอร์ของวิธีนี้ ขึ้นอยู่กับจำนวนของกลุ่มผู้เข้าสอบและข้อสอบถ้ามีจำนวนเพิ่มขึ้น การประมาณค่าก็จะมีความคงที่ไปสู่ค่าพารามิเตอร์เพิ่มมากขึ้น

ส่วนข้อจำกัดของการประมาณค่าด้วยวิธีแมกซิมั่มไลค์ลิฮู้ด คือ การประมาณค่าพารามิเตอร์ในขั้นที่ 2 และ 3 โดยใช้ค่าอนุพันธ์อันดับที่ 2 ในกระบวนการนิวตัน - ราฟสัน (Newton - Raphson) มีโอกาสที่ค่าประมาณที่ได้จะไม่ลู่เข้าสู่ค่าคงที่ ประเด็นถัดมาสำหรับการประมาณค่าในสมการไลค์ลิฮู้ดไม่ใช่สมการเชิงเส้นตรง จะทำให้การหารากของสมการที่ทำให้ฟังก์ชันไลค์ลิฮู้ดมีค่าสูงสุดได้หลายค่าแต่ค่าเหล่านี้ไม่สามารถนำไปใช้หรือประกันได้ว่าเป็นค่าพารามิเตอร์ที่แท้จริงได้ ประเด็นที่สาม ในบางครั้งค่าพารามิเตอร์หรือค่าที่ได้จากการประมาณไม่ตกอยู่ในขอบเขตของค่าพารามิเตอร์ นั่นคืออาจมีค่าใดค่าหนึ่งอยู่ภายนอกขอบเขตที่ยอมรับได้ ในกรณีเช่นนี้ต้องมีการกำหนดขอบเขตจำกัดของค่าประมาณไว้ เพื่อให้ค่าประมาณที่ได้ไม่สูงหรือต่ำเกินไปนัก แต่การกระทำเช่นนี้เป็นจุดอ่อนของการประมาณค่าด้วยวิธีแมกซิมัมไลค์ลิฮู้ด โดยเฉพาะในแบบจำลอง 2 และ 3 พารามิเตอร์ จึงทำให้เกิดปัญหาตามมาเกี่ยวกับความตรง (Validity) ของค่าที่ประมาณได้ (ชนะศึก นิชานนท์, 2553) นอกจากนี้ หากผู้สอบตอบถูกหรือตอบผิดทั้งหมด วิธีนี้จะไม่สามารถประมาณค่าพารามิเตอร์ได้

เนื่องจากการประมาณค่าพารามิเตอร์ด้วยวิธีแมกซิมัมไลค์ลิฮู้ด มีข้อจำกัดและปัญหาเมื่อทำการประมาณค่าพารามิเตอร์ของข้อสอบ และค่าความสามารถของผู้เข้าสอบไปพร้อม ๆ กัน ดังกล่าวแล้ว วิธีของเบส์จึงอาจเป็นวิธีที่เหมาะสมกว่า ทั้งนี้เพราะวิธีของเบส์มีแนวคิดบางประการที่ต่างออกไปจากแนวคิดของวิธีแมกซิมัมไลค์ลิฮู้ด นั่นคือ ค่าพารามิเตอร์ของผู้สอบและค่าพารามิเตอร์ของข้อสอบเป็นตัวแปรสุ่ม (Random variable) จากการแจกแจงที่แสดงได้ด้วยฟังก์ชันความหนาแน่นร่วม (Joint density function) หรือการแจกแจงเริ่ม (Prior distribution) ซึ่งทำให้การใช้ฟังก์ชันไลค์ลิฮู้ดเพียงอย่างเดียวในการประมาณค่าถูกพิจารณาว่าเป็นการใช้ข้อมูล

ที่มีอยู่ อย่างไม่ครบถ้วน เพราะยังมีการแจกแจงเริ่มร่วมกับฟังก์ชันความหนาแน่นร่วมที่ควรนำมาใช้ในการประมาณค่าพารามิเตอร์ด้วย แม้จะมีข้อดีมากกว่าวิธีแมกซิมัมไลค์ลิฮู้ด แต่วิธีของเบส์เวลาใช้ในการวิเคราะห์นานกว่าวิธีแมกซิมัมไลค์ลิฮู้ดมาก

แม้ว่าแต่ละวิธีจะเป็นที่นิยมใช้ในการประมาณค่าพารามิเตอร์ โดยที่วิธีแมกซิมัมไลค์ลิฮู้ดมีข้อดีที่เด่นชัด คือ ประมาณค่าได้เร็วกว่า แต่มีข้อจำกัด คือ มีโอกาสที่ค่าที่ประมาณได้จะไม่ลู่เข้ามากกว่าเช่นกัน และที่สำคัญไม่สามารถประมาณค่าพารามิเตอร์หากผู้สอบตอบถูกหรือตอบผิดทั้งหมด ส่วนวิธีของเบส์ทั้งในแบบที่คำนึงถึงอิทธิพลของเทสต์เลทและไม่คำนึงถึงอิทธิพลของเทสต์เลท แม้จะมีข้อดีที่มีใช้ข้อมูลที่มีอยู่อย่างครบถ้วน ไม่มีปัญหาในส่วนการประมาณค่า แต่ใช้เวลาในการประมวลผลนานกว่าวิธีแมกซิมัมไลค์ลิฮู้ดมาก ดังนั้น จึงยังไม่ชัดเจนว่าวิธีการใดจะมีความเหมาะสมและคุ้มค่าต่อการนำไปใช้มากกว่ากัน ผู้วิจัยจึงสนใจที่จะเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ด้วยวิธีแมกซิมัมไลค์ลิฮู้ดและวิธีของเบส์ทั้งในแบบที่คำนึงถึงอิทธิพลของเทสต์เลทและไม่คำนึงถึงอิทธิพลของเทสต์เลท เพื่อค้นหาว่าวิธีใดจะมีประสิทธิภาพและคุ้มค่าที่สุดในการนำไปใช้วิเคราะห์ข้อสอบที่มีลักษณะเทสต์เลท

การศึกษาข้อสอบที่มีลักษณะเทสต์เลทที่ผ่านมานั้น มีทั้งการศึกษาจากข้อมูลจริงและข้อมูลจำลอง โดยการศึกษาจากข้อมูลจริง มีข้อจำกัดคือ ผู้วิจัยอาจไม่พบตามเงื่อนไขที่สนใจศึกษาจากข้อมูล ดังนั้น ในการศึกษาครั้งนี้ ผู้วิจัยจึงใช้ข้อมูลจำลองเนื่องจากสามารถศึกษาโอกาสที่จะเกิดขึ้นของข้อมูลได้หลากหลายรูปแบบที่น่าจะมีโอกาสเกิดขึ้นจริง จากการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องที่เกี่ยวกับทฤษฎีการวิเคราะห์ข้อสอบการประมาณค่าพารามิเตอร์และการจำลองข้อมูล พบว่ามีปัจจัยต่าง ๆ ที่ส่งผลกระทบต่อการประมาณค่าพารามิเตอร์ ได้แก่

1. อิทธิพลของเทสต์เลท จากการศึกษางานวิจัยที่ผ่านมา พบว่าในการศึกษาเกี่ยวกับอิทธิพลของเทสต์เลท มักกำหนดค่าของอิทธิพลเทสต์เลทไม่แตกต่างกันนัก โดยมักกำหนดช่วงระหว่าง 0 ถึง 1.5 เช่น Wainer et al. (2002) ศึกษาอิทธิพลของเทสต์เลทที่ระดับ 0, 0.5 และ 1 ส่วน Wang & Wilson (2005 a) ศึกษาอิทธิพลของเทสต์เลทที่ระดับ 0.25, 0.5, 0.75 และ 1 นอกจากนี้ Jiao, Wang, & He (2013) ศึกษาอิทธิพลของเทสต์เลทที่ระดับ 0, 0.25, 0.5625, 1 อย่างไรก็ตาม การศึกษาที่ผ่านมานั้น เป็นการศึกษาการประมาณค่าพารามิเตอร์ข้อสอบและพารามิเตอร์ผู้สอบ แต่ไม่ได้ศึกษาในประเด็นของการทำหน้าที่ต่างกันของข้อสอบ ดังนั้นผู้วิจัยจึงสนใจศึกษาขนาดของอิทธิพลของเทสต์เลทที่แตกต่างกันจะส่งผลต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยหรือไม่

2. การแจกแจงของความสามารถ แม้การศึกษาด้านการวัดและวิจัย ส่วนใหญ่จะกำหนดการแจกแจงของความสามารถให้เป็นไปตามสมมติฐาน นั่นคือ การแจกแจงแบบปกติ

แต่ในสถานการณ์จริงแล้ว การแจกแจงความสามารถแบบเบ้ก็มีโอกาสเกิดขึ้นได้กับประชากรของผู้สอบ เช่น เมื่อมีการปรับปรุงการเรียนการสอนหรือความคุ้นเคยกับรูปแบบการสอบในอนาคต ก็อาจทำให้ความสามารถเฉลี่ยของผู้สอบมีมากขึ้น ทำให้เกิดการแจกแจงแบบเบ้ซ้ายได้หรือการแจกแจงแบบเบ้ขวาที่ผู้สอบที่มีความสามารถสูงน้อยกว่าผู้สอบที่มีความสามารถต่ำ นอกจากนี้ การแจกแจงแบบเบ้ยังไม่ตรงกับลักษณะการแจกแจงตามสมมติฐานของโมเดล อาจมีผลกระทบต่อการประมาณค่าพารามิเตอร์ ดังนั้นผู้วิจัยจึงสนใจที่จะศึกษาการแจกแจงของความสามารถที่แตกต่างกัน จะมีผลกระทบต่อการประมาณค่าพารามิเตอร์หรือไม่

3. จำนวนข้อสอบที่มีการทำหน้าที่ต่างกัน จากการศึกษาที่ผ่านมาพบว่า มีการกำหนดสัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันในอัตราส่วนต่าง ๆ เช่น Lee et al. (2009) ศึกษาอำนาจการทดสอบและอัตราความคาดเคลื่อนประเภทที่ 1 ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีลักษณะของเทสต์เลท โดยใช้วิธี SIBTEST และ Poly - SIBTEST และศึกษาสัดส่วนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบ เป็นร้อยละ 0, 10 และ 20 พบว่า ร้อยละของข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบที่แตกต่างกัน ส่งผลต่ออำนาจการทดสอบอย่างไม่มีแบบแผน นอกจากนี้ ผลการศึกษาของ Narayanan & Sawaminathan (1996, อ้างถึงใน สิริรัตน์ วิภาสศิลป์, 2545) พบว่า สัดส่วนของข้อสอบแสดงการทำหน้าที่ต่างกันของข้อสอบในแบบสอบมีผลต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ถ้ามีข้อสอบแสดงการทำหน้าที่ต่างกันของข้อสอบปริมาณมาก จะทำให้ความถูกต้องในการตรวจสอบลดลง และสิริรัตน์ วิภาสศิลป์ (2545) พบว่า หากสัดส่วนของข้อสอบที่แสดงการทำหน้าที่ต่างกันในแบบสอบมากกว่าร้อยละ 20 จะทำให้มีการระบุผิดพลาดในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูง จากผลการศึกษาที่ผ่านมาจะเห็นว่า ยังมีความคลุมเคลือในการศึกษาเกี่ยวกับจำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบ นอกจากนี้ การศึกษาที่ผ่านมาในส่วนที่เป็นข้อสอบที่มีลักษณะของเทสต์เลท ยังเป็นการศึกษาโดยใช้การวิเคราะห์รูปแบบการจัดคะแนน (Scoring) ซึ่งมีจุดอ่อนในการนำไปใช้ ดังนั้น ผู้วิจัยจึงสนใจในการศึกษาเงื่อนไขร้อยละของข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบ เพื่อให้ได้ผลสรุปที่ชัดเจนมากขึ้น

4. อัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ ซึ่งจากการศึกษาที่ผ่านมา พบว่า มีการศึกษาเกี่ยวกับอัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ ซึ่งมีผลต่อวิธีที่ใช้ในการตรวจสอบการทำหน้าที่ต่างกันวิธีต่าง ๆ เช่น จิตติมา วรรณศรี (2538) เปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี Mantel - Heanszel (MH) กับวิธี SIBTEST พบว่า เมื่อกลุ่มอ้างอิงและกลุ่มเปรียบเทียบอัตราส่วน 1: 1 เป็นเงื่อนไขที่ดีที่สุด รองลงมาคือ การใช้อัตราส่วนระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบเป็น 1: 0.75 ศึกษาผลกระทบต่ออำนาจ (Power)

ในการตรวจสอบการทำหน้าที่ ด้วยวิธี SIBTEST และ Mantel - Haenszel (MH) ในระดับต่าง ๆ ผลการศึกษา พบว่า อัตราความคลาดเคลื่อนประเภทที่ 1 (Type I error rate) ของตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมีมากขึ้น เมื่ออัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบมีอัตราส่วนแตกต่างกันมาก (1: 0.1) และควรใช้อัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบไม่น้อยกว่า 1: 0.5 แต่ในทางปฏิบัติการสอบคัดเลือกเข้าศึกษาต่อในสถาบันการศึกษาต่าง ๆ มักมีอัตราส่วนของประชากรระหว่างอัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบที่แตกต่างกัน เช่น การสอบวิชาภาษาอังกฤษสำหรับนิสิตระดับบัณฑิตศึกษา หากในข้อสอบมีบทความที่เนื้อหาเกี่ยวกับการเมืองก็อาจทำให้กลุ่มนิสิตที่เรียนทางด้านรัฐศาสตร์ได้ประโยชน์ ซึ่งเป็นนิสิตกลุ่มน้อยเมื่อเทียบกับนิสิตทั้งหมด (ไม่ใช่อัตราส่วน 1: 1) ดังนั้น ผู้วิจัยจึงสนใจศึกษาอัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบทั้งในอัตราส่วนที่เท่ากันและไม่เท่ากัน เพื่อค้นหาว่าความแตกต่างของอัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบนั้น มีผลกระทบต่อการประมาณค่าพารามิเตอร์หรือไม่

จะเห็นว่าการพัฒนาวิธีการประมาณค่าพารามิเตอร์ ในประเด็นเกี่ยวกับเทสต์เลทหลายโมเดล เช่น TRT แต่ในการใช้งานจริง บางครั้งก็มักใช้โมเดล IRT ในการประมาณค่าพารามิเตอร์แทน เนื่องจากโมเดล IRT เป็นที่รู้จักและมีโปรแกรมสำเร็จรูปรองรับมากกว่า แม้ว่าการไม่คำนึงถึงอิทธิพลของเทสต์เลทส่งผลกระทบต่อการประมาณค่าพารามิเตอร์ดังที่ได้กล่าวมาแล้วก็ตาม ดังนั้น ผลจากการวิจัยในครั้งนี้ เป็นการขยายความรู้ในเชิงทฤษฎีเกี่ยวกับการประมาณค่าพารามิเตอร์และการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อมีการฝ่าฝืนข้อตกลงเบื้องต้น สำหรับการวิเคราะห์ตามแนวทางทฤษฎีการตอบสนองข้อสอบ นอกจากนี้ยังทำให้ทราบถึงความสอดคล้องของวิธีการและโมเดลที่ใช้ในการประมาณค่าพารามิเตอร์ข้อสอบและค่าพารามิเตอร์ผู้สอบทำให้สามารถเลือกใช้วิธีการประมาณค่าพารามิเตอร์ของข้อสอบและพารามิเตอร์ความสามารถของผู้สอบและวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้อย่างเหมาะสม เมื่อเทียบกับทรัพยากรที่ต้องใช้ในการคำนวณและลักษณะของข้อมูล

คำถามการวิจัย

1. การประมาณค่าพารามิเตอร์ข้อสอบ (ความยากและอำนาจจำแนก) และพารามิเตอร์ผู้สอบ (ความสามารถ) ด้วยวิธีแมกซิมัมไลค์ลิฮู้ด (ML) วิธีของเบส์ (Bayes) และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) มีประสิทธิผลต่างกันอย่างไร

2. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมกซิมัมไลค์ลิฮู้ด (ML) วิธีของเบส์ (Bayes) และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) มีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 1 และอำนาจการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบอย่างไร

วัตถุประสงค์ของการวิจัย

เพื่อศึกษาประสิทธิผลในการประมาณค่าพารามิเตอร์และการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่เหมาะสมกับข้อมูลด้วยวิธีแมกซิมัมไลค์ลิฮู้ด (ML) วิธีของเบส์ (Bayes) และวิธีของเบส์แบบมีอิทธิพลของเทสต์เลท (Bayes$_\gamma$) ในประเด็น ดังนี้

1. การประมาณค่าพารามิเตอร์

1.1 เพื่อศึกษาประสิทธิผลในการประมาณค่าพารามิเตอร์ข้อสอบ (ความยากและอำนาจจำแนก) และพารามิเตอร์ผู้สอบ (ความสามารถ) ของวิธีแมกซิมัมไลค์ลิฮู้ด (ML) วิธีของเบส์ (Bayes) และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayes$_\gamma$)

2. การทำหน้าที่ต่างกันของข้อสอบ (DIF)

2.1 เพื่อศึกษาอัตราความคลาดเคลื่อนประเภทที่ 1 (Type I error rate) ของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ของวิธีแมกซิมัมไลค์ลิฮู้ด (ML) วิธีของเบส์ (Bayes) และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayes$_\gamma$)

2.2 เพื่อศึกษาอำนาจ (Power) การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบของวิธีแมกซิมัมไลค์ลิฮู้ด (ML) วิธีของเบส์ (Bayes) และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayes$_\gamma$)

ขอบเขตการวิจัย

1. การศึกษาครั้งนี้ ใช้โมเดลทฤษฎีการตอบข้อสอบแบบ 2 พารามิเตอร์ นั่นคือศึกษาพารามิเตอร์ข้อสอบ 2 พารามิเตอร์ ได้แก่ ความยาก อำนาจจำแนก และพารามิเตอร์ผู้สอบ ได้แก่ ความสามารถของผู้สอบ

2. ข้อมูลที่ใช้ในการศึกษาครั้งนี้ ใช้ข้อมูลจำลองที่ได้จากโปรแกรม R โดยเป็นข้อมูลที่มีวิธีการให้คะแนนรายข้อแบบสองค่า (Dichotomous scoring) ซึ่งทำการจำลองภายใต้เงื่อนไขจำนวน 54 เงื่อนไข (3 x 3 x 3 x 2) ในแต่ละเงื่อนไขจำลองข้อมูลซ้ำ 100 รอบ จำนวนการทำซ้ำภายใต้เงื่อนไขที่แปรเปลี่ยนทั้งหมด 5,400 รอบ โดยมีรายละเอียดของเงื่อนไข ดังนี้

2.1 อิทธิพลของเทสต์เลท ประกอบด้วย 3 เงื่อนไข ได้แก่

2.1.1 อิทธิพลของเทสต์เลทเท่ากันทุกเทสต์เลท

2.1.2 อิทธิพลของเทสต์เลทไม่เท่ากัน

2.1.3 ข้อสอบที่เป็นอิสระและอิทธิพลของเทสต์เลทไม่เท่ากัน

2.2 การแจกแจงของความสามารถของผู้สอบ ประกอบด้วย 3 เงื่อนไข ได้แก่

2.2.1 การแจกแจงแบบปกติ

2.2.2 การแจกแจงแบบเบ้ซ้าย

2.2.3 การแจกแจงแบบเบ้ขวา

2.3 จำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบ ประกอบด้วย 3 เงื่อนไข ได้แก่

2.3.1 ร้อยละ 0 หรือไม่มีข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบ

2.3.2 ร้อยละ 12.5 หรือมีข้อสอบที่ทำหน้าที่ต่างกันจำนวน 5 ข้อ ในแบบสอบ

2.3.3 ร้อยละ 20 หรือมีข้อสอบที่ทำหน้าที่ต่างกันจำนวน 8 ข้อ ในแบบสอบ

2.4 อัตราส่วนของกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ ประกอบด้วย 2 เงื่อนไข ได้แก่

2.4.1 อัตราส่วน 1: 1 หรือ มีกลุ่มอ้างอิงจำนวน 1,000 คน และกลุ่มเปรียบเทียบจำนวน 1,000 คน

2.4.2 อัตราส่วน 1: 0.1 หรือ มีกลุ่มอ้างอิงจำนวน 1,000 คน และกลุ่มเปรียบเทียบจำนวน 100 คน

3. วิธีที่ใช้ในการประมาณค่าพารามิเตอร์ต่าง ๆ สำหรับการศึกษาครั้งนี้ มี 3 วิธี ได้แก่

3.1 วิธีแมกซิมัมไลค์ลิฮู้ด (ML)

3.2 วิธีของเบส์ (Bayes)

3.3 วิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ)

4. การศึกษาครั้งนี้ พิจารณาการวัดประสิทธิผล 2 ส่วน ได้แก่

4.1 การวัดประสิทธิผลการประมาณค่าพารามิเตอร์ พิจารณาจากความเบี่ยงเบนของค่าพารามิเตอร์ที่แท้จริงและค่าที่ประมาณได้ (RMSE)

4.2 การวัดประสิทธิผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบพิจารณาจาก

4.2.1 อัตราความคลาดเคลื่อนประเภทที่ 1 (Type I error rate) ของตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

4.2.2 อำนาจ (Power rate) ของตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

5. การศึกษาครั้งนี้ไม่ได้ศึกษาสาเหตุของการของการละเมิดข้อตกลงเบื้องต้นของ IRT แต่เป็นการศึกษาผลกระทบที่เกิดจากการละเมิดข้อตกลงเบื้องต้นของ IRT เมื่อข้อสอบมีความไม่เป็นอิสระต่อกันจากอิทธิพลของเทสต์เลท

6. การศึกษาครั้งนี้ กำหนดให้แบบสอบมีความยาว 40 ข้อ ประกอบด้วย 4 เทสเลท โดยแต่ละเทสเลทมีขนาดเท่ากัน คือ 10 ข้อทุกเงื่อนไข

ประโยชน์ที่ได้รับจากการวิจัย

1. ผลการศึกษาทำให้ทราบถึงวิธีการและโมเดลที่ใช้ในการประมาณค่าพารามิเตอร์ข้อสอบและค่าพารามิเตอร์ผู้สอบว่าวิธีแมกซิมัมไลค์ลิฮู้ด (ML) วิธีของเบส์ (Bayes) และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) มีลักษณะเป็นอย่างไร ทำให้สามารถเลือกใช้วิธีการประมาณค่าพารามิเตอร์ของข้อสอบและพารามิเตอร์ความสามารถของผู้สอบและวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ได้อย่างเหมาะสมกับลักษณะของข้อมูล

2. เพื่อเป็นแนวทางในการประมาณค่าพารามิเตอร์ข้อสอบและค่าพารามิเตอร์ผู้สอบและวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อีกวิธีหนึ่งให้แก่นักวัดผลการศึกษาในการตรวจสอบและพัฒนาคุณภาพข้อสอบ

3. ผลการวิจัยมีประโยชน์ในการวิเคราะห์ข้อสอบที่มีอิทธิพลเทสต์เลท ซึ่งเป็นการแสดงวิธีการวิเคราะห์ข้อสอบแนวใหม่ ที่สามารถนำไปใช้ได้ทุกประเภทของการวัดผล เนื่องจากข้อสอบลักษณะนี้ มีปรากฏอย่างแพร่หลายทั้งการสอบในระดับสถาบันและระดับชาติ

4. เป็นแนวทางในการศึกษาวิธีการประมาณค่าพารามิเตอร์ และการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ทั้งรูปแบบการวิเคราะห์ที่มีและการไม่มีอิทธิพลเทสต์เลท โดยใช้ Package ต่าง ๆ จากโปรแกรม R

5. เป็นการขยายความรู้ในเชิงทฤษฎีเกี่ยวกับการประมาณค่าพารามิเตอร์และการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อมีการฝ่าฝืนข้อตกลงเบื้องต้น สำหรับการวิเคราะห์ตามแนวทางทฤษฎีการตอบสนองข้อสอบ

นิยามศัพท์เฉพาะ

การทำหน้าที่ต่างกันของข้อสอบ (DIF) หมายถึง ข้อสอบที่ทำให้ผู้สอบซึ่งมีลักษณะต่างกัน แต่มีความสามารถระดับเดียวกัน มีโอกาสตอบถูกไม่เท่ากัน

กลุ่มเปรียบเทียบ (Focal group หรือกลุ่ม F) หมายถึง ผู้สอบกลุ่มที่คาดว่าจะเป็นกลุ่มที่เสียประโยชน์จากข้อสอบที่แสดงออกถึงการทำหน้าที่ต่างกันของข้อสอบ

กลุ่มอ้างอิง (Reference group หรือกลุ่ม R) หมายถึง ผู้สอบกลุ่มที่คาดว่าจะเป็นกลุ่มที่ได้ประโยชน์จากข้อสอบที่แสดงออกถึงการทำหน้าที่ต่างกันของข้อสอบ

ความเป็นอิสระในการตอบข้อสอบ (Local independence) หมายถึง เมื่อควบคุมความสามารถที่มีผลต่อการตอบข้อสอบให้คงที่แล้ว ความน่าจะเป็นในการตอบข้อสอบถูกในแต่ละข้อมีความเป็นอิสระต่อกัน

เทสต์เลท (Testlet) หมายถึง กลุ่มของข้อสอบที่มีความไม่เป็นอิสระของการตอบคำถามในแต่ละข้อ อันเนื่องมาจากมีการใช้ข้อมูล บทความ หรือเหตุการณ์ ที่เป็นตัวกระตุ้นหรือ สิ่งเร้าเดียวกัน

วิธีการประมาณค่าพารามิเตอร์ หมายถึง วิธีที่ใช้ในการคำนวณหาค่าพารามิเตอร์ความยากของข้อสอบและค่าความสามารถของผู้สอบ โดยใช้ค่าสถิติจากข้อมูลที่จำลอง โดยใช้วิธีการประมาณค่าพารามิเตอร์ด้วยวิธีแมกซิมัมไลค์ลิฮู้ด (Maximum likelihood: ML) วิธีของเบส์ (Bayes) และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) โดยใช้แบบจำลองตามทฤษฎีการตอบสนองข้อสอบ และโมเดล Bi-factor MIRT

วิธีแมกซิมัมไลค์ลิฮู้ด (Maximum likelihood: **ML)** หมายถึง การประมาณค่าพารามิเตอร์โดยอาศัยผลที่ได้จากตัวอย่างที่สุ่มเลือกมาจากการแจกแจงที่ทราบรูปแบบของฟังก์ชันความหนาแน่นแต่ไม่ทราบค่าพารามิเตอร์ ดังนั้นจึงใช้หลักของความน่าจะเป็นในการเลือกตัวอย่างและวัดค่าได้จากกลุ่มตัวอย่างที่ถูกเลือก ($U_1 = U_1, U_2 = U_2, \ldots, U_n = U_n$) มาพิจารณาค่าประมาณของค่าพารามิเตอร์ที่ต้องการ ในการวิจัยครั้งนี้ใช้ทำการประมาณค่าพารามิเตอร์ด้วยวิธีแมกซิมัมไลค์ลิฮู้ด โดยเรียกใช้งาน Packages TAM จากโปรแกรม R

วิธีของเบส์ (Bayes) หมายถึง วิธีการประมาณค่าพารามิเตอร์ที่มีแนวคิดว่า ความสามารถของผู้สอบ (θ_i) และค่าพารามิเตอร์ของข้อสอบ ได้แก่ ค่าอำนาจจำแนกของข้อสอบ (a_j) ค่าความยากของข้อสอบ (b_j) เป็นตัวแปรสุ่ม (Random variable) มีเป้าหมายเพื่อหาการแจกแจงภายหลัง (Posterior distribution) ของพารามิเตอร์ที่สนใจ โดยอาศัยฟังก์ชัน Likelihood และการแจกแจงก่อน (Prior distributions) ในการวิจัยครั้งนี้ ใช้ทำการประมาณค่าพารามิเตอร์ด้วยวิธีของเบส์ โดยเรียกใช้งาน Packages R2WinBUGS จากโปรแกรม R แล้วส่งข้อมูลไปประมวลผลที่โปรแกรม WinBUGS เมื่อประมวลผลเสร็จแล้ว โปรแกรมจะส่งผลลัพธ์ที่ได้กลับไปแสดงที่โปรแกรม R โดยใช้แบบจำลองตามทฤษฎีการตอบสนองข้อสอบ (IRT)

วิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) หมายถึง วิธีการประมาณค่าพารามิเตอร์มีแนวคิดว่า ค่าความสามารถของผู้สอบ (θ_i) และค่าพารามิเตอร์ของข้อสอบ ได้แก่ ค่าอำนาจจำแนกของข้อสอบ (a_j) ค่าความยากของข้อสอบ (b_j) เป็นตัวแปรสุ่ม (Random variable) มีเป้าหมายเพื่อหาการแจกแจงภายหลัง (Posterior distribution) ของพารามิเตอร์ที่สนใจ โดยอาศัยฟังก์ชัน Likelihood และการแจกแจงก่อน (Prior distributions) ในการวิจัยครั้งนี้ ใช้ทำการประมาณค่าพารามิเตอร์ด้วยวิธีของเบส์ โดยเรียกใช้งาน Packages R2WinBUGS จากโปรแกรม R แล้วส่งข้อมูลไปประมวลผลที่โปรแกรม WinBUGS เมื่อประมวลผลเสร็จแล้ว โปรแกรมจะส่งผลลัพธ์ที่ได้กลับไปแสดงที่โปรแกรม R โดยใช้โมเดล Bi-factor MIRT

โมเดลตอบสนองข้อสอบพหุมิติแบบสององค์ประกอบ (Bi-factor Multidimensional Item Response Theory Model : Bi-factor MIRT) หมายถึง โมเดลที่ขยายจากทฤษฎีการตอบสนองข้อสอบที่มีเทสต์เลทในแบบสอบ (Testlet response theory) ที่กำหนดให้ความสามารถและอิทธิพลสุ่มเนื่องจากเทสต์เลท มีค่าอำนาจจำแนกเดียวกัน พัฒนาโดย Fukuhara & Kamata (2011) เป็นโมเดลที่นำมาประยุกต์ใช้ในการประมาณค่าพารามิเตอร์ สามารถวิเคราะห์แบบ 2 พารามิเตอร์ได้

พารามิเตอร์อำนาจจำแนก (Discrimination) หมายถึง ค่าที่แสดงความชันที่ตำแหน่งของโค้งลักษณะข้อสอบ (ICC) ณ จุดที่ความสามารถมีโอกาสตอบข้อสอบถูก 0.5 ซึ่งประมาณได้จากวิธีแมกซิมัมไลค์ลิฮู้ด (Maximum likelihood: ML) วิธีของเบส์ (Bayes) และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) โดยใช้แบบจำลองตามทฤษฎีการตอบสนองข้อสอบและโมเดล Bi - factor MIRT

พารามิเตอร์ความยาก (Difficulty) หมายถึง ค่าที่แสดงตำแหน่งของโค้งลักษณะข้อสอบ (ICC) ณ จุดที่ความสามารถมีโอกาสตอบข้อสอบถูก 0.5 ซึ่งประมาณได้จากวิธีแมกซิมัมไลค์ลิฮู้ด (Maximum likelihood: ML) วิธีของเบส์ (Bayes) และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) โดยใช้แบบจำลองตามทฤษฎีการตอบสนองข้อสอบและโมเดล Bi-factor MIRT

พารามิเตอร์ความสามารถของผู้สอบ (Ability) หมายถึง ระดับความสามารถของผู้สอบที่ประมาณได้จากวิธีแมกซิมัมไลค์ลิฮู้ด (Maximum likelihood: ML) วิธีของเบส์ (Bayes) และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) โดยใช้แบบจำลองตามทฤษฎีการตอบสนองข้อสอบและโมเดล Bi - factor MIRT

การจำลองข้อมูล หมายถึง การจัดสถานการณ์การสร้างข้อมูลตามเงื่อนไขเพื่อใช้การประมาณค่าพารามิเตอร์ของวิธีการวิเคราะห์ข้อสอบ โดยการสร้างเลขสุ่มในโปรแกรม R โดยมีเงื่อนไขความแตกต่างของ 4 ปัจจัย ได้แก่

1. อิทธิพลของเทสต์เลท (Testlet effect) หมายถึง ผลกระทบของอิทธิพลแบบสุ่มที่เป็นปฏิสัมพันธ์ของการตอบกับเทสต์เลท คำนวณได้จากการใช้โมเดล Bi - factor MIRT ด้วยวิธีของเบส์ ในการศึกษาครั้งนี้ผู้วิจัยใช้โปรแกรม R ร่วมกับโปรแกรม WinBUGS โดยผ่าน Package R2WinBUGS ในการคำนวณ ในการศึกษาครั้งนี้ ศึกษาอิทธิพลเทสต์เลท 3 เงื่อนไข ได้แก่

1.1 แบบสอบที่มี 4 เทสต์เลทและมีค่าอิทธิพลของเทสต์เลทเท่ากันทุกเทสต์เลท

1.2 แบบสอบที่มี 4 เทสต์เลทและแต่ละเทสต์เลทมีค่าอิทธิพลของเทสต์เลทไม่เท่ากัน

1.3 แบบสอบประกอบด้วยข้อสอบที่เป็นอิสระและเทสต์เลท

2. การแจกแจงของความสามารถ หมายถึง การแจกแจงของข้อมูลที่เป็นโค้งลักษณะต่าง ๆ ในการศึกษาครั้งนี้ ศึกษาการแจกแจงความสามารถที่มีโค้งต่างกัน 3 เงื่อนไข ได้แก่ การแจกแจงแบบปกติ การแจกแจงแบบเบ้ซ้าย และการแจกแจงแบบเบ้ขวา

3. จำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบ หมายถึง จำนวนข้อสอบที่วัดความสามารถของผู้สอบในแต่ละกลุ่มไม่ตรงกันในแบบสอบ ในการศึกษาครั้งนี้ ศึกษาจำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบ 3 เงื่อนไข ได้แก่ ร้อยละ 0 หรือไม่มีข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบ ร้อยละ 12.5 หรือมีข้อสอบที่ทำหน้าที่ต่างกันจำนวน 5 ข้อในแบบสอบ และร้อยละ 20 หรือมีข้อสอบที่ทำหน้าที่ต่างกันจำนวน 8 ข้อ ในแบบสอบ

4. อัตราส่วนของกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ หมายถึง จำนวนผู้สอบที่เป็นกลุ่มอ้างอิงและจำนวนผู้สอบที่เป็นกลุ่มเปรียบเทียบ ในการศึกษาครั้งนี้ ศึกษาอัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ 2 เงื่อนไข ได้แก่

1.1อัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ เป็น 1: 1 หรือมีจำนวนผู้สอบที่เป็นกลุ่มอ้างอิง จำนวน 1,000 คน และจำนวนผู้สอบที่เป็นกลุ่มเปรียบเทียบ จำนวน 1,000 คน

1.2อัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ เป็น 1: 0.1 หรือมีจำนวนผู้สอบที่เป็นกลุ่มอ้างอิง จำนวน 1,000 คน และจำนวนผู้สอบที่เป็นกลุ่มเปรียบเทียบ จำนวน 100 คน

การวัดประสิทธิผลการประมาณค่าพารามิเตอร์ หมายถึง การตรวจสอบความสามารถในการประมาณค่าพารามิเตอร์ (ความยาก อำนาจจำแนกของข้อสอบและความสามารถของผู้สอบ) โดยพิจารณาจาก

1. ความเบี่ยงเบนของค่าพารามิเตอร์ที่แท้จริงและค่าที่ประมาณได้ (Root Mean Square Error: RMSE) หมายถึง ค่าที่วัดจากความแตกต่างระหว่างค่าจริงและค่าที่ประมาณได้จากวิธีต่าง ๆ เป็นการทดสอบความแม่นยำของของค่าที่ประมาณได้ โดยหาก RMSE ของวิธีการใดมีค่าน้อย แสดงว่าวิธีนั้นสามารถประมาณค่าได้ใกล้เคียงกับค่าจริง ดังนั้นหากค่า RMSE มีค่าเท่ากับศูนย์ แสดงว่าไม่เกิดความคลาดเคลื่อนในการประมาณค่า

การวัดประสิทธิผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ หมายถึง ความสามารถการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยพิจารณาจาก

1. อัตราความคลาดเคลื่อนประเภทที่ 1 (Type I error rate) ของตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ หมายถึง จำนวนข้อสอบที่ระบุว่าทำหน้าที่ต่างกันผิดพลาด ทั้งที่ความเป็นจริงข้อสอบเหล่านั้นทำหน้าที่ไม่ต่างกัน ซึ่งคำนวณได้จากผลรวมจำนวนข้อสอบที่ระบุว่าทำหน้าที่ต่างกันผิดพลาดต่อจำนวนข้อสอบที่ทำหน้าที่ไม่ต่างกันในแบบสอบ x 100

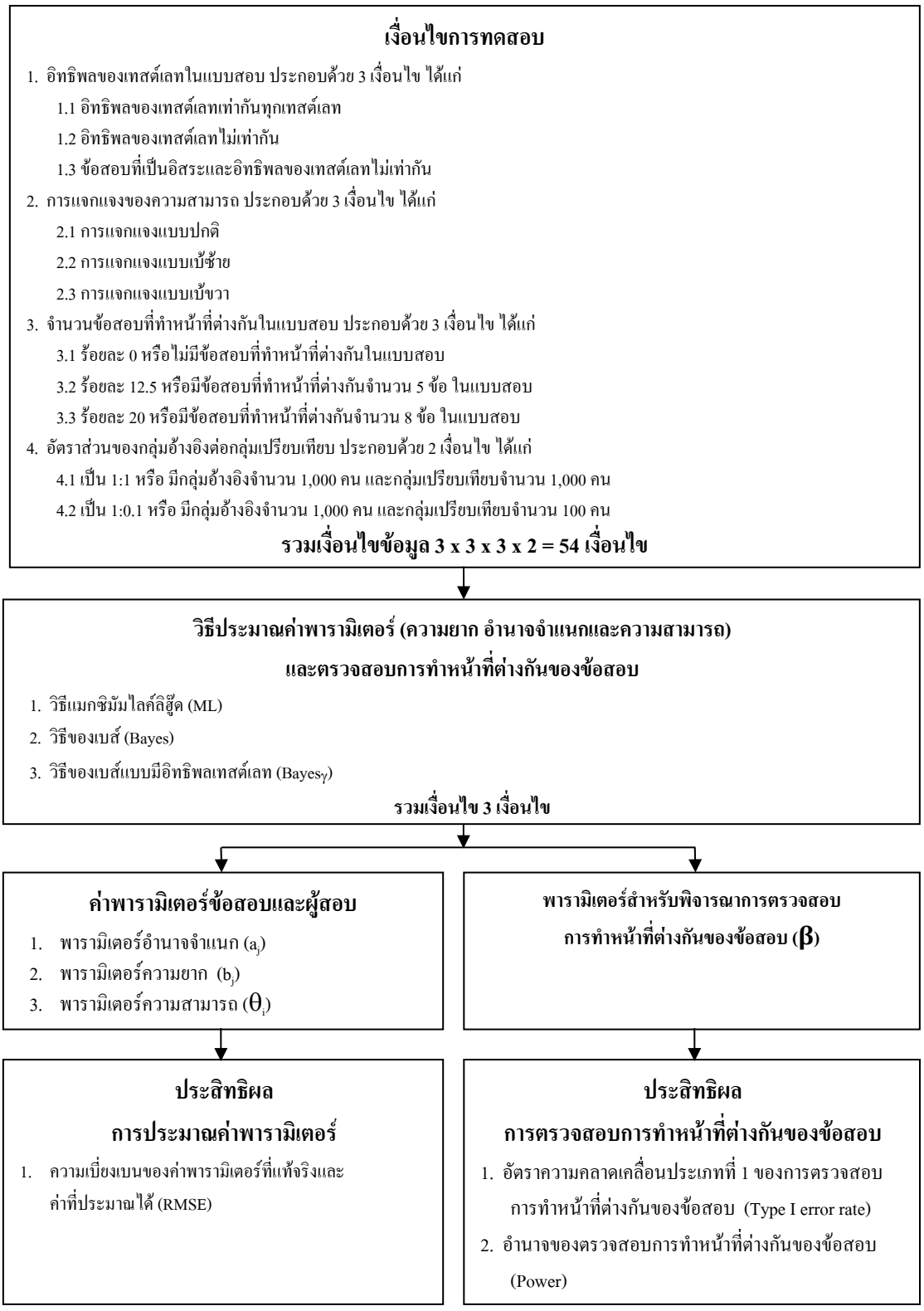
2. อำนาจของตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (Power) หมายถึง จำนวนข้อสอบที่ระบุว่าทำหน้าที่ต่างกันได้ถูกต้อง คำนวณได้จากผลรวมของจำนวนข้อสอบที่ระบุว่าทำหน้าที่ต่างกันได้ถูกต้องต่อจำนวนข้อสอบที่ทำหน้าที่ต่างกันทั้งหมดในแบบสอบ x 100

กรอบแนวคิดในการวิจัย

จากการศึกษาแนวคิดทฤษฎีที่เกี่ยวข้องและงานวิจัยต่าง ๆ ผู้วิจัยได้นำเสนอเป็นกรอบแนวคิดในการวิจัย โดยมีกระบวนการดังนี้

1. จำลองข้อมูล (Simulation) ตามเงื่อนไข เนื่องจากผู้วิจัยได้ศึกษาทฤษฎีและงานวิจัยที่ผ่านมา พบว่า มีลักษณะข้อมูลในเงื่อนไขต่าง ๆ ที่ส่งผลกระทบต่อการประมาณค่าพารามิเตอร์และการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ซึ่งในการตรวจสอบอาจไม่สามารถกำหนดให้ข้อมูลจริงมีเงื่อนไขสอดคล้องกับเงื่อนไขที่ต้องการศึกษาได้ ผู้วิจัยจึงใช้ข้อมูลจำลองในการศึกษา

2. วิเคราะห์ข้อมูล ประกอบด้วย การประมาณค่าพารามิเตอร์และการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยผู้วิจัยจะใช้ข้อมูลที่จำลองได้ ทำการประมวลผลใน 1 ครั้ง เพื่อประมาณค่าพารามิเตอร์และการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีต่าง ๆ ตามเงื่อนไข เมื่อประมาณค่าพารามิเตอร์แล้ว ทำการวัดประสิทธิภาพของการค่าที่ประมาณได้นั้นด้วยดัชนีต่าง ๆ โดยสรุปเป็นกรอบแนวคิดในการวิจัย ดังภาพที่ 1 - 1



ภาพที่ 1-1 กรอบแนวคิดในการวิจัย

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

การวิจัยครั้งนี้ ผู้วิจัยได้ศึกษาเอกสาร รายงานการวิจัย บทความทั้งในและต่างประเทศที่เกี่ยวข้องกับการทำหน้าที่ต่างกันของข้อสอบ วิธีการวิเคราะห์และประมาณค่าพารามิเตอร์ของการวิเคราะห์ข้อสอบ แล้วประมวลเป็นความรู้ โดยนำเสนอตามลำดับ ดังนี้

ตอนที่ 1 แนวคิดเกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบ

ตอนที่ 2 แนวคิดเกี่ยวกับทฤษฎีการตอบสนองข้อสอบ

ตอนที่ 3 แนวคิดเกี่ยวกับทฤษฎีการตอบสนองข้อสอบสำหรับเทสต์เลท

ตอนที่ 4 แนวคิดเกี่ยวกับการประมาณค่าพารามิเตอร์ด้วยวิธีของเบส์และวิธีแมกซิมัมไลค์ลิฮู้ด

ตอนที่ 5 การกำหนดเงื่อนไขสำหรับการจำลองข้อมูล

ตอนที่ 6 งานวิจัยที่เกี่ยวข้อง

ตอนที่ 1 แนวคิดเกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบ

ความยุติธรรมของข้อสอบมีความเกี่ยวข้องกับความตรงในการให้คะแนนของแบบสอบ ซึ่งเป็นหลักฐานที่แสดงว่าแบบสอบสามารถวัดได้ตรงตามวัตถุประสงค์ที่ต้องการวัด ในประเทศสหรัฐอเมริกา ซึ่งเป็นประเทศที่มีเสรีภาพและการเรียกร้องสิทธิของบุคคล ดังนั้น ในการสร้างและการใช้แบบสอบจะอยู่ภายใต้การตรวจสอบอย่างใกล้ชิดของสังคม ทำให้ผู้สร้างและผู้ใช้แบบสอบต้องแสดงให้เห็นว่าแบบสอบเหล่านั้น ปราศจากความลำเอียงต่อผู้ตอบข้อสอบกลุ่มใดกลุ่มหนึ่ง จึงเกิดความพยายามในการตรวจสอบความลำเอียงของข้อสอบมาตั้งแต่ต้นศตวรรษที่ 20 แต่วิธีการตรวจสอบอย่างเป็นระบบได้นำเสนอในปี ค.ศ. 1964 โดย Cardall & Coffman ได้ประยุกต์การวิเคราะห์ความแปรปรวน (Analysis of variance) ทดสอบปฏิกิริยาร่วมของข้อสอบที่ใช้กับผู้ตอบข้อสอบผิวดำและผิวขาวที่เข้ารับการทดสอบ SAT ในปี ค.ศ. 1963

ในกรณีที่ข้อสอบอำนวยประโยชน์ให้กับผู้ตอบข้อสอบที่มีความสามารถเท่าเทียมกัน แต่เป็นสมาชิกของประชากรในกลุ่มย่อยที่มีลักษณะแตกต่างกัน เดิมเรียก ความลำเอียงของข้อสอบ (Item bias) ต่อมามีข้อแย้งว่าความลำเอียงของข้อสอบมีความเกี่ยวข้องกับการรวบรวมหลักฐานเชิงประจักษ์ที่เกี่ยวข้องกับการทำข้อสอบของผู้ตอบข้อสอบจาก 2 กลุ่มย่อย (เช่น เพศ เชื้อชาติ)

หลักฐานเชิงประจักษ์ของผลการสอบที่ต่างกันของผู้ตอบข้อสอบ 2 กลุ่ม ไม่เพียงพอที่จะสรุปว่าข้อสอบมีความลำเอียง เนื่องจากความลำเอียงของข้อสอบให้ความหมายอย่างน้อย 2 อย่าง คือ ความหมายในเชิงสังคมและความหมายในเชิงสถิติ ดังนั้น การตรวจสอบด้วยวิธีการทางสถิติโดยอาศัยข้อมูลเชิงประจักษ์จึงใช้คำว่าการทำหน้าที่เบี่ยงเบนของข้อสอบ (Differential item functioning: DIF) จะเหมาะสมกว่า จากการศึกษามีผู้นิยามความหมายของการทำหน้าที่ต่างกันของข้อสอบไว้ เช่น

การทำหน้าที่ต่างกันของข้อสอบ หมายถึง การสังเกตได้ว่าข้อสอบข้อนั้นแสดงคุณสมบัติทางสถิติที่ต่างกัน เมื่อใช้ข้อมูลจากผู้ตอบข้อสอบที่มีความสามารถเท่ากัน แต่อยู่ในกลุ่มที่ต่างกัน (Angoff, 1993)

การทำหน้าที่ต่างกันของข้อสอบ หมายถึง เหตุการณ์ที่เกิดเมื่อผู้สอบมีความสามารถ (Trait) เท่ากัน แต่มีลักษณะประชากรต่างกัน ทำให้ความน่าจะเป็นที่จะตอบข้อนั้น ๆ ถูกต่างกัน (Roussous & Stout, 1996 cited in Sedivy, 2009)

การทำหน้าที่ต่างกันของข้อสอบ หมายถึง คำที่ใช้อธิบายลักษณะข้อสอบในแบบสอบที่มีพฤติกรรมต่างไปเมื่อกลุ่มของผู้สอบมีความแตกต่างกัน (Wainer, Bradlow & Wang, 2007)

การทำหน้าที่ต่างกันของข้อสอบ หมายถึง ข้อสอบที่แสดงถึงการตอบสนองจากประชากร 2 กลุ่มที่มีความสามารถเหมือนกัน แต่มีความน่าจะเป็นในการตอบถูกต่างกัน (Chaimongkol, Huffer & Kamata, 2007)

การทำหน้าที่ต่างกันของข้อสอบ ในมุมมองของทฤษฎีการตอบสนองข้อสอบ (Item response theory: IRT) แสดงถึงโค้งลักษณะข้อสอบ (Item characteristic curves: ICCs) ที่ต่างกันเมื่อลักษณะของสมาชิกกลุ่มย่อยต่างกัน (Narayanan & Swaminathan, 1996 cited in Fukuhara, 2009)

การทำหน้าที่ต่างกันของข้อสอบ หมายถึง การที่ข้อสอบทำให้ผู้สอบจากต่างกลุ่มกันที่มีความสามารถหรือคุณลักษณะที่มุ่งวัดเท่ากัน มีโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกัน หรือมีฟังก์ชันการตอบสนองข้อสอบแตกต่างกัน (ศิริชัย กาญจนาวาสี, 2550)

การทำหน้าที่ต่างกันของข้อสอบ หมายถึง การที่ข้อสอบทำให้ผู้สอบที่มีลักษณะหรือมาจากต่างกลุ่มกันที่มีความสามารถหรือคุณลักษณะที่มุ่งวัดเท่ากัน มีโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกัน หรือมีฟังก์ชันการตอบสนองข้อสอบแตกต่างกัน (อิทธิฤทธิ์ พงษ์ปิยะรัตน์, 2551)

การทำหน้าที่ต่างกันของข้อสอบ หมายถึง โอกาสของการตอบข้อสอบได้ถูกต้องแตกต่างกัน สำหรับผู้สอบที่มีคุณลักษณะหรือความสามารถในระดับเดียวกัน แต่มาจากกลุ่มประชากรย่อยที่แตกต่างกัน (สุพัฒนา หอมบุปผา, 2556)

จากความหมายของการทำหน้าที่ต่างกันของข้อสอบที่กล่าวมา สรุปได้ว่า การทำหน้าที่ต่างกันของข้อสอบ หมายถึง ข้อสอบที่ทำให้ผู้สอบซึ่งมีลักษณะต่างกัน แต่มีความสามารถระดับเดียวกัน มีโอกาสตอบถูกไม่เท่ากัน

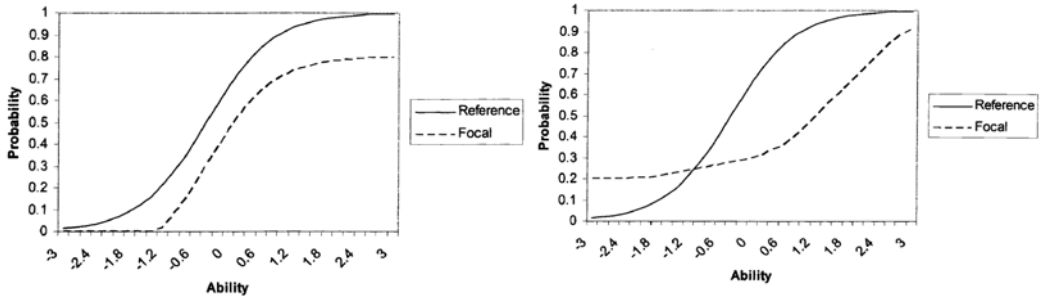
ประเภทของการทำหน้าที่ต่างกันของข้อสอบ

การทำหน้าที่ต่างกันของข้อสอบ เป็นการเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่ม ผู้สอบอย่างน้อย 2 กลุ่มขึ้นไป ปกตินิยมเปรียบเทียบ 2 กลุ่ม คือ กลุ่มแรก เป็นกลุ่มที่คาดว่าจะเป็นกลุ่มที่เสียประโยชน์จากข้อสอบที่แสดงออกถึงการทำหน้าที่ต่างกันของข้อสอบและเป็นกลุ่มที่นักวิจัยสนใจศึกษา เรียกว่า กลุ่มเปรียบเทียบ (Focal group หรือกลุ่ม F) ส่วนกลุ่มที่สองเป็นกลุ่มที่คาดว่าจะเป็นกลุ่มที่ได้ประโยชน์จากข้อสอบที่แสดงออกถึงการทำหน้าที่ต่างกันของข้อสอบ เรียกว่า กลุ่มอ้างอิง (Reference group หรือกลุ่ม R)

การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ ได้แบ่งลักษณะของข้อสอบที่หน้าที่ต่างกันแบ่งเป็น 2 ประเภท ได้แก่ ข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูป (Uniform DIF) และข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูป (Non - uniform DIF) (ศิริชัย กาญจนาวาสี, 2550; Sedivy, 2009)

1. ข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูป (Uniform DIF) หมายถึง ข้อสอบที่ทำให้ผู้สอบกลุ่มหนึ่งมีโอกาสตอบข้อสอบถูกมากกว่าผู้สอบอีกกลุ่มหนึ่งสม่ำเสมอกันทุกระดับความสามารถ เมื่อพิจารณาโค้งคุณลักษณะข้อสอบ (ICC) ของผู้สอบ 2 กลุ่ม พบว่าไม่มีปฏิสัมพันธ์ (Interaction) ระหว่างโค้งคุณลักษณะผู้สอบในทุกระดับความสามารถ ดังภาพที่ 2 - 1 (ก)

2. การทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป (Nonuniform DIF) หมายถึง ข้อสอบที่ทำให้โอกาสการตอบข้อสอบถูกต้องของผู้สอบระหว่างกลุ่มไม่สม่ำเสมอกันทุกระดับความสามารถ เมื่อพิจารณาโค้งคุณลักษณะข้อสอบของผู้สอบทั้ง 2 กลุ่ม พบว่ามีปฏิสัมพันธ์ร่วมกันระหว่างโค้งคุณลักษณะ เช่น ที่ระดับความสามารถหนึ่งผู้สอบกลุ่มอ้างอิงมีโอกาสในการตอบข้อสอบถูกมากกว่าผู้สอบกลุ่มเปรียบเทียบ แต่อีกที่ระดับความสามารถหนึ่ง ผู้สอบกลุ่มเปรียบเทียบมีโอกาสในการตอบข้อสอบถูกมากกว่าผู้สอบกลุ่มอ้างอิง ดังภาพที่ 2 - 1 (ข)



(ก) (ข)

ภาพที่ 2 - 1 โค้งคุณลักษณะข้อสอบที่ทำหน้าที่ต่างกัน (ก) แบบเอกรูป (Uniform DIF)
(ข) แบบอเนกรูป (Nonuniform DIF)

โดยทั่วไปในแบบสอบมาตรฐานมักมีข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูปมากกว่าข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูป แต่ในข้อมูลจริงจะมีข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปได้มากกว่า (สุพัฒนา หอมบุปผา, 2556)

เนื่องจากในการทำหน้าที่ต่างกันของข้อสอบ คือ การที่ข้อสอบวัดความสามารถรอง (Secondary abilities หรือ Nuisance dimension หรือ η) หรือ คุณลักษณะแฝงอื่นนอกเหนือจากความสามารถหลัก (Primary abilities หรือ Primary dimension หรือ θ) หรือ คุณลักษณะแฝงที่ต้องการวัด ที่จะส่งผลให้ผู้สอบต่างกลุ่มที่นำเข้ามาจับคู่เปรียบเทียบกัน มีโอกาสในการตอบข้อสอบได้ถูกต้องต่างกัน ทั้งๆ ที่มีความสามารถหลักที่ต้องการวัดเท่ากัน นั่นคือ การทำหน้าที่ต่างกันจะเกิดเมื่อมีค่าเฉลี่ย η ไม่เท่ากัน ซึ่งการตัดสินการทำหน้าที่ต่างกันของข้อสอบมีได้ 4 สถานการณ์ ดังนี้ (Sedivy, 2009)

1. กลุ่มอ้างอิงและกลุ่มเปรียบเทียบมีค่าเฉลี่ย θ และ η เท่ากัน แสดงว่าไม่มีการทำหน้าที่ต่างกัน (No Bias/DIF)

2. กลุ่มอ้างอิงและกลุ่มเปรียบเทียบมีค่าเฉลี่ย η เท่ากัน แต่มีค่าเฉลี่ย θ ต่างกัน แสดงว่ามี Impact นั่นคือ ความน่าจะเป็นในการตอบข้อสอบถูกมาจากความแตกต่างของความสามารถหลัก

3. กลุ่มอ้างอิงและกลุ่มเปรียบเทียบมีค่าเฉลี่ย θ เท่ากัน แต่มีค่าเฉลี่ย η ต่างกัน แสดงว่ามีการทำหน้าที่ต่างกัน

4. กลุ่มอ้างอิงและกลุ่มเปรียบเทียบมีค่าเฉลี่ยแตกต่างกันทั้ง θ และ η แสดงว่ามีการทำหน้าที่ต่างกัน และมี Impact

หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมีหลายวิธี เนื่องจากมีวิธีการศึกษาและการคิดค้นวิธีการต่าง ๆ เพื่อให้สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้อย่างมีประสิทธิภาพมากที่สุด ซึ่งสามารถแบ่งตามประเภทการวิเคราะห์ได้เป็น 2 กลุ่ม ดังนี้

1. กลุ่มที่ใช้คะแนนสังเกตได้ (Observe score) เป็นกลุ่มที่ใช้คะแนนรวมของแบบสอบเป็นเกณฑ์ในการจับคู่ผลสอบสองกลุ่มตามความรู้ หรือความสามารถที่แท้จริงของผู้สอบ วิธีการในกลุ่มนี้ ได้แก่ วิธีแมนเทล - แฮนส์เซล วิธีถดถอยโลจิสติก และวิธีทำให้เป็นมาตรฐาน จุดเด่นของวิธีการในกลุ่มนี้ คือ กลุ่มตัวอย่างขนาดเล็ก การวิเคราะห์ไม่ยุ่งยากซับซ้อน ส่วนจุดด้อยของวิธีการในกลุ่มนี้ คือ ค่าสถิติจะเปลี่ยนไปตามขนาดของกลุ่มตัวอย่าง เมื่อกลุ่มตัวอย่างที่ศึกษาเปลี่ยนไปผลการศึกษาก็อาจเปลี่ยนแปลงไปด้วย

2. กลุ่มที่ใช้คะแนนที่สังเกตไม่ได้ หรือเป็นตัวแปรแฝง (Latent variable) เป็นกลุ่มวิธีที่มีทฤษฎีการทดสอบเป็นพื้นฐาน ใช้การประมาณค่าคุณลักษณะแฝง (Latent trait) หรือใช้คะแนนจริงของผู้สอบเป็นเกณฑ์ในการจับคู่เปรียบเทียบผู้สอบ วิธีการในกลุ่มนี้ ได้แก่ วิธีการตอบสนองข้อสอบ (IRT) และวิธีซิบเทสท์ (SIBTEST) เป็นต้น

ศิริชัย กาญจนวาสี (2550) ได้เสนอหลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างกลุ่มอ้างอิง และกลุ่มเปรียบเทียบที่ต้องใช้วิธีการจับคู่ตามเกณฑ์ความสามารถ เพราะเป็นเงื่อนไขสำคัญของการทำหน้าที่ต่างกันของข้อสอบ เกณฑ์การจับคู่ที่นิยมมี 2 วิธีที่สำคัญดังนี้

1. เกณฑ์ภายนอก (External criterion)

การทำหน้าที่ต่างกันของข้อสอบโดยใช้เกณฑ์ภายนอกนี้ สามารถนำไปใช้ได้ทั้งข้อสอบรายข้อ และแบบสอบทั้งฉบับ โดยใช้คะแนนจากแบบสอบอื่นเป็นเกณฑ์ภายนอก แล้วใช้เทคนิคการวิเคราะห์ถดถอย เพื่อทำการเปรียบเทียบเส้นกราฟความสัมพันธ์ระหว่างตัวแปรเกณฑ์กับตัวแปรทำนายระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ

หลักการนี้มีจุดมุ่งหมาย เพื่อสร้างสมการทำนายตัวแปรเกณฑ์ ซึ่งเป็นคะแนนของแบบทดสอบอื่นจากตัวแปรทำนายที่เป็นคะแนนรายข้อหรือคะแนนแบบทดสอบ ระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ ในการวิเคราะห์การทำหน้าที่ต่างกันของแบบทดสอบ จะใช้คะแนนรวมของแบบทดสอบทั้งฉบับเป็นตัวแปรทำนาย สำหรับตัวแปรเกณฑ์ที่ใช้เป็นเกณฑ์ภายนอกอาจใช้คะแนนรวมทั้งฉบับหรือเกรดเฉลี่ย หรือคะแนนจากงานที่เกี่ยวข้องของผู้สอบ สมการทำนายสำหรับกลุ่มอ้างอิง คือ $Y_i = A_R + B_R X_i$ และกลุ่มเปรียบเทียบ คือ $Y_i = A_F + B_F X_i$ โดยที่ Y_i เป็นคะแนนของตัวแปรเกณฑ์ภายนอก X_i เป็นคะแนนของตัวแปรทำนาย A เป็นค่าคงที่หรือจุดตัดแกน y (Intercept) และ B เป็นค่าความชัน (Slope)

จากฟังก์ชันการทำนายดังกล่าว สามารถเปรียบเทียบค่าตัดแกน (A) และค่าความชัน (B) ของเส้นกราฟระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบได้ ถ้าเส้นกราฟมีค่าความชันหรือค่าตัดแกนแตกต่างสำหรับข้อสอบใด แสดงว่าข้อสอบหรือแบบสอบนั้น มีการทำหน้าที่ต่างกัน โดยเข้าข้างกลุ่มผู้สอบที่มีค่าตัดแกนหรือค่าความชันที่สูงกว่า

การใช้เกณฑ์ภายนอกมีข้อดี คือ เกณฑ์ที่ใช้มีความเป็นอิสระจากข้อสอบ และแบบสอบที่ต้องการตรวจสอบ แต่มีจุดอ่อน คือ ความเหมาะสมของเกณฑ์ที่จะนำมาใช้ในทางปฏิบัติเป็นการยากที่จะหาเกณฑ์ภายนอกจากแบบสอบฉบับอื่นที่มีความตรงเชิงทำนาย และมีความยุติธรรมสำหรับกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ถ้าเกณฑ์ภายนอกขาดคุณสมบัติดังกล่าว จะทำให้ผลวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบหรือแบบทดสอบขาดความแม่นยำและสมบูรณ์

2. เกณฑ์ภายใน (Internal criterion)

การวิเคราะห์การทำหน้าที่ต่างกันโดยใช้เกณฑ์ภายในเป็นการนำวิธีการทางสถิติตรวจสอบการทำหน้าที่ต่างกันของข้อสอบหรือแบบสอบ โดยเน้นการพิจารณาจากโครงสร้างภายในของแบบสอบเป็นหลัก ด้วยการวิเคราะห์ผลจากการตอบข้อสอบและความสามารถ หรือคะแนนจริงของผู้สอบที่ได้จากแบบสอบฉบับนั้น เพื่อนำมาเปรียบเทียบระหว่างผู้สอบจากกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ที่มีความสามารถหรือคะแนนจริงเท่ากัน ว่าจะมีผลการตอบหรือโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกันหรือไม่ เพื่อบ่งชี้ถึงการทำหน้าที่ต่างกันของข้อสอบการวิเคราะห์ลักษณะนี้นิยมใช้ค่าสถิติต่าง ๆ เป็นตัวบ่งชี้ถึงการทำหน้าที่ต่างกันของข้อสอบ ซึ่งที่นิยมมีดังนี้

2.1 การทดสอบปฏิสัมพันธ์ (Interaction)

ระยะเริ่มต้นของการศึกษาการทำหน้าที่ต่างกันของข้อสอบมีการใช้สถิติทดสอบเอฟ (F - test) จากการวิเคราะห์ความแปรปรวน (ANOVA) เพื่อทดสอบปฏิสัมพันธ์ระหว่างกลุ่มผู้สอบกับข้อสอบ หากผลการทดสอบมีนัยสำคัญก็ชี้ได้ว่ามีการทำหน้าที่ต่างกันของข้อสอบแล้ววิเคราะห์ต่อด้วยวิธี Post Hoc เพื่อระบุข้อสอบที่มีผลต่อการเกิดปฏิสัมพันธ์ ซึ่งเป็นข้อที่ทำหน้าที่ต่างกัน

2.2 การวัดความเบี่ยงเบนสัมพัทธ์ (Relative deviation)

วิธีการนี้เป็นการคำนวณค่าพารามิเตอร์ข้อสอบจำแนกตามกลุ่ม แล้วแปลงให้เป็นค่าความยากมาตรฐาน (Δ) สามารถนำมาเขียนเป็นกราฟเปรียบเทียบเป็นรายข้อ ถ้าข้อใดมีค่าเบี่ยงเบนไปจากแกนหลักที่คาดหมาย หรือเบี่ยงเบนเกินจากความคลาดเคลื่อนมาตรฐานของค่าความยากที่กำหนดก็ย่อมเป็นเครื่องบ่งชี้ถึงการทำหน้าที่ต่างกันของข้อสอบ

วิธีการนี้มีข้อดีและข้อเสียคล้ายกับการทดสอบปฏิสัมพันธ์ และค่าความยากของข้อสอบไม่ใช่ตัวแทนของค่าความยากที่แท้จริงของข้อสอบ และอาจได้รับอิทธิพลจากตัวแปรอื่น เช่น ค่าอำนาจจำแนกและความสามารถของผู้เข้าสอบ

2.3 การเปรียบเทียบน้ำหนักองค์ประกอบ (Factor loading)

วิธีการนี้ใช้การวิเคราะห์องค์ประกอบ (Factor analysis) เป็นเทคนิคทางสถิติที่นิยมใช้ตรวจสอบความตรงเชิงโครงสร้าง (Construct validity) โดยนำการวิเคราะห์องค์ประกอบมาใช้ในการวิเคราะห์โครงสร้างของแบบทดสอบแยกตามกลุ่มผู้เข้าสอบ ความไม่สอดคล้องกันระหว่างน้ำหนักองค์ประกอบบนคุณลักษณะสำคัญในสิ่งที่มุ่งวัดหรือความแตกต่างของค่าเฉลี่ยคะแนนองค์ประกอบ (Factor score) ระหว่างกลุ่มผู้เข้าสอบย่อมสะท้อนการทำหน้าที่ต่างกันของข้อสอบและแบบทดสอบ

ในการใช้เทคนิคการวิเคราะห์องค์ประกอบเชิงสำรวจ (Exploratory factor analysis: EFA) สำหรับศึกษาการทำหน้าที่ต่างกันของข้อสอบ มีจุดอ่อนเรื่องความไม่สอดคล้องระหว่างน้ำหนักองค์ประกอบ อาจเกิดจากความแตกต่างของความสามารถระหว่างกลุ่มก็ได้ แนวทางที่เหมาะสม จึงควรใช้เทคนิคการวิเคราะห์องค์ประกอบเชิงยืนยัน (Confirmatory factor analysis: CFA) นอกจากนี้ยังใช้ CFA สำหรับตรวจสอบความแตกต่างระหว่างกลุ่มในลักษณะความสามารถหลักหรือความสามารถรองได้อีก

2.4 การเปรียบเทียบโอกาสตอบข้อสอบถูก

วิธีการนี้จะเปรียบเทียบโอกาสของการตอบข้อสอบถูกของผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่มีความสามารถเท่ากัน เป็นแนวทางที่ได้รับความนิยมมาก ซึ่งการบ่งชี้การทำหน้าที่ต่างกันของข้อสอบมีการคำนวณค่าสถิติใน 2 แนวทางหลัก ดังนี้

2.4.1 การเปรียบเทียบค่าสัดส่วนความน่าจะเป็นในการตอบข้อสอบถูกของผู้สอบต่างกลุ่มที่มีระดับความสามารถเท่ากัน

2.4.2 การเปรียบเทียบค่าฟังก์ชั่นการตอบสนองข้อสอบหรือโค้งลักษณะข้อสอบระหว่างกลุ่มที่มีระดับความสามารถเท่ากัน ซึ่งวิธีการนี้ตั้งอยู่บนหลักการของทฤษฎี IRT เช่น วิธีการวัดความแตกต่างของค่าพารามิเตอร์ความยาก วิธีการวัดความแตกต่างของพื้นที่ เป็นต้น วิธีการนี้มีข้อดี คือ การคำนวณค่าสถิติมีความแกร่ง เชื่อถือได้ มีกลไกในการควบคุมความสามารถของผู้สอบโดยการจับคู่ความสามารถเพื่อทำการเปรียบเทียบ ณ ตำแหน่งที่มีความสามารถเท่ากัน

วิธีการนี้มีข้อดี คือ การคำนวณค่าสถิติของข้อสอบมีความน่าเชื่อถือ มีการควบคุมความสามารถของผู้เข้าสอบโดยการจับคู่กลุ่มความสามารถ เพื่อเปรียบเทียบ ณ ตำแหน่งต่าง ๆ ที่มี

ความสามารถระดับเท่ากัน จึงเป็นวิธีการที่ยอมรับกันโดยทั่วไป แต่ก็มีข้อจำกัด คือ ความซับซ้อนของแนวคิดพื้นฐาน และการใช้โปรแกรมสำเร็จรูปเพื่อวิเคราะห์ข้อมูลได้บางโปรแกรม

เกณฑ์และวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

จากการศึกษามีผู้จำแนกลักษณะของการทำหน้าที่ต่างกันของข้อสอบ ดังนี้

Potenza & Dorans (1995, อ้างถึงใน ศิริชัย กาญจนาวาสี, 2550) จำแนกวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนแบบสองค่า (Dichotomously score) คือ การให้คะแนน 0 - 1 และข้อสอบที่มีการให้คะแนนแบบหลายค่า (Polytomous score) โดยแบ่งเป็น 2 มิติ ดังนี้

มิติแรก แบ่งตามประเภทการวิเคราะห์ระหว่างกลุ่มที่ใช้คะแนนสังเกตได้ (Observed score) และกลุ่มที่ใช้คะแนนสังเกตไม่ได้ หรือตัวแปรแฝง (Latent variable) โดยกลุ่มที่ใช้คะแนนสังเกตได้ มักวิเคราะห์ตามทฤษฎีการทดสอบแบบดั้งเดิม (Classical test theory: CTT) หรือกลุ่มที่ไม่ใช้ทฤษฎีการตอบสนองข้อสอบ (non - IRT approach) โดยใช้คะแนนรวมของผู้สอบเป็นเกณฑ์การจับกลุ่มผู้สอบ วิธีการตรวจสอบที่สำคัญในกลุ่มนี้ ได้แก่ การวิเคราะห์ความแปรปรวน (Analysis of variance: ANOVA) การวิเคราะห์การถดถอยโลจิสติก (Logistic regression: LR) วิธีแปลงค่าความยากของข้อสอบ (Transformed item difficulty: TID) วิธีแมนเทล - แฮนด์เซล (Mantel-Haenszel: MH) ส่วนกลุ่มที่ใช้คะแนนสังเกตไม่ได้ จะวิเคราะห์บนพื้นฐานของทฤษฎีการตอบสนองข้อสอบ (IRT) สำหรับใช้เป็นเกณฑ์จับคู่ กลุ่มผู้สอบวิธีการตรวจสอบที่สำคัญในกลุ่มนี้ คือ วิธีวัดพื้นที่ความแตกต่างระหว่างโค้งการตอบสนองข้อสอบ (IRT - D2) วิธีอัตราส่วนไลด์ลิฮู้ดลอกลิเนียร์ (Loglinear IRT likelihood ratio) วิธี Lord's χ^2 และวิธี SIBTEST

มิติที่สอง เป็นการแบ่งระหว่าง Parametric Approaches ซึ่งวิเคราะห์โดยมีข้อตกลงเบื้องต้นของโมเดลสำหรับอธิบายความสัมพันธ์ระหว่างคะแนนของข้อสอบและการจับคู่ตัวแปร และ Nonparametric Approaches วิเคราะห์ดัชนีการทำหน้าที่ต่างกันของข้อสอบโดยไม่มีข้อตกลงเบื้องต้นของโมเดลและการจับคู่ตัวแปร

ซึ่งศิริชัย กาญจนาวาสี (2550) แสดงรายชื่อวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่สำคัญ ๆ ดังแสดงในตารางที่ 2 - 1

ตารางที่ 2 - 1 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ที่มีการตรวจให้คะแนนแบบ Dichotomous DIF และ Polytomous DIF

ประเภทและตัวแปรเกณฑ์	Parametric	Nonparametric
1. Dichotomous DIF		
1.1 กลุ่มที่ใช้คะแนนสังเกตได้	ANOVA	TID, MH, STND,
(Observed score)	Logistic regression	SIBTEST
1.2 กลุ่มที่ใช้คะแนนสังเกตไม่ได้	IRT-D^2, Lord's χ^2	
(Latent variable)	General IRTLR	
	Loglinear IRTLR	
2. Polytomous DIF		
2.1 กลุ่มที่ใช้คะแนนสังเกตได้	ANOVA	Polytomous STND
(Observed score)	Polytomous Logistic	GMH
	Regression	
2.2 กลุ่มที่ใช้คะแนนสังเกตไม่ได้	General IRTLR	Polytomous SIBTEST
(Latent variable)	PCM	GPCM

จากการเปรียบเทียบวิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในตารางที่ 2 - 1 พบว่า วิธี ANOVA วิธี χ^2 และวิธี TID เป็นการวิเคราะห์ข้อสอบโดยอาศัยทฤษฎีการวัดแบบดั้งเดิม ซึ่งมีจุดด้อย คือ ค่าพารามิเตอร์ของข้อสอบจะเปลี่ยนไปตามกลุ่มผู้สอบ นอกจากนี้วิธีANOVA และวิธี χ^2 ไม่มีดัชนีบอกระดับการทำหน้าที่ต่างกันของข้อสอบ แต่เป็นวิธีที่ประหยัดและใช้กลุ่มตัวอย่างน้อย วิธี IRT เป็นวิธีที่วิเคราะห์ความแตกต่างของฟังก์ชันการตอบข้อสอบระหว่างกลุ่มผู้สอบ มีดัชนีบอกระดับของการทำหน้าที่ต่างกันของข้อสอบและทดสอบความมีนัยสำคัญทางสถิติ ซึ่งเป็นวิธีที่ให้รายละเอียดมากและมีข้อดี คือ ความไม่แปรเปลี่ยนของค่าพารามิเตอร์ แต่มีข้อเสีย คือ ค่อนข้างสิ้นเปลือง วิธี MH คล้ายกับวิธี χ^2 คือใช้คะแนนรวมจากแบบสอบเป็นตัวแทนของความสามารถ แต่วิธี MH จะวิเคราะห์ทีละระดับความสามารถ และมีดัชนีบอกระดับการทำหน้าที่ต่างกันของข้อสอบและมีการทดสอบความนัยสำคัญทางสถิติ วิธี SIBTEST ใช้คะแนนรวมจากแบบสอบเป็นตัวแทนความสามารถ มีข้อตกลงว่ามีมิติการวัด 2 มิติ ดังนั้น คะแนนจากแบบสอบจึงมี 2 ส่วนคือ คะแนนจากแบบสอบที่มีความตรง (Valid subtest) ซึ่งวัดคุณลักษณะแฝงเป้าหมาย และคะแนนจากแบบสอบที่ศึกษา (Studied subtest) ซึ่งวัดคุณลักษณะแฝงแทรกซ้อน มีดัชนีบอกระดับ

การทำหน้าที่ต่างกันของข้อสอบและทดสอบนัยสำคัญทางสถิติ ทั้งวิธี MH และวิธี SIBTEST เป็นวิธีที่ประหยัด ใช้กลุ่มตัวอย่างน้อย

คมศักดิ์ ชื่นชม (2539) จำแนกวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามทฤษฎีพื้นฐาน ได้เป็น 2 กลุ่ม ดังนี้

กลุ่มที่ 1 กลุ่มที่ใช้หลักการของทฤษฎีการตอบสนองข้อสอบ (IRT) ได้แก่ วิธีโค้งคุณลักษณะข้อสอบแบบ 3 พารามิเตอร์ (Item characteristic curve - 3 parameter: ICC - 3) วิธีโค้งคุณลักษณะข้อสอบแบบ 2 พารามิเตอร์ (Item characteristic curve - 2 parameter: ICC - 2) วิธีโค้งคุณลักษณะข้อสอบแบบ 1 พารามิเตอร์ (Item characteristic curve - 1 parameter: ICC - 1) วิธี Likelihood Ratio Test วิธี SIBTEST และ วิธี Lord's χ^2 test

กลุ่มที่ 2 กลุ่มที่ใช้หลักการของทฤษฎีการทดสอบแบบดั้งเดิม (Classical test theory: CTT) ได้แก่ วิธีแปลงค่าความยากของข้อสอบ (Transformed item difficulty: TID) วิธีวิเคราะห์ความแปรปรวน (Analysis of variance: ANOVA) วิธีวิเคราะห์ด้วยไคสแคว์ (Chi-square: χ^2) วิธีวิเคราะห์องค์ประกอบ (Factor analysis) วิธีวิเคราะห์การถดถอย (Regression analysis) วิธีค่าอำนาจจำแนกของข้อสอบ (Item discrimination indices) วิธีลอกลิเนียร์ (Log - linear) วิธีแมนเทล - เฮนส์เซล (Mantel - Haenszel: MH) วิธีทำให้เป็นมาตรฐาน (Standardization: STND) และวิธีถดถอยโลจิสติก (Logistic regression: LR)

จะเห็นว่าวิธีในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมีหลายวิธี วิธีที่เป็นที่รู้จักและนิยมใช้ในการเปรียบเทียบที่ผ่านมา กาญจนา วัธนสุนทร (2538) ได้สรุปว่าวิธีที่นิยมใช้มี 6 วิธี ได้แก่ วิธีแปลงค่าความยาก (TID) วิธีวิเคราะห์ความแปรปรวน (ANOVA) วิธีไคสแควร์ (χ^2) วิธีทฤษฎีการตอบสนองข้อสอบ (IRT) วิธี Mantel - Haenszel (MH) และวิธี SIBTEST ซึ่งสรุปวิธีการวิเคราะห์ ข้อดี ข้อจำกัดไว้ ดังตารางที่ 2 - 2 (กาญจนา วัธนสุนทร, 2537)

ตารางที่ 2 - 2 เปรียบเทียบวิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

ประเด็น	TID	ANOVA	χ^2	IRT	MH	SIBTEST
ข้อตกลงเบื้องต้น	ผลรวมระหว่างกลุ่มกับข้อกระทงเป็นตัวบ่งชี้ความลำเอียง	ความแปรปรวนและความแปรปรวนร่วมของข้อสอบต้องเท่ากัน	คะแนนรวมจากแบบสอบเป็นตัวแทนความสามารถของผู้สอบ	แบบสอบมีความเป็นมิติเดียวและ ICC สามารถแสดงฟังก์ชันของค่าความสามารถและโอกาสการตอบข้อสอบถูก	คะแนนรวมจากแบบสอบเป็นตัวแทนความสามารถของผู้สอบ	คะแนนรวมจากแบบสอบเป็นตัวแทนความสามารถของผู้สอบและมีมิติการวัด 2 มิติ คือคุณลักษณะแฝงเป้าหมายและคุณลักษณะแฝงแทรกซ้อน
สิ่งที่วิเคราะห์	ผลร่วมระหว่างการเป็นสมาชิกในกลุ่มกับการตอบถูก	ผลร่วมระหว่างการเป็นสมาชิกในกลุ่มกับการตอบถูก	ความแตกต่างของอัตราส่วนการตอบถูกตามระดับคะแนนรวม	ความแตกต่างของฟังก์ชันการตอบข้อสอบที่ระดับความสามารถเดียวกัน	ความแตกต่างของอัตราส่วนการตอบระหว่างผู้ที่มีความสามารถระดับเดียวกัน	ความต่างของคะแนนเฉลี่ยและอัตราส่วนการตอบข้อสอบระหว่างผู้ที่มีความสามารถระดับเดียวกัน
การตัดสิน DIF	ระยะห่างของจุด Δ จากเส้นแกนหลัก	ความมีนัยสำคัญทางสถิติของ F-test	ความมีนัยสำคัญทางสถิติของ χ^2	พื้นที่ระหว่างโค้งลักษณะข้อสอบ	ค่าดัชนี α_{MH} และความมีนัยสำคัญทางสถิติ	ค่าดัชนี β_{SIB} และความมีนัยสำคัญทางสถิติ
ทฤษฎีพื้นฐาน	CTT	CTT	CTT	IRT	CTT	MIRT
ข้อดี	คำนวณง่าย ใช้กลุ่มตัวอย่างน้อย	ใช้กลุ่มตัวอย่างน้อย	คำนวณง่าย มีเกณฑ์ตายตัวในการแปลผล	ค่าพารามิเตอร์ไม่แปรเปลี่ยนตามกลุ่มผู้สอบ	คำนวณง่าย ใช้กลุ่มตัวอย่างน้อย ประหยัด	คำนวณง่าย ใช้กลุ่มตัวอย่างน้อย ตรวจสอบ DIF ได้หลายข้อในครั้งเดียว
ข้อจำกัด	มีความคลาดเคลื่อนเมื่อค่า a สูง และค่า b เปลี่ยนตามกลุ่มผู้สอบ	การคำนวณค่อนข้างยุ่งยากและไม่มีดัชนีบอกระดับความลำเอียง	ไม่มีเกณฑ์ตายตัวในการกำหนดช่วงคะแนนและค่า b เปลี่ยนตามกลุ่มผู้สอบ	มีการคำนวณซับซ้อนหลายรอบ แปลผลยาก ใช้กลุ่มตัวอย่างมาก ค่าใช้จ่ายสูง	ไม่มีความไวในการตรวจสอบ DIF แบบ Nonuniform DIF	อัตราความคลาดเคลื่อนชนิดที่ 1 เพิ่มสูงเมื่อคะแนนเฉลี่ยแตกต่างกันมาก

ในการศึกษาครั้งนี้ ใช้หลักการของทฤษฎีการตอบสนองข้อสอบ (IRT) โดยวิธีแบบ Parametric ด้วยการใช้โมเดล Bifactor MIRT ซึ่งจะอธิบายในตอนที่ 3 แนวคิดเกี่ยวกับทฤษฎีการตอบสนองข้อสอบสำหรับเทสต์เลท

ตอนที่ 2 แนวคิดเกี่ยวกับทฤษฎีการตอบสนองข้อสอบ

การวัดทางจิตวิทยามักนิยมใช้ทฤษฎีการทดสอบแบบดั้งเดิม ซึ่งมีแนวคิดว่า คะแนนที่สังเกตได้ (Observed score: X) เป็นผลรวมของโมเดลเชิงเส้นของคะแนนจริง (True score: T) และคะแนนความคลาดเคลื่อน (Error: E) หรือเขียนเป็นสมการได้ว่า X = T + E แนวคิดนี้มีการพัฒนาสืบเนื่องกันมาและได้รับความนิยมอย่างแพร่หลาย อย่างไรก็ตาม ผู้เชี่ยวชาญด้านการวัดหลายท่านพบว่า ทฤษฎีดังกล่าวมีจุดอ่อนที่สำคัญหลายประการดังต่อไปนี้ (Hambleton, Swaminathan & Rogers, 1991)

1. พารามิเตอร์ของข้อสอบ (Item parameter) ได้แก่ ความยาก (Item difficulty) อำนาจจำแนก (Item discrimination) ไม่คงที่ ขึ้นอยู่กับกลุ่มผู้สอบในแต่ละครั้ง เช่น เมื่อแบบสอบยาก แสดงว่า ความสามารถของผู้สอบน้อย และหากแบบสอบง่าย แสดงว่าความสามารถของผู้สอบมาก เนื่องจากความยากเป็นอัตราส่วนของผู้ที่ตอบถูก ทำให้แบบสอบจะง่ายหรือยากขึ้นอยู่กับความสามารถ เป็นต้น นอกจากนั้นยังไม่สามารถพิจารณาพารามิเตอร์ของข้อสอบแยกจากความสามารถของผู้สอบแต่ละคนได้อย่างอิสระ คุณลักษณะทั้งสองส่วนขึ้นอยู่กับการแจกแจงลักษณะของกลุ่มตัวอย่างที่ใช้ในการทดสอบเฉพาะครั้ง จึงทำให้ความคาดเคลื่อนมาตรฐานเป็นค่าเฉพาะในการสอบแต่ละครั้ง และเป็นค่าเดียวกันสำหรับผู้สอบทุกคน

2. การให้คะแนนผู้สอบแต่ละคน คิดจากคะแนนรวมในการทำแบบสอบ เป็นการพิจารณาระดับแบบสอบมากกว่าการพิจารณารูปแบบการตอบข้อสอบรายข้อ ดังนั้น การแปลความหมายของคะแนน จึงบอกไม่ได้ว่า ผู้สอบมีความสามารถหรือบกพร่องด้านใด นอกจากนี้ยังมีปัญหาในการใช้งานแบบสอบคู่ขนาน นั่นคือ ไม่สามารถนำคะแนนจากการทำแบบสอบต่างฉบับมาเทียบกันได้ เพราะคะแนนจากแบบสอบทั้งสองมีมาตรในการวัด (Scale) ต่างกัน เนื่องจากค่าสถิติในการพิจารณาคัดเลือกข้อสอบมาจัดทำแบบสอบแต่ละฉบับขึ้นอยู่กับกลุ่มผู้สอบ

3. การคำนวณค่าความเที่ยง (Reliability) ของแบบสอบ เป็นความสัมพันธ์ระหว่างคะแนนจริงจากแบบสอบคู่ขนาน ซึ่งยากในทางปฏิบัติที่จะสร้างแบบสอบที่เป็นคู่ขนานกันได้

4. ทฤษฎีการทดสอบแบบดั้งเดิมไม่สามารถใช้ในการสร้างแบบสอบที่เหมาะกับความสามารถของผู้เรียน (Tailor test) และแบบสอบปรับเหมาะ (Computer adaptive test) เนื่องจากไม่มีสารสนเทศที่ชี้ว่าผู้สอบมีโอกาสตอบถูกเท่าใด

จากจุดอ่อนดังกล่าว ทำให้นักจิตวิทยาได้คิดค้นพัฒนาทฤษฎีการตอบสนองข้อสอบ (Item response theory: IRT) ที่ให้ค่าการวัดที่มีความเที่ยงและไม่แปรเปลี่ยนไปตามคุณลักษณะที่เปลี่ยนไปของแบบสอบและกลุ่มผู้สอบ เพราะเมื่อ IRT Model สอดคล้อง (Fit) กับข้อมูลจะทำให้เกิดความไม่แปรเปลี่ยนของพารามิเตอร์ของข้อสอบ (Item parameter) และพารามิเตอร์ความสามารถของผู้สอบ (Ability parameter)

ความคิดพื้นฐานของทฤษฎีการตอบข้อสอบ

ทฤษฎีการตอบสนองข้อสอบตั้งอยู่บนสมมุติฐาน 2 ประการ คือ

1. พฤติกรรมของผู้สอบในการตอบข้อสอบ สามารถพยากรณ์หรืออธิบายคุณลักษณะภายใน (Traits หรือ Latent traits หรือ Abilities) ได้

2. ความสัมพันธ์ระหว่างพฤติกรรมผู้สอบในการตอบข้อสอบและคุณลักษณะภายในอธิบายได้โดยโค้งลักษณะข้อสอบ (Item characteristic curve: ICC) ซึ่งโค้งนี้มีลักษณะเฉพาะ คือ เมื่อผู้สอบมีความสามารถสูง ความน่าจะเป็นในการตอบข้อสอบถูกก็มากขึ้นด้วย

ข้อตกลงเบื้องต้นของทฤษฎีการตอบสนองข้อสอบ

ทฤษฎีการตอบสนองข้อสอบมุ่งอธิบายความสัมพันธ์ระหว่างความสามารถที่แท้จริงของผู้สอบ (θ) กับพฤติกรรมการตอบสนองข้อสอบในแต่ละข้อว่ามีความน่าจะเป็นในการตอบข้อสอบถูกได้มากน้อยเพียงใด $P_i(\theta)$ ด้วยโค้งคุณลักษณะของข้อสอบ (Item characteristic curve: ICC) ซึ่งมีลักษณะเป็นฟังก์ชันคณิตศาสตร์ โดยมีข้อตกลงเบื้องต้นที่สำคัญ ได้แก่ (DeMars, 2010; Hambleton et al., 1991; Fukuhara & Kamata, 2011; ชนะศึก นิชานนท์, 2553)

1. ความเป็นมิติเดียว (Unidimensional)

ความเป็นมิติเดียวเป็นข้อตกลงเบื้องต้นที่ใช้กันทั่วไป สำหรับ IRT หมายถึง แบบสอบต้องวัดความสามารถด้านเดียว (θ) สำหรับผู้สอบแต่ละคน และปัจจัยอื่นที่มีผลต่อการตอบสนองข้อสอบจะถูกกำหนด (Treat) ให้มีความคาดเคลื่อนแบบสุ่ม (Random error) ซึ่งการละเลยข้อตกลงนี้จะนำไปสู่ความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ หรือ Standard Error อย่างไรก็ตาม ข้อตกลงนี้สามารถผ่อนปรนได้ โดยสามารถอธิบายแบบสอบที่มีการใช้ความสามารถพหุมิติ (Multidimensionality) ได้ เช่น Multidimentional Rasch Model เป็นต้น

การตรวจสอบความเป็นมิติเดียวของแบบสอบ ทำได้หลายวิธี ดังนี้

1. การหาค่าความสัมพันธ์ระหว่างค่าน้ำหนักองค์ประกอบรายด้าน (Factor loading) ขององค์ประกอบที่หนึ่งกับค่าสหสัมพันธ์แบบไบซีเรียล (Biserial correlation coefficient) ของข้อสอบรายข้อกับคะแนนรวม ถ้ามีค่าสัมประสิทธิ์สหสัมพันธ์มากกว่า 0.80 สรุปได้ว่าข้อสอบหรือแบบสอบนั้นมีความเป็นมิติเดียว

2. การวิเคราะห์องค์ประกอบ (Factor analysis) ของแบบสอบทั้งฉบับ พิจารณาจากค่าไอเกน (Eigen values) โดยผลการวิเคราะห์องค์ประกอบใดมีค่าไอเกนในองค์ประกอบใดองค์ประกอบหนึ่งสูงกว่าค่าอื่นอย่างชัดเจน สามารถสรุปได้ว่าแบบสอบนั้นมีความเป็นมิติเดียว

3. การใช้โปรแกรม TESTFACT วิเคราะห์องค์ประกอบเชิงยืนยัน (Confirmatory factor analysis) พัฒนาโดย Wilson & Hoskens โดยวิเคราะห์ข้อสอบและทดสอบความตรงของโครงสร้างด้วย χ^2 สำหรับ Likelihood Ratio (G^2) ในการตรวจสอบความเป็นมิติเดียว ดัชนีนี้ใช้ทดสอบด้วยการกำหนดจำนวนองค์ประกอบของชุดข้อมูลไว้ล่วงหน้าแล้วทดสอบด้วย χ^2 ที่มีการประมาณค่าด้วยวิธี G^2 เพื่อทดสอบความเหมาะสมของโมเดล หากค่า G^2 ไม่มีนัยสำคัญแล้ว แสดงว่าข้อมูลมีจำนวนองค์ประกอบเท่าที่กำหนดไว้แต่ต้น

4. การวิเคราะห์การแบ่งกลุ่มแบบลดหลั่น (Hierarchical cluster analysis) เป็นเทคนิคสำหรับทดสอบความเป็นพหุมิติของแบบสอบ โดยพิจารณาการแบ่งกลุ่มของตัวแปร โดยกระบวนการแบ่งกลุ่มนี้เป็นการแบ่งกลุ่มจำนวนข้อสอบที่มีลักษณะคล้ายคลึงกันให้อยู่ในกลุ่มเดียวกัน นอกจากนี้ยังใช้การหมุนซ้ำ (Iteration) จนผลลัพธ์ (Outcome) อยู่ในระดับที่น่าพอใจ ซึ่งการวิเคราะห์ด้วยวิธีนี้สามารถใช้โปรแกรมสำเร็จรูป CCPROX และ HCA ในการวิเคราะห์ได้

5. การใช้โปรแกรม DETECT ในการตรวจสอบมิติแฝงเชิงยืนยันแบบ Nonparametric ซึ่งใช้การประมาณค่าจำนวนมิติแฝงที่มีคุณลักษณะเด่นในชุดข้อมูลและตรวจสอบความเป็นมิติเดียวของแบบสอบ โดยระบุคุณลักษณะเด่นของมิติแฝงในแต่ละข้อ ซึ่งผู้ใช้งานโปรแกรมสามารถระบุจำนวนมิติแฝงที่ต้องการได้ แต่กระบวนการดังกล่าวยังไม่เป็นทางการนัก

6. การใช้โปรแกรม DIMTEST ตรวจสอบสมมติฐานของแบบสอบด้วย Nonparametric Statistical ลักษณะคล้ายกับการตรวจสอบด้วยโปรแกรม DETECT ต่างกันที่โปรแกรม DIMTEST ตรวจสอบความสัมพันธ์ระหว่างชุดข้อสอบย่อยภายใต้เงื่อนไขความแปรปรวนร่วมของข้อสอบ

2. ความเป็นอิสระในการตอบข้อสอบ (Local independence)

ความเป็นอิสระในการตอบข้อสอบมีความเกี่ยวข้องและเชื่อมโยงมาจากความเป็นมิติเดียว หมายถึง เมื่อควบคุมความสามารถที่มีผลต่อการตอบข้อสอบให้คงที่แล้ว ความน่าจะเป็นในการตอบข้อสอบถูกในแต่ละข้อมีความเป็นอิสระต่อกัน หรือกล่าวอีกนัยหนึ่ง เมื่อควบคุมอิทธิพลของที่มีผลต่อการตอบข้อสอบให้คงที่แล้ว ผลการตอบข้อสอบรายข้อไม่มีความสัมพันธ์กัน โดยความเป็นอิสระสามารถจำแนกได้เป็น

2.1 ความเป็นอิสระระหว่างข้อสอบ หมายถึง ข้อสอบแต่ละข้อเป็นอิสระจากกัน กล่าวคือ การตอบข้อสอบข้อหนึ่งไม่มีผลกระทบต่อการตอบข้อสอบข้ออื่น ๆ ในแบบสอบฉบับนั้น

2.2 ความเป็นอิสระระหว่างผู้สอบ หมายถึง ผู้สอบแต่ละคนตอบข้อสอบแต่ละข้ออย่างเป็นอิสระกัน หรือ ถ้า $P_i(\theta_A)$ เป็นอิสระจาก $P_i(\theta_B) \rightarrow P_i(\theta_A) \cap P_i(\theta_B) = P_i(\theta_A) \cdot P_i(\theta_B)$ เมื่อ θ เป็นความสามารถที่มีอิทธิพลต่อการตอบข้อสอบของผู้สอบ U_i เป็นผลการตอบข้อสอบข้อ i ของผู้สอบที่สุ่มเลือกได้ (i = 1, 2,…, n) และ $P(U_i|\theta)$ เป็นความน่าจะเป็นในการตอบข้อสอบของผู้มีความสามารถ θ $P(U_i = 1|\theta)$ เป็นความน่าจะเป็นในการตอบข้อสอบถูก และ $P(U_i = 0|\theta)$ เป็นความน่าจะเป็นในการตอบข้อสอบผิด หรืออธิบายคุณสมบัติของความเป็นอิสระในการตอบข้อสอบในเชิงสถิติ ได้ดังนี้

$$P(U_1, U_2, \dots, U_n|\theta) = P(U_1|\theta)\, P(U_2|\theta) \dots P(U_n|\theta)$$
$$= \prod_{i=1}^{n} P(U_i|\theta) \qquad (1)$$

เมื่อ $U_1, U_2, \dots, U_n$ หมายถึง ผลการตอบของข้อที่ i (i = 1, 2, … , n)

$P(U_1, U_2, \dots, U_n|\theta)$ หมายถึง ความน่าจะเป็นในการตอบข้อที่ i (i = 1, 2, … , n) ถูก เมื่อความสามารถ = θ ($U_i = 1$ เมื่อตอบถูก และ $U_i = 0$ เมื่อตอบไม่ถูก)

$P(U_i|\theta)$ หมายถึง ความน่าจะเป็นในการตอบข้อที่ i (i = 1, 2, … , n) ถูก เมื่อความสามารถ = θ ($U_i = 1$ เมื่อตอบถูกและ $U_i = 0$ เมื่อตอบไม่ถูก)

หมายความว่า สำหรับผู้ตอบข้อสอบที่มีความสามารถ (θ) ความน่าจะเป็นของการตอบข้อสอบทั้งฉบับถูกต้อง จะเท่ากับผลคูณของความน่าจะเป็นของการตอบข้อสอบแต่ละข้อ เช่น ถ้ารูปแบบการตอบข้อสอบ 3 ข้อของผู้สอบคนหนึ่งเป็น (1, 1, 0) นั่นคือ ถ้า $U_1 = 1, U_2 = 1, U_3 = 0$ จะได้

$$P(U_1 = 1|\theta, U_2 = 1|\theta, U_3 = 0|\theta) = P(U_1 = 1|\theta)\, P(U_2 = 1|\theta)\, P(U_3 = 0|\theta)$$

$$= P_1 P_2 Q_3 \qquad (2)$$

เมื่อ $P_i = P(U_i = 1|\theta)$ และ $Q_i = 1 - P_i$

ในการทำข้อสอบแต่ละข้อผู้สอบอาจใช้ความสามารถหลายอย่าง ถ้าสามารถกำจัด (Partialled out) ความสามารถที่ไม่ต้องการวัดออกไป หรือทำให้คงที่ (Held constant) ทำให้การตอบข้อสอบแต่ละข้อของแต่ละคนมีความเป็นอิสระที่เรียกว่า ความเป็นอิสระอย่างมีเงื่อนไข (Conditional independence) ถ้าข้อตกลงเบื้องต้นของความเป็นมิติเดียวเป็นจริง แล้วจะมีคุณสมบัติของความเป็นอิสระในการตอบข้อสอบด้วย

ความเป็นอิสระในการตอบข้อสอบสามารถเกิดขึ้นได้ แม้ว่าแบบสอบไม่ได้มีความเป็นมิติเดียว ความเป็นอิสระในการตอบข้อสอบจะเกิด เมื่อกำหนดคุณลักษณะภายในอย่างสมบูรณ์ และความสามารถทั้งหลายเหล่านั้นต่างส่งผลต่อการตอบข้อสอบ

การตรวจสอบความเป็นอิสระในการตอบข้อสอบ ทำได้หลายวิธี ดังนี้

1. การพิจารณาจาก Variance - Covariance Matrix หรือ Correlation Matrix ของคะแนนคำตอบรายข้อ สำหรับกลุ่มผู้สอบที่มีความสามารถเท่ากัน ค่านอกแนวทแยงมุมของเมทริกซ์ต้องมีค่าต่ำหรือมีค่าเข้าใกล้ศูนย์

2. การตรวจสอบค่าสูงสุดของสหสัมพันธ์ภายในระหว่างข้อสอบ (Magnitude of inter - item correlations) ณ ที่ระดับช่วงคะแนนที่แตกต่างกันของผู้สอบ

3. การสร้างรูปแบบอิทธิพลหลักและอิทธิพลการปฏิสัมพันธ์ของข้อสอบเพื่ออธิบายความไม่เป็นอิสระของข้อสอบ โดยได้เสนอโมเดล 2 รูปแบบในการตรวจสอบความไม่เป็นอิสระของข้อสอบได้แก่ โมเดลปฏิสัมพันธ์คงที่ (Constant interaction model) และโมเดลปฏิสัมพันธ์ของมิติที่ไม่เป็นอิสระ (Dimension - dependent interaction model)

4. การประยุกต์ใช้วิธีการวิเคราะห์พหุระดับ (HLM) พิจารณาจากลักษณะความเป็นเอกพันธ์ของการกระจายค่าความคลาดเคลื่อน (Homoscesdasticity) และความเป็นอิสระของตัวแปรกลุ่มตัวอย่างภายในกลุ่มเดียวกันจะมีลักษณะคล้ายกันมากกว่ากลุ่มตัวอย่างระหว่างกลุ่ม ดังนั้น องค์ประกอบของกลุ่มแต่ละกลุ่มจะเป็นอิสระต่อกัน แต่จะมีความสัมพันธ์กันภายในกลุ่ม บางกลุ่มอาจมีความเป็นเอกพันธ์มากกว่ากลุ่มอื่น ๆ ดังนั้น ความแปรปรวนขององค์ประกอบระหว่างกลุ่ม ต่างกันจากแนวคิดดังกล่าว จึงสามารถนำมาประยุกต์ใช้ในการพิจารณาแบบสอบและลักษณะที่ร่วมกันระหว่างแบบสอบย่อยได้

ความเป็นอิสระในการตอบข้อสอบเป็นข้อตกลงเบื้องต้นที่สำคัญของทฤษฎีการวัด ซึ่งหากละเลยข้อตกลงนี้จะทำให้เกิดผลต่าง ๆ ดังนี้

1. ในมุมของทฤษฎีการวัดแบบดั้งเดิม (Classical test theory: CTT) จะทำให้ประมาณค่าความคาดเคลื่อนมาตรฐานของการวัด (Standard error of measurement) ได้ต่ำกว่าเกินจริง ส่งผลให้ความเที่ยง (Reliability) สูงเกินจริง

2. ในมุมของทฤษฎีการสอบสนองข้อสอบ (IRT) การประมาณค่าสารสนเทศ (Information) จะสูงเกินจริง นั่นหมายถึง ค่าความคาดเคลื่อนมาตรฐานของการประมาณค่าความสามารถจะต่ำเกินจริง นอกจากนี้ อำนาจของอำนาจจำแนก (Item discrimination power) อาจจะลำเอียง ซึ่งทำให้การประมาณค่าความสามารถก็อาจจะลำเอียงไปด้วย เนื่องจากขึ้นอยู่กับอำนาจจำแนก (กรณีฟังก์ชัน 2 พารามิเตอร์และ 3 พารามิเตอร์)

3. โค้งคุณลักษณะของข้อสอบ (Item characteristic curve: ICC)

ทฤษฎีการตอบสนองข้อสอบให้ความสำคัญกับฟังก์ชันคุณลักษณะของข้อสอบหรือโค้งคุณลักษณะข้อสอบโดยเป็นฟังก์ชันทางคณิตศาสตร์ที่แสดงความสัมพันธ์ระหว่างโอกาสในการตอบข้อสอบถูกกับระดับความสามารถของผู้สอบมีลักษณะเป็นฟังก์ชันทางคณิตศาสตร์ เรียกว่าฟังก์ชันโลจิส (Logistic function) หรือใกล้เคียงกับฟังก์ชันปกติสะสม (Normal ogive function) ดังนั้น จะเห็นได้ว่าโอกาสที่ผู้สอบตอบข้อสอบถูกจึงขึ้นอยู่กับโค้งลักษณะข้อสอบ ซึ่งเป็นอิสระจากการกระจายของความสามารถของผู้สอบหรือโอกาสที่ผู้ตอบตอบข้อสอบถูกไม่ขึ้นอยู่กับจำนวนของผู้สอบที่มีความสามารถเหมือนกัน ลักษณะของโค้งคุณลักษณะข้อสอบในแต่ละข้อมีคุณสมบัติไม่แปรเปลี่ยนไปตามกลุ่มของผู้สอบ จึงทำให้โอกาสในการตอบข้อสอบถูกในแต่ละข้อไม่แปรเปลี่ยน ซึ่งไม่เหมือนกับการวิเคราะห์ด้วยทฤษฎีการทดสอบแบบดั้งเดิม (CTT)

4. การทดสอบที่ไม่แข่งขันด้านเวลา (Nonspeed test administration)

ทฤษฎีการตอบสนองข้อสอบมุ่งอธิบายความสัมพันธ์ระหว่างความสามารถที่แท้จริงของผู้สอบกับพฤติกรรมการตอบสนองข้อสอบในแต่ละข้อ ดังนั้น ความสามารถที่แท้จริงของผู้สอบจึงเป็นสิ่งสำคัญสำหรับการวิเคราะห์ตามแนวทฤษฎีนี้ โดยแบบสอบประเภทใช้ความเร็วในการตอบ (Speed test) ผู้สอบต้องใช้ความสามารถอย่างน้อยสองมิติ คือ ความเร็วในการตอบ (Response speed) และความสามารถที่แบบสอบต้องการวัด ดังนั้น เพื่อให้สอดคล้องกับข้อตกลงเบื้องต้นเกี่ยวกับความเป็นมิติเดียว แบบสอบที่นำมาใช้จึงเป็นแบบสอบที่ไม่ใช้ความเร็วในการตอบ (Nonspeed test) ทำให้ทุกคนมีเวลาทำข้อสอบทุกข้อ และสามารถใช้คะแนนรวมจากการทำแบบทดสอบร่วมกับลักษณะข้อสอบเป็นตัวประมาณค่าความสามารถที่แท้จริงของผู้ตอบข้อสอบโดยไม่ต้องพิจารณาเรื่องความเร็ว

การตรวจสอบข้อตกลงเบื้องต้นข้อนี้มีวิธีการตรวจสอบได้หลายวิธี ดังนี้

1. การพิจารณาจากสัดส่วนหรือร้อยละของจำนวนผู้สอบที่สามารถทำข้อสอบได้ครบทุกข้อ โดยผู้สอบส่วนใหญ่ต้องสามารถตอบข้อสอบได้ครบหรือเกือบครบทุกข้อ

2. การเปรียบเทียบความแปรปรวนของจำนวนข้อที่เว้นกับความแปรปรวนของจำนวนข้อที่ตอบผิด ถ้าอัตราส่วนของความแปรปรวนเข้าใกล้ศูนย์ แสดงว่าการดำเนินการจัดการสอบไม่เน้นการแข่งขันด้านเวลา

โมเดลที่ใช้วิเคราะห์ตามทฤษฎีการตอบสนองข้อสอบ

ทฤษฎีการตอบสนองข้อสอบเป็นทฤษฎีการวัดที่เกิดจากการพยายามค้นหาวิธีแก้ไขข้อบกพร่องของทฤษฎีการทดสอบแบบดั้งเดิม (Classical test theory: CTT) ซึ่งทฤษฎีการตอบสนองข้อสอบนี้มีความเชื่อว่าพฤติกรรมการตอบสนองต่อข้อสอบของผู้สอบเป็นสิ่งที่สังเกตได้โดยตรงว่าถูกหรือผิด แต่คุณลักษณะภายในหรือความสามารถที่อยู่ภายในตัวบุคคล ไม่สามารถสังเกตได้โดยตรง ดังนั้น โอกาสที่จะตอบข้อสอบถูกหรือผิดของผู้สอบ จึงขึ้นอยู่กับระดับความสามารถและคุณลักษณะของข้อสอบ ได้แก่ ค่าพารามิเตอร์ประจำข้อสอบแต่ละข้อประกอบด้วย ความยาก (b) อำนาจจำแนก (a) และโอกาสการเดา (c) โดยความสัมพันธ์ของค่าพารามิเตอร์ดังกล่าว อธิบายได้ในรูปของฟังก์ชันคณิตศาสตร์ เรียกว่า ฟังก์ชันโลจิส (Logistic function) ซึ่งมีความใกล้เคียงกับฟังก์ชันปกติสะสม (Normal ogive function) โดยฟังก์ชันโลจิสเป็นฟังก์ชันที่มีความทนต่อความคลาดเคลื่อนที่เกิดกับผู้สอบที่มีความสามารถสูง จึงทำให้ถูกใช้อย่างแพร่หลายและนิยมนำไปใช้จริงมากกว่าฟังก์ชันปกติสะสม โดยโมเดลการตอบสนองข้อสอบแบบสองค่า (Dichotomous item response theory) เป็นดังนี้ (Hambleton et al., 1991; ชนะศึก นิชานนท์, 2553; อิทธิฤทธิ์ พงษ์ปิยะรัตน์, 2551)

1. โมเดลการตอบสนองข้อสอบแบบ 2 พารามิเตอร์

Lord (1952 cited in Hambleton et al., 1991) พัฒนาโมเดลการตอบสนองข้อสอบแบบ 2 พารามิเตอร์ (Two - parameter logistic model: 2PL Model) โดยอยู่บนพื้นฐานของการกระจายแบบปกติสะสม (Cumulative normal distribution หรือ Normal ogive) ซึ่งมีสมการ ดังนี้ (Allen & Yen, 1979)

$$P_i(\theta) = \int_{-\infty}^{a_i(\theta - b_i)} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \qquad (3)$$

เมื่อ b_i หมายถึง ความยากของแบบสอบ

a_i หมายถึง อำนาจจำแนกของแบบสอบ

z หมายถึง ตัวแปรโค้งปกติ (Standard normal variable)

ต่อมา Birnbaum (1968 cited in Hambleton et al., 1991) ได้พัฒนาฟังก์ชันโลจิสแบบ 2 พารามิเตอร์เข้ามาแทนฟังก์ชันปกติสะสมแบบ 2 พารามิเตอร์ ฟังก์ชันโลจิสมีข้อได้เปรียบมากกว่าตรงที่สามารถคำนวณได้สะดวกกว่า โดยในเส้นโค้งโลจิสมีค่าพารามิเตอร์ 2 ตัว คือ ความยากของข้อสอบ (b) และอำนาจจำแนกของข้อสอบ (a)

โมเดลการตอบสนองข้อสอบแบบ 2 พารามิเตอร์มีโค้งลักษณะข้อสอบที่เขียนด้วย

$$P_i(\theta) = \frac{1}{1+e^{-Da_i(\theta-b_i)}} \quad (4)$$

เมื่อ θ หมายถึง ระดับความสามารถของผู้สอบที่ประมาณได้จากแบบจำลองตามทฤษฎีการตอบข้อสอบ

$P_i(\theta)$ หมายถึง ความน่าจะเป็นที่ผู้สอบที่มีระดับความสามารถที่ θ จะสามารถตอบข้อสอบข้อที่ i ได้ถูกต้อง

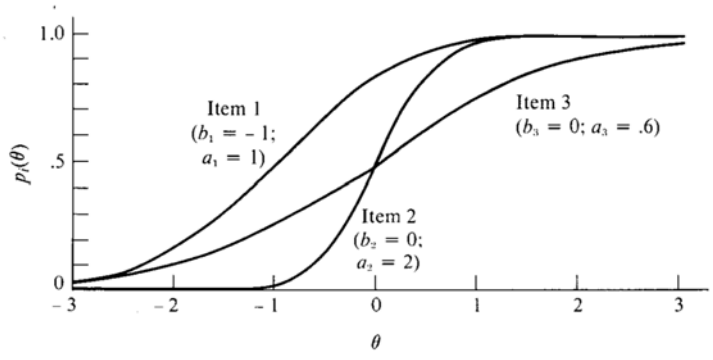
b_i หมายถึง ค่าพารามิเตอร์ความยากของข้อสอบข้อที่ i ซึ่งเป็นค่าที่แสดงตำแหน่งของ ICC ณ จุด θ ที่มีโอกาสตอบข้อสอบถูก 0.5

a_i หมายถึง ค่าพารามิเตอร์อำนาจจำแนกของข้อสอบข้อที่ i ซึ่งเป็นค่าที่แสดงความชันของ ICC ณ ตำแหน่ง b_i

e หมายถึง ค่าคงที่ของ natural logarithm มีค่าประมาณ 2.71828

D หมายถึง ค่าการปรับสเกลให้ Logistic function มีค่าใกล้เคียงกับ Normal ogive function โดยที่ D มีค่าเท่ากับ 1.7

โดยมี ICC ของโมเดลการตอบสนองข้อสอบแบบ 2 พารามิเตอร์ ดังภาพที่ 2 - 2



ภาพที่ 2 - 2 โค้งลักษณะข้อสอบแบบ 2 พารามิเตอร์ 3 ข้อ (Allen & Yen, 1979, p. 255)

2. โมเดลการตอบสนองข้อสอบแบบ 3 พารามิเตอร์

โมเดลการตอบสนองข้อสอบแบบ 3 พารามิเตอร์ (Three - parameter logistic model: 3PL Model) นี้ ได้รับการพัฒนามาจากโมเดลแบบ 2 พารามิเตอร์เพื่อให้เหมาะสมกับ การนำในไปใช้ในการทดสอบ ซึ่งมีอิทธิพลจากโอกาสในการเดา (c) หรือจุดต่ำที่สุดของโค้งการตอบข้อสอบ (Lower asymptote) หมายถึง ความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องของผู้สอบที่มีระดับความสามารถต่ำสุด เข้ามาแฝงอยู่ด้วย โดยมีโค้งลักษณะข้อสอบที่เขียนด้วยฟังก์ชันปกติสะสมแบบ 3 พารามิเตอร์ (Three - parameter normal ogive) ดังสมการ

$$P_i(\theta) = c_i + (1 - c_i)\left[\int_{-\infty}^{a_i(\theta - b_i)} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz\right] \quad (5)$$

เมื่อ b_i หมายถึง ความยากของแบบสอบ

a_i หมายถึง อำนาจจำแนกของแบบสอบ

c_i หมายถึง โอกาสการเดาของแบบสอบ

z หมายถึง ตัวแปรโค้งปกติ (Standard Normal Variable)

หรือโค้งลักษณะข้อสอบที่เขียนด้วยฟังก์ชันโลจิส ดังสมการ

$$P_i(\theta) = c_i + \frac{(1 - c_i)}{1 + e^{-Da_i(\theta - b_i)}} \quad (6)$$

เมื่อ θ หมายถึง ระดับความสามารถของผู้สอบที่ประมาณได้จากแบบจำลองตามทฤษฎีการตอบข้อสอบ

$P_i(\theta)$ หมายถึง ความน่าจะเป็นที่ผู้สอบที่มีระดับความสามารถที่ θ จะสามารถตอบข้อสอบข้อที่ i ได้ถูกต้อง

b_i หมายถึง ค่าพารามิเตอร์ความยากของข้อสอบข้อที่ i ซึ่งเป็นค่าที่แสดงตำแหน่งของ ICC ณ จุด θ ที่มีโอกาสตอบข้อสอบถูก $\frac{1+c_i}{2}$

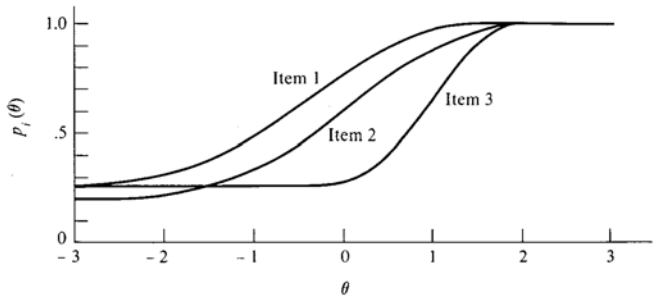
a_i หมายถึง ค่าพารามิเตอร์อำนาจจำแนกของข้อสอบข้อที่ i เป็นค่าความชันของ ICC

c_i หมายถึง ค่าพารามิเตอร์โอกาสในการเดาข้อสอบข้อที่ i ได้ถูก

e หมายถึง ค่าคงที่ของลอกธรรมชาติ (Natural logarithm) มีค่าประมาณ 2.71828

D หมายถึง ค่าการปรับสเกลให้ Logistic function มีค่าใกล้เคียงกับ Normal ogive function โดยที่ D มีค่าเท่ากับ 1.7

โดยมี ICC ของโมเดลการตอบสนองข้อสอบแบบ 3 พารามิเตอร์ ดังภาพที่ 2 - 3



ภาพที่ 2 - 3 โค้งลักษณะข้อสอบแบบ 3 พารามิเตอร์ 3 ข้อ (Allen & Yen, 1979, p. 259)

3. โมเดลการตอบสนองข้อสอบแบบ 1 พารามิเตอร์

โมเดลการตอบสนองข้อสอบแบบ 1 พารามิเตอร์ (One - parameter logistic model: 1PL Model) เป็นโมเดลที่มีลักษณะซับซ้อนน้อยที่สุดของโมเดลในกลุ่มของ IRT หรือ โมเดลราสช์ (Simple rasch model) พัฒนาขึ้นโดย George Rasch นักคณิตศาสตร์ชาวเดนมาร์ก ในปี ค.ศ. 1960 มีความเชื่อว่า องค์ประกอบที่สำคัญที่ทำให้กระบวนการวัดผลเกิดประสิทธิภาพที่ดีที่สุด นั่นคือ การใช้เครื่องมือในการวัดผลต้องมีความเป็นอิสระในตัวเอง หรืออีกนัยหนึ่งคือ คุณภาพของการวัดผลต้องไม่ได้ขึ้นอยู่กับสิ่งที่ต้องการวัด ดังนั้น จึงได้เกิดแนวคิดขึ้นว่าควรมีการพัฒนาการวิเคราะห์ข้อสอบที่ให้ผลการวิเคราะห์เป็นอิสระ โดย Rasch ได้พยายามคิดหาค่าความยากของข้อสอบโดยไม่ต้องไปสัมพันธ์กับผู้สอบและหาค่าความสามารถของผู้สอบโดยไม่ต้องไปสัมพันธ์กับระดับความยากง่ายของข้อสอบ ซึ่งก็คือ เครื่องมือวัดและสิ่งที่ถูกวัดเป็นอิสระต่อกันและกัน

ใน Rasch Model ตัวแปรตามจะเป็นการตอบแบบให้คะแนนสองค่า (Dichotomous response) ของผู้สอบแต่ละคน ส่วนตัวแปรอิสระเป็นคะแนนความสามารถของผู้สอบ (θ) และความยากง่ายของข้อสอบ (b) ทั้งนี้ ตัวแปรอิสระจะเป็นการรวมกันเชิงบวก (Combine additively) ความสัมพันธ์ดังกล่าวจะเป็นกรณีที่ตัวแปรตามมีค่าเป็น log odds หรือความน่าจะเป็น

สำหรับสมการ log odd ของ Rasch Model ตัวแปรจะเป็นลอกธรรมชาติของโอกาสในการตอบข้อสอบได้ถูกต้อง odd จะถูกกำหนดเป็นอัตราส่วนของจำนวนข้อสอบที่ตอบถูกต่อจำนวน

ข้อสอบที่ตอบผิด เช่น odd ของผู้สอบคนหนึ่งเป็น 4/1 นั่นคือข้อสอบ 5 ข้อ ผู้สอบตอบถูก 4 ข้อ และตอบผิด 1 ข้อ นอกจากนี้ odd ยังสามารถเป็นความน่าจะเป็นของการตอบข้อสอบถูกหารด้วยความน่าจะเป็นที่จะตอบข้อสอบผิดได้อีกด้วย สำหรับ Rasch Model ค่า $\log_n$ ของอัตราส่วน odd จะถูกทำให้เป็นโมเดลด้วยผลต่างระหว่าง θ_s และ b_i นั่นคือ อัตราส่วนของ ความน่าจะเป็นที่ผู้สอบจะตอบข้อสอบข้อที่ i ได้ถูกต้องต่ออัตราส่วนที่จะตอบข้อสอบข้อนี้ผิด เขียนเป็นสมการได้ดังนี้

$$\ln\left[\frac{p_{is}}{1-p_{is}}\right] = \theta_s - b_i \qquad (7)$$

หาก θ_s และ b_i เท่ากัน แสดงว่าค่าของ log odd ของการตอบถูกจะเท่ากับศูนย์ เมื่อถอดค่า log ออกมา จะเท่ากับ 1 สามารถแปลความหมายได้ว่าบุคคลมีโอกาสที่จะตอบข้อสอบได้ถูกเท่ากับการตอบข้อสอบผิด

ใน Rasch Model ที่นำเสนอสามารถประมาณค่าความสามารถของผู้สอบได้ เมื่อทราบค่าความยากของข้อสอบ โดยมีการประมาณค่าที่แยกออกจากกัน ซึ่งเป็นคุณสมบัติที่สำคัญของทฤษฎีการวัด

อีกหนึ่งลักษณะของ Rasch Model คือ ตัวแปรตามเป็นความน่าจะเป็นของผู้สอบที่จะตอบข้อสอบถูก โดยค่า θ_s และ b_i มีความสัมพันธ์กันในเชิงบวก แต่โมเดลนี้ความเกี่ยวข้องของตัวแปรตามและตัวแปรอิสระ ซึ่งมีลักษณะไม่เป็นเส้นตรง (Non linear function) เป็นฟังก์ชั่นโลจิส เขียนเป็นสมการได้ ดังนี้

$$P(x_{is} = 1|\theta, b_i) = \frac{\exp(\theta - b_i)}{1+\exp(\theta - b_i)} \qquad (8)$$

หรือ $$P_i(\theta) = \frac{1}{1+e^{-(\theta - b_i)}} \qquad (9)$$

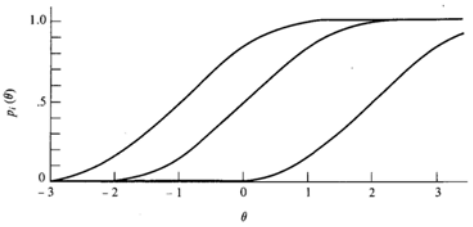
เมื่อ θ หมายถึง ระดับความสามารถของผู้สอบที่ประมาณได้จากแบบจำลองตามทฤษฎีการตอบข้อสอบ ปรับให้เป็นคะแนนมาตรฐานที่มีค่าเฉลี่ยเป็น 0 และส่วนเบี่ยงเบนมาตรฐานเป็น 1 θ มีพิสัยระหว่าง $\pm\alpha$ แต่ในทางปฏิบัติส่วนใหญ่ θ มีค่าระหว่าง ±3

$P_i(\theta)$ หมายถึง ความน่าจะเป็นที่ผู้สอบที่มีระดับความสามารถที่ θ จะสามารถตอบข้อสอบข้อที่ i ได้ถูกต้อง

b_i หมายถึง ค่าพารามิเตอร์ความยากของข้อสอบข้อที่ i ซึ่งเป็นค่าที่แสดงตำแหน่งของ ICC ณ จุด θ ที่มีโอกาสตอบข้อสอบถูก 0.5

e หมายถึง ค่าคงที่ของลอกธรรมชาติมีค่าประมาณ 2.71828

โดยมี ICC ของโมเดลการตอบสนองข้อสอบแบบ 1 พารามิเตอร์ ดังภาพที่ 2 - 4



ภาพที่ 2 - 4 โค้งลักษณะข้อสอบแบบ 1 พารามิเตอร์ 3 ข้อ (Allen & Yen, 1979, p. 261)

โดยสรุปโมเดลการตอบสนองข้อสอบทั้ง 3 แบบ มีความแตกต่างกันดังตารางที่ 2 - 3

ตารางที่ 2 - 3 เปรียบเทียบโมเดลการตอบสนองข้อสอบ 3 แบบ

โมเดล	ฟังก์ชันปกติสะสม	ฟังก์ชันโลจิส	พารามิเตอร์ข้อสอบ
1 พารามิเตอร์	$P_i(\theta) = \int_{-\infty}^{(\theta - b_i)} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$	$P_i(\theta) = \frac{1}{1 + e^{-(\theta - b_i)}}$	b_i มีค่าแปรเปลี่ยนได้ a_i มีค่าคงที่ $c_i = 0$
2 พารามิเตอร์	$P_i(\theta) = \int_{-\infty}^{a_i(\theta - b_i)} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$	$P_i(\theta) = \frac{1}{1 + e^{-Da_i(\theta - b_i)}}$	b_i มีค่าแปรเปลี่ยนได้ a_i มีค่าแปรเปลี่ยนได้ $c_i = 0$
3 พารามิเตอร์	$P_i(\theta) = c_i + (1 - c_i) \left[\int_{-\infty}^{a_i(\theta - b_i)} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \right]$	$P_i(\theta) = c_i + \frac{(1 - c_i)}{1 + e^{-Da_i(\theta - b_i)}}$	b_i มีค่าแปรเปลี่ยนได้ a_i มีค่าแปรเปลี่ยนได้ c_i มีค่าแปรเปลี่ยนได้

คุณลักษณะของทฤษฎีการตอบสนองข้อสอบ

ตามหลักการของทฤษฎีการตอบสนองข้อสอบมีแนวคิดว่า ความสามารถของผู้สอบที่ตอบสนองต่อข้อสอบสามารถอธิบายหรือทำนายได้ด้วยคุณลักษณะของผู้สอบ (Examinee characteristics) ซึ่งอาจเป็นคุณลักษณะแฝงภายในบุคคล (Traits) หรือความสามารถ (Abilities) ซึ่งการประมาณค่าคะแนนสำหรับผู้สอบในความสามารถแฝงอยู่ภายในจะเรียกว่าคะแนนความสามารถ (Ability score) และใช้คะแนนเหล่านี้มาอธิบายความสามารถในการทำข้อสอบ ดังนั้น ทฤษฎีการตอบสนองข้อสอบจึงเป็นทฤษฎีที่มุ่งอธิบายพฤติกรรมการตอบข้อสอบของผู้สอบที่ถูกกำหนดโดยคุณลักษณะภายในหรือความสามารถที่แฝงอยู่ภายในตัวบุคคล และศึกษาความสัมพันธ์ระหว่างการตอบข้อสอบของผู้สอบกับระดับความสามารถที่มีอยู่ด้วยโมเดลที่เป็นฟังก์ชันทางคณิตศาสตร์ เรียกว่า โค้งลักษณะข้อสอบ ซึ่งมีคุณลักษณะสรุปได้ดังนี้

1. ความสามารถของผู้สอบในการทำแบบสอบสามารถอธิบาย/ ทำนายในรูปแบบของคุณลักษณะที่แฝงอยู่ภายในตัวผู้สอบ

2. โมเดลการตอบสนองข้อสอบอธิบายความสัมพันธ์ระหว่างความสามารถในการตอบข้อสอบที่สามารถสังเกตได้โดยตรงว่าผิดหรือถูกกับความสามารถที่แฝงอยู่ภายในตัวผู้สอบ

3. ความสำเร็จของโมเดลการตอบสนองข้อสอบจะให้ค่าเฉลี่ยของการประมาณค่าความสามารถของผู้สอบ

4. การประมาณค่าความสามารถที่แท้จริงของผู้สอบจะประมาณค่าจากความสามารถของผู้สอบที่ตอบสนองข้อสอบ

คุณสมบัติของความไม่แปรเปลี่ยนของค่าพารามิเตอร์

เมื่อโมเดลการตอบสนองข้อสอบมีความสอดคล้อง (Fit) กับข้อมูลเชิงประจักษ์จะทำให้เกิดคุณสมบัติความไม่แปรเปลี่ยน (Invariance) ของค่าพารามิเตอร์ข้อสอบ และพารามิเตอร์ความสามารถของผู้สอบในสองลักษณะดังนี้ (ศิริชัย กาญจนวาสี, 2545)

1. ความไม่แปรเปลี่ยนของพารามิเตอร์ข้อสอบ (Item invariance)

เป็นคุณสมบัติที่ว่า ค่าพารามิเตอร์ของข้อสอบจะมีค่าคงที่ไม่เปลี่ยนแปลง แม้จะเปลี่ยนกลุ่มผู้สอบนั่นคือ พารามิเตอร์ความยาก อำนาจจำแนกและโอกาสการเดาในโค้งลักษณะข้อสอบเดียวกันจะคงที่สำหรับทุกกลุ่มความสามารถผู้สอบ แสดงว่า โค้งลักษณะข้อสอบมีความคงที่ข้ามกลุ่มผู้สอบ

2. ความไม่แปรเปลี่ยนของพารามิเตอร์ความสามารถของผู้สอบ (Ability invariance)

เป็นคุณสมบัติที่ว่าค่าพารามิเตอร์ความสามารถของผู้สอบจะมีค่าคงที่ไม่เปลี่ยนแปลง ถึงแม้จะมีการเปลี่ยนแปลงชุดของแบบสอบ นั่นคือ หากนำแบบสอบที่มุ่งวัดคุณลักษณะเดียวกัน

จำนวน 2 ชุดค่าความสามารถของผู้สอบที่ประมาณค่าได้จากแบบทดสอบ ทั้งสองชุดจะมีค่าที่แตกต่างกันไม่เกินค่าความคลาดเคลื่อนมาตรฐานของการประมาณค่า (SEE) แสดงว่าการประมาณค่าความสามารถมีความคงที่ข้ามชุดของแบบสอบ

ฟังก์ชันสารสนเทศของข้อสอบ (Item information)

การวิเคราะห์ตามทฤษฎีตอบสนองข้อสอบ จะใช้แบบแผนการตอบสนองแบบสอบเป็นรายข้อในการประมาณค่าความสามารถของผู้สอบ ดังนั้น การประเมินคุณภาพของแบบสอบจึงสามารถพิจารณาจากความถูกต้องแม่นยำในการประมาณค่าความสามารถของผู้สอบ ซึ่งมีดัชนีตัวหนึ่งสามารถใช้ชี้ถึงความถูกต้องแม่นยำดังกล่าว เรียกว่า สารสนเทศของแบบสอบเกิดจากผลรวมของค่าฟังก์ชันสารสนเทศของข้อสอบแต่ละข้อรวมเข้าด้วยกัน โดยค่าสารสนเทศของข้อสอบเป็นดัชนีผสมที่สร้างจากดัชนีคุณลักษณะของข้อสอบหลายลักษณะ ได้แก่ ค่าความยาก ค่าอำนาจจำแนก และค่าความแปรปรวนของคะแนนรายข้อ เพื่อบ่งชี้คุณภาพของข้อสอบ (Birnbaum, 1968 อ้างถึงใน ศิริชัย กาญจนวาสี, 2550) เนื่องจากว่าคุณสมบัติความไม่แปรเปลี่ยนไปตามกลุ่มตัวอย่างของค่าพารามิเตอร์จากการวิเคราะห์ด้วยทฤษฎี IRT ทำให้ค่าสารสนเทศเหมาะสมที่จะใช้เป็นตัวบ่งชี้คุณภาพของแบบสอบแทนค่าความตรงและความคลาดเคลื่อนมาตรฐานตามทฤษฎีการวัดแบบดั้งเดิม ซึ่งค่าสารสนเทศมีสูตร ดังนี้

$$I_i(\theta) = \frac{[P_i'(\theta)^2]}{P_i(\theta)Q_i(\theta)}\ ,\ i = 1, 2, \ldots, k \qquad (10)$$

เมื่อ $I_i(\theta)$ หมายถึง ค่าฟังก์ชันสารสนเทศหรือสารสนเทศที่ได้รับจากข้อสอบข้อ i สำหรับผู้ตอบที่มีความสามารถ θ

$P_i'(\theta) = P_i$ = ค่าความชันของฟังก์ชันการตอบสนองข้อสอบข้อ i ณ ตำแหน่งความสามารถ θ

$P_i(\theta) = P_i$ = ความน่าจะเป็นที่ผู้ตอบมีความสามารถ θ จะตอบข้อสอบข้อ i ได้ถูกต้อง

$Q_i(\theta) = Q_i = 1 - P_i(\theta)$

เมื่อพิจารณาสูตรการคำนวณค่าสารสนเทศของข้อสอบ พบว่า ค่าของฟังก์ชันขึ้นอยู่กับความชันของโค้งลักษณะข้อสอบ ถ้าโค้งลักษณะข้อสอบมีค่าความชัน ($P_i'(\theta)$) มากขึ้น ขณะที่ความแปรปรวนของการตอบข้อสอบ ($P_i(\theta)Q_i(\theta)$) มีค่าน้อยลง จะทำให้โค้งสารสนเทศของข้อสอบที่ระดับความสามารถนั้น ๆ มีค่ามากขึ้น ความสูงของโค้งสารสนเทศข้อสอบที่สูงที่สุด

ตรงกับความสามารถในระดับใด แสดงว่า ข้อสอบข้อนั้นจะให้ค่าสารสนเทศสูงสุด ณ ระดับความสามารถนั้น จึงทำให้สามารถเลือกใช้ข้อสอบที่เหมาะสมกับระดับความสามารถของผู้สอบได้อย่างถูกต้อง

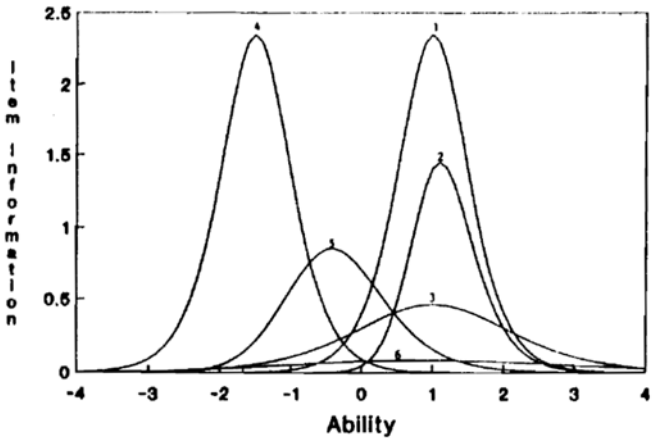
ศิริชัย กาญจนวาสี (2550) อธิบายว่า ค่าฟังก์ชันสารสนเทศของข้อสอบจะมีค่าสูง เมื่อ

1) ผู้สอบมีค่าความสามารถ (θ) ใกล้เคียงกับค่าพารามิเตอร์ความยากของข้อสอบ (b)
2) ค่าพารามิเตอร์อำนาจจำแนก (a) มีค่าสูงขึ้น
3) ค่าพารามิเตอร์โอกาสในการเดาข้อสอบถูก (c) มีค่าเข้าใกล้ศูนย์
4) $I_i(\theta)$ จะมีค่าสูงสุด ณ ตำแหน่ง θ_{max} ถ้า $c_i = 0$, $\theta_{max} = b$ แต่ถ้า $c_i > 0$, $\theta_{max} > b$

ตารางที่ 2 - 4 สูตรการคำนวณค่าสารสนเทศของข้อสอบ $I_i(\theta)$ ค่าสารสนเทศสูงสุดของข้อสอบ $I_i(\theta)_{max}$ และตำแหน่งค่าความสามารถที่ให้สารสนเทศสูงสุด θ_{max}

ค่าประมาณ	1 พารามิเตอร์	2 พารามิเตอร์	3 พารามิเตอร์
$I_i(\theta)$	$D^2 Q_i P_i$	$D^2 a_i^2 Q_i P_i$	$D^2 a_i^2 Q_i (P_i + c_i)^2 / (1 - c_i)^2$
P_i'	$D Q_i P_i$	$D a_i^2 Q_i P_i$	$D a_i^2 Q_i (P_i + c_i) / (1 - c_i)$
$I_i(\theta)_{max}$	$\frac{1}{4} D^2$	$\frac{1}{4} D^2 a_i^2$	$D^2 a_i^2 \frac{\left[1 - 20c_i - 8c_i^2 + (1 + 8c_i)^{\frac{3}{2}}\right]}{8 - (1 - c_i^2)}$
θ_{max}	b_i	b_i	$b_i + \frac{1}{D a_i}\left[Ln1 + \frac{(1 + c_i)^{\frac{1}{2}}}{2}\right]$

จากภาพที่ 2 - 5 เป็นตัวอย่างโค้งสารสนเทศของข้อสอบ จำนวน 6 ข้อ มีความสามารถในการนำไปใช้ในการวัดผู้สอบที่มีความสามารถในช่วงความสามารถที่แตกต่างกัน ตัวอย่างเช่น ข้อสอบข้อที่ 1 มีค่าสารสนเทศของข้อสอบสูงในช่วงความสามารถผู้สอบ +1 ส่วนข้อสอบข้อที่ 4 มีค่าสารสนเทศของข้อสอบสูงในช่วงความสามารถผู้สอบ -1.8 เป็นต้น



ภาพที่ 2 - 5 โค้งสารสนเทศของข้อสอบ 6 ข้อ

ฟังก์ชันสารสนเทศของแบบสอบ (Test information)

ค่าสารสนเทศของแบบสอบ เป็นดัชนีที่แสดงถึงความถูกต้องแม่นยำในการประมาณค่าความสามารถที่แท้จริงของผู้สอบ (θ) ของแบบสอบทั้งฉบับ นั่นคือ หากค่าสารสนเทศของแบบสอบมีค่าสูงอยู่ในช่วง θ ใดก็จะมีความถูกต้องแม่นยำในการประมาณค่าความสามารถของผู้ตอบในช่วง θ นั้นได้สูง โดยฟังก์ชันนี้เป็นผลรวมเชิงพีชคณิตของค่าสารเทศของข้อสอบแต่ละข้อรวมเข้าด้วยกันทั้งฉบับ ณ ตำแหน่ง θ เดียวกัน คำนวณได้จากสูตร

$$I(\theta) = \sum_{i=1}^{k} I_i(\theta) = \frac{\sum[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)} \qquad (11)$$

เมื่อ $I_i(\theta)$ หมายถึง ค่าสารสนเทศหรือค่าสารสนเทศที่ได้รับจากข้อสอบข้อ i สำหรับผู้ตอบที่มีความสามารถ θ

$I(\theta)$ หมายถึง ค่าสารสนเทศที่ได้รับจากแบบสอบสำหรับผู้ที่มีความสามารถเท่ากับ θ

เมื่อพิจารณาสูตรการคำนวณค่าสารสนเทศของแบบสอบ พบว่า ค่าสารสนเทศของแบบสอบ ขึ้นอยู่กับค่าสารสนเทศของข้อสอบแต่ละข้อที่เป็นอิสระต่อกัน ดังนั้น ถ้าต้องการให้แบบสอบมีค่าสารสนเทศสูงจึงต้องให้ออกแบบข้อสอบในแต่ละข้อมีค่าสารสนเทศที่สูง ไม่เหมือนกับทฤษฎีการทดสอบแบบดั้งเดิมที่ค่าความตรงของแบบสอบทั้งฉบับขึ้นอยู่กับค่าความยากและค่าอำนาจจำแนกของข้อสอบแต่ละข้อ ที่ไม่มีความเป็นอิสระต่อกัน ดังนั้น

ค่าความตรงของแบบสอบทั้งฉบับที่ได้จากการวิเคราะห์นี้จึงขึ้นอยู่กับลักษณะสถานการณ์เฉพาะที่นำแบบสอบมาใช้เท่านั้น

ความคลาดเคลื่อนมาตรฐานในการประมาณค่า (Standard error of estimation)

ค่าความคลาดเคลื่อนมาตรฐานการประมาณค่าเป็นค่าที่แสดงถึงความคลาดเคลื่อนที่เกิดขึ้นจากการวัด เป็นค่าส่วนเบี่ยงเบนมาตรฐานของการแจกแจงความน่าจะเป็นของการประมาณค่าความสามารถที่แท้จริง (θ) มีค่าแปรผกผันกับค่าสารสนเทศของข้อสอบ โดยจาก

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}} \tag{12}$$

เมื่อ $SE(\theta)$ หมายถึง ความคลาดเคลื่อนมาตรฐานของการประมาณค่า สำหรับผู้สอบที่มีความสามารถ θ

$I(\theta)$ หมายถึง ค่าสารสนเทศที่ได้รับจากแบบสอบสำหรับผู้ที่มีความสามารถเท่ากับ θ

เมื่อพิจารณาสูตรการคำนวณค่าความคลาดเคลื่อนมาตรฐานในการประมาณค่า พบว่า SE(θ) เป็นส่วนกลับของค่าสารสนเทศของแบบสอบ โดยถ้าแบบสอบใดมีค่าสารสนเทศของแบบสอบสูง แบบสอบนั้นก็จะมีค่า ความคลาดเคลื่อนมาตรฐานในการประมาณค่าต่ำหรือมีความแม่นยำในการประมาณค่าความสามารถสูง ณ ตำแหน่ง θ นั้น

ค่าความคลาดเคลื่อนมาตรฐานในการประมาณค่าเป็นค่าที่มีความหมายคล้ายกับค่าความคลาดเคลื่อนมาตรฐานในการวัด (Standard error of measurement: SEM) ในทฤษฎีการทดสอบแบบดั้งเดิม ซึ่งเป็นค่าที่แสดงถึงความคลาดเคลื่อนในการวัดหรือในการประมาณค่าพารามิเตอร์ของการวัด แต่ต่างกันที่ค่า SE(θ) มีความผันแปรไปตาม θ แต่ค่า SEM เป็นค่าคงที่ของแบบสอบสำหรับผู้สอบทุกคน

ตอนที่ 3 แนวคิดเกี่ยวกับทฤษฎีการตอบสนองข้อสอบสำหรับเทสต์เลท

การศึกษาเกี่ยวกับการจัดการข้อสอบที่มีลักษณะเป็นเทสต์เลท (Testlet) เริ่มเมื่อปี ค.ศ. 1987 โดย Wainer & Kiely เรียกเซตของข้อสอบที่รวมกันเป็นหนึ่งหน่วยการวัด โดยมีโครงสร้างและการจัดการร่วมกันว่าเทสต์เลท (Testlet) เช่น แบบสอบที่เกี่ยวกับการอ่านบทความ การอ่านกราฟ ตารางแสดงจำนวนของข้อมูลต่าง ๆ ซึ่งในการทดสอบลักษณะนี้ คำถามแต่ละข้อมักใช้เวลามาก การใช้เทสต์เลทจะทำให้ประหยัดเวลาได้ โดยวิธีการใช้งานเทสต์เลท คือ ให้ผู้สอบ

ตอบชุดคำถามจากคำถามที่เป็นสถานการณ์หรือสิ่งเร้าเดียวกัน ดังนั้น ผลการตอบสนองต่อข้อสอบที่อยู่ภายในเทสต์เลทเดียวกันจะไม่เป็นอิสระต่อกัน ซึ่งขัดแย้งข้อตกลงเบื้องต้นของ IRT

ทฤษฎีการตอบสนองข้อสอบสำหรับเทสต์เลท (Testlet response theory model)

เทสต์เลท (Testlet) เป็นกลุ่มของข้อสอบที่ใช้ข้อมูล บทความหรือเหตุการณ์ที่เป็นตัวกระตุ้นหรือสิ่งเร้าเดียวกัน เพื่อวัดบางส่วนของลักษณะ (Trait) ที่แบบสอบวัด ซึ่งประหยัดเวลาการสอบ เนื่องจากผู้สอบใช้ข้อมูลเดียวในการทำข้อสอบหลาย ๆ ข้อ และเพิ่มความน่าเชื่อถือของการตอบด้วย ตัวอย่างเช่น แบบสอบที่ทางวิชาคณิตศาสตร์ ที่ออกแบบตัวกระตุ้นเป็นกราฟหรือตาราง โดยใน 1 ข้อคำถาม กำหนดให้ตอบคำถามหลายข้อ ซึ่งแต่ละข้ออาจถามเป็นขั้น ๆ สำหรับการแก้ปัญหาหรือโจทย์ได้ อย่างไรก็ตาม ลักษณะของกลุ่มคำถามที่มีตัวกระตุ้นเดียวกันนี้ อาจทำให้การตอบในแต่ละข้ออาจมีการพึ่งพาข้อมูลกันระหว่างข้อหรือผู้สอบอาจได้รับคำตอบจากข้อสอบข้อถัดไป ดังนั้น การวิเคราะห์และแปลผลจึงเป็นความน่าจะเป็นของเซตของข้อสอบ (ที่ตัวกระตุ้นเดียวกัน) ทำให้การแปลความหมายตามทฤษฎีการตอบสนองข้อสอบไม่ถูกต้องและเป็นการขัดกับข้อตกลงเบื้องต้น นั่นคือ ความเป็นอิสระในการตอบข้อสอบ

แบบสอบที่มีลักษณะของเทสต์เลทจะทำให้มีแหล่งความแปรปรวนและปัจจัย เนื่องจากเทสต์เลทเพิ่มขึ้น ซึ่งหากไม่คำนึงถึงอิทธิพลของเทสต์เลทจะมีผลให้ประมาณค่าความเที่ยง (Reliability) ความคาดเคลื่อนมาตรฐาน (Standard error) และอำนาจจำแนก (Item discrimination) ไม่ถูกต้อง (Christine, 2012) โดย Sereci et al. (1991) อธิบายว่า การละเลยความเป็นอิสระระหว่างข้อสอบและผู้สอบ ทำให้การประมาณค่าความเที่ยงจะเกินจริงประมาณ 10-15% และเพื่อให้การใช้โมเดล IRT ประมาณค่าได้ถูกต้องจะต้องเพิ่มความยาวของแบบสอบเป็น 2 เท่าเพื่อชดเชยค่าความเที่ยงที่ประมาณเกินจริงไป

ในการกำจัดปัญหาความไม่เป็นอิสระในการตอบข้อสอบออกไปนั้น มีหลายวิธี เช่น การนำการวิเคราะห์ของคะแนนแบบหลายค่ามาประยุกต์ (Polytomous scoring) โดยการรวมข้อที่ถูกในหนึ่งเทสต์เลทเป็น 1 Polytomous Item เช่น Graded Response Model (Samejima, 1968) Partial Credit Model (Masters, 1982) Rating Scale Model (Andrich, 1978) Nominal Response Model (Bock, 1972) และการกำหนดให้เทสต์เลทถือเป็นหน่วยในการวิเคราะห์ด้วย โดยคำนวณคะแนนที่ตอบถูกในแต่ละเทสต์เลทแฝงในเทสต์เลท เช่น Lee & Frisbie (1999) Sireci et al. (1991) Thissen et al., (1989) Wainer & Thissen (1996)

การศึกษาที่ผ่านมาพบว่า การกำหนดให้เทสต์เลทถือเป็นหน่วยในการวิเคราะห์ด้วยนั้นสามารถแก้ปัญหาความไม่เป็นอิสระในการตอบข้อสอบออกไปได้อย่างมีประสิทธิภาพ เนื่องจากการกำหนดให้เทสต์เลทเป็นหน่วยในการวิเคราะห์ ทำให้ข้อสอบที่อยู่ระหว่างเทสต์เลทมีความเป็น

อิสระในการตอบ แต่ไม่เป็นอิสระในการตอบภายในเทสต์เลท (Sireci et al., 1991; Thissen et al., 1989; Zenisky, Hambleton & Sireci, 2003) ส่วนการนำการวิเคราะห์ของคะแนนแบบหลายค่ามาประยุกต์มีข้อจำกัด คือ ไม่สามารถวิเคราะห์หาสารสนเทศระดับข้อสอบรายข้อได้

อีกวิธีหนึ่งที่พบจากการศึกษาที่ผ่านมา คือ โมเดลการตอบสนองข้อสอบของเทสต์เลท (Testlet response theory model: TRT) โดยคำนึงถึงอิทธิพลของเทสต์เลท ด้วยการเพิ่มส่วนที่เป็น Random Effect ($\gamma_{id(j)}$) เข้าไปในสมการของโมเดลแบบ IRT ซึ่งเป็นการผ่อนคลายข้อตกลงที่เกี่ยวกับความเป็นอิสระระหว่างข้อสอบและผู้สอบ โมเดล IRT สำหรับเทสต์เลทมีหลายโมเดล ซึ่งมีลักษณะคล้าย โมเดล IRT นั่นคือ มีแบบ 1, 2 และ 3 พารามิเตอร์ ดังนี้

โมเดลการตอบสนองข้อสอบสำหรับเทสต์เลทแบบ 2 พารามิเตอร์ (2PL - TRT)

Bradlow, Wainer & Wang (1999) ศึกษา Parametric Bayesian Model สำหรับแบบสอบที่ประกอบด้วย ข้อสอบที่เป็นอิสระต่อกันและข้อสอบที่มีลักษณะเทสต์เลทผสมกันในแบบสอบเดียวกัน และแสดงให้เห็นถึงความถูกต้องและประสิทธิภาพ โดยใช้การจำลองข้อมูล 2 x 3 factorial โดยใช้จำนวนผู้สอบ 1,000 คน ความยาวแบบสอบ 60 ข้อ ข้อสอบแบบเทสต์เลท ร้อยละ 50 ของแบบสอบ กำหนดให้อิทธิพลของเทสต์เลทซึ่งเป็นอิทธิพลสุ่ม โดยที่ความแปรปรวน (Variance) ของอิทธิพลเทสต์เลท ในแต่ละเทสต์เลทมีค่าต่างกัน หรืออธิบายได้ว่า Random Testlet Effect เป็นปฏิสัมพันธ์ (Interaction) ของข้อสอบรายข้อกับเทสต์เลท โดยที่ค่าความสามารถของแต่ละบุคคลและ Random Testlet Effect ถูกควบคุม ส่วนการตอบสนองของแต่ละบุคคลเป็นอิสระ ด้วยการขยายแนวคิดฟังก์ชันโอไจฟ์ปกติแบบ 2 พารามิเตอร์ เป็นฟังก์ชัน 2PL Testlet Response Theory (2PL - TRT) โดยโมเดล 2PL - TRT มีความพอดี (Fit) กับแบบสอบที่มีการให้คะแนนสองค่า อธิบายลักษณะโมเดลได้ ดังนี้

$$P_{ij}(y_{ij}=1|t_{ij}) = logit^{-1}(t_{ij}) \quad \text{และ} \quad y_{\text{ij}} = \begin{cases} 1 \ \ if \ \ \text{t}_{\text{ij}} > 0 \\ 0 \ \ if \ \ \text{t}_{\text{ij}} \leq 0 \end{cases} \tag{13}$$

เมื่อ t_{ij} หมายถึง ตัวทำนายคะแนนแฝงแบบต่อเนื่องเชิงเส้นตรง

y_{ij} หมายถึง ผลคะแนนของผู้สอบ i ข้อสอบ j มีค่าเป็น 0 หรือ 1 (Observed binary outcome)

$$logit^{-1}(t_{ij}) = \exp(t_{ij})/(1+\exp(t_{ij})) \quad (14)$$

เมื่อ $t_{ij} = a_j(\theta_i - b_j - \gamma_{id(j)})$

θ_i หมายถึง ความสามารถของผู้สอบ i

a_j หมายถึง อำนาจจำแนกของข้อสอบ j

b_j หมายถึง ความยากของข้อสอบ j

$\gamma_{id(j)}$ หมายถึง อิทธิพลสุ่มของ Testlet $d(j)$ ของผู้สอบ i และ ถ้าข้อสอบ j และ j′ อยู่ใน Testlet เดียวกันแล้ว $d(j) = d(j')$

เมื่อกำหนดให้โมเดล 2PL - TRT อยู่ในกรอบของ Bayesian ซึ่งยอมให้มีการแบ่งปันสารสนเทศ (Sharing of information) ระหว่างผู้สอบ ข้อสอบและเทสต์เลท จะได้ Bayesian Hierarchical Structure มีการแจกแจงก่อน (Prior distribution) ดังนี้

$$\theta_i \sim N(0,1)$$
$$a_j \sim N(\mu_a, \sigma_a^2)$$
$$b_j \sim N(\mu_b, \sigma_b^2)$$
$$\gamma_{id(j)} \sim N(0, \sigma_\gamma^2)$$

สำหรับการแจกแจงของค่าเฉลี่ยและความแปรปรวน เป็นดังนี้ $\mu_a \sim N(0, V_a)$, $\mu_b \sim N(0, V_b)$ โดยที่ $\sigma_a^2 \sim \chi_{ga}^{-2}$, $\sigma_b^2 \sim \chi_{gb}^{-2}$ และ $\sigma_\gamma^2 \sim \chi_{g\gamma}^{-2}$ เมื่อ σ_γ^2 หมายถึง ความแปรปรวนของการแจกแจงก่อน สำหรับอิทธิพล Testlet γ และกำหนดให้ $V_a = V_b = 0$, $g_a = g_b = g_\gamma = 0$

สำหรับการวิเคราะห์ ใช้กับเทคนิค MCMC (Markov chain monte carlo) โดยใช้วิธีสุ่มตัวอย่างแบบ Gibbs sampler ในการประมาณค่า (Posterior distribution) ของความสามารถที่แท้จริงของผู้สอบและพารามิเตอร์ของข้อสอบ แล้วทำการเปรียบเทียบโมเดล 2PL - TRT และ 2PL - IRT ทั้งข้อมูลจำลองและข้อมูลจริง พบว่า เมื่อประมาณค่าพารามิเตอร์ด้วยโมเดล 2PL - IRT เกิดความไม่อิสระต่อกันของข้อสอบเนื่องจากเทสต์เลท ส่งผลให้การประมาณค่าพารามิเตอร์ของข้อสอบและพารามิเตอร์ของผู้สอบลำเอียง ขณะที่โมเดล 2PL - TRT มีความลำเอียงน้อยกว่า เช่น เมื่อประมาณค่าด้วยโมเดล 2PL - IRT ค่าอำนาจจำแนกจะต่ำกว่าความจริง ซึ่ง Bradlow et al. (1999) อธิบายปรากฏการณ์ว่า เมื่อใช้โมเดลที่ไม่พอดีกับข้อมูล จึงทำให้ความสัมพันธ์ระหว่างข้อสอบและความสามารถต่ำ ส่งผลให้ประมาณค่าอำนาจจำแนกต่ำเกินจริง

นอกจากนี้ Bradlow et al. (1999) ยังชี้ให้เห็นข้อบกพร่องบางอย่างที่อาจเกิดขึ้นของโมเดล 2PL - TRT ว่าโมเดลนี้ ยังไม่รวมพารามิเตอร์การเดาและใช้ได้กับข้อสอบที่มีการให้คะแนนแบบสองค่าเท่านั้น ซึ่งการประเมินผลในปัจจุบันนิยมใช้ข้อสอบแบบหลายตัวเลือก ทำให้ผู้สอบสามารถเดาคำตอบได้ ดังนั้น การใช้โมเดลแบบ 3 พารามิเตอร์จะมีความพอดีกับสภาพจริงมากกว่าโมเดลที่ไม่มีพารามิเตอร์การเดา นอกจากนี้ แบบสอบมักมีข้อสอบที่มีลักษณะปลายเปิดและมักมีการให้คะแนนแบบหลายค่า ดังนั้น โมเดล 2PLTRT ยังสามารถรองรับแบบสอบที่มีข้อสอบที่มีการให้คะแนนหลายค่าหรือแบบสอบที่มีการให้คะแนนแบบสองค่าและหลายค่าผสมกัน นอกจากนี้ในโมเดล 2PL - TRT ยังมีสมมติว่าความแปรปรวนคงที่ทุกเทสต์เลททั้งหมดในการทดสอบ ซึ่งสถานการณ์จริงอาจไม่ได้เป็นอย่างนั้น โดยที่เทสต์เลทที่ต่างกัน ก็น่าจะมีขนาดของความไม่เป็นอิสระกันของข้อสอบที่แตกต่างกันด้วย

โมเดลการตอบสนองข้อสอบสำหรับเทสต์เลทแบบ 3 พารามิเตอร์ (3PL-TRT)

Wainer et al. (2000) พัฒนาโมเดล 2PL TRT (Bradlow et al., 1999) โดยนำเสนอโมเดล IRT ที่ใช้กับข้อมูลที่มีลักษณะเป็นเทสต์เลทแบบ 3 พารามิเตอร์ ซึ่งประกอบด้วยพารามิเตอร์ของ Testlet Effect โดยที่ความน่าจะเป็นที่จะสำเร็จ (y = 1) ของข้อที่ j สำหรับผู้สอบ i ที่มีระดับความสามารถ θ ของ testlet d(j) และ c_j มีค่าอยู่ระหว่าง [0,1] โดยโมเดล Three - parameter logistic TRT (3PL - TRT) ของ Dichotomous IRT แสดงได้ ดังนี้

$$P_{ij}(y_{ij}=1)=c_j+(1-c_j)\frac{e^{a_j(\theta-b_j-\gamma_{id(j)})}}{1+e^{a_j(\theta-b_j-\gamma_{id(j)})}} \qquad (15)$$

เมื่อ $\gamma_{id(j)}$ หมายถึง อิทธิพลสุ่มของ Testlet d(j) ของผู้สอบ i ที่แฝงใน testlet d(j)

b_j หมายถึง พารามิเตอร์ความยากของข้อ j

a_j หมายถึง พารามิเตอร์อำนาจจำแนกของข้อ j

c_j หมายถึง พารามิเตอร์โอกาสการเดาของข้อ j

และ $\sigma^2_{\gamma d(i)}$ เป็นค่าซึ่งยอมให้ขนาดของความไม่เป็นอิสระกันของข้อสอบที่แตกต่างกันเมื่อเทสต์เลทต่างกัน ส่วนการแจกแจงก่อน (Prior distribution) ใช้กับเทคนิค MCMC โดยใช้วิธีสุ่มตัวอย่างแบบ Gibbs sampler ในการประมาณค่า (Posterior distribution) ของความสามารถที่แท้จริงของผู้สอบและพารามิเตอร์ของข้อสอบ ด้วยวิธีของเบส์ ดังนี้

$$\theta_i \sim N(0,1)$$
$$a_j \sim N(\mu_a, \sigma_a^2)$$
$$b_j \sim N(\mu_b, \sigma_b^2)$$
$$\gamma_{id(j)} \sim N\left(0, \sigma_{\gamma d(j)}^2\right)$$
$$\log[(c_j/(1-c_j)] \sim N(\mu_c, \sigma_c^2)$$

สำหรับการแจกแจงของค่าเฉลี่ย เป็นดังนี้ $\mu_a \sim N(0, V_a), \mu_b \sim N(0, V_b)$ และ $\mu_c \sim N(0, V_c)$ โดยที่ $V_{\mathrm{a}}^{-1} = V_{\mathrm{b}}^{-1} = V_{\mathrm{c}}^{-1} = 0$ และ $\sigma_\gamma^2 \sim \chi_{g\gamma}^{-2}$ สำหรับทุก Prior Variance โดย $\chi_{g_\gamma}^{-2}$ เป็น inverse chi-square random variable ด้วย degree of freedom g_γ ในการศึกษาของ Wainer et al. (2000) นี้ กำหนดให้ g_γ= 0.5 เท่ากันทุกการแจกแจง

โมเดลการตอบสนองข้อสอบสำหรับเทสต์เลทแบบ 1 พารามิเตอร์ (1PL - TRT)

Wang & Wilson (2005b) เสนอ Rasch Testlet Model สำหรับข้อสอบที่มีการให้คะแนนสองค่าและข้อสอบที่มีการให้คะแนนหลายค่า โดยใช้ Random Coefficient Multinominal Logit Model ซึ่งโมเดลมีลักษณะคล้ายกับ 2PL - TRT โดยที่ค่าอำนาจจำแนกเป็น 1 เท่ากันทุกข้อ

Wang & Wilson (2005b) อธิบายว่า ถ้า c_j= 0 และ a_j= 1 แล้ว โมเดลแบบ 2PL - TRT ของ Bradlow et al. (1999) จะลดรูปเป็นโมเดลแบบ 1PL - TRT ดังนี้

$$P_{ij}(y_{ij} = 1) = \frac{e^{(\theta - b_j - \gamma_{id(j)})}}{1 + e^{(\theta - b_j - \gamma_{id(j)})}} \quad (16)$$

โดยที่ความหมายของพารามิเตอร์และการแจกแจงเหมือนกับโมเดลแบบ 2PL – TRT

การตรวจสอบการทำหน้าที่ต่างกันสำหรับเทสต์เลท (Testlet)

เทสต์เลท (Testlet) ถูกนำมาใช้บ่อยครั้งในการสร้างแบบสอบ ตัวอย่างที่เห็นได้บ่อยครั้ง คือ การอ่านหนึ่งบทความเพื่อตอบคำถามหลายข้อ แม้จะทำให้มีประสิทธิภาพมากกว่าการใช้คำถามเพียงข้อเดียว แต่ก็ทำให้ขัดแย้งกับข้อตกลงเบื้องต้นของการใช้โมเดล IRT นั่นคือ ความเป็นอิสระในการตอบข้อสอบ (LID) ทำให้ข้อมูลไม่พอดี (Fit) กับโมเดล IRT แม้จะบังคับให้ชุดของข้อสอบนั้นเป็นอิสระ แต่ก็อาจจะขัดแย้งกับความเป็นมิติเดียว (Unidimentional)

การศึกษาเกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกัน สำหรับเทสต์เลทยังมีไม่มากนั้น จากการศึกษาที่ผ่านมา พบว่ามีการพัฒนาวิธีการและโมเดลต่าง ๆ เพื่อนำมาประยุกต์ใช้

ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบที่มีลักษณะเทสต์เลท โดยผู้วิจัยแบ่งวิธีการตรวจสอบการทำหน้าที่ต่างกัน สำหรับเทสต์เลทเป็น 2 ประเภท คือ

1. การจัดกระทำกับข้อมูล หรือปรับการให้คะแนน (Scoring) เช่น การนำวิธีวิเคราะห์ข้อสอบแบบ Polytomous มาประยุกต์ใช้ (Sireci et al., 1991; Thissen et al., 1989; Wainer, 1995 cited in Zhang, 2010) โดยมีแนวคิด คือ กำจัดความไม่เป็นอิสระในการตอบออกไป โดยรวมคะแนนข้อที่อยู่ใน Testlet เดียวกัน ให้เป็นข้อสอบแบบให้คะแนนหลายค่า (Polytomous) 1 ข้อ วิธีนี้จะกำจัดความไม่เป็นอิสระในการตอบออกไปได้ ทำให้หลีกเลี่ยงการประมาณค่าความเที่ยงของแบบสอบเกินจริง และสารสนเทศทางสถิติของ Polytomous IRT Model ยังมีความคงเส้นคงว่าดีกว่าการใช้ Standard Rasch Model อีกด้วย อย่างไรก็ตาม การนำ Polytomous IRT Model มาประยุกต์ใช้นั้น จะทำให้สารสนเทศในการตอบของผู้สอบที่เป็นรายข้อจะหายไป พารามิเตอร์ของข้อสอบบางข้อจะหายไปเมื่อเทียบกับข้อสอบแบบ Dichotomous ส่งผลให้ไม่มีข้อมูลเพื่อนำไปปรับปรุงแบบสอบ และอาจจะทำให้ประมาณค่า ความเที่ยงของแบบสอบต่ำกว่าความจริง (Yen, 1993) นอกจากนี้ การบังคับให้ผลการตอบสนองข้อสอบภายในเทสต์เลทเป็นการให้คะแนนแบบหลายค่า (Polytomous Scoring) จะสามารถตรวจสอบการทำหน้าที่ต่างกันได้เฉพาะระดับเทสต์เลทเท่านั้น

2. การพัฒนาโมเดลและการประยุกต์ใช้โมเดล เช่น การประยุกต์วิเคราะห์ข้อสอบแบบพหุระดับด้วยโมเดลเชิงเส้นตรงระดับลดหลั่น 3 ระดับ (Jiao et al., 2005) การเพิ่มพารามิเตอร์ DIF ใน Rasch Testlet Model โดย (Wang & Wilson, 2005a) และการประยุกต์ใช้ Bi - factor Multidimensional Item Response Theory Model for Testlets (Fukuhara & Kamata, 2011) เป็นต้น โดยอธิบายได้ดังนี้

Jiao et al. (2005) พัฒนา Testlet Model โดยประยุกต์ใช้การวิเคราะห์ข้อสอบแบบพหุระดับด้วยโมเดลเชิงเส้นตรงระดับลดหลั่น 3 ระดับ (Hierarchical Generalized Linear Model: HGLM - 3L) โดยใส่ Testlet Effect ในโมเดล HGLM-3L ด้วย นั่นคือ

ระดับการวิเคราะห์ที่ 1 ระดับข้อสอบ (Item Scores) แสดงรายละเอียดของข้อที่ j ในเทสต์เลทที่ d สำหรับผู้สอบที่ i ดังสมการ

$$log\left(\frac{p_{jdi}}{1-p_{jdi}}\right) = \eta_{jdi} = \pi_{0di} + \sum_{q=1}^{k} \pi_{qdi} X_{qjdi} \quad (17)$$

เมื่อ p_{jdi} หมายถึง ความน่าเป็นในการตอบถูกของข้อที่ j ในเทสต์เลทที่ d สำหรับผู้สอบที่ i

X_{qjdi} หมายถึง ตัวแปรดัมมี่ที่ q (q = 1, …, k) สำหรับข้อที่ j ใน Testlet ที่ d สำหรับผู้สอบที่ i ซึ่งมีค่าเป็น 1 เมื่อ q = ji และมีค่าเป็น 0 เมื่อ q ≠ j

π_{0di} หมายถึง เป็นจุดตัดแกนตั้ง (Intercept)

π_{qdi} หมายถึง เป็นสัมประสิทธิ์ (Slope) ของ X_{qjdi} เมื่อ q = 1,…,k

ระดับการวิเคราะห์ที่ 2 ระดับเทสต์เลท (Testlet - level) มีสมการเป็น

$$\begin{cases} \pi_{0di} = \gamma_{00i} + \gamma_{0di} \\ \pi_{1di} = \gamma_{10i} \\ \vdots \\ \pi_{kdi} = \gamma_{k0i} \end{cases} \quad (18)$$

เมื่อ γ_{00i} หมายถึง ค่าจุดตัดแกนตั้ง (Intercept) หรือค่าเฉลี่ยของ π_{0di}

γ_{0di} หมายถึง ค่าส่วนเหลือของ π_{0di} และมีการแจกแจง N(0, σ_{β}^{2})

ระดับการวิเคราะห์ที่ 3 ระดับผู้สอบ (Person - level) มีสมการเป็น

$$\begin{cases} \gamma_{00i} = u_{00i} \\ \gamma_{10i} = \beta_{100} \\ \quad \vdots \\ \gamma_{k0i} = \beta_{k00} \end{cases} \quad (19)$$

เมื่อ u_{00i} หมายถึง ตัวแปรสุ่ม (Random Variable) ของจุดตัดแกนตั้ง (Intercept) ระดับที่ 2 หรือ π_{00i} และมีการแจกแจง N(0, σ_{π}^{2})

นั่นคือ จะได้สมการ

$$log\left(\frac{p_{jdi}}{1-p_{jdi}}\right) = \eta_{jdi} = \sum_{q=1}^{k} \beta_{q00} X_{qjdi} + u_{00i} + \gamma_{0di} \quad (20)$$

ดังนั้น ความน่าจะเป็นในการตอบถูกของข้อที่ j ในเทสต์เลทที่ d สำหรับผู้สอบที่ i มีสมการเป็น

$$p_{jdi} = \frac{1}{1+\exp(-\eta_{qdi})} = \frac{1}{1+\exp\{-[(u_{00i}+\gamma_{0di})-\beta_{q00}]\}} \quad (21)$$

ซึ่ง HGLM - 3L นี้เป็นโมเดลอย่างง่ายโดยยังไม่ได้ใส่ตัวแปรทำนายในสมการ สำหรับการศึกษา DIF จะทำการขยายโมเดล HGLM - 3L โดยใส่ตัวแปรกลุ่มเข้าไป ในระดับที่ 2 ดังนี้ (Kamata, 2001)

$$\begin{cases} \pi_{0di} = \gamma_{00i} + \gamma_{01i} G_i + \gamma_{0di} \\ \pi_{1di} = \gamma_{10i} + \gamma_{11i} G_i \\ \quad \vdots \\ \pi_{kdi} = \gamma_{k0i} + \gamma_{k1i} G_i \end{cases} \quad (22)$$

เมื่อ G_i หมายถึง กลุ่มที่มี 2 ค่า คือ 0 และ 1 (Dichotomous group indicator)

γ_{11i} ถึง γ_{k1i} หมายถึง ขนาดของ DIF (DIF magnitude)

อย่างไรก็ตามโมเดล HGLM - 3L สามารถวิเคราะห์ได้เทียบเท่ากับ 1PL - TRT โดยที่ตัวชี้ค่าในโมเดลการวัดแบบพหุระดับที่เป็นพารามิเตอร์ความสามารถ (θ_i) พารามิเตอร์ความยาก (b_j) และพารามิเตอร์อิทธิพลของเทสต์เลท ($\gamma_{id(j)}$) ได้แก่ u_{00i}, β_{q00} และ γ_{0di} ตามลำดับ และโมเดล HGLM - 3L ยังมีข้อจำกัดว่าความแปรปรวนของอิทธิพลของเทสต์เลทมีค่าคงที่

ส่วนการใช้โมเดล Testlet Effect จะเพิ่มพารามิเตอร์อิทธิพลของเทสต์เลทเข้าไปในโมเดล IRT (Wang et al., 2002) และเพื่อเป็นการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ Wang & Wilson (2005a) จึงเพิ่ม DIF parameter ใน Rasch Testlet Model โดยเริ่มจากการประยุกต์ใช้ 3 - parameter Testlet Model (Wainer et al., 2000) ถ้า $c_j = 0$, $a_j = 0$ และ $\sigma^2_{\gamma_{id(j)}} = \sigma^2_{\gamma}$ (Variance เท่ากันทุก Testlets) แล้ว สมการจะลดรูปเหลือเป็น 1 - parameter Rasch Testlet Model นั่นคือ

$$t_{ij} = \theta_i - b_j - \gamma_{id(j)} \quad (23)$$

โดย $\gamma_{id(j)} \sim N\left(0, \sigma^2_{\gamma_{d(j)}}\right)$

ถ้า $\gamma_{id(j)} = 0$ (ไม่มีอิทธิพลของเทสต์เลท) แล้ว สมการจะลดรูปเป็น Dichotomous Rasch Model (Rasch, 1960) เมื่อเกิด DIF ในข้อสอบประเภท Dichotomous (j) ที่มี Testlet แล้ว จะได้สมการ Dichotomous Rasch Testlet Model ดังนี้

$$t_{ij} = \theta_i - b_{gj} - \gamma_{id(j)} \quad (24)$$

เมื่อกำหนดให้ g เป็นกลุ่มที่มีลักษณะต่างกันแล้ว ภายใต้สมการนี้ข้อสอบที่ i จะยอมให้มีความยากต่างกันในต่างกลุ่ม ดังนั้นจะได้สมการเป็น

$$t_{ij} = \theta_i - \left(b_j + \alpha_{jg}\right) - \gamma_{id(j)} \quad (25)$$

เมื่อ b_j เป็นค่าเฉลี่ยความยากของข้อสอบ j ทั้งกลุ่ม และ α_{jg} เป็นความต่างของค่าเฉลี่ย b_j ของกลุ่ม g

$$\alpha_{jg} = b_{jg} - b_j \quad (26)$$

และ $\sum_{g=1}^{G} \alpha_{jg} = 0 \quad (27)$

เมื่อ G เป็นจำนวนกลุ่ม และ $\alpha_{jg} \neq 0$ สำหรับทุกกลุ่ม g หรือความยากจะต้องไม่เป็นค่าคงที่เท่ากันทุกกลุ่ม ดังนั้น α_{jg} จึงใช้ขนาดของ DIF (magnitude of DIF) หรือใช้แสดงพารามิเตอร์ของ DIF สำหรับ 2 กลุ่มเท่านั้น (G = 2) โดยที่ g จะแทนกลุ่มอ้างอิงหรือกลุ่มเปรียบเทียบก็ได้ โดยที่ความต่างของความยากของ 2 กลุ่ม เท่ากับ α_{jg} หรือเขียนสมการได้ดังนี้

$$t_{ij} = \theta_i - b_j - \gamma_{id(j)} - \beta_j G_i \quad (28)$$

เมื่อ $\gamma_{id(j)}$ หมายถึง อิทธิพลสุ่มของ Testlet d(j) ของผู้สอบ i ที่แฝงใน testlet d(j)

b_j หมายถึง พารามิเตอร์ความยากของข้อ j

β_j หมายถึง ขนาดของ DIF (DIF magnitude) ของข้อ j

G_i หมายถึง กลุ่มของผู้สอบ

โดยที่ $P_{ij}(y_{ij} = 1|t_{ij}) = logit^{-1}(t_{ij})$

และ $logit^{-1}(t_{ij}) = \exp(t_{ij})/(1 + \exp(t_{ij}))$

แม้โมเดลของ Wang & Wilson (2005 b) จะรวมอิทธิพลสุ่มของเทสต์เลทเข้าในสมการ แต่โมเดลยังมีข้อจำกัด คือ โมเดลนี้อยู่บนพื้นฐานของ Rasch Model ซึ่งพารามิเตอร์อำนาจจำแนกมีค่าเท่ากันทุกข้อ ทำให้ไม่ยืดหยุ่นเมื่อนำไปใช้กับแบบสอบทั่วไป และไม่มีพารามิเตอร์สำหรับควบคุมความต่างของค่าเฉลี่ยความสามารถระหว่างกลุ่มอ้างอิงและกลุ่มควบคุม ส่งผลให้เกิดอคติ (Bias) ในการประมาณค่าขนาดของ DIF (DIF magnitude) นอกจากนี้ การกำหนดให้ Variance มีค่าเท่ากันทุกเทสต์เลทก็อาจไม่สอดคล้องกับการปฏิบัติจริง

Fukuhara & Kamata (2011) พัฒนา Bi - factor Multidimensional IRT Model for Testlets with Covariates เพื่อรองรับการวิเคราะห์แบบ 2 พารามิเตอร์ได้ และยังไม่ทำให้สารสนเทศรายข้อหายไป ซึ่งสอดคล้องกันความเป็นจริงมากกว่าโมเดลที่ได้กล่าวมาแล้ว โดยโมเดลของ Fukuhara & Kamata (2011) พัฒนาจาก DeMars (2006) และ Li et al. (2006) ซึ่งมีสมการ ดังนี้

$$\ln\left(\frac{P(y_{ij}=1)}{P(y_{ij}=0)}\right) = a_j\theta_i - \delta_j + \lambda_j\gamma_{id(j)} - \beta'_j G_i \quad (29)$$

เมื่อ θ_i หมายถึง มิติที่ 1 หรือ ความสามารถของคนที่ i

$\gamma_{id(j)}$ หมายถึง มิติที่ 2 หรือ อิทธิพลสุ่มของ Testlet d(j)

δ_j หมายถึง พารามิเตอร์ความยากของข้อ j

a_j และ λ_j หมายถึง พารามิเตอร์อำนาจจำแนกของความสามารถและอิทธิพลของเทสต์เลท

β'_j หมายถึง ความต่างของพารามิเตอร์ความยากระหว่างกลุ่ม (ขนาดของ Uniform DIF ของข้อ i)

G_i หมายถึง กลุ่มของผู้สอบ (G_i= 1 เมื่อผู้สอบเป็นกลุ่ม F และ G_i= 0 เมื่อผู้สอบเป็นกลุ่ม R)

โดยที่ให้ความสามารถและอิทธิพลสุ่มเนื่องจากเทสต์เลท เป็นอิสระและมีการแจกแจงแบบปกติมาตรฐาน (Standard Normal Distribution) ขนาดของ Testlet Effect เป็นอัตราส่วนของสัมประสิทธิ์ (Slope) ของมิติที่ 2 ต่อสัมประสิทธิ์ (Slope) ของมิติที่ 1 ดังนั้น ถ้าให้กลุ่มเปรียบเทียบเป็น G_i= 1 แล้วความยากของกลุ่มเปรียบเทียบเป็น $\delta_j - \beta_j'$ ส่วนความยากของกลุ่มอ้างอิง (G_i= 0) เป็น δ_j ทำให้เขียนสมการใหม่ได้เป็น

$$\ln\left(\frac{P(y_{ij}=1)}{P(y_{ij}=0)}\right) = a_j(\theta_i - b_j + C_j\gamma_{id(j)} - \beta_j G_i) \quad (30)$$

เมื่อ $b_j = \frac{\delta_j}{a_j}$, $C_j = \frac{\lambda_j}{a_j}$, $\beta_j = \frac{\beta_j'}{a_j}$ และ $\gamma_{id(j)}$ มีการแจกแจงปกติ (Normal Distribution) ที่มีค่าเฉลี่ยเป็น 0 และความแปรปรวนเป็น $\sigma^2_{\gamma_d}$ สำหรับ Testlet d(j) แทนการแจกแจงแบบปกติมาตรฐาน และเขียนสมการใหม่ได้เป็น

$$\ln\left(\frac{P(y_{ij}=1)}{P(y_{ij}=0)}\right) = a_j(\theta_i - b_j + \frac{C_j}{\sigma_{\gamma_d}}\gamma_{id(j)} - \beta_j G_i) \quad (31)$$

Li et al. (2006) กำหนดให้ ถ้า $C_j = \sigma_{\gamma_d}$ สำหรับทุกข้อภายในเทสต์เลทเดียวกันแล้ว สามารถเขียนสมการใหม่ได้เป็น

$$\ln\left(\frac{P(y_{ij}=1)}{P(y_{ij}=0)}\right) = a_j(\theta_i - b_j + \gamma_{id(j)} - \beta_j G_i) \quad (32)$$

ซึ่งเป็นโมเดลที่ขยายจาก Testlet Response Theory หรือกล่าวได้ว่าเป็นกรณีพิเศษ (Special case) ของ Bi - factor MIRT Model ที่กำหนดให้ความสามารถและอิทธิพลสุ่ม เนื่องจาก Testlet มีระดับอำนาจจำแนกเดียวกัน จากนั้นทำการควบคุมความต่างของค่าเฉลี่ยความสามารถระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบให้อยู่ในระดับเดียวกัน ซึ่งมีสมการเป็น

$$\theta_i = \beta_\theta G_i + \zeta_i \quad (33)$$

เมื่อ β_θ หมายถึง อิทธิพลของกลุ่ม G_i ต่อความสามารถ θ_i

ζ_i หมายถึง ส่วนที่เหลือ (Residual) สำหรับผู้สอบ i

ดังนั้น เขียนสมการใหม่ที่มี 2 latent factors ได้แก่ ζ_i และ $\gamma_{id(j)}$ ซึ่งทำนาย logit ของการตอบข้อสอบแต่ละข้อถูก ได้ดังนี้

$$\ln\left(\frac{P(y_{ij}=1)}{P(y_{ij}=0)}\right) = a_j(\beta_\theta G_i + \zeta_i - b_j + \gamma_{id(j)} - \beta_j G_i) \quad (34)$$

ส่วนกรณีการวิเคราะห์ค่าพารามิเตอร์และการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยไม่คำนึงถึงอิทธิพลของเทสต์เลทนั้น Fukuhara & Kamata (2011) ได้ทำการขยายโมเดลแบบ 2PL - IRT ซึ่งโมเดลจะมีลักษณะคล้ายกับ MIMIC model ซึ่งมีสมการ ดังนี้

$$\ln\left(\frac{P(y_{ij}=1)}{P(y_{ij}=0)}\right) = a_j\left(\theta_i - b_j - \beta_j G_i\right) \quad (35)$$

หาก $\theta_i = \beta_\theta G_i + \zeta_i$ แล้ว จะขยนสมการใหม่ได้เป็น

$$\ln\left(\frac{P(y_{ij}=1)}{P(y_{ij}=0)}\right) = a_j(\beta_\theta G_i + \zeta_i - b_j - \beta_j G_i) \quad (36)$$

โดยที่การพิจารณาการทำหน้าที่ต่างกันของข้อสอบ จะพิจารณาจากค่า β_j ซึ่งถ้าค่า β_j มีค่าขอบล่างของช่วงความเชื่อมั่น 95% ซึ่งตรงกับตำแหน่งเปอร์เซ็นไทล์ที่ 2.5 (val 2.5 pc) และค่าขอบบนของช่วงความเชื่อมั่น 95% ซึ่งตรงกับตำแหน่งเปอร์เซ็นไทล์ที่ 97.5 (val 97.5 pc) ไม่คลุมศูนย์ และค่าสัมบูรณ์ของขนาด β_j มากกว่า 0.426 (Vaughn, 2006 อ้างถึงใน สุพัฒนา หอมบุปผา, 2556) แสดงว่าข้อสอบข้อนั้น ทำหน้าที่ต่างกันของข้อสอบ (DIF)

ผลการวิเคราะห์ใน Fukuhara & Kamata (2011) นั้น กำหนดการแจกแจงก่อนคล้ายกับ Bradlow et al. (1999) และ Li et al. (2006) โดยได้การแจกแจงภายหลังของแต่ละพารามิเตอร์โดยใช้แจกแจงก่อนที่ไม่ให้ข้อมูล (non-informative prior distribution) และกำหนดให้มีความแปรปรวนมากสำหรับการแจกแจงก่อนของแต่ละพารามิเตอร์ ดังตารางที่ 2 - 5

ตารางที่ 2 - 5 Prior Distributions ของพารามิเตอร์และ Hyperparameter

พารามิเตอร์	Prior Distributions ของพารามิเตอร์	Hyperparameter ของพารามิเตอร์	Prior Distributions ของ Hyperparameter
ζ_i	$N(0,1)$		
a_j	$N(\mu_a, \sigma_a^2)I(0,\alpha)$	μ_a	$N(0,1000)$
		σ_a^2	$Inv-\chi^2(0.5)$
b_j	$N(\mu_b, \sigma_b^2)$	μ_b	$N(0,1000)$
		σ_b^2	$Inv-\chi^2(0.5)$
β_j	$N(\mu_\beta, \sigma_\beta^2)$	μ_β	$N(0,1000)$
		σ_β^2	$Inv-\chi^2(0.5)$
β_θ	$N\left(\mu_{\beta_\theta}, \sigma_{\beta_\theta}^2\right)$	μ_{β_θ}	$N(0,1000)$
		$\sigma_{\beta_\theta}^2$	$Inv-\chi^2(0.5)$
$\gamma_{id(j)}$	$N(0, \sigma_{\gamma_d}^2)$	$\sigma_{\gamma_d}^2$	$Inv-\chi^2(0.5)$

จากที่กล่าวมาสามารถสรุปข้อดีและข้อจำกัดของการประมาณค่าพารามิเตอร์และการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสำหรับแบบสอบที่มีลักษณะของเทสต์เลทในรูปแบบต่าง ๆ ได้ดังตารางที่ 2 - 6

ตารางที่ 2 - 6 ข้อดีและข้อจำกัดของการประมาณค่าพารามิเตอร์และการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสำหรับแบบสอบที่มีลักษณะของเทสต์เลทรูปแบบต่าง ๆ

รูปแบบ	วิธีการ	ประโยชน์	ข้อจำกัด
การปรับการให้คะแนน (Scoring)	รวมคะแนนข้อที่อยู่ใน เทสต์เลทเดียวกัน ให้เป็นข้อสอบแบบให้คะแนนหลายค่า (Polytomous) 1 ข้อ แล้วใช้การวิเคราะห์ของ Polytomous IRT Model	- กำจัดความไม่เป็นอิสระในการตอบข้อสอบออกได้ - มีความคงเส้นคงวามากกว่า IRT Model	- ไม่สามารถวิเคราะห์หาสารสนเทศระดับข้อสอบรายข้อได้ - ตรวจสอบ DIF ได้เฉพาะระดับเทสต์เลท

ตารางที่ 2 - 6 (ต่อ)

รูปแบบ	วิธีการ	ประโยชน์	ข้อจำกัด
การประยุกต์วิเคราะห์ข้อสอบแบบพหุระดับด้วยโมเดลเชิงเส้นตรงระดับลดหลั่น 3 ระดับ	ระดับที่ 1 ระดับข้อสอบ ระดับที่ 2 ระดับเทสต์เลท ระดับที่ 3 ระดับผู้สอบ ในการศึกษา DIF ใส่ตัวแปรกลุ่มเข้าไป ในระดับที่ 2	- กำจัดปัญหาความไม่เป็นอิสระในการตอบข้อสอบออกได้ - เพิ่มตัวแปรทำนายในสมการได้	- โมเดล HGLM - 3L วิเคราะห์ได้เทียบเท่ากับ 1PL - TRT - ถือว่าความแปรปรวนของ อิทธิพลแต่ละเทสต์เลท มีค่าคงที่
Testlet Response Theory Models (TRT)	กำหนดให้ เทสต์เลทถือเป็นหน่วยในการวิเคราะห์ด้วย โดยเพิ่มส่วนที่เป็น Random Effect ($\gamma_{id(j)}$) เข้าไปในสมการของโมเดลแบบ IRT	- ผ่อนคลายข้อตกลงที่เกี่ยวกับความเป็นอิสระระหว่างข้อสอบและผู้สอบ - กรณีไม่มี DIF สามารถประมาณค่าพารามิเตอร์ได้ทั้ง 1, 2 และ 3 พารามิเตอร์	- กรณีศึกษ DIF วิเคราะห์ได้เฉพาะ 1 พารามิเตอร์ (b_j) - ไม่มีพารามิเตอร์ในการควบคุมความต่างของค่าเฉลี่ยความสามารถระหว่างกลุ่มอ้างอิงและกลุ่มควบคุม - ความแปรปรวนของอิทธิพลของเทสต์เลทในแต่ละ เทสต์เลทมีค่าเท่ากัน (กรณีศึกษา DIF)
การประยุกต์ใช้ Bi-factor Multidimensional Item Response Theory Model	โมเดลมี 2 คุณลักษณะ ได้แก่ คุณลักษณะหลัก (Primary Trait) เป็นการตอบสนองของข้อสอบแต่ละข้อหรือความสามารถ และคุณลักษณะที่ 2 เป็นคุณลักษณะของเทสต์เลท (Testlet trait) หรืออิทธิพลของเทสต์เลท	- กำจัดปัญหาความไม่เป็นอิสระในการตอบข้อสอบออกได้ - สารสนเทศรายข้อไม่หายไป - มีความพอดีกับข้อมูลที่มีความไม่เป็นอิสระในการตอบโดยมีสาเหตุจากเทสต์เลทมากกว่าโมเดล TRT	- รองรับการวิเคราะห์แบบ 2 พารามิเตอร์ - Fukuhara & Kamata (2011) กำหนดให้ความสามารถและอิทธิพลของเทสต์เลทมีอำนาจจำแนกเท่ากัน ซึ่งหากกำหนดให้มีค่าต่างกันได้จะมีความยืดหยุ่นและพอดีกับข้อมูลมากกว่า

จากตารางที่ 2 - 6 ผู้วิจัยจึงเลือกใช้ Bi-factor Multidimensional IRT Model ของ Fukuhara & Kamata (2011) เนื่องจากเป็นโมเดลที่รองรับการวิเคราะห์แบบ 2 พารามิเตอร์ ซึ่งสอดคล้องกับการศึกษาในครั้งนี้ แม้ความสามารถและอิทธิพลของเทสต์เลทจะมีการแบ่ง (Share*)* อำนาจจำแนกเดียวกัน ซึ่งอาจทำให้มีความยืดหยุ่น (Flexible*)* และพอดีกับข้อมูลน้อยกว่าการกำหนดให้ความสามารถและอิทธิพลของเทสต์เลทมีอำนาจจำแนกที่ต่างกัน อย่างไรก็ตาม ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ Fukuhara & Kamata (2011) ได้ทำการปรับพารามิเตอร์ความสามารถและอิทธิพลของเทสต์เลท ดังสมการ

$$\beta_{\theta}^{adj} = \beta_{\theta} - \bar{\beta} \tag{37}$$

$$\beta_{j}^{adj} = \beta_{j} - \bar{\beta} \tag{38}$$

นอกจากนี้ เมื่อพิจารณาข้อจำกัดของแต่ละโมเดลแล้ว Bi - factor MIRT เป็นโมเดลที่สอดคล้องกับสถานการณ์จริงมากกว่า เช่น ไม่มีข้อจำกัดในเรื่องความแปรปรวนของอิทธิพลของเทสต์เลท นอกจากนี้ ยังสามารถวิเคราะห์หาสารสนเทศและตรวจสอบการทำหน้าที่ต่างกันในระดับข้อสอบได้

ตอนที่ 4 แนวคิดเกี่ยวกับการประมาณค่าพารามิเตอร์ด้วยวิธีของเบส์และวิธีแมกซิมัมไลค์ลิฮู้ด

สำหรับการศึกษาครั้งนี้ ผู้วิจัยใช้วิธีประมาณค่าพารามิเตอร์ในการวิเคราะห์ข้อสอบด้วยทฤษฎีการตอบสนองข้อสอบ 3 วิธี ได้แก่ วิธีแมกซิมัมไลค์ลิฮู้ด วิธีของเบส์ (Baye) และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayeγ) โดยที่วิธีแมกซิมัมไลค์ลิฮู้ดและวิธีของเบส์ (Baye) ใช้โมเดลสำหรับประมาณค่าแบบเดียวกันแต่ต่างวิธี ส่วนวิธีของเบส์ (Baye) และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayeγ) ใช้วิธีการประมาณค่าวิธีเดียวกันแต่ใช้โมเดลต่างกัน นั่นคือ ในการศึกษาครั้งนี้จะใช้วิธีการประมาณค่า 2 วิธี คือ วิธีแมกซิมัมไลค์ลิฮู้ดและวิธีของเบส์ ดังมีรายละเอียด ดังนี้

การประมาณค่าพารามิเตอร์ด้วยวิธีแมกซิมัมไลค์ลิฮู้ด

แนวคิดของวิธีแมกซิมัมไลค์ลิฮู้ด เป็นการประมาณค่าพารามิเตอร์โดยอาศัยผลที่ได้จากการสังเกตค่าที่วัดได้จากตัวอย่างสุ่มที่เลือกมาแจกแจงที่ทราบรูปแบบของฟังก์ชันความหนาแน่นแต่ไม่ทราบค่าพารามิเตอร์ แล้วใช้ค่าของหน่วยตัวอย่าง ($X_1 = x_1, X_2 = x_2, X_3 = x_3, \ldots, X_n = x_n$)

พิจารณาหาค่าประมาณพารามิเตอร์ (β) ซึ่งโอกาสที่จะวัดค่าตัวอย่างสุ่มได้ อาจแสดงได้ด้วยฟังก์ชันความหนาแน่นร่วมของหน่วยตัวอย่าง แต่ฟังก์ชันความหนาแน่นร่วมนี้ขึ้นอยู่กับพารามิเตอร์ β ดังนั้น ค่าประมาณของ β ก็คือ ค่าของ β ที่ทำให้ฟังก์ชันความหนาแน่นร่วมมีค่าสูงสุด ถ้า $f(\mathrm{X};\beta)$ เป็นฟังก์ชันความหนาแน่นของตัวแปรสุ่มและจากตัวอย่างสุ่มขนาด n ได้ค่าสังเกตตัวอย่างเป็น ($X_1 = x_1, X_2 = x_2, X_3 = x_3, \ldots, X_n = x_n$) ดังนั้น ฟังก์ชันความหนาแน่นร่วมของตัวอย่างสามารถแสดงได้ดังนี้

$$f(\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n;\beta) = f(\mathrm{X}_1;\beta)\, f(\mathrm{X}_2;\beta) \ldots f(\mathrm{X}_n;\beta)$$
$$= \prod_{i=1}^{n} f(\mathrm{x}_i;\beta) \qquad (39)$$

เรียกฟังก์ชันนี้ว่า ฟังก์ชันไลค์ลิฮู้ด (Likelihood function) มักใช้สัญลักษณ์ว่า $L(\beta)$ หมายถึง การหาค่า β ที่จะทำให้ $L(\beta)$ มีค่าสูงสุด ซึ่งส่วนใหญ่จะทำโดยการหาค่า β ที่ให้ลอการึทึม (Logarithm) ของฟังก์ชันไลค์ลิฮู้ดมีค่าสูงสุด หรือ $\ln L(\beta)$ ค่าสูงสุด ทั้งนี้เพราะค่าประมาณของพารามิเตอร์ β ที่ทำให้ฟังก์ชันทั้ง 2 ฟังก์ชันมีค่าสูงสุดเป็นค่าเดียวกัน แต่การหาค่าประมาณที่ทำให้ลอการึทึมของฟังก์ชันไลค์ลิฮู้ดมีค่าสูงสุดทำง่ายกว่า

Glas, Wainer, & Bradlow (2000) อธิบาย การคำนวณโดยใช้วิธีวิธีแมกซิมัมไลค์ลิฮู้ดสำหรับโมเดลตามทฤษฎีการตอบสนองข้อสอบแบบเทสต์เลท ดังนี้

จาก likelihood ของโมเดลการตอบสนองข้อสอบสำหรับเทสต์เลท

$$L(Y|T) = \prod_{i=1}^{I} \prod_{j=1}^{J} \left(\frac{e^{t_{ij}}}{1+e^{t_{ij}}}\right)^{y_{ij}} \left(\frac{1}{1+e^{t_{ij}}}\right)^{1-y_{ij}} \qquad (40)$$

เมื่อ $t_{ij} = a_j(\theta_i - b_j + \gamma_{id(j)})$

$\theta_i = (\theta_1, \ldots, \theta_I)$

$a_j = (a_1, \ldots, a_J)$

$b_j = (b_1, \ldots, b_J)$

$\gamma = (\gamma_{1d(1)}, \ldots, \gamma_{Id(J)})$

หาค่าสูงสุดสำหรับ $\Lambda = (\theta, a, b, \gamma)$ โดยการตัดส่วนที่เป็นเทอมของ θ และ γ ซึ่งกระทำได้โดยการ integration จากโมเดล จะได้

$$L\left(Y\middle|\mathrm{a},\mathrm{b},\sigma_\gamma^2\right) = \iint \prod_{i=1}^{I}\prod_{j=1}^{J} p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}}\,dF\,\theta_i F\,\gamma_{id(j)}$$

$$= \iint \prod_{i=1}^{I}\prod_{j=1}^{J}\left(\frac{e^{t_{ij}}}{1+e^{t_{ij}}}\right)^{y_{ij}}\left(\frac{1}{1+e^{t_{ij}}}\right)^{1-y_{ij}} N(0,1)N(0,\sigma_\gamma^2)\,d\,\theta_i\gamma_{id(j)} \quad (41)$$

จากนั้นจึงทำการหาอนุพันธ์ (Differentiate) เทียบกับพารามิเตอร์ที่ต้องการ ได้แก่

$$\mathrm{dL}\left(Y\middle|\mathrm{a},\mathrm{b},\sigma_\gamma^2\right)/\mathrm{da} = 0 \quad (42)$$

$$\mathrm{dL}\left(Y\middle|\mathrm{a},\mathrm{b},\sigma_\gamma^2\right)/\mathrm{db} = 0 \quad (43)$$

$$\mathrm{dL}\left(Y\middle|\mathrm{a},\mathrm{b},\sigma_\gamma^2\right)/\mathrm{d}\sigma_\gamma^2 = 0 \quad (44)$$

สำหรับการประมาณค่าพารามิเตอร์ $\mathrm{a},\mathrm{b},\sigma_\gamma^2$ ที่ทำให้ $\ln \mathrm{L}\left(Y\middle|\mathrm{a},\mathrm{b},\sigma_\gamma^2\right)$ มีค่าสูงสุด ทำได้โดยการหาอนุพันธ์ (Differentiate) เทียบกับพารามิเตอร์ $\mathrm{a},\mathrm{b},\sigma_\gamma^2$ แล้วผลลัพธ์เป็นศูนย์ ดังสมการข้างต้น จะได้สมการไม่เป็นเส้นตรง p + 1 สมการ ดังนั้นการหาค่าพารามิเตอร์นี้ จึงไม่อาจหาได้โดยวิธีทั่วไป แต่หาได้โดยวิธีนิวตัน-ราฟสัน (Newton-Raphson algorithm) ด้วยการประมาณค่าซ้ำ ๆ (Iterative) จนมีค่าคงที่ ซึ่งมีขั้นตอน ดังนี้

สมมติให้ $f \in C^2[a,b]$ คือฟังก์ชัน f หาอนุพันธ์ที่สองได้และต่อเนื่องบนช่วงปิด [a,b] กำหนดให้ $p_0 \in [a,b]$ หรือ p_0 มีค่าในช่วงปิด [a,b] เป็นค่าประมาณของ p ซึ่งทำให้ $f'(p_0) \neq 0$ และ $|p_0 - p|$ มีค่าน้อย ๆ พิจารณา $f(x)$ ที่จุด p_0 จะได้

$$f(x) = f(p_0) + f(x - p_0)\,f'(p_0) + \frac{(p-p_0)^2}{2} f''(\xi(x)) \quad (45)$$

เมื่อ $\xi(x)$ มีค่าระหว่าง x และ p_0 ดังนั้น จาก $f(p) = 0$ จะแทนค่า x ด้วย p จะได้สมการ ดังนี้

$$0 = f(p_0) + f(p - p_0)\,f'(p_0) + \frac{(p-p_0)^2}{2} f''(\xi(p)) \quad (46)$$

เนื่องจาก $|p_0 - p|$ มีค่าน้อย ๆ ดังนั้น $(p - p_0)^2$ จึงไม่มีผลต่อการพิจารณา และจะได้

$$0 \approx f(p_0) + f(p - p_0)\, f'(p_0) \tag{47}$$

เมื่อแก้สมการหาค่าของ p จะได้ $p \approx p_0 - \frac{f(p_0)}{f'(p_0)}$ และใช้จุดเริ่มต้นที่ $p_0, p_1, \ldots, p_{n-1}, p_n$ ตามลำดับ จะได้

$$p_n \approx p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})}, \quad n \geq 1 \tag{48}$$

ประมาณค่าซ้ำ (Iterative) จนกว่าค่าประมาณจะมีค่าลู่ (Convergence) เข้าค่าใดค่าหนึ่ง คือ ค่าประมาณครั้งที่ n - 1 และครั้งที่ n มีค่าต่างกันเข้าใกล้ศูนย์มาก เช่น น้อยกว่า 0.00005 จะได้ค่าที่ p_n เป็นค่าที่ประมาณได้

จุดเด่นของการประมาณค่าด้วยวิธีแมกซิมัมไลค์ลิฮู้ด ได้แก่ (นพดล มีชั้นช่วง, 2544)

1. เมื่อจำนวนข้อสอบและจำนวนผู้สอบเพิ่มขึ้น การประมาณจะได้ค่าที่มีความคงที่เข้าสู่ค่าพารามิเตอร์ที่แท้จริง
2. ฟังก์ชันของค่าสถิติมีความเพียงพอของสารสนเทศทั้งหมดเกี่ยวกับพารามิเตอร์
3. มีความแปรปรวนน้อย
4. มีการกระจายเข้าใกล้การกระจายแบบโค้งปกติ

อย่างไรก็ตามการประมาณค่าด้วยวิธีแมกซิมัมไลค์ลิฮู้ดมีข้อจำกัด ดังนี้

1. กรณีที่จำนวนข้อสอบและผู้สอบไม่มาก จุดเด่นของของการประมาณค่าด้วยวิธีแมกซิมัมไลค์ลิฮู้ด จะเป็นจริงเมื่อประมาณค่าพารามิเตอร์ความสามารถ หรือพารามิเตอร์ข้อสอบเพียงอย่างใดอย่างหนึ่ง
2. การประมาณค่าพารามิเตอร์ในขั้นที่ 2 และ 3 โดยใช้ค่าอนุพันธ์อันดับที่ 2 เป็นตัวหารในกระบวนการนิวตัน-ราฟสัน มีโอกาสที่ค่าประมาณที่ได้จะไม่ลู่เข้าสู่ค่าคงที่ ซึ่งอาจหลีกเลี่ยงปัญหานี้ได้โดยการใช้ค่าสารสนเทศ (Information function) แทน
3. สำหรับการประมาณค่าในสมการไลค์ลิฮูดไม่ใช่สมการเชิงเส้นตรง จะทำให้การหารากของสมการที่ทำให้ฟังก์ชันไลค์ลิฮูดมีค่าสูงสุดได้หลายค่าแต่ค่าเหล่านี้ ไม่สามารถนำไปใช้หรือประกันได้ว่าเป็นค่าพารามิเตอร์ที่แท้จริงได้

4. ในบางครั้งค่าพารามิเตอร์หรือค่าที่ได้จากการประมาณไม่ตกอยู่ในขอบเขตของค่าพารามิเตอร์ กล่าวคืออาจมีค่าใดค่าหนึ่งอยู่ภายนอกขอบเขตที่ยอมรับได้ ในกรณีเช่นนี้ต้องมีการกำหนดขอบเขตจำกัดของค่าประมาณไว้ เพื่อให้ค่าประมาณที่ได้ไม่สูงหรือต่ำเกินไปนัก แต่การกระทำเช่นนี้เป็นจุดอ่อนของกระบวนการประมาณค่าด้วยวิธีแม็กซิมั่มไลค์ลิฮูด โดยเฉพาะในแบบจำลอง 2 และ 3 พารามิเตอร์ จึงทำให้เกิดปัญหาตามมาเกี่ยวกับความตรง (Validity) ของค่าที่ประมาณได้

การประมาณค่าพารามิเตอร์ด้วยวิธีของเบส์

การวิเคราะห์สถิติเชิงอนุมานมี 2 แนวคิดหลัก ได้แก่ แนวคิดของ Frequentists และแนวคิดของ Bayesians โดยแนวคิดของ Frequentists พารามิเตอร์จะถูกกำหนดให้ไม่ทราบค่าและเป็นจำนวนที่แน่นอน (Fixed) วัตถุประสงค์ของ Frequentists คือ การกำหนดค่าประมาณช่วงความเชื่อมั่นรอบจุดของพารามิเตอร์ที่น่าสนใจและการทดสอบสมมติฐานผ่านสมมติฐานศูนย์ (Null hypothesis) ส่วนแนวคิดของ Bayesians พารามิเตอร์จะถูกกำหนดให้ไม่ทราบค่าและเป็นจำนวนสุ่ม (Random) ซึ่งพารามิเตอร์มีการแจกแจงความน่าจะเป็น ซึ่งอาจกำหนดขึ้นโดยใช้ความรู้ก่อนหน้า ความเชื่อ หรือข้อมูลข้างเคียง ในแนวคิดของ Bayesians มีเป้าหมายเพื่อหาการแจกแจงภายหลัง (Posterior distribution) ของพารามิเตอร์ที่สนใจ โดยผ่านฟังก์ชัน Likelihood และการแจกแจงก่อน (Prior distributions) และทดสอบสมมติฐานโดยสร้างการประมาณค่าด้วยช่วงแบบ Bayes และ posterior p - values ซึ่งนำไปใช้ประโยชน์ เช่น การประมาณค่าพารามิเตอร์ การทดสอบสมมติฐานเชิงสถิติ

ปัจจุบันมีการนำแนวคิดของ Bayesians มาใช้อย่างแพร่หลายทั้งด้านวิทยาศาสตร์ สังคมศาสตร์และธุรกิจ แม้ทฤษฎีการประมาณค่าพารามิเตอร์แบบ Bayes จะต้องใช้ความรู้ทางด้านทฤษฎีความน่าจะเป็นและสถิติเชิงอนุมาน ซึ่งหากไม่มีความรู้ทางสถิติ อาจทำให้เกิดปัญหาในการทำความเข้าใจเมื่อนำไปใช้

ทฤษฎีของ Bayes (Bayes' Theorem)

สำหรับแนวคิดแบบ Bayes ตัวแบบที่สังเกตได้ $x = (x_1, x_2, \dots, x_n)^T$ ที่กำหนดเงื่อนไขบนค่าพารามิเตอร์ที่ไม่ทราบค่า $\theta = (\theta_1, \theta_2, \dots, \theta_p)^T$ จะอยู่ในรูปของการแจกแจงความน่าจะเป็น $\int x|\theta$ ซึ่งมาจากการที่กำหนดให้ θ เป็นตัวแปรสุ่ม Random Variable (Θ)

การที่ θ เป็นการแจกแจงก่อน (Prior) นั้น แทนด้วยสัญลักษณ์ของ Prior Density ของ θ เป็น $\pi(\theta)$ นั่นคือ Prior จะถูกแทนค่าของ Density θ ดังนั้น ถ้าเก็บข้อมูล $X = x$ จะได้เงื่อนไขของ Density Given $X = x$ ซึ่งเรียกว่า Posterior เขียนแทนด้วย $x|\theta$ และมีการแจกแจงก่อนเป็น $\pi(\theta|\eta)$

โดยที่ η เป็น เวกเตอร์ของพารามิเตอร์ขั้นที่สอง (Hyperparameter) ในแนวคิดแบบ Bayes นี้ การอนุมานเกี่ยวกับ θ ขึ้นอยู่กับการแจกแจงภายหลัง (Posterior Distribution) ซึ่งมีรูปแบบเป็น

$$p(\theta|x,\eta) = \frac{p(x,\theta|\eta)}{p(x|\eta)} = \frac{\int(x|\theta)\pi(\theta|\eta)}{p(x|\eta)} \tag{49}$$

เมื่อ $p(x|\eta) = \sum_{\theta}\int(x|\theta)\pi(\theta|\eta)$,θ เป็นตัวแปรไม่ต่อเนื่อง

$\int f(x|\theta)\pi(\theta|\eta)d\theta$,θ เป็นตัวแปรต่อเนื่อง

โดยสมการดังกล่าวเรียกว่า ทฤษฎีของ Bayes (Bayes' theorem) จะสังเกตได้ว่าข้อมูลมีสองช่วง นั่นคือ ข้อมูลที่ได้จากการทดลองหรือข้อมูลใหม่ที่ได้รับ (อยู่ในรูปของความน่าจะเป็น f) และค่าเชื่อก่อน (ซึ่งอยู่ในรูปของความน่าจะเป็นก่อนการทดลอง π) โดยค่าอินทิกรัลของตัวส่วนในสมการ เรียกว่า การแจกแจงหน่วยสุดท้าย (Marginal distribution) ของ x เมื่อกำหนดพารามิเตอร์ขั้นที่ 2 เป็น η และเขียนได้ในรูปของ m(x|η)

อย่างไรก็ตาม หากทราบค่า η ก็สามารถตัดตัวแปรนี้ออกจากสมการได้ เนื่องจากไม่จำเป็นต้องกำหนดเงื่อนไข กรณีเป็นค่าคงที่ ดังนั้น สามารถเขียนรูปแบบของการแจกแจงใหม่ให้ง่ายขึ้น ได้เป็น

$$p(\theta|x) = \frac{p(x,\theta)}{p(x)} = \frac{\int(x|\theta)\pi(\theta)}{p(x)} \tag{50}$$

เมื่อ $p(x|\eta) = \sum_{\theta}\int(x|\theta)\pi(\theta|\eta)$, θ เป็นตัวแปรไม่ต่อเนื่อง

$\int f(x|\theta)\pi(\theta|\eta)d\theta$, θ เป็นตัวแปรต่อเนื่อง

เมื่อพิจารณาค่าพารามิเตอร์ที่มีลักษณะต่อเนื่อง จากทฤษฎีของ Bayes สามารถสร้างการกระจายของ θ|x ที่เรียกว่า Posterior Distribution ของ θ โดยให้ Posterior Distribution ของ θ เป็น p(θ|x) เขียนได้เป็น

$$p(\theta|x) = \frac{L(x|\theta)\pi(\theta)}{\int_{\Theta}(x|\theta)\pi(\theta)d\theta} \tag{51}$$

นั่นคือ $L(x|\theta)$ หรือ $f(x|\theta)$ เป็น Likelihood Function และสัญลักษณ์ Θ แทนค่าเป็น Parameter Space ของ θ หรือ Support of $\pi(\theta)$ ซึ่งมีค่าเป็น

$$p(\theta|x) = \int_{\Theta} L(x|\theta)\pi(\theta)d\theta \tag{52}$$

ซึ่งเป็นค่า Normalizing Constant ของ Posterior Distribution ของ θ โดยที่ p(x) เป็น Marginal Probability Distribution ของ x สำหรับการอนุมานปัญหาของ p(x) ไม่มีรูปแบบแน่นอนของการอนุมานแบบ Bayesian ของค่า θ ที่เป็นพื้นฐานของการแจกแจงภายหลังของ θ, p(θ|x) โดยตัว Posterior เอง แล้วสามารถคำนวณค่าข้อมูลได้หลายอย่าง เช่น การค่าเฉลี่ย ค่ามัธยฐาน ฐานนิยม ความแปรปรวนและควอไทล์

การอนุมานแบบ Bayesian โดยสรุปจะหาค่าได้จากการให้ข้อมูล ดังนี้

ข้อมูล $x = (x_1, x_2, \dots, x_n)^T$

ค่าพารามิเตอร์ $\theta = (\theta_1, \theta_2, \dots, \theta_p)^T$

Likelihood $L(x|\theta)$

Prior $\pi(\theta)$

โดยการอนุมานอยู่บนพื้นฐานของ Join Posterior ดังนั้น

$$p(\theta|x) = \frac{L(x|\theta)\pi(\theta)}{\int_{\Theta} L(x|\theta)\pi(\theta)d\theta} \tag{53}$$

ถ้าไม่สนใจค่า x ซึ่งเป็นค่าที่ไม่ขึ้นอยู่กับ θ โดยให้ x เป็นค่าคงที่ จึงเขียนอสมการ ได้เป็นความน่าจะเป็นภายหลังแปรผันตรงกับความควรจะเป็นคูณความน่าจะเป็นก่อนหน้า หรือ $p(\theta|x)\ \alpha\ f(\theta|x)\pi(\theta)$ หรือ Posterior α Likelihood x Prior

ถ้าไม่แน่ใจความเหมาะสมของค่า η สามารถกำหนดให้อยู่ในรูปการแจกแจงก่อนที่ 2 (Second - stage prior distribution หรือ Hyperparameter) โดยเขียนเป็นสมการดังกล่าวด้วย h(η) ซึ่งมีการแจกแจงภายหลังของ θ เป็น

$$p(\theta|x) = \frac{p(x,\theta)}{p(x)} = \frac{\int f(x,\theta,\eta)d\eta}{\iint f(x,\theta,\eta)d\eta d\theta}$$

$$= \frac{\int f(x|\theta)\pi(\theta|\eta)h(\eta)d\eta}{\iint f(x|\theta)\pi(\theta|\eta)h(\eta)d\eta d\theta} \tag{54}$$

สมมติมีเหตุการณ์ 2 เหตุการณ์ คือ A และ B จะได้เงื่อนไขความน่าเป็น คือ

$$Pr(A|B) = \frac{\Pr[A \cap B]}{\Pr[B]} = \frac{\Pr[B|A]\Pr[A]}{\Pr[B]} \quad (55)$$

Markov Chain Monte Carlo

ปี ค.ศ. 1990 กลุ่มนักสถิติคิดค้นวิธีการของ MCMC ซึ่งวิธีการนี้ทำให้การประมาณด้วยวิธี ของเบส์ที่ซับซ้อนสามารถแก้ปัญหาได้ง่ายและสะดวกขึ้น จนเป็นที่นิยมใช้ในปัจจุบัน ทำให้เกิดการพัฒนาวิธีการของ MCMC ในรูปของโปรแกรมเพื่อแก้ปัญหาเมื่อตัวแปรมีความสัมพันธ์กับรูปแบบของการแจกแจงหลายขั้นตอน (Hierarchical model) (อัชฌา อระวีพร, 2554)

การจำลองแบบมอนติคาร์โล (Monte Carlo Simulation) เป็นการใช้ตัวแปรสุ่มแบบ Uniform variable จาก 0 ถึง 1 หรือ U(0, 1) เพื่อแก้ปัญหา การจำลองแบบนี้นิยมใช้ในการประมาณค่าของการหา Integration ซึ่งมีกระบวนการดังนี้ (วุฒิชัย วงษ์ทัศนีย์กร, 2555)

$$I = \int_a^b h(x) \quad (56)$$

โดยให้ h(x) เป็นฟังก์ชันของจำนวนจริงที่ไม่สามารถหา Integral ได้ สามารถทำได้โดยประยุกต์ใช้ตัวแปรสุ่ม Y = (b - a)h(x) โดยให้ X แทนตัวแปรสุ่มแบบ Uniform ต่อเนื่องจาก a ถึง b หรือ U(a, b) ดังนั้น Expected Value ของ Y มีค่าเท่ากับ Integral I ดังนี้

$$\begin{aligned} E[Y] &= E[(b-a)h(x)] \\ &= (b-a)E[h(x)] \\ &= (b-a)\int_a^b h(x)f_x(x)dx \\ &= (b-a)\int_a^b \frac{h(x)dx}{(b-a)} = \int_a^b h(x)dx = I \end{aligned} \quad (57)$$

เมื่อ $f_x(x) = 1/(b-a)$ แทนฟังก์ชันการแจกแจงความน่าจะเป็นของตัวแปรสุ่ม U(a, b) ซึ่งหมายความว่าตัวประมาณค่าของค่าคาดคะเน E[Y] สามารถนำมาใช้ในการประมาณค่าของ Integral ได้โดยใช้ค่าเฉลี่ยตัวอย่างในการประมาณดังนี้

$$\bar{Y}(n) = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{\sum_{i=1}^{n} (b-a)h(X_i)}{n} = (b-a)\frac{\sum_{i=1}^{n} h(a+(b-a)Z_i)}{n} \qquad (58)$$

โดยให้ $X_i = X_1, X_2, X_3, \ldots, X_n$ เป็นตัวแปรสุ่ม *U(a,b)* และ $Z_i = Z_1, Z_2, Z_3, \ldots, Z_n$ เป็นตัวแปรสุ่ม U(0, 1)

นอกไปจากนี้ Monte Carlo Simulation นิยมใช้ในงานวิจัยทางด้านสถิติ เช่น การหาค่าวิกฤตของการทดสอบสมมติฐาน การหาค่าวิกฤตของวิธีการทดสอบตัวแปรสุ่มแบบปกติ และการทดสอบ Kolmogorov - Smirnov (K - S test) เป็นต้น

Gibbs Sampler

Gibbs Sampler หรือ Gibbs Sampling เป็นส่วนหนึ่งของ MCMC Class หลักการของ Gibbs Sampling คือ การจำลองตัวแปรสุ่มจากการแจกแจงแบบมีเงื่อนไข ซึ่งอธิบาย ได้ดังนี้

กำหนดให้ $\theta = [\theta_1\theta_2]$ ซึ่งมี Posterior Density $p(\theta) = p(\theta_1, \theta_2)$ โดยที่ไม่ต้องใช้ตัวแปรตามของ y ได้ ถ้ารู้เงื่อนไข Conditional Density ซึ่งไม่ต้องรับรองในเงื่อนไขมากนัก แต่รู้ได้จาก $p(\theta_1|\theta_2)$ และ $p(\theta_2|\theta_1)$ 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

$$f_1(\theta_1|\theta_2\theta_3, y)$$
$$f_2(\theta_2|\theta_3\theta_1, y)$$
$$f_3(\theta_3|\theta_1\theta_2, y) \quad \text{โดยที่ } y = (y_1, y_2, \ldots, y_n)^T$$

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 Posterior Distribution จะได้อันดับแบบ General Form แบบ i - th เป็น $\{\theta_{1,i}, \theta_{2,i}, \theta_{3,i}\}$ ดังนั้น จะได้

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

3. ทิ้งตัวแรก M ที่ได้จากการรวมในขั้นที่ 2 แล้วใช้ค่า N ที่เป็นตัวสุดท้ายเพื่อสร้างแบบของ Random Sample $\{(\theta_{1,i}, \theta_{2,i}, \theta_{3,i})\}_{i=M+1}^{M+N}$ และทำการประมาณค่า Posterior Marginal โดยใช้ Random Sample

ภายใต้เงื่อนไขอย่างง่าย Geman & Geman (1984) ได้แสดงว่า Join Distribution ของ Above Random Sample จะ Convergence Exponentially สู่ Join Posterior Distribution ของ (θ_1, θ_2, θ_3) ดังนั้น ก่อนที่จะ Realization Form ของ Random Sample จาก Join Distribution $f_1(\theta_1|\theta_2\theta_3, y)$ สามารถทำการอนุมานได้จาก

$$\widehat{\theta_\iota} = \frac{1}{N\ M}\sum_{i=M+1}^{N}\theta_{i,j} \tag{59}$$

$$\hat{\sigma}_i^2 = \frac{1}{N\ M\ 1}\sum_{i=M+1}^{N}(\theta_{i,j}, \hat{\theta}_i)^2 \tag{60}$$

โดยค่า Credible Interval {l, μ} ของ θ_i อยู่ภายใต้เงื่อนไข p(l θ_i μ) = 0.95 เป็นต้น เป็น Algorithm อีกแบบซึ่งเป็นการแทนที่ค่าที่เป็นไปได้ในแต่ละ State เพื่อทำการหาค่าที่ดีที่สุด ในที่นี้เป็นค่าของความเสี่ยงและ Probability Distribution Function (PDF) เพื่อเป็นการตัดสินว่าค่าที่ได้เป็นค่าที่ดีที่สุด นั่นคือ เป็นหนึ่งในวิธีการหาค่าที่ดีที่สุดของข้อมูล (Finding best solution)

การประมาณค่าพารามิเตอร์ด้วยวิธีของเบส์สำหรับทฤษฎีการตอบสนองข้อสอบ

การประมาณค่าพารามิเตอร์ด้วยวิธีของเบส์ในทฤษฎีการตอบสนองข้อสอบ มีแนวคิดว่าค่าความสามารถของผู้สอบ (θ) และค่าพารามิเตอร์ของข้อสอบ ได้แก่ ค่าความยากของข้อสอบ (b) ค่าอำนาจจำแนกของข้อสอบ (a) และค่าโอกาสในการเดาข้อสอบ (c) เป็นตัวแปรสุ่ม (Random variable) จากการแจกแจงที่แสดงได้ด้วยฟังก์ชันความหนาแน่นร่วม (Joint density function) f(θ, b, a, c) โดยเรียกฟังก์ชัน f(θ, b, a, c) นี้ว่า การแจกแจงเริ่ม (Prior distribution) ของค่า θ, b, a และ c ซึ่งทำให้การใช้ฟังก์ชัน Likelihood L(U|θ, b, a, c) เพียงอย่างเดียว ในการประมาณค่า θ, b, a และ c ถูกพิจารณาว่าเป็นการใช้ข้อมูลที่มีอยู่อย่างไม่ครบถ้วน เพราะยังมีการแจกแจงแรกร่วมกับ f(θ, b, a, c) ที่ควรนำมาใช้ในการประมาณค่าพารามิเตอร์ด้วย

จากการนิยามความน่าจะเป็นของผู้เข้าสอบคนที่ i ในการตอบข้อสอบข้อที่ j เมื่อ U_{ij} = 1 สำหรับการตอบถูก และ U_{ij} = 0 สำหรับการตอบผิด แสดงสมการโมเดลแบบ 3 พารามิเตอร์ของโมเดลการตอบสนองข้อสอบ (3PL IRT) ได้ดังนี้

$$P(U_j|\theta, b, a, c) = P(U_j = 1|\theta, b, a, c)P(U_j = 0|\theta, b, a, c)$$

$$= P_j^{U_j}.Q_j^{1-U_j} \quad \text{โดยที่ } \mathrm{Q}_j = 1 - \mathrm{P}_j \tag{61}$$

ถ้าข้อสอบจำนวน n ข้อและแต่ละข้อเป็นอิสะต่อกัน ความน่าจะเป็นของการตอบแสดงด้วยฟังก์ชันความหนาแน่นร่วม ดังสมการต่อไปนี้

$$P(U_1, U_2, \dots, U_n|\theta, b, a, c) = \prod_{j=1}^{n} P_j^{U_j} Q_j^{1-U_j} \quad (62)$$

จากสมการข้างต้น เป็นความน่าจะเป็นของการตอบข้อสอบ n ข้อ ที่วัดหรือสังเกตได้ โดยที่ $U_1, U_2, \dots, U_n$ เป็นตัวแปรสุ่มที่มีค่าเฉพาะเป็น $u_1, u_2, \dots, u_n$ เมื่อ u_{ij} มีค่า 1 หรือ 0 และเนื่องจากสมการนี้เป็นฟังก์ชันทางคณิตศาสตร์ของค่า θ, b, a, c ที่จะบอกว่าตัวแปรสุ่มนี้ มีโอกาสเกิดขึ้นเพียงใด จึงเรียกฟังก์ชันนี้ว่า Likelihood Function ซึ่งแสดงได้ดังนี้

$$L(u_1, u_2, \dots, u_n|\theta, b, a, c) = \prod_{j=1}^{n} P_j^{u_j} Q_j^{1-u_j} \quad (63)$$

เมื่อ $u_{ij} = 1$ ค่าของ Q_1 จะหมดไป และเมื่อ $u_{ij} = 0$ ค่าของ P_1 จะหมดไปและ Likelihood Function ที่มีผู้สอบ N คน ตอบข้อสอบ n ข้อ มีสมการดังนี้

$$L(U|\theta, b, a, c) = \prod_{i=1}^{N} \prod_{j=1}^{n} P_{ij}^{u_{ij}} Q_{ij}^{1-u_{ij}} \quad (64)$$

เมื่อ u = เวคเตอร์ผลการตอบข้อสอบ n ข้อ ของผู้เข้าสอบ N คน

$$P_{ij} = P_i(\theta_j, b_j, a_j, c_j)$$

ถ้าพิจารณาความน่าจะเป็นร่วมของการตอบข้อสอบ P(U|θ, b, a, c) จะเห็นว่าการแจกแจงของตัวแปร U ขึ้นอยู่กับค่าความสามารถ θ และค่าพารามิเตอร์ของข้อสอบ b, a และ c ซึ่งถ้าค่า θ, b, a และ c เปลี่ยนไป โอกาสที่ U ที่มีค่าเท่ากับ u ก็จะเปลี่ยนไปด้วย ดังนั้น การทราบผลการตอบข้อสอบ u จึงน่าจะช่วยทำให้ทราบค่า θ, b, a และ c ได้ดียิ่งขึ้น แสดงได้จากการแจกแจงอย่างมีเงื่อนไขของค่า θ, b, a และ c เมื่อทราบผลการตอบข้อสอบ f(θ, b, a, c|u) และเรียกฟังก์ชันนี้ว่า การแจกแจงหลัง (Posterior Distribution)

การแจกแจงหลังร่วมกันของค่าความสามารถ θ และค่าพารามิเตอร์ของข้อสอบ b, a และ c เมื่อทราบผลการตอบข้อสอบ u ก็คือ

$$f(\theta, b, a, c|U) = L(U|\theta, b, a, c) f(\theta, b, a, c) / f(u) \quad (65)$$

เมื่อ f(u) = การแจกแจงมาร์จินัล (Marginal) ของผลการตอบข้อสอบ

f(θ, b, a, c) = การแจกแจงเริ่มของค่า θ, b, a และ c

f(U|θ, b, a, c) = Likelihood Function

f(θ, b, a, c|U) = การแจกแจงหลังของค่า θ, b, a และ c เมื่อทราบผลการตอบข้อสอบ u

ซึ่งการแจกแจงหลัง เป็นฟังก์ชันที่ใช้ประมาณค่า θ, b, a และ c ด้วยวิธีของเบส์ โดยมีส่วนแตกต่างจากการประมาณด้วยวิธีแมกซิมัมไลค์ลิฮู้ด คือ การแจกแจงเริ่ม f(θ, b, a, c) และการแจกแจงมาร์จินัล f(u) โดยการแจกแจงมาร์จินัลนี้เป็นการแจกแจงที่ไม่ขึ้นอยู่กับค่า θ, b, a และ c จึงถือว่าเป็นค่าคงที่ในการประมาณค่า θ, b, a และ c

จากแนวคิดของการประมาณค่าด้วยวิธีของเบส์ดังกล่าว ทำให้สามารถจำแนกกระบวนการดำเนินการตามแนวคิดของ Bayes ออกเป็น 2 กระบวนการ ดังนี้ (Swaminathan & Gifford, 1985; Swaminathan & Gifford, 1986)

1. กระบวนการกำหนดลักษณะของการแจกแจงเริ่ม (Prior distribution) คือ

1.1 กำหนดให้การแจกแจงเริ่มของค่าความสามารถของผู้สอบ (θ) ค่าความยากของข้อสอบ (b) ค่าอำนาจจำแนกของข้อสอบ (a) และค่าโอกาสในการเดาข้อสอบ (c) เป็นอิสระต่อกัน ดังนี้

$$f(\theta, b, a, c) = f(\theta).f(b).f(a).f(c) \quad (66)$$

โดยกำหนดให้การแจกแจงของ f(θ), f(b), f(a), f(c) เป็นดังนี้

1.1.1 การแจกแจงเริ่มของค่าความสามารถ f(θ) มีข้อตกลงว่าข้อสารสนเทศที่มีมาก่อนของค่าความสามารถของผู้สอบแต่ละคนไม่แตกต่างกัน สามารถใช้แทนกันได้ (Exchangeability) และค่าความสามารถเป็นตัวแปรสุ่มที่มีการแจกแจงเป็นปกติ (Normal distribution) หรือ $f(\theta_i|\mu_\theta, \sigma_\theta^2) = N(\mu_\theta, \sigma_\theta^2)$ เมื่อ $N(\mu_\theta, \sigma_\theta^2)$ หมายถึง การแจกแจงปกติ (Normal distribution) ที่มีค่าเฉลี่ยเท่ากับ μ_θ และค่าความแปรปรวนเท่ากับ σ_θ^2

1.1.2 การแจกแจงเริ่มของค่าความยากของข้อสอบ f(b) อาจใช้กระบวนการเดียวกับการกำหนดการแจกแจงเริ่มแรกของค่าความสามารถ คือ มีข้อตกลงว่า f(b) มีการแจกแจงเป็นปกติ หรืออาจไม่กำหนดการแจกแจงเริ่มแรกไว้ก็ได้

1.1.3 การแจกแจงเริ่มของค่าอำนาจจำแนกของข้อสอบ f(a) ควรเป็นการแจกแจงแบบ Chi-square เนื่องจากค่าอำนาจจำแนกของข้อสอบโดยทั่วไปจะเป็นค่าบวกและเป็นความชันของเส้นโค้งลักษณะของข้อสอบ ณ จุดเปลี่ยนโค้ง

$$f(a_j|v_j w_j) \;\; \alpha \;\; a_j^{v_j-1}\exp[-a_j^2/2w_j]$$

เมื่อ v_j = Degree of Freedom

w_j = Scale Parameter

1.1.4 การแจกแจงเริ่มของค่าการเดาของข้อสอบ f(c) ควรมีการแจกแจงแบบเบต้า (Beta distribution) เนื่องจากค่าพารามิเตอร์ c_j มีขอบเขต [0, 1]

$$f(c_j|s_j t_j) \;\; \alpha \;\; c_j^{a_j}(1-c_j)^{t_j}$$ เมื่อ $s_j t_j$ = Scale Parameter

1.2 กำหนดค่าที่เป็นตัวเลขของพารามิเตอร์สำหรับการแจกแจงเริ่ม

1.2.1 พารามิเตอร์ของการแจกแจงเริ่มของค่าความสามารถ θ ได้แก่ μ_θ และ σ_θ^2 อาจกำหนดให้ μ_θ = 0 และ σ_θ^2 = 1 ซึ่งจะทำให้การประมาณค่าความสามารถ θ มีความสะดวกและประมาณค่าได้รวดเร็วขึ้น

1.2.2 พารามิเตอร์ของการแจกแจงเริ่มของค่าความยาก b หากไม่มีการกำหนดลักษณะการแจกแจงไว้ ก็ใช้ค่าเดียวกับพารามิเตอร์ของการแจกแจงเริ่มของค่าความสามารถ

1.2.3 พารามิเตอร์ของการแจกแจงเริ่มของค่าอำนาจจำแนก a ได้แก่ v_j และ w_j การกำหนดค่าของ v_j และ w_j ที่เหมาะสม อาจจะหาได้จากการกำหนดพิสัย (Range) ของค่า a คือ ถ้าให้ H เป็นขีดจำกัดบนของพิสัยและ L เป็นขีดจำกัดล่างของพิสัย จะหาค่า v_j และ w_j จากสูตร

$$v_j = \frac{1}{2}(1 + z_{\frac{1}{2}}((H+L)/(H-L))^2) \tag{67}$$

$$w_j = \frac{1}{2}((H-L)/Z_{(1/2)\alpha})^2 \tag{68}$$

เมื่อ $Z_{(1/2)\alpha}$ = ค่า Z ของการแจกแจงปกติมาตรฐาน (Standard normal distribution) ที่ระดับนัยสำคัญ α

v_j = Degree of Freedom

นอกจากวิธีดังกล่าว อาจกำหนดให้ v_j และ w_j ของการแจกแจงเริ่ม f(a) ของข้อสอบทุกข้อเท่ากัน คือ $v_j = 10$ และ $w_j = 0.1$ ซึ่งจะทำให้การประมาณค่าพารามิเตอร์มีความสะดวกยิ่งขึ้น และการกำหนดเช่นนี้ ก็ยังคงทำให้ค่าประมาณของค่าพารามิเตอร์ a_j ตกอยู่ในช่วงที่ยอมรับได้ คือ $0.40 < a_j < 1.55$ ด้วยความเชื่อมั่น 99%

1.2.4 พารามิเตอร์ของการแจกแจงเริ่มของค่าการเดา c ได้แก่ s_j และ t_j โดยกำหนด s_j และ t_j ที่เหมาะสมได้จากการสังเกตสัดส่วนการตอบถูกของกลุ่มผู้สอบที่มีความสามารถในระดับต่ำมาก กล่าวคือ ถ้าให้ m แทนจำนวนผู้สอบที่มีระดับความสามารถต่ำมาก และ M แทนสัดส่วนการตอบถูกของผู้สอบกลุ่ม m แล้ว สามารถหาค่า s_j และ t_j ได้จาก $s_j = mM$ และ $t_j = m(1 - M) - 2$ หรืออาจจะกำหนดให้ s_j และ t_j ของการแจกแจงเริ่ม f(c) ของข้อสอบทุกข้อเท่ากัน คือ $s_j = 2$ และ $t_j = 12$ ซึ่งจะทำให้การประมาณค่าพารามิเตอร์มีความสะดวกยิ่งขึ้น และการกำหนดเช่นนี้ ก็ยังคงทำให้ค่าประมาณของค่าพารามิเตอร์ c_j ตกอยู่ในช่วงที่ยอมรับได้ คือ $0.026 < c_j < 0.317$ ด้วยความเชื่อมั่น 99 %

2. กระบวนการประมาณค่าพารามิเตอร์ของข้อสอบและความสามารถของผู้สอบ

การประมาณค่าพารามิเตอร์ด้วยวิธีของเบส์ คือ การหาค่าประมาณ θ_i โดยที่ $i = 1, 2, \ldots, N$ และ b_j, a_j, c_j โดยที่ $j = 1, 2, \ldots, n$ ที่ทำให้ฟังก์ชันการแจกแจงหลัง $f(\theta, b, a, c|u)$ มีค่าสูงสุด ถ้า $\ln f(\theta, b, a, c|u)$ เป็นฟังก์ชันที่หาอนุพันธ์ได้ การหาค่า θ_i, b_j, a_j และ c_j จะหาได้จากการอนุพันธ์ของ $\ln f(\theta, b, a, c|u)$ และกำหนดให้อนุพันธ์ของ $\ln f(\theta, b, a, c|u)$ มีค่าเท่ากับศูนย์ แล้วจึงหาค่ารากของอนุพันธ์ของ $\ln f(\theta, b, a, c|u)$ จาก

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

$$\ln f(\theta, b, a, c|u) = \ln L(u|\theta, b, a, c) + \ln f(\theta) + \ln f(b) + \ln f(a) + \ln f(c) + \text{ค่าคงที่}$$

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

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

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

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

สมการอนุพันธ์ของ ln f(θ, b, a, c|u) = 0 เรียกว่า Modal Equation ค่ารากของ Modal Equation คือ ค่าความสามารถ θ และค่าพารามิเตอร์ของข้อสอบ b, a, c ที่ทำให้ฟังก์ชันการแจกแจงหลังมีค่าสูงสุด อาจทำได้โดยใช้เทคนิค Newton-Raphson ซึ่งเป็นการหาค่าประมาณ โดยการหาค่าซ้ำ (Iterative) มีขั้นตอนดังต่อไปนี้

ขั้นที่ 1 กำหนดค่าเริ่มต้นสำหรับใช้ในการประมาณค่าความสามารถ θ_i และค่าพารามิเตอร์ข้อสอบ b_j, a_j และ c_j ดังนี้

$$\theta_i^{(0)} = ln\left(\frac{X}{n-X_i}\right)$$
$$a_i^{(0)} = R_j/(1-R_j^2)^{1/2}$$
$$b_i^{(0)} = Z_j/R_j$$
$$c_i^{(0)} = 1/m_j$$

เมื่อ ln = Natural Logarithm

X_i = คะแนนสอบของผู้เข้าสอบคนที่ i

n = จำนวนข้อสอบ

R_j = Point - Biserial Correlation

Z_j = ค่า Z ของการแจกแจงปกติมาตรฐานที่พื้นที่ใต้โค้งปกติมาตรฐานด้านขวามือ มีค่าเท่ากับ P_j (U_j / N)

N = จำนวนผู้เข้าสอบทั้งหมด

m_j = จำนวนตัวเลือกในข้อสอบข้อที่ j

ขั้นที่ 2 ประมาณค่าความสามารถของผู้สอบแต่ละคน θ_i โดยกำหนดให้ค่าพารามิเตอร์ของข้อสอบที่ประมาณค่าได้ในครั้งก่อนเป็นค่าคงที่

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

เมื่อ $\theta_i^{m+1}, \theta_i^m$ = ค่าประมาณความสามารถของคนที่ i ครั้งที่ m + 1 และ m

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

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

การประมาณค่าซ้ำ (Iterative) จะกระทำจนกว่าค่าประมาณความสามารถ θ_i จะลู่เข้าค่าคงที่ค่าใดค่าหนึ่ง (Convergence) คือ ค่าประมาณครั้งที่ m + 1 และครั้งที่ m มีค่าแตกต่างกันน้อยกว่าค่าคงที่ ที่กำหนดไว้ เช่น 0.0005

ขั้นที่ 3 ประมาณค่าพารามิเตอร์ของข้อสอบแต่ละข้อ b_j, a_j และ c_j โดยกำหนดให้ค่าความสามารถของผู้เข้าสอบที่ประมาณค่าได้ในครั้งก่อนเป็นค่าคงที่

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

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

a_j^{m+1}, a_j^m = ค่าประมาณอำนาจจำแนกของข้อที่ j ครั้งที่ m + 1 และ m

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

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

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

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

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

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

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

การประมาณค่าซ้ำ (Iterative) จะทำจนกว่าค่าประมาณจะลู่เข้าสู่ค่าคงที่ค่าใดค่าหนึ่ง (Convergence) สำหรับค่าของอนุพันธ์อันดับ 1 g(x) และอนุพันธ์อันดับ 2 h(x) ประกอบด้วย 2 ส่วน คือ ส่วนที่ได้จากฟังก์ชันไลค์ลิฮู้ดและส่วนที่ได้จากการแจกแจงเริ่ม

ขั้นที่ 4 ประมาณค่าซ้ำ ขั้นที่ 2 และขั้นที่ 3 จนกว่าค่าประมาณ θ_i , b_j, a_j และ c_j มีค่าคงที่และถูกต้องเพียงพอ หรือ $f(\theta, b, a, c|u)$ มีค่าสูงสุด

โดยสรุปขั้นตอนการประมาณค่าคือ กำหนดข้อมูลเริ่มต้นที่เป็นไปได้ของตัวแปรสุ่ม จากนั้น นำข้อมูลเริ่มต้นไปจำลอง ครั้งที่ 1 แล้วนำข้อมูลของการจำลองครั้งที่ 1 ไปจำลอง ครั้งที่ 2 เวียนซ้ำเช่นนี้ไปเรื่อย ๆ จะได้ลำดับ n ครั้ง เนื่องจากการจำลองครั้งที่ n+1 จะใช้ข้อมูลครั้งที่ n เท่านั้น ไม่ขึ้นอยู่กับ n-1, n-2, ... , 1 แสดงว่า ข้อมูลที่ได้จากการจำลองครั้งถัดไปขึ้นอยู่กับข้อมูลที่จำลองได้ปัจจุบัน กระบวนการนี้สอดคล้องกับ Stochastic Process ที่เรียกว่า ลูกโซ่มาร์คอฟ

(Markov Chain) ดังนั้นจะได้ว่า ลำดับที่ n มีการแจกแจงเข้าสู่ f(x) ด้วยความคาดเคลื่อนน้อย สำหรับ n ที่มีค่าใหญ่ ในทางปฏิบัติ n ค่าใหญ่นั้น จะเริ่มด้วยค่าใหญ่ตั้งแต่การจำลองครั้งที่ n + 1 เช่น เริ่มที่ 1001 (n = 1000) เป็นต้น โดยไม่ใช้ค่าลำดับที่ 1 - 1000 (จำลองทิ้ง n ค่าแรก) หรือเรียกว่า การ Burn

การศึกษา Monte Carlo ในทฤษฎีการตอบสนองข้อสอบ

การวิเคราะห์ข้อสอบด้วยทฤษฎีการตอบสนองข้อสอบ มีการประยุกต์ใช้เทคนิค Monte Carlo (MC) ในการวิเคราะห์เป็นจำนวนมาก โดยในช่วงปี ค.ศ. 1994 - 1995 ประมาณร้อยละ 25 - 33 ของบทความในวารสาร Applied Psychological Measurement (APM) Psychometrika และ Journal of Educational Measurement (JEM) ใช้เทคนิค MC ในการวิเคราะห์ข้อมูล โดยเฉพาะการประเมินประสิทธิภาพการประมาณค่าพารามิเตอร์ นอกจากนี้ยังมีการประยุกต์ใช้การเปรียบเทียบวิธีการต่าง ๆ ในโมเดลการวิเคราะห์ข้อสอบ ไม่ว่าจะเป็นการทำหน้าที่ต่างกันของข้อสอบหรือการประเมินผลที่ได้จากการวัดหลายมิติ เป็นต้น (Harwell et al., 1996) โดยการนำเทคนิค MC มาใช้ในการวิเคราะห์ IRT มีขั้นตอนดังนี้

1. กำหนดคำถามการวิจัยที่อธิบายถึงวัตถุประสงค์ที่เฉพาะเจาะจง เช่น เพื่อศึกษาความถูกต้องของการประมาณค่าพารามิเตอร์ของโมเดลการวิเคราะห์ข้อสอบ เมื่อมีจำนวนข้อสอบ ผู้สอบ และการแจกแจงเริ่มต้นของพารามิเตอร์ข้อสอบแตกต่างกัน เป็นต้น
2. กำหนดเงื่อนไขของตัวแปรต้นที่ส่งผลต่อตัวแปรตามได้ เช่น จำนวนผู้สอบและข้อสอบ (ตัวแปรต้น) มีผลกระทบต่อค่าพารามิเตอร์ของผู้สอบ (ตัวแปรตาม) หรือไม่
3. ออกแบบการทดลองให้มีความเหมาะสมกับวัตถุประสงค์การวิจัย
4. จำลองข้อมูลให้สอดคล้องกับเงื่อนไขของโมเดลการวิเคราะห์ข้อสอบ เช่น โมเดลแบบ 2 พารามิเตอร์
5. ประมาณค่าพารามิเตอร์โดยใช้ข้อมูลจากการจำลอง
6. เปรียบเทียบผลที่ได้จากการประมาณค่าตามเงื่อนไข โดยใช้ค่าสถิติต่าง ๆ ได้ เช่น ค่ามัธยฐาน ค่าความคลาดเคลื่อนมาตรฐานของการประมาณค่า เป็นต้น
7. กำหนดจำนวนการทำซ้ำ R รอบ
8. คำนวนค่าสถิติที่ได้จากการวิเคราะห์ R รอบ ทั้งสถิติเชิงบรรยายและสถิติเชิงอ้างอิง ซึ่งจะอ้างไปถึงการตอบคำถามการวิจัยและออกแบบการทดลอง

ขั้นตอนการวิเคราะห์ข้อสอบด้วยเทคนิค MC

Naylor, Balintfy, Burdick, & Chu (1968 cited in Harwell et al., 1996) อธิบายขั้นตอนการวิเคราะห์ข้อสอบแบบ IRT ด้วยเทคนิค MC 4 ขั้นตอน ได้แก่ 1) การกำหนดปัญหาการวิจัย

2) ออกแบบการทดลอง ซึ่งรวมถึงการระบุตัวแปรต้นและตัวแปรตาม การออกแบบการวิจัยเชิงทดลอง จำนวนรอบในการคำนวณ และการเลือกโมเดลการวิเคราะห์ข้อสอบ 3) เขียนและระบุโปรแกรมในการจำลองข้อมูลและการประมาณค่าพารามิเตอร์ และ 4) วิเคราะห์ผลจากการจำลองข้อมูล โดยในแต่ละขั้นตอนมีรายละเอียดดังนี้

1. การกำหนดปัญหาการวิจัย ขั้นตอนนี้นับเป็นขั้นตอนที่สำคัญของกระบวนการวิจัย การวิเคราะห์ข้อสอบด้วยเทคนิค MC ก็เช่นเดียวกัน โดยเริ่มจากการกำหนดปัญหาและข้อคำถามการวิจัยก่อน จากนั้นตั้งสมมติฐานการทดสอบ และมีการวัดผลกระทบจากเงื่อนไขต่าง ๆ จากการจำลองข้อมูล โดยทั่วไปการกำหนดปัญหาการวิจัยมักมาจากการทบทวนเอกสารและรายงานการวิจัยที่เกี่ยวข้องกับเรื่องที่สนใจศึกษา มีการทดสอบสมมติฐานที่เป็นตัวแทนของคำถามการวิจัยและผลที่เกิดจากการวัดต้องไวกับตัวแปรที่ศึกษา

2. ออกแบบการศึกษาด้วยเทคนิค MC ต้องออกแบบให้สามารถตอบข้อคำถามของการวิจัยและสมมติฐานการวิจัยได้ โดยต้องมีการออกแบบทั้งตัวแปรต้นหรือตัวเป็นที่เป็นสาเหตุ และ ตัวแปรตามซึ่งเป็นผลกระทบที่เกิดขึ้นจากตัวแปรต้น นอกจากนี้ ยังต้องมีการประเมินผลทั้งในเรื่องของความตรงภายในและความตรงภายนอก

ประเด็นที่เกี่ยวข้องกับการออกแบบของการศึกษา MC รวมถึงการเลือกตัวแปรต้นและตัวแปรตามขึ้นอยู่กับการออกแบบการทดลอง จำนวนการทำซ้ำ และโมเดลของ IRT เพื่อให้ผลการทดลองสามารถอ้างไปสู่ประชากรได้ ซึ่งการออกแบบการศึกษาด้วยด้วยเทคนิค MC มีขั้นตอนดังนี้

2.1 การกำหนดและระบุค่าของตัวแปรต้น คำถามการวิจัยเป็นสิ่งที่กำหนดตัวแปรต้นรวมทั้งเงื่อนไขในการจำลองข้อมูล โดยค่าของตัวแปรถือเป็นค่าคงที่ (Fixed Effect) และไม่ต่อเนื่อง (Discrete) เช่น การศึกษาของ Harwell & Janosky (1991) กำหนดตัวแปรต้น ได้แก่ ขนาดตัวอย่าง (n) ความยาวข้อสอบ (L) ความแปรปรวนของกระจายก่อนหน้า (Prior distribution) ของ a_j หรือความแปรปรวนของตัวแปรต้น โดยที่ค่าของตัวแปรเหล่านี้เกิดจากคำถามวิจัย ซึ่งเน้นไปที่ขนาดตัวอย่างเล็กและความยาวข้อสอบน้อย

พารามิเตอร์เหล่านี้มักแสดงเป็นค่าระยะห่างเท่า ๆ กันในช่วงคงที่หรือเป็นค่าประมาณการจากการทดสอบเทียบกับการทดสอบก่อนหน้า ซึ่งถือเป็นค่าของตัวแปรคงที่ (Fixed Effect) การสุ่มตัวอย่างของค่าอำนาจจำแนกและค่าความยากเป็นค่าของตัวแปรสุ่ม (Random Effect) เนื่องจากถ้ามีการสุ่มค่าอำนาจจำแนกและค่าความยากมาใช้ในการศึกษาจะทำให้โมเดลในการศึกษาเป็นโมเดลแบบสุ่ม ซึ่งสามารถอ้างอิงไปยังประชากร แต่ถ้าโมเดลไม่ได้มีการสุ่มขึ้นมาใช้ในการศึกษา โมเดลนั้นจะไม่สามารถอ้างอิงกลับไปยังประชากรได้ ดังนั้น ค่าพารามิเตอร์

ในโมเดลควรเป็นตัวแทนของตัวแปรต้น อย่างไรก็ตาม นักวิจัยต้องพิจารณาความสัมพันธ์ระหว่างจำนวนของตัวแปรต้น ประสิทธิภาพของการศึกษาและผลการศึกษาด้วย ในขณะที่จำนวนของตัวแปรเพิ่มขึ้น จะทำให้ความรู้ที่ได้มากขึ้น แต่ก็จะใช้เวลาเพื่อการจำลองมากขึ้นตามไปด้วย

2.2 การเลือกแบบการทดลอง โดยทั่วไปตัวแปรอิสระมักจะเป็นตัวกำหนดแบบการทดลองที่เหมาะสม เช่น ถ้าจำนวนตัวแปรอิสระและระดับค่าของตัวแปรน้อย การใช้แบบการทดลองแบบแฟกตอเรียลจะมีความเหมาะสมกว่าแบบการทดลองอื่น ในการศึกษาด้วยเทคนิค MC มักจะใช้แบบการทดลอง โดยยึดเป้าหมายและวัตถุประสงค์ในการศึกษาเป็นหลัก ซึ่งการเลือกแบบทดลองอย่างระมัดระวัง จะช่วยในการวางแผนการวิเคราะห์ผลลัพธ์ได้อย่างถูกต้อง (Lewis, & Ovar, 1989 cited in Harwell et al., 1996) ในงานวิจัยของ Harwell & Jamoskey (1991) มีตัวแปรจัดกระทำหรือตัวแปรต้นเป็น ขนาดกลุ่มตัวอย่าง ความยาวของแบบสอบ และความแปรปรวนของการแจกแจงค่าอำนาจจำแนก ใช้การออกแบบการทดลองแบบแฟกตอเรียลระหว่างกลุ่มตัวอย่างแบบสมบูรณ์ (Completely between - subjects factorial design) นอกจากนี้งานวิจัยของ Yen (1987 cited in Harwel et al., 1996) ได้เปรียบเทียบการใช้โปรแกรมการวิเคราะห์ข้อสอบระหว่างโปแกรม BILOG และโปรแกรม LOGIST โดยมีเงื่อนไขด้านความยาว ของข้อสอบและลักษณะการกระจายของค่าพารามิเตอร์ของผู้สอบ ออกแบบการทดลองแบบแฟกตอเรียล โดยเรื่องของความยาวของข้อสอบและลักษณะการกระจายของค่าพารามิเตอร์ของผู้สอบได้ออกแบบการทดลองแฟกตอเรียลแบบ between - subjects factor และส่วนการเปรียบเทียบระหว่างโปรแกรมคอมพิวเตอร์ทั้งสองได้ออกแบบการทดลองแฟกตอเรียลแบบ within - subjects factor

2.3 การเลือกตัวแปรตาม ไม่เพียงต้องสอดคล้องกับคำถามการวิจัย แต่ต้องเลือกตัวแปรตามที่มีความไวต่อตัวแปรต้น เนื่องจากการเลือกตัวแปรที่มีความไวและควรใช้ประโยชน์ได้ ถ้ามีการแปลงข้อมูลเป็นรูปแบบอื่น เช่น การหาค่า RMSE สามารถแปลงค่าเพื่อให้มีการแจกแจงแบบปกติ ทำให้สามารถนำไปสรุปอ้างอิงได้ นอกจากนี้ ถ้าเป็นเรื่องเกี่ยวกับการศึกษาเปรียบเทียบวิธีการในการศึกษา IRT สามารถใช้คุณลักษณะของแบบสอบ เช่น ความเป็นเอกมิติ การทำหน้าที่ต่างกันของข้อสอบหรือผู้สอบ เป็นตัวแปรตามในการศึกษาถึงผลกระทบของตัวแปรอิสระได้ สำหรับค่าความสัมพันธ์ของค่าจริงกับค่าที่ประมาณได้ ก็สามารถใช้ให้เป็นตัวแปรตามในการใช้เทคนิคมอนติคาร์โล เนื่องจากค่าสัมพันธ์นั้น ใช้เมตริกที่ต่างกันหาความสัมพันธ์ของตัวแปรอิสระและตัวแปรตามได้ เช่น ค่าความสัมพันธ์ระหว่างค่าพารามิเตอร์ที่ประมาณกับความคลาดเคลื่อนมาตรฐาน ส่วนข้อเสียก็คือ ความสัมพันธ์เหล่านี้สะท้อนความสัมพันธ์เฉพาะอันดับของตัวแปรและแสดงอิทธิพลของตัวแปรต้นเท่านั้น เช่น ค่าความสัมพันธ์ระหว่างค่าอำนาจจำแนกที่แท้จริงกับค่า

ที่ประมาณ มีค่าเท่ากับ 0.9 หมายความว่า โดยค่าเฉลี่ยของค่าอำนาจจำแนกที่แท้จริงนั้น อาจสูงกว่าค่าเฉลี่ยของค่าอำนาจจำแนกที่ประมาณได้ แต่ไม่รับรองว่าค่าอำนาจจำแนกที่แท้จริงกับค่าอำนาจจำแนกที่ประมาณจะใกล้เคียงกันหรือดีกว่ามากน้อยเพียงใด เช่น 0.8 กับ 0.9

2.4 การกำหนดจำนวนรอบ สำหรับการศึกษาด้วยเทคนิค MC เปรียบเทียบได้กับการกำหนดขนาดกลุ่มตัวอย่าง โดยมีเกณฑ์ที่ใช้ในการกำหนดประยุกต์ใช้มาจากการกำหนดขนาดกลุ่มตัวอย่างสำหรับในการศึกษาจากข้อมูลเชิงประจักษ์ ในการศึกษาการวิเคราะห์ข้อสอบด้วยทฤษฎีการตอบสนองข้อสอบ จำนวนรอบขึ้นอยู่กับวัตถุประสงค์ในการศึกษา โดยพิจารณาจากความต้องการในลดค่าความแปรปรวนของการสุ่มตัวอย่างในการประมาณค่าพารามิเตอร์ และความต้องของการทดสอบสถิติของผลการจำลองข้อมูลว่าอำนาจในการตรวจสอบผลกระทบที่สนใจเพียงพอหรือไม่

จำนวนรอบมีอิทธิพลโดยตรงกับความแม่นยำในการประมาณค่าพารามิเตอร์ ถ้ากลุ่มตัวอย่างขนาดใหญ่ (มีจำนวนรอบมาก) จะให้การประมาณค่าพารามิเตอร์ด้วยความแปรปรวนของการสุ่มตัวอย่างน้อย ดังนั้น ถ้านักวิจัยไม่กำหนดจำนวนรอบหรือกำหนดจำนวนรอบน้อยจะทำให้ความแปรปรวนของการสุ่มมีมากเพียงพอที่จะให้ทำให้การประมาณค่าพารามิเตอร์มีความลำเอียงมาก ซึ่งจะส่งผลความเที่ยงและความน่าเชื่อถือของผลการวิจัยที่ได้ต่ำ

เทคนิคในการลดความแปรปรวนในการประมาณค่า คือ การเพิ่มจำนวนรอบ ซึ่งจะทำให้ได้ค่าที่คงที่และน่าเชื่อถือได้มากกว่า เนื่องจากสามารถเปรียบเทียบค่าพารามิเตอร์ระหว่างรอบได้ ข้อดีอีกประการ คือ จำนวนรอบสะท้อนความเบี่ยงเบนของค่าประมาณ ซึ่งถ้าความเบี่ยงเบนระหว่างเงื่อนไขมีค่าน้อย แสดงว่า ตัวแปรอิสระส่งผลต่อตัวแปรตามน้อย โดยสมการในการคำนวณความเบี่ยงเบนของค่าพารามิเตอร์ที่แท้จริงและค่าประมาณได้จากค่า RMSD (Root mean square deviation) หรือ RMSE (Root mean square error) มีสมการคือ

$$RMSE = \left[\frac{\sum_{i=1}^{n}(\widehat{a_i}-a_i)^2}{n}\right]^{1/2} \quad (69)$$

เมื่อ $\widehat{a_i}$ = ค่าพารามิเตอร์ที่ได้จากการประมาณค่า

a_i = ค่าพารามิเตอร์แท้จริง

n = จำนวนพารามิเตอร์

นอกจากนี้เมื่อพิจารณาความแปรปรวนของการประมาณค่าจากจำนวนรอบ (The variance of the estimates across replications) และฟังก์ชันความแปรปรวนความคลาดเคลื่อนของข้อมูลเชิงประจักษ์ (Empirical error variance) มีสมการ ดังนี้

$$\frac{\sum_{r=1}^{R}(\hat{a}_{ir}-a_i)^2}{R} = (\bar{\hat{a}}_{\imath} - a_i)^2 + \frac{\sum_{r=1}^{R}(\hat{a}_{ir}-\bar{\hat{a}}_{\imath})^2}{R} \quad (70)$$

จะเห็นว่าถ้าค่า RMSE มีค่าน้อย แสดงว่า การประมาณค่ามีความเที่ยงหรือมีความคงที่ ส่วนถ้า RMSE มีค่ามาก แสดงว่าการประมาณค่าไม่มีความเที่ยงหรือไม่คงที่

เมื่อพิจารณาเรื่องจำนวนรอบกับอำนาจการทดสอบ (Power) พบว่า จำนวนรอบมีความสำคัญกับอำนาจการทดสอบ โดยจำนวนรอบต้องมีขนาดมากพอ โดย Stone (1993; Harwell et al., 1996) ได้ศึกษาความสัมพันธ์ระหว่างจำนวนรอบกับอำนาจการทดสอบ มีขั้นตอน ดังนี้ ขั้นตอนแรกใช้เทคนิค MC ในการศึกษาผลของจำนวนผู้สอบ N (N = 250, 500, 1000) จำนวนข้อสอบ (L = 10, 20, 30) และการแจกแจงของความสามารถ (D = Normal, Skewed, Platykurtic) ทดลองแบบแฟคตอเรียล โดยมีค่า RMSE เป็นตัวแปรตาม ใช้โมเดล 2 พารามิเตอร์ เพื่อประมาณค่าอำนาจจำแนกและค่าความยากแต่ละข้อ กำหนดให้ R = 10 (ข้อมูล 10 ชุด) ในแต่ละเงื่อนไขแสดง RMSE ด้วยกราฟ และทดสอบด้วย ANOVA เพื่อตรวจสอบอิทธิพลของตัวแปรต้น ค่าอิทธิพลสูงสุดแสดงด้วยค่า η^2 (correlation ratio = η^2) ซึ่งมีความเหมาะสมเนื่องจากสอดคล้องกับข้อตกลงเบื้องต้นที่การแจกแจงต้องเป็นโค้งปกติ

ขั้นตอนที่ 2 ใช้ค่า η_s^2 ที่ได้จากขั้นตอนแรก ประมาณค่าอำนาจการทดสอบของ ANOVA เพื่อตรวจสอบอิทธิพลที่เกิดจากจำนวนรอบที่ต่างกัน (R=10, 25, 50, 100) โดยใช้ η_s^2 เป็นตัวประมาณขนาดอิทธิพล ด้วยโปรแกรม STAT - POWER จะใช้ประมาณค่าอำนาจการทดสอบที่ระดับ .05 และองศาอิสระเท่ากับ 27 และ η^2 กระบวนการนี้ทำให้ อำนาจการทดสอบของ ANOVA F - test เปลี่ยน ส่วนความสัมพันธ์ระหว่างจำนวนรอบกับอำนาจการทดสอบ พบว่า อำนาจการทดสอบแปรผันตรงตามจำนวนรอบ

อย่างไรก็ตาม เมื่อจำนวนผู้สอบ (N) และจำนวนข้อสอบ (L) มีขนาดใหญ่พอ อาจน้อยกว่า 100 รอบก็ได้ อาจกำหนดให้จำนวนรอบน้อยกว่า 100 รอบก็ได้ แต่อาจกำหนดจำนวนรอบ (R) ให้เพิ่มขึ้น เพื่อความน่าเชื่อถือของการศึกษา

จะเห็นได้ว่าจำนวนรอบมีความจำเป็นต่อความเที่ยงในการตรวจสอบผลกระทบของผลการจำลองข้อมูลมากในงานวิจัยที่มีลักษณะ ดังนี้

1. งานวิจัยที่สนใจลักษณะการกระจายของการสุ่มตัวอย่างจากข้อมูลเชิงประจักษ์ เช่น งานวิจัยที่ต้องการตรวจสอบคุณสมบัติของค่าสถิติหรือทดสอบนัยสำคัญทางสถิติ

2. งานวิจัยที่ศึกษาค่ากลางของการวิจัยระดับข้อสอบที่มีค่าความแปรปรวนของกลุ่มตัวอย่างมาก

3. งานวิจัยที่มีเป้าหมายในการศึกษาผลกระทบที่ของบริบทที่ซับซ้อน เช่น ผลกระทบที่เกิดขึ้นจากปฏิสัมพันธ์กับอิทธิพลหลัก

4. การวิเคราะห์ด้วยทฤษฎีการตอบสนองควรในการจำลองข้อมูล ควรมีค่าไม่ต่ำกว่า 25 รอบ

3. การเลือกใช้โปรแกรมในการจำลองข้อมูลและการประมาณค่าพารามิเตอร์ อาจใช้โปรแกรมหลายโปรแกรมในการจำลองข้อมูลและวิเคราะห์ผลลัพธ์ก็ได้ ซึ่งมีรายละเอียด ดังนี้

3.1 จำลองคำตอบ เริ่มจากกำหนดค่าเริ่มต้น (Seed) ให้กับตัวเลขสุ่ม ซึ่งจะแปลงเป็นค่าความน่าจะเป็นในการตอบคำถามได้ถูกต้อง

3.2 ตัวเลือกค่าเริ่มต้น ซึ่งผู้วิจัยกำหนดค่าเริ่มต้นได้เอง โดยใช้เติมในช่องว่าง (Prompted) ซึ่งมีประโยชน์ คือ ง่ายต่อการจำลองคำตอบในข้อต่อ ๆ ไป และเป็นค่าที่สัมพันธ์กับความคลาดเคลื่อนในการสุ่ม ซึ่งหลีกเลี่ยงได้ยากในการจำลองข้อมูล เทคนิคอย่างหนึ่งที่จะลดความแปรผัน คือ ใช้พารามิเตอร์ข้อสอบและค่าเริ่มต้นร่วมกันทุกครั้ง เมื่อมีการจำลองข้อมูล เช่น การจำลองข้อสอบ 20 ข้อและ 30 ข้อ ค่าเริ่มต้นที่ใช้ในการจำลองข้อมูล 20 ข้อ ควรจะเป็นค่าเริ่มต้นเดียวกันกับเมื่อจำลองข้อมูล 30 ข้อ เทคนิคอีกประการหนึ่ง คือ การจำลองประชากรข้อมูลคำตอบจำนวนมาก แล้วสุ่มคำตอบมาจากประชากรที่จำลองขึ้นนั้นหลาย ๆ รอบมากกว่าการใช้ ค่าเริ่มต้นหลายตัว เพื่อจะจำลองชุดข้อมูลให้ได้ตามต้องการ การใช้โมเดลพารามิเตอร์ที่ต่างกันในการจำลองข้อสอบ 20 ข้อ 30 ข้อ รวมทั้งค่าเริ่มต้นที่ต่างกันจะทำให้คำตอบมีความเป็นอิสระแก่กันมากขึ้น แต่อาจจะเกิดความคลาดเคลื่อนมากกว่าเมื่อเทียบกับการใช้โมเดลพารามิเตอร์และค่าเริ่มต้นแบบเดียวกัน วิธีที่ควรใช้คือ การจำลองข้อมูลทุกชุดควรใช้โมเดลพารามิเตอร์ต่างกัน แต่ใช้ค่าเริ่มต้นร่วมกัน ในการวิจัยครั้งนี้ ผู้วิจัยใช้ค่าเริ่มต้น (Seed) จากตารางสุ่มโดยดาว์นโหลดได้ที่เว็บไซต์ http://www.rand.org

3.3การจำลองตัวเลขสุ่ม จะใช้ตัวเลขสุ่มที่มีการแจกแจงแบบยูนิฟอร์ม (Uniform distribution) เป็นส่วนใหญ่และใช้วิธีการจำลองข้อมูลแบบ Congruential generators ซึ่งเป็นวิธีที่ใช้โมเดลพีชคณิตในการจำลองตัวเลขที่สุ่มขึ้นมากับตัวเลขสุ่มที่ผ่านมา ตัวเลขสุ่มจะเริ่มจำนวนจาก 0....m ซึ่งจะสุ่มโดยโปรแกรม โดยจะวิ่งเป็นวงจรที่เรียกว่า “ความยาวรอบ” (Period) เมื่อครบรอบก็จะวนกลับมาใช้เลขเดิมอีก จากการศึกษาที่ผ่านมา พบว่า วิธีนี้มีโอกาสเกิดข้อมูลซ้ำเมื่อมีช่วงความยาวรอบมากขึ้น การแจกแจงปกติมาตรฐานจะใช้มากในการใช้เทคนิค MC

โดยการแปลงข้อมูลจาก Uniform(0, 1) เป็น Normal ~ N(0, 1) วิธีที่ใช้แปลงคือ วิธีของ Box - Muller และ Marsaglia

หลักในการเลือกโปรแกรมเพื่อใช้ในการจำลองข้อมูล ประกอบด้วย 4 อย่าง คือ วิธีการควรง่ายต่อการทำความเข้าใจและเขียนโปรแกรม โปรแกรมควรมีความกะทัดรัด รหัสปลายทางควรมีความสมเหตุสมผล (0, 1) และขั้นตอนการคำนวณ ควรใช้ตัวเลขสุ่มเทียม (Pseudorandom numbers) (Ripley, 1987)

3.4 การแปลงตัวเลขสุ่มเป็นคำตอบ เริ่มจากสุ่มค่าจากการแจกแจง (โดยมากเป็นการแจงแจงแบบปกติ) เพื่อให้ได้ N (จำนวนผู้สอบ) และ θ (ความสามารถ) และเป็นโมเดล IRT แบบเอกมิติหรือจากการแจกแจงปกติหลายตัวแปรที่มีค่าความสัมพันธ์ระหว่างตัวแปร เช่น สุ่มความน่าจะเป็นของคำตอบที่ตอบแบบ (0, 1) จะได้จากสมการ ดังนี้

$$P_i(\theta) = c_i + \frac{(1-c_i)}{1+e^{-Da_i(\theta-b_i)}} \tag{71}$$

ค่าความน่าจะเป็นของการตอบ (p) จะแปลงไปเป็นคำตอบ 0, 1 โดยเปรียบเทียบค่าความน่าจะเป็นจากการสุ่มการแจกแจงแบบยูนิฟอร์ม (u) ถ้า (p - u) $\geq$ 0 ให้เป็น 1 (ตอบถูก) และถ้า (p - u) < 0 ให้เป็น 0 (ตอบผิด) ในกรณีเป็นคำตอบหลายค่า ถ้าตัวเลขสุ่มตกอยู่ในช่วงใดก็เป็นคำตอบ K และ K + 1 โดยทำซ้ำด้วยตัวเลขสุ่มที่แตกต่างกันในแต่ละข้อและผู้สอบทั้งหมด

3.5 การประมาณค่าพารามิเตอร์ในโมเดล อาจใช้โปรแกรมสำเร็จรูป เช่น BILOG หรือ MULTILOG หรือสร้างโปรแกรมเอง โดยกำหนดค่าเริ่มต้น (Starting value) และแก้ปัญหาการไม่ลู่เข้า (Non - convergent solutions) ซึ่งหากพบว่าการประมาณค่าไม่ลู่เข้า จะสามารถดำเนินการ 3 ทางเลือก โดยทางเลือกที่ 1 คือ ไม่สนใจการไม่ลู่เข้านั้น แล้วใช้ค่าประมาณจากจำนวนครั้งของการคำนวณซ้ำที่มากที่สุด ทางเลือกที่สอง แยกการประมาณค่าของค่าสถิติ สรุปรวม ได้แก่ RMSDs และทางเลือกที่สามใช้วิธีการคำนวณอื่น ๆ เช่น วิธีของเบส์ (Bayesian) เพื่อบังคับค่าพารามิเตอร์ โดยควรคำนึงถึงค่าพารามิเตอร์ที่ประมาณค่ากับค่าพารามิเตอร์ที่แท้จริงให้มีค่าใกล้เคียงกันด้วย มิฉะนั้นแล้ว ค่าที่ได้จะเป็นค่าที่ลำเอียง

ปัญหาสำคัญของการกำหนดค่าเริ่มต้นและการแก้ปัญหาการไม่ลู่เข้า มักพบในข้อมูลที่มีจำนวนน้อย (N = 200, 40 ข้อ) และในโมเดล IRT ที่ซับซ้อนมากขึ้น ซึ่งแก้ได้โดยใช้ประมาณค่าด้วยวิธีการของเบส์ (Bayesian) ซึ่งมีความเกี่ยวข้องกับการแจกแจงของข้อมูลและการแจกแจงของค่าพารามิเตอร์

4. วิเคราะห์ผลจากการจำลองข้อมูล การวิเคราะห์ผลจะอยู่บนพื้นฐานของคำถามวิจัย การออกแบบการทดสอบสมมติฐานทางสถิติ กระบวนการวิเคราะห์ โดยปกติการวิเคราะห์ผลประกอบด้วย การใช้ตารางสรุปผลรวม สถิติเชิงบรรยายเบื้องต้นหรือการนำเสนอด้วยกราฟแผนภูมิ การดูผลกระทบจากตัวแปรอิสระ อาจต้องใช้สถิติเชิงอ้างอิง ปัญหาของการวิเคราะห์ คือการมีค่าต่าง ๆ กันจำนวนมาก เมื่อต้องการรายงานผล Harwell (1991 cited in Harwell et al., 1996) เสนอว่าควรใช้ทั้งเชิงบรรยายและเชิงอ้างอิง เพื่อเพิ่มโอกาสในการตรวจสอบข้อมูล ซึ่งจะทำให้มีความเชื่อมั่นมากขึ้น

การวิเคราะห์ผลด้วยการใช้สถิติเชิงสรุปอ้างอิงสามารถใช้ได้หลายวิธี แต่มักเป็นการวิเคราะห์ถดถอยพหุคูณและการวิเคราะห์ความแปรปรวน ถ้าตัวแปรต้นเป็นตัวแปรระดับนามบัญญัติ การวิเคราะห์ความแปรปรวนจะดีกว่า แต่ถ้าเป็นตัวแปรระดับช่วงชั้น การใช้การวิเคราะห์ถดถอยพหุคูณจะดีกว่า แต่ส่วนใหญ่แล้วในทฤษฎีการตอบสนองข้อสอบจะใช้ การวิเคราะห์ถดถอยพหุคูณ เนื่องจากตัวแปรส่วนใหญ่จะเป็นระดับช่วงชั้น หรืออัตราส่วน

ข้อควรระวังในการประยุกต์ใช้เทคนิค MC สำหรับการวิเคราะห์ข้อสอบ

ปัจจุบันมีการประยุกต์ใช้ทคนิคมอนติคาร์โลสำหรับการวิเคราะห์ข้อสอบกันอย่างแพร่หลาย โดยตั้งแต่ปี 1981 ถึง ปี 1991 มีงานวิจัยที่ลงตีพิมพ์ในวารสาร Psychometrika และ JEM (Journal of measurement) ซึ่งเป็นวารสารที่มีชื่อเสียงสำหรับการวิจัยทางด้านวัดและประเมินผล 26 เรื่อง (Harwell et al., 1996) บางเรื่องมีการประยุกต์ใช้เทคนิค MC บางส่วน ขณะที่บางเรื่องใช้เทคนิค MC ทั้งหมด นอกจากนี้ งานวิจัยนั้นมักขาดข้อมูลในการสนับสนุนอย่างเพียงพอ เช่น งานวิจัย 16 เรื่อง (ร้อยละ 61.5) ไม่มีการกำหนดจำนวนรอบในการศึกษาซึ่งทำให้ผลการวิเคราะห์เกิดความคลาดเคลื่อน และงานวิจัยจำนวน 20 เรื่อง (ร้อยละ 76.9) ขาดหลักฐานในการแจงแจงความสามารถและค่าพารามิเตอร์ของข้อสอบที่ตรงกับสภาพความจริง ส่งผลให้งานวิจัยขาดความตรงภายนอก ดังนั้น Harwell et al. (1996) จึงเสนอข้อควรระวัง สำหรับงานวิจัยที่จะใช้การวิเคราะห์ข้อสอบด้วยการใช้เทคนิค MC ที่สำคัญก่อนตัดสินใจใช้เทคนิคมอนติคาร์โลสำหรับการวิเคราะห์ข้อสอบ สรุปได้ดังนี้ (Harwell et al., 1996)

1. ปัญหาการวิจัยสามารถแก้ได้ด้วยการวิเคราะห์ด้วยเทคนิค MC จริงหรือไม่

2. การศึกษาในประเด็นนั้นสามารถขยายองค์ความรู้เดิมได้หรือไม่

3. การออกแบบงานวิจัยเหมาะที่จะใช้เทคนิค MC ในการศึกษาหรือไม่

4. การวิเคราะห์ข้อมูลควรจะนำโปรแกรมที่มีอยู่เดิมมาใช้หรือปรับปรุงโปรแกรมใหม่เพื่อให้เหมาะสมกับสภาพปัญหาการวิจัยของเราหรือไม่

5. ผลลัพธ์ในการวิเคราะห์ข้อมูลในงานวิจัยขึ้นอยู่กับค่าเริ่มต้นสำหรับวิธีการประมาณค่าพารามิเตอร์หรือไม่

6. ข้อตกลงเบื้องต้นของการแจกแจงลักษณะตัวแปรต้นและค่าต่างๆ ที่กำหนดในการจำลองข้อมูลตรงกับสภาพความเป็นจริงหรือไม่

ข้อดีและข้อจำกัดของการศึกษาโดยใช้ข้อมูลจำลอง

ส่วนใหญ่งานวิจัยที่เหมาะกับการใช้ข้อมูลจำลองเป็นการศึกษาสถานการณ์ที่ต้องการตรวจสอบการแจกแจงทางสถิติ การเปรียบเทียบตัวประมาณค่าที่เก็บข้อมูลได้ยาก การศึกษาความแกร่งของสถิติ การเปรียบเทียบขั้นตอนการคำนวณของฟังก์ชัน หรือการประเมินขั้นตอนในการคำนวณ (Harwell et al., 1996) โดยการใช้ข้อมูลจำลองในการศึกษามีข้อดี ดังนี้

1. การจำลองข้อมูลสามารถสร้างข้อมูลที่มีเงื่อนไขต่าง ๆ ที่มีความซับซ้อน หรือการเก็บรวมข้อมูลจริงทำได้ยากและมีจำนวนข้อมูลน้อย ซึ่งนำมาวิเคราะห์สถิติในบางประเภทไม่ได้

2. การจำลองข้อมูลสามารถกำหนดและจัดกระทำค่าพารามิเตอร์ ซึ่งนำไปใช้ศึกษาปัจจัยที่ส่งผลต่อองค์ประกอบต่าง ๆ ได้ และค่าใช้จ่ายน้อยกว่าการเก็บข้อมูลจริง

3. การจำลองข้อมูลสามารถกำหนดระยะเวลาที่แน่นอนในการดำเนินการทดลองได้เหมาะสำหรับงานวิจัยที่ต้องการใช้ผลการวิเคราะห์รวดเร็ว

4. การจำลองข้อมูลสามารถทำการสุ่มและสร้างตัวแปรเทียมสำหรับใช้ในตัวแบบได้โดยตรง ซึ่งในการเก็บรวบรวมด้วยข้อมูลจริงบางครั้ง บางสถานการณ์ไม่สามารถกระทำได้โดยตรง

ส่วนข้อจำกัดของการศึกษาโดยการใช้ข้อมูลจำลอง สรุปได้ดังนี้

1. บางครั้งผลที่ได้จากการการจำลองข้อมูล อาจไม่ครอบคลุมทุกกรณีเหมือนกับสถานการณ์ที่เกิดขึ้นจริง

2. การวิจัยบางประเภทมีกระบวนการในการสร้างตัวแบบที่ทำได้ยาก

3. การนำผลจากการจำลองข้อมูลไปใช้ ต้องพิจารณาเงื่อนไขที่ศึกษาว่าสอดคล้องกับสภาพจริงหรือไม่

4. ผลการจำลองข้อมูลนั้นขึ้นอยู่กับจำนวนรอบและความถูกต้องของตัวเลขที่จำลองได้ ดังนั้น คุณธรรมของนักวิจัยที่จำลองข้อมูลจึงเป็นเรื่องที่สำคัญ ซึ่งยากต่อการประเมิน

5. หากเลือกใช้วิธีการที่ไม่เหมาะสมสำหรับจำลองข้อมูล อาจทำให้ผลสรุปในงานวิจัยผิดพลาดและอาจไม่สามารถนำไปใช้ประโยชน์ได้จริง

ดังนั้น สำหรับการศึกษาด้วยการจำลองข้อมูล จึงต้องพิจารณาด้วยงานวิจัยที่ศึกษามีลักษณะอย่างไร เหมาะสมหรือไม่ หรือควรจะใช้ข้อมูลจริง การวิจัยที่ใช้ข้อมูลจำลอง จึงควรเป็น

งานวิจัยที่มีความสลับซับซ้อน ไม่สามารถได้จากการเก็บรวบรวมข้อมูลจริง (Stone, 1993 cited in Harwell et al., 1996)

การแจกแจงความน่าจะเป็นของตัวแปรสุ่ม

การแจกแจงความน่าจะเป็นของตัวแปรสุ่มมีหลายประเภท แต่ผู้วิจัยจะนำเสนอเฉพาะที่มีความเกี่ยวข้องกับสำหรับการศึกษาในครั้งนี้ ได้แก่ การแจกแจงแบบยูนิฟอร์ม (Uniform distribution) การแจกแจงแบบปกติ (Normal distribution) และการแจกแจงแบบ Skew - normal (Skew - normal distribution) มีรายละเอียดเกี่ยวกับการแจกแจงดังนี้

1. การแจกแจงแบบยูนิฟอร์ม (Uniform distribution)

การแจกแจงแบบยูนิฟอร์มมีลักษณะการแจกแจงที่มีค่าได้ทุกค่าจริงในช่วง [a, b], $-\infty < a,b < \infty$ ด้วยความน่าจะเป็นเท่า ๆ กัน โดยสามารถเขียนแทนด้วยสัญลักษณ์ $X \sim U(a, b)$ โดยมีฟังก์ชันความน่าจะเป็นดังนี้ (Berenson, Levine & Krehbiel, 2012)

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{เมื่อ} \quad a \leq x \leq b \\ 0 & \text{เมื่อ} \quad a < x \text{ และ } x < b \end{cases} \tag{72}$$

โดยที่ b = ค่าสูงสุด และ a = ค่าต่ำสุด

การแจกแจงแบบยูนิฟอร์มจะมีค่าเฉลี่ย (Mean) ของ X เท่ากับ

$$\mu = \frac{a+b}{2} \tag{73}$$

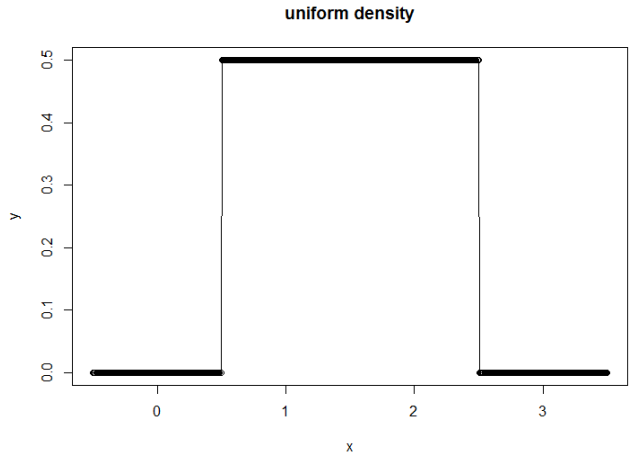
การแจกแจงแบบยูนิฟอร์มจะมีค่าความแปรปรวน (Variance) ของ X คือ

$$\sigma^2 = \frac{(b-a)^2}{12} \tag{74}$$

การแจกแจงแบบยูนิฟอร์มมีค่าความเบี่ยงเบนมาตรฐาน (Standard deviation) ของ X คือ

$$\sigma = \sqrt{\frac{(b-a)^2}{12}} \tag{75}$$

การแจกแจงแบบยูนิฟอร์ม สามารถแสดงตัวอย่างกราฟของฟังก์ชันการแจกแจงแบบยูนิฟอร์ม สำหรับค่าพารามิเตอร์บางค่าได้ดังนี้



ภาพที่ 2 - 6 PDF ของการแจกแจงแบบยูนิฟอร์มด้วยค่า a = 0.5 และ b = 2.5

2. การแจกแจงแบบปกติ (Normal distribution)

การแจกแจงแบบปกติ (Normal distribution) บางครั้งเรียกว่า Gaussian distribution มีลักษณะเป็นรูประฆัง (Classic bell shape) เป็นการแจกแจงความน่าจะเป็นของตัวแปรสุ่มที่ส่วนมากจะมีค่าใกล้เคียงค่าเฉลี่ยของตัวแปรเหล่านั้น จะมีค่าของตัวแปรที่มากกว่าหรือน้อยกว่าค่าเฉลี่ยเป็นส่วนน้อย โดยมีลักษณะที่สำคัญ คือ มีความสมมาตร (Symmetrical) มีค่าเฉลี่ย (Mean) และค่ามัธยฐาน (Median) เท่ากัน มีค่าได้ในช่วงไม่จำกัด (Infinite range) หรือ $-\infty < X < \infty$ เขียนแทนด้วยสัญลักษณ์ $X \sim N(\mu, \sigma^2)$ โดยมีฟังก์ชันความน่าจะเป็นดังนี้ (Berenson et al., 2012)

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{1}{2}\right)[(X-\mu)/\sigma]^2} \tag{76}$$

โดยที่ e = เป็นค่าคงที่ประมาณ 2.71828

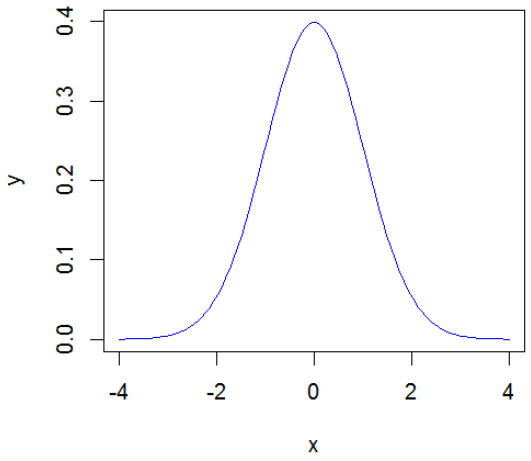
π = เป็นค่าคงที่ประมาณ 3.14159

μ = ค่าเฉลี่ย

σ = ค่าส่วนเบี่ยงเบนมาตรฐาน (Standard deviation)

X = ค่าใด ๆ ของตัวแปรต่อเนื่อง (Continuous variable) ที่อยู่ในช่วง $-\infty < X < \infty$

การแจกแจงแบบปกติมีค่าเฉลี่ยเท่ากับ μ และมีส่วนเบี่ยงเบนมาตรฐานเท่ากับ σ สามารถแสดงตัวอย่างกราฟของฟังก์ชันการแจกแจงปกติ สำหรับค่าพารามิเตอร์บางค่า ดังนี้



ภาพที่ 2 - 7 PDF ของการแจกแจงแบบปกติด้วยค่า $\mu = 0$ และ $\sigma = 1$

3. การแจกแจงแบบ Skew - normal (Skew - normal distribution)

การแจกแจงแบบ Skew - normal เป็นส่วนขยายจากการแจกแจงแบบปกติ มีลักษณะเบ้ โดยมีฟังก์ชันความน่าจะเป็นดังนี้ (Azzalini, 2014; Figueiredo & Gomes, 2013)

$$f(x) = 2\phi(x)\Phi(\alpha x) \tag{77}$$

โดยที่ X = เป็นตัวแปรสุ่มต่อเนื่อง, $-\infty < x < \infty$

α = ค่าคงที่

$$\phi(x) = \exp(-\frac{x^2}{2})\sqrt{2\pi}$$

$$\Phi(\alpha x) = \int_{-\infty}^{\alpha z} \phi(t)dt$$

กำหนด Location parameter และ Scale parameter ให้ Z เป็นตัวแปรสุ่มด้วยฟังก์ชันความน่าจะเป็นดังกล่าวแล้ว จะได้

$$Y = \xi + \omega Z \qquad (\xi \in \mathbf{R}, \omega \in \mathbf{R}^{+}) \tag{78}$$

เรียกตัวแปร Skew - normal (SN) ที่มีพารามิเตอร์ ได้แก่ Location parameter (ξ), Scale parameter (ω) และ Slant parameter (α) มีฟังก์ชันความหนาแน่น ที่ $x \in \mathbf{R}$ เป็น

$$\frac{2}{\omega}\varphi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\alpha\frac{x-\xi}{\omega}\right) = \frac{1}{\omega}\varphi\left(\frac{x-\xi}{\omega};\alpha\right) \tag{79}$$

โดยที่ $Y \sim SN(\xi,\omega^2,\alpha)$ เมื่อ ω^2 เทียบได้กับ $N(\mu, \sigma^2)$ ซึ่งเมื่อ $\alpha = 0$ จะทำให้การเบ้หายไป หรือเป็นการแจกแจงแบบปกตินั่นเอง เมื่อค่าสัมบูรณ์ของ α เพิ่มขึ้น จะทำให้ความเบ้เพิ่มขึ้น และเมื่อ α เข้าสู่ค่านิรันดร์ ($\alpha \to \pm\infty$) ความหนาแน่นจะลู่เข้า Half - normal (Folded normal) density function ถ้าเครื่องหมายของ α เปลี่ยน กราฟจะกลับในทางตรงข้าม

การแจกแจงแบบ Skew - normal จะมีค่าเฉลี่ย (Mean) ของ Y เท่ากับ

$$\mathbf{E}\{Y\} = \xi + \omega\sqrt{2/\pi}\delta \text{ เมื่อ } \delta = \frac{\alpha}{\sqrt{1+\alpha^2}} \tag{80}$$

การแจกแจงแบบ Skew - normal จะมีค่าความแปรปรวน (Variance) ของ Y เท่ากับ

$$\text{var}\{Y\} = \omega^2(1 - 2\delta^2/\pi) \tag{81}$$

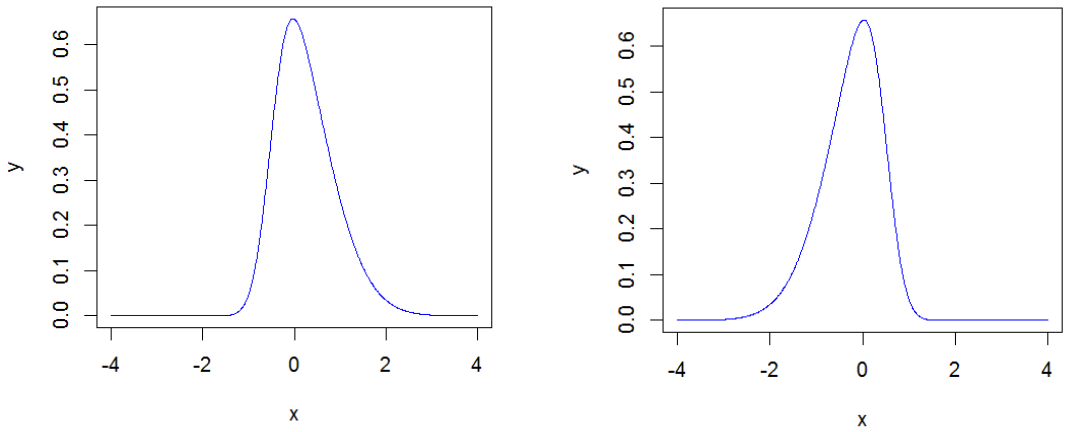
การแจกแจงแบบ Skew - normal จะมีความเบ้ (Skewness) ของ Y เท่ากับ

$$\gamma_1 = \frac{4-\pi}{2}\frac{\mathbf{E}\{X\}^3}{\text{var}\{X\}^{3/2}} \tag{82}$$

การแจกแจงแบบ Skew - normal จะมีความโด่ง (Kurtosis) ของ Y เท่ากับ

$$\gamma_1 = \frac{4-\pi}{2}\frac{\mathbf{E}\{X\}^3}{\text{var}\{X\}^{3/2}} \tag{83}$$

และสามารถแสดงตัวอย่างกราฟของฟังก์ชันความหนาแน่นของความน่าจะเป็น (Probability density function: PDF) ของการแจกแจงแบบ Skew - normal สำหรับค่าพารามิเตอร์บางค่าได้ ดังนี้



(a) $\xi = -0.5, \omega = 1, \alpha = 3$ (b) $\xi = 0.5, \omega = 1, \alpha = -3$

ภาพที่ 2 - 8 PDF ของการแจกแจงแบบ Skew - normal (a) เบ้ขวา และ (b) เบ้ซ้าย

ซึ่งในการวิจัยครั้งนี้ ใช้การจำลองข้อมูลจากการแจกแจงข้อมูลใน 3 รูปแบบ ได้แก่ การแจกแจงแบบยูนิฟอร์ม การแจกแจงแบบปกติ และการแจกแจงแบบ Skew - normal โดยใช้โปรแกรม R ส่วน Model ในการประมาณค่าส่วนที่มีการเรียกใช้โปรแกรม WinBUGS ในส่วนที่เป็น Prior distribution ของพารามิเตอร์เป็นการแจกแจงแบบปกติ

ตอนที่ 5 การกำหนดเงื่อนไขสำหรับการจำลองข้อมูล

เนื่องจากการสร้างเครื่องมือเพื่อเก็บข้อมูลให้ได้ตามเงื่อนไขที่ต้องการศึกษานั้นทำได้ยาก ดังนั้น การศึกษาครั้งนี้จึงใช้การศึกษาจากข้อมูลจำลอง ซึ่งมีรายละเอียดของประเด็นที่ศึกษา ดังนี้

การกำหนดจำนวนและขนาดของเทสต์เลท

การกำหนดจำนวนเทสต์เลท จำนวนข้อสอบใน 1 เทสต์เลทในแบบสอบ พิจารณาจาก Zhang (2010) ได้สรุปข้อมูลงานวิจัยเกี่ยวกับเทสต์เลทในฐานข้อมูลของ EBSCO และ PsychInfo ระหว่างปี ค.ศ. 1989 - 2009 ในประเด็นขนาดของเทสต์เลท พบว่า ในจำนวน 45 บทความที่เกี่ยวข้อง กับเทสต์เลทมี 4 บทความที่ศึกษาเทสต์เลทขนาดเล็ก (ขนาดน้อยกว่า 5 ข้อต่อเทสต์เลท)

ที่เหลือ 41 บทความ ศึกษาขนาดของเทสต์เลทหลายขนาดผสมกัน (ขนาดเล็กและใหญ่) โดยมีรายละเอียดดังตารางที่ 2 - 7

ตารางที่ 2 - 7 ขนาดของเทสต์เลทจากบทความในฐานข้อมูลของ EBSCO และ PsychInfo ระหว่างปี ค.ศ. 1989 - 2009 (Zhang, 2010)

จำนวนข้อใน 1 เทสต์เลท (m)	จำนวนบทความ	ร้อยละ
$m < 5$	16	25.40
$5 \leq m \leq 10$	35	55.56
$11 \leq m \leq 15$	6	9.52
$16 \leq m \leq 20$	2	3.17
$21 \leq m \leq 25$	3	4.76
$m > 25$	1	1.59

จากตารางที่ 2 - 7 จะเห็นว่างานวิจัยส่วนใหญ่ร้อยละ 55.56 ใช้ขนาดของเทสต์เลทเป็น 5 - 10 ข้อ ประกอบกับ Chen (2010) พบว่า งานวิจัยที่ผ่านมาเกี่ยวกับทฤษฎีการตอบสนองข้อสอบที่มีลักษณะเทสต์เลท ส่วนมากทำการศึกษาจำนวนข้อสอบใน 1 เทสต์เลทอยู่ระหว่าง 2 ถึง 25 ข้อต่อ 1 เทสต์เลท ซึ่งเฉลี่ยแล้วใช้จำนวนข้อสอบ 10 ข้อต่อ 1 เทสต์เลท และเพื่อให้สอดคล้องกับการศึกษาเงื่อนไข อิทธิพลของเทสต์เลท ซึ่งงานวิจัยต่าง ๆ ได้กำหนดให้แบบสอบมี 4 เทสต์เลท (Wainer et al., 2007; Jiao et al., 2013) ดังนั้น ในการศึกษาครั้งนี้ จึงกำหนดความยาวแบบสอบเป็น 40 ข้อ ประกอบด้วย 4 เทสต์เลท โดยแต่ละเทสต์เลทมี 10 ข้อ ใช้โมเดลแบบ 2 พารามิเตอร์ (ความยากและอำนาจจำแนก) เหมือนกันทุกเงื่อนไข

การกำหนดรูปแบบและการแจกแจงของพารามิเตอร์ที่ใช้ในการจำลองข้อมูล

สำหรับการกำหนดรูปแบบและการแจกแจงของพารามิเตอร์เพื่อสำหรับใช้ในการจำลองข้อมูล นั้น ผู้วิจัยได้ศึกษาการกำหนดค่าพารามิเตอร์และการแจกแจงโดยผู้วิจัยได้สรุปรูปแบบและการแจกแจงของพารามิเตอร์ของข้อสอบที่ใช้ในการจำลองข้อมูล ได้ดังตารางที่ 2 - 8

ตารางที่ 2 - 8 รูปแบบและการแจกแจงของพารามิเตอร์ของข้อสอบที่ใช้ในการจำลองข้อมูลของงานวิจัยที่ผ่านมา

การศึกษา	อำนาจจำแนก (a_j)	ความยาก (b_j)	จำนวนพารามิเตอร์ที่ศึกษา	ประเด็นที่ศึกษา
Rousos & Stout (1996)	1.32 0.4, 1.0, 2.5	0.03 -1.5, -0.5, 0, 0.5, 1.5	3PL 3PL	ผลของกลุ่มตัวอย่างขนาดเล็กที่มีต่อการตรวจสอบ DIF ระหว่างวิธี SIBTEST และ MH
Bradlow et al. (1999)	$N(0.8, 0.2^2)$	N(0, 1)	2PL	ใช้วิธี Bayessian ในการประมาณค่าพารามิเตอร์ข้อมูลแบบ Testlet (ให้อิทธิพล Testlet เป็น Random effect)
Wang et al. (2002)	$N(1.5, 0.45^2)$ ใช้ค่าเริ่มที่ 0.3	N(0, 1)	3PL	ใช้วิธี Bayessian ในการประมาณค่าพารามิเตอร์ข้อมูลแบบ Testlet
Glas & Meijer (2003)	LN(0, 0.5) 0.5, 1.0, 1.5	N(0, 0.5) b_j =-2.00 + 0.40(j - 1), j = ข้อที่ j	3PL 3PL	ใช้วิธี Bayessian ในการประมาณค่าพารามิเตอร์ของ IRT
Li et al. (2006)	$N(0.8, 0.2^2)$ $N(0.8, 0.2^2)$ 1.2 $N(0.8, 0.2^2)$	N(0, 1) ช่วง [-3, 3] N(0, 1) ช่วง [-3, 3] N(0, 1) ช่วง [-3, 3] N(0, 1) ช่วง [-3, 3]	2PL 2PL 2PL 2PL	เปรียบโมเดลสำหรับประมาณค่าพารามิเตอร์ข้อมูลแบบ Testlet 4 โมเดล
Chaimongkol et al. (2007)	0.2, -0.2, 0.5, 0.8, -0.5, 0, 0.8	-	1PL	ศึกษา DIF โดยใช้ Multilevel logistic regression model 3 ระดับ โดยใช้ WinBUGS

ตารางที่ 2 - 8 (ต่อ)

การศึกษา	อำนาจจำแนก (a_j)	ความยาก (b_j)	จำนวนพารามิเตอร์ที่ศึกษา	ประเด็นที่ศึกษา
Soares, Gonçalves, & Gamerman (2009)	LN(0.2, 0.3)	LN(0.2, 0.3)	3PL	ประยุกต์ Bayessian ในการวิเคราะห์ DIF ของ IRT
Fukuhara & Kamata (2011)	0.8	-1.5, -1.0, -0.5, 0, 0.5, 1.0, 1.5	2PL	การใช้วิธี Bi - factor MIRT ในการตรวจสอบ DIF
	2	-1.5, -1.0, -0.5, 0, 0.5, 1.0, 1.5	2PL	
Huggins (2014)	unif(0.55,1.36)	N(-0.25, 0.65)	3PL	อิทธิพลของ DIF ในข้อสอบ Anchor ที่มีต่อความไม่แปรเปลี่ยนของประชากรในการ Equating

หมายเหตุ N หมายถึง การแจกแจงแบบปกติ (Normal) LN หมายถึง การแจกแจงแบบ Log-normal และ unif หมายถึง การแจกแจงแบบ Uniform

จากตารางที่ 2 - 8 พบว่า มีการกำหนดพารามิเตอร์ของข้อสอบที่ใช้ในการจำลองข้อมูล 2 ลักษณะ ได้แก่ กำหนดเป็นค่าคงที่และกำหนดเป็นการแจกแจง โดยในการกำหนดให้มีการแจกแจงเป็น 3 รูปแบบ ได้แก่ การแจกแจงแบบปกติ (Normal) การแจกแจงแบบ Log - normal และการแจกแจงแบบ Uniform ซึ่งผู้วิจัยกำหนดรูปแบบและการแจกแจงของพารามิเตอร์ของข้อสอบที่ใช้ในการจำลองข้อมูลโดยมีรายละเอียด ดังนี้

พารามิเตอร์อำนาจจำแนก (a) แม้ทางทฤษฎีจะมีค่าระหว่าง $(-\infty, \infty)$ แต่ค่าอำนาจจำแนกควรมีค่าเป็นบวก ซึ่งตามปกติมีค่าไม่เกิน +2.5 ในทางปฏิบัตินิยมใช้ข้อสอบที่มีค่าระหว่าง +0.5 ถึง +2.5 (ศิริชัย กาญจนาวสี, 2550) อย่างไรก็ตาม การกำหนดให้เป็นการแจกแจงแบบปกติ ต้องจำกัดขอบเขตที่เป็นค่าบวกเป็นต้นไปด้วย ส่วนหากการแจกแจงแบบ Log - normal แม้จะให้ค่าเป็นบวกเสมอ แต่ค่าที่ได้อาจเป็นค่าที่ไม่ได้อยู่ในช่วงที่ต้องการ (เช่น มีค่ามากเกินไป) ดังนั้น ผู้วิจัยจึงเลือกกำหนดพารามิเตอร์อำนาจจำแนก (a_j) ให้มีการแจกแจงแบบ uniform ซึ่งเป็นการแจกแจงที่มีโอกาสการเกิดของตัวแปรสุ่มเท่ากันตลอดช่วงที่เป็นไปได้ของตัวแปรสุ่มจากค่าต่ำสุด (a) ไปถึงค่าสูงสุด (b) ทำให้ค่าที่ได้จะมีค่าอยู่ภายในช่วงที่กำหนด โดยให้ค่าต่ำสุด = 0.5 และค่าสูงสุด = 2.5 เพื่อให้อยู่ช่วงที่นิยมใช้ข้อสอบในทางปฏิบัติ หรือเขียนแทนด้วย $a_j \sim \text{unif}(0.5, 2.5)$

พารามิเตอร์ความยาก (b_j) ผู้วิจัยเลือกกำหนดรูปแบบตามที่งานวิจัยส่วนใหญ่เลือกใช้ นั่นคือ การแจกแจงแบบปกติที่มีค่าเฉลี่ยเป็น 0 และส่วนเบี่ยงเบนมาตรฐานเป็น 1 หรือเขียนแทนด้วย $b_j \sim N(0, 1)$

การกำหนดรูปแบบของเงื่อนไขสำหรับการจำลองข้อมูล

ในการศึกษาครั้งนี้ผู้วิจัยมีจุดประสงค์หลัก คือ เปรียบเทียบวิธีที่ใช้ในการประมาณค่าพารามิเตอร์ และวิธีที่ใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อข้อมูลมีลักษณะของอิทธิพลของเทสต์เลท การแจกแจงความสามารถ จำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบ และอัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบที่แตกต่างกัน ซึ่งอธิบายได้ ดังนี้

1. อิทธิพลของเทสต์เลท (Testlet Effect: $\gamma_{id(j)}$)

Wainer et al. (2000) เปรียบเทียบวิธีที่ใช้ในการการประมาณค่าพารามิเตอร์ ได้แก่ วิธีแมกซิมัมไลค์ลิฮู้ด (Maximum likelihood: ML) และวิธีเบส์โดยใช้เทคนิคมาร์คอฟเชนมอนติคาร์โล (Markov chain monte carlo: MCMC) ของข้อสอบลักษณะเทสต์เลท โดยใช้โมเดลแบบ 3 พารามิเตอร์ ทำการจำลองข้อมูล 3 ลักษณะ ได้แก่ 1) ข้อมูลที่ไม่มีอิทธิพลของเทสต์เลท 2) ข้อมูลที่ทุกอิทธิพลของเทสต์เลทมีค่าอิทธิพลของเทสต์เลทเท่ากัน คือ 0.8 และ 3) ข้อมูลที่แต่ละเทสต์เลทมีค่าอิทธิพลของเทสต์เลทไม่เท่ากัน (0.25, 0.5, 1 และ 2) พบว่า ทุกวิธีการประมาณค่าพารามิเตอร์ให้ความสัมพันธ์ระหว่างค่าจริง (True simulated values) และค่าที่คำนวณ (Estimates simulated values) ระดับสูงมาก สำหรับพารามิเตอร์ความยากและความสามารถ (b, θ) ระดับสูง สำหรับพารามิเตอร์อำนาจจำแนก (a) และระดับปานกลางสำหรับพารามิเตอร์การเดา (c) หมายถึง มีความผิดพลาดเชิงระบบ (Systematic bias) ในการทำนาย อย่างไรก็ตาม ในการใช้งานแบบสอบโดยทั่วไป ข้อสอบมักไม่ได้อยู่ในรูปของอิทธิพลของเทสต์เลททุกข้อ และอาจจะไม่ได้มีค่าของอิทธิพลของเทสต์เลทเท่ากันทุกเทสต์เลท เพื่อให้สอดคล้องกับสถานการณ์จริงมากขึ้น ผู้วิจัยจึงศึกษาข้อสอบที่มีลักษณะของเทสต์เลทผสมกับข้อสอบที่มีความเป็นอิสระหรือไม่ขัดแย้งกับข้อตกลงเบื้องต้น (Independent) ด้วย โดยกำหนดอิทธิพลของเทสต์เลทตามการกำหนดของ Jiao et al. (2013) ซึ่งกำหนดระดับของอิทธิพลของเทสต์เลทเป็นไม่มีอิทธิพลของเทสต์เลท ระดับเล็ก ปานกลาง และระดับใหญ่ หรือให้ค่าอิทธิพลของเทสต์เลท เป็น 0, 0.25, 0.5625 และ 1 ตามลำดับ ซึ่งเมื่ออิทธิพลของเทสต์เลทมีค่าเท่ากับศูนย์ หมายถึง ไม่มีอิทธิพลของเทสต์เลทหรือข้อสอบมีความเป็นอิสระนั่นเอง ดังนั้น ผู้วิจัยจึงศึกษาอิทธิพลของเทสต์เลท ซึ่งประกอบด้วย 3 เงื่อนไข ได้แก่

1.1 แบบสอบที่มี 4 เทสต์เลทและมีค่าอิทธิพลของเทสต์เลทเท่ากันทุกเทสต์เลท (Equal effect) นั่นคือ กำหนดให้อิทธิพลของเทสต์เลทมีค่าเท่ากับ 0.8

1.2 แบบสอบที่มี 4 เทสต์เลทและแต่ละเทสต์เลทมีค่าอิทธิพลของเทสต์เลทไม่เท่ากัน (Unequal effect) โดยกำหนดให้อิทธิพลของเทสต์เลทมีค่าเป็น 0.25, 0.5, 1 และ 2 ตามลำดับ

1.3 แบบสอบประกอบด้วยข้อสอบที่เป็นอิสระ (independent) และเทสต์เลท (Unequal effect แบบ Independent + Testlet) โดยกำหนดให้อิทธิพลของเทสต์เลทมีค่าเป็น 0, 0.25, 0.56 และ 1 ตามลำดับ ซึ่งอิทธิพลของเทสต์เลทมีค่าเท่ากับศูนย์ หมายถึง ข้อสอบเป็นอิสระต่อกัน

2. การแจกแจงของความสามารถ (การแจกแจง θ_i)

การวิจัยครั้งนี้ จำลองข้อมูลด้วยการแจกแจงความสามารถของผู้สอบเป็น 3 รูปแบบ นั่นคือ การแจกแจงแบบปกติ การแจกแจงแบบเบ้ซ้ายและการแจกแจงแบบเบ้ขวา แม้การศึกษาด้านการวัดและวิจัยส่วนใหญ่จะใช้การแจกแจงแบบปกติ แต่สถานการณ์จริงอาจเกิดกรณีที่ความสามารถมีการแจกแจงแบบเบ้ได้ โดยที่ผ่านมามีการศึกษาเกี่ยวกับการแจกแจงความสามารถแบบต่าง ๆ เช่น การศึกษาของ Gorin, Dodd, Fitzpatrick, & Shieh (2005) ใช้การแจกแจงแบบเบต้า โดยกำหนดให้มีการแจกแจงเป็นเบ้ซ้าย (Mean = 0.74, SD = 0.16, Skew = -0.73, Kurtosis = 0) ค่าความสามารถจะถูกพยายามแปลงให้เป็นการแจกแจงแบบปกติ ส่งผลให้ค่าเฉลี่ยมีค่าเป็น 1.5 นอกจากนี้ Shin & Wall (2006) ได้ทำการเปรียบเทียบประสิทธิภาพของวิธีที่ใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ 3 วิธี คือ วิธีแมนเทล-แฮนส์เซลไคสแควร์ (Mantel - Haenszel Chi - square Test : MH - $\chi 2$) วิธี ETS'Mantel - Haenszel Data (MH-D) และวิธี IRT (IRT - DIF) ที่โมเดลแบบ 1 พารามิเตอร์และ 3 พารามิเตอร์ เมื่อความสามารถของผู้สอบมีการแจกแจงแบบปกติ การแจกแจงนอนพาราเมทริกซ์ที่เบ้ซ้าย และการแจกแจงไคร์สแควร์เบ้ขวา ผลการวิจัยพบว่า การแจกแจงความสามารถของผู้สอบที่ไม่เป็นไปตามสมมติฐานนั้น มีผลต่อการประมาณค่า DIF และการตรวจสอบค่า DIF โดยจะทำให้อัตราความผิดพลาดประเภทที่ 1 สูงขึ้น

จากผลการวิจัยที่ผ่านมา ผู้วิจัยจึงคาดว่าการแจกแจงความสามารถแบบเบ้ ซึ่งมีโอกาสเกิดขึ้นกับประชากรของผู้สอบในอนาคต เมื่อมีการปรับปรุงการเรียนการสอนหรือความคุ้นเคยกับรูปแบบการสอบ ก็อาจทำให้ความสามารถเฉลี่ยมีมากขึ้น ทำให้เกิดการแจกแจงแบบเบ้ซ้าย หรือการแจกแจงแบบเบ้ขวาที่หากผู้สอบที่มีความสามารถสูงน้อยกว่าผู้สอบที่มีความสามารถต่ำ นอกจากนี้การแจกแจงแบบเบ้ยังไม่ตรงกับลักษณะการแจกแจงตามสมมติฐานของโมเดล อาจมีผลกระทบต่อการประมาณค่าพารามิเตอร์ ดังนั้นผู้วิจัยจึงสนใจที่จะศึกษาเกี่ยวกับการแจกแจงของความสามารถ ในกรณีต่าง ๆ ประกอบด้วย 3 เงื่อนไขตามการศึกษาของ Welkenhuysen - Gybels (2004) ซึ่งศึกษาประสิทธิภาพของวิธีที่ใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนแบบ 2 ค่า (Dichotomously scored items) โดยให้การแจกแจงความสามารถ

ของแต่ละกลุ่มต่างกัน ผลการทดสอบพบว่า วิธี Logistic regression มีความแกร่งที่สุด โดยกำหนดการแจกแจงความสามารถของแต่ละกลุ่มเป็นดังนี้

1. กลุ่มอ้างอิงและกลุ่มเปรียบเทียบมีการแจกแจงแบบปกติ (Normal distribution) เหมือนกัน นั่นคือ มีพารามิเตอร์เป็น (0,1) หรือเขียนแทนด้วย $\theta_{ir} \sim N(0,1)$ และ $\theta_{if} \sim N(0,1)$

2. กลุ่มอ้างอิงมีการแจกแจงแบบปกติ (Normal distribution) มีพารามิเตอร์เป็น (0,1) และกลุ่มเปรียบเทียบมีการแจกแจงแบบเบ้ขวา (Positively skewed distribution) โดยใช้การแจกแจงแบบเบต้า มีพารามิเตอร์เป็น Beta (1.5, 5) หรือเขียนแทนด้วย $\theta_{ir} \sim N(0,1)$ และ $\theta_{if} \sim$ Beta (1.5, 5)

3. กลุ่มอ้างอิงมีการแจกแจงแบบปกติ (Normal distribution) มีพารามิเตอร์เป็น (0,1) และกลุ่มเปรียบเทียบ มีการแจกแจงแบบเบ้ซ้าย (Negatively skewed distribution) โดยใช้การแจกแจงแบบเบต้า มีพารามิเตอร์เป็น Beta (5, 1.5) หรือเขียนแทนด้วย $\theta_{ir} \sim N(0,1)$ และ $\theta_{if} \sim$ Beta (5, 1.5)

แต่เนื่องจากการแจกแจงเบต้า มีค่าอยู่ในช่วง [0, 1] ดังนั้น ผู้วิจัยจึงใช้การแจกแจงแบบ Skew - normal ซึ่งขยายจากการแจกแจงแบบปกติและมีลักษณะเบ้ ซึ่งครอบคลุมทั้งค่าทางบวกและทางลบ สอดคล้องกับค่าความสามารถที่ผลการวิเคราะห์ส่วนใหญ่มักให้ค่าอยู่ในช่วง -3 ถึง +3 โดยกำหนดค่าพารามิเตอร์ของการแจกแจงจากการศึกษางานวิจัยของ Xu & Jia (2011) ซึ่งศึกษาความไวในการประมาณค่าพารามิเตอร์ความสามารถที่มีการแจกแจงแบบปกติ การแจกแจงแบบ Skew-normal เล็กน้อยหรือมีความเบ้เพียงเล็กน้อย โดยกำหนดให้มีค่าพารามิเตอร์ของการแจกแจงของความสามารถเป็น $\theta_i \sim$ SN(-0.5, 1, -1) และการแจกแจงแบบ Skew - normal มาก หรือมีความเบ้มาก โดยกำหนดให้มีค่าพารามิเตอร์ของการแจกแจงของความสามารถเป็น $\theta_i \sim$ SN(-0.5, 1, -3) ในการวิเคราะห์ใช้โมเดลแบบ 1 และ 2 พารามิเตอร์ ผลการศึกษา พบว่า การประมาณค่าพารามิเตอร์ความสามารถโดยใช้โมเดลแบบ 1 พารามิเตอร์ มีความแกร่งในทุกการแจกแจง ส่วนการใช้โมเดลแบบ 2 พารามิเตอร์ มีความแกร่งอยู่บ้าง แต่จะมีความไวเล็กน้อยในกรณีที่ความสามารถมีความเบ้มาก ประกอบกับการศึกษาของ Monaco (1997) ทำการศึกษาการแจกแจงความสามารถแบบเบ้ที่มีผลกับการทำหน้าที่ต่างกันของข้อสอบ ด้วยการเปรียบเทียบวิธี Mantel Haenszel และ DFIT โดยจำลองข้อมูลผู้สอบในกลุ่มเปรียบเทียบ 1,000 และกลุ่มอ้างอิง 500 คน แบบสอบ 40 ข้อ การแจกแจงความสามารถแบบปกติ เบ้ซ้ายและเบ้ขวา พารามิเตอร์ข้อสอบและผู้สอบ ประมาณค่าด้วยการใช้โปรแกรม BILOG พบว่า วิธี Mantel Haenszel วิธี Lord's chi square และวิธี DFIT สามารถตรวจจับ uniform DIF ได้ดี โดยประสิทธิภาพการตรวจจับดีขึ้นเมื่อจำนวนผู้สอบมากขึ้นในทุกวิธี และเมื่อการแจกแจงพารามิเตอร์ความสามารถมีความเบ้มาก

ส่งผลให้ความถูกต้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบลดลง แต่การแจกแจงแบบเบ้ปานกลาง ไม่ส่งผลต่อการตรวจจับมากนัก

ดังนั้น เพื่อให้สอดคล้องกับค่าความสามารถที่ผลการวิเคราะห์ส่วนใหญ่มักให้ค่าอยู่ในช่วง -3 ถึง +3 และค่าพารามิเตอร์ตามการศึกษาของ Xu & Jia (2011) ผู้วิจัยจึงใช้ค่าพารามิเตอร์ของการแจกแจงพารามิเตอร์ความสามารถ ดังนี้

2.1 การแจกแจงแบบปกติ (Normal) ซึ่งเป็นลักษณะการแจกแจงตามสมมติฐานของโมเดล กำหนดให้ค่าพารามิเตอร์ ของค่าเฉลี่ยเป็น 0 และส่วนเบี่ยงเบนมาตรฐานเป็น 1 หรือเขียนแทนด้วย $\theta_i \sim N(0, 1)$

2.2 การแจกแจงแบบเบ้ขวา (Positively skewed ability distribution) โดยให้การแจกแจงของความสามารถในการจำลองข้อมูลมีค่าพารามิเตอร์ Location = -0.5 Scale = 1 และ Slan = 3 หรือเขียนแทนด้วย $\theta_i \sim SN(-0.5, 1, 3)$

2.3 การแจกแจงแบบเบ้ซ้าย (Negatively skewed ability distribution) โดยให้การแจกแจงของความสามารถในการจำลองข้อมูลมีค่าพารามิเตอร์ Location = 0.5 Scale = 1 และ Slan = -3 หรือเขียนแทนด้วย $\theta_i \sim SN(0.5, 1, -3)$

3. จำนวนข้อสอบที่มีการทำหน้าที่ต่างกัน (ร้อยละข้อที่ DIF)

จากการศึกษาที่ผ่านมา พบว่า มีการกำหนดสัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันในอัตราส่วนต่าง ๆ เช่น Okan Bulut (2015) ศึกษาความเป็นเพศที่ทำให้เกิดการทำหน้าที่กันของข้อสอบอันเนื่องจากมาจากอิทธิพลของ Test Booklet โดยใช้ข้อมูลที่เป็นการสอบเข้าศึกษาต่อระดับบัณฑิตศึกษาในประเทศตุรกี ปี 2010 (Entrance examination for graduate studies: EEGS) จำนวน 80 ข้อ มีจำนวนผู้เข้าสอบ 142,178 คน ใช้กลุ่มตัวอย่าง 20,000 คน โดยใช้วิธี Mantel-Haenszel (MH) และวิธี Breslow - Day (BD) ผลการศึกษาพบว่า วิธี MH ตรวจพบการทำหน้าที่ต่างกันของข้อสอบ จำนวน 13 ข้อ (ร้อยละ 16.25) วิธี BD ตรวจพบการทำหน้าที่ต่างกันของข้อสอบ จำนวน 15 ข้อ (ร้อยละ 18.75)

Burhanettin Özdemir (2015) เปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี Lord's Chi-Square วิธี Raju's Area และวิธี Likelihood - Ratio Test โดยใช้ข้อมูลการสอบโครงการ TIMSS 2011 วิชาคณิตศาสตร์ เล่มที่ 2 (Booklet 2) จำนวน 22 ข้อ ของนักเรียนเกรด 8 จำนวน 488 คน พบว่า วิธี Lord's Chi - Square และวิธี Likelihood - Ratio Test ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน 2 ข้อเท่ากัน (ร้อยละ 9) และวิธี Raju's Area ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน 4 ข้อ (ร้อยละ 18)

อัญชลี ธีระวุฒิ (2555) ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบทดสอบ Pre O - NET วิชาคณิตศาสตร์ ชั้นมัธยมศึกษาปีที่ 6 ของสำนักงานเขตพื้นที่การศึกษามัธยมศึกษา เขต 35 จำนวน 40 ข้อ ของนักเรียนจำนวน 1,500 คน ด้วยวิธีซิบเทสท์และวิธีแมนเทล - แฮนส์เซล ผลการวิจัยพบว่า เมื่อจำแนกตามเพศ วิธีซิบเทสท์ตรวจสอบพบข้อสอบทำหน้าที่ต่างกัน 8 ข้อ (ร้อยละ 20) วิธีแมนเทล - แฮนส์เซลตรวจสอบพบข้อสอบทำหน้าที่ต่างกัน 7 ข้อ (ร้อยละ 17.5) เมื่อจำแนกตามเขตที่ตั้งของโรงเรียน วิธีซิบเทสท์พบข้อสอบทำหน้าที่ต่างกัน 1 ข้อ (ร้อยละ 2.5) วิธีแมนเทล - แฮนส์เซลตรวจสอบพบข้อสอบทำหน้าที่ต่างกัน 4 ข้อ (ร้อยละ 10) และเมื่อจำแนกตามขนาดของโรงเรียน วิธีซิบเทสท์และวิธีแมนเทล-แฮนส์เซล ตรวจสอบพบข้อสอบทำหน้าที่ต่างกัน 3 ข้อ (ร้อยละ 7.5) เท่ากัน

สุพัฒนา หอมบุปผา (2556) เปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN โดยใช้ข้อมูลจากการสอบแบบวัดผลสัมฤทธิ์ทางการเรียน เพื่อประเมินคุณภาพการศึกษาระดับชาติ ปีการศึกษา 2553 (O - Net) วิชาภาษาไทย คณิตศาสตร์ และวิทยาศาสตร์ วิชาละ 30 ข้อ กลุ่มตัวอย่าง 1,000 คน เมื่อพิจารณาในส่วนของการประมาณค่าด้วยวิธี BAYESIAN พบว่า สำหรับวิชาภาษาไทย เมื่อจำแนกตามเพศพบข้อสอบที่ทำหน้าที่ต่างกัน 5 ข้อ เมื่อจำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียนพบข้อสอบที่ทำหน้าที่ต่างกัน 5 ข้อ สำหรับวิชาคณิตศาสตร์ เมื่อจำแนกตามเพศ พบข้อสอบที่ทำหน้าที่ต่างกัน 2 ข้อ เมื่อจำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน พบข้อสอบที่ทำหน้าที่ต่างกัน 1 ข้อ และวิชาวิทยาศาสตร์ เมื่อจำแนกตามเพศ พบข้อสอบที่ทำหน้าที่ต่างกัน 7 ข้อ เมื่อจำแนกตามสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน พบข้อสอบที่ทำหน้าที่ต่างกัน 2 ข้อ นั่นคือ พบข้อสอบที่ทำหน้าที่ต่างกันระหว่าง 1 - 7 ข้อ หรือระหว่างร้อยละ 3.3 - 23.3

Narayanan & Swaminathan (1994) ได้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป (Uniform) พบว่า สำหรับการตรวจสอบด้วยวิธีซิปเทสท์และวิธีการถดถอยโลจิสติกส์ เมื่อมีจำนวนข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบในแบบสอบลดลง จะได้อัตราความคาดเคลื่อนประเภทที่ 1 (Type I error) และอำนาจทดสอบ (Power) มากขึ้น

Lee et al. (2009) ศึกษาอำนาจการทดสอบและอัตราความคาดเคลื่อนประเภทที่ 1 ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีลักษณะของเทสต์เลท โดยใช้วิธี SIBTEST และ Poly - SIBTEST และศึกษาสัดส่วนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบ กำหนดให้มีข้อ DIF ร้อยละ 0, 10 และ 20 พบว่า ร้อยละของข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบที่แตกต่างกันส่งผลต่ออำนาจการทดสอบอย่างไม่มีแบบแผน

อย่างไรก็ตาม ผลการศึกษาของ Narayanan & Sawaminathan (1996, อ้างถึงใน สิริรัตน์ วิภาสศิลป์, 2545) พบว่า สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบในแบบสอบมีผลต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ถ้ามีข้อสอบแสดงการทำหน้าที่ต่างกันของข้อสอบปริมาณมาก จะทำให้ความถูกต้องในการตรวจสอบลดลง และสิริรัตน์ วิภาสศิลป์ (2545) พบว่า หากสัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบมากกว่าร้อยละ 20 จะทำให้มีการระบุผิดพลาดในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูง

จากข้อมูลการวิจัยที่ผ่านมา พบว่า ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจากแบบสอบต่าง ๆ มีการตรวจพบข้อสอบที่ทำหน้าที่ต่างกันระหว่าง ร้อยละ 2.5 - 23.3 ประกอบกับการที่แบบสอบมีข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบมาก จะทำให้มีความน่าเชื่อถือในการประมาณค่าความสามารถลดลง ดังนั้น อำนาจการตรวจสอบ (Power) การทำหน้าที่ต่างกันของข้อสอบก็น่าจะลดลงด้วย โดยที่หากแบบสอบมาตรฐานมีข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบร้อยละ 10 - 15 ถือว่าไม่ผิดปกติ แต่หากมีข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบในแบบสอบถึงร้อยละ 20 ถือว่าผิดพลาดมาก (Worst case) (Clauser, 1993 cited in Narayanan & Swaminathan, 1994) ดังนั้น ผู้วิจัยจึงกำหนดเงื่อนไขจำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบ ประกอบด้วย 3 เงื่อนไข ได้แก่

3.1 ร้อยละ 0 หรือไม่มีข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบ

3.2 ร้อยละ 12.5 หรือมีข้อสอบที่ทำหน้าที่ต่างกันจำนวน 5 ข้อ ในแบบสอบ โดยผู้วิจัยเลือกจากจุดกลางของช่วง 10 - 15 นั่นคือ 12.5 เพื่อกำหนดให้เป็นเงื่อนไขตามแบบสอบที่มีข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบที่ถือว่าไม่ผิดปกติ

3.3 ร้อยละ 20 หรือมีข้อสอบที่ทำหน้าที่ต่างกันจำนวน 8 ข้อ ในแบบสอบ เพื่อกำหนดให้เป็นเงื่อนไขตามแบบสอบที่มีข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบที่ถือว่าผิดพลาดมาก

4. อัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ (อัตราส่วน R: F)

จากการศึกษาที่ผ่านมา พบว่า มีการศึกษาเกี่ยวกับอัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ ซึ่งมีผลต่อวิธีที่ใช้ในการตรวจสอบการทำหน้าที่ต่างกันวิธีต่าง ๆ เช่น จิตติมา วรรณศรี (2539) ทำการเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี Mantel - Heanszel (MH) กับวิธี SIBTEST เมื่อขนาดกลุ่มตัวอย่างย่อยระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบเป็นอัตราส่วนต่อกันเป็น 1: 1, 1: 0.9, 1: 0.75 และ 1: 0.5 ภายใต้ขนาดกลุ่มตัวอย่างต่างกันที่ระดับความยากแบบสอบเดียวกัน และระดับความยาวแบบสอบต่างกัน พบว่า ขนาดกลุ่มตัวอย่างที่เป็นเงื่อนไขที่ดีที่สุด คือ เมื่อกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ

มีขนาด 1,000 คน (อัตราส่วน 1: 1) และเงื่อนไขที่ดีรองลงมาคือ การใช้กลุ่มตัวอย่างขนาด 200 คน มีอัตราส่วนระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบเท่ากับ 1: 0.75 นอกจากนี้ Ankenmann, Witt, & Dunbar (1999) ศึกษาการตรวจสอบการทำหน้าที่ต่างกันด้วยวิธี Likelihood Ratio Goodness of Fit Statistics (LR) พบว่า อัตราความคาดเคลื่อนชนิดที่ 1 (Type I error) ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมีค่ามากขึ้น เมื่อขนาดตัวอย่าง 500/ 500 และ Gierl, Bisanz, Boughton & Khaliq (2001 cited in Awuor, 2008) เสนอว่า หากผู้สอบกลุ่มอ้างอิง และกลุ่มเปรียบเทียบมีความสามารถระดับเดียวกันแล้ว และขนาดตัวอย่างที่ต่างกันของกลุ่มอ้างอิง และกลุ่มเปรียบเทียบแตกต่างกันไม่มาก ก็ไม่เป็นอุปสรรคในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบของผู้สอบ

นอกจากนี้ Awuor (2008) ศึกษาผลกระทบต่ออำนาจ (Power) ในการตรวจสอบการทำหน้าที่ ด้วยวิธี SIBTEST และวิธี Mantel - Haenszel (MH) ในระดับต่าง ๆ ผลการศึกษาพบว่า อัตราความคลาดเคลื่อนประเภทที่ 1 (Type I error rate) ของตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมีมากขึ้น เมื่ออัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบมีอัตราส่วนต่างกันมาก (1: 0.1) และควรใช้อัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบไม่น้อยกว่า 1: 0.5 และจากผลการศึกษาของจิติมา วรรณศรี (2539) พบว่า อัตราส่วน 1: 1 เป็นอัตราส่วนที่ให้ประสิทธิภาพในการตรวจสอบดีที่สุด แต่ในทางปฏิบัติการสอบคัดเลือกเข้าศึกษาต่อในสถาบันการศึกษาต่าง ๆ มักมีอัตราส่วนของประชากรระหว่างอัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบที่แตกต่างกัน

จะเห็นว่า การศึกษาผลกระทบต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเมื่ออัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบต่างกันยังมีความคลุมเคลือ ผู้วิจัยจึงสนใจศึกษาขนาดของตัวอย่างที่มีอัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ ทั้งในอัตราส่วนที่เท่ากันและไม่เท่ากัน โดยกำหนดขนาดตัวอย่าง 2 ขนาด ดังนั้น ผู้วิจัยจึงกำหนดให้มีอัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบเป็น 2 เงื่อนไข ได้แก่

4.1 อัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ เป็น 1: 1 หรือ มีกลุ่มอ้างอิง จำนวน 1,000 คน และกลุ่มเปรียบเทียบ จำนวน 1,000 คน

4.2 อัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ เป็น 1: 0.1 หรือ มีกลุ่มอ้างอิง จำนวน 1,000 คน และกลุ่มเปรียบเทียบ จำนวน 100 คน

สำหรับการวิเคราะห์โดยใช้วิธีของเบส์ (Bayes) และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) นั้น จำเป็นต้องมีการกำหนดการแจกแจงก่อนของพารามิเตอร์ การลู่เข้าและจำนวนรอบซึ่งมีรายละเอียด ดังนี้

การแจกแจงก่อนของพารามิเตอร์ (Prior distribution)

การวิเคราะห์โดยใช้วิธีของเบส์ด้วยการ MCMC นั้น ถูกนำมาใช้มากขึ้น จากการศึกษาที่ผ่านมา พบว่า วิธีการนี้มีความถูกต้องและสามารถประมาณค่ากับกลุ่มตัวอย่างขนาดเล็กได้ โดยวิเคราะห์ด้วยวิธีของเบส์ ต้องอาศัยการแจกแจงก่อนของพารามิเตอร์ที่สนใจ โดยต้องระบุในโมเดลที่ใช้วิเคราะห์ด้วย แต่การแจกแจงก่อนของพารามิเตอร์อาจเป็นได้ทั้งการแจกแจงก่อนที่ให้ข้อมูล (Informative prior distribution) และการแจกแจงก่อนที่ไม่ให้ข้อมูล (Noninformative prior distribution) การให้ข้อมูลการแจกแจงก่อนอย่างเพียงพอ มีผลต่อการประมาณค่าพารามิเตอร์ โดยทำให้เกิดความแปรปรวนน้อยและเข้าสู่ค่าเฉลี่ย (มีการลู่เข้าเร็วขึ้น) แต่บางกรณีก็มีการให้ข้อมูลการแจกแจงในอดีตที่มีลักษณะคลุมเครือ (Vague) หรือเป็นการแจกแจงก่อนที่ไม่ให้ข้อมูล ทำให้ยากต่อการระบุการแจกแจงก่อนที่แท้จริงของพารามิเตอร์ที่สนใจ จากการศึกษางานวิจัยที่ผ่านมา พบว่า การกำหนดการแจกแจงก่อนของพารามิเตอร์ข้อสอบและผู้สอบในโมเดลของ IRT มักมีความแตกต่างกัน โดยที่ส่วนใหญ่มักกำหนดให้ความสามารถมีการแจกแจงแบบปกติ หรือเขียนแทนด้วย $\theta_i \sim N(0,1)$ เป็นกำหนดช่วงของความสามารถให้คงที่เพื่อให้แน่ใจว่าสามารถระบุตำแหน่งของโค้งได้ (Ra, 2011)

พารามิเตอร์อำนาจจำแนก (a_j) มีทั้งการกำหนดให้มีการแจกแจงก่อนที่ให้ข้อมูลและการแจกแจงก่อนที่ไม่ให้ข้อมูล ทั้งนี้ ที่สามารถกำหนดให้เป็นการแจกแจงก่อนที่ให้ข้อมูลได้ เนื่องจาก a_j แสดงถึง ผู้ที่มีความสามารถสูงจะตอบข้อสอบถูก ดังนั้น a_j จึงกำหนดให้มีค่ามากกว่า 0 จากการศึกษางานวิจัยที่ผ่านมา พบว่า การกำหนดการแจกแจงก่อนของ a_j ส่วนใหญ่จะกำหนดเป็น $a_j \sim N(\mu_a, \sigma_a^2)$ (Johnson & Albert, 1999; Wang & Wilson, 2005 a, 2005 b; Wang et al., 2002) นอกจากนี้ยังมีการกำหนดการแจกแจงก่อนของ a_j แบบอื่น ๆ อีก เช่น กำหนดให้ a_j มีค่าตั้งแต่ศูนย์ไปทางบวก สำหรับการแจกแจงปกติ หรือ $a_j \sim N(\mu_a, \sigma_a^2)I(0,)$ (Fukuhara & Kamata, 2011; Li et al., 2006) นอกจากนี้ มีการกำหนด a_j ให้มีการแจกแจงแบบ Log-normal หรือ $a_j \sim LN(\mu_a, \sigma_a^2)$ (Sahu, 2002) และกำหนด a_j ให้มีการแจกแจงแบบ Gamma และ Inverted Gamma หรือ $a_j \sim IG(m, n)$ (Bafumi, Gelman, Park & Kaplan, 2005)

สำหรับพารามิเตอร์ความยาก (b_j) โดยทั่วไปกำหนดให้มีการแจกแจงก่อนแบบปกติ หรือ $b_j \sim N(\mu_b, \sigma_b^2)$ อย่างไรก็ตาม Patz & Junker (1999) กำหนดให้ μ_b มีการแจกแจงแบบ Uniform และ σ_b^2 มีการแจกแจงแบบ Inverse chi - square ด้วย

และการแจกแจงก่อนของเทสต์เลท จากการศึกษาของ Bradlow et al. (1999) กำหนดการแจกแจงก่อนแบบไม่ให้ข้อมูล ดังนั้น การแจกแจงของพารามิเตอร์ที่สนใจจึงกำหนดโดยข้อมูลที่วิเคราะห์ อย่างไรก็ตาม เนื่องจากแต่ละเทสต์เลทถือเป็นอิสระต่อกัน ดังนั้น งานวิจัยส่วนใหญ่

จึงมักกำหนดให้ $\gamma_{id(j)} \sim N\left(0, \sigma^2_{\gamma_{d(j)}}\right)$ (Wang & Wilson, 2005 a, 2005 b; Fukuhara & Kamata, 2011; Sedivy, 2009)

จากการศึกษาที่ผ่านมา ผู้วิจัยจึงกำหนดการแจกแจงของพารามิเตอร์ ดังตาราง 2 - 9

ตารางที่ 2 - 9 การแจกแจงของพารามิเตอร์

พารามิเตอร์	Prior Distributions ของพารามิเตอร์	Hyperparameter ของพารามิเตอร์	Prior Distributions ของ Hyperparameter
ζ_i	$N(0,1)$		
a_j	$N(\mu_a, \sigma_a^2)I(0,\alpha)$	μ_a	$N(0,1000)$
		σ_a^2	$Inv-\chi^2(0.5)$
b_j	$N(\mu_b, \sigma_b^2)$	μ_b	$N(0,1000)$
		σ_b^2	$Inv-\chi^2(0.5)$
β_j	$N(\mu_\beta, \sigma_\beta^2)$	μ_β	$N(0,1000)$
		σ_β^2	$Inv-\chi^2(0.5)$
β_θ	$N\left(\mu_{\beta_\theta}, \sigma^2_{\beta_\theta}\right)$	μ_{β_θ}	$N(0,1000)$
		$\sigma^2_{\beta_\theta}$	$Inv-\chi^2(0.5)$
$\gamma_{id(j)}$	$N(0, \sigma^2_{\gamma_d})$	$\sigma^2_{\gamma_d}$	$Inv-\chi^2(0.5)$

การลู่เข้าและจำนวนรอบ (Convergence and iterations)

การวิเคราะห์โดยใช้วิธีของเบส์ จำเป็นต้องมีการตัดข้อมูลส่วนแรกทิ้ง เนื่องจากการประมาณค่ารอบแรก ๆ ค่าพารามิเตอร์ที่ต้องการประมาณค่ายังมีการแกว่งตัว ไม่ลู่เข้าค่าใดค่าหนึ่ง จึงต้องทำการตัดข้อมูลส่วนแรกออก หรือเรียกว่า Burn - in

การตัดสินเกี่ยวกับการลู่เข้านั้น ตรวจสอบได้จากหลายแหล่ง ที่นิยมใช้ ได้แก่ History plot ที่แสดงค่าที่ประมาณได้ในแต่ละรอบ โดยหากกราฟมีแนวโน้มว่ามีการประมาณค่าคงที่จะถือว่าลู่เข้า นอกจากนี้ ยังมีดัชนีที่นิยมใช้ตรวจสอบอีก คือ Density plot และ Autocorrelation plot

การศึกษาเกี่ยวกับการประมาณค่าโมเดลของเทสต์เลทด้วยวิธีของเบส์นั้น การกำหนดจำนวนรอบมีอิทธิพลโดยตรงกับความแม่นยำในการประมาณค่าพารามิเตอร์ โดยหากกำหนดจำนวนรอบน้อยจะทำให้ความแปรปรวนของการสุ่มมีมากเพียงพอที่จะให้ทำให้การประมาณค่าพารามิเตอร์มีความลำเอียงมาก ซึ่งจะส่งผลความเที่ยงและความน่าเชื่อถือของผลการวิจัยที่ได้ต่ำมาก โดยมีนักวิจัยเสนอแนะเกี่ยวกับการกำหนดจำนวนรอบและจำนวนการ Burn - in ต่าง ๆ กัน เช่น Gelman, Carlin, Stern, & Rubin (2003) เสนอว่าควรใช้จำนวนการ Burn - in ครึ่งแรกของจำนวนการวนรอบทั้งหมด (Iteration) Raftery & Lewis (1992) เสนอว่าควรมีจำนวนการ Burn - in

ไม่ต่ำกว่า 500 ครั้ง เป็นต้น และการศึกษาที่ผ่านมามีการกำหนดจำนวนรอบและจำนวนการ Burn - in เช่นตารางที่ 2 - 10

ตารางที่ 2 - 10 การกำหนดจำนวนรอบและจำนวนการ Burn - in ของงานวิจัยที่ผ่านมา

การศึกษา	ขนาดตัวอย่าง	ขนาดแบบสอบ	จำนวนรอบ	จำนวนครั้ง	จำนวนพารามิเตอร์
Wang et al. (2002)	1,000	30 ข้อ, 12 ข้อเดี่ยว (3 ข้อ: 6 Testlet, 6 ข้อ: 3 Testlet, 9 ข้อ: 2 Testlet)	Iteration: 3,000 Burn - in: 2,000	5	3PL
Li et al. (2006)	2,000	20 ข้อ(5 ข้อ: 4 Testlet)	Iteration: 15,000 Burn - in: 1,000	NA	2PL
DeMars (2006)	2,000	25 ข้อ (5 ข้อ: 5 Testlet) 50 ข้อ (5 ข้อ: 10 Testlet)	Iteration: 3,000 Burn - in: 1,000	100	3PL
Bao (2007)	5,000	50 ข้อ 30 ข้อเดี่ยว (10 ข้อ: 2 Testlet)	Iteration: 4,000 Burn - in: 1,500	10	2PL
Sedivy (2009)	500 1,000 4,000	30 ข้อ (5 ข้อ: 6 Testlet, 10 ข้อ: 3 Testlet) (15 ข้อ: 2 Testlet)	Iteration: 2,000 Burn - in: 500	100	2PL
Fukuhara & Kamata (2011)	1,000	42 ข้อ (6 ข้อ: 7 Testlet)	Iteration: 15,000 Burn - in: 5,000	100	2PL
Ra (2011)	1,000 2,000	30 ข้อ (3 ข้อ: 10 Testlet, 6 ข้อ: 5 Testlet) 30 ข้อ (10 ข้อ: 3 Testlet)	Iteration: 50,000 Burn - in: 10,000	25	3PL
Glas (2012)	1,000	20 ข้อ (5 ข้อ: 4 Testlet)	Iteration: 4,000 Burn - in: 1,000	40	2PL

ตารางที่ 2 - 10 (ต่อ)

การศึกษา	ขนาดตัวอย่าง	ขนาดแบบสอบ	จำนวนรอบ	จำนวนครั้ง	จำนวนพารามิเตอร์
Tao, Xu, Shi, & Jiao (2013)	100	10 ข้อ (5 ข้อ: 2 Testlet)	Iteration: 30,000 Burn - in: 5,000	100	2PL
Jiao et al. (2013)	1,000	54 ข้อ (9 ข้อ: 6 Testlet)	Iteration: 2,000 Burn - in: 1,000	25	1PL
Eckes (2014)	2,859 2,214	25 ข้อ (Testlet 1 มี 8 ข้อ Testlet 2 มี 10 ข้อ Testlet 3 มี 7 ข้อ)	Iteration: 4,000 Burn - in: 3,000	ข้อมูลจริง	2PL

เนื่องจาก ผู้วิจัยใช้โมเดลการวิเคราะห์ของ Fukuhara & Kamata (2011) นั่นคือ Bifactor Multidimensional Item Response Theory Model ดังนั้น จึงเลือกกำหนดค่าตาม Fukuhara & Kamata (2011) นั่นคือให้ทำการประมาณค่าซ้ำในแต่ละชุดข้อมูล จำนวน 20,000 รอบ (Number of Iteration) และตัดข้อมูลส่วนแรกออก 5,000 รอบ (Number of Burn-in) เมื่อทำการประมวลผลจะได้ตัวอย่าง กราฟ History กราฟ Density และกราฟ Autocorrelation ดังภาคผนวก ก - ง

การวิเคราะห์สำหรับการวิจัยในครั้งนี้ มีการวิเคราะห์ข้อมูลที่ได้จากการสุ่มโดยการทำกลุ่มข้อมูลซ้ำ 100 ชุดข้อมูล (Data set) และการประมวลผลด้วยวิธี Bayes และ Bayesγ ทำการประมาณค่าซ้ำในแต่ละชุดข้อมูล 20,000 รอบ (Number of iteration) และตัดข้อมูลส่วนแรก 5,000 รอบ (Number of burn-in) เนื่องจากการประมาณค่าในรอบแรก ๆ ค่าพารามิเตอร์ที่ต้องการการประมาณค่ายังมีการแกว่งตัวไม่ลู่เข้าค่าใดค่าหนึ่ง

ตอนที่ 6 งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวข้องกับการประมาณค่าพารามิเตอร์ของข้อมูลแบบเทสต์เลท

สุนทร เทียนงาม และคณะ (2553) ศึกษาผลของระดับความไม่เป็นอิสระของข้อสอบต่อค่าความเที่ยงของแบบสอบ ค่าพารามิเตอร์ของข้อสอบ ค่าความสามารถของผู้สอบและสารสนเทศของแบบสอบด้วยโมเดลทฤษฎีการตอบสนองข้อสอบ ที่มีจำนวนข้อสอบและขนาดกลุ่มตัวอย่างแตกต่างกัน และวิเคราะห์ลักษณะความไม่เป็นอิสระของข้อสอบจากแบบสอบจริง

โดยการจำลองข้อมูลด้วยเทคนิค MC จำนวนตัวแปรประกอบด้วย ระดับความไม่เป็นอิสระของข้อสอบ 9 ระดับโดยแบ่งเป็นกลุ่มย่อย 3 ระดับ (ระดับต่ำ ปานกลางและสูง) โมเดลทฤษฎีการตอบสนองข้อสอบ 3 โมเดล (1PL, 2PL และ 3PL) จำนวนข้อสอบ 3 กลุ่ม (30, 50 และ80 ข้อ) และขนาดกลุ่มตัวอย่าง 3 กลุ่ม (400, 800 และ1200 คน) รวมจำนวนเงื่อนไขทั้งหมด 243 เงื่อนไข และกำหนดจำนวนรอบที่ใช้การประมาณค่าพารามิเตอร์ในแต่ละเงื่อนไข 1,000 รอบ และใช้วิธีการศึกษาจากข้อมูลจริงในวิชาคณิตศาสตร์และภาษาอังกฤษ โปรแกรมที่ใช้ในการจำลองข้อมูล คือ Fortran Power Station 4.0 และโปรแกรมในการตรวจสอบความไม่เป็นอิสระของข้อสอบ คือ LDID (A computer program for local dependence indices for dichotomous item version 1) ผลการวิจัยสรุปได้ดังนี้

1. เมื่อระดับความไม่เป็นอิสระของข้อสอบเพิ่มสูงขึ้น ค่าความเที่ยง (KR20) จะสูงขึ้นในทุกเงื่อนไขของการทดสอบ โดยเมื่อใช้จำนวนข้อสอบ 30 ข้อ ค่าความเที่ยงต่ำสุด - สูงสุด มีค่า 0.671 - 0.740 จำนวนข้อสอบ 50 ข้อ ค่าความเที่ยงต่ำสุด-สูงสุด มีค่า 0.773 - 0.828 และจำนวนข้อสอบ 80 ข้อ ค่าความเที่ยงต่ำสุด - สูงสุดมีค่า 0.845 - 0.886

2. เมื่อวิเคราะห์ข้อสอบตามแนวทฤษฎีการตอบสนองข้อสอบ มีจำนวนผู้สอบ 400 800 และ 1200 คน ทำข้อสอบ 30 50 และ 80 ข้อ ตามลำดับ พบว่า

2.1 เมื่อวิเคราะห์โดยใช้โมเดล 1 พารามิเตอร์ พบว่า เมื่อระดับความไม่เป็นอิสระของข้อสอบเพิ่มสูงขึ้น ค่าความยากและค่าความสามารถไม่แตกต่างกัน และค่าสารสนเทศของแบบสอบส่วนใหญ่มีแนวโน้มคงที่

2.2 เมื่อวิเคราะห์โดยใช้โมเดล 2 พารามิเตอร์ พบว่า ส่วนใหญ่ค่าความยากลดลง แต่ค่าอำนาจจำแนก ค่าความสามารถและสารสนเทศของแบบสอบมีแนวโน้มสูงขึ้น โดยส่วนใหญ่แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05

2.3 เมื่อวิเคราะห์โดยใช้โมเดล 3 พารามิเตอร์ พบว่าค่าความยาก ค่าความสามารถและสารสนเทศของแบบสอบมีแนวโน้มลดลง แต่สำหรับค่าอำนาจจำแนก ค่าการเดาส่วนใหญ่มีแนวโน้มสูงขึ้นและแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05

3. เมื่อวิเคราะห์ความไม่เป็นอิสระของข้อสอบโดยใช้ข้อมูลจริงด้วยโมเดล 1, 2 และ 3 พารามิเตอร์ สำหรับวิชาคณิตศาสตร์ พบว่า ค่าเฉลี่ย $|Q3|$ มีค่าเท่ากับ 0.035 0.026 0.022 ตามลำดับ และในวิชาภาษาอังกฤษ พบว่า ค่าเฉลี่ย $|Q3|$ มีค่าเท่ากับ 0.033 0.028 0.022 ตามลำดับ มีค่าใกล้เคียงกับค่าคาดหวังของ $|Q3|$ ซึ่งมีค่าเท่ากับ .025 แสดงว่า ข้อสอบส่วนใหญ่มีความไม่เป็นอิสระของข้อสอบเล็กน้อย

Cheng (1996) เปรียบเทียบการประมาณค่าของการให้คะแนนแบบ Testlet และการให้คะแนนแบบรายข้อ โดยมีวัตถุประสงค์เพื่อศึกษาอิทธิพลของตัวแปร 3 ตัวแปร ได้แก่ ระดับความเป็นอิสระกันของข้อสอบภายใน Testlet (สูง กลาง ต่ำ และไม่มี) จำนวนข้อสอบในเทสต์เลท (4 ข้อและ 8 ข้อ) จำนวนเทสต์เลทในแบบสอบ (1, 2, 4, 6) ที่มีต่อประสิทธิภาพการให้คะแนนแบบเทสต์เลทและการให้คะแนนแบบรายข้อ ในการประมาณค่าพารามิเตอร์ใช้โปรแกรม MULTILOG (6.0) โดยใช้ Graded Response Model (GRM) มาประยุกต์ใช้ โดยรวมคะแนนในแต่ละเทสต์เลทจะได้คะแนนที่มีลักษณะเป็นแบบ Polytomous และ 2 - parameter Logistic (2 PL) Model ในการคำนวณแบบรายข้อ ในการศึกษาใช้วิธีการตรวจสอบความเป็นอิสระกันของข้อสอบ 2 วิธี ได้แก่ Stout's T procedure และ Yen's Q3 statistic ผลการศึกษา พบว่า

1) เมื่อพิจารณาตัวแปรจำนวนเทสต์เลทในแบบสอบ พบว่า การประมาณค่าความสามารถของการให้คะแนนทั้งสองแบบไม่แตกต่างกัน ยกเว้นกรณีที่จำนวนเทสต์เลทในแบบสอบเท่ากับ 6 การให้คะแนนแบบรายข้อจะมีประสิทธิภาพของการประมาณค่ามากกว่าการให้คะแนนแบบเทสต์เลท ซึ่งสาเหตุที่การให้คะแนนแบบเทสต์เลทไม่ได้ดีไปกว่าการให้คะแนนแบบรายข้อ เนื่องจากข้อมูลที่นำมาใช้ในการวิเคราะห์บางกรณี อาจมีโครงสร้างที่เป็นหลายมิติ (Multidimensional)

2) การให้คะแนนแบบรายข้อจะมีประสิทธิภาพของการประมาณค่ามากกว่าการให้คะแนนแบบเทสต์เลท เมื่อระดับความเป็นอิสระกันของข้อสอบภายในเทสต์เลทต่ำ และการให้คะแนนแบบเทสต์เลท จะมีประสิทธิภาพของการประมาณค่ามากกว่าการให้คะแนนแบบรายข้อ เมื่อระดับความเป็นอิสระกันของข้อสอบภายในเทสต์เลทสูง

3) การวิเคราะห์สารสนเทศของแบบสอบ (Test information) พบว่า เมื่อสัดส่วนของข้อสอบเทสต์เลทเพิ่มขึ้น การให้คะแนนแบบรายข้อสามารถให้สารสนเทศของแบบสอบได้มากกว่าการให้คะแนนแบบเทสต์เลท ทั้งนี้ Cheng (1996) ได้อภิปรายว่า อาจเป็นเพราะการวิเคราะห์การให้คะแนนแบบเทสต์เลทเป็นการรวมคะแนนในเทสต์เลทเข้าด้วยกันเป็นข้อสอบ 1 ข้อ ทำให้ในแบบสอบ มีจำนวนข้อสอบน้อย เช่น กรณี 4 เทสต์เลท จะได้ 4 ข้อ ซึ่งอาจเป็นผลจากความไม่เพียงพอของข้อมูล ทำให้การวิเคราะห์โค้งความคาดเคลื่อนมาตรฐานของการให้คะแนนแบบเทสต์เลทมีความไม่ชัดเจน (Blurriness) เมื่อเทียบกับโค้งความคาดเคลื่อนมาตรฐานที่มีการให้คะแนนแบบรายข้อ อย่างไรก็ตาม การให้คะแนนแบบรายข้อนำไปสู่การประมาณค่าสารสนเทศแบบสอบที่เกินจริง โดยประมาณค่าเกินจริงมากขึ้นเมื่อสัดส่วนของระดับความเป็นอิสระกันของข้อสอบภายในเทสต์เลทสูงขึ้น

Wainer et al. (2000) เปรียบเทียบวิธีที่ใช้ในการการประมาณค่าพารามิเตอร์ 4 วิธี ได้แก่ วิธีแมกซิมัมไลค์ลิฮู้ด (Maximum likelihood: ML) วิธีเบส์โดยไม่คำนึงถึงอิทธิพลของเทสต์เลท วิธีเบส์โดยกำหนดให้อิทธิพลของเทสต์เลทคงที่ (σ_γ^2) และวิธีเบส์แบบมีอิทธิพลเทสต์เลทต่างกัน ($\sigma_{d(j)}^2$) ของข้อสอบลักษณะเทสต์เลท โดยใช้โมเดลแบบ 3 พารามิเตอร์ ทำการจำลองข้อมูล 3 ลักษณะ ได้แก่ 1) ข้อมูลที่ไม่มีอิทธิพลของเทสต์เลท 2) ข้อมูลที่ทุกเทสต์เลทมีค่าอิทธิพลของเทสต์เลทเท่ากัน คือ 0.8 และ 3) ข้อมูลที่แต่ละเทสต์เลทมีค่าอิทธิพลของเทสต์เลทไม่เท่ากัน (0.25, 0.5, 1 และ 2) พบว่า

1. ทุกวิธีการประมาณค่าพารามิเตอร์ให้ความสัมพันธ์ระหว่างค่าจริง (True simulated values) และค่าที่คำนวณ (Estimates simulated values) ในระดับสูงมาก สำหรับพารามิเตอร์ความยากและความสามารถ (b, θ) ระดับสูง สำหรับพารามิเตอร์อำนาจจำแนก (a) และระดับปานกลาง สำหรับพารามิเตอร์การเดา (c) ซึ่งหมายถึง มีความผิดพลาดเชิงระบบ (Systematic bias) ในการทำนาย

2. ทุกวิธีทำนายพารามิเตอร์การเดาได้ไม่ดี

3. กรณีที่ข้อมูลที่ไม่มีอิทธิพลของเทสต์เลท วิธีเบส์ทั้ง 3 โมเดล มีการประมาณค่าได้เทียบเท่ากันในทุกวิธี (ยกเว้น ML)

4. อิทธิพลของเทสต์เลทมีผลกระทบต่อการประมาณค่าพารามิเตอร์ เมื่อใช้วิธีแมกซิมัมไลค์ลิฮู้ดและวิธีเบส์โดยไม่มีอิทธิพลเทสต์เลทในการประมาณค่า นั่นคือ มีความความสัมพันธ์ระหว่างค่าจริงและค่าที่คำนวณต่ำกว่า วิธีเบส์โดยกำหนดให้อิทธิพลเทสต์เลทคงที่และวิธีเบส์แบบมีอิทธิพลของเทสต์เลทต่างกันได้

5. เมื่อข้อมูลมีอิทธิพลเทสต์เลท (เท่ากันและไม่เท่ากัน) พบว่า วิธีเบส์โดยกำหนดให้อิทธิพลของเทสต์เลทคงที่ และวิธีเบส์แบบมีอิทธิพลเทสต์เลท มีค่าความสัมพันธ์ระหว่างค่าจริงและค่าที่คำนวณใกล้กันมากและเกือบเป็น 1

Jiao et al. (2005) พัฒนา Testlet Model โดยประยุกต์ใช้การวิเคราะห์ข้อสอบแบบพหุระดับด้วยโมเดลเชิงเส้นตรงระดับลดหลั่น 3 ระดับ (Hierarchical generalized linear model: HGLM - 3L) โดยใส่อิทธิพลของเทสต์เลท (Testlet effect) ในโมเดล HGLM - 3L และโมเดลเชิงเส้นตรงระดับลดหลั่น 2 ระดับ (HGLM - 2L) ซึ่งเป็นโมเดลที่เทียบเท่ากับโมเดล Rasch เพื่อตรวจสอบความไม่เป็นอิสระของข้อสอบ โดยกำหนดระดับความไม่เป็นอิสระของข้อสอบที่แตกต่างกันอย่างสุ่ม และวิเคราะห์จากข้อมูลจำลอง ผลการศึกษา พบว่า

1. โมเดลการวิเคราะห์แบบพหุระดับ มีความแปรปรวนที่แปรผันตามระดับความไม่เป็นอิสระของข้อสอบ เมื่อค่าความไม่เป็นอิสระของข้อสอบเพิ่มขึ้น การประมาณค่าความแปรปรวน

ของอิทธิพลกลุ่มจะเพิ่มขึ้นด้วยตามลำดับ ดังนั้นค่าที่ประมาณได้กับค่าที่แท้จริงจะมีความใกล้เคียงกันในบางค่า โดยเฉพาะความแปรปรวนในระดับ 2 จะใกล้เคียงกับความไม่เป็นอิสระ ของข้อสอบในข้อมูลจำลองในเงื่อนไข σ_u^2 = 0.0, 0.5, 1.0

2. การประมาณค่าพารามิเตอร์ความยาก จากเงื่อนไขจำลองในแต่ละเงื่อนไข เมื่อเปรียบเทียบระหว่างค่าความยากระหว่างโมเดล HGLM - 2L และ HGLM - 3L ในแต่ละข้อแล้วเมื่อความไม่เป็นอิสระของข้อสอบต่างกัน พบว่า การประมาณค่าพารามิเตอร์ความยากของข้อสอบระหว่างโมเดล HGLM - 2L และ HGLM - 3L ในข้อสอบทุกข้อเหมือนกันเมื่อ σ_u^2 = 0

3. ส่วนเงื่อนไขอื่น ๆ การประมาณค่าพารามิเตอร์ความยากในโมเดล HGLM - 2L ต่ำกว่าค่าพารามิเตอร์ความยากที่ประมาณได้จากโมเดล HGLM - 3L เมื่อค่าความไม่เป็นอิสระของข้อสอบ มีขนาดสูงขึ้น แสดงให้เห็นว่า เมื่อไม่ได้คำนึงถึงค่าอิทธิพลของความไม่เป็นอิสระของข้อสอบ จะทำให้การประมาณค่าความยากต่ำกว่าความเป็นจริง

Yue & Hong - Yun (2012) เปรียบเทียบการประมาณพารามิเตอร์โดยทำการวิเคราะห์ด้วยโมเดล Testlet Random-Effects และโมเดลแบบ Standard 2 - PL ด้วยวิธี Bayesian วัตถุประสงค์ของการศึกษา คือ การตรวจสอบผลกระทบของการเลือกรูปแบบเพื่อวิเคราะห์ข้อมูลในสถานการณ์ต่าง ๆ โดยมีเงื่อนไขการจำลองข้อมูล ได้แก่ ความแปรปรวนของเทสต์เลท (0, 0.5, 1, 2) ขนาดของเทสต์เลท (2, 5, 10) และความยาวของแบบสอบ (20, 40, 60) กำหนดให้มีผู้สอบทุกสถานการณ์เป็น 1,000 คน และมีข้อที่เป็นเทสต์เลท เป็นร้อยละ 50 ของความยาวแบบสอบ ทำการคำนวณซ้ำ 30 ครั้งในแต่ละเงื่อนไข ทำการวิเคราะห์ข้อมูลโดยใช้วิธี MCMC ด้วยโปรแกรม SCORIGHT ทั้ง 2 โมเดลที่ศึกษา การวิเคราะห์แบ่งเป็น 2 ตอน ได้แก่ ตอนที่ 1 เปรียบเทียบการประมาณค่าพารามิเตอร์ของ 2 โมเดล โดยใช้เกณฑ์ในการตรวจสอบ ได้แก่ ความลำเอียง (Bias) MAE (Mean absolute error) RMSE (Root mean square error) ความสัมพันธ์ระหว่างค่าที่ได้จากการคำนวณกับค่าจริง (Correlation between estimated and true values) ค่า 95% posterior interval width และค่า 95% coverage probability ตอนที่ 2 เปรียบเทียบการวิเคราะห์ของ 2 โมเดล ของ 2 ปัจจัย ได้แก่ สัดส่วนของข้อสอบที่เป็นอิสระ (1/ 3, 1/ 2, 2/ 3) และความยาวของข้อสอบ (20, 30, 40, 60) ผลการศึกษา พบว่า

1. ความถูกต้องของการประมาณค่าของพารามิเตอร์ทั้งหมดของ 2 - PL Bayesian testlet random - effect model มีความคงที่ แม้ความแปรปรวนของและขนาดของเทสต์เลทจะเปลี่ยนไป อย่างไรก็ตาม การประมาณค่าความคาดเคลื่อน (Error) ของทุกพารามิเตอร์ของ 2 - PL Bayesian model จะเพิ่มขึ้นอย่างรวดเร็ว เมื่อความแปรปรวนของและขนาดของเทสต์เลทมีค่ามากขึ้น นอกจากนี้ Bayesian testlet random - effect model จะมีความคาดเคลื่อนน้อยกว่า 2 - PL Bayesian

model ดังนั้น จึงควรเลือกใช้ 2 - PL Bayesian testlet random-effects model เมื่อความแปรปรวนของและขนาดของเทสต์เลทมีค่ามาก

2. แม้ว่าความถูกต้องของการประมาณค่าพารามิเตอร์ของข้อสอบและผู้สอบโดยใช้ Bayesian testlet random - effect model จะไม่ได้รับผลกระทบจากความยาวของแบบสอบ แต่หากแบบสอบมีจำนวนข้อน้อย จะทำให้ความคาดเคลื่อนในการประมาณค่าพารามิเตอร์ของ 2 - PL Bayesian model เพิ่มขึ้นอย่างรวดเร็ว และหากข้อสอบมีความเป็นอิสระต่อกันแล้ว การใช้ Bayesian testlet random-effect model จะทำให้ประมาณค่าความสามารถผิดพลาดมากกว่า 2 - PL Bayesian model

3. เมื่อสัดส่วนของจำนวนข้อที่มีความเป็นอิสระกันในแบบสอบมากขึ้น และความยาวของแบบสอบมากกว่า 20 ข้อ การประมาณค่าพารามิเตอร์ของทั้งสองโมเดลไม่ต่างกัน

Lai & Twu (2013) ศึกษาอิทธิพลของเทสต์เลทที่มีผลกับความถูกต้องในการประมาณค่าพารามิเตอร์ผู้สอบ และพารามิเตอร์ข้อสอบ โดยเปรียบเทียบระหว่างการประมาณค่าด้วยโมเดลตาม Item Response Theory (IRT) โมเดลตาม Testlet Response Theory (TRT) และ Bi - factor model เสนอการคำนวณ Q_3 เป็นดัชนีวัดความไม่เป็นอิสระของข้อสอบ ทำการวิเคราะห์โดยจำลองข้อมูลเป็น 6 เทสต์เลท แต่ละเทสต์เลทมีข้อสอบ 5 ข้อ มีเงื่อนไขการจำลองข้อมูล ได้แก่ Testlet Slope (0.0, 0.6, 1.2) ขนาดตัวอย่าง (500, 1000) ทำการคำนวณซ้ำ 100 ครั้งในแต่ละเงื่อนไข ผลการศึกษา พบว่า

1. TRT model สามารถประมาณค่าพารามิเตอร์ได้ถูกต้องมากกว่า IRT model และประสิทธิภาพของ TRT model ดีที่สุด รองลงมาเป็น Bi - factor model และ IRT model ตามลำดับ

2. กลุ่มตัวอย่างขนาดใหญ่จะทำให้การประมาณค่าพารามิเตอร์ถูกต้องกว่ากลุ่มตัวอย่างขนาดเล็ก และเมื่ออิทธิพลของ Testlet มาก ค่าความคาดเคลื่อน (Error) ก็จะมากด้วย

3. ค่าดัชนี Q_3 ที่อยู่ในเทสต์เลทเดียวกันมีค่ามากกว่าข้อที่อยู่ต่างเทสต์เลท แสดงว่าค่าดัชนี Q_3 มีความสามารถในการตรวจสอบความไม่เป็นอิสระของข้อสอบ

งานวิจัยที่เกี่ยวข้องกับการตรวจสอบ DIF

สุทธิพร ศุรธณี (2550) ได้ศึกษาความสามารถในการตรวจสอบการทำหน้าที่ต่างกันองข้อสอบตามตัวแบบเชิงเส้นวางนัยทั่วไประดับลดหลั่น (HGLM) ใช้ข้อมูลจำลองในการศึกษา กำหนดให้มีจำนวนข้อสอบ 20 ข้อ และมีจำนวนผู้สอบเท่ากัน 1,000 คนทุกเงื่อนไข (กำหนดกลุ่มเปรียบเทียบ 500 คนและกลุ่มอ้างอิง 500 คน) โดยเงื่อนไขการจำลองข้อมูล ได้แก่ ความแตกต่างระหว่างค่าเฉลี่ยความสามารถของผู้สอบทั้งสองกลุ่มการแจกแจงความสามารถ 2 เงื่อนไข คือ การแจกแจงแบบปกติ และการแจกแจงแบบฟิชเชอร์ทิปเปทท์ ความแตกต่างระหว่าง

ค่าเฉลี่ยความสามารถของผู้สอบทั้งสองกลุ่ม สัดส่วนของข้อสอบที่มี DIF กำหนดให้มี 3 ระดับคือ 5% 15% และ 30% ขนาด DIF มี 3 ขนาด คือ 0.3 (ขนาดเล็ก) 0.5 (ขนาดปานกลาง) 0.7 (ขนาดใหญ่) และวิธีการตรวจสอบ DIF 2 วิธี คือ วิธีการตรวจสอบทุกข้อพร้อมกัน และวิธีการตรวจสอบทีละข้อ การศึกษาจะใช้การจำลองข้อมูลด้วยเทคนิคมอนติคาล์โล โดยใช้โปรแกรม R เวอร์ชัน 2.01 การวิเคราะห์จะใช้ขั้นตอนวิธี HLM2 โดยใช้โปรแกรมสำเร็จรูป HLM เวอร์ชัน 6.0 โดยแต่ละกรณีทำซ้ำจำนวน 100 ครั้ง และแต่ละครั้งที่ดำเนินการจะคำนวณหาค่าประมาณของ DIF ค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE) กำลังการทดสอบและอัตราความผิดพลาดประเภท ที่ 1 ผลการวิจัยสรุปได้ดังนี้

1. กรณีที่ความสามารถของผู้สอบทั้งสองกลุ่มมีการแจกแจงแบบปกติ พบว่า ความแตกต่างระหว่างค่าเฉลี่ยของความสามารถของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบไม่มีผลต่อค่า RMSE ค่าเฉลี่ยประมาณของ DIF กำลังการทดสอบ และอัตราความผิดพลาดประเภทที่ 1 วิธีการตรวจสอบทีละข้อจะขึ้นอยู่กับสัดส่วนของข้อสอบที่มี DIF โดยที่ค่า RMSE จะมีค่ามากขึ้นเมื่อสัดส่วนของข้อสอบที่มี DIF สูงขึ้น แต่ในทางกลับกันสัดส่วนของข้อสอบที่มี DIF ไม่มีผลต่อวิธีการตรวจสอบทุกข้อพร้อมกัน และวิธีการตรวจสอบทีละข้อยังมีค่า RMSE น้อยกว่าวิธีการตรวจสอบทุกข้อพร้อมกัน กรณีที่มีข้อสอบที่มี DIF จำนวน 1 ข้อ และ 3 ข้อ แต่กรณีที่มีข้อสอบมี DIF จำนวน 6 ข้อ วิธีการตรวจสอบทีละข้อยังมีค่า RMSE มากกว่าวิธีการตรวจสอบทุกข้อพร้อมกัน และค่าเฉลี่ยประมาณของมี DIF จะขึ้นอยู่กับวิธีการที่ใช้ในการตรวจสอบมี DIF โดยที่วิธีการตรวจสอบทีละข้อให้ค่าเฉลี่ยต่ำกว่าค่าจริง ขณะที่วิธีการตรวจสอบทุกข้อพร้อมกันให้ค่าเฉลี่ยทั้งต่ำกว่าและสูงกว่าค่าจริง ส่วนกำลังการทดสอบและอัตราความผิดพลาดประเภทที่ 1 ขึ้นอยู่กับที่อยู่กับวิธีที่ใช้ในการตรวจสอบและสัดส่วนของข้อสอบที่มี DIF ซึ่งวิธีการตรวจสอบทีละข้อมีกำลังการทดสอบสูงกว่าวิธีการตรวจสอบทุกข้อพร้อมกัน กรณีที่มีข้อสอบมี DIF จำนวน 1 ข้อ และ 3 ข้อ แต่กรณีที่มีข้อสอบมี DIF จำนวน 6 ข้อ วิธีการตรวจสอบทีละข้อยังมีกำลังการทดสอบน้อยกว่าวิธีการตรวจสอบทุกข้อพร้อมกัน กำลังการทดสอบของทั้งสองวิธีไม่ขึ้นอยู่กับสัดส่วนของข้อสอบที่มี DIF, อัตราความผิดพลาดประเภทที่ 1 ของวิธีการตรวจสอบทีละข้อส่วนใหญ่มีค่าอยู่นอกช่วงที่กำหนดและยังมีค่ามากกว่าวิธีการตรวจสอบทุกข้อพร้อมกัน

2. กรณีที่ความสามารถของผู้สอบทั้งสองกลุ่มมีการแจกแจงแบบฟิชเชอร์ทิปเปทท์ พบว่า ความแตกต่างระหว่างค่าเฉลี่ยของความสามารถของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบไม่มีผลต่อค่าเฉลี่ยของค่าประมาณของ DIF และกำลังการทดสอบ แต่มีผลต่อค่า RMSE และอัตราความผิดพลาดประเภทที่ 1 โดยที่วิธีการตรวจสอบทีละข้อมีค่าเฉลี่ยของค่าประมาณ DIF ต่ำกว่าค่าจริง และมีค่ากำลังการทดสอบสูงกว่าวิธีการตรวจสอบทุกข้อพร้อมกัน

3. เมื่อเปรียบเทียบทั้งกรณีที่ความสามารถของผู้สอบมีการแจกแจงแบบปกติและการแจกแจงแบบฟิชเชอร์ทิปเปทท์ พบว่า กำลังการทดสอบกรณีที่ความสามารถของผู้สอบมีการแจกแจงแบบปกติมีค่ามากกว่ากรณีที่ความสามารถของผู้สอบมีการแจกแจงแบบฟิชเชอร์ทิปเปทท์ ในกรณีที่มีข้อสอบมี DIF จำนวน 6 ข้อ สำหรับวิธีการตรวจสอบทีละข้อมีค่า RMSE มากกว่าวิธีการตรวจสอบทุกข้อพร้อมกันในกรณีที่มีความแตกต่างระหว่างค่าเฉลี่ยของความสามารถของผู้สอบทั้งสองกลุ่ม และกรณีที่มีข้อสอบมี DIF จำนวน 6 ข้อ อย่างไรก็ตาม วิธีการตรวจสอบทุกข้อพร้อมกัน กรณีที่ความสามารถของผู้สอบมีการแจกแจงแบบฟิชเชอร์ทิปเปทท์ มีค่า RMSE มากกว่ากรณีที่ความสามารถของผู้สอบมีการแจกแจงแบบปกติ สำหรับวิธีการตรวจสอบทุกข้อพร้อมกันในกรณีที่ไม่มีความแตกต่างระหว่างค่าเฉลี่ยของความสามารถของผู้สอบทั้งสองกลุ่ม มีค่าอัตราความผิดพลาดประเภทที่ 1 น้อยกว่ากรณีที่มีความแตกต่างระหว่างค่าเฉลี่ยของความสามารถของผู้สอบทั้งสองกลุ่ม ในขณะที่วิธีการตรวจสอบทีละข้อ มีค่าอัตราความผิดพลาดประเภทที่ 1 อยู่นอกช่วงที่กำหนดทั้งในกรณีที่มีความแตกต่างและไม่มีความแตกต่างระหว่างค่าเฉลี่ยของความสามารถของผู้สอบทั้งสองกลุ่ม

Shin & Wall (2006) เปรียบเทียบประสิทธิภาพของวิธีที่ใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ 3 วิธี คือ วิธีแมนเทล-แฮนส์เซลไคสแควร์ (Mantel - Haenszel chi-square test: MH-χ2) วิธี ETS'Mantel - Haenszel Data (MH-D) และวิธี IRT (IRT - DIF) ที่โมเดลแบบ 1 พารามิเตอร์ และ 3 พารามิเตอร์ เมื่อความสามารถของผู้สอบมีการแจกแจงแบบปกติ การแจกแจงนอนพาราเมทริกซ์ที่เบ้ซ้าย และการแจกแจงไคร์สแควร์เบ้ขวาที่มี Degree of freedom เท่ากับ 10 โดยศึกษาเงื่อนไข ดังนี้ จำนวนผู้สอบที่ต่างกัน 3 ระดับ และขนาด DIF ต่างกัน 3 ระดับ และทำการทดลองซ้ำ 100 ครั้ง ได้ผลสรุปว่า การแจกแจงความสามารถของผู้สอบที่ไม่เป็นไปตามสมมติฐานนั้น มีผลต่อการประมาณค่า DIF และการตรวจสอบค่า DIF โดยจะทำให้อัตราความผิดพลาดประเภทที่ 1 สูงขึ้น

Lee et al. (2009) ได้ศึกษาอัตราความคาดเคลื่อนประเภทที่ 1 (Type I error) และอำนาจการทดสอบ (Power) ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบและเทสต์เลท โดยใช้วิธี SIBTEST และ Poly-SIBTEST (Testlet DIF) โดยรวมคะแนนแต่ละเทสต์เลทให้เป็นข้อสอบแบบ Polutomous 1 ข้อ ข้อมูลที่นำมาศึกษาใช้ทั้งข้อมูลจริงและข้อมูลจำลอง โดยมีเงื่อนไขการจำลองข้อมูล ได้แก่ การแจกแจงความสามารถของกลุ่มเปรียบเทียบและกลุ่มอ้างอิง (เท่ากันทั้ง 2 กลุ่ม คือ N(0,1) และต่างกันโดยกลุ่ม R เป็น N(0,1) กลุ่ม F เป็น N(-1, 1)) ขนาดตัวอย่าง (250, 500, 1000) ความยาวแบบสอบ (30, 60) จำนวนข้อที่ DIF แบบ Uniform

ในแบบสอบ (0%, 10%, 20% ของความยาวแบบสอบ) ขนาดของ DIF (0.2, 0.5, 1.0) รวม 84 เงื่อนไข ผลการศึกษา พบว่า

1. อัตราความคาดเคลื่อนประเภทที่ 1 (Type I error) พบว่า ขนาดตัวอย่างมากจะมีอัตราความคาดเคลื่อนประเภทที่ 1 มากกว่าขนาดตัวอย่างน้อย แต่ถ้าความยาวของแบบสอบน้อย จะมีอัตราความคาดเคลื่อนประเภทที่ 1 มากกว่าความยาวของแบบสอบมาก ซึ่งมีลักษณะเดียวกันทั้งวิธี SIBTEST และ Poly - SIBTEST

2. อำนาจการตรวจสอบ (Power) พบว่า ในกรณีวิธี SIBTEST จะมี Power มากขึ้นเมื่อขนาดตัวอย่างเพิ่มขึ้นและจำนวนข้อที่ DIF ในแบบสอบมากขึ้น ในขณะที่ Power ลดลงเมื่อความยาวของแบบสอบเพิ่มขึ้น ต่างจากวิธี Poly - SIBTEST ที่ Power ลดลง เมื่อขนาดตัวอย่างและยาวของแบบสอบลดลง ส่วนจำนวนข้อที่ DIF ในแบบสอบแสดงผลกับ Power ไม่ชัดเจน

Sedivy (2009) เปรียบเทียบวิธีการตรวจสอบ DIF ของข้อมูลที่มีลักษณะเป็นเทสต์เลท 3 วิธี ได้แก่ วิธี Mantel - Haenszel วิธี Logistic Regression และวิธี SIBTEST โดยศึกษาจากความแกร่ง (Robust) ของการตรวจสอบ DIF จากอัตราความคาดเคลื่อนประเภทที่ 1 (Type I error) และอำนาจในตรวจสอบ DIF (Power) ข้อมูลที่นำมาศึกษาใช้ทั้งข้อมูลจริงและข้อมูลจำลอง โดยมีเงื่อนไขการจำลองข้อมูล ได้แก่ ขนาดตัวอย่าง (500, 1000, 4000) ขนาด DIF (no DIF, 0.2, 0.4, 0.6) จำนวนข้อในเทสต์เลท (5, 10, 15) อิทธิพลของเทสต์เลท (0.5, 1, 2) ระดับความยากของข้อสอบ (ต่ำ, กลาง, สูง) กำหนดให้มีความยาวแบบสอบในทุกสถานการณ์ 30 ข้อเท่ากัน ทำการคำนวณซ้ำ 100 ครั้ง ในแต่ละเงื่อนไข ผลการศึกษา พบว่า

1. อัตราความผิดพลาดประเภทที่ 1 ทั้ง 3 วิธี ได้รับผลกระทบจากข้อสอบแบบเทสต์เลท

2. ในส่วนของประสิทธิภาพหรืออำนาจในตรวจสอบ (Power) พบว่า อัตราการตรวจสอบ DIF สูงสุด สำหรับข้อสอบที่ง่าย เมื่อเทียบกับข้อสอบที่มีความยากปานกลางและยากมาก แสดงว่าวิเคราะห์โดยการรวมอิทธิพลของเทสต์เลทจะทำให้สูญเสียอำนาจในการทดสอบ

3. ทั้ง 3 วิธีมีประสิทธิภาพ นั่นคือมีอำนาจในการตรวจสอบ DIF เกือบร้อยละ 100 เมื่อกำหนดเงื่อนไข ตัวอย่างมีขนาดใหญ่ (n = 4000) และ ขนาดของ DIF อยู่ในระดับปานกลางและใหญ่ (0.4, 0.6) อย่างไรก็ตาม หากความยากเพิ่มขึ้นจะทำให้ Power ลดลง (ภายใต้เงื่อนไขอื่นที่เหมือนกัน)

Fukuhara & 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

อิทธิพลของเทสต์เลท (0.5, 1, 2) ขนาด DIF (0.5, 0.7) ระดับค่าอำนาจจำแนกของข้อสอบ (0.8, 2.0) และจำนวนผู้สอบของ Focal group (500, 250) กำหนดให้มีความยาวแบบสอบในทุกสถานการณ์ 42 ข้อ (7 ข้อต่อ 1 Testlet แบบสอบมี 6 Testlets) และจำนวนผู้สอบ 1,000 คนเท่ากันทุกเงื่อนไข ทำการคำนวณซ้ำ 100 ครั้งในแต่ละเงื่อนไข ผลการศึกษา พบว่า

1. ทั้ง 2 โมเดล ประมาณค่าขนาดของ DIF ในข้อที่ DIF ต่ำกว่าจริง (Underestimated) สำหรับเงื่อนไขที่มีค่าอำนาจจำแนกของข้อสอบ สูง (a = 2.0) และการประมาณค่าความยาก ของทั้ง 2 โมเดลมีค่าเข้าใกล้ 0 เหมือนกัน

2. Bifactor MIRT DIF model มีอัตราการตรวจสอบ DIF สูงกว่าและประมาณค่าขนาด DIF ได้ถูกต้องกว่า IRT DIF model นอกจากนี้ Bifactor MIRT DIF model ยังมีอัตราความผิดพลาดในการตรวจ DIF ที่ต่ำมาก และประมาณค่าพารามิเตอร์อื่นได้ดีกว่า IRT DIF model

3. Bi - factor MIRT DIF model มีอคติ (Bias) ในการประมาณค่าขนาด DIF น้อยกว่า 2PL IRT DIF model

จากผลการศึกษางานวิจัยที่เกี่ยวข้องทั้งในประเทศและต่างประเทศ สรุปได้ว่า มีการศึกษาการประมาณค่าพารามิเตอร์และการทำหน้าที่ต่างกันของข้อสอบที่มีลักษณะของเทสต์เลท โดยใช้ทั้งข้อมูลจำลองและข้อมูลเชิงประจักษ์ งานวิจัยส่วนใหญ่จะเน้นไปที่การใช้ข้อมูลจำลอง เนื่องจากเป็นศึกษาภายใต้เงื่อนไขต่าง ๆ ที่ข้อมูลจริงอาจจะไม่มีลักษณะเป็นตามเงื่อนไขที่ต้องการศึกษา โดยเงื่อนไขที่ศึกษาจะแตกต่างกันไป เช่น ความยาวแบบสอบ ขนาดกลุ่มตัวอย่าง สัดส่วนของขนาดกลุ่มเปรียบเทียบและกลุ่มอ้างอิง ผลการศึกษาที่ผ่านมาส่วนมากจะศึกษาการทำหน้าที่ต่างกันที่ระดับข้อสอบ ส่วนข้อสอบลักษณะของเทสต์เลทยังมีไม่มาก

บทที่ 3
วิธีดำเนินการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์ 2 ประการ คือ ประการแรกเพื่อศึกษาประสิทธิผลการประมาณค่าพารามิเตอร์ข้อสอบ (อำนาจจำแนกและความยาก) กับพารามิเตอร์ความสามารถของผู้สอบ (θ) ด้วยวิธีแมกซิมัมไลค์ลิฮู้ด (ML) วิธีของเบส์ (Bayes) และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) และประการที่สองเพื่อศึกษาผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ระหว่างวิธีแมกซิมัมไลค์ลิฮู้ด (ML) วิธีของเบส์ (Bayes) และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) โดยข้อมูลที่ใช้ในการศึกษาเป็นข้อมูลที่จำลองจากโปรแกรม R โดยประมาณค่าพารามิเตอร์และการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบของวิธีแมกซิมัมไลค์ลิฮู้ดใช้ Package TAM ของโปรแกรม R ส่วนการประมาณค่าพารามิเตอร์ของวิธีแบบ Bayes และ Bayesγ จะทำการเขียนคำสั่งการประมวลผลด้วยโปรแกรม WinBUGS และกลับมาแสดงผลในโปรแกรม R ด้วย Package R2WinBUGS มีรายละเอียดในการดำเนินการวิจัยดังนี้

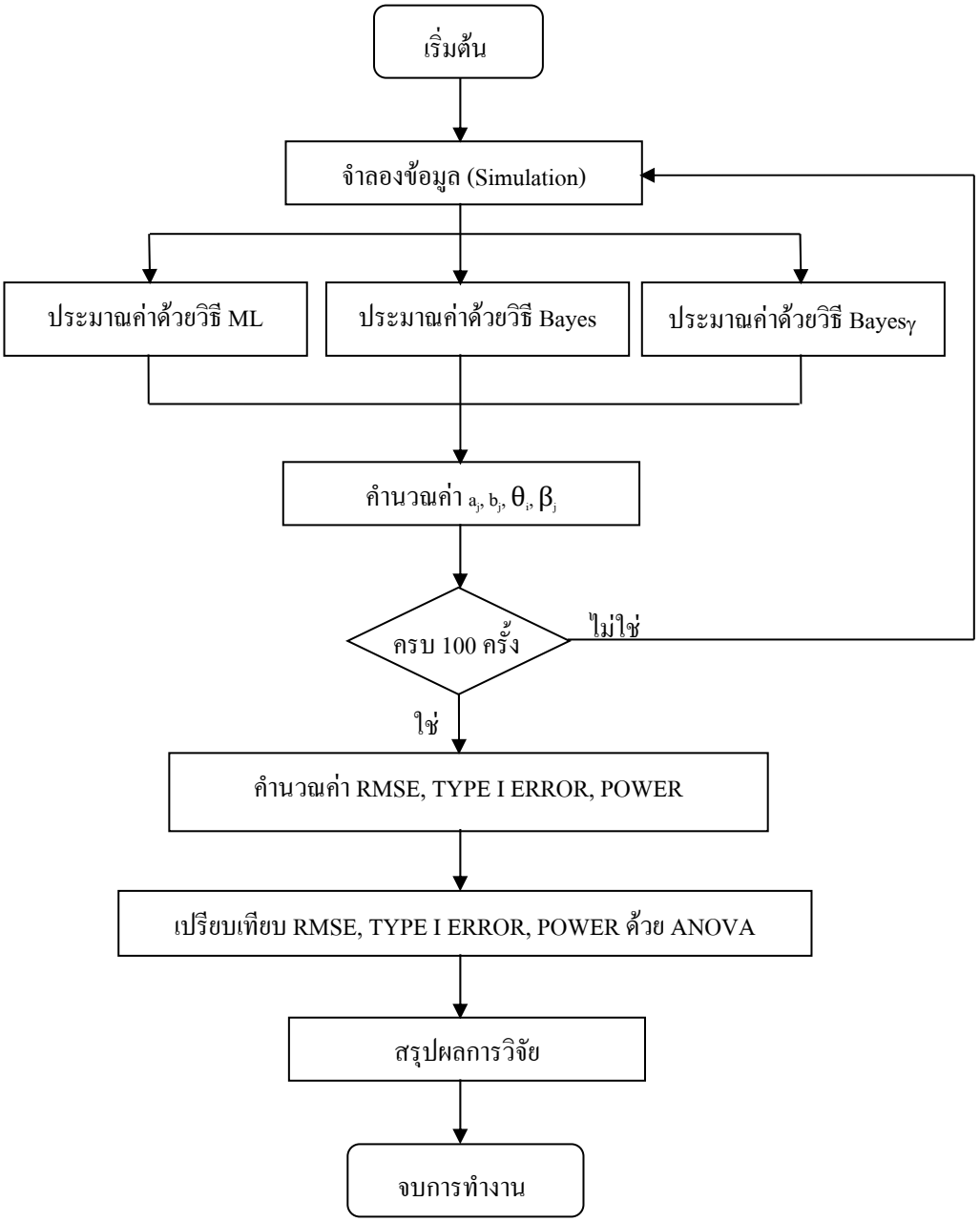
ขั้นตอนการวิจัย

การศึกษาในครั้งนี้มีขั้นตอนการวิจัย โดยแบ่งเป็นขั้นตอนใหญ่ ๆ 3 ขั้นตอน ดังนี้

1. จำลองข้อมูลด้วยโปรแกรม R
2. ประมาณค่าพารามิเตอร์และตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ 3 วิธี
3. การวัดประสิทธิภาพของการประมาณค่า โดยแบ่งเป็น

3.1 วัดประสิทธิภาพของการประมาณค่าพารามิเตอร์ แล้วทดสอบความแตกต่างของ 4 เงื่อนไข และวิธีการประมาณค่าพารามิเตอร์ ด้วยวิธี 5 - Way ANOVA

3.2 วัดประสิทธิภาพของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ แล้วทดสอบความแตกต่างของ 4 เงื่อนไข และวิธีการประมาณค่าพารามิเตอร์ ด้วยวิธี 5 - Way ANOVA โดยแต่ละขั้นตอนมีรายละเอียด ดังนี้



ภาพที่ 3 - 1 ขั้นตอนการดำเนินงานวิจัย

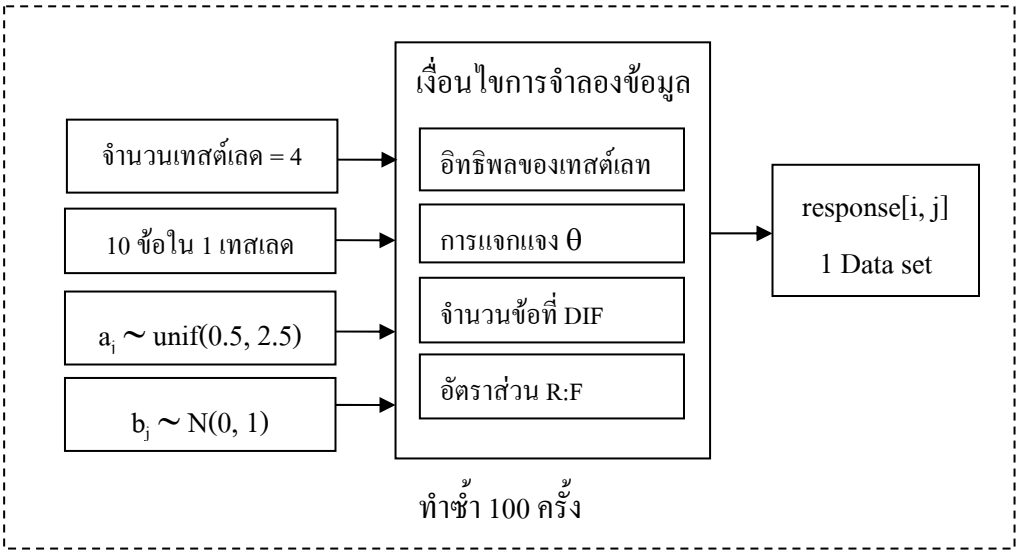
ขั้นตอนที่ 1 จำลองข้อมูลด้วยโปรแกรม R โดยมีรายละเอียดของการจำลองข้อมูล ดังนี้

1. กำหนดให้แบบสอบมีความยาว 40 ข้อ ประกอบด้วย 4 เทสต์เลท แต่ละเทสต์เลทมีขนาดเท่ากัน คือ 10 ข้อทุกเงื่อนไข

2. กำหนดรูปแบบและการแจกแจงพารามิเตอร์ของข้อสอบที่ใช้จำลองข้อมูล ดังนี้

2.1 พารามิเตอร์ความยาก (b_j) มีการแจกแจงแบบปกติที่มีค่าเฉลี่ยเป็น 0 และส่วนเบี่ยงเบนมาตรฐานเป็น 1 หรือเขียนแทนด้วย $b_j \sim N(0, 1)$

2.2 พารามิเตอร์อำนาจจำแนก (a_j) มีการแจกแจงแบบ Uniform มีค่าต่ำสุด = 0.5 และค่าสูงสุด = 2.5 หรือเขียนแทนด้วย $a_j \sim unif(0.5, 2.5)$



ภาพที่ 3 - 2 ขั้นตอนการจำลองข้อมูลด้วยโปรแกรม R

3. กำหนดเงื่อนไขของการจำลองข้อมูลของ 4 ปัจจัย ดังนี้

3.1 อิทธิพลของเทสต์เลท (Testlet Effect: $\gamma_{id(j)}$) 3 เงื่อนไข ได้แก่

3.1.1 แบบสอบที่มีค่าอิทธิพลของเทสต์เลทเท่ากันทุกเทสต์เลท (Equal effect) นั่นคือ กำหนดให้อิทธิพลของเทสต์เลทมีค่าเท่ากับ 0.8

3.1.2 แบบสอบที่แต่ละเทสต์เลทมีค่าอิทธิพลของเทสต์เลทไม่เท่ากัน (Unequal effect) โดยกำหนดให้อิทธิพลของเทสต์เลทมีค่าเป็น 0.25, 0.5, 1 และ 2 ตามลำดับ

3.1.3 แบบสอบประกอบด้วยข้อสอบที่เป็นอิสระและเทสต์เลท (Unequal effect แบบ Independent + Testlet) โดยกำหนดให้อิทธิพลของเทสต์เลทมีค่าเป็น 0, 0.25, 0.56 และ 1 ตามลำดับ ซึ่งอิทธิพลของเทสต์เลทมีค่าเท่ากับศูนย์ หมายถึง ข้อสอบที่เป็นอิสระ

3.2 การแจกแจงของความสามารถ (การแจกแจง θ_i) 3 เงื่อนไข ได้แก่

3.2.1 การแจกแจงแบบปกติ

3.2.2 การแจกแจงแบบเบ้ซ้าย

3.2.3 การแจกแจงแบบเบ้ขวา

3.3 จำนวนข้อสอบที่มีการทำหน้าที่ต่างกัน (ร้อยละข้อที่ DIF) 3 เงื่อนไข ได้แก่

3.3.1 ร้อยละ 0 หรือไม่มีข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบ

3.3.2 ร้อยละ 12.5 หรือมีข้อสอบที่ทำหน้าที่ต่างกันจำนวน 5 ข้อ ในแบบสอบ

3.3.3 ร้อยละ 20 หรือมีข้อสอบที่ทำหน้าที่ต่างกันจำนวน 8 ข้อ ในแบบสอบ

3.4 อัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ (อัตราส่วน R: F) 2 เงื่อนไข ได้แก่

3.4.1 อัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ เป็น 1: 1 หรือ มีกลุ่มอ้างอิง จำนวน 1,000 คน และกลุ่มเปรียบเทียบ จำนวน 1,000 คน

3.4.2 อัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ เป็น 1: 0.1 หรือ มีกลุ่มอ้างอิง จำนวน 1,000 คน และกลุ่มเปรียบเทียบ จำนวน 100 คน

4. การตรวจสอบความถูกต้องของการจำลองข้อมูล

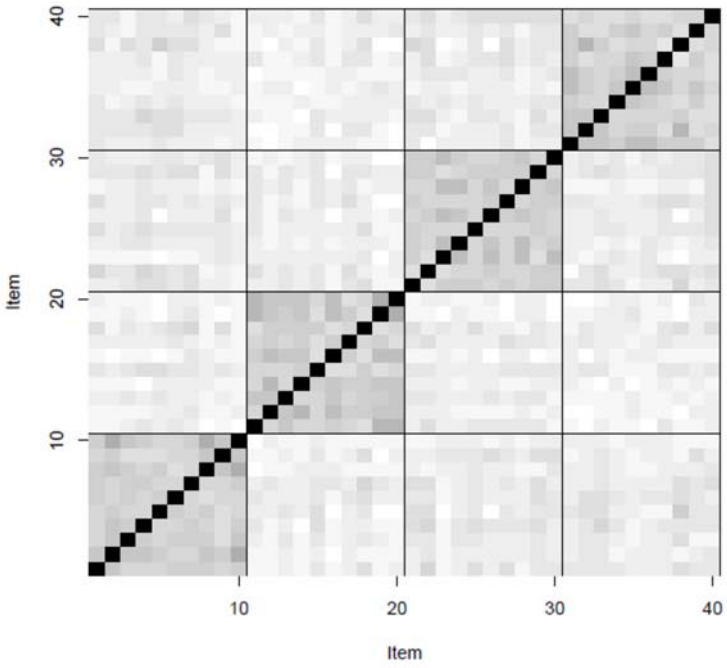
เพื่อให้แน่ใจว่าข้อมูลที่จำลองมามีความถูกต้องตามที่ต้องการ ผู้วิจัยจึงทำการตรวจสอบความถูกต้องของการจำลองข้อมูล โดยแบ่งเป็น

4.1 การตรวจสอบความไม่เป็นอิสระของข้อสอบ (Local item dependence: LID) เพื่อตัดสินว่าข้อสอบที่อยู่ในเทสต์เลทเดียวกันเป็นอิสระต่อกันหรือไม่ จากการใช้สถิติ Q3 (Yen, 1984) ซึ่งเป็นดัชนีสำหรับตรวจสอบความไม่เป็นอิสระกันของข้อสอบ โดยการหาความสัมพันธ์ของส่วนที่เหลือ (Residuals) ของคู่ของข้อสอบหลังจากตัด (Partialling) ส่วนที่เป็นการประมาณค่าคุณลักษณะ (Trait estimate) ออกไป การตัดสินความไม่เป็นอิสระกันของข้อสอบพิจารณาได้จากเกณฑ์ดังนี้ (อนุสรณ์ เกิดศรี, 2557)

1. ถ้าค่าสถิติ Q3 มีค่าเข้าใกล้ 0 แสดงว่าข้อสอบคู่นั้นมีหลักฐานพอเชื่อได้ว่าเป็นอิสระจากกัน

2. ถ้าค่าสถิติ Q3 มีค่าต่างจาก 0 เช่น มีค่าเป็นลบหรือเป็นบวกมาก ๆ แสดงว่าข้อสอบคู่นั้นมีหลักฐานพอเชื่อได้ว่าไม่เป็นอิสระจากกัน

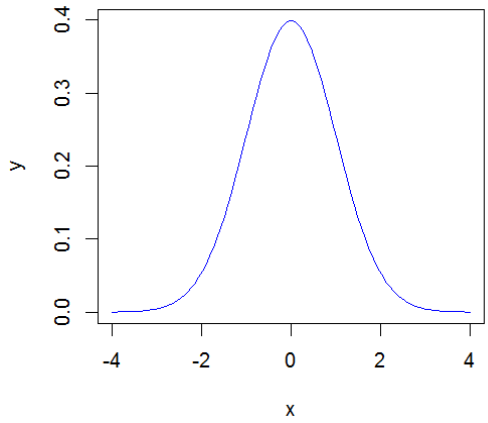
แต่เนื่องจากการศึกษาครั้งนี้เป็นการใช้แบบสอบจำนวน 40 ข้อ ในแต่ละเงื่อนไข ดังนั้นจะมีความสัมพันธ์ของส่วนที่เหลือของคู่ของข้อสอบ จำนวน 780 คู่ต่อการจำลองข้อมูลแต่ละครั้ง ทำให้การพิจาณารายคู่ค่อนข้างยุ่งยาก ดังนั้น ผู้วิจัยจึงเทียบระดับความสัมพันธ์กับระดับสีเทา โดยที่สีเทาเข้มจะมีความสัมพันธ์กันระหว่างคู่มากกว่าสีเทาอ่อน แล้วสร้างกราฟเพื่อตรวจสอบเงื่อนไขความเป็นเทสต์เลทแทน ดังภาพที่ 3 -3



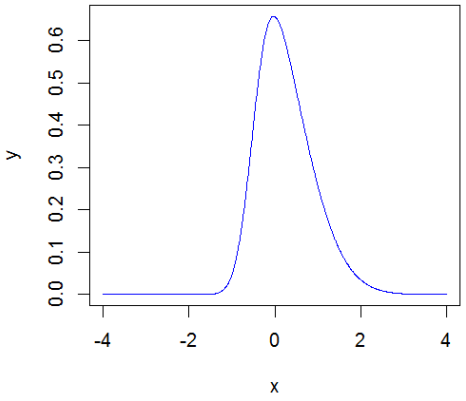
ภาพที่ 3 - 3 กราฟ Q3 statistics จากการจำลองข้อมูล

จากภาพที่ 3 - 3 จะเห็นว่าเส้นทะแยงมุมเป็นสีดำ เนื่องจากเป็นความสัมพันธ์ของข้อเดียวกันซึ่งมีค่าเป็น 1 และในช่วงข้อ 1 - 10 ข้อ 11 - 20 ข้อ 21 - 30 และข้อ 31 - 40 จะเป็นสีเทาเข้มกว่าบริเวณอื่น แสดงว่าในช่วงของข้อสอบที่กำหนดให้มีลักษณะเป็นเทสต์เลทนั้นมีความไม่อิสระต่อกัน

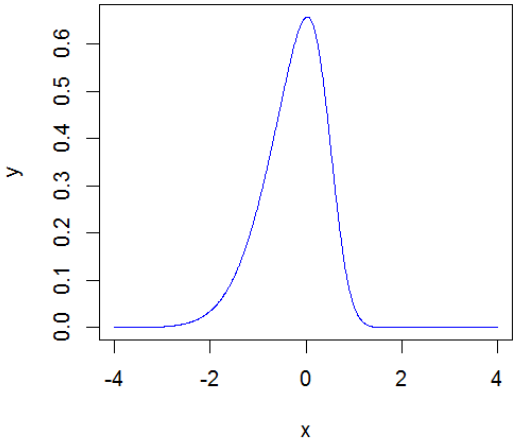
4.2 การตรวจสอบการแจกแจงความสามารถของผู้สอบ มี 3 เงื่อนไข ได้แก่ การแจกแจงแบบปกติ การแจกแจงแบบเบ้ซ้ายและการแจกแจงแบบเบ้ขวา ซึ่งจากภาพที่ 3 - 4 แสดงให้เห็นว่าข้อมูลที่จำลอง มีเงื่อนไขเป็นไปตามที่กำหนด



(a) การแจกแจงแบบปกติ



(b) การแจกแจงแบบเบ้ขวา



(c) การแจกแจงแบบเบ้ซ้าย

ภาพที่ 3 - 4 การแจกแจงความสามารถจากการจำลองข้อมูล

ขั้นตอนที่ 2 ประมาณค่าพารามิเตอร์และตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ประกอบด้วย 3 วิธี ได้แก่ วิธีแมกซิมัมไลค์ลิฮู้ด (Maximum likelihood: ML) วิธีของเบส์ (Bayes) และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) โดยมีรายละเอียด ดังนี้

เนื่องจากในการศึกษาครั้งนี้ มีเงื่อนไขในการวิเคราะห์ข้อมูลซึ่งแบ่งเป็น 2 ส่วน ได้แก่ การประมาณค่าพารามิเตอร์และการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ซึ่งทั้ง 2 ส่วน สามารถประมาณค่าพารามิเตอร์ได้พร้อมกัน ดังนั้นผู้วิจัยจึงอธิบายรายละเอียดของการประมาณค่าพารามิเตอร์ทั้ง 3 วิธี โดยอธิบายทั้งการประมาณค่าพารามิเตอร์ข้อสอบ ผู้สอบ และการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ซึ่งมีรายละเอียด ดังนี้

1. วิธีแมกซิมัมไลค์ลิฮู้ด (Maximum likelihood: ML) มีขั้นตอนการประมาณค่าพารามิเตอร์โดยสรุปดังภาพที่ 3 - 5 ซึ่งมีรายละเอียดดังนี้

1.1 นำข้อมูลที่ทำการจำลองใน**ขั้นตอนที่ 1** ประมาณค่าพารามิเตอร์ด้วยวิธีแมกซิมัมไลค์ลิฮู้ด โดยใช้ฟังก์ชัน tam.mml.2pl สำหรับประมาณค่าพารามิเตอร์ ฟังก์ชัน tam.wle สำหรับประมาณค่าความสามารถและฟังก์ชัน tam.mml.mfr สำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจากการเรียกใช้ Package TAM

ภาพที่ 3 - 5 ขั้นตอนการประมาณค่าพารามิเตอร์ด้วยวิธีแมกซิมัมไลค์ลิฮู้ด

1.2 การพิจารณาค่าพารามิเตอร์ความยากและความสามารถที่ประมาณได้นั้น จะพิจารณาจากค่า modMML$B (พารามิเตอร์อำนาจจำแนก) modMML$xsi (พารามิเตอร์ความยาก) และ Abil$theta (พารามิเตอร์ความสามารถ)

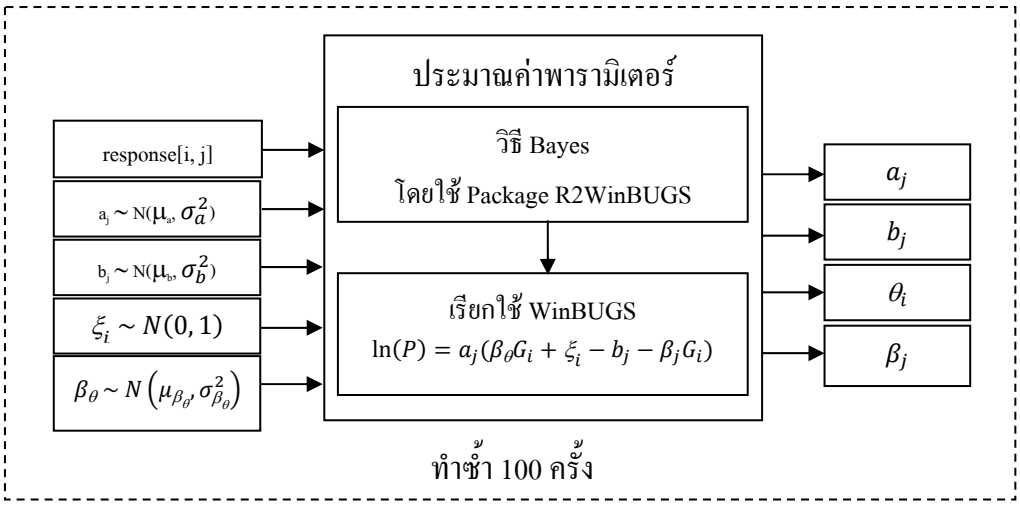
1.3 การตัดสินการทำหน้าที่ต่างกันของข้อสอบ ศึกษาจากค่าปฏิสัมพันธ์ (Interaction term) ของ item:groups จากสมการ formulaA แล้วทำการทดสอบนัยสำคัญด้วย z-statistic (ผลการประมาณค่าปฏิสัมพันธ์หารด้วยความคาดเคลื่อนมาตรฐาน) ที่ระดับนัยสำคัญ .05 หากผลการทดสอบพบว่ามีนัยสำคัญ แสดงว่า ข้อนั้นมีการทำหน้าที่ต่างกันของข้อสอบ เช่น

```
48  I06:Groups0 item:Groups -0.324  0.117
49  I07:Groups0 item:Groups -0.006  0.113
50  I08:Groups0 item:Groups -0.274  0.122
51  I09:Groups0 item:Groups -0.122  0.112
52  I10:Groups0 item:Groups  0.157  0.112
```

แผนภาพที่ 3 - 6 ตัวอย่างผลลัพธ์ของค่าปฏิสัมพันธ์ item:groups

จากแผนภาพที่ 3 - 6 เมื่อพิจารณาค่าปฏิสัมพันธ์ item:groups ของข้อ 8 จะได้ค่า Z score = -0.274/0.122 = -2.25 ซึ่งมีนัยสำคัญทางสถิติ แสดงว่าข้อ 8 มีการทำหน้าที่ต่างกันของข้อสอบ

2. วิธีของเบส์ (Bayes) ประมาณค่าพารามิเตอร์โดยใช้โมเดลการตอบสนองข้อสอบแบบ 2 พารามิเตอร์ กระทำโดยใช้โปรแกรม R ในการจำลองข้อมูลแล้วส่งไปประมวลผลด้วยโปรแกรม WinBUGS และกลับมาแสดงผลลัพธ์ที่โปรแกรม R โดยการเรียกใช้ Package R2WinBUGS สรุปขั้นตอนการประมาณค่าพารามิเตอร์ได้ดังแผนภาพที่ 3 - 7



แผนภาพที่ 3 - 7 ขั้นตอนการประมาณค่าพารามิเตอร์ด้วยวิธีของเบส์ (Bayes)

ขั้นตอนการประมาณค่าพารามิเตอร์ด้วยวิธีของเบส์ (Bayes) มีรายละเอียดดังนี้

2.1 นำข้อมูลที่ทำการจำลองในขั้นตอนที่ 1 ประมาณค่าพารามิเตอร์ด้วยวิธีของเบส์ โดยการเรียกใช้ Package R2WinBUGS ในโปรแกรม R ตามขั้นตอน ดังนี้

2.1.1 สร้างฟังก์ชันสำหรับสร้างตัวแบบเพื่อวิเคราะห์ตัวแบบเบส์ ซึ่งสามารถนำคำสั่งมาจากโปรแกรม WinBUGS แล้วตั้งชื่อเป็น “IRTmodel”

2.1.2 บันทึกไฟล์ในข้อ 2.1.1 ชื่อ “IRT.bug”

2.1.3 กำหนดตัวแปรที่จะนำเข้าไปประมวลผลในโปรแกรม WinBUGS ได้แก่ จำนวนผู้สอบ จำนวนข้อสอบ กลุ่ม และแบบแผนการตอบจากการจำลองข้อมูลตามเงื่อนไขต่าง ๆ

2.1.4 สร้างค่าเริ่มต้นของพารามิเตอร์ในโมเดลในรูปของฟังก์ชัน (Initial value) โดยให้อิทธิพลของความสามารถรวมเป็นศูนย์ ($\beta_\theta = 0$) ส่วนค่าอิทธิพลของกลุ่มที่ทำหน้าที่ต่างกันของข้อสอบ ค่าอำนาจจำแนกและค่าความยากเป็นค่าสุ่มจากการแจกแจงที่กำหนดในการจำลองข้อมูล

2.1.5 กำหนดพารามิเตอร์ที่ต้องการให้ตรวจสอบผล ซึ่งผู้วิจัยพิจารณาเฉพาะค่า b[] (พารามิเตอร์ความยาก) a[] (พารามิเตอร์อำนาจจำแนก) z[] (พารามิเตอร์ความสามารถ) และ β[] (พารามิเตอร์สำหรับตัดสินการทำหน้าที่ต่างกันของข้อสอบ)

2.1.6 ให้โปรแกรม R เริ่มการประมวลผลจากโปรแกรม WinBUGS โดยใส่ค่าที่กำหนดไว้แล้วจากข้อ 2.1.4 - 2.1.6 พร้อมกำหนดจำนวนรอบ (Iteration) และจำนวนตัดออก (Burn - in) ในฟังก์ชัน bugs

2.1.7 หลังจากโปรแกรม R เรียกใช้การประมวลผลจากโปรแกรม WinBUGS แล้ว เมื่อประมวลผลเสร็จ ผลการประมาณค่าพารามิเตอร์จะถูกส่งกลับมาที่โปรแกรม R

2.1.8 บันทึกผลการประมาณค่าในรูปแบบไฟล์ข้อมูล (.txt)

โดยดูรายละเอียดได้ที่ ภาคผนวก จ

2.2 พิจารณาค่าพารามิเตอร์ความยาก อำนาจจำแนกและความสามารถที่ประมาณได้จากค่า b[] (พารามิเตอร์ความยาก) a[] (พารามิเตอร์อำนาจจำแนก) และ z[] (พารามิเตอร์ความสามารถ)

2.3 การตัดสินการทำหน้าที่ต่างกันของข้อสอบ พิจารณาจากสมการ (Fukuhara & Kamata, 2011) ดังนี้

$$\ln\left(\frac{P(y_{ij}=1)}{P(y_{ij}=0)}\right) = a_j(\beta_0 G_i + \zeta_i - b_j - \beta_j G_i)$$

เมื่อ a_j หมายถึง พารามิเตอร์อำนาจจำแนก

β_θ หมายถึง อิทธิพลของกลุ่ม G_i ต่อความสามารถ θ_i

G_i หมายถึง กลุ่มของผู้สอบ (G_i= 1 เมื่อผู้สอบเป็นกลุ่ม F และ G_i= 0 เมื่อผู้สอบเป็นกลุ่ม R)

ζ_i หมายถึง ส่วนที่เหลือ (Residual) สำหรับผู้สอบ i

b_j หมายถึง พารามิเตอร์ความยากของข้อ j

β_j หมายถึง ความต่างของพารามิเตอร์ความยากระหว่างกลุ่ม ใช้ตัดสินการทำหน้าที่ต่างกันของข้อสอบ

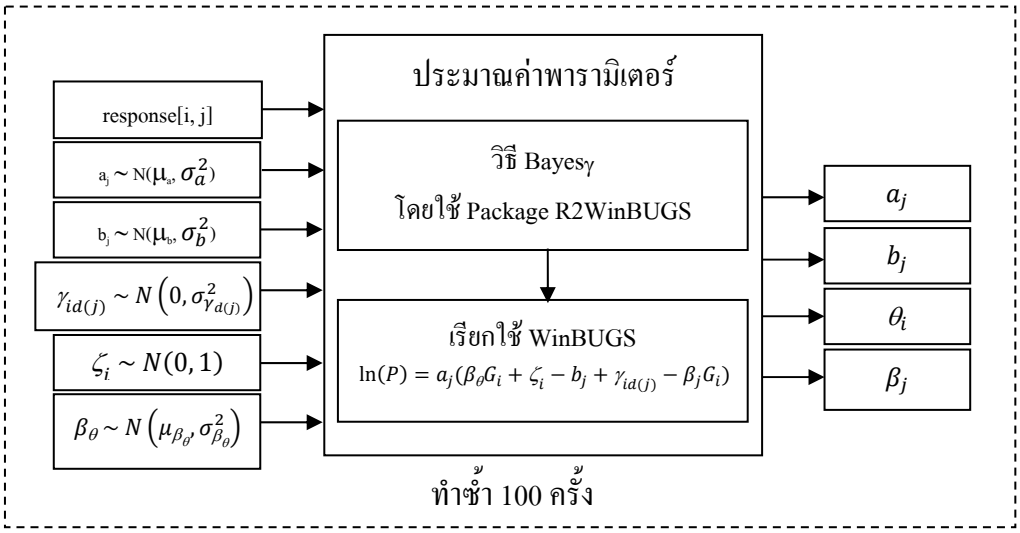
โดยที่การพิจารณาการทำหน้าที่ต่างกันของข้อสอบ จะพิจารณาจากค่า β_j (ในโปรแกรม คือค่า adj.Beta1[]) ซึ่งถ้าค่า β_j มีค่าขอบล่างของช่วงความเชื่อมั่น 95% ซึ่งตรงกับตำแหน่งเปอร์เซ็นไทล์ที่ 2.5 (val 2.5 pc) และค่าขอบบนของช่วงความเชื่อมั่น 95% ซึ่งตรงกับตำแหน่งเปอร์เซ็นไทล์ที่ 97.5 (val 97.5 pc) ไม่คลุม 0 และค่าสัมบูรณ์ของขนาด β_j มากกว่า 0.426 (Vaughn, 2006 อ้างถึงใน สุพัฒนา หอมบุปผา, 2556) แสดงว่าข้อสอบข้อนั้น ทำหน้าที่ต่างกันของข้อสอบ (DIF) ตัวอย่างเช่น

	mean	sd	2.5%	25%	50%	75%	97.5%
adj.Beta1[10]	-0.21064	0.18266	-0.57374	-0.34280	-0.21305	-0.06784	0.10083
adj.Beta1[11]	-0.04309	0.31318	-0.58790	-0.24038	-0.05681	0.13217	0.72551
adj.Beta1[12]	0.06594	0.18803	-0.31055	-0.07541	0.08406	0.18012	0.41110
adj.Beta1[13]	-0.04284	0.24881	-0.50140	-0.21350	-0.05244	0.13582	0.39688
adj.Beta1[14]	0.47768	0.20431	0.15203	0.32005	0.47955	0.63328	0.85216
adj.Beta1[15]	0.31866	0.24969	-0.13800	0.13390	0.32725	0.49782	0.77171
adj.Beta1[16]	-0.00466	0.18656	-0.35751	-0.11980	-0.00205	0.13158	0.33993

แผนภาพที่ 3 - 8 ตัวอย่างผลลัพธ์ของค่า adj.Beta1[] จากการประมาณค่าด้วยวิธี Bayes

จากแผนภาพที่ 3 - 6 เมื่อพิจารณาค่า adj.Beta1[14] หรือการทำหน้าที่ต่างกันของข้อสอบ ข้อที่ 14 พบว่า ค่าขอบล่างของช่วงความเชื่อมั่น 95% และค่าขอบบนของช่วงความเชื่อมั่น 95% ไม่คลุมศูนย์และค่าสัมบูรณ์ของขนาด β_j = 0.477 ซึ่งมากกว่า 0.426 แสดงว่า ข้อที่ 14 มีการทำหน้าที่ต่างกันของข้อสอบ

3. วิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) ประมาณค่าพารามิเตอร์โดยใช้โมเดล Bi - factor MIRT กระทำโดยใช้โปรแกรม R ในการการจำลองข้อมูลแล้วส่งไปประมวลผลด้วยโปรแกรม WinBUGS และกลับมาแสดงผลลัพธ์ที่โปรแกรม R โดยการเรียกใช้ Package R2WinBUGS สรุปขั้นตอนการประมาณค่าพารามิเตอร์ได้ดังแผนภาพที่ 3 - 9



แผนภาพที่ 3 - 9 ขั้นตอนการประมาณค่าพารามิเตอร์ด้วยวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ)

ขั้นตอนการประมาณค่าพารามิเตอร์ด้วยวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) มีรายละเอียด ดังนี้

3.1 นำข้อมูลที่ทำการจำลองในขั้นตอนที่ 1 ประมาณค่าพารามิเตอร์ด้วยวิธีของเบส์แบบมีอิทธิพลเทสต์เลท โดยการเรียกใช้ Package R2WinBUGS ในโปรแกรม R ตามขั้นตอน ดังนี้

3.1.1 สร้างฟังก์ชันสำหรับสร้างตัวแบบเพื่อวิเคราะห์ตัวแบบเบส์ ซึ่งสามารถนำคำสั่งมาจากโปรแกรม WinBUGS แล้วตั้งชื่อเป็น “testletmodel”

3.1.2 บันทึกไฟล์ในข้อ 2.1.1 ชื่อ “Testlet.bug”

3.1.3 กำหนดตัวแปรที่จะนำเข้าไปประมวลผลในโปรแกรม WinBUGS ได้แก่ รูปแบบอิทธิพลเทสต์เลทในแบบสอบ จำนวนผู้สอบ จำนวนข้อสอบ กลุ่ม และแบบแผนการตอบจากการจำลองข้อมูลตามเงื่อนไขต่าง ๆ

3.1.4 สร้างค่าเริ่มต้นของพารามิเตอร์ในโมเดลในรูปของฟังก์ชัน (Initial value) โดยให้อิทธิพลของความสามารถรวมเป็นศูนย์ ($\beta_\theta = 0$) ส่วนค่าอิทธิพลของกลุ่มที่ทำหน้าที่ต่างกันของข้อสอบ ค่าอำนาจจำแนกและค่าความยากเป็นค่าสุ่มจากการแจกแจงที่กำหนดในการจำลองข้อมูล

3.1.5 กำหนดพารามิเตอร์ที่ต้องการให้ตรวจสอบผล ซึ่งผู้วิจัยพิจารณาเฉพาะค่า b[] (พารามิเตอร์ความยาก) a[] (พารามิเตอร์อำนาจจำแนก) z[] (พารามิเตอร์ความสามารถ) และ β[] (พารามิเตอร์สำหรับตัดสินการทำหน้าที่ต่างกันของข้อสอบ)

3.1.6 ให้โปรแกรม R เริ่มการประมวลผลจากโปรแกรม WinBUGS โดยใส่ค่าที่กำหนดไว้แล้วจากข้อ 3.1.4 - 3.1.6 พร้อมกำหนดจำนวนรอบ (Iteration) และจำนวนตัดออก (Burn - in) ในฟังก์ชัน bugs

3.1.7 หลังจากโปรแกรม R เรียกใช้การประมวลผลจากโปรแกรม WinBUGS แล้ว เมื่อประมวลผลเสร็จ ผลการประมาณค่าพารามิเตอร์จะถูกส่งกลับมาที่โปรแกรม R

3.1.8 บันทึกผลการประมาณค่าในรูปแบบไฟล์ข้อมูล (.txt)

โดยดูรายละเอียดได้ที่ ภาคผนวก จ

3.2 จากนั้น พิจารณาการตัดสินการทำหน้าที่ต่างกันของข้อสอบ โดยพิจารณาจากสมการ (Fukuhara & Kamata, 2011) ดังนี้

$$\ln\left(\frac{P(y_{ij}=1)}{P(y_{ij}=0)}\right) = a_j(\beta_\theta G_i + \zeta_i - b_j + \gamma_{id(j)} - \beta_j G_i)$$

เมื่อ a_j หมายถึง พารามิเตอร์อำนาจจำแนก

β_θ หมายถึง อิทธิพลของกลุ่ม G_i ต่อความสามารถ θ_i

G_i หมายถึง กลุ่มของผู้สอบ (G_i= 1 เมื่อผู้สอบเป็นกลุ่ม F และ G_i= 0 เมื่อผู้สอบเป็นกลุ่ม R)

ζ_i หมายถึง ส่วนที่เหลือ (Residual) สำหรับผู้สอบ i

b_j หมายถึง พารามิเตอร์ความยากของข้อ j

$\gamma_{id(j)}$ หมายถึง อิทธิพลสุ่มของ Testlet d(j)

β_j หมายถึง ความต่างของพารามิเตอร์ความยากระหว่างกลุ่ม ใช้พิจารณาการทำหน้าที่ต่างกันของข้อสอบ

ส่วนเกณฑ์การพิจารณาการทำหน้าที่ต่างกันของข้อสอบ ใช้เกณฑ์การพิจารณาเช่นเดียวกับวิธีของเบส์ (Bayes) เช่น

	mean	sd	2.5%	25%	50%	75%	97.5%
adj.Beta1[10]	-0.27727	0.20655	-0.61514	-0.42875	-0.28185	-0.15682	0.16916
adj.Beta1[11]	0.01133	0.31850	-0.61283	-0.19232	-0.01298	0.22700	0.66938
adj.Beta1[12]	0.02916	0.19782	-0.38271	-0.09803	0.03704	0.16568	0.38981
adj.Beta1[13]	-0.04637	0.23685	-0.55977	-0.18800	-0.03482	0.11072	0.36079
adj.Beta1[14]	0.50373	0.21837	0.10368	0.36842	0.50585	0.64440	0.91672
adj.Beta1[15]	0.26375	0.22046	-0.17012	0.12660	0.24865	0.40682	0.70084
adj.Beta1[16]	-0.09600	0.19543	-0.48150	-0.19905	-0.10475	0.02407	0.27676

แผนภาพที่ 3 - 10 ตัวอย่างผลลัพธ์ของค่า adj.Beta1[] จากการประมาณค่าด้วยวิธี Bayesγ

จากแผนภาพที่ 3 - 6 เมื่อพิจารณาค่า adj.Beta1[14] หรือการทำหน้าที่ต่างกันของข้อสอบ ข้อที่ 14 พบว่า ค่าขอบล่างของช่วงความเชื่อมั่น 95% และค่าขอบบนของช่วงความเชื่อมั่น 95% ไม่คลุม 0 และค่าสัมบูรณ์ของขนาด β_j = 0.504 ซึ่งมากกว่า 0.426 แสดงว่า ข้อที่ 14 มีการทำหน้าที่ต่างกันของข้อสอบ

ขั้นตอนที่ 3 การวัดประสิทธิผล แบ่งเป็น 2 ส่วน ได้แก่ การวัดประสิทธิภาพของการประมาณค่าและการวัดประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

1. วัดประสิทธิผลของการประมาณค่า โดยพิจารณาจากความเบี่ยงเบนของค่าพารามิเตอร์ที่แท้จริงและค่าที่ประมาณได้ หรือ Root Mean Square Error (RMSE) ของพารามิเตอร์ความยาก อำนาจจำแนกและพารามิเตอร์ความสามารถของผู้สอบ คำนวณจากสมการ ดังนี้ (Jiao et al., 2013)

$$RMSE(\hat{\beta}) = \sqrt{\frac{1}{N}\sum_{r=1}^{N}(\hat{\beta}_r - \beta)^2}$$

เมื่อ β เป็น พารามิเตอร์จริง

$\hat{\beta}_r$ เป็น พารามิเตอร์ที่ประมาณค่า ครั้งที่ r

N เป็น จำนวนครั้งที่ทำซ้ำในการจำลองข้อมูล

จากนั้นทำการทดสอบด้วยวิธี 5 - way Analyses of Variance (5 - Way ANOVA) เพื่อเปรียบเทียบความแตกต่างของ 4 เงื่อนไข (4 Factors) และวิธีการประมาณค่าพารามิเตอร์ (1 Factor) รวมเป็น 5 Factors ซึ่งกำหนดให้เป็นตัวแปรต้น (Independent variables) ได้แก่ อิทธิพลเทสต์เลท การแจกแจงของความสามารถ จำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบ และอัตราส่วนของกลุ่มเปรียบเทียบต่อกลุ่มอ้างอิง และวิธีการประมาณค่าพารามิเตอร์ และกำหนดให้ตัวแปรตาม (Dependent variables) เป็นความเบี่ยงเบนของค่าพารามิเตอร์ที่แท้จริงและค่าที่ประมาณได้ (RMSE)

2. การวัดประสิทธิผลของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

หลังจากวิเคราะห์ด้วยวิธีทั้งสามแล้ว นำผลการวิเคราะห์ที่ได้มาคำนวณหาอัตราความคลาดเคลื่อนประเภทที่ 1 และค่าอำนาจของตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ดังนี้

2.1 อัตราความคลาดเคลื่อนประเภทที่ 1 (Type I error rate) ของตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ หมายถึง จำนวนครั้งที่ปฏิเสธสมมติฐานหลักเมื่อสมมติฐานหลักเป็นจริง หารด้วยจำนวนครั้งของการทดสอบ ในทางปฏิบัติ คือ จำนวนข้อสอบที่ระบุว่าทำหน้าที่ต่างกันทั้งที่ความเป็นจริงข้อสอบทำหน้าที่ไม่ต่างกัน หารด้วยจำนวนข้อสอบที่ทำหน้าที่ไม่ต่างกันทั้งหมด เขียนเป็นสมการได้เป็น

$$\text{อัตราความคาดเคลื่อนประเภทที่ } 1\ (\mathrm{p}) = \frac{\sum \text{จำนวนข้อสอบที่ระบุว่าทำหน้าที่ต่างกัน}}{\text{จำนวนข้อสอบที่ทำหน้าที่ไม่ต่างกันทั้งหมด}} \times 100$$

จากนั้นทำการทดสอบสมมติฐานเปรียบเทียบอัตราความคลาดเคลื่อนประเภทที่ 1 ที่ได้จากการคำนวณ โดยมีขั้นตอนในการทดสอบสมมติฐานดังนี้

1. ตั้งสมมติฐานเพื่อการทดสอบ ดังนี้

H_0: $p_0 \leq .05$

H_a: $p_0 > .05$

2. สถิติที่ใช้ในการทดสอบ

$$Z = \frac{p - p_0}{\sqrt{[p_0(1-p_0)/n]}}$$

เมื่อ p = สัดส่วนของการปฏิเสธสมมติฐานหลักจากการจำลอง

p_0 = .05

n = จำนวนครั้งที่ทำซ้ำในการจำลองข้อมูล

3. คำนวณค่าสถิติและเปรียบเทียบกับค่าวิกฤตที่ระดับนัยสำคัญ α = .05 หากค่าสถิติที่คำนวณได้ตกอยู่ในบริเวณวิกฤตจะปฏิเสธสมมติฐานศูนย์ (Reject H_0) และยอมรับสมมติฐานทางเลือก (Accept H_a) แต่ถ้าค่าสถิติที่คำนวณได้ตกอยู่นอกบริเวณวิกฤต จะยอมรับสมมติฐานศูนย์ (Accept H_0) หรืออาจสรุปได้ว่ายังไม่มีเหตุผลเพียงพอที่จะปฏิเสธสมมติฐานศูนย์ (Retain H_0)

จากนั้นทำการทดสอบด้วยวิธี 5 - way Analyses of Variance (5 - Way ANOVA) เพื่อเปรียบเทียบความแตกต่างของ 4 เงื่อนไข (4 Factors) และวิธีการประมาณค่าพารามิเตอร์ (1 Factor) รวมเป็น 5 Factors ซึ่งกำหนดให้เป็นตัวแปรต้น (Independent variables) ได้แก่ อิทธิพลเทสต์เลท การแจกแจงของความสามารถ จำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบ และอัตราส่วนของกลุ่มเปรียบเทียบต่อกลุ่มอ้างอิง และวิธีการประมาณค่าพารามิเตอร์ และกำหนดให้ตัวแปรตาม (Dependent variables) เป็นอัตราความคาดเคลื่อนประเภทที่ 1 ดังภาพ ที่ 3 - 11

2.2 อำนาจของตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (Power) หมายถึง จำนวนข้อสอบที่ระบุว่าทำหน้าที่ต่างกันได้ถูก ซึ่งคำนวณได้จาก

$$\text{อำนาจการตรวจสอบ (p)} = \frac{\sum \text{จำนวนข้อสอบที่ระบุว่าทำหน้าที่ต่างกัน}}{\text{จำนวนข้อสอบที่ทำหน้าที่ต่างกันทั้งหมดในแบบสอบ}} \text{ x } 100$$

จากนั้นทำการทดสอบสมมติฐานเปรียบเทียบอัตราอำนาจการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยมีขั้นตอนในการทดสอบสมมติฐานดังนี้

1. ตั้งสมมติฐานเพื่อการทดสอบ ดังนี้

   H_0: $p_0 \geq .80$

   H_a: $p_0 < .80$

2. สถิติที่ใช้ในการทดสอบ

$$Z = \frac{p - p_0}{\sqrt{[p_0(1-p_0)/n]}}$$

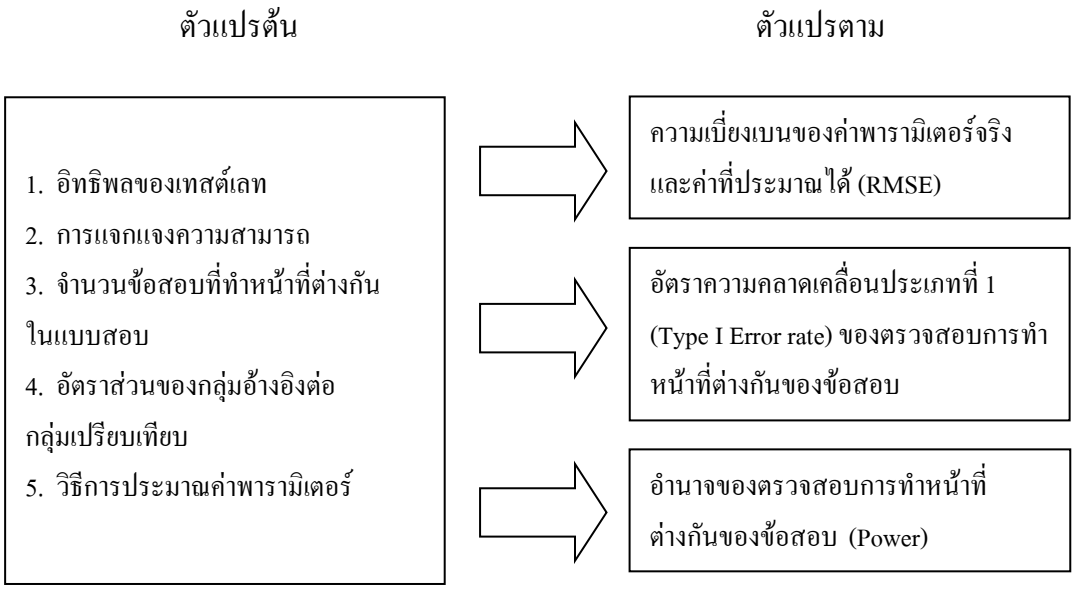
เมื่อ p = สัดส่วนของการยอมรับจากข้อมูลจำลอง

p_0 = .80

n = จำนวนครั้งที่ทำซ้ำในการจำลองข้อมูล

3. คำนวณค่าสถิติและเปรียบเทียบกับค่าวิกฤตที่ระดับนัยสำคัญ α = .05 หากค่าสถิติที่คำนวณได้ตกอยู่ในบริเวณวิกฤตจะปฏิเสธสมมติฐานศูนย์ (Reject H_0) และยอมรับสมมติฐานทางเลือก (Accept H_a) แต่ถ้าค่าสถิติที่คำนวณได้ตกอยู่นอกบริเวณวิกฤต จะยอมรับสมมติฐานศูนย์ (Accept H_0) หรืออาจสรุปได้ว่ายังไม่มีเหตุผลเพียงพอที่จะปฏิเสธสมมติฐานศูนย์ (Retain H_0)

จากนั้นทำการทดสอบด้วยวิธี 5 - way Analyses of Variance (5 - Way ANOVA) เพื่อเปรียบเทียบความแตกต่างของ 4 เงื่อนไขและวิธีการประมาณค่าพารามิเตอร์ (5 Factors) ซึ่งกำหนดให้เป็นตัวแปรต้น (Independent variables) ได้แก่ อิทธิพลของเทสต์เลท การแจกแจงของความสามารถ จำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบ อัตราส่วนของกลุ่มเปรียบเทียบต่อกลุ่มอ้างอิง และวิธีการประมาณค่าพารามิเตอร์ และกำหนดให้ตัวแปรตาม (Dependent variables) เป็นอำนาจการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ดังภาพที่ 3 - 11



ภาพที่ 3 - 11 การเปรียบเทียบความแตกต่างของ 5 เงื่อนไขด้วยวิธี 5 - Way ANOVA

บทที่ 4
ผลการวิเคราะห์ข้อมูล

การวิจัยครั้งนี้มีวัตถุประสงค์ 2 ประการ ประการแรกเพื่อศึกษาผลการประมาณค่าพารามิเตอร์ของข้อสอบ (ความยากและอำนาจจำแนก) กับพารามิเตอร์ความสามารถของผู้สอบ (θ) ด้วยวิธีแมกซิมัมไลค์ลิฮู้ด (ML) วิธีของเบส์ (Bayes) และวิธีของเบส์แบบมีอิทธิพล เทสต์เลท (Bayesγ) และประการที่สองเพื่อศึกษาผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ด้วยวิธีแมกซิมัมไลค์ลิฮู้ด (ML) วิธีของเบส์ (Bayes) และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) โดยการศึกษาข้อมูลที่มีอิทธิพลเทสต์เลท การแจกแจงของความสามารถ จำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบและอัตราส่วนของกลุ่มเปรียบเทียบต่อกลุ่มอ้างอิงที่ต่างกัน ดังนั้น ผู้วิจัยจึงนำเสนอผลการวิเคราะห์เป็น 2 ตอน ดังนี้

ตอนที่ 1 ผลการประมาณค่าพารามิเตอร์ของข้อสอบ (ความยากและอำนาจจำแนก) และพารามิเตอร์ความสามารถของผู้สอบ

ตอนที่ 2 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

เพื่อให้การนำเสนอผลการวิเคราะห์ข้อมูลมีความเข้าใจตรงกัน ผู้วิจัยจึงได้กำหนดสัญลักษณ์และความหมายแทนตัวแปรที่ศึกษาต่าง ๆ ดังนี้

ML	หมายถึง	การประมาณค่าพารามิเตอร์ด้วยวิธีแมกซิมัมไลค์ลิฮู้ด (ML)
Bayes	หมายถึง	การประมาณค่าพารามิเตอร์ด้วยวิธีของเบส์
Bayesγ	หมายถึง	การประมาณค่าพารามิเตอร์ด้วยวิธีของเบส์แบบมีอิทธิพลเทสต์เลท
RMSE	หมายถึง	ความเบี่ยงเบนของค่าพารามิเตอร์ความยากที่แท้จริงและค่าที่ประมาณได้
C1…C54	หมายถึง	เงื่อนไขที่ศึกษา

ตอนที่ 1 ผลการประมาณค่าพารามิเตอร์ของข้อสอบ (ความยากและอำนาจจำแนก) และพารามิเตอร์ความสามารถของผู้สอบ

ผลการประมาณค่าพารามิเตอร์ของข้อสอบ (ความยากและอำนาจจำแนก) และพารามิเตอร์ความสามารถของผู้สอบ ด้วยวิธีแมกซิมัมไลค์ลิฮู้ด วิธีของเบส์และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท ดังตารางที่ 4 - 1 ถึง ตารางที่ 4 - 6

ตารางที่ 4 - 1 ค่าเฉลี่ยดัชนี RMSE ของพารามิเตอร์ความยาก

อิทธิพลเทสต์เลท	การแจกแจงความสามารถ	จำนวนข้อ DIF	จำนวนกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ	เงื่อนไข	ML	Bayes	Bayes$_{\gamma}$
อิทธิพลเท่ากันทุกเทสต์เลท (0.8, 0.8, 0.8, 0.8)	ปกติ	0	1000: 1000	C1	0.1999	0.0701	**0.0085**
			1000: 100	C2	0.3737	0.0463	**0.0406**
		5	1000: 1000	C3	0.1771	0.0136	**0.0089**
			1000: 100	C4	0.3452	0.0495	**0.0455**
		8	1000: 1000	C5	0.1781	0.1916	**0.1669**
			1000: 100	C6	0.3082	**0.0484**	0.0500
	เบ้ซ้าย	0	1000: 1000	C7	0.4374	**0.2628**	0.5219
			1000: 100	C8	0.7567	**0.2892**	0.5554
		5	1000: 1000	C9	0.4126	**0.2520**	0.5032
			1000: 100	C10	0.6863	**0.2910**	0.5492
		8	1000: 1000	C11	0.3741	**0.2687**	0.5304
			1000: 100	C12	0.6410	**0.3021**	0.5825
	เบ้ขวา	0	1000: 1000	C13	**0.1130**	0.2962	1.0238
			1000: 100	C14	**0.1825**	0.2440	0.4539
		5	1000: 1000	C15	**0.1077**	0.2503	0.4583
			1000: 100	C16	**0.1661**	0.2562	0.4547
		8	1000: 1000	C17	**0.1298**	0.2725	0.5149
			1000: 100	C18	**0.1864**	0.3207	0.5829

ตารางที่ 4 - 1 (ต่อ)

อิทธิพลเทสต์เลท	การแจกแจงความสามารถ	จำนวนข้อ DIF	จำนวนกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ	เงื่อนไข	ML	Bayes	$Bayes_{\gamma}$
อิทธิพลเทสต์เลทไม่เท่ากัน (0.25, 0.5, 1, 2)	ปกติ	0	1000: 1000	C19	0.1666	0.0886	**0.0314**
			1000: 100	C20	0.3947	0.1125	**0.0585**
		5	1000: 1000	C21	0.1646	0.0784	**0.0259**
			1000: 100	C22	0.3414	0.1169	**0.0651**
		8	1000: 1000	C23	0.1485	0.0816	**0.0287**
			1000: 100	C24	0.2852	0.1144	**0.0709**
	เบ้ซ้าย	0	1000: 1000	C25	**0.4194**	2.8570	0.5794
			1000: 100	C26	0.7845	2.2627	**0.6001**
		5	1000: 1000	C27	**0.4062**	2.8091	0.5909
			1000: 100	C28	0.6920	2.2610	**0.6552**
		8	1000: 1000	C29	**0.3856**	2.3802	0.5745
			1000: 100	C30	**0.6941**	2.9171	0.7012
	เบ้ขวา	0	1000: 1000	C31	**0.0952**	2.3446	0.5225
			1000: 100	C32	**0.1826**	1.7876	0.5400
		5	1000: 1000	C33	**0.1192**	2.1229	0.4777
			1000: 100	C34	**0.1734**	2.0014	0.5284
		8	1000: 1000	C35	**0.1318**	2.4329	0.5252
			1000: 100	C36	**0.1503**	1.7446	0.4769
ข้อสอบที่เป็นอิสระและเทสต์เลท (0, 0.25, 0.56, 1)	ปกติ	0	1000: 1000	C37	0.2741	0.0316	**0.0262**
			1000: 100	C38	0.4804	0.0627	**0.0588**
		5	1000: 1000	C39	0.2566	0.0330	**0.0269**
			1000: 100	C40	0.4663	0.0691	**0.0635**
		8	1000: 1000	C41	0.2598	0.0386	**0.0335**
			1000: 100	C42	0.4208	0.0829	**0.0798**
	เบ้ซ้าย	0	1000: 1000	C43	**0.5299**	0.5641	0.6251
			1000: 100	C44	0.9617	**0.6331**	0.6971
		5	1000: 1000	C45	**0.5020**	0.6007	0.5921
			1000: 100	C46	0.8238	**0.5312**	0.6064

ตารางที่ 4 - 1 (ต่อ)

อิทธิพลเทสต์เลท	การแจกแจงความสามารถ	จำนวนข้อ DIF	จำนวนกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ	เงื่อนไข	ML	Bayes	Bayesγ
ข้อสอบที่เป็นอิสระและเทสต์เลท (0, 0.25, 0.56, 1)	เบ้ขวา	8	1000: 1000	C47	**0.4825**	0.5045	0.5520
			1000: 100	C48	0.8082	**0.5649**	0.6572
		0	1000: 1000	C49	**0.1714**	0.5145	0.5679
			1000: 100	C50	**0.2463**	0.4692	0.5596
		5	1000: 1000	C51	**0.1858**	0.5285	0.5551
			1000: 100	C52	**0.2334**	0.4872	0.5460
		8	1000: 1000	C53	**0.1920**	0.5805	0.5944
			1000: 100	C54	**0.2697**	0.5420	0.6044

หมายเหตุ ตัวหนา คือ วิธีที่มีค่าเฉลี่ยดัชนี RMSE น้อยที่สุดในเงื่อนไขเดียวกัน

จากตารางที่ 4 - 1 มีค่าเฉลี่ยดัชนี RMSE ของพารามิเตอร์ความยาก มีค่าระหว่าง 0.0085 - 2.9171 เมื่อมองภาพรวมอาจยังไม่เห็นแนวโน้มสำหรับการประมาณค่าพารามิเตอร์ความยากของทั้งสามวิธี ดังนั้น หากพิจารณาตามวิธีที่ใช้ประมาณค่า พบว่า

เมื่อพิจารณาเฉพาะวิธีแมกซิมัมไลค์ลิฮู้ด (ML) พบว่าค่าเฉลี่ยดัชนี RMSE มีค่าระหว่าง 0.0952 - 0.9617 และประมาณค่าพารามิเตอร์ความยากได้ดีที่สุด จำนวน 25 เงื่อนไข ส่วนใหญ่ค่าเฉลี่ยดัชนี RMSE มีค่าน้อย เมื่อความสามารถมีการแจกแจงแบบเบ้ขวา (เงื่อนไข C13 - C18) และเมื่อจำนวนผู้สอบมากจะประมาณค่าได้ดีกว่าจำนวนผู้สอบน้อย

เมื่อพิจารณาวิธีของเบส์ (Bayes) พบว่า ค่าเฉลี่ยดัชนี RMSE มีค่าระหว่าง 0.0136 - 2.9171 และประมาณค่าพารามิเตอร์ความยากได้ดีที่สุด จำนวน 10 เงื่อนไข โดยส่วนใหญ่ค่าเฉลี่ยดัชนี RMSE มีค่าน้อย เมื่อความสามารถมีการแจกแจงแบบปกติ (เงื่อนไข C13 - C18, C19 - C24, C37 - C42) และค่าเฉลี่ยดัชนี RMSE มีค่าสูง เมื่ออิทธิพลของเทสเลทไม่เท่ากัน นั่นคือ มีค่าเป็น 0.25, 0.5, 1, 2 ร่วมกับการแจกแจงความสามารถที่เป็นแบบเบ้ซ้ายและเบ้ขวา (เงื่อนไข C25 - C36)

เมื่อพิจารณาวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) พบว่า ค่าเฉลี่ยดัชนี RMSE มีค่าระหว่าง 0.0085 - 1.0238 ประมาณค่าพารามิเตอร์ความยากได้ดีที่สุด จำนวน 19 เงื่อนไข โดยส่วนใหญ่ค่าเฉลี่ยดัชนี RMSE มีค่าน้อย เมื่อความสามารถมีการแจกแจงแบบปกติ (เงื่อนไข C13 - C18, C19 - C24, C37 - C42)

ตารางที่ 4 - 2 ผลการเปรียบเทียบความแปรปวน 5 ทาง (5 - Way ANOVA) ของค่าเฉลี่ยดัชนี RMSE ของพารามิเตอร์ความยาก

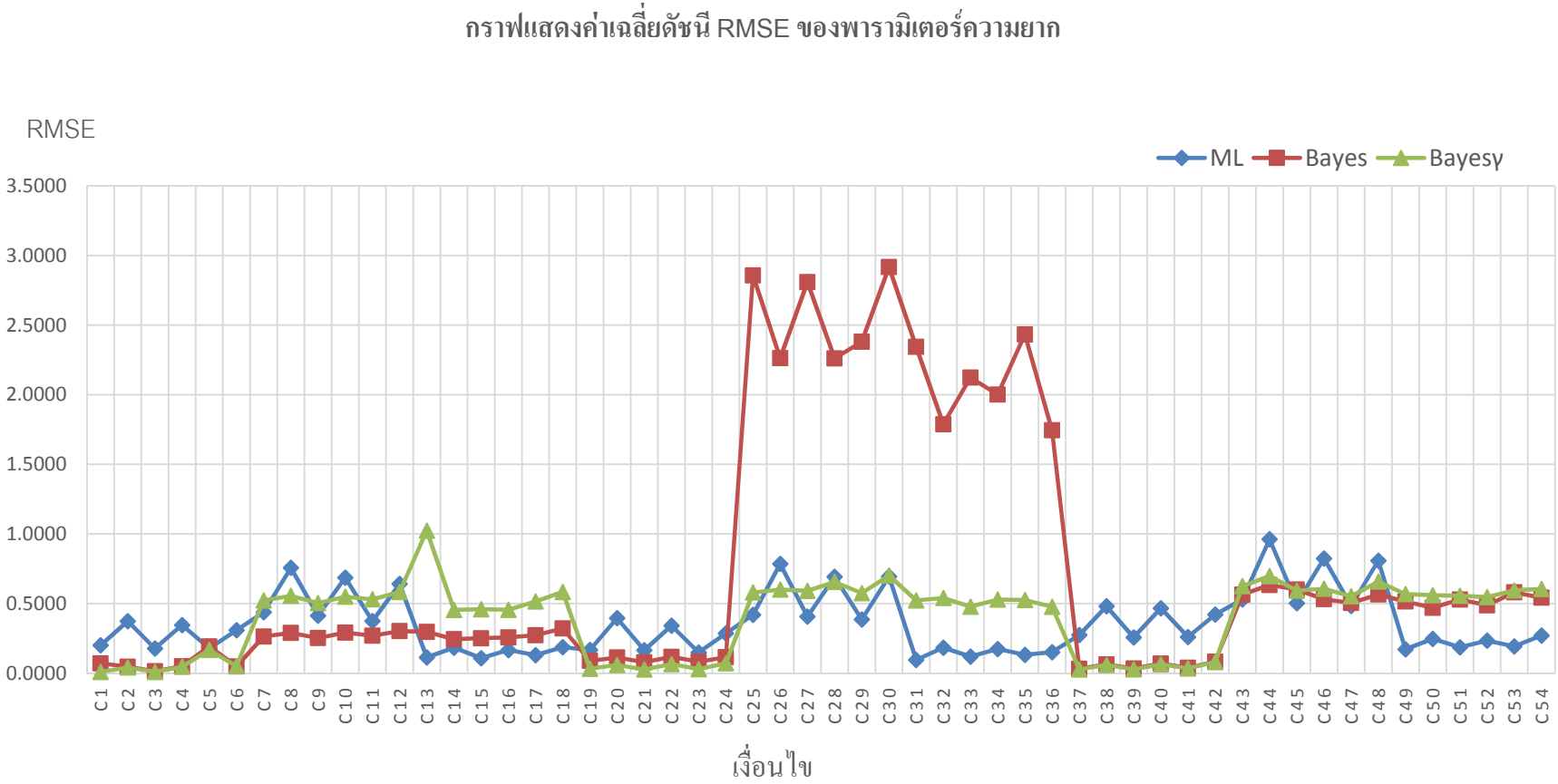
แหล่งความแปรปรวน	SS	df	MS	F	Sig.
TestEff	6.2898	2	3.1449	261.5273**	.0000
Dist	11.4312	2	5.7156	475.3051**	.0000
NumDIF	0.0195	2	0.0098	0.8115	.4617
RperF	0.0710	1	0.0710	5.9006*	.0273
Med	4.4035	2	2.2017	183.0952**	.0000
TestEff * Dist	3.4641	4	0.8660	72.0171**	.0000
TestEff * NumDIF	0.0073	4	0.0018	0.1507	.9599
TestEff * RperF	0.0401	2	0.0201	1.6680	.2198
TestEff * Med	14.0958	4	3.5240	293.0494**	.0000
Dist * NumDIF	0.0068	4	0.0017	0.1415	.9642
Dist * RperF	0.1954	2	0.0977	8.1231**	.0037
Dist * Med	5.3724	4	1.3431	111.6912**	.0000
NumDIF * RperF	0.0129	2	0.0064	0.5351	.5957
NumDIF * Med	0.0238	4	0.0059	0.4941	.7403
RperF * Med	0.4634	2	0.2317	19.2659**	.0001
TestEff * Dist * NumDIF	0.0398	8	0.0050	0.4140	.8960
TestEff * Dist * RperF	0.0314	4	0.0079	0.6538	.6326
TestEff * Dist * Med	6.6378	8	0.8297	68.9991**	.0000
TestEff * NumDIF * RperF	0.0221	4	0.0055	0.4589	.7647
TestEff * NumDIF * Med	0.0122	8	0.0015	0.1272	.9970
TestEff * RperF * Med	0.1153	4	0.0288	2.3976	.0934
Dist * NumDIF * RperF	0.0818	4	0.0204	1.7005	.1990
Dist * NumDIF * Med	0.0338	8	0.0042	0.3515	.9313
Dist * RperF * Med	0.0600	4	0.0150	1.2475	.3308
NumDIF * RperF * Med	0.0366	4	0.0091	0.7599	.5664
TestEff * Dist * NumDIF * RperF	0.1518	8	0.0190	1.5777	.2081
TestEff * Dist * NumDIF * Med	0.0283	16	0.0018	0.1473	.9998
TestEff * Dist * RperF * Med	0.0998	8	0.0125	1.0375	.4492
TestEff * NumDIF * RperF * Med	0.0603	8	0.0075	0.6271	.7439
Dist * NumDIF * RperF * Med	0.1059	8	0.0132	1.1013	.4115
Error	0.1924	16	0.0120		

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

จากตารางที่ 4 - 2 ผลการเปรียบเทียบความแปรปวน 5 ทาง (5 - Way ANOVA) ของค่าเฉลี่ยดัชนี RMSE ของพารามิเตอร์ความยาก พบว่า อิทธิพลหลัก (Main effect) ที่มีนัยสำคัญทางสถิติ ได้แก่ อิทธิพลเทสต์เลท (TestEff) การแจกแจงความสามารถ (Dist) และวิธีที่ใช้ประมาณค่า (Med) มีผลต่อ RMSE อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ส่วนอัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ (RperF) มีผลต่อ RMSE อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นั่นคือ หากปัจจัยหลักมีค่าต่างกัน จะมีค่าเฉลี่ยดัชนี RMSE ของพารามิเตอร์ความยากต่างกัน ส่วนจำนวนข้อที่ทำหน้าที่ต่างกันในแบบสอบค่าเฉลี่ยดัชนี RMSE ของพารามิเตอร์ความยากไม่แตกต่างกัน

เมื่อพิจารณาอิทธิพลร่วม (Interection effect) พบว่า อิทธิพลเทสต์เลทกับการแจกแจงความสามารถ (TestEff * Dist) อิทธิพลเทสต์เลทกับวิธีที่ใช้ประมาณค่า (TestEff * Med) การแจกแจงความสามารถกับอัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ (Dist * RperF) การแจกแจงความสามารถกับวิธีที่ใช้ประมาณค่า (Dist * Med) อัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบกับวิธีที่ใช้ประมาณค่า (RperF * Med) และอิทธิพลเทสต์เลทกับการแจกแจงความสามารถและวิธีที่ใช้ในการประมาณค่า (TestEff * Dist * Med) มีผลต่อ RMSE อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 หมายความว่า เมื่อพิจารณาอิทธิพลเทสต์เลทร่วมกับการแจกแจงความสามารถและวิธีที่ใช้ในการประมาณค่า (TestEff * Dist * Med) พบว่า มีนัยสำคัญทางสถิติ นั่นคือ หากพิจารณาจากภาพที่ 4 - 1 ในช่วงที่อิทธิพลเทสเลทมีค่าไม่เท่ากัน นั่นคือ มีค่าเป็น 0.25, 0.5, 1, 2 ร่วมกับการแจกแจงความสามารถที่เป็นแบบเบ้ซ้ายและเบ้ขวา (เงื่อนไข C25 - C36) ค่าเฉลี่ยดัชนี RMSE ของวิธีของเบส์มีค่าสูงขึ้นอย่างเห็นได้ชัด แสดงว่า ค่าเฉลี่ยดัชนี RMSE ที่เปลี่ยนไปนั้นไม่ได้ขึ้นอยู่กับอิทธิพลของเทสต์เลทอย่างเดียว แต่ต้องพิจารณาอิทธิพลของเทสต์เลทร่วมกับการแจกแจงความสามารถและวิธีที่ใช้ในการประมาณค่าด้วย

เมื่อพิจารณาจากภาพที่ 4 - 1 ในภาพรวมจะพบว่า วิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayes$_\gamma$) ประมาณค่าได้ดี เมื่อข้อมูลมีการแจกแจงความสามารถเป็นแบบปกติ ส่วนวิธีของเบส์ (Bayes) ยังไม่มีแนวโน้มแน่นอน แต่ส่วนใหญ่ประมาณค่าพารามิเตอร์ความยากได้ดี เมื่อข้อมูลมีการแจกแจงความสามารถเป็นแบบเบ้ซ้าย ส่วนวิธีแมกซิมัมไลค์ลิฮู้ด (ML) ประมาณค่าพารามิเตอร์ความยากได้ดี เมื่อข้อมูลมีการแจกแจงความสามารถเป็นแบบเบ้ขวา และทั้ง 3 วิธีมีค่าเฉลี่ยดัชนี RMSE ต่างกันมาก เมื่ออิทธิพลเทสต์เลทมีค่าสูงมากและมีค่าต่างกัน โดยที่ในกรณีนี้วิธีแมกซิมัมไลค์ลิฮู้ด (ML) ประมาณค่าพารามิเตอร์ความยากได้ดีกว่าวิธีอื่น



ภาพที่ 4 - 1 ค่าเฉลี่ยดัชนี RMSE ของพารามิเตอร์ความยาก จำแนกตามเงื่อนไขและวิธีการประมาณค่าพารามิเตอร์

ตารางที่ 4 - 3 ค่าเฉลี่ยดัชนี RMSE ของพารามิเตอร์อำนาจจำแนก

อิทธิพลเทสต์เลท	การแจกแจงความสามารถ	จำนวนข้อ DIF	จำนวนกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ	เงื่อนไข	ML	Bayes	$Bayes_{\gamma}$
อิทธิพลเท่ากันทุกเทสต์เลท (0.8, 0.8, 0.8, 0.8)	ปกติ	0	1000: 1000	C1	0.0219	0.0369	**0.0085**
			1000: 100	C2	0.0314	0.0378	**0.0144**
		5	1000: 1000	C3	0.0255	0.0331	**0.0082**
			1000: 100	C4	0.0314	0.0369	**0.0149**
		8	1000: 1000	C5	0.0237	0.0288	**0.0081**
			1000: 100	C6	0.0369	0.0420	**0.0160**
	เบ้ซ้าย	0	1000: 1000	C7	**0.2147**	0.2651	0.2482
			1000: 100	C8	**0.2397**	0.2627	0.2547
		5	1000: 1000	C9	**0.2176**	0.2538	0.2358
			1000: 100	C10	**0.2550**	0.2699	0.2606
		8	1000: 1000	C11	**0.2157**	0.2414	0.2270
			1000: 100	C12	**0.2443**	0.2571	0.2465
	เบ้ขวา	0	1000: 1000	C13	**0.2003**	0.2516	0.2729
			1000: 100	C14	0.2132	0.2333	**0.2120**
		5	1000: 1000	C15	**0.2086**	0.2413	0.2171
			1000: 100	C16	0.2265	0.2410	**0.2211**
		8	1000: 1000	C17	**0.2251**	0.2520	0.2317
			1000: 100	C18	0.2336	0.2492	**0.2262**
อิทธิพลเทสต์เลทไม่เท่ากัน (0.25, 0.5, 1, 2)	ปกติ	0	1000: 1000	C19	0.1626	0.1836	**0.0117**
			1000: 100	C20	0.1428	0.1365	**0.0176**
		5	1000: 1000	C21	0.1158	0.1194	**0.0130**
			1000: 100	C22	0.1953	0.1605	**0.0175**
		8	1000: 1000	C23	0.1479	0.1409	**0.0121**
			1000: 100	C24	0.2067	0.1708	**0.0201**
	เบ้ซ้าย	0	1000: 1000	C25	1.0726	1.1027	**0.2472**
			1000: 100	C26	1.1288	1.0146	**0.2571**
		5	1000: 1000	C27	1.0723	1.0881	**0.2537**
			1000: 100	C28	1.0833	0.9571	**0.2546**

ตารางที่ 4 - 3 (ต่อ)

อิทธิพลเทสต์เลท	การแจกแจงความสามารถ	จำนวนข้อ DIF	จำนวนกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ	เงื่อนไข	ML	Bayes	$Bayes_{\gamma}$
อิทธิพลเทสต์เลทไม่เท่ากัน (0.25, 0.5, 1, 2)	เบ้ขวา	8	1000: 1000	C29	1.0397	1.0201	**0.2339**
			1000: 100	C30	1.2219	1.0971	**0.2632**
		0	1000: 1000	C31	0.9383	0.9883	**0.2313**
			1000: 100	C32	1.0282	0.9334	**0.2421**
		5	1000: 1000	C33	0.9628	0.9797	**0.2240**
			1000: 100	C34	1.0433	0.9462	**0.2308**
		8	1000: 1000	C35	1.0958	1.1115	**0.2355**
			1000: 100	C36	1.0044	0.9041	**0.2326**
ข้อสอบที่เป็นอิสระและเทสต์เลท (0, 0.25, 0.56, 1)	ปกติ	0	1000: 1000	C37	0.0129	0.0207	**0.0110**
			1000: 100	C38	0.0209	0.0276	**0.0183**
		5	1000: 1000	C39	0.0149	0.0194	**0.0106**
			1000: 100	C40	0.0213	0.0247	**0.0177**
		8	1000: 1000	C41	0.0156	0.0201	**0.0109**
			1000: 100	C42	0.0205	0.0248	**0.0155**
	เบ้ซ้าย	0	1000: 1000	C43	**0.2323**	0.2918	0.2680
			1000: 100	C44	0.2777	0.2844	**0.2679**
		5	1000: 1000	C45	**0.2523**	0.2877	0.2560
			1000: 100	C46	0.2847	0.2917	**0.2726**
		8	1000: 1000	C47	**0.2450**	0.2740	0.2481
			1000: 100	C48	0.2807	0.2820	**0.2655**
	เบ้ขวา	0	1000: 1000	C49	**0.2055**	0.2621	0.2427
			1000: 100	C50	**0.2347**	0.2508	0.2354
		5	1000: 1000	C51	**0.2246**	0.2610	0.2358
			1000: 100	C52	0.2447	0.2554	**0.2371**
		8	1000: 1000	C53	0.2369	0.2617	**0.2322**
			1000: 100	C54	0.2522	0.2586	**0.2426**

หมายเหตุ ตัวหนา คือ วิธีที่มีค่าเฉลี่ยดัชนี RMSE น้อยที่สุดในเงื่อนไขเดียวกัน

จากตารางที่ 4 - 3 ค่าเฉลี่ยดัชนี RMSE ของพารามิเตอร์อำนาจจำแนก โดยรวม มีค่าระหว่าง 0.0081 - 1.2219 หากพิจารณาตามวิธีที่ใช้ในการประมาณค่า พบว่า วิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) มีค่าระหว่าง 0.0081 - 0.2729 ประมาณค่าพารามิเตอร์อำนาจจำแนกได้ดีที่สุด จำนวน 39 เงื่อนไข วิธีแมกซิมัมไลค์ลิฮู้ด (ML) มีค่าระหว่าง 0.0129 - 1.2219 ประมาณค่าพารามิเตอร์อำนาจจำแนกได้ดีที่สุด จำนวน 15 เงื่อนไข ส่วนวิธีของเบส์ (Bayes) มีค่าระหว่าง 0.0194 - 1.1115 และไม่มีเงื่อนไขใดที่วิธีนี้ประมาณค่าพารามิเตอร์อำนาจจำแนกได้ดีที่สุด เมื่อพิจารณาจากภาพที่ 4 – 2 แล้ว พบว่า ส่วนใหญ่วิธีการประมาณค่าด้วยวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) จะประมาณค่าพารามิเตอร์อำนาจจำแนกได้ดีกว่าเมื่อข้อมูลความสามารถมีการแจกแจงแบบปกติ หรือเมื่อค่าอิทธิพลของเทสต์เลทไม่เท่ากัน (Unequal effect) โดยกำหนดให้อิทธิพลของเทสต์เลทมีค่าเป็น 0.25, 0.5, 1 และ 2 ตามลำดับ ส่วนวิธีแมกซิมัมไลค์ลิฮู้ด (ML) ส่วนใหญ่จะประมาณค่าพารามิเตอร์อำนาจจำแนกได้ดีเมื่อข้อมูลความสามารถมีการแจกแจงแบบ เบ้ซ้าย และไม่มีเงื่อนไขใดที่วิธีของเบส์แบบไม่คำนึงถึงอิทธิพลของเทสต์เลท (Bayes) ประมาณได้ดีที่สุด

นอกจากนี้ เมื่อสังเกตเงื่อนไขที่เมื่ออิทธิพลของเทสต์เลทมีค่าเป็น 0.25, 0.5, 1 และ 2 ตามลำดับ ค่าเฉลี่ยดัชนี RMSE ของพารามิเตอร์อำนาจจำแนกเมื่อประมาณค่าด้วย 3 วิธี มีลักษณะคล้ายกันกับการประมาณค่าพารามิเตอร์ความยาก นั่นคือ วิธีแมกซิมัมไลค์ลิฮู้ด (ML) และวิธีของเบส์ (Bayes) ซึ่งเป็นวิธีที่ไม่มีอิทธิพลเทสต์เลทอยู่ในโมเดลการวิเคราะห์มีค่าเฉลี่ยดัชนี RMSE สูงกว่าวิธีการประมาณค่าด้วยวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) อย่างเห็นได้ชัด อย่างไรก็ตาม สำหรับเงื่อนไขอื่น ๆ ทั้ง 3 วิธีมีค่าเฉลี่ยดัชนี RMSE ไม่แตกต่างกันมากนัก

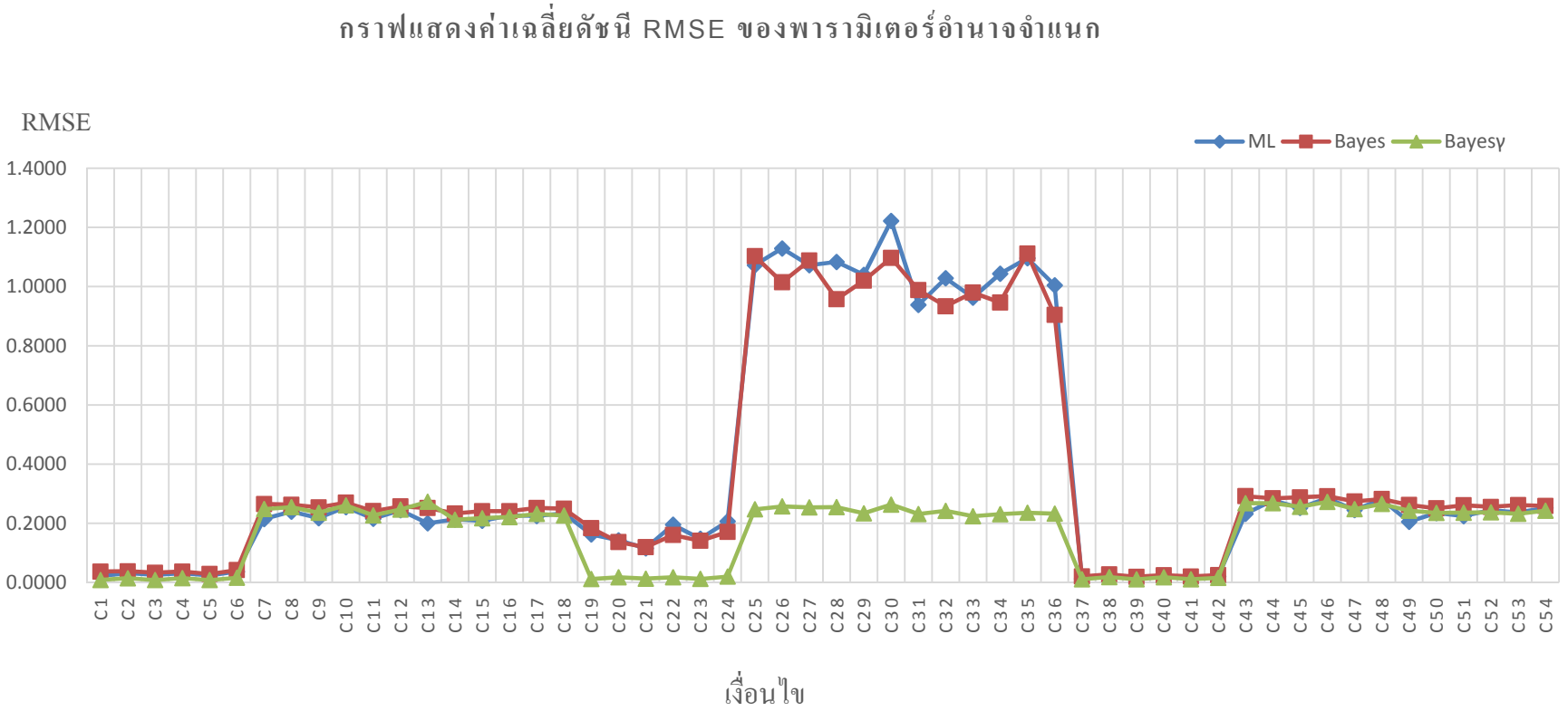
ตารางที่ 4 - 4 ผลการเปรียบเทียบความแปรปวน 5 ทาง (5 - Way ANOVA) ของค่าเฉลี่ยดัชนี RMSE ของพารามิเตอร์อำนาจจำแนก

แหล่งความแปรปรวน	SS	df	MS	F	Sig.
TestEff	5.1514	2	2.5757	5721.4384**	.0000
Dist	4.9753	2	2.4876	5525.7661**	.0000
NumDIF	0.0018	2	0.0009	2.0205	.1651
RperF	0.0019	1	0.0019	4.2388	.0562
Med	1.3967	2	0.6984	1551.2551**	.0000
TestEff * Dist	1.5303	4	0.3826	849.8012**	.0000
TestEff * NumDIF	0.0040	4	0.0010	2.2297	.1116
TestEff * RperF	0.0002	2	0.0001	0.2339	.7941
TestEff * Med	2.5972	4	0.6493	1442.2881**	.0000
Dist * NumDIF	0.0019	4	0.0005	1.0318	.4211
Dist * RperF	0.0044	2	0.0022	4.8859*	.0221
Dist * Med	0.3599	4	0.0900	199.8751**	.0000
NumDIF * RperF	0.0008	2	0.0004	0.9231	.4174
NumDIF * Med	0.0033	4	0.0008	1.8411	.1703
RperF * Med	0.0132	2	0.0066	14.6230**	.0002
TestEff * Dist * NumDIF	0.0013	8	0.0002	0.3606	.9266
TestEff * Dist * RperF	0.0020	4	0.0005	1.0932	.3932
TestEff * Dist * Med	0.7837	8	0.0980	217.6092**	.0000
TestEff * NumDIF * RperF	0.0003	4	0.0001	0.1455	.9624
TestEff * NumDIF * Med	0.0018	8	0.0002	0.4943	.8429
TestEff * RperF * Med	0.0094	4	0.0024	5.2322**	.0069
Dist * NumDIF * RperF	0.0099	4	0.0025	5.4935**	.0056
Dist * NumDIF * Med	0.0014	8	0.0002	0.3754	.9185
Dist * RperF * Med	0.0037	4	0.0009	2.0448	.1363
NumDIF * RperF * Med	0.0004	4	0.0001	0.2026	.9332
TestEff * Dist * NumDIF * RperF	0.0222	8	0.0028	6.1699**	.0010
TestEff * Dist * NumDIF * Med	0.0008	16	0.0001	0.1138	1.0000
TestEff * Dist * RperF * Med	0.0039	8	0.0005	1.0741	.4272
TestEff * NumDIF * RperF * Med	0.0006	8	0.0001	0.1618	.9932
Dist * NumDIF * RperF * Med	0.0053	8	0.0007	1.4716	.2427
Error	0.0072	16	0.0005		

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

จากตารางที่ 4 - 4 ผลการเปรียบเทียบความแปรปวน 5 ทาง (5 - Way ANOVA) ของค่าเฉลี่ยดัชนี RMSE ของพารามิเตอร์อำนาจจำแนก พบว่า อิทธิพลหลัก (Main effect) ที่มีนัยสำคัญทางสถิติ ได้แก่ อิทธิพลของเทสต์เลท (TestEff) การแจกแจงความสามารถ (Dist) และวิธีที่ใช้ประมาณค่าที่ต่างกัน (Med) มีผลต่อ RMSE อย่างมีนัยสำคัญทางสถิติที่ระดับ .01

เมื่อพิจารณาอิทธิพลร่วม (Interection effect) พบว่า อิทธิพลของเทสต์เลทกับการแจกแจงความสามารถ (TestEff * Dist) อิทธิพลของเทสต์เลทกับวิธีที่ใช้ประมาณค่า (TestEff * Med) การแจกแจงความสามารถกับวิธีที่ใช้ในการประมาณค่า (Dist * Med) อัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบกับวิธีที่ใช้ประมาณค่า (RperF * Med) อิทธิพลของเทสต์เลทกับการแจกแจงความสามารถและวิธีที่ใช้ประมาณค่า (TestEff * Dist * Med) อิทธิพลของเทสต์เลทกับอัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบกับวิธีที่ใช้ประมาณค่า (TestEff * RperF * Med) การแจกแจงความสามารถกับจำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบและอัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ (Dist * NumDIF * RperF) อิทธิพลของเทสต์เลทกับการแจกแจงความสามารถกับจำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบและอัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ (TestEff * Dist * NumDIF * RperF) มีผลต่อ RMSE อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ส่วนการแจกแจงความสามารถและอัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ (Dist * RperF) มีผลต่อ RMSE อย่างมีนัยสำคัญทางสถิติที่ระดับ .05



ภาพที่ 4 - 2 ค่าเฉลี่ยดัชนี RMSE ของพารามิเตอร์อำนาจจำแนก จำแนกตามเงื่อนไขและวิธีการประมาณค่าพารามิเตอร์

ตารางที่ 4 - 5 ค่าเฉลี่ยดัชนี RMSE ของพารามิเตอร์ความสามารถ

อิทธิพลเทสต์เลท	การแจกแจงความสามารถ	จำนวนข้อ DIF	จำนวนกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ	เงื่อนไข	ML	Bayes	$Bayes_{\gamma}$
อิทธิพลเท่ากันทุกเทสต์เลท (0.8, 0.8, 0.8, 0.8)	ปกติ	0	1000: 1000	C1	0.3194	0.2820	**0.2408**
			1000: 100	C2	0.2812	0.2490	**0.2412**
		5	1000: 1000	C3	0.2906	0.2431	**0.2349**
			1000: 100	C4	0.2830	0.2532	**0.2452**
		8	1000: 1000	C5	0.2883	0.2443	**0.2364**
			1000: 100	C6	0.2743	0.2476	**0.2386**
	เบ้ซ้าย	0	1000: 1000	C7	0.6373	0.4107	**0.3185**
			1000: 100	C8	0.6537	0.4185	**0.3218**
		5	1000: 1000	C9	0.6371	0.4103	**0.3192**
			1000: 100	C10	0.6407	0.4147	**0.3230**
		8	1000: 1000	C11	0.6339	0.4051	**0.3150**
			1000: 100	C12	0.6372	0.4100	**0.3169**
	เบ้ขวา	0	1000: 1000	C13	0.6146	**0.4021**	0.4488
			1000: 100	C14	0.5986	0.4007	**0.3147**
		5	1000: 1000	C15	0.6165	0.4071	**0.3170**
			1000: 100	C16	0.6013	0.4051	**0.3138**
		8	1000: 1000	C17	0.6212	0.4054	**0.3137**
			1000: 100	C18	0.5908	0.4034	**0.3172**
อิทธิพลเทสต์เลทไม่เท่ากัน (0.25, 0.5, 1, 2)	ปกติ	0	1000: 1000	C19	0.4689	0.3977	**0.2075**
			1000: 100	C20	0.4226	0.3635	**0.2100**
		5	1000: 1000	C21	0.4274	0.3647	**0.2037**
			1000: 100	C22	0.4681	0.3947	**0.2066**
		8	1000: 1000	C23	0.4465	0.3796	**0.2026**
			1000: 100	C24	0.4495	0.3847	**0.2030**
	เบ้ซ้าย	0	1000: 1000	C25	1.3387	0.8024	**0.3031**
			1000: 100	C26	1.4217	0.7889	**0.3078**

ตารางที่ 4 - 5 (ต่อ)

อิทธิพลเทสต์เลท	การแจกแจงความสามารถ	จำนวนข้อ DIF	จำนวนกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ	เงื่อนไข	ML	Bayes	$Bayes_{\gamma}$
อิทธิพลเทสต์เลทไม่เท่ากัน (0.25, 0.5, 1, 2)		5	1000: 1000	C27	1.3090	0.7904	**0.3022**
			1000: 100	C28	1.4421	0.7725	**0.3021**
		8	1000: 1000	C29	1.2461	0.7765	**0.3002**
			1000: 100	C30	1.5201	0.8012	**0.3013**
	เบ้ขวา	0	1000: 1000	C31	1.1767	0.7750	**0.2984**
			1000: 100	C32	1.2748	0.7634	**0.2981**
		5	1000: 1000	C33	1.1811	0.7669	**0.2961**
			1000: 100	C34	1.2790	0.7594	**0.2982**
		8	1000: 1000	C35	1.2323	0.7806	**0.2964**
			1000: 100	C36	1.2325	0.7507	**0.2978**
ข้อสอบที่เป็นอิสระและเทสต์เลท (0, 0.25, 0.56, 1)	ปกติ	0	1000: 1000	C37	0.2384	0.1851	**0.1544**
			1000: 100	C38	0.2147	0.1892	**0.1610**
		5	1000: 1000	C39	0.2305	0.1863	**0.1572**
			1000: 100	C40	0.2109	0.1911	**0.1624**
		8	1000: 1000	C41	0.2236	0.1842	**0.1561**
			1000: 100	C42	0.2038	0.1859	**0.1557**
	เบ้ซ้าย	0	1000: 1000	C43	0.6012	0.3754	**0.2795**
			1000: 100	C44	0.6156	0.3795	**0.2857**
		5	1000: 1000	C45	0.6053	0.3781	**0.2766**
			1000: 100	C46	0.5836	0.3651	**0.2762**
		8	1000: 1000	C47	0.5818	0.3607	**0.2770**
			1000: 100	C48	0.5850	0.3629	**0.2766**
	เบ้ขวา	0	1000: 1000	C49	0.5728	0.3637	**0.2751**
			1000: 100	C50	0.5517	0.3658	**0.2736**
		5	1000: 1000	C51	0.5693	0.3601	**0.2745**
			1000: 100	C52	0.5371	0.3558	**0.2733**
		8	1000: 1000	C53	0.5713	0.3624	**0.2726**
			1000: 100	C54	0.5446	0.3555	**0.2727**

หมายเหตุ ตัวหนา คือ วิธีที่มีค่าเฉลี่ยดัชนี RMSE น้อยที่สุดในเงื่อนไขเดียวกัน

จากตารางที่ 4 - 5 ค่าเฉลี่ยดัชนี RMSE ของพารามิเตอร์ความสามารถ โดยรวมมีค่าเฉลี่ยดัชนี RMSE ระหว่าง 0.1544 - 1.5201 หากพิจารณาตามวิธีที่ใช้ประมาณค่า พบว่า วิธีแมกซิมัมไลค์ลิฮู้ด (ML) มีค่าเฉลี่ยดัชนี RMSE ระหว่าง 0.2038 - 1.5201 โดยไม่มีเงื่อนไขใดที่ประมาณค่าความสามารถได้ดีที่สุด ส่วนวิธีของเบส์ มีค่าเฉลี่ยดัชนี RMSE ระหว่าง 0.1842 - 0.8024 สามารถประมาณค่าพารามิเตอร์ความสามารถได้ดีที่สุด จำนวน 1 เงื่อนไข และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท มีค่าเฉลี่ยดัชนี RMSE ระหว่าง 0.1544 - 0.4488 สามารถประมาณค่าพารามิเตอร์ความสามารถได้ดีที่สุด จำนวน 53 เงื่อนไข โดยผู้วิจัยสรุปผลการศึกษาค่าเฉลี่ยดัชนี RMSE ของพารามิเตอร์อำนาจจำแนก

โดยสรุปส่วนใหญ่การประมาณค่าด้วยวิธีของเบส์แบบคำนึงถึงอิทธิพลของเทสต์เลท (Bayes$_{\gamma}$) สามารถประมาณค่าพารามิเตอร์ความสามารถได้ดีที่สุด ยกเว้นเงื่อนไขข้อมูลที่อิทธิพลของเทสต์เลทเป็น 0.8, 0.8, 0.8, 0.8 ความสามารถมีการแจกแจงแบบเบ้ขวา ไม่มีข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบ และมีอัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบเป็น 1000: 1000 (เงื่อนไข C13) วิธีของเบส์แบบไม่คำนึงถึงอิทธิพลของเทสต์เลท (Bayes) จะประมาณค่าค่าพารามิเตอร์ความสามารถได้ดีกว่า

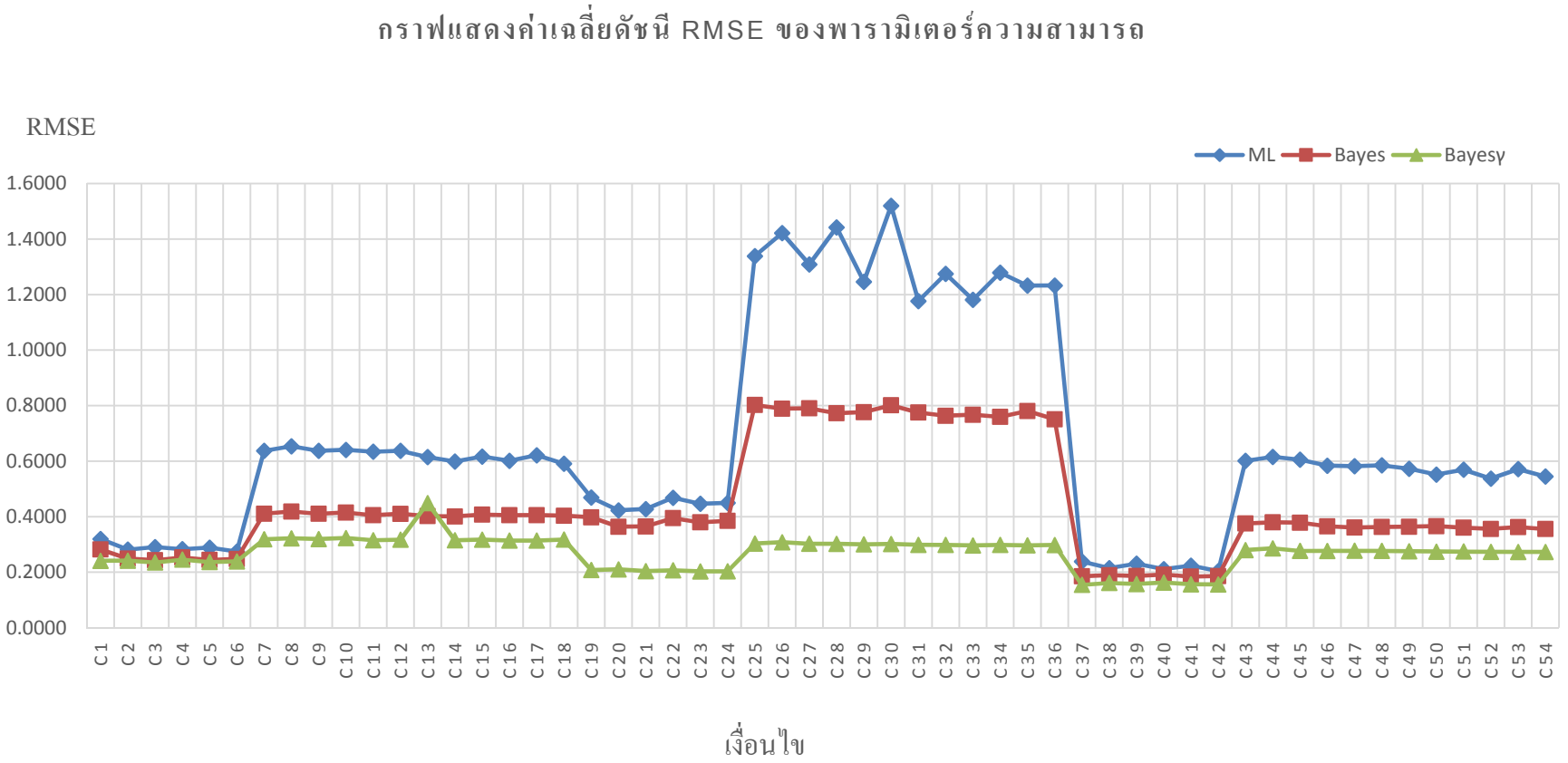
ตารางที่ 4 - 6 ผลการเปรียบเทียบความแปรปวน 5 ทาง (5 - Way ANOVA) ของค่าเฉลี่ยดัชนี RMSE ของพารามิเตอร์ความสามารถ

แหล่งความแปรปรวน	SS	df	MS	F	Sig.
TestEff	2.9645	2	1.4823	4454.9366**	.0000
Dist	2.9772	2	1.4886	4473.9923**	.0000
NumDIF	0.0019	2	0.0009	2.7944	.0910
RperF	0.0005	1	0.0005	1.5335	.2335
Med	4.2428	2	2.1214	6375.9313**	.0000
TestEff * Dist	0.4953	4	0.1238	372.1910**	.0000
TestEff * NumDIF	0.0008	4	0.0002	0.6324	.6466
TestEff * RperF	0.0084	2	0.0042	12.6134**	.0005
TestEff * Med	1.7212	4	0.4303	1293.3023**	.0000
Dist * NumDIF	0.0001	4	0.0000	0.0845	.9860
Dist * RperF	0.0053	2	0.0027	8.0021**	.0039
Dist * Med	1.0981	4	0.2745	825.0917**	.0000
NumDIF * RperF	0.0013	2	0.0007	1.9664	.1724
NumDIF * Med	0.0001	4	0.0000	0.0937	.9831
RperF * Med	0.0033	2	0.0016	4.9018*	.0219
TestEff * Dist * NumDIF	0.0008	8	0.0001	0.2929	.9582
TestEff * Dist * RperF	0.0035	4	0.0009	2.6319	.0731
TestEff * Dist * Med	0.3772	8	0.0472	141.7275**	.0000
TestEff * NumDIF * RperF	0.0016	4	0.0004	1.1784	.3575
TestEff * NumDIF * Med	0.0010	8	0.0001	0.3582	.9278
TestEff * RperF * Med	0.0164	4	0.0041	12.3472**	.0001
Dist * NumDIF * RperF	0.0025	4	0.0006	1.8438	.1698
Dist * NumDIF * Med	0.0013	8	0.0002	0.4824	.8512
Dist * RperF * Med	0.0076	4	0.0019	5.7411**	.0046
NumDIF * RperF * Med	0.0001	4	0.0000	0.0419	.9963
TestEff * Dist * NumDIF * RperF	0.0058	8	0.0007	2.1605	.0903
TestEff * Dist * NumDIF * Med	0.0027	16	0.0002	0.4989	.9124
TestEff * Dist * RperF * Med	0.0081	8	0.0010	3.0302*	.0282
TestEff * NumDIF * RperF * Med	0.0022	8	0.0003	0.8420	.5805
Dist * NumDIF * RperF * Med	0.0054	8	0.0007	2.0443	.1064
Error	0.0053	16	0.0003		

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

จากตารางที่ 4 - 6 ผลการเปรียบเทียบความแปรปวน 5 ทาง (5 - Way ANOVA) ของค่าเฉลี่ยดัชนี RMSE ของพารามิเตอร์ความสามารถ พบว่า อิทธิพลหลัก (Main effect) ที่มีนัยสำคัญทางสถิติ ได้แก่ อิทธิพลของเทสต์เลท (TestEff) การแจกแจงความสามารถ (Dist) และวิธีที่ใช้ประมาณค่า (Med) มีผลต่อ RMSE อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ส่วนจำนวนข้อที่ทำหน้าที่ต่างกันในแบบสอบที่ต่างกัน และอัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบที่ต่างกัน มีค่าเฉลี่ยดัชนี RMSE ของพารามิเตอร์ความสามารถไม่แตกต่างกัน

เมื่อพิจารณาอิทธิพลร่วม (Interection effect) พบว่า อัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบกับวิธีที่ใช้ประมาณค่า (RperF * Med) อิทธิพลของเทสต์เลทกับการแจกแจงความสามารถและวิธีที่ใช้ประมาณค่า (TestEff * Dist * Med) อิทธิพลของเทสต์เลทกับอัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบร่วมกับวิธีที่ใช้ประมาณค่า (TestEff * RperF * Med) การแจกแจงความสามารถกับอัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบและวิธีที่ใช้ประมาณค่า (Dist * RperF * Med) มีผลต่อ RMSE อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ส่วนอิทธิพลของเทสต์เลทกับการแจกแจงความสามารถกับอัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบและวิธีที่ใช้ประมาณค่า (TestEff * Dist * RperF * Med) มีผลต่อ RMSE อย่างมีนัยสำคัญทางสถิติที่ระดับ .05



ภาพที่ 4 - 3 ค่าเฉลี่ยดัชนี RMSE ของพารามิเตอร์ความสามารถ จำแนกตามเงื่อนไขและวิธีการประมาณค่าพารามิเตอร์

ตอนที่ 2 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยพิจารณาจากผลการวิเคราะห์ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error rate) และอำนาจการทดสอบ (Power rate) ของตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ จำแนกตามเงื่อนไขและวิธีการประมาณค่า ดังตารางที่ 4 - 7 ถึง ตารางที่ 4 - 10

ตารางที่ 4 - 7 ผลการวิเคราะห์อัตราความคลาดเคลื่อนประเภทที่ 1 (Type I error rate) ของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

อิทธิพลเทสต์เลท	การแจกแจงความสามารถ	จำนวนข้อ DIF	จำนวนกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ	เงื่อนไข	ML	Bayes	Bayesγ
อิทธิพลเท่ากันทุกเทสต์เลท (0.8, 0.8, 0.8, 0.8)	ปกติ	0	1000: 1000	C1	.3813	.0000*	.0000*
			1000: 100	C2	.5490	.0005*	.0010*
		5	1000: 1000	C3	.3606	.0000*	.0000*
			1000: 100	C4	.5517	.0040*	.0074*
		8	1000: 1000	C5	.4613	.0000*	.0000*
			1000: 100	C6	.5625	.0069*	.0088*
	เบ้ซ้าย	0	1000: 1000	C7	.3763	.0000*	.0000*
			1000: 100	C8	.6175	.0080*	.0125*
		5	1000: 1000	C9	.4149	.0040*	.0040*
			1000: 100	C10	.6000	.0260*	.0291*
		8	1000: 1000	C11	.4275	.0156*	.0181*
			1000: 100	C12	.5338	.0344*	.0319*
	เบ้ขวา	0	1000: 1000	C13	.3673	.0000*	.0008*
			1000: 100	C14	.5583	.0073*	.0093*
		5	1000: 1000	C15	.4214	.0009*	.0023*
			1000: 100	C16	.6137	.0100*	.0231*
		8	1000: 1000	C17	.4675	.0050*	.0084*
			1000: 100	C18	.5969	.0269*	.0244*

ตารางที่ 4 - 7 (ต่อ)

อิทธิพลเทสต์เลท	การแจกแจงความสามารถ	จำนวนข้อ DIF	จำนวนกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ	เงื่อนไข	ML	Bayes	Bayesγ
อิทธิพลเทสต์เลทไม่เท่ากัน (0.25, 0.5, 1, 2)	ปกติ	0	1000: 1000	C19	.3805	.0000*	.0000*
			1000: 100	C20	.5533	.0108*	.0020*
		5	1000: 1000	C21	.3863	.0000*	.0000*
			1000: 100	C22	.5403	.0103*	.0057*
		8	1000: 1000	C23	.4281	.0019*	.0000*
			1000: 100	C24	.5384	.0181*	.0088*
	เบ้ซ้าย	0	1000: 1000	C25	.4880	.4833	.0000*
			1000: 100	C26	.6245	.3969	.0195*
		5	1000: 1000	C27	.4406	.5988	.0057*
			1000: 100	C28	.6140	.3613	.0494*
		8	1000: 1000	C29	.4634	.5579	.0209*
			1000: 100	C30	.5738	.3879	.0603*
	เบ้ขวา	0	1000: 1000	C31	.4110	.4167	.0005*
			1000: 100	C32	.6073	.2904	.0048*
		5	1000: 1000	C33	.4363	.5304	.0069*
			1000: 100	C34	.5757	.2712	.0417*
		8	1000: 1000	C35	.4569	.5726	.0191*
			1000: 100	C36	.6319	.2328	.0444*
ข้อสอบที่เป็นอิสระและเทสต์เลท (0, 0.25, 0.56, 1)	ปกติ	0	1000: 1000	C37	.4023	.0000*	.0000*
			1000: 100	C38	.5718	.0023*	.0023*
		5	1000: 1000	C39	.4071	.0000*	.0000*
			1000: 100	C40	.5637	.0029*	.0020*
		8	1000: 1000	C41	.4809	.0009*	.0009*
			1000: 100	C42	.5481	.0084*	.0091*
	เบ้ซ้าย	0	1000: 1000	C43	.4348	.0090*	.0000*
			1000: 100	C44	.6445	.0328*	.0090*

ตารางที่ 4 - 7 (ต่อ)

อิทธิพลเทสต์เลท	การแจกแจงความสามารถ	จำนวนข้อ DIF	จำนวนกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ	เงื่อนไข	ML	Bayes	Bayes$_{\gamma}$
ข้อสอบที่เป็นอิสระและเทสต์เลท (0, 0.25, 0.56, 1)	เบ้ขวา	0	1000: 1000	C45	.4717	.0254*	.0083*
			1000: 100	C46	.6437	.0394*	.0329*
		5	1000: 1000	C47	.4969	.0400*	.0181*
			1000: 100	C48	.6194	.0516*	.0419*
		8	1000: 1000	C49	.4660	.0015*	.0000*
			1000: 100	C50	.6725	.0175*	.0045*
		0	1000: 1000	C51	.4629	.0123*	.0023*
			1000: 100	C52	.5977	.0297*	.0223*
		5	1000: 1000	C53	.4872	.0388*	.0200*
			1000: 100	C54	.6416	.0413*	.0356*
		8	1000: 1000	C55	.4717	.0254*	.0083*
			1000: 100	C56	.6437	.0394*	.0329*
			1000: 1000	C57	.4969	.0400*	.0181*
			1000: 100	C58	.6194	.0516*	.0419*

* $p < .05$

จากตารางที่ 4 - 7 ผลการวิเคราะห์อัตราความคลาดเคลื่อนประเภทที่ 1 (Type I error rate) ของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยรวมมีค่าระหว่าง 0 - 0.6725 หรือคิดเป็นร้อยละ 0 - 67.25 หากพิจารณาตามวิธีที่ใช้ใน การประมาณค่า พบว่า วิธีแมกซิมัมไลค์ลิฮู้ด มีค่าระหว่าง 0.3606 - 0.6725 หรือคิดเป็นร้อยละ 36.06 - 67.25 วิธีของเบส์ มีค่าระหว่าง 0 - 0.5980 หรือคิดเป็นร้อยละ 0 - 59.80 และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท มีค่าระหว่าง 0 - 0.0603 หรือร้อยละ 0 - 6.03 โดยสรุปผลการศึกษาความคลาดเคลื่อนประเภทที่ 1 ส่วนใหญ่ วิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayes$_{\gamma}$) สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดีที่สุด ส่วนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบสอบด้วยวิธีแมกซิมัมไลค์ลิฮู้ด (ML) ไม่ผ่านตามเกณฑ์ที่กำหนดทุกเงื่อนไข

ตารางที่ 4 - 8 ผลการเปรียบเทียบความแปรปวน 5 ทาง (5 - Way ANOVA) ของความคลาดเคลื่อนประเภทที่ 1

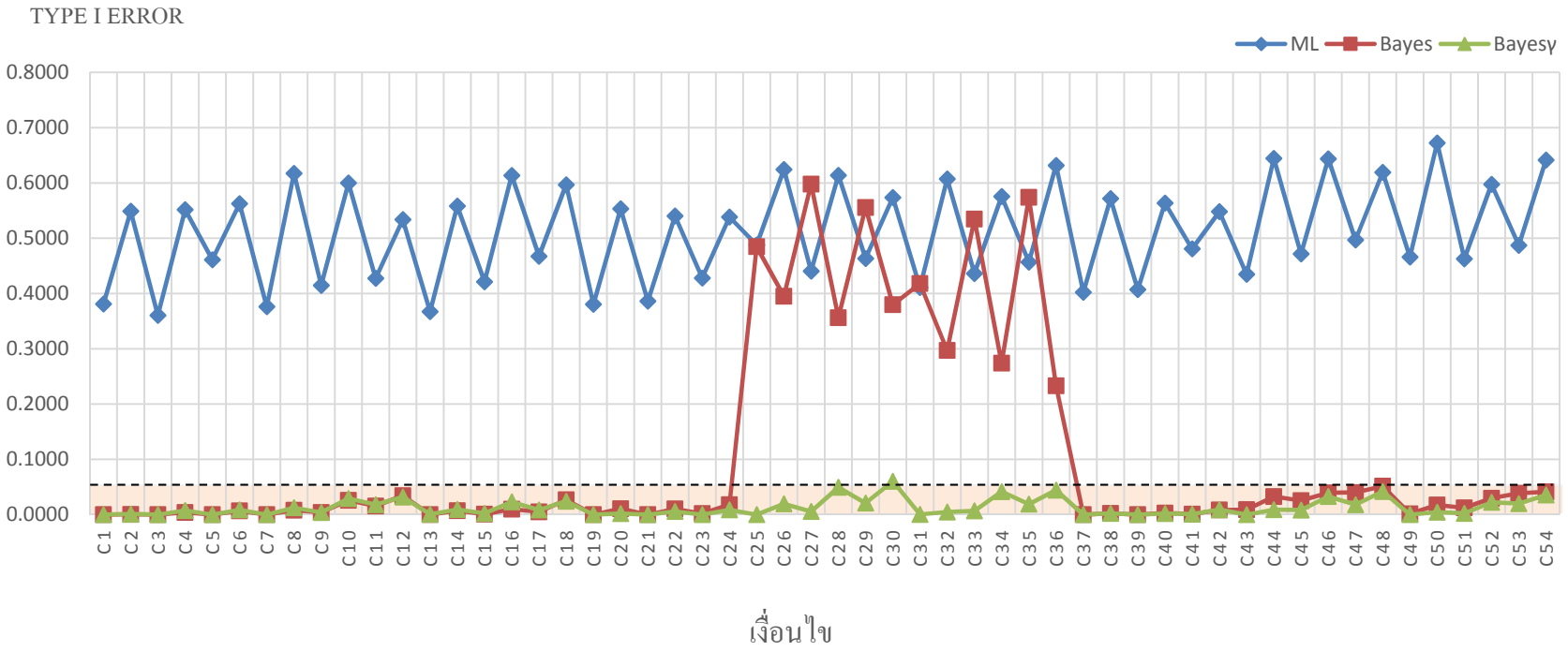
แหล่งความแปรปรวน	SS	df	MS	F	Sig.
TestEff	0.0062	2	0.0031	34.7001**	.0000
Dist	0.0228	2	0.0114	128.1342**	.0000
NumDIF	0.0080	2	0.0040	44.8280**	.0000
RperF	0.1594	1	0.1594	1793.7614**	.0000
Med	8.9631	2	4.4816	50436.0877**	.0000
TestEff * Dist	0.0025	4	0.0006	6.9080**	.0020
TestEff * NumDIF	0.0003	4	0.0001	0.9018	.4860
TestEff * RperF	0.0002	2	0.0001	1.0156	.3844
TestEff * Med	0.0105	4	0.0026	29.5366**	.0000
Dist * NumDIF	0.0017	4	0.0004	4.6718*	.0109
Dist * RperF	0.0027	2	0.0014	15.3134**	.0002
Dist * Med	0.0074	4	0.0019	20.8379**	.0000
NumDIF * RperF	0.0025	2	0.0012	13.9754**	.0003
NumDIF * Med	0.0008	4	0.0002	2.3407	.0992
RperF * Med	0.1839	2	0.0919	1034.6477**	.0000
TestEff * Dist * NumDIF	0.0020	8	0.0003	2.8551*	.0353
TestEff * Dist * RperF	0.0001	4	0.0000	0.3040	.8711
TestEff * Dist * Med	0.0020	8	0.0003	2.8612*	.0350
TestEff * NumDIF * RperF	0.0010	4	0.0002	2.6972	.0684
TestEff * NumDIF * Med	0.0020	8	0.0003	2.8749*	.0344
TestEff * RperF * Med	0.0008	4	0.0002	2.2298	.1116
Dist * NumDIF * RperF	0.0006	4	0.0001	1.6513	.2102
Dist * NumDIF * Med	0.0052	8	0.0006	7.2584**	.0004
Dist * RperF * Med	0.0004	4	0.0001	1.0743	.4016
NumDIF * RperF * Med	0.0094	4	0.0023	26.4336**	.0000
TestEff * Dist * NumDIF * RperF	0.0009	8	0.0001	1.2489	.3343
TestEff * Dist * NumDIF * Med	0.0043	16	0.0003	3.0020*	.0172
TestEff * Dist * RperF * Med	0.0011	8	0.0001	1.5514	.2162
TestEff * NumDIF * RperF * Med	0.0010	8	0.0001	1.3587	.2857
Dist * NumDIF * RperF * Med	0.0020	8	0.0002	2.7634	.0398
Error	0.0014	16	0.0001		

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

จากตารางที่ 4 - 8 ผลการเปรียบเทียบความแปรปวน 5 ทาง (5 - Way ANOVA) ของความคลาดเคลื่อนประเภทที่ 1 พบว่า อิทธิพลหลัก (Main effect) ที่มีนัยสำคัญทางสถิติ ได้แก่ อิทธิพลของเทสต์เลท (TestEff) การแจกแจงความสามารถ (Dist) จำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบ (NumDIF) อัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ (RperF) และวิธีที่ใช้ประมาณค่า (Med) มีผลต่อความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01

เมื่อพิจารณาปฏิสัมพันธ์ของอิทธิพลของเทสต์เลทกับการแจกแจงความสามารถ (TestEff * Dist) อิทธิพลของเทสต์เลทกับวิธีที่ใช้ประมาณค่า (TestEff * Med) การแจกแจงความสามารถกับอัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ (Dist * RperF) การแจกแจงความสามารถกับวิธีที่ใช้ประมาณค่า (Dist * Med) จำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบกับอัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ (NumDIF * RperF) อัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบกับวิธีที่ใช้ในการประมาณค่า (RperF * Med) การแจกแจงความสามารถกับจำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบและวิธีที่ใช้ประมาณค่า (Dist * NumDIF * Med) จำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบกับอัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบและวิธีที่ใช้ประมาณค่า (NumDIF * RperF * Med) อิทธิพลของเทสต์เลทกับการแจกแจงความสามารถกับจำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบและวิธีที่ใช้ประมาณค่า (TestEff * Dist * NumDIF * Med) มีความคลาดเคลื่อนประเภทที่ 1 ต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01

ส่วนการแจกแจงความสามารถกับจำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบ (Dist * NumDIF) อิทธิพลของเทสต์เลทกับการแจกแจงความสามารถและจำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบ (TestEff * Dist * NumDIF) อิทธิพลของเทสต์เลทกับการแจกแจงความสามารถและวิธีที่ใช้ประมาณค่า (TestEff * Dist * Med) อิทธิพลของเทสต์เลทกับจำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบและวิธีที่ใช้ประมาณค่า (TestEff * NumDIF * Med) อิทธิพลของเทสต์เลทกับการแจกแจงความสามารถกับจำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบและวิธีที่ใช้ประมาณค่า (TestEff * Dist * NumDIF * Med) มีความคลาดเคลื่อนประเภทที่ 1 ต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05



ภาพที่ 4 - 4 ผลการวิเคราะห์อัตราความคลาดเคลื่อนประเภทที่ 1 (Type I error rate) ของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

ตารางที่ 4 - 9 ผลการวิเคราะห์ค่าอำนาจการทดสอบ (Power Rate) ของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

อิทธิพลเทสต์เลท	การแจกแจงความสามารถ	จำนวนข้อ DIF	จำนวนกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ	เงื่อนไข	ML	Bayes	Bayesγ
อิทธิพลเท่ากันทุกเทสต์เลท (0.8, 0.8, 0.8, 0.8)	ปกติ	0	1000: 1000	C1	NA	NA	NA
			1000: 100	C2	NA	NA	NA
		5	1000: 1000	C3	.9420*	.1000	.1240
			1000: 100	C4	.7900	.0980	.0900
		8	1000: 1000	C5	.9425*	.1038	.0763
			1000: 100	C6	.7863	.1138	.0450
	เบ้ซ้าย	0	1000: 1000	C7	NA	NA	NA
			1000: 100	C8	NA	NA	NA
		5	1000: 1000	C9	.9800*	.6600	.7180
			1000: 100	C10	.8560*	.3680	.2800
		8	1000: 1000	C11	.9613*	.6000	.5275
			1000: 100	C12	.8113	.2250	.1475
	เบ้ขวา	0	1000: 1000	C13	NA	NA	NA
			1000: 100	C14	NA	NA	NA
		5	1000: 1000	C15	.9720*	.6640	.7000
			1000: 100	C16	.8340*	.2940	.2320
		8	1000: 1000	C17	.9588*	.5450	.4988
			1000: 100	C18	.7800	.2875	.1475
อิทธิพลเทสต์เลทไม่เท่ากัน (0.25, 0.5, 1, 2)	ปกติ	0	1000: 1000	C19	NA	NA	NA
			1000: 100	C20	NA	NA	NA
		5	1000: 1000	C21	.9580*	.1840	.2880
			1000: 100	C22	.7740	.0540	.1120
		8	1000: 1000	C23	.9063*	.0625	.1025
			1000: 100	C24	.7088	.0825	.0975
	เบ้ซ้าย	0	1000: 1000	C25	NA	NA	NA
			1000: 100	C26	NA	NA	NA

ตารางที่ 4 - 9 (ต่อ)

อิทธิพลเทสต์เลท	การแจกแจงความสามารถ	จำนวนข้อ DIF	จำนวนกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ	เงื่อนไข	ML	Bayes	Bayes$_\gamma$
อิทธิพลเทสต์เลทไม่เท่ากัน (0.25, 0.5, 1, 2)	เบ้ซ้าย	5	1000: 1000	C27	.9380*	.8100	.7140
			1000: 100	C28	.8020	.3840	.3960
		8	1000: 1000	C29	.9225*	.7225	.6238
			1000: 100	C30	.7888	.3238	.2413
	เบ้ขวา	0	1000: 1000	C31	NA	NA	NA
			1000: 100	C32	NA	NA	NA
		5	1000: 1000	C33	.9720*	.8420*	.7800
			1000: 100	C34	.8280*	.3240	.3520
		8	1000: 1000	C35	.9113*	.7175	.6163
			1000: 100	C36	.6975	.2625	.2413
ข้อสอบที่เป็นอิสระและเทสต์เลท (0, 0.25, 0.56, 1)	ปกติ	0	1000: 1000	C37	NA	NA	NA
			1000: 100	C38	NA	NA	NA
		5	1000: 1000	C39	.9740*	.2660	.3740
			1000: 100	C40	.7500	.1000	.1260
		8	1000: 1000	C41	.9375*	.1150	.1550
			1000: 100	C42	.7388	.0975	.1000
	เบ้ซ้าย	0	1000: 1000	C43	NA	NA	NA
			1000: 100	C44	NA	NA	NA
		5	1000: 1000	C45	.9400*	.7200	.7820
			1000: 100	C46	.8640*	.3320	.3900
		8	1000: 1000	C47	.9663*	.5588	.6863
			1000: 100	C48	.8275*	.2488	.3063
	เบ้ขวา	0	1000: 1000	C49	NA	NA	NA
			1000: 100	C50	NA	NA	NA
		5	1000: 1000	C51	.9520*	.6140	.7480
			1000: 100	C52	.7460	.2200	.2860
		8	1000: 1000	C53	.9475*	.5913	.7025
			1000: 100	C54	.8063	.2638	.2888

หมายเหตุ NA หมายถึง ไม่สามารถหาค่าได้, * $p < .05$

จากตารางที่ 4 - 9 ผลการวิเคราะห์อำนาจ (Power) การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยรวมมีค่าระหว่าง 0.0450 - 0.9800 หรือคิดเป็นร้อยละ 4.50 - 98 หากพิจารณาตามวิธีที่ใช้ในการประมาณค่า พบว่า วิธีแมกซิมัมไลค์ลิฮู้ดมีค่าระหว่าง 0.6975 - 0.9800 หรือคิดเป็นร้อยละ 69.75 - 98 วิธีของเบส์ มีค่าระหว่าง 0.0540 - 0.8420 หรือคิดเป็นร้อยละ 5.40 - 84.20 และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท มีค่าระหว่าง 0.0450 - 0.7820 หรือคิดเป็นร้อยละ 4.50 - 78.20 โดยสรุปส่วนใหญ่วิธีแมกซิมัมไลค์ลิฮู้ด (ML) ค่าอำนาจในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูงผ่านตามเกณฑ์ที่กำหนด และมีอำนาจการทดสอบมากกว่าวิธีของเบส์ (Bayes) และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) ซึ่งทั้งสองวิธีมีอำนาจในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบต่ำกว่าเกณฑ์ที่กำหนด

ตารางที่ 4 - 10 ผลการเปรียบเทียบความแปรปวน 5 ทาง (5 - Way ANOVA) ของอำนาจในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (Power rate)

แหล่งความแปรปรวน	SS	df	MS	F	Sig.
TestEff	.030	2	0.0150	12.7324**	.0033
Dist	1.426	2	0.7130	606.1911**	.0000
NumDIF	.112	1	0.1117	94.9407**	.0000
RperF	1.571	1	1.5715	1335.9903**	.0000
Med	6.165	2	3.0825	2620.5589**	.0000
TestEff * Dist	.014	4	0.0035	2.9441	.0906
TestEff * NumDIF	.006	2	0.0028	2.3973	.1528
TestEff * RperF	.009	2	0.0043	3.6814	.0735
TestEff * Med	.084	4	0.0211	17.9518**	.0005
Dist * NumDIF	.004	2	0.0019	1.6172	.2571
Dist * RperF	.244	2	0.1221	103.7709**	.0000
Dist * Med	.565	4	0.1412	120.0533**	.0000
NumDIF * RperF	.010	1	0.0100	8.5189*	.0193
NumDIF * Med	.028	2	0.0139	11.8586**	.0040
RperF * Med	.092	2	0.0458	38.9565**	.0001
TestEff * Dist * NumDIF	.024	4	0.0060	5.0893*	.0245
TestEff * Dist * RperF	.010	4	0.0025	2.1173	.1702
TestEff * Dist * Med	.023	8	0.0028	2.4002	.1185
TestEff * NumDIF * RperF	.003	2	0.0013	1.0837	.3833
TestEff * NumDIF * Med	.003	4	0.0007	0.6337	.6526
TestEff * RperF * Med	.012	4	0.0029	2.4790	.1278

ตารางที่ 4 - 10 (ต่อ)

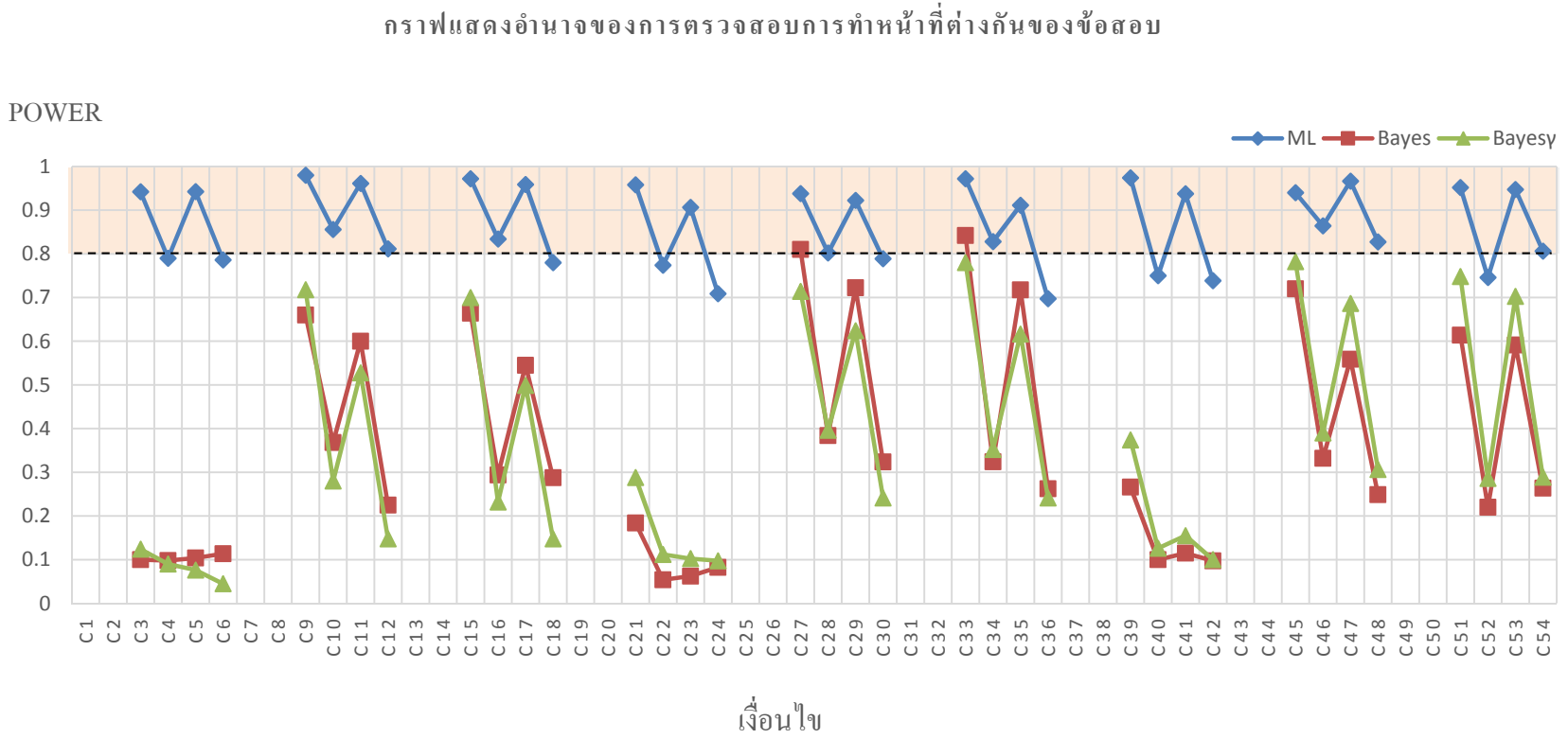
แหล่งความแปรปรวน	SS	df	MS	F	Sig.
Dist * NumDIF * RperF	.008	2	0.0039	3.3441	.0880
Dist * NumDIF * Med	.005	4	0.0013	1.0632	.4340
Dist * RperF * Med	.169	4	0.0422	35.8791**	.0000
NumDIF * RperF * Med	.009	2	0.0046	3.9383	.0645
TestEff * Dist * NumDIF * RperF	.005	4	0.0012	1.0344	.4460
TestEff * Dist * NumDIF * Med	.003	8	0.0004	0.2998	.9459
TestEff * Dist * RperF * Med	.006	8	0.0007	0.6188	.7437
TestEff * NumDIF * RperF * Med	.001	4	0.0003	0.2960	.8726
Dist * NumDIF * RperF * Med	.002	4	0.0004	0.3772	.8190
Error	.009	8	0.0012		

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

จากตารางที่ 4 - 10 ผลการเปรียบเทียบความแปรปวน 5 ทาง (5 - Way ANOVA) ของอำนาจในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (Power rate) พบว่า อิทธิพลหลัก (Main effect) ที่มีนัยสำคัญทางสถิติ ได้แก่ อิทธิพลของเทสต์เลท (TestEff) การแจกแจงความสามารถ (Dist) จำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบ (NumDIF) อัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ (RperF) และวิธีที่ใช้ประมาณค่า (Med) มีผลต่ออำนาจการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อย่างมีนัยสำคัญทางสถิติที่ระดับ .01

เมื่อพิจารณาอิทธิพลร่วม (Interection effect) พบว่า อิทธิพลของเทสต์เลทกับวิธีที่ใช้ประมาณค่า (TestEff * Med) การแจกแจงความสามารถกับอัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ (Dist * RperF) การแจกแจงความสามารถกับวิธีที่ใช้ในการประมาณค่า (Dist * Med) จำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบกับวิธีที่ใช้ประมาณค่า (NumDIF * Med) อัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบกับวิธีที่ใช้ประมาณค่า (RperF * Med) การแจกแจงความสามารถกับอัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบและวิธีที่ใช้ประมาณค่า (Dist * RperF * Med) มีผลต่ออำนาจการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อย่างมีนัยสำคัญทางสถิติที่ระดับ .01

ส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบกับอัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ (NumDIF * RperF) อิทธิพลของเทสต์เลทกับการแจกแจงความสามารถและจำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบ (TestEff * Dist * NumDIF) มีผลต่ออำนาจการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อย่างมีนัยสำคัญทางสถิติที่ระดับ .05



ภาพที่ 4 - 5 ผลการวิเคราะห์อำนาจ (Power Rate) ของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

บทที่ 5
สรุป อภิปรายผล และข้อเสนอแนะ

การวิจัยครั้งนี้มีวัตถุประสงค์ 2 ประการ ประการแรกเพื่อศึกษาผลการประมาณค่าพารามิเตอร์ของข้อสอบ (ความยากและอำนาจจำแนก) กับพารามิเตอร์ความสามารถของผู้สอบด้วยวิธีแมกซิมัมไลค์ลิฮู้ด (ML) วิธีของเบส์ (Bayes) และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayes$_\gamma$) และประการที่สองเพื่อศึกษาผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ระหว่างวิธีแมกซิมัมไลค์ลิฮู้ด (ML) วิธีของเบส์ (Bayes) และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayes$_\gamma$) โดยการศึกษาข้อมูลที่มีอิทธิพลเทสต์เลท การแจกแจงของความสามารถ จำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบและอัตราส่วนของกลุ่มเปรียบเทียบต่อกลุ่มอ้างอิงที่ต่างกัน ภายใต้เงื่อนไข ๆ การจำลองข้อมูล จำนวน 54 เงื่อนไข (3 x 3 x 3 x 2) ในแต่ละเงื่อนไขจำลองข้อมูลซ้ำ 100 ครั้ง จำนวนการทำซ้ำภายใต้เงื่อนไขที่แปรเปลี่ยนทั้งหมด 5,400 ครั้ง จากนั้น ประมาณค่าพารามิเตอร์ด้วยวิธีแมกซิมัมไลค์ลิฮู้ด (ML) วิธีของเบส์ (Bayes) และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayes$_\gamma$)

สรุปผลการวิจัย

จากผลการวิเคราะห์ข้อมูล สามารถสรุปผลการวิจัย จำแนกตามวัตถุประสงค์ โดยแบ่งออกเป็น 2 ตอน คือ ตอนที่ 1 ผลการประมาณค่าพารามิเตอร์ของข้อสอบ (ความยากและอำนาจจำแนก) และพารามิเตอร์ความสามารถของผู้สอบ และตอนที่ 2 ผลการวิเคราะห์ค่าความคลาดเคลื่อนประเภทที่ 1 และอำนาจของตรวจสอบการทำหน้าที่ต่างกันของข้อสอบภายใต้เงื่อนไขการจำลองข้อมูล 4 ปัจจัย ได้แก่ อิทธิพลเทสต์เลท การแจกแจงความสามารถ จำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบ และอัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ และวิธีการประมาณค่าพารามิเตอร์ ซึ่งมีรายละเอียดดังนี้

1. ผลการประมาณค่าพารามิเตอร์ของข้อสอบ (ความยากและอำนาจจำแนก) และพารามิเตอร์ความสามารถของผู้สอบ

1.1 ผลการประมาณค่าพารามิเตอร์ความยาก พบว่า ค่าเฉลี่ยดัชนี RMSE ของพารามิเตอร์ความยาก ซึ่งมีค่าระหว่าง 0.0085 - 2.9171 หากพิจารณาตามวิธีที่ใช้ประมาณค่า พบว่าวิธีแมกซิมัมไลค์ลิฮู้ดมีค่าระหว่าง 0.0952 - 0.9617 วิธีของเบส์แบบไม่คำนึงถึงอิทธิพลของเทสต์

เลท มีค่าระหว่าง 0.0136 - 2.9171 และวิธีของเบส์แบบคำนึงถึงอิทธิพลของเทสต์เลท มีค่าระหว่าง 0.0085 - 1.0238 ส่วนใหญ่วิธีแมกซิมัมไลค์ลิฮู้ด (ML) จะประมาณค่าพารามิเตอร์ความยากได้ดีที่สุด จำนวน 25 เงื่อนไข วิธีที่ประมาณค่าพารามิเตอร์ความยากได้ดีที่สุด รองลงมา คือ วิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) จำนวน 19 เงื่อนไข และวิธีของเบส์ (Bayes) ประมาณค่าค่าพารามิเตอร์ความยากได้ดีที่สุด จำนวน 10 เงื่อนไข ซึ่งเงื่อนไขอิทธิพลเทสต์เลท การแจกแจงความสามารถ อัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบและวิธีที่ใช้ประมาณค่าที่ต่างกัน ทำให้ผลของการประมาณค่าพารามิเตอร์ความยากได้ดีต่างกันด้วย โดยที่วิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) ประมาณค่าได้ดีเมื่อข้อมูลมีการแจกแจงความสามารถเป็นแบบปกติ ส่วนวิธีของเบส์ (Bayes) ยังไม่มีแนวโน้มแน่นอน แต่จะประมาณค่าพารามิเตอร์ความยากได้ดีเป็นส่วนใหญ่เมื่อข้อมูลมีการแจกแจงความสามารถเป็นแบบเบ้ซ้าย และวิธีแมกซิมัมไลค์ลิฮู้ด (ML) จะประมาณค่าพารามิเตอร์ความยากได้ดีเมื่อข้อมูลมีการแจกแจงความสามารถเป็นแบบเบ้ขวา

1.2 ผลการประมาณค่าพารามิเตอร์อำนาจจำแนก พบว่า ค่าเฉลี่ยดัชนี RMSE ของพารามิเตอร์อำนาจจำแนก โดยรวมมีค่าระหว่าง 0.0081 - 1.2219 หากพิจารณาตามวิธีที่ใช้ประมาณค่าพารามิเตอร์ พบว่า วิธีแมกซิมัมไลค์ลิฮู้ด มีค่าระหว่าง 0.0129 - 1.2219 วิธีของเบส์แบบมีอิทธิพลเทสต์เลท มีค่าระหว่าง 0.0194 - 1.1115 และวิธีของเบส์ (Bayes) มีค่าระหว่าง 0.0081 - 0.2729 ส่วนใหญ่วิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) ประมาณค่าได้ดีที่สุด จำนวน 39 เงื่อนไข วิธีที่ประมาณค่าพารามิเตอร์อำนาจจำแนกได้ดีที่สุด รองลงมา คือ วิธีแมกซิมัมไลค์ลิฮู้ด (ML) จำนวน 15 เงื่อนไข และไม่มีเงื่อนไขที่วิธีของเบส์ (Bayes) ประมาณค่าได้ดีที่สุด ซึ่งเงื่อนไขอิทธิพลเทสต์เลท การแจกแจงความสามารถ และวิธีที่ใช้ในการประมาณค่าที่ต่างกันที่ต่างกัน ทำให้ผลการประมาณค่าพารามิเตอร์อำนาจจำแนกได้ดีต่างกันด้วย โดยที่ส่วนใหญ่วิธีการประมาณค่าด้วยวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) จะประมาณค่าพารามิเตอร์อำนาจจำแนกได้ดีกว่าเมื่อข้อมูลความสามารถมีการแจกแจงแบบปกติ หรือเมื่อค่าอิทธิพลของเทสต์เลท ไม่เท่ากัน (Unequal effect) โดยกำหนดให้อิทธิพลของเทสต์เลทมีค่าเป็น 0.25, 0.5, 1 และ 2 ตามลำดับ ส่วนและวิธีแมกซิมัมไลค์ลิฮู้ด (ML) ส่วนใหญ่จะประมาณค่าพารามิเตอร์อำนาจจำแนกได้ดีเมื่อข้อมูลความสามารถมีการแจกแจงแบบเบ้ซ้าย

1.3 ผลการประมาณค่าพารามิเตอร์ความสามารถ พบว่า ค่าเฉลี่ยดัชนี RMSE ของพารามิเตอร์ความสามารถ โดยรวมมีค่าระหว่าง 0.1544 - 1.5201 หากพิจารณาตามวิธีที่ใช้ประมาณค่า พบว่า วิธีแมกซิมัมไลค์ลิฮู้ดมีค่าระหว่าง 0.2038 - 1.5201 วิธีของเบส์แบบไม่คำนึงถึงอิทธิพลของเทสต์เลท มีค่าระหว่าง 0.1842 - 0.8024 และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท มีค่าระหว่าง 0.1544 - 0.4488 ส่วนใหญ่วิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) ประมาณค่าได้ดีที่สุด

จำนวน 53 เงื่อนไข วิธีที่ประมาณค่าพารามิเตอร์อำนาจจำแนกได้ดีที่สุด รองลงมา คือ วิธีของเบส์ (Bayes) จำนวน 1 เงื่อนไข และไม่มีเงื่อนไขใดที่วิธีแมกซิมัมไลค์ลิฮู้ด (ML) ประมาณค่าได้ดีที่สุด ซึ่งเงื่อนไขอิทธิพลเทสต์เลท การแจกแจงความสามารถ และวิธีที่ใช้ในการประมาณค่าที่ต่างกัน ทำให้ผลการประมาณค่าพารามิเตอร์อำนาจจำแนกได้ดีต่างกันด้วย

2. ผลการวิเคราะห์ค่าความคลาดเคลื่อนประเภทที่ 1 และอำนาจของตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

2.1 ผลการวิเคราะห์อัตราความคลาดเคลื่อนประเภทที่ 1 (Type I error rate) ของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ พบว่า โดยรวมมีค่าระหว่าง 0 - 0.6725 หรือคิดเป็นร้อยละ 0 - 67.25 หากพิจารณาตามวิธีที่ใช้ในการประมาณค่า พบว่า วิธีแมกซิมัมไลค์ลิฮู้ด (ML) มีค่าระหว่าง 0.3606 - 0.6725 หรือคิดเป็นร้อยละ 36.06 - 67.25 วิธีของเบส์ (Bayes) มีค่าระหว่าง 0 - 0.5980 หรือคิดเป็นร้อยละ 0 - 59.80 และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) มีค่าระหว่าง 0 - 0.0603 หรือร้อยละ 0 - 6.03 โดยส่วนใหญ่วิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดีที่สุด ส่วนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบสอบด้วยวิธีแมกซิมัมไลค์ลิฮู้ด (ML) ไม่ผ่านตามเกณฑ์ที่กำหนดทุกเงื่อนไข

2.2 ผลการวิเคราะห์ค่าอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (Power rate) พบว่า โดยรวมมีค่าระหว่าง 0.0450 - 0.9800 หรือคิดเป็นร้อยละ 4.50 - 98 หากพิจารณาตามวิธีที่ใช้ประมาณค่า พบว่า วิธีแมกซิมัมไลค์ลิฮู้ดมีค่าระหว่าง 0.6975 - 0.9800 หรือคิดเป็นร้อยละ 69.75 - 98 วิธีของเบส์ มีค่าระหว่าง 0.0540 - 0.8420 หรือคิดเป็นร้อยละ 5.40 - 84.20 และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท มีค่าระหว่าง 0.0450 - 0.7820 หรือคิดเป็นร้อยละ 4.50 - 78.20 โดยสรุป ส่วนใหญ่วิธีแมกซิมัมไลค์ลิฮู้ด (ML) ค่าอำนาจในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูงผ่านตามเกณฑ์ที่กำหนด และมีอำนาจการทดสอบมากกว่าวิธีของเบส์ (Bayes) และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) ซึ่งทั้งสองวิธีมีอำนาจในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบต่ำกว่าเกณฑ์ที่กำหนด

อภิปรายผลการวิจัย

การอภิปรายผลในงานวิจัยนี้นำเสนอ 2 ประเด็นหลักตามวัตถุประสงค์ ได้แก่ ประเด็นแรก ผลการประมาณค่าพารามิเตอร์ข้อสอบ (อำนาจจำแนกและความยาก) พารามิเตอร์ผู้สอบ (ความสามารถ) และประเด็นที่ 2 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยทั้งสองประเด็นทำการประมาณค่าพารามิเตอร์ด้วยวิธีแมกซิมัมไลค์ลิฮู้ด (ML) วิธีของเบส์ (Bayes) และ

วิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) ในเงื่อนไขอิทธิพลเทสต์เลท การแจกแจงของความสามารถ จำนวนข้อสอบที่ทำหน้าที่ต่างกัน ในแบบสอบและอัตราส่วนของกลุ่มเปรียบเทียบต่อกลุ่มอ้างอิงที่ต่างกัน โดยมีรายละเอียด ดังนี้

1. ผลการประมาณค่าพารามิเตอร์ข้อสอบ (ความยากและอำนาจจำแนก) พารามิเตอร์ผู้สอบ (ความสามารถ)

เมื่อพิจารณาวิธีการประมาณค่าพารามิเตอร์ จากผลการเปรียบเทียบความแปรปรวน 5 ทาง (5 - Way ANOVA) ของค่าเฉลี่ยดัชนี RMSE ของพารามิเตอร์ความยาก อำนาจจำแนกและความสามารถ พบว่า วิธีการประมาณค่าพารามิเตอร์ ทำให้ค่าเฉลี่ยดัชนี RMSE ต่างกันอย่างมีนัยสำคัญที่ระดับ .01 ทั้ง 3 พารามิเตอร์ นั่นคือ วิธีการประมาณค่าพารามิเตอร์แต่ละวิธีมีความถูกต้องในการประมาณค่าพารามิเตอร์ต่างกัน โดยเมื่อพิจารณาค่าเฉลี่ยดัชนี RMSE ของพารามิเตอร์ความยากจะเห็นว่า การประมาณค่าด้วยวิธีของเบส์ (Bayes) และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) มีค่าเฉลี่ยดัชนี RMSE ใกล้เคียงกัน ยกเว้นกรณีที่อิทธิพลของเทสต์เลทมีค่าเป็น 0.25, 0.5, 1, 2 ร่วมกับการแจกแจงความสามารถที่เป็นแบบเบ้ซ้ายและเบ้ขวา ซึ่งทำให้วิธีของเบส์ (Bayes) มีค่า RMSE สูงขึ้น ส่วนการประมาณค่าพารามิเตอร์ด้วยวิธีแมกซิมัมไลค์ลิฮู้ด (ML) จะประมาณค่าพารามิเตอร์ความยากได้ดีในช่วงของการแจกแจงความสามารถที่เป็นแบบเบ้ขวา ซึ่งสอดคล้องกับ Vincent & Prathiba (2012) ศึกษาเปรียบเทียบวิธี Marginal Maximum Likelihood และ Markov Chain Monte Carlo ในการประมาณค่าพารามิเตอร์โมเดล Graded Response พบว่า เมื่อการแจกแจงของคุณลักษณะแฝง (Latent trait distribution) เป็นแบบเบ้ (Skew - normal) วิธีที่ประมาณค่าได้ดีกว่า คือ วิธี Marginal Maximum Likelihood อย่างไรก็ตาม เมื่อพิจารณาจากจำนวนตัวอย่างที่แปรเปลี่ยน ไม่มีผลต่อการประมาณค่าด้วยวิธีของเบส์ (Bayes) และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) ซึ่งต่างจากการประมาณค่าพารามิเตอร์ด้วยวิธีแมกซิมัมไลค์ลิฮู้ด (ML) พบว่า เมื่อจำนวนตัวอย่างเพิ่มขึ้นจะทำให้การประมาณค่าพารามิเตอร์ได้ดีกว่าจำนวนตัวอย่างขนาดเล็กกว่า สอดคล้องกับผลการศึกษาของ Wang & Wilson (2005 b) ที่ศึกษา Rasch Testlet Model สำหรับข้อมูลที่มีวิธีการให้คะแนนรายข้อแบบสองค่าและหลายค่า และประมาณค่าพารามิเตอร์ด้วยวิธีแมกซิมัมไลค์ลิฮู้ด พบว่า เมื่อจำนวนตัวอย่างมากขึ้นค่า RMSE ของการประมาณค่าจะลดลงในระดับที่ยอมรับได้

เมื่อพิจารณาที่การประมาณค่าพารามิเตอร์อำนาจจำแนก พบว่า ส่วนใหญ่วิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) ประมาณค่าได้ดีที่สุด เมื่อข้อมูลความสามารถมีการแจกแจงแบบปกติ หรือ เมื่อค่าอิทธิพลเทสต์เลทไม่เท่ากัน (Unequal effect) โดยกำหนดให้อิทธิพลเทสต์เลทมีค่าเป็น 0.25, 0.5, 1 และ 2 ตามลำดับ ส่วนวิธีแมกซิมัมไลค์ลิฮู้ด (ML) ส่วนใหญ่จะประมาณ

ค่าพารามิเตอร์อำนาจจำแนกได้ดีเมื่อข้อมูลความสามารถมีการแจกแจงแบบเบ้ซ้าย และเมื่อพิจารณาที่การประมาณค่าพารามิเตอร์ความสามารถ พบว่า ส่วนใหญ่วิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) ประมาณค่าได้ดีที่สุด

หากมองในภาพรวมทุกเงื่อนไขสำหรับการประมาณค่าพารามิเตอร์ความยาก อำนาจจำแนก และความสามารถจะพบว่า เงื่อนไขส่วนใหญ่ทั้งสามวิธีมีการประมาณค่าได้ไม่ต่างกันมาก ทั้งนี้อาจเป็นเพราะอิทธิพลของเทสต์เลทไม่มากพอที่จะทำให้เกิดผลกระทบต่อการประมาณค่า ซึ่งแม้จะไม่มีกฎเกณฑ์แน่นอนในการตัดสินว่าอิทธิพลของเทสต์เลทเท่าใดจึงจะมีอิทธิพลมากพอ (Ravand, 2015) แต่จากการศึกษาที่ผ่านมา ก็พบว่า ขนาดอิทธิพลเทสต์เลทที่น้อยกว่า 0.25 ไม่มีผลกระทบต่อการประมาณค่าพารามิเตอร์ ส่วนอิทธิพลที่เห็นผลชัดเจนมีค่า 0.5 - 2.0 (เช่น Wainer et al., 2000; Wang et al., 2002; Zhang, 2010) และสอดคล้องกับ Eckes & Baghaei (2015) ได้ตรวจสอบความไม่เป็นอิสระของข้อสอบใน C - Test แล้วพบว่า เมื่ออิทธิพลเทสต์เลทมีค่าน้อย การประมาณค่าด้วย 2 - PL TRT หรือ Standard IRT มีความเหมือนกันค่อนข้างมาก และ Cheng (1996) พบว่า การให้คะแนนแบบรายข้อมีประสิทธิภาพของการประมาณค่ามากกว่าการให้คะแนนแบบเทสต์เลท เมื่อระดับความเป็นอิสระกันของข้อสอบภายในเทสต์เลทต่ำ และการให้คะแนน แบบเทสต์เลทมีประสิทธิภาพของการประมาณค่ามากกว่า การให้คะแนนแบบรายข้อ เมื่อระดับความเป็นอิสระกันของข้อสอบภายในเทสต์เลทสูง ดังนั้น จะเห็นว่า เมื่อค่าอิทธิพลของเทสต์เลท ไม่เท่ากัน (Unequal effect) โดยกำหนดให้อิทธิพลของเทสต์เลทมีค่า เป็น 0.25, 0.5, 1 และ 2 พบว่า การประมาณค่าด้วยวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) จึงสามารถประมาณค่าพารามิเตอร์ได้ดีกว่าอีกสองวิธีอย่างเห็นได้ชัด นอกจากนี้ Purya & Hamdollah (2016) ได้ตรวจสอบความไม่เป็นอิสระของข้อสอบใน C-Test พบว่า ค่า RMSE ของพารามิเตอร์อำนาจจำแนกระหว่างโมเดล TRT (มีอิทธิพลเทสต์เลท) และ IRT (ไม่มีอิทธิพลเทสต์เลท) มีความแม่นยำพอกัน ส่วนพารามิเตอร์ความยากเป็นการวิเคราะห์ด้วยโมเดล TRT มีความแม่นยำมากกว่า

ส่วนใหญ่แล้วทุกวิธีสามารถประมาณค่าพารามิเตอร์ได้ดีเมื่อการแจกแจงของความสามารถเป็นแบบปกติ สำหรับเงื่อนไขที่คาดว่าจะพบในสถานการณ์จริง นั่นคือ กรณีที่แบบสอบมีข้อสอบที่ผสมกันระหว่างข้อสอบที่มีความเป็นอิสระและข้อสอบที่มีอิทธิพลเทสต์เสทผสมอยู่ หากเป็นการสอบที่มีจำนวนผู้สอบมาก เช่น การสอบระดับชาติ คาดว่าการแจกแจงความสามารถน่าจะเป็นแบบปกติ ซึ่งกรณีนี้ พบว่า ทั้งสามวิธีสามารถประมาณค่าพารามิเตอร์ความสามารถและอำนาจจำแนกได้ไม่ต่างกัน ยกเว้นการประมาณค่าพารามิเตอร์ความยาก ซึ่งวิธีของเบส์(Bayes) และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) จะประมาณค่าพารามิเตอร์ได้ดีกว่าวิธีแมกซิมัม

ไลค์ลิฮู้ด (ML) ส่วนกรณีที่ความสามารถมีการแจกแจงแบบเบ้นั้น การประมาณค่าความสามารถของผู้สอบ วิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) จะสามารถประมาณค่าได้ถูกต้องมากกว่าวิธีอื่น ๆ

2. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีลักษณะของความเป็นเทสต์เลทผสมในแบบสอบ ภายใต้เงื่อนไข 4 ปัจจัย ได้แก่ อิทธิพลของเทสต์เลท การแจกแจงของความสามารถ จำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบและอัตราส่วนของกลุ่มเปรียบเทียบต่อกลุ่มอ้างอิงที่ต่างกันนั้น พบว่า วิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) และวิธีของเบส์ (Bayes) สามารถควบคุมอัตราความคาดเคลื่อนประเภทที่ 1 ได้ดี (ยกเว้นกรณีที่อิทธิพลเทสต์เลทมีค่าเป็น 0.25, 0.5, 1, 2 ร่วมกับการแจกแจงความสามารถที่เป็นแบบเบ้ซ้ายและเบ้ขวา วิธีของเบส์มีความคาดเคลื่อนประเภทที่ 1 สูง) และมีอำนาจการตรวจสอบสูงเมื่อมีการแจกแจงความสามารถแบบเบ้ซ้ายและจำนวนตัวอย่างมาก แต่ไม่มากถึงเกณฑ์ทีกำหนด ตรงข้ามกับวิธีแมกซิมัมไลค์ลิฮู้ด (ML) ซึ่งไม่สามารถควบคุมอัตราความคาดเคลื่อนประเภทที่ 1 แต่มีอำนาจการตรวจสอบสูง ทั้งนี้ อาจเป็นเพราะอิทธิพลเทสต์เลททำให้เกิดผลทางลบ นั่นคือ ทำให้การประมาณค่าพารามิเตอร์ไม่ถูกต้องและลำเอียง (แสงหล้า ชัยมงคล, 2551) ซึ่งหากประมาณค่าพารามิเตอร์ที่ใช้ตัดสินใจสูงเกินจริงแล้วก็จะทำให้พบว่าตรวจสอบพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบได้มาก ประกอบกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมกซิมัมไลค์ลิฮู้ด (ML) ตัดสินจากการทดสอบนัยสำคัญด้วย z-statistic (ผลการประมาณค่าปฏิสัมพันธ์หารด้วยความคาดเคลื่อนมาตรฐาน) ส่วนวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) และวิธีของเบส์ (Bayes) ตัดสินการทำหน้าที่ต่างกันของข้อสอบด้วยการเทียบกับเกณฑ์ นั่นคือ ค่าสัมบูรณ์ของขนาดพารามิเตอร์ที่ใช้ตัดสินการทำหน้าที่ต่างกันของข้อสอบ (β_j) มากกว่า 0.426 และช่วงค่าขอบบนและขอบล่างของช่วงความเชื่อมั่น 95% ไม่คลุมศูนย์ ซึ่งมีค่าใกล้เคียงกับขนาดของการทำหน้าที่ต่างกันของข้อสอบที่จำลองข้อมูล (Impact = 0.5) ที่ใช้ประมาณค่า DIF ประกอบกับ Fukuhara & Kamata (2011) พบว่า เมื่ออำนาจจำแนกมีค่าเพิ่มขึ้น วิธีของเบส์ (โมเดล Standart IRT และ Bi - factor MIRT) จะประมาณค่าขนาดของ DIF ในข้อที่ทำหน้าที่ต่างกันต่ำกว่าความเป็นจริง (Underestimated) สอดคล้องกับ Gulsen & Nuri (2015) ศึกษาความเที่ยง โดยใช้ G - theory และการทำหน้าที่ต่างกันของข้อสอบโดยประมาณค่าพารามิเตอร์ด้วยวิธีของเบส์ โดยใช้ WinBUGS พบว่า ขนาดของ DIF ที่ประมาณค่าด้วยโมเดลที่กำจัดอิทธิพลเทสต์เลทจะน้อยกว่าโมเดล Standard IRT ดังนั้น เมื่อทำการประมาณค่าพารามิเตอร์ จึงอาจได้ค่าที่น้อยกว่าเกณฑ์ที่กำหนด จึงส่งผลต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้น้อยลง

อย่างไรก็ตามเมื่อพิจารณาภายใต้เงื่อนไข 4 ปัจจัย มีรายละเอียดแต่ละปัจจัย ดังนี้

2.1 อิทธิพลของเทสต์เลท พบว่า วิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) และวิธีของเบส์ (Bayes) สามารถควบคุมอัตราความคาดเคลื่อนประเภทที่ 1 ได้ดี เมื่อแบบสอบมีค่าอิทธิพลเทสต์เลทเท่ากันทุกเทสต์เลท และแบบสอบประกอบด้วยข้อสอบที่เป็นอิสระผสมกับเทสต์เลท ส่วนกรณีที่แบบสอบมีค่าอิทธิพลของเทสต์เลทไม่เท่ากัน วิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) สามารถควบคุมอัตราความคาดเคลื่อนประเภทที่ 1 ได้ดี ส่วนวิธีของเบส์ (Bayes) สามารถควบคุมอัตราความคาดเคลื่อนประเภทที่ 1 ได้ดีเฉพาะกรณีที่การแจกแจงความสามารถเป็นแบบปกติ อย่างไรก็ตาม ทั้งสองวิธีมีอำนาจการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบต่ำกว่าเกณฑ์ ส่วนวิธีแมกซิมัมไลค์ลิฮู้ด (ML) ไม่สามารถควบคุมอัตราความคาดเคลื่อนประเภทที่ 1 ได้ แต่มีอำนาจการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูง ซึ่งผลของความสามารถในการควบคุมอัตราความคาดเคลื่อนประเภทที่ 1 มีลักษณะคล้ายกับการประมาณค่าพารามิเตอร์ จึงอาจเกิดจากสาเหตุเดียวกัน นั่นคือ ขนาดอิทธิพลเทสต์เลทที่น้อยกว่า 0.25 ไม่มีผลกระทบต่อการประมาณค่าพารามิเตอร์ ส่วนอิทธิพลเทสต์เลทที่เห็นผลชัดเจนมีค่า 0.5 - 2.0 จึงทำให้ขนาดอิทธิพลเทสต์เลทค่าอื่น ๆ มีผลใกล้เคียงกัน ยกเว้นกรณีที่อิทธิพลของเทสต์เลทมีค่าเป็น 0.25, 0.5, 1, 2

2.2 การแจกแจงของความสามารถ พบว่า วิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) และวิธีของเบส์ (Bayes) สามารถควบคุมอัตราความคาดเคลื่อนประเภทที่ 1 ได้ดี แต่มีอำนาจการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบต่ำกว่าเกณฑ์ โดยที่การแจกแจงแบบเบ้จะมีอำนาจการตรวจสอบมากกว่าการแจกแจงแบบปกติ ทั้งนี้ขัดแย้งกับผลการศึกษาของ Monaco (1997) ทำการศึกษาการแจกแจงความสามารถแบบเบ้ที่มีผลกับการทำหน้าที่ต่างกันของข้อสอบด้วยการเปรียบเทียบวิธี Mantel - Haenszel และวิธี DFIT โดยใช้ข้อมูลจำลอง พบว่า เมื่อการแจกแจงพารามิเตอร์ความสามารถมีความเบ้มาก ส่งผลให้ความถูกต้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบลดลง แต่การแจกแจงแบบเบ้ปานกลางไม่ส่งผลต่อการตรวจจับมากนัก ส่วนวิธีแมกซิมัมไลค์ลิฮู้ด (ML) ไม่สามารถควบคุมอัตราความคาดเคลื่อนประเภทที่ 1 ได้ แต่มีอำนาจการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูงกว่าอีกสองวิธีในทุกการแจกแจง

2.3 จำนวนข้อสอบที่มีการทำหน้าที่ต่างกัน พบว่า อัตราความคาดเคลื่อนประเภทที่ 1 มีผลกระทบเฉพาะกรณีที่อิทธิพลเทสต์เลทมีค่าเป็น 0.25, 0.5, 1, 2 ที่การแจกแจงความสามารถแบบเบ้ มีความไม่แน่นอน ส่วนอำนาจในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ พบว่า จำนวนข้อสอบที่มีการทำหน้าที่ต่างกันปกติ (5 ข้อ หรือ DIF ร้อยละ 12.5 ในแบบสอบ) จะมีอำนาจในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้มากกว่า จำนวนข้อสอบที่มีการทำหน้าที่

ต่างกันมาก (8 ข้อ หรือ DIF ร้อยละ 20 ในแบบสอบ) สอดคล้องกับ Narayanan & Swaminathan (1994) อธิบายว่า ปกติแล้วแบบสอบยาวจะมีความน่าเชื่อถือมากกว่าแบบสอบสั้น เนื่องจากผลของการประมาณค่าความสามารถมีความน่าเชื่อถือ ในทางกลับกันถ้าสัดส่วนของข้อสอบที่ DIF มากขึ้น (จำนวนข้อที่ทำหน้าที่ต่างกันในแบบสอบ) จะทำให้การประมาณค่าความสามารถมีความน่าเชื่อถือน้อยลง ซึ่งส่งผลให้มีความถูกต้องน้อยลง ดังนั้น อำนาจของการตรวจสอบการทำหน้าที่ต่างกันก็จะลดลงตามไปด้วย

2.4 อัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ พบว่า ทั้งสามวิธี เมื่อจำนวนผู้สอบมาก (อัตราส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบเป็น 1000: 1000) อัตราความคาดเคลื่อนประเภทที่ 1 (Type I error) จะน้อยกว่าจำนวนผู้สอบน้อย และอำนาจทดสอบ (Power) มากกว่าจำนวนผู้สอบน้อย สอดคล้องกับ Narayanan & Swaminathan (1994) อธิบายว่า อัตราการตรวจจับข้อสอบที่ทำหน้าที่ต่างกันจะเพิ่มขึ้น เมื่อจำนวนตัวอย่างมากขึ้น และ Hambleton (1991, อ้างถึงใน ศิริชัย กาญจนวาสี, 2550) ระบุว่าการประมาณค่าพารามิเตอร์ด้วยวิธีแมกซิมัมไลค์ลิฮู้ด เมื่อวิเคราะห์โมเดล 2 และ 3 พารามิเตอร์ ค่าพารามิเตอร์จะมีความคงเส้นคงวาได้ ก็ต่อเมื่อมีจำนวนข้อสอบมากและกลุ่มผู้สอบมีขนาดใหญ่

หากพิจารณาเงื่อนไขที่คาดว่าจะพบในสถานการณ์จริงบ่อยที่สุด ซึ่งอาจเป็นกรณีที่แบบสอบมีข้อสอบที่ผสมกันระหว่างข้อสอบที่มีความเป็นอิสระและข้อสอบที่มีอิทธิพลเทสต์เลทผสมอยู่ นั้น พบว่า วิธีของเบส์ (Bayes) และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) สามารถควบคุมอัตราความคาดเคลื่อนประเภทที่ 1 ได้ดีกว่าวิธีแมกซิมัมไลค์ลิฮู้ด (ML) อย่างไรก็ตาม เมื่อพิจารณาที่อำนาจการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ กลับพบว่า วิธีแมกซิมัมไลค์ลิฮู้ด (ML) ซึ่งไม่สามารถควบคุมอัตราความคาดเคลื่อนประเภทที่ 1 ได้นั้น กลับมีอำนาจการตรวจสอบสูง ส่วนวิธีของเบส์ (Bayes) และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) มีอำนาจการตรวจสอบสูงเมื่อมีการแจกแจงความสามารถแบบเบ้ซ้ายและจำนวนตัวอย่างมาก แต่ไม่มากถึงเกณฑ์ที่กำหนด

ข้อเสนอแนะ

ข้อเสนอแนะในการนำผลการวิจัยไปใช้

1. ผลการศึกษาครั้งนี้ พบว่า การประมาณค่าพารามิเตอร์ความสามารถของผู้สอบจากแบบสอบที่มีลักษณะของเทสต์เลท ส่วนใหญ่วิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) ประมาณค่าได้ดีที่สุด ดังนั้น สำหรับการสอบที่ต้องตัดสินความสามารถของผู้สอบ ควรใช้วิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) เนื่องจากสามารถประมาณค่าได้ถูกต้องมากกว่าวิธีอื่น ๆ

2. ในการวิเคราะห์ข้อสอบที่มีลักษณะของความเป็นเทสต์เลทผสมในแบบสอบ ควรตรวจสอบข้อมูลก่อนการวิเคราะห์คุณภาพข้อสอบ เช่น หากข้อมูลความสามารถมีการแจกแจงแบบปกติ ควรประมาณค่าพารามิเตอร์ของข้อสอบโดยใช้วิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) หากข้อมูลความสามารถมีการแจกแจงแบบเบ้ สามารถใช้การประมาณค่าพารามิเตอร์ข้อสอบโดยใช้วิธีแมกซิมัมไลค์ลิฮู้ด (ML) แทนได้ เนื่องจากจะได้ผลการวิเคราะห์ใกล้เคียงกัน แต่ประหยัดเวลาในการคำนวณมากกว่า

3. ผลการศึกษาครั้งนี้ พบว่า เมื่อแบบสอบประกอบด้วยอิทธิพลของเทสต์เลทมีค่าเป็น 0.25, 0.5, 1, 2 (ระดับของอิทธิพลเทสเลดต่างกันและอิทธิพลเทสเลดอยู่ในระดับมาก) จะมีผลการประมาณค่าพารามิเตอร์ที่ต่างกันที่สุด โดยที่วิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) สามารถประมาณค่าได้ดีที่สุด ดังนั้น ก่อนการวิเคราะห์ข้อสอบควรมีการตรวจสอบระดับของความไม่เป็นอิสระของข้อสอบก่อน เพื่อให้แน่ใจว่าจะใช้โมเดลใดที่เหมาะสมที่สุดในการวิเคราะห์ข้อสอบ

4. ผลการศึกษาครั้งนี้ พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีลักษณะของความเป็นเทสต์เลทผสมในแบบสอบ ภายใต้เงื่อนไข 4 ปัจจัย แม้วิธีแมกซิมัมไลค์ลิฮู้ด (ML) จะมีอำนาจการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูง แต่ไม่สามารถควบคุมอัตราความคาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ที่กำหนด ตรงข้ามกับวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) ที่มีความสามารถในการควบคุมอัตราความคาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ และมีอำนาจการตรวจสอบสูงหากข้อมูลมีการแจกแจงความสามารถแบบเบ้ซ้ายและจำนวนตัวอย่างมาก แม้อำนาจการตรวจสอบจะไม่ถึงเกณฑ์ที่กำหนดก็ตาม ดังนั้น สำหรับ การตัดสินการทำหน้าที่ต่างกันของข้อสอบ จึงควรใช้วิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) เนื่องจากตัดสินใจผิดพลาดน้อยกว่า

ข้อเสนอแนะในการวิจัยครั้งต่อไป

1. เนื่องจากการศึกษาครั้งนี้เป็นการศึกษาการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป (Uniform DIF) ดังนั้นการศึกษาครั้งต่อไป ควรศึกษาการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป (Non - uniform DIF) เพื่อเปรียบเทียบผลการประมาณค่าพารามิเตอร์และการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบของลักษณะข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบที่มีลักษณะต่างกัน

2. การศึกษาครั้งนี้ กำหนดให้แบบสอบประกอบด้วยเทสต์เลทที่มีจำนวนข้อเท่ากันทุกเทสต์เลท ดังนั้น การศึกษาครั้งต่อไป ควรศึกษาแบบสอบที่ประกอบด้วยเทสต์เลทที่มีจำนวนข้อในแต่ละเทสต์เลทในหลายรูปแบบ

3. โปรแกรมที่ใช้ในการประมาณค่าพารามิเตอร์ด้วยวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) ในการศึกษาครั้งนี้ใช้โปรแกรม R ร่วมกับโปรแกรม WinBUGS แม้จะง่ายต่อการนำไปใช้ และมีความยืดหยุ่น เนื่องจากสามารถเรียกใช้งานได้จาก R2WinBUGS Package แต่ก็มีอุปสรรค คือ เวลาที่ใช้ในการประมวลผลค่อนข้างนาน (Time - consuming process) สำหรับการศึกษาครั้งนี้ใช้เวลาประมาณ 8 - 10 ชั่วโมงต่อครั้ง (ขึ้นอยู่กับเงื่อนไขและสภาพแวดล้อมของคอมพิวเตอร์) ดังนั้น สำหรับการศึกษาครั้งต่อไป อาจเลือกใช้โปรแกรม (Software) หรือสำรวจกระบวนการอื่น ๆ ในการนำวิธีประมาณค่านี้ไปใช้ โดยให้มีเวลาการประมวลผลที่น้อยลง

4. ควรศึกษาลักษณะของข้อมูลในด้านอื่น ๆ ที่มีผลกระทบต่อการประมาณค่าพารามิเตอร์หรือการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เช่น แบบสอบที่ประกอบด้วยข้อมูลที่มีการสูญหาย (Missing data) ข้อมูลสุดโต่ง (Outlier) แบบสอบที่มีข้อมูลผสมกันของวิธีการให้คะแนนรายข้อแบบสองค่าและหลายค่า (Dichotomous and polytomous data) เป็นต้น

5. แม้ว่าวิธีของเบส์แบบมีอิทธิพลเทสต์เลท (Bayesγ) จะกำจัดปัญหาความไม่เป็นอิสระในการตอบข้อสอบออกได้รองรับการวิเคราะห์แบบ 2 พารามิเตอร์ และไม่ทำให้สารสนเทศรายข้อหายไป ซึ่งสอดคล้องกันความเป็นจริงมากกว่าโมเดลอื่น ๆ แต่อุปสรรคอย่างหนึ่งของโมเดลนี้ คือ มีจำนวนพารามิเตอร์ที่ต้องประมาณค่ามาก ซึ่งอาจทำให้ความพอดี (Fit) ของโมเดลอาจจะไม่ดีนัก (Li, Li & Wang, 2010) ดังนั้น การศึกษาครั้งต่อไปอาจทำการเปรียบเทียบการประมาณค่าพารามิเตอร์และตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับโมเดลอื่น ๆ เช่น Four - level IRT model (Jiao, et al., 2012) เป็นต้น

6. ควรศึกษาลักษณะความเป็นเทสต์เลทที่มีผลต่อการประมาณค่าความสามารถในกรณีของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (Computerized adaptive testing: CAT) เนื่องจากการสุ่มเลือกข้อสอบอาจเกิดกรณีที่ผลการตอบของข้อสอบข้อหนึ่งมาจากโจทย์ของอีกข้อหนึ่ง (Cross - information) ซึ่งแสดงถึงความไม่เป็นอิสระของการตอบคำถามในแต่ละข้อ ทำให้เกิดการคำนวณค่าความสามารถของผู้สอบคาดเคลื่อน แล้วส่งผลต่อการยุติการสอบก่อนกำหนด

บรรณานุกรม

กาญจนา วัธนสุนทร. (2537). *การพัฒนาเกณฑ์ตัดสินข้อสอบลำเอียงทางเพศ*. วิทยานิพนธ์ครุศาสตรดุษฎีบัณฑิต, สาขาวิชาการวัดและประเมินผลการศึกษา, บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย.

คมศักดิ์ ชื่นชม. (2539). *การศึกษาผลการวิเคราะห์ความลำเอียงที่ใช้วิธีต่างกันของแบบทดสอบวัดจริยธรรมด้านความซื่อสัตย์*. วิทยานิพนธ์การศึกษามหาบัณฑิต, สาขาวิชาการวัดผลการศึกษา, บัณฑิตวิทยาลัย, มหาวิทยาลัยศรีนครินทรวิโรฒ.

จิติมา วรรณศรี. (2539). *การเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมนเทล - แฮนส์เซลกับวิธีซิบเทสท์ เมื่อความยาวแบบสอบ ขนาดกลุ่มตัวอย่างและอัตราส่วน ของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบต่างกัน*. วิทยานิพนธ์ครุศาสตรมหาบัณฑิต, สาขาวิชาการวัดและประเมินผลการศึกษา, คณะครุศาสตร์, จุฬาลงกรณ์มหาวิทยาลัย.

ชนะศึก นิชานนท์. (2553). *ประสิทธิภาพของการประมาณค่าพารามิเตอร์แบบเบส์ โดยใช้การสรุปอ้างอิงความน่าเชื่อถือของโมเดลการตอบสนองข้อสอบ*. วิทยานิพนธ์ครุศาสตรดุษฎีบัณฑิต, สาขาวิชาการวัดและประเมินผลการศึกษา, คณะครุศาสตร์, จุฬาลงกรณ์มหาวิทยาลัย.

นพดล มีชั้นช่วง. (2544). *การเปรียบเทียบผลของการประมาณค่าพารามิเตอร์ตามทฤษฎีการตอบสนอง ข้อสอบระหว่างวิธีแมกซิมัมไลค์ลิฮูด วิธีฮิวริสติก และวิธีของเบย์ของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนคณิตศาสตร์ชั้นมัธยมศึกษาปีที่ 1*. วิทยานิพนธ์การศึกษามหาบัณฑิต, การวัดผลการศึกษา, คณะศึกษาศาสตร์, มหาวิทยาลัยมหาสารคาม

วุฒิชัย วงษ์ทัศนีย์กร. (2555). *การวิเคราะห์แบบจำลอง*. เข้าถึงได้จาก http://wuthichai.ie.engr.tu.ac.th

ศิริชัย กาญจนาวสี. (2550). *ทฤษฎีการทดสอบแนวใหม่ (Modern test theory)* (พิมพ์ครั้งที่ 3). กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

แสงหล้า ชัยมงคล. (2551). การตรวจสอบความไม่เป็นอิสระเฉพาะที่ระหว่างคู่ของข้อสอบในกรณีที่ผลตอบสนองของข้อสอบเป็นแบบพหุวิภาค โดยใช้หลักการเอนโทรปีสารสนเทศ. *วารสารวิทยาศาสตร์และเทคโนโลยี*, *16*(1), 1 - 9.

สุนทร เทียนงาม, ศิริชัย กาญจนวาสี และดิเรก ศรีสุโข. (2553). ผลของความไม่เป็นอิสระของข้อสอบที่มีต่อค่าความเที่ยง ค่าพารามิเตอร์ของข้อสอบ ค่าความสามารถของผู้สอบและสารสนเทศของแบบสอบเมื่อมีเงื่อนไขการทดสอบที่แตกต่างกัน. *วารสารวิธีวิทยาการวิจัย*, *23*(3), 247 - 271.

สุพัฒนา หอมบุปผา. (2556). *การเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN*. วิทยานิพนธ์ปรัชญาดุษฎีบัณฑิต, สาขาวิชาวิจัย วัดผลและสถิติการศึกษา, คณะศึกษาศาสตร์, มหาวิทยาลัยบูรพา.

สุทธิพร ศุรธณี. (2550). *การศึกษาความสามารถในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามตัวแบบเชิงเส้นวางนัยทั่วไประดับลดหลั่น*. วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต, สาขาวิชาสถิติประยุกต์, คณะวิทยาศาสตร์และเทคโนโลยี, มหาวิทยาลัยธรรมศาสตร์.

สิริรัตน์ วิภาสศิลป์. (2545). *การเปรียบเทียบวิธีซิบเทสท์และดีเอฟไอทีในการตรวจสอบการทำหน้าที่เบี่ยงเบน ของข้อสอบ หมวดข้อสอบและแบบทดสอบจากข้อมูลการตอบข้อสอบที่ใช้ความสามารถหลายมิติ*. วิทยานิพนธ์การศึกษาดุษฎีบัณฑิต, การทดสอบและวัดผลการศึกษา, บัณฑิตวิทยาลัย, มหาวิทยาลัยศรีนครินทรวิโรฒ.

อนุสรณ์ เกิดศรี. (2557). *ประสิทธิภาพของวิธีการคัดเลือกข้อสอบสองวิธีในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ สำหรับโมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย: การเปรียบเทียบระหว่างวิธีมอนติ คาร์โล ซีเอที และวิธีแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ*. วิทยานิพนธ์ครุศาสตรดุษฎีบัณฑิต, สาขาวิชาการวัดและประเมินผลการศึกษา, คณะครุศาสตร์, จุฬาลงกรณ์มหาวิทยาลัย.

อัชฌา อระวีพร. (2554). การหาค่าตัวประมาณเบส์ด้วยโปรแกรมวินบั๊ก. *วารสารวิทยาศาสตร์ลาดกระบัง*, *20*(2), 45 - 60.

อัญชลี ธีระวุฒิ. (2555). *การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ Pre O - NET ชั้นมัธยมศึกษาปีที่ 6 ของสำนักงานเขตพื้นที่การศึกษามัธยมศึกษาเขต 35*. วิทยานิพนธ์ครุศาสตรมหาบัณฑิต, การวัด ประเมินและวิจัยทางการศึกษา, มหาวิทยาลัยราชภัฏลำปาง.

อิทธิฤทธิ์ พงษ์ปิยะรัตน์. (2551). *การวิเคราะห์ข้อสอบและการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ: การวิเคราะห์พหุระดับ*. วิทยานิพนธ์ครุศาสตรดุษฎีบัณฑิต, สาขาวิชาการวัดและประเมินผลการศึกษา, คณะครุศาสตร์, จุฬาลงกรณ์มหาวิทยาลัย.

Ackerman, T. A. (1987). The Robustness of Logist and Bilog IRT Estimation Programs to Violations of Local Independence. *Paper Presented at the Annual Meeting of 1987 AERA*. Washington, D.C. April 20 - 24, 1987.

Allen, M. J., & Yen, W. M. (1979). *Introduction to Measurement Theory*. CA: Brooks/ Cole Publishing Company.

Andrich, D. A. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561 - 573.

Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1999). An investigation of the power of the likelihood ratio goodness of fit statistic in detecting differential item functioning. *Journal of Educational Measurement*, *36*(4), 277 - 300.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In Paul, W.H., and Howard, W. (Eds.), *Differential Item Functioning*. pp. 3 - 23. New Jersey: America Lawrence Erlbaum Association.

Awuor, A. R. (2008). *Effect of unequal sample sizes on the power of DIF detection: An IRT -Based Monte Carlo study with SIBTEST and Mantel - Haenszel procedures*. Doctor of Philosophy in Educational Research and Evaluation, Virginia Polytechnic Institute and State University.

Azzalini, A. (2014). *The Skew - Normal and Related Families*. New York: Cambridge University Press.

Bafumi, J., Gelman, A., Park, D., & Kaplan, N. (2005). Pactical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis*, *13*(2), 171 - 187.

Bao, H. (2007). *Investigating differential item function amplifcation and cancellation in application of item response testlet models*. Doctoral dissertation in Philosophy, University of Maryland.

Berenson, M. L., Levine, D. M., & Krehbiel, T. C. (2012). *Basic Business Statistics; Concepts and Applications*. (12th ed.). New Jersey: Prentice Hall.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(29), 29 - 51.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*(2)*,* 153 - 168.

Bulut, O. (2015). An empirical analysis of gender-based DIF due to test booklet effect. *European Journal of Research on Education*, *3*(1), 7 - 16.

Burhanettin Özdemir. (2015). A comparison of IRT - based methods for examining differential item functioning in TIMSS 2011 mathematics subtest. *Procedia - Social and Behavioral Sciences*, *174*, 2075 - 2083.

Chaimongkol, S., Huffer, F. W., & Kamata, A. (2007). *An explanatory differential item functioning (DIF) model by the WinBUG 1.4*. Journal of Science and Technology, *29*(2), 449 - 458.

Chen, T. (2010). *Random or fixed testlet effects: A comparison of two multilevel testlet models*. Doctoral Dissertation in Philosophy. University of Texas at Austin.

Cheng, S. C. (1996). *A Comparison of testlet and individual item scoring*. Doctoral Dissertation in Philosophy, University of Texas at Austin.

Christine, E. (2012). Confirming testlet effects. *Applied Psychological Measurement*, *36*(2), 104 -121.

Clauser, B. E. (1991). Examination of various influence on the Mantel - Haenszel statistic. *Paper Presented at the Annual Meeting of American Educational Research Association*. Chicago Illinois, *11*(4), 3 - 7.

DeMars, C. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement*, *43*(2), 145 - 168.

DeMars, C. (2010). *Item response theory*. New York: Oxford University Press.

Eckes, T., & Baghaei, P. (2015). Using testlet response theory to examine local dependence in C - Tests. *Applied Measurement in Education*, *28*(2), 85 - 98.

Eckes, T., (2014). Examining testlet effects in the TestDaF listening section: A testlet response theory modeling approach. *Language Testing*, *31*(1), 39 - 61.

Figueiredo, F., & Gomes, M. I. (2013). The Skew - Normal distribution in SPC. *Revstat: Statistical Journal*, *11*(1), 83 - 104.

Fukuhara, H. (2009). *A differential item functioning model for testlet - based items using bi - factor multidimensional item response theory model: A Bayesian approach*. Doctoral Dissertation in Philosophy. Florida State University.

Fukuhara, H., & Kamata, A. (2011). A differential item functioning model for testlet - based items using a bi - factor multidimensional item response theory model: A bayesian approach. *Applied Psychological Measurement*, *35*(8), 604 - 622.

Gelman, A. E., Carlin, J. B., Stern, H. S., & Rubin, R. D. (2003). *Bayesian data analysis* (2nd ed.). New York: Chapman & Hall/ CRC.

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.

Glas, C. A. W. (2012). *Estimating and testing the extended testlet model*. (LSAC Research Report, RR 12-03). Newtown, PA: Law School Admission Council, Inc.

Glas, C. A. W., & Meijer R. R. (2003). A bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement*, *27*(3), 217 - 233

Glas, C. A. W., Wainer, H., & Bradlow, E. T. (2000). MML and EAP estimation in testlet-based adaptive testing. In W. J . van der Linden & C. A. W. Glas (Eds.), *ComputerAdaptive Testing: Theory and Practice* (pp. 271 - 288). Netherlands: Kluwer.

Gorin, J. S., Dodd, B. G., Fitzpatrick, S. J., & Shieh, Y. Y. (2005). Computerized adaptive testing with the partial credit model: Estimation procedures, population distributions, and item pool characteristics. *Applied Psychological Measurement*, *29*(6), 433 - 456.

Gulsen, T. T. & Nuri D. (2015). The effects of testlets on reliability and differential item functioning. *Educational Sciences: Theory & Practice*, *15*(4), 969 - 980.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. California: Sage Publications.

Harwell, M., Stone, C. A., Hsu T., Kirisci L. (1996). Monte carlo studies in item response theory. *Applied Psychologiscal Measurement*, *20*(2), 101 - 125.

Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement*, *15*(3), 279 - 291.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Huggins, A. C. (2014). The effect of differential item functioning in anchor items on population invariance of equating. *Educational and Psychological Measurement*, *74*(4), 627 - 658.

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement*, *49*(1), 82 - 100.

Jiao, H., Wang S., & Kamata, A. (2005). Modeling local item dependence with the hierarchical generalized linear model. *Journal of Applied Measurement*, *6*(3), 311 - 321.

Jiao, H., Wang S., & He, W., (2013). Estimation methods for one - parameter testlet models. *Journal of Educational Measurement*, *50*(2), 186 - 203.

Johnson, V., & Albert, J. (1999). *Ordinal data modeling*. New York: Springer - Verlag.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*(1), 79.

Lai, P., & Twu., B. (2013). The influence of testlet effect on the accuracy of parameter estimation. *Taiwan Academic Online*, *47*(1), 113 - 134.

Lee, Y., Cohen, A., Toro, M. (2009). *Examining type I error and power for detection for differential item and testlet functioning*. Asia Pacific Education Review, *10*(3), 365 - 375.

Lee, G., & Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education*, *12*(3), 237 - 255.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, *30*(1), 3 - 21.

Li, Y., Li, S., & Wang, L. (2010). *Application of a general polytomous testlet model to the reading section of a large-scale English language assessment* (Research Report). Princeton, New Jersey: Educational Testing Service.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Pscychometrika*, 47, 149 - 174.

Monaco, M. (1997). *A monte carlo study of skewed theta distribution on DIF indicies*. retrieved from http://eric.ed.gov/?id=ED411246

Narayanan, P. & Swaminathan, H. (1994). Performance of the Mantel - Haenszel and simultaneous item bias procedures for detecting differentail item functioning. *Applied Psychological Measurement*, *18*(4), 315 - 328.

Patz, R. J., & Junker, B. W. (1999). A straightforward and approach to markov chain monte carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*(2), 146-178.

Purya B., & Hamdollah R., (2016). Modeling local item dependence in cloze and reading comprehension test items using testlet response theory. *Psicologica*, *37*, 85 - 104.

Ra, J. (2011). *Sensitivity of prior specification within testlet model*. Doctoral Dissertation in Philosophy, University of Georgia.

Raftery, A. L., & Lewis, S. (1992). How many iterations in the Gibbs sampler? In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4* (pp. 763 - 773). Oxford, UK: Oxford University Press.

Ratana, N. (1993). Simultaneous DIF amplification and cancellation: Shealy - Stout' s test for DIF. *Journal of Educational Measurement*, *30*(4), 293 - 311.

Raju, S., Linden, J. & Fleer F. (1995). IRT - based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, *19*(4), 353 - 368.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Institute of Educational Research.

Ravand, H., (2015). Assessing testlet effect, impact, differential testlet, and item functioning using cross-classified multilevel measurement modeling. *SAGE Open*, *5*(2), 1 - 9.

Ripley, B. D. (1987). *Stochastic simulation*. New York: Wiley.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel - Haenszel type I error performance. *Journal of Educational Measurement*, *33*(2), 215 - 230.

Sahu, S. K. (2002). Bayesian estimation and model choice in item response models. *Journal of Statistical Computation and Simulation*, *72*(3), 217 - 232.

Samejima, F. (1968). *Estimation of Latent Ability Using a Response Pattern of Graded Scores*. (Research Report). Princeton, New Jersey: Educational Testing Service.

Sedivy, S.K. (2009). *Using Traditional methods to detect differential item functioning in testlet data*. Doctoral Dissertation in Educational Philosophy, University of Wisconsin - Milwaukee.

Shin, S. H., & Wall, N. L. (2006). *Three differential item functioning detection methods with three different ability distributions*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet - based tests. *Journal of Educational Measurement*, *28*(3), 237 - 247.

Soares, M. T., Gonçalves, B. F., & Gamerman, D. (2009). An Integrated Bayesian Model for DIF Analysis. *Journal of Educational and Behavioral Statistics*, *34*(3), 348 - 377.

Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two - parameter logistic model. *Psychometrika*, *50*(3), 349 - 364.

Swaminathan, H., & Gifford, J. A. (1986). Bayesian estimation in the three - parameter logistic model. *Psychometrika*, *51*(4), 599 - 601.

Swaminathan, H., & Roger, J., H. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361 - 370.

Tao, J., Xu, B., Shi, N., & Jiao, H. (2013). Refining the two-parameter testlet response model by introducing testlet discrimination parameters. *Japanese Psychological Research*, *55*(3), 284 - 291.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiplecategorical-response models. *Journal of Educational Measurement*, *26*, 247 - 260.

Vincent, K., & Prathiba, N. (2012). Recovery of graded response model parameters: A Comparison of Marginal. Applied Psychological Measurement, *36*(5), 399 - 419.

Wainer H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet - based adaptive testing. In W. J. van der Linden & C. A. W. Glass (Eds.), *Computerized Adaptive Testing: Theory and Practice.* 245 - 269. Netherlands: Kluwer Academic Publishers.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge: Cambridge University Press.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability?. *Educational Measurement: Issues and Practice*, *15*(1), 22 - 29.

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and application. *Applied Psychological Measurement, 26*, 109 - 128.

Wang, W., & Wilson, M. (2005a). Assessment of differential item functioning in testlet - based items using the Rasch testlet model. *Educational and Psychological Measurement*, *65*, 549 - 576.

Wang, W., & Wilson, M. (2005b). The rasch testlet model. *Applied Psychological Measurement*, *29*, 126 - 149.

Welkenhuysen - Gybels, J. (2004). The performance of some observed and unobserved conditional invariance techniques for the detection of differential item functioning. *Quality & Quantity*, *38*, 681 - 702.

Xu, X., & Jia, Y. (2011). *The sensitivity of parameter estimates to the latent ability distribution* (Research Report 11 - 40). Princeton, NJ: Educational Testing Service.

Yen, W. (1993). Scaling performance assessment: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187 - 213.

Yue, L., Hong-Yun, L. (2012). When should we use testlet model? A comparison study of bayesian testlet random - effects model and standard 2 - PL bayesian model. *Acta Psychologica Sinca*, *44*(2), 263 - 275.

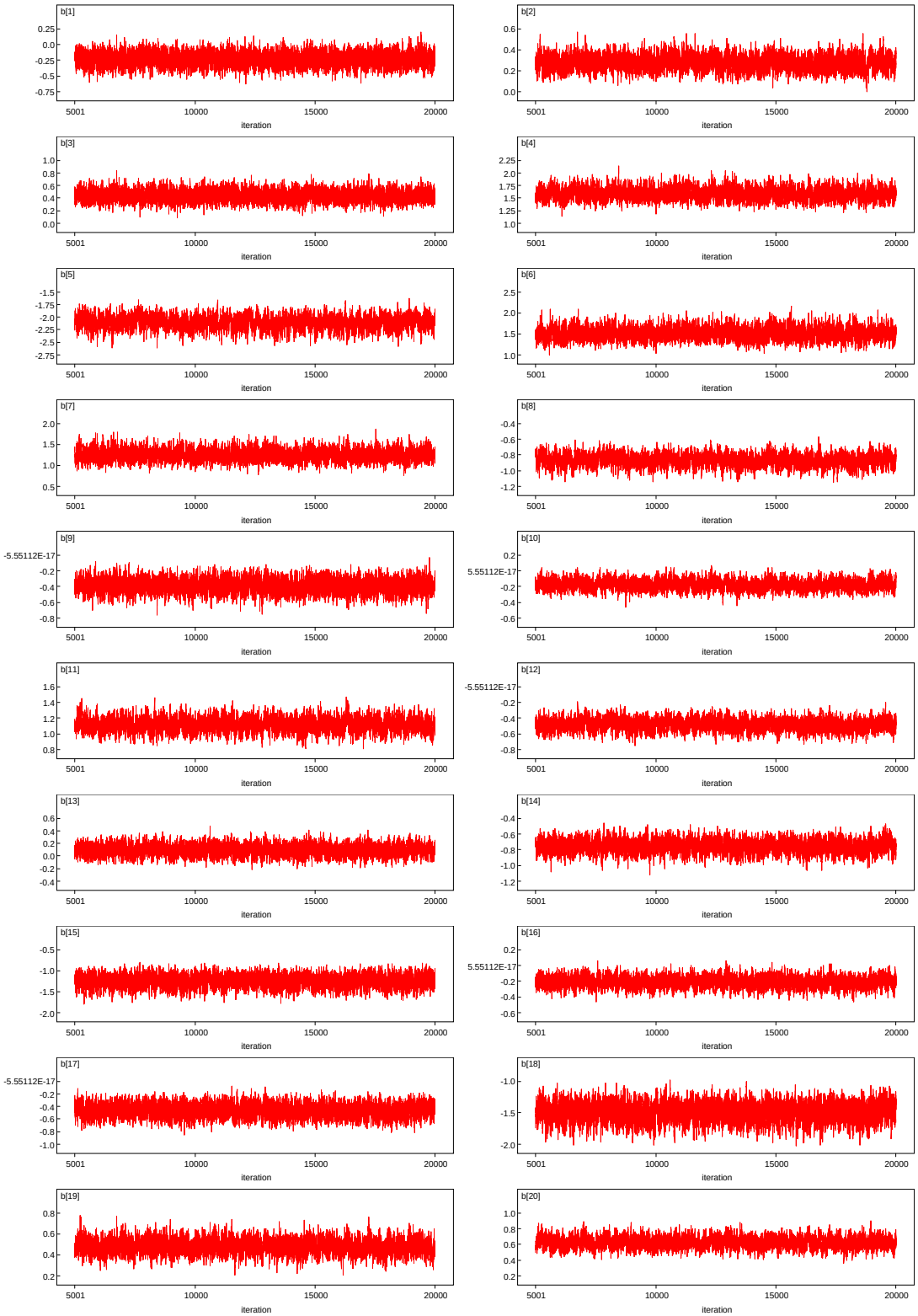
Zenisky, A. L., Hambleton, R. K., & Sireci, S. G.(2003). *Effect of Local Item Dependence on the Validity of IRT Item, Test, and Ability Statistics*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Zhang, O. (2010). *Polytomous IRT or testlet model: An evaluation of scoring models in small testlet size situations*. Master of Arts in Education, University of Florida.

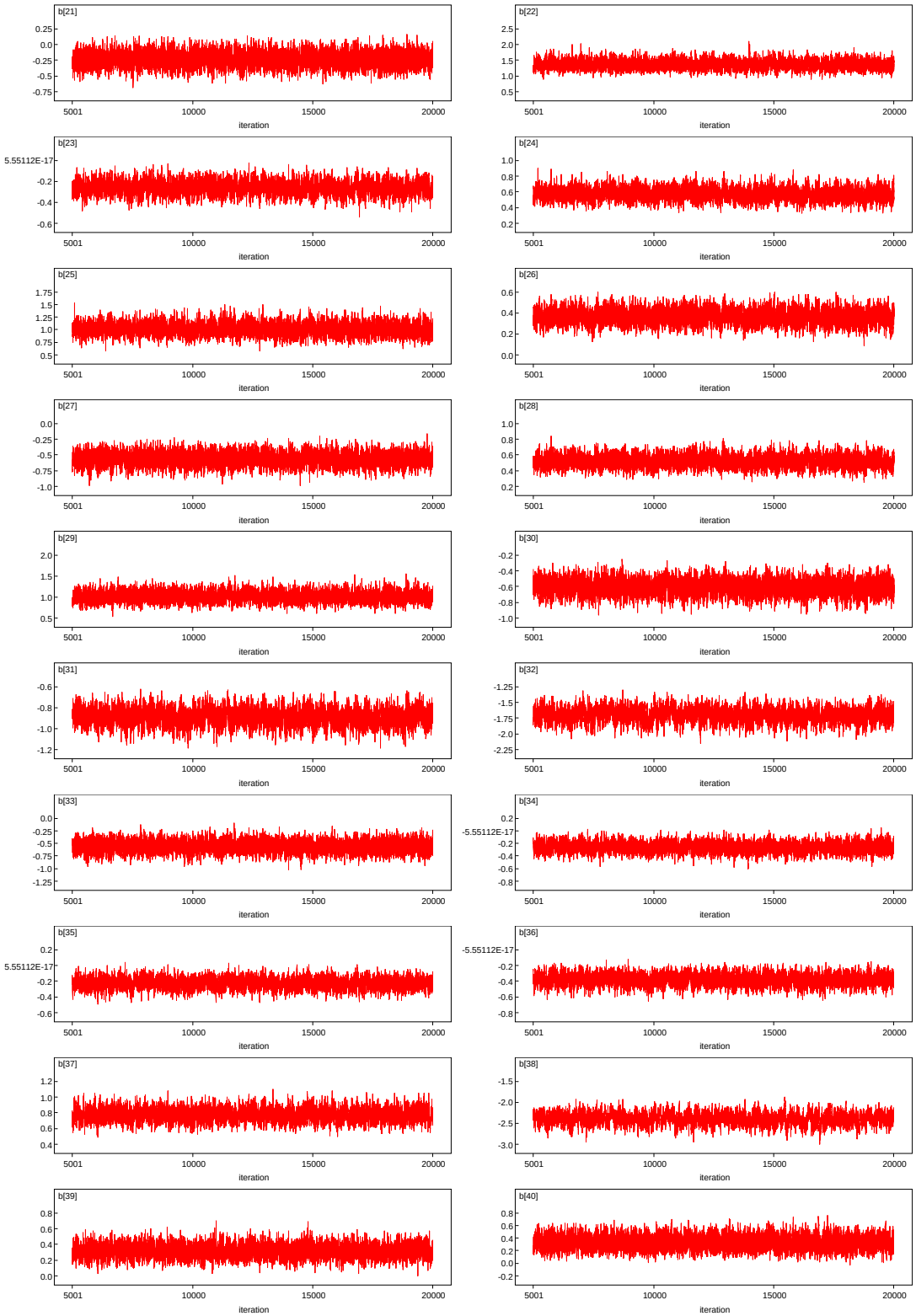
ภาคผนวก

ภาคผนวก ก

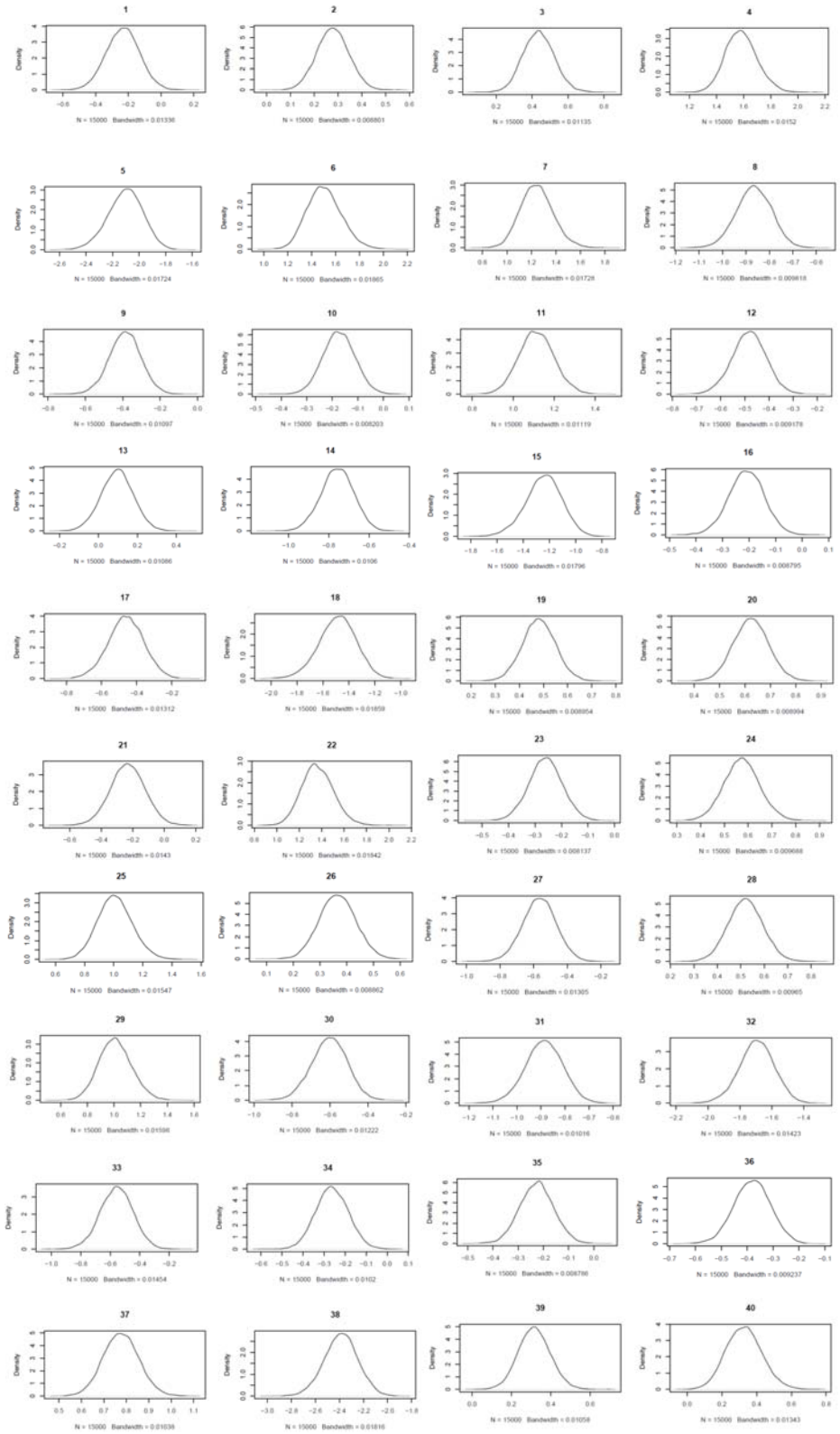
ตัวอย่าง History plot, density plot และ acf (Autocorrelation function) plot
ของพารามิเตอร์ความยาก ที่ประมาณค่าด้วยวิธี Bayes และวิธี Bayes$_{\gamma}$



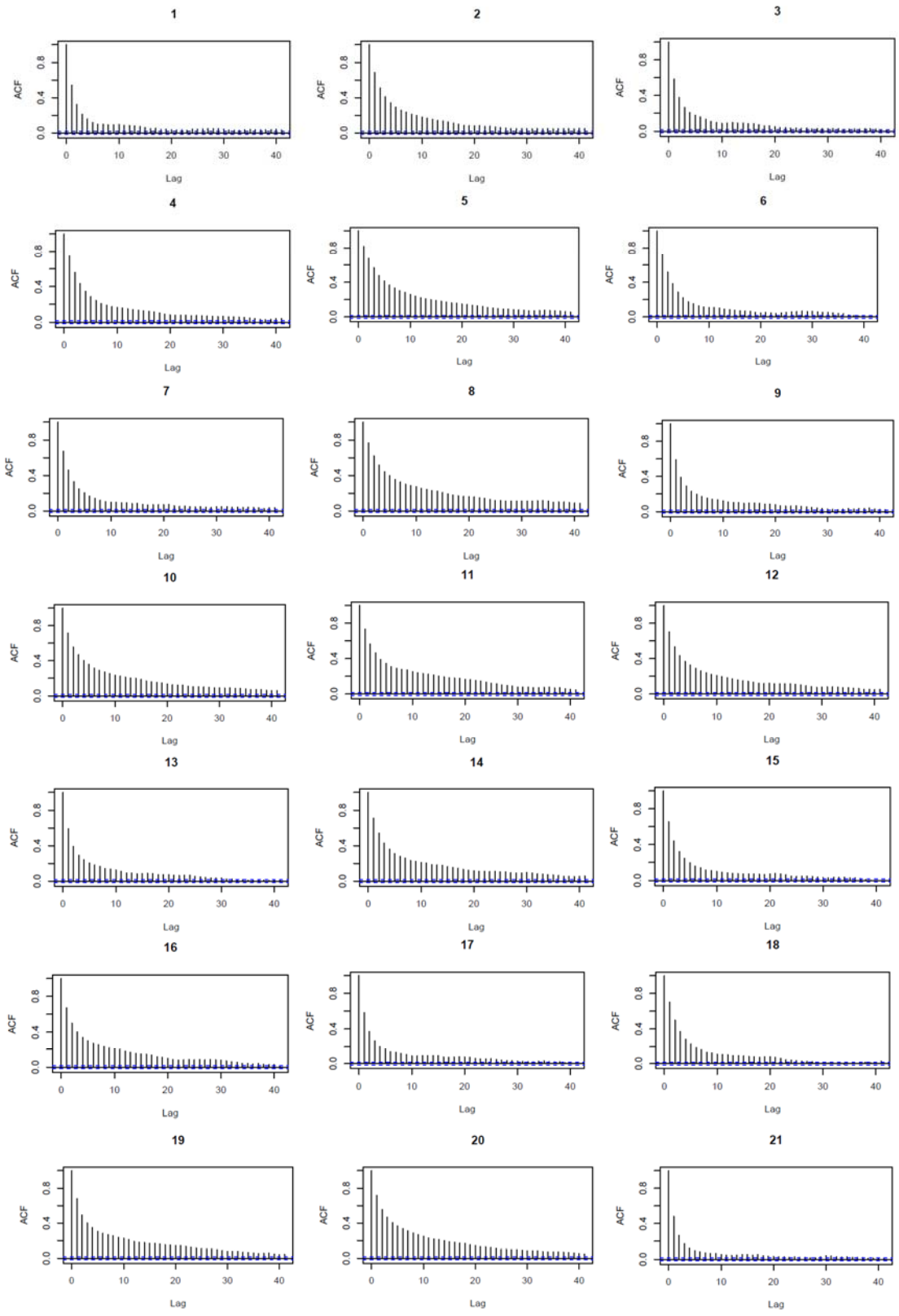
ภาพที่ ก - 1 ตัวอย่าง History plot ของพารามิเตอร์ความยากที่ประมาณค่าด้วยวิธี Bayesγ


b[21]
b[22]
b[23]
b[24]
b[25]
b[26]
b[27]
b[28]
b[29]
b[30]
b[31]
b[32]
b[33]
b[34]
b[35]
b[36]
b[37]
b[38]
b[39]
b[40]
iteration
5001
10000
15000
20000

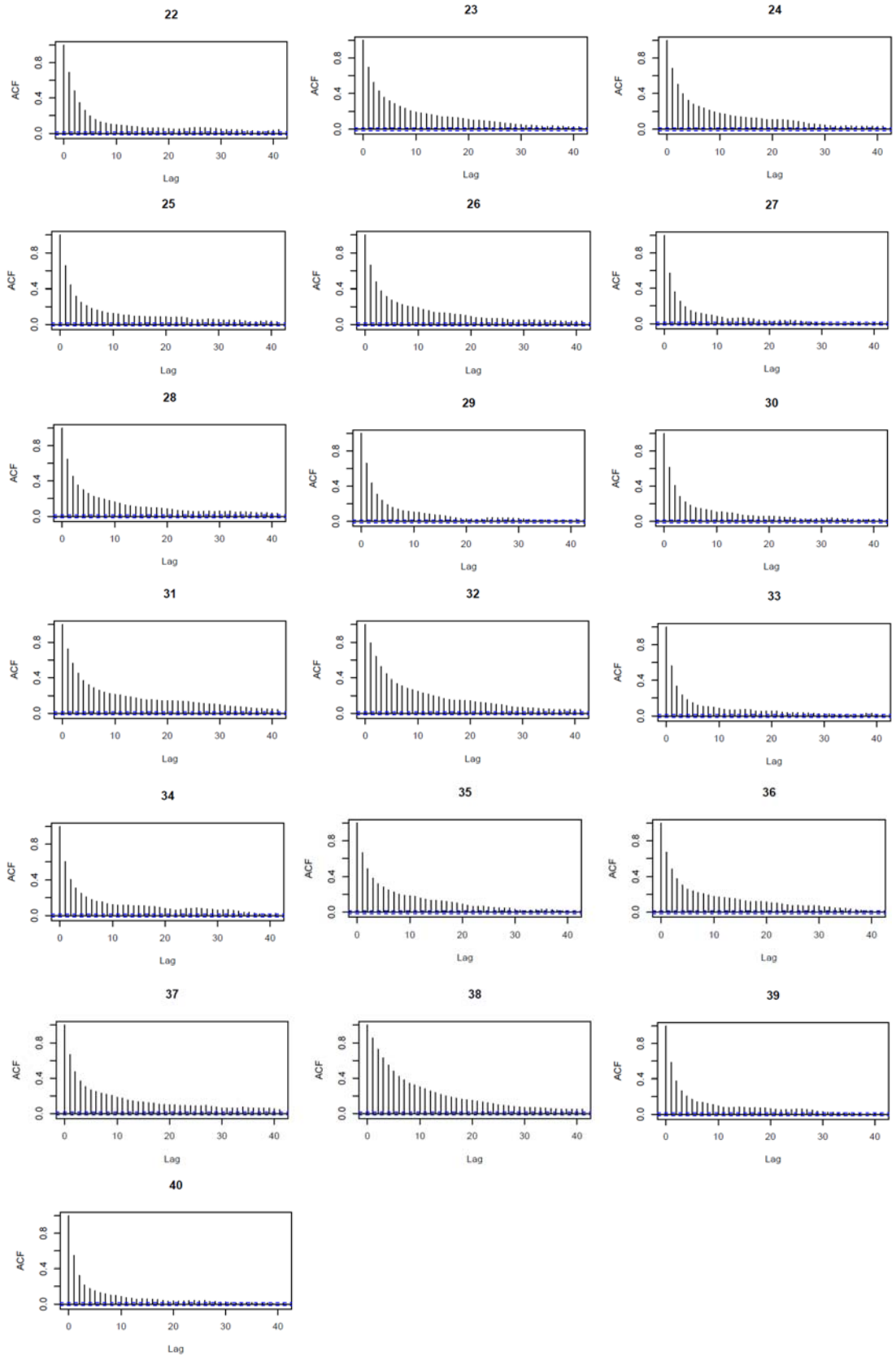
ภาพที่ ก - 1 (ต่อ)



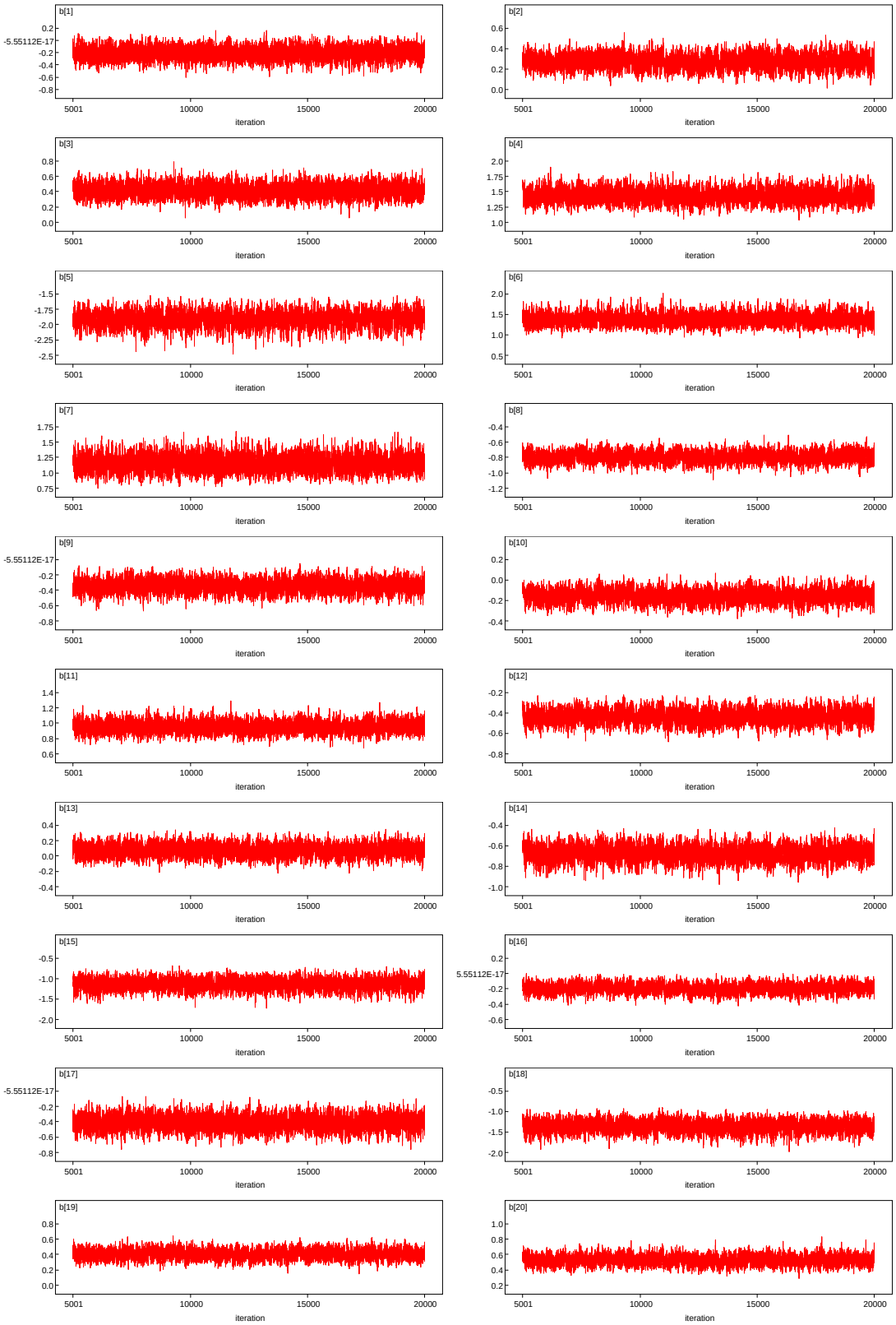
ภาพที่ ก - 2 ตัวอย่าง density plot ของพารามิเตอร์ความยากที่ประมาณค่าด้วยวิธี Bayesγ



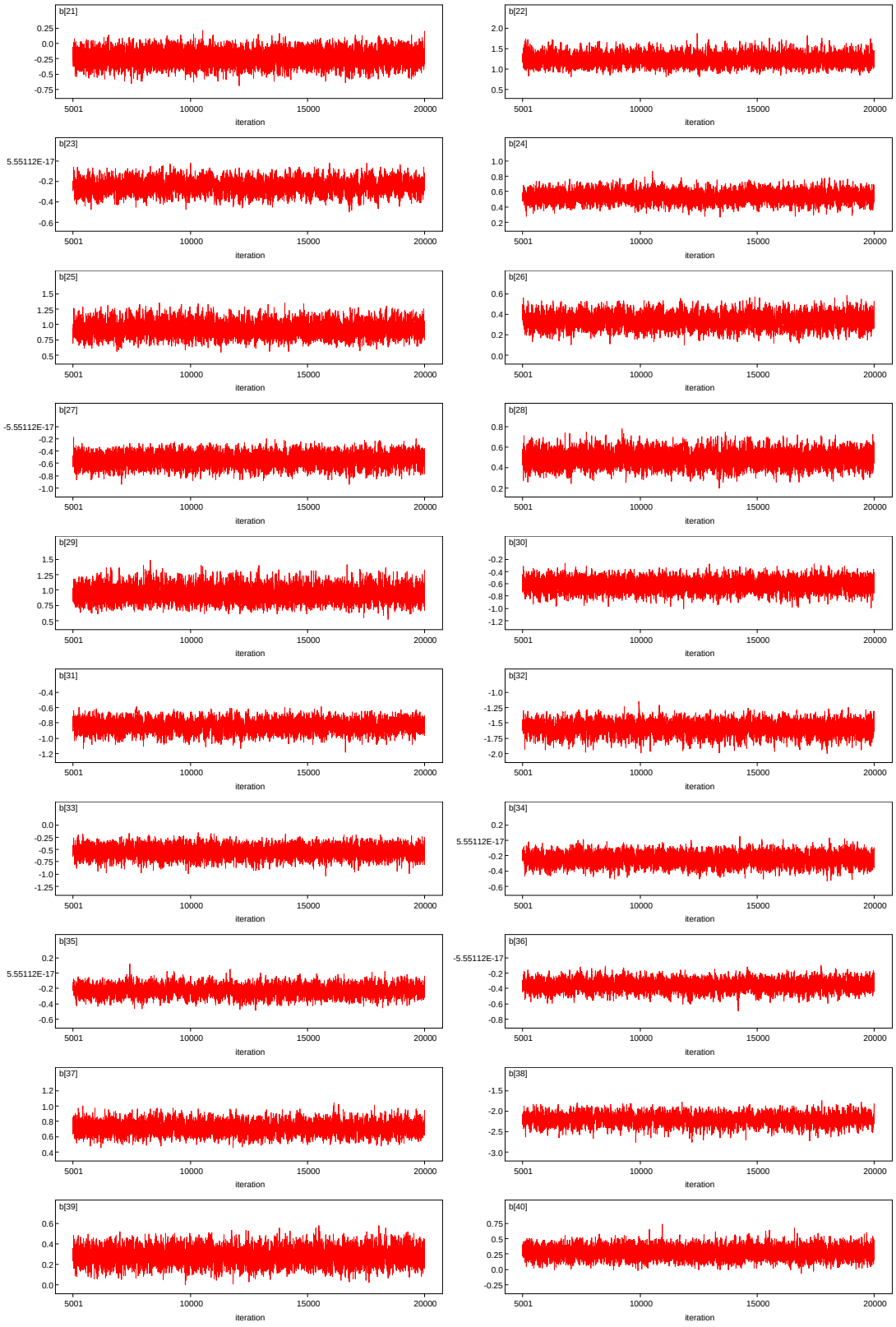
ภาพที่ ก - 3 ตัวอย่าง acf plot ของพารามิเตอร์ความยากที่ประมาณค่าด้วยวิธี Bayesγ



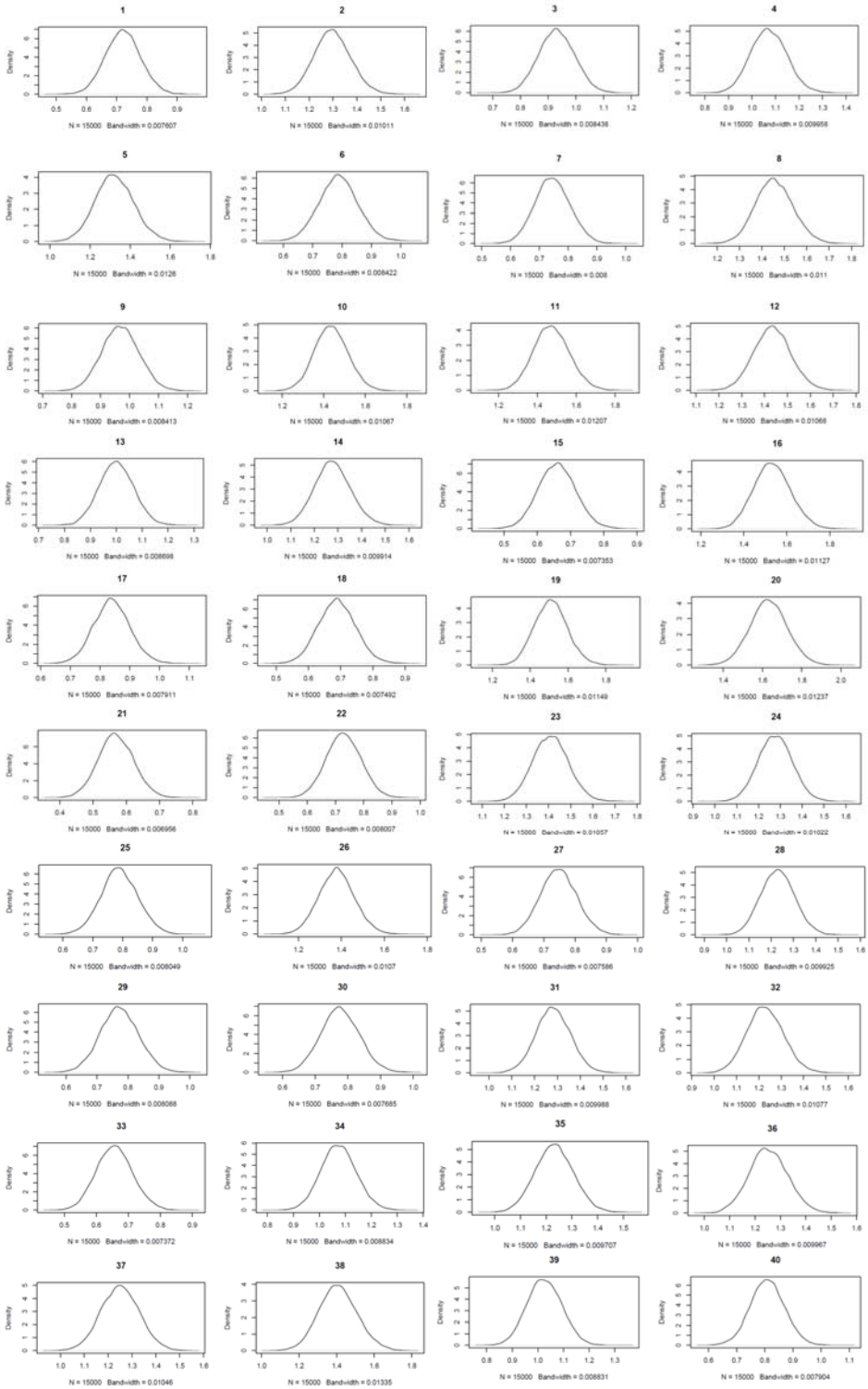
ภาพที่ ก - 3 (ต่อ)



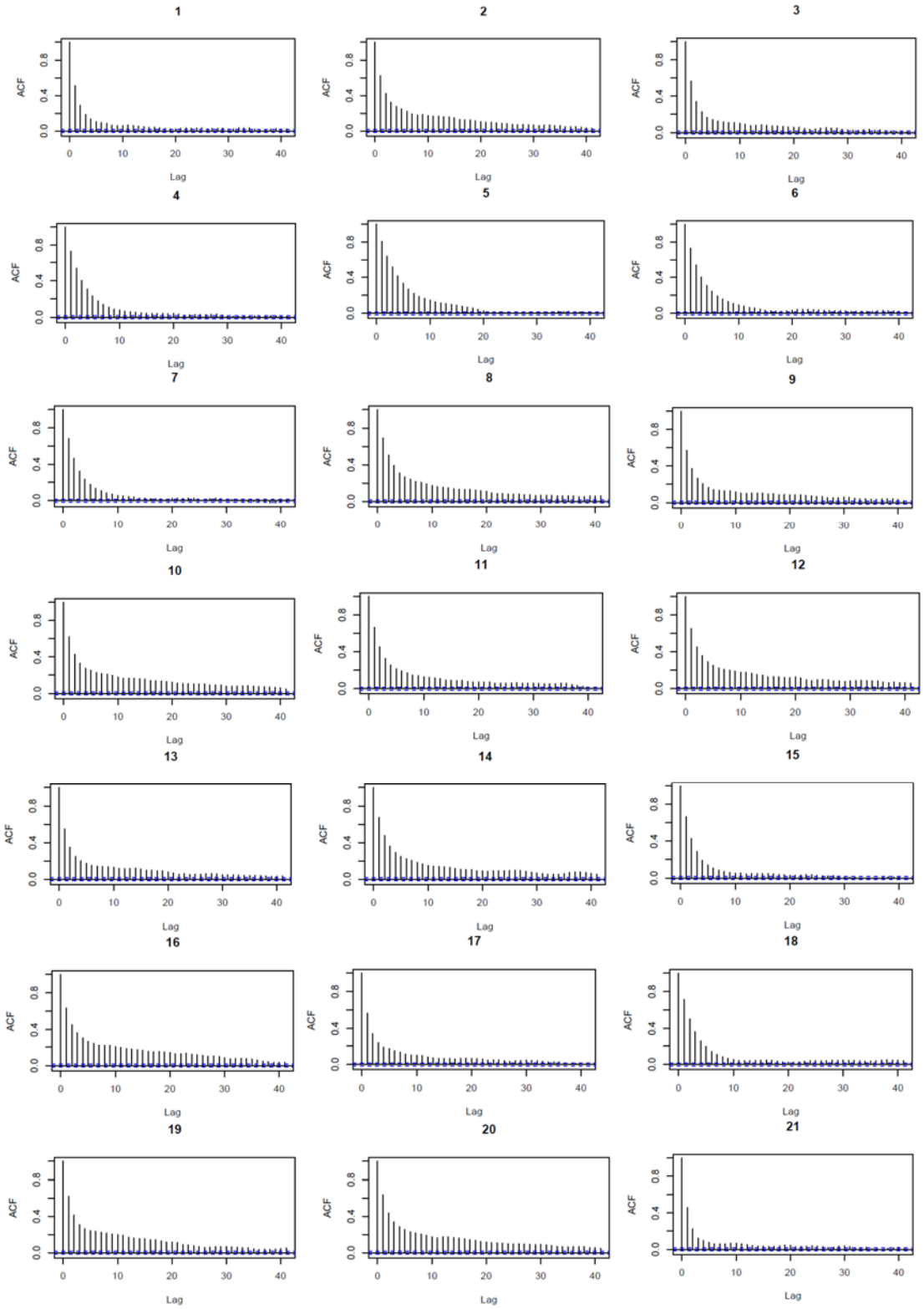
ภาพที่ ก - 4 ตัวอย่าง History plot ของพารามิเตอร์ความยากที่ประมาณค่าด้วยวิธี Bayes


b[21]
b[22]
b[23]
b[24]
b[25]
b[26]
b[27]
b[28]
b[29]
b[30]
b[31]
b[32]
b[33]
b[34]
b[35]
b[36]
b[37]
b[38]
b[39]
b[40]
iteration
5001
10000
15000
20000

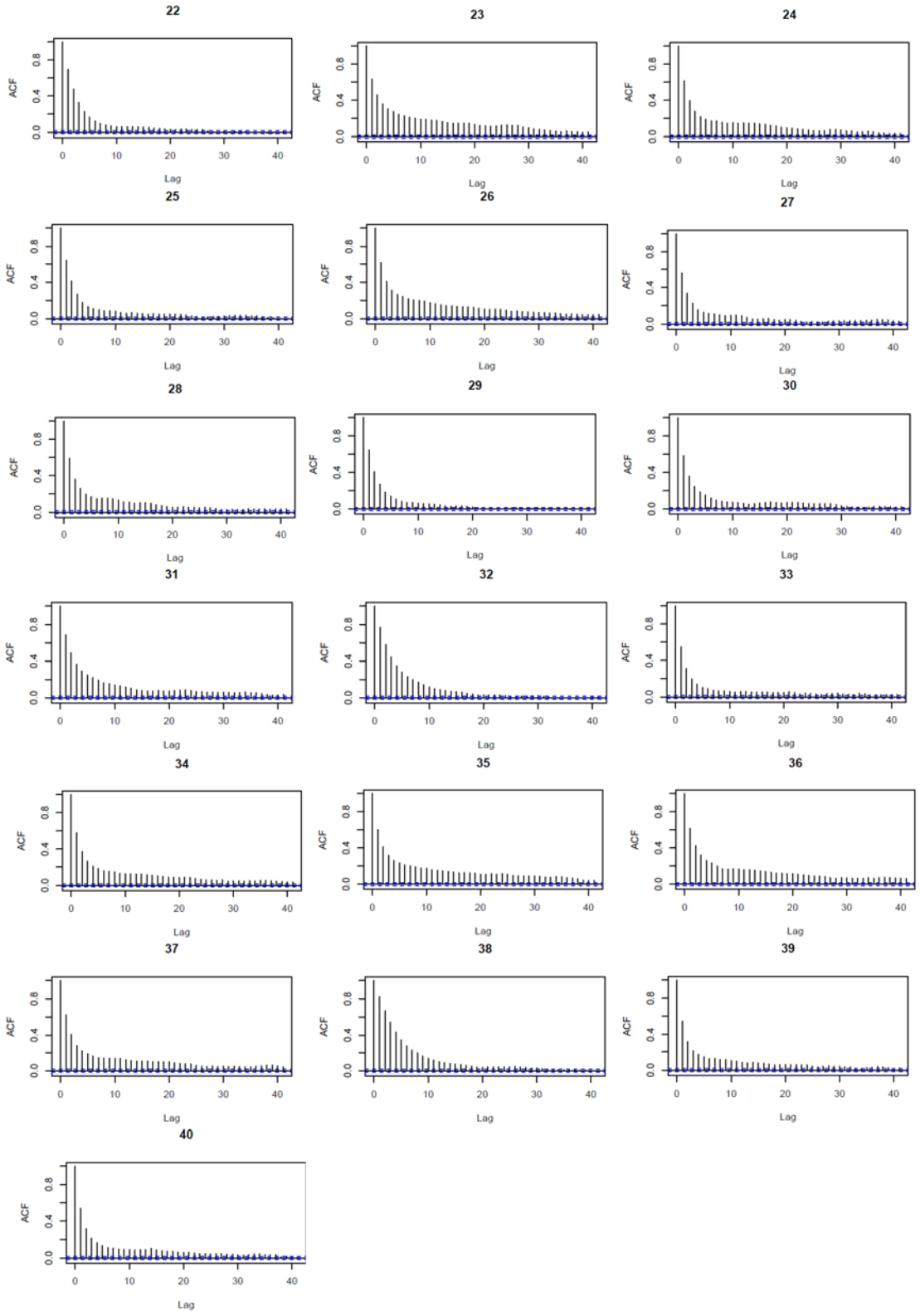
ภาพที่ ก - 4 (ต่อ)



ภาพที่ ก - 5 ตัวอย่าง density plot ของพารามิเตอร์ความยากที่ประมาณค่าด้วยวิธี Bayes



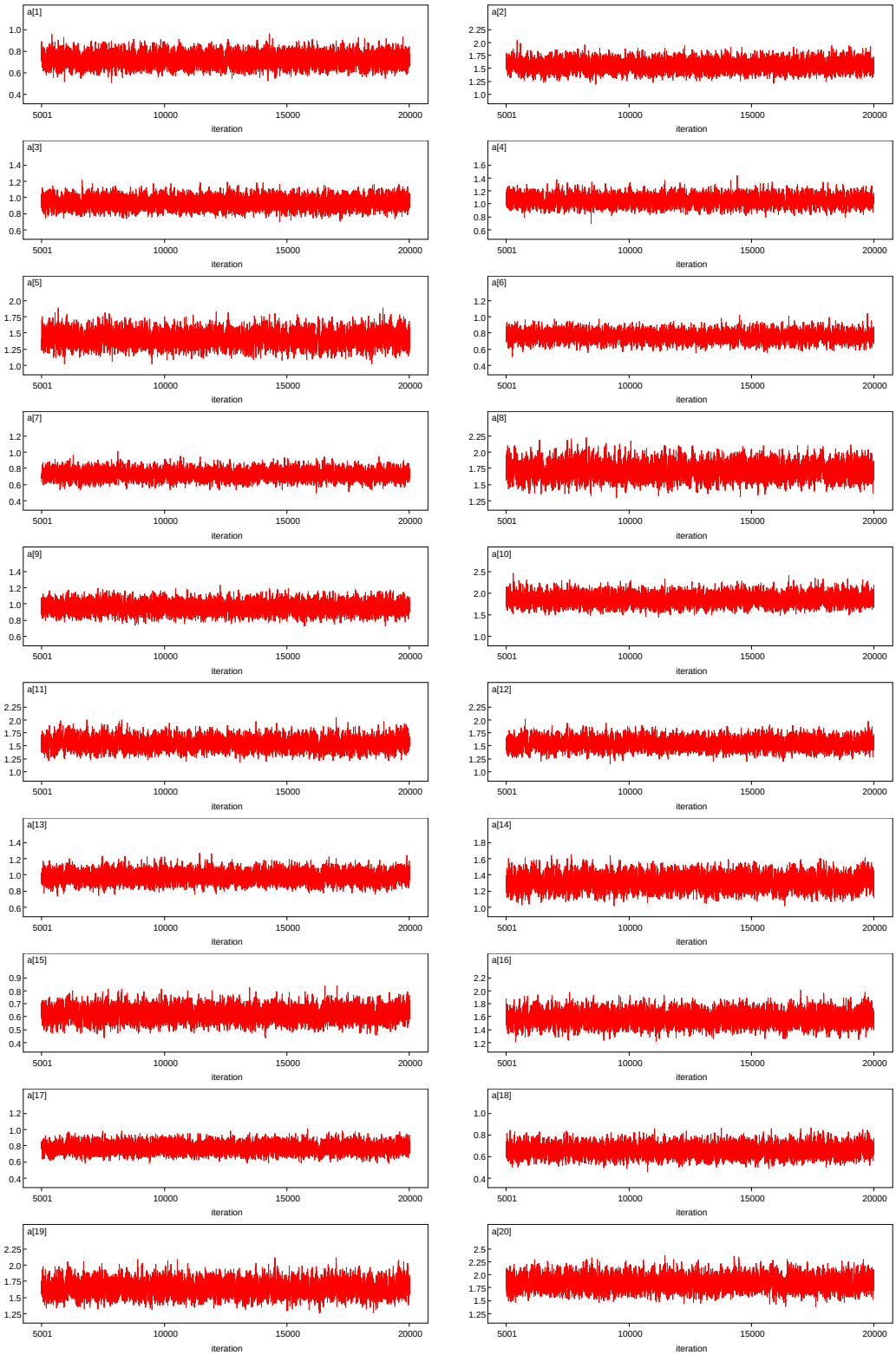
ภาพที่ ก - 6 ตัวอย่าง acf plot ของพารามิเตอร์ความยากที่ประมาณค่าด้วยวิธี Bayes



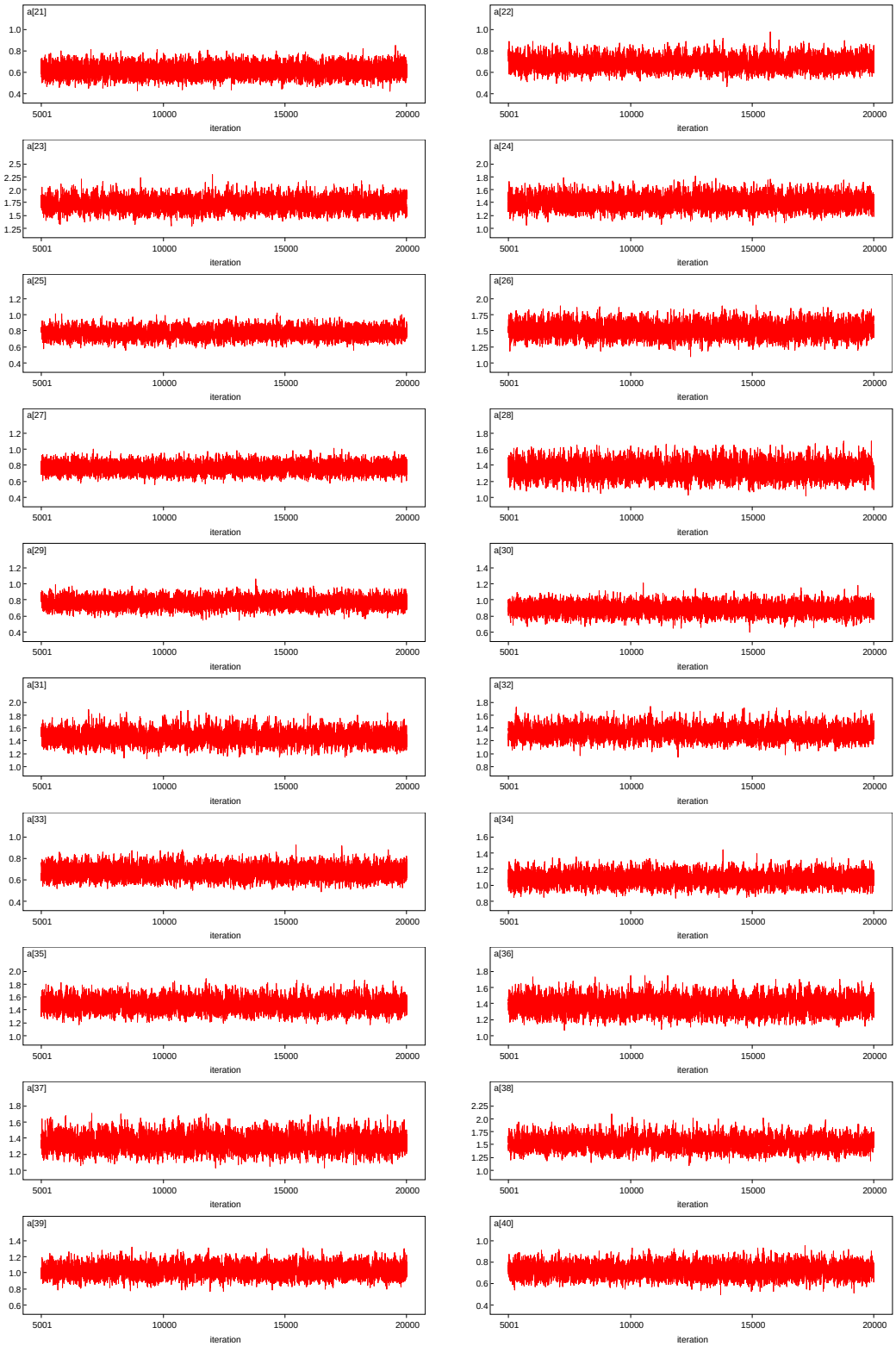
ภาพที่ ก - 6 (ต่อ)

ภาคผนวก ข

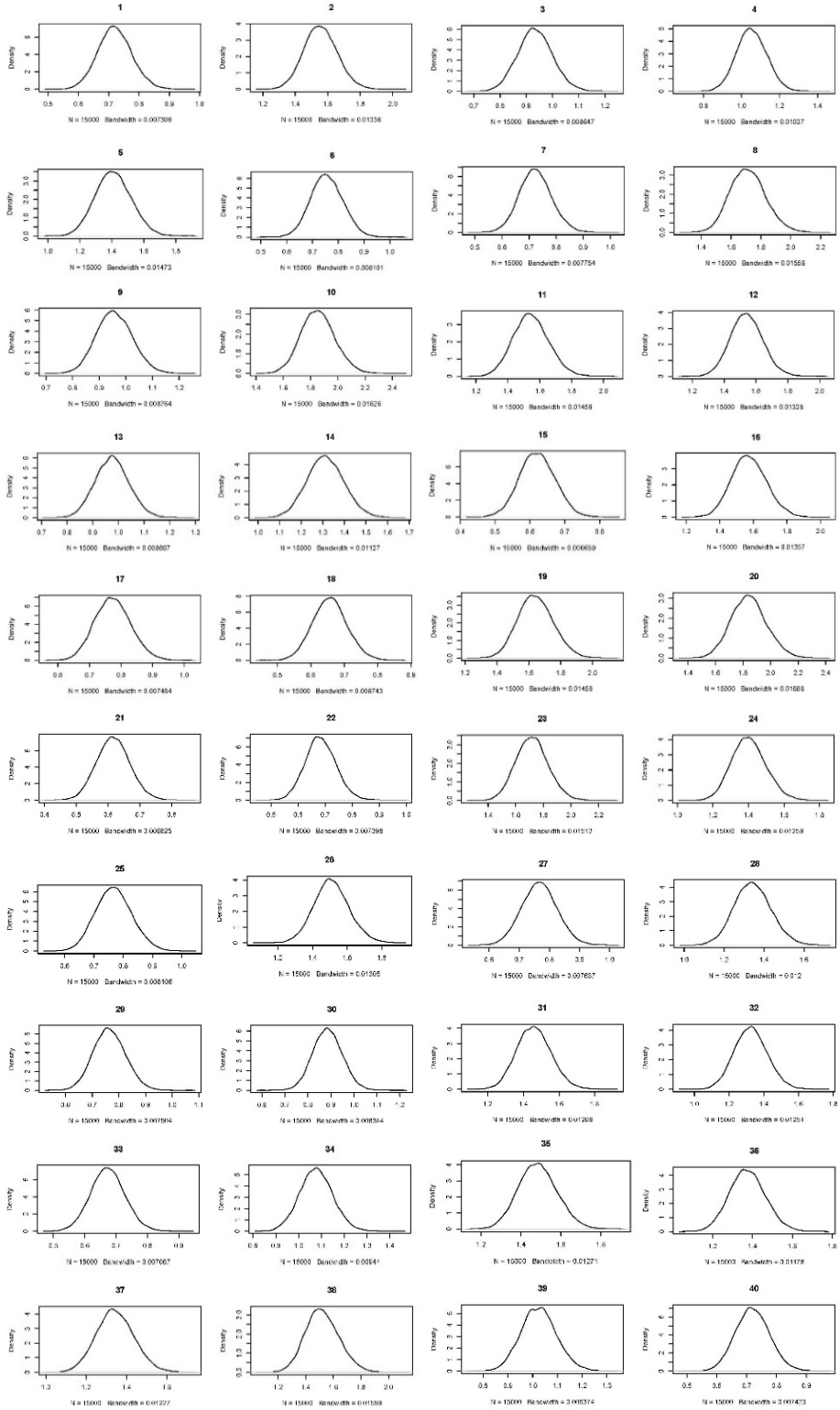
ตัวอย่าง History plot, density plot และ acf (Autocorrelation function) plot ของพารามิเตอร์อำนาจจำแนกประมาณค่าด้วยวิธี Bayes และวิธี Bayes$_{\gamma}$



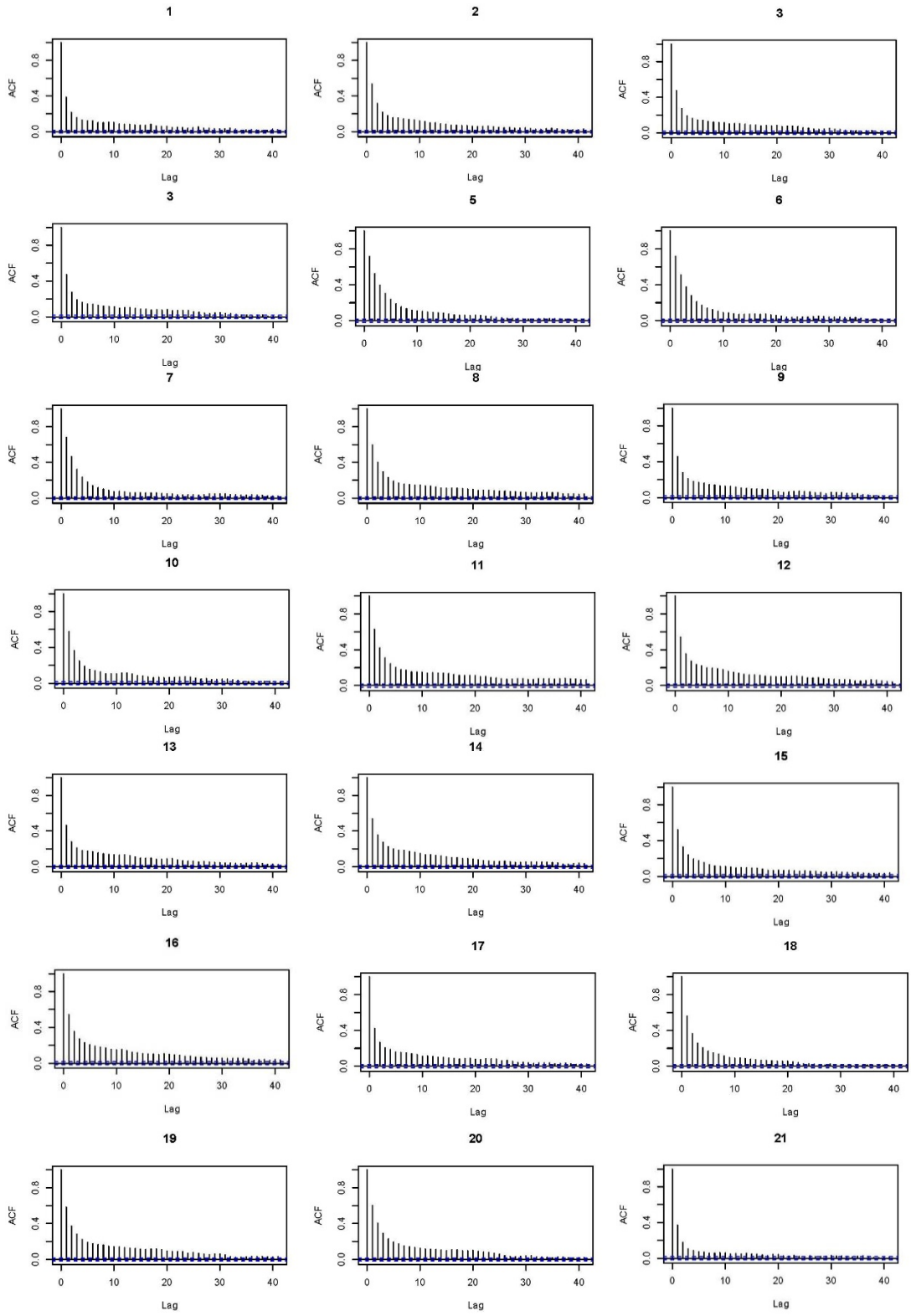
ภาพที่ ข - 1 ตัวอย่าง History plot ของพารามิเตอร์อำนาจจำแนกที่ประมาณค่าด้วยวิธี Bayesγ


a[21]
a[22]
a[23]
a[24]
a[25]
a[26]
a[27]
a[28]
a[29]
a[30]
a[31]
a[32]
a[33]
a[34]
a[35]
a[36]
a[37]
a[38]
a[39]
a[40]
iteration
5001
10000
15000
20000

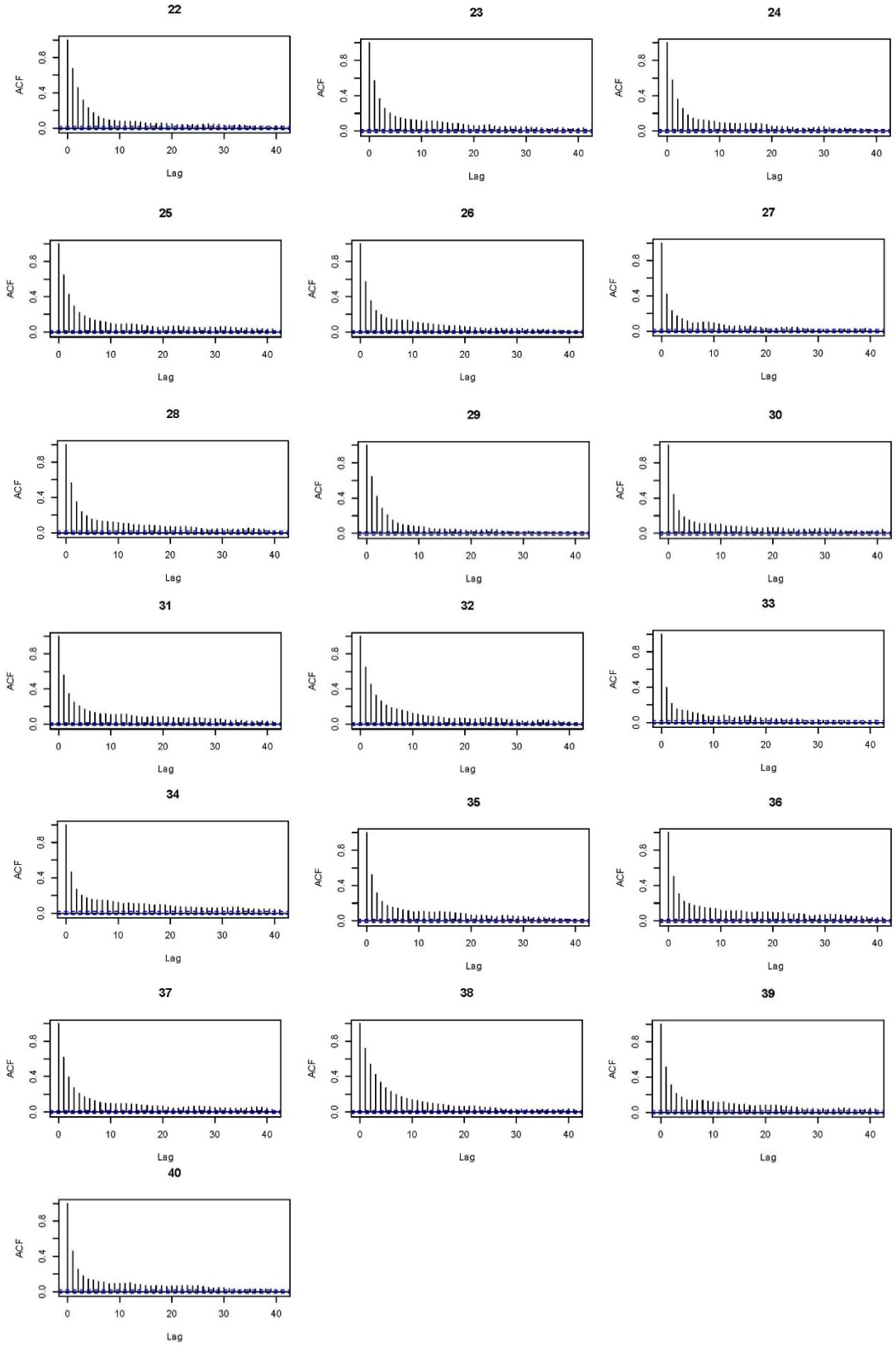
ภาพที่ ข - 1 (ต่อ)



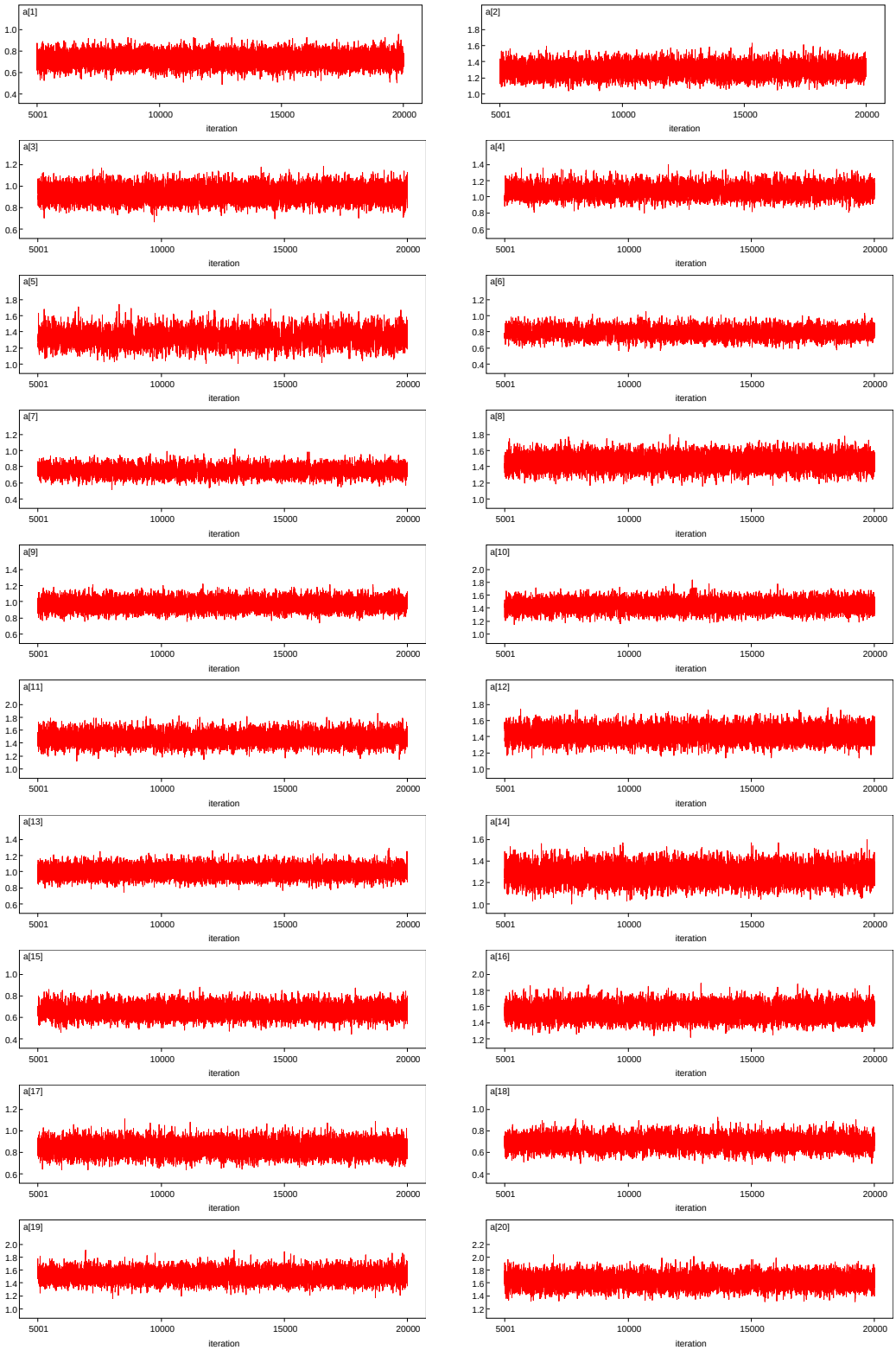
ภาพที่ ข - 2 ตัวอย่าง density plot ของพารามิเตอร์อำนาจจำแนกที่ประมาณค่าด้วยวิธี Bayes$_{\gamma}$



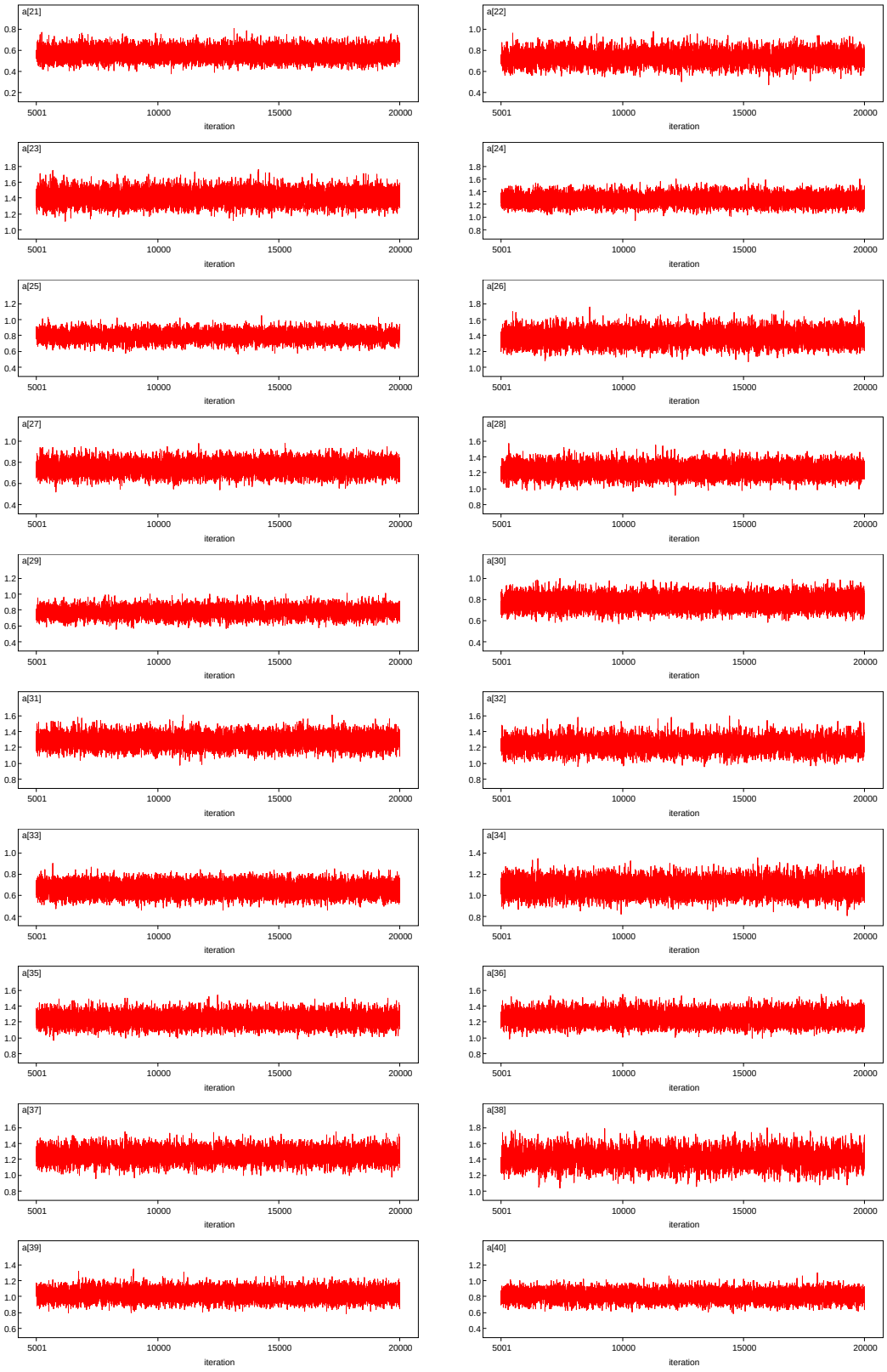
ภาพที่ ข - 3 ตัวอย่าง acf plot ของพารามิเตอร์อำนาจจำแนกที่ประมาณค่าด้วยวิธี Bayesγ



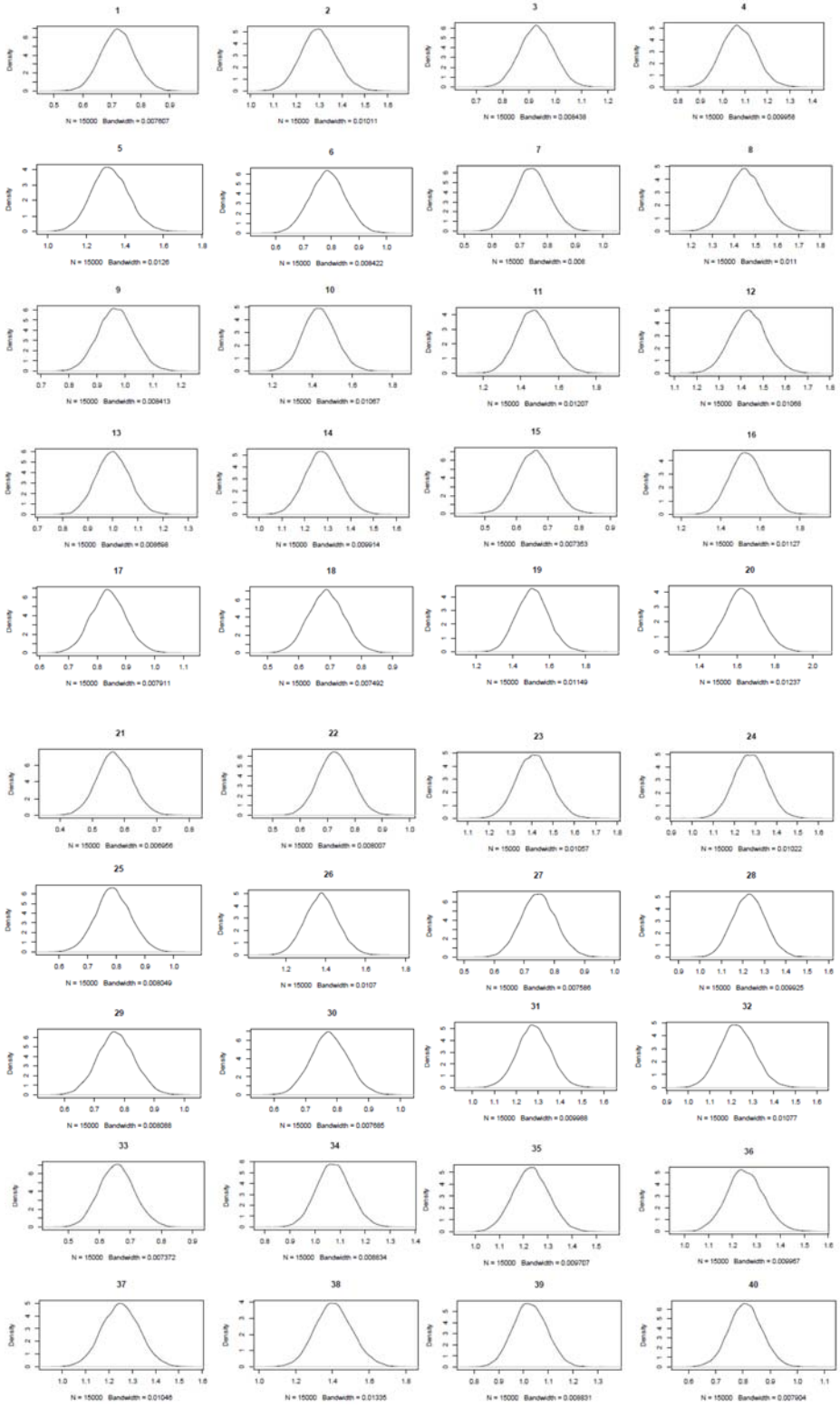
ภาพที่ ข - 3 (ต่อ)



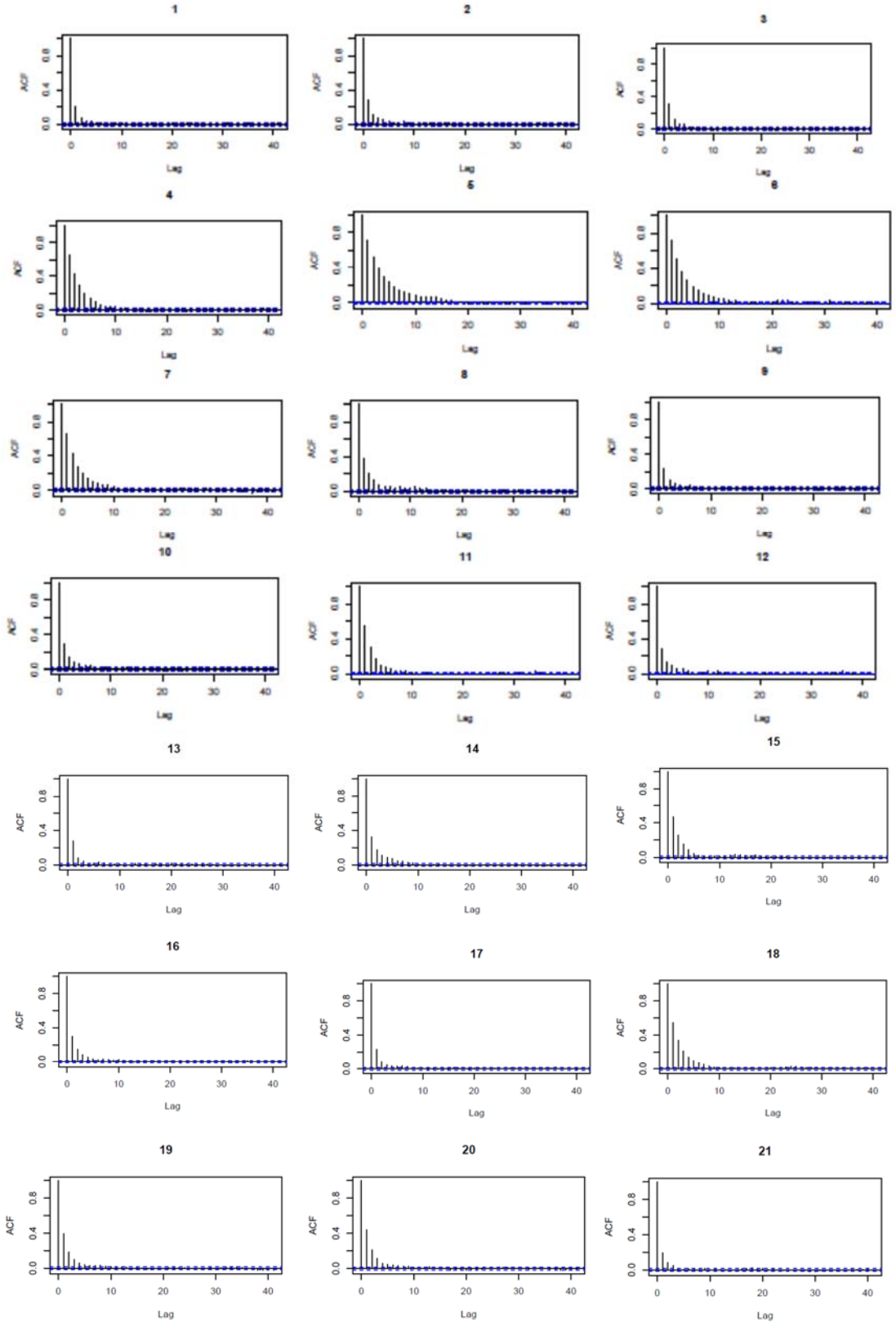
ภาพที่ ข - 4 ตัวอย่าง History plot ของพารามิเตอร์อำนาจจำแนกที่ประมาณค่าด้วยวิธี Bayes


a[21]
a[22]
a[23]
a[24]
a[25]
a[26]
a[27]
a[28]
a[29]
a[30]
a[31]
a[32]
a[33]
a[34]
a[35]
a[36]
a[37]
a[38]
a[39]
a[40]
iteration
5001
10000
15000
20000

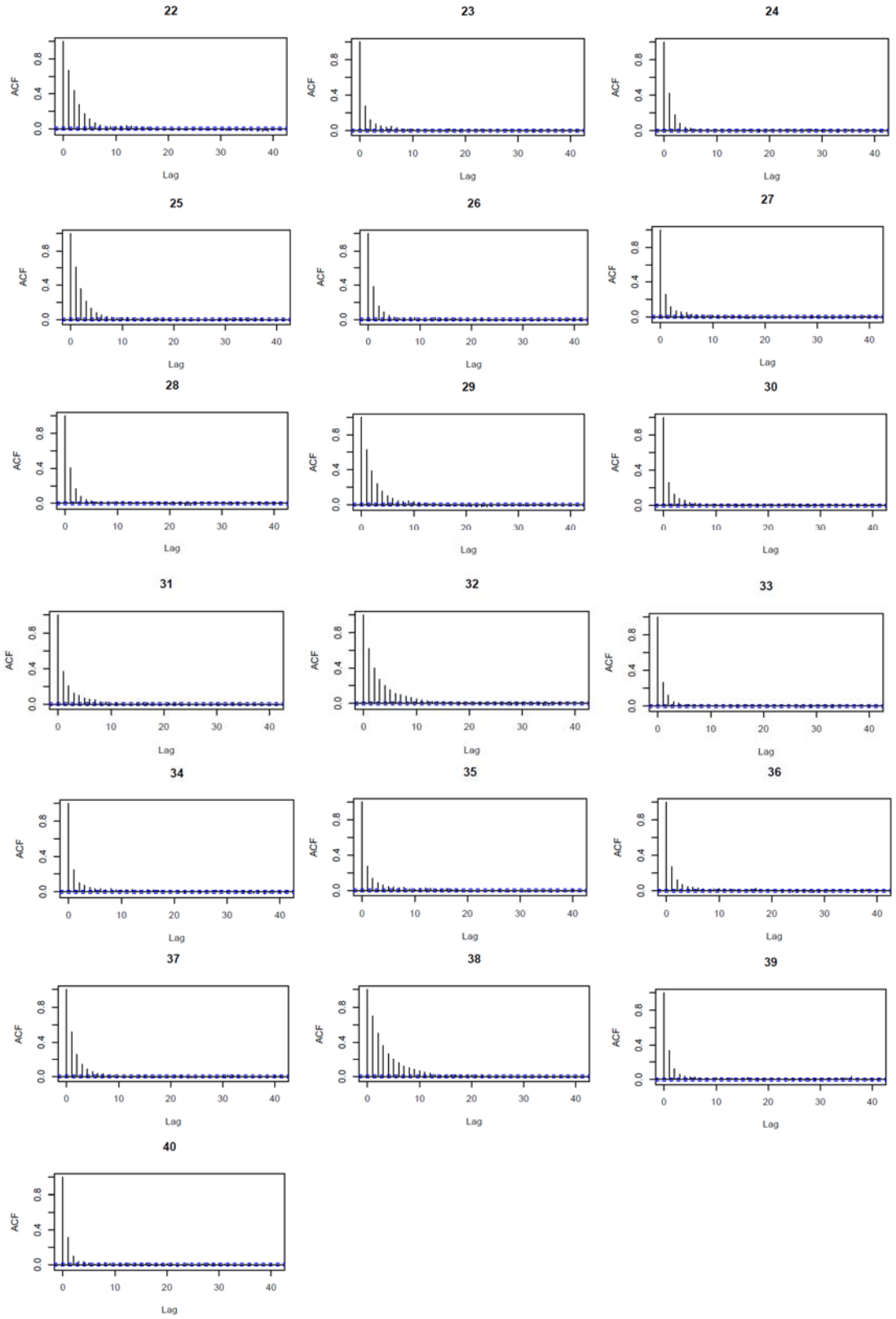
ภาพที่ ข - 4 (ต่อ)



ภาพที่ ข - 5 ตัวอย่าง density plot ของพารามิเตอร์อำนาจจำแนกที่ประมาณค่าด้วยวิธี Bayes



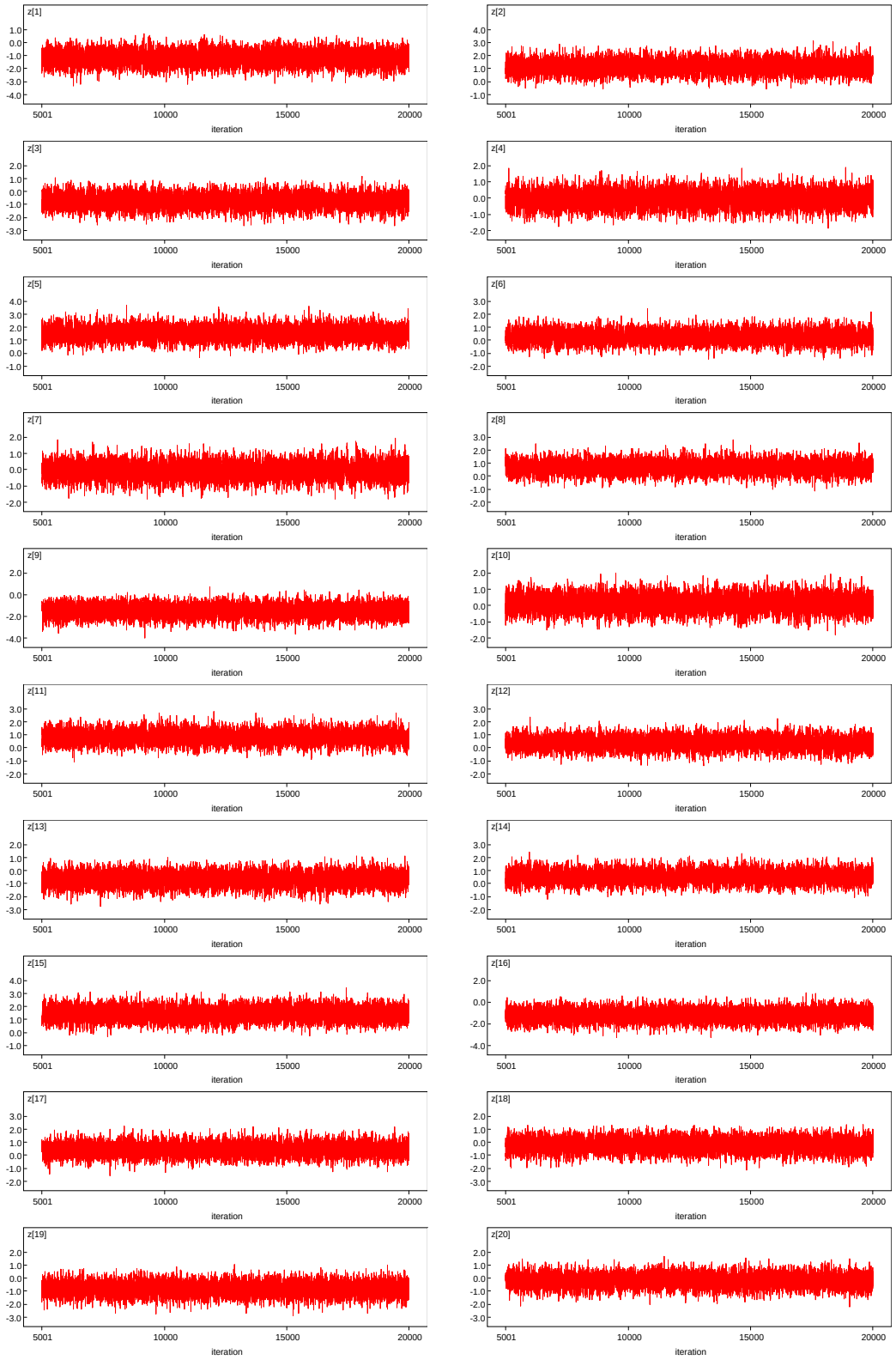
ภาพที่ ข - 6 ตัวอย่าง acf plot ของพารามิเตอร์อำนาจจำแนกที่ประมาณค่าด้วยวิธี Bayes



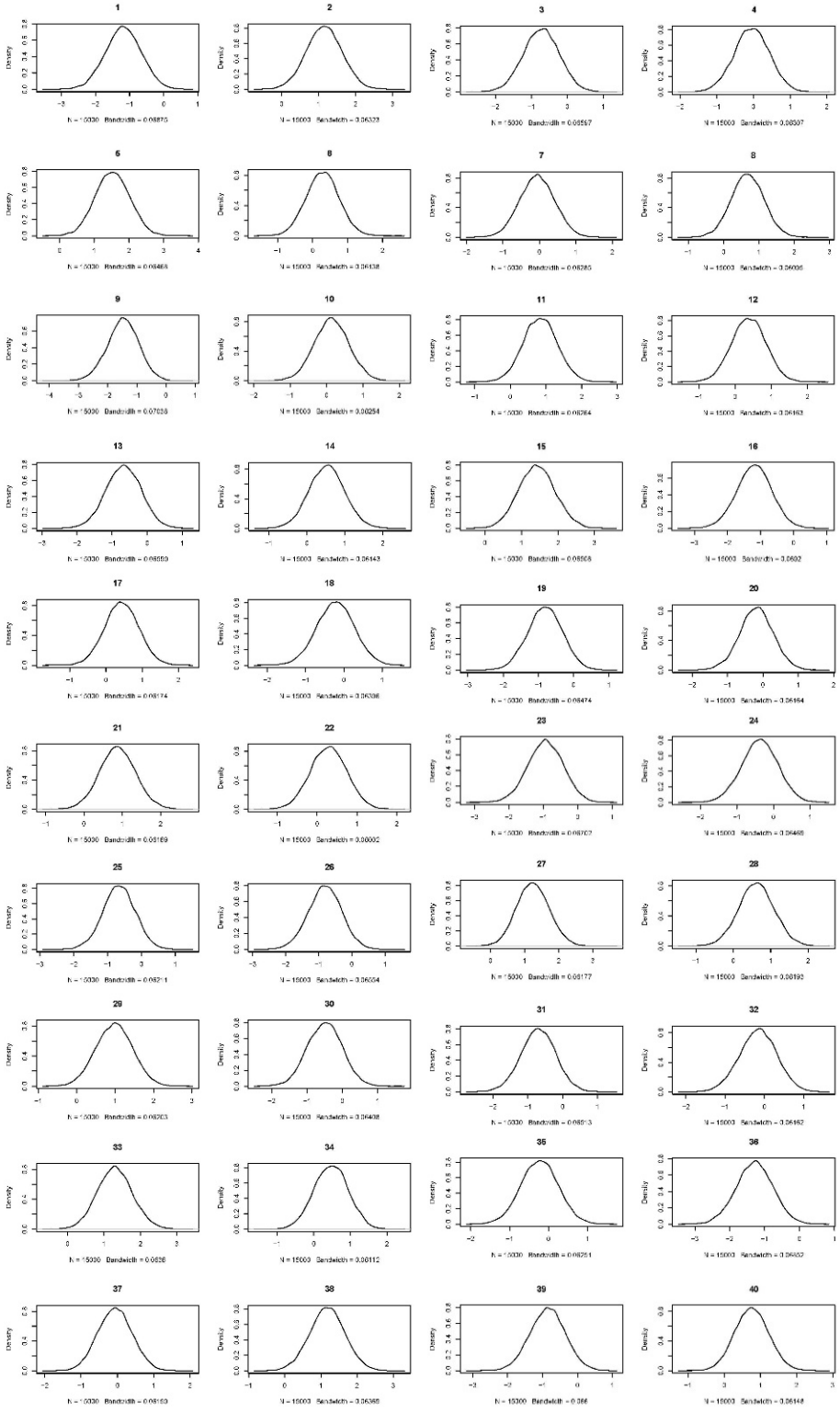
ภาพที่ ข - 6 (ต่อ)

ภาคผนวก ค

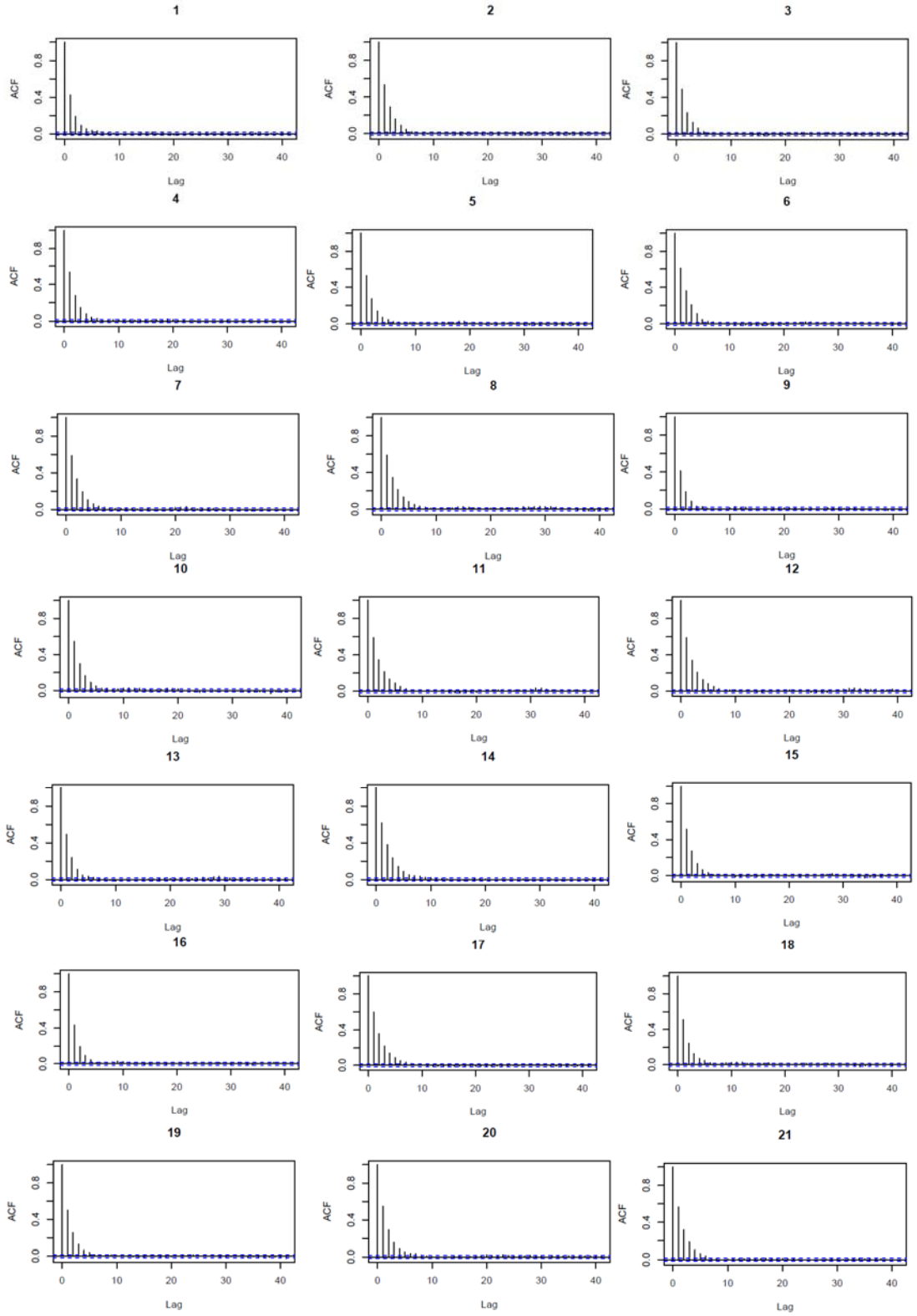
ตัวอย่าง History plot, density plot และ acf (autocorrelation function) plot ของพารามิเตอร์ความสามารถ ที่ประมาณค่าด้วยวิธี Bayes และวิธี Bayes$_\gamma$



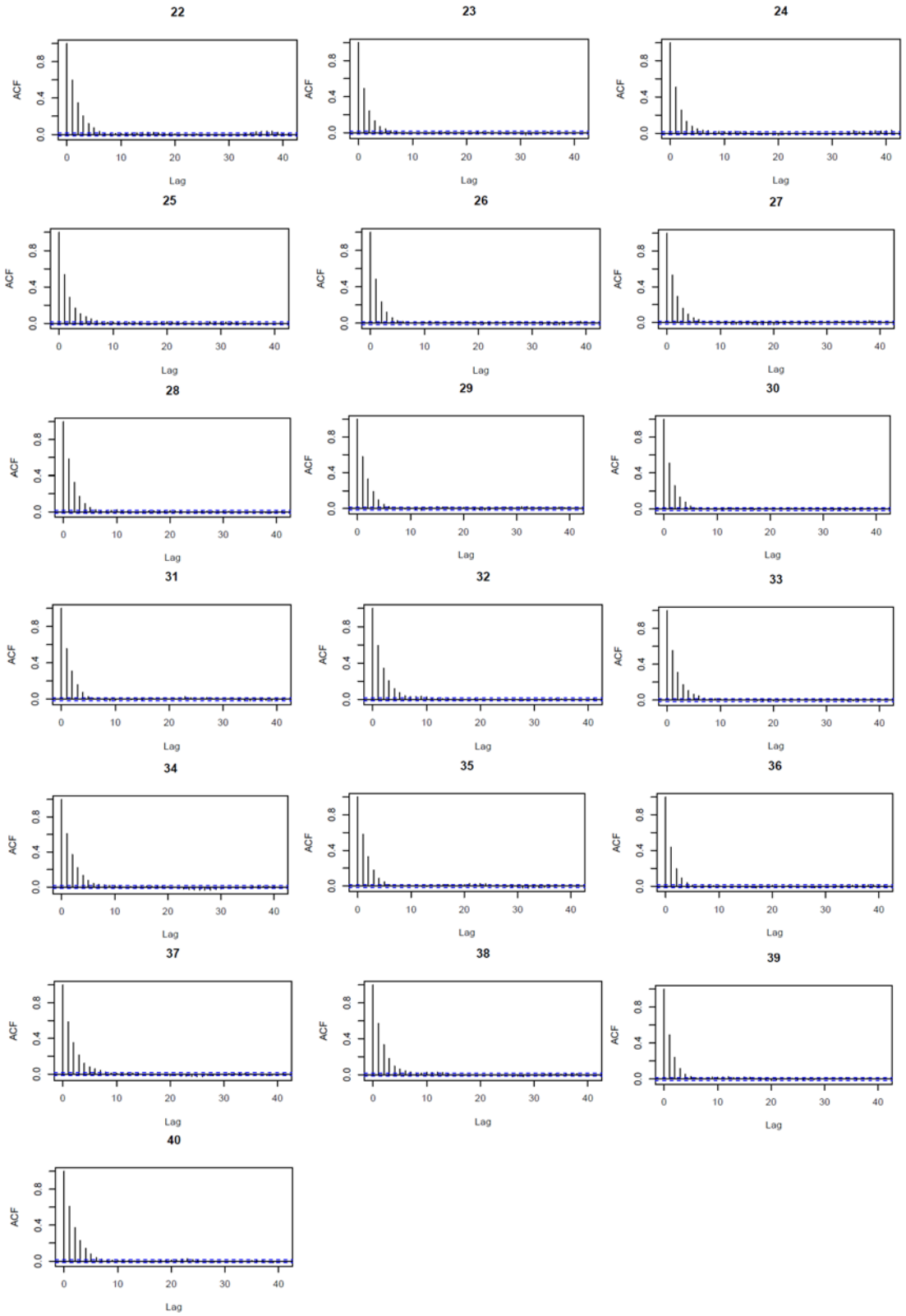
ภาพที่ ค - 1 ตัวอย่าง History plot ของพารามิเตอร์ความสามารถที่ประมาณค่าด้วยวิธี Bayesγ



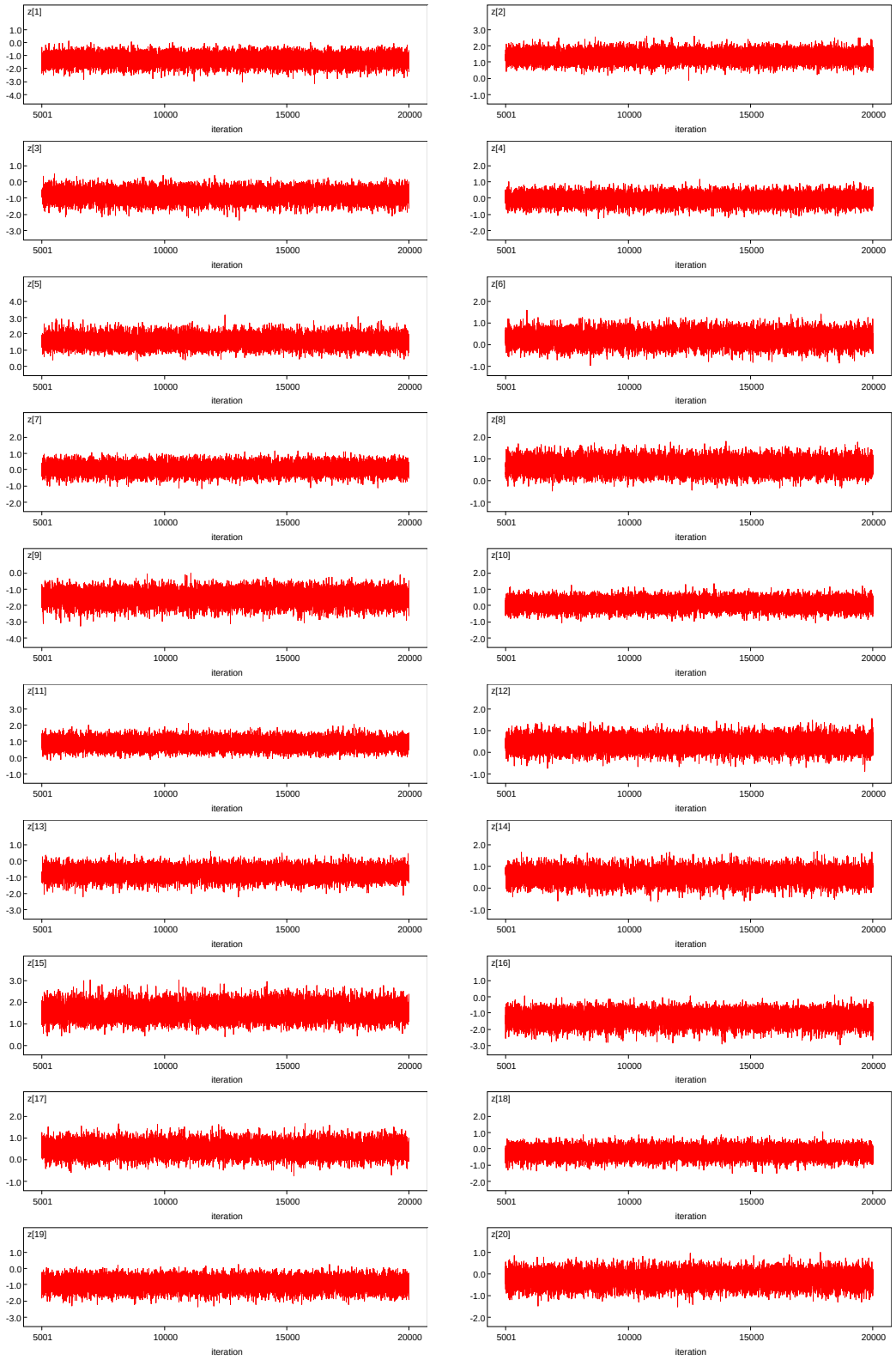
ภาพที่ ค - 2 ตัวอย่าง density plot ของพารามิเตอร์ความสามารถที่ประมาณค่าด้วยวิธี Bayes$_{\gamma}$



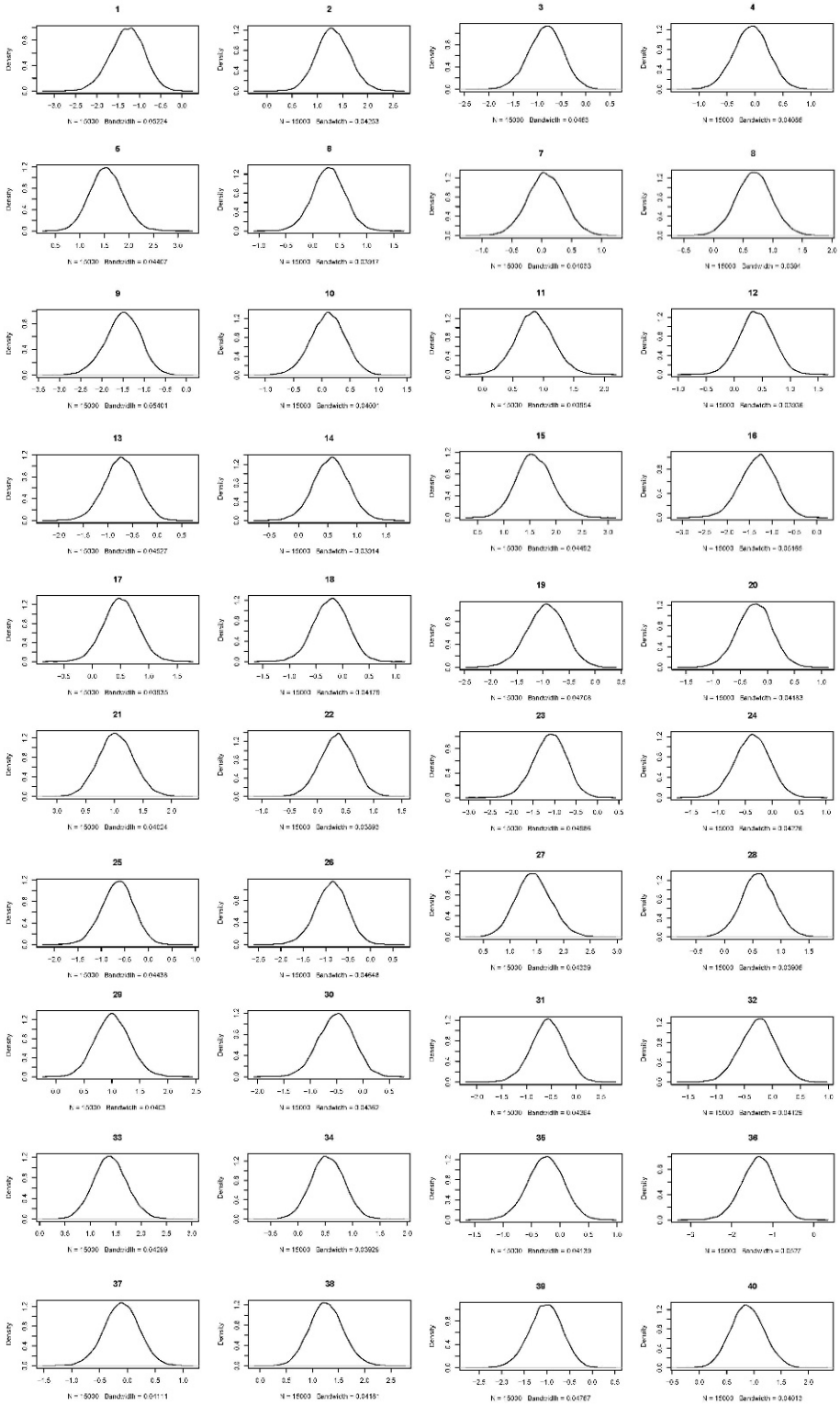
ภาพที่ ค - 3 ตัวอย่าง acf plot ของพารามิเตอร์ความสามารถที่ประมาณค่าด้วยวิธี Bayesγ



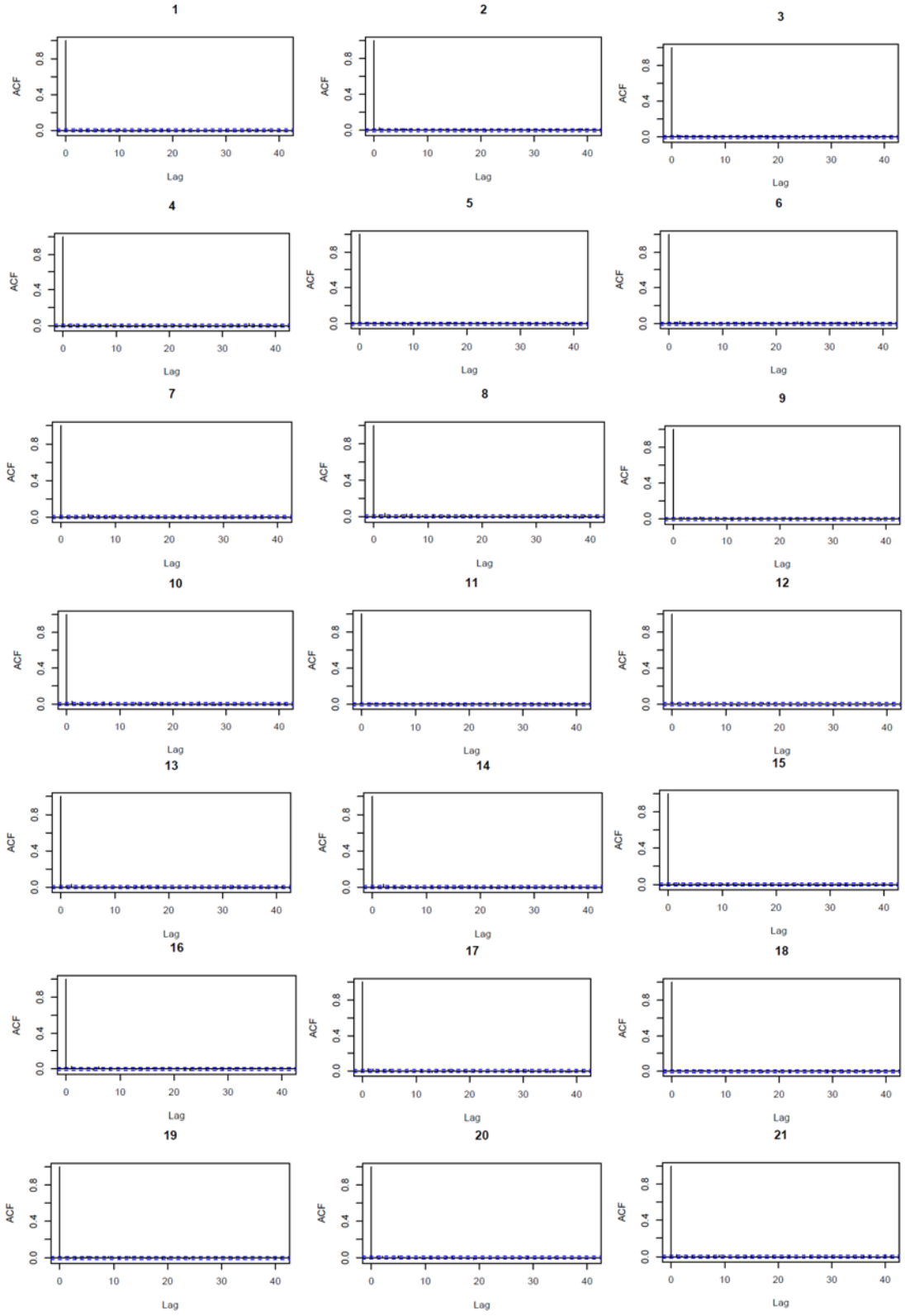
ภาพที่ ค - 3 (ต่อ)



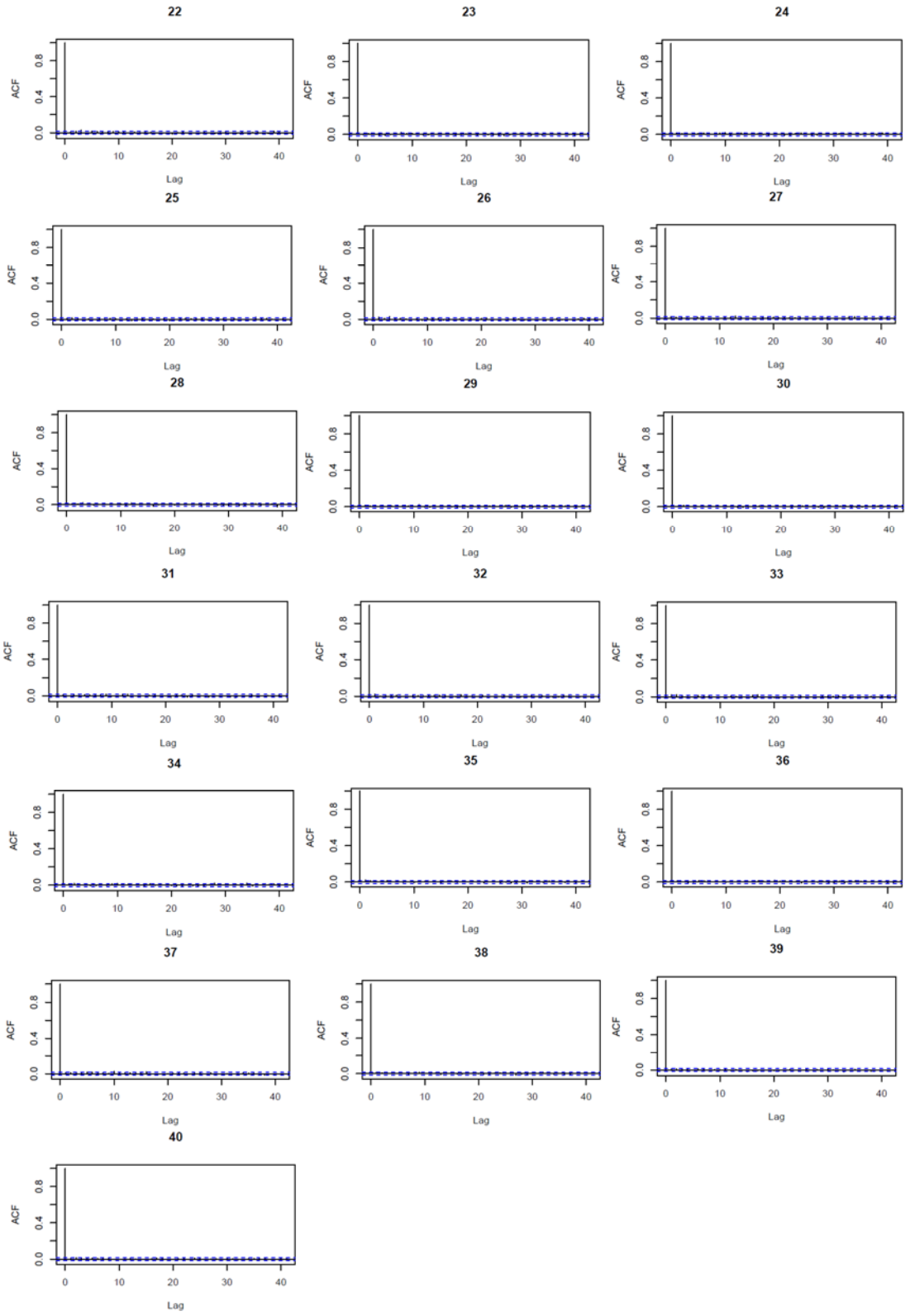
ภาพที่ ค - 4 ตัวอย่าง History plot ของพารามิเตอร์ความสามารถที่ประมาณค่าด้วยวิธี Bayes



ภาพที่ ค - 5 ตัวอย่าง density plot ของพารามิเตอร์ความสามารถที่ประมาณค่าด้วยวิธี Bayes



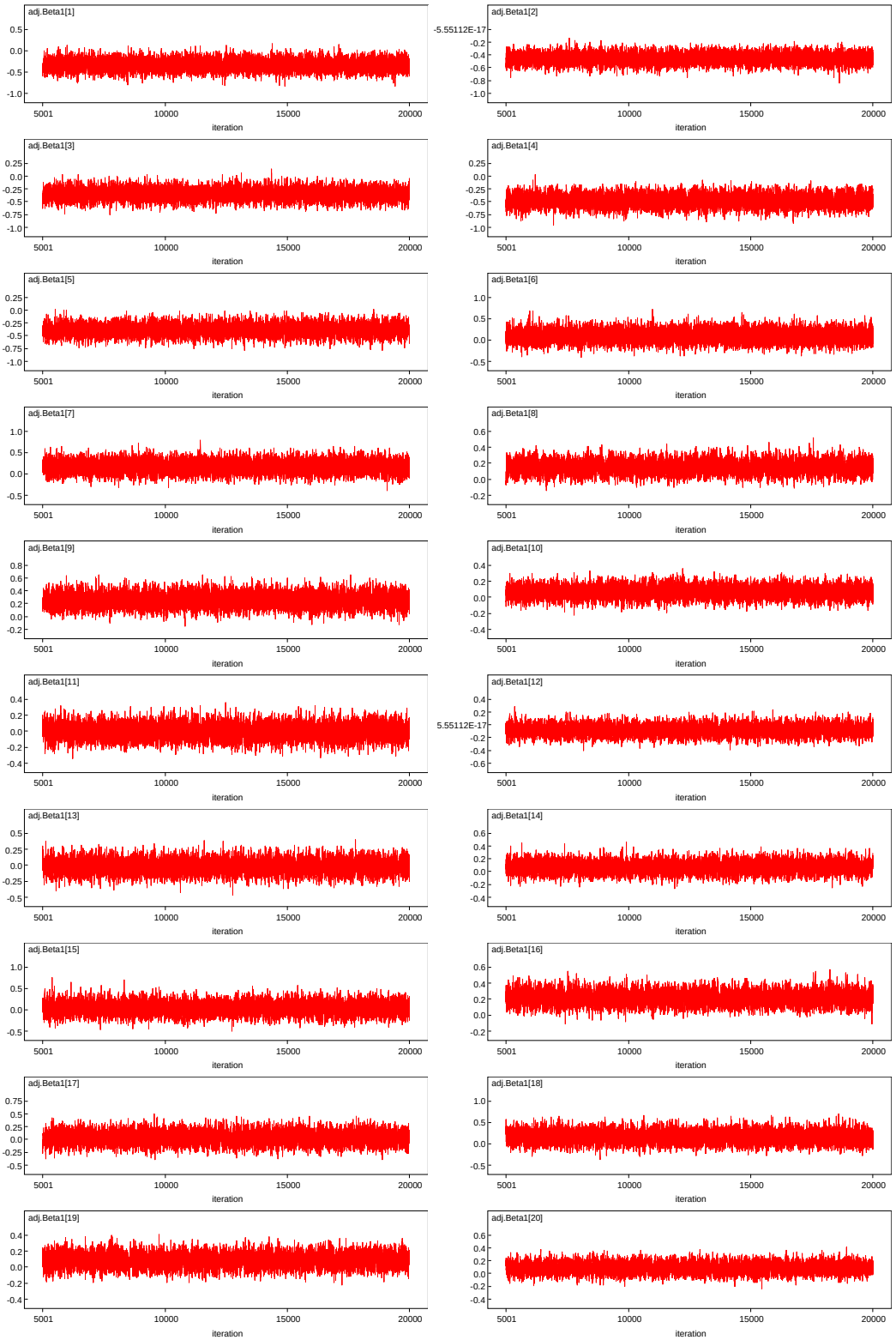
ภาพที่ ค - 6 ตัวอย่าง acf plot ของพารามิเตอร์ความสามารถที่ประมาณค่าด้วยวิธี Bayes



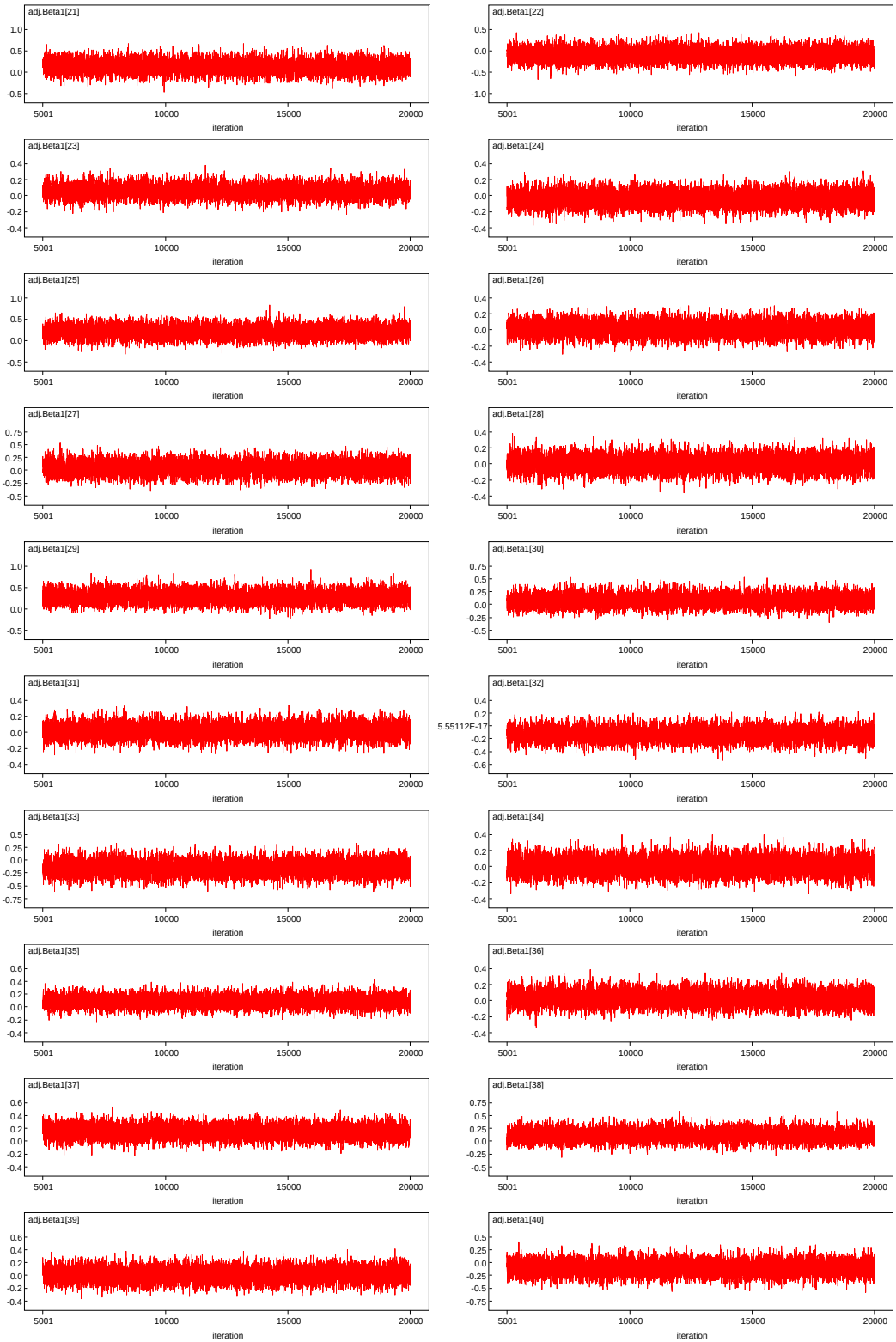
ภาพที่ ค - 6 (ต่อ)

ภาคผนวก ง

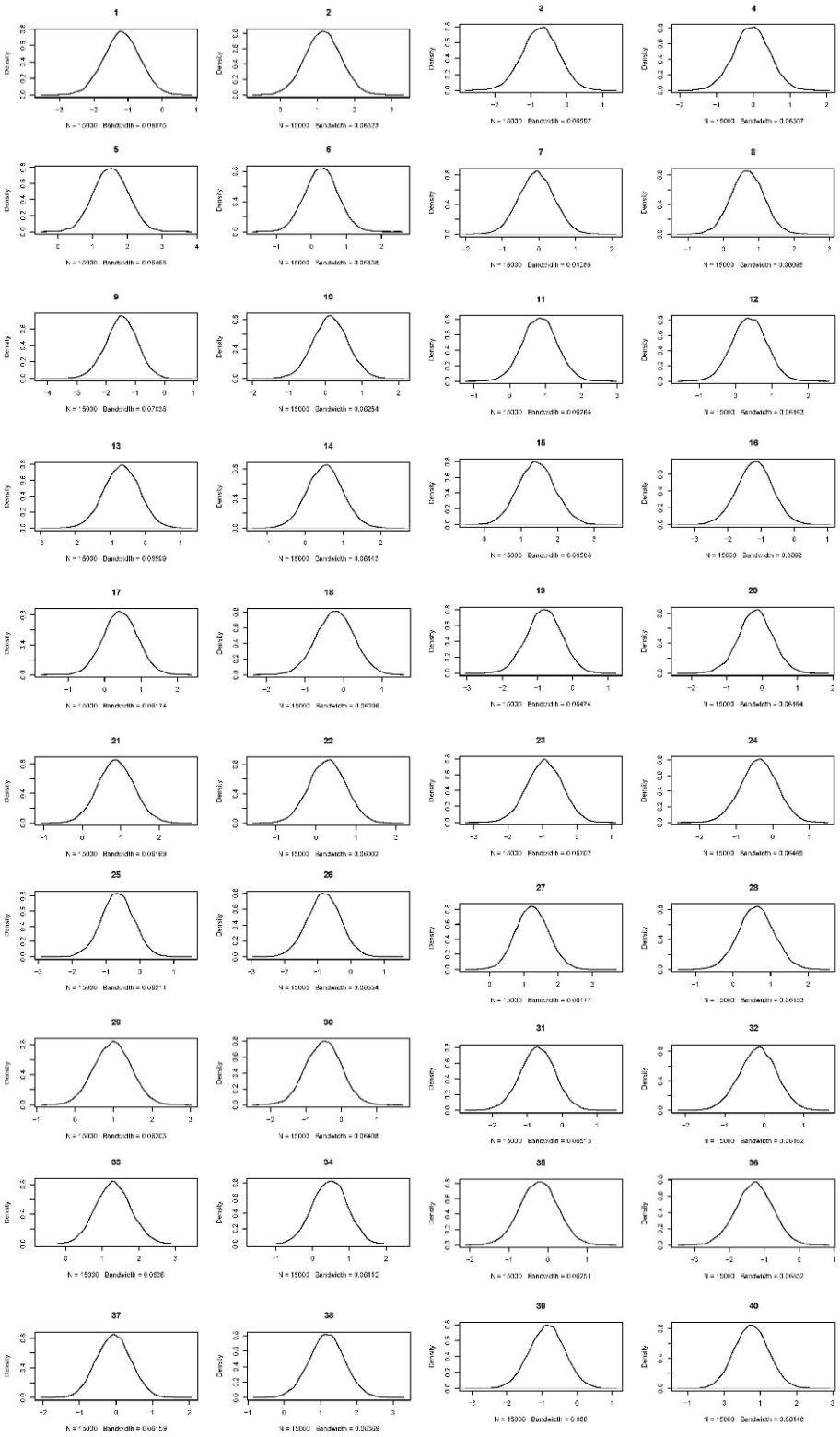
ตัวอย่าง History plot, density plot และ acf (autocorrelation function) plot
ของพารามิเตอร์ที่ใช้ในการตัดสิน DIF ที่ประมาณค่าด้วยวิธี Bayes และวิธี Bayes$_{\gamma}$



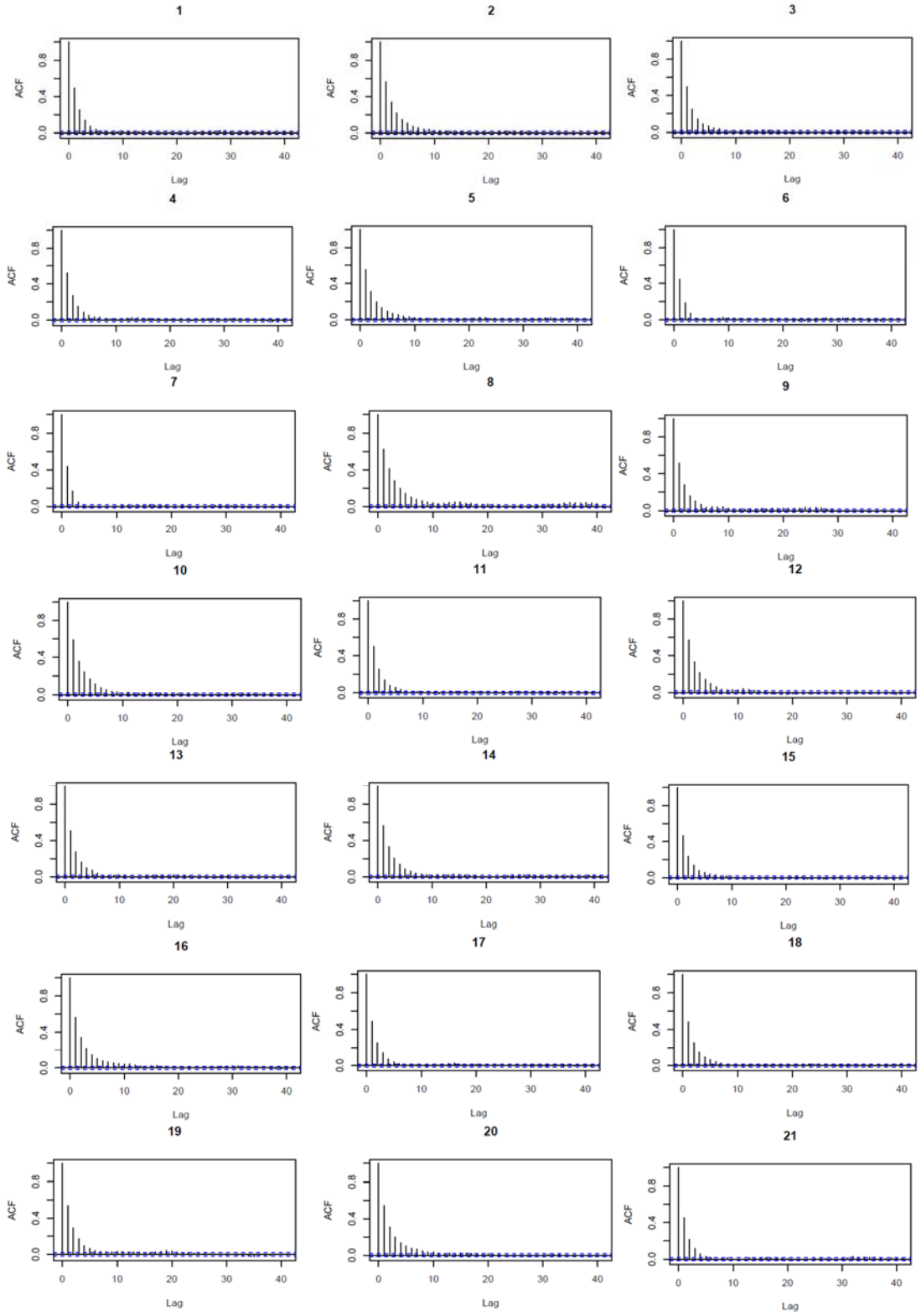
ภาพที่ ง - 1 ตัวอย่าง History plot ของพารามิเตอร์ที่ใช้ในการตัดสิน DIF ที่ประมาณค่าด้วยวิธี Bayes$_{\gamma}$


adj.Beta1[21]
adj.Beta1[22]
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]
5.55112E-17
5001
10000
15000
20000
iteration

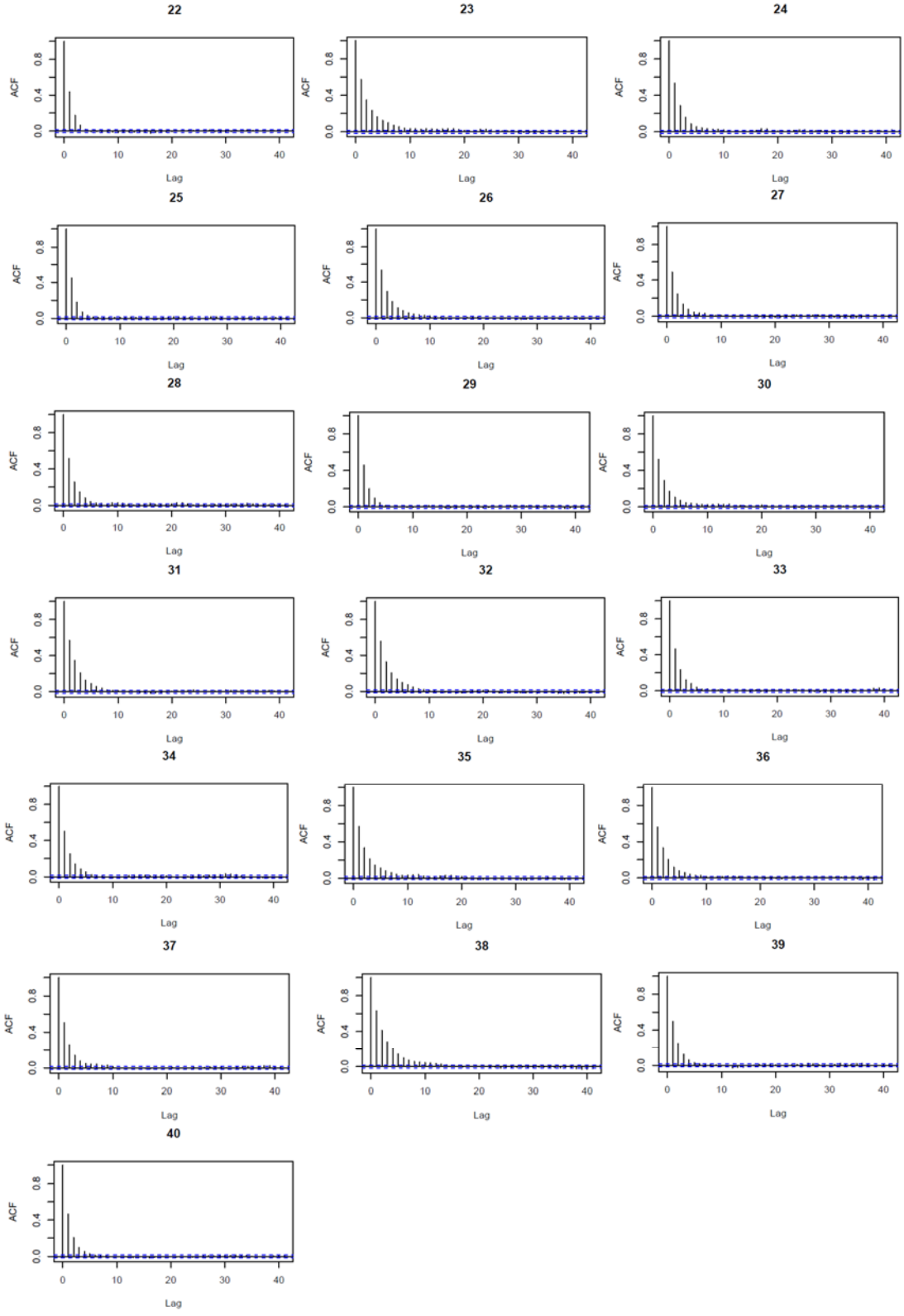
ภาพที่ ง - 1 (ต่อ)



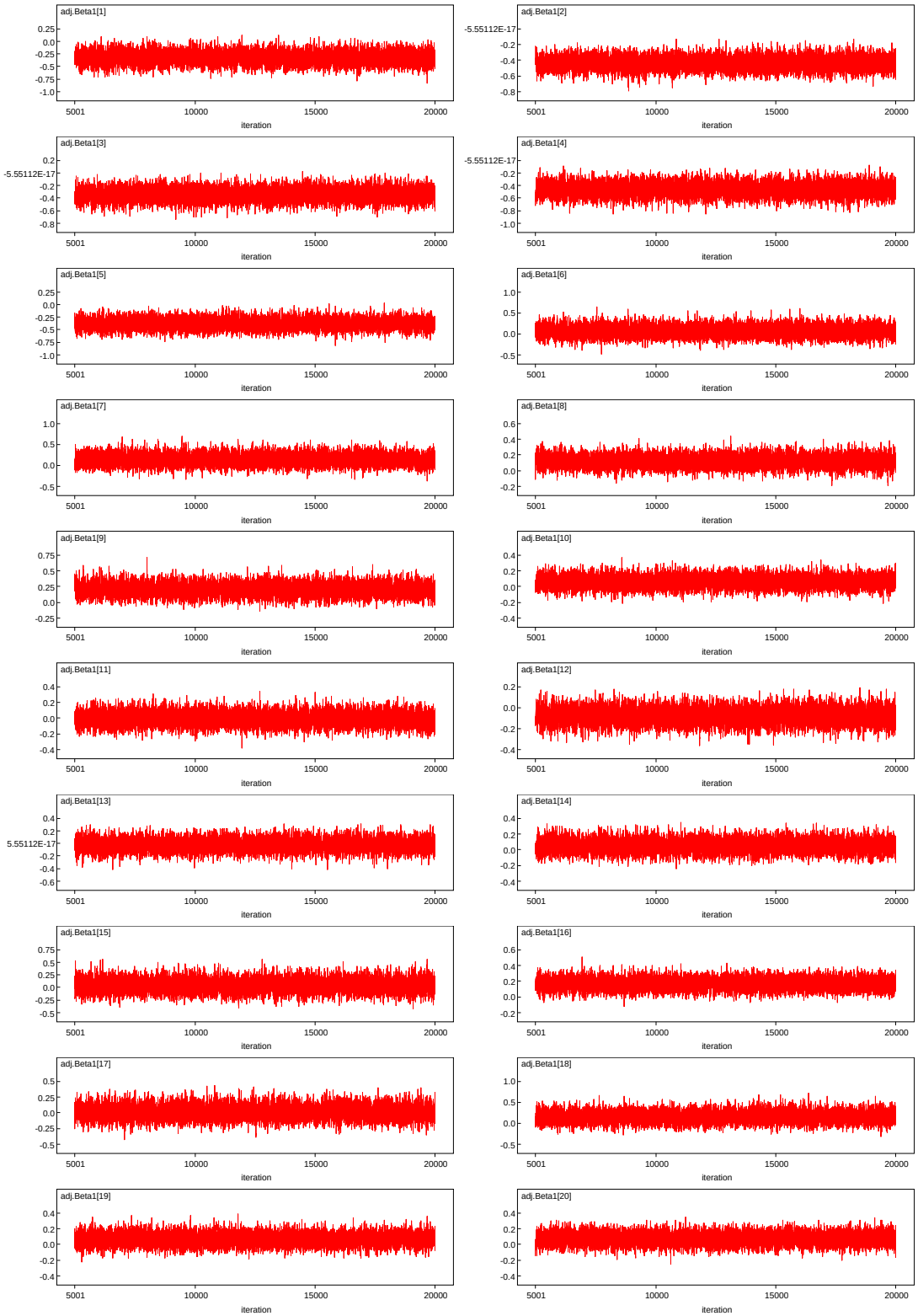
ภาพที่ ง - 2 ตัวอย่าง density plot ของพารามิเตอร์ที่ใช้ในการตัดสิน DIF ที่ประมาณค่าด้วยวิธี Bayes$_{\gamma}$



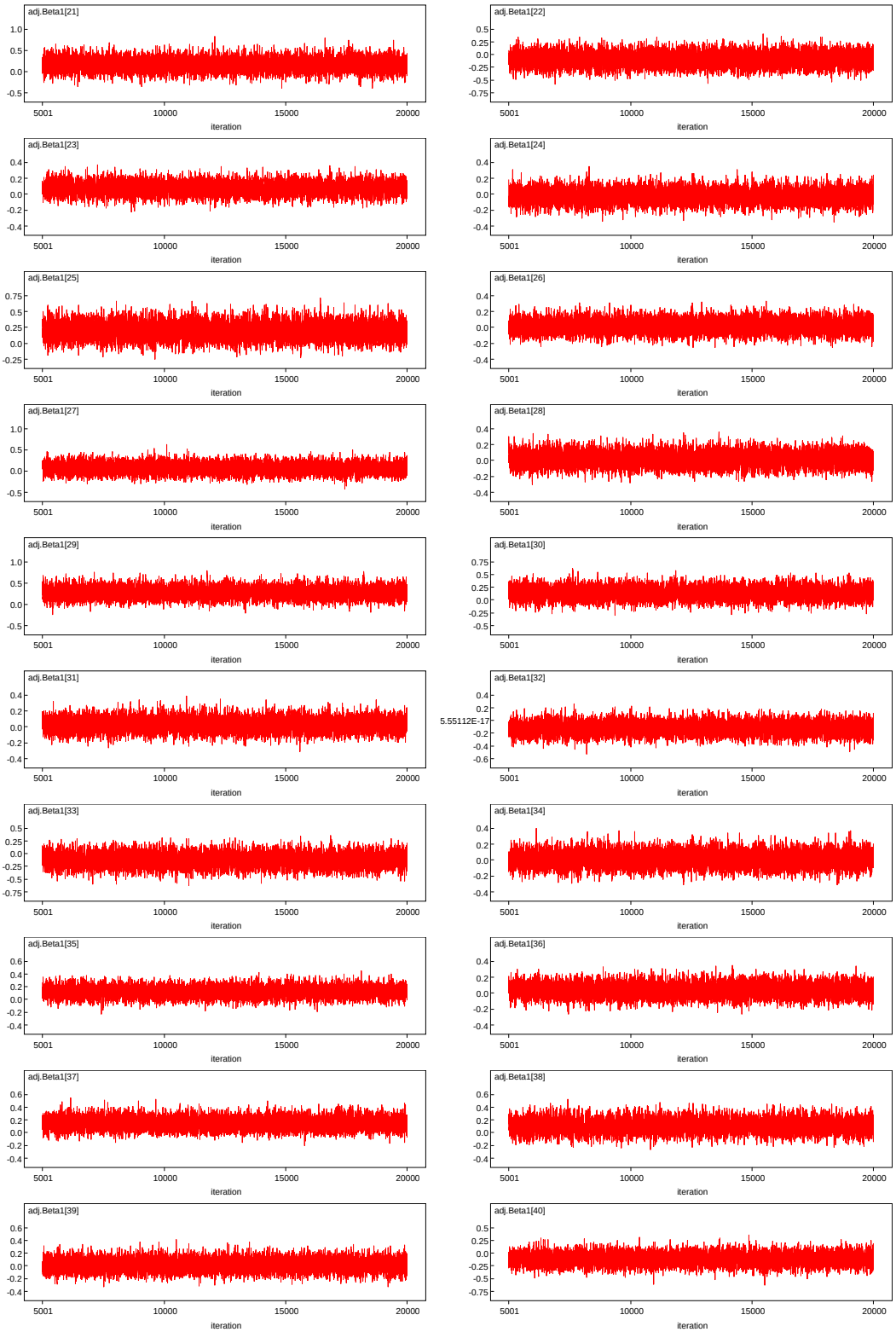
ภาพที่ ง - 3 ตัวอย่าง acf plot ของพารามิเตอร์ที่ใช้ในการตัดสิน DIF ที่ประมาณค่าด้วยวิธี Bayesγ



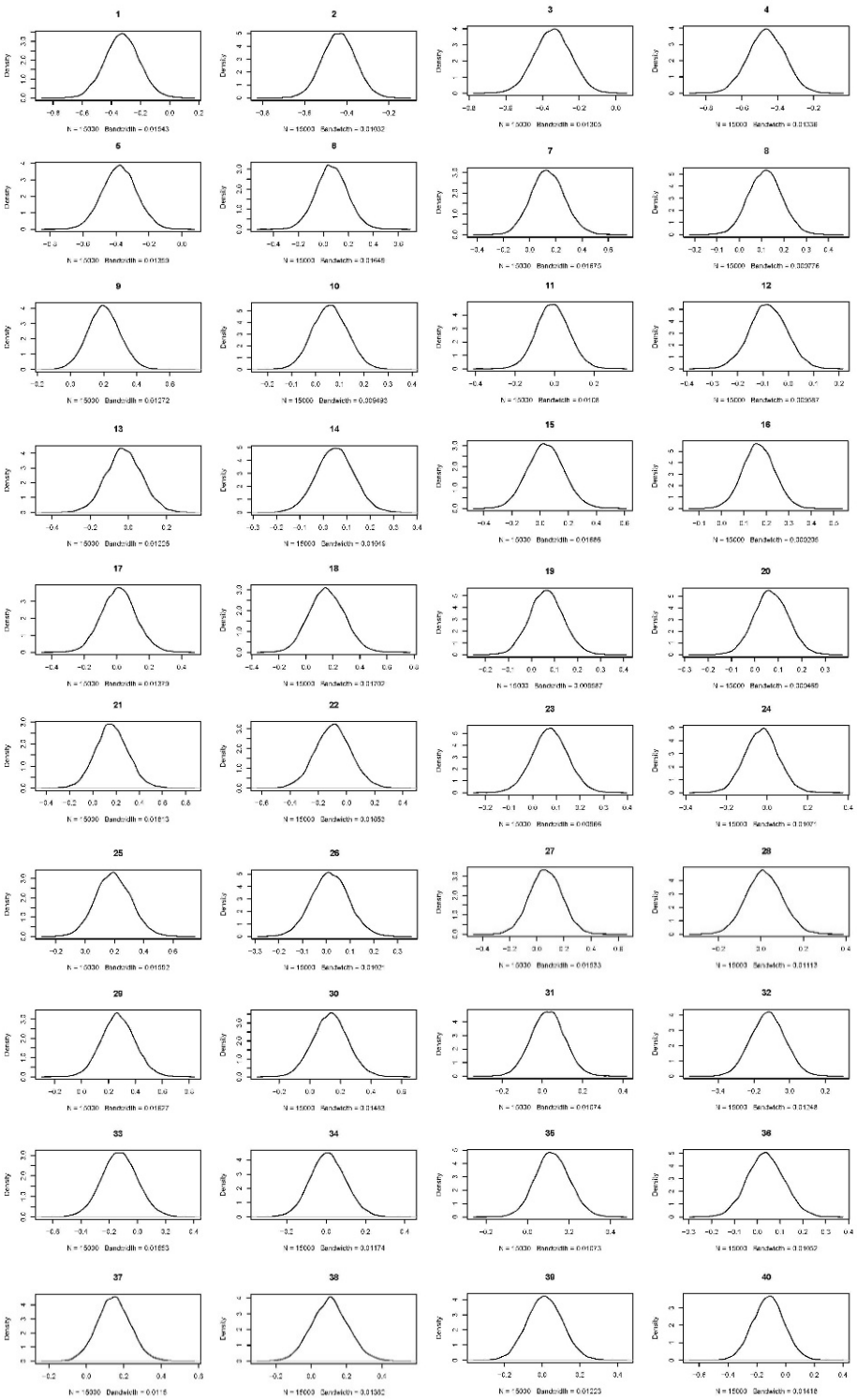
ภาพที่ ง - 3 (ต่อ)



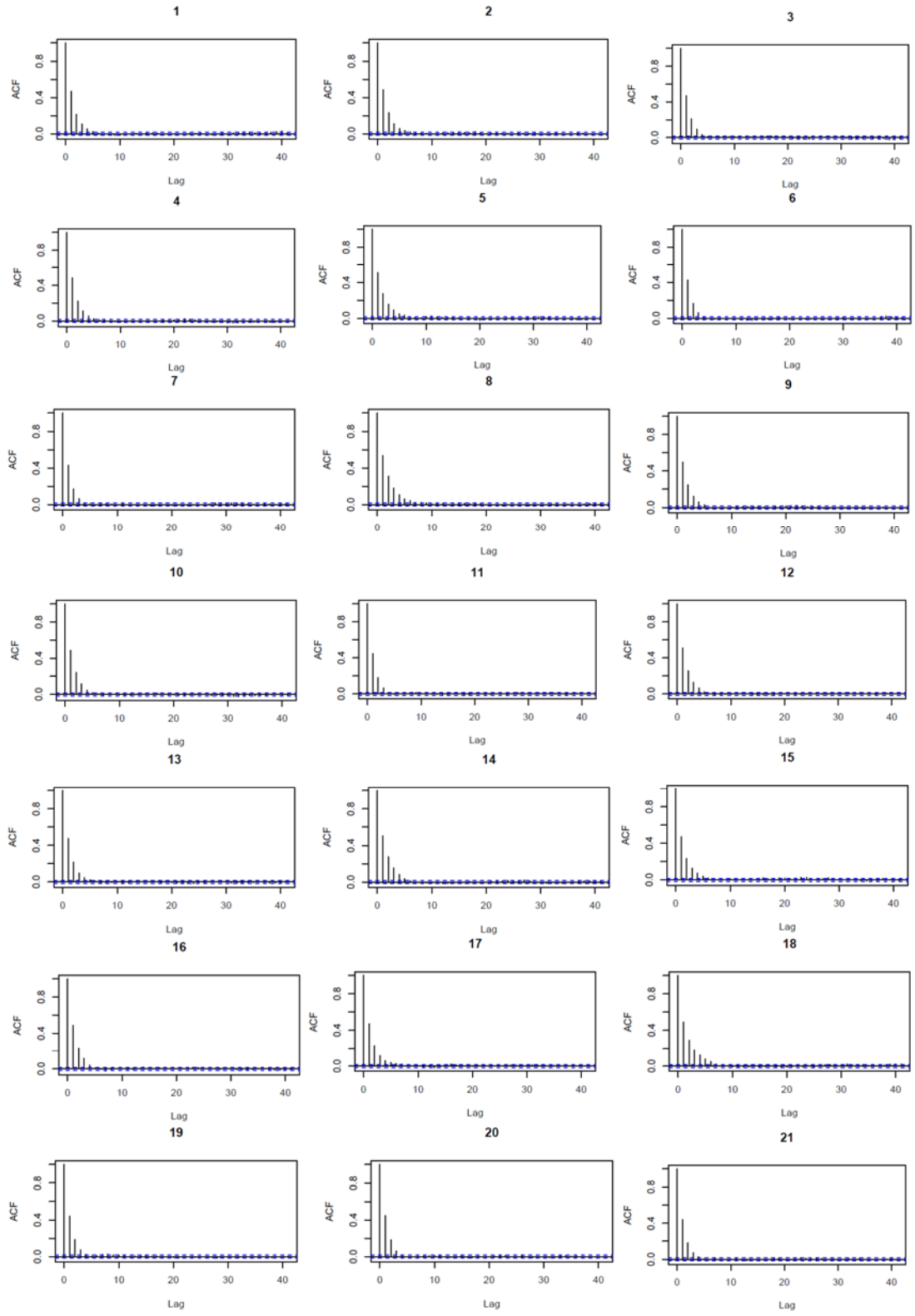
ภาพที่ ง - 4 ตัวอย่าง History plot ของพารามิเตอร์ที่ใช้ในการตัดสิน DIF ที่ประมาณค่าด้วยวิธี Bayes


adj.Beta1[21]
adj.Beta1[22]
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]
5.55112E-17
adj.Beta1[33]
adj.Beta1[34]
adj.Beta1[35]
adj.Beta1[36]
adj.Beta1[37]
adj.Beta1[38]
adj.Beta1[39]
adj.Beta1[40]
5001
10000
15000
20000
iteration

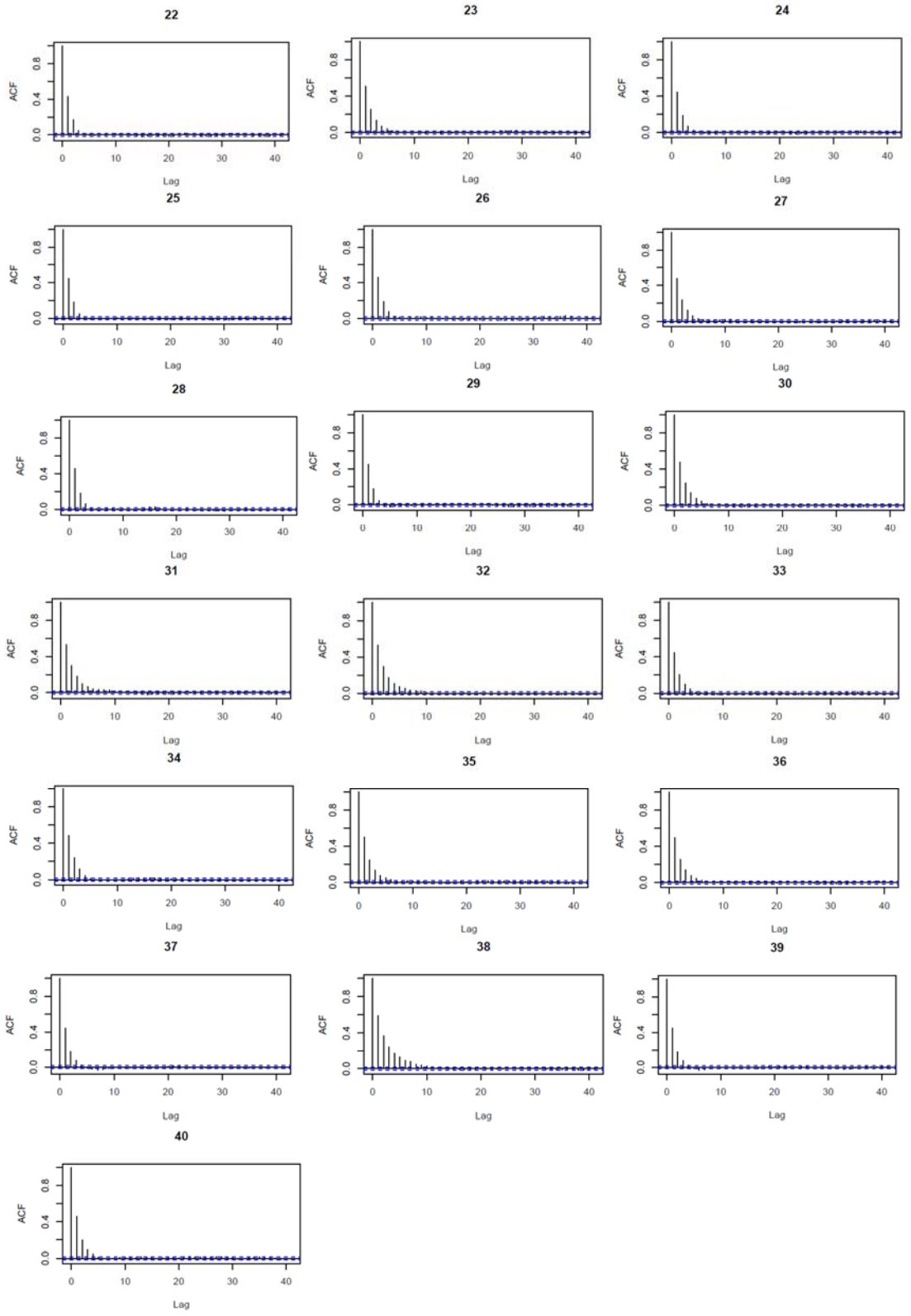
ภาพที่ ง - 4 (ต่อ)



ภาพที่ ง - 5 ตัวอย่าง density plot ของพารามิเตอร์ที่ใช้ในการตัดสิน DIF ที่ประมาณค่าด้วยวิธี Bayes



ภาพที่ ง - 6 ตัวอย่าง acf plot ของพารามิเตอร์ที่ใช้ในการตัดสิน DIF ที่ประมาณค่าด้วยวิธี Bayes



ภาพที่ ง - 6 (ต่อ)

ภาคผนวก จ

คำสั่งในโปรแกรม R สำหรับการประมวลผลด้วยวิธีของเบส์
และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท

คำสั่งในโปรแกรม R สำหรับการประมวลผลด้วยวิธีของเบส์

```
IRTmodel <- function(){
mu.Beta1<-mean(Beta1[])
mu.z<-mean(z[])
sd.z<-sd(z[])
for (i in 1:n){
        a[i]~dnorm(mua,siga) %_%I(0,)
        b[i]~dnorm(mub,sigb)
        Beta1[i]~dnorm(muBeta1,sigBeta1)
        adj.Beta1[i]<-(Beta1[i]-mu.Beta1) }
mua~dnorm(0,.0001)
mub~dnorm(0,.0001)
Beta~dnorm(muBeta,sigBeta)
muBeta~dnorm(0,.0001)
muBeta1~dnorm(0,.0001)
siga~dchisqr(.5)
sigb~dchisqr(.5)
sigBeta~dchisqr(.5)
sigBeta1~dchisqr(.5)
for (j in 1:N){
        for (i in 1:n){
            p[j,i]<- 1/(1+exp(-(a[i]*(((Beta*Group[j])+z[j])-b[i]-(Beta1[i]*Group[j])))))
            U[j,i]~dbern(p[j,i])
            }
        z[j] ~ dnorm(0,1)
        }
adj.Beta<-(Beta-mu.Beta1)
}
```

```
filename <- file.path("C:/A1", "IRT.bug")
write.model(IRTmodel, filename)
data <- list("N", "n", "Group", "U")
inits <- function(){
    list(Beta = 0, Beta1 =dd, a = a,  b = b, muBeta = 0.0, muBeta1 = 0.0, mua = 0.0,  mub = 0.0,
                sigBeta = 0.5, sigBeta1 = 0.5, siga = 0.5, sigb = 0.5)
                }
parameters <- c("b", "a", "adj.Beta", "adj.Beta1", "z")
dataIRT.sim <- bugs(data, inits, parameters , model.file = "C:/A1/IRT.bug", n.thin=1,
                 n.chains = 1, n.iter = 20000, n.burnin = 5000,  digits=5, DIC = TRUE,
                 bugs.directory = "c:/WinBUGS14/")
```

คำสั่งในโปรแกรม R สำหรับการประมวลผลด้วยวิธีของเบส์แบบมีอิทธิพลเทสต์เลท

```
tesletsmodel <- function(){
        mu.Beta1<-mean(Beta1[])
        Beta~dnorm(muBeta,sigBeta)
        mu.z<-mean(z[])
        sd.z<-sd(z[])
        for (i in 1:n){
            a[i]~dnorm(mua,siga) %_%I(0,)
            b[i]~dnorm(mub,sigb)
            Beta1[i]~dnorm(muBeta1,sigBeta1)
            adj.Beta1[i]<-(Beta1[i]-mu.Beta1)
        }
mua~dnorm(0,.0001)
mub~dnorm(0,.0001)
muBeta~dnorm(0,.0001)
muBeta1~dnorm(0,.0001)
siga~dchisqr(.5)
sigb~dchisqr(.5)
sigBeta~dchisqr(.5)
sigBeta1~dchisqr(.5)
for (j in 1:N){
        for (i in 1:n){
            p[j,i]<- 1/(1+exp(-a[i]*(((Beta*Group[j])+z[j])-b[i]+gamma[j,i]
            -Beta1[i]*Group[j])))
            U[j,i]~dbern(p[j,i])
            gamma[j,i]<-gamtes[j,test[i]]}
            z[j]~dnorm(0,1)
```

```
            for (i in 1:NT){
                gamtes[j,i]~dnorm(0,siggam[i])
                }
            }
adj.Beta<-(Beta-mu.Beta1)
for (i in 1:NT){
            siggam[i]~dchisqr(.5)
            }
}
filename <- file.path("C:/", "Testlet.bug")
write.model(tesletsmodel, filename)
dataTest <- list("test", "NT", "N", "n", "Group", "U")
initsTest <- function(){
    list(Beta = 0, Beta1 = dd, a = a,  b = b, muBeta = 0.0, muBeta1 = 0.0,    mua = 0.0, mub = 0.0,
                sigBeta = 0.5, sigBeta1 = 0.5, siga = 0.5, sigb = 0.5,  siggam = VarTest)
             }
parameters <- c("b", "a",  "adj.Beta", "adj.Beta1", "z")
dataTestlet.sim <- bugs(dataTest ,inits = initsTest, parameters, model.file = "C:/Testlet.bug",
                n.thin=1, n.chains = 1, n.iter = 20000, n.burnin = 5000,   digits=5,  DIC = TRUE,
                bugs.directory = "C:/WinBUGS14/")
```