# การเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนีพาสเกลจีสแควร์และดัชนีเอสไคสแควร์ทั่วไป

นางสาวธีรนุช จาบประไพ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาครุศาสตรมหาบัณฑิต
สาขาวิชาการวัดและประเมินผลการศึกษา ภาควิชาวิจัยและจิตวิทยาการศึกษา
คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2554

# THE COMPARISON OF ITEM FIT INDEX EFFICIENCY BETWEEN PARSCALE $G^2$ AND GENERALIZED $S-\chi^2$

Miss Teeranuch Jabprapai

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Education Program in Educational Measurement and Evaluation

Department of Educational Research and Psychology

Faculty of Education

Chulalongkorn University

Academic Year 2011

| | |
|---|---|
| หัวข้อวิทยานิพนธ์ | การเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนีพาสเกลจีสแควร์และดัชนีเอสไคสแควร์ทั่วไป |
| โดย | นางสาวธีรนุช จาบประไพ |
| สาขาวิชา | การวัดและประเมินผลการศึกษา |
| อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก | ผู้ช่วยศาสตราจารย์ ดร.ณัฏฐภรณ์ หลาวทอง |

---

คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาครุศาสตรมหาบัณฑิต

…….. คณบดีคณะครุศาสตร์
(ศาสตราจารย์ ดร.ศิริชัย กาญจนวาสี)

คณะกรรมการสอบวิทยานิพนธ์

……..…… ประธานกรรมการ
(ศาสตราจารย์ ดร.ศิริชัย กาญจนวาสี)

…….... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผู้ช่วยศาสตราจารย์ ดร.ณัฏฐภรณ์ หลาวทอง)

…… กรรมการภายนอกมหาวิทยาลัย
(รองศาสตราจารย์ ดร.สุภมาส อังศุโชติ)

ธีรนุช จาบประไพ : การเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนีพาสเกลจีสแควร์และดัชนีเอสไคสแควร์ทั่วไป. (THE COMPARISON OF ITEM FIT INDEX EFFICIENCY BETWEEN PARSCALE $G^2$ AND GENERALIZED $S-\chi^2$) อ. ที่ปรึกษาวิทยานิพนธ์หลัก : ผศ.ดร.ณัฏฐภรณ์ หลาวทอง, 175 หน้า.

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถาม (item fit index) สองชนิดคือ ดัชนีพาสเกลจีสแควร์และดัชนีเอสไคสแควร์ทั่วไป ข้อมูลที่ใช้ในการศึกษาจำลองภายใต้โมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค 2 โมเดล คือ Grade Response Model (GRM) และ Generalized Partial Credit Model (GPCM) จัดกระทำข้อมูลตาม 3 เงื่อนไข คือ 1) ความยาวแบบวัด 3 ระดับ คือ 10, 20, และ 40 ข้อ 2) ขนาดกลุ่มตัวอย่าง 3 ระดับ คือ 500, 1000, และ 2000 คน 3) จำนวนรายการคำตอบ 4 ระดับ คือ 3, 5, 7, และ 9 รายการ รวมข้อมูลที่ศึกษาทั้งหมด 72 สถานการณ์ เกณฑ์ที่นำมาใช้ในการประเมินประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถาม คือค่าความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบ โดยใช้การเปรียบเทียบ 2 กรณี คือ 1.การเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดตามเงื่อนไขของ Kang และ Chen (2008) และ 2.การเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดโดยใช้สถิติการวิเคราะห์ความแปรปรวนแบบ 2 ทาง

ผลการวิจัยพบว่า

1. ในการเปรียบเทียบประสิทธิภาพตามเงื่อนไขของ Kang และ Chen (2008) ดัชนีเอสไคสแควร์ทั่วไปมีประสิทธิภาพในการบ่งชี้ความสอดคล้องของข้อคำถามมากกว่าดัชนีพาสเกลจีสแควร์ในเกือบทุกสถานการณ์ที่ทำการศึกษาเนื่องจากดัชนีเอสไคสแควร์ทั่วไปมีค่าความคลาดเคลื่อนประเภทที่ 1 ในการบ่งชี้ความสอดคล้องของข้อคำถามน้อยกว่าดัชนีพาสเกลจีสแควร์ ใน 70 สถานการณ์ จากทั้งหมด 72 สถานการณ์ (ค่าเฉลี่ยของค่าความคลาดเคลื่อนประเภทที่ 1 ใน 72 สถานการณ์ของดัชนีพาสเกลจีสแควร์ = 0.1535, ค่าเฉลี่ยของค่าความคลาดเคลื่อนประเภทที่ 1 ใน 72 สถานการณ์ของดัชนีเอสไคสแควร์ทั่วไป = 0.0216)

2. ในการเปรียบเทียบประสิทธิภาพโดยใช้สถิติการวิเคราะห์ความแปรปรวนแบบ 2 ทาง ดัชนีเอสไคสแควร์ทั่วไปให้ค่าความคลาดเคลื่อนประเภทที่ 1 น้อยกว่าดัชนีพาสเกลจีสแควร์ใน 5 กรณีที่ทำการศึกษา และดัชนีพาสเกลจีสแควร์ให้อำนาจการทดสอบที่สูงกว่าดัชนีเอสไคสแควร์ทั่วไปทั้ง 6 กรณีที่ทำการศึกษา

3. ดัชนีเอสไคสแควร์ทั่วไปมีโอกาสในการบ่งชี้ข้อคำถามที่สอดคล้องกับโมเดลว่าเป็นข้อคำถามที่ไม่สอดคล้องกับโมเดลน้อยกว่าดัชนีพาสเกลจีสแควร์ (ดัชนีเอสไคสแควร์ทั่วไปมีค่าความคลาดเคลื่อนประเภทที่ 1 น้อยกว่าดัชนีพาสเกลจีสแควร์) ในขณะที่ดัชนีพาสเกลจีสแควร์มีโอกาสในการบ่งชี้ข้อคำถามที่ไม่สอดคล้องกับโมเดลว่าเป็นข้อคำถามที่ไม่สอดคล้องกับโมเดลมากกว่าดัชนีเอสไคสแควร์ทั่วไป (ดัชนีพาสเกลจีสแควร์มีอำนาจการทดสอบมากกว่าดัชนีเอสไคสแควร์ทั่วไป)

ภาควิชา วิจัยและจิตวิทยาการศึกษา ลายมือชื่อนิสิต .....

สาขาวิชา การวัดและประเมินผลการศึกษา ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก .....

ปีการศึกษา 2554

## 5283360127 : MAJOR EDUCATIONAL MEASUREMENT AND EVALUATION
KEYWORDS : ITEM FIT INDEX / PARSCALE $G^2$ / GENERALIZED $S-\chi^2$ / TYPE I ERROR / POWER OF THE TEST

TEERANUCH JABPRAPAI : THE COMPARISON OF ITEM FIT INDEX EFFICIENCY BETWEEN PARSCALE $G^2$ AND GENERALIZED $S-\chi^2$. ADVISOR : ASST.PROF. NUTTAPORN LAWTHONG, Ph.D., 175 pp.

The purpose of this research was to compare the efficiency of two item fit indices- PARSCALE $G^2$ and GENERALIZED $S-\chi^2$. Data were simulated under two Polytomous item response theory models- Grade response model (GRM) and Generalized partial credit model (GPCM). Three conditions were manipulated: 1) three levels of test length (10, 20 and 40), 2) three levels of sample size (500, 1000 and 2000), 3) four levels of category (3, 5, 7 and 9). Seventy-two situations were analyzed. Type I error and power of the test were used as criteria to evaluate the efficiency of item fit index in this research in 1) the comparisons of two item fit indices efficiency by Kang and Chen (2008)' s condition, and 2) the comparisons of two item fit indices efficiency in two-way ANOVA.

The results of this research were :

1. The comparisons of two item fit indices efficiency in Kang and Chen (2008)' s condition, GENERALIZED $S-\chi^2$ had more efficiency than PARSCALE $G^2$ in most situations due to the fact that the type I error of GENERALIZED $S-\chi^2$ was less than the type I error of PARSCALE $G^2$ in 70 situations out of 72 situations (mean of PARSCALE $G^2$ ' s type I error = 0.1535, mean of GENERALIZED $S-\chi^2$ 's type I error = 0.0216).

2. The comparisons of two item fit indices efficiency in two-way ANOVA, type I error of GENERALIZED $S-\chi^2$ was less than the type I error of PARSCALE $G^2$ in 5 cases of 6 cases, and the power of the test of PARSCALE $G^2$ was greater than the power of the test of GENERALIZED $S-\chi^2$ in all 6 cases.

3. GENERALIZED $S-\chi^2$ had probability to indicate the fitted items were misfitted items less than PARSCALE $G^2$ (GENERALIZED $S-\chi^2$ had type I error less than PARSCALE $G^2$). While PARSCALE $G^2$ had probability to indicate the misfitted items were misfitted items more than GENERALIZED $S-\chi^2$ (PARSCALE $G^2$ had power of the test more than GENERALIZED $S-\chi^2$).

Department : Educational Research and Psychology Student's Signature .....

Field of Study : Educational Measurement and Evaluation Advisor's Signature .....

Academic Year : 2011

# กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จลุล่วงได้ โดยได้รับความกรุณาและเอื้อเฟื้ออย่างดียิ่งของ ผู้ช่วยศาสตราจารย์ ดร.ณัฏฐภรณ์ หลาวทอง อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ได้สละเวลาให้ความช่วยเหลือ แนะนำ ให้คำปรึกษาที่เป็นประโยชน์ นับตั้งแต่เริ่มแรกในการศึกษาวิจัย จวบจนกระทั่งวิทยานิพนธ์เสร็จสมบูรณ์โดยไม่เห็นแก่ความเหน็ดเหนื่อย อีกทั้งยังสนับสนุนส่งเสริมและให้กำลังใจแก่ผู้วิจัยเสมอมา ผู้วิจัยรู้สึกซาบซึ้งเป็นอย่างยิ่งและขอกราบขอบพระคุณท่านอาจารย์เป็นอย่างสูงไว้ ณ โอกาสนี้

ขอกราบขอบพระคุณ ศาสตราจารย์ ดร.ศิริชัย กาญจนวาสี และรองศาสตราจารย์ ดร.สุภมาส อังศุโชติ ที่ได้ให้ความกรุณาในการเป็นกรรมการสอบวิทยานิพนธ์และให้คำชี้แนะในการปรับปรุงวิทยานิพนธ์นี้ให้มีความสมบูรณ์มากยิ่งขึ้น

ขอกราบขอบพระคุณคณาจารย์ภาควิชาวิจัยและจิตวิทยาการศึกษาทุกท่าน ที่ได้ประสิทธิ์ประสาทความรู้แก่ผู้วิจัยอย่างดียิ่งตลอดระยะเวลาที่ศึกษา ตลอดทั้งคุณครู อาจารย์ทุกท่านที่ได้อบรมสั่งสอนวิชาความรู้ให้แก่ผู้วิจัยตั้งแต่วัยเยาว์จวบจนปัจจุบัน

สุดท้ายนี้ ขอกราบขอบพระคุณบิดามารดา พี่ชาย ที่เป็นกำลังใจและเอาใจใส่ช่วยเหลือผู้วิจัยเป็นอย่างดีเสมอมา ทุกท่านล้วนเป็นกำลังใจสำคัญที่ช่วยให้ผู้วิจัยผ่านพ้นอุปสรรคปัญหา จนสามารถดำเนินภารกิจต่างๆ ที่เกี่ยวข้องในการศึกษาวิจัยจนสำเร็จลุล่วงด้วยดี หากวิทยานิพนธ์ฉบับนี้จะเป็นประโยชน์ทางวิชาการแก่ผู้หนึ่งผู้ใด ผู้วิจัยขออนุโมทนาคุณความดีที่เกิดขึ้นนี้แด่คุณตา คุณยายผู้ล่วงลับ

# สารบัญ

หน้า

บทที่

หน้า

# สารบัญตาราง

ตารางที่ หน้า

ตารางที่ หน้า

ตารางที่ หน้า

ตารางที่ หน้า

ตารางที่ หน้า

# สารบัญแผนภาพ

# บทที่ 1

# บทนำ

## ความเป็นมาและความสำคัญของปัญหา

นับตั้งแต่ทฤษฎีการทดสอบแนวใหม่ (Modern Test Theory) ได้เข้ามามีบทบาทในการวัดคุณลักษณะภายในตัวบุคคลมากขึ้นและเป็นที่นิยมแพร่หลายในปัจจุบัน ก่อให้เกิดการค้นพบแนวคิดและองค์ความรู้เพิ่มเติมจากทฤษฎีดังกล่าวมาเป็นลำดับ ดังที่นักวิชาการได้ให้ความสำคัญในศาสตร์ทางด้านการวัด อาทิเช่น Wright และ Stone (1979) ได้กล่าวถึงความสำคัญของการวัดคุณลักษณะของบุคคลว่า การวัดคุณลักษณะในด้านใดด้านหนึ่งของบุคคล ผู้วัดจะต้องระบุคุณลักษณะที่ต้องการวัดให้ชัดเจนเพื่อที่จะสร้างข้อคำถามได้ตรงและครอบคลุมคุณลักษณะนั้น เพื่อให้ข้อคำถามที่สร้างขึ้นสามารถให้ผลการวัดที่มีความสม่ำเสมอคงเส้นคงวา ดังนั้นความสอดคล้องของข้อคำถามจึงเป็นเรื่องที่นักวัดผลจะต้องตระหนักและให้ความสำคัญ

ในการประยุกต์ทฤษฎีการทดสอบแนวใหม่ (Modern Test Theory) โดยเฉพาะอย่างยิ่งทฤษฎีการตอบสนองข้อสอบ (Item Response Theory) เพื่อใช้ในทางปฏิบัติให้ประสบผลสำเร็จขึ้นอยู่กับความสอดคล้อง (fit) ระหว่างโมเดลกับข้อมูลที่นำมาวิเคราะห์ (Dodeen, 2004; Liang & Wells, 2009) ซึ่ง Embretson และ Reise (2000) ได้ให้ความเห็นว่า ในการประยุกต์ใช้โมเดลทฤษฎีการตอบสนองข้อสอบ (IRT model) นั้นไม่มีความจำเป็นที่โมเดลเดียวกันจะนำไปประยุกต์ใช้กับข้อคำถามทุกข้อในแบบวัด เช่น แบบวัดมักประกอบด้วยข้อคำถามทั้งแบบที่มีการตรวจให้คะแนนเพียง 2 ค่าและมากกว่า 2 ค่า ซึ่งข้อคำถามบางข้ออาจนำเสนอโดยโมเดลโลจีสติกแบบ 2 พารามิเตอร์ (2 – parameter logistic model) และบางข้ออาจถูกนำเสนอโดย Grade response model (GRM) โมเดลทฤษฎีการตอบสนองข้อสอบที่แตกต่างกันถูกประมาณขึ้นสำหรับข้อคำถามแต่ละข้อในแบบวัด ซึ่งงานวิจัยจำนวนมากได้ให้ความสนใจในการศึกษาในเรื่องเกี่ยวกับการตัดสินความสอดคล้อง (fit) ของโมเดลทฤษฎีการตอบสนองข้อสอบบนพื้นฐานของการศึกษาข้อคำถามเป็นรายข้อ และไม่เห็นด้วยที่จะตัดสินความสอดคล้องโดยใช้การศึกษาทั้งโมเดล (Over – all model data fit)

อย่างไรก็ตาม ข้อคำถามที่มีลักษณะไม่ค่อยสอดคล้อง (poor item fit) อาจนำไปสู่ความล้มเหลวในการประมาณค่าพารามิเตอร์ของข้อคำถาม เช่น โมเดล 1 พารามิเตอร์สอดคล้องกับข้อมูลที่เป็น 2 พารามิเตอร์ หรือเกิดปัญหา nonmonotonicity of item–trait relation ซึ่งขัดแย้งกับกรณีทั่วไปที่ parametric IRT model มีข้อสมมติว่า เมื่อระดับคุณลักษณะ (trait–

level) เพิ่มขึ้นความน่าจะเป็นของการตอบถูกก็จะเพิ่มขึ้นด้วย นอกจากนี้ข้อคำถามที่มีลักษณะไม่ค่อยสอดคล้องจะให้คำอธิบายต่าง ๆ เกี่ยวกับคุณลักษณะของข้อคำถามได้เล็กน้อย (poor item construction) สารสนเทศที่ได้จากแบบวัดลดลง การทดสอบจึงไม่สามารถให้ประโยชน์ในการบ่งชี้ความสามารถเพื่อจำแนกบุคคลออกตามระดับคุณลักษณะภายในได้ดีเท่าที่ควรจะเป็น (Embretson & Reise, 2000)

เมื่อพิจารณาบริบทในเมืองไทย การศึกษาเกี่ยวกับดัชนีความสอดคล้องของข้อคำถามกับโมเดลทฤษฎีการตอบสนองข้อสอบ ยังไม่มีผู้ศึกษาไว้มากนัก แต่เมื่อพิจารณาในต่างประเทศจะพบว่ามีผู้สนใจศึกษากันในประเด็นนี้อย่างแพร่หลายโดยอาศัยแนวคิดพื้นฐานในการประเมินความสอดคล้องของข้อคำถาม คือ โมเดลทฤษฎีการตอบสนองข้อสอบนั้นสามารถอธิบายหรือทำนายการตอบข้อคำถามที่มีความจำเพาะเจาะจงได้ดีเพียงไร ซึ่งวิธีตรวจสอบเพื่อบ่งชี้ลงไป อาจใช้กราฟโดยการเปรียบเทียบระหว่าง โค้งการตอบสนองข้อสอบ (Item response curve) ที่ทำนายขึ้นกับโค้งการตอบสนองข้อสอบที่สังเกตได้จากข้อมูลจริง การเปรียบเทียบใช้วิธีการทางคณิตศาสตร์และสถิติเข้าช่วย วิธีที่เป็นที่นิยมมากคือการทดสอบความสอดคล้องเหมาะสม (Goodness of fit test) ที่ใช้ตัวสถิติไคสแควร์ในการทดสอบ แม้ในยุคต่อมาได้เกิดดัชนีความสอดคล้องของข้อคำถามตัวอื่น ๆ ตามมา ซึ่งได้มีการปรับและลดข้อจำกัดของดัชนีในยุคแรก แต่ก็ยังคงอาศัยแนวคิดหลักจากยุคแรกเริ่ม

นอกจากนี้ในยุคปัจจุบัน โปรแกรมคอมพิวเตอร์ที่ใช้ในการวิเคราะห์ข้อคำถามล้วนมีการบรรจุดัชนีความสอดคล้องของข้อคำถามลงไปในกระบวนการวิเคราะห์ข้อสอบด้วย เช่น PARSCALE และ BILOG ในบางโปรแกรมก็ถูกพัฒนาขึ้นมาเพื่อใช้ในการทดสอบเกี่ยวกับความสอดคล้องของข้อคำถามกับโมเดลการตอบข้อสอบโดยเฉพาะ เช่น IRTFIT macros ซึ่งประมวลผลบนโปรแกรม SAS จะเห็นได้ว่า ประเด็นในเรื่องความสอดคล้องของข้อคำถามกับโมเดลทฤษฎีการตอบสนองข้อสอบเป็นเรื่องที่น่าสนใจควรให้ความสำคัญเพื่อให้เกิดองค์ความรู้เพิ่มเติมในองค์ความรู้เดิมที่มีอยู่แล้วเป็นการช่วยให้ศาสตร์ในด้านการวัดและประเมินผลมีพัฒนาการก้าวหน้ายิ่งขึ้นไป

ถึงแม้ว่าแบบวัดที่ใช้ประเมินคุณลักษณะภายในตัวบุคคลจะเป็นเพียงสิ่งเร้าอย่างหนึ่งที่กระตุ้นให้ผู้ตอบแสดงความคิดหรือพฤติกรรมออกมา แต่อาจไม่สามารถบ่งบอกความสามารถของผู้ตอบได้ครบตามความเป็นจริงทุกประการ ดังนั้นการพัฒนาแบบวัดและกระบวนการในการวัด การวิเคราะห์ข้อคำถามให้มีคุณภาพจึงเป็นสิ่งจำเป็นที่จะต้องให้ความสำคัญ เพื่อให้ผลการวัดสอดคล้องกับความสามารถที่แท้จริงของผู้ตอบมากที่สุด ซึ่งการ

ตรวจสอบความสอดคล้องของข้อคำถามกับโมเดลทฤษฎีการตอบสนองข้อสอบ เป็นกระบวนการวิเคราะห์ข้อคำถามขั้นตอนหนึ่งที่นักวัดผลและผู้สนใจในศาสตร์การวัดไม่ควรมองข้ามนอกเหนือจากการศึกษาคุณภาพของข้อคำถามในด้านอื่นๆ

เมื่อศึกษาประเด็นที่เกี่ยวข้องกับดัชนีความสอดคล้องของข้อคำถาม (item fit index) ทำให้เห็นถึงวิวัฒนาการของดัชนีความสอดคล้องของข้อคำถามในภาพรวม (ค.ศ.1969-2010) ซึ่งในช่วงนี้ได้เกิดการคิดค้นดัชนีความสอดคล้องของข้อคำถามขึ้นมาจำนวนมากจากนักวิชาการต่างๆ ประกอบกับในปัจจุบันที่ข้อสอบทางการศึกษาและข้อคำถามทางจิตวิทยาให้ความสนใจนำโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค (Polytomous item response theory model) มาช่วยในการวิเคราะห์ข้อคำถามมากขึ้น เนื่องจากแบบวัดเจตคติ แบบวัดบุคลิกภาพ แบบวัดความสนใจในอาชีพโดยส่วนใหญ่มักอยู่ในลักษณะของมาตรประมาณค่า (ศิริชัย กาญจนวาสี, 2550) หรือแม้กระทั่งการทดสอบในปัจจุบันที่ให้ความสำคัญในการนำวิธีการตรวจให้คะแนนความรู้บางส่วน (Partial credit) มาใช้ดังปรากฏจากปริมาณงานวิจัยที่ศึกษาเกี่ยวกับวิธีการให้คะแนนความรู้บางส่วนที่มีเพิ่มขึ้นประกอบกับผลการวิจัยของเอมอร จังศิริพรปกรณ์ ในปี พ.ศ. 2545 ที่ทำการศึกษาเปรียบเทียบคุณภาพของแบบสอบในด้านความตรงตามสภาพ อำนาจจำแนก ความยาก ฟังก์ชันสารสนเทศของข้อสอบโดยใช้ทฤษฎีการตอบสนองข้อสอบเมื่อแบบสอบตรวจให้คะแนนแบบประเพณีนิยมและวิธีให้คะแนนความรู้บางส่วน ซึ่งพบว่าถ้าพิจารณาที่ฟังก์ชันสารสนเทศของข้อสอบวิธีประเพณีนิยมจะให้ค่าต่ำสุด แสดงให้เห็นถึงประสิทธิภาพของวิธีให้คะแนนความรู้บางส่วนที่จะเป็นประโยชน์มากขึ้นในการวิเคราะห์ข้อสอบ จึงเป็นเหตุผลหนึ่งที่สนับสนุนให้ผู้วิจัยเลือกใช้โมเดล GPCM มาใช้ในการวิจัย

นอกจากนี้จากการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง พบว่า ในการศึกษาวิจัยเกี่ยวกับดัชนีความสอดคล้องของข้อคำถาม (item fit index) โดยส่วนใหญ่จะใช้เทคนิคการจำลองแบบ (simulation) อย่างไรก็ตาม สถานการณ์ต่าง ๆ ที่จำลองขึ้นล้วนจำลองโดยอ้างอิงจากข้อมูลในสถานการณ์จริง ซึ่งเกิดขึ้นในการทดสอบทางการศึกษาและจิตวิทยา เช่น ในการจำลองข้อมูลการตอบโดยให้ค่าความยาก มีการแจกแจงแบบปกติที่มีค่าเฉลี่ยเป็น 0 และมีค่าความแปรปรวนเป็น 1 ซึ่งการแจกแจงนี้คล้ายกับลักษณะการแจกแจงของค่าพารามิเตอร์ที่พบในการทดสอบทางการศึกษา (Orlando & Thissen, 2003) แม้ในการจำลองปัจจัยที่ส่งผลต่อดัชนีความสอดคล้องของข้อคำถาม เช่น ความยาวแบบวัด และขนาดกลุ่มตัวอย่าง งานวิจัยที่ได้ศึกษาผ่านมาล้วนพยายามจำลองข้อมูลโดยให้เงื่อนไขเหล่านี้สอดคล้องกับบริบทความเป็นจริงมากที่สุด

ในการศึกษาเกี่ยวกับดัชนีความสอดคล้องของข้อคำถามมีความจำเป็นต้องใช้การจำลองแบบ (simulation) เนื่องจากไม่สามารถกระทำได้ในสถานการณ์จริง หรือถ้าทำได้ก็อาจใช้เวลามาก นอกจากนี้อาจมีปัจจัยภายนอกบางประการที่ส่งผลทำให้ผลการศึกษาวิจัยมีความคลาดเคลื่อนด้วย เช่น หากต้องการศึกษาว่า จำนวนรายการคำตอบ (number of category) ที่แตกต่างกันจะมีผลต่อประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามหรือไม่ หากทำการศึกษากับข้อมูลจริงโดยการเก็บข้อมูลจากกลุ่มตัวอย่างซึ่งต้องเก็บหลายครั้ง แต่ละครั้งใช้ข้อคำถามเดิมแต่เปลี่ยนจำนวนรายการคำตอบ การทำเช่นนี้กลุ่มตัวอย่างอาจเบื่อหน่าย ส่งผลให้ข้อมูลที่ได้ไม่น่าเชื่อถือ นอกจากนี้แล้วในการศึกษาเกี่ยวกับทฤษฎีการตอบสนองข้อสอบในประเด็นที่เกี่ยวกับผลที่เกิดขึ้นจากการละเมิดข้อตกลงเบื้องต้นก็มักใช้เทคนิคการจำลองแบบเพื่อให้ครอบคลุมสถานการณ์ที่เป็นไปได้ในสภาพจริงให้มากที่สุด

จากการอภิปรายผลวิจัยของ Kang และ Chen (2008) ซึ่งได้ทำการศึกษาประสิทธิภาพของดัชนี Generalized $S-\chi^2$ สำหรับโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค (polytomous item response theory model) โดยทำการเปรียบเทียบกับดัชนี PARSCALE $G^2$ โดยใช้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) และอำนาจการทดสอบ (Power of the test) ได้ให้ข้อเสนอแนะในการทำวิจัยต่อไปว่า ควรทำการศึกษาในรายการคำตอบอื่น ๆ นอกเหนือจาก 5 รายการคำตอบ เพื่อศึกษาว่าจำนวนรายการคำตอบ (number of category) ที่แตกต่างกันจะส่งผลอย่างไรต่อดัชนีความสอดคล้องของข้อคำถาม หรือให้ศึกษาในกรณีค่าพารามิเตอร์ความสามารถผู้ตอบมีการแจกแจงแบบอื่น ที่ไม่ใช่การแจกแจงแบบปกติ เช่น การแจกแจงแบบเอกรูป (uniform distribution) หรือกรณีที่ข้อมูลมีการแจกแจงเป็นโค้งที่เบ้ (skewed) ซึ่งสอดคล้องกับข้อเสนอแนะของ Liang และ Wells (2009) และงานวิจัยอื่น ๆ ที่ได้ศึกษารวบรวมมา นอกจากนี้ Lattuis, Clark และ O'Brien (2009) ได้ทำการศึกษาดัชนี $S-\chi^2$, $\chi^{2*}$ และ $\chi^2/dfs$ สำหรับ Grade Response Model (GRM) เช่นเดียวกับ Kang และ Chen (2010) ที่ศึกษาประสิทธิภาพของ Generalized $S-\chi^2$ สำหรับ Grade Response Model (GRM) โดยใช้ข้อมูลที่มีลักษณะเป็นพหุวิภาค (Polytomous Data) มีจำนวนรายการคำตอบ 3 และ 5 ตามลำดับ ซึ่งที่เขาเลือกจำนวนรายการคำตอบนี้ เนื่องจากจำนวนรายการคำตอบดังกล่าวเป็นที่นิยมใช้ในแบบสอบและแบบวัดทางจิตวิทยาโดยทั่วไป ซึ่งเขาได้ให้ข้อเสนอแนะว่าในการศึกษาครั้งต่อไปอาจทำการศึกษาในโมเดลอื่น ๆ เช่น PCM หรือ GPCM และจำนวนรายการคำตอบอื่น ๆ เพิ่มเติม

เมื่อพิจารณาการศึกษาวิจัยของนักวิชาการที่ผ่านมาเกี่ยวกับตัวแปรที่ส่งผลต่อประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามยังให้สารสนเทศที่สอดคล้องกัน กล่าวคือ เมื่อความยาวแบบวัด (Test length) มีค่าน้อยหรือเป็นแบบวัดที่สั้น ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) จะมีค่าค่อนข้างมาก (Orlando & Thissen, 2000; Demars, 2005; Kang & Chen, 2008; Liang & Wells, 2009) แต่เมื่อพิจารณาที่อำนาจการทดสอบ (Power of the test) พบว่า เมื่อความยาวแบบวัดเพิ่มมากขึ้น มีแนวโน้มที่อำนาจการทดสอบจะเพิ่มขึ้นด้วย (Kang & Chen, 2010) นอกจากนี้ เมื่อพิจารณาที่ขนาดกลุ่มตัวอย่าง (sample size) พบว่าเมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้น มีแนวโน้มที่ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) จะมีค่าลดลง (Stone & Zhang, 2003; Kang & Chen, 2010) แต่เมื่อพิจารณาที่อำนาจการทดสอบ (Power of the test) พบว่า เมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้น มีแนวโน้มที่อำนาจการทดสอบจะเพิ่มขึ้นตามไปด้วย (Stone & Zhang, 2003; Liang & Wells, 2009; Lattuis, Clark & O'Brein, 2009; Kang & Chen, 2010) อย่างไรก็ตามข้อค้นพบนี้อาจไม่เป็นจริงเสมอไปในทุกกรณี ในบางครั้งอาจมีอิทธิพลของขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ และชนิดของดัชนีความสอดคล้องของข้อสอบที่ส่งผลให้การเพิ่มขึ้นหรือลดลงของค่าอำนาจการทดสอบและค่าความคลาดเคลื่อนประเภทที่ 1 ไม่เป็นไปตามที่ค้นพบในทุกกรณี เช่น ผลการศึกษาของ Stone และ Zhang (2003) ที่อำนาจการทดสอบลดลงเล็กน้อยเมื่อความยาวแบบวัดเพิ่มขึ้น นอกจากนี้ผลการศึกษาวิจัยของ Kang และ Chen (2010) ยังให้สารสนเทศเพิ่มเติมว่าโดยส่วนใหญ่เมื่อมี 5 รายการคำตอบค่าความคลาดเคลื่อนประเภทที่ 1 จะมากกว่าเมื่อมี 3 รายการคำตอบ ซึ่งสันนิษฐานได้ว่าจำนวนรายการคำตอบอาจส่งผลต่อประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามจึงมีความจำเป็นที่ควรมีการศึกษาวิจัยประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามเพิ่มเติมในกรณีที่จำนวนรายการคำตอบนอกเหนือไปจาก 3 และ 5 เพราะการสร้างแบบวัดทางจิตวิทยายังนิยมใช้จำนวนรายการคำตอบเป็น 7 และ 9 อาทิ The High School Chemistry Self-Efficacy Scale (HCSS) ของ Aydin และ Uzuntiryaki (2009) ซึ่งใช้ 9 รายการคำตอบ และ The Schutte Self-Report Emotional Intelligence Scale on International Studentsของ Mun, Wang, Kim, และ Bodenhorn (2009)ซึ่งใช้ 7 รายการคำตอบเป็นต้น

นอกจากนี้ จากการศึกษางานวิจัยที่ผ่านมาในการจำลองข้อมูลเพื่อใช้ในการวิจัยเกี่ยวกับดัชนีความสอดคล้องของข้อคำถาม (Orlando & Thissen, 2000, 2003; Stone & Zhang, 2003; Dodeen, 2004; DeMars, 2005; Kang & Chen, 2008, 2010; Liang & Wells, 2009; Lattuis, Clark & O'Brien, 2009) นิยมใช้ความยาวของแบบวัด (test length) 10, 20, และ 40 ข้อ

เป็นจำนวนมาก ความยาวของแบบวัดสูงที่สุดคือ 80 ข้อ ความยาวของแบบวัดน้อยที่สุดคือ 5 ข้อ ขนาดกลุ่มตัวอย่างหรือจำนวนผู้ตอบข้อคำถาม เริ่มต้นที่ 500 คน ขนาดกลุ่มตัวอย่างที่นิยมใช้มากคือ 500, 1000, และ2000 คน ขนาดกลุ่มตัวอย่างสูงที่สุดคือ 5000 คน ซึ่งนักวิชาการส่วนใหญ่ให้ความเห็นว่าขนาดกลุ่มตัวอย่างที่ใช้ในการวิเคราะห์เกี่ยวกับทฤษฎีการตอบสนองข้อสอบควรใช้ขนาดกลุ่มตัวอย่างไม่ต่ำกว่า 500 คน สอดคล้องกับ Reise และ Yu (1990 อ้างถึงใน ศิริชัย กาญจนวาสี, 2550) ที่กล่าวว่าสามารถใช้ขนาดกลุ่มตัวอย่าง 250 คนในการประมาณค่าพารามิเตอร์ของโมเดล GRM โดยใช้โปรแกรม MULTILOG แต่ถ้าต้องการให้ได้ผลดีควรใช้ขนาดตัวอย่างไม่ต่ำกว่า 500 คน ซึ่งตามหลักการแล้วควรกำหนดขนาดกลุ่มตัวอย่างให้มีขนาดใหญ่พอที่จะทำให้ความคลาดเคลื่อนมาตรฐานของการประมาณค่าพารามิเตอร์มีค่าน้อยและอยู่ในระดับที่ยอมรับได้ตามเป้าหมายของการนำผลไปใช้ในทางปฏิบัติ (ศิริชัย กาญจนวาสี, 2550) ซึ่งKang และ Chen (2008) ให้เหตุผลเพิ่มเติมในการเลือกใช้ขนาดกลุ่มตัวอย่าง 500, 1000, และ 2000 คนว่าเป็นตัวแทนของขนาดตัวอย่างขนาดเล็ก ขนาดกลาง และขนาดใหญ่ตามลำดับ และได้ให้เหตุผลเพิ่มเติมในการเลือกใช้ความยาวแบบวัด 10 และ 20 ข้อว่าเป็นตัวแทนของความยาวแบบวัดขนาดกลางและขนาดใหญ่ จึงเป็นสาเหตุอย่างหนึ่งที่ทำให้ผู้วิจัยเลือกใช้ขนาดกลุ่มตัวอย่าง 500, 1000, และ 2000 คนกับความยาวแบบวัด 10, 20, และ 40 ข้อในการวิจัยครั้งนี้

เมื่อพิจารณาข้อคำถามที่มีการตรวจให้คะแนนแบบพหุวิภาค (polytomous item) พบว่า ดัชนีความสอดคล้องของข้อคำถามที่ใช้กับข้อคำถามลักษณะนี้จะมีไม่มากนักเมื่อเปรียบเทียบกับข้อคำถามที่มีการตรวจให้คะแนนแบบทวิวิภาค (dichotomous item) ซึ่งเท่าที่ผู้วิจัยได้ศึกษารวบรวมมามีดัชนีที่น่าสนใจ 2 ตัว คือ PARSCALE $G^2$ และ Generalized $S-\chi^2$ เนื่องจากดัชนีทั้งสองเป็นที่นิยมใช้ในการวิเคราะห์ข้อสอบหรือข้อคำถามโดยทั่วไป (Kang & Chen, 2008) และดัชนี PARSCALE $G^2$ ยังสามารถคำนวณได้อย่างสะดวกรวดเร็วจากโปรแกรม PARSCALE (Liang & Wells, 2009) โดยจากการศึกษางานวิจัยที่เกี่ยวข้องให้สารสนเทศที่คล้ายกันว่าดัชนี Generalized $S-\chi^2$ เป็นดัชนีที่มีประสิทธิภาพในการบ่งชี้ความสอดคล้องของข้อคำถามเนื่องจากมักจะให้ค่าความคลาดเคลื่อนประเภทที่ 1 และค่าอำนาจการทดสอบอยู่ในเกณฑ์ที่เหมาะสมมากกว่าดัชนีตัวอื่น ๆ ที่นำมาใช้ในการศึกษาเปรียบเทียบ (Kang & Chen, 2008; Liang & Wells, 2009; Lattuis, Clark & O'Brien, 2009) จึงเป็นที่น่าสนใจว่าหากสถานการณ์เปลี่ยนแปลงไปจากการทดลองที่นักวิชาการท่านอื่น ๆ ได้ศึกษามา ดัชนี Generalized $S-\chi^2$ จะยังคงมีประสิทธิภาพเช่นเดิมหรือไม่

จากข้อมูลที่ได้กล่าวมาทั้งหมดทำให้ผู้วิจัยเกิดความสนใจที่จะทำการศึกษาวิจัยเกี่ยวกับการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามสำหรับโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค ภายใต้ความแตกต่างของความยาวแบบวัด ขนาดกลุ่มตัวอย่างและจำนวนรายการคำตอบ โดยผู้วิจัยมีความสนใจว่า ณ ที่สถานการณ์ทั้ง 72 สถานการณ์ที่ทำการศึกษา ประสิทธิภาพของดัชนี Generalized $S-\chi^2$ และดัชนี PARSCALE $G^2$ ดัชนีใดที่มีความเหมาะสมที่จะนำไปใช้บ่งชี้ความสอดคล้องของข้อคำถามในสถานการณ์นั้นๆ มากกว่ากัน ซึ่งโมเดลที่นำมาใช้ในการวิเคราะห์ประกอบด้วย GRM และ GPCM เนื่องจากโมเดลทั้งสองพัฒนาบนพื้นฐานของโมเดลแบบ 2 พารามิเตอร์เหมือนกัน ซึ่งเหมาะสำหรับข้อสอบหรือข้อคำถามที่มีความยากและค่าอำนาจจำแนกที่แตกต่างกัน (ศิริชัย กาญจนวาสี, 2550) นอกจากนี้โมเดลทั้งสองยังเป็นโมเดลที่มีผู้สนใจศึกษากันมาก ไม่เข้มงวดเกี่ยวกับข้อตกลงเบื้องต้นและสามารถใช้กับแบบสอบและแบบวัดหลายลักษณะ (Donoghue, 1994; De Ayala, 1994; Muraki, 1992, 1993; Reise & Yu, 1990; Koch & De Ayala, 1989; Koch, 1983 อ้างถึงใน เอมอร จังศิริพรปกรณ์, 2545) โดยในการศึกษาวิจัยครั้งนี้ ผู้วิจัยมีความมุ่งหวังว่าผลการวิจัยจะเป็นแนวทางในการเลือกใช้ดัชนีความสอดคล้องของข้อคำถามสำหรับโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค ในสถานการณ์ที่ความยาวแบบวัด ขนาดกลุ่มตัวอย่างและจำนวนรายการคำตอบมีความแตกต่างกัน

**คำถามการวิจัย**

ในสถานการณ์ ณ ความยาวแบบวัดเป็น 10, 20, และ 40 ข้อ ขนาดกลุ่มตัวอย่างเป็น 500, 1000 และ 2000 คน จำนวนรายการคำตอบเป็น 3, 5, 7, และ 9 รายการ ดัชนี Generalized $S-\chi^2$ และดัชนี PARSCALE $G^2$ มีประสิทธิภาพในการบ่งชี้ความสอดคล้องของข้อคำถามแตกต่างกันหรือไม่ อย่างไร

**วัตถุประสงค์ของการวิจัย**

เพื่อเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถาม (item fit index) สำหรับโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค สองชนิดคือ Generalized $S-\chi^2$ และ PARSCALE $G^2$ ในสถานการณ์ ณ ความยาวแบบวัดเป็น 10, 20, และ 40 ข้อ ขนาดกลุ่มตัวอย่างเป็น 500, 1000, และ 2000 คน จำนวนรายการคำตอบเป็น 3, 5, 7, และ 9 รายการ

**สมมติฐานการวิจัย**

ดัชนี Generalized $S-\chi^2$ มีประสิทธิภาพในการบ่งชี้ความสอดคล้องของข้อคำถามมากกว่า ดัชนี PARSCALE $G^2$ ในสถานการณ์ ณ ความยาวแบบวัดเป็น 10, 20, และ 40 ข้อ ขนาดกลุ่มตัวอย่างเป็น 500, 1000, และ 2000 คน จำนวนรายการคำตอบเป็น 3, 5, 7, และ 9 รายการ

**ขอบเขตของการวิจัย**

1. การวิจัยครั้งนี้ใช้วิธีการศึกษาโดยการจำลองข้อมูลเพื่อศึกษาเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถาม 2 ชนิด คือ Generalized $S-\chi^2$ และ PARSCALE $G^2$ ซึ่งสถานการณ์การจำลองข้อมูลเป็นไปตามเงื่อนไขคือ ตัวแปรอิสระ 3 ตัวคือ 1) ความยาวแบบวัด 3 ขนาดคือ 10, 20, และ 40 ข้อ 2) ขนาดกลุ่มตัวอย่าง 3 ขนาดคือ 500, 1000, และ 2000 คน 3) จำนวนรายการคำตอบ 4 ขนาดคือ 3, 5, 7, และ 9 รายการ โดยมีประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามเป็นตัวแปรตาม ภายใต้โมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค
2. โมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค ที่ใช้ในการศึกษาวิจัยนี้เป็นโมเดลทฤษฎีการตอบสนองข้อสอบแบบเอกมิติ (Unidimensional IRT model) ประกอบด้วย 2 โมเดลคือ Grade Response Model (GRM) และ Generalized Partial Credit Model (GPCM) เนื่องจากโมเดลทั้งสองพัฒนาบนพื้นฐานของโมเดลแบบ 2 พารามิเตอร์เหมือนกัน ซึ่งเหมาะสำหรับข้อสอบหรือข้อคำถามที่มีความยากและค่าอำนาจจำแนกที่แตกต่างกัน (ศิริชัย กาญจนวาสี, 2550) นอกจากนี้โมเดลทั้งสองยังเป็นโมเดลที่มีผู้สนใจศึกษากันมาก ไม่เข้มงวดเกี่ยวกับข้อตกลงเบื้องต้นและสามารถใช้กับแบบสอบและแบบวัดหลายลักษณะ (Donoghue, 1994; De Ayala, 1994; Muraki, 1992, 1993; Reise & Yu, 1990; Koch & De Ayala, 1989; Koch, 1983 อ้างถึงใน เอมอร จังศิริพรปกรณ์, 2545)
3. เกณฑ์ที่นำมาใช้ในการประเมินประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถาม คือ ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) และอำนาจการทดสอบ (Power of the test) ซึ่งขั้นตอนในการพิจารณา จะพิจารณาที่ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ก่อน ซึ่งถ้าดัชนีชนิดใดมีค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) น้อยกว่าดัชนีนั้นจะเป็นดัชนีที่มีประสิทธิภาพในการบ่งชี้ความสอดคล้องของข้อคำถามกับโมเดล

ทฤษฎีการตอบสนองข้อสอบในสถานการณ์นั้น แต่ถ้าหากดัชนีทั้งสองมีค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) เท่ากันจึงจะพิจารณาอำนาจการทดสอบ (Power of the test) โดยถ้าดัชนีชนิดใดมีอำนาจการทดสอบที่มากกว่า ดัชนีนั้นจะเป็นดัชนีที่มีประสิทธิภาพในการบ่งชี้ความสอดคล้องของข้อคำถามกับโมเดลทฤษฎีการตอบสนองข้อสอบในสถานการณ์นั้น นอกจากนี้ในการวิเคราะห์ข้อมูลในการศึกษาวิจัยครั้งนี้ใช้เทคนิคการจำลองแบบ (simulation) ในการจำลองข้อมูลที่ใช้ในการศึกษาวิจัยทั้ง 72 สถานการณ์ โดยในแต่ละสถานการณ์ที่ทำการศึกษามีการกระทำซ้ำทั้งหมด 30 ครั้ง ซึ่งหลังจากที่ได้ทำการจำลองข้อมูลตามสถานการณ์ที่ใช้ในการศึกษาด้วยโปรแกรม WINGEN แล้วจะทำให้ผู้วิจัยได้โมเดลจำลอง (Generating Model: GM) มา 30 โมเดลในแต่ละสถานการณ์ ดังนั้นในการศึกษาวิจัยนี้จึงมีโมเดลจำลอง (Generating Model: GM) ทั้งหมด 30 x 72 = 2160 โมเดล จากนั้นจึงนำโมเดลจำลองที่ได้ไปตรวจสอบค่าพารามิเตอร์และประมาณค่าดัชนีความสอดคล้องของข้อคำถามในแต่ละโมเดลโดยถ้าเป็นดัชนี Generalized $S-\chi^2$ จะวิเคราะห์ด้วย IRTFIT macros ในโปรแกรม SAS แต่ถ้าหากเป็นดัชนี PARSCALE $G^2$ จะวิเคราะห์ด้วยโปรแกรม PARSCALE ซึ่งหลังจากผ่านขั้นตอนในการวิเคราะห์นี้แล้ว โมเดลจำลอง (Generating Model : GM) จะกลายเป็นโมเดลเทียบมาตรฐาน (Calibrating Model: CM) ดังนั้น ณ สถานการณ์ใด ๆ ที่ใช้ในการศึกษาจะแบ่งโมเดลเป็น 2 ประเภท คือ โมเดลจำลอง (Generating Model : GM) และโมเดลเทียบมาตรฐาน (Calibrating Model: CM) ถ้า GM กับ CM เป็นโมเดลที่ตรงกัน (เช่นเป็น GRM เหมือนกัน) จะประเมินประสิทธิภาพโดยดูที่ Type I error แต่ถ้า CM ไม่ตรงกับ GM จะใช้ในการคำนวณ Power of the test ดังตารางที่ 1.1

**ตารางที่ 1.1 การออกแบบในการประเมินประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถาม**

| Calibrating Model (CM) | Generating Model (GM) | |
|---|---|---|
| | GRM | GPCM |
| GRM | Type I error | Power of the test |
| GPCM | Power of the test | Type I error |

**ข้อตกลงเบื้องต้น**

1. การวิจัยครั้งนี้ใช้วิธีการศึกษาโดยการจำลองข้อมูล ซึ่งข้อมูลที่จำลองขึ้นนั้น อยู่บนข้อตกลงเบื้องต้นดังนี้

   1.1 กลุ่มตัวอย่างซึ่งเป็นผู้ตอบข้อคำถามได้มาจากการสุ่มของประชากร ซึ่งระดับความสามารถของผู้ตอบข้อคำถามมีการแจกแจงแบบปกติที่มีค่าเฉลี่ยเป็น 0 ส่วนเบี่ยงเบนมาตรฐานเป็น 1

   1.2 ข้อมูลที่ใช้ในการศึกษาวิเคราะห์จำลองขึ้นโดยใช้โปรแกรม WINGEN ซึ่งข้อมูลที่จำลองขึ้นนั้น กำหนดให้มีลักษณะเป็นไปตาม Grade Response Model (GRM) และ Generalized Partial Credit Model (GPCM) ดังนั้นการวิจัยนี้จึงดำเนินการภายใต้ข้อสมมติที่ว่า ข้อมูลที่จำลองขึ้นจากโปรแกรม WINGEN นั้น มีลักษณะสอดคล้องและเป็นไปตามโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาคที่ใช้ในการศึกษาวิจัยทั้ง 2 โมเดล
2. โมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค ที่ใช้ในการศึกษาวิจัยนี้เป็นโมเดลทฤษฎีการตอบสนองข้อสอบแบบเอกมิติ (Unidimensional IRT model) ประกอบด้วย 2 โมเดลคือ Grade Response Model (GRM) และ Generalized Partial Credit Model (GPCM)

**ข้อจำกัดของการวิจัย**

1. การวิจัยนี้ใช้วิธีการศึกษาโดยการจำลองข้อมูลด้วยเทคนิคการจำลองแบบ (Simulation) ทั้งนี้เนื่องมาจากวัตถุประสงค์ของการวิจัยที่ต้องการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถาม (item fit index) ทั้งสองชนิดในหลากหลายสถานการณ์ ซึ่งทำให้ไม่สามารถทำการศึกษาโดยใช้ข้อมูลจริงได้
2. ในการศึกษาวิจัยครั้งนี้จำลองข้อมูลโดยใช้โปรแกรม WINGEN ซึ่งอยู่ภายใต้ข้อสมมติที่ว่า “ข้อมูลที่จำลองขึ้นนั้นมีลักษณะเป็นไปตามโมเดลทฤษฎีการตอบสนองข้อสอบแบบเอกมิติ 2 โมเดล คือ Grade Response Model (GRM) และ Generalized Partial Credit Model (GPCM)” ซึ่งในขั้นตอนของการวิจัยมีเพียงการตรวจสอบข้อมูลที่จำลองขึ้นว่ามีลักษณะเป็นไปตามเงื่อนไขที่กำหนดขึ้นหรือไม่ในเรื่องลักษณะการแจกแจงของค่าพารามิเตอร์ความสามารถผู้สอบ($\theta$) และลักษณะการแจกแจงของค่าพารามิเตอร์ข้อคำถาม โดยยังไม่ได้มีการตรวจสอบข้อมูลใน

ประเด็นเกี่ยวกับความสอดคล้องของข้อมูลที่จำลองขึ้นกับโมเดลทฤษฎีการตอบสนองข้อสอบโดยใช้ค่าสถิติทดสอบ

**คำจำกัดความที่ใช้ในการวิจัย**

ดัชนีความสอดคล้องของข้อคำถาม (item fit index) หมายถึง ค่าสถิติที่บ่งบอกว่าข้อคำถามข้อนั้นมีความสอดคล้องหรือแตกต่างจากค่าที่คาดหวังในโมเดลทฤษฎีการตอบสนองข้อสอบซึ่งเป็นไปตามทฤษฎีการตอบสนองข้อสอบ โดยในการศึกษาวิจัยนี้ศึกษาดัชนี 2 ชนิดคือ Generalized $S-\chi^2$ และ PARSCALE $G^2$

ดัชนี Generalized $S-\chi^2$ หมายถึง ดัชนีความสอดคล้องของข้อคำถามซึ่งพัฒนามาจากดัชนี $S-\chi^2$ ของ Orlando และ Thissen (2000) ซึ่งนำมาใช้กับข้อคำถามที่มีการตรวจให้คะแนนแบบพหุวิภาค

ดัชนี PARSCALE $G^2$ หมายถึง ดัชนีความสอดคล้องของข้อคำถามของ Muraki และ Bock (1997) ซึ่งเป็นดัชนีที่นิยมใช้ทั้งในข้อคำถามที่มีการตรวจให้คะแนนแบบทวิวิภาคและพหุวิภาค

ประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถาม หมายถึง ประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามในการตรวจสอบความสอดคล้องของข้อคำถามกับโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค ซึ่งวัดจากความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบ

ความคลาดเคลื่อนประเภทที่ 1 (Type I error : $\alpha$) หมายถึง โอกาสในการปฏิเสธสมมติฐานที่เป็นจริง ในการวิจัยนี้คือ การบ่งชี้ว่าข้อคำถามข้อนั้นไม่สอดคล้อง (misfit) ทั้ง ๆ ที่ข้อคำถามข้อนั้นมีความสอดคล้อง (fit) กับโมเดลทฤษฎีการตอบสนองข้อสอบ

อำนาจการทดสอบ (Power of the test : 1-$\beta$) หมายถึง โอกาสในการปฏิเสธสมมติฐานที่เป็นเท็จ ในการวิจัยนี้คือ การบ่งชี้ว่าข้อคำถามข้อนั้นไม่สอดคล้อง (misfit) ซึ่งในความเป็นจริงข้อคำถามข้อนั้นไม่มีความสอดคล้อง (misfit) กับโมเดลทฤษฎีการตอบสนองข้อสอบ

โมเดลจำลอง (Generating Model: GM) หมายถึง โมเดลการตอบข้อคำถามของผู้ตอบแบบวัดซึ่งผู้วิจัยจำลองขึ้นโดยใช้โปรแกรม WINGEN ซึ่งจะมีทั้งหมด 30 โมเดลในแต่ละ

สถานการณ์ที่ทำการศึกษา และโมเดลนี้ประกอบด้วย 2 โมเดลคือ Grade Response Model (GRM) และ Generalized Partial Credit Model (GPCM)

โมเดลเทียบมาตรฐาน (Calibrating Model: CM) หมายถึง โมเดลจำลอง (Generating Model: GM) ที่นำไปตรวจสอบค่าพารามิเตอร์และประมาณค่าดัชนีความสอดคล้องของข้อคำถามในแต่ละโมเดล โดยถ้าเป็นดัชนี Generalized $S-\chi^2$ จะวิเคราะห์ด้วย IRTFIT macros ในโปรแกรม SAS แต่ถ้าหากเป็นดัชนี PARSCALE $G^2$ จะวิเคราะห์ด้วยโปรแกรม PARSCALE ซึ่งหลังจากผ่านขั้นตอนในการวิเคราะห์นี้แล้ว โมเดลจำลอง (Generating Model: GM) จะกลายเป็นโมเดลเทียบมาตรฐาน (Calibrating Model: CM)

โมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค หมายถึง โมเดลทฤษฎีการตอบสนองข้อสอบแบบเอกมิติ (Unidimensional IRT model) ประกอบด้วย 2 โมเดลคือ Grade Response Model (GRM) และ Generalized Partial Credit Model (GPCM)

ค่าพารามิเตอร์ของข้อคำถาม หมายถึง ค่าสถิติที่บ่งบอกคุณภาพของข้อคำถามตามทฤษฎีการตอบสนองข้อสอบ ได้แก่ ค่าพารามิเตอร์ระดับความยากของขั้นการตอบที่ j ในข้อที่ i ($\delta_{ij}$) ค่าพารามิเตอร์ความชันของข้อคำถามที่ i ($\alpha_i$) และค่าพารามิเตอร์ Threshold ของแต่ละรายการคำตอบของข้อที่ i ($\beta_{ij}$)

ขนาดกลุ่มตัวอย่าง หมายถึง จำนวนผู้ตอบข้อคำถามมี 3 ขนาดคือ 500, 1000, และ 2000 คน

ความยาวแบบวัด หมายถึง จำนวนข้อคำถามในแต่ละแบบวัดมี 3 ขนาดคือ 10, 20, และ 40 ข้อ

จำนวนรายการคำตอบ หมายถึง จำนวนรายการคำตอบในแต่ละข้อคำถามมี 4 ขนาดคือ 3, 5, 7, และ 9 รายการ

การแจกแจงของค่าพารามิเตอร์ความสามารถผู้สอบ หมายถึง ค่าพารามิเตอร์ความสามารถผู้สอบ ($\theta$) ซึ่งมีการแจกแจงแบบปกติมาตรฐาน (standard normal distribution) ที่มีค่าเฉลี่ยเป็น 0 และค่าความแปรปรวนเป็น 1

**ประโยชน์ที่คาดว่าจะได้รับ**

1. ทำให้ทราบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามสำหรับ Grade Response Model (GRM) และ Generalized Partial Credit Model (GPCM) สองตัวคือ Generalized $S-\chi^2$ และ PARSCALE $G^2$ ในสถานการณ์ที่ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง และจำนวนรายการคำตอบ มีความแตกต่างกัน ซึ่งจะเป็นแนวทางในการกำหนดความยาวแบบวัด ขนาดกลุ่มตัวอย่าง และจำนวนรายการคำตอบในการสร้างแบบวัดทางจิตวิทยาหรือแบบสอบทางการศึกษาที่มีการตรวจให้คะแนนแบบพหุวิภาคภายใต้เงื่อนไขว่า แบบวัดทางจิตวิทยาหรือแบบสอบทางการศึกษานั้นมีความตรงและความเที่ยงในการวัดหรือจำแนกลักษณะของบุคคล
2. เป็นแนวทางในการเลือกใช้ดัชนีความสอดคล้องของข้อคำถามสำหรับ Grade Response Model (GRM) และ Generalized Partial Credit Model (GPCM) ในสถานการณ์ที่จำนวนรายการคำตอบ ขนาดกลุ่มตัวอย่างและความยาวแบบวัดมีความแตกต่างกัน เพื่อให้การบ่งชี้ความสอดคล้องของข้อสอบมีความคลาดเคลื่อนน้อยที่สุด
3. ให้สารสนเทศ เพื่อเสริมสร้างองค์ความรู้ที่เกี่ยวกับการศึกษาดัชนีความสอดคล้องของข้อคำถามต่อไป

# บทที่ 2

# เอกสารและงานวิจัยที่เกี่ยวข้อง

ผู้วิจัยนำเสนอแนวคิดและทฤษฎีที่เกี่ยวข้องกับดัชนีความสอดคล้องของข้อคำถาม (item fit index) และโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค โดยแบ่งการนำเสนอออกเป็น 8 ตอน คือ

ตอนที่ 1 ความเป็นมาของดัชนีความสอดคล้องของข้อคำถามกับโมเดลทฤษฎีการตอบสนองข้อสอบ

ตอนที่ 2 แนวคิดและการคำนวณดัชนีความสอดคล้องของข้อคำถามที่ใช้กับโมเดลทฤษฎีการตอบสนองข้อสอบแบบทวิวิภาค (Dichotomous Item Response Theory Model)

ตอนที่ 3 แนวคิดและการคำนวณดัชนีความสอดคล้องของข้อคำถามที่ใช้กับโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค (Polytomous Item Response Theory Model)

ตอนที่ 4 การตรวจสอบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถาม (item fit index)

ตอนที่ 5 โมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค (Polytomous item response theory model) ที่ใช้ในการวิจัย

ตอนที่ 6 เอกสารและงานวิจัยที่เกี่ยวข้อง

ตอนที่ 7 การทดสอบความสอดคล้องของข้อคำถามกับโมเดลทฤษฎีการตอบสนองข้อสอบ

ตอนที่ 8 กรอบแนวคิดที่ใช้ในการวิจัย

## ตอนที่ 1 ความเป็นมาของดัชนีความสอดคล้องของข้อคำถามกับโมเดลทฤษฎีการตอบสนองข้อสอบ

ความสอดคล้องของข้อคำถามกับโมเดลทฤษฎีการตอบสนองข้อสอบ เป็นเรื่องที่นักวิชาการได้สนใจศึกษามาอย่างยาวนาน มีการใช้การทดสอบสมมติฐานทางสถิติที่มีชื่อเรียกว่า Goodness of fit test เป็นเครื่องมือช่วยในการตัดสินใจถึงความสอดคล้องของข้อคำถาม เป็นที่มาของการเกิดดัชนีความสอดคล้องของข้อคำถามกับโมเดลทฤษฎีการตอบสนองข้อสอบ ในระยะแรก ดัชนีความสอดคล้องของข้อคำถามจะมีเพียงข้อคำถามที่ให้คะแนนแบบทวิวิภาค

(dichotomous item) เท่านั้น ในระยะหลังจึงได้มีการคิดค้นเพิ่มเติมให้เกิดดัชนีที่สามารถใช้กับข้อคำถามที่ให้คะแนนแบบพหุวิภาค (polytomous item) ได้

ดัชนีความสอดคล้องของข้อคำถาม กรณีข้อคำถามที่ให้คะแนนแบบทวิวิภาค (dichotomous item) มีนักวิชาการคิดค้นขึ้นมาหลายท่าน (Hambleton & Swaminathan, 1996; Demars, 2005; Stone & Zhang, 2003; Liang & Wells, 2009; Dodeen, 2004; Kang & Chen, 2008; Embretson & Reise, 2000) ซึ่งได้มีการพัฒนาเป็นลำดับตั้งแต่อดีตจนถึงปัจจุบัน ดังในปี ค.ศ. 1969 ที่ Wright และPanchapakesan ได้คิดค้นค่าสถิติไคสแควร์ที่ใช้ในการทดสอบความสอดคล้องของข้อสอบหรือข้อคำถามเป็นรายข้อซึ่งใช้กับโมเดล 1 พารามิเตอร์ ซึ่งวิธีนี้สามารถขยายไปสู่โมเดล 2 และ 3 พารามิเตอร์ได้

ต่อมา Divgi และ Wollenberg (1980,1981,1982 อ้างถึงใน Hambleton & Swaminathan,1996) ได้เสนอว่า วิธีการทดสอบความสอดคล้องของข้อคำถามของ Wright–Panchapakesan ตัวสถิติที่ใช้ในการทดสอบไม่มีการแจกแจงแบบไคสแควร์ และองศาความเป็นอิสระ (degree of freedom) มีค่ามากกว่าความเป็นจริง ในช่วงปี ค.ศ. 1970 – 1973 ได้มีการเสนอการทดสอบความสอดคล้องของ Rasch model และ Normal ogive model แบบ 2 พารามิเตอร์โดย Anderson, Bock และ Liebermann

Yen's (1981 อ้างถึงใน Demars 2005) ได้เสนอดัชนี Q1 ที่ใช้ในการตรวจสอบความสอดคล้องของข้อคำถามที่มีการตรวจให้คะแนนเพียง 2 ค่า ซึ่งมีความคล้ายคลึงกับดัชนี $\chi_B^2$ ของ Bock (1972) ที่สร้างมาสำหรับใช้กับข้อคำถามที่มีการประมาณค่าพารามิเตอร์โดยวิธี joint maximum likelihood ซึ่งองศาความเป็นอิสระถูกปรับค่าโดยจำนวนพารามิเตอร์ของข้อคำถาม นอกจากนี้กระบวนการอีกอย่างหนึ่งที่ใช้ในการตรวจสอบเพื่อบ่งชี้ว่าข้อคำถามมีความสอดคล้องกับโมเดลทฤษฎีการตอบสนองข้อสอบ คือการใช้วิธี Standardize Residuals (SRs) โดยเป็นการหาความแตกต่างระหว่างค่าคาดหวังกับค่าที่สังเกตได้จริง ในแต่ละระดับของความสามารถของผู้สอบ ซึ่งค่าความแตกต่างนี้ เรียกว่า residual นำค่า residual มาหารด้วยส่วนเบี่ยงเบนมาตรฐานของค่าคาดหวัง (expected performance) ซึ่งถ้าข้อมูลมีความสอดคล้องกับโมเดล การแจกแจงของค่า standardize residual จะมีการแจกแจงเข้าใกล้การแจกแจงแบบปกติมาตรฐาน ทำให้การทดสอบความสอดคล้องเหมาะสมของข้อคำถาม สามารถใช้ค่าสถิติซึ่งอ้างอิงมาจากตารางการแจกแจงปกติมาตรฐานได้

นอกเหนือจากดัชนีที่ได้กล่าวมาแล้ว ในการบ่งชี้ความสอดคล้องของข้อคำถามยังมีดัชนีอื่น ๆ อีก เช่น $G^2$ ของ McMinley and Mill's ซึ่ง Kang และ Chen (2008) ได้ให้ความเห็นว่าดัชนีนี้มีความคล้ายคลึงกับ PARSCALE $G^2$ ของ Muraki และ Bock (1997) ในแง่ของการให้ค่า Type I Error ที่ค่อนข้างมาก และมีความไว (sensitivity) ต่อความยาวของแบบวัด (Test Length) และขนาดกลุ่มตัวอย่าง (Sample Size) สอดคล้องกับผลในการทำวิจัยด้วยการจำลองแบบของ Orlando และ Thissen (2000 อ้างถึงใน Demars, 2005) ที่แสดงให้เห็นว่า $Q_1$ ของ Yen's (1981) และ $G^2$ ของ McKinley และ Mill's (1985) ให้ค่า Type I error ที่มากเมื่อแบบวัดมีการตรวจให้คะแนนเพียง 2 ค่า และเป็นแบบวัดที่สั้น จากปัญหาเกี่ยวกับความไว (sensitivity) ต่อความยาวของแบบวัดและขนาดกลุ่มตัวอย่าง ประกอบกับการมีค่าความผิดพลาดประเภทที่ 1 (Type I error) ที่มากจึงทำให้เกิดการใช้ดัชนีข้อคำถามที่เหมาะสมที่มีชื่อเรียกว่า $S-\chi^2$ (Orlando & Thissen, 2000 อ้างถึงใน Kang & Chen, 2008) ซึ่งมีกระบวนการที่คล้ายคลึงกับดัชนี $Q_1$ แต่มีข้อดีมากกว่า $Q_1$ และ $G^2$ เนื่องจาก $Q_1$ และ $G^2$ มีกระบวนการจัดกลุ่ม (grouping procedure) ขึ้นอยู่บน model-dependent ability estimate ในขณะที่ $S-\chi^2$ มีกระบวนการจัดกลุ่ม (grouping procedure) ขึ้นอยู่บน test score เช่น จำนวนข้อสอบที่ตอบถูก เป็นต้น

ดัชนีที่กล่าวมาแล้วทั้งหมดในข้างต้น เป็นดัชนีความสอดคล้องของข้อคำถามที่ให้คะแนนแบบทวิวิภาค (dichotomous) เท่านั้น ในการที่จะประเมินความสอดคล้องของข้อคำถามแบบที่ให้คะแนนแบบพหุวิภาค (polytomous) นั้น ในระยะแรกใช้ดัชนี PARSCALE $G^2$ ซึ่งสามารถใช้ได้ทั้งข้อคำถามที่ให้คะแนนแบบทวิวิภาค (dichotomous) และข้อคำถามแบบที่ให้คะแนนแบบพหุวิภาค (polytomous) ซึ่งดัชนีตัวนี้เป็นที่ใช้กันอย่างแพร่หลายเนื่องจากมีโปรแกรมคอมพิวเตอร์ช่วยในการคำนวณ อย่างไรก็ตาม ปัญหาที่พบมีความคล้ายคลึงกับที่ผ่านมา คือ ดัชนีมีความไวต่อความยาวแบบวัดและขนาดกลุ่มตัวอย่าง (Kang & Chen, 2008) ในปี ค.ศ.2008 Kang และ Chen ได้นำดัชนี Generalized $S-\chi^2$ มาใช้ในการบ่งชี้ข้อคำถามที่สอดคล้องกับโมเดลทฤษฎีการตอบสนองข้อสอบแบบให้คะแนนมากกว่า 2 ค่า ประกอบด้วย Generalized partial credit model (GPCM) , Rating scale model (RSM), Partial credit model (PCM) โดยทำการศึกษาเปรียบเทียบดัชนี PARSCALE $G^2$ ผลการศึกษาให้ข้อมูลที่น่าสนใจว่า ประสิทธิภาพของดัชนี Generalized $S-\chi^2$ จะมีความคงเส้นคงวามากกว่าดัชนี PARSCALE $G^2$ ในทุกเงื่อนไขที่ทำการศึกษา แม้ในกรณีความยาวแบบวัดที่สั้น และขนาดกลุ่มตัวอย่างมีจำนวนน้อย

ดังนั้นในปัจจุบัน ดัชนีความสอดคล้องของข้อคำถามกับโมเดลทฤษฎีการตอบสนองข้อสอบสำหรับข้อคำถามที่ให้คะแนนแบบพหุวิภาคที่น่าสนใจในการนำไปใช้ในทางปฏิบัติให้มากขึ้น คือ ดัชนี Generalized $S-\chi^2$ ซึ่งมีโปรแกรมคอมพิวเตอร์ที่สามารถช่วยในการคำนวณได้เช่นเดียวกับดัชนี PARSCALE's $G^2$ ที่มีโปรแกรม PARSCALE ช่วยในการคำนวณ ในขณะที่ Generalized $S-\chi^2$ สามารถใช้โปรแกรม SAS ช่วยในการคำนวณได้

จากการศึกษาเอกสารงานวิจัยที่เกี่ยวข้องพบว่า ในระยะแรก ดัชนีความสอดคล้องของข้อคำถาม (Item fit index) ได้ถูกนำมาใช้กับข้อคำถามที่ให้คะแนนแบบทวิวิภาค (dichotomous) แล้วจึงได้มีการพัฒนาแนวคิดเพื่อนำไปสู่การคำนวณหาดัชนีความสอดคล้องของข้อคำถามที่ให้คะแนนแบบพหุวิภาค (polytomous item) ดังนั้นเพื่อความเข้าใจดัชนีที่ถูกพัฒนาขึ้นในภายหลัง จึงขอนำเสนอรายละเอียดของดัชนีที่ใช้กับข้อคำถามที่ให้คะแนนแบบทวิวิภาค (dichotomous) พอสังเขป ดังตอนที่ 2

**ตอนที่ 2 แนวคิดและการคำนวณดัชนีความสอดคล้องของข้อคำถามที่ใช้กับโมเดลทฤษฎีการตอบสนองข้อสอบแบบทวิวิภาค (Dichotomous Item Response Theory Model)**

1) $\chi^2$ ของ Wright และ Panchapakesan (1969)

ดัชนี $\chi^2$ ของ Wright และ Panchapakesan (1969) พัฒนาขึ้นเพื่อทดสอบความสอดคล้องของข้อสอบหรือข้อคำถามภายใต้โมเดล 1 พารามิเตอร์และสามารถขยายไปสู่โมเดล 2 และ 3 พารามิเตอร์ได้ ซึ่งดัชนีนี้มีการแจกแจงแบบไคสแควร์ (Chi-square distribution) ที่มีองศาอิสระ (degree of freedom) เท่ากับ n-2 โดยมีสูตรการคำนวณดังนี้

$$\chi_j^2 = \sum_{i=1}^{n-1} y_{ij}^2$$

เมื่อ $y_{ij} = \{f_{ij} - \mathrm{E}(f_{ij})\}/\{var\ f_{ij}\}^{1/2}$

$f_{ij}$ คือ ความถี่ของผู้สอบที่ระดับความสามารถ (ability) ที่ $i$ ที่ตอบข้อสอบข้อที่ $j$ ได้ถูกต้อง ซึ่ง $f_{ij}$ มีการแจกแจงแบบทวินาม (binomial distribution) ซึ่งมีความน่าจะเป็นในการตอบถูก $\theta_i^*/(\theta_i^* + b_j^*)$ ในกรณีโมเดล 1 พารามิเตอร์

$n$ คือ จำนวนข้อสอบ/ข้อคำถาม ในแบบสอบ/แบบวัด

2) $Q_1$ ของ Yen (1981)

ดัชนี $Q_1$ ของ Yen (1981) เป็นดัชนีที่พัฒนาขึ้นโดยมีการแจกแจงแบบไคสแควร์ (Chi-square distribution) ที่มีองศาอิสระ (degree of freedom) เท่ากับ 10-m เมื่อ m คือ จำนวนพารามิเตอร์ของข้อสอบหรือข้อคำถามซึ่งในการคำนวณ $Q_1$ ค่าพารามิเตอร์ของข้อคำถามและค่าพารามิเตอร์ความสามารถของผู้สอบ ($\theta$) จะถูกพยากรณ์ภายใต้โมเดลทฤษฎีการตอบสนองข้อสอบที่เลือก (เช่น โมเดล 1, 2 หรือ 3 พารามิเตอร์) และมีการเรียงลำดับผู้สอบ โดยใช้ค่าพารามิเตอร์ความสามารถผู้สอบ ($\theta$) เป็นเกณฑ์ในการเรียงลำดับ และมีการแบ่งกลุ่มผู้สอบมีลักษณะเหมือนกัน ออกเป็น 10 กลุ่ม (k = 10) ซึ่งในแต่ละกลุ่มที่ k จะถูกกำหนดให้มีจำนวนผู้สอบอยู่ในกลุ่มเท่า ๆ กัน $E_{ik1}$ คำนวณโดยใช้ค่าเฉลี่ยของสัดส่วนที่คาดหวังของการตอบถูกในแต่ละกลุ่ม เนื่องจาก k = 10 ดังนั้นค่าองศาอิสระ (degree of freedom) ของ $Q_1$ คือ 10 – m เมื่อ m คือ จำนวนพารามิเตอร์ของข้อสอบหรือข้อคำถามที่ได้ประมาณขึ้น โดยมีสูตรการคำนวณดังนี้

$$Q_{ij} = \sum_{j=1}^{10} \sum_{z=0}^{1} N_k \frac{(O_{ikz} - E_{ikz})^2}{E_{ikz}}$$

$$= \sum_{k=1}^{10} N_k \frac{(O_{ik1} - E_{ik1})^2}{E_{ik1}(1 - E_{ik1})}$$

$z$ คือ คะแนนข้อสอบหรือคะแนนข้อคำถาม (item score)

$k$ คือ จำนวนกลุ่มผู้สอบที่มีลักษณะเหมือนกัน (homogeneous group of examination)

$N_k$ คือ จำนวนผู้สอบในกลุ่มที่ k (group k)

$O_{ik1} = 1 - O_{ik0}$ คือ สัดส่วนที่สังเกตได้ของการตอบถูกในกลุ่มที่ k

$E_{ik1} = 1 - E_{ik0}$ คือ สัดส่วนที่คาดหวังของการตอบถูกในกลุ่มที่ k

3) $\chi_B^2$ ของ Bock (1972)

ดัชนี $\chi_B^2$ ของ Bock (1972) เป็นดัชนีที่ใช้กับข้อคำถามที่มีการตรวจให้คะแนนแบบทวิวิภาค (dichotomous item) หรือข้อคำถามที่มีลักษณะเป็น Polytomous nominal item โดยมีการแจกแจงแบบไคสแควร์ (Chi-square distribution) ที่มีองศาอิสระ (degree of freedom) เท่ากับ G-m เมื่อ G คือ จำนวนกลุ่มผู้สอบที่ถูกแบ่ง ค่าองศาอิสระนี้นำไปใช้ประโยชน์ในการหา

ค่าวิกฤต (Critical value) ในการทดสอบสมมติฐานเพื่อบ่งชี้ความสอดคล้องของข้อคำถาม ค่า $\chi_B^2$ นี้สามารถคำนวณโดยใช้โปรแกรม BILOG ของ Mislevy และ Bock (1990) มีสูตรการคำนวณดังนี้

$$\chi_B^2 = \sum_{j=1}^{G} \frac{N_j(O_{ij}-E_{ij})}{E_{ij}(1-E_{ij})}$$

เมื่อ $N_j$ คือ จำนวนผู้สอบในแต่ละกลุ่มที่เกิดจากการแบ่ง $\theta$ ออกเป็นช่วง ๆ

$O_{ij}$ คือ สัดส่วนที่สังเกตได้บนข้อคำถามข้อที่ $i$ กลุ่มที่ $j$

$E_{ij}$ คือ สัดส่วนที่คาดหวังบนข้อคำถามข้อที่ $i$ กลุ่มที่ $j$ ซึ่งขึ้นอยู่กับการคำนวณ item response function (IRF) ขณะที่มีการประมาณค่า median $\theta$ ภายในช่วง $\theta$ ที่ถูกแบ่ง

$G$ คือ จำนวนกลุ่มผู้สอบที่ถูกแบ่ง

4) A likelihood ratio $G^2$ ของ McKinley และ Mills (1985)

ดัชนี A likelihood ratio $G^2$ ของ McKinley และ Mills (1985) มีการแจกแจงแบบไคสแควร์ (Chi-square distribution) ที่มีองศาอิสระ (degree of freedom) เท่ากับ 10-m มีวิธีการคำนวณที่คล้ายคลึงกับ $Q_1$ ของ Yen's (1981) เนื่องจากผู้สอบถูกจัดลำดับตามระดับความสามารถ($\theta$) และแบ่งออกเป็น 10 กลุ่มที่เท่าๆ กันแล้วจึงพิจารณาคำตอบที่ถูกหรือผิดในแต่ละกลุ่มเพื่อนำไปใช้ในการคำนวณ ดังสูตรการคำนวณ

$$G_i^2 = 2\sum_{k=1}^{10} N_k \left[O_{ik} \ln\left(\frac{O_{ik}}{E_{ik}}\right) + (1-O_{ik}) \ln\left(\frac{1-O_{ik}}{1-E_{ik}}\right)\right]$$

เมื่อ $O_{ik}$ คือ สัดส่วนที่สังเกตได้ (observed proportion) ซึ่งได้จากการแบ่งช่วง $\theta$ ออกเป็น 10 ส่วน

$E_{ik}$ คือ ค่าเฉลี่ยความน่าจะเป็นที่ตอบถูกในแต่ละช่วงของ $\theta$

$N_k$ คือ จำนวนผู้สอบในกลุ่มที่ k

5) The Standardized Residuals (SRs) ของ Hamberton, Swaminathan และ Roger (1991)

The Standardized Residuals (SRs) เป็นดัชนีอย่างหนึ่งที่ใช้บ่งชี้ความสอดคล้องของข้อคำถาม ข้อดีของดัชนีตัวนี้คือ มีความไว (sensitive) ต่อขนาดกลุ่มตัวอย่างน้อยกว่าดัชนีที่มีแนวคิดจากการทดสอบแบบไคสแควร์ ในการคำนวณดัชนีนี้จะแบ่งค่าระดับความสามารถ

ออกเป็นช่วงเท่า ๆ กัน 10 ส่วน แล้วคำนวณหาความแตกต่างระหว่างค่าที่แท้จริงกับค่าคาดหวังของผู้สอบในแต่ละช่วงระดับความสามารถ ซึ่งค่าความแตกต่างนี้เองเรียกว่า residual ดัชนี SRs ได้มาจากการหารค่าความแตกต่างด้วยค่าความคลาดเคลื่อนมาตรฐานของค่าคาดหวัง ดังในสูตรการคำนวณ

$$Z_{ij} = \frac{P_{ij} - \mathrm{E}(P_{ij})}{\sqrt{\mathrm{E}(P_{ij}) - [1 - \mathrm{E}(P_{ij})]/N_j}}$$

เมื่อ $P_{ij}$ คือ สัดส่วนที่สังเกตได้ของคำตอบถูกบนข้อสอบข้อที่ $i$ รายการคำตอบที่ $j$

$E(P_{ij})$ คือ สัดส่วนที่คาดหวังของคำตอบถูก

$N_j$ คือ จำนวนผู้สอบในกลุ่มที่เกิดจากการแบ่ง $\theta$ ออกเป็นกลุ่ม

เนื่องจาก The Standardized Residuals (SRs) มีการแจกแจงเข้าใกล้การแจงแจงปกติมาตรฐาน ดังนั้น ค่าวิกฤต (critical value) ที่ใช้ในการทดสอบสมมติฐาน เพื่อระบุความสอดคล้องของข้อคำถาม จึงเป็นค่าที่ได้จากตารางแจกแจงปกติมาตรฐาน โดยมีค่าวิกฤต (Critical value) ที่ใช้ในการทดสอบคือ 2 และ -2

6) A log – likelihood $\chi^2$ หรือ PARSCALE $G^2$ ของ Muraki และ Bock (1997)

ศึกษารายละเอียดได้ในตอนที่ 3 หัวข้อ PARSCALE $G^2$ ของ Muraki และ Bock (1997)

7) $S - \chi^2$ ของ Orlando และ Thissen (2000)

ดัชนี $S - \chi^2$ ของ Orlando และ Thissen (2000) มีองศาอิสระ $(\text{degree of freedom}) = \mathrm{I} - 1 - \mathrm{m}$ มีรูปแบบและวิธีการที่คล้ายคลึงกับกระบวนการของ $Q_1$ ของ Yen's (1981) แต่มีข้อดีมากกว่า $Q_1$ และ $G^2$ เนื่องจาก $Q_1$ และ $G^2$ มีกระบวนการจัดกลุ่ม (grouping procedure) ขึ้นอยู่บน model-dependent ability estimate ในขณะที่ $S - \chi^2$ มีกระบวนการจัดกลุ่ม (grouping procedure) ขึ้นอยู่บน test score เช่น จำนวนข้อสอบที่ตอบถูก ดัชนีนี้มีสูตรการคำนวณดังนี้

$$S - \chi^2 = \sum_{k=1}^{I-1} \sum_{z=0}^{1} N_k \frac{(O_{ikz} - E_{ikz})^2}{E_{ikz}}$$

$$= \sum_{k=1}^{I-1} N_k \frac{(O_{ikz} - E_{ik1})^2}{E_{ik1}(1 - E_{ik1})}$$

โดยที่ $$E_{ik1} = \frac{\int P_{i1}(\theta) f^{*i}(k-1/\theta)\phi(\theta)\partial\theta}{\int f\left(\frac{k}{\theta}\right)\phi(\theta)\partial(\theta)}$$

เมื่อ $E_{ik1}$ คือ สัดส่วนที่คาดหวังของผู้สอบในกลุ่ม $k$ ซึ่งตอบข้อสอบข้อที่ $i$ ถูกต้อง

$f(k/\theta)$ คือ การแจกแจงคะแนนสอบที่ $(x = k)$ ทำนายแบบมีเงื่อนไขเมื่อกำหนด $\theta$

$f^{*i}(k/\theta)$ คือ การแจกแจงของคะแนนสอบที่ทำนายแบบมีเงื่อนไขโดยปราศจากข้อสอบข้อที่ $i$

$\phi(\theta)$ คือ การแจกแจงของประชากร $\theta$ (population distribution of $\theta$)

ซึ่ง $f(k/\theta)$ และ $f^{*i}(k/\theta)$ สามารถคำนวณโดยใช้ recursive algorithm ซึ่งพัฒนาโดย Lord และ Wingersky (1984)

จากการศึกษาดัชนีทั้ง 7 ตัว พบว่าดัชนีทุกตัวยกเว้น $\chi^2$ ของ Wright and Panchapakesan ล้วนอาศัยแนวคิดจากดัชนี $Q_1$ ของ Yen's (1981) ซึ่งกระบวนการโดยทั่วไปก่อนจะนำหลักสูตรการคำนวณมาใช้คือ

1. ประมาณค่าความสามารถผู้สอบ ($\theta$) และค่าพารามิเตอร์ข้อสอบ (item parameter) จากข้อมูลการตอบข้อสอบ
2. เรียงลำดับผู้สอบโดยใช้ค่าความสามารถผู้สอบ ($\theta$) เป็นเกณฑ์ในการเรียงลำดับ
3. แบ่งผู้สอบออกเป็นกลุ่มย่อยตาม $\theta$ ที่จัดเรียงโดยพยายามให้แต่ละกลุ่ม $\theta$ ที่ถูกแบ่งมีจำนวนสมาชิกในกลุ่มเท่า ๆ กัน
4. คำนวณสัดส่วนของผู้สอบในแต่ละกลุ่มย่อย จำแนกตามการตอบถูกหรือผิด ในข้อสอบแต่ละข้อ
5. เปรียบเทียบ สัดส่วนที่สังเกตได้ (observed proportion) กับสัดส่วนที่คาดหวัง (expected proportion) ซึ่งได้มาจากการพยากรณ์โดยโมเดลที่เลือกใช้ การเปรียบเทียบนี้ใช้ดัชนีที่กล่าวมาแล้ว อาทิ $Q_1$ ของ Yen (1981), $\chi_B^2$ ของ Bock (1972) และ $G^2$ ของ McKinley and Mills (1985)

อย่างไรก็ตาม $\chi^2$ ของ Bock (1972) แตกต่างจาก $Q_1$ ของ Yen's (1981) ในเรื่องที่ $\chi^2$ ของ Bock's (1972) มีจำนวนของช่วงของ $\theta$ ที่ถูกแปรผันได้ ในขณะที่ $Q_1$ ถูกกำหนดให้คงที่เท่ากับ 10 นอกจากนี้ สัดส่วนที่คาดหวัง (expected proportion) ของ $\chi^2$ ของ Bock's (1972) ยังคำนวณโดยใช้มัธยฐาน (median) ซึ่งแตกต่างจาก $Q_1$ ที่ใช้ ค่าเฉลี่ย (mean) ส่วนวิธี Standardized Residuals (SRs) จะมีข้อดีกว่าวิธีอื่น ๆ ที่ใช้สถิติ $\chi^2$ เนื่องจากลดปัญหาความไว (sensitivity) กับขนาดกลุ่มตัวอย่าง (sample size) ลงไป Orlando และ Thissen (2000) ให้ข้อเสนอว่า $S-\chi^2$ นั้นมีข้อดีมากกว่า $Q_1$ และ $G^2$ เพราะทั้ง $Q_1$ และ $G^2$ มีกระบวนการจัดกลุ่ม (grouping procedure) บน model–dependent ability estimates ในขณะที่ $S-\chi^2$ มีกระบวนการจัดกลุ่มบนคะแนนสอบ (test score)

ดัชนีที่นำเสนอไปข้างต้น เป็นดัชนีความสอดคล้องของข้อคำถามกับโมเดลทฤษฎีการตอบสนองข้อสอบที่ให้คะแนนแบบทวิวิภาค (dichotomous item response theory model) ต่อไปนี้จะเป็นดัชนีความสอดคล้องของข้อคำถามที่ให้คะแนนแบบพหุวิภาค (polytomous item) ดังในตอนที่ 3

**ตอนที่ 3 แนวคิดและการคำนวณดัชนีความสอดคล้องของข้อคำถามที่ใช้กับโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค (Polytomous Item Response Theory Model)**

1) PARSCALE $G^2$ ของ Muraki และ Bock (1997)

PARSCALE $G^2$ เป็นดัชนีความสอดคล้องของข้อคำถามที่สามารถใช้ได้ทั้งข้อคำถามหรือข้อสอบที่ให้คะแนนแบบทวิวิภาคและแบบพหุวิภาคซึ่งประมาณขึ้นโดยใช้การทำหน้าที่ตอบสนองข้อคำถาม (item response function) ที่ค่าเฉลี่ยความสามารถของผู้สอบ (mean ability of examinees) ในช่วงหรือกลุ่มที่ k โดยใช้ expected a posteri ability estimated มีสูตรการคำนวณดังนี้

$$G_i^2 = 2\sum_{k=1}^{k_i}\sum_{z=0}^{z_i} r_{ikz} \ln \frac{r_{ikz}}{N_{ik}P_{iz}(\overline{\theta_k})}$$

เมื่อ $z$ คือ คะแนนของข้อสอบที่มีค่าตั้งแต่ 0 จนถึงคะแนนสูงสุดของข้อสอบนั้น คือ $z_i$

$r_{ikz}$ คือ จำนวนผู้สอบที่สังเกตได้ที่ได้คะแนน $z$ ในกลุ่มที่ $k$

$N_{ik}$ คือ จำนวนผู้สอบทั้งหมดในกลุ่มที่ $k$

$P_{iz}(\overline{\theta_k})$ คือ response function สำหรับคะแนน $z$ ที่ค่าเฉลี่ยความสามารถของผู้สอบในกลุ่มที่ $k$

ค่าองศาอิสระ (degree of freedom) ของ $G_i^2$ ในกรณีเป็นข้อสอบที่มีการตรวจให้คะแนนแบบทวิวิภาค (dichotomous item) จะมีค่าเท่ากับ $k_i$ (จำนวนกลุ่ม) ซึ่งมีความแตกต่างจาก $Q_1$ ของ Yen's (1981) ที่ไม่มีการปรับค่าสำหรับจำนวนพารามิเตอร์ (m) ที่ประมาณค่าเนื่องจาก Mislevy และ Bock (1990) มีความคิดเห็นว่า จำนวนพารามิเตอร์ (m) ไม่ควรถูกนำมาพิจารณาเกี่ยวกับองศาอิสระ (degree of freedom) ของ $G_i^2$

2) The Root Integrated Squared Error (RISE)

The Root Integrated Squared Error (RISE) เป็นวิธีการในการทดสอบความสอดคล้องของโมเดล ซึ่งมีแนวคิดมาจากการเปรียบเทียบ nonparametric item response function (IRF) กับ parametric IRF ซึ่งขั้นตอนในการคำนวณโดยวิธีนี้มีดังนี้

1. ประมาณ item $\boldsymbol{i}$'s IRF nonparametrically
2. ค้นหา best – fitting IRF สำหรับ parametric model ที่สนใจ
3. ทดสอบว่ามีความแตกต่างระหว่าง IRF ทั้งสองอย่างมีนัยสำคัญหรือไม่ ซึ่งคำนวณโดยใช้ RISE จากสูตร

$$RISE_i = \sqrt{\frac{\sum_{q=1}^{Q}(\hat{P}_q - \hat{P}_q non)^2}{Q}}$$

เมื่อ $\hat{P}_q non$ คือ จุดบน non – parametric item response function (IRF)

$\hat{P}_q$ คือ จุดบน parametric item response function (IRF)

$Q$ คือ จำนวนจุดที่ประเมิน (evaluation points) ที่เคยได้รับมาจาก kernel – smooth IRF

ระดับความมีนัยสำคัญของ RISE อาจประมาณโดย parametric bootstrapping procedure ซึ่งมีกระบวนการจำลองข้อมูล m ชุด ภายใต้เงื่อนไขที่ parametric model มีความสอดคล้อง (fit) การแจกแจงตัวอย่าง (sample distribution) ของ RISE ได้จากการคำนวณ RISE สำหรับแต่ละข้อมูลที่จำลอง ในการประมาณค่า p–value สำหรับข้อสอบข้อที่ $i$ ได้มาจากสัดส่วนของ RISE ของข้อมูลจำลองที่มากกว่า RISE ที่สังเกตได้ ซึ่ง Liang และ Wells (2009) กล่าวว่า RISE ให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) และค่าความคลาดเคลื่อนประเภทที่ 2 (Type II error) ในการทดสอบความสอดคล้องในระดับที่ยอมรับได้ ทั้งในกรณีข้อมูลแบบทวิวิภาค (dichotomous data) และข้อมูลแบบพหุวิภาค (polytomous data)

3) Generalized $S-\chi^2$ พัฒนาโดย Kang และ Chen (2008)

Generalized $S-\chi^2$ พัฒนาโดย Kang และ Chen (2008) เป็นดัชนีที่ได้ปรับปรุงมาจาก $S-\chi^2$ ของ Orlando และ Thissen (2000) เพื่อนำมาใช้กับข้อคำถามที่ให้คะแนนแบบพหุวิภาคที่มีจำนวนรายการคำตอบทั้งหมด $Z_i+1$ รายการคำตอบ ก็คือคะแนนรายการคำตอบ (category score) มีค่าตั้งแต่ 0 จนถึง $Z_i$ ซึ่งดัชนีนี้สามารถคำนวณโดยใช้สูตรดังนี้

$$Generalized\ S-\chi^2=\sum_{k=Z_i}^{F-Z_i}\sum_{z=0}^{Z_i}N_k\frac{(O_{ikz}-E_{ikz})^2}{E_{ikz}}$$

เมื่อ $Z_i$ คือ คะแนนสูงสุดของข้อคำถามข้อที่ $i$ ซึ่งมี $Z_i+1$ รายการคำตอบ

$F$ คือ คะแนนสอบที่สมบูรณ์แบบ เช่น $F=\sum_{i=1}^{n}Z_i$

$E_{ikz}$ คือ สัดส่วนรายการคำตอบที่คาดหวัง (Expected category proportion)

โดยที่ $$E_{ikz}=\frac{\int P_i(Z/\theta)f^{*i}(k-z/\theta)\phi(\theta)\partial\theta}{\int f(k/\theta)\phi(\theta)\partial(\theta)}$$

ซึ่ง $f(Z/\theta)$ และ $f^{*i}(Z/\theta)$ สามารถคำนวณโดยใช้ generalized recursive algorithm ที่พัฒนาโดย Thissen, Pommerich, Billeaud, และ Williums (1995)

ในการคำนวณ Generalized $S-\chi^2$ จะยกเว้นกลุ่มที่มีคะแนน 0 ($k=0$) และคะแนนที่สมบูรณ์แบบหรือคะแนนเต็ม ($k=F$) ดังนั้นผลรวมของ $k$ จึงดำเนินการไปถึงเพียง $F-Z_i$ สาเหตุที่เป็นเช่นนี้เพราะภายในกลุ่มที่มีคะแนนสูงหรือต่ำผิดปกติ (extremely low or high test score) ค่า $E_{ikz}$ สำหรับบางรายการคำตอบจะมีค่าเป็น 0 ในกรณี $S-\chi^2$ ในข้อคำถาม

แบบทวิวิภาค (dichotomous item) คะแนนกลุ่มข้างเคียง (neighboring test score group) ต้องยุบลงไปเพื่อรักษาความถี่ของ expected category ที่น้อยที่สุดให้เป็น 1 ($N_k E_{ikz} = 1$) อย่างไรก็ตาม วิธียุบนี้ไม่อาจเหมาะสมสำหรับข้อคำถามแบบพหุวิภาค (polytomous item) ดังนั้น ถ้าจำเป็นในการยุบกลุ่มของรายการคำตอบ ควรแน่ใจว่า $N_k E_{ikz} = 1$ ในขั้นตอนวิธียุบกลุ่ม (collapsing algorithm) ดัชนี Generalized $S - \chi^2$ มีองศาอิสระ $(degree\ of\ freedom) = K_i Z_i - m - C_i$ เมื่อ $m$ คือจำนวนพารามิเตอร์ของข้อคำถามที่ประมาณขึ้น และ $C_i$ คือ จำนวนทั้งหมดของรายการคำตอบที่ถูกยุบ

**ตอนที่ 4 การตรวจสอบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถาม (item fit index)**

การตรวจสอบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามกับโมเดลทฤษฎีการตอบสนองข้อสอบ คือการตรวจสอบว่าดัชนีความสอดคล้องของข้อคำถามนั้นมีประสิทธิภาพเพียงใดในการบ่งชี้ว่าข้อคำถามนั้นมีความสอดคล้อง (fit) หรือไม่สอดคล้อง (misfit) กับโมเดลทฤษฎีการตอบสนองข้อสอบที่ใช้ในการวิเคราะห์ โดยอาศัยแนวคิดพื้นฐานในการทดสอบสมมติฐานทางสถิติเข้าช่วย เพื่อทดสอบสมมติฐานเกี่ยวกับความสอดคล้องของข้อคำถามดังนี้

$\mathrm{H_0}$ : ข้อคำถามมีความสอดคล้องกับโมเดลทฤษฎีการตอบสนองข้อสอบ

$\mathrm{H_1}$ : ข้อคำถามไม่มีความสอดคล้องกับโมเดลทฤษฎีการตอบสนองข้อสอบ

การทดสอบสมมติฐานนี้จะใช้ดัชนีความสอดคล้องของข้อคำถามเป็นค่าสถิติทดสอบ ซึ่งเกณฑ์ที่ใช้ในการตัดสินว่าจะปฏิเสธหรือยอมรับสมมติฐานศูนย์ (null hypothesis) ที่ได้ตั้งไว้จะมี 2 แบบคือ

1. กรณีใช้ตารางสถิติ จะปฏิเสธสมมติฐานศูนย์ (null hypothesis) เมื่อค่าวิกฤต (critical value) ในการทดสอบมีค่ามากกว่าค่าสถิติที่อ่านค่าได้จากตารางสถิติ เช่น ถ้าในกรณีที่ดัชนีความสอดคล้องของข้อคำถามมีการแจกแจงแบบไคสแควร์ ค่าสถิติที่ใช้ทดสอบจะอ่านค่าได้จากตารางสถิติการแจกแจงแบบไคสแควร์ โดยจะปฏิเสธสมมติฐานศูนย์เมื่อค่าดัชนีความสอดคล้องของข้อคำถามที่คำนวณได้มีค่ามากกว่าค่าสถิติที่อ่านค่าได้จากตารางสถิติการแจกแจงแบบไคสแควร์ นั่นคือข้อคำถามข้อนั้นไม่มีความสอดคล้องกับโมเดลทฤษฎีการตอบสนองข้อสอบ

2. กรณีใช้ค่า p-value จะปฏิเสธสมมติฐานศูนย์ (null hypothesis) เมื่อค่า p-value ที่คำนวณได้มีค่าน้อยกว่าค่าระดับนัยสำคัญที่ใช้ในการทดสอบ ซึ่งโดยส่วนใหญ่ในการทดสอบทางสถิติค่าระดับนัยสำคัญที่นิยมใช้คือ .05

จากการทดสอบสมมติฐานดังกล่าวจึงเป็นที่มาของแนวคิดที่ใช้ในการตรวจสอบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถาม (item fit index) ซึ่งใช้เกณฑ์ในการบ่งชี้และเปรียบเทียบประสิทธิภาพคือ ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) และอำนาจการทดสอบ (Power of the test) โดยมีแนวคิดว่า คุณสมบัติของดัชนีความสอดคล้องของข้อคำถาม (item fit index) ที่ดีควรมีค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ที่น้อยและมีอำนาจการทดสอบ (Power of the test) ที่มาก เนื่องจากค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) เป็นค่าที่แสดงถึงโอกาสในการปฏิเสธสมมติฐานที่เป็นจริงซึ่งในการวิจัยนี้คือ การบ่งชี้ว่าข้อคำถามข้อนั้นไม่สอดคล้อง (misfit) ทั้ง ๆ ที่ข้อคำถามข้อนั้นมีความสอดคล้อง (fit) กับโมเดลทฤษฎีการตอบสนองข้อสอบ และอำนาจการทดสอบ (Power of the test) เป็นค่าที่แสดงถึงโอกาสในการปฏิเสธสมมติฐานที่เป็นเท็จซึ่งในการวิจัยนี้คือ การบ่งชี้ว่าข้อคำถามข้อนั้นไม่สอดคล้อง (misfit) ซึ่งในความเป็นจริงข้อคำถามข้อนั้นไม่มีความสอดคล้อง (misfit) กับโมเดลทฤษฎีการตอบสนองข้อสอบ

**ตอนที่ 5 โมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค (Polytomous item response model) ที่ใช้ในการวิจัย**

ในการวิจัยครั้งนี้เลือกใช้โมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค (Polytomous item response model) 2 โมเดลคือ Grade Response Model (GRM) และ Generalized Partial Credit Model (GPCM) เนื่องจากโมเดลทั้งสองพัฒนาบนพื้นฐานของโมเดลแบบ 2 พารามิเตอร์เหมือนกัน ซึ่งเหมาะสำหรับข้อสอบหรือข้อคำถามที่มีความยากและค่าอำนาจจำแนกที่แตกต่างกัน (ศิริชัย กาญจนวาสี, 2550) นอกจากนี้โมเดลทั้งสองยังเป็นโมเดลที่มีผู้สนใจศึกษากันมาก ไม่เข้มงวดเกี่ยวกับข้อตกลงเบื้องต้นและสามารถใช้กับแบบสอบและแบบวัดหลายลักษณะ (Donoghue, 1994; De Ayala, 1994; Muraki, 1992, 1993; Reise & Yu, 1990; Koch & De Ayala, 1989; Koch, 1983 อ้างถึงใน เอมอร จังศิริพรปกรณ์, 2545) ดังนั้นจึงขอนำเสนอแนวคิดและทฤษฎีที่เกี่ยวข้องกับโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค (Polytomous item response model) เฉพาะโมเดล GRM และ GPCM พอสังเขปดังนี้

1. Grade Response Model (GRM)

GRM พัฒนาโดย Samejima (1969; 1996 อ้างถึงในศิริชัย กาญจนวาสี, 2550; Embretson & Reise, 2000) สำหรับใช้กับแบบวัดหรือแบบสอบที่ข้อคำถามแต่ละข้อมีรายการคำตอบแบบมาตรเรียงอันดับ (order categorical response) ซึ่งแต่ละข้ออาจมีจำนวนรายการคำตอบแตกต่างกันได้ โดยมีการพัฒนาบนพื้นฐานของโมเดลของราสช์แบบ 2 พารามิเตอร์ ใช้หลักการคำนวณความน่าจะเป็นของการตอบแต่ละรายการคำตอบ 2 ขั้นตอน (indirect model) ในขั้นตอนแรกคำนวณค่าความชันร่วมของแต่ละข้อคำถาม แล้วจึงคำนวณค่าพารามิเตอร์ของแต่ละรายการคำตอบในแต่ละข้อคำถาม โมเดล GRM มักใช้กับแบบวัดที่เป็นมาตรประมาณค่า (Rating scale) ที่ไม่จำเป็นต้องมีรายการคำตอบเท่ากันทุกข้อ หรือแบบสอบที่มีการตรวจให้คะแนนความรู้บางส่วนที่มีจำนวนลำดับขั้นของการให้คะแนนแตกต่างกัน มีสูตรในการคำนวณ (ศิริชัย กาญจนวาสี, 2550) ดังนี้

$$P_{ix}^{*}(\theta) = \frac{\exp\left[\alpha_i\left(\theta - \beta_{ij}\right)\right]}{1 + \exp\left[\alpha_i\left(\theta - \beta_{ij}\right)\right]}$$

เมื่อ $x = j = 1,2, \ldots, m_i$

$P_{ix}^{*}(\theta)$ คือ ความน่าจะเป็นของผู้ตอบซึ่งมีคุณลักษณะระดับ $\theta$ จะตอบข้อคำถามที่ i ด้วยการเลือกรายการคำตอบที่ $x$ เมื่อ $x = 1,2, \ldots, m_i$

$\alpha_i$ คือ ค่าพารามิเตอร์ความชันร่วม (slope parameter) ของข้อคำถามที่ i

$\beta_{ij}$ คือ ค่าพารามิเตอร์ threshold ของแต่ละรายการคำตอบ (threshold parameter) ของข้อที่ i

2. Generalized Partial Credit Model (GPCM)

GPCM พัฒนาโดย Muraki (1992; 1993 อ้างถึงในศิริชัย กาญจนวาสี, 2550; Embretson & Reise, 2000) สำหรับใช้กับแบบสอบที่มีการตรวจให้คะแนนตามลำดับขั้นความสำเร็จเป็นคะแนนเรียงลำดับ 0, 1, 2, 3,… (Master, 1982 อ้างถึงในเอมอร จังศิริพรปกรณ์, 2545) หรือใช้กับแบบวัดเจตคติที่มีคะแนนเรียงลำดับหลายค่า (Dodd & Koch, 1982 อ้างถึงในเอมอร จังศิริพรปกรณ์, 2545) มีลักษณะเป็นโมเดลที่ยอมให้ข้อคำถามแต่ละข้อสามารถมี

ค่าพารามิเตอร์ความชันแตกต่างกันได้ และใช้หลักการคำนวณความน่าจะเป็นของการตอบแต่ละระดับขั้นการตอบโดยตรงแบบขั้นตอนเดียว (direct IRT method) มีสูตรในการคำนวณ (ศิริชัย กาญจนวาสี, 2550) ดังนี้

$$P_{ix}(\theta) = \frac{\exp\left[\sum_{j=0}^{x} \alpha_i(\theta - \delta_{ij})\right]}{\sum_{r=0}^{m}[exp \sum_{j=0}^{r}(\theta - \delta_{ij})]}$$

เมื่อ $\sum_{j=0}^{0} \alpha_i(\theta - \delta_{ij}) \equiv 0$

$P_{ix}(\theta)$ คือ ความน่าจะเป็นที่ผู้ตอบซึ่งมีคุณลักษณะ $\theta$ จะตอบข้อคำถามที่ i ด้วยการเลือกหรือสามารถทำรายการคำตอบขั้นที่ $x$ จากจำนวน $m_i$ ขั้น (step)

$\delta_{ij}$ คือ ค่าพารามิเตอร์ระดับความยากของขั้นการตอบที่ j ในข้อคำถามที่ i (item step difficulty) เมื่อ $j = 1,2, \ldots, m_i$

$\alpha_i$ คือ ค่าพารามิเตอร์ความชันของข้อคำถามที่ i

**ตอนที่ 6 เอกสารและงานวิจัยที่เกี่ยวข้อง**

แนวความคิดที่เกี่ยวกับการวิเคราะห์ข้อคำถามที่สอดคล้องกับโมเดลการตอบข้อสอบที่มีการศึกษากันมาตั้งแต่ ค.ศ. 1969 (Wright & Panchapakesan, 1969 อ้างถึงใน Hamberton & Swaminathan, 1996) โดยเริ่มจากการพิจารณาโมเดลการตอบข้อสอบแบบ 1 พารามิเตอร์ โดยใช้ดัชนีที่เป็นค่าสถิติที่มีการแจกแจงแบบไคสแควร์ และมีการพัฒนาอย่างต่อเนื่องมาเป็นลำดับ มีนักวิชาการในศาสตร์ด้านการวัดหลายท่านที่มีความสนใจและทำวิจัยในประเด็นนี้ ซึ่งผู้วิจัยได้ค้นคว้าและรวบรวมมานำเสนอโดยเรียงลำดับตามปีที่นักวิชาการแต่ละท่านได้ทำการวิจัย โดยมีจุดมุ่งหมายเพื่อให้เห็นพัฒนาการในการศึกษาเกี่ยวกับดัชนีความสอดคล้องของข้อคำถาม ดังนี้

Orlando และ Thissen (2000) ได้ศึกษาการตรวจสอบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามกับโมเดลทฤษฎีการตอบสนองข้อสอบ โดยเปรียบเทียบกันระหว่างดัชนี 4 ตัว คือ $S-\chi^2, S-G^2, Q_1-\chi^2$ และ $Q_1-G^2$ ซึ่งดัชนี $S-\chi^2$ เป็นดัชนีที่ Orlando และ Thissen ได้คิดค้นขึ้น ซึ่งดัชนี $Q_1-G^2$ คือ ค่า Pearson $\chi^2$ และ $Q_1-\chi^2$ คือ ค่า Likelihood ratio $G^2$ ของ McKinley และ Mills (1985) ในการวิจัยใช้การจำลองแบบ (Simulation) จำลอง

ข้อมูลการตอบเป็นโมเดล 3 พารามิเตอร์ โดยกำหนดค่าพารามิเตอร์ทั้ง 3 ตัว ให้มีความคล้ายคลึงกับค่าพารามิเตอร์ที่พบในข้อมูลจริงในแบบทดสอบทางการศึกษา ใช้ความยาวแบบสอบ 10, 40 และ 80 ข้อตามลำดับ มีการกระทำซ้ำ 100 ครั้ง เกณฑ์ในการพิจารณาคือ ใช้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) และอำนาจการทดสอบ (Power of the test) ผลที่ได้จากการศึกษาคือ $S-\chi^2$ มีค่า Type I error ที่ค่อนข้างคงเส้นคงวา ไม่แปรเปลี่ยนไปตามความยาวของแบบสอบ และพบว่า $S-\chi^2$ มีอำนาจในการตรวจสอบความไม่สอดคล้อง (misfit) ได้ดีเมื่อเทียบกับดัชนีตัวอื่น ๆ ที่ใช้ในการศึกษา

ต่อมาในปี ค.ศ. 2003 Orlando และ Thissen ได้ทำการศึกษาเพื่อขยายองค์ความรู้ในการประเมินในระยะแรกของวิธีการใช้คะแนนรวม (summed score) ในการสร้างดัชนีความสอดคล้องของข้อคำถาม ที่ใช้กับข้อสอบที่มีการตรวจให้คะแนนเพียง 2 ค่า ซึ่งวิธีการใช้คะแนนรวมส่งผลให้เกิดค่า Type I error ที่มากใน Likelihood ratio $G^2$ ดังนั้นการศึกษาครั้งนี้จึงใช้ดัชนีเพียง 2 ตัวคือ $S-\chi^2$ และ $Q_1-\chi^2$ หรือ Pearson $\chi^2$ จาก Yen's $Q_1$ ในการจำลองข้อมูลใช้ความยาวแบบสอบ 10, 40, และ 80 ข้อ จำนวนผู้สอบ 500, 1000, และ 2000 คน มีการใช้ Receiver Operation Characterictic (ROC) ช่วยในการประเมินประสิทธิภาพของดัชนีทั้ง 2 ตัว ซึ่ง ROC เป็นกราฟแสดงความสัมพันธ์ระหว่าง อัตราการทำนายถูกและอัตราการทำนายผิด ซึ่งรูปร่างของกราฟ ROC สามารถตรวจสอบเพื่อประเมินอำนาจการทดสอบ (Power of the test) และ ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนีทั้ง 2 ได้ ซึ่งผลการศึกษาบ่งชี้ว่าประสิทธิภาพของ $S-\chi^2$ ดีกว่า $Q_1-\chi^2$ ในหลายเงื่อนไขที่ใช้ในการศึกษา ซึ่งสอดคล้องกับผลการศึกษาในปี ค.ศ.2000 ของ Orlando และ Thissen ที่กล่าวว่า $S-\chi^2$ เป็นเครื่องมืออย่างหนึ่งที่เป็นประโยชน์ในการตรวจสอบความไม่สอดคล้อง (misfit) ได้ดี

Stone และ Zhang (2003) ได้ทำการศึกษาเปรียบเทียบ 4 วิธีการที่ใช้ในการประเมินความสอดคล้องของโมเดลการตอบข้อสอบที่มีการตรวจให้คะแนนเพียง 2 ค่า ใช้เทคนิคการจำลองแบบ (simulation) จำลองข้อมูลในการตอบ โดยให้ค่าพารามิเตอร์ความสามารถของผู้สอบมีการแจงแจงแบบปกติที่มีค่าเฉลี่ยเป็น 0 และส่วนเบี่ยงเบนมาตรฐานเป็น 1 ความยาวแบบสอบที่ใช้คือ 10, 20, และ 40 ข้อ ตามลำดับจำนวนผู้สอบ 500, 1000, และ 2000 คน ตามลำดับ เกณฑ์ที่ใช้ในการพิจารณาเพื่อตัดสินว่าวิธีใดดีที่สุด คือค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) และอำนาจการทดสอบ (Power of the test) ผลการศึกษาให้ผลว่าวิธีดั้งเดิม (Traditional Procedure) คือวิธีการของ Donoghue และ Hombo (2001) สามารถใช้ได้กับแบบสอบความยาว

10 ข้อ ส่วนในกรณีความยาว 20 และ 40 ข้อ ไม่สามารถนำมาใช้ได้ ต้องใช้วิธีของ Bock (1972) แทน ซึ่งวิธีของ Bock จะให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ที่มากเมื่อเปรียบเทียบกับวิธีทางเลือก (Alternative Procedure) ซึ่งประกอบด้วยวิธีของ Orlando และ Thissen (2000) และวิธีของ Stone (2000) เมื่อใช้เกณฑ์อำนาจการทดสอบ (Power of the test) พบว่า ทั้ง 3 วิธีให้ผลสอดคล้องกัน คือ อำนาจการทดสอบจะเพิ่มมากขึ้น เมื่อจำนวนตัวอย่าง (sample size) เพิ่มขึ้น และ อำนาจการทดสอบจะลดลงเมื่อค่า $\alpha$ หรือค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ลดลง วิธีที่มีอำนาจในการตรวจสอบความไม่สอดคล้อง (misfit) มากที่สุดคือวิธีของ Stone (2000) อย่างไรก็ตามวิธีของ Orlando และ Thissen มีอำนาจเพียงพอ (adequate power) ในการตรวจสอบความไม่สอดคล้องของข้อคำถาม ที่ขนาดกลุ่มตัวอย่างขนาดใหญ่ (2000 คน) ในขณะที่วิธีของ Stone (2000) จะมีอำนาจที่เพียงพอในกลุ่มตัวอย่างซึ่งมีขนาดเล็ก

Dodeen (2004) ให้ความสนใจความสัมพันธ์ระหว่างค่าพารามิเตอร์ของข้อสอบกับดัชนีความสอดคล้องของข้อคำถาม (item fit index) เขาจึงได้ศึกษาอิทธิพลของอำนาจจำแนก ความยากและค่าการเดา ที่มีต่อดัชนีความสอดคล้องของข้อคำถาม (item fit index) 2 ตัว คือ $\chi_B^2$ ของ Bock (1972) และ the Standardized Residuals (SRs) โดยพิจารณาจากค่าสหสัมพันธ์ของเพียร์สัน (Pearson Correlation) และศึกษาอิทธิพลของแต่ละระดับของพารามิเตอร์ (อำนาจจำแนก ความยากและค่าการเดา) บนค่าเฉลี่ยของดัชนีความสอดคล้องของข้อคำถาม (item fit index) โดยใช้การวิเคราะห์ความแปรปรวน (ANOVA) ตามด้วยการเปรียบเทียบรายคู่ (pairwise comparison) ในการศึกษาได้มีการจำลองข้อมูลการตอบเป็นแบบทวิวิภาค (dichotomous item) ความยาวแบบสอบ 50 ข้อ จำนวนผู้สอบ 1000 คน ค่าพารามิเตอร์ความสามารถผู้สอบมีการแจกแจงปกติที่มีค่าเฉลี่ยเป็น 0 และส่วนเบี่ยงเบนมาตรฐาน 1 มีการทำซ้ำ 100 ครั้ง ใน 9 สถานการณ์ มีการพิจารณาจำนวนและสัดส่วนของข้อสอบที่มีลักษณะไม่สอดคล้อง (misfit) ที่ระดับความเชื่อมั่น 99% ผลการศึกษาแสดงให้เห็นว่าค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error)ของ $\chi_B^2$ มากกว่า SRs ทั้งนี้เนื่องจากดัชนี $\chi_B^2$ มีความไว (sensitivity) ต่อขนาดกลุ่มตัวอย่างใหญ่ และมีความสัมพันธ์เชิงบวกระหว่างดัชนีความสอดคล้องของข้อคำถาม (item fit index) กับค่าอำนาจจำแนกและค่าการเดา แต่ไม่มีความสัมพันธ์กับค่าความยาก นอกจากนี้เมื่อค่าพารามิเตอร์ของข้อสอบทั้งอำนาจจำแนก ค่าความยาก และค่าการเดาเพิ่มขึ้น จำนวนข้อสอบที่ไม่สอดคล้อง (misfit) จะมีค่าเพิ่มขึ้นด้วย ทั้งในกรณีของค่าดัชนี $\chi_B^2$ และ Standardized Residuals (SRs)

ดังนั้นในการใช้ดัชนี item fit จึงควรใช้อย่างระมัดระวังเมื่อข้อสอบมีค่าอำนาจจำแนกสูง และพบว่ามีการตอบข้อสอบด้วยการเดา

ในปี ค.ศ. 2005 DeMars มีความสนใจในประสิทธิภาพของดัชนี PARSCALE $G^2$ ที่ใช้ทดสอบเพื่อชี้วัดความสอดคล้องของข้อคำถาม เนื่องจากดัชนีตัวนี้มีอยู่ในโปรแกรม PARSCALE ซึ่งใช้กันอย่างแพร่หลาย และสามารถวิเคราะห์ได้ทั้งข้อสอบที่มีการตรวจให้คะแนนแบบทวิวิภาคและแบบพหุวิภาค แต่ในการศึกษาทดสอบกับข้อคำถามที่มีการตรวจให้คะแนนแบบพหุวิภาค (polytomous item) โดยใช้การจำลองข้อมูลให้ค่าพารามิเตอร์ความสามารถของผู้ตอบมีการแจกแจงแบบปกติ (normal distribution) และการแจกแจงแบบยูนิฟอร์ม (uniform distribution) ขนาดกลุ่มตัวอย่าง 1000 คน แบบวัดความยาว 10 และ 20 ข้อ ทำซ้ำ 100 ครั้ง โดยกำหนดให้แต่ละสถานการณ์ที่จำลองขึ้นมีแบบวัดที่แตกต่างกัน 100 ฉบับ โมเดลทฤษฎีการตอบสนองข้อสอบที่เลือกใช้มี 2 โมเดลคือ Partial Credit Model (PCM) และ Grade Response Model (GRM) ข้อคำถามในแต่ละข้อมี 5 สเกลคำตอบ (category) เกณฑ์ที่ใช้ในการพิจารณาคือสัดส่วนของข้อตอบที่ถูกบ่งชี้ว่าไม่สอดคล้อง (misfit) นั่นคือ ค่าความคลาดเคลื่อนประเภทที่ 1มีค่าเท่ากับ 0.05 ($\alpha = 0.05$) ซึ่งกรณีค่าพารามิเตอร์ความสามารถผู้ตอบมีการแจกแจงแบบปกติ ค่า $\alpha$ ที่พบจะมีค่าเข้าใกล้ 0.05 ในทั้งโมเดล PCM และ GRM ในขณะที่เมื่อพารามิเตอร์ความสามารถของผู้ตอบมีการแจกแจงแบบยูนิฟอร์ม (uniform) ค่า $\alpha$ จะมีค่าค่อนข้างมากโดยเฉพาะเมื่อความยาวของแบบวัดเป็น 10 ข้อ นอกจากนี้เมื่อพิจารณาสัดส่วนของข้อคำถามที่ถูกบ่งชี้ว่า misfit จำแนกตามองศาอิสระ (degree of freedom) พบว่า กรณีโมเดล PCM เมื่อความยาวแบบวัด 20 ข้อค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ค่อนข้างคงที่ และเข้าใกล้ 0.05 ในสถานการณ์อื่น ๆ ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ลดลงเมื่อค่าองศาอิสระ (degree of freedom) เพิ่มขึ้น ดังนั้นจากการวิจัยนี้ของ DeMars สรุปได้ว่า ดัชนี PARSCALE $G^2$ ไม่เหมาะสมที่จะนำไปใช้กับแบบวัดที่สั้น แต่เหมาะที่จะใช้กับแบบวัดที่ยาว

Kang และ Chen (2008) ได้ทำการศึกษาประสิทธิภาพของดัชนี Generalized $S - \chi^2$ ซึ่งเป็นดัชนีที่ได้ทำการปรับปรุงมาจากดัชนี $S - \chi^2$ เพื่อให้สามารถนำมาใช้กับข้อคำถามที่เป็น Polytomous ได้ โดยทำการเปรียบเทียบกับดัชนี PARSCALE $G^2$ เกณฑ์ที่นำมาใช้ในการเปรียบเทียบคือ ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) และค่าอำนาจการทดสอบ (Power of the test) โมเดลที่ใช้ในการศึกษา มี 3 โมเดล คือ Generalized Partial Credit Model (GPCM), Partial Credit Model (PCM) และ Rating Scale Model (RSM) ทำการจำลองข้อมูล

การตอบข้อคำถามโดยมีความยาวแบบวัด 5, 10, 20 ข้อ ขนาดกลุ่มตัวอย่าง 500, 1000, 2000 และ 5000 คนตามลำดับ ในแต่ละสถานการณ์ทั้งหมด 36 สถานการณ์ที่ได้จำลองขึ้น มีการทำซ้ำ 100 ครั้ง การศึกษาค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) และอำนาจการทดสอบ (Power of the test) จะพิจารณาโดยใช้วิธีแบ่งโมเดลเป็น 2 ประเภท คือ โมเดลจำลอง (Generating Model : GM) และโมเดลเทียบมาตรฐาน (Calibrating Model : CM) ซึ่งจำนวนของพารามิเตอร์ข้อคำถาม (Item parameter) ของ CM จะน้อยกว่าหรือเท่ากับ GM เสมอ ถ้า GM กับ CM เป็นโมเดลที่ตรงกัน (เช่นเป็น RSM เหมือนกัน) จะประเมินประสิทธิภาพโดยดูที่ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) แต่ถ้า CM ง่ายกว่า GM จะใช้ในการคำนวณค่าอำนาจการทดสอบ (Power of the test) ซึ่งข้อคำถามจะมีลักษณะไม่สอดคล้อง (misfit) เมื่อค่า P – value ที่สังเกตได้น้อยกว่า 0.05 ผลการวิเคราะห์นำไปสู่ข้อสรุปได้ว่า PARSCALE $G^2$ ไม่ค่อยมีประสิทธิภาพในการตรวจสอบความสอดคล้องของข้อสอบในสถานการณ์ทดลองส่วนใหญ่ ทั้งเงื่อนไขความยาวแบบวัดที่สั้น และปานกลาง หรือในกรณีกลุ่มตัวอย่างขนาดใหญ่ ในขณะที่ Generalized $S-\chi^2$ มีประสิทธิภาพมาก โดยเฉพาะในกรณีความยาวแบบวัด 5 หรือ 10 ข้อ ซึ่งคล้ายกับที่พบ ในขนาดกลุ่มตัวอย่างที่มากโดยไม่คำนึงถึงความยาวแบบวัดดังนั้น Generalized $S-\chi^2$ จึงเป็นดัชนีที่เหมาะสมในการบ่งชี้ความสอดคล้องของข้อคำถามกับโมเดลทฤษฎีการตอบสนองข้อสอบที่ให้คะแนนแบบพหุวิภาค (Polytomous item response theory model)

Liang และ Wells (2009) สนใจศึกษาดัชนีความสอดคล้องของข้อคำถามกับ Generalized Partial Credit Model (GPCM) โดยดัชนีที่เขาสนใจคือ $S-\chi^2$, PARSCALE $G^2$ และ Root Integrated Square Error (RISE) ซึ่ง RISE เป็นความแตกต่างระหว่าง 2 item response function เป็นวิธี non–parametrice ที่ใช้ในการประเมินความสอดคล้องของโมเดล ในการศึกษาใช้การจำลองข้อมูลโดยแบ่งเป็น 2 ครั้ง ครั้งแรกข้อมูลจำลองภายใต้เงื่อนไขความยาวแบบสอบ 10, 20, และ40 ข้อ ขนาดกลุ่มตัวอย่าง 500, 1000, และ 2000 คน ค่าพารามิเตอร์ความสามารถของผู้สอบมีการแจกแจงแบบปกติ การวิเคราะห์ค่า ความคลาดเคลื่อนประเภทที่ 1 (Type I error) ขึ้นอยู่กับสัดส่วนของข้อสอบที่ถูกจำลองว่าเป็น GPCM ที่ถูกตรวจสอบโดยให้ผลว่า misfit ที่ $\alpha = 0.05$ ค่าอำนาจการทดสอบ (power of the test) ขึ้นอยู่กับสัดส่วนของข้อสอบที่ถูกจำลองโดยโค้งรายการคำตอบที่ได้จากการสังเกต (empirically derived category curve) ครั้งที่ 2 ศึกษาเพื่อตรวจสอบประสิทธิภาพของ RISE ในบริบทความเป็นจริงของการทดสอบมาตรฐานขนาดใหญ่ (large – scale standardized test) ข้อมูลจำลองเป็นแบบสอบ 40 ข้อ ประกอบด้วยข้อสอบหลายตัวเลือก 30 ข้อ ตอบสั้น 5 ข้อเป็นโมเดลโลจิสติกแบบ 2 พารามิเตอร์ และ 5 ข้อเป็น

GPCM ที่มีลักษณะข้อสอบแบบตรวจให้คะแนนบางส่วนมี 5 รายการคำตอบ ผลการศึกษา 2 ครั้ง ให้ผลที่คล้ายกันคือ RISE และ $S-\chi^2$ จะให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ที่น้อยกว่า เมื่อเทียบกับ PARSCALE $G^2$ และเมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้น ดัชนีทั้ง 3 ตัวก็จะให้อำนาจการทดสอบ (Power of the test) ที่มากขึ้นด้วยเช่นกัน แต่ RISE จะให้ค่าอำนาจการทดสอบ (Power of the test) สูงที่สุด อย่างไรก็ตามการนำค่าอำนาจการทดสอบ (Power of the test) มาพิจารณาเพียงอย่างเดียวอาจไม่สามารถเชื่อถือได้ ควรพิจารณาที่ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ด้วยบทสรุปจากการศึกษานี้ คือ RISE มีข้อได้เปรียบวิธีอื่น ๆ คือ จะให้ Graphical model misfit ซึ่งจะช่วยบอกตำแหน่งและประเภทของ misfit ได้ ซึ่งก็จะเป็นประโยชน์หากนำไปใช้ในทางปฏิบัติกับข้อมูลจริง

Lattuis, Clark และ O'Brien (2009) ได้ตรวจสอบประสิทธิภาพของวิธีทางเลือก (Alternative Procedure) ในการตรวจสอบเพื่อบ่งชี้ความสอดคล้องของข้อคำถาม ซึ่งประกอบด้วยดัชนี 3 ตัวคือ $S-\chi^2$ ของ Orlando และ Thissen (2000) ซึ่งถูกปรับเป็น Generalized $S-\chi^2$ โดย Kang และ Chen (2008), $\chi^{2*}$ ของ Stone (2000) และ Adjusted $\chi^2$ to degree of freedom ratio ($\chi^2/dfs$) ของ Drasgow, Levine, Tsien, Williams และ Mead (1995) โดยใช้ Grade Response Model (GRM) ในการวิเคราะห์โดยการจำลองสถานการณ์ที่ผันแปรไปตามบริบทของจำนวนข้อคำถามและขนาดกลุ่มตัวอย่าง ซึ่งแบบวัดยาว 10 และ 20 ข้อ ขนาดกลุ่มตัวอย่าง 500, 1000, และ 2000 คน ค่าพารามิเตอร์ความสามารถผู้สอบถูกเลือกอย่างสุ่มจากการแจกแจงแบบปกติมาตรฐาน การจำลองข้อมูลที่มีลักษณะไม่สอดคล้อง (misfit) จำลองโดยใช้โมเดล GGUM เกณฑ์ที่ใช้ในการตัดสินใจคือ ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) และอำนาจการทดสอบ (Power of the test) โดยก่อนที่จะมีการตรวจสอบดัชนีความสอดคล้องของข้อคำถาม (item fit index) ได้มีการตรวจสอบความถูกต้องของ recovering the population parameter โดยใช้ root mean square error (RMSE) ซึ่งเป็นการเปรียบเทียบค่าประมาณ (estimated value) กับค่าประชากร (population value) ของพารามิเตอร์ข้อคำถาม เพื่อเป็นการยืนยันว่าค่าพารามิเตอร์ข้อคำถามที่จำลองขึ้นมีความถูกต้องสมเหตุสมผล ผลการศึกษาวิจัยพบว่า เมื่อโมเดลการตอบเป็น Grade Response Model (GRM) ทั้ง $S-\chi^2$ , $\chi^{2*}$ และ Adjusted $\chi^2/dfs$ ที่ไม่ได้ทำการทดสอบข้ามกลุ่ม (cross validation) ให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ที่ค่อนข้างมาก เมื่อพิจารณาอำนาจการทดสอบ (Power of the test) กรณีมีเงื่อนไขความชัน (Slope) ลดลง 0.25 มี $S-\chi^2$ มีอำนาจการทดสอบที่เพียงพอ (adequate power) กรณีเดียวเมื่อขนาดกลุ่มตัวอย่าง 2000 คน และความยาวแบบวัด 10 ข้อ ส่วนในกรณีที่

ค่าความชันลดลงมาก $S-\chi^2$ ให้อำนาจการทดสอบที่เพียงพอในทุกเงื่อนไขขนาดกลุ่มตัวอย่าง และความยาวแบบสอบ $\chi^{2*}$ ให้อำนาจการทดสอบ (Power of the test) ในระดับที่ยอมรับได้ ทั้ง 2 เงื่อนไขความชัน ในทุกขนาดกลุ่มตัวอย่างและความยาวแบบวัด ในกรณี last threshold ลดลง 0.25 ดัชนี $S-\chi^2$ ให้อำนาจการทดสอบ (power of the test) ที่ยอมรับได้ กรณี แบบวัดยาว 10 ข้อ ขนาดกลุ่มตัวอย่าง 1000 และ 2000 คน ดัชนี $\chi^{2*}$ ให้อำนาจการทดสอบ (Power of the test) ที่ยอมรับได้ เมื่อขนาดกลุ่มตัวอย่าง อย่างน้อย 1000 คน เช่นเดียวกับ Adjusted $\chi^2/dfs$ กรณี item single ในกรณี GGUM generated ดัชนี $S-\chi^2$ ให้อำนาจการทดสอบ (Power of the test) ที่ยอมรับได้เมื่อขนาดกลุ่มตัวอย่างอย่างน้อย 1000 คน ในขณะที่ดัชนี $\chi^2/dfs$ กรณี item single ไม่ได้ให้อำนาจการทดสอบ (Power of the test) เป็นที่ยอมรับ แสดงให้เห็นถึงอิทธิพลของขนาดกลุ่มตัวอย่างที่ยังคงส่งผลต่ออำนาจการทดสอบ (Power of the test) จากผลการวิจัยให้ข้อสรุปว่า $S-\chi^2$ และ $\chi^{2*}$ให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ต่ำกว่าที่คาดไว้ ส่วนดัชนี Adjusted $\chi^2/dfs$ มีค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ที่มากเมื่อใช้กับการทดสอบข้ามกลุ่ม และมีค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ต่ำมากเมื่อไม่มีการทดสอบข้ามกลุ่ม นอกจากนี้เมื่อไม่มีการทดสอบข้ามกลุ่ม Adjusted $\chi^2/dfs$ จะมีอำนาจการทดสอบสูงสุดในทุกกรณี

ในปี ค.ศ. 2010 Kang และ Chen ได้ทำการศึกษาประสิทธิภาพของดัชนี Generalized $S-\chi^2$ สำหรับ Grade response model (GRM) โดยมีวัตถุประสงค์ที่จะพัฒนาดัชนีที่ใช้กับข้อคำถามที่ให้คะแนนแบบพหุวิภาค (Polytomous item) ที่สามารถควบคุมค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ในระดับที่พอเพียงและมีอำนาจการทดสอบ (Power of the test) ที่เหมาะสมในการตรวจสอบข้อคำถามที่ไม่สอดคล้อง (misfit item) ซึ่งในการศึกษาวิจัยใช้เทคนิคการจำลองแบบ (simulation) จำลองข้อมูลโดยมีความยาวแบบวัด ขนาดสั้นคือ 5 ข้อ ความยาวปานกลาง10 ข้อ และความยาวมาก 20 ข้อ ขนาดกลุ่มตัวอย่างเล็ก 500 คน ขนาดปานกลาง 1000 คน และขนาดใหญ่ 2000 คน จำนวนรายการคำตอบ 2 ขนาด คือ 3 และ 5 ค่าพารามิเตอร์ความสามารถผู้สอบมีการแจกแจงแบบปกติ (Normal distribution) และการแจกแจงแบบเอกรูป (Uniform distribution) ดังนั้นสถานการณ์จำลองที่ใช้ในการวิจัยมีทั้งหมด 36 สถานการณ์ (ความยาวแบบสอบ 3 ขนาด x ขนาดกลุ่มตัวอย่าง 3 ขนาด x จำนวนรายการคำตอบ 2 ขนาด x การแจกแจงของค่าพารามิเตอร์ความสามารถผู้สอบ 2 แบบ) ในแต่ละสถานการณ์มีการกระทำซ้ำ 100 ครั้ง เกณฑ์ที่ใช้ในการพิจารณาคือค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) และอำนาจการทดสอบ (Power of the test) ซึ่งวิธีการที่ใช้ในการศึกษาจะมีความแตกต่างจาก

งานวิจัยอื่นๆที่ผ่านมา ตรงที่ในการศึกษาโดยใช้อำนาจการทดสอบเป็นเกณฑ์ในการพิจารณาได้มีการจำลองข้อมูลในลักษณะของข้อคำถามที่ไม่ดี (bad item) 2 ประเภทคือ M-type misfit ซึ่งก็คือ multidimentionality และ D-type misfit ซึ่งเป็นความแตกต่างระหว่างโค้งรายการคำตอบ (curve discrepancy) การจำลองกรณี M-type misfit จะใช้ two-dimensional GRM กรณี D-type misfit จะใช้ Guttman step function นอกจากนี้ ในการศึกษาเกี่ยวกับอำนาจการทดสอบจะใช้จำนวนรายการคำตอบเป็น 5 เท่านั้นและค่าพารามิเตอร์ความสามารถผู้สอบจะใช้เฉพาะการแจกแจงแบบปกติมาตรฐาน ดังนั้นจำนวนสถานการณ์ที่จำลองขึ้นจึงมีเพียง 18 สถานการณ์ (ความยาวแบบสอบ 3 ขนาด x ขนาดกลุ่มตัวอย่าง 3 ขนาด x ประเภทของ misfit 2 ประเภท) ผลการวิจัยพบว่า ทั้งในกรณีค่าพารามิเตอร์ความสามารถผู้สอบมีการแจกแจงแบบปกติและการแจกแจงแบบยูนิฟอร์ม ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) จะลดลงเมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้น ย่อมแสดงถึงประสิทธิภาพที่เพิ่มขึ้นของดัชนี Generalized $S-\chi^2$ เมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้น แต่ถ้าพิจารณาจากจำนวนรายการคำตอบที่แตกต่างกันจะพบว่าโดยส่วนใหญ่เมื่อมี 5 รายการคำตอบค่าความคลาดเคลื่อนประเภทที่ 1จะมากกว่าเมื่อมี 3 รายการคำตอบ แต่โดยส่วนใหญ่แล้วดัชนี Generalized $S-\chi^2$ จะสามารถควบคุมค่าความคลาดเคลื่อนประเภทที่ 1ได้ในระดับที่พอเพียงเนื่องจากมีเพียง 13 สถานการณ์จากทั้งหมด 36 สถานการณ์ที่ค่าความคลาดเคลื่อนประเภทที่ 1 อยู่นอกเหนือจากช่วงความเชื่อมั่น 95% โดยเฉพาะในกรณีแบบวัด 20 ข้อที่มี 5 รายการคำตอบ ขนาดกลุ่มตัวอย่าง 500 คน ที่ค่าความคลาดเคลื่อนประเภทที่ 1มีค่าสูงค่อนข้างมากอย่างชัดเจน เมื่อพิจารณาจากเกณฑ์อำนาจการทดสอบภายใต้ M-type misfit ดัชนี Generalized $S-\chi^2$ ไม่มีความไว (insensitive) ในการตรวจสอบความไม่สอดคล้องของข้อคำถามเมื่อขนาดกลุ่มตัวอย่างใหญ่ ซึ่งจากสิ่งที่พบบ่งชี้ว่าดัชนี Generalized $S-\chi^2$ อาจจะไม่เหมาะสมนักที่จะนำมาใช้เป็นเครื่องมือในการประเมินความสอดคล้องของข้อคำถามเมื่อเกิด M-type misfit ขณะที่ภายใต้ข้อคำถามที่มีลักษณะ D-type misfit จะมีอำนาจการทดสอบที่เหมาะสมในสถานการณ์ที่แบบวัดยาว 20 ข้อ ขนาดกลุ่มตัวอย่างอย่างน้อย 1000 คน ซึ่งแสดงให้เห็นว่าค่าอำนาจการทดสอบนั้นเป็นฟังก์ชันของขนาดกลุ่มตัวอย่างนั่นคืออำนาจการทดสอบมีความสัมพันธ์กับขนาดกลุ่มตัวอย่าง จากผลวิจัยทั้งหมดสรุปว่า ภายใต้โมเดล GRM ดัชนี Generalized $S-\chi^2$ ให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) และอำนาจการทดสอบ (Power of the test) ที่เหมาะสมในการตรวจสอบความไม่สอดคล้อง (misfit) ทั้งสองประเภทของข้อคำถามในขนาดกลุ่มตัวอย่างที่ใหญ่ เพราะฉะนั้น Generalized $S-\chi^2$ จึงเป็น

ดัชนีที่มีแนวโน้มที่ดีในการนำไปใช้บ่งชี้ความสอดคล้องของข้อคำถามที่มีการตรวจให้คะแนนมากกว่า 2 ค่าในการประเมินทางการศึกษาและจิตวิทยา

จากการศึกษางานวิจัยทั้งหมดที่กล่าวมา สามารถสรุปสาระสำคัญที่เกี่ยวข้องกับดัชนีความสอดคล้องของข้อคำถาม (item fit index) เพื่อนำสารสนเทศที่ได้ไปใช้ในการวิจัยได้ดังตารางที่ 2.1

**ตารางที่ 2.1 สรุปสาระสำคัญของงานวิจัยที่เกี่ยวข้องกับดัชนีความสอดคล้องของข้อคำถาม** (item fit index)

| นักวิชาการ | ชื่องานวิจัย | ประเภทของข้อสอบหรือข้อคำถาม | วิธีหรือดัชนีที่ใช้ในการวิจัย | โมเดลที่ใช้ในการวิจัย | ขนาดกลุ่มตัวอย่าง | ความยาวแบบวัดหรือแบบสอบ | เกณฑ์ที่ใช้ในการเปรียบเทียบ | ผลการวิจัย |
|---|---|---|---|---|---|---|---|---|
| Orlando & Thissen (2000) | Likelihood-based item fit indices for dichotomous item response theory models. | Dichotomous item | $S-\chi^2$,$S-G^2$, $Q_1-\chi^2$ และ $Q_1-G^2$ | โมเดล 3 พารามิเตอร์ | 1000 คน | 10, 40 และ 80 ข้อ | ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) และอำนาจการทดสอบ (Power of the test) | $S-\chi^2$ มีค่า Type I error ที่ค่อนข้างคงเส้นคงวา มีอำนาจการทดสอบ (power of the test) ในการตรวจสอบความไม่สอดคล้อง (misfit) ของข้อสอบ ได้ดีเมื่อเทียบกับดัชนีตัวอื่น ๆ ที่ใช้ในการศึกษา |
| Orlando & Thissen (2003) | Further investigation of the performance of $S-\chi^2$ : An item fit index for use with dichotomous item response theory model | Dichotomous item | $S-\chi^2$ และ $Q_1-\chi^2$ | โมเดล 3 พารามิเตอร์ | 500, 1000 และ 2000 คน | 10, 40 และ 80 ข้อ | ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) และอำนาจการทดสอบ (Power of the test) ที่ได้จาก Receiver Operation Characterictic (ROC) | ประสิทธิภาพของ $S-\chi^2$ ดีกว่า $Q_1-\chi^2$ ในหลายเงื่อนไขที่ใช้ในการศึกษา |

**ตารางที่** 2.1 (ต่อ)

| นักวิชาการ | ชื่องานวิจัย | ประเภทของข้อสอบหรือข้อคำถาม | วิธีหรือดัชนีที่ใช้ในการวิจัย | โมเดลที่ใช้ในการวิจัย | ขนาดกลุ่มตัวอย่าง | ความยาวแบบวัดหรือแบบสอบ | เกณฑ์ที่ใช้ในการเปรียบเทียบ | ผลการวิจัย |
|---|---|---|---|---|---|---|---|---|
| Stone และ Zhang (2003) | Assessing goodness of fit of Item Response Theory Models: A comparison of traditional and alternative procedure. | Dichotomous item | วิธีดั้งเดิม (Traditional Procedure) คือวิธีของ Donoghue และ Hombo (2001) กับวิธีของ Bock (1972) วิธีทางเลือก (Alternative Procedure) ซึ่งประกอบด้วยวิธีของ Orlando และ Thissen (2000) และวิธีของ Stone (2000) | โมเดล 1, 2 และ 3 พารามิเตอร์ | 500, 1000 และ 2000 คน | 10, 20 และ 40 ข้อ | ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) และอำนาจการทดสอบ (Power of the test) | วิธีการของ Donoghue และ Hombo (2001) สามารถใช้ได้กับแบบสอบความยาว 10 ข้อ ส่วนความยาว 20 และ 40 ข้อต้องใช้วิธีของ Bock (1972) แทนซึ่ง จะให้ค่าความคลาดเคลื่อนประเภทที่ 1ที่มาก วิธีที่มีอำนาจในการตรวจสอบความไม่สอดคล้อง (misfit) มากที่สุดคือวิธีของ Stone(2000) อย่างไรก็ตามวิธีของ Orlando และ Thissen มีอำนาจเพียงพอ (adequate power) ในการตรวจสอบความไม่สอดคล้อง ที่ขนาดกลุ่มตัวอย่างขนาดใหญ่ในขณะที่วิธีของ Stone (2000) จะมีอำนาจที่เพียงพอในกลุ่มตัวอย่างซึ่งมีขนาดเล็ก |

**ตารางที่** 2.1 (ต่อ)

| นักวิชาการ | ชื่องานวิจัย | ประเภทของข้อสอบหรือข้อคำถาม | วิธีหรือดัชนีที่ใช้ในการวิจัย | โมเดลที่ใช้ในการวิจัย | ขนาดกลุ่มตัวอย่าง | ความยาวแบบวัดหรือแบบสอบ | เกณฑ์ที่ใช้ในการเปรียบเทียบ | ผลการวิจัย |
|---|---|---|---|---|---|---|---|---|
| Dodeen (2004) | The relationship between item parameter and item fit | Dichotomous item | $\chi_B^2$ ของ Bock (1972) และ the Standardized Residuals (SRs) | โมเดล 3 พารามิเตอร์ | 1000 คน | 50 ข้อ | ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ค่าสหสัมพันธ์ของเพียร์สัน (Pearson Correlation) และการวิเคราะห์ความแปรปรวน (ANOVA) ตามด้วยการเปรียบเทียบรายคู่ (pairwise comparison) | ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error)ของ $\chi_B^2$ มากกว่า SRs นอกจากนี้มีความสัมพันธ์เชิงบวกระหว่างดัชนีความสอดคล้องของข้อคำถาม (item fit index)กับค่าอำนาจจำแนกและค่าการเดา แต่ไม่มีความสัมพันธ์กับค่าความยาก นอกจากนี้เมื่อค่าพารามิเตอร์ของข้อสอบทั้งอำนาจจำแนก ค่าความยาก และค่าการเดาเพิ่มขึ้น จำนวนข้อสอบที่ไม่สอดคล้อง (misfit) จะมีค่าเพิ่มขึ้นด้วย ทั้งในกรณีของค่าดัชนี $\chi_B^2$ และ Standardized Residuals (SRs) |

**ตารางที่** 2.1 (ต่อ)

| นักวิชาการ | ชื่องานวิจัย | ประเภทของข้อสอบหรือข้อคำถาม | วิธีหรือดัชนีที่ใช้ในการวิจัย | โมเดลที่ใช้ในการวิจัย | ขนาดกลุ่มตัวอย่าง | ความยาวแบบวัดหรือแบบสอบ | เกณฑ์ที่ใช้ในการเปรียบเทียบ | ผลการวิจัย |
|---|---|---|---|---|---|---|---|---|
| DeMars (2005) | Type I error rate for PARSCALE's fit index | Polytomous item | ดัชนี PARSCALE $G^2$ | PCM และ GRM ที่มี 5 รายการคำตอบ | 1000 คน | 10 และ 20 ข้อ | สัดส่วนของข้อคำถามที่ถูกบ่งชี้ว่าไม่สอดคล้อง (misfit) นั่นคือ ค่าความคลาดเคลื่อนประเภทที่ 1มีค่าเท่ากับ 0.05 ($\alpha = 0.05$) | ดัชนี PARSCALE $G^2$ ไม่เหมาะสมที่จะนำไปใช้กับแบบสอบที่สั้น แต่เหมาะที่จะใช้กับแบบสอบที่ยาว |
| Kang และ Chen (2008) | Performance of the Generalized $S-\chi^2$ item fit index for polytomous IRT models. | Polytomous item | ดัชนี PARSCALE $G^2$ และ ดัชนี Generalized $S-\chi^2$ | GPCM, PCM และ RSM ที่มี 5 รายการคำตอบ | 500, 1000, 2000 และ 5000 คน | 5, 10, 20 ข้อ | ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) และค่าอำนาจการทดสอบ (Power of the test) | PARSCALE $G^2$ ไม่ค่อยมีประสิทธิภาพในการตรวจสอบ item fit ในสถานการณ์ ทั้งเงื่อนไขความยาวแบบวัดที่สั้น และปานกลาง หรือในกรณีตัวอย่างขนาดใหญ่ ขณะที่ Generalized $S-\chi^2$ มีประสิทธิภาพมาก โดยเฉพาะ ในกรณีความยาวแบบวัด 5 หรือ 10 ข้อ ดังนั้นGeneralized $S-\chi^2$ จึงเป็นดัชนีที่เหมาะสมในการบ่งชี้ความสอดคล้องของข้อคำถาม |

**ตารางที่** 2.1 (ต่อ)

| นักวิชาการ | ชื่องานวิจัย | ประเภทของข้อสอบหรือข้อคำถาม | วิธีหรือดัชนีที่ใช้ในการวิจัย | โมเดลที่ใช้ในการวิจัย | ขนาดกลุ่มตัวอย่าง | ความยาวแบบวัดหรือแบบสอบ | เกณฑ์ที่ใช้ในการเปรียบเทียบ | ผลการวิจัย |
|---|---|---|---|---|---|---|---|---|
| Liang และ Wells (2009) | A model fit statistic for generalized partial credit model. | Polytomous item | $S-\chi^2$, PARSCALE $G^2$ และ Root Integrated Square Error (RISE) | GPCM | ครั้งแรก 500, 1000 และ 2000 คน ครั้งที่ 2 500, 1000 และ 2000 คน | ครั้งแรก10, 20 และ40 ข้อ ครั้งที่ 2แบบสอบ 40 ข้อ ประกอบด้วยข้อสอบหลายตัวเลือก 30 ข้อ ตอบสั้น 5 ข้อเป็นโมเดลโลจิสติกแบบ 2 พารามิเตอร์ และ 5 ข้อเป็น GPCM | ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) และค่าอำนาจการทดสอบ (Power of the test) | RISE และ $S-\chi^2$ จะให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error)ที่น้อยกว่า เมื่อเทียบกับ PARSCALE $G^2$ และเมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้นดัชนีทั้ง 3 ตัวก็จะให้อำนาจการทดสอบ (Power of the test) ที่มากขึ้นด้วย แต่ RISE มีข้อได้เปรียบวิธีอื่น ๆ คือ จะให้ Graphical model misfit ซึ่งจะช่วยบอกตำแหน่งและประเภทของ misfit ได้ |

**ตารางที่** 2.1 (ต่อ)

| นักวิชาการ | ชื่องานวิจัย | ประเภทของข้อสอบหรือข้อคำถาม | วิธีหรือดัชนีที่ใช้ในการวิจัย | โมเดลที่ใช้ในการวิจัย | ขนาดกลุ่มตัวอย่าง | ความยาวแบบวัดหรือแบบสอบ | เกณฑ์ที่ใช้ในการเปรียบเทียบ | ผลการวิจัย |
|---|---|---|---|---|---|---|---|---|
| Lattuis, Clark และ O'Brien (2009) | An examination of item response theory item fit indices for Grade Response Model. | Polytomous item | Generalized $S-\chi^2$, $\chi^{2*}$ และ Adjusted $\chi^2$ to degree of freedom ratio ($\chi^2/dfs$) | GRM, GGUM | 500, 1000 และ 2000 คน | 10 และ20 ข้อ | ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) และอำนาจการทดสอบ (Power of the test) | $S-\chi^2$ และ $\chi^{2*}$ให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error)ต่ำกว่าที่คาดไว้ ส่วนดัชนี Adjusted $\chi^2/dfs$ มีค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error)ที่มากเมื่อใช้กับการทดสอบข้ามกลุ่ม และมีค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ต่ำมากเมื่อไม่มีการทดสอบข้ามกลุ่ม นอกจากนี้เมื่อไม่มีการทดสอบข้ามกลุ่ม Adjusted $\chi^2/dfs$ จะมีอำนาจการทดสอบสูงสุดในทุกกรณี |

**ตารางที่** 2.1 (ต่อ)

| นักวิชาการ | ชื่องานวิจัย | ประเภทของข้อสอบหรือข้อคำถาม | วิธีหรือดัชนีที่ใช้ในการวิจัย | โมเดลที่ใช้ในการวิจัย | ขนาดกลุ่มตัวอย่าง | ความยาวแบบวัดหรือแบบสอบ | เกณฑ์ที่ใช้ในการเปรียบเทียบ | ผลการวิจัย |
|---|---|---|---|---|---|---|---|---|
| Kang และ Chen (2010) | Performance of the generalized $S-\chi^2$ item fit index for graded response model. | Polytomous item | Generalized $S-\chi^2$ | GRM ที่มี 3 และ 5 รายการคำตอบ | 500, 1000 และ 2000 คน | 5, 10 และ 20 ข้อ | ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) และอำนาจการทดสอบ (Power of the test) | ดัชนี Generalized $S-\chi^2$ ให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error)และอำนาจการทดสอบ (Power of the test) ที่เหมาะสมในการตรวจสอบความไม่สอดคล้อง (misfit)ทั้งสองประเภท (M-type misfit และ D-type misfit) ของข้อคำถามในขนาดกลุ่มตัวอย่างที่ใหญ่ แต่ถ้าพิจารณาจากจำนวนรายการคำตอบที่แตกต่างกันจะพบว่าโดยส่วนใหญ่เมื่อมี 5 รายการคำตอบค่าความคลาดเคลื่อนประเภทที่ 1จะมากกว่าเมื่อมี 3 รายการคำตอบ แต่โดยส่วนใหญ่แล้วดัชนี Generalized $S-\chi^2$จะสามารถควบคุมค่าความคลาดเคลื่อนประเภทที่ 1ได้ในระดับที่พอเพียง |

**ตอนที่ 7 การทดสอบความสอดคล้องของข้อคำถามกับโมเดลทฤษฎีการตอบสนองข้อสอบ**

ในการทดสอบความสอดคล้องของข้อคำถามกับโมเดลทฤษฎีการตอบสนองข้อสอบมีสมมติฐานในการทดสอบดังต่อไปนี้

$H_0$ : ข้อคำถามมีความสอดคล้องกับโมเดลทฤษฎีการตอบสนองข้อสอบ

$H_1$ : ข้อคำถามไม่มีความสอดคล้องกับโมเดลทฤษฎีการตอบสนองข้อสอบ

จากสมมติฐานดังกล่าวนี้จะมีตัวสถิติที่ใช้ในการทดสอบ ซึ่งในการวิจัยครั้งนี้ก็คือดัชนีความสอดคล้องของข้อคำถาม ซึ่งการทดสอบจะมีขั้นตอนเหมือนกับการทดสอบสมมติฐานทางสถิติโดยทั่ว ๆ ไป คือ มีการกำหนดระดับความเชื่อมั่นที่ใช้ในการทดสอบ ซึ่งในการศึกษาวิจัยครั้งนี้ใช้ระดับความเชื่อมั่น 95% เกณฑ์ที่ใช้ในการตัดสินใจว่าจะปฏิเสธหรือยอมรับสมมติฐานศูนย์ที่ได้ตั้งไว้มี 2 วิธี คือ

1) การใช้ค่า p-value ที่ได้จากโปรแกรม ซึ่งถ้าค่า p-value ที่ได้จากโปรแกรมมีค่าน้อยกว่าระดับนัยสำคัญที่กำหนดไว้ซึ่งในการวิจัยครั้งนี้คือ .05 ก็จะทำการปฏิเสธสมมติฐานศูนย์ที่ได้ตั้งไว้กล่าวคือข้อคำถามไม่มีความสอดคล้องกับโมเดลทฤษฎีการตอบสนองข้อสอบ
2) การใช้ค่าสถิติที่อ่านค่าได้จากตารางสถิติ กล่าวคือ ดัชนีที่ใช้ในการบ่งชี้ความสอดคล้องของข้อคำถามนั้นเป็นตัวสถิติที่มีการแจกแจงแบบไคสแควร์ ดังนั้นจะปฏิเสธสมมติฐานศูนย์ที่ได้ตั้งไว้เมื่อค่าดัชนีที่คำนวณได้มีค่ามากกว่าค่าสถิติที่อ่านค่าได้จากตารางไคสแควร์

จากการศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้องซึ่งได้นำเสนอทั้งหมด ทำให้ได้องค์ความรู้ที่สามารถนำไปใช้ในการวิจัยต่อไป โดยเฉพาะอย่างยิ่งองค์ความรู้เกี่ยวกับตัวแปรที่ใช้ในการศึกษาประสิทธิภาพของดัชนีความสอดคล้องของข้อสอบ (item fit index) ดังต่อไปนี้

1. ความยาวแบบวัด หรือความยาวแบบสอบจากการศึกษางานวิจัยที่ผ่านมาในการจำลองข้อมูลเพื่อใช้ในการวิจัย นิยมใช้ความยาวของแบบวัด (test length) 10, 20, 40 ข้อเป็นจำนวนมาก ความยาวของแบบวัดสูงที่สุดคือ 80 ข้อ ความยาวของแบบวัดน้อยที่สุดคือ 5 ข้อ โดยสามารถพิจารณารายละเอียดได้จากตารางที่ 2.2

**ตารางที่ 2.2 สรุปการเลือกใช้ความยาวแบบวัดหรือแบบสอบในการวิจัยของนักวิชาการ**

| นักวิชาการ | ความยาวแบบวัดหรือแบบสอบ | | | | | |
|---|---|---|---|---|---|---|
| | 5 ข้อ | 10 ข้อ | 20 ข้อ | 40 ข้อ | 50 ข้อ | 80 ข้อ |
| Orlando และ Thissen (2000) | | ✓ | | ✓ | | ✓ |
| Orlando และ Thissen (2003) | | ✓ | | ✓ | | ✓ |
| Stone และ Zhang (2003) | | ✓ | ✓ | ✓ | | |
| Dodeen (2004) | | | | | ✓ | |
| DeMars (2005) | | ✓ | ✓ | | | |
| Kang และ Chen (2008) | ✓ | ✓ | ✓ | | | |
| Liang และ Wells (2009) ครั้งที่ 1 | | ✓ | ✓ | ✓ | | |
| Liang และ Wells (2009) ครั้งที่ 2 | | | | ✓ | | |
| Lattuis, Clark และ O'Brien (2009) | | ✓ | ✓ | | | |
| Kang และ Chen (2010) | ✓ | ✓ | ✓ | | | |

✓ หมายถึง นักวิจัยเลือกใช้ความยาวแบบวัดหรือแบบสอบนั้นในการวิจัย

นอกจากนี้ ในการศึกษาวิจัยของนักวิชาการที่ผ่านมายังให้สารสนเทศที่สอดคล้องกัน กล่าวคือ เมื่อความยาวแบบวัด (Test length) มีค่าน้อยหรือเป็นแบบวัดที่สั้น ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) จะมีค่าค่อนข้างมาก (Orlando & Thissen, 2000; Demars, 2005; Kang & Chen, 2008; Liang & Wells, 2009) แต่เมื่อพิจารณาที่อำนาจการทดสอบ (Power of the test) พบว่าเมื่อความยาวแบบวัดเพิ่มมากขึ้น มีแนวโน้มที่อำนาจการทดสอบจะเพิ่มขึ้นด้วย (Kang & Chen, 2010) อย่างไรก็ตามข้อค้นพบนี้อาจไม่เป็นจริงเสมอไปในทุกกรณี ในบางครั้งอาจมีอิทธิพลของขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ และชนิดของดัชนีความสอดคล้องของข้อคำถามที่ส่งผลให้การเพิ่มขึ้นหรือลดลงของค่าอำนาจการทดสอบและค่าความคลาดเคลื่อนประเภทที่ 1 ไม่เป็นไปตามที่ค้นพบในทุกกรณี เช่น ผลการศึกษาของ Stone และ Zhang (2003) ที่อำนาจการทดสอบลดลงเล็กน้อยเมื่อความยาวแบบวัดเพิ่มขึ้น ดังนั้น Liang และ Wells (2009) จึงให้ความเห็นว่าถ้าค่าความคลาดเคลื่อนประเภทที่ 1 มีค่ามาก อำนาจการทดสอบอาจไม่น่าเชื่อถือ เนื่องจากค่าความคลาดเคลื่อนประเภทที่ 1 ที่มีค่ามากจะส่งผลให้อำนาจการทดสอบมีค่ามากผิดปกติได้ ดังนั้นในการเลือกเกณฑ์ในการบ่งชี้ความสอดคล้องของข้อคำถามกับโมเดลทฤษฎีการตอบสนองข้อสอบของ Kang และ Chen (2008) จึงพิจารณาที่ค่าความ

คลาดเคลื่อนประเภทที่ 1 ก่อนถ้าหากค่าความคลาดเคลื่อนประเภทที่ 1 มีค่ามาก ก็จะไม่พิจารณาอำนาจการทดสอบ

2. ขนาดกลุ่มตัวอย่างหรือจำนวนผู้ตอบข้อคำถาม เริ่มต้นที่ 500 คน ขนาดกลุ่มตัวอย่างที่นิยมใช้มากคือ 500, 1000, 2000 คน ขนาดกลุ่มตัวอย่างสูงที่สุดคือ 5000 คน โดยสามารถพิจารณารายละเอียดได้จากตารางที่ 2.3

**ตารางที่ 2.3 สรุปการเลือกใช้ขนาดกลุ่มตัวอย่างในการวิจัยของนักวิชาการ**

| นักวิชาการ | ขนาดกลุ่มตัวอย่าง | | | |
|---|---|---|---|---|
| | 500 คน | 1000 คน | 2000 คน | 5000 คน |
| Orlando และ Thissen (2000) | | ✓ | | |
| Orlando และ Thissen (2003) | ✓ | ✓ | ✓ | |
| Stone และ Zhang (2003) | ✓ | ✓ | ✓ | |
| Dodeen (2004) | | ✓ | | |
| DeMars (2005) | | ✓ | | |
| Kang และ Chen (2008) | ✓ | ✓ | ✓ | ✓ |
| Liang และ Wells (2009) ครั้งที่ 1 | ✓ | ✓ | ✓ | |
| Liang และ Wells (2009) ครั้งที่ 2 | ✓ | ✓ | ✓ | |
| Lattuis, Clark และ O'Brien (2009) | ✓ | ✓ | ✓ | |
| Kang และ Chen (2010) | ✓ | ✓ | ✓ | |

✓ หมายถึง นักวิชาการเลือกใช้ขนาดกลุ่มตัวอย่างนั้นในการวิจัย

นอกจากนี้ ในการศึกษาวิจัยของนักวิชาการที่ผ่านมายังให้สารสนเทศที่สอดคล้องกัน กล่าวคือ เมื่อขนาดกลุ่มตัวอย่าง (sample size) เพิ่มขึ้น ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) จะมีค่าลดลง (Stone & Zhang, 2003; Kang & Chen, 2010) แต่เมื่อพิจารณาที่อำนาจการทดสอบ (Power of the test) พบว่าเมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้น มีแนวโน้มที่อำนาจการทดสอบจะเพิ่มขึ้นตามไปด้วย (Stone & Zhang, 2003; Liang & Wells, 2009; Lattuis, Clark & O'Brein, 2009; Kang & Chen, 2010) แต่ข้อค้นพบนี้อาจไม่เป็นจริงเสมอไปในทุกกรณี ในบางครั้งอาจมีอิทธิพลของความยาวแบบวัด จำนวนรายการคำตอบ และชนิดของดัชนีความ

สอดคล้องของข้อสอบที่ส่งผลให้การเพิ่มขึ้นหรือลดลงของค่าอำนาจการทดสอบและค่าความคลาดเคลื่อนประเภทที่ 1 ไม่เป็นไปตามที่ค้นพบในทุกกรณี

3. จำนวนรายการคำตอบ งานวิจัยที่ศึกษาในกรณีรายการคำตอบมีจำนวนที่หลากหลายยังมีไม่มากนัก เท่าที่พบจะทำการศึกษาเพียงจำนวนรายการคำตอบเป็น 3 และ 5 เท่านั้น ซึ่ง Kang และ Chen (2010) ได้ให้เหตุผลในการศึกษาที่จำนวนรายการคำตอบเป็น 3 และ 5 เนื่องจากเป็นจำนวนรายการคำตอบที่นิยมใช้ในการสร้างแบบวัด ซึ่งจากผลการวิจัยพบว่าโดยส่วนใหญ่เมื่อมี 5 รายการคำตอบค่าความคลาดเคลื่อนประเภทที่ 1จะมากกว่าเมื่อมี 3 รายการคำตอบ ซึ่งสันนิษฐานได้ว่าจำนวนรายการคำตอบอาจส่งผลต่อประสิทธิภาพของดัชนีความสอดคล้องของข้อสอบ นอกจากนี้ ในการสร้างแบบวัดทางจิตวิทยายังนิยมใช้จำนวนรายการคำตอบเป็น 7 และ 9 อาทิ The High School Chemistry Self-Efficacy Scale (HCSS) ของ Aydin และ Uzuntiryaki (2009) ซึ่งใช้ 9 รายการคำตอบ และ The Schutte Self-Report Emotional Intelligence Scale on International Students ของ Mun, Wang, Kim, และ Bodenhorn (2009) ซึ่งใช้ 7 รายการคำตอบ

นอกจากนี้แล้วในการทำซ้ำในแต่ละสถานการณ์ที่ได้ทำการศึกษา นักวิชาการส่วนใหญ่ใช้การทำซ้ำประมาณ 20-100 ครั้ง ซึ่งจากการศึกษาของ Harwell, Hsu และ Kirisci (1996 อ้างถึงใน พัชรี จันทร์เพ็ง, 2550) พบว่าถ้าศึกษาโดยใช้โมเดลทฤษฎีการตอบสนองข้อสอบ (Item Response Theory Model) เป็นฐานควรมีการทำซ้ำอย่างน้อย 20-25 ครั้งเพื่อให้การประมาณค่ามีความเที่ยงมากยิ่งขึ้น ดังนั้นในการศึกษาวิจัยครั้งนี้จึงเลือกใช้จำนวนครั้งในการทำซ้ำ 30 ครั้งในแต่ละสถานการณ์

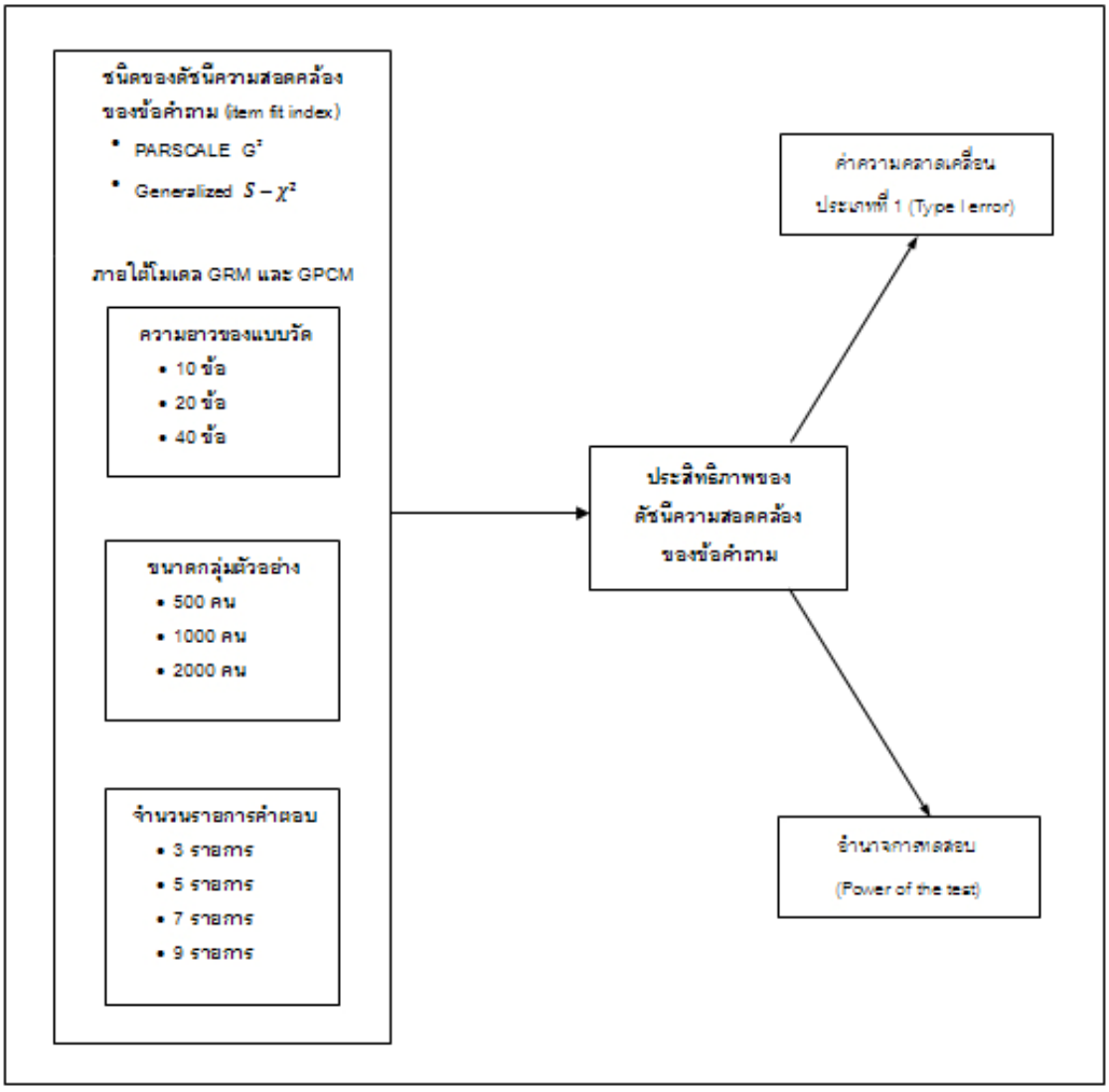
จากการศึกษา รวบรวมข้อมูลเกี่ยวกับแนวคิดและทฤษฎีต่างๆ ดังที่ได้กล่าวมาข้างต้น ทำให้ผู้วิจัยเกิดแนวความคิดในการทำการศึกษาวิจัยเกี่ยวกับดัชนีความสอดคล้องของข้อคำถาม (item fit index) ซึ่งสามารถสรุปได้ตามกรอบแนวคิดที่ใช้ในการวิจัย ดังนี้

**ตอนที่ 8 กรอบแนวคิดที่ใช้ในการวิจัย**

การศึกษาวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามสองชนิด คือ Generalized $S-\chi^2$ และ PARSCALE $G^2$ ในสถานการณ์ที่ความยาวแบบวัด ขนาดกลุ่มตัวอย่างและจำนวนรายการคำตอบมีความแตกต่างกันภายใต้โมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค 2 โมเดลคือ Grade Response Model

(GRM) และ Generalized Partial Credit Model (GPCM) เกณฑ์ที่นำมาใช้ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถาม คือ ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) และอำนาจการทดสอบ (Power of the test) ซึ่งผลการวิจัยจะเป็นแนวทางในการเลือกใช้ดัชนีความสอดคล้องของข้อคำถามในสถานการณ์จริงต่อไปในอนาคต รวมทั้งยังอาจเป็นแนวทางในการกำหนดความยาวแบบวัด ขนาดกลุ่มตัวอย่าง และจำนวนรายการคำตอบในการสร้างแบบวัดทางจิตวิทยาหรือแบบสอบทางการศึกษาที่มีการตรวจให้คะแนนแบบพหุวิภาคโดยสามารถแสดงกรอบแนวคิดในการวิจัยดังแผนภาพที่ 2.1

**แผนภาพที่ 2.1 กรอบแนวคิดในการวิจัย**



# บทที่ 3

# วิธีดำเนินการวิจัย

การวิจัยครั้งนี้เป็นการวิจัยเชิงทดลองโดยมีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถาม (item fit index) สำหรับโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค สองชนิดคือ Generalized $S-\chi^2$ และ PARSCALE $G^2$ ในสถานการณ์ ณ ความยาวแบบวัดเป็น 10, 20, และ 40 ข้อ ขนาดกลุ่มตัวอย่างเป็น 500, 1000, และ 2000 คน จำนวนรายการคำตอบเป็น 3, 5, 7, และ 9 รายการ โดยใช้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) กับ อำนาจการทดสอบ (Power of the test) เป็นเกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถาม (item fit index) ซึ่งผลการวิจัยจะเป็นแนวทางในการเลือกใช้ดัชนีความสอดคล้องของข้อคำถามสำหรับ Grade Response Model (GRM) และ Generalized Partial Credit Model (GPCM) ในสถานการณ์ที่จำนวนรายการคำตอบ ขนาดกลุ่มตัวอย่างและความยาวแบบวัดมีความแตกต่างกัน เพื่อให้การบ่งชี้ความสอดคล้องของข้อสอบมีความคลาดเคลื่อนน้อยที่สุด และเป็นแนวทางในการกำหนดความยาวแบบวัด ขนาดกลุ่มตัวอย่าง และจำนวนรายการคำตอบในการสร้างแบบวัดทางจิตวิทยาหรือแบบสอบทางการศึกษาที่มีการตรวจให้คะแนนแบบพหุวิภาค ซึ่งมีวิธีดำเนินการวิจัยดังต่อไปนี้

1. ศึกษาค้นคว้าทฤษฎีและหลักการของดัชนีความสอดคล้องของข้อคำถาม (item fit index) จากเอกสาร ตำราทางวิชาการ วารสาร และงานวิจัยที่เกี่ยวข้องซึ่งส่วนใหญ่เป็นงานวิจัยทางด้านทฤษฎีการวัด (Measurement Theory)
2. จำลองข้อมูลการตอบข้อคำถามของผู้ตอบแบบวัดด้วยโปรแกรม WINGEN จำแนกตามเงื่อนไขสถานการณ์ที่ทำการศึกษาประกอบด้วย 1) ความยาวแบบวัด 3 ขนาดคือ 10, 20 และ 40 ข้อ 2) ขนาดกลุ่มตัวอย่าง 3 ขนาดคือ 500, 1000 และ 2000 คน 3) จำนวนรายการคำตอบ 4 ขนาดคือ 3, 5, 7 และ 9 รายการ ภายใต้โมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค 2 โมเดล คือ Grade Response Model (GRM) และ Generalized Partial Credit Model (GPCM) นอกจากนี้กำหนดให้ค่าพารามิเตอร์ความสามารถของผู้ตอบแบบวัดมีการแจกแจงแบบปกติซึ่งเป็นสถานการณ์ที่พบบ่อยในการทดสอบทางการศึกษาและการวัดทางจิตวิทยา และในแต่ละสถานการณ์ที่ทำการศึกษามีการกระทำซ้ำ 30 ครั้ง ในขั้นตอนนี้จะทำให้ได้โมเดลจำลอง (Generating Model: GM) จำนวน 30 โมเดลในแต่ละสถานการณ์ ดังนั้นจะมีโมเดลจำลองทั้งหมด 72 x 30 = 2160 โมเดล

3. วิเคราะห์ข้อมูลเบื้องต้นเพื่อทำให้ทราบลักษณะของข้อมูลที่ได้จำลองขึ้นในแต่ละสถานการณ์ที่ทำการศึกษา โดยพิจารณาค่าสถิติพื้นฐานได้แก่ ค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐาน ของค่าพารามิเตอร์ความสามารถผู้สอบ($\theta$) และค่าพารามิเตอร์ข้อคำถาม
4. จากโมเดลจำลอง (Generating Model: GM) ทำการตรวจสอบข้อมูลที่ได้จำลองขึ้นในแต่ละสถานการณ์ เพื่อตรวจสอบว่าข้อมูลการตอบแบบวัดเป็นไปตามเงื่อนไขที่กำหนดไว้หรือไม่ โดยพิจารณาลักษณะการแจกแจงของค่าพารามิเตอร์ความสามารถผู้สอบ($\theta$) และค่าพารามิเตอร์ข้อคำถาม
5. ทำการปรับเทียบโมเดล ซึ่งโมเดลที่ได้จากการปรับเทียบ (Calibrate) นี้ เรียกว่าโมเดลเทียบมาตรฐาน (Calibrating Model: CM) ทำเพื่อปรับเทียบหรือประมาณค่าพารามิเตอร์ในโมเดลจำลอง (Generating Model: GM) เพื่อประโยชน์ในการคำนวณหาค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) กับ อำนาจการทดสอบ (Power of the test)
6. คำนวณดัชนี PARSCALE $G^2$ และหาค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) กับ อำนาจการทดสอบ (Power of the test) โดยใช้โปรแกรม PARSCALE ของ Muraki และ Bock (1990)
7. คำนวณดัชนี Generalized $S-\chi^2$ และหาค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) กับ อำนาจการทดสอบ (Power of the test) โดยใช้ IRTFIT macros ของ Bjorner, Smith, Stone และ Sun (2007) ซึ่งประมวลผลบนโปรแกรม SAS
8. พิจารณาค่าความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ จำแนกตามโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาคที่ใช้ในการศึกษาว่ามีการเปลี่ยนแปลงไปตามความยาวแบบวัด ขนาดกลุ่มตัวอย่าง และจำนวนรายการคำตอบที่เปลี่ยนแปลงไปอย่างไร
9. เปรียบเทียบประสิทธิภาพของดัชนี Generalized $S-\chi^2$ กับ PARSCALE $G^2$ โดยใช้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) กับ อำนาจการทดสอบ (Power of the test)
10. สรุปผลการวิเคราะห์ข้อมูล เพื่อเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถาม (item fit index) ในแต่ละสถานการณ์ที่ทำการศึกษา

**ตัวแปรที่ศึกษา**

ในการศึกษาวิจัยนี้มีตัวแปรอิสระและตัวแปรตามที่ใช้ในการศึกษาดังต่อไปนี้

1. ตัวแปรอิสระ (independent variable) มี 1 ตัว คือ ชนิดของดัชนีความสอดคล้องของข้อคำถามซึ่งมี 2 ชนิด คือ Generalized $S-\chi^2$ และ PARSCALE $G^2$ โดยมีเงื่อนไขในการศึกษา คือ
   1) ความยาวของแบบวัดมี 3 ขนาดคือ 10, 20 และ 40 ข้อ
   2) ขนาดกลุ่มตัวอย่าง 3 ขนาดคือ 500, 1000 และ 2000 คน
   3) จำนวนรายการคำตอบ 4 ขนาดคือ 3, 5, 7 และ 9 รายการ

   โดยทำการศึกษาภายใต้โมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค 2 โมเดล คือ Grade Response Model (GRM) และ Generalized Partial Credit Model (GPCM)
2. ตัวแปรตาม (dependent variable) มี 1 ตัว คือ ประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถาม ซึ่งสามารถวัดได้จากค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) กับอำนาจการทดสอบ (Power of the test)

จากที่กล่าวมานั้น เป็นการสรุปขั้นตอนการดำเนินการศึกษาวิจัยโดยรวม ตั้งแต่การศึกษาค้นคว้าทฤษฎีและหลักการของดัชนีความสอดคล้องของข้อคำถาม (item fit index) จากเอกสาร ตำราทางวิชาการ วารสาร และงานวิจัยที่เกี่ยวข้อง จนถึงการวิเคราะห์ข้อมูลและสรุปผลการวิจัย ซึ่งรายละเอียดนั้นจะนำเสนอวิธีดำเนินการวิจัยโดยละเอียดใน 2 ตอน ได้แก่

ตอนที่ 1 การศึกษาการจำลองข้อมูล

ตอนที่ 2 การวิเคราะห์ข้อมูลที่ได้จากการจำลองข้อมูล

ซึ่งแต่ละตอนมีรายละเอียดดังนี้

**ตอนที่ 1 การศึกษาการจำลองข้อมูล**

ผู้วิจัยจำลองข้อมูลการตอบข้อคำถามของผู้ตอบแบบวัดตามตัวแปรที่กำหนดในการวิจัยโดยดำเนินการจำลองข้อมูลโดยใช้โปรแกรม WINGEN ของ Hambleton และ Han (2007) ตามเงื่อนไขที่ใช้ในการศึกษาคือ

1. ความยาวแบบวัด 3 ขนาดคือ 10, 20 และ 40 ข้อ
2. ขนาดกลุ่มตัวอย่าง 3 ขนาดคือ 500, 1000 และ 2000 คน
3. จำนวนรายการคำตอบ 4 ขนาดคือ 3, 5, 7 และ 9 รายการ

โดยเงื่อนไขทั้งหมดนี้อยู่ภายใต้โมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค สองโมเดลคือ Grade Response Model (GRM) และ Generalized Partial Credit Model (GPCM) ซึ่ง

ค่าพารามิเตอร์ความสามารถของผู้ตอบแบบวัดมีการแจกแจงแบบปกติ ดังนั้นสถานการณ์ทั้งหมดที่ทำการศึกษามีทั้งหมด 72 สถานการณ์ดังในตารางที่ 3.1

**ตารางที่ 3.1 สถานการณ์ทั้งหมดที่ทำการศึกษาทั้ง 72 สถานการณ์**

| ลำดับที่ | โมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค | ความยาวแบบวัด | ขนาดกลุ่มตัวอย่าง | จำนวนรายการคำตอบ |
|---|---|---|---|---|
| 1 | Grade Response Model (GRM) | 10 | 500 | 3 |
| 2 | Grade Response Model (GRM) | 10 | 500 | 5 |
| 3 | Grade Response Model (GRM) | 10 | 500 | 7 |
| 4 | Grade Response Model (GRM) | 10 | 500 | 9 |
| 5 | Grade Response Model (GRM) | 10 | 1000 | 3 |
| 6 | Grade Response Model (GRM) | 10 | 1000 | 5 |
| 7 | Grade Response Model (GRM) | 10 | 1000 | 7 |
| 8 | Grade Response Model (GRM) | 10 | 1000 | 9 |
| 9 | Grade Response Model (GRM) | 10 | 2000 | 3 |
| 10 | Grade Response Model (GRM) | 10 | 2000 | 5 |
| 11 | Grade Response Model (GRM) | 10 | 2000 | 7 |
| 12 | Grade Response Model (GRM) | 10 | 2000 | 9 |
| 13 | Grade Response Model (GRM) | 20 | 500 | 3 |
| 14 | Grade Response Model (GRM) | 20 | 500 | 5 |
| 15 | Grade Response Model (GRM) | 20 | 500 | 7 |
| 16 | Grade Response Model (GRM) | 20 | 500 | 9 |
| 17 | Grade Response Model (GRM) | 20 | 1000 | 3 |
| 18 | Grade Response Model (GRM) | 20 | 1000 | 5 |
| 19 | Grade Response Model (GRM) | 20 | 1000 | 7 |
| 20 | Grade Response Model (GRM) | 20 | 1000 | 9 |
| 21 | Grade Response Model (GRM) | 20 | 2000 | 3 |
| 22 | Grade Response Model (GRM) | 20 | 2000 | 5 |
| 23 | Grade Response Model (GRM) | 20 | 2000 | 7 |
| 24 | Grade Response Model (GRM) | 20 | 2000 | 9 |
| 25 | Grade Response Model (GRM) | 40 | 500 | 3 |
| 26 | Grade Response Model (GRM) | 40 | 500 | 5 |
| 27 | Grade Response Model (GRM) | 40 | 500 | 7 |
| 28 | Grade Response Model (GRM) | 40 | 500 | 9 |
| 29 | Grade Response Model (GRM) | 40 | 1000 | 3 |
| 30 | Grade Response Model (GRM) | 40 | 1000 | 5 |
| 31 | Grade Response Model (GRM) | 40 | 1000 | 7 |
| 32 | Grade Response Model (GRM) | 40 | 1000 | 9 |

**ตารางที่** 3.1 (ต่อ)

| ลำดับที่ | โมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค | ความยาวแบบวัด | ขนาดกลุ่มตัวอย่าง | จำนวนรายการคำตอบ |
|---|---|---|---|---|
| 33 | Grade Response Model (GRM) | 40 | 2000 | 3 |
| 34 | Grade Response Model (GRM) | 40 | 2000 | 5 |
| 35 | Grade Response Model (GRM) | 40 | 2000 | 7 |
| 36 | Grade Response Model (GRM) | 40 | 2000 | 9 |
| 37 | Generalized Partial Credit Model (GPCM) | 10 | 500 | 3 |
| 38 | Generalized Partial Credit Model (GPCM) | 10 | 500 | 5 |
| 39 | Generalized Partial Credit Model (GPCM) | 10 | 500 | 7 |
| 40 | Generalized Partial Credit Model (GPCM) | 10 | 500 | 9 |
| 41 | Generalized Partial Credit Model (GPCM) | 10 | 1000 | 3 |
| 42 | Generalized Partial Credit Model (GPCM) | 10 | 1000 | 5 |
| 43 | Generalized Partial Credit Model (GPCM) | 10 | 1000 | 7 |
| 44 | Generalized Partial Credit Model (GPCM) | 10 | 1000 | 9 |
| 45 | Generalized Partial Credit Model (GPCM) | 10 | 2000 | 3 |
| 46 | Generalized Partial Credit Model (GPCM) | 10 | 2000 | 5 |
| 47 | Generalized Partial Credit Model (GPCM) | 10 | 2000 | 7 |
| 48 | Generalized Partial Credit Model (GPCM) | 10 | 2000 | 9 |
| 49 | Generalized Partial Credit Model (GPCM) | 20 | 500 | 3 |
| 50 | Generalized Partial Credit Model (GPCM) | 20 | 500 | 5 |
| 51 | Generalized Partial Credit Model (GPCM) | 20 | 500 | 7 |
| 52 | Generalized Partial Credit Model (GPCM) | 20 | 500 | 9 |
| 53 | Generalized Partial Credit Model (GPCM) | 20 | 1000 | 3 |
| 54 | Generalized Partial Credit Model (GPCM) | 20 | 1000 | 5 |
| 55 | Generalized Partial Credit Model (GPCM) | 20 | 1000 | 7 |
| 56 | Generalized Partial Credit Model (GPCM) | 20 | 1000 | 9 |
| 57 | Generalized Partial Credit Model (GPCM) | 20 | 2000 | 3 |
| 58 | Generalized Partial Credit Model (GPCM) | 20 | 2000 | 5 |
| 59 | Generalized Partial Credit Model (GPCM) | 20 | 2000 | 7 |
| 60 | Generalized Partial Credit Model (GPCM) | 20 | 2000 | 9 |
| 61 | Generalized Partial Credit Model (GPCM) | 40 | 500 | 3 |
| 62 | Generalized Partial Credit Model (GPCM) | 40 | 500 | 5 |
| 63 | Generalized Partial Credit Model (GPCM) | 40 | 500 | 7 |
| 64 | Generalized Partial Credit Model (GPCM) | 40 | 500 | 9 |
| 65 | Generalized Partial Credit Model (GPCM) | 40 | 1000 | 3 |
| 66 | Generalized Partial Credit Model (GPCM) | 40 | 1000 | 5 |
| 67 | Generalized Partial Credit Model (GPCM) | 40 | 1000 | 7 |
| 68 | Generalized Partial Credit Model (GPCM) | 40 | 1000 | 9 |

**ตารางที่** 3.1 (ต่อ)

| ลำดับที่ | โมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค | ความยาวแบบวัด | ขนาดกลุ่มตัวอย่าง | จำนวนรายการคำตอบ |
|---|---|---|---|---|
| 69 | Generalized Partial Credit Model (GPCM) | 40 | 2000 | 3 |
| 70 | Generalized Partial Credit Model (GPCM) | 40 | 2000 | 5 |
| 71 | Generalized Partial Credit Model (GPCM) | 40 | 2000 | 7 |
| 72 | Generalized Partial Credit Model (GPCM) | 40 | 2000 | 9 |

โดยในแต่ละสถานการณ์ที่จำลองข้อมูลจะมี 30 โมเดลซึ่งเท่ากับจำนวนครั้งในการกระทำซ้ำ โมเดลนี้เรียกว่า โมเดลจำลอง (Generating Model: GM) ซึ่งข้อมูลที่ได้จากการจำลองนี้จะนำไปใช้ในการวิเคราะห์ต่อไป

โปรแกรม WINGEN ของ Hambleton และ Han (2007) ที่ใช้ในการจำลองข้อมูลนั้น เป็นโปรแกรมที่พัฒนาขึ้นเพื่อตอบสนองการศึกษาวิเคราะห์เกี่ยวกับทฤษฎีการตอบสนองข้อสอบ (Item Response Theory) ซึ่งใช้แนวทางในการศึกษาด้วยเทคนิคมอนติคาร์โล คือการจำลองข้อมูลการตอบข้อคำถามเพื่อให้ได้ค่าพารามิเตอร์ที่แท้จริงตามที่ต้องการ ในการจำลองข้อมูลประกอบด้วย 3 ขั้นตอน ได้แก่

1. การจำลองข้อมูลค่าพารามิเตอร์ความสามารถผู้สอบ (Generating in ability parameter values)
2. การจำลองข้อมูลค่าพารามิเตอร์ข้อคำถาม (Generating in item parameter values)
3. การจำลองข้อมูลการตอบข้อคำถาม (Simulating item response data)

ซึ่งในการดำเนินการทั้ง 3 ขั้นตอนดังกล่าวนี้ สามารถกระทำแยกกันได้ในแต่ละขั้นตอนและทั้ง 3 ขั้นตอนนี้ดำเนินการผ่านหน้าต่างโปรแกรมเพียงหน้าต่างเดียว (one-stop interface screen) ซึ่งรายละเอียดในการจำลองข้อมูลในแต่ละขั้นตอนเป็นดังนี้

**1.1 การจำลองข้อมูลค่าพารามิเตอร์ความสามารถผู้สอบ (Generating in ability parameter values)**

ในขั้นตอนนี้ มีการระบุขนาดกลุ่มตัวอย่างหรือจำนวนผู้ตอบข้อคำถาม ลักษณะการแจกแจงของค่าพารามิเตอร์ความสามารถของผู้สอบ ($\theta$) ซึ่งในการวิจัยนี้ผู้วิจัยเลือกใช้ขนาดกลุ่มตัวอย่าง 3 ขนาด คือ 500, 1000, 2000 คน และกำหนดให้ค่าพารามิเตอร์ความสามารถของผู้สอบ ($\theta$) มีการแจกแจงแบบปกติ ที่มีค่าเฉลี่ย (*M*) เป็น 0 และส่วนเบี่ยงเบนมาตรฐาน (*SD*) เป็น 1

**1.2 การจำลองข้อมูลค่าพารามิเตอร์ข้อคำถาม** (Generating in item parameter values)

ในขั้นตอนนี้ มีการระบุความยาวของแบบวัด จำนวนรายการคำตอบ โมเดลทฤษฎีการตอบสนองข้อสอบ และลักษณะการแจกแจงของค่าพารามิเตอร์ข้อคำถาม ซึ่งในการวิจัยนี้ผู้วิจัยเลือกใช้ความยาวของแบบวัด (test length) 3 ขนาด คือ 10, 20, 40 ข้อ จำนวนรายการคำตอบ 4 ขนาดคือ 3, 5, 7, 9 รายการ โมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค (Polytomous item response model) 2 โมเดลคือ Grade Response Model (GRM) และ Generalized Partial Credit Model (GPCM) และลักษณะการแจกแจงของค่าพารามิเตอร์ข้อคำถามซึ่งมีรายละเอียดจำแนก ดังตารางที่ 3.2

**ตารางที่ 3.2 ลักษณะการแจกแจงของค่าพารามิเตอร์ข้อคำถาม**

| **โมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค** | **ลักษณะการแจกแจงของค่าพารามิเตอร์ข้อคำถาม** |
|---|---|
| Grade Response Model (GRM) | a มีการแจกแจงแบบ Lognormal ที่มีค่าเฉลี่ย 0 ส่วนเบี่ยงเบนมาตรฐาน 0.5 |
| | b มีการแจกแจงแบบ Uniform ที่มีค่าน้อยที่สุด -2 ค่ามากที่สุด 1 |
| Generalized Partial Credit Model (GPCM) | a มีการแจกแจงแบบ Lognormal ที่มีค่าเฉลี่ย 0 ส่วนเบี่ยงเบนมาตรฐาน 0.5 |
| | b มีการแจกแจงแบบ Uniform ที่มีค่าน้อยที่สุด -2 ค่ามากที่สุด 1 |

**1.3 การจำลองข้อมูลการตอบข้อคำถาม** (Simulating item response data)

ในขั้นตอนนี้ ต้องมีการอ้างอิงแฟ้มข้อมูล (Files) ค่าพารามิเตอร์ความสามารถผู้สอบและค่าพารามิเตอร์ข้อคำถาม จากขั้นตอนที่ 1 และ 2 พร้อมทั้งระบุจำนวนครั้งในการทำซ้ำ ซึ่งในการศึกษาวิจัยนี้ผู้วิจัยใช้การกระทำซ้ำ 30 ครั้งในแต่ละสถานการณ์

**ตอนที่ 2 การวิเคราะห์ข้อมูลที่ได้จากการจำลองข้อมูล**

การวิเคราะห์ข้อมูลในการศึกษาวิจัยครั้งนี้ใช้ดัชนีความสอดคล้องของข้อคำถาม (item fit index) 2 ชนิดคือ ดัชนี Generalized $S-\chi^2$ และดัชนี PARSCALE $G^2$ ซึ่งหลังจากที่ได้

ทำการจำลองข้อมูลตามสถานการณ์ที่ใช้ในการศึกษาแล้วจะทำให้ผู้วิจัยได้โมเดลจำลอง (Generating Model: GM) มา 30 โมเดลในแต่ละสถานการณ์ ดังนั้นในการศึกษาวิจัยนี้จึงมีโมเดลจำลอง (Generating Model: GM) ทั้งหมด 30 x 72 = 2160 โมเดล ซึ่งในขั้นตอนต่อไปผู้วิจัยดำเนินการวิเคราะห์ข้อมูลตามลำดับดังต่อไปนี้

1. วิเคราะห์ข้อมูลเบื้องต้นเพื่อทำให้ทราบลักษณะของข้อมูลที่ได้จำลองขึ้นในแต่ละสถานการณ์ที่ทำการศึกษา โดยพิจารณาค่าสถิติพื้นฐานได้แก่ ค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐาน ของค่าพารามิเตอร์ความสามารถผู้สอบ ($\theta$) และค่าพารามิเตอร์ข้อคำถาม (a,b)
2. จากโมเดลจำลอง (Generating Model : GM) ทำการตรวจสอบข้อมูลที่ได้จำลองขึ้นในแต่ละสถานการณ์ เพื่อตรวจสอบว่าข้อมูลการตอบแบบวัดเป็นไปตามเงื่อนไขที่กำหนดไว้หรือไม่ โดยพิจารณาใน 2 หัวข้อ ดังนี้
    2.1 พิจารณาการแจกแจงของค่าพารามิเตอร์ความสามารถผู้สอบ ($\theta$) จากรูปหลายเหลี่ยมความถี่ (Histogram) ในโปรแกรม WINGEN เพื่อตรวจสอบว่าข้อมูลการตอบแบบวัดเป็นไปตามเงื่อนไขที่ว่าค่าพารามิเตอร์ความสามารถของผู้ตอบแบบวัดมีการแจกแจงแบบปกติหรือไม่
    2.2 พิจารณาการแจกแจงของค่าพารามิเตอร์ข้อคำถามจากการทดสอบทางสถิติโดยใช้ค่าสถิติ Kolmogorov-Smirnov Z เพื่อตรวจสอบว่า ค่าพารามิเตอร์ข้อคำถาม a มีการแจกแจงแบบ Lognormal ที่มีค่าเฉลี่ย 0 ส่วนเบี่ยงเบนมาตรฐาน 0.5 หรือไม่ และค่าพารามิเตอร์ข้อคำถาม b มีการแจกแจงแบบเอกรูป (Uniform distribution) ที่มีค่าน้อยที่สุด -2 ค่ามากที่สุด 1 หรือไม่ ซึ่งเงื่อนไขดังกล่าวนี้ได้กำหนดขึ้นโดยระบุลักษณะการแจกแจงของค่าพารามิเตอร์ข้อคำถามในการจำลองข้อมูลด้วยโปรแกรม WINGEN
3. ทำการปรับเทียบโมเดล ซึ่งโมเดลที่ได้จากการปรับเทียบ (Calibrate) นี้ เรียกว่าโมเดลเทียบมาตรฐาน (Calibrating Model: CM) ทำเพื่อปรับเทียบหรือประมาณค่าพารามิเตอร์ในโมเดลจำลอง (Generating Model: GM) เพื่อประโยชน์ในการคำนวณหาค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) กับ อำนาจการทดสอบ (Power of the test) เนื่องจากถ้า GM กับ CM เป็นโมเดลที่ตรงกัน (เช่นเป็น GRM เหมือนกัน) จะประเมินประสิทธิภาพโดยดูที่ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) แต่ถ้า CM ไม่ตรงกับ GM (เช่น GM เป็น GRM แต่ CM เป็น GPCM) จะใช้ในการคำนวณอำนาจการทดสอบ (Power of the test)

4. คำนวณค่าดัชนี PARSCALE $G^2$ ด้วยโปรแกรม PARSCALE และคำนวณค่าดัชนี Generalized $S-\chi^2$ ด้วย IRTFIT macros ซึ่งประมวลผลบนโปรแกรม SAS จากนั้นจึงพิจารณาหาข้อคำถามที่ไม่สอดคล้อง (misfit) กับโมเดลทฤษฎีการตอบสนองข้อสอบในแต่ละสถานการณ์ที่ทำการศึกษา โดยการทดสอบสมมติฐานทางสถิติเกี่ยวกับความสอดคล้องของข้อคำถามดังสมมติฐาน

$H_0$ : ข้อคำถามมีความสอดคล้องกับโมเดลทฤษฎีการตอบสนองข้อสอบ

$H_1$ : ข้อคำถามไม่มีความสอดคล้องกับโมเดลทฤษฎีการตอบสนองข้อสอบ

เกณฑ์การตัดสินใจพิจาณาจากค่า p-value ที่ได้จากโปรแกรม ซึ่งถ้าค่า p-value ที่ได้มีค่าน้อยกว่า .05 แสดงว่าข้อคำถามข้อนั้นไม่สอดคล้อง (misfit) กับโมเดลทฤษฎีการตอบสนองข้อสอบซึ่งจะทำการเก็บรวบรวมข้อมูลโดยนับจำนวนข้อคำถามที่ถูกบ่งชี้ว่าไม่สอดคล้อง (misfit) แล้วรวมไว้เป็นผลรวมของจำนวนข้อคำถามที่ไม่สอดคล้อง (misfit) กับโมเดลทฤษฎีการตอบสนองข้อสอบ ต่อจากนั้นจึงนำผลคูณระหว่างความยาวแบบวัดกับจำนวนครั้งในการทำซ้ำมาหารผลรวมข้างต้น ซึ่งในกรณีที่โมเดลจำลอง (Generating Model: GM)กับโมเดลเทียบมาตรฐาน (Calibrating Model: CM) เป็นโมเดลที่ตรงกัน (เช่นเป็น GRM เหมือนกัน) จะเป็นการคำนวณหาค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ในแต่ละสถานการณ์ทั้ง 72 สถานการณ์ที่ทำการศึกษา การคำนวณค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ในการวิจัยครั้งนี้สรุปได้ดังแผนภาพที่ 3.1

**แผนภาพที่ 3.1 การคำนวณค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนีความสอดคล้องของข้อคำถามในแต่ละสถานการณ์ที่ทำการศึกษา**

**ค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนีความสอดคล้องของข้อคำถามในแต่ละสถานการณ์ที่ทำการศึกษา** ***(เมื่อ GM และ CM เป็นโมเดลที่ตรงกัน)***

$$= \frac{\text{ผลรวมของจำนวนข้อคำถามที่ไม่สอดคล้อง } \mathbf{(misfit)} \text{ กับโมเดลทฤษฎีการตอบสนองข้อสอบ}}{\text{ความยาวแบบวัด} * \text{จำนวนครั้งในการทำซ้ำคือ } \mathbf{30}}$$

จากแผนภาพที่ 3.1 ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) คือสัดส่วน (Proportion) ของจำนวนข้อคำถามที่ถูกบ่งชี้ว่าไม่สอดคล้องกับโมเดลทฤษฎีการตอบสนองข้อสอบที่ใช้ในการวิเคราะห์ ในกรณีที่โมเดลจำลอง (Generating Model: GM)

กับโมเดลเทียบมาตรฐาน (Calibrating Model: CM) เป็นโมเดลที่ตรงกัน (เช่นเป็น GRM เหมือนกัน) ตัวอย่างเช่น การคำนวณค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนี PARSCALE $G^2$ ในสถานการณ์ที่ 1 ซึ่งเป็นโมเดลจำลอง (GM) ชนิด GRM ที่มีความยาวแบบวัด 10 ข้อ ขนาดกลุ่มตัวอย่าง 500 คน จำนวนรายการคำตอบ 3 รายการ จะคำนวณค่าความคลาดเคลื่อนประเภทที่ 1 โดยการปรับเป็นโมเดลเทียบมาตรฐาน (CM) ชนิด GRM และเมื่อทำการหาผลรวมของจำนวนข้อคำถามที่ไม่สอดคล้อง (misfit) กับโมเดลทฤษฎีการตอบสนองข้อสอบในแต่ละรอบการทำซ้ำทั้ง 30 ครั้งพบว่า จำนวนข้อคำถามที่ไม่สอดคล้อง (misfit) กับโมเดลทฤษฎีการตอบสนองข้อสอบในแต่ละรอบการทำซ้ำมีทั้งหมด 164 ข้อ และจำนวนข้อคำถามทั้งหมดในสถานการณ์ที่ 1 คือ 10*30 = 300 ข้อ ดังนั้นค่าความคลาดเคลื่อนประเภทที่ 1 ในสถานการณ์ที่ 1 คือ 164/300 = 0.5467

ในการคำนวณอำนาจการทดสอบ (Power of the test) มีลักษณะเช่นเดียวกันกับการคำนวณค่าความคลาดเคลื่อนประเภทที่ 1 กล่าวคือจะทำการเก็บรวบรวมข้อมูลโดยนับจำนวนข้อคำถามที่ถูกบ่งชี้ว่าไม่สอดคล้อง (misfit) แล้วรวมไว้เป็นผลรวมของจำนวนข้อคำถามที่ไม่สอดคล้อง (misfit) กับโมเดลทฤษฎีการตอบสนองข้อสอบ ต่อจากนั้นจึงนำผลคูณระหว่างความยาวแบบวัดกับจำนวนครั้งในการทำซ้ำมาหารผลรวมข้างต้นเพื่อคำนวณหาอำนาจการทดสอบ ซึ่งในกรณีที่โมเดลจำลอง (Generating Model: GM)กับโมเดลเทียบมาตรฐาน (Calibrating Model: CM) เป็นโมเดลที่ไม่ตรงกัน (เช่น GM เป็น GRM แต่ CM เป็น GPCM) จะเป็นการคำนวณหาอำนาจการทดสอบ (Power of the test) ของดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ในแต่ละสถานการณ์ทั้ง 72 สถานการณ์ที่ทำการศึกษา การคำนวณอำนาจการทดสอบ (Power of the test) ในการวิจัยครั้งนี้สรุปได้ดังแผนภาพที่ 3.2

**แผนภาพที่ 3.2 การคำนวณอำนาจการทดสอบ (Power of the test) ของดัชนีความสอดคล้องของข้อคำถามในแต่ละสถานการณ์ที่ทำการศึกษา**

**อำนาจการทดสอบของดัชนีความสอดคล้องของข้อคำถามในแต่ละสถานการณ์ที่ทำการศึกษา**
***(เมื่อ GM และ CM เป็นโมเดลที่ไม่ตรงกัน)***

$$= \frac{\text{ผลรวมของจำนวนข้อคำถามที่ไม่สอดคล้อง } \mathbf{(misfit)} \text{ กับโมเดลทฤษฎีการตอบสนองข้อสอบ}}{\text{ความยาวแบบวัด} * \text{จำนวนครั้งในการทำซ้ำคือ } \mathbf{30}}$$

จากแผนภาพที่ 3.2 อำนาจการทดสอบ (Power of the test) คือสัดส่วน (Proportion) ของจำนวนข้อคำถามที่ถูกบ่งชี้ว่าไม่สอดคล้องกับโมเดลทฤษฎีการตอบสนองข้อสอบที่ใช้ในการวิเคราะห์ ในกรณีที่โมเดลจำลอง (Generating Model: GM) กับโมเดลเทียบมาตรฐาน (Calibrating Model: CM) เป็นโมเดลที่ไม่ตรงกัน (เช่น GM เป็น GRM แต่ CM เป็น GPCM) ตัวอย่างเช่น การคำนวณอำนาจการทดสอบของดัชนี PARSCALE $G^2$ ในสถานการณ์ที่ 1 ซึ่งเป็นโมเดลจำลอง (GM) ชนิด GRM ที่มีความยาวแบบวัด 10 ข้อ ขนาดกลุ่มตัวอย่าง 500 คน จำนวนรายการคำตอบ 3 รายการ จะคำนวณอำนาจการทดสอบโดยการปรับเป็นโมเดลเทียบมาตรฐาน (CM) ชนิด GPCM และเมื่อทำการหาผลรวมของจำนวนข้อคำถามที่ไม่สอดคล้อง (misfit) กับโมเดลทฤษฎีการตอบสนองข้อสอบในแต่ละรอบการทำซ้ำทั้ง 30 ครั้งพบว่า จำนวนข้อคำถามที่ไม่สอดคล้อง (misfit) กับโมเดลทฤษฎีการตอบสนองข้อสอบในแต่ละรอบการทำซ้ำมีทั้งหมด 181 ข้อ และจำนวนข้อคำถามทั้งหมดในสถานการณ์ที่ 1 คือ 10*30 = 300 ข้อ ดังนั้นค่าความคลาดเคลื่อนประเภทที่ 1 ในสถานการณ์ที่ 1 คือ 181/300 = 0.6033

รายละเอียดในการพิจารณาเพื่อใช้ในการคำนวณค่าความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบที่กล่าวมาทั้งหมดนั้น แสดงในตารางที่ 3.3

**ตารางที่ 3.3 การคำนวณหาค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) กับอำนาจการทดสอบ (Power of the test) ของดัชนีความสอดคล้องของข้อคำถาม**

| Calibrating Model (CM) | Generating Model (GM) | |
|---|---|---|
| | GRM | GPCM |
| GRM | **Type I error** (สัดส่วนของจำนวนข้อคำถามที่ถูกบ่งชี้ว่าไม่สอดคล้องกับโมเดลทฤษฎีการตอบสนองข้อสอบที่ใช้ในการวิเคราะห์เมื่อ GM และ CM เป็นโมเดล GRM แบบเดียวกัน) | **Power of the test** (สัดส่วนของจำนวนข้อคำถามที่ถูกบ่งชี้ว่าไม่สอดคล้องกับโมเดลทฤษฎีการตอบสนองข้อสอบที่ใช้ในการวิเคราะห์เมื่อ GM เป็นโมเดล GPCM และ CM เป็นโมเดล GRM) |
| GPCM | **Power of the test** (สัดส่วนของจำนวนข้อคำถามที่ถูกบ่งชี้ว่าไม่สอดคล้องกับโมเดลทฤษฎีการตอบสนอง | **Type I error** (สัดส่วนของจำนวนข้อคำถามที่ถูกบ่งชี้ว่าไม่สอดคล้องกับโมเดลทฤษฎีการตอบสนอง |

**ตารางที่** 3.3 (ต่อ)

| Calibrating Model (CM) | Generating Model (GM) | |
|---|---|---|
| | GRM | GPCM |
| | ข้อสอบที่ใช้ในการวิเคราะห์เมื่อ GM เป็นโมเดล GRM และ CM เป็นโมเดล GPCM) | ข้อสอบที่ใช้ในการวิเคราะห์เมื่อ GM และ CM เป็นโมเดล GPCM แบบเดียวกัน) |

จากตารางที่ 3.3 พบว่า ในการคำนวณค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) นั้นอาศัยแนวคิดจากนิยามของความคลาดเคลื่อนประเภทที่ 1 ในการวิจัยนี้คือ การบ่งชี้ว่าข้อคำถามข้อนั้นไม่สอดคล้อง (misfit) ทั้ง ๆ ที่ข้อคำถามข้อนั้นมีความสอดคล้อง (fit) กับโมเดลทฤษฎีการตอบสนองข้อสอบ ดังนั้น ในการคำนวณค่าความคลาดเคลื่อนประเภทที่ 1 จึงจะคำนวณตามแผนภาพที่ 3.1 เมื่อโมเดลจำลอง (GM) เป็นโมเดลทฤษฎีการตอบสนองข้อสอบชนิดเดียวกันกับโมเดลเทียบมาตรฐาน (CM)

นอกจากนี้ในการคำนวณอำนาจการทดสอบ (Power of the test) ได้อาศัยแนวคิดจากนิยามของอำนาจการทดสอบ ในการวิจัยนี้คือ การบ่งชี้ว่าข้อคำถามข้อนั้นไม่สอดคล้อง (misfit) ซึ่งในความเป็นจริงข้อคำถามข้อนั้นไม่มีความสอดคล้อง (misfit) กับโมเดลทฤษฎีการตอบสนองข้อสอบ ดังนั้น ในการคำนวณอำนาจการทดสอบจึงจะคำนวณตามแผนภาพที่ 3.2 เมื่อโมเดลจำลอง (GM) เป็นโมเดลทฤษฎีการตอบสนองข้อสอบคนละชนิดกันกับโมเดลเทียบมาตรฐาน (CM)

5. พิจารณาค่าความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ จำแนกตามโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาคที่ใช้ในการศึกษาว่ามีการเปลี่ยนแปลงไปตามความยาวแบบวัด ขนาดกลุ่มตัวอย่าง และจำนวนรายการคำตอบที่เปลี่ยนแปลงไปอย่างไร โดยมีการนำค่าความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ที่จำแนกตามโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาคในการศึกษา ไปวิเคราะห์เปรียบเทียบเพื่อทดสอบว่าเมื่อเงื่อนไขที่ใช้ในการศึกษา (ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ) เปลี่ยนแปลงไป จะส่งผลต่อค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนี PARSCALE $G^2$ อย่างมีนัยสำคัญทางสถิติหรือไม่ โดยใช้การทดสอบสมมติฐานทางสถิติที่ระดับความเชื่อมั่น 95 % เพื่อเปรียบเทียบค่าเฉลี่ยของค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) และอำนาจการทดสอบของดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ในแต่ละเงื่อนไขที่ใช้ในการศึกษาจำแนก

ตามโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค โดยใช้การวิเคราะห์ความแปรปรวนทางเดียว (One-way ANOVA) และทำการทดสอบเปรียบเทียบพหุคูณ (Multiple Comparison) โดยถ้าความแปรปรวนของประชากรในแต่ละกลุ่มมีค่าความแปรปรวนไม่แตกต่างกันจะใช้วิธี LSD เนื่องจากวิธี LSD เป็นวิธีที่ควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ตามเกณฑ์ที่กำหนดในภาพรวมทั้งหมดของการทดลอง (ปุณยนุช พินชู, 2548) แต่ถ้าความแปรปรวนของประชากรในแต่ละกลุ่มมีค่าความแปรปรวนแตกต่างกันจะใช้วิธี Dunnett's T3

6. เปรียบเทียบประสิทธิภาพของดัชนี Generalized $S-\chi^2$ กับดัชนี PARSCALE $G^2$ โดยใช้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) กับ อำนาจการทดสอบ (Power of the test) เป็นเกณฑ์ในการพิจารณา

6.1 เปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามสองชนิดคือ Generalized $S-\chi^2$ และ PARSCALE $G^2$ ตามเงื่อนไขของ Kang และ Chen (2008) การเปรียบเทียบโดยวิธีนี้ มีเงื่อนไขว่าจะพิจารณาที่ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ก่อน ซึ่งถ้าค่าความคลาดเคลื่อนประเภทที่ 1 ที่คำนวณได้มีค่าที่มาก (Type I error rate are considerably inflated) ก็จะไม่พิจารณาอำนาจการทดสอบ (Power of the test) ทั้งนี้เนื่องจาก Kang และ Chen มีแนวคิดว่า ดัชนีความสอดคล้องของข้อคำถามที่ดีนั้น ควรมีโอกาสในการบ่งชี้ข้อคำถามที่มีความสอดคล้อง (fit) กับโมเดลว่าเป็นข้อคำถามที่ไม่สอดคล้อง (misfit) กับโมเดลให้น้อยที่สุด

ดังนั้น ผู้วิจัยทำการทดสอบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามที่ระดับความเชื่อมั่น 95% ดังนั้นเกณฑ์ในการพิจารณาเพื่อใช้เปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามจึงมี 2 ขั้นตอนตามลำดับ คือ

1) พิจารณากรณีที่ GM กับ CM เป็นโมเดลที่ตรงกันคือการพิจารณาที่ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ก่อนโดยค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ที่ได้จากโปรแกรมทั้งสองคือโปรแกรม PARSCALE และ IRTFIT macros ซึ่งประมวลผลบนโปรแกรม SAS ควรมีค่ามากกว่าหรือเท่ากับ 0.05 ซึ่งเป็นระดับนัยสำคัญของการทดสอบว่าข้อคำถามแต่ละข้อมีความสอดคล้องกับโมเดลทฤษฎีการตอบสนองข้อสอบหรือไม่ ถ้าหากค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ที่ได้จากโปรแกรมทั้งสองมีค่าน้อยกว่า .05 แสดงว่าข้อคำถามข้อนั้นไม่สอดคล้องจะทำการนับจำนวนข้อคำถามที่ไม่สอดคล้องนั้นเพื่อนำไปสู่การคำนวณสัดส่วนของจำนวนข้อคำถามที่

ถูกบ่งชี้ว่าไม่สอดคล้องกับโมเดลทฤษฎีการตอบสนองข้อสอบ เพื่อให้ได้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ซึ่งถ้าดัชนีชนิดใดมีค่าความคลาดเคลื่อนประเภทที่ 1 น้อยกว่าจะสรุปว่าดัชนีนั้นมีประสิทธิภาพในการบ่งชี้ความสอดคล้องของข้อคำถามมากกว่าดัชนีอีกชนิดหนึ่ง แต่ถ้าหากดัชนีทั้งสองชนิดมีค่าความคลาดเคลื่อนประเภทที่ 1 เท่ากันจึงจะพิจารณาอำนาจการทดสอบ (power of the test) ในขั้นตอนต่อไป

2) พิจารณากรณีที่ GM กับ CM เป็นโมเดลที่ไม่ตรงกันคือการพิจารณาที่อำนาจการทดสอบ (power of the test) โดยอำนาจการทดสอบ (power of the test) นี้เป็นค่าที่ได้จากโปรแกรมทั้งสองคือโปรแกรม PARSCALE และ IRTFIT macros ซึ่งประมวลผลบนโปรแกรม SAS ในกรณีที่ GM กับ CM เป็นโมเดลที่ไม่ตรงกัน การที่จะพิจารณาอำนาจการทดสอบ (power of the test) ต้องเป็นกรณีที่ดัชนีทั้งสองชนิดมีค่าความคลาดเคลื่อนประเภทที่ 1 เท่ากันเท่านั้น โดยถ้าดัชนีชนิดใดมีอำนาจการทดสอบมากกว่า จะสรุปว่าดัชนีนั้นมีประสิทธิภาพในการบ่งชี้ความสอดคล้องของข้อคำถามมากกว่าดัชนีอีกชนิดหนึ่ง

6.2 เปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิด คือ Generalized $S-\chi^2$ และ PARSCALE $G^2$ โดยใช้สถิติการวิเคราะห์ความแปรปรวนแบบ 2 ทาง (two-way ANOVA) โดยนำค่าความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบมาทำการวิเคราะห์ต่อไปโดยการทดสอบผลของชนิดของดัชนีความสอดคล้องของข้อคำถาม และผลของปฏิสัมพันธ์ระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามกับเงื่อนไขที่ใช้ในการศึกษา(ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ) ว่ามีนัยสำคัญทางสถิติหรือไม่ ซึ่งถ้าชนิดของดัชนีความสอดคล้องของข้อคำถามกับเงื่อนไขที่ใช้ในการศึกษา (ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ) นั้นมีอิทธิพลปฏิสัมพันธ์ (interaction effect) กันย่อมแสดงให้เห็นว่า ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดภายใต้เงื่อนไขที่ใช้ในการศึกษา (ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ) ที่ต่างกัน จะส่งผลให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) และอำนาจการทดสอบ (Power of the test) มีค่าแตกต่างกัน นำไปสู่การสรุปผลเกี่ยวกับประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามที่แตกต่างกัน ซึ่งในการวิเคราะห์ในขั้นตอนนี้ จำแนกได้ทั้งหมด 12 กรณี แบ่งออกเป็นการวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อน

ประเภทที่ 1 (Type I error) 6 กรณีและการวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบ (Power of the test) 6 กรณี ดังนี้

<u>การวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1</u>

1.การวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขความยาวแบบวัดที่ต่างกัน 3 ระดับ กรณีโมเดล GRM

2.การวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างที่ต่างกัน 3 ระดับ กรณีโมเดล GRM

3.การวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขจำนวนรายการคำตอบที่ต่างกัน 4 ระดับ กรณีโมเดล GRM

4.การวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขความยาวแบบวัดที่ต่างกัน 3 ระดับ กรณีโมเดล GPCM

5.การวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างที่ต่างกัน 3 ระดับ กรณีโมเดล GPCM

6.การวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขจำนวนรายการคำตอบที่ต่างกัน 4 ระดับ กรณีโมเดล GPCM

การวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบ

1.การวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขความยาวแบบวัดที่ต่างกัน 3 ระดับ กรณีโมเดล GRM

2.การวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างที่ต่างกัน 3 ระดับ กรณีโมเดล GRM

3.การวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขจำนวนรายการคำตอบที่ต่างกัน 4 ระดับ กรณีโมเดล GRM

4.การวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขความยาวแบบวัดที่ต่างกัน 3 ระดับ กรณีโมเดล GPCM

5.การวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างที่ต่างกัน 3 ระดับ กรณีโมเดล GPCM

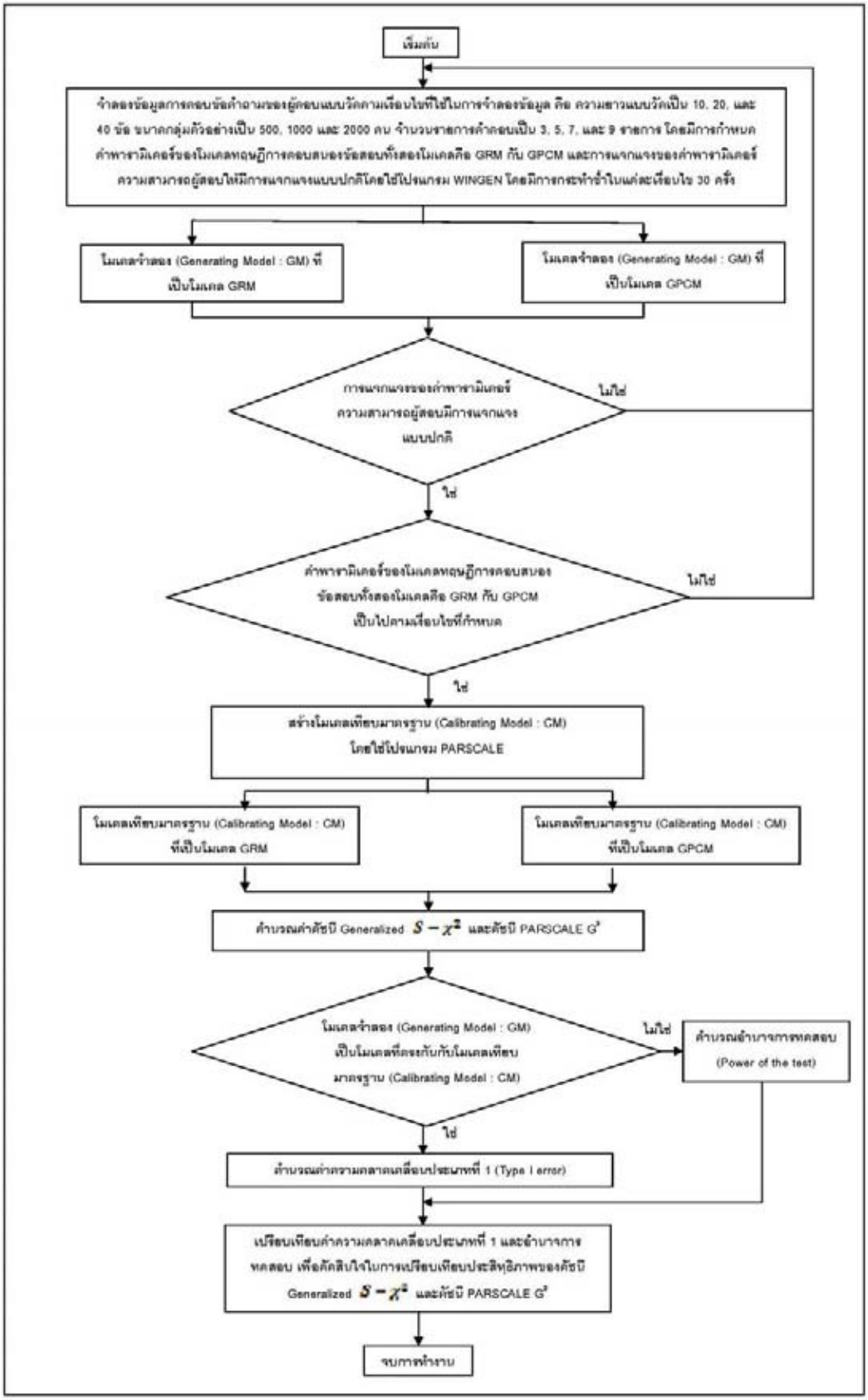
6.การวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขจำนวนรายการคำตอบที่ต่างกัน 4 ระดับ กรณีโมเดล GPCM

การวิเคราะห์ความแปรปรวนแบบ 2 ทาง (two-way ANOVA) ในการศึกษานี้ดำเนินการโดยนำค่าความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบมาทำการวิเคราะห์เพื่อทดสอบผลของชนิดของดัชนีความสอดคล้องของข้อคำถาม และผลของปฏิสัมพันธ์ระหว่าง

ชนิดของดัชนีความสอดคล้องของข้อคำถามกับเงื่อนไขที่ใช้ในการศึกษา(ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ) ว่ามีนัยสำคัญทางสถิติหรือไม่ ซึ่งถ้าชนิดของดัชนีความสอดคล้องของข้อคำถามกับเงื่อนไขที่ใช้ในการศึกษา (ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ) นั้นมีอิทธิพลปฏิสัมพันธ์ (interaction effect) กัน ผู้วิจัยจะวิเคราะห์ต่อไปด้วยเทคนิค Simple effect โดยศึกษาเปรียบเทียบค่าเฉลี่ยของค่าความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของดัชนีทั้งสองชนิดสำหรับแต่ละระดับของเงื่อนไขที่ใช้ในการศึกษา (ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ)ที่มีอิทธิพลปฏิสัมพันธ์กับชนิดของดัชนีความสอดคล้องของข้อคำถาม แต่จะไม่ทำการวิเคราะห์ต่อไปด้วยเทคนิค Simple effect ในกรณีเปรียบเทียบค่าเฉลี่ยของค่าความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบระหว่างแต่ละระดับของเงื่อนไขที่ใช้ในการศึกษาจำแนกตามชนิดของดัชนี เนื่องจากได้ทำการศึกษาวิเคราะห์ไปแล้วในขั้นตอนการพิจารณาค่าความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ จำแนกตามโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาคที่ใช้ในการศึกษา แต่ถ้าหากไม่มีอิทธิพลปฏิสัมพันธ์ (interaction effect) กันระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามกับเงื่อนไขที่ใช้ในการศึกษา จะทำการเปรียบเทียบค่าเฉลี่ยของค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) และอำนาจการทดสอบของดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ จำแนกตามโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค โดยไม่เปรียบเทียบค่าเฉลี่ยของค่าความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของดัชนีทั้งสองชนิดสำหรับแต่ละระดับของเงื่อนไขที่ใช้ในการศึกษา (ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ) แต่จะใช้การทดสอบ t-test กรณีกลุ่มตัวอย่างเป็นอิสระกัน (independent sample t-test)

จากวิธีการดำเนินงานวิจัยที่กล่าวมาทั้งหมดสามารถสรุปขั้นตอนการจำลองข้อมูลและการวิเคราะห์ข้อมูลดังแผนภาพที่ 3.3

**แผนภาพที่ 3.3 ขั้นตอนการจำลองข้อมูลและการวิเคราะห์ข้อมูล**



# บทที่ 4

# ผลการวิเคราะห์ข้อมูล

การศึกษาวิจัยครั้งนี้เป็นการวิจัยเชิงทดลองซึ่งใช้เทคนิคการจำลองแบบ (Simulation) โดยมีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถาม (item fit index) สำหรับโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค สองชนิดคือ Generalized $S-\chi^2$ และ PARSCALE $G^2$ ในสถานการณ์ ณ ความยาวแบบวัดเป็น 10, 20, และ 40 ข้อ ขนาดกลุ่มตัวอย่างเป็น 500, 1000, และ 2000 คน จำนวนรายการคำตอบเป็น 3, 5, 7, และ 9 รายการ ซึ่งเกณฑ์ที่นำมาใช้ในการประเมินประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามในการวิจัยนี้มี 2 ประการ ดังนี้

1. **ความคลาดเคลื่อนประเภทที่** 1 (Type I error : $\alpha$) เป็นโอกาสในการปฏิเสธสมมติฐานที่เป็นจริง ในการวิจัยนี้คือ การบ่งชี้ว่าข้อคำถามข้อนั้นไม่สอดคล้อง (misfit) ทั้ง ๆ ที่ข้อคำถามข้อนั้นมีความสอดคล้อง (fit) กับโมเดลทฤษฎีการตอบสนองข้อสอบ

2. **อำนาจการทดสอบ** (Power of the test : 1-$\beta$) เป็นโอกาสในการปฏิเสธสมมติฐานที่เป็นเท็จ ในการวิจัยนี้คือ การบ่งชี้ว่าข้อคำถามข้อนั้นไม่สอดคล้อง (misfit) ซึ่งในความเป็นจริงข้อคำถามข้อนั้นไม่มีความสอดคล้อง (misfit) กับโมเดลทฤษฎีการตอบสนองข้อสอบ

ขั้นตอนในการพิจารณาเพื่อประเมินประสิทธิภาพของของดัชนีความสอดคล้องของข้อคำถาม (item fit index) จะพิจารณาที่ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ก่อน ซึ่งถ้าดัชนีชนิดใดมีค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) น้อยกว่า ดัชนีนั้นจะเป็นดัชนีที่มีประสิทธิภาพในการบ่งชี้ความสอดคล้องของข้อคำถามกับโมเดลทฤษฎีการตอบสนองข้อสอบในสถานการณ์นั้น แต่ถ้าหากดัชนีทั้งสองมีค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) เท่ากันจึงจะพิจารณาอำนาจการทดสอบ (Power of the test) โดยถ้าดัชนีชนิดใดมีอำนาจการทดสอบที่มากกว่า ดัชนีนั้นจะเป็นดัชนีที่มีประสิทธิภาพในการบ่งชี้ความสอดคล้องของข้อคำถามกับโมเดลทฤษฎีการตอบสนองข้อสอบในสถานการณ์นั้น

ในการนำเสนอผลการวิเคราะห์ข้อมูลจึงมุ่งเน้นนำเสนอค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) และอำนาจการทดสอบ (Power of the test) ซึ่งเป็นเกณฑ์ที่ใช้ในการประเมินประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามเป็นหลัก นอกจากนี้ยังนำเสนอผลการวิเคราะห์เพื่อตรวจสอบข้อมูลที่ได้จำลองขึ้นในแต่ละสถานการณ์ ว่าข้อมูลมีลักษณะเป็นไป

ตามเงื่อนไขที่กำหนดไว้หรือไม่ ซึ่งการตรวจสอบข้อมูลที่ได้จำลองขึ้นในการศึกษาวิจัยด้วยเทคนิคการจำลองแบบ (Simulation) นั้นเป็นสิ่งจำเป็นและมีความสำคัญอย่างยิ่ง ดังนั้น ผลการวิเคราะห์ข้อมูลจึงนำเสนอออกเป็น 4 ตอน ได้แก่

ตอนที่ 1 ผลการวิเคราะห์ข้อมูลเบื้องต้นของข้อมูลที่ได้จำลองขึ้นในแต่ละสถานการณ์ที่ทำการศึกษา

ตอนที่ 2 ผลการวิเคราะห์ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error)

ตอนที่ 3 ผลการวิเคราะห์ค่าอำนาจการทดสอบ (Power of the test)

ตอนที่ 4 ผลการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามสองชนิดคือ Generalized $S-\chi^2$ และ PARSCALE $G^2$

ในการนำเสนอผลการวิเคราะห์ข้อมูลในบทนี้ ผู้วิจัยได้กำหนดสัญลักษณ์และอักษรย่อภาษาอังกฤษที่ใช้สื่อความหมายแทนชุดข้อมูลเงื่อนไขที่ทำการศึกษา และค่าสถิติ ดังนี้

| | | |
|---|---|---|
| $\theta$ | แทน | ค่าพารามิเตอร์ความสามารถผู้สอบ |
| a | แทน | ค่าความชัน หรือค่าอำนาจจำแนกของข้อคำถาม |
| b | แทน | ค่าระดับความยากของขั้นการตอบ |
| *M* | แทน | ค่าเฉลี่ย |
| *SD* | แทน | ส่วนเบี่ยงเบนมาตรฐาน |
| Min. | แทน | ค่าน้อยที่สุด |
| Max. | แทน | ค่ามากที่สุด |
| *Sk* | แทน | ความเบ้ |
| *Ku* | แทน | ความโด่ง |
| KS_Z | แทน | ค่าสถิติทดสอบ Kolmogorov-Smirnov Z |
| *F* | แทน | ค่าสถิติทดสอบ F |
| *p* | แทน | ค่า p-value ที่ใช้ในการทดสอบระดับนัยสำคัญของสมมติฐาน |
| Ln(a) | แทน | ค่าลอการิทึมธรรมชาติ (natural logarithm) ของ a |

**ตอนที่ 1 ผลการวิเคราะห์ข้อมูลเบื้องต้นของข้อมูลที่ได้จำลองขึ้นในแต่ละสถานการณ์ที่ทำการศึกษา**

1.1 ค่าสถิติพื้นฐานของค่าพารามิเตอร์ความสามารถผู้สอบ($\theta$) และค่าพารามิเตอร์ข้อคำถาม

การวิเคราะห์ข้อมูลในส่วนนี้ เป็นการวิเคราะห์ด้วยค่าสถิติพื้นฐานเพื่อบรรยายลักษณะของค่าพารามิเตอร์ความสามารถผู้สอบ($\theta$) และค่าพารามิเตอร์ข้อคำถามในแต่ละสถานการณ์ที่ทำการศึกษา ซึ่งประกอบด้วย ค่าเฉลี่ย (*M*) ส่วนเบี่ยงเบนมาตรฐาน (*SD*) โดยมีรายละเอียดดังตารางในภาคผนวก ข

รายละเอียดของค่าเฉลี่ย (*M*) ส่วนเบี่ยงเบนมาตรฐาน (*SD*) ที่ใช้พรรณนาค่าพารามิเตอร์ความสามารถผู้สอบ($\theta$) และค่าพารามิเตอร์ข้อคำถามในแต่ละสถานการณ์ที่ทำการศึกษา ดังปรากฏในภาคผนวก ข พบว่า ค่าเฉลี่ยของค่าความชันหรือค่าอำนาจจำแนกของข้อคำถาม (a) ทั้ง 72 สถานการณ์ (3X3X4X2 สถานการณ์) มีค่าตั้งแต่ 0.792 ถึง 1.434 และค่าเฉลี่ยของค่าระดับความยากของขั้นการตอบ ทั้ง 72 สถานการณ์ (3X3X4X2 สถานการณ์) มีค่าตั้งแต่ -0.742 ถึง 0.579 แสดงให้เห็นว่าข้อมูลที่จำลองขึ้นมีอำนาจจำแนกอยู่ในระดับที่เหมาะสม มีระดับความยากของขั้นการตอบอยู่ในระดับปานกลาง

นอกจากนี้ เมื่อพิจารณาค่าพารามิเตอร์ความสามารถผู้สอบ($\theta$) พบว่า ค่าเฉลี่ยของค่าพารามิเตอร์ความสามารถผู้สอบทั้ง 72 สถานการณ์ (3X3X4X2 สถานการณ์) มีค่าตั้งแต่ -0.109 ถึง 0.097 แสดงให้เห็นว่าข้อมูลที่จำลองขึ้นนั้นผู้ตอบข้อคำถามมีความสามารถระดับปานกลาง ซึ่งสามารถสรุปสาระสำคัญได้ดังในตารางที่ 4.1

**ตารางที่ 4.1 สรุปรายละเอียดของค่าเฉลี่ย (*M*) ส่วนเบี่ยงเบนมาตรฐาน (*SD*) ของค่าพารามิเตอร์ความสามารถผู้สอบ และค่าพารามิเตอร์ข้อคำถามในแต่ละสถานการณ์ที่ทำการศึกษา**

| ค่าพารามิเตอร์ | n | | Min. | | Max. | | *M* | |
|---|---|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| a | 72 | 72 | 0.792 | 0.281 | 1.434 | 1.666 | 1.117 | 0.575 |
| b | 72 | 72 | -0.742 | 0.657 | 0.579 | 1.000 | -0.493 | 0.865 |
| $\theta$ | 72 | 72 | -0.109 | 0.892 | 0.097 | 1.069 | -0.004 | 1.001 |

1.2 ลักษณะการแจกแจงของค่าพารามิเตอร์ความสามารถผู้สอบ($\theta$) และค่าพารามิเตอร์ข้อคำถาม

การวิเคราะห์ข้อมูลในส่วนนี้ เป็นการวิเคราะห์เพื่อตรวจสอบข้อมูลที่จำลองขึ้นว่ามีลักษณะเป็นไปตามเงื่อนไขที่กำหนดหรือไม่ โดยพิจารณาลักษณะการแจกแจงของค่าพารามิเตอร์ความสามารถผู้สอบ($\theta$) จากรูปหลายเหลี่ยมความถี่ (Histogram) ว่ามีการแจกแจงแบบปกติหรือไม่ และพิจารณาการแจกแจงของค่าพารามิเตอร์ข้อคำถามจากการทดสอบทางสถิติโดยใช้ค่าสถิติ Kolmogorov-Smirnov Z ว่าค่าพารามิเตอร์ข้อคำถามเป็นไปตามเงื่อนไขที่ได้กำหนดขึ้นก่อนการจำลองข้อมูลหรือไม่ ในการตรวจสอบการแจกแจงของค่าพารามิเตอร์โดยใช้ค่าสถิติ Kolmogorov-Smirnov Z ในโปรแกรมสำเร็จรูปทางสถิตินั้น ในปัจจุบันโปรแกรมสามารถตรวจสอบการแจกแจงได้ 4 ลักษณะ คือ 1.การแจกแจงแบบปกติ (Normal distribution) 2.การแจกแจงแบบเอกรูป (Uniform distribution) 3.การแจกแจงแบบปัวซง (Poisson distribution) 4.การแจกแจงแบบเอ็กซ์โพเนลเชียล (Exponential distribution) แต่เนื่องจากในโปรแกรมสำเร็จรูปที่ใช้ในการคำนวณค่าสถิติ Kolmogorov-Smirnov Z นั้น ไม่สามารถตรวจสอบการแจกแจงแบบ Lognormal โดยตรงได้ จึงได้อาศัยนิยามเกี่ยวกับการแจกแจงแบบ Lognormal ของตัวแปรทางสถิติเข้ามาช่วย ซึ่งนิยามดังกล่าว กล่าวว่า “ถ้า Ln (X) มีการแจกแจงแบบปกติ จะได้ว่าตัวแปรสุ่ม X มีการแจกแจงแบบ Lognormal” ดังนั้น ในการตรวจสอบการแจกแจงของค่าพารามิเตอร์ a ว่ามีการแจกแจงแบบ Lognormal หรือไม่ จึงใช้วิธีการตรวจสอบว่า Ln (a) มีการแจกแจงแบบปกติหรือไม่ ถ้า Ln (a) มีการแจกแจงแบบปกติ สามารถสรุปได้ว่า a มีการแจกแจงแบบ Lognormal

ค่าสถิติ Kolmogorov-Smirnov Z พร้อมทั้งค่า p-value ที่ใช้ในการทดสอบระดับนัยสำคัญของสมมติฐาน ($p$) ที่ใช้ในการทดสอบลักษณะการแจกแจงของค่าพารามิเตอร์ข้อคำถาม (a และ b) พบว่า ค่าสถิติ Kolmogorov-Smirnov Z ที่ใช้ในการทดสอบลักษณะการแจกแจงของค่าพารามิเตอร์ข้อคำถาม (a และ b) ไม่มีนัยสำคัญทางสถิติที่ระดับ .05 ($p > .05$) ในทุกสถานการณ์ แสดงว่า ค่าพารามิเตอร์ข้อคำถามของข้อมูลที่จำลองขึ้น มีการแจกแจงตามลักษณะที่กำหนดไว้ กล่าวคือ a มีการแจกแจงแบบ Lognormal และ b มีการแจกแจงเอกรูป (Uniform distribution) ด้วยระดับความเชื่อมั่น 95% ซึ่งรายละเอียดในแต่ละสถานการณ์ที่ทำการศึกษาสามารถพิจารณาได้จากตารางที่ 4.2

**ตารางที่ 4.2 ค่าสถิติทดสอบของค่าพารามิเตอร์ข้อคำถาม**

| สถานการณ์ที่ | Ln(a) | | b | |
|---|---|---|---|---|
| | KS_Z | *p* | KS_Z | *p* |
| 1 | 0.5370 | 0.9352 | 0.5560 | 0.9166 |
| 2 | 0.5727 | 0.8983 | 0.9399 | 0.3401 |

**ตารางที่** 4.2 (ต่อ)

| สถานการณ์ที่ | Ln(a) | | b | |
|---|---|---|---|---|
| | KS_Z | *p* | KS_Z | *p* |
| 3 | 0.5226 | 0.9476 | 0.6575 | 0.7803 |
| 4 | 0.4698 | 0.9800 | 0.7564 | 0.6164 |
| 5 | 0.5603 | 0.9121 | 0.9507 | 0.3266 |
| 6 | 0.8364 | 0.4862 | 0.5635 | 0.9086 |
| 7 | 0.5143 | 0.9541 | 1.3340 | 0.0569 |
| 8 | 0.5522 | 0.9206 | 1.0278 | 0.2414 |
| 9 | 0.4031 | 0.9969 | 0.6791 | 0.7457 |
| 10 | 0.6623 | 0.7727 | 1.3178 | 0.0620 |
| 11 | 0.7189 | 0.6795 | 0.7646 | 0.6026 |
| 12 | 0.4456 | 0.9887 | 0.9025 | 0.3894 |
| 13 | 0.5511 | 0.9217 | 0.9664 | 0.3078 |
| 14 | 0.4338 | 0.9918 | 0.8977 | 0.3960 |
| 15 | 0.4000 | 0.9972 | 0.7266 | 0.6666 |
| 16 | 0.7471 | 0.6321 | 0.6019 | 0.8618 |
| 17 | 0.5856 | 0.8828 | 1.1374 | 0.1504 |
| 18 | 0.8218 | 0.5091 | 0.6584 | 0.7788 |
| 19 | 0.6392 | 0.8086 | 0.8472 | 0.4695 |
| 20 | 0.5946 | 0.8714 | 0.6794 | 0.7451 |
| 21 | 0.6256 | 0.8288 | 0.6634 | 0.7709 |
| 22 | 0.6660 | 0.7669 | 0.4149 | 0.9953 |
| 23 | 0.3839 | 0.9985 | 1.3340 | 0.0569 |
| 24 | 0.5129 | 0.9551 | 0.8026 | 0.5398 |
| 25 | 0.5073 | 0.9591 | 1.2108 | 0.1066 |
| 26 | 0.5610 | 0.9113 | 0.6273 | 0.8263 |
| 27 | 0.5830 | 0.8859 | 1.3436 | 0.0541 |
| 28 | 0.7741 | 0.5868 | 0.7678 | 0.5973 |
| 29 | 0.5550 | 0.9177 | 0.8624 | 0.4467 |
| 30 | 0.4441 | 0.9892 | 0.7425 | 0.6398 |
| 31 | 0.4856 | 0.9724 | 0.5512 | 0.9216 |
| 32 | 0.5000 | 0.9640 | 0.7007 | 0.7100 |
| 33 | 0.4363 | 0.9912 | 0.5331 | 0.9388 |
| 34 | 0.9135 | 0.3743 | 1.0514 | 0.2189 |
| 35 | 0.6478 | 0.7953 | 1.0352 | 0.2341 |
| 36 | 0.4920 | 0.9688 | 0.5313 | 0.9404 |
| 37 | 0.4939 | 0.9677 | 0.8033 | 0.5388 |
| 38 | 0.3610 | 0.9995 | 0.7141 | 0.6876 |

**ตารางที่** 4.2 (ต่อ)

| สถานการณ์ที่ | Ln(a) | | b | |
|---|---|---|---|---|
| | KS_Z | *p* | KS_Z | *p* |
| 39 | 0.5523 | 0.9205 | 0.7795 | 0.5779 |
| 40 | 0.5803 | 0.8893 | 0.8370 | 0.4853 |
| 41 | 0.4665 | 0.9814 | 0.6935 | 0.7221 |
| 42 | 0.4182 | 0.9948 | 0.4030 | 0.9969 |
| 43 | 0.4581 | 0.9847 | 1.0231 | 0.2460 |
| 44 | 0.6481 | 0.7949 | 0.8402 | 0.4803 |
| 45 | 0.8376 | 0.4843 | 0.7199 | 0.6779 |
| 46 | 0.8272 | 0.5006 | 0.7388 | 0.6460 |
| 47 | 0.4219 | 0.9942 | 1.3184 | 0.1618 |
| 48 | 0.8552 | 0.4575 | 0.8619 | 0.4474 |
| 49 | 0.5781 | 0.8919 | 0.9135 | 0.3744 |
| 50 | 0.6735 | 0.7548 | 1.1709 | 0.1288 |
| 51 | 0.7040 | 0.7046 | 0.6088 | 0.8524 |
| 52 | 0.6476 | 0.7957 | 0.8357 | 0.4873 |
| 53 | 0.5660 | 0.9059 | 0.5516 | 0.9212 |
| 54 | 0.6147 | 0.8443 | 0.6583 | 0.7791 |
| 55 | 0.5650 | 0.9070 | 0.6950 | 0.7195 |
| 56 | 0.5331 | 0.9388 | 0.9716 | 0.3017 |
| 57 | 0.6039 | 0.8591 | 0.7668 | 0.5990 |
| 58 | 0.5211 | 0.9489 | 0.9426 | 0.3367 |
| 59 | 0.4924 | 0.9686 | 0.7324 | 0.6568 |
| 60 | 0.5896 | 0.8778 | 1.2906 | 0.0715 |
| 61 | 0.8849 | 0.4139 | 0.9928 | 0.2778 |
| 62 | 0.6878 | 0.7314 | 0.7032 | 0.7060 |
| 63 | 0.5659 | 0.9060 | 0.7928 | 0.5559 |
| 64 | 0.4011 | 0.9971 | 0.5128 | 0.9552 |
| 65 | 0.6372 | 0.8115 | 1.2489 | 0.1884 |
| 66 | 0.4068 | 0.9964 | 0.5273 | 0.9438 |
| 67 | 0.7101 | 0.6944 | 0.7195 | 0.6785 |
| 68 | 0.3710 | 0.9991 | 0.8873 | 0.4105 |
| 69 | 0.4802 | 0.9752 | 0.9113 | 0.3773 |
| 70 | 0.6681 | 0.7635 | 1.2819 | 0.1748 |
| 71 | 0.7312 | 0.6588 | 0.7920 | 0.5572 |
| 72 | 0.5556 | 0.9171 | 0.7143 | 0.6873 |

นอกจากนี้ เมื่อพิจารณาลักษณะการแจกแจงของค่าพารามิเตอร์ความสามารถผู้สอบ($\theta$) จากรูปหลายเหลี่ยมความถี่ (Histogram) ในโปรแกรม WINGEN ดังปรากฏในภาคผนวก ก พบว่า ค่าพารามิเตอร์ความสามารถผู้สอบ($\theta$) มีการแจกแจงแบบปกติ (Normal distribution)

**ตอนที่ 2 ผลการวิเคราะห์ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error)**

2.1 ค่าสถิติพื้นฐานของค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$

การวิเคราะห์ข้อมูลในส่วนนี้ เป็นการวิเคราะห์ด้วยค่าสถิติพื้นฐานเพื่อบรรยายลักษณะของค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ทั้ง 72 สถานการณ์ที่ทำการศึกษา ซึ่งประกอบด้วย ค่าน้อยที่สุด ค่ามากที่สุด ค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐาน ค่าความเบ้ ค่าความโด่ง ซึ่งพบว่า ดัชนี PARSCALE $G^2$ มีค่าความคลาดเคลื่อนประเภทที่ 1 ตั้งแต่ 0.0033 ถึง 0.9100 แตกต่างจากดัชนี Generalized $S-\chi^2$ ซึ่งมีค่าความคลาดเคลื่อนประเภทที่ 1 ตั้งแต่ 0.0000 ถึง 0.1533 จะเห็นได้ว่า ค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนี Generalized $S-\chi^2$ มีค่าสูงสุดเพียง 0.1533 เท่านั้น แตกต่างจากดัชนี PARSCALE $G^2$ ที่มีค่าสูงสุดถึง 0.9100

นอกจากนี้ค่าเฉลี่ยของค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนี Generalized $S-\chi^2$ มีค่า 0.0216 ซึ่งน้อยกว่าดัชนี PARSCALE $G^2$ ที่มีค่า 0.1535 รวมทั้งค่าส่วนเบี่ยงเบนมาตรฐานของค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนี Generalized $S-\chi^2$ มีค่า 0.0379 ซึ่งน้อยกว่าดัชนี PARSCALE $G^2$ ที่มีค่า 0.1974 แสดงให้เห็นว่าค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนี Generalized $S-\chi^2$ มีการกระจายน้อยกว่าดัชนี PARSCALE $G^2$ และค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนีสองชนิดมีการแจกแจงในลักษณะเบ้ขวา กล่าวคือ ค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนีทั้งสองชนิดส่วนใหญ่มีค่าน้อยกว่าค่าเฉลี่ย ซึ่งรายละเอียดที่กล่าวมาทั้งหมด สามารถพิจารณาได้จากตารางที่ 4.3

**ตารางที่ 4.3 ค่าสถิติพื้นฐานของค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$**

| ดัชนีความสอดคล้องของข้อคำถาม | ค่าสถิติพื้นฐาน | | | | | |
|---|---|---|---|---|---|---|
| | Min. | Max. | *M* | *SD* | *Sk* | *Ku* |
| PARSCALE $G^2$ | 0.0033 | 0.9100 | 0.1535 | 0.1974 | 1.887 | 3.328 |
| Generalized $S-\chi^2$ | 0.0000 | 0.1533 | 0.0216 | 0.0379 | 2.232 | 4.303 |

2.2 ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี PARSCALE $G^2$

การวิเคราะห์ข้อมูลในส่วนนี้ เป็นการนำเสนอข้อมูลค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนี PARSCALE $G^2$ จำแนกตามโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาคที่ใช้ในการศึกษา การนำเสนอข้อมูลนี้มีวัตถุประสงค์เพื่อการพิจารณาลักษณะของค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนี PARSCALE $G^2$ ทั้ง 72 สถานการณ์ที่ทำการศึกษา ว่ามีการเปลี่ยนแปลงไปตามความยาวแบบวัด ขนาดกลุ่มตัวอย่าง และจำนวนรายการคำตอบที่เปลี่ยนแปลงไปอย่างไร

เมื่อพิจารณาที่ค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนี PARSCALE $G^2$ ทั้ง 72 สถานการณ์ ที่โมเดล GRM พบว่า เมื่อความยาวแบบวัดเพิ่มขึ้น ก็มีแนวโน้มที่ค่าความคลาดเคลื่อนประเภทที่ 1 จะลดลงด้วยในเกือบทุกสถานการณ์ ยกเว้นในกรณี ณ ความยาวแบบวัดเปลี่ยนแปลงจาก 20 ข้อไป 40 ข้อ ที่ขนาดกลุ่มตัวอย่าง 500 คน ในทุกจำนวนรายการคำตอบ และที่ขนาดกลุ่มตัวอย่าง 1000 คน ในจำนวนรายการคำตอบ 7 และ 9 รายการ และเมื่อพิจารณาที่โมเดล GPCM พบว่า เมื่อความยาวแบบวัดเพิ่มขึ้น ก็มีแนวโน้มที่ค่าความคลาดเคลื่อนประเภทที่ 1 จะลดลงด้วยเช่นกันกับโมเดล GRM แต่โดยส่วนใหญ่ ค่าความคลาดเคลื่อนประเภทที่ 1 จะลดลง ณ ความยาวแบบวัดเปลี่ยนแปลงจาก 10 ข้อไป 20 ข้อ ที่ขนาดกลุ่มตัวอย่าง 1000 และ 2000 คน ในทุกจำนวนรายการคำตอบ และที่ขนาดกลุ่มตัวอย่าง 500 คน ในจำนวนรายการคำตอบ 3 และ 9 รายการ นอกจากนี้ เมื่อความยาวแบบวัดเปลี่ยนแปลงจาก 20 ข้อไป 40 ข้อ จะมีเพียง 3 กรณีคือ ณ ขนาดกลุ่มตัวอย่าง 500 คน ในจำนวนรายการคำตอบ 7 รายการ ณ ขนาดกลุ่มตัวอย่าง 1000 คน ในจำนวนรายการคำตอบ 7 รายการ และ ณ ขนาดกลุ่มตัวอย่าง 2000 คน ในจำนวนรายการคำตอบ 3 รายการที่ค่าความคลาดเคลื่อนประเภทที่ 1 จะลดลงเมื่อความยาวแบบวัดเพิ่มขึ้น

เมื่อพิจารณาที่ขนาดกลุ่มตัวอย่าง พบว่า โดยส่วนใหญ่โมเดล GRM เมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้น ค่าความคลาดเคลื่อนประเภทที่ 1 จะไม่ลดลง มีเพียงกรณี ณ ขนาดกลุ่มตัวอย่างเปลี่ยนแปลงจาก 500 คนไป 1000 คน ที่ความยาวแบบวัด 40 ข้อ ในจำนวนรายการคำตอบ 3 และ 5 รายการ ที่ความยาวแบบวัด 20 ข้อ ในจำนวนรายการคำตอบ 3 รายการและกรณี ณ ขนาดกลุ่มตัวอย่างเปลี่ยนแปลงจาก 1000 คนไป 2000 คน ที่ความยาวแบบวัด 40 ข้อ ในจำนวนรายการคำตอบ 7 และ 9 รายการ เมื่อพิจารณาที่โมเดล GPCM พบว่า โดยส่วนใหญ่โมเดล GPCM เมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้น ค่าความคลาดเคลื่อนประเภทที่ 1 จะไม่ลดลงเช่นเดียวกันกับโมเดล GRM มีเพียงกรณี ณ ขนาดกลุ่มตัวอย่างเปลี่ยนแปลงจาก 500 คนไป 1000 คน ที่ความยาวแบบวัด 20 ข้อ ในจำนวนรายการคำตอบ 5 และ 7 รายการที่ความยาวแบบวัด 40 ข้อ ในจำนวนรายการคำตอบ 9 รายการ และกรณี ณ ขนาดกลุ่มตัวอย่างเปลี่ยนแปลงจาก 1000 คนไป 2000 คน ที่ความยาวแบบวัด 10 ข้อ ในจำนวนรายการคำตอบ 7 รายการ ที่ความยาวแบบวัด 20 ข้อ ในจำนวนรายการคำตอบ 7 รายการ ที่ความยาวแบบวัด 40 ข้อ ในจำนวนรายการคำตอบ 3 รายการ ที่ค่าความคลาดเคลื่อนประเภทที่ 1 จะลดลงเมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้น

เมื่อพิจารณาที่จำนวนรายการคำตอบ พบว่า โดยส่วนใหญ่โมเดล GRM เมื่อจำนวนรายการคำตอบเพิ่มขึ้น ค่าความคลาดเคลื่อนประเภทที่ 1 จะลดลงโดยเฉพาะในกรณีจำนวนรายการคำตอบเปลี่ยนแปลงจาก 3 ไป 5 รายการ ในทุกความยาวแบบวัดและขนาดตัวอย่าง และในกรณีจำนวนรายการคำตอบเปลี่ยนแปลงจาก 5 ไป 7 รายการ ในเกือบทุกความยาวแบบวัดและขนาดตัวอย่างยกเว้นในกรณีความยาวแบบวัด 40 ข้อ ขนาดตัวอย่าง 1000 คน ส่วนในกรณีจำนวนรายการคำตอบเปลี่ยนแปลงจาก 7 ไป 9 รายการนั้น มีเพียง 2 กรณีคือ กรณีความยาวแบบวัด 10 ข้อ ขนาดตัวอย่าง 1000 คนและกรณีความยาวแบบวัด 10 ข้อ ขนาดตัวอย่าง 2000 คน พิจารณาที่โมเดล GPCM เมื่อจำนวนรายการคำตอบเปลี่ยนแปลงจาก 3 ไป 5 รายการ พบว่า เมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้น ค่าความคลาดเคลื่อนประเภทที่ 1 จะลดลงในกรณีความยาวแบบวัด 10 ข้อ ในทุกขนาดกลุ่มตัวอย่าง กรณีความยาวแบบวัด 20 ข้อ ขนาดกลุ่มตัวอย่าง 1000 คน และ 2000 คน เมื่อจำนวนรายการคำตอบเปลี่ยนแปลงจาก 5 ไป 7 รายการ พบว่า มีเพียงกรณีขนาดกลุ่มตัวอย่าง 2000 คน ความยาวแบบวัด 10 และ 40 ข้อ เท่านั้นที่ค่าความคลาดเคลื่อนประเภทที่ 1 ลดลง และเมื่อจำนวนรายการคำตอบเปลี่ยนแปลงจาก 7 ไป 9 รายการ พบว่า มีเพียงกรณีขนาดกลุ่มตัวอย่าง 500 คน ความยาวแบบวัด 20 ข้อ และกรณีขนาดกลุ่มตัวอย่าง 1000 คน ความยาวแบบวัด 10 และ 20 ข้อ เท่านั้นที่ค่าความคลาดเคลื่อนประเภทที่ 1 ลดลง เมื่อจำนวนรายการคำตอบเพิ่มขึ้น

ซึ่งรายละเอียดของค่าความคลาดเคลื่อนประเภทที่ 1 ทั้ง 72 สถานการณ์ ของดัชนี PARSCALE $G^2$ ดังแสดงในตารางที่ 4.4

**ตารางที่ 4.4 ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี** PARSCALE $G^2$

| โมเดล | ความยาวแบบวัด | ขนาดกลุ่มตัวอย่าง | จำนวนรายการคำตอบ | | | |
|---|---|---|---|---|---|---|
| | | | 3 | 5 | 7 | 9 |
| GRM | 10 | 500 | 0.5467 | 0.0767 | 0.0133 | 0.0200 |
| | | 1000 | 0.6900 | 0.1467 | 0.1167 | 0.0567 |
| | | 2000 | 0.9100 | 0.6067 | 0.3333 | 0.1133 |
| GRM | 20 | 500 | 0.1167 | 0.0183 | 0.0050 | 0.0133 |
| | | 1000 | 0.0500 | 0.0183 | 0.0067 | 0.0133 |
| | | 2000 | 0.1217 | 0.0500 | 0.0233 | 0.0267 |
| GRM | 40 | 500 | 0.1317 | 0.0692 | 0.0092 | 0.0158 |
| | | 1000 | 0.0475 | 0.0133 | 0.0167 | 0.0183 |
| | | 2000 | 0.0783 | 0.0200 | 0.0125 | 0.0142 |
| GPCM | 10 | 500 | 0.0367 | 0.0033 | 0.0167 | 0.1200 |
| | | 1000 | 0.4833 | 0.1033 | 0.2733 | 0.2367 |
| | | 2000 | 0.5367 | 0.4300 | 0.1067 | 0.3433 |
| GPCM | 20 | 500 | 0.0083 | 0.0267 | 0.2083 | 0.0567 |
| | | 1000 | 0.0217 | 0.0200 | 0.1700 | 0.1400 |
| | | 2000 | 0.3083 | 0.0533 | 0.0867 | 0.2833 |
| GPCM | 40 | 500 | 0.0158 | 0.0308 | 0.1058 | 0.7225 |
| | | 1000 | 0.0275 | 0.1283 | 0.1542 | 0.1942 |
| | | 2000 | 0.0175 | 0.3692 | 0.2842 | 0.3867 |

จากการพิจารณาค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนี PARSCALE $G^2$ จำแนกตามโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาคที่ใช้ในการศึกษาว่ามีการเปลี่ยนแปลงไปตามความยาวแบบวัด ขนาดกลุ่มตัวอย่าง และจำนวนรายการคำตอบที่เปลี่ยนแปลงไปอย่างไร ตามรายละเอียดดังที่กล่าวมาแล้วนั้น ผู้วิจัยจึงได้นำค่าความคลาดเคลื่อน

ประเภทที่ 1 ของดัชนี PARSCALE $G^2$ ไปวิเคราะห์เปรียบเทียบโดยมีวัตถุประสงค์เพื่อต้องการทดสอบว่าเมื่อเงื่อนไขที่ใช้ในการศึกษา (ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ) เปลี่ยนแปลงไป จะส่งผลต่อค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนี PARSCALE $G^2$ อย่างมีนัยสำคัญทางสถิติหรือไม่ ผู้วิจัยจึงได้ทำการทดสอบสมมติฐานทางสถิติที่ระดับความเชื่อมั่น 95 % เพื่อเปรียบเทียบค่าเฉลี่ยของค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี PARSCALE $G^2$ ในแต่ละเงื่อนไขที่ใช้ในการศึกษาจำแนกตามโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาคที่ใช้ในการศึกษา

ผลการทดสอบสมมติฐานทางสถิติ พบว่า เมื่อเป็นโมเดล GRM จะมีเพียงเงื่อนไขความยาวแบบวัด และจำนวนรายการคำตอบ ที่ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี PARSCALE $G^2$ มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ ($p < .05$) แสดงว่า มีค่าเฉลี่ยของค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี PARSCALE $G^2$ อย่างน้อย 2 ขนาดความยาวแบบวัด (1 คู่) และ 2 จำนวนรายการคำตอบ (1 คู่) ที่แตกต่างกันที่ระดับนัยสำคัญ .05 ส่วนในโมเดล GPCM นั้นไม่มีเงื่อนไขใดที่ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี PARSCALE $G^2$ มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ ซึ่งรายละเอียดค่าสถิติที่ใช้ในการทดสอบ ปรากฏดังในตารางที่ 4.5

**ตารางที่ 4.5 ผลการวิเคราะห์เปรียบเทียบค่าเฉลี่ยของค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี PARSCALE $G^2$ ในแต่ละเงื่อนไขที่ใช้ในการศึกษา**

| โมเดล | เงื่อนไขที่ใช้ในการศึกษา | ค่าสถิติ | | | | |
|---|---|---|---|---|---|---|
| | | SS | df | MS | *F* | *p* |
| GRM | ความยาวแบบวัด (10, 20, 40 ข้อ) | | | | | |
| | ระหว่างกลุ่ม | 0.560 | 2 | 0.280 | 8.593* | 0.001 |
| | ภายในกลุ่ม | 1.075 | 33 | 0.033 | | |
| | รวม | 1.635 | 35 | | | |
| GRM | ขนาดกลุ่มตัวอย่าง (500, 1000, 2000 คน) | | | | | |
| | ระหว่างกลุ่ม | 0.080 | 2 | 0.040 | 0.853 | 0.435 |
| | ภายในกลุ่ม | 1.555 | 33 | 0.047 | | |
| | รวม | 1.653 | 35 | | | |

**ตารางที่** 4.5 (ต่อ)

| โมเดล | เงื่อนไขที่ใช้ในการศึกษา | ค่าสถิติ | | | | |
|---|---|---|---|---|---|---|
| | | SS | df | MS | *F* | *p* |
| GRM | จำนวนรายการคำตอบ (3, 5, 7, 9 รายการ) | | | | | |
| | ระหว่างกลุ่ม | 0.390 | 3 | 0.130 | 3.338* | 0.031 |
| | ภายในกลุ่ม | 1.246 | 32 | 0.039 | | |
| | รวม | 1.635 | 35 | | | |
| GPCM | ความยาวแบบวัด (10, 20, 40 ข้อ) | | | | | |
| | ระหว่างกลุ่ม | 0.080 | 2 | 0.040 | 1.323 | 0.280 |
| | ภายในกลุ่ม | 0.998 | 33 | 0.030 | | |
| | รวม | 1.078 | 35 | | | |
| GPCM | ขนาดกลุ่มตัวอย่าง (500, 1000, 2000 คน) | | | | | |
| | ระหว่างกลุ่ม | 0.149 | 2 | 0.075 | 2.649 | 0.086 |
| | ภายในกลุ่ม | 0.929 | 33 | 0.028 | | |
| | รวม | 1.078 | 35 | | | |
| GPCM | จำนวนรายการคำตอบ (3, 5, 7, 9 รายการ) | | | | | |
| | ระหว่างกลุ่ม | 0.114 | 3 | 0.038 | 1.259 | 0.305 |
| | ภายในกลุ่ม | 0.965 | 32 | 0.030 | | |
| | รวม | 1.078 | 35 | | | |

* $p < .05$

จากผลการทดสอบที่พบว่า ในโมเดล GRM เงื่อนไขความยาวแบบวัดที่แตกต่างกัน ส่งผลให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี PARSCALE $G^2$ มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ ดังนั้นจึงต้องทำการทดสอบต่อไปว่า ณ ขนาดความยาวแบบวัดคู่ใดบ้างที่ให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี PARSCALE $G^2$ แตกต่างกัน โดยใช้การทดสอบเปรียบเทียบพหุคูณ ซึ่งผลการทดสอบพบว่า ที่โมเดล GRM ความยาวแบบวัด 10 ข้อจะให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี PARSCALE $G^2$ ที่มากกว่าความยาวแบบวัด 20 และ 40 ข้อ อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ซึ่งรายละเอียดค่าสถิติที่ใช้ในการทดสอบ ปรากฏดังในตารางที่ 4.6

**ตารางที่ 4.6 ผลการวิเคราะห์ความแตกต่างของค่าเฉลี่ยค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี PARSCALE $G^2$ โมเดล GRM ในเงื่อนไขความยาวแบบวัด (10, 20, 40 ข้อ)**

| ความยาวแบบวัด | ค่าเฉลี่ย | ความแตกต่างระหว่างความยาวแบบวัด | | |
|---|---|---|---|---|
| | | 10 ข้อ | 20 ข้อ | 40 ข้อ |
| 10 ข้อ | 0.303 | - | 0.264* | 0.266* |
| 20 ข้อ | 0.039 | | - | 0.002 |
| 40 ข้อ | 0.037 | | | - |
| $F$ = 8.593 | $p$ = 0.001 | | | |

* $p < .05$

จากผลการทดสอบที่พบว่า ในโมเดล GRM เงื่อนไขจำนวนรายการคำตอบที่แตกต่างกัน ส่งผลให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี PARSCALE $G^2$ มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ ดังนั้นจึงต้องทำการทดสอบต่อไปว่า ณ ขนาดจำนวนรายการคำตอบคู่ใดบ้างที่ให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี PARSCALE $G^2$ แตกต่างกัน โดยใช้การทดสอบเปรียบเทียบพหุคูณ ซึ่งผลการทดสอบพบว่า ที่โมเดล GRM จำนวนรายการคำตอบ 3 รายการ จะให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี PARSCALE $G^2$ ที่มากกว่าจำนวนรายการคำตอบ 7 และ 9 รายการ อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ซึ่งรายละเอียดค่าสถิติที่ใช้ในการทดสอบ ปรากฏดังในตารางที่ 4.7

**ตารางที่ 4.7 ผลการวิเคราะห์ความแตกต่างของค่าเฉลี่ยค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี PARSCALE $G^2$ โมเดล GRM ในเงื่อนไขจำนวนรายการคำตอบ (3, 5, 7, 9 รายการ)**

| จำนวนรายการคำตอบ | ค่าเฉลี่ย | ความแตกต่างระหว่างจำนวนรายการคำตอบ | | | |
|---|---|---|---|---|---|
| | | 3 รายการ | 5 รายการ | 7 รายการ | 9 รายการ |
| 3 รายการ | 0.299 | - | 0.186 | 0.240* | 0.267* |
| 5 รายการ | 0.113 | | - | 0.054 | 0.081 |
| 7 รายการ | 0.059 | | | - | 0.027 |
| 9 รายการ | 0.032 | | | | - |
| $F$ = 3.338 | $p$ = 0.031 | | | | |

* $p < .05$

2.3 ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี Generalized $S-\chi^2$

การวิเคราะห์ข้อมูลในส่วนนี้ เป็นการนำเสนอข้อมูลค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนี Generalized $S-\chi^2$ จำแนกตามโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาคที่ใช้ในการศึกษา การนำเสนอข้อมูลนี้มีวัตถุประสงค์เพื่อการพิจารณาลักษณะของค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนี Generalized $S-\chi^2$ ทั้ง 72 สถานการณ์ที่ทำการศึกษา ว่ามีการเปลี่ยนแปลงไปตามความยาวแบบวัด ขนาดกลุ่มตัวอย่าง และจำนวนรายการคำตอบที่เปลี่ยนแปลงไปอย่างไร

เมื่อพิจารณาที่ค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนี Generalized $S-\chi^2$ ทั้ง 72 สถานการณ์ ที่โมเดล GRM พบว่า เมื่อความยาวแบบวัดเพิ่มขึ้น ก็มีแนวโน้มที่ค่าความคลาดเคลื่อนประเภทที่ 1 จะลดลงด้วยในกรณี ณ ความยาวแบบวัดเปลี่ยนแปลงจาก 10 ข้อไป 20 ข้อ ที่ขนาดกลุ่มตัวอย่าง 500 คน ในทุกจำนวนรายการคำตอบ และที่ขนาดกลุ่มตัวอย่าง 1000 และ 2000 คน ในจำนวนรายการคำตอบ 3, 5 และ 7 รายการ นอกจากนี้ ค่าความคลาดเคลื่อนประเภทที่ 1 จะลดลงด้วยในกรณี ณ ความยาวแบบวัดเปลี่ยนแปลงจาก 20 ข้อไป 40 ข้อในกรณีที่ขนาดกลุ่มตัวอย่าง 500 คน ในจำนวนรายการคำตอบ 3 และ 5 รายการ ที่ขนาดกลุ่มตัวอย่าง 1000 คน ในจำนวนรายการคำตอบ 3 รายการ ที่ขนาดกลุ่มตัวอย่าง 2000 คน ในจำนวนรายการคำตอบ 3 และ 5 เมื่อพิจารณาที่โมเดล GPCM พบว่า เมื่อความยาวแบบวัดเพิ่มขึ้น ก็มีแนวโน้มที่ค่าความคลาดเคลื่อนประเภทที่ 1 จะลดลงด้วยเช่นกันกับโมเดล GRM แต่โดยส่วนใหญ่ ค่าความคลาดเคลื่อนประเภทที่ 1 จะลดลง ณ ความยาวแบบวัดเปลี่ยนแปลงจาก 10 ข้อไป 20 ข้อ ที่ขนาดกลุ่มตัวอย่าง 1000 และ 2000 คน ยกเว้นในขนาดกลุ่มตัวอย่าง 2000 คน จำนวนรายการคำตอบ 7 รายการ ที่ค่าความคลาดเคลื่อนประเภทที่ 1 จะคงเดิม ในขณะที่ ณ ความยาวแบบวัดเปลี่ยนแปลงจาก 20 ข้อไป 40 ข้อ จะมี ณ ขนาดกลุ่มตัวอย่าง 500 คน ในจำนวนรายการคำตอบ 3, 5 และ 7 รายการ ณ ขนาดกลุ่มตัวอย่าง 1000 คน ในจำนวนรายการคำตอบ 3 และ 7 รายการ ณ ขนาดกลุ่มตัวอย่าง 2000 คน ในจำนวนรายการคำตอบ 3 รายการ ที่ค่าความคลาดเคลื่อนประเภทที่ 1 จะลดลงเมื่อความยาวแบบวัดเพิ่มขึ้น

เมื่อพิจารณาที่ขนาดกลุ่มตัวอย่าง พบว่า โดยส่วนใหญ่โมเดล GRM เมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้น ค่าความคลาดเคลื่อนประเภทที่ 1 จะไม่ลดลง มีเพียงกรณี ณ ขนาดกลุ่มตัวอย่างเปลี่ยนแปลงจาก 500 คนไป 1000 คน ที่ความยาวแบบวัด 10 ข้อ ในจำนวนรายการคำตอบ 5, 7 และ 9 รายการ ที่ความยาวแบบวัด 20 ข้อ ในจำนวนรายการคำตอบ 5 รายการ และกรณี ณ ขนาดกลุ่มตัวอย่างเปลี่ยนแปลงจาก 1000 คนไป 2000 คน ที่ความยาวแบบวัด 10 ข้อ

จำนวนรายการคำตอบ 3 และ 7 รายการ ที่ความยาวแบบวัด 20 ข้อ จำนวนรายการคำตอบ 3 รายการ และที่ความยาวแบบวัด 40 ข้อ ในจำนวนรายการคำตอบ 3 และ 9 รายการ เมื่อพิจารณาที่โมเดล GPCM พบว่า โดยส่วนใหญ่โมเดล GPCM เมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้น ค่าความคลาดเคลื่อนประเภทที่ 1 จะไม่ลดลงเช่นเดียวกันกับโมเดล GRM มีเพียงกรณี ณ ขนาดกลุ่มตัวอย่างเปลี่ยนแปลงจาก 500 คนไป 1000 คน ที่ความยาวแบบวัด 20 ข้อ ในจำนวนรายการคำตอบ 5 รายการ ที่ความยาวแบบวัด 40 ข้อ ในจำนวนรายการคำตอบ 7 และ 9 รายการ กรณี ณ ขนาดกลุ่มตัวอย่างเปลี่ยนแปลงจาก 1000 คนไป 2000 คน ที่ความยาวแบบวัด 10 ข้อ ในจำนวนรายการคำตอบ 7 และ 9 รายการ ที่ความยาวแบบวัด 20 ข้อ ในจำนวนรายการคำตอบ 7 และ 9 รายการ ที่ความยาวแบบวัด 40 ข้อ ในจำนวนรายการคำตอบ 5, 7 และ 9 รายการ ที่ค่าความคลาดเคลื่อนประเภทที่ 1 จะลดลงเมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้น

เมื่อพิจารณาที่จำนวนรายการคำตอบ พบว่า โดยส่วนใหญ่โมเดล GRM เมื่อจำนวนรายการคำตอบเพิ่มขึ้น ค่าความคลาดเคลื่อนประเภทที่ 1 จะลดลง ณ กรณีจำนวนรายการคำตอบเปลี่ยนแปลงจาก 3 ไป 5 รายการ ในขนาดตัวอย่าง 1000 และ 2000 คน ณ กรณีจำนวนรายการคำตอบเปลี่ยนแปลงจาก 5 ไป 7 รายการ ในขนาดตัวอย่าง 500 และ 2000 คน และ ณ กรณีจำนวนรายการคำตอบเปลี่ยนแปลงจาก 7 ไป 9 รายการ ในขนาดตัวอย่าง 2000 คน พิจารณาที่โมเดล GPCM พบว่า โดยส่วนใหญ่ เมื่อจำนวนรายการคำตอบเพิ่มขึ้น ค่าความคลาดเคลื่อนประเภทที่ 1 จะไม่ลดลงมีเพียง เมื่อจำนวนรายการคำตอบเปลี่ยนแปลงจาก 3 ไป 5 รายการ พบว่า เมื่อจำนวนรายการคำตอบเพิ่มขึ้น ค่าความคลาดเคลื่อนประเภทที่ 1 จะลดลงในเกือบทุกขนาดกลุ่มตัวอย่างและทุกความยาวแบบวัด ยกเว้นที่ กรณีความยาวแบบวัด 20 ข้อ ขนาดกลุ่มตัวอย่าง 500 คน ที่เมื่อจำนวนรายการคำตอบเพิ่มขึ้น ค่าความคลาดเคลื่อนประเภทที่ 1 จะลดลง

รายละเอียดของค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนี Generalized $S-\chi^2$ ทั้ง 72 สถานการณ์ ดังแสดงในตารางที่ 4.8

**ตารางที่ 4.8 ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี** Generalized $S-\chi^2$

| โมเดล | ความยาวแบบวัด | ขนาดกลุ่มตัวอย่าง | จำนวนรายการคำตอบ | | | |
|---|---|---|---|---|---|---|
| | | | 3 | 5 | 7 | 9 |
| GRM | 10 | 500 | 0.0167 | 0.0367 | 0.0400 | 0.0033 |
| | | 1000 | 0.0633 | 0.0033 | 0.0167 | 0.0000 |
| | | 2000 | 0.0167 | 0.0067 | 0.0033 | 0.0000 |

**ตารางที่** 4.8 (ต่อ)

| โมเดล | ความยาวแบบวัด | ขนาดกลุ่มตัวอย่าง | จำนวนรายการคำตอบ | | | |
|---|---|---|---|---|---|---|
| | | | 3 | 5 | 7 | 9 |
| GRM | 20 | 500 | 0.0017 | 0.0033 | 0.0000 | 0.0000 |
| | | 1000 | 0.0167 | 0.0000 | 0.0000 | 0.0000 |
| | | 2000 | 0.0100 | 0.0033 | 0.0000 | 0.0000 |
| GRM | 40 | 500 | 0.0008 | 0.0008 | 0.0000 | 0.0017 |
| | | 1000 | 0.0058 | 0.0008 | 0.0008 | 0.0333 |
| | | 2000 | 0.0025 | 0.0025 | 0.0008 | 0.0000 |
| GPCM | 10 | 500 | 0.0200 | 0.0000 | 0.0033 | 0.1167 |
| | | 1000 | 0.0200 | 0.0033 | 0.1167 | 0.1533 |
| | | 2000 | 0.0700 | 0.0133 | 0.0000 | 0.1400 |
| GPCM | 20 | 500 | 0.0033 | 0.0050 | 0.0317 | 0.0517 |
| | | 1000 | 0.0050 | 0.0000 | 0.1033 | 0.0533 |
| | | 2000 | 0.0100 | 0.0000 | 0.0000 | 0.0000 |
| GPCM | 40 | 500 | 0.0025 | 0.0017 | 0.0267 | 0.1492 |
| | | 1000 | 0.0033 | 0.0017 | 0.0192 | 0.0717 |
| | | 2000 | 0.0050 | 0.0000 | 0.0025 | 0.0625 |

จากการพิจารณาค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนี Generalized $S-\chi^2$ จำแนกตามโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาคที่ใช้ในการศึกษาว่ามีการเปลี่ยนแปลงไปตามความยาวแบบวัด ขนาดกลุ่มตัวอย่าง และจำนวนรายการคำตอบที่เปลี่ยนแปลงไปอย่างไร ตามรายละเอียดดังที่กล่าวมาแล้วนั้น ผู้วิจัยจึงได้นำค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนี Generalized $S-\chi^2$ ไปวิเคราะห์เปรียบเทียบโดยมีวัตถุประสงค์เพื่อต้องการทดสอบว่าเมื่อเงื่อนไขที่ใช้ในการศึกษา (ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ) เปลี่ยนแปลงไป จะส่งผลต่อค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนี Generalized $S-\chi^2$ อย่างมีนัยสำคัญทางสถิติหรือไม่ ผู้วิจัยจึงได้ทำการทดสอบสมมติฐานทางสถิติที่ระดับความเชื่อมั่น 95 % เพื่อเปรียบเทียบค่าเฉลี่ยของค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี Generalized $S-\chi^2$ ในแต่ละเงื่อนไขที่ใช้ในการศึกษาจำแนกตามโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาคที่ใช้ในการศึกษา

ผลการทดสอบสมมติฐานทางสถิติ พบว่า เมื่อเป็นโมเดล GRM จะมีเพียงเงื่อนไขความยาวแบบวัด ที่ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี Generalized $S-\chi^2$ มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ ($p < .05$) แสดงว่า มีค่าเฉลี่ยของค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี Generalized $S-\chi^2$ อย่างน้อย 2 ขนาดความยาวแบบวัด (1 คู่) ที่แตกต่างกันที่ระดับนัยสำคัญ .05 ส่วนในโมเดล GPCM นั้นจะมีเพียงเงื่อนไขจำนวนรายการคำตอบ ที่ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี Generalized $S-\chi^2$ มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ ($p < .05$) แสดงว่า มีค่าเฉลี่ยของค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี Generalized $S-\chi^2$ อย่างน้อย 2 รายการคำตอบ (1 คู่) ที่แตกต่างกันที่ระดับนัยสำคัญ .05 ซึ่งรายละเอียดค่าสถิติที่ใช้ในการทดสอบ ปรากฏดังในตารางที่ 4.9

**ตารางที่ 4.9 ผลการวิเคราะห์เปรียบเทียบค่าเฉลี่ยของค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี Generalized $S-\chi^2$ ในแต่ละเงื่อนไขที่ใช้ในการศึกษา**

| โมเดล | เงื่อนไขที่ใช้ในการศึกษา | ค่าสถิติ | | | | |
|---|---|---|---|---|---|---|
| | | SS | df | MS | *F* | *p* |
| GRM | ความยาวแบบวัด (10, 20, 40 ข้อ) | | | | | |
| | ระหว่างกลุ่ม | 0.002 | 2 | 0.001 | 4.468* | 0.019 |
| | ภายในกลุ่ม | 0.006 | 33 | 0.000 | | |
| | รวม | 0.007 | 35 | | | |
| GRM | ขนาดกลุ่มตัวอย่าง (500, 1000, 2000 คน) | | | | | |
| | ระหว่างกลุ่ม | 0.000 | 2 | 0.000 | 0.947 | 0.398 |
| | ภายในกลุ่ม | 0.007 | 33 | 0.000 | | |
| | รวม | 0.007 | 35 | | | |
| GRM | จำนวนรายการคำตอบ (3, 5, 7, 9 รายการ) | | | | | |
| | ระหว่างกลุ่ม | 0.001 | 3 | 0.000 | 0.972 | 0.418 |
| | ภายในกลุ่ม | 0.006 | 32 | 0.000 | | |
| | รวม | 0.007 | 35 | | | |

**ตารางที่** 4.9 (ต่อ)

| โมเดล | เงื่อนไขที่ใช้ในการศึกษา | ค่าสถิติ | | | | |
|---|---|---|---|---|---|---|
| | | SS | df | MS | *F* | *p* |
| | ระหว่างกลุ่ม | 0.007 | 2 | 0.004 | 1.592 | 0.219 |
| | ภายในกลุ่ม | 0.074 | 33 | 0.002 | | |
| | รวม | 0.082 | 35 | | | |
| GPCM | ขนาดกลุ่มตัวอย่าง (500, 1000, 2000 คน) | | | | | |
| | ระหว่างกลุ่ม | 0.003 | 2 | 0.001 | 0.536 | 0.590 |
| | ภายในกลุ่ม | 0.079 | 33 | 0.002 | | |
| | รวม | 0.082 | 35 | | | |
| GPCM | จำนวนรายการคำตอบ (3, 5, 7, 9 รายการ) | | | | | |
| | ระหว่างกลุ่ม | 0.039 | 3 | 0.013 | 9.668* | 0.000 |
| | ภายในกลุ่ม | 0.043 | 32 | 0.001 | | |
| | รวม | 0.082 | 35 | | | |

* $p < .05$

จากผลการทดสอบที่พบว่า ในโมเดล GRM เงื่อนไขความยาวแบบวัดที่แตกต่างกัน ส่งผลให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี Generalized $S-\chi^2$ มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ ดังนั้นจึงต้องทำการทดสอบต่อไปว่า ณ ขนาดความยาวแบบวัดคู่ใดบ้างที่ให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี Generalized $S-\chi^2$ แตกต่างกัน โดยใช้การทดสอบเปรียบเทียบพหุคูณ ซึ่งผลการทดสอบพบว่า ที่โมเดล GRM ความยาวแบบวัด 10 ข้อจะให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี Generalized $S-\chi^2$ ที่มากกว่าความยาวแบบวัด 20 และ 40 ข้อ อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ซึ่งรายละเอียดค่าสถิติที่ใช้ในการทดสอบ ปรากฏดังในตารางที่ 4.10

**ตารางที่ 4.10 ผลการวิเคราะห์ความแตกต่างของค่าเฉลี่ยค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี Generalized $S-\chi^2$ โมเดล GRM ในเงื่อนไขความยาวแบบวัด (10, 20, 40 ข้อ)**

| ความยาวแบบวัด | ค่าเฉลี่ย | ความแตกต่างระหว่างความยาวแบบวัด | | |
|---|---|---|---|---|
| | | 10 ข้อ | 20 ข้อ | 40 ข้อ |
| 10 ข้อ | 0.017 | - | 0.014* | 0.013* |
| 20 ข้อ | 0.003 | | - | -0.001 |
| 40 ข้อ | 0.004 | | | - |
| $F$ = 4.468 | $p$ = 0.019 | | | |

* $p < .05$

จากผลการทดสอบที่พบว่า ในโมเดล GPCM เงื่อนไขจำนวนรายการคำตอบที่แตกต่างกัน ส่งผลให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี Generalized $S-\chi^2$ มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ ดังนั้นจึงต้องทำการทดสอบต่อไปว่า ณ จำนวนรายการคำตอบคู่ใดบ้างที่ให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี Generalized $S-\chi^2$ แตกต่างกัน โดยใช้การทดสอบเปรียบเทียบพหุคูณ ซึ่งผลการทดสอบพบว่าที่โมเดล GPCM จำนวนรายการคำตอบ 9 รายการจะให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี Generalized $S-\chi^2$ ที่มากกว่าจำนวนรายการคำตอบ 3 และ 5 รายการอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่แตกต่างจากจำนวนรายการคำตอบ 7 รายการอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ซึ่งรายละเอียดค่าสถิติที่ใช้ในการทดสอบ ปรากฏดังในตารางที่ 4.11

**ตารางที่ 4.11 ผลการวิเคราะห์ความแตกต่างของค่าเฉลี่ยค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี Generalized $S-\chi^2$ โมเดล GPCM ในเงื่อนไขจำนวนรายการคำตอบ (3, 5, 7, 9 รายการ)**

| จำนวนรายการคำตอบ | ค่าเฉลี่ย | ความแตกต่างระหว่างจำนวนรายการคำตอบ | | | |
|---|---|---|---|---|---|
| | | 3 รายการ | 5 รายการ | 7 รายการ | 9 รายการ |
| 3 รายการ | 0.015 | - | 0.012 | -0.019 | -0.074* |
| 5 รายการ | 0.003 | | - | -0.031 | -0.086* |
| 7 รายการ | 0.034 | | | - | -0.055 |
| 9 รายการ | 0.089 | | | | - |
| $F$ = 9.668 | $p$ = 0.000 | | | | |

* $p < .05$

ผลการวิเคราะห์ค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนี PARSCALE $G^2$ สรุปสาระสำคัญได้ว่า เมื่อนำค่าความคลาดเคลื่อนประเภทที่ 1 ไปทดสอบเพื่อสังเกตความเปลี่ยนแปลงของค่าความคลาดเคลื่อนประเภทที่ 1 เมื่อเงื่อนไขที่ใช้ในการศึกษา (ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ) มีการเปลี่ยนแปลงไป จะพบว่าในโมเดล GRM มีเพียงเงื่อนไขความยาวแบบวัดและจำนวนรายการคำตอบเท่านั้นที่มีการเปลี่ยนแปลงของค่าความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติ โดยสอดคล้องกับผลการพิจารณาของผู้วิจัยว่า ที่โมเดล GRM เมื่อความยาวแบบวัดเพิ่มขึ้น ก็มีแนวโน้มที่ค่าความคลาดเคลื่อนประเภทที่ 1 จะลดลงด้วย และที่โมเดล GRM เมื่อจำนวนรายการคำตอบเพิ่มขึ้น ค่าความคลาดเคลื่อนประเภทที่ 1 จะลดลง ดังในผลการทดสอบเปรียบเทียบพหุคูณ (Multiple Comparison) ที่ให้สารสนเทศว่า ในโมเดล GRM ความยาวแบบวัด 10 ข้อ จะให้ค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนี PARSCALE $G^2$ มากกว่าความยาวแบบวัด 20 และ 40 ข้อ และในโมเดล GRM จำนวนรายการคำตอบ 3 รายการ จะให้ค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนี PARSCALE $G^2$ มากกว่าจำนวนรายการคำตอบ 7 และ 9 รายการ

ผลการวิเคราะห์ค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนี Generalized $S-\chi^2$ สรุปสาระสำคัญได้ว่า เมื่อนำค่าความคลาดเคลื่อนประเภทที่ 1 ไปทดสอบเพื่อสังเกตความเปลี่ยนแปลงของค่าความคลาดเคลื่อนประเภทที่ 1 เมื่อเงื่อนไขที่ใช้ในการศึกษา (ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ) มีการเปลี่ยนแปลงไป จะพบว่าในโมเดล GRM มีเพียงเงื่อนไขความยาวแบบวัดเท่านั้นที่มีการเปลี่ยนแปลงของค่าความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติ โดยสอดคล้องกับผลการพิจารณาของผู้วิจัยว่า ที่โมเดล GRM เมื่อความยาวแบบวัดเพิ่มขึ้น ก็มีแนวโน้มที่ค่าความคลาดเคลื่อนประเภทที่ 1 จะลดลงด้วย และที่โมเดล GPCM มีเพียงเงื่อนไขจำนวนรายการคำตอบเท่านั้นที่มีการเปลี่ยนแปลงของค่าความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติ โดยสอดคล้องกับผลการพิจารณาของผู้วิจัยว่า ในโมเดล GPCM เมื่อจำนวนรายการคำตอบเพิ่มขึ้น ค่าความคลาดเคลื่อนประเภทที่ 1 จะไม่ลดลง ดังในผลการทดสอบเปรียบเทียบพหุคูณ (Multiple Comparison) ที่ให้สารสนเทศว่า ในโมเดล GRM ความยาวแบบวัด 10 ข้อ จะให้ค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนี Generalized $S-\chi^2$ มากกว่าความยาวแบบวัด 20 และ 40 ข้อ และในโมเดล GPCM จำนวนรายการคำตอบ 9 รายการ จะให้ค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนี PARSCALE $G^2$ มากกว่าจำนวนรายการคำตอบ 3 และ 5 รายการ แต่ไม่แตกต่างจากจำนวนรายการคำตอบ 7 รายการ

จะเห็นได้ว่า ผลการทดสอบเปรียบเทียบพหุคูณ (Multiple Comparison) ของดัชนีทั้งสองชนิดให้สารสนเทศที่สอดคล้องกันในโมเดล GRM ว่าความยาวแบบวัด 10 ข้อ จะให้ค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนีทั้งสองชนิดมากกว่าความยาวแบบวัด 20 และ 40 ข้อ แสดงให้เห็นว่า ความยาวแบบวัดมีผลต่อค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนีทั้งสองชนิดเหมือนกัน

**ตอนที่ 3 ผลการวิเคราะห์อำนาจการทดสอบ (Power of the test)**

3.1 ค่าสถิติพื้นฐานของอำนาจการทดสอบ (Power of the test) ของดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$

การวิเคราะห์ข้อมูลในส่วนนี้ เป็นการวิเคราะห์ด้วยค่าสถิติพื้นฐานเพื่อบรรยายลักษณะของอำนาจการทดสอบ (Power of the test) ของดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ทั้ง 72 สถานการณ์ที่ทำการศึกษา ซึ่งประกอบด้วย ค่าน้อยที่สุด ค่ามากที่สุด ค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐาน ค่าความเบ้ ค่าความโด่ง ซึ่งพบว่า ดัชนี PARSCALE $G^2$ มีอำนาจการทดสอบตั้งแต่ 0.0233 ถึง 0.9983 แตกต่างจากดัชนี Generalized $S-\chi^2$ ซึ่งมีอำนาจการทดสอบตั้งแต่ 0.0008 ถึง 0.7567 จะเห็นได้ว่า ค่ามากที่สุดและค่าน้อยที่สุดของอำนาจการทดสอบของดัชนี PARSCALE $G^2$ มีค่ามากกว่าดัชนี Generalized $S-\chi^2$

นอกจากนี้ค่าเฉลี่ยของอำนาจการทดสอบของดัชนี Generalized $S-\chi^2$ มีค่า 0.1064 ซึ่งน้อยกว่าดัชนี PARSCALE $G^2$ ที่มีค่า 0.4614 รวมทั้งค่าส่วนเบี่ยงเบนมาตรฐานของอำนาจการทดสอบของดัชนี Generalized $S-\chi^2$ มีค่า 0.1256 ซึ่งน้อยกว่าดัชนี PARSCALE $G^2$ ที่มีค่า 0.2935 แสดงให้เห็นว่าอำนาจการทดสอบของดัชนี Generalized $S-\chi^2$ มีการกระจายน้อยกว่าดัชนี PARSCALE $G^2$ และอำนาจการทดสอบของดัชนีสองชนิดมีการแจกแจงในลักษณะเบ้ขวา กล่าวคือ อำนาจการทดสอบของดัชนีทั้งสองชนิดส่วนใหญ่มีค่าน้อยกว่าค่าเฉลี่ย ซึ่งรายละเอียดที่กล่าวมาทั้งหมด สามารถพิจารณาได้จากตารางที่ 4.12

**ตารางที่ 4.12 ค่าสถิติพื้นฐานของอำนาจการทดสอบ (Power of the test) ของดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$**

| ดัชนีความสอดคล้องของข้อคำถาม | ค่าสถิติพื้นฐาน | | | | | |
|---|---|---|---|---|---|---|
| | Min. | Max. | *M* | *SD* | *Sk* | *Ku* |
| PARSCALE $G^2$ | 0.0233 | 0.9983 | 0.4614 | 0.2935 | 0.2650 | -1.138 |
| Generalized $S-\chi^2$ | 0.0008 | 0.7567 | 0.1064 | 0.1256 | 2.6110 | 9.584 |

3.2 อำนาจการทดสอบ (Power of the test) ของดัชนี PARSCALE $G^2$

การวิเคราะห์ข้อมูลในส่วนนี้ เป็นการนำเสนอข้อมูลอำนาจการทดสอบของดัชนี PARSCALE $G^2$ จำแนกตามโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาคที่ใช้ในการศึกษา การนำเสนอข้อมูลนี้มีวัตถุประสงค์เพื่อการพิจารณาลักษณะของอำนาจการทดสอบของดัชนี PARSCALE $G^2$ ทั้ง 72 สถานการณ์ที่ทำการศึกษา ว่ามีการเปลี่ยนแปลงไปตามความยาวแบบวัด ขนาดกลุ่มตัวอย่าง และจำนวนรายการคำตอบที่เปลี่ยนแปลงไปอย่างไร

เมื่อพิจารณาที่อำนาจการทดสอบของดัชนี PARSCALE $G^2$ ทั้ง 72 สถานการณ์ ที่โมเดล GRM พบว่า เมื่อความยาวแบบวัดเพิ่มขึ้น โดยส่วนใหญ่ อำนาจการทดสอบจะมีค่าลดลง มีเพียงบางกรณีเท่านั้นที่อำนาจการทดสอบเพิ่มขึ้น เช่น ในกรณี ณ ความยาวแบบวัดเปลี่ยนแปลงจาก 10 ข้อไป 20 ข้อ และ ณ ความยาวแบบวัดเปลี่ยนแปลงจาก 20 ข้อไป 40 ข้อ ที่ขนาดกลุ่มตัวอย่าง 1000 คน ในจำนวนรายการคำตอบ 9 รายการ และเมื่อพิจารณาที่โมเดล GPCM พบว่า โดยส่วนใหญ่ เมื่อความยาวแบบวัดเพิ่มขึ้น ก็มีแนวโน้มที่อำนาจการทดสอบจะเพิ่มขึ้นด้วย โดยเฉพาะเมื่อจำนวนรายการคำตอบเป็น 7 และ 9 รายการ แต่โดยส่วนใหญ่ อำนาจการทดสอบจะเพิ่มขึ้นเมื่อความยาวแบบวัดเปลี่ยนแปลงจาก 20 ข้อไป 40 ข้อ

เมื่อพิจารณาที่ขนาดกลุ่มตัวอย่าง พบว่า โมเดล GRM เมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้น อำนาจการทดสอบจะมีค่าเพิ่มมากขึ้นในทุกกรณีที่ทำการศึกษา และเมื่อพิจารณาที่โมเดล GPCM พบว่า โดยส่วนใหญ่โมเดล GPCM เมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้น อำนาจการทดสอบก็จะเพิ่มขึ้นเช่นเดียวกันกับโมเดล GRM แต่จะมีเพียงบางกรณีเท่านั้นที่อำนาจการทดสอบลดลง ณ ขนาดกลุ่มตัวอย่างเปลี่ยนแปลงจาก 500 คนไป 1000 คน ที่ความยาวแบบวัด 20 ข้อ ในจำนวนรายการคำตอบ 5 รายการ ที่ความยาวแบบวัด 40 ข้อ ในจำนวนรายการคำตอบ 3 รายการ และกรณี ณ ขนาดกลุ่มตัวอย่างเปลี่ยนแปลงจาก 1000 คนไป 2000 คน ที่ความยาวแบบวัด 40 ข้อ ในจำนวนรายการคำตอบ 3 รายการ

เมื่อพิจารณาที่จำนวนรายการคำตอบ พบว่า โดยส่วนใหญ่โมเดล GRM เมื่อจำนวนรายการคำตอบเพิ่มขึ้น อำนาจการทดสอบจะเพิ่มขึ้น มีเพียงบางกรณีเท่านั้นที่อำนาจการทดสอบไม่เพิ่มขึ้น โดยเฉพาะในกรณีความยาวแบบวัด 10 ข้อ เช่น ณ กรณีจำนวนรายการคำตอบเปลี่ยนแปลงจาก 3 ไป 5 รายการ ในขนาดตัวอย่าง 500, 1000 และ 2000 คน ณ กรณีจำนวนรายการคำตอบเปลี่ยนแปลงจาก 5 ไป 7 รายการ ที่ความยาวแบบวัด 20 ข้อ ขนาดตัวอย่าง 500 คน และที่ความยาวแบบวัด 40 ข้อ ขนาดตัวอย่าง 2000 คน และ ณ กรณีจำนวนรายการคำตอบเปลี่ยนแปลงจาก 7 ไป 9 รายการ ที่ความยาวแบบวัด 10 ข้อ ในทุกขนาดตัวอย่าง พิจารณาที่

โมเดล GPCM พบว่า โดยส่วนใหญ่ เมื่อจำนวนรายการคำตอบเพิ่มขึ้น อำนาจการทดสอบจะเพิ่มขึ้น มีเพียงบางกรณีเท่านั้นที่อำนาจการทดสอบไม่เพิ่มขึ้น เช่น เมื่อจำนวนรายการคำตอบเปลี่ยนแปลงจาก 3 ไป 5 รายการ พบว่า เมื่อจำนวนรายการคำตอบเพิ่มขึ้น อำนาจการทดสอบจะลดลงในกรณีขนาดกลุ่มตัวอย่าง 500 คน ความยาวแบบวัด 10 และ 40 ข้อ กรณีขนาดกลุ่มตัวอย่าง 1000 คน ความยาวแบบวัด 10 และ 20 ข้อและกรณีขนาดกลุ่มตัวอย่าง 2000 คน ความยาวแบบวัด 10 ข้อ นอกจากนี้ เมื่อจำนวนรายการคำตอบเปลี่ยนแปลงจาก 5 ไป 7 รายการ พบว่า เมื่อจำนวนรายการคำตอบเพิ่มขึ้น อำนาจการทดสอบจะลดลงในกรณีขนาดกลุ่มตัวอย่าง 500 คน ความยาวแบบวัด 20 ข้อ

ซึ่งรายละเอียดของอำนาจการทดสอบ ทั้ง 72 สถานการณ์ ของดัชนี PARSCALE $G^2$ ดังแสดงในตารางที่ 4.13

**ตารางที่ 4.13 อำนาจการทดสอบ (Power of the test) ของดัชนี PARSCALE $G^2$**

| โมเดล | ความยาวแบบวัด | ขนาดกลุ่มตัวอย่าง | จำนวนรายการคำตอบ | | | |
|---|---|---|---|---|---|---|
| | | | 3 | 5 | 7 | 9 |
| GRM | 10 | 500 | 0.6033 | 0.2400 | 0.2967 | 0.2069 |
| | | 1000 | 0.7700 | 0.4467 | 0.6967 | 0.3933 |
| | | 2000 | 0.9567 | 0.8933 | 0.9067 | 0.8833 |
| GRM | 20 | 500 | 0.0233 | 0.1700 | 0.1217 | 0.3183 |
| | | 1000 | 0.0517 | 0.3233 | 0.4417 | 0.4650 |
| | | 2000 | 0.3167 | 0.6217 | 0.7367 | 0.8200 |
| GRM | 40 | 500 | 0.0300 | 0.1525 | 0.2100 | 0.2467 |
| | | 1000 | 0.0450 | 0.3842 | 0.4258 | 0.5333 |
| | | 2000 | 0.1242 | 0.6200 | 0.5417 | 0.6383 |
| GPCM | 10 | 500 | 0.0567 | 0.0400 | 0.0767 | 0.1667 |
| | | 1000 | 0.4800 | 0.2200 | 0.2367 | 0.2433 |
| | | 2000 | 0.6500 | 0.5700 | 0.5833 | 0.6633 |
| GPCM | 20 | 500 | 0.0550 | 0.2817 | 0.1517 | 0.1700 |
| | | 1000 | 0.2600 | 0.1850 | 0.5300 | 0.7867 |
| | | 2000 | 0.3850 | 0.5133 | 0.8850 | 0.9383 |

**ตารางที่** 4.13 (ต่อ)

| โมเดล | ความยาวแบบวัด | ขนาดกลุ่มตัวอย่าง | จำนวนรายการคำตอบ | | | |
|---|---|---|---|---|---|---|
| | | | 3 | 5 | 7 | 9 |
| GPCM | 40 | 500 | 0.5242 | 0.3158 | 0.6950 | 0.9650 |
| | | 1000 | 0.4325 | 0.6817 | 0.8208 | 0.9867 |
| | | 2000 | 0.2133 | 0.8758 | 0.9308 | 0.9983 |

จากการพิจารณาอำนาจการทดสอบของดัชนี PARSCALE $G^2$ จำแนกตามโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาคที่ใช้ในการศึกษาว่ามีการเปลี่ยนแปลงไปตามความยาวแบบวัด ขนาดกลุ่มตัวอย่าง และจำนวนรายการคำตอบที่เปลี่ยนแปลงไปอย่างไร ตามรายละเอียดดังที่กล่าวมาแล้วนั้น ผู้วิจัยจึงได้นำอำนาจการทดสอบของดัชนี PARSCALE $G^2$ ไปวิเคราะห์เปรียบเทียบโดยมีวัตถุประสงค์เพื่อต้องการทดสอบว่าเมื่อเงื่อนไขที่ใช้ในการศึกษา (ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ) เปลี่ยนแปลงไป จะส่งผลต่ออำนาจการทดสอบของดัชนี PARSCALE $G^2$ อย่างมีนัยสำคัญทางสถิติหรือไม่ ผู้วิจัยจึงได้ทำการทดสอบสมมติฐานทางสถิติที่ระดับความเชื่อมั่น 95 % เพื่อเปรียบเทียบค่าเฉลี่ยของอำนาจการทดสอบของดัชนี PARSCALE $G^2$ ในแต่ละเงื่อนไขที่ใช้ในการศึกษาจำแนกตามโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาคที่ใช้ในการศึกษา

ผลการทดสอบสมมติฐานทางสถิติ พบว่า เมื่อเป็นโมเดล GRM จะมีเพียงเงื่อนไขความยาวแบบวัดและขนาดกลุ่มตัวอย่าง ที่อำนาจการทดสอบของดัชนี PARSCALE $G^2$ มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ ($p < .05$) แสดงว่า มีอำนาจการทดสอบของดัชนี PARSCALE $G^2$ อย่างน้อย 2 ขนาดความยาวแบบวัด (1 คู่) และอย่างน้อย 2 ขนาดกลุ่มตัวอย่าง (1 คู่) ที่แตกต่างกันที่ระดับนัยสำคัญ .05 เช่นเดียวกันกับโมเดล GPCM ที่มีเพียงเงื่อนไขความยาวแบบวัดและขนาดกลุ่มตัวอย่าง ที่อำนาจการทดสอบของดัชนี PARSCALE $G^2$ มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ ($p < .05$) แสดงว่า มีอำนาจการทดสอบของดัชนี PARSCALE $G^2$ อย่างน้อย 2 ขนาดความยาวแบบวัด (1 คู่) และอย่างน้อย 2 ขนาดกลุ่มตัวอย่าง (1 คู่) ที่แตกต่างกันที่ระดับนัยสำคัญ .05 ซึ่งรายละเอียดค่าสถิติที่ใช้ในการทดสอบ ปรากฏดังในตารางที่ 4.14

**ตารางที่ 4.14 ผลการวิเคราะห์เปรียบเทียบค่าเฉลี่ยของอำนาจการทดสอบ (Power of the test) ของดัชนี PARSCALE $G^2$ ในแต่ละเงื่อนไขที่ใช้ในการศึกษา**

| โมเดล | เงื่อนไขที่ใช้ในการศึกษา | ค่าสถิติ | | | | |
|---|---|---|---|---|---|---|
| | | SS | df | MS | *F* | *p* |
| GRM | ความยาวแบบวัด (10, 20, 40 ข้อ) | | | | | |
| | ระหว่างกลุ่ม | 0.547 | 2 | 0.274 | 4.185* | 0.024 |
| | ภายในกลุ่ม | 2.157 | 33 | 0.065 | | |
| | รวม | 2.704 | 35 | | | |
| GRM | ขนาดกลุ่มตัวอย่าง (500, 1000, 2000 คน) | | | | | |
| | ระหว่างกลุ่ม | 1.240 | 2 | 0.620 | 13.985* | 0.000 |
| | ภายในกลุ่ม | 1.463 | 33 | 0.044 | | |
| | รวม | 2.704 | 35 | | | |
| GRM | จำนวนรายการคำตอบ (3, 5, 7, 9 รายการ) | | | | | |
| | ระหว่างกลุ่ม | 0.173 | 3 | 0.058 | 0.728 | 0.543 |
| | ภายในกลุ่ม | 2.531 | 32 | 0.079 | | |
| | รวม | 2.704 | 35 | | | |
| GPCM | ความยาวแบบวัด (10, 20, 40 ข้อ) | | | | | |
| | ระหว่างกลุ่ม | 0.890 | 2 | 0.445 | 5.943* | 0.006 |
| | ภายในกลุ่ม | 2.471 | 33 | 0.075 | | |
| | รวม | 3.361 | 35 | | | |
| GPCM | ขนาดกลุ่มตัวอย่าง (500, 1000, 2000 คน) | | | | | |
| | ระหว่างกลุ่ม | 0.924 | 2 | 0.462 | 6.252* | 0.005 |
| | ภายในกลุ่ม | 2.438 | 33 | 0.074 | | |
| | รวม | 3.361 | 35 | | | |
| GPCM | จำนวนรายการคำตอบ (3, 5, 7, 9 รายการ) | | | | | |
| | ระหว่างกลุ่ม | 0.543 | 3 | 0.181 | 2.053 | 0.126 |
| | ภายในกลุ่ม | 2.819 | 32 | 0.088 | | |
| | รวม | 3.361 | 35 | | | |

* $p < .05$

จากผลการทดสอบที่พบว่า ในโมเดล GRM เงื่อนไขความยาวแบบวัดที่แตกต่างกัน ส่งผลให้อำนาจการทดสอบของดัชนี PARSCALE $G^2$ มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ ดังนั้นจึงต้องทำการทดสอบต่อไปว่า ณ ขนาดความยาวแบบวัดคู่ใดบ้างที่ให้อำนาจการทดสอบของดัชนี PARSCALE $G^2$ แตกต่างกัน โดยใช้การทดสอบเปรียบเทียบพหุคูณ ซึ่งผลการทดสอบพบว่า ที่โมเดล GRM ความยาวแบบวัด 10 ข้อจะให้อำนาจการทดสอบของดัชนี PARSCALE $G^2$ ที่มากกว่าความยาวแบบวัด 40 ข้อ อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ซึ่งรายละเอียดค่าสถิติที่ใช้ในการทดสอบ ปรากฏดังในตารางที่ 4.15

**ตารางที่ 4.15 ผลการวิเคราะห์ความแตกต่างของค่าเฉลี่ยอำนาจการทดสอบ (Power of the test) ของดัชนี PARSCALE $G^2$ โมเดล GRM ในเงื่อนไขความยาวแบบวัด (10, 20, 40 ข้อ)**

| ความยาวแบบวัด | ค่าเฉลี่ย | ความแตกต่างระหว่างความยาวแบบวัด | | |
|---|---|---|---|---|
| | | 10 ข้อ | 20 ข้อ | 40 ข้อ |
| 10 ข้อ | 0.608 | - | 0.240 | 0.279* |
| 20 ข้อ | 0.368 | | - | 0.039 |
| 40 ข้อ | 0.329 | | | - |
| $F = 4.185$ | $p = 0.024$ | | | |

* $p < .05$

จากผลการทดสอบที่พบว่าในโมเดล GRM เงื่อนไขขนาดกลุ่มตัวอย่างที่แตกต่างกัน ส่งผลให้อำนาจการทดสอบของดัชนี PARSCALE $G^2$ มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ ดังนั้นจึงต้องทำการทดสอบต่อไปว่า ณ ขนาดกลุ่มตัวอย่างคู่ใดบ้างที่ให้อำนาจการทดสอบของดัชนี PARSCALE $G^2$ แตกต่างกัน โดยใช้การทดสอบเปรียบเทียบพหุคูณ ซึ่งผลการทดสอบพบว่า ที่โมเดล GRM ขนาดกลุ่มตัวอย่าง 2000 คนจะให้อำนาจการทดสอบของดัชนี PARSCALE $G^2$ ที่มากกว่าขนาดกลุ่มตัวอย่าง 500 และ 1000 คนอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นอกจากนี้ ขนาดกลุ่มตัวอย่าง 1000 คนจะให้อำนาจการทดสอบของดัชนี PARSCALE $G^2$ ที่มากกว่าขนาดกลุ่มตัวอย่าง 500 คนอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ด้วย ซึ่งรายละเอียดค่าสถิติที่ใช้ในการทดสอบ ปรากฏดังในตารางที่ 4.16

**ตารางที่ 4.16 ผลการวิเคราะห์ความแตกต่างของค่าเฉลี่ยอำนาจการทดสอบ** (Power of the test)**ของดัชนี** PARSCALE $G^2$ **โมเดล** GRM **ในเงื่อนไขขนาดกลุ่มตัวอย่าง** (500, 1000, 2000 **คน**)

| **ขนาดกลุ่มตัวอย่าง** | **ค่าเฉลี่ย** | **ความแตกต่างระหว่างขนาดกลุ่มตัวอย่าง** | | |
|---|---|---|---|---|
| | | 500 คน | 1000 คน | 2000 คน |
| 500 คน | 0.218 | - | -0.197* | -0.454* |
| 1000 คน | 0.415 | | - | -0.257* |
| 2000 คน | 0.672 | | | - |
| $F$ = 13.985 | $p$ = 0.000 | | | |

* $p < .05$

จากผลการทดสอบที่พบว่าในโมเดล GPCM เงื่อนไขความยาวแบบวัดที่แตกต่างกัน ส่งผลให้อำนาจการทดสอบของดัชนี PARSCALE $G^2$ มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ ดังนั้นจึงต้องทำการทดสอบต่อไปว่า ณ ขนาดความยาวแบบวัดคู่ใดบ้างที่ให้อำนาจการทดสอบของดัชนี PARSCALE $G^2$ แตกต่างกัน โดยใช้การทดสอบเปรียบเทียบพหุคูณ ซึ่งผลการทดสอบพบว่า ที่โมเดล GPCM ความยาวแบบวัด 40 ข้อจะให้อำนาจการทดสอบของดัชนี PARSCALE $G^2$ ที่มากกว่าความยาวแบบวัด 10 และ 20 ข้อ อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ซึ่งรายละเอียดค่าสถิติที่ใช้ในการทดสอบ ปรากฏดังในตารางที่ 4.17

**ตารางที่ 4.17 ผลการวิเคราะห์ความแตกต่างของค่าเฉลี่ยอำนาจการทดสอบ** (Power of the test) **ของดัชนี** PARSCALE $G^2$ **โมเดล** GPCM **ในเงื่อนไขความยาวแบบวัด** (10, 20, 40 **ข้อ**)

| **ความยาวแบบวัด** | **ค่าเฉลี่ย** | **ความแตกต่างระหว่างความยาวแบบวัด** | | |
|---|---|---|---|---|
| | | 10 ข้อ | 20 ข้อ | 40 ข้อ |
| 10 ข้อ | 0.332 | - | -0.096 | -0.371* |
| 20 ข้อ | 0.428 | | - | -0.275* |
| 40 ข้อ | 0.703 | | | - |
| $F$ = 5.943 | $p$ = 0.006 | | | |

* $p < .05$

จากผลการทดสอบที่พบว่าในโมเดล GPCM เงื่อนไขขนาดกลุ่มตัวอย่างที่แตกต่างกัน ส่งผลให้อำนาจการทดสอบของดัชนี PARSCALE $G^2$ มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ ดังนั้นจึงต้องทำการทดสอบต่อไปว่า ณ ขนาดกลุ่มตัวอย่างคู่ใดบ้างที่ให้อำนาจการทดสอบของดัชนี PARSCALE $G^2$ แตกต่างกัน โดยใช้การทดสอบเปรียบเทียบพหุคูณ ซึ่งผลการทดสอบพบว่า ที่โมเดล GPCM ขนาดกลุ่มตัวอย่าง 2000 คนจะให้อำนาจการทดสอบของดัชนี PARSCALE $G^2$ ที่มากกว่าขนาดกลุ่มตัวอย่าง 500 คนอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ซึ่งรายละเอียดค่าสถิติที่ใช้ในการทดสอบ ปรากฏดังในตารางที่ 4.18

**ตารางที่ 4.18 ผลการวิเคราะห์ความแตกต่างของค่าเฉลี่ยอำนาจการทดสอบ** (Power of the test) **ของดัชนี** PARSCALE $G^2$ **โมเดล** GPCM **ในเงื่อนไขขนาดกลุ่มตัวอย่าง** (500, 1000, 2000 **คน**)

| **ขนาดกลุ่มตัวอย่าง** | **ค่าเฉลี่ย** | **ความแตกต่างระหว่างขนาดกลุ่มตัวอย่าง** | | |
|---|---|---|---|---|
| | | 500 คน | 1000 คน | 2000 คน |
| 500 คน | 0.292 | - | -0.197 | -0.392* |
| 1000 คน | 0.489 | | - | -0.195 |
| 2000 คน | 0.684 | | | - |
| $F$ = 6.252 | $p$ = 0.005 | | | |

* $p < .05$

3.3 อำนาจการทดสอบ (Power of the test) ของดัชนี Generalized $S-\chi^2$

การวิเคราะห์ข้อมูลในส่วนนี้ เป็นการนำเสนอข้อมูลอำนาจการทดสอบของดัชนี Generalized $S-\chi^2$ จำแนกตามโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาคที่ใช้ในการศึกษา การนำเสนอข้อมูลนี้มีวัตถุประสงค์เพื่อการพิจารณาลักษณะของอำนาจการทดสอบของดัชนี Generalized $S-\chi^2$ ทั้ง 72 สถานการณ์ที่ทำการศึกษา ว่ามีการเปลี่ยนแปลงไปตามความยาวแบบวัด ขนาดกลุ่มตัวอย่าง และจำนวนรายการคำตอบที่เปลี่ยนแปลงไปอย่างไร

เมื่อพิจารณาที่อำนาจการทดสอบของดัชนี Generalized $S-\chi^2$ ทั้ง 72 สถานการณ์ ที่โมเดล GRM พบว่า เมื่อความยาวแบบวัดเพิ่มขึ้น โดยส่วนใหญ่ อำนาจการทดสอบจะมีค่าลดลง มีเพียงบางกรณีเท่านั้นที่อำนาจการทดสอบเพิ่มขึ้น เช่น ในกรณี ณ ความยาวแบบวัดเปลี่ยนแปลงจาก 10 ข้อไป 20 ข้อ และ ณ ความยาวแบบวัดเปลี่ยนแปลงจาก 20 ข้อไป 40 ข้อ ที่ขนาดกลุ่มตัวอย่าง 1000 คน ในจำนวนรายการคำตอบ 9 รายการ และเมื่อพิจารณาที่โมเดล

GPCM พบว่า โดยส่วนใหญ่ เมื่อความยาวแบบวัดเพิ่มขึ้น อำนาจการทดสอบจะมีค่าลดลง โดยเฉพาะ ณ ความยาวแบบวัดเปลี่ยนแปลงจาก 20 ข้อไป 40 ข้อ ขนาดกลุ่มตัวอย่าง 1000 และ 2000 คนในทุกจำนวนรายการคำตอบ และ ณ ความยาวแบบวัดเปลี่ยนแปลงจาก 10 ข้อไป 20 ข้อ ขนาดกลุ่มตัวอย่าง 500 และ 2000 คนในทุกจำนวนรายการคำตอบ

เมื่อพิจารณาที่ขนาดกลุ่มตัวอย่าง พบว่า โมเดล GRM เมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้น อำนาจการทดสอบจะมีค่าเพิ่มมากขึ้นในเกือบทุกกรณีที่ทำการศึกษา ยกเว้น ณ ขนาดกลุ่มตัวอย่างเปลี่ยนแปลงจาก 500 คนไป 1000 คน ที่ความยาวแบบวัด 20 ข้อ ในจำนวนรายการคำตอบ 3 และ 9 รายการ และ ณ ขนาดกลุ่มตัวอย่างเปลี่ยนแปลงจาก 1000 คนไป 2000 คน ที่ความยาวแบบวัด 40 ข้อ ในจำนวนรายการคำตอบ 9 รายการ และเมื่อพิจารณาที่โมเดล GPCM พบว่า โดยส่วนใหญ่โมเดล GPCM เมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้น อำนาจการทดสอบก็จะเพิ่มขึ้นเช่นเดียวกันกับโมเดล GRM แต่จะมีบางกรณีเท่านั้นที่อำนาจการทดสอบลดลง เช่น ณ ขนาดกลุ่มตัวอย่างเปลี่ยนแปลงจาก 500 คนไป 1000 คน ที่ความยาวแบบวัด 40 ข้อ ในจำนวนรายการคำตอบ 7 และ 9 รายการ และกรณี ณ ขนาดกลุ่มตัวอย่างเปลี่ยนแปลงจาก 1000 คนไป 2000 คน ที่ความยาวแบบวัด 10 ข้อ ในจำนวนรายการคำตอบ 7 และ 9 รายการ ที่ความยาวแบบวัด 20 ข้อ ในจำนวนรายการคำตอบ 3 และ 5 รายการ ที่ความยาวแบบวัด 40 ข้อ ในจำนวนรายการคำตอบ 3 และ 5 รายการ

เมื่อพิจารณาที่จำนวนรายการคำตอบ พบว่า โดยส่วนใหญ่โมเดล GRM เมื่อจำนวนรายการคำตอบเพิ่มขึ้น อำนาจการทดสอบจะเพิ่มขึ้น มีเพียงบางกรณีเท่านั้นที่อำนาจการทดสอบไม่เพิ่มขึ้น เช่น ในกรณีความยาวแบบวัด 10 ข้อ เช่น ณ กรณีจำนวนรายการคำตอบเปลี่ยนแปลงจาก 3 ไป 5 รายการ ความยาวแบบวัด 20 ข้อ ในขนาดตัวอย่าง 500 คน ณ กรณีจำนวนรายการคำตอบเปลี่ยนแปลงจาก 5 ไป 7 รายการ ที่ความยาวแบบวัด 20 ข้อ ขนาดตัวอย่าง 500 คน และที่ความยาวแบบวัด 20 และ 40 ข้อ ขนาดตัวอย่าง 2000 คน และ ณ กรณีจำนวนรายการคำตอบเปลี่ยนแปลงจาก 7 ไป 9 รายการ ที่ความยาวแบบวัด 10 ข้อ ในขนาดตัวอย่าง 500 และ 1000 คน ที่ความยาวแบบวัด 20 ข้อที่ขนาดตัวอย่าง 1000 คน เมื่อพิจารณาที่โมเดล GPCM พบว่า โดยส่วนใหญ่ เมื่อจำนวนรายการคำตอบเพิ่มขึ้น อำนาจการทดสอบจะเพิ่มขึ้น มีเพียงบางกรณีเท่านั้นที่อำนาจการทดสอบไม่เพิ่มขึ้น เช่น เมื่อจำนวนรายการคำตอบเปลี่ยนแปลงจาก 3 ไป 5 รายการ พบว่า เมื่อจำนวนรายการคำตอบเพิ่มขึ้น อำนาจการทดสอบจะลดลงในกรณีขนาดกลุ่มตัวอย่าง 500 คน ความยาวแบบวัด 10 ข้อ กรณีขนาดกลุ่มตัวอย่าง 1000 คน ความยาวแบบวัด 10 และ 20 ข้อ และกรณีขนาดกลุ่มตัวอย่าง 2000 คน ความยาวแบบวัด 10

และ 20 ข้อ นอกจากนี้ เมื่อจำนวนรายการคำตอบเปลี่ยนแปลงจาก 5 ไป 7 รายการ พบว่า เมื่อจำนวนรายการคำตอบเพิ่มขึ้น อำนาจการทดสอบจะลดลงในกรณีขนาดกลุ่มตัวอย่าง 1000 คน ความยาวแบบวัด 40 ข้อด้วย

รายละเอียดของอำนาจการทดสอบ ของดัชนี Generalized $S-\chi^2$ ทั้ง 72 สถานการณ์ ดังแสดงในตารางที่ 4.19

**ตารางที่** 4.19 **อำนาจการทดสอบ** (Power of the test) **ของดัชนี** Generalized $S-\chi^2$

| **โมเดล** | **ความยาวแบบวัด** | **ขนาดกลุ่มตัวอย่าง** | **จำนวนรายการคำตอบ** | | | |
|---|---|---|---|---|---|---|
| | | | 3 | 5 | 7 | 9 |
| GRM | 10 | 500 | 0.0200 | 0.0433 | 0.0633 | 0.0200 |
| | | 1000 | 0.0533 | 0.0767 | 0.1733 | 0.0433 |
| | | 2000 | 0.0600 | 0.2800 | 0.3400 | 0.3967 |
| GRM | 20 | 500 | 0.0350 | 0.0233 | 0.0233 | 0.1267 |
| | | 1000 | 0.0067 | 0.0417 | 0.1217 | 0.0850 |
| | | 2000 | 0.0617 | 0.1567 | 0.1450 | 0.2833 |
| GRM | 40 | 500 | 0.0008 | 0.0150 | 0.0217 | 0.0433 |
| | | 1000 | 0.0025 | 0.0200 | 0.0650 | 0.1125 |
| | | 2000 | 0.0083 | 0.0875 | 0.0750 | 0.0933 |
| GPCM | 10 | 500 | 0.0267 | 0.0233 | 0.0267 | 0.1900 |
| | | 1000 | 0.0667 | 0.0333 | 0.2600 | 0.4300 |
| | | 2000 | 0.1700 | 0.1333 | 0.2200 | 0.2967 |
| GPCM | 20 | 500 | 0.0100 | 0.0150 | 0.0233 | 0.0467 |
| | | 1000 | 0.0800 | 0.0750 | 0.1733 | 0.7567 |
| | | 2000 | 0.0417 | 0.0383 | 0.2033 | 0.2817 |
| GPCM | 40 | 500 | 0.0075 | 0.0150 | 0.0800 | 0.1708 |
| | | 1000 | 0.0200 | 0.0442 | 0.0292 | 0.1108 |
| | | 2000 | 0.0133 | 0.0267 | 0.1083 | 0.1867 |

จากการพิจารณาอำนาจการทดสอบของดัชนี Generalized $S-\chi^2$ จำแนกตามโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาคที่ใช้ในการศึกษาว่ามีการเปลี่ยนแปลงไป

ตามความยาวแบบวัด ขนาดกลุ่มตัวอย่าง และจำนวนรายการคำตอบที่เปลี่ยนแปลงไปอย่างไรตามรายละเอียดดังที่กล่าวมาแล้วนั้น ผู้วิจัยจึงได้นำอำนาจการทดสอบของดัชนี Generalized $S-\chi^2$ ไปวิเคราะห์เปรียบเทียบโดยมีวัตถุประสงค์เพื่อต้องการทดสอบว่าเมื่อเงื่อนไขที่ใช้ในการศึกษา (ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ) เปลี่ยนแปลงไป จะส่งผลต่ออำนาจการทดสอบของดัชนี Generalized $S-\chi^2$ อย่างมีนัยสำคัญทางสถิติหรือไม่ ผู้วิจัยจึงได้ทำการทดสอบสมมติฐานทางสถิติที่ระดับความเชื่อมั่น 95 % เพื่อเปรียบเทียบค่าเฉลี่ยของอำนาจการทดสอบ ของดัชนี Generalized $S-\chi^2$ ในแต่ละเงื่อนไขที่ใช้ในการศึกษาจำแนกตามโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาคที่ใช้ในการศึกษา

ผลการทดสอบสมมติฐานทางสถิติ พบว่า เมื่อเป็นโมเดล GRM จะมีเพียงเงื่อนไขขนาดกลุ่มตัวอย่างที่อำนาจการทดสอบของดัชนี Generalized $S-\chi^2$ มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ ($p<$ .05) แสดงว่า มีอำนาจการทดสอบของดัชนี Generalized $S-\chi^2$ อย่างน้อย 2 ขนาดกลุ่มตัวอย่าง (1 คู่) ที่แตกต่างกันที่ระดับนัยสำคัญ .05 ในขณะที่โมเดล GPCM ที่มีเพียงเงื่อนไขจำนวนรายการคำตอบที่อำนาจการทดสอบของดัชนี Generalized $S-\chi^2$ มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ ($p<$ .05) แสดงว่า มีอำนาจการทดสอบของดัชนี Generalized $S-\chi^2$ อย่างน้อย 2 จำนวนรายการคำตอบ (1 คู่) ที่แตกต่างกันที่ระดับนัยสำคัญ .05 ซึ่งรายละเอียดค่าสถิติที่ใช้ในการทดสอบ ปรากฏดังในตารางที่ 4.20

**ตารางที่ 4.20 ผลการวิเคราะห์เปรียบเทียบค่าเฉลี่ยของอำนาจการทดสอบ (Power of the test) ของดัชนี Generalized $S-\chi^2$ ในแต่ละเงื่อนไขที่ใช้ในการศึกษา**

| โมเดล | เงื่อนไขที่ใช้ในการศึกษา | ค่าสถิติ | | | | |
|---|---|---|---|---|---|---|
| | | SS | df | MS | *F* | *p* |
| GRM | ความยาวแบบวัด (10, 20, 40 ข้อ) | | | | | |
| | ระหว่างกลุ่ม | 0.044 | 2 | 0.022 | 2.560 | 0.093 |
| | ภายในกลุ่ม | 0.283 | 33 | 0.009 | | |
| | รวม | 0.327 | 35 | | | |
| GRM | ขนาดกลุ่มตัวอย่าง (500, 1000, 2000 คน) | | | | | |
| | ระหว่างกลุ่ม | 0.110 | 2 | 0.055 | 8.321* | 0.001 |
| | ภายในกลุ่ม | 0.217 | 33 | 0.007 | | |
| | รวม | 0.327 | 35 | | | |

**ตารางที่** 4.20 (ต่อ)

| โมเดล | เงื่อนไขที่ใช้ในการศึกษา | ค่าสถิติ | | | | |
|---|---|---|---|---|---|---|
| | | SS | df | MS | *F* | *p* |
| GRM | จำนวนรายการคำตอบ (3, 5, 7, 9 รายการ) | | | | | |
| | ระหว่างกลุ่ม | 0.058 | 3 | 0.019 | 2.303 | 0.096 |
| | ภายในกลุ่ม | 0.269 | 32 | 0.008 | | |
| | รวม | 0.327 | 35 | | | |
| GPCM | ความยาวแบบวัด (10, 20, 40 ข้อ) | | | | | |
| | ระหว่างกลุ่ม | 0.056 | 2 | 0.028 | 1.290 | 0.289 |
| | ภายในกลุ่ม | 0.717 | 33 | 0.022 | | |
| | รวม | 0.773 | 35 | | | |
| GPCM | ขนาดกลุ่มตัวอย่าง (500, 1000, 2000 คน) | | | | | |
| | ระหว่างกลุ่ม | 0.094 | 2 | 0.047 | 2.289 | 0.117 |
| | ภายในกลุ่ม | 0.679 | 33 | 0.021 | | |
| | รวม | 0.773 | 35 | | | |
| GPCM | จำนวนรายการคำตอบ (3, 5, 7, 9 รายการ) | | | | | |
| | ระหว่างกลุ่ม | 0.311 | 3 | 0.104 | 7.189* | 0.001 |
| | ภายในกลุ่ม | 0.462 | 32 | 0.014 | | |
| | รวม | 0.773 | 35 | | | |

* $p < .05$

จากผลการทดสอบที่พบว่าในโมเดล GRM เงื่อนไขขนาดกลุ่มตัวอย่างที่แตกต่างกัน ส่งผลให้อำนาจการทดสอบของดัชนี Generalized $S-\chi^2$ มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ ดังนั้นจึงต้องทำการทดสอบต่อไปว่า ณ ขนาดกลุ่มตัวอย่างคู่ใดบ้างที่ให้อำนาจการทดสอบของดัชนี Generalized $S-\chi^2$ แตกต่างกัน โดยใช้การทดสอบเปรียบเทียบพหุคูณ ซึ่งผลการทดสอบพบว่า ที่โมเดล GRM ขนาดกลุ่มตัวอย่าง 2000 คน จะให้อำนาจการทดสอบของดัชนี Generalized $S-\chi^2$ ที่มากกว่าขนาดกลุ่มตัวอย่าง 500 คนอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ซึ่งรายละเอียดค่าสถิติที่ใช้ในการทดสอบ ปรากฏดังในตารางที่ 4.21

**ตารางที่ 4.21 ผลการวิเคราะห์ความแตกต่างของค่าเฉลี่ยอำนาจการทดสอบ (Power of the test) ของดัชนี Generalized $S-\chi^2$ โมเดล GRM ในเงื่อนไขขนาดกลุ่มตัวอย่าง (500, 1000, 2000 คน)**

| ขนาดกลุ่มตัวอย่าง | ค่าเฉลี่ย | ความแตกต่างระหว่างขนาดกลุ่มตัวอย่าง | | |
|---|---|---|---|---|
| | | 500 คน | 1000 คน | 2000 คน |
| 500 คน | 0.036 | - | -0.031 | -0.130* |
| 1000 คน | 0.067 | | - | -0.099 |
| 2000 คน | 0.166 | | | - |
| $F$ = 8.321 | $p$ = 0.001 | | | |

* $p < .05$

จากผลการทดสอบที่พบว่าในโมเดล GPCM เงื่อนไขจำนวนรายการคำตอบที่แตกต่างกัน ส่งผลให้อำนาจการทดสอบของดัชนี Generalized $S-\chi^2$ มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ ดังนั้นจึงต้องทำการทดสอบต่อไปว่า ณ จำนวนรายการคำตอบคู่ใดบ้างที่ให้อำนาจการทดสอบของดัชนี Generalized $S-\chi^2$ แตกต่างกัน โดยใช้การทดสอบเปรียบเทียบพหุคูณ ซึ่งผลการทดสอบพบว่า ที่โมเดล GPCM จำนวนรายการคำตอบ 9 รายการจะให้อำนาจการทดสอบของดัชนี Generalized $S-\chi^2$ ที่มากกว่าจำนวนรายการคำตอบ 5 รายการอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ซึ่งรายละเอียดค่าสถิติที่ใช้ในการทดสอบ ปรากฏดังในตารางที่ 4.22

**ตารางที่ 4.22 ผลการวิเคราะห์ความแตกต่างของค่าเฉลี่ยอำนาจการทดสอบ (Power of the test) ของดัชนี Generalized $S-\chi^2$ โมเดล GPCM ในเงื่อนไขจำนวนรายการคำตอบ (3, 5, 7, 9 รายการ)**

| จำนวนรายการคำตอบ | ค่าเฉลี่ย | ความแตกต่างระหว่างจำนวนรายการคำตอบ | | | |
|---|---|---|---|---|---|
| | | 3 รายการ | 5 รายการ | 7 รายการ | 9 รายการ |
| 3 รายการ | 0.048 | - | 0.003 | -0.077 | -0.226 |
| 5 รายการ | 0.045 | | - | -0.080 | -0.229* |
| 7 รายการ | 0.125 | | | - | -0.149 |
| 9 รายการ | 0.274 | | | | - |
| $F$ = 7.189 | $p$ = 0.001 | | | | |

* $p < .05$

ผลการวิเคราะห์อำนาจการทดสอบของดัชนี PARSCALE $G^2$ สรุปสาระสำคัญได้ว่า เมื่อนำอำนาจการทดสอบไปทดสอบเพื่อสังเกตความเปลี่ยนแปลงของอำนาจการทดสอบเมื่อเงื่อนไขที่ใช้ในการศึกษา (ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ) มีการเปลี่ยนแปลงไป จะพบว่าในโมเดล GRM มีเพียงเงื่อนไขความยาวแบบวัดและขนาดกลุ่มตัวอย่างเท่านั้นที่มีการเปลี่ยนแปลงของอำนาจการทดสอบอย่างมีนัยสำคัญทางสถิติ โดยสอดคล้องกับผลการพิจารณาของผู้วิจัยว่า ที่โมเดล GRM เมื่อความยาวแบบวัดเพิ่มขึ้น โดยส่วนใหญ่อำนาจการทดสอบจะมีค่าลดลงมีเพียงบางกรณีเท่านั้นที่อำนาจการทดสอบเพิ่มขึ้น และที่โมเดล GRM เมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้น อำนาจการทดสอบจะมีค่าเพิ่มมากขึ้นในทุกกรณีที่ทำการศึกษา ดังในผลการทดสอบเปรียบเทียบพหุคูณ (Multiple Comparison) ที่ให้สารสนเทศว่า ในโมเดล GRM ความยาวแบบวัด 10 ข้อ จะให้อำนาจการทดสอบของดัชนี PARSCALE $G^2$ มากกว่าความยาวแบบวัด 40 ข้อ และในโมเดล GRM ขนาดกลุ่มตัวอย่าง 2000 คน จะให้อำนาจการทดสอบของดัชนี PARSCALE $G^2$ มากกว่าขนาดกลุ่มตัวอย่าง 500 และ 1000 คน นอกจากนี้ในโมเดล GRM ขนาดกลุ่มตัวอย่าง 1000 คน จะให้อำนาจการทดสอบของดัชนี PARSCALE $G^2$ มากกว่าขนาดกลุ่มตัวอย่าง 500 คนด้วย เมื่อพิจารณาที่โมเดล GPCM จะมี 2 เงื่อนไขเช่นเดียวกันกับโมเดล GRM คือเงื่อนไขความยาวแบบวัดและขนาดกลุ่มตัวอย่างเท่านั้นที่มีการเปลี่ยนแปลงของอำนาจการทดสอบอย่างมีนัยสำคัญทางสถิติ โดยสอดคล้องกับผลการพิจารณาของผู้วิจัยว่า ที่โมเดล GPCM เมื่อความยาวแบบวัดเพิ่มขึ้น ก็มีแนวโน้มที่อำนาจการทดสอบจะเพิ่มขึ้นด้วย และที่โมเดล GPCM เมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้นอำนาจการทดสอบก็จะเพิ่มขึ้นเช่นเดียวกับโมเดล GRM ดังในผลการทดสอบเปรียบเทียบพหุคูณ (Multiple Comparison) ที่ให้สารสนเทศว่า ในโมเดล GPCM ความยาวแบบวัด 40 ข้อ ให้อำนาจการทดสอบมากกว่าความยาวแบบวัด 10 และ 20 ข้อ และในโมเดล GPCM ขนาดกลุ่มตัวอย่าง 2000 คน ให้อำนาจการทดสอบมากกว่าขนาดกลุ่มตัวอย่าง 500 คน

ผลการวิเคราะห์อำนาจการทดสอบของดัชนี Generalized $S-\chi^2$ สรุปสาระสำคัญได้ว่า เมื่อนำอำนาจการทดสอบไปทดสอบเพื่อสังเกตความเปลี่ยนแปลงของอำนาจการทดสอบเมื่อเงื่อนไขที่ใช้ในการศึกษา (ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ) มีการเปลี่ยนแปลงไป จะพบว่าในโมเดล GRM มีเพียงเงื่อนไขขนาดกลุ่มตัวอย่างเท่านั้นที่มีการเปลี่ยนแปลงของอำนาจการทดสอบอย่างมีนัยสำคัญทางสถิติ โดยสอดคล้องกับผลการพิจารณาของผู้วิจัยว่า ที่โมเดล GRM เมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้น อำนาจการทดสอบจะมีค่าเพิ่มมากขึ้น และที่โมเดล GPCM มีเพียงเงื่อนไขจำนวนรายการคำตอบเท่านั้นที่มีการเปลี่ยนแปลง

ของอำนาจการทดสอบอย่างมีนัยสำคัญทางสถิติ โดยสอดคล้องกับผลการพิจารณาของผู้วิจัยว่าโดยส่วนใหญ่ เมื่อจำนวนรายการคำตอบเพิ่มขึ้น อำนาจการทดสอบจะเพิ่มขึ้น ดังในผลการทดสอบเปรียบเทียบพหุคูณ (Multiple Comparison) ที่ให้สารสนเทศว่า ในโมเดล GRM ขนาดกลุ่มตัวอย่าง 2000 คน จะให้อำนาจการทดสอบของดัชนี Generalized $S-\chi^2$ มากกว่าขนาดกลุ่มตัวอย่าง 500 คน และในโมเดล GPCM จำนวนรายการคำตอบ 9 รายการ จะให้อำนาจการทดสอบของดัชนี Generalized $S-\chi^2$ มากกว่าจำนวนรายการคำตอบ 5 รายการ

จะเห็นได้ว่า ผลการทดสอบเปรียบเทียบพหุคูณ (Multiple Comparison) ของดัชนีทั้งสองชนิดให้สารสนเทศที่สอดคล้องกันในโมเดล GRM ว่าขนาดกลุ่มตัวอย่างมาก (2000 คน) จะให้อำนาจการทดสอบของดัชนีทั้งสองชนิดมากกว่าขนาดกลุ่มตัวอย่างน้อย (500 คน) แสดงให้เห็นว่า ขนาดกลุ่มตัวอย่างมีผลต่ออำนาจการทดสอบของดัชนีทั้งสองชนิดเหมือนกัน

**ตอนที่ 4 ผลการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามสองชนิดคือ** Generalized $S-\chi^2$ **และ** PARSCALE $G^2$

จากตอนที่ 2 และตอนที่ 3 ของการนำเสนอผลการวิเคราะห์ในบทนี้ ได้นำเสนอผลการวิเคราะห์ค่าความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของดัชนีทั้ง 2 ชนิด คือ ดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ จำแนกตามโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค โดยมีการพิจารณาลักษณะของค่าความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของดัชนีทั้ง 2 ชนิด ใน 72 สถานการณ์ที่ทำการศึกษา ว่ามีการเปลี่ยนแปลงไปตามความยาวแบบวัด ขนาดกลุ่มตัวอย่าง และจำนวนรายการคำตอบที่เปลี่ยนแปลงไปอย่างไร นอกจากนี้ยังได้มีการศึกษาว่า เมื่อเงื่อนไขที่ใช้ในการศึกษา (ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ) เปลี่ยนแปลงไป จะส่งผลต่อค่าความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของดัชนีทั้ง 2 ชนิดอย่างมีนัยสำคัญทางสถิติหรือไม่ โดยทดสอบสมมติฐานทางสถิติที่ระดับความเชื่อมั่น 95 % เพื่อเปรียบเทียบค่าเฉลี่ยของค่าความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของดัชนีทั้ง 2 ชนิด ในแต่ละเงื่อนไขที่ใช้ในการศึกษา (ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ)

ในตอนที่ 4 นี้จะนำเสนอผลการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามสองชนิด จำแนกออกเป็น 2 ส่วน คือ 1.ผลการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดคือ Generalized $S-\chi^2$ และ PARSCALE $G^2$ ตามเงื่อนไขของ Kang และ Chen (2008) 2.ผลการเปรียบเทียบประสิทธิภาพของดัชนีความ

สอดคล้องของข้อคำถามทั้งสองชนิด คือ Generalized $S-\chi^2$ และ PARSCALE $G^2$ โดยใช้สถิติการวิเคราะห์ความแปรปรวนแบบ 2 ทาง (two-way ANOVA)

4.1 ผลการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามสองชนิดคือ Generalized $S-\chi^2$ และ PARSCALE $G^2$ ตามเงื่อนไขของ Kang และ Chen (2008)

ขั้นตอนในการพิจารณาเพื่อเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถาม (item fit index) ตามเงื่อนไขของ Kang และ Chen (2008) จะพิจารณาที่ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ก่อน ซึ่งถ้าดัชนีชนิดใดมีค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) น้อยกว่า ดัชนีนั้นจะเป็นดัชนีที่มีประสิทธิภาพในการบ่งชี้ความสอดคล้องของข้อคำถามในสถานการณ์นั้น แต่ถ้าหากดัชนีทั้งสองมีค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) เท่ากันจึงจะพิจารณาอำนาจการทดสอบ (Power of the test) โดยถ้าดัชนีชนิดใดมีอำนาจการทดสอบที่มากกว่า ดัชนีนั้นจะเป็นดัชนีที่มีประสิทธิภาพในการบ่งชี้ความสอดคล้องของข้อคำถามกับโมเดลทฤษฎีการตอบสนองข้อสอบในสถานการณ์นั้น ซึ่งหลังจากที่ได้ดำเนินการคำนวณค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) และอำนาจการทดสอบ (Power of the test) ของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดตามขั้นตอนการดำเนินการวิจัย ปรากฏผลค่าความคลาดเคลื่อนประเภทที่ 1และอำนาจการทดสอบเพื่อนำไปใช้ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถาม (item fit index) ดังตารางที่ 4.23

**ตารางที่ 4.23 การเปรียบเทียบค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) และอำนาจการทดสอบ (Power of the test) ของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิด คือ Generalized $S-\chi^2$ และ PARSCALE $G^2$**

| ลำดับที่ | โมเดล | ความยาวแบบวัด | ขนาดกลุ่มตัวอย่าง | จำนวนรายการคำตอบ | ค่าความคลาดเคลื่อนประเภทที่ 1 | | อำนาจการทดสอบ | |
|---|---|---|---|---|---|---|---|---|
| | | | | | PARSCALE $G^2$ | Generalized $S-\chi^2$ | PARSCALE $G^2$ | Generalized $S-\chi^2$ |
| 1 | GRM | 10 | 500 | 3 | 0.5467 | 0.0167 | 0.6033 | 0.0200 |
| 2 | GRM | 10 | 500 | 5 | 0.0767 | 0.0367 | 0.2400 | 0.0433 |
| 3 | GRM | 10 | 500 | 7 | 0.0133 | 0.0400 | 0.2967 | 0.0633 |
| 4 | GRM | 10 | 500 | 9 | 0.0200 | 0.0033 | 0.2069 | 0.0200 |
| 5 | GRM | 10 | 1000 | 3 | 0.6900 | 0.0633 | 0.7700 | 0.0533 |
| 6 | GRM | 10 | 1000 | 5 | 0.1467 | 0.0033 | 0.4467 | 0.0767 |
| 7 | GRM | 10 | 1000 | 7 | 0.1167 | 0.0167 | 0.6967 | 0.1733 |

**ตารางที่** 4.23 (ต่อ)

| ลำดับที่ | โมเดล | ความยาวแบบวัด | ขนาดกลุ่มตัวอย่าง | จำนวนรายการคำตอบ | ค่าความคลาดเคลื่อนประเภทที่ 1 | | อำนาจการทดสอบ | |
|---|---|---|---|---|---|---|---|---|
| | | | | | PARSCALE $G^2$ | Generalized $S-\chi^2$ | PARSCALE $G^2$ | Generalized $S-\chi^2$ |
| 8 | GRM | 10 | 1000 | 9 | 0.0567 | 0.0000 | 0.3933 | 0.0433 |
| 9 | GRM | 10 | 2000 | 3 | 0.9100 | 0.0167 | 0.9567 | 0.0600 |
| 10 | GRM | 10 | 2000 | 5 | 0.6067 | 0.0067 | 0.8933 | 0.2800 |
| 11 | GRM | 10 | 2000 | 7 | 0.3333 | 0.0033 | 0.9067 | 0.3400 |
| 12 | GRM | 10 | 2000 | 9 | 0.1133 | 0.0000 | 0.8833 | 0.3967 |
| 13 | GRM | 20 | 500 | 3 | 0.1167 | 0.0017 | 0.0233 | 0.0350 |
| 14 | GRM | 20 | 500 | 5 | 0.0183 | 0.0033 | 0.1700 | 0.0233 |
| 15 | GRM | 20 | 500 | 7 | 0.0050 | 0.0000 | 0.1217 | 0.0233 |
| 16 | GRM | 20 | 500 | 9 | 0.0133 | 0.0000 | 0.3183 | 0.1267 |
| 17 | GRM | 20 | 1000 | 3 | 0.0500 | 0.0167 | 0.0517 | 0.0067 |
| 18 | GRM | 20 | 1000 | 5 | 0.0183 | 0.0000 | 0.3233 | 0.0417 |
| 19 | GRM | 20 | 1000 | 7 | 0.0067 | 0.0000 | 0.4417 | 0.1217 |
| 20 | GRM | 20 | 1000 | 9 | 0.0133 | 0.0000 | 0.4650 | 0.0850 |
| 21 | GRM | 20 | 2000 | 3 | 0.1217 | 0.0100 | 0.3167 | 0.0617 |
| 22 | GRM | 20 | 2000 | 5 | 0.0500 | 0.0033 | 0.6217 | 0.1567 |
| 23 | GRM | 20 | 2000 | 7 | 0.0233 | 0.0000 | 0.7367 | 0.1450 |
| 24 | GRM | 20 | 2000 | 9 | 0.0267 | 0.0000 | 0.8200 | 0.2833 |
| 25 | GRM | 40 | 500 | 3 | 0.1317 | 0.0008 | 0.0300 | 0.0008 |
| 26 | GRM | 40 | 500 | 5 | 0.0692 | 0.0008 | 0.1525 | 0.0150 |
| 27 | GRM | 40 | 500 | 7 | 0.0092 | 0.0000 | 0.2100 | 0.0217 |
| 28 | GRM | 40 | 500 | 9 | 0.0158 | 0.0017 | 0.2467 | 0.0433 |
| 29 | GRM | 40 | 1000 | 3 | 0.0475 | 0.0058 | 0.0450 | 0.0025 |
| 30 | GRM | 40 | 1000 | 5 | 0.0133 | 0.0008 | 0.3842 | 0.0200 |
| 31 | GRM | 40 | 1000 | 7 | 0.0167 | 0.0008 | 0.4258 | 0.0650 |
| 32 | GRM | 40 | 1000 | 9 | 0.0183 | 0.0333 | 0.5333 | 0.1125 |
| 33 | GRM | 40 | 2000 | 3 | 0.0783 | 0.0025 | 0.1242 | 0.0083 |
| 34 | GRM | 40 | 2000 | 5 | 0.0200 | 0.0025 | 0.6200 | 0.0875 |
| 35 | GRM | 40 | 2000 | 7 | 0.0125 | 0.0008 | 0.5417 | 0.0750 |
| 36 | GRM | 40 | 2000 | 9 | 0.0142 | 0.0000 | 0.6383 | 0.0933 |
| 37 | GPCM | 10 | 500 | 3 | 0.0367 | 0.0200 | 0.0567 | 0.0267 |
| 38 | GPCM | 10 | 500 | 5 | 0.0033 | 0.0000 | 0.0400 | 0.0233 |
| 39 | GPCM | 10 | 500 | 7 | 0.0167 | 0.0033 | 0.0767 | 0.0267 |
| 40 | GPCM | 10 | 500 | 9 | 0.1200 | 0.1167 | 0.1667 | 0.1900 |
| 41 | GPCM | 10 | 1000 | 3 | 0.4833 | 0.0200 | 0.4800 | 0.0667 |

**ตารางที่** 4.23 (ต่อ)

| ลำดับที่ | โมเดล | ความยาวแบบวัด | ขนาดกลุ่มตัวอย่าง | จำนวนรายการคำตอบ | ค่าความคลาดเคลื่อนประเภทที่ 1 | | อำนาจการทดสอบ | |
|---|---|---|---|---|---|---|---|---|
| | | | | | PARSCALE $G^2$ | Generalized $S-\chi^2$ | PARSCALE $G^2$ | Generalized $S-\chi^2$ |
| 42 | GPCM | 10 | 1000 | 5 | 0.1033 | 0.0033 | 0.2200 | 0.0333 |
| 43 | GPCM | 10 | 1000 | 7 | 0.2733 | 0.1167 | 0.2367 | 0.2600 |
| 44 | GPCM | 10 | 1000 | 9 | 0.2367 | 0.1533 | 0.2433 | 0.4300 |
| 45 | GPCM | 10 | 2000 | 3 | 0.5367 | 0.0700 | 0.6500 | 0.1700 |
| 46 | GPCM | 10 | 2000 | 5 | 0.4300 | 0.0133 | 0.5700 | 0.1333 |
| 47 | GPCM | 10 | 2000 | 7 | 0.1067 | 0.0000 | 0.5833 | 0.2200 |
| 48 | GPCM | 10 | 2000 | 9 | 0.3433 | 0.1400 | 0.6633 | 0.2967 |
| 49 | GPCM | 20 | 500 | 3 | 0.0083 | 0.0033 | 0.0550 | 0.0100 |
| 50 | GPCM | 20 | 500 | 5 | 0.0267 | 0.0050 | 0.2817 | 0.0150 |
| 51 | GPCM | 20 | 500 | 7 | 0.2083 | 0.0317 | 0.1517 | 0.0233 |
| 52 | GPCM | 20 | 500 | 9 | 0.0567 | 0.0517 | 0.1700 | 0.0467 |
| 53 | GPCM | 20 | 1000 | 3 | 0.0217 | 0.0050 | 0.2600 | 0.0800 |
| 54 | GPCM | 20 | 1000 | 5 | 0.0200 | 0.0000 | 0.1850 | 0.0750 |
| 55 | GPCM | 20 | 1000 | 7 | 0.1700 | 0.1033 | 0.5300 | 0.1733 |
| 56 | GPCM | 20 | 1000 | 9 | 0.1400 | 0.0533 | 0.7867 | 0.7567 |
| 57 | GPCM | 20 | 2000 | 3 | 0.3083 | 0.0100 | 0.3850 | 0.0417 |
| 58 | GPCM | 20 | 2000 | 5 | 0.0533 | 0.0000 | 0.5133 | 0.0383 |
| 59 | GPCM | 20 | 2000 | 7 | 0.0867 | 0.0000 | 0.8850 | 0.2033 |
| 60 | GPCM | 20 | 2000 | 9 | 0.2833 | 0.0000 | 0.9383 | 0.2817 |
| 61 | GPCM | 40 | 500 | 3 | 0.0158 | 0.0025 | 0.5242 | 0.0075 |
| 62 | GPCM | 40 | 500 | 5 | 0.0308 | 0.0017 | 0.3158 | 0.0150 |
| 63 | GPCM | 40 | 500 | 7 | 0.1058 | 0.0267 | 0.6950 | 0.0800 |
| 64 | GPCM | 40 | 500 | 9 | 0.7225 | 0.1492 | 0.9650 | 0.1708 |
| 65 | GPCM | 40 | 1000 | 3 | 0.0275 | 0.0033 | 0.4325 | 0.0200 |
| 66 | GPCM | 40 | 1000 | 5 | 0.1283 | 0.0017 | 0.6817 | 0.0442 |
| 67 | GPCM | 40 | 1000 | 7 | 0.1542 | 0.0192 | 0.8208 | 0.0292 |
| 68 | GPCM | 40 | 1000 | 9 | 0.1942 | 0.0717 | 0.9867 | 0.1108 |
| 69 | GPCM | 40 | 2000 | 3 | 0.0175 | 0.0050 | 0.2133 | 0.0133 |
| 70 | GPCM | 40 | 2000 | 5 | 0.3692 | 0.0000 | 0.8758 | 0.0267 |
| 71 | GPCM | 40 | 2000 | 7 | 0.2842 | 0.0025 | 0.9308 | 0.1083 |
| 72 | GPCM | 40 | 2000 | 9 | 0.3867 | 0.0625 | 0.9983 | 0.1867 |

จากนั้นได้ดำเนินการตามขั้นตอนในการพิจารณาเพื่อเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถาม (item fit index) โดยใช้หลักการพิจารณาตามเงื่อนไขของ

Kang และ Chen (2008) พบว่า ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนี Generalized $\boldsymbol{S-\chi^2}$ มีค่าน้อยกว่าดัชนี PARSCALE $G^2$ ในเกือบทุกสถานการณ์ที่ทำการศึกษา ยกเว้นสถานการณ์ที่ 3 และสถานการณ์ที่ 32 ที่ดัชนี PARSCALE $G^2$ จะมีค่าความคลาดเคลื่อนประเภทที่ 1 น้อยกว่าดัชนี Generalized $\boldsymbol{S-\chi^2}$ ดังนั้น ณ การพิจารณาโดยใช้เกณฑ์ค่าความคลาดเคลื่อนประเภทที่ 1 จะพบว่า ในโมเดล GRM นั้น ดัชนี Generalized $\boldsymbol{S-\chi^2}$ มีประสิทธิภาพในการบ่งชี้ความสอดคล้องของข้อคำถามมากกว่าดัชนี PARSCALE $G^2$ ในเกือบทุกสถานการณ์ที่ทำการศึกษา ยกเว้น ใน 2 สถานการณ์ ประกอบด้วยสถานการณ์ที่ 3 กรณีโมเดล GRM ความยาวแบบวัด 10 ข้อ ขนาดกลุ่มตัวอย่าง 500 คน จำนวนรายการคำตอบ 7 รายการ และสถานการณ์ที่ 32 กรณีโมเดล GRM ความยาวแบบวัด 40 ข้อ ขนาดกลุ่มตัวอย่าง 1000 คน จำนวนรายการคำตอบ 9 รายการ ที่ดัชนี PARSCALE $G^2$ มีประสิทธิภาพในการบ่งชี้ความสอดคล้องของข้อคำถามมากกว่าดัชนี Generalized $\boldsymbol{S-\chi^2}$ นอกจากนี้เมื่อพิจารณาในโมเดล GPCM นั้น จะพบว่า ดัชนี Generalized $\boldsymbol{S-\chi^2}$ มีประสิทธิภาพในการบ่งชี้ความสอดคล้องของข้อคำถามมากกว่าดัชนี PARSCALE $G^2$ ในทุกสถานการณ์ที่ทำการศึกษา

ในการพิจารณาเพื่อเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามโดยใช้หลักการพิจารณาตามเงื่อนไขของ Kang และ Chen (2008) จึงไม่มีสถานการณ์ใดเลยที่ใช้อำนาจการทดสอบ (Power of the test) เป็นเกณฑ์ในการเปรียบเทียบ เนื่องจาก ณ การพิจารณาโดยใช้เกณฑ์ค่าความคลาดเคลื่อนประเภทที่ 1 สามารถสรุปผลการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามได้แล้ว

ดังนั้น ในการศึกษาวิจัยเพื่อเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้ง 72 สถานการณ์ที่ทำการศึกษา สามารถสรุปผลได้ว่า ทั้ง 70 สถานการณ์ที่ทำการศึกษาดัชนี Generalized $S-\chi^2$ มีประสิทธิภาพในการบ่งชี้ความสอดคล้องของข้อคำถามมากกว่าดัชนี PARSCALE $G^2$ ยกเว้น สถานการณ์ที่ 3 กรณีโมเดล GRM ความยาวแบบวัด 10 ข้อ ขนาดกลุ่มตัวอย่าง 500 คน จำนวนรายการคำตอบ 7 รายการ และสถานการณ์ที่ 32 กรณีโมเดล GRM ความยาวแบบวัด 40 ข้อ ขนาดกลุ่มตัวอย่าง 1000 คน จำนวนรายการคำตอบ 9 รายการ ที่ดัชนี PARSCALE $G^2$ มีประสิทธิภาพในการบ่งชี้ความสอดคล้องของข้อคำถามมากกว่าดัชนี Generalized $S-\chi^2$

4.2 ผลการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิด คือ Generalized $S-\chi^2$ และ PARSCALE $G^2$ โดยใช้สถิติการวิเคราะห์ความแปรปรวนแบบ 2 ทาง (two-way ANOVA)

จากผลการวิเคราะห์ข้อมูลดังที่กล่าวมาแล้ว ผู้วิจัยมีความสนใจศึกษาต่อไปโดยการทดสอบผลของชนิดของดัชนีความสอดคล้องของข้อคำถาม และผลของปฏิสัมพันธ์ระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามกับเงื่อนไขที่ใช้ในการศึกษา(ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ) ว่ามีนัยสำคัญทางสถิติหรือไม่ ซึ่งถ้าชนิดของดัชนีความสอดคล้องของข้อคำถามกับเงื่อนไขที่ใช้ในการศึกษา (ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ) นั้นมีอิทธิพลปฏิสัมพันธ์ (interaction effect) กันย่อมแสดงให้เห็นว่า ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดภายใต้เงื่อนไขที่ใช้ในการศึกษา (ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ) ที่ต่างกัน จะส่งผลให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) และอำนาจการทดสอบ (Power of the test) มีค่าแตกต่างกัน นำไปสู่การสรุปผลเกี่ยวกับประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามที่แตกต่างกัน

ในการทดสอบเพื่อเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามโดยใช้สถิติการวิเคราะห์ความแปรปรวนแบบ 2 ทาง (two-way ANOVA) จำแนกได้ทั้งหมด 12 กรณี แบ่งออกเป็นการวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) 6 กรณีและการวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบ (Power of the test) 6 กรณี ดังนี้

4.2.1 การวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error)

4.2.1.1 การวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขความยาวแบบวัดที่ต่างกัน 3 ระดับ กรณีโมเดล GRM

4.2.1.2 การวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างที่ต่างกัน 3 ระดับ กรณีโมเดล GRM

4.2.1.3 การวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 ในการเปรียบเทียบประสิทธิภาพของดัชนีความ

สอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขจำนวนรายการคำตอบที่ต่างกัน 4 ระดับ กรณีโมเดล GRM

4.2.1.4 การวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขความยาวแบบวัดที่ต่างกัน 3 ระดับ กรณีโมเดล GPCM

4.2.1.5 การวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างที่ต่างกัน 3 ระดับ กรณีโมเดล GPCM

4.2.1.6 การวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขจำนวนรายการคำตอบที่ต่างกัน 4 ระดับ กรณีโมเดล GPCM

4.2.2 การวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบ (Power of the test)

4.2.2.1 การวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขความยาวแบบวัดที่ต่างกัน 3 ระดับ กรณีโมเดล GRM

4.2.2.2 การวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างที่ต่างกัน 3 ระดับ กรณีโมเดล GRM

4.2.2.3 การวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อ

คำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขจำนวนรายการคำตอบที่ต่างกัน 4 ระดับ กรณีโมเดล GRM

4.2.2.4 การวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขความยาวแบบวัดที่ต่างกัน 3 ระดับ กรณีโมเดล GPCM

4.2.2.5 การวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างที่ต่างกัน 3 ระดับ กรณีโมเดล GPCM

4.2.2.6 การวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขจำนวนรายการคำตอบที่ต่างกัน 4 ระดับ กรณีโมเดล GPCM

ซึ่งในแต่ละกรณี มีรายละเอียดดังต่อไปนี้

4.2.1 การวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error)

<u>4.2.1.1 การวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขความยาวแบบวัดที่ต่างกัน 3 ระดับ กรณีโมเดล GRM</u>

ในโมเดล GRM เมื่อพิจารณาผลการทดสอบปฏิสัมพันธ์ 2 ทาง พบว่า มีอิทธิพลปฏิสัมพันธ์ (interaction effect) กันระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามและความยาวแบบวัดที่ระดับนัยสำคัญ.05 ($F$ = 7.688, $p$ = 0.001) ดังรายละเอียดในตารางที่ 4.24 แสดงว่า ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดภายใต้ความยาวแบบวัดที่ต่างกัน จะส่งผลให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) มีค่าแตกต่างกันด้วย นั่นคือ ชนิดของดัชนีความสอดคล้องของข้อคำถามและความยาวแบบวัดมีผลร่วมกันต่อค่าความคลาดเคลื่อนประเภทที่ 1

**ตารางที่ 4.24 ผลการวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขความยาวแบบวัดที่ต่างกัน 3 ระดับ กรณีโมเดล GRM**

| แหล่งของความแปรปรวน | SS | df | MS | *F* | *p* |
|---|---|---|---|---|---|
| A (ชนิดของดัชนีความสอดคล้องของข้อคำถาม) | 0.251 | 1 | 0.251 | 15.305* | 0.000 |
| B (ความยาวแบบวัด) | 0.310 | 2 | 0.155 | 9.455* | 0.000 |
| AB | 0.252 | 2 | 0.126 | 7.688* | 0.001 |
| ความคลาดเคลื่อน (e) | 1.081 | 66 | 0.016 | | |
| ทั้งหมด (Total) | 2.217 | 72 | | | |

test of homogeneity of variances (F = 35.070, *p*= 0.000)

* $p < .05$

เนื่องจากในหัวข้อนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิด คือ Generalized $S-\chi^2$ และ PARSCALE $G^2$ ดังนั้นเมื่อพบว่ามีปฏิสัมพันธ์ระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามและความยาวแบบวัดในโมเดล GRM ผู้วิจัยจะเลือกวิเคราะห์ด้วย Simple effect โดยศึกษาเปรียบเทียบค่าเฉลี่ยของค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนีทั้งสองชนิดสำหรับแต่ละระดับของความยาวแบบวัดโดยใช้การทดสอบ t-test กรณีกลุ่มตัวอย่างเป็นอิสระกัน (independent sample t-test) ในแต่ละระดับของความยาวแบบวัด ซึ่งผลการวิเคราะห์พบว่าในทุกขนาดของความยาวแบบวัด (10, 20 และ 40 ข้อ) ดัชนี Generalized $S-\chi^2$ จะให้ค่าความคลาดเคลื่อนประเภทที่ 1 น้อยกว่าดัชนี PARSCALE $G^2$ อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ผลปรากฏดังในตารางที่ 4.25

**ตารางที่ 4.25 ผลการวิเคราะห์เปรียบเทียบค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ สำหรับแต่ละขนาดของความยาวแบบวัด กรณีโมเดล GRM**

| ความยาวแบบวัด | ดัชนี | N | *M* | *SD* | *F* | *p* | *t* | *p* |
|---|---|---|---|---|---|---|---|---|
| 10 ข้อ | PARSCALE $G^2$ | 12 | 0.303 | 0.308 | 37.224* | 0.000 | 3.205* | 0.008 |
| | Generalized $S-\chi^2$ | 12 | 0.017 | 0.019 | | | | |

**ตารางที่** 4.25 (ต่อ)

| ความยาวแบบวัด | ดัชนี | N | *M* | *SD* | *F* | *p* | *t* | *p* |
|---|---|---|---|---|---|---|---|---|
| 20 ข้อ | PARSCALE $G^2$ | 12 | 0.039 | 0.040 | 14.296* | 0.001 | 3.042* | 0.011 |
| | Generalized $S-\chi^2$ | 12 | 0.003 | 0.005 | | | | |
| 40 ข้อ | PARSCALE $G^2$ | 12 | 0.037 | 0.038 | 13.480* | 0.001 | 2.937* | 0.012 |
| | Generalized $S-\chi^2$ | 12 | 0.004 | 0.009 | | | | |

* $p < .05$

4.2.1.2 การวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างที่ต่างกัน 3 ระดับ กรณีโมเดล GRM

ในโมเดล GRM เมื่อพิจารณาผลการทดสอบปฏิสัมพันธ์ 2 ทาง พบว่า ไม่มีอิทธิพลปฏิสัมพันธ์ (interaction effect) กันระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามและขนาดกลุ่มตัวอย่างที่ระดับนัยสำคัญ.05 ($F$ = 0.956, $p$ = 0.390) ดังรายละเอียดในตารางที่ 4.26 แสดงว่า ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดภายใต้ขนาดกลุ่มตัวอย่างที่ต่างกัน ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) จะมีค่าไม่แตกต่างกันด้วย นั่นคือ ชนิดของดัชนีความสอดคล้องของข้อคำถามและความยาวแบบวัดไม่มีผลร่วมกันต่อค่าความคลาดเคลื่อนประเภทที่ 1

เนื่องจากในโมเดล GRM ไม่มีอิทธิพลปฏิสัมพันธ์ (interaction effect) กันระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามและขนาดกลุ่มตัวอย่าง ผู้วิจัยจึงสามารถสรุปอ้างอิงไปยังอิทธิพลหลักได้ โดยเมื่อพิจารณาที่ขนาดกลุ่มตัวอย่างพบว่า ในโมเดล GRM ขนาดกลุ่มตัวอย่างที่แตกต่างกันไม่มีผลต่อค่าความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ($F$ = 0.750, $p$ = 0.476) แต่เมื่อพิจารณาที่ชนิดของดัชนีความสอดคล้องของข้อคำถามพบว่า ในโมเดล GRM ชนิดของดัชนีความสอดคล้องของข้อคำถามส่งผลต่อค่าความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ($F$ = 10.593, $p$ = 0.002) ซึ่งผลการวิเคราะห์แสดงดังในตารางที่ 4.26

**ตารางที่ 4.26 ผลการวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างที่ต่างกัน 3 ระดับ กรณีโมเดล GRM**

| แหล่งของความแปรปรวน | SS | df | MS | *F* | *p* |
|---|---|---|---|---|---|
| A (ชนิดของดัชนีความสอดคล้องของข้อคำถาม) | 0.251 | 1 | 0.251 | 10.593* | 0.002 |
| B (ขนาดกลุ่มตัวอย่าง) | 0.036 | 2 | 0.018 | 0.750 | 0.476 |
| AB | 0.045 | 2 | 0.023 | 0.956 | 0.390 |
| ความคลาดเคลื่อน (e) | 1.562 | 66 | 0.024 | | |
| ทั้งหมด (Total) | 2.217 | 72 | | | |

test of homogeneity of variances ($F$ = 6.717, $p$= 0.000)

* $p < .05$

เนื่องจากในโมเดล GRM ชนิดของดัชนีความสอดคล้องของข้อคำถามส่งผลต่อค่าความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 จึงต้องมีการเปรียบเทียบดัชนีทั้งสองชนิดที่ใช้ในการศึกษาว่าดัชนีใดให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ที่น้อยกว่าซึ่งดัชนีนั้นจะเป็นดัชนีที่มีประสิทธิภาพในการบ่งชี้ความเหมาะสมของข้อคำถาม โดยใช้การทดสอบ t-test กรณีกลุ่มตัวอย่างเป็นอิสระกัน (independent sample t-test) ซึ่งจากการทดสอบผลปรากฏดังในตารางที่ 4.27 ซึ่งพบว่าในโมเดล GRM ดัชนี Generalized $S-\chi^2$ จะให้ค่าความคลาดเคลื่อนประเภทที่ 1 น้อยกว่าดัชนี PARSCALE $G^2$ อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

**ตารางที่ 4.27 ผลการวิเคราะห์เปรียบเทียบค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ กรณีโมเดล GRM**

| ดัชนี | N | *M* | *SD* | *F* | *p* | *t* | *p* |
|---|---|---|---|---|---|---|---|
| PARSCALE $G^2$ | 36 | 0.126 | 0.216 | 21.651* | 0.000 | 3.268* | 0.002 |
| Generalized $S-\chi^2$ | 36 | 0.008 | 0.014 | | | | |

* $p < .05$

4.2.1.3 การวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขจำนวนรายการคำตอบที่ต่างกัน 4 ระดับ กรณีโมเดล GRM

ในโมเดล GRM เมื่อพิจารณาผลการทดสอบปฏิสัมพันธ์ 2 ทาง พบว่า มีอิทธิพลปฏิสัมพันธ์ (interaction effect) กันระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามและจำนวนรายการคำตอบที่ระดับนัยสำคัญ.05 ($F$ = 3.074, $p$ = 0.034) ดังรายละเอียดในตารางที่ 4.28 แสดงว่า ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดภายใต้จำนวนรายการคำตอบที่ต่างกัน จะส่งผลให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) มีค่าแตกต่างกันด้วย นั่นคือ ชนิดของดัชนีความสอดคล้องของข้อคำถามและจำนวนรายการคำตอบมีผลร่วมกันต่อค่าความคลาดเคลื่อนประเภทที่ 1

**ตารางที่ 4.28 ผลการวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี** PARSCALE $G^2$ **และดัชนี** Generalized $S-\chi^2$ **ภายใต้เงื่อนไขจำนวนรายการคำตอบที่ต่างกัน 4 ระดับ กรณีโมเดล** GRM

| แหล่งของความแปรปรวน | SS | df | MS | $F$ | $p$ |
|---|---|---|---|---|---|
| A (ชนิดของดัชนีความสอดคล้องของข้อคำถาม) | 0.251 | 1 | 0.251 | 12.813* | 0.001 |
| B (จำนวนรายการคำตอบ) | 0.210 | 3 | 0.070 | 3.578* | 0.019 |
| AB | 0.180 | 3 | 0.060 | 3.074* | 0.034 |
| ความคลาดเคลื่อน (e) | 1.252 | 64 | 0.020 | | |
| ทั้งหมด (Total) | 2.217 | 72 | | | |

test of homogeneity of variances ($F$ = 13.558, $p$= 0.000)

* $p < .05$

เนื่องจากในหัวข้อนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิด คือ Generalized $S-\chi^2$ และ PARSCALE $G^2$ ดังนั้นเมื่อพบว่ามีปฏิสัมพันธ์ระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามและจำนวนรายการคำตอบในโมเดล GRM ผู้วิจัยจะเลือกวิเคราะห์ด้วย Simple effect โดยศึกษาเปรียบเทียบค่าเฉลี่ยของค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนีทั้งสองชนิดสำหรับแต่ละระดับของจำนวนรายการคำตอบ โดยใช้การทดสอบ t-test กรณีกลุ่มตัวอย่างเป็นอิสระกัน (independent sample t-test) ในแต่ละระดับของจำนวนรายการคำตอบ ซึ่งผลการวิเคราะห์พบว่าใน 2 ระดับของจำนวนรายการคำตอบ (จำนวนรายการคำตอบ 3 และ 9 รายการ) ดัชนี Generalized $S-\chi^2$ จะให้ค่าความคลาดเคลื่อนประเภทที่ 1 น้อยกว่าดัชนี PARSCALE $G^2$ อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ส่วนใน 2 ระดับของจำนวนรายการคำตอบ (จำนวนรายการคำตอบ 5 และ 7 รายการ) ดัชนี Generalized $S-\chi^2$

จะให้ค่าความคลาดเคลื่อนประเภทที่ 1 ไม่แตกต่างจากดัชนี PARSCALE $G^2$ ที่ระดับ .05 ผลปรากฏดังในตารางที่ 4.29

**ตารางที่ 4.29 ผลการวิเคราะห์เปรียบเทียบค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ สำหรับแต่ละระดับของจำนวนรายการคำตอบ กรณีโมเดล GRM**

| รายการคำตอบ | ดัชนี | N | *M* | *SD* | *F* | *p* | *t* | *p* |
|---|---|---|---|---|---|---|---|---|
| 3 รายการ | PARSCALE $G^2$ | 9 | 0.299 | 0.327 | 31.295* | 0.000 | 2.605* | 0.031 |
| | Generalized $S-\chi^2$ | 9 | 0.014 | 0.019 | | | | |
| 5 รายการ | PARSCALE $G^2$ | 9 | 0.113 | 0.189 | 5.281* | 0.035 | 1.685 | 0.130 |
| | Generalized $S-\chi^2$ | 9 | 0.006 | 0.012 | | | | |
| 7 รายการ | PARSCALE $G^2$ | 9 | 0.059 | 0.108 | 6.393* | 0.022 | 1.449 | 0.184 |
| | Generalized $S-\chi^2$ | 9 | 0.007 | 0.014 | | | | |
| 9 รายการ | PARSCALE $G^2$ | 9 | 0.032 | 0.033 | 4.551* | 0.049 | 2.410* | 0.037 |
| | Generalized $S-\chi^2$ | 9 | 0.004 | 0.010 | | | | |

* $p < .05$

4.2.1.4 การวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขความยาวแบบวัดที่ต่างกัน 3 ระดับ กรณีโมเดล GPCM

ในโมเดล GPCM เมื่อพิจารณาผลการทดสอบปฏิสัมพันธ์ 2 ทาง พบว่า ไม่มีอิทธิพลปฏิสัมพันธ์ (interaction effect) กันระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามและความยาวแบบวัดที่ระดับนัยสำคัญ.05 ($F$= 0.761, $p$= 0.471) ดังรายละเอียดในตารางที่ 4.30 แสดงว่า ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดภายใต้ความยาวแบบวัดที่ต่างกัน ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) จะมีค่าไม่แตกต่างกันด้วย นั่นคือ ชนิดของดัชนีความสอดคล้องของข้อคำถามและความยาวแบบวัดไม่มีผลร่วมกันต่อค่าความคลาดเคลื่อนประเภทที่ 1

เนื่องจากในโมเดล GPCM ไม่มีอิทธิพลปฏิสัมพันธ์ (interaction effect) กันระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามและความยาวแบบวัด ผู้วิจัยจึงสามารถสรุป

อ้างอิงไปยังอิทธิพลหลักได้ โดยเมื่อพิจารณาที่ความยาวแบบวัดพบว่า ในโมเดล GPCM ความยาวแบบวัดที่แตกต่างกันไม่มีผลต่อค่าความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ($F$ = 1.922, $p$ = 0.154) แต่เมื่อพิจารณาที่ชนิดของดัชนีความสอดคล้องของข้อคำถามพบว่า ในโมเดล GPCM ชนิดของดัชนีความสอดคล้องของข้อคำถามส่งผลต่อค่าความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ($F$ = 23.501, $p$ = 0.000) ซึ่งผลการวิเคราะห์แสดงดังในตารางที่ 4.30

**ตารางที่** 4.30 **ผลการวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี** PARSCALE $G^2$ **และดัชนี** Generalized $\boldsymbol{S-\chi^2}$ **ภายใต้เงื่อนไขความยาวแบบวัดที่ต่างกัน 3 ระดับ กรณีโมเดล** GPCM

| แหล่งของความแปรปรวน | SS | df | MS | *F* | *p* |
|---|---|---|---|---|---|
| A (ชนิดของดัชนีความสอดคล้องของข้อคำถาม) | 0.382 | 1 | 0.382 | 23.501* | 0.000 |
| B (ความยาวแบบวัด) | 0.062 | 2 | 0.031 | 1.922 | 0.154 |
| AB | 0.025 | 2 | 0.012 | 0.761 | 0.471 |
| ความคลาดเคลื่อน (e) | 1.073 | 66 | 0.016 | | |
| ทั้งหมด (Total) | 2.382 | 72 | | | |

test of homogeneity of variances ($F$ = 9.155, $p$= 0.000 )

* $p < .05$

เนื่องจากในโมเดล GPCM ชนิดของดัชนีความสอดคล้องของข้อคำถามส่งผลต่อค่าความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 จึงต้องมีการเปรียบเทียบดัชนีทั้งสองชนิดที่ใช้ในการศึกษาว่าดัชนีใดให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ที่น้อยกว่าซึ่งดัชนีนั้นจะเป็นดัชนีที่มีประสิทธิภาพในการบ่งชี้ความเหมาะสมของข้อคำถาม โดยใช้การทดสอบ t-test กรณีกลุ่มตัวอย่างเป็นอิสระกัน (independent sample t-test) ซึ่งจากการทดสอบผลปรากฏดังในตารางที่ 4.31 ซึ่งพบว่าในโมเดล GPCM ดัชนี Generalized $S-\chi^2$ จะให้ค่าความคลาดเคลื่อนประเภทที่ 1 น้อยกว่าดัชนี PARSCALE $G^2$ อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

**ตารางที่ 4.31 ผลการวิเคราะห์เปรียบเทียบค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี** PARSCALE $G^2$ **และดัชนี** Generalized $S-\chi^2$ **กรณีโมเดล** GPCM

| ดัชนี | N | *M* | *SD* | *F* | *p* | *t* | *p* |
|---|---|---|---|---|---|---|---|
| PARSCALE $G^2$ | 36 | 0.181 | 0.176 | 32.859* | 0.000 | 4.801* | 0.000 |
| Generalized $S-\chi^2$ | 36 | 0.035 | 0.048 | | | | |

* $p < .05$

4.2.1.5 การวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างที่ต่างกัน 3 ระดับ กรณีโมเดล GPCM

ในโมเดล GPCM เมื่อพิจารณาผลการทดสอบปฏิสัมพันธ์ 2 ทาง พบว่า ไม่มีอิทธิพลปฏิสัมพันธ์ (interaction effect) กันระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามและขนาดกลุ่มตัวอย่างที่ระดับนัยสำคัญ.05 ($F = 2.872$, $p = 0.064$) ดังรายละเอียดในตารางที่ 4.32 แสดงว่า ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดภายใต้ความยาวแบบวัดที่ต่างกัน ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) จะมีค่าไม่แตกต่างกันด้วย นั่นคือ ชนิดของดัชนีความสอดคล้องของข้อคำถามและความยาวแบบวัดไม่มีผลร่วมกันต่อค่าความคลาดเคลื่อนประเภทที่ 1

เนื่องจากในโมเดล GPCM ไม่มีอิทธิพลปฏิสัมพันธ์ (interaction effect) กันระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามและขนาดกลุ่มตัวอย่าง ผู้วิจัยจึงสามารถสรุปอ้างอิงไปยังอิทธิพลหลักได้ โดยเมื่อพิจารณาที่ขนาดกลุ่มตัวอย่างพบว่า ในโมเดล GPCM ขนาดกลุ่มตัวอย่างที่แตกต่างกันไม่มีผลต่อค่าความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ($F = 2.095$, $p = 0.131$) แต่เมื่อพิจารณาที่ชนิดของดัชนีความสอดคล้องของข้อคำถามพบว่า ในโมเดล GPCM ชนิดของดัชนีความสอดคล้องของข้อคำถามส่งผลต่อค่าความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ($F = 25.005$, $p = 0.000$) ซึ่งผลการวิเคราะห์แสดงดังในตารางที่ 4.32

**ตารางที่ 4.32 ผลการวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างที่ต่างกัน 3 ระดับ กรณีโมเดล GPCM**

| แหล่งของความแปรปรวน | SS | df | MS | *F* | *p* |
|---|---|---|---|---|---|
| A (ชนิดของดัชนีความสอดคล้องของข้อคำถาม) | 0.382 | 1 | 0.382 | 25.005* | 0.000 |
| B (ขนาดกลุ่มตัวอย่าง) | 0.064 | 2 | 0.032 | 2.095 | 0.131 |
| AB | 0.088 | 2 | 0.044 | 2.872 | 0.064 |
| ความคลาดเคลื่อน (e) | 1.008 | 66 | 0.015 | | |
| ทั้งหมด (Total) | 2.382 | 72 | | | |

test of homogeneity of variances ($F$ = 3.315, $p$= 0.010)

* $p < .05$

เนื่องจากในโมเดล GPCM ชนิดของดัชนีความสอดคล้องของข้อคำถามส่งผลต่อค่าความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 จึงต้องมีการเปรียบเทียบดัชนีทั้งสองชนิดที่ใช้ในการศึกษาว่าดัชนีใดให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ที่น้อยกว่าซึ่งดัชนีนั้นจะเป็นดัชนีที่มีประสิทธิภาพในการบ่งชี้ความเหมาะสมของข้อคำถาม โดยใช้การทดสอบ t-test กรณีกลุ่มตัวอย่างเป็นอิสระกัน (independent sample t-test) ซึ่งจากการทดสอบปรากฏผลในตารางที่ 4.31 ที่ผ่านมา จะพบว่าในโมเดล GPCM ดัชนี Generalized $S-\chi^2$ จะให้ค่าความคลาดเคลื่อนประเภทที่ 1 น้อยกว่าดัชนี PARSCALE $G^2$ อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

4.2.1.6 การวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขจำนวนรายการคำตอบที่ต่างกัน 4 ระดับ กรณีโมเดล GPCM

ในโมเดล GPCM เมื่อพิจารณาผลการทดสอบปฏิสัมพันธ์ 2 ทาง พบว่า ไม่มีอิทธิพลปฏิสัมพันธ์ (interaction effect) กันระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามและจำนวนรายการคำตอบที่ระดับนัยสำคัญ.05 ($F$ = 0.250, $p$ = 0.861) ดังรายละเอียดในตารางที่ 4.33 แสดงว่า ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสอง

ชนิดภายใต้จำนวนรายการคำตอบที่ต่างกัน ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) จะมีค่าไม่แตกต่างกัน นั่นคือ ชนิดของดัชนีความสอดคล้องของข้อคำถามและจำนวนรายการคำตอบไม่มีผลร่วมกันต่อค่าความคลาดเคลื่อนประเภทที่ 1

เนื่องจากในโมเดล GPCM ไม่มีอิทธิพลปฏิสัมพันธ์ (interaction effect) กันระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามและขนาดกลุ่มตัวอย่าง ผู้วิจัยจึงสามารถสรุปอ้างอิงไปยังอิทธิพลหลักได้ โดยเมื่อพิจารณาที่ขนาดกลุ่มตัวอย่างพบว่า ในโมเดล GPCM ขนาดกลุ่มตัวอย่างที่แตกต่างกันมีผลต่อค่าความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ($F$ = 2.983, $p$ = 0.038) และเมื่อพิจารณาที่ชนิดของดัชนีความสอดคล้องของข้อคำถามพบว่า ในโมเดล GPCM ชนิดของดัชนีความสอดคล้องของข้อคำถามส่งผลต่อค่าความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ($F$ = 24.269, $p$ = 0.000) ซึ่งผลการวิเคราะห์แสดงดังในตารางที่ 4.33

**ตารางที่ 4.33 ผลการวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี** PARSCALE $G^2$ **และดัชนี** Generalized $\boldsymbol{S-\chi^2}$ **ภายใต้เงื่อนไขจำนวนรายการคำตอบที่ต่างกัน 4 ระดับ กรณีโมเดล** GPCM

| แหล่งของความแปรปรวน | SS | df | MS | $F$ | $p$ |
|---|---|---|---|---|---|
| A (ชนิดของดัชนีความสอดคล้องของข้อคำถาม) | 0.382 | 1 | 0.382 | 24.269* | 0.000 |
| B (จำนวนรายการคำตอบ) | 0.141 | 3 | 0.047 | 2.983* | 0.038 |
| AB | 0.012 | 3 | 0.004 | 0.250 | 0.861 |
| ความคลาดเคลื่อน (e) | 1.007 | 64 | 0.016 | | |
| ทั้งหมด (Total) | 2.382 | 72 | | | |

test of homogeneity of variances ($F$ = 8.064, $p$= 0.000)

* $p < .05$

เนื่องจากในโมเดล GPCM ชนิดของดัชนีความสอดคล้องของข้อคำถามส่งผลต่อค่าความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 จึงต้องมีการเปรียบเทียบดัชนีทั้งสองชนิดที่ใช้ในการศึกษาว่าดัชนีใดให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ที่น้อยกว่าซึ่งดัชนีนั้นจะเป็นดัชนีที่มีประสิทธิภาพในการบ่งชี้ความเหมาะสมของข้อคำถาม โดยใช้การทดสอบ t-test กรณีกลุ่มตัวอย่างเป็นอิสระกัน (independent sample t-test) ซึ่งจากการ

ทดสอบปรากฏผลในตารางที่ 4.31 ที่ผ่านมา จะพบว่าในโมเดล GPCM ดัชนี Generalized $S-\chi^2$ จะให้ค่าความคลาดเคลื่อนประเภทที่ 1 น้อยกว่าดัชนี PARSCALE $G^2$ อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

เนื่องจากในโมเดล GPCM จำนวนรายการคำตอบส่งผลต่อค่าความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 จึงต้องมีการเปรียบเทียบรายคู่ ว่าจำนวนรายการคำตอบคู่ใดบ้างที่ให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) แตกต่างกันโดยใช้การเปรียบเทียบพหุคูณ (multiple comparison) แต่เนื่องจากในหัวข้อนี้มีวัตถุประสงค์เพื่อเปรียบเทียบค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิด คือ Generalized $S-\chi^2$ และ PARSCALE $G^2$ ดังนั้นเมื่อพบว่าไม่มีปฏิสัมพันธ์ระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามและจำนวนรายการคำตอบในโมเดล GPCM ผู้วิจัยจะเลือกวิเคราะห์เฉพาะการเปรียบเทียบค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ของดัชนีทั้งสองชนิดที่ใช้ในการศึกษาเท่านั้น โดยไม่พิจารณาว่าในโมเดล GPCM จำนวนรายการคำตอบที่แตกต่างกันส่งผลอย่างไรต่อค่าความคลาดเคลื่อนประเภทที่ 1 อีก เนื่องจากได้ศึกษาวิเคราะห์ไปแล้วใน ตอนที่ 2 ผลการวิเคราะห์ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error)

4.2.2 การวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบ (Power of the test)

4.2.2.1 การวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขความยาวแบบวัดที่ต่างกัน 3 ระดับ กรณีโมเดล GRM

ในโมเดล GRM เมื่อพิจารณาผลการทดสอบปฏิสัมพันธ์ 2 ทาง พบว่า ไม่มีอิทธิพลปฏิสัมพันธ์ (interaction effect) กันระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามและความยาวแบบวัดที่ระดับนัยสำคัญ.05 ($F$ = 2.114, $p$ = 0.129) ดังรายละเอียดในตารางที่ 4.34 แสดงว่า ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดภายใต้ความยาวแบบวัดที่ต่างกัน อำนาจการทดสอบจะมีค่าไม่แตกต่างกัน นั่นคือ ชนิดของดัชนีความสอดคล้องของข้อคำถามและความยาวแบบวัดไม่มีผลร่วมกันต่ออำนาจการทดสอบ

เนื่องจากในโมเดล GRM ไม่มีอิทธิพลปฏิสัมพันธ์ (interaction effect) กันระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามและความยาวแบบวัด ผู้วิจัยจึงสามารถสรุปอ้างอิงไปยังอิทธิพลหลักได้ โดยเมื่อพิจารณาที่ความยาวแบบวัดพบว่า ในโมเดล GRM ความยาว

แบบวัดที่แตกต่างกันมีผลต่ออำนาจการทดสอบอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ($F$ = 5.879, $p$ = 0.004) และเมื่อพิจารณาที่ชนิดของดัชนีความสอดคล้องของข้อคำถามพบว่า ในโมเดล GRM ชนิดของดัชนีความสอดคล้องของข้อคำถามส่งผลต่ออำนาจการทดสอบอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ($F$ = 58.053, $p$ = 0.000) ซึ่งผลการวิเคราะห์แสดงดังในตารางที่ 4.34

**ตารางที่** 4.34 **ผลการวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี** PARSCALE $G^2$ **และดัชนี** Generalized $S-\chi^2$ **ภายใต้เงื่อนไขความยาวแบบวัดที่ต่างกัน 3 ระดับ กรณีโมเดล** GRM

| แหล่งของความแปรปรวน | SS | df | MS | *F* | *p* |
|---|---|---|---|---|---|
| A (ชนิดของดัชนีความสอดคล้องของข้อคำถาม) | 2.146 | 1 | 2.146 | 58.053* | 0.000 |
| B (ความยาวแบบวัด) | 0.435 | 2 | 0.217 | 5.879* | 0.004 |
| AB | 0.158 | 2 | 0.078 | 2.114 | 0.129 |
| ความคลาดเคลื่อน (e) | 2.440 | 66 | 0.037 | | |
| ทั้งหมด (Total) | 10.128 | 72 | | | |

test of homogeneity of variances ($F$ = 10.504, $p$ = 0.000)

* $p < .05$

เนื่องจากในโมเดล GRM ชนิดของดัชนีความสอดคล้องของข้อคำถามส่งผลต่ออำนาจการทดสอบอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 จึงต้องมีการเปรียบเทียบ ดัชนีทั้งสองชนิดที่ใช้ในการศึกษาว่าดัชนีใดให้อำนาจการทดสอบที่มากกว่า ซึ่งดัชนีนั้นจะเป็นดัชนีที่มีประสิทธิภาพในการบ่งชี้ความเหมาะสมของข้อคำถาม โดยใช้การทดสอบ t-test กรณีกลุ่มตัวอย่างเป็นอิสระกัน (independent sample t-test) ซึ่งจากการทดสอบผลปรากฏดังในตารางที่ 4.35 ซึ่งพบว่าในโมเดล GRM ดัชนี Generalized $S-\chi^2$ จะให้อำนาจการทดสอบน้อยกว่าดัชนี PARSCALE $G^2$ อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

**ตารางที่** 4.35 **ผลการวิเคราะห์เปรียบเทียบอำนาจการทดสอบของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี** PARSCALE $G^2$ **และดัชนี** Generalized $S-\chi^2$ **กรณีโมเดล** GRM

| ดัชนี | N | *M* | *SD* | *F* | *p* | *t* | *p* |
|---|---|---|---|---|---|---|---|
| PARSCALE $G^2$ | 36 | 0.434 | 0.278 | 36.684* | 0.000 | 7.040* | 0.000 |
| Generalized $S-\chi^2$ | 36 | 0.089 | 0.097 | | | | |

* $p < .05$

เนื่องจากในโมเดล GRM ความยาวแบบวัดส่งผลต่ออำนาจการทดสอบอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 จึงต้องมีการเปรียบเทียบรายคู่ ว่าความยาวแบบวัดคู่ใดบ้างที่ให้อำนาจการทดสอบแตกต่างกันโดยใช้การเปรียบเทียบพหุคูณ (multiple comparison) แต่เนื่องจากในหัวข้อนี้มีวัตถุประสงค์เพื่อเปรียบเทียบอำนาจการทดสอบของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิด คือ Generalized $S-\chi^2$ และ PARSCALE $G^2$ ดังนั้นเมื่อพบว่าไม่มีปฏิสัมพันธ์ระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามและความยาวแบบวัดในโมเดล GPCM ผู้วิจัยจะเลือกวิเคราะห์เฉพาะการเปรียบเทียบอำนาจการทดสอบ ของดัชนีทั้งสองชนิดที่ใช้ในการศึกษาเท่านั้น โดยไม่พิจารณาว่าในโมเดล GRM ความยาวแบบวัดที่แตกต่างกันส่งผลอย่างไรต่ออำนาจการทดสอบอีก เนื่องจากได้ศึกษาวิเคราะห์ไปแล้วใน ตอนที่ 3 ผลการวิเคราะห์อำนาจการทดสอบ (Power of the test)

4.2.2.2 การวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างที่ต่างกัน 3 ระดับ กรณีโมเดล GRM

ในโมเดล GRM เมื่อพิจารณาผลการทดสอบปฏิสัมพันธ์ 2 ทาง พบว่า มีอิทธิพลปฏิสัมพันธ์ (interaction effect) กันระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามและขนาดกลุ่มตัวอย่างที่ระดับนัยสำคัญ.05 ($F$ = 6.185, $p$ = 0.003) ดังรายละเอียดในตารางที่ 4.36 แสดงว่า ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดภายใต้ขนาดกลุ่มตัวอย่างที่ต่างกัน จะส่งผลให้อำนาจการทดสอบมีค่าแตกต่างกันด้วย นั่นคือ ชนิดของดัชนีความสอดคล้องของข้อคำถามและขนาดกลุ่มตัวอย่างมีผลร่วมกันต่ออำนาจการทดสอบ

**ตารางที่ 4.36 ผลการวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างที่ต่างกัน 3 ระดับ กรณีโมเดล GRM**

| แหล่งของความแปรปรวน | SS | df | MS | *F* | *p* |
|---|---|---|---|---|---|
| A (ชนิดของดัชนีความสอดคล้องของข้อคำถาม) | 2.146 | 1 | 2.146 | 84.266* | 0.000 |
| B (ขนาดกลุ่มตัวอย่าง) | 1.035 | 2 | 0.517 | 20.319* | 0.000 |
| AB | 0.315 | 2 | 0.158 | 6.185* | 0.003 |
| ความคลาดเคลื่อน (e) | 1.681 | 66 | 0.025 | | |
| ทั้งหมด (Total) | 10.128 | 72 | | | |

test of homogeneity of variances ($F$ = 4.822, $p$= 0.001)

* $p < .05$

เนื่องจากในหัวข้อนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิด คือ Generalized $S-\chi^2$ และ PARSCALE $G^2$ ดังนั้นเมื่อพบว่ามีปฏิสัมพันธ์ระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามและขนาดกลุ่มตัวอย่างในโมเดล GRM ผู้วิจัยจะเลือกวิเคราะห์ด้วย Simple effect โดยศึกษาเปรียบเทียบอำนาจการทดสอบของดัชนีทั้งสองชนิดสำหรับแต่ละระดับของขนาดกลุ่มตัวอย่างโดยใช้การทดสอบ t-test กรณีกลุ่มตัวอย่างเป็นอิสระกัน (independent sample t-test) ในแต่ละระดับของขนาดกลุ่มตัวอย่าง ซึ่งผลการวิเคราะห์พบว่าในทุกระดับของขนาดกลุ่มตัวอย่าง (500, 1000 และ 2000 คน) ดัชนี Generalized $S-\chi^2$ จะให้อำนาจการทดสอบน้อยกว่าดัชนี PARSCALE $G^2$ อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ผลปรากฏดังในตารางที่ 4.37

**ตารางที่ 4.37 ผลการวิเคราะห์เปรียบเทียบอำนาจการทดสอบของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ สำหรับแต่ละระดับของขนาดกลุ่มตัวอย่าง กรณีโมเดล GRM**

| ขนาดกลุ่มตัวอย่าง | ดัชนี | N | *M* | *SD* | *F* | *p* | *t* | *p* |
|---|---|---|---|---|---|---|---|---|
| 500 คน | PARSCALE $G^2$ | 12 | 0.218 | 0.152 | 6.274* | 0.020 | 4.045* | 0.002 |
| | Generalized $S-\chi^2$ | 12 | 0.036 | 0.033 | | | | |
| 1000 คน | PARSCALE $G^2$ | 12 | 0.414 | 0.213 | 5.877* | 0.024 | 5.487* | 0.000 |
| | Generalized $S-\chi^2$ | 12 | 0.066 | 0.051 | | | | |

**ตารางที่** 4.37 (ต่อ)

| ขนาดกลุ่มตัวอย่าง | ดัชนี | N | *M* | *SD* | *F* | *p* | *t* | *p* |
|---|---|---|---|---|---|---|---|---|
| 2000 คน | PARSCALE $G^2$ | 12 | 0.671 | 0.253 | 3.503 | 0.075 | 6.186* | 0.000 |
| | Generalized $S-\chi^2$ | 12 | 0.165 | 0.127 | | | | |

* $p < .05$

4.2.2.3 การวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขจำนวนรายการคำตอบที่ต่างกัน 4 ระดับ กรณีโมเดล GRM

ในโมเดล GRM เมื่อพิจารณาผลการทดสอบปฏิสัมพันธ์ 2 ทาง พบว่า ไม่มีอิทธิพลปฏิสัมพันธ์ (interaction effect) กันระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามและจำนวนรายการคำตอบที่ระดับนัยสำคัญ.05 ($F$ = 0.121, $p$ = 0.948) ดังรายละเอียดในตารางที่ 4.38 แสดงว่า ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดภายใต้จำนวนรายการคำตอบที่ต่างกัน อำนาจการทดสอบจะมีค่าไม่แตกต่างกัน นั่นคือ ชนิดของดัชนีความสอดคล้องของข้อคำถามและจำนวนรายการคำตอบไม่มีผลร่วมกันต่ออำนาจการทดสอบ

เนื่องจากในโมเดล GRM ไม่มีอิทธิพลปฏิสัมพันธ์ (interaction effect) กันระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามและจำนวนรายการคำตอบ ผู้วิจัยจึงสามารถสรุปอ้างอิงไปยังอิทธิพลหลักได้ โดยเมื่อพิจารณาที่จำนวนรายการคำตอบพบว่า ในโมเดล GRM จำนวนรายการคำตอบที่แตกต่างกันไม่มีผลต่ออำนาจการทดสอบอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ($F$ = 1.638, $p$ = 0.189) แต่เมื่อพิจารณาที่ชนิดของดัชนีความสอดคล้องของข้อคำถาม พบว่า ในโมเดล GRM ชนิดของดัชนีความสอดคล้องของข้อคำถามส่งผลต่ออำนาจการทดสอบอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ($F$ = 49.053, $p$ = 0.000) ซึ่งผลการวิเคราะห์แสดงดังในตารางที่ 4.38

**ตารางที่ 4.38 ผลการวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขจำนวนรายการคำตอบที่ต่างกัน 4 ระดับ กรณีโมเดล GRM**

| แหล่งของความแปรปรวน | SS | df | MS | *F* | *p* |
|---|---|---|---|---|---|
| A (ชนิดของดัชนีความสอดคล้องของข้อคำถาม) | 2.146 | 1 | 2.146 | 49.053* | 0.000 |
| B (จำนวนรายการคำตอบ) | 0.215 | 3 | 0.072 | 1.638 | 0.189 |
| AB | 0.016 | 3 | 0.005 | 0.121 | 0.948 |
| ความคลาดเคลื่อน (e) | 2.800 | 64 | 0.044 | | |
| ทั้งหมด (Total) | 10.128 | 72 | | | |

test of homogeneity of variances ($F$ = 6.888, $p$= 0.000)

* $p < .05$

เนื่องจากในโมเดล GRM ชนิดของดัชนีความสอดคล้องของข้อคำถามส่งผลต่ออำนาจการทดสอบอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 จึงต้องมีการเปรียบเทียบ ดัชนีทั้งสองชนิดที่ใช้ในการศึกษาว่าดัชนีใดให้อำนาจการทดสอบที่มากกว่า ซึ่งดัชนีนั้นจะเป็นดัชนีที่มีประสิทธิภาพในการบ่งชี้ความเหมาะสมของข้อคำถาม โดยใช้การทดสอบ t-test กรณีกลุ่มตัวอย่างเป็นอิสระกัน (independent sample t-test) ซึ่งจากการทดสอบปรากฏผลในตารางที่ 4.35 ที่ผ่านมา จะพบว่าในโมเดล GRM ดัชนี Generalized $S-\chi^2$ จะให้อำนาจการทดสอบน้อยกว่าดัชนี PARSCALE $G^2$ อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

4.2.2.4 การวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขความยาวแบบวัดที่ต่างกัน 3 ระดับ กรณีโมเดล GPCM

ในโมเดล GPCM เมื่อพิจารณาผลการทดสอบปฏิสัมพันธ์ 2 ทาง พบว่า มีอิทธิพลปฏิสัมพันธ์ (interaction effect) กันระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามและความยาวแบบวัดที่ระดับนัยสำคัญ.05 ($F$ = 7.187, $p$ = 0.002) ดังรายละเอียดในตารางที่ 4.39 แสดงว่า ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดภายใต้ความยาวแบบวัดที่ต่างกัน จะส่งผลให้อำนาจการทดสอบมีค่าแตกต่างกันด้วย นั่นคือ ชนิดของดัชนีความสอดคล้องของข้อคำถามและความยาวแบบวัดมีผลร่วมกันต่ออำนาจการทดสอบ

**ตารางที่ 4.39 ผลการวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขความยาวแบบวัดที่ต่างกัน 3 ระดับ กรณีโมเดล GPCM**

| แหล่งของความแปรปรวน | SS | df | MS | *F* | *p* |
|---|---|---|---|---|---|
| A (ชนิดของดัชนีความสอดคล้องของข้อคำถาม) | 2.396 | 1 | 2.396 | 49.593* | 0.000 |
| B (ความยาวแบบวัด) | 0.252 | 2 | 0.126 | 2.606 | 0.081 |
| AB | 0.694 | 2 | 0.347 | 7.187* | 0.002 |
| ความคลาดเคลื่อน (e) | 3.189 | 66 | 0.048 | | |
| ทั้งหมด (Total) | 13.254 | 72 | | | |

test of homogeneity of variances ($F$ = 5.594, $p$= 0.000 )

* $p < .05$

เนื่องจากในหัวข้อนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิด คือ Generalized $S-\chi^2$ และ PARSCALE $G^2$ ดังนั้นเมื่อพบว่ามีปฏิสัมพันธ์ระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามและความยาวแบบวัดในโมเดล GPCM ผู้วิจัยจะเลือกวิเคราะห์ด้วย Simple effect โดยศึกษาเปรียบเทียบอำนาจการทดสอบของดัชนีทั้งสองชนิดสำหรับแต่ละระดับของความยาวแบบวัดโดยใช้การทดสอบ t-test กรณีกลุ่มตัวอย่างเป็นอิสระกัน (independent sample t-test) ในแต่ละระดับของความยาวแบบวัด ซึ่งผลการวิเคราะห์พบว่าในทุกระดับของความยาวแบบวัด (10, 20 และ 40 ข้อ) ดัชนี Generalized $S-\chi^2$ จะให้อำนาจการทดสอบน้อยกว่าดัชนี PARSCALE $G^2$ อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ผลปรากฏดังในตารางที่ 4.40

**ตารางที่ 4.40 ผลการวิเคราะห์เปรียบเทียบอำนาจกาทดสอบของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ สำหรับแต่ละระดับของความยาวแบบวัด กรณีโมเดล GPCM**

| ความยาวแบบวัด | ดัชนี | N | *M* | *SD* | *F* | *p* | *t* | *p* |
|---|---|---|---|---|---|---|---|---|
| 10 ข้อ | PARSCALE $G^2$ | 12 | 0.332 | 0.240 | 11.398* | 0.003 | 2.229* | 0.040 |
| | Generalized $S-\chi^2$ | 12 | 0.156 | 0.129 | | | | |
| 20 ข้อ | PARSCALE $G^2$ | 12 | 0.428 | 0.303 | 3.344 | 0.081 | 2.659* | 0.015 |
| | Generalized $S-\chi^2$ | 12 | 0.145 | 0.211 | | | | |
| 40 ข้อ | PARSCALE $G^2$ | 12 | 0.703 | 0.274 | 17.752* | 0.000 | 7.824* | 0.000 |
| | Generalized $S-\chi^2$ | 12 | 0.068 | 0.063 | | | | |

* $p < .05$

4.2.2.5 การวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างที่ต่างกัน 3 ระดับ กรณีโมเดล GPCM

ในโมเดล GPCM เมื่อพิจารณาผลการทดสอบปฏิสัมพันธ์ 2 ทาง พบว่า ไม่มีอิทธิพลปฏิสัมพันธ์ (interaction effect) กันระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามและขนาดกลุ่มตัวอย่างที่ระดับนัยสำคัญ.05 ($F$ = 3.129, $p$ = 0.050) ดังรายละเอียดในตารางที่ 4.41 แสดงว่า ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดภายใต้ขนาดกลุ่มตัวอย่างที่ต่างกัน อำนาจการทดสอบจะมีค่าไม่แตกต่างกัน นั่นคือ ชนิดของดัชนีความสอดคล้องของข้อคำถามและขนาดกลุ่มตัวอย่างไม่มีผลร่วมกันต่ออำนาจการทดสอบ

เนื่องจากในโมเดล GPCM ไม่มีอิทธิพลปฏิสัมพันธ์ (interaction effect) กันระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามและขนาดกลุ่มตัวอย่าง ผู้วิจัยจึงสามารถสรุปอ้างอิงไปยังอิทธิพลหลักได้ โดยเมื่อพิจารณาที่ขนาดกลุ่มตัวอย่างพบว่า ในโมเดล GRM ขนาดกลุ่มตัวอย่างที่แตกต่างกันมีผลต่ออำนาจการทดสอบอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ($F$= 7.648, $p$ = 0.001) และเมื่อพิจารณาที่ชนิดของดัชนีความสอดคล้องของข้อคำถามพบว่า ในโมเดล GPCM ชนิดของดัชนีความสอดคล้องของข้อคำถามส่งผลต่ออำนาจการทดสอบอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ($F$ = 50.733, $p$ = 0.000) ซึ่งผลการวิเคราะห์แสดงดังในตารางที่ 4.41

**ตารางที่ 4.41 ผลการวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าอำนาจการทดสอบในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี** PARSCALE $G^2$ **และดัชนี** Generalized $S-\chi^2$ **ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างที่ต่างกัน 3 ระดับ กรณีโมเดล** GPCM

| แหล่งของความแปรปรวน | SS | df | MS | $F$ | $p$ |
|---|---|---|---|---|---|
| A (ชนิดของดัชนีความสอดคล้องของข้อคำถาม) | 2.396 | 1 | 2.396 | 50.733* | 0.000 |
| B (ขนาดกลุ่มตัวอย่าง) | 0.722 | 2 | 0.361 | 7.648* | 0.001 |
| AB | 0.296 | 2 | 0.148 | 3.129 | 0.050 |
| ความคลาดเคลื่อน (e) | 3.117 | 66 | 0.047 | | |
| ทั้งหมด (Total) | 13.254 | 72 | | | |

test of homogeneity of variances ($F$ = 4.453, $p$= 0.001)

* $p < .05$

เนื่องจากในโมเดล GPCM ชนิดของดัชนีความสอดคล้องของข้อคำถามส่งผลต่ออำนาจการทดสอบอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 จึงต้องมีการเปรียบเทียบ ดัชนีทั้งสองชนิดที่ใช้ในการศึกษาว่าดัชนีใดให้อำนาจการทดสอบที่มากกว่า ซึ่งดัชนีนั้นจะเป็นดัชนีที่มีประสิทธิภาพในการบ่งชี้ความเหมาะสมของข้อคำถาม โดยใช้การทดสอบ t-test กรณีกลุ่มตัวอย่างเป็นอิสระกัน (independent sample t-test) ซึ่งจากการทดสอบผลปรากฏดังในตารางที่ 4.42 ซึ่งพบว่าในโมเดล GPCM ดัชนี Generalized $S-\chi^2$ จะให้อำนาจการทดสอบน้อยกว่าดัชนี PARSCALE $G^2$ อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

**ตารางที่ 4.42 ผลการวิเคราะห์เปรียบเทียบอำนาจการทดสอบของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ กรณีโมเดล GPCM**

| ดัชนี | N | *M* | *SD* | *F* | *p* | *t* | *p* |
|---|---|---|---|---|---|---|---|
| PARSCALE $G^2$ | 36 | 0.488 | 0.309 | 28.771* | 0.000 | 6.369* | 0.000 |
| Generalized $S-\chi^2$ | 36 | 0.123 | 0.149 | | | | |

* $p < .05$

เนื่องจากในโมเดล GPCM ขนาดกลุ่มตัวอย่างส่งผลต่ออำนาจการทดสอบอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 จึงต้องมีการเปรียบเทียบรายคู่ ว่าขนาดกลุ่มตัวอย่างคู่ใดบ้างที่ให้อำนาจการทดสอบแตกต่างกันโดยใช้การเปรียบเทียบพหุคูณ (multiple comparison) แต่เนื่องจากในหัวข้อนี้มีวัตถุประสงค์เพื่อเปรียบเทียบอำนาจการทดสอบของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิด คือ Generalized $S-\chi^2$ และ PARSCALE $G^2$ ดังนั้นเมื่อพบว่าไม่มีปฏิสัมพันธ์ระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามและขนาดกลุ่มตัวอย่างในโมเดล GPCM ผู้วิจัยจะเลือกวิเคราะห์เฉพาะการเปรียบเทียบอำนาจการทดสอบ ของดัชนีทั้งสองชนิดที่ใช้ในการศึกษาเท่านั้น โดยไม่พิจารณาว่าในโมเดล GPCM ขนาดกลุ่มตัวอย่างที่แตกต่างกันส่งผลอย่างไรต่ออำนาจการทดสอบอีก เนื่องจากได้ศึกษาวิเคราะห์ไปแล้วใน ตอนที่ 3 ผลการวิเคราะห์อำนาจการทดสอบ (Power of the test)

4.2.2.6 การวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขจำนวนรายการคำตอบที่ต่างกัน 4 ระดับ กรณีโมเดล GPCM

ในโมเดล GPCM เมื่อพิจารณาผลการทดสอบปฏิสัมพันธ์ 2 ทาง พบว่า ไม่มีอิทธิพลปฏิสัมพันธ์ (interaction effect) กันระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามและจำนวนรายการคำตอบที่ระดับนัยสำคัญ.05 ($F$ = 0.260, $p$ = 0.854) ดังรายละเอียดในตารางที่ 4.43 แสดงว่า ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดภายใต้จำนวนรายการคำตอบที่ต่างกัน อำนาจการทดสอบจะมีค่าไม่แตกต่างกัน นั่นคือ ชนิดของดัชนีความสอดคล้องของข้อคำถามและจำนวนรายการคำตอบไม่มีผลร่วมกันต่ออำนาจการทดสอบ

เนื่องจากในโมเดล GPCM ไม่มีอิทธิพลปฏิสัมพันธ์ (interaction effect) กันระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามและจำนวนรายการคำตอบ ผู้วิจัยจึงสามารถสรุปอ้างอิงไปยังอิทธิพลหลักได้ โดยเมื่อพิจารณาที่จำนวนรายการคำตอบพบว่า ในโมเดล GRM จำนวนรายการคำตอบที่แตกต่างกันมีผลต่ออำนาจการทดสอบอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ($F$ = 5.293, $p$ = 0.003) และเมื่อพิจารณาที่ชนิดของดัชนีความสอดคล้องของข้อคำถาม พบว่า ในโมเดล GPCM ชนิดของดัชนีความสอดคล้องของข้อคำถามส่งผลต่ออำนาจการทดสอบอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ($F$ = 46.738, $p$ = 0.000) ซึ่งผลการวิเคราะห์แสดงดังในตารางที่ 4.43

**ตารางที่ 4.43 ผลการวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี** PARSCALE $G^2$ **และดัชนี** Generalized $S-\chi^2$ **ภายใต้เงื่อนไขจำนวนรายการคำตอบที่ต่างกัน 4 ระดับ กรณีโมเดล** GPCM

| แหล่งของความแปรปรวน | SS | df | MS | $F$ | $p$ |
|---|---|---|---|---|---|
| A (ชนิดของดัชนีความสอดคล้องของข้อคำถาม) | 2.396 | 1 | 2.396 | 46.738* | 0.000 |
| B (จำนวนรายการคำตอบ) | 0.816 | 3 | 0.271 | 5.293* | 0.003 |
| AB | 0.040 | 3 | 0.013 | 0.260 | 0.854 |
| ความคลาดเคลื่อน (e) | 3.281 | 64 | 0.051 | | |
| ทั้งหมด (Total) | 13.254 | 72 | | | |

test of homogeneity of variances ($F$ = 7.830, $p$= 0.000)

* $p < .05$

เนื่องจากในโมเดล GPCM ชนิดของดัชนีความสอดคล้องของข้อคำถามส่งผลต่ออำนาจการทดสอบอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 จึงต้องมีการเปรียบเทียบ ดัชนีทั้งสอง

ชนิดที่ใช้ในการศึกษาว่าดัชนีใดให้อำนาจการทดสอบที่มากกว่า ซึ่งดัชนีนั้นจะเป็นดัชนีที่มีประสิทธิภาพในการบ่งชี้ความเหมาะสมของข้อคำถาม โดยใช้การทดสอบ t-test กรณีกลุ่มตัวอย่างเป็นอิสระกัน (independent sample t-test) ซึ่งจากการทดสอบปรากฏผลในตารางที่ 4.42 ที่ผ่านมา จะพบว่าในโมเดล GPCM ดัชนี Generalized $S-\chi^2$ จะให้อำนาจการทดสอบน้อยกว่าดัชนี PARSCALE $G^2$ อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

เนื่องจากในโมเดล GPCM จำนวนรายการคำตอบส่งผลต่ออำนาจการทดสอบอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 จึงต้องมีการเปรียบเทียบรายคู่ว่าจำนวนรายการคำตอบคู่ใดบ้างที่ให้อำนาจการทดสอบแตกต่างกันโดยใช้การเปรียบเทียบพหุคูณ (multiple comparison) แต่เนื่องจากในหัวข้อนี้มีวัตถุประสงค์เพื่อเปรียบเทียบอำนาจการทดสอบของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิด คือ Generalized $S-\chi^2$ และ PARSCALE $G^2$ ดังนั้นเมื่อพบว่าไม่มีปฏิสัมพันธ์ระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามและจำนวนรายการคำตอบในโมเดล GPCM ผู้วิจัยจะเลือกวิเคราะห์เฉพาะการเปรียบเทียบอำนาจการทดสอบ ของดัชนีทั้งสองชนิดที่ใช้ในการศึกษาเท่านั้น โดยไม่พิจารณาว่าในโมเดล GPCM จำนวนรายการคำตอบที่แตกต่างกันส่งผลอย่างไรต่ออำนาจการทดสอบอีก เนื่องจากได้ศึกษาวิเคราะห์ไปแล้วใน ตอนที่ 3 ผลการวิเคราะห์อำนาจการทดสอบ (Power of the test)

จากรายละเอียดในข้อ 4.2 ผลการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิด คือ Generalized $S-\chi^2$ และ PARSCALE $G^2$ โดยใช้สถิติการวิเคราะห์ความแปรปรวนแบบ 2 ทาง (two-way ANOVA) พบว่า ในการใช้เกณฑ์ค่าความคลาดเคลื่อนประเภทที่ 1 พบว่ามีเพียง 2 กรณี คือ 1.การวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขความยาวแบบวัดที่ต่างกัน 3 ระดับ กรณีโมเดล GRM และ 2.การวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขจำนวนรายการคำตอบที่ต่างกัน 4 ระดับ กรณีโมเดล GRM เท่านั้นที่มีอิทธิพลปฏิสัมพันธ์ระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามกับเงื่อนไขที่ใช้ในการศึกษา ซึ่งแสดงว่าในโมเดล GRM ประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดภายใต้เงื่อนไขความยาวแบบวัดและเงื่อนไขจำนวนรายการคำตอบที่แตกต่างกัน จะส่งผลให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) มีค่าแตกต่างกันด้วย นั่นคือ ชนิดของดัชนีความสอดคล้องของข้อคำถามและ

เงื่อนไขความยาวแบบวัดและเงื่อนไขจำนวนรายการคำตอบมีผลร่วมกันต่อค่าความคลาดเคลื่อนประเภทที่ 1

ในการใช้เกณฑ์อำนาจการทดสอบ พบว่ามีเพียง 2 กรณีเช่นกัน คือ 1.การวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างที่ต่างกัน 3 ระดับ กรณีโมเดล GRM และ 2.การวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขความยาวแบบวัดที่ต่างกัน 3 ระดับ กรณีโมเดล GPCM เท่านั้นที่มีอิทธิพลปฏิสัมพันธ์ระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามกับเงื่อนไขที่ใช้ในการศึกษา ซึ่งแสดงว่าในโมเดล GRM ประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างที่แตกต่างกัน จะส่งผลให้อำนาจการทดสอบมีค่าแตกต่างกันด้วย นั่นคือ ชนิดของดัชนีความสอดคล้องของข้อคำถามและเงื่อนไขขนาดกลุ่มตัวอย่างมีผลร่วมกันต่ออำนาจการทดสอบ และในโมเดล GPCM ประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดภายใต้เงื่อนไขความยาวแบบวัดที่แตกต่างกัน จะส่งผลให้อำนาจการทดสอบมีค่าแตกต่างกันด้วย นั่นคือ ชนิดของดัชนีความสอดคล้องของข้อคำถามและเงื่อนไขความยาวแบบวัดมีผลร่วมกันต่ออำนาจการทดสอบ ซึ่งสามารถสรุปได้ดังตารางที่ 4.44

**ตารางที่ 4.44 สรุปผลการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ โดยการใช้สถิติการวิเคราะห์ความแปรปรวนแบบ 2 ทาง (two-way ANOVA)**

| การวิเคราะห์ | อิทธิพลปฏิสัมพันธ์ | ผลการเปรียบเทียบตามชนิดของดัชนี |
|---|---|---|
| **ค่าความคลาดเคลื่อนประเภทที่ 1** | | |
| โมเดล GRM | | |
| ชนิดของดัชนี X ความยาวแบบวัด | มี | ดัชนี Generalized $S-\chi^2$ < ดัชนี PARSCALE $G^2$ |
| ชนิดของดัชนี X ขนาดกลุ่มตัวอย่าง | ไม่มี | ดัชนี Generalized $S-\chi^2$ < ดัชนี PARSCALE $G^2$ |
| ชนิดของดัชนี X จำนวนรายการคำตอบ | มี | ดัชนี Generalized $S-\chi^2$ < ดัชนี PARSCALE $G^2$ เฉพาะที่ 3 และ 9 รายการคำตอบ แต่ที่ 5 และ 7 รายการคำตอบไม่แตกต่างกันอย่างมีนัยสำคัญทางสถิติ |
| โมเดล GPCM | | |
| ชนิดของดัชนี X ความยาวแบบวัด | ไม่มี | ดัชนี Generalized $S-\chi^2$ < ดัชนี PARSCALE $G^2$ |
| ชนิดของดัชนี X ขนาดกลุ่มตัวอย่าง | ไม่มี | ดัชนี Generalized $S-\chi^2$ < ดัชนี PARSCALE $G^2$ |
| ชนิดของดัชนี X จำนวนรายการคำตอบ | ไม่มี | ดัชนี Generalized $S-\chi^2$ < ดัชนี PARSCALE $G^2$ |
| **อำนาจการทดสอบ** | | |
| โมเดล GRM | | |
| ชนิดของดัชนี X ความยาวแบบวัด | ไม่มี | ดัชนี PARSCALE $G^2$ > ดัชนี Generalized $S-\chi^2$ |
| ชนิดของดัชนี X ขนาดกลุ่มตัวอย่าง | มี | ดัชนี PARSCALE $G^2$ > ดัชนี Generalized $S-\chi^2$ |
| ชนิดของดัชนี X จำนวนรายการคำตอบ | ไม่มี | ดัชนี PARSCALE $G^2$ > ดัชนี Generalized $S-\chi^2$ |
| โมเดล GPCM | | |
| ชนิดของดัชนี X ความยาวแบบวัด | มี | ดัชนี PARSCALE $G^2$ > ดัชนี Generalized $S-\chi^2$ |
| ชนิดของดัชนี X ขนาดกลุ่มตัวอย่าง | ไม่มี | ดัชนี PARSCALE $G^2$ > ดัชนี Generalized $S-\chi^2$ |
| ชนิดของดัชนี X จำนวนรายการคำตอบ | ไม่มี | ดัชนี PARSCALE $G^2$ > ดัชนี Generalized $S-\chi^2$ |

โดยเมื่อพิจารณารายละเอียดตามที่ได้สรุปดังตารางที่ 4.44 ในการใช้เกณฑ์ค่าความคลาดเคลื่อนประเภทที่ 1 พบว่าดัชนี Generalized $S-\chi^2$ ให้ค่าความคลาดเคลื่อนประเภทที่ 1 น้อยกว่าดัชนี PARSCALE $G^2$ ในเกือบทั้ง 6 กรณีที่ทำการศึกษา ยกเว้น ในโมเดล GRM ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขจำนวนรายการคำตอบที่ต่างกัน 4 ระดับ ที่จำนวนรายการคำตอบ 5 และ 7 รายการที่ค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนีทั้งสองชนิด

ไม่แตกต่างกันอย่างมีนัยสำคัญทางสถิติ แสดงว่า ดัชนี Generalized $S-\chi^2$ มีโอกาสในการบ่งชี้ข้อคำถามที่สอดคล้อง (fit) กับโมเดลว่าเป็นข้อคำถามที่ไม่สอดคล้อง (misfit) กับโมเดลน้อยกว่าดัชนี PARSCALE $G^2$

อย่างไรก็ตามเมื่อพิจารณาที่เกณฑ์อำนาจการทดสอบ กลับพบว่าดัชนี PARSCALE $G^2$ ให้อำนาจการทดสอบที่สูงกว่าดัชนี Generalized $S-\chi^2$ แสดงว่าดัชนี PARSCALE $G^2$ มีโอกาสในการบ่งชี้ข้อคำถามที่ไม่สอดคล้อง (misfit) กับโมเดลว่าเป็นข้อคำถามที่ไม่สอดคล้อง (misfit) กับโมเดลมากกว่าดัชนี Generalized $S-\chi^2$

ถ้าอ้างอิงหลักการของ Kang และ Chen (2008) มาช่วยในการพิจารณา โดยหลักการของ Kang และ Chen (2008) ได้ให้ความสำคัญกับเกณฑ์ค่าความคลาดเคลื่อนประเภทที่ 1 มาเป็นอันดับแรก โดยพวกเขาคิดว่า ดัชนีความสอดคล้องของข้อคำถามที่ดีนั้น ควรมีโอกาสในการบ่งชี้ข้อคำถามที่มีความสอดคล้อง (fit) กับโมเดลว่าเป็นข้อคำถามที่ไม่สอดคล้อง (misfit) กับโมเดลให้น้อยที่สุด ดังปรากฏในงานวิจัยของ Kang และ Chen ในปี ค.ศ. 2008 ว่าถ้าค่าความคลาดเคลื่อนประเภทที่ 1 ที่คำนวณได้มีค่าที่มาก (Type I error rate are considerably inflated) ก็จะไม่พิจารณาอำนาจการทดสอบ (Power of the test) มาประกอบในการตัดสินประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถาม จะพบว่า ดัชนี Generalized $S-\chi^2$ จะเป็นดัชนีที่มีประสิทธิภาพในการบ่งชี้ความเหมาะสมของข้อคำถามมากกว่าดัชนี PARSCALE $G^2$

# บทที่ 5

# สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

การวิจัยครั้งนี้เป็นการวิจัยเชิงทดลองโดยมีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถาม (item fit index) สำหรับโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค สองชนิดคือ Generalized $S-\chi^2$ และ PARSCALE $G^2$ ภายใต้สถานการณ์ ณ ความยาวแบบวัดเป็น 10, 20, และ 40 ข้อ ขนาดกลุ่มตัวอย่างเป็น 500, 1000, และ 2000 คน จำนวนรายการคำตอบเป็น 3, 5, 7, และ 9 รายการ โดยในแต่ละสถานการณ์ที่ทำการศึกษามีการกระทำซ้ำ 30 ครั้ง

โมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาคที่ใช้ในการวิจัยครั้งนี้เป็นโมเดลทฤษฎีการตอบสนองข้อสอบแบบเอกมิติ (Unidimensional IRT model) ประกอบด้วย 2 โมเดลคือ Grade Response Model (GRM) และ Generalized Partial Credit Model (GPCM) เนื่องจากโมเดลทั้งสองพัฒนาบนพื้นฐานของโมเดลแบบ 2 พารามิเตอร์เหมือนกัน ซึ่งเหมาะสำหรับข้อสอบหรือข้อคำถามที่มีความยากและค่าอำนาจจำแนกที่แตกต่างกัน รวมทั้งโมเดลทั้งสองยังเป็นโมเดลที่มีผู้สนใจศึกษากันมาก ไม่เข้มงวดเกี่ยวกับข้อตกลงเบื้องต้นและสามารถใช้กับแบบสอบและแบบวัดหลายลักษณะ

การจำลองข้อมูลประกอบด้วย 3 ขั้นตอน ได้แก่ 1) การจำลองข้อมูลค่าพารามิเตอร์ความสามารถผู้สอบ (Generating in ability parameter values) 2) การจำลองข้อมูลค่าพารามิเตอร์ข้อคำถาม (Generating in item parameter values) และ 3) การจำลองข้อมูลการตอบข้อคำถาม (Simulating item response data) ซึ่งทั้ง 3 ขั้นตอนนี้ดำเนินการโดยใช้โปรแกรม WINGEN ของ Hambleton และ Han (2007)

เกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถาม (item fit index) ในการวิจัยนี้คือ 1) ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ซึ่งเป็นการบ่งชี้ว่าข้อคำถามข้อนั้นไม่สอดคล้อง (misfit) ทั้ง ๆ ที่ข้อคำถามข้อนั้นมีความสอดคล้อง (fit) กับโมเดลทฤษฎีการตอบสนองข้อสอบและ 2) อำนาจการทดสอบ (Power of the test) ซึ่งเป็นการบ่งชี้ว่าข้อคำถามข้อนั้นไม่สอดคล้อง (misfit) ซึ่งในความเป็นจริงข้อคำถามข้อนั้นไม่มีความสอดคล้อง (misfit) กับโมเดลทฤษฎีการตอบสนองข้อสอบ

ในการพิจารณาเพื่อเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถาม (item fit index) จำแนกเป็น 2 กรณี คือ 1.การเปรียบเทียบประสิทธิภาพของดัชนีความ

สอดคล้องของข้อคำถามสองชนิดคือ Generalized $\boldsymbol{S-\chi^2}$ และ PARSCALE $G^2$ ตามเงื่อนไขของ Kang และ Chen (2008) และ 2.การเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิด คือ Generalized $\boldsymbol{S-\chi^2}$ และ PARSCALE $G^2$ โดยใช้สถิติการวิเคราะห์ความแปรปรวนแบบ 2 ทาง (two-way ANOVA)

ถ้าเป็นการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดตามเงื่อนไขของ Kang และ Chen (2008) จะพิจารณาที่ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ก่อน ซึ่งถ้าดัชนีชนิดใดมีค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) น้อยกว่า ดัชนีนั้นจะเป็นดัชนีที่มีประสิทธิภาพในการบ่งชี้ความสอดคล้องของข้อคำถาม ในสถานการณ์นั้น แต่ถ้าหากดัชนีทั้งสองมีค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) เท่ากันจึงจะพิจารณาอำนาจการทดสอบ (Power of the test) โดยถ้าดัชนีชนิดใดมีอำนาจการทดสอบที่มากกว่า ดัชนีนั้นจะเป็นดัชนีที่มีประสิทธิภาพในการบ่งชี้ความสอดคล้องของข้อคำถามกับโมเดลทฤษฎีการตอบสนองข้อสอบในสถานการณ์นั้น

ถ้าเป็นการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดโดยใช้สถิติการวิเคราะห์ความแปรปรวนแบบ 2 ทาง (two-way ANOVA) โดยการทดสอบผลของชนิดของดัชนีความสอดคล้องของข้อคำถาม และผลของปฏิสัมพันธ์ระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามกับเงื่อนไขที่ใช้ในการศึกษา (ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ) ว่ามีนัยสำคัญทางสถิติหรือไม่ ซึ่งถ้าชนิดของดัชนีความสอดคล้องของข้อคำถามกับเงื่อนไขที่ใช้ในการศึกษา มีอิทธิพลปฏิสัมพันธ์ (interaction effect) กันย่อมแสดงให้เห็นว่า ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดภายใต้เงื่อนไขที่ใช้ในการศึกษาที่ต่างกัน จะส่งผลให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) และอำนาจการทดสอบ (Power of the test) มีค่าแตกต่างกัน ซึ่งถ้าชนิดของดัชนีความสอดคล้องของข้อคำถามกับเงื่อนไขที่ใช้ในการศึกษา (ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ) นั้นมีอิทธิพลปฏิสัมพันธ์ (interaction effect) กัน ผู้วิจัยจะวิเคราะห์ต่อไปด้วยเทคนิค Simple effect โดยศึกษาเปรียบเทียบค่าเฉลี่ยของค่าความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของดัชนีทั้งสองชนิดสำหรับแต่ละระดับของเงื่อนไขที่ใช้ในการศึกษา (ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ)ที่มีอิทธิพลปฏิสัมพันธ์กับชนิดของดัชนีความสอดคล้องของข้อคำถาม แต่จะไม่ทำการวิเคราะห์ต่อไปด้วยเทคนิค Simple effect ในกรณีเปรียบเทียบค่าเฉลี่ยของค่าความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบระหว่างแต่ละระดับของเงื่อนไขที่ใช้ในการศึกษาจำแนกตามชนิดของดัชนี เนื่องจากได้ทำการศึกษาวิเคราะห์ไป

แล้วในขั้นตอนการพิจารณาค่าความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ จำแนกตามโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาคที่ใช้ในการศึกษา แต่ถ้าหากไม่มีอิทธิพลปฏิสัมพันธ์ (interaction effect) กันระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามกับเงื่อนไขที่ใช้ในการศึกษา จะทำการเปรียบเทียบค่าเฉลี่ยของค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) และอำนาจการทดสอบของดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ จำแนกตามโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค โดยไม่เปรียบเทียบค่าเฉลี่ยของค่าความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของดัชนีทั้งสองชนิดสำหรับแต่ละระดับของเงื่อนไขที่ใช้ในการศึกษา (ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ) แต่จะใช้การทดสอบ t-test กรณีกลุ่มตัวอย่างเป็นอิสระกัน (independent sample t-test) ซึ่งในการวิเคราะห์ความแปรปรวนแบบ 2 ทาง (two-way ANOVA) จำแนกได้ทั้งหมด 12 กรณี แบ่งออกเป็นการวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) 6 กรณีและการวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบ (Power of the test) 6 กรณี ดังที่ได้กล่าวไปแล้วในบทที่ 4 ในบทนี้ ผู้วิจัยสรุปผลการวิจัย อภิปรายผลการวิจัย รวมทั้งข้อเสนอแนะ ดังต่อไปนี้

**สรุปผลการวิจัย**

ในการสรุปผลการวิจัยครั้งนี้อยู่ภายใต้ข้อสมมติว่า “ข้อมูลที่จำลองขึ้นและนำมาวิเคราะห์นั้นได้มาจากแบบวัดที่มีความตรงตามเนื้อหา (Content Validity) และมีความเที่ยง (Reliability)” กล่าวคือ แบบวัดมีความเพียงพอและมีความเป็นตัวแทนคุณลักษณะที่มุ่งวัด รวมทั้งยังให้ผลการวัดที่มีความคงเส้นคงวา ดังนั้นการสรุปผลการวิจัยในการศึกษาครั้งนี้จึงอยู่ภายใต้เงื่อนไขข้อสมมติที่ได้กล่าวไปแล้ว ในการพิจารณาผลการวิจัยหรือนำผลการวิจัยไปใช้จึงต้องอยู่ภายใต้ข้อสมมติดังกล่าวด้วย ในการสรุปผลการวิจัยนี้ จำแนกออกเป็น 2 ส่วน ดังนี้

<u>ส่วนที่ 1 ผลการวิเคราะห์ค่าความคลาดเคลื่อนประเภทที่ 1และอำนาจการทดสอบ</u>

ผลการวิเคราะห์ค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนี PARSCALE $G^2$ สรุปสาระสำคัญได้ว่า เมื่อนำค่าความคลาดเคลื่อนประเภทที่ 1 ไปทดสอบเพื่อสังเกตความเปลี่ยนแปลงของค่าความคลาดเคลื่อนประเภทที่ 1 เมื่อเงื่อนไขที่ใช้ในการศึกษา (ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ) มีการเปลี่ยนแปลงไป จะพบว่าในโมเดล GRM มีเพียงเงื่อนไขความยาวแบบวัดและจำนวนรายการคำตอบเท่านั้นที่มีการเปลี่ยนแปลงของค่าความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติ โดยสอดคล้องกับผลการพิจารณาของผู้วิจัยว่า ที่โมเดล GRM เมื่อความยาวแบบวัดเพิ่มขึ้น ก็มีแนวโน้มที่ค่าความคลาดเคลื่อนประเภทที่ 1

จะลดลงด้วย และที่โมเดล GRM เมื่อจำนวนรายการคำตอบเพิ่มขึ้น ค่าความคลาดเคลื่อนประเภทที่ 1 จะลดลง ดังในผลการทดสอบเปรียบเทียบพหุคูณ (Multiple Comparison) ที่ให้สารสนเทศว่า ในโมเดล GRM ความยาวแบบวัด 10 ข้อ จะให้ค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนี PARSCALE $G^2$ มากกว่าความยาวแบบวัด 20 และ 40 ข้อ และในโมเดล GRM จำนวนรายการคำตอบ 3 รายการ จะให้ค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนี PARSCALE $G^2$ มากกว่าจำนวนรายการคำตอบ 7 และ 9 รายการ

ผลการวิเคราะห์ค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนี Generalized $S-\chi^2$ สรุปสาระสำคัญได้ว่า เมื่อนำค่าความคลาดเคลื่อนประเภทที่ 1 ไปทดสอบเพื่อสังเกตความเปลี่ยนแปลงของค่าความคลาดเคลื่อนประเภทที่ 1 เมื่อเงื่อนไขที่ใช้ในการศึกษา (ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ) มีการเปลี่ยนแปลงไป จะพบว่าในโมเดล GRM มีเพียงเงื่อนไขความยาวแบบวัดเท่านั้นที่มีการเปลี่ยนแปลงของค่าความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติ โดยสอดคล้องกับผลการพิจารณาของผู้วิจัยว่า ที่โมเดล GRM เมื่อความยาวแบบวัดเพิ่มขึ้น ก็มีแนวโน้มที่ค่าความคลาดเคลื่อนประเภทที่ 1 จะลดลงด้วย และที่โมเดล GPCM มีเพียงเงื่อนไขจำนวนรายการคำตอบเท่านั้นที่มีการเปลี่ยนแปลงของค่าความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติ โดยสอดคล้องกับผลการพิจารณาของผู้วิจัยว่า ในโมเดล GPCM เมื่อจำนวนรายการคำตอบเพิ่มขึ้น ค่าความคลาดเคลื่อนประเภทที่ 1 จะไม่ลดลง ดังในผลการทดสอบเปรียบเทียบพหุคูณ (Multiple Comparison) ที่ให้สารสนเทศว่า ในโมเดล GRM ความยาวแบบวัด 10 ข้อ จะให้ค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนี Generalized $S-\chi^2$ มากกว่าความยาวแบบวัด 20 และ 40 ข้อ และในโมเดล GPCM จำนวนรายการคำตอบ 9 รายการ จะให้ค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนี PARSCALE $G^2$ มากกว่าจำนวนรายการคำตอบ 3 และ 5 รายการ แต่ไม่แตกต่างจากจำนวนรายการคำตอบ 7 รายการ

จะเห็นได้ว่า ผลการทดสอบเปรียบเทียบพหุคูณ (Multiple Comparison) ของดัชนีทั้งสองชนิดให้สารสนเทศที่สอดคล้องกันในโมเดล GRM ว่าความยาวแบบวัด 10 ข้อ จะให้ค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนีทั้งสองชนิดมากกว่าความยาวแบบวัด 20 และ 40 ข้อ แสดงให้เห็นว่า ความยาวแบบวัดมีผลต่อค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนีทั้งสองชนิดเหมือนกัน

ผลการวิเคราะห์อำนาจการทดสอบของดัชนี PARSCALE $G^2$ สรุปสาระสำคัญได้ว่า เมื่อนำอำนาจการทดสอบไปทดสอบเพื่อสังเกตความเปลี่ยนแปลงของอำนาจการทดสอบเมื่อ

เงื่อนไขที่ใช้ในการศึกษา (ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ) มีการเปลี่ยนแปลงไป จะพบว่าในโมเดล GRM มีเพียงเงื่อนไขความยาวแบบวัดและขนาดกลุ่มตัวอย่างเท่านั้นที่มีการเปลี่ยนแปลงของอำนาจการทดสอบอย่างมีนัยสำคัญทางสถิติ โดยสอดคล้องกับผลการพิจารณาของผู้วิจัยว่า ที่โมเดล GRM เมื่อความยาวแบบวัดเพิ่มขึ้น โดยส่วนใหญ่อำนาจการทดสอบจะมีค่าลดลงมีเพียงบางกรณีเท่านั้นที่อำนาจการทดสอบเพิ่มขึ้น และที่โมเดล GRM เมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้น อำนาจการทดสอบจะมีค่าเพิ่มมากขึ้นในทุกกรณีที่ทำการศึกษา ดังในผลการทดสอบเปรียบเทียบพหุคูณ (Multiple Comparison) ที่ให้สารสนเทศว่า ในโมเดล GRM ความยาวแบบวัด 10 ข้อ จะให้อำนาจการทดสอบของดัชนี PARSCALE $G^2$ มากกว่าความยาวแบบวัด 40 ข้อ และในโมเดล GRM ขนาดกลุ่มตัวอย่าง 2000 คน จะให้อำนาจการทดสอบของดัชนี PARSCALE $G^2$ มากกว่าขนาดกลุ่มตัวอย่าง 500 และ 1000 คน นอกจากนี้ในโมเดล GRM ขนาดกลุ่มตัวอย่าง 1000 คน จะให้อำนาจการทดสอบของดัชนี PARSCALE $G^2$ มากกว่าขนาดกลุ่มตัวอย่าง 500 คนด้วย เมื่อพิจารณาที่โมเดล GPCM จะมี 2 เงื่อนไขเช่นเดียวกันกับโมเดล GRM คือเงื่อนไขความยาวแบบวัดและขนาดกลุ่มตัวอย่างเท่านั้นที่มีการเปลี่ยนแปลงของอำนาจการทดสอบอย่างมีนัยสำคัญทางสถิติ โดยสอดคล้องกับผลการพิจารณาของผู้วิจัยว่า ที่โมเดล GPCM เมื่อความยาวแบบวัดเพิ่มขึ้น ก็มีแนวโน้มที่อำนาจการทดสอบจะเพิ่มขึ้นด้วย และที่โมเดล GPCM เมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้นอำนาจการทดสอบก็จะเพิ่มขึ้นเช่นเดียวกับโมเดล GRM ดังในผลการทดสอบเปรียบเทียบพหุคูณ (Multiple Comparison) ที่ให้สารสนเทศว่า ในโมเดล GPCM ความยาวแบบวัด 40 ข้อ ให้อำนาจการทดสอบมากกว่าความยาวแบบวัด 10 และ 20 ข้อ และในโมเดล GPCM ขนาดกลุ่มตัวอย่าง 2000 คน ให้อำนาจการทดสอบมากกว่าขนาดกลุ่มตัวอย่าง 500 คน

ผลการวิเคราะห์อำนาจการทดสอบของดัชนี Generalized $S-\chi^2$ สรุปสาระสำคัญได้ว่า เมื่อนำอำนาจการทดสอบไปทดสอบเพื่อสังเกตความเปลี่ยนแปลงของอำนาจการทดสอบเมื่อเงื่อนไขที่ใช้ในการศึกษา (ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ) มีการเปลี่ยนแปลงไป จะพบว่าในโมเดล GRM มีเพียงเงื่อนไขขนาดกลุ่มตัวอย่างเท่านั้นที่มีการเปลี่ยนแปลงของอำนาจการทดสอบอย่างมีนัยสำคัญทางสถิติ โดยสอดคล้องกับผลการพิจารณาของผู้วิจัยว่า ที่โมเดล GRM เมื่อขนาดกลุ่มตัวอย่างเพิ่มขึ้น อำนาจการทดสอบจะมีค่าเพิ่มมากขึ้น และที่โมเดล GPCM มีเพียงเงื่อนไขจำนวนรายการคำตอบเท่านั้นที่มีการเปลี่ยนแปลงของอำนาจการทดสอบอย่างมีนัยสำคัญทางสถิติ โดยสอดคล้องกับผลการพิจารณาของผู้วิจัยว่าโดยส่วนใหญ่ เมื่อจำนวนรายการคำตอบเพิ่มขึ้น อำนาจการทดสอบจะเพิ่มขึ้น ดังในผลการ

ทดสอบเปรียบเทียบพหุคูณ (Multiple Comparison) ที่ให้สารสนเทศว่า ในโมเดล GRM ขนาดกลุ่มตัวอย่าง 2000 คน จะให้อำนาจการทดสอบของดัชนี Generalized $S-\chi^2$ มากกว่าขนาดกลุ่มตัวอย่าง 500 คน และในโมเดล GPCM จำนวนรายการคำตอบ 9 รายการ จะให้อำนาจการทดสอบของดัชนี Generalized $S-\chi^2$ มากกว่าจำนวนรายการคำตอบ 5 รายการ

จะเห็นได้ว่า ผลการทดสอบเปรียบเทียบพหุคูณ (Multiple Comparison) ของดัชนีทั้งสองชนิดให้สารสนเทศที่สอดคล้องกันในโมเดล GRM ว่าขนาดกลุ่มตัวอย่างมาก (2000 คน) จะให้อำนาจการทดสอบของดัชนีทั้งสองชนิดมากกว่าขนาดกลุ่มตัวอย่างน้อย (500 คน) แสดงให้เห็นว่า ขนาดกลุ่มตัวอย่างมีผลต่ออำนาจการทดสอบของดัชนีทั้งสองชนิดเหมือนกัน

<u>ส่วนที่ 2 ผลการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถาม</u>

ผลการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถาม (item fit index) จำแนกเป็น 2 กรณี คือ 1.ผลการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามสองชนิดคือ Generalized $\boldsymbol{S-\chi^2}$ และ PARSCALE $G^2$ ตามเงื่อนไขของ Kang และ Chen (2008) และ 2.ผลการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิด คือ Generalized $\boldsymbol{S-\chi^2}$ และ PARSCALE $G^2$ โดยใช้สถิติการวิเคราะห์ความแปรปรวนแบบ 2 ทาง (two-way ANOVA)

1. ผลการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามสองชนิดคือ Generalized $S-\chi^2$ และ PARSCALE $G^2$ ตามเงื่อนไขของ Kang และ Chen (2008)

ผลการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามสองชนิดในทั้ง 72 สถานการณ์ที่ทำการศึกษา ทั้ง 70 สถานการณ์ที่ทำการศึกษาดัชนี Generalized $S-\chi^2$ มีประสิทธิภาพในการบ่งชี้ความสอดคล้องของข้อคำถามมากกว่าดัชนี PARSCALE $G^2$ ยกเว้นใน 2 สถานการณ์คือ สถานการณ์ที่ 3 กรณีโมเดล GRM ความยาวแบบวัด 10 ข้อ ขนาดกลุ่มตัวอย่าง 500 คน จำนวนรายการคำตอบ 7 รายการ และสถานการณ์ที่ 32 กรณีโมเดล GRM ความยาวแบบวัด 40 ข้อ ขนาดกลุ่มตัวอย่าง 1000 คน จำนวนรายการคำตอบ 9 รายการ ที่ดัชนี PARSCALE $G^2$ มีประสิทธิภาพในการบ่งชี้ความสอดคล้องของข้อคำถามมากกว่าดัชนี Generalized $S-\chi^2$ ทั้งนี้ผลการเปรียบเทียบดังกล่าว ใช้เพียงค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) เท่านั้นเป็นเกณฑ์ในการเปรียบเทียบ ไม่ได้พิจารณาที่อำนาจการทดสอบ (Power of the test) เนื่องจากขั้นตอนในการพิจารณาเพื่อเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถาม (item fit index) ที่จะพิจารณาค่าความคลาดเคลื่อนประเภทที่ 1 ก่อน ซึ่ง

ถ้าดัชนีชนิดใดมีค่าความคลาดเคลื่อนประเภทที่ 1 น้อยกว่า ดัชนีนั้นจะเป็นดัชนีที่มีประสิทธิภาพในการบ่งชี้ความสอดคล้องของข้อคำถามในสถานการณ์นั้น

2. ผลการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิด คือ Generalized $S-\chi^2$ และ PARSCALE $G^2$ โดยใช้สถิติการวิเคราะห์ความแปรปรวนแบบ 2 ทาง (two-way ANOVA)

ผลการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามสองชนิดในทั้ง 72 สถานการณ์ที่ทำการศึกษา โดยการใช้เกณฑ์ค่าความคลาดเคลื่อนประเภทที่ 1 พบว่ามีเพียง 2 กรณี คือ 1.การวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขความยาวแบบวัดที่ต่างกัน 3 ระดับ กรณีโมเดล GRM และ 2.การวิเคราะห์ความแปรปรวนแบบ 2 ทางของค่าความคลาดเคลื่อนประเภทที่ 1 ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขจำนวนรายการคำตอบที่ต่างกัน 4 ระดับ กรณีโมเดล GRM เท่านั้นที่มีอิทธิพลปฏิสัมพันธ์ระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามกับเงื่อนไขที่ใช้ในการศึกษา ซึ่งแสดงว่าในโมเดล GRM ประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดภายใต้เงื่อนไขความยาวแบบวัดและเงื่อนไขจำนวนรายการคำตอบที่แตกต่างกัน จะส่งผลให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) มีค่าแตกต่างกันด้วย นั่นคือ ชนิดของดัชนีความสอดคล้องของข้อคำถามและเงื่อนไขความยาวแบบวัดและจำนวนรายการคำตอบมีผลร่วมกันต่อค่าความคลาดเคลื่อนประเภทที่ 1

ในการใช้เกณฑ์อำนาจการทดสอบ พบว่ามีเพียง 2 กรณีเช่นกัน คือ 1.การวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างที่ต่างกัน 3 ระดับ กรณีโมเดล GRM และ 2.การวิเคราะห์ความแปรปรวนแบบ 2 ทางของอำนาจการทดสอบในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขความยาวแบบวัดที่ต่างกัน 3 ระดับ กรณีโมเดล GPCM เท่านั้นที่มีอิทธิพลปฏิสัมพันธ์ระหว่างชนิดของดัชนีความสอดคล้องของข้อคำถามกับเงื่อนไขที่ใช้ในการศึกษา ซึ่งแสดงว่าในโมเดล GRM ประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดภายใต้เงื่อนไข

ขนาดกลุ่มตัวอย่างที่แตกต่างกัน จะส่งผลให้อำนาจการทดสอบมีค่าแตกต่างกันด้วย นั่นคือ ชนิดของดัชนีความสอดคล้องของข้อคำถามและเงื่อนไขขนาดกลุ่มตัวอย่างมีผลร่วมกันต่ออำนาจการทดสอบ และในโมเดล GPCM ประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดภายใต้เงื่อนไขความยาวแบบวัดที่แตกต่างกัน จะส่งผลให้อำนาจการทดสอบมีค่าแตกต่างกันด้วย นั่นคือ ชนิดของดัชนีความสอดคล้องของข้อคำถามและเงื่อนไขความยาวแบบวัดมีผลร่วมกันต่ออำนาจการทดสอบ

นอกจากนี้ ในการใช้เกณฑ์ค่าความคลาดเคลื่อนประเภทที่ 1 พบว่า ดัชนี Generalized $S-\chi^2$ ให้ค่าความคลาดเคลื่อนประเภทที่ 1 น้อยกว่าดัชนี PARSCALE $G^2$ ในเกือบทั้ง 6 กรณีที่ทำการศึกษา ยกเว้น ในโมเดล GRM ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขจำนวนรายการคำตอบที่ต่างกัน 4 ระดับ ที่จำนวนรายการคำตอบ 5 และ 7 รายการที่ค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนีทั้งสองชนิดไม่แตกต่างกันอย่างมีนัยสำคัญทางสถิติ แสดงว่า ดัชนี Generalized $S-\chi^2$ มีโอกาสในการบ่งชี้ข้อคำถามที่สอดคล้อง (fit) กับโมเดลว่าเป็นข้อคำถามที่ไม่สอดคล้อง (misfit) กับโมเดลน้อยกว่าดัชนี PARSCALE $G^2$

อย่างไรก็ตามเมื่อพิจารณาที่เกณฑ์อำนาจการทดสอบ กลับพบว่าดัชนี PARSCALE $G^2$ ให้อำนาจการทดสอบที่สูงกว่าดัชนี Generalized $S-\chi^2$ แสดงว่าดัชนี PARSCALE $G^2$ มีโอกาสในการบ่งชี้ข้อคำถามที่ไม่สอดคล้อง (misfit) กับโมเดลว่าเป็นข้อคำถามที่ไม่สอดคล้อง (misfit) กับโมเดลมากกว่าดัชนี Generalized $S-\chi^2$

**อภิปรายผลการวิจัย**

ในการอภิปรายผลในงานวิจัยนี้อยู่ภายใต้ข้อสมมติว่า “ข้อมูลที่จำลองขึ้นและนำมาวิเคราะห์นั้นได้มาจากแบบวัดที่มีความตรงตามเนื้อหา (Content Validity) และมีความเที่ยง (Reliability)” กล่าวคือ แบบวัดมีความเพียงพอและมีความเป็นตัวแทนคุณลักษณะที่มุ่งวัด รวมทั้งยังให้ผลการวัดที่มีความคงเส้นคงวา นอกจากนี้ ข้อมูลที่จำลองขึ้นจากโปรแกรม WINGEN นั้น มีลักษณะสอดคล้องและเป็นไปตามโมเดลทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค ที่ใช้ในการศึกษาวิจัยทั้ง 2 โมเดลคือ GRM และ GPCM โดยนำเสนอการอภิปรายผลการวิจัยตามสมมติฐานการวิจัยและผลการวิจัยดังนี้

จากสมมติฐานการวิจัยที่เชื่อว่าดัชนี Generalized $S-\chi^2$ มีประสิทธิภาพในการบ่งชี้ความสอดคล้องของข้อคำถามมากกว่าดัชนี PARSCALE $G^2$ ในสถานการณ์ทั้งหมด 72 สถานการณ์นั้น (ณ ความยาวแบบวัดเป็น 10, 20, และ 40 ข้อ ขนาดกลุ่มตัวอย่างเป็น 500,

1000, และ 2000 คน จำนวนรายการคำตอบเป็น 3, 5, 7, และ 9 รายการ) จากผลการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถาม (item fit index) ซึ่งจำแนกเป็น 2 กรณี คือ 1.ผลการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามสองชนิดตามเงื่อนไขของ Kang และ Chen (2008) และ 2.ผลการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดโดยใช้สถิติการวิเคราะห์ความแปรปรวนแบบ 2 ทาง (two-way ANOVA) พบว่าในกรณีที่ 1 ผลการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามสองชนิดตามเงื่อนไขของ Kang และ Chen (2008) พบว่าจากสถานการณ์ทั้งหมดที่ทำการศึกษา 72 สถานการณ์ ผลการวิจัยสอดคล้องกับสมมติฐาน 70 สถานการณ์ มีเพียง 2 สถานการณ์เท่านั้นที่ไม่สอดคล้องกับสมมติฐาน เนื่องจากดัชนี PARSCALE $G^2$ มีค่าความคลาดเคลื่อนประเภทที่ 1 น้อยกว่าดัชนี Generalized $S-\chi^2$ ซึ่งแสดงว่าดัชนี PARSCALE $G^2$ มีประสิทธิภาพในการบ่งชี้ความสอดคล้องของข้อคำถามมากกว่าดัชนี Generalized $S-\chi^2$ กล่าวคือ ในสถานการณ์ที่ 3 กรณีโมเดล GRM ความยาวแบบวัด 10 ข้อ ขนาดกลุ่มตัวอย่าง 500 คน จำนวนรายการคำตอบ 7 รายการ ดัชนี PARSCALE $G^2$ มีค่าความคลาดเคลื่อนประเภทที่ 1 น้อยกว่าดัชนี Generalized $S-\chi^2$ (0.0133 < 0.0400) และในสถานการณ์ที่ 32 กรณีโมเดล GRM ความยาวแบบวัด 40 ข้อ ขนาดกลุ่มตัวอย่าง 1000 คน จำนวนรายการคำตอบ 9 รายการ ดัชนี PARSCALE $G^2$ มีค่าความคลาดเคลื่อนประเภทที่ 1 น้อยกว่าดัชนี Generalized $S-\chi^2$ (0.0183 < 0.0333) ซึ่งสอดคล้องกับผลการวิจัยของ Kang และ Chen (2008) ที่ ในกรณีความยาวแบบวัดที่ยาว ขนาดกลุ่มตัวอย่าง 1000 คน ดัชนี PARSCALE $G^2$ มีค่าความคลาดเคลื่อนประเภทที่ 1 น้อยกว่าดัชนี Generalized $S-\chi^2$ จะเห็นได้ว่า หลักการของ Kang และ Chen (2008) นั้นได้ให้ความสำคัญกับเกณฑ์ค่าความคลาดเคลื่อนประเภทที่ 1 มาเป็นอันดับแรก โดยพวกเขาคิดว่า ดัชนีความสอดคล้องของข้อคำถามที่ดีนั้น ควรมีโอกาสในการบ่งชี้ข้อคำถามที่มีความสอดคล้อง (fit) กับโมเดลว่าเป็นข้อคำถามที่ไม่สอดคล้อง (misfit) กับโมเดลให้น้อยที่สุด

ในกรณีที่ 2 ผลการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดโดยใช้สถิติการวิเคราะห์ความแปรปรวนแบบ 2 ทาง (two-way ANOVA) พบว่าในเกณฑ์ค่าความคลาดเคลื่อนประเภทที่ 1 ในโมเดล GRM ประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดภายใต้เงื่อนไขความยาวแบบวัดและเงื่อนไขจำนวนรายการคำตอบที่แตกต่างกัน จะส่งผลให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) มีค่าแตกต่างกันด้วย นั่นคือ ชนิดของดัชนีความสอดคล้องของข้อคำถามและเงื่อนไขความยาวแบบวัดและจำนวนรายการคำตอบมีผลร่วมกันต่อค่าความคลาดเคลื่อนประเภทที่ 1 ในเกณฑ์อำนาจการทดสอบ ในโมเดล

GRM ประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างที่แตกต่างกัน จะส่งผลให้อำนาจการทดสอบมีค่าแตกต่างกันด้วย นั่นคือ ชนิดของดัชนีความสอดคล้องของข้อคำถามและเงื่อนไขขนาดกลุ่มตัวอย่างมีผลร่วมกันต่ออำนาจการทดสอบ และในโมเดล GPCM ประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามทั้งสองชนิดภายใต้เงื่อนไขความยาวแบบวัดที่แตกต่างกัน จะส่งผลให้อำนาจการทดสอบมีค่าแตกต่างกันด้วย นั่นคือ ชนิดของดัชนีความสอดคล้องของข้อคำถามและเงื่อนไขความยาวแบบวัดมีผลร่วมกันต่ออำนาจการทดสอบ และเมื่อเปรียบเทียบค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนีทั้งสองชนิดพบว่า ดัชนี Generalized $S-\chi^2$ ให้ค่าความคลาดเคลื่อนประเภทที่ 1 น้อยกว่าดัชนี PARSCALE $G^2$ ใน 5 กรณีที่ทำการศึกษา ซึ่งสอดคล้องกับสมมติฐานในการวิจัย ยกเว้น ในโมเดล GRM ในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามระหว่างดัชนี PARSCALE $G^2$ และดัชนี Generalized $S-\chi^2$ ภายใต้เงื่อนไขจำนวนรายการคำตอบที่ต่างกัน 4 ระดับ ที่จำนวนรายการคำตอบ 5 และ 7 รายการที่ค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนีทั้งสองชนิดไม่แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05

เมื่อพิจารณาที่เกณฑ์อำนาจการทดสอบ กลับพบว่าดัชนี PARSCALE $G^2$ ให้อำนาจการทดสอบที่สูงกว่าดัชนี Generalized $S-\chi^2$ และโดยส่วนใหญ่อำนาจการทดสอบของดัชนี Generalized $S-\chi^2$ จะมีค่าค่อนข้างน้อย (มีค่าตั้งแต่ 0.0008 ถึง 0.7567) เมื่อเทียบกับอำนาจการทดสอบของดัชนี PARSCALE $G^2$ (มีค่าตั้งแต่ 0.0233 ถึง 0.9983) แสดงว่าดัชนี PARSCALE $G^2$ มีโอกาสในการบ่งชี้ข้อคำถามที่ไม่สอดคล้อง (misfit) กับโมเดลว่าเป็นข้อคำถามที่ไม่สอดคล้อง (misfit) กับโมเดลมากกว่าดัชนี Generalized $S-\chi^2$ ผลการวิจัยนี้สอดคล้องกับ Kang และ Chen (2010) ที่ได้ทำการศึกษาประสิทธิภาพของดัชนี Generalized $S-\chi^2$ สำหรับ Grade response model (GRM) เมื่อพิจารณาจากเกณฑ์อำนาจการทดสอบภายใต้ M-type misfit ดัชนี Generalized $S-\chi^2$ ไม่มีความไว (insensitive) ในการตรวจสอบความไม่สอดคล้องของข้อคำถามเมื่อขนาดกลุ่มตัวอย่างใหญ่ ซึ่งจากสิ่งที่พบบ่งชี้ว่าดัชนี Generalized $S-\chi^2$ อาจจะไม่เหมาะสมนักที่จะนำมาใช้เป็นเครื่องมือในการประเมินความสอดคล้องของข้อคำถามเมื่อเกิด M-type misfit แต่ถ้าเป็นข้อคำถามที่มีลักษณะ D-type misfit ดัชนี Generalized $S-\chi^2$ จะมีอำนาจการทดสอบที่เหมาะสมในบางสถานการณ์ อย่างไรก็ตามในภาพรวม Kang และ Chen (2010) ยังสรุปผลการวิจัยว่า ภายใต้โมเดล GRM ดัชนี Generalized $S-\chi^2$ ให้ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) และอำนาจการทดสอบ (Power of the test) ที่เหมาะสมในการตรวจสอบความไม่สอดคล้อง (misfit) ทั้งสองประเภทของข้อคำถามในขนาดกลุ่มตัวอย่างที่

ใหญ่ เพราะฉะนั้น Generalized $S-\chi^2$ จึงเป็นดัชนีที่มีแนวโน้มที่ดีในการนำไปใช้บ่งชี้ความสอดคล้องของข้อคำถามที่มีการตรวจให้คะแนนมากกว่า 2 ค่าในการประเมินทางการศึกษาและจิตวิทยา เนื่องจาก Kang และ Chen (2010) มีความเห็นว่า อำนาจการทดสอบนั้นเป็นฟังก์ชันของขนาดกลุ่มตัวอย่างนั่นคืออำนาจการทดสอบมีความสัมพันธ์กับขนาดกลุ่มตัวอย่าง สอดคล้องกับ Lattuis, Clark และ O'Brien (2009) ที่กล่าวว่าอิทธิพลของขนาดกลุ่มตัวอย่างยังคงส่งผลต่ออำนาจการทดสอบ ซึ่ง Liang และ Wells (2009) เสนอว่าในการเปรียบเทียบประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามโดยนำค่าอำนาจการทดสอบ (Power of the test) มาพิจารณาเพียงอย่างเดียวอาจไม่สามารถเชื่อถือได้ ควรพิจารณาที่ค่าความคลาดเคลื่อนประเภทที่ 1 (Type I error) ด้วย เนื่องจากอิทธิพลของขนาดกลุ่มตัวอย่างที่มีต่ออำนาจการทดสอบ ดังที่กล่าวกันว่าอำนาจกาทดสอบจะแปรผันตรงกับขนาดตัวอย่าง ถ้าขนาดตัวอย่างใหญ่ขึ้น อำนาจการทดสอบจะสูงขึ้นด้วย (สุชาดา บวรกิติวงศ์, 2541)

จากการศึกษาวิเคราะห์เพื่อสังเกตความเปลี่ยนแปลงของค่าความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบเมื่อเงื่อนไขที่ใช้ในการศึกษา (ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง จำนวนรายการคำตอบ) มีการเปลี่ยนแปลงไป และได้ทดสอบสมมติฐานทางสถิติเพื่อดูว่าความเปลี่ยนแปลงนั้นเป็นการเปลี่ยนแปลงที่มีนัยสำคัญหรือไม่ พบว่า ในกรณีค่าความคลาดเคลื่อนประเภทที่ 1 ดัชนีทั้งสองชนิดให้ผลสอดคล้องกันว่า ความยาวแบบวัดและจำนวนรายการคำตอบเป็นปัจจัยที่ส่งผลให้เกิดความเปลี่ยนแปลงของค่าความคลาดเคลื่อนประเภทที่ 1 อย่างมีนัยสำคัญทางสถิติ กล่าวคือ เมื่อความยาวแบบวัดเพิ่มขึ้นมีแนวโน้มที่ค่าความคลาดเคลื่อนประเภทที่ 1จะลดลง นอกจากนี้โดยส่วนใหญ่แล้ว จากการพิจารณาค่าความคลาดเคลื่อนประเภทที่ 1 ในโมเดล GRM ของดัชนี PARSCALE $G^2$ พบว่า ค่าความคลาดเคลื่อนประเภทที่ 1จะลดลงเมื่อจำนวนรายการคำตอบเปลี่ยนแปลงจาก 3 ไป 5 รายการเป็นจำนวนครั้งมากที่สุด และแทบไม่ลดลงเลยเมื่อจำนวนรายการคำตอบเปลี่ยนแปลงจาก 7 ไป 9 รายการ สอดคล้องกับดัชนี Generalized $S-\chi^2$ ซึ่งขัดแย้งกับผลการวิจัยของ Kang และ Chen (2010) ที่กล่าวว่าโดยส่วนใหญ่ในโมเดล GRM เมื่อมี 5 รายการคำตอบค่าความคลาดเคลื่อนประเภทที่ 1จะมากกว่าเมื่อมี 3 รายการคำตอบ ที่เป็นเช่นนี้อาจเป็นสาเหตุเนื่องมาจากการออกแบบการทดลองของผู้วิจัยที่มีความแตกต่างบางประการจากงานวิจัยของ Kang และ Chen (2010) เนื่องจากงานวิจัยของ Kang และ Chen (2010) มีการกระทำซ้ำมากถึง 100 ครั้ง ในขณะที่ค่าความคลาดเคลื่อนประเภทที่ 1 ในโมเดล GPCM ของดัชนี PARSCALE $G^2$ พบว่า ค่าความคลาดเคลื่อนประเภทที่ 1

จะลดลงเมื่อจำนวนรายการคำตอบเปลี่ยนแปลงจาก 3 ไป 5 รายการเป็นจำนวนครั้งมากที่สุด สอดคล้องกับดัชนี Generalized $S-\chi^2$ เช่นกัน

นอกจากนี้ เมื่อพิจารณาที่ความยาวแบบวัด พบว่าดัชนีทั้งสองชนิดให้ผลที่สอดคล้องกันว่า โดยส่วนใหญ่เมื่อความยาวแบบวัดเพิ่มขึ้น ค่าความคลาดเคลื่อนประเภทที่ 1 จะลดลง โดยเฉพาะเมื่อมีการเปลี่ยนแปลงความยาวแบบวัดจาก 10 ข้อไป 20 ข้อ ส่วนในการพิจารณาที่ขนาดกลุ่มตัวอย่างโดยใช้ค่าความคลาดเคลื่อนประเภทที่ 1 นั้นจะไม่มีรูปแบบการเปลี่ยนแปลงที่ชัดเจน แต่การเปลี่ยนแปลงของขนาดกลุ่มตัวอย่างจะส่งผลอย่างชัดเจนต่ออำนาจการทดสอบ ทั้งนี้อาจเนื่องจากอำนาจการทดสอบมีความสัมพันธ์กับขนาดกลุ่มตัวอย่างดังที่ได้กล่าวไปแล้ว

**ข้อเสนอแนะ**

ผู้วิจัยมีข้อเสนอแนะใน 2 ประเด็นหลัก คือ ข้อเสนอแนะในการนำไปใช้และข้อเสนอแนะสำหรับการวิจัยต่อไป ดังต่อไปนี้

1. ข้อเสนอแนะในการนำไปใช้

1.1 จากผลการวิจัย พบว่า โดยส่วนใหญ่ดัชนี Generalized $S-\chi^2$ มีประสิทธิภาพในการบ่งชี้ความสอดคล้องของข้อคำถามมากกว่า ดัชนี PARSCALE $G^2$ ดังนั้น ในการวิเคราะห์คุณภาพของแบบสอบหรือชุดของข้อคำถามเกี่ยวกับการบ่งชี้ความสอดคล้องของข้อคำถามจึงควรใช้ดัชนี Generalized $S-\chi^2$ ในการวิเคราะห์เกี่ยวกับประเด็นดังกล่าว ซึ่งผลการวิจัยนี้อยู่ภายใต้เงื่อนไขของ Kang และ Chen (2008) ที่ให้ความสำคัญกับแนวคิดว่า ดัชนีความสอดคล้องของข้อคำถามที่ดีนั้น ควรมีโอกาสในการบ่งชี้ข้อคำถามที่มีความสอดคล้อง (fit) กับโมเดลว่าเป็นข้อคำถามที่ไม่สอดคล้อง (misfit) กับโมเดลให้น้อยที่สุด หากแนวคิดในการตัดสินความสอดคล้องของข้อคำถามเปลี่ยนแปลงไป ก็อาจทำให้การตัดสินประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถามเปลี่ยนแปลงได้ เช่น ในการวิจัยนี้ หากพิจารณาโดยกำหนดเงื่อนไขว่าดัชนีความสอดคล้องของข้อคำถามที่ดี ควรมีคุณสมบัติในการบ่งชี้ข้อคำถามที่ไม่สอดคล้องกับโมเดล (misfit) ได้ดี โดยไม่สนใจคุณสมบัติเกี่ยวกับการบ่งชี้ข้อคำถามผิดพลาดในกรณีบ่งชี้ว่าข้อคำถามที่สอดคล้องกับโมเดล (fit) เป็นข้อคำถามที่ไม่สอดคล้องกับโมเดล (misfit) นั้น เกณฑ์ที่นำมาใช้ประเมินประสิทธิภาพของดัชนีความสอดคล้องของข้อคำถาม ก็จะพิจารณาแต่อำนาจการทดสอบเพียงอย่างเดียว ซึ่งดัชนี PARSCALE $G^2$ จะมีความเหมาะสมในการนำไปใช้ในกรณีดังกล่าวมากกว่า ดัชนี Generalized $S-\chi^2$ อย่างไรก็ตาม ข้อเสนอแนะดังกล่าวนี้อยู่ภายใต้

สถานการณ์ของข้อมูลที่จำลองขึ้นในการศึกษา ซึ่งอาจเป็นข้อจำกัดอย่างหนึ่งว่า หากนำไปทดลองวิเคราะห์กับข้อมูลจริง ผลการวิจัยอาจแตกต่างกับผลการวิจัยนี้ได้

นอกจากนี้ในการพิจารณากำหนดจำนวนรายการคำตอบในแบบวัดนั้น แม้ว่าจากผลการศึกษาวิจัยโดยการพิจารณาค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนีทั้งสองชนิดจะพบว่า โดยส่วนใหญ่ในโมเดล GRM ค่าความคลาดเคลื่อนประเภทที่ 1 จะลดลงเกือบทุกครั้งเมื่อจำนวนรายการคำตอบเปลี่ยนแปลงจาก 3 ไป 5 รายการคำตอบและจะคงที่หรืออาจเพิ่มขึ้นเมื่อมีการเปลี่ยนแปลงจาก 7 ไป 9 รายการคำตอบ ซึ่งอาจนำไปสู่ข้อสรุปที่ว่า “ในโมเดล GRM ควรพิจารณาเลือกใช้จำนวนรายการคำตอบ 5 และ 7 รายการคำตอบ” อย่างไรก็ตาม ในทางปฏิบัตินั้นอาจไม่สามารถสรุปเช่นนั้นได้ เนื่องจากต้องพิจารณาปัจจัยอื่นๆ ที่เกี่ยวข้องด้วย อาทิ คุณภาพของแบบวัดในแง่ของความตรงตามเนื้อหา (Content Validity) ความเที่ยง (Reliability) รวมทั้งความสามารถในการจำแนกคุณลักษณะผู้สอบของแบบวัดด้วย เช่นเดียวกันกับเมื่อพิจารณาค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนีทั้งสองชนิดในโมเดล GPCM ที่พบว่า โดยส่วนใหญ่ในโมเดล GPCM ค่าความคลาดเคลื่อนประเภทที่ 1 จะลดลงเกือบทุกครั้งเมื่อจำนวนรายการคำตอบเปลี่ยนแปลงจาก 3 ไป 5 รายการคำตอบแต่จะมีค่าเพิ่มขึ้นเมื่อมีการเปลี่ยนแปลงจาก 5 ไป 7 รายการคำตอบ และจะเพิ่มขึ้นอีกเมื่อมีการเปลี่ยนแปลงจาก 7 ไป 9 รายการคำตอบ ซึ่งอาจนำไปสู่ข้อสรุปที่ว่า “ในโมเดล GPCM ควรพิจารณาเลือกใช้จำนวนรายการคำตอบ 3 และ 5 รายการคำตอบ” อย่างไรก็ตาม ในทางปฏิบัตินั้น อาจไม่สามารถสรุปเช่นนั้นได้ เนื่องจากต้องพิจารณาปัจจัยอื่นๆ ที่เกี่ยวข้อง อาทิ คุณภาพของแบบวัดในแง่ของความตรงตามเนื้อหา (Content Validity) ความเที่ยง (Reliability) รวมทั้งความสามารถในการจำแนกคุณลักษณะผู้สอบของแบบวัด รวมทั้งปัจจัยอื่นๆ ที่เกี่ยวข้องในบริบทของการเก็บรวบรวมข้อมูลด้วย เช่น ถึงแม้ว่าการใช้ 7 รายการคำตอบจะให้ค่าความคลาดเคลื่อนประเภทที่ 1 น้อยกว่า 5 รายการคำตอบ แต่ในทางปฏิบัติ อาจใช้เวลาทำแบบวัดนาน ทำให้ผู้ตอบข้อคำถามเบื่อหน่ายในการตอบ ส่งผลถึงความตรงของผลการวัดจากแบบวัดนั้นได้

จากผลการวิจัยที่พบว่า ทั้งในโมเดล GRM และ GPCM ค่าความคลาดเคลื่อนประเภทที่ 1 จะลดลง เมื่อขนาดความยาวแบบวัดเพิ่มขึ้น โดยเฉพาะเมื่อมีการเปลี่ยนแปลงความยาวแบบวัดจาก 10 ข้อไป 20 ข้อนั้น ดังนั้นจึงควรเลือกใช้ความยาวแบบวัด 20 ข้อในการสร้างแบบวัด ข้อสรุปดังกล่าวนี้อาจไม่สามารถนำไปใช้ในสถานการณ์จริงได้ เพราะในทางปฏิบัติ สิ่งที่สำคัญมากอย่างหนึ่งและไม่อาจละเลยได้ในการสร้างหรือพัฒนาแบบวัดก็คือ ความตรง โดยเฉพาะความตรงตามเนื้อหา (Content Validity) การที่จะกำหนดความยาวแบบวัด หรือจำนวน

รายการคำตอบในแบบวัด จึงต้องพิจารณาความตรงเป็นหลัก ว่าขนาดความยาวแบบวัด หรือจำนวนรายการคำตอบเท่าไร จึงจะทำให้แบบวัดมีความตรง ความเที่ยง และความสามารถในการจำแนกคุณลักษณะของผู้ตอบแบบวัดได้มากที่สุด

1.2 ในปัจจุบัน IRTFIT macros ของ Bjorner, Smith, Stone และ Sun (2007) ซึ่งประมวลผลบนโปรแกรม SAS นั้น ไม่สามารถประมาณค่าพารามิเตอร์ข้อคำถามได้เอง จำเป็นต้องใช้ค่าพารามิเตอร์ข้อคำถามจากโปรแกรม PARSCALE ซึ่งทำให้การคำนวณดัชนี Generalized $S-\chi^2$ อาจไม่ได้รับความสะดวกนักเมื่อเทียบกับดัชนี PARSCALE $G^2$ เนื่องจากต้องมีการจัดเตรียมไฟล์ข้อมูลพารามิเตอร์ข้อคำถามตามรูปแบบที่คู่มือการใช้โปรแกรมกำหนด เพื่อให้ IRTFIT macros สามารถอ่านค่าพารามิเตอร์จากโปรแกรม PARSCALE และนำไปคำนวณดัชนี Generalized $S-\chi^2$ ได้ ดังนั้น ในอนาคตผู้พัฒนาโปรแกรมควรพัฒนาให้ IRTFIT macros สามารถประมาณค่าพารามิเตอร์ข้อคำถามและคำนวณดัชนี Generalized $S-\chi^2$ ได้ในขั้นตอนเดียวโดยไม่ต้องใช้ค่าพารามิเตอร์ข้อคำถามจากโปรแกรมอื่น หรืออาจนำดัชนี Generalized $S-\chi^2$ ไปบรรจุไว้ในโปรแกรมที่ใช้ในการวิเคราะห์ข้อสอบอื่นๆ อาทิ โปรแกรม PARSCALE เพื่อให้ดัชนี Generalized $S-\chi^2$ ถูกนำไปใช้อย่างแพร่หลายมากขึ้น

2. ข้อเสนอแนะสำหรับการวิจัยต่อไป

2.1 ในการวิจัยครั้งนี้ ในแต่ละสถานการณ์ที่ทำการศึกษามีการกระทำซ้ำเพียง 30 ครั้ง ซึ่งเป็นไปตามผลการศึกษาของ Harwell, Hsu และ Kirisci (1996 อ้างถึงใน พัชรี จันทร์เพ็ง, 2550) พบว่าถ้าศึกษาโดยใช้โมเดลทฤษฎีการตอบสนองข้อสอบ (Item Response Theory Model) เป็นฐานควรมีการทำซ้ำอย่างน้อย 20-25 ครั้งเพื่อให้การประมาณค่ามีความเที่ยงมากยิ่งขึ้น ซึ่งจากผลการวิจัยที่ค่าความคลาดเคลื่อนประเภทที่ 1 และค่าอำนาจการทดสอบในบางสถานการณ์ ที่มีการเปลี่ยนแปลงของ ความยาวแบบวัด ขนาดกลุ่มตัวอย่าง และจำนวนรายการคำตอบแล้ว แต่ค่าความคลาดเคลื่อนประเภทที่ 1 หรือค่าอำนาจการทดสอบก็ยังไม่เปลี่ยนแปลง รวมทั้งกรณีที่ค่าความคลาดเคลื่อนประเภทที่ 1 ของดัชนีทั้งสองมีค่าที่แตกต่างกันเพียงเล็กน้อยเท่านั้น ในกรณีต่างๆ เหล่านี้ เป็นที่น่าสนใจว่าหากมีการกระทำซ้ำในแต่ละสถานการณ์ที่เพิ่มมากขึ้นจากการวิจัยครั้งนี้ ผลการวิจัยจะยังคงเป็นเช่นเดิมหรือไม่ ดังนั้น ในการศึกษาวิจัยครั้งต่อไป จึงควรเพิ่มจำนวนรอบในการกระทำซ้ำให้มากขึ้น เพื่อให้ผลการศึกษามีความถูกต้องมากที่สุด

2.2 ในการวิจัยครั้งนี้ จำลองข้อมูลที่ใช้ในการศึกษาวิจัยด้วยโปรแกรม WINGEN ซึ่งกำหนดให้มีลักษณะเป็นไปตามโมเดลที่ใช้ในการศึกษาวิจัยทั้ง 2 โมเดลคือ Grade Response Model (GRM) และ Generalized Partial Credit Model (GPCM) แม้ว่าการวิจัยนี้จะได้มีการ

ตรวจสอบคุณลักษณะของข้อมูลที่ได้จำลองขึ้นว่าเป็นไปตามเงื่อนไขที่กำหนดหรือไม่ (ค่าพารามิเตอร์ความสามารถของผู้สอบ ($\theta$) มีการแจกแจงแบบปกติ และค่าพารามิเตอร์ข้อคำถาม a มีการแจกแจงแบบ Lognormal ค่าพารามิเตอร์ข้อคำถาม b มีการแจกแจงแบบ Uniform) แต่การตรวจสอบดังกล่าวนี้ ยังไม่สามารถยืนยันได้ว่า ข้อมูลที่จำลองขึ้นนั้นมีลักษณะสอดคล้องเป็นไปตามคุณลักษณะของ Grade Response Model (GRM) และ Generalized Partial Credit Model (GPCM) ดังนั้น ในการวิจัยครั้งต่อไป ผู้วิจัยจึงควรมีการตรวจสอบข้อมูลที่ได้จำลองขึ้นว่ามีลักษณะสอดคล้องเป็นไปตามคุณลักษณะของ Grade Response Model (GRM) และ Generalized Partial Credit Model (GPCM) หรือไม่ด้วย ซึ่งจะช่วยลดข้อจำกัดของผลการวิจัย ทำให้ผลการวิจัยมีความน่าเชื่อถือมากยิ่งขึ้น

2.3 ในการวิจัยนี้ มีการตรวจสอบคุณลักษณะของข้อมูลที่ได้จำลองขึ้นว่าเป็นไปตามเงื่อนไขที่กำหนดหรือไม่ โดยในการตรวจสอบค่าพารามิเตอร์ความสามารถของผู้สอบ ($\theta$) ว่ามีการแจกแจงแบบปกติหรือไม่ ได้ดำเนินการโดยใช้การพิจารณาการแจกแจงของค่าพารามิเตอร์ความสามารถผู้สอบ ($\theta$) จากรูปหลายเหลี่ยมความถี่ (Histogram) ซึ่งพบว่าในบางสถานการณ์เมื่อพิจารณาจากรูปหลายเหลี่ยมความถี่ยังไม่อาจมั่นใจได้ว่าค่าพารามิเตอร์ความสามารถผู้สอบ ($\theta$) มีการแจกแจงแบบปกติ ดังนั้น ในการวิจัยครั้งต่อไป ควรใช้การตรวจสอบด้วยค่าสถิติสอบจะมีความเหมาะสมมากกว่าการใช้รูปหลายเหลี่ยมความถี่ (Histogram)

2.4 ในการศึกษาวิจัยครั้งนี้ไม่ได้นำเสนอผลการวิจัยในลักษณะว่า ดัชนีทั้งสองชนิดนี้ให้ค่าความสอดคล้องของข้อคำถามร่วมกันร้อยละเท่าใด แตกต่างกันร้อยละเท่าใดในแต่ละสถานการณ์ที่ทำการศึกษา ซึ่งการนำเสนอในลักษณะนี้จะต้องมีการวิเคราะห์เพื่อเปรียบเทียบดัชนีทั้งสองชนิดเป็นรายข้อในแต่ละสถานการณ์ที่ทำการศึกษา ซึ่งอาจทำให้ขั้นตอนการวิเคราะห์มีความซับซ้อนขึ้น แต่ถ้าจะมีการวิจัยในประเด็นที่เกี่ยวข้องกับการวิจัยในลักษณะเดียวกันกับงานวิจัยนี้ก็ควรดำเนินการในประเด็นนี้ด้วย เพื่อให้ผลการวิจัยสามารถตอบข้อสงสัยหรือประเด็นที่ยังไม่ชัดเจนได้อย่างถูกต้อง ชัดเจน เข้าใจได้อย่างกว้างขวาง

# รายการอ้างอิง

**ภาษาไทย**


กุสุมา สุวรรณแก้ว. *การเปรียบเทียบผลการตรวจสอบความเหมาะสมของบุคคลระหว่างดัชนีแอลแซดกับดัชนีดัลเบิ้ลยูวัน ตามทฤษฎีการตอบสนองข้อสอบ*. วิทยานิพนธ์ปริญญามหาบัณฑิต, ภาควิชาวิจัยการศึกษา บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย, 2540.

ปุณยนุช พินชู. *การเปรียบเทียบค่าความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของวิธีการเปรียบเทียบค่าเฉลี่ยรายคู่ สำหรับแผนการทดลองแบบสุ่มสมบูรณ์*. วิทยานิพนธ์ปริญญามหาบัณฑิต, สาขาวิชาสถิติการศึกษา ภาควิชาวิจัยและจิตวิทยาการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย, 2548.

พัชรี จันทร์เพ็ง. *การเปรียบเทียบคุณภาพของวิธีการเชื่อมโยงคะแนนตามทฤษฎีการตอบสนองข้อสอบแบบพหุมิติภายใต้การหมุนแกน โครงสร้างเชิงมิติและระดับความสัมพันธ์ที่แตกต่างกัน*. วิทยานิพนธ์ปริญญาดุษฎีบัณฑิต, สาขาวิชาการวัดและประเมินผลการศึกษา ภาควิชาวิจัยและจิตวิทยาการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย, 2550.

รัชกมล กบิลจิตต์. *ความน่าจะเป็น ทฤษฎีและการประยุกต์*. นครปฐม: โรงพิมพ์มหาวิทยาลัยศิลปากร, 2545.

วลีมาศ แซ่อึ้ง. *การเปรียบเทียบอำนาจการทดสอบและอัตราความคลาดเคลื่อนประเภทที่1 ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูประหว่างวิธีซิปเทสท์ปรับใหม่ วิธีซิปเทสท์ วิธีแมนเทล-แฮนส์เซล และวิธีการถดถอยโลจิสติก*. วิทยานิพนธ์ปริญญาดุษฎีบัณฑิต, สาขาวิชาการวัดและประเมินผลการศึกษา ภาควิชาวิจัยการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย, 2543.

ศิริชัย กาญจนวาสี. *ทฤษฎีการทดสอบแนวใหม่ (Modern Test Theory)*. กรุงเทพมหานคร: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย, 2550.

สุชาดา บวรกิติวงศ์. *สถิติประยุกต์ทางพฤติกรรมศาสตร์*. กรุงเทพมหานคร: สำนักพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย, 2548.

สุชาดา บวรกิติวงศ์. ทำไม $\alpha$ ต้อง .05 ?. *วารสารวิธีวิทยาการวิจัย*. 11 (2540): 13-20.

เอมอร จังศิริพรปกรณ์. *การเปรียบเทียบคุณภาพของแบบสอบเลือกตอบเมื่อตรวจด้วยวิธีการให้คะแนนความรู้บางส่วนกับวิธีประเพณีนิยม*. กรุงเทพมหานคร: งานวิจัยกองทุนรัชดาภิเษกสมโภช จุฬาลงกรณ์มหาวิทยาลัย, 2545.

**ภาษาอังกฤษ**

Aydin, C. Y., and Uzuntiryaki, E. Development and psychometric evaluation of the high school chemistry self-efficacy scale. *Educational and Psychological Measurement,* 69 (2009), 868-880.

Bjorner, J. B., Smith, K. J., Stone, C., and Sun, X. *IRTFIT: A macro for item fit and local dependence tests under IRT model* [computer software and manual]. 2007. Available from : http://outcomes.cancer.gov/areas/measurement/irt_model _fit.html. [June 2010]

DeMars, C. Type I error rate for PARSCALE's fit index. *Educational and Psychological Measurement,* 65(2005), 42-50.

Dodeen, H. The relationship between item parameter and item fit. *Journal of educational measurement,* 41(2004), 261-270.

Du Toit, M. *IRT form SSI: BILOG-MG MULTILOG PARSCALE TESTFACT* [computer software and manual]. Mooresville, IN: Scientific Software International, 2003.

Embretson, S. E., and Reise, S. P. *Item response theory for psychologists.* New Jersey: Lawrence Erlbaum Associates Publisher, 2000.

Hambleton, R. K., Swaminathan, H., and Roger, H. J. *Fundamentals of Item Response Theory* (1st ed.,). California : SAGE Publication,1991.

Hambleton, R. K., and Swaminathan, H. *Item response theory principle and application*. Norwell: Kluwer Academic Publishers, 1996.

Han, K. T., and Hambleton, R. K. *User's manual for WinGen: Windows software that generates IRT model parameters and item response* [computer software and manual]. 2007. Available form: http://www.hantest.net/wingen. [July, 2010]

Kang, T. and Chen, T. T. Performance of the Generalized $S - \chi^2$ item fit index for polytomous IRT models. *Journal of educational measurement,* 45 (2008), 391-406.

Kang, T., and Chen, T. T. Performance of the generalized $S - \chi^2$ item fit index for graded response model. *Asia Pacific Educ.Rev,* 12 (2010), 89-96.

Liang, T., and Wells, C. S. A model fit statistic for generalized partial credit model. *Educational and Psychological Measurement,* 69 (2009), 913-928.

LaHuis, D. M., Clark, P., and O'Brien, E. An examination of item response theory item fit indices for grade response model. *Organizational Research Method, 14 (2009),* 10-23.

Mun Ng, K., Wang, C., Kim, D., and Bodenhorn, N. Factor structure analysis of the schutte self-report emotional intelligence scale on international students. *Educational and Psychological Measurement* [Online]. 2009. Available form: http://epm.sagepub.com/content/70/4/695. [December, 2009]

Orlando, M., and Thissen, D. Likelihood-based item fit indices for dichotomous item response theory models. *Applied psychological measurement,* 24 (2000), 50-64.

Orlando, M., and Thissen, D. Further investigation of the performance of $S - \chi^2$: An item fit index for use with dichotomous item response theory models. *Applied psychological measurement,* 27 (2003), 289-298.

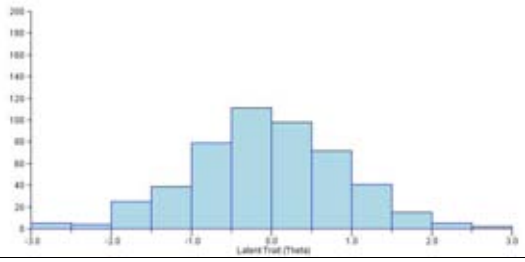
Stone, C. A., and Zhang, B. Assessing goodness of fit of item response theory models : A comparison of traditional and alternative procedure. *Journal of Educational Measurement,* 40 (2003), 331-352.

Wright, B. D., and Stone, M. H. *Best test design*. Chicago: MESA Press, 1979.

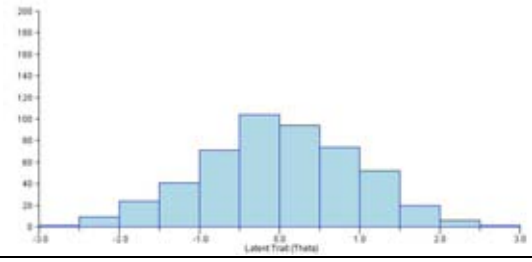
# ภาคผนวก

**ภาคผนวก ก รูปแสดงการแจกแจงของค่าพารามิเตอร์ความสามารถผู้สอบ ในแต่ละสถานการณ์ที่ทำการศึกษา**

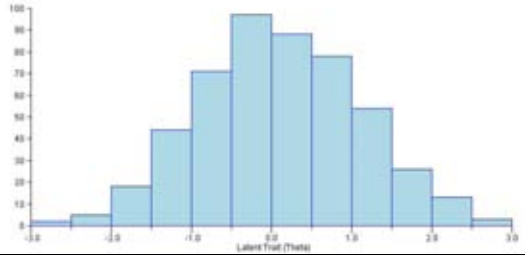
สถานการณ์ที่ 1



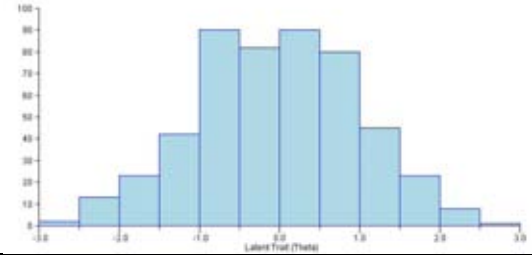
สถานการณ์ที่ 2



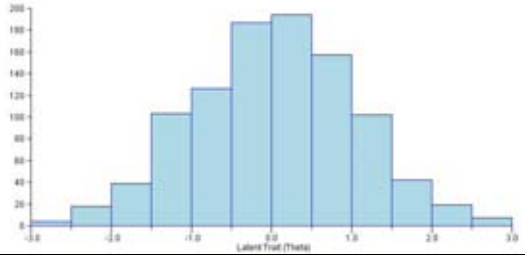
สถานการณ์ที่ 3



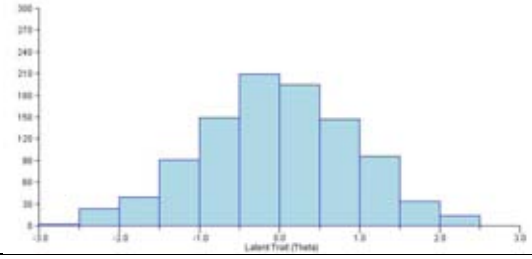
สถานการณ์ที่ 4



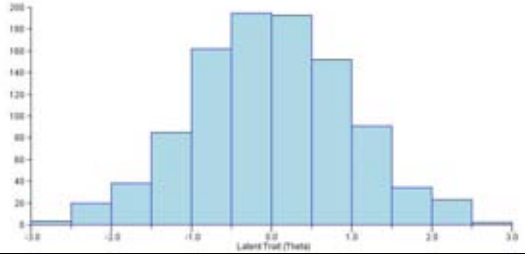
สถานการณ์ที่ 5



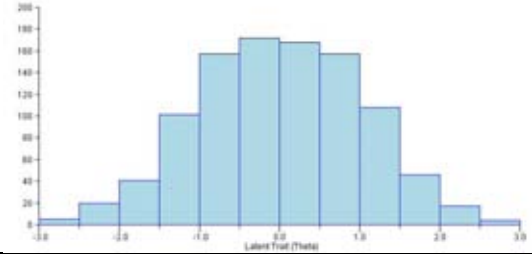
สถานการณ์ที่ 6



สถานการณ์ที่ 7

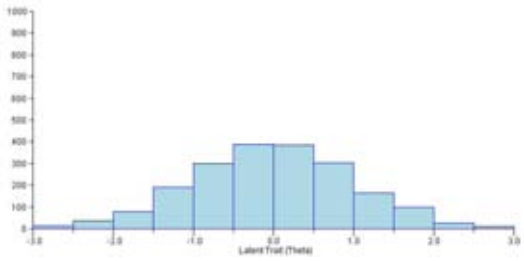


สถานการณ์ที่ 8

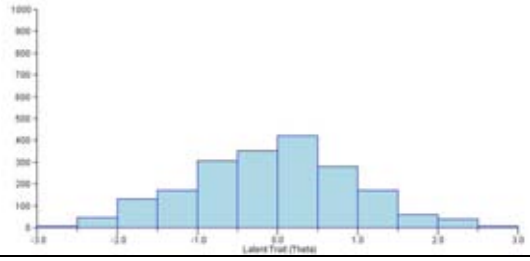


**ภาคผนวก ก รูปแสดงการแจกแจงของค่าพารามิเตอร์ความสามารถผู้สอบ**
**ในแต่ละสถานการณ์ที่ทำการศึกษา**

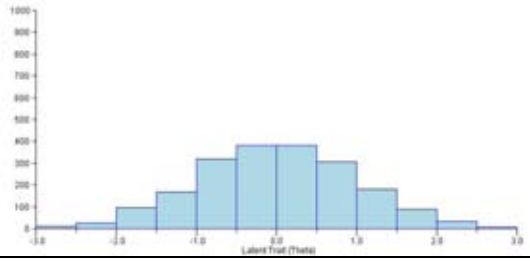
สถานการณ์ที่ 9



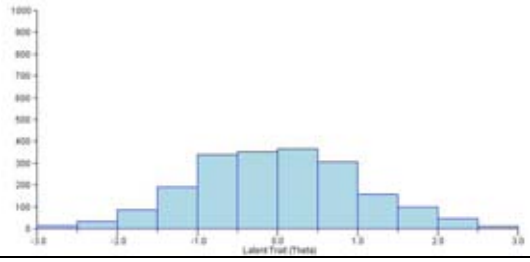
สถานการณ์ที่ 10



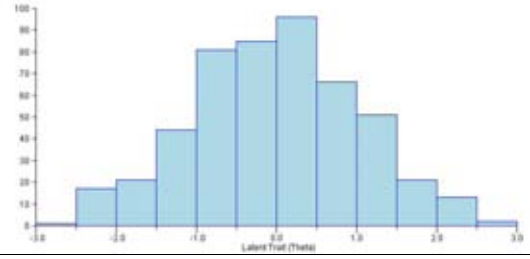
สถานการณ์ที่ 11



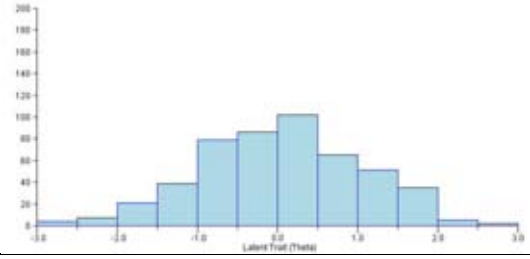
สถานการณ์ที่ 12



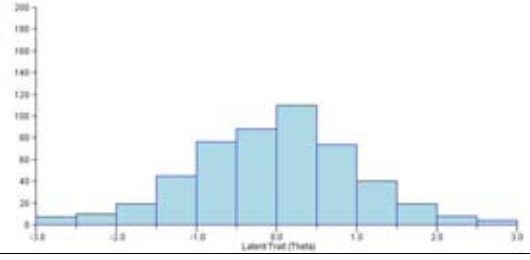
สถานการณ์ที่ 13



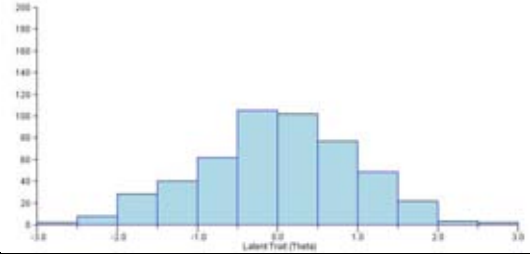
สถานการณ์ที่ 14



สถานการณ์ที่ 15

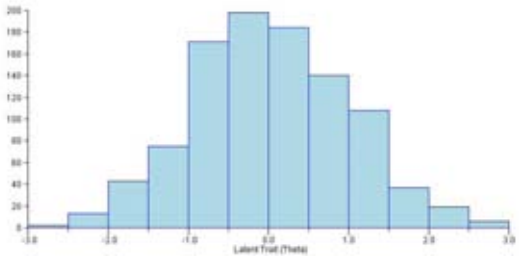


สถานการณ์ที่ 16

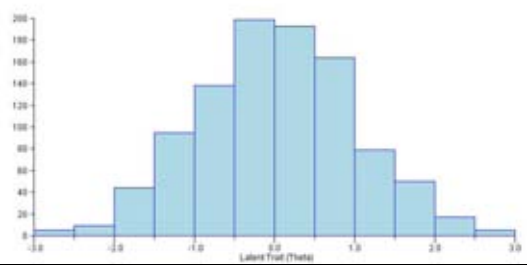


**ภาคผนวก ก รูปแสดงการแจกแจงของค่าพารามิเตอร์ความสามารถผู้สอบ ในแต่ละสถานการณ์ที่ทำการศึกษา**

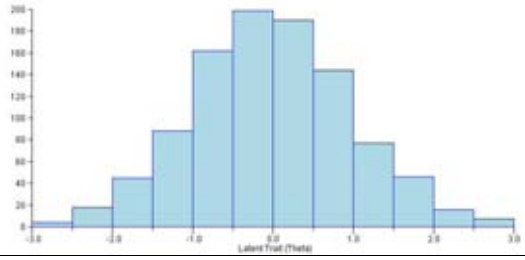
สถานการณ์ที่ 17



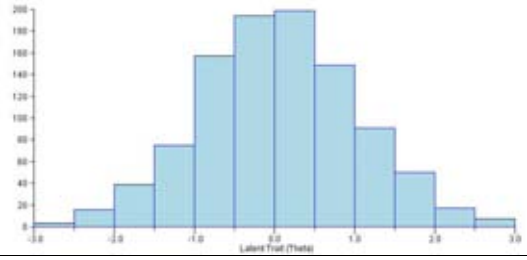
สถานการณ์ที่ 18



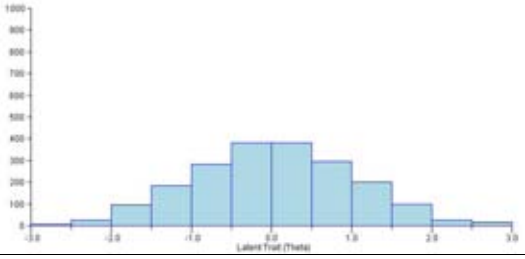
สถานการณ์ที่ 19



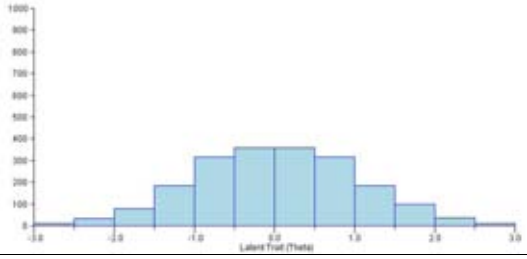
สถานการณ์ที่ 20



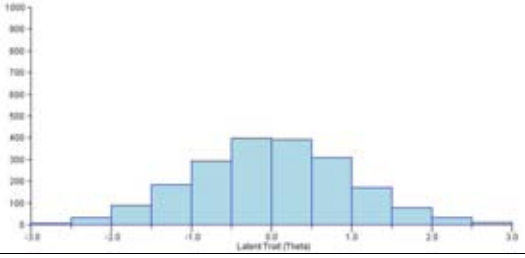
สถานการณ์ที่ 21



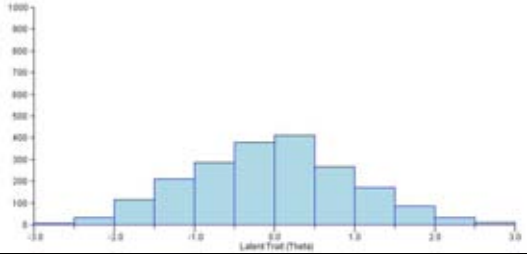
สถานการณ์ที่ 22



สถานการณ์ที่ 23

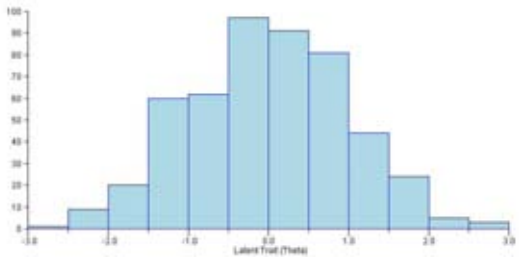


สถานการณ์ที่ 24

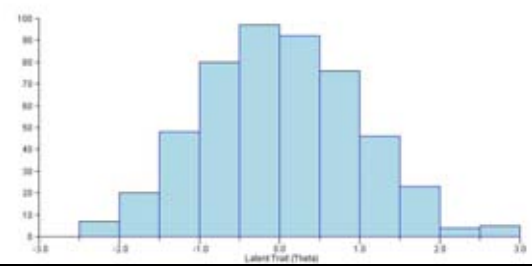


**ภาคผนวก ก รูปแสดงการแจกแจงของค่าพารามิเตอร์ความสามารถผู้สอบ**
**ในแต่ละสถานการณ์ที่ทำการศึกษา**

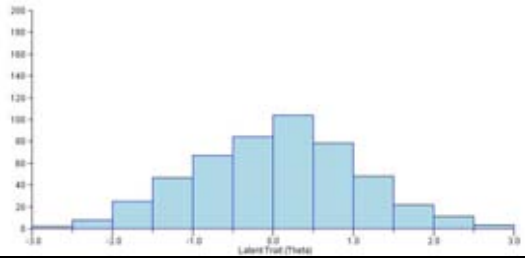
สถานการณ์ที่ 25



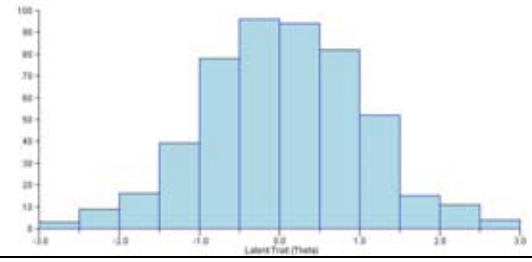
สถานการณ์ที่ 26



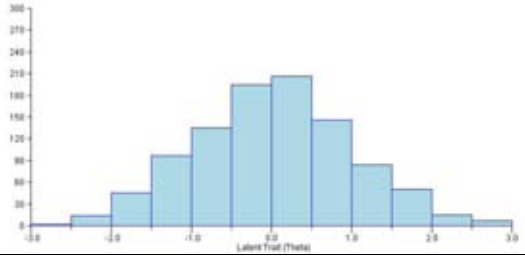
สถานการณ์ที่ 27



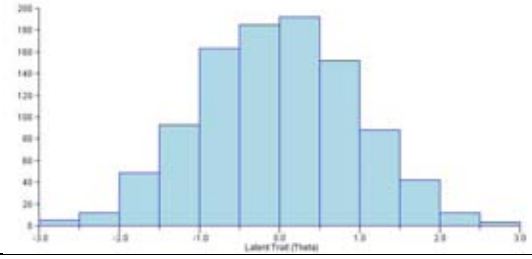
สถานการณ์ที่ 28



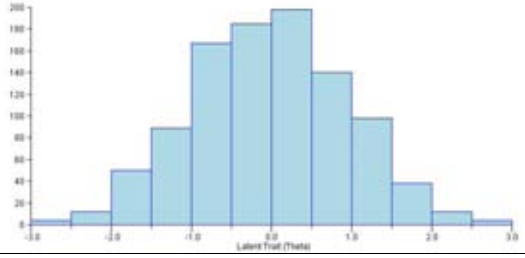
สถานการณ์ที่ 29



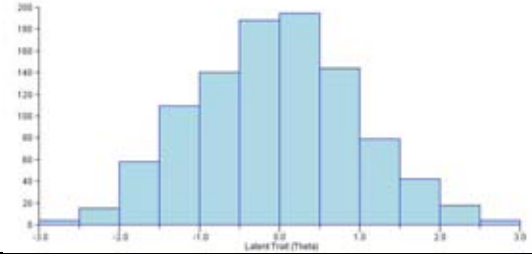
สถานการณ์ที่ 30



สถานการณ์ที่ 31

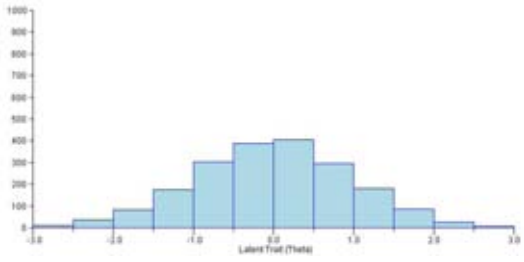


สถานการณ์ที่ 32

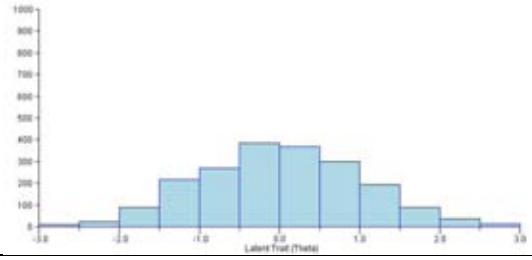


**ภาคผนวก ก รูปแสดงการแจกแจงของค่าพารามิเตอร์ความสามารถผู้สอบ**
**ในแต่ละสถานการณ์ที่ทำการศึกษา**

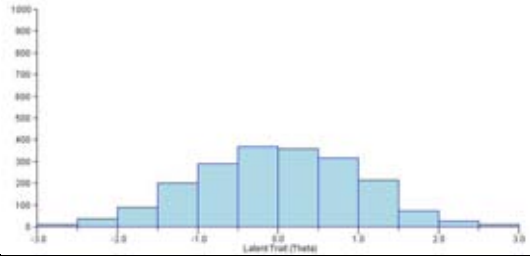
**สถานการณ์ที่** 33



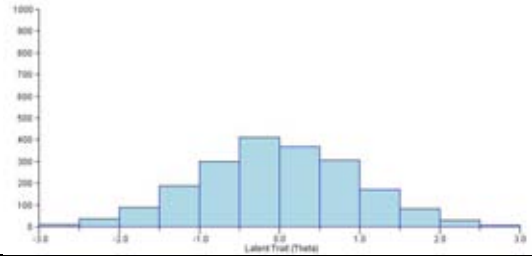
**สถานการณ์ที่** 34



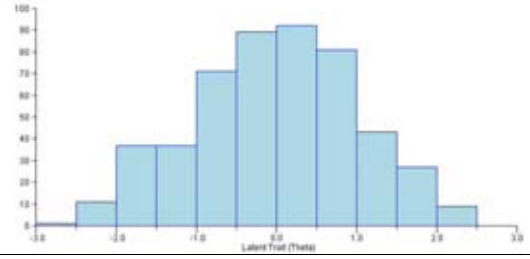
**สถานการณ์ที่** 35



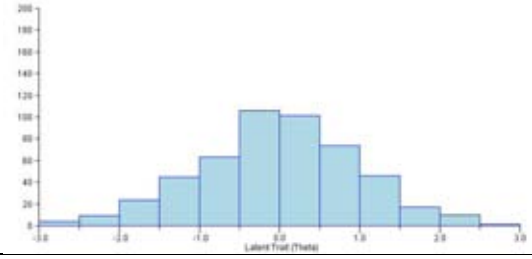
**สถานการณ์ที่** 36



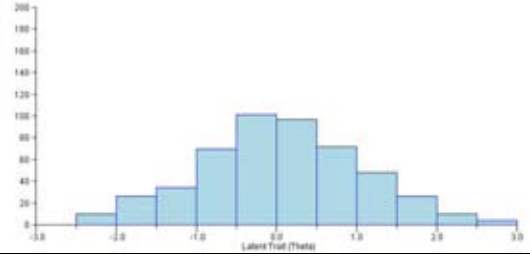
**สถานการณ์ที่** 37



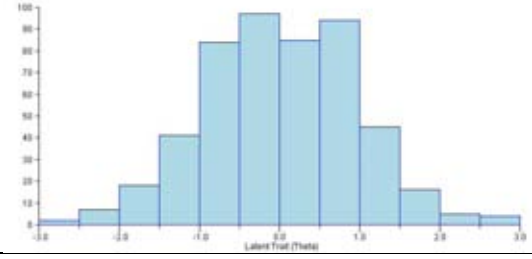
**สถานการณ์ที่** 38



**สถานการณ์ที่** 39

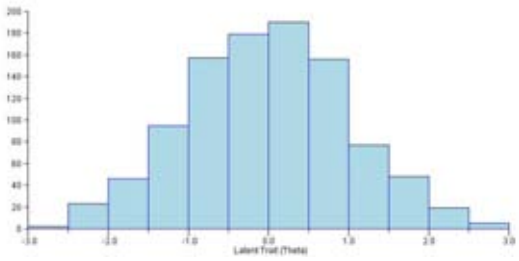


**สถานการณ์ที่** 40

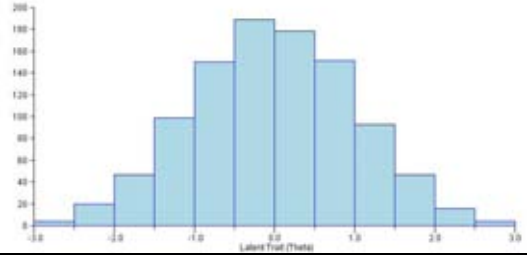


**ภาคผนวก ก รูปแสดงการแจกแจงของค่าพารามิเตอร์ความสามารถผู้สอบ**
**ในแต่ละสถานการณ์ที่ทำการศึกษา**

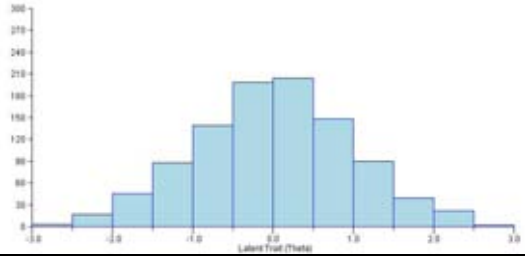
สถานการณ์ที่ 41



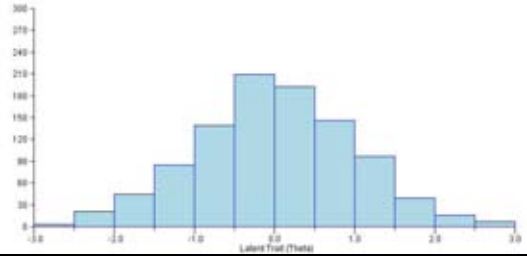
สถานการณ์ที่ 42



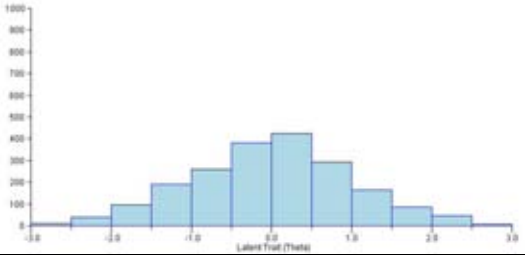
สถานการณ์ที่ 43



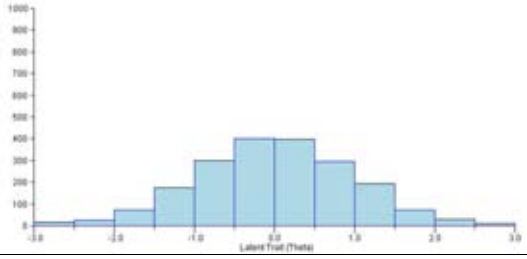
สถานการณ์ที่ 44



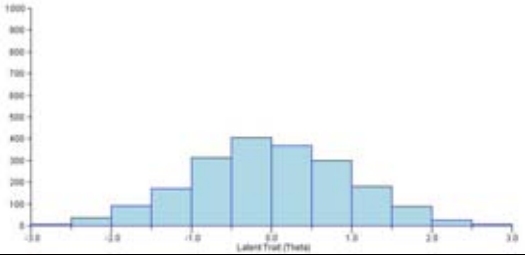
สถานการณ์ที่ 45



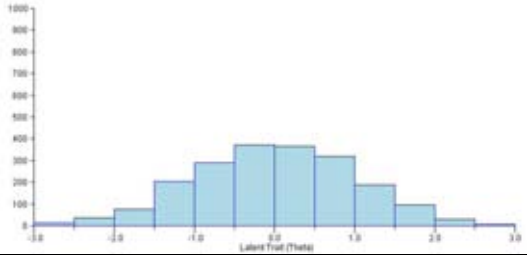
สถานการณ์ที่ 46



สถานการณ์ที่ 47

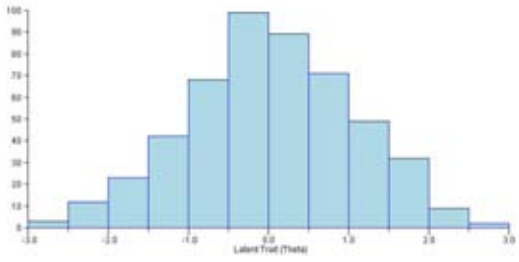


สถานการณ์ที่ 48

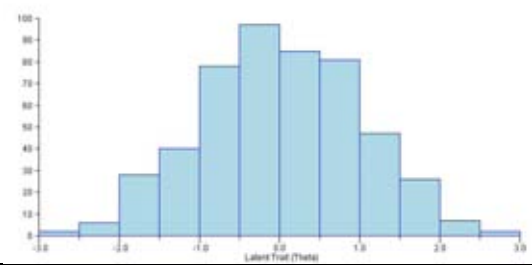


**ภาคผนวก ก รูปแสดงการแจกแจงของค่าพารามิเตอร์ความสามารถผู้สอบ**
**ในแต่ละสถานการณ์ที่ทำการศึกษา**

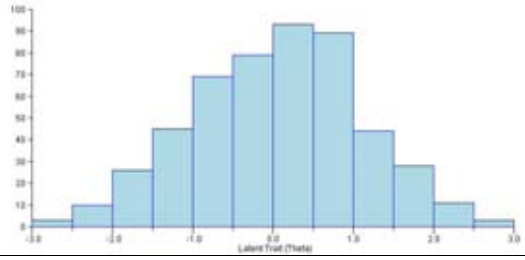
**สถานการณ์ที่** 49



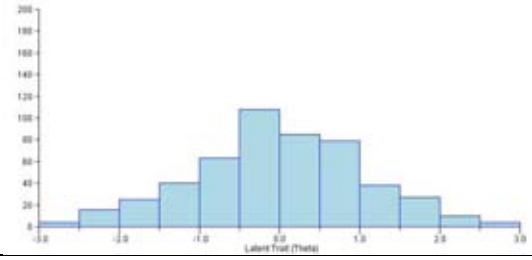
**สถานการณ์ที่** 50



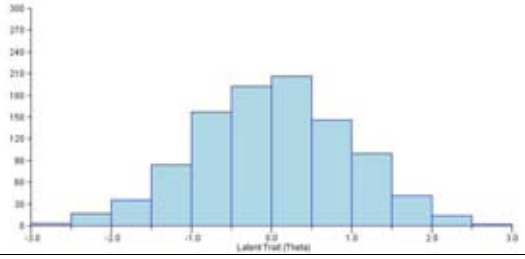
**สถานการณ์ที่** 51



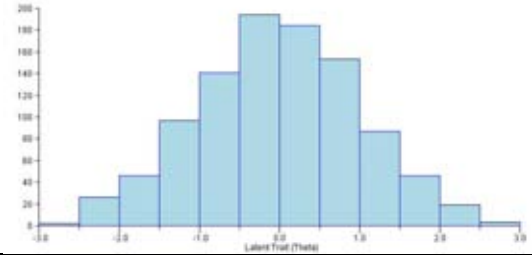
**สถานการณ์ที่** 52



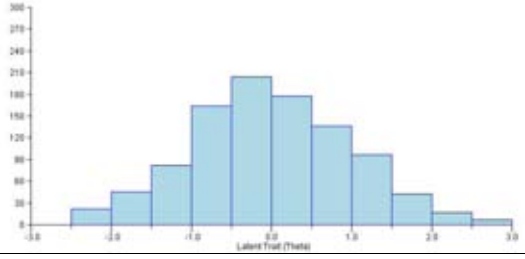
**สถานการณ์ที่** 53



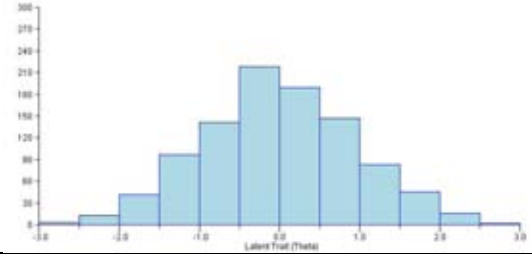
**สถานการณ์ที่** 54



**สถานการณ์ที่** 55

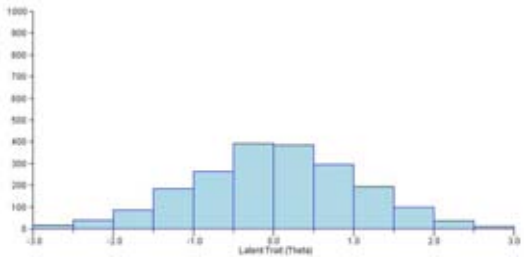


**สถานการณ์ที่** 56

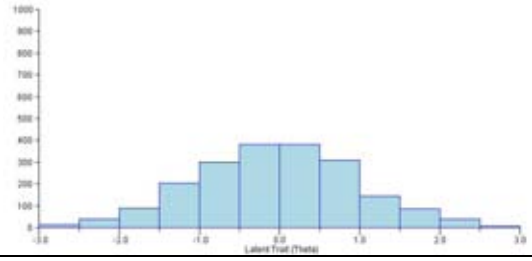


**ภาคผนวก ก รูปแสดงการแจกแจงของค่าพารามิเตอร์ความสามารถผู้สอบ**
**ในแต่ละสถานการณ์ที่ทำการศึกษา**

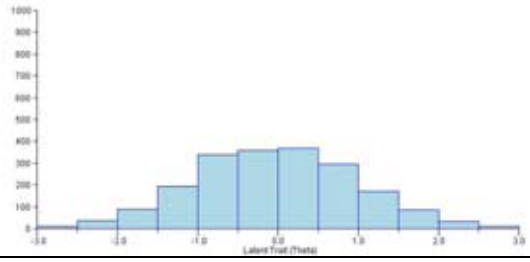
สถานการณ์ที่ 57



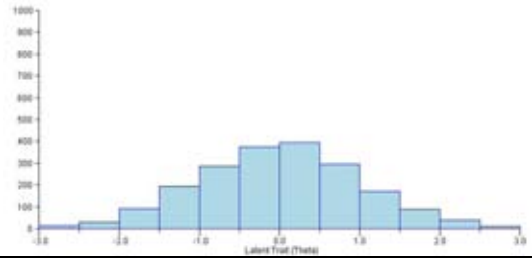
สถานการณ์ที่ 58



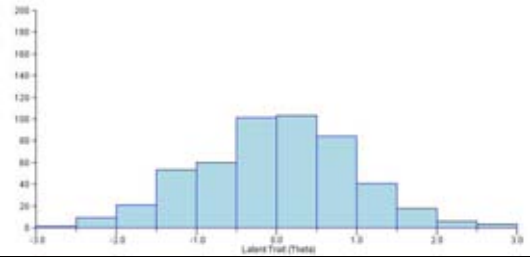
สถานการณ์ที่ 59



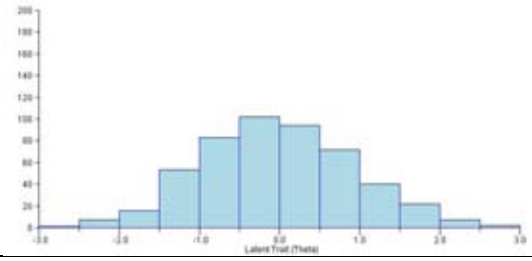
สถานการณ์ที่ 60



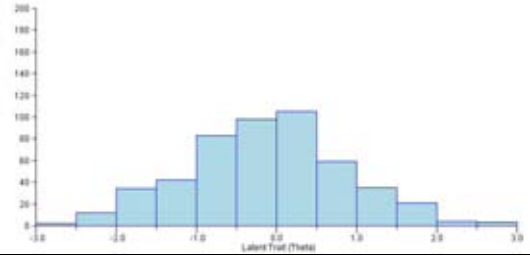
สถานการณ์ที่ 61



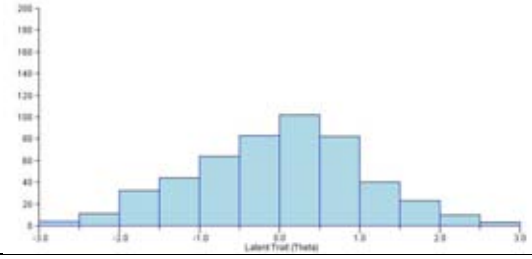
สถานการณ์ที่ 62



สถานการณ์ที่ 63

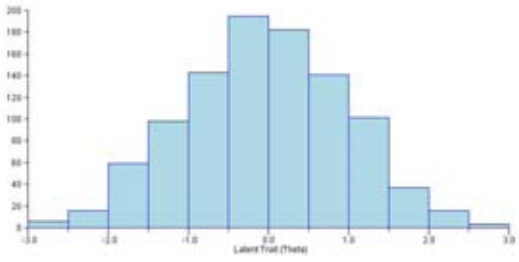


สถานการณ์ที่ 64

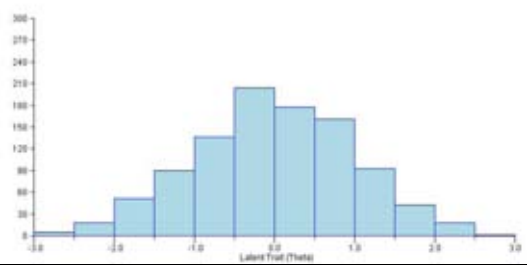


**ภาคผนวก ก รูปแสดงการแจกแจงของค่าพารามิเตอร์ความสามารถผู้สอบ**
**ในแต่ละสถานการณ์ที่ทำการศึกษา**

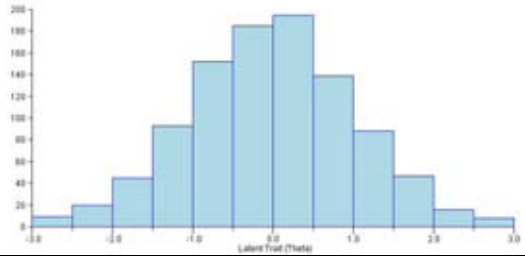
สถานการณ์ที่ 65



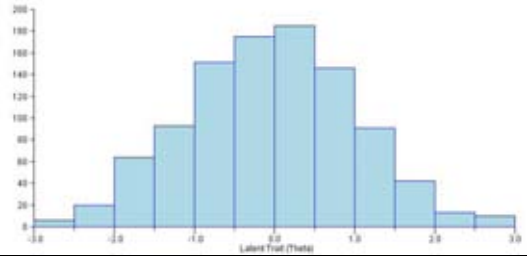
สถานการณ์ที่ 66



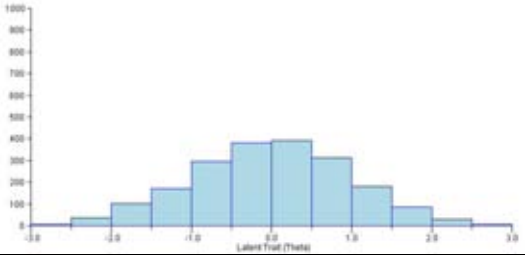
สถานการณ์ที่ 67



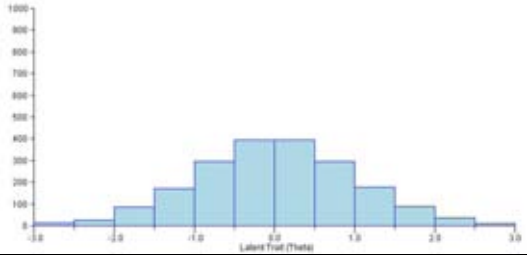
สถานการณ์ที่ 68



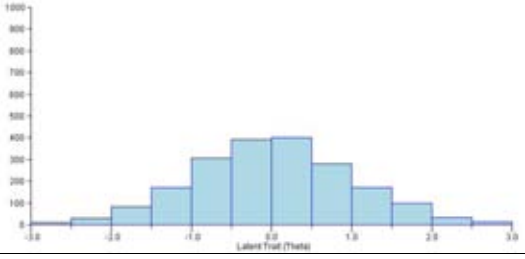
สถานการณ์ที่ 69



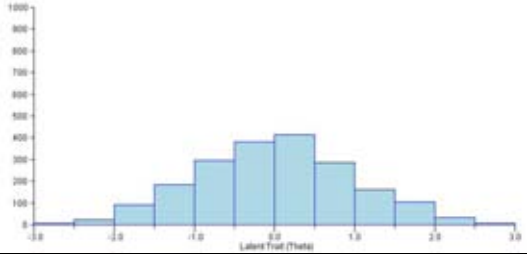
สถานการณ์ที่ 70



สถานการณ์ที่ 71



สถานการณ์ที่ 72



**ภาคผนวก ข ค่าสถิติพื้นฐานของค่าพารามิเตอร์ความสามารถผู้สอบ และค่าพารามิเตอร์ข้อคำถาม**

| สถานการณ์ที่ | $\theta$ | | a | | b | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD |
| 1 | -0.073 | 0.967 | 1.392 | 1.293 | -0.338 | 0.895 |
| 2 | 0.006 | 0.988 | 1.147 | 0.460 | -0.632 | 0.829 |
| 3 | 0.097 | 1.001 | 1.137 | 0.382 | -0.499 | 0.877 |
| 4 | -0.022 | 1.008 | 0.942 | 0.344 | 0.579 | 0.860 |
| 5 | 0.050 | 1.011 | 1.195 | 0.632 | -0.647 | 1.000 |
| 6 | -0.044 | 0.955 | 0.878 | 0.332 | 0.422 | 0.852 |
| 7 | 0.012 | 0.981 | 1.081 | 0.346 | -0.722 | 0.901 |
| 8 | -0.003 | 1.041 | 0.914 | 0.302 | -0.466 | 0.938 |
| 9 | -0.013 | 1.001 | 1.061 | 0.536 | -0.513 | 0.734 |
| 10 | -0.078 | 1.008 | 1.067 | 0.756 | -0.737 | 0.769 |
| 11 | 0.000 | 0.994 | 0.940 | 0.444 | -0.511 | 0.950 |
| 12 | -0.015 | 1.026 | 1.191 | 0.338 | -0.395 | 0.838 |
| 13 | -0.024 | 1.054 | 1.146 | 0.551 | -0.355 | 0.918 |
| 14 | 0.053 | 1.029 | 1.199 | 0.632 | -0.549 | 0.881 |
| 15 | -0.033 | 1.013 | 1.120 | 0.639 | -0.587 | 0.857 |
| 16 | 0.000 | 0.952 | 1.190 | 0.546 | -0.523 | 0.838 |
| 17 | 0.031 | 0.987 | 1.077 | 0.458 | -0.280 | 0.935 |
| 18 | 0.015 | 0.892 | 1.122 | 0.464 | -0.542 | 0.846 |
| 19 | -0.013 | 1.004 | 1.280 | 0.724 | -0.588 | 0.874 |
| 20 | 0.049 | 0.987 | 1.120 | 0.580 | -0.504 | 0.916 |
| 21 | 0.033 | 1.007 | 1.113 | 0.484 | -0.352 | 0.889 |
| 22 | 0.030 | 1.019 | 0.955 | 0.324 | -0.527 | 0.846 |
| 23 | -0.003 | 0.981 | 1.097 | 0.533 | -0.634 | 0.807 |
| 24 | -0.050 | 1.005 | 1.199 | 0.531 | -0.552 | 0.893 |
| 25 | 0.019 | 1.019 | 1.247 | 0.711 | -0.275 | 0.894 |
| 26 | 0.007 | 0.995 | 1.208 | 0.617 | -0.514 | 0.856 |
| 27 | 0.031 | 1.018 | 1.268 | 0.689 | -0.331 | 0.839 |
| 28 | 0.050 | 0.999 | 1.148 | 0.691 | -0.468 | 0.689 |
| 29 | 0.013 | 1.004 | 1.129 | 0.568 | -0.549 | 0.917 |
| 30 | -0.046 | 0.980 | 1.059 | 0.536 | -0.568 | 0.887 |
| 31 | -0.025 | 0.973 | 1.101 | 0.620 | -0.526 | 0.877 |
| 32 | -0.048 | 1.016 | 1.154 | 0.650 | -0.524 | 0.863 |
| 33 | -0.003 | 0.979 | 1.146 | 0.524 | -0.464 | 0.874 |
| 34 | 0.021 | 1.008 | 1.030 | 0.462 | -0.584 | 0.871 |
| 35 | -0.007 | 1.007 | 0.966 | 0.471 | -0.467 | 0.918 |
| 36 | -0.029 | 0.985 | 0.999 | 0.517 | -0.500 | 0.858 |
| 37 | -0.021 | 1.040 | 0.948 | 0.388 | -0.522 | 0.976 |
| 38 | -0.014 | 0.980 | 1.030 | 0.443 | -0.402 | 0.894 |

| สถานการณ์ที่ | $\theta$ | | a | | b | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD |
| 39 | 0.074 | 1.034 | 0.986 | 0.515 | -0.423 | 0.905 |
| 40 | 0.005 | 0.967 | 1.368 | 0.768 | -0.598 | 0.808 |
| 41 | -0.014 | 1.012 | 1.012 | 0.460 | -0.742 | 0.657 |
| 42 | -0.019 | 1.011 | 1.061 | 0.712 | -0.540 | 0.846 |
| 43 | 0.003 | 0.980 | 0.859 | 0.485 | -0.724 | 0.792 |
| 44 | 0.006 | 0.983 | 1.353 | 0.772 | -0.609 | 0.861 |
| 45 | -0.011 | 1.006 | 1.098 | 0.281 | -0.734 | 0.900 |
| 46 | 0.009 | 0.976 | 0.957 | 0.659 | -0.658 | 0.810 |
| 47 | -0.007 | 0.976 | 0.792 | 0.337 | -0.577 | 0.973 |
| 48 | 0.003 | 0.999 | 1.329 | 0.646 | -0.486 | 0.807 |
| 49 | 0.031 | 1.050 | 1.357 | 0.631 | -0.694 | 0.766 |
| 50 | 0.017 | 0.985 | 1.184 | 0.499 | -0.415 | 0.966 |
| 51 | 0.045 | 1.037 | 1.297 | 0.649 | -0.518 | 0.895 |
| 52 | -0.016 | 1.069 | 1.227 | 0.485 | -0.561 | 0.912 |
| 53 | 0.014 | 0.958 | 1.124 | 0.632 | -0.383 | 0.857 |
| 54 | -0.033 | 1.010 | 1.020 | 0.578 | -0.553 | 0.845 |
| 55 | 0.000 | 1.007 | 1.171 | 0.707 | -0.479 | 0.867 |
| 56 | 0.005 | 0.975 | 1.144 | 0.544 | -0453 | 0.911 |
| 57 | 0.024 | 1.034 | 1.069 | 0.483 | -0.529 | 0.765 |
| 58 | -0.043 | 1.012 | 1.035 | 0.511 | -0.432 | 0.884 |
| 59 | -0.042 | 1.000 | 1.071 | 0.601 | -0.531 | 0.882 |
| 60 | -0.003 | 1.002 | 0.965 | 0.437 | -0.620 | 0.894 |
| 61 | -0.008 | 0.972 | 1.153 | 0.520 | -0.425 | 0.878 |
| 62 | -0.031 | 0.957 | 1.047 | 0.532 | -0.500 | 0.851 |
| 63 | -0.109 | 1.002 | 1.142 | 1.666 | -0.549 | 0.883 |
| 64 | -0.035 | 1.067 | 1.434 | 0.851 | -0.473 | 0.869 |
| 65 | -0.041 | 1.015 | 1.217 | 0.631 | -0.643 | 0.794 |
| 66 | -0.018 | 0.992 | 1.218 | 0.706 | -0.531 | 0.848 |
| 67 | -0.037 | 1.044 | 1.059 | 0.593 | -0.544 | 0.854 |
| 68 | -0.050 | 1.052 | 1.088 | 0.582 | -0.451 | 0.865 |
| 69 | -0.014 | 0.989 | 1.208 | 0.565 | -0.397 | 0.921 |
| 70 | 0.010 | 0.996 | 1.149 | 0.775 | -0.613 | 0.865 |
| 71 | 0.021 | 1.000 | 0.977 | 0.450 | -0.470 | 0.861 |
| 72 | 0.012 | 0.991 | 1.232 | 0.624 | -0.486 | 0.851 |

## ภาคผนวก ค ตัวอย่างข้อมูลการตอบข้อคำถามที่จำลองขึ้นเพื่อใช้ในการวิเคราะห์

| | |
|---|---|
| 0001 | 43334323433424444444 |
| 0002 | 30230300020414002111 |
| 0003 | 44431334434424443442 |
| 0004 | 43442410433414342042 |
| 0005 | 03304344434212343344 |
| 0006 | 04443044233224404344 |
| 0007 | 43242124044414444444 |
| 0008 | 44444444444444444441 |
| 0009 | 44433444444244444444 |
| 0010 | 43232004434314343342 |
| 0011 | 04442323434414444340 |
| 0012 | 31032013432413412342 |
| 0013 | 34214014034424344342 |
| 0014 | 33402423034212343442 |
| 0015 | 44434423434444424441 |
| 0016 | 02200402210010300210 |
| 0017 | 34242444044434444444 |
| 0018 | 44304424444444344444 |
| 0019 | 03400400022212044040 |
| 0020 | 04244044044444414444 |
| 0021 | 33300011300212003020 |
| 0022 | 33002422232422312341 |
| 0023 | 33230203114210304341 |
| 0024 | 43243023444444444441 |
| 0025 | 02202022133212304340 |
| 0026 | 43434424344422414444 |
| 0027 | 21401212033212322021 |
| 0028 | 43100200121102312041 |
| 0029 | 13401224434213342044 |
| 0030 | 23043243433414412444 |
| 0031 | 40040013133111143142 |
| 0032 | 44404244444424444444 |
| 0033 | 43024323030411343341 |
| 0034 | 44344444444344443342 |
| 0035 | 24443424433434404344 |
| 0036 | 44402323434222342341 |
| 0037 | 44334444444244444444 |
| 0038 | 34444444444444444344 |
| 0039 | 33302223434414313342 |
| 0040 | 03332314321224313340 |
| 0041 | 33302113233414413021 |
| 0042 | 44344124044223444444 |
| 0043 | 34443444434424414344 |
| 0044 | 20000301210203313011 |
| 0045 | 43332224043410302344 |
| 0046 | 43334344434444403343 |
| 0047 | 43432324333422403341 |
| 0048 | 13342314033213302441 |
| 0049 | 43202423433414313441 |
| 0050 | 44402444444444444444 |

## ภาคผนวก ง คำสั่งที่ใช้ในการวิเคราะห์ด้วยโปรแกรม PARSCALE

```
      GRADED MODEL - NORMAL REPONSE FUNCTION: EAP SCALE SCORES
>COMMENTS

   All items have 5 categories 20 items 500 persons.
   This command use to calculate type I error.

>FILE   DFNAME='data14_1.DAT', SAVE;
>SAVE   PARM='data14_1.PAR', SCORE='data14_1.SCO';
>INPUT  NIDW=4, NTOTAL=20, NTEST=1, LENGTH=(20), NFMT=1, MAXCAT=5;
(4A1,5X,20A1)
>TEST1 TNAME=testdata14_1, ITEM=(1(1)20), NBLOCK=20;
>BLOCK BNAME=SBLOCK1, NITEMS=1, NCAT=5, CADJUST=0,
ORIGINAL=(0,1,2,3,4), REPEAT=20;
>CALIB  GRADED, LOGISTIC, SCALE=1.702, CYCLES=(100,1,1,1), NQPTS=40,
        ITEMFIT=10;
>SCORE  NOSCORE;
```

## ภาคผนวก จ คำสั่งที่ใช้วิเคราะห์ด้วยโปรแกรม IRTFIT Macros

```
libname macros 'C:\IRTFITMACRO';
options mstored sasmstore=macros;

data data14_1;
infile 'E:\sx2\data14_1.dat';
input id 1-4 it1 10 it2 11 it3 12 it4 13 it5 14 it6 15 it7 16 it8 17 it9 18 it10 19 it11 20 it12 21 it13 22 it14
23 it15 24 it16 25 it17 26 it18 27 it19 28 it20 29;
run;

data par14_1;
infile 'E:\sx2\par14_1.txt';
input name $ choices model $ slope location category1 category2 category3 category4;
proc print data=data14_1;
run;
proc print data=par14_1;
run;

%IRTFIT(DATA=data14_1
,PARFILE=par14_1
,ITEMLIST=it1-it20
,TESTMETHOD=SUM
,outfmt=rtf
,outlib=E:\sx2\data14
,outcore=data14_1
,d=1.7
,mincode=0
,maxchoice=9
,missing=.
);
```

**ภาคผนวก ฉ ตัวอย่างแฟ้มข้อมูลพารามิเตอร์ข้อสอบที่ใช้ในการวิเคราะห์บนโปรแกรม IRTFIT Macros**

it1 5 GRM 0.754 -0.553 0.607 0.303 -0.053 -0.858

it2 5 GRM 2.320 -0.607 0.752 0.293 0.236 -1.281

it3 5 GRM 0.580 -0.444 0.814 0.682 -0.162 -1.334

it4 5 GRM 0.696 -0.234 0.415 0.285 0.088 -0.788

it5 5 GRM 1.351 -0.161 0.932 0.678 -0.583 -1.027

it6 5 GRM 0.599 -0.633 1.049 0.721 -0.284 -1.486

it7 5 GRM 1.806 -0.036 1.171 0.170 -0.638 -0.704

it8 5 GRM 1.576 -0.714 0.795 0.277 -0.184 -0.888

it9 5 GRM 0.558 -0.846 0.757 0.250 -0.427 -0.581

it10 5 GRM 1.887 -0.799 1.119 0.535 0.008 -1.661

it11 5 GRM 0.968 -0.923 0.610 0.545 -0.015 -1.141

it12 5 GRM 0.968 -0.893 0.856 0.679 -0.680 -0.854

it13 5 GRM 2.535  0.158 1.401 -0.026 -0.630 -0.746

it14 5 GRM 1.071 -0.862 0.884 0.576 -0.484 -0.976

it15 5 GRM 1.013 -1.037 0.632 0.438 0.222 -1.293

it16 5 GRM 0.659  0.169 1.036 -0.129 -0.449 -0.459

it17 5 GRM 1.256 -0.852 0.742 0.685 -0.149 -1.278

it18 5 GRM 0.977 -0.666 0.645 0.451 0.230 -1.326

it19 5 GRM 3.017 -1.291 0.450 -0.024 -0.165 -0.261

it20 5 GRM 1.199 -0.051 1.090 -0.099 -0.478 -0.513

## ภาคผนวก ช ตัวอย่างผลลัพธ์จากโปรแกรม PARSCALE ในการวิเคราะห์ค่าความคลาดเคลื่อนประเภทที่ 1

```
                                    PARSCALE V4.1

     MAXIMUM LIKELIHOOD ITEM ANALYSIS AND TEST SCORING: POLYTOMOUS MODEL


    [ PHASE 2 ]


    CURRENT DATE: 12-24-2010
    CURRENT TIME: 16:00:27

                ***  POLYTOMOUS MODEL ITEM ANALYSER  ***

                          *** PHASE 2  ***

              GRADED MODEL - NORMAL REPONSE FUNCTION: EAP SCALE SCORES

MAINTEST: TESTDATA

                        ******************************
                           CALIBRATION OF MAINTEST
                                   TESTDATA
                        ******************************


[E-M CYCLES]   GRADED RESPONSE MODEL

CATEGORY AND ITEM PARAMETERS AFTER CYCLE    0

  LARGEST CHANGE=  0.000
  -2 LOG LIKELIHOOD =      21498.703

CATEGORY AND ITEM PARAMETERS AFTER CYCLE    1

  LARGEST CHANGE=  1.819 (  1.442-> -0.378) at Category of Block: 19 BLOCK
  -2 LOG LIKELIHOOD =      23590.746

CATEGORY AND ITEM PARAMETERS AFTER CYCLE    2

  LARGEST CHANGE=  2.963 ( -1.179-> -4.142) at Category of Block: 15 BLOCK
  -2 LOG LIKELIHOOD =      21798.491

CATEGORY AND ITEM PARAMETERS AFTER CYCLE    3

  LARGEST CHANGE=  0.688 ( -2.402-> -1.714) at Location of Item: 19 0019
  -2 LOG LIKELIHOOD =      21575.864

CATEGORY AND ITEM PARAMETERS AFTER CYCLE    4

  LARGEST CHANGE=  0.552 ( -4.448-> -3.895) at Category of Block: 15 BLOCK
  -2 LOG LIKELIHOOD =      21391.067

CATEGORY AND ITEM PARAMETERS AFTER CYCLE    5

  LARGEST CHANGE=  0.524 ( -3.895-> -3.371) at Category of Block: 15 BLOCK
  -2 LOG LIKELIHOOD =      21262.678

CATEGORY AND ITEM PARAMETERS AFTER CYCLE    6
```

```
  LARGEST CHANGE=  0.472 ( -3.371-> -2.899) at Category of Block: 15 BLOCK
  -2 LOG LIKELIHOOD =      21175.675

CATEGORY AND ITEM PARAMETERS AFTER CYCLE    7

  LARGEST CHANGE=  0.401 ( -2.899-> -2.497) at Category of Block: 15 BLOCK
  -2 LOG LIKELIHOOD =      21117.354

CATEGORY AND ITEM PARAMETERS AFTER CYCLE    8

  LARGEST CHANGE=  0.324 ( -2.497-> -2.173) at Category of Block: 15 BLOCK
  -2 LOG LIKELIHOOD =      21078.667

CATEGORY AND ITEM PARAMETERS AFTER CYCLE    9

  LARGEST CHANGE=  0.247 ( -2.173-> -1.927) at Category of Block: 15 BLOCK
  -2 LOG LIKELIHOOD =      21054.284

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   10

  LARGEST CHANGE=  0.178 ( -1.927-> -1.749) at Category of Block: 15 BLOCK
  -2 LOG LIKELIHOOD =      21039.578

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   11

  LARGEST CHANGE=  0.121 ( -1.749-> -1.628) at Category of Block: 15 BLOCK
  -2 LOG LIKELIHOOD =      21030.627

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   12

  LARGEST CHANGE=  0.079 ( -1.628-> -1.549) at Category of Block: 15 BLOCK
  -2 LOG LIKELIHOOD =      21024.745

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   13

  LARGEST CHANGE=  0.051 ( -1.549-> -1.498) at Category of Block: 15 BLOCK
  -2 LOG LIKELIHOOD =      21020.456

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   14

  LARGEST CHANGE=  0.033 ( -1.498-> -1.464) at Category of Block: 15 BLOCK
  -2 LOG LIKELIHOOD =      21017.072

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   15

  LARGEST CHANGE=  0.028 (  2.693->  2.721) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      21014.287

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   16

  LARGEST CHANGE=  0.025 (  2.721->  2.746) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      21011.950

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   17

  LARGEST CHANGE=  0.022 (  2.746->  2.769) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      21009.973

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   18

  LARGEST CHANGE=  0.020 (  2.769->  2.789) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      21008.292

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   19

  LARGEST CHANGE=  0.019 (  2.789->  2.808) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      21006.860

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   20

  LARGEST CHANGE=  0.017 (  2.808->  2.825) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      21005.636
```

```
CATEGORY AND ITEM PARAMETERS AFTER CYCLE   21

  LARGEST CHANGE=  0.016 (  2.825->  2.840) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      21004.588

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   22

  LARGEST CHANGE=  0.014 (  2.840->  2.855) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      21003.689

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   23

  LARGEST CHANGE=  0.013 (  2.855->  2.868) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      21002.916

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   24

  LARGEST CHANGE=  0.012 (  2.868->  2.880) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      21002.250

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   25

  LARGEST CHANGE=  0.011 (  2.880->  2.892) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      21001.676

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   26

  LARGEST CHANGE=  0.010 (  2.892->  2.902) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      21001.180

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   27

  LARGEST CHANGE=  0.010 (  2.902->  2.912) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      21000.750

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   28

  LARGEST CHANGE=  0.009 (  2.912->  2.921) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      21000.378

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   29

  LARGEST CHANGE=  0.008 (  2.921->  2.929) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      21000.056

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   30

  LARGEST CHANGE=  0.008 (  2.929->  2.937) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      20999.775

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   31

  LARGEST CHANGE=  0.007 (  2.937->  2.944) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      20999.531

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   32

  LARGEST CHANGE=  0.007 (  2.944->  2.950) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      20999.318

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   33

  LARGEST CHANGE=  0.006 (  2.950->  2.956) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      20999.132

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   34

  LARGEST CHANGE=  0.006 (  2.956->  2.962) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      20998.970

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   35
```

```
  LARGEST CHANGE=  0.005 (  2.962->  2.967) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      20998.828

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   36

  LARGEST CHANGE=  0.005 (  2.967->  2.972) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      20998.704

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   37

  LARGEST CHANGE=  0.004 (  2.972->  2.976) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      20998.595

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   38

  LARGEST CHANGE=  0.004 (  2.976->  2.980) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      20998.499

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   39

  LARGEST CHANGE=  0.004 (  2.980->  2.984) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      20998.415

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   40

  LARGEST CHANGE=  0.003 (  2.984->  2.987) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      20998.341

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   41

  LARGEST CHANGE=  0.003 (  2.987->  2.991) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      20998.276

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   42

  LARGEST CHANGE=  0.003 (  2.991->  2.994) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      20998.218

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   43

  LARGEST CHANGE=  0.003 (  2.994->  2.996) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      20998.167

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   44

  LARGEST CHANGE=  0.003 (  2.996->  2.999) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      20998.122

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   45

  LARGEST CHANGE=  0.002 (  2.999->  3.001) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      20998.082

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   46

  LARGEST CHANGE=  0.002 (  3.001->  3.003) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      20998.047

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   47

  LARGEST CHANGE=  0.002 (  3.003->  3.005) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      20998.015

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   48

  LARGEST CHANGE=  0.002 (  3.005->  3.007) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      20997.988

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   49

  LARGEST CHANGE=  0.002 (  3.007->  3.009) at Slope    of Item: 19 0019
```

```
  -2 LOG LIKELIHOOD =      20997.963

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   50

  LARGEST CHANGE=  0.002 (  3.009->  3.010) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      20997.941

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   51

  LARGEST CHANGE=  0.001 (  3.010->  3.012) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      20997.921

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   52

  LARGEST CHANGE=  0.001 (  3.012->  3.013) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      20997.903

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   53

  LARGEST CHANGE=  0.001 (  3.013->  3.014) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      20997.888

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   54

  LARGEST CHANGE=  0.001 (  3.014->  3.015) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      20997.874

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   55

  LARGEST CHANGE=  0.001 (  3.015->  3.016) at Slope    of Item: 19 0019
  -2 LOG LIKELIHOOD =      20997.861

CATEGORY AND ITEM PARAMETERS AFTER CYCLE   56

  LARGEST CHANGE=  0.001 (  3.016->  3.017) at Slope    of Item: 19 0019


  ITEM BLOCK   1  SBLOCK1

  CATEGORY PARAMETER  :     0.607     0.303    -0.053    -0.858
  S.E.                :     0.091     0.085     0.080     0.080
  ITEM BLOCK   2  BLOCK

  CATEGORY PARAMETER  :     0.752     0.293     0.236    -1.281
  S.E.                :     0.056     0.043     0.042     0.041
  ITEM BLOCK   3  BLOCK

  CATEGORY PARAMETER  :     0.814     0.682    -0.162    -1.334
  S.E.                :     0.111     0.108     0.098     0.104
  ITEM BLOCK   4  BLOCK

  CATEGORY PARAMETER  :     0.415     0.285     0.088    -0.788
  S.E.                :     0.088     0.086     0.085     0.087
  ITEM BLOCK   5  BLOCK

  CATEGORY PARAMETER  :     0.932     0.678    -0.583    -1.027
  S.E.                :     0.063     0.058     0.052     0.058
  ITEM BLOCK   6  BLOCK

  CATEGORY PARAMETER  :     1.049     0.721    -0.284    -1.486
  S.E.                :     0.119     0.110     0.096     0.101
  ITEM BLOCK   7  BLOCK

  CATEGORY PARAMETER  :     1.171     0.170    -0.638    -0.704
  S.E.                :     0.058     0.042     0.045     0.046
  ITEM BLOCK   8  BLOCK

  CATEGORY PARAMETER  :     0.795     0.277    -0.184    -0.888
  S.E.                :     0.071     0.055     0.048     0.046
  ITEM BLOCK   9  BLOCK

  CATEGORY PARAMETER  :     0.757     0.250    -0.427    -0.581
```

```
   S.E.                :     0.122     0.111     0.103     0.102
   ITEM BLOCK  10  BLOCK

   CATEGORY PARAMETER  :     1.119     0.535     0.008    -1.661
   S.E.                :     0.086     0.059     0.047     0.049
   ITEM BLOCK  11  BLOCK

   CATEGORY PARAMETER  :     0.610     0.545    -0.015    -1.141
   S.E.                :     0.088     0.086     0.072     0.065
   ITEM BLOCK  12  BLOCK

   CATEGORY PARAMETER  :     0.856     0.679    -0.680    -0.854
   S.E.                :     0.097     0.090     0.065     0.065
   ITEM BLOCK  13  BLOCK

   CATEGORY PARAMETER  :     1.401    -0.026    -0.630    -0.746
   S.E.                :     0.051     0.035     0.039     0.041
   ITEM BLOCK  14  BLOCK

   CATEGORY PARAMETER  :     0.884     0.576    -0.484    -0.976
   S.E.                :     0.093     0.081     0.061     0.060
   ITEM BLOCK  15  BLOCK

   CATEGORY PARAMETER  :     0.632     0.438     0.222    -1.293
   S.E.                :     0.092     0.084     0.078     0.064
   ITEM BLOCK  16  BLOCK

   CATEGORY PARAMETER  :     1.036    -0.129    -0.449    -0.459
   S.E.                :     0.095     0.089     0.091     0.091
   ITEM BLOCK  17  BLOCK

   CATEGORY PARAMETER  :     0.742     0.685    -0.149    -1.278
   S.E.                :     0.081     0.079     0.058     0.055
   ITEM BLOCK  18  BLOCK

   CATEGORY PARAMETER  :     0.645     0.451     0.230    -1.326
   S.E.                :     0.081     0.076     0.072     0.068
   ITEM BLOCK  19  BLOCK

   CATEGORY PARAMETER  :     0.450    -0.024    -0.165    -0.261
   S.E.                :     0.065     0.046     0.042     0.040
   ITEM BLOCK  20  BLOCK

   CATEGORY PARAMETER  :     1.090    -0.099    -0.478    -0.513
   S.E.                :     0.069     0.055     0.057     0.057
```

| ITEM | BLOCK | SLOPE | S.E. | LOCATION | S.E. | GUESSING | S.E. |
|---|---|---|---|---|---|---|---|
| 0001 | 1 | 0.754 | 0.046 | -0.553 | 0.069 | 0.000 | 0.000 |
| 0002 | 2 | 2.320 | 0.142 | -0.607 | 0.028 | 0.000 | 0.000 |
| 0003 | 3 | 0.580 | 0.034 | -0.444 | 0.085 | 0.000 | 0.000 |
| 0004 | 4 | 0.696 | 0.046 | -0.234 | 0.074 | 0.000 | 0.000 |
| 0005 | 5 | 1.351 | 0.070 | -0.161 | 0.039 | 0.000 | 0.000 |
| 0006 | 6 | 0.599 | 0.033 | -0.633 | 0.082 | 0.000 | 0.000 |
| 0007 | 7 | 1.806 | 0.095 | -0.036 | 0.031 | 0.000 | 0.000 |
| 0008 | 8 | 1.576 | 0.085 | -0.714 | 0.036 | 0.000 | 0.000 |
| 0009 | 9 | 0.558 | 0.041 | -0.846 | 0.103 | 0.000 | 0.000 |
| 0010 | 10 | 1.887 | 0.104 | -0.799 | 0.032 | 0.000 | 0.000 |
| 0011 | 11 | 0.968 | 0.055 | -0.923 | 0.056 | 0.000 | 0.000 |
| 0012 | 12 | 0.968 | 0.059 | -0.893 | 0.060 | 0.000 | 0.000 |

| 0013 | 13 | 2.535 | 0.147 | 0.158 | 0.025 | 0.000 | 0.000 |
|---|---|---|---|---|---|---|---|
| 0014 | 14 | 1.071 | 0.060 | -0.862 | 0.052 | 0.000 | 0.000 |
| 0015 | 15 | 1.013 | 0.058 | -1.037 | 0.055 | 0.000 | 0.000 |
| 0016 | 16 | 0.659 | 0.042 | 0.169 | 0.077 | 0.000 | 0.000 |
| 0017 | 17 | 1.256 | 0.067 | -0.852 | 0.043 | 0.000 | 0.000 |
| 0018 | 18 | 0.977 | 0.054 | -0.666 | 0.054 | 0.000 | 0.000 |
| 0019 | 19 | 3.017 | 0.285 | -1.291 | 0.037 | 0.000 | 0.000 |
| 0020 | 20 | 1.199 | 0.067 | -0.051 | 0.045 | 0.000 | 0.000 |

SUMMARY STATISTICS OF PARAMETER ESTIMATES

| PARAMETER | MEAN | STN DEV | N |
|---|---|---|---|
| SLOPE | 1.289 | 0.700 | 20 |
| LOG(SLOPE) | 0.128 | 0.508 | 20 |
| THRESHOLD | -0.564 | 0.412 | 20 |
| GUESSING | 0.000 | 0.000 | 0 |

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| POINT | -0.4000E+01 | -0.3795E+01 | -0.3590E+01 | -0.3385E+01 | -0.3179E+01 |
| WEIGHT | 0.3872E-04 | 0.8283E-04 | 0.1687E-03 | 0.3289E-03 | 0.6218E-03 |

| | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|
| POINT | -0.2974E+01 | -0.2769E+01 | -0.2564E+01 | -0.2359E+01 | -0.2154E+01 |
| WEIGHT | 0.1158E-02 | 0.2135E-02 | 0.3800E-02 | 0.6156E-02 | 0.8686E-02 |

| | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|
| POINT | -0.1949E+01 | -0.1744E+01 | -0.1538E+01 | -0.1333E+01 | -0.1128E+01 |
| WEIGHT | 0.1156E-01 | 0.1689E-01 | 0.2280E-01 | 0.2619E-01 | 0.3813E-01 |

| | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|
| POINT | -0.9231E+00 | -0.7179E+00 | -0.5128E+00 | -0.3077E+00 | -0.1026E+00 |
| WEIGHT | 0.5571E-01 | 0.6689E-01 | 0.7131E-01 | 0.7691E-01 | 0.8548E-01 |

| | 21 | 22 | 23 | 24 | 25 |
|---|---|---|---|---|---|
| POINT | 0.1026E+00 | 0.3077E+00 | 0.5128E+00 | 0.7179E+00 | 0.9231E+00 |
| WEIGHT | 0.8531E-01 | 0.7796E-01 | 0.7132E-01 | 0.6211E-01 | 0.5206E-01 |

| | 26 | 27 | 28 | 29 | 30 |
|---|---|---|---|---|---|
| POINT | 0.1128E+01 | 0.1333E+01 | 0.1538E+01 | 0.1744E+01 | 0.1949E+01 |
| WEIGHT | 0.4398E-01 | 0.3477E-01 | 0.2569E-01 | 0.1835E-01 | 0.1267E-01 |

| | 31 | 32 | 33 | 34 | 35 |
|---|---|---|---|---|---|
| POINT | 0.2154E+01 | 0.2359E+01 | 0.2564E+01 | 0.2769E+01 | 0.2974E+01 |
| WEIGHT | 0.8376E-02 | 0.5283E-02 | 0.3183E-02 | 0.1836E-02 | 0.1016E-02 |

| | 36 | 37 | 38 | 39 | 40 |
|---|---|---|---|---|---|
| POINT | 0.3179E+01 | 0.3385E+01 | 0.3590E+01 | 0.3795E+01 | 0.4000E+01 |
| WEIGHT | 0.5390E-03 | 0.2745E-03 | 0.1341E-03 | 0.6288E-04 | 0.2827E-04 |

TOTAL WEIGHT: 1.00000
MEAN : 0.00000
S.D. : 0.99965

ITEM FIT STATISTICS

| BLOCK | ITEM | CHI-SQUARE | D.F. | PROB. |
|---|---|---|---|---|
| SBLOCK1 | 0001 | 21.10732 | 22. | 0.514 |
| BLOCK | 0002 | 8.03859 | 13. | 0.842 |
| BLOCK | 0003 | 26.80828 | 21. | 0.177 |
| BLOCK | 0004 | 12.80302 | 18. | 0.804 |
| BLOCK | 0005 | 8.64121 | 18. | 0.967 |
| BLOCK | 0006 | 14.14946 | 23. | 0.923 |
| BLOCK | 0007 | 6.28122 | 13. | 0.935 |
| BLOCK | 0008 | 7.88456 | 16. | 0.952 |
| BLOCK | 0009 | 15.65992 | 17. | 0.548 |
| BLOCK | 0010 | 4.74970 | 14. | 0.989 |
| BLOCK | 0011 | 8.33410 | 17. | 0.959 |
| BLOCK | 0012 | 11.31202 | 14. | 0.662 |
| BLOCK | 0013 | 5.99176 | 14. | 0.966 |
| BLOCK | 0014 | 9.65336 | 17. | 0.918 |
| BLOCK | 0015 | 9.29316 | 16. | 0.901 |
| BLOCK | 0016 | 25.71953 | 18. | 0.106 |
| BLOCK | 0017 | 21.16664 | 16. | 0.172 |
| BLOCK | 0018 | 8.61147 | 18. | 0.968 |
| BLOCK | 0019 | 4.00569 | 7. | 0.781 |
| BLOCK | 0020 | 8.96270 | 16. | 0.915 |
| TOTAL | | 239.17371 | 328. | 1.000 |

225372 BYTES OF NUMERICAL WORKSPACE USED OF 8192000 AVAILABLE IN PHASE 2
1104 BYTES OF CHARACTER WORKSPACE USED OF 2048000 AVAILABLE IN PHASE 2
NORMAL END

**ภาคผนวก ซ ตัวอย่างผลลัพธ์จาก IRTFIT Macros ในการวิเคราะห์ค่าความคลาดเคลื่อนประเภทที่ 1**

| *item_no* | *name* | *df* | *G2* | *Prob_G2* | *X2* | *Prob_X2* |
|---|---|---|---|---|---|---|
| 1 | IT1 | 54 | 47.57 | 0.7192 | 46.89 | 0.7427 |
| 2 | IT2 | 49 | 21.39 | 0.9998 | 21.48 | 0.9998 |
| 3 | IT3 | 58 | 61.91 | 0.3383 | 57.26 | 0.5029 |
| 4 | IT4 | 56 | 33.22 | 0.9934 | 32.69 | 0.9946 |
| 5 | IT5 | 53 | 32.20 | 0.9893 | 32.00 | 0.9901 |
| 6 | IT6 | 59 | 57.70 | 0.5236 | 56.76 | 0.5585 |
| 7 | IT7 | 52 | 37.59 | 0.9335 | 35.05 | 0.9656 |
| 8 | IT8 | 53 | 38.26 | 0.9363 | 36.31 | 0.9613 |
| 9 | IT9 | 55 | 38.52 | 0.9553 | 35.76 | 0.9794 |
| 10 | IT10 | 53 | 23.74 | 0.9998 | 24.04 | 0.9998 |
| 11 | IT11 | 55 | 34.75 | 0.9851 | 30.71 | 0.9967 |
| 12 | IT12 | 54 | 32.56 | 0.9908 | 30.31 | 0.9962 |
| 13 | IT13 | 53 | 25.48 | 0.9995 | 24.52 | 0.9997 |
| 14 | IT14 | 55 | 40.15 | 0.9336 | 39.23 | 0.9466 |
| 15 | IT15 | 54 | 25.24 | 0.9997 | 24.78 | 0.9998 |
| 16 | IT16 | 58 | 46.42 | 0.8628 | 44.88 | 0.8963 |
| 17 | IT17 | 52 | 42.81 | 0.8142 | 41.58 | 0.8491 |
| 18 | IT18 | 54 | 41.63 | 0.8907 | 40.08 | 0.9207 |
| 19 | IT19 | 44 | 11.88 | 1.0000 | 12.03 | 1.0000 |
| 20 | IT20 | 54 | 30.88 | 0.9952 | 30.65 | 0.9956 |

# ประวัติผู้เขียนวิทยานิพนธ์

นางสาวธีรนุช จาบประไพ เกิดเมื่อวันที่ 26 ตุลาคม 2528 ที่อำเภอพระพุทธบาท จังหวัดสระบุรี สำเร็จการศึกษาปริญญาวิทยาศาสตรบัณฑิต สาขาวิชาสถิติ จากคณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์ เมื่อปีการศึกษา 2550 และเข้าศึกษาต่อในหลักสูตรครุศาสตรมหาบัณฑิต สาขาวิชาการวัดและประเมินผลการศึกษา ภาควิชาวิจัยและจิตวิทยาการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2552 ปัจจุบันรับราชการในตำแหน่งนักวิชาการสถิติปฏิบัติการ สังกัดกรมทางหลวง กระทรวงคมนาคม