การเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบให้คะแนนหลายค่า โดยวิธีทดสอบอัตราส่วนความควรจะเป็น วิธีเบส์เซียน และวิธีโพลี-ซิปเทสท์

A Comparison of the Efficiency of Likelihood Ratio Test, Batesian and Poly-SIBTEST Procedures in Detecting Differential Item Functioning for Polytomous Scored Items

อาวีพร ปานทอง [1]/ ไพรัตน์ วงษ์นาม [2]/ เกตุจันทร์ จำปาไชยศรี [3]

Aweeporn Panthong / Pairatana Wongnam / Katechan Jampachaisri

[1, 2] สาขาวิชาวิจัย วัดผล และสถิติการศึกษา ภาควิชาวิจัยและจิตวิทยาประยุกต์ คณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา
Program in Educational Research, Measurement, and Statistics, Department of Research and Applied Psychology, Faculty of Education, Burapha University.

[3] สาขาวิชาสถิติ ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยนเรศวร
Program in Statistics, Department of Mathematics, Faculty of Science, Naresuan University.

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบอัตราความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนหลายค่า โดยวิธีทดสอบอัตราส่วนความควรจะเป็น วิธีเบส์เซียน และวิธีโพลี-ซิปเทสท์ ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย ข้อมูลที่ศึกษาเป็นข้อมูลจำลองโดยใช้โมเดลพาเชียลเครดิตทั่วไปภายใต้ทฤษฎีการตอบสนองข้อสอบ จำลองแบบทดสอบที่มีโครงสร้างวัดความสามารถมิติเดียว โดยข้อสอบแต่ละข้อวัดความสามารถหลักข้อสอบทุกข้อมีตัวเลือกให้เลือกจำนวน 5 ตัวเลือก ในการจำลองข้อมูลผลการตอบภายใต้ปัจจัยที่แตกต่างกัน คือ ความยาวของแบบสอบ 3 รูปแบบ ขนาดของการทำหน้าที่ต่างกันของข้อสอบ 3 ขนาด ความแตกต่างของการแจกแจงความสามารถ 2 ระดับ และขนาดตัวอย่าง 3 รูปแบบ รวมข้อมูลทั้งหมดที่ต้องจัดกระทำจำนวน 54 เงื่อนไข ($3 \times 3 \times 3 \times 2$) ในแต่ละเงื่อนไขจำลองข้อมูลทำซ้ำ 500 รอบ ผลการวิจัยพบว่า เมื่อความยาวของข้อสอบและขนาดตัวอย่างเพิ่มขึ้น วิธีทดสอบอัตราส่วนความควรจะเป็น และวิธีเบส์เซียน สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดีกว่าวิธีโพลี-ซิปเทสท์ โดยภาพรวมวิธีทดสอบอัตราส่วนความควรจะเป็น และวิธีเบส์เซียนมีอัตราความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบใกล้เคียงกัน และอยู่ในเกณฑ์ที่กำหนดมากกว่าวิธีโพลี-ซิปเทสท์ ผลการศึกษาครั้งนี้เสนอแนะให้ใช้วิธีทดสอบอัตราส่วนความควรจะเป็น และวิธีเบส์เซียนเนื่องจากสามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ และมีอำนาจการทดสอบสูง

คำสำคัญ: การทำหน้าที่ต่างกันของข้อสอบ, วิธีทดสอบอัตราส่วนความควรจะเป็น, วิธีเบส์เซียน, วิธีโพลี-ซิปเทสท์

Abstract

The purpose of this research was to compare Type I error rate and the power of likelihood ratio test (LRT), Bayesian, and the Poly-SIBTEST procedures in the detecting of differential item functioning (DIF) for polytomous scored items. In this study, data were simulated under the generalized partial credit model, and responses were simulated from one dimensional test. All items were in five response categories scoring. These data were simulated under a variety of four factors: three levels forms of length test, three levels forms of DIF magnitudes, two levels of ability distribution differences, and three levels of sample size proportions. A total of 54 (3x3x3x2) conditions were studied. The data were replicated 500 times for each condition. Results of the study were as follows: When length test increased, LRT and Bayesian procedure had better control of type I error rate than Poly-SIBTEST procedure. In general, the Type I error rates of LRT and Bayesian procedures were within the nominal limits. They were higher power than Poly-SIBTEST procedures. The results of this study suggested LRT and Bayesian procedures to control the Type I error rate and high power.

Keywords: the detecting of differential item functioning, likelihood ratio test, Bayesian, the Poly-SIBTEST

บทนำ

การวัดผลและประเมินผลการศึกษา เป็นเครื่องมือสำคัญที่ช่วยพัฒนาคุณภาพการศึกษา ผลที่ได้จากการวัดผลและประเมินผลการศึกษานับเป็นเครื่องมือสำคัญที่ช่วยในการปรับปรุงคุณภาพของการศึกษาเพราะผลจากการวัดและประเมินผลการศึกษาจะเป็นพื้นฐานสำหรับการตัดสินผลการเรียน เพื่อใช้ในการปรับปรุงการเรียนการสอนและช่วยให้ได้ข้อมูลสารสนเทศเกี่ยวกับพัฒนาการ การแนะแนว การประเมินผลหลักสูตร แบบเรียน การจัดระบบบริหารของโรงเรียนตลอดจนการปรับปรุงวิธีการเรียนของผู้เรียนให้มีประสิทธิภาพยิ่งขึ้น สำหรับการวัดและประเมินผลเพื่อใช้สำหรับการตัดสินผลการเรียน เป็นการประเมินสรุปผลการเรียนรู้เพื่อตัดสินและให้การรับรองความรู้ความสามารถของผู้เรียนว่าอยู่ในระดับใด เครื่องมือที่ใช้ในการประเมินจึงมีความหลากหลาย ยกตัวอย่างเช่น แบบทดสอบ แบบสังเกต แบบสอบถาม และแบบวัดทางจิตวิทยา เป็นต้น

แบบทดสอบเลือกตอบ (Multiple-choice Test) เป็นเครื่องมือวัดผลที่นิยมใช้มากที่สุด เนื่องจากสามารถวัดได้ครอบคลุมเนื้อหา มีความเป็นปรนัยในการตรวจให้คะแนน ตรวจให้คะแนนง่าย รวดเร็ว และสะดวกในการดำเนินการสอบ เนื่องจากแบบทดสอบที่ให้คะแนนตอบถูกให้ 1 คะแนน ตอบผิดให้ 0 คะแนน มีจุดบกพร่องคือสามารถเดาคำตอบถูกได้ เนื่องจากมีการกำหนดคำตอบให้เลือก ดังนั้นเพื่อลดโอกาสการเดาถูก นักวัดผลจึงพัฒนาแบบทดสอบที่สามารถวัดความรู้ที่แท้จริงของผู้เรียนมากที่สุด กระบวนการพัฒนาแบบทดสอบเลือกตอบจึงเน้นที่วิธีการตอบและการให้คะแนน โดยมีเป้าหมายเพื่อแก้ไขจุดบกพร่องของแบบทดสอบแบบเลือกตอบ วิธีการปรับการตรวจให้คะแนนรายข้อมากกว่า 2 ค่า ข้อสอบที่ให้คะแนนหลายค่า (Polytomously scored items) ปัจจุบันนักการศึกษาให้ความสนใจการสร้างเครื่องมือวัดแบบให้คะแนนหลายค่ามากยิ่งขึ้น เนื่องจากในวงการศึกษาและหน่วยงานต่างๆ ต้องการนำไปใช้พิจารณาตัดสินใจในเรื่องต่างๆ มากยิ่งขึ้น เช่น ข้อสอบที่ให้คะแนนหลายค่ามีหลายประเภท เช่น ข้อสอบความเรียง (Essay items) การตัดสินคุณภาพของแฟ้มสะสมงาน (Portfolio) การพิสูจน์ทางคณิตศาสตร์ และการทดลองทางวิทยาศาสตร์ เป็นต้น การหาคุณภาพของเครื่องมือหรือแบบทดสอบเป็นส่วนสำคัญของการวัดผลและประเมินผลทางการศึกษา ค่าความตรงเป็นคุณสมบัติที่สำคัญที่สุดของเครื่องมือทุกชนิดที่จะบ่งชี้ถึงคุณภาพของเครื่องมือวัด ซึ่งความตรงเป็นคุณสมบัติที่แสดงถึงความสามารถในการวัดหรือแบบสอบนั้นทำหน้าที่ได้ตามวัตถุประสงค์ที่กำหนดไว้หรือไม่ และวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นอีกหนึ่งวิธีที่ใช้เพื่อตรวจสอบความตรงของข้อสอบ เมื่อนำแบบทดสอบไปสอบกับ

ผู้สอบกลุ่มย่อยที่มีลักษณะแตกต่างกันตั้งแต่ 2 กลุ่มขึ้นไป ที่มีความสามารถหรือคุณลักษณะหลักเท่ากัน ความน่าจะเป็นในการตอบข้อสอบถูกข้อนั้นไม่เท่ากัน แสดงว่าแบบทดสอบนั้นขาดความตรง หรือมีการทำหน้าที่ต่างกันของข้อสอบ เพราะไม่ได้วัดเฉพาะคุณลักษณะหลักที่เป็นเป้าหมายตามที่ต้องการวัดเท่านั้น แต่ยังวัดคุณลักษณะแฝงแทรกซ้อนที่ไม่ต้องการวัดของผู้สอบอีกด้วย วิธีการนี้เรียกว่า “การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ” ผู้วิจัยมีความสนใจที่จะเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบให้คะแนนหลายค่าโดยวิธีอัตราส่วนความควรจะเป็น (Likelihood ratio test) วิธีเบส์เซียน (Bayesian) และวิธีโพลี-ซิปเทสท์ (Poly-SIBTEST) ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย เพื่อศึกษาประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนหลายค่า

วัตถุประสงค์ของการวิจัย

1. เปรียบเทียบอัตราความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนหลายค่าระหว่างวิธีทดสอบอัตราส่วนความควรจะเป็น วิธีเบส์เซียน และวิธีโพลี-ซิปเทสท์ ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัยคือ ความยาวของแบบสอบ ขนาดของการทำหน้าที่ต่างกันของข้อสอบ ความแตกต่างของการแจกแจงความสามารถ ขนาดของกลุ่มตัวอย่าง และวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

ขอบเขตการวิจัย

1. ข้อมูลที่ศึกษาเป็นข้อมูลจำลอง โดยใช้โมเดลพาเชียลเครดิตทั่วไป (Generalized partial credit model) ภายใต้ทฤษฎีการตอบสนองข้อสอบ จำลองแบบทดสอบที่มีโครงสร้างวัดความสามารถมิติเดียว (Unidimensional) โดยข้อสอบแต่ละข้อวัดความสามารถเป้าหมายหลักกำหนดให้ แทนความสามารถหลัก ข้อสอบทุกข้อมีตัวเลือกให้เลือกตอบจำนวน 5 ตัวเลือก โดยให้คะแนนเป็น 0, 1, 2, 3 และ 4 คะแนน ในการจำลองข้อมูลดังกล่าวจะจำลองผลการตอบข้อสอบ ภายใต้ปัจจัยที่แตกต่างจำนวน 5 ปัจจัย คือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ 3 วิธี ความยาวของแบบสอบ 3 รูปแบบ ขนาดของการทำหน้าที่ต่างกันของข้อสอบ 3 ขนาด ความแตกต่างของการแจกแจงความสามารถ 2 ระดับ และขนาดของกลุ่มตัวอย่าง 3 รูปแบบ รวมข้อมูลทั้งหมดที่ต้องจัดกระทำเพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจำนวน 54 เงื่อนไข (3 × 3 × 3 × 2) ในแต่ละเงื่อนไขจำลองข้อมูลทำซ้ำ 500 รอบ จำนวนการทำซ้ำภายใต้เงื่อนไขที่แปรเปลี่ยนทั้งหมด 27,000 รอบ

วิธีดำเนินการวิจัย

การจัดกระทำตัวแปร

การวิจัยครั้งนี้ศึกษาในสถานการณ์จำลอง โดยใช้ทฤษฎีการตอบข้อสอบแบบมิติเดียว (Unidimensional item response theory) โมเดลพาเชียลเครดิตทั่วไป (Generalized partial credit Model) จำลองข้อมูลภายใต้การจัดกระทำตัวแปรอิสระ 4 ตัวแปร คือ ความยาวของข้อสอบ 3 ขนาด ขนาดของข้อสอบที่ทำหน้าที่ต่างกัน 3 ขนาด ความแตกต่างของการแจกแจงความสามารถ 2 ระดับ และขนาดกลุ่มตัวอย่าง 3 ขนาด โดยจำลองแบบทดสอบที่มีโครงสร้างแบบมิติเดียว (Unidimensional) ข้อสอบแต่ละข้อมีรายการตอบ 5 รายการ ผลการตอบในรายการตอบ 1, 2, 3, 4 และ 5 ให้คะแนนเป็น 0, 1, 2, 3 และ 4 ตามลำดับ สำหรับการจัดกระทำตัวแปรอิสระทั้ง 4 ตัวแปรดังนี้

1. ตัวแปรอิสระ มี 4 ตัวแปร ดังนี้

1.1 ความยาวของแบบทดสอบ 3 รูปแบบ คือ จำนวน 30 ข้อ, 50 ข้อ และ 100 ข้อ

1.2 ขนาดของข้อสอบทำหน้าที่ต่างกันมี 3 ขนาด คือ 0.25, 0.50 และ 1.00

1.3 ความแตกต่างของการแจกแจงความสามารถ มี 2 ระดับ คือ เท่ากัน และไม่เท่ากัน $\left(\bar{\theta}_R - \bar{\theta}_F = 1.0SD\right)$

1.4 ขนาดตัวอย่าง สัดส่วนจำนวนผู้สอบ 1 : 2 คือ 50 : 100 คน, 100 : 200 คน และ 200 : 400 คน

2. ตัวแปรตาม มี 2 ตัวแปร คือ ประสิทธิภาพของการตรวจสอบ

2.1 อัตราความคลาดเคลื่อนประเภทที่ 1

2.2 อำนาจการทดสอบ

การวิเคราะห์ข้อมูล

การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ ผู้วิจัยนำรายการคำตอบของผู้สอบที่เป็นการตรวจให้คะแนนหลายค่าที่ได้จากการจำลองข้อมูลที่มีการระบุขนาดของการทำหน้าที่ต่างกันในกลุ่มอ้างอิงและกลุ่มเปรียบเทียบมาวิเคราะห์ค่าทางสถิติเพื่อระบุถึงการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ภายใต้โมเดลพาเชียลเครดิตทั่วไป (Generalized Partial Credit Model) เป็นการศึกษาความน่าจะเป็นของการเลือกตอบข้อคำถามของผู้สอบ (j) โดยมีค่าคะแนนในแต่ละตัวเลือก (k) ในแต่ละข้อคำถาม (i) มีสมการดังนี้ (Hyun & Taehoon. 2006. p.5)

$$P_{jk}(\theta) = \frac{e^{\sum_{k=0}^{m} a_i\left(\theta_j - b_i + \tau_k\right)}}{\sum_{y=0}^{m} e^{\sum_{k=0}^{m} a_i\left(\theta_j - b_i + \tau_k\right)}} \quad \text{เมื่อ } \tau_1 \equiv 0$$

เมื่อ i แทน ผู้ตอบข้อคำถาม หรือ ผู้สอบคนที่ 1,2,...,I

j แทน ข้อคำถาม มีค่า j = 1, 2, 3,...,J

k แทน ตัวเลือกในแต่ละข้อคำถาม มีค่า k = 0, 1, 2,..., m

a_j แทน ค่าอำนาจจำแนกของข้อคำถามข้อที่ j

b_j แทน ค่าความยากของข้อคำถามข้อที่ j

τ_k แทน ค่าจุดตัดในการเปลี่ยนตัวเลือกในแต่ละข้อคำถาม (threshold)

อัตราความคลาดเคลื่อนประเภทที่ 1

เกณฑ์การพิจารณาหากมีค่าความคลาดเคลื่อนประเภทที่ 1 ต่ำกว่าหรือเท่ากับ 0.05 ถือว่าควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดี (Atar, B. & Kamata, A. 2011. P. 40) นั่นคือวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบไม่ระบุการทำหน้าที่ต่างกันของข้อสอบในข้อที่ไม่มีการทำหน้าที่ต่างกันของข้อสอบได้จริง

อำนาจการทดสอบ

เกณฑ์ที่ใช้พิจารณาอำนาจการทดสอบ จะพิจารณาอำนาจการทดสอบเมื่อสามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ก่อน และอำนาจการทดสอบต้องมีค่าเฉลี่ยตั้งแต่ 0.80 ขึ้นไป จึงถือว่ามีอำนาจการทดสอบเพียงพอ (Sufficient power) หากต่ำกว่า 0.80 ถือว่าวิธีนั้นๆ ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ไม่ดี (Atar, B. & Kamata, A. 2011. P. 40)

ผลการวิจัย

การวิเคราะห์ประสิทธิภาพเมื่อพิจารณาอัตราความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนหลายค่าระหว่างวิธีทดสอบอัตราส่วนความควรจะเป็น วิธีเบส์เซียน และวิธีโพลี-ซิปเทสท์ ภายใต้ปัจจัยที่แตกต่างกัน 4 ปัจจัย สำหรับผลการวิเคราะห์ประสิทธิภาพเมื่อพิจารณาทั้งความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสามวิธี ภายใต้ปัจจัยที่แตกต่างกัน 4 ปัจจัย คือ ความยาวของแบบสอบ ขนาดของการทำหน้าที่ต่างกันของข้อสอบ ความแตกต่างของการแจกแจงความสามารถ และขนาดตัวอย่าง มีดังนี้

ตารางที่ 1 ผลการวิเคราะห์ประสิทธิภาพเมื่อพิจารณาทั้งความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของวิธีทดสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธีทดสอบอัตราส่วนความควรจะเป็น วิธีเบส์เซียน และวิธีโพลี-ซิปเทสท์

ความยาวของข้อสอบ	ขนาดของการทำหน้าที่ต่างกันของข้อสอบ	ความแตกต่างของการแจกแจงความสามารถ	ขนาดของกลุ่มตัวอย่าง ($N_F : N_R$)		
			50 : 100	100 : 200	200 : 400
30 ข้อ	0.25	เท่ากัน	-	-	-
		แตกต่างกัน 1.0SD	POLYS	BAYES/POLYS	LRT
	0.50	เท่ากัน	BAYES	BAYES/POLYS	BAYES
		แตกต่างกัน 1.0SD	LRT/BAYES	LRT/BAYES	BAYES
	1.00	เท่ากัน	BAYES	LRT/BAYES	LRT/BAYES
		แตกต่างกัน 1.0SD	LRT/BAYES	LRT/BAYES	LRT/BAYES
50 ข้อ	0.25	เท่ากัน	-	BAYES	BAYES
		แตกต่างกัน 1.0SD	POLYS	LRT	LRT/BAYES
	0.50	เท่ากัน	BAYES	LRT/BAYES	LRT/BAYES/POLYS
		แตกต่างกัน 1.0SD	LRT/BAYES	LRT/BAYES	LRT
	1.00	เท่ากัน	LRT/BAYES	BAYES	LRT/BAYES
		แตกต่างกัน 1.0SD	BAYES	LRT/BAYES	LRT/BAYES
100 ข้อ	0.25	เท่ากัน	POLYS	BAYES/POLYS	BAYES
		แตกต่างกัน 1.0SD	POLYS	LRT	LRT
	0.50	เท่ากัน	BAYES	BAYES	LRT/BAYES
		แตกต่างกัน 1.0SD	LRT	LRT	LRT
	1.00	เท่ากัน	POLYS	LRT/BAYES	LRT/BAYES
		แตกต่างกัน 1.0SD	LRT/BAYES	LRT/BAYES/POLYS	LRT/BAYES/POLYS

จากตารางที่ 1 ผลการวิเคราะห์ประสิทธิภาพเมื่อพิจารณาทั้งความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของวิธีทดสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธีทดสอบอัตราส่วนความควรจะเป็น วิธีเบส์เซียน และวิธีโพลี-ซิปเทสท์ ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัยหลัก พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธีเบส์เซียน มีอำนาจการทดสอบมากที่สุด รองลงมาคือ วิธีทดสอบอัตราส่วนความควรจะเป็น และวิธีโพลี-ซิปเทสท์ตามลำดับ ภายใต้ปัจจัยที่แปรเปลี่ยนทุกๆ ปัจจัย

อภิปรายผลการวิจัย

1. ปัจจัยความยาวของแบบสอบ

ผลการตรวจสอบพบว่า ปัจจัยความยาวของแบบสอบที่ทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนหลายค่า ในการตรวจสอบด้วยวิธีทดสอบอัตราส่วนความควรจะเป็น วิธีเบส์เซียน และวิธีโพลี-ซิปเทสท์ เมื่อพิจารณาปัจจัยความยาวข้อสอบ 30 ข้อ และ 50 ข้อ ภายใต้ปัจจัยอื่นที่แปรเปลี่ยน สรุปได้ว่า เมื่อความยาวของข้อสอบเพิ่มมากขึ้น การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนหลายค่า พบว่า การตรวจสอบด้วยวิธีทดสอบอัตราส่วนความควรจะเป็น วิธีเบส์เซียน และ

วิธีโพลี-ซิปเทสท์ ทั้งสามวิธีมีอัตราความคลาดเคลื่อนประเภทที่ 1 ลดลง โดยวิธีเบส์เซียนสามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดีมากที่สุด รองลงมาคือวิธีทดสอบอัตราส่วนความควรจะเป็น ผลดังกล่าว สอดคล้องกับผลการศึกษาของโคเฮ็นและคิม (Cohen & Kim, 1993) ที่พบว่าเมื่อเพิ่มความยาวของแบบทดสอบแล้วอัตราความคลาดเคลื่อนประเภทที่ 1 ในการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีโพลี-ซิบเทสท์จะเพิ่มขึ้นเมื่อพิจารณาค่าอำนาจการทดสอบในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนหลายค่าทั้งสามวิธี ผลการตรวจสอบพบว่าปัจจัยความยาวของแบบสอบที่ทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนหลายค่า ในการตรวจสอบด้วยวิธีเบส์เซียน มีอำนาจการทดสอบในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมากกว่า วิธีทดสอบอัตราส่วนความควรจะเป็น และวิธีโพลี-ซิปเทสท์ ในทุกๆปัจจัยอื่นที่แปรเปลี่ยน วิธีทดสอบอัตราส่วนความควรจะเป็น และวิธีเบส์เซียน มีอำนาจการทดสอบในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนหลายค่าเพิ่มขึ้น เมื่อความยาวแบบสอบเพิ่มขึ้น โดยวิธีเบส์เซียนมีอำนาจการทดสอบในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนหลายค่าได้ดีมากที่สุด สอดคล้องกับผลการศึกษาของ Swaminathan and Rogers (1990) พบว่าเมื่อความยาวของแบบสอบเพิ่มขึ้น จะทำให้อัตราความถูกต้องของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูงขึ้น แต่เมื่อความยาวข้อสอบ 60 ข้อ อัตราความถูกต้องของวิธีโพลี-ซิบเทสท์จะลดลง

2. ปัจจัยขนาดของการทำหน้าที่ต่างกันของข้อสอบ

ผลการตรวจสอบพบว่า ปัจจัยขนาดของการทำหน้าที่ต่างกันของข้อสอบมีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 1 เมื่อขนาดของการทำหน้าที่ต่างกันของข้อสอบเพิ่มมากขึ้น การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนหลายค่า ด้วยวิธีทดสอบอัตราส่วนความควรจะเป็น และวิธีเบส์เซียน อัตราความคลาดเคลื่อนประเภทที่ 1 ลดลงและสามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ดี ผลดังกล่าว สอดคล้องกับผลการศึกษาของของแอทตา และคามาตะ (Atar & Kamata, 2011, p. 40) พบว่าเมื่อขนาดการทำหน้าที่ต่างกันของข้อสอบมี ขนาดลดลง วิธีการตรวจสอบด้วยวิธีอัตราส่วนความควรจะเป็น เมื่อขนาดกลุ่มตัวอย่างมีขนาดเล็ก (N = 400R/200F) หรือขนาดกลาง Medium (N =800R/400F) และวิธีการถดถอยโลจิสติกเมื่อขนาดกลุ่มตัวอย่างมีขนาดใหญ่ (N =1200R/1200F) ทั้งสองวิธีจะมีอัตราส่วนความคลาดเคลื่อนประเภทที่ 1 ลดลง และค่าอำนาจการทดสอบในการตรวจสอบของทั้งสามวิธีเมื่อขนาดของการทำหน้าที่ต่างกันของข้อสอบเพิ่มมากขึ้น การตรวจสอบการทำหน้าที่ต่างกันด้วยวิธีทดสอบอัตราส่วนความควรจะเป็น และวิธีเบส์เซียนมีอำนาจการทดสอบเพิ่มขึ้น โดยวิธีเบส์เซียนมีอำนาจการทดสอบในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนหลายค่าได้ดีมากที่สุด รองลงมาคือ วิธีทดสอบอัตราส่วนความควรจะเป็น

3. ปัจจัยความแตกต่างของการแจกแจงความสามารถ

ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนหลายค่า พบว่า ปัจจัยความแตกต่างของการแจกแจงความสามารถของกลุ่มผู้สอบ ภายใต้ปัจจัยอื่นทุกๆ ปัจจัย เมื่อการแจกแจงความสามารถระหว่างกลุ่มเท่ากัน $\left(\bar{\theta}_R - \bar{\theta}_F = 1.0SD\right)$ การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนหลายค่าด้วยวิธีเบส์เซียนสามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ดี และเมื่อการแจกแจงความสามารถระหว่างกลุ่มไม่เท่ากัน การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนหลายค่าด้วยวิธีทดสอบอัตราส่วนความควรจะเป็นสามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ดี และค่าอำนาจการทดสอบในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อการแจกแจงความสามารถของกลุ่มผู้สอบระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบเท่ากัน การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนหลายค่าด้วยวิธีเบส์เซียนมีอำนาจการทดสอบสูง และการแจกแจงความสามารถระหว่างกลุ่มไม่เท่ากัน $\left(\bar{\theta}_R - \bar{\theta}_F = 1.0SD\right)$ วิธีทดสอบอัตราส่วนความควรและวิธีเบส์เซียน มีอำนาจการทดสอบสูงในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนหลายค่าสอดคล้องกับผลการศึกษาของ Li-An Wu and Rung-Ching Tsai (2010) พบว่าเมื่อขนาดของค่าเฉลี่ยความสามารถผู้สอบเพิ่มขึ้น อัตราความถูกต้อง

ของวิธีDFIT และวิธี LRT จะเพิ่มขึ้น

4. ปัจจัยขนาดของตัวอย่าง

ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนหลายค่า พบว่า ขนาดของตัวอย่างมีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 1 เมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้น วิธีทดสอบอัตราส่วนความควรจะเป็นและวิธีเบส์เซียน มีอัตราความคลาดเคลื่อนประเภทที่ 1 ลดลง และสามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ ตามเกณฑ์ที่กำหนด และวิธีโพลี-ซิปเทสท์ ควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้กรณีขนาดตัวอย่างมีขนาดเล็ก คือ 50 : 100 คน และค่าอำนาจการทดสอบในการตรวจสอบพบว่า เมื่อขนาดของตัวอย่างมีสัดส่วน 25 : 50 คน (NF : NR) การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนหลายค่า ด้วยวิธีเบส์เซียนมีอำนาจการทดสอบสูงกว่าวิธีอื่นๆ และเมื่อขนาดตัวอย่างเพิ่มขึ้น สัดส่วน 100 : 200 คน และ 200 : 400 คน (NF : NR) วิธีทดสอบอัตราส่วนความควรจะเป็น และวิธีเบส์เซียนจะมีอำนาจการทดสอบสูงขึ้น สอดคล้องกับผลการศึกษาของ Narayanan and Swaminathan (1996) ที่พบว่าเมื่อขนาดตัวอย่างเพิ่มขึ้นจะมีผลทำให้อัตราความถูกต้องเพิ่มขึ้นตามด้วย

ข้อเสนอแนะในการนำผลการวิจัยไปใช้

1. ผลการศึกษาครั้งนี้พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนหลายค่าภายใต้เงื่อนไขของปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ ความยาวของแบบสอบ 3 รูปแบบ ขนาดของการทำหน้าที่ต่างกันของข้อสอบ 3 ขนาด ความแตกต่างของการแจกแจงความสามารถ 2 ระดับ และขนาดของตัวอย่าง 3 รูปแบบ โดยภาพรวมภายใต้ทุกเงื่อนไขวิธีทดสอบอัตราส่วนความควรจะเป็น และวิธีเบส์เซียนมีความเหมาะสมมากที่สุดในการตรวจสอบการทำหน้าที่ต่างกัน ทำให้สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้และมีอำนาจการทดสอบสูง

2. ผลการศึกษาครั้งนี้พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนหลายค่า ด้วยวิธีโพลี-ซิปเทสท์ ไม่เหมาะสมกับขนาดของตัวอย่างขนาดใหญ่จำนวนข้อสอบที่มีจำนวนมาก และข้อสอบที่มีขนาดของการทำหน้าที่ต่างกันขนาดกลางและขนาดใหญ่ ซึ่งจะมีผลทำให้อัตราความคลาดเคลื่อนประเภทที่ 1 สูงเกินปกติ (Inflate) ดังนั้นควรใช้ในขนาดกลุ่มตัวอย่างที่มีขนาดเล็กและจำนวนข้อสอบที่มีจำนวนน้อยข้อซึ่งจะมีผลทำให้การตรวจสอบมีความถูกต้องและแม่นยำมากขึ้น

3. ผลการศึกษาครั้งนี้พบว่า ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนหลายค่าภายใต้ขนาดตัวอย่าง ซึ่งประกอบด้วยกลุ่มเปรียบเทียบและกลุ่มอ้างอิงที่มีสัดส่วนจำนวนผู้สอบในแต่ละกลุ่ม คือ สัดส่วน 1 : 2 ได้แก่ 25 : 50 คน, 50 : 100 คน และ 100 : 200 คน (กลุ่มเปรียบเทียบ : กลุ่มอ้างอิง) เมื่อวิเคราะห์ด้วยวิธีทดสอบอัตราส่วนความควรจะเป็นสามารถตรวจพบข้อสอบทำหน้าที่ต่างกันของข้อสอบที่มีขนาดกลุ่มตัวอย่างขนาด 200 : 400 คน และเหมาะสมกับขนาดของการทำหน้าที่ต่างกันของข้อสอบที่มีขนาดใหญ่ (1.00DIF) ในขณะที่วิธีเบส์เซียนสามารถตรวจพบการทำหน้าที่ต่างกันของข้อสอบได้ที่มีขนาดตัวอย่าง 50 : 100 คน และ 100 : 200 คน และเหมาะสมกับขนาดของการทำหน้าที่ต่างกันของข้อสอบที่มีขนาดเล็กและขนาดกลาง (0.25DIF, 0.5DIF ตามลำดับ) ดังนั้นวิธีเบส์เซียนมีความเหมาะสมในการนำไปใช้มากกว่า ส่วนวิธีโพลี-ซิปเทสท์เป็นวิธีในกลุ่มทฤษฎีการตอบข้อสอบมีรูปแบบนอนพาราเมทริก (Nonparametric form) ไม่จำเป็นต้องใช้โมเดลประมาณค่าพารามิเตอร์ ดังนั้นจึงไม่มีข้อตกลงของโมเดลที่ใช้อธิบายความสัมพันธ์ระหว่างผลการตอบข้อสอบกับตัวแปรการจับคู่ (Chang, Mazzeo; & Roussos, 1996 : 334) ดังนั้นวิธีโพลี-ซิปเทสท์จึงมีจุดเด่นที่ไม่จำเป็นต้องใช้ตัวอย่างขนาดใหญ่ จึงเหมาะกับขนาดตัวอย่างขนาดเล็กและความยาวข้อสอบจำนวนน้อยข้อ

4. ผลการศึกษาครั้งนี้พบว่า ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีทดสอบอัตราส่วนความควรจะเป็น วิธีเบส์เซียน และวิธีโพลี-ซิปเทสท์ มีข้อดีข้อจำกัดแตกต่างกัน ข้อดีของวิธีทดสอบอัตราส่วนความควรจะเป็น คือ โปรแกรมที่ใช้ในการวิเคราะห์ไม่ยุ่งยาก มีให้เลือกหลากหลายโปรแกรม และเวลาที่ใช้ในการวิเคราะห์ข้อมูลได้รวดเร็ว ข้อจำกัดของวิธีทดสอบอัตราส่วนความควรจะเป็น คือ การแจกแจงความสามารถของผู้สอบ และพารามิเตอร์ของข้อสอบต้องมีการแจกแจงแบบปกติ ถึง

จะประมาณค่าพารามิเตอร์ได้อย่างน่าเชื่อถือ ข้อดีของวิธีเบส์เซียน คือ สามารถประมาณค่าพารามิเตอร์โดยระบุการแจกแจงของความสามารถผู้สอบ และการแจกแจงพารามิเตอร์ข้อสอบได้ทุกรูปแบบ และให้ค่าอำนาจการทดสอบสูง ควบคุมความคลาดเคลื่อนประเภทที่ 1 ต่ำกว่าอีกสองวิธี ข้อจำกัดของวิธีเบส์เซียน คือ ใช้เวลาในการประมาณค่าพารามิเตอร์มาก มีความยุ่งยากในการเขียนโปรแกรมที่ใช้ในการวิเคราะห์ข้อมูลและ มีโปรแกรมที่ใช้ในการวิเคราะห์ยังไม่หลากหลายมากนัก ข้อดีของวิธีโพลี-ซิปเทสท์ คือ สามารถตรวจสอบได้กับขนาดตัวอย่างที่มีขนาดเล็ก และความยาวข้อสอบจำนวนน้อย โปรแกรมที่ใช้ในการวิเคราะห์ใช้งานง่าย สะดวกรวดเร็ว ข้อจำกัดของการตรวจสอบด้วยวิธีโพลี-ซิปเทสท์ คือ ไม่เหมาะกับขนาดตัวอย่างขนาดใหญ่ ความยาวของแบบสอบจำนวนมาก และขนาดของการทำหน้าที่ต่างกันของข้อสอบมีค่าสูง

เอกสารอ้างอิง

Atar, B. & Kamata, A. (2011, April). Comparison of IRT likelihood ratio test and logistic regression DIF detection procedures. **Hacettepe University Journal of Education**. 41(1) : 36-47.

Chang, H. H., Mazzeo, J., & Roussos, L. (1996, Autumn). Detecting DIF for polytornously scored items: An adaptation of the SIBTEST procedure. **Journal of Educational Measurement**. 33(3): 333-353.

Cohen, A. S., & Kim, S. H. (1993, March). A comparison of Lord's chi-square and Raju's area measures in detection of DIF. **Applied Psychological Measurement**. 17(1) : 39-52.

Hyun Jung Sung & Taehoon Kang. (2006). **Choosing a Polytomous IRT Model using Bayesian Model Selection Methods**. San Francisco : University of Wisconsin-Madiso.

Li-An Wu & Rung-Ching Tsai. (2010, January). **A comparison of three polytomous DIF detection methods**. The National Science Council of Taiwan (NSC 92-2413-H-003-068).

Narayanan, P. Y., & Swaminathan, H. (1996, September). Identification of items that show nonuniform DIF. **Applied Psychological Measurement**. 20(3) : 257-274.

Swaminathan, H., & Rogers, H. J. (1990, Winter). Detecting differential item functioning using logistic regression procedures. **Journal of Educational Measurement**. 27(4) : 361-370.