ประสิทธิภาพของการประมาณค่าพารามิเตอร์แบบเบส์โดยใช้การสรุปอ้างอิงความน่าเชื่อถือของโมเดลการตอบสนองข้อสอบ

The Efficiency of Bayesian Parameter Estimation Based on Generalizability in Item Response Modeling

ชนะศึก นิชานนท์*[1] ศิริชัย กาญจนวาสี[1] และ Mark Wilson[2]

[1]จุฬาลงกรณ์มหาวิทยาลัย

[2]University of California, Berkeley USA

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าพารามิเตอร์ของวิธีการสรุปอ้างอิงความน่าเชื่อถือของโมเดลการตอบสนองข้อสอบ (Generalizability in Item Response) 4 รูปแบบ ได้แก่ รูปแบบที่ 1 Original GIRM พัฒนาโดย Brigg and Wilson (2007) รูปแบบที่ 2 AGIRM A รูปแบบที่ 3 AGIRM B และรูปแบบที่ 4 Numerical Bayesian GIRM ผู้วิจัยเป็นผู้พัฒนาขึ้น นอกจากนี้ยังศึกษาถึงอิทธิพลของขนาดกลุ่มอย่างและจำนวนข้อสอบ รวมทั้งยังศึกษาความไว (Sensitivity) ของรูปแบบต่างๆ ต่อการกำหนดลักษณะการแจกแจงเริ่มแรกของค่าพารามิเตอร์ของข้อสอบและผู้สอบที่ส่งผลต่อประสิทธิภาพของวิธีการประมาณ ซึ่งวัดได้จากดัชนี 3 ประเภท ได้แก่ ความลำเอียงในการประมาณค่า (Biased estimator) คำนวณจากการวิเคราะห์ค่าความคลาดเคลื่อนเฉลี่ย (Mean Average Deviation) ความไม่แน่นอนในการประมาณค่า (Uncertainty estimator) คำนวณจากการวิเคราะห์ค่าส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation) และประสิทธิภาพในการประมาณค่าองค์ประกอบความแปรปรวนซึ่งคำนวณจากการวิเคราะห์ระยะทางยุคลิด (Euclidean Distance) ข้อมูลที่ใช้ในการศึกษาครั้งนี้ใช้ข้อมูลจำลอง (simulation) จากโปรแกรม R และทำการประมวลผลภายใต้จากการเขียนคำสั่งการประมวลผลด้วยโปรแกรม WinBUGS ด้วย Package R2 WinBUGS

ผลการวิจัยพบว่า เมื่อเปรียบเทียบประสิทธิภาพในการประมาณค่าของรูปแบบของวิธีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัดด้วยโมเดลการตอบสนองข้อสอบ (GIRM) พบว่า ความลำเอียงในการประมาณค่า รูปแบบที่ 1 กับรูปแบบที่ 4 ให้ค่าประสิทธิภาพสูงที่สุด โดยรูปแบบที่ 4 สามารถประมาณค่าพารามิเตอร์ได้เฉพาะลักษณะการแจกแจงเริ่มแรกของผู้สอบและข้อสอบแบบปกติ สำหรับความไม่แน่นอนในการประมาณค่า พบว่า รูปแบบที่ 4 ให้ค่าประสิทธิภาพสูงที่สุด สำหรับลักษณะการแจกแจงเริ่มแรกของผู้สอบและข้อสอบแบบปกติส่วนลักษณะการแจกแจกเริ่มแรกของค่าพารามิเตอร์ตัวใดตัวหนึ่งของผู้สอบหรือข้อสอบที่ไม่มีลักษณะการแจกแจงเริ่มแรกแบบปกติ พบว่า

* **ผู้ประสานงานหลัก** (Corresponding Author)

รูปแบบที่ 1 ให้ค่าประสิทธิภาพสูงที่สุด และเมื่อพิจารณาในด้านประสิทธิภาพขององค์ประกอบความแปรปรวนยุคลิด พบว่า รูปแบบที่ 2 ให้ค่าประสิทธิภาพสูงที่สุด การศึกษาอิทธิพลของขนาดกลุ่มตัวอย่างและจำนวนข้อสอบ พบว่า ขนาดกลุ่มตัวอย่าง ส่งผลต่อประสิทธิภาพในด้านความลำเอียงและความไม่แน่นอน ในการประมาณค่าในทุกรูปแบบ และด้านประสิทธิภาพขององค์ประกอบความแปรปรวนยุคลิด พบว่า ขนาดกลุ่มตัวอย่างไม่ส่งผลต่อประสิทธิภาพในการประมาณค่า สำหรับความยาวแบบสอบ พบว่า ส่งผลต่อการการวัดประสิทธิภาพในความลำเอียงในการประมาณค่าและการวิเคราะห์ประสิทธิภาพองค์ประกอบความแปรปรวนยุคลิด ในทุกรูปแบบ ส่วนความไม่แน่นอนในการประมาณค่า พบว่า ขนาดกลุ่มตัวอย่างส่งผลต่อประสิทธิภาพในการประมาณค่าเฉพาะรูปแบบที่ 3 เท่านั้น ในเรื่องการศึกษาความไว พบว่า การแจกแจงเริ่มแรกของผู้สอบ ไม่ส่งผลต่อประสิทธิภาพด้านความลำเอียง ส่วนความไม่แน่นอนในการประมาณค่าและประสิทธิภาพขององค์ประกอบความแปรปรวนยุคลิด พบว่าการแจกแจงเริ่มแรกของผู้สอบ ส่งผลต่อการวัดประสิทธิภาพความลำเอียงในทุกรูปแบบ ส่วนการแจกแจงเริ่มแรกของข้อสอบ พบว่า ส่งผลต่อการวัดประสิทธิภาพความลำเอียงในการประมาณค่าและความไม่แน่นอนในการประมาณค่าทุกรูปแบบ และส่งผลการวิเคราะห์ประสิทธิภาพองค์ประกอบความแปรปรวนยุคลิดเฉพาะในกรณีที่การแจกแจงเริ่มแรกของผู้สอบเป็นแบบแกมม่าเท่านั้น

คำสำคัญ : การวิเคราะห์ข้อสอบ การประมาณค่าพารามิเตอร์ วิธีการสรุปอ้างอิงความน่าเชื่อถือของโมเดลการตอบสนองข้อสอบ

Abstract

This research aimed to compare the efficiency of the 4 forms of GIRM; Form 1 Original GIRM developed by Brigg and Wilson (2007), Form 2 AGIRM A, Form 3 AGIRM B, and Form 4 Numerical Bayesian GIRM developed by the researcher. The research also investigated the impact of the sample size, the test length, and the sensitivity of each form toward the prior distribution setting of parameter of the items and the examiners which affected the efficiency of the estimation which could be measured by 3 types of indicators; Biased estimator calculated from Mean Average Deviation (MAD), Uncertainty estimator calculated from Standard Deviation (S.D.), and efficiency of constituent variance estimation calculated from Euclidean Distance (EUC). The data was simulation data generated using program R and was assessed with program WinBUGS with Package R2 WinBUGS. The results of research were as follows : 1.From the comparison of GIRM, it was found that when Bias estimated from MAD was considered, Form 1 Original GIRM and Form 4 Numerical Bayesian GIRM had the best efficiency. Form 4 could estimate normal prior distribution of the items and the examiners. For Uncertainty, it was found that Form 4 is the most efficiency for normal prior distribution. For the prior distribution of one parameter without normal prior distribution, Form 1 Original GIRM had the best efficiency. And when the efficiency of constituent variance estimation calculated from EUC was considered, it was found that Form 2 AGIRM A had the most efficiency. 2. From the study of the influence of the sample size and the test

length, it was found that the sample size affected the measure of bias efficiency in Form 1, Form 2, and Form 3. For the uncertainty in the estimation, it was found that the sample size affected all forms of efficiency estimation. And when the efficiency of constituent variance EUC was considered, it was found that the sample size did not affect efficiency estimation. The test length affected the bias efficiency and efficiency of constituent variance EUC in Form 1, Form 2, and Form 3. For uncertainty of estimation, it was found that the sample size affected all forms of efficiency estimation. 3. According to the sensitivity of the test analysis, it was found that the prior distribution of the examiners did not affect the measure of bias efficiency. For the uncertainty of estimation and analysis of constituent variance EUC, it was found that the prior distribution of the items affected the measure of bias efficiency. It was found that the prior distribution affected the measure of Bias estimator and all forms of uncertainty estimator. It also affected the efficiency of constituent variance EUC only when the prior distribution of the examiners was gamma.

Keywords : Item Analysis, Parameter Estimation, Generalization in Item Response Modeling

บทนำ

ปัจจุบันเราสามารถแบ่งโมเดลการวิเคราะห์ข้อสอบออกได้เป็น 2 ประเภทหลัก ได้แก่ การวิเคราะห์ข้อสอบตามหลักทฤษฎีการทดสอบแบบดั้งเดิม (Classical Test Theory; CTT) และทฤษฎีการทดสอบแนวใหม่ (Modern Test Theory; MTT) ซึ่งในปัจจุบันที่ใช้กันอย่างแพร่หลายมี 2 ทฤษฎี ได้แก่ ทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัด (Generalizability Theory; GT) และทฤษฎีการตอบสนองข้อสอบ (Item Response Theory; IRT) โดยทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัด (GT) เป็นทฤษฎีที่อธิบายแหล่งของความคลาดเคลื่อนที่เกิดจากหลายแหล่ง (Multiple sources of error) โดยใช้หลักการของการวิเคราะห์ความแปรปรวน (ANOVA) ส่วนทฤษฎีการตอบสนองข้อสอบเป็นทฤษฎีที่อธิบายความสัมพันธ์ระหว่างคุณลักษณะหรือความสามารถที่แท้จริงของบุคคลกับพฤติกรรมการตอบสนองข้อสอบในแต่ละข้อว่ามีความน่าจะเป็นในการตอบถูกได้มากน้อยเพียงใด อธิบายโดยใช้โค้งคุณลักษณะของข้อสอบ (Item Characteristic Curve; ICC) ซึ่งมีลักษณะเป็นฟังก์ชันทางคณิตศาสตร์ที่เรียกว่า ฟังก์ชันโลจิส (logistic function) หรือใกล้เคียงกับฟังก์ชันปกติสะสม (normal ogive function)

จากทฤษฎีการทดสอบแนวใหม่ ทั้งสองทฤษฎีสามารถทำให้ ผลการวิเคราะห์ข้อสอบมีความถูกต้องมากยิ่งขึ้น โดยทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัดเป็นการวิเคราะห์ที่สามารถอธิบายแหล่งความคลาดเคลื่อนของการวัดได้มากขึ้น และสามารถสรุปอ้างอิงไปยังสถานการณ์ต่างๆ ของการวัดในแต่ละแหล่งความคลาดเคลื่อนได้ ซึ่งอาจเรียกทฤษฎีนี้ได้ว่าเป็นโมเดลแบบสุ่ม (sampling model) (Brennan, 2001) แต่ทฤษฎียังไม่สามารถวิเคราะห์ค่าพารามิเตอร์ของข้อสอบเป็นรายข้อและพารามิเตอร์ของผู้สอบเป็นรายบุคคลได้ ในขณะที่ทฤษฎีการตอบสนองข้อสอบ สามารถให้ทั้งสารสนเทศที่เป็นค่าพารามิเตอร์ของข้อสอบเป็นรายข้อ (item parameter) และพารามิเตอร์ของผู้สอบเป็นรายบุคคล (person parameter)ได้ ซึ่งอาจเรียกทฤษฎีนี้ได้ว่าเป็นโมเดลการกำหนดมาตรวัด (scaling model) (Brennan, 2001)

จากจุดเด่นและข้อจำกัดในแต่ละทฤษฎี จึงทำให้นักวัดผลมีความพยายามในการรวมทั้งสองทฤษฎีเข้าไว้ด้วยกันเพื่อรวมจุดเด่นและลดข้อจำกัดในแต่ละทฤษฎี เพื่อการวิเคราะห์ข้อสอบที่มีความถูกต้องมากที่สุด (Kolen and Harris, 1987; Mislevy, 1993; Verhelst and Verstralen, 2001; Patz and others, 2002; Bock, Brennan and Muraki, 2002; Glas and Van Der Linder, 2003; Briggs and Wilson, 2007)

จนกระทั่งมาถึงในปัจจุบันได้มีการพัฒนาวิธีการสรุปอ้างอิงความน่าเชื่อถือของโมเดลการตอบสนองข้อสอบ (GIRM) (Briggs and Wilson, 2007) ซึ่งเป็นการรวมโมเดลการสุ่ม ของทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัดเข้าไปในโมเดลการกำหนดมาตรวัด (Scaling Model) ของทฤษฎีการตอบสนองข้อสอบ โดยอาศัยสมมติฐานของการกระจายไปยังแหล่งความคลาดเคลื่อนจากการวัดที่เกี่ยวข้อง โดยวิเคราะห์ด้วยวิธีการสรุปอ้างอิงความน่าเชื่อถือของโมเดลการตอบสนองข้อสอบ ไม่ต้องมีการปรับข้อมูลที่ขาดหายไป (missing data) ในเมทริกซ์ของข้อมูล ในขณะที่ทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัดต้องมีการออกแบบข้อมูลแทนที่ขาดหายไป ทั้งนี้เนื่องจากการประมาณค่าองค์ประกอบความแปรปรวนตั้งอยู่บนพื้นฐานของการคาดคะเน (expected) มากกว่าข้อมูลจริงที่อยู่ในตารางเมทริกซ์ของข้อมูล จึงทำให้ส่งผลต่อการวิเคราะห์ข้อมูล เพราะการกำหนดข้อมูลแทนที่ (specified) แบบไม่สมดุลสำหรับการออกแบบแบบสุ่ม (unbalanced random effects design) จะไม่สามารถประมาณค่าในขั้นตอนการศึกษาเชิงสรุปอ้างอิง (Generalizability Study; G-Study) และการศึกษาเชิงตัดสินใจ (Decision Study; D-study) ในโปรแกรม GENOVA[1] ได้

การประมาณค่าของวิธีการสรุปอ้างอิงความน่าเชื่อถือของโมเดลการตอบสนองข้อสอบจะประมาณค่าเมทริกซ์ที่คาดหวังโดยใช้เทคนิค MCMC (Markov Chain Monte Carlo) แบบ Metropolis-Hastings algorithm ภายในวิธีการ Gibbs sampler ซึ่งสามารถประมาณค่าการแจกแจงภายหลัง (Posterior distribution) ของความสามารถที่แท้จริงของผู้สอบแต่ละคน (θ) และค่าพารามิเตอร์ของข้อสอบแต่ละข้อ (β) สามารถแสดงสมการทางคณิตศาสตร์ได้ดังนี้

สมการสำหรับประมาณค่าความสามารถที่แท้จริงของผู้สอบ (θ)

$$P(\theta|\beta,X) = \frac{L(X|\theta,\beta)f(\theta)g(\beta)}{\int_{\theta} L(X|\theta,\beta)f(\theta)g(\beta)d\theta}$$

สมการสำหรับประมาณค่าพารามิเตอร์ของข้อสอบ (β)

$$P(\beta|\theta,X) = \frac{L(X|\theta,\beta)f(\theta)g(\beta)}{\int_{\beta} L(X|\theta,\beta)f(\theta)g(\beta)d\beta}$$

จากสมการ $P(\theta|\beta,X)$ และสมการ $P(\beta|\theta,X)$ ที่สามารถประมาณได้จากเทคนิค MCMC จะทำให้สามารถหาค่าของ $P(\theta_p,\beta_i)$ หรือความน่าจะเป็นในการตอบข้อสอบถูกของผู้สอบที่มีระดับความสามารถ θ_p ของข้อสอบที่มีระดับความยาก β_i ของแต่ละคนในแต่ละข้อของฟังก์ชันการตอบสนองข้อสอบโมเดลแบบราสช์ (Rasch model item response function) จากนั้นจึงเอา มาสร้างเป็นเมทริกซ์ที่คาดหวังประมาณองค์ประกอบความแปรปรวนและ

ค่าสัมประสิทธิ์การสรุปอ้างอิงความน่าเชื่อถือของการวัด โดยประมาณได้จากการวิเคราะห์ทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัดด้วยกรอบแนวคิดของทฤษฎีการตอบสนองข้อสอบที่ได้จากการเชื่อมโยงสูตรของ Kolen and Harris (1987) ดังนี้

$$\mu = \int_{\theta}\int_{\beta} P(\theta_p,\beta_i) f(\beta) g(\theta) d\beta d\theta \quad \text{(คำอธิบาย } \mu \text{ ใน GT)}$$

$$\pi(\theta) = \int_{\beta} P(\theta_p,\beta_i) f(\beta) d\beta - \mu \quad \text{(คำอธิบาย } v_p \text{ ใน GT)}$$

$$\iota(\theta) = \int_{\theta} P(\theta_p,\beta_i) g(\theta) d\theta - \mu \quad \text{(คำอธิบาย } v_i \text{ ใน GT)}$$

$$v(\theta,\beta) = P(\theta_p,\beta_i) - \pi(\theta) - \iota(\beta) - \mu \quad \text{(คำอธิบาย } v_{pi} \text{ ใน GT)}$$

$$\sigma^2(e) = \int_{\theta}\int_{\beta} P(\theta_p,\beta_i)[1-\mathrm{P}(\theta_p,\beta_i)] f(\beta) g(\theta) d\beta d\theta \quad \text{(คำอธิบาย } v_e \text{ ใน GT)}$$

ดังนั้นจะเห็นได้ว่าการวิเคราะห์ด้วยวิธีการสรุปอ้างอิงความน่าเชื่อถือของโมเดลการตอบสนองข้อสอบ (GIRM) สามารถวิเคราะห์ความแปรปรวนแยกระหว่างองค์ประกอบความแปรปรวนของปฏิสัมพันธ์ระหว่างบุคคล ($v(\theta,\beta)$) กับข้อสอบกับความคลาดเคลื่อนแบบสุ่ม ($\sigma^2(e)$) ได้ ในขณะที่การวิเคราะห์ตามทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัด (GT) ไม่สามารถแยกองค์ประกอบดังกล่าวได้ โดยคำนวณรวมองค์ประกอบความแปรปรวนของปฏิสัมพันธ์ระหว่างบุคคลกับข้อสอบกับความคลาดเคลื่อนแบบสุ่ม ($\sigma^2(pi,e)$) เข้าไว้ด้วยกัน

ในทางปฏิบัติค่าพารามิเตอร์ θ_p และ β_i เป็นค่าที่ไม่ทราบและจะต้องถูกประมาณขึ้น โดยสามารถประมาณค่าพารามิเตอร์ดังกล่าว ได้จากสมการการแจกแจงภายหลัง (Posterior distribution) ด้วยวิธีการประมาณค่า MCMC ของโมเดลการตอบสนองข้อสอบ ในแต่ละขั้นของการประมาณค่าด้วยวิธีการ Markov Chain จะทำให้สามารถประมาณค่า θ_p และ β_i ได้ เป็น $\hat{\theta}_p^{(m)}$ และ $\hat{\beta}_i^{(m)}$ ซึ่งค่าที่ได้จากการประมาณค่านี้จะเป็นองค์ประกอบในเมทริกซ์การตอบสนองที่คาดหวัง (Expected response matrix) และนำมาใช้ในการคำนวณตามสูตรของ Kolen and Harris จะทำให้ได้องค์ประกอบของความแปรปรวน $\hat{\sigma}^2(\pi)^{(m)}$ $\hat{\sigma}^2(i)^{(m)}$ $\hat{\sigma}^2(\gamma)^{(m)}$ และ $\hat{\sigma}^2(e)^{(m)}$ ดังนั้นจะเห็นได้ว่าการประมาณค่าตามวิธีการดังกล่าวของวิธีการวิเคราะห์ข้อสอบด้วยวิธีการสรุปอ้างอิงความน่าเชื่อถือของโมเดลการตอบสนองข้อสอบ (GIRM) นำไปสู่การฝ่าฝืนข้อตกลงเบื้องต้นแบบสุ่มของการวิเคราะห์ ทั้งนี้สาเหตุน่าจะมาจากการป้อนคำสั่งในกระบวนการการประมาณค่าความแปรปรวนขององค์ประกอบความคลาดเคลื่อนของขั้นตอนการประมาณด้วยวิธี Markov Chain ไม่สามารถทำให้เป็นอิสระจากกันได้ จึงนำไปสู่การฝ่าฝืนข้อตกลงเบื้องต้นซึ่งมีต่อผลกระทบแบบสุ่ม (random effects) เช่น การประมาณค่าองค์ประกอบความแปรปรวนของ $\sigma^2(i)$ ตามวิธีการสรุปอ้างอิงความน่าเชื่อถือของโมเดลการตอบสนองข้อสอบ (GIRM) แบบ p X i design ด้วยวิธีการ Markov Chain เพื่อนำมาสร้างเป็นเมทริกซ์คะแนนที่สังเกตได้ของบุคคล ในการคำนวณดังกล่าวต้องการคำนวณค่า $\sigma^2(i)$ ในการประมาณค่าจะต้องมีการเปลี่ยนแปลงเฉพาะคอลัมน์ในเมทริกซ์ โดยค่าในแถวของเมทริกซ์จะต้องคงที่ แต่ในการวิเคราะห์ด้วยโมเดลการตอบสนองข้อสอบนั้นใช้การประมาณค่าด้วยวิธี Markov Chain จะทำให้ทั้งคอลัมน์ และแถวในเมทริกซ์มีค่าที่ผันแปรร่วมกัน จึงทำให้การประมาณพารามิเตอร์ของวิธีการดังกล่าวมีแนวโน้มในการเกิดความไม่แน่นอนในการประมาณค่า (Briggs and Wilson, 2007)

ดังนั้นในการศึกษาครั้งนี้ ผู้วิจัยจึงสนใจศึกษาวิธีการประมาณค่าของวิธีการวิเคราะห์ข้อสอบด้วยวิธีการสรุป

อ้างอิงความน่าเชื่อถือของโมเดลการตอบสนองข้อสอบ เพื่อไม่ให้นำไปสู่การฝ่าฝืนข้อตกลงเบื้องต้นแบบสุ่ม โดยศึกษารูปแบบการประมาณค่าวิธีการวิเคราะห์ข้อสอบของวิธีการสรุปอ้างอิงความน่าเชื่อถือของโมเดลการตอบสนองข้อสอบ (GIRM) 4 รูปแบบ ซึ่งได้แก่ รูปแบบที่ 1 Original GIRM ซึ่งเป็นรูปแบบการประมาณค่าเริ่มแรกของวิธีการ GIRM ที่ Briggs and Wilson (2007) ได้พัฒนาขึ้นสำหรับรูปแบบที่ 2 Alternative GIRM A รูปแบบที่ 3 Alternative GIRM B ทั้งสองรูปแบบเช่นรูปแบบที่ผู้วิจัยปรับวิธีการประมาณค่าพารามิเตอร์ขึ้นมาใหม่ โดยการนำค่าการแจกแจงภายหลัง (Posterior distribution) ที่ประมาณค่าพารามิเตอร์ (θ_p) และ (β_i) ได้มาจำลองเมทริกซ์ค่าสังเกตใหม่ด้วยการคำนวณคะแนนการแจกแจงการทำนายภายหลัง (Posterior predictive distribution) ซึ่งเป็นค่าเฉลี่ยของคะแนนที่สังเกตได้โดยได้มาจากการวิเคราะห์ได้ใช้เทคนิค MCMC (Markov Chain Monte Carlo) รูปแบบที่ 4 Numerical Bayesian GIRM ดังนั้นจะทำให้ได้ Predictive Observed Matrix แทนเมทริกซ์การตอบสนองที่คาดหวัง (Expected Response Matrix) ที่ได้จากการประมาณค่าด้วยรูปแบบที่ 1 Original GIRM ซึ่งน่าจะช่วยแก้ไขปัญหา การประมาณค่าพารามิเตอร์ที่ไม่เป็นอิสระต่อกันระหว่างแถวและคอลัมน์ในเมทริกซ์การตอบสนองที่คาดหวัง และส่งผลทำให้การประมาณค่าที่ได้จากการวิเคราะห์ในรูปแบบที่ 1 เกิดความไม่แน่นอนในการประมาณค่า โดยรูปแบบที่ 2 และรูปแบบที่ 3 จะมีลักษณะต่างกันเฉพาะในขั้นตอนการประมาณค่าองค์ประกอบความแปรปรวน (variance components) โดยรูปแบบที่ 2 จะใช้การประมาณด้วยการวิเคราะห์ความแปรปรวน (ANOVA) เหมือนกับการวิเคราะห์ในทฤษฎี GT ส่วนรูปแบบที่ 3 จะประมาณค่าด้วยสูตรของ Kolen and Harris (1987) เช่นเดียวกับการประมาณค่าในรูปแบบที่ 1 และรูปแบบที่ 4 เป็นรูปแบบที่ผู้วิจัยประยุกต์ใช้การประมาณค่าแบบเบส์ในการประมาณค่าพารามิเตอร์ของ IRT ที่ Swaminathan and Gifford (1982, 1985, 1986) พัฒนาขึ้น ซึ่งเป็นการประมาณค่าที่ไม่ได้ใช้เทคนิค MCMC ซึ่งน่าจะช่วยแก้ไขปัญหาการประมาณค่าพารามิเตอร์ที่ไม่เป็นอิสระต่อกันระหว่างแถวและคอลัมน์ได้ โดยในการศึกษาครั้งนี้ทำการศึกษาประสิทธิภาพของการประมาณค่าใน 3 มิติ ได้แก่ มิติที่ 1 ความลำเอียงในการประมาณค่า (Biased estimator) คำนวณได้จากการวิเคราะห์ค่าความคลาดเคลื่อนเฉลี่ย (Mean Average Deviation : MAD) มิติที่ 2 ความไม่แน่นอนในการประมาณค่า (Uncertainty estimator) ซึ่งคำนวณได้จากการวิเคราะห์ค่าส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation : S.D.) และมิติที่ 3 ประสิทธิภาพในการประมาณค่าองค์ประกอบความแปรปรวน ซึ่งคำนวณได้จากการวิเคราะห์ค่าองค์ประกอบความแปรปรวนระยะทางยุคลิด (Euclidean Distance : EUC) ในการศึกษาครั้งนี้ จึงสนใจศึกษาวิธีการประมาณค่าของวิธีการวิเคราะห์ข้อสอบด้วยวิธีการสรุปอ้างอิงความน่าเชื่อถือของโมเดลการตอบสนองข้อสอบ เพื่อไม่ให้นำไปสู่การฝ่าฝืนข้อตกลงเบื้องต้นแบบสุ่ม โดยศึกษารูปแบบการประมาณค่าวิธีการวิเคราะห์ข้อสอบของวิธีการสรุปอ้างอิงความน่าเชื่อถือของโมเดลการตอบสนองข้อสอบ 4 รูปแบบ โดยวัดได้จากประสิทธิภาพของการประมาณค่าทั้ง 3 มิติ รวมถึงสรรหาอิทธิพลของกลุ่มตัวอย่างและจำนวนข้อสอบที่ส่งผลต่อประสิทธิภาพนอกจากนี้ยังศึกษาความไวของรูปแบบต่างๆที่มีต่อลักษณะการแจกแจงเริ่มแรกของค่าพารามิเตอร์ของข้อสอบและผู้สอบด้วย

วัตถุประสงค์

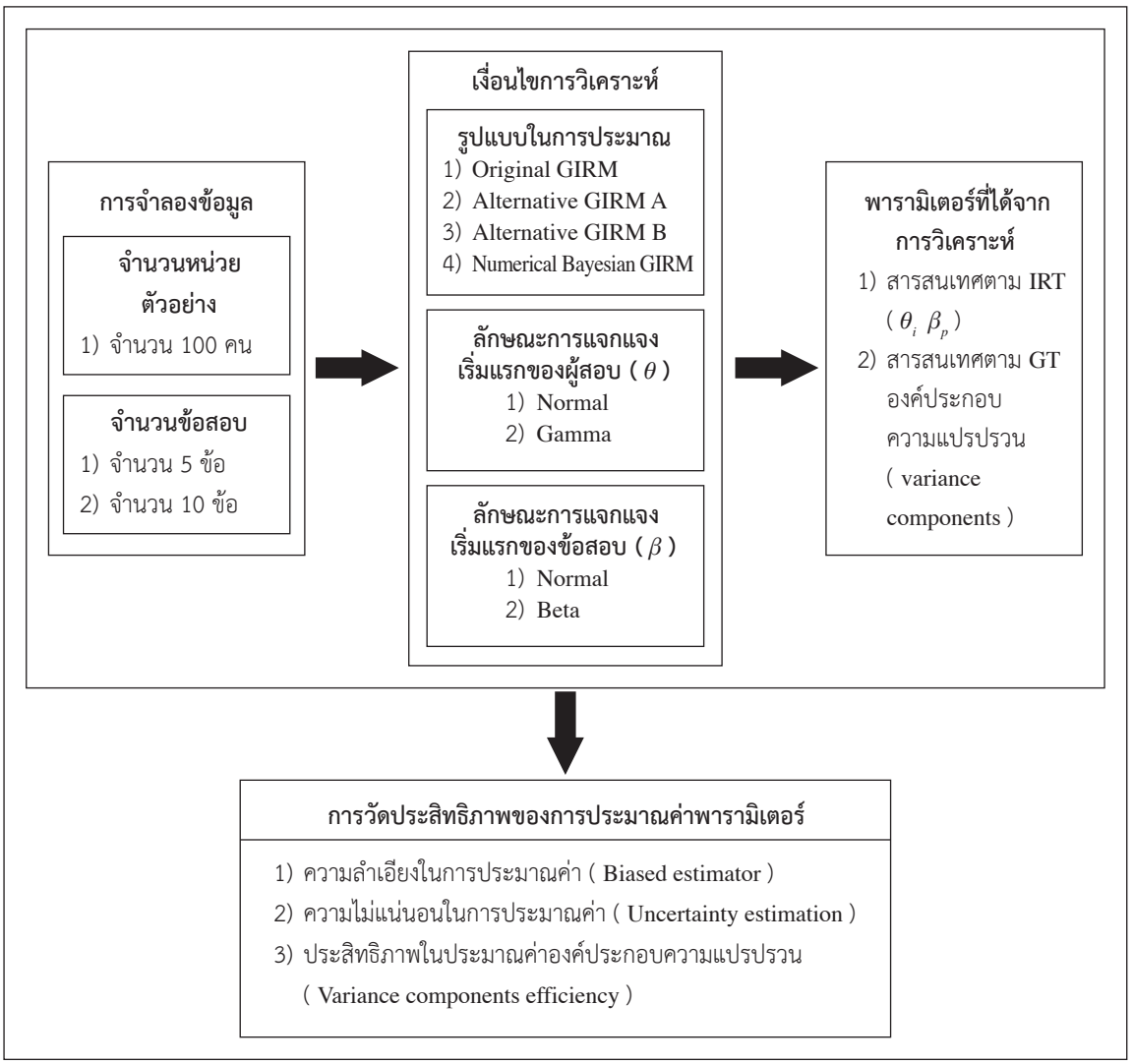
1. เพื่อเปรียบเทียบประสิทธิภาพของวิธีวิเคราะห์ข้อสอบด้วยวิธีการสรุปอ้างอิงความน่าเชื่อถือของโมเดลการตอบสนองข้อสอบ (GIRM) 4 รูปแบบ ระหว่างรูปแบบที่ 1 Original GIRM รูปแบบที่ 2 Alternative GIRM A

รูปแบบที่ 3 Alternative GIRM B และรูปแบบที่ 4 Numerical Bayesian GIRM

2. เพื่อศึกษาอิทธิพลของขนาดกลุ่มตัวอย่างและจำนวนข้อสอบที่มีต่อประสิทธิภาพของวิธีวิเคราะห์ข้อสอบด้วยวิธีการสรุปอ้างอิงความน่าเชื่อถือของโมเดลการตอบสนองข้อสอบ 4 รูปแบบ

3. เพื่อศึกษาความไว (Sensitivity) ของผลการวิเคราะห์ข้อสอบด้วยวิธีการสรุปอ้างอิงความน่าเชื่อถือของโมเดลการตอบสนองข้อสอบ 4 รูปแบบที่มีต่อการกำหนดลักษณะการแจกแจงเริ่มแรกของค่าพารามิเตอร์ของข้อสอบ (β) และลักษณะการแจกแจงเริ่มแรกของค่าพารามิเตอร์ของผู้สอบ (θ)

กรอบแนวคิด



ระเบียบวิธีวิจัย

การศึกษาในครั้งนี้ใช้ข้อมูลจำลองที่ได้จากโปรแกรม R และทำการประมวลผลภายใต้จากการเขียนคำสั่งการประมวลผลด้วยโปรแกรม WinBUGS และกลับมาแสดงผลจะแสดงผลในโปรแกรม R โดยใช้โปรแกรมร่วมกันระหว่างโปรแกรม R และโปรแกรม WinBUGS[2] ด้วย Package R2 WinBUGS[3] โดยมีขั้นตอนการวิจัย ดังนี้

1. การศึกษาเอกสารและงานวิจัยทั้งในและต่างประเทศที่เกี่ยวกับแนวคิด วิธีการและจุดเด่น ข้อจำกัดในแต่ละทฤษฎีการวิเคราะห์ข้อสอบ ซึ่งได้แก่ ทฤษฎีการทดสอบแบบดั้งเดิม ทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัด ทฤษฎีการตอบสนองข้อสอบ และวิธีการสรุปอ้างอิงความน่าเชื่อถือของโมเดลการตอบสนองข้อสอบ

2. การศึกษาแนวคิด หลักการ วิธีการใช้งาน รวมทั้งวิธีการเขียนคำสั่งสำหรับการวิเคราะห์ด้วยโปรแกรม R โปรแกรม WinBUGS และ Package R2 WinBUGS

3. การจำลองข้อมูลด้วยโปรแกรม R จากการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง พบว่า ตัวแปรที่ส่งผลต่อการประมาณค่าพารามิเตอร์ของข้อสอบ ได้แก่ จำนวนหน่วยตัวอย่างในการวิเคราะห์ ข้อมูลและจำนวนข้อสอบที่ใช้ในการวิเคราะห์ข้อมูล ดังนั้นในการจำลองข้อมูล จึงมีเงื่อนไขในการจำลองข้อมูล 2 ตัวแปร ได้แก่ จำนวนกลุ่มตัวอย่างในการวิเคราะห์ 2 เงื่อนไข ได้แก่ 100 คนและ 300 คน จำนวนข้อสอบที่ใช้ในการวิเคราะห์ข้อมูล 2 เงื่อนไข ได้แก่ 5 ข้อ และ 10 ข้อ รวมเงื่อนไขในการจำลองข้อมูลทั้งสิ้น 4 เงื่อนไข

4. การสร้างโมเดลการวิเคราะห์ข้อมูล โดยการศึกษาครั้งนี้ ตัวแปรที่เกี่ยวข้องกับการวิเคราะห์ มีจำนวน 3 ตัวแปร ได้แก่ ตัวแปรรูปแบบการวิเคราะห์ข้อสอบของวิธีการสรุปอ้างอิงความน่าเชื่อถือของโมเดลการตอบสนองข้อสอบ 4 รูปแบบ ประกอบด้วย รูปแบบที่ 1 ซึ่งเป็นรูปแบบต้นแบบที่ Briggs and Willson (2007) พัฒนาขึ้น รูปแบบที่ 2 และรูปแบบที่ 3 เป็นรูปแบบที่ผู้วิจัยปรับวิธีการประมาณค่าพารามิเตอร์ขึ้นมาใหม่ และรูปแบบที่ 4 เป็นรูปแบบที่ผู้วิจัยประยุกต์ใช้การประมาณค่าแบบเบส์ในการประมาณค่าพารามิเตอร์ของ IRT ที่ Swaminathan and Gifford (1982, 1985, 1986) พัฒนาขึ้น ตัวแปรลักษณะการแจกแจงเริ่มแรกของค่าพารามิเตอร์ของผู้สอบ (θ) 2 ลักษณะ ประกอบด้วย ลักษณะการแจกแจงแบบปกติ และลักษณะการแจกแจงแบบแกมมา และตัวแปรลักษณะการแจกแจงเริ่มแรกของค่าพารามิเตอร์ของข้อสอบ (β) 2 ลักษณะ ประกอบด้วย ลักษณะการแจกแจงแบบปกติ และลักษณะการแจกแจงแบบเบต้า รวมโมเดลการวิเคราะห์ที่ผู้วิจัยต้องตรวจสอบทั้งสิ้น 16 เงื่อนไข (4 รูปแบบการวิเคราะห์ × 2 ลักษณะการแจกแจงเริ่มแรกของค่าพารามิเตอร์ของผู้สอบ (θ) 2 ลักษณะการแจกแจงเริ่มแรกของค่าพารามิเตอร์ของข้อสอบ (β))

5. การตรวจสอบข้อมูลที่จำลองขึ้นกับโมเดลการวิเคราะห์ข้อมูลด้วยโปรแกรม WinBUGS ภายใต้การเขียนคำสั่ง การประมวลผล และกลับมาแสดงผลในโปรแกรม R โดยใช้โปรแกรมร่วมกันระหว่างโปรแกรม R และโปรแกรม WinBUGS ด้วยโปรแกรม Package R2 WinBUGS ดังนั้นจะมีเงื่อนไขที่ผู้วิจัยต้องตรวจสอบทั้งสิ้น 64 เงื่อนไข (ข้อมูลจำลอง 4 เงื่อนไข × โมเดลการวิเคราะห์ 16 เงื่อนไข) การวิเคราะห์สำหรับการวิจัยในครั้งนี้ มีการวิเคราะห์ข้อมูลที่ได้จากการสุ่มโดยการทำกลุ่มข้อมูลซ้ำ 500 ชุดข้อมูล โดยการประมาณค่าซ้ำในแต่ละชุดข้อมูล 10,000 รอบ (number of iteration) หลังจากตัดข้อมูลในส่วนแรกออก 1,000 รอบ (number of burnin) เนื่องจากการประมาณค่าในรอบแรกๆ ค่าพารามิเตอร์ที่ต้องการการประมาณค่ายังมีการแกว่งตัวไม่ลู่เข้าค่าใดค่าหนึ่ง

6. การวัดประสิทธิภาพของการประมาณค่า ทำการวัดประสิทธิภาพตัวประมาณค่า 3 ดัชนี ได้แก่ ดัชนีแรก

ความลำเอียงในการประมาณค่า (Biased estimation) ซึ่งคำนวณได้จากการวิเคราะห์ค่าความคลาดเคลื่อนเฉลี่ย (Mean Average Deviation; MAD) ดัชนีที่สอง ความไม่แน่นอนในการประมาณค่า (Uncertainty estimation) ซึ่งคำนวณได้จากการวิเคราะห์ค่าส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation; S.D.) และดัชนีสุดท้าย ประสิทธิภาพในการประมาณค่าองค์ประกอบความแปรปรวน ซึ่งคำนวณได้จากการวิเคราะห์ค่าองค์ประกอบความแปรปรวนระยะทางยุคลิด (Euclidean Distance; EUC)

สรุปผลการวิจัย

1. ผลการวิเคราะห์เพื่อเปรียบเทียบประสิทธิภาพของตัวประมาณค่าในแต่ละรูปแบบการประมาณค่าพารามิเตอร์ของวิธีการสรุปอ้างอิงความน่าเชื่อถือของโมเดลการตอบสนองข้อสอบ 4

รูปแบบเมื่อพิจารณาความลำเอียงในการประมาณค่าโดยพิจารณาจากค่า MAD รูปแบบที่ 1 กับรูปแบบที่ 4 มีประสิทธิภาพสูงที่สุด แต่รูปแบบที่ 4 เป็นรูปแบบที่ประมาณค่าพารามิเตอร์ได้เฉพาะลักษณะการแจกแจงเริ่มแรกแบบปกติ เท่านั้น โดยเมื่อพิจารณาในแต่ละกรณีพบว่า รูปแบบที่ 1สามารถประมาณค่าได้มีประสิทธิภาพในกรณีที่ลักษณะการแจกแจงเริ่มแรกของผู้สอบแบบปกติ และลักษณะการแจกแจงเริ่มแรกของข้อสอบแบบเบต้า ส่วนรูปแบบที่ 2 และรูปแบบที่ 3 จะประมาณค่าได้มีประสิทธิภาพในกรณีที่ลักษณะการแจกแจงเริ่มแรกของผู้สอบแบบแกมมา และลักษณะการแจกแจงเริ่มแรกของข้อสอบแบบเบต้า ส่วนรูปแบบที่ 4 ประมาณค่าได้มีประสิทธิภาพในกรณีที่ลักษณะการแจกแจงเริ่มแรกของผู้สอบและข้อสอบแบบปกติ เมื่อพิจารณาประสิทธิภาพในการประมาณค่าด้านความไม่แน่นอนในการประมาณค่า พบว่า รูปแบบที่ 4 มีประสิทธิภาพสูงที่สุด สำหรับลักษณะการแจกแจงเริ่มแรกแบบปกติ ส่วนถ้าลักษณะการแจกแจงเริ่มแรกของค่าพารามิเตอร์ตัวหนึ่งตัวใดไม่ได้มีลักษณะการแจกแจงเริ่มแรกแบบปกติ รูปแบบที่ 1 มีประสิทธิภาพสูงที่สุด และเมื่อพิจารณาประสิทธิภาพในการประมาณค่าด้านประสิทธิภาพขององค์ประกอบความแปรปรวนยุคลิด พบว่า รูปแบบที่ 2 มีประสิทธิภาพสูงที่สุด สามารถสรุปได้ดังตารางที่ 1

ตารางที่ 1 ผลการสรุปประสิทธิภาพของการประมาณค่าในแต่ละดัชนีชี้วัดจำแนกตามรูปแบบการประมาณค่าพารามิเตอร์ และกรณีศึกษาทั้ง 16 กรณี

กรณีศึกษา**	การวัดประสิทธิภาพของการประมาณค่า											
	MAD				SD				EUC			
	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
Nor-Nor-300-5				*				*		*		
Nor-Nor-300-10	*							*		*		
Nor-Nor-100-5				*				*		*		
Nor-Nor-100-10	*							*		*		
Nor-Beta-300-5	*			NA	*			NA	*		NA	
Nor-Beta-300-10	*			NA	*			NA	*		NA	
Nor-Beta-100-5	*			NA	*			NA	*		NA	
Nor-Beta-100-10		*	*	NA	*			NA	*		NA	
Gam-Nor-300-5		*	*	NA	*			NA	*		NA	
Gam-Nor-300-10		*	*	NA	*			NA	*		NA	
Gam-Nor-100-5	*			NA	*			NA	*		NA	
Gam-Nor-100-10	*			NA	*			NA	*		NA	
Gam-Beta-300-5		*	*	NA	*			NA	*		NA	
Gam-Beta-300-10		*	*	NA	*			NA	*		NA	
Gam-Beta-100-5	*			NA	*			NA	*		NA	
Gam-Beta-100-10		*	*	NA	*			NA	*		NA	
รวม	8	6	6	2	12	0	0	4	0	16	0	0
	(50%)	(37.5%)	(37.5%)	(50%)	(75%)	(0%)	(0%)	(100%)	(0%)	(100%)	(0%)	(0%)

NA ไม่สามารถวิเคราะห์ได้

* เป็นวิธีการที่ให้ประสิทธิภาพดีสุดในแต่ละเงื่อนไข

** กรณีศึกษา หมายถึง การแจกแจงเริ่มแรกของผู้สอบ การแจกแจงเริ่มแรกของข้อสอบ จำนวนผู้สอบ จำนวนข้อสอบ ยกตัวอย่าง เช่น กรณีศึกษา Gam-Nor-300-10 หมายถึง การแจกแจงเริ่มแรกของผู้สอบมีลักษณะแบบแกมมา การแจกแจงเริ่มแรกของข้อสอบแบบปกติ จำนวนหน่วยตัวอย่าง 300 คน และจำนวนข้อสอบ 10 ข้อ

2. ผลการวิเคราะห์ความแตกต่างของขนาดกลุ่มตัวอย่างที่มีต่อประสิทธิภาพในการประมาณค่าพารามิเตอร์ ขนาดของกลุ่มตัวอย่างส่งผลต่อประสิทธิภาพในด้านความลำเอียงในการประมาณค่าเกือบทุกรูปแบบ ยกเว้นรูปแบบที่ 4 ส่วนความไม่แน่นอนในการประมาณค่า พบว่า ขนาดกลุ่มตัวอย่างส่งผลต่อประสิทธิภาพในการประมาณค่าในทุกรูปแบบ และเมื่อมองในเรื่องของการวิเคราะห์ประสิทธิภาพขององค์ประกอบความแปรปรวนยุคลิด พบว่า ขนาดของกลุ่มตัวอย่างไม่ส่งผลประสิทธิภาพในการประมาณค่า เมื่อพิจารณาในเรื่องของความยาวของแบบสอบ พบว่า ส่งผลต่อการวัดประสิทธิภาพในด้านความลำเอียงในการประมาณค่าและการวิเคราะห์ประสิทธิภาพขององค์ประกอบความแปรปรวนยุคลิด เกือบทุกรูปแบบ ยกเว้นรูปแบบที่ 4 ส่วนความไม่แน่นอนในการประมาณค่า พบว่า ขนาดกลุ่มตัวอย่างส่งผลต่อประสิทธิภาพในการประมาณค่าในทุกรูปแบบ สามารถสรุปได้ดังตารางที่ 2

ตารางที่ 2 ผลการศึกษาอิทธิพลของขนาดตัวอย่าง ความยาวของแบบสอบ การแจกแจงเริ่มแรกของผู้สอบ และการแจกแจงเริ่มแรกของข้อสอบจำแนกตามตัววัดประสิทธิภาพและรูปแบบการประมาณค่า

ประสิทธิภาพ	รูปแบบการประมาณค่า	จำนวนหน่วยตัวอย่าง	ความยาวของแบบสอบ	การแจกแจงเริ่มแรกของผู้สอบ	การแจกแจงเริ่มแรกของข้อสอบ
MAD	Original GIRM	✓	✓	X	✓
	AGIRM A	✓	✓	X	✓
	AGIRM A	✓	✓	X	✓
	Numerical GIRM	X	X	X	✓
SD	Original GIRM	✓	✓	✓	✓
	AGIRM A	✓	✓	✓	✓
	AGIRM A	✓	✓	✓	✓
	Numerical GIRM	✓	✓	✓	✓
EUC	Original GIRM	X	✓	✓	*
	AGIRM A	X	✓	✓	*
	AGIRM A	X	✓	✓	*
	Numerical GIRM	X	X	NA	NA

✓ ส่งผลต่อการประมาณค่าประสิทธิภาพ

X ไม่ส่งผลต่อการประมาณค่าประสิทธิภาพ

NA ไม่สามารถทำการวิเคราะห์ได้

* เฉพาะการแจกแจงเริ่มแรกของผู้สอบแบบแกมมา การระบุการแจกแจงเริ่มแรกของข้อสอบจึงจะส่งผลต่อการประมาณค่า

3. ผลการวิเคราะห์ความไว (Sensitivity) ของผลการประมาณค่าของวิธีการสรุปอ้างอิงความน่าเชื่อถือของโมเดลการตอบสนองข้อสอบที่มีต่อการกำหนดลักษณะการแจกแจงเริ่มแรกของค่าพารามิเตอร์ของผู้สอบและข้อสอบที่มีต่อประสิทธิภาพในการประมาณค่าพารามิเตอร์พบว่าการแจกแจงเริ่มแรกของผู้สอบ พบว่า ไม่ส่งผลต่อการวัดประสิทธิภาพในด้านความลำเอียง ส่วนในเรื่องของส่วนความไม่แน่นอนในการประมาณค่าและการวิเคราะห์ประสิทธิภาพขององค์ประกอบความแปรปรวนยุคลิด พบว่าการแจกแจงเริ่มแรกของผู้สอบ ส่งผลต่อการวัดประสิทธิภาพในด้านความลำเอียงในรูปแบบ ส่วนในเรื่องการแจกแจงเริ่มแรกของข้อสอบ พบว่า ส่งผลต่อการวัดประสิทธิภาพในด้านความลำเอียงในการประมาณค่าและความไม่แน่นอนในการประมาณค่าในทุกรูปแบบ และส่งผลการวิเคราะห์ประสิทธิภาพขององค์ประกอบความแปรปรวนยุคลิดเฉพาะในกรณีที่การแจกแจงเริ่มแรกของผู้สอบเป็นแบบแกมมาเท่านั้น

อภิปรายผลการวิจัย

1. การประมาณค่าพารามิเตอร์ตามแนวการวิเคราะห์ด้วยทฤษฎีการตอบสนองข้อสอบด้วยวิธีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัดด้วยโมเดลการตอบสนองข้อสอบ จากการวิเคราะห์ค่าพารามิเตอร์ตามแนวทฤษฎีการตอบสนองข้อสอบของรูปแบบการประมาณของวิธีการ GIRM พบว่า ค่าพารามิเตอร์ของข้อสอบและค่าพารามิเตอร์ของผู้สอบด้วยวิธีการ MCMC ซึ่งใช้ในการประมาณค่าของวิธีการ GIRM ในรูปแบบที่ 1 รูปแบบที่ 2 และรูปแบบที่ 3 ให้ค่าเฉลี่ยของพารามิเตอร์ของข้อสอบและพารามิเตอร์ของผู้สอบไม่แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05 กับการประมาณค่าพารามิเตอร์ด้วยวิธีการ Numerical bayesian method ในรูปแบบที่ 4 แสดงว่า การประมาณค่าพารามิเตอร์ของผู้สอบทั้งสองวิธีให้ผลที่ไม่แตกต่างกัน แต่เมื่อพิจารณาค่าส่วนเบี่ยงเบนมาตรฐาน พบว่า ค่าส่วนเบี่ยง-เบนมาตรฐานที่ได้จากการประมาณค่าด้วยวิธีการ Numerical bayesian method ในรูปแบบที่ 4 ให้ค่าที่ต่ำกว่าค่าส่วนเบี่ยงเบนมาตรฐานที่ได้จากการประมาณค่าวิธีการ MCMC ในรูปแบบที่ 1 รูปแบบที่ 2 และ รูปแบบที่ 3 แสดงให้เห็นว่า ค่าเฉลี่ยของพารามิเตอร์ที่ได้จากการประมาณค่าตามแนวแนวทฤษฎีการตอบสนองข้อสอบไม่แตกต่างกันระหว่างการประมาณค่าด้วยวิธี MCMC กับวิธี Numerical แต่การประมาณค่าด้วยวิธีการ Numerical มีความแน่นอนในการประมาณค่ามากกว่าโดยพิจารณาจากค่าส่วนเบี่ยงเบนมาตรฐาน ทั้งนี้สาเหตุเป็นเพราะในการประมาณค่าด้วยวิธีการประมาณค่าด้วยวิธีการ MCMC ของการประมาณค่าแบบเบส์นั้นการประมาณค่าพารามิเตอร์ของผู้สอบจะขึ้นอยู่กับการประมาณค่าของข้อสอบ และการประมาณค่าของข้อสอบจะขึ้นอยู่กับการประมาณค่าผู้สอบ จึงทำให้ค่าพารามิเตอร์ทั้งสองไม่เป็นอิสระต่อกัน ซึ่งฝ่าฝืนข้อตกลงเบื้องต้นในการประมาณค่าของค่าพารามิเตอร์ของทฤษฎีการตอบสนองข้อสอบ เกี่ยวกับความเป็นอิสระของแบบสอบ (Local Independent)

นอกจากนี้เมื่อพิจารณาการประมาณค่าพารามิเตอร์ของข้อสอบและผู้สอบตามแนวทฤษฎีการตอบสนองของข้อสอบ พบว่า สามารถประมาณค่าพารามิเตอร์ได้เฉพาะกรณีการแจกแจงเริ่มแรกผู้สอบและการแจกแจงเริ่มแรกของข้อสอบมีลักษณะการกระจายแบบโค้งปกติเท่านั้น ทั้งนี้เนื่องมาจากการประมาณค่าพารามิเตอร์ของรูปแบบที่ 1 รูปแบบที่ 2 และรูปแบบที่ 3 ใช้การประมาณค่าด้วยวิธีการ MCMC ซึ่งเป็นการประมาณค่าพารามิเตอร์โดยการสุ่มค่าพารามิเตอร์ขึ้นมาทำซ้ำในหลายๆ ครั้ง ซึ่งต่างจากการประมาณค่าพารามิเตอร์ของรูปแบบที่ 4 ซึ่งใช้การประมาณค่าด้วยวิธีการ Numerical method ซึ่งใช้วิธีการทำให้ฟังก์ชันการแจกแจงภายหลังมีค่าสูงสุด จากการหาอนุพันธ์ของ

ฟังก์ชันและกำหนดให้ค่าของอนุพันธ์มีค่าเท่ากับศูนย์ ซึ่งการประมาณค่าในลักษณะนี้ถ้าลักษณะการแจกแจงไม่ใช่รูปแบบการแจกแจงที่เข้ากันได้ เช่น ลักษณะการแจกแจงแบบแกมมา กับลักษณะการแจกแจงแบบเบต้า จะทำให้ไม่สามารถมีการดำเนินการทางคณิตศาสตร์ (Operation) ได้ (Ntzoufras, 2009) จึงเป็นเหตุผลที่รูปแบบการประมาณค่าในรูปแบบที่ 4 จึงสามารถประมาณค่าพารามิเตอร์ได้เฉพาะลักษณะการแจกแจงเริ่มแรกของข้อสอบแบบปกติกับลักษณะการแจกแจงเริ่มแรกของผู้สอบแบบปกติเท่านั้น

ดังนั้นจะเห็นได้ว่าในเรื่องการประมาณค่าพารามิเตอร์ตามแนวทฤษฎีการตอบสนองข้อสอบ ทั้งการประมาณค่าจากรูปแบบที่ 1 รูปแบบที่ 2 และรูปแบบที่ 3 ซึ่งใช้การประมาณค่าด้วยวิธี MCMC กับการประมาณค่าจากรูปแบบที่ 4 ซึ่งใช้การประมาณค่าด้วยวิธี Numerical method ให้ผลที่ไม่แตกต่างกัน แต่จากการประมาณค่าจากวิธีการที่ 4 จะให้ค่าประมาณที่คงที่มากกว่า และสามารถทำการประมาณค่าได้เฉพาะกรณีที่มีลักษณะการแจกแจงเริ่มแรกของข้อสอบและผู้สอบแบบปกติเท่านั้น

2. การประมาณค่าพารามิเตอร์ตามแนวการวิเคราะห์ด้วยทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัดด้วยวิธีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัดด้วยโมเดลการตอบสนองข้อสอบ จากการวิเคราะห์ค่าพารามิเตอร์ตามแนวทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัดของรูปแบบการประมาณค่าของวิธีการสรุปอ้างอิงความน่าเชื่อถือของโมเดลการตอบสนองข้อสอบ พบว่า การประมาณค่าองค์ประกอบความแปรปรวนในรูปแบบที่ 1 และรูปแบบที่ 4 ซึ่งใช้การประมาณค่าด้วยสูตร Kolen and Harris กับรูปแบบที่ 2 ซึ่งใช้การประมาณค่าด้วยสูตร ANOVA และรูปแบบที่ 3 ซึ่งใช้การประมาณค่าด้วยสูตร Kolen and Harris และ ANOVA ให้ผลการประมาณค่าองค์ประกอบความแปรปรวนไม่แตกต่างกัน ซึ่งสอดคล้องกับงานวิจัยของ Briggs and Wilson (2007) ที่เปรียบเทียบค่าองค์ประกอบความแปรปรวนของค่าสารสนเทศที่ได้จากวิธีการวิเคราะห์ด้วยวิธี GIRM กับการวิเคราะห์ด้วยทฤษฎีการสรุปอ้างความน่าเชื่อถือของผลการวัดแบบดั้งเดิม พบว่า ค่าองค์ประกอบความแปรปรวนที่ได้จากสองวิธีการวิเคราะห์ให้ผลที่ไม่แตกต่างกัน

เมื่อพิจารณาค่าส่วนเบี่ยงเบนมาตรฐานของค่าองค์ประกอบความแปรปรวนในแต่ละชุดข้อมูล พบว่า การประมาณค่าองค์ประกอบความแปรปรวนจากรูปแบบที่ 4 ให้ค่าส่วนเบี่ยงเบนมาตรฐานต่ำที่สุด แสดงว่ามีความคงที่ในการประมาณค่าองค์ประกอบความแปรปรวนมากที่สุด ทั้งนี้สาเหตุน่าจะเนื่องมาจากขั้นตอนการวิเคราะห์องค์ประกอบความแปรปรวนเป็นขั้นตอนที่ 3 ของวิธีการสรุปอ้างอิงความน่าเชื่อถือของโมเดลการตอบสนองข้อสอบ ดังนั้นข้อมูลก่อนนำมาวิเคราะห์เป็นข้อมูลที่ได้จากผลการวิเคราะห์ในขั้นตอนที่ 1 ซึ่งในรูปแบบที่ 4 นี้ใช้การประมาณค่าแบบ Numerical Bayesian ซึ่งไม่ฝ่าฝืนข้อตกลงเบื้องต้นของความไม่เป็นอิสระต่อกัน เหมือนการประมาณค่าด้วยวิธีการ MCMC ดังที่ได้ให้เหตุผลไว้ในอภิปรายผลการวิจัยในข้อที่ 1

นอกจากนี้เมื่อพิจารณาผลการวิเคราะห์องค์ประกอบความแปรปรวนของรูปแบบที่ 2 พบว่า ในการประมาณค่าองค์ประกอบความแปรปรวนไม่สามารถวิเคราะห์แยกองค์ประกอบความแปรปรวนของปฏิสัมพันธ์ (interaction) กับความคลาดเคลื่อนส่วนที่เหลือ (residual) ทั้งนี้เนื่องจากการวิเคราะห์ในรูปแบบที่ 2 ใช้การวิเคราะห์องค์ประกอบความแปรปรวนด้วย การวิเคราะห์ความแปรปรวนซึ่งการวิเคราะห์ด้วย ANOVA ไม่สามารถวิเคราะห์แยกแหล่งความแปรปรวนที่มาจากปฏิสัมพันธ์ กับความคลาดเคลื่อนส่วนที่เหลือ (residual)

3. การเปรียบเทียบประสิทธิภาพในการประมาณค่าของรูปแบบการประมาณค่าทั้ง 4 รูปแบบของวิธีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัดด้วยโมเดลการตอบสนองข้อสอบการเปรียบเทียบประสิทธิภาพของการประมาณค่าในแต่ละรูปแบบการประมาณค่าพารามิเตอร์ของวิธีการสรุปอ้างอิงความน่าเชื่อถือของโมเดลการตอบสนองข้อสอบทั้ง 4 รูปแบบ ซึ่งได้แก่ รูปแบบที่ 1 ซึ่งพัฒนาโดย Brigg and Wilson (2007) และรูปแบบที่ 2 รูปแบบที่ 3 รูปแบบที่ 4 ซึ่งพัฒนาวิธีการประมาณค่าโดยผู้วิจัย โดยในการวัดประสิทธิภาพของตัวประมาณค่าจะวัดจากดัชนี 3 ประเภท ซึ่งได้แก่ ความลำเอียงในการประมาณค่า ซึ่งคำนวณได้จากการวิเคราะห์ค่าความคลาดเคลื่อนเฉลี่ย (MAD) ความไม่-แน่นอนในการประมาณค่า ซึ่งคำนวณได้จากการวิเคราะห์ค่าส่วนเบี่ยงเบนมาตรฐาน และประสิทธิภาพในการประมาณค่าองค์ประกอบความแปรปรวน ซึ่งคำนวณได้จากการวิเคราะห์ค่าองค์ประกอบความแปรปรวนระยะทางยุคลิด (EUC)

ผลการศึกษา พบว่า เมื่อพิจารณาความลำเอียงในการประมาณค่าโดยพิจารณาจากค่า MAD รูปแบบที่ 1 กับรูปแบบที่ 4 มีประสิทธิภาพสูงที่สุด แต่รูปแบบที่ 4 เป็นรูปแบบที่ประมาณค่าพารามิเตอร์ได้เฉพาะลักษณะการแจกแจงเริ่มแรกแบบปกติเท่านั้น โดยเมื่อพิจารณาในแต่ละกรณีพบว่า รูปแบบที่ 1 จะประมาณค่าได้มีประสิทธิภาพในกรณีที่ลักษณะการแจกแจงเริ่มแรกของผู้สอบแบบปกติ และลักษณะการแจกแจงเริ่มแรกของข้อสอบแบบเบต้า ส่วนรูปแบบที่ 2 และรูปแบบที่ 3 จะประมาณค่าได้มีประสิทธิภาพในกรณีที่ลักษณะการแจกแจงเริ่มแรกของผู้สอบแบบแกมมา และลักษณะการแจกแจงเริ่มแรกของข้อสอบแบบเบต้า ส่วนรูปแบบที่ 4 จะประมาณค่าได้มีประสิทธิภาพในกรณีที่ลักษณะการแจกแจงเริ่มแรกของผู้สอบและข้อสอบแบบปกติ ดังนั้นจะเห็นได้ว่า ถ้าลักษณะการแจกแจงเริ่มแรกที่ไม่ปกติ เช่น ลักษณะการแจกแจงเริ่มแรกแบบแกมมา กับลักษณะการแจกแจงเริ่มแรกแบบเบต้า รูปแบบที่ 2 และรูปแบบที่ 3 จะให้การประมาณค่าที่มีประสิทธิภาพดีกว่า ทั้งนี้สาเหตุมาจากในขั้นตอนที่ 2 ของการประมาณในรูปแบบที่ 2 และรูปแบบที่ 3 มีการคำนวณหา Predictive Observed Matrix มีลักษณะคล้ายคลึงกับข้อมูลจริงแบบสองค่า (dichotomous) ซึ่งถ้าผู้ตอบถูกได้ 1 คะแนน ตอบผิดได้ 0 คะแนน ในขณะที่รูปแบบที่ 1 แบะรูปแบบที่ 4 มีการคำนวณหา Expected Response Matrix ซึ่งข้อมูลมีลักษณะเป็นความน่าจะเป็น ซึ่งได้มาจากการคำนวณจากสูตร Logistic function ดังนั้นค่าที่ได้จึงไม่เป็นอิสระจากกัน จึงทำให้การคำนวณในขั้นตอนที่ 3 จึงไม่เป็นอิสระต่อกัน นอกจากนี้ในการคำนวณ Predictive Observed Matrix ยังเป็นการปรับลักษณะการกระจายจึงทำให้สามารถประมาณค่า ลดความลำเอียงให้มากที่สุดเมื่อลักษณะการแจกแจงไม่ใช่ลักษณะแบบปกติ (Unnormal distribution)

เมื่อพิจารณาประสิทธิภาพในการประมาณค่าด้านความไม่แน่นอนในการประมาณค่า (Uncertainty) พบว่า รูปแบบที่ 4 มีประสิทธิภาพสูงที่สุด สำหรับลักษณะการแจกแจงเริ่มแรกแบบปกติ ถ้าลักษณะการแจกแจกเริ่มแรกของค่าพารามิเตอร์ ตัวหนึ่งตัวใดไม่ได้มีลักษณะการแจกแจงเริ่มแรกแบบปกติ รูปแบบที่ 1 มีประสิทธิภาพสูงที่สุด ทั้งนี้น่าจะมีสาเหตุมาจากการประมาณค่าพารามิเตอร์ในรูปแบบที่ 1 รูปแบบที่ 2 และรูปแบบที่ 3 ประมาณค่าพารามิเตอร์ของข้อสอบและค่าพารามิเตอร์ของผู้สอบด้วยวิธีการ MCMC ในขณะที่รูปแบบที่ 4 ประมาณค่าพารามิเตอร์ของข้อสอบและค่าพารามิเตอร์ของผู้สอบด้วยวิธีการ Numerical bayesian method ทั้งนี้สาเหตุอาจจะเป็นเพราะในการประมาณค่าด้วยวิธีการประมาณค่าด้วยวิธีการ MCMC ของการประมาณค่าแบบเบส์นั้นการประมาณค่าพารามิเตอร์ของผู้สอบจะขึ้นอยู่กับการประมาณค่าของข้อสอบ และการประมาณค่าของข้อสอบจะขึ้นอยู่กับการประมาณค่าผู้สอบ จึงทำให้ค่าพารามิเตอร์ทั้งสองไม่เป็นอิสระต่อกัน ซึ่งฝ่าฝืนข้อตกลงเบื้องต้นในการประมาณค่าของค่าพารามิเตอร์ของ

ทฤษฎีการตอบสนองข้อสอบ จึงทำให้การประมาณค่าด้วยรูปแบบที่ 1 รูปแบบที่ 2 และรูปแบบที่ 3 มีความไม่แน่นอนในการประมาณค่า (Uncertainty) มากกว่าการประมาณค่าด้วยรูปแบบที่ 4 ซึ่งสอดคล้องกับข้อเสนอแนะของบทความ GIRM ของ Briggs and Wilson (2007) ว่าการประเมินค่าด้วยวิธีการ MCMC จะทำให้ฝ่าฝืนข้อมูลเบื้องต้นของความเป็นอิสระของพารามิเตอร์ความสามารถของผู้สอบ และค่าพารามิเตอร์ความยากของข้อสอบ

ข้อเสนอแนะในการนำผลการวิจัยนี้ไปใช้

1. การประมาณค่ารูปแบบที่ 1 เหมาะสมกับสถานการณ์ที่มีลักษณะการแจกแจงเริ่มแรกของค่าพารามิเตอร์ตัวใดตัวหนึ่งไม่ใช่การแจกแจงแบบปกติ และลักษณะการแจกแจงเริ่มแรกของค่าพารามิเตอร์อีกตัวเป็นแบบปกติ เช่น ลักษณะการแจกแจงเริ่มแรกของข้อสอบมีลักษณะแบบปกติ และลักษณะการแจกแจงของผู้สอบแบบเบต้า เป็นต้น เนื่องจากจะให้ค่าพารามิเตอร์ที่ได้มีความลำเอียงต่ำและมีความคงที่ในการประมาณค่าของค่าพารามิเตอร์

นอกจากนี้สิ่งที่ต้องคำนึงถึงเมื่อใช้รูปแบบการประมาณค่ารูปแบบนี้คือ จำนวนหน่วยตัวอย่าง และความยาวของแบบสอบ โดยถ้าใช้จำนวนหน่วยและความยาวของข้อสอบจำนวนมากจะยิ่งทำให้ประสิทธิภาพของการประมาณค่ายิ่งเพิ่มสูงขึ้น และเมื่อพิจารณาถึงความไวที่มีต่อการกำหนดลักษณะการแจกแจงเริ่มแรกของค่าพารามิเตอร์ พบว่ารูปแบบนี้มีความทนทาน (Robustness) ต่อการกำหนดลักษณะการแจกแจงเริ่มแรกของผู้สอบ แต่มีความไวต่อการกำหนดลักษณะการแจกแจงเริ่มแรกของข้อสอบ

2. การประมาณค่ารูปแบบที่ 2 รูปแบบที่ 3 เหมาะสมกับสถานการณ์ที่มีลักษณะการ แจกแจงเริ่มแรกของค่าพารามิเตอร์ไม่ใช่การแจกแจงแบบปกติ เช่นลักษณะการแจกแจงเริ่มแรกของข้อสอบมีลักษณะแบบแกมมา และลักษณะการแจกแจงของผู้สอบแบบเบต้า เป็นต้น เนื่องจากจะให้ค่าพารามิเตอร์ที่ได้มีความคงที่ในการประมาณค่าของค่าพารามิเตอร์และมีประสิทธิภาพในการประมาณค่าองค์ประกอบความแปรปรวนของยุคลิดสูง

นอกจากนี้สิ่งที่ต้องคำนึงถึงเมื่อใช้รูปแบบการประมาณค่ารูปแบบนี้คือ ขนาดกลุ่มตัวอย่าง และความยาวของแบบสอบ โดยถ้าใช้ขนาดกลุ่มตัวอย่างและจำนวนข้อสอบจำนวนมากจะยิ่งทำให้ประสิทธิภาพของการประมาณค่ายิ่งเพิ่มสูงขึ้น และเมื่อพิจารณาถึงความไวที่มีต่อการกำหนดลักษณะการแจกแจงเริ่มแรกของค่าพารามิเตอร์พบว่ารูปแบบนี้มีความไวทั้งต่อการกำหนดลักษณะการแจกแจงเริ่มแรกของข้อสอบและผู้สอบ

3. การประมาณค่ารูปแบบที่ 4 สามารถประมาณค่าพารามิเตอร์เฉพาะสถานการณ์ที่มีลักษณะการแจกแจงเริ่มแรกของค่าพารามิเตอร์แบบปกติเท่านั้น เช่นลักษณะการแจกแจงเริ่มแรกของข้อสอบมีลักษณะแบบปกติ และลักษณะการแจกแจงของผู้สอบแบบปกติ แต่จะให้ค่าพารามิเตอร์ที่ประมาณได้มีประสิทธิภาพสูงมาก เนื่องจากจะให้ค่าพารามิเตอร์ที่ได้มีความลำเอียงต่ำ มีความคงที่ในการประมาณค่าของค่าพารามิเตอร์และมีประสิทธิภาพในการประมาณค่าองค์ประกอบความแปรปรวนของยุคลิดสูง

รูปแบบที่ 4 เป็นรูปแบบที่ไม่ผันแปรต่อขนาดตัวอย่างและความยาวของแบบสอบ ดังนั้น การวิเคราะห์ด้วยรูปแบบนี้จึงสามารถวิเคราะห์จำนวนของข้อสอบหรือผู้สอบน้อยได้มีประสิทธิภาพ นอกจากนี้ รูปแบบนี้มีความทนทานต่อการกำหนดลักษณะการแจกแจงเริ่มแรกของผู้สอบ แต่มีความไว ต่อการกำหนดลักษณะการแจกแจงเริ่มแรกของข้อสอบ

หมายเหตุ

[1] โปรแกรม GENOVA (GENeralized Analysis of Variance System) เป็นโปรแกรมสำหรับการวิเคราะห์ค่าสารสนเทศของทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัด (Generalizability Theory; GT) (Crick & Brennan, 1983)

[2, 3] โปรแกรม Win BUGS เป็นโปรแกรมสำหรับประมาณค่าพารามิเตอร์ด้วยวิธีการ MCMC (Markov Chain Monte Carlo) ในการประมาณค่าพารามิเตอร์ของวิธีการ สรุปอ้างอิงความน่าเชื่อถือของโมเดลการตอบสนองข้อสอบ ใช้โปรแกรม WinBUGS ในการประมาณโดยการเขียนคำสั่งในโปรแกรม R โดยให้วิเคราะห์ในโปรแกรม WinBUGS และกลับมารายงานผลในโปรแกรม R ด้วย Package R2WinBUGS (Sturtz, Ligges, & Gelman, 205)

References

Bock Brennan & Muraki. (2002). *The Information in Multiple Rating.* Applied Psychological Measurement. 26 (4), p. 364-375.

Brennan, R. (2001). An Essay on the History and Future of Reliability from the Perspective of Replications. *Journal of Educational Measurement.* 38 (4), p. 295-317.

Briggs, C. D. & Wilson, M. (2007). Generalizability in Item Response Modeling. *Journal of Educational Measurement,* 44 (2), p. 131-155.

Glas, C. A. W. , & Van der Linnden, W. J. (2003). Computerized Adaptive Testing with Item Cloning. *Applied Psychological Measurement,* 27 (4), p. 247-261.

Kolen, M. & Harris, D. (1987). *A Multivariate Test Theory Model Based on Item Response Theory and Generalizability.* Paper presented at the American Educational Research Association, Washington DC., USA.

Mislevy, R. J. (1993). Some formuls for use with Bayesian ability estiamates. *Educational and Psychology Measurement,* 52, p. 315-328.

Ntzoufras, I. (2009). Bayesian A John Wiley & Sons, inc., Publication.

Patz, R. J. , Junker, B. W., Johnson, M. S. & Mariano, L. T. (2002). The Hierarchical Rater Model for Rated Test Items and its Application to Large-Scale Educational Assessment Data. *Journal of Educational and Behavioral Statistics.* 27 (4), p. 341-384.

Spiegelhalter, D., Thomas, A., Best N., & Lunn D. (2004). WinBUGS 1.4 (available at http://www.mrc-bus.com.ac.uk/buge).

Sturtz, S., Ligges, U. & Gelman, A. (2005). R2WinBUGS : A package for running WinBUGS from R. *Journal of statistical Sofwave, 12(3). Available online at http://www.jstatsoft.org/v12/i03/v12i03.pdf.*

Swaminathan, H. & Gifford, J. A. (1982). Bayesina Estimation in the Rasch Model. *Journal of Edicational Statistics.* 7, p. 175-192.

Swaminathan, H. & Gifford, J. A. (1985). Bayesina Estimation in the Two-Parameter Logistic Model. *Psychometrika.* 50 (September) ,p. 349-364.

Swaminathan, H. & Gifford, J. A. (1986). Bayesina Estimation in the Three-Parameter Logistic Model. *Psychometrika.* 51 (December) p. 599-601.

Verhelst, N. & Verstralen, H. (2001). An IRT model for multiple raters. In A. Boomsma, M. van Duijn, and T. Snijders (Eds.) *Essays on item response theory* (pp. 88-108). New York : Springer-Verlag.

คณะผู้เขียน

ดร.ชนะศึก นิชานนท์

ดุษฎีบัณฑิตสาขาการวัดและประเมินผลการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ผู้อำนวยการสถาบันวิจัยและพัฒนา มหาวิทยาลัยราชภัฏสวนดุสิต

Email : chanasuek.nic@.dusit.ac.th

ศาสตราจารย์ ดร.ศิริชัย กาญจนวาสี

คณบดีคณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

Email : skanjanawasee@hotmail.com

Professor Mark Wilson, Ph.D

University of California, Berkeley USA

Email : MarkW @berkeley.edu