Chulalongkorn University
Chula Digital Collections

Chulalongkorn University Theses and Dissertations (Chula ETD)

2019

การวิเคราะห์ฉันทามติและการทำหน้าที่ต่างกันระหว่างผู้ประเมินในการวิเคราะห์ความสอดคล้องในแนวเดียวกัน: การประยุกต์ใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม

ศิธรา จุฑารัตน์
คณะครุศาสตร์

Follow this and additional works at: https://digital.car.chula.ac.th/chulaetd

Part of the Educational Assessment, Evaluation, and Research Commons

Recommended Citation

จุฑารัตน์, ศิธรา, "การวิเคราะห์ฉันทามติและการทำหน้าที่ต่างกันระหว่างผู้ประเมินในการวิเคราะห์ความสอดคล้องในแนวเดียวกัน: การประยุกต์ใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม" (2019). *Chulalongkorn University Theses and Dissertations (Chula ETD)*. 9082.
https://digital.car.chula.ac.th/chulaetd/9082

This Thesis is brought to you for free and open access by Chula Digital Collections. It has been accepted for inclusion in Chulalongkorn University Theses and Dissertations (Chula ETD) by an authorized administrator of Chula Digital Collections. For more information, please contact ChulaDC@car.chula.ac.th.

การวิเคราะห์ฉันทามติและการทำหน้าที่ต่างกันระหว่างผู้ประเมินในการวิเคราะห์ความสอดคล้องในแนวเดียวกัน: การประยุกต์ใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม

น.ส.ศิธรา จุฑารัตน์

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาครุศาสตรดุษฎีบัณฑิต
สาขาวิชาการวัดและประเมินผลการศึกษา ภาควิชาวิจัยและจิตวิทยาการศึกษา
คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2562

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

ANALYSIS OF RATERS CONSENSUS AND DIFFERENTIAL RATER FUNCTIOING IN ALIGNMENT ANALYSIS: THE APPLICATION OF CULTURAL CONSENSUS ANALYSIS MODEL

Miss Sitara Jutharat

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in Educational Measurement and Evaluation
Department of Educational Research and Psychology
FACULTY OF EDUCATION
Chulalongkorn University
Academic Year 2019
Copyright of Chulalongkorn University

หัวข้อวิทยานิพนธ์	การวิเคราะห์ฉันทามติและการทำหน้าที่ต่างกันระหว่างผู้ประเมินในการวิเคราะห์ความสอดคล้องในแนวเดียวกัน: การประยุกต์ใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม
โดย	น.ส.ศิธรา จุฑารัตน์
สาขาวิชา	การวัดและประเมินผลการศึกษา
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	รองศาสตราจารย์ ดร.ศิริเดช สุชีวะ
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	อาจารย์ ดร.สิวะโชติ ศรีสุทธิยากร

คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาครุศาสตรดุษฎีบัณฑิต

..... คณบดีคณะครุศาสตร์

(รองศาสตราจารย์ ดร.ศิริเดช สุชีวะ)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(ศาสตราจารย์ ดร.ศิริชัย กาญจนวาสี)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(รองศาสตราจารย์ ดร.ศิริเดช สุชีวะ)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม

(อาจารย์ ดร.สิวะโชติ ศรีสุทธิยากร)

..... กรรมการ

(รองศาสตราจารย์ ดร.โชติกา ภาษีผล)

..... กรรมการ

(รองศาสตราจารย์ ดร.ณัฏฐภรณ์ หลาวทอง)

..... กรรมการภายนอกมหาวิทยาลัย

(รองศาสตราจารย์ ดร.สังวรณ์ งัดกระโทก)

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ศิธรา จุฑารัตน์ : การวิเคราะห์ฉันทามติและการทำหน้าที่ต่างกันระหว่างผู้ประเมินในการวิเคราะห์ความสอดคล้องในแนวเดียวกัน: การประยุกต์ใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม. (ANALYSIS OF RATERS CONSENSUS AND DIFFERENTIAL RATER FUNCTIOING IN ALIGNMENT ANALYSIS: THE APPLICATION OF CULTURAL CONSENSUS ANALYSIS MODEL) อ.ที่ปรึกษาหลัก : รศ. ดร.ศิริเดช สุชีวะ, อ.ที่ปรึกษาร่วม : อ. ดร.สิวะโชติ ศรีสุทธิยากร

การวิจัยนี้มีวัตถุประสงค์ 1) เพื่อประยุกต์ใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมเพื่อวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินในการศึกษาความสอดคล้องในแนวเดียวกัน 2) เพื่อตรวจสอบประสิทธิภาพของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมในการการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินในการศึกษาความสอดคล้องในแนวเดียวกัน 3) เพื่อศึกษาผลการประยุกต์ใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมในการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินและการทำหน้าที่ต่างกันของผู้ประเมินในการประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียนกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น

การดำเนินการศึกษาแบ่งเป็น 3 ระยะ คือ ระยะที่ 1 ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง ระยะที่ 2 ศึกษาผลการประมาณความสอดคล้องระหว่างผู้ประเมินด้วยการจำลองแบบมอนติคาร์โล แบ่งเป็น การจำลองข้อมูล การประเมินประสิทธิภาพของการประมาณค่าของโมเดล และการวิเคราะห์ผลการจำลองข้อมูล ระยะที่ 3 ศึกษาผลการประยุกต์ใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมในการวิเคราะห์ความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบ ในการประเมินระดับชั้นเรียน กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น ผลการวิจัยสรุปได้ดังนี้

1) โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมในการศึกษาครั้งนี้มี 2 โมเดล โมเดล GCM หรือ MC-GCM เป็นโมเดลวิเคราะห์ข้อมูลในบริบทการประเมินที่ให้คะแนนแบบ (0, 1) จะมีพารามิเตอร์ของข้อคำถาม ซึ่งประกอบด้วยพารามิเตอร์ตำแหน่งคะแนนฉันทามติกับพารามิเตอร์ความยากของข้อคำถาม พารามิเตอร์ของผู้ประเมิน ประกอบด้วยพารามิเตอร์ความสามารถของผู้ประเมินกับพารามิเตอร์ความลำเอียงในการประเมิน โมเดล LTRM หรือ MC-LTRM เป็นโมเดลการวิเคราะห์ข้อมูลในบริบทการประเมินที่ให้คะแนนแบบมาตรประมาณค่า ประกอบด้วยพารามิเตอร์ 2 กลุ่ม คือ พารามิเตอร์ของข้อคำถามประกอบด้วยพารามิเตอร์ตำแหน่งคะแนนฉันทามติของการประเมินกับพารามิเตอร์ความยากของคำถามประเมิน พารามิเตอร์ของผู้ประเมิน ประกอบด้วยพารามิเตอร์ความสามารถของผู้ประเมิน กับพารามิเตอร์ความลำเอียงในการประเมิน

2) ผลการตรวจสอบประสิทธิภาพของการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินด้วยการจำลองข้อมูล พบว่า โมเดล MC-GCM และโมเดล MC-LTRM สามารถประมาณค่าพารามิเตอร์ได้ใกล้เคียงกับค่าจริงของพารามิเตอร์ที่กำหนดไว้ โดยมีค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสอง และค่าความลำเอียงในการประมาณค่าที่เข้าใกล้ 0 และมีค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าจริงและค่าที่ได้จากการประมาณค่าของโมเดลในระดับสูง ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าจริงและค่าที่ได้จากการประมาณค่ามีความสัมพันธ์กันในระดับสูง และมีนัยสำคัญทางสถิติ ผลการวิเคราะห์อิทธิพลของตัวแปรอิสระที่ส่งผลต่อการประมาณค่า พบว่า ปัจจัยที่ส่งผลต่อประสิทธิภาพในการประมาณค่าของโมเดลทั้งสอง คือ การทำหน้าที่ต่างกันของผู้ประเมิน ซึ่งส่งผลต่อค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสองและค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าจริงและค่าที่ได้จากการประมาณค่าของโมเดลในการประมาณค่าพารามิเตอร์ความสามารถของผู้ประเมิน และความยากของรายการประเมินอย่างมีนัยสำคัญทางสถิติ

3) ผลการวิเคราะห์ความสอดคล้องในแนวเดียวกันด้วยโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม พบว่า ไม่มีการทำหน้าที่ต่างกันระหว่างผู้ประเมิน โดยมีผลการประเมิน ดังนี้ 3.1) ผลคะแนนการประเมินระดับความซับซ้อนทางปัญญาของข้อสอบในการประเมินระดับชาติ กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับชั้นมัธยมศึกษาปีที่ 3 มีตำแหน่งคะแนนการประเมินอยู่ในตำแหน่งคะแนนประเมินระดับ 2 (เข้าใจ) ถึงระดับ 4 (ประยุกต์ใช้) 3.2) ผลการประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียน มีตำแหน่งคะแนนการประเมินจะอยู่ในเทรชโฮลด์ที่ 4 หรือระดับการประเมินที่ 5 (สอดคล้องโดยตรง)

สาขาวิชา	การวัดและประเมินผลการศึกษา	ลายมือชื่อนิสิต .....
ปีการศึกษา	2562	ลายมือชื่อ อ.ที่ปรึกษาหลัก .....
		ลายมือชื่อ อ.ที่ปรึกษาร่วม .....

5884221327 : MAJOR EDUCATIONAL MEASUREMENT AND EVALUATION
KEYWORD: CULTURAL CONSENSUS MODELS, ALIGNMENT, DIFFERENTIAL RATER FUNCTIONING

Sitara Jutharat : ANALYSIS OF RATERS CONSENSUS AND DIFFERENTIAL RATER FUNCTIOING IN ALIGNMENT ANALYSIS: THE APPLICATION OF CULTURAL CONSENSUS ANALYSIS MODEL . Advisor: Assoc. Prof. SIRIDEJ SUJIVA, Ph.D. Co-advisor: SIWACHOAT SRISUTTIYAKORN, Ph.D.

The purposes of this study were to 1) apply Cultural Consensus Models in analyzing raters' consensus for educational alignment analysis 2) investigate and verify model efficiency in estimating raters' consensus for alignment analysis 3) examine the result of applying Cultural Consensus model in the analysis of raters' consensus and differential rater functioning in the evaluation of the standards and indicators-classroom test items alignment in junior secondary school education.

This study is divided into 3 phrases. First, research papers and documents involved in the study were reviewed. Second, Monte Carlo study was conducted to investigate models' performance. Third, the 2 Cultural Consensus Models were implemented with evaluation of science standards and indicators-classroom test items alignment in the junior secondary school education to study alignment between classroom test items and science standards and indicators. Research results were as follows:

1) This research studied 2 Cultural Consensus models. GCM/MC-GCM for dichotomous data has 2 sets of parameters; item parameter and rater parameter. The item parameter includes consensus score and item difficulty while rater parameter includes rater competency and rater's bias. LTRM/MC-LTRM for polytomous data also has 2 sets of parameters. The item parameter includes consensus score and item difficulty. The rater parameter includes rater competency and rater's bias.

2) The simulation study showed that both MC-GCM and MC-LTRM can yield the parameters estimation close to the specified parameter values with MSE and Bias of estimator values close to zero and had high correlation between the actual and estimated value. Partial ETA squared showed that differential rater functioning had significant effect on model capability in estimating rater's competency and item difficulty parameters which resulted in higher MSE and lower correlation between the actual and estimated values

3) Alignment analysis results showed no evidence of differential rater functioning. Raters' consensus results are; 3.1) the cognitive demand evaluation of national science test items was in the 2nd (understanding) and the 4th (applying) category. 3.2) the posterior of consensus score parameter of standard and indicators-classroom test items alignment score indicated that all raters gave the evaluation score at the 5th (completely align)

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Field of Study:	Educational Measurement and Evaluation	Student's Signature .....
Academic Year:	2019	Advisor's Signature .....
		Co-advisor's Signature .....

กิตติกรรมประกาศ

ผู้วิจัยขออุทิศการศึกษาครั้งนี้ แด่ Professor William H. Batchelder, PhD (1940 - 2018) ผู้พัฒนาโมเดล ซึ่งเคยให้คำแนะนำเกี่ยวกับทฤษฎีการวิเคราะห์ฉันทามติเชิงวัฒนธรรมแก่ผู้วิจัยตั้งแต่เริ่มพัฒนาหัวข้อวิทยานิพนธ์ และขอขอบคุณ ดร.รอยส์ แอนเดอร์ส (Royce Anders, PhD) สำหรับคำแนะนำเกี่ยวกับโมเดลการวิเคราะห์ข้อมูลและการใช้แพ็กเกจสำเร็จรูปสำหรับการวิเคราะห์ฉันทามติเชิงวัฒนธรรม

ผู้วิจัยขอกราบขอบพระคุณ รองศาสตราจารย์ ดร.ศิริเดช สุชีวะ อาจารย์ที่ปรึกษา และอาจารย์ ดร.สิวะโชติ ศรีสุทธิยากร อาจารย์ที่ปรึกษาร่วม ผู้ให้คำแนะนำและข้อเสนอแนะแก่ผู้วิจัยในการทำวิทยานิพนธ์ฉบับนี้ ตั้งแต่เริ่มพัฒนาหัวข้อวิทยานิพนธ์จนกระทั่งการจัดทำรูปเล่มวิทยานิพนธ์ฉบับสมบูรณ์ และขอกราบขอบพระคุณประธานกรรมการกรรมการสอบวิทยานิพนธ์ ศาสตราจารย์ ดร.ศิริชัย กาญจนวาสี และคณะกรรมการสอบวิทยานิพนธ์ รองศาสตราจารย์ ดร.โชติกา ภาษีผล รองศาสตราจารย์ ดร.ณัฏฐภรณ์ หลาวทอง รองศาสตราจารย์ ดร.สังวรณ์ งัดกระโทก สำหรับการตรวจสอบความถูกต้อง และให้คำแนะนำและข้อเสนอแนะที่เป็นประโยชน์แก่ผู้วิจัย

ขอกราบขอบพระคุณคณาจารย์สาขาการวัดและประเมินผลการศึกษา และคณาจารย์ภาควิชาวิจัยและจิตวิทยาการศึกษาทุกท่านที่ประสิทธิ์ประสาทความรู้ด้านการวิจัย สถิติ และการวัดและประเมินผล ทำให้ผู้วิจัยสามารถนำความรู้ต่าง ๆ ที่ได้รับจากการศึกษามาใช้ประโยชน์ในการทำวิทยานิพนธ์

ขอขอบคุณ บุษยารัตน์ จันทร์ประเสริฐ ที่อนุเคราะห์ข้อมูลตัวอย่างที่ใช้ในการศึกษาครั้งนี้ และขอบคุณเพื่อน พี่ น้อง นิสิตสาขาวิชาการวัดและประเมินผลทุกคนที่ได้ร่วมเรียนและให้กำลังใจแก่ผู้วิจัย

ขอขอบคุณพี่ต่าย พี่ป้อง และเจ้าหน้าที่คณะครุศาสตร์ทุกท่านที่ได้อำนวยความสะดวกแก่ผู้วิจัยในการดำเนินการด้านเอกสารต่าง ๆ ตลอดการเรียนในหลักสูตร

ขอขอบคุณมหาวิทยาลัยมหิดลและคณะศิลปศาสตร์ สำหรับการสนับสนุนทุนการศึกษาในการศึกษาครั้งนี้ และขอบคุณเจ้าหน้าที่ที่เกี่ยวข้องทุกท่านที่ช่วยดำเนินการและอำนวยความสะดวกเรื่องการลาศึกษาต่อของผู้วิจัย ขอบคุณอาจารย์ ดร.อัญชลี ภู่ผะกา และเพื่อนคณาจารย์สาขาวิชาภาษาไทยทุกคน สำหรับกำลังใจที่ดีตลอดการลาศึกษา

ศิธรา จุฑารัตน์

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ค
บทคัดย่อภาษาอังกฤษ ..... ง
กิตติกรรมประกาศ ..... จ
สารบัญ ..... ฉ
สารบัญตาราง ..... ญ
สารบัญรูป ..... ฎ
บทที่ 1 บทนำ 1
ความเป็นมาและความสำคัญของปัญหา 1
คำถามวิจัย 7
วัตถุประสงค์การวิจัย 8
สมมติฐานการวิจัย 8
ขอบเขตการวิจัย 9
คำจำกัดความในการวิจัย 13
ประโยชน์ที่ได้รับ 14
ประโยชน์ทางวิชาการ 14
ประโยชน์ทางปฏิบัติ 14
บทที่ 2 เอกสารและงานวิจัยที่เกี่ยวข้อง 16
ตอนที่ 1 ความเที่ยงระหว่างผู้ประเมิน 17
1.1 ความหมายและความสำคัญของความเที่ยงระหว่างผู้ประเมิน 17
1.2 ประเภทของความเที่ยงระหว่างผู้ประเมิน 20
1.3 วิธีการตรวจสอบความเที่ยงระหว่างผู้ประเมิน 24

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ตอนที่ 2 การทำหน้าที่ต่างกันของผู้ประเมิน 45
2.1 ความหมายของการทำหน้าที่ต่างกันของผู้ประเมิน 45
2.2 การตรวจสอบการทำหน้าที่ต่างกันของผู้ประเมิน 46
ตอนที่ 3 ทฤษฎีฉันทามติเชิงวัฒนธรรม 50
จุดเริ่มต้นและความหมายของทฤษฎีฉันทามติเชิงวัฒนธรรม 51
ข้อตกลงเบื้องต้นของทฤษฎีฉันทามติเชิงวัฒนธรรม 53
โมเดลทางสถิติและวิธีการประมาณค่าของการวิเคราะห์ฉันทามติเชิงวัฒนธรรม 54
ซอฟท์แวร์ทางสถิติสำหรับการวิเคราะห์โมเดลฉันทามติเชิงวัฒนธรรม 74
ตอนที่ 4 การศึกษาความสอดคล้องในแนวเดียวกัน 83
4.1 ความหมายและความสำคัญของการศึกษาความสอดคล้องในแนวเดียวกัน 85
4.2 วิธีการศึกษาความสอดคล้องในแนวเดียวกัน 87
กรอบแนวคิดในการวิจัย 96
บทที่ 3 วิธีดำเนินการวิจัย 97
ระยะที่ 1 การศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง 98
ระยะที่ 2 การศึกษาผลการประมาณความสอดคล้องระหว่างผู้ประเมินด้วยการจำลองแบบมอนติคาร์โล 98
ระยะที่ 3 ศึกษาผลการประยุกต์ใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมในการวิเคราะห์ความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบ ในการประเมินระดับชั้นเรียนกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น 110
บทที่ 4 ผลการวิเคราะห์ข้อมูล 112
ตอนที่ 1 ผลการพัฒนากระบวนการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินสำหรับการวิเคราะห์ความสอดคล้องในแนวเดียวกันทางการศึกษาด้วยโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม 113
กรอบแนวคิดของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม 113
พารามิเตอร์ในการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินด้วยโมเดล MC-GCM 114

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

พารามิเตอร์ในการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินด้วยโมเดล MC-LTRM 116
ตอนที่ 2 ผลการวิเคราะห์ประสิทธิภาพของการประมาณค่าของโมเดลจากข้อมูลจำลอง 120
2.1 ประสิทธิภาพของการประมาณค่าของโมเดล MC-GCM 121
2.2 ประสิทธิภาพของการประมาณค่าโมเดล MC-LTRM 127
2.3 ผลการวิเคราะห์ขนาดอิทธิพลของตัวแปรอิสระที่ส่งผลต่อการประมาณค่าของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม 131
ตอนที่ 3 การวิเคราะห์ความเที่ยงระหว่างผู้ประเมินโดยใช้ข้อมูลจริง 135
ผลการตรวจสอบจำนวนกลุ่มวัฒนธรรม 135
ผลการวิเคราะห์ข้อมูล 137
ผลการวิเคราะห์ผลการประเมินจากแบบบันทึกระดับความซับซ้อนทางปัญญาของข้อสอบในการประเมินระดับชาติ กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับชั้นมัธยมศึกษาปีที่ 3 138
ผลการวิเคราะห์ผลการประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียน กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น 143
บทที่ 5 สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ 148
สรุปผลการวิจัย 149
อภิปรายผลการวิจัย 156
ข้อเสนอแนะ 159
บรรณานุกรม 161
ภาคผนวก 172
ผลการวิเคราะห์การประเมินจากแบบบันทึกระดับความซับซ้อนทางปัญญาของข้อสอบใน การประเมินระดับชาติ กลุ่มสาระการเรียนรู้วิทยาศาสตร์ 173
ผลการวิเคราะห์ผลการประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียน กลุ่มสาระการเรียนรู้วิทยาศาสตร์ 178

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

กระบวนการวิเคราะห์ความสอดคล้องในแนวเดียวกัน ด้วยโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม..... 183

ประวัติผู้เขียน..... 185

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

สารบัญตาราง

หน้า

ตาราง 2.1 เกณฑ์การพิจารณาความสอดคล้องของสัมประสิทธิ์แคปปา 28
ตาราง 2.2 เกณฑ์การพิจารณาความสอดคล้องของสัมประสิทธิ์แคปปาของ 29
ตาราง 2.3 แหล่งความผันแปรและองค์ประกอบความแปรปรวนของการวัดหนึ่งฟาเซท 36
ตาราง 2.4 ขนาดตัวอย่างสำหรับการวิเคราะห์โดยทฤษฎีการตอบสนองข้อสอบ 41
ตาราง 2.5 ขนาดตัวอย่างสำหรับการวิเคราะห์โดยทฤษฎีการตอบสนองข้อสอบ 41
ตาราง 2.6 ผลการวิเคราะห์ค่าไอเกนกลุ่มวัฒนธรรมของผู้ประเมิน โมเดล MC-LTRM 80
ตาราง 2.7 การเปรียบเทียบวิธีการศึกษาความสอดคล้องในแนวเดียวกัน 93
ตาราง 3.1 ค่าสัมประสิทธิ์สหสัมพันธ์และค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสองของการประมาณค่าฉันทามติเชิงวัฒนธรรมด้วยโมเดล MC-GCM กรณีไม่มีการทำหน้าที่ต่างกันของผู้ประเมิน 103
ตาราง 3.2 ค่าสัมประสิทธิ์สหสัมพันธ์และค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสองของการประมาณค่าฉันทามติเชิงวัฒนธรรมด้วยโมเดล MC-GCM กรณีมีการทำหน้าที่ต่างกันของผู้ประเมิน 104
ตาราง 3.3 ค่าสัมประสิทธิ์สหสัมพันธ์และ MSE ของการประมาณค่าฉันทามติเชิงวัฒนธรรมด้วยโมเดล LTRM กรณีที่ไม่มีการทำหน้าที่ต่างกันของผู้ประเมิน 106
ตาราง 3.4 ค่าสัมประสิทธิ์สหสัมพันธ์และ MSE ของการประมาณค่าฉันทามติเชิงวัฒนธรรมด้วยโมเดล LTRM กรณีมีการทำหน้าที่ต่างกันของผู้ประเมิน 106
ตาราง 4.1 ผลการประมาณค่าพารามิเตอร์โดยโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม MC-GCM กรณีไม่มีการทำหน้าที่ต่างกันของผู้ประเมิน 126
ตาราง 4.2 ผลการประมาณค่าพารามิเตอร์โดยโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม MC-GCM กรณีมีการทำหน้าที่ต่างกันของผู้ประเมิน 126
ตาราง 4.3 ผลการประมาณค่าพารามิเตอร์ฉันทามติเชิงวัฒนธรรมโดยโมเดล MC-LTRM กรณีไม่มีการทำหน้าที่ต่างกันของผู้ประเมิน 130
ตาราง 4.4 ผลการประมาณค่าพารามิเตอร์ฉันทามติเชิงวัฒนธรรมโดยโมเดล MC-LTRM กรณีมีการทำหน้าที่ต่างกันของผู้ประเมิน 131
ตาราง 4.5 ผลการวิเคราะห์ขนาดอิทธิพลของตัวแปรอิสระที่ส่งผลต่อการประมาณค่าของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม 134

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ตาราง 4.6 ผลการประมาณค่าพารามิเตอร์ตำแหน่งคะแนนการประเมินระดับความซับซ้อนทางปัญญาของข้อสอบ (Tk) 138
ตาราง 4.7 ผลการประมาณค่าพารามิเตอร์เทรชโฮลด์ร่วมระหว่างผู้ประเมินระดับความซับซ้อนทางปัญญาของข้อสอบ (γc) 139
ตาราง 4.8 ผลการประมาณค่าพารามิเตอร์ตำแหน่งคะแนนการประเมินความสอดคล้องในแนวเดียวกัน (Tk) 144
ตาราง 4.9 ผลการประมาณค่าพารามิเตอร์เทรชโฮลด์ร่วมระหว่างผู้ประเมินระดับความสอดคล้องในแนวเดียวกัน (γc) 145

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

สารบัญรูป

หน้า

รูป 2.1 ปริมาณความถูกต้องของข้อมูลเทียบกับสัดส่วนความสอดคล้อง 29
รูป 2.2 ผลรวมความสอดคล้องและความแตกต่างของการประเมิน 31
รูป 2. 3 ผลการวิเคราะห์ Bland-Altman plots 32
รูป 2.4 ความแตกต่างของเกณฑ์การประเมินระหว่างผู้ตรวจ 34
รูป 2.5 การกำหนดพารามิเตอร์ในโมเดล GCM 60
รูป 2.6 การตรวจสอบการประมาณค่าจำนวนกลุ่มฉันทามติของโมเดล MC-GCM 65
รูป 2.7 ผลการประมาณค่าพารามิเตอร์ของโมเดล MC-GCM 66
รูป 2.8 ผลการประมาณค่าสมาชิกกลุ่มวัฒนธรรมและรูปแบบการตอบคำถามของโมเดล MC-GCM67
รูป 2.9 โมเดล LTRM แบบไม่มี bias (ซ้าย) และแบบมี bias (ขวา) 71
รูป 2.10 หน้าต่างวิเคราะห์ข้อมูลจากชุดคำสั่ง CCTpack 76
รูป 2.11 ผลการวิเคราะห์องค์ประกอบทางวัฒนธรรมของผู้ประเมิน 77
รูป 2.12 การแจกแจงความน่าจะเป็นภายหลังของค่าพารามิเตอร์ของโมเดล MC-LTRM 78
รูป 2.13 ผลการวิเคราะห์องค์ประกอบทางวัฒนธรรมของผู้ประเมิน 79
รูป 2.14 การแจกแจงความน่าจะเป็นภายหลังของค่าพารามิเตอร์ของโมเดล MC-LTRM 81
รูป 2.15 กรอบแนวคิดในการวิจัย 96
รูป 3.1 ขั้นตอนการจำลองข้อมูล 101
รูป 3.2 ความเบี่ยงเบนของค่าประมาณความยากของรายการประเมิน (δi) 103
รูป 3.3 ค่า MSE ของการประมาณค่าด้วยโมเดล MC-GCM เมื่อกำหนดให้ไม่มีการทำหน้าที่ต่างกันของผู้ประเมิน 103
รูป 3.4 ค่า MSE ของการประมาณค่าด้วยโมเดล MC-GCM เมื่อกำหนดให้มีการทำหน้าที่ต่างกันของผู้ประเมิน 105
รูป 3.5 สัมประสิทธิ์สหสัมพันธ์ระหว่างค่าจริงกับค่าประมาณของโมเดล 107
รูป 3.6 ความลำเอียงในการประมาณค่ามารามิเตอร์ของโมเดล 108
รูป 3.7 การวิเคราะห์ข้อมูลจำลอง 110
รูป 4.1 กรอบแนวคิดของโมเดล GCM 115
รูป 4.2 กรอบแนวคิดของโมเดล LTRM 117
รูป 4.3 ขั้นตอนการตรวจสอบความสอดคล้องระหว่างผู้ประเมิน ในการวิเคราะห์ความสอดคล้องในแนวเดียวกัน 119

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

รูป 4.4 ผลการวิเคราะห์การทำหน้าที่ต่างกันระหว่างผู้ประเมิน 120
รูป 4.5 ผลการประมาณค่าความสามารถของผู้ประเมินโดยโมเดล MC-GCM 122
รูป 4.6 ผลการประมาณค่าพารามิเตอร์ความลำเอียงในการประเมินโดยโมเดล MC-GCM 123
รูป 4.7 ผลการประมาณค่าพารามิเตอร์ความยากของรายการประเมินโดยโมเดล MC-GCM 124
รูป 4.8 ผลการประมาณค่าฉันทามติเชิงวัฒนธรรมระหว่างผู้ประเมิน 125
รูป 4.9 เปรียบเทียบการประมาณค่าระหว่างข้อมูลที่มี DRF กับไม่มี DRF 128
รูป 4.10 การเปรียบเทียบผลการประมาณค่าพารามิเตอร์ ai , bi ระหว่างกรณีมี DRF กับไม่มี DRF 130
รูป 4.11 อิทธิพลของจำนวนผู้ประเมินส่งผลต่อความลำเอียงในการประมาณค่า 132
รูป 4.12 อิทธิพลของจำนวนข้อคำถามประเมินส่งผลต่อค่า MSE และ correlation 133
รูป 4.13 การทำหน้าที่ต่างกันของผู้ประเมิน ส่งผลต่อค่า MSE และ correlation ในโมเดล MC-GCM 133
รูป 4.14 การทำหน้าที่ต่างกันระหว่าผู้ประเมิน ส่งผลต่อค่า MSE และ correlation ในโมเดล MC-LTRM 134
รูป 4.15 จำนวนองค์ประกอบที่แสดงจำนวนกลุ่มวัฒนธรรมของผู้ประเมิน 136
รูป 4.16 ผลการวิเคราะห์การทำหน้าที่ระหว่างผู้ประเมินด้วยโมเดล LTRM 137
รูป 4.17 ค่าเฉลี่ยภายหลังการประมาณค่าพารามิเตอร์ Tk และ γc 140
รูป 4.18 ค่าเฉลี่ยการแจกแจงภายหลังของการประมาณค่าพารามิเตอร์ λk 141
รูป 4.19 ค่าเฉลี่ยการแจกแจงภายหลังของการประมาณค่าพารามิเตอร์ Ei 141
รูป 4.20 ค่าเฉลี่ยการแจกแจงภายหลังของการประมาณค่าพารามิเตอร์ ai และ bi 142
รูป 4.21 ค่าเฉลี่ยภายหลังการประมาณค่าพารามิเตอร์ Tk และ γc 145
รูป 4.22 ค่าเฉลี่ยการแจกแจงภายหลังของการประมาณค่าพารามิเตอร์ λk 146
รูป 4.23 ค่าเฉลี่ยการแจกแจงภายหลังของการประมาณค่าพารามิเตอร์ Ei 146
รูป 4.24 ค่าเฉลี่ยการแจกแจงภายหลังของการประมาณค่าพารามิเตอร์ ai และ bi 147

บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

ความสอดคล้องทางการศึกษาเป็นคุณสมบัติสำคัญของระบบการศึกษา ความสอดคล้องระหว่างมาตรฐานของหลักสูตร การจัดการเรียนรู้ และการวัดการประเมินผลเป็นหลักฐานบ่งชี้ของความตรงของการวัดและประเมินผล หรืออีกนัยหนึ่ง คือ ระดับความแม่นยำของการแปลความหมายของการทดสอบทางการศึกษา ทั้งนี้ การวัดและประเมินผลทางการศึกษาเป็นการให้สารสนเทศเกี่ยวกับการที่ผู้เรียนได้รับการพัฒนาความรู้และทักษะที่คาดหวังเพียงใด เมื่อปัจจัยด้านหลักสูตร การเรียนการสอน และการวัดและประเมินผลมีความสอดคล้องกันก็จะนำไปสู่ประสิทธิภาพในการจัดการศึกษาและโอกาสที่ผู้เรียนจะได้รับการจัดประสบการณ์การเรียนรู้ที่คาดหวัง ผลการวิเคราะห์ความสอดคล้องทางการศึกษาจึงสะท้อนให้ผู้เกี่ยวข้องกับหลักสูตรและการจัดการเรียนการสอนทราบถึงปฏิสัมพันธ์ระหว่างมาตรฐานการเรียนรู้กับการวัดและประเมินผลทางการศึกษา อันนำไปสู่การกำหนดนโยบายทางการศึกษาระดับประเทศ

การวิเคราะห์ความสอดคล้องในแนวเดียวกันทางการศึกษาที่ใช้ในปัจจุบันมีอยู่ด้วยกัน 3 รูปแบบ คือ 1) Webb Model (Webb, 1997a; 1997b; 1999) ประกอบด้วยขั้นตอนการศึกษา 2 ขั้นตอน คือ การใช้ผู้เชียวชาญลงรหัสระดับความลึกซึ้งของความรู้ จากนั้น ให้ผู้เชี่ยวชาญลงรหัสระดับของความลึกซึ้งของความรู้ในข้อสอบวัดความรู้มาตรฐานหลักสูตรและวัตถุประสงค์การเรียนรู้ 2) Surveys of Enacted Curriculum (SEC) Model (Porter และ Smithson, 2002) ศึกษาความสอดคล้องในเชิงปริมาณใน 3 มิติองค์ประกอบทางการศึกษา ได้แก่ ความสอดคล้องของเน้อหาสมรรถนะที่คาดหวังของผู้เรียน และเนื้อหาในการสอน โดยมีขั้นตอนการศึกษา คือให้ผู้เชี่ยวชาญลงรหัสในตารางประเมินความสอดคล้องแล้วนำคะแนนจากการประเมินของผู้เชี่ยวชาญมาคิดดัชนีความสอดคล้อง 3) Achieve Model พัฒนาขึ้นโดย Achieve, Inc. (CCSSO, 2002) ประกอบด้วยเกณฑ์ 5 เกณฑ์ คือ ความเหมาะสมด้านเนื้อหา (Content centrality) ความเหมาะสมด้านสมรรถภาพ (Performance centrality) ความท้าทาย (Challenge) ความสมดุล (Balance) และการกระจายของเนื้อหา (Range)

รูปแบบการวิเคราะห์ความสอดคล้องทั้ง 3 รูปแบบ ประกอบด้วยวิธีการในการวิเคราะห์ด้วยกัน 3 วิธี คือ 1) Sequential development เป็นการพัฒนามาตรฐานและการวัดและประเมินผลขึ้นตามลำดับ โดยพัฒนามาตรฐานการเรียนรู้ขึ้นมาก่อนเพื่อใช้เป็นแม่แบบในการกำหนดโครงสร้างและวิธีการวัดและประเมินผลในลำดับถัดไป 2) Expert review เป็นวิธีการใช้ความคิดเห็นจากผู้เชี่ยวชาญเกี่ยวกับเนื้อหาในกรอบมาตรฐานการเรียนรู้และการวัดและประเมินผล วิธีการนี้จะใช้วิเคราะห์ความสอดคล้องระหว่างการประเมินผลและมาตรฐานที่ถูกสร้างขึ้นมาแล้ว หรืออาจใช้เป็นส่วนหนึ่งของการพัฒนามาตรฐานและการประเมินผลก็ได้ และ 3) Document analysis เป็นการวิเคราะห์เอกสารด้วยวิธีการเข้ารหัส ใช้ในการวิเคราะห์ความสอดคล้องในแนวเดียวกันที่ซับซ้อนหรือการวิเคราะห์ความสอดคล้องของหลักสูตรระหว่างประเทศ

เนื่องจากการวิเคราะห์ความสอดคล้องในแนวเดียวกันเป็นการวิเคราะห์ความสอดคล้องกันระหว่างมาตรฐานการศึกษากับการวัดและประเมินผล ซึ่งจำเป็นต้องใช้ความคิดเห็นของผู้เชี่ยวชาญ ดังนั้นการตรวจสอบความน่าเชื่อถือในการประเมินของผู้เชี่ยวชาญจึงเป็นหนึ่งในกระบวนการสำคัญที่จะต้องดำเนินการในการวิเคราะห์ความสอดคล้องในแนวเดียวกันทุกรูปแบบ โดยการวิเคราะห์ความน่าเชื่อถือในการประเมินของผู้เชี่ยวชาญหรือผู้ประเมินนี้จำแนกได้เป็นสองประเภท ได้แก่ การประเมินความสอดคล้องภายในตัวผู้ประเมิน ที่เป็นการประเมินคุณสมบัติความคงที่ของการใช้เครื่องมือโดยผู้ประเมินกลุ่มเดิมข้ามช่วงเวลา และ การประเมินความสอดคล้องระหว่างผู้ประเมินที่เป็นการประเมินคุณสมบัติความคงที่ของการใช้เครื่องมือระหว่างผู้ประเมินหลายคน (Stephens, Vos, Stevens และ Moore, 2006) อย่างไรก็ตาม จากการศึกษาเกี่ยวกับการวิเคราะห์ความเที่ยงระหว่างผู้ประเมิน พบว่า วิธีการวิเคราะห์ความเที่ยงระหว่างผู้ประเมินที่กล่าวมามีข้อสังเกตเกี่ยวกับการใช้งานและสารสนเทศที่ได้ สรุปได้ดังต่อไปนี้

ประการแรก สถิติบางตัวไม่สามารถใช้กับการวิเคราะห์ข้อมูลได้ทุกระดับ ยกตัวอย่างเช่น การหาสัดส่วนความสอดคล้อง (percentage agreement) สัมประสิทธิ์แคปปา (Kappa Coefficient) และ Many-facets Rasch model (MFRM) เป็นวิธีการที่ออกแบบมาให้ใช้ได้กับข้อมูลแบบจัดประเภท แต่ไม่รองรับการวิเคราะห์กับข้อมูลแบบต่อเนื่อง (Linacre, 1994; Stemler, 2004; Farrokhi, Esfandiari และ Schaefer, 2012; Salzberger, 2010)

ประการที่สอง สถิติบางตัวมีความไวต่อความลำเอียงของผู้ประเมิน เช่น สัมประสิทธิ์แคปปา (Kappa coefficient) แม้จะเป็นสถิติที่มีการปรับแก้ปัญหาการประมาณค่าเกินจริงของคำนวณสัดส่วนความสอดคล้อง (percentage agreement) แต่ สัมประสิทธิ์แคปปายังเป็นสถิติที่อ่อนไหวต่อ

ความลำเอียงระหว่างผู้ประเมินรวมถึงความถี่ของการให้คะแนน ทั้งนี้ การแจกแจงที่ไม่สมมาตรหรือลำเอียงจะส่งผลให้ค่าสัมประสิทธิ์แคปปาสูงเกินจริง (Byrt, Bishop และ Carlin, 1993) นอกจากนี้สัมประสิทธิ์แคปปามีปัญหาในการตีความจากสารสนเทศที่ได้รับ (Andres & Marzo, 2004; Stemler, 2004; Kottner และคณะ, 2011; Gisev และคณะ, 2013 โดย McHugh (2012) เสนอว่าควรตีความสัมประสิทธิ์แคปปาเป็นปริมาณความแปรปรวนภายในตัวแปรตามที่อธิบายได้ด้วยตัวแปรอิสระแต่ก็มีผู้เสนอเกณฑอื่น ๆ สำหรับแปลผลค่าสัมประสิทธิ์แคปปาไว้หลายเกณฑ์ (Landis & Koch, 1977; Krippendorff, 1980; Fleiss Levin & Paik, 2003; McHugh, 2012) ซึ่งก่อให้เกิดปัญหาความเป็นปรนัยในการตีความหมายและใช้สารสนเทศจากค่าสัมประสิทธิ์แคปปา

ประการที่สาม การวิเคราะห์ทางสถิติบางประเภทมีข้อตกลงเกี่ยวกับความแปรปรวนและการแจกแจงของข้อมูล เช่น การวิเคราะห์ Intraclass Correlation Coefficient (ICC) ซึ่งใช้การวิเคราะห์ความแปรปรวน (ANOVA) ในการประมาณค่าสถิติ จุดเด่นของ ICC คือ สามารถวิเคราะห์ความเที่ยงระหว่างผู้ประเมินหลายคนโดยเป็นการวัดความสอดคล้องกัน (agreement) ของการใช้เครื่องมือระหว่างผู้ประเมิน ซึ่งดีกว่าการใช้สัมประสิทธิ์สหสัมพันธ์ (correlation coefficients) และสัมประสิทธิ์อัลฟาที่เป็นเพียงการวัดความคงที่ (consistency) ของการใช้เครื่องมือระหว่างผู้ประเมิน และสามารถวิเคราะห์ความเที่ยงระหว่างผู้ประเมินได้ทีละสองคนเท่านั้น นอกจากนี้ยังมีความแกร่งต่อการเกิดค่าสูญหายมากกว่าการใช้สัมประสิทธิ์สหสัมพันธ์ หรือสัมประสิทธิ์อัลฟาในกรณีที่ค่าสูญหายดังกล่าวเป็นการสูญหายแบบสุ่มสมบูรณ์ (missing completely at random: MCAR) และยังสามารถวิเคราะห์ตัวแปรได้ทุกระดับ (Gisev และคณะ, 2013) อย่างไรก็ตาม เนื่องจากการวิเคราะห์ ICC อยู่บนพื้นฐานของการวิเคราะห์ความแปรปรวน หากข้อมูลในการวิเคราะห์มาจากกลุ่มตัวอย่างที่มีลักษณะเป็นวิวิธพันธ์ (heterogeneous) สูง อาจทำให้การประมาณส่วนประกอบความแปรปรวนของตัวแปรมีความคลาดเคลื่อนและส่งผลให้ค่า ICC ผิดพลาดและไม่สามารถใช้สรุปสารสนเทศเกี่ยวกับความเที่ยงระหว่างผู้ประเมินได้ (Muller และคณะ, 1994; Costa-Santos และคณะ, 2009)

ประการที่สี่ วิธีการทางสถิติส่วนใหญ่มักให้สารสนเทศเพียงด้านเดียว เช่น ให้สารสนเทศเกี่ยวกับความสอดคล้อง (agreement) หรือความคงที่ (consistency) ของการใช้เครื่องมือระหว่างผู้ประเมิน และไม่ได้นำปัจจัยเกี่ยวกับพฤติกรรมการตรวจของผู้ประเมิน หรือปัจจัยอื่น ๆ ที่เกี่ยวข้องกับการตรวจให้คะแนนมาวิเคราะห์ร่วมด้วย ปัจจัยดังกล่าวเป็นสาเหตุของการทำหน้าที่ต่างกันของผู้ประเมิน (differential rater functioning: DRF) ซึ่งทำให้ผู้ประเมินมีรูปแบบการให้คะแนนที่

แตกต่างกันในการประเมินเนื้อหาเดียวกัน การวิเคราะห์เพื่อตรวจสอบความแตกต่างของการตอบสนองของผู้ประเมินที่มีต่อเนื้อหา และการแก้ปัญหาการประเมินความเที่ยงระหว่างผู้ประเมินภายใต้สถานการณ์ที่เกิดการทำหน้าที่ต่างกันของผู้ประเมินจึงเป็นกระบวนการที่มีความจำเป็นในการวิเคราะห์ความเที่ยงระหว่างผู้ประเมินให้มีความถูกต้อง (Stemler, 2004; Gisev และคณะ, 2013) จากประเด็นปัญหาดังกล่าวจึงมีการพัฒนาวิธีการทางสถิติในการตรวจสอบและประมาณค่าความเที่ยงระหว่างผู้ประเมินภายใต้สถานการณ์ที่มีการทำหน้าที่ต่างกันของผู้ประเมิน วิธีการที่เป็นที่นิยมและพบว่ามีการใช้บ่อยในการวิจัยต่าง ๆ คือ การวิเคราะห์ Many Facet Rasch Model (MFRM) (Schaefer, 2008; Muckle และ Karabatsos, 2009; Farrokhi, Esfandiani และ Schaefer, 2012; Myford และ Wolfe, 2009; Xun Yan, 2014; Engelhard Jr., Wind, Kobin และ Chajewski, 2013; Wesolowski, Wind และ Engelhard Jr., 2015) นอกจากนี้ยังมีการใช้วิธีการ Multilevel Analysis (Leckie และ Baird, 2011) และ Mantel-Haenzel method (Johanson และ Osborn, 2004) เป็นต้น

อย่างไรก็ตาม จากการศึกษางานวิจัยที่เกี่ยวข้องพบว่าวิธีการข้างต้นยังมีข้อจำกัดเกี่ยวกับขนาดตัวอย่างของการวิเคราะห์ MFRM โดย Farrokhi และคณะ (2012) เสนอว่า การวิเคราะห์ MFRM ควรใช้กลุ่มตัวอย่างที่มีขนาดใหญ่ในการวิเคราะห์ นอกจากนี้ การศึกษาของ Muckle และ Karabatsos (2009) ยังพบข้อจำกัดเกี่ยวกับข้อตกลงเบื้องต้นเกี่ยวกับการแจกแจงปกติของข้อมูลในกรณีที่การแจกแจงของประชากรไม่เป็นการแจกแจงปกติจะทำให้โมเดลมีความสอดคล้องกับข้อมูลต่ำ และการประมาณค่าประชากรของอิทธิพลสุ่มจะไม่แม่นยำ ซึ่งส่งผลเสียต่อการวิเคราะห์ข้อมูลเกี่ยวกับความเที่ยงของการประเมิน รวมถึงการคำนวณค่า p-value นั้นจะมีปัญหาความคลาดเคลื่อนแบบที่ 1 เฟ้อเมื่อทำการทดสอบซ้ำหลายครั้ง และการคำนวณ p-value แยกกันแต่ละพารามิเตอร์นั้นถือว่าพารามิเตอร์มีความแปรปรวนร่วมเป็น 0 ซึ่งเป็นข้อตกลงที่ไม่เหมาะสมอันนำไปสู่ข้อสรุปที่ผิดพลาด ทั้งนี้ Muckle และ Karabatsos (2009) ได้เสนอให้ใช้วิธีการประมาณค่าแบบเบส์ซึ่งสามารถกำหนดการแจกแจงความน่าจะเป็นก่อนหน้าได้ทุกรูปแบบ

จากข้อจำกัดต่าง ๆ ดังที่กล่าวมา ผู้วิจัยจึงได้ศึกษาวิธีการทางเลือกที่จะสามารถใช้เพื่อแก้ปัญหาหรือลดทอนข้อจำกัดต่าง ๆ ในข้างต้น และพบว่าทฤษฎีการวิเคราะห์ฉันทามติเชิงวัฒนธรรม (Cultural Consensus Theory) ซึ่งพัฒนาโดย A. Kimball Romney Susan C. Weller และ William H. Batchelder (1986) เป็นวิธีการวิเคราะห์ความสอดคล้องระหว่างผู้ให้ข้อมูล (informants) ที่สามารถนำมาประยุกต์ใช้ในการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินได้

การวิเคราะห์ฉันทามติเชิงวัฒนธรรม เป็นวิธีการวิเคราะห์ข้อมูลที่ได้จากการตอบแบบสอบถาม หรือการประเมินที่ให้สารสนเทศหลัก คือ “คำตอบเชิงวัฒนธรรม” (Cultural consensus answer) คำตอบเชิงวัฒนธรรมนี้เป็นคำตอบที่สะท้อนฉันทามติของกลุ่ม เปรียบเทียบได้กับผลคะแนนการประเมินที่สอดคล้องกันระหว่างผู้ประเมิน ซึ่งใช้เป็นข้อสรุปหรือข้อยุติของการประเมิน

การวิเคราะห์ฉันทามติเชิงวัฒนธรรมยังสามารถลดข้อจำกัดของวิธีการทางสถิติที่ใช้ในการประมาณค่าความสอดคล้องระหว่างผู้ประเมินที่มีอยู่ในปัจจุบัน ได้แก่ (1) สามารถรองรับข้อมูลของตัวแปรทั้งแบบจัดประเภท และแบบต่อเนื่อง (2) โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมสามารถจำแนกกลุ่มผู้ให้ข้อมูลตามความสอดคล้องของการให้ข้อมูล นอกจากนี้ยังมีการประมาณค่าพารามิเตอร์ความลำเอียงในการประเมิน พารามิเตอร์กลุ่มวัฒนธรรม และความน่าจะเป็นของการเป็นสมาชิกกลุ่มวัฒนธรรม สามารถจำแนกความสอดคล้องหรือฉันทามติของผู้ให้ข้อมูลตามตัวแปรต่าง ๆ ได้ เช่น ภูมิหลัง เพศ ศาสนา หรือความเชื่อทางการเมือง (Anders และ Batchelder, 2012) เป็นต้น (3) การวิเคราะห์ฉันทามติทางวัฒนธรรมประมาณค่าพารามิเตอร์ที่เกี่ยวกับคุณลักษณะของผู้ตอบ (หรือผู้ประเมิน) ได้แก่ ความลำเอียงในการเลือกตอบ (เช่น การเลือกให้คะแนนสูงหรือต่ำในมาตรประมาณค่าหรือรูบริค) แนวโน้มในการเลือกตอบ (เช่น ผู้ประเมินบางคนมักให้คะแนนระดับปานกลาง หรือพอใช้มากกว่าระดับอื่น) นอกจากนี้ยังสามารถวิเคราะห์ความยาก ในการให้คะแนนคำถามแต่ละข้อ ซึ่งค่าพารามิเตอร์ที่ได้จะสามารถอธิบายลักษณะการให้คะแนนของผู้ประเมินแต่ละคนหรือในภาพรวมได้ (4) การประมาณค่าพารามิเตอร์ในโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมเป็นการประมาณค่าแบบเบส์ ที่มีประสิทธิภาพสูงกว่าการประมาณแบบดั้งเดิมในกรณีที่โมเดลมีความซับซ้อนสูง สามารถให้ค่าประมาณที่มีความน่าเชื่อถือมากกว่าวิธีการแบบดั้งเดิมในกรณีที่ตัวอย่างมีขนาดเล็ก ซึ่งเหมาะสมกับการวิเคราะห์การตรวจให้คะแนนโดยผู้ประเมินหรือผู้เชี่ยวชาญหลายคน ที่มักใช้จำนวนผู้ประเมินหรือผู้เชี่ยวชาญจำนวนไม่มาก นอกจากนี้การสรุปอ้างอิงหรือตีความหมายพารามิเตอร์ของโมเดลใช้การแจกแจงความน่าจะเป็นภายหลัง (posterior distribution) ซึ่งทำให้การแปลผลการสรุปอ้างอิงสามารถทำได้อย่างตรงไปตรงมา และชัดเจนกว่าการใช้วิธีการแบบดั้งเดิม ซึ่งมีจุดเด่นที่ยอมให้นำสารสนเทศที่เกี่ยวข้องนอกเหนือจากข้อมูลเชิงประจักษ์มาใช้ประกอบการวิเคราะห์ด้วย (5) ผลการวิเคราะห์หลักส่วนหนึ่งของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมได้รับการพัฒนาให้อยู่ในรูปแบบทัศนภาพข้อมูลที่ช่วยให้ผู้วิเคราะห์สามารถแปลความหมายได้ง่ายและเห็นภาพชัดเจน

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมมีพารามิเตอร์หลักของแต่ละโมเดล ประกอบด้วย พารามิเตอร์คำตอบฉันทามติ ความสามารถของผู้ตอบ และความลำเอียงในการตอบ นอกจากนี้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมยังสามารถจำแนกออกได้เป็นหลายโมเดล โดยแต่ละโมเดลยังมีพารามิเตอร์เฉพาะที่ให้สารสนเทศที่แตกต่างกันเพื่อรองรับวัตถุประสงค์ในการวิเคราะห์ข้อมูลต่าง ๆ

จากคุณลักษณะดังกล่าว ผู้วิจัยจึงมีสมมติฐานว่า มีความเป็นไปได้ที่จะประยุกต์ใช้โมเดลฉันทามติเชิงวัฒนธรรมกับการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินสำหรับการวิเคราะห์ความสอดคล้องในแนวเดียวกันทางการศึกษาภายใต้สถานการณ์ที่มีการทำหน้าที่ต่างกันของผู้ประเมิน ทั้งนี้ จากการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องในปัจจุบันพบว่า ยังไม่มีการศึกษาใดที่มีการประยุกต์ใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมสำหรับสถานการณ์ในข้างต้น งานวิจัยนี้จึงทำการศึกษาการประยุกต์ใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมกับการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินสำหรับการวิเคราะห์ความสอดคล้องในแนวเดียวกันทางการศึกษา ด้วยเหตุนี้ เพื่อให้ได้องค์ความรู้เกี่ยวกับวิธีวิทยาใหม่ทางการวัดและประเมินผลทางการศึกษาในการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินสำหรับการวิเคราะห์ความสอดคล้องในแนวเดียวกันทางการศึกษา

นอกจากนี้ ประเด็นสำคัญอีกประการหนึ่ง คือ การตรวจสอบประสิทธิภาพของผลการวิเคราะห์ที่ได้จากโมเดลฉันทามติเชิงวัฒนธรรมที่ผู้วิจัยนำมาประยุกต์ใช้ในการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินสำหรับการวิเคราะห์ความสอดคล้องในแนวเดียวกันทางการศึกษา ผลการศึกษาในส่วนนี้จะใช้เพื่อยืนยันสมมุติฐานของผู้วิจัยเกี่ยวกับความเป็นไปได้ในการประยุกต์ใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมเพื่อแก้ปัญหาหรือลดทอนข้อจำกัดต่าง ๆ ในการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินสำหรับการวิเคราะห์ความสอดคล้องในแนวเดียวกันทางการศึกษา และยังให้สารสนเทศที่สามารถนำมาสร้างเป็นข้อเสนอแนะในการประยุกต์ใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมในบริบทดังกล่าวอีกด้วย

การดำเนินงานขั้นสุดท้ายของการวิจัยนี้เป็นการประยุกต์ใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมในการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินสำหรับการวิเคราะห์ความสอดคล้องในแนวเดียวกันทางการศึกษากับข้อมูลจริง โดยเป็นการวิเคราะห์ความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียนกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น จากการศึกษาของ บุษยารัตน์ จันทร์ประเสริฐ (2561) การศึกษาความสอดคล้อง

ในแนวเดียวกันในการประเมินระดับชาติกับการประเมินระดับชั้นเรียน ในกลุ่มสาระการเรียนรู้วิทยาศาสตร์ระดับมัธยมศึกษาตอนต้น ประกอบด้วยการศึกษาและการใช้สถิติในการวิเคราะห์ข้อมูลทั้งสิ้น 3 ขั้นตอนเพื่อที่จะได้สารสนเทศเกี่ยวกับความน่าเชื่อถือของผู้ประเมินและผลของการประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียนกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น ได้แก่ การตรวจสอบอิทธิพลของผู้ประเมิน การเปรียบเทียบผลการประเมินความสอดคล้องในแนวเดียวกันก่อนและหลังควบคุมอิทธิพลการกดหรือปล่อยคะแนน และการศึกษาความสอดคล้องในแนวเดียวกันตามแนวคิดของ Porter ทั้งนี้จากแนวคิดของการวิเคราะห์ฉันทามติเชิงวัฒนธรรม ผลการศึกษาที่ได้จากการวิเคราะห์ด้วยโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมสามารถให้สารสนเทศเกี่ยวกับอิทธิพลการกดหรือปล่อยคะแนนของผู้ประเมิน ความน่าเชื่อถือของผู้ประเมิน รวมถึงผลคะแนนฉันทามติในการประเมินได้พร้อมกันจากการวิเคราะห์ข้อมูลในคราวเดียว นอกจากนี้ การวิเคราะห์ด้วยโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมยังให้สารสนเทศในมิติอื่นที่แตกต่างไปจากการวิเคราะห์แบบดั้งเดิม ซึ่งช่วยให้สารสนเทศป้อนกลับแก่ผู้เกี่ยวข้อง ทั้งในระดับผู้บริหารทางการศึกษา หรือนักวัดประเมินผลทางการศึกษาที่เกี่ยวข้อง สำหรับอ้างอิงและสนับสนุนการตัดสินใจทางการศึกษาได้

คำถามวิจัย

1. การวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินในการศึกษาความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียนกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้นด้วยโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมมีขั้นตอนอย่างไรและให้สารสนเทศใดบ้าง และสารสนเทศดังกล่าวสนับสนุนการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินอย่างไร

2. ประสิทธิภาพของการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินในการศึกษาความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียนกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้นด้วยโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมเป็นอย่างไร

3. ผลการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินและการทำหน้าที่ต่างกันของผู้ประเมินในการประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมิน

ระดับชั้นเรียนกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้นด้วยการวิเคราะห์ฉันทามติเชิงวัฒนธรรมให้สารสนเทศอย่างไรบ้าง

วัตถุประสงค์การวิจัย

1. เพื่อประยุกต์ใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมเพื่อวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินในการศึกษาความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียนกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น

2. เพื่อตรวจสอบประสิทธิภาพของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมในการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินในการศึกษาความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียนกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น

3. เพื่อศึกษาผลการประยุกต์ใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมในการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินและการทำหน้าที่ต่างกันของผู้ประเมินในการประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียนกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น

สมมติฐานการวิจัย

การวิเคราะห์ฉันทามติเชิงวัฒนธรรมมีจุดเด่นเรื่องความเสถียรของการประมาณค่าในกลุ่มตัวอย่างที่มีจำนวนน้อย รวมถึงได้รับการทดสอบว่าโมเดลมีความเสถียรและสามารถประมาณค่าพารามิเตอร์ที่มีความแตกต่างกันของรูปแบบในการประเมิน (มีการทำหน้าที่ต่างกันระหว่างผู้ประเมิน) ได้อย่างมีประสิทธิภาพ (Romney, Weller และ Batchelder, 1986; Anders และ Batchelder, 2012; 2015; France และBatchelder, 2015) อย่างไรก็ตาม เนื่องจากยังมีเอกสารและงานวิจัยที่สนับสนุนในประเด็นดังกล่าวน้อย ผู้วิจัยจึงทำการศึกษาเพิ่มเติมเพื่อหาข้อสนับสนุนเกี่ยวกับปัจจัยที่ส่งผลต่อการประมาณค่าของโทเดล และได้กำหนดสมมติฐานในการวิจัยครั้งนี้ ดังนี้

1. ปัจจัยที่ส่งผลต่อประสิทธิภาพของการประมาณค่าของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม ได้แก่ จำนวนผู้ประเมิน จำนวนรายการประเมิน และการทำหน้าที่ต่างกันของผู้ประเมิน

2. โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมสามารถประมาณค่าพารามิเตอร์ที่เกี่ยวข้องกับการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินในการศึกษาความสอดคล้องในแนวเดียวกันได้อย่างมีประสิทธิภาพ (Anders และ Batchelder, 2012; 2015)

ขอบเขตการวิจัย

การศึกษาครั้งนี้ ผู้วิจัยสนใจศึกษาผลการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินโดยการประยุกต์ใช้การวิเคราะห์ฉันทามติเชิงวัฒนธรรม เพื่อศึกษาความเหมาะสมของโมเดล ประสิทธิภาพการวิเคราะห์ข้อมูล และสารสนเทศที่ได้จากการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมิน โดยแบ่งการศึกษาออกเป็น 3 ระยะ ดังนี้

ระยะที่ 1 การศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง ประกอบด้วย การศึกษาแนวคิดทฤษฎีเกี่ยวกับการวิเคราะห์ความเที่ยงระหว่างผู้ประเมิน การทำหน้าที่ต่างกันระหว่างผู้ประเมิน การวิเคราะห์ฉันทามติเชิงวัฒนธรรม และการศึกษาความสอดคล้องในแนวเดียวกันทางการศึกษา

ระยะที่ 2 การศึกษาจากสถานการณ์จำลองข้อมูลในการประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบ ผู้วิจัยกำหนดความเป็นไปได้ของสถานการณ์ในการประเมิน 2 แบบ คือ การประเมินที่มีการให้คะแนนแบบ 0, 1 กับการประเมินที่มีการให้คะแนนแบบมาตรประมาณค่า ดังนั้น ผู้วิจัยจึงเลือกศึกษาประสิทธิภาพการประมาณค่าพารามิเตอร์ของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม 2 โมเดลซึ่งเหมาะสมกับข้อมูลจากการประเมินข้างต้น ได้แก่

1. Multi-culture General Condorcet Model: MC-GCM ซึ่งเป็นโมเดลการวิเคราะห์ข้อมูลหรือการให้คะแนนแบบ (0,1) รายละเอียดของการกำหนดสถานการณ์จำลองของโมเดล MC-GCM มีดังนี้

1) กำหนดจำนวนผู้ประเมิน (N)ดังนี้ $N = (15,\ 30, 45)$

2) กำหนดจำนวนรายการประเมินความสอดคล้องในแนวเดียวกัน (M) ดังนี้ $M = (25,\ 55, 85)$

3) กำหนดเงื่อนไขการทำหน้าที่ต่างกันของผู้ประเมิน (T) โดย T_1 เป็นรายการประเมินที่มีความยากของคำถามประเมินเท่ากัน $(\delta_k = 0.5)$ และ T_2 เป็นรายการประเมินที่มีความยากของคำถามประเมินไม่เท่ากัน $(\delta_k \neq 0.5)$

4) สร้างคำตอบฉันทามติของผู้ประเมิน 2 กลุ่มขนาด $(Z_{tk})_{nxm}$ โดยการสุ่มจากการแจกแจงความน่าจะเป็นแบบเบอร์นูลี $Z_{tk} \sim Bernoulli\ (m, P_t)$ โดย $P_t \sim Uniform(0,1)$

5) สร้างค่าพารามิเตอร์กลุ่มวัฒนธรรม $e_i \sim Categorical(\lambda)$ โดย $\lambda \sim Dirichlet(L)$, $L = (1)_{1xT}$

6) กำหนดค่าความสามารถของผู้ประเมิน (θ_i) โดยการสุ่มจากการแจกแจงความน่าจะเป็นแบบเบต้า $\theta_i \sim Beta\ (n, 2, 2)$

7) กำหนดค่าความยากของคำถามประเมิน (δ_k) โดยการสุ่มจากการแจกแจงความน่าจะเป็นแบบเบต้า $\delta_k \sim Beta\ (n, 2, 2)$

8) กำหนดค่าความลำเอียงในการประเมิน (g_i) โดยการสุ่มจากการแจกแจงความน่าจะเป็นแบบเบต้า $g_i \sim Beta\ (n, 2, 2)$

9) สร้างเมตริกซ์คำตอบของการประเมิน X_{ik} จากค่าพารามิเตอร์ในข้อ 3 ถึง 8 จากสมการ (1.1) เมื่อกำหนดให้คำถามประเมินมีความยากเท่าเทียมกัน หรือสมการ (1.2) เมื่อกำหนดให้คำถามประเมินมีความยากแตกต่างกัน

$$Pr(X_{ik}) = \left(D_i Z_{k,e_i}\right) + ((1 - D_i) g_i) \tag{1.1}$$

$$D_{ik} = \frac{\theta_i(1 - \delta_k)}{\theta_i(1 - \delta_k) + \delta_k(1 - \theta_i)} \tag{1.2}$$

จากข้อ 1 ถึง 9 มีสถานการณ์จำลองข้อมูลทั้งหมด 3x3x2 = 18 สถานการณ์จำลองที่ใช้ในการศึกษาประสิทธิภาพการประมาณค่าฉันทามติเชิงวัฒนธรรมของผู้ประเมินด้วยโมเดล MC-GCM

2. Multi-culture Latent Truth Rater Model: MC-LTRM เป็นโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมสำหรับวิเคราะห์การประเมินการตอบสนองบนสเกลที่มากกว่า 2 ค่า เช่น การให้คะแนนแบบเรียงอันดับ (5 = มากที่สุด, 4 = มาก, 3 = ปานกลาง, 2 = น้อย, 1 = น้อยที่สุด) ซึ่งผู้ตอบมีลักษณะของกลุ่มวัฒนธรรมย่อย (subculture) ผู้วิจัยประยุกต์โมเดลดังกล่าวสำหรับการวิเคราะห์การประเมินที่ให้คะแนนบนมาตรประมาณค่า (Rating scale) ในการประเมินที่มีการทำหน้าที่ต่างกันของผู้ประเมิน มีรายละเอียดดังนี้

1) กำหนดจำนวนผู้ประเมิน (N) ดังนี้ $N = (15,\ 30, 45)$

2) กำหนดจำนวนรายการประเมินความสอดคล้องในแนวเดียวกัน (M) ดังนี้ $M = (25,\ 55, 85)$

3) กำหนดตำแหน่งของการตอบ (Item Location Value) T_{vk} โดยการสุ่มจากการแจกแจงแบบปกติ $T_{vk} \sim Normal\left(\mu_{T_v}, \tau_{T_v}\right)$ โดย v คือ จำนวนกลุ่มผู้ประเมิน มีค่า $V = 2$

4) กำหนดเทรชโฮลด์ร่วม (Shared Thresholds) (γ_c) โดย $\gamma_c \sim Normal\ (0, 0.1)$ โดย c คือจำนวนสเกลบนมาตรประมาณค่า มีค่า $C - 1$ โดย $C = \{c_1, \dots, c_5\}s$

5) กำหนดค่าความสามารถของผู้ประเมิน (E_i) โดยการสุ่มจากการแจกแจงความน่าจะเป็นแบบแกมมา $E_i \sim Gamma\ (\mu_E^2\tau_E, \mu_E\tau_E)$ โดย $\mu_E \sim Gamma\ (4,4), \tau_E \sim Gamma\ (4,4)$

6) กำหนดค่าความยากของคำถามประเมิน (λ_k) โดยการสุ่มจากการแจกแจงความน่าจะเป็นแบบแกมมา $\lambda_k \sim Gamma\left(\mu_\lambda^2\tau_\lambda, \mu_\lambda\tau_\lambda\right)$ โดย $\mu_\lambda = 1,\ \tau_\lambda \sim Gamma\ (1, 0.1)$

7) กำหนดค่าความลำเอียงในการประเมิน $(a_i,\ b_i)$ โดย $a_i \sim Gamma(\mu_a^2\tau_a, \mu_a\tau_a)$, และ $b_i \sim Normal\ (\mu_b, \tau_b)$

จากนั้น สร้างเมตริกซ์คำตอบของการประเมิน X_{ik} จากค่าพารามิเตอร์ในข้อ 2) ถึง 7) จากสมการ (1.3)

$$Pr(X_{ik} = x_{ik} \mid T_k, \lambda_k, \gamma_c, E_i, a_i, b_i) = \begin{cases} \Pr\left(Y_{ik} \le \delta_{i,x_{ik}}\right) & \text{หาก } x_{ik} = 1 \\ \Pr\left(\delta_{i,x_{ik}} - 1 < Y_{ik} \le \delta_{i,x_{ik}}\right) & \text{หาก } x_{ik} = c, 1 < c < C \\ \Pr\left(Y_{ik} > \delta_{i,x_{ik}-1}\right) & \text{หาก } x_{ik} = C \end{cases} \tag{1.3}$$

จากข้อ 1 ถึง 7 มีสถานการณ์จำลองข้อมูลทั้งหมด 3x3x2 = 18 สถานการณ์จำลองที่ใช้ในการศึกษาประสิทธิภาพการประมาณค่าฉันทามติเชิงวัฒนธรรมของผู้ประเมินด้วยโมเดล MC-LTRM

การศึกษาอิทธิพลที่ส่งผลต่อประสิทธิภาพของการประมาณค่าพารามิเตอร์ของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมในครั้งนี้ ผู้วิจัยกำหนดตัวแปรอิสระในการศึกษา คือ จำนวนผู้ประเมิน (N) จำนวนแบบสอบ (M) จำนวนกลุ่มการทำหน้าที่ต่างกันของผู้ประเมิน (T, V) ซึ่งตัวแปรดังกล่าวเป็นพารามิเตอร์หลักที่ส่งผลต่อการประมาณค่าของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมและเป็นตัวแปรที่มีการศึกษามากที่สุดในการพัฒนาโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม

ตัวแปรตามในการศึกษาครั้งนี้ คือ ประสิทธิภาพของการประมาณค่าพารามิเตอร์ของโมเดลฉันทามติเชิงวัฒนธรรมในสถานการณ์จำลองข้อมูล ได้แก่ ค่าสัมประสิทธิ์สหสัมพันธ์ (Pearson Correlation Coefficient) ใช้สำหรับหาความสัมพันธ์ระหว่างค่าจริงของพารามิเตอร์ที่ได้จากการสุ่มและค่าที่ได้จากการประมาณค่าของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม ค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Mean square error: MSE) ใช้สำหรับตรวจสอบความแตกต่างระหว่างตัวประมาณค่าและค่าประมาณที่ได้ ความลำเอียงในการประมาณค่า (Bias of Estimators) เป็นการพิจารณาความคงที่ของการประมาณค่าในแง่ของความแม่นยำของโมเดล

ระยะที่ 3 การศึกษาจากข้อมูลจริง ข้อมูลที่ใช้ในการศึกษาประสิทธิภาพของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมในการประมาณค่าความสอดคล้องระหว่างผู้ประเมินและการทำหน้าที่ต่างกันของผู้ประเมินในการศึกษาครั้งนี้ คือ ผลการประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียน กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น (บุษยรัตน์ จันทร์ประเสริฐ, 2560) ประกอบด้วย

1) ผลการประเมินจากแบบบันทึกระดับความซับซ้อนทางปัญญาของข้อสอบในการประเมินระดับชาติ กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับชั้นมัธยมศึกษาปีที่ 3 เป็นการประเมินลำดับขั้นของกระบวนการทางปัญญาจำแนกตามพฤติกรรมการเรียนรู้ด้านพุทธิพิสัยของ Bloom (Bloom’s Taxonomy) จำนวน 40 ตัวชี้วัด จาก 12 มาตรฐานการเรียนรู้ ผลการประเมินเป็นการจัดกลุ่มตัวบ่งชี้ตามพฤติกรรมการเรียนรู้ 6 กลุ่ม ได้แก่ จำ เข้าใจ ประยุกต์ใช้ วิเคราะห์ ประเมินค่า และสร้างสรรค์

2) ผลการประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียน กลุ่มสารการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น เป็นการประเมินความสอดคล้องระหว่างข้อสอบกับตัวชี้วัด จำนวน 40 ข้อ

สาเหตุที่ผู้วิจัยเลือกใช้ข้อมูลดังกล่าว เนื่องจาก การประเมินความสอดคล้องระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียนมีความสำคัญต่อการประเมินผลทางการศึกษาและจำเป็นต้องใช้ผู้ประเมินที่มีความเชี่ยวชาญด้านหลักสูตรและการวัดและประเมินผล ดังนั้น ผลการประเมินจากผู้เชี่ยวชาญควรเป็นผลการประเมินที่มีคุณภาพและสามารถใช้เป็นข้อมูลสำหรับอ้างอิงหรือตัดสินใจทางการศึกษาได้

คำจำกัดความในการวิจัย

คะแนนฉันทามติของการประเมิน (Cultural Consensus Score) หมายถึง ค่าเฉลี่ยความน่าจะเป็นของพารามิเตอร์ Z_k ในโมเดล MC-GCM และพารามิเตอร์ T_k ในโมเดลMC-LTRM ซึ่งเป็นความน่าจะเป็นของตำแหน่งคะแนนของรายการประเมินข้อที่ k ในการประเมินความสอดคล้องในแนวเดียวกัน

การทำหน้าที่ต่างกันของผู้ประเมิน (Differential Rater Functioning) หมายถึง จำนวนรูปแบบความสอดคล้องของผลการประเมินซึ่งได้จากการคำนวณค่าไอเกนของน้ำหนักองค์ประกอบ หากไม่มีการทำหน้าที่ต่างกันระหว่างผู้ประเมินการวิเคราะห์ scree plot จะมีค่าน้ำหนักองค์ประกอบเท่ากับ 1 องค์ประกอบ

ความสามารถของผู้ประเมิน (Rater competency) หมายถึง ระดับความแม่นยำในการระบุตำแหน่งคะแนนการประเมินที่สอดคล้องกับผลการประเมินของกลุ่มผู้ประเมิน โดย E_i จะมีค่าสูงเมื่อผู้ประเมินมีการประเมินที่สอดคล้องกับคำตอบของกลุ่มผู้ประเมิน

ความยากของคำถามประเมิน (Item difficulty) หมายถึง ความแปรปรวนของผลการประเมินข้ามกลุ่มผู้ประเมินในแต่ละข้อคำถามการประเมิน หากคำถามมีความยากไม่เท่าเทียมกัน ความแปรปรวนของเมตริกซ์การประเมิน (X_{ik}) จะสูง หากคำถามมีความยากเท่าเทียมกัน ความแปรปรวนของ X_{ik} จะต่ำ การคำนวณค่าความยากของคำถามประเมินในโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมจะใช้การคำนวณค่าสถิติ VDI (Batchelder และ Anders, 2012) โดย $VDI(X) = \sum_{k=1}^{M} V_k^2/M - (\sum_{k=1}^{M} V_k/M)^2$ เมื่อ $V_k = \sum_{i=1}^{N} X_{ik}^2/N - (\sum_{i=1}^{N} X_{ik}/N)^2$

ความลำเอียงในการประเมิน (Rating bias) หมายถึง ความคลาดเคลื่อนในการให้คะแนนของผู้ประเมิน i ในรายการประเมินข้อ k ความคลาดเคลื่อนดังกล่าวมี 2 ลักษณะ คือ พารามิเตอร์ a_i คือ ความกว้างของเทรชโฮลด์ของผู้ประเมิน หาก $a_i < 1$ แสดงว่าผู้ประเมินมีแนวโน้มที่จะเปลี่ยนระดับคะแนนได้ยากหรือมีความไวในการเปลี่ยนระดับการประเมินต่ำ ในขณะที่พารามิเตอร์ b_i คือ แนวโน้มการกดหรือปล่อยคะแนนในการประเมิน

ประสิทธิภาพของการประมาณค่า (Model performance) หมายถึง ความถูกต้อง (precision) และแม่นยำ (accuracy) ของการประมาณค่าความเที่ยงระหว่างผู้ประเมินโดยโมเดลฉันทามติเชิงวัฒนธรรม ความถูกต้องของการประมาณค่าพิจารณาจากค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (MSE) ความแม่นยำของการประมาณค่าพิจารณาจากค่าความลำเอียงในการประมาณค่า (Bias of estimators) และพิจารณาความสอดคล้อง

ระหว่างค่าพารามิเตอร์ที่กำหนดกับค่าพารามิเตอร์ที่ได้จากการประมาณค่าจาก ค่าสัมประสิทธิ์สหสัมพันธ์ (Pearson Correlation Coefficient)

ประโยชน์ที่ได้รับ

ประโยชน์ทางวิชาการ

ผลการศึกษาครั้งนี้เป็นการขยายองค์ความรู้เกี่ยวกับการประยุกต์ใช้การวิเคราะห์ฉันทามติเชิงวัฒนธรรมในการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินซึ่งเป็นขั้นตอนที่สำคัญในการศึกษาความสอดคล้องระหว่างผู้ประเมิน ทั้งนี้ จากการศึกษา พบว่า โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมสามารถนำไปประยุกต์ใช้ในการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินได้จริง และให้สารสนเทศที่เกี่ยวข้องกับการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินที่สำคัญ ได้แก่ การระบุการทำหน้าที่ต่างกันของผู้ประเมิน คะแนนฉันทามติในการประเมิน ความสามารถของผู้ประเมิน และความลำเอียงของผู้ประเมิน ทั้งนี้ จากการศึกษาวิธีการตรวจสอบความสอดคล้องระหว่างผู้ประเมินในการศึกษาความสอดคล้องในแนวเดียวกันทางการศึกษา ยังไม่พบวิธีการตรวจสอบความคล้องคล้องระหว่างผู้ประเมินที่ใช้เป็นมาตรฐานเดียวกันในการศึกษาความสอดคล้องในแนวเดียวกัน ผู้วิจัยจึงเสนอการโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมเพื่อเป็นส่วนหนึ่งของการตรวจสอบความสอดคล้องระหว่างผู้ประเมินในการตรวจสอบความสอดคล้องในแนวเดียวกันทางการศึกษา

ประโยชน์ทางปฏิบัติ

ผลการศึกษาครั้งนี้เสนอโมเดลทางเลือกสำหรับการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินในการศึกษาความสอดคล้องในแนวเดียวกันทางการศึกษา เนื่องจากโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมใช้การประมาณค่าแบบเบส์ซึ่งใช้กระบวนการของลูกโซ่มาร์คอฟ ทำให้สามารถลดจำนวนตัวอย่างลงได้มากกว่าการวิเคราะห์ด้วยสถิติแบบดั้งเดิม นอกจากผลการประมาณค่าของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมจะให้สารสนเทศที่สำคัญ คือ ผลคะแนนการประเมินที่เป็นฉันทามิขิงการประเมินแล้ว ยังให้สารสนเทศที่สามารถนำไปให้พิจารณาประกอบผลการประเมินของผู้เชี่ยวชาญ ได้แก่ สารสนเทศเกี่ยวกับความสามารถของผู้ประเมิน ใช้พิจารณาความน่าเชื่อถือของผู้ประเมินว่ามีความคลาดเคลื่อนในการประเมินต่างไปจากผู้เชี่ยวชาญคนอื่น ๆ หรือไม่อย่างไรอีกด้วย จากการศึกษาในครั้งนี้สรุปได้ว่า การวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินด้วยโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมสามารถนำมาประยุกต์ใช้กับการวิเคราะห์ความสอดคล้อง

ระหว่างผู้ประเมินในการศึกษาความสอดคล้องระหว่างข้อสอบกับตัวชี้วัดทางการศึกษา ทั้งการศึกษาความสอดคล้องในแนวเดียวกันของรายวิชาในระดับหลักสูตรสถานศึกษา หรือการศึกษาความสอดคล้องในแนวเดียวกันของข้อสอบกับตัวชี้วัดของหลักสูตรในการประเมินผลสัมฤทธิ์ทางการเรียนในระดับชาติ ซึ่งหากการศึกษาความสอดคล้องในแนวเดียวกันดังกล่าวมีการใช้ความคิดเห็นหรือผลการประเมินจากผู้เชี่ยวชาญเป็นส่วนหนึ่งของขั้นตอนการศึกษา ก็สามารถนำการวิเคราะห์ฉันทามติเชิงวัฒนธรรมไปประยุกต์ใช้ในการตรวจสอบความสอดคล้องระหว่างผู้ประเมิน รวมถึงวิเคราะห์ผลคะแนนฉันทามติในการประเมินความสอดคล้องได้

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

การศึกษาครั้งนี้ ผู้วิจัยได้ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องกับการวิเคราะห์ความเที่ยงระหว่างผู้ปะเมิน การทำหน้าที่ต่างกันของผู้ประเมิน และทฤษฎีฉันทามติเชิงวัฒนธรรม (Cultural Consensus Theory) โดยนำเสนอรายละเอียดเป็น 3 ตอน ดังนี้

ตอนที่ 1 ความเที่ยงระหว่างผู้ประเมิน

1.1 ความหมายและความสำคัญของความเที่ยงระหว่างผู้ประเมิน

1.2 ประเภทของความเที่ยงระหว่างผู้ประเมิน

1.3 วิธีการตรวจสอบความเที่ยงระหว่างผู้ประเมิน

ตอนที่ 2 การทำหน้าที่ต่างกันของผู้ประเมิน

2.1 ความหมายของการทำหน้าที่ต่างกันของผู้ประเมิน

2.2 วิธีการตรวจสอบการทำหน้าที่ต่างกันของผู้ประเมิน

ตอนที่ 3 ทฤษฎีฉันทามติเชิงวัฒนธรรม

3.1 จุดเริ่มต้นและความหมายของทฤษฎีฉันทามติเชิงวัฒนธรรม

3.2 ข้อตกลงเบื้องต้นของทฤษฎีฉันทามติเชิงวัฒนธรรม

3.3 โมเดลทางสถิติและวิธีการประมาณค่าของการวิเคราะห์ฉันทามติเชิงวัฒนธรรม

3.3.1 General Condorcet Model (GCM)

3.3.2 Multicultural General Condorcet Model (MC-GCM)

3.3.3 Latent Truth Rater Model (LTRM)

3.3.4 Multicultural Latent Truth Rater Model (MC-LTRM)

ตอนที่ 4 การศึกษาความสอดคล้องในแนวเดียวกัน

4.1 ความหมายและความสำคัญของการศึกษาความสอดคล้องในแนวเดียวกัน

4.2 วิธีการศึกษาความสอดคล้องในแนวเดียวกัน

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

ตอนที่ 1 ความเที่ยงระหว่างผู้ประเมิน

การประเมินผลโดยผู้ประเมินหลายคนเป็นวิธีการที่นำมาใช้ในการวัดและประเมินผลในหลายสถานการณ์ เช่น การประเมินการสอนของนิสิตโดยอาจารย์นิเทศ การประเมินทักษะการสื่อสาร การประเมินการประกวดผลงานของนักเรียน การประเมินงานเขียนโดยเพื่อนและครู เป็นต้น คุณภาพของผู้ประเมินและการประเมินมีบทบาทสำคัญต่อการแปลผลคะแนนการทดสอบ เนื่องจากเป็นตัวบ่งชี้ว่าผู้สอบมีคุณลักษณะทางจิตวิทยาตามที่มุ่งวัดจริงหรือไม่ นอกจากนี้ ผู้ประเมินยังเปรียบเสมือนแว่นขยายที่สะท้อนสมรรถนะของผู้สอบหรือผู้เรียนตามโครงสร้างด้วย (Thompson, Foster, Cole และ Dowding, 2005; Hogarth และ Karelaia, 2007; Engelhard, 2013) ถึงแม้ว่าโดยทั่วไปแล้วการทดสอบที่เป็นปรนัยจะมีความเที่ยงสูงกว่าการทดสอบที่เป็นอัตนัยเนื่องจากมีกระบวนการให้คะแนนที่ชัดเจนมากกว่า อย่างไรก็ตาม การทดสอบแบบปรนัยจะมีข้อจำกัดในการวัดคุณลักษณะหรือความสามารถของผู้รับการประเมินในขณะที่การทดสอบแบบอัตนัยจะส่งเสริมการแสดงความคิดเห็น ความคิดสร้างสรรค์ รวมถึงการประมวลและเรียบเรียงความรู้ การใช้ภาษาในการถ่ายทอดความรู้อย่างเป็นระบบ (ศิริชัย กาญจนวาสี, 2556) รวมถึงการแสดงความสามารถเชิงทักษะและการแก้ปัญหาเฉพาะหน้าของผู้ได้รับการประเมินได้เป็นอย่างดี ยกตัวอย่างเช่น ในกรณีของการสอบปฏิบัติที่ต้องใช้ทักษะที่หลากหลาย ด้วยเหตุนี้ การทดสอบแบบอัตนัย ไม่ว่าจะเป็นการสอบข้อเขียนหรือการทดสอบทักษะด้านอื่น ๆ ที่เกี่ยวข้องกับการเรียนรู้ของผู้เรียนจึงมีความจำเป็น รวมไปถึงการพัฒนากระบวนการในการตรวจให้คะแนนการทดสอบแบบอัตนัยเพื่อให้ได้ผลการทดสอบที่มีคุณภาพและมีความเที่ยงสูงจึงมีความสำคัญและเป็นประเด็นที่นักวิจัยด้านการวัดและประเมินผลสนใจศึกษามาเป็นเวลานาน

CHULALONGKORN UNIVERSITY

ผู้วิจัยได้ศึกษาเอกสารงานวิจัยที่เกี่ยวข้องกับการศึกษาและตรวจสอบความเที่ยงระหว่างผู้ประเมิน หรือ Inter-rater reliability โดยสรุปได้เป็น 3 หัวข้อใหญ่ ได้แก่ 1) ความหมายและความสำคัญของความเที่ยงระหว่างผู้ประเมิน 2) ประเภทของความเที่ยงระหว่างผู้ประเมิน และ 3) วิธีการตรวจสอบความเที่ยงระหว่างผู้ประเมิน

1.1 ความหมายและความสำคัญของความเที่ยงระหว่างผู้ประเมิน

ความเที่ยงระหว่างผู้ประเมิน (Inter-rater reliability) หมายถึง ระดับความสอดคล้องของผู้ประเมิน ผู้ตรวจ หรือผู้ให้คะแนนที่มีตั้งแต่ 2 คนขึ้นไป ซึ่งแสดงถึงความคงที่ของการกำหนดคะแนนในกระบวนการให้คะแนนข้อมูลชุดเดียวกันบนเกณฑ์การประเมินเดียวกันและในช่วงเวลาเดียวกัน

(Lange, 2011; Stemler, 2004) Gwet (2014) ชี้ให้เห็นว่าสิ่งสำคัญของการประเมิน คือ ระดับความสามารถของผู้ได้รับการประเมินควรเป็นผลจากคุณลักษณะของผู้รับการประเมินการและการทดสอบที่สะท้อนคุณลักษณะดังกล่าว โดยไม่ขึ้นอยู่กับคุณลักษณะของผู้ประเมิน หากความเที่ยงระหว่างผู้ประเมินอยู่ในระดับสูงแสดงว่าผู้ประเมินกลุ่มนั้นสามารถทำการประเมิน (หรือให้ข้อมูล) แทนกันได้โดยไม่ส่งผลกระทบต่อผลการประเมินหรือผลการเก็บข้อมูลของผู้วิจัย ดังนั้น ความเที่ยงระหว่างผู้ประเมินจึงเป็นการตรวจสอบประสิทธิภาพการทำซ้ำ (reproducibility) การประเมินของผู้ประเมินที่แตกต่างกัน Saal, Downey และ Lahey (1980) ได้กล่าวว่า ความสอดคล้องหรือฉันทามติระหว่างผู้ประเมิน (interrater reliability หรือ rater agreement) นอกจากจะหมายถึงความสอดคล้องในการให้คะแนนระหว่างผู้ประเมินหลายคน ระดับความสอดคล้องระหว่างผู้ประเมินแต่ละคนที่ให้คะแนนพฤติกรรมหรือมิติในการทดสอบเดียวกันอย่างเป็นอิสระต่อกันถือเป็นหลักฐานของความตรงเชิงโครงสร้างอย่างหนึ่ง Bernardin, Alvares และ Cranny (1976) กล่าวว่าความเที่ยงระหว่างผู้ประเมินหมายถึง สัมประสิทธิ์สหสัมพันธ์ของคะแนนระหว่างผู้ตรวจในการประเมินผู้สอบในคุณลักษณะเดียวกันหรือมิติเดียวกัน หากคะแนนระหว่างผู้ประเมินมีความสัมพันธ์กันสูงแสดงถึงความสอดคล้องกันระหว่างผู้ประเมิน

การศึกษาเกี่ยวกับความเที่ยงระหว่างผู้ประเมินมีความสำคัญต่อความน่าเชื่อถือในการประเมินและการสรุปผลการประเมินในฐานะที่เป็นตัวบ่งชี้คุณภาพของกระบวนการประเมินทั้งในการทดสอบขนาดใหญ่และการทดสอบภายในโรงเรียนหรือชั้นเรียน Meadows และ Billington (2005) นอกจากด้านการวัดและประเมินผลทางการศึกษา การวิเคราะห์ความเที่ยงระหว่างผู้ประเมินได้รับการประยุกต์ใช้กับการตรวจสอบคุณภาพของผู้ตรวจในศาสตร์หลายแขนง โดยเฉพาะการศึกษาวิจัยด้านสุขภาพมักนิยมใช้การตรวจสอบความเที่ยงระหว่างผู้ประเมินประกอบการตัดสินผลการรักษา McHugh (2012) ชี้ให้เห็นความสำคัญของความเที่ยงระหว่างผู้ประเมินในการการเก็บข้อมูลในการศึกษาวิจัยทางการแพทย์ในกรณีการวินิจฉัยความรุนแรงของแผลกดทับจากตัวแปรต่าง ๆ เช่น รอยแดง การบวม แผลเปื่อย ซึ่งบางครั้งการตีความผลการวินิจฉัยจากเครื่องมือวัดอาจแตกต่างกัน Gwet (2014) ยกตัวอย่างกรณีที่ผู้ป่วยบางรายได้รับการรักษาผิดพลาดเนื่องมาจากการวินิจฉัยที่คลาดเคลื่อน Gisev และคณะ (2013) ยกตัวอย่างการวิเคราะห์ความเที่ยงของผู้ประเมินในการประเมินทักษะการสื่อสารของเภสัชกร ทั้งนี้ การรายงานผลการศึกษาทางการแพทย์จำเป็นต้องรายงานระดับความสอดคล้องระหว่างผู้ประเมินเสมอ นอกกจากนี้ การศึกษาวิจัยเกี่ยวกับการวิเคราะห์เนื้อหา (content analysis) เป็นอีกสาขาหนึ่งที่นิยมใช้การวิเคราะห์ความเที่ยงระหว่าง

ผู้ประเมินเนื่องจากการศึกษาดังกล่าวเป็นการให้ผู้ประเมินประเมินลักษณะของข้อความแล้วจัดลงในหมวดหมู่ที่กำหนดโดยใช้เกณฑ์การลงรหัสที่กำหนดขึ้น การลงรหัสดังกล่าวจะมีการตรวจสอบความน่าเชื่อถือ เรียกว่า การตรวจสอบความเที่ยงระหว่างผู้ลงรหัส (intercoder reliability) โดยการให้ผู้ลงรหัสแต่ละคนทดลองลงรหัสในข้อมูลเดียวกันแล้วนำมาเปรียบเทียบความสอดคล้องในการลงรหัส (Krippendorff, 2012; Zhao, Liu, และ Deng, 2013)

จากตัวอย่างที่ยกมาสามารถสรุปได้ว่า การตรวจสอบความเที่ยงระหว่างผู้ประเมินมีบทบาทสำคัญในการตรวจสอบคุณภาพของผู้ประเมิน เนื่องจาก การประเมินโดยบุคคลนั้นมีความคลาดเคลื่อนอันเนื่องมาจากตัวผู้ประเมินเองรวมถึงปัจจัยภาพนอกที่ส่งผลต่อความแม่นยำของการประเมินด้วย Downey และ Lahey (1980) ได้สรุปเกี่ยวกับการศึกษาคุณภาพของข้อมูลที่ได้จากการประเมิน (rating data) รวมไปถึงความคลาดเคลื่อนของการประเมินที่ส่งผลต่อความไม่คงที่ของผลการประเมิน และได้ศึกษาเอกสารในช่วงปี 1975 ถึง 1977 การศึกษาดังกล่าวแสดงถึงการให้ความสำคัญต่อคุณภาพของผู้ประเมินในการประเมินที่มีอยู่เป็นเวลานานแล้ว ในบทความดังกล่าว Saal และคณะ ได้สรุปความคลาดเคลื่อนที่เกิดจากผู้ตรวจซึ่งสอดคล้องกับการศึกษาในเวลาต่อมา (Myford และ Wolfe, 2003; Meadows และ Billington, 2005; Tisi, Whitehouse, Maughan, Burdett, 2013) โดยจำแนกได้ดังต่อไปนี้

1) อิทธิพลฮาโล (Halo effect) (Thorndike, 1920) เป็นปรากฏการณ์ที่เกิดจากผู้ประเมินมีทัศนคติทางบวกต่อผู้รับการประเมินอยู่ก่อนแล้ว เช่น ผู้รับการประเมินเป็นนักเรียนที่มีความสามารถหรือตั้งใจเรียนชั้นเรียน ความประทับใจดังกล่าวส่งผลต่อการให้คะแนนในการทดสอบที่ผู้ประเมินมีแนวโน้มจะให้คะแนนนักเรียนคนดังกล่าวสูงกว่าความเป็นจริง (Nisbett และ Wilson, 1977; Cooper, 1981; Malouff, Emmerton และ Schutte, 2013)

2) การกด/ปล่อยคะแนน (Leniency and Severity) คือ แนวโน้มที่ผู้ประเมินให้คะแนนต่ำหรือสูงกว่าความเป็นจริง อีกนัยหนึ่งคือ การกดหรือปล่อยคะแนนของผู้ประเมิน โดยทั่วไปแล้วความคลาดเคลื่อนในลักษณะนี้จะคงที่ในผู้ประเมินแต่ละคนโดยไม่ได้ขึ้นอยู่กับทัศนคติที่มีต่อผู้รับการประเมินแต่เป็นพฤติกรรมเฉพาะของผู้ประเมินที่มีมโนทัศน์เกี่ยวกับมาตรฐานของการให้คะแนนผิดไปจากความเป็นจริง (Saal และ Landy, 1977; Decotiis, 1977; Bernardin, LaShells, Smith และ Alvares, 1976)

3) การให้คะแนนในช่วงกลางหรือช่วงจำกัด (Central tendency/range restriction) คือ สถานการณ์ที่ผู้ประเมินให้คะแนนในช่วงจำกัด เช่น การให้คะแนนในระดับปานกลางในผู้รับการประเมินทุกคน หรือให้คะแนนที่อยู่ในช่วงใดช่วงหรือแบบคงที่ การให้คะแนนในระดับปานกลางแสดงถึงความไม่มั่นใจหรือลังเลในการตัดสินคะแนนซึ่งส่งผลให้อำนาจจำแนกในการตรวจต่ำ (Saal, Downey และ Lahey, 1980)

4) ความคลาดเคลื่อนในการให้คะแนนที่เกิดจากการที่ผู้ประเมินให้คะแนนคุณลักษณะที่คล้ายคลึงกันด้วยคะแนนใกล้เคียงกัน (Logical error) (Newcomb, 1931)

5) การให้คะแนนโดยเปรียบเทียบความสามารถของผู้รับการประเมินกับความสามารถของตนเอง (Contrast error) (Murray, 1938)

6) อิทธิพลของข้อคำถามใกล้เคียงที่ส่งผลต่อการให้คะแนนของผู้ประเมิน (Proximity error) (Stockford และ Bissell, 1949)

เมื่อพิจารณาปัจจัยที่ส่งผลต่อความแม่นยำในการประเมินของผู้ประเมินที่กล่าวมา จะเห็นว่าความคลาดเคลื่อนของการประเมินมีสาเหตุทั้งจากตัวผู้ประเมินเอง และยังมีสาเหตุจากปัจจัยเกี่ยวกับแบบสอบด้วย ปัจจัยเหล่านี้ไม่สามารถควบคุมหรือขจัดออกได้ทั้งหมด ดังนั้นการตรวจสอบคุณภาพของการประเมินของผู้ประเมินจึงสำคัญเพราะให้สารสนเทศเกี่ยวกับความคลาดเคลื่อนของการประเมินซึ่งใช้เป็นข้อมูลประกอบการรายงานผลการประเมินให้มีความน่าเชื่อถือและโปร่งใสมากขึ้น

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

1.2 ประเภทของความเที่ยงระหว่างผู้ประเมิน

Stemler (2004) อธิบายเกี่ยวกับประเภทของความเที่ยงระหว่างผู้ประเมินว่า อันที่จริงแล้วความเที่ยงระหว่างผู้ประเมินไม่ได้มีความหมายเพียงแค่ความคงที่ของการให้คะแนนระหว่างผู้ประเมิน แต่แบ่งย่อยออกเป็น 3 ลักษณะ ตามวิธีการทางสถิติในการประมาณค่าความเที่ยงระหว่างผู้ประเมิน ได้แก่ การประมาณค่าความฉันทามติ (Consensus estimate) การประมาณค่าความคงที่ (Consistency estimate) และการประมาณค่าการวัด (Measurement estimate) การประมาณค่าทั้ง 3 ลักษณะนำไปสู่สารสนเทศที่แตกต่างกันและมีวิธีการประมาณค่าที่แตกต่างกันดังรายละเอียดต่อไปนี้

1) การประมาณค่าความสอดคล้อง (Agreement/Consensus estimate) เป็นการประมาณค่าความเที่ยงระหว่างผู้ประเมินที่อยู่บนพื้นฐานข้อตกลงว่าผู้ประเมินแต่ละคนต้องให้คะแนนตรงกันทุกคุณลักษณะที่ประเมิน ซึ่งหากผู้ประเมินสามารถให้คะแนนตรงกันได้ทั้งหมดจะเป็นเครื่องสะท้อนว่าผู้ประเมินเข้าใจเกณฑ์การประเมินตรงกันหรือมีการตีความโครงสร้างทางปัญญาร่วมกัน การประมาณค่าฉันทามติเป็นการประมาณค่าที่เหมาะสมสำหรับข้อมูลแบบจัดกลุ่มหรือข้อมูลการประเมินระดับคุณลักษณะที่แต่ละระดับบนมาตรประมาณค่าแสดงถึงคุณลักษณะที่แตกต่างกันเชิงคุณภาพ ยกตัวอย่างเช่น อาจารย์นิเทศประเมินการฝึกสอนของนิสิตฝึกสอนว่าจัดกิจกรรมการเรียนการสอน เหมาะสม/ไม่เหมาะสม กับผู้เรียน หรือตัวอย่างการศึกษาเกี่ยวกับวัตถุประสงค์ในการจัดการศึกษาต่อเนื่องของโรงเรียนของ Stemler และ Bebell (1999) ซึ่งให้ผู้ประเมินจำแนกพันธกิจของโรงเรียนให้เข้ากับวัตถุประสงค์ของการวัดการศึกษาต่อเนื่อง กรณีดังกล่าว หากผู้ประเมินจำแนกพันธกิจเข้ากับวัตถุประสงค์ได้ตรงกันทุกคนจะถือว่าผู้ประเมินเข้าใจและสื่อสารตรงกันเกี่ยวกับมโนทัศน์ของพันธกิจและวัตถุประสงค์ของการจัดการศึกษาดังกล่าว เป็นต้น การประมาณค่าฉันทามติสามารถใช้ประมาณค่าข้อมูลแบบเรียงลำดับที่ต่อเนื่องกันได้ เช่น การประเมินบนมาตรประมาณค่าแบบลิเคิร์ท (Likert scale) เป็นต้น ซึ่งการประมาณค่าลักษณะนี้เป็นการประมาณค่าสำหรับการประเมินที่จำแนกบุคคลเป็นระดับตามคุณภาพของคุณลักษณะหรือโครงสร้างทางจิตวิทยาในเชิงปริมาณ การประมาณค่าความสอดคล้องมุ่งเน้นการตรวจสอบความสอบคล้องระหว่างผู้ประเมินโดยอยู่บนข้อตกลงที่ว่าผู้ประเมินต้องให้คะแนนตรงกันในแต่ละคุณลักษณะที่ประเมิน สถิติที่เกี่ยวข้องกับการประมาณค่าฉันทามติที่นิยมใช้มากที่สุด คือ สัดส่วนความสอดคล้อง (percent-agreement) ซึ่งเป็นสถิติที่คำนวณและแปลผลง่าย อย่างไรก็ตาม การคิดสัดส่วนความสอดคล้องอาจมีแนวโน้มในการเกิดค่าฉันทามติที่เกินจริง (Hayes & Hatch, 1999; Stemler, 2004) และการที่จะทำให้ผู้ประเมินให้คะแนนตรงกันจำเป็นต้องทำความเข้าใจกับผู้ประเมินหรืออาจต้องอบรมเกี่ยวกับการใช้เกณฑ์ในการประเมินซึ่งเป็นกระบวนการที่ใช้เวลาและสิ้นเปลืองทรัพยากร เนื่องจากข้อจำกัดดังกล่าวจึงได้มีการปรับปรุงการคำนวณสัดส่วนความสอดคล้องโดยใช้การรวมคะแนนที่ใกล้เคียงกันบนมาตรประมาณค่า กล่าวคือ หากผู้ประเมินให้คะแนนต่างกันไม่เกิน 1 ระดับของมาตรประมาณค่าจะถือว่าผู้ประเมินดังกล่าวให้คะแนนสอดคล้องกัน แต่การคำนวณคะแนนดังกล่าวก็ยังไม่สามารถลดการเกิดค่าฉันทามติที่เกินจริงได้ โดยเฉพาะหากมาตรประมาณค่าที่ใช้มีช่วงจำกัด เช่น ให้คะแนนเพียง 3 ระดับ (ดีมาก พอใช้ ควรปรับปรุง) วิธีการทางสถิติที่นำมาใช้เพื่อลดข้อจำกัดเกี่ยวกับการประมาณค่าที่เกินจริงของการประมาณค่าฉันทามติของผู้ประเมิน คือ สัมประสิทธิ์แคปปา

(Kappa coefficient) (Stemler, 2004; Kottner และคณะ, 2011; Gisev และคณะ, 2013) สถิติดังกล่าวเป็นการประมาณค่าระดับฉันทามติจากสัดส่วนของคะแนนสังเกตได้จากการให้คะแนนระหว่างผู้ประเมิน 2 คน กับสัดส่วนความเป็นไปได้ที่ผู้ประเมินจะให้คะแนนตรงกัน ซึ่งมีการปรับสัดส่วนของระดับความสอดคล้องที่เกิดขึ้นโดยบังเอิญ Cohen's kappa มีค่าตั้งแต่ -1 ถึง +1 (Stemler, 2004; Gisev และคณะ, 2013) สัมประสิทธิ์แคปปาเท่ากับ 0 หมายความว่า ผู้ประเมิน 2 คนมีความเห็นไม่ตรงกันเท่ากับสัดส่วนความสอดคล้องที่เกิดขึ้นโดยบังเอิญ สัมประสิทธิ์แคปปามีค่า +1 แสดงว่าผู้ประเมินมีความเห็นสอคล้องกันทั้งหมด สัมประสิทธิ์แคปปาเป็นสถิติที่ได้รับความนิยมในการประมาณค่าความเที่ยงระหว่างผู้ประเมินที่ใช้กันโดยทั่วไปในการศึกษาวิจัยต่าง ๆ อย่างไรก็ตาม ข้อจำกัดที่สำคัญประการหนึ่งของสถิติดังกล่าว คือ ความยุ่งยากและซับซ้อนในการแปลความหมาย (Andres & Marzo, 2004; Stemler, 2004; Kottner และคณะ, 2011; Gisev และคณะ, 2013) ถึงแม้จะมีผู้เสนอเกณฑ์ในการแปลผลค่าสัมประสิทธิ์แคปปาไว้แล้ว ได้แก่ Landis & Koch (1977) Krippendorff (1980) Fleiss Levin & Paik (2003) และ McHugh (2012) แต่นักวิจัยจำเป็นต้องพิจารณาเลือกเกณฑ์ที่เหมาะสมแก่การศึกษาของตน (ประสพชัย พสุนนท์, 2558) Conger (2016) ตั้งข้อสังเกตว่าเกณฑ์ดังกล่าวไม่ได้จำแนกระหว่างดัชนีกับความสอดคล้อง ไม่คำนึงถึงการแจกแจงความถี่ระหว่างระดับ และไม่ได้จำแนกระหว่างความแม่นยำกับความสอดคล้องของผู้ประเมิน โดยแสดงให้เห็นว่าค่าของสัมประสิทธิ์แคปปาจะลดลงรวมถึงมีคุณภาพต่ำลงเมื่อความแปรปรวนในแต่ละระดับเพิ่มขึ้น LeBreton และ Senter (2008) ได้กล่าวว่าขอบเขตบนของสัมประสิทธิ์แคปปามีค่าสูงเกินความจำเป็นและเป็นไปได้ยากที่จะได้ค่าสัมประสิทธิ์แคปปาที่สูงในการศึกษาบางกรณี ดังนั้น ค่าสัมประสิทธิ์แคปปาต่ำไม่ได้แสดงถึงการมีความสอดคล้องต่ำเสมอไป นอกจากนี้ สัมประสิทธิ์แคปปายังเป็นสถิติที่อ่อนไหวต่อความลำเอียงระหว่างผู้ประเมินรวมถึงความถี่ของการให้คะแนน ทั้งนี้การแจกแจงที่ไม่สมมาตรหรือลำเอียงจะส่งผลให้ค่าสัมประสิทธิ์แคปปามีค่าสูงเกินจริง (Byrt, Bishop และ Carlin, 1993)

2) การประมาณค่าความคงที่ (Consistency estimate) มีจุดประสงค์เพื่อตรวจสอบความคงเส้นคงวาของการให้คะแนนซึ่งอยู่บนข้อตกลงที่ว่าผู้ประเมินไม่จำเป็นต้องให้คะแนนตรงกันทุกคุณลักษณะ แต่การให้คะแนนนั้นต้องมีความคงที่ในการจำแนกระดับคุณลักษณะที่มุ่งวัดได้ สถิติที่ใช้ในการตรวจสอบความคงที่ที่เป็นที่รู้จักและใช้กับอย่างกว้างขวางคือ สัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน (Pearson correlation coefficient) ในกรณีที่มีผู้ประเมิน 2 คน ข้อจำกัดของสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน คือ ข้อมูลที่นำมาวิเคราะห์ต้องมีการแจกแจงปกติและสามารถเปรียบเทียบ

ความสัมพันธ์ได้ครั้งละ 1 คู่ เท่านั้น เนื่องจากข้อจำกัดของการแจกแจงข้อมูล สัมประสิทธิ์สหสัมพันธ์แบบสเปียร์แมนจึงได้ถูกนำมาใช้ในกรณีที่ข้อมูลแจกแจงไม่ปกติ แต่จำเป็นต้องปรับข้อมูลให้เป็นแบบเรียงลำดับ (rank) ก่อนทำการวิเคราะห์ ในกรณีที่มีผู้ประเมินตั้งแต่ 3 คนขึ้นไป จะใช้สัมประสิทธิ์อัลฟา (Cronbach's Alpha coefficient) ในการประมาณค่าความคงที่ โดยเป็นการประมาณค่าความคงที่แบบสอดคล้องภายในซึ่งเป็นประโยชน์ในการตรวจสอบระดับความคงที่ของการให้คะแนนของผู้ประเมินแต่ละคนโดยเปรียบเทียบระหว่างผู้ประเมินอย่างไรก็ตาม เช่นเดียวกับการใช้สัมประสิทธิ์สหสัมพันธ์ การวิเคราะห์ความเที่ยงโดยสัมประสิทธิ์อัลฟาจำเป็นต้องให้ผู้ประเมินประเมินทุกคุณลักษณะที่มุ่งวัดของผู้รับการประเมินทุกคน วิธีการทางสถิติอีกวิธีหนึ่งที่ได้รับความนิยมในการตรวจสอบความเที่ยงระหว่างผู้ประเมินคือ Intraclass Correlation Coefficient (ICC) ซึ่งใช้วิธีการวิเคราะห์ความแปรปรวน (ANOVA) จุดเด่นของ ICC คือ สามารถใช้วิเคราะห์ความเที่ยงระหว่างผู้ประเมินหลายคน โดยสามารถวิเคราะห์ข้อมูลที่มีค่าสูญหายได้ รวมถึงสามารถวิเคราะห์ตัวแปรได้ทุกระดับ (Gisev และคณะ, 2013) อย่างไรก็ตาม เนื่องจากการวิเคราะห์ ICC อยู่บนพื้นฐานของการวิเคราะห์ความแปรปรวนและมีข้อตกลงเกี่ยวกับการแจกแจงปกติของข้อมูล ค่า ICC อาจสูงกินจริงได้หากข้อมูลในการวิเคราะห์มาจากกลุ่มตัวอย่างที่มีลักษณะเป็นวิวิธพันธ์ (heterogeneous) สูง ส่งผลให้ไม่สามารถสรุปได้ว่าผู้ประเมินมีความเที่ยงอย่างแท้จริง (Muller และคณะ, 1994; Costa-Santos และคณะ, 2009)

3) การประมาณค่าการวัด (Measurement estimate) ตั้งอยู่บนข้อตกลงว่าการสรุปผลคะแนนของผู้ประเมินต้องใช้สารสนเทศทั้งหมดที่ได้จากการประเมินรวมถึงความแตกต่างของการให้คะแนนระหว่างผู้ประเมินในการวิเคราะห์ ซึ่ง Linacre (2002) ชี้ว่าปัจจัยที่ทำให้เกิดความคลาดเคลื่อนในการวิเคราะห์และแปลความหมายนั้นไม่ได้เกิดจากการให้คะแนน แต่เกิดจากกระบวนการเก็บรวบรวมข้อมูล ดังนั้น การประมาณค่าการวัดจึงไม่จำเป็นที่ผู้ประเมินต้องมีผลประเมินตรงกัน เนื่องจากวิธีการประมาณค่าดังกล่าวสามารถประมาณค่าการกดคะแนนของผู้ประเมินได้ในการวิเคราะห์ขั้นสุดท้าย (Stemler, 2004) สถิติที่นิยมใช้สำหรับการประมาณค่าการวัด ได้แก่ การวิเคราะห์องค์ประกอบหลัก (Principal component analysis) ซึ่งเป็นการตรวจสอบปริมาณความแปรปรวนร่วมกันของคะแนนในองค์ประกอบ หากความแปรปรวนร่วมกันสูงกว่าร้อยละ 60 แสดงว่าผู้ประเมินแต่ละคนประเมินคุณลักษณะเดียวกัน ซึ่งถือว่าผลการประเมินมีความเที่ยง นอกจากนั้น การวิเคราะห์องค์ประกอบหลักยังให้สารสนเทศที่เป็นผลคะแนนโดยดูจากค่าน้ำหนัก

องค์ประกอบ ข้อจำกัดของวิธีการนี้ คือ เป็นวิธีการที่ตั้งอยู่บนข้อตกลงเบื้องต้นว่าการให้คะแนนของผู้ประเมินต้องไม่มีความคลาดเคลื่อน

ในปัจจุบัน มีวิธีการประมาณค่าการวัดที่ใช้กันแพร่หลายในการวิเคราะห์ความเที่ยงระหว่างผู้ประเมิน คือ Many-facets Rasch model (MFRM) ซึ่งเป็นวิธีการที่ใช้ในการวิเคราะห์การทำหน้าที่ต่างกันของผู้ประเมินในการทดสอบขนาดใหญ่ เช่น การทดสอบของ Educational Testing Service (ETS) (Engelhard Jr., Wind, Kobrin และ Chajewski, 2013) จุดเด่นของโมเดลดังกล่าวคือการรวมความเข้มงวด (severity) ในการให้คะแนนของผู้ตรวจไว้บนสเกลเดียวกับความสามารถของผู้ให้ข้อมูลและความยากของข้อคำถาม ทำให้สามารถเปรียบเทียบคะแนนระหว่างผู้ประเมินแต่ละคน รวมถึงทราบได้ว่าผู้ประเมินคนใดมีความเข้มงวดในการให้คะแนนมากกว่ากัน (Stemler, 2004) นอกจากนี้ MFRM ยังให้สารสนเทศที่เป็นประโยชน์ต่อการประเมินความเข้มงวดในการให้คะแนนหรือความคงที่ในการจำแนกระดับคุณลักษณะจากเกณฑ์การให้คะแนนว่าผู้ประเมินสามารถจำแนกระดับคุณลักษณะจากเกณฑ์การให้คะแนนได้คงที่เพียงใด อย่างไรก็ตาม ข้อจำกัดที่สำคัญประการหนึ่งของวิธีการประมาณค่าการวัดทั้งการวิเคราะห์องค์ประกอบและ Many-facets Rasch model คือ ไม่รองรับข้อมูลแบบจัดกลุ่ม และใช้กลุ่มตัวอย่างขนาดใหญ่ในการวิเคราะห์ (Linacre, 1994; Stemler, 2004; Farrokhi, Esfandiari และ Schaefer, 2012)

1.3 วิธีการตรวจสอบความเที่ยงระหว่างผู้ประเมิน

การตรวจสอบความเที่ยงระหว่างผู้ประเมินใช้สถิติเป็นเครื่องมือช่วยในการวิเคราะห์ข้อมูลเกี่ยวกับความสอดคล้องในการให้คะแนนของผู้ประเมินแต่ละคนในการประเมินคุณลักษณะหรือมิติของการวัดเดียวกัน เนื่องจากการตรวจสอบความเที่ยงระหว่างผู้ประเมินเป็นการตรวจสอบคุณภาพของการประเมินที่สำคัญและมีนักวิจัยให้ความสนใจศึกษาเป็นจำนวนมากตั้งแต่อดีตจนปัจจุบัน รวมถึงมีการสังเคราะห์เอกสารงานวิจัยที่เกี่ยวกับความเที่ยงของผู้ประเมินที่สำคัญ โดยสรุปได้ดังนี้

<u>ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐาน</u>

ค่าเฉลี่ย (mean) และส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation: SD) เป็นสถิติพื้นฐานที่ใช้ตรวจสอบการกระจายของการให้คะแนนระหว่างผู้ตรวจ ซึ่ง Bramley (2007) ให้ข้อเสนอแนะว่าสถิติที่เข้าใจง่ายและตีความได้ง่ายที่สุดควรไม่ซับซ้อน โดยตั้งอยู่บนพื้นฐานของการแจกแจงความแตกต่างระหว่างคะแนนจากการวัดและคะแนนจริง หรือคะแนนระหว่างผู้ตรวจกับผู้เชี่ยวชาญ ทั้งนี้ ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของความแตกต่างดังกล่าวให้สารสนเทศที่

ตรงไปตรงมาตามแนวคิดของการทดสอบแบบดั้งเดิม การศึกษาเกี่ยวกับความเที่ยงระหว่างผู้ประเมินในอดีตได้มีการอธิบายเกี่ยวกับการเปรียบเทียบค่าเฉลี่ยระหว่างมิติการประเมินกับค่ากลางของสเกล การประเมินเป็นวิธีการหนึ่งที่สามารถใช้ตรวจสอบการกดหรือปล่อยคะแนนของผู้ประเมินได้ หากค่าเฉลี่ยของคะแนนการประเมินเกินกว่าค่ากลางถือว่ามีการกดหรือปล่อยคะแนน (Bernardin, Alvares และ Cranny, 1976) และในกรณีของการตรวจสอบการให้คะแนนในช่วงจำกัด (Restriction of range) สามารถพิจารณาจากส่วนเบี่ยงเบนมาตรฐานของคะแนนการประเมิน หากส่วนเบี่ยงเบนมาตรฐานมีค่าต่ำ แสดงถึงการมีแนวโน้มของการให้คะแนนในช่วงจำกัด การศึกษาเกี่ยวกับการกดหรือปล่อยคะแนน และการให้คะแนนในช่วงจำกัดในช่วงแรก พบว่ามีการใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานเป็นเกณฑ์ในการพิจารณาความเที่ยงของผู้ตรวจ เช่น James (1974) ศึกษาเกี่ยวกับการให้คะแนนข้อสอบวิชาฟิสิกส์โดยพิจารณาส่วนเบี่ยงเบนมาตรฐานและสัมประสิทธิ์สหสัมพันธ์ระหว่างผู้ตรวจ Baird (1988) ทำการทดลองการให้คะแนนข้อสอบ A Level ในวิชาเคมีและวิชาวรรณคดีอังกฤษโดยการเปรียบเทียบส่วนเบี่ยงเบนมาตรฐานของการให้คะแนนผู้สอบเพศชายและเพศหญิง การศึกษาไม่พบความแตกต่างระหว่างเพศที่ส่งผลต่อการให้คะแนน และ Dennis (1990) ศึกษาการตรวจแบบปกปิดข้อมูลและเปิดเผยข้อมูลผู้สอบ พบว่าการปกปิดหรือเปิดเผยข้อมูลไม่มีผลต่อการให้คะแนน แต่เมื่อพิจารณาส่วนเบี่ยงเบนมาตรฐานพบว่า ส่วนเบี่ยงเบนมาตรฐานของคะแนนผู้สอบที่เป็นเพศหญิงสูงกว่าผู้สอบที่เป็นเพศชาย

<u>สัมประสิทธิ์สหสัมพันธ์ (Correlation coefficients)</u>

สถิติที่นำมาใช้เป็นเกณฑ์การพิจารณาความสอดคล้องระหว่างผู้ประเมินอย่างแพร่หลาย คือ สัมประสิทธิ์สหสัมพันธ์ (Pearson Correlation Coefficient) เป็นการพิจารณาความสอดคล้องของการให้คะแนนที่อยู่ในช่วง 0 ถึง 1 หากค่าสัมประสิทธิ์สหสัมพันธ์สูง แสดงว่ามีความสอดคล้องในการให้คะแนนในระดับสูง ค่าสัมประสิทธิ์สหสัมพันธ์ใช้ประกอบการพิจารณาคุณภาพของการตรวจให้คะแนนในหลายประเด็น ได้แก่ ความสอดคล้องของการให้คะแนนระหว่างผู้ประเมิน ความสอดคล้องของการให้คะแนนของผู้ประเมินข้ามช่วงเวลา รวมถึงความคงที่ในการให้คะแนนของผู้ประเมิน (Bernardin, Alvares และ Cranny, 1976; Stemler, 2004) เนื่องจากเป็นสถิติที่คำนวณได้ง่ายทั้งการใช้เครื่องคำนวณหรือคำนวณด้วยมือ มีสูตรการคำนวณดังสมการ

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)}\sqrt{(\sum y_i^2 - n\bar{y}^2)}} \quad (2.1)$$

อย่างไรก็ตาม Meadows และ Billington (2005) ชี้ว่าการใช้ค่าสัมประสิทธิ์สหสัมพันธ์ในการวัดความเที่ยงนั้นมีข้อจำกัดเนื่องจากสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สันเป็นสถิติที่ไม่สะท้อนความสอดคล้องระหว่างข้อมูลอย่างแท้จริง แต่เป็นการสะท้อนระดับความสัมพันธ์ของข้อมูล หากข้อมูลมีความสัมพันธ์กันสูงแม้ในทิศทางตรงข้ามกันค่าสัมประสิทธิ์สหสัมพันธ์จะมีค่าสูงทั้งที่จริงแล้วมีระดับความสอดคล้องระหว่างข้อมูลต่ำ (Gisev และคณะ, 2013; Giavarina, 2015) ทั้งนี้สัมประสิทธิ์สหสัมพันธ์ไม่สามารถสะท้อนถึงความแตกต่างของคุณลักษณะภายใต้การแจกแจงของตัวแปรที่มีความสัมพันธ์กันได้ และการใช้ค่าสัมประสิทธิ์สหสัมพันธ์นั้นทำให้เกิดการประมาณค่าเกินจริงเพราะไม่ได้ให้ความสำคัญกับค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของคะแนน ดังเช่น Lunz และ Wright (1994 อ้างใน Meadows และ Billington, 2005) แสดงให้เห็นในการศึกษาว่าค่าสัมประสิทธิ์สหสัมพันธ์ที่สูงอาจเป็นผลมาจากการให้คะแนนที่ต่างกันอย่างเป็นระบบของผู้ตรวจ สอดคล้องกับความเห็นของ Stemler (2004) ที่ชี้ว่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สันมีข้อตกลงเบื้องต้นเกี่ยวกับการแจกแจงปกติของข้อมูล ดังนั้น หากข้อมูลมีการแจกแจงแบบเบ้ทางใดทางหนึ่งอาจส่งผลต่อการอ่อนตัวของขอบเขตบนของค่าสถิติดังกล่าวได้

<u>Percentage agreement</u>

การหาสัดส่วนความสอดคล้องเป็นวิธีการคำนวณความสอดงคล้องจากคะแนนดิบของผู้ตรวจในการให้คะแนนแบบจัดกลุ่มหรือเรียงอันดับ คำนวณจาก

$$PA = \frac{\text{จำนวนการให้คะแนนที่ตรงกัน}}{\text{จำนวนการให้คะแนนทั้งหมด}} \text{x } 100 \quad (2.2)$$

ตัวอย่างเช่น ผู้ตรวจ 2 คน ให้คะแนนการเขียนความเรียง 5 ฉบับ โดยมีระดับคะแนน 1 – 5 ระดับ ดังนี้

ความเรียง	ผู้ตรวจ 1	ผู้ตรวจ 2	จำนวนความสอดคล้อง
1	5	5	1
2	4	5	0
3	4	4	1
4	3	2	0
5	3	3	1
	รวม		3

จากผลการให้คะแนนดังกล่าว ผู้ตรวจให้คะแนนสอดคล้องกันทั้งหมดจำนวน 3 ฉบับ จากการตรวจทั้งหมด 5 ฉบับ คิดเป็น (3/5) x 100 = 60 ดังนั้นจึงสรุปว่า การตรวจความเรียงครั้งนี้ผู้ตรวจให้คะแนนสอดคล้องกันคิดเป็นร้อยละ 60 หากมีจำนวนผู้ตรวจมากกว่า 2 คน จะคำนวณความสัดส่วนสอดคล้องเป็นรายคู่ก่อนนำมาหาค่าเฉลี่ยของสัดส่วนความสอดคล้องทั้งหมดอีกครั้ง ดังนี้

ความเรียง	ผู้ตรวจ 1	ผู้ตรวจ 2	ผู้ตรวจ 3	ผู้ตรวจ 1:2	ผู้ตรวจ 1:3	ผู้ตรวจ 2:3	จำนวนความสอดคล้อง
1	5	5	5	1	1	1	3/3
2	4	5	3	0	0	0	0/3
3	4	4	4	1	1	1	3/3
4	3	2	2	0	0	1	1/3
5	3	3	2	1	0	0	1/3

จากนั้นนำจำนวนความสอดคล้องมาหาค่าเฉลี่ย จะได้ ((3/3)+(0/3)+(3/3)+(1/3)+(1/3)) / 5 = 0.53 หรือคิดเป็นร้อยละ 53

จากตัวอย่างดังกล่าวจะเห็นว่า การหาสัดส่วนความสอดคล้องสามารถคำนวณและแปลผลได้ง่ายสำหรับผู้ตรวจจำนวน 2-3 คน หากมีจำนวนผู้ตรวจมากขึ้น การคำนวณจะใช้เวลานานเนื่องจากต้องคำนวณความสอดคล้องรายคู่ นอกจากนี้ ผลการคำนวณที่ได้แสดงถึงจำนวนความสอดคล้องในการให้คะแนนเพียงอย่างเดียว แต่ไม่สามารถบอกถึงสารสนเทศอื่นที่อยู่เบื้องหลังการให้คะแนนได้ เช่น ความลำเอียงในการตรวจ การกด/ปล่อยคะแนนของผู้ตรวจ รวมถึงไม่สามารถบอกเกี่ยวกับอิทธิพลของผู้ตรวจที่มีต่อการให้คะแนนได้

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

<u>สัมประสิทธิ์แคปปา (Kappa statistics)</u>

สัมประสิทธิ์แคปปาถูกนำเสนอขึ้นโดย Jacob Cohen (Cohen, 1960) เป็นสถิติที่ใช้สำหรับการพิจารณาความสอดคล้องระหว่างผู้ตรวจในแง่ของความสอดคล้อง (agreement) และความคงเส้นคงวา (consistency) ของการให้คะแนนแบบจัดกลุ่มและได้รับการนำไปใช้อย่างแพร่หลายในการวิเคราะห์ความเที่ยงระหว่างผู้ประเมิน สัมประสิทธิ์แคปปาได้รับการพัฒนาแนวคิดและวิธีการคำนวณจนเกิดเป็นกลุ่มของสถิติที่ใช้เป็นตัวชี้วัดความสอดคล้องและความคงที่ของการประเมินที่มีผู้ประเมินจำนวน 2 คน การคำนวณสัมประสิทธิ์แคปปามีขั้นตอนดังนี้

สมมติให้ผู้ตรวจ 2 คน ประเมินพฤติกรรมของผู้เรียน โดยกำหนดให้ 0 คือ ไม่มีพฤติกรรม และ 1 มีพฤติกรรมที่คาดหวัง ได้ผลดังตาราง

ผู้ประเมิน A	**ผู้ประเมิน B**		
	0	**1**	**รวม**
0	A	B	B0 = A+B
1	C	D	B1 = C+D
รวม	A0 = A+C	A1 = B+D	**N**

จากนั้น คำนวณสัมประสิทธิ์แคปปาดังสมการ

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{2.3}$$

เมื่อ $p_o = (A + D)/N$ และ $p_e = \left(\frac{A0}{N}\right)\left(\frac{B0}{N}\right) + (\frac{A1}{N})(\frac{B1}{N})$

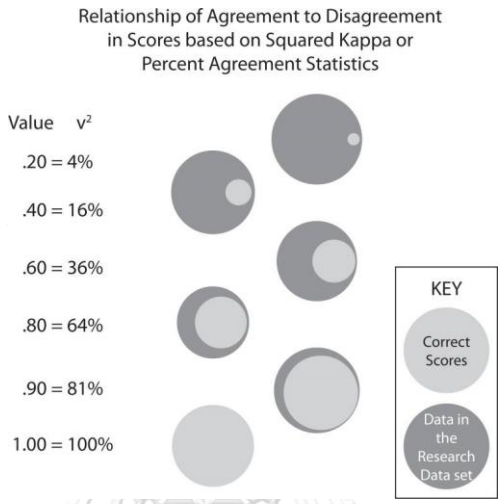
จากสมการ p_o คือ ความสอดคล้องระหว่างผู้ประเมิน และ p_e คือ ความน่าจะเป็นของความสอดคล้องกันโดยบังเอิญ การแปลความหมายของสัมประสิทธิ์แคปปาสามารถแปลได้ตามเกณฑ์ที่ตั้งไว้ซึ่งมีหลายเกณฑ์ ดังตาราง 2.1

ตาราง 2.1 เกณฑ์การพิจารณาความสอดคล้องของสัมประสิทธิ์แคปปา

Landis & Koch (1977)		**Krippendorff (1980)**		**Fleiss, Levin & Piak (2003)**	
Kappa (κ)	ความหมาย	Kappa (κ)	ความหมาย	Kappa (κ)	ความหมาย
0.81 – 1.00	Almost perfect	0.80 ขึ้นไป	ยอมรับได้	0.75 – 1.00	ดีมาก
0.61 – 0.80	Substantial				
0.41 – 0.60	Moderate	0.67 – 0.80	สอดคล้องปานกลาง	0.40 – 0.74	ดี
0.21 – 0.40	Fair				
0.00 – 0.20	Slight	0.00 – 0.67	ไม่สอดคล้อง	0.00 – 0.39	ต่ำ
ต่ำกว่า 0.00	Poor				

McHugh (2012) แสดงความเห็นว่า สถิติแคปปาอยู่ในรูปของค่าสัมประสิทธิ์ซึ่งไม่สามารถแปลความหมายได้โดยตรงแต่ควรตีความในรูปของ Coefficient of Determination (COD) ซึ่งเป็นการอธิบายปริมาณความแปรปรวนภายในตัวแปรตามที่สามารถอธิบายได้ด้วยตัวแปรอิสระ การคำนวณ COD ใช้กับสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน แต่ McHugh ได้นำมาปรับใช้กับ

การแปลความหมายของค่าสัมประสิทธิ์แคปปาด้วย โดยอธิบายว่าข้อมูลของคะแนนที่ได้จากการรวบรวมคะแนนนั้นมีทั้งส่วนที่เป็นคะแนนจริง (correct scores) และส่วนของความคลาดเคลื่อน (errors) ทั้งนี้ McHugh ได้นำค่าสัมประสิทธิ์แคปปามายกกำลังสองเพื่อแสดงถึงร้อยละของความถูกต้องแม่นยำของคะแนนที่ได้จากการเก็บรวบรวม ดังรูปที่ 2.1 และได้เสนอแนวทางการแปลความหมายของค่าสถิติแคปปา ดังตาราง 2.2



รูป 2.1 ปริมาณความถูกต้องของข้อมูลเทียบกับสัดส่วนความสอดคล้อง (McHugh, 2012)

ตาราง 2.2 เกณฑ์การพิจารณาความสอดคล้องของสัมประสิทธิ์แคปปาของ McHugh (2012)

Kappa (κ)	**ระดับความสอดคล้องระหว่างผู้ประเมิน**	**ร้อยละของความถูกต้องของข้อมูล**
0 – 0.20	ไม่สอดคล้อง	0 – 4
0.21 – 0.39	ความสอดคล้องต่ำมาก	4 – 15
0.40 – 0.59	ความสอดคล้องต่ำ	15 – 35
0.60 – 0.79	ความสอดคล้องปานกลาง	36 - 63
0.80 – 0.90	สอดคล้องสูง	64 – 81
มากกว่า 0.90	สอดคล้องสูงที่สุด	82 – 100

แนวคิดของการคำนวณสถิติแคปปาของโคเฮน คือ การคำนวณความไม่สอดคล้องระหว่างผู้ประเมินแต่ไม่ได้บอกถึงระดับของความไม่สอดคล้อง ดังนั้น จึงมีการปรับปรุงวิธีการคำนวณแบบถ่วงน้ำหนัก หรือ weighted kappa (κ_w) สำหรับตัวแปรแบบจัดกลุ่มหรือเรียงอันดับที่ได้รับการประเมินจากผู้ประเมิน สถิติแคปปาแบบถ่วงน้ำหนักจะคำนวณน้ำหนักการให้คะแนนที่แตกต่างกันโดยใช้การถ่วงน้ำหนักแบบเชิงเส้น หากมีความแตกต่างกันสูงจะให้น้ำหนักมากขึ้น ดังตัวอย่าง

สมมติให้นักเรียนประเมินงานเขียนของเพื่อน ให้คะแนน 5 ระดับ ผลการประเมินเป็นดังนี้

ระดับคะแนน	1	2	3	4	5	รวม
1	1	0	1	0	0	2
2	0	2	0	0	0	2
3	0	0	1	0	1	2
4	0	0	0	1	0	1
5	0	1	0	0	2	3
รวม	1	0	2	1	3	10

จากนั้น คำนวณค่าคาดหมาย (Expected value) ของคะแนนการประเมิน

ระดับคะแนน	1	2	3	4	5	รวม
1	0.2	0.6	0.4	0.2	0.6	2
2	0.2	0.6	0.4	0.2	0.6	2
3	0.2	0.6	0.4	0.2	0.6	2
4	0.1	0.3	0.2	0.1	0.3	1
5	0.3	0.9	0.6	0.3	0.9	3
รวม	1	0	2	1	3	10

กำหนดน้ำหนักตามความแตกต่างของการให้คะแนนตามแนวทแยงของเมตริกซ์โดยคะแนนที่ต่างกันมากที่สุดมีน้ำหนักมากที่สุด

ระดับคะแนน	1	2	3	4	5
1	0	1	2	3	4
2	1	0	1	2	3
3	2	1	0	1	2
4	3	2	1	0	1
5	4	3	2	1	0

คำนวณสัมประสิทธิ์แคปปาแบบถ่วงน้ำหนัก (κ_w) ดังสมการ

$$\kappa_w = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}x_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}m_{ij}} \tag{2.4}$$

เมื่อ $\mathrm{w_{ij}}$ คือ คะแนนถ่วงน้ำหนัก x_{ij} คือ คะแนนสังเกตได้ และ m_{ij} คือ ค่าคาดหมาย

จะได้ $\kappa_w = 0.573171$

การแปลผลค่าสัมประสิทธิ์แคปปาแบบถ่วงน้ำหนักใช้หลักเกณฑ์เดียวกับการแปลผลสัมประสิทธิ์แคปปา ทั้งนี้ ผลการประมาณค่าของสัมประสิทธิ์แคปปาแบบถ่วงน้ำหนักจะมีค่าสูงกว่าการประมาณค่าของสัมประสิทธิ์แคปปาแบบปกติ (Soeken และ Prescott, 1986; Zaiontz, มปป.; Gisev และคณะ 2013)

ข้อสังเกตที่มักพบเกี่ยวกับสถิตินี้ คือ สัมประสิทธิ์แคปปาเป็นสถิติที่เหมาะสมที่สุดสำหรับการวัดความสอดคล้องจริงหรือไม่ หรือค่า κ เป็นดัชนีในการวัดความเที่ยงหรือไม่ (Conger, 2016) เนื่องจากสัมประสิทธิ์แคปปาคำนวณความสอดคล้องโดยตัดความสอดคล้องโดยบังเอิญออกและเปรียบเทียบกับความน่าจะเป็นสูงสุดในการเกิดความสอดคล้องที่ไม่บังเอิญ ทำให้สัมประสิทธิ์แคปปามักถูกนำไปเปรียบเทียบกับการคำนวณสัดส่วนความสอดคล้อง (Percentage agreement) ในแง่ของความยุ่งยากซับซ้อนของการคำนวณและการแปลความหมายค่าที่ได้ซึ่งมีเกณฑ์การแปลความหมายอยู่หลายเกณฑ์ดังที่ได้กล่าวไปแล้ว นอกจากนี้ ประเด็นที่ได้รับการวิพากษ์วิจารณ์บ่อยครั้งเกี่ยวกับสถิติแคปปา คือ ความไม่สมมาตรของริมขอบ (marginal asymmetry) (Warrens, 2014; Conger, 2016) ดังรูป 2.2 ซึ่ง Warrens (2014) ได้อธิบายเกี่ยวกับความสัมพันธ์ระหว่างการแจกแจงตามขอบกับค่าสัมประสิทธิ์แคปปาไว้ว่า หากผลรวมการประเมิน A และ B มีค่าใกล้เคียงกัน ค่าสัมประสิทธิ์แคปปาจะมีค่าต่ำลง ในทางตรงกันข้าม หากการแจกแจง A และ B ในรูปแตกต่างกันสูง ค่าสัมประสิทธิ์แคปปาจะมีค่าสูงขึ้น

คะแนน	1	2	รวม
1	3	2	5
2	1	4	5
รวม	4	6	10

A

B

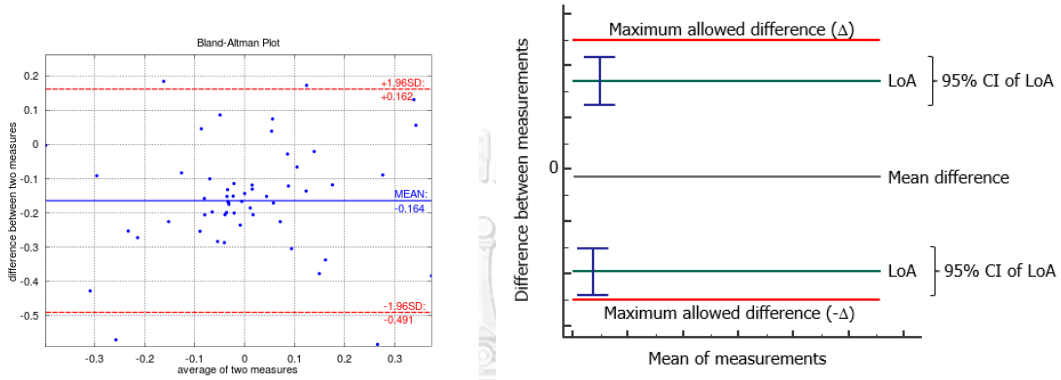
Rater marginal

รูป 2.2 ผลรวมความสอดคล้องและความแตกต่างของการประเมิน

<u>Limits of agreement</u>

Limits of agreement หรือ Bland-Altman plots เป็นวิธีการวิเคราะห์ความสอดคล้องในรูปแบบของการพล็อตคะแนนที่เสนอโดย Bland และ Altman (1986, 1999) สำหรับการวิเคราะห์ความสอดคล้องระหว่างวิธีการวัด 2 วิธี ในการศึกษาวิจัยทางการแพทย์ (Sedgwick, 2013; Gisev และคณะ, 2013) เพื่อใช้แทนการวิเคราะห์สัมประสิทธิ์สหสัมพันธ์แบบเพียร์สันที่มี

ข้อจำกัดเกี่ยวกับการประมาณค่าความสอดคล้องดังที่ได้กล่าวไปแล้ว ลักษณะของ Bland-Altman plots เป็นดังรูป 2.3 (ซ้าย) โดยแกน x เป็นค่าเฉลี่ยของคะแนนระหว่างผู้ประเมิน ((A+B)/2) และแกน y เป็นผลต่างของคะแนนระหว่างผู้ประเมิน (A-B) โดย 95% ของความสอดคล้อง หรือ limits of agreement จะอยู่ในช่วง $\pm 1.96\,SD$ (Myles และ Cui, 2007; Costa-Santos และคณะ, 2009; Gisev และคณะ, 2013)



รูป 2. 3 ผลการวิเคราะห์ Bland-Altman plots

(MedCalc Software, 2018)

รูป 2.3 (ซ้าย) แสดงให้เห็นว่าการประเมินของผู้ประเมิน A และ B ต่างกันโดยเฉลี่ย -0.164 คะแนน หมายความว่าการประเมินครั้งนี้มีความลำเอียงเกิดขึ้น ของขอบเขตความสอดคล้องในการประเมินนี้ คือ +0.162 และ -0.491 การพิจารณาความสอดคล้องถือว่าผลการประเมินระหว่างผู้ประเมินทั้ง 2 คนสอดคล้องกันเมื่อผลต่างระหว่างคะแนนอยู่ในช่วง 95% ความเชื่อมั่นของขอบเขตความสอดคล้อง ดังรูปที่ 2.3 (ขวา) ทั้งนี้ ช่วงความเชื่อมั่น (Confident Interval) ของค่าเฉลี่ยผลต่าง และ limits of agreement เป็นการอธิบายความน่าจะเป็นที่จะเกิดความคลาดเคลื่อนในการประมาณค่า (Giavarina, 2015)

<u>การวิเคราะห์สหสัมพันธ์ภายในชั้น (Intra-class correlation: ICC)</u>

สหสัมพันธ์ภายในชั้น หรือ Intra-class correlation (ICC) เป็นสถิติที่ใช้ในการรายงานความเที่ยงระหว่างผู้ประเมินที่นำเสนอโดย Fisher (1954) ซึ่งใช้อย่างแพร่หลายในการศึกษาวิจัยเกี่ยวกับการแพทย์เพื่อวัดความสอดคล้องระหว่างวิธีการวัดหรือเครื่องมือวัดหลายวิธีว่าได้ผลการวัดที่สอดคล้องกันหรือไม่ ต่อมาได้รับการประยุกต์ใช้กับการศึกษาเกี่ยวกับคุณภาพของเครื่องมือและวิธีการวัด เช่น การตรวจสอบความเที่ยงของเครื่องมือวัดด้วยวิธีการวัดซ้ำ (Test-retest

reliability) การตรวจสอบความเที่ยงภายในผู้ประเมิน (Intra-rater reliability) และการตรวจสอบความน่าเชื่อถือระหว่างผู้ประเมิน (Inter-rater reliability)

การวิเคราะห์ ICC อยู่บนพื้นฐานของการวิเคราะห์ความแปรปรวน (ANOVA) จุดเด่นของ ICC คือ ใช้วิเคราะห์ข้อมูลที่มีผู้ประเมินได้มากกว่า 2 คนขึ้นไป รวมถึงสามารถวิเคราะห์ข้อมูลที่มีค่าสูญหายได้ ค่า ICC มีค่าตั้งแต่ 0 (ไม่มีความสอดคล้อง) ถึง 1 (มีความสอดคล้องมากที่สุด) การวิเคราะห์ ICC นักวิจัยต้องออกแบบการวิจัยและเลือกโมเดลสำหรับการวิเคราะห์ตามเกณฑ์ต่อไปนี้ (Gisev และคณะ, 2013)

1) ผู้ประเมินเป็นกลุ่มเดียวกันในทุกหน่วยตัวอย่างหรือไม่

2) ผู้ประเมินถูกเลือกอย่างเจาะจงหรือสุ่มจากประชากร

3) มุ่งเน้นการตรวจสอบคุณภาพของผู้ประเมินคนเดียวหรือหลายคน

4) มุ่งเน้นการตรวจสอบความสอดคล้อง (Agreement) หรือความคงที่ (Consistency)

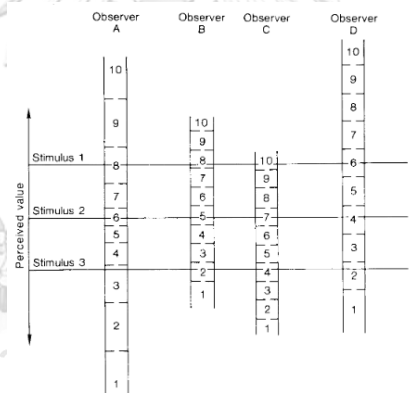
ทั้งนี้ เกณฑ์ข้อที่ 1) และ 2) จะใช้สำหรับการเลือกโมเดลในการวิเคราะห์ เกณฑ์ข้อที่ 3) จะใช้สำหรับการเลือกประเภทของผู้ประเมิน และเกณฑ์ข้อที่ 4) จะสะท้อนการนิยามความหมายของสัมประสิทธิ์สหสัมพันธ์ภายในชั้นที่ต้องการศึกษา (สายวรุณ สุกก่ำ เอกสิริ แก่นศักดิ์ศิริ และอุทุมพร โดมทอง, มปป.) โมเดลการวิเคราะห์ ICC มีด้วยกัน 3 โมเดลใหญ่ ได้แก่ 1) One-way random effects model เป็นโมเดลที่หน่วยตัวอย่างได้รับการประเมินจากผู้ประเมินต่างกัน และผู้ประเมินได้รับการสุ่มจากกลุ่มประชากร 2) Two-way random effects model เป็นโมเดลที่หน่วยตัวอย่างได้รับการประเมินโดยผู้ประเมินคนเดียวหรือกลุ่มเดียวกันซึ่งสุ่มจากกลุ่มประชากร และ 3) Two-way mixed model เป็นโมเดลในการวิเคราะห์ที่หน่วยตัวอย่างได้รับการประเมินจากผู้ประเมินคนเดียวหรือกลุ่มเดียวกันซึ่งผู้ประเมินได้มาจากการเลือกแบบเจาะจง (McGraw และ Wong, 1996; Gisev และคณะ, 2013)

Muller และ Buttner (1994) ชี้ว่าการวิเคราะห์สหสัมพันธ์ภายในชั้นตั้งอยู่บนพื้นฐานของการวิเคราะห์ความแปรปรวนนั้นเป็นข้อจำกัดที่สำคัญของ ICC ประการหนึ่ง ข้อมูลที่มีความแปรปรวนสูงจะให้ค่า ICC ที่สูงเมื่อเทียบกับข้อมูลที่มีระดับความสอดคล้องใกล้เคียงกันแต่มีความแปรปรวนต่ำกว่า Costa-Santos และคณะ (2009) ศึกษาเปรียบเทียบการแปลผลการประมาณค่าความสอดคล้องระหว่างสถิติ Limits of Agreement (LA) และ ICC ในการให้คะแนนสุขภาพของ

ทารกแรกคลอด (Apgar score) โดยสูตินรีแพทย์เชี่ยวชาญ พบว่า การแปลผลของ ICC ให้การประมาณค่าความสอดคล้องสูงกว่า LA ในขณะที่ LA ให้การประมาณค่าที่สมเหตุสมผลกับความเป็นจริงมากกว่าเนื่องจากข้อมูลในการศึกษาครั้งนี้มีความแปรปรวนสูงทำให้การประมาณค่าสูงเกินจริง ทั้งนี้ การศึกษาดังกล่าวพบความไม่คงที่ในการแปลผลของสถิติ LA และ ICC โดยผู้วิจัยแนะนำให้ใช้การประมาณค่าดังกล่าวร่วมกับการประมาณค่าโดยวิธีอื่นเพื่อผลการประมาณค่าที่น่าเชื่อถือมากยิ่งขึ้น

<u>การปรับคะแนนการประเมิน (Score adjustment)</u>

Meadows และ Billington (2005) กล่าวถึงวิธีการตรวจสอบและปรับแก้ความไม่คงที่ของการให้คะแนนในการสังเคราะห์เอกสารเกี่ยวกับวิธีการตรวจสอบความเที่ยงระหว่างผู้ประเมิน ได้แก่ การปรับคะแนน (Score adjustments) ซึ่งมีแนวคิดว่าผู้ตรวจแต่ละคนมีมโนทัศน์เกี่ยวกับคุณลักษณะหรือมิติการประเมินตรงกัน แต่มีเกณฑ์ในการให้คะแนนไม่เท่ากัน ดังรูป 2.4



รูป 2.4 ความแตกต่างของเกณฑ์การประเมินระหว่างผู้ตรวจ

(Brown และ Daniel, 1990)

จากรูปจะเห็นว่าผู้ประเมิน A มีช่วงการให้คะแนนกว้างแต่มีเทรชโฮลด์ในการเปลี่ยนระดับคะแนนไม่เท่ากัน การเปลี่ยนระดับคะแนนจาก 1 ไป 2 และ 9 ไป 10 นั้นมีช่วงกว้างกว่าการเปลี่ยนระดับคะแนนในช่วง 4 ไป 5 ผู้ประเมิน B และ C มีช่วงการให้คะแนนในช่วงจำกัด และผู้ประเมิน B มีแนวโมกดคะแนนมากกว่าผู้ประเมิน C ในขณะที่ผู้ประเมิน D มีเทรชโฮลด์ในการเปลี่ยนระดับคะแนนใกล้เคียงกันในแต่ละระดับ แต่มีแนวโน้มกดคะแนนมากกว่าผู้ประเมิน A ดังนั้นในการให้คะแนนคุณลักษณะที่ 3 ดังตัวอย่างในรูป 2.4 ผู้ประเมิน A ให้ 3 คะแนน ผู้ประเมิน C ให้ 4 คะแนน ในขณะที่ผู้ประเมิน B และ D ให้เพียง 2 คะแนน Brown และ Daniel (1990) เสนอกระบวนการปรับ

คะแนนทั้งหมด 11 วิธี ได้แก่ 1) Median rating 2) Mean rating 3) Origin-Adjusted rating (OAR) 4) Baseline-Adjusted OAR 5) Z-Score 6) Baseline-Adjusted Z-Score (BZ-Score) 7) Least Squares rating (LSR) 8) Baseline-Adjusted LSR 9) Scenic Beauty Estimate (SBE) 10) By-stimulus SBE และ 11) By-Observer SBE Brown และ Daniel แนะนำว่าการเลือกใช้กระบวนการปรับคะแนนขึ้นอยู่กับการออกแบบการให้คะแนน วัตถุประสงค์ของการวัด และข้อตกลงเบื้องต้นของกระบวนการปรับคะแนน ยกตัวอย่างเช่น การปรับคะแนนแบบ median rating เหมาะกับการประเมินแบบเรียงอันดับ ในขณะที่หากการประเมินเป็นการให้คะแนนแบบต่อเนื่อง ควรใช้การปรับคะแนนแบบ Z-Score, LSR หรือ SBE มากกว่า ซึ่งการปรับคะแนนแบบ LSR มีข้อดีคือให้ผลการปรับคะแนนบนสเกลดั้งเดิมของการประเมิน ในขณะที่ Z-Score จะให้ผลการปรับคะแนนบนสเกลคะแนนมาตรฐาน นอกจากนี้ การปรับคะแนนแบบ LSR และ Z-Score มีข้อตกลงเบื้องต้นเกี่ยวกับความเท่ากันของช่วงการประเมิน แต่ SBE ไม่มีข้อตกลงนี้ ดังนั้นนักวิจัยควรทราบลักษณะของข้อมูลเพื่อเลือกกระบวนการปรับคะแนนที่เหมาะสม ในการศึกษาของ Meadows และ Billington (2005) กล่าวถึงข้อสังเกตของการปรับคะแนนจากเอกสารงานวิจัยต่างๆ ได้แก่ ความแตกต่างของคะแนนที่มีนัยสำคัญหรือจำเป็นต้องใช้กระบวนการปรับคะแนน ความสัมพันธ์ระหว่างสัดส่วนของจำนวนตัวอย่างคะแนนกับการประมาณค่าความเที่ยง Baird และ Mac (1999) เสนอว่าควรทำการปรับคะแนนทุกกรณีแม้ว่าจะมีความแตกต่างน้อย ส่วนการศึกษาของ Al-Bayatti (2005) พบว่าการประมาณค่าความเที่ยงมีความสัมพันธ์ทางบวกกับจำนวนตัวอย่างคะแนนในการวิเคราะห์และระดับความสามารถของผู้ประเมิน โดยการศึกษาสัดส่วนของจำนวนตัวอย่างคะแนนกับผู้ประเมินที่เป็นนักศึกษาปริญญาตรี ครูฝึกสอน และผู้ประเมินเชี่ยวชาญ พบว่ากลุ่มผู้ประเมินเชี่ยวชาญใช้จำนวนตัวอย่างคะแนนน้อยที่สุดในการประมาณค่าความเที่ยง

<u>ทฤษฎีสรุปอ้างอิงความน่าเชื่อถือของผลการวัด (Generalizability Theory: G-Theory)</u>

ทฤษฎีสรุปอ้างอิงความน่าเชื่อถือของผลการวัด เป็นการวิเคราะห์ความเที่ยงหรือความน่าเชื่อถือของผลการวัดในสถานการณ์ต่าง ๆ โดยใช้วิธีการวิเคราะห์ความแปรปรวน (ANOVA) ทฤษฎีสรุปอ้างอิงความน่าเชื่อถือของผลการวัดเป็นวิธีการที่เสนอโดยครอนบาคและคณะ (Cronbach, Gleser, Nanda & Rajaratnam, 1972) (Shavelson และ Webb, 1991) มีแนวคิดเกี่ยวกับการแยกความคลาดเคลื่อนของการทดสอบออกเป็นแหล่งความคลาดเคลื่อนต่าง ๆ ทั้งความคลาดเคลื่อนอย่างเป็นระบบ (systematic error) และความคลาดเคลื่อนอย่างสุ่ม (unsystematic error) และประมาณค่าความคลาดเคลื่อนเหล่านั้น โดยถือว่าคะแนนจากการวัดถือ

เป็นการสุ่มจากเอกภาพของการทดสอบซึ่งประกอบด้วยค่าสังเกตที่เป็นไปได้ทั้งหมดของการวัดที่เกิดขึ้นในสถานการณ์ต่างๆ ที่เกี่ยวข้องกับการวัดหรือการทดสอบ เช่น ลักษณะของแบบสอบ ข้อสอบ ผู้ตรวจ หรือบริบทของการทดสอบ องค์ประกอบเหล่านี้เรียกรวมกันว่า ฟาเซท (facet) ของการวัด ทฤษฎีสรุปอ้างอิงความน่าเชื่อถือของผลการวัดทำการประเมินเงื่อนไขที่เกี่ยวข้องกับการวัดด้วยการจำแนกและประมาณค่าความแปรปรวนของผลการวัดที่เกิดขึ้นภายใต้แหล่งความคลาดเคลื่อนต่างๆ เพื่อศึกษาสถานการณ์ของการวัดหรือการทดสอบที่คุ้มค่าและนำไปใช้ได้จริงในการทดสอบที่จะให้ผลการวัดที่ถูกต้องและแม่นยำมากที่สุด จากนั้นนำผลการวิเคราะห์ไปใช้ประกอบการตัดสินใจเลือกหรือออกแบบสถานการณ์การทดสอบที่เหมาะสมที่สุดสำหรับแต่ละวัตถุประสงค์ของการทดสอบ (Shavelson และ Webb, 2001; 2005; Webb, Shavelson และ Haertel, 2006)

การออกแบบการวิเคราะห์โดยทฤษฎีสรุปอ้างอิงความน่าเชื่อถือของผลการวัดขึ้นอยู่กับองค์ประกอบที่ผู้วิเคราะห์คาดว่าจะส่งผลต่อความคลาดเคลื่อนของการวัด หรือ ฟาเซท (facet) ตัวอย่างเช่น ต้องการศึกษาความน่าเชื่อถือของผลการทดสอบที่มาจากผู้ตรวจจำนวน 1, 2, 3, 4, หรือ 5 คน ในการตรวจการเขียนความเรียงภาษาไทยของนักเรียน ฟาเซทในการศึกษาครั้งนี้ก็คือ จำนวนผู้ประเมิน (rater) หรือในกรณีที่ต้องการศึกษาความน่าเชื่อถือของการทดสอบวิชาคณิตศาสตร์ที่สอบในช่วงเวลาต่างกันและใช้ข้อสอบที่มีความยาวต่างกัน ฟาเซทในการศึกษาจะเป็น ช่วงเวลาในการทดสอบ (occasion) และความยาวของแบบสอบ (item) จากนั้นออกแบบเอกภพของการทดสอบสำหรับฟาเซทที่ต้องการศึกษาในรูปแบบ crossed design คือ ผู้สอบทุกคนได้รับเงื่อนไขของการวัดเดียวกันเหมือนกันทุกคน หรือรูปแบบ nested design คือ ผู้สอบแต่ละคนได้รับเงื่อนไขของการวัดแตกต่างกัน (Shavelson และ Webb, 2001; 2005; ศิริชัย กาญจนวาสี, 2555) ตัวอย่างการจำแนกแหล่งความผันแปรและองค์ประกอบความแปรปรวนของการวัดหนึ่งฟาเซทเป็นดังตาราง 2.3

ตาราง 2.3 แหล่งความผันแปรและองค์ประกอบความแปรปรวนของการวัดหนึ่งฟาเซท (Shavelson และ Webb, 1991)

แหล่งความผันแปร	ประเภทของความผันแปร	องค์ประกอบความแปรปรวน
Persons (p)	Universe score	σ_p^2
Items (i)	Error	σ_i^2
Residual (pi, e)	Error	$\sigma_{pi,e}^2$

กำหนดให้คะแนนสังเกตได้ของผู้สอบเป็น X_{pi} จะได้จำแนกส่วนประกอบของคะแนนได้ดังนี้

$$
\begin{aligned}
X_{pi} = \quad & \mu && \text{(ค่าเฉลี่ยรวม)} \\
& + \mu_p - \mu && \text{(อิทธิพลของบุคคล)} \\
& + \mu_i - \mu && \text{(อิทธิพลของข้อสอบ)} \\
& + X_{pi} - \mu_p - \mu_i + \mu && \text{(เศษเหลือ)}
\end{aligned}
\tag{2.5}
$$

จากสมการ 2.5 จะได้องค์ประกอบของความแปรปรวน เป็น $\sigma^2_{(X_{pi})} = \sigma^2_p + \sigma^2_i + \sigma^2_{pi,e}$ หรือความแปรปรวนของคะแนนที่สังเกตได้ประกอบด้วยความแปรปรวนของคะแนนจากเอกภพ ความแปรปรวนจากคะแนนเฉลี่ยรายข้อ และความแปรปรวนจากความคลาดเคลื่อน

Baker และคณะ (2008) ใช้การวิเคราะห์ตามทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัดเพื่อศึกษาความน่าเชื่อถือของการตรวจให้คะแนนการทดสอบ KS3 ระหว่างผู้ประเมินชาวอังกฤษกับผู้ประเมินในประเทศอื่น พบว่าผู้ประเมินชาวออสเตรเลียมีรูปแบบการประเมินรวมถึงให้คะแนนการประเมินใกล้เคียงกับผู้ประเมินชาวอังกฤษซึ่งเป็นเจ้าของแบบทดสอบซึ่งสรุปได้ว่าเนื่องมาจากความใกล้เคียงกันด้านสังคมและวัฒนธรรม นอกจากนี้ยังพบว่าอิทธิพลของข้อสอบเป็นแหล่งความคลาดเคลื่อนที่สำคัญที่สุดในการศึกษาดังกล่าว เขาได้แสดงความคิดเห็นในการศึกษาว่า G-Theory เป็นวิธีการที่ใช้ในการสร้างรูปแบบของอิทธิพลในการสร้างความแตกต่างให้กับปัจจัยในการทดสอบ เช่น จำนวนข้อสอบ การกำหนดเกณฑ์ หรือการให้คะแนน เป็นต้น นอกจากนี้ G-Theory ยังให้สารสนเทศที่เป็นประโยชน์ต่อการปรับปรุงและพัฒนารูปแบบการทดสอบและเพิ่มความน่าเชื่อถือของผลการทดสอบในภาพรวม ซึ่งสอดคล้องกับความคิดเห็นของ Newton (2009) ที่กล่าวว่า ทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัด (G-Theory) เป็นวิธีการศึกษาความเที่ยงที่ให้ประโยชน์จากการวิเคราะห์มากกว่าการศึกษาที่ใช้การวิเคราะห์ความแปรปรวนอื่นๆ เนื่องจากมีข้อตกลงเบื้องต้นไม่มาก และให้สารสนเทศเชิงปริมาณเกี่ยวกับแหล่งความคลาดเคลื่อนของการวัดที่แตกต่างกัน รวมถึงสามารถรวมความคลาดเคลื่อนที่เกี่ยวกับข้อสอบและการให้คะแนนมารวมในการศึกษาในโมเดลเดียวกันได้ (Baird และคณะ, 2012; Bramley และ Dhawan, 2012)

<u>ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT)</u>

ทฤษฎีการตอบสนองข้อสอบแบบตรวจให้คะแนนหลายค่าได้รับการยอมรับว่าเป็นวิธีการที่มีประสิทธิภาพในการตรวจสอบความหลากหลายของปัจจัยที่ส่งผลต่อการประเมินอันเนื่องมาจากผู้ประเมิน และได้มีการศึกษาการประยุกต์ใช้ทฤษฎีการตอบสนองข้อสอบโมเดลต่าง ๆ สำหรับการตรวจสอบคุณภาพการประเมินของผู้ประเมิน Peterson (2013) อธิบายเกี่ยวกับการสร้างรูปแบบความสอดคล้องระหว่างผู้ประเมินด้วยโมเดลการตอบสนองข้อสอบในกรณีที่ผู้ประเมินให้คะแนนแบบ 2 ค่า ว่ามีลักษณะเป็นโมเดลแบบ 2 พารามิเตอร์ ดังสมการ

$$P(X_{vi} = 1|\theta_v) = \frac{\exp\,(\alpha_i\theta_v - \beta_i)}{1 + exp(\alpha_i\theta_v - \beta_i)} \qquad (2.6)$$

เมื่อ $\theta = (\theta_1, \dots, \theta_n)^t \in R^n$ เป็นลักษณะของบุคคล และ $\beta = (\beta_1, \dots, \beta_k)^t \in R^k, \alpha = (\alpha_1, \dots, \alpha_k)^t \in R^k$ เป็นลักษณะของผู้ประเมิน โดยถือว่า X_{vi} ทุกตัวเป็นเงื่อนไขความเป็นอิสระของ α, β และ θ

พารามิเตอร์ θ_v มีค่าตั้งแต่ $-\infty$ ถึง ∞ และเป็นตัวกำหนดความยากของการประเมินในกรณีต่างๆ กรณีที่การประเมินในข้อนั้นง่าย θ จะมีค่าอยู่บริเวณสุดทางด้านใดด้านหนึ่ง ซึ่งความน่าจะเป็นของการประเมินทางบวกจะอยู่บริเวณใกล้ 0 หรือใกล้ 1 ในกรณีของข้อที่ประเมินได้ยาก θ จะมีค่าใกล้ 0 ซึ่งความน่าจะเป็นของการประเมินทางบวกจะอยู่บริเวณ 1.5 ขึ้นอยู่กับพารามิเตอร์ของผู้ประเมิน $\alpha = (\alpha_1, \dots, \alpha_k)^t \in R^k$ เป็นพารามิเตอร์อำนาจจำแนก แสดงความน่าจะเป็นของผู้ตรวจในการตอบสนองต่อความสามารถของผู้สอบต่างกัน พารามิเตอร์นี้อาจแปลความหมายเป็นระดับความเชี่ยวชาญของผู้ประเมิน ในการประเมินที่มีผู้เชี่ยวชาญสูงกว่า จะมีความน่าจะเป็นของการประเมินทางบวกเมื่อ θ เพิ่มขึ้นมากกว่า นอกจากนั้น เมื่อมีความลังเลหรือไม่แน่ใจในการประเมินผู้ประเมินที่มีความเชี่ยวชาญจะมีการกระจายของ θ ต่ำกว่า กราฟที่มีความชันมากกว่าแสดงว่ามีความสอดคล้องของการประเมินมากกว่ากราฟที่มีความชันน้อย

Wang, Engelhard และ Wolfe (2015) ศึกษาการใช้โมเดลการตอบสนองข้อสอบแบบ Unfolding model ในการตรวจสอบความแม่นยำของผู้ประเมิน โดยเปรียบเทียบความแตกต่างระหว่างคะแนนการประเมินกับการให้คะแนนของผู้เชี่ยวชาญในการตรวจให้คะแนนแบบ 2 ค่า โดยศึกษาใน 3 กลุ่ม คือ การประเมินที่มีความคลาดเคลื่อนที่ต่ำกว่าคะแนนผู้เชี่ยวชาญ การประเมินที่แม่นยำ และการประเมินที่มีความคลาดเคลื่อนสูงกว่าคะแนนผู้เชี่ยวชาญ ใช้ Hyperbolic cosine

model (HCM) ประเมินความแม่นยำของการให้คะแนน Unfolding model ใช้ในการวิเคราะห์ระยะห่างระหว่างคะแนนสังเกตได้กับคะแนนจากผู้เชี่ยวชาญ Engelhard (1996, อ้างใน Wang, Engelhard และ Wolfe, 2015) เสนอโมเดลความแตกต่างระหว่างการประเมินดังสมการ

$$A_{ni} = \max\{|R_{ni} - B_i|\} - |R_{ni} - B_i| \tag{2.7}$$

เมื่อ $\mathrm{R_{ni}}$ เป็นคะแนนการประเมินที่สังเกตได้ของผู้ประเมิน n ที่ประเมินข้อสอบฉบับที่ i และ $\mathrm{B_i}$ คือคะแนนจากผู้เชี่ยวชาญที่ประเมินข้อสอบฉบับที่ i ดังนี้ ค่าที่เป็นไปได้ทั้งหมดของการประเมินที่แม่นยำจะมีทิศทางบวก โมเดล Unfolding แตกต่างจากโมเดลการตอบสนองข้อสอบแบบปกติหรือโมเดลสะสม (cumulative response model) โมเดลสะสมจะมีลักษณะเป็นฟังก์ชันที่มีค่าเพิ่มขึ้นทิศทางเดียว แต่โมเดล Unfolding จะมีลักษณะของฟังก์ชันเพิ่มและลด โดยผู้ประเมินอาจจะมีความแม่นยำเมื่อประเมินข้อสอบหรืองานเขียนบางลักษณะแต่อาจจะลดความแม่นยำลงเมื่อประเมินงานที่ประเมินได้ยากกว่า

Hyperbolic Cosine Model (HCM) เป็นวิธีการสร้างรูปแบบการตอบสนองข้อสอบของ Unfolding IRT วิธีหนึ่งซึ่งใช้ฟังก์ชันทางคณิตศาสตร์ $\cosh(x)$ ในการคลี่คลายรูปแบบการประเมิน HCM มีที่มาจากราสช์โมเดล ทั้งนี้ การศึกษาของ Wang และคณะ (2015) เสนอว่าวิธีการที่แก้ปัญหาข้อจำกัดของการวิเคราะห์ความเที่ยงระหว่างผู้ประเมินด้วย Many Facet Rasch Model ได้

Wind และ Engelhard Jr. (2015) การศึกษาการใช้ Mokken Scale Analysis (MSA) ในการประเมินคุณภาพของผู้ประเมิน Mokken Scale Analysis เป็นแนวคิดของ Mokken (Mokken, 1971 อ้างใน Wind และ Engelhard Jr., 2015) ที่พัฒนาจากเทคนิคการประเมินคุณภาพของข้อมูลที่ได้จากมาตรประมาณค่าในการศึกษาวิจัยทางรัฐศาสตร์ สาธารณสุข และการศึกษาตัวแปรด้านจิตพิสัย MSA เป็นโมเดลการตอบสนองข้อสอบแบบนอนพาราเมตริกซ์ซึ่งฟังก์ชันการตอบสนองข้อสอบ (item response function: IRF) ไม่มีรูปแบบตายตัว จึงถือว่าเป็นโมเดลการวิเคราะห์ที่เข้มงวดน้อยกว่าโมเดลการตอบสนองข้อสอบแบบพาราเมตริกซ์ในแง่ของข้อตกลงเบื้องต้นระหว่างตัวแปรแฝงและความน่าจะเป็นของการเลือกคำตอบ แต่ยังคงตั้งอยู่บนข้อตกลงเกี่ยวกับความเป็นเอกมิติ (unidimensionality) ความสัมพันธ์แบบทิศทางเดียว (monotonicity) และความไม่แปรเปลี่ยน (invariance) เช่นเดียวกับโมเดลการตอบสนองข้อสอบอื่น ๆ Mokken model มีข้อกำหนดว่าความน่าจะเป็นของการตอบต้องเป็นฟังก์ชันแบบไม่ลด (nondecreasing function) ทั้งนี้ MSA สำหรับการตรวจให้คะแนนแบบ 2 ค่า ประกอบด้วยโมเดล Monotone Homogeneity model

(MH) ที่มีข้อตกลงเบื้องต้นเกี่ยวกับความเป็นเอกมิติ เงื่อนไขความเป็นอิสระ และความสัมพันธ์แบบทิศทางเดียว และ Double Monotonicity model (DM) มีข้อตกลงเพิ่มเติมเกี่ยวกับความคงที่ของระดับความยากข้ามระดับความสามารถของผู้สอบ ส่วน MSA สำหรับการตรวจให้คะแนนแบบหลายค่าประกอบด้วย 2 โมเดล เช่นเดียวกันและมีลักษณะใกล้เคียงกับ polytomous Rasch model (Wind, 2015) ในการศึกษาของ Wind และ Engelhard Jr. (2015) ได้ประยุกต์ใช้ MSA สำหรับการตรวจให้คะแนนแบบหลายค่าในการประเมินคุณภาพของการตรวจให้คะแนนบนมาตรประมาณค่า (rating scale) การศึกษาดังกล่าวได้เสนอตัวบ่งชี้คุณภาพของผู้ประเมินที่ได้จากการประยุกต์ Monotone Homogeneity model (MH) เป็นโมเดลวิเคราะห์คุณภาพผู้ประเมิน Monotone Homogeneity for Ratings model (MH-R) และ Double Monotonicity for Ratings model (DM-R) ที่ได้จากการประยุกต์ใช้ Double Monotonicity model (DM) ดัชนีคุณภาพผู้ประเมินดังกล่าวให้สารสนเทศเกี่ยวกับการวินิจฉัยคุณภาพการประเมินในแง่ของการตรวจสอบความไม่แปรเปลี่ยน (invariance) ของผู้ประเมิน Wind (2017) ได้เสนอกระบวนการในการประยุกต์ใช้ MSA ในการวัดและประเมินผลทางการศึกษา พร้อมเสนอแนะว่า ควรใช้ MSA สำหรับการประเมินทางการศึกษาเมื่อนักวิจัยสนใจศึกษาระดับความสอดคล้องระหว่างการประเมินกับลักษณะพื้นฐานของการวัด เช่น ในกระบวนการพัฒนาหรือการปรับปรุงการประเมิน หรือในกรณีที่ต้องการศึกษาเกี่ยวกับคุณภาพของการประเมินและต้องการสารสนเทศเกี่ยวกับการจัดลำดับของบุคคลหรือข้อสอบแต่มีตัวอย่างไม่เพียงพอสำหรับการวิเคราะห์การตอบสนองข้อสอบแบบพาราเมตริกซ์

การศึกษาการวิเคราะห์ความเที่ยงระหว่างผู้ประเมินโดยใช้โมเดลการตอบสนองข้อสอบ IRT Facets models (Bock, Brennan และ Muraki, 1999; Wilson และ Hoskens, 2001) พบว่า เมื่อจำนวนผู้ประเมินเพิ่มขึ้น IRT Facets models จะประมาณค่าความสามารถของผู้สอบ (θ) ไปอย่างไม่สิ้นสุดแม้จะไม่มีการตอบข้อสอบแล้วก็ตาม และยังพบว่า IRT Facets models มีความลำเอียงในการประมาณค่าที่ต่ำลงเมื่อจำนวนผู้ประเมินเพิ่มขึ้น โดยประมาณค่าต่ำกว่าค่า (θ) ของการตอบถูกเมื่อไม่มีผู้ประเมิน Patz และคณะ (2002) ได้ศึกษาการประยุกต์ใช้ Hierarchical Rater Model (HRM) สำหรับการตรวจให้คะแนนแบบหลายค่าโดยผู้ประเมินหลายคนเพื่อศึกษาการสร้างรูปแบบฉันทามติระหว่างผู้ประเมิน และสร้างรูปแบบของอิทธิพลการกดคะแนนและความคงที่ของผู้ประเมินรายบุคคล ซึ่งเป็นวิธีการหนึ่งที่ใช้ปรับแก้ปัญหาวิธีการรวบรวมสารสนเทศในการประเมินโดยผู้ประเมินหลายคนของรูปแบบฟาเซทภายในทฤษฎีการตอบสนองข้อสอบเพื่อประมาณค่าความสามารถของผู้สอบ

ทฤษฎีการตอบสนองข้อสอบมีจุดเด่นเกี่ยวกับคุณสมบัติความไม่แปรเปลี่ยนของพารามิเตอร์ (Meadows และ Billington, 2005; ศิริชัย กาญจนวาสี, 2555) ซึ่งทำให้สามารถปรับเทียบคะแนนของผู้สอบไปยังการทดสอบที่มีคุณลักษณะใกล้เคียงกันได้ นอกจากนี้ ยังเป็นโมเดลการวิเคราะห์ที่เสถียรต่อการละเมิดข้อตกลงเบื้องต้น อย่างไรก็ตาม IRT มีข้อจำกัดเช่นเดียวกับโมเดลการวิเคราะห์ขนาดใหญ่และซับซ้อนที่ต้องการกลุ่มตัวอย่างขนาดใหญ่สำหรับการวิเคราะห์ โดย Kean และ Reilly (2014) เสนอขนาดตัวอย่างสำหรับการวิเคราะห์โมเดล IRT ไว้ดังตาราง 2.4

ตาราง 2.4 ขนาดตัวอย่างสำหรับการวิเคราะห์โดยทฤษฎีการตอบสนองข้อสอบ (Kean & Reilly, 2014)

IRT models	Participants
One-parameter model (1PL)	150
Two-parameter model (2PL)	500
Three-parameter model (3PL)	มากกว่า 1000

Sahin และ Anil (2016) ศึกษาเกี่ยวกับอิทธิพลของขนาดตัวอย่างและความยาวของแบบสอบที่ส่งผลต่อพารามิเตอร์ของข้อสอบในการวิเคราะห์ IRT โดยศึกษาความยาวแบบสอบจำนวน 10 20 และ 30 ข้อ กับผู้สอบ 150 250 350 500 750 1,000 2,000 3,000 และ 5,000 คน ได้ผลสรุปของขนาดตัวอย่างที่เหมาะสมสำหรับการวิเคราะห์ IRT โมเดลต่าง ๆ ดังตาราง 2.5

ตาราง 2.5 ขนาดตัวอย่างสำหรับการวิเคราะห์โดยทฤษฎีการตอบสนองข้อสอบ (Sahin & Anil, 2016)

ความยาวแบบสอบ (ข้อ)	**จำนวนตัวอย่าง (คน)**		
	1PL	2PL	3PL
10	150	750	750
20	150	500	750
30	150	250	350

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

<u>การวิเคราะห์ Rasch Model</u>

ราชส์โมเดล (Rasch model) เป็นโมเดลการวัดทางจิตวิทยาสำหรับการวิเคราะห์ข้อมูลแบบจัดกลุ่มมีแนวคิดว่า ความน่าจะเป็นของการที่บุคคลจะตอบข้อสอบได้ถูกต้องขึ้นอยู่กับความสามารถ (θ) และระดับความยากของข้อสอบ (b_i) โดยมีข้อตกลงเบื้องต้นว่าข้อสอบมีค่าอำนาจจำแนกเท่าเทียมกัน ราสช์โมเดลเป็นรูปแบบหนึ่งของโมเดลการตอบสนองข้อสอบซึ่งโมเดลตั้งต้นเป็นฟังก์ชันโลจิสของความน่าจะเป็นในการตอบคำตอบได้ถูกต้องของบุคคลที่มีความสามารถต่างกันในข้อสอบที่มีความยากต่างกัน เมื่อนำมาประยุกต์ใช้กับการประเมินที่ใช้ผู้ประเมิน ราสช์โมเดลจะแปลความหมายเป็นความน่าจะเป็นของการประเมินในประเภทหรือระดับคะแนนบนมาตรการประเมินในฐานะฟังก์ชันการกดคะแนนของผู้ประเมินและผลสัมฤทธิ์ของผู้เรียน เมื่อทำการวัดความไม่แปรเปลี่ยนของผู้ประเมินแล้ว จะสามารถอธิบายความสามารถของผู้เรียนได้อย่างอิสระจากผู้ประเมินเช่นเดียวกับการกดคะแนนของผู้ประเมินที่สามารถอธิบายได้อย่างเป็นอิสระจากผู้เรียน (Wind และ Peterson, 2017)

การศึกษาเกี่ยวกับราสช์โมเดลกับการประเมินโดยใช้ผู้ประเมินมุ่งเน้นการประเมินระดับของการวัดความไม่แปรเปลี่ยนของผู้ประเมินโดยมีจุดประสงค์เพื่อปรับปรุงกระบวนการวัดและเป็นแนวทางในการตรวจสอบคุณภาพการประเมิน ในช่วงปี ค.ศ. 1990 Linacre Wright และ Lunz (1990) Engelhard (1994) Myford Marr และ Linacre (1996) ได้เสนอวิธีการประมาณค่าความเที่ยงโดยใช้การวิเคราะห์ Many-Facets Rasch model ซึ่งให้สารสนเทศมากกว่าและสามารถอ้างอิงผลการประมาณค่าได้อย่างมีประสิทธิภาพกว่าการวิเคราะห์ความเที่ยงระหว่างผู้ประเมินแบบเดิม

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Engelhard (1994) ศึกษาการปรับเทียบคะแนนโดยใช้ Many-Facets Rasch model (MFRM) ในความคลาดเคลื่อนของการประเมินรูปแบบต่าง ๆ ได้แก่ การกดคะแนน อิทธิพลฮาโล และการให้คะแนนในช่วงจำกัด พบว่าวิธีการดังกล่าวสามารถระบุความคลามดเคลื่อนของการประเมินและสามารถปรับเทียบคะแนนประเมินระหว่างผู้ประเมินในกลุ่มความคลาดเคลื่อนต่าง ๆ ได้ อย่างไรก็ตาม Myford Marr และ Linacre (1996) ทดลองใช้การวิเคราะห์ฟาเซทเพื่อเทียบการตรวจให้คะแนนการทดสอบการเขียนภาษาอังกฤษ (TWE) ของ Educational Testing Service (ETS) ในการตรวจความเรียงภาษาอังกฤษจำนวน 100 ฉบับ มีผู้ตรวจ 105 คน ในการตรวจครั้งที่ 1 และ 50 คนในการตรวจครั้งที่ 2 พบว่าข้อมูลมีจำนวนผู้ตรวจน้อยเกินไปทำให้การประมาณค่าการกดและปล่อยคะแนนขาดความเที่ยงตรงและขาดความคงที่ รวมถึงทำให้ไม่สามารถสรุปอิทธิพลของ

ผู้ตรวจที่มีต่อการให้คะแนนได้อย่างชัดเจน และไม่สามารถปรับเทียบอิทธิพลของผู้ตรวจได้ ทั้งนี้ Myford Marr และ Linacre ได้เสนอแนะว่าการวิเคราะห์ดังกล่าวควรใช้ข้อมูลที่มีขนาดใหญ่ขึ้นเพื่อทำให้การประมาณค่าพารามิเตอร์ทำได้อย่างมีประสิทธิภาพ

การศึกษาเอกสารเกี่ยวกับวิธีการประเมินคุณภาพการวัดและประเมินผลทางภาษาโดย Wind และ Peterson (2017) พบว่า ราสช์โมเดลเป็นวิธีการที่นิยมใช้มากเป็นอันดับแรกของการตรวจสอบคุณภาพระหว่างผู้ประเมินในรูปคะแนนมาตรฐานในการศึกษาที่เน้นการตรวจสอบคุณภาพของผู้ประเมิน เน้นการศึกษาโครงสร้างความสามารถทางภาษา (ฟัง พูด อ่าน เขียน) การศึกษาด้วยข้อมูลจำลอง และการออกแบบการประเมิน Myford และ Wolfe (2003; 2004) เสนอวิธีการตรวจสอบอิทธิพลของผู้ประเมิน เรียกว่า Many-Facet Rasch Measurement (MFRM) เป็นโมเดลที่ขยายจากราสช์โมเดลทำให้การกดคะแนนของผู้ประเมินเปรียบเทียบบนสเกลเดียวกันในรูปของความสามารถของผู้ประเมินและความยากของการประเมินทำให้เปรียบเทียบได้ว่าผู้ประเมินที่ให้คะแนนสังเกตได้เท่ากันอาจจะมีนัยความหมายของคะแนนต่างกันได้ เช่น คะแนน 3 คะแนนจากผู้ประเมิน A อาจเท่ากับคะแนน 5 คะแนนที่ประเมินโดยผู้ประเมิน B เมื่อนำมาเปรียบเทียบบนสเกล Schaefer (2008) วิเคราะห์ MFRM เพื่อสำรวจรูปแบบความลำเอียงของผู้ประเมินที่เป็นเจ้าของภาษาในการประเมินการเขียนของผู้เรียนภาษาอังกฤษเป็นภาษาต่างประเทศ พบรูปแบบความลำเอียงในการประเมินภายในกลุ่มย่อยของผู้ประเมิน โดยผู้ประเมินจะมีแนวโน้มกดหรือปล่อยคะแนนกับผู้สอบที่มีความสามารถสูงมากกว่าผู้สอบที่มีความสามารถต่ำกว่า Kaliski และคณะ (2012) ใช้การวิเคราะห์ MFRM ตรวจสอบคุณภาพการประเมินของคณะกรรมการจำนวน 15 คน ในการทดสอบการเลื่อนระดับวิชาวิทยาศาสตร์สิ่งแวดล้อม โดยศึกษาอิทธิพลเกี่ยวกับผู้ประเมิน 4 ฟาเซท คือ 1) การกด/ปล่อยคะแนน 2) ความยากของการประเมิน 3) รอบการประเมิน 4) จุดตัดของคะแนนการประเมินโดยคณะกรรมการ การวิเคราะห์ MFRM ให้สารสนเทศเกี่ยวกับคุณภาพของผู้ประเมินในฟาเซททั้ง 4 ฟาเซทดังกล่าว รวมถึงให้สารสนเทศเกี่ยวกับการทำหน้าที่ต่างกันของผู้ประเมินเพศชายกับเพศหญิง ผู้ประเมินที่เป็นอาจารย์มหาวิทยาลัยกับผู้ประเมินที่เป็นอาจารย์โรงเรียนมัธยม

Wang และคณะ (2015) แสดงความเห็นเกี่ยวกับการวิเคราะห์ MFRM ว่ามีข้อจำกัดเกี่ยวที่ไม่จำแนกทิศทางของความไม่แม่นยำว่าการประเมินที่สังเกตได้นั้นต่ำกว่าหรือสูงกว่าการประเมินของผู้เชี่ยวชาญ นอกจากนี้ยังไม่ให้สารสนเทศเกี่ยวกับขอบเขตของความแม่นยำของผู้ประเมิน และได้เสนอการวิเคราะห์ความแม่นยำของการประเมินโดยใช้ Unfolding model ดังที่กล่าวไปแล้ว

การสรุปอ้างอิงความน่าเชื่อถือของผลการวัด (G-Theory) และ Many-Facets Rasch model (MFRM) มีจุดเด่น คือ สามารถใช้วิเคราะห์ข้อมูลที่มาจากหลายแหล่งความคลาดเคลื่อนได้ โดย G-Theory ใช้การวิเคราะห์ความคลาดเคลื่อนจากแหล่งความคลาดเคลื่อนต่าง ๆ แล้วเปรียบเทียบอิทธิพลของแต่ละแหล่งเพื่อประมาณค่าความเที่ยงในสถานการณ์การทดสอบในกรณีต่างๆ เช่น จำนวนผู้ตรวจต่างกัน ความยาวข้อสอบต่างกัน เป็นต้น ในขณะที่ MFRM ใช้วิธีการเปรียบเทียบความสอดคล้องระหว่างโมเดลกับข้อมูลเพื่อประมาณค่าความเที่ยงของผู้ตรวจและแสดงให้เห็นว่าผู้ตรวจแต่ละคนให้คะแนนข้อสอบแตกต่างกันอย่างไร ทั้งนี้ การวิเคราะห์ MFRM อยู่บนข้อตกลงเบื้องต้นว่าข้อมูลต้องไม่มีความคลาดเคลื่อน Baird และคณะ (2012) ศึกษาเปรียบเทียบผลการวิเคราะห์ระหว่าง G-Theory MFRM และ Multilevel modelling (MLM) พบว่า G-Theory และ MLM ให้สารสนเทศเกี่ยวกับความอิทธิพลของความแปรปรวนของผู้ตรวจที่ ข้อคำถาม และแหล่งอิทธิพลอื่นที่ส่งผลต่อความแปรปรวนของคะแนน ในขณะที่ MFRM ให้สารสนเทศเกี่ยวกับอิทธิพลระหว่างผู้ตรวจและความเที่ยงของการวัด อย่างไรก็ตาม MFRM ในการศึกษาครั้งนี้พบว่ามีความสอดคล้องกับข้อมูลต่ำเนื่องจากการมีปฏิสัมพันธ์ระหว่างตัวแปรดังที่ปรากฏในการวิเคราะห์ G-Theory ซึ่งโมเดล MFRM ไม่สามารถวิเคราะห์ได้

การศึกษาดังกล่าวทำให้เห็นถึงจุดเด่นและข้อจำกัดของการวิเคราะห์ความเที่ยงของ G-Theory และ MFRM ที่มีข้อตกลงเบื้องต้นต่างกันและให้สารสนเทศที่ต่างกัน ดังนั้น การที่นักวิจัยต้องการสารสนเทศที่หลากหลายเพื่อตอบคำถามวิจัยเกี่ยวกับความเที่ยงระหว่างผู้ประเมิน จึงจำเป็นต้องให้สถิติวิเคราะห์หลายขั้นตอนเพื่อตอบคำถามวิจัยให้ครบถ้วน จากเอกสารที่เกี่ยวข้องสรุปได้ว่า การศึกษาความสอดคล้องระหว่างผู้ประเมินจะใช้การศึกษาสัดส่วนความสอดคล้อง การศึกษาความคงที่ระหว่างผู้ประเมินใช้การวิเคราะห์สัมประสิทธิ์แคปปา การวิเคราะห์สหสัมพันธ์ภายในชั้น (ICC) และการหาสัมประสิทธิ์สหสัมพันธ์ ในขณะที่การวิเคราะห์ความคลาดเคลื่อนของการวัดจะใช้โมเดลทางสถิติขั้นสูงที่ซับซ้อนขึ้น เช่น โมเดลการตอบสนองข้อสอบ ราสช์โมเดล รวมถึงการวิเคราะห์ MFRM เพื่อศึกษาการทำหน้าที่ต่างกันของผู้ประเมิน

ตอนที่ 2 การทำหน้าที่ต่างกันของผู้ประเมิน

2.1 ความหมายของการทำหน้าที่ต่างกันของผู้ประเมิน

การศึกษาโมเดลการตอบสนองข้อสอบมีการพัฒนาวิธีการรวบรวมข้อมูลจากฟาเซทต่าง ๆ เรียกว่า ฟาเซทโมเดล (Linacre, 1989) ซึ่งฟาเซทที่สนใจจะถูกกำหนดให้เป็นพารามิเตอร์ที่มีอิทธิพลต่อการตอบสนองข้อสอบ ยกตัวอย่างเช่น การศึกษา 3 ฟาเซท ประกอบด้วย ผู้สอบ ข้อสอบ และผู้ตรวจ จะกำหนดพารามิเตอร์ ดังนี้ 1) ความสามารถของผู้สอบ 2) ความยากของข้อคำถาม และ 3) การกดคะแนนของผู้ตรวจ โดยทั่วไปแล้วการศึกษาฟาเซทจะถือว่าฟาเซทต่าง ๆไม่มีปฏิสัมพันธ์ต่อกัน อย่างไรก็ตาม ในความเป็นจริงจะพบว่าฟาเซทสามารถมีปฏิสัมพันธ์ต่อกันได้ เรียกว่า Differential Facet Functioning (DFF) เช่น ผู้ตรวจอาจมีรูปแบบการกดหรือปล่อยคะแนนต่างกันเมื่อตรวจข้อสอบหรือคำถามต่างข้อหรือต่างเนื้อหากัน (ปฏิสัมพันธ์ระหว่างผู้ตรวจ - ข้อสอบ) บางกรณี ผู้ตรวจอาจมีรูปแบบการให้คะแนนต่างกันสำหรับผู้สอบต่างกลุ่มกัน (ปฏิสัมพันธ์ระหว่างผู้ตรวจ – ผู้สอบ) หรือกรณีที่ข้อสอบบางข้อส่งผลต่อการตอบสนองข้อสอบของผู้สอบต่างกัน (ปฏิสัมพันธ์ระหว่างข้อสอบ – ผู้สอบ) ในกรณีของปฏิสัมพันธ์ระหว่างข้อสอบ – ผู้สอบ รู้จักกันในชื่อการทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) ในกรณีของปฏิสัมพันธ์ระหว่างผู้ตรวจ – ผู้สอบ จะเรียกว่า *การทำหน้าที่ต่างกันของผู้ประเมิน (Differential Rater Functioning)*

การทำหน้าที่ต่างกันของผู้ประเมิน (Differential Rater Functioning: DRF) เป็นการศึกษาหลักฐานของการประเมินที่แตกต่างกันระหว่างผู้ประเมินในแง่ของการกด/ปล่อยคะแนนเมื่อให้คะแนนผู้สอบที่มีคุณลักษณะต่างกัน เช่น ผู้ประเมินบางคนให้คะแนนผู้สอบเพศหญิงสูงกว่าเพศชาย หรือกดคะแนนนักเรียนที่มีความสามารถสูง แต่ปล่อยคะแนนให้กับนักเรียนที่มีความสามารถต่ำกว่า มโนทัศน์ของการทำหน้าที่ต่างกันของผู้ประเมิน (DRF) ใกล้เคียงกับมโนทัศน์ของการทำหน้าที่ต่างกันของข้อสอบ (DIF) กล่าวคือ DIF พิจารณาปฏิสัมพันธ์ระหว่างข้อสอบที่ส่งผลต่อตอบของผู้สอบต่างกลุ่ม ในขณะที่ DRF พิจารณาปฏิสัมพันธ์ระหว่างการให้คะแนนของผู้ตรวจที่มีต่อผู้สอบต่างกลุ่มหรือผู้ตรวจมีแนวโน้มจะให้กดคะแนนในการตรวจคำตอบของผู้ให้ข้อมูลกลุ่มหนึ่งและปล่อยคะแนนในการตรวจคำตอบของผู้ให้ข้อมูลอีกกลุ่มหนึ่ง การศึกษา DRF ส่วนใหญ่มีข้อตกลงเกี่ยวกับกลุ่มที่ทราบอยู่แล้ว (known group) เช่น เพศ หรือ เชื้อชาติที่เหมือนกันระหว่างผู้ตรวจกับผู้สอบจะทำให้เกิดความลำเอียงในการให้คะแนน อย่างไรก็ตามมีบางกรณีที่ผู้ตรวจมีระดับการกด/ปล่อยคะแนนที่

ต่างกันระหว่างผู้สอบแต่ละกลุ่มที่ไม่ได้อยู่ในกลุ่มที่ทราบแน่ชัด เช่น ผู้ตรวจอาจให้คะแนนผู้สอบที่ลายมือดีหรือเขียนตอบเป็นระเบียบสูงกว่าผู้สอบที่มีลายมืออ่านยาก (Jin และ Wang, 2017)

2.2 การตรวจสอบการทำหน้าที่ต่างกันของผู้ประเมิน

การวิเคราะห์ DRF ได้รับการพัฒนาขึ้นเพื่อปรับปรุงการตรวจสอบการประเมินและระบุความแตกต่างของการตอบสนองของผู้ตรวจที่มีต่อกลุ่มข้อสอบ ซึ่งเป็นประโยชน์ต่อการวิเคราะห์ความคลาดเคลื่อนในการประเมิน เช่น การกดหรือปล่อยคะแนน วิธีการตรวจสอบ DRF ที่พบมากที่สุด คือ การวิเคราะห์ Many Facet Rasch Model (MFRM) (Schaefer, 2008; Muckle และ Karabatsos, 2009; Farrokhi, Esfandiani และ Schaefer, 2012; Myford และ Wolfe, 2009; Xun Yan, 2014; Engelhard Jr., Wind, Kobin และ Chajewski, 2013; Wesolowski, Wind และ Engelhard Jr., 2015) นอกจากนี้ยังมีการใช้วิธีการ Multilevel Analysis (Leckie และ Baird, 2011) และ Mantel-Haenzel method (Johanson และ Osborn, 2004)

จากการศึกษาเอกสารที่เกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของผู้ประเมิน พบว่า Many Facet Rasch Model (MFRM) เป็นวิธีการที่ได้รับความนิยมในการตรวจสอบพฤติกรรมการให้คะแนนของผู้ประเมินมากที่สุด เนื่องจาก การวิเคราะห์ MFRM สามารถแยกพิจารณาฟาเซทต่างๆ ได้อย่างเป็นอิสระ ผลการวิเคราะห์ให้สารสนเทศเกี่ยวกับอิทธิพลของผู้ประเมินในระดับกลุ่มได้ เช่น ความแตกต่างของการกดหรือปล่อยคะแนนในการประเมินผู้สอบ หรือการตรวจสอบว่าผู้ประเมินแต่ละชุดสามารถทำการประเมินแทนกันได้หรือไม่ รวมถึงประสิทธิภาพของผู้ประเมินในการจำแนกความสามารถของผู้สอบและจำแนกความแตกต่างระหว่างคุณลักษณะที่ต้องประเมิน (Myford และ Wolfe, 2004) Many Facet Rasch Model (MFRM) การวิเคราะห์ MRFM รายงานค่าสถิติเกี่ยวกับความคงที่ของการประเมินด้วย infit และ outfit mean square residual โดย infit mean square residual เป็นเศษเหลือค่าเฉลี่ยกำลังสองถ่วงน้ำหนักที่อ่อนไหวต่อรูปแบบการประเมินที่ไม่คาดหวังและใช้เป็นตัวบ่งชี้ความคงที่ภายในของผู้ประเมิน ในขณะที่ outfit mean square residual เป็นค่ามาตรฐานของเศษเหลือค่าเฉลี่ยกำลังสองแบบไม่ถ่วงน้ำหนักซึ่งอ่อนไหวต่อการให้คะแนนที่ไม่คาดหวังแบบสุดโต่ง หรือ outliers โดยสถิติทั้งสองตัวควรมีค่าใกล้ 1 ค่า infit ที่น้อยกว่า 1 บ่งชี้ว่าผู้ประเมินให้คะแนนระดับเดียว หาก infit มากกว่า 1 แสดงถึงการให้คะแนนลักษณะสุ่ม เกณฑ์ที่มักใช้ในการพิจารณาอยู่ระหว่าง 0.5 ถึง 1.5 (Linacre, 2002; Xun Yan, 2014)

นอกจากการวิเคราะห์พฤติกรรมการประเมินและการทำหน้าที่ต่างกันของผู้ประเมินแล้ว MFRM ยังได้รับการพัฒนาโมเดลให้สามารถวิเคราะห์การตอบสนองต่อการตอบของผู้ประเมินในระดับกลุ่มต่าง ๆ รวมถึงสามารถสร้างรูปแบบการให้คะแนนแบบ 2 ค่า และหลายค่าได้พร้อมกัน เรียกว่า Many Facet Partial Credit Model (MFR-PC) ซึ่งเป็นการสรุปอ้างอิงราสช์โมเดลสำหรับการตรวจให้คะแนน 2 ค่า และสามารถใช้กับการประเมินบนมาตรประมาณค่าที่มีระดับคะแนน 2 ระดับขึ้นไป (Engelhard Jr. และคณะ, 2013) Wesolowski และคณะ (2015) ใช้ MFR-PC ศึกษาการทำหน้าที่ต่างกันของผู้ประเมินในการประเมินด้านดนตรี โดยศึกษาความไม่แปรเปลี่ยนของความเข้มงวดในการประเมินของผู้ประเมินในการประเมินโรงเรียนระดับต่าง ๆ การศึกษาของ Wesolowski และคณะ (2015) พบว่า MFR-PC สามารถตรวจสอบรูปแบบย่อยของการกด/ปล่อยคะแนนภายในกลุ่มย่อยของผู้ประเมินและรายบุคคลได้ โดยพบการทำหน้าที่ต่างกันของผู้ประเมินทั้งในระดับกลุ่มผู้ประเมินและระดับบุคคล

Wolfe และคณะ (1999) เสนอวิธีการตรวจสอบการทำหน้าที่ต่างกันของผู้ประเมินข้ามช่วงเวลา (Differential rater functioning over time: DRIFT) โดยใช้การวิเคราะห์ Multi-Faceted Rating Scale Model (MFRSM) สำหรับศึกษากรณีการประเมินที่ใช้เวลานานซึ่งทำให้เกิดความคลาดเคลื่อนระหว่างการประเมินได้ โดยแบ่ง DRIFT ในการศึกษาเป็น 3 ลักษณะ คือ 1) Primacy/Recency คือ การเพิ่มขึ้นหรือลดลงของการให้คะแนนในภาพรวมของผู้ประเมินเมื่อเวลาผ่านไป ในกรณีของ Primacy ผู้ประเมินจะให้คะแนนสูงในการประเมินตอนต้นและจะกดคะแนนหรือให้คะแนนยากขึ้นเมื่อเวลาผ่านไป ในขณะที่ Recency จะเป็นไปในทิศทางตรงกันข้าม2) Practical/Fatigue คือ การที่ผู้ประเมินเพิ่มหรือลดความแม่นยำของการประเมินเมื่อเวลาผ่านไป Practical คือ การที่ผู้ประเมินมีความสอดคล้องในการประเมินมากขึ้นเมื่อเวลาผ่านไป ส่วน Fatigue คือ ผู้ประเมินมีความแม่นยำลดลงเมื่อเวลาผ่านไป 3) Differential Centrality/Differential Extremism คือ แนวโน้มการให้คะแนนในช่วงกลางหรือช่วงสุดโต่งของสเกลเมื่อเวลาผ่านไป Differential Centrality เป็นการที่ผู้ประเมินมีแนวโน้มที่จะให้คะแนนในช่วงกลางของสเกลมากขึ้นเมื่อประเมินติดต่อกันเป็นเวลานาน Differential Extremism เป็นแนวโน้มที่ผู้ประเมินจะให้คะแนนแบบสุดโต่งทางซ้ายหรือขวาของสเกลมากขึ้นเมื่อเวลาผ่านไป ทั้งนี้ Multi-Faceted Rating Scale Model (MFRSM) ถูกนำมาใช้สำหรับการวิเคราะห์การทำหน้าที่ต่างกันของผู้ประเมินข้ามช่วงเวลาดังกล่าว

Wolfe และคณะ (1999) ศึกษาโดยการจำลองสถานการณ์ใน 3 เงื่อนไข คือ 1) ประเภทของ DRIFT 2) ขนาดของ DRIFT และ 3) ช่วงเวลา เพื่อดูการเปลี่ยนแปลงของการให้คะแนนเมื่อเวลาผ่านไป เปรียบเทียบความแตกต่างระหว่างช่วงเวลาโดยทำให้อยู่ในรูปของคะแนนมาตรฐาน ความแตกต่างในรูปคะแนนมาตรฐานมีค่าเฉลี่ยเท่ากับ 0 และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 1.00 คะแนนมาตรฐานมีความแตกต่างกันสูง (มากกว่า 2.00) แสดงว่าความเข้มงวดของผู้ประเมินเพิ่มขึ้นเมื่อเวลาผ่านไป หากมีค่าต่ำกว่า 2.00 แสดงว่าความเข้มงวดของผู้ประเมินลดลงเมื่อเวลาผ่านไป จากนั้น ดูความสอดคล้องระหว่างคะแนนจากผู้ประเมินและการประมาณค่าของโมเดลโดย fit statistics ได้แก่ infit หรือ เศษเหลือของค่าเฉลี่ยกำลังสองแบบถ่วงน้ำหนักโดยความแปรปรวน และ outfit หรือ เศษเหลือของค่าเฉลี่ยกำลังสองแบบไม่ถ่วงน้ำหนัก ค่า fit statistics ทั้งสองอยู่ในรูปของไค-สแควร์ มีค่าตั้งแต่ 0 ถึง ∞ ค่า infit มากกว่า 1 แสดงว่ามีความแตกต่างระหว่างค่าคาดหมายและค่าสังเกตได้ของการประเมินบริเวณช่วงกลางของการแจกแจงของคะแนน ค่า outfit มากกว่า 1.00 แสดงว่ามีเศษเหลือที่ไม่คาดหมายบริเวณหางของการแจกแจงของคะแนน ค่า fit statistics น้อยกว่า 1.00 แสดงว่ามีความแปรผันต่ำกว่าที่คาดหมายจากการประมาณค่าของ MFRSM โดยทั่วไปแล้วค่าสถิติที่แสดงถึงความสอดคล้องจะอยู่ระหว่าง 0.8 ถึง 1.4 ทั้งนี้ จุดตัดของระดับความสอดคล้องจะแตกต่างกันไปตามลักษณะของการประเมิน (Wolfe และคณะ, 1999) การทดสอบความแตกต่างระหว่าง fit statistics ทั้งสอง คือ การหาสัดส่วนระหว่างค่าเฉลี่ยกำลังสองของค่าสถิติความสอดคล้องจาก 2 ช่วงเวลาใช้เป็นตัวบ่งชี้ความแตกต่างของความแม่นยำ และสัดส่วนระหว่างความแปรปรวนของการประเมินของผู้ประเมินจาก 2 ช่วงเวลา ใช้เป็นตัวบ่งชี้ของการใช้ระดับสเกลการประเมินที่ต่างกัน ความแตกต่างระหว่างคะแนนที่ได้อยู่ในรูปการแจกแจง F การทดสอบสมมติฐาน หาก F-ratio มีค่าต่ำกว่า 0.05 แสดงถึงการเกิด DRIFT ในการประเมิน

ในปี 2009 Myford และ Wolfe ได้ศึกษาการตรวจสอบ DRIFT โดยใช้โมเดลที่พัฒนาขึ้นชื่อ Time Facet model ที่ทำให้เห็นภาพการกด/ปล่อยคะแนนของผู้ประเมินเป็นลักษณะที่คงที่ข้ามช่วงเวลาแต่ปรับแก้การเคลื่อนของค่าเฉลี่ยการประเมินข้ามช่วงเวลา Time Facet Model มีประโยชน์ในการระบุการเปลี่ยนแปลงของค่าเฉลี่ยข้ามกลุ่มผู้ประเมินเมื่อเวลาผ่านไปซึ่งให้ภาพของการเปลี่ยนแปลงโดยเฉลี่ยในภาพรวมระหว่างช่วงเวลาต่าง ๆ อันเนื่องมาจากความแตกต่างของความเข้มงวดในการให้คะแนนต่างช่วงเวลาหรือความยากของการให้คะแนนในแต่ละช่วงเวลา Time Facet Model เหมาะกับการประเมินที่ประเมินการทดสอบคุณลักษณะเดียว (Single performance task)

การศึกษาของ Myford และ Wolfe (2009) ได้เสนอดัชนีสำหรับระบุความแตกต่างของความแม่นยำและระดับของสเกล ได้แก่ SR-ROR ($\mathrm{r_{SR-ROR}}$) เป็น สหสัมพันธ์แบบเพียร์สันแสดงความแม่นยำในรูปของลักษณะคงที่ของประสิทธิภาพผู้ประเมินภายใต้กรอบการอ้างอิงที่สร้างขึ้นคำนวณจากการประเมินของผู้ประเมินรายบุคคลและการประเมินจากผู้ประเมินทั้งหมด ค่าสัมประสิทธิ์ $\mathrm{r_{SR-ROR}}$ ให้สารสนเทศโดยย่อของระดับความคงที่ระหว่างผู้ประเมินในการจัดลำดับความสามารถของผู้สอบ หากผู้ประเมินจัดลำดับความสามารถของผู้สอบเป็นลักษณะสุ่มเมื่อเทียบกับผู้ประเมินคนอื่นค่า $\mathrm{r_{SR-ROR}}$ จะมีค่าใกล้ 0 และหากการจัดลำดับของผู้ประเมินสอดคล้องกับผู้ประเมินคนอื่น ค่า r_{SR-ROR} จะมีค่าใกล้ 1 ดัชนีอีกตัวหนึ่งคือ $r_{res,exp}$ เป็นดัชนีที่มีประโยชน์สำหรับการตรวจการให้คะแนนในช่วงกลางของสเกล (Central tendency) $r_{res,exp}$ เป็นสหสัมพันธ์ระหว่างเศษเหลือและค่าคาดหมายของการประเมินในโมเดล

นอกจากการพัฒนาวิธีการตรวจสอบการทำหน้าที่ต่างกันของผู้ประเมินที่กล่าวมาแล้ว ยังมีการศึกษาที่นำวิธีการดังกล่าวไปใช้ในการตรวจสอบการทำหน้าที่ต่างกันของผู้ประเมิน อาทิ การศึกษาของ Schaefer (2008) ที่ศึกษารูปแบบความลำเอียงของผู้ประเมินโดยใช้การวิเคราะห์ MFRM โดยวิเคราะห์การประเมินของผู้ประเมินที่เป็นเจ้าของภาษาจำนวน 40 คน ที่ประเมินงานเขียน 40 ฉบับของนักเรียนชาวญี่ปุ่น พบว่า MFRM มีประสิทธิภาพในการวิเคราะห์แหล่งความผันแปรของคะแนนการเขียนที่มาจากความลำเอียงของผู้ประเมิน ซึ่งสอดคล้องกับการศึกษาของ McNamara (1996) และข้อสรุปของ Linacre (1998) ว่า MFRM ให้ผลการวิเคราะห์ความตรงของกระบวนการประเมินที่ตรงไปตรงมา โดยการลดอิทธิพลความผันแปรของผู้ประเมินที่มีต่อระดับของผู้สอบและประมาณค่าความสามารถของผู้สอบที่เป็นอิสระจากการกดและปล่อยคะแนน Farrokhi และคณะ (2012) ศึกษาการทำหน้าที่ต่างกันของผู้ตรวจโดย MFRM พบว่า การตรวจให้คะแนนของครู เพื่อน และตนเอง แตกต่างกันในแง่ของการกด/ปล่อยคะแนน ทั้งนี้ การตรวจให้คะแนนโดยเพื่อนและตนเองมีรูปแบบการประเมินแบบกด/ปล่อยคะแนนที่ต่างกันในข้อที่ต่างจากการให้คะแนนของครู และพบว่าประเภทหรือกลุ่มของผู้ประเมินจะกดคะแนนกับนักเรียนที่มีความสามารถสูงและปล่อยคะแนนกับนักเรียนที่มีความสามารถต่ำ นอกจากนี้ยังพบว่า การประเมินตนเองของนักเรียนจะกดคะแนนต่อข้อคำถามในขณะที่การประเมินโดยเพื่อนจะมีความลำเอียงในการกดคะแนนต่อผู้ให้ข้อมูล นอกจากนี้ Farrokhi และคณะ (2012) ได้ให้ข้อเสนอแนะของการศึกษาว่า ควรใช้กลุ่มตัวอย่างที่มีขนาดใหญ่ในการวิเคราะห์ ทั้งนี้ กลุ่มตัวอย่างในการศึกษาครั้งนี้มีผู้ประเมิน 194 คน ประเมินงานเขียน 188 ฉบับ การศึกษาของ Xun Yan (2014) เพื่อตรวจสอบประสิทธิภาพของผู้ประเมินใน

การทดสอบความสามารถทางภาษาอังกฤษ พบการทำหน้าที่ต่างกันของผู้ประเมินในการกดคะแนนเนื่องมาจากมโนทัศน์ที่ต่างกันต่อความสามารถทางการพูดของผู้สอบชาวอินเดียและผู้สอบชาวจีนที่มีความสามารถต่ำ

เนื่องจาก MFRM มีปัญหาเกี่ยวกับการประมาณค่าความคลาดเคลื่อนมาตรฐาน ดังเช่น Patz และคณะ (2002) กล่าวว่า การบิดเบือนของการประมาณค่าความคลาดเคลื่อนมาตรฐานในการประมาณค่าความสามารถของผู้สอบอันเนื่องมาจากการที่ MFRM มองข้ามความไม่เป็นอิสระระหว่างการประเมินคำตอบในข้อเดียวกันของผู้สอบ ดังนั้น Muckle และ Karabatsos (2009) จึงได้เสนอให้ขยายโมเดล MFRM โดยเพิ่มส่วนของโมเดล Hierarchical Generalized Linear Model (HGLM) ที่มี random intercept แสดงค่าความสามารถของผู้สอบในการทดสอบ และอิทธิพลคงที่สำหรับข้อสอบ ผู้ประเมิน และฟาเซทอื่นๆ HGLM สามารถสร้างรูปแบบอิทธิพลแบบสุ่มสำหรับข้อสอบและผู้ประเมินเพื่อที่จะประเมินความคงที่ข้ามผู้สอบซึ่งจะแก้ปัญหาการประมาณค่าของ MFRM ดังกล่าวได้ อย่างไรก็ตาม ผลการศึกษาของ Muckle และ Karabatsos (2009) ยังพบข้อจำกัดเกี่ยวกับข้อตกลงเบื้องต้นของการแจกแจงของโมเดลที่ต้องเป็นการแจกแจงปกติ ในกรณีที่การแจกแจงของประชากรไม่เป็นการแจกแจงปกติจะทำให้โมเดลมีความสอดคล้องกับข้อมูลต่ำ และการประมาณค่าประชากรของอิทธิพลสุ่มจะไม่แม่นยำ ซึ่งส่งผลเสียต่อการวิเคราะห์ข้อมูลเกี่ยวกับความเที่ยงของการประเมิน และการประมาณค่าด้วย Marginal Maximum Likelihood (MMLE) ให้การประมาณค่าแบบจุด ซึ่งมีข้อจำกัดเกี่ยวกับความน่าเชื่อถือของการประมาณค่าพารามิเตอร์รวมถึงการแปลความหมายของค่า p-value ซึ่งเป็นสถิติที่ขึ้นอยู่กับตัวอย่างนั้นมีปัญหาความคลาดเคลื่อนแบบที่ 1 เฟ้อ เมื่อทำการทดสอบซ้ำหลายครั้ง และการคำนวณ p-value แยกกันแต่ละพารามิเตอร์นั้นถือว่าพารามิเตอร์มีความแปรปรวนร่วมเป็น 0 ซึ่งเป็นข้อตกลงที่ไม่เหมาะสมอันนำไปสู่ข้อสรุปที่ผิดพลาด ทั้งนี้ Muckle และ Karabatsos (2009) ได้เสนอให้ใช้วิธีการประมาณค่าแบบเบส์ซึ่งสามารถกำหนดการแจกแจงความน่าจะเป็นก่อนหน้าได้ทุกรูปแบบ

ตอนที่ 3 ทฤษฎีฉันทามติเชิงวัฒนธรรม

ทฤษฎีฉันทามติเชิงวัฒนธรรม (Cultural Consensus Theory) มีจุดเริ่มต้นมาจากการวิจัยด้านมานุษยวิทยาและชาติพันธุ์วรรณนา โดยมีวัตถุประสงค์เพื่อตรวจสอบความสอดคล้องของการให้คำตอบของผู้ให้ข้อมูลโดยการนำโมเดลทางสถิติมาช่วยในการวิเคราะห์ข้อมูล ทฤษฎีฉันทามติเชิงวัฒนธรรมได้รับการพัฒนาในช่วงปี 1986 ในฐานะเครื่องมือของระเบียบวิธีที่สนับสนุนการศึกษาเชิง

ชาติพันธุ์วรรณนา โดยมีจุดมุ่งหมายเพื่อระบุองค์ความรู้เชิงวัฒนธรรมมีสมาชิกของชุมชนหรือสังคมมีร่วมกัน และได้รับการพัฒนาอย่างต่อเนื่องและนำไปประยุกต์ใช้ในศาสตร์หลายสาขา

จุดเริ่มต้นและความหมายของทฤษฎีฉันทามติเชิงวัฒนธรรม

ทฤษฎีฉันทามติเชิงวัฒนธรรม (Cultural Consensus Theory) เป็นทั้งทฤษฎี (Theory) และวิธีการ (Method) ในการวิเคราะห์ความสอดคล้องหรือฉันทามติ (Consensus) ทางความคิดหรือความเห็นของคนส่วนใหญ่ ในบทความของ Borgatti และ Halgin (2011) ได้แยกคำจำกัดความของ Cultural Consensus Theory ไว้ 2 ทาง ดังนี้

ทฤษฎีฉันทามติเชิงวัฒนธรรมในทางทฤษฎี หมายถึง สถานการณ์ที่เกิดขึ้นภายใต้ความเห็นพ้องต้องกันภายในกลุ่มคนนั้นถือว่าเป็นสัญญาณที่แสดงถึง “องค์ความรู้” หรือ “ความถูกต้อง” ภายใต้บริบทของวัฒนธรรมหนึ่ง ทั้งนี้ คติหรือความเชื่อในระบบของญาณวิทยา (Epistemological system) ต่างขึ้นอยู่กับการเชื่อมโยงระหว่างความเห็นพ้องต้องกันและความจริง (Truth) ตัวอย่างที่เห็นได้ชัดเจนคือการตัดสินของศาลที่จะไม่ตัดสินให้โจทก์เป็นฝ่ายถูกจนกว่าคณะลูกขุนจะเห็นชอบร่วมกัน เช่นเดียวกับการทดลองทางวิทยาศาสตร์ที่ใช้ค่าเฉลี่ยของผลการทดลองหลายๆ ครั้งในการประมาณค่าความถูกต้องของผลการศึกษานั้น อย่างไรก็ตาม มีข้อถกเถียงว่าการตัดสินบนพื้นฐานของฉันทมตินั้นสามารถนำไปสู่ข้อสรุปที่ผิดได้ ซึ่ง Borgatti และ Halgin (2011) ชี้ให้เห็นว่าความเห็นพ้องต้องกันนั้นไม่สามารถอนุมานว่าคำตอบนั้นถูกต้องเสมอไป แต่การวิเคราะห์ฉันทามติ (Consensus Analysis) จะช่วยแก้ปัญหาในบางสถานการณ์ที่ความเห็นพ้องต้องกันสามารถอนุมานถึงองค์ความรู้หรือคำตอบได้อย่างแท้จริง

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ทฤษฎีฉันทามติเชิงวัฒนธรรมในทางปฏิบัติ หรือในแง่ของวิธีการนั้น Borgatti และ Halgin ได้อธิบายว่าหมายถึง วิธีการในการสร้างกรอบความคิด และรวมความหลากหลายของบุคคลให้เป็นอันหนึ่งอันเดียวกัน การวิเคราะห์ฉันทามติ (Consensus Analysis) ให้ข้อสรุป 3 ประการ คือ 1) แนวทางในการระบุว่าความคิดหรือทัศนคติที่หลากหลายนั้นอยู่ภายใต้วัฒนธรรมเดียวกันหรือไม่ ยกตัวอย่างเช่น เมื่อถามคำถามคนไทยเกี่ยวกับวัฒนธรรมการกราบในสังคมไทย อาจมีผู้ให้คำตอบหรือแสดงความคิดเห็นที่ต่างกันออกไป แต่เมื่อวิเคราะห์ฉันทามติจะพบว่าคำตอบที่แตกต่างนั้นจะยังคงแสดงถึงความเชื่อหรือองค์ความรู้เกี่ยวกับวัฒนธรรมการกราบภายใต้บริบทของสังคมไทยที่เป็นจุดร่วมกัน ซึ่งแสดงว่าผู้ให้ข้อมูลอยู่ภายใต้วัฒนธรรมเดียวกัน 2) การวิเคราะห์ฉันทามติสามารถวัดระดับความรู้ความสามารถทางวัฒนธรรม หรือ cultural competence ของบุคคลภายใน

วัฒนธรรมเดียวกันได้ และ 3) วิธีการนี้จะค้นหาคำตอบที่เป็นคำตอบที่ถูกต้อง (culturally correct answer) ให้กับข้อคำถามที่ตั้งขึ้น

ทฤษฎีฉันทามติเชิงวัฒนธรรมได้รับการพัฒนาโดย A. Kimball Romney, Susan C. Weller และ William H. Batchelder และได้รับการนำเสนอในรูปแบบบทความตีพิมพ์ครั้งแรกในวารสาร American Anthropologist ในปี 1986 ในบทความดังกล่าวได้เสนอวิธีการสร้างเกณฑ์ที่เป็นปรนัยสำหรับการวัดความถูกต้องของการสรุปคำตอบให้กับคำถามเชิงวัฒนธรรมในการวิเคราะห์ข้อมูลทางชาติพันธุ์วรรณนาและมานุษยวิทยาวัฒนธรรม แนวคิดสำคัญของทฤษฎีนี้ คือ การใช้รูปแบบของความสอดคล้องทางความคิด หรือฉันทามติระหว่างผู้ให้ข้อมูลเพื่อนำมาใช้อนุมานเกี่ยวกับความสามารถที่แตกต่างกันในองค์ความรู้เดียวกัน เมื่อผู้ให้ข้อมูลแต่ละกลุ่มมีการตอบสนองที่มีลักษณะร่วมกันทางวัฒนธรรม (common culture) ต่อชุดคำถามที่เกี่ยวกับองค์ความรู้ที่มีร่วมกันจะถือว่าลักษณะร่วมนี้เป็นฉันทามติ (consensus/culturally correct) ลักษณะเด่นของทฤษฎีการวิเคราะห์ฉันทามติเชิงวัฒนธรรม คือ การไม่ได้มุ่งวัดฉันทามติของกลุ่มโดยเปรียบเทียบกับองค์ความรู้ที่เป็นแนวคิดทฤษฎีที่ถูกต้องหรือมีอยู่ แต่มีจุดมุ่งหมายเพื่อค้นหาว่าผู้ให้ข้อมูลซึ่งเป็นสมาชิกของกลุ่มหรือสังคมนั้นมีความคิดเห็นที่เป็นฉันทามติภายในกลุ่มอย่างไร โดยไม่สนใจว่าฉันทามติดังกล่าว “ถูกต้อง” หรือ “เป็นจริง” ตามนิยามทางทฤษฎี ซึ่งแตกต่างจากทฤษฎีการวัดทางจิตวิทยาอื่น ๆ ที่มุ่งศึกษาระดับความรู้หรือความสามารถของบุคคลโดยอิงกับนิยามของสิ่งที่มุ่งวัด ทฤษฎีการวิเคราะห์ฉันทามติเชิงวัฒนธรรมถือว่าคำตอบที่ถูกต้องในเชิงฉันทามติของกลุ่มเป็นตัวแปรแฝงที่ต้องใช้การประมาณค่าทางสถิติ และมีการประมาณค่าพารามิเตอร์ระดับความรู้หรือความสามารถของผู้ให้ข้อมูลหรือผู้ให้ข้อมูลโดยใช้ข้อมูลจากการตอบคำถามหรือแบบสอบถาม

แนวคิดดังกล่าวของทฤษฎีการวิเคราะห์ฉันทามติเชิงวัฒนธรรมสอดคล้องกับธรรมชาติของการประเมินโดยใช้ความคิดเห็นของผู้เชี่ยวชาญในแง่ที่ว่า ผลการประเมินนั้นเป็นตัวแปรแฝงของผู้เชี่ยวชาญ ซึ่งมีความเบี่ยงเบนของการให้คะแนนขึ้นอยู่กับปัจจัยต่าง ๆ ที่เกี่ยวข้องกับการประเมินหรือตัวผู้ประเมินเอง ซึ่งประมาณค่าได้จากค่าสังเกตได้ของผลการให้คะแนนในการประเมิน

ข้อตกลงเบื้องต้นของทฤษฎีฉันทามติเชิงวัฒนธรรม

เนื่องจากการวิเคราะห์ฉันทามติโดยทฤษฎีฉันทามติเชิงวัฒนธรรมเป็นการศึกษาความสอดคล้องขององค์ความรู้ในผู้ให้ข้อมูลที่อยู่ร่วมวัฒนธรรมเดียวกัน จึงมีข้อตกลงเบื้องต้นเกี่ยวกับการวิเคราะห์อยู่ 3 ประการ ดังนี้ (Romney, Weller, and Batchelder 1986 และ Borgatti and Halgin (2011)

1. Common Truth มีคำตอบที่ถูกต้องเพียงคำตอบเดียวสำหรับคำถามแต่ละข้อ เช่นเดียวกับการเลือกคำตอบแบบหลายตัวเลือก (multiple-choice) จะมีคำตอบที่ถูกต้องเพียงคำตอบเดียวในขณะที่ตัวเลือกอื่นเป็นตัวลวงหรือคำตอบผิด ซึ่งเป็นแสดงว่าผู้ให้ข้อมูลอยู่ภายใต้วัฒนธรรมเดียวกัน

2. Local Independence หรือ Conditional Independence คำตอบของผู้ให้ข้อมูลแต่ละคนเป็นอิสระจากกัน นั่นคือ ผู้ให้ข้อมูลจะไม่ลอกเลียนคำตอบจากผู้ให้ข้อมูลคนอื่น และการตอบคำถามข้อใดข้อหนึ่งจะไม่มีผลต่อการตอบคำถามในข้อถัดไป แสดงให้เห็นว่า เมื่อผู้ให้ข้อมูลไม่ทราบคำตอบ เขาจะเลือกคำตอบจากตัวเลือกที่กำหนดเท่านั้น ดังนั้น การที่ผู้ให้ข้อมูล 2 คน หรือทั้งกลุ่มเลือกคำตอบเดียวกันจะมีเหตุผลมาจากผู้ให้ข้อมูลทราบคำตอบหรือเดาคำตอบถูกด้วยตนเองโดยไม่มีปัจจัยอื่นที่มีผลต่อการเลือกคำตอบนั้น Romney, Weller, and Batchelder (1986) อธิบายข้อตกลงนี้ให้เห็นในสมการ

$$\Pr[(X_{ik})_{NxM}|(Z_k)_{IxM}] = \prod_{i=1}^{N}\prod_{k=1}^{M}\Pr(X_{ik}\,|Z_k) \tag{2.8}$$

ความสัมพันธ์ระหว่างรูปแบบคำตอบของผู้ให้ข้อมูลเป็นผลลัพธ์ที่เกิดจากระดับความสอดคล้องกับคำตอบที่แท้จริง คือ Z เมื่อใดก็ตามที่ข้อมูลสอดคล้องกับโมเดล ความสัมพันธ์ระหว่างผู้ให้ข้อมูลจะสูงหากคำนวณจาก Response Profile Data หรือคำตอบที่ผู้ให้ข้อมูลคนที่ i ตอบคำถามข้อที่ k แต่จะเข้าใกล้ 0 เมื่อคำนวณจาก Performance Profile Data หรือการตอบข้อที่ k ถูก หรือ ผิด

3. Item Homogeneity ความเป็นเอกพันธ์ของข้อคำถาม หมายถึงคำถามแต่ละข้อเป็นคำถามในรูปแบบของการสุ่มจากเอกภพของคำถามในหัวข้อเดียวกันทั้งหมด ดังนั้น ความน่าจะเป็นที่ผู้ให้ข้อมูล i จะทราบคำตอบของคำถามข้อใด ๆ มีค่าเท่ากัน หรืออีกนัยหนึ่งคือ ข้อคำถามทุกข้อมีความยากใกล้เคียงกัน

ข้อตกลงเบื้องต้นทั้ง 3 ประการนี้ เป็นข้อตกลงพื้นฐานของการวิเคราะห์ด้วยโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมเพื่อที่จะสกัดคำตอบที่เป็นองค์ความรู้ร่วมกันภายใต้บริบทของวัฒนธรรมเดียวกัน ทั้งนี้ เมื่อนำไปใช้กับข้อมูลเชิงประจักษ์แล้วพบว่าไม่สอดคล้องกัน สามารถอนุมานได้ใน 3 ประเด็น คือ 1) ข้อมูลนั้นไม่ได้เกิดขึ้นภายใต้วัฒนธรรมเดียวกัน 2) ผู้ให้ข้อมูลหรือผู้ให้ข้อมูลไม่ได้มีความรู้หรือประสบการณ์ร่วมกันภายใต้วัฒนธรรมเดียวกัน หรือ 3) มีความผิดพลาดหรือปัจจัยอื่นที่ส่งผลต่อผลการวิเคราะห์ข้อมูล

โมเดลทางสถิติและวิธีการประมาณค่าของการวิเคราะห์ฉันทามติเชิงวัฒนธรรม

แนวคิดหลักของทฤษฎีฉันทามติเชิงวัฒนธรรมคือการใช้รูปแบบของความเห็นพ้องต้องกันหรือฉันทามติภายในกลุ่มผู้ให้ข้อมูลอันนำไปสู่การสรุปอ้างอิงเกี่ยวกับความรู้ความสามารถที่แตกต่างกันของผู้ให้ข้อมูลที่มีเกี่ยวกับองค์ความรู้ร่วม (shared knowledge) การวิเคราะห์ฉันทามติเชิงวัฒนธรรมถือว่าการรวบรวมข้อมูลที่มีลักษณะเป็นองค์ความรู้ร่วมกันทางสังคมจะมีค่าความสอดคล้องระดับสูง ซึ่งผู้ให้ข้อมูลจะมีระดับความสอดคล้องในการตอบระหว่างกันที่แตกต่างกันไปตามระดับความรู้เกี่ยวกับประเด็นนั้น ๆ เมื่อเราทราบระดับความรู้ความสามารถของผู้ให้ข้อมูล เราก็จะสามารถหาคำตอบที่ถูกต้องเชิงวัฒนธรรมได้โดยการถ่วงน้ำหนักคำตอบของผู้ให้ข้อมูลแต่ละคนและหาความน่าจะเป็นของคำตอบที่ถูกต้องและเป็นตัวแทนของคำตอบเชิงวัฒนธรรม

การวิเคราะห์ฉันทามติเชิงวัฒนธรรมเป็นการวิเคราะห์ข้อมูลโดยโมเดลสถิติแบบพาราเมตริกในการประมาณค่าพารามิเตอร์จากข้อมูลสังเกตได้ไปยังผลลัพธ์ที่อยู่ใน sample space ทั้งหมด โดยแทนปริภูมิตัวอย่างด้วย T โดยที่ $\mathrm{x} \in \mathrm{T}$ คือความเป็นไปได้ของผลลัพธ์ทั้งหมดซึ่งอยู่ในรูปของเมตริกซ์ เรียกว่า response profile matrix มีค่า $\mathrm{X} = (\mathrm{X_{ik}})_{\mathrm{NxM}}$ เมื่อ $\mathrm{X_{ik}}$ เป็นตัวแปรสุ่มแทนการตอบของผู้ให้ข้อมูลคนที่ i ในคำถามข้อ k

แนวคิดของโมเดลฉันทามติเชิงวัฒนธรรมเป็นการเปรียบเทียบความสอดคล้องหรือใกล้เคียงกันระหว่างผลลัพธ์ $\mathrm{x} \in \mathrm{T}$ กับผลลัพธ์ที่ได้จากการจำลองข้อมูลของโมเดล ดังนั้น การอนุมานทางสถิติของโมเดลฉันทามติเชิงวัฒนธรรมจึงพิจารณาจากความสอดคล้องระหว่างโมเดลกับข้อมูลซึ่งสามารถทำได้ทั้งวิธีการทางสถิติแบบดั้งเดิมและแบบเบส์ ฟังก์ชันภาวะน่าจะเป็น (Likelihood function) ของโมเดลจะมีลักษณะเดียวกันกับฟังก์ชันความน่าจะเป็นของโมเดล ดังนี้

$$L(X|\Phi) = Pr(X|\Phi) \tag{2.9}$$

ในกรณีของฟังก์ชันภาวะน่าจะเป็น X เป็นค่าตายตัว และ Φ จะแปรตามขอบเขตใน Ω_Φ การประมาณค่าด้วยวิธีการทางสถิติแบบดั้งเดิม เมื่อสร้างข้อมูลผลลัพธ์ X จากปริภูมิพารามิเตอร์แล้ว จะทำการประมาณค่าพารามิเตอร์แบบ point estimate โดยใช้วิธีการภาวะน่าจะเป็นสูงสุด (Maximum Likelihood: ML) อย่างไรก็ตาม การใช้วิธีภาวะน่าจะเป็นสูงสุดอาจไม่เหมาะสมในกรณีของโมเดลฉันทามติเชิงวัฒนธรรมเนื่องจากข้อกำหนดของสถิติดังกล่าวที่ข้อมูลต้องเป็นตัวแปรสุ่มที่มีการแจกแจงที่เหมือนกัน (identically distributed) และเป็นอิสระต่อกัน (independent) มีขนาดใหญ่รวมถึงสามารถเพิ่มจำนวนได้ ในขณะที่ข้อมูลของโมเดลฉันทามติเชิงวัฒนธรรมเป็นเมตริกซ์การให้คะแนนของผู้ประเมิน i ในข้อสอบข้อที่ k โดย $X = (X_{ik})_{NxM}$ ซึ่งประกอบด้วยผู้ประเมินจำนวน N คน ในการประเมินข้อสอบ M ฉบับ ทั้งนี้ ตัวแปรสุ่มในเมตริกซ์ X นั้นไม่มีลักษณะของความเป็นอิสระหรือการแจกแจงที่เหมือนกัน แต่มีลักษณะของเงื่อนไขความเป็นอิสระเมื่อกำหนดพารามิเตอร์ นอกจากนี้ การเพิ่มจำนวนตัวแปรในเมตริกซ์ X หมายถึงการเพิ่มจำนวนผู้ประเมินและจำนวนข้อสอบ ซึ่งตัวแปรที่เพิ่มขึ้นมีพารามิเตอร์องค์ประกอบที่ต้องเพิ่มขึ้นตามไปด้วย ด้วยเหตุนี้ วิธีการทางสถิติแบบเบส์จึงเหมาะสมมากกว่าในการใช้อ้างอิงทางสถิติสำหรับโมเดลฉันทามติเชิงวัฒนธรรม

การอนุมานทางสถิติแบบเบส์อยู่บนพื้นฐานการประยุกต์ทฤษฎีบทของเบส์ (Bayes theorem) ซึ่งถือว่าพารามิเตอร์เป็นตัวแปรสุ่มที่มีการแจกแจงอยู่บนปริภูมิพารามิเตอร์ Ω_Φ โดยขึ้นอยู่กับความเชื่อพื้นฐานของนักวิจัยว่าค่าพารามิเตอร์ใดจะเป็นตัวแทนของค่าสังเกต (Kruschke, 2011; Batchelder, Anders และ Oravecz, 2018) สมมติให้ A และ B เป็นเหตุการณ์ที่อยู่ในปริภูมิตัวอย่าง จะเขียนความสัมพันธ์ของเหตุการณ์ดังกล่าวในรูปความน่าจะเป็นได้ว่า

$$Pr(A|\mathrm{B}) = \frac{Pr(B|\mathrm{A})\mathrm{Pr}\,(A)}{\mathrm{Pr}\,(B)} \tag{2.10}$$

ตามแนวคิดของการอนุมานทางสถิติแบบเบส์ นักวิจัยสามารถใช้สารสนเทศเพิ่มเติมจากแหล่งอื่นเพื่อเปลี่ยนแปลงความเชื่อเกี่ยวกับการแจกแจงพารามิเตอร์ที่สนใจได้ ซึ่งเขียนในรูปการอนุมานค่าพารามิเตอร์ ดังนี้

$$\Pr(\Phi|X) = \frac{Pr(X|\Phi)Pr(\Phi)}{Pr(X)} \tag{2.11}$$

เมื่อ $Pr\ (\Phi|X)$ คือ การแจกแจงความน่าจะเป็นก่อนหน้าที่ต้องการเมื่อกำหนดค่าสังเกต X $Pr(X|\Phi)$ คือ ฟังก์ชันความน่าจะเป็นที่ได้จากฟังก์ชันภาวะน่าจะเป็น $L(X|\Phi)$ $Pr(\Phi)$ คือ การแจกแจงความน่าจะเป็นก่อนหน้าของพารามิเตอร์ ซึ่งแสดงถึงความเชื่อหรือสมมติฐานเบื้องต้นของนักวิจัย $Pr(X)$ คือ ความน่าจะเป็นของค่าสังเกต X ได้จากการคำนวณค่าเฉลี่ยของผลคูณภาวะความน่าจะเป็นกับความน่าจะเป็นก่อนหน้าทั้งหมดในปริภูมิตัวอย่าง $Pr(X)$ ในสมการ 2.12 เป็นค่าคงที่บวกเนื่องจากพารามิเตอร์ถูกอินทิเกรทออก ทำให้การแจกแจงความน่าจะเป็นภายหลังเป็นสัดส่วนเดียวกับผลคูณของฟังก์ชันภาวะน่าจะเป็นกับความน่าจะเป็นก่อนหน้า

$$\Pr(X) = \int_{\Omega_\Phi} L(X||\Phi)P(\Phi)d\Phi \tag{2.12}$$

จากสมการ 2.12 สามารถเขียนเป็นรูปแบบเฉพาะของโมเดลฉันทามติเชิงวัฒนธรรมได้ ดังนี้

$$\Pr\,[Z = (Z_k)|X = (x_{ik})] = \frac{\Pr\,[(X = (x_{ik})|Z = (Z_k)]\Pr\,[Z = (Z_k)]}{\Pr\,[X = (x_{ik})]} \tag{2.13}$$

การวิเคราะห์โมเดลฉันทามติเชิงวัฒนธรรม นักวิจัยสามารถกำหนดการแจกแจงได้หลายรูปแบบ (Kruschke, 2011; Gelman และคณะ 2013, Batchelder, Anders และ Oravecz, 2018) โดยทั่วไปแล้วการกำหนดความน่าจะเป็นก่อนหน้าในการศึกษาโมเดลฉันทามติเชิงวัฒนธรรมจะกำหนดแบบ uninformative เนื่องจากวัตถุประสงค์ของการศึกษาโมเดลนี้การหาคำตอบเกี่ยวกับฉันทามติระหว่างผู้ประเมินโดยไม่กำหนดคำตอบหรือความเชื่อใดไว้ล่วงหน้า อย่างไรก็ตาม การกำหนดรูปแบบการแจกแจงความหน้าจะเป็นก่อนหน้าส่งผลต่อการแจกแจงความน่าจะเป็นภายหลังของการประมาณค่า นักวิจัยจึงจำเป็นต้องให้ความสำคัญกับปัจจัยต่อไปนี้ 1) ควรมีข้อมูลที่เพียงพอเพื่อลดอิทธิพลของการแจกแจงความน่าจะเป็นก่อนหน้าที่ส่งผลต่อการแจกแจงความน่าจะเป็นภายหลัง และ 2) หากต้องการศึกษาเพื่อสร้างข้อสรุปที่สำคัญเกี่ยวกับพารามิเตอร์ ควรกำหนดการแจกแจงหลาย ๆ รูปแบบในการวิเคราะห์ความอ่อนไหว (Sensitivity analysis) (Gelman, 2013) เพื่อดูว่ารูปแบบการแจกแจงที่เลือกส่งผลต่อข้อสมมติฐานหลักของการวิจัยหรือไม่ ทั้งนี้ หากนักวิจัยวิเคราะห์ข้อมูลโดยใช้ Graphic User Interface ใน

การวิเคราะห์ โปรแกรมจะเลือกการแจกแจงก่อนหน้าที่เหมาะสมให้ตามค่าตั้งต้นของโมเดล หากต้องการเปลี่ยนแปลงรูปแบบการแจกแจงก่อนหน้า รวมถึงกำหนดค่าต่าง ๆ ของโมเดลเพิ่มเติม นักวิจัยสามารถทำได้โดยการเพิ่มคำสั่งในชุดคำสั่ง CCTPack ในโปรแกรม R (Anders, 2017; Batchelder, Anders และ Oravecz, 2018)

เนื่องจากโมเดลฉันทามติเชิงวัฒนธรรมเป็นโมเดลขนาดใหญ่ซึ่งมีพารามิเตอร์จำนวนมากทำให้เป็นไปได้ยากในทางปฏิบัติที่จะคำนวณสมการ 2.12 จึงต้องใช้วิธีการอื่นมาช่วยในการคำนวณ คือ การใช้เทคนิคลูกโซ่มาร์คอฟมอนติคาร์โล (Markov Chain Monte Carlo: MCMC) ซึ่งเป็นเทคนิคการจำลองข้อมูลเพื่อสุ่มตัวอย่างของพารามิเตอร์ที่ต้องการศึกษา โดยการสร้างลูกโซ่ของการเปลี่ยนสถานะในปริภูมิพารามิเตอร์ Ω_{Φ} ความน่าจะเป็นของการเปลี่ยนสถานะของลูกโซ่ขึ้นอยู่กับสถานะก่อนหน้าโดยการเปรียบเทียบความน่าจะเป็นระหว่างสถานะในปริภูมิสถานะ ดังสมการ

$$\frac{\Pr(\Phi_1|x)}{\Pr(\Phi_2|x)} = \frac{L(x|\Phi_1)\Pr(\Phi_1)}{L(x|\Phi_2)\Pr(\Phi_2)} \tag{2.14}$$

ในระบบการจำลองลูกโซ่ ลูกโซ่มาร์คอฟแต่ละสายจะเริ่มต้นจากค่าพารามิเตอร์ที่ต่างกัน หลังจากตัดการเปลี่ยนสถานะลำดับต้น ๆ ออก (burn-in) ลูกโซ่ลำดับที่เหลือจะเป็นการประมาณการแจกแจงของลูกโซ่มาร์คอฟและถูกนำมาประมาณค่าความน่าจะเป็นภายหลัง จากนั้นจึงประเมินการลู่เข้าของลูกโซ่ว่าเข้าสู่การประมาณค่าที่คงที่ของการแจกแจงความน่าจะเป็นภายหลังหรือไม่ (Karabatsos และ Batchelder, 2003; Oravecz และคณะ, 2015; Batchelder, Anders และ Oravecz, 2018)

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

การพัฒนาโมเดลภายใต้ทฤษฎีฉันทามติเชิงวัฒนธรรมในปัจจุบันมีด้วยกัน 8 โมเดล ซึ่งพัฒนาเพื่อรองรับการวิเคราะห์ข้อมูลที่มีลักษณะต่างกันไป ได้แก่ 1) General Condorcet Model (GCM) เป็นโมเดลสำหรับวิเคราะห์ข้อมูลแบบจัดกลุ่มและมีการให้คะแนนแบบ [0, 1] มีพารามิเตอร์หลัก 3 พารามิเตอร์ คือ Z_k (ฉันทามติเชิงวัฒนธรรม) D_i (ความสามารถของตัวอย่าง) และ g_i (ความลำเอียงในการเดาคำตอบ) 2) Multi-culture General Condorcet Model (MC-GCM) เป็นโมเดลที่ขยายจากโมเดล GCM สำหรับวิเคราะห์ข้อมูลที่ตัวอย่างเป็นสมาชิกของต่างกลุ่มวัฒนธรรม มีพารามิเตอร์เพิ่มเติม คือ δ_k (ความยากของข้อคำถาม) และ e_i (สมาชิกกลุ่มวัฒนธรรม/subgroup) 3) Latent Truth Model (LTM) เป็นโมเดลสำหรับวิเคราะห์ข้อมูลแบบต่อเนื่องที่มีค่าอยู่ในช่วง [0,1] เช่น สัดส่วนของความน่าจะเป็นที่จะเกิดเหตุการณ์ต่าง ๆ มีพารามิเตอร์ คือ T_k (ฉันทามติเชิงวัฒนธรรม)

E_i (ความสามารถของตัวอย่าง) และ b_i (เทรชโฮลด์ความลำเอียงในการเลือกคำตอบ) 4) Latent Truth Rater Model (LTRM) เป็นโมเดลการวิเคราะห์ข้อมูลแบบเรียงอันดับ ประกอบด้วยพารามิเตอร์ 6 พารามิเตอร์ ดังนี้ T_k (ตำแหน่งของค่าประจำคำตอบหรือฉันทามติเชิงวัฒนธรรม) λ_k (ความยากของข้อคำถาม) γ_c (เทรชโฮลด์ร่วม) E_i (ความสามารถของตัวอย่าง) a_i (Scaling bias) b_i (Shifting bias) 5) Multi-culture Latent Truth Rater Model (MC-LTRM) เป็นโมเดลที่ขยายจาก LTRM สำหรับการวิเคราะห์ข้อมูลที่ตัวอย่างเป็นสมาชิกของต่างกลุ่มวัฒนธรรม (subgroups) มีพารามิเตอร์เพิ่มเติมจาก LTRM คือ Ω_i (กลุ่มวัฒนธรรม) และ π (ความน่าจะเป็นของการเป็นสมาชิกกลุ่มวัฒนธรรม) 6) Continuous Response Model (CRM) เป็นโมเดลสำหรับวิเคราะห์ข้อมูลที่เป็นค่าต่อเนื่อง มีพารามิเตอร์ทั่วไปคล้ายกับโมเดล LTRM 7) Extended Condorcet Model (ECM) เป็นโมเดลการวิเคราะห์ข้อมูลแบบจัดกลุ่มเช่นเดียวกับโมเดล GCM แต่เพิ่มพารามิเตอร์ในการประมาณค่าการเลือกตอบตัวเลือก “ไม่ทราบคำตอบ” (B) สำหรับตัวอย่างที่ทราบคำตอบและไม่ต้องการเดาคำตอบ และ 8) Item Easiness Model (IEM) เป็นโมเดลสำหรับการวิเคราะห์การให้คะแนนที่เป็นจำนวนเต็ม เช่น การให้คะแนนตามเกณฑ์รูบริค

การศึกษาครั้งนี้ ผู้วิจัยได้ศึกษารายละเอียดของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม 2 โมเดลใหญ่ คือ General Condorcet Model (GCM) และ Latent Truth Rater Model (LTRM) ซึ่งทั้งสองโมเดลประกอบด้วยโมเดลย่อยซึ่งเป็นส่วนขยายเพิ่มเติมสำหรับการวิเคราะห์ข้อมูลกรณีที่ผู้ตอบมีรูปแบบการตอบที่แตกต่างกันตามกลุ่มวัฒนธรรม ซึ่งในการศึกษานี้ ผู้วิจัยได้ศึกษาโมเดลส่วนขยายทั้งสองโมเดล ได้แก่ multi-culture GCM (MC-GCM) multi-culture LTRM (MC-LTRM) โดยมีวัตถุประสงค์เพื่อประยุกต์ใช้โมเดลดังกล่าวในการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินในกรณีที่มีการทำหน้าที่ต่างกันของผู้ประเมิน โดยมีรายละเอียดของโมเดล ดังนี้

1. General Condorcet Model (GCM) เป็นโมเดลการวิเคราะห์ความสอดคล้องของการตอบ ซึ่งในการศึกษานี้เป็นการให้คะแนนในการประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัดโดยผู้เชี่ยวชาญ ในกรณีที่การประเมินนั้นเป็นการให้คะแนนแบบ (0, 1) โดย

$$Z_k = \begin{cases} 1 & \text{เมื่อผลการประเมินในข้อ } k \text{ คือ "สอดคล้อง"} \\ 0 & \text{เมื่อผลการประเมินในข้อ } k \text{ คือ "ไม่สอดคล้อง"} \end{cases}$$

ข้อกำหนดของ General Condorcet Model (GCM)

หากกำหนดให้ผู้ประเมินแต่ละคน (N) ประเมินคุณลักษณะของข้อสอบว่า “สอดคล้อง” หรือ “ไม่สอดคล้อง” กับตัวชี้วัดของหลักสูตรในการประเมิน M ข้อ และคำตอบทั้งหมดเป็นค่าที่สังเกตได้ response profile matrix X = $(X_{ik})_{NxM}$ โดยที่

$$X_{ik} = \begin{cases} 1 \text{ หากผู้ประเมิน } i \text{ ประเมินว่า "สอดคล้อง" ในข้อที่ } k \\ 0 \text{ หากผู้ประเมิน } i \text{ ประเมินว่า "ไม่สอดคล้อง" ในข้อที่ } k \end{cases} \tag{2.15}$$

โมเดล GCM ระบุพารามิเตอร์สำหรับคำตอบที่ได้รับการประเมินว่า “สอดคล้อง” ให้กับชุดคำถาม (Z) = $(Z_k)_{1xm}$ ด้วยช่วงห่าง $\{0, 1\}^M$ นอกจากนี้ โมเดลยังระบุพารามิเตอร์อัตราความแม่นยำ (Hit rate) และ อัตราการแจ้งเตือนที่ผิดพลาด (false alarm rate) สำหรับผู้ให้ข้อมูลแต่ละคน แทนด้วย H = $(H_i)_{1xN}$ และ F = $(F_i)_{1xN}$ ตามลำดับ โดยที่ $\forall i, 0 \leq F_i \leq H_i \leq 1$ ข้อกำหนดที่จำเป็นสำหรับโมเดล คือ อัตราการแจ้งเตือนที่ผิดพลาดของผู้ให้ข้อมูลจะต้องไม่เกินอัตราความแม่นยำ โมเดลพื้นฐานสามารถอธิบายได้ดังต่อไปนี้

1) คะแนนฉันทามติในการประเมินของผู้ประเมินจะมีรูปแบบเดียวต่อหนึ่งชุดรายการประเมิน

2) เงื่อนไขความเป็นอิสระของการตอบที่ปรากฏใน response profile matrix กำหนดโดย

$$\Pr\,[X = (x_{ik})_{NxM}|Z, G, \theta, \delta| = \prod_{i=1}^{N}\prod_{k=1}^{M} \Pr\,(X_{ik} = x_{ik}|Z_k, g_i, \theta_i, \delta_k) \tag{2.16}$$

สำหรับค่าที่สังเกตได้ที่เป็นไปได้ทั้งหมด (X_{ik}) ของ response profile matrix

3) ขอบเขตความน่าจะเป็นของการประเมิน กำหนดโดย

$$\Pr(X_{ik} = X_{ik}|Z_k, g_i, D_{ik}) = \begin{cases} D_{ik} + (1 - D_{ik})g_i \text{ เมื่อ } x_{ik} = 1 \text{ และ } Z_k = 1 \\ (1 - D_{ik})g_i \text{ เมื่อ } x_{ik} = 1 \text{ และ } Z_k = 0 \\ (1 - D_{ik})(1 - g_i) \text{ เมื่อ } x_{ik} = 0 \text{ และ } Z_k = 1 \\ D_{ik} + (1 - D_{ik})x(1 - g_i) \text{เมื่อ } x_{ik} = 0 \text{ และ } Z_k = 0 \end{cases} \tag{2.17}$$

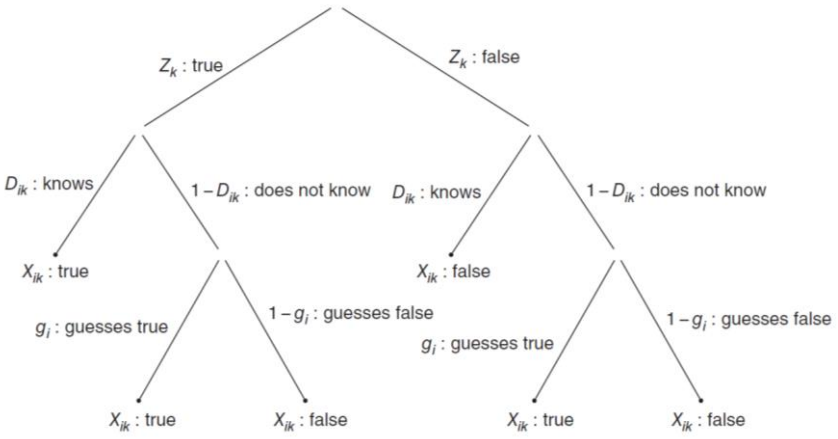
คะแนนฉันทามติในการประเมินในข้อ 1) เป็นข้อตกลงเบื้องต้นที่สำคัญของโมเดลว่าผลคะแนนที่ได้จากการประเมินของผู้เชี่ยวชาญนั้นจะมีคำตอบที่สอดคล้องกันเพียงรูปแบบเดียวต่อผู้เชี่ยวชาญหนึ่งกลุ่ม ในขณะที่เงื่อนไขความเป็นอิสระเป็นข้อตกลงที่คล้ายคลึงกับทฤษฎี

การตอบสนองข้อสอบ (Item Response Theory) รวมถึงโมเดลอื่นที่มีการประมาณค่าพารามิเตอร์ของผู้ตอบและข้อคำถาม โดยมีจุดประสงค์ในการกำหนดให้ความไม่เป็นอิสระระหว่างตัวแปรแบบสุ่มทั้งหลายถูกขจัดออกไปด้วยเงื่อนไขของโมเดล สุดท้าย โมเดลกำหนดการแจกแจงริมขอบในสมการ (2.16) ในแง่ของอัตราความแม่นยำและการแจ้งเตือนที่ผิดพลาด สมการ (2.17) แสดงถึงอัตราความแม่นยำ (Hits) อัตราการแจ้งเตือนที่ผิดพลาด (False alarms) ความคลาดเคลื่อน (Misses) และการปฏิเสธอย่างถูกต้อง (Correct rejections) (Batchelder และ Anders, 2012)

4) โมเดล GCM ประกอบด้วยพารามิเตอร์ความสามารถของผู้ประเมิน (Competence: $D = (Di)1\ x\ n$) และพารามิเตอร์ความลำเอียงในการให้คะแนนประเมิน ($G = (gi)1\ x\ N$) สำหรับผู้ให้ข้อมูลแต่ละคน โดย $\forall \mathrm{i}, 0 \leq \mathrm{D_i} \leq 1,\ 0 \leq \mathrm{g_i} \leq 1$ ความเชื่อมโยงระหว่างพารามิเตอร์ดังกล่าวกับอัตราความแม่นยำและอัตราการแจ้งเตือนที่ผิดพลาด เป็นดังสมการ (2.18)

$$\forall \mathrm{i}, \quad \begin{aligned} \mathrm{H_i} &= \mathrm{D_i} + (1 - \mathrm{D_i})\mathrm{g_i} \\ \mathrm{F_i} &= (1 - \mathrm{D_i})\mathrm{g_i} \end{aligned} \tag{2.18}$$

สมการ (2.18) แสดงในรูปแบบของแผนภูมิการกำหนดพารามิเตอร์ใหม่จาก Hit และ False alarm rate ให้เป็นพารามิเตอร์ความสามารถ (D) และพารามิเตอร์ความลำเอียงในการเดาคำตอบ (g) ได้ดังรูป 2.5



รูป 2.5 การกำหนดพารามิเตอร์ในโมเดล GCM

การกำหนดพารามิเตอร์ใน General Condorcet Model แสดงให้เห็นความน่าจะเป็นในการประเมินของผู้เชี่ยวชาญในสามกรณี กรณีแรก คือ ผู้ประเมินทราบตำแหน่งที่แท้จริงของคะแนนประเมินและประเมินได้ตรงตามฉันทามติของกลุ่มผู้ประเมินด้วยความน่าจะเป็น D_{ik} กรณีที่สอง คือ ผู้ประเมินไม่ทราบตำแหน่งของคะแนนประเมินในข้อ i ด้วยความน่าจะเป็น $1 - Di$ ทำให้มีการเดา

หรือการให้คะแนนที่มีความลำเอียงเกิดขึ้น โดยแยกออกเป็น การเดาคำตอบถูก (gi) กับการเดาคำตอบผิด $(1-gi)$ ดังแสดงในสมการ (2.17)

พารามิเตอร์ D_{ik} ของโมเดล GCM เป็นการประมาณค่าความน่าจะเป็นของทั้งตัวผู้ประเมินและข้อคำถาม ซึ่งทำให้มีจำนวนพารามิเตอร์เกินกว่าจำนวนข้อมูลใน response profile matrix วิธีการแก้ไขการระบุค่าพารามิเตอร์จำนวนมากกระทำโดยการนำรูปแบบของโมเดลการตอบสนองข้อสอบแบบ 1 พารามิเตอร์ หรือราสช์โมเดลเข้ามาใช้ในการประมาณค่า โดยกำหนด α_i เป็นพารามิเตอร์ความสามารถของผู้ประเมิน และ β_k เป็นความยากของรายการประเมิน โดยให้มีขอบเขต $-\infty < \alpha_i, \beta_k < \infty$ และความน่าจะเป็นที่ผู้ประเมิน i จะระบุตำแหน่งคะแนนฉันทามติของข้อคำถาม k ได้ถูกต้อง คือ

$$\Pr(Y_{ik} = 1|\alpha_i, \beta_k) = [1 + e^{-(\alpha_i - \beta_k)}]^{-1} \tag{2.19}$$

การนำราสช์โมเดลเข้ามาช่วยในการระบุโมเดลมีจุดประสงค์เพื่อขจัดปฏิสัมพันธ์ระหว่างผู้ประเมินกับรายการประเมิน ในกรณีที่ผู้ประเมินมีความน่าจะเป็นในการประเมินรายการประเมินได้สอดคล้องกับฉันทามติในข้อใดข้อหนึ่งสูงกว่าข้ออื่น ๆ ราสช์โมเดลจะกำหนดให้ผู้ประเมินคนนั้นมีความน่าจะเป็นที่จะประเมินได้สอดคล้องกับคะแนนฉันทามติมากกว่าผู้ประเมินคนอื่นในทุกรายการประเมิน

ในกรณีที่มีการทำหน้าที่ต่างกันระหว่างผู้ประเมิน ราสช์โมเดลจะถูกนำมาใช้เพื่อระบุความยากของรายการประเมิน ทำให้สามารถระบุความน่าจะเป็นของการรู้ตำแหน่งคะแนนฉันทามติในการประเมินของผู้ประเมิน (D_{ik}) ซึ่งแสดงถึงความแม่นยำในการประเมิน โดยที่ $1 \le i \le N, 1 \le k \le M$

$$D_{ik} = \frac{\theta_i(1-\delta_k)}{\theta_i(1-\delta_k) + \delta_k(1-\theta_i)} \tag{2.20}$$

โดย $0 < \theta_i,\ \delta_k < 1$

เมื่อ พารามิเตอร์ θ_i คือ ความสามารถของผู้ประเมิน และ δ_k คือ ความยากของรายการประเมิน โดยสมการ (2.20) เป็นการปรับพารามิเตอร์ใหม่ซึ่งให้การประมาณค่าที่เท่าเทียมกับราสช์โมเดลในสมการ (2.19) หากผู้ประเมินมีความน่าจะเป็นของความลำเอียงในการประเมินเท่ากับ 0.5 และรายการประเมินมีความยากเท่าเทียมกันแล้ว $D_{ik} = D_i$ ทำให้ความน่าจะเป็นของการประเมินในสมการ (2.17) ปรับเป็นดังนี้

$$\Pr(X_{ik} = x_{ik}|Z_k, D_i) = \begin{cases} \frac{1+D_i}{2} & \text{เมื่อ } X_{ik} = Z_k \\ \frac{1+D_i}{2} & \text{เมื่อ } X_{ik} \neq Z_k \end{cases} \quad (2.21)$$

ตัวอย่างการศึกษาโดยใช้การวิเคราะห์ด้วยโมเดล GCM ได้แก่ การศึกษาของ Trotter, Weller, Baer, Pachter, Glazer, Garcia De Alba Garcia, และ Klein (1999) ใช้โมเดล GCM ในการศึกษาความเชื่อและการรับรู้เกี่ยวกับโรค AIDS ของชาวละตินในรัฐคอนเน็กติคัทและเท็กซัส และในเม็กซิโกและกัวเตมาลา พบว่า กลุ่มผู้ให้ข้อมูลมีความเชื่อและการรับรู้เกี่ยวกับโรค AIDS ร่วมกันในแต่ละพื้นที่ เมื่อเปรียบเทียบคำตอบพบว่าสัดส่วนองค์ความรู้ร่วมของแต่ละพื้นที่ลดลงอย่างมีนัยสำคัญทางสถิติ ในขณะที่องค์ความรู้เกี่ยวกับโรค AIDS จะเข้มข้นและมีรายละเอียดมากขึ้นในเขตที่มีความชุกของโรค นอกจากนี้ องค์ความรู้เกี่ยวกับโรค AIDS ของชาวละตินจะรวมถึงความรู้เกี่ยวกับชีวการแพทย์ของโรค AIDS ผลจากากรวิเคราะห์ข้อมูลของกลุ่มตัวอย่างทั้ง 4 กลุ่ม พบว่า มีกลุ่มตัวอย่าง 3 กลุ่มตอบแสดงความมีฉันทามติในองค์ความรู้เกี่ยวกับโรค AIDS คิดเป็นร้อยละ 90 ดังแสดงในตัวอย่างความสอดคล้องกันของความรู้เกี่ยวกับความเสี่ยงในการติดต่อ

Weller, Baer, Garcia และ Rocha (2012) ได้วิเคราะห์ข้อมูลเกี่ยวกับความเชื่อเกี่ยวกับโรคเบาหวานโดยเก็บข้อมูลจากแพทย์ ผู้ป่วย และตัวแทนจากชุมชนในสหรัฐอเมริกาและเม็กซิโก โดยศึกษาว่าการรับรู้เกี่ยวกับโรคเบาหวานนั้นแตกต่างกันในแต่ละเชื้อชาติ พรมแดนทางภาษาศาสตร์และกลุ่มแพทย์กับคนทั่วไปหรือไม่ การศึกษาครั้งนี้ใช้การวิเคราะห์ตามทฤษฎี Cultural Consensus พบว่า กลุ่มแพทย์มีการรับรู้เกี่ยวกับโรคเบาหวานแตกต่างจากกลุ่มผู้ป่วยและกลุ่มคนทั่วไป

2. Multicultural General Condorcet Model (MC-GCM) เป็นโมเดล GCM ที่ปรับให้รองรับการวิเคราะห์ข้อมูลที่มีคำตอบแฝงได้หลายคำตอบตามกลุ่มผู้ให้ข้อมูล ในกรณีการประเมินมีผู้ประเมินที่มีพฤติกรรมการประเมินที่แตกต่างกัน รวมถึงยอมให้มีความแตกต่างกันของความยากในการประเมิน โมเดล MC-GCM เป็นการขยายผลการศึกษาของ Batchelder และ Anders (2012) เกี่ยวกับการปรับปรุงการอนุมานทางสถิติแบบเบส์แบบลดหลั่น (Hierarchical Bayesian Inference) และได้นำมาใช้เป็นส่วนหนึ่งของโมเดลนี้ การเพิ่มพารามิเตอร์จำนวนกลุ่มวัฒนธรรมทำให้สามารถประเมินได้ว่าผู้ให้ข้อมูลมีลักษณะของการแบ่งกลุ่มย่อยภายใน (subgroup) หรือไม่ ในกรณีของการวิเคราะห์ความสอดคล้องของผู้ประเมิน MC-GCM จะช่วยให้ทราบได้ว่าผู้ประเมินมีการทำหน้าที่แตกต่างกันในการประเมินหรือไม่ และมีลักษณะการทำหน้าที่ต่างกันอย่างไร รวมถึงการประมาณค่าความน่าเชื่อถือ (D_{ik}) ของผู้ประเมินทั้งรายบุคคลและกลุ่ม

ข้อกำหนดของ Multicultural General Condorcet Model (MC-GCM)

MC-GCM ใช้ข้อมูลจากเมตริกซ์ X_{ik} ในสมการ (2.15) เช่นเดียวกับ GCM แต่ผ่อนปรนข้อตกลงเกี่ยวกับการมีฉันทามติร่วมเพียงหนึ่งเดียวของผู้ประเมินให้สามารถมีความแตกต่างของฉันทามติได้ตามกลุ่มผู้ประเมินโดยผู้ประเมินจะเป็นสมาชิกของกลุ่มฉันทามติกลุ่มหนึ่งกลุ่มใดเพียงกลุ่มเดียว นอกจากนี้ยังยินยอมให้มีความแตกต่างของความยากในการประเมินได้ ลักษณะของโมเดลและข้อตกลงเบื้องต้นของ MC-GCM เป็นไปตามที่ Batchelder และ Anders (2012) อธิบายไว้ในการศึกษา GCM ในส่วนของฉันทามติของกลุ่ม MC-GCM ระบุไว้ดังนี้

$$\forall k, \qquad Z_k \begin{cases} 1 \text{ ถ้าคำตอบข้อ } k \text{ เป็นคำตอบที่กลุ่มเห็นว่าเป็นคำตอบที่ถูก} \\ 0 \text{ ถ้าคำตอบข้อ } k \text{ เป็นคำตอบที่กลุ่มเห็นว่าเป็นคำตอบผิด} \end{cases} \tag{2.22}$$

GCM มีข้อตกลงเกี่ยวกับอัตราความแม่นยำ H_i และการแจ้งเตือนที่ผิดพลาด F_i (Hit and False Alarm Rate) เช่นเดียวกับโมเดล GCM โดย

$$\Pr(X_{ik} = 1 | H_i, F_i, Z_k) = \begin{cases} H_i \text{ ถ้า } Z_k = 0 \\ F_i \text{ ถ้า } Z_k = 1 \end{cases} \tag{2.23}$$

ข้อมูลในการวิเคราะห์ MC-GCM ได้จากเมตริกซ์ $\mathrm{X_{ik}}$ ซึ่งในกรณีของ MC-GCM จะถือว่าข้อมูลดังกล่าวเป็นรูปแบบผสมของ GCM ซึ่งมีคุณลักษณะที่แตกต่างกันตามฉันทามติ โดย MC-GCM ถือว่ามีฉันทามติจำนวน $\mathrm{T} \geq 1$ กลุ่ม และคำตอบที่เป็นฉันทามติของกลุ่มมีค่า $\boldsymbol{z} = \{\mathrm{Z_1}, \dots, \mathrm{Z_T}\}$ แทนที่จะเป็นฉันทามติค่าเดียวเหมือนดังโมเดล GCM นอกจากนี้ MC-GCM ยังกำหนดให้มีพารามิเตอร์แฝงของกลุ่มวัฒนธรรม เป็น $\mathbf{E} = (\mathrm{e_i})_{1\mathrm{xN}}$ เมื่อ $\mathbf{e_i} \in \{1, \dots, \mathrm{T}\}$ เมื่อ Z_{ei} คือการประเมินที่เป็นฉันทามติของผู้ประเมิน i

โมเดล MC-GCM มีพารามิเตอร์ทั้งหมด $3\mathrm{N}$ พารามิเตอร์ เมื่อ N คือ จำนวนผู้ประเมินรวมกับพารามิเตอร์ความสามารถ ความลำเอียง และสมาชิกกลุ่มฉันทามติ และ พารามิเตอร์ของกลุ่มวัฒนธรรมจำนวน $\mathrm{T} \cdot \mathrm{M} + \mathrm{M}$ ในกรณีความยากในการประเมินแตกต่างกัน หรือ $\mathrm{T} \cdot \mathrm{M}$ ในกรณีความยากในการประเมินเท่าเทียมกัน ในการประมาณค่าเมตริกซ์ $\mathrm{X_{ik}}$ ที่มีจำนวนข้อมูลขนาด NxM ในการประมาณค่าที่มีข้อมูลเพียงพอต่อพารามิเตอร์ Anders และ Batchelder (2012) กำหนดให้ $\mathrm{Min}\{\mathrm{N}, \mathrm{M}\} \geq \mathrm{T} + 4$ เมื่อ N คือ จำนวนผู้ประเมิน M คือจำนวนข้อสอบหรือผู้รับการประเมิน และ T คือ จำนวนชุดคำตอบที่เป็นฉันทามติของผู้ประเมิน

Anders และ Batchelder (2012) เสนอสถิติสำหรับการตรวจสอบการรวมพารามิเตอร์ความยากของข้อคำถามในการวิเคราะห์ MC-GCM และโมเดลฉันทามติเชิงวัฒนธรรมอื่น ๆ เรียกว่า Variance Dispersion Index (VDI) ซึ่งสะท้อนความยากของการเลือกคำตอบหรือการให้คะแนนของผู้ให้ข้อมูลในแต่ละข้อคำถาม ค่า VDI คำนวณได้จากความแปรปรวนของการตอบของผู้ให้ข้อมูลทั้งหมดในแต่ละข้อคำถาม แล้วนำมาคำนวณความแปรปรวนของข้อคำถามในแต่ละข้ออีกครั้งหนึ่ง ดังนี้

$$VDI(X) = \sum_{K=1}^{M} \frac{V_k^2}{M} - \left(\sum_{k=1}^{M} V_k / M\right)^2 \qquad (2.24)$$

เมื่อ

$$V_k = \sum_{i=1}^{N} \frac{X_{ik}^2}{N} - \left(\sum_{i=1}^{N} X_{ik} / N\right)^2$$

ค่า VDI สามารถใช้ในการตรวจสอบความน่าจะเป็นภายหลังว่าควรรวมพารามิเตอร์ความยากของข้อคำถามเข้าในการตรวจสอบความสอดคล้องของโมเดลหรือไม่ ทั้งนี้ ค่า VDI ควรอยู่ในช่วงเปอร์เซ็นไทล์ที่ 10 ถึง 90 อย่างไรก็ตาม ในกรณีของการวิเคราะห์ข้อมูลพหุวัฒนธรรม นักวิจัยควรใช้การเปรียบเทียบค่า Deviance Information Criterion (DIC) เพื่อเลือกโมเดลที่เหมาะสมโดยพิจารณาจากโมเดลที่มีค่า DIC ต่ำกว่า

การประมาณค่าของโมเดล MC-GCM ใช้วิธีการประมาณค่าแบบเบส์แบบลดหลั่น (Bayesian hierarchical framework) ซึ่งอธิบายรายละเอียดใน Batchelder และ Anders (2012) ในการศึกษาเกี่ยวกับ MC-GCM นั้น Anders และ Batchelder (2012) ได้กำหนดการแจกแจงของพารามิเตอร์ ดังต่อไปนี้

$$Z_{tk} \sim Bernoulli(p_t)$$
$$\delta_k \sim Beta(\mu_\delta \tau_\delta, (1-\mu_\delta)\tau_\delta)$$
$$\theta_k \sim Beta\ (\mu_{\theta_{ei}} \tau_{\theta_{ei}}, (1-\mu_{\theta_{ei}})\tau_{\theta_{ei}})$$
$$g_i \sim Beta(\mu_{g_{ei}} \tau_{g_{ei}}, (1-\mu_{g_{ei}})\tau_{g_{ei}})$$
$$e_i \sim Categorical\ (\lambda)$$

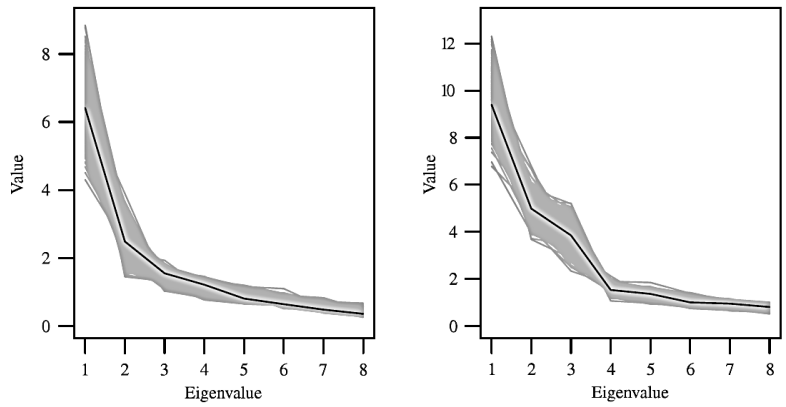
เมื่อ $t \in \{1, \ldots, T\}$ และกำหนดค่า T ไว้แล้ว

การแจกแจงความน่าจะเป็นของ $\mathrm{Z_{tk}}$ เป็นการแจกแจง Bernoulli เมื่อ $Z_{tk} \in \{0,1\}$ เช่นเดียวกับโมเดล GCM และเมื่อพารามิเตอร์ δ_k θ_k และ g_i มีปริภูมิตัวอย่างในขอบเขต [0, 1] จึงได้กำหนดการแจกแจงความน่าจะเป็นของพารามิเตอร์เป็นการแจกแจงเบตาเนื่องจากการแจกแจง

ดังกล่าวจะทำให้การประมาณค่าความน่าจะเป็นของค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของผู้ประเมินแต่ละกลุ่มเป็นอิสระต่อกัน (Anders และ Batchelder, 2012) จากนั้น กำหนดการแจงแจงความน่าจะเป็นก่อนหน้าของไฮเปอร์พารามิเตอร์ ดังนี้ (Anders และ Batchelder, 2012)

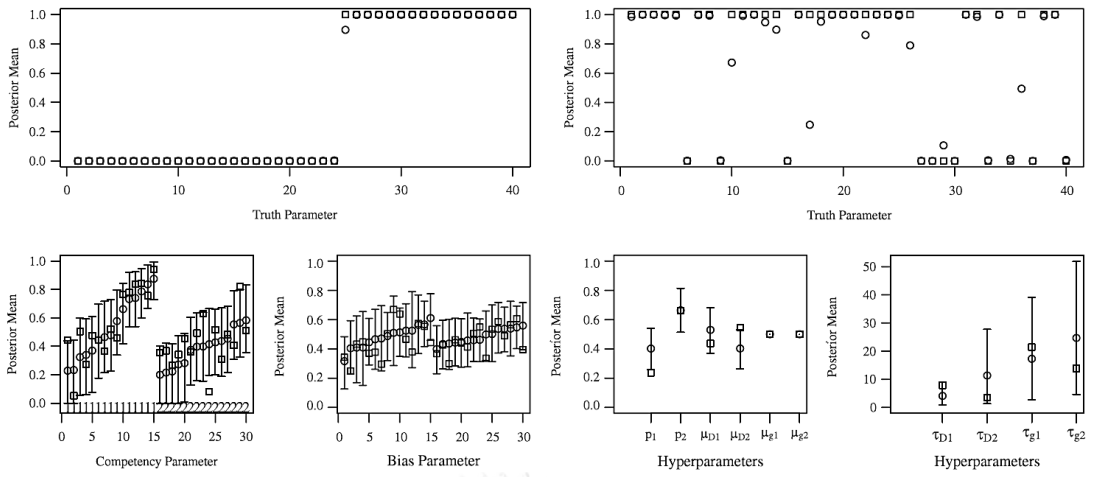
$$p_t \sim Uniform\ (0,1)$$
$$\mu_{\theta t} \sim Beta\ (\alpha, \alpha), \alpha = 2$$
$$\tau_{\theta_t} \sim Gamma\ (\mu_{\tau\theta}^2/\sigma_{\tau\theta}^2, \mu_{\tau\theta}/\sigma_{\tau\theta}^2), \mu_{\tau\theta} = 10, \sigma_{\tau\theta} = 10$$
$$\mu_{gt} = 1/2$$
$$\tau_{gt} \sim Gamma\ (\mu_{\tau g}^2/\sigma_{\tau g}^2, \mu_{\tau g}/\sigma_{\tau g}^2), \mu_{\tau g} = 10, \sigma_{\tau g} = 10$$
$$\mu_{\delta} = 1/2$$
$$\tau_{\delta} \sim Gamma\ (\mu_{\tau\delta}^2/\sigma_{\tau\delta}^2, \mu_{\tau\delta}/\sigma_{\tau\delta}^2), \mu_{\tau\delta} = 10, \sigma_{\tau\delta} = 10$$
$$\lambda \sim Dirichlet(L), L = (1)_{1xT}$$

Anders และ Batchelder (2012) ศึกษาประสิทธิภาพของ MC-GCM ในการประมาณค่าฉันทามติเชิงวัฒนธรรมด้วยข้อมูลจำลองและข้อมูลจริง การศึกษากับข้อมูลจำลองพบว่า MC-GCM ประมาณค่าความแตกต่างของกลุ่มฉันทามติได้ตรงตามที่กำหนด ดังรูป 2.6 แสดงให้เห็นผลการประมาณค่าจำนวนกลุ่มฉันทามติเชิงวัฒนธรรมจากข้อมูลจำลองเมื่อกำหนด $\mathrm{T} = 2$ และ $T = 3$ และให้การประมาณค่าพารามิเตอร์ต่าง ๆ ของโมเดล



รูป 2.6 การตรวจสอบการประมาณค่าจำนวนกลุ่มฉันทามติของโมเดล MC-GCM (Anders และ Batchelder, 2012)

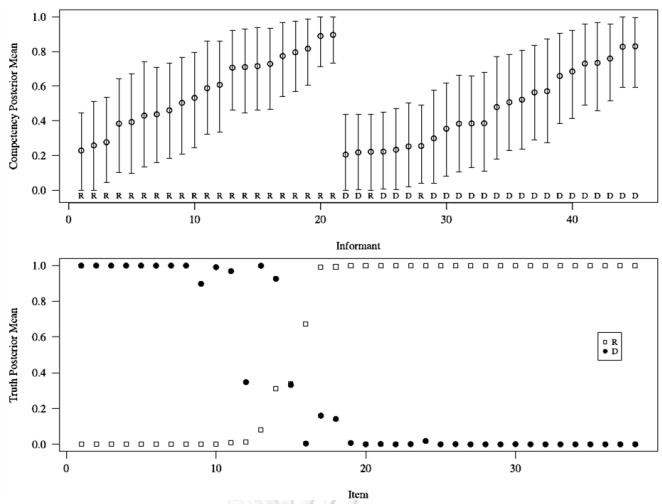
ในรูป 2.7 เป็นการประมาณค่ากลุ่มวัฒนธรรมจากข้อมูลจำลอง เมื่อกำหนด $\mathrm{T} = 1$ (ซ้าย) และ $T = 2$ (ขวา) ในขณะที่ด้านล่างเป็นผลการประมาณค่าพารามิเตอร์ความสามารถของผู้ประเมิน ความลำเอียง และไฮเปอร์พารามิเตอร์ ในช่วง 95% HDI เห็นได้ว่าโมเดลสามารถประมาณค่าข้อมูลที่มีความแตกต่างของกลุ่มวัฒนธรรมได้อย่างมีประสิทธิภาพ



รูป 2.7 ผลการประมาณค่าพารามิเตอร์ของโมเดล MC-GCM

(Anders และ Batchelder, 2012)

Anders และ Batchelder (2012) ยังศึกษาประสิทธิภาพของการประมาณค่าของโมเดล MC-GCM กับข้อมูลจริง 2 ชุด พบว่าโมเดลสามารถประมาณค่าจำนวนกลุ่มฉันทามติได้ดี รวมถึงสามารถจำแนกการเป็นสมาชิกกลุ่มฉันทามติของผู้ให้ข้อมูลได้อย่างมีประสิทธิภาพ นอกจากนี้ ในการทดลองวิเคราะห์ข้อมูลเกี่ยวกับความคิดเห็นทางการเมืองของกลุ่มคนที่เป็นสมาชิกพรรคการเมืองของสหรัฐอเมริกา 2 พรรคการเมือง พบว่าโมเดลสามารถจำแนกความแตกต่างของความเชื่อระหว่างสมาชิกพรรคทั้ง 2 ได้อย่างมีประสิทธิภาพ รวมถึงระบุรูปแบบของความเชื่อจากการตอบคำถามได้ ดังรูปที่ 2.8 (บน) แสดงการจำแนกกลุ่มของผู้ให้ข้อมูลคำถามระหว่างพรรคเดโมแครทและพรรครีพับลิกัน โดยกลุ่มแรก (R) เป็นกลุ่มของผู้ให้ข้อมูลคำถามที่เป็นสมาชิกพรรครีพับลิกัน และกลุ่มที่สอง (D) เป็นกลุ่มของสมาชิกพรรคเดโมแครท และรูป 2.8 (ล่าง) เป็นรูปแบบการตอบคำถามที่เป็นแบบจัดกลุ่ม 2 ตัวเลือก เห็นได้ว่าผู้ให้ข้อมูลที่เป็นสมาชิกพรรคการเมืองทั้งสองพรรคมีรูปแบบการเลือกตอบคำตอบที่แตกต่างกันอย่างชัดเจน กล่าวคือ ในข้อที่สมาชิกพรรคเดโมแครทเลือกตอบตัวเลือกที่ 1 สมาชิกพรรครีพับลิกันจะเลือกตอบตัวเลือกที่สอง



รูป 2.8 ผลการประมาณค่าสมาชิกกลุ่มวัฒนธรรมและรูปแบบการตอบคำถามของโมเดล MC-GCM (Anders และ Batchelder, 2012)

3. Latent Truth Rater Model (LTRM) เป็นโมเดลที่พัฒนาต่อจาก LTM ใช้สำหรับข้อมูลแบบเรียงลำดับ เพื่อที่จะทำให้โมเดลของ Cultural Consensus Theory ใช้ได้กับข้อมูลเรียงลำดับที่เป็นแบบ 2 ค่า และหลายค่า นอกเหนือไปจากการประมาณค่าความสามารถและความยากของการตอบข้อคำถามแล้ว LTRM ยังถูกกำหนดให้ใช้เพื่อการประมาณค่าตัวเลือกของคำถามที่มีร่วมกันเชิงวัฒนธรรมในรูปแบบของมาตรเรียงลำดับ และประมาณค่าแนวโน้มของความลำเอียงในการตอบคำถามของผู้ให้ข้อมูลแต่ละคนในวัฒนธรรมนั้น โครงสร้างของ LTRM จะเหมือนกับโมเดลตรวจจับสัญญาณแบบมาตรฐาน แต่มีส่วนประกอบเพิ่มเติมสำหรับแต่ละข้อคำถามขึ้นอยู่กับตำแหน่งของความจริงแฝงร่วมและผลของการตอบข้อคำถาม LTRM ถือว่าการตอบคำถามแต่ละครั้งของผู้ให้ข้อมูลสะท้อนคุณค่าของตำแหน่งคำตอบที่อยู่ภายในข้อคำถาม แม้ว่าความน่าจะเป็นของการเลือกตอบจะแตกต่างกันระหว่างผู้ให้ข้อมูลซึ่งขึ้นอยู่กับปัจจัยที่มีต่อการตอบข้อคำถาม เช่น ความรู้ความสามารถของผู้ให้ข้อมูล

ข้อกำหนดของ Latent Truth Rater Model (LTRM)

สมมติว่าผู้ให้ข้อมูล N แต่ละคนตอบคำถามเป็นข้อมูลที่เป็นการจัดกลุ่มแบบเรียงลำดับ $\mathrm{C} \in \{1, \dots, \mathrm{C}\}$ ในแต่ละข้อคำถาม และให้คำตอบเป็นค่าที่สังเกตได้ใน response profile matrix $\mathrm{X} = (\mathrm{X_{ik}})_{\mathrm{NxM}}$ เมื่อ

$\mathrm{X_{ik}} = \mathrm{c}$ เมื่อ ผู้ให้ข้อมูล i เลือกคำตอบ c ให้กับข้อ k (2.25)

LTRM กำหนดว่าแต่ละข้อคำถามจะมี “ค่าความจริงแท้” ในรูปของตำแหน่งฉันทามติแฝงซึ่งตำแหน่งของคำตอบเหล่านี้ถูกกำหนดให้เป็นพารามิเตอร์ในโมเดล ค่าตัวเลือกในคำถามไม่ว่าจะในแง่ของความจริง ความน่าจะเป็น หรือระดับองศาจะถูกแปลให้เป็นค่าในรูป (0,1) ยกตัวอย่างเช่น การเลือกตอบคำถามในมาตรประมาณค่า (Likert scale) ที่มี ระดับ 1 ถึง 5 นัยของการวิเคราะห์ LTRM ตามทฤษฎี Consensus Analysis คือ มีตัวเลือกเดียว หรือ ค่าเพียงค่าเดียวในจำนวน 5 ระดับ ที่เป็นคำตอบที่แท้จริงของกลุ่ม ดังนั้น สมมติว่าในคำถาม “การฝึกอบรมมีความจำเป็นเพียงใดต่อหน่วยงาน” ผู้ให้ข้อมูล i เลือกตอบระดับที่ 2 (น้อย) แต่ผู้ให้ข้อมูลคนอื่น ๆ เลือกตอบระดับที่ 5 (มากที่สุด) LTRM จะแปลค่าคำตอบ 2 เป็น 0 (คำตอบที่ผิด หรือไม่ตรงกับฉันทามติ) และแปลค่าคำตอบ 5 เป็น 1 (คำตอบที่ถูก หรือความจริงที่สอดคล้องกับฉันทามติของกลุ่ม) ทั้งนี้ การเลือกตอบในระดับใดขึ้นอยู่กับความรู้ของผู้ให้ข้อมูล (ในกรณีตัวอย่างอาจเป็นความรู้เกี่ยวกับการอบรม หรือความรู้เกี่ยวกับการพัฒนาองค์กร) หากผู้ให้ข้อมูลมีความรู้ความสามารถสูงหรือเข้าใจบริบท/สถานการณ์ หรือวัฒนธรรมที่เกี่ยวข้อง ก็จะเลือกคำตอบที่สอดคล้องกับคำตอบที่ “เป็นจริง” มากที่สุด นอกจากความสามารถของผู้ให้ข้อมูลแล้ว ความยากของคำถามก็เป็นอีกพารามิเตอร์หนึ่งที่ถูกนำมาวิเคราะห์ด้วย เช่นเดียวกับความลำเอียงในการเลือกคำตอบ และระดับของการเปลี่ยนคำตอบ (thresholds) ดังนั้นการที่ผู้ให้ข้อมูลเลือกว่าจะตอบข้อใดจึงขึ้นอยู่กับการประเมิน (appraisal) ของผู้ให้ข้อมูล ซึ่งการประเมินนี้มีความคลาดเคลื่อน $Y_{ik} = T_k + \epsilon_{ik}$ การประเมินนี้เป็นตัวแปรแฝงในรูปของการแจกแจงอิสระที่มีค่าเฉลี่ยเลขคณิตเป็นระดับหรือค่าของคำตอบที่แท้จริง และมีค่าความแม่นยำ $T_{ik} > 0$ เพื่อแสดงปริมาณความคลาดเคลื่อนที่อยู่รอบค่าเฉลี่ยเลขคณิต ความแม่นยำ (precision) นี้ขึ้นอยู่กับความสามารถของผู้ให้ข้อมูลและความยากของคำถามดังที่กล่าวมาแล้ว นอกจากนี้ยังมีพารามิเตอร์ที่ระบุระดับของการเปลี่ยนคำตอบ (thresholds) จากนั้น คำตอบ X_{ik} ถูกระบุโดยตำแหน่งของ Y_{ik} ตามเทรชโฮลด์ของผู้ให้ข้อมูล หรือ $-\infty < \delta_{i1} < \delta_{i2} < \cdots < \delta_{iC-1} < \infty$

ข้อกำหนดของ LTRM อยู่ภายใต้สัจพจน์ 6 สัจพจน์ ดังต่อไปนี้

สัจพจน์ที่ 1 ความจริงเชิงวัฒนธรรม (Cultural Truth) เช่นเดียวกับโมเดลอื่นๆ ที่กล่าวมาแล้ว มีความจริงเชิงวัฒนธรรมร่วมกันระหว่างผู้ให้ข้อมูลซึ่งเป็นของตำแหน่งของคำตอบร่วมกับของผู้ให้ข้อมูล หรือ $T = (T_k)_{1xM}$ เมื่อ $T_k \in (-\infty, \infty)$

สัจพจน์ที่ 2 ขอบเขตของการประเมินคำตอบ ผู้ให้ข้อมูลทุกคนมีการประเมินการเลือก Y_{ik} ตอบแฝงอยู่ในการตอบแต่ละคำตอบ การประเมินคำตอบนี้สอดคล้องกับ $Y_{ik} = T_k + \epsilon_{ik}$ ในขณะที่ความคลาดเคลื่อนของการประเมินคำตอบสอดคล้องกับ

$\forall i, k;\ \epsilon_{ik} \sim \mathrm{Normal}(0, \tau_{ik})$ เมื่อ τ_{ik} คือ ความแม่นยำ (precision) (2.26)

สัจพจน์ที่ 3 ความคลาดเคลื่อน (Error Precision) กำหนดโดยพารามิเตอร์ความสามารถของผู้ให้ข้อมูล $E = (Ei)_{1xn}$ และพารามิเตอร์ความยากของข้อคำถาม $\Lambda = (\lambda_k)_{1xM}$ ความคลาดเคลื่อนนี้ได้จาก

$$\forall i, k;\ \tau_{ik} = E_i/\lambda_k) \tag{2.27}$$

เมื่อ $E_i, \lambda_k > 0$ หากข้อคำถามมีความยากเท่ากันในแต่ละข้อแล้ว จะได้เป็น $\forall k, \lambda_k = 1$

สัจพจน์ที่ 4 เงื่อนไขความเป็นอิสระ ค่าการประเมินคำตอบแฝง $Y = (Y_{ik})_{NxM}$ เป็นเงื่อนไขความเป็นอิสระที่ให้พารามิเตอร์ตำแหน่งข้อคำถามและความแม่นยำ ซึ่งคือค่าที่สังเกตได้ (Y_{ik}) ของ Y องค์ประกอบร่วม $h(\cdot)$ ที่มีเงื่อนไขบนพารามิเตอร์คือ

$$h[(y_{ik}) \mid \mathbf{T}, \mathbf{E}, \mathbf{\Lambda}] = \prod_{i=1}^{N}\prod_{k=1}^{M} f(y_{ik} \mid T_k, \tau_{ik}) \tag{2.28}$$

เมื่อ $f(Y_{ik} \mid T_k, \tau_{ik})$ เป็นการแจกแจงชายขอบของการประเมินคำตอบบนเงื่อนไขความสอดคล้องกันของพารามิเตอร์

สัจพจน์ที่ 5 เทรชโฮลด์การเปลี่ยนกลุ่มคำตอบ (Category Thresholds) ผู้ให้ข้อมูลแต่ละคนจะมีช่วงการเปลี่ยนแปลงคำตอบ (tresholds) จำนวน C-1 หรือ $-\infty < \delta_{i1} < \delta_{i2} < \cdots < \delta_{iC-1} < \infty$ คำตอบของผู้ให้ข้อมูล X_{ik} ถูกกำหนดโดยตำแหน่งของ Y_{ik} ตามกลุ่มของเทรชโฮลด์ ดังนี้

$$X_{ik} = \begin{cases} 1 & \text{if } Y_{ik} \le \delta_{i1}, \\ c & \text{if } \delta_{i,c-1} < Y_{ik} \le \delta_{ic} \\ C & \text{if } Y_{ik} > \delta_{i,C-1}. \end{cases} \tag{2.29}$$

สัจพจน์ที่ 6 ความลำเอียงของการตอบ (Response Biases) ประกอบด้วยขอบร่วมของกลุ่ม $G = (\gamma_c)xC - 1$ และพารามิเตอร์ความลำเอียงของผู้ให้ข้อมูล 2 พารามิเตอร์ที่ทำงานร่วมกับ G คือ scaling bias $A = (a_i)_{1xN}$ ด้วยขอบเขต $(0, \infty)$ และ shifting bias $B = (b_i)_{1xn}$ ด้วยขอบเขต $(-\infty, \infty)$ ขอบร่วมของกลุ่มแต่ละอันกำหนดโดย

$$\delta_{ic} = a_i\, \gamma_c + b_i$$
เมื่อ $A = (a_i)_{1xN}$ ในช่วง $(0, \infty)$ และ $B = (b_i)_{1xN}$ ในช่วง $(-\infty, \infty)$ (2.30)

สัจพจน์เหล่านี้ได้รับการออกแบบเพื่อสร้างรูปแบบการตอบในแต่ละระดับคำตอบของผู้ให้ข้อมูลที่มีความสามารถ (E_i) ต่างกันให้เข้ากับข้อคำถามที่มีตำแหน่งของคำตอบที่แท้จริงแฝงอยู่ร่วมกัน

(T_k) สัจพจน์ที่ 1 ระบุตำแหน่งของคำตอบที่เป็นจริง สัจพจน์ที่ 2 แสดงว่าผู้ให้ข้อมูลแต่ละคนมีการประเมินแฝงสำหรับคำตอบที่เป็นจริงแต่ละข้อพร้อมความคลาดเคลื่อนในรูปของการแจกแจงปกติในบริเวณรอบ ๆ ค่าที่แท้จริงด้วยความแม่นยำ τ_{ik} ซึ่งขึ้นอยู่กับผู้ให้ข้อมูล i และข้อที่ k ข้อตกลงนี้ใกล้เคียงกับข้อตกลงเกี่ยวกับคะแนนจริงในทฤษฎีการทดสอบแบบดั้งเดิม หรือ X = T + E นั่นคือการเลือกคำตอบแต่ละข้อ ผู้ให้ข้อมูลจะประเมินค่าของคำตอบด้วยความรู้ความสามารถรวมถึงความคลาดเคลื่อนในการประเมินการเลือกตอบ สัจพจน์ที่ 3 จัดให้ τ_{ik} ขึ้นอยู่กับความสามารถของผู้ให้ข้อมูลและความยากของข้อคำถามแล้วคำนวณจากสมการที่ (2.27) สัจพจน์ที่ 4 เป็นเงื่อนไขความเป็นอิสระเช่นเดียวกับโมเดลการตอบสนองข้อสอบและโมเดลพาราเมตริกอื่น ๆ ที่ตัวแปรแฝงมีโครงสร้างคล้ายคลึงกับ Y สัจพจน์ที่ 5 และ 6 เกี่ยวข้องกับการกำหนดเทรชโฮลด์สำหรับผู้ให้ข้อมูลแต่ละคนที่จะคัดกรองการปะเมินคำตอบแฝง Y_{ik} ให้เป็นข้อมูลแบบเรียงลำดับที่มีความชัดเจน สัจพจน์ดังกล่าวยอมรับว่ามีชุดร่วมของกลุ่มเทรชโฮลด์ G อย่างไรก็ตาม ผู้ให้ข้อมูลแต่ละคนอาจสร้างความลำเอียงให้กับเทรชโฮลด์ร่วมเหล่านี้ด้วย scaling factor (a_i) และ shift factor (b_i) ยกตัวอย่างเช่น ครูมีสัดส่วนในการให้คะแนนนักเรียนไว้อยู่แล้ว หรือการพยายามตอบให้มีการกระจายของการเลือกคำตอบเท่าๆ กัน เป็นต้น

รูปที่ 2.9 แสดงโมเดล LTRM ตามที่ระบุในสัจพจน์ที่ 1 – 6 สำหรับคำถามที่มีคำตอบ 5 ระดับ จำนวน 3 ข้อ ที่ตอบโดยผู้ให้ข้อมูลคนเดียวกัน เมื่อ

T_k คือ ค่าประจำตำแหน่งของคำตอบ

λ_k คือ ความยากของข้อคำถาม

Y_{ik} คือ การประเมินการเลือกคำตอบ

γ_c คือ เทรชโฮลด์ร่วม (no bias)

δ_{ic} คือ เทรชโฮลด์ร่วม (with bias)

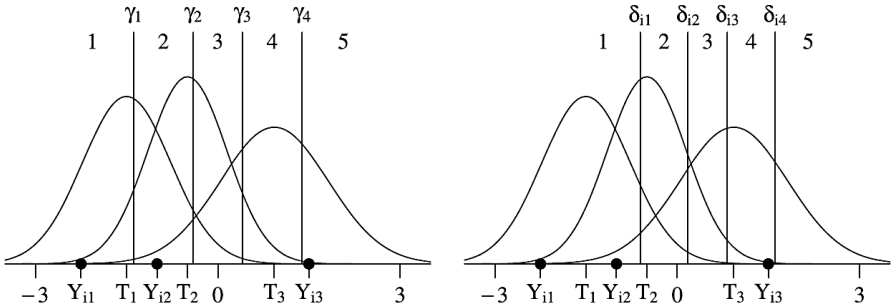
E_i คือ ความแม่นยำในการระบุตำแหน่งคะแนนของผู้ประเมิน

a_i คือ scaling bias

b_i คือ shifting bias

จะเห็นว่าการประเมินการเลือกคำตอบในเมตริกซ์คุณลักษณะแฝง Y_{ik} ของกราฟที่ไม่มีความลำเอียงและมีความลำเอียงตกอยู่ในระดับต่างกัน ส่งผลต่อการเลือกตอบที่ปรากฏในเมตริกซ์สังเกตได้ X_{ik} ของทั้งกราฟทั้งสองในค่าการประเมินเดียวกัน ในกราฟด้านขวาจะพบการหดตัวของระยะเทรชโฮลด์ รวมถึงเทรชโฮลด์ร่วม δ_{ic} ที่เลื่อนตัวไปทางด้านขวาของกราฟ จากพารามิเตอร์

ความลำเอียงทั้ง 2 พารามิเตอร์ แสดงให้เห็นว่าผู้ประเมิน (rater) มีแนวโน้มให้คะแนนแบบสุดโต่ง และมักจะให้คะแนนในระดับต่ำ



รูป 2.9 โมเดล LTRM แบบไม่มี bias (ซ้าย) และแบบมี bias (ขวา)
(Anders และ Batchelder, 2015)

คุณสมบัติของ Latent Truth Rater Model (LTRM)

1) Spearman's Law Property เช่นเดียวกับ GCM และ LTM คุณสมบัติที่สำคัญของ LTRM คือการแสดงถึงการมีวัฒนธรรมร่วมกันเพียงหนึ่งเดียว ดังแสดงให้เห็นในการพิสูจน์คุณสมบัติของทั้งสองโมเดลที่ผ่านมา คุณสมบัตินี้ได้ถูกนำมาใช้ในการพัฒนาโมเดลตรวจสอบ Bayesian posterior-predictive model สำหรับ GCM เช่นเดียวกับใน LTRM

2) The Inclusion of Item Difficulty การรวมค่าความยากของข้อคำถามในโมเดลเป็นคุณสมบัติสำคัญรองจากคุณสมบัติข้อแรก คุณสมบัตินี้เกี่ยวข้องกับสัจพจน์ที่ 3 และใช้ในการตรวจสอบโมเดลเพื่อแยกความแตกต่างของข้อตกลงเบื้องต้นเกี่ยวกับความเป็นกลางและความเป็นวิวิธพันธ์ของความยากของข้อคำถาม Batchelder และ Anders (2012) ได้เสนอสถิติตัวใหม่ที่เรียกว่า Variance Dispersion Index (VDI) ซึ่งทำหน้าที่สะท้อนความแตกต่างของข้อคำถามอันเนื่องมาจากความผันแปรของการตอบ ทั้งนี้ ใน LTRM การคำนวณค่าความยากมีแนวโน้มที่จะทำให้ค่า VDI เพิ่มขึ้น และทำให้สามารถตรวจสอบความจำเป็นที่จะรวมค่าความยากเข้ามาเพื่อให้สอดคล้องกับความแม่นยำที่แตกต่างกันของการตอบข้ามข้อคำถาม

4. Multi-culture Latent Truth Rater Model (MC-LTRM) เป็นโมเดลที่ขยายจาก Latent Truth Rater Model (LTRM) ซึ่งใช้สำหรับวิเคราะห์ฉันทามติของผู้ให้ข้อมูลสำหรับการตอบคำถามที่เป็นลักษณะข้อมูลแบบเรียงลำดับ LTRM ถูกกำหนดให้ใช้เพื่อการประมาณค่าตัวเลือกของคำถามที่มีร่วมกันเชิงวัฒนธรรมในรูปแบบของมาตรเรียงลำดับ และประมาณค่าแนวโน้มของ

ความลำเอียงในการตอบคำถามของผู้ให้ข้อมูลแต่ละคนในวัฒนธรรมนั้น LTRM เป็นโมเดลที่ออกแบบเพื่อประมาณค่าวัฒนธรรมรวมระหว่างข้อคำถามบนมาตรเรียงอันดับและความลำเอียงในการตอบของผู้ให้ข้อมูลแต่ละคนในกลุ่มวัฒนธรรมนอกเหนือไปจากความรู้หรือความเชี่ยวชาญของผู้ให้ข้อมูลและความยาก (ในการให้คะแนน) ของข้อคำถาม ข้อกำหนดที่สำคัญของ LTRM และโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม คือ ข้อมูลต้องมีฉันทามติร่วมเพียงหนึ่งเดียว ทั้งนี้ รายละเอียดและข้อกำหนดของโมเดลระบุไว้ในการศึกษาของ Anders และ Batchelder (2015)

Multi-Culture Latent Truth Rater Model (MC-LTRM) ได้รับการขยายต่อจาก LTRM โดยปรับข้อกำหนดของโมเดลเกี่ยวกับคำตอบเชิงวัฒนธรรมให้แตกต่างกันตามกลุ่มวัฒนธรรมของผู้ให้ข้อมูล สำหรับการวิเคราะห์ข้อมูลที่มีสมมติฐานว่าผู้ให้ข้อมูลมีชุดขององค์ความรู้ที่ต่างกัน รวมถึงปรับพารามิเตอร์ความลำเอียงในการเลือกตอบให้รองรับการเลือกตอบของผู้ให้ข้อมูลต่างกลุ่มวัฒนธรรมกัน ดังนี้

สัจพจน์ที่ 1 ความจริงเชิงวัฒนธรรม (Multiple Cultural Truth) MC-LTRM กำหนดให้มีค่าคุณลักษณะแฝงเชิงวัฒนธรรม (latent cultural truth) $V \geq 1$ โดย $\mathcal{T} = \{T_1, \dots, T_v, \dots, T_V\}$ เมื่อ $T_v \in \prod_{K=1}^{M}(-\infty, \infty)$ โดยผู้ประเมินแต่ละคนเป็นสมาชิกของกลุ่มวัฒนธรรมใดกลุ่มวัฒนธรรมหนึ่งเพียงกลุ่มเดียว (T_{Ω_i}) โดยพารามิเตอร์ $\boldsymbol{\Omega} = (\Omega_i)_{1xN}$ แทนกลุ่มวัฒนธรรมที่ผู้ประเมินแต่ละคนเป็นสมาชิก โดย $\Omega_i \in \{1, \dots, V\}$

สัจพจน์ที่ 2 ความลำเอียงของการตอบ (Multiple Cultural Ordinal Scale) ประกอบด้วยชุดของขอบเขตร่วมของสเกลแบบเรียงอันดับสำหรับกลุ่มวัฒนธรรมแต่ละกลุ่ม โดย $\boldsymbol{\mathcal{G}} = \{G_1, \dots, G_v, \dots, G_V\}$ เทรชโฮลด์ความลำเอียงของผู้ประเมินคำนวณจาก

$$\delta_{ic} = a_i \, \gamma_{\Omega_i c} + b_i \tag{2.31}$$

เมื่อ $A = (a_i)_{1xN}$ ในช่วง $(0, \infty)$ และ $B = (b_i)_{1xN}$ ในช่วง $(-\infty, \infty)$

สัจพจน์ที่ 1 ยอมให้มีความน่าจะเป็นของคำตอบมากกว่าหนึ่งชุดคำตอบ โดยมีเงื่อนไขว่าผู้ให้ข้อมูล i เป็นสมาชิกของกลุ่มวัฒนธรรมกลุ่มใดกลุ่มหนึ่งเพียงกลุ่มเดียว ยกตัวอย่างเช่น การให้คะแนนที่ผู้ประเมินเป็นผู้เชี่ยวชาญจากต่างสถาบันซึ่งอาจมีแนวทางในการตีความคุณลักษณะที่ต้องประเมินต่างกัน หรือตีความเกณฑ์การประเมินต่างกัน ในกรณีดังกล่าวอาจอนุมานได้ว่า คะแนนที่ได้จากผู้ประเมินต่างสถาบันอาจไม่ได้อยู่ภายใต้วัฒนธรรมร่วม (T_k) เพียงหนึ่งเดียวแต่เป็นกลุ่มพหุวัฒนธรรม

(T_{vk}) ในขณะที่สัจพจน์ที่ 2 เป็นการสร้างสเกลแบบเรียงอันดับสำหรับแต่ละกลุ่มวัฒนธรรม โดย G_V แทนแนวโน้มในการเลือกตอบของกลุ่มวัฒนธรรม v ซึ่งขอบเขตของสเกล γ_c ในการวิเคราะห์กลุ่มวัฒนธรรมเดี่ยวถูกแทนที่ด้วย γ_{vc} เมื่อ v เป็นค่าที่ได้จาก Ω_i ภาวะความน่าจะเป็น (Likelihood function) ของโมเดล MC-LTRM คำนวณจาก

$$L[X = (x_{ik})_{NxM} | \mathcal{T}, \Lambda, \mathcal{G}, E, A, B, \Omega] \\ = \prod_{i=1}^{N} \prod_{k=1}^{M} [F(\delta_{i,x_{ik}} | T_{\Omega_i k}, \tau_{ik}) - F(\delta_{i,x_{ik}-1} | T_{\Omega_i k}, \tau_{ik})] \tag{2.32}$$

เมื่อ $\delta_{i,0} = -\infty, \delta_{i,C} = \infty$

การประมาณค่าพารามิเตอร์ของ MC-LTRM กระทำภายใต้การวิเคราะห์ทางสถิติแบบเบส์ โดยการระบุโมเดลทางคณิตศาสตร์แบบลดหลั่น (Hierarchical Model Specification) โดยมีพารามิเตอร์ของโมเดล ดังนี้

$T_{vk} \sim Normal\ (\mu T_v, \tau T_v)$	ค่าประจำตำแหน่งของคำตอบ หรือคะแนนฉันทามติของการประเมิน
$\lambda_k \sim Gamma\ (\mu_\lambda^2\ \tau_\lambda, \mu_\lambda \tau_\lambda)$	ความยากของข้อคำถาม
$\gamma_{vc} \sim Normal\ (\mu_\gamma, \tau_\gamma)$	เทรชโฮลด์ร่วม
$E_i \sim Gamma\ (\mu_{E_{\Omega i}}^2, \tau_{E_{\Omega i}}, \mu_{E_{\Omega i}}, \tau_{E_{\Omega i}})$	ความสามารถของผู้ประเมิน
$a_i \sim Gamma\ (\mu_{a_{\Omega i}}^2, \tau_{a_{\Omega i}}, \mu_{a_{\Omega i}}, \tau_{a_{\Omega i}})$	scaling bias
$b_i \sim Normal\ (\mu_{b_{\Omega i}}, \tau_{b_{\Omega i}})$	shifting bias

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

โดยมีการกำหนดการแจกแจงพารามิเตอร์ ดังนี้

$\mu_T \sim Normal\ (0,0.1)$	$\tau_T \sim Gamma\ (1,0.1)$
$\mu_\lambda = 1$	$\tau_\lambda \sim Gamma\ (1,0.1)$
$\mu_\gamma = 0$	$\tau_\gamma = 0.1$
$\mu_E \sim Gamma\ (4,4)$	$\tau_E \sim Gamma\ (4,4)$
$\mu_a = 1$	$\tau_a \sim Gamma\ (1,0.1)$
$\mu_b = 0$	$\tau_b \sim Gamma\ (1,0.1)$

ซอฟท์แวร์ทางสถิติสำหรับการวิเคราะห์โมเดลฉันทามติเชิงวัฒนธรรม

การวิเคราะห์ฉันทามติเชิงวัฒนธรรมสามารถทำได้โดยซอฟท์แวร์สำหรับวิเคราะห์สถิติแบบเบส์ เช่น WinBUGS JAGS และ Stan นอกจากนี้ สถิติวิเคราะห์แบบเบส์ช่วยให้สามารถวิเคราะห์โมเดลที่ซับซ้อนได้ดีขึ้น โดยเฉพาะโมเดลแบบลำดับขั้น เช่น MC-GCM ซึ่งเป็นโมเดลที่วิเคราะห์ข้อมูลพหุวัฒนธรรม หรือโมเดลที่วิเคราะห์ตัวแปรร่วมโดยการวิเคราะห์การถดถอยของพารามิเตอร์ ในปัจจุบัน การประมาณค่าภาวะความควรจะเป็นสูงสุด(maximum likelihood) เพียงอย่างเดียวไม่เพียงพอต่อการประมาณค่าโมเดลฉันทามติเชิงวัฒนธรรมที่มีความซับซ้อน ดังนั้น การประมาณค่าพารามิเตอร์โดยกระบวนทัศน์ทางสถิติแบบเบส์จึงเหมาะสมมากกว่า ซอฟท์แวร์ในการประมาณค่าแบ่งออกเป็น 2 ลักษณะ คือ ซอฟท์แวร์สำหรับการวิเคราะห์โมเดลแบบทั่วไป และซอฟท์แวร์สำหรับวิเคราะห์โมเดลแบบลำดับขั้น

1. ซอฟท์แวร์สำหรับการวิเคราะห์โมเดลแบบทั่วไป ซอฟท์แวร์แรกได้รับการพัฒนาโดย Karabatos และ Batchelder (2003) ใช้ร่วมกับโปรแกรม S-PLUS สำหรับการประมาณค่าโมเดล GCM ภายใต้ข้อตกลงเบื้องต้นว่าข้อมูลต้องมีคำตอบฉันทามติเพียงคำตอบเดียว (single consensus answer key) จากนั้น Oravecz Vandekerckhove และ Batchelder (2014) ได้พัฒนาโปรแกรมที่ใช้ graphic user interface (GUI) ในการสั่งการ มีชื่อว่า Bayesian Cultural Consensus Toolbox (BCCT) ซึ่งเป็นโปรแกรมที่เขียนภายใต้โปรแกรม MATLAB และ JAGS ทั้งนี้ ผู้ใช้ไม่จำเป็นต้องมีความรู้เบื้องต้นเกี่ยวกับการเขียนคำสั่งทางสถิติแต่อย่างใด โดยผู้ใช้สามารถดาวน์โหลดโปรแกรมและคู่มือได้จากเว็บไซต์ https://git.psu.edu/ zzo1/BCCTToolbox.git. BCCTสามารถวิเคราะห์ข้อมูลที่ตัวอย่างมีความสามารถต่างกัน มีพารามิเตอร์ความลำเอียง และมีความยากของคำถามไม่เท่ากันได้ ผู้ใช้สามารถเลือกกำหนดค่าพารามิเตอร์บางค่าให้เป็นค่าคงที่ได้ การตรวจสอบความสอดคล้องของโมเดลใช้การเปรียบเทียบค่า Deviance Information Criterion (DIC) และ posterior predictive model check 2 ค่าในโปรแกรม คือ Bayesian p-value และ eigenvalue ratio test ผลการวิเคราะห์ข้อมูลจะอยู่ในรูปของกราฟที่ผู้ใช้สามารถเข้าใจได้ง่าย

2. ซอฟท์แวร์สำหรับวิเคราะห์โมเดลแบบลำดับขั้น โมเดลแบบลำดับขั้นถือว่าพารามิเตอร์ที่เกี่ยวข้องได้มาจากการสุ่มจากการแจกแจงแบบลำดับขั้น โดยพารามิเตอร์แต่ละค่ามีการแจกแจงก่อนหน้าของตัวเอง โมเดลแบบลำดับขั้นเหมาะสำหรับการวิเคราะห์ข้อมูลที่มีการร่วมตัวแปรร่วมและการวิเคราะห์การถดถอย เช่น การตอบคำถามว่า เพศ อายุ หรือระดับการศึกษาของผู้ประเมินมีผลต่อ

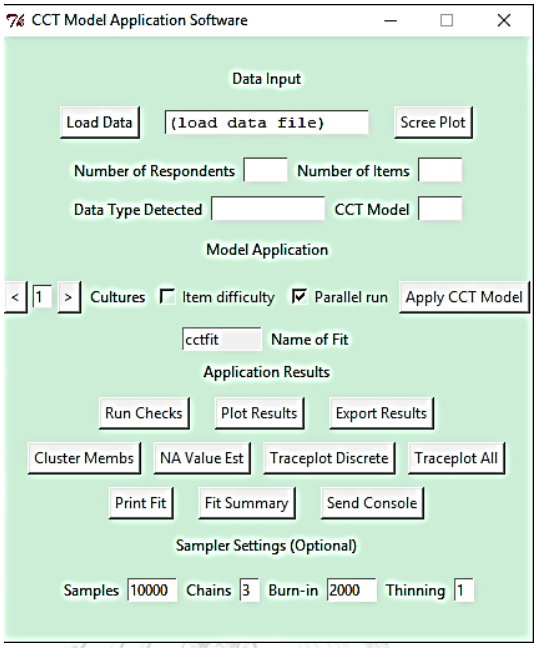
การให้คะแนนในการประเมินหรือไม่ ซอฟท์แวร์ที่ใช้สำหรับการวิเคราะห์ ได้แก่ Hierarchical Condorcet Modeling Toolbox (HCMT) เป็นซอฟท์แวร์แบบ stand alone สามารถดาวน์โหลดได้จาก https://git.psu.edu/zzo1/ HierarchicalCondorcetModelingToolbox.git HCMT เป็นซอฟท์แวร์ประเภท graphic user interface (GUI) เช่นเดียวกับ BCCT แต่ได้รับการพัฒนาให้พารามิเตอร์ในโมเดลมีลักษณะลำดับขั้นซึ่งใช้วิเคราะห์ตัวแปรร่วมได้ นอกจากนี้ยังสามารถวิเคราะห์โมเดล GCM ที่มีการให้เลือก “ไม่ทราบคำตอบ” ในกรณีของการตอบแบบสอบถามที่ผู้ให้ข้อมูลไม่ทราบคำตอบและไม่ต้องการเดาคำตอบซึ่งเป็นพารามิเตอร์ส่วนขยายจากโมเดล GCM ปกติ โดยมีตัวอย่างการวิเคราะห์ในการศึกษาของ Oravecz Anders และ Batchelder (2015)

3. ซอฟท์แวร์ในการวิเคราะห์โมเดลฉันทามติเชิงวัฒนธรรมที่ได้กล่าวมาก่อนหน้านี้ มีข้อจำกัดประการหนึ่งคือสามารถวิเคราะห์ได้เฉพาะข้อมูลที่เป็นกลุ่มวัฒนธรรมเดี่ยวตามข้อตกลงเบื้องต้นของโมเดล GCM $\mathrm{Z} = (\mathrm{Z_k})$ นั่นคือตัวอย่างต้องมีฉันทามติร่วมกันเพียงหนึ่งเดียว ซึ่งในปัจจุบันได้มีการพัฒนาโมเดลฉันทามติเชิงวัฒนธรรมที่อนุญาตให้ตัวอย่างมีฉันทามติจำแนกตามกลุ่มย่อย (subgroup) ได้ ดังนั้นจึงได้มีการพัฒนาแพ็คเกจชื่อ CCTpack (Anders, 2017) ซึ่งเป็นแพ็คเกจที่ทำงานร่วมกันโปรแกรม R สำหรับวิเคราะห์ฉันทามติเชิงวัฒนธรรม CCTpack สามารถใช้วิเคราะห์โมเดลฉันทามติเชิงวัฒนธรรมได้ทั้งโมเดลแบบวัฒนธรรมเดี่ยวและพหุวัฒนธรรมและใช้วิเคราะห์ได้กับข้อมูลทุกระดับและเป็นซอฟท์แวร์เดียวที่สามารถวิเคราะห์โมเดลทั้งหมดได้ในปัจจุบัน ผู้ใช้งาน CCTpack สามารถเขียนคำสั่งของโปรแกรม R ด้วยตนเอง หรือใช้คำสั่ง cctgui( ) ซึ่งแสดงหน้าจอแบบ graphic user interface ซึ่งผู้ใช้สามารถนำเข้าไฟล์ข้อมูลและเลือกกำหนดค่าต่าง ๆ ในการวิเคราะห์ได้ ดังรูป 2.10 ทั้งนี้ ผู้ใช้สามารถติดตั้งชุดคำสั่ง CCTpack ได้โดยการเรียกคำสั่ง install.package (“CCTpack”) ได้จากภายในโปรแกรมและเรียกใช้งานได้โดยคำสั่ง library (“CCTpack”) ผู้ใช้สามารถดาวน์โหลดคู่มือการใช้งานได้จากเว็บไซต์ https://cran.rproject.org/web/packages/ CCTpack/CCTpack.pdf (Anders, 2017)

ขั้นตอนในการวิเคราะห์ข้อมูลด้วยชุดคำสั่ง CCTpack ประกอบด้วย ขั้นตอน ดังนี้

1) นักวิจัยต้องเตรียมข้อมูลให้อยู่ในรูปของเมตริกซ์ NxM โดยแถว (row) คือผู้ประเมิน และหลัก (column) คือข้อคำถามหรือรายการประเมิน และแทนค่าสูญหายด้วย NA จากนั้นนำข้อมูลเข้าสู่โปรแกรม โดยโปรแกรมจะวิเคราะห์เมตริกซ์ข้อมูลดิบและวิเคราะห์ข้อมูลด้วยโมเดลที่เหมาะสม

กล่าวคือ หากข้อมูลเป็นการให้คะแนนแบบ [0,1] โปรแกรมจะวิเคราะห์ด้วยโมเดล GCM หากข้อมูลเป็นการให้คะแนนแบบเรียงอันดับจะวิเคราะห์ด้วยโมเดล LTRM เป็นต้น



รูป 2.10 หน้าต่างวิเคราะห์ข้อมูลจากชุดคำสั่ง CCTpack
(Anders, 2017)

2) กำหนดกลุ่มวัฒนธรรมตั้งต้น (Culture group) โดยนักวิจัยต้องทราบสมมติฐานเบื้องต้นเกี่ยวกับกลุ่มวัฒนธรรมว่ามีกลุ่มย่อย (cluster/subgroup) หรือไม่ นักวิจัยสามารถเปรียบเทียบจำนวนกลุ่มฉันทามติได้จากการวิเคราะห์ scree plot เพื่อดูจำนวนองค์ประกอบ นักวิจัยสามารถเลือกให้โปรแกรมวิเคราะห์ความยากของข้อคำถามในกรณีที่มีกลุ่มฉันทามติมากกว่า 1 กลุ่ม เพื่อดูความแตกต่างของการให้คะแนนระหว่างผู้ประเมินได้ เช่น ผู้ประเมินที่มีประสบการณ์ต่างกันอาจมีความถูกต้องแม่นยำในการให้คะแนนต่างกัน

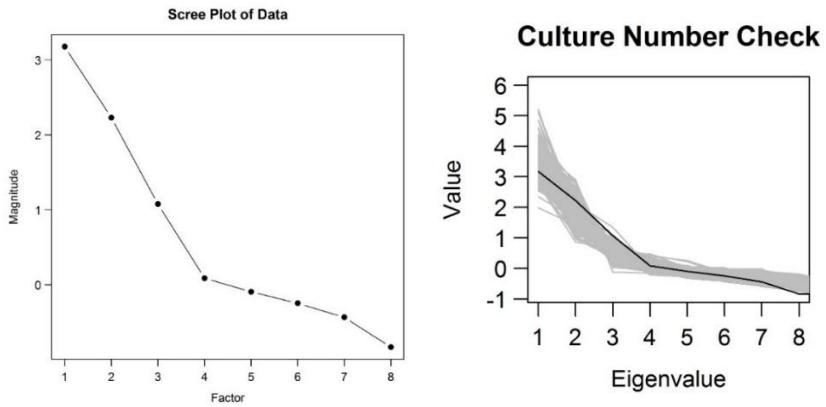
3) จากนั้น นักวิจัยเลือกคำสั่ง Apply CCT model หรือพิมพ์คำสั่ง cctfit() เพื่อวิเคราะห์ข้อมูล โดยคำสั่งเป็นดังนี้

```
cctfit<-cctapply(data=DRFtest3, clusters = 2, itemdiff = TRUE, samples = 500, chains = 3, burnin = 1000, plotr = TRUE)
```

จากคำสั่ง แสดงการวิเคราะห์ข้อมูลจากไฟล์ DRFtest3 ซึ่งเป็นไฟล์การประเมินงานเขียนที่ผู้วิจัยรวบรวม จากการวิเคราะห์ scree plot พบว่าค่าองค์ประกอบของตัวอย่างจำแนกออกเป็น 2

กลุ่ม ผู้วิจัยจึงกำหนดให้โมเดลวิเคราะห์แบบพหุวัฒนธรรมโดยใช้โมเดล MC-LTRM รวมถึงให้วิเคราะห์ความยากในการให้คะแนนเพื่อดูความแม่นยำของผู้ประเมินแต่ละกลุ่มด้วยคำสั่งย่อย itemdiff = TRUE

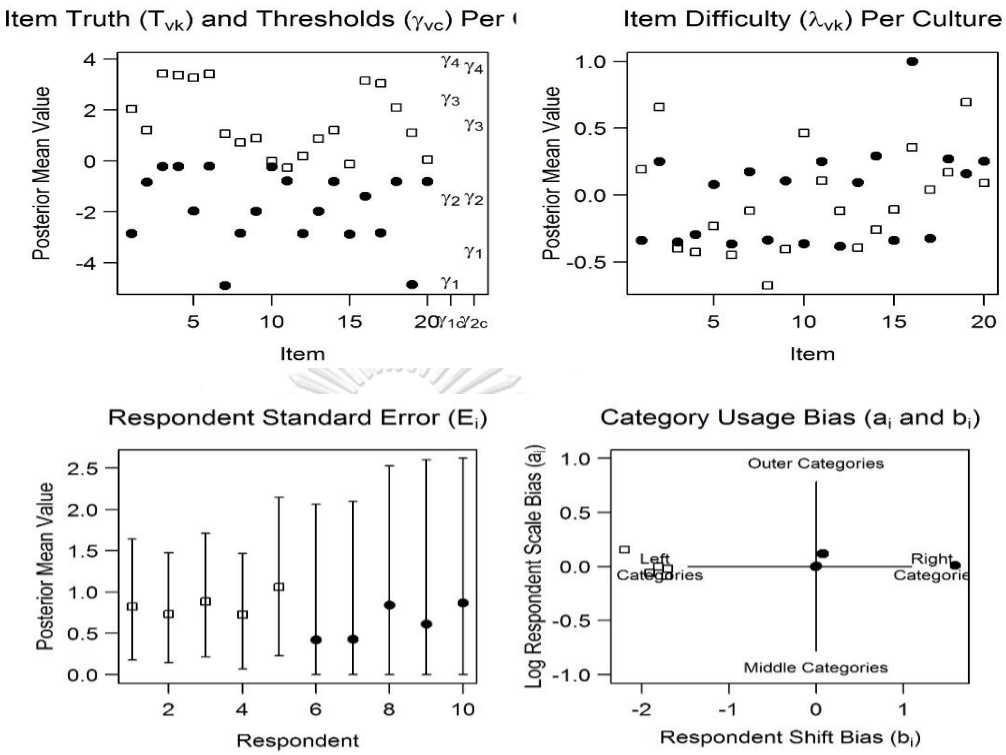
4) ตรวจสอบความสอดคล้องของโมเดล สิ่งสำคัญคือการตรวจสอบจำนวนกลุ่มฉันทามติ (cultural number check) ว่าตรงตามที่นักวิจัยระบุในขั้นตอนที่ 2 หรือไม่ หากกราฟแสดงว่าเส้นทึบเป็นแนวเดียวกันกับแถบสีเทาดังรูป 2.11 (ขวา) แสดงว่าโมเดลสอดคล้องกับข้อมูลเชิงประจักษ์ หากเส้นทึบไม่แนบสนิทกับแถบสีเทาแสดงว่าโมเดลยังไม่สอดคล้องกับข้อมูล นักวิจัยจำเป็นต้องระบุกลุ่มวัฒนธรรมใหม่อีกครั้งหนึ่ง ทั้งนี้ หากเส้นทึบไม่แนบสนิทกับแถบสีเทาบนกราฟ นักวิจัยอาจจำเป็นต้องพิจารณาค่า DIC เพื่อเลือกโมเดลที่มีค่า DIC ต่ำสุดในการวิเคราะห์ ต่อมาคือการตรวจสอบว่าข้อคำถามมีความยากเท่าเทียมกันหรือไม่ตามข้อตกลงเบื้องต้นของโมเดลเกี่ยวกับความเป็นเอกพันธ์ของข้อคำถาม โดยผู้ประเมินมีความน่าจะเป็นในการเลือกให้คะแนนระดับใด ๆ เท่ากับ $\mathbf{g_i = 1/2}$ การตรวจสอบความเป็นเอกพันธ์ของข้อคำถามพิจารณาจากค่า VDI (Batchelder และ Anders, 2012) อยู่ในช่วงเปอร์เซ็นไทล์ที่ 10 และ 90 ถือว่าข้อคำถามมีความยากเท่าเทียมกัน อย่างไรก็ตาม ในกรณีของการวิเคราะห์ข้อมูลพหุวัฒนธรรม นักวิจัยควรใช้การเปรียบเทียบค่า Deviance Information Criterion (DIC) เพื่อเลือกโมเดลที่เหมาะสมโดยพิจารณาจากโมเดลที่มีค่า DIC ต่ำกว่า



รูป 2.11 ผลการวิเคราะห์องค์ประกอบทางวัฒนธรรมของผู้ประเมิน

หลังจากโปรแกรมวิเคราะห์ข้อมูลเรียบร้อยแล้ว จะรายงานผลการวิเคราะห์ในรูปของกราฟดังรูป 2.12 จะเห็นได้ว่าโมเดลจำแนกผู้ประเมินออกเป็น 2 กลุ่ม (สี่เหลี่ยมสีขาวและจุดสีดำ) ซึ่งมีฉันทามติของคะแนนประเมินต่างกันดังรูป 2.12 (บนซ้าย) และมีความแม่นยำในการให้คะแนนต่างกัน

(บนขวา) นอกจากนี้ยังให้ผลการประมาณค่าพารามิเตอร์ความสามารถของผู้ประเมิน (E_i) ของผู้ประเมินทั้งสองกลุ่ม (ล่างซ้าย) และพารามิเตอร์ความลำเอียงในการให้คะแนน (a_i, b_i) (ล่างขวา)

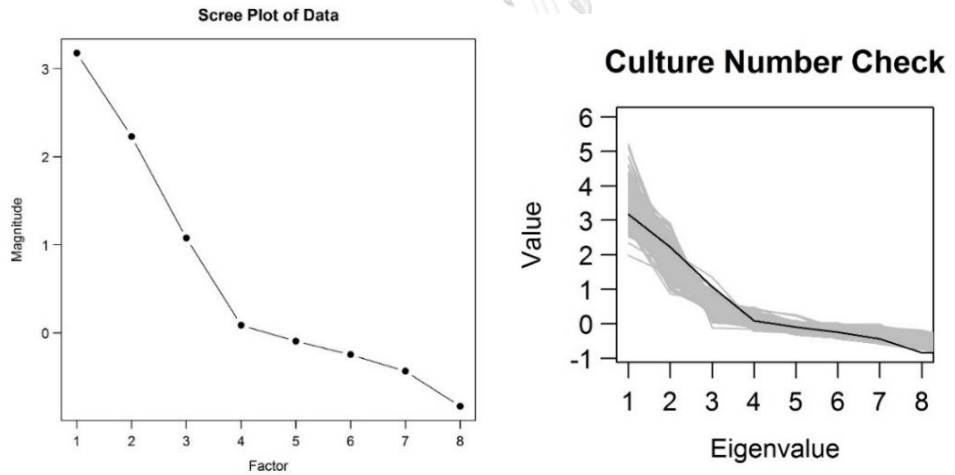


รูป 2.12 การแจกแจงความน่าจะเป็นภายหลังของค่าพารามิเตอร์ของโมเดล MC-LTRM

ระยะเวลาในการวิเคราะห์ข้อมูลโดยใช้ชุดคำสั่ง CCTpack ขึ้นอยู่กับปัจจัยหลายประการ ได้แก่ 1) ความซับซ้อนของโมเดล โมเดล CCT ที่ซับซ้อนกว่า เช่น LTRM หรือ MC-LTRM ที่มีการวิเคราะห์จำแนกกลุ่มฉันทามติจะใช้เวลานานกว่าโมเดลพื้นฐานอย่าง GCM 2) ขนาดของข้อมูลและการกำหนดค่าของลูกโซ่มาร์คอฟ หากข้อมูลมีขนาดใหญ่จะใช้เวลานานในการวิเคราะห์ ทั้งนี้นักวิจัยสามารถกำหนดค่าของกระบวนการ MCMC ให้ต่ำลงเพื่อลดเวลาในการวิเคราะห์ลงได้

ผู้วิจัยได้การศึกษาผลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมจากข้อมูลแบบเรียงอันดับด้วยโมเดล MC-LTRM ใช้ข้อมูลการประเมินการเขียนความเรียงของนักศึกษาระดับปริญญาตรี จำนวน 20 ฉบับ โดยผู้ประเมิน 5 คน การให้คะแนนใช้เกณฑ์การให้คะแนนแบบรูบริค จำนวน 5 ระดับ การวิเคราะห์ข้อมูลใช้ผลการประเมินคะแนนภาพรวมงานเขียนทั้งฉบับ คะแนนเต็ม 5 วิเคราะห์ข้อมูลภายใต้กรอบแนวคิดการวิเคราะห์ฉันทามติเชิงวัฒนธรรม (Cultural Consensus Theory) โดยใช้โมเดลสำหรับการวิเคราะห์ข้อมูลจากมาตรเรียงอันดับสำหรับข้อมูลพหุวัฒนธรรม (Multi-culture

Latent Truth Rater Model: MC-LTRM) การวิเคราะห์ข้อมูลทำโดยโปรแกรม R และ JAGS โดยสามารถวิเคราะห์ได้ 3 วิธี วิธีแรกคือการใช้ชุดคำสั่ง CCTpack (Anders, 2017) ในการวิเคราะห์ วิธีที่สอง นักวิจัยสามารถเรียกใช้คำสั่ง Graphic User Interface ในชุดคำสั่ง CCTpack โดยพิมพ์คำสั่ง cctgui() วิธีที่สามคือการเขียนคำสั่งในการวิเคราะห์เอง ซึ่งนักวิจัยจะสามารถปรับค่าการแจกแจงของพารามิเตอร์ต่าง ๆ ได้ตามความเหมาะสมมากกว่าการใช้ชุดคำสั่งสำเร็จรูป คำสั่งที่เขียนขึ้นนี้จะทำงานร่วมกับคำสั่ง rjags (Plumer, 2012; Anders และ Batchelder, 2015) ผลการตรวจสอบความสอดคล้องของโมเดล จากข้อมูลการตรวจให้คะแนนการเขียนความเรียงของนักศึกษาระดับปริญญาตรีที่นำมาทดลองใช้ในการวิเคราะห์ข้อมูล พบว่า ข้อมูลที่นำมาวิเคราะห์แสดงถึงกลุ่มวัฒนธรรมที่ต่างกันระหว่างผู้ประเมิน โดยผลการวิเคราะห์แสดงการแบ่งกลุ่มผู้ประเมินออกเป็นสองกลุ่ม ดังแสดงในรูป scree plot องค์ประกอบทางวัฒนธรรมของผู้ประเมิน ดังรูปที่ 2.13 (ซ้าย) สอดคล้องกับผลการคำนวณค่าไอเกนขององค์ประกอบทางวัฒนธรรมทางด้านขวา แสดงให้เห็นว่าผู้ประเมินมีวัฒนธรรมในการตรวจให้คะแนนงานเขียนต่างกัน ดังนั้นโมเดลที่เหมาะสมในการวิเคราะห์ข้อมูลชุดนี้คือโมเดล MC-LTRM



รูป 2.13 ผลการวิเคราะห์องค์ประกอบทางวัฒนธรรมของผู้ประเมิน

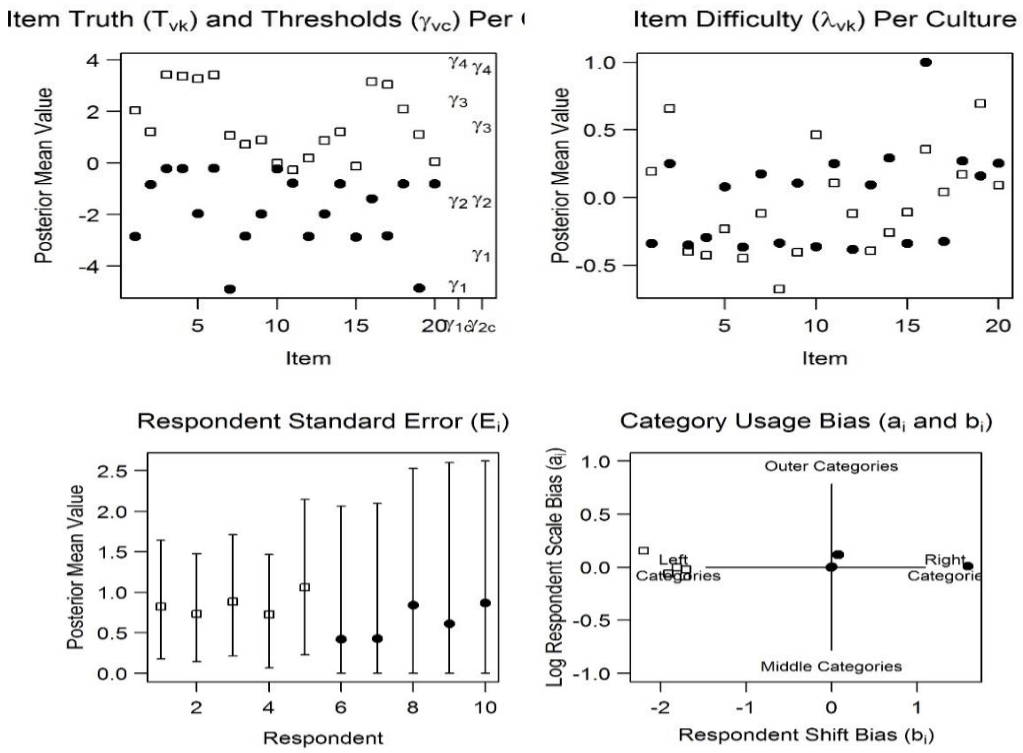
จากนั้น ตรวจสอบความเท่ากันของพารามิเตอร์ความยากของข้อคำถามในการวิเคราะห์ ในกรณีของข้อมูลที่มีหนึ่งกลุ่มวัฒนธรรม นักวิจัยสามารถใช้ผลการวิเคราะห์ค่าสถิติ VDI เพื่อพิจารณาความเท่ากันของพารามิเตอร์ความยากของการตรวจให้คะแนนในการวิเคราะห์ แต่สำหรับโมเดลพหุวัฒนธรรม Anders และ Batchelder (2015) แนะนำให้ใช้การเปรียบเทียบค่า DIC ระหว่างการวิเคราะห์ข้อคำถามที่มีความยากเท่ากัน ($\lambda_k = 1$) และแตกต่างกัน ($\lambda_k \neq 1$) ดังตาราง 2.6

ตาราง 2.6 ผลการวิเคราะห์ค่าไอเกนกลุ่มวัฒนธรรมของผู้ประเมิน โมเดล MC-LTRM

Eigenvalue	Deviance Information Criterion	
3.178	$DIC_{MC-LTRM_{V=2}^{\lambda k=1}}$ 472.61	$DIC_{MC-LTRM_{V=2}^{\lambda k\neq 1}}$ 242.05
2.232		
1.077		
0.088		

จากตาราง 2.6 แสดงให้เห็นว่าโมเดลที่มีข้อคำถามมีความยากเท่ากัน ($MC-LTRM_{V=2}^{\lambda k=1}$) มีค่า DIC สูงกว่าโมเดลที่มีค่าคำถามไม่เท่ากัน ($MC-LTRM_{V=2}^{\lambda k\neq 1}$) ดังนั้น โมเดลที่เหมาะสมในการวิเคราะห์ข้อมูลจึงได้แก่โมเดล $MC-LTRM_{V=2}^{\lambda k\neq 1}$

ผลการวิเคราะห์การทำหน้าที่ต่างกันของผู้ประเมิน พบว่า การแจกแจงความน่าจะเป็นภายหลังของคะแนนฉันทามติของผู้ตรวจ (T_{vk}) และความยากของการให้คะแนนเรียงความแต่ละฉบับ (λ_{vk}) โดยการแจกแจงความน่าจะเป็นภายหลังของพารามิเตอร์คะแนนฉันทามติจากผู้ประเมินกลุ่มที่ 1 มีค่าระหว่าง -4.889 ถึง -0.205 ผู้ประเมินกลุ่มที่ 2 มีค่าระหว่าง -0.263 ถึง 3.422 ค่าสัมประสิทธิ์สหสัมพันธ์ (R_{xy}) ระหว่างทั้งสองกลุ่มเท่ากับ 2.30 แสดงให้เห็นว่าคะแนนฉันทามติของทั้ง 2 กลุ่มมีความสัมพันธ์กันในระดับต่ำการแจกแจงความน่าจะเป็นภายหลังของพารามิเตอร์ความยากในการประเมินงานเขียนรายข้อของผู้กลุ่มที่ 1 มีค่าระหว่าง –0.383 ถึง 0.999 และความยากในการประเมินงานเขียนรายข้อของผู้ประเมินกลุ่มที่ 2 มีค่าระหว่าง -0.677 ถึง 0.693 ในรูปที่ 2.14 ภาพขวาแสดงการแจกแจงความน่าจะเป็นภายหลังของพารามิเตอร์ความยากในการประเมินงานเขียนแต่ละฉบับ (λ_{vk}) พบว่างานเขียนฉบับที่ 16 เป็นงานเขียนที่ประเมินได้ยากสำหรับผู้ประเมินกลุ่มที่ 1 ในขณะที่งานเขียนฉบับที่ 19 เป็นงานเขียนที่ประเมินยากสำหรับผู้ประเมินกลุ่มที่ 2 จากรูปที่ 2. ภาพซ้ายจะเห็นว่าผู้ประเมินกลุ่มที่ 1 มีแนวโน้มที่จะประเมินในระดับคะแนนที่ต่ำกว่า และให้คะแนนในช่วงเทรชโฮลด์ที่ 1 และ 2 ในขณะที่ผู้ประเมินกลุ่มที่ 2 ประเมินในระดับคะแนนที่สูงกว่า โดยให้คะแนนในช่วงเทรชโฮลด์ที่ 3 และ 4 ทั้งนี้ ค่าเฉลี่ยการแจกแจงความน่าจะเป็นภายหลังของเทรชโฮลด์ผู้ประเมินกลุ่มที่ 1 มีค่าระหว่าง -4.784 ถึง 3.903 ในขณะที่เทรชโฮลด์ของผู้ประเมินกลุ่มที่ 2 มีค่าระหว่าง –3.580 ถึง 3.658 โดยมีขอบเขตบนและล่างของแต่ละเทรชโฮลด์ของแต่ละกลุ่ม



รูป 2.14 การแจกแจงความน่าจะเป็นภายหลังของค่าพารามิเตอร์ของโมเดล MC-LTRM

ความสามารถของผู้ให้ข้อมูลแสดงด้วยพารามิเตอร์ E_i มีค่าระหว่าง 0.65 ถึง 2.27 แบ่งออกเป็น 2 กลุ่ม โดยกลุ่มที่ 1 ประกอบด้วยผู้ประเมิน 5 คน และกลุ่มที่ 2 ประกอบด้วยผู้ประเมิน 5 คน ดังแสดงในรูปที่ 2.13 (ล่างซ้าย) โดยค่าเฉลี่ยการแจกแจงความน่าจะเป็นภายหลังของผู้ประเมินกลุ่มที่ 1 อยู่ระหว่าง 0.422 ถึง 0.869 กลุ่มที่ 2 อยู่ระหว่าง 0.884 ถึง 1.061 ค่าเฉลี่ยการแจกแจงความน่าจะเป็นภายหลังของพารามิเตอร์ความสามารถ (μ_E) และความแม่นยำ (τ_E) ของผู้ประเมินกลุ่มที่ 1 และผู้ประเมินกลุ่มที่ 2 มีค่า μ_E = [-4.515, -0.259] และ τ_E = [7.657, 31.892] แสดงให้เห็นว่าในภาพรวมผู้ประเมินกลุ่ม 2 มีความสามารถสูงกว่าผู้ประเมินกลุ่มที่ 1 แต่ในขณะเดียวกันก็มีความแปรปรวนสูงกว่าเช่นกัน

รูปที่ 2.14 (ล่างซ้าย) แสดงความสามารถของผู้ประเมิน ($\mathrm{E_i}$) 2 กลุ่ม โดยกลุ่มที่ 1 ประกอบด้วยผู้ประเมิน 5 คน และกลุ่มที่ 2 ประกอบด้วยผู้ประเมิน 5 คน ค่าเฉลี่ยการแจกแจงความน่าจะเป็นภายหลังของผู้ประเมินกลุ่มที่ 1 อยู่ระหว่าง 0.884 ถึง 1.061 กลุ่มที่ 2 อยู่ระหว่าง 0.422 ถึง 0.869 แสดงให้เห็นว่าในภาพรวมผู้ประเมินกลุ่ม 1 มีความสามารถต่ำกว่าผู้ประเมินกลุ่มที่ 2 เนื่องจากมีความคลาดเคลื่อนในการประเมิน (error) สูงกว่า ภาพล่างขวาแสดงรูปแบบความลำเอียงในการประเมินของผู้ประเมินทั้งสองกลุ่ม จะเห็นได้จากพารามิเตอร์ a_i ว่าผู้ประเมินกลุ่ม

ที่ 1 มีความไวในการเปลี่ยนระดับการประเมินต่ำเช่นเดียวกับผู้ประเมินกลุ่มที่ 2 อย่างไรก็ตาม เมื่อพิจารณาพารามิเตอร์ $\mathbf{b_i}$ จะเห็นว่าผู้ประเมินกลุ่มที่ 2 มีแนวโน้มที่จะเปลี่ยนระดับการประเมินไปทางบวก หรือมีแนวโน้มที่จะปล่อยคะแนน ในขณะที่ผู้ประเมินกลุ่มที่ 1 มีแนวโน้มที่จะเปลี่ยนระดับการประเมินไปทางลบ หรือมีแนวโน้มที่จะกดคะแนน

การศึกษาครั้งนี้ พบว่า สารสนเทศที่ได้จากการวิเคราะห์วิเคราะห์การทำหน้าที่ต่างกันของผู้ประเมินประกอบด้วยพารามิเตอร์ฉันทามติเชิงวัฒนธรรม ซึ่งสะท้อนระดับคะแนนที่เป็นฉันทามติของผู้ประเมินทั้งกลุ่ม โดยพิจารณาจากค่าฉันทามติเชิงวัฒนธรรม ($\mathrm{T_{vk}}$) ที่อยู่ในช่วงของเทรชโฮลด์ของสเกล (γ_{vc}) พารามิเตอร์ความยาก (λ_{vk}) ของข้อคำถาม แสดงถึงระดับความยากในการให้คะแนนแต่ละคุณลักษณะซึ่งส่งผลต่อความแม่นยำในการประเมินการให้คะแนนของผู้ประเมินแต่ละคน พารามิเตอร์ความสามารถของผู้ประเมิน (E_i) สะท้อนความแม่นยำและความน่าเชื่อถือของผู้ประเมินแต่ละคน และพารามิเตอร์ความลำเอียง (a_i, b_i) แสดงแนวโน้มในการเลือกประเมินในระดับคะแนนต่าง ๆ บนมาตรประมาณค่า เช่น ผู้ประเมินบางคนมีแนวโน้มที่จะประเมินในระดับกลางเสมอ ในขณะที่ผู้ประเมินบางคนมีแนวโน้มจะประเมินในระดับคะแนนสุดโต่งทางด้านซ้ายหรือขวาของสเกล นอกจากนี้ ผลการวิเคราะห์ข้อมูลแสดงให้เห็นถึงมุมมองอีกมุมมองหนึ่งในการตอบปัญหาเกี่ยวกับคุณภาพของผู้ประเมินที่ส่งผลต่อคะแนนที่ได้จากการประเมิน หากนักวิจัยนำมุมมองนี้มาใช้ในการตรวจสอบคุณภาพของผู้ประเมินร่วมกับการตรวจสอบด้วยวิธีดั้งเดิมอาจช่วยตอบคำถามเกี่ยวกับคุณภาพของการประเมินและการคัดเลือกผู้ประเมินได้อีกทางหนึ่ง

การวิเคราะห์ข้อมูลเกี่ยวกับฉันทามติเชิงวัฒนธรรมภายใต้ทฤษฎี Cultural Consensus Theory ในช่วงแรกนั้นมีเพียงโมเดลที่สนับสนุนข้อมูล 2 ค่า แบบ (0,1) เท่านั้น ต่อมาได้มีการพัฒนาและแก้ไขสถิติที่ใช้กับโมเดลให้สามารถวิเคราะห์ข้อมูลแบบมาตรประมาณค่า แบบเรียงลำดับ และแบบต่อเนื่องได้ เพื่อให้ครอบคลุมลักษณะของข้อมูลมากยิ่งขึ้น การพัฒนาโมเดลภายใต้ทฤษฎี Cultural Consensus Theory นี้ส่งผลให้เกิดการศึกษาวิจัยที่เกี่ยวกับการพัฒนาโมเดล และการประยุกต์ใช้โมเดลในการศึกษาในสาขาอื่น ๆ จำนวนมาก

การศึกษาทฤษฎีการวิเคราะห์ฉันทามติเชิงวัฒนธรรมข้างต้นให้ข้อสรุปว่า การวิเคราะห์ฉันทามติเชิงวัฒนธรรมเป็นวิธีการที่ให้สารสนเทศเกี่ยวกับรูปแบบและพฤติกรรมของการเลือกตอบในคำถามที่ไม่มีการกำหนดคำตอบไว้ล่วงหน้า เช่น การสำรวจความคิดเห็น หรือการสำรวจทางสังคมศาสตร์ เนื่องจากมีพารามิเตอร์ที่ประมาณค่าคำตอบที่เป็นฉันทามติของกลุ่มผู้ตอบที่เป็นสมาชิก

หรือส่วนหนึ่งของกลุ่มวัฒนธรรมหรือความเชื่อเดียวกัน พารามิเตอร์ที่สะท้อนความเชี่ยวชาญหรือความน่าเชื่อถือของผู้ตอบ รวมถึงพารามิเตอร์ที่บ่งชี้ปัจจัยร่วมที่ส่งผลต่อการเลือกคำตอบ ได้แก่ ความลำเอียง ความยากของคำถาม แนวโน้มการเดา การจำแนกกลุ่มฉันทามติ เมื่อนำมาเปรียบเทียบกับตัวแปรที่เกี่ยวข้องกับความเที่ยงระหว่างผู้ประเมิน อาทิ ความสามารถของผู้ประเมิน ความลำเอียงของผู้ประเมิน และการทำหน้าที่ต่างกันของผู้ประเมิน จะเห็นได้ว่า พารามิเตอร์ของโมเดลเหล่านี้สามารถสะท้อนค่าของตัวแปรดังกล่าวได้เช่นกัน นอกจากนี้ จุดเด่นที่สำคัญของการวิเคราะห์ฉันทามติเชิงวัฒนธรรม คือ การแยกตัวแปรแฝง “สมรรถภาพ” ออกจาก “ความรู้ความเชี่ยวชาญ” ทั้งนี้เนื่องจากทฤษฎีฉันทามติเชิงวัฒนธรรมถือว่าหลักฐานที่แสดงถึงสมรรถภาพ (เช่น การตอบคำตอบได้ถูกต้อง) นั้นอาจเป็นผลมาจากการเดาส่วนหนึ่ง เช่นเดียวกับการวิเคราะห์ความเที่ยงระหว่างผู้ประเมินและการทำหน้าที่ต่างกันของผู้ประเมินที่แสดงให้เห็นว่าผู้ประเมินที่มีความรู้เท่ากัน หากมีความลำเอียง หรือเกิดความคลาดเคลื่อนในการประเมินเมื่อเวลาผ่านไป ก็ทำให้ผลการประเมินระหว่างผู้ประเมินไม่คงที่ได้ จุดเด่นอีกประการหนึ่งของโมเดลฉันทามติเชิงวัฒนธรรม คือ เป็นโมเดลที่ออกแบบมาสำหรับการวิเคราะห์ตัวอย่างจำนวนน้อย ซึ่งสอดคล้องกับบริบทของการประเมินขนาดเล็ก หรือการประเมินภายในสถานศึกษาที่ใช้ผู้ประเมินจำนวนไม่มากซึ่งไม่สามารถวิเคราะห์ด้วยโมเดลการวิเคราะห์ เช่น โมเดลการตอบสนองข้อสอบ หรือราสช์โมเดลได้ แต่ยังคงต้องการสารสนเทศเกี่ยวกับผู้ประเมินดังกล่าว ดังนั้น ผู้วิจัยจึงสนใจศึกษาการพัฒนากระบวนการวิเคราะห์ความเที่ยงระหว่างผู้ประเมินและการทำหน้าที่ต่างกันของผู้ประเมินโดยใช้โมเดลฉันทามติเชิงวัฒนธรรม รวมถึงศึกษาประสิทธิภาพในการประมาณค่าของโมเดลเมื่อนำมาใช้กับบริบทในการวิเคราะห์ความเที่ยงระหว่างผู้ประเมินดังกล่าว

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

ตอนที่ 4 การศึกษาความสอดคล้องในแนวเดียวกัน

การศึกษาความสอดคล้องในแนวเดียวกันเริ่มมาจากการปฏิรูปการศึกษาในประเทศสหรัฐอเมริกาตั้งแต่ช่วงปี ค.ศ. 1900 เป็นต้นมา สหรัฐอเมริกาได้ออกกฎหมายเกี่ยวกับการศึกษาที่เน้นการจัดระบบการศึกษาที่สอดคล้องกันโดยเฉพาะด้านการวัดและประเมินผลการศึกษาของผู้เรียนที่ต้องสอดคล้องกับมาตรฐานการจัดการเรียนรู้

ในปี ค.ศ. 1997 Norman Webb นักการศึกษาของสหรัฐอเมริกาได้เสนอวิธีการและเกณฑ์การประเมินความสอดคล้องเสนอต่อ National Institute for Science Education ซึ่งมีเกณฑ์การประเมินความสอดคล้อง 5 ประเด็น ได้แก่ 1) Content focus คือ การมุ่งเน้นเนื้อหาใน

การจัดการเรียนการสอน 2) Pedagogical implication คือ วิธีการจัดการเรียนการสอน 3) Equity คือ ความเสมอภาคของผู้เรียน 4) Articulation across graded คือ ความเชื่อมโยงระหว่างเนื้อหานแต่ละระดับชั้น และ 5) System applicability คือ การสามารถนำความรู้ที่ได้รับไปใช้ได้อย่างเป็นระบบ ในปี ค.ศ. 1998 วิธีการประเมินความสอดคล้องของ Webb ได้รับการนำไปใช้ในการประเมินทางการศึกษาใน 4 รัฐ ของสหรัฐอเมริกา และได้มีการปรับปรุงวิธีการประเมินให้เหมาะสมมากขึ้นเพื่อนำไปใช้ในการประเมินในรัฐอื่น ๆ ต่อไป การพัฒนาวิธีการประเมินของ Webb นั้นได้รับการสนับสนุนจากหน่วยงานหลายฝ่าย ไม่ว่าจะเป็น Council of Chief State School Officers (CCSSO) และ National Science Foundation (NSF) ในปี ค.ศ. 2004 Webb ได้พัฒนาเครื่องมือการประเมินความสอดคล้องสำหรับการใช้ประเมินผ่านทางเว็บไซต์มีชื่อว่า Web Alignment Tool (WAT) และได้มีการสาธิตการใช้งานไปเมื่อปี ค.ศ. 2005

นอกจากการพัฒนาวิธีการศึกษาความสอดคล้องในแนวเดียวกันของ Webb แล้ว ยังมีวิธีการศึกษาความสอดคล้องในแนวเดียวกันที่เสนอโดย Porter ในปี ค.ศ. 2002 มีชื่อว่า Survey of the Enacted Curriculum หรือ SEC วิธีการของ Porter ใช้การวิเคราะห์เมตริกซ์ความสอดคล้องเนื้อหากับระดับความซับซ้อนทางปัญญาโดยนำเสนอในรูปแบบของเมตริกซ์

จากการศึกษาของ Web (1997) และ Porter (2002) แสดงให้เห็นความสำคัญและจำเป็นของการศึกษาความสอดคล้องในแนวเดียวกันทางการศึกษา ซึ่งทั้งสองได้เสนอว่าการศึกษาความสอดคล้องเป็นสิ่งเดียวกับการศึกษาเกี่ยวกับความตรง (validity) ทั้งนี้ ผลที่ได้จากการประเมินความสอดคล้องระหว่างการวัดและประเมินผลกับมาตรฐานหลักสูตรจะเป็นหลักฐานที่แสดงถึงความตรงของการประเมิน (Evidence of assessment’s validity)

ผู้วิจัยได้ศึกษาเอกสารงานวิจัยที่เกี่ยวข้องกับการศึกษาความสอดคล้องในแนวเดียวกัน และได้แบ่งผลการสังเคราะห์เอกสารออกเป็น 2 หัวข้อ ได้แก่ 1) ความหมายและความสำคัญของการศึกษาความสอดคล้องในแนวเดียวกัน และ 2) วิธีการศึกษาความสอดคล้องในแนวเดียวกัน โดยมีรายละเอียดดังต่อไปนี้

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

4.1 ความหมายและความสำคัญของการศึกษาความสอดคล้องในแนวเดียวกัน

ความสอดคล้องในแนวเดียวกัน มาจากคำว่า “alignment” ตามความหมายในพจนานุกรม หมายถึง การจัดวางของสองสิ่งในตำแหน่งเส้นตรงเดียวกันหรือขนานกัน (Cambridge Dictionary, online) การจัดกิจกรรมหรือระบบเพื่อให้เข้ากันหรือสอดคล้องไปด้วยกัน (Macmillan Dictionary, online)

ความสอดคล้องในแนวเดียวกันทางการศึกษา หมายถึง ระดับความสอดคล้องของส่วนประกอบต่าง ๆ ของระบบการศึกษา อันได้แก่ มาตรฐาน หลักสูตร การประเมิน และการเรียนการสอน ดำเนินไปในทางเดียวกันเพื่อมุ่งสู่เป้าหมายหลักทางการศึกษา (Ananda, 2003; Resnick, Rothman, Slattery, and Vranek, 2003; Webb, 1997b) เช่นเดียวกับนิยามของ Case, Jorgensen และ Zucker (2004) ที่กล่าวว่า ความสอดคล้องในแนวเดียวกันทางการศึกษา หมายถึง ระดับของการทำงานประสานกันขององค์ประกอบในระบบการศึกษา อาทิ มาตรฐาน หลักสูตร การประเมิน และการเรียนการสอน เพื่อมุ่งสู่เป้าหมายที่ตั้งไว้

Webb (1997a; 1999) ได้ให้ความหมายเพิ่มเติมว่า การศึกษาความสอดคล้องในแนวเดียวกันเป็นการตรวจสอบว่ามาตรฐานการศึกษากับการวัดและประเมินผลสะท้อนถึงเนื้อหาบทเรียนเดียวกัน ซึ่งสอดคล้องกับ La Marca (2001) ที่ให้ความหมายว่า ระดับความสอดคล้องระหว่างเนื้อหาของการวัดและประเมินผลกับเนื้อหาที่ระบุในมาตรฐานการเรียนรู้

เว็บไซต์อภิทานศัพท์ทางการศึกษา (www.edglossary.org) กล่าวถึงการนิยามความหมายของความสอดคล้องในแนวเดียวกันว่า หากกล่าวถึง ‘ความสอดคล้องในแนวเดียวกัน’ โดยไม่มีบริบทที่เกี่ยวข้องอาจไม่สามารถระบุความหมายที่แท้จริงของความสอดคล้องในแนวเดียวกันได้ เนื่องจากคำจำกัดความของคำดังกล่าวแตกต่างไปตามระดับของผู้เกี่ยวข้อง อาทิ ความสอดคล้องในแนวเดียวกันในเชิงนโยบาย มีความหมายถึงกระบวนการตรวจสอบทบทวนรายละเอียดและความชัดเจนระหว่างการพัฒนาข้อกฎหมายที่เกี่ยวกับการศึกษา รวมถึงการนำนโยบายทางการศึกษาไปปฏิบัติ ในทางยุทธศาสตร์ ความสอดคล้องในแนวเดียวกัน หมายถึง แผนปฏิบัติการที่โรงเรียนใช้ในการจัดการเรียนการสอนเพื่อให้บรรลุเป้าหมายของแผนการพัฒนา ในขณะเดียวกันก็ให้ความสำคัญกับกระบวนการทำงานร่วมกันด้วยความสอดคล้องและมีประสิทธิภาพ ในแง่ของการวัดและประเมินผล ความสอดคล้องในแนวเดียวกัน หมายถึง ความเป็นอันหนึ่งอันเดียวกันระหว่างการวัดและประเมินผล มาตรฐานการเรียนรู้ และวิธีการจัดการเรียนการสอน เพื่อที่จะสามารถวัดผลผู้เรียนได้ตรงกับมโน

ทัศน์และทักษะที่ระบุในมาตรฐานการเรียนรู้ของแต่ละระดับชั้น ขณะที่ความสอดคล้องในแนวเดียวกันในมุมมองของหลักสูตร หมายถึง การจัดองค์ความรู้ ทักษะ หัวข้อและมโนทัศน์สำหรับผู้เรียน รวมถึงการจัดบทเรียน หน่วยการเรียนรู้ ใบงานและบทอ่าน รวมทั้งสื่อการเรียนการสอนให้สอดคล้องกับมาตรฐานการเรียนรู้ที่กำหนด รวมไปถึงการจัดหลักสูตรการเรียนการสอนของทุกรายวิชาให้เชื่อมโยงและสอดคล้องกัน

จากนิยามข้างต้นจะเห็นได้ว่า การศึกษาความสอดคล้องในแนวเดียวกัน ไม่ได้มีความหมายเพียงแค่มุมมองทางด้านการวัดและประเมินผลเพียงอย่างเดียว แต่ยังรวมถึงการให้ความสำคัญกับความเชื่อมโยงกันของการจัดการศึกษาทั้งระบบ จึงอาจสรุปได้ว่า การศึกษาความสอดคล้องในแนวเดียวกัน หมายถึง การศึกษาเกี่ยวกับระดับความสัมพันธ์ระหว่างองค์ประกอบทั้งหมดของระบบการศึกษาในเชิงนโยบายและการปฏิบัติ ได้แก่ เป้าหมายของหลักสูตร มาตรฐานการเรียนรู้ วิธีการจัดการเรียนการสอน เนื้อหา กิจกรรม การวัดและประเมินผล ซึ่งทั้งหมดนี้ต้องสอดคล้องกับวัยและศักยภาพของผู้เรียนด้วย

ความสำคัญของการศึกษาความสอดคล้องในแนวเดียวกันสะท้อนในนิยามที่กล่าวไปแล้วข้างต้น กล่าวคือ ความสอดคล้องในแนวเดียวกันเป็นสิ่งสำคัญและจำเป็นสำหรับการจัดการศึกษาแบบอิงมาตรฐาน เพราะเป็นเหมือนกับไม้บรรทัดที่กำกับให้การจัดการศึกษาทั้งระบบเป็นไปในทิศทางเดียวกัน ความสอดคล้องขององค์ประกอบต่าง ๆ ของระบบการศึกษาจะช่วยอำนวยความสะดวกให้ทั้งครูและผู้เรียนพัฒนาการเรียนรู้ได้อย่างมีประสิทธิภาพและมีประสิทธิผลดียิ่งขึ้น (Webb, 1997) ดังนั้น การศึกษาความสอดคล้องในแนวเดียวกันจึงมีความสำคัญ ดังที่ สังวรณ์ งัดกระโทก (2555) ได้กล่าวถึงความสำคัญของการศึกษาความสอดคล้องในแนวเดียวกันไว้ว่า

1) ทำให้ระบบการศึกษาทั้งหลายมุ่งไปสู่เป้าหมายเดียวกัน ผู้เกี่ยวข้องในแต่ละระบบมีความเข้าใจตรงกัน สามารถขับเคลื่อนการจัดการศึกษาให้สอดคล้องกันจนบรรลุถึงเป้าหมายเดียวกัน

2) ช่วยลดการประเมินอื่น ๆ ที่ซ้ำซ้อน

3) ช่วยในการวางแผนและปรับปรุงที่เกี่ยวกับการจัดการศึกษาให้มีความสอดคล้องกัน เช่น การพัฒนาครู การเลือกสื่อและตำราเรียน

กล่าวโดยสรุป การศึกษาความสอดคล้องในแนวเดียวกันเป็นการประเมินความน่าเชื่อถือของการจัดการศึกษาแบบอิงมาตรฐานซึ่งมีบทบาทสำคัญในการปฏิรูปการศึกษา การประเมินดังกล่าวทำให้มาตรฐานหลักสูตรเป็นรูปเป็นร่างและสร้างแรงกระตุ้นให้ผู้สอนสอนเนื้อหาที่ผู้เรียนสามารถนำไปใช้ได้อย่างแท้จริง

4.2 วิธีการศึกษาความสอดคล้องในแนวเดียวกัน

วิธีการศึกษาความสอดคล้องในแนวเดียวกันมีอยู่ด้วยกัน 3 วิธี ได้แก่ 1) Webb Model ของ Webb 2) The Survey of Enacted Curriculum Model (SEC) ของ Porter และ 3) Achieve Model ของบริษัท Achieve โดยมีรายละเอียดของแต่ละวิธีการ ดังต่อไปนี้

4.2.1 Webb Model

Webb Model เป็นวิธีการตรวจสอบความสอดคล้องที่ได้รับความนิยมมากที่สุดในการศึกษาความสอดคล้องในแนวเดียวกันทางการศึกษา การศึกษาความสอดคล้องของ Webb Model ใช้การลงรหัสของผู้เชี่ยวชาญและการจับคู่ความสอดคล้องระหว่างเนื้อหาบทเรียนที่อยู่ในมาตรฐานการเรียนรู้กับเนื้อหาที่ระบุในเครื่องมือการประเมิน เกณฑ์การพิจารณาความสอดคล้องของ Webb ประกอบด้วย 4 เกณฑ์ ได้แก่

1) เกณฑ์ความสอดคล้องของเนื้อหา (Categorical Concurrence) เป็นการเปรียบเทียบความเหมือนกันระหว่างเนื้อหาของเนื้อหาในการประเมินกับเนื้อหาที่อยู่ในผลการเรียนรู้ที่คาดหวัง เกณฑ์ดังกล่าวมีลักษณะเดียวกับการตรวจสอบความตรงเชิงเนื้อหา (Content validity) ที่ใช้การเปรียบเทียบความตรงจากผังของข้อสอบ วิธีการประเมินทำโดยการให้ผู้เชี่ยวชาญพิจารณาและระบุความสอดคล้องของเนื้อหาระหว่างคำถามในแบบทดสอบหรือกิจกรรมที่เกี่ยวข้องกับการประเมินในแต่ละจุดประสงค์ หากมีข้อคำถามที่สอดคล้องกับจุดประสงค์ 1 ตำแหน่ง ถือเป็น 1 ฮิต ในแต่ละข้อคำถามอาจสอดคล้องกับวัตถุประสงค์ได้มากกว่า 1 ข้อก็ได้ การตัดสินความสอดคล้องของเนื้อหาพิจารณาจากจำนวนฮิต (hit) ของแต่ละมาตรฐาน ระดับความสอดคล้องที่ยอมรับได้ คือ แต่ละมาตรฐานจะต้องมีจำนวนความสอดคล้องอย่างน้อย 6 ข้อ

2) เกณฑ์ความสอดคล้องด้านความลึกซึ้งของความรู้ (Depth-of-Knowledge consistency: DOK) แนวคิดหลักของเกณฑ์นี้ คือ ผลการประเมินการเรียนรู้ของผู้เรียนอยู่ในระดับเดียวกับผลการเรียนรู้ที่คาดหวังที่ผู้เรียนต้องรู้และปฏิบัติได้ จึงจะถือว่าการประเมินนั้นมี

ความสอดคล้องในแนวเดียวกัน ระดับของ DOK ในเกณฑ์นี้คำนวณจากความถี่ของความสอดคล้อง (hit) ระหว่างคำถามหรือกิจกรรมในทุกจุดประสงค์ของแต่ละมาตรฐาน หากมีคำถามหรือกิจกรรมการประเมินผู้เรียนที่สอดคล้องกับ DOK อย่างน้อยร้อยละ 50 ของคำถามหรือกิจกรรมทั้งหมด ถือว่าเป็นระดับความสอดคล้องที่ยอมรับได้ ทั้งนี้ การกำหนดระดับความลึกซึ้งของความรู้แบ่งเป็น 4 ระดับ คือ ระดับที่ 1 Recall เป็นระดับของการจดจำข้อเท็จจริงต่าง ๆ คำศัพท์ นิยาม หรือขั้นตอนและการแสดงกระบวนการหรือวิธีทำง่าย ๆ ระดับที่ 2 Skill/Concept เป็นระดับการประมวลผลทางปัญญาที่เกี่ยวข้องกับการตัดสินใจเลือกวิธีการในการแก้ปัญหาต่าง ๆ ระดับที่ 3 Strategic thinking เป็นระดับของการใช้การคิดเชิงเหตุผลในการวางแผน การอธิบายเชิงตรรกะของสิ่งต่าง ๆ ระดับที่ 4 Extended thinking เป็นระดับของการใช้ความคิดเชิงเหตุผลที่ซับซ้อนมากขึ้นในการวางแผน พัฒนา

3) ความสอดคล้องด้านขอบเขตของความรู้ (Range of Knowledge Correspondence) พิจารณาจากความถี่ของความสัมพันธ์ระหว่างวัตถุประสงค์ภายในมาตรฐานกับข้อคำถามหรือกิจกรรมการวัดประเมินผล ซึ่งต้องมีความสัมพันธ์กันอย่างน้อย 1 ข้อ โดยระดับของการยอมรับได้ในเกณฑ์นี้ คือ มีจุดประสงค์อย่างน้อยร้อยละ 50 ที่มีความสัมพันธ์กับคำถามหรือกิจกรรมการประเมินผู้เรียน

4) เกณฑ์ความสมดุลของเนื้อหา (Balance pf Representation) เกณฑ์ดังกล่าวจะดูการกระจายของเนื้อหาที่เหมาะสมและเท่าเทียมกันในแต่ละมาตรฐาน โดยคำนวณดัชนีความสมดุล (balance index) กำหนดให้มีค่าตั้งแต่ 0.7 ขึ้นไป จะถือว่ามีการกระจายของคำถามและกิจกรรมการประเมินที่สมดุลในแต่ละมาตรฐาน

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

ขั้นตอนการศึกษาความสอดคล้องของ Webb Model ประกอบด้วย 2 ขั้นตอน ขั้นตอนแรก คือการให้ผู้เชี่ยวชาญลงรหัสระดับความลึกซึ้งของความรู้ (Depth-of-Knowledge: DOK) ในมาตรฐานการเรียนรู้ ขั้นตอนที่สอง คือการให้ผู้เชี่ยวชาญลงรหัสระดับของ DOK ในข้อสอบวัดความรู้ มาตรฐานหลักสูตร และวัตถุประสงค์การเรียนรู้ จากนั้นตัดสินผลการประเมินโดยเปรียบเทียบกับระดับความสอดคล้องที่ยอมรับได้ที่กำหนดไว้ในเกณฑ์การพิจารณาความสอดคล้องทั้ง 4 เกณฑ์

ในปี ค.ศ. 1999 Webb ได้พัฒนาการตรวจสอบความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานการเรียนรู้ของรัฐกับเครื่องมือประเมินรายวิชาคณิตศาสตร์และวิทยาศาสตร์ ผลการศึกษาพบว่ามีความสอดคล้องที่หลากหลายในระดับชั้นต่าง ๆ และการพิจารณาเกณฑ์ความสมดุลของเนื้อหา พบว่า มีข้อคำถามของแบบทดสอบกระจายอยู่ตามจุดประสงค์ต่าง ๆ ของมาตรฐานการเรียนรู้อย่างเท่าเทียมกัน ในส่วนของเกณฑ์ความสอดคล้องด้านความลึกซึ้งของความรู้ พบว่าข้อ

คำถามของแบบทดสอบมีจุดมุ่งหมายที่ต่ำกว่าระดับความรู้ที่กำหนดในมาตรฐานการเรียนรู้ และมีขอบเขตเนื้อหาไม่ครอบคลุมตามที่มาตรฐานระบุไว้

ต่อมาในปี ค.ศ. 2002 Webb ได้ศึกษาความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและการประเมินวิชาคณิตศาสตร์ใน 3 รัฐ โดยใช้ผู้เชี่ยวชาญระดับชั้นละ 4 คน มีการตรวจสอบความเที่ยงระหว่างผู้ประเมินโดยใช้สหสัมพันธ์ภายในชั้น (ICC) พบว่ามีความสอดคล้องของการให้คะแนนในระดับปานกลาง ผลการศึกษาพบว่าข้อสอบที่ใช้ในการประเมินมีความสอดคล้องกับวัตถุประสงค์มากกว่า 1 ข้อ และระดับความลึกของความรู้ (DOK) สอดคล้องกับเนื้อหาในจุดประสงค์ แต่ไม่พบความสอดคล้องตามเกณฑ์ด้านขอบเขตของความรู้

Andrew และ Dawn (2010) ศึกษาความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานการเรียนรู้ระดับประถมวัยของรัฐอินเดียนา โดยใช้วิธีการศึกษาความสอดคล้องของ Webb ในการเปรียบเทียบตัวชี้วัดของหลักสูตรจากระดับที่สูงกว่ากับข้อรายการประเมินที่เป็นมาตรประมาณค่าของ ISTAR ในระดับพื้นฐาน ใช้ผู้เชี่ยวชาญจำนวน 13 คน พบว่า มีความไม่สอดคล้องกันในเกณฑ์ของ Webb ที่ใช้ในการศึกษา คือ เกณฑ์ด้านความลึกซึ้งของความรู้ และเกณฑ์ด้านขอบเขตของความรู้

4.2.2 The Survey of Enacted Curriculum Model (SEC)

The Survey of Enacted Curriculum Model (SEC) เป็นวิธีการศึกษาความสอดคล้องในแนวเดียวกันของ Porter ซึ่งพัฒนาจากแนวคิดของการประเมินหลักสูตร วิธีการนี้เน้นการให้ความสำคัญกับเนื้อหาของหลักสูตร เพราะเนื้อหาเป็นสิ่งที่ผู้เรียนได้เรียนรู้จึงเป็นสิ่งที่สมควรได้รับการประเมินเป็นอันดับแรก คำว่า ‘เนื้อหา’ ในความหมายของ Porter และ Smithson (2001) หมายถึง หัวเรื่อง (topic) ที่ครูสอน หรือเรื่องที่ใช้ในการทดสอบ

วิธีการศึกษาความสอดคล้องของ Porter และ Smithson เป็นการศึกษาความสอดคล้องในเชิงปริมาณใน 3 องค์ประกอบของระบบการศึกษา จึงเป็นวิธีการที่ช่วยให้ผู้เกี่ยวข้องกับกระบวนการทางการศึกษาเห็นความเชื่อมโยงระหว่างเป้าหมายของการจัดการเรียนรู้กับสิ่งที่สอนในชั้นเรียนกับสิ่งที่ประเมิน รวมถึงเปรียบเทียบระดับความสอดคล้องได้ในระดับโรงเรียนและระดับรัฐ วิธีการของ Porter จะใช้ผู้เชี่ยวชาญประเมินความสอดคล้องระหว่างองค์ประกอบต่าง ๆ แล้วนำผลการประเมินมาสร้างเมตริกซ์จำนวน 2 เมตริกซ์ แล้วคำนวณค่าความสอดคล้องระหว่างเมตริกซ์การประเมินทั้งสอง เรียกว่า ดัชนีความสอดคล้อง (alignment index)

มิติการประเมินความสอดคล้องของ Porter ประกอบด้วย 3 มิติ ได้แก่ ความสอดคล้องของเนื้อหา (content match) สมรรถนะที่คาดหวังของผู้เรียน (expectations for student performance) และเนื้อหาในการสอน (instructional content)

มิติความสอดคล้องของเนื้อหานั้น Porter จะสร้างเมตริกซ์ของเนื้อหาขึ้นมาโดยแยกออกเป็นหัวข้อตามระดับชั้นต่าง ๆ โดยสามารถทำได้ทั้งแบบละเอียดซึ่งแสดงหัวเรื่องทั้งหมด หรือแบบหยาบซึ่งแสดงเฉพาะหัวเรื่องหลักก็ได้ วิธีการนี้ใกล้เคียงกับการตรวจสอบความตรงเชิงเนื้อหา แต่ให้สารสนเทศจะแตกย่อยลงไปมากกว่า ในส่วนของมิติสมรรถนะที่คาดหวังของผู้เรียนนั้นจะคล้ายกับการวัดระดับความลึกซึ้งของความรู้ (DOK) ของ Webb Model แต่วิธีการของ Porter (2002) จะจำแนกความซับซ้อนทางปัญญา (cognitive demand) ออกเป็น 5 ระดับ โดยตั้งอยู่บนพื้นฐานของ revised Bloom's taxonomy ได้แก่ การจำ (memorize) การปฏิบัติตามขั้นตอน (perform procedure) การสื่อสารความเข้าใจ (communicate understanding) การแก้ปัญหาในสถานการณ์จำเพาะ (solve non-routine problems) และสุดท้ายคือ การคาดคะเน/สรุปอ้างอิง/พิสูจน์ (conjecture/generalize/prove)

มิติสุดท้าย คือ มิติเนื้อหาในการสอน ซึ่งเป็นมิติที่ Porter และ Smithson ให้ความสำคัญมากที่สุด เพราะเป็นตัวแปรแทรกแซงในการศึกษาผลสัมฤทธิ์ของผู้เรียน การเก็บข้อมูลเกี่ยวกับเนื้อหาในการสอนทำได้โดยการประเมินเนื้อหาที่ครูสอน และการเน้นด้านความลึกของความรู้ซึ่งวัดว่าครูเน้นให้ผู้เรียนใช้ความซับซ้อนทางปัญญาระดับใด

ขั้นตอนในการศึกษาความสอดคล้องในแนวเดียวกันของ Porter คือ การให้ผู้เชี่ยวชาญพิจารณาเนื้อหาของเครื่องมือประเมิน และเนื้อหามาตรฐาน จากนั้นใช้กระบวนการลงรหัสในตารางแบบ 2 มิติ โดยให้แถวเป็นหัวเรื่องของเนื้อหา และคอลัมน์เป็นสมรรถนะที่คาดหวังของผู้เรียน ค่าในแต่ละช่อง (cell) คือสัดส่วนของเนื้อหาในช่องนั้นต่อเนื้อหาทั้งหมด ในการประเมินความสอดคล้องระหว่างเนื้อหาของมาตรฐานการเรียนรู้กับเนื้อหาของแบบสอบนั้น ผู้เก็บข้อมูลต้องทำเมตริกซ์แยกกัน 2 เมตริกซ์ โดยเมตริกซ์ X เป็นความสัมพันธ์ระหว่างเนื้อหา (แถว) กับสมรรถนะของผู้เรียน (คอลัมน์) ส่วนเมตริกซ์ Y เป็นความสัมพันธ์ระหว่างคำถามในแบบทดสอบ (แถว) กับสมรรถนะของผู้เรียน (คอลัมน์) จากนั้นให้ผู้เชี่ยวชาญประเมินความสอดคล้องระหว่างรายการประเมินทั้ง 2 เมตริกซ์ แล้วนำคะแนนการประเมินที่ได้จากทั้ง 2 เมตริกซ์มาคิดดัชนีความสอดคล้องดังสมการ

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

$$\text{ดัชนีความสอดคล้อง} = 1 - \frac{\sum|X_i - Y_i|}{2}$$

เมื่อ X_i คือคะแนนประเมินเมตริกซ์ X ตำแหน่งที่ i

Y_i คือคะแนนประเมินเมตริกซ์ Y ตำแหน่งที่ i

ค่าดัชนีความสอดคล้องมีค่าตั้งแต่ 0 ถึง 1 โดย 0 หมายถึง ไม่สอดคล้อง และ 1 หมายถึง สอดคล้องอย่างสมบูรณ์

Blank, Porter และ Smithson (2001) ศึกษาระดับความสอดคล้องในแนวเดียวกันระหว่างการสอนและการวัดและประเมินผลใน 6 รัฐ ของสหรัฐอเมริกา โดยใช้ผู้เชี่ยวชาญจำนวน 4 คน ประชุมเพื่อประเมินความสอดคล้องระหว่างข้อคำถามในการประเมินผู้เรียนกับมาตรฐานการเรียนรู้ พบว่า ความสอดคล้องในแนวเดียวกันระหว่างการประเมินกับการสอนภายในรัฐมีค่าใกล้เคียงกับความสอดคล้องในแนวเดียวกันระหว่างรัฐ เมื่อศึกษาความสอดคล้องในการประเมินของรัฐกับการประเมินความก้าวหน้าทางการศึกษาแห่งชาติ พบว่ามีค่าความสอดคล้องเพิ่มขึ้นเมื่อเทียบกับค่าความสอดคล้องภายในรัฐ

Porter, Smithson, Blank และ Zeidner (2007) ศึกษาเปรียบเทียบแผนที่เนื้อหาและดัชนีความสอดคล้องในการใช้โปรแกรมการพัฒนาวิชาชีพครูคณิตศาสตร์และครูวิทยาศาสตร์ระหว่างโรงเรียนที่เข้าร่วมโปรงแกรมและไม่ได้เข้าร่วมโปรแกรม พบว่า ดัชนีความสอดคล้องสามารถใช้เป็นตัวแปรตามในการศึกษานัยสำคัญของกลุ่มทดลองเทียบกับกลุ่มควบคุม นอกจากนี้ ดัชนีความสอดคล้องที่เป็นข้อมูลเชิงปริมาณมีประโยชน์ต่อการพิจารณาความสอดคล้องทั้งระหว่างครูเป็นรายบุคคล ระหว่างกลุ่มครู ระหว่างมาตรฐาน และองค์ประกอบอื่น ๆ ที่เกี่ยวข้องกับการจัดการศึกษาที่สามารถวิเคราะห์เนื้อหาได้

Polikoff และ Porter (2014) ศึกษาบทบาทของความสอดคล้องในแนวเดียวกันกับการเรียนการสอนในการทำนายตัวชี้วัดในการจัดการเรียนการสอน โดยใช้ตัวอย่างเป็นครูจำนวน 327 คน และศึกษาจากการวิเคราะห์เนื้อหาโดยกลุ่มผู้เชี่ยวชาญ จำนวน 3 คน พบว่า ดัชนีความสอดคล้องมีความสัมพันธ์กับคะแนนโมเดลมูลค่าเพิ่มในระดับต่ำ และพบอิทธิพลปฏิสัมพันธ์เล็กน้อย รวมถึงพบว่าความสอดคล้องของการสอนไม่มีความสัมพันธ์กับองค์ประกอบในการวัดประสิทธิภาพของการเรียนการสอน

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Polikoff (2015) ศึกษาความสอดคล้องของเนื้อหาตำราเรียนวิชาคณิตศาสตร์กับมาตรฐานแกนกลาง โดยใช้ข้อมูลจากการสำรวจหลักสูตรที่นำไปใช้จริง (SEC) ที่ Porter จัดทำขึ้นในปี 2002 การศึกษานี้ใช้ผู้เชี่ยวชาญจำนวน 3-4 คนในการลงรหัสความสอดคล้องในแนวเดียวกัน ผลการศึกษาพบว่า มีเนื้อหาจำนวนมากที่ไม่สอดคล้องกันระหว่างเอกสารตำราเรียนกับมาตรฐานการเรียนรู้กลาง นอกจากนี้ยังพบว่า ตำราเรียนวิชาคณิตศาสตร์เน้นเนื้อหาที่เป็นความจำมากกว่าด้านอื่น ๆ

4.2.3 Achieve Model

การศึกษาความสอดคล้องที่พัฒนาโดยบริษัท Achieve หรือ Achieve Model เป็นการวัดความสอดคล้องระหว่างเนื้อหาของมาตรฐานกับเนื้อหาในการประเมิน ซึ่งให้สารสนเทศทั้งเชิงปริมาณและคุณภาพ Achieve Model ได้รับการพัฒนาขึ้นเพื่อเปรียบเทียบความสอดคล้องระหว่างมาตรฐานการศึกษาของรัฐหรือประเทศต่าง ๆ ซึ่งผลดงกล่าวนำไปใช้พัฒนาศักยภาพของบุคลากรทางการศึกษาและใช้เป็นข้อมูลประกอบการปฏิรูปการศึกษาในระดับประเทศได้ การวัดความสอดคล้องโดย Achieve Model วิเคราะห์ความสอดคล้องใน 4 มิติ ดังนี้

1) Content Centrality เป็นการวิเคราะห์ความสอดคล้องระหว่างเนื้อหาของข้อสอบแต่ละข้อกับเนื้อหาของมาตรฐาน โดยการตรวจสอบคุณภาพความสอดคล้องจะใช้ผู้เชี่ยวชาญเป็นผู้กำหนดระดับความสอดคล้องของข้อสอบแต่ละข้อบนมาตรแบบ 5 ระดับ

2) Performance Centrality เป็นการวิเคราะห์ความสอดคล้องระหว่างสมรรถนะหรือความซับซ้อนทางปัญญาของผู้เรียนในข้อสอบแต่ละข้อกับสมรรถนะที่กำหนดในมาตรฐาน

3) Challenge เป็นการวิเคราะห์ว่าข้อสอบสามารถวัดความรอบรู้ของผู้เรียนได้หรือไม่ โดยพิจารณาใน 2 ปัจจัย คือ แหล่งความท้าทาย เป็นการวัดว่าข้อสอบได้รับการออกแบบขึ้นอย่างยุติธรรมและไม่มีข้อผิดพลาด ระดับความท้าทาย เป็นการเปรียบเทียบน้ำหนักความสำคัญของสมรรถนะที่ระบุในมาตรฐาน โดยประเมินว่าข้อสอบนั้นยากง่ายเหมาะสมกับผู้เรียนหรือไม่

4) Balance and Range เป็นการพิจารณาความสมดุลและครอบคลุมของมาตรฐาน โดยความสมดุล (balance) เป็นการประเมินภาพรวมในระดับมาตรฐาน เพื่อเปรียบเทียบน้ำหนักความสำคัญของเนื้อหาที่ปรากฏในแบบทดสอบกับน้ำหนักความสำคัญของเนื้อหาที่ระบุในมาตรฐาน ส่วนความครอบคลุม (range) เป็นการวัดขอบเขตของเนื้อหาโดยวัดจากสัดส่วนของจุดประสงค์

การเรียนรู้ในมาตรฐานที่วัดโดยข้อสอบอย่างน้อย 1 ข้อ สัดส่วนดังกล่าวควรมีค่าระหว่าง 0.50 – 0.66 หากสูงกว่า 0.67 จะถือว่ามีความครอบคลุมและมีความหลากหลายระดับที่ดี

ขั้นตอนการศึกษาความสอดคล้องในแนวเดียวกันของ Achieve Model จะมีขั้นตอนการดำเนินงาน 3 ขั้นตอนหลัก (Rothman, Slatterry และ Vranek, 2002) ขั้นตอนที่ 1 เป็นการจับคู่ความสอดคล้อง (matching) ระหว่างเครื่องมือการประเมินกับมาตรฐานการเรียนรู้เป็นรายข้อ ขั้นตอนที่ 2 เป็นการตรวจสอบความท้าทายของแบบทดสอบ และขั้นตอนที่ 3 เป็นการตรวจสอบความสมดุลและความครอบคลุมของแบบทดสอบ

Rothman, Slattery, Vranek และ Resnick (2002) ประยุกต์วิธีการศึกษาความสอดคล้องในแนวเดียวกันของบริษัท Achieve ในการตรวจสอบการประเมินผลการเรียนรู้ใน 5 รัฐ โดยอบรมผู้เชี่ยวชาญด้านการสอน ด้านหลักสูตร และด้านเนื้อหาและการวัดและประเมินผล ผลการศึกษาพบว่า ข้อสอบมีความสอดคล้องกับเนื้อหาและสมรรถนะที่ระบุในมาตรฐานการเรียนรู้ อย่างไรก็ตาม การประเมินของรัฐต่าง ๆ ไม่คำนึงถึงการตรวจสอบเกณฑ์ด้านแหล่งความท้าทาย รวมถึงทุกรัฐไม่มีการตรวจสอบเกณฑ์ความครอบคลุมของเนื้อหา ทั้งนี้พบว่าแบบทดสอบให้ความสำคัญมากเกินไปกับจุดประสงค์ที่มีน้ำหนักความสำคัญน้อย การศึกษานี้ได้สรุปเกี่ยวกับจุดเด่นและข้อจำกัดของวิธีการ Achieve Model ว่า เป็นวิธีการศึกษาที่ละเอียดและสามารถแสดงจุดเด่นและสิ่งที่ต้องปรับปรุงของระบบการประเมินทางการศึกษาในแต่ละรัฐได้อย่างดี อย่างไรก็ตาม วิธีการ Achieve Model นั้นใช้เวลาค่อนข้างมากและมีค่าใช้จ่ายในการดำเนินการค่อนข้างสูง

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Case, Jorgensen และ Zucker (2004) ได้เปรียบเทียบความแตกต่างระหว่างวิธีการศึกษาความสอดคล้องในแนวเดียวกันทางการศึกษาทั้ง 3 วิธี ดังแสดงในตาราง 2.7

ตาราง 2.7 การเปรียบเทียบวิธีการศึกษาความสอดคล้องในแนวเดียวกัน

โมเดล	จุดเด่น	ระยะเวลาการเก็บข้อมูลและการวิเคราะห์	ระยะเวลาการอบรมผู้เชี่ยวชาญ
Webb	1. เป็นการประเมินเชิงปริมาณ 2. ให้ผลการประเมินเชิงคุณภาพ 3. สามารถวัดความสอดคล้องระหว่างผู้ประเมินได้จากสถิติความสอดคล้อง	- 1 วันต่อทีม สำหรับการจับคู่เนื้อหากับแบบทดสอบ - 1 เดือน สำหรับการวิเคราะห์และรายงานผลโดยทั่วไป - ในกรณีที่ใช้ Webb Alignment Tool ใช้เวลา 2 ชั่วโมงสำหรับ	ใช้เวลาประมาณครึ่งวันในการอบรมผู้เชี่ยวชาญ

โมเดล	จุดเด่น	ระยะเวลาการเก็บข้อมูลและการวิเคราะห์	ระยะเวลาการอบรมผู้เชี่ยวชาญ
		การคำนวณฉันทามติระหว่างผู้ประเมินในขั้นตอนแรก และใช้เวลาประมาณ 60 ถึง 90 นาทีในการวิเคราะห์ DOK จากนั้นทำการประมวลผลสรุปและรายงานภายใน 2 สัปดาห์	
SEC	1. ใช้การประเมินโดยเมตริกซ์เนื้อหา 2. ผลการวัดความสอดคล้องมีความตรงเชิงพยากรณ์ที่สอดคล้องกับคะแนนผลสัมฤทธิ์ของผู้เรียน 3. สารสนเทศที่ได้สามารถช่วยให้โรงเรียนและครูปรับปรุงการจัดการเรียนการสอน	- ใช้เวลา 1 วันในการลงรหัสและเปรียบเทียบเมตริกซ์ - ใช้เวลาครึ่งวันสำหรับผู้เชี่ยวชาญประเมินเชิงคุณภาพ - ใช้เวลา 1 สัปดาห์ สำหรับการวิเคราะห์และรายงานผล	ใช้เวลาประมาณครึ่งวันในการอบรมผู้เชี่ยวชาญ
Achieve	1. ผู้เชี่ยวชาญต้องกำหนดเกณฑ์ด้วยตนเอง 2. มีการสัมภาษณ์เชิงลึก 3. มีการรายงานผลเชิงเทคนิค	- การพิจารณาความสอดคล้องใช้เวลา 1 วัน - การวิเคราะห์และรายงานผลใช้เวลาประมาณ 4 – 6 สัปดาห์	มีผู้เชี่ยวชาญขององค์กรเอง ไม่จำเป็นต้องอบรมเพิ่ม

Chulalongkorn University

นอกจากวิธีการศึกษาความสอดคล้องในแนวเดียวกันทั้ง 3 วิธี ที่ได้กล่าวมาแล้ว ยังมีการศึกษาเกี่ยวกับวิธีการศึกษาความสอดคล้องในแนวเดียวกันวิธีอื่น ๆ ได้แก่ การศึกษาของ Eckhout, Plake, Smith และ Larsen (2007) ที่ศึกษาความสอดคล้องในแนวเดียวกันของมาตรฐานการเรียนรู้ทางเลือกในการศึกษาสำหรับเด็กที่มีความต้องการพิเศษ กับมาตรฐานการเรียนรู้แกนกลางซึ่งเป็นมาตรฐานสำหรับผู้เรียนทั่วไป การศึกษานี้ใช้ผู้เชี่ยวชาญซึ่งเป็นครูด้านการศึกษาพิเศษจำนวน 38 คน แบ่งเป็น 6 กลุ่มย่อย กลุ่มละ 6-7 คน ผู้เชี่ยวชาญประเมินระดับความสอดคล้องในแนวเดียวกันระหว่างเนื้อหาและมาตรฐานการเรียนรู้ทางเลือก กับเนื้อหาและมาตรฐานการเรียนรู้กลางโดยใช้มาตรประมาณค่า 3 ระดับ จากนั้นประชุมกลุ่มผู้เชี่ยวชาญเพื่อหาฉันทามติของการประเมิน

ผลการศึกษาพบว่าผู้เชี่ยวชาญเห็นตรงกันว่าทักษะที่ระบุในมาตรฐานส่วนใหญ่สอดคล้องกับระดับความสมรรถนะที่มุ่งหวังของผู้เรียน

Lombardi, Sebuen, Conley และ Snow (2010) ประยุกต์การศึกษา G-Theory ในการศึกษาความเที่ยงของผู้ประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบคัดเลือกเข้ามหาวิทยาลัยกับมาตรฐานการเตรียมความพร้อมในวิชาคณิตศาสตร์และวิชาภาษาอังกฤษ โดยให้ผู้เชี่ยวชาญ 6 คน ประเมินความสอดคล้องในแนวเดียวกันทางระบบออนไลน์ ใช้รูปแบบการประเมินแบบ i x r คือ ผู้เชี่ยวชาญทุกคนประเมินรายการประเมินทุกข้อ ผลการศึกษาพบว่า ข้อสอบและมาตรฐานวิชาคณิตศาสตร์ รวมถึงประเภทของมาตรที่ใช้ในการประเมินส่งผลต่อความเที่ยงของการประเมิน

Herman, Webb และ Stephen (2007) ทำการศึกษาเพื่อตรวจสอบผลกระทบของฉันทามติระหว่างผู้ประเมินที่มีต่อการพิจารณาตัดสินความสอดคล้องในแนวเดียวกันระหว่างการทดสอบทางคณิตศาสตร์ของโรงเรียนมัธยมกับประกาศของมหาวิทยาลัยแคลิฟอร์เนียเกี่ยวกับสมรรถนะที่คาดหวังทางคณิตศาสตร์ของนักศึกษา โดยให้อาจารย์ของมหาวิทยาลัยและครูของโรงเรียน จำนวน 20 คน ประเมินข้อสอบคณิตศาสตร์ของการทดสอบของรัฐ ผลการศึกษาพบว่า มีความแปรปรวนในการตัดสินของผู้ประเมิน และพบความแตกต่างระหว่างรูปแบบการประเมินของผู้ประเมิน

การศึกษาความสอดคล้องในแนวเดียวกัน เป็นการศึกษาความสัมพันธ์เชื่อมโยงเป็นอันหนึ่งอันเดียวกันระหว่างองค์ประกอบต่าง ๆ ทางการศึกษา ซึ่งแหล่งข้อมูลที่สำคัญในการศึกษาความสอดคล้อง ได้แก่ ผู้เชี่ยวชาญในการประเมินระดับความสอดคล้อง ดังนั้น ฉันทามติระหว่างผู้ประเมินจึงเป็นประเด็นสำคัญของการศึกษาความสอดคล้องในการประเมิน หากผู้ประเมินมีความแปรปรวนในการให้คะแนนสูง หรือมีรูปแบบการให้คะแนนที่ไม่เป็นอันหนึ่งอันเดียวกันก็จะส่งผลต่อผลการศึกษาและให้สารสนเทศที่ผิดพลาดได้ ดังนั้น การวิเคราะห์ความเที่ยงระหว่างผู้ประเมินจึงมีความสำคัญในกระบวนการการศึกษาดังกล่าว เนื่องจากเป็นการสะท้อนภาพของรูปแบบการให้คะแนนของผู้ประเมินว่ามีฉันทามติเดียวกันหรือไม่ ก่อนที่จะนำผลการประเมินไปสรุปอ้างอิงต่อไป การศึกษาครั้งนี้เป็นการนำโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมมาประยุกต์ใช้เพื่อวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินในการศึกษาความสอดคล้องในแนวเดียวกันทางการศึกษา ซึ่งยังไม่เคยมีการนำโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมมาใช้วิเคราะห์ในบริบทการประเมินลักษณะนี้มาก่อน ดังนั้น ผู้วิจัยจึงศึกษาความเป็นไปได้ของการประยุกต์ใช้โมเดลดังกล่าวโดยการจำลอง

สถานการณ์ในบริบทของการประเมินความสอดคล้องในแนวเดียวกันทางการศึกษาที่มีความแตกต่างกันของจำนวนผู้ประเมิน จำนวนคำถามประเมิน และการทำหน้าที่ต่างกันระหว่างผู้ประเมิน จากนั้นทำการทดสอบประสิทธิภาพในการประมาณค่าพารามิเตอร์ของโมเดล MC-GCM และ MC-LTRM และตรวจสอบความถูกต้องและแม่นยำของการประมาณค่าของโมเดลโดยพิจารณาจากค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสอง ค่าความลำเอียงในการประมาณค่า และค่าสัมประสิทธิ์สหสัมพันธ์ ดังแสดงในรูป 2.15

กรอบแนวคิดในการวิจัย

รูป 2.15 กรอบแนวคิดในการวิจัย

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

บทที่ 3
วิธีดำเนินการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อ 1) เพื่อประยุกต์ใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมเพื่อวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินในการศึกษาความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียนกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น 2) เพื่อตรวจสอบประสิทธิภาพของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมในการการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินในการศึกษาความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียน กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น 3) เพื่อศึกษาผลการประยุกต์ใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมในการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินและการทำหน้าที่ต่างกันของผู้ประเมินในการประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียนกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น การตัดสินประสิทธิภาพของโมเดลพิจารณาจาก 1) ความสอดคล้องระหว่างค่าที่กำหนดกับค่าที่ได้จากการประมาณค่าของโมเดล โดยพิจารณาค่าสัมประสิทธิ์สหสัมพันธ์ (Pearson correlation coefficient) 2) ความถูกต้องและน่าเชื่อถือของผลการประมาณค่า โดยพิจารณาจาก ความลำเอียงในการประมาณค่า (Bias of estimators) และค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Mean Square Error: MSE)

วิธีการดำเนินการวิจัยมีทั้งหมด 3 ระยะ ได้แก่ <u>ระยะที่ 1</u> การศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องกับการวิเคราะห์ความเที่ยงระหว่างผู้ประเมิน การทำหน้าที่ต่างกันของผู้ประเมิน โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม และการศึกษาความสอดคล้องในแนวเดียวกันทางการศึกษา <u>ระยะที่ 2</u> การศึกษาผลการประมาณความสอดคล้องระหว่างผู้ประเมินด้วยการจำลองแบบมอนติคาร์โลแบ่งเป็น การจำลองข้อมูล การประเมินประสิทธิภาพของการประมาณค่าของโมเดล และการวิเคราะห์ผลการจำลองข้อมูล <u>ระยะที่ 3</u> ศึกษาผลการประยุกต์ใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมในการวิเคราะห์ความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบ ในการประเมินระดับชั้นเรียน กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น รายละเอียดของการดำเนินการวิจัย มีดังต่อไปนี้

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ระยะที่ 1 การศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง

การศึกษาในระยะที่ 1 มีมีวัตถุประสงค์เพื่อศึกษาแนวทางในการประยุกต์ใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมในการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินในการวิเคราะห์ความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบ ในการประเมินระดับชั้นเรียน กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น การศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องประกอบด้วย การศึกษาแนวคิดทฤษฎีเกี่ยวกับการวิเคราะห์ความเที่ยงระหว่างผู้ประเมิน แนวคิดทฤษฎีเกี่ยวกับการทำหน้าที่ต่างกันของผู้ประเมิน แนวคิดและทฤษฎีที่เกี่ยวข้องกับการวิเคราะห์ฉันทามติเชิงวัฒนธรรม แนวคิดทฤษฎีเกี่ยวกับการศึกษาความสอดคล้องในแนวเดียวกันทางการศึกษา

การดำเนินการวิจัยในระยะที่ 1 ประกอบด้วย 2 ส่วน ส่วนที่ 1 เป็นการวิเคราะห์จุดเด่นและข้อจำกัดของวิธีการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมิน การวิเคราะห์การทำหน้าที่ต่างกันของผู้ประเมิน และการศึกษาความสอดคล้องในแนวเดียวกันทางการศึกษา โดยเน้นการศึกษาข้อจำกัดเพื่อนำเสนอวิธีการวิเคราะห์ที่สามารถให้สารสนเทศความสอดคล้องระหว่างผู้ประเมิน การทำหน้าที่ต่างกันระหว่างผู้ประเมิน และผลการศึกษาความสอดคล้องในแนวเดียวกันทางการศึกษาได้พร้อมกันจากการวิเคราะห์ข้อมูลในคราวเดียว ส่วนที่ 2 เป็นการศึกษาโมเดลการวิเคราะห์การทำหน้าที่ต่างกันระหว่างผู้ประเมิน โดยศึกษาแนวคิดของโมเดลการวิเคราะห์ข้อมูล คำจำกัดความของพารามิเตอร์ วิธีการคำนวณ การประมาณค่าทางสถิติ แล้วนำมาเชื่อมโยงกับขั้นตอนการศึกษาความสอดคล้องในแนวเดียวกัน และนำมาประยุกต์ใช้กับการวิเคราะห์ความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบ ในการประเมินระดับชั้นเรียน กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น

CHULALONGKORN UNIVERSITY

ระยะที่ 2 การศึกษาผลการประมาณความสอดคล้องระหว่างผู้ประเมินด้วยการจำลองแบบมอนติคาร์โล

การศึกษานี้มีวัตถุประสงค์เพื่อตรวจสอบประสิทธิภาพของการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินในการวิเคราะห์ความสอดคล้องในแนวเดียวกัน 2 โมเดล ได้แก่ MC-GCM และ MC-LTRM ในการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินในสถานการณ์ที่มีจำนวนผู้ประเมิน จำนวนคำถามหรือรายการประเมิน และการทำหน้าที่ต่างกันของผู้ประเมิน

ผู้วิจัยกำหนดจำนวนผู้ประเมิน และจำนวนรายการประเมินจากการศึกษาเอกสารงานวิจัยที่เกี่ยวข้องกับการศึกษาโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม (Anders and Batchelder, 2012; 2015, Anders และคณะ, 2014) และงานวิจัยของบุษยารัตน์ จันทร์ประเสริฐ (2561) ในส่วนของจำนวนผู้ประเมินและจำนวนรายการประเมินในสถานการณ์จริง โดยผู้วิจัยได้จำลองข้อมูลสำหรับการวิเคราะห์ทั้ง 2 โมเดล โดยแบ่งรายละเอียดตามโมเดลการวิเคราะห์ ดังนี้

การจำลองข้อมูล

1. การจำลองข้อมูลสำหรับการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินด้วยโมเดล MC-GCM

ผู้วิจัยทำการกำหนดค่าพารามิเตอร์ตามที่ระบุในขอบเขตการวิจัย จากนั้น ประมาณค่าพารามิเตอร์จากเมตริกซ์ X_{ik} ได้จากการจำลองข้อมูล โดยใช้โมเดล MC-GCM โดยกำหนดการประมาณค่าตามค่าตั้งต้นของโมเดล ซึ่งใช้จำนวนตัวอย่าง 10000 หน่วย จำนวนลูกโซ่ 3 ลูกโซ่ แล้วบันทึกค่าประมาณพารามิเตอร์ที่ได้จากการประมาณค่าของโมเดลลงในเมตริกซ์ที่สร้างไว้ จำนวน 6 เมตริกซ์ ดังนี้

1) theta.est เป็นเมตริกซ์เก็บค่าประมาณความสามารถของผู้ประเมิน (θ_i) ขนาด RxN เมื่อ R คือ จำนวนรอบของการกระทำซ้ำการจำลองข้อมูล

2) g.est เป็นเมตริกซ์เก็บค่าประมาณความลำเอียงในการประเมิน (g_i) ขนาด RxN

3) del1.est เป็นเมตริกซ์เก็บค่าประมาณความยากของรายการประเมิน (δ_{k1}) สำหรับผู้ประเมินกลุ่มที่ 1 ขนาด RxM

4) del2.est เป็นเมตริกซ์เก็บค่าประมาณความยากของรายการประเมิน (δ_{k2}) สำหรับผู้ประเมินกลุ่มที่ 2 ขนาด RxM

5) Z1.est เป็นเมตริกซ์เก็บค่าคำตอบฉันทามติของผู้ประเมินกลุ่มที่ 1 $(Z_{t_1,k})$ สำหรับผู้ประเมินกลุ่มที่ 1 ขนาด RxM

6) Z2.est เป็นเมตริกซ์เก็บค่าคำตอบฉันทามติของผู้ประเมินกลุ่มที่ 1 $(Z_{t_2,k})$ สำหรับผู้ประเมินกลุ่มที่ 2 ขนาด RxM

ทำซ้ำขั้นตอนที่ 1 ถึง 9 จำนวน 100 ครั้ง

2. การจำลองข้อมูลสำหรับการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินด้วยโมเดล MC-LTRM

ผู้วิจัยประมาณค่าพารามิเตอร์จากเมตริกซ์ X_{ik} ที่ได้จากการจำลองข้อมูลตามขอบเขตที่ระบุในขอบเขตการวิจัย โดยใช้โมเดล MC-LTRM โดยกำหนดการประมาณค่าตามค่าตั้งต้นของโมเดล ซึ่งใช้จำนวนตัวอย่าง 10000 หน่วย จำนวนลูกโซ่ 3 ลูกโซ่ จากนั้น บันทึกค่าประมาณพารามิเตอร์ที่ได้จากการประมาณค่าของโมเดลลงในเมตริกซ์ที่สร้างไว้ จำนวน 5 เมตริกซ์ ดังนี้

1) E.est เป็นเมตริกซ์เก็บค่าประมาณความสามารถของผู้ประเมิน (E_i) -ขนาดRxN เมื่อ R คือ จำนวนรอบของการกระทำซ้ำการจำลองข้อมูล

2) a.est เป็นเมตริกซ์เก็บค่าประมาณความลำเอียงในการประเมิน (a_i) ขนาด RxN

3) b.est เป็นเมตริกซ์เก็บค่าประมาณความลำเอียงในการประเมิน (b) ขนาด RxN

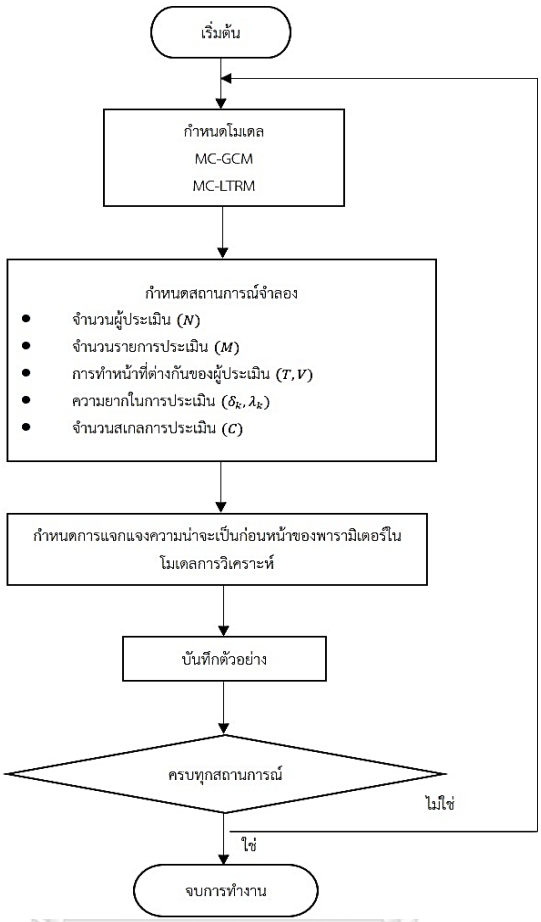
4) T.est เป็นเมตริกซ์เก็บค่าประมาณตำแหน่งของคะแนนการประเมิน (T_{vk}) สำหรับผู้ประเมินขนาด RxM

5) lam.est เป็นเมตริกซ์เก็บค่าประมาณความยากของรายการประเมิน (λ_k) สำหรับผู้ประเมิน ขนาด RxM

ทำซ้ำขั้นตอนที่ 1 ถึง 7 จำนวน 100 ครั้ง

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

ขั้นตอนการศึกษาจากสถานการณ์จำลอง มีขั้นตอนดังรูป 3.1



รูป 3.1 ขั้นตอนการจำลองข้อมูล

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

การตรวจสอบความถูกต้องของข้อมูลจำลอง

1. การตรวจสอบความถูกต้องของการจำลองข้อมูลในโมเดล MC-GCM

จากการทดลองจำลองและประมาณค่าข้อมูลจากสถานการณ์จำลอง พบว่าการประมาณค่าข้อมูลจำลองมีค่าสูญหาย (NA) เนื่องจากโมเดล MC-GCM ไม่สามารถประมาณค่าพารามิเตอร์ในบางรอบของการจำลองข้อมูลได้ ปรากฏการณ์ดังกล่าวเกิดขึ้นกับกรณีกำหนดข้อคำถามมีความยากแตกต่างกันเป็นส่วนใหญ่ ผู้วิจัยจึงทำการทดสอบความคงที่ของการประมาณค่าพารามิเตอร์ของโมเดลโดยทำการจำลองข้อมูลจำนวน 300 รอบ แล้วพล็อตกราฟเพื่อดูความคงที่ของการประมาณค่าพารามิเตอร์ ใน 2 กรณี ได้แก่ 1) กรณีการจำลองข้อมูลที่คำถามประเมินมีความยากเท่าเทียมกัน และ 2) กรณีการจำลองข้อมูลคำถามประเมินมีความยากแตกต่างกัน

ผู้วิจัยทดสอบความคงที่ของการประมาณค่าฉันทามติเชิงวัฒนธรรมกรณีข้อคำถามการประเมินมีความยากแตกต่างกัน โดยเลือกสถานการณ์จำลองที่มีผู้ประเมิน 15 คน รายการประเมิน 25 ข้อ ซึ่งเป็นสถานการณ์จำลองที่มีจำนวนผู้ประเมินและรายการประเมินน้อยที่สุดในการศึกษาครั้งนี้ เนื่องจากถึงแม้จะมีการศึกษาว่าโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมมีความเสถียรในการวิเคราะห์ตัวอย่างจำนวนน้อย ขึ้นอยู่กับความสามารถของผู้ตอบและจำนวนข้อคำถาม หากจำนวนตัวอย่างและข้อคำถามมีจำนวนน้อยเกินไป รวมถึงหากการประเมินนั้นมีความแปรปรวนต่ำ อาจทำให้ผลการวิเคราะห์ที่ได้ในแต่ละครั้งมีความคลาดเคลื่อนได้(Batchelder และคณะ,1986) การทดสอบความคงที่ของการประมาณค่าฉันทามติเชิงวัฒนธรรมด้วยโมเดล MC-GCM มีรายละเอียดดังนี้

<u>การทดสอบความคงที่ของการประมาณค่ากรณีไม่มีการทำหน้าที่ต่างกันของผู้ประเมิน</u>

ผู้วิจัยกำหนดจำนวนผู้ประเมิน $N = 15$ จำนวนรายการประเมิน $M = 25$ จำนวนกลุ่มวัฒนธรรม $T = 2$ เวกเตอร์ความสามารถของผู้ประเมินที่ได้จากการสุ่มมีค่าตั้งแต่ 0.037 ถึง 0.793 เวกเตอร์ความลำเอียงในการเดามีค่าตั้งแต่ 0.286 ถึง 0.793 เวกเตอร์ความยากของรายการประเมินคือ $[0.5, \ldots, 0.5]$ เวกเตอร์ฉันทามติของผู้ประเมินมีค่าดังนี้

ผลการประเมินของผู้ประเมินกลุ่มที่ 1 $\mathbf{1\,1\,1\,0\,1\,0\,1\,0\,0\,1\,1\,0\,1\,0\,0\,1\,1\,1\,0\,1\,0\,1\,0\,0\,1}$

ผลการประเมินของผู้ประเมินกลุ่มที่ 2 $\mathbf{0\,0\,0\,1\,0\,1\,0\,1\,1\,0\,0\,1\,0\,1\,1\,0\,0\,0\,1\,0\,1\,0\,1\,1\,0}$

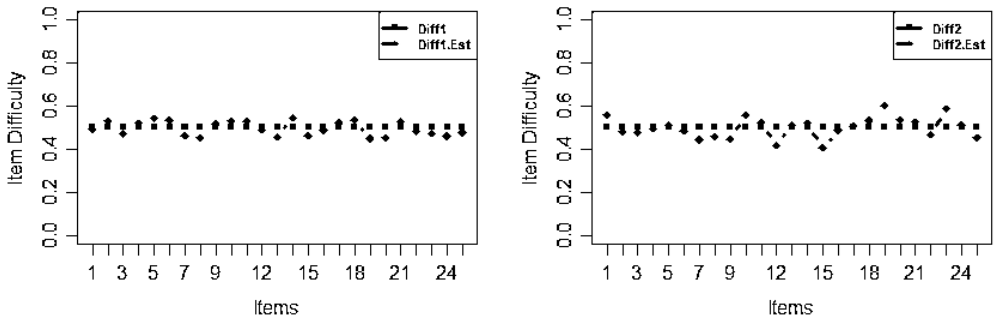
จากนั้น ทำการจำลองข้อมูลซ้ำ 500 ครั้ง ได้ผลการประมาณค่าจำนวน 421 หน่วย คิดเป็นร้อยละ 84.2 ของจำนวนรอบในการประมาณค่าทั้งหมด นำผลการประมาณค่าที่จำนวน 50 100 150 200 250 และ 300 รอบ มาคำนวณค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน (Pearson Correlation Coefficient) ค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสอง (Mean Square Error: MSE) ผลการทดสอบความคงที่ของการประมาณค่าฉันทามติเชิงวัฒนธรรมด้วยโมเดล MC-GCM พบว่า การกระทำซ้ำตั้งแต่ 50 ถึง 300 รอบ โดยค่าพารามิเตอร์และค่าประมาณของโมเดล MC-GCM มีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติ เมื่อกระทำซ้ำตั้งแต่ 50 รอบขึ้นไป การประมาณค่าในโมเดล MC-GCM ที่ข้อคำถามมีความยากเท่าเทียมกันมีความคงที่มากกว่า และให้ผลการประมาค่าที่มีประสิทธิภาพกว่าการประมาณค่าของโมเดลในกรณีที่คำถามประเมินมีความยากแตกต่างกัน ดังแสดงในตาราง 3.1

ตาราง 3.1 ค่าสัมประสิทธิ์สหสัมพันธ์และค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสองของการประมาณค่าฉันทามติเชิงวัฒนธรรมด้วยโมเดล MC-GCM กรณีไม่มีการทำหน้าที่ต่างกันของผู้ประเมิน

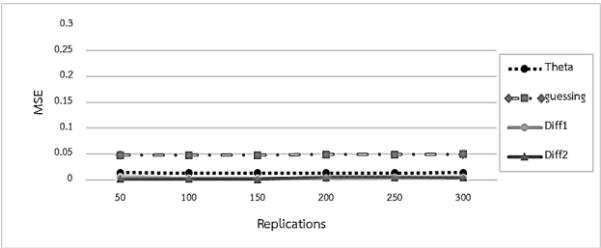
Param	Replication											
	50		100		150		200		250		300	
	Corr.	MSE	Corr.	MSE	Corr.	MSE	Corr.	MSE	Corr.	MSE	Corr.	MSE
θ	0.973**	0.014	0.971**	0.013	0.968**	0.013	0.968**	0.134	0.964**	0.129	0.963**	0.014

**p-value ≤ 0.01

ในส่วนของพารามิเตอร์คำตอบฉันทามติเชิงวัฒนธรรมของผู้ประเมินทั้งสองกลุ่ม และพารามิเตอร์ความยากของรายการประเมินนั้น จะเห็นได้ว่าโมเดล MC-GCM สามารถประมาณค่าพารามิเตอร์ความยากของรายการประเมิน (δ_i) ได้อย่างมีประสิทธิภาพและคงที่ในกรณีที่ไม่มีความแตกต่างกันระหว่างผู้ประเมิน ดังแสดงในรูป 3.2 และรูป 3.3 ผลการประมาณค่าคำตอบฉันทามติของผู้ประเมินกลุ่ม 1 และกลุ่ม 2 มีความสัมพันธ์กันในระดับสูง $(r_{Z1,Est_1} = 0.919, r_{Z2,Est_2} = 1.000)$ ผลการประมาณค่าความยากของรายการประเมินเมื่อกำหนด $\delta_i = 0.5$ พบว่าส่วนเบี่ยงเบนมาตรฐานของการประมาณค่า $SD_{\delta 1} = 0.133$ และ $SD_{\delta 2} = 0.150$ และมีค่าเฉลี่ยเท่ากับ $\bar{X}_{\delta_1} = 0.496, \bar{X}_{\delta_2} = 0.499$ ตามลำดับ



รูป 3.2 ความเบี่ยงเบนของค่าประมาณความยากของรายการประเมิน (δ_i)



รูป 3.3 ค่า MSE ของการประมาณค่าด้วยโมเดล MC-GCM
เมื่อกำหนดให้ไม่มีการทำหน้าที่ต่างกันของผู้ประเมิน

<u>การทดสอบความคงที่ของการประมาณค่ากรณีมีการทำหน้าที่ต่างกันของผู้ประเมิน</u>

ผู้วิจัยกำหนดจำนวนผู้ประเมิน $N = 15$ จำนวนรายการประเมิน $\mathrm{M} = 25$ จำนวนกลุ่มวัฒนธรรม $T = 2$ เวกเตอร์ความสามารถของผู้ประเมินมีค่าตั้งแต่ 0.025 ถึง 0.933 เวกเตอร์ความลำเอียงในการเดามีค่าตั้งแต่ 0.211 ถึง 0.830 เวกเตอร์ความยากของรายการประมินกลุ่มที่ 1 มีค่าตั้งแต่ 0.197 ถึง 0.482 เวกเตอร์ความยากของรายการประเมินกลุ่มที่ 2 มีค่าตั้งแต่ 0.535 ถึง 0.951 เวกเตอร์ฉันทามติของผู้ประเมินมีค่าดังนี้

ผลการประเมินของผู้ประเมินกลุ่มที่ 1 $\mathbf{0\,1\,0\,1\,0\,1\,1\,1\,1\,0\,1\,1\,1\,0\,1\,0\,0\,1\,0\,1\,0\,1\,0\,1\,1}$

ผลการประเมินของผู้ประเมินกลุ่มที่ 2 $\mathbf{1\,0\,1\,0\,1\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,1\,1\,0\,1\,0\,1\,0\,1\,0\,0}$

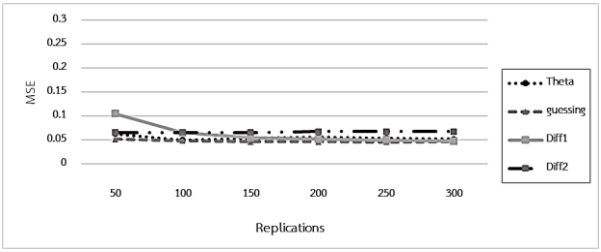
จากนั้น ทำการจำลองข้อมูลซ้ำ 500 ครั้ง ได้ผลการประมาณค่าจำนวน 347 หน่วย คิดเป็นร้อยละ 69.4 ของจำนวนรอบในการประมาณค่าทั้งหมด นำผลการประมาณค่าที่จำนวน 50 100 150 200 250 และ 300 รอบ มาคำนวณค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน (Pearson Correlation Coefficient) ค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสอง (Mean Square Error: MSE)

ผลการทดสอบความคงที่ของการประมาณค่าฉันทามติเชิงวัฒนธรรมด้วยโมเดล MC-GCM พบว่า การกระทำซ้ำตั้งแต่ 100 ถึง 300 รอบ ให้ผลการประมาณค่าใกล้เคียงกัน เมื่อพิจารณาจากค่าสหสัมพันธ์และค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสองจะเห็นว่าผลการประมาณค่ามีแนวโน้มคงที่ โดยค่าพารามิเตอร์และค่าประมาณของโมเดล MC-GCM มีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติ เมื่อกระทำซ้ำตั้งแต่ 50 รอบขึ้นไป ดังแสดงในตาราง 3.2 และรูป 3.4 แสดงผลการทดสอบความคงที่ของการประมาณค่าของโมเดล MC-GCM สรุปได้ว่า โมเดลสามารถประมาณค่าได้คงที่เมื่อกระทำซ้ำตั้งแต่ 50 รอบ ขึ้นไป อย่างไรก็ตาม เพื่อผลการประมาณค่าที่มีประสิทธิภาพ ผู้วิจัยจึงกำหนดการกระทำซ้ำการจำลองข้อมูลจำนวน 100 รอบ

ตาราง 3.2 ค่าสัมประสิทธิ์สหสัมพันธ์และค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสองของการประมาณค่าฉันทามติเชิงวัฒนธรรมด้วยโมเดล MC-GCM กรณีมีการทำหน้าที่ต่างกันของผู้ประเมิน

Param	Replications											
	50		100		150		200		250		300	
	Corr.	MSE	Corr.	MSE	Corr.	MSE	Corr.	MSE	Corr.	MSE	Corr.	MSE
θ	0.619*	0.063	0.735**	0.050	0.700**	0.054	0.673**	0.056	0.707**	0.054	0.728**	0.052
g	0.426	0.052	0.461	0.049	0.475	0.047	0.477	0.047	0.494	0.046	0.487	0.047
δ_1	0.416**	0.105	0.453**	0.065	0.475**	0.055	0.467**	0.052	0.501**	0.050	0.541**	0.048
δ_2	0.309	0.066	0.412**	0.066	0.438**	0.066	0.414**	0.068	0.430**	0.068	0.433**	0.068

*p-value ≤ 0.05, **p-value ≤ 0.01



รูป 3.4 ค่า MSE ของการประมาณค่าด้วยโมเดล MC-GCM
เมื่อกำหนดให้มีการทำหน้าที่ต่างกันของผู้ประเมิน

2. การตรวจสอบความถูกต้องของการจำลองข้อมูลในโมเดล MC-LTRM

เนื่องจากโมเดล MC-LTRM เป็นโมเดลที่ใช้ในการวิเคราะห์ข้อมูลจริงในการศึกษาครั้งนี้ และการประมาณค่าของโมเดล MC-LTRM มีการประมาณค่าพารามิเตอร์จำนวนมากทำให้ใช้เวลานานในการวิเคราะห์ข้อมูล ผู้วิจัยจึงจำเป็นต้องทำการทดสอบความถูกต้องของการประมาณค่าเพื่อตรวจสอบความถูกต้องและแม่นยำของการประมาณค่าด้วยจำนวนรอบการทำซ้ำที่น้อยที่สุดที่มีประสิทธิภาพในการประมาณค่ามากที่สุด โดยการจำลองข้อมูลที่มีลักษณะใกล้เคียงกับข้อมูลจริง คือ เป็นข้อมูลการประเมินที่มีผู้ประเมินจำนวน 20 คน และมีรายการประเมินจำนวน 40 ข้อ ทำการทดสอบเฉพาะกรณีการประเมินที่ข้อคำถามมีความยากเท่าเทียมกันเพื่อให้สอดคล้องกับการประเมินจากข้อมูลจริง จากนั้น ตรวจสอบความถูกต้องและแม่นยำของการประเมินโดยคำนวณค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน (Pearson Correlation Coefficient) ค่าความลำเอียงในการประมาณค่า (Bias of estimators) และค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสอง (Mean Square Error: MSE)

การกำหนดจำนวนการทำซ้ำในการจำลองข้อมูล

ผลการจำลองข้อมูลจำนวน 300 รอบ พบว่าโมเดล MC-LTRM สามารถประมาณค่าได้อย่างมีประสิทธิภาพตั้งแต่ 50 รอบ ขึ้นไป และประมาณค่าได้คงที่เมื่อทำซ้ำตั้งแต่ 100 รอบขึ้นไป อย่างไรก็ดี ผลการประมาณค่าจากการทำซ้ำ 50 ถึง 300 รอบให้ค่าความถูกต้องและแม่นยำที่ใกล้เคียงกัน โดยมีค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าจริงกับค่าที่ได้จากการประมาณค่าของโมเดลระหว่าง 0.984 ถึง 1.00 ดังรายละเอียดในตาราง 3.3 ในกรณีที่ไม่มีการทำหน้าที่ต่างกันของผู้ประเมิน และตาราง 3.4 ในกรณีที่มีการทำหน้าที่ต่างกันของผู้ประเมิน

ตาราง 3.3 ค่าสัมประสิทธิ์สหสัมพันธ์และ MSE ของการประมาณค่าฉันทามติเชิงวัฒนธรรมด้วยโมเดล LTRM กรณีที่ไม่มีการทำหน้าที่ต่างกันของผู้ประเมิน

Param	Replications											
	50		100		150		200		250		300	
	Corr.	MSE	Corr.	MSE	Corr.	MSE	Corr.	MSE	Corr.	MSE	Corr.	MSE
E	0.989**	0.004	0.992**	0.003	0.995**	0.001	0.997**	0.000	0.998**	0.000	0.998**	0.000
T	0.984**	0.023	0.993**	0.007	0.992**	0.008	0.994**	0.006	0.997**	0.000	0.996**	0.000
λ	0.984**	0.001	0.991*	0.000	0.963**	0.000	0.997*	0.000	0.999**	0.000	1.000**	0.000
a	0.998**	0.001	0.998**	0.000	0.998**	0.000	0.999**	0.000	0.999**	0.000	0.999**	0.000
b	00.993**	0.001	0.995**	0.001	0.995**	0.000	0.996**	0.000	0.997**	0.000	0.997**	0.000

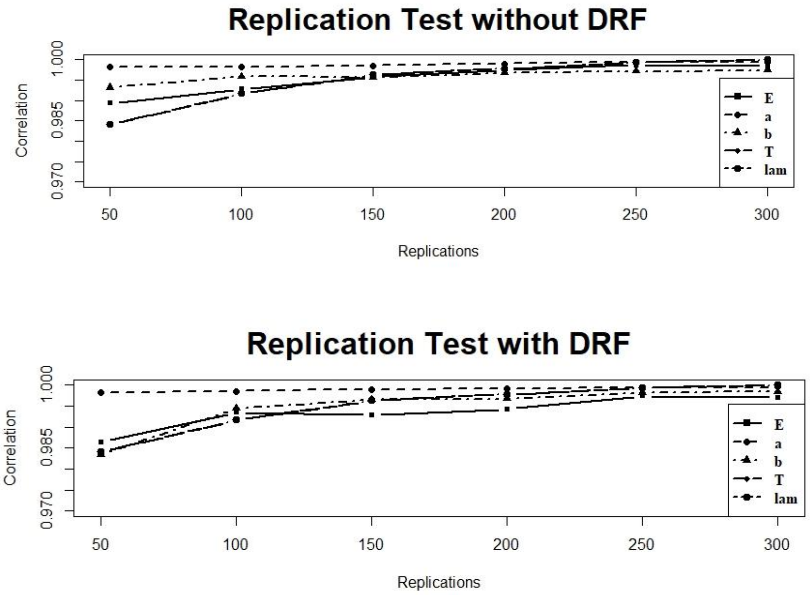
**p-value ≤ 0.01

ตาราง 3.4 ค่าสัมประสิทธิ์สหสัมพันธ์และ MSE ของการประมาณค่าฉันทามติเชิงวัฒนธรรมด้วยโมเดล LTRM กรณีมีการทำหน้าที่ต่างกันของผู้ประเมิน

Param	Replications											
	50		100		150		200		250		300	
	Corr.	MSE	Corr.	MSE	Corr.	MSE	Corr.	MSE	Corr.	MSE	Corr.	MSE
E	0.986**	0.025	0.993**	0.001	0.992**	0.003	0.994**	0.002	0.997**	0.000	0.996**	0.000
T	0.984**	0.023	0.991**	0.007	0.996**	0.008	0.997**	0.006	0.999**	0.000	1.000**	0.000
λ	0.984**	0.001	0.991*	0.000	0.996**	0.000	0.997*	0.000	0.999**	0.000	1.000**	0.000
a	0.998**	0.001	0.998**	0.001	0.999**	0.000	0.999**	0.000	0.999**	0.000	0.999**	0000
b	0.983**	0.001	0.994**	0.001	0.996**	0.000	0.996**	0.000	0.998**	0.000	0.998**	0.000

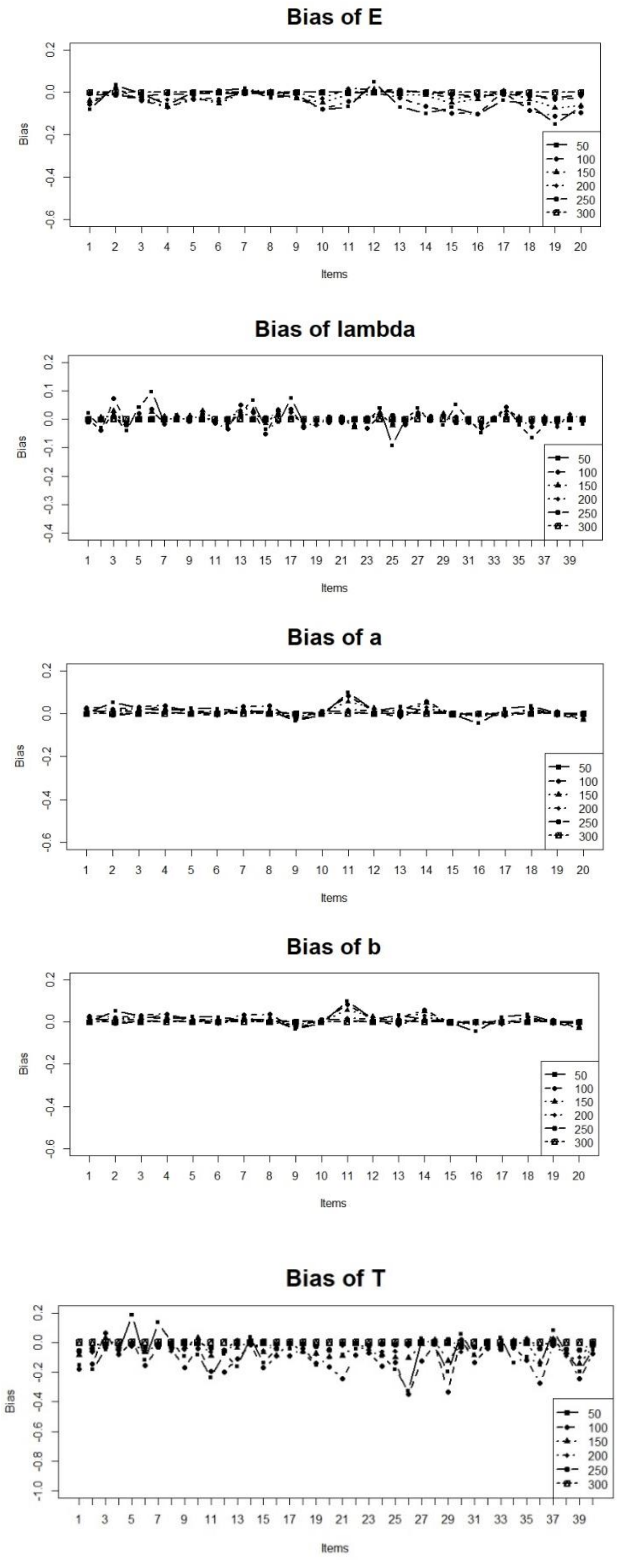
**p-value ≤ 0.01

เมื่อพิจารณาจากรูป 3.5 จะเห็นว่าโมเดลประมาณค่าพารามิเตอร์ได้แม่นยำมากกว่าในกรณีที่ไม่มีการทำหน้าที่ต่างกันของผู้ประเมิน โดยมีค่าประมาณคงที่ตั้งแต่การทำซ้ำรอบที่ 100 เป็นต้นไป ในขณะที่กรณีที่มีการทำหน้าที่ต่างกันของผู้ประเมินจะมีความไม่เสถียรของการประมาณค่าความสามารถของผู้ประเมิน



รูป 3.5 สัมประสิทธิ์สหสัมพันธ์ระหว่างค่าจริงกับค่าประมาณของโมเดล

รูป 3.6 แสดงผลความลำเอียงในการประมาณค่าของพารามิเตอร์ทั้ง 5 พารามิเตอร์ จะเห็นได้ว่า ในการประมาณค่าพารามิเตอร์ตั้งแต่รอบที่ 50 ถึงรอบที่ 300 มีค่าระหว่าง -0.2 ถึง 0.2 ซึ่งเป็นค่าที่เข้าใกล้ 0 และมีความเที่ยงตรงในการประมาณค่าใกล้เคียงกันทุกรอบการทำซ้ำ มีข้อสังเกตในการประมาณค่าพารามิเตอร์ฉันทามติเชิงวัฒนธรรม (T_k) ที่การประมาณค่าไม่แม่นยำในการทำซ้ำรอบที่ 50 และ 100 โดยมีค่าระหว่าง -0.4 ถึง 0.2 อย่างไรก็ตาม เมื่อพิจารณาจากผลการประมาณค่าพารามิเตอร์ที่มีความสัมพันธ์กันในระดับสูงอย่างมีนัยสำคัญทางสถิติในทุกรอบการประเมินแล้ว สามารถสรุปได้ว่าการทำซ้ำจำนวน 50 รอบ สามารถให้ผลการประมาณค่าที่มีประสิทธิภาพ ดังนั้น ในการศึกษาประสิทธิภาพของการประมาณค่าของโมเดล MC-LTRM จากการจำลองข้อมูลในครั้งนี้ ผู้วิจัยจึงกำหนดการกระทำซ้ำของข้อมูลจำลองจำนวน 50 รอบ



รูป 3.6 ความลำเอียงในการประมาณค่ามารามิเตอร์ของโมเดล

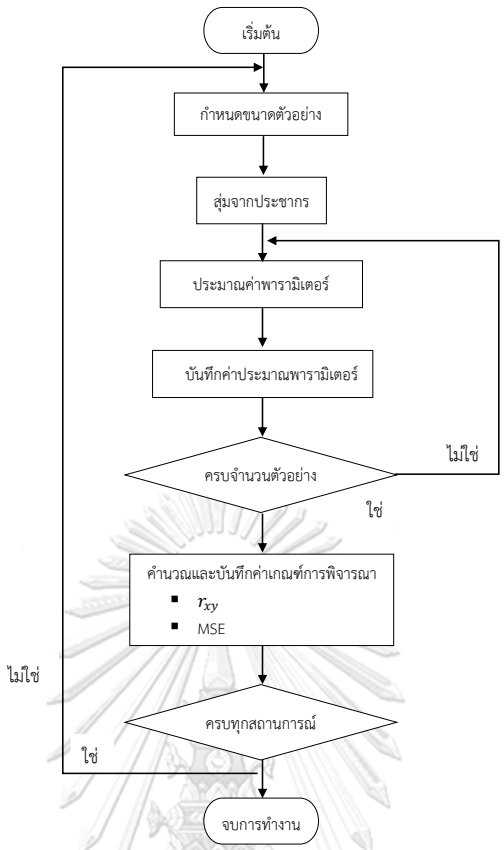
การวิเคราะห์ผลการจำลองข้อมูล

ผู้วิจัยวิเคราะห์ผลการประมาณค่าพารามิเตอร์จากข้อมูลจำลองจำนวน 4 พารามิเตอร์ ได้แก่ ฉันทามติเชิงวัฒนธรรม ($\mathrm{Z_k}$) ความสามารถของผู้ประเมิน (θ_i) ความยากของรายการประเมิน (λ_k) และความลำเอียงในการประเมิน (g_i) โดยกำหนดเงื่อนไขการจำลองข้อมูลที่แตกต่างกัน 18 สถานการณ์สำหรับการประมาณค่าในหนึ่งโมเดล ตัวแปรอิสระที่ศึกษา คือ จำนวนผู้ประเมิน (N) แบ่งเป็น 15 30 และ 45 คน จำนวนรายการประเมิน (M) แบ่งเป็น 25 55 และ 85 ข้อ และการทำหน้าที่ต่างกันของผู้ประเมิน แบ่งเป็น กรณีที่ไม่มีการทำหน้าที่ต่างกันของผู้ประเมิน กำหนดให้ข้อคำถามมีความยากเท่าเทียมกัน และกรณีมีการทำหน้าที่ต่างกันของผู้ประเมิน กำหนดให้ข้อคำถามมีความยากแตกต่างกัน ตัวแปรตาม คือ ประสิทธิภาพของการประมาณค่าของโมเดล MC-GCM ได้แก่ ค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (MSE) ค่าความลำเอียงในการประมาณค่า (Bias) และค่าสัมประสิทธิ์สหสัมพันธ์ (Pearson Correlation Coefficient) ดังรูป 3.7 การวิเคราะห์ข้อมูลประกอบด้วย 2 ส่วน ดังนี้

1. การวิเคราะห์สถิติบรรยาย ได้แก่ ค่าต่ำสุด ค่าสูงสุด ค่าเฉลี่ย (mean) พิสัย (range) โดยนำเสนอในรูปของตารางและกราฟ เพื่อศึกษาแนวโน้มของความถูกต้องและความแม่นยำของการประมาณค่าพารามิเตอร์ของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม

2. การวิเคราะห์ขนาดอิทธิพล (R - squared) เป็นการวิเคราะห์ค่าความผันแปรของตัวแปรอิสระ ได้แก่ ค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (MSE) ค่าความลำเอียงในการประมาณค่า (Bias) และค่าสัมประสิทธิ์สหสัมพันธ์ (Pearson Correlation Coefficient) ที่อธิบายโดยตัวแปรต้น ได้แก่ จำนวนผู้ประเมิน (N) จำนวนแบบสอบ (M) และการทำหน้าที่ต่างกันของผู้ประเมิน เพื่อศึกษาปัจจัยที่ส่งผลต่อประสิทธิภาพในการประมาณค่าของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY



รูป 3.7 การวิเคราะห์ข้อมูลจำลอง

ระยะที่ 3 ศึกษาผลการประยุกต์ใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมในการวิเคราะห์ความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบ ในการประเมินระดับชั้นเรียนกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น

กลุ่มเป้าหมาย

ข้อมูลที่ใช้ในการศึกษาครั้งนี้เป็นข้อมูลทุติยภูมิจากการศึกษาเรื่อง *ความสอดคล้องในแนวเดียวกันระหว่างข้อสอบในการประเมินระดับชาติกับข้อสอบในการประเมินระดับชั้นเรียน กลุ่มสาระการเรียนรู้วิทยาศาสตร์: การประยุกต์ใช้โมเดลหลายองค์ประกอบของราส์ช และทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัด* ศึกษาโดย บุษยารัตน์ จันทร์ประเสริฐ (2560) ประกอบด้วยข้อมูล 2 ส่วน คือ

1. ผลการประเมินจากแบบบันทึกระดับความซับซ้อนทางปัญญาของข้อสอบในการประเมินระดับชาติ กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับชั้นมัธยมศึกษาปีที่ 3 เป็นการประเมินลำดับขั้น

ของกระบวนการทางปัญญาจำแนกตามพฤติกรรมการเรียนรู้ด้านพุทธิพิสัยของ Bloom (Bloom's Taxonomy) ผลการประเมินเป็นการจัดกลุ่มตัวบ่งชี้ตามพฤติกรรมการเรียนรู้ 6 กลุ่ม ได้แก่ จำ เข้าใจ ประยุกต์ใช้ วิเคราะห์ ประเมินค่า และสร้างสรรค์

2. ผลการประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียน กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น เป็นการประเมินความสอดคล้องระหว่างข้อสอบกับตัวชี้วัด จำนวน 40 ข้อ ผลการประเมินเป็นมาตรประมาณค่า 5 ระดับ ได้แก่ ไม่สอดคล้อง ค่อนข้างไม่สอดคล้อง ไม่แน่ใจ ค่อนข้างสอดคล้อง สอดคล้องโดยตรง

การวิเคราะห์ข้อมูลในการศึกษาครั้งนี้ ผู้วิจัยใช้ผลการประเมินความสอดคล้องในแนวเดียวกันฯ จำนวน 40 ข้อ ที่ประเมินโดยผู้ประเมิน จำนวน 20 คน

การวิเคราะห์ข้อมูล

การวิเคราะห์ข้อมูลใช้โมเดลฉันทามติเชิงวัฒนธรรมสำหรับการวิเคราะห์ข้อมูลแบบจัดกลุ่มและเรียงอันดับ (LTRM) เพื่อ 1) ศึกษาผลการประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียน กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น 2) วิเคราะห์การทำหน้าที่ต่างกันระหว่างผู้ประเมิน 3) ศึกษารูปแบบความลำเอียงในการประเมินของผู้ประเมิน ทั้งนี้ โมเดล LTRM จะประมาณค่าพารามิเตอร์คะแนนฉันทามติของการประเมิน (T) ความสามารถของผู้ประเมิน (E) ความยากของรายการประเมิน (λ) และความลำเอียงในการประเมิน (a, b) โดยใช้วิธีการประมาณค่าทางสถิติแบบเบส์ และรายงานผลการวิเคราะห์ข้อมูลในรูปแบบแผนภูมิและกราฟ

บทที่ 4
ผลการวิเคราะห์ข้อมูล

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อ 1) เพื่อประยุกต์ใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมเพื่อวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินในการศึกษาความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียนกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น 2) เพื่อตรวจสอบประสิทธิภาพของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมในการการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินในการศึกษาความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียนกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น 3) เพื่อศึกษาผลการประยุกต์ใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมในการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินและการทำหน้าที่ต่างกันของผู้ประเมินในการประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียนกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น การตัดสินประสิทธิภาพของโมเดลพิจารณาจาก 1) ความสอดคล้องระหว่างค่าที่กำหนดกับค่าที่ได้จากการประมาณค่าของโมเดล โดยพิจารณาค่าสัมประสิทธิ์สหสัมพันธ์ (Pearson correlation coefficient) 2) ความถูกต้องและน่าเชื่อถือของผลการประมาณค่า โดยพิจารณาจาก ความลำเอียงในการประมาณค่า (Bias of estimators) และค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Mean Square Error: MSE)

ผลการวิเคราะห์ข้อมูลแบ่งเป็น 3 ตอน ได้แก่ ตอนที่ 1 ผลการพัฒนากระบวนการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินสำหรับการวิเคราะห์ความสอดคล้องในแนวเดียวกันทางการศึกษาด้วยโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม ตอนที่ 2 ผลการวิเคราะห์ประสิทธิภาพของการประมาณค่าของโมเดลจากข้อมูลจำลอง ประกอบด้วย 1) ประสิทธิภาพของการประมาณค่าของโมเดล MC-GCM 2) ประสิทธิภาพของการประมาณค่าของโมเดล MC-LTRM และตอนที่ 3 การวิเคราะห์ความเที่ยงระหว่างผู้ประเมินโดยใช้ข้อมูลจริง แบ่งเป็น 2 ส่วน คือ 1) ขั้นตอนการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินโดยโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม 2) ผลการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมิน

ตอนที่ 1 ผลการพัฒนากระบวนการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินสำหรับการวิเคราะห์ความสอดคล้องในแนวเดียวกันทางการศึกษาด้วยโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม

การวิเคราะห์ฉันทามติเชิงวัฒนธรรม (Cultural Consensus Analysis) เป็นวิธีการวิเคราะห์ข้อมูลแบบจัดกลุ่มและแบบเรียงอันดับ โดยวิเคราะห์ข้อมูลซึ่งอยู่ในรูปของเมตริกซ์ $N\ x\ M$ เมื่อ N คือ ผู้ประเมิน และ M คือ คำถามประเมิน โดยโมเดลจะประมาณค่าตำแหน่งคะแนนของคำถามประเมินแต่ละข้อซึ่งได้จากการให้คะแนนของผู้ประเมินออกมาเป็นคะแนนฉันทามติของการประเมิน (Consensus score)

การวิเคราะห์ฉันทามติเชิงวัฒนธรรมตามแนวคิดของ Romney และคณะ (1986) มีแนวคิดว่า คำถามแต่ละคำถามมีคำตอบอยู่ในตัวเอง ซึ่งคำตอบนั้นแตกต่างกันไปตามบริบทของผู้ตอบ ผู้ตอบที่มีความรู้ ความเชื่อ แนวคิด หรือเติบโตภายใต้วัฒนธรรมที่ต่างกันจะมีคำตอบที่แตกต่างกัน ในทางกลับกัน หากผู้ตอบมีภูมิหลังใกล้เคียงกัน แนวโน้มของคำตอบก็จะสอดคล้องกันมากกว่า การนำแนวคิดการวิเคราะห์ฉันทามติเชิงวัฒนธรรมมาประยุกต์ใช้กับการประเมินความสอดคล้องในแนวเดียวกันในการศึกษานี้ มีข้อตกลงเบื้องต้นเกี่ยวกับการนำโมเดลไปใช้ ดังต่อไปนี้ 1) ผู้ประเมินเป็นสมาชิกกลุ่มวัฒนธรรมเดียวกัน กล่าวคือ ผู้ประเมินแต่ละคนมีภูมิหลังที่เกี่ยวข้องกับการประเมิน ได้แก่ ความรู้ความเชี่ยวชาญในหัวข้อที่ประเมิน และความเป็นกลางในการประเมินใกล้เคียงกัน และระบุตำแหน่งคะแนนการประเมินในแต่ละข้อคำถามประเมินได้สอดคล้องกับคะแนนฉันทามติของกลุ่ม และ 2) คำถามประเมินมีการตอบในรูปของข้อมูลจัดกลุ่ม หรือเรียงอันดับ

กรอบแนวคิดของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม

การวิเคราะห์ฉันทามติเชิงวัฒนธรรมมีแนวคิดคล้ายกับโมเดลการวัดแบบดั้งเดิม คือ ผลการตอบหรือผลการปะเมินของผู้ประเมินมาจากผลรวมของคะแนนจริงกับความคลาดเคลื่อน $Y_{ik} = T_k + \epsilon_{ik}$ และมีความใกล้เคียงกับทฤษฎีการตอบสนองข้อสอบในแง่ของความสัมพันธ์ระหว่างความน่าจะเป็นที่ผู้ประเมินจะให้คะแนนการประเมินได้ตรงกับผลการประเมินของกลุ่ม (ตอบถูก) กับปัจจัยที่เกี่ยวข้องกับการให้คะแนนในการประเมิน (ความเชี่ยวชาญ ความลำเอียง) ความแตกต่างระหว่างโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมกับโมเดลการตอบสนองข้อสอบ คือ โมเดลการตอบสนองข้อสอบเทียบความน่าจะเป็นในการตอบถูกของผู้สอบโดยเทียบผลการตอบกับเฉลยที่มี

อยู่แล้ว แต่โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมเป็นการศึกษาความน่าจะเป็นในการ “ตอบถูก” ของผู้ประเมินโดยเทียบกับค่าเฉลี่ยความน่าจะเป็นภายหลังของการตอบของกลุ่มผู้ประเมินทั้งหมด

สารสนเทศหลักของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม คือ คะแนนฉันทามติของการประเมิน (Consensus score) ซึ่งเป็นผลมาจากการประมาณค่าพารามิเตอร์ที่เกี่ยวข้องกับการให้คะแนนในการประเมินของผู้ประเมิน โดยมีรายละเอียดที่แตกต่างกันในแต่ละโมเดลการวิเคราะห์ดังต่อไปนี้

พารามิเตอร์ในการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินด้วยโมเดล MC-GCM

กรอบแนวคิดในการวิเคราะห์ข้อมูลของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม GCM ในรูป 4.1 แสดงพารามิเตอร์ของโมเดล MC-GCM มีรายละเอียด ดังนี้

1. การทำหน้าที่ต่างกันของผู้ประเมิน (T) เป็นจำนวนความแตกต่างของรูปแบบการประเมินที่สอดคล้องกันของกลุ่มผู้ประเมิน ซึ่งได้จากการคำนวณจำนวนองค์ประกอบซึ่งระบุโดยค่าไอเกน ค่าไอเกนเป็นข้อมูลเบื้องต้นให้นักวิจัยเลือกโมเดลการวิเคราะห์ในขั้นตอนต่อไปได้อย่างเหมาะสม กล่าวคือ หากค่าไอเกนแสดงองค์ประกอบของผลการประเมินว่ามีองค์ประกอบเดียว หมายถึงผลการประเมินในครั้งนั้นไม่มีการทำหน้าที่ต่างกันของผู้ประเมิน ซึ่งนักวิจัยจะทำการเคราะห์ข้อมูลต่อไปด้วยโมเดล GCM หากค่าไอเกนแสดงจำนวนองค์ประกอบมากกว่า 1 องค์ประกอบ นักวิจัยจะต้องเลือกโมเดลสำหรับการวิเคราะห์การทำหน้าที่แตกต่างกันระหว่างผู้ประเมินด้วย นั่นคือนักวิจัยต้องใช้โมเดล MC-GCM ในการวิเคราะห์ข้อมูลในขั้นต่อไป

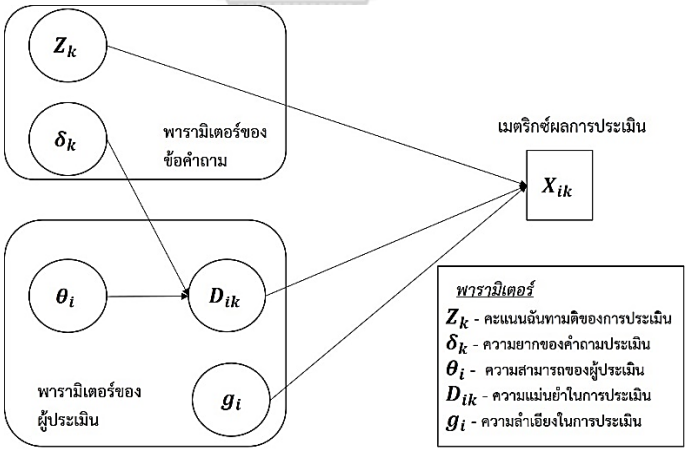
2. คะแนนฉันทามติของการประเมิน (Z_k) เป็นข้อสรุปของผลคะแนนในการประเมินของผู้ประเมิน ซึ่งแสดงถึงความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับมาตรฐานและตัวชี้วัด ในกรณีของโมเดล MC-GCM หรือ โมเดล GCM คะแนนฉันทามติระหว่าผู้ประเมินจะมีค่าในช่วง (0,1) $Z_{k1} = 1$ หมายถึง คะแนนประเมินที่เป็นฉันทามติของรายการประเมินข้อที่ 1 มีค่าเท่ากับ 1 หรือหมายถึง ผู้ประเมินมีความเห็นว่ารายการประเมินในข้อดังกล่าวมีความสอดคล้องในแนวเดียวกัน อย่างไรก็ตาม เนื่องจากค่าพารามิเตอร์ Z_k เป็นค่าเฉลี่ยของการแจกแจงความน่าจะเป็นภายหลังของพารามิเตอร์ในช่วง 0 ถึง 1 ดังนั้น การแปลความหมายของสารสนเทศที่ได้จะอยู่ในรูปของความน่าจะเป็นที่คะแนนฉันทามติในการประเมินในข้อที่ i จะเท่ากับ 1 หรือหมายถึง “ค่าเฉลี่ยของการแจกแจงความน่าจะเป็นภายหลังของผลการประเมินความสอดคล้องในแนวเดียวกันระหว่างหลักสูตรและ

การประเมินผลในรายการประเมินข้อที่ 1 มีความสอดคล้องในแนวเดียวกันระหว่างหลักสูตรกับการวัดและประเมินผล”

3. ความแม่นยำในการประเมิน (D_{ik}) เป็นความน่าจะเป็นที่ผู้ประเมินคนที่ i จะให้คะแนนการประเมินในรายการประเมินข้อ k ตรงกับผลการประเมินที่เป็นฉันทามติของกลุ่ม พารามิเตอร์ความสามารถของผู้ประเมินคำนวณผ่านพารามิเตอร์ความยากของคำถามประเมิน (δ_k) และพารามิเตอร์ความสามารถของผู้ประเมิน (θ_i) ซึ่งเป็นมโนทัศน์เดียวกันกับทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) ซึ่ง Batchelder และคณะ (2018) นำสมการของราชส์โมเดลมาใช้ในการประมาณค่าพารามิเตอร์ดังกล่าว โดย

$$D_{ik} = \frac{\theta_i(1-\delta_k)}{\theta_i(1-\delta_k)+\delta_k(1-\theta_i)}$$

4. ความลำเอียงในการให้คะแนนการประเมิน (g_i) เป็นพารามิเตอร์ที่สะท้อนแนวโน้มในการให้คะแนนเมื่อผู้ประเมินไม่แน่ใจหรือขาดความเชี่ยวชาญในการประเมินบางหัวข้อแต่จำเป็นต้องตัดสินคะแนน พารามิเตอร์ดังกล่าวเป็นการระบุความน่าจะเป็นที่ผู้ประเมิน i จะให้คะแนนการประเมินในรายการประเมินข้อ k เท่ากับ 1 เมื่อมีความไม่แน่ใจในการให้คะแนนการประเมิน



รูป 4.1 กรอบแนวคิดของโมเดล GCM

พารามิเตอร์ในการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินด้วยโมเดล MC-LTRM

กรอบแนวคิดในการวิเคราะห์ข้อมูลของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม LTRM แสดงในรูป 4.2 ประกอบด้วยพารามิเตอร์ 2 กลุ่ม ได้แก่ 1) พารามิเตอร์ของข้อคำถาม ประกอบด้วยพารามิเตอร์ตำแหน่งคะแนนฉันทามติของการประเมิน ($\mathrm{T_k}$) และความยากของรายการประเมิน (λ_k) 2) พารามิเตอร์ของผู้ประเมิน ประกอบด้วยพารามิเตอร์ความสามารถของผู้ประเมิน (E_i) และพารามิเตอร์ความลำเอียงในการประเมิน (a_i, b_i)

พารามิเตอร์ของข้อคำถามกับพารามิเตอร์ความแม่นยำในการประเมินของผู้ประเมินจะส่งผลต่อการให้คะแนนประเมินของผู้ประเมินในเมตริกซ์คุณลักษณะแฝงของผู้ประเมิน (Y_{ik}) ส่วนพารามิเตอร์ความลำเอียงกับพารามิเตอร์จำนวนตัวเลือก/ระดับคะแนน (γ_c) จะส่งผลต่อการตัดสินใจระบุตำแหน่งของการให้คะแนนประเมินในสเกล ซึ่งค่าพารามิเตอร์ทั้งหมดจะสะท้อนในเมตริกซ์การให้คะแนนประเมิน X_{ik} ซึ่งเป็นผลคะแนนของผู้ประเมินคนที่ i ในการประเมินรายการประเมินข้อที่ k รายละเอียดของพารามิเตอร์ของโมเดลมีดังนี้

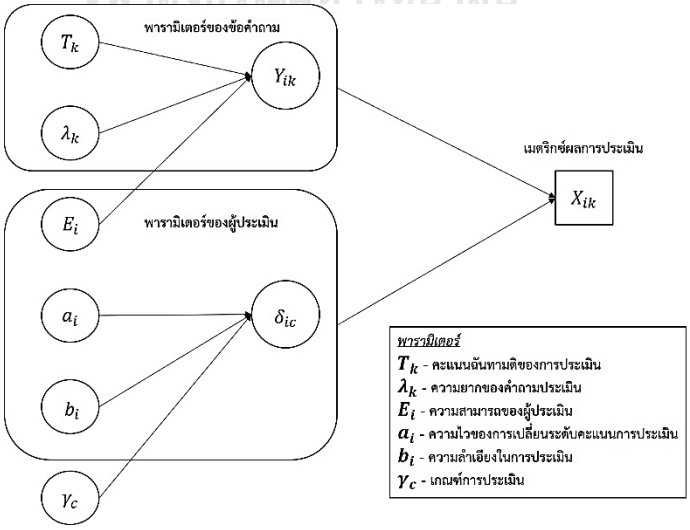
1. การทำหน้าที่ต่างกันของผู้ประเมิน (V) เป็นการระบุจำนวนของความแตกต่างของรูปแบบในการประเมินของผู้ประเมินจากการคำนวณค่าไอเกน หากค่าไอเกนแสดงว่าไม่มีการทำหน้าที่ต่างกันของผู้ประเมิน นักวิจัยจะเลือกใช้โมเดล LTRM ในการวิเคราะห์ข้อมูล แต่ถ้าค่าไอเกนแสดงจำนวนรูปแบบการประเมินที่มากกว่า 1 กลุ่ม นักวิจัยจะเลือกใช้โมเดล MC-LTRM ในการวิเคราะห์ข้อมูล

2. คะแนนฉันทามติของการประเมิน (T_{vk}) เป็นการประมาณค่าพารามิเตอร์เทรชโฮลด์ร่วมระหว่างผู้ประเมิน (γ_c) ควบคู่กับผลการประมาณค่าพารามิเตอร์ T_{vk} หรือ T_k (ในกรณีของโมเดล LTRM) พารามิเตอร์เทรชโฮลด์ร่วมระหว่างผู้ประเมิน (γ_c) เป็นพารามิเตอร์ที่แสดงแนวโน้มการเปลี่ยนระดับการให้คะแนนของสเกลการตอบ เมื่อ c เป็นจำนวนสเกล ในขณะที่พารามิเตอร์คะแนนฉันทามติระหว่างผู้ประเมิน (T_{vk}) เป็นพารามิเตอร์ที่ระบุค่าตำแหน่งของผลการประเมิน (item location values) บนเส้นจำนวนจริง โดย $T_v \in \prod_{k=1}^{M}(-\infty, \infty)$ เมื่อ v คือ กลุ่มการทำหน้าที่ต่างกันของผู้ประเมิน และ M คือ จำนวนคำถามประเมิน โดยผู้ประเมินแต่ละคนจะเป็นสมาชิกของกลุ่มการทำหน้าที่เพียง 1 กลุ่ม (Anders และ Batchelder, 2015) สารสนเทศของการประมาณค่าพารามิเตอร์คำตอบฉันทามติเชิงวัฒนธรรม เป็นการระบุตำแหน่งของผลการประเมินที่ได้จากการประมาณค่าเมตริกซ์การให้คะแนนประเมินของผู้ประเมิน (X_{ik}) ซึ่งต้องพิจารณาร่วมกับค่า γ_c ว่ารายการประเมินข้อใดตกอยู่ในตำแหน่งใดของสเกลการประเมิน

3. ความยากของรายการประเมิน (λ_k) เป็นการประมาณค่าว่าคำถามประเมินในข้อใดที่ส่งผลต่อความแม่นยำในการประเมินของผู้ประเมิน จากบทความของ Anders และ Batchelder (2015) หากเฉลี่ยภายหลังของการประมาณค่า (posterior mean) พารามิเตอร์ความยากของรายการประเมินมีค่าระหว่างครึ่งหนึ่งถึงสองเท่าของค่าเฉลี่ยภายหลังของการประมาณค่าพารามิเตอร์ความสามารถของผู้ประเมิน ถือว่ารายการประเมินนั้นมีระดับความยากของการประเมินที่เหมาะสม

4. ความสามารถของผู้ประเมิน (E_i) เป็นความแม่นยำในการประเมินระบุตำแหน่งคะแนนบนสเกลการประเมินของผู้ประเมินแต่ละคน หากค่า E_i ของผู้ประเมินคนใดค่าสูงแสดงว่าผู้ประเมินคนนั้นมีความแม่นยำในการประเมินได้ตรงกับฉันทามติมากกว่าผู้ประเมินคนอื่น

5. ความลำเอียงในการประเมิน $(a_i,\ b_i)$ แสดงความน่าจะเป็นในการให้คะแนนการประเมินเมื่อผู้ประเมินมีความไม่แน่ใจในการระบุคะแนนซึ่งอาจเป็นผลจากความไม่ชัดเจนของข้อคำถามหรือมีอคติต่อสิ่งที่ต้องประเมิน พารามิเตอร์ a_i เป็นการบอกแนวโน้มว่าผู้ประเมินจะเลือกประเมินในตำแหน่งซ้าย กลาง หรือขวาของสเกล ในขณะที่พารามิเตอร์ b_i บอกแนวโน้มว่าผู้ประเมินจะเลือกขยับตำแหน่งการประเมินไปทางบวกหรือลบ เมื่อนำข้อมูลจากการวิเคราะห์ข้อมูลมาพล็อตตำแหน่งของพารามิเตอร์ทั้งสองเข้าด้วยกัน $(a_i,\ b_i)$ จะสามารถมองเห็นรูปแบบของการประเมินของผู้ประเมินแต่ละคนได้ เทรชโฮลด์ความลำเอียงของผู้ประเมินคำนวณจากพารามิเตอร์ขอบเขตของสเกล (category boundaries) $\delta_{ic} = a_i\,\gamma_{\Omega_i c} + b_i$ เมื่อ $A = (a_i)_{1xN}$ ในช่วง $(0, \infty)$ และ $B = (b_i)_{1xN}$ ในช่วง $(-\infty, \infty)$



รูป 4.2 กรอบแนวคิดของโมเดล LTRM

ขั้นตอนการวิเคราะห์ความเที่ยงระหว่างผู้ประเมินโดยโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม ประกอบขั้นตอน ดังนี้

1. การตรวจสอบการทำหน้าที่ต่างกันระหว่างผู้ประเมิน

สิ่งที่ต้องทำเป็นอันดับแรกในการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินด้วยโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม คือ การตรวจสอบการทำหน้าที่ต่างกันระหว่างผู้ประเมิน เพื่อเลือกโมเดลการวิเคราะห์ที่เหมาะสมกับข้อมูล โดยพิจารณาค่าไอเกนของจำนวนองค์ประกอบที่แสดงถึงจำนวนรูปแบบในการประเมิน หากมีจำนวนรูปแบบการประเมินที่สอดคล้องกันหรือเป็นฉันทามติเดียวกันจำนวน 1 กลุ่ม นักวิจัยสามารถวิเคราะห์ข้อมูลได้ด้วยโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมแบบปกติ หากค่าไอเกนแสดงจำนวนองค์ประกอบมากกว่า 1 กลุ่ม นักวิจัยจึงใช้การวิเคราะห์ด้วยโมเดลสำหรับการวิเคราะห์กลุ่มพหุวัฒนธรรม (Multicultural Model)

2. การวิเคราะห์พารามิเตอร์ความยากของรายการประเมิน

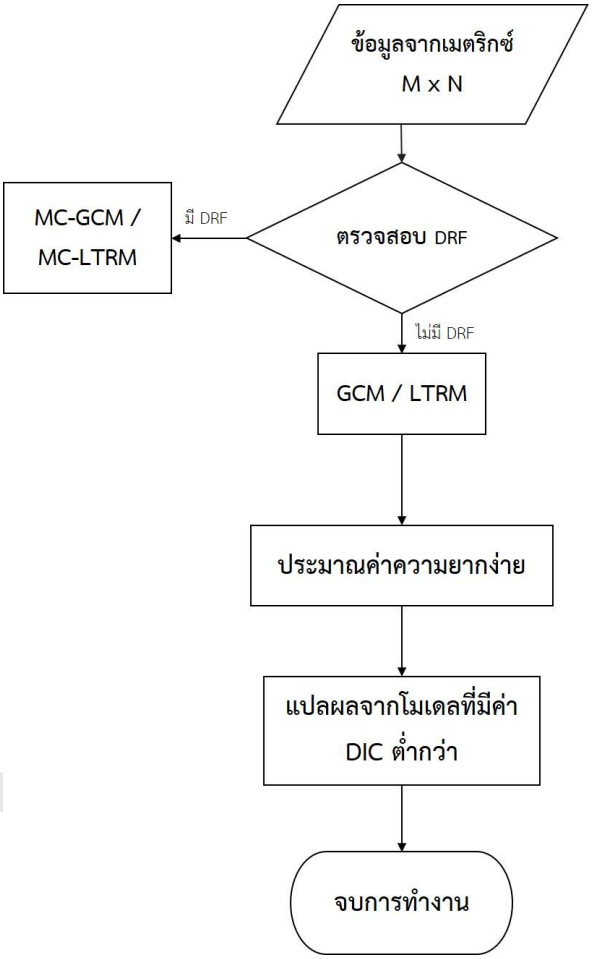
ข้อตกลงเบื้องต้นของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม คือ ความเป็นเอกพันธ์ของคำถาม (Item Homogeneity) อย่างไรก็ตาม จากบทความที่นำเสนอโมเดล LTRM ของ Batchelder และ Anders (2012) ได้เสนอสถิติ Variance Dispersion Index (VDI) ซึ่งทำหน้าที่สะท้อนความแตกต่างของข้อคำถามอันเนื่องมาจากความผันแปรของการตอบ โดยคำนวณความแปรปรวนในการให้คะแนนของผู้ประเมินทั้งหมด

3. การแปลผลและสรุปผลการวิเคราะห์

เมื่อเลือกโมเดลที่เหมาะสมกับข้อมูลแล้ว ผู้วิจัยแปลผลการวิเคราะห์ข้อมูลจากผลการประมาณค่าพารามิเตอร์ ทั้งนี้ โมเดล MC-GCM จะให้ผลการประมาณค่าพารามิเตอร์ 3 พารามิเตอร์ คือ 1) คะแนนฉันทามติของการประเมิน (Z_k) เป็นพารามิเตอร์ค่าเฉลี่ยของผลคะแนนจากการประเมินของผู้ประเมินในแต่ละข้อคำถาม 2) ความแม่นยำในการประเมิน (D_{ik}) เป็นพารามิเตอร์แสดงความถูกต้องแม่นยำในการประเมินของผู้ประเมิน ว่าผู้ประเมินสามารถประเมินได้สอดคล้องกับการประเมินของผู้ประเมินคนอื่นภายในกลุ่มหรือไม่ และ 3) ความลำเอียงในการให้คะแนนการประเมิน (g_i) แสดงแนวโน้มที่ผู้ประเมินจะเลือกให้ผลการประเมิน = 1 (สอดคล้อง) ในกรณีที่ผู้ประเมินไม่แน่ใจในกรระบุคะแนนประเมิน

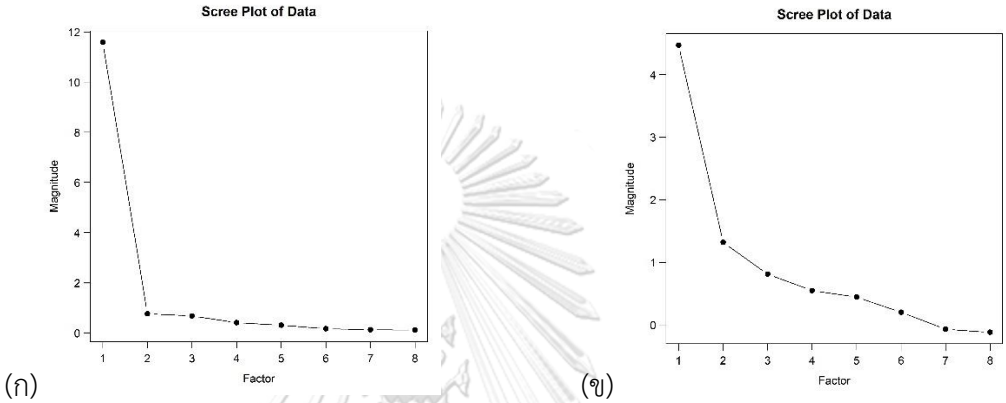
การประยุกต์ใช้การวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินโดยใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมในการศึกษาความสอดคล้องในแนวเดียวกันทางการศึกษานั้น สามารถนำวิธีการวิเคราะห์ไปใช้ในขั้นตอนของการตรวจสอบความเที่ยงของผู้ประเมิน ภายหลังจาก

การจัดเตรียมเนื้อหาของการประเมิน ซึ่งประกอบด้วยตัวชี้วัดและมาตรฐานของหลักสูตรในแต่ละกลุ่มสาระการเรียนรู้ และข้อสอบวัดผลสัมฤทธิ์ทางการเรียน เมื่อผู้เชี่ยวชาญทำการประเมินความสอดคล้องระหว่างเนื้อหาของข้อสอบกับตัวชี้วัดทางการศึกษาตามแบบฟอร์มที่สถานศึกษาหรือหน่วยงานกำหนดแล้ว นำผลการประเมินของผู้เชี่ยวชาญมาวิเคราะห์ความสอดคล้องระหว่างผู้ประเมิน โดยมีขั้นตอนตามรูป 4.3



รูป 4.3 ขั้นตอนการตรวจสอบความสอดคล้องระหว่างผู้ประเมิน
ในการวิเคราะห์ความสอดคล้องในแนวเดียวกัน

จากรูป 4.3 สิ่งที่นักวิจัยหรือผู้วิเคราะห์ข้อมูลต้องเตรียม คือ ข้อมูลผลคะแนนการประเมิน ซึ่งอยู่ในรูปของเมตริกซ์ $N \ x \ M$ เมื่อนำคะแนนการประเมินความสอดคล้องในแนวเดียวกันเข้าสู่โปรแกรมการวิเคราะห์แล้ว ให้ตรวจสอบว่ามีการทำหน้าที่ต่างกันระหว่างผู้ประเมินหรือไม่ โดยการพิจารณาจากผลการคำนวณค่าไอเกน ดังรูป 4.4 หากมีการทำหน้าที่ต่างกันระหว่างผู้ประเมิน (รูป 4.4 (ข)) ให้เลือกวิเคราะห์ด้วยโมเดลสำหรับการทำหน้าที่ต่างกันระหว่างผู้ประเมิน (MC-GCM หรือ MC-LTRM)



(ก) (ข)

รูป 4.4 ผลการวิเคราะห์การทำหน้าที่ต่างกันระหว่างผู้ประเมิน

ตอนที่ 2 ผลการวิเคราะห์ประสิทธิภาพของการประมาณค่าของโมเดลจากข้อมูลจำลอง

การวิเคราะห์ประสิทธิภาพของการประมาณค่าของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม มีวัตถุประสงค์เพื่อศึกษาแนวโน้มในการประมาณค่าของโมเดลเมื่อมีการเปลี่ยนแปลงบริบทของการประเมิน ได้แก่ จำนวนผู้ประเมิน จำนวนรายการประเมิน และการทำหน้าที่ต่างกันของผู้ประเมิน

ประสิทธิภาพการประมาณค่าของโมเดล หมายถึง ความสามารถในการประมาณค่าพารามิเตอร์ของโมเดลมีความน่าเชื่อถือ และสอดคล้องกับค่าที่กำหนดไว้ในการจำลองข้อมูล ผู้วิจัยจำลองข้อมูลด้วยวิธีการจำลองแบบมอนติคาร์โลดังรายละเอียดในบทที่ 3 แล้วคำนวณค่าความลำเอียงในการประมาณค่า และค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง จากนั้นเปรียบเทียบค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าพารามิเตอร์ที่กำหนดกับค่าที่ได้จากการประมาณค่าของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม และวิเคราะห์ค่า R – squared เพื่อศึกษาปัจจัยที่ส่งผลต่อการประมาณค่า ผลการวิเคราะห์ข้อมูลจำแนกตามโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมมีรายละเอียดดังต่อไปนี้

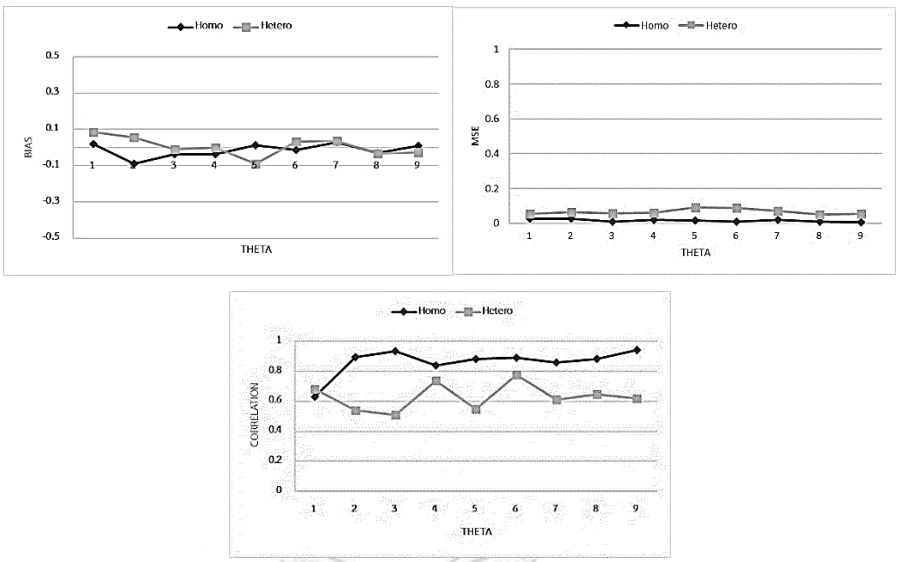
2.1 ประสิทธิภาพของการประมาณค่าของโมเดล MC-GCM

ผลการตรวจสอบความถูกต้องของการจำลองข้อมูลการประมาณค่าความเที่ยงระหว่างผู้ประเมินที่มีการทำหน้าที่ต่างกันของผู้ประเมินจำนวน 2 กลุ่ม ด้วยโมเดล MC-GCM มีรายละเอียดดังนี้

2.1.1 ผลการประมาณค่าพารามิเตอร์ความสามารถของผู้ประเมิน ($\boldsymbol{\theta_i}$)

ผลการประมาณค่าพารามิเตอร์ความสามารถของผู้ประเมินจากการจำลองข้อมูลผลการประเมินแบบ 2 ค่า (1 = ใช่, 0 = ไม่ใช่) จากจำนวนผู้ประเมินและจำนวนรายการประเมินตามรายละเอียดในบทที่ 3 พบว่า โมเดล MC-GCM สามารถประมาณค่าพารามิเตอร์ได้ใกล้เคียงกับค่าจริงของพารามิเตอร์ที่กำหนดไว้ ในกรณีที่คำถามประเมินมีความยากเท่าเทียมกัน โมเดล MC-GCM สามารถประมาณค่าพารามิเตอร์ความสามารถของผู้ประเมิน (θ_i) โดยมีค่าความลำเอียงในการประมาณค่าความสามารถของผู้ประเมิน ระหว่าง -0.088 ถึง 0.028 ค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสองจากการประมาณค่าความสามารถของผู้ประเมิน ระหว่าง 0.009 ถึง 0.028 ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าจริงและค่าที่ได้จากการประมาณค่าความสามารถของผู้ประเมินมีค่าระหว่าง 0.630 ถึง 0.940 และมีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติ ดังแสดงในตารางที่ 4.1 และรูป 4.5

ตาราง 4.2 แสดงให้เห็นว่า กรณีที่คำถามประเมินมีความยากไม่เท่าเทียมกัน การประมาณค่าพารามิเตอร์ความสามารถของผู้ประเมินของโมเดล MC-GCM จะมีความคลาดเคลื่อนสูงกว่ากรณีที่รายการประเมินมีความยากเท่าเทียมกัน โดยมีค่าความลำเอียงในการประมาณค่าความสามารถของผู้ประเมินมีค่าระหว่าง -0.090 ถึง 0.079 ค่าความลำเอียงในการประมาณค่าฉันทามติเชิงวัฒนธรรมของผู้ประเมินกลุ่มที่ 2 ระหว่าง -0.027 ถึง 0.086 มีค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสองของการประมาณค่าความสามารถของผู้ประเมิน ระหว่าง 0.055 ถึง 0.091 และมีค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าจริงและค่าที่ได้จากการประมาณค่าพารามิเตอร์ความสามารถของผู้ประเมินที่ต่ำกว่ารายการประเมินที่มีความยากเท่าเทียมกัน คือมีค่าระหว่าง 0.510 ถึง 0.775

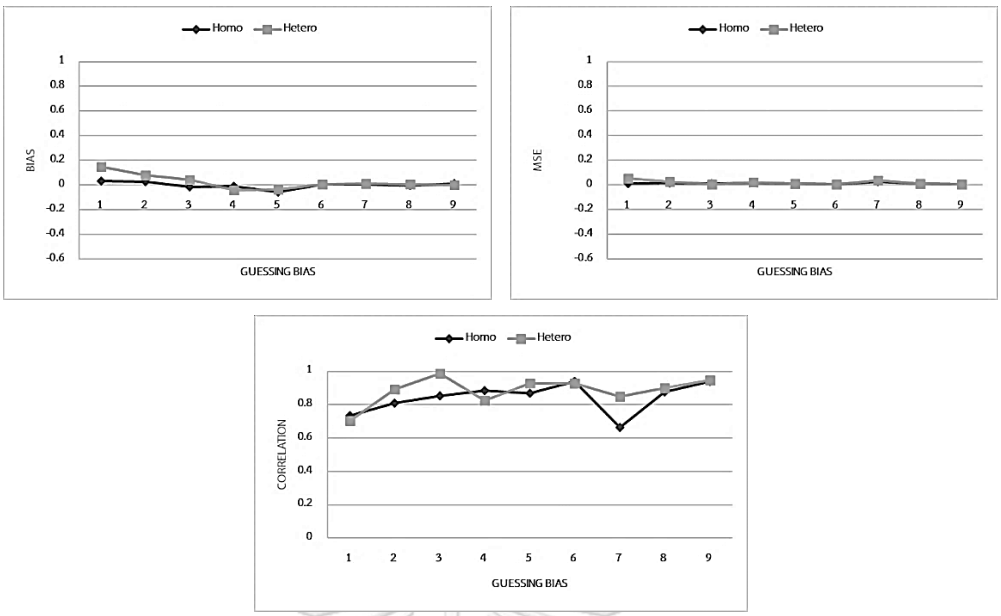


รูป 4.5 ผลการประมาณค่าความสามารถของผู้ประเมินโดยโมเดล MC-GCM

2.1.2 ผลการประมาณค่าพารามิเตอร์ความลำเอียงในการประเมิน (g_i)

โมเดล MC-GCM สามารถประมาณค่าพารามิเตอร์ความลำเอียงในการประเมิน (g_i) จากการจำลองข้อมูลผลการประเมินแบบ 2 ค่า (1 = ใช่, 0 = ไม่ใช่) ได้ใกล้เคียงกับค่าจริงของพารามิเตอร์ที่กำหนดไว้ ค่าประมาณพารามิเตอร์ความลำเอียงในการประเมิน (g_i) มีค่าความลำเอียงในการประมาณค่า ระหว่าง -0.058 ถึง 0.034 ค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสองจากการประมาณค่า ระหว่าง 0.005 ถึง 0.027 ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าจริงและค่าที่ได้จากการประมาณค่าความสามารถของผู้ประเมิน มีค่าระหว่าง 0.663 ถึง 0.940 และมีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติ ดังแสดงในตาราง 4.1 และ 4.2

จากรูป 4.6 จะเห็นได้ว่า กรณีที่รายการประเมินมีความยากของคำถามประเมินแตกต่างกันจะมีความคลาดเคลื่อนของการประมาณค่าสูงกว่ากรณีที่รายการประเมินมีความยากของคำถามประเมินเท่าเทียมกัน โดยมีค่าความลำเอียงในการประมาณค่า ระหว่าง -0.038 ถึง 0.152 ค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสอง ระหว่าง 0.005 ถึง 0.057 ค่า และค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าจริงและค่าที่ได้จากการประมาณค่า มีค่าระหว่าง 0.702 ถึง 0.985

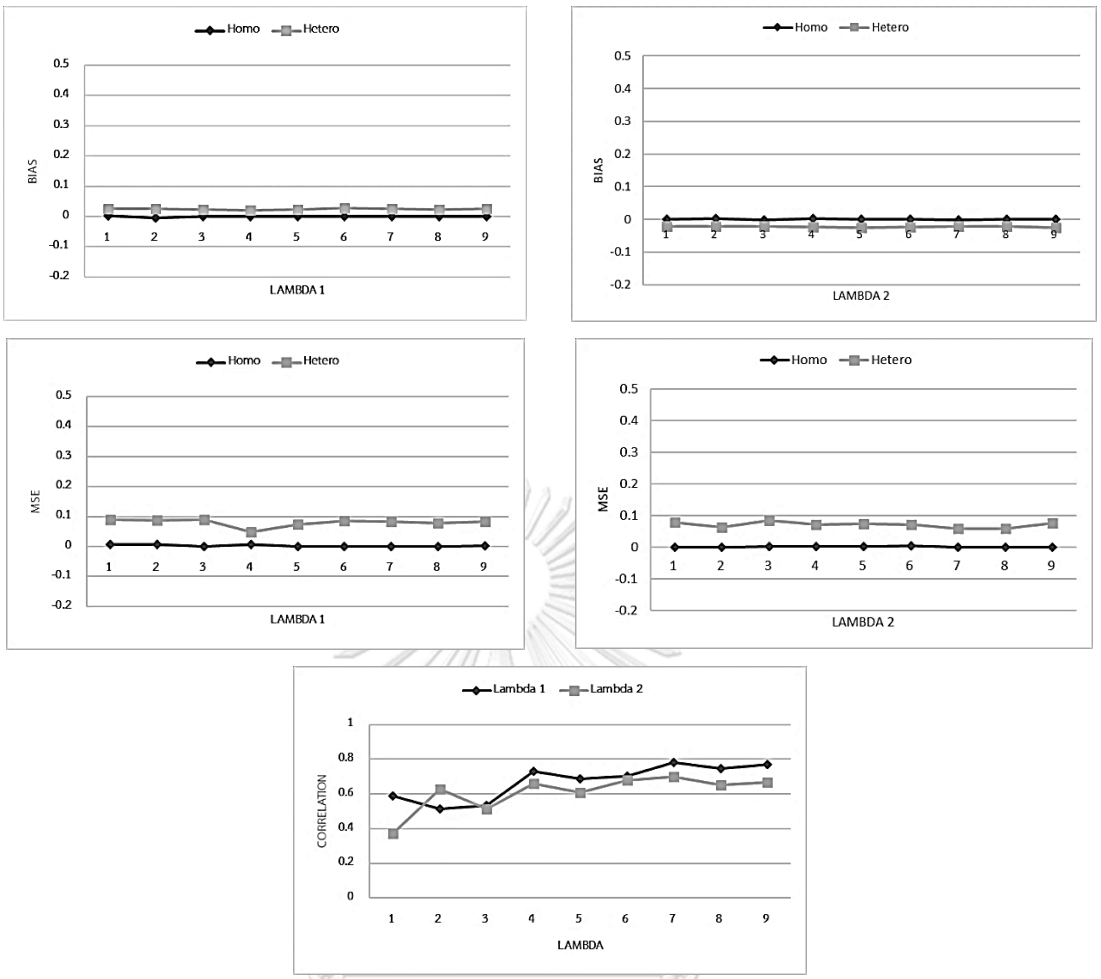


รูป 4.6 ผลการประมาณค่าพารามิเตอร์ความลำเอียงในการประเมินโดยโมเดล MC-GCM

2.1.3 ผลการประมาณค่าพารามิเตอร์ความยากของรายการประเมิน (δ_k)

ผลการตรวจสอบความถูกต้องของการประมาณค่าพารามิเตอร์ความยากของรายการประเมิน (δ_k) จากการจำลองข้อมูลผลการประเมินแบบ 2 ค่า (1 = ใช่, 0 = ไม่ใช่) พบว่าค่าพารามิเตอร์ความยากของรายการประเมิน (δ_k) มีค่าความลำเอียงในการประมาณค่าความยากของรายการประเมินของผู้ประเมินกลุ่มที่ 1 ระหว่าง -0.000 ถึง 0.003 ค่ากลุ่มที่ 2 ระหว่าง -0.002 ถึง 0.002 มีค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสองของการประมาณค่าความยากของรายการประเมินของผู้ประเมินกลุ่มที่ 1 ระหว่าง 0.000 ถึง 0.008 กลุ่มที่ 2 มีค่าระหว่าง 0.000 ถึง 0.005 ดังแสดงในตารางที่ 4.1 และ 4.2

การประมาณค่าของโมเดล MC-GCM ในกรณีที่คำถามประเมินมีความยากแตกต่างกันมีความคลาดเคลื่อนสูงกว่ากรณีที่รายการประเมินมีความยากเท่าเทียมกัน ดังแสดงในรูป 4.7 โดยมีค่าความลำเอียงในการประมาณค่าพารามิเตอร์ความยากของรายการประเมินของผู้ประเมินกลุ่มที่ 1 ระหว่าง 0.022 ถึง 0.026 ผู้ประเมินกลุ่มที่ 2 ระหว่าง -0.027 ถึง -0.022 มีค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสองจากการประมาณค่าความยากของรายการประเมินของผู้ประเมินกลุ่มที่ 1 ระหว่าง 0.049 ถึง 0.091 ผู้ประเมินกลุ่มที่ 2 มีค่าระหว่าง 0.060 ถึง 0.086 ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าจริงและค่าที่ได้จากการประมาณค่าความยากของรายการประเมินของผู้ประเมินกลุ่มที่ 1 มีค่าระหว่าง 0.513 ถึง 0.783 ผู้ประเมินกลุ่มที่ 2 มีค่าระหว่าง 0.370 ถึง 0.697



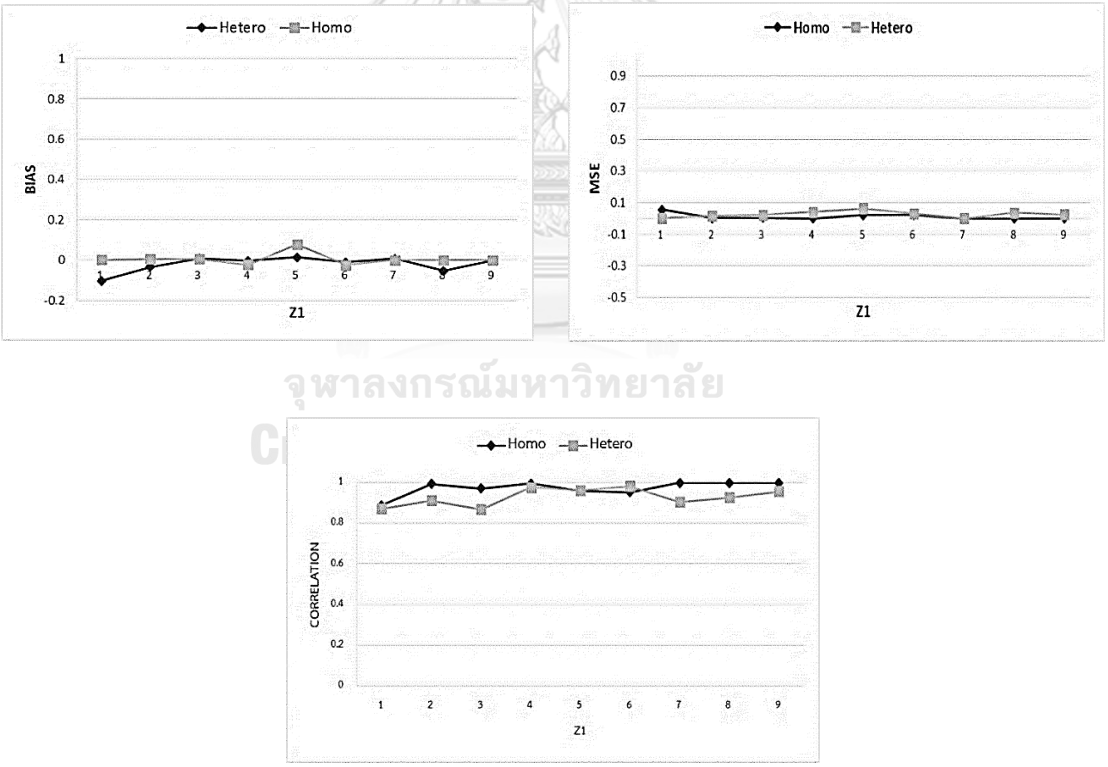
รูป 4.7 ผลการประมาณค่าพารามิเตอร์ความยากของรายการประเมินโดยโมเดล MC-GCM

2.1.4 ผลการประมาณค่าพารามิเตอร์ฉันทามติเชิงวัฒนธรรม (Z_k)

ผู้วิจัยจำลองข้อมูลผลการประเมินแบบ 2 ค่า (1 = ใช่, 0 = ไม่ใช่) จากจำนวนผู้ประเมินและจำนวนรายการประเมินตามรายละเอียดในบทที่ 3 แล้วตรวจสอบความถูกต้องของการประมาณค่าพารามิเตอร์ฉันทามติเชิงวัฒนธรรมของผู้ประเมินด้วยโมเดล MC-GCM ใน 2 กรณี คือ 1) รายการประเมินที่มีความยากเท่าเทียมกันและ 2) รายการประเมินที่มีความยากแตกต่างกัน ผลการตรวจสอบความถูกต้อง พบว่า โมเดล MC-GCM สามารถประมาณค่าพารามิเตอร์ได้ใกล้เคียงกับค่าจริงของพารามิเตอร์ที่กำหนดไว้ ในกรณีที่คำถามประเมินมีความยากเท่าเทียมกัน โมเดล MC-GCM สามารถประมาณค่าพารามิเตอร์ฉันทามติเชิงวัฒนธรรม (Z_k) โดยมีค่าความลำเอียงในการประมาณค่าฉันทามติเชิงวัฒนธรรมของผู้ประเมินกลุ่มที่ 1 ระหว่าง -0.024 ถึง 0.079 ค่าความลำเอียงในการประมาณค่าฉันทามติเชิงวัฒนธรรมของผู้ประเมินกลุ่มที่ 2 ระหว่าง -0.027 ถึง 0.053 มีค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสองจากการประมาณค่าฉันทามติเชิงวัฒนธรรมของผู้ประเมินกลุ่มที่ 1 ระหว่าง 0.000 ถึง 0.056 ผู้ประเมินกลุ่มที่ 2 มีค่าระหว่าง 0.000 ถึง 0.167 ค่าสัมประสิทธิ์

สหสัมพันธ์ระหว่างค่าจริงและค่าที่ได้จากการประมาณค่าฉันทามติเชิงวัฒนธรรมของผู้ประเมินกลุ่มที่ 1 มีค่าระหว่าง 0.889 ถึง 1.000 ผู้ประเมินกลุ่มที่ 2 มีค่าระหว่าง 0.695 ถึง 1.000 และมีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติ ดังแสดงในตารางที่ 4.1 และรูปที่ 4.8

อย่างไรก็ตาม การประมาณค่าของโมเดล MC-GCM ในกรณีที่รายการประเมินมีความยากของข้อคำถามไม่เท่าเทียมกันจะมีความคลาดเคลื่อนสูงกว่ากรณีที่รายการประเมินมีความยากเท่าเทียมกันโดยมีค่าความลำเอียงในการประมาณค่าฉันทามติเชิงวัฒนธรรมของผู้ประเมินกลุ่มที่ 1 ระหว่าง -0.003 ถึง 0.015 ค่าความลำเอียงในการประมาณค่าฉันทามติเชิงวัฒนธรรมของผู้ประเมินกลุ่มที่ 2 ระหว่าง -0.006 ถึง 0.075 มีค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสองจากการประมาณค่าฉันทามติเชิงวัฒนธรรมของผู้ประเมินกลุ่มที่ 1 ระหว่าง 0.008 ถึง 0.059 ผู้ประเมินกลุ่มที่ 2 มีค่าระหว่าง 0.003 ถึง 0.062 ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าจริงและค่าที่ได้จากการประมาณค่าฉันทามติเชิงวัฒนธรรมของผู้ประเมินกลุ่มที่ 1 มีค่าระหว่าง 0.868 ถึง 0.982 ผู้ประเมินกลุ่มที่ 2 มีค่าระหว่าง 0.766 ถึง 0.992 ดังรายละเอียดในตาราง 4.2



รูป 4.8 ผลการประมาณค่าฉันทามติเชิงวัฒนธรรมระหว่างผู้ประเมิน

ตาราง 4.1 ผลการประมาณค่าพารามิเตอร์โดยโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม MC-GCM กรณีไม่มีการทำหน้าที่ต่างกันของผู้ประเมิน

	N	15			30			45		
	M	25	55	85	25	55	85	25	55	85
Theta	bias	0.020	-0.088	-0.035	-0.038	0.013	-0.015	0.028	-0.03	0.007
	mse	0.027	0.028	0.01	0.021	0.015	0.009	0.019	0.011	0.007
	corr	0.630**	0.894**	0.932**	0.838**	0.884**	0.891**	0.860**	0.884**	0.940**
g	bias	0.034	0.026	-0.011	-0.006	-0.058	0.006	0.008	-0.004	0.012
	mse	0.015	0.017	0.012	0.019	0.015	0.005	0.027	0.011	0.008
	corr	0.735**	0.809**	0.854**	0.882**	0.869**	0.939**	0.663**	0.877**	0.940**
delta	bias	0.002	0.001	0.001	0.001	0.000	0.000	0.001	0.000	0.000
	mse	0.004	0.003	0.002	0.004	0.002	0.003	0.001	0.000	0.002
Z	Bias	0.028	0.025	0.004	-0.012	-0.021	-0.026	0.000	-0.003	0.000
	MSE	0.054	0.085	0.004	0.002	0.015	0.023	0.000	0.001	0.000
	corr	0.893	0.845	0.981	0.998	0.971	0.955	1.000	0.998	0.999

**p-value ≤ 0.01

ตาราง 4.2 ผลการประมาณค่าพารามิเตอร์โดยโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม MC-GCM กรณีมีการทำหน้าที่ต่างกันของผู้ประเมิน

	N	15			30			45		
	M	25	55	85	25	55	85	25	55	85
theta	bias	0.086	0.055	-0.012	-0.001	-0.09	0.032	0.037	-0.033	-0.027
	mse	0.055	0.063	0.058	0.061	0.091	0.088	0.071	0.052	0.053
	corr	0.683**	0.538**	0.510**	0.735**	0.546**	0.775**	0.612**	0.646**	0.618**
g	bias	0.152	0.082	0.042	-0.038	-0.034	0.009	0.014	0.008	0.004
	mse	0.057	0.028	0.005	0.022	0.011	0.009	0.037	0.015	0.007
	corr	0.702**	0.891**	0.985**	0.825**	0.926**	0.928**	0.848**	0.898**	0.947**
delta1	bias	0.026	0.025	0.024	0.020	0.024	0.027	0.026	0.023	0.025
	mse	0.089	0.088	0.091	0.049	0.073	0.086	0.084	0.078	0.083
	corr	0.588**	0.513**	0.535**	0.732**	0.686**	0.701**	0.783**	0.746**	0.769**
Delta2	bias	-0.022	-0.022	-0.022	-0.024	-0.025	-0.024	-0.021	-0.022	-0.026
	mse	0.079	0.063	0.086	0.071	0.075	0.071	0.060	0.060	0.076
	corr	0.370	0.629**	0.512**	0.659**	0.610**	0.679**	0.697**	0.652**	0.667*
Z	Bias	0.047	0.027	0.032	0.021	0.045	0.004	0.004	0.052	0.008
	MSE	0.031	0.008	0.006	0.026	0.038	0.020	0.004	0.030	0.022
	corr	0.932	0.914	0.817	0.948	0.941	0.959	0.927	0.938	0.954

**p-value ≤ 0.01

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

2.2 ประสิทธิภาพของการประมาณค่าโมเดล MC-LTRM

ผลการตรวจสอบความถูกต้องของการประมาณค่าความเที่ยงระหว่างผู้ประเมินที่มีการทำหน้าที่ต่างกันของผู้ประเมิน 2 กลุ่ม ด้วยโมเดล MC-LTRM มีรายละเอียดดังนี้

2.2.1 ผลการประมาณค่าประมาณตำแหน่งของการตอบ (T_{vk})

ผู้วิจัยจำลองข้อมูลผลการประเมินแบบ 5 ระดับ จากจำนวนผู้ประเมินและจำนวนรายการประเมินตามรายละเอียดในตอนที่ 1 ข้อ 1.1 แล้วตรวจสอบความถูกต้องของการประมาณค่าพารามิเตอร์ฉันทามติเชิงวัฒนธรรมของผู้ประเมินด้วยโมเดล MC-LTRM ใน 2 กรณี คือ 1) กรณีไม่มีการทำหน้าที่ต่างกันของผู้ประเมิน และ 2) กรณีมีการทำหน้าที่ต่างกันของผู้ประเมิน ผลการตรวจสอบความถูกต้องในกรณีที่รายการประเมินมีความยากเท่าเทียมกันของคำถามประเมิน โมเดล MC-LTRM สามารถประมาณค่าพารามิเตอร์ฉันทามติเชิงวัฒนธรรม (T_{vk}) โดยมีค่าความลำเอียงในการประมาณค่าฉันทามติเชิงวัฒนธรรมของผู้ประเมิน ระหว่าง -0.140 ถึง 0.220 ค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสองจากการประมาณค่าฉันทามติเชิงวัฒนธรรมของผู้ประเมินอยู่ระหว่าง 0.010 ถึง 0.064 ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าจริงและค่าที่ได้จากการประมาณค่าฉันทามติเชิงวัฒนธรรมของผู้ประเมินมีค่าระหว่าง 0.995 ถึง 0.998 ดังรายละเอียดในตาราง 4.3

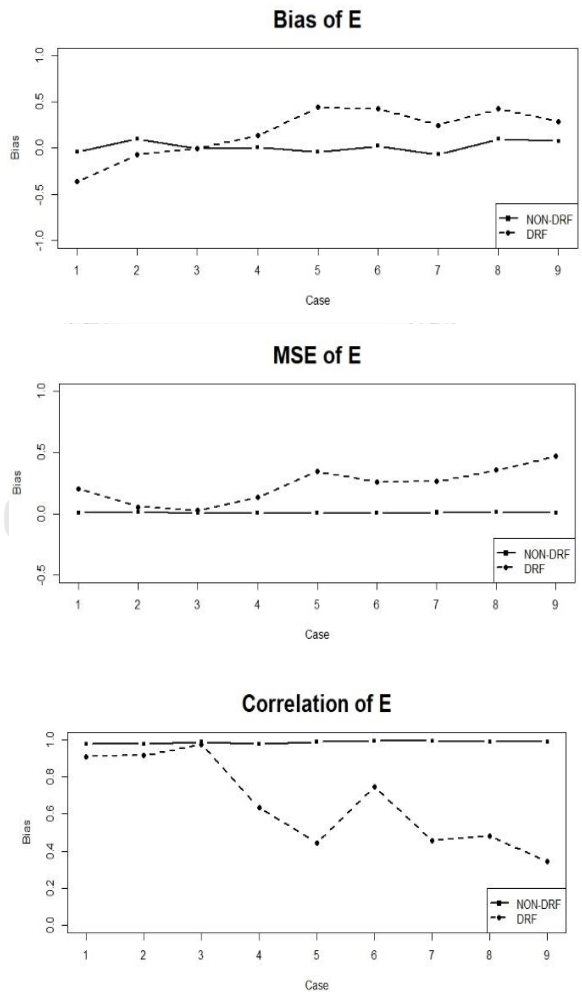
กรณีที่มีการทำหน้าที่ต่างกันของผู้ประเมิน พบว่า การประมาณค่าของโมเดล MC-LTRM มีความคลาดเคลื่อนสูงกว่า โดยมีค่าความลำเอียงในการประมาณค่าฉันทามติเชิงวัฒนธรรมของผู้ประเมินกลุ่มที่ 1 ระหว่าง -1.122 ถึง 0.053 ค่าความลำเอียงในการประมาณค่าฉันทามติเชิงวัฒนธรรมของผู้ประเมินกลุ่มที่ 2 ระหว่าง -1.003 ถึง 2.012 มีค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสองจากการประมาณค่าฉันทามติเชิงวัฒนธรรมของผู้ประเมินกลุ่มที่ 1 ระหว่าง 0.282 ถึง 1.473 ผู้ประเมินกลุ่มที่ 2 มีค่าระหว่าง -0.022 ถึง 4.141 ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าจริงและค่าที่ได้จากการประมาณค่าฉันทามติเชิงวัฒนธรรมของผู้ประเมินกลุ่มที่ 1 มีค่าระหว่าง 0.937 ถึง 0.989 ผู้ประเมินกลุ่มที่ 2 มีค่าระหว่าง 0.936 ถึง 0.986 และมีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติ ดังแสดงในตารางที่ 4.4

2.2.2 ผลการประมาณค่าพารามิเตอร์ความสามารถของผู้ประเมิน (E_i)

ผลการประมาณค่าพารามิเตอร์ความสามารถของผู้ประเมินจากการจำลองข้อมูลผลการประเมิน พบว่า โมเดล MC-LTRM สามารถประมาณค่าพารามิเตอร์ได้ใกล้เคียงกับค่าจริงของพารามิเตอร์ที่กำหนดไว้ ในกรณีที่คำถามประเมินมีความยากเท่าเทียมกัน โมเดล MC-LTRM สามารถประมาณค่าพารามิเตอร์ความสามารถของผู้ประเมิน (E_i) โดยมีค่าความลำเอียงในการประมาณค่าความสามารถของผู้ประเมิน ระหว่าง -0.003 ถึง 0.439 ค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสองจาก

การประมาณค่าความสามารถของผู้ประเมิน ระหว่าง 0.027 ถึง 0.470 ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าจริงและค่าที่ได้จากการประมาณค่าความสามารถของผู้ประเมิน มีค่าระหว่าง 0.916 ถึง 0.990 และมีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติ ดังแสดงในตารางที่ 4.3

กรณีที่มีการทำหน้าที่ต่างกันของผู้ประเมิน การประมาณค่าพารามิเตอร์ความสามารถของผู้ประเมินของโมเดล MC-LTRM จะมีความคลาดเคลื่อนสูงกว่ากรณีที่ไม่มีการทำหน้าที่ต่างกันของผู้ประเมิน โดยมีค่าความลำเอียงในการประมาณค่าความสามารถของผู้ประเมินมีค่าระหว่าง -0.042 ถึง 0.101 ค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสองของการประมาณค่าความสามารถของผู้ประเมิน ระหว่าง 0.004 ถึง 0.015 และมีค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าจริงและค่าที่ได้จากการประมาณค่าพารามิเตอร์ความสามารถของผู้ประเมินที่ต่ำกว่ากรณีไม่มีการทำหน้าที่ต่างกันของผู้ประเมิน คือมีค่าระหว่าง 0.345 ถึง 0.976 และค่าสัมประสิทธิ์สหสัมพันธ์จะต่ำลงเมื่อจำนวนรายการประเมินมีจำนวนมากขึ้น ดังตาราง 4.4 และรูป 4.9



รูป 4.9 เปรียบเทียบการประมาณค่าระหว่างข้อมูลที่มี DRF กับไม่มี DRF

2.2.3 ผลการประมาณค่าความยากของรายการประเมิน $(\boldsymbol{\lambda_k})$

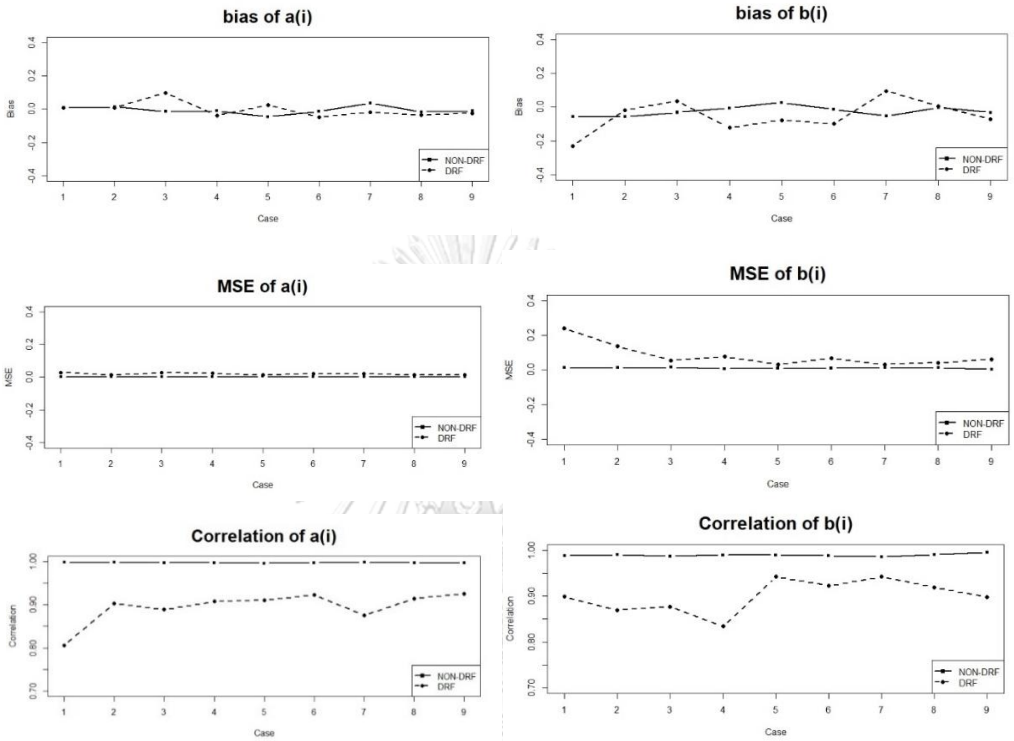
ผลการประมาณค่าพารามิเตอร์ยากของการประเมินจากการจำลองข้อมูลผลการประเมิน พบว่า ในกรณีที่รายการประเมินมีความยากเท่าเทียมกัน โมเดล MC-LTRM สามารถประมาณค่าพารามิเตอร์ยากของรายการประเมิน $(\boldsymbol{\lambda_k})$ โดยมีค่าความลำเอียงในการประมาณค่าความสามารถของผู้ประเมิน ระหว่าง -0.009 ถึง 0.009 ค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสองจากการประมาณค่าความสามารถของผู้ประเมินมีค่าในช่วง 0.001 ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าจริงและค่าที่ได้จากการประมาณค่าความสามารถของผู้ประเมินมีค่าระหว่าง 0.984 ถึง 0.990 และมีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติ

ผลการประมาณค่าของโมเดล MC-LTRM กรณีมีการทำหน้าที่ต่างกันของผู้ประเมินมีค่าความลำเอียงในการประมาณค่าความยากของการประเมินจากผู้ประเมินกลุ่มที่ 1 ระหว่าง -0.29 ถึง 0.002 กลุ่มที่ 1 ระหว่าง -0.25 ถึง 0.068 ค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสองจากการประมาณค่าความยากของรายการประเมินของผู้ประเมินกลุ่มที่ 1 มีค่าระหว่าง 0.006 ถึง 0.013 กลุ่มที่ 2 มีค่าระหว่าง 0.006 ถึง 0.037 ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าจริงและค่าที่ได้จากการประมาณค่าความยากของรายการประเมินของผู้ประเมินกลุ่มที่ 1 มีค่าระหว่าง 0.712 ถึง 0.873 กลุ่มที่ 2 มีค่าระหว่าง 0.905 ถึง 0.953 และมีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติ ดังแสดงในตารางที่ 4.4

2.2.4 ผลการประมาณค่าประมาณความลำเอียงในการประเมิน $(\boldsymbol{a_i})$ และ $(\boldsymbol{b_i})$

ผลการประมาณค่าพารามิเตอร์ความลำเอียงในการประเมิน $(a_i\,, b_i)$ โดยโมเดล MC-LTRM พลว่าโมเดลสามารถประมาณค่าได้ใกล้เคียงกับค่าจริงของพารามิเตอร์ที่กำหนดไว้ ในกรณีไม่มีการทำหน้าที่ต่างกันของผู้ประเมิน ค่าพารามิเตอร์ a_i มีค่าความลำเอียงในการประมาณค่า ระหว่าง -0.047 ถึง 0.036 ค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสองจากการประมาณค่าระหว่าง 0.002 ถึง 0.005 ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าจริงและค่าที่ได้จากการประมาณค่ามีค่าระหว่าง 0.996 ถึง 0.998 และมีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติ ผลการประมาณค่าพารามิเตอร์ b_i มีค่าความลำเอียงในการประมาณค่า ระหว่าง -0.058 ถึง 0.027 ค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสองจากการประมาณค่าระหว่าง 0.008 ถึง 0.015 ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าจริงและค่าที่ได้จากการประมาณค่ามีค่าระหว่าง 0.987 ถึง 0.998 และมีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติ ดังแสดงในตาราง 4.3

รูป 4.10 แสดงให้เห็นความแตกต่างของผลการประมาณค่ากรณีมีการทำหน้าที่ต่างกันของผู้ประเมิน จะเห็นได้ว่า เมื่อมีการทำหน้าที่ต่างกันของผู้ประเมินในการให้คะแนนการประเมิน จะทำให้การประมาณค่ามีความถูกต้องและแม่นยำต่ำว่าการประมาณค่าของโมเดลในกรณีการประเมินที่ไม่มีการทำหน้าที่ต่างกันของผู้ประเมิน



รูป 4.10 การเปรียบเทียบผลการประมาณค่าพารามิเตอร์ $(a_i\ , b_i)$
ระหว่างกรณีมี DRF กับไม่มี DRF

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

ตาราง 4.3 ผลการประมาณค่าพารามิเตอร์ฉันทามติเชิงวัฒนธรรมโดยโมเดล MC-LTRM กรณีไม่มีการทำหน้าที่ต่างกันของผู้ประเมิน

	N	15			30			45		
	M	25	55	85	25	55	85	25	55	85
E	bias	-0.042	0.101	-0.007	0.009	-0.040	0.020	-0.069	0.095	0.074
	mse	0.007	0.015	0.004	0.005	0.004	0.006	0.010	0.011	0.008
	corr	0.978	0.978	0.987	0.976	0.989	0.993	0.994	0.990	0.990
T	bias	-0.102	0.142	0.017	-0.011	-0.025	0.000	-0.140	0.220	0.117
	mse	0.018	0.037	0.012	0.010	0.013	0.011	0.027	0.064	0.030
	corr	0.997	0.996	0.996	0.997	0.996	0.996	0.998	0.996	0.995

	N	15			30			45		
	M	25	55	85	25	55	85	25	55	85
lambda	bias	-0.001	-0.001	0.009	-0.012	0.004	0.002	0.005	0.006	-0.001
	mse	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
a	bias	0.008	0.013	-0.014	-0.012	-0.047	-0.014	0.036	-0.016	-0.010
	mse	0.002	0.002	0.002	0.003	0.005	0.002	0.003	0.002	0.002
	corr	0.998	0.998	0.997	0.997	0.996	0.997	0.998	0.997	0.997
b	bias	-0.056	-0.058	-0.032	-0.006	0.027	-0.014	-0.053	-0.003	-0.032
	mse	0.013	0.012	0.015	0.008	0.008	0.009	0.014	0.011	0.005
	corr	0.988	0.990	0.987	0.990	0.990	0.988	0.986	0.990	0.996

**p-value ≤ 0.01

ตาราง 4.4 ผลการประมาณค่าพารามิเตอร์ฉันทามติเชิงวัฒนธรรมโดยโมเดล MC-LTRM กรณีมีการทำหน้าที่ต่างกันของผู้ประเมิน

	N	15			30			45		
	M	25	55	85	25	55	85	25	55	85
E	bias	-0.361	-0.073	-0.003	0.132	0.439	0.425	0.248	0.423	0.289
	mse	0.203	0.056	0.027	0.132	0.345	0.260	0.264	0.355	0.470
	corr	0.910	0.916	0.976	0.634	0.445	0.746	0.457	0.481	0.345
T	bias	-1.063	-0.377	-0.376	0.116	0.802	0.655	0.463	0.906	0.707
	mse	1.286	0.259	0.531	0.554	1.342	1.784	2.117	2.344	2.360
	corr	0.960	0.984	0.981	0.978	0.973	0.959	0.976	0.981	0.987
lambda	bias	0.028	-0.003	-0.018	-0.008	-0.016	0.003	0.006	0.011	0.008
	mse	0.022	0.010	0.008	0.009	0.009	0.009	0.007	0.007	0.006
	corr	0.858	0.908	0.872	0.855	0.842	0.809	0.890	0.878	0.891
a	bias	0.009	0.009	0.098	-0.038	0.024	-0.048	-0.018	-0.036	-0.025
	mse	0.028	0.014	0.027	0.023	0.014	0.021	0.022	0.014	0.016
	corr	0.806	0.903	0.889	0.908	0.910	0.922	0.876	0.914	0.925
b	bias	-0.231	-0.018	0.035	-0.122	-0.078	-0.099	0.096	0.004	-0.070
	mse	0.242	0.137	0.056	0.077	0.032	0.067	0.033	0.040	0.063
	corr	0.899	0.869	0.877	0.834	0.943	0.923	0.942	0.919	0.898

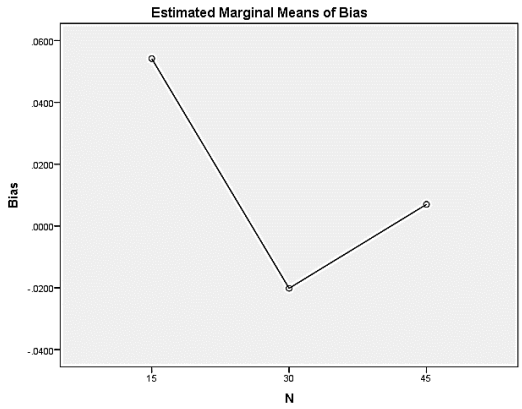
**p-value ≤ 0.01

2.3 ผลการวิเคราะห์ขนาดอิทธิพลของตัวแปรอิสระที่ส่งผลต่อการประมาณค่าของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม

จากการศึกษาเอกสารงานวิจัยที่เกี่ยวข้องกับประสิทธิภาพการประมาณค่าของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมมีข้อสรุปจากการศึกษาก่อนหน้าว่า โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมได้รับการพัฒนามาให้เหมาะกับการวิเคราะห์ข้อมูลจำนวนน้อย โดยมีการศึกษาโดยการจำลองข้อมูลและการศึกษาจากข้อมูลจริงว่าโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมสามารถ

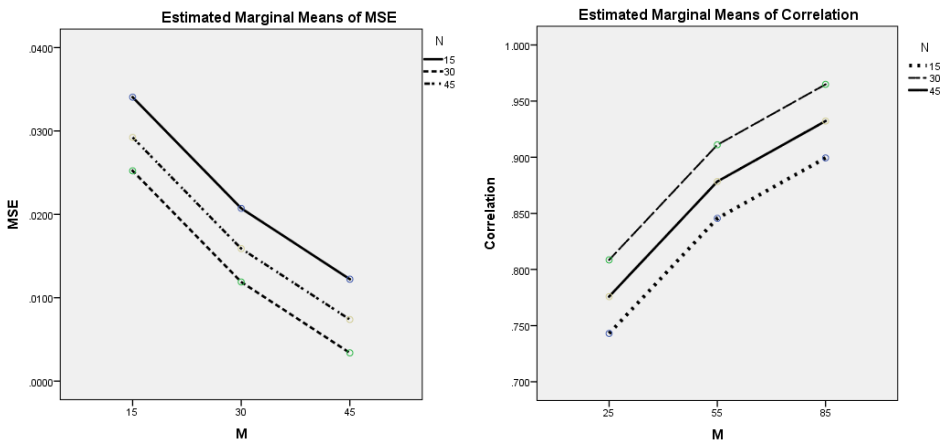
วิเคราะห์ข้อมูลที่มีจำนวนผู้ให้ข้อมูลขั้นต่ำ 6 คนต่อกลุ่มวัฒนธรรม การศึกษาขนาดอิทธิพลในงานวิจัยนี้มีจุดมุ่งหมายเพื่อสนับสนุนข้อสรุปของการศึกษาที่ผ่านมาว่า จำนวนผู้ประเมิน จำนวนข้อคำถามการประเมิน และการทำหน้าที่ต่างกันของผู้ประเมิน ไม่ส่งผลต่อประสิทธิภาพในการประมาณค่าของโมเดล ผลการวิเคราะห์ขนาดอิทธิพลของตัวแปรอิสระที่ส่งผลต่อการประมาณค่าของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม พบว่า ตัวแปรอิสระที่ส่งผลต่อประสิทธิภาพในการประมาณค่าของโมเดลมีดังนี้

จำนวนผู้ประเมิน (N) ส่งผลต่อค่าความลำเอียงในการการประมาณค่า (Bias) ของพารามิเตอร์ความลำเอียงของผู้ประเมิน (g) ในการประมาณค่าของโมเดล MC-GCM โดยมีขนาดอิทธิพลเท่ากับ 0.531 ในรูป 4.11 จะเห็นว่าเมื่อจำนวนผู้ประเมินเพิ่มขึ้น ทำให้ค่าความลำเอียงในการประมาณค่าพารามิเตอร์ g ลดลง



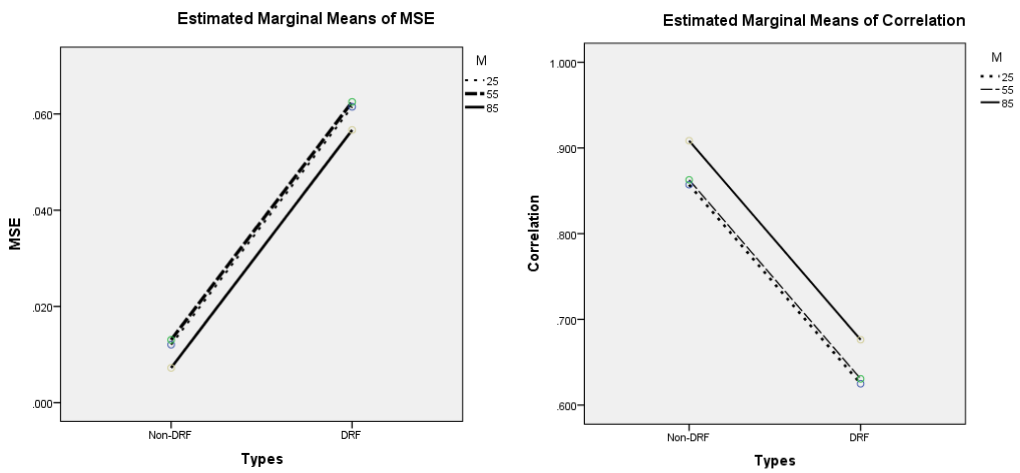
รูป 4.11 อิทธิพลของจำนวนผู้ประเมินส่งผลต่อความลำเอียงในการประมาณค่า

จำนวนข้อคำถามประเมิน (M) ส่งผลต่อค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (MSE) และค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าจริงกับค่าประมาณพารามิเตอร์ ในการประมาณค่าความลำเอียงของผู้ประเมิน (g) ของโมเดล MC-GCM โดยมีขนาดอิทธิพลต่อค่าเฉลี่ยความคลาดเคลื่อนกำลังสองเท่ากับ 0.597 และมีขนาดอิทธิพลต่อค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าจริงกับค่าที่ได้จากการประมาณค่าพารามิเตอร์ของโมเดลเท่ากับ 0.685 โดยเมื่อมีจำนวนข้อคำถามประเมินมากขึ้นจะส่งผลให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองลดลง และส่งผลให้ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าจริงกับค่าประเมินเพิ่มขึ้น ดังรูป 4.12



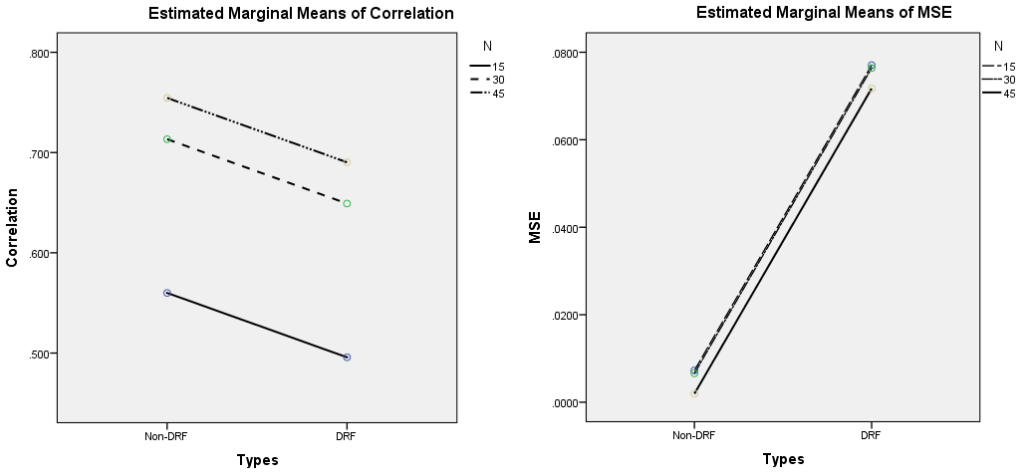
รูป 4.12 อิทธิพลของจำนวนข้อคำถามประเมินส่งผลต่อค่า MSE และ correlation

การทำหน้าที่ต่างกันของผู้ประเมิน (Types) ส่งผลต่อค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง และค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าจริงกับค่าประมาณ ในการประมาณค่าพารามิเตอร์ความสามารถของผู้ประเมิน (Theta) โดยมีและส่งผลต่อความลำเอียงในการประมาณค่า และค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของการประมาณค่าพารามิเตอร์ความยากของรายการประเมิน (Delta) ในการประมาณค่าของโมเดล MC-GCM โดยมีขนาดอิทธิพลของตัวแปรอิสระที่ส่งผลต่อค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง และค่าสัมประสิทธิ์สหสัมพันธ์มีค่าเท่ากับ 0.869 และ .702 ตามลำดับ รูป 4.13 จะเห็นว่า เมื่อมีการทำหน้าที่ต่างกันของผู้ประเมินมีผลให้ความคลาดเคลื่อนในการประมาณค่าสูงขึ้น เช่นเดียวกับการประมาณค่าของโมเดล MC-LTRM ซึ่งมีค่า Bias และ MSE สูงขึ้นเมื่อมีการทำหน้าที่ต่างกันของผู้ประเมิน ดังรูป 4.14



รูป 4.13 การทำหน้าที่ต่างกันของผู้ประเมิน

ส่งผลต่อค่า MSE และ correlation ในโมเดล MC-GCM



รูป 4.14 การทำหน้าที่ต่างกันระหว่าผู้ประเมิน
ส่งผลต่อค่า MSE และ correlation ในโมเดล MC-LTRM

ตาราง 4.5 ผลการวิเคราะห์ขนาดอิทธิพลของตัวแปรอิสระที่ส่งผลต่อการประมาณค่าของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม

ตัวแปรอิสระที่ศึกษา	พารามิเตอร์	Partial Eta-squared		
		Bias	MSE	Corr
โมเดล MC-GCM				
จำนวนผู้ประเมิน (N)	Theta	0.052	0.21	0.171
	ʒ	0.531*	0.193	0.27
	Delta	0.44	0.415	0.609
	Z	0.078	0.169	0.306
จำนวนคำถามประเมิน (M)	Theta	0.244	0.066	0.085
	ʒ	0.109	0.597**	0.685**
	Delta	0.288	0.459	0.137
	Z	0.134	0.17	0.092
การทำหน้าที่ต่างกันระหว่างผู้ประเมิน (Type)	Theta	0.072	0.869**	0.702**
	ʒ	0.166	0.179	0.189
	Delta	0.989**	0.982**	0.296
	Z	0.004	0.004	0.035
โมเดล MC-LTRM				
จำนวนผู้ประเมิน (N)	E	0.422	0.324	0.422
	lamda	0.128	0.131	0.117
	a(i)	0.294	0.117	0.215
	b(i)	0.409	0.283	0.118
จำนวนคำถามประเมิน (M)	E	0.255	0.024	0.018
	lamda	0.131	0.131	0.132
	a(i)	0.012	0.277	0.243
	b(i)	0.09	0.109	0.016

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

ตัวแปรอิสระที่ศึกษา	พารามิเตอร์	Partial Eta-squared		
		Bias	MSE	Corr
การทำหน้าที่ต่างกันระหว่างผู้ประเมิน (Type)	E	0.259	0.672	0.652
	lamda	0.979**	0.950**	0.961**
	a(i)	0.004	0.876**	0.875**
	b(i)	0.189	0.497	0.808**

*p-value ≤ 0.05, **p-value ≤ 0.01

ตอนที่ 3 การวิเคราะห์ความเที่ยงระหว่างผู้ประเมินโดยใช้ข้อมูลจริง

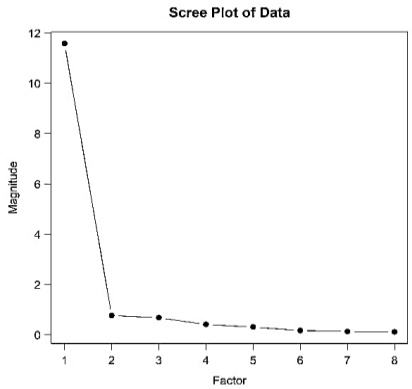
ข้อมูลที่ใช้ในการศึกษาครั้งนี้ คือ ผลการประเมินความสอดคล้องในแนวเดียวกันระว่างมาตรฐานและตัวชี้วัด กับข้อสอบในการประเมินระดับชั้นเรียน กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับชั้นมัธยมศึกษาตอนต้น ข้อมูลดังกล่าวเป็นข้อมูลทุติยภูมิจากการศึกษาเรื่อง *ความสอดคล้องในแนวเดียวกันระหว่างข้อสอบในการประเมินระดับชาติกับข้อสอบในการประเมินระดับชั้นเรียน กลุ่มสาระการเรียนรู้วิทยาศาสตร์: การประยุกต์ใช้โมเดลหลายองค์ประกอบของราส์ช และทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัด* ศึกษาโดย บุษยารัตน์ จันทร์ประเสริฐ (2560)

การวิเคราะห์ข้อมูลในการศึกษาครั้งนี้ ผู้วิจัยใช้ผลการประเมินความสอดคล้องในแนวเดียวกันฯ จำนวน 40 ข้อ ที่ประเมินโดยผู้ประเมิน จำนวน 20 คน นำมาวิเคราะห์โดยใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมสำหรับข้อมูลเรียงอันดับ หรือ Latent Truth Rater Model (LTRM) รายละเอียดของการวิเคราะห์ข้อมูลแบ่งเป็น 1. ขั้นตอนการวิเคราะห์ความเที่ยงระหว่างผู้ประเมินโดยโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม และ 2. ผลการวิเคราะห์ข้อมูล

ผลการตรวจสอบจำนวนกลุ่มวัฒนธรรม

ผู้วิจัยทำการตรวจสอบ scree plot ของผลการประเมินความสอดคล้องในแนวเดียวกันฯ รูป 4.15 จะเห็นว่ากราฟแสดงค่าไอเกนจำนวน 1 องค์ประกอบ แสดงให้เห็นว่าคำตอบของการประเมินในครั้งนี้ไม่มีการทำหน้าที่ต่างกันของผู้ประเมิน ดังนั้น ผู้วิจัยจึงใช้โมเดล LTRM สำหรับวิเคราะห์ข้อมูลที่ไม่มีความแตกต่างกันของกลุ่มวัฒนธรรมของผู้ตอบ/ผู้ประเมินในการวิเคราะห์ข้อมูลขั้นต่อไป

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

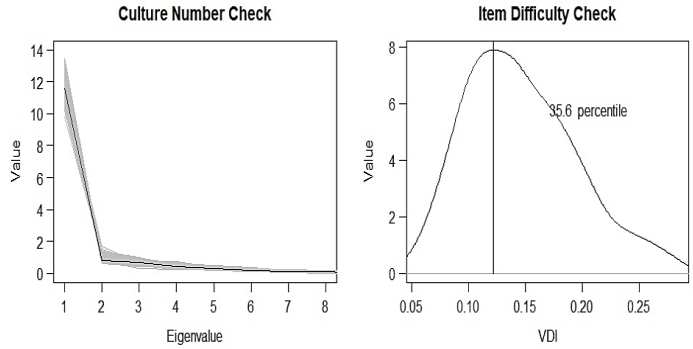


รูป 4.15 จำนวนองค์ประกอบที่แสดงจำนวนกลุ่มวัฒนธรรมของผู้ประเมิน

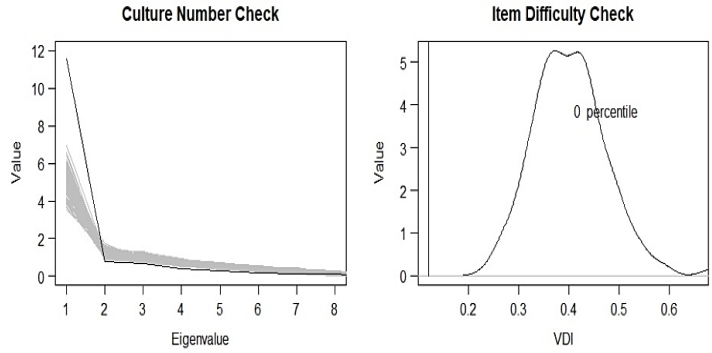
ผลการวิเคราะห์พารามิเตอร์ความยากของรายการประเมิน

รูป 4.16 แสดงการเปรียบเทียบผลการวิเคราะห์ความยากของรายการประเมินจากคำสั่ง cctapply (X) ในชุดคำสั่ง CCTpack ซึ่งคำสั่ง itemdiff = T เป็นโมเดลการวิเคราะห์ที่รวมพารามิเตอร์ความยากของรายการประเมิน $(\lambda_{k\neq1})$ ในการวิเคราะห์ ส่วนคำสั่ง itemdiff = F จะไม่รวมพารามิเตอร์ความยากของรายการประเมินในการวิเคราะห์ $(\lambda_{k=1})$ ผลการวิเคราะห์พบว่า เมื่อกำหนดการประมาณค่า $(\lambda_{k\neq1})$ แล้ว โมเดลมีความสอดคล้องกับข้อมูลมากกว่าโมเดลที่ไม่รวมการประมาณค่า$(\lambda_{k=1})$ เมื่อพิจารณาแผนภาพของค่าไอเกนจะเห็นว่าภาพ (ก.) แสดงผลการประมาณค่า (เส้นทึบสีดำ) ที่สอดคล้องกับข้อมูลมากกว่าภาพ (ข.) ค่า VDI ของการประมาณค่าด้วยโมเดล LTRM$\lambda_{k\neq1}$ อยู่ที่เปอร์เซ็นไทล์ที่ 35.6 และมีค่า DIC เท่ากับ 1171.26 ในขณะที่ผลการประมาณค่าของโมเดล LTRM$\lambda_{k=1}$ มีค่า VDI อยู่ที่เปอร์เซ็นไทล์ที่ 0 และมีค่า DIC 2696.6

จากผลการประมาณค่าของโมเดลดังที่กล่าวมาข้างต้น ผู้วิจัยจังเลือกวิเคราะห์ข้อมูลตามผลการประมาณค่าของโมเดล LTRM$\lambda_{k\neq1}$ โดยพิจารณาโมเดลที่มีค่า DIC ต่ำกว่า และดำเนินการแปลผลการวิเคราะห์ข้อมูลต่อไป



(ก.)



(ข.)

รูป 4.16 ผลการวิเคราะห์การทำหน้าที่ระหว่างผู้ประเมินด้วยโมเดล LTRM

ผลการวิเคราะห์ข้อมูล

เมื่อเลือกโมเดลที่เหมาะสมกับข้อมูลแล้ว ผู้วิจัยแปลผลการวิเคราะห์ข้อมูลจากผลการประมาณค่าในข้อ 2) ทั้งนี้ โมเดล LTRM ให้ผลการประมาณค่าพารามิเตอร์ทั้งสิ้น 5 พารามิเตอร์ ในรูปของค่าเฉลี่ยการแจกแจงภายหลังของพารามิเตอร์ (Posterior mean) ประกอบด้วย **1) คะแนนฉันทามติของการประเมิน** $(\boldsymbol{T_k})$ เป็นพารามิเตอร์ที่ใช้สำหรับพิจารณาฉันทามติของคะแนนการประเมินว่าผลการประเมินแต่ละรายการควรมีคะแนนเท่าใด **2) เทรชโฮลด์ร่วมระหว่างผู้ประเมิน** $(\boldsymbol{\gamma_c})$ แสดงการแจกแจงของระดับการเปลี่ยนช่วงของการให้คะแนนที่เพิ่มขึ้นหรือลดลงของผู้ประเมิน **3) ความยากของรายการประเมิน** $(\boldsymbol{\lambda_k})$ แสดงระดับความยากของคำถามประเมินที่ส่งผลต่อการให้คะแนนของผู้ประเมิน **4) ความสามารถของผู้ประเมิน** $(\boldsymbol{E_i})$ แสดงระดับความสามารถของผู้เชี่ยวชาญในการประเมินได้ตรงกับฉันทามติของกลุ่ม และ **5) ความลำเอียงในการประเมิน** $(\boldsymbol{a_i},\ \boldsymbol{b_i})$ เป็นพารามิเตอร์ที่แสดงแนวโน้มการกดและปล่อยคะแนนการประเมินของผู้ประเมิน

ผลการวิเคราะห์ผลการประเมินจากแบบบันทึกระดับความซับซ้อนทางปัญญาของข้อสอบในการประเมินระดับชาติ กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับชั้นมัธยมศึกษาปีที่ 3

1. ผลการประมาณค่าพารามิเตอร์ตำแหน่งคะแนนการประเมินระดับความซับซ้อนทางปัญญาของข้อสอบ (T_k) และเทรชโฮลด์ร่วมระหว่างผู้ประเมิน (γ_c)

ผลคะแนนการประเมินระดับความซับซ้อนทางปัญญาของข้อสอบในการประเมินระดับชาติ กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับชั้นมัธยมศึกษาปีที่ 3 พิจารณาจากการประมาณค่าพารามิเตอร์ตำแหน่งคะแนนการประเมิน (T_k) ของผู้ประเมินจำนวน 20 คน ต่อรายการประเมิน 40 ข้อ ดังแสดงในรูป 4.15 นอกจากนี้ โมเดล LTRM แตกต่างจากโมเดล GCM เนื่องจากเพิ่มพารามิเตอร์เทรชโฮลด์ของผู้ประเมิน (γ_c) เป็นแนวโน้มการเปลี่ยนระดับคะแนนจากระดับหนึ่งไปสู่ระดับที่สูงขึ้นหรือต่ำลง ตาราง 4.6 แสดงการประมาณค่าพารามิเตอร์ T_k มีค่าเฉลี่ยการแจกแจงภายหลัง (posterior mean) ระหว่าง -6.334 ถึง 1.160 และมีค่าเฉลี่ยของค่าเฉลี่ยการแจกแจงภายหลัง μ_{T_k} เท่ากับ -1.689 และค่าเฉลี่ยความแม่นยำในการประมาณค่า τ_{T_k} เท่ากับ 0.247 ส่วนพารามิเตอร์ γ_c มีค่าเฉลี่ยการแจกแจงภายหลังอยู่ระหว่าง -3.793 ถึง 2.522

ตาราง 4.6 ผลการประมาณค่าพารามิเตอร์ตำแหน่งคะแนนการประเมินระดับความซับซ้อนทางปัญญาของข้อสอบ (T_k)

Item	Post. Mean	Post. SD	HDI 2.50%	25%	50%	75%	97.50%
1	-0.887	0.507	-2.038	-1.190	-0.826	-0.521	-0.075
2	-2.999	0.825	-4.579	-3.571	-3.011	-2.424	-1.398
3	-5.025	1.346	-7.666	-5.923	-5.050	-4.111	-2.411
4	-0.282	0.455	-1.298	-0.545	-0.235	0.020	0.511
5	0.706	0.305	0.160	0.490	0.690	0.900	1.338
6	0.469	0.169	0.160	0.350	0.464	0.584	0.814
7	-3.680	0.949	-5.432	-4.351	-3.722	-3.044	-1.758
8	0.492	0.177	0.168	0.367	0.485	0.611	0.855
9	-1.206	0.589	-2.537	-1.572	-1.142	-0.768	-0.263
10	-5.193	1.413	-8.018	-6.118	-5.201	-4.253	-2.433
11	-0.350	0.450	-1.347	-0.609	-0.304	-0.049	0.422
12	-4.505	1.198	-6.850	-5.306	-4.533	-3.706	-2.144
13	-2.356	0.782	-3.975	-2.876	-2.322	-1.791	-0.961
14	-2.938	1.108	-5.383	-3.637	-2.850	-2.129	-1.064
15	-5.200	1.378	-7.869	-6.120	-5.239	-4.261	-2.511

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

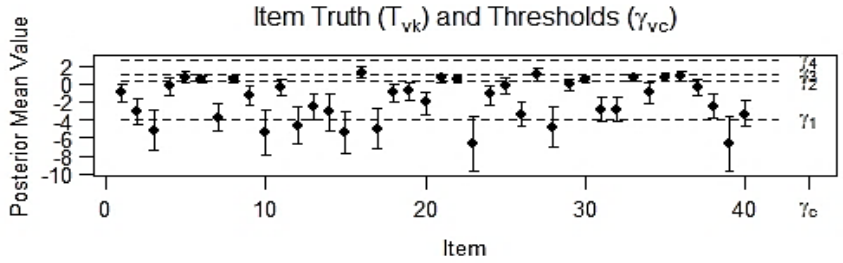
Item	Post. Mean	Post. SD	HDI				
			2.50%	25%	50%	75%	97.50%
16	1.160	0.370	0.484	0.901	1.151	1.402	1.921
17	-4.792	1.288	-7.319	-5.657	-4.808	-3.926	-2.253
18	-0.925	0.569	-2.226	-1.271	-0.846	-0.493	-0.063
19	-0.810	0.486	-1.921	-1.102	-0.751	-0.457	-0.038
20	-1.979	0.690	-3.439	-2.422	-1.940	-1.485	-0.773
21	0.560	0.209	0.185	0.411	0.551	0.698	0.985
22	0.431	0.177	0.116	0.306	0.425	0.549	0.790
23	-6.334	1.793	-10.063	-7.481	-6.302	-5.111	-2.949
24	-1.138	0.556	-2.380	-1.476	-1.077	-0.733	-0.242
25	-0.244	0.454	-1.272	-0.501	-0.198	0.054	0.549
26	-3.216	0.818	-4.732	-3.797	-3.262	-2.676	-1.543
27	0.993	0.371	0.334	0.729	0.977	1.234	1.764
28	-4.651	1.292	-7.259	-5.504	-4.650	-3.773	-2.169
29	-0.048	0.302	-0.771	-0.204	-0.008	0.151	0.442
30	0.435	0.164	0.138	0.318	0.426	0.545	0.769
31	-2.789	0.780	-4.292	-3.341	-2.797	-2.244	-1.284
32	-2.761	0.778	-4.271	-3.298	-2.771	-2.214	-1.262
33	0.601	0.190	0.249	0.467	0.598	0.733	0.977
34	-0.938	0.557	-2.179	-1.274	-0.877	-0.542	-0.014
35	0.712	0.217	0.301	0.562	0.711	0.859	1.141
36	0.826	0.276	0.326	0.631	0.815	1.008	1.394
37	-0.392	0.447	-1.413	-0.650	-0.340	-0.086	0.364
38	-2.374	0.724	-3.835	-2.864	-2.356	-1.858	-1.052
39	-6.317	1.782	-9.986	-7.461	-6.291	-5.113	-2.925
40	-3.213	0.816	-4.715	-3.794	-3.260	-2.660	-1.547
μ_{T_k}	-1.689	0.537	-2.780	-2.054	-1.669	-1.303	-0.709
τ_{T_k}	0.247	0.198	0.113	0.138	0.181	0.269	0.819

ตาราง 4.7 ผลการประมาณค่าพารามิเตอร์เทรชโฮลด์ร่วมระหว่างผู้ประเมินระดับความซับซ้อนทางปัญญาของข้อสอบ (γ_c)

γ_c	Post. Mean	Post. SD	HDI				
			2.50%	25%	50%	75%	97.50%
γ_1	-3.793	0.933	-5.483	-4.453	-3.868	-3.181	-1.847
γ_2	0.233	0.126	0.006	0.144	0.228	0.318	0.493
γ_3	1.038	0.264	0.493	0.861	1.057	1.221	1.528
γ_4	2.522	0.647	1.237	2.100	2.535	2.973	3.771

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

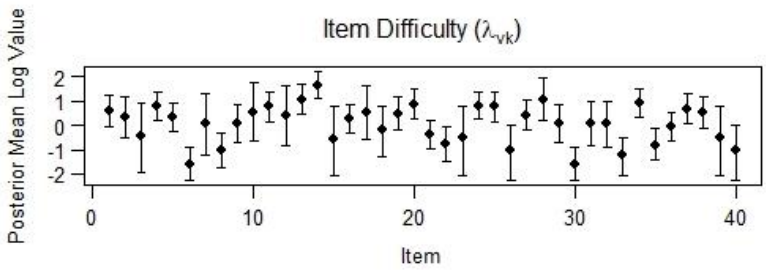
เมื่อพิจารณาจากรูป 4.17 จะเห็นภาพรวมของการประเมินทั้ง 40 รายการ (จุดทึบสีดำ) ว่ามีตำแหน่งคะแนนการประเมินอยู่ในช่วงเทรชโฮลด์ 2 และ 3 คือในตำแหน่งคะแนนประเมินระดับ 2 (เข้าใจ) ถึงระดับ 4 (ประยุกต์ใช้) นอกจากนี้ เทรชโฮลด์ร่วมของผู้ประเมินมีระยะห่างไม่เท่ากันในแต่ละระดับการประเมิน แสดงถึงความไม่สม่ำเสมอของการเปลี่ยนระดับการให้คะแนนในการประเมินที่ผู้ประเมินมีแนวโน้มจะเปลี่ยนไปสู่การให้คะแนนในระดับที่สูงขึ้น γ_2 และ γ_3 มีตำแหน่งใกล้เคียงกัน แสดงว่าผู้ประเมินมีความน่าจะเป็นที่จะเปลี่ยนระดับการประเมินระหว่างคะแนนในสองระดับ คือ ระหว่างระดับ “เข้าใจ” ไปสู่ระดับ “ประยุกต์ใช้” (γ_2) และระดับ “ประยุกต์ใช้” ไปสู่ระดับ “วิเคราะห์” (γ_3) ในการประเมินข้อที่ 4, 5, 6, 8, 16, 21, 22, 27, 29, 30, 33, 35 และ 36 นอกจากนี้ รายการประเมินที่มีค่า Posterior Mean ของ T_k ต่ำกว่าค่า Posterior Mean ของ γ_1 (-3.793) หมายถึงเป็นข้อคำถามที่ได้รับการประเมินในระดับ “ความจำ” ได้แก่ ข้อ 3, 10, 11, 14, 16, 22, 27 และ 29



รูป 4.17 ค่าเฉลี่ยภายหลังการประมาณค่าพารามิเตอร์ T_k และ γ_c

2. ผลการประมาณค่าความยากของรายการประเมินระดับความซับซ้อนทางปัญญาของข้อสอบ (λ_k)

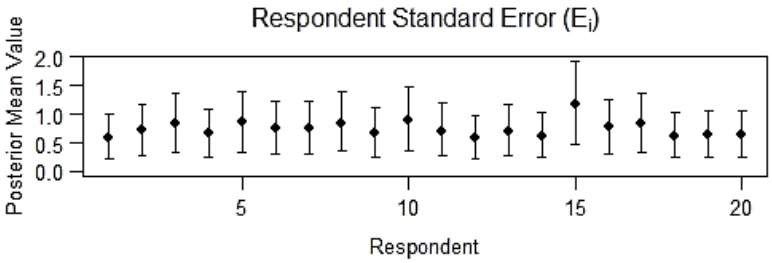
จากการพิจารณาค่า DIC ระหว่างโมเดล LTRM$\lambda_{k\neq1}$ และโมเดล LTRM$\lambda_{k=1}$ ในข้อ 2.1 (2) ผู้วิจัยจึงเลือกวิเคราะห์ข้อมูลตามผลการประมาณค่าของโมเดล LTRM$\lambda_{k\neq1}$ ซึ่งมีการประมาณค่าพารามิเตอร์ความยากของรายการประเมิน (λ_k) ผลการประมาณค่าพารามิเตอร์ความยากของรายการประเมินความระดับความซับซ้อนทางปัญญาของข้อสอบ จำนวน 40 ข้อ มีค่าเฉลี่ยภายหลังของการประมาณค่าระหว่าง -1.593 ถึง 1.603 ทั้งนี้ รายการประเมินที่มีค่าความยากต่ำสุด ได้แก่ ข้อ 30 ส่วนรายการประเมินที่มีค่าสูงสุด ได้แก่ข้อ 14 ดังแสดงในรูป 4.18



รูป 4.18 ค่าเฉลี่ยการแจกแจงภายหลังของการประมาณค่าพารามิเตอร์ λ_k

3. ผลการประมาณค่าความสามารถของผู้ประเมินในการประเมินระดับความซับซ้อนทางปัญญาของข้อสอบ (E_i)

ความสามารถของผู้ประเมิน (E_i) เป็นพารามิเตอร์ที่เกี่ยวกับความเที่ยงตรงในการกำหนดตำแหน่งคะแนนในการประเมิน โดยความเที่ยงตรงของการประเมินมีค่า $\tau_{ik} = \frac{E_i}{\lambda_k}$ ผลการประมาณค่าความสามารถของผู้ประเมินมีค่าเฉลี่ยภายหลังของการประมาณค่าพารามิเตอร์ E_i มีค่าระหว่าง 0.580 ถึง 1.158 มีค่าเฉลี่ยความแม่นยำในการประมาณค่า τ_{E_i} เท่ากับ 22.541 จากรูป 4.19 จะเห็นว่าค่าเฉลี่ยภายหลังของการประมาณค่าพารามิเตอร์ความสามารถของผู้ประเมินแต่ละคนมีค่าใกล้เคียงกัน ยกเว้นผู้ประเมินคนที่ 15 เมื่อพิจารณาจากจุดสีดำซึ่งเป็นค่าเฉลี่ยความน่าจะเป็นภายหลังของความแม่นยำในการประเมิน พบว่ามีค่าสูงที่สุด แต่ก็พบว่ามีการกระจายของการประมาณค่าสูงกว่าผู้ประเมินคนอื่นในกลุ่มเดียวกัน แสดงให้เห็นว่า แม้ผู้ประเมินคนที่ 15 จะมีค่าเฉลี่ยภายหลังของความสามารถสูง แต่ก็มีความแปรปรวนในการให้คะแนนมากกว่าผู้ประเมินคนอื่น



รูป 4.19 ค่าเฉลี่ยการแจกแจงภายหลังของการประมาณค่าพารามิเตอร์ E_i

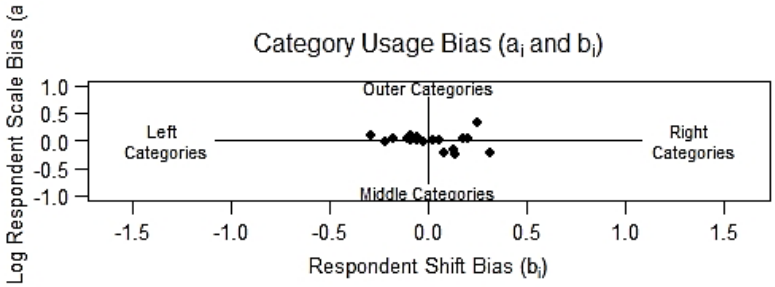
4. ผลการประมาณค่าความลำเอียงในการประเมิน $(a_i,\ b_i)$

Response Biases หรือพารามิเตอร์ความลำเอียงในการประเมิน $(a_i,\ b_i)$ เป็นการประมาณค่าแนวโน้มการกำหนดตำแหน่งคะแนนการประเมินของผู้ประเมิน พารามิเตอร์ a_i (scaling bias) ประมาณค่าการหดหรือขยายช่วงระหว่างเทรชโฮลด์ร่วมของผู้ประเมิน ในขณะที่พารามิเตอร์ b_i (shifting bias) ประมาณค่าการเปลี่ยนตำแหน่งของการให้คะแนนไปทางด้านซ้ายหรือขวาของสเกล

ผลการประมาณค่าพารามิเตอร์ a_i และ b_i ของผู้ประเมินจำนวน 20 คน มีค่า scaling bias ระหว่าง 0.800 ถึง 1.400 และมีค่า shifting bias ระหว่าง -0.210 ถึง 0.186 เมื่อพิจารณาจากรูป 4.20 แสดงแผนภาพที่พล็อตระหว่าง a_i กับ b_i โดยแบ่งพื้นที่เป็น 4 ส่วน ค่าเฉลี่ยภายหลังของการประมาณค่าแสดงว่าผู้ประเมินส่วนใหญ่ประเมินระดับความซับซ้อนทางปัญญาของข้อสอบอยู่ในช่วงกลางของมาตรประมาณค่าและมีการประเมินที่ใกล้เคียงกัน รวมถึงมีน้ำหนักการให้คะแนนนระดับต่ำและสูงที่ใกล้เคียงกัน

เมื่อพิจารณาร่วมกับผลการประเมินระดับความซับซ้อนทางปัญญาของข้อสอบ T_k จะเห็นว่าผลการประเมินอยู่ระหว่างตำแหน่งคะแนนประเมินระดับ 2 (เข้าใจ) ถึงระดับ 4 (ประยุกต์ใช้) จากระดับการประเมิน 6 ระดับ ซึ่งเป็นตำแหน่งที่อยู่ในช่วงกลางของมาตรการประเมิน

จากรูป 4.20 พบว่า ในการประเมินระดับความซับซ้อนทางปัญญาของข้อสอบ ผู้ประเมินกลุ่มนี้มีความไวในการเปลี่ยนระดับคะแนนในการประเมินใกล้เคียงกัน โดยมีค่า a_i ระหว่าง 0.774 – 1.100 และมีแนวโน้มของการกดหรือปล่อยคะแนนในระดับปานกลาง โดยมีค่า b_i ระหว่าง -0.289 – 0.317



รูป 4.20 ค่าเฉลี่ยการแจกแจงภายหลังของการประมาณค่าพารามิเตอร์ a_i และ b_i

ผลการวิเคราะห์ผลการประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียน กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น

การประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียน กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น เป็นการประเมินระดับความสอดคล้องระหว่างข้อสอบกับตัวชี้วัด โดยมีระดับความสอดคล้อง 5 ระดับ คือ 0 = ไม่สอดคล้อง 1 =ค่อนข้างไม่สอดคล้อง 2 = ไม่แน่ใจ 3 = ค่อนข้างสอดคล้อง 4 = สอดคล้องโดยตรง ผู้วิจัยนำข้อมูลการประเมินความสอดคล้องในแนวเดียวกันของข้อสอบจำนวน 20 ข้อ จากผู้ประเมินจำนวน 40 คน มาวิเคราะห์ด้วยโมเดล LTRM โดยแปลงระดับการให้คะแนน จาก 0 – 4 เป็น 1 – 5 ดังนี้ 1 = ไม่สอดคล้อง 2 =ค่อนข้างไม่สอดคล้อง 3 = ไม่แน่ใจ 4 = ค่อนข้างสอดคล้อง 5 = สอดคล้องโดยตรง จากนั้นประมาณค่าพารามิเตอร์โดยโมเดล LTRM$\lambda_{k\neq 1}$ และโมเดล LTRM$\lambda_{k=1}$ มีค่า DIC LTRM$\lambda_{k\neq 1}$ เท่ากับ 993.519 และ LTRM$\lambda_{k=1}$ เท่ากับ 1643.501 ผู้วิจัยจึงเลือกผลการประมาณค่าจากโมเดล LTRM$\lambda_{k\neq 1}$ ในการแปลและสรุปผลการวิเคราะห์ข้อมูล

1. ผลการประมาณค่าพารามิเตอร์ตำแหน่งคะแนนการประเมินความสอดคล้องในแนวเดียวกัน (T_k) และเทรชโฮลด์ร่วมระหว่างผู้ประเมิน (γ_c)

ผลการประมาณค่าพารามิเตอร์ตำแหน่งคะแนนการประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานกับตัวชี้วัด (T_k) จากผู้ประเมิน 40 คน เป็นดังตารางที่ 4.8 ค่าเฉลี่ยภายหลังของการแจกแจงการประมาณค่าพารามิเตอร์ T_k มีค่าระหว่าง 1.540 ถึง 6.104 มีค่าเฉลี่ยของค่าเฉลี่ยการแจกแจงภายหลัง(μ_{T_k}) เท่ากับ 4.126 และค่าเฉลี่ยความถูกต้องแม่นยำของการประมาณค่า (τ_{T_k}) เท่ากับ 0.703 เมื่อพิจารณาจากระดับเทรชโฮลด์ในตาราง 4.9 มีค่าระหว่าง -1.623 ถึง 1.921 แสดงให้เห็นว่าระยะห่างของเทรชโฮลด์ร่วมอยู่ใกล้กันมาก ดังจะเห็นได้จากรูป 4.21 เมื่อพิจารณาเทียบกับพารามิเตอร์ T_k แล้ว ตำแหน่งคะแนนการประเมินจะอยู่ในเทรชโฮลด์ที่ 4 หรือระดับการประเมินที่ 5 (สอดคล้องโดยตรง)

ตาราง 4.8 ผลการประมาณค่าพารามิเตอร์ตำแหน่งคะแนนการประเมินความสอดคล้องในแนวเดียวกัน (T_k)

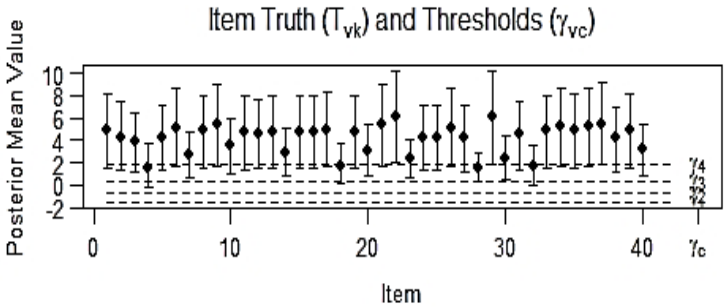
Item	Post. Mean	Post. SD	HDI 2.50%	25%	50%	75%	97.50%
1	4.867	1.697	1.846	3.670	4.761	5.926	8.547
2	4.307	1.597	1.578	3.163	4.176	5.300	7.814
3	3.907	1.389	1.464	2.935	3.788	4.758	6.928
4	1.578	1.033	0.034	0.849	1.412	2.137	4.035
5	4.296	1.537	1.607	3.205	4.191	5.246	7.647
6	5.119	1.796	1.947	3.864	4.993	6.239	9.025
7	2.663	1.075	0.925	1.893	2.550	3.291	5.095
8	4.889	1.688	1.872	3.710	4.780	5.928	8.507
9	5.482	1.904	2.087	4.138	5.361	6.671	9.574
10	3.484	1.263	1.307	2.604	3.385	4.249	6.266
11	4.748	1.738	1.745	3.520	4.613	5.804	8.642
12	4.534	1.603	1.714	3.417	4.426	5.515	8.036
13	4.749	1.656	1.824	3.592	4.620	5.779	8.343
14	2.955	1.091	1.078	2.183	2.861	3.618	5.391
15	4.840	1.678	1.847	3.666	4.735	5.881	8.501
16	4.832	1.680	1.826	3.665	4.703	5.873	8.474
17	4.956	1.734	1.868	3.755	4.835	6.030	8.766
18	1.698	0.930	0.294	1.032	1.568	2.207	3.864
19	4.839	1.681	1.810	3.677	4.733	5.872	8.438
20	3.118	1.225	1.097	2.258	2.996	3.825	5.884
21	5.446	1.912	2.042	4.098	5.319	6.668	9.551
22	6.097	2.099	2.319	4.636	5.972	7.416	10.646
23	2.364	0.888	0.858	1.733	2.288	2.902	4.344
24	4.266	1.523	1.588	3.193	4.160	5.195	7.615
25	4.283	1.501	1.600	3.235	4.185	5.204	7.528
26	5.176	1.798	1.988	3.914	5.048	6.316	9.037
27	4.322	1.525	1.634	3.269	4.201	5.266	7.643
28	1.540	0.669	0.460	1.056	1.476	1.943	3.044
29	6.104	2.131	2.309	4.596	5.960	7.441	10.748
30	2.425	1.057	0.744	1.658	2.296	3.046	4.827
31	4.505	1.599	1.718	3.380	4.401	5.479	7.963
32	1.631	0.958	0.213	0.937	1.481	2.164	3.895
33	4.900	1.697	1.852	3.716	4.782	5.963	8.505
34	5.285	1.810	2.026	4.012	5.175	6.437	9.148
35	4.851	1.716	1.841	3.646	4.709	5.885	8.641

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Item	Post. Mean	Post. SD	HDI 2.50%	25%	50%	75%	97.50%
36	5.287	1.818	1.995	4.008	5.178	6.460	9.154
37	5.501	1.911	2.107	4.162	5.367	6.698	9.600
38	4.288	1.500	1.614	3.253	4.178	5.207	7.574
39	4.859	1.700	1.824	3.658	4.742	5.920	8.590
40	3.159	1.226	1.137	2.282	3.039	3.906	5.890
μ_{T_k}	4.126	1.244	1.650	3.282	4.132	4.962	6.590
τ_{T_k}	0.703	0.833	0.142	0.287	0.458	0.773	3.099

ตาราง 4.9 ผลการประมาณค่าพารามิเตอร์เทรชโฮลด์ร่วมระหว่างผู้ประเมินระดับความสอดคล้องในแนวเดียวกัน (γ_c)

Item	Post. Mean	Post. SD	HDI 2.50%	25%	50%	75%	97.50%
γ_1	-1.623	0.574	-2.856	-1.980	-1.579	-1.224	-0.596
γ_2	-0.704	0.256	-1.263	-0.868	-0.684	-0.522	-0.252
γ_3	0.406	0.184	0.121	0.272	0.382	0.513	0.829
γ_4	1.921	0.638	0.719	1.488	1.890	2.345	3.254

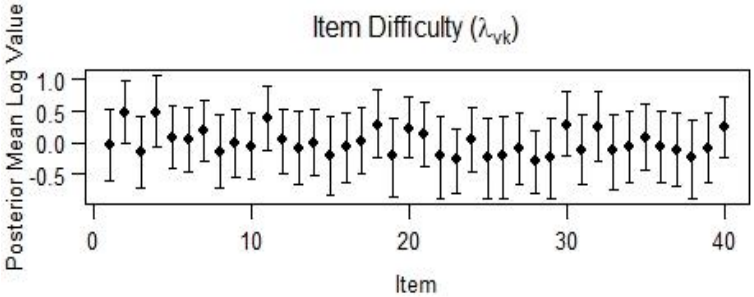


รูป 4.21 ค่าเฉลี่ยภายหลังการประมาณค่าพารามิเตอร์ T_k และ γ_c

2. ผลการประมาณค่าความยากของรายการประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานกับตัวชี้วัด (λ_k)

ค่าเฉลี่ยภายหลังการแจกแจงพารามิเตอร์ความยากของรายการประเมินความสอดคล้องในแนวเดียวกัน มีค่าเฉลี่ยการแจกแจงภายหลังของพารามิเตอร์ λ_k ระหว่าง -0.295 ถึง 0.476 รายการประเมินที่มีค่าเฉลี่ยภายหลังการแจกแจงพารามิเตอร์สูงสุด ได้แก่ข้อ 4 และต่ำสุดได้แก่ข้อ 28 เมื่อ

พิจารณารูปที่ 4.22 จะเห็นว่ารายการประเมินแต่ละรายการมีค่าเฉลี่ยการแจกแจงภายหลังของพารามิเตอร์ความยากของรายการประเมินใกล้เคียงกัน โดยมีค่าเฉลี่ยการแจกแจงภายหลังของความถูกต้องแม่นยำในการประมาณค่า (τ_{λ_k}) เท่ากับ 9.132

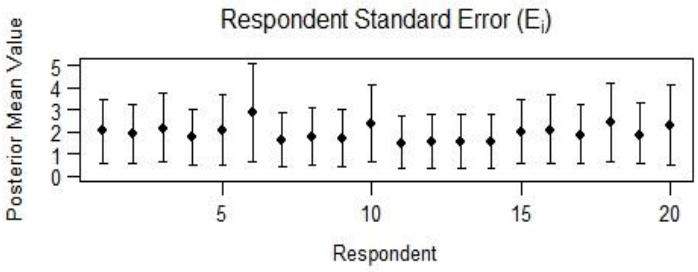


รูป 4.22 ค่าเฉลี่ยการแจกแจงภายหลังของการประมาณค่าพารามิเตอร์ λ_k

3. ผลการประมาณค่าความสามารถของผู้ประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานกับตัวชี้วัด (E_i)

ผลการประมาณค่าความสามารถของผู้ประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานกับตัวชี้วัด (E_i) ของผู้ประเมินจำนวน 20 คน มีค่าระหว่าง 1.504 ถึง 2.856 มีค่าเฉลี่ยของค่าเฉลี่ยการแจกแจงภายหลังของพารามิเตอร์ (μ_{E_i}) เท่ากับ 0.564 และค่าเฉลี่ยของค่าเฉลี่ยความถูกต้องแม่นยำในการประมาณค่า (τ_{E_i}) เท่ากับ 24.266 จากรูป 4.23 จะเห็นว่าผู้ประเมินคนที่ 6 มีค่าเฉลี่ยภายหลังของการประมาณค่าความสามารถในการประเมินสูงที่สุด แต่ก็มีการกระจายของการประมาณค่าความสามารถในการประเมินสูงที่สุดในกลุ่มด้วย รองลงมาคือผู้ประเมินคนที่ 18 10 20 และ 3 ตามลำดับ

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY



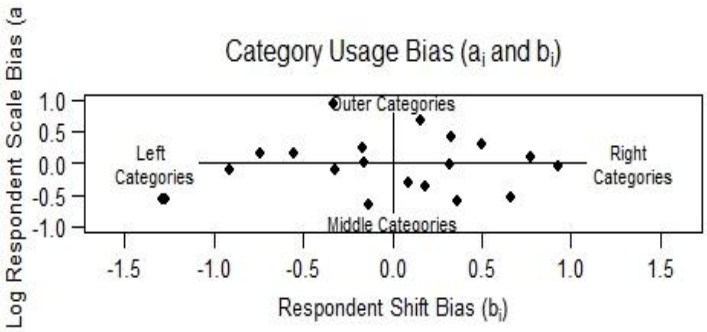
รูป 4.23 ค่าเฉลี่ยการแจกแจงภายหลังของการประมาณค่าพารามิเตอร์ E_i

4. ผลการประมาณค่าความลำเอียงในการประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานกับตัวชี้วัด $(\boldsymbol{a_i},\ \boldsymbol{b_i})$

ผลการประมาณค่าพารามิเตอร์ความลำเอียงในการประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานกับตัวชี้วัด $(a_i,\ b_i)$ จากรายการประเมินจำนวน 40 ข้อ ของผู้ประเมินจำนวน 20 คน มีค่า scaling bias (a_i) ระหว่าง 0.552 ถึง 2.592 และมีค่า shifting bias (b_i) ระหว่าง -1.290 ถึง 0.926

รูป 4.24 แสดงแผนภาพที่พล็อตระหว่าง a_i กับ b_i แสดงการกระจายตัวของค่าเฉลี่ยภายหลังของการประมาณค่าในพื้นที่ของกราฟ ส่วนที่น่าสนใจ ได้แก่ ผู้ประเมินคนที่ 15 ความไวในการเปลี่ยนระดับคะแนนสูงที่สุดในกลุ่ม และมีแนวโน้มการเปลี่ยนระดับคะแนนไปยังระดับคะแนนที่ต่ำกว่า (0.552, -0.334) ผู้ประเมินคนที่ 18 มีความไวในการเปลี่ยนระดับคะแนนต่ำกว่าผู้ประเมินคนอื่น ๆ ในกลุ่มและมีแนวโน้มที่จะเปลี่ยนระดับการประเมินไปในทางบวก (1.575, 0.327) ในขณะที่ผู้ประเมินคนที่ 4 มีความไวในการเปลี่ยนระดับคะแนนต่ำและมีแนวโน้มที่จะเปลี่ยนระดับการประเมินไปในทางลบ หรือมีแนวโน้มที่จะกดคะแนนในการประเมิน (2.592, -0.322)

กล่าวโดยสรุป ผู้ประเมินที่มีค่าเฉลี่ยภายหลังของพารามิเตอร์ $-0.5 > a_i\ , |b_i| > 0.5$ มีแนวโน้มที่จะมีความแปรปรวนในการให้คะแนนในการประเมินมากกว่าผู้ประเมินที่มีค่าเฉลี่ยภายหลังของพารามิเตอร์ $-0.5 \leq a_i\ , |b_i| \leq 0.5$



รูป 4.24 ค่าเฉลี่ยการแจกแจงภายหลังของการประมาณค่าพารามิเตอร์ a_i และ b_i

บทที่ 5
สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

การวิจัยครั้งนี้มีวัตถุประสงค์ 1) เพื่อประยุกต์ใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม เพื่อวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินในการศึกษาความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียนกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น 2) เพื่อตรวจสอบประสิทธิภาพของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมในการการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินในการศึกษาความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียนกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น 3) เพื่อศึกษาผลการประยุกต์ใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมในการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินและการทำหน้าที่ต่างกันของผู้ประเมินในการประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียนกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น

วิธีการศึกษาแบ่งเป็น 3 ระยะ คือ <u>ระยะที่ 1</u> ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง <u>ระยะที่ 2</u> การศึกษาผลการประมาณความสอดคล้องระหว่างผู้ประเมิน โดยศึกษาโมเดลวิเคราะห์ฉันทามติเชิงวัฒนธรรม 2 โมเดล ได้แก่ 1) Multi-culture General Condorcet Model และ 2) Multi-culture Latent Truth Rater Model ด้านกระบวนการในการวิเคราะห์ ปัจจัยที่ส่งผลต่อการประมาณค่าและประสิทธิภาพของโมเดลในการวิเคราะห์ความเที่ยงระหว่างผู้ประเมิน โดยใช้วิธีการจำลองข้อมูลแบบมอนติคาร์โล การตัดสินประสิทธิภาพของโมเดลพิจารณาจากค่าสัมประสิทธิ์สหสัมพันธ์ (Pearson Correlation Coefficient) ระหว่างค่าพารามิเตอร์ที่ผู้วิจัยกำหนดกับค่าที่ได้จากการประมาณของการจำลองข้อมูล ความลำเอียงในการประมาณค่า (Bias) และค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Mean Square Error: MSE) <u>ระยะที่ 3</u> ศึกษาผลการประยุกต์ใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมในการวิเคราะห์ความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบ ในการประเมินระดับชั้นเรียนกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น ซึ่งเป็นการวิเคราะห์ความเที่ยงระหว่างผู้ประเมินในบริบทการประเมินจริงด้วยโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมสำหรับมาตรประมาณค่า (LTRM)

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ข้อมูลที่ใช้ในการศึกษาในระยะที่ 3 คือ ผลการประเมินความสอดคล้องในแนวเดียวกันระว่างมาตรฐานและตัวชี้วัด กับข้อสอบในการประเมินระดับชั้นเรียน กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับชั้นมัธยมศึกษาตอนต้น ข้อมูลดังกล่าวเป็นข้อมูลทุติยภูมิจากการศึกษาเรื่อง *ความสอดคล้องในแนวเดียวกันระหว่างข้อสอบในการประเมินระดับชาติกับข้อสอบในการประเมินระดับชั้นเรียน กลุ่มสาระการเรียนรู้วิทยาศาสตร์: การประยุกต์ใช้โมเดลหลายองค์ประกอบของราส์ช และทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัด* ศึกษาโดย บุษยารัตน์ จันทร์ประเสริฐ (2560) ผู้วิจัยใช้ผลการประเมินจำนวน 40 ข้อ ที่ประเมินโดยผู้ประเมิน จำนวน 20 คน นำมาวิเคราะห์โดยใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมสำหรับข้อมูลเรียงอันดับ หรือ Latent Truth Rater Model (LTRM) ข้อมูลแบ่งเป็น 2 ส่วน คือ 1) ผลการประเมินจากแบบบันทึกระดับความซับซ้อนทางปัญญาของข้อสอบในการประเมินระดับชาติ กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับชั้นมัธยมศึกษาปีที่ 3 เป็นการประเมินลำดับขั้นของกระบวนการทางปัญญาจำแนกตามพฤติกรรมการเรียนรู้ด้านพุทธิพิสัยของ Bloom (Bloom’s Taxonomy) ผลการประเมินเป็นการจัดกลุ่มตัวบ่งชี้ตามพฤติกรรมการเรียนรู้ 6 กลุ่ม ได้แก่ จำ เข้าใจ ประยุกต์ใช้ วิเคราะห์ ประเมินค่า และสร้างสรรค์ 2) ผลการประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียน กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น เป็นการประเมินความสอดคล้องระหว่างข้อสอบกับตัวชี้วัด จำนวน 40 ข้อ ผลการประเมินเป็นมาตรประมาณค่า 5 ระดับ ได้แก่ ไม่สอดคล้อง ค่อนข้างไม่สอดคล้อง ไม่แน่ใจ ค่อนข้างสอดคล้อง สอดคล้องโดยตรง

สรุปผลการวิจัย

1. ขั้นตอนในการวิเคราะห์และสารสนเทศที่ได้จากการศึกษาความสอดคล้องระหว่างผู้ประเมินในการศึกษาความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียน กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น ด้วยโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม

โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมประกอบด้วยพารามิเตอร์ 2 กลุ่ม คือ พารามิเตอร์ของข้อคำถาม และพารามิเตอร์ของผู้ประเมิน ซึ่งมีรายละเอียดต่างกันในแต่ละโมเดล โมเดล GCM หรือ MC-GCM ซึ่งเป็นโมเดลวิเคราะห์ข้อมูลในบริบทการประเมินที่ให้คะแนนแบบ (0, 1) จะมีพารามิเตอร์ของข้อคำถาม ซึ่งประกอบด้วยพารามิเตอร์ตำแหน่งคะแนนฉันทามติซึ่งเป็นคุณลักษณะแฝงของคำถามประเมินกับพารามิเตอร์ความยากของข้อคำถาม พารามิเตอร์ของผู้ประเมิน

ประกอบด้วยพารามิเตอร์ความสามารถของผู้ประเมิน หมายถึง ความแม่นยำในการระบุตำแหน่งของคะแนนฉันทามติได้สอดคล้องกับกลุ่มผู้ประเมินคนอื่น ๆ กับพารามิเตอร์ความลำเอียงในการประเมิน ซึ่งเป็นความน่าจะเป็นที่ผู้ประเมินจะเลือกประเมินได้ตรงกับตำแหน่งของคะแนนฉันทามติเมื่อผู้ประเมินมีความลังเลในการให้คะแนนประเมิน

โมเดล LTRM หรือ MC-LTRM เป็นโมเดลการวิเคราะห์ข้อมูลในบริบทการประเมินที่ให้คะแนนแบบมาตรประมาณค่า ประกอบด้วยพารามิเตอร์ 2 กลุ่มเช่นเดียวกัน พารามิเตอร์กลุ่มแรก คือ พารามิเตอร์ของข้อคำถาม ประกอบด้วยพารามิเตอร์ตำแหน่งคะแนนฉันทามติของการประเมิน และพารามิเตอร์ความยากของคำถามประเมิน พารามิเตอร์กลุ่มที่สอง คือ พารามิเตอร์ของผู้ประเมิน ประกอบด้วยพารามิเตอร์ความสามารถของผู้ประเมิน ซึ่งแสดงความแม่นยำในการระบุตำแหน่งคะแนนฉันทามติที่อยู่บนสเกลคุณลักษณะแฝงของคำถามประเมิน ส่วนพารามิเตอร์ความลำเอียงในการประเมินจะเป็นการบอกความไวในการเปลี่ยนระดับของการให้คะแนนกับรูปแบบของการกดหรือปล่อยคะแนนในการประเมิน

ผลการศึกษาครั้งนี้สรุปสารสนเทศที่ได้จากการวิเคราะห์ด้วยโมเดลฉันทามติเชิงวัฒนธรรมออกเป็น 2 ส่วนตามโมเดลที่ศึกษา ได้แก่ สารสนเทศที่ได้จากโมเดล MC-GCM และสารสนเทศที่ได้จากโมเดล MC-LTRM

สารสนเทศที่ได้จากการวิเคราะห์ข้อมูลการประเมินของผู้ประเมินด้วยโมเดล MC-GCM จะมีสารสนเทศ 4 ประการหลัก คือ **1) การทำหน้าที่ต่างกันระหว่างผู้ประเมิน** ซึ่งได้จากการคำนวณจำนวนองค์ประกอบซึ่งระบุโดยค่าไอเกน สารสนเทศนี้จะช่วยให้นักวิจัยได้ข้อมูลเบื้องต้นเกี่ยวกับรูปแบบของผลการประเมินของผู้ประเมินว่ามีความสอดคล้องกันเป็นหนึ่งเดียว หรือมีความแตกต่างของการให้คะแนนการประเมินอันเนื่องมาจากปัจจัยที่เกี่ยวเนื่องกับการประเมิน ค่าไอเกนนี้ยังเป็นข้อมูลเบื้องต้นให้นักวิจัยเลือกโมเดลการวิเคราะห์ในขั้นตอนต่อไปได้อย่างเหมาะสม **2) คะแนนฉันทามติระหว่างผู้ประเมิน** ซึ่งได้จากการประมาณค่าพารามิเตอร์คำตอบฉันทามติของผู้ประเมิน ซึ่งจำแนกคำตอบฉันทามติของผู้ประเมินตามจำนวนกลุ่มวัฒนธรรมหรือการทำหน้าที่ต่างกันของผู้ประเมิน **3) ความสามารถของผู้ประเมิน** (D_{ik}) เป็นความน่าจะเป็นที่ผู้ประเมินคนที่ i จะให้คะแนนการประเมินในรายการประเมินข้อ k ตรงกับผลการประเมินที่เป็นฉันทามติของกลุ่ม พารามิเตอร์ความสามารถของผู้ประเมินคำนวณผ่านพารามิเตอร์ความยากของคำถามประเมิน (δ_k) และพารามิเตอร์ความรู้ความสามารถของผู้ประเมิน และ **4) ความลำเอียงในการประเมิน** ได้จาก

การประมาณค่าพารามิเตอร์ความลำเอียงในการเดาคำตอบ (g_i) เมื่อผู้ประเมินไม่แน่ใจหรือขาดความเชี่ยวชาญในการประเมินบางหัวข้อแต่จำเป็นต้องตัดสินคะแนน พารามิเตอร์ดังกล่าวเป็นการระบุความน่าจะเป็นที่ผู้ประเมิน i จะให้คะแนนการประเมินในรายการประเมินข้อ k เท่ากับ 1 เมื่อมีความไม่แน่ใจในการให้คะแนนการประเมิน

สารสนเทศที่ได้จากการวิเคราะห์ข้อมูลการประเมินของผู้ประเมินในการศึกษาความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบด้วยโมเดล MC-LTRM มีสารสนเทศ 4 ประการหลัก คือ **1) การทำหน้าที่ต่างกันระหว่างผู้ประเมิน** ซึ่งเป็นการระบุจำนวนของความแตกต่างของรูปแบบในการประเมินของผู้ประเมินว่ามีรูปแบบความสอดคล้องกันระหว่างผู้ประเมินทั้งสิ้นกี่รูปแบบ สารสนเทศที่ได้นี้จะนำไปสู่การเลือกโมเดลในการวิเคราะห์ในขั้นตอนต่อไป **2) คำตอบฉันทามติเชิงวัฒนธรรม** ได้จากการผลการประมาณค่าพารามิเตอร์เทรชโฮลด์ร่วมระหว่างผู้ประเมิน (γ_c) ควบคู่กับผลการประมาณค่าพารามิเตอร์ T_{vk} หรือ T_k ในกรณีของโมเดล LTRM พารามิเตอร์เทรชโฮลด์ร่วมระหว่างผู้ประเมิน (γ_c) เป็นพารามิเตอร์ที่แสดงแนวโน้มการเปลี่ยนระดับการให้คะแนนของสเกลการตอบ เมื่อ c เป็นจำนวนสเกล **3) ความยากของรายการประเมิน** แสดงให้เห็นว่าคำถามประเมินในข้อใดที่ส่งผลต่อความแม่นยำในการประเมินของผู้ประเมิน จากบทความของ Anders และ Batchelder (2015) หากเฉลี่ยภายหลังของการประมาณค่า (posterior mean) พารามิเตอร์ความยากของรายการประเมินมีค่าระหว่างครึ่งหนึ่งถึงสองเท่าของค่าเฉลี่ยภายหลังของการประมาณค่าพารามิเตอร์ความสามารถของผู้ประเมิน ถือว่ารายการประเมินนั้นมีระดับความยากของการประเมินที่เหมาะสม **4) ความแม่นยำในการประเมิน** ซึ่งพิจารณาจากผลการประมาณค่าความสามารถในการประเมินของผู้ประเมิน (E_i) สารสนเทศนี้หมายถึง ความน่าจะเป็นที่ผู้ประเมินจะให้คะแนนประเมินได้ตรงกับผลการประเมินซึ่งเป็นฉันทามติของกลุ่มผู้ประเมิน ซึ่งเป็นการพิจารณาว่าผู้ประเมินคนใดมีแนวโน้มที่จะให้คะแนนประเมินแตกต่างไปจากกลุ่ม **5) ความลำเอียงในการประเมิน** แสดงความน่าจะเป็นในการให้คะแนนการประเมินเมื่อผู้ประเมินมีความไม่แน่ใจในการระบุคะแนนหรือมีอคติต่อสิ่งที่ต้องประเมิน พารามิเตอร์ a_i เป็นการบอกความไวในการเปลี่ยนระดับการประเมินของผู้ประเมินคนที่ i ในขณะที่พารามิเตอร์ b_i บอกรูปแบบการกดหรือปล่อยคะแนนของผู้ประเมิน เมื่อนำข้อมูลจากการวิเคราะห์ข้อมูลมาพล็อตตำแหน่งของพารามิเตอร์ทั้งสองเข้าด้วยกัน $(a_i,\ b_i)$ จะสามารถมองเห็นรูปแบบของการประเมินของผู้ประเมินแต่ละคนได้

การวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินในการวิเคราะห์ความโดยโมเดลฉันทามติเชิงวัฒนธรรมมีขั้นตอนในการวิเคราะห์ข้อมูล 3 ขั้นตอน คือ 1) การตรวจสอบการทำหน้าที่ต่างกันระหว่างผู้ประเมินโดยการตรวจสอบจากค่าน้ำหนักองค์ประกอบ มีจุดประสงค์เพื่อเลือกโมเดลในการวิเคราะห์ หากมีการทำหน้าที่ต่างกันระหว่างผู้ประเมินให้เลือกใช้โมเดลการวิเคราะห์กลุ่มพหุวัฒนธรรม (multiculture - MC) 2) การวิเคราะห์พารามิเตอร์ความยากของรายการประเมิน เพื่อตรวจสอบข้อตกลงเบื้องต้นว่าคำถามประเมินควรมีความยากเท่าเทียมกัน (Anders และ Batchelder, 2012) หากมีความยากไม่เท่าเทียมกันนักวิจัยควรรวมพารามิเตอร์ความยากเข้าไปในการวิเคราะห์ข้อมูลด้วย 3) การแปลผลและสรุปผลการวิเคราะห์ นักวิจัยสรุปผลการวิเคราะห์ข้อมูลตามพารามิเตอร์ของโมเดล

2. ประสิทธิภาพของการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินในการวิเคราะห์ความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียนกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้นด้วยโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม

ประสิทธิภาพของการประมาณค่าของโมเดล MC-GCM

โมเดล MC-GCM สามารถประมาณค่าพารามิเตอร์ได้ใกล้เคียงกับค่าจริงของพารามิเตอร์ที่กำหนดไว้ โดยเฉพาะอย่างยิ่งในกรณีที่รายการประเมินมีความเป็นเอกพันธ์ของคำถามประเมิน โมเดล MC-GCM สามารถประมาณค่าพารามิเตอร์ฉันทามติเชิงวัฒนธรรม (Z_k) พารามิเตอร์ความสามารถของผู้ประเมิน (θ_i) พารามิเตอร์ความยากของรายการประเมิน (δ_k) และพารามิเตอร์ความลำเอียงในการประเมิน (g_i) ได้ใกล้เคียงกับค่าจริงของพารามิเตอร์ที่กำหนด โดยมีค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสอง และค่าความลำเอียงในการประมาณค่าที่เข้าใกล้ 0 และมีค่าสัมประสิทธิ์สหสันพันธ์ระหว่างค่าจริงและค่าที่ได้จากการประมาณค่าของโมเดลในระดับสูง ในกรณีที่รายการประเมินมีลักษณะเป็นวิวิธพันธ์จะมีความคลาดเคลื่อนของการประมาณค่าพารามิเตอร์ต่าง ๆ สูงกว่ากรณีที่รายการประเมินมีความเป็นเอกพันธ์ แต่ยังคงอยู่ในระดับที่ยอมรับได้ นอกจากนี้ ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าจริงและค่าที่ได้จากการประมาณค่ายังคงมีความสัมพันธ์กันในระดับสูง และมีนัยสำคัญทางสถิติ

ผลการวิเคราะห์ปัจจัยที่ส่งผลต่อการประมาณค่าโดยการวิเคราะห์อิทธิพลของตัวแปรอิสระ ได้แก่ จำนวนผู้ประเมิน (N) จำนวนรายการประเมิน (M) และการทำหน้าที่ต่างกันของผู้ประเมิน ที่ส่งผลต่อประสิทธิภาพของการประมาณค่าของโมเดล MC-GCM พบว่า ปัจจัยที่ส่งผลต่อประสิทธิภาพในการประมาณค่าของโมเดล คือ การทำหน้าที่ต่างกันของผู้ประเมิน ซึ่งส่งผลต่อค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสองและค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าจริงและค่าที่ได้จากการประมาณค่าของโมเดลในการประมาณค่าพารามิเตอร์ความสามารถของผู้ประเมิน และความยากของรายการประเมินอย่างมีนัยสำคัญทางสถิติ สอดคล้องกับผลการศึกษาประสิทธิภาพในการประมาณค่าของโมเดลที่พบว่าความถูกต้องแม่นยำในการประมาณค่าของโมเดลลดลงเมื่อกำหนดให้รายการประเมินมีความยากแตกต่างกัน

ประสิทธิภาพของการประมาณค่าของโมเดล MC-LTRM

ผลการตรวจสอบความถูกต้องในกรณีที่รายการประเมินมีความเป็นเอกพันธ์ของคำถามประเมิน โมเดล MC-LTRM สามารถประมาณค่าพารามิเตอร์ฉันทามติเชิงวัฒนธรรม (T_k) พารามิเตอร์ความสามารถของผู้ประเมิน (E_i) และค่าพารามิเตอร์ยากของข้อคำถาม ($\boldsymbol{\lambda_k}$) ได้ใกล้เคียงกับค่าพารามิเตอร์ที่กำหนด อย่างไรตาม กรณีที่มีการทำหน้าที่ต่างกันของผู้ประเมิน การประมาณค่าพารามิเตอร์ความสามารถของผู้ประเมินของโมเดล MC-LTRM จะมีความคลาดเคลื่อนสูงกว่ากรณีที่รายการประเมินมีความเป็นเอกพันธ์ และมีค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าจริงและค่าที่ได้จากการประมาณค่าพารามิเตอร์ความสามารถของผู้ประเมินที่ต่ำกว่ากรณีไม่มีการทำหน้าที่ต่างกันของผู้ประเมิน และค่าสัมประสิทธิ์สหสัมพันธ์จะต่ำลงเมื่อจำนวนรายการประเมินมีจำนวนมากขึ้น ผลการประมาณค่าพารามิเตอร์ความลำเอียงในการประเมิน ($\mathrm{a_i}$, $\mathrm{b_i}$) โดยโมเดล MC-LTRM พลว่าโมเดลสามารถประมาณค่าได้ใกล้เคียงกับค่าจริงของพารามิเตอร์ที่กำหนดไว้ในกรณีไม่มีการทำหน้าที่ต่างกันของผู้ประเมิน เมื่อมีการทำหน้าที่ต่างกันของผู้ประเมินในการให้คะแนนการประเมิน จะทำให้การประมาณค่ามีความถูกต้องและแม่นยำต่ำว่าการประมาณค่าของโมเดลในกรณีการประเมินที่ไม่มีการทำหน้าที่ต่างกันของผู้ประเมิน

ผลการวิเคราะห์อิทธิพลของตัวแปรอิสระ ได้แก่ จำนวนผู้ประเมิน (N) จำนวนรายการประเมิน (M) และการทำหน้าที่ต่างกันของผู้ประเมินที่ส่งผลต่อประสิทธิภาพของการประมาณค่าของโมเดล MC-LTRM พบว่า ปัจจัยที่ส่งผลต่อประสิทธิภาพในการประมาณค่าของโมเดล คือ *การทำหน้าที่ต่างกันของผู้ประเมิน* ซึ่งส่งผลต่อค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสองและค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าจริงและค่าที่ได้จากการประมาณค่าของโมเดลในการประมาณค่าพารามิเตอร์ความสามารถของผู้ประเมิน และพารามิเตอร์ความลำเอียงในการประเมินอย่างมีนัยสำคัญทางสถิติ

สอดคล้องกับผลการศึกษาประสิทธิภาพในการประมาณค่าของโมเดลที่พบว่าความถูกต้องแม่นยำในการประมาณค่าของโมเดลลดลงเมื่อกำหนดให้รายการประเมินมีความยากแตกต่างกัน

3. ผลการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินและการทำหน้าที่ต่างกันของผู้ประเมินในการประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียนกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้นด้วยการวิเคราะห์ฉันทามติเชิงวัฒนธรรม

3.1. ผลการวิเคราะห์ความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียน กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้นด้วยโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม

ผลคะแนนการประเมินระดับความซับซ้อนทางปัญญาของข้อสอบในการประเมินระดับชาติ กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับชั้นมัธยมศึกษาปีที่ 3 พิจารณาจากการประมาณค่าพารามิเตอร์ตำแหน่งคะแนนการประเมิน (T_k) ของผู้ประเมินจำนวน 20 คน ต่อรายการประเมิน 40 ข้อ ให้ผลภาพรวมของการประเมินทั้ง 40 รายการ ว่ามีตำแหน่งคะแนนการประเมินอยู่ในช่วงเทรชโฮลด์ 2 และ 3 คือในตำแหน่งคะแนนประเมินระดับ 2 (เข้าใจ) ถึงระดับ 4 (ประยุกต์ใช้) นอกจากนี้ เทรชโฮลด์ร่วมของผู้ประเมินมีระยะห่างไม่เท่ากันในแต่ละระดับการประเมิน แสดงถึงความไม่สม่ำเสมอของการเปลี่ยนระดับการให้คะแนนในการประเมินที่ผู้ประเมินมีแนวโน้มจะเปลี่ยนไปสู่การให้คะแนนในระดับที่สูงขึ้น γ_2 และ γ_3 มีตำแหน่งใกล้เคียงกันแสดงว่าผู้ประเมินมีความน่าจะเป็นที่จะเปลี่ยนระดับการประเมินระหว่างคะแนนในสองระดับ คือ ระหว่างระดับ “เข้าใจ” ไปสู่ระดับ “ประยุกต์ใช้” (γ_2) และระดับ “ประยุกต์ใช้” ไปสู่ระดับ “วิเคราะห์” (γ_3) ในการประเมินข้อที่ 4, 5, 6, 8, 16, 21, 22, 27, 29, 30, 33, 35 และ 36

ผลการประมาณค่าพารามิเตอร์ความยากของรายการประเมินความระดับความซับซ้อนทางปัญญาของข้อสอบ จำนวน 40 ข้อ พบว่า รายการประเมินที่มีค่าความยากต่ำสุด ได้แก่ ข้อ 30 ส่วนรายการประเมินที่มีค่าสูงสุด ได้แก่ข้อ 14

ผลการประมาณค่าพารามิเตอร์ a_i และ b_i ของผู้ประเมินจำนวน 20 คน มีค่า scaling bias ระหว่าง 0.800 ถึง 1.400 และมีค่า shifting bias ระหว่าง -0.210 ถึง 0.186 ค่าเฉลี่ยภายหลังของการประมาณค่าแสดงว่าผู้ประเมินส่วนใหญ่มีความไวในการเปลี่ยนระดับการให้คะแนนการประเมินในการประเมินระดับความซับซ้อนทางปัญญาของข้อสอบที่ใกล้เคียงกัน และไม่แสดงรูปแบบการกดหรือปล่อยคะแนนที่เด่นชัด เมื่อพิจารณาร่วมกับผลการประเมินระดับความซับซ้อนทางปัญญาของข้อสอบ T_k จะเห็นว่าผลการประเมินอยู่ระหว่างตำแหน่งคะแนนประเมินระดับ 2 (เข้าใจ) ถึงระดับ 4 (ประยุกต์ใช้) จากระดับการประเมิน 6 ระดับ ซึ่งเป็นตำแหน่งที่อยู่ในช่วงกลางของมาตรการประเมิน

3.2. ผลการวิเคราะห์การทำหน้าที่ต่างกันของผู้ประเมินในการวิเคราะห์ความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียนกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น

การประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียน กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น เป็นการประเมินระดับความสอดคล้องระหว่างข้อสอบกับตัวชี้วัด โดยมีระดับความสอดคล้อง 5 ระดับ คือ 0 = ไม่สอดคล้อง 1 =ค่อนข้างไม่สอดคล้อง 2 = ไม่แน่ใจ 3 = ค่อนข้างสอดคล้อง 4 = สอดคล้องโดยตรง ผู้วิจัยนำข้อมูลการประเมินความสอดคล้องในแนวเดียวกันของข้อสอบจำนวน 20 ข้อ จากผู้ประเมินจำนวน 40 คน มาวิเคราะห์ด้วยโมเดล LTRM

ค่าเฉลี่ยภายหลังของการแจกแจงการประมาณค่าพารามิเตอร์ $\mathrm{T_k}$ มีค่าระหว่าง 1.540 ถึง 6.104 มีค่าเฉลี่ยของค่าเฉลี่ยการแจกแจงภายหลัง $(\mu_{\mathrm{T_k}})$ เท่ากับ 4.126 และค่าเฉลี่ยความถูกต้องแม่นยำของการประมาณค่า $(\tau_{\mathrm{T_k}})$ เท่ากับ 0.703 เมื่อพิจารณาเทียบกับพารามิเตอร์ $\mathrm{T_k}$ แล้ว ตำแหน่งคะแนนการประเมินจะอยู่ในเทรชโฮลด์ที่ 4 หรือระดับการประเมินที่ 5 (สอดคล้องโดยตรง)

ค่าเฉลี่ยภายหลังการแจกแจงพารามิเตอร์ความยากของรายการประเมินความสอดคล้องในแนวเดียวกัน มีค่าเฉลี่ยการแจกแจงภายหลังของพารามิเตอร์ λ_{k} ระหว่าง -0.295 ถึง 0.476 รายการประเมินที่มีค่าเฉลี่ยภายหลังการแจกแจงพารามิเตอร์สูงสุด ได้แก่ข้อ 4 และต่ำสุดได้แก่ข้อ 28 รายการประเมินแต่ละรายการมีค่าเฉลี่ยการแจกแจงภายหลังของพารามิเตอร์ความยากของรายการประเมินใกล้เคียงกัน โดยมีค่าเฉลี่ยการแจกแจงภายหลังของความถูกต้องแม่นยำในการประมาณค่า $(\tau_{\lambda_{\mathrm{k}}})$ เท่ากับ 9.132

ผลการประมาณค่าความสามารถของผู้ประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานกับตัวชี้วัด (E_i) ของผู้ประเมินจำนวน 20 คน มีค่าระหว่าง 1.504 ถึง 2.856 มีค่าเฉลี่ยของค่าเฉลี่ยการแจกแจงภายหลังของพารามิเตอร์ $(\mu_{\mathrm{E_i}})$ เท่ากับ 0.564 และค่าเฉลี่ยของค่าเฉลี่ยความถูกต้องแม่นยำในการประมาณค่า $(\tau_{\mathrm{E_i}})$ เท่ากับ 24.266 จากผลการประมาณค่า สรุปได้ว่าผู้ประเมินคนที่ 6 มีค่าเฉลี่ยความสามารถในการประเมินสูงที่สุดในกลุ่ม ในขณะเดียวกันก็มีความแปรปวนในการให้คะแนนสูงกว่าผู้ประเมินคนอื่น

ผลการประมาณค่าพารามิเตอร์ความลำเอียงในการประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานกับตัวชี้วัด $(\mathrm{a_i},\ \mathrm{b_i})$ จากรายการประเมินจำนวน 40 ข้อ ของผู้ประเมินจำนวน 20 คน มีค่า scaling bias $(\mathrm{a_i})$ ระหว่าง 0.552 ถึง 2.592 และมีค่า shifting bias (b_i) ระหว่าง -1.290 ถึง 0.926 แสดงให้เห็นการกระจายของรูปแบบการประเมินของผู้ประเมิน จากผลการประมาณค่าแสดงว่า ผู้ประเมินคนที่ 15 มีแนวโน้มการเปลี่ยนระดับคะแนนไปในทางลบซึ่งแสดงถึงการกดคะแนนในการประเมิน ผู้ประเมินคนที่ 18 มีความไวในการเปลี่ยนระดับการให้คะแนนต่ำกว่าผู้ประเมินคนอื่น

และมีแนวโน้มที่จะเปลี่ยนระดับการประเมินไปทางบวก ผู้ประเมินคนที่ 4 มีแนวโน้มที่จะเปลี่ยนระดับการประเมินไปทางลบซึ่งแสดงถึงการกดคะแนน

ผลการวิเคราะห์ข้อมูลจริงสรุปได้ว่า ผลการวิเคราะห์ความสอดคล้องในแนวเดียวกันด้วยโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมในการประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียน กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น เป็นดังนี้

1) ไม่พบการทำหน้าที่ต่างกันของผู้ประเมินทั้ง 20 คน

2) คำตอบฉันทามติของผลการประเมินจากแบบบันทึกระดับความซับซ้อนทางปัญญาของข้อสอบในการประเมินระดับชาติ กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับชั้นมัธยมศึกษาปีที่ 3 จากผู้ประเมิน 20 คน สรุปได้ดังตาราง 5.1

ตาราง 5.1 ผลการประเมินจากแบบบันทึกระดับความซับซ้อนทางปัญญาของข้อสอบในการประเมินระดับชาติ กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับชั้นมัธยมศึกษาปีที่ 3

ระดับความซับซ้อนทางปัญญา	ข้อสอบ (ข้อที่)
จำ	3, 10, 12, 15, 17, 23, 28, 39
เข้าใจ	1, 2, 4, 7, 9, 11, 13, 14, 18, 19, 20, 24, 25, 26, 29, 31, 32, 34, 37, 38, 40
ประยุกต์ใช้	5, 6, 8, 21, 22, 27, 30, 33, 35, 36
วิเคราะห์	16
ประเมินค่า	-
สร้างสรรค์	-

3) คำตอบฉันทามติของผลการวิเคราะห์ผลการประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียน กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับมัธยมศึกษาตอนต้น จากผู้ประเมิน 20 คน พบว่า <u>ผู้ประเมินมีความเห็นว่าข้อสอบทุกข้อมีความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัด</u>

อภิปรายผลการวิจัย

จากการศึกษาเรื่อง การวิเคราะห์ฉันทามติของผู้ประเมินและการทำหน้าที่ต่างกันของผู้ประเมินในการวิเคราะห์ความสอดคล้องในแนวเดียวกัน: การประยุกต์ใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม ครั้งนี้ ผู้วิจัยพบประเด็นที่น่าสนใจเกี่ยวกับจุดเด่น ข้อจำกัด และสารสนเทศที่ได้รับจากการศึกษาครั้งนี้ ดังต่อไปนี้

1. การศึกษาครั้งนี้กำหนดขอบเขตของการจำลองข้อมูลโดยอ้างอิงจากการศึกษาครั้งก่อนหน้าซึ่งทำการศึกษาเกี่ยวกับประสิทธิภาพในการประมาณค่าของโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม ซึ่งส่วนมากจะเป็นการศึกษาซึ่งกระทำโดยผู้พัฒนาโมเดล รวมถึงผู้ที่นำโมเดลนี้ไปประยุกต์ใช้ในการศึกษาวิจัยทางมานุษยวิทยา และทางสังคมศาสตร์ ทำให้ขอบเขตของตัวอย่างขั้นต่ำ ($N = 25$) นั้นยังมีจำนวนสูงเมื่อเทียบกับจำนวนตัวอย่างที่ใช้ในการวิเคราะห์ข้อมูลจริง อย่างไรก็ตาม เมื่อพิจารณาจากการศึกษาของ Romney และคณะ (1986) ซึ่งได้สรุปความแม่นยำของการประมาณค่าพารามิเตอร์ในขนาดตัวอย่างขนาดต่าง ๆ ไว้ว่า หากตัวอย่างมีความเชี่ยวชาญในระดับสูง (มีความน่าจะเป็นในการประเมินได้สอดคล้องกับผลการประเมินของกลุ่มในระดับสูง) จะใช้จำนวนตัวอย่างอย่างน้อยที่สุด 4 คน ($N = 4$) ต่อหนึ่งกลุ่มวัฒนธรรมก็สามารถให้ผลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมที่น่าเชื่อถือได้ที่ระดับนัยสำคัญทางสถิติที่ .99

2. จากการศึกษาโดยการจำลองข้อมูล พบว่า เมื่อกำหนดให้มีการเลือกตัวอย่างแบบสุ่มในกระบวนการลูกโซ่มาร์คอฟ ทำให้บางครั้งโมเดลไม่สามารถประมาณค่าข้อมูลบางชุดได้ อันเนื่องมาจากการสุ่มผลการตอบของผู้ประเมินเข้าสู่เมตริกซ์ X_{ik} มีบางครั้งที่การสุ่มผลการตอบระหว่างผู้ประเมิน i กับ j มีคำตอบเหมือนกันส่งผลต่อการคำนวณค่าความแปรปรวนของ correlation matrix ทำให้ไม่สามารถคำนวณค่าไอเกนในการวิเคราะห์องค์ประกอบในการระบุจำนวนกลุ่มผู้ประเมินได้ ในกรณีดังกล่าว หากเป็นการวิเคราะห์ข้อมูลจริงนักวิจัยสามารถข้ามขั้นตอนการระบุจำนวนองค์ประกอบแล้วใช้การเปรียบเทียบค่า DIC ระหว่างโมเดลวิเคราะห์กลุ่มวัฒนธรรม 1 กลุ่ม กับโมเดลที่วิเคราะห์ข้อมูลที่มีหลายกลุ่มวัฒนธรรมได้ โดยหากค่า DIC ของโมเดลใดมีค่าน้อยกว่า นักวิจัยก็สามารถเลือกใช้โมเดลนั้นในการวิเคราะห์ อย่างไรก็ตาม ในการจำลองข้อมูลหากมีข้อมูลชุดใดที่มีความแปรปรวนเป็น 0 โมเดลจะข้ามการวิเคราะห์ข้อมูลชุดนั้นและจะเลือกข้อมูลชุดใหม่ขึ้นมาแทน ทำให้เกิดค่าสูญหายในการวิเคราะห์ซึ่งทำให้ผู้วิจัยต้องนำเข้าข้อมูลมากขึ้นเพื่อให้ได้จำนวนหน่วยตัวอย่างที่ต้องการ

3. ผลการวิเคราะห์ข้อมูลการประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบฯ พบว่า ผู้ประเมินคนที่ 4 มีแนวโน้มที่จะกดคะแนน ($a_i = 2.592, b_i = -0.332$) สอดคล้องกับผลการศึกษาของบุษยารัตน์ จันทร์ประเสริฐ (2560) นอกจากนี้ ผลการศึกษาในครั้งนี้ยังพบว่า ผู้ประเมินคนดังกล่าวมีความไวในการเปลี่ยนระดับการให้คะแนนการประเมินต่ำโดยพิจารณาจากค่าพารามิเตอร์ a_i ซึ่งแสดงช่วงห่างของเทรชโฮลด์ของผู้ประเมิน หากระยะห่างของเทรชโฮลด์กว้าง แสดงว่าผู้ประเมินคนดังกล่าวมีความไวในการเปลี่ยนระดับคะแนนการประเมินต่ำ ในบริบทของ

การประเมิน หากข้อคำถามคลุมเครือหรือไม่แสดงถึงความสอดคล้องในแนวเดียวกันระหว่างตัวชี้วัดกับข้อสอบอย่างชัดเจน ผู้ประเมินคนที่ 4 มีแนวโน้มที่จะประเมินในระดับคะแนนที่เป็นทางลบ เช่น “ค่อนข้างไม่สอดคล้อง” หรือ “ไม่สอดคล้อง” ในขณะที่ผู้ประเมินคนที่ 13 มีแนวโน้มที่จะประเมินในทางบวกตั้งแต่ระดับ “ไม่แน่ใจ” ขึ้นไป $(a_i = 1.391, b_i = 0.250)$ นอกจากนี้ จากการศึกษาของบุษยารัตน์ จันทร์ประเสริฐ พบว่าไม่สามารถจัดผู้ประเมินคนดังกล่าวอยู่ในกลุ่มรูปแบบของผู้ประเมินในการศึกษาก่อนหน้า กล่าวคือ ความแม่นยำในการให้คะแนน อิทธิพลแนวโน้มการให้คะแนนค่ากลาง อิทธิพลการจำกัดช่วง และอิทธิพลของความไม่มีแบบแผน ทั้งนี้ จากผลการวิเคราะห์ด้วยโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมผู้ประเมินคนที่ 18 เป็นผู้ประเมินที่มีค่าเฉลี่ยความสามารถในการประเมินสูง แต่ก็มีค่าเฉลี่ยความแปรปรวนในการประเมินสูงด้วยเช่นกัน ผู้ประเมินคนดังกล่าวมีความไวในการเปลี่ยนระดับคะแนนในการประเมินต่ำเป็นลำดับสามรองจากผู้ประเมินคนที่ 4 และ 2 และมีแนวโน้มที่จะเปลี่ยนระดับการให้คะแนนไปในทางบวก ซึ่งเป็นการแปลความหมายของพารามิเตอร์ที่แสดงให้เห็นรูปแบบของการประเมินของผู้ประเมินที่ต่างไปจากการแปลความหมายของการวิเคราะห์ด้วยสถิติแบบดั้งเดิม

4. ถึงแม้ว่าโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมจะมีลักษณะที่คล้ายคลึงกับโมเดลการตอบสนองข้อสอบ (Item Response Theory: IRT) หรือราส์ชโมเดลในแง่ของการประมาณค่าความสามารถแฝงของผู้ตอบ รวมถึงมีการประมาณค่าพารามิเตอร์ความยากของข้อคำถาม แต่โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมเป็นโมเดลที่พัฒนาขึ้นโดยมีวัตถุประสงค์เพื่อวิเคราะห์ความสอดคล้องของรูปแบบการตอบของผู้ให้ข้อมูล/ผู้ประเมิน ในกรณีที่รายการคำถามนั้นอยู่ในรูปของแบบสอบถาม หรือคำถามที่ไม่มีคำตอบหรือเฉลยเอาไว้ล่วงหน้า ดังเช่น การหาข้อสรุปของคำตอบให้แก่คำถามในการศึกษาวิจัยทางสังคมศาสตร์ หรือทางมานุษยวิทยา โดยประมาณค่าความน่าจะเป็นที่ผู้ให้ข้อมูลจะให้ข้อมูลสอดคล้องกับสมาชิกคนอื่นของกลุ่มภายใต้ปัจจัยความคลาดเคลื่อนต่าง ๆ เช่น ความยากของข้อคำถาม ความลำเอียงในการเดาคำตอบ และความรู้เดิม / ภูมิหลัง หรือคติของผู้ให้ข้อมูลที่มีต่อข้อคำถาม ดังที่ Anders และ Batchelder (2015) ได้ให้ข้อสังเกตในตอนท้ายของผลการศึกษา นอกจากนี้ โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมยังสามารถนำมาประยุกต์ใช้กับการประเมินที่มีผู้ประเมินหลายคนเพื่อหาผลการประเมินฉันทามติของการประเมินได้

ข้อเสนอแนะ

<u>ข้อเสนอแนะในการนำผลการวิจัยไปใช้</u>

1. นักวิจัย นักวัดและประเมินผลสามารถนำขั้นตอนในการวิเคราะห์ข้อมูลไปประยุกต์ใช้กับการศึกษาความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินของกลุ่มสาระการเรียนรู้อื่น หรือในบริบทของการประเมินอื่นที่มีการวิเคราะห์ความสอดคล้องหรือความเที่ยงระหว่างผู้ประเมิน โดยการวิเคราะห์ข้อมูลจะให้สารสนเทศเกี่ยวกับคำถามการประเมินและสารสนเทศเกี่ยวกับความสามารถและความลำเอียงของผู้ประเมินซึ่งสถานศึกษาและหน่วยงานที่เกี่ยวข้องกับการทดสอบทางการศึกษาสามารถนำไปใช้ประกอบการพิจารณาความน่าเชื่อถือของผู้เชี่ยวชาญในการคัดเลือกมาประเมินความสอดคล้องในแนวเดียวกันทางการศึกษา

2. การประเมินประสิทธิภาพของการวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินจากการจำลองข้อมูลแสดงให้เห็นว่าการทำหน้าที่ต่างกันระหว่างผู้ประเมินส่งผลต่อประสิทธิภาพในการประมาณค่าของโมเดล ดังนั้น หากนักวิจัยต้องการใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมในการศึกษาความสอดคล้องในแนวเดียวกันทางการศึกษา ควรตรวจสอบการทำหน้าที่ต่างกันระหว่างผู้ประเมินและเลือกใช้โมเดลการวิเคราะห์ที่เหมาะสมกับข้อมูล รวมถึงควรเพิ่มจำนวนตัวอย่าง จำนวนลูกโซ่ โดยการเข้าไปกำหนดค่าเพิ่มเติมในการวิเคราะห์ข้อมูล เพื่อทำให้โมเดลสามารถประมาณค่าได้ถูกต้องและแม่นยำมากขึ้น

3. การคัดเลือกผู้เชี่ยวชาญในการประเมินความสอดคล้องในแนวเดียวกันทางการศึกษานอกจากต้องคำนึงถึงความรู้ความสามารถของผู้เชี่ยวชาญแล้ว ยังควรคำนึงถึงภูมิหลังของผู้ประเมิน อาทิ รูปแบบการกดหรือปล่อยคะแนน ความไวในการเปลี่ยนระดับคะแนนเมื่อมีการเปลี่ยนแปลงรายละเอียดของคำถามประเมิน รวมถึงอคติของผู้ประเมิน ซึ่งสิ่งเหล่านี้อาจเป็นปัจจัยที่ส่งผลต่อคุณภาพของการประเมินนอกเหนือไปจากความเชี่ยวชาญในบริบทของการประเมิน

<u>ข้อเสนอแนะในการทำวิจัยครั้งต่อไป</u>

1. การศึกษาครั้งนี้เป็นการศึกษาโดยใช้ข้อมูลทุติยภูมิ ซึ่งผู้วิจัยไม่ได้ทำการจัดบริบทของการสอดคล้องในแนวเดียวกันด้วยตนเอง จึงอาจมีความคลาดเคลื่อนของผลการวิเคราะห์ซึ่งเกิดจากปัจจัยแฝงอื่น ดังนั้นจึงควรมีการศึกษาการใช้โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมในบริบทของการประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินในบริบทจริง เพื่อให้ได้ผลการศึกษาที่ถูกต้องและแม่นยำมากขึ้น

2. ควรมีการศึกษาเกี่ยวกับประสิทธิภาพของการประมาณค่าพารามิเตอร์อื่น ๆ ของโมเดลซึ่งอาจส่งผลต่อการประมาณค่า หรือศึกษาแนวโน้มของประสิทธิภาพการประมาณค่าเมื่อกำหนดการแจกแจงของพารามิเตอร์ในรูปแบบอื่นนอกเหนือจากรูปแบบที่เป็นค่าตั้งต้นของโมเดล

3. การศึกษาครั้งนี้ไม่พบการทำหน้าที่ต่างกันระหว่างผู้ประเมินในการศึกษาข้อมูลจริง เนื่องจากเป็นข้อมูลที่มาจากผู้ประเมินที่ผ่านการคัดเลือกและได้รับการอบรมเกี่ยวกับการประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบมาแล้ว จึงเป็นผู้ประเมินที่มีความน่าเชื่อถือและมีความเที่ยงสูง อย่างไรก็ตาม หากมีการศึกษากับข้อมูลที่มีการทำหน้าที่ต่างกันระหว่างผู้ประเมินก็จะทำให้เห็นรูปแบบการประเมินที่แตกต่างออกไป แล้วอาจจะได้สารสนเทศเกี่ยวกับคำถามประเมินและผู้ประเมินที่แตกต่างกันมากขึ้น

4. ควรมีการศึกษาความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียนของกลุ่มสาระการเรียนรู้อื่น ๆ

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

บรรณานุกรม

ภาษาไทย

บุษยารัตน์ จันทร์ประเสริฐ. (2561). *ความสอดคล้องในแนวเดียวกันระหว่างข้อสอบในการประเมินระดับชาติกับข้อสอบในการประเมินระดับชั้นเรียน กลุ่มสาระการเรียนรู้วิทยาศาสตร์: การประยุกต์ใช้โมเดลหลายองค์ประกอบของราสช์และทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัด*. (ปริญญาดุษฎีบัณฑิต). จุฬาลงกรณ์มหาวิทยาลัย

ประสพชัย พสุนนท์. (2558). การประเมินความเชื่อมั่นระหว่างผู้ประเมินโดยใช้สัมประสิทธิ์แคปปา. *วารสารวิชาการศิลปศาสตร์ประยุกต์*, 8(1), 2-20.

ศิริชัย กาญจนวาสี. (2555). *ทฤษฎีการทดสอบแนวใหม่*. กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

ศิริชัย กาญจนวาสี. (2556). *ทฤษฎีการทดสอบแบบดั้งเดิม*. กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

ภาษาอังกฤษ

Al-Bayatti, M. and Jones, B. (2005). *NAA Enhancing the Quality of Marking Project: the Effect of Sample Size on Increased Precision in Detecting Errant Marking*. London: QCA. Retrieved from http://dera.ioe.ac.uk/9451/1/ The_effect_ of_sample_size_on_increased_precision_in _detecting_errant_marking.pdf

Andres, A. M., Marzo, P. F. (2004). Delta: A new measure of agreement between two raters. British Journal of Mathematical and Statistiscal Psychology, 57, 1-19. Retrieved from https://onlinelibrary.wiley.com/doi/epdf/ 10.1348/000711004849268

Anders R., Oravecz Z., Batchelder W. H. (2014). Cultural Consensus Theory for Continuous Responses: A Latent Appraisal Model for Information Pooling. *Journal of Mathematical Psychology,* 61, 1-13.

Anders, R., & Batchelder, W. H. (2015). Cultural consensus theory for the ordinal data case. *Psychometrika*, 80(1), 151-181.

Anders, R. (2017). Package ‘CCTpack’. Consensus Analysis, Model-Based Clustering, and Cultural Consensus Theory Applications. R package version 1.5.2.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Baker, E., Ayres, P., O'Neil, H.F., Chli, K., Sawyer, W., Sylvester, R.M. and Carroll, B. (2008). *KS3 English Test Marker Study in Australia: Final Report to the National Assessment Agency of England*. Sherman Oaks, CA: University of Southern California.

Baird, J. (1988). What's in a name? Experiments with blind marking in A-level examinations. *Educational Research*. 40(2), 191-202.

Baird, J. & Mac, Q. (1999) How should examiner adjustments be calculated? - A discussion paper. *AEB Research Report*, RC13.

Baird, J.A., Hayes, M., Johnson, R., Johnson, S. and Lamprianou, L. (2012). *Marker Effects and Examination Reliability: a Comparative Exploration from the Perspectives of Generalizability Theory, Rasch Modelling and Multilevel Modelling*. Coventry: Ofqual.

Batchelder W. H., Strashny A., Romney A. K. (2010). Cultural consensus theory: Aggregating continuous responses in a finite interval. In Chai S.-K., Salerno J. J., Mabrey P. L. (Eds.), *Social computing, behavioral modeling, and prediction 2010* (pp. 98-107). New York, NY: Springer-Verlag.

Batchelder W. H., Anders R. (2012). Cultural Consensus Theory: Comparing Different Concepts of Cultural Truth. *Journal of Mathematical Psychology*, 56, 316-332.

Batchelder, W.H., Anders, R., and Oravecvz, Z. In Wixted, J. T. & Wagenmakers, E. J. (Eds., in press). *Stevens' handbook of experimental psychology and cognitive neuroscience (4th ed.)*. Volume 4: Methodology. New York: Wiley.

Bernardin, H. J., Alvares, K. M., & Cranny, C. J. (1976). A recomparison of behavioral expectation scales to summated scales. *Journal of Applied Psychology*, 61(5), 564-570.

Bland J.M., Altman D.G. (1999). Measuring agreement in method comparison studies. *Stat Methods Med Res*, 8, 135–160.

Borgatti, S. P., & Halgin, D. S. (2011). Consensus analysis. *A companion to cognitive Anthropology*, 171-190.

Bramley, T. (2007). Quantifying marker agreement: terminology, statistics and issues. Research Matters, 4, 22–27.

Bramley, T. and Dhawan, V. (2010). Estimates of Reliability of Qualifications. *Ofqual*

Reliability Compendium (Chapter 7). Coventry: Ofqual. Retrieved from http://webarchive.nationalarchives.gov.uk/20140813100205/http://ofqual.gov.uk/standards/research/reliability/compendium/

Brown, T. C., and Daniel, T. C. (1990). *Scaling of Ratings: Concepts and Methods*. Research paper. United States Department of Agriculture. Retrieved from https://www.fs.fed.us/ rm/pubs_rm/rm_rp293.pdf

Byrt, T., Bishop, J., Carlin, J. B. (1993). Bias, Prevalence and Kappa. *J Clin Epdemiol*, 46, 423-429.Cohen, J. (1960). A coefficient for agreement for nominal scales. *Educational and Psychological Measurement*, Vol. 20, 37-46.

Conger, A. J. (2016). Kappa and Rater Acuuracy: Paradigms and Parameters. *Educational and Psychological Measurement*, 77(6), 1-29.

Costa-Santos et al. (2009). The Limits of Agreement and the Intraclass Correlation Coefficient may be Inconsistent in the Interpretation of Agreement. *Journal of Clinical Epidemiology*, 64(2011), 264-269.

Dantas, Clarissa de Rosalmeida, & Banzato, Cláudio Eduardo Muller. (2007). Inter-rater reliability and factor analysis of the Brazilian version of the Schedule for the Assessment of Insight: Expanded Version (SAI-E). *Revista Brasileira de Psiquiatria*, 29(4), 359-362. Epub May 11, 2007 .https://dx.doi.org/10.1590/S1516-44462006005000041

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education*, 4, 289–303.

Engelhard, G. (1994). Examining Rater Errors in the Assessment of Written Composition with a Many-Faceted Rasch Model. *Journal of Educational Measurement*, 31(2), 93-112. Retrieved from http://www.jstor.org/stable/1435170

Engelhard, G. (2013). *Invariant Measurement: Using Rasch models in the social, behavioral, and health science*. New York: Routledge.

Engelhard, G. Jr., Wind, S. A., Jennifer, L. K., Chajewski, M. (2013). *Differential Item and Person Functioning in Large-Scale Writing Assessments within the Context of the SAT*. Research report. College Board.

Farrokhi et al. (2012). A Many-Facet Rasch Measurement of Differential Rater

Severity/Leniency in Three Types of Assessment. *JALT Journal*, Vol. 34(1), 79-102.

Fleiss, J. L., Levin, B., Paik, M. C., (2003). *Statistical Methods for Raters and Proportions*. 3rd Edition. New Jersey: Wiley.

Gelman et al. (2014). *Bayesian Data Analysis*. 3rd Edition. FL: Taylor & Francis Group.

Giavarina, D. (2015). Understanding Bland Altman analysis. *Biochemia Medica*, 25(2), 141–151. http://doi.org/10.11613/BM.2015.015

Gisev et al. (2013). Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, 9(2013), 330-338.

Gorman, B. S. (1976). Principal Components Analysis as an Alternative to Kendall's Coefficient of Concordance, W. Education and Psychological Measurement, 36(1976), 627-629

Gunilla Näsström. (2008). *Measurement of Alignment between Standards and Assessment*. Department of Educational Measurement, Umeå University

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Hayes, J. R., & Hatch, J. A. (1999). Issues in measuring reliability: Correlation versus percentage of agreement. *Written Communication*, 16(3), 354-367.

Hogarth, R. M., and Karelaia, N. (2007). Heuristic and Linear Models of Judgment: Matching Rules and Environment. *Psychological Review*, 114(3), 733-758. Retrieved from https://www.researchgate.net/profile/ Robin_Hogarth/ publication/6199093_Heuristic_and_Linear_Models_of_Judgment_Matching_Rules_and_Environments/links/544e2c9f0cf29473161a1dbf/Heuristic-and-Linear-Models-of-Judgment-Matching-Rules-and-Environments.pdf

Hruschka, D. J., Maupin, J. N. (2012). Competence, Agreement, and, Luck: Testing Whether Some People Agree More with a Cultural Truth then Do Others. *Field Methods*, 25(2), 107-123.

James, C. (1974). The consistency of marking a physics examination. *Physics Education*, Vol. 9, 271-274.

Jin, Kuan-Yu and Wang, Wen-Chung. (2017). Assessment of Differential Rater

Functioning in Latent Classes with New Mixture Facets Models. *Multivariate Behavioral Research*, 52(3), 391-402. Retrieved from http://dx.doi.org/10.1080/00273171.2017.1299615

Johanson, G. A., and Osborn, C. J. (2004). Acquiescence as Differential Person Functioning. *Assessment & Evaluation in Higher Education*, 29(5), 535-548.

Kaliski et al. (2012). Using the Many-Faceted Rasch Model to Evaluate Standard Setting Judgments: An Illustration with the Advance Placement Environmental Science Exam. *Educational and Psychological Measurement*, 73(3), 386-411.

Kean, J., & Reilly, J. (2014). Item response theory. *Handbook for Clinical Research: Design, Statistics and Implementation*. pp. 195-198. New York: Demos Medical Publishing.

Kottner et al. (2011). Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *International Journal of Nursing Studies*, 48(2011), 661-671.

Krippendorff, K. (1980). *Content Analysis: An Introduction to Its Methodology*. CA: Sage Publications, Inc.

Krippendorff, K. (2012). *Content Analysis: An Introduction to Its Methodology*. 3rd Edition. CA : Sage Publications Inc.

Kruschke, J. k. (2011). *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. New York: Academic Press.

La Marca, P. M. (2001). Alignment of standards and assessments as an accountability criterion. *Practical Assessment, Research & Evaluation*, 7(21).

Landis, J. R., Koch, G. G. (1977). The Measures of observer agreement for categorical data. *Biometrics*, 33(1), 159-174.

LeBreton JM, Senter JL. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organ Res Methods*, 11, 815–852.

Leckie, G., and Baird, J. (2011). Rater Effects on Essay Scoring: A Multilevel Analysis of Severity Drift, Central Tendency, and Rater Experience. *Journal of Educational Measurement*, 48(4), 399-418.

Linacre, J. M., Wright, B.D. & Lunz, M.E. (1990) A Facets Model for judgmental scoring. [Online] retrieved from https://www.rasch.org/memo61.htm

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Linacre, J. M. (1994). *Many-facet Rasch Measurement*. Chicago: MESA Press.

Linacre, J. M. (2002).Judge ratings with forced agreement. *Rasch Measurement Transactions*, 16(1), 857-858.

Malouff, J. M., Emmerton, A. J., & Schutte, N. S. (2013). The risk of a halo bias as a reason to keep students anonymous during grading. *Teaching of Psychology*, 40, 233–237.

MedCalc Software. 2018. *Bland-Altmand Plot*. [Online] Available: https://www.medcalc.org/manual/blandaltman.php

McGraw, O. K. and Wong, S. P. (1996). Forming Inferences about Some Intraclass Correlation Coefficients. *Psychological Methods*, 1(1), 30-46.

McHugh, M. L. (2012). Interrater Reliability: The Kappa Statistic. *Biochemia Medica*, 22(3), 276-282.

McNamara, T. F.(1996). *Measuring Second Language Performance*. New York: Longman.

Meadows, M. and Billington, L. (2005). *A Review of the Literature on Marking Reliability. Manchester*: AQA [Online]. Available: https://orderline.education.gov.uk/gempdf/1849625344/QCDA104983_review_of_the_literature_on_marking_reliability.pdf

Muckle, T. J., Karabatsos, G. (2009). Hierarchical Generalized Linear Models for the Analysis of Judge Ratings. *Journal of Education Measurement*, 46(2), 198-219.

Muller, R., Buttner, P. (1994). A Critical Discussion of Intraclass Correlation Coefficients. *Stat Med*, 13, 2465-76.

Myford, C., Marr, D.B., and Linacre, J.M. (1996) Reader calibration and its potential role for equating for the test of written English. Research Report No. 52 Princeton, NJ: Educational Testing Service.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, 4(4), 386–422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, 5(2), 189–227.

Myford, C. M., & Wolfe, E. W. (2009). Monitoring Rater Performance over Time: A Framework for Detecting Differential Accuracy and Differential Scale Category Use. *Journal of Educational Measurement*. 46(4), 371-389.

Myles, P.S., Cui, J. (2007). I. Using the Bland–Altman method to measure agreement with repeated measures. *British Journal of Anaesthesia*, 99(3), 309-311. https://doi.org/10.1093/bja/aem214.

Newstead, S. E. & Dennis, I. (1990). Blind marking and sex bias in student assessment. *Assessment and Evaluation in Higher Education*, 15(12), 132-139.

Newton, P.E. (2009). The reliability of results from national curriculum testing in England, testing in England, *Educational Research*, 51(2), 181–212.

Oravecz, Z., Faust, K., and Batchelder, W. H. (2014). An extended Cultural Consensus Theory model to account for cognitive processes in decision making in social surveys. *Sociological Methodology*, 26, 185-228.

Oravecz, Z., Vandekerckhove, J. & Batchelder, W. (2014). Bayesian Cultural Consensus Theory. *Field Methods*, 26, 207-222.

Patz, R. J., Junker, B. W., Johnson, M. S., and Mariano, L. T. (2002). The Hierarchical Rater Model for Rated Test Items and its Application to Large-Scale Educational Assessment Data. *Journal of Educational and Behavioral Statistics*, 27(4), 341-384.

Peterson, J. H. (2013). Analysis of Rater Agreement by Rasch and IRT Models. In Karl el al. (Eds.), *Rasch Models in Health*. New Jersey: John Wiley & Sons, Inc.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice.
Educational Researcher, 31(7), 3-14.

Porter, A. C., & Smithson, J. L. (2001). Are content standards being implemented in the classroom? A methodology and some tentative answers. In S. H. Fuhrman (Ed.), *From the Capitol to the classroom*. *Standards-based reform in the States* (pp. 60-80). Chicago: National Society for the Study of Education, University of Chicago press.

Porter, A. C., Smithson, J., Blank, R., & Zeidner, T. (2007). Alignment as a teacher variable. *Applied Measurement in Education*, 20(1), 27-51.

Roach, A. T., Niebling, B. C., & Kurz, A. (2008). Evaluating the alignment among curriculum, instruction, and assessments: Implications and applications for research and practice. *Psychology in the Schools*, 45(2), 158-176.

Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American anthropologist*, 88(2), 313-338.

Rothman, R. (2003). Imperfect matches: The alignment of standards and tests. *Paper commissioned by the Committee on Test Design for K-12 Science Achievement*, March 2003.

Saal, F. E., Downey, R. G., and Lahey, M. A. (1980). Rating the ratings: Assessing the Psychometric Quality of Rating Data. *Psychological Bulletin*, 88(2), 413-428.

Sahin, A., and Anil, D. (2016). The Effects of Test Length and Sample Size on Item Parameters in Item Response Theory. *Educational Science: Theory and Practice*. 17(1), 321-335.

Salzberger, Thomas. (2010). Does the Rasch Model Convert an Ordinal Scale into an Interval Scale? *Rasch Measurement Transactions*, 24:2, 1273-1228. Retrieved from https://www.rasch.org/rmt/rmt242.pdf

Schaefer, E. (2008). Rater Bias Patterns in an EFL Writing Assessment. *Language Testing*, 28(4), 465-493.

Sedwick, P. (2013). Limits of Agreement (Bland-Altman method). *BMJ,* 2013; 346:f1630 doi: 10.1136/bmj.f1630

Shavelson, R. J., Webb, N. M., and Rowley, G. (1989). *Generalizability Theory*. American Psychologist, 44, 922-932. DOI: 10.1037/0003-066X.44.6.922.

Shavelson, R. J., and Webb, N. M. (1991). Generalizability Theory: A Primer. CA: SAGE Publications.

Soeken K. L., Prescott P. A. (1986). Issues in the use of kappa to estimate reliability. *Med Care*, 24, 733–741.

Stephens, J-P, Vos, G. A., Stevens, E. M., & Moore, J. S. (2006). Test-retest repeatability of the Strain index. *Applied Ergonomics*, 37(3), 275-281.

Stephen L. France and William H. Batchelder. (2015). Maximum Likelihood Item Easiness Models for Test Theory without an Answer Key. *Educational and Psychological Measurement*_2015, 75(1), 57–77.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Stemler, S. E., & Bebell, D. (1999, April). *An empirical approach to understanding and analyzing the mission statements of selected educational institutions*. Paper presented at the New England Educational Research Organization (NEERO), Portsmouth, NH.

Stemler, S.E. (2004). A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability. *Practical Assessment Research & Evaluation*, Vol. 9, No. 4. Retrieved from http://pareonline.net/getvn.asp?v=9&n=4

Tisi, J., Whitehouse, G., Maughan, S., and Burdett, N. (2013). *A Review ofLiterature on Marking Reliability Research* (Report commissioned by the Office of Qualifications and Examinations Regulation). Slough, UK: NFER. Retrieved from https://www.nfer.ac.uk/publications/MARK01/MARK01.pdf

Thompson, C. A. Foster, A., Cole, I., & Dowding, D. W. (2005). Using Social Judgement Theory to Model Nurses' Use of Clinical Information in Critical Care Education. *Nurse Education Today*, 25(1), 68-77. Retrieved from https://www.sciencedirect.com/science/article/pii/S026069170400139X?via%3Dihub

Trotter, R. T2nd, Weller, S. C., Baer, R. D., Pachter, L. M., Glazer, M., Garcia de Alba Garcia, J. E., Klein, R. E. (1999). Consensus theory model of AIDS/SIDA beliefs in four Latino populations. *AIDS Educ Prev.*, Oct, 11(5), 414-26. Retrieved from http://jan.ucc.nau.edu/rtt/pdf%20format%20pubs/Trotter%201990s%20pdf%20Pubs/Consensus%20Theory%20Model%20of%20AIDS.pdf

Wang, J., Engelgard G. Jr., Wolfe, E. W. (2015). Evaluating Rater Accuracy in Rater-Mediated Assessments Using an Unfolding Model. *Educational and Psychological Measurement*, 76(6), 1005-1025.

Watkins, M. W., & Pacheco, M. (2000). Interobserver agreement in behavioral research: Importance and calculation. *Journal of Behavioral Education*, 10(4), 205-212.

Warrens, M. J. (2014). On Marginal Dependencies of the 2 × 2 Kappa. *Advances in Statistics*, vol. 2014, Article ID 759527, 6 pages. https://doi.org/10.1155/2014/759527

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education (Research monograph, No. 6)*. Madison: National Institute for Science Education.

Webb, N. L. (2002). An analysis of the alignment between mathematics standards and assessments for three states. Paper presented at the annual meeting of the American *Educational Research Association*, April 1-5, in New Orleans, USA.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education*, 20(1), 7-25.

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). 4 Reliability Coefficients and Generalizability Theory. *Handbook of statistics*, 26, 81-124. Retrieved from https://web.stanford.edu/dept/SUSE/SEAL/Reports_Papers/methods_papers/G%20Theory%20Hdbk%20of%20Statistics.pdf

Weller, S. C. (2007). Cultural Consensus Theory: Applications and Frequently Asked Questions. *Field Methods*, 19, 339-368.

Weller, S. C. (2007). Explanatory Models of Diabetes in the U.S. and Mexico: The Patient-provider gap and cultural competence. *Social Science & Medicine*, 75, 1088-1096.

Wesoloaski, B. C., Wind, S. A., and Engelhard, G. Jr. (2015). Rater Fairness in Music Performance Assessment: Evaluating Model-Data Fit and Differential Rater Functioning. *Musicae Scientiae*, 19920, 147-170.

Wind, S. A. (2015). Evaluating the quality of analytic ratings with Mokken scaling. *Psychological Test and Assessment Modeling*, 57(3), 423-444. Retrieved from https://pdfs.semanticscholar.org/ad9c/748e9f28137c2976801367be1d026f2a209b.pdf

Wind, S. A. (2017). An instructional module on Mokken scale analysis. *Educational Measurement: Issues and Practice*, 36(2), 50-66. Retrieved from https://onlinelibrary.wiley.com/doi/epdf/10.1111/emip.12153

Wind, S.A. and Engelhard G. Jr. (2015). Exploring Rating Quality in Rater-Mediated Assessments Using Mokken Scale Analysis. *Educational and Psychological Measurement*, Vol. 76(4), 685-706.

Wind, S. A., and Peterson, M. E. (2017). A Systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, 35(2), 1-32.

Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics*, 26, 283–306. Retrieved from https://www.jstor.org/stable/pdf/3648159.pdf?refreqid=excelsior%3Aa307e0b8e04d93f7eb325e07c5bc780f

Wolfe, E.W., Moulder, B. C., Myford, C. M. (1999). Detecting Differential Rater Functioning over Time (DRIFT) Using a Rasch Multi-Faceted Rating Scale Model. [Washington D.C.] : Distributed by ERIC Clearinghouse, http://www.eric.ed.gov/contentdelivery/servlet/ ERICServlet?accno=ED434115

Xun Yan. (2014). An Examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *Language Testing*, Vol. 31(4), 501-527.

Zaiontz, C. (n.d.). Reliability. *Real Statistics Using Excel*. [Online] Retrieved from http://www.real-statistics.com/reliability/

Zhao, X., Liu, J. S., & Deng, K. (2012). Assumptions behind intercoder reliability indices. In C. T. Salmon (eds.), *Communication Yearbook 36*, pp. 419 - 480. New York: Routledge. Retrieved from https://repository.hkbu.edu.hk/cgi/viewcontent.cgi?article=1000&context=coms_bkch.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

ภาคผนวก

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ผลการวิเคราะห์การประเมินจากแบบบันทึกระดับความซับซ้อนทางปัญญาของข้อสอบใน การประเมินระดับชาติ กลุ่มสาระการเรียนรู้วิทยาศาสตร์

	mean	sd	2.50%	25%	50%	75%	97.50%	Rhat	n.eff
E[1]	0.591	0.199	0.265	0.447	0.567	0.709	1.045	1.053	43
E[2]	0.717	0.230	0.333	0.550	0.691	0.855	1.233	1.053	42
E[3]	0.827	0.270	0.380	0.633	0.796	0.988	1.435	1.050	44
E[4]	0.669	0.226	0.302	0.507	0.640	0.802	1.183	1.052	44
E[5]	0.856	0.279	0.397	0.654	0.826	1.023	1.479	1.048	45
E[6]	0.738	0.240	0.342	0.565	0.710	0.882	1.275	1.051	43
E[7]	0.745	0.239	0.347	0.572	0.719	0.890	1.280	1.050	43
E[8]	0.845	0.271	0.395	0.651	0.819	1.008	1.446	1.055	40
E[9]	0.673	0.224	0.302	0.512	0.647	0.806	1.185	1.054	41
E[10]	0.900	0.295	0.415	0.687	0.869	1.079	1.557	1.047	46
E[11]	0.708	0.236	0.322	0.540	0.679	0.849	1.239	1.056	41
E[12]	0.591	0.203	0.261	0.445	0.566	0.711	1.057	1.053	44
E[13]	0.702	0.240	0.314	0.531	0.671	0.845	1.254	1.049	45
E[14]	0.623	0.208	0.284	0.474	0.598	0.747	1.097	1.054	44
E[15]	1.180	0.392	0.545	0.900	1.134	1.411	2.054	1.043	47
E[16]	0.778	0.253	0.360	0.595	0.750	0.927	1.352	1.047	46
E[17]	0.831	0.269	0.387	0.639	0.799	0.987	1.437	1.052	41
E[18]	0.622	0.209	0.280	0.474	0.596	0.746	1.097	1.051	44
E[19]	0.629	0.214	0.275	0.475	0.604	0.758	1.113	1.051	44
E[20]	0.628	0.214	0.282	0.475	0.602	0.753	1.110	1.053	43
Emu	-0.368	0.300	-1.057	-0.557	-0.351	-0.155	0.159	1.074	34
Etau	22.700	20.978	5.625	11.459	17.082	26.410	73.521	1.002	6000
Om[1]	1	0	1	1	1	1	1	1	1
Om[2]	1	0	1	1	1	1	1	1	1
Om[3]	1	0	1	1	1	1	1	1	1
Om[4]	1	0	1	1	1	1	1	1	1
Om[5]	1	0	1	1	1	1	1	1	1
Om[6]	1	0	1	1	1	1	1	1	1
Om[7]	1	0	1	1	1	1	1	1	1
Om[8]	1	0	1	1	1	1	1	1	1
Om[9]	1	0	1	1	1	1	1	1	1
Om[10]	1	0	1	1	1	1	1	1	1
Om[11]	1	0	1	1	1	1	1	1	1
Om[12]	1	0	1	1	1	1	1	1	1
Om[13]	1	0	1	1	1	1	1	1	1
Om[14]	1	0	1	1	1	1	1	1	1

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

	mean	sd	2.50%	25%	50%	75%	97.50%	Rhat	n.eff
Om[15]	1	0	1	1	1	1	1	1	1
Om[16]	1	0	1	1	1	1	1	1	1
Om[17]	1	0	1	1	1	1	1	1	1
Om[18]	1	0	1	1	1	1	1	1	1
Om[19]	1	0	1	1	1	1	1	1	1
Om[20]	1	0	1	1	1	1	1	1	1
T[1,1]	-0.909	0.509	-2.031	-1.221	-0.856	-0.547	-0.054	1.015	180
T[2,1]	-3.135	0.726	-4.562	-3.626	-3.145	-2.631	-1.766	1.066	37
T[3,1]	-5.215	1.146	-7.671	-5.903	-5.192	-4.479	-3.075	1.088	28
T[4,1]	-0.281	0.474	-1.333	-0.562	-0.240	0.033	0.551	1.002	4500
T[5,1]	0.754	0.297	0.222	0.551	0.740	0.943	1.371	1.023	97
T[6,1]	0.508	0.159	0.219	0.393	0.503	0.616	0.828	1.025	98
T[7,1]	-3.840	0.808	-5.408	-4.387	-3.876	-3.278	-2.261	1.084	29
T[8,1]	0.532	0.168	0.228	0.411	0.526	0.646	0.873	1.027	82
T[9,1]	-1.238	0.582	-2.502	-1.611	-1.187	-0.816	-0.255	1.020	110
T[10,1]	-5.411	1.272	-8.190	-6.163	-5.351	-4.564	-3.134	1.058	40
T[11,1]	-0.355	0.466	-1.387	-0.637	-0.317	-0.037	0.466	1.007	450
T[12,1]	-4.691	1.024	-6.826	-5.322	-4.696	-4.007	-2.769	1.083	30
T[13,1]	-2.466	0.748	-4.048	-2.941	-2.424	-1.937	-1.145	1.039	58
T[14,1]	-3.072	1.064	-5.365	-3.721	-2.984	-2.323	-1.258	1.025	89
T[15,1]	-5.409	1.180	-7.882	-6.135	-5.392	-4.637	-3.208	1.082	30
T[16,1]	1.220	0.342	0.610	0.980	1.209	1.441	1.932	1.047	49
T[17,1]	-4.994	1.115	-7.342	-5.695	-4.972	-4.237	-2.938	1.069	34
T[18,1]	-0.947	0.582	-2.259	-1.302	-0.876	-0.515	-0.048	1.026	96
T[19,1]	-0.824	0.490	-1.929	-1.116	-0.772	-0.474	-0.016	1.016	170
T[20,1]	-2.054	0.650	-3.430	-2.468	-2.012	-1.595	-0.923	1.034	65
T[21,1]	0.601	0.201	0.231	0.458	0.594	0.738	1.004	1.024	90
T[22,1]	0.467	0.171	0.150	0.347	0.462	0.582	0.814	1.017	130
T[23,1]	-6.583	1.586	-9.985	-7.571	-6.484	-5.490	-3.754	1.053	44
T[24,1]	-1.172	0.542	-2.354	-1.505	-1.125	-0.790	-0.245	1.017	130
T[25,1]	-0.243	0.464	-1.284	-0.517	-0.201	0.072	0.565	1.006	1000
T[26,1]	-3.365	0.698	-4.727	-3.835	-3.399	-2.889	-1.989	1.086	29
T[27,1]	1.046	0.347	0.417	0.806	1.034	1.268	1.761	1.024	91
T[28,1]	-4.843	1.153	-7.266	-5.554	-4.810	-4.048	-2.752	1.063	38
T[29,1]	-0.033	0.317	-0.772	-0.198	0.006	0.178	0.480	1.012	950
T[30,1]	0.472	0.155	0.188	0.359	0.466	0.578	0.786	1.023	97
T[31,1]	-2.908	0.695	-4.285	-3.381	-2.911	-2.425	-1.597	1.073	33
T[32,1]	-2.887	0.696	-4.267	-3.356	-2.885	-2.400	-1.576	1.070	34
T[33,1]	0.644	0.175	0.329	0.517	0.640	0.762	0.995	1.042	56
T[34,1]	-0.974	0.558	-2.174	-1.313	-0.927	-0.588	-0.004	1.013	200

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

	mean	sd	2.50%	25%	50%	75%	97.50%	Rhat	n.eff
T[35,1]	0.757	0.196	0.398	0.615	0.755	0.892	1.149	1.048	49
T[36,1]	0.873	0.253	0.421	0.694	0.865	1.040	1.397	1.040	57
T[37,1]	-0.394	0.464	-1.439	-0.662	-0.346	-0.074	0.392	1.006	540
T[38,1]	-2.477	0.657	-3.819	-2.914	-2.456	-2.004	-1.280	1.051	45
T[39,1]	-6.580	1.602	-10.061	-7.564	-6.466	-5.483	-3.730	1.056	41
T[40,1]	-3.363	0.701	-4.718	-3.832	-3.400	-2.880	-1.990	1.081	30
Tmu	-1.753	0.501	-2.783	-2.084	-1.733	-1.398	-0.843	1.041	55
Ttau	0.204	0.103	0.113	0.136	0.171	0.233	0.492	1.090	37
a[1]	0.808	0.066	0.674	0.765	0.810	0.852	0.931	1.001	2500
a[2]	1.020	0.083	0.870	0.964	1.016	1.072	1.196	1.001	6300
a[3]	1.088	0.090	0.927	1.027	1.083	1.144	1.282	1.000	5400
a[4]	1.032	0.080	0.884	0.978	1.028	1.081	1.202	1.002	1200
a[5]	1.016	0.084	0.863	0.959	1.012	1.068	1.193	1.001	3800
a[6]	1.074	0.088	0.917	1.014	1.068	1.128	1.265	1.001	3600
a[7]	0.970	0.076	0.832	0.920	0.966	1.018	1.133	1.001	2200
a[8]	1.053	0.087	0.896	0.993	1.047	1.106	1.242	1.000	7400
a[9]	0.863	0.066	0.731	0.820	0.863	0.905	0.992	1.000	24000
a[10]	1.048	0.087	0.891	0.989	1.043	1.102	1.234	1.000	4500
a[11]	1.012	0.082	0.867	0.956	1.006	1.063	1.188	1.000	7100
a[12]	1.064	0.083	0.913	1.008	1.060	1.116	1.238	1.000	13000
a[13]	1.391	0.125	1.181	1.304	1.379	1.464	1.676	1.002	1100
a[14]	1.069	0.089	0.908	1.007	1.065	1.126	1.258	1.001	7600
a[15]	1.100	0.101	0.918	1.031	1.094	1.161	1.317	1.001	1600
a[16]	0.803	0.080	0.637	0.750	0.807	0.859	0.951	1.001	3400
a[17]	1.021	0.085	0.870	0.962	1.016	1.074	1.204	1.000	30000
a[18]	1.028	0.080	0.884	0.972	1.024	1.078	1.196	1.000	4600
a[19]	0.973	0.075	0.835	0.922	0.970	1.021	1.130	1.000	12000
a[20]	0.774	0.073	0.625	0.725	0.776	0.825	0.909	1.000	7600
amu	0	0	0	0	0	0	0	1	1
atau	47.516	25.624	15.377	29.879	41.817	58.755	112.069	1.001	30000
b[1]	0.079	0.103	-0.118	0.012	0.076	0.142	0.293	1.006	1700
b[2]	0.020	0.112	-0.208	-0.049	0.020	0.091	0.241	1.005	30000
b[3]	-0.088	0.124	-0.354	-0.163	-0.081	-0.006	0.141	1.005	890
b[4]	0.197	0.110	0.000	0.122	0.189	0.264	0.437	1.012	220
b[5]	-0.057	0.122	-0.314	-0.131	-0.052	0.021	0.172	1.006	660
b[6]	-0.099	0.122	-0.361	-0.174	-0.093	-0.020	0.127	1.007	700
b[7]	-0.029	0.116	-0.269	-0.101	-0.026	0.045	0.198	1.004	7400
b[8]	-0.177	0.136	-0.471	-0.259	-0.165	-0.083	0.060	1.010	320
b[9]	0.130	0.106	-0.068	0.061	0.124	0.195	0.352	1.006	750
b[10]	0.173	0.121	-0.043	0.093	0.164	0.245	0.432	1.008	440

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

	mean	sd	2.50%	25%	50%	75%	97.50%	Rhat	n.eff
b[11]	0.057	0.113	-0.169	-0.012	0.056	0.126	0.288	1.004	12000
b[12]	-0.056	0.114	-0.295	-0.127	-0.051	0.018	0.160	1.006	840
b[13]	0.250	0.138	0.024	0.153	0.234	0.330	0.564	1.020	130
b[14]	-0.078	0.119	-0.333	-0.150	-0.070	0.000	0.141	1.004	1300
b[15]	-0.289	0.163	-0.648	-0.388	-0.274	-0.173	-0.011	1.012	210
b[16]	0.317	0.132	0.094	0.224	0.304	0.397	0.608	1.016	150
b[17]	-0.091	0.128	-0.366	-0.168	-0.083	-0.007	0.141	1.005	950
b[18]	-0.103	0.118	-0.359	-0.176	-0.097	-0.025	0.113	1.005	650
b[19]	-0.217	0.131	-0.498	-0.298	-0.208	-0.127	0.017	1.010	330
b[20]	0.132	0.106	-0.066	0.062	0.126	0.198	0.355	1.005	820
bmu	0	0	0	0	0	0	0	1	1
btau	34.073	25.812	8.026	17.496	26.991	42.212	102.542	1.043	67
deviance	969.180	19.979	932.143	955.284	968.299	982.140	1011.076	1.005	430
gam[1,1]	-3.968	0.783	-5.462	-4.499	-4.034	-3.426	-2.408	1.107	24
gam[2,1]	0.260	0.126	0.024	0.174	0.254	0.345	0.520	1.009	290
gam[3,1]	1.093	0.224	0.657	0.934	1.106	1.247	1.519	1.102	26
gam[4,1]	2.616	0.556	1.548	2.239	2.623	2.974	3.794	1.094	27
lam[1,1]	0.584	0.323	-0.046	0.367	0.580	0.794	1.239	1.005	430
lam[2,1]	0.337	0.432	-0.518	0.045	0.342	0.635	1.168	1.001	1500
lam[3,1]	-0.457	0.734	-1.951	-0.953	-0.436	0.053	0.927	1.001	3700
lam[4,1]	0.781	0.311	0.188	0.569	0.774	0.986	1.410	1.003	670
lam[5,1]	0.309	0.309	-0.268	0.099	0.298	0.507	0.954	1.003	790
lam[6,1]	-1.590	0.392	-2.248	-1.894	-1.605	-1.310	-0.793	1.000	30000
lam[7,1]	0.061	0.643	-1.394	-0.324	0.109	0.504	1.191	1.001	3600
lam[8,1]	-1.049	0.360	-1.774	-1.286	-1.041	-0.808	-0.366	1.001	3600
lam[9,1]	0.096	0.381	-0.689	-0.150	0.104	0.350	0.830	1.002	1300
lam[10,1]	0.541	0.603	-0.654	0.143	0.542	0.942	1.727	1.010	200
lam[11,1]	0.761	0.314	0.165	0.550	0.754	0.964	1.406	1.005	410
lam[12,1]	0.385	0.623	-0.913	-0.005	0.411	0.802	1.562	1.001	2000
lam[13,1]	1.082	0.310	0.519	0.869	1.061	1.278	1.739	1.005	600
lam[14,1]	1.626	0.298	1.070	1.411	1.619	1.840	2.203	1.005	430
lam[15,1]	-0.587	0.749	-2.032	-1.110	-0.582	-0.064	0.854	1.000	9500
lam[16,1]	0.245	0.310	-0.334	0.032	0.235	0.450	0.884	1.003	690
lam[17,1]	0.555	0.560	-0.572	0.191	0.567	0.931	1.623	1.004	470
lam[18,1]	-0.183	0.522	-1.361	-0.492	-0.134	0.178	0.724	1.002	2000
lam[19,1]	0.492	0.340	-0.170	0.264	0.487	0.716	1.172	1.004	630
lam[20,1]	0.870	0.305	0.324	0.658	0.852	1.064	1.519	1.006	390
lam[21,1]	-0.378	0.303	-0.948	-0.583	-0.388	-0.179	0.252	1.003	640
lam[22,1]	-0.772	0.352	-1.474	-1.005	-0.770	-0.537	-0.082	1.000	4200
lam[23,1]	-0.524	0.728	-2.004	-1.018	-0.491	-0.013	0.845	1.000	7300

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

	mean	sd	2.50%	25%	50%	75%	97.50%	Rhat	n.eff
lam[24,1]	0.816	0.293	0.275	0.615	0.804	1.004	1.434	1.005	410
lam[25,1]	0.775	0.311	0.185	0.564	0.767	0.977	1.416	1.003	640
lam[26,1]	-1.061	0.618	-2.174	-1.518	-1.071	-0.632	0.163	1.000	26000
lam[27,1]	0.401	0.310	-0.174	0.187	0.388	0.603	1.038	1.003	850
lam[28,1]	1.045	0.447	0.186	0.735	1.040	1.349	1.935	1.003	700
lam[29,1]	0.051	0.401	-0.716	-0.225	0.049	0.323	0.840	1.001	3700
lam[30,1]	-1.597	0.404	-2.252	-1.915	-1.622	-1.315	-0.766	1.000	8000
lam[31,1]	0.089	0.470	-0.884	-0.211	0.108	0.409	0.975	1.005	440
lam[32,1]	0.071	0.474	-0.915	-0.234	0.092	0.395	0.950	1.001	1500
lam[33,1]	-1.248	0.387	-2.063	-1.499	-1.232	-0.985	-0.512	1.001	4400
lam[34,1]	0.910	0.303	0.356	0.702	0.893	1.102	1.556	1.006	360
lam[35,1]	-0.810	0.334	-1.466	-1.031	-0.811	-0.587	-0.154	1.002	1400
lam[36,1]	-0.040	0.298	-0.596	-0.246	-0.049	0.157	0.569	1.004	520
lam[37,1]	0.676	0.321	0.066	0.456	0.670	0.885	1.337	1.004	610
lam[38,1]	0.533	0.330	-0.083	0.308	0.522	0.747	1.223	1.006	410
lam[39,1]	-0.533	0.732	-2.014	-1.028	-0.507	-0.016	0.845	1.000	30000
lam[40,1]	-1.052	0.618	-2.173	-1.512	-1.055	-0.626	0.186	1.001	1500
lammu	0	0	0	0	0	0	0	1	1
lamtau	1.098	0.527	0.305	0.705	1.037	1.411	2.295	1.005	370
pi	1	0	1	1	1	1	1	1	1

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

ผลการวิเคราะห์ผลการประเมินความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินระดับชั้นเรียน กลุ่มสาระการเรียนรู้วิทยาศาสตร์

	mean	sd	2.50%	25%	50%	75%	97.50%	Rhat	n.eff
E[1]	2.050	0.771	0.751	1.503	1.973	2.519	3.787	1.030	75
E[2]	1.908	0.709	0.687	1.402	1.841	2.340	3.464	1.030	71
E[3]	2.142	0.824	0.784	1.548	2.069	2.633	3.995	1.032	73
E[4]	1.734	0.678	0.621	1.248	1.664	2.133	3.261	1.027	89
E[5]	2.058	0.842	0.706	1.451	1.955	2.558	3.972	1.033	71
E[6]	2.856	1.210	1.013	1.984	2.700	3.544	5.666	1.039	62
E[7]	1.644	0.644	0.583	1.186	1.572	2.029	3.089	1.019	110
E[8]	1.746	0.683	0.623	1.258	1.671	2.159	3.270	1.022	93
E[9]	1.726	0.673	0.602	1.250	1.658	2.127	3.249	1.030	77
E[10]	2.363	0.899	0.862	1.727	2.273	2.887	4.393	1.027	76
E[11]	1.504	0.628	0.500	1.051	1.434	1.874	2.931	1.019	130
E[12]	1.518	0.660	0.467	1.036	1.439	1.911	3.025	1.013	180
E[13]	1.520	0.667	0.466	1.039	1.440	1.911	3.025	1.015	160
E[14]	1.559	0.640	0.533	1.109	1.481	1.929	3.029	1.024	98
E[15]	1.985	0.776	0.707	1.432	1.902	2.445	3.724	1.028	78
E[16]	2.045	0.812	0.712	1.473	1.950	2.523	3.869	1.035	65
E[17]	1.835	0.712	0.661	1.321	1.754	2.266	3.438	1.022	90
E[18]	2.454	0.939	0.892	1.772	2.369	3.024	4.557	1.031	68
E[19]	1.873	0.733	0.658	1.356	1.796	2.312	3.507	1.041	59
E[20]	2.313	0.978	0.777	1.612	2.185	2.858	4.607	1.042	58
Emu	0.564	0.365	-0.353	0.362	0.611	0.820	1.150	1.043	61
Etau	24.266	33.958	2.921	7.580	13.262	26.206	116.510	1.012	430
Om[1]	1	0	1	1	1	1	1	1	1
Om[2]	1	0	1	1	1	1	1	1	1
Om[3]	1	0	1	1	1	1	1	1	1
Om[4]	1	0	1	1	1	1	1	1	1
Om[5]	1	0	1	1	1	1	1	1	1
Om[6]	1	0	1	1	1	1	1	1	1
Om[7]	1	0	1	1	1	1	1	1	1
Om[8]	1	0	1	1	1	1	1	1	1
Om[9]	1	0	1	1	1	1	1	1	1
Om[10]	1	0	1	1	1	1	1	1	1
Om[11]	1	0	1	1	1	1	1	1	1
Om[12]	1	0	1	1	1	1	1	1	1
Om[13]	1	0	1	1	1	1	1	1	1
Om[14]	1	0	1	1	1	1	1	1	1

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

	mean	sd	2.50%	25%	50%	75%	97.50%	Rhat	n.eff
Om[15]	1	0	1	1	1	1	1	1	1
Om[16]	1	0	1	1	1	1	1	1	1
Om[17]	1	0	1	1	1	1	1	1	1
Om[18]	1	0	1	1	1	1	1	1	1
Om[19]	1	0	1	1	1	1	1	1	1
Om[20]	1	0	1	1	1	1	1	1	1
T[1,1]	4.867	1.697	1.846	3.670	4.761	5.926	8.547	1.030	70
T[2,1]	4.307	1.597	1.578	3.163	4.176	5.300	7.814	1.025	82
T[3,1]	3.907	1.389	1.464	2.935	3.788	4.758	6.928	1.025	78
T[4,1]	1.578	1.033	0.034	0.849	1.412	2.137	4.035	1.010	230
T[5,1]	4.296	1.537	1.607	3.205	4.191	5.246	7.647	1.031	64
T[6,1]	5.119	1.796	1.947	3.864	4.993	6.239	9.025	1.027	73
T[7,1]	2.663	1.075	0.925	1.893	2.550	3.291	5.095	1.017	110
T[8,1]	4.889	1.688	1.872	3.710	4.780	5.928	8.507	1.032	67
T[9,1]	5.482	1.904	2.087	4.138	5.361	6.671	9.574	1.029	69
T[10,1]	3.484	1.263	1.307	2.604	3.385	4.249	6.266	1.022	82
T[11,1]	4.748	1.738	1.745	3.520	4.613	5.804	8.642	1.028	74
T[12,1]	4.534	1.603	1.714	3.417	4.426	5.515	8.036	1.030	70
T[13,1]	4.749	1.656	1.824	3.592	4.620	5.779	8.343	1.030	70
T[14,1]	2.955	1.091	1.078	2.183	2.861	3.618	5.391	1.021	89
T[15,1]	4.840	1.678	1.847	3.666	4.735	5.881	8.501	1.031	68
T[16,1]	4.832	1.680	1.826	3.665	4.703	5.873	8.474	1.027	73
T[17,1]	4.956	1.734	1.868	3.755	4.835	6.030	8.766	1.031	69
T[18,1]	1.698	0.930	0.294	1.032	1.568	2.207	3.864	1.013	210
T[19,1]	4.839	1.681	1.810	3.677	4.733	5.872	8.438	1.031	69
T[20,1]	3.118	1.225	1.097	2.258	2.996	3.825	5.884	1.023	87
T[21,1]	5.446	1.912	2.042	4.098	5.319	6.668	9.551	1.031	67
T[22,1]	6.097	2.099	2.319	4.636	5.972	7.416	10.646	1.028	70
T[23,1]	2.364	0.888	0.858	1.733	2.288	2.902	4.344	1.023	84
T[24,1]	4.266	1.523	1.588	3.193	4.160	5.195	7.615	1.032	65
T[25,1]	4.283	1.501	1.600	3.235	4.185	5.204	7.528	1.033	64
T[26,1]	5.176	1.798	1.988	3.914	5.048	6.316	9.037	1.028	71
T[27,1]	4.322	1.525	1.634	3.269	4.201	5.266	7.643	1.032	67
T[28,1]	1.540	0.669	0.460	1.056	1.476	1.943	3.044	1.024	98
T[29,1]	6.104	2.131	2.309	4.596	5.960	7.441	10.748	1.024	80
T[30,1]	2.425	1.057	0.744	1.658	2.296	3.046	4.827	1.017	140
T[31,1]	4.505	1.599	1.718	3.380	4.401	5.479	7.963	1.029	68
T[32,1]	1.631	0.958	0.213	0.937	1.481	2.164	3.895	1.016	160
T[33,1]	4.900	1.697	1.852	3.716	4.782	5.963	8.505	1.028	71
T[34,1]	5.285	1.810	2.026	4.012	5.175	6.437	9.148	1.030	67

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

	mean	sd	2.50%	25%	50%	75%	97.50%	Rhat	n.eff
T[35,1]	4.851	1.716	1.841	3.646	4.709	5.885	8.641	1.030	75
T[36,1]	5.287	1.818	1.995	4.008	5.178	6.460	9.154	1.029	69
T[37,1]	5.501	1.911	2.107	4.162	5.367	6.698	9.600	1.028	73
T[38,1]	4.288	1.500	1.614	3.253	4.178	5.207	7.574	1.029	70
T[39,1]	4.859	1.700	1.824	3.658	4.742	5.920	8.590	1.029	70
T[40,1]	3.159	1.226	1.137	2.282	3.039	3.906	5.890	1.024	83
Tmu	4.126	1.244	1.650	3.282	4.132	4.962	6.590	1.038	59
Ttau	0.703	0.833	0.142	0.287	0.458	0.773	3.099	1.093	73
a[1]	1.336	0.363	0.739	1.081	1.296	1.542	2.166	1.004	550
a[2]	2.011	0.475	1.232	1.679	1.957	2.283	3.106	1.005	950
a[3]	0.734	0.232	0.350	0.568	0.711	0.872	1.257	1.002	1200
a[4]	2.592	0.587	1.645	2.180	2.519	2.920	3.970	1.007	800
a[5]	0.625	0.211	0.276	0.474	0.603	0.752	1.090	1.001	1300
a[6]	1.147	0.336	0.619	0.909	1.101	1.338	1.929	1.002	1500
a[7]	1.215	0.409	0.594	0.930	1.148	1.428	2.219	1.004	730
a[8]	0.961	0.350	0.425	0.718	0.908	1.150	1.799	1.003	1600
a[9]	1.029	0.287	0.540	0.825	1.004	1.201	1.659	1.001	5800
a[10]	1.052	0.327	0.531	0.823	1.009	1.237	1.806	1.002	1500
a[11]	1.240	0.403	0.616	0.964	1.183	1.450	2.198	1.006	760
a[12]	0.626	0.276	0.202	0.431	0.588	0.777	1.288	1.002	1000
a[13]	0.624	0.274	0.202	0.429	0.587	0.775	1.265	1.001	1300
a[14]	0.945	0.310	0.447	0.728	0.910	1.119	1.658	1.001	2700
a[15]	0.552	0.203	0.218	0.406	0.532	0.674	1.007	1.001	2200
a[16]	0.588	0.203	0.252	0.442	0.570	0.710	1.039	1.001	3300
a[17]	0.774	0.240	0.367	0.604	0.752	0.917	1.308	1.004	550
a[18]	1.575	0.414	0.904	1.281	1.525	1.814	2.524	1.002	3100
a[19]	1.377	0.357	0.784	1.128	1.339	1.587	2.188	1.001	11000
a[20]	0.988	0.287	0.528	0.783	0.954	1.154	1.641	1.000	5400
amu	0	0	0	0	0	0	0	1	1
atau	4.021	2.295	1.141	2.482	3.541	4.977	9.853	1.006	3200
b[1]	-0.168	0.693	-1.798	-0.502	-0.094	0.229	1.083	1.004	2600
b[2]	0.155	0.684	-1.332	-0.193	0.147	0.538	1.524	1.001	14000
b[3]	0.182	0.666	-1.183	-0.172	0.161	0.563	1.530	1.002	3000
b[4]	-0.332	0.813	-2.251	-0.691	-0.205	0.139	0.985	1.016	660
b[5]	0.659	0.727	-0.596	0.151	0.574	1.100	2.251	1.002	2100
b[6]	0.777	0.831	-0.629	0.191	0.684	1.287	2.606	1.003	1200
b[7]	-0.745	0.899	-2.980	-1.178	-0.556	-0.129	0.504	1.006	870
b[8]	-0.913	0.931	-3.261	-1.379	-0.716	-0.242	0.341	1.010	420
b[9]	0.316	0.656	-0.988	-0.057	0.267	0.683	1.729	1.002	1600
b[10]	-0.158	0.715	-1.817	-0.505	-0.081	0.248	1.154	1.000	30000

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

	mean	sd	2.50%	25%	50%	75%	97.50%	Rhat	n.eff
b[11]	-0.551	0.825	-2.627	-0.938	-0.388	-0.013	0.682	1.005	650
b[12]	-1.276	1.148	-4.072	-1.852	-1.035	-0.437	0.179	1.005	960
b[13]	-1.290	1.200	-4.206	-1.848	-1.040	-0.436	0.184	1.016	670
b[14]	-0.324	0.749	-2.125	-0.685	-0.210	0.122	0.947	1.009	610
b[15]	-0.134	0.723	-1.831	-0.468	-0.064	0.272	1.176	1.006	1400
b[16]	0.362	0.688	-0.990	-0.042	0.309	0.753	1.850	1.002	5500
b[17]	0.087	0.681	-1.350	-0.261	0.075	0.466	1.472	1.002	1500
b[18]	0.327	0.711	-1.113	-0.074	0.283	0.740	1.820	1.005	6000
b[19]	0.501	0.674	-0.710	0.067	0.435	0.898	1.977	1.003	2100
b[20]	0.926	0.807	-0.389	0.332	0.850	1.427	2.699	1.004	1500
bmu	0	0	0	0	0	0	0	1	1
btau	5.559	17.666	0.208	0.663	1.385	3.485	41.113	1.016	550
deviance	851.568	16.849	820.062	840.079	851.180	862.550	885.823	1.000	7900
gam[1,1]	-1.623	0.574	-2.856	-1.980	-1.579	-1.224	-0.596	1.048	49
gam[2,1]	-0.704	0.256	-1.263	-0.868	-0.684	-0.522	-0.252	1.037	61
gam[3,1]	0.406	0.184	0.121	0.272	0.382	0.513	0.829	1.033	74
gam[4,1]	1.921	0.638	0.719	1.488	1.890	2.345	3.254	1.050	47
lam[1,1]	-0.030	0.282	-0.586	-0.216	-0.029	0.155	0.528	1.000	11000
lam[2,1]	0.468	0.255	0.009	0.293	0.453	0.628	1.012	1.000	24000
lam[3,1]	-0.134	0.287	-0.713	-0.321	-0.131	0.056	0.426	1.000	15000
lam[4,1]	0.476	0.289	-0.034	0.278	0.454	0.649	1.114	1.001	3600
lam[5,1]	0.067	0.252	-0.408	-0.104	0.060	0.233	0.582	1.000	25000
lam[6,1]	0.048	0.258	-0.452	-0.126	0.047	0.218	0.561	1.000	30000
lam[7,1]	0.189	0.243	-0.259	0.023	0.180	0.343	0.698	1.000	8400
lam[8,1]	-0.136	0.298	-0.748	-0.325	-0.127	0.062	0.427	1.001	3900
lam[9,1]	0.004	0.272	-0.529	-0.176	0.004	0.181	0.547	1.000	30000
lam[10,1]	-0.054	0.264	-0.564	-0.230	-0.057	0.119	0.475	1.001	4700
lam[11,1]	0.393	0.259	-0.076	0.215	0.378	0.556	0.941	1.000	30000
lam[12,1]	0.040	0.260	-0.457	-0.135	0.035	0.211	0.561	1.000	30000
lam[13,1]	-0.085	0.295	-0.691	-0.275	-0.078	0.112	0.474	1.000	10000
lam[14,1]	-0.013	0.266	-0.539	-0.189	-0.014	0.162	0.510	1.000	6800
lam[15,1]	-0.198	0.317	-0.859	-0.396	-0.183	0.011	0.387	1.000	30000
lam[16,1]	-0.070	0.286	-0.648	-0.258	-0.064	0.120	0.480	1.000	5300
lam[17,1]	0.023	0.267	-0.505	-0.155	0.020	0.198	0.555	1.000	30000
lam[18,1]	0.269	0.272	-0.220	0.081	0.253	0.438	0.858	1.000	13000
lam[19,1]	-0.201	0.313	-0.860	-0.396	-0.191	0.010	0.382	1.000	30000
lam[20,1]	0.232	0.244	-0.221	0.065	0.223	0.386	0.741	1.000	8100
lam[21,1]	0.134	0.258	-0.362	-0.039	0.128	0.302	0.660	1.000	25000
lam[22,1]	-0.216	0.327	-0.927	-0.414	-0.198	0.000	0.384	1.001	2500
lam[23,1]	-0.265	0.261	-0.791	-0.436	-0.261	-0.091	0.234	1.000	6600

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

	mean	sd	2.50%	25%	50%	75%	97.50%	Rhat	n.eff
lam[24,1]	0.059	0.256	-0.434	-0.111	0.055	0.226	0.571	1.000	7100
lam[25,1]	-0.242	0.322	-0.911	-0.443	-0.231	-0.026	0.364	1.000	5200
lam[26,1]	-0.215	0.333	-0.939	-0.416	-0.199	0.007	0.390	1.000	8800
lam[27,1]	-0.103	0.292	-0.696	-0.292	-0.099	0.092	0.462	1.000	4700
lam[28,1]	-0.295	0.253	-0.799	-0.460	-0.291	-0.126	0.189	1.001	2600
lam[29,1]	-0.220	0.327	-0.912	-0.418	-0.204	0.001	0.378	1.001	4300
lam[30,1]	0.282	0.260	-0.185	0.102	0.269	0.442	0.832	1.000	11000
lam[31,1]	-0.111	0.281	-0.674	-0.296	-0.108	0.077	0.429	1.000	6400
lam[32,1]	0.263	0.285	-0.251	0.068	0.247	0.439	0.882	1.001	4600
lam[33,1]	-0.130	0.298	-0.738	-0.323	-0.123	0.072	0.436	1.000	6300
lam[34,1]	-0.061	0.287	-0.633	-0.248	-0.059	0.130	0.496	1.001	5000
lam[35,1]	0.078	0.265	-0.437	-0.099	0.072	0.251	0.611	1.000	8000
lam[36,1]	-0.060	0.290	-0.646	-0.246	-0.055	0.132	0.502	1.001	4900
lam[37,1]	-0.109	0.296	-0.713	-0.298	-0.103	0.088	0.458	1.000	30000
lam[38,1]	-0.242	0.318	-0.914	-0.440	-0.232	-0.025	0.351	1.000	5800
lam[39,1]	-0.080	0.282	-0.646	-0.267	-0.075	0.108	0.471	1.000	7600
lam[40,1]	0.236	0.246	-0.210	0.067	0.223	0.392	0.761	1.000	5500
lammu	0	0	0	0	0	0	0	1	1
lamtau	9.132	3.500	3.232	6.317	8.934	11.923	15.503	1.001	3300
pi	1	0	1	1	1	1	1	1	1

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

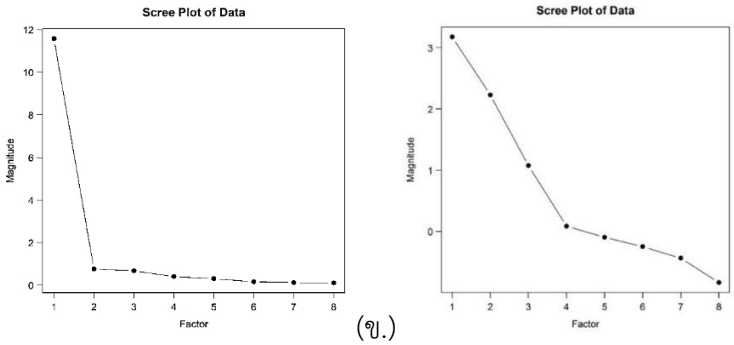
กระบวนการวิเคราะห์ความสอดคล้องในแนวเดียวกัน ด้วยโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรม

1. การเตรียมข้อมูลสำหรับการวิเคราะห์

ข้อมูลที่ใช้ในการวิเคราะห์ฉันทามติเชิงวัฒนธรรมมีลักษณะเป็นเมตริกซ์ขนาด $N \, x \, M$ โดย N เป็นจำนวนผู้ประเมินอยู่ในตำแหน่งแถว และ M เป็นจำนวนข้อคำถามประเมินอยู่ในตำแหน่งคอลัมน์ โดยข้อมูลดังกล่าวไม่จำเป็นต้องมีตำแหน่งหัวตาราง นักวิจัยสามารถเตรียมข้อมูลในรูปแบบไฟล์ .xslx หรือ .csv เพื่อสามารถนำเข้าในโปรแกรม R ได้ ในส่วนของการเตรียมการวิเคราะห์ข้อมูล นักวิจัยต้องติดตั้งโปแกรม R และติดตั้งชุดคำสั่ง CCTPack สำหรับใช้ในการวิเคราะห์ข้อมูล

2. การตรวจสอบจำนวนกลุ่มวัฒนธรรม

การตรวจสอบจำนวนกลุ่มของผู้ประเมินที่มีความสอดคล้องกันเพื่อเลือกโมเดลการวิเคราะห์ที่เหมาะสม ขั้นตอนนี้เป็นขั้นตอนที่นักวิจัยสามารถตรวจสอบการทำหน้าที่ต่างกันของผู้ประเมินเบื้องต้นได้ว่าการให้คะแนนการประเมินมีการทำหน้าที่แตกต่างกันระหว่างผู้ประเมินหรือไม่ โดยทั่วไปแล้ว โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมมีข้อตกลงเบื้องต้นว่าคำตอบของผู้ให้ข้อมูล (หรือผู้ประเมิน ในการศึกษานี้) มีความเป็นเอกพันธ์ หรือมีฉันทามติเดียวกัน (single-culture truth) อย่างไรก็ตาม ได้มีการศึกษาและพัฒนาโมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมเพื่อวิเคราะห์ข้อมูลที่มีการตอบ/การประเมินที่ไม่สอดคล้องกันระหว่างผู้ประเมินจำนวนมากกว่า 1 กลุ่มวัฒนธรรม ได้แก่ โมเดล MC-GCM และโมเดล MC-LTRM (Batchelder และ Anders, 2012; Anders และ Batchelder, 2015) ผลการตรวจสอบจำนวนกลุ่มวัฒนธรรมพิจารณาจากจำนวนองค์ประกอบดังแสดงในรูป 4.1 (ก.) แสดงว่าคำตอบของผู้ประเมินมีความสอดคล้องเป็นอันหนึ่งอันเดียวกันทั้งหมด ในขณะที่รูป 4.1 (ข.) แสดงว่าคำตอบของผู้ประเมินมีความแตกต่างกันระหว่างกลุ่มของผู้ประเมินมากกว่า 1 กลุ่ม



(ก.) (ข.)

รูป 4.1 จำนวนกลุ่มการทำหน้าที่ต่างกันของผู้ประเมิน

3. การเลือกโมเดลสำหรับวิเคราะห์ข้อมูล

หลังจากนักวิจัยตรวจสอบจำนวนกลุ่มฉันทามติของผู้ประเมินแล้วว่ามีการทำหน้าที่ต่างกันของผู้ประเมินหรือไม่ ขั้นตอนต่อไป คือ การเลือกโมเดลสำหรับการวิเคราะห์ข้อมูล สำหรับการวิจัยทางสังคมศาสตร์ และการศึกษาเกี่ยวกับความเที่ยงระหว่างผู้ประเมิน ข้อมูลส่วนใหญ่อยู่ในรูปของการให้คะแนนแบบ 0 – 1 หรือให้คะแนนบนมาตรเรียงอันดับ โมเดลที่เหมาะสมสำหรับการวิเคราะห์ข้อมูลดังกล่าว ได้แก่

การประเมินที่ให้คะแนนแบบ 0 – 1 ที่**ไม่มี**การทำหน้าที่ต่างกันของผู้ประเมิน	General Condorcet Model (GCM)
การประเมินที่ให้คะแนนแบบ 0 – 1 ที่**มี**การทำหน้าที่ต่างกันของผู้ประเมิน	Multi-Culture General Condorcet Model (MC-GCM)
การประเมินที่ให้คะแนนแบบเรียงอันดับ ที่**ไม่มี**การทำหน้าที่ต่างกันของผู้ประเมิน	Latent Truth Rater Model (LTRM)
การประเมินที่ให้คะแนนแบบเรียงอันดับ ที่**มี**การทำหน้าที่ต่างกันของผู้ประเมิน	Multi-Culture Latent Truth Rater Model (MC-LTRM)

นักวิจัยไม่จำเป็นต้องเลือกโมเดลในการวิเคราะห์ด้วยตนเอง เมื่อนำเข้าข้อมูลและพิมพ์คำสั่ง cctapply(data) หากไม่มีการปรับแต่งอาร์กิวเมนต์เพิ่มเติม โมเดลจะประมาณค่าข้อมูลโดยเลือกโมเดลที่เหมาะสมกับข้อมูลโดยอัตโนมัติ อย่างไรก็ตาม นักวิจัยสามารถกำหนดรายละเอียดการจำลองข้อมูลและการวิเคราะห์ข้อมูลได้ตามรายละเอียดที่ปรากฏในคู่มือของชุดคำสั่ง CCTpack (Anders, 2017) หรือดูรายละเอียดเพิ่มเติมได้ในส่วนช่วยเหลือ (help) ของโปรแกรม R

4. การแปลผลและสรุปผลการวิเคราะห์

โมเดลการวิเคราะห์ฉันทามติเชิงวัฒนธรรมแต่ละโมเดลมีพารามิเตอร์ที่แตกต่างกัน สำหรับโมเดล GCM จะมีพารามิเตอร์ที่สำคัญที่ต้องแปลความหมายอยู่ 4 พารามิเตอร์ คือ พารามิเตอร์คำตอบของการประเมิน (Z_k) พารามิเตอร์ความสามารถของผู้ประเมิน (D_i) พารามิเตอร์ความลำเอียงในการประเมิน (g_i) และพารามิเตอร์ความยากของการประเมิน (δ_k) หากมีการทำหน้าที่ต่างกันของผู้ประเมิน จะมีพารามิเตอร์จำนวนกลุ่มวัฒนธรรม (e_i) เพิ่มขึ้นมาอีกพารามิเตอร์ สำหรับโมเดล LTRM จะประกอบด้วยพารามิเตอร์ตำแหน่งคะแนนการประเมิน $(\mathrm{T_k})$ พารามิเตอร์เทรชโฮลด์ร่วมระหว่างผู้ประเมิน (γ_c) พารามิเตอร์ความยากของรายการประเมิน (λ_k) พารามิเตอร์ระดับความสามารถของผู้ประเมิน (E_i) และพารามิเตอร์ความลำเอียงในการประเมิน $(a_i,\ b_i)$ และในกรณีที่มีการทำหน้าที่ต่างกันของผู้ประเมินจะมีพารามิเตอร์ (Ω_i) แสดงจำนวนกลุ่มวัฒนธรรมของผู้ประเมิน

ประวัติผู้เขียน

ชื่อ-สกุล	นางสาวศิธรา จุฑารัตน์
วัน เดือน ปี เกิด	13 ธันวาคม 2524
สถานที่เกิด	กรุงเทพมหานคร
วุฒิการศึกษา	สำเร็จการศึกษาปริญญาครุศาสตรบัณฑิต จากคณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ปีการศึกษา 2547 สำเร็จการศึกษาปริญญาครุศาสตรมหาบัณฑิต สาขาการสอนภาษาไทย จากคณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ปีการศึกษา 2550 ศึกษาระดับปริญญาดุษฎีบัณฑิต สาขาวิชาการวัดและประเมินผลการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2558

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY