การเปรียบเทียบประสิทธิผลการประมาณค่าพารามิเตอร์และการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธีแมกซิมัมไลค์ลิฮู้ด วิธีของเบย์แบบไม่คำนึงถึงอิทธิพลของเทสเลท และวิธีของเบย์แบบคำนึงถึงอิทธิพลของเทสเลท*

Comparing the Effectiveness of Parameter Estimation and Differential Item Functioning by Using Maximum Likelihood, Bayesian, and Bayesian with Testlet

อริสฬา เตหลิ่ม**
ดร.ไพรัตน์ วงษ์นาม***
ดร.สมพงษ์ ปั้นหุ่น****

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาประสิทธิผลในการประมาณค่าพารามิเตอร์ของผู้สอบ (ความสามารถ) และเพื่อศึกษาผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ระหว่างวิธีแมกซิมัมไลค์ลิฮู้ด (ML) วิธีของเบย์แบบไม่คำนึงถึงอิทธิพลของเทสเลท (Bayes) และวิธีของเบย์แบบคำนึงถึงอิทธิพลของเทสเลท (Bayesγ) ข้อมูลที่ศึกษาเป็นข้อมูลจำลองภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ อิทธิพลของเทสเลท (เท่ากันทุกเทสเลท และข้อสอบที่เป็นอิสระในเทสเลท) การแจกแจงของความสามารถ (ปกติ เบ้ซ้าย เบ้ขวา) จำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบ (5 ข้อและ 8 ข้อจากแบบสอบ 40 ข้อ) และอัตราส่วนของกลุ่มเปรียบเทียบต่อกลุ่มอ้างอิง (1000:1000, 1000:100) รวมจำนวนเงื่อนไขทั้งหมด 24 เงื่อนไข (2 x 3 x 2 x 2) กำหนดจำนวนรอบในการประมาณค่าพารามิเตอร์ในแต่ละเงื่อนไข 100 รอบ ผลการวิจัยสรุปได้ ดังนี้

1. เมื่อเปรียบเทียบประสิทธิภาพในการประมาณค่าพารามิเตอร์ของผู้สอบระหว่างวิธี ML วิธี Bayes และวิธี Bayesγ พบว่า วิธี Bayesγ ประมาณค่าพารามิเตอร์ได้ดีที่สุด

2. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ พบว่า วิธี Bayes และวิธี Bayesγ สามารถควบคุมอัตราความคาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ที่กำหนด และมีอำนาจการตรวจสอบสูงเมื่อมีการแจกแจงความสามารถแบบเบ้ซ้ายและจำนวนตัวอย่างมาก แต่ไม่มากถึงเกณฑ์ที่กำหนด ตรงข้ามกับวิธี ML ซึ่งไม่สามารถควบคุมอัตราความคาดเคลื่อนประเภทที่ 1 แต่มีอำนาจการตรวจสอบสูง

คำสำคัญ : การทำหน้าที่ต่างกันของข้อสอบ/ เทสเลท/ วิธีของเบย์

---

*ดุษฎีนิพนธ์ปรัชญาดุษฎีบัณฑิต สาขาวิชาวิจัย วัดผลและสถิติการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยบูรพานักวิชาการศึกษากองทะเบียนและประมวลผลการศึกษา สำนักงานอธิการบดี มหาวิทยาลัยบูรพา

**นิสิตหลักสูตรปรัชญาดุษฎีบัณฑิต สาขาวิชาวิจัย วัดผลและสถิติการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา

***อาจารย์ที่ปรึกษาหลัก รองศาสตราจารย์ ภาควิชาวิจัยและจิตวิทยาประยุกต์ คณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา

****อาจารย์ที่ปรึกษาร่วม ภาควิชาวิจัยและจิตวิทยาประยุกต์ คณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา

ABSTRACT

The objectives of this research were (1) to study the effectiveness of person parameter estimation, and (2) to study the differential item functioning by Using Maximum Likelihood (ML), Bayesian (Bayes), and Bayesian with Testlet (Bayesγ). In this study, the data were simulated. All conditions were consists of 2 levels of testlet effect (equal effect, independent + testlet), 3 levels of ability distribution (normal, negative, and positive skew distributions), 2 amount of items with DIF (5, and 8 items in the 40-item test length), and 2 levels of ratio of a reference and focal group (1000:1000, 1000:100) The entire total of testing conditions was 24 (2 x 3 x 2 x 2). The 100 replications used to estimate the item parameters and test statistics in each condition. The research results were as follows:

1. From the comparison of ML, Bayes, and Bayesγ for person parameter estimation, it was found that Bayesγ had the best estimator.

2. From the study of the detection of Differential Item Functioning, it was found that the Bayes, and Bayes γ estimate procedures had well-control of Type I error rate and higher the power rate when negatively skewed ability distributions and sample sizes were increased but it is not adequate for criteria, whereas ML estimate procedures had not control of Type I error rate and high power of DIF detection.

Keywords : Differential Item Functioning/ Testlet/ Bayesian estimator

บทนำ

ทฤษฎีการตอบสนองข้อสอบ มีการตกลงเกี่ยวกับการยอมรับความจริงเบื้องต้นข้อหนึ่ง นั่นคือ ความเป็นอิสระระหว่างข้อสอบและผู้สอบ (Item Local Independent) หมายถึง เมื่อควบคุมความสามารถ (Latent Trait หรือ Ability) ที่ส่งผลต่อข้อสอบให้คงที่แล้ว ผลการตอบข้อสอบแต่ละข้อต้องเป็นอิสระกัน ซึ่งหากไม่คำนึงถึงข้อตกลงเบื้องต้นนี้แล้ว จะนำไปสู่ข้อผิดพลาดต่าง ๆ สามารถพบข้อผิดพลาดในสถานการณ์จริง บ่อยครั้งข้อตกลงเบื้องต้นดังกล่าวมักขัดแย้งกับการนำไปใช้ โดยที่ผลจากการตอบข้อสอบข้อหนึ่งถูกจะมีผลต่อการตอบข้อสอบข้ออื่นถูกด้วย หรือเมื่อแบบสอบมีลักษณะเป็นเทสเลท (Testlet) เช่น ข้อสอบหลายข้อที่ต้องใช้ข้อมูลในการตอบจากกราฟ หรือแผนภูมิ หรือบทความเดียวกัน หรือแบบสอบที่เป็นการสอบทักษะการอ่าน (Reading Comprehension Test) ซึ่งแบ่งแบบสอบเป็นตอน ๆ ในตอนหนึ่งอาจประกอบด้วยข้อคำถาม 4 - 12 ข้อ ดังนั้น ผลการตอบสนองของข้อคำถามเหล่านี้ ขึ้นอยู่กับผู้สอบเข้าใจบทความที่ใช้เป็นคำถามมากน้อยเพียงใด หรือ กรณีของคอมพิวเตอร์ปรับเหมาะ (Computerized Adaptive Testing: CAT) ในการสุ่มเลือกข้อสอบอาจเกิดกรณีที่ผลการตอบของข้อสอบข้อหนึ่งมาจากโจทย์ของอีกข้อหนึ่ง (Cross-information) ซึ่งแสดงถึงความไม่เป็นอิสระของการตอบคำถามในแต่ละข้อ เช่น ในมุมของทฤษฎีการทดสอบแบบดั้งเดิม หากเกิดความไม่เป็นอิสระระหว่าง

ข้อสอบและผู้สอบ เนื่องจากเทสเลท (Testlet) แล้วจะทำให้ประมาณค่าความคาดเคลื่อนมาตรฐานของการวัด (Standard Error of Measurement) น้อยเกินจริง ซึ่งทำให้การประมาณค่าความเที่ยง (Reliability) สูงเกินจริง ส่วนในมุมของทฤษฎีการตอบสนองข้อสอบ (IRT) จะส่งผลให้ประมาณค่าสารสนเทศของแบบสอบสูงขึ้น นั่นหมายถึง มีการประมาณค่าความคลาดเคลื่อนมาตรฐานต่ำเกินจริง นอกจากนี้ ยังเกิดอคติในการประมาณค่าความยากและค่าอำนาจจำแนกของข้อสอบด้วย

สำหรับวิธีการประมาณค่าพารามิเตอร์ นั้น Zhang (2010) ได้สรุปข้อมูลบทความเกี่ยวกับเทสเลทระหว่างปี 1989-2009 ในฐานข้อมูลของ EBSCO และ PsychInfo พบว่า ร้อยละ 82.76 ประมาณค่าพารามิเตอร์ด้วยวิธี ML และร้อยละ 17.24 ประมาณค่าพารามิเตอร์ด้วยวิธี Bayes จะเห็นว่าในการวิเคราะห์ข้อสอบด้วยทฤษฎีการตอบสนองข้อสอบ มักนำวิธี ML และวิธี Bayes มาใช้ในการประมาณค่าพารามิเตอร์ โดยที่ข้อดีของการประมาณค่าด้วยวิธี ML คือ ให้สารสนเทศของค่าพารามิเตอร์ที่ต้องการได้ทั้งหมด ไม่ว่าจะเป็นพารามิเตอร์ของผู้สอบและพารามิเตอร์ของข้อสอบ แต่การประมาณค่าพารามิเตอร์ของวิธีนี้ ขึ้นอยู่กับจำนวนของกลุ่มผู้เข้าสอบและข้อสอบ ถ้ามีจำนวนเพิ่มขึ้น การประมาณค่าก็จะมีความคงที่ไปสู่ค่าพารามิเตอร์เพิ่มมากขึ้น ส่วนข้อจำกัดของการประมาณค่าด้วยวิธี ML คือ การประมาณค่าพารามิเตอร์ในขั้นที่ 2 และ 3 โดยใช้ค่าอนุพันธ์อันดับที่ 2 ในกระบวนการนิวตัน-ราฟสัน มีโอกาสที่ค่าประมาณที่ได้จะไม่ลู่เข้าสู่ค่าคงที่ ประเด็นถัดมาสำหรับการประมาณค่าในสมการไลค์ลิฮูดไม่ใช่สมการเชิงเส้นตรง จะทำให้การหารากของสมการที่ทำให้ฟังก์ชันไลค์ลิฮูดมีค่าสูงสุดได้หลายค่าแต่ค่าเหล่านี้ไม่สามารถนำไปใช้หรือประกันได้ว่าเป็นค่าพารามิเตอร์ที่แท้จริงได้ ประเด็นที่สาม ในบางครั้งค่าพารามิเตอร์หรือค่าที่ได้จากการประมาณไม่ตกอยู่ในขอบเขตของค่าพารามิเตอร์ กล่าวคืออาจมีค่าใดค่าหนึ่งอยู่ภายนอกขอบเขตที่ยอมรับได้ ในกรณีเช่นนี้ต้องมีการกำหนดขอบเขตจำกัดของค่าประมาณไว้ เพื่อให้ค่าประมาณที่ได้ไม่สูงหรือต่ำเกินไปนัก

ดังนั้น วิธี Bayes จึงอาจเป็นวิธีที่เหมาะสมกว่า ทั้งนี้เพราะวิธี Bayes มีแนวคิดบางประการที่ต่างออกไปจากแนวคิดของวิธี ML นั่นคือ ค่าพารามิเตอร์ของผู้สอบและค่าพารามิเตอร์ของข้อสอบเป็นตัวแปรสุ่ม (Random Variable) จากการแจกแจงที่แสดงได้ด้วยฟังก์ชันความหนาแน่นร่วม (Joint Density Function) หรือ การแจกแจงเริ่มแรก (Prior Distribution) ซึ่งทำให้การใช้ฟังก์ชันไลค์ลิฮูดเพียงอย่างเดียวในการประมาณค่าถูกพิจารณาว่าเป็นการใช้ข้อมูลที่มีอยู่อย่างไม่ครบถ้วน เพราะยังมีการแจกแจงเริ่มต้นร่วมกับฟังก์ชันความหนาแน่นร่วมที่ควรนำมาใช้ในการประมาณค่าพารามิเตอร์ด้วย แม้จะมีข้อดีมากกว่าวิธี ML แต่วิธี Bayes ใช้เวลาในการวิเคราะห์นานกว่าวิธี ML มาก

นอกจากการหาคุณภาพของแบบสอบที่วิเคราะห์ค่าพารามิเตอร์ของข้อสอบด้วยวิธีต่าง ๆ แล้ว ยังมีสิ่งที่ต้องคำนึงถึงในการพัฒนาแบบสอบอีกกรณีหนึ่ง คือ แบบสอบต้องมีความยุติธรรมกับผู้สอบทุกกลุ่ม ไม่เข้าข้างกลุ่มใดกลุ่มหนึ่ง ลักษณะการทำหน้าที่ของข้อสอบประเภทนี้เรียกว่า การทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) และจากที่กล่าวมาแล้วว่าอิทธิพลของเทสเลทส่งผลกระทบต่อการประมาณค่าพารามิเตอร์ ผู้วิจัยจึงคาดว่าอิทธิพลของเทสเลทจะส่งผลต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วย

แม้จะมีผลกระทบดังที่กล่าวมาแล้ว แต่การวิเคราะห์ข้อสอบก็มักจะไม่คำนึงถึงอิทธิพลของเทสเลท ข้อจำกัดประการหนึ่งคือ โปรแกรมสำเร็จรูปที่รองรับ

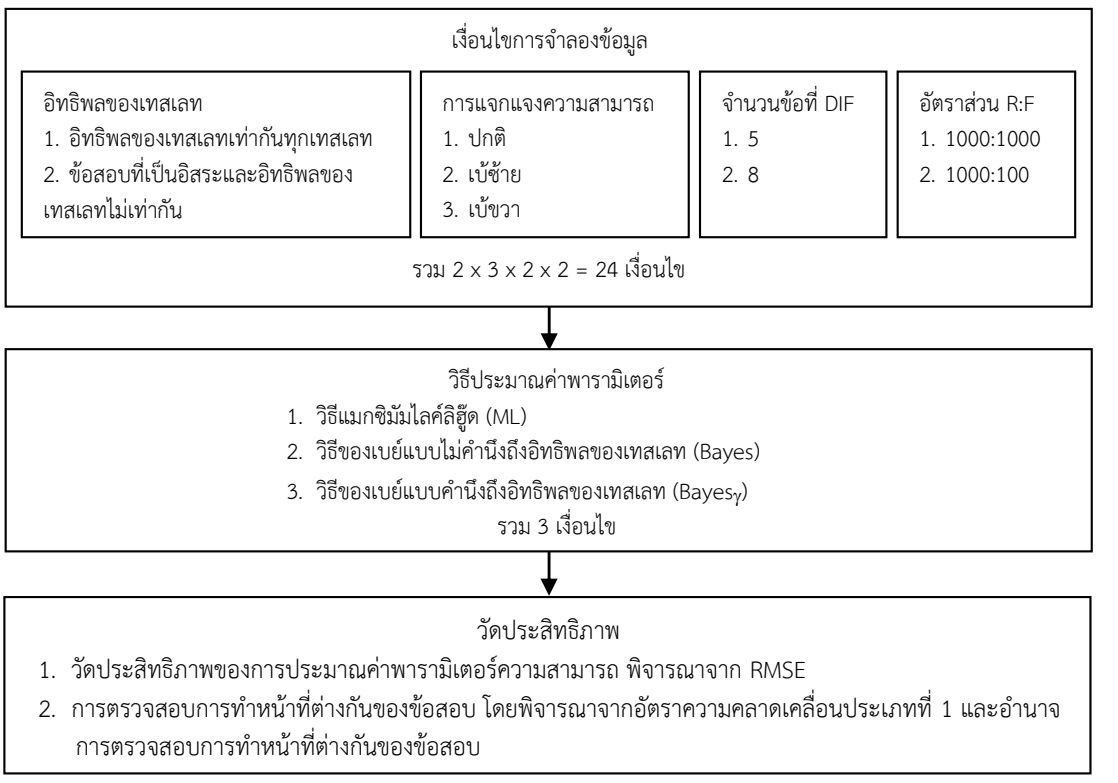
การวิเคราะห์มีน้อย และยังไม่สะดวกต่อการใช้งาน เช่น SCORIGHT แม้จะรองรับการวิเคราะห์แบบสอบที่ส่วนประกอบของเทสเลทได้ แต่หน้าจอที่ติดต่อกับกับผู้ใช้งานเป็นแบบ Command Prompt (โปรแกรมสำเร็จรูปส่วนใหญ่จะมีโหมดกราฟฟิค GUI) รวมทั้งการแสดงผลลัพท์ที่เป็นไฟล์ข้อความ (Text Files) ยังไม่สามารถแสดงรูปภาพได้ ดังนั้นส่วนใหญ่แล้ว วิธีที่ใช้ในการประมาณค่าพารามิเตอร์ในการวิเคราะห์ข้อสอบจึงยังคงใช้ทฤษฎีการตอบสนองข้อสอบที่ยังคงไม่คำนึงถึงอิทธิพลของเทสเลท (traditional IRT) โดยที่วิธีวิเคราะห์ที่โปรแกรมสำเร็จรูปใช้ ส่วนมากจะเป็นวิธี ML และวิธี Bayes จากปัญหาข้อจำกัดดังกล่าว ผู้วิจัยจึงต้องการตรวจสอบว่าการวิเคราะห์โดยละเลยอิทธิพลของเทสเลท จากการวิเคราะห์ข้อสอบด้วยวิธี ML และวิธี Bayes เมื่อเปรียบเทียบกับวิธีของเบย์แบบคำนึงถึงอิทธิพลของเทสเลท (Bayesγ) นั้นส่งผลเหมือนหรือต่างกันอย่างไร

วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาประสิทธิภาพในการประมาณค่าพารามิเตอร์ผู้สอบ (ความสามารถ) ด้วยวิธี ML วิธี Bayes และวิธี Bayesγ

2. เพื่อศึกษาประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ด้วยวิธี ML วิธี Bayes และวิธี Bayesγ

กรอบแนวคิดในการวิจัย



ภาพ 1 กรอบแนวคิดในการวิจัย

การวิเคราะห์ข้อมูล

การวิจัยครั้งนี้กำหนดให้ความยาวแบบสอบเป็น 40 ข้อ ประกอบด้วย 4 เทสเลท โดยแต่ละเทสเลทประกอบด้วยข้อสอบ 10 ข้อ และกำหนดให้ข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบมีความเข้ม (Magnitude of DIF) เท่ากับ 0.5 โดยใช้ข้อมูลที่ได้จากการจำลอง 24 เงื่อนไข (2 x 3 x 2 x 2) ด้วยโปรแกรม R ในส่วนของประมาณค่าพารามิเตอร์และการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี ML ทำการประมวลผลและแสดงผลโดยใช้ Package TAM ของโปรแกรม R ส่วนการประมาณค่าพารามิเตอร์ด้วยวิธี Bayes และ Bayesγ จะทำการเขียนคำสั่งการประมวลผลด้วยโปรแกรม WinBUGS และกลับมาแสดงผลจะแสดงผลในโปรแกรม R โดยใช้ Package R2WinBUGS มีรายละเอียดในการดำเนินการวิจัยดังนี้

1. ทำการจำลองข้อมูลจากเงื่อนไข ดังนี้

1.1 อิทธิพลของเทสเลทในแบบสอบ (TestletEff.) ประกอบด้วย 2 เงื่อนไข ได้แก่

1.1.1 อิทธิพลของเทสเลทเท่ากันทุกเทสเลท (0.8, 0.8, 0.8, 0.8)

1.1.2 ข้อสอบที่เป็นอิสระและอิทธิพลของเทสเลทไม่เท่ากัน (0, 0.25, 0.56, 1)

1.2 การแจกแจงของความสามารถ (Dist. theta) ประกอบด้วย 3 เงื่อนไข ได้แก่

1.2.1 การแจกแจงแบบปกติ

1.2.2 การแจกแจงแบบเบ้ซ้าย

1.2.3 การแจกแจงแบบเบ้ขวา

1.3 จำนวนข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบ (NumDIF) ประกอบด้วย 2 เงื่อนไข ได้แก่

1.3.1 ร้อยละ 10 หรือมีข้อสอบที่ทำหน้าที่ต่างกันจำนวน 5 ข้อ ในแบบสอบ

1.3.2 ร้อยละ 20 หรือมีข้อสอบที่ทำหน้าที่ต่างกันจำนวน 8 ข้อ ในแบบสอบ

1.4 อัตราส่วนของกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ (R:F) ประกอบด้วย 2 เงื่อนไข ได้แก่

1.4.1 เป็น 1:1 หรือกลุ่มอ้างอิง 1,000 คน และกลุ่มเปรียบเทียบ 1,000 คน

1.4.2 เป็น 1:0.1 หรือ กลุ่มอ้างอิง 1,000 คน และกลุ่มเปรียบเทียบ 100 คน

2. ประมาณค่าพารามิเตอร์ความสามารถและพารามิเตอร์ที่ตัดสินการทำหน้าที่ต่างกันของข้อสอบมี 3 วิธี ได้แก่ วิธี ML วิธี Bayes และวิธี Bayesγ

2.1 วิธี ML ทำการประมาณค่าพารามิเตอร์ด้วยการใช้โปรแกรม R จากการเรียกใช้ Package TAM ด้วยการใช้โมเดลการตอบสนองข้อสอบแบบ 2 พารามิเตอร์ (traditional IRT)

2.2 วิธี Bayes ประมาณค่าพารามิเตอร์ด้วยการใช้โปรแกรม R จากการเรียกใช้ Package R2WinBUGS ด้วยการใช้โมเดลการตอบสนองข้อสอบแบบ 2 พารามิเตอร์ (traditional IRT)

2.3 วิธี Bayesγ ประมาณค่าพารามิเตอร์ด้วยการใช้โปรแกรม R จากการเรียกใช้ Package R2WinBUGS ด้วยการใช้โมเดลการตอบสนองข้อสอบแบบ 2 พารามิเตอร์ โดยใช้ bi-factor MIRT (Fukuhara & Kamata, 2011) ดังสมการ

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

เมื่อ a_j หมายถึง พารามิเตอร์อำนาจจำแนก

$\hat{a}_\theta$ หมายถึง อิทธิพลของกลุ่ม G_i ต่อความสามารถ θ_i

G_i หมายถึง กลุ่มของผู้สอบ

ζ_i หมายถึง ส่วนที่เหลือ (Residual) สำหรับผู้สอบ i

b_j หมายถึง พารามิเตอร์ความยากของข้อ j

$\tilde{a}_{id(j)}$ หมายถึง อิทธิพลสุ่มของ Testlet *d(j)*

$\hat{a}_j$ หมายถึง ความต่างของพารามิเตอร์ความยากระหว่างกลุ่ม ใช้พิจารณาการทำหน้าที่ต่างกันของข้อสอบ โดยตัดสินว่าข้อสอบทำหน้าที่ต่างกันเมื่อมีค่าสัมบูณ์มากกว่า 0.426 และค่าขอบล่างของช่วงความเชื่อมั่น 95% และค่าขอบบนของช่วงความเชื่อมั่น 95% ไม่คลุม 0

3. วัดประสิทธิภาพของการประมาณค่า

3.1 วัดประสิทธิภาพของการประมาณค่าพารามิเตอร์ความสามารถ พิจารณาจากความเบี่ยงเบนของค่าพารามิเตอร์ที่แท้จริงและค่าที่ประมาณได้ หรือ Root Mean Square Error (RMSE) ของพารามิเตอร์ความสามารถของผู้สอบ

3.2 การวัดประสิทธิภาพของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ พิจารณาจากอัตราความคลาดเคลื่อนประเภทที่ 1 (Type I Error rate) ซึ่งหากมีค่าไม่เกิน 0.05 หรือร้อยละ 5 ถือว่าสามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ และอำนาจ (Power) ของตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

สรุปผลการวิจัย

1. ประสิทธิภาพในการประมาณค่าพารามิเตอร์ความสามารถ ด้วยวิธี ML วิธี Bayes และวิธี Bayesγ

1.1 ผลวิเคราะห์ความเบี่ยงเบนของค่าพารามิเตอร์ที่แท้จริงและค่าที่ประมาณได้ หรือ Root Mean Square Error (RMSE) ของพารามิเตอร์ความสามารถของผู้สอบ มีค่าดังนี้

ตารางที่ 1 ค่าเฉลี่ยดัชนี RMSE ของพารามิเตอร์ความสามารถของผู้สอบ

TestletEff.	Dist.theta	NumDIF	R:F	ML	Bayes	Bayesγ
0.8, 0.8, 0.8, 0.8	ปกติ	5	1000:1000	0.2906	0.2431	0.2349
0.8, 0.8, 0.8, 0.8	ปกติ	5	1000:100	0.2830	0.2532	0.2452
0.8, 0.8, 0.8, 0.8	ปกติ	8	1000:1000	0.2883	0.2443	0.2364
0.8, 0.8, 0.8, 0.8	ปกติ	8	1000:100	0.2743	0.2476	0.2386
0.8, 0.8, 0.8, 0.8	เบ้ซ้าย	5	1000:1000	0.6371	0.4103	0.3192
0.8, 0.8, 0.8, 0.8	เบ้ซ้าย	5	1000:100	0.6407	0.4147	0.3230
0.8, 0.8, 0.8, 0.8	เบ้ซ้าย	8	1000:1000	0.6339	0.4051	0.3150
0.8, 0.8, 0.8, 0.8	เบ้ซ้าย	8	1000:100	0.6372	0.4100	0.3169
0.8, 0.8, 0.8, 0.8	เบ้ขวา	5	1000:1000	0.6165	0.4071	0.3170
0.8, 0.8, 0.8, 0.8	เบ้ขวา	5	1000:100	0.6013	0.4051	0.3138
0.8, 0.8, 0.8, 0.8	เบ้ขวา	8	1000:1000	0.6212	0.4054	0.3137
0.8, 0.8, 0.8, 0.8	เบ้ขวา	8	1000:100	0.5908	0.4034	0.3172
0, 0.25, 0.56, 1	ปกติ	5	1000:1000	0.2305	0.1863	0.1572
0, 0.25, 0.56, 1	ปกติ	5	1000:100	0.2109	0.1911	0.1624
0, 0.25, 0.56, 1	ปกติ	8	1000:1000	0.2236	0.1842	0.1561
0, 0.25, 0.56, 1	ปกติ	8	1000:100	0.2038	0.1859	0.1557
0, 0.25, 0.56, 1	เบ้ซ้าย	5	1000:1000	0.6053	0.3781	0.2766
0, 0.25, 0.56, 1	เบ้ซ้าย	5	1000:100	0.5836	0.3651	0.2762
0, 0.25, 0.56, 1	เบ้ซ้าย	8	1000:1000	0.5818	0.3607	0.2770
0, 0.25, 0.56, 1	เบ้ซ้าย	8	1000:100	0.5850	0.3629	0.2766
0, 0.25, 0.56, 1	เบ้ขวา	5	1000:1000	0.5693	0.3601	0.2745
0, 0.25, 0.56, 1	เบ้ขวา	5	1000:100	0.5371	0.3558	0.2733
0, 0.25, 0.56, 1	เบ้ขวา	8	1000:1000	0.5713	0.3624	0.2726
0, 0.25, 0.56, 1	เบ้ขวา	8	1000:100	0.5446	0.3555	0.2727

จากตารางที่ 1 พบว่า ผลการประมาณค่าพารามิเตอร์ความสามารถ พบว่า ค่าเฉลี่ยดัชนี RMSE ของพารามิเตอร์ความสามารถ โดยรวมมีค่าระหว่าง 0.1557 - 0.6407 หากพิจารณาตามวิธีที่ใช้ในการประมาณค่า พบว่า วิธีแมกซิมัมไลค์ลิฮู้ดมีค่าระหว่าง 0.2038 - 0.6407 วิธี Bayes มีค่าระหว่าง 0.1842 - 0.4147 และวิธี Bayesg สามารถประมาณค่าพารามิเตอร์ความสามารถได้ดีที่สุด

2. ประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ด้วยวิธี ML วิธี Bayes และวิธี Bayesγ

2.1 ผลการวิเคราะห์อัตราความคลาดเคลื่อนประเภทที่ 1 (Type I Error Rate) ของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เป็นดังนี้

ตารางที่ 2 ผลการวิเคราะห์อัตราความคลาดเคลื่อนประเภทที่ 1 ของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

Testl etEff.	Dist.theta	NumDIF	R:F	ML	Bayes	Bayesγ
0.8, 0.8, 0.8, 0.8	ปกติ	5	1000:1000	0.3606	0.0040*	0.0000*
0.8, 0.8, 0.8, 0.8	ปกติ	5	1000:100	0.5517	0.0040*	0.0074*
0.8, 0.8, 0.8, 0.8	ปกติ	8	1000:1000	0.4613	0.0000*	0.0000*
0.8, 0.8, 0.8, 0.8	ปกติ	8	1000:100	0.5625	0.0069*	0.0088*
0.8, 0.8, 0.8, 0.8	เบ้ซ้าย	5	1000:1000	0.4149	0.0040*	0.0040*
0.8, 0.8, 0.8, 0.8	เบ้ซ้าย	5	1000:100	0.6000	0.0260*	0.0291*
0.8, 0.8, 0.8, 0.8	เบ้ซ้าย	8	1000:1000	0.4275	0.0156*	0.0181*
0.8, 0.8, 0.8, 0.8	เบ้ซ้าย	8	1000:100	0.5338	0.0344*	0.0319*
0.8, 0.8, 0.8, 0.8	เบ้ขวา	5	1000:1000	0.4214	0.0009*	0.0023*
0.8, 0.8, 0.8, 0.8	เบ้ขวา	5	1000:100	0.6137	0.0100*	0.0231*
0.8, 0.8, 0.8, 0.8	เบ้ขวา	8	1000:1000	0.4675	0.0050*	0.0084*
0.8, 0.8, 0.8, 0.8	เบ้ขวา	8	1000:100	0.5969	0.0269*	0.0244*
0, 0.25, 0.56, 1	ปกติ	5	1000:1000	0.4071	0.0000*	0.0000*
0, 0.25, 0.56, 1	ปกติ	5	1000:100	0.5637	0.0029*	0.0020*
0, 0.25, 0.56, 1	ปกติ	8	1000:1000	0.4809	0.0009*	0.0009*
0, 0.25, 0.56, 1	ปกติ	8	1000:100	0.5481	0.0084*	0.0091*
0, 0.25, 0.56, 1	เบ้ซ้าย	5	1000:1000	0.4717	0.0254*	0.0083*
0, 0.25, 0.56, 1	เบ้ซ้าย	5	1000:100	0.6437	0.0394*	0.0329*
0, 0.25, 0.56, 1	เบ้ซ้าย	8	1000:1000	0.4969	0.0400*	0.0181*
0, 0.25, 0.56, 1	เบ้ซ้าย	8	1000:100	0.6194	0.0516*	0.0419*
0, 0.25, 0.56, 1	เบ้ขวา	5	1000:1000	0.4629	0.0123*	0.0023*
0, 0.25, 0.56, 1	เบ้ขวา	5	1000:100	0.5977	0.0297*	0.0223*
0, 0.25, 0.56, 1	เบ้ขวา	8	1000:1000	0.4872	0.0388*	0.0200*
0, 0.25, 0.56, 1	เบ้ขวา	8	1000:100	0.6416	0.0413*	0.0356*

*สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ไม่เกิน 0.0500 (ร้อยละ 5)

จากตารางที่ 2 พบว่า ผลการวิเคราะห์อัตราความคลาดเคลื่อนประเภทที่ 1 (Type I Error Rate) ของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ พบว่า โดยรวมมีค่าระหว่าง 0 - 0.6437 หรือคิดเป็นร้อยละ 0 - 64.38 หากพิจารณาตามวิธีที่ใช้ในการประมาณค่า พบว่า วิธี ML มีค่าระหว่าง 0.3606 - 0.6437 หรือคิดเป็นร้อยละ 36.06 - 64.38 วิธี Bayes มีค่าระหว่าง 0 - 0.0516 หรือคิดเป็นร้อยละ 0 - 5.16 และวิธี Bayesγ มีค่าระหว่าง 0 - 0.0419 หรือร้อยละ 0 - 4.19 โดยวิธี Bayes และวิธี Bayesγ สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ทุกเงื่อนไข ส่วนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบสอบด้วยวิธี ML ไม่ผ่านตามเกณฑ์ที่กำหนดทุกเงื่อนไข

2.2 ผลการวิเคราะห์ค่าอำนาจการทดสอบ (Power Rate) ของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เป็นดังนี้

ตารางที่ 3 ผลการวิเคราะห์ค่าอำนาจการทดสอบ (Power Rate) ของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

TestletEff.	Dist.theta	NumDIF	R:F	ML	Bayes	Bayesγ
0.8, 0.8, 0.8, 0.8	ปกติ	5	1000:1000	0.9420*	0.1000	0.1240
0.8, 0.8, 0.8, 0.8	ปกติ	5	1000:100	0.7900	0.0980	0.0900
0.8, 0.8, 0.8, 0.8	ปกติ	8	1000:1000	0.9425*	0.1038	0.0763
0.8, 0.8, 0.8, 0.8	ปกติ	8	1000:100	0.7863	0.1138	0.0450
0.8, 0.8, 0.8, 0.8	เบ้ซ้าย	5	1000:1000	0.9800*	0.6600	0.7180
0.8, 0.8, 0.8, 0.8	เบ้ซ้าย	5	1000:100	0.8560*	0.3680	0.2800
0.8, 0.8, 0.8, 0.8	เบ้ซ้าย	8	1000:1000	0.9613*	0.6000	0.5275
0.8, 0.8, 0.8, 0.8	เบ้ซ้าย	8	1000:100	0.8113*	0.2250	0.1475
0.8, 0.8, 0.8, 0.8	เบ้ขวา	5	1000:1000	0.9720*	0.6640	0.7000
0.8, 0.8, 0.8, 0.8	เบ้ขวา	5	1000:100	0.8340*	0.2940	0.2320
0.8, 0.8, 0.8, 0.8	เบ้ขวา	8	1000:1000	0.9588*	0.5450	0.4988
0.8, 0.8, 0.8, 0.8	เบ้ขวา	8	1000:100	0.7800*	0.2875	0.1475
0, 0.25, 0.56, 1	ปกติ	5	1000:1000	0.9740*	0.2660	0.3740
0, 0.25, 0.56, 1	ปกติ	5	1000:100	0.7500	0.1000	0.1260
0, 0.25, 0.56, 1	ปกติ	8	1000:1000	0.9375*	0.1150	0.1550
0, 0.25, 0.56, 1	ปกติ	8	1000:100	0.7388	0.0975	0.1000
0, 0.25, 0.56, 1	เบ้ซ้าย	5	1000:1000	0.9400*	0.7200	0.7820
0, 0.25, 0.56, 1	เบ้ซ้าย	5	1000:100	0.8640*	0.3320	0.3900
0, 0.25, 0.56, 1	เบ้ซ้าย	8	1000:1000	0.9663*	0.5588	0.6863

TestletEff.	Dist.theta	NumDIF	R:F	ML	Bayes	Bayesγ
0, 0.25, 0.56, 1	เบ้ซ้าย	8	1000:100	0.8275*	0.2488	0.3063
0, 0.25, 0.56, 1	เบ้ขวา	5	1000:1000	0.9520*	0.6140	0.7480
0, 0.25, 0.56, 1	เบ้ขวา	5	1000:100	0.7460	0.2200	0.2860
0, 0.25, 0.56, 1	เบ้ขวา	8	1000:1000	0.9475*	0.5913	0.7025
0, 0.25, 0.56, 1	เบ้ขวา	8	1000:100	0.8063*	0.2638	0.2888

*อำนาจจำแนกตั้งแต่ 0.8000 ขึ้นไป (ร้อยละ 80)

จากตารางที่ 3 พบว่า ผลการวิเคราะห์ค่าอำนาจการทดสอบ (Power Rate) ของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ พบว่า โดยรวมมีค่าระหว่าง 0.0450-0.9800 หรือคิดเป็นร้อยละ 4.50-98 หากพิจารณาตามวิธีที่ใช้ในการประมาณค่า พบว่า วิธี ML มีค่าระหว่าง 0.7388-0.9800 หรือคิดเป็นร้อยละ 73.88 - 98 วิธี Bayes มีค่าระหว่าง 0.0975-0.7200 หรือคิดเป็นร้อยละ 9.75-72 และวิธี Bayesg ผลการศึกษาครั้งนี้ พบว่า การประมาณค่าพารามิเตอร์ความสามารถของผู้สอบจากแบบสอบที่มีลักษณะของเทสเลท ส่วนใหญ่วิธี Bayesg ในการวิเคราะห์ข้อสอบที่มีลักษณะของความเป็นเทสเลทผสมในแบบสอบควรมีการตรวจสอบข้อมูลก่อนการวิเคราะห์คุณภาพข้อสอบ เช่น หากข้อมูลความสามารถมีการแจกแจงแบบปกติ ควรประมาณค่าพารามิเตอร์ของข้อสอบโดยใช้วิธี Bayesgผลการศึกษาครั้งนี้ พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีลักษณะของความเป็นเทสเลทผสมในแบบสอบ ภายใต้เงื่อนไข 4 ปัจจัย แม้วิธี ML จะมีอำนาจการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูง แต่ไม่สามารถควบคุมอัตราความคาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ที่กำหนด ตรงข้ามกับวิธี Bayesg ควรศึกษารายละเอียดของตัวแปรเพิ่มเติม โดยให้มีค่าหลายหลายมากขึ้น เช่น ลักษณะของอิทธิพลของเทสเลท และอิทธิพลของการทำหน้าที่ต่างกันของข้อสอบ เป็นต้น

ควรศึกษาลักษณะของข้อมูลในด้านอื่นๆ ที่มีผลกระทบต่อการประมาณค่าพารามิเตอร์หรือการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เช่น แบบสอบที่ประกอบด้วยข้อมูลที่มีการสูญหาย (Missing data) ข้อมูลสุดโต่ง (Outlier) แบบสอบที่มีข้อมูลผสมกันของวิธีการให้คะแนนรายข้อแบบสองค่าและหลายค่า (Dichotomous and Polytomous data) เป็นต้น

แม้ว่าวิธี Bayesg ควรศึกษาลักษณะความเป็นเทสเลทที่มีผลต่อการประมาณค่าความสามารถในกรณีของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (Computerized Adaptive Testing: CAT) เนื่องจากการสุ่มเลือกข้อสอบอาจเกิดกรณีที่ผลการตอบของข้อสอบข้อหนึ่งมาจากโจทย์ของอีกข้อหนึ่ง (Cross-information) ซึ่งแสดงถึงความไม่เป็นอิสระของการตอบคำถามในแต่ละข้อ ทำให้เกิดการคำนวณค่าความสามารถของผู้สอบคาดเคลื่อน แล้วส่งผลต่อการยุติการสอบก่อนกำหนด

เอกสารอ้างอิง

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

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.

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

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.

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.

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

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.

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.