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

วาสนา กลมอ่อน

ดุษฎีนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปรัชญาดุษฎีบัณฑิต
สาขาวิชาวิจัย วัดผลและสถิติการศึกษา
คณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา
สิงหาคม 2560

คณะกรรมการควบคุมดุษฎีนิพนธ์ และคณะกรรมการสอบดุษฎีนิพนธ์ ได้พิจารณาดุษฎีนิพนธ์ของ วาสนา กลมอ่อน ฉบับนี้แล้ว เห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปรัชญาดุษฎีบัณฑิต สาขาวิชาวิจัย วัดผลและสถิติการศึกษา ของมหาวิทยาลัยบูรพาได้

คณะกรรมการควบคุมดุษฎีนิพนธ์

..... อาจารย์ที่ปรึกษาหลัก
(รองศาสตราจารย์ ดร.ไพรัตน์ วงษ์นาม)

..... อาจารย์ที่ปรึกษาร่วม
(ผู้ช่วยศาสตราจารย์ ดร.สุรีพร อนุศาสนนันท์)

คณะกรรมการสอบดุษฎีนิพนธ์

..... ประธาน
(ดร.อาวีพร ปานทอง)

..... กรรมการ
(รองศาสตราจารย์ ดร.ไพรัตน์ วงษ์นาม)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.สุรีพร อนุศาสนนันท์)

..... กรรมการ
(ดร.ณัฐกฤตา งามมีฤทธิ์)

คณะศึกษาศาสตร์อนุมัติให้รับดุษฎีนิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปรัชญาดุษฎีบัณฑิต สาขาวิชาวิจัย วัดผลและสถิติการศึกษา ของมหาวิทยาลัยบูรพา

..... คณบดีคณะศึกษาศาสตร์
(รองศาสตราจารย์ ดร.วิชิต สุรัตน์เรืองชัย)
วันที่ 4 เดือน พฤษภาคม พ.ศ. 2560

กิตติกรรมประกาศ

ดุษฎีนิพนธ์ฉบับนี้สำเร็จได้ด้วยความกรุณาจาก รองศาสตราจารย์ ดร.ไพรัตน์ วงษ์นาม อาจารย์ที่ปรึกษาหลัก และผู้ช่วยศาสตราจารย์ ดร.สุรีพร อนุศาสนนันท์ อาจารย์ที่ปรึกษาร่วม ที่กรุณาให้คำปรึกษาแนะนำแนวทางที่ถูกต้อง ตลอดจนแก้ไขข้อบกพร่องต่าง ๆ ด้วยความละเอียดถี่ถ้วนและเอาใจใส่ด้วยดีเสมอมา ผู้วิจัยรู้สึกซาบซึ้งเป็นอย่างยิ่ง จึงขอกราบขอบพระคุณเป็นอย่างสูงมา ณ โอกาสนี้

ขอขอบพระคุณ ดร.อาวีพร ปานทอง ประธานสอบปากเปล่า และ ดร.ณัฐกฤตา งามมีฤทธิ์ กรรมการสอบปากเปล่า ที่ได้กรุณาให้ข้อเสนอแนะในการปรับปรุงแก้ไขจนทำให้ดุษฎีนิพนธ์ฉบับนี้มีความถูกต้องและสมบูรณ์ยิ่งขึ้น

ขอขอบพระคุณคณาจารย์ทุกท่านที่ได้ให้คำแนะนำสั่งสอนความรู้ต่าง ๆ ตั้งแต่เริ่มเข้ารับการศึกษา ณ มหาวิทยาลัยแห่งนี้ จนกระทั่งผู้วิจัยได้ทำดุษฎีนิพนธ์ฉบับนี้เสร็จสมบูรณ์

ขอขอบพระคุณ รองศาสตราจารย์ ดร.มนตรี แย้มกสิกร รองศาสตราจารย์ ดร.วิชิต สุรัตน์เรืองชัย ผู้ช่วยศาสตราจารย์ ดร.ระพินทร์ ฉายวิมล ดร.จันทร์พร พรหมมาศ และคณะผู้บริหารทุกท่านที่ให้ความเมตตาและให้โอกาสด้านการศึกษาแก่ผู้วิจัย จนสำเร็จลุล่วงด้วยดี

ขอขอบคุณ รองศาสตราจารย์ ดร.ฉลอง ทับศรี ผู้ช่วยศาสตราจารย์ ดร.เชษฐ์ ศิริสวัสดิ์ ดร.สมพงษ์ ปั้นหุ่น และ ดร.ดลดาว ปูรณานนท์ พี่ เพื่อน และน้อง ๆ สาขาวิชาวิจัย วัดผลและสถิติการศึกษา และเพื่อนร่วมงานทุกคนที่ให้คำแนะนำและเป็นกำลังใจแก่ผู้วิจัยเสมอมา โดยเฉพาะอย่างยิ่งน้องในฝ่ายยุทธศาสตร์ และน้อง ๆ ทีมงานกัลยาณมิตรทุกคน ที่ให้กำลังใจและให้ความช่วยเหลือในขั้นตอนต่าง ๆ ระหว่างการวิจัยอย่างดียิ่ง จนกระทั่งดุษฎีนิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดี

ขอขอบคุณ คุณนุปกรณ์ ทองคำ พี่สาวที่แสนดี และน้อง ๆ ในครอบครัวทุกคน ที่ให้ความรัก ความห่วงใย คอยดูแล ช่วยเหลือในทุกสิ่งทุกอย่าง และเป็นกำลังใจแก่ผู้วิจัยเสมอมา โดยเฉพาะอย่างยิ่ง คุณสุรสิทธิ์ กลมอ่อน และคุณมณีรัตน์ กลมอ่อน ที่เป็นขวัญกำลังใจสำคัญยิ่ง และเป็นแรงผลักดันให้ผู้วิจัยมีความเพียรมุ่งมั่นทำให้ดุษฎีนิพนธ์ฉบับนี้สำเร็จได้ด้วยดีและมีวันนี้

ขอกราบขอบพระคุณ คุณพ่ออำนวย ฐิติวรารักษ์ และคุณแม่ดวงใจ ฐิติวรารักษ์ ที่เป็นผู้อบรมเลี้ยงดู มอบความรัก ความห่วงใย และเป็นพลังอันยิ่งใหญ่ของผู้วิจัยเสมอมา

คุณค่าและคุณประโยชน์ของดุษฎีนิพนธ์ฉบับนี้ ผู้วิจัยขอมอบเป็นกตัญญูกตเวทีแด่บุพการี บูรพาจารย์ และผู้มีพระคุณทุกท่านทั้งในอดีตและปัจจุบันที่ทำให้ข้าพเจ้าเป็นผู้มีการศึกษาและประสบความสำเร็จมาจนตราบเท่าทุกวันนี้

วาสนา กลมอ่อน

52810151: สาขาวิชา: วิจัย วัดผลและสถิติการศึกษา; ปร.ด. (วิจัย วัดผล และสถิติการศึกษา)
คำสำคัญ: การทำหน้าที่ต่างกันของข้อสอบ/ อัตราความคลาดเคลื่อนประเภทที่ 1/ อัตรา
อำนาจการทดสอบ/ IRT LR/ IRT Poly-SIBTEST

วาสนา กลมอ่อน: 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) คณะกรรมการควบคุมดุษฎีนิพนธ์: ไพรัตน์ วงษ์นาม, ค.ด., สุรีพร อนุศาสนนันท์, ค.ด. 214 หน้า. ปี พ.ศ. 2560.

การวิจัยนี้มีวัตถุประสงค์เพื่อ 1) ตรวจสอบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย คือ ขนาดการทำหน้าที่ต่างกันของข้อสอบ 2 รูปแบบ ความยาวของแบบสอบ 2 รูปแบบ และขนาดของกลุ่มตัวอย่าง 3 ขนาด และ 2) เปรียบเทียบอัตราความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสามวิธีภายใต้เงื่อนไขปัจจัยที่แตกต่าง 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 และอัตราอำนาจการทดสอบสูงกว่าเกณฑ์ที่กำหนดภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบที่มีขนาดกลาง สำหรับวิธี Poly-SIBTEST มีอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบ ไม่อยู่ในเกณฑ์ที่กำหนดเกือบทุกเงื่อนไขปัจจัย และวิธี Multiple-groups CFA มีอัตราความคลาดเคลื่อนประเภทที่ 1 สูงกว่าเกณฑ์ที่กำหนด และอัตราอำนาจการทดสอบ สูงกว่าเกณฑ์ที่กำหนดภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบที่มีขนาดกลาง

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

52810151: MAJOR: EDUCATIONAL RESEARCH, MEASUREMENT AND STATISTICS;
Ph.D. (EDUCATIONAL RESEARCH MEASUREMENT AND STATISTICS)
KEYWORDS: DETECTING DIFFERENTIAL ITEM FUNCTIONING/ TYPE I ERROR RATE/
POWER RATE/ IRT LR/ IRT POLY-SIBTEST

WASANA KLOM-ON: COMPARISON OF THE EFFICIENCY OF DIFFERENTIAL ITEM FUNCTIONING FOR POLYTOMOUS SCORED ITEMS: IRT LR, POLY-SIBTEST AND MULTIPLE-GROUPS CFA METHOD. ADVISORY COMMITTEE: PAIRATANA WONGNAM, Ph.D., SUREEPORN ANUSASANANAN, Ph.D., 214 P. 2017.

The purpose of this research were: 1) to detect 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 the investigated differential item functioning under a variety of three factors differences under 3 different conditional 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). The data were simulated under the unidimensional Graded-Response Model, and all items were in five response categories scoring of 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 was that the 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 at nominal limits on almost all conditions. The type I error rate on Multiple-groups CFA procedure was higher than nominal limit on overall conditions and power rate was higher than the nominal limit under medium magnitude of DIF.

2. The 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 different $(\alpha = .001)$. Moreover, the result of methods depended on magnitude of DIF, test length, and sample size.

สารบัญ

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญตาราง (ต่อ)

สารบัญภาพ

บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

การวัดทางจิตวิทยาส่วนใหญ่จะเป็นการวัดทางอ้อมผ่านตัวแปรแฝงคุณลักษณะภายในของแต่ละบุคคล โดยวัดจากพฤติกรรมที่สังเกตได้ผ่านทางเนื้องานหรือข้อคำถามที่เกี่ยวข้อง ซึ่งทั้งตัวบุคคลและข้อคำถามในมิติทางจิตวิทยาจึงต้องสรุปลงความเห็นตามพฤติกรรมที่สังเกตได้ (Embretson & Reise, 2000, pp. 40-41) และเนื่องจากการวัดคุณลักษณะภายในของมนุษย์มีความสำคัญและจำเป็นต้องศึกษา เพื่อให้เข้าใจถึงการเกิดพฤติกรรมภายนอกของมนุษย์ อันจะนำไปสู่การทำนาย ควบคุม และพัฒนาพฤติกรรมมนุษย์ จึงจำเป็นต้องอาศัยทฤษฎีการทดสอบเพื่อช่วยให้นักวัดผลสามารถทำการสร้างและพัฒนาแบบสอบให้มีคุณภาพ แปลความหมายผลการวัดได้อย่างถูกต้อง และใช้เป็นสารสนเทศสำหรับการตัดสินใจทางการศึกษาและจิตวิทยาได้อย่างเหมาะสม ทั้งนี้ ในการสร้างแบบสอบต่าง ๆ เช่น แบบสอบผลสัมฤทธิ์ แบบวัดความคิดแบบวัดความสามารถ เป็นต้น ตามแนวคิดของทฤษฎีการทดสอบแบบดั้งเดิม (Classical Test Theory: CTT) นิยมสร้างและคัดเลือกข้อสอบตามเนื้อเรื่องรวมทั้งพิจารณาค่าพารามิเตอร์ความยาก (Difficulty parameter) และค่าพารามิเตอร์อำนาจจำแนก (Discriminattion parameter) ของข้อสอบ ซึ่งข้อสอบที่มีค่าอำนาจจำแนกสูงมักเป็นข้อที่พึงประสงค์ แต่สำหรับระดับความยากของข้อสอบมักเลือกตามการคาดคะเนถึงระดับความสามารถของกลุ่มผู้สอบที่มาทำการทดสอบ โดยมีข้อจำกัดที่สำคัญคือ ค่าความยากและค่าอำนาจจำแนกของข้อสอบผันแปรตามกลุ่มผู้สอบที่มีระดับความสามารถต่างกัน ดังนั้น คุณภาพของการเลือกข้อสอบตามแนว CTT จึงขึ้นอยู่กับความใกล้เคียงระหว่างกลุ่มผู้สอบที่นำมาใช้ในการคำนวณค่าความยากและค่าอำนาจจำแนก กับกลุ่มเป้าหมายที่จะทำการทดสอบ ถ้ากลุ่มทั้งสองแตกต่างกันมาก ค่าความยากและค่าอำนาจจำแนกที่เคยคำนวณไว้จะไม่เหมาะสมกับกลุ่มเป้าหมายที่จะทำการทดสอบ นอกจากนี้ยังไม่สามารถควบคุมความคลาดเคลื่อนของการวัดได้อย่างมีประสิทธิภาพ เพราะข้อสอบแต่ละข้อมีผลต่อค่าความเที่ยงของแบบสอบที่ไม่เป็นอิสระกัน เนื่องจากค่าความเที่ยงของแบบสอบขึ้นอยู่กับระดับความสัมพันธ์ซึ่งกันและกันของข้อสอบทุกข้อที่อยู่ในแบบสอบฉบับนั้น จึงไม่สามารถจำแนกอิทธิพลของข้อสอบแต่ละข้อที่มีต่อค่าความเที่ยงของแบบสอบทั้งฉบับอย่างเป็นอิสระจากกันได้ และจากข้อจำกัดของค่าพารามิเตอร์ของข้อสอบผันแปรตามกลุ่มผู้สอบ และคะแนนสังเกตได้หรือค่าประมาณความสามารถของผู้ตอบไม่เป็นอิสระหรือขึ้นอยู่กับข้อสอบและแบบสอบที่นำมาใช้

ในวงการศึกษาและหน่วยงานต่าง ๆ มักนำผลการทดสอบมาใช้เป็นข้อมูลเพื่อตัดสินใจในเรื่องต่าง ๆ มากขึ้น ซึ่งแบบทดสอบที่นำมาใช้ควรเป็นแบบทดสอบมาตรฐานที่มีความเที่ยงตรง ความเชื่อมั่น ความยาก ที่เหมาะสม ตลอดจนจำแนกระดับความสามารถของผู้สอบได้ นอกจากนี้ ยังต้องคำนึงถึงความยุติธรรมต่อผู้สอบด้วย (อรินทร์ น่วมถนอม, 2549, หน้า 137) การศึกษาคุณภาพด้านความยุติธรรมของข้อสอบหรือแบบสอบระหว่างผู้สอบกลุ่มต่าง ๆ เป็นการตรวจสอบความลำเอียงของข้อสอบ (Item bias) ความลำเอียงของแบบสอบ (Test bias) และความลำเอียงในการคัดเลือก (Selection bias) เพื่อจำแนกข้อสอบที่ทำหน้าที่ไม่เหมาะสมหรือไม่ยุติธรรมออกจากแบบสอบ ซึ่งเป็นการพัฒนาแบบสอบให้มีคุณภาพเหมาะสมสำหรับนำไปใช้ทดสอบต่อไป ทั้งนี้ ความยุติธรรมของแบบสอบเป็นประเด็นสำคัญในการทดสอบทางการศึกษาและจิตวิทยา ผลการทดสอบที่ได้ไม่เพียงเกี่ยวข้องกับข้อสอบ/ ข้อคำถามที่วัดความสามารถของผู้สอบเท่านั้น แต่ยังมีปัจจัยอื่น ๆ ที่เกี่ยวข้องด้วย เช่น การสอบคัดเลือกเข้าศึกษาระดับนานาชาติ สิ่งสำคัญของการสอบต้องไม่มีข้อสอบ/ ข้อคำถามที่เข้าข้างนักเรียนในเขตพื้นที่ตั้งหรือนักเรียนที่มีฐานะทางเศรษฐกิจและสังคม (Chang, Huang, & Tsai, 2015, p. 181) ข้อสอบที่เข้าข้างผู้สอบกลุ่มใดกลุ่มหนึ่งมีผลทำให้ผู้สอบกลุ่มนั้นได้เปรียบ ส่วนผู้สอบอีกกลุ่มหนึ่งเสียเปรียบ ทั้ง ๆ ที่ผู้สอบทั้งสองกลุ่มมีระดับความสามารถเท่ากัน แบบสอบนั้นคือแบบสอบที่ขาดความยุติธรรม ซึ่งเป็นผลมาจากความลำเอียงของข้อสอบ (Item bias) และความลำเอียงของแบบสอบ (Test bias) (อรินทร์ น่วมถนอม, 2549, หน้า 137) ดังนั้นจึงจำเป็นต้องมีการตรวจสอบความลำเอียงเพื่อจำแนกข้อสอบที่ทำหน้าที่ไม่เหมาะสมหรือไม่ยุติธรรม ทั้งนี้ ความลำเอียงดังกล่าวก็คือ ความคลาดเคลื่อนอย่างเป็นระบบ (Systematic error) ที่เกิดจากการวัด

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) โดยในระยะแรกของการพัฒนาโมเดล IRT จะเน้นที่การวิเคราะห์ข้อสอบที่มีรูปแบบการตอบแบบสองค่า (Binary response data) แต่ภายหลังได้มีการพัฒนาขยายไปสู่การตอบในรูปแบบต่าง ๆ เช่น รูปแบบการตอบที่ตัวเลือกเป็นแบบมาตรประมาณค่า (Rating scale) รูปแบบการตอบที่เป็นการให้คะแนนความรู้บางส่วน (Partial credit scoring) หรือการให้คะแนนแบบหลายกลุ่ม (Multiple category scoring) เป็นต้น (อิทธิฤทธิ์ พงษ์ปิยะรัตน์, 2551, หน้า 4) นอกจากนี้ ค่าพารามิเตอร์ของข้อสอบ

ตามทฤษฎีการตอบสนองข้อสอบมีลักษณะไม่แปรเปลี่ยนตามกลุ่มผู้สอบ การคัดเลือกข้อสอบแต่ละข้อจึงสามารถทำได้อย่างอิสระ รวมทั้งค่าพารามิเตอร์ความยากและความสามารถ (θ) ของผู้สอบได้รับการประมาณค่าอยู่บนสเกลเดียวกัน จึงทำให้เลือกข้อสอบแต่ละข้อให้ทำหน้าที่ดีที่สุด ณ บริเวณใดบริเวณหนึ่งบนสเกลของความสามารถได้ เช่น กำหนดจุดตัดสำหรับ ผู้รอบรู้/ ไม่รอบรู้ ณ ตำแหน่ง θ ที่ต้องการ เป็นต้น นอกจากนี้ข้อได้เปรียบที่สำคัญของ IRT ก็คือ สามารถเลือกข้อสอบเป็นรายข้อบนพื้นฐานของปริมาณสารสนเทศที่จะได้รับ สำหรับผู้สอบที่มี θ ต่าง ๆ กัน สารสนเทศของข้อสอบสะท้อนความถูกต้องแม่นยำของการประมาณค่า และเมื่อนำมารวมกันจะเป็นสารสนเทศของแบบสอบ จึงสามารถใช้เป็นหลักประกันว่า จะได้แบบสอบตามเป้าหมายที่สนองต่อการนำไปใช้ที่ให้ผลแม่นยำตามที่ต้องการ (ศิริชัย กาญจนวาสี, 2555, หน้า 9; 78-85) โค้งลักษณะข้อสอบ (Item Characteristic Curve: ICC) เป็นพื้นฐานของทฤษฎี IRT และส่วนใหญ่กำหนดลักษณะความสัมพันธ์เป็นแบบฟังก์ชันโลจิสติก (Logistic fuction) ซึ่งเป็นโมเดลความสัมพันธ์ระหว่างผลการตอบข้อสอบหรือข้อคำถามกับความสามารถที่มีอยู่ของแต่ละบุคคล (Edelen & Reeve, 2007, p. 6) สำหรับทฤษฎี IRT จำแนกเป็น 2 ประเภท คือ 1) ทฤษฎีการตอบสนองข้อสอบแบบตรวจให้คะแนนสองค่า (Dichotomous item response theory) ใช้กับการตรวจคะแนนรายข้อแบบสองค่า เช่น ข้อสอบหรือข้อคำถามที่ตรวจให้คะแนนแบบ 0 หรือ 1 (ตอบผิดได้ 0, ตอบถูกได้ 1) แบบถูกหรือผิด แบบเห็นด้วยหรือไม่เห็นด้วย แบบใช่หรือไม่ใช่ เป็นต้น และ 2) ทฤษฎีการตอบสนองข้อสอบแบบตรวจให้คะแนนมากกว่าสองค่า (Polytomous item response theory) ใช้กับการตรวจคะแนนรายข้อแบบมากกว่าสองค่า เช่น ข้อสอบหรือข้อคำถามมาตรประมาณค่า (Rating scale) การตรวจข้อสอบแบบให้คะแนนความรู้บางส่วน (Partial scale) เป็นต้น ซึ่งโมเดลการตอบสนองข้อสอบแบบตรวจให้คะแนนที่นิยมใช้กัน ได้แก่ โมเดลแบบหนึ่งพารามิเตอร์ (One-parameter model) โมเดลแบบสองพารามิเตอร์ (Two-parameter model) และโมเดลแบบสามพารามิเตอร์ (Three-parameter model) การใช้โมเดลการตอบสนองข้อสอบที่สอดคล้องกับข้อมูล (Model-Data Fit) จะทำให้ค่าพารามิเตอร์ที่ได้มีลักษณะไม่แปรเปลี่ยน (Invariance) ทั้งความไม่แปรเปลี่ยนของค่าประมาณความสามารถและความไม่แปรเปลี่ยนของค่าประมาณพารามิเตอร์ของข้อสอบ จึงทำให้การเลือกข้อสอบแต่ละข้อ เพื่อสร้างชุดข้อสอบเป็นแบบสอบกระทำได้อย่างมีประสิทธิภาพ นอกจากนี้ ปริมาณสารสนเทศที่ได้จากข้อสอบ เมื่อนำมารวมกันเป็นสารสนเทศของแบบสอบจะสะท้อนถึงความถูกต้องแม่นยำของการประมาณค่าความสามารถ (θ) เพื่อให้ได้แบบสอบตามเป้าหมายที่สนองต่อการนำไปใช้อย่างมีประสิทธิผลและประสิทธิภาพ (ศิริชัย กาญจนวาสี, 2555, หน้า 77-86)

วิธีการวัดของ IRT อยู่บนพื้นฐานของโมเดลและการเปรียบเทียบค่าพารามิเตอร์ที่เปลี่ยนไปในแต่ละโมเดล เช่น โมเดล 1-พารามิเตอร์ จะเป็นการเปรียบเทียบค่าพารามิเตอร์ความยากของข้อสอบ (b) ของแต่ละกลุ่ม โมเดล 2-พารามิเตอร์ เป็นการเปรียบเทียบค่าพารามิเตอร์อำนาจจำแนก (a) และค่าพารามิเตอร์ความยากของข้อสอบ (b) ในแต่ละกลุ่ม แต่ถ้าเป็นการเปรียบเทียบระหว่างกลุ่มค่าพารามิเตอร์ความยากของข้อสอบ (b) แตกต่างกัน แสดงให้เห็นว่าเป็นโมเดลการตอบสนองข้อสอบที่เป็นรูปแบบเดียวกัน (Uniform DIF) และหากค่าพารามิเตอร์อำนาจจำแนก (a) แตกต่างกัน แสดงให้เห็นว่าเป็นโมเดลการตอบสนองข้อสอบที่ไม่เป็นรูปแบบเดียวกัน (Non-uniform DIF) เป็นต้น ทั้งนี้ โมเดลการตอบสนองข้อสอบมีทั้งแบบตรวจให้คะแนน 2 ค่า และมากกว่า 2 ค่า ซึ่งโมเดลการตอบสนองข้อสอบแบบตรวจให้คะแนนมากกว่า 2 ค่า มักพบได้ในแบบสอบทางการศึกษาและแบบทดสอบทางจิตวิทยา (Nering & Ostini, 2010, p. 3) โดยเฉพาะโมเดล Graded-response (GRM) ที่นำเสนอโดย Samejima (1969, 1996) พัฒนามาเพื่อใช้กับแบบสอบหรือแบบวัดที่แต่ละข้อคำถามมีรายการคำตอบแบบมาตรเรียงลำดับ (Ordered categorical responses) และใช้หลักการคำนวณความน่าจะเป็นของการตอบแต่ละรายการคำตอบแบบ 2 ขั้นตอน โดยขั้นตอนแรก คือ คำนวณค่าความชันร่วมของแต่ละข้อคำถาม (α) และขั้นตอนที่สอง คำนวณค่าพารามิเตอร์ของแต่ละรายการคำตอบในแต่ละข้อคำถาม (β) (ศิริชัย กาญจนวาสี, 2555, หน้า 89)

ในปัจจุบันนักวิจัยนิยมใช้คำว่า “การทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF)” แทน ความลำเอียงของข้อสอบ (Item bias) ซึ่งมีความเหมาะสมมากกว่าในการอธิบายสารสนเทศเชิงสถิติ (อรินทร์ น่วมถนอม, 2549, หน้า 138) และเน้นการใช้วิธีการทางสถิติในการตรวจสอบข้อสอบ เพื่อเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มผู้สอบที่เป็นกลุ่มเปรียบเทียบ (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 อ้างถึงใน อรินทร์ น่วมถนอม, 2549 หน้า 138) สอดคล้องกับ Embretson and Reise (2000, pp. 251-252) กล่าวว่า สำหรับวิธีการตรวจสอบ DIF ง่าย ๆ ก็คือการใช้กลุ่มผู้สอบหรือกลุ่มตัวอย่างขนาดใหญ่ ทั้งในกลุ่มอ้างอิง (Reference group) และกลุ่มเปรียบเทียบ (Focal group) และประมาณค่าพารามิเตอร์แยกเป็น 2 กลุ่ม แล้วเปรียบเทียบโค้งการตอบสนองข้อสอบ (IRCs) หรือใช้กระบวนการทางสถิติที่ซับซ้อน

การเลือกวิธีตรวจสอบ DIF ไม่ได้มุ่งวิธีที่หลากหลายแต่มุ่งศึกษาเชิงลึกในการตรวจสอบเพื่อให้เกิดสารสนเทศมากยิ่งขึ้น (Gómez-Benito, Hidalgo, & Padilla, 2009) ถ้าในแบบสอบมีจำนวนข้อสอบที่ทำหน้าที่ต่างกันหลายข้อจะส่งผลให้ความตรงของแบบสอบลดน้อยลง ข้อสอบที่ทำหน้าที่ต่างกัน คือ การเกิดความสัมพันธ์ทางสถิติระหว่างผลการตอบข้อสอบในกลุ่มผู้สอบเมื่อกำหนดระดับคะแนนที่ใช้เป็นเกณฑ์ในการจับคู่ทำให้เกิดรูปแบบของการทำหน้าที่ต่างกันโดยพิจารณาจากขนาดและทิศทางที่แปรเปลี่ยนไปตามระดับความสามารถที่ต่างกันของกลุ่มผู้สอบ (ธเกียรติกมล ทองงอก, 2554, หน้า 5) ซึ่ง ศิริชัย กาญจนวาสี (2555, หน้า 120-122) เสนอแนะว่าเงื่อนไขสำคัญของหลักการตรวจสอบ DIF จำเป็นต้องจับคู่ (Matching) ผู้สอบตามความสามารถและเกณฑ์การจับคู่ (Macthing criteria) ที่นิยมใช้กันมี 2 วิธี คือ 1) เกณฑ์ภายนอก (External criterion) โดยใช้คะแนนจากแบบสอบอื่นเป็นเกณฑ์ภายนอกแล้วใช้เทคนิคการวิเคราะห์ถดถอย (Regression analysis) ทำการเปรียบเทียบกราฟความสัมพันธ์ระหว่างตัวแปรเกณฑ์ กับตัวแปรทำนายระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ โดยฟังก์ชันการทำนายสมการของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ สามารถเปรียบเทียบค่าตัดแกน (Intercept: a) และ ค่าความชัน (Slope: b) ของเส้นกราฟระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบได้ ถ้าเส้นกราฟมีค่าตัดแกน (a) หรือค่าความชัน (b) แตกต่างกันแสดงว่าข้อสอบหรือแบบสอบนั้นมีการทำหน้าที่ต่างกัน โดยลำเอียงเข้าข้างกลุ่มผู้สอบที่มีค่าตัดแกนตั้งสูงกว่าหรือค่าความชันที่สูงกว่า โดยมีข้อดีคือ เกณฑ์ที่ใช้มีความเป็นอิสระจากข้อสอบและแบบสอบ แต่มีจุดอ่อนที่ความเหมาะสมของเกณฑ์ที่จะนำมาใช้ แต่ในทางปฏิบัติยากที่จะหาตัวแปรเกณฑ์ภายนอกจากแบบสอบฉบับอื่นที่มีความตรงเชิงทำนายและมีความยุติธรรมสำหรับกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ซึ่งหากขาดคุณสมบัติดังกล่าวนี้ จะทำให้ผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบหรือแบบสอบขาดความแม่นยำและความสมบูรณ์ สำหรับวิธีที่ 2) เกณฑ์ภายใน (Internal criterion) เป็นการนำวิธีการทางสถิติมาตรวจสอบการทำหน้าที่ต่างกันของข้อสอบหรือแบบสอบ โดยเน้นการพิจารณาโครงสร้างภายในของแบบสอบเป็นหลักด้วยการวิเคราะห์ผลจากการตอบข้อสอบและความสามารถหรือคะแนนจริงของผู้สอบที่ได้จากแบบสอบฉบับนั้น ๆ เพื่อนำมาเปรียบเทียบผู้สอบจากกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบที่มีความสามารถหรือคะแนนจริงเท่ากันว่า จะมีผลการตอบหรือโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกันหรือไม่ เพื่อบ่งชี้ถึงการทำหน้าที่ต่างกันของข้อสอบ

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมีหลากหลายวิธีตามรูปแบบการตรวจให้คะแนนแบบ 2 ค่า หรือมากกว่า 2 ค่า เช่น การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีอัตราส่วนความควรจะเป็น (Likelihood ratio test) วิธี Poly-SIBTEST ที่ขยายมาจากวิธี SIBTEST วิธีการถดถอยโลจิสติก วิธีแมนเทล-แฮนด์เซล เป็นต้น (Potenza & Dorans, 1995) ซึ่งในปัจจุบัน

มีการนำวิธีในกลุ่มของ SEM มาใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมากขึ้น เช่น งานวิจัยของ Meade and Lautenschlager (2004) และ Muthén and Asparouhov (2014) เป็นต้น และการวิเคราะห์องค์ประกอบเชิงยืนยัน (Confirmatory Factor Analysis: CFA) เป็นการประเมินผลการวัดความไม่แปรเปลี่ยน (Measurment Invariance: MI) ซึ่งรู้จักกันดีในนามของความไม่แปรเปลี่ยนของค่าองค์ประกอบ (Factorial Invariance) หรือ ความเท่ากันของการวัด (Measurement equivalence) ในการศึกษาคุณสมบัติตัวบ่งชี้ที่สังเกตได้ทางจิตวิทยาข้ามกลุ่มต่าง ๆ เช่น กลุ่มเพศ กลุ่มภาษา เป็นต้น หรือช่วงเวลาหรือเงื่อนไขที่ต่างกัน นอกจากนี้ยังเป็นวิธีการทดสอบการวัดตัวบ่งชี้ที่มีโครงสร้างเดียวกันด้วยวิธีการวัดเดียวกันในกลุ่มที่แตกต่างกันหรือช่วงเวลาหรือเงื่อนไขที่ต่างกัน ดังนั้น ผลการตอบของตัวบ่งชี้จะขึ้นอยู่กับคะแนนของคุณลักษณะแฝง โดยไม่ขึ้นอยู่กับการเป็นสมาชิกของกลุ่มที่แตกต่างกัน ดังนั้น ความแตกต่างของผลการตอบที่สังเกตได้ จึงเป็นผลที่เกิดจากความแตกต่างที่แท้จริงของคุณลักษณะที่ต้องวัด ดังนั้น ความไม่แปรเปลี่ยนในการวัด (MI) จึงหมายถึง ไม่มีข้อสอบที่ทำหน้าที่ต่างกัน (Non-DIF) ตามทฤษฎีการตอบสนองข้อสอบ และหากมีความแปรเปลี่ยนในการวัด (Measurement non-invariance) จะหมายถึง พบการทำหน้าที่ต่างกันของข้อสอบ (DIF) (Hoffman, 2014) นอกจากนี้ ในกระบวนการของการตรวจสอบ DIF พบปัจจัยที่เกี่ยวข้องหลายปัจจัย เช่น ปัจจัยความยาวของแบบสอบ ขนาดของกลุ่มตัวอย่าง ความแตกต่างของค่าเบี่ยงเบนมาตรฐาน ความแตกต่างของการแจกแจงข้อมูล และปฏิสัมพันธ์ของปัจจัยต่าง ๆ (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, Arsan, Gök, & Kelecioglu, 2014, p. 2187)

จากการศึกษางานวิจัยที่ผ่านมา มีประเด็นปัญหาที่น่าสนใจ ดังนี้

1. ปัจจุบันการเลือกวิธีตรวจสอบ DIF ไม่ได้มุ่งเน้นวิธีที่หลากหลายเพียงอย่างเดียว แต่มุ่งศึกษาเชิงลึกในการตรวจสอบเพื่อให้เกิดสารสนเทศมากยิ่งขึ้น อันจะเป็นประโยชน์ต่อการพัฒนาคุณภาพของแบบสอบให้มีความตรงและยุติธรรมกับผู้สอบมากที่สุด โดยมีวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่หลากหลายวิธี เช่น วิธีอัตราส่วนความควรจะเป็น (Likelohood ratio test: LR) วิธีโพลี-ซิปเทสท์ วิธีซิปเทสท์ วิธีการถดถอยโลจิสติก วิธีการวิเคราะห์พหุระดับ วิธี MIMIC วิธี Multiple-groups CFA เป็นต้น ซึ่งแต่ละวิธีการมีความเหมาะสม ข้อมูลที่ใช้ในการวิเคราะห์แตกต่างกันและมีประสิทธิภาพในการตรวจสอบภายใต้ปัจจัยเงื่อนไขต่าง ๆ ที่แตกต่างกัน

วิธีอัตราส่วนความควรจะเป็น (Likelihood ratio test: LR) โดย Thissen, Steinberg, and Wainer (1988) เป็นวิธีการในรูปแบบพาราเมตริกและเป็นวิธีการพื้นฐานในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ทั้งข้อสอบที่ทำหน้าที่ต่างกันที่เป็นรูปแบบเดียวกัน (Uniform DIF) และไม่เป็นรูปแบบเดียวกัน (Nonuniform DIF) ซึ่งตามทฤษฎีการตอบสนองข้อสอบ การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเกี่ยวข้องกับฟังก์ชันคะแนนจริงของข้อสอบทั้งข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบสองค่า และแบบหลายค่า ถ้าการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบพบว่า ข้อสอบไม่ได้ทำหน้าที่ต่างกันของข้อสอบ แสดงว่า ฟังก์ชันคะแนนจริงของข้อสอบในกลุ่มอ้างอิงและฟังก์ชันคะแนนจริงของข้อสอบในกลุ่มเปรียบเทียบ เหมือนกัน และการทดสอบสมมติฐานในกระบวนการของวิธีอัตราส่วนความควรจะเป็น เป็นการทดสอบค่าพารามิเตอร์ของข้อสอบระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ ซึ่งถ้ามีความแตกต่างของค่าพารามิเตอร์ความยากของข้อสอบระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ จะเป็นการทดสอบการทำหน้าที่ต่างกันของข้อสอบที่เป็นรูปแบบเดียวกัน และหากมีความแตกต่างของค่าพารามิเตอร์อำนาจจำแนกของข้อสอบระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ จะเป็นการทดสอบการทำหน้าที่ต่างกันของข้อสอบที่ไม่เป็นรูปแบบเดียวกัน และจากการศึกษาเปรียบเทียบวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี IRT LR (IRT likelihood ratio test) และวิธีถดถอยโลจิสติก พบว่า วิธี IRT LR และวิธีถดถอยโลจิสติก สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดี ในขณะที่วิธี IRT LR มีอำนาจการทดสอบสูงภายใต้ขนาดของกลุ่มตัวอย่างที่มีขนาดกลางและขนาดใหญ่ และเมื่อขนาดของกลุ่มตัวอย่างมีขนาดใหญ่ขึ้น วิธี IRT LR มีอำนาจการทดสอบสูงขึ้นด้วย (Atar & Kamata, 2011, p. 36)

วิธี Poly-SIBTEST โดย Chang, Mazzeo, & Roussos (1996) เป็นวิธีที่ปรับขยายมาจากวิธีซิปเทสท์ (SIBTEST) ของ Shealy and Stout, 1993 เพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนหลายค่า (Potenza & Dorans, 1995)โดยวิธีนี้มีจุดเด่นหลายประการ เช่น ใช้เทคนิคการตรวจสอบแบบหลายมิติ โดยแยกข้อสอบที่ใช้เป็นเกณฑ์การจับคู่ออกจากข้อสอบที่ใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบอย่างชัดเจน ทำให้เกณฑ์มีความเที่ยงตรงและมีความเชื่อมั่นสูง คะแนนเกณฑ์การจับคู่ค่อนข้างเป็นคุณลักษณะแฝงมากกว่าคะแนนที่ได้จากการสอบ ทำให้มีความถูกต้องและแม่นยำ มีการคำนวณทวนซ้ำหลายรอบ (Iterative algorithm) เพื่อคัดเลือกข้อสอบที่ทำหน้าที่ต่างกันออกจากคะแนนเกณฑ์การจับคู่ ทำให้เกณฑ์มีความบริสุทธิ์ (Purification) มีการปรับแก้ค่าการถดถอย (Regression correction) เพื่อลดความแตกต่างของค่าความสามารถเป้าหมายระหว่างกลุ่มผู้สอบ ทำให้สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ที่สูงเกินปกติได้ (อรินทร์ น่วมถนอม, 2549, หน้า 13)

Chang et al. (2015) ได้ศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) โดยใช้ Multiple-group categorical CFA (MCCFA) ด้วยวิธีการ Minimum free baseline โดยเน้นการประเมินประสิทธิภาพการวิเคราะห์ Multiple-group categorical CFA (MCCFA) กับความแกร่งของค่าความแตกต่างของค่าไค-สแควร์ในการทดสอบ DIF ของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าภายใต้วิธีการ Minimum Free Baseline เพื่อแสดงให้เห็นถึงประสิทธิภาพของอำนาจการทดสอบ (Power rate) และอัตราความคลาดเคลื่อนประเภทที่ 1 (Type I error rate) ในการตรวจสอบ DIF ที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ซึ่งผลการวิจัย พบว่า วิธี MCCFA ภายใต้วิธีการ Minimum free baseline มีประโยชน์สำหรับการตรวจสอบ DIF ที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ซึ่งวิธีการวิเคราะห์ MCCFA เป็นการวิเคราะห์ในลักษณะเดียวกันกับการวิเคราะห์ CFA แต่เป็นการวิเคราะห์แบบหลายกลุ่ม ซึ่งจะเน้นการทดสอบในการวัดความไม่แปรเปลี่ยน (Measurement Invariance: MI) และหากเป็นการศึกษาตามทฤษฎี IRT จะเน้นการทดสอบการทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) นั่นก็คือ เมื่อศึกษาพบว่า มีความไม่แปรเปลี่ยนในการวัด (MI) ก็มีความหมายเดียวกันกับ การทำหน้าที่ต่างกันของข้อสอบไม่แตกต่างกัน (No-DIF)

นอกจากนี้ Yoon and Millsap (2007) ได้ศึกษาการวิเคราะห์ความไม่แปรเปลี่ยนขององค์ประกอบในแบบสอบ โดยใช้การจำลองข้อมูล และกำหนดเงื่อนไขขนาดการทำหน้าที่ต่างกันของข้อสอบที่ใช้ในการศึกษาเป็น 3 ขนาด คือ ขนาด 0.01, 0.02 และ0.03 และ Lopez Rivas, Stark, and Chernyshenko (2009) ได้ศึกษาผลกระทบของค่าพารามิเตอร์ข้อสอบในกลุ่มอ้างอิง ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธีอัตราส่วนความควรจะเป็นโดยกำหนดเงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ 2 ขนาด คือ ขนาดเล็ก (0.25) และขนาดใหญ่ (0.5) โดยผลการศึกษาพบว่า ขนาดการทำหน้าที่ต่างกันของข้อสอบที่มีขนาดใหญ่และจำนวนกลุ่มตัวอย่างมีขนาดเพิ่มขึ้น มีอำนาจการทดสอบสูง และสามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดี นอกจากนี้ Atar and Kamata (2011, p. 40) ได้ศึกษาเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการทดสอบอัตราส่วนความควรจะเป็นกับวิธีการถดถอยโลจิสติก พบว่า วิธีการทดสอบอัตราส่วนความควรจะเป็น มีอำนาจการทดสอบการทำหน้าที่ต่างกันของข้อสอบสูงเมื่อขนาดของการทำหน้าที่ต่างกันของข้อสอบมีขนาดกลาง (0.43) และอำนาจการทดสอบการทำหน้าที่ต่างกันของข้อสอบจะเพิ่มขึ้นเมื่อขนาดของการทำหน้าที่ต่างกันเพิ่มขึ้น

2. จากการศึกษางานวิจัยในต่างประเทศปัจจุบันให้ความสนใจศึกษาความไม่แปรเปลี่ยนในการวัดข้ามกลุ่ม (Measurement invariance across groups) เพิ่มมากขึ้น ซึ่งคำจำกัดความของ “ความไม่แปรเปลี่ยนในการวัด (Measurement invariance) ก็คือ บุคคลที่มีความสามารถเดียวกัน

ในกลุ่มที่ต่างกันมีความน่าจะเป็นที่ตัวแปรจะเหมือนกันโดยไม่คำนึงถึงการเป็นสมาชิกของกลุ่ม (Schmitt & Kuljanin, 2008; Mellenbergh, 1989 อ้างถึงใน Kim & Yoon, 2011) ซึ่งหากกล่าวถึงกลุ่มตัวอย่าง 2 กลุ่ม กลุ่มตัวอย่าง 2 กลุ่มนี้ อาจจะแตกต่างกัน เช่น เพศ (ชาย-หญิง) ประเทศ (ฝรั่งเศส-ออสเตรเลีย) เป็นต้น และหากต้องการทราบว่า แบบทดสอบที่สร้างขึ้นมานั้นสามารถนำไปใช้และเปรียบเทียบกับกลุ่มตัวอย่าง 2 กลุ่ม ได้หรือไม่ ก็ต้องเปรียบเทียบคุณลักษณะแฝงบางประการของทั้ง 2 กลุ่ม โดยเน้นที่โครงสร้างองค์ประกอบ และความไม่แปรเปลี่ยนในการวัด ซึ่งบ่งบอกถึงโครงสร้างองค์ประกอบภายใน (Factor structure) ของทั้ง 2 กลุ่ม ประกอบด้วย ค่าน้ำหนักองค์ประกอบ (Factor loading) การแจกแจงค่าเฉลี่ย และความคลาดเคลื่อน ซึ่งการทดสอบความไม่แปรเปลี่ยนในการวัดเป็นการทดสอบความเท่ากันของโมเดลสมการโครงสร้างแบบข้ามกลุ่ม (Hortensius, 2012, pp. 7-8)

การศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ พอสรุปได้ว่า มีหลากหลายวิธีตามรูปแบบการตรวจให้คะแนนแบบ 2 ค่า หรือมากกว่า 2 ค่า ดังเช่น การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีอัตราส่วนความควรจะเป็น (Likelihood ratio test) วิธี Poly-SIBTEST ที่ขยายมาจากวิธี SIBTEST วิธีการถดถอยโลจิสติก วิธีแมนเทล-แฮนด์เซล เป็นต้น (Potenza & Dorans, 1995) ซึ่งในปัจจุบันมีการนำวิธีในกลุ่มของ SEM มาใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมากขึ้น ดังเช่น งานวิจัยของ Meade and Lautenschlager (2004), Kim and Yoon (2011) และ Muthén and Asparouhov (2014) เป็นต้น และในกระบวนการของการตรวจสอบ DIF พบปัจจัยที่เกี่ยวข้องหลายปัจจัย เช่น ปัจจัยความยาวของแบบสอบ ขนาดของกลุ่มตัวอย่าง ความแตกต่างของค่าเบี่ยงเบนมาตรฐาน ความแตกต่างของการแจกแจงข้อมูล และปฏิสัมพันธ์ของปัจจัยต่าง ๆ (Ackerman & Evans, 1992; Finch, 2005; Finch & French, 2007; Kim, 2010; Narayanan & Swaminathan, 1994; Prieto et al., 1997; Rogers & Swaminathan, 1993; Roussos & Stout, 1996; Shealy & Stout, 1993 cited in Atalay Kabasakal et al., 2014, p. 2187)

จากแนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ผู้วิจัยจึงมีความสนใจการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันสามวิธี คือ วิธี IRT LR (Likelihood ratio test) (Thissen et al., 1988) วิธี Poly-SIBTEST (Chang et al., 1996) และวิธี Multiple-groups CFA (Kim & Yoon, 2011) ภายใต้โมเดล GRM และเงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย ซึ่งประกอบด้วย ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง ว่าประสิทธิภาพในการตรวจสอบด้วยวิธีการตรวจสอบสามวิธีมีความแตกต่างกันหรือไม่ อย่างไร โดยพิจารณาจากอัตรา

ความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบการทำหน้าที่ต่างกันของข้อสอบ และใช้วิธีการศึกษาจำลองข้อมูล (Simulation) การทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าภายใต้โมเดล GRM ที่มีรายการคำตอบเป็น 1, 2, 3, 4 และ 5 และให้คะแนนคำตอบเป็น 0, 1, 2, 3 และ 4 ภายใต้เงื่อนไขปัจจัยที่แตกต่างกัน 3 ปัจจัย รวม 12 เงื่อนไข (2x2x3) ประกอบด้วย ขนาดการทำหน้าที่ต่างกันของข้อสอบ (ขนาดเล็ก 0.10 และขนาดกลาง 0.50) ความยาวของแบบสอบ 2 รูปแบบ (10 ข้อ และ 20 ข้อ) และขนาดของกลุ่มตัวอย่าง 3 ขนาด (200 คน, 500 คน และ 1,000 คน)

วัตถุประสงค์ของการวิจัย

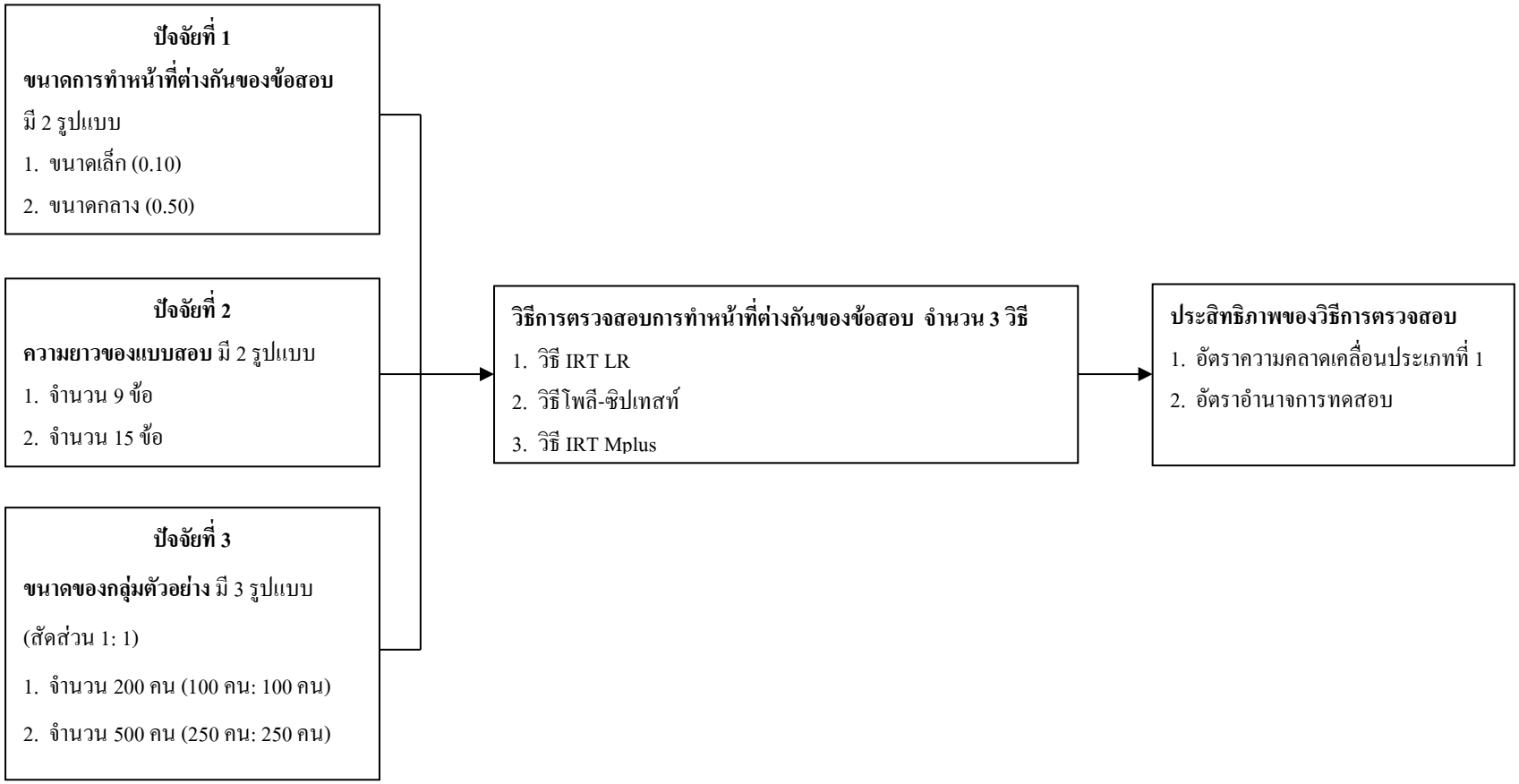
1. เพื่อศึกษาประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบการสนองข้อสอบแบบมิติเดียว ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย คือ ขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง

1.1 เพื่อศึกษาอัตราความคลาดเคลื่อนประเภทที่ 1 (Type I error rate) ของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA

1.2 เพื่อศึกษาอัตราอำนาจการทดสอบ (Power rate) ของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA

2. เพื่อเปรียบเทียบอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบสนองแบบสอบมิติเดียว ด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสามวิธี ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย คือ ขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง

กรอบแนวคิดในการวิจัย



ภาพที่ 1 กรอบแนวคิดในการวิจัย กรณีการจำลองข้อมูลภายใต้ทฤษฎีการตอบสนองข้อสอบแบบมิติเดียว

ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

1. ทำให้ทราบถึงประสิทธิภาพของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่เหมาะสมกับข้อมูลการตอบสนองข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในแบบทดสอบมิติเดียว ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย

2. เป็นแนวทางในการเลือกใช้วิธีการตรวจสอบให้เหมาะสมกับลักษณะของข้อสอบทางการศึกษาและจิตวิทยา

3. ทำให้ทราบถึงข้อจำกัดของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย ได้แก่ ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง

4. ทำให้ทราบถึงปัจจัยที่มีผลต่อวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสำหรับการประกอบการตัดสินข้อดีและข้อจำกัดในการเลือกใช้วิธีการตรวจสอบที่ได้สารสนเทศสูงสุด

5. เป็นแนวทางให้กับนักวิชาการทางด้านการศึกษาในการตัดสินใจเลือกใช้วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบภายใต้สถานการณ์การสอบใบบริบทจริงได้อย่างเหมาะสม

ขอบเขตของการวิจัย

การศึกษาในครั้งนี้มีขอบเขตของการศึกษา ดังนี้

1. การศึกษาในครั้งนี้เป็นการศึกษาโดยการใช้ข้อมูลจำลอง (Simulation data) ภายใต้ทฤษฎีการตอบสนองข้อสอบแบบมิติเดียว (Unidimensional item response theory) และการทำหน้าที่ต่างกันของข้อสอบไม่เป็นรูปแบบเดียวกัน (Nonuniform DIF) โดยจำลองข้อมูลการตอบข้อสอบที่มีโครงสร้างวัดความสามารถแบบมิติเดียวที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า (Polytomous item) ในโมเดล GRM โดยรูปแบบของแบบทดสอบ 1 ฉบับ ประกอบด้วย ข้อสอบจำนวน 9 ข้อ และ 15 ข้อ และข้อสอบทุกข้อมีรูปแบบรายการคำตอบแบบ 5 ตัวเลือก ซึ่งมีรายการคำตอบเป็น 0, 1, 2, 3 และ 4 โดยใช้การจำลองผลการตอบข้อสอบภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย คือ ขนาดการทำหน้าที่ต่างกันของข้อสอบ 2 รูปแบบ ความยาวของแบบสอบ 2 รูปแบบ และขนาดของกลุ่มตัวอย่าง 3 ขนาด ในสัดส่วน 1: 1 รวมข้อมูลเงื่อนไขที่ศึกษาทั้งสิ้น 12 เงื่อนไข (2x2x3) และในแต่ละเงื่อนไขจำลองข้อมูลวนซ้ำ 100 รอบ

2. ตัวแปรที่ศึกษา ประกอบด้วย

2.1 ตัวแปรอิสระ มี 3 ตัวแปร ดังนี้

2.1.1 ขนาดการทำหน้าที่ต่างกันของข้อสอบ มี 2 ขนาด

2.1.1.1 ขนาดเล็ก (0.10)

2.1.1.2 ขนาดกลาง (0.50)

2.1.2 ความยาวของแบบสอบ มี 2 รูปแบบ

2.1.2.1 จำนวน 9 ข้อ

2.1.2.2 จำนวน 15 ข้อ

2.1.3 ขนาดกลุ่มตัวอย่าง มี 3 ขนาด สัดส่วน 1: 1

2.1.3.1 จำนวน 200 คน (100 คน: 100 คน)

2.1.3.2 จำนวน 500 คน (250 คน: 250 คน)

2.1.3.3 จำนวน 1,000 คน (500 คน: 500 คน)

2.2 ตัวแปรตาม มี 2 ตัวแปร ดังนี้

2.2.1 อัตราความคลาดเคลื่อนประเภทที่ 1 (Type I error rate) ของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

2.2.2 อัตราอำนาจการทดสอบ ข้อสอบ (Power Rate) ของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

นิยามศัพท์เฉพาะ

การทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) หมายถึง ข้อสอบหรือข้อคำถามที่ทำให้ผู้สอบที่มีความสามารถหรือคุณลักษณะที่มุ่งวัดเท่ากัน มาจากต่างกลุ่มกัน มีโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกัน

ข้อสอบที่มีการตรวจให้คะแนนแบบหลายค่า หมายถึง ข้อสอบหรือข้อคำถามที่มีการกำหนดลำดับขั้นของการให้คะแนนผลการตอบข้อสอบหรือข้อคำถามเป็นจำนวนเต็มหลายระดับคะแนน ซึ่งการศึกษาครั้งนี้กำหนดให้ข้อสอบทุกข้อมีตัวเลือกให้เลือกตอบจำนวน 5 ตัวเลือก และมีรายการคำตอบเป็น 0, 1, 2, 3 และ 4

แบบทดสอบมิติเดียว (Unidimensional test) หมายถึง แบบทดสอบที่วัดคุณลักษณะแฝงของผู้สอบเพียง 1 คุณลักษณะ โดยการจำลองแบบทดสอบออกเป็น 2 ฉบับ และมีจำนวนข้อสอบที่แตกต่างกัน คือ ฉบับที่หนึ่ง เป็นแบบทดสอบที่มีข้อสอบ จำนวน 9 ข้อ และอีกฉบับหนึ่งเป็นแบบทดสอบที่มีข้อสอบ จำนวน 15 ข้อ

กลุ่มเปรียบเทียบ (Focal group: กลุ่ม F) หมายถึง กลุ่มผู้สอบที่คาดว่าจะเป็นกลุ่มที่เสียประโยชน์จากการตอบข้อสอบเมื่อข้อสอบทำหน้าที่ต่างกัน โดยมีความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องต่ำกว่าผู้สอบในกลุ่มอ้างอิงที่มีความสามารถเท่ากัน

กลุ่มอ้างอิง (Reference group: กลุ่ม R) หมายถึง กลุ่มผู้สอบที่คาดว่าจะเป็นกลุ่มที่ได้ประโยชน์จากการตอบข้อสอบเมื่อข้อสอบทำหน้าที่ต่างกัน โดยมีความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องสูงกว่าผู้สอบในกลุ่มเปรียบเทียบที่มีความสามารถเท่ากัน

ข้อสอบทำหน้าที่ต่างกันไม่เป็นรูปแบบเดียวกัน (Nonuniform DIF) หมายถึง ความแตกต่างของความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องระหว่างผู้สอบกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบมีค่าไม่เท่ากัน ทุกระดับของค่าพารามิเตอร์ความสามารถ θ_1 และ θ_2 การศึกษาครั้งนี้กำหนดให้ค่าพารามิเตอร์อำนาจจำแนกของข้อสอบระหว่างกลุ่มมีค่าแตกต่างกัน 0.10 และ 0.50 และค่าพารามิเตอร์รายการคำตอบระหว่างกลุ่มเท่ากัน

ขนาดการทำหน้าที่ต่างกันของข้อสอบ (Magnitude of DIF) หมายถึง ขนาดของการทำหน้าที่ต่างกันของข้อสอบอย่างมีนัยสำคัญทางสถิติ การวิจัยในครั้งนี้กำหนดใช้ขนาดเล็ก (0.10) และขนาดกลาง (0.50)

ขนาดของกลุ่มตัวอย่าง (Sample size) หมายถึง จำนวนผู้สอบของกลุ่มเปรียบเทียบและกลุ่มอ้างอิงที่ใช้ในการศึกษา การวิจัยในครั้งนี้กำหนดใช้จำนวนผู้สอบของกลุ่มเปรียบเทียบและอ้างอิง มีสัดส่วนในแต่ละกลุ่มที่เท่ากันคือ 1: 1 ($N_F : N_R$) และมีจำนวนผู้สอบ 3 ขนาด คือ 200 คน, 500 คน และ 1,000 คน นั่นคือ 100 คน: 100 คน, 250 คน: 250 คน และ 500 คน: 500 คน

ความยาวของแบบสอบ (Test length) หมายถึง จำนวนข้อสอบในแบบสอบทั้งฉบับ การวิจัยในครั้งนี้กำหนดความยาวของแบบสอบออกเป็น 2 ขนาด คือ 9 ข้อ และ 15 ข้อ

วิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ หมายถึง วิธีการที่ใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ การวิจัยในครั้งกำหนดวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ มี 3 วิธี คือ วิธีอัตราส่วนความควรจะเป็น (Likelihood ratio test: LR) วิธี Poly-DIBTEST และ วิธี Multiple-groups CFA

วิธีทดสอบอัตราส่วนความควรจะเป็น (Likelihood Ratio Test: LRT) หมายถึง วิธีการทดสอบอัตราส่วนความควรจะเป็นตามทฤษฎีการตอบสนองข้อสอบในรูปของล็อกเชิงเส้น โดยประมาณค่าพารามิเตอร์ข้อสอบด้วยวิธีความควรจะเป็นสูงสุด (Maximum Likelihood: ML) โดยการทดสอบอัตราส่วนความควรจะเป็นเพื่อทดสอบนัยสำคัญของการทำหน้าที่ต่างกันของข้อสอบ (De Ayala, 2013)

วิธี (Poly-SIBTEST หมายถึง วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในกลุ่ม วิธี IRT ที่พัฒนาโดย Chang et al. (1996) เพื่อใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนแบบหลายค่า ในการศึกษาครั้งนี้นำมาตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่วัดความสามารถมิติเดียวและให้คะแนนแบบหลายค่า

วิธี Multiple-Groups CFA หมายถึง วิธีการที่ใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกรณีกลุ่มตัวอย่างมีมากกว่า 1 กลุ่ม เพื่อทดสอบความไม่แปรเปลี่ยนของค่าพารามิเตอร์ ได้แก่ ค่า Factor loading, ค่า Thresholds และค่า Residual variances โดยอยู่บนพื้นฐานของการวิเคราะห์องค์ประกอบเชิงยืนยันของข้อสอบที่มีการให้คะแนนแบบหลายค่า (5 ตัวเลือก)

ประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ หมายถึง ความถูกต้องของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA โดยพิจารณาจากอัตราความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของการทำหน้าที่ต่างกันของข้อสอบ

อัตราความคลาดเคลื่อนประเภทที่ 1 (Type I error rate) หมายถึง จำนวนข้อสอบที่ตรวจสอบผิดพลาดว่า ทำหน้าที่ต่างกัน (False positive: α) ทั้งที่ในความเป็นจริงแล้วข้อสอบทำหน้าที่ไม่ต่างกัน ต่อจำนวนข้อสอบที่ทำหน้าที่ไม่ต่างกันทั้งหมดในแบบสอบ โดยคำนวณเป็นร้อยละ

อัตราอำนาจการทดสอบ (Power rate) หมายถึง จำนวนข้อสอบที่ตรวจสอบได้ถูกต้องว่าทำหน้าที่ต่างกัน ต่อจำนวนข้อสอบที่ทำหน้าที่ต่างกันทั้งหมดในแบบสอบ โดยคำนวณเป็นร้อยละ

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

การศึกษาเรื่อง การเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA ผู้วิจัยได้ศึกษาแนวคิดทฤษฎี ตลอดจนเอกสารและงานวิจัยที่เกี่ยวข้อง โดยนำเสนอเป็น 3 ตอน ดังต่อไปนี้

ตอนที่ 1 ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT)

ตอนที่ 2 การทำหน้าที่ต่างกันของข้อสอบ

ตอนที่ 3 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ และงานวิจัยที่เกี่ยวข้อง

ตอนที่ 1 ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT)

ทฤษฎีการตอบสนองข้อสอบ (IRT) เป็นทฤษฎีที่มีแนวคิดว่า ความน่าจะเป็นของการตอบข้อสอบได้ถูกต้องขึ้นอยู่กับความสามารถจริงของผู้ตอบ และคุณลักษณะของข้อสอบที่ประกอบด้วย พารามิเตอร์ความยาก อำนาจจำแนก และโอกาสการเดาข้อสอบได้ถูก ซึ่งสามารถแสดงด้วยโมเดลการตอบสนองข้อสอบ อาจเป็นโมเดล 1 พารามิเตอร์ โมเดล 2 พารามิเตอร์ หรือโมเดล 3 พารามิเตอร์ โดยถือว่า ค่าพารามิเตอร์ของข้อสอบและความสามารถจริงของผู้สอบมีความสัมพันธ์กัน ดังนั้น การประมาณค่าพารามิเตอร์ของข้อสอบ ได้แก่ ค่าความยาก ค่าอำนาจจำแนก และค่าความน่าจะเป็นในการเดาข้อสอบได้ถูก จะต้องกระทำพร้อม ๆ ไปกับการประมาณค่าความสามารถจริงของผู้สอบ จึงจะทำให้ได้ค่าพารามิเตอร์ที่มีนัยทั่วไป มีความน่าเชื่อถือและไม่แปรผันไปตามความสามารถของกลุ่มผู้สอบ นอกจากนี้ ทฤษฎี IRT สามารถวิเคราะห์ความคลาดเคลื่อนในการวัดข้อสอบแต่ละข้อและแบบสอบทั้งฉบับ จำแนกตามระดับความสามารถจริงของผู้สอบ และไม่มีข้อจำกัดว่าแบบสอบต้องเป็นแบบสอบคู่ขนาน (ศิริชัย กาญจนวาสี, 2555, หน้า 7) โดยสรุปคือ ทฤษฎี IRT เป็นทฤษฎีการวัดที่อธิบายความสัมพันธ์ระหว่างความสามารถที่มีอยู่ภายในบุคคล (Latent trait or ability) กับผลการตอบข้อสอบหรือข้อคำถาม โดยใช้โค้งลักษณะข้อสอบ (Item Characteristic Curve: ICC) ที่มีการกำหนดลักษณะของข้อสอบด้วยพารามิเตอร์ความยาก (b) อำนาจจำแนก (a) และโอกาสการเดาข้อสอบถูก (c) ที่เป็นฟังก์ชันทางคณิตศาสตร์ คือ ฟังก์ชันโลจิส (Logistic function) หรือฟังก์ชันปกติสะสม (Normal ogive function) โค้งลักษณะข้อสอบมีหลายลักษณะ ขึ้นอยู่กับโมเดล (Model) หรือแบบจำลองที่ใช้

อธิบายความสัมพันธ์ โมเดลที่นิยมใช้คือ โมเดลแบบหนึ่งพารามิเตอร์ (One-parameter model) โมเดลแบบสองพารามิเตอร์ (Two-parameter model) และโมเดลแบบสามพารามิเตอร์ (Three-parameter model)

ความสำคัญของทฤษฎีการทดสอบและแบบจำลอง

Hambleton and Jones (2012) กล่าวว่า ทฤษฎีการทดสอบและแบบจำลองที่สัมพันธ์กันเป็นสิ่งสำคัญในการศึกษาการวัดทางการศึกษาและทางจิตวิทยา เพราะทั้งสองสิ่งเป็นกรอบแนวคิดในการพิจารณาปัญหาทางด้านเนื้อหาและเทคนิค ซึ่งเนื้อหาที่สำคัญมากที่สุดอย่างหนึ่งก็คือความคลาดเคลื่อนในการวัด โดยทฤษฎีหรือแบบจำลองที่ดีจะทำให้เข้าใจถึงความคลาดเคลื่อนในการวัด ได้แก่ 1) การประมาณค่าพารามิเตอร์ความสามารถของผู้สอบและผลของความคลาดเคลื่อนที่อาจจะทำให้เกิดขึ้นได้น้อยที่สุด (เช่น ความยาวของแบบสอบ เป็นต้น) 2) ความสัมพันธ์ระหว่างตัวแปร และ 3) ผลของคะแนนจริงหรือคะแนนความสามารถ และความน่าเชื่อถือ ความแตกต่างของทฤษฎีกับแบบจำลองจะช่วยจัดการกับความคลาดเคลื่อนที่เกิดขึ้นได้ เช่น ความคลาดเคลื่อนที่อาจจะเกิดจากการแจกแจงแบบปกติในโมเดล หรือความคลาดเคลื่อนอาจจะไม่เกิดจากการแจกแจงแต่อาจจะเกิดจากสิ่งอื่น โดยในหนึ่งโมเดล ขนาดของความคลาดเคลื่อนในการวัดอาจจะมีความคงที่ข้ามกลุ่มของมาตรวัดคะแนนของแบบสอบ (เช่น ความคลาดเคลื่อนมาตรฐานของการวัด) นอกจากนี้ ขนาดของความคลาดเคลื่อนอาจจะสัมพันธ์กับคะแนนจริงของผู้สอบ ทั้งนี้ ข้อกำหนดต่างๆ ที่เกี่ยวกับความคลาดเคลื่อนในแบบจำลองจะมีผลอย่างมากต่อคะแนนความคลาดเคลื่อนในการประมาณค่าและการรายงานผล นอกจากนี้ ทฤษฎีการทดสอบหรือแบบจำลองที่ดีจะสามารถเป็นกรอบอ้างอิงการออกแบบแบบสอบหรือแก้ไขปัญหาในทางปฏิบัติได้ และแบบจำลองการสอบที่ดีอาจจะกำหนดความสัมพันธ์ระหว่างข้อสอบกับคะแนนความสามารถของแบบสอบได้อย่างแม่นยำเพื่อให้การออกแบบแบบสอบเป็นไปด้วยความรอบคอบเพื่อให้ได้คะแนนการแจกแจงของแบบสอบและขนาดของความคลาดเคลื่อนที่สามารถยอมรับได้

ตารางที่ 1 ความแตกต่างระหว่างทฤษฎีการทดสอบแบบดั้งเดิม (Classical Test Theory: CTT) กับทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) และโมเดลที่เกี่ยวข้อง

ประเด็น	ทฤษฎี CTT	ทฤษฎี IRT
โมเดล	โมเดลเชิงเส้น (Linear)	โมเดลไม่เชิงเส้น (Nonlinear)
ระดับ	แบบสอบ	ข้อสอบ
ข้อตกลงเบื้องต้น	มีจุดอ่อน	มีจุดแข็ง
ความสัมพันธ์ระหว่างข้อสอบกับความสามารถของผู้สอบ	ไม่เฉพาะเจาะจง	ฟังก์ชันคุณลักษณะของข้อสอบ (Item characteristic functions)
ความสามารถของผู้สอบ	คะแนนของแบบสอบหรือการประมาณค่าคะแนนจริงที่อยู่บนพื้นฐานของมาตรวัดคะแนนแบบสอบ	คะแนนความสามารถที่อยู่บนพื้นฐานของมาตรวัด $-\infty$ ถึง $+\infty$
สถิติความไม่แปรเปลี่ยนของข้อสอบและผู้สอบ	มีความแปรเปลี่ยนของค่าพารามิเตอร์ข้อสอบและผู้สอบ	มีความไม่แปรเปลี่ยนของค่าพารามิเตอร์ข้อสอบและผู้สอบ
สถิติข้อสอบ	ค่า p, r	ค่า b, a และ c (กรณี โมเดล 3 พารามิเตอร์) และฟังก์ชันสารสนเทศของข้อสอบ
ขนาดของกลุ่มตัวอย่าง (สำหรับการประมาณค่าพารามิเตอร์ข้อสอบ)	200 ถึง 500 (โดยทั่วไป)	อยู่บนพื้นฐานของโมเดล IRT แต่ถ้ากลุ่มตัวอย่างขนาดใหญ่โดยทั่วไปก็จะมากกว่า 500 คน ขึ้นไป

ข้อตกลงเบื้องต้นของทฤษฎีการตอบสนองข้อสอบ

ข้อตกลงเบื้องต้นที่สำคัญของ IRT มี 4 ประเด็นคือ ความเป็นเอกมิติ (Unidimensionality) ความเป็นอิสระ (Local independence) โมเดลการตอบสนองข้อสอบ (Item response models) และการสอบที่ไม่แข่งขันด้านเวลา (Nonspeeded test administration) (ศิริชัย กาญจนวาสี, 2555, หน้า 75-78) ดังนี้

1. ความเป็นเอกมิติ (Unidimensionality) ข้อคำถามหรือข้อสอบทุกข้อในเครื่องมือ/แบบสอบ มุ่งวัดเพียงคุณลักษณะเดียว หรือความสามารถเดียว (One ability) ตรวจสอบโดยใช้เทคนิคการวิเคราะห์ทางสถิติ ได้แก่ การวิเคราะห์ตัวประกอบ (Factor analysis) เพื่อคำนวณค่าไอเกน (Eigen value) และอัตราส่วนระหว่างค่าไอเกนของตัวประกอบแรกกับตัวประกอบถัดไป ถ้าอัตราส่วนสูงแสดงว่า เครื่องมือหรือแบบสอบวัดคุณลักษณะเด่นเดียว (Single dominant factor) หรือการวิเคราะห์องค์ประกอบเชิงยืนยัน (Confirmatory factor analysis) เพื่อยืนยันว่า เครื่องมือหรือแบบสอบมุ่งวัดเพียงคุณลักษณะเดียวหรือความสามารถเดียว

2. ความเป็นอิสระ (Local independence) เมื่อควบคุมอิทธิพลความสามารถ (θ) แล้วผลการตอบข้อสอบรายข้อ ไม่มีความสัมพันธ์กัน โมเดลการตอบสนองข้อสอบจึงมีเพียง θ ปัจจัยเดียว ที่มีอิทธิพลต่อผลการตอบข้อสอบรายข้อ ดังนั้น ความเป็นอิสระจึงจำแนกเป็น 2 ลักษณะ คือ ความอิสระระหว่างข้อสอบ และความอิสระระหว่างผู้สอบ ดังนี้

2.1 ความเป็นอิสระระหว่างข้อสอบ เมื่อสุ่มผู้สอบที่มีความสามารถ (θ) มา 1 คน ในการตอบข้อสอบ k ข้อ ให้ U_j เป็นผลการตอบหรือคะแนนข้อสอบข้อที่ j หลังจากควบคุม θ ของผู้สอบแล้ว คะแนนผลการตอบของผู้สอบในแต่ละข้อจะไม่สัมพันธ์กัน

$$\text{ดังนั้น } P(U_1, U_2, \ldots, U_k/\ \theta) = P(U_1/\ \theta)\, P(U_2/\ \theta) \ldots P(U_k/\ \theta)$$
$$= \prod_{j=1}^{k} P(U_j/\ \theta)$$

แต่ถ้าผลการตอบข้อสอบรายข้อของผู้สอบคนเดียวกันเป็นอิสระจากกัน ความน่าจะเป็นของแบบแผนการตอบข้อสอบ k ข้อ ของผู้สอบที่มีความสามารถ θ จะเท่ากับ ผลคูณระหว่างความน่าจะเป็นของผลการตอบข้อสอบแต่ละข้อ

2.2 ความเป็นอิสระระหว่างผู้สอบ เมื่อสุ่มข้อสอบขึ้นมา 1 ข้อ ในการตอบข้อสอบของผู้สอบ n คน ให้ U_i เป็นผลการตอบหรือคะแนนข้อสอบของผู้สอบคนที่ i หลังจากควบคุม θ ของผู้สอบแต่ละคนแล้ว คะแนนผลการตอบข้อสอบข้อนั้นของผู้สอบแต่ละคนไม่สัมพันธ์กัน

$$\text{ดังนั้น } P(U_1, U_2, \ldots, U_n/\ \theta) = P(U_1/\ \theta)\, P(U_2/\ \theta) \ldots P(U_n/\ \theta)$$
$$= \prod_{i=1}^{n} P(U_i/\ \theta)$$

ถ้าผลการตอบข้อสอบข้อเดียวกันของผู้สอบแต่ละคนเป็นอิสระจากกัน ความน่าจะเป็นของแบบแผนการตอบข้อสอบของผู้สอบ n คน จะเท่ากับ ผลคูณระหว่างความน่าจะเป็นของผลการตอบข้อสอบข้อนั้นของผู้สอบแต่ละคน

สำหรับการตรวจสอบความเป็นอิสระระหว่างข้อสอบและผู้สอบ สามารถพิจารณาได้จากเมตริกซ์ความแปรปรวนและความแปรปรวนร่วม (Variance-covaraince matrix) หรือ

เมตริกสหสัมพันธ์ (Correlation matrix) ของคะแนนคำตอบรายข้อ สำหรับกลุ่มผู้สอบที่มีช่วงความสามารถเท่ากัน ค่านอกแนวทแยงมุมควรมีค่าต่ำหรือเข้าใกล้ศูนย์

3. โมเดลการตอบสนองข้อสอบ (Item response models) เป็นโมเดลที่แสดงความสัมพันธ์ระหว่างระดับความสามารถ คุณลักษณะของข้อสอบ และโอกาสของการตอบข้อสอบได้ถูก ในรูปของโค้งลักษณะข้อสอบ (ICC) ที่มีลักษณะเป็นฟังก์ชันโลจิส (Logistic fuction) ที่แตกต่างกันตามจำนวนพารามิเตอร์ที่ใช้บรรยายลักษณะของข้อสอบ ได้แก่ โมเดลการตอบสนองข้อสอบแบบ 1-พารามิเตอร์, 2-พารามิเตอร์ และ 3-พารามิเตอร์ ดังนี้

โมเดลการตอบสนองข้อสอบแบบ 1-พารามิเตอร์ มีข้อตกลงเบื้องต้นว่า ข้อสอบแต่ละข้อมีพารามิเตอร์ c = 0 และพารามิเตอร์ a เท่ากัน แต่มีความแตกต่างกันเฉพาะพารามิเตอร์ b เท่านั้น จึงเหมาะกับข้อสอบอิงเกณฑ์ที่ไม่สลับซับซ้อน ข้อสอบที่ค่อนข้างเรียบง่ายสำหรับพัฒนาคลังข้อสอบที่มีความเป็นเอกพันธ์

โมเดลการตอบสนองข้อสอบแบบ 2-พารามิเตอร์ มีข้อตกลงเบื้องต้นว่า ข้อสอบแต่ละข้อมีพารามิเตอร์ c = 0 มีความแตกต่างกันของพารามิเตอร์ a และ b จึงเหมาะกับข้อสอบที่ต้องเติมคำตอบ หรือข้อสอบแบบเลือกตอบที่ไม่ยากมากนักและกลุ่มผู้สอบมีความพร้อมในการตอบ

โมเดลการตอบสนองข้อสอบแบบ 3-พารามิเตอร์ มีข้อตกลงเบื้องต้นว่า ข้อสอบแต่ละข้อมีความแตกต่างกันได้ทั้งพารามิเตอร์ a, b และ c จึงเหมาะกับข้อสอบแบบเลือกตอบทั่วไป ข้อสอบแบบหลายตัวเลือก เนื่องจากผู้สอบสามารถเดาคำตอบได้

การตรวจสอบความเหมาะสมของโมเดลการตอบสนองข้อสอบกับข้อมูล (Model-data fit) พิจารณาใน 2 ประเด็น คือ 1) ความไม่แปรเปลี่ยนของค่าประมาณความสามารถ (Invariance of ability parameter estimates) ซึ่งตรวจสอบได้โดยการเปรียบเทียบค่าประมาณความสามารถของผู้สอบที่ได้จากกลุ่มข้อสอบที่แตกต่างกัน เช่น กลุ่มข้อสอบยาก กลุ่มข้อสอบง่าย หรือกลุ่มข้อสอบจากคลังข้อสอบเดียวกันว่า มีความครอบคลุมเนื้อหาแตกต่างกัน เป็นต้น และค่าประมาณความสามารถจะถือว่าไม่แปรเปลี่ยน เมื่อความแตกต่างเกิดขึ้นไม่เกินความคลาดเคลื่อนมาตรฐานของการประมาณค่า และ 2) ความไม่แปรเปลี่ยนของค่าประมาณพารามิเตอร์ของข้อสอบ (Invariance of item parameter estimates) ตรวจสอบได้โดยเปรียบเทียบค่าประมาณพารามิเตอร์แต่ละตัวของข้อสอบที่ได้จากกลุ่มตัวอย่างผู้สอบหลายกลุ่ม เช่น กลุ่มผู้สอบชาย/ หญิง กลุ่มผู้สอบจำแนกตามภูมิภาค เป็นต้น ค่าประมาณพารามิเตอร์ของข้อสอบจะถือว่าไม่แปรเปลี่ยน เมื่อผลการพล๊อตกราฟออกมาเป็นเส้นตรง โดยมีการกระจายไม่แตกต่างจากผลที่ได้จากกลุ่มตัวอย่าง 2 กลุ่ม ซึ่งเป็นกลุ่มที่ทัดเทียมกัน

4. การสอบที่ไม่แข่งขันด้านเวลา (Nonspeeded test administration) ความเร็วในการตอบจะต้องไม่มีอิทธิพลต่อผลการสอบ การจัดการสอบจึงต้องไม่อยู่ในสถานการณ์ที่สอบแข่งขันกันด้วยเวลา การสอบจะต้องอยู่ในลักษณะที่ผู้สอบมีเวลาเพียงพอในการทำข้อสอบ (Power test administration) การตรวจสอบความเหมาะสมของมิติด้านเวลาในการดำเนินการสอบ พิจารณาได้จาก สัดส่วนหรือร้อยละของจำนวนผู้สอบที่ทำข้อสอบได้ครบทุกข้อ โดยผู้สอบส่วนใหญ่ (เช่น ร้อยละ 80 เป็นต้น) สามารถตอบข้อสอบได้ครบหรือเกือบครบทุกข้อ นอกจากนี้ควรพิจารณาเปรียบเทียบระหว่างความแปรปรวนของจำนวนข้อที่เว้น กับความแปรปรวนของจำนวนข้อที่ตอบผิด ถ้าอัตราส่วนของความแปรปรวนเข้าใกล้ศูนย์ แสดงว่า การจัดการสอบเป็นไปตามข้อตกลงเบื้องต้นในการเปรียบเทียบ

ค่าพารามิเตอร์ในทฤษฎีการตอบสนองข้อสอบ

ตามทฤษฎีการตอบสนองข้อสอบ ประกอบด้วย ค่าพารามิเตอร์ของผู้สอบ ค่าพารามิเตอร์ของข้อสอบ และค่าคงที่ (ศิริชัย กาญจนวาสี, 2555) ดังนี้

ค่าพารามิเตอร์ของผู้สอบ คือ ค่าระดับความสามารถของผู้สอบ (θ) ซึ่งประมาณได้จากโมเดลการตอบสนองข้อสอบ โดยนิยมปรับให้เป็นคะแนนมาตรฐานที่มีค่าเฉลี่เป็น 0 และมีพิสัยอยู่ระหว่าง $-\infty$ ถึง ∞ แต่ผลการวิเคราะห์ส่วนใหญ่ให้ค่าอยู่ในช่วง -3 ถึง +3 โดย $P_i(\theta)$ จะหมายถึงความน่าจะเป็นที่ผู้ตอบที่มีความสามารถระดับ θ จะตอบข้อสอบข้อที่ i ได้ถูกต้อง

ค่าพารามิเตอร์ของข้อสอบ ประกอบด้วย ค่าพารามิเตอร์ความยาก (b_i) ค่าพารามิเตอร์อำนาจจำแนก (a_i) และค่าพารามิเตอร์การเดา (c_i) โดยมีรายละเอียดดังนี้

1. ค่าพารามิเตอร์ความยาก (b_i) คือ ค่าพารามิเตอร์ความยาก (Difficulty parameter) ของข้อสอบข้อที่ i โดยตำแหน่งของโค้งบนสเกลของความสามารถ (θ) ที่ทำให้มีโอกาสตอบข้อสอบได้ถูกต้องเท่ากับ $\frac{1+c_i}{2}$ สำหรับโมเดล 1-พารามิเตอร์ และ 2-พารามิเตอร์ ค่า $P_i(\theta)$ จะเท่ากับ 0.50 สำหรับโมเดล 3-พารามิเตอร์ ค่า $P_i(\theta)$ จะเท่ากับ $\frac{1+c_i}{2}$

ในทางทฤษฎีค่าพารามิเตอร์ความยาก (b_i) มีค่าอยู่ระหว่าง $-\infty$ ถึง $+\infty$ แต่ในทางปฏิบัติจะใช้ข้อสอบที่มีค่า b_i อยู่ระหว่าง -2.50 ถึง +2.50 โดยค่า b_i ที่อยู่ใกล้ -2.50 แสดงว่าเป็นข้อสอบที่ง่าย และค่า b_i ที่อยู่ใกล้ +2.50 แสดงว่า เป็นข้อสอบที่ยาก

2. ค่าพารามิเตอร์อำนาจจำแนก (a_i) คือ ค่าพารามิเตอร์อำนาจจำแนก (Discrimination parameter) ของข้อสอบข้อที่ i โดยการจำแนกค่าความต่างของค่า $P_i(\theta)$ ระหว่างผู้สอบที่มีความสามารถ $\leq \theta$ กับ $> \theta$ และมีค่าเป็นสัดส่วนโดยตรงของค่าความชันของ ICC ที่ตำแหน่ง b_i ค่า a_i ที่มีค่าสูงแสดงถึงการจำแนกผู้สอบที่มีความสามารถแตกต่างกันได้ดี และในทางทฤษฎี

ค่า a_i มีค่าอยู่ระหว่าง $-\infty$ ถึง $+\infty$ โดยมีค่าเป็นบวก ซึ่งตามปกติมีค่าไม่เกิน +2.50 และในทางปฏิบัตินิยมใช้ข้อสอบที่มีค่า a_i อยู่ระหว่าง +0.50 ถึง +2.50

3. ค่าพารามิเตอร์โอกาสในการเดาข้อสอบได้ถูก (c_i) คือ ค่าพารามิเตอร์โอกาสในการเดาข้อสอบได้ถูก (Guessing parameter) โดยโอกาสในการตอบถูกของผู้สอบที่มีความสามารถต่ำ เป็นค่ากำกับต่ำสุด (Lower asseymtote) และในทางทฤษฎี มีค่าอยู่ระหว่าง 0 ถึง 1 โดยทั่วไปนิยมใช้ข้อสอบที่มีค่า c_i ไม่เกิน 0.30 ตามปกติควรมีค่าต่ำกว่าโอกาสในการตอบถูกโดยการเดาตามทฤษฎี CTT

ค่าคงที่ ประกอบด้วย ค่าคงที่ e คือ ค่าคงที่ของลอการิทึมธรรมชาติ (Natural log) โดยมีค่าเท่ากับ 2.71828 และค่าคงที่ D คือ ค่าองค์ประกอบของการปรับสเกล (Scaling factor) เป็นค่าการปรับสเกลเพื่อทำให้ฟังก์ชัน Logistic กับ ฟังก์ชัน Normal ogive ใกล้เคียงกัน หรือมีค่าประมาณ θ ต่างกันไม่เกิน 0.01

โมเดลการตอบสนองข้อสอบใน IRT

1. โมเดลการตอบสนองข้อสอบแบบตรวจให้คะแนน 2 ค่า (Dichotomous IRT models)

โมเดลการตอบสนองข้อสอบแบบตรวจให้คะแนน 2 ค่า เป็นโมเดลการตอบสนองข้อสอบที่ใช้กับการตรวจให้คะแนนข้อสอบรายข้อแบบ 2 ค่า เช่น ข้อสอบหรือข้อคำถามที่ตรวจให้คะแนนแบบ 0, 1 (ตอบผิดได้ 0, ตอบถูกได้ 1) แบบถูก/ ผิด ใช่/ ไม่ใช่ เป็นต้น โดยโมเดลการตอบสนองข้อสอบแบบตรวจให้คะแนน 2 ค่า ที่นิยมใช้คือ โมเดลโลจิส (Logistic model) ที่มีวิธีคำนวณที่ง่ายและสะดวกว่าโมเดลปกติสะสม (Normal ogive model) มีฟังก์ชันทางคณิตศาสตร์ (ศิริชัย กาญจนวาสี, 2555) ดังนี้

ตารางที่ 2 ฟังก์ชันทางคณิตศาสตร์ของโมเดลการตอบสนองข้อสอบ (ศิริชัย กาญจนวาสี, 2555)

Models	Normal Ogive Function	Logistic Function
1-Parameter	$P_i(\theta) = \int_{-\infty}^{\theta-bi} \frac{1}{\sqrt{2\pi}} e^{-z^2/2dz}$	$P_i(\theta) = \frac{1}{1+e^{-(\theta-b_i)}}$
2-Parameter	$P_i(\theta) = \int_{-\infty}^{a_i(\theta-bi)} \frac{1}{\sqrt{2\pi}} e^{-z^2/2dz}$	$P_i(\theta) = \frac{1}{1+e^{-Da_i(\theta-b_i)}}$
3-Parameter	$P_i(\theta) = c_i + (1+c_i) \int_{-\infty}^{a_i(\theta-bi)} \frac{1}{\sqrt{2\pi}} e^{-z^2/2dz}$	$P_i(\theta) = c_i + \frac{1}{1+e^{-Da_i(\theta-b_i)}}$

Embreton and Reise (2000, p. 66-72) ได้กล่าวว่า ในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ตัวแปรแฝงที่มีเพียงหนึ่งคุณลักษณะเพียงพอที่จะอธิบายความแตกต่างของบุคคลได้ และมีความเหมาะสมสอดคล้องกับข้อมูลที่มีองค์ประกอบเดียว อย่างไรก็ตามโมเดลการตอบสนองข้อสอบแบบมิติเดียวไม่เหมาะสมกับข้อมูลที่มีลักษณะดังนี้ 1) ข้อมูลที่มีคุณลักษณะแฝงสองคุณลักษณะขึ้นไปที่มีผลกระทบต่อข้อคำถามแตกต่างกัน 2) ผู้สอบที่มีความแตกต่างกันอย่างเป็นระบบทั้งด้านกลยุทธวิธี โครงสร้างความรู้ หรือ การแปลความในการตอบข้อคำถาม ซึ่งคุณลักษณะตามข้อ 1) และ 2) ข้างต้น จะมีความเหมาะสมกับโมเดลการตอบสนองข้อสอบแบบหลายมิติ โดยอธิบายถึงโมเดลการตอบสนองข้อสอบแบบ 1 พารามิเตอร์, 2 พารามิเตอร์ และ 3 พารามิเตอร์ ไว้ดังนี้

1.1 โมเดลการตอบสนองข้อสอบแบบ 1 พารามิเตอร์ (1-Parameter model)

โมเดลการตอบสนองข้อสอบแบบ 1 พารามิเตอร์ หรือ โมเดล Rasch (Rasch model) เป็นโมเดลทำนายความน่าจะเป็นของผลการตอบขของผู้สอบคนที่ s ที่จะตอบข้อคำถามข้อที่ i ได้ถูกต้อง ($P(X_{is}=1)$) โดยมีสมการดังนี้

$$P(X_{is}=1|\theta_s,\beta_i) = \frac{\exp(\theta_s-\beta_i)}{1+\exp(\theta_s-\beta_i)} \quad (1)$$

โดย X_{is} หมายถึง ผลการตอบข้อคำถามของผู้ตอบคนที่ s ที่จะตอบข้อคำถามข้อที่ i

θ_s หมายถึง ระดับคุณลักษณะแฝงหรือความสามารถของผู้ตอบคนที่ s

β_i หมายถึง ค่าความยากของข้อคำถามข้อที่ i

เช่นเดียวกับสูตร

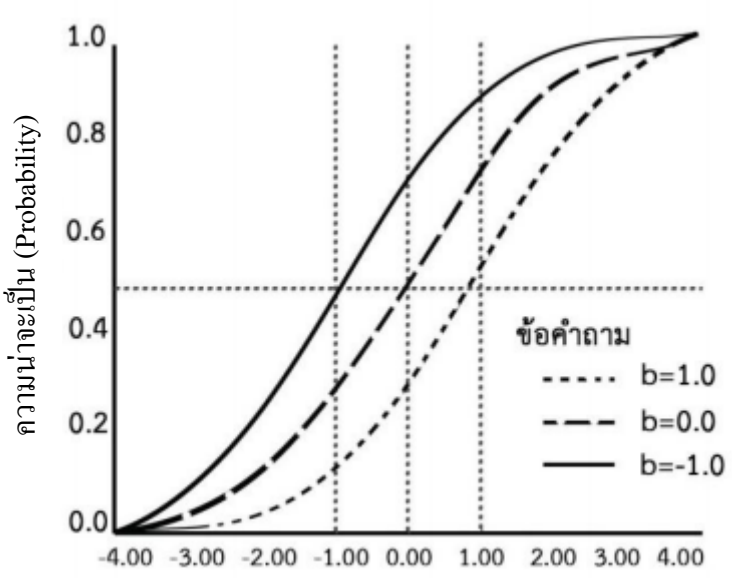
$$P_i(\theta) = \frac{1}{1+e^{-(\theta-b_i)}} \quad (2)$$

เมื่อ $P_i(\theta)$ = ความน่าจะเป็นที่ผู้ตอบซึ่งมีความสามารถ θ จะตอบข้อสอบข้อที่ i ได้ถูกต้อง

b_i = ค่าพารามิเตอร์ความยากของข้อสอบข้อที่ i ที่แสดงตำแหน่งของ ICC ณ จุด θ ที่มีโอกาสตอบข้อสอบถูก 0.50

e = ค่าคงที่ของลอการิทึมธรรมชาติ (Natural log) ซึ่งมีค่าเท่ากับ 2.718

ทั้งนี้ ค่าพารามิเตอร์ b_i มีค่าแปรเปลี่ยนตามลักษณะของข้อสอบแต่ละข้อ โดยค่าพารามิเตอร์ a_i มีค่าคงที่ และค่าพารามิเตอร์ c_i มีค่าเท่ากับ 0 และมีโค้งลักษณะข้อสอบดังภาพที่ 2



ภาพที่ 2 โค้งลักษณะข้อสอบของโมเดลโลจิสแบบ 1-พารามิเตอร์ (Embretson & Reise, 2000, p. 68)

1.2 โมเดลการตอบสนองข้อสอบแบบ 2 พารามิเตอร์ (2-Parameter model)

โมเดลการตอบสนองข้อสอบแบบ 2 พารามิเตอร์ มีค่าพารามิเตอร์ค่าอำนาจจำแนกของข้อสอบ (Item discrimination parameter) เพิ่มขึ้นอีก 1 พารามิเตอร์ ซึ่งความน่าจะเป็นของผู้ตอบคนที่ s ที่จะตอบข้อคำถามข้อที่ i โดยมีสมการดังนี้

$$P(X_{is}=1|\theta_s,\beta_i,\alpha_i)=\frac{\exp[\alpha_i(\theta_s-\beta_i)]}{1+\exp[\alpha_i(\theta_s-\beta_i)]} \quad (3)$$

โดย X_{is} หมายถึง ผลการตอบข้อคำถามของผู้ตอบคนที่ s ที่จะตอบข้อคำถามข้อที่ i

θ_s หมายถึง ระดับคุณลักษณะแฝงหรือความสามารถของผู้ตอบคนที่ s

β_i หมายถึง ค่าความยากของข้อคำถามข้อที่ i

α_i หมายถึง ค่าอำนาจจำแนกของข้อคำถามข้อที่ i

เช่นเดียวกันกับสูตร

$$P_i(\theta) = \frac{1}{1+e^{-Da_i(\theta-b_i)}} \quad (4)$$

เมื่อ $P_i(\theta)$ = ความน่าจะเป็นที่ผู้ตอบซึ่งมีความสามารถ θ จะตอบข้อสอบข้อที่ i ได้ถูกต้อง

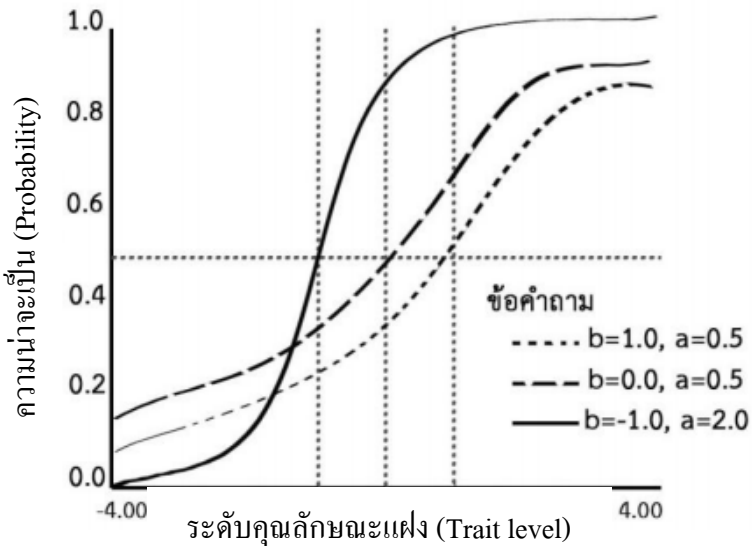
b_i = ค่าพารามิเตอร์ความยากของข้อสอบข้อที่ i ที่แสดงตำแหน่งของ ICC ณ จุด θ ที่มีโอกาสตอบข้อสอบถูก 0.50

a_i = ค่าพารามิเตอร์อำนาจจำแนกของข้อสอบข้อที่ i ซึ่งเป็นค่าความชันของ ICC ณ ตำแหน่ง b_i

e = ค่าคงที่ของลอการิทึมธรรมชาติ (Natural log) ซึ่งมีค่าเท่ากับ 2.718

D = ค่าองค์ประกอบของการปรับสเกล (Scaling factor) เป็นค่าการปรับสเกลเพื่อทำให้ฟังก์ชัน Logistic กับฟังก์ชัน Normal ogive ใกล้เคียงกัน หรือมีค่าประมาณ θ ต่างกันไม่เกิน 0.01 ซึ่งมีค่าเท่ากับ 1.70

ทั้งนี้ ค่าพารามิเตอร์ b_i และ a_i มีค่าแปรเปลี่ยนตามลักษณะของข้อสอบแต่ละข้อ สำหรับค่าพารามิเตอร์ $C_i = 0$ และมีโค้งลักษณะข้อสอบดังภาพที่ 3



ภาพที่ 3 โค้งลักษณะข้อสอบของโมเดลโลจิสแบบ 2-พารามิเตอร์ (Embretson & Reise,

2000, p. 71)

1.3 โมเดลการตอบสนองข้อสอบแบบ 3 พารามิเตอร์ (3-Parameter model)

โมเดลการตอบสนองข้อสอบแบบ 3 พารามิเตอร์ มีค่าพารามิเตอร์เพิ่มขึ้นมาอีกหนึ่งค่า โดยเป็นค่าพารามิเตอร์ที่แสดงให้เห็นถึงโค้งลักษณะข้อสอบว่า ไม่เข้าใกล้ศูนย์เมื่อข้อคำถามมีโอกาสการเดา เช่น ข้อคำถามแบบหลายตัวเลือก ซึ่งมีความน่าจะเป็นในการตอบข้อคำถามได้ถูกซึ่งอาจมีค่ามากกว่าศูนย์ แม้ผู้ตอบจะมีคุณลักษณะแฝงหรือความสามารถต่ำก็ตาม โดยมีสมการดังนี้

$$P(X_{is}=1|\theta_s,\beta_i,\alpha_i\gamma_i) = \gamma_i + (1-\gamma_i)\frac{\exp[\alpha_i(\theta_s-\beta_i)]}{1+\exp[\alpha_i(\theta_s-\beta_i)]} \quad (5)$$

โดย X_{is} หมายถึง ผลการตอบข้อคำถามของผู้ตอบคนที่ s ที่จะตอบข้อคำถามข้อที่ i

θ_s หมายถึง ระดับคุณลักษณะแฝงหรือความสามารถของผู้ตอบคนที่ s

β_i หมายถึง ค่าความยากของข้อคำถามข้อที่ i

α_i หมายถึง ค่าอำนาจจำแนกของข้อคำถามข้อที่ i

γ_i หมายถึง ค่าพารามิเตอร์โอกาสการเดาของข้อคำถามข้อที่ i

เช่นเดียวกันกับสูตร $$P_i(\theta) = c_i + \frac{1}{1+e^{-Da_i(\theta-b_i)}} \quad (6)$$

เมื่อ $P_i(\theta)$ = ความน่าจะเป็นที่ผู้ตอบซึ่งมีความสามารถ θ จะตอบข้อสอบข้อที่ i ได้ถูกต้อง

b_i = ค่าพารามิเตอร์ความยากของข้อสอบข้อที่ i ที่แสดงตำแหน่งของ ICC ณ จุด θ ที่มีโอกาสตอบข้อสอบถูก $\frac{1+c_i}{2}$

a_i = ค่าพารามิเตอร์อำนาจจำแนกของข้อสอบข้อที่ i ซึ่งเป็นค่าความชันของ ICC ณ ตำแหน่ง b_i

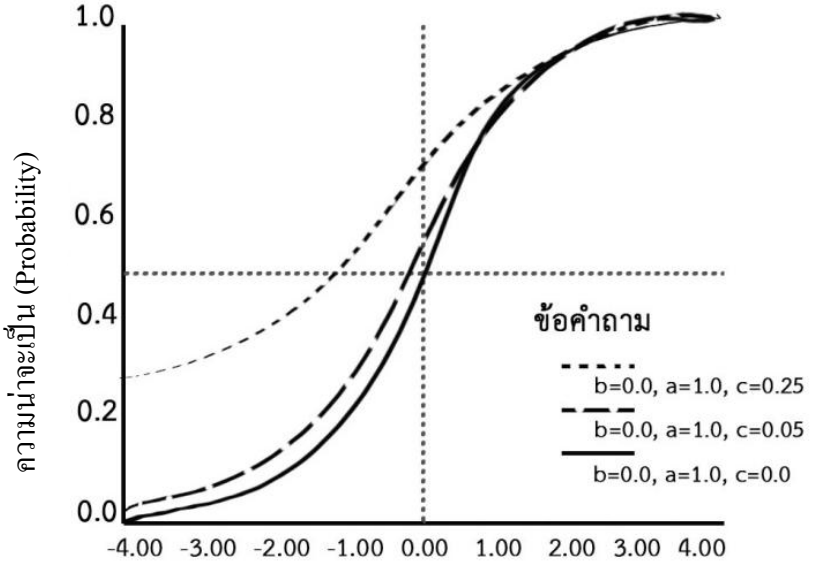
c_i = ค่าพารามิเตอร์โอกาสเดาข้อสอบได้ถูกต้อง

e = ค่าคงที่ของลอการิทึมธรรมชาติ (Natural log) ซึ่งมีค่าเท่ากับ 2.718

D = ค่าองค์ประกอบของการปรับสเกล (Scaling factor) เป็นค่าการปรับสเกลเพื่อทำให้ฟังก์ชัน Logistic กับฟังก์ชัน Normal ogive ใกล้เคียงกัน

หรือมีค่าประมาณ θ ต่างกันไม่เกิน 0.01 ซึ่งมีค่าเท่ากับ 1.70

ทั้งนี้ ค่าพารามิเตอร์ b_i, a_i และ c_i มีค่าแปรเปลี่ยนตามลักษณะของข้อสอบแต่ละข้อ และมีโค้งลักษณะข้อสอบดังภาพที่ 4



ภาพที่ 4 โค้งลักษณะข้อสอบของโมเดลโลจิสแบบ 3-พารามิเตอร์ (Emvretson & Reise, 2000, p. 72)

2. โมเดลการตอบสนองข้อสอบแบบตรวจให้คะแนนมากกว่า 2 ค่า (Polytomous IRT models)

โมเดลการตอบสนองข้อสอบแบบตรวจให้คะแนนมากกว่า 2 ค่า หรือโมเดลการตอบสนองข้อสอบแบบพหุวิภาค เป็นโมเดลความสัมพันธ์ที่ไม่ใช่เชิงเส้นตรงระหว่างความสามารถหรือคุณลักษณะของผู้ตอบ (θ) กับความน่าจะเป็นของการเลือกตอบแต่ละรายการคำตอบของข้อสอบหรือข้อคำถาม การประมาณค่าพารามิเตอร์สำหรับแต่ละรายการคำตอบของข้อสอบ หรือข้อคำถามนำไปสู่การคำนวณค่าฟังก์ชันสารสนเทศของข้อสอบ เมื่อนำมารวมกัน ณ ตำแหน่ง θ ทำให้ได้ค่าฟังก์ชันสารสนเทศของแบบสอบ และสามารถคำนวณค่าความคลาดเคลื่อนมาตรฐานของการประมาณค่า θ ได้ในลักษณะเดียวกันกับโมเดลการตอบสนองข้อสอบแบบตรวจให้คะแนน 2 ค่า หรือโมเดลการตอบสนองข้อสอบแบบทวิภาค ทั้งนี้ โมเดล

การตอบสนองข้อสอบแบบตรวจให้คะแนนมากกว่า 2 ค่า ที่พัฒนาบนพื้นฐานของราสช์โมเดล (Rasch model) หรือโมเดลแบบ 1-พารามิเตอร์ ได้แก่ Partial credit model (PCM) และ Rating scale model (RSM) ซึ่งเหมาะสำหรับใช้กับข้อสอบหรือข้อคำถามที่แต่ละข้อมีค่าอำนาจจำแนกเท่ากัน สำหรับ Graded-response model (GRM), Modified graded-response model (M-GRM), Generalized partial credit model (G-PCM) และ Normial response model (NRM) พัฒนาบนพื้นฐานของโมเดลแบบ 2-พารามิเตอร์ ที่เหมาะสำหรับข้อสอบหรือข้อคำถามที่มีค่าความยากและค่าอำนาจจำแนกแตกต่างกัน และการเลือกใช้โมเดลการตอบสนองข้อสอบแบบตรวจให้คะแนนมากกว่า 2 ค่า ควรพิจารณาใน 4 ประเด็นหลัก คือ

1) ปรัชญาความเชื่อเกี่ยวกับโมเดล และจุดมุ่งหมายการนำผลไปใช้ของผู้พัฒนาแบบสอบ

2) ควรใช้กลุ่มตัวอย่างที่มีความเป็นวิวิธพันธ์ (Heterogeneous sample) และขนาดกลุ่มตัวอย่างต้องใหญ่เพียงพอที่จะทำให้ความคลาดเคลื่อนมาตรฐานของการประมาณค่าอยู่ในระดับที่ยอมรับได้ตามเป้าหมายของการนำผลไปใช้

3) ควรเลือกใช้แบบแผนการตอบที่สะดวก และสามารถตรวจให้คะแนนได้ง่ายอย่างเป็นปรนัย

4) ข้อมูลที่นำมาวิเคราะห์จะต้องมีการตอบทุกข้อ และแต่ละข้อจะต้องมีการตอบทุกรายการ จึงจะทำให้สามารถประมาณค่าพารามิเตอร์ตามโมเดลที่เลือกใช้ได้

โมเดล Graded-Response (Graded-Response Model: GRM)

Samejima (1969, 1996 cited in Embreton & Reise, 2000, p. 97-102; ศิริชัย กาญจนวาสี, 2555) ได้พัฒนา Graded-Response Model (GRM) สำหรับใช้กับแบบสอบหรือแบบวัดที่แต่ละข้อคำถามมีรายการคำตอบแบบมาตรเรียงลำดับ (Ordered categorical responses) ซึ่งแต่ละข้ออาจมีรายการคำตอบที่แตกต่างกันได้ เป็นโมเดลการตอบสนองข้อสอบแบบ 2-พารามิเตอร์ และใช้หลักการคำนวณความน่าจะเป็นของการตอบแต่ละรายการคำตอบแบบ 2 ขั้นตอน (Indirect IRT model) โดยขั้นตอนแรก คำนวณค่าความชันร่วมของแต่ละข้อคำถาม จากนั้นจึงคำนวณค่าพารามิเตอร์ของแต่ละรายการคำตอบในแต่ละข้อคำถาม ใน GRM คำถามแต่ละข้อ (i) อธิบายได้ด้วย ความชันร่วมของข้อคำถาม (Common item slope parameter, α) และค่า Threshold ของแต่ละรายการคำตอบ (Category threshold, β_{ij}) เมื่อ j = 1, …, m_j โดย m_i คือ จำนวนของ Threshold ของข้อที่ $_i$ และจำนวนรายการคำตอบของข้อที่ i (K) = m_i + 1 การวิเคราะห์ตามโมเดล GRM มีเป้าหมายเพื่อประมาณค่า α_i และตำแหน่งของ β_i ของผู้ตอบที่มีคุณลักษณะ (θ) บนสเกลที่ต่อเนื่องกัน โดยมีสูตรดังนี้

$$P^{*}_{ix}(\theta) = \frac{\exp[\alpha_i(\theta-\beta_{ij})]}{1+\exp[\alpha_i(\theta-\beta_{ij})]} \quad (7)$$

เมื่อ $x = j = 1, \ldots, m_i$

$P^{*}_{ix}(\theta)$ = ความน่าจะเป็นที่ผู้ตอบซึ่งมีคุณลักษณะระดับ θ ตอบข้อ i ด้วยการเลือกรายการคำตอบที่ x เมื่อ $x = 1, 2, \ldots, m_i$

α_i = ค่าพารามิเตอร์ความชันร่วม (Slope parameter) ของข้อที่ i

β_i = ค่าพารามิเตอร์ Threshold ของแต่ละรายการคำตอบ (Threashold parameter) ของข้อที่ i

ตัวอย่างข้อคำถาม: ท่านมีความสำคัญต่อหน่วยงานของท่านเพียงใด

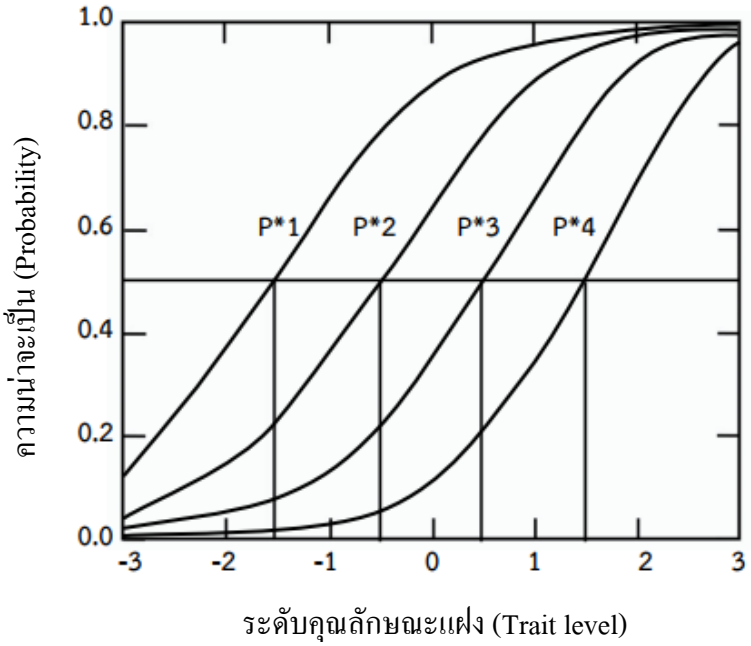
	น้อยที่สุด	น้อย	ปานกลาง	มาก	มากที่สุด
คะแนน (x)	0	1	2	3	4
	|----------|	|----------|	|----------|	|----------|	
Threshold		1	2	3	4

จากตัวอย่างข้อคำถามข้างต้น จะมีรายการคำตอบ (K) จำนวน 5 รายการคำตอบ ซึ่งแต่ละรายการคำตอบกำหนดให้คะแนนเป็น 0, 1, 2, 3 และ 4 คือ Categories 0, 1, 2, 3, 4 และมีค่า Threshold เท่ากับ 4 Theshold คือ 1, 2, 3 และ 4

ค่า α_i คล้ายกับค่าอำนาจจำแนกของข้อสอบตามทฤษฎีการทดสอบแบบดั้งเดิม แต่ไม่ควรพิจารณาโดยตรงว่า เป็นค่าอำนาจจำแนกของข้อสอบ เพราะการประเมินขาดความสามารถในการจำแนก จำเป็นต้องคำนวณจากค่าสารสนเทศของข้อสอบที่ระดับ θ ของผู้สอบ

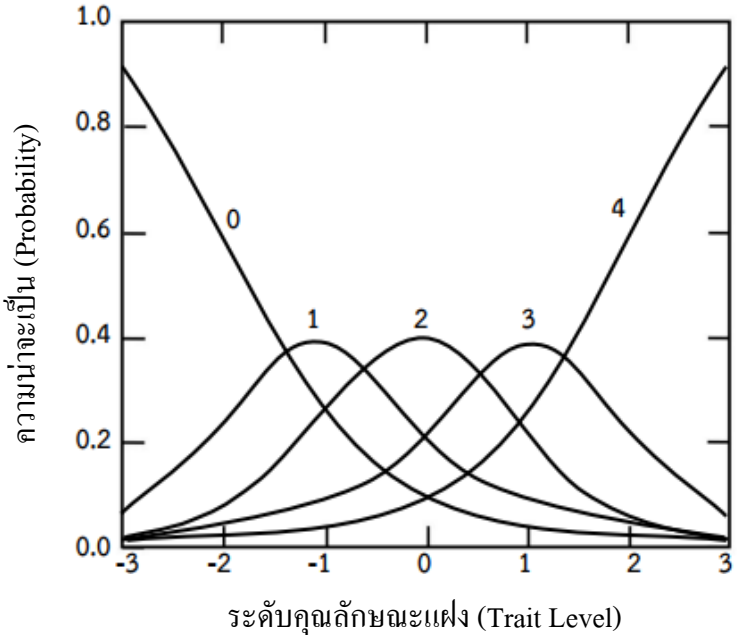
โค้งแสดงฟังก์ชันของ $P^{*}_{ix}(\theta)$ เรียกว่า โค้งลักษณะปฏิบัติการ (Operating Characteristic Curves: OCC) ต้องคำนวณแต่ละโค้งแยกระหว่างรายการคำตอบ ดังนั้น จึงต้องประมาณค่า β_{ij} โดย β_{ij} มีความหมายคล้ายเป็นระดับค่า θ ที่จำเป็นจะต้องมีเพื่อให้มีโอกาสตอบเหนือ Threshold j ด้วยความน่าจะเป็น 0.50 หรือ 50% และโค้งแสดงฟังก์ชันของความน่าจะเป็นในการเลือกรายการ

คำตอบต่าง ๆ ของผู้ที่มีคุณลักษณะ θ เรียกว่า โค้งการเลือกรายการคำตอบ (Category Response Curves: CRC) ลักษณะโค้ง CRC มีความเกี่ยวข้องสัมพันธ์กับโค้ง OCC และค่าพารามิเตอร์ความชันร่วมของข้อคำถาม (α_i)



ภาพที่ 5 โค้งลักษณะปฏิบัติการ (Operating characteristic curve) สำหรับข้อคำถามที่มี 5 รายการคำตอบ ตามแนวคิดโมเดล GRM (Embretson, & Reise, 2000, p. 100)

จากภาพที่ 5 แสดงให้เห็นโค้ง OCC จำนวน 4 โค้ง ของข้อคำถามที่มี 5 รายการคำตอบ ตามแนวคิดของ Graded-response model ซึ่งมีค่าพารามิเตอร์ความชันร่วมของข้อคำถาม $\alpha_i = 1.50$ และมีค่าพารามิเตอร์ Threshold ของแต่ละรายการคำตอบเป็น $\beta_{i1} = -1.50$, $\beta_{i2} = -0.50$, $\beta_{i3} = 0.50$ และ $\beta_{i4} = 1.50$ บนสเกลคุณลักษณะแฝง θ ที่แสดงถึงผู้ตอบมีความน่าจะเป็น 0.50 ในการเลือกตอบรายกรคำตอบที่ j = x หรือเหนือกว่า และเมื่อพิจารณาจากโค้ง P*1 ผู้ตอบที่มีค่า θ เท่ากับ -1.50 มีความน่าจะเป็น 0.50 ในการเลือกรายการคำตอบ 0 หรือ 1, 2, 3, 4 พอ ๆ กัน แต่ถ้าผู้ตอบที่มีค่า θ มากกว่า -1.50



ภาพที่ 6 โค้งการเลือกรายการคำตอบ (Category response curves) ของข้อคำถามที่มีตัวเลือกรายการคำตอบ 5 รายการ ตามแนวคิดของโมเดล GRM (Embretson & Reise, 2000, p. 101)

ค่าพารามิเตอร์ของข้อคำถามในโมเดล GRM เป็นค่าที่กำหนดรูปร่างและตำแหน่งของโค้ง CRC และ OCC ถ้าค่าพารามิเตอร์ความชันร่วมของข้อคำถาม (α_i) มีค่าสูงขึ้น จะทำให้โค้ง OCC มีความชันมากขึ้น เป็นผลให้ช่วงการกระจายของโค้ง CRC แคบลงและมียอดสูงขึ้น แสดงว่า รายการคำตอบสามารถจำแนกระหว่างระดับ θ ของผู้ตอบได้ดี ส่วนค่าพารามิเตอร์ Threshold ของรายการคำตอบ (β_{ij}) บอกถึงตำแหน่งของโค้ง OCC และตำแหน่งบริเวณที่พบกันของโค้ง CRC ของรายการคำตอบ 2 รายการที่อยู่ถัดกัน

ตอนที่ 2 การทำหน้าที่ต่างกันของข้อสอบ

ในการสร้างและการตรวจสอบคุณภาพของแบบสอบจะต้องคำนึงถึงคุณภาพด้านความตรงเป็นสำคัญ เนื่องจากความตรงเป็นหัวใจสำคัญของคุณภาพแบบสอบและแสดงถึง

ความสามารถในการวัดได้ถูกต้องแม่นยำ หากผลของการวัดได้ค่าที่ใกล้เคียงกับค่าคุณลักษณะที่แท้จริงเพียงใดก็ถือว่า การวัดนั้นมีความตรงมากขึ้นเพียงนั้น แต่การทำหน้าที่ต่างกันของข้อสอบก็เป็นลักษณะหนึ่งของการตรวจสอบคุณภาพด้านความตรงในประเด็นของความยุติธรรมของข้อสอบและแบบสอบ (Item and test unfairness) ธเกียรติกมล ทองงอก (2554) ได้สรุปเกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบแบบ 2 ค่า (Dichotomous) กับ แบบมากกว่า 2 ค่า (Polytomous) ไว้ว่า เป็นการศึกษาการทำหน้าที่ต่างกันของข้อสอบโดยพิจารณาในประเด็นรูปแบบการให้คะแนนในแบบสอบเป็นหลัก มีความยืดหยุ่นเกี่ยวกับรูปแบบของข้อสอบว่า เป็นการให้คะแนนแบบ 2 ค่า หรือแบบมากกว่า 2 ค่า ซึ่งวิธีการที่นำใช้ต้องมีข้อตกลงเบื้องต้นเกี่ยวข้องกับคะแนนของแบบสอบว่า สามารถนำวิธีการมาในการคำนวณได้ เช่น การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี MH, LR, SIBTEST, LRT, general IRT-LR, LLM และวิธีการผสม (Mixed effect models) โดยแรกเริ่มนั้นการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี MH ไม่ได้มีไว้สำหรับตรวจสอบข้อสอบที่ให้คะแนนแบบ 2 ค่า แต่มีการปรับขยายสูตรเพื่อศึกษาเกี่ยวกับข้อสอบที่ให้คะแนนแบบ 2 ค่า ต่อมาวิธี LR ได้ถูกนำมาปรับใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบ 2 ค่า

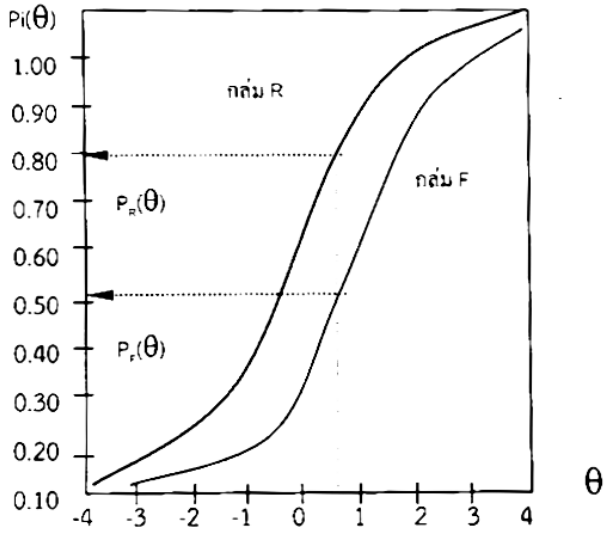
ทั้งนี้ การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยพื้นฐานแล้วมาจากทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) หรือการวิเคราะห์องค์ประกอบเชิงยืนยัน (Confirmatory Factor Analysis: CFA) ซึ่งการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบของโมเดล IRT มีอยู่หลายวิธี เช่น การทดสอบอัตราส่วนไลด์ลิฮู้ด (The IRT-based llihood ratio test) การเปรียบเทียบพื้นที่ใต้โค้งลักษณะข้อสอบ การทดสอบค่า Chi-Square นอกจากนี้ยังมีโมเดล MIMIC (Multiple causes model) และโมเดลโครงสร้างความแปรปรวนร่วมและค่าเฉลี่ย (The mean and covariance structure model: MACS) ซึ่งต่างก็เป็นพื้นฐานของการวิเคราะห์องค์ประกอบเชิงยืนยัน (CFA) ที่ใช้ในการตรวจสอบ DIF (Chang et al., 2015, p. 182)

รูปแบบของการทำหน้าที่ต่างกันของข้อสอบ

การทำหน้าที่ต่างกันของข้อสอบ เป็นการเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มผู้สอบอย่างน้อย 2 กลุ่มขึ้นไป โดยนิยมเปรียบเทียบ 2 กลุ่ม คือ กลุ่มแรก เรียกว่า กลุ่มเปรียบเทียบ (Focal group หรือกลุ่ม F) ซึ่งเป็นกลุ่มที่สนใจศึกษาและคาดว่าจะเป็นกลุ่มที่เสียเปรียบในการตอบข้อสอบ และกลุ่มที่สอง เรียกว่า กลุ่มอ้างอิง (Reference group หรือกลุ่ม R) เป็นกลุ่มที่คาดว่าจะได้เปรียบในการตอบข้อสอบได้ถูกต้อง โดยแบ่งลักษณะข้อสอบที่ทำหน้าที่ต่างกันได้ 2 รูปแบบ (ศิริชัย กาญจนวาสี, 2555) ดังนี้

1. ข้อสอบทำหน้าที่ต่างกันแบบเอกรูป (Uniform DIF) หรือข้อสอบทำหน้าที่ต่างกันที่เป็นรูปแบบเดียวกัน หมายถึง ข้อสอบที่ทำให้ผู้สอบกลุ่มหนึ่งมีโอกาสในการตอบข้อสอบ

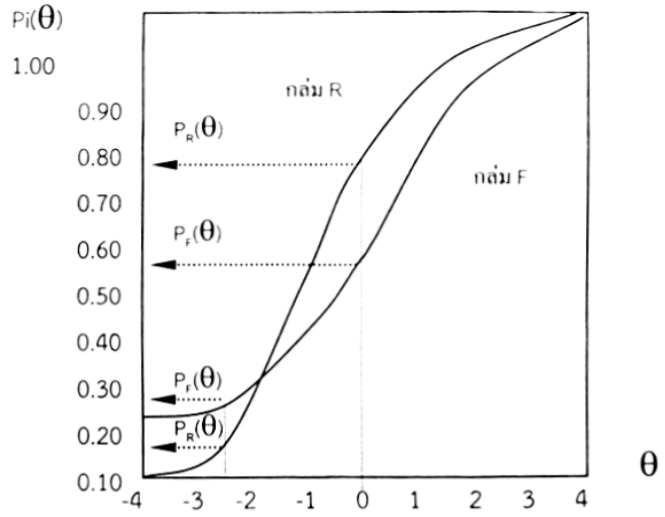
ถูกมากกว่าผู้สอบอีกกลุ่มหนึ่งอย่างสม่ำเสมอกันในทุกระดับความสามารถ และเมื่อพิจารณาโค้งคุณลักษณะข้อสอบของผู้สอบทั้ง 2 กลุ่ม จะพบว่า ไม่มีปฏิสัมพันธ์ระหว่างความสามารถของผู้สอบกับการเป็นสมาชิกของกลุ่ม (Group membership)



ภาพที่ 7 ข้อสอบทำหน้าที่กันแบบเอกรูป (Uniform DIF)

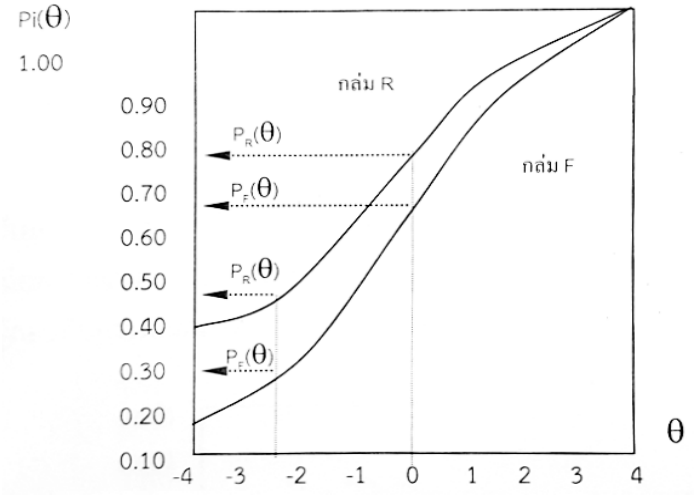
2. ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป (Nonuniform DIF) หรือข้อสอบทำหน้าที่ต่างกันที่ไม่เป็นรูปแบบเดียวกัน หมายถึง ข้อสอบที่ทำให้โอกาสในการตอบข้อสอบถูกของผู้สอบระหว่างกลุ่มแตกต่างกันอย่างไม่สม่ำเสมอกันในทุกระดับความสามารถ เมื่อพิจารณาโค้งคุณลักษณะข้อสอบของผู้สอบ 2 กลุ่ม พบว่ามีปฏิสัมพันธ์ร่วมกันระหว่างระดับความสามารถของผู้สอบ กับการเป็นสมาชิกของกลุ่ม เช่น ที่ระดับความสามารถหนึ่ง กลุ่มผู้สอบกลุ่ม R มีโอกาสในการตอบข้อสอบถูกมากกว่ากลุ่มผู้สอบกลุ่ม F แต่ที่ระดับความสามารถอีกระดับหนึ่ง กลุ่มผู้สอบกลุ่ม F มีโอกาสในการตอบข้อสอบถูกมากกว่ากลุ่มผู้สอบกลุ่ม R โดยจำแนกข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูป ได้ 2 ลักษณะ (Swaminathan & Rogers, 1990) ดังนี้

2.1 ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป มีปฏิสัมพันธ์ไม่เป็นลำดับ (Disordinal interaction) เป็นการทำหน้าที่ต่างกันของกลุ่มผู้สอบที่เกิดขึ้นเมื่อโค้งลักษณะข้อสอบตัดกันระหว่างช่วงความสามารถของผู้สอบ หรือเรียกว่า ข้อสอบทำหน้าที่ต่างกันแบบไม่มีทิศทาง (Non-unidirectional DIF) ดังภาพที่ 8



ภาพที่ 8 ข้อสอบทำหน้าที่ต่างกันแบบไม่มีทิศทาง (Non-unidirectional DIF) (ศิริชัย กาญจนวาสี, 2555, หน้า 118)

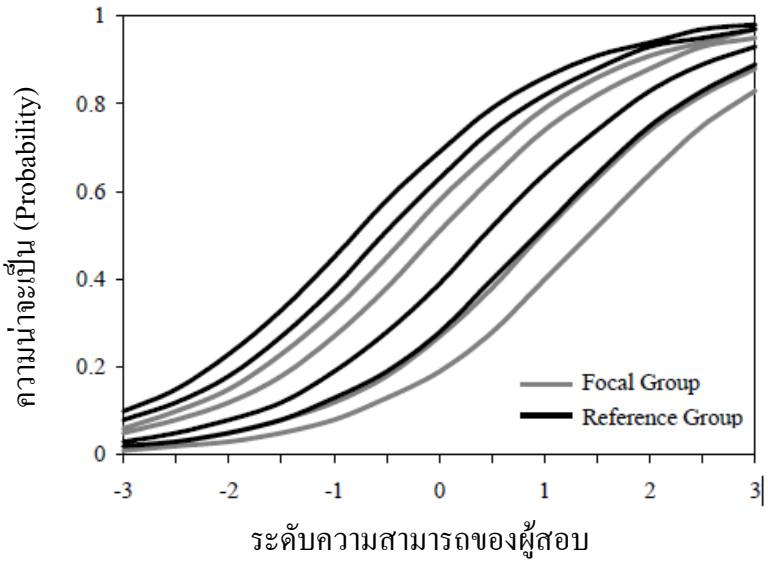
2.2 ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป มีปฏิสัมพันธ์เป็นลำดับ (Ordinal interaction) เป็นการทำหน้าที่ต่างกันของกลุ่มผู้สอบที่เกิดขึ้นเมื่อโค้งลักษณะข้อสอบต่างกันอย่างไม่สม่ำเสมอแบบไม่ตัดกัน หรืออาจตัดกันนอกช่วงความสามารถของผู้สอบตรงปลายสุดของช่วงความสามารถต่ำหรือสูง หรือเรียกว่า ข้อสอบทำหน้าที่ต่างกันแบบมีทิศทางเดียว (Unidirection DIF) ดังภาพที่ 9



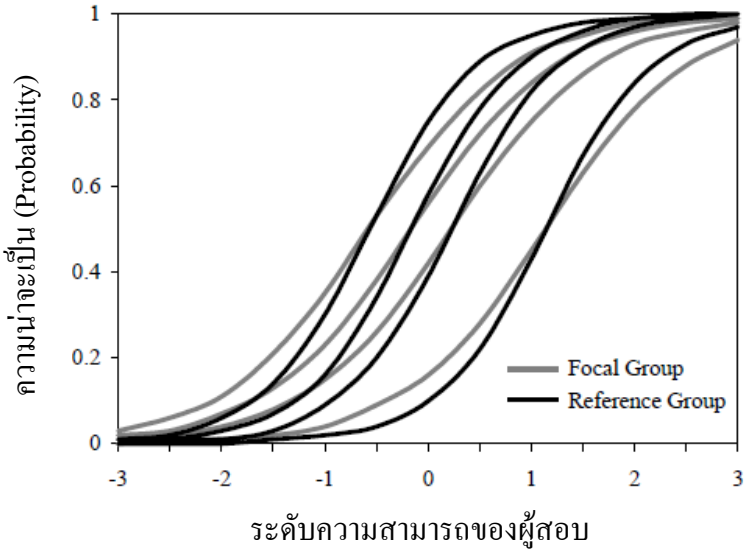
ภาพที่ 9 ข้อสอบทำหน้าที่ต่างกันแบบมีทิศทางเดียวกัน (Unidirectional DIF) (ศิริชัย กาญจนวาสี, 2555, หน้า 119)

ตามทฤษฎีการตอบสนองข้อสอบ สามารถพิจารณา “ปฏิสัมพันธ์” ได้จากความแตกต่างของค่าพารามิเตอร์อำนาจจำแนกของข้อสอบ ระหว่างผู้สอบกลุ่มย่อย 2 กลุ่ม ถ้าข้อสอบทำหน้าที่ต่างกันแบบเอกรูป แล้วโค้งลักษณะข้อสอบ (ICCs) ระหว่างผู้สอบกลุ่มย่อย 2 กลุ่ม จะขนานกัน หรือมีฟังก์ชันการตอบสนองข้อสอบ (Item Response Functions: IRFs) เหมือนกัน แต่ถ้าข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปแล้วโค้งลักษณะข้อสอบระหว่างผู้สอบกลุ่มย่อย 2 กลุ่ม จะไม่ขนานกัน หรือมีฟังก์ชันการตอบสนองข้อสอบต่างกัน ดังนั้นความแตกต่างระหว่างโค้งลักษณะข้อสอบทั้ง 2 แบบ จะบอกถึงขนาดและทิศทางของข้อสอบที่ทำหน้าที่ต่างกัน ซึ่งสามารถคำนวณได้โดยใช้สูตรการคำนวณพื้นที่ของ Raju (1990)

ทั้งนี้ French and Miller (1996, p. 135 อ้างถึงใน อรินทร์ น่วมถนอม, 2549) กล่าวว่า เงื่อนไขของ “ปฏิสัมพันธ์” ของรูปแบบการทำหน้าที่ต่างกันของข้อสอบที่มีการตรวจให้คะแนนแบบหลายค่า (Polytomous) จะมีความสมบูรณ์มากกว่า เพราะว่าไม่เพียงจะเกิดปฏิสัมพันธ์ระหว่างระดับความสามารถกับการเป็นสมาชิกของกลุ่มเท่านั้น แต่ยังมีตัวแปรที่สาม คือ ระดับคะแนนของข้อสอบเข้ามาเกี่ยวข้องด้วย ดังนั้นการทำหน้าที่เบี่ยงเบนของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าจะเกิดขึ้นภายในรายการคะแนนทั้งหมด (Scoe categories) และรูปแบบของข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าจะมีลักษณะคล้ายกับรูปแบบของข้อสอบที่ทำหน้าที่ต่างกันแบบสองค่า ซึ่งสามารถแสดงด้วยฟังก์ชันการตอบขอบรายการ (Boundary Response Functions: BRFs) ภายใต้โมเดล GRM ดังภาพที่ 10-11



ภาพที่ 10 ข้อสอบทำหน้าที่ต่างกันที่เป็นรูปแบบเดียวกัน (Uniform DIF) ภายใต้โมเดล GRM



ภาพที่ 11 ข้อสอบทำหน้าที่ต่างกันที่ไม่เป็นรูปแบบเดียวกัน (Nonuniform DIF) ภายใต้โมเดล GRM

หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF detecting) เป็นการเปรียบเทียบผลการตอบข้อสอบเป็นรายข้อระหว่างกลุ่มผู้สอบอย่างน้อย 2 กลุ่ม (กลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ) ที่มีความสามารถหลักที่มุ่งวัดเท่ากัน และในการเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ จำเป็นต้องจับคู่ (Matching) ผู้สอบตามความสามารถ ซึ่งเป็นเงื่อนไขสำคัญของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยเกณฑ์การจับคู่ (Matching criteria) ที่นิยมใช้มี 2 วิธี (ศิริชัย กาญจนวาสี, 2555) คือ

1. เกณฑ์ภายนอก (External criterion)

การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยเกณฑ์ภายนอกนี้ สามารถนำไปใช้ได้ทั้งข้อสอบรายข้อและแบบสอบทั้งฉบับ โดยการใช้คะแนนจากแบบสอบอื่นเป็นเกณฑ์ภายนอก แล้วใช้เทคนิคการวิเคราะห์ถดถอย (Regression analysis) เพื่อเปรียบเทียบเส้นกราฟความสัมพันธ์ระหว่างตัวแปรเกณฑ์ กับตัวแปรทำนายระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ หลักการนี้มีจุดมุ่งหมายเพื่อสร้างสมการทำนายตัวแปรเกณฑ์ ซึ่งเป็นคะแนนของแบบสอบอื่นจากตัวแปรทำนายซึ่งเป็นคะแนนรายข้อ หรือคะแนนแบบสอบระหว่างกลุ่มกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ ในการวิเคราะห์จะใช้คะแนนรายข้อเป็นตัวแปรทำนาย แต่ถ้าเป็นการวิเคราะห์การทำหน้าที่ต่างกันของแบบสอบ จะใช้คะแนนรวมของแบบสอบทั้งฉบับเป็นตัวแปรทำนาย สำหรับตัวแปรเกณฑ์ที่ใช้

เป็นเกณฑ์ภายนอก อาจใช้คะแนนรวมทั้งฉบับ หรือเกรดเฉลี่ย หรือผลสัมฤทธิ์จากงานที่เกี่ยวข้องกับผู้สอบ โดยสมการทำนายสำหรับกลุ่มอ้างอิง คือ $Y_i = A_R + B_R X_i$ และกลุ่มเปรียบเทียบ คือ $Y_i = A_F + B_F X_i$ โดยที่ Y_i เป็นคะแนนของตัวแปรเกณฑ์ภายนอก X_i เป็นคะแนนของตัวแปรทำนาย A เป็นค่าคงที่หรือค่าตัดแกน y (Intercept) และ B เป็นค่าความชัน (Slope) และจากฟังก์ชันการทำนายดังกล่าว สามารถเปรียบเทียบค่าตัดแกน (A) และค่าความชัน (B) ของเส้นกราฟระหว่างกลุ่ม R กับกลุ่ม F ได้ ถ้าเส้นกราฟมีค่าความชันหรือค่าตัดแกนแตกต่างสำหรับข้อสอบใด แสดงว่าข้อสอบหรือแบบสอบนั้น มีการทำหน้าที่ต่างกัน โดยเข้าข้างกลุ่มผู้สอบที่มีค่าตัดแกนหรือค่าความชันที่สูงกว่า

การใช้เกณฑ์ภายนอกมีข้อดี คือ เกณฑ์ที่ใช้มีความเป็นอิสระจากข้อสอบ และแบบสอบที่ต้องการตรวจสอบ แต่มีจุดอ่อน คือ ความเหมาะสมของเกณฑ์ที่จะนำมาใช้ ซึ่งในทางปฏิบัติยากที่จะหาเกณฑ์ภายนอกจากแบบสอบฉบับอื่นที่มีความตรงเชิงทำนาย และมีความยุติธรรมสำหรับกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ถ้าเกณฑ์ภายนอกขาดคุณสมบัติดังกล่าว จะทำให้ผลวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบหรือแบบทดสอบขาดความแม่นยำและสมบูรณ์

2. เกณฑ์ภายใน (Internal criterion)

การวิเคราะห์การทำหน้าที่ต่างกัน โดยใช้เกณฑ์ภายในเป็นการนำวิธีการทางสถิติมาตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ หรือแบบสอบ โดยเน้นการพิจารณาจากโครงสร้างภายในของแบบสอบเป็นหลัก ด้วยการวิเคราะห์ผลจากการตอบข้อสอบและความสามารถหรือคะแนนจริงของผู้สอบที่ได้จากแบบสอบฉบับนั้น เพื่อนำมาเปรียบผู้สอบจากกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบที่มีความสามารถหรือคะแนนจริงเท่ากัน ว่าจะมีจะผลการตอบหรือโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกันหรือไม่ เพื่อบ่งชี้ถึงการทำหน้าที่ต่างกันของข้อสอบ การวิเคราะห์โดยใช้เกณฑ์ภายในนิยมใช้ค่าสถิติต่าง ๆ เป็นตัวบ่งชี้ถึงการทำหน้าที่ต่างกันของข้อสอบ

2.1 การทดสอบปฏิสัมพันธ์ (Interaction) มีการใช้สถิติทดสอบเอฟ (F-test) จากการวิเคราะห์ความแปรปรวน (ANOVA) เพื่อทดสอบปฏิสัมพันธ์ระหว่างกลุ่มผู้สอบกับข้อสอบ ถ้าการทดสอบมีนัยสำคัญเป็นสัญญาณของการทำหน้าที่ต่างกันของข้อสอบ จากนั้นจึงทำการวิเคราะห์ต่อด้วยวิธีการ Post hoc เพื่อระบุข้อสอบที่มีผลต่อการเกิดปฏิสัมพันธ์ ซึ่งเป็นข้อที่ทำหน้าที่ต่างกัน โดยวิธีนี้มีข้อดีที่สามารถศึกษาผู้สอบหลาย ๆ ได้สะดวก แต่มีจุดอ่อนในการควบคุมกลุ่มต่างๆ ให้มีความสามารถที่ทัดเทียมกัน โดยอัตราความคลาดเคลื่อนประเภทที่ 1 จะสูงขึ้น เมื่อจำนวนข้อสอบเพิ่มมากขึ้น

2.2 การวัดความเบี่ยงเบนสัมพัทธ์ (Relative deviation) ในการคำนวณค่าความยากของข้อสอบ เมื่อคำนวณแยกระหว่างกลุ่ม และแปลงให้เป็นค่าความยากมาตรฐาน (Δ) สามารถนำมาพล๊อตเปรียบเทียบเป็นรายข้อคำถาม ถ้าข้อใดเบี่ยงเบนไปจากแกนหลักที่คาดหมาย หรือเบี่ยงเบนเกินจากความคลาดเคลื่อนมาตรฐานของค่าความยากที่กำหนด จะแสดงถึงการทำหน้าที่ต่างกันของข้อสอบ รวมทั้งสามารถคำนวณค่าสหสัมพันธ์ระหว่างค่าความยากรายข้อระหว่างกลุ่มเพื่อแสดงถึงการทำหน้าที่ต่างกันของข้อสอบ ถ้าค่าสหสัมพันธ์เข้าใกล้ 1.00 แสดงว่าค่าความยากสัมพัทธ์ของข้อสอบมีค่าใกล้เคียงกันระหว่างกลุ่ม วิธีการนี้มีข้อดีและข้อเสียคล้ายกับการทดสอบปฏิสัมพันธ์ นอกจากนี้ค่ายากของข้อสอบไม่ใช่ตัวแทนของค่าความยากจริงของข้อสอบ และได้รับอิทธิพลจากค่าแทรกซ้อนของค่าอำนาจจำแนกและความสามารถของผู้สอบ

2.3 การเปรียบเทียบน้ำหนักองค์ประกอบ (Factor loading)

การวิเคราะห์องค์ประกอบ (Factor analysis) เป็นเทคนิคทางสถิติที่นิยมใช้ในการตรวจสอบความตรงเชิงทฤษฎีหรือโครงสร้าง (Construct validity) และเมื่อนำมาใช้ในการวิเคราะห์โครงสร้างของแบบสอบแยกตามกลุ่มผู้สอบ ความไม่สอดคล้องกันระหว่างค่าน้ำหนักตัวประกอบบนคุณลักษณะที่มุ่งวัด หรือความแตกต่างของค่าเฉลี่ยคะแนนตัวประกอบ (Factor scores) ระหว่างกลุ่มผู้สอบ จะสะท้อนถึงการทำหน้าที่ต่างกันของข้อสอบและแบบสอบ การใช้เทคนิคการวิเคราะห์องต์ประกอบเชิงสำรวจ (Exporatory Factor Analysis: EFA) สำหรับศึกษาการทำหน้าที่ต่างกันของข้อสอบ มีจุดอ่อนในประเด็นความไม่สอดคล้องกันระหว่างค่าน้ำหนักตัวประกอบ ซึ่งอาจเกิดจากความแตกต่างของความสามารถระหว่างกลุ่ม แนวทางที่เหมาะสมจึงควรใช้เทคนิคการวิเคราะห์องค์ประกอบเชิงยืนยัน (Confirmatory Factor Analysis: CFA) โดยสามารถตรวจสอบความแตกต่างระหว่างกลุ่มผู้สอบ ด้านคุณลักษณะหรือความสามารถหลักและความสามารถรองได้ด้วย

2.4 การเปรียบเทียบโอกาสการตอบข้อสอบได้ถูก

การวิเคราะห์โอกาสการตอบข้อสอบได้ถูกของผู้สอบระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบที่มีระดับความสามารถเท่ากัน สำหรับการศึกษาการทำหน้าที่ต่างกันของข้อสอบเป็นแนวทางสำคัญที่นิยมใช้ มีการคำนวณค่าสถิติ 2 แนวทางคือ

2.4.1 เปรียบเทียบค่าสัดส่วนหรือความน่าจะเป็นในการตอบข้อสอบได้ถูกของผู้สอบต่างกลุ่มที่มีความสามารถเท่ากัน เช่น วิธีแมนเทล-แฮนส์เซล (MH) เป็นต้น

2.4.2 เปรียบเทียบค่าฟังก์ชันการตอบสนองข้อสอบ หรือโค้งลักษณะข้อสอบระหว่างกลุ่มผู้สอบที่มีระดับความสามารถเท่ากัน โดยเป็นวิธีที่อยู่บนพื้นฐานของทฤษฎี IRT เช่น วิธีวัดความแตกต่างของพื้นที่ วิธีวัดความแตกต่างของค่าพารามิเตอร์ความยาก วิธีการทดสอบ

ไค-สแควร์ของลอร์ด (Lord's χ^2-test) เป็นต้น วิธีการนี้มีข้อดีคือ การคำนวณค่าสถิติของข้อสอบมีความน่าเชื่อถือ มีกลไกควบคุมความสามารถของผู้สอบโดยการจับคู่กลุ่มความสามารถ เพื่อทำการเปรียบเทียบ ณ ตำแหน่งต่าง ๆ ที่มีความสามารถเท่ากัน แต่มีข้อจำกัดด้านความซับซ้อนของแนวคิดพื้นฐาน และการวิเคราะห์มีความจำเป็นต้องใช้โปรแกรมคอมพิวเตอร์โดยเฉพาะ

ตอนที่ 3 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ และงานวิจัยที่เกี่ยวข้อง

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ สามารถจำแนกตามลักษณะของเกณฑ์ประกอบด้วย ลักษณะการตรวจให้คะแนน (แบบสองค่าและแบบหลายค่า) มิติของตัวแปรเกณฑ์ (กลุ่มวิธีที่ใช้คะแนนสังเกตได้และกลุ่มวิธีที่ใช้คะแนนของตัวแปรแฝง) มิติลักษณะของสถิติวิเคราะห์ (กลุ่มที่ใช้สถิติพาราเมตริกและกลุ่มที่ใช้สถิตินันพาราเมตริก) (ธเกียรติกมล ทองงอก, 2554) โดยมีรายละเอียดดังนี้

1. จำแนกตามลักษณะการตรวจให้คะแนน

1.1 กลุ่มวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการตรวจให้คะแนนแบบสองค่า (Dichotomous DIF method) หรือการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการตรวจให้คะแนนแบบสองค่า (Dichotomous DIF procedure) แบบสอบที่มีลักษณะของการตรวจให้คะแนนแบบนี้ ได้แก่ แบบสอบชนอดเลือกตอบที่มีการให้คะแนนในการตอบถูกเป็น 1 คะแนน ในขณะที่ตอบผิดได้ 0 คะแนน การนำเสนอการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบสอบถูกกำหนดอย่างจริงจัง เพื่อให้เกิดความยุติธรรมในการใช้แบบสอบ และมีความตรงต่อการแสดงความหมายที่แฝงอยู่ของคะแนนเหล่านั้น งานวิจัยที่เกี่ยวข้องกับการทำหน้าที่ต่างกันของข้อสอบมีความหลากหลายมากขึ้น ซึ่งเดิมเน้นไปที่ข้อสอบที่มีการตรวจให้คะแนนแบบสองค่า แต่ในช่วงระยะเวลาที่ผ่านมามีความพยายามพัฒนาแนวทางเลือกใหม่ของวิธีการวัดที่ช่วยจุดประกายให้เดประเด็นที่น่าสนใจ และยังมีการทำหน้าที่ต่างกันของข้อสอบชนิดอื่นที่นอกเหนือจากข้อสอบที่มีการตรวจให้คะแนนแบบทวิภาค

1.2 กลุ่มวิธีการตรวจสอบการทำน้าที่ต่างกันของข้อสอบที่มีการตรวจให้คะแนนแบบหลายค่า (Polytomously methods) เช่น ข้อสอบวัดภาคปฏิบัติ (Perfoemance assessment) ข้อสอบความเรียง (Essay items) การตัดสินคุณภาพของแฟ้มสะสมผลงาน (Portfolio assessment) ข้อสอบที่วัดการอ่าน (Reading item) และข้อสอบที่วัดการเขียน (Writing item) รวมไปถึงข้อสอบปลายเปิด (Open-ended item) เป็นต้น

2. จำแนกตามมิติของตัวแปรเกณฑ์

2.1 กลุ่มวิธีที่ใช้คะแนนสังเกตได้ (Observed score) ค่าพารามิเตอร์แปรเปลี่ยนไปตามกลุ่มผู้สอบ วิธีในกลุ่มนี้มักวิเคราะห์ตามทฤษฎีการทดสอบแบบดั้งเดิม และเรียกกลุ่มที่ไม่ใช้ทฤษฎีการตอบสนองข้อสอบ (Non-IRT approach) ใช้คะแนนรวมของผู้สอบเป็นเกณฑ์การจับคู่ของกลุ่มผู้สอบ วิธีการตรวจสอบที่สำคัญในกลุ่มนี้ ได้แก่ การวิเคราะห์ความแปรปรวน (ANOVA) การวิเคราะห์การถดถอยโลจิสติกพหุวิภาค (Polytomous logistic regression) วิธีแมนเทล-แฮนส์เซลทั่วไป (Gerneral mantel-haenszel) และวิธีดัชนีมาตรฐานพหุวิภาค (Polytomous standardization)

2.2 กลุ่มวิธีที่ใช้คะแนนของคุณลักษณะหรือตัวแปรแฝง (Latent variable) โดยวิเคราะห์อยู่บนพื้นฐานของทฤษฎีการตอบสนองข้อสอบ ตัวแปรแฝงหรือตัวแปรคุณลักษณะดังกล่าวจะถูกใช้เป็นเกณฑ์การจับคู่กลุ่มผู้สอบและค่าพารามิเตอร์คงที่ไม่ว่าจะเป็นกลุ่มผู้สอบใด วิธีการตรวจสอบที่สำคัญในกลุ่มนี้ ได้แก่ วิธีการใช้คะแนนบางส่วนทั่วไป (Generalized Partial Credit Model: GPCM) วิธีอัตราส่วนความควรจะเป็นในรูปทั่วไป (General IRT likelihood ratio) วิธีการให้คะแนนบางส่วน (Partial Credit Model: PCM) และวิธีซิปเทสท์แบบหลายค่า (Polytomous SIBTEST)

3. จำแนกตามมิติลักษณะของสถิติวิเคราะห์

3.1 กลุ่มที่ใช้สถิติพาราเมตริก (Parametric approach) การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบมีข้อตกลงเบื้องต้นของโมเดลที่มุ่งเน้นอธิบายความสัมพันธ์ระหว่างคะแนนของข้อสอบและการจับคู่ตัวแปร

3.2 กลุ่มที่ใช้สถิตินันพาราเมตริก (Nonparametric approach) การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบไม่มีข้อตกลงเบื้องต้นของโมเดล และมีความหลากหลายวิธีการทางสถิติที่ถูกพัฒนามาใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบสองค่าและแบบหลายค่า

การศึกษาจำลองข้อมูลในทฤษฎีการตอบสนองข้อสอบ (Monte carlo studies in item response theory)

การศึกษาจำลองข้อมูลที่เริ่มใช้ในทฤษฎี IRT เพื่อให้ข้อมูลเกี่ยวกับความตรงของวิธีการที่สามารถนำไปประยุกต์ใช้กับข้อมูลจริง เช่น ข้อมูลที่มีจำนวนผู้สอบน้อย และข้อมูลแบบหลายมิติ เป็นต้น และจำนวนการทำวนซ้ำเป็นสิ่งที่ควรทำในการศึกษาจำลองข้อมูลนี้ ทฤษฎี IRT เป็นทฤษฎีที่มีความสำคัญและเป็นวิธีการที่รู้จักกันโดยทั่วไปสำหรับข้อมูลโมเดลการตอบสนองข้อสอบในการประยุกต์ใช้สำหรับปัญหาทางการวัดจำนวนมาก ซึ่งในช่วงปี ค.ศ. 1994-1995 มีบทความ

ในวารสาร Applied psychological measurement (APM) Psychometrika และ Journal of Educational Measurement (JEM) ที่ใช้เทคนิคการศึกษาจำลองข้อมูล (Monte Carlo: MC) (Harwell, Stone, Hsu, & Kirisci, 1996) และอริสพา เตหลิ่ม (2559, หน้า 79) ได้สรุปประเด็นขั้นตอนในการนำเทคนิค MC มาใช้ในการวิเคราะห์ IRT ตามผลการศึกษาของ Harwell et al. (1996) ดังนี้

1. กำหนดคำถามการวิจัยที่อธิบายถึงวัตถุประสงค์ที่เฉพาะเจาะจง เช่น เพื่อศึกษาความถูกต้องของการประมาณค่าพารามิเตอร์ของโมเดลการวิเคราะห์ข้อสอบ เมื่อมีจำนวนข้อสอบ ผู้สอบ และการแจกแจงเริ่มต้นของพารามิเตอร์ข้อสอบแตกต่างกัน เป็นต้น

2. กำหนดเงื่อนไขของตัวแปรต้นที่ส่งผลต่อตัวแปรตามได้ เช่น จำนวนผู้สอบและข้อสอบ (ตัวแปรต้น) มีผลกระทบต่อค่าพารามิเตอร์ของผู้สอบ (ตัวแปรตาม) หรือไม่

3. ออกแบบการทดลองให้มีความเหมาะสมกับวัตถุประสงค์การวิจัย

4. จำลองข้อมูลให้สอดคล้องกับเงื่อนไขของโมเดลการวิเคราะห์ข้อสอบ เช่น โมเดลแบบ 2 พารามิเตอร์

5. ประมาณค่าพารามิเตอร์โดยใช้ข้อมูลจากการจำลอง

6. เปรียบเทียบผลที่ได้จากการประมาณค่าตามเงื่อนไข โดยใช้ค่าสถิติต่าง ๆ ได้ เช่น ค่ามัธยฐาน ค่าความคลาดเคลื่อนมาตรฐานของการประมาณค่า เป็นต้น

7. กำหนดจำนวนการทำซ้ำ R รอบ

8. คำนวณค่าสถิติที่ได้จากการวิเคราะห์ R รอบ ทั้งสถิติเชิงบรรยายและสถิติเชิงอ้างอิง ซึ่งจะอ้างไปถึงการตอบคำถามการวิจัยและออกแบบการทดลอง

ทั้งนี้ ขั้นตอนการวิเคราะห์ข้อสอบใน IRT ด้วยเทคนิค MC มี 4 ขั้นตอน ดังนี้

1. การกำหนดปัญหาการวิจัย ขั้นตอนนี้เป็นขั้นตอนสำคัญของกระบวนการวิจัย การวิเคราะห์ข้อสอบด้วยเทคนิค MC ก็เช่นเดียวกัน ซึ่งจะเริ่มจากการกำหนดปัญหาและข้อคำถามการวิจัยก่อน จากนั้นตั้งสมมติฐานการทดสอบ และมีการวัดผลกระทบจากเงื่อนไขต่าง ๆ จากการจำลองข้อมูล โดยทั่วไปการกำหนดปัญหาการวิจัยมักมาจากการทบทวนเอกสารและรายงานการวิจัยที่เกี่ยวข้องกับเรื่องที่สนใจศึกษา มีการทดสอบสมมติฐานที่เป็นตัวแทนของคำถามการวิจัยและผลที่เกิดจากการวัดต้องไวกับตัวแปรที่ศึกษา

2. ออกแบบการศึกษาด้วยเทคนิค MC ต้องออกแบบให้สามารถตอบข้อคำถามของการวิจัยและสมมติฐานการวิจัยได้ โดยต้องมีการออกแบบทั้งตัวแปรต้นหรือตัวเป็นที่เป็นสาเหตุ และ ตัวแปรตามซึ่งเป็นผลกระทบที่เกิดขึ้นจากตัวแปรต้น นอกจากนี้ ยังต้องมีการประเมินผลทั้งในเรื่องของความตรงภายในและความตรงภายนอก ประเด็นที่เกี่ยวข้องกับการออกแบบของ

การศึกษา MC รวมถึงการเลือกตัวแปรต้น และตัวแปรตามขึ้นอยู่กับการออกแบบการทดลอง จำนวนการทำซ้ำ และโมเดลของ IRT เพื่อให้ผลการทดลองสามารถอ้างไปสู่ประชากรได้ ซึ่งการออกแบบการศึกษาด้วยเทคนิค MC มีขั้นตอนดังนี้

2.1 การกำหนดและระบุค่าของตัวแปรต้น คำถามการวิจัยเป็นสิ่งที่กำหนดตัวแปรต้น รวมทั้งเงื่อนไขในการจำลองข้อมูล โดยค่าของตัวแปรถือเป็นค่าคงที่ (Fixed effect) และไม่ต่อเนื่อง (Discrete) เช่น การศึกษาของ Harwell and Janosky (1991) กำหนดตัวแปรต้น ได้แก่ ขนาดตัวอย่าง (n) ความยาวข้อสอบ (L) ความแปรปรวนของกระจายก่อนหน้า (Prior distribution) ของ a_j หรือความแปรปรวนของตัวแปรต้น โดยที่ค่าของตัวแปรเหล่านี้เกิดจากคำถามวิจัย ซึ่งเน้นไปที่ขนาดตัวอย่างเล็กและความยาวข้อสอบน้อย

ค่าพารามิเตอร์เหล่านี้มักแสดงเป็นค่าระยะห่างเท่า ๆ กันในช่วงคงที่หรือเป็นค่าประมาณการจากการทดสอบเทียบกับการทดสอบก่อนหน้า ซึ่งถือเป็นค่าของตัวแปรคงที่ (Fixed effect) การสุ่มตัวอย่างของค่าอำนาจจำแนกและค่าความยากเป็นค่าของตัวแปรสุ่ม (Random effect) เนื่องจากถ้ามีการสุ่มค่าอำนาจจำแนกและค่าความยากมาใช้ในการศึกษาจะทำให้โมเดลในการศึกษาเป็นโมเดลแบบสุ่ม ที่สามารถอ้างอิงไปยังประชากรได้ แต่ถ้าโมเดลไม่ได้มีการสุ่มมาใช้ในการศึกษา โมเดลนั้นจะไม่สามารถอ้างอิงกลับไปยังประชากรได้ ดังนั้น ค่าพารามิเตอร์ในโมเดลควรเป็นตัวแทนของตัวแปรต้น อย่างไรก็ตาม นักวิจัยต้องพิจารณาความสัมพันธ์ระหว่างจำนวนของตัวแปรต้น ประสิทธิภาพของการศึกษาและผลการศึกษาด้วย ในขณะที่จำนวนของตัวแปรเพิ่มขึ้น จะทำให้ได้ความรู้ที่มากขึ้น แต่ก็จะใช้เวลาในการจำลองมากขึ้นตามไปด้วย

2.2 การเลือกแบบการทดลอง โดยทั่วไปตัวแปรอิสระมักจะเป็นตัวกำหนดแบบการทดลองที่เหมาะสม เช่น ถ้าจำนวนตัวแปรอิสระและระดับค่าของตัวแปรน้อย การใช้แบบการทดลองแบบแฟกตอเรียลจะมีความเหมาะสมกว่าแบบการทดลองอื่น ในการศึกษาด้วยเทคนิค MC มักจะใช้แบบการทดลอง โดยยึดเป้าหมายและวัตถุประสงค์ในการศึกษาเป็นหลัก ซึ่งการเลือกแบบทดลองอย่างระมัดระวัง จะช่วยในการวางแผนการวิเคราะห์ผลลัพธ์ได้อย่างถูกต้อง (Lewis & Ovar, 1989 cited in Harwell et al., 1996) ในงานวิจัยของ Harwell and Jamoskey (1991) มีตัวแปรจัดกระทำหรือตัวแปรต้นเป็น ขนาดกลุ่มตัวอย่าง ความยาวของแบบสอบ และความแปรปรวนของการแจกแจงค่าอำนาจจำแนก ใช้การออกแบบการทดลองแบบแฟกตอเรียลระหว่างกลุ่มตัวอย่างแบบสมบูรณ์ (Completely between-subjects factorial design) นอกจากนี้งานวิจัยของ Yen (1987 cited in Harwel et al., 1996) ได้เปรียบเทียบการใช้โปรแกรมการวิเคราะห์ข้อสอบระหว่างโปแกรม BILOG และโปรแกรม LOGIST โดยมีเงื่อนไขด้านความยาว ของข้อสอบและลักษณะการกระจายของค่าพารามิเตอร์ของผู้สอบ ออกแบบการทดลองแบบแฟกตอเรียล

โดยเรื่องของความยาวของข้อสอบและลักษณะการกระจายของค่าพารามิเตอร์ของผู้สอบ ได้ออกแบบการทดลองแฟกตอเรียลแบบ Between-subjects factor และส่วนการเปรียบเทียบระหว่างโปรแกรมคอมพิวเตอร์ทั้งสองได้ออกแบบการทดลองแฟกตอเรียลแบบ Within-subjects factor

2.3 การเลือกตัวแปรตาม ไม่เพียงต้องสอดคล้องกับคำถามการวิจัย แต่ต้องเลือกตัวแปรตามที่มีความไวต่อตัวแปรต้น เนื่องจากการเลือกตัวแปรที่มีความไวและควรใช้ประโยชน์ได้ถ้ามีการแปลงข้อมูลเป็นรูปแบบอื่น เช่น การหาค่า RMSE สามารถแปลงค่าเพื่อให้มีการแจกแจงแบบปกติ ทำให้สามารถนำไปสรุปอ้างอิงได้ นอกจากนี้ ถ้าเป็นเรื่องเกี่ยวกับการศึกษาเปรียบเทียบวิธีการในการศึกษา IRT สามารถใช้คุณลักษณะของแบบสอบ เช่น ความเป็นเอกมิติ การทำหน้าที่ต่างกันของข้อสอบหรือผู้สอบ เป็นตัวแปรตามในการศึกษาถึงผลกระทบของตัวแปรอิสระได้ สำหรับค่าความสัมพันธ์ของค่าจริงกับค่าที่ประมาณได้ ก็สามารถใช้ให้เป็นตัวแปรตามในการใช้เทคนิคมอนติคาร์โล เนื่องจากค่าสัมพันธ์นั้น ใช้เมตริกที่ต่างกันหาความสัมพันธ์ของตัวแปรอิสระและตัวแปรตามได้ เช่น ค่าความสัมพันธ์ระหว่างค่าพารามิเตอร์ที่ประมาณกับความคลาดเคลื่อนมาตรฐาน ส่วนข้อเสียก็คือ ความสัมพันธ์เหล่านี้สะท้อนความสัมพันธ์เฉพาะอันดับของตัวแปรและแสดงอิทธิพลของตัวแปรต้นเท่านั้น เช่น ค่าความสัมพันธ์ระหว่างค่าอำนาจจำแนกที่แท้จริงกับค่าที่ประมาณ มีค่าเท่ากับ 0.9 หมายความว่า โดยค่าเฉลี่ยของค่าอำนาจจำแนกที่แท้จริงนั้น อาจสูงกว่าค่าเฉลี่ยของค่าอำนาจจำแนกที่ประมาณได้ แต่ไม่รับรองว่าค่าอำนาจจำแนกที่แท้จริงกับค่าอำนาจจำแนกที่ประมาณจะใกล้เคียงกันหรือดีกว่ามากน้อยเพียงใด เช่น 0.8 กับ 0.9

2.4 การกำหนดจำนวนรอบ สำหรับการศึกษาด้วยเทคนิค MC เปรียบเทียบได้กับการกำหนดขนาดกลุ่มตัวอย่าง โดยมีเกณฑ์ที่ใช้ในการกำหนดประยุกต์ใช้มาจากการกำหนดขนาดกลุ่มตัวอย่างสำหรับในการศึกษาจากข้อมูลเชิงประจักษ์ ในการศึกษาการวิเคราะห์ข้อสอบด้วยทฤษฎีการตอบสนองข้อสอบ จำนวนรอบขึ้นอยู่กับวัตถุประสงค์ในการศึกษา โดยพิจารณาจากความต้องการในลดค่าความแปรปรวนของการสุ่มตัวอย่างในการประมาณค่าพารามิเตอร์ และความต้องของการทดสอบสถิติของผลการจำลองข้อมูลว่าอำนาจในการตรวจสอบผลกระทบที่สนใจเพียงพอหรือไม่

จำนวนรอบมีอิทธิพลโดยตรงกับความแม่นยำในการประมาณค่าพารามิเตอร์ ถ้ากลุ่มตัวอย่างมีขนาดใหญ่ (มีจำนวนรอบมาก) จะทำให้การประมาณค่าพารามิเตอร์ความแปรปรวนของการสุ่มตัวอย่างน้อย ดังนั้น ถ้านักวิจัยไม่กำหนดจำนวนรอบหรือกำหนดจำนวนรอบน้อยจะทำให้ความแปรปรวนของการสุ่มมีมาพอที่จะทำให้การประมาณค่าพารามิเตอร์มีความลำเอียงมาก ซึ่งจะส่งผลต่อความเที่ยงและความน่าเชื่อถือของผลการวิจัยที่ต่ำ จะเห็นได้ว่า จำนวนรอบมีความจำเป็น

ต่อความเที่ยงในการตรวจสอบผลกระทบต่อผลการจำลองข้อมูลมากในงานวิจัยที่มีลักษณะดังต่อไปนี้

2.4.1 งานวิจัยที่สนใจลักษณะการกระจายของการสุ่มตัวอย่างจากข้อมูลเชิงประจักษ์ เช่น งานวิจัยที่ต้องการตรวจสอบคุณสมบัติของค่าสถิติหรือทดสอบนัยสำคัญทางสถิติ

2.4.2 งานวิจัยที่ศึกษาค่ากลางของการวิจัยในระดับข้อสอบที่มีค่าความแปรปรวนของกลุ่มตัวอย่างมาก

2.4.3 งานวิจัยที่มีเป้าหมายในการศึกษาผลกระทบในบริบทที่ซับซ้อน เช่น ผลกระทบที่เกิดจากปฏิสัมพันธ์กับอิทธิหลัก

2.4.4 การวิเคราะห์ข้อมูลการจำลองข้อมูลตามทฤษฎีการตอบสนองข้อสอบควรมีการคำนวณวนซ้ำไม่ต่ำกว่า 25 รอบ

3. การเลือกใช้โปรแกรมการจำลองข้อมูลและการประมาณค่าพารามิเตอร์ อาจใช้โปรแกรมหลายโปรแกรมในการจำลองข้อมูลและวิเคราะห์ผลลัพธ์ ดังมีรายละเอียดดังนี้

3.1 การจำลองคำตอบ เริ่มจากกำหนดค่าเริ่มต้น (Seed) ให้กับตัวเลขสุ่ม ซึ่งจะแปลงเป็นค่าความน่าจะเป็นในการตอบคำถามได้ถูกต้อง

3.2 ตัวเลือกค่าเริ่มต้น ผู้วิจัยกำหนดค่าเริ่มต้นได้เอง โดยใช้การเติมช่องว่าง (Prompted) ซึ่งมีประโยชน์ คือ ง่ายต่อการจำลองคำตอบในข้อต่อ ๆ ไป และเป็นค่าที่สัมพันธ์กับความคลาดเคลื่อนในการสุ่ม ซึ่งหลีกเลี่ยงได้ยากในการจำลองข้อมูล ทั้งนี้ เทคนิคที่จะช่วยลดความแปรปรวนก็คือ การใช้ค่าพารามิเตอร์ข้อสอบและค่าเริ่มต้นร่วมกันทุกครั้ง เมื่อมีการจำลองข้อมูล เช่น การจำลองข้อสอบ 20 ข้อ และ 30 ข้อ ค่าเริ่มต้นที่ใช้ในการจำลองข้อมูล 20 ข้อ ควรจะเป็นค่าเริ่มต้นเดียวกันกับเมื่อจำลองข้อมูล 30 ข้อ เทคนิคอีกประการหนึ่ง คือ จำลองประชากรข้อมูลผลการตอบจำนวนมาก แล้วสุ่มคำตอบมาจากประชากรที่จำลองขึ้นหลาย ๆ รอบมากกว่าการใช้ค่าเริ่มต้นหลายตัว เพื่อให้ได้ชุดข้อมูลจำลองตามต้องการ การใช้โมเดลพารามิเตอร์ที่ต่างกันในการจำลองข้อสอบ 20 ข้อ หรือ 30 ข้อ รวมทั้งค่าเริ่มต้นที่ต่างกันจะทำให้คำตอบมีความเป็นอิสระกันมากขึ้น แต่อาจจะเกิดความคลาดเคลื่อนมากกว่าเมื่อเทียบกับการใช้โมเดลพารามิเตอร์และค่าเริ่มต้นที่เป็นแบบเดียวกัน

3.3 การจำลองตัวเลขสุ่ม จะใช้ตัวเลขสุ่มที่มีการแจกแจงแบบยูนิฟอร์ม (Uniform distribution) เป็นส่วนใหญ่และใช้วิธีการจำลองข้อมูลแบบ Congruential generators ซึ่งเป็นวิธีที่ใช้โมเดลพีชคณิตในการจำลองตัวเลขที่สุ่มขึ้นมากับตัวเลขสุ่มที่ผ่านมา ตัวเลขสุ่มจะเริ่มจำนวนจาก 0....m ซึ่งจะสุ่มโดยโปรแกรม โดยจะวิ่งเป็นวงจรที่เรียกว่า “ความยาวรอบ” (Period) เมื่อครบรอบก็จะวนกลับมาใช้เลขเดิมอีก จากการศึกษาที่ผ่านมา พบว่า วิธีนี้มีโอกาสเกิดข้อมูลซ้ำ

เมื่อมีช่วงความยาวรอบมากขึ้น เทคนิค MC จะใช้การแจกแจงปกติมาตรฐานเป็นส่วนใหญ่

3.4 การแปลงตัวเลขสุ่มเป็นคำตอบ เริ่มจากสุ่มค่าจากการแจกแจง (โดยมากเป็นการแจงแจงแบบปกติ) เพื่อให้ได้ N (จำนวนผู้สอบ) และ θ (ความสามารถ) และเป็นโมเดล IRT แบบเอกมิติหรือจากการแจกแจงปกติหลายตัวแปรที่มีค่าความสัมพันธ์ระหว่างตัวแปร เช่น สุ่มความน่าจะเป็นของคำตอบที่ตอบแบบ (0, 1) จะได้จากสมการ ดังนี้

$$P_i(\theta) = c_i + \frac{(1 \quad c_i)}{1 + e^{\ Da_i(\theta \ \ b_i)}} \tag{8}$$

ค่าความน่าจะเป็นของการตอบ (p) จะแปลงไปเป็นคำตอบ 0, 1 โดยเปรียบเทียบค่าความน่าจะเป็นจากการสุ่มการแจกแจงแบบยูนิฟอร์ม (u) ถ้า (p - u) > 0 ให้เป็น 1 (ตอบถูก) และถ้า (p - u) < 0 ให้เป็น 0 (ตอบผิด) ในกรณีเป็นคำตอบหลายค่า ถ้าตัวเลขสุ่มตกอยู่ในช่วงใดก็เป็นคำตอบ K และ K + 1 โดยทำซ้ำด้วยตัวเลขสุ่มที่แตกต่างกันในแต่ละข้อและผู้สอบทั้งหมด

3.5 การประมาณค่าพารามิเตอร์ในโมเดล อาจใช้โปรแกรมสำเร็จรูป เช่น BILOG หรือ MULTILOG หรือสร้างโปรแกรมเอง โดยกำหนดค่าเริ่มต้น (Starting value) และแก้ปัญหาการไม่ลู่เข้า (Non-convergent solutions) ซึ่งหากพบว่าการประมาณค่าไม่ลู่เข้า จะสามารถดำเนินการ 3 ทางเลือก โดยทางเลือกที่ 1 คือ ไม่สนใจการไม่ลู่เข้านั้น แล้วใช้ค่าประมาณจากจำนวนครั้งของการคำนวณซ้ำที่มากที่สุด ทางเลือกที่สอง แยกการประมาณค่าของค่าสถิติ สรุปรวม ได้แก่ RMSDs และทางเลือกที่สามใช้วิธีการคำนวณอื่น ๆ เช่น วิธีของเบส์ (Bayesian) เพื่อบังคับค่าพารามิเตอร์ โดยควรคำนึงถึงค่าพารามิเตอร์ที่ประมาณค่ากับค่าพารามิเตอร์ที่แท้จริงให้มีค่าใกล้เคียงกันด้วย มิฉะนั้นแล้วค่าที่ได้จะเป็นค่าที่มีความลำเอียง

4. การวิเคราะห์ผลการจำลองข้อมูล การวิเคราะห์ผลจะอยู่บนพื้นฐานของคำถามการวิจัย การออกแบบการทดสอบสมมติฐานทางสถิติ กระบวนการวิเคราะห์ และโดยปกติการวิเคราะห์ผลจะประกอบด้วย การใช้ตารางสรุปผลรวม สถิติเชิงบรรยายเบื้องต้น หรือการนำเสนอด้วยกราฟแผนภูมิ การดูผลกระทบจากตัวแปรอิสระ อาจต้องใช้สถิติเชิงอ้างอิง ปัญหาของการวิเคราะห์คือการมีค่าต่าง ๆ จำนวนมากเมื่อต้องการรายงานผล จึงควรใช้ทั้งสถิติเชิงบรรยายและสถิติเชิงอ้างอิงเพื่อเพิ่มโอกาสในการตรวจสอบข้อมูล ซึ่งจะทำให้มีความน่าเชื่อถือมากขึ้น

คุณภาพของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

ดัชนีบ่งชี้คุณภาพของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (ศิริชัย กาญจนวาสี, 2555, หน้า 151-152) ที่นิยมใช้พิจารณาจาก อำนาจการทดสอบ (Power Rate) และอัตราความคลาดเคลื่อนประเภทที่ 1 (Type I error rate) ซึ่งการคำนวณค่าสถิติตามวิธีการตรวจสอบ

การทำหน้าที่ต่างกันของข้อสอบ มีจุดมุ่งหมายเพื่อทดสอบนัยสำคัญของผลการทดสอบ โดยมีสมมติฐานศูนย์คือ ข้อสอบไม่ได้ทำหน้าที่ต่างกัน (H_0: No DIF) ส่วนผลการทดสอบสมมติฐานของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีต่าง ๆ นำไปสู่การตัดสินใจว่า จะยอมรับสมมติฐานศูนย์ (Accept H_0) หรือปฏิเสธสมมติฐานศูนย์ (Reject H_0) สำหรับความคลาดเคลื่อน (Error) ที่เกิดขึ้นจากการทดสอบความมีนัยสำคัญมี 2 ประเภท (ศิริชัย กาญจนวาสี, ทวีวัฒน์ ปิตยานนท์ และดิเรก ศรีสุโข, 2551, หน้า 48-49) คือ

1. ความคลาดเคลื่อนประเภทที่ 1 หรือเรียกว่า Type I error คือความคลาดเคลื่อนที่จากการปฏิเสธสมมติฐานศูนย์ (Null hypothesis) เมื่อสมมติฐานศูนย์เป็นจริง

2. ความคลาดเคลื่อนประเภทที่ 2 หรือเรียกว่า Type II error คือความคลาดเคลื่อนที่เกิดจากการยอมรับหรือคงสมมติฐานศูนย์ (Null hypothesis) ไว้ ทั้ง ๆ ที่สมมติฐานศูนย์ไม่เป็นจริง

	H_0 เป็นจริง	H_0 ไม่เป็นจริง
ปฏิเสธ H_0	ความคลาดเคลื่อน ประเภทที่ 1 (α)	การตัดสินใจที่ถูกต้อง
ยอมรับ H_0	การตัดสินใจที่ถูกต้อง	ความคลาดเคลื่อน ประเภทที่ 2 (β)

ภาพที่ 12 ประเภทของความคลาดเคลื่อนของการทดสอบทางสถิติ

การทดสอบนัยสำคัญทางสถิติสำหรับการทำหน้าที่ต่างกันของข้อสอบ (Test of significance for DIF) ด้วยสถิติไค-สแควร์ ในวิธีการถดถอยโลจิสติก ได้กำหนดวิธีการศึกษาเป็นรูปแบบโมเดลที่เป็นระดับขั้น (Hierachy) โดยมีระดับขั้นสำหรับการทดสอบ 3 ระดับ (Zumbo, 1999) คือ ระดับที่ 1 เป็นระดับขั้นการกำหนดเงื่อนไขของตัวแปร ระดับที่ 2 การนำกลุ่มตัวแปรเข้าสู่สมการ และระดับที่ 3 เป็นการศึกษาปฏิสัมพันธ์ในสมการสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการถดถอยโลจิสติก โดยทดสอบค่านัยสำคัญทางสถิติไค-สแควร์ (Chi-square: χ^2) ค่าสัมประสิทธิ์ของวิธีการถดถอยโลจิสติกประมาณค่าโดยวิธี Maximum likelihood ซึ่งในอดีตการประมาณค่าโมเดลพิจารณาจากค่าความเพียงพอหรือความสัมพันธ์ส่วนประกอบของตัวแปรที่ต้องการตรวจสอบ การตรวจสอบระหว่างระดับที่ 1 กับระดับที่ 2 เป็นค่าการเปลี่ยนแปลงที่เกิดจากการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป การตรวจสอบความแตกต่างระหว่างระดับที่ 2 กับระดับที่ 3 เป็นค่าการเปลี่ยนแปลงที่เกิดจากการทำหน้าที่ต่างกัน

ของข้อสอบแบบอเนกรูป การตรวจสอบความแตกต่างใช้สถิติ G^2 ส่วนการทดสอบนัยความสำคัญใช้สถิติ χ^2 ที่มี df = 1 เมื่อ Swaminatha and Rogers (1990) ได้เสนอวิธีการทดสอบนัยสำคัญใช้สถิติ χ^2 ที่มี df = 2 ซึ่งยอมให้มีทั้งการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปและแบบอเนกรูป ที่สามารถตรวจสอบไปพร้อม ๆ กันได้

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีทดสอบอัตราส่วนความควรจะเป็น (Likelihood Ratio Test: LR Test)

Atar and Kamata (2011) กล่าวว่า กระบวนการของวิธีการทดสอบอัตราส่วนความควรจะเป็น ของ Thissen, Steinberg & Wainer (1993) ก็คือสถิติพาราเมตริกและกระบวนการพื้นฐานสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ซึ่งสามารถใช้ทดสอบได้ทั้งข้อสอบที่ทำหน้าที่ต่างกันที่เป็นรูปแบบเดียวกันและไม่เป็นรูปแบบเดียวกัน และในบริบทของทฤษฎีการตอบสนองข้อสอบ การทำหน้าที่ต่างกันของข้อสอบมีทั้งฟังก์ชันคะแนนจริงของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบสองค่าและแบบหลายค่า ถ้าการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบพบว่า ข้อสอบไม่ได้ทำหน้าที่ต่างกัน แสดงว่า ฟังก์ชันคะแนนจริงของข้อสอบทั้งกลุ่มอ้างอิงและกลุ่มเปรียบเทียบมีรูปแบบเดียวกัน แต่ถ้าตรวจสอบพบว่า ข้อสอบทำหน้าที่ต่างกัน แสดงว่า ฟังก์ชันคะแนนจริงของข้อสอบทั้งกลุ่มอ้างอิงและกลุ่มเปรียบเทียบแตกต่างกัน

แนวคิดและหลักการ

Thissen et al. (1988) ได้เสนอวิธีการทดสอบอัตราส่วนความเป็นไปได้สำหรับใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยการทดสอบความแตกต่างของผลการตอบข้อสอบระหว่างผู้สอบกลุ่มย่อย 2 กลุ่ม วิธีทดสอบอัตราส่วนความเป็นไปได้ แบ่งออกเป็น 3 วิธีย่อย ๆ คือ 1) วิธีการทดสอบอัตราส่วนความเป็นไปได้ในทฤษฎีการตอบข้อสอบในรูปทั่วไป (General LR) เป็นวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบซึ่งใช้การประมาณค่าพารามิเตอร์ในโมเดลการตอบข้อสอบด้วยวิธีความเป็นไปได้สูงสุดแบบมาร์จินอล (Marginal Maximum Likelihood; MML) 2) วิธีการทดสอบอัตราส่วนความเป็นไปได้ในทฤษฎีการตอบข้อสอบในรูปลอกลิเนียร์ (Loglinear LR) เป็นการประมาณค่าพารามิเตอร์ด้วยวิธีความเป็นไปได้สูงสุด (Maximum Likelihood; ML) และ 3) วิธีการทดสอบอัตราส่วนความเป็นไปได้ในทฤษฎีการตอบข้อสอบในรูปสารสนเทศที่มีขอบเขตจำกัด (Limited information LR) สำหรับวิธีนี้จะใช้การประมาณค่าพารามิเตอร์ในโมเดลการตอบข้อสอบแบบ Normal ogive ด้วยวิธีกำลังสองน้อยที่สุดในรูปทั่วไป (Generalized Least Squares: GLS) ทั้ง 3 วิธีดังกล่าวจะใช้ทดสอบอัตราส่วนความเป็นไปได้เพื่อทดสอบนัยสำคัญของการทำหน้าที่ต่างกันของข้อสอบซึ่งผลความกลมกลืนของโมเดลเป็นที่รู้จักโดยทั่วไปในความกลมกลืนของฟังก์ชันในทฤษฎีการตอบข้อสอบ (IRT)

ค่าความกลมกลืนของฟังก์ชันเป็นดัชนีบอกว่าโมเดลความเหมาะสมกับข้อมูลที่จะใช้กระบวนการประมาณค่าความเป็นไปได้สูงสุดในการประมาณค่าพารามิเตอร์ข้อสอบ (Camili & Shepard, 1994)

หลักการตรวจสอบความเท่าเทียมกันของการวัดด้วยวิธีการทดสอบอัตราส่วนความเป็นไปได้ จะเปรียบเทียบระหว่าง 2 โมเดล คือ โมเดลพื้นฐาน (Compact) และโมเดลเปรียบเทียบ (Augmented) ในโมเดลแรกสมมติให้มีกลุ่มผู้สอบที่แตกต่างกัน ดังนั้นจึงบังคับให้พารามิเตอร์ของข้อสอบระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบมีค่าเท่ากัน (Group 1-Item 1 และ Group 2-Item 1) (Thissen et al., 1998, 1993; Atar and Kamata, 2011, p. 37-38) โมเดลพื้นฐานนี้เป็นการจัดเตรียมค่าความเป็นไปได้ที่เกี่ยวข้องกับการประมาณค่าพารามิเตอร์ข้อสอบข้อที่ i ในแต่ละกลุ่มประชากรค่าความเป็นไปได้นี้สามารถเปรียบเทียบทุกพารามิเตอร์ของข้อสอบทั้งหมดที่บังคับให้เท่ากันข้ามกลุ่มในโมเดลพื้นฐาน

สมมติฐานในการทดสอบสำหรับโมเดลของพารามิเตอร์ของข้อสอบ โดยใช้สูตรดังนี้

$H_0: a_{jR} = a_{jF}$ และ $b_{jR} = b_{jF}$ สำหรับทุก j

H_A : พารามิเตอร์ของข้อสอบในข้อที่ j ของทั้งสองกลุ่มไม่เท่ากัน อย่างน้อย 1 พารามิเตอร์

สูตรการทดสอบอัตราส่วนความเป็นไปได้ของสองโมเดล แสดงได้ดังนี้

$$LR = \frac{L^*(Model_C)}{L^*(Model_A)} \tag{9}$$

เมื่อ $L^*(Model_C)$ แทน ฟังก์ชันความเป็นไปได้ของโมเดลพื้นฐาน (ค่าพารามิเตอร์ที่น้อยกว่า)

$L^*(Model_A)$ แทน ฟังก์ชันความเป็นไปได้ของโมเดลเปรียบเทียบที่ยอมให้พารามิเตอร์ข้อสอบของข้อสอบที่ j ข้ามกลุ่มผู้สอบมีความหลากหลาย

นั่นคือในโมเดลพื้นฐาน (Compact) จะประกอบด้วยข้อสอบที่ทำหน้าที่ไม่ต่างกัน สำหรับในโมเดลเปรียบเทียบจะประกอบด้วยข้อสอบที่มีค่าพารามิเตอร์ระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่มีค่าแปรเปลี่ยนไปตามกลุ่มโมเดลเปรียบเทียบ (Augmented) อาจมีข้อสอบจำนวน 1 ข้อหรือมากกว่าที่ทำหน้าที่ต่างกัน นอกจากนี้ระหว่าง 2 โมเดลจะต้องมีข้อสอบร่วมซึ่งเป็นข้อสอบที่สมมติว่าทำหน้าที่ไม่ต่างกัน จากนั้นจึงทำการเปรียบเทียบระหว่าง 2 โมเดลด้วยสถิติอัตราส่วนความเป็นไปได้ดังนี้

$$G_j^2 = -2\log(LR)$$
$$G^2 = -2\ln(L^*(Model_C) - [-2Ln(L^*(Model_A))] \quad (10)$$

เมื่อ G_j^2 แทน สถิติอัตราส่วนความเป็นไปได้ของข้อสอบข้อที่ i

$L^*(Model_C)$ แทน ฟังก์ชันความเป็นไปได้ของโมเดลพื้นฐาน (Compact)

$L^*(Model_A)$ แทน ฟังก์ชันความเป็นไปได้ของโมเดลเปรียบเทียบ (Augmented)

โดยทั่วไปแล้ว $L^*(Model_C) < L^*(Model_A)$ และสถิติ $G_j^2 > 0$ มีการแจกแจงแบบ chi-squared ซึ่งระดับของความเป็นอิสระเท่ากับผลต่างของจำนวนพารามิเตอร์ในโมเดล $L^*(Model_C)$ และโมเดล $L^*(Model_A)$ ดัชนีความไม่กลมกลืนในระดับนัยสำคัญทางสถิติของผลการทดสอบจะชี้บอกว่าโมเดลพื้นฐานกลมกลืนน้อยกว่าโมเดลเปรียบเทียบอย่างมีนัยสำคัญทางสถิติในการใช้การทดสอบอัตราส่วนความควรจะเป็น ค่าไค-สแควร์ ที่องศาความเป็นอิสระเท่ากับจำนวนพารามิเตอร์ของข้อสอบในการประมาณค่าของข้อสอบเหล่านั้นมีนัยสำคัญทางสถิติ แสดงว่ามีการทำหน้าที่ต่างกันของข้อสอบ (Atar & Kamata, 2011)

ในการเปรียบเทียบค่าพารามิเตอร์ในโมเดลที่ตั้งสมมติฐานให้มีค่าเท่ากันระหว่างกลุ่มเปรียบเทียบกลุ่มอ้างอิง สถิติที่ใช้ทดสอบอัตราส่วนความควรจะเป็น คือ -2log likelihood (-2LL) และนำค่าที่ได้ไปเทียบการแจกแจง χ^2 กับ df เท่ากับจำนวนค่าพารามิเตอร์ที่ทดสอบ ถ้าผลปรากฏว่าปฏิเสธสมมติฐานว่างแสดงว่าข้อสอบทำหน้าที่ต่างกัน นั่นคือค่าพารามิเตอร์อำนาจจำแนก (a_i) หรือค่าพารามิเตอร์ลำดับขั้น (b_{jk}) หรือทั้งสองค่า มีค่าไม่เท่ากันระหว่างกลุ่มอ้างอิงกับกลุ่มสนใจ (Woods, 2011 อ้างถึงใน อาวิพร ปานทอง, 2558) ค่า chi-squared (χ^2) จะมี Degree of freedom (df) เท่ากับจำนวนตัวแปรอิสระหรือจำนวนพารามิเตอร์ของข้อสอบ ค่า χ^2 สามารถคำนวณได้ดังนี้

$$\chi^2 = -2[LL(C)-LL(A)] \quad (11)$$

โดยที่ LL(C) คือ ค่า Log-likelihood เมื่อ โมเดลไม่มีข้อจำกัด (Unrestricted model) และ LL(A) คือ ค่า Log-likelihood เมื่อโมเดลมีข้อจำกัด (Restricted model) เนื่องจากค่า LL(C) มากกว่า LL(A) ดังนั้นค่า χ^2 จึงมีค่าเป็นบวกเสมอ

วิธีซิปเทสท์ (SIBTEST)

วิธีซิปเทสท์" (SIBTEST) ของ Shealy & Stout (1993) เป็นวิธีที่ใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ และแบบทดสอบ (Differential item/ test functioning; DIF/ DTF) ในข้อสอบที่มีรูปแบบหารตรวจให้คะแนนแบบสองค่า พัฒนามาจากโมเดลความลำเอียงของแบบสอบภายใต้ทฤษฎีการตอบข้อสอบแบบหลายมิติ (Multidimensional IRT) มีรูปแบบนันพาราเมทริก (Nonparametric form) ไม่ต้องใช้ฟังก์ชันการตอบข้อสอบประมาณค่าความสามารถ วิธีซิปเทสท์เป็นวิธีที่ใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีทิศทางเดียวกัน (Unidirectional DIF) โดยเฉพาะ ข้อได้เปรียบของวิธีซิปเทสท์ก็คือ สามารถคำนวณได้ง่าย เสียค่าใช้จ่ายไม่มาก และไม่จำเป็นต้องใช้ตัวอย่างขนาดใหญ่ สามารถใช้สถิติทดสอบทดสอบนัยสำคัญ เพื่อตัดสินการทำหน้าที่ต่างกัน ของข้อสอบครั้งละหนึ่งข้อ หรือมากกว่าหนึ่งข้อพร้อมกัน (Simultaneous) ผลการวิเคราะห์ทำให้ทราบขนาดและทิศทางของการทำหน้าที่ต่างกันของข้อสอบ (Nandakumar, 1993, p. 295) นอกจากนี้ Li and Stout (1996) ได้พัฒนาวิธี Crossing SIBTEST เพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบตัดกัน รุสโซและสเตาท์ (Roussos & Stout, 1996) ได้ศึกษาอัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีซิปเทสท์ในกลุ่มตัวอย่างขนาดเล็ก และ Chang et al. (1996) ได้พัฒนาวิธีโพลี-ซิปเทสท์ (Poly-SIBTEST) เพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า สเตาท์และคนอื่น ๆ (Stout, Li, Nandakumar, & Bolt, 1997) ได้พัฒนาวิธีมัลติซิป (MULTISIB) เพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบสองมิติ เป็นต้น (อรินทร์ น่วมถนอม, 2549)

แนวคิดและหลักการ

Shealy and Stout (1993, pp. 163-164) ได้อธิบายการทำหน้าที่ต่างกันของข้อสอบโดยใช้ขอบของฟังก์ชันการตอบข้อสอบ (Marginal IRFs) ของความสามารถเป้าหมาย ที่ต้องการวัด θ สำหรับกลุ่ม g (กลุ่มอ้างอิง หรือกลุ่มสนใจ) ดังนี้

$$M_{ig}(\theta) = E\left[P_i(\Theta, \eta)\middle| = \theta, G = g\right] \tag{12}$$

ถ้า $\eta|\Theta = \theta, G = g$ มีความหนาแน่นแบบมีเงื่อนไขของ η เมื่อกำหนดความสามารถ θ ของกลุ่ม g มีค่าคงที่ซึ่งแทนด้วย $f_g(\eta|\theta)$ ดังนั้นการกำหนดนิยามในสมการ (1) สามารถคำนวณได้ดังนี้

$$M_{ig}(\theta) = \int_{+\infty}^{-\infty} P_i(\theta, \eta) f_g(\eta|\theta) d\eta \tag{13}$$

จากสมการ (13) ถ้าการแจกแจงแบบมีเงื่อนไขของความสามารถ η มีค่าเท่ากันสำหรับผู้สอบสองกลุ่ม แล้วข้อสอบจะทำหน้าที่ต่างกัน (Mo-DIF) เพราะว่ามีความสามารถ θ เท่ากัน จะทำให้ความน่าจะเป็นในการตอบข้อสอบถูกเท่ากัน (Ackeman, 1992, p. 76) จากแนวคิดดังกล่าว สามารถนิยามการทำหน้าที่ต่างกันของข้อสอบ (DIF) โดยใช้ Marginal IRFs ได้ว่า “ถ้าฟังก์ชันการตอบข้อสอบของความสามารถเป้าหมายสำหรับกลุ่มอ้างอิงมีค่ามากกว่าฟังก์ชันการตอบข้อสอบสำหรับกลุ่มสนใจ แล้วข้อสอบจะทำหน้าที่ต่างกัน โดยข้อสอบเข้าข้างกลุ่มอ้างอิง” ซึ่งแสดงในรูปสัญลักษณ์ทางคณิตศาสตร์ ดังนี้

$$M_{iR}(\theta) > M_{iF}(\theta) \tag{14}$$

ฟังก์ชันการตอบข้อสอบของแบบสอบที่ต้องการศึกษา สำหรับผู้สอบกลุ่มอ้างอิงและกลุ่มสนใจ กำหนดดังนี้

$$M_{SR}(\theta) = \sum_{i=n+1}^{N} M_{iR}(\theta) \tag{15}$$

$$M_{SF}(\theta) = \sum_{i=n+1}^{N} M_{iF}(\theta) \tag{16}$$

เมื่อ $M_{SR}(\theta)$ และ $M_{SF}(\theta)$ แทนผลรวม Marginal IRFs ของข้อสอบที่ต้องการศึกษา ณ ระดับความสามารถ θ จากผู้สอบกลุ่มอ้างอิงและกลุ่มสนใจ ตามลำดับ สำหรับปริมาณของการทำหน้าที่ต่างกันของข้อสอบ (Amount of DIF) สามารถคำนวณจากความแตกต่างระหว่าง $M_{SR}(\theta)$ และ $M_{SF}(\theta)$ ดังนี้

$$B(\theta) = M_{SR}(\theta) - M_{SF}(\theta) \tag{17}$$

-

ขนาดของความแตกต่างดังกล่าว แสดงถึงปริมาณของการทำหน้าที่ต่างกันของข้อสอบจากแบบทดสอบชุดย่อยที่ต้องการศึกษา ณ ระดับความสามารถ θ ซึ่งเข้าข้างกลุ่ม ได้คำนวณค่าเฉลี่ยของปริมาณการทำหน้าที่ต่างกันของข้อสอบที่มีทิศทางเดียวกัน (Unidirectional DIF)

$$\beta_{uni} = \int_{+\infty}^{-\infty} B(\theta) f_F(\theta) d\theta \qquad (18)$$

เมื่อ β_{uni} แทนดัชนีการทำหน้าที่ต่างกันของข้อสอบที่มีทิศทางเดียวกัน และ $f_F(\theta)$ แทนฟังก์ชันความหนาแน่นความน่าจะเป็นของการแจกแจงความสามารถเป้าหมาย จากผู้สอบกลุ่มรวมทั้งหมด

กระบวนการตรวจสอบ

ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีทิศทางเดียวกันตามแนวคิดของ Shealy and Stout (1993) จะเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มอ้างอิงและกลุ่มสนใจ โดยใช้แบบทดสอบจำนวน N ข้อ แล้วแบ่งแบบทดสอบดังกล่าวออกเป็นสองชุด คือ แบบทดสอบชุดย่อยที่มีความถูกต้อง (Valid subtests) และแบบทดสอบชุดย่อยที่ใช้ในการศึกษา (Studied subtests) กล่าวคือ แบบทดสอบชุดแรกใช้ในการจับคู่เปรียบเทียบ (Matching subtests) ประกอบด้วยข้อสอบข้อที่ 1 ถึง n ซึ่งเป็นข้อสอบที่ไม่สงสัยว่าทำหน้าที่ต่างกัน โดยวัดความสามารถเป้าหมาย θ เพียงความสามารถเดียว ส่วนแบบทดสอบชุดหลังเป็นส่วนที่เหลือจากชุดแรกประกอบด้วยข้อสอบข้อที่ $n+1$ ถึง N ข้อสอบดังกล่าวสงสัยว่าทำหน้าที่ต่างกัน โดยวัดทั้งความสามารถเป้าหมาย θ และความสามารถแทรกซ้อน η

การทดสอบสมมติฐาน

ในการทดสอบสมมติฐานของการทำหน้าที่ต่างกันของข้อสอบที่มีทิศทางเดียวกัน เมื่อข้อสอบเข้าข้างกลุ่มอ้างอิง นำดัชนี β_{uni} มากำหนดสมมติฐานศูนย์ (H_0) และสมมติฐานทางเลือก (H_1) ดังนี้

$$H_0 : \beta_{uni} = 0$$
$$H_1 : \beta_{uni} > 0 \qquad (19)$$

การทดสอบสมมติฐานการทำหน้าที่ต่างกันของข้อสอบที่มีทิศทางเดียวกันจะประมาณค่าดัชนี β_{uni} โดยคำนวณจากคะแนนของแบบทดสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบ และแบบทดสอบชุดย่อยที่ต้องการศึกษา ดังนี้

$$X = \sum_{i=1}^{n} U_i \qquad (20)$$

$$Y = \sum_{i=n+1}^{N} U_i \qquad (21)$$

เมื่อ

X แทน คะแนนรวมจากแบบทดสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบ

Y แทน คะแนนรวมจากแบบทดสอบชุดย่อยที่ต้องการศึกษา

U_i แทน ผลการตอบข้อสอบข้อที่ i (ตอบถูกได้ 1 คะแนน และตอบผิดได้ 0 คะแนน)

การคำนวณคะแนนเฉลี่ยจากผลการตอบข้อสอบในแบบสอบชุดย่อยที่ต้องการศึกษาของผู้สอบกลุ่มอ้างอิงและกลุ่มสนใจที่มีความสามารถระดับเดียวกัน แล้วนำคะแนนเฉลี่ยดังกล่าวมาจับคู่เปรียบเทียบ โดยพิจารณาจากคะแนนรวมที่เท่ากันของแบบสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบ $(X = k)$ ดังนี้

$$\overline{Y}_{Rk} - \overline{Y}_{Fk} \qquad (22)$$

เมื่อ $\overline{Y}_{Rk}$ และ $\overline{Y}_{Fk}$ แทนค่าเฉลี่ยของคะแนน Y จากการตอบแบบทดสอบชุดย่อยที่ใช้ในการจับคู่เปรียบเทียบ แล้วได้คะแนนรวม $X = k$ สำหรับผู้สอบในกลุ่มอ้างอิงและกลุ่มสนใจตามลำดับ คะแนนเฉลี่ยที่ใช้ในการเปรียบเทียบดังกล่าวอาจทำให้การตรวจสอบผิดพลาดจากความเป็นจริง กล่าวคือ เมื่อเกิดความแตกต่างของการแจกแจงค่าความสามารถ (Ability distribution) ของกลุ่มอ้างอิงและกลุ่มสนใจจะมีผลทำให้ $\overline{Y}_{Rk} - \overline{Y}_{Fk}$ มีค่าแตกต่างจาก 0 อย่างเป็นระบบ ทำให้ตรวจพบว่าข้อสอบทำหน้าที่ต่างกัน ซึ่งความเป็นจริงแล้วข้อสอบทำหน้าที่ไม่ต่างกัน ดังนั้น ความแตกต่างของการแจกแจงค่าความสามารถของกลุ่มอ้างอิงและกลุ่มสนใจที่เกิดขึ้นสามารถปรับแก้ค่าการถดถอย (Regression correction) เพื่อกำจัดค่าที่สูงเกิดปกติ (Inflate) สำหรับค่าเฉลี่ย $\overline{Y}_{Rk}$ และ $\overline{Y}_{Fk}$ ที่ปรับแก้วแล้วแทนด้วย $\overline{Y}^{*}_{Rk}$ และ $\overline{Y}^{*}_{Fk}$ ตามลำดับ

ค่า $\overline{Y}^{*}_{Rk} - \overline{Y}^{*}_{Fk}$ เป็นความแตกต่างของผลการตอบข้อสอบในแบบทดสอบชุดย่อยที่ศึกษาระหว่างกลุ่มผู้สอบที่มีความสามารถระดับเดียวกัน ถ้า $\overline{Y}^{*}_{Rk} - \overline{Y}^{*}_{Fk} = 0$ ทุกคะแนน k แสดงว่าข้อสอบที่สงสัยในแบบทดสอบชุดย่อยช่วยทำหน้าที่ต่างกัน (No-DIF) และถ้า $\overline{Y}^{*}_{Rk} - \overline{Y}^{*}_{Fk} > 0$ ทุกคะแนน k แสดงว่าข้อสอบที่สงสัยในแบบสอบชุดย่อยทำหน้าที่ต่างกันที่มีทิศทางเดียวกัน (Unidirectional DIF) โดยข้อสอบเข้าข้างกลุ่มอ้างอิง ถ้า $\overline{Y}^{*}_{Rk} - \overline{Y}^{*}_{Fk} < 0$ ทุกคะแนน k แสดงว่าข้อสอบที่สงสัยในแบบทดสอบชุดย่อยทำหน้าที่ต่างกันที่มีทิศทางเดียวกัน โดยข้อสอบเข้าข้างกลุ่มสนใจ สำหรับค่าความต่างของผลการตอบข้อสอบดังกล่าว สามารถนำมาประมาณค่าในรูป $\hat{\beta}_{uni}$ ดังนี้

$$\beta_{uni} = \sum_{k=0}^{n} \hat{P}_k (\overline{Y}^*_{Rk} - \overline{Y}^*_{Fk}) \quad (23)$$

เมื่อ $\hat{P}_k$ แทนสัดส่วนของผู้สอบทั้งหมด (กลุ่มอ้างอิงและกลุ่มสนใจ) ผู้ตอบแบบทดสอบชุดย่อยที่ใช้จับคู่เปรียบเทียบ แล้วได้คะแนนรวม $X = k$ สัดส่วนของผู้สอบดังกล่าวสามารถเขียนในรูปสัญลักษณ์ ดังนี้

$$\hat{P}_k = \frac{(J_{Rk} + J_{Fk})}{\sum_{k=0}^{n} (J_{Rk} + J_{Fk})} \quad (24)$$

เมื่อ J_{Rk} และ J_{Fk} แทนจำนวนผู้สอบซึ่งตอบแบบทดสอบชุดย่อยที่ใช้จับคู่เปรียบเทียบแล้วได้คะแนนรวม $X = k$ สำหรับกลุ่มอ้างอิงและกลุ่มสนใจ ตามลำดับ จากนั้นจึงนำค่าประมาณ $\hat{\beta}_{uni}$ ที่คำนวณในสมการ (23) มาทดสอบสมมติฐานศูนย์ (No-DIF) โดยใช้สถิติ B_{uni} ดังนี้

$$B_{uni} = \frac{\hat{\beta}_{uni}}{\hat{\sigma}(\hat{\beta}_{uni})} \quad (25)$$

$\hat{\sigma}(\hat{\beta}_{uni})$ เป็นค่าประมาณความคลาดเคลื่อนมาตรฐานของ B_{uni} คำนวณจาก

$$\hat{\sigma}(\hat{\beta}_{uni}) = \sqrt{\sum_{k=0}^{n} \hat{P}_k^2 \left[\frac{1}{J_{Rk}} \hat{\sigma}^2(Y|k,R) + \frac{1}{J_{Fk}} \hat{\sigma}^2(Y|k,F) \right]} \quad (26)$$

เมื่อ $\hat{\sigma}^2(Y|k,R)$ และ $\hat{\sigma}^2(Y|k,F)$ แทนค่าประมาณความแปรปรวนของคะแนนจากแบบทดสอบชุดย่อยที่ต้องการศึกษาในกลุ่มอ้างอิงและกลุ่มสนใน ตามลำดับ สำหรับสถิติ B_{uni} มีการแจกแจงใกล้เคียงการแจกแจงแบบปกติมาตรฐาน [N(0,1)] เมื่อข้อสอบทำหน้าที่ไม่ต่างกัน และถ้าผลการทดสอบพบว่า $B_{uni} > Z_{\alpha}$ อย่างมีนัยสำคัญที่ระดับ α โดยที่ $P[N(0,1) > Z_{\alpha}] = \alpha$ แสดงว่า ปฏิเสธ H_0 นั่นคือ ข้อสอบทำหน้าที่ต่างกันที่มีทิศทางเดียวกัน (Unidirectional DIF) เมื่อ $B_{uni} > 0$ แสดงว่าข้อสอบเข้าข้างกลุ่มอ้างอิง และเมื่อ $B_{uni} < 0$ แสดงว่าข้อสอบเข้าข้างกลุ่มสนใจ โดยเกณฑ์ที่ใช้จำแนกขนาดของการทำหน้าที่ต่างกันของข้อสอบมีดังนี้

- DIF ระดับ A ขนาดเล็ก: ปฏิเสธสมมติฐานศูนย์ และ $\left|\hat{B}_{uni}\right| < 0.059$
- DIF ระดับ B ขนาดปานกลาง: ปฏิเสธสมมติฐานศูนย์ และ $0.059 \le \left|\hat{B}_{uni}\right| < 0.088$
- DIF ระดับ C ขนาดใหญ่: ปฏิเสธสมมติฐานศูนย์ และ $\left|\hat{B}_{uni}\right| \ge 0.088$

วิธีโพลี-ซิปเทสท์ (Poly-SIBTEST)

วิธีซิปเทสท์ (Shealy & Stout, 1993) เป็นวิธีที่พัฒนาสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนสองค่า ต่อมา Chang et al. (1996) ได้ปรับขยายวิธีซิปเทสท์ (Modified SIBTEST) เพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนหลายค่า และเรียกวิธีที่พัฒนาใหม่นี้ว่า “วิธีโพลี-ซิปเทสท์” (Poly-SIBTEST) (อรินทร์ น่วมถนอม, 2549) มีรายละเอียดดังนี้

แนวคิดและหลักการ

นิยามการทดสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนสองค่าของตัวแปรแฝง (Latent variable) กำหนดไว้ว่า เมื่อกำหนดให้ $E_R[Y|\theta]$ และ $E_F[Y|\theta]$ แทนการถดถอยของ Y บนตัวแปรแฝง θ ของกลุ่มอ้างอิงและกลุ่มสนใจ ตามลำดับ ข้อสอบจะไม่แสดงทำหน้าที่ต่างกัน (Not exhibit DIF) ถ้าทุกค่าของความสามารถ θ ที่มี

$$E_R[Y|\theta] = E_F[Y|\theta] \tag{27}$$

และเมื่อกำหนด $E_R[Y|X]$ และ $E_F[Y|X]$ แทนการถดถอยของ Y บนคะแนนสังเกต (Observed score) ของกลุ่มอ้างอิงและกลุ่มสนใจ ตามลำดับ ข้อสอบจะไม่แสดงทำหน้าที่ต่างกัน เบนถ้าทุกค่าของคะแนน X ที่มี

$$E_R[Y|X] = E_F[Y|X] \tag{28}$$

จากนิยามการทดสอบสมมติฐานศูนย์ของการทำหน้าที่ต่างกันของข้อสอบ (Null-DIF) โดยใช้ตัวแปรแฝงและคะแนนสังเกตดังที่กล่าวมา การศึกษาของรุสโซและสเตาท์ (Chang et al., 1996, p. 335; Roussos & Stout, 1996) ได้เสนอแนะว่า ในการนำนิยามทั้งสองกรณีมาใช้ควรพิจารณาแยกส่วนกัน

การนำนิยามของตัวแปรแฝงในสมการ (28) มาใช้ทดสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า สามารถกำหนดได้ว่า “ข้อสอบไม่แสดงการทำหน้าที่ต่างกัน (Not exhibit DIF)” ถ้าการถดถอยของคะแนนข้อสอบบนตัวแปรแฝง

เหมือนกัน สำหรับกลุ่มผู้สอบ เมื่อกำหนดให้ Y แทนคะแนนของข้อสอบที่ต้องการศึกษา โดยเป็นคะแนนในรายการแบบจัดอันดับ (Ordered categories) ซึ่งมี $m+1$ รายการ $(Y = k,\ 0 \le k \le m)$ และ $P_{k,g}(\theta)$ แทนฟังก์ชันการตอบรายการของข้อสอบ (Item-Category Response Function; ICRF) ที่ระดับคะแนน k ในกลุ่ม g ดังนั้นการถดถอยของคะแนนข้อสอบบนความสามารถ θ จะกำหนดในรูปผลรวมของฟังก์ชันการตอบรายการของข้อสอบแบบถ่วงน้ำหนัก ดังนี้

$$E_g[Y|\theta] = \sum_{k=1}^{m} kP_{k,g}(\theta) \tag{29}$$

สำหรับวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า จะเปรียบเทียบความแตกต่างของ ICRFs โดยสามารถเขียนในรูปสัญลักษณ์ (อาวีพร ปานทอง, 2558) ได้ดังนี้

$$P_{kR}(\theta) = P_{kF}(\theta) \tag{30}$$

โดย $k = 1, 2, \ldots, m$

เมื่อ $P_{kR}(\theta)$ และ $P_{kF}(\theta)$ แทน ICRFs ที่ระดับคะแนน k ของผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบตามลำดับ โดยข้อสอบ 1 ข้อ จะมี ICRFs จำนวน m ฟังก์ชัน สามารถนำมาใช้กับข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าแบบเรียงลำดับ สำหรับวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ซึ่งใช้คะแนนสังเกตได้ในสมการ ดังเช่น วิธีแมนเทล-แฮนส์เซล (Mantel-Haenszel; MH) (Holland & Thayer, 1998) และวิธีการทำให้เป็นมาตรฐาน (Standardization; STD) (Dorans & Kulick, 1986) สามารถกำหนดเป็นนิยามได้ว่า “ข้อสอบไม่แสดงทำหน้าที่ต่างกัน (Not exhibit DIF) ถ้าการถดถอยของคะแนนข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า บนคะแนนสังเกตของแบบสอบที่ใช้ในการจับคู่เหมือนกันสำหรับกลุ่มผู้สอบ

กระบวนการตรวจสอบ

ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า โดยใช้วิธีโพลี-ซิปเทสท์จะใช้กรอบแนวคิดเช่นเดียวกับวิธีซิปเทสท์ต้นฉบับเดิม (Original SIBTEST) ที่ตรวจสอบในกรณีให้คะแนนสองค่า โดยการประยุกต์ใช้สถิติเพื่อทดสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนในรายการแบบจัดอันดับ (Ordered categories) ดังนี้

กำหนดให้

Y แทนคะแนนของข้อสอบที่ต้องการศึกษา ซึ่งให้คะแนนตามรายการแบบจัดอันดับจำนวน $m+1$ รายการ $(Y = 0,\ 1,\ 2,\ ...,\ m)$

$X_1,\ X_2,\ ...,\ X_n$ แทนคะแนนของข้อสอบที่ใช้ในการจับคู่เปรียบเทียบ จำนวน *n* ข้อ

$m_1,\ m_2,\ ...,\ m_n$ แทนคะแนนมากสุดที่เป็นไปได้ของ $X_1,\ X_2,\ ...,\ X_n$ ตามลำดับ

X แทนคะแนนที่ใช้ในการจับคู่เปรียบเทียบ คำนวณจากสูตร

$$X = \sum_{j=1}^{n} X_j \tag{31}$$

เมื่อ $X = 0,\ 1,\ 2,\ ...,\ n_H$ โดยที่ n_H เป็นคะแนนของข้อสอบที่ใช้ในการจับคู่เปรียบเทียบ ซึ่งมีค่ามากที่สุดที่เป็นไปได้ สามารถคำนวณได้ดังนี้

$$n_H = \sum_{j=1}^{n} m_j \tag{32}$$

$\overline{Y}_{gk}$ แทนคะแนนเฉลี่ยของข้อสอบที่ต้องการศึกษาสำหรับผู้สอบทั้งหมดในกลุ่ม g (R หรือ F) ซึ่งได้คะแนน $X = k$ ถึงแม้ว่าวิธีซิปเทสท์พัฒนามาจากโมเดลของทฤษฎีการตอบข้อสอบ (IRT) แต่ในที่นี้จะนำทฤษฎีการทดสอบแบบมาตรฐานเดิม (CTT) มาอธิบายเพื่อทำความเข้าใจกระบวนการตรวจสอบของวิธีดังกล่าว ตามข้อตกลงของทฤษฎีการทดสอบมาตรฐานเดิมเกี่ยวกับคะแนน X กำหนดว่า $X = T + E$ เมื่อ T แทนคะแนนจริงของแบบทดสอบที่ใช้ในการจับคู่ ดังนั้น ตัวแปร $E = X - T$ แทนความคลาดเคลื่อนในการวัด และสมมติว่ามีค่าเฉลี่ยเป็นศูนย์ทั้งสองกลุ่มเมื่อให้ $f_g(t)$ แทนความหนาแน่นของคะแนนจริงของแบบทดสอบที่ใช้ในการจับคู่ในกลุ่ม g (R หรือ F) ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจะพิจารณาการถดถอยของคะแนนข้อสอบที่ต้องการศึกษาบนคะแนนจริงของแบบทดสอบที่ใช้ในการจับคู่ในกลุ่ม g (R หรือ F) ซึ่งการถดถอยดังกล่าวเขียนในรูปสัญลักษณ์ได้ว่า

$E_g[Y|t] = [Y|T = t,\ G = g]$

นิยามของข้อสอบที่ต้องการศึกษาไม่แสดงการทำหน้าที่ต่างกัน (Not exhibit DIF) ตามทฤษฎีการทดสอบแบบมาตรฐานเดิมกำหนดได้ว่า “ถ้า $E_R[Y|t] = E_F[Y|t]$ สำหรับทุกค่าของคะแนนจริง t จากแบบทดสอบที่ใช้ในการจับคู่” นิยามดังกล่าวสมมูลกับนิยามที่กำหนดโดยใช้

ตัวแปรแฝงตามทฤษฎีการตอบสนองข้อสอบ (Chang et al., 1996, p. 337; citing Chang & Mazzeo,1994) ดังนั้นสามารถนำนิยามดังกล่าวไปอธิบายการทำหน้าที่เบี่ยงเบนของข้อสอบที่ให้คะแนนหลายค่าด้วยวิธีโพลี-ซิปเทสท์ ซึ่งมีรายละเอียดดังนี้

วิธีซิปเทสท์พัฒนามาจากโมเดลในทฤษฎีการตอบข้อสอบ ดังนั้นการกำหนดนิยามการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนสองค่าจะใช้ตัวแปรแฝง ซึ่งสามารถกำหนดได้ว่า "การทำหน้าที่ต่างกันของข้อสอบเกิดขึ้นเมื่อ $E_R[Y|\theta] \neq E_F[Y|\theta]$ ที่ระดับความสามารถ θ" และเมื่อ $f_F(\theta)$ แทนความหนาแน่นของความสามารถ θ ในกลุ่มเปรียบเทียบ ดัชนีการทำหน้าที่ต่างกันของข้อสอบสามารถคำนวณได้จาก $B_o = \int B_o(\theta) f_F(\theta) d\theta$

นิยามการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าได้ว่า $B(t) = E_R[Y|t] - E_F[Y|t]$ โดยดัชนีการทำหน้าที่ต่างกันของข้อสอบสามารถคำนวณจาก $f_F(t)$ $\beta = \int B_o(t) f_F(t) dt$ ซึ่งสามารถประมาณค่าได้ดังนี้

$$d_k = \overline{Y}_{Rk} - \overline{Y}_{Fk} \tag{33}$$

เมื่อ $k = 1, 2, \ldots, n_H$

เมื่อ Y_{Rk} และ Y_{Fk} แทนค่าเฉลี่ยของคะแนนข้อสอบที่ต้องการศึกษาของผู้สอบทั้งหมดซึ่งได้คะแนน $X = k$ ของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ตามลำดับ ค่า $Y_{Rk} - Y_{Fk}$ เป็นความแตกต่างของผลการตอบข้อสอบที่ต้องการศึกษาระหว่างกลุ่มผู้สอบ ที่มีคะแนนสังเกตได้จากแบบสอบที่ใช้ในการจับคู่เท่ากัน ถ้าสมมติว่าผู้สอบมีคะนนสังเกตได้เท่ากับหรือใกล้เคียงกับคะแนนจริงจากแบบสอบที่มช้ในการจับคู่ซึ่งจะเป็นคะแนนจริงได้ ถ้าแบบสอบดังกล่าวมีความยาวมากพอเพื่อทำให้มีความเชื่อมั่นสูง หรือกลุ่มผู้สอบที่ศุกษามีการแจกแจงความสามารถคล้ายคลึงกัน ดังนั้นสมการ (33) ถือได้ว่าเป็นความแตกต่างในคะแนนข้อสอบที่ระดับคะแนนจริงเท่ากัน แต่ถ้าข้ออบที่ศึกษาไม่มีคะแนนสังเกตได้ที่ทำให้ข้อสอบทำหน้าที่ต่างกันแล้วคาดว่า $d_k \approx 0$ สำหรับการประมาณค่าดัชนีการทำหน้าที่ต่างกันของข้อสอบ สามารถคำนวณได้ดังนี้

$$\hat{\beta} = \sum_{k=0}^{n_H} p_k d_k \tag{34}$$

โดยที่

$$p_k = \frac{N_{Rk} + N_{Fk}}{N} \tag{35}$$

เมื่อ p_k แทนสัดส่วนของผู้สอบทั้งหมด (กลุ่มอ้างอิงและกลุ่มสนใจ) ซึ่งตอบแบบสอบที่ใช้ในการจับคู่ X_1, X_2, ..., X_n แล้วได้คะแนน $X = k$ ต่อจากนั้นจะนำค่าประมาณ $\hat{\beta}$ มาทดสอบสมมติฐาน

การทดสอบสมมติฐาน

นำดัชนี $\hat{\beta}$ มาทดสอบสมมติฐานศูนย์ (No-DIF) โดยใช้สถิติ B (อรินทร์ น่วมถนอม, 2549) ดังนี้

$$B = \frac{\hat{\beta}}{\hat{\sigma}(\hat{\beta})} \tag{36}$$

โดยที่

$$\hat{\sigma}(\hat{\beta}) = \sqrt{\sum_{k=0}^{n_H} P_k^2 \left[\frac{\hat{\sigma}^2(Y|k,\ R)}{N_{Rk}} + \frac{\hat{\sigma}^2(Y|k,\ F)}{N_{Fk}} \right]} \tag{37}$$

เมื่อ $\hat{\sigma}(\hat{\beta})$ แทนค่าประมาณความคลาดเคลื่อนมาตรฐานของ $\hat{\beta}$ และ $\hat{\sigma}^2(Y|k,\ g)$ แทนค่าประมาณความแปรปรวนของคะแนนจากแบบสอบที่ต้องการศึกษาในกลุ่ม g (R หรือ F) ซึ่งมีคะแนนรวมเท่ากับ *k* สำหรับสถิติ *B* มีการแจกแจงใกล้เคียงปกติมาตฐาน [N(0, 1)] เมื่อข้อสอบทำหน้าที่ไม่ต่างกัน (No-DIF) ถ้าผลการทดสอบพบว่า $|B| > Z_{1\ \frac{\alpha}{2}}$ อย่างมีนัยสำคัญที่ระดับ α แสดงว่า ปฏิเสธ H_0 นั่นคือ ข้อสอบทำหน้าที่ต่างกันน (DIF) โดยเข้าข้างผู้สอบกลุ่มใดกลุ่มหนึ่ง ในการทดสอบด้วยสถิติ B อาจมีความคลาดเคลื่อนประเภทที่ 1 สูง ซึ่งเกิดจากความแตกต่างของการแจกแจงความสามารถระหว่างกลุ่มผู้สอบ ความแตกต่างดังกล่าวจะส่งผลกระทบ (Impact) ต่อการทดสอบสถิติ B_{uni} โดยทำให้มีค่าเฟ้อ (Inflate) หรือสูงผิดปกติ ปัญหาดังกล่าวสามารถแก้ไขโดยใช้การถดถอยเชิงเส้นของคะแนนจริงบนคะแนนสังเกตจากทฤษฎีการทดสอบแบบดั้งเดิม ในกรณีข้อสอบที่มีรูปแบบการตรวจให้คะแนนหลายค่าจะใช้การคำนวณแอลฟาของครอนบาค (Cronbach's alpha) ซึ่งเป็นการประมาณค่าความชันของเส้นการถดถอยของแต่ละกลุ่ม แต่ในวิธีซิปเทสท์ต้นฉบับเดิมจะคำนวณด้วย K-R 20 (Kuder-Richardson formula 20)

การวิเคราะห์องค์ประกอบเชิงยืนยันกลุ่มพหุ (Multiple-groups Confirmatory Factor Analysis: Multiple-groups CFA)

วิธีการวิเคราะห์องค์ประกอบเชิงยืนยัน (Confirmatory Factor Analysis: CFA) เป็นการวิเคราะห์องค์ประกอบที่ปรับปรุงจุดอ่อนของการวิเคราะห์องค์ประกอบเชิงสำรวจ (EFA) โดยมีประโยชน์ที่สำคัญ (นงลักษณ์ วิรัชชัย, 2540, หน้า 217-218) ดังนี้

1. ช่วยลดจำนวนตัวแปรจากตัวแปรจำนวนมากที่มีความร่วมกันหรือมีความสัมพันธ์กัน ทำให้เกิดปัญหาในการวิเคราะห์ข้อมูลและการสรุปผลการวิเคราะห์ องค์ประกอบที่สร้างขึ้นใหม่จะประกอบด้วยค่าความร่วมกันของตัวแปรต่าง ๆ ทำให้สามารถหาค่าขององค์ประกอบแต่ละหน่วยตัวอย่างได้และเรียกค่าขององค์ประกอบว่า ค่าคะแนนองค์ประกอบ ซึ่งมีผลให้สามารถนำองค์ประกอบที่สร้างขึ้นไปวิเคราะห์เทคนิคทางสถิติอื่น ๆ ต่อไป

2. จากองค์ประกอบที่สร้างขึ้นทำให้สามารถแก้ปัญหาที่ข้อสมมติหรือเงื่อนไขของเทคนิคการวิเคราะห์ทางสถิติบางเทคนิคที่ไม่เป็นจริง เช่น เทคนิคการวิเคราะห์ถดถอยเชิงพหุ การวิเคราะห์จำแนกกลุ่ม การวิเคราะห์ถดถอยโลจิสติก ซึ่งมีเงื่อนไขว่าตัวแปรอิสระจะต้องไม่มีความสัมพันธ์กัน แต่ในทางปฏิบัติมักพบว่าตัวแปรอิสระหลาย ๆ ตัว มักมีความสัมพันธ์กัน วิธีการแก้ปัญหาวิธีหนึ่งก็คือ การใช้เทคนิคการวิเคราะห์องค์ประกอบเพื่อรวมตัวแปรที่มีความสัมพันธ์กันไว้ในองค์ประกอบเดียวกัน และถ้าสามารถทำให้แต่ละองค์ประกอบไม่มีความสัมพันธ์กันแล้วจะสามารถใช้องค์ประกอบที่สร้างขึ้นใหม่เป็นตัวแปรอิสระในเทคนิคดังกล่าวข้างต้น โดยที่องค์ประกอบต่าง ๆ นั้นไม่มีความสัมพันธ์กัน ทำให้สามารถแก้ปัญหาของเงื่อนไขของเทคนิคดังกล่าวได้

3. ทำให้ผู้ศึกษาทราบถึงโครงสร้างความสัมพันธ์ของตัวแปร ทำให้ทราบว่ามีตัวแปรใดบ้างที่มีความสัมพันธ์กันมากหรือมีความร่วมกันสูง ตัวแปรใดบ้างที่ไม่สัมพันธ์กันหรือมีความสัมพันธ์กันน้อย

4. เมื่อได้องค์ประกอบไปเป็นตัวแปรเพื่อวิเคราะห์ต่อไปนั้นเป็นสิ่งที่มีประโยชน์มากในทางปฏิบัติ เพราะทำให้สามารถอธิบายความหมายหรือเปรียบเทียบองค์ประกอบด้านต่าง ๆ ได้

5. ทำให้สามารถตรวจสอบหรือยืนยันโครงสร้างของตัวแปรว่ามีลักษณะเป็นไปตามที่คาดไว้หรือไม่ เช่น การวิเคราะห์ถดถอย สัมพันธ์ มีวัตถุประสงค์สำคัญ 4 ประการ คือ 1) เพื่อตรวจสอบทฤษฎีที่ใช้เป็นพื้นฐานในการวิเคราะห์องค์ประกอบ 2) เพื่อสำรวจและระบุองค์ประกอบ 3) ใช้เป็นเครื่องมือในการสร้างตัวแปรใหม่ และ 4) วิเคราะห์ข้อมูล

การวิจัยทางสังคมศาสตร์และพฤติกรรมศาสตร์ เทคนิควิธีที่มีความโดดเด่นและได้รับการยอมรับในเชิงการยืนยันโครงสร้างความสัมพันธ์ของโมเดลการวิจัยที่มีต่อข้อมูลเชิงประจักษ์ใน

กลุ่มประชากรที่มีคุณลักษณะที่แตกต่างกันตั้งแต่ 2 กลุ่มขึ้นไป คือ การวิเคราะห์กลุ่มพหุ (Multiple group analysis หรือ Multi-sample analysis)

จุดมุ่งหมายของการวิเคราะห์กลุ่มพหุ

นงลักษณ์ วิรัชชัย (2540) ได้อธิบายถึงการวิเคราะห์กลุ่มพหุและขั้นตอนการวิเคราะห์โมเดลกลุ่มพหุ ว่าการวิเคราะห์กลุ่มพหุมีจุดมุ่งหมายที่สำคัญในการวิเคราะห์ คือการตรวจสอบว่าโมเดลกรอบความคิด (Conceptual framework) ที่สร้างขึ้นจากทฤษฎีมีความสอดคล้องกับข้อมูลเชิงประจักษ์ของแต่ละกลุ่มหรือไม่อย่างไร โดยข้อตกลงที่สำคัญของการวิเคราะห์กลุ่มพหุคือ กลุ่มตัวอย่างแต่ละกลุ่มต้องเป็นอิสระจากกันและได้มาโดยการสุ่มจากประชากรแต่ละกลุ่ม และหน่วยตัวอย่างทุกหน่วยต้องเป็นสมาชิกของกลุ่มใดกลุ่มหนึ่งเพียงกลุ่มเดียวโดยไม่เป็นสมาชิกร่วมกันในสองกลุ่ม (Mutually exclusive) โดยหัวใจสำคัญของการวิเคราะห์กลุ่มพหุ คือ การวิเคราะห์ข้อมูลที่ได้รับการรวบรวมมาจากกลุ่มตัวอย่างทุกกลุ่ม โดยมีการกำหนดเงื่อนไขบังคับ (Constraints) ให้โมเดลที่เป็นกรอบความคิดในการวิจัยมีลักษณะเป็นแบบเดียวกันสำหรับการตรวจสอบความสอดคล้องกลมกลืนระหว่างโมเดลและข้อมูลเชิงประจักษ์ หากผลการวิเคราะห์พบว่า ค่าไค-สแควร์ (χ^2) ในการทดสอบความกลมกลืนมีค่าต่ำกว่าค่าวิกฤติอย่างไม่มีนัยสำคัญทางสถิติ จะสรุปว่า โมเดลที่สร้างขึ้นตามทฤษฎีสอดคล้องกับข้อมูลเชิงประจักษ์ทุกกลุ่ม และโมเดลมีลักษณะเป็นแบบเดียวกัน เรียกตามศัพท์สถิติวิเคราะห์โมเดลไม่แปรเปลี่ยน หรือมีความยืนยงระหว่างกลุ่ม (Invariance across group)

ขั้นตอนการวิเคราะห์โมเดลกลุ่มพหุ

ขั้นตอนการวิเคราะห์กลุ่มพหุ สรุปได้เป็น 3 ขั้นตอน ดังนี้

ขั้นที่ 1 การวิเคราะห์กลุ่มพหุไม่มีการกำหนดเงื่อนไขบังคับ เป็นการวิเคราะห์ข้อมูลจากกลุ่มตัวอย่างหลายกลุ่มเพื่อประมาณค่าพารามิเตอร์ในโมเดลลิสเรล υ แต่ละกลุ่มประชากรแยกกัน และเพื่อทดสอบว่าโมเดลสำหรับกลุ่มประชากรแต่ละกลุ่มนั้นสอดคล้องกลมกลืนกับข้อมูลเชิงประจักษ์ โดยผลที่ได้จากการวิเคราะห์จะมีรายงานดัชนีวัดระดับความกลมกลืนของทุกกลุ่มประชากรเป็นภาพรวม ซึ่งได้มาจากดัชนีวัดระดับความกลมกลืนจากกลุ่มประชากรแต่ละกกลุ่มรวมกัน ถ้าผลการวิเคราะห์ข้อมูลนั้นในขั้นนี้ได้ค่าไค-สแควร์รวมไม่มีนัยสำคัญทางสถิติ แสดงว่าโมเดลแต่ละกลุ่มประชากรสอดคล้องกลมกลืนกับข้อมูลเชิงประจักษ์ทุกกลุ่ม แต่หากค่าไค-สแควร์รวมมีนัยสำคัญทางสถิติ แสดงว่า โมเดลของประชากรอย่างน้อย 1 กลุ่ม ไม่สอดคล้องกลมกลืนกับข้อมูลเชิงประจักษ์ และต้องดำเนินการปรับโมเดลแล้ววิเคราะห์ใหม่เพื่อให้ได้โมเดลที่มีโครงสร้างความสัมพันธ์ระหว่างตัวแปรสอดคล้องกับข้อมูลเชิงประจักษ์

ขั้นที่ 2 การวิเคราะห์กลุ่มพหุมีการกำหนดเงื่อนไขบังคับ เป็นการวิเคราะห์จากกลุ่มตัวอย่างหลายกลุ่มต่อเนื่องจากขั้นตอนที่ 1 โดยมีการกำหนดเงื่อนไขบังคับเพื่อทดสอบความไม่แปรเปลี่ยนของโมเดลระหว่างกลุ่มประชากรแต่ละกลุ่ม และต้องทำการวิเคราะห์หลายครั้งตามจำนวนสมมติฐานที่ต้องการตรวจสอบ เพื่อสรุปว่าโมเดลมีความไม่แปรเปลี่ยนอย่างไรบ้างระหว่างกลุ่มประชากรในการทดสอบแต่ละครั้ง

ขั้นตอนที่ 3 การวิเคราะห์สรุป เป็นการวิเคราะห์คำนวณหาผลต่างของดัชนีวัดระดับความกลมกลืนที่ได้จากการทดสอบสมมติฐานในขั้นที่ 2 ระหว่างคู่ที่มีเงื่อนไขบังคับน้อยและคู่ที่มีเงื่อนไขบังคับมาก จากผลต่างของดัชนีวัดระดับความกลมกลืนที่ได้นำมาตีความหมายสรุปผลการวิเคราะห์เกี่ยวกับโมเดลกลุ่มพหุทั้งหมด อย่างไรก็ดี ในการวิเคราะห์ที่มีวัตถุประสงค์มุ่งตอบปัญหาว่า มีความไม่แปรเปลี่ยนระหว่างกลุ่มประชากรหรือไม่ อย่างไร การตีความหมายจะเน้นที่ลักษณะผลการทดสอบสมมติฐานว่าโมเดลที่ไม่แปรเปลี่ยนมีลักษณะอย่างไร แต่หากในการวิเคราะห์มุ่งตอบปัญหาเกี่ยวกับปฏิสัมพันธ์หรือสนใจตอบปัญหาเกี่ยวกับอิทธิพลของตัวแปรปรับ จะต้องตีความหมายเพิ่มเติมจากการวิเคราะห์ความไม่แปรเปลี่ยนให้สามารถตอบคำถามวิจัยได้ด้วย

การตรวจสอบการวัดความไม่แปรเปลี่ยนใน IRT

การใช้ IRT ในการตรวจสอบการวัดความไม่แปรเปลี่ยนมีประวัติอันยาวนาน (Millsap, Gunn, Everson, & Zautra, 2014, p. 366-382) ซึ่งแนวคิดทั่วไปเริ่มต้นจากโมเดล IRT ที่มีความสอดคล้องกับข้อมูลแบบข้ามกลุ่ม แล้วทำการตรวจสอบความไม่แปรเปลี่ยนโดยการกำหนดเงื่อนไขบังคับค่าพารามิเตอร์ในโมเดลตามลำดับขั้นที่ซ้อนกัน (Nested sequence) หรือทดสอบความแตกต่างของค่าพารามิเตอร์ในโมเดลของแต่ละกลุ่มโดยตรง ซึ่งวัตถุประสงค์แรกของการวิเคราะห์ความไม่แปรเปลี่ยน (Invariance) คือ การหาโมเดล IRT ที่มีความสอดคล้องพอดีกับข้อมูล โดยมี 2 สิ่งสำคัญและควรพิจารณาในการเลือกโมเดลคือ 1) ความเป็นธรรมชาติของมาตรวัดในการตอบสนองข้อสอบ และ 2) จำนวนตัวแปรแฝงของข้อสอบที่คาดคะเนไว้ และในการประยุกต์ใช้ IRT เพื่อตรวจสอบความไม่แปรเปลี่ยนมีลำดับขั้นตอนดังนี้

ขั้นตอนที่ 1 การเลือกโมเดลพื้นฐาน โดยเป้าหมายของการวิเคราะห์ความไม่แปรเปลี่ยนคือการศึกษาโมเดล IRT ที่มีความเหมาะสมสอดคล้องกับข้อมูลในทุกกลุ่ม และสิ่งสำคัญในการเลือกโมเดลก็คือ 1) ธรรมชาติของมาตรวัดรายการคำตอบของข้อสอบ เช่น มาตรวัดของรายการคำตอบของข้อสอบเป็นแบบสองค่า (จริงหรือไม่จริง) หรือแบบหลายค่า (มาตรวัดลิเคิร์ท) 2) จำนวนตัวแปรแฝงภายใต้ข้อคำถามที่คาดหวัง ซึ่งการวิเคราะห์องค์ประกอบมักถูกใช้เป็นเครื่องมือในการตรวจสอบมิติในทางจิตวิทยา ซึ่งการวิเคราะห์องค์ประกอบเชิงสำรวจ (Exploratory

Factor Analysis: EFA) เป็นวิเคราะห์องค์ประกอบพื้นฐานเพื่อสำรวจมิติข้อมูลในระดับข้อคำถาม แต่การวิเคราะห์ CFA จะให้ข้อมูลสารสนเทศที่เหมาะสมหลากหลายทั้งรายละเอียดการประมาณค่าเศษเหลือเมื่อโมเดลมีความไม่สอดคล้องกับข้อมูล ดังนั้น ถ้าโมเดล CFA มีความสอดคล้องกับข้อมูล โมเดล IRT ก็ควรจะมีความสอดคล้องกับข้อมูลด้วย โดยในการศึกษาของ Millsap, Gunn, Everson, and Zautra (2014) ภายใต้โมเดล CFA กำหนดเงื่อนไขบังคับดังนี้

1. กำหนดค่าเฉลี่ยองค์ประกอบ (Factor mean) เท่ากับศูนย์ ในกลุ่มเพศชาย และประมาณค่าพารามิเตอร์อิสระในกลุ่มเพศหญิง

2. กำหนดค่าความแปรปรวนขององค์ประกอบ (Factor variance) เท่ากับ 1 ในกลุ่มเพศชาย และประมาณค่าพารามิเตอร์อิสระในกลุ่มเพศหญิง

3. กำหนดค่าน้ำหนักองค์ประกอบ (Factor loading) ในแบบสอบที่ 1 และแบบสอบที่ 2 บังคับให้มีความไม่แปรเปลี่ยน โดยไม่บังคับค่าใด ๆ และกำหนดค่าพารามิเตอร์ Threshold ในแบบสอบที่ 1 บังคับให้มีความไม่แปรเปลี่ยน

4. กำหนดค่าความแปรปรวนขององค์ประกอบ (Factor variance) ของข้อคำถาม 3 ข้อ ในแบบสอบที่ 1 บังคับให้มีความไม่แปรเปลี่ยน

5. กำหนดค่าพารามิตอร์ Theshold ที่ 1 และ Theshold ที่ 2 ของข้อคำถามในแบบสอบที่ 2 บังคับให้มีความไม่แปรเปลี่ยน

การกำหนดเงื่อนไขตามข้อที่ 1 ถึง ข้อที่ 5 ใช้สำหรับการวิเคราะห์ Multiple-groups CFA (Millsap and Tein, 2004; Millsap, 2011 cited in Millsap, Gunn, Everson, & Zautra, 2014, p. 372)

ขั้นตอนที่ 2 การประเมินการวัดความไม่แปรเปลี่ยน เป็นการทดสอบความไม่แปรเปลี่ยนของการกำหนดเงื่อนไขบังคับ โดยพิจารณากค่าพารามิเตอร์น้ำหนักองค์ประกอบใน CFA และค่าพารามิเตอร์อำนาจจำแนกใน MIRT (Multidimensional IRT)

ขั้นตอนที่ 3 ขนาดอิทธิพลและการตัดสิน ในการศึกษาการวัดขนาดอิทธิพลในแบบสอบมิติเดียวที่มีข้อคำถามแบบให้คะแนนหลายค่าที่มีการละเมิดความไม่แปรเปลี่ยนยังมีมากนัก และหากความไม่แปรเปลี่ยนถูกละเมิด ฟังก์ชันคะแนนที่คาดหวังจะมีความแตกต่างระหว่างกลุ่ม

งานวิจัยที่เกี่ยวข้องกับการทำหน้าที่ต่างกันของข้อสอบ (DIF)

อาวีพร ปานทอง (2558) ศึกษาการเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบให้คะแนนหลายค่าโดยวิธีทดสอบอัตราส่วนความควรจะเป็น วิธีเบส์เซียน และวิธีโพลี-ซิปเทสท์ โดยใช้การจำลองข้อมูลโมเดลพาเชียลเครดิตทั่วไป (Generalized partial credit model) ตามทฤษฎีการตอบสนองข้อสอบแบบมิติเดียว (Unidimensional item response theory) ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ ความยาวของแบบสอบ 3 รูปแบบ ขนาดของ

การทำหน้าที่ต่างกันของข้อสอบ 3 ขนาด ความแตกต่างของการแจกแจงความสามารถ 2 ระดับ และขนาดของกลุ่มตัวอย่าง 3 รูปแบบ เพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ รวมจำนวน 54 เงื่อนไข (3x3x3x2) และในแต่ละเงื่อนไขจำลองข้อมูลทวนซ้ำ 500 รอบ พบว่า วิธีทดสอบอัตราส่วนความควรจะเป็น และวิธีเบส์เซียน มีอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ใกล้เคียงกัน ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนหลายค่าที่มีรูปแบบเดียวกัน ทั้งอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 สูงกว่าวิธีโพลี-ซิปเทสท์ ภายใต้เงื่อนไขปัจจัยที่แปรเปลี่ยน และเมื่อความยาวของข้อสอบเพิ่มขึ้น วิธีทดสอบอัตราส่วนความควรจะเป็น และวิธีเบส์เซียน สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดีกว่าวิธีโพลี-ซิปเทสท์ และเมื่อขนาดของกลุ่มตัวอย่างเพิ่มขึ้นมีผลทำให้ทุกวิธี มีอำนาจการทดสอบเพิ่มขึ้นในทุกปัจจัย

สุพัฒนา หอมบุปผา (2556) ศึกษาการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN ข้อมูลที่ใช้ในการวิเคราะห์เป็นคะแนนการสอบวัดผลสัมฤทธิ์ทางการเรียนเพื่อประเมินคุณภาพการศึกษาระดับชาติ ของนักเรียนชั้นประถมศึกษาปีที่ 3 ในรายวิชาภาษาไทย คณิตศาสตร์ และวิทยาศาสตร์ ของสำนักทดสอบการศึกษา กระทรวงศึกษาธิการ ปีการศึกษา 2553 จำนวน 1,000 คน จำแนกเป็นเพศชายและเพศหญิง ที่อยู่ในเขตกรุงเทพมหานครและปริมณฑล และนอกเขตกรุงเทพมหานครและปริมณฑล โดยใช้โปรแกรมสำเร็จรูป 3 โปรแกรม ได้แก่ โปรแกรม HLM โปรแกรม Mplus และโปรแกรม WinBUGS พบว่าการวิเคราะห์ค่าพารามิเตอร์ความยากของข้อสอบ ค่าพารามิเตอร์ความสามารถของผู้สอบ และผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ในวิชาภาษาไทย คณิตศาสตร์ และวิทยาศาสตร์ ด้วยวิธี HGLM-2L วิธี MIMIC และวิธี BAYESIAN มีความสัมพันธ์กันในระดับสูงมาก อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และวิธี HGLM-2L เป็นวิธีที่ตรวจพบการทำหน้าที่ต่างกันของข้อสอบมากที่สุด และวิธี MIMIC เป็นวิธีที่ตรวจพบการทำหน้าที่ต่างกันของข้อสอบน้อยที่สุด

อรินทร์ น่วมถนอม (2549) ศึกษาการเปรียบเทียบวิธีโพลี-ซิปเทสท์ (Poly-SIBTEST) วิธีการถดถอยโลจิสติกแบบจัดอันดับ (Ordinal logistic regression) และวิธีการถดถอยโลจิสติกแบบจัดอันดับหลายมิติ (Multidimensional ordinal logistic regression) ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่วัดความสามารถหลายมิติและให้คะแนนหลายค่า โดยการจำลองข้อมูลภายใต้โมเดลพาเชียลเครดิตทั่วไปแบบหลายมิติ จำลองผลการตอบจากแบบสอบที่วัดความสามารถ 2 มิติ จำนวน 40 ข้อ ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ รูปแบบของข้อสอบทำหน้าที่ต่างกัน 2 รูปแบบ สัดส่วนของข้อสอบทำหน้าที่ต่างกัน 3 ขนาด ความแตกต่างของการแจกแจงความสามารถ 3 ระดับ และขนาดกลุ่มตัวอย่าง 4 ขนาด รวมข้อมูลที่ใช้ในการศึกษา จำนวน 72 เงื่อนไข วิเคราะห์

ข้อมูลในแต่ละเงื่อนไขด้วยวิธีโพลี-ซิปเทสท์ วิธีการถดถอยโลจิสติกแบบจัดอันดับ และวิธีการถดถอยโลจิสติกแบบจัดอันดับหลายมิติ พบว่า วิธีการถดถอยโลจิสติกแบบจัดอันดับและวิธีการถดถอยโลจิสติกแบบจัดอันดับหลายมิติ มีอัตราความถูกต้องใกล้เคียงกัน การตรวจสอบการทำน้าที่ต่างกันของข้อสอบแบบเอกรูปและแบบอเนกรูป มีอัตราความถูกต้องสูงกว่าวิธีโพลี-ซิปเทสท์ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป สัดส่วนของข้อสอบทำหน้าที่ต่างกันของข้อสอบ ไม่มีผลต่อวิธีโพลี-ซิปเทสท์และวิธีการถดถอยโลจิสติกแบบจัดอันดับหลายมิติ แต่มีผลต่อวิธีการถดถอยโลจิสติกแบบจัดอันดับ และเมื่อความแตกต่างของการแจกแจงความสามารถเพิ่มขึ้น วิธีโพลี-ซิปเทสท์ สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 สูงเกินปกติได้ดีกว่าวิธีอื่น และเมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้นมีผลทำให้ทุกวิธีมีอัตราความถูกต้องเพิ่มขึ้นเกือบทุกเงื่อนไข

อุทัยวรรณ สายพัฒนะ (2547) ศึกษาการเปรียบเทียบประสิทธิภาพของผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบสอบที่มีการให้คะแนนแบบหลายค่า โดยมีปัจจัยเงื่อนไขความยาวของแบบสอบและกลุ่มตัวอย่างที่แตกต่างกันระหว่างวิธี GMH กับวิธี Polytomous SIBTEST พบว่า การตรวจสอบ DIF ด้วยวิธี GMH ในเงื่อนไขความยาวของแบบสอบ จำนวน 40 ข้อ, 30 ข้อ และ 20 ข้อ และกลุ่มตัวอย่างขนาด 1,000 คน, 500 คน และ 250 คน ส่งผลต่อความถูกต้องและความผิดพลาดในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแตกต่างกัน ยกเว้น แบบสอบที่มีความยาว จำนวน 30 ข้อ และ 20 ข้อ กลุ่มตัวอย่างขนาด 500 คน และ 250 คน ส่งผลต่อความผิดพลาดในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ไม่แตกต่างกัน และการตรวจสอบ DIF ด้วยวิธี Polytomous SIBTEST ในเงื่อนไขความยาวของแบบสอบ จำนวน 40 ข้อ, 30 ข้อ และ 20 ข้อ และกลุ่มตัวอย่างขนาด 1,000 คน, 500 คน และ 250 คน ส่งผลต่อความถูกต้องและความผิดพลาดในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแตกต่างกัน ยกเว้น แบบสอบที่มีความยาว จำนวน 20 ข้อ กลุ่มตัวอย่างขนาด 1,000 คน และ 500 คน ส่งผลต่อความผิดพลาดในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ไม่แตกต่างกัน และเพื่อพิจารณาการเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธี GMH และวิธี Polytomous SIBTEST ในทุกเงื่อนไข พบว่า ประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบของทั้งสองวิธี ไม่แตกต่างกัน ยกเว้นกรณีกลุ่มตัวอย่างขนาด 1,000 คนในแบบสอบขนาด 20 ข้อ วิธี Polytomous SIBTEST มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูงกว่าวิธี GMH

วลีมาศ แซ่อึ้ง (2543) ศึกษาการเปรียบเทียบอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูประหว่างวิธี

ซิปเทสท์ปรับใหม่ วิธีซิปเทสท์ วิธีแมนเทล-แฮนส์เซล และวิธีการถดถอยโลจิสติก ข้อมูลที่ใช้ในการศึกษาจำลองภายใต้โมเดลโลจิสติกแบบ 3 พารามิเตอร์ชนิดกำหนดค่าการคาดเดา (c) คงที่ แล้วจัดกระทำข้อมูลตามปัจจัย 4 ตัว คือ (1) ลักษณะของข้อสอบที่มีค่าความความยาก (b) และค่าอำนาจจำแนก (a) ระดับต่ำ ปานกลาง และสูง (2) ความยาวของแบบทดสอบ 2 ระดับ (3) สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบ 3 ระดับ และ (4) ขนาดกลุ่มตัวอย่าง 6 ระดับ รวมข้อมูลที่ศึกษาทั้งหมด 324 เงื่อนไข แล้วนำข้อมูลของแต่ละเงื่อนไขมาคำนวณค่าอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป พบว่า อำนาจการทดสอบในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอนเกรูปของวิธีซิปเทสท์ปรับใหม่และวิธีการถดถอยโลจิสติกมีค่าเท่าเทียมกันภายใต้เกือบทุกเงื่อนไข และทั้งสองวิธีดังกล่าว มีอำนาจการทดสอบสูงกว่าวิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซลภายใต้เกือบทุกเงื่อนไข ส่วนอัตราความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปของวิธีซิปเทสท์ปรับใหม่ วิธีซิปเทสท์ วิธีแมนเทล-แฮนส์เซล และวิธีการถดถอยโลจิสติก มีค่าอยู่ภายในเกณฑ์ของอัตราความคลาดเคลื่อนประเภทที่ 1 ที่ระดับ 10% เกือบทุกเงื่อนไข

นิคม กีรติวากูร (2542) ศึกษาการเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธีการวิเคราะห์องค์ประกอบจำกัด (RFA) วิธีแมนเทล-แฮนส์เซล (MH) และวิธีการตอบสนองข้อสอบ (IRT) แบบ 2 พารามิเตอร์ โดยเปรียบเทียบกับเกณฑ์ที่กำหนดและศึกษาจากข้อมูลจำลอง ซึ่งปัจจัยที่ศึกษาคือ ขนาดกลุ่มตัวอย่าง 2 ขนาด คือ ขนาดเล็ก (300 คน) และขนาดใหญ่ (1,000 คน) ความยาวแบบทดสอบ 2 ขนาด คือ แบบสอบสั้น (25 ข้อ) และแบบสอบยาว (75 ข้อ) ค่าความยากของข้อสอบแบ่งออกเป็น 3 ระดับคือ กลุ่มข้อสอบที่มีความยากสูง ปานกลางและต่ำ ขนาดความลำเอียงของข้อสอบแบ่งออกเป็น 2 ขนาด คือ กลุ่มข้อสอบที่มีความลำเอียงสูงและต่ำ พบว่า วิธี RFA มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูงที่สุด โดยวิธี MH มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูงภายใต้เงื่อนไขแบบสอบที่มีค่าความยากต่ำ ค่าอำนาจจำแนกสูง ส่วนวิธี IRT แบบ 2 พารามิเตอร์ มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูงภายใต้เงื่อนไขแบบสอบที่มีค่าความยากต่ำ และวิธี IRT มีอัตราความคลาดเคลื่อนประเภทที่ 1 สูงกว่าวิธี MH และวิธี RFA ตามลำดับ

ญาณภัทร สีหะมงคล (2540) ศึกษาการเปรียบเทียบความสอดคล้องของผลการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันระหว่างวิธี Lord's χ^2 วิธี Raju's area measures และวิธี Closed interval area เมื่อขนาดของกลุ่มตัวอย่าง ความยาวของแบบทดสอบ และสัดส่วนจำนวน

ข้อสอบที่ทำหน้าที่ต่างกันในแบบทดสอบต่างกัน ข้อมูลที่ใช้ในการศึกษาเป็นผลการสอบประเมินคุณภาพและความก้าวหน้าทางการศึกษา วิชาคณิตศาสตร์ของนักเรียนชั้นประถมศึกษาปีที่ 4 ปีการศึกษา 2536 ของสำนักงานการประถมศึกษาแห่งชาติ จำนวน 11,404 คน เครื่องมือที่ใช้เป็นแบบทดสอบแบบเลือกตอบ จำนวน 80 ข้อ ผลการศึกษาพบว่า จำนวนข้อสอบที่ทำหน้าที่ต่างกันจากการตรวจสอบด้วยวิธีการทั้งสามแตกต่างกันเมื่อขนาดของกลุ่มตัวอย่างและความยาวของแบบทดสอบต่างกัน ส่วนความสัมพันธ์ระหว่างวิธีการทั้งสามมีค่าสัมประสิทธิ์สหสัมพันธ์ค่อนข้างสูงมากและมีนัยสำคัญทางสถิติเกือบทุกเงื่อนไขของการศึกษา และสำหรับความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบส่วนมากจะมีค่าปานกลางถึงต่ำเกือบทุกเงื่อนไขของการศึกษา

Wood (2011) ศึกษาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าเมื่อขนาดของกลุ่มผู้สอบมีขนาดเล็ก เพื่อศึกษาอัตราความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบด้วยวิธีการ 3 วิธี คือ วิธีแมนเทล (Mantel test/ Cox's β) วิธีสถิติทดสอบ Liu-Agrsti และวิธีสถิติทดสอบ HW3 โดยการจำลองข้อูลภายใต้เงื่อนไขขนาดของกลุ่มตัวอย่างที่มีขนาดเล็ก 2 รูปแบบ คือ ขนาดกลุ่มตัวอย่าง 40: 40 คน ในกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ และขนาดกลุ่มตัวอย่าง 400: 40 คน ในกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ วิธีแมนเทล (Mantel test/ Cox's β) วิธีสถิติทดสอบ Liu-Agrsti มีอำนาจการทดสอบสูง

Penfield and Algina (2006) ศึกษาการประมาณค่าอิทธิพลของการทำหน้าที่ต่างกันของแบบสอบ (DTF) โดยไม่คิดเครื่องหมายในแบบสอบแบบผสม (Mixed format test) โดยการจำลองข้อมูล ประกอบด้วย ข้อสอบที่ตรวจให้คะแนนแบบ 2 ค่า (Dichotomously) และข้อสอบที่ตรวจให้คะแนนแบบหลายค่า (Polytomously) จำนวน 4 ตัวเลือก การวิเคราะห์ข้อมูลแบ่งออกเป็น 2 กรณี คือ กรณีแรก แบบสอบที่ประกอบด้วยข้อสอบที่ตรวจให้คะแนนแบบ 2 ค่า จำนวน 20 ข้อ และข้อสอบที่ตรวจให้คะแนนแบบหลายค่า จำนวน 8 ข้อ กรณีที่ 2 แบบสอบที่ประกอบด้วยข้อสอบที่ตรวจให้คะแนนแบบหลายค่า จำนวน 8 ข้อ และข้อสอบที่ตรวจให้คะแนนแบบหลายค่า จำนวน 12 ข้อ โดยข้อสอบที่ตรวจให้คะแนนแบบ 2 ค่า วิเคราะห์แบบ 3 พารามิเตอร์ สำหรับข้อสอบที่ตรวจให้คะแนนแบบหลายค่า วิเคราะห์ด้วยวิธีแมนเทล-แฮนด์เซลแบบทั่วไป (GMH) กลุ่มตัวอย่าง จำนวน 1,000 คน แบ่งเป็นกลุ่มอ้างอิง จำนวน 500 คน และกลุ่มเปรียบเทียบ จำนวน 500 คน โดยพิจารณาจากการแจกแจงแบบปกติที่มีค่าส่วนเบี่ยงเบนมาตรฐานเป็น 1 และค่าเฉลี่ยขึ้นอยู่กับเงื่อนไข จำนวน 40 เงื่อนไข (2 ระดับของค่าเฉลี่ยการแจกแจงความสามารถ x2 ชนิดของแบบสอบ x2 พารามิเตอร์โอกาสในการเดา x5 ขนาดอิทธิพลของการทำหน้าที่กัน) พบว่าแบบสอบที่มีข้อสอบที่ตรวจให้คะแนนแบบ 2 ค่า จำนวนมากจะส่งผลต่อความลำเอียงทางลบ

แต่แบบสอบที่มีข้อที่ตรวจให้คะแนนแบบหลายค่าจำนวนมากจะส่งผลต่อความลำเอียงทางบวกเพียงเล็กน้อยเท่านั้น

Stark et al. (2006) ศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการวิเคราะห์องค์ประกอบเชิงยืนยัน (Confirmatory factor analysis) และวิธีการทดสอบอัตราส่วนไลค์ลิฮู้ด (LR) ตามทฤษฎีการตอบสนองข้อสอบ โดยใช้ข้อมูลการจำลองที่ตั้งอยู่บนพื้นฐานของวิธี Mean and covariance structures (MACS) โดยใช้แบบวัดมิติเดียว จำนวนข้อคำถาม 15 ข้อ มีตัวแปรที่เกี่ยวข้อง จำนวน 8 ตัวแปร จำนวนเงื่อนไขที่ใช้จำลองข้อมูล คือ 320 เงื่อนไข ($2_6 + 2_8$) แต่ละเงื่อนไขจะมีการทำซ้ำ 50 ครั้ง การวิเคราะห์ด้วย MACS ใช้วิธีโปรแกรมลิสเรล 8 และการทดสอบด้วยวิธี LR อยู่บนพื้นฐานของโมเดล Graded reponse โดยใช้โปรแกรม MULTILOG ผลการศึกษาพบว่า IRT ด้วยวิธี LR ให้ผลดีกว่าวิธี MACS ในกลุ่มตัวอย่างขนาดเล็ก และในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบวัดมิติเดียวเมื่อกลุ่มตัวอย่างขนาดเล็ก การวิเคราะห์ MACS ให้ผลดีกว่า

Chang et al. (1996) ศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรจให้คะแนนแบบหลายค่า โดยใช้วิธี SIBTEST เปรียบเทียบกับวิธีแมนเทล-แฮนส์เซล และวิธี SMD ซึ่งแบ่งการศึกษาออกเป็น2 ส่วน คือ ส่วนที่ 1 จำลองข้อมูลเพื่อเรียบเทียบวิธี SIBTEST แบบประยุกต์กับวิธีแมนเทล-แฮนส์เซล และวิธี SMD ผลการศึกษาพบว่า วิธี SIBTEST มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ดี ส่วนที่ 2 ใช้การจำลองข้อสอบที่มีอำนาจจำแนกแตกต่างกัน 11 ค่า (ตั้งแต่ 1.5-2.00) ขนาดของกลุ่มตัวอย่างที่แตกต่างกัน (500, 1,000 และ 2,000 คน) ความยาวของแบบสอบ 24 ข้อ สำหรับวิธี SIBTEST และ 25 ข้อ สำหรับวิธีแมนเทล-แฮนส์เซล และวิธี SMD ผลการศึกษาพบว่า วิธีแมนเทล-แฮนส์เซล และวิธี SMD มีความคลาดเคลื่อนประเภทที่ 1 ค่อนข้างสูง เมื่อค่าอำนาจจำแนกของข้อสอบมีค่าแตกต่างจากค่าเฉลี่ยของค่าอำนาจจำแนกของแบบสอบที่มีความตรง และอัตราการปฏิเสธของทั้งสามวิธี จะมีค่าสูงขึ้นเมื่อมีค่าอำนาจจำแนกสูงขึ้น

Mellor (1995) ศึกษาการเปรียบเทียบวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ด้วยวิธีการตรวจสอบ 4 วิธี คือ วิธีแมนเทล-แฮนส์เซล (GMH) วิธี Poly-SIBTEST วิธี LR วิธี CRL(Continuation-Ratio Logistic: CRL) และวิธีการวิเคราะห์ฟังก์ชันการจำแนกโลจิสติส (Logistic Discriminant Facuntion Analysis: LDFA) โดยการจำลองข้อมูลภายใต้เงื่อนไขการทำหน้าที่ต่างกัน 6 เงื่อนไข และแต่ละวิธีจำลองวนซ้ำ 100 รอบ และนำไปศึกษากับข้อมูลจริง ผลการศึกษาพบว่า วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ด้วยวิธี GMH และวิธี Poly-SIBTEST

ให้ผลดีกว่าวิธี CRL และวิธี LDFA เล็กน้อย กรณี การแจกแจงความสามารถของผู้สอบไม่เท่ากัน วิธี GMH จะให้ผลการตรวจสอบดีกว่าและสามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดี

งานวิจัยที่เกี่ยวข้องกับการวิเคราะห์องค์ประกอบเชิงยืนยันแบบหลายกลุ่ม (MGCFA)

Steinmetz, Schmidt, Tina-Booh, Wieczorek, and Schwartz (2009) ศึกษาความไม่แปรเปลี่ยนในการวัดแบบสอบโดยใช้ Multigroup CFA: ความแตกต่างระหว่างกลุ่มการเรียนการสอนในการวัดคุณค่าความเป็นมนุษย์ ซึ่งการศึกษาครั้งนี้เป็นการประยุกต์ใช้กระบวนการวัดความไม่แปรเปลี่ยนโดยใช้วิธีการวิเคราะห์องค์ประกอบเชิงยืนยันแบบหลายกลุ่ม (MGCFA) ซึ่งกระบวนการดังกล่าวเป็นการตรวจสอบโครงสร้างองค์ประกอบและความไม่แปรเปลี่ยนของโครงสร้างองค์ประกอบของแบบสอบถาม Portraits value (Portraits Value Questionaire: PVQ) ของ Schwartz et al. (2001)โดยศึกษาข้ามกลุ่ม 3 กลุ่มการศึกษา โดยกลุ่มตัวอย่าง มีจำนวน 1,677 คน และการวัดในแบบสอบถาม PVQ เป็นการวัดคุณค่าพื้นฐาน จำนวน 10 ข้อ ตามสมมติฐานของ Schwartz et al. (2001) ที่ครอบคลุมคุณค่าความเป็นมนุษย์ในสังคม ได้แก่ ความสำเร็จ (Achievement) ความชอบ (Hedonism) การกำกับตนเอง (Self-direction) ความเมตตากรุณา (Benevolence) ความสอดคล้อง (Conformity) ความปลอดภัย (Security) การกระตุ้น (Stimulation) อำนาจ (Power) ประเพณี (Tradition) และ ความเป็นสากล (Universalism) และทำการประมาณค่าและเปรียบเทียบค่าเฉลี่ยตัวแปรแฝง ของทั้ง 3 กลุ่มการศึกษา ผลการวิเคราะห์พบว่ามีความไม่แปรเปลี่ยนของคุณค่าความเป็นมนุษย์ และความไม่แปรเปลี่ยนของค่าพารามิเตอร์ และค่าเฉลี่ยของตัวแปรแฝง แสดงให้เห็นว่า ผู้ตอบแบบสอบถามที่มีการศึกษาน้อยจะให้ความสำคัญกับคุณค่าของความปลอดภัย ประเพณี และความสอดคล้อง มากกว่า

Kim, Chohen, and Kim (2007) ศึกษาการทำหน้าที่ต่างกันของขนาดอิทธิพลของข้อสอบที่ตรวจให้คะแนนแบบหลายค่า (Polytomous) โดยมีกลุ่มตัวอย่างขนาดใหญ่ (N = 105,731) เพื่อเปรียบเทียบความสอดคล้องตามวิธีการตรวจสอบ 5 วิธี คือ วิธีการทดสอบอัตราส่วนไลด์ลิฮู้ดแบบ IRT วิธีการถดถอยโลจิสติก วิธีการทดสอบอัตราส่วนไลด์ลิฮู้ด (Likelihood ratio test) วิธีแมนเทล วิธีแมนเทล-แฮนส์เซลแบบทั่วไป (GMH) โดยใช้โปรแกรม MULTILOG และโปรแกรม IRTLRDIF วิเคราะห์ด้วยวิธีการทดสอบอัตราส่วนไลด์ลิฮู้ด ส่วนวิธีแมนเทลและวิธีแมนเทล-แฮนส์เซลแบบทั่วไป เขียนโปรแกรมด้วยภาษาฟอร์แทน ผลการวิจัยพบว่า สามารถตรวจพบตรวจพบข้อสอบที่ทำหน้าที่ต่างกันจากทั้ง 5 วิธี และข้อค้นพบที่สำคัญ คือ การใช้กลุ่มตัวอย่างที่มีขนาดใหญ่เกินไปจะไม่มีประโยชน์ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

Oishi (2006) ศึกษาตรวจสอบความเท่าเทียมกันของการวัดความพึงพอใจด้วยแบบวัดความพึงพอใจในชีวิตระหว่างกลุ่มตัวอย่างชาวอเมริกันและชาวจีนโดยใช้ Multigroup Structural Eqation Modeling (SEM), Multiple Indicator Multiple Cause Model (MIMIC) และทฤษฎีการตอบสนองข้อสอบ (IRT) กลุ่มตัวอย่างเป็นนักศึกษา จำนวน 556 คน ในสถาบันเทคโนโลยี Zhejiang ประเทศสาธารณรัฐประชาชนจีน ผู้วิจัยให้นักศึกษาที่ลงทะเบียนเรียนวิชาจิตวิทยาเบื้องต้นในมหาวิทยาลัยอิลินอยส์ ทำแบบสอบถามในชั้นเรียนโดยใช้แบบวัด SWLS ที่ใช้ประเมินความพึงพอใจในชีวิตของคนทั่วโลก โดยแบบวัดนี้ประกอบด้วย 5 ข้อคำถาม สเกลการตอบสนองข้อคำถามมี 7 ระดับ โดยเรียงจากสเกล 1 (ไม่เห็นด้วยมากที่สุด) ถึง 7 (เห็นด้วยมากที่สุด) ผลการวิจัยพบว่า การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบแสดงให้เห็นถึงแง่มุมความแตกต่างของวิธีการแบบดั้งเดิมในการวัดประเด็นการวิจัยทางวัฒนธรรมและความเป็นอยู่ที่ดี การวิเคราะห์ IRT แสดงให้เห็นถึง ความแตกต่างของค่าเฉลี่ยระหว่างกลุ่มชาวจีนกับชาวอเมริกัน ข้อสอบที่มีความลำเอียงจะให้คะแนนค่าน้ำหนักน้อยกว่า ดังนั้นความแตกต่างของค่าเฉลี่ยที่ค้นพบก่อนหน้าระหว่างกลุ่มชาวจีนกับชาวอเมริกัน อาจจะไม่ค้นพบได้ง่ายในข้อสอบที่มีความลำเอียง สุดท้ายการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ (IRT) แสดงข้อมูลที่เกี่ยวข้องกับแนวคิดของความพึงพอใจในชีวิต การศึกษาวิจัยในครั้งแสดงให้เห็นความสำคัญและประโยชน์ของการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในโครงสร้างอื่น ๆ ซึ่งจะเป็นประโยชน์ในอนาคต

ตารางที่ 3 สรุปผลการศึกษางานวิจัยที่ศึกษาเกี่ยวกับวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ภายใต้เงื่อนไขต่าง ๆ (ขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง) ในช่วงปี พ.ศ. 2540-2558

ปี พ.ศ./ค.ศ.	ผู้วิจัย	ประเด็นที่ศึกษา	ผลการศึกษา
งานวิจัยภายในประเทศ			
2540	ญาณภัทร สีหะมงคล	เปรียบเทียบความสอดคล้องของผลการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันระหว่าง 1) วิธี Lord's χ^2 2) วิธี Raju's Area Measures 3) วิธี Closed Interval Area	จำนวนข้อสอบที่ทำหน้าที่ต่างกันจากการตรวจสอบด้วยวิธีการทั้งสามวิธี แตกต่างกันเมื่อขนาดของกลุ่มตัวอย่างและความยาวของแบบสอบต่างกัน
2540	พรรณี จิตมาศ	วิเคราะห์ความลำเอียงต่อเพศของแบบสอบคณิตศาสตร์ด้วย 3 วิธี คือ 1) วิธีแปลงค่าความยาก 2) วิธีแมนเทล-แฮนส์เซล 3) วิธีซิปเทสท์	วิเคราะห์จากกลุ่มตัวอย่างขนาด 500 คน วิธีซิปเทสท์พบข้อสอบที่มีความลำเอียงมากที่สุดและเมื่อวิเคราะห์จากกลุ่มผู้สอบขนาด 1,000 คน วิธีแมนเทล-แฮนส์เซลพบข้อสอบมีความลำเอียงมากที่สุด
2540	รัชรินทร์ มุคดา	เปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของแบบสอบอเนกรูปในกรณีที่จัดกลุ่มความสามารถ ค่าความยากของข้อสอบ ค่าอำนาจจำแนกของข้อสอบต่างกัน ด้วยวิธีแมนเทล-แฮนส์เซล และวิธีถดถอยโลจิสติก	วิธีแมนเทล-แฮนส์เซล กับวิธีถดถอยโลจิสติกมีประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของแบบสอบอเนกรูปเท่ากัน ในกลุ่มผู้สอบที่มีความสามารถสูง ปานกลาง และต่ำ

ตารางที่ 3 (ต่อ)

ปี พ.ศ./ ค.ศ.	ผู้วิจัย	ประเด็นที่ศึกษา	ผลการศึกษา
2542	นิคม กีรติวรางกูร	เปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ระหว่างวิธีการวิเคราะห์องค์ประกอบจำกัด (RFA) วิธีแมนเทล-แฮนส์เซล (MH) และวิธีการตอบสนองข้อสอบ (IRT) แบบ 2 พารามิเตอร์	วิธี RFA มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูงที่สุด และวิธี IRT มีอัตราความคลาดเคลื่อนประเภทที่ 1 สูงกว่าวิธีแมนเทล-แฮนส์เซล และวิธี RFA และวิธี IRT แบบ 2 พารามิเตอร์ มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูงภายใต้เงื่อนไขแบบสอบที่มีความยากต่ำ ที่ขนาดความยาวของแบบสอบ 75 ข้อ เมื่อใช้กลุ่มตัวอย่าง 1,000 คน
2543	อารี วัชรโสตถิกุล	เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้รูปแบบและวิธีการตรวจสอบต่างกัน ด้วยวิธีซิปเทสท์ กับวิธีถดถอยโลจิสติก	จำนวนข้อสอบที่ทำหน้าที่ต่างกันโดยใช้รูปแบบต่างกัน (รูปแบบคะแนนรวมทั้งฉบับ แยกตามเนื้อหา และแยกตามระดับพฤติกรรม) แตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติทั้งวิธีวิธีซิปเทสท์ กับวิธีถดถอยโลจิสติก

ตารางที่ 3 (ต่อ)

ปี พ.ศ./ ค.ศ.	ผู้วิจัย	ประเด็นที่ศึกษา	ผลการศึกษา
2543	ทองอยู่ สาระ	เปรียบเทียบอำนาจการตรวจสอบและจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอและแบบไม่สม่ำเสมอระหว่างวิธีแมนเทล-แฮนส์เซลและวิธีถดถอยโลจิสติก	ความยาวของแบบสอบไม่มีผลต่ออำนาจการตรวจสอบและการจำแนกผิดพลาดทั้ง 2 วิธี แต่เมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้น อำนาจการตรวจสอบจะเพิ่มขึ้น
2543	วลีมาศ แซ่อึ้ง	เปรียบเทียบอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป ระหว่าง 1) วิธีซิปเทสท์ปรับใหม่ 2) วิธีซิปเทสท์ 3) วิธีแมนเทล-แฮนส์เซล 4) วิธีถดถอยโลจิสติก	อำนาจการทดสอบในการตรวจสอบข้อสอบแบบอเนกรูปของวิธีซิปเทสท์ปรับใหม่และวิธีถดถอยโลจิสติกมีค่าเท่าเทียมกันเกือบทุกเงื่อนไขและทั้งสองวิธีมีอำนาจการทดสอบสูงกว่าวิธีซิปเทสท์ และวิธี MH ภายใต้เกือบทุกเงื่อนไข ส่วนอัตราความคลาดเคลื่อนประเภทที่ 1 ทั้ง 4 วิธี มีค่าอยู่ภายในเกณฑ์
2545	สิริรัตน์ วิภาสศิลป์	เปรียบเทียบวิธีซิปเทสท์และดีเอฟไอที (DFIT) ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบหมวดข้อสอบและแบบสอบ จากข้อมูลการตอบข้อสอบที่ใช้ความสามารถหลายมิติ	เมื่อแบบสอบประกอบด้วยข้อสอบ 30, 40 และ 50 ให้ผลแตกต่างกันในเงื่อนไขขนาดกลุ่มตัวอย่างแตกต่างกันและวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีซิปเทสท์ ภายใต้กลุ่มตัวอย่างที่มีขนาด 50, 100 และ 200 คน ส่งผลต่อความถูกต้องในการตรวจสอบการทำ

ตารางที่ 3 (ต่อ)

ปี พ.ศ./ ค.ศ.	ผู้วิจัย	ประเด็นที่ศึกษา	ผลการศึกษา
			หน้าที่ต่างกันของข้อสอบไม่แตกต่างกัน และเมื่อกลุ่มตัวอย่างเพิ่มขึ้นเป็น 500 และ 1,000 คน ส่งผลต่อความถูกต้องและการระบุผิดพลาดในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูงกว่ากลุ่มตัวอย่างที่มีขนาด 50, 100 และ 200 คน
2547	อุทัยวรรณ สายพัฒนะ	เปรียบเทียบประสิทธิภาพของผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบที่มีการให้คะแนนแบบหลายค่า ระหว่างวิธี GMH และวิธี Polytomous SIBTEST ภายใต้เงื่อนไขความยาวของแบบสอบและกลุ่มตัวอย่างที่แตกต่างกัน	การเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ระหว่างวิธี GMH และวิธี Polytomous SIBTEST ในทุกเงื่อนไขความยาวของแบบทดสอบและกลุ่มตัวอย่างทุกขนาด ส่งผลต่อประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการตรวจสอบทั้งสองวิธี ไม่แตกต่างกัน ยกเว้นเงื่อนไขกลุ่มตัวอย่างที่มีขนาด 1,000 คน ของแบบทดสอบขนาด 20 ข้อ วิธี Polytomous SIBTEST มีประสิทธิภาพในการตรวจสอบสูงกว่าวิธี GMH

ตารางที่ 3 (ต่อ)

ปี พ.ศ./ ค.ศ.	ผู้วิจัย	ประเด็นที่ศึกษา	ผลการศึกษา
2549	อรินทร์ น่วมถนอม	เปรียบเทียบวิธีโพลี-ซิปเทสท์ วิธีการถดถอยโลจิสติกแบบจัดอันดับ และวิธีการถดถอยโลจิสติกแบบจัดอันดับหลายมิติและให้คะแนนหลายค่า โดยการจำลองข้อมูลภายใต้โมเดลพาเชียลเครดิตทั่วไปแบบหลายมิติ ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ รูปแบบของข้อสอบทำหน้าที่ต่างกัน 2 รูปแบบ สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน 3 ขนาด ความแตกต่างของการแจกแจงความสามารถ 3 ระดับ และขนาดกลุ่มตัวอย่าง 4 ขนาด และจำลองข้อมูลวนซ้ำ 50 ครั้ง	วิธีการถดถอยโลจิสติกแบบจัดอันดับ และวิธีการถดถอยโลจิสติกแบบจัดอันดับหลายมิติ มีอัตราความถูกต้องใกล้เคียงกัน ภายใต้เงื่อนไขปัจจัยรูปแบบของข้อสอบที่ทำหน้าที่ต่างกันที่เป็นรูปแบบเดียวกันและไม่เป็นรูปแบบเดียวกัน และมีอัตราความถูกต้องสูงกว่าวิธีโพลี-ซิปเทสท์ ภายใต้เงื่อนไขปัจจัยข้อสอบที่ทำหน้าที่ต่างกันที่ไม่เป็นรูปแบบเดียวกัน และสัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันไม่มีผลต่อวิธีโพลี-ซิปเทสท์และวิธีการถดถอยโลจิสติกแบบจัดอันดับหลายมิติ แต่มีผลต่อวิธีการถดถอยโลจิสติกแบบจัดอันดับ และเมื่อความแตกต่างของการแจกแจงความสามารถเพิ่มขึ้น วิธีโพลี-ซิปเทสท์ สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ดีกว่าวิธีอื่น และเมื่อขนาดของกลุ่มตัวอย่างเพิ่มขึ้น มีผลทำให้ทุกวิธีมีอัตราความถูกต้องเพิ่มขึ้น ภายใต้เงื่อนไขเกือบทุกเงื่อนไข

ตารางที่ 3 (ต่อ)

ปี พ.ศ./ ค.ศ.	ผู้วิจัย	ประเด็นที่ศึกษา	ผลการศึกษา
2554	ธเกียรติกมล ทองงอก	เปรียบเทียบอัตราความถูกต้องและอัตราความคลาดเคลื่อนประเภทที่ 1 ของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบทวิวิภาค โดยการจำลองข้อมูลและข้อมูลเชิงประจักษ์ ในวิธีถดถอยโลจิสติก ระหว่างการวัดขนาดอิทธิพลตามเกณฑ์ Jodoin and Gierl กับเกณฑ์ Zumbo and Thomas	วิธีถดถอยโลจิสติก โดยการวัดขนาดอิทธิพลตามเกณฑ์ Jodoin and Gierl มีอัตราความถูกต้องในการตรวจสอบการทำหน้าที่คต่างกันของข้อสอบสูงกว่าเกณฑ์ Zumbo and Thomas เกือบทุกเงื่อนไข โดยข้อสอบที่ทำหน้าที่ต่างกันแบบอนเกรูปมีอัตราความถูกต้องจากการวัดขนาดอิทธิพลทั้ง 2 เกณฑ์ สูงกว่าแบบเอกรูป และแบบสอบที่มีจำนวนข้อสอบทำหน้าที่ต่างกันทั้งฉบับคิดเป็นร้อยละ 20 มีอัตราความถูกต้องจากการวัดขนาดอิทธิพลทั้ง 2 เกณฑ์ สูงกว่าในแบบสอบที่มีจำนวนข้อสอบที่ทำหน้าที่ต่างกันทั้งฉบับคิดเป็นร้อยละ 10 และเมื่อขนาดอิทธิพลของข้อสอบที่การทำหน้าที่ต่างกันเพิ่มขึ้น มีผลทำให้อัตราความถูกต้องจากการวัดขนาดอิทธิพลทั้ง 2 เกณฑ์ เพิ่มขึ้นเกือบทุกเงื่อนไข

ตารางที่ 3 (ต่อ)

ปี พ.ศ./ ค.ศ.	ผู้วิจัย	ประเด็นที่ศึกษา	ผลการศึกษา
2558	อาวีพร ปานทอง	เปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบให้คะแนนหลายค่า โดยวิธีทดสอบอัตราส่วนความควรจะเป็น วิธีเบส์เซียน และวิธีโพลี-ซิปเทสท์ ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ ความยาวของแบบสอบ ขนาดการทำหน้าที่ต่างกันของข้อสอบ ความแตกต่างของการแจกแจงความสามารถของผู้ตอบข้อคำถาม และขนาดของกลุ่มตัวอย่าง โดยใช้โมเดลพาเชียลเครดิตทั่วไป และใช้การจำลองข้อมูลวนซ้ำ 500 รอบ	วิธีทดสอบอัตราส่วนความควรจะเป็น และวิธีเบส์เซียน มีอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ใกล้เคียงกัน ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ห้คะแนนหลายค่าที่เป็นรูปแบบเดียวกัน และมีอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 สูงกว่าวิธีโพลี-ซิปเทส์ และเมื่อความยาวของข้อสอบเพิ่มขึ้น วิธีทดสอบอัตราส่วนความควรจะเป็น และวิธีเบส์เซียน สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดีกว่าวิธีโพลี-ซิปเทสท์ และเมื่อขนาดของกลุ่มตัวอย่างเพิ่มขึ้น มีผลทำให้ทุกวิธีมีอำนาจการทดสอบเพิ่มขึ้นทุกเงื่อนไขปัจจัย

ตารางที่ 3 (ต่อ)

ปี พ.ศ./ ค.ศ.	ผู้วิจัย	ประเด็นที่ศึกษา	ผลการศึกษา
งานวิจัยในต่างประเทศ			
1990	Swaminathan and Roger	เปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธีถดถอยโลจิสติกกับวิธีแมนเทล-แฮนส์เซล	วิธีถดถอยโลจิสติกและวิธีแมนเทล-แฮนส์เซล ให้ผลการตรวจสอบการทำหน้าที่ต่างกันแบบเอกรูปตรวจสอบได้ถูกต้องร้อยละ 100 กรณีกลุ่มตัวอย่าง 500 คน ในทุกความยาวของแบบสอบ สำหรับการตรวจสอบการทำหน้าที่ต่างกันแบบอเนกรูป พบว่า วิธีแมนเทล-แฮนส์เซล ตรวจสอบได้เล็กน้อย ส่วนวิธีถดถอยโลจิสติก ตรวจสอบถูกต้องร้อยละ 50 กรณีกลุ่มตัวอย่างน้อย ข้อสอบสั้นถูกต้องร้อยละ 75 กรณีแบบสอบยาวและกลุ่มตัวอย่างขนาดใหญ่
1992	Mazor, Clauser, and Hambleton	ศึกษาผลกระทบของขนาดกลุ่มตัวอย่างที่มีต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมนเทล-แฮนส์เซล	1. เมื่อตัวอย่างขนาดใหญ่ขึ้นจะสามารถระบุข้อสอบที่ทำหน้าที่ต่างกันได้ถูกต้องได้มากขึ้น 2. ข้อสอบที่ไม่สามารถตรวจสอบพบหรือระบุว่าทำหน้าที่ต่างกันได้ เนื่องมาจากข้อสอบเหล่านั้นมีความยากมากหรือมีความยากต่างกันเพียงเล็กน้อยระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบอีกทั้งเป็นข้อที่มีค่าอำนาจจำแนกต่ำ

ตารางที่ 3 (ต่อ)

ปี พ.ศ./ ค.ศ.	ผู้วิจัย	ประเด็นที่ศึกษา	ผลการศึกษา
1994	Mazor, Clauser, and Hambleton	ศึกษาการใช้วิธีแมนเทล-แฮนส์เซล ในการตรวจสอบลการทำหน้าที่ต่างกันของข้อสอบ	ภายใต้เงื่อนไขข้อสอบที่ทำหน้าที่ต่างกัน ค่าความยาก ค่าอำนาจจำแนก ค่าโอกาสในการเดา การกระจาย ความสามารถและปฏิสัมพันธ์ระหว่างการกระจายความสามารถกับค่าพารามิเตอร์ของแบบสอบมีผลต่อการประมาณค่า α_{MH} และข้อสอบที่ทำหน้าที่ต่างกันส่วนใหญ่เป็นข้อสอบ DIF แบบเอกรูปมากกว่าแบบอเนกรูป
1996	Narayanan and Swaminathan	ศึกษาเปรียบเทียบประสิทธิภาพการตรวนจสอบการทำหน้าที่ต่างกันแบบอเนกรูประหว่าง 1) วิธีแมนเทล-แฮนส์เซล 2) วิธีถดถอยโลจิสติก 3) วิธีโครซิปท์ (CRO-SIB)	วิธีถดถอยโลจิสติกและวิธีโครซิปท์ ให้ผลในการตรวจสอบการทำหน้าที่ต่างกันแบบอเนกรูปใกล้เคียงกันและทั้ง 2 วิธีตรวจจับการทำหน้าที่ต่างกันได้ดีกว่าวิธีแมนเทล-แฮนส์เซล ปัจจัยที่ส่งผลต่อ DIF แบบอเนกรูป คือ ขนาดกลุ่มตัวอย่าง เมื่อเพิ่มขนาดกลุ่มตัวอย่างทั้ง 3 วิธีสามารถตรวจสอบได้มากขึ้น การกระจายความสามารถระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบแบบเท่ากัน ทำให้ตรวจสอบได้มากขึ้น พื้นที่ความแตกต่างระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ

ตารางที่ 3 (ต่อ)

ปี พ.ศ./ ค.ศ.	ผู้วิจัย	ประเด็นที่ศึกษา	ผลการศึกษา
			เพิ่มขึ้นจาก .04 เป็น 1.0 ทั้ง 3 วิธี สามารถตรวจสอบได้มากขึ้น ข้อสอบที่พบว่าทำหน้าที่ต่างกันด้วยวิธีถดถอยโลจิสติกและวิธีโครซิปท์ ส่วนใหญ่เป็นข้อสอบที่มีค่าความยากต่ำ ค่าอำนาจจำแนกสูง ส่วนวิธีแมนเทล-แฮนส์เซล ตรวจสอบข้อสอบที่ DIF แบบอเนกรูปได้ดีเฉพาะกรณีข้อสอบยากและข้อสอบง่ายซึ่งโค้งลักษณะข้อสอบ (ICC) ของผู้สอบ 2 กลุ่มตัดกันที่ระดับความสามารถสูงหรือความสามารถต่ำเท่านั้น
1996	Roussos and Stout	ศึกษาผลของกลุ่มตัวอย่างขนาดเล็กที่มีต่อความคลาดเคลื่อนนิดที่ 1 ระหว่าง วิธีซิปเทสท์ กับวิธีแมนเทล-แฮนส์เซล	ค่าสถิติของวิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล มีแนวโน้มที่จะมีความคลาดเคลื่อนชนิดที่ 1 เพิ่มขึ้นเมื่อขนาดกลุ่มตัวอย่างมีความแตกต่างของค่าเฉลี่ยระหว่างกลุ่มเพิ่มขึ้น ส่วนการศึกษาครั้งที่ 2 กลุ่มตัวอย่าง 500, 1,000 และ 3,000 คน ความแตกต่างของค่าเฉลี่ยการกระจายความสามารถระหว่างกลุ่มเป็น 0 และ 1.0 ค่าอำนาจจำแนก 3 ระดับ ค่าความยาก 5 ระดับ ค่าโอกาสในการเดา 3 ระดับ

ตารางที่ 3 (ต่อ)

ปี พ.ศ./ ค.ศ.	ผู้วิจัย	ประเด็นที่ศึกษา	ผลการศึกษา
1996	Chang et al.	ได้เปรียบเทียบประสิทธิภาพของวิธีซิปเทสท์ปรับใหม่กับวิธี Mantel และวิธี Standardized mean difference (SMD)	ค่าความยาก 5 ระดับ ค่าโอกาสในการเดา 3 ระดับ พบว่าเมื่อความแตกต่างของค่าเฉลี่ยการกระจายความสามารถเป็น 1.0 ความคลาดเคลื่อนชนิดที่ 1 เพิ่มขึ้นทั้ง 2 วิธี วิธี SIBTEST ตรวจสอบ DIF ได้ดี แต่วิธี Mantel และ SMD ดีกว่า SIBTEST เล็กน้อย วิธี SIBTEST สามารถควบคุมผลกระทบที่ก่อให้เกิดอัตราความคลาดเคลื่อนประเภทที่ 1 ได้เหนือกว่าวิธี Mantel และวิธี SMD ภายใต้เงื่อนไขทั่ว ๆ ไปของการตรวจสอบ DIF เมื่อขนาดของกลุ่มตัวอย่างเพิ่มขึ้น อำนาจในการตรวจสอบของวิธี Mantel และ SMD จะเพิ่มขึ้นอย่างรวดเร็ว

ตารางที่ 3 (ต่อ)

ปี พ.ศ./ ค.ศ.	ผู้วิจัย	ประเด็นที่ศึกษา	ผลการศึกษา
1996	French and Miller	ศึกษาความเป็นไปได้ของการใช้วิธีถดถอยโลจิสติกในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนแบบพหุวิภาค	เมื่อกลุ่มตัวอย่างมีขนาดเล็กลง อำนาจในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบลดลงและเมื่อพารามิเตอร์อำนาจจำแนกของข้อสอบยิ่งแตกต่างกันมาก อำนาจในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปยิ่งเพิ่มขึ้น
1997	Flowers	ศึกษาการบรรยาย DFIT ข้อข้อสอบที่ให้คะแนนแบบ Polytomous และประเมินรวมถึงเปรียเทียบการทำ DFIT ในการแผ่ขยายของขั้นตอน SIBTEST และ Lord's chi-square	อัตราความคลาดเคลื่อนประเภทที่ 1 ใกล้เคียงกับระดับแอลฟา ยกเว้นเมื่อจำนวนข้อสอบที่มี DIF 20% และจำนวน DIF ที่มากที่สุดมีค่าสูง ปัจจัย ที่มีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 1 คือค่าของพารามิเตอร์
1997	Oshima, Raju, and Flowers	ศึกษาการทำหน้าที่ต่างกันของข้อสอบและแบบสอบแบบพหุมิติ (Multidimensional DIF) โดยใช้กรอบแนวคิดของวิธีดีเอฟไอที (DFIT) ในการศึกษานี้ใช้ข้อมูลจำลองจากโมเดลโลจิสติกแบบพหุมิติ แบบ 2 พารามิเตอร์ (M2PL)	ข้อสอบที่จำลองเพิ่มขึ้นไม่ทำหน้าที่ต่างกัน (No DIF) นอกจากนี้ยังพบว่า เมื่อค่าความยากของทั้ง 2 มิติแตกต่างกันจะทำให้ค่าดัชนี CDIF และค่าดัชนี NCDIF มีค่าเพิ่มมากขึ้น หากค่าความยากของทั้ง 2 มิติแตกต่างกันในทิศทางตรงกันข้ามจะทำให้ค่าดัชนี CDIF มีค่าเท่ากัน

ตารางที่ 3 (ต่อ)

ปี พ.ศ./ ค.ศ.	ผู้วิจัย	ประเด็นที่ศึกษา	ผลการศึกษา
			ค่าความยากของทั้ง 2 มิติแตกต่างกันแต่เป็นไปในทิศทางเดียวกันจะทำให้ค่าดัชนี NCDIF จะมีค่าเพิ่มขึ้น
1998	Oshima, Raju, Flowers, and Slinde	ศึกษาสาเหตุของการทำหน้าที่ต่างกันของกลุ่มข้อสอบ (Differential Bundle Functioning: DFIT-DBF) โดยแบ่งข้อสอบออกเป็นกลุ่ม ๆ ที่แตกต่างกัน	เมื่อเปรียบเทียบระหว่างเพศหญิงและเพศชาย เมื่อตัดข้อสอบซึ่งเป็นข้อที่ตรวจสอบพบว่าเกิดการทำหน้าที่ต่างกัน (DIF) ออกจากแบบสอบแล้วตรวจสอบการทำหน้าที่ต่างกันของแบบสอบ (DTF) แบบสอบไม่ทำหน้าที่ต่างกัน นั่นคือ ดัชนี DTF ไม่แตกต่างจากศูนย์อย่างมีนัยสำคัญ เมื่อเปรียบเทียบเศษฐานะทางสังคม พบว่าไม่พบข้อสอบที่ทำหน้าที่ต่างกัน (No DIF) และเมื่อแบ่งวิเคราะห์ตามกลุ่มข้อสอบ พบว่ากลุ่มข้อสอบที่ 5 มีค่าดัชนี NCDIF สูงที่สุดโดยเข้าข้างเพศชายมากกว่าเพศหญิงแต่ค่าดัชนี bundle NCDIF มีค่าไม่แตกต่างกันเมื่อแบ่งกลุ่มตามเศรษฐานะทางสังคม

ตารางที่ 3 (ต่อ)

ปี พ.ศ./ ค.ศ.	ผู้วิจัย	ประเด็นที่ศึกษา	ผลการศึกษา
2000	Kim	ศึกษาเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่าง 1) วิธีการทดสอบอัตราส่วนไลค์ลิฮู้ด (Likelihood Ratio Test) 2) วิธีแมนเทล 3) วิธีแมนเทล-แฮนส์เซลแบบทั่วไป (GMH)	ทั้ง 3 วิธี ให้ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ดีเมื่อกลุ่มตัวอย่างขนาด 100 คน ยังมีข้อค้นพบที่สำคัญคือการใช้กลุ่มตัวอย่างที่มีขนาดใหญ่เกินไปจะไม่มีประโยชน์ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ
2001	Penfield	ศึกษาการทำหน้าที่ต่างกันของข้อสอบในหลายกลุ่มด้วยวิธีแมนเทล-แฮนส์เซล 3 แบบ โดยมีวัตถุประสงค์เพื่อศึกษาขนาดของความคลาดเคลื่อนชนิดที่ 1 เมื่อมีกลุ่มตัวอย่างที่ศึกษาพร้อมกันหลายกลุ่ม เปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบด้วย 1) วิธีแมนเทล-แฮนส์เซลแบบไคสแควร์ที่ไม่ปรับระดับของ α 2) วิธีแมนเทล-แฮนส์เซลแบบไคสแควร์ที่ไม่ปรับระดับของ α ด้วย Bonferroni 3) วิธีแมนเทล-แฮนส์เซลแบบทั่วไป	วิธีแมนเทล-แฮนส์เซลแบบทั่วไปดีที่สุดในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในหลายกลุ่ม

ตารางที่ 3 (ต่อ)

ปี พ.ศ./ ค.ศ.	ผู้วิจัย	ประเด็นที่ศึกษา	ผลการศึกษา
2001	Walker and Beretvas	ศึกษาการสืบสอบเชิงประจักษ์กระบวนทัศน์ของการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบแบบพหุมิติ: การอธิบายทางพุทธิปัญญาสำหรับการทำหน้าที่ต่างกันของข้อสอบ เป็นการศึกษาเพื่อเปรียบเทียบผลการทำหน้าที่ต่างกันของข้อสอบแบบเอกมิติ (Unidimensional) กับผลการทำหน้าที่ต่างกันของข้อสอบแบบพหุมิติ (Multidimensional) ว่าแบบใดจะสอดคล้องกับข้อมูลเชิงประจักษ์มากกว่ากัน ระหว่าง 1) วิธีโพลีซิปเทสท์ และ 2) วิธี LISREL	การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบแบบพหุมิติสอดคล้องกับข้อมูลเชิงประจักษ์มากกว่าการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบแบบเอกมิติ
2002	Bolt	เปรียบเทียบการตรวจจับการทำหน้าที่ต่างกันของข้อสอบที่ตรวจให้คะแนนแบบหลายค่าด้วยสถิติพาราเมติก และนันพาราเมตริก เป็นการศึกษาโดยการจำลองข้อมูลด้วยเทคนิคมอลติคาร์โล ระหว่าง1) วิธี GRM 2) วิธี GRM-LR และ 3) วิธี GRM-DFIT	1. วิธี GRM ให้ผลที่สอดคล้องกับข้อมูลเชิงประจักษ์มากที่สุด 2. วิธี GRM-LR ให้ค่าความคลาดเคลื่อนประเภทที่ 1 สูงกว่าวิธีอื่น ๆ 3. วิธี GRM-DFIT ให้ค่าความคลาดเคลื่อนประเภทที่ 1 ต่ำกว่าวิธีอื่น

ตารางที่ 3 (ต่อ)

ปี พ.ศ./ ค.ศ.	ผู้วิจัย	ประเด็นที่ศึกษา	ผลการศึกษา
2003	Gierl, Bisanz, Bisanz, and Boughton	ศึกษาการระบุเนื้อหาและทักษะทางพุทธิปัญญาที่ทำให้เกิดความแตกต่างทางเพศที่มีต่อวิชาคณิตศาสตร์โดยใช้กระบวนทัศน์ของการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบแบบพหุมิติ (Multidimensional DIF) ในข้อสอบที่ให้คะแนนแบบ 2 ค่า กำหนดความสามารถทางคณิตศาสตร์เป็นมิติความสามารถที่ 1 ความสามารถทางพุทธิปัญญาเป็นมิติความสามารถที่ 2 วิเคราะห์ข้อมูลด้วยวิธิซิปเทสท์ และ DIMTEST โดยวิธีซิปเทสท์ในการตรวจสอบเพื่อหาขนาดของการเกิดการทำหน้าที่ต่างกันของข้อสอบ ส่วนวิธีดิมเทสท์ ใช้ในการวิเคราะห์มิติของแบบสอบ	ข้อสอบบางข้อเข้าข้างนักเรียนชาย ส่วนข้อสอบบางข้อเข้าข้างนักเรียนหญิง โดยนักเรียนชายทำคะแนนในส่วนของมิติสัมพันธ์ (Spatial) ได้ดีกว่านักเรียนหญิงในขณะที่นักเรียนหญิงทำคะแนนในส่วนของทักษะความจำ (Memorization) ได้ดีกว่านักเรียนชาย

ตารางที่ 3 (ต่อ)

ปี พ.ศ./ ค.ศ.	ผู้วิจัย	ประเด็นที่ศึกษา	ผลการศึกษา
2005	Cohen and Bolt	วิเคราะห์โมเดลแบบผสมในการทำหน้าที่ต่างกันของข้อสอบโดยวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบตามเพศและใช้แบบสอบแบบผสม (Mixed format test) ตามแนวคิด IRT ด้วยวิธีการทดสอบอัตราส่วนไลด์ลิฮู้ด (Likelihood ratio test) ด้วยโปรแกรม Multilog	พบการทำหน้าที่ต่างกันของข้อสอบโดยมีข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 5 ข้อ โดยข้อสอบจำนวน 4 ข้อ เข้าข้างเพศชายและข้อสอบอีก 1 ข้อ เข้าข้างเพศหญิง
2005	Su and Wang	จำลองข้อมูลในการสืบสอบปัจจัยที่มีผลต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิธีแมนเทล วิธีแมนเทล-แฮนส์เซลทั่วไป วิธี Logistic Discriminant Function Analysis (LDFA)	ทั้ง 3 วิธีมีการควบคุมค่าอัตราความคลาดเคลื่อนประเภทที่ 1 ได้เป็นอย่างดี วิธี Mantel และวิธี LDFA มีอำนาจการตรวจสอบสูงกว่าวิธี LDFA
2005	Finch	ศึกษาเปรียบเทียบในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โมเดล MIMIC โดยการจำลองข้อมูลจำนวนผู้สอบและจำนวนข้อสอบด้วยวิธีมอลติคาร์โล ระหว่าง 1) วิธีซิปเทสท์ 2) วิธีแมนเทล-แฮนส์เซล 3) วิธีการทดสอบอัตราส่วนความควรจะเป็นแบบ IRT (IRT Likelihood Ratio Test)	โมเดล MIMIC ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ดีในกรณีที่ข้อสอบมีจำนวน 50 ข้อ แบบ 2 พารามิเตอร์ และโมเดล MIMIC สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้สูงในกรณีที่ข้อสอบมีจำนวน 20 ข้อ แบบ 3 พารามิเตอร์ โลจิสติก ส่วนความคลาดเคลื่อนชนิดที่ 1 มีค่าต่ำสุดในวิธีแมนเทล-แฮนส์เซล นอกจากนี้ยังได้ข้อค้นพบ

ตารางที่ 3 (ต่อ)

ปี พ.ศ./ ค.ศ.	ผู้วิจัย	ประเด็นที่ศึกษา	ผลการศึกษา
			ว่า วิธีซิปเทสท์ ให้ผลคล้ายวิธีแมนเทล-แฮนส์เซล แต่มีขนาดของความคลาดเคลื่อนชนิดที่ 1 สูงกว่า
2006	Lei, Chen, & Yu	ศึกษาการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบแบบปรับเหมาะโดยใช้คอมพิวเตอร์ การจำลองข้อมูลการทำหน้าที่ต่างกันของข้อสอบทั้งแบบมีทิศทางและไม่มีทิศทางระหว่าง 1) วิธีถดถอยโลจิสติก 2) วิธีการทดสอบอัตราส่วนความควรจะเป็นแบบ IRT และ 3) วิธีแคทซิบ	วิธีถดถอยโลจิสติกและวิธีการทดสอบอัตราส่วนความควรจะเป็น แบบ IRT ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้งแบบมีทิศทาง และแบบไม่มีทิศทางได้ดีเท่ากันและทั้ง 2 วิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ดีกว่าวิธีแคทซิบ ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบมีทิศทางได้ดีกว่าแบบไม่มีทิศทาง
2006	Park	ตรวจสอบ DIF แบบเอกรูปและแบบอเนกรูปจากข้อสอบด้านภาษาและเพศในการทดสอบการเขียนความเรียง MELAB วัดความสามารถภาษาอังกฤษ รัฐมิชิแกน สหรัฐอเมริกา ด้านการวัดทักษะการอ่าน การฟังและการไวยากรณ์ด้วยวิธีถดถอยโลจิสติกแบบ 3 ขั้นตอน	ไม่เกิดการทำหน้าที่ต่างกันของแบบสอบ MELAB

ตารางที่ 3 (ต่อ)

ปี พ.ศ./ ค.ศ.	ผู้วิจัย	ประเด็นที่ศึกษา	ผลการศึกษา
2006	Penfield and Algina	ศึกษาการประมาณค่าอิทธิพลของการทำหน้าที่ต่างกันของแบบสอบ (DTF) โดยไม่คิดเครื่องหมายในแบบสอบแบบผสม (Mixed format test) ศึกษาจำลองข้อมูล ประกอบด้วยข้อสอบที่ตรวจให้คะแนนแบบ 2 ค่า และแบบหลายค่า 4 ตัวเลือก การวิเคราะห์ข้อมูลแบ่งออกเป็น 2 กรณี คือ แบบสอบที่ประกอบด้วยข้อสอบที่ตรวจให้คะแนนแบบ 2 ค่า 20 ข้อ และแบบหลายค่าจำนวน 8 ข้อ กรณีที่ 2 แบบสอบที่ตรวจให้คะแนนแบบ 2 ค่า 8 ข้อ และข้อสอบที่ให้คะแนนแบบหลายค่า 12 ข้อ โดยข้อสอบ 2 ค่าวิเคราะห์แบบ 3 พารามิเตอร์ ส่วนข้อสอบหลายค่าวิเคราะห์ด้วยวิธีแมนเทล-แฮนส์เซลแบบทั่วไป (GMH)กลุ่มตัวอย่างจำนวน 1,000 คน แบ่งเป็นกลุ่มอ้างอิง 500 คน และกลุ่มเปรียบเทียบ 500 คน พิจารณาจากการแจกแจงแบบปกติที่มีส่วนเบี่ยงเบนมาตรฐานเป็น 1 และค่าเฉลี่ยขึ้นอยู่กับเงื่อนไขทั้งหมด 40 เงื่อนไขที่แตกต่างกัน (2 ระดับค่าเฉลี่ย	แบบสอบที่มีข้อสอบที่ตรวจให้คะแนนแบบ 2 ค่า จำนวนมากจะส่งผลต่อความลำเอียงทางลบแต่แบบสอบที่มีข้อที่ตรวจให้คะแนนแบบหลายค่าจำนวนมากจะส่งผลต่อความลำเอียงทางบวกเพียงเล็กน้อยเท่านั้น

ตารางที่ 3 (ต่อ)

ปี พ.ศ./ ค.ศ.	ผู้วิจัย	ประเด็นที่ศึกษา	ผลการศึกษา
		การแจกแจงความสามารถ x2 ชนิดของแบบสอบ x2 พารามิเตอร์โอกาสในการเดา x5 ขนาดอิทธิพลของการทำหน้าที่ต่างกัน)	
2006	Oishi	ตรวจสอบความเท่าเทียมของการวัดความพึงพอใจด้วยแบบวัดความพึงพอใจในชีวิตระหว่างกลุ่มตัวอย่างชาวอเมริกันและชาวจีนโดยใช้ Multigroup Structural Equation Modeling (SEM), Multiple indicator multiple cause model (MIMIC) ทฤษฎีการตอบสนองข้อสอบ	การวิเคราะห์ IRT แสดงให้เห็นความแตกต่างของค่าเฉลี่ยระหว่างกลุ่มชาวจีนและชาวอเมริกัน
2006	Stark et al.	ได้พัฒนาและทดสอบแผนการร่วมที่ใช้ในการระบุการทำหน้าที่ต่างกันของข้อสอบ โดยใช้วิธี MACS และวิธี LR โดยใช้ข้อมูลจำลองในการตรวจสอบความเที่ยงตรงของทั้งสองวิธี	IRT วิธี LR ให้ผลดีกว่าวิธี MACS ในขณะที่กลุ่มตัวอย่างมีขนาดเล็กในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบวัดมิติเดียว เมื่อใช้กลุ่มตัวอย่างขนาดเล็ก การวิเคราะห์ MACS ให้ผลดีกว่า

ตารางที่ 3 (ต่อ)

ปี พ.ศ./ ค.ศ.	ผู้วิจัย	ประเด็นที่ศึกษา	ผลการศึกษา
2007	Kim, Cohen Alagoz, and Kim	ศึกษาการทำหน้าที่ต่างกันของขนาดอิทธิพลของข้อสอบที่ตรวจให้คะแนนแบบหลายค่า ใช้กลุ่มตัวอย่างขนาดใหญ่ (N = 105, 731) เพื่อเปรียบเทียบถึงความสอดคล้องตามวิธี 1. วิธีการทดสอบอัตราส่วนความควรจะเป็นแบบ IRT 2. วิธีถดถอยโลจิสติก 3. วิธีการทดสอบอัตราส่วนไลด์ลิฮู้ด 4. วิธีแมนเทล 5. วิธีแมนเทล-แฮนส์เซลแบบทั่วไป	ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้ง 10 ข้อจากทั้ง 5 วิธีและได้ข้อค้นพบที่สำคัญ คือการใช้กลุ่มตัวอย่างที่มีขนาดใหญ่เกินไปจะไม่มีประโยชน์ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ
2007	Elosua, and López-Jaúregui	ศึกษาแหล่งของการทำหน้าที่ต่างกันของข้อสอบที่ส่งผลต่อการแปลแบบสอบ การศึกษาในครั้งนี้มีวัตถุประสงค์เพื่อหาแหล่งของการทำหน้าที่ต่างกันของข้อสอบที่ส่งผลต่อการแปลแบบสอบในข้อคำถามที่ตรวจให้คะแนนแบบ 2 ค่า วิเคราะห์ข้อมูลด้วยวิธีแมนเทล-แฮนส์เซล และจากความเห็นของผู้เชี่ยวชาญ (Expert judgment)	เกณฑ์ทั้ง 4 แบบ คือ ความเกี่ยวข้องทางวัฒนธรรม (Cultural relevance) ปัญหาในการแปล (Translation problems) ไวยากรณ์ (Grammar) และการตีความหมายคำ (Semantic differences) ส่งผลต่อการทำหน้าที่ต่างกันของข้อสอบ วิธีแมนเทล-แฮนส์เซล ตรวจพบ 32 ข้อ ผู้เชี่ยวชาญตรวจสอบ พบ 28 ข้อ และมีข้อคำถามที่ทั้งผู้เชี่ยวชาญและ

ตารางที่ 3 (ต่อ)

ปี พ.ศ./ ค.ศ.	ผู้วิจัย	ประเด็นที่ศึกษา	ผลการศึกษา
			วิธีแมนเทล-แฮนส์เซล ตรวจสอบพบว่าเกิด DIF ตรงกัน 22 ข้อ การทำหน้าที่ต่างกันของข้อสอบมี 29 แหล่ง
2008	Walker, Zhang and Surber	ศึกษาการใช้กรอบแนวคิดกระบวนทัศน์ของการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบแบบพหุมิติในการตัดสินผลความสามารถในการอ่านที่ส่งผลต่อความสามารถทางคณิตศาสตร์ วิเคราะห์ข้อมูลด้วยโปรแกรม NOHARM	ความสามารถในการอ่านส่งผลต่อความสามารถทางคณิตศาสตร์ในทางบวก นั้นคือนักเรียนที่มีความสามารถในการอ่านสูงจะสามารถทำคะแนนในส่วนของคณิตศาสตร์ได้สูงด้วยและมีนักเรียนเพียงส่วนหนึ่งเท่านั้นที่มีความสามารถในการอ่านสูงแต่ทำคะแนนในส่วนของคณิตศาสตร์ได้ไม่ค่อยดี
2009	Wiberg	ศึกษาการทำหน้าที่ต่างกันของข้อสอบของแบบวัดความสามารถระดับสูง Mastery tests ทำการเปรียบเทียบ 3 วิธีโดยใช้ข้อมูลจริง เพื่อต้องการเปรียบเทียบวิธีการลอกลิเนียร์โมเดลโลจิสติก กรีเกรสชัน และวิธีแมนเทล-แฮนส์เซล สถิติในการตรวจสอบ DIF 1. ลอกลิเนียร์โมเดล (LLM) 2. วิธีถดถอยโลจิสติก	การตรวนสอบ DIF ในแบบวัดความสามารถระดับสูงข้อมูลที่ใช้ในการวิเคราะห์เป็นผลการสอบจากเครื่องมือ Swedish theory driving license test (SDLT) และ Mastery test ประกอบด้วยข้อสอบจำนวน 65 ข้อ ในระดับยากซึ่งผู้เข้าร่วมต้องทำข้อสอบได้อย่างน้อย 52 ข้อขึ้นไปถึงจะผ่านการทดสอบและจากผู้เข้าสอบ 5,404 คนและสุ่มคัดเลือกข้อสอบมา 15 ข้อ ที่ครอบคลุมหลักสูตร เพื่อนำมาตรวจสอบ

ตารางที่ 3 (ต่อ)

ปี พ.ศ./ ค.ศ.	ผู้วิจัย	ประเด็นที่ศึกษา	ผลการศึกษา
		3. วิธีแมนเทล-แฮนส์เซลใช้โปรแกรม R package (R-development-core-team, 2007)	DIF พบว่า มีความสัมพันธ์กันค่อนข้างสูงเกี่ยวกับขนาดของการเกิดการทำหน้าที่ต่างกันของข้อสอบ วิธีโลจิสติกรีเกรสชันและลอกลิเนียร์โมเดลจะมีประโยชน์ในการให้ค่าช่วงคะแนนในการสอบที่แน่นอนซึ่งถือเป็นสิ่งที่น่าสนใจเป็นพิเศษในแบบวัดความสามารถระดับสูงนี้ ซึ่งในการทดสอบคะแนนส่วนนี้วิธีการโลจิสติกกรีเกรสชันและวิธีแทนเทล-แฮนส์เซลให้ผลลัพธ์ที่แตกต่างกัน
2009	Kahraman, De Boeck, and Janssen	ศึกษารูปแบบ DIF จากข้อมูลที่มีผลการตอบซับซ้อน โดยใช้ยุทธวิธีในการออกแบบแบบสอบ สถิติที่ใช้ในการทดสอบ DIF ใช้ประมาณการปรับเหมาะของวิธีถดถอยโลจิสติก ภายใต้วิธีการของทฤษฎีการตอบสนองข้อสอบ (IRT Approach)	จุดหมายของการศึกษา เพื่อเสนอวิธีการสร้างรูปแบบของการตอบสนองข้อมูลพหุมิติกับกลุ่มโครงสร้างที่เกี่ยวข้องและปัจจัยหลักของกระบวนการประมาณค่าระดับพารามิเตอร์เพื่อขยายรวมผลกระทบของมิติของแบบสอบและปัจจัยจากกลุ่ม ความแตกต่างในสมรรถนะของการทำข้อสอบนอกเหนือจากกลุ่ม จำแนกความแตกต่างของ

ตารางที่ 3 (ต่อ)

ปี พ.ศ./ ค.ศ.	ผู้วิจัย	ประเด็นที่ศึกษา	ผลการศึกษา
			ของการเกิดการทำหน้าที่ต่างกันของข้อสอบ 2 ระดับ ใช้ข้อมูลจริงจากการสุ่มนักเรียนประถมศึกษาเกรด 3, 4 ระดับละ 269 คน ใช้แบบเขียนตอบคำศัพท์ที่กำหนดให้ ผลวิจัยพบว่าตัวอย่างประกอบนี้เป็นการนำเสนอการใช้มาตรวัดความเชี่ยวชาญหรือชำนาญในการสะกดคำของชาวดัสช์ จากสองกลุ่มย่อยคือปฏิสัมพันธ์ระหว่างกลุ่มกับข้อสอบและปฏิสัมพันธ์ระหว่างกลุ่มกับข้อสอบ ในแต่ละด้านโมเดลหลักโดยเฉพาะข้อสอบแต่ละข้อ
2009	Gómez-Benito et al.	ประสิทธิภาพของขนาดอิทธิพลในวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีถดถอยโลจิสติก จำลองข้อมูล 5 ปัจจัย คือรูปแบบของการทำหน้าที่ต่างกันของข้อสอบ ขนาดอิทธิพลของการทำหน้าที่ต่างกันของข้อสอบ จำนวนข้อสอบที่เกิดการทำหน้าที่ต่างกันในแบบสอบแต่ละฉบับ ขนาดกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบและความยาวของข้อสอบทั้งฉบับเงื่อนไขที่ทำการศึกษา จำนวน 225 เงื่อนไข	วิธีการตรวจสอบการทำหน้าที่ต่างกันเลือกใช้วิธีถดถอยโลจิสติกภายใต้โมเดลทฤษฎีการตอบสนองข้อสอบ 2 พารามิเตอร์ ผลการวิจัยสนับสนุนให้ศึกษาการวัดขนาดอิทธิพลโดยสถิติ R^2 รวมกับการทดสอบนัยสำคัญทางสถิติจะทำให้ได้สารสนเทศมากยิ่งขึ้น

บทที่ 3
วิธีดำเนินการวิจัย

การวิจัยครั้งนี้ มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 12 เงื่อนไข (2x2x3) คือ ขนาดการทำหน้าที่ต่างกันของข้อสอบ 2 ขนาด (ขนาดเล็ก และขนาดกลาง) ความยาวของแบบสอบ 2 รูปแบบ (9 ข้อ และ 15 ข้อ) และขนาดของกลุ่มตัวอย่าง 3 ขนาด (200 คน, 500 คน และ 1,000 คน) ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA โดยในแต่ละเงื่อนไขจำลองข้อมูลวนซ้ำ 100 รอบ โดยพิจารณาจากอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ และเปรียบเทียบอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่า ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก คือ ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ปัจจัยความยาวของแบบสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง โดยมีขั้นตอนในการดำเนินการวิจัย แบ่งเป็น 5 ขั้นตอน ดังนี้

ขั้นตอนที่ 1 การจัดกระทำข้อมูลตามตัวแปรและเงื่อนไขที่ศึกษา

ขั้นตอนที่ 2 การจำลองข้อมูล

ขั้นตอนที่ 3 การวิเคราะห์ข้อมูล

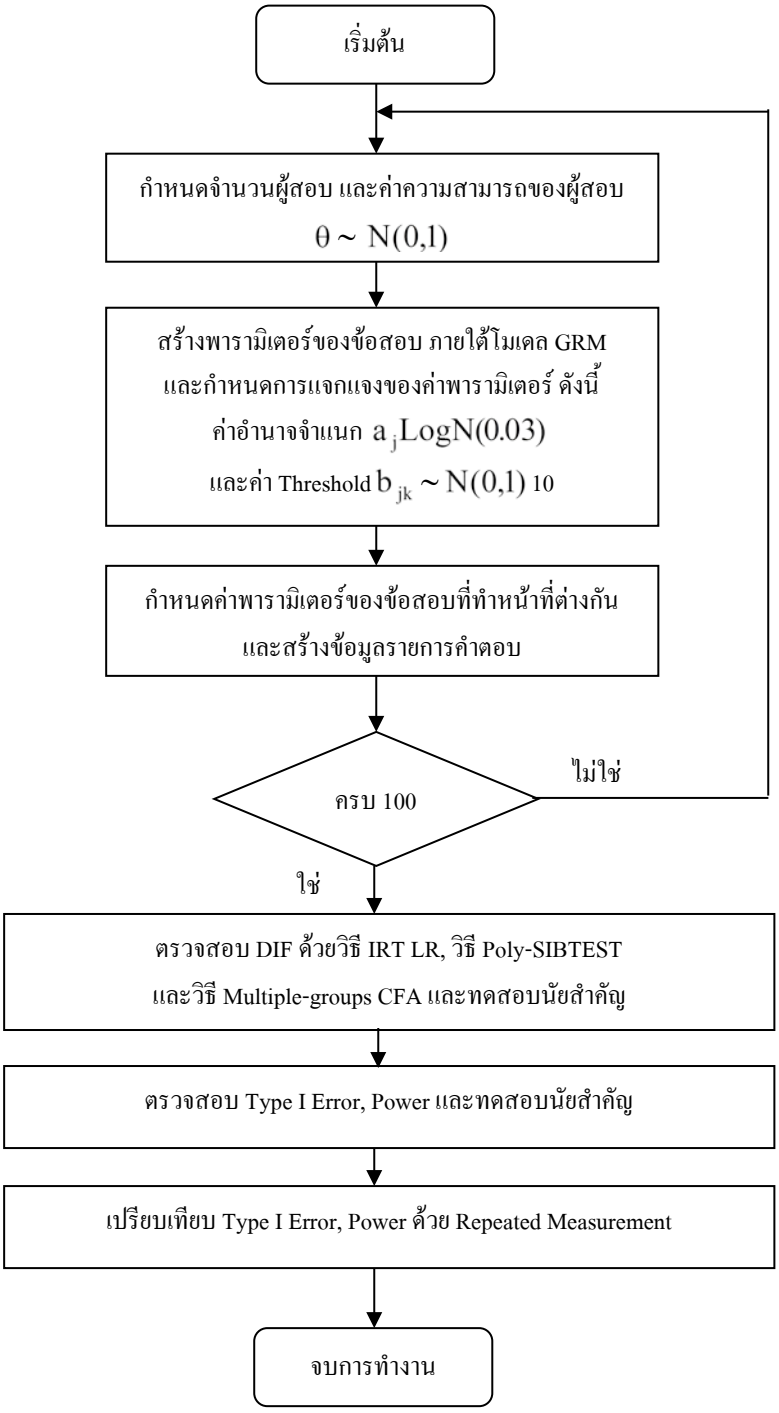
ขั้นตอนที่ 4 การเปรียบเทียบประสิทธิภาพของผลการตรวจสอบ DIF

ขั้นตอนที่ 5 การเปรียบเทียบอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย

ขั้นตอนที่ 1 การจัดกระทำข้อมูลตามตัวแปรและเงื่อนไขที่ศึกษา

การศึกษาในครั้งนี้เป็นการศึกษาในสถานการณ์จำลอง โดยใช้ทฤษฎีการตอบสนองข้อสอบแบบมิติเดียว (Unidimensional item response theory) ด้วยโมเดล Graded Response (GRM) และ ข้อสอบทำหน้าที่ต่างกันที่ไม่เป็นรูปแบบเดียวกัน (Nonuniform DIF) ประกอบด้วยการจัดกระทำข้อมูลใน 2 ขั้นตอน คือ ขั้นตอนที่ 1 เป็นการจัดกระทำข้อมูลภายใต้การจัดกระทำตัวแปรอิสระ 3 ตัวแปร คือ ขนาดการทำหน้าที่ต่างกันของข้อสอบ 2 รูปแบบ ความยาวของแบบสอบ 2 รูปแบบ และขนาดของกลุ่มตัวอย่าง 3 ขนาด มีข้อมูลที่ศึกษาทั้งสิ้น 12 เงื่อนไข (2x2x3)

ทุกเงื่อนไขจำลองข้อมูลวนซ้ำ 100 รอบ และใช้การจำลองแบบสอบในผู้สอบที่มีระดับความสามารถเฉลี่ยของกลุ่มเปรียบเทียบและกลุ่มอ้างอิงเท่ากัน (M = 0, SD = 1) รวมจำนวนข้อมูลจำลองที่เป็นไปตามเงื่อนไข 1,200 ชุด ดังมีรายละเอียดดังนี้



ภาพที่ 13 ขั้นตอนการดำเนินงานวิจัย

ขั้นตอนที่ 2 การจำลองข้อมูล

การศึกษาครั้งนี้ได้ทำการศึกษาโดยการจำลองกลุ่มตัวอย่างที่ประกอบด้วยกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ และแบบสอบที่จำลองขึ้นเป็นแบบสอบที่มีโครงสร้างการวัดมิติเดียว (Unidimensional) ข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าและทำหน้าที่ต่างกันแบบ Nonuniform โดยใช้โมเดล Graded response ประกอบด้วย ข้อสอบที่มีรายการคำตอบ 5 รายการ ผลการตอบข้อสอบในแต่ละรายการ คือ 1, 2, 3, 4 และ 5 โดยให้คะแนนเป็น 0, 1, 2, 3 และ 4 ตามลำดับ และกำหนดให้ความสามารถของข้อสอบมีแจกแจงแบบปกติ โดยใช้โปรแกรม WinGen ที่พัฒนาโดย Han (2007) ในการจำลองข้อมูล โดยมีประเด็นเงื่อนไขปัจจัยที่ใช้ในการศึกษาครั้งนี้ จำนวน 3 ปัจจัย คือ ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ปัจจัยความยาวของแบบสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง ดังนี้

ขนาดการทำหน้าที่ต่างกันของข้อสอบ (Magnitude of DIF)

Atar and Kamata (2011, p. 40) ได้ศึกษาเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการทดสอบอัตราส่วนความควรจะเป็นกับวิธีการถดถอยโลจิสติก พบว่า วิธีการทดสอบอัตราส่วนความควรจะเป็น มีอำนาจการทดสอบการทำหน้าที่ต่างกันของข้อสอบสูงเมื่อขนาดของการทำหน้าที่ต่างกันของข้อสอบมีขนาดกลาง (0.43) และอำนาจการทดสอบการทำหน้าที่ต่างกันของข้อสอบจะเพิ่มขึ้นเมื่อขนาดของการทำหน้าที่ต่างกันเพิ่มขึ้น Yoon and Millsap (2007) ได้ศึกษาการวิเคราะห์ความไม่แปรเปลี่ยนขององค์ประกอบในแบบสอบ โดยใช้การจำลองข้อมูลและกำหนดเงื่อนไขขนาดการทำหน้าที่ต่างกันของข้อสอบที่ใช้ในการศึกษาเป็น 3 ขนาด คือ ขนาด 0.01, 0.02 และ 0.03 และ Lopez Rivas et al. (2009) ได้ศึกษาผลกระทบของค่าพารามิเตอร์ข้อสอบในกลุ่มอ้างอิง ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธีอัตราส่วนความควรจะเป็นโดยกำหนดเงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ 2 ขนาด คือ ขนาดเล็ก (0.25) และขนาดใหญ่ (0.5) โดยผลการศึกษาพบว่า ขนาดการทำหน้าที่ต่างกันของข้อสอบที่มีขนาดใหญ่และจำนวนกลุ่มตัวอย่างมีขนาดเพิ่มขึ้น มีอำนาจการทดสอบสูง และสามารถควบคุมความคลาดเคลื่อนประเภทที่ 1ได้ดี ธเกียรติกมล ทองงอก (2554) ได้ศึกษาประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในวิธีถดถอยโลจิสติก โดยใช้เกณฑ์ขนาดอิทธิพล 2 วิธี สำหรับข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบทวิภาค ด้วยข้อมูลจำลองและข้อมูลเชิงประจักษ์ พบว่า ปฏิสัมพันธ์ระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบภายใต้วิธีถดถอยโลจิสติกกับเงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ 3 รูปแบบ คือ 0.1, 0.2 และ 0.4 มีอัตราความถูกต้องและอัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .001 และอาวีพร ปานทอง (2558) ได้ศึกษาเปรียบเทียบประสิทธิภาพ

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบให้คะแนนหลายค่าโดยวิธีทดสอบอัตราส่วนความควรจะเป็น วิธีเบส์เซียน และวิธีโพลี-ซิปเทสท์ พบว่า ปัจจัยขนาดของการทำทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนหลายค่า 3 ขนาด คือ 0.25 0.50 และ 1.00 มีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 1 และวิธี Poly-SIBTEST มีความคลาดเคลื่อนประเภทที่ 1 สูงเกินปกติเมื่อขนาดการทำหน้าที่ต่างกันของข้อสอบเพิ่มขึ้น จากผลการศึกษาข้างต้นพอสรุปได้ว่า ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบมีผลกระทบต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ดังนั้น ในการศึกษาครั้งนี้ผู้วิจัยจึงกำหนดปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบเป็นเงื่อนไขในการศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในครั้งนี้ โดยสนใจศึกษาขนาดของการทำหน้าที่ต่างกันของข้อสอบที่มีขนาดเล็ก คือ 0.10 และ ขนาดกลาง คือ 0.50

ความยาวของแบบสอบ

Chang et al. (2015) ได้ศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้วิธี Multiple-group categorical CFA ด้วยกระบวนการ Free baseline น้อยที่สุด ได้ออกแบบการวิจัยโดยกำหนดความยาวของแบบสอบที่ใช้ในการศึกษา เป็นแบบสอบถามลิเคิร์ท (Likert scale) จำนวน 15 ข้อ เพื่อใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ รวมถึง รวมถึง Lopez Rivas et al. (2009) ได้ศึกษาอิทธิพลของค่าพารามิเตอร์ข้อสอบที่เป็นกลุ่มอ้างอิงบนพื้นฐานของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้วิธีการ Free baseline ด้วยการวิเคราะห์อัตราส่วนความควรจะเป็นพบว่า ในเงื่อนไขปัจจัยขนาดของกลุ่มตัวอย่างที่มีขนาดเล็ก ค่าอำนาจการทดสอบจะเพิ่มขึ้นเมื่อค่าอำนาจจำแนกมีค่าสูงขึ้น ซึ่งจากข้อค้นพบนี้จึงได้เสนอแนะให้มีการศึกษาเพิ่มเติมเกี่ยวกับแบบสอบที่มีขนาดสั้นและกลุ่มตัวอย่างมีจำนวนน้อย นอกจากนี้ในการศึกษาครั้งนี้ผู้วิจัยศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีโครงสร้างการวัดแบบมิติเดียว ดังนั้น การกำหนดความยาวของแบบสอบจึงเป็นแบบสอบสั้นที่มีจำนวนข้อคำถามไม่มาก จึงสนใจศึกษาการจัดกระทำความยาวของแบบสอบ 2 รูปแบบ คือ จำนวน 9 ข้อ และ จำนวน 15 ข้อ

ขนาดของกลุ่มตัวอย่าง (Sample size)

ขนาดของกลุ่มตัวอย่างที่ใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ใช้การจำลองกลุ่มตัวอย่างออกเป็น 2 กลุ่ม คือ กลุ่มเปรียบเทียบ (N_F) และกลุ่มอ้างอิง (N_R) ตามปกติจะกำหนดขนาดของกลุ่มตัวอย่างของกลุ่มอ้างอิงเท่ากับหรือมากกว่ากลุ่มเปรียบเทียบ และวิธีโพลี-ซิปเทสท์ มักใช้ขนาดของกลุ่มตัวอย่างที่มีขนาดเล็กจนถึงขนาดใหญ่ในช่วง 100 ถึง 3,000 คน

(อรินทร์ น่วมถนอม, 2549) นอกจากนี้ Rogers and Swaminathan (1993) ได้ศึกษาพบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเมื่อขนาดของกลุ่มตัวอย่างเพิ่มขึ้น อัตราอำนาจการทดสอบจะเพิ่มขึ้น และอัตราความคลาดเคลื่อนประเภทที่ 1 จะลดลง รวมถึง Lopez Rivas et al. (2009) ได้ศึกษาอิทธิพลของค่าพารามิเตอร์ข้อสอบที่เป็นกลุ่มอ้างอิงบนพื้นฐานของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้วิธีการ Free baseline ด้วยการวิเคราะห์อัตราส่วนความควรจะเป็น พบว่า ในเงื่อนไขปัจจัยขนาดของกลุ่มตัวอย่างที่มีขนาดเล็ก ค่าอำนาจการทดสอบจะเพิ่มขึ้นเมื่อค่าอำนาจจำแนกมีค่าสูงขึ้น ซึ่งจากข้อค้นพบนี้จึงได้เสนอแนะให้มีการศึกษาเพิ่มเติมเกี่ยวกับแบบสอบที่มีขนาดสั้นและกลุ่มตัวอย่างมีจำนวนน้อย และ Nunnally (1978 อ้างถึงใน ศิริชัย กาญจนวาสี, 2548, หน้า 220) กล่าวว่า จำนวนกลุ่มตัวอย่างขั้นต่ำสำหรับใช้ในการวิเคราะห์ข้อสอบยังไม่มีกฎเกณฑ์ที่ตายตัว แต่สามารถกล่าวโดยทั่ว ๆ ไปได้ว่ากลุ่มตัวตัวอย่างขนาด 200 คน จะทำให้ค่าสถิติที่คำนวณได้มีความคงที่น่าเชื่อถือ หรืออย่างน้อยควรใช้กลุ่มตัวอย่าง 5 ถึง 10 เท่าของจำนวนข้อสอบ จากข้อค้นพบและข้อเสนอแนะข้างต้น ในการศึกษาครั้งนี้ผู้วิจัยจึงสนใจศึกษาขนาดของกลุ่มตัวอย่างที่มีจำนวน 200 คน, 500 คน และ 1,000 คน ในสัดส่วน 1: 1 ซึ่ง Chang et al. (2015) กล่าวว่า การศึกษาก่อนหน้านี้ แสดงให้เห็นว่าการกำหนดขนาดของกลุ่มตัวอย่างให้มีขนาดเท่ากันในกลุ่มเปรียบเทียบและกลุ่มอ้างอิงให้ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบดีกว่าการกำหนดขนาดของกลุ่มตัวอย่างไม่เท่ากัน ดังนั้น $N_F : N_R$ จะเป็น 100 คน: 100 คน, 250 คน: 250 คน และ 500 คน: 500 คน

จากเงื่อนไขปัจจัยที่แตกต่าง 4 ปัจจัย คือ ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ 2 รูปแบบ (0.10, 0.50) ความยาวของแบบสอบ 2 รูปแบบ (จำนวน 9 ข้อ, 15 ข้อ) และขนาดของกลุ่มตัวอย่าง 3 ขนาด (200 คน, 500 คน, 1,000 คน) รวมจำนวนเงื่อนไขที่ศึกษา 12 เงื่อนไข (2x2x3) และในแต่ละเงื่อนไขจำลองข้อมูลวนซ้ำ 100 ครั้ง ดังมีแบบแผนการจำลองข้อมูล ดังนี้

ตารางที่ 4 แบบแผนการจำลองข้อมูล

ขนาดการทำหน้าที่ต่างกันของข้อสอบ	ความยาวของแบบสอบ	ขนาดกลุ่มตัวอย่าง	เงื่อนไข
ขนาดเล็ก (0.10)	จำนวน 9 ข้อ	จำนวน 200 คน	เงื่อนไขที่ 1
		จำนวน 500 คน	เงื่อนไขที่ 2
		จำนวน 1,000 คน	เงื่อนไขที่ 3
	จำนวน 15 ข้อ	จำนวน 200 คน	เงื่อนไขที่ 4
		จำนวน 500 คน	เงื่อนไขที่ 5
		จำนวน 1,000 คน	เงื่อนไขที่ 6
ขนาดกลาง (0.50)	จำนวน 9 ข้อ	จำนวน 200 คน	เงื่อนไขที่ 7
		จำนวน 500 คน	เงื่อนไขที่ 8
		จำนวน 1,000 คน	เงื่อนไขที่ 9
	จำนวน 15 ข้อ	จำนวน 200 คน	เงื่อนไขที่ 10
		จำนวน 500 คน	เงื่อนไขที่ 11
		จำนวน 1,000 คน	เงื่อนไขที่ 12

โดยมีขั้นตอนการจำลองข้อมูลดังนี้

1. กำหนดขนาดกลุ่มตัวอย่างในกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบในสัดส่วน 1:1 คือ

1.1 กรณีกลุ่มตัวอย่าง ($N_F : N_R$) ขนาด 200 คน แบ่งกลุ่มตัวอย่างในกลุ่มเปรียบเทียบ จำนวน 100 คน และกลุ่มอ้างอิง จำนวน 100 คน (100: 100)

1.2 กรณีกลุ่มตัวอย่าง ($N_F : N_R$) ขนาด 500 คน แบ่งกลุ่มตัวอย่างในกลุ่มเปรียบเทียบ จำนวน 250 คน และกลุ่มอ้างอิง จำนวน 250 คน (250: 250)

1.3 กรณีกลุ่มตัวอย่าง ($N_F : N_R$) ขนาด 1,000 คน แบ่งกลุ่มตัวอย่างในกลุ่มเปรียบเทียบ จำนวน 500 คน และกลุ่มอ้างอิง จำนวน 500 คน (500: 500)

2. สร้างข้อมูลความสามารถของผู้สอบทั้งหมดตามขนาดของกลุ่มตัวอย่าง โดยกำหนดให้ทำการจำลองข้อมูลมีการแจกแจงแบบปกติ ที่มีค่าเฉลี่ยเป็น 0 และส่วนเบี่ยงเบนมาตรฐานเป็น 1 หรือเขียนแทนด้วย $\theta \sim N(0,1)$ ซึ่งค่าความสามารถของผู้สอบ (θ) จะมีพิสัยอยู่ระหว่าง $-\alpha$ ถึง $+\alpha$ แต่ผลการวิเคราะห์ส่วนใหญ่มักให้ค่าอยู่ในช่วง -3 ถึง +3 (ศิริชัย กาญจนวาสี, 2555, หน้า 54) ดังนั้น ผู้วิจัยจึงกำหนดให้การแจกแจงความสามารถของผู้สอบมีการแจกแจงแบบปกติ

โดยกำหนดให้ค่าเฉลี่ยความสามารถเป็น 0 และส่วนเบี่ยงเบนมาตรฐาน เป็น 1 เหมือนกันทั้งกลุ่มเปรียบเทียบและกลุ่มอ้างอิง

3. สร้างข้อมูลการตอบข้อสอบเพื่อให้ได้รายการคำตอบของผู้สอบตามรูปแบบของแบบสอบที่มีโครงสร้างแบบมิติเดียว (Unidimensional) และข้อสอบมีรูปแบบการตรวจให้คะแนนแบบหลายค่า ภายใต้โมเดล Graded response ซึ่งมีลักษณะเป็นโมเดลทั่วไปของโมเดลการตอบสนองข้อสอบที่มี 2 พารามิเตอร์ ซึ่งอธิบายได้ด้วยความชันร่วมของข้อคำถาม (Common item slop parameter, α_i) และค่า Threshold ของแต่ละรายการคำตอบ (Category threshold parameter, β_{ij}) เมื่อ j = 1, …, m_i โดยที่ m_i เป็นจำนวนของ Threshold ของข้อ i และจำนวนรายการคำตอบของข้อ i (K_i) = $m_i + 1$ โดยใช้หลักการคำนวณความน่าจะเป็นของการตอบแต่ละรายการคำตอบแบบ 2 ขั้นตอน (Indirect IRT model) โดยขั้นตอนแรกคำนวณค่าความชันร่วมของแต่ละข้อคำถาม จากนั้นจึงคำนวณค่าพารามิเตอร์ของแต่ละรายการคำตอบในแต่ละข้อคำถาม (ศิริชัย กาญจนวาสี, 2555, หน้า 89) โดยขั้นตอนการสร้างข้อมูลการตอบข้อสอบมีดังนี้

3.1 กำหนดจำนวนข้อสอบในแบบสอบที่มีความยาว จำนวน 9 ข้อ และจำนวน 15 ข้อ

3.2 กำหนดจำนวนรายการคำตอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า โดยกำหนดผลการตอบข้อสอบ 5 รายการ โดยให้คะแนนเป็น 0, 1, 2, 3 และ 4 ตามลำดับ และใช้โมเดล GRM ในการประมาณค่าพารามิเตอร์

3.3 กำหนดการแจกแจงของค่าพารามิเตอร์ข้อสอบ ได้แก่ ค่าความชันร่วม หรือค่าอำนาจจำแนก (a_i) เนื่องจากในทางทฤษฎีค่าพารามิเตอร์อำนาจจำแนกของข้อสอบ เป็นสัดส่วนโดยตรงของค่าความชันของโค้งลักษณะข้อสอบ (Item Charateristic Curve: ICC) ที่ตำแหน่ง b_j ซึ่งค่า a_i สูง แสดงถึงการจำแนกผู้สอบที่มีความสามารถแตกต่างกันได้ดี และในทางทฤษฎีค่า a_i มีค่าระหว่าง $(-\infty - +\infty)$ และควรมีค่าเป็น + และในทางปฏิบัตินิยมใช้ข้อสอบที่มีค่า a_i อยู่ระหว่าง +0.50 ถึง +2.50 ในการวิจัยครั้งนี้ ผู้วิจัยกำหนดให้ค่าพารามิเตอร์อำนาจจำแนก (a_i) มีการแจกแจงแบบล็อกนอร์มอล มีค่าเฉลี่ยเป็น 0 และค่าเบี่ยงเบนมาตรฐาน เป็น 0.03 หรือเขียนแทนด้วย $a_j \text{LogN}(0.03)$ และค่าพารามิเตอร์ Threshold (b_j) มีการแจกแจงแบบปกติ มีค่าเฉลี่ยเป็น 0 และค่าเบี่ยงเบนมาตรฐาน เป็น 1

4. สร้างข้อมูลข้อสอบที่ทำหน้าที่ต่างกัน เนื่องจากการศึกษาครั้งนี้เป็นการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ผู้วิจัยจึงจำลองข้อสอบที่ทำหน้าที่ต่างกันตามขนาดการทำหน้าที่ต่างกันของข้อสอบ 2 รูปแบบ คือ ขนาดเล็ก (0.10) และขนาดกลาง (0.50) โดยใช้สัดส่วนข้อสอบที่ทำหน้าที่ต่างกัน 10% (โดยไม่เป็นเงื่อนไขในการศึกษาครั้งนี้) ดังนั้น เงื่อนไขปัจจัยความยาวของ

แบบสอบ จำนวน 9 ข้อ จึงมีข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 1 ข้อ (กำหนดให้ข้อที่ 1 เป็นข้อสอบที่ทำหน้าที่ต่างกัน) และความยาวของแบบสอบ จำนวน 15 ข้อ จึงมีข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 2 ข้อ (กำหนดให้ข้อที่ 1 และข้อที่ 8 เป็นข้อสอบที่ทำหน้าที่ต่างกัน) และข้อสอบที่ทำหน้าที่ต่างกันไม่เป็นรูปแบบเดียวกัน (Nonuniform DIF) ดังนั้น เงื่อนไขการจำลองรูปแบบของข้อสอบที่ทำหน้าที่ต่างกันที่ไม่เป็นรูปแบบเดียวกัน กำหนดค่าพารามิเตอร์อำนาจจำแนกของข้อสอบ (a_i) ให้มีค่าแปรเปลี่ยน ดังนี้

ตารางที่ 5 การจำลองรูปแบบการทำหน้าที่ต่างกันของข้อสอบทำหน้าที่ไม่เป็นรูปแบบเดียวกัน ภายใต้ความยาวของแบบสอบ จำนวน 9 ข้อ ขนาดการทำหน้าที่ต่างกันของข้อสอบ 2 รูปแบบ (0.10 และ 0.50)

ข้อสอบข้อที่	ขนาด DIF	รูปแบบการทำหน้าที่ต่างกันของข้อสอบ
X_1	0.10	$a_{i1F} - 0.10 = a_{i1R}$, $b_{j1F} = b_{j1R} + 0.10$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R} + 0.10$, $b_{j4F} = b_{j4R}$
X_2	0.10	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_3	0.10	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_4	0.10	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_5	0.10	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_6	0.10	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_7	0.10	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_8	0.10	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$

ตารางที่ 5 (ต่อ)

ข้อสอบข้อที่	ขนาด DIF	รูปแบบการทำหน้าที่ต่างกันของข้อสอบ
X_9	0.10	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_1	0.50	$a_{i1F} - 0.50 = a_{i1R}$, $b_{j1F} = b_{j1R} + 0.50$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R} + 0.50$, $b_{j4F} = b_{j4R}$
X_2	0.50	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_3	0.50	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_4	0.50	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_5	0.50	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_6	0.50	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_7	0.50	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_8	0.50	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_9	0.50	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$

ตารางที่ 6 การจำลองรูปแบบการทำหน้าที่ต่างกันของข้อสอบทำหน้าที่ไม่เป็นรูปแบบเดียวกันภายใต้ความยาวของแบบสอบ จำนวน 15 ข้อ ขนาดการทำหน้าที่ต่างกันของข้อสอบ 2 รูปแบบ (0.10 และ 0.50)

ข้อสอบข้อที่	ขนาด DIF	รูปแบบการทำหน้าที่ต่างกันของข้อสอบ
X_1	0.10	$a_{i1F} - 0.10 = a_{i1R}$, $b_{j1F} = b_{j1R} + 0.10$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R} + 0.10$, $b_{j4F} = b_{j4R}$
X_2	0.10	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_3	0.10	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_4	0.10	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_5	0.10	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_6	0.10	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_7	0.10	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_8	0.10	$a_{i1F} - 0.10 = a_{i1R}$, $b_{j1F} = b_{j1R} + 0.10$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R} + 0.10$, $b_{j4F} = b_{j4R}$
X_9	0.10	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_{10}	0.10	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_{11}	0.10	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_{12}	0.10	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$

ตารางที่ 6 (ต่อ)

ข้อสอบข้อที่	ขนาด DIF	รูปแบบการทำหน้าที่ต่างกันของข้อสอบ
X_{13}	0.10	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_{14}	0.10	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_{15}	0.10	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_{1}	0.50	$a_{i1F} - 0.50 = a_{i1R}$, $b_{j1F} = b_{j1R} + 0.50$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R} + 0.50$, $b_{j4F} = b_{j4R}$
X_{2}	0.50	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_{3}	0.50	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_{4}	0.50	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_{5}	0.50	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_{6}	0.50	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_{7}	0.50	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_{8}	0.50	$a_{i1F} - 0.50 = a_{i1R}$, $b_{j1F} = b_{j1R} + 0.50$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R} + 0.50$, $b_{j4F} = b_{j4R}$
X_{9}	0.50	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_{10}	0.50	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$

ตารางที่ 6 (ต่อ)

ข้อสอบข้อที่	ขนาด DIF	รูปแบบการทำหน้าที่ต่างกันของข้อสอบ
X_{11}	0.50	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_{12}	0.50	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_{13}	0.50	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_{14}	0.50	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$
X_{15}	0.50	$a_{i1F} = a_{i1R}$, $b_{j1F} = b_{j1R}$, $b_{j2F} = b_{j2R}$, $b_{j3F} = b_{j3R}$, $b_{j4F} = b_{j4R}$

รูปแบบการทำหน้าที่ต่างกันของข้อสอบในตารางที่ 3 และ 4 เป็นการออกแบบจำลองค่าพารามิเตอร์รายการคำตอบของกลุ่มเปรียบเทียบและกลุ่มอ้างอิง โดยกำหนดให้มีค่าแตกต่างกัน 0.10 และ 0.50 ซึ่งเป็นค่าขนาดการทำหน้าที่ต่างกันของข้อสอบตามเงื่อนไขปัจจัยของการศึกษาในครั้งนี้

ขั้นตอนที่ 3 การวิเคราะห์ข้อมูล

การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA และเพื่อเปรียบเทียบอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่า ด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสามวิธี ผู้วิจัยดำเนินการดังนี้

1. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี IRT LR โดยใช้โปรแกรม IRTPRO Version 4.1 (Cai, Thissen, & du Toit, 2011) โดยมีสถิติการทดสอบอัตราส่วนความควรจะเป็น (G^2) (Atar & Kamata, 2011, p. 37) ดังนี้

$$G^2 = -2LL_C - (-2LL_A) \quad (38)$$

เมื่อ G^2 แทน ค่าการแจกแจง χ^2 ของแต่ละข้อคำถาม ที่ df เท่ากัน

LL_C แทน ฟังก์ชันความเป็นไปได้ของโมเดลพื้นฐาน (Compact)

LL_A แทน ฟังก์ชันความเป็นไปได้ของโมเดลเปรียบเทียบ (Augmented)

ถ้าผลการทดสอบพบว่า มีนัยสำคัญทางสถิติ แสดงว่า ตรวจพบการทำหน้าที่ต่างกันของข้อสอบ

2. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธีโพลี-ซิปเทสท์ โดยใช้โปรแกรม DIF Analysis Version 1.7

3. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี IRT Mplus โดยใช้โปรแกรม Mplus Version 6.12 ซึ่งเน้นที่การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามทฤษฎี IRT

ขั้นตอนที่ 4 การเปรียบเทียบประสิทธิภาพของผลการตรวจสอบ DIF

ประสิทธิภาพของการตรวจสอบ DIF พิจารณาจากอัตราความคลาดเคลื่อนประเภทที่ 1 (Type I error rate) และอัตราอำนาจการทดสอบการทำหน้าที่ต่างกันของข้อสอบ (Power rate) ซึ่งการวิเคราะห์อัตราความถูกต้อง $(1-\beta)$ กับความคลาดเคลื่อนประเภทที่ 2 (β) และการวิเคราะห์ระดับความเชื่อมั่น $(1-\alpha)$ กับความคลาดเคลื่อนประเภทที่ 1 (α) เป็นดัชนีที่มีสเกลผกผันกัน ดังนั้นการพิจารณาดัชนีบ่งชี้คุณภาพ 2 ตัว คือ อัตราความถูกต้องของการตรวจพบ DIF (Power rate) และอัตราความคลาดเคลื่อนประเภทที่ 1 (Type I error rate) ก็เพียงพอที่จะให้สารสนเทศครบทั้ง 4 เหตุการณ์ (ศิริชัย กาญจนวาสี, 2555)

อัตราความคลาดเคลื่อนประเภทที่ 1 (Type I error rate) เป็นการระบุผิดพลาดว่าข้อสอบหน้าที่ต่างกัน (False positive) ทั้งที่ความเป็นจริงข้อสอบทำหน้าที่ไม่ต่างกัน คำนวณจากสัดส่วนของจำนวนข้อสอบที่ตรวจสอบผิดพลาดว่า ทำหน้าที่ต่างกันทั้งที่ในความเป็นจริงข้อสอบไม่ได้ทำหน้าที่ต่างกัน คำนวณเป็นค่าร้อยละ ดังนี้

$$E_1 = \frac{n_2}{N_2} \times 100 \quad (39)$$

เมื่อ E_1 แทน อัตราความคลาดเคลื่อนประเภทที่ 1

n_2 แทน จำนวนข้อสอบที่ระบุผิดว่า DIF

N_2 แทน จำนวนข้อสอบที่ไม่ DIF ทั้งหมดที่ตรวจสอบด้วยวิธีเกณฑ์

อัตราอำนาจการทดสอบ คำนวณจากจำนวนของข้อสอบที่ตรวจสอบได้ถูกต้องว่าทำหน้าที่ต่างกัน ต่อจำนวนข้อสอบที่ทำน้าที่ต่างกันทั้งหมดในแบบสอบ คำนวณเป็นค่าร้อยละ ดังนี้

$$P = \frac{n_1}{N_1} \times 100 \quad (40)$$

เมื่อ P แทน อัตราอำนาจการทดสอบ

n_1 แทน จำนวนข้อสอบที่ตรวจสอบได้ถูกต้องว่า DIF

N_1 แทน จำนวนข้อสอบที่ DIF ทั้งหมดที่ตรวจสอบด้วยวิธีเกณฑ์

เกณฑ์การพิจารณาค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบการทำหน้าที่ต่างกันของข้อสอบ

ผู้วิจัยได้นำข้อมูลที่ได้จากการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่เหมาะสมกับข้อมูลการตอบข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA ภายใต้ปัจจัยที่แตกต่าง 3 ปัจจัย คือ ขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง ไปคำนวณอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบ และนำผลที่ได้มาพิจารณาตามเกณฑ์ต่อไปนี้

อัตราความคลาดเคลื่อนประเภทที่ 1

เกณฑ์การพิจารณาหากมีค่าความคลาดเคลื่อนประเภทที่ 1 ต่ำกว่าหรือเท่ากับ 0.05 ถือว่าควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดี (Atar & Kamata, 2011, p. 40 และ อาวีพร ปานทอง, 2558, หน้า 81) นั่นคือวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบไม่ระบุการทำหน้าที่ต่างกันของข้อสอบในข้อที่ไม่มีการทำหน้าที่ต่างกันของข้อสอบได้จริง การทดสอบสมมติฐานความคลาดเคลื่อนประเภทที่ 1 มีดังนี้

ตั้งสมมติฐานการทดสอบ

$H_0 : P \leq .05$
$H_1 : P > .05$

สถิติทดสอบ ดังนี้ $$Z = \frac{\hat{P} \quad P}{\sqrt{\frac{P(1 \quad P)}{n}}}$$

$\hat{P}$ แทน สัดส่วนการเกิดความคลาดเคลื่อนประเภทที่ 1 ของกลุ่มตัวอย่าง
P แทน สัดส่วนการเกิดความคลาดเคลื่อนประเภทที่ 1 ของประชากร
n แทน จำนวนของการทำซ้ำในการจำลองข้อมูล

การกำหนดขอบเขตวิกฤต

เนื่องจากเป็นการทดสอบสมมติฐานทางเดียว จึงนำ Z ที่คำนวณได้จากสูตรเทียบกับ $Z_\alpha = Z_{.05}$ ถ้า Z ที่คำนวณได้น้อยกว่า $Z_\alpha = 1.645$ จะยอมรับสมมติฐาน H_0 แสดงว่าวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบนั้นสามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

อัตราอำนาจการทดสอบ

เกณฑ์ที่ใช้พิจารณาอัตราอำนาจการทดสอบ จะพิจารณาอัตราอำนาจการทดสอบเมื่อสามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ก่อน และอำนาจการทดสอบต้องมีค่าเฉลี่ยตั้งแต่ 0.80 ขึ้นไป จึงถือว่ามีอำนาจการทดสอบเพียงพอ (Sufficient power) หากต่ำกว่า 0.80 ถือว่าวิธีนั้น ๆ ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ไม่ดี (Atar & Kamata, 2011, p. 40 และอาวีพร ปานทอง, 2558, หน้า 81) การทดสอบสมมติฐานอำนาจการทดสอบมีดังนี้

$H_0 : P \geq .80$
$H_1 : P < .80$

สถิติทดสอบ ดังนี้ $$Z = \frac{\hat{P} \quad P}{\sqrt{\frac{P(1 \quad P)}{n}}}$$

$\hat{P}$ แทน สัดส่วนอำนาจการทดสอบของกลุ่มตัวอย่าง

P แทน สัดส่วนอำนาจการทดสอบของประชากร

n แทน จำนวนของการทำซ้ำในการจำลองข้อมูล

การกำหนดขอบเขตวิกฤต

เนื่องจากเป็นการทดสอบสมมติฐานทางเดียว จึงนำ Z ที่คำนวณได้จากสูตรเทียบกับ $Z_{\alpha} = Z_{.05}$ ถ้า Z ที่คำนวณได้มากกว่า $Z_{.05} = 1.645$ จะยอมรับสมมติฐาน H_0 แสดงว่าวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบนั้นมีอำนาจการทดสอบอย่างมีนัยสำคัญทางสถิติที่ระดับ .05

ทั้งนี้ เกณฑ์ที่ใช้ในการตัดสินประสิทธิภาพของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ พิจารณาจากผลการตรวจสอบที่มีอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำ และอัตราอำนาจการทดสอบสูง ตามเกณฑ์ที่กำหนดข้างต้น

ขั้นตอนที่ 5 การเปรียบเทียบอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย

การวิเคราะห์ในขั้นตอนนี้ ผู้วิจัยดำเนินการวิเคราะห์เปรียบเทียบอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบ GRM ภายใต้เงื่อนไขปัจจัยหลักที่แตกต่าง 3 ปัจจัยหลัก คือ ขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง ด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ 3 วิธี ด้วยวิธีการวิเคราะห์ความแปรปรวนแบบวัดซ้ำ (Repeated measurement) ซึ่งมีตัวแปรวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นตัวแปรวัดซ้ำ (Within-subject factor) และปัจจัยหลัก 3 ปัจจัย เป็นตัวแปรวัดต่างกลุ่ม (Between-subjects factors) โดยใช้โปรแกรม SPSS ดังนี้

1. การวิเคราะห์เปรียบเทียบค่าเฉลี่ยอัตราความคลาดเคลื่อนประเภทที่ 1 ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก ด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสามวิธี (Tests of within-subjects effects)

2. การวิเคราะห์เปรียบเทียบค่าเฉลี่ยอัตราความคลาดเคลื่อนประเภทที่ 1 ภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง (Tests of between-subjects effects)

3. การวิเคราะห์เปรียบเทียบค่าเฉลี่ยอำนาจการทดสอบ ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก ด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสามวิธี (Tests of within-subjects effects)

4. การวิเคราะห์เปรียบเทียบค่าเฉลี่ยอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง (Tests of between-subjects effects)

บทที่ 4
ผลการวิเคราะห์ข้อมูล

การวิจัยครั้งนี้ มีจุดมุ่งหมายเพื่อ 1) ศึกษาประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 12 เงื่อนไข (2x2x3) คือ ขนาดการทำหน้าที่ต่างกันของข้อสอบ (ขนาดเล็ก และขนาดกลาง) ความยาวของแบบสอบ 2 รูปแบบ (9 ข้อ และ 15 ข้อ) และขนาดของกลุ่มตัวอย่าง 3 ขนาด (200 คน, 500 คน และ 1,000 คน) ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA โดยในแต่ละเงื่อนไขจำลองข้อมูลวนซ้ำ 100 ครั้ง โดยพิจารณาจากอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ และ 2) เปรียบเทียบอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่า ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก คือ ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ปัจจัยความยาวของแบบสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง ดังนั้น การเสนอผลการวิเคราะห์ข้อมูล ผู้วิจัยจึงนำเสนอตามขั้นตอนของการวิเคราะห์ข้อมูลเพื่อตอบวัตถุประสงค์ของการวิจัยดังกล่าว โดยนำเสนอแยกเป็น 2 ตอน ดังนี้

ตอนที่ 1 ผลการตรวจสอบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA โดยพิจารณาจากอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

ตอนที่ 2 ผลการเปรียบเทียบอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก

ในการนำเสนอผลการวิเคราะห์ข้อมูล ผู้วิจัยได้กำหนดสัญลักษณ์ที่ใช้ในการแสดงผลการวิเคราะห์ข้อมูล ดังนี้

GRM	หมายถึง โมเดล Graded-response
DIF	หมายถึง การทำหน้าที่ต่างกันของข้อสอบ
IRTLR	หมายถึง วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี IRT LR

POLYSIB	หมายถึง	วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี Poly-SIBTEST
MGCFA	หมายถึง	วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี Multiple-groups CFA
C1…C12	หมายถึง	เงื่อนไขที่ศึกษา (รวม 12 เงื่อนไข: 2x2x3)
method	หมายถึง	วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ 3 วิธี คือ วิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-Groups CFA
DIFsize	หมายถึง	ขนาดการทำหน้าที่ต่างกันของข้อสอบ 2 รูปแบบ คือ ขนาดเล็ก (0.10) และขนาดกลาง (0.50)
LTESTsize	หมายถึง	ความยาวของแบบสอบ 2 รูปแบบ คือ จำนวน 9 ข้อ และ จำนวน 15 ข้อ
SAMPLEsize	หมายถึง	ขนาดของกลุ่มตัวอย่าง 3 ขนาด คือ 200 คน (100 คน: 100 คน, 250 คน: 250 คน และ 500 คน: 500 คน)

ตอนที่ 1 ผลการตรวจสอบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ภายใต้เงื่อนไขปัจจัยที่แตกต่างกัน 3 ปัจจัย ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA

การตรวจสอบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบ GRM ภายใต้เงื่อนไขปัจจัยที่แตกต่างกัน 3 ปัจจัยหลัก คือ ขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA พิจารณาจากอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ จำแนกเป็น 2 ประเด็น คือ ประเด็นที่ 1 คือ ผลการวิเคราะห์ค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก ด้วยวิธีการตรวจสอบการทำหน้าต่างกันของข้อสอบ 3 วิธี และประเด็นที่ 2 คือ ผลการทดสอบค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 และค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก ด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ 3 วิธี

1. ผลการวิเคราะห์ค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก ด้วยวิธีการตรวจสอบการทำหน้าต่างกันของข้อสอบ 3 วิธี

ผลการวิเคราะห์ค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก ด้วยวิธีการตรวจสอบการทำหน้าต่างกันของข้อสอบ 3 วิธี แสดงดังตารางที่ 6-9 และแผนภาพที่ 14-15

ตารางที่ 7 ค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 ภายใต้ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ขนาดเล็ก (0.10) ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA

ขนาดการทำหน้าที่ต่างกันของข้อสอบ	ความยาวของแบบสอบ	ขนาดของกลุ่มตัวอย่าง (N_F: N_R)	ค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1		
			IRTLR	POLYSIB	MGCFA
ขนาดเล็ก (0.10)	จำนวน 9 ข้อ	100: 100	0.015	0.059	0.111
		250: 250	0.015	0.048	0.080
		500: 500	0.013	0.051	0.061
	จำนวน 15 ข้อ	100: 100	0.012	0.061	0.132
		250: 250	0.027	0.051	0.118
		500: 500	0.045	0.056	0.116

จากผลการวิเคราะห์ตามตารางที่ 7 ค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ภายใต้ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ขนาดเล็ก (0.10) ความยาวของแบบสอบ จำนวน 9 ข้อ และ 15 ข้อ และขนาดของกลุ่มตัวอย่าง สัดส่วน 1: 1 (กลุ่มเปรียบเทียบ: กลุ่มอ้างอิง) จำนวน 100: 100 คน, 250: 250 คน และ 500: 500 คน ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี IRT LR มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 อยู่ในช่วงประมาณ 0.012-0.045

หรือคิดเป็นร้อยละ 1.20-4.50 วิธี Poly-SIBTEST มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 อยู่ในช่วงประมาณ 0.048-0.061 หรือคิดเป็นร้อยละ 4.80-6.10 และวิธี Multiple-groups CFA มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 อยู่ในช่วงประมาณ 0.061-0.132 หรือคิดเป็นร้อยละ 6.10-13.20

1. เมื่อพิจารณาภายใต้ปัจจัยความยาวของแบบสอบ จำนวน 9 ข้อ และขนาดของกลุ่มตัวอย่าง (N_F: N_R) ขนาด 100 คน: 100 คน พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียวด้วยวิธี IRT LR มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำสุด คิดเป็นร้อยละ 1.50 สำหรับวิธี Poly-SIBTEST คิดเป็นร้อยละ 5.90 และวิธี Multiple-groups CFA คิดเป็นร้อยละ 11.10

2. เมื่อพิจารณาภายใต้ปัจจัยความยาวของแบบสอบ จำนวน 9 ข้อ และขนาดของกลุ่มตัวอย่าง (N_F: N_R) ขนาด 250 คน: 250 คน พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียวด้วยวิธี IRT LR มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำสุด คิดเป็นร้อยละ 1.50 สำหรับวิธี Poly-SIBTEST คิดเป็นร้อยละ 4.80 และวิธี Multiple-groups CFA คิดเป็นร้อยละ 8.00

3. เมื่อพิจารณาภายใต้ปัจจัยความยาวของแบบสอบ จำนวน 9 ข้อ และขนาดของกลุ่มตัวอย่าง (N_F: N_R) ขนาด 500 คน: 500 คน พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียวด้วยวิธี IRT LR มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำสุด คิดเป็นร้อยละ 1.30 สำหรับวิธี Poly-SIBTEST คิดเป็นร้อยละ 5.10 และวิธี Multiple-groups CFA คิดเป็นร้อยละ 6.10

4. เมื่อพิจารณาภายใต้ปัจจัยความยาวของแบบสอบ จำนวน 15 ข้อ และขนาดของกลุ่มตัวอย่าง (N_F: N_R) ขนาด 100 คน: 100 คน พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียวด้วยวิธี IRT LR มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำสุด คิดเป็นร้อยละ 1.20 สำหรับวิธี Poly-SIBTEST คิดเป็นร้อยละ 6.10 และวิธี Multiple-groups CFA คิดเป็นร้อยละ 13.20

5. เมื่อพิจารณาภายใต้ปัจจัยความยาวของแบบสอบ จำนวน 15 ข้อ และขนาดของกลุ่มตัวอย่าง (N_F: N_R) ขนาด 250 คน: 250 คน พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียวด้วยวิธี IRT LR มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำสุด คิดเป็นร้อยละ 2.70 สำหรับวิธี Poly-SIBTEST คิดเป็นร้อยละ 5.10 และวิธี Multiple-groups CFA คิดเป็นร้อยละ 11.80

6. เมื่อพิจารณาภายใต้ปัจจัยความยาวของแบบสอบ จำนวน 15 ข้อ และขนาดของกลุ่มตัวอย่าง (N_F: N_R) ขนาด 500 คน: 500 คน พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียวด้วยวิธี IRT LR มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำสุด คิดเป็นร้อยละ 4.50 สำหรับวิธี Poly-SIBTEST คิดเป็นร้อยละ 5.60 และวิธี Multiple-groups CFA คิดเป็นร้อยละ 11.60

7. โดยภาพรวม การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ขนาดเล็ก (0.10) ด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสามวิธี พบว่า วิธี IRT LR มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 อยู่ในเกณฑ์ที่กำหนดทุกเงื่อนไข (ต่ำกว่า 0.05) และมีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำสุด คิดเป็นร้อยละ 1.20 ภายใต้เงื่อนไขปัจจัยความยาวของแบบสอบ จำนวน 15 ข้อ และขนาดของกลุ่มตัวอย่าง 100 คน: 100 คน สำหรับวิธี Poly-SIBTEST มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 อยู่ในเกณฑ์ที่กำหนด (ต่ำกว่า 0.05) เพียง 1 เงื่อนไข คิดเป็นร้อยละ 4.80 ภายใต้เงื่อนไขปัจจัยความยาวของแบบสอบจำนวน 9 ข้อ และขนาดของกลุ่มตัวอย่าง 250 คน: 250 คน สำหรับวิธี Multiple-groups CFA มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 สูงกว่าเกณฑ์ที่กำหนดในทุกเงื่อนไขปัจจัย

ตารางที่ 8 ค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 ภายใต้ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ขนาดกลาง (0.50) ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA

ขนาดการทำหน้าที่ต่างกันของข้อสอบ	**ความยาวของแบบสอบ**	**ขนาดของกลุ่มตัวอย่าง (N_F: N_R)**	**ค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1**		
			IRTLR	**POLYSIB**	**MGCFA**
ขนาดกลาง (0.50)	จำนวน 9 ข้อ	100: 100	0.005	0.053	0.108
		250: 250	0.018	0.050	0.061
		500: 500	0.024	0.065	0.073
	จำนวน 15 ข้อ	100: 100	0.016	0.069	0.155
		250: 250	0.024	0.061	0.189
		500: 500	0.023	0.050	0.117

จากผลการวิเคราะห์ตามตารางที่ 8 ค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ภายใต้ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ขนาดกลาง (0.50) ความยาวของแบบสอบ จำนวน 9 ข้อ และ 15 ข้อ และขนาดของกลุ่มตัวอย่าง สัดส่วน 1: 1 (กลุ่มเปรียบเทียบ: กลุ่มอ้างอิง) จำนวน 100: 100 คน, 250: 250 คน และ 500: 500 คน ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี IRT LR มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 อยู่ในช่วงประมาณ 0.005-0.024 หรือคิดเป็นร้อยละ 0.50-2.40 วิธี Poly-SIBTEST มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 อยู่ในช่วงประมาณ 0.050-0.069 หรือคิดเป็นร้อยละ 5.50-6.90 และวิธี Multiple-groups CFA มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 อยู่ในช่วงประมาณ 0.061-0.089 หรือคิดเป็นร้อยละ 6.10-18.90

1. เมื่อพิจารณาภายใต้ปัจจัยความยาวของแบบสอบ จำนวน 9 ข้อ และขนาดของกลุ่มตัวอย่าง (N_F: N_R) ขนาด 100 คน: 100 คน พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี IRT LR มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำสุด คิดเป็นร้อยละ 0.50 สำหรับวิธี Poly-SIBTEST คิดเป็นร้อยละ 5.30 และวิธี Multiple-groups CFA คิดเป็นร้อยละ 10.80

2. เมื่อพิจารณาภายใต้ปัจจัยความยาวของแบบสอบ จำนวน 9 ข้อ และขนาดของกลุ่มตัวอย่าง (N_F: N_R) ขนาด 250 คน: 250 คน พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี IRT LR มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำสุด คิดเป็นร้อยละ 1.80 สำหรับวิธี Poly-SIBTEST คิดเป็นร้อยละ 5.00 และวิธี Multiple-groups CFA คิดเป็นร้อยละ 6.10

3. เมื่อพิจารณาภายใต้ปัจจัยความยาวของแบบสอบ จำนวน 9 ข้อ และขนาดของกลุ่มตัวอย่าง (N_F: N_R) ขนาด 500 คน: 500 คน พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี IRT LR มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำสุด คิดเป็นร้อยละ 2.40 สำหรับวิธี Poly-SIBTEST คิดเป็นร้อยละ 6.50 และวิธี Multiple-groups CFA คิดเป็นร้อยละ 7.30

4. เมื่อพิจารณาภายใต้ปัจจัยความยาวของแบบสอบ จำนวน 15 ข้อ และขนาดของกลุ่มตัวอย่าง (N_F: N_R) ขนาด 100 คน: 100 คน พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว

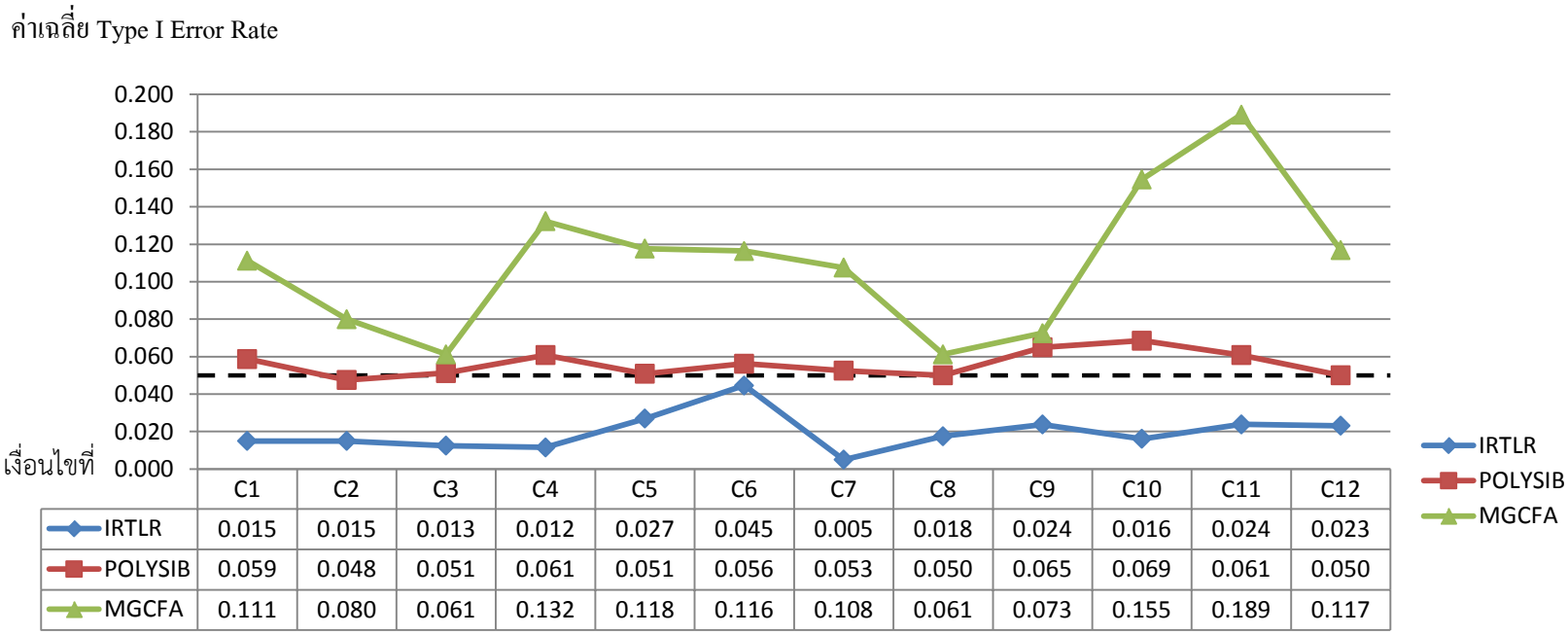
ด้วยวิธี IRT LR มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำสุด คิดเป็นร้อยละ 1.60 สำหรับวิธี Poly-SIBTEST คิดเป็นร้อยละ 6.90 และวิธี Multiple-groups CFA คิดเป็นร้อยละ 15.50

5. เมื่อพิจารณาภายใต้ปัจจัยความยาวของแบบสอบ จำนวน 15 ข้อ และขนาดของกลุ่มตัวอย่าง (N_F: N_R) ขนาด 250 คน: 250 คน พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี IRT LR มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำสุด คิดเป็นร้อยละ 2.40 สำหรับวิธี Poly-SIBTEST คิดเป็นร้อยละ 6.10 และวิธี Multiple-groups CFA คิดเป็นร้อยละ 18.90

6. เมื่อพิจารณาภายใต้ปัจจัยความยาวของแบบสอบ จำนวน 15 ข้อ และขนาดของกลุ่มตัวอย่าง (N_F: N_R) ขนาด 500 คน: 500 คน พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี IRT LR มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำสุด คิดเป็นร้อยละ 2.30 สำหรับวิธี Poly-SIBTEST คิดเป็นร้อยละ 5.00 และวิธี Multiple-groups CFA คิดเป็นร้อยละ 11.70

7. โดยภาพรวม การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจการให้คะแนนแบบหลายค่า ภายใต้ปัจจัยการทำหน้าที่ต่างกันของข้อสอบ ขนาดกลาง (0.50) ด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสามวิธี พบว่า วิธี IRT LR มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 อยู่ในเกณฑ์ที่กำหนด (ต่ำกว่า 0.05) ทุกเงื่อนไขปัจจัย และมีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำสุด คิดเป็นร้อยละ 0.50 ภายใต้เงื่อนไขปัจจัยความยาวของแบบสอบ จำนวน 9 ข้อ และขนาดของกลุ่มตัวอย่าง 100 คน: 100 คน สำหรับวิธี Poly-SIBTEST มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 อยู่ในเกณฑ์ที่กำหนดเพียง 2 เงื่อนไข (เท่ากับ 0.05) และมีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำสุด คิดเป็นร้อยละ 5.00 ภายใต้เงื่อนไขปัจจัยความยาวของแบบสอบ จำนวน 9 ข้อ และขนาดของกลุ่มตัวอย่าง 250 คน: 250 คน และภายใต้เงื่อนไขปัจจัยความยาวของแบบสอบ จำนวน 15 ข้อ และขนาดของกลุ่มตัวอย่าง 500 คน: 500 คน สำหรับวิธี Multiple-groups CFA มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 สูงกว่าเกณฑ์ที่กำหนดทุกเงื่อนไขปัจจัย

กราฟแสดงค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 ของการทดสอบการทำหน้าที่ต่างกันของข้อสอบภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA



IRTLR คือ วิธีทดสอบอัตราส่วนความควรจะเป็น, POLYSIB คือ วิธี Poly-SIBTEST และ MGCFA คือ วิธี Multiple-groups CFA

ภาพที่ 14 กราฟแสดงค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 (Type I error rate) ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA

จากภาพที่ 14 ผลการวิเคราะห์ค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 ภายใต้ปัจจัยที่แตกต่าง 3 ปัจจัยหลัก ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA พบว่า โดยภาพรวมของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี IRT LR มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 อยู่ในเกณฑ์ที่กำหนด (ต่ำกว่า 0.05) ทุกเงื่อนไขปัจจัย และวิธี Poly-SIBTEST มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 อยู่ในเกณฑ์ที่กำหนด (ต่ำกว่าหรือเท่ากับ 0.05) เพียง 3 เงื่อนไข คือ เงื่อนไขที่ 2 (C2) ภายใต้ปัจจัยเงื่อนไขขนาดการทำหน้าที่ต่างกันของข้อสอบ ขนาดเล็ก (0.10) ความยาวของแบบสอบจำนวน 9 ข้อ และขนาดของกลุ่มตัวอย่าง 250 คน: 250 คน เงื่อนไขที่ 8 (C8) ภายใต้ปัจจัยเงื่อนไขขนาดการทำหน้าที่ต่างกันของข้อสอบ ขนาดกลาง (0.50) ความยาวของแบบสอบ จำนวน 9 ข้อ และขนาดของกลุ่มตัวอย่าง 250 คน: 250 คน และเงื่อนไขที่ 12 (C12) ภายใต้ปัจจัยเงื่อนไขขนาดการทำหน้าที่ต่างกันของข้อสอบ ขนาดกลาง (0.50) ความยาวของแบบสอบ จำนวน 15 ข้อ และขนาดของกลุ่มตัวอย่าง 500 คน: 500 คน ตามลำดับ สำหรับวิธี Multiple-groups CFA มีค่าเฉลี่ยอัตราความคลาดเคลื่อนประเภทที่ 1 สูงกว่าเกณฑ์ที่กำหนดในทุกเงื่อนไขปัจจัย

ตารางที่ 9 ค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบภายใต้ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ขนาดเล็ก (0.10) ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA

ขนาดการทำหน้าที่ต่างกันของข้อสอบ	**ความยาวของแบบสอบ**	**ขนาดของกลุ่มตัวอย่าง (N_F: N_R)**	**ค่าเฉลี่ยของอัตราอำนาจการทดสอบ**		
			IRTLR	**POLYSIB**	**MGCFA**
ขนาดเล็ก (0.10)	จำนวน 9 ข้อ	100: 100	0.000	0.050	0.080
		250: 250	0.120	0.080	0.360
		500: 500	0.200	0.230	0.070
	จำนวน 15 ข้อ	100: 100	0.115	0.075	0.245
		250: 250	0.295	0.065	0.190
		500: 500	0.645	0.090	0.460

จากผลการวิเคราะห์ตามตารางที่ 9 ค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ภายใต้ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบขนาดเล็ก (0.10) ความยาวของแบบสอบ จำนวน 9 ข้อ และ 15 ข้อ และขนาดของกลุ่มตัวอย่างสัดส่วน 1: 1 (กลุ่มเปรียบเทียบ: กลุ่มอ้างอิง) จำนวน 100: 100 คน, 250: 250 คน และ 500: 500 คน ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี IRT LR มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อยู่ในช่วงประมาณ 0.000-0.645 หรือคิดเป็นร้อยละ 0.00-64.50 สำหรับวิธี Poly-SIBTEST มีค่าเฉลี่ยของอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อยู่ในช่วงประมาณ 0.050-0.230 หรือคิดเป็นร้อยละ 5.00-23.00 และวิธี Multiple-groups CFA มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อยู่ในช่วงประมาณ 0.070-0.460 หรือคิดเป็นร้อยละ 7.00-46.00

1. เมื่อพิจารณาภายใต้ปัจจัยความยาวของแบบสอบ จำนวน 9 ข้อ และขนาดของกลุ่มตัวอย่าง (N_F: N_R) ขนาด 100 คน: 100 คน พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี Multiple-groups CFA มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูงสุด คิดเป็นร้อยละ 8.00 สำหรับวิธี Poly-SIBTEST มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ คิดเป็นร้อยละ 5.00 และวิธี IRT LR มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ คิดเป็นร้อยละ 0.00

2. เมื่อพิจารณาภายใต้ปัจจัยความยาวของแบบสอบ จำนวน 9 ข้อ และขนาดของกลุ่มตัวอย่าง (N_F: N_R) 250 คน: 250 คน พบว่า วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี Multiple-groups CFA มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูงสุด คิดเป็นร้อยละ 36.00 สำหรับวิธี IRT LR มีค่าเฉลี่ยของอัตราอำนาจการทดสอบการทำหน้าที่ต่างกันของข้อสอบ คิดเป็นร้อยละ 12.0 และวิธี Poly-SIBTEST มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ คิดเป็นร้อยละ 8.00

3. เมื่อพิจารณาภายใต้ปัจจัยความยาวของแบบสอบ จำนวน 9 ข้อ และขนาดของกลุ่มตัวอย่าง (N_F: N_R) 500 คน: 500 คน พบว่า วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี Poly-SIBTEST มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูงสุด คิดเป็นร้อยละ 23.00 สำหรับวิธี IRT LR มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ คิดเป็นร้อยละ 20.00 และวิธี Multiple-groups CFA มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ คิดเป็นร้อยละ 7.00

4. เมื่อพิจารณาภายใต้ปัจจัยความยาวของแบบสอบ จำนวน 15 ข้อ และขนาดของกลุ่มตัวอย่าง (N_F: N_R) 100 คน: 100 คน พบว่า วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี Multiple-groups CFA มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูงสุด คิดเป็นร้อยละ 24.50 สำหรับวิธี IRT LR มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ คิดเป็นร้อยละ 11.50 และวิธี Poly-SIBTEST มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ คิดเป็นร้อยละ 7.50

5. เมื่อพิจารณาภายใต้ปัจจัยความยาวของแบบสอบ จำนวน 15 ข้อ และขนาดของกลุ่มตัวอย่าง (N_F: N_R) 250 คน: 250 คน พบว่า วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี IRT LR มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูงสุด คิดเป็นร้อยละ 29.50 สำหรับวิธี Multiple-groups CFA มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ คิดเป็นร้อยละ 19.00 และวิธี Poly-SIBTEST มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ คิดเป็นร้อยละ 6.50

6. เมื่อพิจารณาภายใต้ปัจจัยความยาวของแบบสอบ จำนวน 15 ข้อ และขนาดของกลุ่มตัวอย่าง (N_F: N_R) 500 คน: 500 คน พบว่า วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี IRT LR มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูงสุด คิดเป็นร้อยละ 64.50 สำหรับวิธี Multiple-groups CFA มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ คิดเป็นร้อยละ 46.00 และวิธี Poly-SIBTEST

มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ คิดเป็นร้อยละ 9.00

7. โดยภาพรวม การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจการให้คะแนนแบบหลายค่า ภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ขนาดเล็ก (0.10) ด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสามวิธี พบว่า ภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ขนาดเล็ก (0.10) วิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ต่ำกว่าเกณฑ์ที่กำหนดในทุกเงื่อนไขปัจจัย โดยวิธี IRT LR มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูงสุด คิดเป็นร้อยละ 64.50 ภายใต้เงื่อนไขปัจจัยความยาวของแบบสอบ 15 ข้อ และขนาดของกลุ่มตัวอย่าง 500 คน: 500 คน สำหรับวิธี Multiple-groups CFA มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูงสุด คิดเป็นร้อยละ 46.00 ภายใต้เงื่อนไขปัจจัยความยาวของแบบสอบ 15 ข้อ และขนาดของกลุ่มตัวอย่าง 500 คน: 500 คน และวิธี Poly-SIBTEST มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูงสุด คิดเป็นร้อยละ 23.00 ภายใต้เงื่อนไขปัจจัยความยาวของแบบสอบ 9 ข้อ และขนาดของกลุ่มตัวอย่าง 500 คน: 500 คน

ตารางที่ 10 ค่าเฉลี่ยของอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบภายใต้ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ขนาดกลาง (0.50) ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA

ขนาดการทำหน้าที่ต่างกันของข้อสอบ	ความยาวของแบบสอบ	ขนาดของกลุ่มตัวอย่าง (N_F: N_R)	ค่าเฉลี่ยของอัตราอำนาจการทดสอบ		
			IRTLR	POLYSIB	MGCFA
ขนาดกลาง (0.50)	จำนวน 9 ข้อ	100: 100	1.000	0.040	0.920
		250: 250	1.000	0.210	0.950
		500: 500	1.000	0.890	1.000
	จำนวน 15 ข้อ	100: 100	0.930	0.250	0.865
		250: 250	0.980	0.455	0.980
		500: 500	1.000	0.325	0.995

จากผลการวิเคราะห์ตามตารางที่ 10 ค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ภายใต้ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบขนาดกลาง (0.50) ความยาวของแบบสอบ จำนวน 9 ข้อ และ 15 ข้อ และขนาดของกลุ่มตัวอย่างสัดส่วน 1: 1 (กลุ่มเปรียบเทียบ: กลุ่มอ้างอิง) จำนวน 100: 100 คน, 250: 250 คน และ 500: 500 คน ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA พบว่า วิธี IRT LR มีค่าเฉลี่ยของอัตราอำนาจการทดสอบการทำหน้าที่ต่างกันของข้อสอบ อยู่ในช่วงประมาณ 0.930-1.000 หรือคิดเป็นร้อยละ 93.00-100.00 สำหรับวิธี Poly-SIBTEST มีค่าเฉลี่ยของอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อยู่ในช่วงประมาณ 0.040-0.890 หรือคิดเป็นร้อยละ 4.00-89.00 และวิธี Multiple-groups CFA มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อยู่ในช่วงประมาณ 0.865-1.000 หรือคิดเป็นร้อยละ 86.50-100.00

1. เมื่อพิจารณาภายใต้ปัจจัยความยาวของแบบสอบ จำนวน 9 ข้อ และขนาดของกลุ่มตัวอย่าง (N_F: N_R) ขนาด 100 คน: 100 คน พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี IRT LR มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูงสุด คิดเป็นร้อยละ 100.00 สำหรับวิธี MGCFA มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ คิดเป็นร้อยละ 92.00 และวิธี Poly-SIBTEST มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ คิดเป็นร้อยละ 4.00

2. เมื่อพิจารณาภายใต้ปัจจัยความยาวของแบบสอบ จำนวน 9 ข้อ และขนาดของกลุ่มตัวอย่าง (N_F: N_R) ขนาด 250 คน: 250 คน พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี IRT LR มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ สูงสุด คิดเป็นร้อยละ 100.00 สำหรับวิธี MGCFA มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ คิดเป็นร้อยละ 95.00 และวิธี Poly-SIBTEST มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ คิดเป็นร้อยละ 21.00

3. เมื่อพิจารณาภายใต้ปัจจัยความยาวของแบบสอบ จำนวน 9 ข้อ และขนาดของกลุ่มตัวอย่าง (N_F: N_R) ขนาด 500 คน: 500 คน พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี IRT LR และวิธี MGCFA มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำ

หน้าที่ต่างกันของข้อสอบ สูงสุด คิดเป็นร้อยละ 100.00 สำหรับวิธี Poly-SIBTEST มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ คิดเป็นร้อยละ 89.00

4. เมื่อพิจารณาภายใต้ปัจจัยความยาวของแบบสอบ จำนวน 15 ข้อ และขนาดของกลุ่มตัวอย่าง (N_F: N_R) ขนาด 100 คน: 100 คน พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี IRT LR มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ สูงสุด คิดเป็นร้อยละ 93.00 สำหรับวิธี MGCFA มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ คิดเป็นร้อยละ 86.50 และวิธี Poly-SIBTEST มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ คิดเป็นร้อยละ 25.00

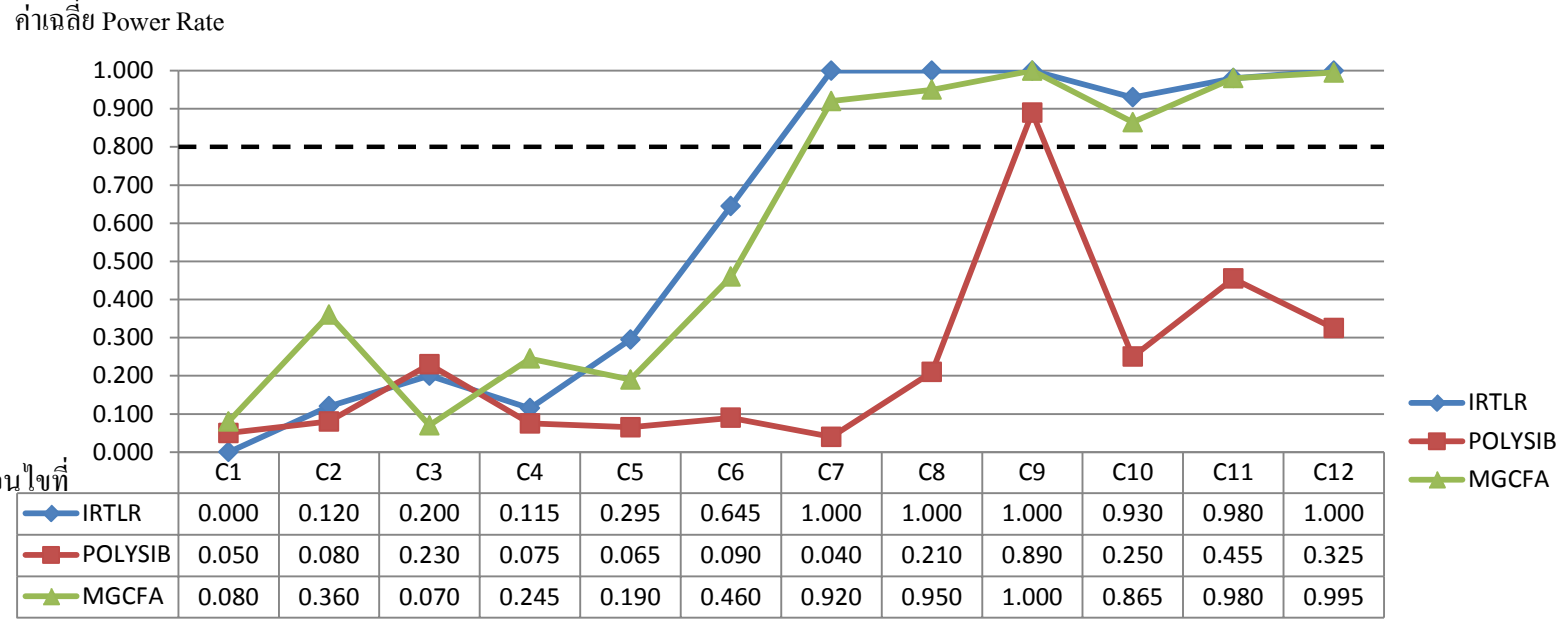
5. เมื่อพิจารณาภายใต้ปัจจัยความยาวของแบบสอบ จำนวน 15 ข้อ และขนาดของกลุ่มตัวอย่าง (N_F: N_R) ขนาด 250 คน: 250 คน พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี IRT LR และวิธี MGCFA มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ สูงสุด คิดเป็นร้อยละ 98.00 เท่ากัน สำหรับวิธี Poly-SIBTEST มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ คิดเป็นร้อยละ 45.50

6. เมื่อพิจารณาภายใต้ปัจจัยความยาวของแบบสอบ จำนวน 15 ข้อ และขนาดของกลุ่มตัวอย่าง (N_F: N_R) ขนาด 500 คน: 500 คน พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี IRT LR มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ สูงสุด คิดเป็นร้อยละ 100.00 สำหรับวิธี MGCFA มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ คิดเป็นร้อยละ 99.50 และวิธี Poly-SIBTEST มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ คิดเป็นร้อยละ 32.50

7. โดยภาพรวม การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจการให้คะแนนแบบหลายค่า ภายใต้เงื่อนไขปัจจัยการทำหน้าที่ต่างกันของข้อสอบ ขนาดกลาง (0.50) ด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสามวิธี พบว่า วิธี IRT LR มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อยู่ในเกณฑ์ที่กำหนด (สูงกว่า 0.80) ในทุกเงื่อนไขปัจจัย และมีค่าเฉลี่ยของอัตราอำนาจการทดสอบของ

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ สูงสุด คิดเป็นร้อยละ 100.00 จำนวน 4 เงื่อนไขปัจจัย คือ เงื่อนไขปัจจัยความยาวของแบบสอบ จำนวน 9 ข้อ และขนาดกลุ่มตัวอย่าง100 คน: 100 คน เงื่อนไขปัจจัยความยาวของแบบสอบ จำนวน 9 ข้อ และขนาดกลุ่มตัวอย่าง 250 คน: 250 คน เงื่อนไขปัจจัยความยาวของแบบสอบ จำนวน 9 ข้อ และขนาดกลุ่มตัวอย่าง 500 คน: 500 คน และเงื่อนไขปัจจัยความยาวของแบบสอบ จำนวน 15 ข้อ และขนาดกลุ่มตัวอย่าง 500 คน: 500 คน สำหรับวิธี Multiple-groups CFA มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อยู่ในเกณฑ์ที่กำหนด (สูงกว่า 0.80) ในทุกเงื่อนไขปัจจัย และมีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ สูงสุด คิดเป็นร้อยละ 100.00 จำนวน 1 เงื่อนไขปัจจัย คือ เงื่อนไขปัจจัยความยาวของแบบสอบ จำนวน 9 ข้อ และขนาดกลุ่มตัวอย่าง500 คน: 500 คน และวิธี Poly-SIBTEST มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อยู่ในเกณฑ์ที่กำหนด (สูงกว่า 0.80) เพียง 1 เงื่อนไขปัจจัย คือ เงื่อนไขปัจจัยความยาวของแบบสอบ จำนวน 9 ข้อ และขนาดกลุ่มตัวอย่าง 500 คน: 500 คน

กราฟแสดงค่าเฉลี่ยของอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบภายใต้ปัจจัยที่แตกต่าง 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

IRTLR คือ วิธีทดสอบอัตราส่วนความควรจะเป็น, POLYSIB คือ วิธี Poly-SIBTEST และ MGCFA คือ วิธี Multiple-groups CFA

ภาพที่ 15 กราฟแสดงค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (Power Rate) ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA

จากภาพที่ 15 ผลการวิเคราะห์ค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA พบว่า โดยภาพรวมของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก ด้วยวิธี IRT LR และวิธี Multiple-groups CFA มีอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ 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

2. ผลการทดสอบค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 และค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก ด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ 3 วิธี

ผลการทดสอบค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 และค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก ด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ 3 วิธี แสดงดังตารางที่ 10-13

ตารางที่ 11 ผลการทดสอบอัตราความคลาดเคลื่อนประเภทที่ 1 ภายใต้ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ขนาดเล็ก (0.10) ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA

ขนาดการทำหน้าที่ต่างกันของข้อสอบ	ความยาวของแบบสอบ	ขนาดของกลุ่มตัวอย่าง (N_F: N_R)	ผลการทดสอบ Z-test อัตราความคลาดเคลื่อนประเภทที่ 1		
			IRTLR	**POLYSIB**	**MGCFA**
ขนาดเล็ก (0.10)	จำนวน 9 ข้อ	100: 100	-1.606*	0.401*	2.810
		250: 250	-1.606*	-0.115*	1.376*
		500: 500	-1.721*	0.057*	0.516*
	จำนวน 15 ข้อ	100: 100	-1.765*	0.497*	3.776
		250: 250	-1.059*	0.038*	3.105
		500: 500	-0.248*	0.285*	3.048

* $Z_{.05} < 1.645$

จากตารางที่ 11 ผลการทดสอบอัตราความคลาดเคลื่อนประเภทที่ 1 ภายใต้ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ขนาดเล็ก (0.10) ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA พบว่า วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ขนาดเล็ก (0.10) ด้วยวิธี IRT LR และวิธี Poly-SIBTEST สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ในทุกเงื่อนไขปัจจัย สำหรับวิธี Multiple-groups CFA สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 เพียง 2 เงื่อนไขปัจจัย คือ เงื่อนไขปัจจัยความยาวของแบบสอบ จำนวน 9 ข้อ และขนาดของกลุ่มตัวอย่าง 250 คน: 250 คน และ 500 คน: 500 คน

ตารางที่ 12 ผลการทดสอบอัตราความคลาดเคลื่อนประเภทที่ 1 ภายใต้ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ขนาดกลาง (0.50) ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA

ขนาดการทำหน้าที่ต่างกันของข้อสอบ	ความยาวของแบบสอบ	ขนาดของกลุ่มตัวอย่าง (N_F: N_R)	ผลการทดสอบ Z-test อัตราความคลาดเคลื่อนประเภทที่ 1		
			IRTLR	**POLYSIB**	**MGCFA**
ขนาดกลาง (0.50)	จำนวน 9 ข้อ	100: 100	-2.065*	0.115*	2.638
		250: 250	-1.491*	0.000*	0.516*
		500: 500	-1.204*	0.688*	1.032*
	จำนวน 15 ข้อ	100: 100	-1.553*	0.850*	4.799
		250: 250	-1.200*	0.497*	6.387
		500: 500	-1.236*	0.002*	3.071

* $Z_{.05} < 1.645$

จากตารางที่ 12 ผลการทดสอบอัตราความคลาดเคลื่อนประเภทที่ 1 ภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ขนาดกลาง (0.50) ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA พบว่า วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ขนาดกลาง (0.50) ด้วยวิธี IRT LR และวิธี Poly-SIBTEST สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ในทุกเงื่อนไขปัจจัย สำหรับวิธี Multiple-groups CFA สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 เพียง 2 เงื่อนไขปัจจัย คือ เงื่อนไขปัจจัยความยาวของแบบสอบ จำนวน 9 ข้อ และขนาดของกลุ่มตัวอย่าง 250 คน: 250 คน และขนาดของกลุ่มตัวอย่าง 500 คน: 500 คน

ตารางที่ 13 ผลการทดสอบอัตราอำนาจการทดสอบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ภายใต้ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ขนาดเล็ก (0.10) ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA

ขนาดการทำหน้าที่ต่างกันของข้อสอบ	ความยาวของแบบสอบ	ขนาดของกลุ่มตัวอย่าง (N_F: N_R)	ผลการทดสอบ Z-test อัตราอำนาจการทดสอบ		
			IRTLR	**POLYSIB**	**MGCFA**
ขนาดเล็ก (0.10)	จำนวน 9 ข้อ	100: 100	-20.000	-18.750	-18.000
		250: 250	-17.000	-18.000	-11.000
		500: 500	-15.000	-14.250	-18.250
	จำนวน 15 ข้อ	100: 100	-17.125	-18.125	-13.875
		250: 250	-12.625	-18.375	-15.250
		500: 500	-3.875	-17.750	-8.500

* $Z_{.05} > -1.645$

จากตารางที่ 13 ผลการทดสอบอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ขนาดเล็ก (0.10) ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA พบว่า วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ภายใต้ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ขนาดเล็ก (0.10) ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA ไม่มีอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ทั้งสามวิธีในทุกเงื่อนไขปัจจัย

ตารางที่ 14 ผลการทดสอบอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ขนาดกลาง (0.50) ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA

ขนาดการทำหน้าที่ต่างกันของข้อสอบ	ความยาวของแบบสอบ	ขนาดของกลุ่มตัวอย่าง (N_F: N_R)	ผลการทดสอบ Z-test อัตราอำนาจการทดสอบ		
			IRTLR	**POLYSIB**	**MGCFA**
ขนาดกลาง (0.50)	จำนวน 9 ข้อ	100: 100	5.000*	-19.000	3.000*
		250: 250	5.000*	-14.750	3.750*
		500: 500	5.000*	2.250*	5.000*
	จำนวน 15 ข้อ	100: 100	3.250*	-13.750	1.625*
		250: 250	4.500*	-8.625	4.500*
		500: 500	5.000*	-11.875	4.875*

* $Z_{.05} > -1.645$

จากตารางที่ 14 ผลการทดสอบอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ขนาดกลาง (0.50) ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA พบว่า วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ขนาดกลาง (0.50) ด้วยวิธี IRT LR และวิธี Multiple-groups CFA มีอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อย่างมีนัยสำคัญทางสถิติระดับ .05 ในทุกเงื่อนไขปัจจัย สำหรับวิธี Poly-SIBTEST มีอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 เพียง 1 เงื่อนไข คือ เงื่อนไขปัจจัยความยาวของแบบสอบ จำนวน 9 ข้อ และขนาดของกลุ่มตัวอย่าง 500 คน: 500 คน

ตอนที่ 2 ผลการเปรียบเทียบอัตราคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก

การวิเคราะห์ในขั้นตอนนี้ ผู้วิจัยดำเนินการวิเคราะห์เปรียบเทียบอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบ GRM ภายใต้เงื่อนไขปัจจัยหลักที่แตกต่าง 3 ปัจจัยหลัก คือ ขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง ด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ 3 วิธี ด้วยวิธีการวิเคราะห์ความแปรปรวนแบบวัดซ้ำ (Repeated measurement) ซึ่งมีตัวแปรวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นตัวแปรวัดซ้ำ (Within-subject factor) และปัจจัยหลัก 3 ปัจจัย เป็นตัวแปรวัดต่างกลุ่ม (Between-subjects factors) และจากการตรวจสอบตามข้อตกลงเบื้องต้นเกี่ยวกับ Compound Symmetry โดยใช้สถิติ Mauchly's test of sphericity พบว่า ไม่เป็น Compound symmetry หมายความว่า ค่าความสัมพันธ์ (Correlation) ของตัวแปรตามแต่ละคู่ที่วัดซ้ำ และความแปรปรวน (Variance) ของตัวแปรตามในการวัดซ้ำแต่ละครั้งของแต่ละกลุ่มแตกต่างกัน ซึ่งถือว่าเป็นการละเมิดข้อตกลงเบื้องต้น ทั้งนี้สามารถปรับแก้โดยใช้สถิติ Greenhouse-geisser epsilon, Huynh-feldt epsilon หรือ Lower-bound epsilon และในการวิจัยครั้งนี้ผู้วิจัยเลือกใช้สถิติ Greenhouse-geisser epsilon (Munro, 2005, p. 215, 233; วลัยภรณ์ อารีรักษ์, 2554, หน้า 99) ดังผลการวิเคราะห์ตามตารางที่ 15-18

ตารางที่ 15 ผลการวิเคราะห์เปรียบเทียบค่าเฉลี่ยอัตราความคลาดเคลื่อนประเภทที่ 1 ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก ด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสามวิธี (Tests of within-subjects effects)

แหล่งของความแปรปรวน	SS	df	MS	F
method	4.981	1.576	3.162	408.973***
method * DIFsize	.042	1.576	.027	3.458*
method * LTESTsize	.486	1.576	.308	39.865***
method * SAMPsize	.254	3.151	.081	10.430***
method * DIFsize * LTESTsize	.079	1.576	.050	6.456**
method * DIFsize * SAMPsize	.011	3.151	.004	.455
method * LTESTsize * SAMPsize	.094	3.151	.030	3.849**
method * DIFsize * LTESTsize * SAMPsize	.065	3.151	.021	2.665*
Error	14.469	1871.639	.008	

*** $p < .001$, ** $p < .01$, * $p < .05$

จากตารางที่ 15 ผลการวิเคราะห์เปรียบเทียบค่าเฉลี่ยความคลาดเคลื่อนประเภทที่ 1 ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก ด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ 3 วิธี คือ วิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA พบว่า วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ มีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .001 (F = 408.973) นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแตกต่างกัน มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกัน

เมื่อพิจารณาผลการตรวจสอบปฏิสัมพันธ์ระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ พบว่า ปฏิสัมพันธ์ระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ มีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 (F = 3.458) นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบและปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ แตกต่างกัน มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกัน

เมื่อพิจารณาผลการตรวจสอบปฏิสัมพันธ์ระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับบัจจัยความยาวของแบบสอบ พบว่า ปฏิสัมพันธ์ระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยความยาวของแบบสอบ มีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .001 (F = 39.865) นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบและปัจจัยความยาวของแบบสอบ แตกต่างกัน มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกัน

เมื่อพิจารณาผลการตรวจสอบปฏิสัมพันธ์ระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยขนาดของกลุ่มตัวอย่าง พบว่า ปฏิสัมพันธ์ระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยขนาดของกลุ่มตัวอย่าง มีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .001 (F = 10.403) นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบและปัจจัยขนาดของกลุ่มตัวอย่าง แตกต่างกัน มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกัน

เมื่อพิจารณาผลการตรวจสอบปฏิสัมพันธ์ระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ และปัจจัยความยาวของแบบสอบ พบว่า ปฏิสัมพันธ์ระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ และปัจจัยความยาวของแบบสอบ มีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 (F = 6.456) นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ กับปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ และปัจจัยความยาวของแบบสอบ มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกัน

เมื่อพิจารณาผลการตรวจสอบปฏิสัมพันธ์ระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง พบว่า ปฏิสัมพันธ์ระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้งสามวิธี กับปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง ไม่มีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 1 นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ กับปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง แตกต่างกัน ไม่มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกัน

เมื่อพิจารณาผลการตรวจสอบปฏิสัมพันธ์ระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยความยาวของแบบสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง พบว่า ปฏิสัมพันธ์ระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้งสามวิธี กับปัจจัยความยาวของแบบสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง มีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 1

อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 (F =3.849) นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ กับปัจจัยความยาวของแบบสอบและปัจจัยขนาดของกลุ่มตัวอย่าง แตกต่างกัน มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกัน

เมื่อพิจารณาผลการตรวจสอบปฏิสัมพันธ์ระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ปัจจัยความยาวของแบบสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง พบว่า ปฏิสัมพันธ์ระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้งสามวิธี กับปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ปัจจัยความยาวของแบบสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง มีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 (F = 2.665) นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ กับปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ปัจจัยความยาวของแบบสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง แตกต่างกัน มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกัน

ตารางที่ 16 ผลการวิเคราะห์เปรียบเทียบค่าเฉลี่ยอัตราความคลาดเคลื่อนประเภทที่ 1 ภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง (Tests of between-subjects effects)

แหล่งของความแปรปรวน	SS	df	MS	F
DIFsize	.022	1	.022	2.969
LTESTsize	.475	1	.475	64.604***
SAMPsize	.042	2	.021	2.881
DIFsize * LTESTsize	.019	1	.019	2.643
DIFsize * SAMPsize	.016	2	.008	1.057
LTESTsize * SAMPsize	.049	2	.024	3.307*
DIFsize * LTESTsize * SAMPsize	.110	2	.055	7.447**
Error	8.734	1188	.007	

*** $p < .001$, ** $p < .01$, * $p < .05$

จากตารางที่ 16 ผลการวิเคราะห์เปรียบเทียบค่าเฉลี่ยอัตราความคลาดเคลื่อนประเภทที่ 1 ภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง พบว่า เงื่อนไขปัจจัยความยาวของแบบสอบ มีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .001 (F = 64.604) สำหรับเงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง ไม่มีผลต่อความเคลื่อนประเภทที่ 1 นั่นคือความยาวของแบบสอบ แตกต่างกัน มีผลให้ความคลาดเคลื่อนประเภทที่ 1 แตกต่างกัน สำหรับเงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง แตกต่างกัน ไม่มีผลให้ความคลาดเคลื่อนประเภทที่ 1 แตกต่างกัน

เมื่อพิจารณาผลการตรวจสอบปฏิสัมพันธ์ระหว่างปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยความยาวของแบบสอบ พบว่า ปฏิสัมพันธ์ระหว่างปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยความยาวของแบบสอบ ไม่มีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 1 นั่นคือ ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบและปัจจัยความยาวของแบบสอบ แตกต่างกัน ไม่มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกัน

เมื่อพิจารณาผลการตรวจสอบปฏิสัมพันธ์ระหว่างปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยขนาดของกลุ่มตัวอย่าง พบว่า ปฏิสัมพันธ์ระหว่างปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยขนาดของกลุ่มตัวอย่าง ไม่มีผลต่อความคลาดเคลื่อนประเภทที่ 1 นั่นคือ ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบและปัจจัยขนาดของกลุ่มตัวอย่าง แตกต่างกัน ไม่มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกัน

เมื่อพิจารณาผลการตรวจสอบปฏิสัมพันธ์ระหว่างปัจจัยความยาวของแบบสอบกับปัจจัยขนาดของกลุ่มตัวอย่าง พบว่า ปฏิสัมพันธ์ระหว่างปัจจัยความยาวของแบบสอบกับปัจจัยขนาดของกลุ่มตัวอย่าง มีผลต่อความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 (F = 3.307) นั่นคือ ปัจจัยความยาวของแบบสอบและปัจจัยขนาดของกลุ่มตัวอย่าง แตกต่างกัน มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกัน

เมื่อพิจารณาผลการตรวจสอบปฏิสัมพันธ์ระหว่างปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยความยาวของแบบสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง พบว่า ปฏิสัมพันธ์ระหว่างปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยความยาวของแบบสอบ ปัจจัยความยาวของแบบสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง มีผลต่อความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 (F = 7.447) นั่นคือ ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยความยาวของแบบสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง แตกต่างกัน มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกัน

ตารางที่ 17 ผลการวิเคราะห์เปรียบเทียบค่าเฉลี่ยอำนาจการทดสอบ ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก ด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสามวิธี (Tests of within-subjects effects)

แหล่งของความแปรปรวน	SS	df	MS	F
method	112.850	1.949	57.893	858.326***
method * DIFsize	42.986	1.949	22.052	326.946***
method * LTESTsize	2.689	1.949	1.380	20.455***
method * SAMPsize	4.065	1.949	1.043	15.461***
method * DIFsize * LTESTsize	2.219	1.949	1.139	16.880***
method * DIFsize * SAMPsize	12.127	3.899	3.111	46.120***
method * LTESTsize * SAMPsize	20.442	3.899	5.243	77.738***
method * DIFsize * LTESTsize * SAMPsize	1.926	3.899	.494	7.323***
Error	156.195	2315.759	.067	

*** $p < .001$, ** $p < .01$, * $p < .05$

จากตารางที่ 17 ผลการวิเคราะห์เปรียบเทียบค่าเฉลี่ยอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก ด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ 3 วิธี คือ วิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA พบว่า วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ มีผลต่ออำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อย่างมีนัยสำคัญทางสถิติที่ระดับ .001 (F = 858.326) นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแตกต่างกัน มีผลให้อำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ แตกต่างกัน

เมื่อพิจารณาผลการตรวจสอบปฏิสัมพันธ์ระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ พบว่า ปฏิสัมพันธ์ระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ มีผลต่ออำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อย่างมีนัยสำคัญทางสถิติที่ระดับ .001 (F = 326.946) นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ และปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ แตกต่างกัน มีผลให้อัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ แตกต่างกัน

เมื่อพิจารณาผลการตรวจสอบปฏิสัมพันธ์ระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับบัจจัยความยาวของแบบสอบ พบว่า ปฏิสัมพันธ์ระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยความยาวของแบบสอบ มีผลต่ออัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อย่างมีนัยสำคัญทางสถิติที่ระดับ .001 (F = 20.455) นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบและปัจจัยความยาวของแบบสอบ แตกต่างกัน มีผลให้อัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ แตกต่างกัน

เมื่อพิจารณาผลการตรวจสอบปฏิสัมพันธ์ระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยขนาดของกลุ่มตัวอย่าง พบว่า ปฏิสัมพันธ์ระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยขนาดของกลุ่มตัวอย่าง มีผลต่ออัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อย่างมีนัยสำคัญทางสถิติที่ระดับ .001 (F = 15.461) นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบและปัจจัยขนาดของกลุ่มตัวอย่าง แตกต่างกัน มีผลให้อัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ แตกต่างกัน

เมื่อพิจารณาผลการตรวจสอบปฏิสัมพันธ์ระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ และปัจจัยความยาวของแบบสอบ พบว่า ปฏิสัมพันธ์ระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ และปัจจัยความยาวของแบบสอบ มีผลต่ออัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อย่างมีนัยสำคัญทางสถิติที่ระดับ .001 (F = 16.880) นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ กับปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ และปัจจัยความยาวของแบบสอบ มีผลให้อัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ แตกต่างกัน

เมื่อพิจารณาผลการตรวจสอบปฏิสัมพันธ์ระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง พบว่า ปฏิสัมพันธ์ระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้งสามวิธี กับปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง มีผลต่ออัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อย่างมีนัยสำคัญทางสถิติที่ระดับ .001 (F = 46.120) นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ กับปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง แตกต่างกัน มีผลให้อัตราอำนาจการทดสอบการทำหน้าที่ต่างกันของข้อสอบ แตกต่างกัน

เมื่อพิจารณาผลการตรวจสอบปฏิสัมพันธ์ระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยความยาวของแบบสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง พบว่า ปฏิสัมพันธ์ระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้งสามวิธี กับปัจจัยความยาวของแบบสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง มีผลต่ออัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกัน อย่างมีนัยสำคัญทางสถิติที่ระดับ .001 (F = 77.738) นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยความยาวของแบบสอบ และปัจจัยขนาดของกลุ่มตัวอย่างแตกต่างกัน มีผลให้อัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแตกต่างกัน

เมื่อพิจารณาผลการตรวจสอบปฏิสัมพันธ์ระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ปัจจัยความยาวของแบบสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง พบว่า ปฏิสัมพันธ์ระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้งสามวิธี กับปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ปัจจัยความยาวของแบบสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง มีผลต่ออัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อย่างมีนัยสำคัญทางสถิติที่ระดับ .001 (F = 7.323) นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ กับปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ปัจจัยความยาวของแบบสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง แตกต่างกัน มีผลให้อัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ แตกต่างกัน

ตารางที่ 18 ผลการวิเคราะห์เปรียบเทียบค่าเฉลี่ยอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง (Tests of between-subjects effects)

แหล่งของความแปรปรวน	SS	df	MS	F
DIFsize	294.694	1	294.694	3057.616***
LTESTsize	1.138	1	1.138	11.805**
SAMPsize	22.743	2	11.371	117.985***
DIFsize * LTESTsize	3.361	1	3.361	34.873***
DIFsize * SAMPsize	.121	2	.060	.628
LTESTsize * SAMPsize	.390	2	.195	2.024
DIFsize * LTESTsize * SAMPsize	11.613	2	5.806	60.244***
Error	114.500	1188	.096	

*** $p < .001$, ** $p < .01$, * $p < .05$

จากตารางที่ 18 ผลการวิเคราะห์เปรียบเทียบค่าเฉลี่ยอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง พบว่า เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง มีผลต่ออำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อย่างมีนัยสำคัญทางสถิติที่ระดับ .001 ($F = 3057.616$ และ $F = 177.985$ ตามลำดับ) และเงื่อนไขปัจจัยความยาวของแบบสอบ มีผลต่ออำนาจการทดสอบการทำหน้าที่ต่างกันของข้อสอบ อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ($F = 11.805$) นั่นคือ ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง แตกต่างกัน มีผลให้อำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ แตกต่างกัน

เมื่อพิจารณาผลการตรวจสอบปฏิสัมพันธ์ระหว่างปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยความยาวของแบบสอบ พบว่า ปฏิสัมพันธ์ระหว่างปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยความยาวของแบบสอบ มีผลต่ออำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อย่างมีนัยสำคัญทางสถิติที่ระดับ .001 ($F = 34.873$) นั่นคือ ปัจจัย

ขนาดการทำหน้าที่ต่างกันของข้อสอบ และปัจจัยความยาวของแบบสอบ แตกต่างกัน มีผลให้อัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ แตกต่างกัน

เมื่อพิจารณาผลการตรวจสอบปฏิสัมพันธ์ระหว่างปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยขนาดของกลุ่มตัวอย่าง พบว่า ปฏิสัมพันธ์ระหว่างปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยขนาดของกลุ่มตัวอย่าง ไม่มีผลต่ออำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ นั่นคือ ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบและปัจจัยขนาดของกลุ่มตัวอย่าง แตกต่างกัน ไม่มีผลให้อัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต้างกันของข้อสอบ แตกต่างกัน

เมื่อพิจารณาผลการตรวจสอบปฏิสัมพันธ์ระหว่างปัจจัยความยาวของแบบสอบกับปัจจัยขนาดของกลุ่มตัวอย่าง พบว่า ปฏิสัมพันธ์ระหว่างปัจจัยความยาวของแบบสอบกับปัจจัยขนาดของกลุ่มตัวอย่าง ไม่มีผลต่ออำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ นั่นคือ ปัจจัยความยาวของแบบสอบและปัจจัยขนาดของกลุ่มตัวอย่าง แตกต่างกัน ไม่มีผลให้อัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ แตกต่างกัน

เมื่อพิจารณาผลการตรวจสอบปฏิสัมพันธ์ระหว่างปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยความยาวของแบบสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง พบว่า ปฏิสัมพันธ์ระหว่างปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยความยาวของแบบสอบ ปัจจัยความยาวของแบบสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง มีผลต่ออำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ อย่างมีนัยสำคัญทางสถิติที่ระดับ .001 ($F = 60.244$) นั่นคือ ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบกับปัจจัยความยาวของแบบสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง แตกต่างกัน มีผลให้อัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ แตกต่างกัน

บทที่ 5
สรุปผล อภิปรายผล และข้อเสนอแนะ

การวิจัยครั้งนี้มีวัตถุประสงค์ที่สำคัญ 2 ประการ คือ ประการแรก เพื่อเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 12 เงื่อนไข (2x2x3) คือ ขนาดการทำหน้าที่ต่างกันของข้อสอบ (ขนาดเล็ก และขนาดกลาง) ความยาวของแบบสอบ 2 รูปแบบ (9 ข้อ และ 15 ข้อ) และขนาดของกลุ่มตัวอย่าง 3 ขนาด (200 คน, 500 คน และ 1,000 คน) ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA โดยในแต่ละเงื่อนไขจำลองข้อมูลวนซ้ำ 100 ครั้ง โดยพิจารณาจากอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ และประการที่สอง เพื่อเปรียบเทียบอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่า ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก คือ ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ปัจจัยความยาวของแบบสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง

การดำเนินการวิจัยครั้งนี้ เป็นการศึกษาข้อมูลจำลอง โดยใช้โมเดล Graded-Response (GRM) ภายใต้ทฤษฎีการตอบสนองข้อสอบ จำลองรูปแบบของผลการตอบของแบบสอบที่มีโครงสร้างวัดความสามารถแบบมิติเดียว จำนวน 9 ข้อ และ 15 ข้อ ข้อสอบทุกข้อมีรายการคำตอบ 5 รายการ ให้คะแนนเป็น 0, 1, 2, 3 และ 4 ตามลำดับ การจำลองข้อมูลใช้โปรแกรม WinGen ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก รวมข้อมูลทั้งหมดที่ต้องจัดกระทำเพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ จำนวน 12 เงื่อนไข (2x2x3) และในแต่ละเงื่อนไขจำลองข้อมูลวนซ้ำ 100 รอบ จากนั้นนำข้อมูลที่ได้จากการจำลองทั้งหมดมาตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสามวิธี คือ วิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA โดยใช้โปรแกรม IRTPRO โปรแกรม SIBTEST และโปรแกรม Mplus

สรุปผลการวิจัย

การสรุปผลการวิจัยจำแนกตามวัตถุประสงค์โดยแบ่งออกเป็น 2 ตอน คือ ตอนที่ 1 ผลการตรวจสอบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA และตอนที่ 2 ผลการเปรียบเทียบอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก โดยมีรายละเอียดดังนี้

1. ผลการตรวจสอบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA

1.1 ผลการวิเคราะห์ค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 ของการทดสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA ภายใต้ปัจจัยที่แตกต่าง 3 ปัจจัย คือ ขนาดการทำหน้าที่แตกต่างกัน ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง พบว่า โดยภาพรวม วิธี IRT LR มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำสุด คิดเป็นร้อยละ 0.50 ภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบขนาดกลาง ความยาวของแบบสอบ จำนวน 9 ข้อ และขนาดของกลุ่มตัวอย่าง 100 คน: 100 คน สำหรับวิธี Poly-SIBTEST มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำสุด คิดเป็นร้อยละ 4.80 ภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ขนาดเล็ก ความยาวของแบบสอบ จำนวน 9 ข้อ และขนาดของกลุ่มตัวอย่าง 250 คน: 250 คน และวิธี Multiple-groups CFA มีค่าเฉลี่ยของอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำสุด คิดเป็นร้อยละ 6.10 ภายใต้เงื่อนไขปัจจัยความยาวของแบบสอบ จำนวน 9 ข้อ และขนาดของกลุ่มตัวอย่างเพิ่มมากขึ้น

1.2 ผลการตรวจสอบประสิทธิภาพการควบคุมความคลาดเคลื่อนประเภทที่ 1 ของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก คือ ขนาดการทำหน้าที่แตกต่างกัน ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA พบว่า วิธี IRT LR และวิธี Poly-SIBTEST สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดีในทุกเงื่อนไขปัจจัย สำหรับวิธี Multiple-groups CFA สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดี ภายใต้เงื่อนไขปัจจัยความยาวของแบบสอบ จำนวน 9 ข้อ และขนาดของกลุ่มตัวอย่าง 250 คน: 250 คน และ 500 คน: 500 คน

1.3 ผลการวิเคราะห์ค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA ภายใต้ปัจจัยที่แตกต่าง 3 ปัจจัย คือ ขนาดการทำหน้าที่แตกต่างกัน ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง พบว่า โดยภาพรวม วิธี IRT LR มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ สูงสุด คิดเป็นร้อยละ 100.00 ภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบและขนาดความยาวของแบบสอบ เพิ่มมากขึ้น สำหรับวิธี Poly-SIBTEST มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ สูงสุด คิดเป็นร้อยละ 89.00 ภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบขนาดเล็ก และขนาดของกลุ่มตัวอย่าง เพิ่มมากขึ้น และวิธี Multiple-groups CFA มีค่าเฉลี่ยของอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ สูงสุด คิดเป็นร้อยละ 100.00 ภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ขนาดกลาง ความยาวของแบบสอบและขนาดของกลุ่มตัวอย่าง เพิ่มมากขึ้น

1.4 ผลการตรวจสอบประสิทธิภาพอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก คือ ขนาดการทำหน้าที่แตกต่างกัน ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA พบว่า วิธี IRT LR และวิธี Multiple-groups CFA มีอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ สูงกว่าวิธี Poly-SIBTEST ภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบขนาดกลางในทุกเงื่อนไขปัจจัย สำหรับวิธี Poly-SIBTEST มีอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูง ภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบขนาดกลาง ความยาวของแบบสอบ จำนวน 9 ข้อ และขนาดของกลุ่มตัวอย่าง เพิ่มมากขึ้น

2. ผลการเปรียบเทียบอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก

2.1 ผลการวิเคราะห์เปรียบเทียบค่าเฉลี่ยอัตราความคลาดเคลื่อนประเภทที่ 1 ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก ด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสามวิธี พบว่า วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ แตกต่างกัน มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 แตกต่างกัน สำหรับผลการวิเคราะห์เปรียบเทียบค่าเฉลี่ยอัตราความคลาดเคลื่อนประเภทที่ 1 ภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ

ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง พบว่า เงื่อนไขปัจจัยความยาวของแบบสอบแตกต่างกัน มีผลให้ความคลาดเคลื่อนประเภทที่ 1 แตกต่างกัน สำหรับเงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ และปัจจัยขนาดของกลุ่มตัวอย่าง แตกต่างกัน ไม่มีผลให้ความคลาดเคลื่อนประเภทที่ 1 แตกต่างกัน

2.2 ผลการวิเคราะห์เปรียบเทียบค่าเฉลี่ยอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก ด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสามวิธี พบว่า วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแตกต่างกัน มีผลให้อำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแตกต่างกัน และผลการวิเคราะห์เปรียบเทียบค่าเฉลี่ยอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง พบว่า ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง แตกต่างกัน มีผลให้อำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ แตกต่างกัน

อภิปรายผลการวิจัย

การอภิปรายผลการวิจัยครั้งนี้ นำเสนอในสองประเด็นหลักตามวัตถุประสงค์ คือ ประเด็นแรก ผลการตรวจสอบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก คือ ขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA และประเด็นที่สอง ผลการเปรียบเทียบอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก โดยมีรายละเอียดดังนี้

1. ผลการตรวจสอบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA

1.1 จากการตรวจสอบประสิทธิภาพการควบคุมความคลาดเคลื่อนประเภทที่ 1 ของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก คือ ขนาดการทำหน้าที่แตกต่างกัน ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง ด้วยวิธี IRT LR และวิธี Poly-SIBTEST สามารถควบคุม

ความคลาดเคลื่อนประเภทที่ 1 ได้ดีในทุกเงื่อนไขปัจจัย สอดคล้องกับผลการศึกษาความไม่แปรเปลี่ยนในการวัดแบบสอบโดยการเปรียบเทียบด้วยวิธี 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 et al. (2009) พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีขนาดการทำหน้าที่ต่างกันของข้อสอบขนาดเล็ก และค่าอำนาจจำแนกมีค่าต่ำ ($a = 0.6$) ด้วยวิธี IRT LR สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดี และมีอำนาจการทดสอบสูง นอกจากนี้ ผลการศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในโมเดล GRM ของ Cohen, Kim, and Baker (1993) พบว่า เมื่อความยาวของแบบสอบเพิ่มขึ้น วิธี Poly-SIBTEST มีอัตราความคลาดเคลื่อนประเภทที่ 1 เพิ่มขึ้นด้วย ซึ่งสอดคล้องกับงานวิจัยในประเทศ ดังเช่น งานวิจัยของ อาวีพร ปานทอง (2558) ที่ได้ศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีโครงสร้างวัดความสามารถมิติเดียว (Unidimensional) ที่ให้คะแนนแบบหลายค่า พบว่า วิธี IRT LR มีอัตราความคลาดเคลื่อนประเภทที่ 1 ต่ำ และมีอำนาจการทดสอบ อยู่ในขอบเขตที่กำหนด และ 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 ความแตกต่างของการแจกแจงความสามารถและขนาดของกลุ่มตัวอย่าง จึงมีผลกระทบต่ออัตราความคลาดเคลื่อนประเภทที่ 1 น้อยมาก รวมถึงมีการคำนวณที่ง่าย ไม่ซับซ้อน และไม่จำเป็นต้องใช้กลุ่มตัวอย่างขนาดใหญ่ (Chang et al., 1995 cilted in Potenza & Doran, 1995, p. 31 (อ้างถึงใน อุทัยวรรณ สายพัฒนะ, 2547, หน้า 19) นอกจากนี้ วิธี Poly-SIBTEST มีข้อได้เปรียบคือ สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ดีเกือบทุกเงื่อนไขปัจจัยที่ศึกษา (อรินทร์ น่วมถนอม, 2548, หน้า 277) และการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี IRT LR ให้ผลดีในกลุ่มตัวอย่างขนาดเล็ก (Stark et al., 2006) รวมถึงในการศึกษาวิจัยครั้งนี้ผู้วิจัยกำหนดสัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันไว้ในช่วง 10% แต่ไม่ได้กำหนดเป็นเงื่อนไขในการศึกษา ซึ่ง อรินทร์ น่วมถนอม (2548) ได้ศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่วัดความสามารถหลายมิติและให้คะแนนหลายค่าด้วยวิธี Poly-SIBTEST พบว่า ปัจจัยสัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันในช่วง 10% ถึง 30% ไม่มีผลต่อประสิทธิภาพของวิธี Poly-SIBTEST สำหรับวิธี Multiple-groups CFA สามารถควบคุม

ความคลาดเคลื่อนประเภทที่ 1 ได้ดี ภายใต้เงื่อนไขปัจจัยความยาวของแบบสอบ จำนวน 9 ข้อ และขนาดของกลุ่มตัวอย่าง 250 คน: 250 คน และ 500 คน: 500 คน

1.2 จากการตรวจสอบประสิทธิภาพอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก คือ ขนาดการทำหน้าที่แตกต่างกัน ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง ด้วยวิธี IRT LR และวิธี Multiple-groups CFA มีอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูง เมื่อปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบเพิ่มขึ้น และมีอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูงกว่าวิธี Poly-SIBTEST เกือบทุกเงื่อนไข ซึ่งสอดคล้องกับผลการศึกษาของ Kim and Yoon (2011) ที่พบว่า ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ด้วยวิธี MCCFA และวิธี IRT LR มีอำนาจการทดสอบสูงเมื่อขนาดการทำหน้าที่ต่างกันของข้อสอบมีขนาดใหญ่ (0.04) และมีขนาดของกลุ่มตัวอย่างเพิ่มมากขึ้นเกือบทุกเงื่อนไขปัจจัย และสอดคล้องกับผลการศึกษาของทองอยู่ สาระ (2543), วลีมาศ แซ่อึ้ง (2543), French and Miller (1996), Krisjansson, Aylesworth, Mcdowell, and Zombo (2005), Narayanan and Swaminathan (1994, 1996), Rogers and Swaminathan (1993), Whitmore and Schumacker (1999) อ้างถึงใน อรินทร์ น่วมถนอม (2549) พบว่า เมื่อขนาดของกลุ่มตัวอย่างเพิ่มขึ้นมีผลทำให้วิธี Poly-SIBTEST มีอัตราอำนาจการทดสอบเพิ่มขึ้นด้วย นอกจากนี้ Kim and Yoon (2011) พบว่า วิธี Multiple-group Categorical CFA (MCCFA) มีอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ดีเมื่อขนาดการทำหน้าที่มีขนาดใหญ่ ทั้งนี้ วิธี Poly-SIBTEST เป็นวิธีที่มีแนวคิดพื้นฐานมาจากทฤษฎีการตอบสนองข้อสอบแบบหลายมิติ (Multidimensional) ซึ่ง Chang et al. (1996) ได้ปรับขยายมาจากวิธี SIBTEST ของ Shealy and Stout (1993) เพื่อใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ซึ่งมีข้อจำกัดคือไม่สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบที่ไม่เป็นรูปแบบเดียวกัน (Nonuniform DIF) จึงส่งผลให้อัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ มีค่าต่ำ แต่มีข้อได้เปรียบคือสามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดี (อรินทร์ น่วมถนอม, 2549, หน้า 143-144)

2. ผลการเปรียบเทียบอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสามวิธี ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย

2.1 จากการวิเคราะห์เปรียบเทียบค่าเฉลี่ยอัตราความคลาดเคลื่อนประเภทที่ 1 ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย คือ ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง (Tests of between-subjects effects) ที่แตกต่างกัน มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 มีค่าเฉลี่ยแตกต่างกัน โดยเฉพาะปัจจัยความยาวของแบบสอบ มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 1 มีค่าเฉลี่ยแตกต่างกัน สอดคล้องกับผลการศึกษาของ Cohen et al. (1993) พบว่า เมื่อความยาวของแบบสอบเพิ่มขึ้น วิธีโพลี-ซิปเทสท์มีอัตราความคลาดเคลื่อนประเภทที่ 1 เพิ่มขึ้นด้วย โดย Potenza and Dorans (1995) และ อาวีพร ปานทอง (2558) พบว่า วิธี Poly-SIBTEST สามารถใช้ได้ดีกับแบบสอบสั้นที่มีข้อสอบหรือข้อคำถามจำนวนน้อย และ Chang et al. (1996) ยังพบว่า เมื่อขนาดของกลุ่มตัวอย่างเพิ่มขึ้น ไม่มีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 1 ซึ่งสอดคล้องกับผลการศึกษาของ Bolt (2002) พบว่า เมื่อขนาดของกลุ่มตัวอย่างเพิ่มขึ้นจาก 300 คน เป็น 1,000 คนต่อกลุ่ม การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี Poly-SIBTEST มีอัตราความคลาดเคลื่อนประเภทที่ 1 ค่อนข้างคงที่ ทั้งนี้ อรินทร์ น่วมถนอม (2549) ได้กล่าวว่า วิธี Poly-SIBTEST เป็นวิธีในกลุ่มทฤษฎีการตอบสนองข้อสอบรูปแบบนันพาราเมตริก (Nonparametric form) จึงไม่จำเป็นต้องใช้โมเดลประมาณค่าพารามิเตอร์ ดังนั้น จึงไม่มีข้อตกลงของโมเดลที่ใช้อธิบายความสัมพันธ์ระหว่างผลการตอบข้อสอบกับตัวแปรจับคู่ (Ackerman, 1994, p. 76; Chang et al., 1996, p. 334; Potenza, & Dorans, 1995, p. 24 อ้างถึงใน อรินทร์ น่วมถนอม, 2549)

2.2 ผลการวิเคราะห์เปรียบเทียบค่าเฉลี่ยอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย คือ ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง (Tests of between-subjects effects) ที่แตกต่างกัน มีผลให้อัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ มีค่าเฉลี่ยแตกต่างกัน ในทุกเงื่อนไขปัจจัย ซึ่งสอดคล้องกับ Lopez Rivas et al. (2009) พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี IRT LR สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดี และมีอำนาจการทดสอบสูง เมื่อขนาดการทำหน้าที่ต่างกันของข้อสอบมีขนาดใหญ่ และขนาดของกลุ่มตัวอย่างเพิ่มขึ้น และผลการศึกษาของ Narayanan and Swaminathan (1994, pp. 315-328) ที่ได้ศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่เป็น

รูปแบบเดียวกัน และไม่เป็นรูปแบบเดียวกัน ระว่างวิธีแมนเทล-แฮนส์เซล กับวิธีซิปเทสท์ พบว่า ปัจจัยขนาดของกลุ่มตัวอย่าง การแจกแจงความสามารถสัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน ขนาดการทำหน้าที่ต่างกันของข้อสอบ และประเภทของข้อสอบ มีผลต่ออัตราอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีซิปเทสท์ อย่างมีนัยสำคัญ และผลการศึกษาของ อุทัยวรรณ สายพัฒนะ (2547) พบว่า ขนาดของกลุ่มตัวอย่างมีผลต่อประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยเสนอแนะว่า ในการศึกษาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการให้คะแนนแบบหลายค่าด้วยวิธี Poly-SIBTEST ควรใช้กลุ่มตัวอย่างที่มีขนาดไม่ต่ำกว่า 500 คน หรือกลุ่มตัวอย่างที่มีขนาดใหญ่ จะทำให้ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมีอำนาจการทดสอบสูง ถึงแม้จะมีความคลาดเคลื่อนประเภทที่ 1 สูงขึ้นด้วย ซึ่งสอดคล้องกับ อาวีพร ปานทอง (2558) พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนแบบหลายค่า ด้วยวิธี Poly-SIBTEST ไม่เหมาะสมกับขนาดของกลุ่มตัวอย่างขนาดใหญ่

ข้อเสนอแนะ

ข้อเสนอแนะในการนำผลการวิจัยไปใช้

1. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดล Graded-response แบบมิติเดียว ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย คือ ขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง ในทางปฏิบัติสถานการณ์จริง ควรใช้วิธี IRT LR ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เนื่องจากมีประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูง โดยสามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดี และมีอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสูง

2. ผลการศึกษาครั้งนี้ พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี Poly-SIBTEST ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดีในทุกเงื่อนไขปัจจัย ซึ่งมีประสิทธิภาพใกล้เคียงกับวิธี IRT LR แต่เนื่องด้วยวิธี Poly-SIBTEST มีอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ สูงภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบที่มีขนาดเล็ก และขนาดของกลุ่มตัวอย่างมีขนาดเพิ่มขึ้น (500 คน: 500 คน) ดังนั้น ในทางปฏิบัติ การศึกษาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า หากพบว่า ขนาดการทำหน้าที่ต่างกันของข้อสอบมีขนาดเล็ก และกลุ่มตัวอย่างมีขนาดใหญ่ (ไม่ควรต่ำกว่า 500 คนต่อกลุ่ม)

ขึ้นไป ควรใช้วิธี Poly-SIBTEST ซึ่งจะทำให้ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมีอำนาจการทดสอบในการตรวจสอบสูง

3. ผลการศึกษาครั้งนี้ พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี Multiple-Groups CFA ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดีภายใต้เงื่อนไขปัจจัยความยาวของแบบสอบที่มีข้อสอบหรือข้อคำถามจำนวนน้อย (จำนวน 9 ข้อ) และขนาดของกลุ่มตัวอย่างมีจำนวนมาก (จำนวน 500 คนต่อกลุ่ม) โดยมีอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ สูงภายใต้เงื่อนไขปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบที่มีขนาดกลาง (0.50) ในทุกเงื่อนไขปัจจัย และเมื่อขนาดของกลุ่มตัวอย่างเพิ่มขึ้น ค่าอำนาจการทดสอบมีแนวโน้มสูงขึ้นด้วย ดังนั้น ในการศึกษาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า ด้วยวิธี Multiple-groups CFA ควรจะเลือกใช้กับแบบสอบที่มีขนาดการทำหน้าที่ต่างกันของข้อสอบที่มีขนาดกลาง ความยาวของแบบสอบจำนวนน้อย และขนาดของกลุ่มตัวอย่างมีจำนวนมาก จะทำให้สามารถควบคุมความคลาดเคลื่อนได้ดี และมีอำนาจการทดสอบสูง

4. ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบค่าในโมเดล Graded-response แบบมิติเดียว ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัย ด้วยวิธี IRT LR วิธี Poly-SIBTEST และวิธี Multiple-groups CFA ควรคำนึงถึงปัจจัยความยาวของแบบสอบที่เหมาะสมสำหรับการนำไปใช้ เนื่องจากการศึกษาวิจัยครั้งนี้พบว่า ความยาวของแบบสอบ มีผลให้อัตราความคลาดเคลื่อนประเภทที่ 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 มีแนวโน้มเพิ่มขึ้นเมื่อความยาวของแบบสอบเพิ่มขึ้น รวมถึงมีอำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ต่ำลง ในขณะที่วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี Multiple-groups CFA ไม่สามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้เมื่อความยาวของแบบสอบเพิ่มขึ้น นอกจากนี้ ปัจจัยขนาดการทำหน้าที่ต่างกันของข้อสอบ ความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่าง ทั้งสามปัจจัย มีผลให้อำนาจการทดสอบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ แตกต่างกัน 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

3. ควรมีการศึกษาในลักษณะเดียวกันนี้โดยใช้ข้อมูลจริงเพื่อเปรียบเทียบกับการศึกษาในครั้งนี้ และการใช้ตัวแปรเพศ อายุ ภูมิลำเนา หรือตัวแปรอื่น ๆ ที่เกี่ยวกับคุณลักษณะของผู้สอบมาเป็นเกณฑ์ในการแบ่งกลุ่มผู้สอบ แล้วพิจารณาผลการตรวจสอบให้ผลแตกต่าง หรือสอดคล้องกับการศึกษาครั้งนี้หรือไม่ อย่างไร รวมถึงการศึกษาประสิทธิภาพของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเมื่อมีผู้สอบหลายกลุ่ม ด้วยวิธีการตรวจสอบในกลุ่ม IRT และ non-IRT เพื่อให้ได้ข้อมูลสารสนเทศที่เหมาะสม น่าเชื่อถือ และสะดวกรวดเร็วสำหรับการนำไปใช้ในทางปฏิบัติได้จริง

บรรณานุกรม

ญาณภัทร สีหะมงคล. (2540). *การเปรียบเทียบผลการตรวจข้อสอบที่ทำหน้าที่ต่างกัน ระหว่างวิธี Lord's χ^2 วิธี Raju's Area Measures และวิธี Closed Interval Area.* ปริญญานิพนธ์การศึกษาดุษฎีบัณฑิต, สาขาวิชาการทดสอบและการวัดผลการศึกษา, บัณฑิตวิทยาลัย, มหาวิทยาลัยศรีนครินทรวิโรฒ.

ทองอยู่ สาระ. (2543). *การเปรียบเทียบอำนาจการการตรวจสอบการจำแนกผิดพลาดในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอและแบบไม่สม่ำเสมอระหว่างวิธีแมนเทล-แฮนส์เซล และวิธีถดถอยโลจิสติกโดยใช้ความยาวของแบบสอบและขนาดกลุ่มตัวอย่างต่างกัน.* วิทยานิพนธ์การศึกษามหาบัณฑิต, สาขาวิชาการวัดผลการศึกษา, บัณฑิตวิทยาลัย, มหาวิทยาลัยศรีนครินทรวิโรฒ.

ธเกียรติกมล ทองงอก. (2554). *ประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในวิธีถดถอยโลจิสติก โดยใช้เกณฑ์อิทธิพล 2 วิธี สำหรับข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบทวิภาค: ข้อมูลจำลองและข้อมูลเชิงประจักษ์.* วิทยานิพนธ์ครุศาสตรดุษฎีบัณฑิต, สาขาวิชาการวัดและประเมินผลการศึกษา, บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย.

นงลักษณ์ วิรัชชัย. (2540). *ความไม่แปรเปลี่ยนของแบบจำลองการเป็นสมาชิกด้วยในรักของครูระหว่างบุคลากรครู 2 กลุ่ม: การประยุกต์ใช้การสร้างแบบจำลองสมการโครงสร้างชนิดกลยุทธ์กลุ่มพหุ.* กรุงเทพฯ: จุฬาลงกรณ์มหาวิทยาลัย.

นิคม กีรติวรางกูร. (2542). *การเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ระหว่างวิธีการวิเคราะห์องค์ประกอบจำกัด แมนเทล-แฮนน์เซล และการตอบสนองข้อสอบ.* วิทยานิพนธ์ครุศาสตรมหาบัณฑิต, สาขาวิชาการวัดผลและประเมินผลการศึกษา, บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย.

พรรณี จินตมาศ. (2540). *การเปรียบเทียบผลการวิเคราะห์ความลำเอียงของข้อสอบโดยใช้ขนาดกลุ่มผู้สอบและวิธีวิเคราะห์ต่างกัน.* วิทยานิพนธ์การศึกษามหาบัณฑิต, สาขาวิชาการวัดผลการศึกษา, บัณฑิตวิทยาลัย, มหาวิทยาลัยศรีนครินทรวิโรฒ.

รัชรินทร์ มุคดา. (2540). *การเปรียบเทียบประสิทธิภาพระหว่างวิธีแมนเทล-แฮนส์เซลกับวิธีถดถอยโลจิสติกในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปในกรณีการจัดกลุ่มความสามารถ ค่าความยากของข้อสอบ และค่าอำนาจจำแนกของข้อสอบต่างกัน.* วิทยานิพนธ์ครุศาสตรมหาบัณฑิต, สาขาวิชาการวัดผลและประเมินผลการศึกษา, บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย.

วลัยภรณ์ อารีรักษ์. (2554). *ผลของโปรแกรมการส่งเสริมพฤติกรรมการป้องกันการหกล้มต่อการรับรู้ความสามารถตนเอง ความคาดหวังผลดีจากการปฏิบัติและพฤติกรรมการป้องกันการหกล้มของผู้สูงอายุที่มีความเสี่ยงต่อการหกล้มในชุมชน*. วิทยานิพนธ์พยาบาลศาสตรมหาบัณฑิต, สาขาวิชาการพยาบาลผู้สูงอายุ, คณะพยาบาลศาสตร์, มหาวิทยาลัยบูรพา.

วลีมาศ แซ่อึ้ง. (2543). *การเปรียบเทียบอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูประหว่างวิธีซิปเทสท์ปรับใหม่ วิธีซิปเทสท์ วิธีแมนเทล-แฮนส์เซล และวิธีการถดถอยโลจิสติก*. วิทยานิพนธ์ครุศาสตรดุษฎีบัณฑิต, สาขาวิชาการวัดและประเมินผลการศึกษา, บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย.

ศิริชัย กาญจนวาสี, ทวีวัฒน์ ปิตยานนท์ และดิเรก ศรีสุโข. (2551). *การเลือกใช้สถิติที่เหมาะสมสำหรับการวิจัย* (พิมพ์ครั้งที่ 5). กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

ศิริชัย กาญจนวาสี. (2548). *ทฤษฎีการประเมิน*. กรุงเทพฯ: จุฬาลงกรณ์มหาวิทยาลัย.

ศิริชัย กาญจนวาสี. (2555). *ทฤษฎีการทดสอบแนวใหม่ (Modern test theories)* (พิมพ์ครั้งที่ 4). กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

สิริรัตน์ วิภาสศิลป์. (2545). *การเปรียบเทียบวิธีซิปเทสท์และดีเอฟไอทีในการตรวจสอบการทำหน้าที่เบี่ยงเบนของข้อสอบ หมวดข้อสอบ และแบบทดสอบ จากข้อมูลการตอบข้อสอบที่ใช้ความสามารถหลายมิติ*. ปริญญานิพนธ์การศึกษาดุษฎีบัณฑิต, สาขาวิชาการทดสอบและวัดผลการศึกษา, บัณฑิตวิทยาลัย, มหาวิทยาลัยศรีนครินทรวิโรฒ.

สุพัฒนา หอมบุปผา. (2556). *การเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN*. ดุษฎีนิพนธ์ปรัชญาดุษฎีบัณฑิต, สาขาวิชาวิจัย วัดผลและสถิติการศึกษา, คณะศึกษาศาสตร์, มหาวิทยาลัยบูรพา.

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

อรินทร์ น่วมถนอม. (2549). *การเปรียบเทียบวิธีโพลี-ซิปเทสท์ วิธีการถดถอยโลจิสติกแบบจัดอันดับ และวิธีการถดถอยดลจจิสติกแบบจัดอันดับหลายมิติ ในการตรวจสอบการทำหน้าที่เบี่ยงเบนของข้อสอบที่วัดความสามารถหลายมิติและให้คะแนนหลายค่า*. วิทยานิพนธ์การศึกษาดุษฎีบัณฑิต, สาขาวิชาการทดสอบและวัดผลการศึกษา, บัณฑิตวิทยาลัย, มหาวิทยาลัยศรีนครินทรวิโรฒ.

อริสพา เตหลิ่ม. (2559). *การเปรียบเทียบประสิทธิผลการประมาณค่าพารามิเตอร์และการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมกซิมัมไลค์ลิฮู้ด วิธีของเบส์และวิธีของเบส์แบบมีอิทธิพลเทสต์เลท*. ดุษฎีนิพนธ์ปรัชญาดุษฎีบัณฑิต, สาขาวิชาวิจัย วัดผลและสถิติการศึกษา, คณะศึกษาศาสตร์, มหาวิทยาลัยบูรพา.

อารี วัชรโสตถิกุล. (2543). *การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้รูปแบบและวิธีการแตกต่างกัน*. วิทยานิพนธ์การศึกษามหาบัณฑิต, สาขาวิชาการวัดผลการศึกษา, บัณฑิตวิทยาลัย, มหาวิทยาลัยศรีนครินทรวิโรฒ.

อาวิพร ปานทอง. (2558). *การเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบให้คะแนนหลายค่าโดยวิธีทดสอบอัตราส่วนความควรจะเป็น วิธีเบส์เซียนและวิธีโพลี-ซิปเทสท์*. วิทยานิพนธ์ปรัชญาดุษฎีบัณฑิต, สาขาวิชาวิจัย วัดผลและสถิติการศึกษา, คณะศึกษาศาสตร์, มหาวิทยาลัยบูรพา.

อิทธิฤทธิ์ พงษ์ปิยะรัตน์. (2551). *การวิเคราะห์ข้อสอบและการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ: การวิเคราะห์พหุระดับ*. วิทยานิพนธ์ครุศาสตรดุษฎีบัณฑิต, สาขาวิชาการวัดและประเมินผลการศึกษา, บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย.

อุทัยวรรณ สายพัฒนะ. (2547). *การเปรียบเทียบประสิทธิภาพของผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ในแบบสอบที่มีการให้คะแนนแบบหลายค่า ระหว่างวิธี GMH และวิธี Polytomous SIBTEST*. วิทยานิพนธ์การศึกษาดุษฎีบัณฑิต, สาขาวิชาการทดสอบและวัดผลการศึกษา, บัณฑิตวิทยาลัย, มหาวิทยาลัยศรีนครินทรวิโรฒ.

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of educational measurement*, *29*(1), 67-91.

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items are measuring. *Applied Measurement in Education, 7*(4), 255-278.

Ackerman, T. A., & Evans, J. A. (1992). *An investigation of the relationship between reliability, power, and the Type I error rate of the Mantel-Haenszel and simultaneous item bias detection procedures*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

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 Universitesity Journal of Educaiton*, *41*, 36-47.

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education*, *15*(2), 113-141.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO for windows [Computer software]*. Lincolnwiid, IL: Scientific Software International.

Camilli, G. & Shepard, L. A. (1994). *Methods of identifying biased test items*. Thousand Oaks: SAGE.

Chang, H. H., & Mazzeo, J. (1994). The unique correspondence of the item response function and item category response functions in polytomously scored item response models. *Psychometrika*, *59*(3), 391-404.

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

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., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, *42*(2), 133-148.

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.

De Ayala, R. J. (2013). *The theory and practice of item response theory*. New York London: The Guilford Press.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. *Journal of Educational Measurement, 23*(4), 355-368.

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, *16*(1), 5-18.

Elosua, P., & López-Jaúregui, A. (2007). Potential sources of differential item functioning in the adaptation of tests. *International Journal of Testing*, *7*(1), 39-52.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahawah, NJ: Lawrence Erlbaum.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29,* 278-295.

Finch, W. H., & French, B. F. (2007). Detection of crossing differential item functioning: A comparison of four methods. *Educational and Psychological Measurement, 67*(4), 565-582.

Flowers, C. P. (1997). The relationship between polytomous DFIT and other polytomous DIF procedures. In *Annual Meeting of the National Council on Measurement in Education (NCME)*.

French, A.W., & Miller, T.R. (1996). Logistics regresstion and its use in detection differential item functioning in polytomous items. *Journal of Education Measurement, 33*, 315-332.

Gierl, M. J., Bisanz, J., Bisanz, G. L., & Boughton, K. A. (2003). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the multidimensionality‐based DIF analysis paradigm. *Journal of Educational Measurement*, *40*(4), 281-306.

Gómez-Benito, J. Hidalgo, M. D. & Padilla, J. L. (2009). Efficacy of effect size measures in logistic regression an application for detecting DIF. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 15*(1), 18-25.

Han, K. T. (2007). WinGen: Windows software that generates item response theory parameters and item responses. *Applied Psychological Measurement*, *31*(5), 457-459.

Hambleton, R. K., & Jones, R. W. (2012). Comparison of classical test theory and item response theory and their applications to test development, Instructional Topics in Educational Measurement Series 16.

Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement*, *15*(3), 279-291.

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied psychological measurement*, *20*(2), 101-125.

Hoffman, L. (2014). *Measurement Invariance in CFA and Differential Item Functioning in IRT/ IFA*. Retrieved from http://www.lesahoffman.com/PSYC948/ 948_Lecture9_Invariance.pdf

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. *Test validity*, 129-145.

Hortensius, L. (2012). *Project for introduction to multivariate statistics: Measurement invariance*.

Kahraman, N., De Boeck, P., & Janssen, R. (2009). Modeling DIF in complex response data using test design strategies. *International Journal of Testing*, *9*(2), 151-166.

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.

Kim, M. (2010). *Impact of strategic sourcing, e-procurement and integration on supply chain risk mitigation and performance*. Buffalo, NY: State University of New York at Buffalo.

Kim, S. H. (2000). *An investigation of the likelihood ratio test, the Mantel test, and the generalized Mantel-Haenszel test of DIF*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Kim, S. H., Cohen, A. S., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Educational Measurement*, *44*(2), 93-116.

Kristjansson, E., Aylesworth, R., Mcdowell, I., & Zombo, B. D. (2005). A Comparison of you methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement, 65*, 935-953.

Lei, P. W., Chen, S. Y., & Yu, L. (2006). Comparing methods of assessing differential item functioning in a computerized adaptive testing environment. *Journal of Educational Measurement*, *43*(3), 245-264.

Li, H. H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika*, *61*(4), 647-677.

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.

Mellor, L. T. (1996). A comparison of four differential item functioning (DIF) methods for polytomously scored items.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement*, *52*(2), 443-451.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel procedure. *Educational and Psychological Measurement*, *54*(2), 284-291.

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.

Millsap, R. E., Gunn, H., Everson, H. T., & Zautra, A. (2014). Using item response theory to evaluate measurement invariance in Health-Related Measures. *Handbook of Item Response Theory Modeling*, 364-385.

Munro, B. H. (2005). *Statistical methods for health care research* (Vol. 1). Lippincott Williams & Wilkins.

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

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement*, *30*(4), 293-311.

Narayanan, P. & Sawaminathan, H. (1994). Perfromance of the Mantel-Haenzel and simulthaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18*(4)*,* 315-328.

Narayanan, P., & Swaminathan, H. (1996). Identification of Items that show Nonuniform DIF. *Applied Psychological Measurement*, *20*(3), 257-274.

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

Oishi, S. (2006). The concept of life satisfaction across cultures: An IRT analysis. *Journal of Research in Personality*, *40*(4), 411-423.

Oshima, T. C., Raju, N. S., & Flowers, C. P. (1997). Development and demonstration of multidimensional IRT-based internal measures of differential functioning of Items and tests. *Journal of Educational Measurement*, *34*(3), 253-272.

Oshima, T. C., Raju, N. S., Flowers, C. P., & Slinde, J. A. (1998). Differential bundle functioning using the DFIT framework: Procedures for identifying possible sources of differential functioning. *Applied Measurement in Education*, *11*(4), 353-369.

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Sage.

Park, T. (2006). Detecting DIF across different language and gender groups in the MELAB essay test using the logistic regression method. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *4*, 81-96.

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: A comparison of three Mantel-Haenszel procedures. *Applied Measurement in Education*, *14*(3), 235-259.

Penfield, R. D., & Algina, J. (2006). A generalized DIF effect variance estimator for measuring unsigned differential test functioning in mixed format tests. *Journal of Educational Measurement*, *43*(4), 295-312.

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.

Prieto, P., Barbero, M. I., & San Luis, C. (1997). Identification of nonuniform DIF: A comparison of Mantel-Haenszel and IRT analysis procedure. *Educational and Psychological Measurement, 57*(4), 559-568.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*(2), 197-207.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*(2), 105-116.

Roussos, L. & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*(4), 355-371.

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

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

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, *18*(4), 210-222.

Schwartz, S. H., Melech, G., Lehmann, A., Burgess, S., Harris, M., & Owens, V. (2001). Extending the cross-cultural validity of the theory of basic human values with a different method of measurement. *Journal of cross-cultural psychology*, *32*(5), 519-542.

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.

Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., & Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic acids research*, *34*(suppl_1), D535-D539.

Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorek, S., & Schwartz, S. H. (2009). Testing measurement invariance using multigroup CFA: Differences between educational groups in human values measurement. *Quality & Quantity*, *43*(4), 599-616.

Stout, W., Li, H. H., Nandakumar, R., & Bolt, D. (1997). MULTISIB: A procedure to investigate DIF when a test is intentionally two-dimensional. *Applied Psychological Measurement*, *21*(3), 195-213.

Stroud, A. H., & Secrest, D. (1996). *Gaussian quadrature formulas.* Englewood Cliffs, NJ: Prentice Hall.

Su, Y. H., & Wang, W. C. (2005). Efficiency of the Mantel, generalized Mantel-Haenszel, and logistic discriminant function analysis methods in detecting differential item functioning for polytomous items. *Applied Measurement in Education*, *18*(4), 313-350.

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

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.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In W. P. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Erlbaum.

Walker, C. M., & Beretvas, S. N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: A cognitive explanation for DIF. *Journal of Educational Measurement*, *38*(2), 147-163.

Wang, W. C., & Yeh, Y. L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, *27*(6), 479-498.

Walker, C. M., Zhang, B., & Surber, J. (2008). Using a multidimensional differential item functioning framework to determine if reading ability affects student performance in mathematics. *Applied Measurement in Education*, *21*(2), 162-181.

Wiberg, M. (2009). Differential item functioning in mastery tests: A comparison of three methods using real data. *International Journal of Testing*, *9*(1), 41-59.

Wood, S. W. (2011). Differential item functioning procedures for polytomous items when examinee sample sizes are small.

Yoon, M., & Millsap, R.E. (2007). Detecting violation of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling, 14*, 435-463

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item score*. Ottawa, Ontario, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

ภาคผนวก

ภาคผนวก ก

ตัวอย่างการจำลองรูปแบบผลการตอบข้อสอบ
ที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว
และข้อสอบทำหน้าที่ต่างกันที่ไม่เป็นรูปแบบเดียวกัน (Non-Uniform)

ตัวอย่างการจำลองรูปแบบการตอบข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่า
ในโมเดลการตอบสนองข้อสอบแบบมิติเดียว ด้วยโปรแกรม WinGen

1	GRM	5	1.551	-0.915	0.281	1.040	1.864
2	GRM	5	0.797	-0.039	0.670	0.752	1.649
3	GRM	5	0.496	-1.906	-1.105	-0.325	-0.113
4	GRM	5	1.174	-0.489	-0.021	0.307	0.825
5	GRM	5	1.403	-1.493	-1.230	0.062	0.638
6	GRM	5	1.061	-1.000	-0.264	0.107	0.444
7	GRM	5	0.823	-1.325	-1.002	0.224	0.333
8	GRM	5	1.508	-1.283	-0.527	0.719	0.847
9	GRM	5	1.094	-0.283	0.334	1.114	1.191

1	GRM	5	1.451	-0.815	0.281	1.14	1.864
2	GRM	5	0.797	-0.039	0.670	0.752	1.649
3	GRM	5	0.496	-1.906	-1.105	-0.325	-0.113
4	GRM	5	1.174	-0.489	-0.021	0.307	0.825
5	GRM	5	1.403	-1.493	-1.230	0.062	0.638
6	GRM	5	1.061	-1.000	-0.264	0.107	0.444
7	GRM	5	0.823	-1.325	-1.002	0.224	0.333
8	GRM	5	1.508	-1.283	-0.527	0.719	0.847
9	GRM	5	1.094	-0.283	0.334	1.114	1.191

ภาคผนวก ข

ตัวอย่างผลการวิเคราะห์ข้อมูลด้วยโปรแกรม IRTPRO

ตัวอย่างผลการวิเคราะห์การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี IRT LR

DIF Statistics for Graded Items

Item numbers in:										
Group 1	**Group 2**	**Total X^2**	**d.f.**	**p**	**X^2_a**	**d.f.**	**p**	**$X^2_{c\|a}$**	**d.f.**	**p**
1	1	1.2	5	0.9426	0.3	1	0.5691	0.9	4	0.9248
2	2	6.2	5	0.2885	2.1	1	0.1512	4.1	4	0.3894
3	3	12.5	5	0.0280	8.7	1	0.0032	3.8	4	0.4306
4	4	3.8	5	0.5813	0.1	1	0.8094	3.7	4	0.4449
5	5	8.1	5	0.1513	0.1	1	0.8051	8.0	4	0.0905
6	6	8.5	5	0.1286	0.7	1	0.4144	7.9	4	0.0962
7	7	7.6	5	0.1821	0.4	1	0.5240	7.1	4	0.1280
8	8	2.2	5	0.8231	0.1	1	0.7546	2.1	4	0.7199
9	9	3.4	5	0.6422	1.8	1	0.1772	1.6	4	0.8176

DIF Statistics for Graded Items

Item numbers in:										
Group 1	**Group 2**	**Total X^2**	**d.f.**	**p**	**X^2_a**	**d.f.**	**p**	**$X^2_{c\|a}$**	**d.f.**	**p**
1	1	11.3	5	0.0454	2.4	1	0.1241	8.9	4	0.0626
2	2	5.1	5	0.4039	1.6	1	0.2135	3.6	4	0.4704
3	3	6.8	5	0.2347	0.4	1	0.5477	6.5	4	0.1661
4	4	2.2	5	0.8198	0.7	1	0.3967	1.5	4	0.8288
5	5	7.2	5	0.2071	0.7	1	0.4062	6.5	4	0.1651
6	6	1.6	5	0.9032	1.0	1	0.3090	0.5	4	0.9688
7	7	3.9	5	0.5717	2.6	1	0.1072	1.2	4	0.8711
8	8	2.1	5	0.8359	0.3	1	0.5709	1.8	4	0.7774
9	9	3.6	5	0.6073	0.0	1	0.8570	3.6	4	0.4669

DIF Statistics for Graded Items

Item numbers in:										
Group 1	**Group 2**	**Total X^2**	**d.f.**	**p**	**X^2_a**	**d.f.**	**p**	**$X^2_{c\|a}$**	**d.f.**	**p**
1	1	2.4	5	0.7974	0.8	1	0.3862	1.6	4	0.8073
2	2	18.0	5	0.0029	3.8	1	0.0513	14.2	4	0.0066
3	3	1.2	5	0.9469	0.1	1	0.7407	1.1	4	0.8991
4	4	2.7	5	0.7472	0.1	1	0.8157	2.6	4	0.6201
5	5	4.7	5	0.4592	1.1	1	0.3029	3.6	4	0.4638
6	6	4.2	5	0.5162	0.1	1	0.7343	4.1	4	0.3903
7	7	6.3	5	0.2806	0.1	1	0.7056	6.1	4	0.1881
8	8	5.5	5	0.3545	0.2	1	0.6714	5.4	4	0.2534
9	9	6.5	5	0.2644	0.1	1	0.7808	6.4	4	0.1711

DIF Statistics for Graded Items

Item numbers in:										
Group 1	**Group 2**	**Total X^2**	**d.f.**	**p**	**X^2_a**	**d.f.**	**p**	**$X^2_{c\|a}$**	**d.f.**	**p**
1	1	4.0	5	0.5562	0.3	1	0.5554	3.6	4	0.4623
2	2	3.3	5	0.6585	0.5	1	0.4966	2.8	4	0.5907
3	3	3.3	5	0.6580	0.0	1	0.9778	3.3	4	0.5135
4	4	4.2	5	0.5278	0.6	1	0.4387	3.6	4	0.4702
5	5	1.8	5	0.8760	1.0	1	0.3288	0.8	4	0.9323
6	6	3.0	5	0.6954	0.3	1	0.5732	2.7	4	0.6072
7	7	2.6	5	0.7590	0.0	1	0.8670	2.6	4	0.6292
8	8	9.6	5	0.0877	2.1	1	0.1497	7.5	4	0.1114
9	9	5.9	5	0.3181	0.6	1	0.4284	5.3	4	0.2626
10	10	4.1	5	0.5374	1.5	1	0.2256	2.6	4	0.6245
11	11	2.4	5	0.7943	0.1	1	0.7532	2.3	4	0.6842
12	12	7.3	5	0.1958	0.0	1	0.8930	7.3	4	0.1194
13	13	0.8	5	0.9783	0.3	1	0.5622	0.4	4	0.9787
14	14	5.0	5	0.4180	0.0	1	0.8735	5.0	4	0.2920
15	15	2.5	5	0.7766	0.7	1	0.4172	1.8	4	0.7651

DIF Statistics for Graded Items

Item numbers in:										
Group 1	**Group 2**	**Total X^2**	**d.f.**	**p**	**X^2_a**	**d.f.**	**p**	**$X^2_{c\|a}$**	**d.f.**	**p**
1	1	4.0	5	0.5501	0.0	1	0.9176	4.0	4	0.4084
2	2	7.6	5	0.1793	0.6	1	0.4499	7.0	4	0.1341
3	3	2.2	5	0.8246	1.0	1	0.3237	1.2	4	0.8784
4	4	4.6	5	0.4687	0.6	1	0.4345	4.0	4	0.4098
5	5	1.9	5	0.8681	0.6	1	0.4305	1.2	4	0.8717
6	6	9.3	5	0.0964	0.6	1	0.4570	8.8	4	0.0668
7	7	2.6	5	0.7660	0.1	1	0.7094	2.4	4	0.6572
8	8	5.5	5	0.3566	0.0	1	0.9630	5.5	4	0.2392
9	9	8.4	5	0.1344	0.8	1	0.3707	7.6	4	0.1066
10	10	4.8	5	0.4411	0.3	1	0.5961	4.5	4	0.3409
11	11	1.2	5	0.9453	0.4	1	0.5145	0.8	4	0.9424
12	12	5.7	5	0.3380	0.0	1	0.9670	5.7	4	0.2244
13	13	0.8	5	0.9767	0.0	1	0.8808	0.8	4	0.9407
14	14	1.7	5	0.8903	0.1	1	0.7543	1.6	4	0.8105
15	15	3.8	5	0.5814	0.6	1	0.4270	3.2	4	0.5332

ภาคผนวก ค

ตัวอย่างผลการวิเคราะห์ข้อมูลด้วยโปรแกรม Poly-SIBTEST

ตัวอย่างผลการวิเคราะห์การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี Poly-SIBTEST

SIBTEST-pooled weighting

Run	Suspect Subtest	**Beta**	**standard**	**p-value**	**p-elim**		**MS/ SSD**	**FLAGG**
no.	**Item Numbers**	**estimate**	**errors**		**R**	**F**		
1	1	0.084	0.146	0.566 E	.14	.30	0.02	0
2	2	0.452	0.260	0.082 E	.19	.18	-0.02	0
3	3	-0.200	0.263	0.448 E	.27	.23	0.04	0
4	4	-0.238	0.241	0.323 E	.30	.24	0.03	0
5	5	-0.328	0.164	0.045 E	.23	.29	0.06	0
6	6	-0.241	0.239	0.314 E	.20	.26	0.08	0
7	7	-0.083	0.227	0.715 E	.18	.24	0.01	0
8	8	0.169	0.152	0.266 E	.19	.17	0.01	0
9	9	-0.009	0.193	0.962 E	.25	.29	0.03	0

SIBTEST-pooled weighting

Run	Suspect Subtest	**Beta**	**standard**	**p-value**	**p-elim**		**MS/ SSD**	**FLAGG**
no.	**Item Numbers**	**estimate**	**errors**		**R**	**F**		
1	1	0.296	0.140	0.034 E	.36	.35	-0.06	0
2	2	-0.245	0.266	0.358 E	.30	.44	-0.01	0
3	3	0.459	0.233	0.049 E	.24	.31	-0.08	0
4	4	-0.384	0.243	0.115 E	.27	.22	0.00	0
5	5	0.075	0.157	0.630 E	.22	.11	-0.04	0
6	6	0.095	0.219	0.663 E	.20	.18	-0.04	0
7	7	-0.103	0.224	0.645 E	.17	.17	-0.02	0
8	8	-0.017	0.187	0.929 E	.25	.21	-0.03	0
9	9	-0.209	0.223	0.349 E	.23	.21	0.00	0

SIBTEST-pooled weighting

Run no.	Suspect Subtest Item Numbers	Beta estimate	standard errors	p-value	p-elim R	p-elim F	MS/ SSD	FLAGG
1	1	0.115	0.086	0.178 E	.08	.07	0.01	0
2	2	0.009	0.076	0.910 E	.07	.10	0.01	0
3	3	0.091	0.112	0.419 E	.13	.04	0.01	0
4	4	0.001	0.093	0.990 E	.08	.05	0.02	0
5	5	-0.080	0.064	0.206 E	.08	.05	0.02	0
6	6	0.130	0.081	0.106 E	.06	.01	0.01	0
7	7	0.101	0.096	0.296 E	.08	.05	0.01	0
8	8	0.207	0.118	0.080 E	.07	.02	0.01	0
9	9	0.135	0.106	0.203 E	.04	.04	0.00	0
10	10	-0.217	0.109	0.046 E	.12	.05	0.03	0
11	11	-0.131	0.102	0.197 E	.05	.06	0.02	0
12	12	-0.110	0.103	0.288 E	.08	.05	0.02	0
13	13	-0.082	0.121	0.502 E	.04	.05	0.02	0
14	14	-0.112	0.111	0.311 E	.07	.05	0.02	0
15	15	0.048	0.110	0.660 E	.09	.06	0.01	0

SIBTEST-pooled weighting

Run no.	Suspect Subtest Item Numbers	Beta estimate	standard errors	p-value	p-elim R	p-elim F	MS/ SSD	FLAGG
1	1	-0.080	0.089	0.365 E	.07	.11	0.00	0
2	2	-0.127	0.083	0.124 E	.09	.06	0.00	0
3	3	-0.012	0.104	0.906 E	.03	.12	0.00	0
4	4	0.050	0.084	0.550 E	.10	.09	-0.01	0
5	5	-0.037	0.068	0.590 E	.04	.10	0.00	0
6	6	0.088	0.083	0.291 E	.04	.04	-0.01	0
7	7	0.113	0.108	0.295 E	.08	.08	-0.01	0
8	8	-0.060	0.118	0.614 E	.02	.08	0.00	0
9	9	-0.070	0.107	0.512 E	.04	.04	0.00	0
10	10	0.086	0.112	0.444 E	.14	.10	-0.01	0
11	11	-0.114	0.110	0.299 E	.11	.14	0.00	0
12	12	-0.147	0.085	0.083 E	.08	.11	0.00	0
13	13	0.007	0.123	0.955 E	.04	.12	-0.01	0
14	14	-0.027	0.116	0.815 E	.01	.08	0.00	0
15	15	0.035	0.114	0.756 E	.02	.07	0.00	0

ภาคผนวก ง

ตัวอย่างผลการวิเคราะห์ข้อมูลด้วยโปรแกรม Mplus

ตัวอย่างผลการวิเคราะห์การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ
ด้วยวิธี Multiple-groups CFA

```
Mplus VERSION 6.12
MUTHEN & MUTHEN
08/02/2017   7:28 PM

INPUT INSTRUCTIONS

 Title:
  Run condition9 dif item 1
 Data:
      file is F_MERGE_R_1.txt;
 Variable: names are group q1-q9;
          Usevariables are All;
          Categorical are q1-q9;
          grouping is group (1=FG 2=RG);
 Analysis:
         Estimator is wlsmv;
         parameterization=theta;

 Model: f1 by q1* (a1)
       q2-q9 (a2-a9);
       f1@1;
       [f1@0];
       q1-q9@1;
 Model RG: !Q2 Q4 Q1$1 Q1$3
       f1 by q1* (a1)
       q2 (a2d)
       q3 (a3)
```

```
    q4 (a4d)
    q5-q9 (a5-a9);
    [Q1$1*];
    [Q1$3*];
    f1*;
    [f1@0];
    q1-q9@1;
Output: modindices(4);
```

INPUT READING TERMINATED NORMALLY

Run condition9 dif item 1

SUMMARY OF ANALYSIS

Number of groups	2
Number of observations	
Group FG	500
Group RG	500
Number of dependent variables	9
Number of independent variables	0
Number of continuous latent variables	1

Observed dependent variables

Binary and ordered categorical (ordinal)

Q1 Q2 Q3 Q4 Q5 Q6
Q7 Q8 Q9

Continuous latent variables

F1

Variables with special functions

Grouping variable GROUP

Estimator	WLSMV
Maximum number of iterations	1000
Convergence criterion	0.500D-04
Maximum number of steepest descent iterations	20
Parameterization	THETA

Input data file(s)

F_MERGE_R_1.txt

Input data format FREE

UNIVARIATE PROPORTIONS AND COUNTS FOR CATEGORICAL VARIABLES

Group FG

Q1		
Category 1	0.390	195.000
Category 2	0.030	15.000
Category 3	0.268	134.000
Category 4	0.102	51.000
Category 5	0.210	105.000
Q2		
Category 1	0.288	144.000
Category 2	0.130	65.000

Category 3	0.036	18.000
Category 4	0.062	31.000
Category 5	0.484	242.000
Q3		
Category 1	0.110	55.000
Category 2	0.256	128.000
Category 3	0.282	141.000
Category 4	0.002	1.000
Category 5	0.350	175.000
Q4		
Category 1	0.436	218.000
Category 2	0.242	121.000
Category 3	0.062	31.000
Category 4	0.018	9.000
Category 5	0.242	121.000
Q5		
Category 1	0.158	79.000
Category 2	0.046	23.000
Category 3	0.408	204.000
Category 4	0.162	81.000
Category 5	0.226	113.000
Q6		
Category 1	0.226	113.000
Category 2	0.330	165.000
Category 3	0.028	14.000
Category 4	0.244	122.000
Category 5	0.172	86.000
Q7		
Category 1	0.466	233.000
Category 2	0.066	33.000

Category 3	0.254	127.000
Category 4	0.158	79.000
Category 5	0.056	28.000
Q8		
Category 1	0.460	230.000
Category 2	0.016	8.000
Category 3	0.138	69.000
Category 4	0.254	127.000
Category 5	0.132	66.000
Q9		
Category 1	0.188	94.000
Category 2	0.074	37.000
Category 3	0.354	177.000
Category 4	0.012	6.000
Category 5	0.372	186.000
Group RG		
Q1		
Category 1	0.210	105.000
Category 2	0.180	90.000
Category 3	0.182	91.000
Category 4	0.270	135.000
Category 5	0.158	79.000
Q2		
Category 1	0.264	132.000
Category 2	0.112	56.000
Category 3	0.032	16.000
Category 4	0.054	27.000
Category 5	0.538	269.000

Q3		
Category 1	0.106	53.000
Category 2	0.248	124.000
Category 3	0.314	157.000
Category 4	0.002	1.000
Category 5	0.330	165.000
Q4		
Category 1	0.444	222.000
Category 2	0.242	121.000
Category 3	0.054	27.000
Category 4	0.032	16.000
Category 5	0.228	114.000
Q5		
Category 1	0.170	85.000
Category 2	0.050	25.000
Category 3	0.410	205.000
Category 4	0.130	65.000
Category 5	0.240	120.000
Q6		
Category 1	0.210	105.000
Category 2	0.332	166.000
Category 3	0.024	12.000
Category 4	0.252	126.000
Category 5	0.182	91.000
Q7		
Category 1	0.470	235.000
Category 2	0.066	33.000
Category 3	0.304	152.000
Category 4	0.104	52.000
Category 5	0.056	28.000

Q8		
Category 1	0.470	235.000
Category 2	0.018	9.000
Category 3	0.170	85.000
Category 4	0.236	118.000
Category 5	0.106	53.000
Q9		
Category 1	0.186	93.000
Category 2	0.076	38.000
Category 3	0.332	166.000
Category 4	0.002	1.000
Category 5	0.404	202.000

THE MODEL ESTIMATION TERMINATED NORMALLY

MODEL FIT INFORMATION

Number of Free Parameters 50

Chi-Square Test of Model Fit

Value	91.776*
Degrees of Freedom	94
P-Value	0.5457

Chi-Square Contributions From Each Group

FG	44.805
RG	46.970

* The chi-square value for MLM, MLMV, MLR, ULSMV, WLSM and WLSMV cannot be used for chi-square difference testing in the regular way. MLM, MLR and WLSM chi-square difference testing is described on the Mplus website. MLMV, WLSMV, and ULSMV difference testing is done using the DIFFTEST option.

RMSEA (Root Mean Square Error Of Approximation)

Estimate	0.000
90 Percent C.I.	0.000 0.023
Probability RMSEA <= .05	1.000

CFI/TLI

CFI	1.000
TLI	1.000

Chi-Square Test of Model Fit for the Baseline Model

Value	10677.183
Degrees of Freedom	72
P-Value	0.0000

WRMR (Weighted Root Mean Square Residual)

Value	0.934

MODEL RESULTS

			Two-Tailed
Estimate	S.E.	Est./S.E.	P-Value

Group FG

F1 BY				
Q1	0.980	0.063	15.480	0.000
Q2	1.494	0.139	10.733	0.000
Q3	1.406	0.088	16.066	0.000
Q4	1.312	0.104	12.656	0.000
Q5	1.231	0.079	15.645	0.000
Q6	1.265	0.075	16.911	0.000
Q7	1.263	0.081	15.637	0.000
Q8	1.139	0.076	15.081	0.000
Q9	1.339	0.089	15.024	0.000
Means				
F1	0.000	0.000	999.000	999.000
Thresholds				
Q1$1	-0.391	0.079	-4.933	0.000
Q1$2	-0.339	0.057	-5.991	0.000
Q1$3	0.686	0.082	8.339	0.000
Q1$4	1.267	0.067	18.767	0.000
Q2$1	-0.973	0.073	-13.319	0.000
Q2$2	-0.433	0.067	-6.433	0.000
Q2$3	-0.291	0.066	-4.392	0.000
Q2$4	-0.055	0.065	-0.840	0.401
Q3$1	-2.161	0.105	-20.648	0.000
Q3$2	-0.626	0.072	-8.649	0.000
Q3$3	0.710	0.072	9.868	0.000
Q3$4	0.720	0.072	9.988	0.000
Q4$1	-0.236	0.062	-3.793	0.000

Q4$2	0.742	0.066	11.176	0.000
Q4$3	1.008	0.070	14.445	0.000
Q4$4	1.134	0.071	15.886	0.000
Q5$1	-1.570	0.084	-18.752	0.000
Q5$2	-1.283	0.077	-16.586	0.000
Q5$3	0.494	0.066	7.515	0.000
Q5$4	1.170	0.074	15.891	0.000
Q6$1	-1.270	0.075	-17.033	0.000
Q6$2	0.201	0.065	3.095	0.002
Q6$3	0.309	0.065	4.727	0.000
Q6$4	1.513	0.079	19.217	0.000
Q7$1	-0.131	0.064	-2.033	0.042
Q7$2	0.139	0.065	2.138	0.033
Q7$3	1.442	0.082	17.606	0.000
Q7$4	2.590	0.117	22.119	0.000
Q8$1	-0.135	0.061	-2.227	0.026
Q8$2	-0.069	0.061	-1.145	0.252
Q8$3	0.531	0.064	8.359	0.000
Q8$4	1.804	0.087	20.699	0.000
Q9$1	-1.504	0.087	-17.343	0.000
Q9$2	-1.078	0.078	-13.785	0.000
Q9$3	0.451	0.068	6.597	0.000
Q9$4	0.482	0.069	7.036	0.000

Variances

F1	1.000	0.000	999.000	999.000

Residual Variances

Q1	1.000	0.000	999.000	999.000
Q2	1.000	0.000	999.000	999.000

Q3	1.000	0.000	999.000	999.000
Q4	1.000	0.000	999.000	999.000
Q5	1.000	0.000	999.000	999.000
Q6	1.000	0.000	999.000	999.000
Q7	1.000	0.000	999.000	999.000
Q8	1.000	0.000	999.000	999.000
Q9	1.000	0.000	999.000	999.000

Group RG

F1 BY

Q1	0.980	0.063	15.480	0.000
Q2	1.083	0.101	10.748	0.000
Q3	1.406	0.088	16.066	0.000
Q4	1.073	0.094	11.385	0.000
Q5	1.231	0.079	15.645	0.000
Q6	1.265	0.075	16.911	0.000
Q7	1.263	0.081	15.637	0.000
Q8	1.139	0.076	15.081	0.000
Q9	1.339	0.089	15.024	0.000

Means

F1	0.000	0.000	999.000	999.000

Thresholds

Q1$1	-1.150	0.090	-12.754	0.000
Q1$2	-0.339	0.057	-5.991	0.000
Q1$3	0.259	0.080	3.219	0.001
Q1$4	1.267	0.067	18.767	0.000
Q2$1	-0.973	0.073	-13.319	0.000

Q2$2	-0.433	0.067	-6.433	0.000
Q2$3	-0.291	0.066	-4.392	0.000
Q2$4	-0.055	0.065	-0.840	0.401
Q3$1	-2.161	0.105	-20.648	0.000
Q3$2	-0.626	0.072	-8.649	0.000
Q3$3	0.710	0.072	9.868	0.000
Q3$4	0.720	0.072	9.988	0.000
Q4$1	-0.236	0.062	-3.793	0.000
Q4$2	0.742	0.066	11.176	0.000
Q4$3	1.008	0.070	14.445	0.000
Q4$4	1.134	0.071	15.886	0.000
Q5$1	-1.570	0.084	-18.752	0.000
Q5$2	-1.283	0.077	-16.586	0.000
Q5$3	0.494	0.066	7.515	0.000
Q5$4	1.170	0.074	15.891	0.000
Q6$1	-1.270	0.075	-17.033	0.000
Q6$2	0.201	0.065	3.095	0.002
Q6$3	0.309	0.065	4.727	0.000
Q6$4	1.513	0.079	19.217	0.000
Q7$1	-0.131	0.064	-2.033	0.042
Q7$2	0.139	0.065	2.138	0.033
Q7$3	1.442	0.082	17.606	0.000
Q7$4	2.590	0.117	22.119	0.000
Q8$1	-0.135	0.061	-2.227	0.026
Q8$2	-0.069	0.061	-1.145	0.252
Q8$3	0.531	0.064	8.359	0.000
Q8$4	1.804	0.087	20.699	0.000
Q9$1	-1.504	0.087	-17.343	0.000
Q9$2	-1.078	0.078	-13.785	0.000
Q9$3	0.451	0.068	6.597	0.000

Q9$4	0.482	0.069	7.036	0.000

Variances

F1	1.079	0.124	8.719	0.000

Residual Variances

Q1	1.000	0.000	999.000	999.000
Q2	1.000	0.000	999.000	999.000
Q3	1.000	0.000	999.000	999.000
Q4	1.000	0.000	999.000	999.000
Q5	1.000	0.000	999.000	999.000
Q6	1.000	0.000	999.000	999.000
Q7	1.000	0.000	999.000	999.000
Q8	1.000	0.000	999.000	999.000
Q9	1.000	0.000	999.000	999.000

QUALITY OF NUMERICAL RESULTS

Condition Number for the Information Matrix 0.311E-01
(ratio of smallest to largest eigenvalue)

MODEL MODIFICATION INDICES

NOTE: Modification indices for direct effects of observed dependent variables regressed on covariates and residual covariances among observed dependent variables may not be included. To include these, request MODINDICES (ALL).

Minimum M.I. value for printing the modification index 4.000

	M.I.	E.P.C.	Std E.P.C.	StdYX E.P.C.

Group FG

Variances/Residual Variances

Q1	7.621	0.498	0.498	0.254

Group RG

WITH Statements

Q6	WITH Q1	4.345	0.274	0.274	0.274

Variances/Residual Variances

Q1	7.626	-0.498	-0.498	-0.245

Beginning Time: 19:28:54
Ending Time: 19:28:55
Elapsed Time: 00:00:01

MUTHEN & MUTHEN
3463 Stoner Ave.
Los Angeles, CA 90066

Tel: (310) 391-9971
Fax: (310) 391-8971
Web: www.StatModel.com
Support: Support@StatModel.com

Mplus VERSION 6.12

MUTHEN & MUTHEN

08/02/2017 7:39 PM

INPUT INSTRUCTIONS

```
Title:
 Run condition dif item 1 and 8
Data:
     file is F_MERGE_R_6.txt;
Variable: names are group q1-q15;
         Usevariables are All;
         Categorical are q1-q15;
         grouping is group (1=FG 2=RG);
Analysis:
        Estimator is wlsmv;
        parameterization=theta;

Model: f1 by q1* (a1)
      q2-q15 (a2-a15);
      f1@1;
      [f1@0];
      q1-q15@1;
Model RG: !Q2
      f1 by q1* (a1)
      q2 (a2d)
      q3-q15 (a3-a15);
      f1*;
      [f1@0];
      q1-q15@1;
```

Output: modindices(4);

INPUT READING TERMINATED NORMALLY

Run condition dif item 1 and 8

SUMMARY OF ANALYSIS

Number of groups	2
Number of observations	
Group FG	100
Group RG	100
Number of dependent variables	15
Number of independent variables	0
Number of continuous latent variables	1

Observed dependent variables

Binary and ordered categorical (ordinal)

Q1 Q2 Q3 Q4 Q5 Q6

Q7 Q8 Q9 Q10 Q11 Q12

Q13 Q14 Q15

Continuous latent variables

F1

Variables with special functions

Grouping variable GROUP

Estimator	WLSMV
Maximum number of iterations	1000
Convergence criterion	0.500D-04
Maximum number of steepest descent iterations	20
Parameterization	THETA

Input data file(s)

F_MERGE_R_6.txt

Input data format FREE

UNIVARIATE PROPORTIONS AND COUNTS FOR CATEGORICAL VARIABLES

Group FG

Q1		
Category 1	0.250	25.000
Category 2	0.050	5.000
Category 3	0.090	9.000
Category 4	0.170	17.000
Category 5	0.440	44.000
Q2		
Category 1	0.200	20.000
Category 2	0.280	28.000
Category 3	0.030	3.000
Category 4	0.330	33.000
Category 5	0.160	16.000
Q3		
Category 1	0.450	45.000
Category 2	0.130	13.000
Category 3	0.170	17.000

Category 4	0.040	4.000
Category 5	0.210	21.000
Q4		
Category 1	0.250	25.000
Category 2	0.520	52.000
Category 3	0.130	13.000
Category 4	0.010	1.000
Category 5	0.090	9.000
Q5		
Category 1	0.080	8.000
Category 2	0.240	24.000
Category 3	0.010	1.000
Category 4	0.210	21.000
Category 5	0.460	46.000
Q6		
Category 1	0.180	18.000
Category 2	0.210	21.000
Category 3	0.150	15.000
Category 4	0.370	37.000
Category 5	0.090	9.000
Q7		
Category 1	0.200	20.000
Category 2	0.270	27.000
Category 3	0.230	23.000
Category 4	0.100	10.000
Category 5	0.200	20.000
Q8		
Category 1	0.170	17.000
Category 2	0.340	34.000
Category 3	0.110	11.000

Category 4	0.100	10.000
Category 5	0.280	28.000
Q9		
Category 1	0.190	19.000
Category 2	0.010	1.000
Category 3	0.030	3.000
Category 4	0.010	1.000
Category 5	0.760	76.000
Q10		
Category 1	0.330	33.000
Category 2	0.020	2.000
Category 3	0.270	27.000
Category 4	0.040	4.000
Category 5	0.340	34.000
Q11		
Category 1	0.390	39.000
Category 2	0.070	7.000
Category 3	0.140	14.000
Category 4	0.050	5.000
Category 5	0.350	35.000
Q12		
Category 1	0.410	41.000
Category 2	0.020	2.000
Category 3	0.120	12.000
Category 4	0.160	16.000
Category 5	0.290	29.000
Q13		
Category 1	0.410	41.000
Category 2	0.160	16.000
Category 3	0.160	16.000

Category 4	0.210	21.000
Category 5	0.060	6.000
Q14		
Category 1	0.320	32.000
Category 2	0.100	10.000
Category 3	0.140	14.000
Category 4	0.050	5.000
Category 5	0.390	39.000
Q15		
Category 1	0.500	50.000
Category 2	0.030	3.000
Category 3	0.030	3.000
Category 4	0.030	3.000
Category 5	0.410	41.000
Group RG		
Q1		
Category 1	0.270	27.000
Category 2	0.100	10.000
Category 3	0.010	1.000
Category 4	0.150	15.000
Category 5	0.470	47.000
Q2		
Category 1	0.310	31.000
Category 2	0.230	23.000
Category 3	0.060	6.000
Category 4	0.210	21.000
Category 5	0.190	19.000
Q3		
Category 1	0.430	43.000

Category 2	0.090	9.000
Category 3	0.200	20.000
Category 4	0.070	7.000
Category 5	0.210	21.000
Q4		
Category 1	0.200	20.000
Category 2	0.480	48.000
Category 3	0.220	22.000
Category 4	0.010	1.000
Category 5	0.090	9.000
Q5		
Category 1	0.080	8.000
Category 2	0.200	20.000
Category 3	0.010	1.000
Category 4	0.210	21.000
Category 5	0.500	50.000
Q6		
Category 1	0.280	28.000
Category 2	0.180	18.000
Category 3	0.080	8.000
Category 4	0.360	36.000
Category 5	0.100	10.000
Q7		
Category 1	0.210	21.000
Category 2	0.280	28.000
Category 3	0.210	21.000
Category 4	0.040	4.000
Category 5	0.260	26.000
Q8		
Category 1	0.170	17.000

Category 2	0.280	28.000
Category 3	0.070	7.000
Category 4	0.110	11.000
Category 5	0.370	37.000
Q9		
Category 1	0.190	19.000
Category 2	0.010	1.000
Category 3	0.120	12.000
Category 4	0.010	1.000
Category 5	0.670	67.000
Q10		
Category 1	0.270	27.000
Category 2	0.030	3.000
Category 3	0.370	37.000
Category 4	0.050	5.000
Category 5	0.280	28.000
Q11		
Category 1	0.410	41.000
Category 2	0.040	4.000
Category 3	0.210	21.000
Category 4	0.040	4.000
Category 5	0.300	30.000
Q12		
Category 1	0.410	41.000
Category 2	0.010	1.000
Category 3	0.140	14.000
Category 4	0.130	13.000
Category 5	0.310	31.000
Q13		
Category 1	0.390	39.000

Category 2	0.170	17.000
Category 3	0.210	21.000
Category 4	0.200	20.000
Category 5	0.030	3.000
Q14		
Category 1	0.340	34.000
Category 2	0.040	4.000
Category 3	0.210	21.000
Category 4	0.030	3.000
Category 5	0.380	38.000
Q15		
Category 1	0.540	54.000
Category 2	0.030	3.000
Category 3	0.010	1.000
Category 4	0.060	6.000
Category 5	0.360	36.000

THE MODEL ESTIMATION TERMINATED NORMALLY

MODEL FIT INFORMATION

Number of Free Parameters 77

Chi-Square Test of Model Fit

Value	242.049*
Degrees of Freedom	253
P-Value	0.6788

Chi-Square Contributions From Each Group

FG	120.346
RG	121.703

* The chi-square value for MLM, MLMV, MLR, ULSMV, WLSM and WLSMV cannot be used for chi-square difference testing in the regular way. MLM, MLR and WLSM chi-square difference testing is described on the Mplus website. MLMV, WLSMV, and ULSMV difference testing is done using the DIFFTEST option.

RMSEA (Root Mean Square Error Of Approximation)

Estimate	0.000
90 Percent C.I.	0.000 0.034
Probability RMSEA <= .05	0.998

CFI/TLI

CFI	1.000
TLI	1.004

Chi-Square Test of Model Fit for the Baseline Model

Value	2788.882
Degrees of Freedom	210
P-Value	0.0000

WRMR (Weighted Root Mean Square Residual)

Value	1.017

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Group FG				
F1 BY				
Q1	0.648	0.116	5.581	0.000
Q2	0.796	0.135	5.910	0.000
Q3	1.098	0.154	7.122	0.000
Q4	1.425	0.186	7.664	0.000
Q5	1.932	0.249	7.755	0.000
Q6	0.826	0.110	7.535	0.000
Q7	0.934	0.126	7.411	0.000
Q8	0.905	0.129	6.991	0.000
Q9	1.082	0.191	5.665	0.000
Q10	0.915	0.126	7.258	0.000
Q11	1.362	0.174	7.806	0.000
Q12	0.762	0.120	6.346	0.000
Q13	1.399	0.184	7.598	0.000
Q14	1.010	0.148	6.811	0.000
Q15	0.632	0.114	5.559	0.000
Means				
F1	0.000	0.000	999.000	999.000
Thresholds				
Q1$1	-0.761	0.114	-6.644	0.000
Q1$2	-0.502	0.109	-4.593	0.000

Q1$3	-0.346	0.107	-3.230	0.001
Q1$4	0.133	0.105	1.272	0.203
Q2$1	-0.929	0.134	-6.949	0.000
Q2$2	0.026	0.123	0.208	0.835
Q2$3	0.176	0.123	1.433	0.152
Q2$4	1.298	0.141	9.227	0.000
Q3$1	-0.222	0.130	-1.710	0.087
Q3$2	0.182	0.131	1.393	0.163
Q3$3	0.919	0.142	6.459	0.000
Q3$4	1.182	0.149	7.956	0.000
Q4$1	-1.294	0.173	-7.499	0.000
Q4$2	1.014	0.171	5.916	0.000
Q4$3	2.194	0.224	9.799	0.000
Q4$4	2.295	0.233	9.853	0.000
Q5$1	-2.996	0.317	-9.442	0.000
Q5$2	-1.121	0.211	-5.300	0.000
Q5$3	-1.060	0.210	-5.047	0.000
Q5$4	0.105	0.188	0.557	0.577
Q6$1	-0.939	0.125	-7.509	0.000
Q6$2	-0.241	0.115	-2.107	0.035
Q6$3	0.129	0.114	1.131	0.258
Q6$4	1.682	0.160	10.506	0.000
Q7$1	-1.114	0.137	-8.153	0.000
Q7$2	-0.067	0.120	-0.562	0.574
Q7$3	0.709	0.126	5.641	0.000
Q7$4	0.995	0.133	7.465	0.000
Q8$1	-1.273	0.145	-8.800	0.000
Q8$2	-0.068	0.118	-0.573	0.566
Q8$3	0.232	0.119	1.950	0.051
Q8$4	0.601	0.123	4.884	0.000

Q9$1	-1.276	0.166	-7.700	0.000
Q9$2	-1.223	0.164	-7.461	0.000
Q9$3	-0.862	0.151	-5.692	0.000
Q9$4	-0.819	0.150	-5.479	0.000
Q10$1	-0.703	0.125	-5.645	0.000
Q10$2	-0.609	0.123	-4.942	0.000
Q10$3	0.499	0.121	4.118	0.000
Q10$4	0.665	0.123	5.409	0.000
Q11$1	-0.420	0.148	-2.832	0.005
Q11$2	-0.188	0.148	-1.274	0.203
Q11$3	0.553	0.150	3.693	0.000
Q11$4	0.756	0.151	5.009	0.000
Q12$1	-0.284	0.111	-2.549	0.011
Q12$2	-0.236	0.111	-2.124	0.034
Q12$3	0.173	0.111	1.557	0.120
Q12$4	0.653	0.116	5.622	0.000
Q13$1	-0.429	0.152	-2.831	0.005
Q13$2	0.276	0.151	1.828	0.068
Q13$3	1.142	0.166	6.871	0.000
Q13$4	2.858	0.279	10.249	0.000
Q14$1	-0.617	0.129	-4.797	0.000
Q14$2	-0.356	0.126	-2.840	0.005
Q14$3	0.266	0.125	2.122	0.034
Q14$4	0.411	0.127	3.238	0.001
Q15$1	0.059	0.104	0.569	0.570
Q15$2	0.148	0.105	1.415	0.157
Q15$3	0.207	0.105	1.978	0.048
Q15$4	0.344	0.106	3.242	0.001

Variances

F1	1.000	0.000	999.000	999.000

Residual Variances

Q1	1.000	0.000	999.000	999.000
Q2	1.000	0.000	999.000	999.000
Q3	1.000	0.000	999.000	999.000
Q4	1.000	0.000	999.000	999.000
Q5	1.000	0.000	999.000	999.000
Q6	1.000	0.000	999.000	999.000
Q7	1.000	0.000	999.000	999.000
Q8	1.000	0.000	999.000	999.000
Q9	1.000	0.000	999.000	999.000
Q10	1.000	0.000	999.000	999.000
Q11	1.000	0.000	999.000	999.000
Q12	1.000	0.000	999.000	999.000
Q13	1.000	0.000	999.000	999.000
Q14	1.000	0.000	999.000	999.000
Q15	1.000	0.000	999.000	999.000

Group RG

F1 BY

Q1	0.648	0.116	5.581	0.000
Q2	1.202	0.180	6.679	0.000
Q3	1.098	0.154	7.122	0.000
Q4	1.425	0.186	7.664	0.000
Q5	1.932	0.249	7.755	0.000
Q6	0.826	0.110	7.535	0.000
Q7	0.934	0.126	7.411	0.000

Q8	0.905	0.129	6.991	0.000
Q9	1.082	0.191	5.665	0.000
Q10	0.915	0.126	7.258	0.000
Q11	1.362	0.174	7.806	0.000
Q12	0.762	0.120	6.346	0.000
Q13	1.399	0.184	7.598	0.000
Q14	1.010	0.148	6.811	0.000
Q15	0.632	0.114	5.559	0.000

Means

F1	0.000	0.000	999.000	999.000

Thresholds

Q1$1	-0.761	0.114	-6.644	0.000
Q1$2	-0.502	0.109	-4.593	0.000
Q1$3	-0.346	0.107	-3.230	0.001
Q1$4	0.133	0.105	1.272	0.203
Q2$1	-0.929	0.134	-6.949	0.000
Q2$2	0.026	0.123	0.208	0.835
Q2$3	0.176	0.123	1.433	0.152
Q2$4	1.298	0.141	9.227	0.000
Q3$1	-0.222	0.130	-1.710	0.087
Q3$2	0.182	0.131	1.393	0.163
Q3$3	0.919	0.142	6.459	0.000
Q3$4	1.182	0.149	7.956	0.000
Q4$1	-1.294	0.173	-7.499	0.000
Q4$2	1.014	0.171	5.916	0.000
Q4$3	2.194	0.224	9.799	0.000
Q4$4	2.295	0.233	9.853	0.000
Q5$1	-2.996	0.317	-9.442	0.000

Q5$2	-1.121	0.211	-5.300	0.000
Q5$3	-1.060	0.210	-5.047	0.000
Q5$4	0.105	0.188	0.557	0.577
Q6$1	-0.939	0.125	-7.509	0.000
Q6$2	-0.241	0.115	-2.107	0.035
Q6$3	0.129	0.114	1.131	0.258
Q6$4	1.682	0.160	10.506	0.000
Q7$1	-1.114	0.137	-8.153	0.000
Q7$2	-0.067	0.120	-0.562	0.574
Q7$3	0.709	0.126	5.641	0.000
Q7$4	0.995	0.133	7.465	0.000
Q8$1	-1.273	0.145	-8.800	0.000
Q8$2	-0.068	0.118	-0.573	0.566
Q8$3	0.232	0.119	1.950	0.051
Q8$4	0.601	0.123	4.884	0.000
Q9$1	-1.276	0.166	-7.700	0.000
Q9$2	-1.223	0.164	-7.461	0.000
Q9$3	-0.862	0.151	-5.692	0.000
Q9$4	-0.819	0.150	-5.479	0.000
Q10$1	-0.703	0.125	-5.645	0.000
Q10$2	-0.609	0.123	-4.942	0.000
Q10$3	0.499	0.121	4.118	0.000
Q10$4	0.665	0.123	5.409	0.000
Q11$1	-0.420	0.148	-2.832	0.005
Q11$2	-0.188	0.148	-1.274	0.203
Q11$3	0.553	0.150	3.693	0.000
Q11$4	0.756	0.151	5.009	0.000
Q12$1	-0.284	0.111	-2.549	0.011
Q12$2	-0.236	0.111	-2.124	0.034
Q12$3	0.173	0.111	1.557	0.120

Q12$4	0.653	0.116	5.622	0.000
Q13$1	-0.429	0.152	-2.831	0.005
Q13$2	0.276	0.151	1.828	0.068
Q13$3	1.142	0.166	6.871	0.000
Q13$4	2.858	0.279	10.249	0.000
Q14$1	-0.617	0.129	-4.797	0.000
Q14$2	-0.356	0.126	-2.840	0.005
Q14$3	0.266	0.125	2.122	0.034
Q14$4	0.411	0.127	3.238	0.001
Q15$1	0.059	0.104	0.569	0.570
Q15$2	0.148	0.105	1.415	0.157
Q15$3	0.207	0.105	1.978	0.048
Q15$4	0.344	0.106	3.242	0.001

Variances

F1	0.906	0.220	4.124	0.000

Residual Variances

Q1	1.000	0.000	999.000	999.000
Q2	1.000	0.000	999.000	999.000
Q3	1.000	0.000	999.000	999.000
Q4	1.000	0.000	999.000	999.000
Q5	1.000	0.000	999.000	999.000
Q6	1.000	0.000	999.000	999.000
Q7	1.000	0.000	999.000	999.000
Q8	1.000	0.000	999.000	999.000
Q9	1.000	0.000	999.000	999.000
Q10	1.000	0.000	999.000	999.000
Q11	1.000	0.000	999.000	999.000
Q12	1.000	0.000	999.000	999.000

Q13	1.000	0.000	999.000	999.000
Q14	1.000	0.000	999.000	999.000
Q15	1.000	0.000	999.000	999.000

QUALITY OF NUMERICAL RESULTS

Condition Number for the Information Matrix 0.697E-02
(ratio of smallest to largest eigenvalue)

MODEL MODIFICATION INDICES

NOTE: Modification indices for direct effects of observed dependent variables regressed on covariates and residual covariances among observed dependent variables may not be included. To include these, request MODINDICES (ALL).

Minimum M.I. value for printing the modification index 4.000

		M.I.	E.P.C.	Std E.P.C.	StdYX E.P.C.
Group FG					
No modification indices above the minimum value.					
Group RG					
WITH Statements					
Q14	WITH Q4	5.163	0.671	0.671	0.671

Beginning Time: 19:39:33

Ending Time: 19:39:33

Elapsed Time: 00:00:00

MUTHEN & MUTHEN

3463 Stoner Ave.

Los Angeles, CA 90066

Tel: (310) 391-9971

Fax: (310) 391-8971

Web: www.StatModel.com

Support: Support@StatModel.com

ภาคผนวก จ

ผลการวิเคราะห์ข้อมูลด้วยวิธีการวิเคราะห์ความแปรปรวนแบบวัดซ้ำ

(Repeated Measurement)

ผลการวิเคราะห์เปรียบเทียบค่าเฉลี่ยอัตราความคลาดเคลื่อนประเภทที่ 1 ภายใต้เงื่อนไขปัจจัยที่แตกต่าง 3 ปัจจัยหลัก ด้วยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสามวิธี

Mauchly's Test of Sphericity[a]

Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[b]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
method	.731	372.600	2	.000	.788	.796	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. Design: Intercept + DIFsize + LTESTsize + SAMPsize + DIFsize * LTESTsize + DIFsize * SAMPsize + LTESTsize * SAMPsize + DIFsize * LTESTsize * SAMPsize

Within Subjects Design: method

b. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

ภาคผนวก ฉ

แบบรายงานผลการพิจารณาจริยธรรมการวิจัย คณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา

แบบรายงานผลการพิจารณาจริยธรรมการวิจัย
คณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา

---

๑. **ชื่อวิทยานิพนธ์**

ชื่อเรื่องวิทยานิพนธ์ (ภาษาไทย) ประสิทธิภาพการตรวจสอบความไม่แปรเปลี่ยนของค่าพารามิเตอร์การทำหน้าที่ต่างกันของข้อสอบที่มีรูปแบบการตรวจให้คะแนนแบบหลายค่าในโมเดลการตอบสนองข้อสอบแบบมิติเดียว: วิธีการถดถอยโลจิสติก วิธีโพลีโตมัสซิบเทสท์ และวิธีการวิเคราะห์ องค์ประกอบเชิงยืนยันกลุ่มพหุ

ชื่อเรื่องวิทยานิพนธ์ (ภาษาอังกฤษ) EFFICIENCY OF A DETECTION INVARIANCE MEASUREMENT PARAMETERS OF DIF FOR POLYTOMOUSLY SCORED ITEMS IN UNIDIMENSIONAL IRT MODEL: LOGISTIC TEGRESSION, POLYTOMOUSE-SIBTEST AND MULTIPLE-GROUP CFA

๒. **ชื่อนิสิต** นางวาสนา กลมอ่อน **หลักสูตร** ปรัชญาดุษฎีบัณฑิต
รหัสประจำตัว ๕๒๘๑๐๑๕๑ **สาขาวิชา** วิจัย วัดผลและสถิติการศึกษา **คณะศึกษาศาสตร์**
☑ ภาคปกติ ☐ ภาคพิเศษ

๓. **ผลการพิจารณาของคณะกรรมการจริยธรรมการวิจัย:**

คณะกรรมการจริยธรรมการวิจัย ได้พิจารณารายละเอียดวิทยานิพนธ์ เรื่องดังกล่าวข้างต้นแล้ว ในประเด็นที่เกี่ยวกับ

๑) การเคารพในศักดิ์ศรี และสิทธิของมนุษย์ที่ใช้เป็นตัวอย่างการวิจัย

๒) วิธีการอย่างเหมาะสมในการได้รับความยินยอมจากกลุ่มตัวอย่างก่อนเข้าร่วมโครงการวิจัย (Informed consent) รวมทั้งการปกป้องสิทธิประโยชน์และรักษาความลับของกลุ่มตัวอย่างในการวิจัย

๓) การดำเนินการวิจัยอย่างเหมาะสม เพื่อไม่ก่อความเสียหายต่อสิ่งที่ศึกษาวิจัยไม่ว่าจะเป็นสิ่งที่มีชีวิตหรือไม่มีชีวิต

คณะกรรมการจริยธรรมการวิจัย มีมติเห็นชอบ ดังนี้

(✓) อนุมัติโครงการวิจัย

( ) ไม่อนุมัติ

๔. **วันที่ให้การอนุมัติ:**.....๒๔.....**เดือน** มีนาคม พ.ศ. ๒๕๖๐

(รองศาสตราจารย์ ดร.วิชิต สุรัตน์เรืองชัย)
คณบดีคณะศึกษาศาสตร์
ประธานคณะกรรมการพิจารณาจริยธรรมการวิจัย

ประวัติย่อของผู้วิจัย

ชื่อ-สกุล	นางวาสนา กลมอ่อน
วัน เดือน ปีเกิด	13 ตุลาคม พ.ศ. 2512
สถานที่เกิด	อำเภอศรีราชา จังหวัดชลบุรี
สถานที่อยู่ปัจจุบัน	35/ 4 ถนนบางแสนล่าง ตำบลแสนสุข อำเภอเมืองชลบุรี จังหวัดชลบุรี 20130
ตำแหน่งและประวัติการทำงาน	
พ.ศ. 2536-2537	ครู โรงเรียนเซนต์ปอลคอนแวนต์ อำเภอศรีราชา จังหวัดชลบุรี
พ.ศ. 2537-2550	นักวิเคราะห์นโยบายและแผน คณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา
พ.ศ. 2551-ปัจจุบัน	นักวิเคราะห์นโยบายและแผนชำนาญการ คณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา
ประวัติการศึกษา	
พ.ศ. 2534	การศึกษาบัณฑิต (วิทยาศาสตร์-คณิตศาสตร์) มหาวิทยาลัยบูรพา
พ.ศ. 2551	วิทยาศาสตรมหาบัณฑิต (เทคโนโลยีวิจัยการศึกษา) มหาวิทยาลัยบูรพา
พ.ศ. 2560	ปรัชญาดุษฎีบัณฑิต (วิจัย วัดผลและสถิติการศึกษา) มหาวิทยาลัยบูรพา