วารสารจันทรเกษมสาร ปีที่ 23 ฉบับที่ 45 กรกฎาคม – ธันวาคม 2560
Journal of Chandrakasemsarn Vol. 23 No. 45 July – December, 2017

ประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามของแบบทดสอบพหุมิติให้คะแนนหลายค่าด้วยวิธีโพลี–ซิปเทสท์ วิธีทดสอบวอลด์ และวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ

Performance of Differential Item Function Detection for Multidimensional Polytomous Item Using Poly–Sibtest, Wald Test and Multi Group Confirmatory Factor Analysis

ณัฐพร ภักดี[1] ไพรัตน์ วงษ์นาม[2] และสุรีพร อนุศาสนนันท์[3]

Nuttaporn Phakdee[1] Pairatana Wongnam[2] and Sureeporn Anusananun[3]

[1]ดุษฎีบัณฑิต สาขาวิจัย วัดผล และสถิติการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา

[2]ที่ปรึกษาหลัก สาขาวิจัย วัดผล และสถิติการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา

[3]ที่ปรึกษาร่วม สาขาวิจัย วัดผล และสถิติการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาประสิทธิภาพของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามของแบบทดสอบพหุมิติให้คะแนนหลายค่าด้วยวิธีโพลี–ซิปเทสท์ วิธีทดสอบวอลด์ และวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ ในการศึกษาครั้งนี้ใช้ข้อมูลจำลอง โดยจำลองภายใต้โมเดลเกรดเรสพอนสองมิติ ซึ่งแต่ละข้อจะมีรายการตอบ 5 รายการ ให้คะแนนแต่ละรายการเป็น 1, 2, 3, 4 หรือ 5 คะแนน ข้อมูลดังกล่าวจำลองภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ ความยาวของแบบทดสอบต่างกัน 2 ขนาด ขนาดของการทำหน้าที่ต่างกัน 3 ระดับ สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 2 ขนาด และขนาดของกลุ่มตัวอย่างต่างกัน 5 รูปแบบ รวมเงื่อนไขที่แต่ละวิธีต้องทำการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามทั้งหมด 60 เงื่อนไข (2X3X2X5) ในแต่ละเงื่อนไขกระทำซ้ำ 100 รอบ

ผลการวิจัยพบว่า ภาพรวมสำหรับความยาวของแบบทดสอบจำนวน 20 ข้อ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามดีกว่าวิธีทดสอบวอลด์ แต่เมื่อความยาวของแบบทดสอบมากขึ้นวิธีทดสอบวอลด์มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามดีกว่าวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ โดยทั้งสองวิธีมีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามดีกว่าวิธีโพลี–ซิปเทสท์ ในทุกเงื่อนไข

คำสำคัญ: การทำหน้าที่ต่างกันของข้อคำถาม วิธีการวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ วิธีทดสอบวอลด์

Abstract

The purpose of this research was to compare performance of differential item functioning (DIF) for multidimensional polytomous items of Poly–SIBTEST, Wald test and Multi group confirmatory factor analysis procedures in the detecting. The type of all item were in five response categories scoring as 1, 2, 3, 4 or 5. This data was simulated under the multidimensional graded response model. A variety of four factors were two differing levels of test length, three differing levels of magnitude of DIF, two differing levels of proportion of DIF items, and five differing levels of sample sizes. A total of 60 conditions were studied. with 100 replications each.

The major findings are as follow: Multi group confirmatory factor analysis is more efficient than the Wald test for test length are 20 items when test length increased Wald test more efficient than Multi group confirmatory factor analysis and Multi group confirmatory factor analysis and Wald test is more efficient than Poly–SIBTEST procedure on detecting of differential item functioning (DIF) for multidimensional polytomous items under all condition.

Keywords: Differential item function, Multi group confirmatory factor analysis, Wald test

บทนำ

แบบทดสอบเป็นเครื่องมือที่สำคัญสำหรับการวัดผลทางการศึกษาและจิตวิทยา การวัดจะถูกต้องแม่นยำขึ้นอยู่กับแบบทดสอบที่ดีมีคุณภาพ ความมีคุณภาพของแบบทดสอบประกอบไปด้วยคุณภาพของแบบทดสอบรายข้อ ได้แก่ ความยากง่าย (Difficulty index) และอำนาจจำแนก (Discriminant index) และคุณภาพแบบทดสอบทั้งฉบับ ได้แก่ ความเที่ยง (Reliability) และความตรง (Validity) ความตรงนับว่าเป็นหัวใจที่สำคัญยิ่งของแบบทดสอบที่แสดงถึงความสามารถในการวัดได้ถูกต้องแม่นยำ คะแนนที่วัดได้จากแบบทดสอบสามารถวัดค่าได้ใกล้เคียงกับค่าคุณลักษณะภายในที่แท้จริงมากเพียงใด ก็ถือว่า แบบทดสอบความตรงมากขึ้นเพียงนั้น (Ayala, 2008; Demars, 2010) นอกจากนี้ยังมีคุณสมบัติอีกประการหนึ่งของแบบทดสอบ ที่นักวิจัยให้ความสำคัญ คือ ความยุติธรรมของแบบทดสอบ (Test fairness) เนื่องจากในปัจจุบันมีการนำแบบทดสอบมาใช้ในการวัดความสามารถของบุคคล เพื่อคัดเลือกเข้าศึกษาต่อ บรรจุงาน เลื่อนตำแหน่ง หรือออกใบอนุญาต ดังนั้น แบบทดสอบที่นำมาใช้ควรเป็นแบบทดสอบที่มีความคลาดเคลื่อนน้อยที่สุด จำแนกความสามารถของผู้สอบได้ และไม่เอนเอียงเข้าข้างผู้สอบกลุ่มใดกลุ่มหนึ่งตามคุณลักษณะบางอย่างที่ต่างกัน เช่น เพศ เชื้อชาติ วัฒนธรรม หรือภูมิหลังอื่น ของผู้สอบ เพราะข้อสอบที่มีความลำเอียงจะทำให้คะแนนการทดสอบแตกต่างกันระหว่างกลุ่ม ทั้งๆ ที่ทั้งสองกลุ่มมีความสามารถเท่ากัน การตรวจสอบความลำเอียงของข้อสอบ (Item bias) ถือเป็นส่วนหนึ่งของการพัฒนาแบบทดสอบ นักพัฒนาจึงจำเป็นต้องศึกษาเพื่อเป็นหลักฐานแสดงความยุติธรรม โดยทำการคัดเลือกข้อที่มีความลำเอียงออกจากแบบทดสอบ (สุชาติ สิริมีนนันท์, 2555)

ปัจจุบันนักวิจัยใช้คำว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม (Differential item functioning) หรือเรียกว่า “DIF” แทนคำว่า “ความลำเอียงของข้อสอบ” หลักการตรวจสอบเป็นการเปรียบเทียบผลการตอบระหว่างผู้สอบกลุ่มย่อยสองกลุ่มที่มีความสามารถระดับเดียวกัน โดยผู้สอบกลุ่มหนึ่งเป็นตัวแทนกลุ่มหลักในประชากร เรียกว่า “กลุ่มอ้างอิง” (Reference group: R) ส่วนอีกกลุ่มหนึ่งเป็นตัวแทนกลุ่มรองในประชากร เรียกว่า “กลุ่มสนใจ” (Focal group: F) ซึ่งเป็นกลุ่มเป้าหมายที่ต้องการศึกษา ข้อคำถามที่ใช้ในการตรวจสอบเรียกว่า “ข้อคำถามศึกษา” (Studies item) เมื่อข้อคำถามเกิดการทำหน้าที่ต่างกัน กลุ่มอ้างอิงจะได้เปรียบในการตอบ ส่วนกลุ่มสนใจคาดว่าจะเสียเปรียบในการตอบ เกณฑ์ที่ใช้ในการเปรียบเทียบผลการตอบระหว่างกลุ่มอ้างอิงและกลุ่มสนใจจะใช้เกณฑ์การจับคู่ (Matching) ตามความสามารถ (ศิริชัย กาญจนวาสี, 2550ก) ซึ่งถ้าตรวจสอบพบว่าข้อคำถามใดทำหน้าที่ต่างกัน ข้อคำถามข้อนั้นจะถูกพิจารณาเพื่อทำการปรับปรุง หรือตัดออกจากแบบทดสอบ

นักวิจัยทางการศึกษาและจิตวิทยาให้ความสนใจศึกษาเพื่อหาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม โดยระยะแรกสนใจแบบทดสอบที่วัดความสามารถแบบเอกมิติให้คะแนนสองค่า (Unidimensional dichotomous) วิเคราะห์ภายใต้ทฤษฎีการทดสอบแบบดั้งเดิม (Classical test theory methods) และใช้คะแนนสังเกตเป็นเกณฑ์การจับคู่ เช่น วิธีแมนเทลแฮนเซล วิธีทำให้เป็นมาตรฐาน และวิธีการถดถอยโลจิสติก เป็นต้น แต่วิธีเหล่านี้มีจุดอ่อนที่การประมาณค่าพารามิเตอร์ของข้อคำถามและแบบทดสอบ และค่าความคลาดเคลื่อนมีค่าเปลี่ยนไปตามกลุ่มผู้สอบ ต่อมานักวิจัยจึงพัฒนาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามด้วยทฤษฎีการตอบสนองข้อคำถาม (Item response theory) ใช้ค่าประมาณระดับความสามารถของผู้สอบซึ่งเป็นตัวแปรคุณลักษณะแฝงเป็นเกณฑ์ในการจับคู่ ซึ่งเป็นวิธีที่ช่วยแก้ไขจุดอ่อนดังกล่าว เช่น วิธีไคสแควร์ของลอร์ด และวิธีการทดสอบอัตราส่วนความน่าจะเป็น เป็นต้น ต่อมามีการพัฒนาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามจากให้คะแนนสองค่าเป็นหลายค่า เช่น วิธีวิเคราะห์ฟังก์ชันการจำแนกโลจิสติก การถดถอยโลจิสติกแบบจัดอันดับ เป็นต้น แต่วิธีการส่วนใหญ่มักอยู่บนโมเดลการวัดคุณลักษะแบบเอกมิติ แต่สำหรับแบบทดสอบทางด้านการศึกษาและจิตวิทยา มักสร้างแบบทดสอบบนโมเดลการวัดแบบหลายมิติ (Multidimensional model) และให้คะแนนหลายค่า เช่น วิธีโพลี–ซิปเทสท์ (Poly–Sibtest) วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ (Multi group confirmatory factor analysis: MG–CFA) และ วิธีการทดสอบแบบวอลด์ (Wald test) จะเห็นว่าวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามได้มีการคิดค้นและพัฒนามาอย่างต่อเนื่องเป็นลำดับ ซึ่งมีประเด็นที่น่าสนใจ 3 ประเด็น กล่าวคือ

ประเด็นแรก นักวัดผลได้ปรับเปลี่ยนวิธีการวัดและประเมินผลจากให้คะแนนสองค่าเป็นให้คะแนนหลายค่า เพราะเชื่อว่าการให้คะแนนข้อคำถามแบบหลายค่าให้สารสนเทศได้มากกว่าและเป็นการประเมินตามสภาพจริง อีกทั้งแบบทดสอบทางด้านการศึกษาและจิตวิทยา มักเป็นแบบทดสอบที่มีลักษณะการวัดความสามารถหลายมิติและให้คะแนนหลายค่า แต่ยังคงถูกวิเคราะห์ภายใต้ความเป็นเอกมิติ จึงทำให้เกิดความลำเอียงในการประมาณค่าพารามิเตอร์ของข้อคำถาม และการประมาณค่าความสามารถของผู้สอบได้ (Wilson & Hoskens, 2005) ดังนั้นการวิเคราะห์ข้อคำถามของแบบทดสอบพหุมิติให้คะแนนหลายค่า จึงเป็นประเด็นที่สนใจในการวิจัยครั้งนี้

ประเด็นสอง ความตรงของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามขึ้นอยู่กับวิธีการตรวจสอบที่ใช้คุณลักษณะแฝงเป็นเกณฑ์ ถ้าแบบทดสอบสร้างบนพื้นฐานการวัดแบบพหุมิติตามแบบโมเดลเกรดเรสพอนพหุมิติ (MGRM) และให้คะแนนหลายค่า พบว่ามีหลายวิธี แต่สำหรับงานวิจัยนี้ ผู้วิจัยสนใจวิธีโพลี–ซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบแบบวอลด์

ประเด็นที่สาม การเลือกใช้วิธีการตรจสอบการทำหน้าที่ต่างกันของข้อคำถาม ผลการตรวจสอบจะมีประสิทธิภาพ ควรคำนึงถึงปัจจัยร่วมที่นำมาใช้ในการตรวจสอบ เช่น ขนาดของการทำหน้าที่ต่างกันของข้อคำถาม สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน ความยาวของแบบทดสอบ และขนาดตัวอย่าง เป็นต้น เพื่อเป็นสารสนเทศส่วนหนึ่งในการตัดสินใจเลือกใช้วิธีการตรวจสอบที่มีประสิทธิภาพของแต่ละเงื่อนไข และทั้งสามวิธีตรวจสอบได้ดีในเงื่อนไขใด

จากประเด็นดังกล่าว ผู้วิจัยจึงได้นำวิธีโพลี-ซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และ วิธีการทดสอบแบบวอลด์ มาใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแบบทดสอบพหุมิติที่ให้คะแนนหลายค่า เมื่อตรวจสอบภายใต้ปัจจัยขนาดของการทำหน้าที่ต่างกันของข้อคำถาม ความยาวของแบบทดสอบ สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง ซึ่งผลที่ได้จากการศึกษาในครั้งนี้จะทำให้ทราบว่า วิธีใดมีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามในแต่ละเงื่อนไขเป็นอย่างไร และในแต่ละวิธีตรวจสอบได้ดีในเงื่อนไขใด อีกทั้งยังเป็นแนวทางในการเลือกใช้วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามของแบบทดสอบพหุมิติให้คะแนนหลายค่า และเป็นทางเลือกสำหรับการศึกษาการทำหน้าที่ต่างกันของข้อคำถามในเครื่องมือการวัดอื่น นอกเหนือจากแบบทดสอบอีกด้วย

วัตถุประสงค์ของการวิจัย

เพื่อทดสอบประสิทธิภาพของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามของแบบทดสอบพหุมิติให้คะแนนหลายค่า ด้วยวิธีโพลี-ซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีทดสอบวอลด์ ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ ความยาวของแบบทดสอบ ขนาดของการทำหน้าที่ต่างกันของข้อคำถาม สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และขนาดตัวอย่าง

ขอบเขตของการวิจัย

การวิจัยครั้งนี้เป็นการศึกษาประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม ข้อมูลทั้งหมดที่ใช้ในการตรวจสอบเป็นข้อมูลจำลอง ภายใต้โมเดลเกรดเรสพอนพหุมิติ (Multidimensional graded response model) สำหรับสร้างแบบทดสอบพหุมิติ (Multidimensional test) ที่มีลักษณะการวัดแบบสองมิติ (Two-dimensional) และเป็นแบบทดสอบชนิดความสามารถหลายมิติระหว่างข้อคำถาม (Multidimensional between-item test) โดยข้อคำถามแต่ละข้อวัดความสามารถมิติใดมิติหนึ่ง ข้อคำถามแต่ละข้อมีรายการคำตอบแบบ 5 รายการ โดยให้คะแนนแต่ละรายการเป็น 1, 2, 3, 4 หรือ 5 จำลองผลการตอบภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย ได้แก่ ความยาวของแบบทดสอบ 2 ขนาด ขนาดของการทำหน้าที่ต่างกันของคำถาม 3 ระดับ สัดส่วน

ข้อคำถามที่ทำหน้าที่ต่างกัน 2 ขนาด และ ขนาดตัวอย่างที่แตกต่างกัน 5 รูปแบบ รวมข้อมูลทั้งหมดที่ตรวจสอบ จำนวน 60 เงื่อนไข ในแต่ละเงื่อนไขจำลองซ้ำ 100 ครั้ง

ตัวแปรที่ใช้ในการวิจัย

1. ตัวแปรต้น มี 5 ตัวแปร คือ

1.1 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม มี 3 วิธี ซึ่งเป็นวิธีสำหรับตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามของแบบทดสอบพหุมิติและให้คะแนนหลายค่า ได้แก่ โพลี–ซิปเทสท์ (Chang, Mazzeo, & Roussos, 1996) วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ (Reise, Widaman, & Pugh, 1993; Vandenberg & Lance, 2000) และวิธีทดสอบวอลด์ (Cai, 2008)

1.2 ความยาวของแบบทดสอบ (Test length) สำหรับการศึกษาที่ผ่านมา พบว่าแบบทดสอบให้คะแนนหลายค่าด้วยการจำลองส่วนใหญ่จะศึกษาในช่วง 10 ถึง 40 ข้อ (Wang & Su, 2004; Williams & Beretvas, 2006) และ แบบทดสอบ จำนวน 20 ข้อ และ 40 ข้อ มีความแกร่งสำหรับการทดสอบการทำหน้าที่ต่างกัน (Wu & Lei, 2009) ดังนั้น ผู้วิจัยจึงใช้ความยาวของแบบทดสอบ คือ 20 ข้อ และ 40 ข้อ

1.3 ขนาดของการทำหน้าที่ต่างกัน (Magnitude DIF) เป็นค่าความแตกต่างของข้อคำถามที่ทำหน้าที่ต่างกัน ชนิดไม่เป็นรูปแบบเดียวกัน (Non–uniform) ซึ่งมีค่าอำนาจจำแนกของกลุ่มอ้างอิงสูงกว่ากลุ่มสนใจ แต่มีค่าเทสโฮล (ความยากรายคำตอบ) น้อยกว่ากลุ่มสนใจ จากการศึกษาของ Meade, Lautenschlager and Johnson (2006) พบว่า ขนาดของการทำหน้าที่ต่างกันของข้อคำถามขนาดใหญ่ ($\geq$ .40) จะมีอำนาจในการทดสอบสูง ผู้วิจัยจึงกำหนดขนาดของการทำหน้าที่ต่างกัน 3 ขนาด คือ .40, .70 และ 1.0

1.4 สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน (Proportion DIF) วัดจากสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกันต่อจำนวนข้อคำถามทั้งหมดในแบบทดสอบ ข้อคำถามที่ทำหน้าที่ต่างกันนี้มีผลให้ผู้สอบที่มาจากกลุ่มที่แตกต่างกัน มีความสามารถที่ต้องการวัดเท่ากันจะมีความน่าจะเป็นในการตอบข้อคำถามได้ถูกต้องไม่เท่ากัน การศึกษาครั้งนี้ผู้วิจัยกำหนดสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน 2 ขนาด คือ ร้อยละ 10 และร้อยละ 20 เนื่องจาก การวิจัยที่ผ่านมาศึกษาทำหน้าที่ต่างกันของข้อคำถามที่มีสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกันระหว่างร้อยละ 10 ถึงร้อยละ 20 เพราะถ้าเกินกว่านี้จะทำให้เกิดความผิดพลาดเชิงบวก (False positives) และ ความผิดพลาดเชิงลบ (False negatives) ขนาดใหญ่ (Gonzales–Roma, Hernandez, & Gomez–Benito, 2006; Stark, Chernyshenko, & Drasgow,2006; Wu & Lei, 2009) ถ้าความยาวแบบทดสอบเป็น 20 ข้อ จะมีข้อคำถามที่ทำหน้าที่ต่างกันจำนวน 2 ข้อ (ร้อยละ 10) และ 4 ข้อ (ร้อยละ 20) สำหรับกรณี

ความยาวแบบทดสอบเป็น 40 ข้อ จะมีข้อที่ทำหน้าที่ต่างกันจำนวน 4 ข้อ (ร้อยละ 10) และ 8 ข้อ (ร้อยละ 20)

1.5. ขนาดตัวอย่าง (Sample size) การศึกษาที่ผ่านมามีการกำหนดขนาดตัวอย่างสำหรับการศึกษาทั้งแบบที่มีสัดส่วนเท่ากัน (Meade & Lautenschlager, 2004, Gonzalez–Roma, Hernandez, & Gomez–Benito, 2006) และสัดส่วนที่ต่างกัน (Finch & French, 2007) โดยการศึกษามีช่วงของการกำหนดตัวอย่างระหว่าง 250 ถึง 1000 คน ดังนั้น ผู้วิจัยจึงได้กำหนดตัวอย่าง สำหรับการศึกษาในครั้งนี้ โดยใช้สัดส่วนของกลุ่มสนใจต่อกลุ่มอ้างอิง โดยใช้อัตราส่วน 1:1 และ 1:2 จำนวน 5 รูปแบบ คือ 250:250, 500:500, 1000:1000, 250:500 และ 500:1000 คน (Meade, Lautenschlager, & Johnson, 2006; Stark, Chernyshenko, & Drasgow, 2006)

2. ตัวแปรตาม ได้แก่ ประสิทธิภาพของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม พิจารณาจากอัตราความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบที่ผ่านการทดสอบตามเกณฑ์ที่กำหนด มี 2 ตัวแปร คือ

2.1 ความคลาดเคลื่อนประเภทที่ 1 (Type I error) คำนวณจากสัดส่วนจำนวนข้อคำถามที่ตรวจสอบได้ผิดพลาดว่าทำหน้าที่ต่างกันทั้งที่จริงข้อคำถามข้อนั้นไม่ทำหน้าที่ต่างกันต่อจำนวนข้อคำถามที่ไม่ทำหน้าที่ต่างกันทั้งหมดในแบบทดสอบ โดยมีเกณฑ์ในการทดสอบ คือ ค่าความคลาดเคลื่อนประเภทที่ 1 ต่ำกว่าหรือเท่ากับ .05 ถือว่าควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดี (Atar & Kamata, 2011)

2.2 อำนาจการทดสอบ (Power of test) คำนวณจากสัดส่วนจำนวนข้อคำถามที่ตรวจสอบได้ถูกต้องว่าทำหน้าที่ต่างกันต่อจำนวนข้อคำถามที่ทำหน้าที่ต่างกันทั้งหมดในแบบทดสอบ โดยมีเกณฑ์ในการทดสอบ คือ อำนาจการทดสอบต้องมีค่าตั้งแต่ .80 ขึ้นไป จึงถือว่ามีอำนาจการทดสอบเพียงพอ (Sufficient power) (Atar & Kamata, 2011)

สมมติฐานของการวิจัย

1. การตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามของแบบทดสอบพหุมิติที่ให้คะแนนหลายค่าของทั้ง 3 วิธี มีความคลาดเคลื่อนประเภทที่ 1 ไม่เกิน .05 ในแต่ละเงื่อนไขภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ ความยาวของแบบทดสอบ ขนาดของการทำหน้าที่ต่างกัน สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และ ขนาดตัวอย่าง

2. การตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามของแบบทดสอบพหุมิติที่ให้คะแนนหลายค่าของทั้ง 3 วิธี มีอำนาจการทดสอบมากกว่า .80 ในแต่ละเงื่อนไขภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ ความยาวของแบบทดสอบ ขนาดของการทำหน้าที่ต่างกัน สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน และ ขนาดตัวอย่าง

วิธีดำเนินการวิจัย

การวิจัยครั้งนี้มีจุดมุ่งหมายเพื่อศึกษาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามของแบบทดสอบพหุมิติให้คะแนนหลายค่า ได้แก่ วิธีโพลี–ซิปเทสท์ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์ ซึ่งเป็นงานวิจัยเชิงทดลอง ทดสอบภายใต้ข้อมูลจำลอง

เครื่องมือการวิจัย

โปรแกรมสำหรับการจำลองข้อมูล และการวิเคราะห์ข้อมูล ได้แก่ โปรแกรม R (3.3.1)

การรวบรวมข้อมูล

1. การจำลองความสามารถผู้สอบ กำหนดจำนวนผู้สอบของกลุ่มสนใจ:กลุ่มอ้างอิง (NF:NR) ตามสัดส่วนของขนาดตัวอย่าง (250:250, 500:500, 1000 :1000, 250:500 และ 500:1000 คน) โดยข้อมูลความสามารถของผู้สอบมีการแจกแจงดังนี้ $\theta_j \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0.5\\0.5 & 1\end{bmatrix}\right)$

2. กำหนดข้อคำถามตามจำนวนข้อคำถามของแบบทดสอบ (20 ข้อ และ 40 ข้อ) แต่ละข้อมีจำนวนรายการคำตอบ 5 รายการ และสัดส่วนจำนวนข้อคำถามในแต่ละมิติเป็น 1:1 สำหรับแบบทดสอบ 20 ข้อ และสัดส่วนของการทำหน้าที่ต่างกัน ร้อยละ 10 และ ร้อยละ 20 จะมีข้อคำถามที่ทำหน้าที่ต่างกัน คือ ข้อ 10, 20 และ 9, 10, 19, 20 ตามลำดับ ส่วนกรณีแบบทดสอบจำนวน 40 ข้อ จะมีข้อคำถามที่ทำหน้าที่ต่างกัน คือ ข้อ 9, 10, 19, 20 และ 7, 8, 9, 10, 17, 18, 19, 20 ตามลำดับ โดยในแต่ละข้อที่กล่าวมาจะมีอำนาจจำแนกในข้อคำถามที่ทำหน้าที่ต่างกันของกลุ่มอ้างอิง (R) มีค่ามากกว่ากลุ่มสนใจ (F) เป็น $(a_{iR} = a_{iF} + Mag\ DIF)$ และเทสโฮลของกลุ่มอ้างอิง (R) มีค่าน้อยกว่ากลุ่มสนใจ (F) $\left(\tau_{ikR} = \tau_{ikF} - Mag\ DIF\right)$ ในทุกรายการคำตอบ

3. การจำลองพารามิเตอร์ข้อคำถาม นำข้อมูลที่ได้จากการจำลองในข้อ 1 และ 2 ที่มีการระบุขนาดของการทำหน้าที่ต่างกันในกลุ่มอ้างอิง และกลุ่มสนใจ มาจำลองพารามิเตอร์ของข้อคำถาม ซึ่งความน่าจะเป็น $(P(u_{vij}))$ ของการเลือกตอบข้อคำถามของผู้สอบคนที่ j มีความสามารถ θ ในมิติ ν $(\nu = 1,2)$ โดยมีค่าคะแนนในแต่ละตัวเลือก (k) ในแต่ละข้อคำถามที่ i ภายใต้โมเดล MGRM ดังสมการ

$$P(u_{vij} = k|\boldsymbol{\theta_{vj}}) = \frac{1}{\sqrt{2\pi}} \int_{\boldsymbol{a'_{vi}\theta_{vj}}+\tau_{vi,k+1}}^{\boldsymbol{a'_{vi}\theta_{vj}}+\tau_{vik}} e^{\frac{-t^2}{2}} dt$$

θ_{vj} คือ พารามิเตอร์ความสามารถของผู้สอบคนที่ j ในมิติที่ ν , a_{vi} คือ พารามิเตอร์อำนาจจำแนกของข้อคำถามข้อที่ i ในมิติที่ ν โดย $0 \le a_{vi} \sim N(0,1) \le 1$ และ τ_{ik} คือ พารามิเตอร์ข้อคำถาม เทสโฮล ของข้อคำถามข้อที่ i ในรายการที่ k โดย $\tau_{ik} \sim N(0,1)$

4. การสร้างข้อมูลผลการตอบข้อคำถาม ในขั้นตอนนี้เป็นการนำเซตข้อมูลที่เกิดจากการจำลองในขั้นตอนที่ 1, 2 และ 3 มาจัดทำข้อมูลรายการคำตอบของกลุ่มตัวอย่างที่จะทำการศึกษาในแต่ละเงื่อนไข จำนวน 60 เงื่อนไข กระทำซ้ำ 100 ชุด ซึ่งจะได้ข้อมูลเป็นไปตามเงื่อนไขทั้งสิ้น 6,000 ชุด

การวิเคราะห์ข้อมูล

1. วิธีโพลี–ซิปเทสท์ (Poly–Sibtest)

Chang, Mazzeo and Roussos (1996) ได้พัฒนาวิธีนี้เพื่อใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามที่ให้คะแนนหลายค่า ดังสมการ

$$B = \frac{\sum_{k=0}^{n_H} P_k \, d_k}{\sqrt{\sum_{k=0}^{n_H} P_k^2 \left[\frac{\hat{\sigma}^2(Y|k,R)}{N_{Rk}} + \frac{\hat{\sigma}^2(Y|k,F)}{N_{Fk}}\right]}}, k = 0, \ldots, n_H$$

เมื่อ P_k แทนสัดส่วนของผู้สอบของกลุ่มอ้างอิงและกลุ่มสนใจ ซึ่งตอบแบบทดสอบเดียวกันและได้คะแนนเท่ากัน $x = k$ และ d_k แทนผลต่างของค่าเฉลี่ยคะแนนสอบของกลุ่มอ้างอิงและกลุ่มสนใจ ภายใต้ข้อคำถามศึกษา ณ คะแนนรวมในแบบทดสอบ $x = k$ (ณ ที่คะแนนรวมเท่ากัน อนุมานว่ามีความสามารถเท่ากัน) ถ้าข้อคำถามศึกษาใดไม่มีคะแนนสังเกตได้แสดงว่า $d_k \approx 0$ ถ้าผลการทดสอบพบว่า $|B| > Z_{1-\frac{\alpha}{2}}$ อย่างมีนัยสำคัญที่ระดับ .05 แสดงว่าข้อคำถามทำหน้าที่ต่างกันโดยเข้าข้างผู้สอบกลุ่มใดกลุ่มหนึ่ง

2. วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ (MG–CFA)

วิธีการวิเคราะห์นี้เป็นการนำความรู้จากการวิเคราะห์สมการเชิงโครงสร้างมาใช้ในการวิเคราะห์ความแตกต่างกันของข้อคำถามโดยยึดหลักแนวคิดความเท่าเทียม/ไม่แปรเปลี่ยน (Measurement Equivalence/Invariance) ของโมเดล (Reise, Widaman, & Pugh, 1993; Vandenberg & Lance, 2000) ดังสมการ

$$x^g = \Lambda^g \xi^g + \delta^g$$

$$\Sigma^g = \Lambda^g \Phi \Lambda^{g\prime} + \Theta_\delta^g$$

การทดสอบความไม่แปรเปลี่ยนจะมีสมมติฐานที่สำคัญ 5 ข้อ คือ

1. รูปแบบของโครงสร้าง CFA เหมือนกัน และมีรูปแบบเดียวกันระหว่างกลุ่ม

2. ชุดข้อคำถามของแบบทดสอบมีคุณลักษณะแฝง $(\xi^g = \xi^{g\prime})$ ในแต่ละกลุ่มมีค่าเท่ากัน

3. น้ำหนักองค์ประกอบ $(\lambda^g = \lambda^{g\prime})$ ที่มีค่าเท่ากันระหว่างกลุ่ม

4. เทสโฮล $(\tau^g = \tau^{g\prime})$ ของรายการคำตอบมีค่าเท่ากันระหว่างกลุ่ม $(\tau^g = \tau^{g\prime})$

5. ความแปรปรวน ($\Theta_{\delta}^{g} = \Theta_{\delta}^{g'}$) เท่ากันระหว่างกลุ่ม

6. ความแปรปรวนและความแปรปรวนร่วม ($\boldsymbol{\phi}^{g} = \boldsymbol{\phi}^{g'}$) ของตัวแปรแฝงเท่ากันระหว่างกลุ่ม สำหรับการศึกษาครั้งนี้ ผู้วิจัยทำการทดสอบความไม่แปรเปลี่ยนของข้อคำถามแต่ละข้อในแบบทดสอบโดยใช้การกำหนดให้สมมติฐานเพียงสองข้อ คือ 3. และ 4. ไม่เป็นจริง ส่วนข้ออื่น ๆ ให้เป็นไปตามสมมติฐานเดิม และถ้าพบว่าค่าความแตกต่างของไคสแควร์ที่ใช้ในการทดสอบของโมเดลเริ่มต้น (Baseline model) กับโมเดลทดสอบ (Hypothesis model) พบนัยสำคัญในข้อคำถามใด แสดงว่าข้อคำถามข้อนั้นทำหน้าที่ต่างกัน

3. วิธีทดสอบวอลด์ (Wald test)

Wald (Cai, 2008) พัฒนามาจาก Lord (1980) โดยปรับปรุงประสิทธิภาพในการทดสอบโดยใช้การประมาณค่าเมทริกซ์ความแปรปรวน–ความแปรปรวนร่วมด้วยวิธี Supplemented Expectation Maximization (SEM) โดยมีสถิติทดสอบ คือ

$$\chi^2 = \hat{\boldsymbol{\nu}}'\Sigma^{-1}\hat{\boldsymbol{\nu}}\ ,\ df = \text{จำนวนพารามิเตอร์ในการทดสอบ}$$

เมื่อ $\hat{\boldsymbol{\nu}} = \left[\alpha_{f1} - \alpha_{r1}, \beta_{f1} - \beta_{r1}, \beta_{f2} - \beta_{r2}, \beta_{f3} - \beta_{r3}, \beta_{f4} - \beta_{r4}\right]$ เป็นเวกเตอร์ผลต่างระหว่างพารามิเตอร์ของข้อคำถามทุกข้อ ระหว่างกลุ่มอ้างอิงและกลุ่มสนใจ และ Σ^{-1} เป็นเมทริกซ์ความแปรปรวน–ความแปรปรวนร่วมของความคลาดเคลื่อนของกลุ่มสนใจ (F) และกลุ่มอ้างอิง (R) $(\Sigma^{-1} = (\Sigma_F + \Sigma_R)^{-1})$

4. คำนวณอำนาจการทดสอบ และความคลาดเคลื่อนประเภทที่ 1 พร้อมทดสอบสมมติฐานเพื่อหาประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามตามเกณฑ์ที่กำหนดด้วยสถิติ Z

ผลการวิจัย

ผลการวิเคราะห์ประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม โดยพิจารณาจากการทดสอบการควบคุมความคลาดเคลื่อนประเภทที่ 1 (น้อยกว่า .05) และอำนาจการทดสอบ (มากกว่า .80) ตามเกณฑ์ที่กำหนด ณ ระดับนัยสำคัญ .05 ของวิธีโพลี–ซิปเทสท์ (P) วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ (M) และวิธีทดสอบวอลด์ (W) ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย ได้แก่ ความยาวของแบบทดสอบ (Test length) ขนาดของการทำหน้าที่ต่างกันของข้อคำถาม (MAG DIF) สัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน (PROP DIF) และขนาดตัวอย่าง แสดงดังตารางที่ 1

ผลการทดสอบ พบว่า สำหรับความยาวของแบบทดสอบเป็น 20 ข้อ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันได้ดีกว่า

วิธีทดสอบวอลด์ และเมื่อความยาวของแบบทดสอบเพิ่มขึ้นเป็น 40 ข้อ วิธีทดสอบวอลด์มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันได้ดีกว่าวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ ส่วนวิธีโพลี–ซิปเทสท์ ไม่ผ่านเกณฑ์ในการตรวจสอบประสิทธิภาพ

ตารางที่ 1 ผลการวิเคราะห์อำนาจการทดสอบ และการควบคุมความคลาดเคลื่อนประเภทที่ 1 ของ DIF 3 วิธี ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย

TEST LENGTH	MAG DIF	PROP DIF	SAMPSIZE (NF:NR)														
			250:250			500:500			1000:1000			250:500			500:1000		
			P	M	W	P	M	W	P	M	W	P	M	W	P	M	W
20 ข้อ	.40	10%	–	✓	✓	–	✓	✓	–	✓	✓	–	✓	✓	–	–	✓
		20%	–	✓	✓	–	✓	✓	–	–	–	–	✓	✓	–	–	–
	.70	10%	–	✓	✓	–	✓	✓	–	✓	–	–	✓	–	–	✓	–
		20%	–	✓	–	–	✓	–	–	–	–	–	✓	–	–	✓	–
	1.00	10%	–	✓	✓	–	✓	✓	–	✓	–	–	✓	✓	–	✓	–
		20%	–	–	–	–	✓		–	–	–	–	–	–	–	✓	–
40 ข้อ	.40	10%	–	–	✓	–	✓	✓	–	–	✓	–	✓	✓	–	✓	✓
		20%	–	✓	✓	–	–	✓	–	–	–	–	–	✓	–	–	–
	.70	10%	–	✓	✓	–		✓	–	–	✓	–	✓	–	–	✓	✓
		20%	–	–	✓	–	–	–	–	–	–	–	–	–	–	–	–
	1.00	10%	–	✓	✓	–	✓	✓	–	–	✓	–	✓	✓	–	–	✓
		20%	–	–	✓	–	–	–	–	–	–	–	–	–	–	–	–

หมายเหตุ เครื่องหมาย ✓ หมายถึง ผ่านการทดสอบประสิทธิภาพเป็นไปตามเกณฑ์ทั้ง P และ E

เมื่อพิจารณาปัจจัยขนาดของการทำหน้าที่ต่างกัน พบว่า วิธีการวิเคราะห์องค์ประกอบเชิงยืนยันมีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันดีกว่าวิธีทดสอบวอลด์ในทุกปัจจัยขนาดของการทำหน้าที่ต่างกัน ส่วนวิธีโพลี–ซิปเทสท์ ไม่ผ่านเกณฑ์การตรวจสอบประสิทธิภาพ ส่วนปัจจัยสัดส่วนของข้อคำถามที่ทำหน้าที่ต่างกัน พบว่า วิธีการวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันใกล้เคียงกัน และเมื่อพิจารณาปัจจัยขนาดตัวอย่าง พบว่า เมื่อขนาดของตัวอย่างเพิ่มขึ้นทั้งอัตราส่วนแบบ 1:1 และ 1:2 มีค่าเพิ่มขึ้น ประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันลดลงทั้งวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ และวิธีการทดสอบวอลด์

อภิปรายผลการวิจัย

1. ปัจจัยความยาวของแบบทดสอบ ผลการทดสอบพบว่า วิธีโพลี–ซิปเทสท์ไม่สามารถควบคุมประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามได้ตามเกณฑ์ที่กำหนด ซึ่งไม่สอดคล้องกับ สุชาติ สิริมีนนันท์ (2555) ที่พบว่า วิธีโพลี–ซิปเทสท์สามารถควบคุมประสิทธิภาพได้เมื่อความยาวของแบบทดสอบเพิ่มขึ้น ส่วนวิธีการวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ พบว่า เมื่อความยาวของแบบทดสอบเพิ่มมากขึ้น การควบคุมประสิทธิภาพของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามได้ตามเกณฑ์ที่กำหนดลดลง ซึ่งไม่สอดคล้องกับการศึกษาของ French and Finch (2008) ที่พบว่า เมื่อความยาวของแบบทดสอบเพิ่มขึ้นการควบคุมประสิทธิภาพของวิธีการวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุมีประสิทธิภาพมากขึ้น และเมื่อพิจารณาวิธีการทดสอบวอลด์ พบว่า สามารถควบคุมประสิทธิภาพของการตรวจสอบการทำหน้าที่ต่างกันได้มากขึ้น เมื่อความยาวของแบบทดสอบมากขึ้น ผลที่ได้ไม่สอดคล้องกับ Cao, Tay and Liu (2017) ที่พบว่า ความยาวของแบบทดสอบไม่มีผลต่อประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกัน ความไม่สอดคล้องของผลการทดสอบครั้งนี้ของทั้ง 3 วิธี อาจสืบเนื่องมาจากการพิจารณาเฉพาะความยาวของแบบทดสอบเพียงอย่างเดียว อาจจะไม่ครอบคลุมถึงเงื่อนไขที่มีปฏิสัมพันธ์ของปัจจัย อันได้แก่ ขนาดของการทำหน้าที่ต่างกันของข้อคำถาม สัดส่วนของการทำหน้าที่ต่างกันของข้อคำถาม และขนาดกลุ่มตัวอย่าง ซึ่งนำผลของปฏิสัมพันธ์มาใช้เป็นตัวแปรที่ควรทำการศึกษาต่อไป

2. ปัจจัยขนาดของการทำหน้าที่ต่างกันของข้อคำถาม การควบคุมประสิทธิภาพของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามของวิธีโพลี–ซิปเทสท์ พบว่า ไม่สามารถควบคุมประสิทธิภาพของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามได้ตามเกณฑ์ ซึ่งสอดคล้องกับ Kilmen (2016) ที่พบว่า เมื่อขนาดของการทำหน้าที่ต่างกันของข้อคำถามมากขึ้น ($\geq$ .40) วิธีโพลี–ซิปเทสท์จะไม่สามารถควบคุมการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามตามเกณฑ์ที่กำหนดได้ สำหรับวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ พบว่า เมื่อขนาดของการทำหน้าที่ต่างกันของข้อคำถามมีขนาดเพิ่มขึ้น ประสิทธิภาพของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามลดลง ซึ่งไม่สอดคล้องกับ Kim and Yoon (2011) และ Stark, Chernyshenko, and Drasgow (2006) ที่พบว่า ขนาดของการทำหน้าที่ต่างกันเพิ่มขึ้นประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันเพิ่มขึ้น ทั้งนี้เป็นเพราะการศึกษาของนักวิจัยทั้งสอง ศึกษาขนาดของการทำหน้าที่ต่างกันโดยมีการกำหนดให้ค่าอำนาจจำแนกและค่าเทสโฮลของข้อคำถามที่ต่างกันในกลุ่มสนใจและกลุ่มอ้างอิงแตกต่างกันไม่เท่ากัน คือ อำนาจจำแนกเป็น .60 .70 .80 และ .90 ส่วนค่าเทสโฮล คือ .10 .15 และ 2.50 แต่สำหรับผู้วิจัยกำหนดให้ค่าทั้งสองในกลุ่มสนใจและกลุ่มอ้างอิงแตกต่างกันเท่ากัน คือเท่ากับ .40 .70 และ 1.00 ส่วนวิธีการทดสอบวอลด์ พบว่า

มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามเพิ่มขึ้นเมื่อขนาดของการทำหน้าที่ต่างกันเพิ่มขึ้น ซึ่งสอดคล้องกับ Cao, Tay and Liu (2017) พบว่า ขนาดของการทำหน้าที่ต่างกันเพิ่มขึ้นจาก .40 เป็น .70 และ .70 เป็น 1.00 มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามเพิ่มขึ้น

3. ปัจจัยสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกัน ผลการทดสอบพบว่า วิธีโพลี–ซิปเทสท์ไม่สามารถควบคุมประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามได้ตามเกณฑ์ที่กำหนด ซึ่งสอดคล้องกับ Kilmen (2016) ที่พบว่า วิธีโพลี–ซิปเทสท์มีประสิทธิภาพลดลง เมื่อสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกันมากขึ้น ส่วนวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ พบว่า เมื่อสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกันมากขึ้น ประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามลดลง ซึ่งไม่สอดคล้องกับ French and Finch (2008) ทำการศึกษาการทำหน้าที่ต่างกันของข้อคำถามด้วยวิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ โดยที่สัดส่วนการทำหน้าที่ต่างกันเท่ากับ ร้อยละ 17 และร้อยละ 34 พบว่า สัดส่วนข้อคำถามที่ทำหน้าที่ต่างกันที่มีค่าเพิ่มขึ้นไม่มีผลต่อประสิทธิภาพของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม และสำหรับการทดสอบวอลด์ พบว่า เมื่อสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกันมากขึ้น ประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามเพิ่มขึ้น สอดคล้องกับการศึกษาของ Cao, Tay and Liu (2017); Woods, Cai and Wang (2013) และ Carroll (2015) ที่ทำการทดสอบภายใต้ปัจจัยสัดส่วนที่อยู่ตั้งแต่ร้อยละ 10 ถึงร้อยละ 20

4. ปัจจัยขนาดตัวอย่าง ผลการทดสอบพบว่า วิธีโพลี–ซิปเทสท์ ไม่สามารถควบคุมประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามได้ตามเกณฑ์ที่กำหนด ซึ่งสอดคล้องกับการศึกษาของ Gonzales–Roma, Hernandez and Gomez–Benito, (2006) ที่พบว่า เมื่อขนาดของตัวอย่างที่ใช้ศึกษามีค่ามากขึ้น วิธีโพลี–ซิปเทสท์ไม่มีประสิทธิภาพในการทดสอบการทำหน้าที่ต่างกันของข้อคำถาม ส่วนวิธีการวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ พบว่า เมื่ออัตราส่วนของกลุ่มสนใจและกลุ่มอ้างอิงมีค่าเพิ่ม ประสิทธิภาพของการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามจะมีค่าลดลง สอดคล้องกับการศึกษาของ French and Finch (2008); Kim and Yoon (2011); Stark, Chernyshenko, and Drasgow (2006) พบว่า เมื่อขนาดตัวอย่างเพิ่มขึ้นประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามเพิ่มขึ้น ส่วนวิธีการทดสอบวอลด์ พบว่า เมื่อขนาดตัวอย่างมีค่ามากขึ้น ประสิทธิภาพของการตรวจสอบทำหน้าที่ต่างกันของข้อคำถามเพิ่มขึ้น เช่นเดียวกับการศึกษาของ Cao, Tay, and Liu (2017); Langer (2008); Woods, Cai and Wang (2013) ที่พบว่า วิธีการทดสอบวอลด์มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามเพิ่มขึ้น เมื่อขนาดตัวอย่างเพิ่มขึ้น โดยความสอดคล้องของการศึกษาของทั้งสามวิธีนั้นเป็นผลจากการศึกษาของนักวิจัยที่ได้

กล่าวถึง ทำการศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามภายใต้ขนาดตัวอย่างขนาด 250 ถึง 1000

ข้อเสนอแนะการวิจัย

ข้อเสนอแนะสำหรับการปฏิบัติ

1. ถ้าผู้ใช้จะนำวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามไปใช้กับแบบทดสอบสองมิติให้คะแนน 5 ค่า ที่มีความยาวแบบทดสอบ 20 ข้อ วิธีวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุมีความเหมาะสมกว่าอีกสองวิธี สำหรับความยาวแบบทดสอบ 40 ข้อ วิธีทดสอบวอลด์จะมีความเหมาะสมกว่าอีกสองวิธี

2. การตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามที่ให้คะแนนหลายค่า ด้วยวิธีโพลีโตมัส-ซิปเทสท์ อาจไม่เหมาะสมกับความยาวแบบทดสอบ 20 ข้อ และ 40 ข้อ ด้วยขนาดของการทำหน้าที่ต่างกัน ตั้งแต่ .40 ขึ้นไป โดยมีสัดส่วนข้อคำถามที่ทำหน้าที่ต่างกันร้อยละ 10 และร้อยละ 20 ทุกขนาดตัวอย่าง เนื่องจากมีอัตราความคลาดเคลื่อนประเภทที่ 1 สูงเกินปกติ (Inflate)

ข้อเสนอแนะสำหรับการวิจัยครั้งต่อไป

1. ควรทำการศึกษากับเครื่องมือวัดแบบหลายมิติที่มีมิติของการศึกษามากกว่า 2 มิติ หรือแบบทดสอบที่ประกอบด้วยข้อคำถามที่ใช้โจทย์หรือสถานการณ์ร่วมกัน (Testlets)

2. ควรศึกษากลุ่มตัวอย่างที่มีมากกว่าสองกลุ่ม เช่น ใช้เกณฑ์เชื้อชาติ ภูมิลำเนา แล้วพิจารณาผลการตรวจสอบว่าแตกต่าง หรือสอดคล้องกับผลการศึกษาครั้งนี้หรือไม่

3. ควรมีการศึกษาเพิ่มเติมเกี่ยวกับปฏิสัมพันธ์ระหว่างปัจจัยที่แปรเปลี่ยน เช่น ความยาวของแบบทดสอบ ขนาดตัวอย่าง สัดส่วนและขนาดของการทำหน้าที่ต่างกัน เพื่อตรวจสอบว่าปัจจัยดังกล่าว มีผลต่อการตรวจสอบโดยตรง หรือเกิดจากปฏิสัมพันธ์ระหว่างปัจจัยอื่น

เอกสารอ้างอิง

ศิริชัย กาญจนวาสี. (2550ก). *ทฤษฎีการทดสอบแนวใหม่*. (พิมพ์ครั้งที่ 3) กรุงเทพ ฯ: โรงพิมพ์จุฬาลงกรณ์มหาวิทยาลัย.

สุชาติ สิริมีนนันท์. (2555). การเปรียบเทียบวิธีโพลีโตมัสซิบเทสท์ วิธีการวิเคราะห์ฟังก์ชันการจำแนกโลจิสติก และวิธีการถดถอยโลจิสติกแบบจัดอันดับ ในการตรวจสอบการทำหน้าที่เบี่ยงเบนของข้อสอบในแบบทดสอบที่มีการให้คะแนนหลายค่า. *วารสารจันทรเกษมสาร*, 35(18), 101–110

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

Ayala, R. J. (2008). *The theory and practice of Item Response Theory*. New York: Guilford.

Cai, L. (2008). SEM of another flavor: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology*, *61*, 309–329.

Cao, M., Tay, L., & Liu, Y. (2017). A Monte Carlo Study of an iterative Wald test procedure for DIF analysis. *Educational and Psychological measurement*, *77*(1), 104–118.

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

Carroll, H. F. C. (2015). *An examination of the improved Wald test for differential item functioning detection with multiple groups*. Doctoral dissertation, Educational Psychology, University of Kansas.

Demars, C. (2010). *Item response theory: Understanding statistics measurement*. Oxford University Press.

Finch, H., & French, B. F. (2007). Detection of crossing differential item functioning item: A comparison of four methods. Educational and Psychological Measurement, 67,565–582.

French, B. F., & Finch, W. H. (2008). Multi–group confirmatory factor analysis: Locating the invariant referent. *Structural Equation Modeling*, *15*, 96–113.

Gonzales–Roma, V., Hernandez, A., & Gomez–Benito, J. (2006). Power and Type–I error of the mean and covariance structure analysis model for detecting differential item functioning in graded response items. *Multivariate Behavioral Research*, *41*(1), 29–53.

Kilmen, S. (2016). Effect of DIF magnitudes, focal group sample size, and DIF ratio on the performance of SIBTEST. *International Journal of Social Sciences and Education*, *6*(1), 91–97.

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

Langer, M. (2008). *A reexamination of Lord's Wald test for differential item functioning using item response theory and modern error estimation*. North Carolina: University of North Carolina, Chapel Hill.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and

confirmatory factor analytic methodologies for establishing measurement equivalence/ invariance. Organizational Research Methods, *7*(4), 361–388.

Meade, A. W., Lautenschlager, G. J., & Johnson, E. C. (2006). Alternative cutoff values and DFIT tests of measurement invariance. Paper presented at the 21st annual conference of the Society for Industrial and Organizational Psychology. Dallas, TX.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*, 552–566.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, *91*(6), 1292–1306.

Vandenberg, R. J. & Lance, C. E. (2000). A review and synthesis of measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–70.

Wang, W. C., & Su, Y. Y. (2004). Factors influencing the Mantel and generalized Mantel–Haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement*, *28*, 450–480.

Williams, N. J., & Beretvas, S. N. (2006). DIF identification using HGLM for polytomous items. *Applied Psychological Measurement*, *30*, 22–42.

Wilson, M., & Hoskens, M. (2005). *Multidimensional item response: Multimethod/ Multitrait perspective*. In, S. Alagumalai, D.D. Curtis, and N. Hungi (Eds.). Applied rasch measurement: A Book exemplars, Springer, 287–307.

Woods, C. M., Cai, L., & Wang, M. (2013). The Langer–Improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two–group IRT. *Educational and Psychological Measurement. 73*(3), 532–547.

Wu, Q., & Lei, P. W. (2009). Using multi–group confirmatory factor analysis to detect differential item functioning when tests are multidimensional. Paper presented at the Annual Meeting of the National Council for Measurement in Education, San Diego: CA.