การเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA

Comparison of the Efficiency of Differential Item Functioning for Polytomous scored items: IRT LR, Poly-SIBTEST and Multiple-groups CFA Method

วาสนา กลมอ่อน*
klomoon@buu.ac.th
ไพรัตน์ วงษ์นาม**
สุรีพร อนุศาสนนันท์***

บทคัดย่อ

การวิจัยนี้มีวัตถุประสงค์เพื่อ 1) ตรวจสอบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-Groups CFA ภายใต้เงื่อนไขปัจจัยที่แตกต่าง3 ปัจจัย คือ ขนาดการทำหน้าที่ต่างกันของข้อสอบ 2 รูปแบบ ความยาวของแบบสอบ 2 รูปแบบ และขนาดของกลุ่มตัวอย่าง 3 ขนาด และ 2) เปรียบเทียบอัตราความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสามวิธี ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย โดยการจำลองข้อมูลภายใต้โมเดล Graded-Response และข้อสอบทุกข้อมีรายการคำตอบ 5 ตัวเลือก ให้คะแนนเป็น 0, 1, 2, 3 และ 4 คะแนน รวมจำนวน 12 เงื่อนไข (2x2x3) และในแต่ละเงื่อนไขจำลองข้อมูลวนซ้ำ 100 รอบ

ผลการวิจัยสรุปได้ดังนี้

1. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ภายใต้เงื่อนไขปัจจัยหลักที่แตกต่าง 3 ปัจจัย ด้วยวิธี IRT LR มีอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำกว่าเกณฑ์ที่กำหนด และอัตราอำนาจการทดสอบสูงกว่าเกณฑ์ที่กำหนดภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบที่มีขนาด

* นิสิตระดับดุษฎีบัณฑิต สาขาวิจัย วัดผล และสถิติการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา

** รองศาสตราจารย์ ดร. ภาควิชาวิจัยและจิตวิทยาประยุกต์ คณะศึกษาศาสตร์

*** ผู้ช่วยศาสตราจารย์ ดร. ภาควิชาวิจัยและจิตวิทยาประยุกต์ คณะศึกษาศาสตร์

กลาง สำหรับวิธี Poly-SIBTEST มีอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบ ไม่อยู่ในเกณฑ์ที่กำหนดเกือบทุกเงื่อนไขปัจจัย และวิธี Multiple-groups CFA มีอัตราความคลาดเคลื่อนประเภทที่ 1 สูงกว่าเกณฑ์ที่กำหนด และอัตราอำนาจการทดสอบ สูงกว่าเกณฑ์ที่กำหนดภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบที่มีขนาดกลาง

2. ผลการเปรียบเทียบอัตราความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบ การทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันสามวิธี พบว่า ความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของทั้งสามวิธีโดยรวม แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .001 นอกจากนี้ ผลของวิธีการตรวจสอบยังขึ้นอยู่กับปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง

คำสำคัญ : การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อัตราความคลาดเคลื่อนประเภทที่ 1 อัตราอำนาจการทดสอบ IRT LR Poly-SIBTEST

Abstract

The purpose of this research were: (1) to detecting of the efficiency of differential item functioning for polytomous scored items by using IRT LR, poly-SIBTEST and multiple-groups CFA method, and (2) to compare the Type I error rate and the power rate of investigated differential item functioning under a variety of three factors differences 3 factors: two levels forms of DIF magnitudes (small, medium), two levels forms of length test (9 items, 15 items), and three levels forms of sample size (200, 500, 1,000). These data were simulated under the unidimensional Graded-Response Model, and all items were in five response categories scoring as 0, 1, 2, 3 and 4. A total of 12 (2x2x3) conditions were studied.The data were replicated 100 times for each condition.

The research results were as follows:

1. The performance in differential item functioning (DIF) for polytomous scored items detecting under a variety of three factors differences 3 factors. Type I Error rate on IRT LR procedure was less than nominal limit and power rate was higher than nominal limit under medium magnitude of DIF. For Poly-SIBTEST procedure, Type I Error rate and Power rate were not nominal limits on almost conditions. And Type I Error rate on Multiple-groups CFA procedure was higher than nominal limit on overall conditions and power rate was higher than nominal limit under medium magnitude of DIF.

2. Results of the comparison of Type I error rate and Power rate by using DIF procedure on three methods found that Type I Error and Power on overall methods was statistically significant (α = 0.001). Moreover, result of methods depended on Magnitude of DIF, test length, and sample size.

Keywords : Detecting differential item functioning, Type I error rate, Power rate, IRT LR, Poly-SIBTEST

บทนำ

โครงสร้างการวัดทางจิตวิทยาไม่สามารถวัดได้โดยตรง ตัวแปรแฝงที่เป็นคุณลักษณะภายในจะถูกวัดโดยการสังเกตพฤติกรรมที่แสดงออกหรือการใช้ข้อสอบ/ข้อคำถาม ดังนั้นคุณสมบัติของตัวบุคคลและข้อสอบ/ข้อคำถามในมิติบนพื้นฐานทางจิตวิทยาจึงเป็นการอนุมานมาจากพฤติกรรม (Embretson & Reise, 2000, p.41) และเนื่องจากการวัดคุณลักษณะภายในของมนุษย์มีความสำคัญและจำเป็นต้องศึกษา เพื่อให้เข้าใจถึงการเกิดพฤติกรรมภายนอกของมนุษย์ อันจะนำไปสู่การทำนาย ควบคุม และพัฒนาพฤติกรรมมนุษย์ จึงจำเป็นต้องอาศัยทฤษฎีการทดสอบ เพื่อช่วยให้นักวัดผลสามารถทำการสร้างและพัฒนาแบบสอบให้มีคุณภาพ แปลความหมายผลการวัดได้อย่างถูกต้อง และใช้เป็นข้อมูลสารสนเทศได้อย่างเหมาะสม (ศิริชัย กาญจนวาสี, 2555, หน้า 9) ในวงการศึกษาและหน่วยงานต่างๆ มักนำผลการทดสอบมาใช้เป็นข้อมูลเพื่อตัดสินใจในเรื่องต่างๆ มากขึ้น ซึ่งแบบทดสอบที่นำมาใช้ควรเป็นแบบทดสอบมาตรฐานที่มีความเที่ยงตรง ความเชื่อมั่น ความยากที่เหมาะสม ตลอดจนจำแนกระดับความสามารถของผู้สอบได้ นอกจากนี้ ยังต้องคำนึงถึงความยุติธรรมต่อผู้สอบด้วย (อรินทร์ น่วมถนอม, 2549, หน้า 137) ความยุติธรรมของแบบสอบเป็นประเด็นสำคัญในการทดสอบทางการศึกษาและจิตวิทยา ผลการทดสอบที่ได้ไม่เพียงเกี่ยวข้องกับข้อสอบ/ ข้อคำถามที่วัดความสามารถของผู้สอบเท่านั้น แต่ยังมีปัจจัยอื่นๆ ที่เกี่ยวข้องด้วย เช่น การสอบคัดเลือกเข้าศึกษาระดับนานาชาติ สิ่งสำคัญของการสอบต้องไม่มีข้อสอบ/ ข้อคำถามที่เข้าข้างนักเรียนในเขตพื้นที่ตั้งหรือนักเรียนที่มีฐานะทางเศรษฐกิจและสังคม (Chang, Huang &Tsai, 2015, p. 181) ข้อสอบที่เข้าข้างผู้สอบกลุ่มใดกลุ่มหนึ่ง มีผลทำให้ผู้สอบกลุ่มนั้นได้เปรียบ ส่วนผู้สอบอีกกลุ่มหนึ่งเสียเปรียบทั้งๆ ที่ผู้สอบทั้งสองกลุ่มมีระดับความสามารถเท่ากันแบบสอบนั้น คือแบบสอบที่ขาดความยุติธรรม ซึ่งเป็นผลจากความลำเอียงของข้อสอบ (Item bias) และความลำเอียงของแบบสอบ (Test bias) (อรินทร์ น่วมถนอม, 2549, หน้า 137) ดังนั้นจึงจำเป็นต้องมีการตรวจสอบความลำเอียงเพื่อจำแนกข้อสอบที่ทำหน้าที่ไม่เหมาะสมหรือไม่ยุติธรรม ซึ่งต่อมาได้มีการเปลี่ยนมาใช้คำว่า การทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) โดยเน้นการใช้วิธีการทางสถิติสำหรับการตรวจสอบข้อสอบเพื่อเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มผู้สอบที่เป็นกลุ่มเปรียบเทียบ (Focal group: กลุ่ม F) เป็นกลุ่มที่สนใจศึกษาและคาดว่าจะเป็นกลุ่มที่เสียเปรียบในการตอบข้อสอบ กับกลุ่มอ้างอิง (Reference group: กลุ่ม R) เป็นกลุ่มที่คาดว่าจะได้เปรียบในการตอบข้อสอบได้ถูกต้อง (ศิริชัย กาญจนวาสี, 2555, หน้า 115-118) ทั้งนี้ การทำหน้าที่ต่างกันของข้อสอบเกิดขึ้นเมื่อผู้สอบจากกลุ่มที่แตกต่างกัน และมีการจับคู่ความสามารถตามที่ข้อสอบหรือแบบทดสอบต้องการวัดเท่ากัน มีความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องไม่เท่ากัน (Clauser & Mazor, 1998; Zumbo, 1999 อ้างถึงในอรินทร์ น่วมถนอม, 2549 หน้า 138)

ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) เป็นทฤษฎีที่ขยายแนวคิดมาจากทฤษฎีการทดสอบแบบดั้งเดิม (Classical Test Theory: CTT) (Ostini & Nering, 2006,p. 2) ซึ่งหลักการสำคัญของ IRT มุ่งอธิบายความสัมพันธ์ระหว่างคุณลักษณะภายในหรือความสามารถที่มีอยู่ภายในตัวบุคคล กับพฤติกรรมการตอบสนองข้อสอบของบุคคลนั้น ในรูปของฟังก์ชันทางคณิตศาสตร์หรือโมเดลที่แสดงความสัมพันธ์ระหว่างระดับความสามารถ คุณลักษณะของข้อสอบ และโอกาสของการตอบข้อสอบได้ถูก (ศิริชัย กาญจนวาสี, 2555, หน้า 53) วิธีการวัดของ IRT อยู่บนพื้นฐานของโมเดลและการเปรียบเทียบค่าพารามิเตอร์ที่เปลี่ยนไปในแต่ละโมเดลเช่น โมเดล 1-พารามิเตอร์ จะเป็นการเปรียบเทียบค่าพารามิเตอร์ความยากของข้อสอบ (b) ของแต่ละกลุ่ม

โมเดล 2-พารามิเตอร์ เป็นการเปรียบเทียบค่าพารามิเตอร์อำนาจจำแนก (a) และค่าพารามิเตอร์ความยากของข้อสอบ (b) ในแต่ละกลุ่ม แต่ถ้าเป็นการเปรียบเทียบระหว่างกลุ่มค่าพารามิเตอร์ความยากของข้อสอบ (a) แตกต่างกัน แสดงให้เห็นว่าเป็นโมเดลการตอบสนองข้อสอบที่เป็นรูปแบบเดียวกัน (Uniform DIF) และหากค่าพารามิเตอร์อำนาจจำแนก (a) แตกต่างกัน แสดงให้เห็นว่าเป็นโมเดลการตอบสนองข้อสอบที่ไม่เป็นรูปแบบเดียวกัน (Non-Uniform DIF) เป็นต้น ทั้งนี้ โมเดลการตอบสนองข้อสอบมีทั้งแบบตรวจให้คะแนน 2 ค่า และมากกว่า 2 ค่า ซึ่งโมเดลการตอบสนองข้อสอบแบบตรวจให้คะแนนมากกว่า 2 ค่า มักพบในแบบสอบทางการศึกษาและแบบทดสอบทางจิตวิทยา (Nering & Ostini, 2010, p.3) โดยเฉพาะโมเดล Graded- Response (GRM) ที่นำเสนอโดย Samejima (1969, 1996) พัฒนามาเพื่อใช้กับแบบสอบหรือแบบวัดที่แต่ละข้อคำถามมีรายการคำตอบแบบมาตรเรียงลำดับ (Ordered Categorical Responses) และใช้หลักการคำนวณความน่าจะเป็นของการตอบแต่ละรายการคำตอบแบบ 2 ขั้นตอน โดยขั้นตอนแรกคือ คำนวณค่าความชันร่วมของแต่ละข้อคำถาม (α) และขั้นตอนที่สอง คำนวณค่าพารามิเตอร์ของแต่ละรายการคำตอบในแต่ละข้อคำถาม (β) (ศิริชัย กาญจนวาสี, 2555, หน้า 89)

จากการศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี IRT LR (Likelihood Ratio Test) และวิธีถดถอยโลจิสติก พบว่า วิธี IRT LR และวิธีถดถอยโลจิสติก สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดี และวิธี IRT LR มีอำนาจการทดสอบสูงภายใต้ขนาดของกลุ่มตัวอย่างที่มีขนาดกลางและขนาดใหญ่ และเมื่อขนาดของกลุ่มตัวอย่างมีขนาดใหญ่ขึ้น วิธี IRT LR มีอำนาจการทดสอบสูงขึ้นด้วย (Atar & Kamata, 2011, p.36) สำหรับวิธี Poly-SIBTEST นี้มีจุดเด่นหลายประการ เช่น ใช้เทคนิคการตรวจสอบแบบหลายมิติ โดยแยกข้อสอบที่ใช้เป็นเกณฑ์การจับคู่ออกจากข้อสอบที่ใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบอย่างชัดเจน ทำให้เกณฑ์มีความเที่ยงตรงและมีความเชื่อมั่นสูง คะแนนเกณฑ์การจับคู่ค่อนข้างเป็นคุณลักษณะแฝงมากกว่าคะแนนที่ได้จากการสอบ ทำให้มีความถูกต้องและแม่นยำ มีการคำนวณทวนซ้ำหลายรอบ (Iterative algorithm) เพื่อคัดเลือกข้อสอบที่ทำหน้าที่ต่างกันออกจากคะแนนเกณฑ์การจับคู่ ทำให้เกณฑ์มีความบริสุทธิ์ (Purification) มีการปรับแก้ค่าการถดถอย (Regression correction) เพื่อลดความแตกต่างของค่าความสามารถเป้าหมายระหว่างกลุ่มผู้สอบ ทำให้สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ที่สูงเกินปกติได้ (อรินทร์ น่วมถนอม, 2549, หน้า 13) นอกจากนี้ในปัจจุบันยังมีการนำวิธีในกลุ่มของ SEM มาใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมากขึ้น ดังเช่น งานวิจัยของ Meade & Lautenschlager (2004) และMuthen & Asparouhov (2014) เป็นต้น ซึ่งในกระบวนการตรวจสอบ DIF พบปัจจัยที่เกี่ยวข้องหลายปัจจัย ดังเช่น ปัจจัยความยาวของแบบสอบ ขนาดของกลุ่มตัวอย่าง ความแตกต่างของค่าเบี่ยงเบนมาตรฐาน ความแตกต่างของการแจกแจงข้อมูล และปฏิสัมพันธ์ของปัจจัยต่างๆ (Ackerman & Evans, 1992; Finch, 2005; Finch & French, 2007; Kim, 2010; Narayanan & Swaminathan, 1994; Prieto, Barbero, & San Luis, 1997; Rogers & Swaminathan, 1993; Roussos & Stout, 1996; Shealy & Stout, 1993 cited in Atalay Kabasakal, K., Arsan, N., Gök, B., & Elecioglu, H., 2014, p. 2187) ทั้งนี้ การแสดงหลักฐานเกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบเป็นประเด็นสำคัญหนึ่งในการทดสอบทางการศึกษาและจิตวิทยา ซึ่งมีความสำคัญไม่น้อยไปกว่าการแสดงหลักฐานเกี่ยวกับความเที่ยงตรง ความเชื่อมั่นหรือคุณลักษณะของข้อสอบ (อรินทร์ น่วมถนอม, 2549, หน้า 138)จากแนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ผู้วิจัย

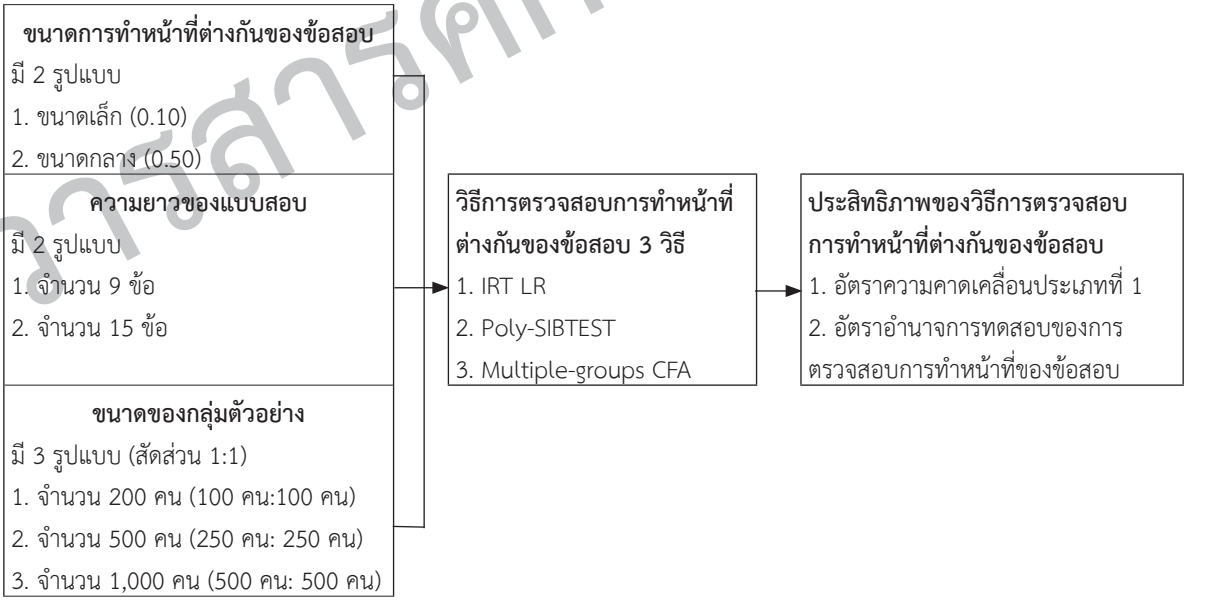
จึงมีความสนใจการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันสามวิธี คือ วิธี IRT LR (Likelihood Ratio Test) (Thissen, Steinberg, & Wainer, 1986) วิธี Poly-SIBTEST (Chang, Mazzeo, & Roussos, 1996) และวิธี Multiple-groups CFA (Kim & Yoon, 2011) ภายใต้โมเดล GRM และเงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย คือ ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง

คำถามวิจัย

1 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-Groups CFA ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 4 ปัจจัย คือ ขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง มีประสิทธิภาพในการตรวจสอบอัตราความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ แตกต่างกันอย่างไร

2 อัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ 3 วิธี ภายใต้เงื่อนไขปัจจัยหลักที่แตกต่างกัน 3 ปัจจัย แตกต่างกันอย่างไร

กรอบแนวคิดในการวิจัย



ภาพที่ 1 กรอบแนวคิดในการวิจัย

วัตถุประสงค์ของการวิจัย

1.เพื่อตรวจสอบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบการสนองข้อสอบแบบมิติเดียว ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-Groups CFA ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย คือ ขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง

2. เพื่อเปรียบเทียบอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบสนองแบบสอบมิติเดียว ด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสามวิธี

ขอบเขตการวิจัย

1. การศึกษาในครั้งนี้เป็นการศึกษาในสถานการณ์จำลอง โดยจำลองข้อมูล (Simulation Data) ภายใต้โมเดล Graded Response (GRM) ตามทฤษฎีการตอบสนองข้อสอบแบบมิติเดียว (Unidimensional Item Response Theory) และจำลองข้อมูลการตอบข้อสอบที่มีโครงสร้างวัดความสามารถแบบมิติเดียวที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า (Polytomous Item) โดยรูปแบบของแบบทดสอบ มี 2 รูปแบบ ประกอบด้วย รูปแบบที่ 1 ข้อสอบมีจำนวน 9 ข้อ และรูปแบบที่ 2 ข้อสอบมีจำนวน 15 ข้อ และข้อสอบทุกข้อมีรูปแบบรายการคำตอบแบบ 5 ตัวเลือก ซึ่งมีรายการคำตอบเป็น 0, 1, 2, 3 และ 4 โดยใช้การจำลองผลการตอบข้อสอบภายใต้เงื่อนไขปัจจัยที่แตกต่างกัน 3 ปัจจัย คือ ขนาดการทำหน้าที่ต่างกันของข้อสอบ 2 รูปแบบ (ขนาดเล็ก และขนาดกลาง) ความยาวของแบบสอบ 2 รูปแบบ (9 ข้อ และ 15 ข้อ) และขนาดกลุ่มตัวอย่าง 3 ขนาด (200 คน, 500 คน และ 1,000 คน) ในสัดส่วน 1:1 เพื่อใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ รวมเป็น 12 เงื่อนไข (2x2x3) และในแต่ละเงื่อนไขจำลองข้อมูลวนซ้ำ 100 รอบ

2. ตัวแปรที่ศึกษา ประกอบด้วย ตัวแปรอิสระ มี 3 ตัวแปร คือ ขนาดการทำหน้าที่ต่างกันของข้อสอบ มี 2 รูปแบบ (ขนาดเล็ก (0.10), ขนาดกลาง (0.50)) ความยาวของแบบสอบ มี 2 รูปแบบ (จำนวน9 ข้อ, จำนวน 15 ข้อ) และขนาดของกลุ่มตัวอย่าง มี 3 ขนาด สัดส่วน 1:1 (100 คน:100 คน, 250 คน:250 คน และ 500 คน:500 คน) และตัวแปรตาม มี 2 ตัวแปร คือ อัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

วิธีดำเนินการวิจัย

1. จำลองข้อมูลผลการตอบของผู้สอบที่มีโครงสร้างการวัดแบบมิติเดียว (Unidimensional) ที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า โดยใช้โมเดล Graded-Response ซึ่งผลการตอบข้อสอบแต่ละรายการให้คะแนนเป็น 0, 1, 2, 3 และ 4 และใช้การจำลองข้อมูลที่มีการแจกแจงแบบปกติ และผู้สอบมีระดับความสามารถของกลุ่มเปรียบเทียบและกลุ่มอ้างอิงเท่ากัน (M=0, SD=1) โดยเลือกการแจกแจงของค่าพารามิเตอร์ข้อสอบ ได้แก่ ค่าอำนาจจำแนก $a_j \sim LogN$ (0,0.03) และค่าความยาก $b \sim N$ (0,1) และสัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน กำหนดให้มีจำนวน 10% ด้วยโปรแกรม WinGen ภายใต้เงื่อนไขปัจจัยหลักที่แตกต่างกัน 3 ปัจจัย ประกอบด้วย ขนาดการทำหน้าที่ต่างกันของข้อสอบ 2 รูปแบบ (ขนาดเล็ก = 0.10, ขนาดกลาง = 0.5) ความยาวของแบบสอบ 2 รูปแบบ (จำนวน 9 ข้อ, จำนวน 15 ข้อ) และขนาดของกลุ่มตัวอย่าง 3 ขนาด (200 คน, 500 คน, 1,000 คน) ในสัดส่วน 1:1 รวมข้อมูลที่ศึกษาทั้งหมด 12 เงื่อนไข (2x2x3) และในแต่ละเงื่อนไขจำลองข้อมูลวนซ้ำ 100 รอบ ได้จำนวนชุดข้อมูลทั้งหมดที่ใช้ในการศึกษา 1,200 ชุด

2. ตรวจสอบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-Groups CFA ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยโดยใช้โปรแกรม IRTPRO Version 4.1, โปรแกรม DIF Analysis Version 1.7 และโปรแกรม Mplus Version 7.14 และสำหรับเกณฑ์ที่ใช้ในการ

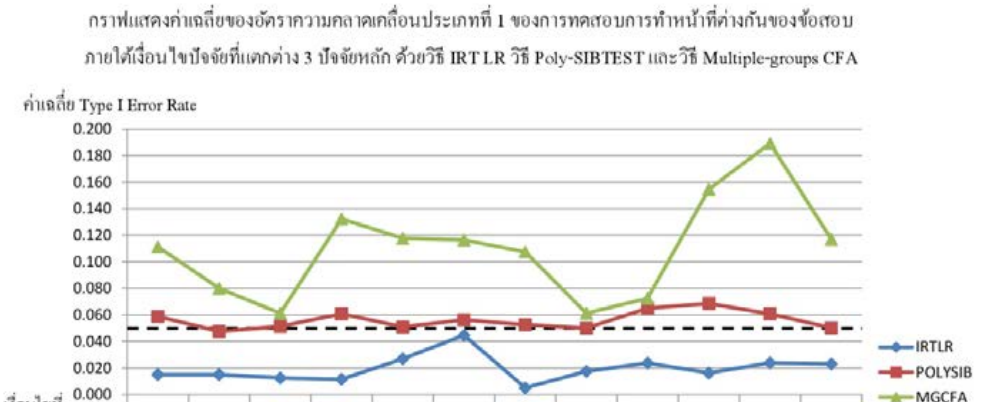
ตัดสินประสิทธิภาพของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ พิจารณาจากผลการตรวจสอบค่าอัตราความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ใช้เกณฑ์การพิจารณาดังนี้ ค่าความคลาดเคลื่อนประเภทที่ 1 ต้องมีค่าเฉลี่ยต่ำกว่าหรือเท่ากับ 0.05 ซึ่งถือว่า ควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดี และการพิจารณาค่าอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ จะพิจารณาเมื่อสามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ก่อน โดยค่าอำนาจการทดสอบ ต้องมีค่าเฉลี่ยตั้งแต่ 0.80 ขึ้นไป จึงถือว่ามีอำนาจการทดสอบเพียงพอ (Sufficient power) และหากมีค่าเฉลี่ยต่ำกว่า 0.80 ถือว่าวิธีนั้นๆ ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ไม่ดี (Atar & Kamata, 2011, p.40 และอาวีพร ปานทอง, 2558, หน้า 81) และดำเนินการทดสอบสมมติฐานที่ระดับนัยสำคัญทางสถิติ .05

3. เปรียบเทียบอัตราความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสามวิธีภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย โดยใช้วิธีการวิเคราะห์ความแปรปรวนแบบวัดซ้ำ (Repeated Measurement) ซึ่งมีตัวแปรวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นตัวแปรวัดซ้ำ (Within-Subject Factor) และปัจจัยที่แตกต่าง 3 ปัจจัย เป็นตัวแปรวัดต่างกลุ่ม (Between-Subjects Factors)

ผลการวิจัย

ผลการวิจัยขอนำเสนอเป็น 2 ส่วน ดังนี้

1. ผลการตรวจสอบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-Groups CFA ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย มีรายละเอียดดังนี้

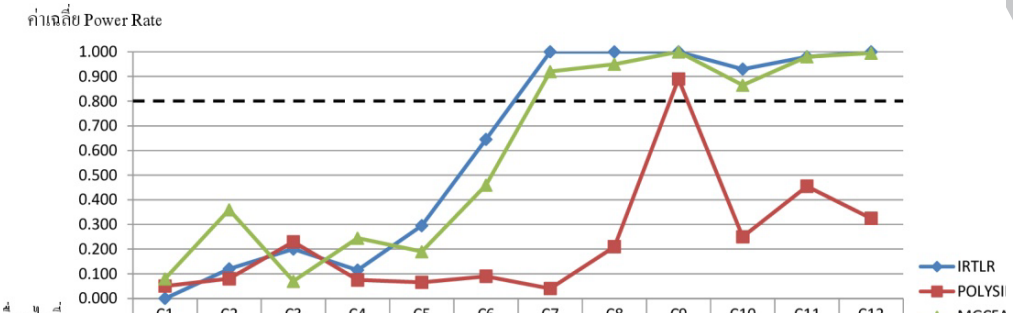


เงื่อนไขที่	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
IRTLR	0.015	0.015	0.013	0.012	0.027	0.045	0.005	0.018	0.024	0.016	0.024	0.023
POLYSIB	0.059	0.048	0.051	0.061	0.051	0.056	0.053	0.050	0.065	0.069	0.061	0.050
MGCFA	0.111	0.080	0.061	0.132	0.118	0.116	0.108	0.061	0.073	0.155	0.189	0.117

ภาพที่ 1 กราฟแสดงค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 (Type I error Rate) ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA

จากภาพที่ 1 ผลการวิเคราะห์ค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA ภายใต้ปัจจัยที่แตกต่าง 3 ปัจจัย พบว่า โดยภาพรวมของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี IRT LR มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 อยู่ในเกณฑ์ที่กำหนด (ต่ำกว่า 0.05) ทุกเงื่อนไขปัจจัย และวิธี Poly-SIBTEST มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 อยู่ในเกณฑ์ที่กำหนด (ต่ำกว่าหรือเท่ากับ 0.05) เพียง 3 เงื่อนไข สำหรับวิธี Multiple-groups CFA มีค่าเฉลี่ยอัตราความคลาดเคลื่อนประเภทที่ 1 สูงกว่าเกณฑ์ที่กำหนดในทุกเงื่อนไขปัจจัย

กราฟแสดงค่าเฉลี่ยของอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ
ภายใต้ปัจจัยที่แตกต่าง 3 ปัจจัยหลัก ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA



เงื่อนไขที่	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
IRTLR	0.000	0.120	0.200	0.115	0.295	0.645	1.000	1.000	1.000	0.930	0.980	1.000
POLYSIB	0.050	0.080	0.230	0.075	0.065	0.090	0.040	0.210	0.890	0.250	0.455	0.325
MGCFA	0.080	0.360	0.070	0.245	0.190	0.460	0.920	0.950	1.000	0.865	0.980	0.995

ภาพที่ 2 กราฟแสดงค่าเฉลี่ยของอัตราอำนาจการทดสอบ (Power Rate) ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA

จากภาพที่ 2 ผลการวิเคราะห์ค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย พบว่า โดยภาพรวมของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี IRT LR และวิธี Multiple-groups CFA ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อยู่ในเกณฑ์ที่กำหนด (สูงกว่า 0.80) เพียง 6 เงื่อนไขปัจจัย คือ เงื่อนไขปัจจัยที่ 7-12 (C7-C12) ภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ขนาดกลาง 0.50 ในทุกเงื่อนไขปัจจัย สำหรับวิธี Poly-SIBTEST มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อยู่ในเกณฑ์ที่กำหนด (สูงกว่า 0.80) เพียง 1 เงื่อนไขปัจจัย คือ เงื่อนไขที่ 9 (C9) ภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ขนาดกลาง (0.50) ความยาวของแบบสอบ จำนวน 9 ข้อ และขนาดของกลุ่มตัวอย่าง 500 คน:500 คน

ตารางที่ 1 ผลการทดสอบนัยสำคัญของอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบ ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย

ขนาดการทำหน้าที่ต่างกันของข้อสอบ	ความยาวของแบบสอบ	ขนาดของกลุ่มตัวอย่าง (N_F:N_R)	อัตราความคลาดเคลื่อนประเภทที่ 1			อัตราอำนาจการทดสอบ		
			IRTLR	POLYSIB	MGCFA	IRTLR	POLYSIB	MGCFA
ขนาดเล็ก (0.10)	จำนวน 9 ข้อ	100:100	-1.606*	0.401*	2.810	-20.000	-18.750	-18.000
		250:250	-1.606*	-0.115*	1.376*	-17.000	-18.000	-11.000
		500:500	-1.721*	0.057*	0.516*	-15.000	-14.250	-18.250
	จำนวน 15 ข้อ	100:100	-1.765*	0.497*	3.776	-17.125	-18.125	-13.875
		250:250	-1.059*	0.038*	3.105	-12.625	-18.375	-15.250
		500:500	-0.248*	0.285*	3.048	-3.875	-17.750	-8.500
ขนาดกลาง (0.50)	จำนวน 9 ข้อ	100:100	-2.065*	0.115*	2.638	5.000**	-19.000	3.000**
		250:250	-1.491*	0.000*	0.516*	5.000**	-14.750	3.750**
		500:500	-1.204*	0.688*	1.032*	5.000**	2.250**	5.000**
	จำนวน 15 ข้อ	100:100	-1.553*	0.850*	4.799	3.250**	-13.750	1.625**
		250:250	-1.200*	0.497*	6.387	4.500**	-8.625	4.500**
		500:500	-1.236*	0.002*	3.071	5.000**	-11.875	4.875**

*$Z_{.05} < 1.645$, **$Z_{.05} > -1.645$

จากตารางที่ 1 พบว่า วิธี IRT LR และวิธี Poly-SIBTEST สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ในทุกเงื่อนไขปัจจัย สำหรับวิธี Multiple-groups CFA สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 เพียง 2 เงื่อนไขปัจจัย และเมื่อพิจารณาอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ของข้อสอบ ขนาดเล็ก (0.10) ทั้งสามวิธี ไม่มีอำนาจการทดสอบ แต่เมื่อขนาดการทำหน้าที่ต่างกันของข้อสอบ เพิ่มขึ้น คือ ขนาดกลาง (0.50) วิธี IRT LR และวิธี Multiple-groups CFA มีอำนาจการทดสอบในทุกเงื่อนไขปัจจัย อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 สำหรับวิธี Poly-SIBTEST มีอำนาจการทดสอบ เพียง 1 เงื่อนไข คือ เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ขนาดกลาง (0.50) ความยาวของแบบสอบจำนวน 9 ข้อ และขนาดของกลุ่มตัวอย่าง 500 คน: 500 คน

2. ผลการเปรียบเทียบอัตราคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสามวิธี ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก

2.1 ผลการวิเคราะห์เปรียบเทียบค่าเฉลี่ยอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสามวิธี (Tests of Within-Subjects Effects) พบว่า วิธี

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ มีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .001 (F = 408.973) และมีผลต่ออัตราอำนาจการทดสอบ อย่างมีนัยสำคัญทางสถิติที่ระดับ .001 (F = 858.326) นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแตกต่างกัน มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบ แตกต่างกัน

2.2 ผลการวิเคราะห์เปรียบเทียบค่าเฉลี่ยอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง (Tests ofBetween-Subjects Effects) พบว่า เงื่อนไขปัจจัยความยาวของแบบสอบมีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .001 (F = 64.604) สำหรับเงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบและปัจจัยขนาดของกลุ่มตัวอย่าง ไม่มีผลต่อความเคลื่อนประเภทที่ 1 นั่นคือ ความยาวของแบบสอบแตกต่างกัน มีผลให้ความคลาดเคลื่อนประเภทที่ 1 แตกต่างกัน ในขณะที่ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง แตกต่างกัน ไม่มีผลให้ความคลาดเคลื่อนประเภทที่ 1 แตกต่างกัน และเมื่อพิจารณาด้านอำนาจการทดสอบ พบว่า เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง ทั้งสามปัจจัยมีผลต่ออัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ นั่นคือ ขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง ทั้งสามปัจจัย มีผลให้อำนาจการทดสอบแตกต่างกัน

อภิปรายผลการวิจัย

1. ผลการตรวจสอบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ด้วยวิธีIRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย

1.1 จากการตรวจสอบประสิทธิภาพการควบคุมความคลาดเคลื่อนประเภทที่ 1 ของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย คือ ขนาดการทำหน้าที่แตกต่างกัน ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง ด้วยวิธี IRT LR และวิธี Poly-SIBTEST สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดีในทุกเงื่อนไขปัจจัย สอดคล้องกับผลการศึกษาความไม่แปรเปลี่ยนในการวัดแบบสอบ โดยการเปรียบเทียบด้วยวิธี Multiple-group Categorical CFA (MCCFA) กับวิธีการทดสอบ Likelihood Ratio Chi-Square Difference (IRT LR) ของ Kim & Yoon (2011) พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ด้วยวิธี IRT LR มีอัตราความความเคลื่อนประเภทที่ 1 ต่ำกว่าวิธี MCCFA เกือบทุกเงื่อนไข และ Lopez Rivas, Stark, & Chernyshenko (2009) พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีขนาดการทำหน้าที่ต่างกันของข้อสอบขนาดเล็ก และค่าอำนาจจำแนกมีค่าต่ำ (a = 0.6) ด้วยวิธี IRT LR สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดี และมีอำนาจการทดสอบสูง นอกจากนี้ผลการศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในโมเดล GRM ของ Cohen, Kim, & Baker (1993) พบว่า เมื่อความยาวของแบบสอบเพิ่มขึ้น วิธี Poly-SIBTEST มีอัตราความคลาดเคลื่อนประเภทที่ 1 เพิ่มขึ้นด้วย สำหรับผลการตรวจสอบการทำหน้าที่ต่างกัน

ของข้อสอบด้วยวิธี Poly-SIBTEST สอดคล้องกับ อุทัยวรรณ สายพัฒนะ (2547) พบว่า วิธี Polytomous SIBTEST มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูงกว่าวิธี Generalized Mantel-Haenszel โดย อรินทร์ น่วมถนอม (2549) กล่าวว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี Poly-SIBTEST มีอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำ เนื่องมาจากวิธี Poly-SIBTEST ใช้เทคนิคการตรวจสอบแบบหลายมิติ ปัจจัยเกี่ยวกับสัดส่วนของการทำหน้าที่ต่างกันของข้อสอบ ความแตกต่างของการแจกแจงความสามารถ และขนาดของกลุ่มตัวอย่าง จึงมีผลกระทบต่ออัตราความคลาดเคลื่อนประเภทที่ 1 น้อยมาก รวมถึงมีการคำนวณที่ง่าย ไม่ซับซ้อน และไม่จำเป็นต้องใช้กลุ่มตัวอย่างขนาดใหญ่ (Chang, Mazzeo and Roussos, 1995 cilted in Potenza and Doran, 1995, p.31 อ้างถึงใน อุทัยวรรณ สายพัฒนะ, 2547, หน้า 19)

1.2 จากการตรวจสอบประสิทธิภาพอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย คือ ขนาดการทำหน้าที่แตกต่างกัน ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง ด้วยวิธี IRT LR และวิธี Multiple-groups CFA มีอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูง เมื่อปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบเพิ่มขึ้น และมีอำนาจการทดสอบสูงกว่าวิธี Poly-SIBTEST เกือบทุกเงื่อนไข ซึ่งสอดคล้องกับผลการศึกษาของ Kim & Yoon (2011) ที่พบว่า ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ด้วยวิธี MCCFA และวิธี IRT LR มีอำนาจการทดสอบสูงเมื่อขนาดการทำหน้าที่ต่างกันของข้อสอบมีขนาดใหญ่ (0.04) และมีขนาดของกลุ่มตัวอย่างเพิ่มมากขึ้นเกือบทุกเงื่อนไขปัจจัย และสอดคล้องกับผลการศึกษาของ ทองอยู่ สาระ (2543), วลีมาศ แซ่อึ้ง (2543), French & Miller (1996), Krisjansson และคณะ (2005), Narayanan & Swaminathan (1994, 1996), Roger & Swaminathan (1993), Whitmore & Schumacker (1999) อ้างถึงใน อรินทร์ น่วมถนอม (2549) พบว่า เมื่อขนาดของกลุ่มตัวอย่างเพิ่มขึ้นมีผลทำให้วิธี Poly-SIBTEST มีอัตราอำนาจการทดสอบเพิ่มขึ้นด้วย นอกจากนี้ Kim & Yoon (2011) พบว่า วิธี Multiple-group Categorical CFA (MCCFA) มีอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ดีเมื่อขนาดการทำหน้าที่มีขนาดใหญ่ ทั้งนี้ Chang, Mazzeo & Roussos (1996) ได้ปรับขยายวิธี Poly-SIBTEST มาจากวิธี SIBTEST ของ Shealy & Stout (1993) เพื่อใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ซึ่งมีข้อจำกัดคือไม่สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบที่ไม่เป็นรูปแบบเดียวกัน (Nonuniform DIF) จึงส่งผลให้อัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ มีค่าต่ำ แต่มีข้อได้เปรียบคือ สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดี (อรินทร์ น่วมถนอม, 2549, หน้า 143-144)

2. ผลการเปรียบเทียบอัตราคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสามวิธี ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย

จากการวิเคราะห์เปรียบเทียบค่าเฉลี่ยอัตราความคลาดเคลื่อนประเภทที่ 1 ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย คือ ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง (Tests of Between-Subjects Effects) ที่แตกต่างกัน มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 มีค่าเฉลี่ยแตกต่างกัน โดยเฉพาะปัจจัยความยาวของแบบสอบ มีผลให้อัตราความคลาดเคลื่อน

ประเภทที่ 1 มีค่าเฉลี่ยแตกต่างกัน สอดคล้องกับผลการศึกษาของ Cohen & Kim (1993) พบว่า เมื่อความยาวของแบบสอบเพิ่มขึ้น วิธีโพลี-ซิปเทสท์มีอัตราความคลาดเคลื่อนประเภทที่ 1 เพิ่มขึ้นด้วย โดย Potenza & Dorans (1995) และ อาวีพร ปานทอง (2558) พบว่า วิธี Poly-SIBTEST สามารถใช้ได้ดีกับแบบสอบสั้นที่มีข้อสอบหรือข้อคำถามจำนวนน้อย และ Chang, Mazzeo & Roussos (1996) ยังพบว่า เมื่อขนาดของกลุ่มตัวอย่างเพิ่มขึ้น ไม่มีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 1 ซึ่งสอดคล้องกับผลการศึกษาของ Bolt (2002) พบว่า เมื่อขนาดของกลุ่มตัวอย่างเพิ่มขึ้นจาก 300 คน เป็น 1,000 คนต่อกลุ่ม การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี Poly-SIBTEST มีอัตราความคลาดเคลื่อนประเภทที่ 1ค่อนข้างคงที่ สำหรับการวิเคราะห์เปรียบเทียบค่าเฉลี่ยอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย คือ ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง (Tests of Between-Subjects Effects) ที่แตกต่างกัน มีผลให้อัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ มีค่าเฉลี่ยแตกต่างกัน ในทุกเงื่อนไขปัจจัย ซึ่งสอดคล้องกับ Lopez Rivas, Stark, & Chernyshenko (2009) พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี IRT LR สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดี และมีอำนาจการทดสอบสูง เมื่อขนาดการทำหน้าที่ต่างของข้อสอบมีขนาดใหญ่ และขนาดของกลุ่มตัวอย่างเพิ่มขึ้น และผลการศึกษาของ Narayanan & Swaminathan (1994, p.315-328) ที่ได้ศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่เป็นรูปแบบเดียวกัน และไม่เป็นรูปแบบเดียวกัน ระว่างวิธีแมนเทล-แฮนส์เซล กับวิธีซิปเทสท์ พบว่า ปัจจัยขนาดของกลุ่มตัวอย่าง การแจกแจงความสามารถสัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน ขนาดการทำหน้าที่ต่างกันของข้อสอบและประเภทของข้อสอบ มีผลต่ออัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีซิปเทสท์ อย่างมีนัยสำคัญ และผลการศึกษาของอุทัยวรรณ สายพัฒนะ (2547) พบว่า ขนาดของกลุ่มตัวอย่างมีผลต่อประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยเสนอแนะว่า ในการศึกษาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าด้วยวิธี Poly-SIBTEST ควรใช้กลุ่มตัวอย่างที่มีขนาดไม่ต่ำกว่า 500 คน หรือกลุ่มตัวอย่างที่มีขนาดใหญ่ จะทำให้ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมีอำนาจการทดสอบสูง ถึงแม้จะมีความคลาดเคลื่อนประเภทที่ 1 สูงขึ้นด้วย

ข้อเสนอแนะ

ข้อเสนอแนะสำหรับการนำไปใช้

1. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดล Graded-Response แบบมิติเดียว ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย คือขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง ในทางปฏิบัติสถานการณ์จริง ควรใช้วิธี IRT LR ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเนื่องจากมีประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูง โดยสามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดี และมีอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูงในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบค่าในโมเดล Graded-Response แบบมิติเดียว ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA ควรคำนึงถึงปัจจัยความยาวของแบบสอบที่เหมาะสมสำหรับการนำไปใช้ เนื่องจากการศึกษาวิจัยครั้งนี้พบว่า ความยาวของแบบสอบ มีผล

ให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกัน ซึ่งวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี IRT LR และวิธี Poly-SIBTEST ทั้ง 2 วิธี สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ดีภายใต้เงื่อนไขปัจจัยความยาวของแบบสอบ จำนวน 9 ข้อ และ 15 ข้อ แต่อัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธี Poly-SIBTEST มีแนวโน้มเพิ่มขึ้น เมื่อความยาวของแบบสอบเพิ่มขึ้น รวมถึงมีอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบต่ำลง ในขณะที่วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี Multiple-groups CFA ไม่สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้เมื่อความยาวของแบบสอบเพิ่มขึ้น ดังนั้น ในทางปฏิบัติหากทำการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีความยาวของแบบสอบสั้น มีข้อสอบหรือข้อคำถามจำนวนน้อย และขนาดการทำหน้าที่ต่างกันของข้อสอบมีขนาดเล็กหรือขนาดกลาง จึงควรเลือกใช้วิธี IRT LR ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ซึ่งจะทำให้การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมีประสิทธิภาพสูง

ข้อเสนอแนะในการวิจัยครั้งต่อไป

1.จากการศึกษาครั้งนี้พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี Multiple-groups CFA มีประสิทธิภาพต่ำกว่าวิธี IRT LR และวิธี Poly-SIBTEST แต่มีอำนาจการทดสอบ ใกล้เคียงกับวิธี IRT LR จึงควรทำการศึกษาในลักษณะเดียวกัน โดยศึกษาปัจจัยอื่นที่คาดว่าจะมีผลต่อความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบ เช่น จำนวนข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ ค่าอำนาจจำแนก ค่าความยากของข้อสอบ ประเภทของข้อสอบที่ทำหน้าที่ต่างกันแบบ Uniform และแบบ Nonuniform เป็นต้น เพื่อให้ได้ข้อมูลสารสนเทศที่ครอบคลุมและมีความน่าเชื่อถือสำหรับการนำไปใช้ได้จริงมากขึ้น

2. การศึกษาครั้งนี้เป็นการศึกษาโดยใช้โมเดลการตอบสนองข้อสอบแบบมิติเดียว (Unidimensional) ดังนั้น เพื่อให้เกิดความหลากหลายในการศึกษาข้อมูลเชิงลึก จึงควรมีการศึกษาโดยใช้โมเดลการตอบสนองข้อสอบแบบหลายมิติ (Multidimensional) เพื่อศึกษาข้อจำกัดและความเป็นไปได้ในการนำไปใช้จริง ทั้งในส่วนของความเป็นพหุมิติภายในแบบสอบหรือความเป็นพหุมิติระดับข้อสอบ เพื่อให้มีความสอดคล้องกับสถานการณ์การใช้ข้อสอบที่เป็นจริงและเป็นประโยชน์ต่อการวัดผลทางการศึกษาได้กว้างขวางมากขึ้น

เอกสารอ้างอิง

ศิริชัย กาญจนวาสี. (2555). ทฤษฎีการทดสอบแนวใหม่ (Modern Test Theories) (พิมพ์ครั้งที่ 4). กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

อรินทร์ น่วมถนอม. (2549). การเปรียบเทียบวิธีโพลี-ซิปเทสท์ วิธีการถดถอยโลจิสติกแบบจัดอันดับ และวิธีการถดถอยโลจิสติกแบบจัดอันดับหลายมิติ ในการตรวจสอบการทำหน้าที่เบี่ยงเบนของข้อสอบที่วัดความสามารถหลายมิติและให้คะแนนหลายค่า. วิทยานิพนธ์ปริญญาการศึกษาดุษฎีบัณฑิต, สาขาวิชาการทดสอบและวัดผลการศึกษา, บัณฑิตวิทยาลัย, มหาวิทยาลัยศรีนครินทรวิโรฒ.

อรินทร์ น่วมถนอม. (2549). การเปรียบเทียบวิธีโพลี-ซิปเทสท์ วิธีการถดถอยโลจิสติกแบบจัดอันดับ และวิธีการถดถอยโลจิสติกแบบจัดอันดับหลายมิติ ในการตรวจสอบการทำหน้าที่เบี่ยงเบนของข้อสอบที่วัดความสามารถหลายมิติและให้คะแนนหลายค่า. วารสารวิจัยทางการศึกษา คณะศึกษาศาสตร์ มศว. ปีที่ 1 (1), 136-145.

อาวีพร ปานทอง (2558) การเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบให้คะแนนหลายค่าโดยวิธีทดสอบอัตราส่วนความควรจะเป็น วิธีเบส์เซียนและวิธีโพลี-ซิปเทสท์. วิทยานิพนธ์ปริญญาดุษฎีบัณฑิต, สาขาวิชาวิจัย วัดผลและสถิติการศึกษา, คณะศึกษาศาสตร์, มหาวิทยาลัยบูรพา.

อุทัยวรรณ สายพัฒนะ. (2547). การเปรียบเทียบประสิทธิภาพของผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบสอบที่มีการให้คะแนนแบบหลายค่า ระหว่างวิธี GMH และวิธี Polytomous SIBTEST, วิทยานิพนธ์ปริญญาการศึกษาดุษฎีบัณฑิต, สาขาวิชาการทดสอบและวัดผลการศึกษา, บัณฑิตวิทยาลัย, มหาวิทยาลัยศรีนครินทรวิโรฒ.

Atalay Kabasakal, K., Arsan, N., Gök, B., & Kelecioglu, H. (2014). Comparing Performances (Type I Error and Power) of IRT Likelihood Ratio SIBTEST and Mantel-Haenszel Methods in the Determination of Differential Item Functioning. Educational Sciences: Theory and Practice, 14(6), 2186-2193.

Atar, B., & Kamata, A. (2011). Comparison of IRT likelihood ratio test and logistic regression DIF detection procedures. Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 41(41).

Chang, H. H., Mazzeo, J., & Roussos, L. (1995). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. ETS Research Report Series, 1995(1).

Chang, Y. W., Huang, W. K., & Tsai, R. C. (2015). DIF detection using multiple–group categorical CFA with minimum free baseline approach. Journal of Educational Measurement, 52(2), 181-199.

Clauser, R. E.; & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items.Educational Measurement: Issues and Practice. 17(1): 31-44.

Cohen, A.S., Kim, S., & Baker, F. B. (1993). Detection of Differential Item Functioning in the Graded Response Model. Applied Psychological Measurement. 17(4): 335-350.

Embretson, S. E., & Reise, S. P. (2013). Item response theory. Psychology Press.

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. Structural Equation Modeling, 18(2), 212-228.

Lopez Rivas, G. E., Stark, S., & Chernyshenko, O. S. (2009). The effects of referent item parameters on differential item functioning detection using the free baseline likelihood ratio test. Applied Psychological Measurement, 33(4), 251-265.

Meade, A. W., & Lautenschlager, G. J. (2004). A Comparison of item response theoryand confirmatory factor analytic methodologies for establishing measurement equivalence/ invariance. Organizational Research Methods, 7(4), 361-388.

Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: the alignment method. Frontiers in psychology, 5.

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

Nering, M. L., & Ostini, R. (Eds.). (2011). Handbook of polytomous item response theory models. Taylor & Francis.

Ostini, R., & Nering, M. L. (2006). Polytomous item response theory models(No. 144). Sage.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. Applied Psychological Measurement, 19(1), 23-37.

Samejima, F. (1969). Estimation of a latent ability using a response pattern of gradescores. Pschometric Monographs, 17.

Samejima, F. (1996). Estimation of a latent ability using a response pattern of gradescores. Pschometric Monograph Supplement, 34, 100-114.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/ DIF. Psychometrika, 58(2), 159-194.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.),Test validity (pp. 67-113). Hillsdale, NJ: Erlbaum.