ประสิทธิภาพของการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบภายใต้เงื่อนไขที่แตกต่างกัน

นางสาวนริศรา เสือคล้าย

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาครุศาสตรดุษฎีบัณฑิต
สาขาวิชาการวัดและประเมินผลการศึกษา ภาควิชาวิจัยและจิตวิทยาการศึกษา
คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2559

THE EFFICIENCY OF CLASSIFICATION INDICES ESTIMATIONS BASED ON ITEM RESPONSE THEORY UNDER DIFFERENT CONDITIONS

Miss Narissara Suaklay

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy Program in Educational Measurement and Evaluation
Department of Educational Research and Psychology
Faculty of Education
Chulalongkorn University
Academic Year 2016

หัวข้อวิทยานิพนธ์	ประสิทธิภาพของการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบภายใต้เงื่อนไขที่แตกต่างกัน
โดย	นางสาวนริศรา เสือคล้าย
สาขาวิชา	การวัดและประเมินผลการศึกษา
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.ณัฏฐภรณ์ หลาวทอง
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	ศาสตราจารย์ ดร.ศิริชัย กาญจนวาสี

คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาดุษฎีบัณฑิต

.....คณบดีคณะครุศาสตร์

(รองศาสตราจารย์ ดร. ศิริเดช สุชีวะ)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ

(รองศาสตราจารย์ ดร.ศิริเดช สุชีวะ)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(ผู้ช่วยศาสตราจารย์ ดร.ณัฏฐภรณ์ หลาวทอง)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม

(ศาสตราจารย์ ดร.ศิริชัย กาญจนวาสี)

.....กรรมการ

(รองศาสตราจารย์ ดร.โชติกา ภาษีผล)

.....กรรมการ

(รองศาสตราจารย์ ดร. วรรณี แกมเกตุ)

.....กรรมการภายนอกมหาวิทยาลัย

(ผู้ช่วยศาสตราจารย์ ดร.สังวรณ์ งัดกระโทก)

นริศรา เสือคล้าย : ประสิทธิภาพของการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบภายใต้เงื่อนไขที่แตกต่างกัน (THE EFFICIENCY OF CLASSIFICATION INDICES ESTIMATIONS BASED ON ITEM RESPONSE THEORY UNDER DIFFERENT CONDITIONS) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ผศ. ดร.ณัฏฐภรณ์ หลาวทอง, อ.ที่ปรึกษาวิทยานิพนธ์ร่วม: ศ. ดร.ศิริชัย กาญจนวาสี, 234 หน้า.

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อประมาณค่าและเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทระหว่างวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบสามวิธี คือ วิธีการของ Rudner (2005) วิธีการของ Guo (2006) และวิธีการของ Lee (2010) ดำเนินการวิจัยโดยการจำลองข้อมูลด้วยโปรแกรม WINGEN ภายใต้เงื่อนไขของตัวแปรต้น 3 ตัวแปร ได้แก่ ความยาวของแบบสอบ (25, 50 ข้อ) โมเดลการวัด (1PL, 2PL, 3PL) และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ (10%, 20%) ทำการประมาณค่าดัชนีด้วยโปรแกรม R และพิจารณาประสิทธิภาพของวิธีการประมาณค่าจากค่าเฉลี่ยดัชนีการจำแนกประเภทจากการทำซ้ำ 100 รอบ นอกจากนี้ยังนำวิธีการประมาณค่าดัชนีการจำแนกประเภททั้งสามวิธีไปใช้กับข้อมูลเชิงประจักษ์ คือ คะแนนสอบของนักเรียนชั้นมัธยมศึกษาปีที่ 3 จากการสอบ O-NET ปีการศึกษา 2556 จำนวน 8,000 คน เพื่อประมาณค่าดัชนีการจำแนกประเภทและหาประสิทธิภาพของวิธีการต่อไป

ผลการวิจัยสรุปได้ดังนี้

1. ผลจากการจำลองข้อมูลภายใต้สถานการณ์เงื่อนไขทั้งหมดพบว่า วิธีการของ Rudner มีค่าดัชนีความถูกต้องสูง (0.8234-0.9086) และดัชนีความสอดคล้องค่อนข้างสูง (0.7550-0.8749) วิธีการของ Guo มีค่าดัชนีความถูกต้องสูง (0.9987- 1) และดัชนีความสอดคล้องสูง (0.9982- 1) และวิธีการของ Lee มีค่าดัชนีความถูกต้องค่อนข้างสูง (0.6285- 0.7496) และดัชนีความสอดคล้องปานกลาง (0.5372- 0.6938)

2. ผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทพบว่า วิธีการของ Guo เป็นวิธีการที่มีประสิทธิภาพสูงกว่าวิธีการของ Rudner และวิธีการของ Lee อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 โดยวิธีการประมาณค่า ความยาวของแบบสอบ โมเดลการวัด และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบมีอิทธิพลร่วมกันต่อดัชนีความถูกต้องและดัชนีความสอดคล้องอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ด้วยขนาดอิทธิพล .624 และ .656 ตามลำดับ

3. ผลจากการนำไปใช้กับข้อมูลเชิงประจักษ์พบว่า วิธีการของ Rudner มีค่าดัชนีความถูกต้องค่อนข้างสูง (0.6676-0.7581) และดัชนีความสอดคล้องค่อนข้างสูง (0.5579-0.6628) วิธีการของ Guo มีค่าดัชนีความถูกต้องสูงมีค่าเท่ากับ 1 และดัชนีความสอดคล้องสูงมีค่าเท่ากับ 1 และวิธีการของ Lee มีค่าดัชนีความถูกต้องค่อนข้างสูง (0.6223-0.6445) และดัชนีความสอดคล้องปานกลาง (0.5173- 0.5570)

ผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทพบว่า วิธีการของ Guo เป็นวิธีการที่มีประสิทธิภาพสูงกว่าวิธีการของ Rudner และวิธีการของ Lee อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

ภาควิชา วิจัยและจิตวิทยาการศึกษา ลายมือชื่อนิสิต .....

สาขาวิชา การวัดและประเมินผลการศึกษา ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

ปีการศึกษา 2559 ลายมือชื่อ อ.ที่ปรึกษาร่วม .....

5584214727 : MAJOR EDUCATIONAL MEASUREMENT AND EVALUATION
KEYWORDS: CLASSIFICATION ACCURACY / CLASSIFICATION CONSISTENCY / ITEM RESPONSE THEORY (IRT)

NARISSARA SUAKLAY: THE EFFICIENCY OF CLASSIFICATION INDICES ESTIMATIONS BASED ON ITEM RESPONSE THEORY UNDER DIFFERENT CONDITIONS. ADVISOR: ASST. PROF. NUTTAPORN LAWTHONG, Ph.D., CO-ADVISOR: PROF. SIRICHAI KANJANAWASEE, Ph.D., 234 pp.

This study purposed to estimate the classification indices with three methods (Rudner, Guo, Lee) based on Item Response Theory (IRT) and to compare the efficiency of those. The data was simulated with WINGEN program under different experimental conditions: test length (25, 50 items), measurement model (1PL, 2PL, 3PL) and model misfit (10%, 20%). R program was used to estimate the classification indices. The efficiency of classification methods was evaluated through mean of classification indices over the 100 replications. In addition, three methods were applied to the empirical data --O-NET Examination-- purposing to estimate the classification indices and to investigate the efficiency of those.

The results indicated that:

1. In simulation study, it was found that Rudner's method was effectiveness as high level of classification accuracy (0.8234-0.9086) and classification consistency (0.7550-0.8749), Guo's method was effectiveness as high level of classification accuracy (0.9987- 1) and classification consistency (0.9982- 1) and Lee's method effectiveness rather high level of classification accuracy (0.6285- 0.7496) and as moderate classification consistency (0.5372- 0.6938).

2. The comparison results for three methods were found that Guo's method had higher effectively than Rudner's method and Lee's method of .05 as statistical significance level. Test length, measurement model and model misfit had interaction influence to accuracy and consistency classification indices of .05 as statistical level with effect size .624 and .656 respectively.

3. For empirical data, it was found that Rudner's method was effectiveness rather high level of classification accuracy (0.6676-0.7581) and classification consistency (0.5579-0.6628), Guo's method was effectiveness as high level of classification accuracy (1) and classification consistency (1) and Lee's method effectiveness rather high level of classification accuracy (0.6223-0.6445) and as moderate classification consistency (0.5173- 0.5570).

The comparison results for three methods were found that Guo's method had higher effectively than Rudner's method and Lee's method of .05 as statistical significance level.

Department:	Educational Research and Psychology	Student's Signature .....
		Advisor's Signature .....
Field of Study:	Educational Measurement and Evaluation	Co-Advisor's Signature .....
Academic Year:	2016	

กิตติกรรมประกาศ

วิทยานิพนธ์เล่มนี้สำเร็จลงได้ด้วยความเมตตากรุณา และความเอาใจใส่เป็นอย่างดียิ่งของอาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก ผู้ช่วยศาสตราจารย์ ดร.ณัฏฐภรณ์ หลาวทอง และอาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม ศาสตราจารย์ ดร.ศิริชัย กาญจนวาสี ที่ได้สละเวลาในการให้คำปรึกษาแนะนำแนวทางที่ถูกต้อง ตลอดจนแก้ไขข้อบกพร่องต่างๆ ด้วยความละเอียดถี่ถ้วน อีกทั้งยังคอยกำกับติดตามอย่างสม่ำเสมอ และให้กำลังใจในการทำวิทยานิพนธ์จนสำเร็จลุล่วงไปด้วยดี ผู้วิจัยรู้สึกซาบซึ้งเป็นอย่างยิ่ง และขอกราบขอบพระคุณท่านอาจารย์เป็นอย่างสูงมา ณ โอกาสนี้

ขอกราบขอบพระคุณ รองศาสตราจารย์ ดร.ศิริเดช สุชีวะ ประธานกรรมการสอบวิทยานิพนธ์ รองศาสตราจารย์ ดร.โชติกา ภาษีผล รองศาสตราจารย์ ดร.วรรณี แกมเกตุ รองศาสตราจารย์ ดร.กมลวรรณ ตังธนกานนท์ และผู้ช่วยศาสตราจารย์ ดร.สังวรณ์ งัดกระโทก กรรมการสอบวิทยานิพนธ์และกรรมการสอบหัวข้อวิทยานิพนธ์ ที่ได้เมตตาให้ความรู้ คำแนะนำ และข้อเสนอแนะที่เป็นประโยชน์ต่อการปรับปรุงแก้ไขวิทยานิพนธ์ให้มีความสมบูรณ์ชัดเจนมากยิ่งขึ้น

ขอขอบพระคุณอาจารย์ ดร.สิวะโชติ ศรีสุทธิยากร และอาจารย์ ดร.สุรศักดิ์ เก้าเอี้ยน อาจารย์ประจำสาขาสถิติการศึกษา ภาควิชาวิจัยและจิตวิทยาการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย สำหรับความเป็นกัลยาณมิตรที่ดีในการถ่ายทอดความรู้ ให้คำแนะนำและคำปรึกษาในการเขียนคำสั่งที่ใช้ในโปรแกรม R จนทำให้การวิเคราะห์ข้อมูลสำเร็จไปได้ด้วยดี และขอขอบพระคุณอาจารย์ ดร.ชุติวัฒน์ สุวัตถิพงศ์ อาจารย์ประจำสำนักเทคโนโลยีการศึกษา มหาวิทยาลัยสุโขทัยธรรมาธิราช สำหรับความเป็นกัลยาณมิตรที่ดีในการให้คำแนะนำและคำปรึกษาในการดำเนินกิจกรรมที่เกี่ยวข้องกับระบบคอมพิวเตอร์และออนไลน์

ขอกราบขอบพระคุณสถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน) ที่ได้อนุเคราะห์เอื้อเฟื้อข้อมูลคะแนนการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET) ระดับชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2556 วิชาคณิตศาสตร์และภาษาไทยเพื่อใช้เป็นข้อมูลในการทำวิจัยครั้งนี้อย่างดียิ่ง

ขอขอบพระคุณโครงการ ทุน 90 ปี จุฬาลงกรณ์มหาวิทยาลัย กองทุนรัชดาภิเษกสมโภช ที่สนับสนุนทุนสำหรับดำเนินการวิทยานิพนธ์ในครั้งนี้

ขอขอบคุณพี่นิอร ไชยพรพัฒนา พี่อุทุมพร ชาติเผือก คุณชรินรัตน์ พุ่มเกษม คุณอังค์วรา วงษ์รักษา พี่นัทชา ปรีชาวัฒนสกุล พี่นฤนาท เข่งเงิน พี่สิริวิมล ศุกรศร พี่สุนทรัตร์ แสงงาม พี่ธนิสา สังขจันทร์ และพี่ๆ ที่สถาบันภาษาไทยสิรินธร ที่คอยหยิบยื่นความช่วยเหลือให้ทุกเมื่อที่ต้องการและคอยเป็นกำลังใจให้เสมอมา

ขอขอบคุณพี่อาภาพรรณ ประทุมไทย พี่นิลวิศาล เฮงสมบูรณ์ น้องธนิยา เยาดำ น้องกุลรตี พันธุ์แฉล้ม เพื่อนนิสิตทุกคนในรุ่น รุ่นพี่ และรุ่นน้องสาขาการวัดและประเมินผลการศึกษา ที่แบ่งปันความรู้ซึ่งกันและกัน คอยช่วยเหลือ และเป็นกำลังใจที่ดีให้กันเสมอมา รวมถึงผู้มีส่วนเกี่ยวข้องทุกท่านที่มิได้ปรากฏชื่อในที่นี้ ที่มีส่วนช่วยเหลือในการทำวิทยานิพนธ์ฉบับนี้จนสำเร็จลุล่วงไปด้วยดี

ท้ายที่สุดขอกราบขอบพระคุณบิดา มารดา และบุคคลในครอบครัว ที่เป็นผู้วางรากฐานทางการศึกษา ให้การสนับสนุนในทุกๆ ด้าน อีกทั้งยังหล่อหลอมสิ่งดีๆ ให้แก่ผู้วิจัยมาโดยตลอด

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ง
บทคัดย่อภาษาอังกฤษ ..... จ
กิตติกรรมประกาศ ..... ฉ
สารบัญ ..... ช
สารบัญตาราง ..... ญ
สารบัญภาพ ..... ฑ
บทที่ 1 บทนำ 1
ความเป็นมาและความสำคัญของปัญหา 1
คำถามการวิจัย 7
วัตถุประสงค์การวิจัย 8
สมมติฐานการวิจัย 8
ขอบเขตการวิจัย 9
คำจำกัดความที่ใช้ในการวิจัย 11
ประโยชน์ที่ได้รับ 14
บทที่ 2 เอกสารและงานวิจัยที่เกี่ยวข้อง 16
ตอนที่ 1 ความรู้เบื้องต้นเกี่ยวกับดัชนีการจำแนกประเภท 16
ตอนที่ 2 แนวคิดด้านการวัดผลทางการศึกษาที่อธิบายเกี่ยวกับดัชนีการจำแนกประเภท 21
ตอนที่ 3 วิธีการประมาณค่าดัชนีการจำแนกประเภท 23
ตอนที่ 4 งานวิจัยที่เกี่ยวข้องกับดัชนีการจำแนกประเภท 61
ตอนที่ 5 กรอบแนวคิดในการวิจัย 74
บทที่ 3 วิธีดำเนินการวิจัย 78

หน้า

ขั้นตอนที่ 1 การศึกษาผลการประมาณค่าดัชนีการจำแนกประเภทด้วยวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบ 3 วิธี คือ วิธีการของ Rudner (2005) วิธีการของ Guo (2006) และวิธีการของ Lee (2010) 80

ขั้นตอนที่ 2 การเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภททั้งสามวิธี 94

ขั้นตอนที่ 3 การศึกษาผลการประมาณค่าและเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบทั้งสามวิธี เมื่อใช้กับข้อมูลเชิงประจักษ์ 95

บทที่ 4 ผลการวิเคราะห์ข้อมูล 101

ตอนที่ 1 ผลการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบสามวิธีภายใต้การศึกษาการจำลองข้อมูล (simulation study) 101

ตอนที่ 2 ผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทตามแนวคิดทฤษฎีการตอบสนองข้อสอบทั้งสามวิธีภายใต้การศึกษาการจำลองข้อมูล (simulation study) 150

ตอนที่ 3 ผลการประมาณค่าและผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบทั้งสามวิธี เมื่อนำไปใช้กับข้อมูลเชิงประจักษ์ (empirical data) 156

บทที่ 5 สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ 191

สรุปผลการวิจัย 192

อภิปรายผลการวิจัย 198

ข้อเสนอแนะจากการวิจัย 202

รายการอ้างอิง 205

ภาคผนวก 210

ภาคผนวก ก การจำลองข้อมูลการตอบข้อสอบด้วยโปรแกรม WINGEN3 211

ภาคผนวก ข คำสั่งที่ใช้ในโปรแกรม R 217

หน้า

ประวัติผู้เขียนวิทยานิพนธ์ 234

สารบัญตาราง

หน้า

ตารางที่ 2.1 ความสอดคล้องของการจำแนกประเภท (classification consistency)..... 19

ตารางที่ 2.2 ความถูกต้องของการจำแนกประเภท (classification accuracy)..... 20

ตารางที่ 2.3 การให้คะแนนที่ใช้ในการประมาณค่า..... 30

ตารางที่ 2.4 ลักษณะเฉพาะของวิธีการประมาณค่าความสอดคล้องของการจำแนกที่พัฒนาโดยนักวัดผลทางการศึกษา..... 41

ตารางที่ 2.5 ผลการตัดสินผ่าน/ตกแบบสองคูณสอง..... 42

ตารางที่ 2.6 ตารางการจำแนกประเภท (Classification table)..... 47

ตารางที่ 2.7 พารามิเตอร์ของข้อสอบตาม Generalized Partial Credit Model สำหรับข้อสอบการอ่าน 10 ข้อ..... 52

ตารางที่ 2.8 ร้อยละของผู้สอบในแต่ละกลุ่มคะแนนจากการจำแนกประเภทที่คาดหวัง (expected classification table)..... 53

ตารางที่ 2.9 ตัวอย่างของตารางการจำแนกประเภท (classification table)..... 55

ตารางที่ 2.10 ลักษณะเฉพาะของวิธีการประมาณค่าความถูกต้องของการจำแนกที่พัฒนาโดยนักวัดผลทางการศึกษา..... 60

ตารางที่ 2.11 การแจกแจงความถี่ตามประเด็นที่พบจากงานวิจัยที่ศึกษาเกี่ยวกับดัชนีการจำแนกประเภท..... 63

ตารางที่ 3.1 ขั้นตอนในการดำเนินการวิจัย..... 79

ตารางที่ 3.2 ข้อตกลงเบื้องต้นและสูตรที่ใช้ในการประมาณค่าดัชนีการจำแนกประเภทของวิธีการประมาณค่าตามทฤษฎีการตอบสนองข้อสอบ..... 84

ตารางที่ 3.3 รูปแบบข้อสอบที่ใช้ในการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐานสำหรับนักเรียนชั้นมัธยมศึกษาปีที่ 3 ประจำปีการศึกษา 2556 จำแนกตามวิชา..... 87

ตารางที่ 3.4 ระดับการประเมินผลการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET) สำหรับนักเรียนชั้นมัธยมศึกษาปีที่ 3..... 87

ตารางที่ 3.5 คะแนนจุดตัดวิชาคณิตศาสตร์ ชุด A ที่ปรับเทียบตามสเกลของ θ 89

ตารางที่ 3.6 คะแนนจุดตัดวิชาคณิตศาสตร์ ชุด B ที่ปรับเทียบตามสเกลของ θ 89

ตารางที่ 3.7 คะแนนจุดตัดวิชาภาษาไทย ชุด A ที่ปรับเทียบตามสเกลของ θ 90

ตารางที่ 3.8 คะแนนจุดตัดวิชาภาษาไทย ชุด B ที่ปรับเทียบตามสเกลของ θ 90

ตารางที่ 3.9 จำนวนประชากรและกลุ่มตัวอย่างจำแนกตามรายวิชา 96

ตารางที่ 4.1 สถานการณ์เงื่อนไขในการจำลองข้อมูล 103

ตารางที่ 4.2 ค่าพารามิเตอร์ข้อสอบของข้อมูลจำลองสถานการณ์ที่ 1-2 105

ตารางที่ 4.3 ค่าพารามิเตอร์ข้อสอบของข้อมูลจำลองสถานการณ์ที่ 3-4 108

ตารางที่ 4.4 ค่าพารามิเตอร์ข้อสอบของข้อมูลจำลองสถานการณ์ที่ 5-6 111

ตารางที่ 4.5 ค่าพารามิเตอร์ข้อสอบของข้อมูลจำลองสถานการณ์ที่ 7-8 115

ตารางที่ 4.6 ค่าพารามิเตอร์ข้อสอบของข้อมูลจำลองสถานการณ์ที่ 9-10 119

ตารางที่ 4.7 ค่าพารามิเตอร์ข้อสอบของข้อมูลจำลองสถานการณ์ที่ 11-12 124

ตารางที่ 4.8 ค่าสถิติพื้นฐานของค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) ที่ได้จากการจำลองข้อมูล 129

ตารางที่ 4.9 ค่าเฉลี่ยดัชนีความถูกต้องและความสอดคล้องของการจำแนกประเภทจากการทำซ้ำจำนวน 100 รอบ ในสถานการณ์จำลองจำแนกตามวิธีการประมาณค่า 135

ตารางที่ 4.10 ผลการวิเคราะห์ความแปรปรวนแบบสามทาง (3-WAY ANOVA) ของค่าเฉลี่ยดัชนีความถูกต้องและดัชนีความสอดคล้องของการจำแนกประเภทจากการทำซ้ำจำนวน 100 รอบ ในสถานการณ์จำลองโดยใช้วิธีการของ Rudner 141

ตารางที่ 4.11 ผลการวิเคราะห์ความแปรปรวนแบบสามทาง (3-WAY ANOVA) ของค่าเฉลี่ยดัชนีความถูกต้องและดัชนีความสอดคล้องของการจำแนกประเภทจากการทำซ้ำจำนวน 100 รอบ ในสถานการณ์จำลองโดยใช้วิธีการของ Guo 143

ตารางที่ 4.12 ผลการวิเคราะห์ความแปรปรวนแบบสามทาง (3-WAY ANOVA) ของค่าเฉลี่ยดัชนีความถูกต้องและดัชนีความสอดคล้องของการจำแนกประเภทจากการทำซ้ำจำนวน 100 รอบ ในสถานการณ์จำลองโดยใช้วิธีการของ Lee 144

ตารางที่ 4.13 ผลการเปรียบเทียบค่าเฉลี่ยดัชนีความถูกต้องและดัชนีความสอดคล้องของการจำแนกประเภทจากการทำซ้ำของปัจจัยที่มีอิทธิพลต่อค่าเฉลี่ยดัชนีการจำแนกประเภท จำแนกตามวิธีการประมาณค่า 146

ตารางที่ 4.14 ผลการวิเคราะห์ความแตกต่างของค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทจากการทำซ้ำจำนวน 100 รอบ ในสถานการณ์จำลองจำแนกตามวิธีการประมาณค่า 150

ตารางที่ 4.15 ผลการวิเคราะห์ความแตกต่างของค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทจากการทำซ้ำจำนวน 100 รอบ ในสถานการณ์จำลองจำแนกตามวิธีการประมาณค่า 152

ตารางที่ 4.16 ผลการวิเคราะห์ความแปรปรวนแบบสี่ทาง (4-WAY ANOVA) ของค่าเฉลี่ยดัชนีความถูกต้องและดัชนีความสอดคล้องของการจำแนกประเภทจากการทำซ้ำจำนวน 100 รอบ ในสถานการณ์จำลอง..... 154

ตารางที่ 4.17 ผลการเปรียบเทียบขนาดอิทธิพลของปฏิสัมพันธ์ร่วมระหว่างปัจจัยที่มีต่อค่าเฉลี่ยดัชนีการจำแนกประเภทจำแนกตามวิธีการประมาณค่า 156

ตารางที่ 4.18 ค่าสถิติพื้นฐานของคะแนนผลการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET) 158

ตารางที่ 4.19 ผลการวิเคราะห์ค่าสถิติของข้อสอบ 25 ข้อ จากแบบสอบวิชาคณิตศาสตร์ ชุด A (n=2,000) จำแนกตามโมเดลการวัด 160

ตารางที่ 4.20 ผลการวิเคราะห์ค่าสถิติของข้อสอบ 25 ข้อ จากแบบสอบรายวิชาคณิตศาสตร์ ชุด B (n=2,000) จำแนกตามโมเดลการวัด 163

ตารางที่ 4.21 ผลการวิเคราะห์ค่าสถิติของข้อสอบ 50 ข้อ จากแบบสอบรายวิชาภาษาไทย ชุด A (n=2,000) จำแนกตามโมเดลการวัด 166

ตารางที่ 4.22 ผลการวิเคราะห์ค่าสถิติของข้อสอบ 50 ข้อ จากแบบสอบรายวิชาภาษาไทย ชุด B (n=2,000) จำแนกตามโมเดลการวัด 170

ตารางที่ 4.23 ค่าสถิติพื้นฐานของค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) และค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (se) ของคะแนนผลการทดสอบรายวิชาคณิตศาสตร์ ชุด A 173

ตารางที่ 4.24 ค่าสถิติพื้นฐานของค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) และค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (se) ของคะแนนผลการทดสอบรายวิชาคณิตศาสตร์ ชุด B 175

ตารางที่ 4.25 ค่าสถิติพื้นฐานของค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) และค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (se) ของคะแนนผลการทดสอบรายวิชาภาษาไทย ชุด A 176

ตารางที่ 4.26 ค่าสถิติพื้นฐานของค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) และค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (se) ของคะแนนผลการทดสอบรายวิชาภาษาไทย ชุด B 177

ตารางที่ 4.27 ลักษณะของแบบสอบจำแนกตามข้อตกลงเบื้องต้น 179

ตารางที่ 4.28 ค่าดัชนีความถูกต้องและความสอดคล้องของการจำแนกประเภทในสถานการณ์จริงจำแนกตามวิธีการประมาณค่า 181

ตารางที่ 4.29 ค่าเฉลี่ยดัชนีความถูกต้องและความสอดคล้องของการจำแนกประเภทจากการทำซ้ำจำนวน 100 รอบ ในสถานการณ์จริงจำแนกตามวิธีการประมาณค่า 184

ตารางที่ 4.30 ผลการวิเคราะห์ความแตกต่างของค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภท (classification accuracy) จากการทำซ้ำจำนวน 100 รอบ ในสถานการณ์จริงจำแนกตามวิธีการประมาณค่า 188

ตารางที่ 4.31 ผลการวิเคราะห์ความแตกต่างของค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภท (classification consistency) จากการทำซ้ำจำนวน 100 รอบ ในสถานการณ์จริงจำแนกตามวิธีการประมาณค่า 189

สารบัญภาพ

หน้า

ภาพที่ 2.1 ความผิดพลาดเชิงบวกสำหรับผู้สอบที่ระดับความสามารถหนึ่ง 49

ภาพที่ 2.2 กรอบแนวคิดในการวิจัย 77

ภาพที่ 4.1 ฟังก์ชันสารสนเทศของแบบสอบของข้อมูลจำลองชุดที่ 1 (251PL10)..... 106

ภาพที่ 4.2 ฟังก์ชันการตอบสนองข้อสอบของข้อมูลจำลองชุดที่ 1 (251PL10)..... 106

ภาพที่ 4.3 ฟังก์ชันสารสนเทศของแบบสอบของข้อมูลจำลองชุดที่ 2 (251PL20)..... 107

ภาพที่ 4.4 ฟังก์ชันการตอบสนองข้อสอบของข้อมูลจำลองชุดที่ 2 (251PL20)..... 107

ภาพที่ 4.5 ฟังก์ชันสารสนเทศของแบบสอบของข้อมูลจำลองชุดที่ 3 (252PL10)..... 109

ภาพที่ 4.6 ฟังก์ชันการตอบสนองข้อสอบของข้อมูลจำลองชุดที่ 3 (252PL10)..... 110

ภาพที่ 4.7 ฟังก์ชันสารสนเทศของแบบสอบของข้อมูลจำลองชุดที่ 4 (252PL20)..... 110

ภาพที่ 4.8 ฟังก์ชันการตอบสนองข้อสอบของข้อมูลจำลองชุดที่ 4 (252PL20)..... 111

ภาพที่ 4.9 ฟังก์ชันสารสนเทศของแบบสอบของข้อมูลจำลองชุดที่ 5 (253PL10)..... 113

ภาพที่ 4.10 ฟังก์ชันสารสนเทศของแบบสอบและฟังก์ชันการตอบสนองข้อสอบของข้อมูลจำลอง 113

ภาพที่ 4.11 ฟังก์ชันสารสนเทศของแบบสอบของข้อมูลจำลองชุดที่ 6 (253PL20)..... 114

ภาพที่ 4.12 ฟังก์ชันการตอบสนองข้อสอบของข้อมูลจำลองชุดที่ 6 (253PL20)..... 114

ภาพที่ 4.13 ฟังก์ชันสารสนเทศของแบบสอบของข้อมูลจำลองชุดที่ 7 (501PL10)..... 117

ภาพที่ 4.14 ฟังก์ชันการตอบสนองข้อสอบของข้อมูลจำลองชุดที่ 7 (501PL10)..... 118

ภาพที่ 4.15 ฟังก์ชันสารสนเทศของแบบสอบของข้อมูลจำลองชุดที่ 8 (501PL20)..... 118

ภาพที่ 4.16 ฟังก์ชันการตอบสนองข้อสอบของข้อมูลจำลองชุดที่ 8 (501PL20)..... 119

ภาพที่ 4.17 ฟังก์ชันสารสนเทศของแบบสอบของข้อมูลจำลองชุดที่ 9 (502PL10)..... 122

ภาพที่ 4.18 ฟังก์ชันการตอบสนองข้อสอบของข้อมูลจำลองชุดที่ 9 (502PL10)..... 122

ภาพที่ 4.19 ฟังก์ชันสารสนเทศของแบบสอบของข้อมูลจำลองชุดที่ 10 (502PL20) 123

ภาพที่ 4.20 ฟังก์ชันการตอบสนองข้อสอบของข้อมูลจำลองชุดที่ 10 (502PL20) 123
ภาพที่ 4.21 ฟังก์ชันสารสนเทศของแบบสอบของข้อมูลจำลองชุดที่ 11 (503PL10) 126
ภาพที่ 4.22 ฟังก์ชันการตอบสนองข้อสอบของข้อมูลจำลองชุดที่ 11 (503PL10) 127
ภาพที่ 4.23 ฟังก์ชันสารสนเทศของแบบสอบของข้อมูลจำลองชุดที่ 12 (503PL20) 127
ภาพที่ 4.24 ฟังก์ชันการตอบสนองข้อสอบของข้อมูลจำลองชุดที่ 12 (503PL20) 128
ภาพที่ 4.25 ลักษณะการแจกแจงค่าความสามารถผู้สอบของข้อมูลจำลองชุดที่ 1 (251PL10) 130
ภาพที่ 4.26 ลักษณะการแจกแจงค่าความสามารถผู้สอบของข้อมูลจำลองชุดที่ 2 (251PL20) 130
ภาพที่ 4.27 ลักษณะการแจกแจงค่าความสามารถผู้สอบของข้อมูลจำลองชุดที่ 3 (252PL10) 130
ภาพที่ 4.28 ลักษณะการแจกแจงค่าความสามารถผู้สอบของข้อมูลจำลองชุดที่ 4 (252PL20) 131
ภาพที่ 4.29 ลักษณะการแจกแจงค่าความสามารถผู้สอบของข้อมูลจำลองชุดที่ 5 (253PL10) 131
ภาพที่ 4.30 ลักษณะการแจกแจงค่าความสามารถผู้สอบของข้อมูลจำลองชุดที่ 6 (253PL20) 131
ภาพที่ 4.31 ลักษณะการแจกแจงค่าความสามารถผู้สอบของข้อมูลจำลองชุดที่ 7 (501PL10) 132
ภาพที่ 4.32 ลักษณะการแจกแจงค่าความสามารถผู้สอบของข้อมูลจำลองชุดที่ 8 (501PL20) 132
ภาพที่ 4.33 ลักษณะการแจกแจงค่าความสามารถผู้สอบของข้อมูลจำลองชุดที่ 9 (502PL10) 132
ภาพที่ 4.34 ลักษณะการแจกแจงค่าความสามารถผู้สอบของข้อมูลจำลองชุดที่ 10 (502PL20) ... 133
ภาพที่ 4.35 ลักษณะการแจกแจงค่าความสามารถผู้สอบของข้อมูลจำลองชุดที่ 11 (503PL10) ... 133
ภาพที่ 4.36 ลักษณะการแจกแจงค่าความสามารถผู้สอบของข้อมูลจำลองชุดที่ 12 (503PL20) ... 133
ภาพที่ 4.37 ค่าเฉลี่ยดัชนีการจำแนกประเภทของข้อมูลจำลองจากการทำซ้ำ 100 รอบ..... 139
ภาพที่ 4.38 ค่าเฉลี่ยดัชนีการจำแนกประเภทของข้อมูลจำลองจากการทำซ้ำ 100 รอบ..... 139
ภาพที่ 4.39 ค่าเฉลี่ยดัชนีการจำแนกประเภทของข้อมูลจำลองจากการทำซ้ำ 100 รอบ..... 140
ภาพที่ 4.40 กราฟเปรียบเทียบค่าเฉลี่ยดัชนีการจำแนกประเภทจากการทำซ้ำ 100 รอบ..... 147
ภาพที่ 4.41 กราฟเปรียบเทียบค่าเฉลี่ยดัชนีการจำแนกประเภทจากการทำซ้ำ 100 รอบ..... 148
ภาพที่ 4.42 กราฟเปรียบเทียบค่าเฉลี่ยดัชนีการจำแนกประเภทจากการทำซ้ำ 100 รอบ 149

ภาพที่ 4.43 กราฟเปรียบเทียบค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทจากการทำซ้ำ 100 รอบ ในสถานการณ์จำลองด้วยวิธีการประมาณค่า 3 วิธี 151

ภาพที่ 4.44 กราฟเปรียบเทียบค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทจากการทำซ้ำ 100 รอบ ในสถานการณ์จำลองด้วยวิธีการประมาณค่า 3 วิธี 153

ภาพที่ 4.45 โค้งสารสนเทศและความคลาดเคลื่อนมาตรฐานของแบบสอบวิชาคณิตศาสตร์ชุด A. 161

ภาพที่ 4.46 โค้งสารสนเทศและความคลาดเคลื่อนมาตรฐานของแบบสอบวิชาคณิตศาสตร์ชุด B. 164

ภาพที่ 4.47 โค้งสารสนเทศและความคลาดเคลื่อนมาตรฐานของแบบสอบวิชาภาษาไทยชุด A 168

ภาพที่ 4.48 โค้งสารสนเทศและความคลาดเคลื่อนมาตรฐานของแบบสอบวิชาภาษาไทยชุด B 172

ภาพที่ 4.49 ฮิสโตแกรมและ Normal P-P Plot ของค่าพารามิเตอร์ความสามารถของผู้สอบ..... 174

ภาพที่ 4.50 ฮิสโตแกรมและ Normal P-P Plot ของค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (se) รายวิชาคณิตศาสตร์ ชุด A..... 174

ภาพที่ 4.51 ฮิสโตแกรมและ Normal P-P Plot ของค่าพารามิเตอร์ความสามารถของผู้สอบ..... 175

ภาพที่ 4.52 ฮิสโตแกรมและ Normal P-P Plot ของค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (se) รายวิชาคณิตศาสตร์ ชุด B..... 175

ภาพที่ 4.53 ฮิสโตแกรมและ Normal P-P Plot ของค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) รายวิชาภาษาไทย ชุด A..... 176

ภาพที่ 4.54 ฮิสโตแกรมและ Normal P-P Plot ของค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (se) รายวิชาภาษาไทย ชุด A..... 177

ภาพที่ 4.55 ฮิสโตแกรมและ Normal P-P Plot ของค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) รายวิชาภาษาไทย ชุด B..... 178

ภาพที่ 4.56 ฮิสโตแกรมและ Normal P-P Plot ของค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (se) รายวิชาภาษาไทย ชุด B..... 178

ภาพที่ 4.57 ค่าเฉลี่ยดัชนีการจำแนกประเภทของข้อมูลจริงจากการทำซ้ำ 100 รอบ 186

ภาพที่ 4.58 ค่าเฉลี่ยดัชนีการจำแนกประเภทของข้อมูลจริงจากการทำซ้ำ 100 รอบ 186

ภาพที่ 4.59 ค่าเฉลี่ยดัชนีการจำแนกประเภทของข้อมูลจริงจากการทำซ้ำ 100 รอบ 187

ภาพที่ 4.60 กราฟเปรียบเทียบค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทของข้อมูลจริง .. 189

ภาพที่ 4.61 กราฟเปรียบเทียบค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทของข้อมูลจริง .. 190

บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

การทดสอบทางการศึกษานับว่าเป็นกระบวนการที่สำคัญกระบวนการหนึ่งในการตัดสินใจทางการศึกษา ไม่ว่าจะเป็นการทดสอบขนาดใหญ่ (large-scale testing) ที่มีวัตถุประสงค์ของการทดสอบเพื่อให้ได้มาซึ่งคะแนนที่เชื่อถือได้ สามารถนำไปตีความได้อย่างสมเหตุสมผล และเพื่อนำคะแนนที่ได้ไปสร้างการตัดสินใจเกี่ยวกับการจัดตำแหน่ง (placement) การปรับปรุงแก้ไข (remediation) และการรับรองผล (certification) (Wainer & Kiely, 1987) หรือจะเป็นการทดสอบตามหลักสูตรที่จัดขึ้นภายในสถานศึกษา เพื่อให้สอดคล้องกับการจัดการเรียนการสอนแบบอิงวัตถุประสงค์ตามหลักสูตร (objective-based instructional) โดยมีวัตถุประสงค์ของการทดสอบเพื่อจัดนักเรียนเข้าสู่กลุ่มสมรรถภาพหรือกลุ่มรอบรู้ตามแต่ละจุดประสงค์ของหลักสูตรการเรียนการสอน และนำข้อมูลที่ได้ไปใช้ในการกำกับ ติดตาม และแก้ไขพฤติกรรมการเรียนรู้ของนักเรียน (Swaminathan, Hambleton, & Algina, 1974) ซึ่งการตัดสินใจทางการศึกษาที่กล่าวมาข้างต้นนั้นต้องนำคะแนนที่ได้จากการทดสอบมาทำการเปรียบเทียบกับเกณฑ์ตามวัตถุประสงค์ของการทดสอบ หรือกล่าวอีกนัยหนึ่งว่าคะแนนจุดตัด ตัวอย่างเช่น การตัดสินการสอบผ่านหรือตกของนักเรียนที่ต้องนำคะแนนสอบของนักเรียนแต่ละคนมาเทียบกับเกณฑ์ที่กำหนดไว้ โดยพิจารณาว่าคะแนนของนักเรียนสูงกว่าหรือต่ำกว่าคะแนนเกณฑ์ที่กำหนด ถ้านักเรียนมีคะแนนสูงกว่าเกณฑ์นักเรียนคนนั้นจะได้รับการตัดสินว่าเป็นผู้มีความรอบรู้สามารถผ่านการทดสอบนี้ได้ แต่ในทางกลับกันถ้านักเรียนมีคะแนนต่ำกว่าเกณฑ์นักเรียนคนนั้นจะได้รับการตัดสินว่าเป็นผู้ไม่มีความรอบรู้พอที่จะผ่านการทดสอบ จากตัวอย่างจะเห็นได้ว่าการตัดสินนักเรียนโดยพิจารณาจากคะแนนที่ได้จากการทดสอบหรือคะแนนที่สังเกตได้เพียงอย่างเดียวอาจนำไปสู่ความคลาดเคลื่อนของการจำแนกประเภทอันเนื่องมาจากคะแนนที่ได้จากแบบสอบนั้นมีความคลาดเคลื่อนของการวัดเกิดขึ้น ซึ่งจะมีความคลาดเคลื่อนไปจากคะแนนที่แท้จริงของคุณลักษณะของบุคคลที่มุ่งวัด (Spearman, 1904 อ้างถึงใน ศิริชัย กาญจนวาสี, 2555)

ความคลาดเคลื่อนของการจำแนกประเภทสามารถแบ่งได้เป็นสองกรณี คือ กรณีแรกเป็นความคลาดเคลื่อนของการจำแนกประเภทที่เกิดขึ้นในสถานการณ์ของการบริหารจัดการการทดสอบสองสถานการณ์ที่คู่ขนานกัน ผู้บริหารจัดการการทดสอบต้องทำการตัดสินใจจำแนกหรือจัดผู้สอบเข้าสู่กลุ่มสมรรถภาพใดสมรรถภาพหนึ่ง โดยพิจารณาจากคะแนนที่ได้จากการทดสอบ โอกาสที่จะเกิดความคลาดเคลื่อนของการจำแนกประเภทสามารถเป็นไปได้สองรูปแบบ คือ โอกาสที่ผู้สอบได้รับ

การจำแนกเข้าสู่กลุ่มระดับสมรรถภาพที่สูงกว่ามาตรฐานในการทดสอบที่หนึ่ง แต่ได้รับการจำแนกเข้าสู่กลุ่มระดับสมรรถภาพที่ต่ำกว่ามาตรฐานในการทดสอบที่สอง และโอกาสที่ผู้สอบได้รับการจำแนกเข้าสู่กลุ่มระดับสมรรถภาพที่ต่ำกว่ามาตรฐานในการทดสอบครั้งที่หนึ่ง แต่ได้รับการจำแนกเข้าสู่กลุ่มระดับสมรรถภาพที่สูงกว่ามาตรฐานในการทดสอบครั้งที่สอง ความเป็นไปได้สองรูปแบบนี้เรียกว่า การจำแนกที่ไม่สอดคล้องกัน (inconsistent classification) แต่ถ้าผู้สอบได้รับการจำแนกจากทั้งสองสถานการณ์ให้อยู่ในกลุ่มสมรรถภาพเดียวกันไม่ว่าจะเป็นกลุ่มที่สูงกว่าหรือต่ำกว่ามาตรฐานก็ตาม เรียกว่าการจำแนกประเภทที่สอดคล้องกันหรือความสอดคล้องของการจำแนกประเภท (consistent classification) ซึ่งแสดงให้เห็นว่าไม่มีความคลาดเคลื่อนของการจำแนกประเภทเกิดขึ้น

กรณีที่สองเป็นความคลาดเคลื่อนของการจำแนกประเภทที่เกิดขึ้นในสถานการณ์การทดสอบทั่วไป ผู้บริหารจัดการการทดสอบต้องตัดสินใจเพื่อจำแนกหรือจัดผู้สอบเข้าสู่กลุ่มสมรรถภาพใดสมรรถภาพหนึ่ง โดยพิจารณาจากคะแนนที่ได้จากการทดสอบหรือคะแนนที่สังเกตได้และพิจารณาจากสถานะจริงของผู้สอบหรือคะแนนจริง โอกาสที่จะเกิดความคลาดเคลื่อนของการจำแนกประเภทสามารถเป็นไปได้สองรูปแบบ คือ โอกาสที่ผู้สอบได้รับการจำแนกเข้าสู่กลุ่มสมรรถภาพที่สูงกว่ามาตรฐานเมื่อพิจารณาจากคะแนนที่สังเกตได้ แต่ได้รับการจำแนกเข้าสู่กลุ่มสมรรถภาพที่ต่ำกว่ามาตรฐานเมื่อพิจารณาจากคะแนนจริง และโอกาสที่ผู้สอบได้รับการจำแนกเข้าสู่กลุ่มสมรรถภาพที่ต่ำกว่ามาตรฐานเมื่อพิจารณาจากคะแนนที่สังเกตได้ แต่ได้รับการจำแนกเข้าสู่กลุ่มสมรรถภาพที่สูงกว่ามาตรฐานเมื่อพิจารณาจากคะแนนจริง ความเป็นไปได้สองรูปแบบนี้เรียกว่า การจำแนกที่ผิดพลาด (misclassification) แต่ถ้าผู้สอบได้รับการจำแนกให้อยู่ในกลุ่มสมรรถภาพเดียวกัน เมื่อพิจารณาจากทั้งคะแนนที่สังเกตได้และคะแนนจริง ไม่ว่าจะเป็นกลุ่มที่สูงกว่าหรือต่ำกว่ามาตรฐานก็ตาม เรียกว่าการจำแนกประเภทที่ถูกต้องหรือความถูกต้องของการจำแนกประเภท (correct classification) ซึ่งแสดงให้เห็นว่าไม่มีความคลาดเคลื่อนของการจำแนกประเภทเกิดขึ้น

ความคลาดเคลื่อนของการจำแนกประเภทที่เกิดขึ้นทำให้นักวัดผลทางจิตวิทยาเกิดความตระหนักถึงความสำคัญของปัญหานี้ เนื่องด้วยการไม่ทราบขนาดของความคลาดเคลื่อนของการวัดที่เกิดขึ้นของผู้สอบแต่ละคนในแต่ละสถานการณ์การทดสอบ (Wainer & Kiely, 1987) และพบว่า การประมาณค่าความเที่ยงแบบดั้งเดิมซึ่งได้รับการพัฒนาบนพื้นฐานของคะแนนสอบอาจจะไม่เหมาะสมสำหรับการประเมินความสอดคล้องและความถูกต้องของการจำแนกประเภท ดังนั้นจึงจำเป็นที่จะต้องใช้เทคนิคในการประเมินความเที่ยงแบบใหม่ที่มีความเหมาะสมในการใช้มากกว่าและมีวัตถุประสงค์เพื่อทำให้ความคลาดเคลื่อนในการจำแนกประเภทเกิดขึ้นได้น้อยที่สุด โดยให้นิยามว่าดัชนีการจำแนกประเภท (classification indices) อันประกอบด้วย ดัชนีความสอดคล้องของการจำแนก

ประเภท (classification consistency) และดัชนีความถูกต้องของการจำแนกประเภท (classification accuracy) ซึ่งทำการพัฒนาควบคู่กันไปตั้งแต่ปี ค.ศ. 1970 เป็นต้นมา

ช่วงแรกของการพัฒนาวิธีการประมาณค่าดัชนีการจำแนกประเภท ส่วนใหญ่เป็นการพัฒนาภายใต้พื้นฐานของแนวคิดทฤษฎีการทดสอบแบบดั้งเดิม (CTT-based) โดยเริ่มจากการใช้โมเดลทวินาม (binomial model) ในการจำแนกในฐานะที่เป็นโมเดลของคะแนนจริง (true-score model) ที่แข็งแกร่ง เนื่องจากข้อตกลงเบื้องต้นที่สร้างขึ้นภายใต้โมเดลนี้มีความสัมพันธ์กับข้อตกลงเบื้องต้นของโมเดลการทดสอบแบบดั้งเดิม (CTT model) ที่กำหนดขึ้นสำหรับการแจกแจงของคะแนนที่สังเกตได้ของแบบสอบที่ประกอบด้วยข้อสอบแบบให้คะแนนได้สองค่าเท่านั้น (dichotomous items) หลังจากนั้นในปี ค.ศ. 1976 Huynh (1976) และ Subkoviak (1976) ได้พัฒนาวิธีการประมาณค่าขึ้นและเป็นวิธีการประมาณค่าที่เป็นที่นิยมใช้กันอย่างแพร่หลายในเวลาต่อมา ถัดมาในปี ค.ศ. 1995 Livingston และ Lewis (1995) ได้ทำการขยายโมเดลทวินาม (binomial model) เพื่อสามารถนำไปใช้ได้กับข้อสอบแบบให้คะแนนได้มากกว่าสองค่า (polytomous items) หลังจากนั้น Lee (2007) และ Lee, Brennan, และ Wan (2009) ได้เสนอวิธีการที่ใช้โมเดลอเนกนาม (multinomial model) และโมเดลอเนกนามเชิงซ้อน (compound multinomial model) ในการประมาณค่าสำหรับข้อสอบแบบให้คะแนนได้มากกว่าสองค่า (polytomous items)

ในช่วงเวลาเดียวกันวิธีการที่อยู่ภายใต้พื้นฐานของแนวคิดทฤษฎีการตอบสนองข้อสอบได้รับการพัฒนาขึ้นสำหรับแบบสอบที่มีการให้คะแนนเป็นสเกลของคะแนนดิบ (raw scores scale) การศึกษาในช่วงแรกซึ่งรวมถึง Huynh (1990) ใช้โมเดลของ Rasch (Rasch model) และ Wang, Kolen และ Harris (2000 cited in Lee, 2010) ใช้โมเดล polytomous IRT (polytomous IRT model) ในการประมาณค่า ต่อมาในปี ค.ศ. 2005 Rudner (2005) ได้พัฒนาวิธีการประมาณค่าโดยใช้ IRT model ในการคำนวณค่าความน่าจะเป็นเกี่ยวกับเวคเตอร์ของการตอบสนองแบบมีเงื่อนไขบนความสามารถแฝง (latent ability) ถัดจากนั้นไม่นาน Guo (2006) ได้พัฒนาวิธีการประมาณค่าด้วยวิธีการแจกแจงแฝง (latent distribution method) ซึ่งแตกต่างจากวิธีการของ Rudner ในเรื่องของข้อตกลงเบื้องต้นเกี่ยวกับการแจกแจงของความคลาดเคลื่อนของการประมาณค่าคะแนนจริง หลังจากนั้นในปี ค.ศ. 2010 Lee (2010) ได้พัฒนาวิธีการที่นำไปใช้สำหรับโมเดล mixture IRT (mixture IRT model)

สังเกตได้ว่าในช่วงแรกของการพัฒนาวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการทดสอบแบบดั้งเดิม โดยเฉพาะการประมาณค่าดัชนีความสอดคล้องของการจำแนกประเภทเป็นการประมาณค่าที่ต้องดำเนินการภายใต้การบริหารจัดการการทดสอบสองสถานการณ์ เนื่องจากการประมาณค่าความสอดคล้องของการจำแนกประเภทเป็นการเปรียบเทียบความเห็นพ้องต้องกันของการจำแนกผู้สอบระหว่างการทดสอบสองสถานการณ์ที่คู่ขนานกัน แต่ในสถานการณ์การทดสอบ

จริงนั้นเป็นไปได้ยากที่จะจัดการทดสอบสองครั้งที่ใช้แบบสอบคู่ขนานในภายหลังจึงมีนักวัดผลจำนวนหนึ่งพัฒนาวิธีการประมาณค่าดัชนีการจำแนกประเภทสำหรับการทดสอบเดียว (single administration) ขึ้นมา

การศึกษางานวิจัยทางด้านการวัดผลทางการศึกษาที่เกี่ยวข้องพบว่า งานวิจัยส่วนใหญ่เป็นการศึกษาที่นำวิธีการประมาณค่าดัชนีการจำแนกประเภทไปใช้เพื่อตรวจสอบหาค่าดัชนี รองลงมาคือการศึกษาเพื่อพัฒนาวิธีการประมาณค่าดัชนีการจำแนกประเภท การศึกษาเปรียบเทียบวิธีการประมาณค่าดัชนีการจำแนกประเภทต่างๆ และการศึกษาถึงปัจจัยที่ส่งผลต่อค่าดัชนีการจำแนกประเภท ตามลำดับ จากงานวิจัยทั้งหลายข้างต้นสังเกตเห็นได้ว่าการที่ผู้วิจัยจะตัดสินใจเลือกใช้วิธีการประมาณค่าแบบใดในการตรวจสอบความสอดคล้องและความถูกต้องของการจำแนกประเภทในเงื่อนไขที่แตกต่างกันของการทดสอบทางการศึกษานั้น ควรพิจารณาถึงประสิทธิภาพที่แตกต่างกันของแต่ละวิธีการภายใต้สถานการณ์การประยุกต์ใช้ ไม่ว่าจะเป็นเงื่อนไขเกี่ยวกับโมเดลการวัด ขนาดกลุ่มตัวอย่าง ตำแหน่งของคะแนนจุดตัด จำนวนตำแหน่งของคะแนนจุดตัด คุณลักษณะทางจิตมิติของแบบสอบ คุณลักษณะทางจิตมิติของข้อสอบ ความยาวของแบบสอบ จำนวนคุณลักษณะที่ต้องการจะวัด ความเป็นอิสระระหว่างคุณลักษณะที่ต้องการจะวัด รูปแบบการแจกแจงความสามารถของผู้สอบ ความสูงของมาตรฐานในการจำแนกประเภท สหสัมพันธ์ภายในข้อสอบ และความเป็นไปได้ของการชดเชย (Cui, Gierl & Chang, 2012; Hubregtse & Eggen, 2012; Lathrop & Cheng, 2013; Lee, 2010; MacCann & Stanley, 2010; Martineau, 2007; Suaklay, Lawthong & Kanjanawasee, 2016; Wyse, 2011) โดยปัจจัยเหล่านี้จะส่งผลให้ค่าดัชนีที่ประมาณค่าได้มีความแข็งแกร่ง มีประสิทธิภาพและมีความคลาดเคลื่อนเกิดขึ้นน้อยที่สุด

ดังนั้นการเปรียบเทียบค่าดัชนีที่ได้จากการประมาณค่าที่มีแนวคิดและทฤษฎีพื้นฐานที่แตกต่างกันเพื่อตรวจสอบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทเหล่านั้นจึงเป็นสิ่งสำคัญ เพื่อให้ได้มาซึ่งสารสนเทศที่เป็นประโยชน์ต่อการนำไปประยุกต์ใช้ในสถานการณ์การทดสอบที่หลากหลายได้อย่างเหมาะสม โดยเฉพาะอย่างยิ่งสำหรับการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET) ซึ่งเป็นการทดสอบขนาดใหญ่ที่มีผลได้ผลเสียสูง (high stakes test) และมีวิธีการประเมินที่ได้มาตรฐาน

นอกจากนี้การศึกษาแนวคิด ทฤษฎีพื้นฐาน วิธีการคำนวณ และข้อตกลงเบื้องต้นของวิธีการประมาณค่าดัชนีการจำแนกประเภทแต่ละวิธีแล้ว พบว่าวิธีการประมาณค่าที่มีแนวคิดพื้นฐานตามทฤษฎีการทดสอบแบบดั้งเดิม (CTT-based) นั้นยากต่อการนำไปประยุกต์ใช้ให้เหมาะสมกับสถานการณ์การทดสอบในปัจจุบันที่ส่วนใหญ่จัดขึ้นภายใต้แนวคิดของทฤษฎีการตอบสนองข้อสอบ ผู้วิจัยจึงสนใจที่จะศึกษาเฉพาะวิธีการประมาณค่าดัชนีการจำแนกประเภทที่มีแนวคิดพื้นฐานตามทฤษฎีการตอบสนองข้อสอบเท่านั้น และเนื่องจากวิธีการประมาณค่าดัชนีการจำแนกประเภทตาม

ทฤษฎีการตอบสนองข้อสอบนั้นพัฒนาขึ้นพร้อมกับความนิยมที่เพิ่มขึ้นของการใช้งานทฤษฎีการตอบสนองข้อสอบในด้านต่างๆ ของการปฏิบัติเกี่ยวกับการทดสอบ โดยพื้นฐานแล้วทุกวิธีการที่พัฒนาโดยใช้ทฤษฎีการตอบสนองข้อสอบเป็นฐานล้วนตั้งอยู่บนพื้นฐานของการใช้ข้อตกลงเบื้องต้นเดียวกันกับการประยุกต์ใช้ทฤษฎีการตอบสนองข้อสอบในด้านต่างๆ รวมถึงความเป็นเอกมิติ (unidimensionality) ความเป็นอิสระระหว่างข้อสอบ (local item independency) และความเหมาะสมของโมเดลการตอบสนองข้อสอบกับข้อมูล (model fit) (Lee, 2010)

เมื่อพิจารณาแนวคิดและทฤษฎีพื้นฐานของวิธีการประมาณค่าดัชนีการจำแนกประเภททั้งหมด พบว่า วิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบ (IRT) มีจำนวน 3 วิธีการ คือ วิธีการของ Rudner (2005) วิธีการของ Guo (2006) และ วิธีการของ Lee (2010) และจากการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง พบว่า วิธีการประมาณค่าทั้งสามวิธีนี้มีข้อตกลงเบื้องต้นเกี่ยวกับโมเดลตามทฤษฎีการตอบสนองข้อสอบ (IRT model) เหมือนกัน แต่ต่างกันในเรื่องของข้อตกลงเบื้องต้นเกี่ยวกับลักษณะการแจกแจง ลักษณะของข้อมูลที่ใช้ในการประมาณค่าและความน่าจะเป็นที่คาดหวังในการจำแนกผู้สอบเข้าสู่แต่ละระดับความสามารถ โดยข้อตกลงเบื้องต้นเกี่ยวกับลักษณะการแจกแจงนั้น วิธีการของ Rudner (2005) มีข้อตกลงเบื้องต้นเกี่ยวกับการแจกแจงของความคลาดเคลื่อนของการประมาณค่าความสามารถของผู้สอบว่าต้องเป็นการแจกแจงแบบปกติ (normal distribution) และวิธีการของ Lee (2010) มีข้อตกลงเบื้องต้นเกี่ยวกับการแจกแจงของคะแนนสอบว่าต้องเป็นการแจกแจงแบบอเนกนามเชิงซ้อน (compound binomial distribution) ส่วนวิธีการของ Guo (2006) ไม่มีข้อตกลงเบื้องต้นเกี่ยวกับการแจกแจงแต่จะใช้การแจกแจงภายหลังของฟังก์ชันความน่าจะเป็นของรูปแบบการตอบข้อสอบในการประมาณค่า

ด้านลักษณะของข้อมูลหรือคะแนนที่ใช้ในการประมาณค่านั้นมีความแตกต่างกัน คือ วิธีการของ Rudner (2005) และวิธีการของ Guo (2006) ใช้ข้อมูลในลักษณะเดียวกันคือใช้คะแนนในสเกลของคะแนนความสามารถหรือคุณลักษณะแฝง (theta scale) ส่วนวิธีการของ Lee (2010) ใช้คะแนนที่อยู่บนสเกลของคะแนนรวม (summed score scale) สำหรับด้านความน่าจะเป็นที่คาดหวังในการจำแนกผู้สอบเข้าสู่แต่ละระดับความสามารถนั้น แต่ละวิธีมีความแตกต่างกันคือวิธีการของ Lee (2010) คะแนนสอบในการหาความน่าจะเป็นที่คาดหวังในการจำแนกผู้สอบ วิธีการของ Rudner (2005) ใช้พื้นที่ใต้โค้งปกติของความคลาดเคลื่อนมาตรฐานในการประมาณค่าความสามารถของผู้สอบในการหาความน่าจะเป็น ส่วนวิธีการของ Guo (2006) ใช้พื้นที่ใต้โค้งปกติของฟังก์ชันความน่าจะเป็นในการตอบข้อสอบของผู้สอบในการประมาณค่าความน่าจะเป็นที่คาดหวัง เนื่องด้วยความแตกต่างเหล่านี้ทำให้ผู้วิจัยเลือกที่จะศึกษาถึงประสิทธิภาพในการประมาณค่าดัชนีการจำแนกประเภทของวิธีการประมาณค่าตามทฤษฎีการตอบสนองข้อสอบทั้งสามวิธี

สำหรับตัวแปรที่เป็นเงื่อนไขในการศึกษาประสิทธิภาพของการประมาณค่าดัชนีการจำแนกประเภท จากการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องพบว่าได้มีการศึกษาเกี่ยวกับตัวแปรที่น่าจะส่งผลต่อค่าดัชนีการจำแนกประเภทไว้ทั้งสิ้น 12 ตัวแปร คือ โมเดลการวัด ขนาดกลุ่มตัวอย่าง ตำแหน่งของคะแนนจุดตัด จำนวนจุดตัด คุณลักษณะทางจิตมิติของแบบสอบ คุณลักษณะทางจิตมิติของข้อสอบ ความยาวของแบบสอบ จำนวนคุณลักษณะที่ต้องการจะวัด ความเป็นอิสระระหว่างคุณลักษณะที่ต้องการจะวัด รูปแบบการแจกแจงความสามารถของผู้สอบ ความสูงของมาตรฐานในการจำแนกประเภท สหสัมพันธ์ภายในข้อสอบ และความเป็นไปได้ของการชดเชย (Cui, Gierl & Chang, 2012; Hubregtse & Eggen, 2012; Lathrop & Cheng, 2013; Lee, 2010; MacCann & Stanley, 2010; Martineau, 2007; Suaklay, Lawthong & Kanjanawasee, 2016; Wyse, 2011) ซึ่งในงานวิจัยที่ผ่านมาได้ทำการศึกษาตัวแปรเหล่านี้ภายใต้เงื่อนไข สถานการณ์ และการใช้วิธีการประมาณค่าดัชนีการจำแนกประเภทที่แตกต่างกันไป โดยในการวิจัยครั้งนี้ผู้วิจัยเลือกตัวแปรที่เป็นเงื่อนไขในการศึกษาจำนวน 3 ตัวแปร คือ โมเดลการวัด ความยาวของแบบสอบ และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ ส่วนตัวแปรอื่นจะควบคุมให้อยู่ในสถานการณ์มาตรฐานทั่วไป เนื่องจากทั้งสามตัวแปรนี้เป็นปัจจัยที่มีความสำคัญต่อการประมาณค่าดัชนี ถึงแม้ว่าตัวแปรความยาวของแบบสอบนั้นพบว่า มีงานวิจัยของ Lathrop และ Cheng (2013) เพียงเรื่องเดียวที่ทำการศึกษาในเรื่องนี้ ซึ่งก็ยังไม่ได้ทำการศึกษาที่ครอบคลุมวิธีการประมาณค่าดัชนีการจำแนกประเภทครบทั้งสามวิธีที่ใช้ในการศึกษาครั้งนี้ และเนื่องจากในการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET) ซึ่งเป็นการทดสอบที่ได้มาตรฐาน มีการกำหนดความยาวของแบบสอบที่ใช้ในแต่ละวิชาอยู่ในช่วง 25 ถึง 50 ข้อ ผู้วิจัยจึงนำจำนวนข้อที่น้อยที่สุดและมากที่สุดมากำหนดเป็นเงื่อนไขในการศึกษาจำลองครั้งนี้ โดยในการวิจัยครั้งนี้กำหนดความยาวของแบบสอบที่ใช้เป็นเงื่อนไขในการจำลองข้อมูลจำนวน 2 เงื่อนไข คือ แบบสอบที่มีความยาว 25 ข้อ และแบบสอบที่มีความยาว 50 ข้อ

สำหรับตัวแปรโมเดลการวัดนั้นจะมีงานวิจัยที่ทำการศึกษาไว้จำนวนหนึ่ง เช่นงานวิจัยของ Lathrop และ Cheng (2013), Lee (2010) และ Zhang (2008, 2010) แต่ก็ยังมีไม่มากนัก และงานวิจัยเหล่านี้ก็ยังไม่ได้ทำการศึกษาที่ครอบคลุมวิธีการประมาณค่าดัชนีการจำแนกประเภทครบทั้งสามวิธีที่ใช้ในการศึกษาครั้งนี้ และเนื่องจากโมเดลการวัดภายใต้โมเดลการตอบสนองข้อสอบแบบสองค่ามีทั้งสิ้น 3 โมเดล คือ โมเดลโลจิสติกแบบหนึ่ง สอง และสามพารามิเตอร์ โดยในการวิจัยครั้งนี้ได้ทำการกำหนดโมเดลการวัดที่ใช้เป็นเงื่อนไขในการจำลองข้อมูลจำนวน 3 โมเดล คือ โมเดลโลจิสติกแบบหนึ่งพารามิเตอร์ (one-parameter logistic model: 1PL) โมเดลโลจิสติกแบบสองพารามิเตอร์ (two-parameter logistic model: 2PL) และโมเดลโลจิสติกแบบสามพารามิเตอร์ (three-parameter logistic model: 3PL) ส่วนตัวแปรความไม่เหมาะสมของโมเดลการวัดกับข้อสอบนั้นพบว่า ยังไม่มีงานวิจัยใดที่ทำการศึกษาเกี่ยวกับเรื่องนี้ แต่เนื่องจากในสถานการณ์การทดสอบทั่วไป

นั้นมีโอกาสที่จะเกิดความไม่เหมาะสมของโมเดลการวัดกับข้อสอบได้สูง ดังนั้นจึงมีความจำเป็นที่ต้องพิจารณาถึงปัจจัยสำคัญนี้ด้วย และเนื่องจากในสถานการณ์การทดสอบที่ได้มาตรฐานนั้น ระดับความไม่เหมาะสมของโมเดลการวัดกับข้อสอบที่ยอมรับได้ไม่ควรเกินร้อยละ 20 ผู้วิจัยจึงนำระดับความไม่เหมาะสมของโมเดลการวัดกับข้อสอบมากำหนดเป็นเงื่อนไขในการศึกษาจำลองครั้งนี้ โดยในการวิจัยครั้งนี้กำหนดความไม่เหมาะสมของโมเดลการวัดกับข้อสอบที่ใช้เป็นเงื่อนไขในการจำลองข้อมูลจำนวน 2 เงื่อนไข คือ แบบสอบที่มีความไม่เหมาะสมของโมเดลการวัดกับข้อสอบร้อยละ 10 และแบบสอบที่มีความไม่เหมาะสมของโมเดลการวัดกับข้อสอบร้อยละ 20

ด้วยเหตุผลดังที่กล่าวมาข้างต้น ผู้วิจัยจึงมีความสนใจที่จะศึกษาประสิทธิภาพของวิธีการประมาณค่าตามทฤษฎีการตอบสนองข้อสอบภายใต้การศึกษาจำลอง (simulation study) ที่มีเงื่อนไขด้านตัวแปรต้นที่แตกต่างกัน เพื่อให้ได้สารสนเทศเกี่ยวกับวิธีการประมาณค่าดัชนีการจำแนกประเภทที่ให้ผลการประมาณค่าที่มีประสิทธิภาพดีที่สุดภายใต้สถานการณ์หรือเงื่อนไขดังกล่าว ซึ่งจะเป็นประโยชน์ต่อการเลือกใช้วิธีการประมาณค่าดัชนีการจำแนกประเภทที่เหมาะสมทั้งดัชนีความสอดคล้องและดัชนีความถูกต้องของการจำแนกประเภท สำหรับสถานการณ์การทดสอบที่มีผลได้ผลเสียสูงต่อไปไม่ว่าจะเป็นการทดสอบระดับชาติหรือระดับชั้นเรียนก็ตาม

คำถามการวิจัย

1. เมื่อกำหนดเงื่อนไขในการประมาณค่าที่แตกต่างกันในการศึกษาจำลอง วิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบสามวิธี คือ วิธีการของ Rudner (2005) วิธีการของ Guo (2006) และวิธีการของ Lee (2010) จะให้ผลการประมาณค่าดัชนีที่แตกต่างกันหรือไม่ อย่างไร ทั้งดัชนีความสอดคล้องของการจำแนกประเภทและดัชนีความถูกต้องของการจำแนกประเภท

2. เมื่อกำหนดเงื่อนไขในการประมาณค่าที่แตกต่างกันในการศึกษาจำลองสำหรับวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบทั้งสามวิธี วิธีการใดมีประสิทธิภาพในการประมาณค่าดัชนีการจำแนกประเภทสูงที่สุด ทั้งดัชนีความสอดคล้องของ การจำแนกประเภทและดัชนีความถูกต้องของการจำแนกประเภท

3. เมื่อนำวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบสามวิธีไปใช้กับข้อมูลเชิงประจักษ์ ผลการประมาณค่าดัชนีการจำแนกประเภทที่ได้จะมีลักษณะอย่างไร ทั้งดัชนีความสอดคล้องของการจำแนกประเภทและดัชนีความถูกต้องของการจำแนกประเภท

วัตถุประสงค์การวิจัย

1. เพื่อประมาณค่าดัชนีการจำแนกประเภทโดยใช้วิธีการประมาณค่าตามทฤษฎีการตอบสนองข้อสอบสามวิธี ได้แก่ วิธีการของ Rudner (2005) วิธีการของ Guo (2006) และวิธีการของ Lee (2010) จากการจำลองข้อมูลภายใต้เงื่อนไขของการศึกษา

2. เพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบสามวิธี ได้แก่ วิธีการของ Rudner (2005) วิธีการของ Guo (2006) และวิธีการของ Lee (2010) จากการจำลองข้อมูลภายใต้เงื่อนไขของการศึกษา

3. เพื่อประมาณค่าและเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบทั้งสามวิธี เมื่อใช้กับข้อมูลเชิงประจักษ์

สมมติฐานการวิจัย

การศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องกับดัชนีการจำแนกประเภท ทำให้สามารถตั้งสมมติฐานการวิจัยตามวัตถุประสงค์การวิจัยข้อที่สองเพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบสามวิธี คือ วิธีการของ Rudner (2005) วิธีการของ Guo (2006) และวิธีการของ Lee (2010) ภายใต้เงื่อนไขของการศึกษา โดยมีสมมติฐานการวิจัยคือ **วิธีการประมาณค่าดัชนีการจำแนกประเภทของ Guo (2006) น่าจะมีประสิทธิภาพในการประมาณค่ามากที่สุด รองลงมาน่าจะเป็นวิธีการของ Rudner (2005) และวิธีการของ Lee (2010) ตามลำดับ**

ทั้งนี้เนื่องจากงานวิจัยของ Guo (2006) ที่ได้ทำการศึกษาความถูกต้องของการจำแนกประเภทที่คาดหวังโดยใช้การแจกแจงแฝง (latent distribution) ซึ่งในการศึกษาครั้งนี้ Guo ได้พัฒนาวิธีการประมาณค่าด้วยการแจกแจงแฝง (latent distribution method) โดยทำการศึกษาเปรียบเทียบกับวิธีการประมาณค่าด้วยการแจกแจงของความคลาดเคลื่อนมาตรฐานของการประมาณค่าความสามารถของผู้สอบที่ Rudner พัฒนาขึ้นในปี ค.ศ. 2001 และปรับปรุงเพิ่มเติมในปี ค.ศ. 2005 พบว่า วิธีการของ Guo แตกต่างจากวิธีการของ Rudner ตรงการคำนวณจำนวนผู้สอบที่คาดหวังในแต่ละกลุ่มความสามารถด้วยการแจกแจงภายหลังของผู้สอบ (the normalized likelihood function) เป็นผลให้ข้อตกลงเบื้องต้นของวิธีการของ Rudner เกี่ยวกับการแจกแจงปกติของความคลาดเคลื่อนมาตรฐานในการประมาณค่าความสามารถของผู้สอบไม่มีความจำเป็น ดังนั้น latent distribution method อาจจะเป็นวิธีการที่แข็งแกร่งกว่า

และงานวิจัยของ Lathrop และ Cheng (2013) ที่ได้ทำการเปรียบเทียบวิธีการประมาณค่าดัชนีความถูกต้องของการจำแนกประเภท (classification accuracy) สองวิธีตามทฤษฎีการตอบสนอง

ข้อสอบ คือวิธีการของ Lee ซึ่งประมาณค่าความถูกต้องของการจำแนกประเภทด้วยคะแนนรวมทั้งหมด (total sum scores) และวิธีการของ Rudner ประมาณค่าด้วยคุณลักษณะแฝง (latent trait estimates) โดยได้ดำเนินการศึกษาจำลองภายใต้เงื่อนไขของโมเดลตามทฤษฎีการตอบสนองข้อสอบ (IRT model) ขนาดตัวอย่าง (sample size) ความยาวของแบบสอบ (test length) และตำแหน่งของคะแนนจุดตัด (cut score location) ผลการวิจัยพบว่า เมื่อโมเดลมีความเหมาะสมกับข้อมูล การจำแนกประเภทที่ได้จากการประมาณค่าคุณลักษณะแฝง (latent trait estimates) จะมีความถูกต้องพอกันหรือมากกว่าการจำแนกประเภทที่ได้จากการประมาณค่าด้วยคะแนนรวม (total score)

จากทั้งสองงานวิจัยนี้จะเห็นได้ว่าวิธีการของ Rudner น่าจะให้ค่าดัชนีที่สูงกว่าวิธีการของ Lee และวิธีการของ Guo ก็น่าจะให้ค่าดัชนีที่สูงกว่าวิธีการของ Rudner จึงทำให้สามารถตั้งสมมติฐานได้ว่า “วิธีการประมาณค่าดัชนีการจำแนกประเภทของ Guo (2006) น่าจะมีประสิทธิภาพในการประมาณค่ามากที่สุด รองลงมาน่าจะเป็นวิธีการของ Rudner (2005) และวิธีการของ Lee (2010) ตามลำดับ”

ขอบเขตการวิจัย

1. การวิจัยครั้งนี้เป็นการศึกษาจำลอง (simulation study) เพื่อศึกษาประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภท โดยในการศึกษาครั้งนี้ประกอบด้วยวิธีการประมาณค่าทั้งสิ้น 3 วิธี ซึ่งเป็นวิธีการตามทฤษฎีการตอบสนองข้อสอบทั้งหมด และสามารถใช้ประมาณค่าได้ทั้งดัชนีความสอดคล้องของการจำแนกประเภทและดัชนีความถูกต้องของการจำแนกประเภท ได้แก่

1.1 วิธีการที่พัฒนาขึ้นโดย Rudner (2005)

1.2 วิธีการที่พัฒนาขึ้นโดย Guo (2006)

1.3 วิธีการที่พัฒนาขึ้นโดย Lee (2010)

2. การจำลองข้อมูลในการศึกษาครั้งนี้มีตัวแปรที่เป็นเงื่อนไขในการจำลองข้อมูลทั้งสิ้น 3 ตัวแปร ส่วนตัวแปรอื่นจะควบคุมให้อยู่ในสถานการณ์มาตรฐานของการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET) โดยตัวแปรที่เป็นเงื่อนไขในการจำลองข้อมูล ได้แก่

2.1 ความยาวของแบบสอบ สำหรับตัวแปรความยาวของแบบสอบพบว่ามีงานวิจัยของ Lathrop และ Cheng (2013) เพียงเรื่องเดียวที่ทำการศึกษาในเรื่องนี้ ซึ่งก็ยังไม่ได้ทำการศึกษาที่ครอบคลุมวิธีการประมาณค่าดัชนีการจำแนกประเภทครบทั้งสามวิธีที่ใช้ในการศึกษาครั้งนี้ และเนื่องจากในการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET) ซึ่งเป็นการทดสอบที่ได้มาตรฐานมีการกำหนดความยาวของแบบสอบที่ใช้ในแต่ละวิชาอยู่ในช่วง 25 ถึง 50 ข้อ ผู้วิจัยจึงนำจำนวนข้อที่น้อยที่สุดและมากที่สุดมากำหนดเป็นเงื่อนไขในการศึกษาจำลองครั้งนี้ โดยในการวิจัยครั้งนี้กำหนดความยาวของแบบสอบที่ใช้เป็นเงื่อนไขในการจำลองข้อมูลจำนวน 2 เงื่อนไข ได้แก่

2.2.1 แบบสอบสั้น (25 ข้อ)

2.2.2 แบบสอบยาว (50 ข้อ)

2.2 โมเดลการวัด สำหรับตัวแปรโมเดลการวัดนี้มีงานวิจัยที่ทำการศึกษาไว้จำนวนหนึ่ง เช่นงานวิจัยของ Lathrop และ Cheng (2013), Lee (2010) และ Zhang (2008, 2010) แต่ก็ยังมีไม่มากนัก และงานวิจัยเหล่านี้ก็ยังไม่ได้ทำการศึกษาที่ครอบคลุมวิธีการประมาณค่าดัชนีการจำแนกประเภทครบทั้งสามวิธีที่ใช้ในการศึกษาครั้งนี้ และเนื่องจากโมเดลการวัดภายใต้โมเดลการตอบสนองข้อสอบแบบสองค่ามีทั้งสิ้น 3 โมเดล คือ โมเดลโลจิสติกแบบหนึ่ง สอง และสามพารามิเตอร์ ดังนั้นในการวิจัยครั้งนี้จึงกำหนดโมเดลการวัดที่ใช้เป็นเงื่อนไขในการจำลองข้อมูลจำนวน 3 โมเดล ได้แก่

2.1.1 โมเดลโลจิสติกแบบหนึ่งพารามิเตอร์ (one-parameter logistic model: 1PL)

2.1.2 โมเดลโลจิสติกแบบสองพารามิเตอร์ (two-parameter logistic model: 2PL)

2.1.3 โมเดลโลจิสติกแบบสามพารามิเตอร์ (three-parameter logistic model: 3PL)

2.3 ความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ จากงานวิจัยที่ทำการศึกษาพบว่ายังไม่มีการศึกษาเกี่ยวกับตัวแปรนี้ แต่เนื่องจากในสถานการณ์การทดสอบทั่วไปนั้นมีโอกาสที่จะเกิดความไม่เหมาะสมของโมเดลการวัดกับข้อสอบได้ ดังนั้นจึงมีความจำเป็นที่ต้องพิจารณาถึงปัจจัยสำคัญนี้ด้วย และเนื่องจากในสถานการณ์การทดสอบที่ได้มาตรฐานนั้น ระดับความไม่เหมาะสมของโมเดลการวัดกับข้อสอบที่ยอมรับได้ไม่ควรเกินร้อยละ 20 ผู้วิจัยจึงนำระดับความไม่เหมาะสมของโมเดลการวัดกับข้อสอบมากำหนดเป็นเงื่อนไขในการศึกษาจำลองครั้งนี้ โดยในการวิจัยครั้งนี้กำหนดความไม่เหมาะสมของโมเดลการวัดกับข้อสอบที่ใช้เป็นเงื่อนไขในการจำลองข้อมูลจำนวน 2 เงื่อนไข ได้แก่

2.3.1 แบบสอบที่มีจำนวนข้อสอบที่ไม่เหมาะสมกับโมเดลน้อยกำหนดเป็นร้อยละ 10

2.3.2 แบบสอบที่มีจำนวนข้อสอบที่ไม่เหมาะสมกับโมเดลมากกำหนดเป็นร้อยละ 20

3. การตรวจสอบประสิทธิภาพของวิธีการประมาณค่าในการศึกษาครั้งนี้ใช้การเปรียบเทียบค่าเฉลี่ยดัชนีการจำแนกประเภททั้งดัชนีความถูกต้องและดัชนีความสอดคล้องของการจำแนกประเภท โดยคำนวณได้จากการหาค่าเฉลี่ยของดัชนีการจำแนกประเภทจากการทำซ้ำ (replication) จำนวน 100 รอบ

4. ข้อมูลเชิงประจักษ์หรือข้อมูลจริงที่ใช้ในการศึกษาครั้งนี้ ใช้ข้อมูลการตอบข้อสอบและคะแนนสอบของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ที่ทำการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน

(O-NET) ในปีการศึกษา 2556 จำนวน 2 รายวิชาหลัก ได้แก่ วิชาคณิตศาสตร์และวิชาภาษาไทย และเกณฑ์การเทียบคะแนนทีปกติ (Normalized T-score) จากผลการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET) ของแต่ละวิชา มีแนวการให้ระดับผลการประเมินเป็น 8 ระดับ ซึ่งดำเนินการสร้างข้อสอบ กำหนดคะแนนจุดตัด และบริหารจัดการการทดสอบโดยสถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน)

คำจำกัดความที่ใช้ในการวิจัย

1. ดัชนีการจำแนกประเภท หมายถึง ค่าความน่าจะเป็นที่แสดงถึงการจำแนกหรือจัดผู้สอบเข้าสู่กลุ่มระดับความสามารถที่กำหนดไว้ โดยใช้คะแนนสอบเป็นข้อมูลในการเทียบกับเกณฑ์หรือคะแนนจุดตัดของแต่ละระดับสมรรถภาพ ประกอบด้วย 2 ดัชนี ได้แก่

1.1 ดัชนีความสอดคล้องของการจำแนกประเภท หมายถึง ความน่าจะเป็นในการตัดสินผลคะแนนของผู้สอบตามเกณฑ์ที่กำหนดไว้ได้สอดคล้องกัน คำนวณได้จากสูตรต่อไปนี้ (Wyse & Hao, 2012)

$$\hat{\gamma} = \frac{\Sigma\left(\hat{\mathrm{P}} * \hat{\mathrm{P}}\right)}{\mathrm{N}_e}$$

เมื่อ $\hat{\gamma}$ คือ ดัชนีความสอดคล้องของการจำแนกประเภท

$\hat{P}$ คือ เมตริกซ์ $N_e \times C$ ของความน่าจะเป็นที่คาดหวัง

N_e คือ จำนวนผู้สอบ

ดัชนีความสอดคล้องของการจำแนกประเภทมีค่าตั้งแต่ 0-1 โดยถ้ามีค่าเข้าใกล้ 1 แสดงว่ามีความน่าจะเป็นในการจำแนกผู้สอบได้สอดคล้องกันสูง ถ้าค่าเข้าใกล้ 0 แสดงว่ามีความน่าจะเป็นในการจำแนกผู้สอบได้สอดคล้องกันต่ำ

1.2 ดัชนีความถูกต้องของการจำแนกประเภท หมายถึง ความน่าจะเป็นในการตัดสินผลคะแนนที่สังเกตได้ของผู้สอบตามเกณฑ์ที่กำหนดไว้ตามความสามารถที่แท้จริงได้อย่างถูกต้อง คำนวณได้จากสูตรต่อไปนี้ (Wyse & Hao, 2012)

$$\hat{\tau} = \frac{\Sigma\left(\hat{\mathrm{P}} * W\right)}{\mathrm{N}_e}$$

เมื่อ $\hat{\tau}$ คือ ดัชนีความถูกต้องของการจำแนกประเภท

$\hat{P}$ คือ เมตริกซ์ $N_e \times C$ ของความน่าจะเป็นที่คาดหวัง

W คือ เมตริกซ์ $N_e \times$ W ของน้ำหนักซึ่งใช้ในการกำหนดกลุ่มระดับความสามารถที่ผู้สอบได้รับในการประเมิน

N_e คือ จำนวนผู้สอบ

ดัชนีความถูกต้องของการจำแนกประเภทมีค่าตั้งแต่ 0-1 โดยถ้ามีค่าเข้าใกล้ 1 แสดงว่ามีความน่าจะเป็นในการจำแนกผู้สอบได้ถูกต้องตามความสามารถที่แท้จริงสูง ถ้าค่าเข้าใกล้ 0 แสดงว่ามีความน่าจะเป็นในการจำแนกผู้สอบได้ถูกต้องตามความสามารถที่แท้จริงต่ำ

2. วิธีการประมาณค่าดัชนีการจำแนกประเภท หมายถึง การดำเนินการทางคณิตศาสตร์เพื่อให้ได้มาซึ่งค่าที่แสดงถึงความเห็นพ้องต้องกันการจำแนกหรือจัดผู้สอบเข้าสู่กลุ่มระดับสมรรถภาพที่กำหนดไว้ ซึ่งสามารถใช้ประมาณค่าได้ทั้งดัชนีความสอดคล้องของการจำแนกประเภทและดัชนีความถูกต้องของการจำแนกประเภท โดยในการศึกษาครั้งนี้ใช้วิธีการตามทฤษฎีการตอบสนองข้อสอบ 3 วิธีการ ได้แก่

2.1 วิธีการของ Rudner หมายถึง วิธีการที่พัฒนาขึ้นโดย Rudner (2005) มีข้อตกลงเบื้องต้นเกี่ยวกับการแจกแจงปกติของความคลาดเคลื่อนมาตรฐานในการประมาณค่าคะแนนจริง เป็นวิธีการที่ใช้ได้กับทั้งข้อมูลที่มีลักษณะของข้อสอบที่มีการให้คะแนนได้สองค่า (dichotomous item) ข้อสอบที่มีการให้คะแนนได้มากกว่าสองค่า (polytomous item) และข้อสอบที่มีการให้คะแนนในรูปแบบคะแนนตามทฤษฎีการตอบสนองข้อสอบ (IRT pattern score) และใช้คะแนนความสามารถตามสเกลของ θ (test scored on theta scale) เป็นโมเดลในการประมาณค่าดัชนี

วิธีการของ Rudner มีสูตรในการหาค่าความน่าจะเป็นในการจำแนกผู้สอบที่มีความสามารถ θ เข้ากลุ่มความสามารถ (Wyse & Hao, 2012) ดังนี้

$$\hat{p}_{iC} = \phi\left(\kappa_{C_i}, \kappa_{C_{i+1}}, \hat{\theta}_i, \hat{\sigma}_{\theta_i}\right)$$

สูตรในการคำนวณค่าดัชนีความสอดคล้องของการจำแนกประเภท (Wyse & Hao, 2012) คือ

$$\hat{\gamma} = \frac{\Sigma\left(\hat{\mathrm{P}} * \hat{\mathrm{P}}\right)}{\mathrm{N}_e}$$

สูตรในการคำนวณค่าดัชนีความถูกต้องของการจำแนกประเภท (Wyse & Hao, 2012) คือ

$$\hat{\tau} = \frac{\Sigma\left(\hat{\mathrm{P}} * W\right)}{\mathrm{N}_e}$$

2.2 วิธีการของ Guo หมายถึง วิธีการที่พัฒนาขึ้นโดย Guo (2006) มีข้อตกลงเบื้องต้นเกี่ยวกับการแจกแจงปกติของฟังก์ชันความน่าจะเป็น (likelihood functions) ในการตอบข้อสอบของผู้สอบ เป็นวิธีการที่ใช้ได้กับทั้งข้อมูลที่มีลักษณะของข้อสอบที่มีการให้คะแนนได้สองค่า

(dichotomous item) ข้อสอบที่มีการให้คะแนนได้มากกว่าสองค่า (polytomous item) และข้อสอบที่มีการให้คะแนนในรูปแบบคะแนนตามทฤษฎีการตอบสนองข้อสอบ (IRT pattern score) และใช้ latent distribution เป็นโมเดลในการประมาณค่าดัชนี

วิธีการของ Guo มีสูตรในการหาค่าความน่าจะเป็นในการจำแนกผู้สอบที่มีความสามารถ θ เข้ากลุ่มความสามารถ (Wyse & Hao, 2012) ดังนี้

$$\hat{p}_{ic} = \frac{\sum_{\theta=\kappa_{c_i}}^{\kappa_{c_{i+1}}} L(u_{1i}, u_{2i}, \ldots, u_{ni} \mid \theta)}{\sum_{h=1}^{C+1} \sum_{\theta=\kappa_h}^{\kappa_{h+1}} L(u_{1i}, u_{2i}, \ldots, u_{ni} \mid \theta)}$$

ส่วนสูตรที่ใช้ในการคำนวณค่าดัชนีความสอดคล้องและความถูกต้องของการจำแนกประเภทนั้นใช้สูตรเดียวกับวิธีการของ Rudner ซึ่งพัฒนาขึ้นโดย Wyse และ Hao (2012)

2.3 วิธีการของ Lee หมายถึง วิธีการที่พัฒนาขึ้นโดย Lee (2010) มีข้อตกลงเบื้องต้นเกี่ยวกับการแจกแจงแบบอเนกนามเชิงซ้อน (compound binomial) ของคะแนนจริง เป็นวิธีการที่ใช้ได้กับทั้งข้อมูลที่มีลักษณะของข้อสอบที่มีการให้คะแนนได้สองค่า (dichotomous item) ข้อสอบที่มีการให้คะแนนได้มากกว่าสองค่า (polytomous item) และข้อสอบที่มีการให้คะแนนในรูปแบบคะแนนรวม (summed score) และใช้ mixture IRT model เป็นโมเดลในการประมาณค่าดัชนี

วิธีการของ Lee มีสูตรในการหาค่าความน่าจะเป็นในการจำแนกผู้สอบที่มีความสามารถ θ เข้ากลุ่มความสามารถ (Wyse & Hao, 2012) ดังนี้

$$\hat{p}_{ic} = \sum_{x=\kappa_C}^{\kappa_{C+1}} fn\left(X = x \mid \hat{\theta}\right)$$

ส่วนสูตรที่ใช้ในการคำนวณค่าดัชนีความสอดคล้องและความถูกต้องของการจำแนกประเภทนั้นใช้สูตรเดียวกับวิธีการของ Rudner ซึ่งพัฒนาขึ้นโดย Wyse และ Hao (2012)

3. ประสิทธิภาพของวิธีการประมาณค่า หมายถึง ค่าที่ใช้ในการระบุถึงคุณภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภท ซึ่งในงานวิจัยนี้คือค่าเฉลี่ยของค่าดัชนีการจำแนกประเภทคำนวณได้จากการหาค่าเฉลี่ยค่าดัชนีการจำแนกประเภท ทั้งดัชนีความสอดคล้อง (consistency) และดัชนีความถูกต้อง (accuracy) ของการจำแนกประเภทจากการทำวนซ้ำ (replication) จำนวน 100 รอบ โดยหากวิธีการใดมีค่าเฉลี่ยดัชนีการจำแนกประเภทสูงจะหมายความว่าวิธีการประมาณค่านั้นมีประสิทธิภาพสูงที่สุด

4. ความยาวของแบบสอบ หมายถึง จำนวนข้อสอบที่บรรจุอยู่ในแบบสอบที่ใช้เป็นข้อมูลในการจำลองข้อมูลและการประมาณค่าดัชนีการจำแนกประเภท โดยตัวแปรความยาวของแบบสอบที่ใช้เป็นเงื่อนไขในการจำลองข้อมูลมี 2 เงื่อนไข คือ 1) แบบสอบสั้นที่มีข้อสอบจำนวน 25 ข้อ และ 2) แบบสอบยาวที่มีข้อสอบจำนวน 50 ข้อ

5. โมเดลการวัด หมายถึง ลักษณะของโมเดลการวัดที่ใช้ในการประมาณค่าดัชนีการจำแนกประเภท โดยตัวแปรโมเดลการวัดที่ใช้เป็นเงื่อนไขในการจำลองข้อมูลมี 3 โมเดล คือ 1) โมเดลโลจิสติกแบบหนึ่งพารามิเตอร์ (one-parameter logistic model: 1PL) 2) โมเดลโลจิสติกแบบสองพารามิเตอร์ (two-parameter logistic model: 2PL) และ 3) โมเดลโลจิสติกแบบสามพารามิเตอร์ (three-parameter logistic model: 3PL)

6. ความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ หมายถึง ลักษณะของข้อสอบที่ไม่เหมาะสมกับโมเดลการวัดที่ใช้ในการประมาณค่าพารามิเตอร์ข้อสอบและพารามิเตอร์ความสามารถผู้สอบ โดยพิจารณาจากจำนวนข้อสอบที่ไม่เหมาะสมกับโมเดลเทียบกับจำนวนข้อสอบทั้งหมด ตัวอย่างเช่นกรณีที่ใช้โมเดลการวัดแบบ 3PL ในการประมาณค่าพารามิเตอร์กับแบบสอบที่ประกอบด้วยข้อสอบที่เหมาะสมกับทั้งโมเดลแบบ 1PL, 2PL และ 3PL ในแบบสอบชุดเดียวกัน แสดงว่ามีความไม่เหมาะสมของโมเดลการวัดกับข้อสอบเกิดขึ้น โดยตัวแปรความไม่เหมาะสมของโมเดลการวัดกับข้อสอบที่ใช้เป็นเงื่อนไขในการจำลองข้อมูลมี 2 เงื่อนไข คือ 1) ความไม่เหมาะสมของโมเดลการวัดกับข้อสอบร้อยละ 10 และ 2) ความไม่เหมาะสมของโมเดลการวัดกับข้อสอบร้อยละ 20

7. ผลการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET) หมายถึง ผลการประเมินระดับชาติขั้นพื้นฐานปีการศึกษา 2556 ที่จัดทดสอบโดยสถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน) ใน 2 รายวิชาหลัก คือ ภาษาไทยและคณิตศาสตร์ โดยมีขั้นตอนในการกำหนดคะแนนจุดตัดดังนี้ 1) กำหนดระดับคะแนนเป็น 8 ระดับ 2) กำหนดช่วงคะแนนในแต่ละระดับด้วยวิธี Normalized T-Score 3) กำหนดเกณฑ์คะแนนต่ำสุดระดับผ่านหรือระดับ 1 ที่ควรสูงกว่าคะแนนค่าของโอกาสการเดา เช่น แบบสอบปรนัยแบบ 4 ตัวเลือก คะแนนเต็ม 100 คะแนน เกณฑ์คะแนนต่ำสุดระดับผ่านควรสูงกว่า 25 คะแนน 4) กำหนดเกณฑ์คะแนนต่ำสุดที่ได้รับ 4 ควรมีคะแนนตั้งแต่ร้อยละ 80 และ 5) ช่วงคะแนนในแต่ละระดับแต่ละวิชาจะไม่กำหนดคงที่ โดยจะผันแปรไปตามการกระจายของคะแนนวิชานั้นๆ และระดับความยากง่ายของข้อสอบ (สถาบันทดสอบทางการศึกษาแห่งชาติ, 2556)

ประโยชน์ที่ได้รับ

1. ทำให้ได้สารสนเทศเกี่ยวกับวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบที่มีประสิทธิภาพดีสุด ซึ่งจะเป็นประโยชน์ต่อการเลือกใช้วิธีการจำแนกประเภท

ที่เหมาะสมในการนำไปประยุกต์ใช้ในการประมาณค่าดัชนีการจำแนกประเภท ทั้งดัชนีความสอดคล้อง และดัชนีความถูกต้องของการจำแนกประเภทสำหรับสถานการณ์การทดสอบที่มีผลได้ผลเสียสูง

2. ทำให้ทราบถึงโมเดลการวัดที่เหมาะสมสำหรับการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบภายใต้สถานการณ์การทดสอบที่มีผลได้ผลเสียสูง เพื่อให้สามารถเลือกใช้วิธีการประมาณค่าที่สอดคล้องกับโมเดลการวัดได้อย่างเหมาะสม และสามารถควบคุมความคลาดเคลื่อนที่อาจเกิดขึ้นจากการเลือกใช้โมเดลการวัดที่ไม่เหมาะสมได้

3. ทำให้ทราบถึงวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบที่เหมาะสมกับความยาวของแบบสอบภายใต้สถานการณ์การทดสอบที่มีผลได้ผลเสียสูง เพื่อให้สามารถเลือกใช้วิธีการประมาณค่าที่สอดคล้องกับแบบสอบที่มีความยาวที่แตกต่างกันได้อย่างเหมาะสม และสามารถควบคุมความคลาดเคลื่อนที่อาจเกิดขึ้นจากการเลือกใช้วิธีการประมาณค่าที่ไม่เหมาะสมกับความยาวของแบบสอบได้

4. ทำให้ทราบถึงวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบที่เหมาะสมกับสถานการณ์การทดสอบที่มีความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ เพื่อให้สามารถเลือกใช้วิธีการประมาณค่าที่สอดคล้องกับสถานการณ์การทดสอบที่มีความไม่เหมาะสมของโมเดลการวัดกับข้อสอบที่แตกต่างกันได้อย่างเหมาะสม และสามารถควบคุมความคลาดเคลื่อนที่อาจเกิดขึ้นจากการเลือกใช้วิธีการประมาณค่าที่ไม่เหมาะสมกับสถานการณ์การทดสอบที่ข้อสอบมีความไม่เหมาะสมกับโมเดลการตอบสนองข้อสอบ

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

การศึกษาประสิทธิภาพของการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบภายใต้เงื่อนไขที่แตกต่างกันในครั้งนี้ ผู้วิจัยทำการค้นคว้าและศึกษาตำรา เอกสาร และงานวิจัยที่เกี่ยวข้องจนได้มาซึ่งมวลความรู้เชิงทฤษฎี โดยแบ่งการนำเสนอออกเป็น 5 ตอน ดังนี้ ตอนแรกเป็นการนำเสนอความรู้เบื้องต้นเกี่ยวกับดัชนีการจำแนกประเภท สาระในตอนนี้กล่าวถึงความหมาย ความเป็นมา และความจำเป็นที่ต้องมีการประมาณค่าดัชนีการจำแนกประเภท ตอนที่ 2 แนวคิดด้านการวัดผลทางการศึกษาที่อธิบายเกี่ยวกับดัชนีการจำแนกประเภท อันเป็นที่มาของแนวคิดในการพัฒนาวิธีการประมาณค่าดัชนีการจำแนกประเภทต่างๆ ที่นักวัดผลทางการศึกษาพัฒนาขึ้น ตอนที่ 3 การประมาณค่าดัชนีการจำแนกประเภทด้วยวิธีการต่างๆ ที่นักวัดผลทางการศึกษาพัฒนาขึ้นตั้งแต่เริ่มต้นจนถึงปัจจุบัน ตอนที่ 4 งานวิจัยที่ศึกษาเกี่ยวกับดัชนีการจำแนกประเภทเพื่อนำไปประยุกต์ใช้ในสถานการณ์ต่างๆ และตอนสุดท้ายนำเสนอกรอบแนวคิดที่ใช้ในการศึกษาครั้งนี้

ตอนที่ 1 ความรู้เบื้องต้นเกี่ยวกับดัชนีการจำแนกประเภท

สาระในตอนนี้จะบรรยายถึงความเป็นมาและความสำคัญของดัชนีการจำแนกประเภท และความหมายของดัชนีการจำแนกประเภท ดังรายละเอียดต่อไปนี้

1.1 ความเป็นมาและความสำคัญของดัชนีการจำแนกประเภท

การทดสอบทางการศึกษานับว่าเป็นกระบวนการที่สำคัญกระบวนการหนึ่งในการตัดสินใจทางการศึกษา ไม่ว่าจะเป็นการทดสอบขนาดใหญ่ทางการศึกษา (large-scale testing) ที่มีวัตถุประสงค์ของการทดสอบเพื่อให้ได้มาซึ่งคะแนนที่เชื่อถือได้ โดยสามารถนำไปตีความได้อย่างสมเหตุสมผล และเพื่อสร้างการตัดสินใจเกี่ยวกับการจัดตำแหน่ง (placement) การปรับปรุงแก้ไข (remediation) และการรับรองผล (certification) (Wainer & Kiely, 1987) หรือจะเป็นการทดสอบตามหลักสูตรที่จัดขึ้นในสถานศึกษา ที่บริหารจัดการการทดสอบให้สอดคล้องกับการจัดการเรียนการสอนแบบอิงวัตถุประสงค์ตามหลักสูตร (objective-based instructional) โดยมีวัตถุประสงค์ของการทดสอบเพื่อจัดนักเรียนเข้าสู่กลุ่มสมรรถภาพหรือกลุ่มรอบรู้ตามแต่ละจุดประสงค์ของหลักสูตรการเรียนการสอนและนำข้อมูลไปใช้ในการกำกับ ติดตาม และแก้ไขพฤติกรรมการเรียนรู้ของนักเรียน (Swaminathan, Hambleton, & Algina, 1974) ซึ่งการตัดสินใจเกี่ยวกับผู้สอบนั้นต้องนำคะแนนที่ได้จากการทดสอบมาทำการเปรียบเทียบกับเกณฑ์ตามวัตถุประสงค์ของการทดสอบหรือ

กล่าวอีกนัยหนึ่งว่าคะแนนจุดตัด เช่น การตัดสินการสอบผ่านหรือตกของนักเรียน โดยกระบวนการดังกล่าวข้างต้นคือการจำแนกประเภทนั่นเอง (Lathrop & Cheng, 2013)

นอกจากนี้ในบริบทของการทดสอบจำนวนมากมีความจำเป็นที่จะต้องจำแนกผู้สอบเข้าสู่กลุ่มที่มีทักษะการปฏิบัติเฉพาะตัวร่วมกัน ซึ่งขึ้นอยู่กับชุดของมาตรฐานต่างๆ ที่กำหนดไว้ โดยมาตรฐานเหล่านี้สามารถให้นิยามได้ว่าเป็นชุดของคะแนนจุดตัดที่ได้มาจากกระบวนการกำหนดมาตรฐานนั่นเอง ตัวอย่างที่เห็นได้อย่างเด่นชัดคือ การจำแนกประเภทของทักษะการปฏิบัติ ซึ่งเป็นวิธีการที่ทำได้ง่ายและสะดวกในการอธิบายและการตีความระดับความสามารถของผู้สอบจากผลการปฏิบัติที่ผู้สอบแสดงพฤติกรรมเพื่อตอบสนองต่อสิ่งเร้าที่กำหนดไว้ และมีการใช้กันอย่างแพร่หลายทั้งในการทดสอบทางการศึกษาและการทดสอบเพื่อขอใบอนุญาตต่างๆ การจำแนกประเภทนั้นมีทั้งการจำแนกประเภทออกเป็นสองกลุ่ม (binary classification) ที่ต้องการแบ่งผู้สอบออกเป็นสองกลุ่ม คือกลุ่มรอบรู้และกลุ่มไม่รอบรู้ (mastery/non-mastery) หรือกลุ่มผ่านและกลุ่มตก (pass/fail) ซึ่งตัดสินโดยใช้คะแนนจุดตัดเพียงจุดตัดเดียวเท่านั้น และการจำแนกประเภทออกเป็นหลายกลุ่ม (multiple classifications) ที่ต้องการแบ่งผู้สอบเข้าสู่กลุ่มมากกว่าสองกลุ่ม เช่น อาจแบ่งเป็นสี่กลุ่มดังนี้ กลุ่มจำเป็นต้องปรับปรุง กลุ่มขั้นพื้นฐาน กลุ่มมีความเชี่ยวชาญ และกลุ่มดีมาก เป็นต้น ซึ่งตัดสินโดยใช้คะแนนจุดตัดหลายจุดตัดขึ้นอยู่กับจำนวนกลุ่มที่ต้องการ

ในการประเมินจำเป็นจะต้องทราบว่า เครื่องมือวัดใดมีความคลาดเคลื่อนที่เกิดขึ้นโดยธรรมชาติบ้าง ดังนั้นคะแนนสอบที่ใช้ในการจำแนกนักเรียนตามความสามารถของนักเรียนแต่ละคนจึงรวมอยู่ในความคลาดเคลื่อนนั้นด้วย ดังนั้นจึงเป็นสิ่งสำคัญที่ผู้เชี่ยวชาญทางการศึกษาจะต้องประเมินความถูกต้องของการจำแนกประเภท (classification accuracy) และความสอดคล้องของการจำแนกประเภท (classification consistency) เพื่อตรวจสอบความคลาดเคลื่อนของการวัดที่เกิดขึ้น อันนำไปสู่การจำแนกนักเรียนเข้าสู่กลุ่มระดับสมรรถภาพที่ตรงกับความสามารถที่แท้จริง โดยการใช้คะแนนเหล่านี้จะมีความสำคัญและมีความยุติธรรมถ้าการตัดสินใจที่สร้างขึ้นมีทั้งความถูกต้อง (accurate) และความสอดคล้องกัน (consistent) (Wainer & Kiely, 1987)

1.2 ความหมายของดัชนีการจำแนกประเภท

ส่วนนี้จะแบ่งการอธิบายความหมายของดัชนีการจำแนกประเภทออกเป็นสองส่วนคือ ส่วนแรกเป็นการอธิบายความหมายของความสอดคล้องของการจำแนกประเภท และส่วนที่สองเป็นการอธิบายความหมายของดัชนีความถูกต้องของการจำแนกประเภท

1.2.1 ความหมายของดัชนีความสอดคล้องของการจำแนกประเภท

ความสอดคล้องของการจำแนกประเภท (classification consistency) อ้างถึงการใช้แบบสอบเพื่อสร้างการตัดสินใจในการจัดผู้สอบเข้ากลุ่มสมรรถภาพ ในขอบเขตของการจำแนก

ประเภทที่เห็นพ้องต้องกันในสถานการณ์ของการบริหารจัดการการทดสอบสองสถานการณ์ที่เป็นอิสระต่อกัน หรือรูปแบบของการใช้แบบสอบคู่ขนาน (Hambleton & Novick, 1973)

Livingston และ Lewis (1995) ได้ให้ความหมายของความสอดคล้องของการจำแนกประเภท (classification consistency) ว่าหมายถึง ความสอดคล้องกันระหว่างการจำแนกประเภทต่างๆ ที่มีพื้นฐานอยู่บนสถานการณ์การทดสอบสองสถานการณ์ที่ไม่เกี่ยวเนื่องกัน และใช้แบบสอบที่มีความยากเท่าเทียมกัน

Young และ Yoon (1998) ได้กล่าวถึงความสอดคล้องของการตัดสินใจ (consistency of a decision) ไว้ว่า ความสอดคล้องของการตัดสินใจ (consistency of a decision) เป็นขอบเขตที่การตัดสินใจดังกล่าวจะเห็นพ้องกับการตัดสินใจที่ควรจะเป็น เมื่อนักเรียนได้รับรูปแบบการทดสอบที่แตกต่างกันด้วยการทดสอบในมาตรฐานใหม่ ซึ่งมีความยากเท่าเทียมกันและครอบคลุมเนื้อหาเดียวกันกับรูปแบบที่ได้รับการทดสอบจริง

Brennan (2001 cited in Lee, Hanson, & Brennan, 2002) กล่าวว่า ความสอดคล้องของการจำแนกประเภท (classification consistency) มักถูกอ้างถึงความเที่ยง (reliability) ของการจำแนกประเภท เพราะความหมายของความสอดคล้องของการจำแนกประเภท (classification consistency) ต้องใช้แนวคิดของการทดสอบซ้ำ ซึ่งเป็นองค์ประกอบที่สำคัญที่สุดของการวิเคราะห์ความเที่ยง (reliability)

Lee et al. (2002) ให้นิยามของความสอดคล้องของการจำแนกประเภท (classification consistency) ว่าเป็นขนาดของความสอดคล้องของการจำแนกบนฐานของการบริหารจัดการการทดสอบสองสถานการณ์ หรือชุดของแบบสอบสองชุดที่คู่ขนานกัน อันเนื่องมาจากข้อมูลที่ได้จากการวัดซ้ำ การคำนวณดัชนีความสอดคล้องของการจำแนกประเภท (classification consistency) ในเชิงจิตมิติจะใช้คะแนนสอบจากการจัดการทดสอบเพียงครั้งเดียว ผลลัพธ์หนึ่งที่ได้จากการวิเคราะห์ดัชนีความสอดคล้องของการจำแนกประเภทที่มีหลายกลุ่มคือตารางความสอดคล้องขนาด H×H สมาชิกของตาราง H×H คือ ความน่าจะเป็นร่วม (joint probabilities) ของการจำแนกที่สังเกตได้ในแถวและคอลัมน์ ถึงแม้ว่าจะมีความเป็นไปได้ที่จะได้จำนวนดัชนีที่แตกต่างกันจากตารางการณ์จรหลายประเภท (multiple contingency table)

Clauser, Margolis และ Case (2006) อธิบายว่า ดัชนีความสอดคล้องของการจำแนกประเภท (classification consistency) เป็นตัวแทนของการประเมินที่คล้ายกับการประเมินคุณสมบัติด้านความเที่ยงของแบบสอบที่ใช้สำหรับการตัดสินใจต่างๆ ในการจำแนกประเภท ถ้าแบบสอบหรือแบบสอบที่คู่ขนานกันได้รับการบริหารจัดการสำหรับกลุ่มตัวอย่างผู้สอบในสองสถานการณ์

ความสอดคล้องของการจำแนกประเภท (classification consistency) สามารถสรุปให้อยู่ในขอบเขตของสัดส่วนของผู้สอบที่ได้รับการจำแนกด้วยวิธีเดียวกันกับทั้งสองสถานการณ์

Lee (2010) กล่าวว่าจากเอกสารทางการวัดผลได้อ้างถึงความสอดคล้องของการจำแนกประเภท (classification consistency) ว่าเป็นระดับที่ผู้สอบถูกแบ่งออกเป็นหมวดหมู่หรือประเภทเข้าไปอยู่ในกลุ่มสมรรถภาพเดียวกันด้วยการตอบข้อสอบในแบบสอบที่คู่ขนานกันในการประเมินเดียวกัน

Lathrop และ Cheng (2013) ให้นิยามของความสอดคล้องของการจำแนกประเภท (classification consistency) ไว้ว่าเป็นสัดส่วนของผู้สอบที่ได้รับการจำแนกเข้าสู่กลุ่มสมรรถภาพเดียวกันจากการบริหารจัดการการทดสอบสองสถานการณ์โดยใช้แบบสอบที่เป็นอิสระต่อกันและคู่ขนานกัน

กล่าวโดยสรุป ความสอดคล้องของการจำแนกประเภท (classification consistency) หรือความสอดคล้องของการตัดสิน (decision consistency) หมายถึง ความสอดคล้องของการตัดสินผลคะแนนของผู้สอบตามเกณฑ์ที่กำหนดไว้จากการทดสอบด้วยแบบสอบสมมูลหรือแบบสอบคู่ขนาน ซึ่งสามารถนำเสนอให้เข้าใจได้ง่ายขึ้นดังตารางต่อไปนี้

ตารางที่ 2.1 ความสอดคล้องของการจำแนกประเภท (classification consistency)

		การตัดสินใจบนพื้นฐานของการทดสอบรูปแบบที่สอง	
		ต่ำกว่ามาตรฐาน (Below the Standard)	**สูงกว่ามาตรฐาน (Above the Standard)**
การตัดสินใจบนพื้นฐานของการทดสอบรูปแบบที่หนึ่ง	**ต่ำกว่ามาตรฐาน (Below the Standard)**	การจำแนกที่สอดคล้อง (Consistent Classification)	การจำแนกที่ไม่สอดคล้อง (Inconsistent Classification)
	สูงกว่ามาตรฐาน (Above the Standard)	การจำแนกที่ไม่สอดคล้อง (Inconsistent Classification)	การจำแนกที่สอดคล้อง (Consistent Classification)

1.2.2 ความหมายของดัชนีความถูกต้องของการจำแนกประเภท

ความถูกต้องของการจำแนกประเภท (classification accuracy) อ้างถึงการใช้แบบสอบเพื่อสร้างการตัดสินใจในการจัดผู้สอบเข้ากลุ่มสมรรถภาพ ในขอบเขตของการจำแนกที่เกิดขึ้นจริง (actual classification) อยู่บนฐานของคะแนนที่สังเกตได้ (observed scores) เห็นพ้องต้องกันกับการจำแนกที่แท้จริง (true classification) ที่อยู่บนพื้นฐานของคะแนนจริง (true scores) (Hambleton & Novick, 1973)

Livingston และ Lewis (1995) ได้ให้ความหมายของความถูกต้องของการจำแนกประเภท (classification accuracy) ว่าหมายถึงขอบเขตในการจำแนกประเภทที่แท้จริงของผู้สอบ ซึ่งสอดคล้องกับคะแนนที่น่าจะได้บนฐานของคะแนนจริง หากคะแนนจริงของผู้สอบถูกแสดงออกมาด้วยวิธีการใดวิธีการหนึ่ง

Young และ Yoon (1998) ได้กล่าวถึงความถูกต้องของการตัดสินใจ (accuracy of a decision) ไว้ว่า ความถูกต้องของการตัดสินใจเป็นขอบเขตที่การตัดสินใจดังกล่าวจะเห็นพ้องกับการตัดสินใจที่ควรจะเป็น เมื่อนักเรียนแต่ละคนได้รับการทดสอบด้วยรูปแบบใดรูปแบบหนึ่งจากรูปแบบที่เป็นไปได้ทั้งหมดของการทดสอบ

Lee et al. (2002) อธิบายว่า ความถูกต้องของการจำแนกประเภท (classification accuracy) มีความสัมพันธ์ที่ใกล้เคียงกับความตรง (validity) ของระบบการจำแนกประเภท การประเมินระดับของความถูกต้อง (accuracy) ของการจำแนกประเภทอยู่บนพื้นฐานของคะแนนที่สังเกตได้ (observed scores) ว่าเป็นความพยายามที่จะจำแนกตามคะแนนที่แท้จริง (true scores)

Lee (2010) กล่าวว่า ความถูกต้องของการจำแนกประเภท (classification accuracy) อ้างถึงการประมาณค่าการจำแนกประเภทที่เกิดขึ้นจริงโดยใช้คะแนนจุดตัดที่สังเกตได้ที่สอดคล้องกับการจำแนกประเภทที่แท้จริงบนพื้นฐานของคะแนนจุดตัดที่รู้จริง

Lathrop และ Cheng (2013) ให้นิยามของความถูกต้องของการจำแนกประเภท (classification accuracy) ไว้ว่าเป็นสัดส่วนของผู้สอบที่ได้รับการจำแนกเข้าสู่กลุ่มสมรรถภาพที่ตรงกับสมรรถภาพที่แท้จริงของผู้สอบ

กล่าวโดยสรุป ความถูกต้องของการจำแนกประเภท (classification accuracy) หรือความถูกต้องของการตัดสินใจ (decision accuracy) หมายถึง ความถูกต้องของการตัดสินผลคะแนนที่สังเกตได้ของผู้สอบตามเกณฑ์ที่กำหนดไว้ตามความสามารถที่แท้จริงของผู้สอบ ซึ่งสามารถแสดงให้เข้าใจได้ง่ายขึ้นดังตารางต่อไปนี้

ตารางที่ 2.2 ความถูกต้องของการจำแนกประเภท (classification accuracy)

		การตัดสินใจจากการทดสอบ	
		ต่ำกว่ามาตรฐาน (Below the Standard)	**สูงกว่ามาตรฐาน (Above the Standard)**
"สถานะจริง" บนพื้นฐานของค่าเฉลี่ย	**ต่ำกว่ามาตรฐาน (Below the Standard)**	การจำแนกที่ถูกต้อง (Correct Classification)	การจำแนกที่ผิดพลาด (Misclassification)
	สูงกว่ามาตรฐาน (Above the Standard)	การจำแนกที่ผิดพลาด (Misclassification)	การจำแนกที่ถูกต้อง (Correct Classification)

ตอนที่ 2 แนวคิดด้านการวัดผลทางการศึกษาที่อธิบายเกี่ยวกับดัชนีการจำแนกประเภท

เนื่องด้วยการทดสอบในปัจจุบันมีวัตถุประสงค์เพื่อจัดประเภทความสามารถหรือสมรรถภาพของผู้สอบเพิ่มขึ้น ทำให้ผู้สอบอาจจะมีความกังวลบางอย่าง เช่น ความกังวลเกี่ยวกับสัดส่วนของผู้สอบที่คาดหวังว่าควรจะได้รับการจำแนกอย่างสอดคล้องกันเมื่อมีการทดสอบซ้ำ หรือความกังวลเกี่ยวกับสัดส่วนของผู้สอบที่มีคะแนนจริง (true score) สูงกว่าคะแนนจุดตัด แต่อาจจะได้รับการจำแนกเป็นผู้ไม่รอบรู้แทนที่จะเป็นผู้รอบรู้ เป็นต้น ดังนั้นความสอดคล้องและความถูกต้องของการจำแนกประเภทจึงกลายมาเป็นความกังวลที่สำคัญมากกว่าความกังวลเกี่ยวกับคะแนน และความกังวลนี้จะกลายเป็นที่น่าสนใจมากยิ่งขึ้นอันเนื่องมาจากผลที่ตามมาจากการตัดสินใจในแง่ของระดับสมรรถภาพของผู้สอบ ตัวอย่างเช่น การตัดสินใจอาจสร้างขึ้นเพื่อประเมินสมรรถภาพของครูและโรงเรียน เพื่อตรวจสอบระดับความสามารถของนักเรียนที่จะสำเร็จการศึกษา หรือเพื่อตัดสินว่าควรจะออกใบรับรองให้หรือไม่ เป็นต้น การประมาณค่าความเที่ยงแบบดั้งเดิมอาจจะไม่เหมาะสมสำหรับการประเมินความสอดคล้องและความถูกต้องของการจำแนกประเภท ดังนั้นจึงจำเป็นที่จะต้องใช้เทคนิคในการประเมินความเที่ยงแบบใหม่ที่มีความเหมาะสมในการใช้มากกว่า ซึ่งพัฒนาขึ้นโดยนักวัดผลทางการศึกษาตั้งแต่ปี ค.ศ. 1970 เป็นต้นมา

ดังนั้นความสอดคล้องและความถูกต้องของการจำแนกประเภทจึงเป็นสิ่งที่น่าสนใจอย่างมากเมื่อมีการนำแบบสอบมาใช้ในการจำแนกประเภทสมรรถภาพของผู้สอบ แนวคิดของความสอดคล้องและความถูกต้องของการจำแนกประเภทได้รับการเสนอให้เป็นดัชนีที่ใช้อธิบายถึงความเที่ยง (reliability) และความตรง (validity) ของการจำแนกประเภท (Hambleton & Novick, 1973) การนำเสนอในตอนนี้จะแบ่งเนื้อหาออกเป็นสองส่วนเพื่อบรรยายถึงแนวคิดด้านการวัดผลทางการศึกษาที่อธิบายเกี่ยวกับดัชนีการจำแนกประเภท โดยส่วนแรกจะอธิบายถึงดัชนีความสอดคล้องของการจำแนกประเภท (classification consistency index) และส่วนที่สองจะอธิบายถึงดัชนีความถูกต้องของการจำแนกประเภท (classification accuracy index)

2.1 แนวคิดเกี่ยวกับดัชนีความสอดคล้องของการจำแนกประเภท

ในหัวข้อนี้เป็นการอธิบายถึงความเป็นมาของความสอดคล้องของการจำแนกประเภท (classification consistency) หรือความสอดคล้องของการตัดสิน (decision consistency) ก่อนที่การอธิบายถึงความถูกต้องของการจำแนกประเภท (classification accuracy) หรือความถูกต้องของการตัดสินใจ (decision accuracy) เนื่องจากประวัติของการศึกษาพบว่ามีการพิจารณาถึงความสอดคล้องของการจำแนกประเภทก่อนความถูกต้องของการจำแนกประเภท

ดัชนีความสอดคล้องของการจำแนกประเภท (classification consistency) เป็นตัวแทนของการประเมินที่คล้ายกับการประเมินความเที่ยง ซึ่งเป็นคุณสมบัติหนึ่งของคะแนนสอบที่ใช้ใน

การตัดสินใจต่างๆ เกี่ยวกับการจำแนกประเภทหรือจัดกลุ่มผู้สอบ ถ้าทำการบริหารจัดการการทดสอบสำหรับกลุ่มตัวอย่างผู้สอบสองสถานการณ์โดยใช้แบบสอบที่คู่ขนานกัน ความสอดคล้องของการจำแนกประเภทสามารถสรุปให้อยู่ในขอบเขตของสัดส่วนของผู้สอบที่ได้รับการจำแนกด้วยวิธีเดียวกันกับทั้งสองสถานการณ์การบริหารจัดการการทดสอบ (Clauser, Margolis, & Case, 2006)

การประมาณค่าความสอดคล้องของการจำแนกประเภทนั้น คะแนนที่สังเกตได้หรือคะแนนที่ปรับปรุงใหม่ในการบริหารจัดการการทดสอบสองสถานการณ์ด้วยรูปแบบการทดสอบเดียวกันอาจมีความคลาดเคลื่อนเกิดขึ้น ตัวอย่างเช่น ผู้สอบบางคนอาจจะมีผลคะแนนที่ได้จากการบริหารจัดการการทดสอบแรกสูงกว่าคะแนนที่ได้จากการบริหารจัดการการทดสอบที่สองด้วยรูปแบบการทดสอบเดียวกัน อันนำไปสู่การที่ผู้สอบอาจจะได้รับการจำแนกหรือจัดเข้าสู่กลุ่มระดับสมรรถภาพที่สูงกว่าจากการบริหารจัดการการทดสอบแรก แต่ได้รับการจัดเข้าสู่กลุ่มระดับสมรรถภาพที่ต่ำกว่าจากการบริหารจัดการการทดสอบที่สองด้วยรูปแบบการทดสอบเดียวกัน ในขณะที่ผู้สอบอื่นอาจจะมีผลคะแนนเท่ากันในทั้งสองสถานการณ์การทดสอบ อันนำไปสู่การที่ผู้สอบอาจจะได้รับการจำแนกที่สอดคล้องกัน และยังคงมีผู้สอบอื่นๆ ที่อาจจะมีคะแนนจากการบริหารจัดการการทดสอบแรกต่ำกว่าคะแนนที่ได้จากการบริหารจัดการการทดสอบที่สองด้วยรูปแบบการทดสอบเดียวกัน อันนำไปสู่การที่ผู้สอบอาจจะได้รับการจำแนกหรือจัดเข้าสู่กลุ่มระดับสมรรถภาพที่ต่ำกว่าจากการบริหารจัดการการทดสอบแรก แต่ได้รับการจัดเข้าสู่กลุ่มที่สูงกว่าจากการบริหารจัดการการทดสอบที่สองด้วยรูปแบบการทดสอบเดียวกัน (Lee, Brennan & Kolen, 2000)

ความแตกต่างนี้มีสาเหตุเนื่องมาจากความคลาดเคลื่อนของการวัด (measurement error) อย่างไรก็ตามความแตกต่างบางประการอาจเกิดขึ้น และนำไปสู่ความไม่สอดคล้องกันในการจำแนกผู้สอบเข้าสู่กลุ่มระดับสมรรถภาพที่แตกต่างกัน ปัญหาก็คือการที่ไม่ทราบขนาดของ ความคลาดเคลื่อนของการวัดสำหรับผู้สอบแต่ละคนในแต่ละสถานการณ์ของการบริหารจัดการการทดสอบ (Wainer & Kiely, 1987) อย่างไรก็ตามมีนักวัดผลทางการศึกษาจำนวนหนึ่งตระหนักถึงความสำคัญของปัญหานี้ จึงได้พัฒนาวิธีการประมาณค่าความสอดคล้องของการจำแนกประเภทขึ้นมาโดยการประยุกต์ใช้ทฤษฎีทางการวัดผลที่หลากหลายและแตกต่างกันไปตามความเชื่อพื้นฐานของแต่ละคน เพื่อให้ได้วิธีการประมาณค่าที่เหมาะสมและมีความคลาดเคลื่อนในการประมาณค่าเกิดขึ้นน้อยที่สุด โดยจะอธิบายถึงวิธีการประมาณค่าที่นักวัดผลทางการศึกษาได้พัฒนาไว้ในตอนที่สาม ซึ่งจะบรรยายถึงวิธีการประมาณค่าดัชนีการจำแนกประเภท

2.2 แนวคิดเกี่ยวกับดัชนีความถูกต้องของการจำแนกประเภท

ความถูกต้องของการจำแนกประเภท (classification accuracy) หรือความถูกต้องของการตัดสินใจ (decision accuracy) ต่างจากความสอดคล้องของการจำแนกประเภท (cassification consistency) ตรงที่ความถูกต้องของการจำแนกประเภทนั้นต้องการเกณฑ์ที่ชัดเจน ถ้าความสอดคล้อง

ของการจำแนกประเภทสามารถเข้าใจว่าเป็นการจัดกระทำข้อมูลที่เหมือนกับความเที่ยงแล้ว ความถูกต้องของการจำแนกประเภทก็สามารถเข้าใจว่าเป็นการจัดกระทำข้อมูลที่เหมือนกับความตรงได้ ความแปลกใหม่ในมุมมองนี้มีความหมายโดยนัยที่สำคัญหลายประการ สิ่งที่เห็นได้อย่างเด่นชัดที่สุดคือ ถ้าความสอดคล้องของการจำแนกประเภทไม่มีความถูกต้องในตัวเองแล้ว การจำแนกประเภทที่สอดคล้องกันอาจจะผิดพลาดอย่างต่อเนื่อง (Young & Yoon, 1998)

แม้ว่าคะแนนสอบ (test scores) จะไม่ใช่ข้อผิดพลาดอันเนื่องมาจากความคลาดเคลื่อนของการวัด (measurement error) แต่คะแนนจริง (true scores) ที่สอดคล้องกันคือข้อผิดพลาดนั้น ดังนั้นการตัดสินใจในการจำแนกประเภทที่เกิดจากคะแนนจริง (true scores) คือการจำแนกประเภทที่ถูกต้อง ในทางกลับกันการตัดสินใจในการจำแนกประเภทที่เกิดจากคะแนนสอบจะไม่ถูกต้อง เนื่องจากประกอบไปด้วยความคลาดเคลื่อนจากการวัด (Keller, Swaminathan & Sireci, 2003) บางครั้งคะแนนของผู้สอบอาจจะต่ำเกินไปนำไปสู่ความคลาดเคลื่อนของการวัดเชิงลบ (negative measurement error) ซึ่งหมายความว่าผู้สอบอาจจะได้รับการจำแนกหรือจัดเข้าสู่กลุ่มที่ต่ำกว่าเมื่อคะแนนจริงของผู้สอบบ่งชี้ว่าผู้สอบควรจะอยู่ในกลุ่มที่สูงกว่าถัดไป ในขณะที่คะแนนของผู้สอบอื่นอาจถูกต้อง นำไปสู่ความคลาดเคลื่อนของการวัดที่เป็นศูนย์ (zero measurement error) ในกรณีนี้การจำแนกหรือจัดกลุ่มผู้สอบจะมีเหตุผลหรือเป็นไปตามความจริง และยังคงมีคะแนนของผู้สอบคนอื่นๆ ที่อาจจะสูงเกินไป นำไปสู่ความคลาดเคลื่อนของการวัดเชิงบวก (positive measurement error) ซึ่งหมายความว่าผู้สอบอาจจะได้รับการจำแนกหรือจัดเข้าสู่กลุ่มที่สูงกว่าเมื่อคะแนนจริงของผู้สอบบ่งชี้ว่าผู้สอบควรจะอยู่ในกลุ่มที่ต่ำกว่าถัดไป (Lee, Brennan, & Kolen, 2000) ปัญหาคือการที่ไม่ทราบถึงขนาดของความคลาดเคลื่อนของการวัดนั่นเอง ซึ่งเป็นไปในแนวทางเดียวกับความสอดคล้องของการจำแนกประเภทที่มีนักวัดผลทางการศึกษาจำนวนหนึ่งตระหนักถึงความสำคัญของปัญหานี้ จึงได้พัฒนาวิธีการประมาณค่าความถูกต้องของการจำแนกประเภทขึ้นมาโดยการประยุกต์ใช้ทฤษฎีทางการวัดผลที่หลากหลายและแตกต่างกันไปตามความเชื่อพื้นฐานของแต่ละคน เพื่อให้ได้วิธีการประมาณค่าที่เหมาะสมและมีความคลาดเคลื่อนในการประมาณค่าเกิดขึ้นน้อยที่สุด โดยจะอธิบายถึงวิธีการประมาณค่าที่นักวัดผลทางการศึกษาได้พัฒนาไว้ในตอนที่สาม ซึ่งจะบรรยายถึงวิธีการประมาณค่าดัชนีการจำแนกประเภท

ตอนที่ 3 วิธีการประมาณค่าดัชนีการจำแนกประเภท

เนื้อหาในตอนนี้จะแบ่งออกเป็นสองส่วนเพื่ออธิบายถึงวิธีการประมาณค่าดัชนีการจำแนกประเภททั้งดัชนีความสอดคล้องและความถูกต้องของการจำแนกประเภท โดยส่วนแรกอธิบายถึงวิธีการประมาณค่าดัชนีความสอดคล้องของการจำแนกประเภท (classification consistency

index) และส่วนที่สองอธิบายถึงวิธีการประมาณค่าดัชนีความถูกต้องของการจำแนกประเภท (classification accuracy index) ดังรายละเอียดต่อไปนี้

3.1 วิธีการประมาณค่าดัชนีความสอดคล้องของการจำแนกประเภท

การขับเคลื่อนด้วยแนวโน้มของการพัฒนาหลักสูตรช่วงปี ค.ศ. 1970 ทำให้นักวัดผลทางการศึกษาได้พัฒนาวิธีการสำหรับประเมินการตัดสินใจโดยใช้แบบสอบวัดความรอบรู้แบบอิงเกณฑ์ (criterion-referenced) หรืออิงโดเมน (domain-referenced) ซึ่งวิธีการเหล่านี้แบ่งได้เป็น 2 ประเภทตามแนวคิดพื้นฐาน คือ วิธีการที่ใช้แนวคิดของ squared-error loss functions และวิธีการที่ใช้แนวคิดของ threshold loss functions

ในปี ค.ศ. 1973 Hambleton และ Novick (1973) ได้เสนอดัชนีความเห็นพ้องต้องกัน (agreement index: P) ให้เป็นการวัดความสอดคล้องของการจำแนกประเภท โดยให้นิยามของดัชนีความเห็นพ้องต้องกันว่าเป็นสัดส่วนของผู้สอบที่ได้รับการจำแนกที่สอดคล้องกันในการบริหารจัดการการทดสอบสองสถานการณ์ และทำการขยายสูตร p_0 และ κ สำหรับใช้ในสถานการณ์ที่มีจำนวนกลุ่มสมรรถภาพมากกว่าสองกลุ่ม สูตรทั่วไปของ p_0 คือ

$$p_0 = \sum_{i=1}^{k} p_{ii}$$

เมื่อ p_{ii} คือ ร้อยละของผู้สอบที่ได้รับการจำแนกเข้าสู่กลุ่มระดับสมรรถภาพที่ i^{th} ในสองสถานการณ์ที่ใช้รูปแบบการทดสอบที่สามารถสับเปลี่ยนกันได้สองรูปแบบ (two interchangeable test forms) อย่างสอดคล้องกัน และ $k \geq 2$ คือจำนวนของกลุ่มระดับสมรรถภาพ

สูตรทั่วไปของ p_c คือ

$$p_c = \sum_{i=1}^{k} p_{i\bullet} p_{\bullet i}$$

เมื่อ $p_{i\bullet}$ และ $p_{\bullet i}$ คือ ร้อยละของผู้สอบแต่ละคนที่ได้รับการจำแนกเข้าสู่กลุ่มของระดับสมรรถภาพที่ i^{th} ในสองสถานการณ์ที่ใช้รูปแบบการทดสอบที่สามารถสับเปลี่ยนกันได้สองรูปแบบ

ถัดมาในปี ค.ศ. 1974 Swaminathan, Hambleton และ Algina (1974) ได้เสนอวิธีการของโคเฮน (Cohen's K) ไว้เป็นทางเลือกหนึ่งสำหรับดัชนีความสอดคล้องของการจำแนกประเภท เนื่องจากวิธีการนี้พิจารณาถึงความแตกต่างของโอกาสที่จะเกิดความเห็นพ้องต้องกันในแต่ละระดับความสามารถ สูตรของโคเฮน (Cohen's K) มีลักษณะดังนี้

$$\kappa = \frac{P_O - P_C}{1 - P_C}$$

P_o แทนสัดส่วนที่สังเกตได้ของผู้สอบที่ได้รับการจำแนกด้วยวิธีการเดียวกันกับทั้งสองสถานการณ์การทดสอบ P_c แทนระดับของความเห็นพ้องต้องกันที่คาดว่าจะเกิดขึ้นจากความน่าจะเป็น ค่าสัมประสิทธิ์สะท้อนให้เห็นถึงความน่าจะเป็นที่คาดหวังของความสอดคล้องของการจำแนกประเภท

ต่อมา Huynh (1976), Subkoviak (1976, 1988) และ Livingston และ Lewis (1995) ได้เสนอวิธีการสำหรับการประมาณค่าความสอดคล้องของการจำแนกประเภทในรูปแบบที่มีการบริหารจัดการการทดสอบเดียว วิธีการแรกๆ ที่นำเสนอโดย Huynh (1976) และ Subkoviak (1976, 1988) อยู่ภายใต้ข้อตกลงเบื้องต้นที่ว่าข้อสอบต้องมีความยากที่เท่าเทียมกัน โมเดลพื้นฐานที่นำเสนอโดย Huynh ได้รับการขยายให้กว้างออกไปในภายหลังโดย Hanson และ Brennan (1990) โดยใช้การแจกแจงของเบต้าแบบสี่พารามิเตอร์ (four-parameter beta distribution) ที่ซับซ้อนมากกว่ากับโมเดลความสามารถพื้นฐานต่างๆ เช่น คะแนนจริง

Breyer และ Lewis (1994 cited in Clauser, Margolis, & Case, 2006) ใช้วิธีแบ่งครึ่งข้อสอบ (split-half) เพื่อประมาณค่าดัชนีความสอดคล้องของการจำแนกประเภท โดยทำการแบ่งแบบสอบออกเป็นสองส่วน คะแนนจุดตัด (cut score) ของข้อสอบครึ่งหนึ่งจะถูกนำไปใช้กับข้อสอบทั้งสองส่วนที่แบ่งไว้ และทำการเปรียบเทียบค่าที่ได้จากทั้งสองส่วน จากนั้นนำค่าการประมาณที่ได้ขยายให้ครอบคลุมถึงความยาวของแบบสอบเต็มฉบับ จากนั้นในปี 1995 Livingston และ Lewis (Clauser, Margolis & Case, 2006) ได้เสนอวิธีการที่ใช้โมเดลเบต้าแบบสี่พารามิเตอร์ขึ้น (four-parameter beta model) ซึ่งแตกต่างจากวิธีการที่มีมาก่อนหน้าหลายด้าน รวมถึงการที่ได้พัฒนาวิธีการประมาณค่าสำหรับทั้งความถูกต้องและความสอดคล้องของการจำแนกประเภทไว้อีกด้วย

วิธีการทั้งหลายที่กล่าวถึงข้างต้นแตกต่างกันที่ระดับความซับซ้อนในการคำนวณของแต่ละวิธีการ ในช่วงปี ค.ศ. 1970 เป็นประเด็นสำคัญที่ขัดขวางต่อการพัฒนาวิธีการประมาณค่าต่างๆ คือจำนวนคอมพิวเตอร์สมรรถภาพสูงที่มีอย่างจำกัด แม้ว่าการประเมินเกี่ยวกับการประมาณค่าเหล่านี้จะแสดงให้เห็นถึงประโยชน์ที่ได้ก็ตาม แต่ในปัจจุบันความสามารถในการคำนวณไม่ได้เป็นปัญหาที่สำคัญอีกต่อไป ข้อจำกัดหนึ่งของความสอดคล้องของการจำแนกประเภทในฐานะที่เป็นดัชนีที่มีประโยชน์ต่อแบบสอบสำหรับการตัดสินใจในการจำแนกประเภทคือ ดัชนีนี้มักมีค่าสูงแม้ว่าการจำแนกประเภทของแบบสอบที่ใช้เป็นแบบสุ่มก็ตาม ตัวอย่างเช่น ถ้า 10% ของผู้สอบไม่ผ่านการทดสอบในสถานการณ์ใดสถานการณ์หนึ่งจากสองสถานการณ์ แม้ว่าการจำแนกประเภทต่างๆ เป็นแบบสุ่มสมบูรณ์ 82% ของผู้สอบจะถูกคาดหวังให้ได้รับการจำแนกให้สอดคล้องกับอีกสถานการณ์หนึ่ง (Clauser, Margolis & Case, 2006)

วิธีการประมาณค่าความสอดคล้องของการจำแนกประเภทที่นักวัดผลหลายท่านได้ทำการศึกษาและพัฒนาขึ้นสามารถอธิบายรายละเอียดของแต่ละวิธีได้ดังนี้

1) Huynh's method

วิธีการของ Huynh (1976) ใช้การแจกแจงเบต้า (beta distribution) ของพารามิเตอร์ α และ β สำหรับคะแนนจริง (true scores) และการแจกแจงเบต้าแบบทวินามของตัวแปรสองตัว (bivariate beta-binomial distribution) สำหรับคะแนนที่สังเกตได้ (observed scores) ให้ x และ y เป็นคะแนนสอบที่ได้มาจากแบบสอบสองฉบับที่คู่ขนานกัน X และ Y ภายใต้ข้อตกลงเบื้องต้นเกี่ยวกับความเป็นอิสระระหว่างข้อสอบ (local independence) x และ y เป็นผลมาจากการแจกแจงเบต้าแบบทวินามของตัวแปรสองค่า (bivariate beta-binomial distribution) ด้วยความหนาแน่นของความน่าจะเป็นร่วม (joint proability) ดังนี้ (Huynh, 1976)

$$f(x,y)=\frac{\binom{n}{x}\binom{n}{y}}{B(\alpha,\beta)}B(\alpha+x+y,2n+\beta-x-y) \tag{1}$$

เมื่อ B คือ ฟังก์ชันของเบต้าที่มีพารามิเตอร์ α และ β และ n คือจำนวนข้อสอบทั้งหมด สมมติให้ C เป็นคะแนนจุดตัด (cut score) ที่ใช้แบ่งผู้สอบเข้าสู่กลุ่มที่แบ่งไว้สองกลุ่ม ค่าดัชนีความสอดคล้องของการจำแนกประเภท (classification consistency index: P) สามารถคำนวณได้ดังนี้

$$\begin{aligned} P &= P(x\le C-1, y\le C-1)+P(x\ge C, y\ge C) \\ &= \sum_{x=0}^{C-1}\sum_{y=0}^{C-1}P(x,y)+\sum_{x=C}^{n}\sum_{y=C}^{n}P(x,y) \end{aligned} \tag{2}$$

ต่อมา Hanson และ Brennan (1990) ได้ขยายโมเดลนี้โดยการนำไปใช้กับการแจกแจงเบต้าแบบสี่พารามิเตอร์ (four-parameter beta distribution) ซึ่งเป็นหลักการโดยทั่วไปของการแจกแจงเบต้าแบบสองพารามิเตอร์ (two-parameter beta distribution) โดยนอกจากพารามิเตอร์ α และ β แล้ว ยังได้เพิ่มอีกสองพารามิเตอร์คือขีดจำกัดล่าง (lower limits of the distribution: a) และขีดจำกัดบน (upper limits of the distribution: b) ของการแจกแจงเข้าไปด้วย คะแนนจริง (true score: T) เป็นผลมาจากการแจกแจงด้วยความหนาแน่นดังนี้

$$g(T\mid\alpha,\beta,a,b)=\frac{1}{B(\alpha+1,\beta+1)}\frac{(T-a)^{\alpha}(b-T)^{\beta}}{(b-a)^{\alpha+\beta+1}} \tag{3}$$

จากสมการพบว่าโมเดลเบต้าแบบทวินาม (beta-binomial model) โดยทั่วไปจะเหมาะสมกับการแจกแจงของคะแนนที่สังเกตได้มากกว่า ซึ่งเป็นสิ่งที่คาดหวังจะให้เกิดขึ้นเนื่องจากพารามิเตอร์ที่เพิ่มเข้าไปนั้นสามารถใช้ประโยชน์ในการหาการแจกแจงที่เหมาะสมที่สุดได้

วิธีการที่ใช้แนวคิดของโมเดลเบต้าแบบทวินาม (beta-binomial model) คือคุณสมบัติที่ดีทางคณิตศาสตร์ ซึ่งได้รับการค้นพบว่าวิธีการเหล่านั้นมีความคลาดเคลื่อนมาตรฐานที่เกิดขึ้นเพียงเล็กน้อย (Subkoviak, 1976) นอกจากนี้ยังพบว่าการฝ่าฝืนข้อตกลงเบื้องต้นในเรื่องความยากของข้อสอบที่ทัดเทียมกันนั้นส่งผลต่อการประมาณค่าเพียงเล็กน้อย

2) Subkoviak's method

วิธีการของ Subkoviak (1976) เป็นวิธีที่คล้ายกับวิธีการของ Huynh โดยมีการกำหนดโมเดลทวินาม (binomial model) ในการแจกแจงของคะแนนที่สังเกตได้ (observed score distributions) แต่แทนที่จะสร้างข้อตกลงเบื้องต้นเกี่ยวกับการแจกแจงของคะแนนจริง (true scores) Subkoviak กลับประมาณค่าดัชนีความสอดคล้องสำหรับผู้สอบแต่ละคนในแต่ละช่วงเวลา และทำการหาค่าเฉลี่ยของผู้สอบทั้งหมด โดยเฉพาะอย่างยิ่งการประมาณค่าสัดส่วนที่ถูกต้องของคะแนนจริงของผู้สอบแต่ละคนด้วยการนำไปประมาณค่าถดถอยเชิงเส้น (linear regression approximation) โดยใช้สัดส่วนที่ถูกต้องของคะแนนที่สังเกตได้ของผู้สอบ และสัมประสิทธิ์ KR-20 มีการสร้างการแจกแจงของคะแนนที่สังเกตได้แบบมีเงื่อนไขขึ้น (conditional observed score distribution) ภายใต้การประมาณค่าคะแนนจริง (true score) และโมเดล ทวินาม (binomial model) ค่าดัชนีความสอดคล้อง (consistency index) ผู้สอบแต่ละคนจะได้รับการคำนวณและนำดัชนีความสอดคล้องของผู้สอบทุกคนหาค่าเฉลี่ย ในทางคณิตศาสตร์ค่าดัชนีความสอดคล้อง (consistency index) สำหรับผู้สอบคนที่ i จะกำหนดให้เป็น $P_c^{(i)}$ ดังนี้

$$P_c^{(i)} = P(x_i \geq C)^2 + [1 - P(x_i \geq C)]^2 \quad (1)$$

เมื่อ

$$P(x_i \geq C) = \sum_{x_i=C}^{n} \binom{n}{x_i} \hat{\pi}_i^{x_i} (1 - \hat{\pi}_i)^{n-x_i} \quad (2)$$

เมื่อ $\hat{\pi}_i$ คือสัดส่วนที่ถูกต้องของคะแนนจริงที่ได้จากการประมาณค่าสำหรับผู้สอบคนที่ i, x_i คือคะแนนที่สังเกตได้ของผู้สอบ, C คือคะแนนจุดตัด และ n คือจำนวนข้อสอบทั้งหมด $P(x_i \geq C)$ คือความน่าจะเป็นที่ผู้สอบจะได้คะแนนเท่ากับหรือมากกว่าคะแนนจุดตัด C ค่าดัชนีความสอดคล้องของการจำแนกประเภท (classification consistency index (P)) สูงสุดคือค่าเฉลี่ยของ $P_c^{(i)}$ ระหว่างผู้สอบทั้งหมด อย่างไรก็ตามวิธีการนี้ก็ยังยากที่จะนำไปใช้ได้ในกรณีที่เป็นแบบสอบสั้น เนื่องจากจะนำไปสู่

การประมาณค่าคะแนนขอบเขต (domain scores) ที่ไม่ดีนัก และเมื่อประมาณค่าได้เป็น 0 หรือ 1 การประมาณค่าความสอดคล้องของการจำแนกประเภทที่ได้จะสูงเกินไป

Lee และคณะ (2002) จึงได้เสนอโมเดลอเนกนามเชิงซ้อน (compound multinomial model: CM model) สำหรับแบบสอบที่ประกอบด้วยข้อสอบที่มีทั้งข้อสอบที่ให้คะแนนได้สองค่า (dichotomous items) และข้อสอบที่ให้คะแนนได้หลายค่า (polytomous items) วิธีการอเนกนามเชิงซ้อน สามารถมองได้ว่าเป็นชุดทั่วไปของวิธีการของ Subkoviak ในแง่ที่ว่าวิธีการนี้จะลดขั้นตอนของ Subkoviak เมื่อข้อสอบทั้งหมดมีการให้คะแนนแบบสองค่า วิธีการลดความลำเอียงของ Brennan และ Lee (2006) มีการนำไปใช้ในการสร้างการแจกแจงของคะแนนที่สังเกตได้ที่มีปริมาณความแปรปรวนเท่ากับความแปรปรวนที่มีในการแจกแจงของคะแนนจริง

ในปี 2004 Brennan และ Wan (2004) ได้ขยายวิธีการของ Subkoviak โดยการพัฒนาวิธีการ bootstrap ขึ้นมา วิธีการของ Brennan และ Wan นี้มีกรอบแนวคิดที่สัมพันธ์กับวิธีการของ Subkoviak ในแง่ของการที่ไม่ต้องสร้างข้อตกลงเบื้องต้นเกี่ยวกับการแจกแจงความสามารถที่แท้จริงเหมือนกัน ในทางกลับกันวิธีการนี้ได้สร้างจำนวนของการทำซ้ำขนาดใหญ่ (ที่เรียกว่า bootstrap sampling) และคำนวณสัดส่วนของการตัดสินใจที่สอดคล้องกันสำหรับผู้สอบแต่ละคน แล้วจึงหาค่าเฉลี่ยของผู้สอบทั้งหมด วิธีการ bootstrap ได้รับการกล่าวถึงว่าเป็นวิธีที่ง่ายและยืดหยุ่นมากสำหรับการประเมินที่ซับซ้อน เมื่อการแจกแจกของคะแนนที่สังเกตได้ยากต่อการประมาณค่า Wan, Brennan และ Lee (2007) ค้นพบว่าวิธีการอเนกนามเชิงซ้อน (compound multinomial procedure) และวิธีการ bootstrap (bootstrap procedure) ให้ค่าการประมาณที่ใกล้เคียงกันมาก และทั้งสองวิธีการเป็นวิธีการที่ขยายเพิ่มเติมมาจากวิธีการของ Subkoviak

3) Livingston and Lewis's method

Livingston และ Lewis (1995) เสนอแนวคิดที่เรียกว่า ความยาวของแบบสอบที่มีประสิทธิผล (effective test length) วิธีการนี้มีพื้นฐานอยู่บนโมเดลทวินาม (binomial model) ซึ่งจะสามารถนำไปใช้ได้กับแบบสอบที่มีข้อสอบแบบให้คะแนนได้หลายค่า หรือมีการถ่วงน้ำหนักไม่เท่ากัน เช่น แบบสอบที่ประกอบด้วยการรวมกันของข้อสอบแบบให้คะแนนได้หลายค่า (polytomous items) และข้อสอบแบบให้คะแนนได้สองค่า (dichotomous items) หรือแบบสอบที่ใช้คะแนนที่ประกอบขึ้นจากหลายส่วน (composite scores) วิธีการนี้มีข้อมูลที่จำเป็นต้องใช้ในการประมาณค่าคือ การแจกแจงของคะแนนที่สังเกตได้ สัมประสิทธิ์ความเที่ยงของแบบสอบ คะแนนสอบสูงสุดและต่ำสุดที่เป็นไปได้ และคะแนนจุดตัด

โดย Livingston และ Lewis ได้พัฒนาวิธีการประมาณค่าความสอดคล้องของการจำแนกประเภทตามทฤษฎีการทดสอบแบบดั้งเดิม (CTT) (Livingston & Lewis, 1995 cited in Zhang, 2010) โดยใช้ข้อมูลจากแบบสอบรูปแบบเดียวในการประมาณค่าความสอดคล้องของการจำแนก

ประเภทบนฐานของคะแนนสอบ ซึ่งวิธีการนี้ไม่เพียงใช้กับคะแนนสอบแบบที่พิจารณาด้วยการนับคำตอบถูกเท่านั้น แต่ยังใช้กับคะแนนสอบแบบอื่นๆ ที่สามารถประมาณค่าความเที่ยงได้อีกด้วย โดยมีการกำหนดค่าต่างๆ ในการคำนวณดังนี้

เมื่อกำหนดให้ X แทนคะแนนของผู้สอบจากแบบสอบฉบับที่ 1 ทำตัวเลขให้เป็นจำนวนเต็มที่ใกล้ที่สุด ในกรณีที่จำเป็นจะระบุคะแนนจากแบบสอบต่างฉบับกันเป็น X_0, X_1 และ X_2 โดย X_0 แทนรูปแบบของการทดสอบสำหรับข้อมูลที่มีอยู่ คะแนนต่ำสุดและสูงสุดที่เป็นไปได้แทนด้วย X_{min} และ X_{max} ซึ่ง X_{min} อาจจะมีค่าเป็นลบได้

สัมประสิทธิ์ความเที่ยง (reliability coefficient) ของคะแนนสอบแทนด้วย r การประมาณความยาวของแบบสอบที่เหมาะสมจะแทนด้วย n บนสเกลตั้งแต่ 0 ถึง 1 สัดส่วนของคะแนนแทนด้วย p ดังนั้น

$$p = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

เมื่อคะแนนจริงร่วมกับคะแนนที่สังเกตได้ (X) จะแสดงได้ด้วยสัดส่วนบนสเกลตั้งแต่ 0 ถึง 1 และแทนด้วย T_p ดังนั้น

$$T_p = \frac{E(X) - X_{\min}}{X_{\max} - X_{\min}} \tag{2}$$

โดยที่ T_p เป็นสัดส่วนของคะแนนจริง

บางส่วนของวิธีการประมาณค่าเกี่ยวข้องกับการปรับการให้คะแนนของแบบสอบฉบับใหม่จากการให้คะแนนแบบเดิม ซึ่งประเมินค่าจาก X_{min} ถึง X_{max} เป็นการให้คะแนนแบบใหม่ซึ่งประเมินค่าจาก 0 ถึง n คะแนนที่ปรับใหม่แทนด้วย X' ดังนั้น

$$X' = np = n\frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{3}$$

โดยมีตารางสรุปถึงการให้คะแนนที่ใช้ในการประมาณค่าดังนี้

ตารางที่ 2.3 การให้คะแนนที่ใช้ในการประมาณค่า

ชื่อ (Name)	**สัญลักษณ์ (Symbol)**	**ช่วง (Range)**	**อันตรภาค (Interval)**
คะแนนจากแบบสอบฉบับเดียว (Single-form score)	X (X_0, X_1, X_2)	X_{min} ถึง X_{max}	1
สัดส่วนของคะแนนจากแบบสอบฉบับเดียว (Proportional single-form score)	p	.00 ถึง 1.00	.01
คะแนนปรับใหม่จากแบบสอบฉบับเดียว (Transformed single-form score)	X'	0 ถึง n	1
สัดส่วนของคะแนนจริง (Proportional true score)	T_p	.00 ถึง 1.00	.01

เมื่อกำหนดค่าต่างๆ แล้วก็สามารถนำไปใช้ในการประมาณค่า โดยมีขั้นตอนในการประมาณค่าดังต่อไปนี้

3.1) ประมาณค่าความยาวของแบบสอบที่มีประสิทธิผล (n)

3.2) ประมาณค่าการแจกแจงของคะแนนจริงที่พอเหมาะ (T_p)

แปลงการแจกแจงแบบต่อเนื่อง (continuous distribution) ให้เป็นการแจกแจงแบบไม่ต่อเนื่อง (discrete distribution) ด้วยการแบ่งช่วง (0,1) ให้เป็นระดับอันตรภาค (intervals) ด้วยขนาด .01 และคำนวณสัดส่วนของการแจกแจงในแต่ละระดับของ T_p ให้ g_i แทนการประมาณค่าความถี่ที่พอเหมาะสำหรับระดับของคะแนนจริงที่ i (i^{th}) และให้ t_i แทนคะแนนตรงจุดกึ่งกลางของช่วง

3.3) ประมาณค่าการแจกแจงแบบมีเงื่อนไขของการจำแนกประเภทในแบบสอบรูปแบบอื่นสำหรับผู้สอบในแต่ละระดับของคะแนนจริง

สร้างการแจกแจงความน่าจะเป็น (probability distribution) ของคะแนนในการทดสอบสมมติฐานของผู้สอบที่เป็นอิสระกัน (n) โดยสร้างการแจกแจงแบบทวินาม (binomial distribution) ด้วยพารามิเตอร์ n และ t_i ในแต่ละระดับของการแจกแจงของคะแนนจริง (T_p) กำหนดให้ t_i คือข้อสอบที่ให้คะแนนแบบสองค่าสำหรับผู้สอบที่มีความน่าจะเป็นของการตอบถูกในแต่ละข้อ ให้ X' แทนคะแนนของตัวแปรที่มีการแจกแจงแบบทวินาม (binomial distribution) หาการแจกแจงความน่าจะเป็นสะสม (cumulative probability distribution) ของ X'ที่แต่ละระดับของการแจกแจงของคะแนนจริง (T_p) แปลงการแจกแจงนี้เป็นสเกล X ใช้สมการ 3 และการประมาณค่าแบบช่วงเชิงเส้น (linear interpolation) แปลงการแจกแจงความน่าจะเป็นสะสมที่เปลี่ยนรูปแล้วให้เป็น

การแจกแจงความน่าจะเป็นสำหรับ X ที่กำหนดให้ ใช้ขอบเขตของกลุ่มในการกำหนดการแจกแจงของคะแนนจริง (T_p) สำหรับแต่ละระดับของคะแนนจริง (true-score) ความน่าจะเป็นแบบมีเงื่อนไขว่าผู้สอบที่มีคะแนนจริง t_i เมื่อทำแบบสอบฉบับอื่นจะได้รับการจำแนกให้อยู่ในแต่ละกลุ่ม ให้ X_1 แทนคะแนนของผู้สอบจากแบบสอบฉบับอื่น และให้ x_j^* แทนคะแนนจุดตัดที่ j (j^{th}) ให้ x_j เป็นจำนวนเต็มที่ใกล้ x_j^* ที่สุด (โดยที่ x_j คือจำนวนเต็ม ดังนั้น $x_j - 0.5 < x_j^* \leq x_j + 0.5$) ดังนั้นความน่าจะเป็นที่คะแนนของผู้สอบจะต่ำกว่าคะแนนจุดตัดที่ j (j^{th}) ในแต่ละระดับของการแจกแจงของคะแนนจริง (T_p) จะกำหนดได้ด้วยสมการที่แก้ไขแล้วดังนี้

$$\hat{P}\left(X < x_j^* \mid T_p = t_i\right) = P\left(X < x_j \mid T_p = t_i\right) + \frac{x_j^* - \left(x_j - 0.5\right)}{1.0} P\left(X = x_j \mid T_p = t_i\right) \qquad (4)$$

3.4) ประมาณค่าการแจกแจงร่วม (joint distribution) ของการจำแนกประเภทในแบบสอบที่ต่างกันสองฉบับ

ในแต่ละระดับของ T_p ใช้การจำแนกแบบมีเงื่อนไขบน X_1 (จากขั้นตอนที่ 3.3) เพื่อกำหนดเงื่อนไขในการจำแนกแบบสองทางบน X_1 และ X_2 เมื่อ X_1 และ X_2 คือคะแนนจากแบบสอบสองฉบับที่เป็นรูปแบบอื่น (ที่นอกเหนือจากรูปแบบที่ใช้ในการบริหารจัดการการทดสอบ) สมมติฐานคือการจำแนกนั้นมีพื้นฐานอยู่บน X_1 และ X_2 ที่เป็นอิสระต่อกัน (independent) และมีการแจกแจงเหมือนกัน (identically distribute) สำหรับผู้สอบที่ให้ระดับของ T_p (เช่น ให้อันตรภาคที่มีขนาด .01) ดังนั้นสัดส่วนของความถี่ในแต่ละเซลล์เงื่อนไขของการจำแนกแบบสองทางคือผลลัพธ์ของสัดส่วนแบบกำหนดขอบเขต (marginal proportions) สำหรับแต่ละชุดของกลุ่มบน X_1 และ X_2 ผลรวมความถี่ของเซลล์สูงกว่าระดับของ T_p เพื่อให้ได้ผลการประมาณค่าการจำแนกแบบสองทางที่ขึ้นอยู่กับ X_1 และ X_2 สำหรับประชากรผู้สอบทั้งหมด

3.5) ประมาณค่าการแจกแจงร่วม (joint distribution) ของการจำแนกประเภทในแบบสอบฉบับอื่นและแบบสอบที่ใช้ในการบริหารจัดการการทดสอบจริง

เมื่อปรับการประมาณค่าการจำแนกแบบสองทางที่ขึ้นอยู่กับ X_1 และ X_2 (ผลลัพธ์จากขั้นตอนที่ 3.4) เพื่อที่จะทำให้ความถี่จำนวนเล็กน้อยที่ไม่สำคัญของ X_1 ตรงกับความถี่ที่สังเกตได้สำหรับ X_0 การปรับประกอบด้วยการกำหนดตัวคูณ อัตราส่วนของความถี่ที่สังเกตได้ไปยังความถี่ที่คาดหวัง สำหรับแต่ละกลุ่มของ X_1 และนำไปใช้กับความถี่ของเซลล์ทั้งหมดสำหรับกลุ่มนั้นๆ ของ X_1 การปรับการจำแนกแบบสองทางเป็นการประมาณค่าการจำแนกแบบสองทางที่ขึ้นอยู่กับ X_0 และ X_2 และเป็นพื้นฐานสำหรับการประมาณค่าทางสถิติที่ใช้อธิบายความสอดคล้องของการจำแนกประเภท

4) Rudner's method

Wyse และ Hao (2012) ได้นำเสนอวิธีการประมาณค่าดัชนีความสอดคล้องของการจำแนกประเภทใหม่ โดยประยุกต์มาจากวิธีการประมาณค่าดัชนีความถูกต้องของการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบของ Rudner ซึ่ง Wyse และ Hao (2012) ได้อธิบายถึงวิธีการประมาณค่าดัชนีความถูกต้องของการจำแนกประเภทไว้ดังนี้

วิธีการประมาณค่าดัชนีความถูกต้องของการจำแนกประเภทของ Rudner (Rudner, 2001, 2005 cited in Wyse & Hao, 2012) ใช้ข้อมูลเวกเตอร์ในการคำนวณความถูกต้องของการจำแนกประเภทจำนวนสามเวกเตอร์ โดยเวกเตอร์แรกเป็นเวกเตอร์ของคะแนนจุดตัด (C+1)

$$\kappa = \left[\kappa_1 \kappa_2 \cdots \kappa_{C+1}\right] \tag{1}$$

เมื่อ $\kappa_1 < \kappa_2 < \cdots < \kappa_{C+1}$

และ $\kappa_1 = -\infty, \kappa_{C+1} = \infty$

เวกเตอร์ของคะแนนจุดตัดนี้ประกอบด้วยจำนวนจุดตัดในการประเมิน และช่วงต่ำสุดและสูงสุดสำหรับแต่ละกลุ่มความสามารถ ตัวอย่างเช่น ถ้ามีคะแนนจุดตัดจำนวน 3 จุดตัด เวกเตอร์ในสมการที่ (1) จะประกอบด้วย 3 คะแนนจุดตัดและค่าอนันต์เชิงบวก (∞) และลบ (-∞)

เวกเตอร์ที่สองคือเวกเตอร์ของคะแนนที่ประมาณค่าได้ของผู้สอบหรือเวกเตอร์ความสามารถของผู้สอบ แสดงได้ดังนี้

$$\hat{\theta} = \left[\hat{\theta}_1 \hat{\theta}_2 \cdots \hat{\theta}_{N_e}\right]' \tag{2}$$

เมื่อ N_e คือ จำนวนผู้สอบ และ $\hat{\theta}_i$ คือ ค่าความสามารถที่ประมาณค่าได้ตามทฤษฎีการตอบสนองข้อสอบสำหรับผู้สอบคนที่ i เวกเตอร์ในสมการที่ (2) ประกอบด้วยค่าความสามารถที่ประมาณค่าได้ของผู้สอบแต่ละคน

เวกเตอร์ที่สามคือเวกเตอร์ของค่าความคลาดเคลื่อนมาตรฐานในการประมาณค่าความสามารถของผู้สอบ (standard error estimates) เขียนได้ดังนี้

$$\hat{\sigma}_{\hat{\theta}} = \left[\hat{\sigma}_{\hat{\theta}_1} \hat{\sigma}_{\hat{\theta}_2} \cdots \hat{\sigma}_{\hat{\theta}_{N_e}}\right]' \tag{3}$$

เมื่อ N_e คือ จำนวนผู้สอบ และ $\hat{\sigma}_{\hat{\theta}_i}$ คือ ค่าความคลาดเคลื่อนมาตรฐานในการประมาณค่าความสามารถของผู้สอบตามทฤษฎีการตอบสนองข้อสอบสำหรับผู้สอบคนที่ i ความคลาดเคลื่อนมาตรฐานในสมการที่ (3) สามารถคำนวณได้จากฟังก์ชันสารสนเทศของแบบสอบตามทฤษฎีการตอบสนองข้อสอบ (test information function) ในกรณีนี้ความคลาดเคลื่อนมาตรฐานของผู้สอบแต่ละคนคือ

$$\hat{\sigma}_{\hat{\theta}i} = \frac{1}{\sqrt{I(\hat{\theta}_i)}} \tag{4}$$

เมื่อ I($\hat{\theta}_i$) คือค่าฟังก์ชันสารสนเทศของแบบสอบสำหรับผู้สอบคนที่ i

เนื่องด้วยวิธีการของ Rudner มีข้อตกลงเบื้องต้นเกี่ยวกับการแจกแจงปกติของความคลาดเคลื่อนมาตรฐานในการประมาณค่าคะแนนจริง โดยใช้วิธีการประมาณค่าแบบ ความเป็นไปได้สูงสุด (maximum likelihood: ML) ทำให้ความน่าจะเป็นที่คาดหวัง (expected probability) ของการให้คะแนนในแต่ละกลุ่มระดับความสามารถ (C) เป็นดังนี้

$$\hat{p}_{iC} = \phi\left(\kappa_{C_i}, \kappa_{C_{i+1}}, \hat{\theta}_i, \hat{\sigma}_{\theta_i}\right) \tag{5}$$

เมื่อ $\hat{p}_{iC}$ คือ ความน่าจะเป็นในการจำแนกผู้สอบที่มีความสามารถ θ เข้าสู่กลุ่มความสามารถ

ϕ คือ พื้นที่ใต้โค้งปกติ

κ คือ คะแนนจุดตัด

$_{C_i}$ คือ ตำแหน่งของคะแนนจุดตัดที่ i

ϕ คือ พื้นที่ใต้โค้งปกติ

$\hat{\theta}_i$ คือ ความสามารถที่ประมาณค่าได้ของผู้สอบคนที่ i

$\hat{\sigma}_{\theta_i}$ คือ ความคลาดเคลื่อนในการประมาณค่าความสามารถของคนที่ i

โดยที่ $\phi(a, b, \mu, \sigma)$ คือพื้นที่ใต้โค้งปกติจาก a ถึง b ด้วยค่าเฉลี่ย (μ) และส่วนเบี่ยงเบนมาตรฐานของ (σ)

ความสอดคล้องของการจำแนกประเภทซึ่งให้การประมาณค่าสัดส่วนของผู้สอบที่ถูกจำแนกให้อยู่ในกลุ่มระดับความสามารถเดียวกันจากการทดสอบซ้ำที่คู่ขนานกันนั้น เกี่ยวข้องกับการใช้ผลลัพธ์ของเมตริกซ์ $\hat{P}$ กับ $\hat{P}$ และไม่ได้เกี่ยวข้องกับเมตริกซ์ที่ใช้กำหนดระดับความสามารถที่สังเกตได้ของ

ผู้สอบ ดังนั้นสูตรของดัชนีความสอดคล้องของการจำแนกประเภทใหม่ ($\hat{\gamma}$) ที่ Wyse และ Hao (2012) ได้พัฒนาขึ้นจึงเป็นดังนี้

$$\hat{\gamma} = \frac{\Sigma\left(\hat{\mathrm{P}} * \hat{\mathrm{P}}\right)}{\mathrm{N}_e} \tag{6}$$

เมื่อ $\hat{\gamma}$ คือ ดัชนีความสอดคล้องของการจำแนกประเภท

$\hat{P}$ คือ เมตริกซ์ $N_e \times C$ ของความน่าจะเป็นที่คาดหวัง

N_e คือ จำนวนผู้สอบ

5) Guo's method

Wyse และ Hao (2012) ได้นำเสนอวิธีการประมาณค่าดัชนีความสอดคล้องของการจำแนกประเภทวิธีการใหม่ โดยประยุกต์มาจากวิธีการประมาณค่าดัชนีความถูกต้องของการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบของ Guo (Guo, 2006 cited in Wyse & Hao, 2012) ซึ่งไม่ได้สร้างข้อตกลงเบื้องต้นเกี่ยวกับการแจกแจงปกติของความคลาดเคลื่อนมาตรฐานในการประมาณค่าคะแนนจริง แต่ใช้การคำนวณความน่าจะเป็นของการจำแนกที่คาดหวังและเมตริกซ์ $\hat{P}$ ที่อยู่บนฐานของฟังก์ชันความน่าจะเป็นของผู้สอบแต่ละคนจากโมเดลการตอบสนองข้อสอบ (IRT model) จะเห็นได้ว่ารูปแบบในการพัฒนาวิธีการประมาณค่าความสอดคล้องของการจำแนกประเภทนี้มีความคล้ายคลึงกับรูปแบบในการพัฒนาวิธีการประมาณค่าตามวิธีการของ Rudner แต่ต่างกันที่ข้อตกลงเบื้องต้นเกี่ยวกับการแจกแจงของความคลาดเคลื่อนมาตรฐานในการประมาณค่าคะแนนจริง

ฟังก์ชันความน่าจะเป็นสำหรับข้อสอบแบบให้คะแนนได้สองค่าเป็นดังนี้

$$L\left(u_{1i}, u_{2i}, \ldots, u_{ni} \mid \theta\right) = \prod_{j=1}^{n} P_{ij}^{u_{ij}} Q^{1-u_{ij}} \tag{1}$$

เมื่อ i คือ ผู้สอบ

j คือ ข้อสอบที่อยู่ในแบบสอบ

u_{ij} คือ การตอบข้อสอบข้อที่ j ของผู้สอบคนที่ i โดยให้แทนเป็น 1 ถ้าตอบข้อสอบถูก และเป็น 0 ถ้าตอบข้อสอบผิด

P_{ij} คือ ความน่าจะเป็นของการตอบข้อสอบข้อที่ j ถูก

Q_{ij} คือ ความน่าจะเป็นของการตอบข้อสอบข้อที่ j ผิด คำนวณได้จาก 1 - P_{ij}

ความน่าจะเป็นที่คาดหวังของการจำแนกผู้สอบเข้าสู่แต่ละกลุ่มระดับความสามารถหาได้โดยใช้ฟังก์ชันความน่าจะเป็นดังนี้

$$\hat{p}_{ic} = \frac{\sum_{\theta=\kappa_{C_i}}^{\kappa_{C_{i+1}}} L(u_{1i}, u_{2i}, \ldots, u_{ni} \mid \theta)}{\sum_{h=1}^{C+1} \sum_{\theta=\kappa_h}^{\kappa_{h+1}} L(u_{1i}, u_{2i}, \ldots, u_{ni} \mid \theta)} \quad (2)$$

ในการคำนวณความน่าจะเป็นที่คาดหวังสำหรับกลุ่มระดับความสามารถสูงสุดและต่ำสุดนั้น คะแนนจุดตัดสูงสุดและต่ำสุดในสมการที่ (1) จากวิธีการของ Rudner จำเป็นต้องนำมากำหนดค่า θ สูงสุดและต่ำสุดก่อน เช่น θ = 6 และ -6 ดังนั้นเวกเตอร์ของคะแนนจุดตัดจึงแสดงได้ดังนี้

$$\kappa = [\kappa_1 \kappa_2 \cdots \kappa_{C+1}] \quad (3)$$

เมื่อ $\kappa_1 < \kappa_2 < \cdots < \kappa_{C+1}$

และ $\kappa_1 = -6, \kappa_{C+1} = 6$

จะเห็นได้ว่าการใช้การกำหนดค่าสูงสุดและค่าต่ำสุดสำหรับคะแนนจุดตัดแทนที่ค่าอนันต์เชิงบวกและลบนั้น เป็นข้อแตกต่างเพียงเล็กน้อยระหว่างวิธีการของ Guo กับวิธีการของ Rudner จากนั้นคำนวณดัชนีความสอดคล้องของการจำแนกประเภทได้ดังสมการที่ (6) ตามวิธีการของ Rudner ที่อธิบายไว้ข้างต้น

6) Lee's method

ในปี ค.ศ. 2002 Lee และคณะ (2002) ได้สร้างวิธีการประมาณค่าความสอดคล้องของการจำแนกประเภทตามทฤษฎีการทดสอบแบบดั้งเดิมขึ้น และต่อมาในปี 2010 Lee ได้พัฒนาวิธีการสำหรับแบบสอบที่มีการให้คะแนนโดยการสรุปรวมคะแนนจากข้อสอบทั้งหมดตามทฤษฎีการตอบสนองข้อสอบ (IRT) โดยมีความเชื่อพื้นฐานว่าโมเดลตามทฤษฎีการตอบสนองข้อสอบ (IRT) ที่เลือกมามีความเหมาะสมเป็นอย่างดี และพารามิเตอร์ของข้อสอบก็ได้รับการปรับเทียบมาเป็นอย่างดีเช่นเดียวกัน

โดย Lee และคณะ (2002) ได้ทำการศึกษาวิธีการประมาณค่าดัชนีความสอดคล้องและความถูกต้องสำหรับการจำแนกประเภท โดยกำหนดให้กระบวนการทดสอบต้องการวัดคุณลักษณะแฝงเป็น ϕ และให้ Φ แทนตัวแปรแฝงแบบสุ่ม (latent random variable) ให้ $g(\phi)$ และ ω แทนความหนาแน่น (density) และระยะห่าง (space) ของตัวแปรแฝงแบบสุ่ม (Φ) กำหนดให้ข้อมูลที่สร้างขึ้น

ประกอบด้วย คะแนนสอบ (test scores) แทนด้วยคะแนน x จากข้อคำถามที่มีการให้คะแนนแบบสองค่า (dichotomous) จำนวน K ข้อ โดยความน่าจะเป็นแบบกำหนดขอบเขตของคะแนนดิบ (marginal probability) ดังสูตร

$$\Pr(X = x) = \int_{\Omega} \Pr(X = x \mid \Phi = \phi) g(\phi) d\phi \,, x = 0,1,\ldots,K$$

การแจกแจงแบบกำหนดขอบเขต (marginal distribution: Pr(X=x)) ในที่นี้เท่ากับ f(x) และการแจกแจงความคลาดเคลื่อนแบบมีเงื่อนไข (conditional error distribution) คือ Pr(X = x| Φ = ϕ) เท่ากับ f (x| ϕ)

สถานการณ์การวัดปกติที่ใช้คือผู้สอบแต่ละคนจะถูกจำแนกเข้าสู่กลุ่มหนึ่งกลุ่มใดใน H กลุ่ม โดยใช้คะแนนจุดตัดที่สังเกตได้ H-1 ได้แก่ c_1, c_2, ..., $c_{(H-1)}$ ผู้สอบที่มีคะแนนที่สังเกตได้มากกว่าหรือเท่ากับ 0 และน้อยกว่า c1 จะถูกจำแนกเข้าสู่กลุ่มแรก และดำเนินการเช่นนี้ไปเรื่อยๆ จนถึงกลุ่มที่ H^{th} โดยคะแนนการทดสอบระหว่าง $c_{(H-1)}$ และ K ให้ I_h (h = 1, 2, . . . , H) แทน h^{th} กลุ่มของผู้สอบ โดย $c_{(h-1)} \le x \le c_h - 1$ จะได้รับการจำแนก เมื่อ c_o=0 และ c_h=K+1 ดังนั้นความน่าจะเป็นแบบมีเงื่อนไข (conditional probability) และความน่าจะเป็นแบบกำหนดขอบเขตของแต่ละกลุ่ม (marginal probability) ใช้สูตรดังนี้

$$\Pr(X \in I_h \mid \Phi = \phi) = \sum_{x=c_{(h-1)}}^{c_h - 1} f(x \mid \phi), \; h = 1,2,\ldots,H$$

และ

$$\Pr(X \in I_h) = \int_{\Omega} \sum_{x=c_{(h-1)}}^{c_h - 1} f(x \mid \phi) g(\phi) d\phi \,, \; h = 1,2,\ldots,H$$

ในการศึกษาครั้งนี้ Lee และคณะ (2002) ได้กล่าวถึงดัชนีทั่วไปสองตัวคือ ดัชนีความเห็นพ้องต้องกัน (agreement index: P) และสัมประสิทธิ์แคปปา (coefficient kappa) (Cohen, 1960 cited in Lee et al., 2002) ค่า P คือ ผลรวมของสมาชิกบนเส้นทแยงมุมในตารางการณ์จร H×H และสัมประสิทธิ์แคปปา (k) คือ ดัชนีความสอดคล้องสำหรับโอกาสในการเกิดความเห็นพ้องต้องกัน ความน่าจะเป็นของความไม่สอดคล้องของการจำแนกประเภท (inconsistent classification) คือ 1-P นอกจากนั้นยังแยกคะแนนจุดตัด H-1 ไปยังข้อมูลตรงข้ามเพื่อนำไปใช้กับคะแนนจุดตัดทุกคะแนนในเวลาเดียวกัน ดังนั้นจะได้ดัชนีย่อย (subindices) H-1 ค่าจากการวิเคราะห์ตารางการณ์จรแบบสองทาง H-1 ตาราง ได้เป็น P_m และ K_m เมื่อ m = 1, 2, ..., H-1 ดัชนีย่อยเหล่านี้จะมีประโยชน์

เมื่อคะแนนจุดตัดใดจุดตัดหนึ่งได้รับการพิจารณาให้เป็นระดับคะแนนการผ่านที่น้อยที่สุด ดัชนีความสอดคล้องของการจำแนกประเภท P และ P_m จะสามารถนำมาคำนวณเงื่อนไขของคุณลักษณะแฝง (ϕ) ซึ่งจะทำให้ได้ข้อมูลที่เป็นประโยชน์ต่อผู้ใช้แบบสอบ

โดยมีขั้นตอนของการประมาณค่าดัชนีความสอดคล้องของการจำแนกประเภทดังนี้

ตามข้อตกลงของเงื่อนไขบนคุณลักษณะแฝง (ϕ) ตัวแปรสุ่มของคะแนนดิบ คือ x_1 และ x_2 สำหรับการบริหารจัดการการทดสอบสองสถานการณ์ที่เป็นอิสระต่อกันและมีการแจกแจงเหมือนกัน ดังนั้นการแจกแจงร่วมแบบมีเงื่อนไข (conditional joint distribution) ของ x_1 และ x_2 เป็นดังสมการ

$$f(x_1, x_2 \mid \phi) = f(x_1 \mid \phi) f(x_2 \mid \phi) \tag{1}$$

ขนาดของการแจกแจงร่วมแบบกำหนดขอบเขต (marginal joint distribution) ของ x_1 และ x_2 ได้จากการอินทิเกรทความน่าจะเป็นของสมการที่ (1) บนการแจกแจงของตัวแปรแฝงแบบสุ่ม (Φ)

$$f(x_1, x_2) = \int_{\Omega} f(x_1, x_2 \mid \phi) g(\phi) d\phi \tag{2}$$

การจำแนกที่สอดคล้องกันจะเกิดขึ้นถ้า x_1 และ x_2 ของผู้สอบที่อยู่ในกลุ่มเดียวกัน (I_h) ความน่าจะเป็นแบบมีเงื่อนไข (conditional probability) ของการตกอยู่ในกลุ่มเดียวกันบนโอกาสของการทดสอบสองครั้งเป็น

$$\Pr(X_1 \in I_h, X_2 \in I_h \mid \Phi = \phi) = \left[\sum_{x_1 = c_{(h-1)}}^{c_h - 1} f(x_1 \mid \phi) \right]^2, \quad h = 1,2,\ldots,H \tag{3}$$

ดังนั้นดัชนีความสอดคล้องของความเห็นพ้องต้องกัน (P) บนเงื่อนไขของคุณลักษณะแฝง (ϕ) คือ

$$P(\phi) = \sum_{h=1}^{H} \Pr(X_1 \in I_h, X_2 \in I_h \mid \Phi = \phi) \tag{4}$$

เมื่อแยกแต่ละจุดตัดของดัชนีย่อย (subindices) ของความเห็นพ้องต้องกัน (P) บนเงื่อนไขของคุณลักษณะแฝง (ϕ) คือ

$$P_m(\phi) = \left[\sum_{j=1}^{m} \Pr(X_1 \in I_j \mid \Phi = \phi) \right]^2 + \left[\sum_{j=m+1}^{H} \Pr(X_1 \in I_j \mid \Phi = \phi) \right]^2, m = 1,2,\ldots,H-1$$

(5)

ค่าขอบเขตของดัชนีความสอดคล้อง (agreement indices) คำนวณได้โดย

$$P = \int_{\Omega} P(\phi) g(\phi) d(\phi) \tag{6}$$

และ

$$P_m = \int_{\Omega} P_m(\phi) g(\phi) d(\phi), m = 1,2,\ldots,H-1 \tag{7}$$

สัมประสิทธิ์ P และ P_m แทนความน่าจะเป็นของผู้สอบที่ได้มาแบบสุ่ม ซึ่งได้รับการจำแนกให้อยู่ในกลุ่มที่สังเกตได้กลุ่มเดียวกันบนโอกาสของการทดสอบสองสถานการณ์ ความน่าจะเป็นของความไม่สอดคล้องกันของการจำแนกสามารถคำนวณได้โดยการหักออกจากความน่าจะเป็นของการจำแนกที่สอดคล้องกัน

สัมประสิทธิ์แคปปา (coefficient's kappa) ในภาพรวมเมื่อนำคะแนนจุดตัดไว้ด้วยกัน คือ

$$\kappa = \frac{P - P_c}{1 - P_c} \tag{8}$$

เมื่อ P_c แทนความน่าจะเป็นของการจำแนกที่สอดคล้องกันที่เกิดขึ้นโดยบังเอิญ การเกิดความเห็นพ้องต้องกันโดยบังเอิญคือผลรวมของกำลังสองของความน่าจะเป็นของการจำแนกแต่ละกลุ่ม เขียนได้เป็น

$$P_c = \sum_{h=1}^{H} \Pr(X_1 \in I_h) \Pr(X_2 \in I_h) = \sum_{h=1}^{H} \left[\Pr(X_1 \in I_h)\right]^2 \tag{9}$$

ความน่าจะเป็น P_c ถูกกำหนดภายใต้กระบวนการอย่างสุ่มสองกระบวนการ ซึ่งผู้สอบได้รับการกำหนดให้เข้าสู่กลุ่มตามกฎของความน่าจะเป็นแต่ละกลุ่ม เนื่องจาก $1/H \le P_c \le P$ ดังนั้น κ จึงเป็นการปรับสเกลอีกรูปแบบหนึ่งของ P สุดท้ายการแยกจุดตัดดัชนีย่อยของ κ ทำได้โดย

$$\kappa_m = \frac{P_m - P_{mc}}{1 - P_{mc}} \tag{10}$$

เมื่อ

$$P_{mc} = \left[\sum_{j=1}^{m} \Pr\left(X_1 \in I_j\right)\right]^2 + \left[\sum_{j=m+1}^{H} \Pr\left(X_1 \in I_j\right)\right]^2, m = 1,2,\ldots,H-1 \qquad (11)$$

ต่อมาในปี ค.ศ. 2010 Lee ได้ศึกษาวิธีการประมาณค่าดัชนีความสอดคล้องและความถูกต้องของการจำแนกประเภทสำหรับการประเมินที่ซับซ้อนตามทฤษฎีการตอบสนองข้อสอบ (IRT) โดยนำไปใช้กับข้อมูลจากการทดสอบจริงที่ประกอบไปด้วยข้อสอบแบบให้คะแนนสองค่า (dichotomous) และข้อสอบแบบให้คะแนนได้มากกว่าสองค่า (polytomous) ซึ่งมีขั้นตอนของการประมาณค่าดัชนีความสอดคล้องของการจำแนกประเภทดังนี้

ให้ $x_1, x_2, \ldots, x_{K-1}$ แทนชุดของคะแนนจุดตัดที่สังเกตได้ ซึ่งนำไปใช้ในการจำแนกผู้สอบเข้าสู่กลุ่มที่มีความสามารถพิเศษที่คล้ายคลึงกันจำนวน K กลุ่ม นั่นคือ ผู้สอบที่มีคะแนนที่สังเกตได้ต่ำกว่า x_1 จะถูกระบุให้อยู่ในกลุ่มแรก ผู้สอบที่มีคะแนนมากกว่าหรือเท่ากับ x_1 และน้อยกว่า x_2 จะถูกระบุให้อยู่ในกลุ่มที่สอง และเป็นแบบนี้ไปเรื่อยๆ กำหนดการแจกแจงแบบมีเงื่อนไขของคะแนนรวมและคะแนนจุดตัด ความน่าจะเป็นแบบมีเงื่อนไขของการจัดกลุ่มสามารถคำนวณได้จากผลรวมของความน่าจะเป็นแบบมีเงื่อนไขของคะแนนรวมสำหรับ x ทั้งหมดที่อยู่ในกลุ่ม h ดังนี้

$$p_\theta(h) = \sum_{x=x_{(h-1)}}^{x_h 1} \Pr\left(X = x \mid \theta\right) \qquad (1)$$

เมื่อ h = 1, 2, ..., K กลุ่มแรกจะประกอบด้วยคะแนนต่ำสุดที่เป็นไปได้ และกลุ่มสุดท้ายจะประกอบด้วยคะแนนสูงสุดที่เป็นไปได้

ดัชนีความสอดคล้องของการจำแนกแบบมีเงื่อนไข (ϕ_θ) สามารถกำหนดให้เป็นความน่าจะเป็นที่ผู้สอบจะมี θ ที่จะถูกจำแนกเข้าสู่กลุ่มเดียวกันในการบริหารจัดการการทดสอบที่เป็นอิสระกันด้วยรูปแบบของแบบสอบที่คู่ขนานกันสองฉบับ (Lee et al., 2002 cited in Lee, 2010) ดังนั้น ϕ_θ สามารถคำนวณได้ดังนี้

$$\phi_\theta = \sum_{h=1}^{K} \left[p_\theta(h)\right]^2 \qquad (2)$$

ดัชนีความสอดคล้องของการจำแนกประเภทแบบมีเงื่อนไขคำนวณค่าความสอดคล้องของการจำแนกสำหรับระดับ θ ที่แตกต่างกัน ดัชนีความสอดคล้องของการจำแนกประเภทแบบ marginal แทนได้ด้วย ϕ ดังนี้

$$\phi = \int_{-\infty}^{\infty} \phi_\theta g(\theta) d\theta \qquad (3)$$

ดัชนีอีกดัชนีหนึ่งที่เป็นที่รู้จักกันดี คือ สัมประสิทธิ์แคปปา (κ) คำนวณได้ดังนี้

$$\kappa = \frac{\phi - \phi_c}{1 - \phi_c} \tag{4}$$

เมื่อ ϕ_c คือความน่าจะเป็นที่เป็นไปได้ โดยทั่วไปความน่าจะเป็นที่เป็นไปได้จะคำนวณด้วย $\phi_c = \sum_{h-1}^{K} [p(h)]^2$ เมื่อ p(h) คือความน่าจะเป็นของกลุ่มแบบกำหนดของเขต (marginal category proability) ซึ่งได้มาจากคะแนนจุดตัดของ θ ดังนี้

$$E(X \mid \theta = \theta^*) = \sum_i \sum_j j \Pr(U_i = j \mid \theta = \theta^*) \tag{5}$$

เมื่อ θ^* คือคะแนนจุดตัดที่อยู่ในสเกลของ θ, U_i คือตัวแปรสุ่มที่แสดงถึงการตอบข้อสอบข้อที่ i, $Pr(U_i = j \mid \theta = \theta^*)$ คือความน่าจะเป็นแบบมีเงื่อนไขสำหรับคะแนน j สำหรับข้อสอบข้อที่ i ซึ่งขึ้นอยู่กับโมเดลตามทฤษฎีการตอบสนองข้อสอบ (IRT model) สมการที่ 5 เป็นการแปลงคะแนนจุดตัดของ θ ทั้งหมด $\theta^*_1, , ..., \theta^*_{K-1}$ ให้เป็นคะแนนจุดตัดของคะแนนรวม $x_1, x_2, ..., x_{K-1}$

ในการรวมคะแนนจริง (true scores) ทั้งหมดเข้าด้วยกันนั้น มีวิธีที่ใช้ดำเนินการสองวิธี วิธีแรกคือการใช้คะแนนและค่าน้ำหนักของการหาตำแหน่งหรือบริเวณที่ได้จากการประมาณค่า (the estimated quadrature points and weights) ซึ่งได้จากผลการปรับเทียบคะแนนตามทฤษฎีการตอบสนองข้อสอบ เรียกว่า D-method ซึ่งวิธีการนี้เกิดขึ้นภายใต้ข้อตกลงเบื้องต้นเกี่ยวกับการแจกแจงความสามารถที่แท้จริงของผู้สอบ วิธีการที่สองคือการคำนวณค่าดัชนีการจำแนกประเภทสำหรับผู้สอบแต่ละคนในแต่ละช่วงเวลา และค่าเฉลี่ยของประชากรทั้งหมด เรียกว่า P-method

จากที่กล่าวมา จะเห็นได้ว่าวิธีการประมาณค่าดัชนีความสอดคล้องของการจำแนกประเภทมีความแตกต่างกันที่เห็นได้อย่างชัดเจนคือแนวคิดพื้นฐานในการพัฒนาวิธีการประมาณค่า ซึ่งมีทั้งวิธีการตามทฤษฎีการทดสอบแบบดั้งเดิม (CTT) อันประกอบด้วยวิธีการที่พัฒนาโดย Huynh (1976), Subkoviak (1976, 1988) และ Livingtion และ Lewis (1995) และวิธีการที่อยู่ภายใต้แนวคิดของทฤษฎีการตอบสนองข้อสอบ (IRT) อันประกอบด้วยวิธีการที่พัฒนาโดย Rudner (2001, 2005), Guo (2006) และ Lee (2010) ซึ่งแต่ละวิธีการจะมีความแตกต่างกันในเรื่องของทฤษฎีการทดสอบข้อตกลงเบื้องต้น วิธีการคำนวณ และลักษณะของข้อมูลที่ใช้ในการประมาณค่าโดยแสดงได้ดังตารางต่อไปนี้

ตารางที่ 2.4 ลักษณะเฉพาะของวิธีการประมาณค่าความสอดคล้องของการจำแนกที่พัฒนาโดยนักวัดผลทางการศึกษา

วิธีการประมาณค่า	ทฤษฎีการทดสอบ		ข้อตกลงเบื้องต้น	ลักษณะข้อมูล					
				รูปแบบของข้อสอบ		รูปแบบของคะแนนสอบ		รูปแบบของคะแนนจุดตัด	
	CTT	IRT		Dichotomous item	Polytomous item	Raw score	Scale score	Raw score	Scale score
Huynh (1976)	✓		คะแนนที่สังเกตได้มีการแจกแจงเบต้าแบบทวินาม (beta-binomial distribution), คะแนนจริงมีการแจกแจงแบบเบต้า (beta distribution)	✓		✓		✓	
Subkoviak (1976,1988)	✓		คะแนนที่สังเกตได้มีการแจกแจงแบบทวินาม (binomial distribution)	✓		✓		✓	
Livington & Lewis (1995)	✓		คะแนนที่สังเกตได้มีการแจกแจงเบต้าแบบทวินาม (beta-binomial distribution), คะแนนจริงมีการแจกแจงแบบเบต้า เบต้าแบบทวินาม (beta distribution)	✓	✓	✓	✓	✓	
Rudner (2005)		✓	ความคลาดเคลื่อนของการประมาณค่าคะแนนจริงมีการแจกแจงแบบปกติ (normal distribution)	✓	✓		✓		✓
Guo (2006)		✓	ฟังก์ชันความน่าจะเป็นในการตอบข้อสอบของผู้สอบมีการแจกแจงแบบปกติ (normal distribution)	✓	✓		✓		✓
Lee (2010)		✓	คะแนนจริงมีการแจกแจงแบบทวินามเชิงซ้อน เบต้าแบบทวินาม	✓	✓		✓	✓	

วิธีการประมาณค่า	ทฤษฎีการทดสอบ		ข้อตกลงเบื้องต้น	ลักษณะข้อมูล					
				รูปแบบของข้อสอบ		รูปแบบของคะแนนสอบ		รูปแบบของคะแนนจุดตัด	
	CTT	IRT		Dichotomous item	Polytomous item	Raw score	Scale score	Raw score	Scale score
			(compound binomial distribution)						

3.2 วิธีการประมาณค่าดัชนีความถูกต้องของการจำแนกประเภท

ภายใต้ข้อตกลงเบื้องต้นของทฤษฎีการทดสอบแบบดั้งเดิม (Classical Test Theory: CTT) ความถูกต้องของการจำแนก (classification accuracy) จะเป็นฟังก์ชันเกี่ยวกับความเที่ยงของแบบสอบ การแจกแจงของคะแนน สัดส่วนของผู้สอบในแต่ละระดับความสามารถที่เหมาะสมกับมาตรฐานและตำแหน่งของคะแนนจุดตัด ซึ่งในกรอบแนวคิดของทฤษฎีการทดสอบแบบดั้งเดิมนั้น ผู้สอบที่มีระดับความสามารถที่เหมาะสมกับมาตรฐานคือผู้สอบที่มีคะแนนจริง (true score) เท่ากับหรือสูงกว่าคะแนนจุดตัดที่สัมพันธ์กับมาตรฐานนั้น ภายใต้กรอบแนวคิดนี้ความผิดพลาดของการจำแนก (classification errors) เกิดขึ้นเมื่อผู้สอบมีคะแนนจริงสูงกว่ากว่าคะแนนจุดตัด (cut score) บนสเกลของคะแนนจริง (true-score scale) และคะแนนที่สังเกตได้ (observed score) ต่ำกว่าคะแนนจุดตัดบนสเกลของคะแนนที่สังเกตได้ (observed-score scale) หรือเมื่อผู้สอบมีคะแนนจริงต่ำกว่าคะแนนจุดตัดบนสเกลของคะแนนจริงและคะแนนที่สังเกตได้สูงกว่าคะแนนจุดตัดบนสเกลของคะแนนที่สังเกตได้ (Young & Yoon, 1998)

เมื่อกำหนดให้ x เป็นคะแนนที่อยู่บนสเกลของคะแนนที่สังเกตได้, τ เป็นคะแนนที่อยู่บนสเกลของคะแนนจริง และให้ x_0 เป็นคะแนนจุดตัดที่อยู่บนสเกลของคะแนนที่สังเกตได้, τ_0 เป็นคะแนนจุดตัดที่อยู่บนสเกลของคะแนนจริง ในทางทฤษฎีแล้ว x จะเป็นค่าที่ไม่ต่อเนื่อง (discrete) ส่วน τ จะเป็นค่าที่ต่อเนื่อง (continuous) (Clauser, Margolis, & Case, (2006)) ตารางสำหรับผลการตัดสินผ่าน/ตกแบบสองคูณสองแสดงได้ดังนี้

ตารางที่ 2.5 ผลการตัดสินผ่าน/ตกแบบสองคูณสอง

		คะแนนจริง (True Scores)		ขอบเขต (Marginal)
		0	1	
คะแนนที่สังเกตได้ (Observed Scores)	0	P00	P01	P0•
	1	P10	P11	P1•
ขอบเขต (Marginal)		P•0	P•1	1

โดยแต่ละเซลล์ในตารางสามารถกำหนดและให้นิยามได้ดังนี้

$P00 = Pr(x < x_0$ และ $\tau < \tau_0)$

$P01 = Pr(x < x_0$ และ $\tau > \tau_0)$

$P10 = Pr(x \geq x_0$ และ $\tau < \tau_0)$

$P11 = Pr(x \geq x_0$ และ $\tau > \tau_0)$

P00 คือ ความน่าจะเป็นที่คะแนนที่สังเกตได้มีค่าน้อยกว่าคะแนนจุดตัดบนสเกลของคะแนนที่สังเกตได้ และคะแนนจริงมีค่าน้อยกว่าคะแนนจุดตัดบนสเกลของคะแนนจริง

P01 คือ ความน่าจะเป็นที่คะแนนที่สังเกตได้มีค่าน้อยกว่าคะแนนจุดตัดบนสเกลของคะแนนที่สังเกตได้ และคะแนนจริงมีค่ามากกว่าคะแนนจุดตัดบนสเกลของคะแนนจริง

P10 คือ ความน่าจะเป็นที่คะแนนที่สังเกตได้มีค่ามากกว่าหรือเท่ากับคะแนนจุดตัดบนสเกลของคะแนนที่สังเกตได้ และคะแนนจริงมีค่าน้อยกว่าคะแนนจุดตัดบนสเกลของคะแนนจริง

P11 คือ ความน่าจะเป็นที่คะแนนที่สังเกตได้มีค่ามากกว่าหรือเท่ากับคะแนนจุดตัดบนสเกลของคะแนนที่สังเกตได้ และคะแนนจริงมีค่ามากกว่าคะแนนจุดตัดบนสเกลของคะแนนจริง

และขอบเขต (Marginal) สามารถกำหนดและให้นิยามได้ดังนี้

$P0\bullet = Pr(x < x_0)$

$P1\bullet = Pr(x \geq x_0)$

$P\bullet0 = Pr(\tau < \tau_0)$

$P\bullet1 = Pr(\tau > \tau_0)$

P0• คือ ความน่าจะเป็นที่คะแนนที่สังเกตได้มีค่าน้อยกว่าคะแนนจุดตัดบนสเกลของคะแนนที่สังเกตได้

P1• คือ ความน่าจะเป็นที่คะแนนที่สังเกตได้มีค่ามากกว่าหรือเท่ากับคะแนนจุดตัดบนสเกลของคะแนนที่สังเกตได้

P•0 คือ ความน่าจะเป็นที่คะแนนจริงมีค่าน้อยกว่าคะแนนจุดตัดบนสเกลของคะแนนจริง

P•1 คือ ความน่าจะเป็นที่คะแนนจริงมีค่ามากกว่าคะแนนจุดตัดบนสเกลของคะแนนจริง

หรือจะอธิบายให้เข้าใจได้ง่ายขึ้นคือ P10 เป็นสัดส่วนของผู้สอบที่ผ่านบนสเกลของคะแนนที่สังเกตได้แต่ตกบนสเกลของคะแนนจริง หรือคะแนนผ่านแต่ความสามารถที่แท้จริงคือตก สำหรับทางการศึกษามีผู้นำไปใช้ คือ Hanson และ Brennan (1990 cited in Clauser, Margolis, & Case, 2006) และ Brennan (2004 cited in Clauser, Margolis, & Case, 2006) โดยได้ให้นิยามอัตราความผิดพลาดเชิงบวกเป็น fp; fp = P10 และอัตราความผิดพลาดเชิงลบ คือ fn; fn = P01

แรกเริ่มในช่วงต้นปี ค.ศ. 1990 ทั้งความสอดคล้องของการจำแนกประเภท (classification consistency) และความถูกต้องของการจำแนกประเภท (classification accuracy) ได้รับการพิจารณาอย่างละเอียด วิธีการแรกในการประมาณค่าดัชนีทั้งสองค่าสำหรับข้อสอบที่มีการให้คะแนนแบบสองค่า (dichotomously scored items) ถูกเสนอขึ้นโดย Hanson และ Brennan (1990) ต่อมา Livingston และ Lewis (1995) ได้ขยายวิธีการของ Hanson และ Brennan เพิ่มเติมสำหรับแบบสอบที่มีทั้งข้อสอบที่มีการให้คะแนนแบบสองค่า (dichotomously scored items) และข้อสอบที่มีการให้คะแนนได้หลายค่า (polytomously scored items) โดยใช้ความยาวของแบบสอบที่มีประสิทธิภาพ (effective test length) เพื่ออำนวยความสะดวกให้กับข้อสอบที่มีการให้คะแนนได้หลายค่า (polytomously scored items) จึงทำให้วิธีการนี้ดูเป็นวิธีที่มีความซับซ้อนในเชิงการคำนวณทางคณิตศาสตร์เป็นอย่างมาก

ดังนั้น Lee (2002) จึงได้เสนอวิธีการอเนกนามเชิงซ้อนขึ้น (compound multinomial procedure) ซึ่งพัฒนามาจากวิธีการของ Livingston และ Lewis โดยหลีกเลี่ยงการใช้ความยาวของแบบสอบที่มีประสิทธิภาพ (effective test length) ในการพัฒนาครั้งนี้ Lee ได้เสนอไว้ 2 โมเดล คือ โมเดลอเนกนาม (multinomial model) ที่สามารถใช้ประมาณค่าความสอดคล้องและความถูกต้องของการจำแนกประเภทสำหรับข้อสอบที่มีการให้คะแนนได้สองค่า (dichotomously scored items) และสำหรับข้อสอบที่มีการให้คะแนนได้หลายค่า (polytomously scored items) ที่มีแต้มของคะแนนเดียวกันระหว่างข้อสอบทั้งหมด ส่วนอีกโมเดลหนึ่งคือโมเดลอเนกนามเชิงซ้อน (compound multinomial model) ซึ่งสามารถใช้กับข้อสอบที่มีแต้มของคะแนนที่หลากหลายได้

ในทำนองเดียวกันก็มีการศึกษาเพื่อพัฒนาวิธีการประมาณค่าตามทฤษฎีการตอบสนองข้อสอบ (IRT) ด้วยเช่นกัน โดยในปี ค.ศ. 2001 Rudner (2001) ได้เสนอวิธีการสำหรับคำนวณค่าดัชนีความถูกต้องของการจำแนกประเภทสำหรับข้อสอบแบบให้คะแนนได้สองค่า (dichotomous items) ต่อมาในปี ค.ศ. 2005 Rudner ได้ศึกษาความถูกต้องของการจำแนกประเภทที่คาดหวัง (expected classification accuracy) โดยทำการขยายผลการศึกษาวิธีการประมาณค่าความถูกต้องของการจำแนกประเภทที่คาดหวังจากการประมาณค่าสำหรับการให้คะแนนแบบสองค่า (dichotomous items) มาศึกษากับการให้คะแนนแบบหลายค่า (polytomous items) โดยใช้การประมาณค่าแบบ categorical approach ซึ่งเป็นการสร้างตารางจำแนกประเภทของคะแนนจริง (true scores) และคะแนนที่คาดหวัง (expected scores) สำหรับข้อสอบที่ให้คะแนนแบบหลายค่า นอกจากนี้ Guo (2006) ยังได้ทำการศึกษาความถูกต้องของ การจำแนกประเภทที่คาดหวังเพิ่มเติมโดยใช้การแจกแจงแฝง (the latent distribution) ซึ่ง Guo ได้พัฒนาวิธีการประมาณค่าแบบวิธีการการแจกแจงแฝง (latent distribution method) โดยทำการศึกษาเปรียบเทียบกับวิธีการตามทฤษฎีการตอบสนองข้อสอบที่ Rudner พัฒนาขึ้นในปี ค.ศ. 2001 และ 2005

รายละเอียดของวิธีการประมาณค่าความถูกต้องของการจำแนกประเภทที่นักวัดผลดังกล่าวข้างต้นทำการศึกษาและพัฒนาขึ้นอธิบายได้ดังนี้

1) Livingston & Lewis's method

Livingston และ Lewis (1995) ได้พัฒนาวิธีการประมาณค่าความถูกต้องของการจำแนกประเภทแบบเดียวกับการประมาณค่าความสอดคล้องของการจำแนกประเภทคือพัฒนาบนพื้นฐานของทฤษฎีการทดสอบแบบดั้งเดิม (CTT) เช่นเดียวกันกัน โดยมีขั้นตอนในการประมาณค่าดังต่อไปนี้

1.1) ประมาณค่าความยาวของแบบสอบที่มีประสิทธิผล (n)

1.2) ประมาณค่าการแจกแจงของคะแนนจริงที่พอเหมาะ (T_p)

เปลี่ยนการแจกแจงแบบต่อเนื่อง (continuous distribution) ให้เป็นการแจกแจงแบบไม่ต่อเนื่อง (discrete distribution) ด้วยการแบ่งช่วง (0,1) ให้เป็นระดับอันตรภาค (intervals) ด้วยขนาด .01 และคำนวณสัดส่วนของการกระจายในแต่ละระดับของ T_p ให้ g_i แทนการประมาณค่าความถี่ที่พอเหมาะสำหรับระดับของคะแนนจริงที่ i (i^{th}) และให้ t_i แทนคะแนนตรงจุดกึ่งกลางของช่วง

1.3) ประมาณค่าการแจกแจงแบบมีเงื่อนไขของการจำแนกประเภทในแบบสอบรูปแบบอื่น สำหรับผู้สอบในแต่ละระดับของคะแนนจริง

สร้างการแจกแจงความน่าจะเป็น (probability distribution) ของคะแนนในการทดสอบสมมติฐานของผู้สอบที่เป็นอิสระกัน (n) โดยสร้างการแจกแจงแบบทวินาม (binomial distribution) ด้วยพารามิเตอร์ n และ t_i ในแต่ละระดับของการแจกแจงของคะแนนจริง (T_p) กำหนดให้ t_i คือข้อสอบที่ให้คะแนนได้สองค่าสำหรับผู้ทำข้อสอบที่มีความน่าจะเป็นของการตอบถูกในแต่ละข้อ ให้ X' แทนคะแนนของตัวแปรที่มีการแจกแจงแบบทวินาม หาการแจกแจงความน่าจะเป็นสะสม (cumulative probability distribution) ของ X' ที่แต่ละระดับของการแจกแจงของคะแนนจริง (T_p) แปลงการกระจายนี้เป็นสเกล X ใช้สมการ 3 และการประมาณค่าแบบช่วงเชิงเส้น (linear interpolation) แปลงการแจกแจงความน่าจะเป็นสะสม (cumulative probability distribution) ที่เปลี่ยนรูปแล้วให้เป็นการแจกแจงความน่าจะเป็นสำหรับ X ที่กำหนดให้ ใช้ขอบเขตของกลุ่มในการกำหนดการแจกแจงของคะแนนจริง (T_p) สำหรับแต่ละระดับของคะแนนจริง (true-score) ความน่าจะเป็นแบบมีเงื่อนไขว่าผู้สอบที่มีคะแนนจริง t_i เมื่อทำแบบสอบฉบับอื่นจะได้รับการจำแนกให้อยู่ในแต่ละกลุ่ม ให้ X_1 แทนคะแนนของผู้สอบจากแบบสอบฉบับอื่น และให้ x_j^* แทนคะแนนจุดตัดที่ j (j^{th}) ให้ x_j เป็นจำนวนเต็มที่ใกล้ x_j^* ที่สุด (โดยที่ x_j คือจำนวนเต็ม ดังนั้น $x_j - 0.5 < x_j^* \leq x_j + 0.5$) ดังนั้นความน่าจะเป็นที่คะแนนของผู้สอบจะต่ำกว่าคะแนนจุดตัดที่ j (j^{th}) ในแต่ละระดับของการแจกแจงของคะแนนจริง (T_p) จะกำหนดด้วยสมการที่แก้ไขแล้วดังนี้

$$\hat{P}\left(X < x_j^* \mid T_p = t_i\right) = P\left(X < x_j \mid T_p = t_i\right) + \frac{x_j^* - \left(x_j - 0.5\right)}{1.0} P\left(X = x_j \mid T_p = t_i\right) \quad (4)$$

1.4) ประมาณค่าการแจกแจงร่วม (joint distribution) ของการจำแนกประเภทบนฐานของคะแนนจริงและคะแนนจากแบบสอบฉบับอื่น

แปลงขอบเขตของกลุ่มเชิงเส้น (category boundaries linearly) จากสเกลแบบดั้งเดิมของ X (มีค่าตั้งแต่ X_{min} จนถึง X_{max}) ไปยังสเกลของสัดส่วน (ตั้งแต่ .00 ถึง 1.00) ด้วยการประยุกต์มาจากสมการที่ 1 ($p = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$) ใช้ขอบเขตที่แปลงแล้วนี้ ประมาณค่าการแจกแจงของ T_p จากขั้นตอนที่ 1.2 และประมาณค่าการจำแนกแบบมีเงื่อนไขที่แต่ละระดับของ T_p จากขั้นตอนที่ 1.3 เพื่อประมาณค่าการแจกแจงร่วม (joint distribution) ของการจำแนกบนสัดส่วนของคะแนนจริง (T_p) และคะแนนสอบที่ได้จากแบบสอบรูปแบบอื่น (X_1) ให้ t_j^* แทนผลของการใช้สมการที่ 1 ไปยังคะแนนจุดตัด x_j^* และให้ t_{i^*} เป็นค่าที่ไม่ต่อเนื่องของ t_i ที่ใกล้กับ t_j^* ที่สุด (ถูกต้องมากกว่านั้น t_{i^*} คือค่าที่ $t_{i^*} - .005 < t_j^* \leq t_{i^*} + .005$) ดังนั้นการประมาณค่าความถี่ที่สัมพันธ์กันในส่วนหนึ่งของการแจกแจงร่วมสำหรับการที่ $T_p < t_j^*$ และ $X_1 < x_j^*$ คือ

$$\sum_{i<i^*} g_i \hat{P}\left(X < x_j^* \mid T_p = t_i\right) + \frac{t_j^* - \left(t_{i^*} - .005\right)}{.01} g_{i^*} \hat{P}\left(X < x_j^* \mid T_p = t_{i^*}\right)$$

1.5) ประมาณค่าการแจกแจงร่วม (joint distribution) ของการจำแนกประเภทบนฐานของคะแนนจริงและคะแนนจากแบบสอบที่ใช้ในการบริหารจัดการการทดสอบจริง

ปรับการประมาณค่าการจำแนกแบบสองทางที่ขึ้นอยู่กับ T_p และ X_1 (ผลลัพธ์จากขั้นตอนที่ 1.4) เพื่อที่จะทำให้ความถี่จำนวนเล็กน้อยที่ไม่สำคัญของ X ตรงกับความถี่ที่สังเกตได้สำหรับการบริหารจัดการการทดสอบจริง (X_0) การปรับประกอบด้วยการกำหนดตัวคูณสำหรับแต่ละกลุ่มของ X และนำไปใช้กับความถี่ของเซลล์ทั้งหมดสำหรับกลุ่มนั้นๆ ของ X ตัวคูณสำหรับแต่ละกลุ่มเป็นช่วงของความถี่ที่สังเกตได้สำหรับกลุ่มนั้นๆ (ในการแจกแจงของ X_0) เพื่อประมาณค่าความถี่สำหรับกลุ่มนั้นๆ (ในการแจกแจงแบบมาร์จินอลของ X_1) การปรับการจำแนกแบบสองทางเป็นการประมาณค่าการจำแนกแบบสองทางที่ขึ้นอยู่กับ T_p และ X_0 และเป็นพื้นฐานสำหรับการประมาณค่าทางสถิติที่ใช้อธิบายความถูกต้องของการจำแนกประเภท

2) Rudner's method

Rudner (2001) ได้เสนอวิธีการสำหรับคำนวณค่าดัชนีความถูกต้องของการจำแนกประเภทสำหรับข้อสอบแบบให้คะแนนได้สองค่า (dichotomous items) ภายใต้กรอบแนวคิดของทฤษฎีการตอบสนองข้อสอบ (IRT) วิธีการของ Rudner นี้ แบบสอบจะได้รับการให้คะแนนบนสเกลของความสามารถแฝง (latent ability scale) โดยให้ θ แทนคะแนนจริง (true score) $\hat{\theta}$ แทนคะแนนที่สังเกตได้ (observed score) และ θ_c แทนคะแนนจุดตัด (cut score) ซึ่งมีข้อตกลงเบื้องต้นว่าคะแนนจริงใดๆ ที่มีความสัมพันธ์กับคะแนนที่สังเกตได้ตามการแจกแจงแบบปกติด้วยค่าเฉลี่ยของคะแนนจริงและส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนมาตรฐานของการประมาณค่าคะแนนที่สังเกตได้ (standard error of estimation)

โดย Rudner (2001) ได้ทำการพัฒนาวิธีการที่ใช้ทฤษฎีการตอบสนองข้อสอบแบบ 3 พารามิเตอร์ (three-parameter item response theory) และการจำแนกประเภทแบบสองลักษณะ (รอบรู้ และไม่รอบรู้) วิธีการเก็บข้อมูลที่ได้จากการวัดในลักษณะต่อเนื่องตามทฤษฎี การทดสอบแบบดั้งเดิม (classical test theory) และตรรกะนี้สามารถขยายไปสู่การจัดกลุ่มที่มากกว่านี้ได้โดยง่าย เริ่มจากแบบสอบที่กำหนดคะแนนของผู้สอบแต่ละคน (θ_i's) ให้อยู่ในสเกลแบบต่อเนื่อง (continuous scale: θ scale) และคะแนนจุดตัดบนสเกลนั้น (θ_c) ที่ใช้เพื่อจำแนกผู้สอบเข้าสู่หนึ่งในสองกลุ่มที่แบ่งไว้ให้แตกต่างกันโดยสิ้นเชิง ผู้สอบที่มีคะแนนสูงกว่าคะแนนจุดตัดจะได้รับการจำแนกให้อยู่ในกลุ่มรอบรู้ (master) ส่วนผู้สอบที่มีคะแนนต่ำกว่าคะแนนจุดตัดจะได้รับ การจำแนกให้อยู่ในกลุ่มไม่รอบรู้ (non-master)

ระบุความแตกต่างระหว่างกลุ่มผู้สอบที่ควรจะได้รับการจัดเข้าสู่กลุ่มระดับความสามารถที่ขึ้นอยู่กับคะแนนจริง (true scores) ของแต่ละคนและกลุ่มที่ผู้สอบถูกจัดตำแหน่งจะขึ้นอยู่กับคะแนนที่สังเกตได้ (observed score) เป้าหมายคือเพื่อพัฒนาและวิเคราะห์ตารางการจำแนกประเภทแบบสองต่อสอง ดังตารางที่ 6 ซึ่งแสดงให้เห็นถึงสัดส่วนที่คาดหวังของการจำแนกที่ถูกต้องและการจำแนกที่ไม่ถูกต้อง

ตารางที่ 2.6 ตารางการจำแนกประเภท (Classification table)

ถูกจำแนกให้รอบรู้	*ถูกจำแนกให้ไม่รอบรู้*
รอบรู้จริง	*รอบรู้จริง**
ถูกจำแนกให้รอบรู้	ถูกจำแนกให้ไม่รอบรู้
*ไม่รอบรู้จริง**	ไม่รอบรู้จริง

ในตารางที่ 2.6 ควอแดนท์บนซ้ายและควอแดนท์ล่างขวาแสดงถึงการจำแนกที่ถูกต้อง ที่เหลืออีกสองควอแดนท์แสดงถึงการจำแนกที่ไม่ถูกต้อง

สัดส่วนที่คาดหวังของการจำแนกผู้สอบทั้งหมดที่รอบรู้เป็นผู้ไม่รอบรู้จริง คือ

$$P(cm,n)=\sum_{\theta_i<\theta_c} P\left(\hat{\theta}>\theta_c \mid \theta_i\right) f\left(\theta_i\right)/n \qquad (1)$$

ตั้งข้อสังเกตได้ว่า คำที่ใช้แทน “ผู้รอบรู้จริง” (true masters) และ “ผู้ไม่รอบรู้จริง” (true non-masters) ในความหมายทางสถิติคือ ความสามารถที่แท้จริงที่อยู่สูงกว่าหรือต่ำกว่าคะแนนจุดตัดที่ตั้งขึ้นมาอย่างไม่มีกฎเกณฑ์ ในสมการ (1) $P(\hat{\theta}>\theta_c \mid \theta_i)$ คือความน่าจะเป็นของการมีคะแนนที่สังเกตได้ $\hat{\theta}$ อยู่สูงกว่าคะแนนจุดตัด ให้คะแนนจริงเท่ากับ θ_i , $f(\theta_i)$ คือจำนวนที่คาดหวังของคนที่มีคะแนนจริงเป็น θ_i และ n คือ จำนวนผู้สอบทั้งหมด ดังนั้น $P(\hat{\theta}>\theta_c \mid \theta_i)$ $f(\theta_i)$ คือจำนวนที่คาดหวังของผู้สอบที่มีคะแนนจริงเป็น θ_i ซึ่งจะได้รับการจำแนกเป็นผู้รอบรู้ เช่น จะมีคะแนนที่สังเกตได้มากกว่าคะแนนจุดตัด รวมค่านี้ของผู้สอบทุกคนที่มีคะแนนจริงน้อยกว่าคะแนนจุดตัดและหารด้วย n เพื่อให้ได้ค่าความน่าจะเป็นของการจำแนกที่ไม่ถูกต้องเป็นผู้รอบรู้ (ความผิดพลาดเชิงบวก: false positive)

ในทำนองเดียวกัน สัดส่วนที่คาดหวังของความผิดพลาดเชิงลบ (false negative) คือ

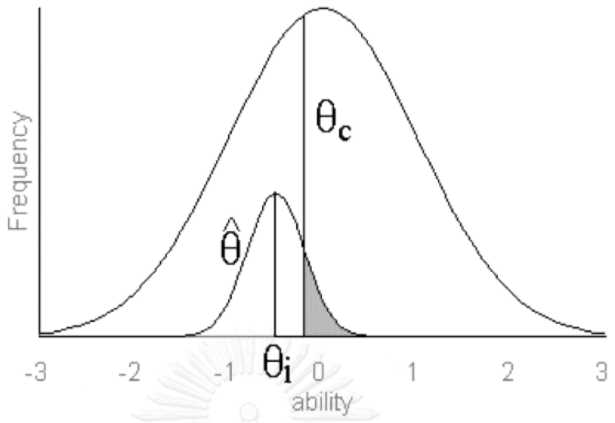
$$P(cn,m)=\sum_{\theta_i>\theta_c} P\left(\hat{\theta}<\theta_c \mid \theta_i\right) f\left(\theta_i\right)/n \qquad (2)$$

จากสมการ (1) ความน่าจะเป็นของการมีคะแนนที่สังเกตได้สูงกว่าคะแนนจุดตัดให้เป็น θ_i, $P(\hat{\theta}>\theta_c \mid \theta_i)$ คือ พื้นที่ภายใต้โค้งปกติ

$$z=\frac{\theta_c-\theta_i}{se(\theta_i)} \qquad (3)$$

ซึ่งอธิบายได้ในภาพที่ 1 โค้งระฆังคว่ำที่สูงกว่าแสดงถึงการแจกแจงของความสามารถภายในประชากรทั้งหมด และคะแนนจุดตัดถูกกำหนดที่ θ_c = -.2 โค้งที่เล็กกว่าแสดงถึงการแจกแจงที่คาดหวังของค่าที่สังเกตได้ของ theta สำหรับผู้สอบที่มีค่าจริงของ θ_i = -.5 ตรวจสอบด้วยคะแนนจริงของ θ_i = -.5 คือผู้ไม่รอบรู้ และควรจะถูกจำแนกด้วยวิธีการนั้น อย่างไรก็ตามคะแนน ที่สังเกตได้จะเปลี่ยนแปลงโดยประมาณ θ_i = -.5 บริเวณที่แรเงาที่อยู่ทางขวาของคะแนนจุดตัดแสดงถึงความน่าจะเป็นของผู้สอบที่มีคะแนนจริง -.5 ซึ่งสามารถได้รับการคาดหวังให้เป็นการจำแนกที่ไม่

ถูกต้องในฐานะที่เป็นผู้รอบรู้ ภาพที่ 1 เป็นเพียงหนึ่งค่าของ theta เพื่อตรวจสอบ P(cm,n) โดยบุคคลหนึ่งจะมีเส้นโค้งสำหรับแต่ละค่าของ theta น้อยกว่า θ_c



ภาพที่ 2.1 ความผิดพลาดเชิงบวกสำหรับผู้สอบที่ระดับความสามารถหนึ่ง
(ที่มา: Rudner, 2001)

ในสมการที่ (3) se(θ_i) แทนความคลาดเคลื่อนมาตรฐานของการประมาณค่าความสามารถที่คะแนนของ θ_i ได้เป็น

$$se(\theta_i) = \frac{1}{\sqrt{I(\theta_i)}} \tag{4}$$

เมื่อ I(θ_i) คือฟังก์ชันสารสนเทศของแบบสอบ (test information function) หาค่าที่คะแนนของ θ_i Lord (1980 cited in Rudner, 2001) สร้างสมการสำหรับ I(θ_i) โดยใช้คะแนนรวมที่ถ่วงน้ำหนักแล้ว (weighted composite scoring) จำนวนการให้คะแนนที่ถูกต้อง และการให้คะแนนตามทฤษฎีการตอบสนองข้อสอบ เมื่อใช้ทฤษฎีการตอบสนองข้อสอบ (IRT) ฟังก์ชันสารสนเทศของแบบสอบที่ θ_i คือ ผลรวมของฟังก์ชันสารสนเทศของข้อสอบที่ θ_i ซึ่งสามารถคำนวณได้จากค่าพารามิเตอร์ของข้อสอบ (a, b และ c) ตามทฤษฎีการตอบสนองข้อสอบ (IRT)

สัดส่วนที่คาดหวังของผู้สอบที่มีคะแนนจริงเป็น θ_i ค่า f(θ_i)/n สามารถประมาณค่าได้จากตัวอย่างที่ใช้ในการศึกษานำร่อง สังเกตว่าความน่าจะเป็นของคะแนนที่ได้รับของ θ_i แทนด้วย P(θ_i) และดูที่ θ ในฐานะที่เป็นตัวแปรต่อเนื่องมากกว่าที่เป็นตัวแปรแบ่งกลุ่ม ดังนั้นสมการ (1) และ (2) จึงเป็น

$$P(cm,n)=\int_{\theta_i=-\infty}^{\theta_c} P(\hat{\theta}>\theta_c \mid \theta_i)P(\theta_i)d\theta_i \quad (5)$$

และ

$$P(cn,m)=\int_{\theta_i=\theta_c}^{\infty} P(\hat{\theta}<\theta_c \mid \theta_i)P(\theta_i)d\theta_i \quad (6)$$

การทำกลุ่มของสมการให้สมบูรณ์ตามต้องการสำหรับตารางการจำแนกแบบสองต่อสอง (two-by-two classification table) ความน่าจะเป็นของการจำแนกที่ถูกต้องให้เป็นผู้รอบรู้ และความน่าจะเป็นของการจำแนกที่ถูกต้องให้เป็นผู้ไม่รอบรู้ คือ

$$P(cm,n)=\int_{\theta_i=\theta_c}^{\infty} P(\hat{\theta}>\theta_c \mid \theta_i)P(\theta_i)d\theta_i \quad (7)$$

และ

$$P(cn,n)=\int_{\theta_i=-\infty}^{\theta_c} P(\hat{\theta}<\theta_c \mid \theta_i)P(\theta_i)d\theta_i \quad (8)$$

ถ้าสมมติว่าค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานมีการแจกแจงแบบปกติ ดังนั้น $P(\theta_i)$ เป็นจุดสูงสุดของโค้ง Gaussian (Gaussian curve) ที่ทำการประมาณค่าที่ $(\theta_i - \mu)/\sigma$

ต่อมาในปี ค.ศ. 2005 Rudner ได้ศึกษาความถูกต้องของการจำแนกประเภทที่คาดหวัง (expected classification accuracy) โดยทำการขยายผลการศึกษาการประมาณค่าความถูกต้องของการจำแนกประเภทที่คาดหวังจากการประมาณค่าสำหรับการให้คะแนนได้สองค่า (dichotomous items) มาศึกษากับการให้คะแนนได้หลายค่า (polytomous items) โดยใช้การประมาณค่าแบบ categorical approach ซึ่งเป็นการสร้างตารางจำแนกประเภทของคะแนนจริง (true scores) และคะแนนที่คาดหวัง (expected scores) สำหรับข้อสอบที่ให้คะแนนได้หลายค่าภายใต้ทฤษฎีการตอบสนองข้อสอบ (IRT)

ถ้าแบบสอบประกอบด้วยข้อสอบจำนวน N ข้อ ต้องการจำแนกผู้สอบให้อยู่ในกลุ่มใดกลุ่มหนึ่งของคะแนนจากจำนวน K กลุ่ม จากนิยามให้คะแนนจริง (true score) เป็น θ คะแนนที่สังเกตได้ (observed score) ที่สอดคล้องกันเป็น $\hat{\theta}$ ซึ่งคาดหวังว่ามีการแจกแจงแบบปกติด้วยค่าเฉลี่ยของ θ และส่วนเบี่ยงเบนมาตรฐานของ se(θ) ความน่าจะเป็นของผู้สอบที่ทำให้คะแนนจริงของ θ มีคะแนนที่สังเกตได้อยู่ในช่วง [a, b] บนระดับ theta เป็นดังนี้

$$\text{Prob}(a<\hat{\theta}<b \mid \theta)=\phi\left(\frac{b-\theta}{se(\theta)}\right)-\phi\left(\frac{a-\theta}{se(\theta)}\right) \quad (1)$$

เมื่อ ϕ(z) คือฟังก์ชันการแจกแจงปกติสะสม (cumulative normal distribution function) ซึ่งเป็นพื้นที่ใต้โค้งปกติระหว่าง a และ b ด้วยค่าเฉลี่ย θ และส่วนเบี่ยงเบนมาตรฐาน se(θ)

คูณสมการ (1) ด้วยสัดส่วนที่คาดหวังของผู้สอบที่มีคะแนนจริงเท่ากับ θ ทำให้ได้สัดส่วนที่คาดหวังของผู้สอบที่มีคะแนนจริงเป็น θ คาดหวังว่าจะอยู่ในช่วง [a, b] รวมหรืออินทิเกรทผู้สอบทั้งหมดที่อยู่ในช่วง [c, d] จะได้สัดส่วนที่คาดหวังของผู้สอบทั้งหมดซึ่งมีคะแนนจริงอยู่ใน [c, d] และคะแนนที่สังเกตได้อยู่ใน [a, b]

$$\sum_{\theta=c}^{d} P\left(a < \hat{\theta} < b \mid \theta\right) f(\theta)$$

เมื่อ f(θ) คือสัดส่วนที่คาดหวังของผู้สอบที่มีคะแนนจริง (true score) เป็น θ ถ้าสมมติให้ θ เป็น N(μ, σ) ดังนั้น f(θ) คือฟังก์ชันของความหนาแน่นปกติมาตรฐาน φ(z)

กำหนด [a,b] และ [c,d] เพื่อให้สอดคล้องกับช่วงของคะแนนจริงที่กำหนดจากคะแนนจุดตัด ทำให้ได้องค์ประกอบของตารางการจำแนกที่แสดงให้เห็นถึงสัดส่วนที่คาดหวังของผู้สอบทั้งหมดกับคะแนนที่สังเกตได้และคะแนนจริงในแต่ละเซลล์ องค์ประกอบแต่ละส่วนของตารางการจำแนกเป็นดังนี้

$$\sum_{\theta=c}^{d} P\left(a < \hat{\theta} < b \mid \theta\right) f(\theta) = \sum_{\theta=c}^{d} \left(\phi\left(\frac{b-\theta}{se(\theta)} \right) - \phi\left(\frac{a-\theta}{se(\theta)} \right) \right) \varphi\left(\frac{\theta-\mu}{\sigma} \right) \qquad (2)$$

เมื่อ f(θ) คือสัดส่วนที่คาดหวังของผู้สอบที่มีคะแนนจริง (true score) เป็น θ คำนวณสมการที่ (2) สำหรับแต่ละเซลล์เพื่อทำให้ค่า K × K ในตารางการจำแนกสมบูรณ์ ค่าความถูกต้องโดยรวมแล้วคือผลรวมของสมาชิกทั้งหมดในแนวเส้นทแยงมุม

ตัวอย่าง

ตารางที่ 2.7 ประกอบด้วยพารามิเตอร์ของข้อสอบวัดการอ่านสำหรับนักเรียนเกรด 8 ในโปรแกรมการประเมินภาคปฏิบัติของรัฐแมรีแลนด์ปี 2001 รูปแบบ A จำนวน 10 ข้อ แบบสอบนี้ได้รับการปรับเทียบและให้คะแนนโดย CTB-McGraw Hill โดยใช้ Generalized Partial Credit Model (Muraki, 1992 cited in Rudner, 2005) เมื่อ K = 5 ช่วงคะแนนคือ [(375,489), (490,529), (530,579), (580,619), (620,650)] ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานถูกกำหนดมาให้เป็น 500 และ 50 ตามลำดับ

ตารางที่ 2.7 พารามิเตอร์ของข้อสอบตาม Generalized Partial Credit Model สำหรับข้อสอบการอ่าน 10 ข้อ

ข้อที่	a	b1	b2	b3
1	0.040	20.103	18.650	22.206
2	0.040	21.231	19.442	
3	0.037	19.573	18.674	
4	0.044	22.838	21.573	
5	0.043	22.941	21.357	
6	0.042	21.926	18.325	
7	0.051	26.644	23.579	
8	0.049	25.684	23.270	
9	0.052	27.247	25.523	
10	0.037	20.104	19.191	

เนื่องจากความคลาดเคลื่อนมาตรฐานที่ θ เป็นส่วนกลับของกำลังสองของฟังก์ชันสารสนเทศของแบบสอบ (test information function) ที่ θ โดยที่ฟังก์ชันสารสนเทศของแบบสอบคือผลรวมของฟังก์ชันสารสนเทศของข้อสอบ และภายใต้ Generalized Partial Credit Model (Donoghue, 1994 cited in Rudner, 2005) ฟังก์ชันสารสนเทศของข้อสอบคือ

$$I_i(\theta) = a_i^2\left[\sum_{k=0}^{m_j} k^2 P_{ik}(\theta) - \left(\sum_{k=0}^{m_j} kP_{ik}(\theta)\right)^2\right]$$

เมื่อ a_i คือ ดัชนีอำนาจจำแนกของข้อสอบ (item discrimination index) ความคลาดเคลื่อนมาตรฐานสำหรับค่าที่ให้สำหรับแต่ละ θ คือ

$$se(\theta) = 1\Big/\sqrt{\sum_{i=1}^{n} a_i^2\left[\sum_{k=0}^{m_j} k^2 P_{ik}(\theta) - \left(\sum_{k=0}^{m_j} kP_{ik}(\theta)\right)^2\right]} \qquad (3)$$

เมื่อใช้ตารางที่ 2.7 และสมการที่ (2) และ (3) ในการหาค่าการจำแนกได้ดังตารางที่ 2.8 ผลรวมตามแนวเส้นทแยงมุมในตารางที่ 2.8 ได้เท่ากับ 81.4% ซึ่งเป็นค่าความถูกต้องที่คาดหวัง ดังนั้นข้อสอบ 10 ข้อนี้พบว่ามีความเพียงพอสำหรับการรายงานผลคะแนนรวม

ตารางที่ 2.8 ร้อยละของผู้สอบในแต่ละกลุ่มคะแนนจากการจำแนกประเภทที่คาดหวัง (expected classification table)

		กลุ่มของคะแนนที่คาดหวัง (Expected score category)					
		(375-489) 0	(490-529) 1	(530-579) 2	(580-619) 3	(620-650) 4	
กลุ่มของคะแนนจริง (True score category)	0	33.4	4.9	0.0	0.0	0.0	38.3
	1	4.7	33.3	3.8	0.0	0.0	41.8
	2	0.0	3.5	14.2	1.1	0.0	18.8
	3	0.0	0.0	0.4	0.5	0.2	1.1
	4	0.0	0.0	0.0	0.0	0.0	0.0
		38.1	41.7	18.4	1.6	0.2	100

3) Guo's method

Guo (2006) ได้ทำการศึกษาความถูกต้องของการจำแนกประเภทที่คาดหวัง (expected classification accuracy) โดยใช้การแจกแจงแฝง (latent distribution) โดยที่ Guo ได้พัฒนาวิธีการประมาณค่าแบบวิธีการแจกแจงแฝง (latent distribution method) จากการศึกษาเปรียบเทียบกับวิธีการประมาณค่าตามทฤษฎีการตอบสนองข้อสอบที่ Rudner พัฒนาขึ้นในปี ค.ศ. 2001 และ 2005 (Rudner's method) พบว่า latent distribution method แตกต่างจากวิธีการของ Rudner (2005) ในส่วนของการคำนวณจำนวนผู้สอบที่คาดหวังในแต่ละกลุ่มความสามารถด้วยการแจกแจงภายหลัง (posterior distributions) โดยใช้ฟังก์ชันความน่าจะเป็นปกติ (normalized likelihood function) เป็นผลให้ข้อตกลงเบื้องต้นเกี่ยวกับการแจกแจงปกติของความคลาดเคลื่อนในการประมาณค่าคะแนนจริงของวิธีการของ Rudner (2005) ไม่มีความจำเป็น ดังนั้นวิธีการแจกแจงแฝง (latent distribution method) อาจจะเป็นวิธีการที่แข็งแกร่งกว่าแม้ว่าการประมาณค่าความสามารถจะมีความถูกต้องเพียงเล็กน้อยอันเนื่องมาจากแบบสอบที่มีข้อสอบจำนวนน้อย หรือค่าสารสนเทศของแบบสอบที่มีค่าต่ำในบางระดับความสามารถ

การประมาณค่าความถูกต้องของการจำแนกประเภท (classification accuracy) ด้วยวิธีการแจกแจงแฝง (latent distribution method) มีขั้นตอนดังนี้

3.1) กำหนดคะแนนจุดตัดไปยังสเกล θ (θ scale) หลังจากกำหนดค่ามาตรฐานสำหรับแบบสอบ แปลงคะแนนจุดตัด x ที่เลือกมาเพื่อแบ่งผลคะแนน (reporting scale) เข้าสู่ช่วงคะแนนของกลุ่ม r (r = x+1) ให้อยู่ในสเกลเดียวกับ θ และทำการเพิ่มเติมค่า θ สำหรับคะแนนที่น่าจะเป็นต่ำสุดและสูงสุดของแบบสอบ กำหนดให้ค่า θ เป็น M (M = x+2) เมื่อ m ใช้แทนค่า θ

ดังนั้น m = 1, 2, ..., M ในที่นี้ θ_1 คือ ค่า θ ของคะแนนที่น่าจะเป็นต่ำสุด และ θ_M คือ ค่า θ ของคะแนนที่น่าจะเป็นสูงสุดของแบบสอบ

3.2) คำนวณค่าความน่าจะเป็น ทำการคำนวณค่าความน่าจะเป็นของผู้สอบคนที่ i สำหรับกลุ่มคะแนนที่อยู่ในช่วง r โดยใช้คะแนนที่ได้จากการตอบข้อสอบของผู้สอบ (μ_1 ถึง μ_n) และค่าพารามิเตอร์ของข้อสอบ $a_1, b_1, c_1, ..., a_n, b_n, c_n$ ของโมเดลการตอบสนองข้อสอบโลจิสติกแบบสามพารามิเตอร์ ตัวอย่างเช่น

$$L_{ri} = \int_{m}^{m+1} L(u_1, u_2, \ldots, u_n \mid \theta) d\theta \tag{1}$$

ความน่าจะเป็นของการสร้างกลุ่มของคะแนน θ ให้แยกจากกันโดยมีระยะห่างเท่าๆ กันสามารถคำนวณได้ดังนี้

$$L_{ri} = \sum_{\theta=m}^{m+1} L(u_1, u_2, \ldots, u_n \mid \theta) \tag{2}$$

3.3) ปรับความน่าจะเป็นให้เป็นค่าปกติ (normalize the likelihood) ทำการปรับค่าความน่าจะเป็นให้เป็นค่าปกติเพื่อให้ผลรวมของความน่าจะเป็นสำหรับผู้สอบแต่ละคนมีค่าเท่ากับ 1 ขั้นตอนนี้มีความจำเป็นเพราะการแจกแจงต้องถูกลดทอนให้เท่ากับคะแนนต่ำสุดและสูงสุดที่ได้มาไม่ใช่ที่ระดับ -∞ และ ∞ เพื่อที่จะนำไปคำนวณ หลังจากทำให้เป็นค่าปกติแล้วผลรวมของความน่าจะเป็นระหว่างผู้สอบทั้งหมดจะเท่ากับหรือใกล้เคียงกับผลรวมของจำนวนผู้สอบทั้งหมด การทำตามลำดับนี้จะทำให้แปลผลได้ง่ายขึ้น

$$NormL_{ri} = \frac{L_{ri}}{\int_{1}^{M} L(u_1, u_2, \ldots, u_n \mid \theta) d\theta} \tag{3}$$

ในทำนองเดียวกัน ผลรวมของความน่าจะเป็นของตัวหารสามารถคำนวณได้เช่นเดียวกับผลรวมระหว่างค่า θ ตั้งแต่ θ_1 ถึง θ_M ที่แยกจากกันโดยสิ้นเชิง

3.4) คำนวณจำนวนผู้สอบที่สังเกตได้ในช่วงคะแนนของกลุ่ม s จำนวนผู้สอบที่สังเกตได้ในช่วงคะแนนของกลุ่ม s เป็นจำนวนของผู้สอบที่มีการประมาณค่าความน่าจะเป็นสูงสุด (maximum likelihood point estimates) ของ $\hat{\theta}$ ที่ตกอยู่ในช่วงคะแนนของกลุ่มนี้

$$O_s = N_{(m \le \hat{\theta} < m+1)} \tag{4}$$

3.5) คำนวณจำนวนผู้สอบที่คาดหวังในแต่ละเซลล์ $N_{(s,r)}$ ของตารางจำแนกประเภท สำหรับผู้สอบที่มีคะแนนที่สังเกตได้ตกอยู่ในช่วงคะแนนของกลุ่ม s นั้น จำนวนผู้สอบที่คาดหวังในแต่ละช่วงคะแนนของกลุ่ม r สามารถคำนวณได้ดังนี้

$$N_{sr} = \sum_{i \in s} NormL_{ri} \tag{5}$$

3.6) รวบรวมตารางการจำแนกประเภท ดังตัวอย่างในตารางที่ 2.9 สังเกตว่าตำแหน่งของจุดทศนิยมจะพบในเซลล์ที่เป็นผลมาจากการแบ่งสัดส่วนของผู้สอบบางคนให้อยู่ในกลุ่มของคะแนนที่คาดหวังมากกว่าหนึ่งกลุ่ม

ตารางที่ 2.9 ตัวอย่างของตารางการจำแนกประเภท (classification table)

		จำนวนผู้สอบที่คาดหวังในแต่ละช่วงคะแนน (Expected N)		
		r_1	**...**	**r_M**
จำนวนผู้สอบที่สังเกตได้ในแต่ละช่วงคะแนน (Observed N)	**s_1**	$N_{(1,1)}$	$N_{(1,...)}$	$N_{(1,M)}$
	...	$N_{(...,1)}$	$N_{(...,...)}$	$N_{(...,M)}$
	s_M	$N_{(M,1)}$	$N_{(M,...)}$	$N_{(M,M)}$

3.7) คำนวณดัชนีความถูกต้อง (accuracy index) ดัชนีความถูกต้องสามารถคำนวณได้ในลักษณะเดียวกับที่ Rudner ได้เสนอไว้ ซึ่งเป็นเพียงการหาค่าร้อยละของผลรวมตามแนวเส้นทแยงมุมหารด้วยผลรวมของจำนวนผู้สอบทั้งหมด

$$\frac{\sum_{s=r=1}^{M} N_{(s,r)}}{\sum_{s=1}^{M}\sum_{r=1}^{M} N_{sr}} \tag{6}$$

ดัชนีความถูกต้องชี้ให้เห็นถึงร้อยละของผู้เข้าสอบที่ได้รับการจำแนกได้อย่างถูกต้อง ดัชนีที่มีค่าสูงกว่าแสดงให้เห็นถึงความถูกต้องของคะแนนจากแบบสอบที่ใช้ในการจำแนกผู้สอบเข้าสู่กลุ่มที่ถูกต้องได้มากกว่า

เนื่องจากฟังก์ชันของความน่าจะเป็นคำนวณขึ้นโดยใช้พารามิเตอร์ของข้อสอบ สมมติฐานที่สำคัญคือการรู้ค่าพารามิเตอร์ของข้อสอบ นั่นคือการประมาณค่าพารามิเตอร์ของข้อสอบต้องสมเหตุสมผลมีค่าใกล้เคียงกับค่าที่แท้จริงของการประมาณค่าเหล่านั้น จึงต้องมั่นใจว่าพารามิเตอร์ของข้อสอบถูกปรับเทียบด้วยผู้สอบจำนวนมากเพื่อผลที่ดีที่สุด ซึ่งมีความสามารถครอบคลุมช่วงทั้งหมดของคะแนน θ (θ scale) เมื่อมีการรายงานผลคะแนน สำหรับวิธีการแจกแจง

แฝงนี้ (latent distribution method) พบว่าบางส่วนของข้อตกลงเบื้องต้นที่จำเป็นสำหรับวิธีการของ Rudner เช่น การทำการประมาณค่าความคลาดเคลื่อนมาตรฐานของการประมาณค่าให้เป็นค่าปกติ และการประมาณค่าที่เหมาะสมของคะแนนที่สังเกตได้ (θ) ไปยังคะแนนจริง ($\hat{\theta}$) เป็นสิ่งที่ไม่จำเป็น

4) Lee's method

ในปี ค.ศ. 2002 Lee และคณะ ได้ทำศึกษาวิธีการประมาณค่าดัชนีความสอดคล้องและความถูกต้องสำหรับการจำแนกประเภทหลายประเภท (multiple classifications) นั้น นอกจากจะให้นิยามและพัฒนาวิธีการประมาณค่าดัชนีความสอดคล้องของการจำแนกประเภท แล้วยังได้ให้นิยามและพัฒนาวิธีการประมาณค่าของดัชนีความถูกต้องของการจำแนกประเภทไว้อีกด้วย โดยให้นิยามของดัชนีความถูกต้องของการจำแนกประเภท (classification accuracy) ว่าเป็นขนาดของการจำแนกที่แท้จริง โดยใช้คะแนนจุดตัดที่สังเกตได้ที่สอดคล้องกับการจำแนกที่ "ถูก" (true) โดยมีพื้นฐานอยู่บนคะแนนจุดตัดของคะแนนที่รู้จริง (Livingston และ Lewis, 1995 cited in Lee et al., 2002) ในขณะที่ความสอดคล้องของการจำแนกประเภทถูกนิยามบนพื้นฐานของการแจกแจงของคะแนนที่สังเกตได้ของการแจกแจงของคะแนนสอบที่ได้จากชุดข้อสอบที่วัดคุณลักษณะเดียวกันสองชุด ความถูกต้องของการจำแนกประเภทถูกนิยามบนพื้นฐานของการแจกแจงของตัวแปรสองตัว (bivariate distribution) ของการแจกแจงของคะแนนที่สังเกตได้และคะแนนจริง

ในการคำนวณดัชนีความถูกต้องของการจำแนกประเภทต้องแบ่งคะแนนจุดตัดของคะแนนจริง ϕ_1, ϕ_2, ..., $\phi_{(H-1)}$ ซึ่งแบ่งส่วนประชากรเป็นจำนวน H กลุ่ม ตามคะแนนคุณลักษณะจริงของผู้สอบ ให้ Γ_l (l = 1, 2, ..., H) แทนจำนวนกลุ่มการจำแนกที่แท้จริง ซึ่งนิยามว่าเป็นสถานะที่แท้จริงของผู้สอบ ซึ่ง $\phi_{(l-1)} \leq \phi_l$ สำหรับ l = 1 เงื่อนไขคือ $\min(\phi) \leq \phi \leq \phi_l$ และ l = H เงื่อนไขคือ $\phi_{(H-1)} \leq \phi \leq \max(\phi)$ เช่นเดียวกับกรณีของความสอดคล้องของการจำแนกประเภท ตารางการณ์จรขนาด H×H จะถูกสร้างขึ้น ซึ่งบรรจุด้วยความน่าจะเป็นร่วมของการจำแนกกลุ่มที่สังเกตได้และกลุ่มที่เป็นจริงคือ $\Pr(X \in I_h, \Phi \in \Gamma_l)$ ตารางการณ์จรสำหรับความถูกต้องของการจำแนกประเภทจะไม่เป็นแบบสมมาตร

อย่างไรก็ตามเนื่องจากการแจกแจงสองการแจกแจงที่นำไปใช้ในการสร้างตารางนั้นแตกต่างกัน ดัชนีความถูกต้องของการจำแนกประเภท ในภาพรวมแทนด้วยสัญลักษณ์ γ คือ ผลรวมของสมาชิกในแนวเส้นทแยงมุมในตารางการณ์จร H×H ผลบวกของสมาชิกที่อยู่เหนือแนวเส้นทแยงมุมจะชี้ให้เห็นถึงความน่าจะเป็นในภาพรวมของกลุ่มที่สังเกตได้ของผู้สอบว่ามีค่าสูงกว่ากลุ่มจริงของผู้สอบ แทนด้วย P^+ ตรงกันข้ามผลบวกของสมาชิกที่อยู่ใต้แนวเส้นทแยงมุมจะชี้ให้เห็นถึงความน่าจะเป็นในภาพรวมของกลุ่มที่สังเกตได้ของผู้สอบว่ามีค่า สูงกว่ากลุ่มจริงของผู้สอบ แทนด้วย P^- เมื่อแต่ละคู่ของ

คะแนนจุดตัดจริงถูกนำมาแยกส่วนกันเป็นดัชนีย่อย H-1 ของ γ ซึ่งก็คือ γ_m จะถูกคำนวณขึ้น แล้ว P^+ และ P^- จะกลายเป็นอัตราความคลาดเคลื่อนเชิงบวกและอัตรา ความคลาดเคลื่อนเชิงลบ

โดยมีขั้นตอนในการประมาณค่าดัชนีความถูกต้องของการจำแนกประเภทดังนี้

ถ้าผู้สอบมีคะแนนสังเกตได้ (observed score) และคะแนนคุณลักษณะ (latent score) เป็น $X \in I_h$ (h = 1, 2, ..., H) และ $\phi \in \Gamma_l$ (l = 1, 2, ..., H) ความถูกต้องของการจำแนกจะเกิดขึ้นเมื่อ h = l ความน่าจะเป็นแบบมีเงื่อนไขของความถูกต้องของการจำแนกสร้างโดย

$$\gamma(\phi) = \Pr(X \in I_l \mid \Phi = \phi) \tag{1}$$

เมื่อ l = 1, 2, ..., H เป็นกลุ่มที่ $\phi \in \Gamma_l$ และ $\Pr(X \in I_l \mid \Phi = \phi)$ คำนวณได้โดยใช้สมการนี้

$$\Pr(X \in I_h \mid \Phi = \phi) = \sum_{x=c_{(h-1)}}^{c_h - 1} f(x \mid \phi), h = 1,2,\ldots,H \tag{2}$$

ให้ $P^+(\phi)$ คือ ความน่าจะเป็นแบบมีเงื่อนไขของผู้สอบด้วยคะแนนคุณลักษณะ (latent score) ϕ ภายในช่วงของกลุ่มจริงที่มีคะแนนที่สังเกตได้ตกอยู่ในกลุ่มสังเกตได้ ซึ่งเป็น 1 หรือมากกว่ากลุ่มจริง อีกทั้ง $P^-(\phi)$ คือ ความน่าจะเป็นแบบมีเงื่อนไขของผู้สอบที่ตกอยู่ในกลุ่มสังเกตได้ 1 หรือน้อยกว่ากลุ่มจริง จะได้อัตราความคลาดเคลื่อนของสองเงื่อนไขนี้เป็น

$$P^+(\phi) = \sum_{h=l+1}^{H} \Pr(X \in I_h \mid \Phi = \phi) \tag{3}$$

และ

$$P^-(\phi) = \sum_{h=1}^{l-1} \Pr(X \in I_h \mid \Phi = \phi) \tag{4}$$

เมื่อ l = 1, 2, ..., H คือ กลุ่มที่ $\phi \in \Gamma_l$

จากที่กล่าวมาข้างต้นจะเห็นได้ว่าดัชนีความถูกต้องของการจำแนกนั้นสามารถคำนวณดัชนีย่อยของ γ เมื่อแยกจุดตัดเป็น

$$\gamma_{m(\phi)} \begin{cases} \sum_{h=1}^{m} Pr(X \in I_h \mid \Phi = \phi), \ \phi \in \Gamma_l \ and \ l \leq m \\ \sum_{h=m+1}^{H} Pr(X \in I_h \mid \Phi = \phi), \ \phi \in \Gamma_l \ and \ l > m \end{cases} \tag{5}$$

สำหรับ m = 1, 2, ..., H-1 ที่คล้ายกันกับดัชนีย่อยของอัตราความคลาดเคลื่อนแบบมีเงื่อนไข จะแทนด้วย

$$P_m^+(\phi) = \sum_{h=m+1}^{H} \Pr(X \in I_h \mid \Phi = \phi), \quad \phi \in \Gamma_l \quad and \quad l \le m \tag{6}$$

และ

$$P_m^-(\phi) = \sum_{h=1}^{m} \Pr(X \in I_h \mid \Phi = \phi), \quad \phi \in \Gamma_l \quad and \quad l > m \tag{7}$$

สำหรับ m = 1, 2, ..., H-1 ขนาดความน่าจะเป็นของ γ_m, P_m^+ และ P_m^- คำนวณได้ตามสมการที่ 5, 6 และ 7 บนการแจกแจง Φ และผลของ P_m^+ และ P_m^- จะสามารถเปรียบเทียบกันได้ ซึ่งก็คือ อัตราการคลาดเคลื่อนเชิงบวกและเชิงลบ

ต่อมาในปี ค.ศ. 2010 Lee ได้ศึกษาวิธีการประมาณค่าดัชนีความสอดคล้องและความถูกต้องของการจำแนกประเภทสำหรับการประเมินที่ซับซ้อนตามทฤษฎีการตอบสนองข้อสอบ (IRT) โดยนำไปใช้กับข้อมูลจากการทดสอบจริงที่ประกอบไปด้วยข้อสอบแบบให้คะแนนได้สองค่า (dichotomous) และให้คะแนนได้หลายค่า (polytomous) ซึ่งมีขั้นตอนในการประมาณค่าความถูกต้องของการจำแนกประเภทดังต่อไปนี้

สมมติว่าความน่าจะเป็นของการจัดกลุ่มแบบมีเงื่อนไข $p_\theta(h)$ ถูกคำนวณบนพื้นฐานของคะแนนจุดตัดที่สังเกตได้โดยใช้สมการ $\phi_\theta = \sum_{h=1}^{K} [p_\theta(h)]^2$ แล้วให้สมมติชุดของคะแนนจุดตัดที่แท้จริงในเมทริกซ์ของคะแนนรวม $\tau_1, \tau_2, ..., \tau_{K-1}$ พิจารณาสถานะของกลุ่มที่แท้จริงของผู้สอบแต่ละคนด้วย θ และ τ (เช่น คะแนนรวมที่คาดหวัง) ถ้าสถานะของกลุ่มที่แท้จริงของผู้สอบที่ทราบได้เป็น η(= 1, 2, ..., K) ความน่าจะเป็นแบบมีเงื่อนไขของความถูกต้องของการจำแนกประเภทแบบง่ายๆ คือ

$$\gamma_\theta = p_\theta(\eta), \ for \ \theta \in \eta \tag{1}$$

สังเกตว่ากลุ่มจริง η สามารถพิจารณาได้จากการเปรียบเทียบคะแนนรวมที่คาดหวังของ θ โดยคำนวณจาก $E(X \mid \theta = \theta^*) = \sum_i \sum_j j \Pr(U_i = j \mid \theta = \theta^*)$ ด้วยคะแนนจุดตัดจริง สมการที่ 1 ที่กล่าวถึงนี้คือดัชนีความถูกต้องของการจำแนกประเภทแบบมีเงื่อนไข โดยดัชนีความถูกต้องของการจำแนกประเภทแบบ marginal แทนด้วย γ ดังนี้

$$\gamma = \int_{-\infty}^{\infty} \gamma_\theta g(\theta) d\theta \tag{2}$$

ความถูกต้องของการจำแนกประเภทมักจะได้รับการประมาณค่าด้วยอัตราความคลาดเคลื่อนของความผิดพลาดเชิงบวก (false positive) และความผิดพลาดเชิงลบ (false negative) (Hanson และ Brennan, 1990; Lee et al., 2002 cited in Lee, 2010) อัตราความคลาดเคลื่อนของความผิดพลาดเชิงบวก (false positive) แบบมีเงื่อนไขถูกกำหนดในที่นี้ว่าเป็นความน่าจะเป็นที่ผู้สอบถูกจำแนกเข้าสู่กลุ่มที่สูงกว่ากลุ่มที่แท้จริงของผู้สอบ ซึ่งแสดงได้ดังนี้

$$\gamma_\theta^+ = \sum_{\eta=\eta^*+1}^{K} p_\theta(\eta),\ for \theta \in \eta^* \tag{3}$$

ในทางตรงกันข้ามอัตราความคลาดเคลื่อนของความผิดพลาดเชิงลบ (false negative) แบบมีเงื่อนไขคือความน่าจะเป็นที่ผู้สอบถูกจำแนกเข้าสู่กลุ่มซึ่งต่ำกว่ากลุ่มที่แท้จริงของผู้สอบ แทนด้วย

$$\gamma_\theta^- = \sum_{\eta=1}^{\eta^*-1} p_\theta(\eta),\ for \theta \in \eta^* \tag{4}$$

ขอบเขตของอัตราความคลาดเคลื่อนของความผิดพลาดเชิงบวก (false positive) และความผิดพลาดเชิงลบ (false negative) แบบ marginal ของ γ^+ และ γ^- คือ

$$\gamma^+ = \int_{-\infty}^{\infty} \gamma_\theta^+ g(\theta) d\theta \tag{5}$$

และ

$$\gamma^- = \int_{-\infty}^{\infty} \gamma_\theta^- g(\theta) d\theta \tag{6}$$

เมื่อคะแนนจุดตัดที่แท้จริงถูกกำหนดบนเมทริกซ์ของ θ สมการ

$$E\left(X \mid \theta = \theta^*\right) = \sum_i \sum_j j \Pr(U_i = j \mid \theta = \theta^*)$$

สามารถนำไปใช้เพื่อหาคะแนนจุดตัดที่แท้จริงบนเมทริกซ์ของคะแนนรวม τ_1, τ_2, ..., τ_{K-1} ดังนั้นขั้นตอนที่เหมือนกับที่อธิบายไว้ข้างต้นสามารถนำไปใช้ได้

ดังที่กล่าวมาจะเห็นได้ว่าวิธีการประมาณค่าดัชนีความถูกต้องของการจำแนกประเภทมีความแตกต่างกันที่เห็นได้อย่างชัดเจนคือแนวคิดพื้นฐานในการพัฒนาวิธีการประมาณค่า ซึ่งมีทั้งวิธีการตามทฤษฎีการทดสอบแบบดั้งเดิม (CTT) อันประกอบด้วยวิธีการที่พัฒนาโดย Livingtion และ Lewis (1995) และวิธีการตามทฤษฎีการตอบสนองข้อสอบ (IRT) อันประกอบด้วยวิธีการที่พัฒนาโดย Rudner (2005), Guo (2006) และ Lee (2010) ซึ่งแต่ละวิธีการจะมีความแตกต่างกันในเรื่องของ

ทฤษฎีการทดสอบ ข้อตกลงเบื้องต้น วิธีการคำนวณ และลักษณะของข้อมูลที่ใช้ในการประมาณค่า โดยแสดงได้ดังตารางต่อไปนี้

ตารางที่ 2.10 ลักษณะเฉพาะของวิธีการประมาณค่าความถูกต้องของการจำแนกที่พัฒนาโดยนักวัดผลทางการศึกษา

วิธีการประมาณค่า	ทฤษฎีการทดสอบ		ข้อตกลงเบื้องต้น	ลักษณะข้อมูล					
				รูปแบบของข้อสอบ		รูปแบบของคะแนนสอบ		รูปแบบของคะแนนจุดตัด	
	CTT	IRT		Dichotomous item	Polytomous item	Raw score	Scale score	Raw score	Scale score
Livington & Lewis (1995)	✓		คะแนนที่สังเกตได้มีการแจกแจงเบต้าแบบทวินาม (beta-binomial distribution), คะแนนจริงมีการแจกแจงแบบเบต้า เบต้าแบบทวินาม (beta distribution)	✓	✓	✓		✓	
Rudner (2005)		✓	ความคลาดเคลื่อนของการประมาณค่าคะแนนจริงมีการแจกแจงแบบปกติ (normal distribution)	✓	✓		✓		✓
Guo (2006)		✓	ฟังก์ชันความน่าจะเป็นในการตอบข้อสอบของผู้สอบมีการแจกแจงแบบปกติ (normal distribution)	✓	✓		✓		✓
Lee (2010)		✓	คะแนนจริงมีการแจกแจงแบบทวินามเชิงซ้อน เบต้าแบบทวินาม (compound binomial distribution)	✓	✓		✓	✓	

ตอนที่ 4 งานวิจัยที่เกี่ยวข้องกับดัชนีการจำแนกประเภท

การศึกษางานวิจัยที่เกี่ยวกับดัชนีการจำแนกประเภท (classification indices) ทางด้านการศึกษาของต่างประเทศมีวัตถุประสงค์เพื่อได้ข้อมูลพื้นฐานเกี่ยวกับดัชนีการจำแนกประเภทในบริบทการศึกษาของต่างประเทศ ดำเนินการโดยรวบรวมบทความวิจัยหรือบทคัดย่อที่เกี่ยวกับดัชนีการจำแนกประเภทจากฐานข้อมูล CU Reference Databases ของศูนย์วิทยทรัพยากร จุฬาลงกรณ์มหาวิทยาลัย ผลจากการวิเคราะห์เนื้อหา พบว่า ส่วนใหญ่เป็นการพัฒนาวิธีการประมาณค่าดัชนีการจำแนกประเภท โดยงานวิจัยที่เกี่ยวกับการพัฒนาวิธีการประมาณค่าดัชนีการจำแนกประเภทนั้น มีทั้งการพัฒนาวิธีการประมาณค่าดัชนีความสอดคล้องของการจำแนกประเภท (classification consistency indices) และดัชนีความถูกต้องของการจำแนกประเภท (classification accuracy indices) สามารถแบ่งได้เป็นการพัฒนาภายใต้พื้นฐานของทฤษฎีการทดสอบแบบดั้งเดิม (CTT) และทฤษฎีการตอบสนองข้อสอบ (IRT) ซึ่งภายใต้แต่ละพื้นฐานนั้นก็มีการพัฒนาวิธีการประมาณค่าที่แตกต่างกันไปตามแนวคิดที่นักวิจัยนำมาใช้

การพัฒนาภายใต้พื้นฐานของทฤษฎีการทดสอบแบบดั้งเดิม (CTT) ประกอบด้วยงานวิจัยของ Huynh (1976), Subkoviak (1976, 1988), Hanson และ Brennan (1990) และ Livingston และ Lewis (1995) ส่วนการพัฒนาภายใต้พื้นฐานของทฤษฎีการตอบสนองข้อสอบ (IRT) นั้น ประกอบด้วยงานวิจัยของ Rudner (2001, 2005), Lee, Hanson และ Brennan (2002), Lee (2010), Guo (2006) และ Wyse และ Hao (2012) โดยรายละเอียดของงานวิจัยในส่วนนี้ได้กล่าวถึงแล้วในหัวข้อวิธีการประมาณค่าดัชนีความสอดคล้องของการจำแนกประเภท และวิธีการประมาณค่าดัชนีความถูกต้องของการจำแนกประเภท ข้างต้น

นอกจากการพัฒนาภายใต้พื้นฐานของทฤษฎีการทดสอบทั้งสองแล้วยังมีงานวิจัยของ Cui, Gierl และ Chang (2012) ที่ได้ทำการพัฒนาวิธีการประมาณค่าความสอดคล้องของการจำแนกประเภทและความถูกต้องของการจำแนกประเภทสำหรับการประเมินวินิจฉัยทางพุทธิปัญญาอีกด้วย เนื่องจากในการประเมินวินิจฉัยทางพุทธิปัญญาดำเนินการจำแนกกลุ่มผู้สอบอยู่ภายใต้กรอบแนวคิดของโมเดลวินิจฉัยทางพุทธิปัญญา (Cognitive Diagnostic Model: CDM) แทนการใช้คะแนนจุดตัด (cut score) ทำให้การประมาณค่าดัชนีความสอดคล้องและดัชนีความถูกต้องของการจำแนกประเภทนั้นดำเนินการภายใต้แนวคิดเบต้าแบบทวินาม (beta binomial) และทฤษฎีการตอบสนองข้อสอบ (IRT) ไม่ได้ จึงต้องพัฒนาวิธีการประมาณค่าดัชนีความสอดคล้องและความถูกต้องของการจำแนกประเภทขึ้นมาใหม่ภายใต้กรอบแนวคิดของโมเดลวินิจฉัยทางพุทธิปัญญา (cognitive diagnostic model: CDM) ซึ่งใช้การศึกษาจำลองเพื่อประเมินประสิทธิภาพของดัชนีความสอดคล้อง (classification consistency) และดัชนีความถูกต้องของการจำแนกประเภท (classification

accuracy) โดยดำเนินการจัดการภายใต้ 4 ปัจจัยดังนี้ 1) ค่าอำนาจจำแนกของข้อสอบ (item discrimination power) มีสองเงื่อนไข คือเงื่อนไขที่มีค่าอำนาจจำแนกของข้อสอบต่ำและมีค่าอำนาจจำแนกของข้อสอบสูง 2) จำนวนคุณลักษณะรวมที่ต้องการจะวัด (total number of attributes measured by the test) มีสามเงื่อนไข คือเงื่อนไขที่ต้องการวัดคุณลักษณะ 3, 5 และ 8 คุณลักษณะ 3) ความเป็นอิสระระหว่างคุณลักษณะที่ต้องการจะวัด (dependency among the attributes) มีสองเงื่อนไขคือจะใช้ dichotomized multivariate normal distribution เมื่อคุณลักษณะมีความสัมพันธ์กัน และใช้ uniform distribution เมื่อคุณลักษณะไม่สัมพันธ์กัน 4) ขนาดกลุ่มตัวอย่าง (sample size) มีสามเงื่อนไขคือมีกลุ่มตัวอย่างจำนวน 100, 500 และ 1,000 คน ผลการวิจัยได้เสนอแนะขั้นตอนในการคำนวณและการสรุปอ้างอิงเชิงเส้นทางสถิติสำหรับดัชนีความสอดคล้อง (classification consistency) และดัชนีความถูกต้องของการจำแนกประเภท (classification accuracy) โดยเฉพาะอย่างยิ่งการออกแบบสำหรับการประเมินวินิจฉัยทางพุทธิพิสัย (cognitive diagnostic assessment) ดัชนีการจำแนกประเภทแบบใหม่สามารถนำไปใช้ในฐานะที่เป็นตัวบ่งชี้ที่สำคัญของความเที่ยง (reliability) และความตรง (validity) ของผลการจำแนกประเภทที่เกิดจากการประเมินวินิจฉัยทางพุทธิปัญญา (cognitive diagnostic assessment) สำหรับแบบสอบที่รู้ค่าพารามิเตอร์ของข้อสอบ หรือแบบสอบที่ได้รับการปรับค่าพารามิเตอร์ของข้อสอบให้เป็นมาตรฐานแล้ว การแจกแจงของกลุ่มตัวอย่าง (sampling distribution) ของสองดัชนีใหม่แสดงเป็นปกติวิสัยเชิงเส้นกำกับ (asymptotically normal) เพื่ออธิบายถึงวิธีการคำนวณดัชนีใหม่นี้ ซึ่งสามารถนำไปใช้กับข้อมูลในการวินิจฉัยจริงจากการทดสอบการลบเศษส่วน (Tatsuoka) ทั้งยังใช้กับข้อมูลจำลองเพื่อประเมินผลการดำเนินงานและคุณสมบัติการแจกแจงของสองดัชนีได้อีกด้วย

รองลงมาคือ การนำวิธีการประมาณค่าดัชนีการจำแนกประเภทไปใช้เพื่อตรวจสอบหาค่าดัชนี งานวิจัยในส่วนนี้จะเป็นการนำวิธีการประมาณค่าดัชนีการจำแนกประเภทไปใช้เพื่อตรวจสอบหาค่าดัชนี ซึ่งมีทั้งค่าดัชนีความสอดคล้องของการจำแนกประเภท (classification consistency indices) และดัชนีความถูกต้องของการจำแนกประเภท (classification accuracy indices) โดยทำการศึกษาในบริบทของการบริหารจัดการการทดสอบที่แตกต่างกันไป มีงานวิจัยดังนี้Hubregtse และ Eggen (2012), Kapoor และ Welch (2011), Wheadon และ Stockford (2010), Bramley (2010), Lee, Brennan และ Wan (2009), Betebenner, Zhang (2008) และ Cheng (2008)

อันดับสามคือ การศึกษาเปรียบเทียบวิธีการประมาณค่าดัชนีการจำแนกประเภทต่างๆ งานวิจัยในส่วนนี้จะเป็นการนำวิธีการประมาณค่าดัชนีการจำแนกประเภททั้งค่าดัชนีความสอดคล้องของการจำแนกประเภท (classification consistency indices) และดัชนีความถูกต้องของการจำแนกประเภท (classification accuracy indices) มาเปรียบเทียบกัน โดยทำการศึกษาจำลองภายใต้เงื่อนไขต่างๆ ที่ผู้วิจัยได้ศึกษามาเป็นอย่างดีแล้ว รวมถึงการเปรียบเทียบค่าดัชนีการจำแนก

ประเภทภายใต้บริบทของการบริหารจัดการการทดสอบที่แตกต่างกันด้วย งานวิจัยเหล่านี้ได้แก่ งานวิจัยของ Lathrop และ Cheng (2013) และงานวิจัยของ Zhang (2010)

สุดท้ายคือ การศึกษาถึงปัจจัยที่ส่งผลต่อค่าดัชนีการจำแนกประเภท งานวิจัยในส่วนนี้จะเป็นการศึกษาถึงปัจจัยที่ส่งผลต่อค่าดัชนีการจำแนกประเภท (classification indices) โดยทำการศึกษาถึงสิ่งที่ทำให้ค่าดัชนีความสอดคล้อง (classification consistency) และความถูกต้องของการจำแนกประเภท (classification accuracy) มีค่าสูงหรือต่ำ เพื่อนำผลที่ได้ไปปรับปรุงการบริหารจัดการการทดสอบหรือวิธีการที่ใช้ในการจำแนกประเภทต่อไป ประกอบด้วยงานวิจัยของ Wyse (2011) และ Suaklay, Lawthong & Kanjanawasee (2016)

สามารถนำงานวิจัยที่ศึกษามาสรุปเป็นตารางแจกแจงตามประเด็นที่ศึกษาได้ดังนี้

ตารางที่ 2.11 การแจกแจงความถี่ตามประเด็นที่พบจากงานวิจัยที่ศึกษาเกี่ยวกับดัชนีการจำแนกประเภท

ประเด็นในการวิจัย	ดัชนีการจำแนกประเภท		ทฤษฎีการทดสอบ		ปัจจัยที่ส่งผลต่อดัชนีการจำแนกประเภท
	ความสอดคล้อง	ความถูกต้อง	CTT	IRT	
1. การพัฒนาวิธีการประมาณค่าดัชนีการจำแนกประเภท					
Huynh (1976)		✓	✓		
Subkoviak (1976, 1988)		✓	✓		
Hanson & Brennan (1990)	✓		✓		
Livingston & Lewis (1995)	✓	✓	✓		
Rudner (2001, 2005)		✓		✓	
Lee, Hanson & Brennan (2002)	✓	✓		✓	
Lee (2010)	✓	✓		✓	
Guo (2006)		✓		✓	
Wyse & Hao (2012)	✓			✓	
Cui, Gierl และ Chang (2012)	✓	✓		✓ *สำหรับ CDM	- ค่าอำนาจจำแนกของข้อสอบ - จำนวนคุณลักษณะที่ต้องการวัด - ความเป็นอิสระระหว่างคุณลักษณะฯ - ขนาดกลุ่มตัวอย่าง
2. การตรวจสอบหาค่าดัชนีการจำแนกประเภท					
Hubregtse และ Eggen (2012)		✓		✓	- รูปแบบการแจกแจงความสามารถ - การกำหนดมาตรฐาน - ความเป็นไปได้ของการชดเชย

ประเด็นในการวิจัย	ดัชนีการจำแนกประเภท		ทฤษฎีการทดสอบ		ปัจจัยที่ส่งผลต่อดัชนีการจำแนกประเภท
	ความสอดคล้อง	ความถูกต้อง	CTT	IRT	
Kapoor และ Welch (2011)	✓		✓		- รูปแบบการบริหารจัดการการทดสอบ
Wheadon และ Stockford (2010)	✓	✓	✓	✓	
Bramley (2010)	✓	✓	✓		
Lee, Brennan และ Wan (2009)	✓	✓	✓		
Zhang (2008)		✓	✓	✓	- โมเดลการวัด Ctt&irt
Cheng (2008)	✓	✓	✓		- วิธีการคัดเลือกข้อสอบในการทดสอบ CAT
3. การเปรียบเทียบวิธีการประมาณค่าดัชนีการจำแนกประเภท					
Lathrop และ Cheng (2013)		✓		✓	- โมเดลการวัด - ขนาดกลุ่มตัวอย่าง - ความยาวของแบบสอบ - ตำแหน่งของคะแนนจุดตัด
Zhang (2010)	✓	✓	✓	✓	- โมเดลการวัด
4. การศึกษาปัจจัยที่ส่งผลต่อค่าดัชนีการจำแนกประเภท					
Wyse (2011)		✓		✓	- สารสนเทศของแบบสอบ - วิธีการกำหนดคะแนนจุดตัด - ขนาดกลุ่มตัวอย่าง
Suaklay, Lawthong & Kanjanawasee (2016)	✓	✓		✓	- ขนาดกลุ่มตัวอย่าง - สหสัมพันธ์ภายในแบบสอบ

การศึกษางานวิจัยที่เกี่ยวข้องข้างต้นสามารถนำข้อค้นพบที่ได้จากงานวิจัยดังกล่าวมาสรุปเป็นประเด็นที่ควรพิจารณาในการดำเนินการประมาณค่าดัชนีการจำแนกประเภท เพื่อให้สามารถเลือกวิธีการประมาณค่าได้สอดคล้องกับข้อมูลที่มี อันจะนำไปสู่ผลการประมาณค่าดัชนีที่มีประสิทธิภาพ โดยองค์ประกอบที่ควรพิจารณามีดังต่อไปนี้

4.1 โมเดลสำหรับการประมาณค่าดัชนีการจำแนกประเภท

บทบาทของโมเดลในการประมาณค่าดัชนีการจำแนกประเภทคือ เพื่อที่จะประมาณค่าการแจกแจงของคะแนนจริง และเพื่อที่จะทำนายการแจกแจงของคะแนนที่สังเกตได้ของการบริหารจัดการการทดสอบที่แตกต่างกันในเรื่องเงื่อนไขของแบบสอบเกี่ยวกับในระดับของคะแนน

จริง โดยสมมติว่าโมเดลการวัด (measurement models) สำหรับข้อมูลของแบบสอบ วิธีการบริหารจัดการการทดสอบเดียว (single-administration) ที่ประมาณค่าการแจกแจงของคะแนนจริง (true score) และการแจกแจงของคะแนนที่สังเกตได้แบบมีเงื่อนไข (conditional observed score) เมื่อกำหนดตารางการจรณ์ของการจำแนกประเภทเป็นแบบ J × J และดัชนีความเห็นพ้องต้องกัน P และ kappa สามารถคำนวณได้จากตารางนั้น ค่าพารามิเตอร์ของโมเดล การแจกแจงของคะแนนจริงและคะแนนที่สังเกตได้ และดัชนีการจำแนกประเภททั้งหมดที่ได้รับการประมาณค่าบนฐานของข้อมูลจริงจากการบริหารจัดการการทดสอบเดียว (Lee et al., 2002) โดยโมเดลการวัดที่เป็นที่นิยมใช้ในวิธีการประมาณค่าสำหรับการบริหารจัดการการทดสอบเดียว (single administration) มีทั้งโมเดลที่อยู่บนพื้นฐานของทฤษฎีการทดสอบแบบดั้งเดิม อันประกอบไปด้วย โมเดลทวินาม (binomial model), โมเดลเบต้าแบบทวินามสองพารามิเตอร์ (two-parameter beta binomial model: 2PB), โมเดลเบต้าแบบทวินามสี่พารามิเตอร์ (four-parameter beta binomial model: 4PB), โมเดลอเนกนาม (multinomial model), โมเดลอเนกนามเชิงซ้อน (compound multinomial model: CM) และโมเดลที่อยู่บนพื้นฐานของทฤษฎีตอบสนองข้อสอบ (item response theory modesl: IRT models)

4.2 ข้อตกลงเบื้องต้นสำหรับการประมาณค่าดัชนีการจำแนกประเภท

ส่วนนี้เป็นการสรุปถึงข้อตกลงเบื้องต้นของวิธีการประมาณค่าดัชนีการจำแนกประเภทที่ได้กล่าวถึงไปข้างต้น จะเห็นได้ว่าวิธีการประมาณค่าดัชนีการจำแนกประเภทสามารถแบ่งโดยใช้พื้นฐานทางทฤษฎีการทดสอบออกได้เป็น 2 แนวคิด คือ วิธีการที่ใช้แนวคิดของทฤษฎีการทดสอบแบบดั้งเดิมเป็นฐาน (CTT-based method) และวิธีการที่ใช้แนวคิดของทฤษฎีการตอบสนองข้อสอบเป็นฐาน (IRT-based method) โดยในแต่ละแนวคิดก็มีข้อตกลงเบื้องต้นในการใช้ที่แตกต่างกันสามารถอธิบายได้ดังนี้

4.2.1 โมเดลของคะแนนจริง

ช่วงแรกของการพัฒนาวิธีการประมาณค่าดัชนีการจำแนกประเภทนั้น ส่วนใหญ่เป็นการพัฒนาภายใต้แนวคิดของทฤษฎีการทดสอบแบบดั้งเดิม โดยเริ่มจากการใช้โมเดลทวินาม (binomial model) ในการจำแนกในฐานะที่เป็นโมเดลของคะแนนจริง (true-score model) ที่แข็งแกร่ง เนื่องจากข้อตกลงเบื้องต้นที่สร้างขึ้นภายใต้โมเดลนี้มีความสัมพันธ์กับข้อตกลงเบื้องต้นของโมเดลการทดสอบแบบดั้งเดิม (CTT model) ที่กำหนดขึ้นสำหรับการแจกแจงของคะแนนที่สังเกตได้ของแบบสอบที่ประกอบด้วยข้อสอบแบบให้คะแนนได้สองค่า (dichotomous items) เท่านั้น วิธีการที่เป็นที่นิยมพัฒนาขึ้นโดย Huynh (1976) และ Subkoviak (1976) ต่อมา Livingston และ Lewis (1995) ได้ทำการขยายโมเดลทวินาม (binomial model) เพื่อสามารถนำไปใช้ได้กับข้อสอบแบบให้

คะแนนได้หลายค่า (polytomous items) ต่อมา Lee (2007) และ Lee et al. (2009) ได้เสนอโมเดลอเนกนาม (multinomial model) และโมเดลอเนกนามเชิงซ้อน (compound multinomial model) ขึ้น เพื่อใช้สำหรับข้อสอบแบบให้คะแนนได้หลายค่า (polytomous items)

ส่วนวิธีการที่อยู่ภายใต้กรอบแนวคิดของทฤษฎีการตอบสนองข้อสอบวิธีการต่างๆ ได้รับการพัฒนาสำหรับแบบสอบที่มีการให้คะแนนเป็นสเกลของคะแนนดิบ การศึกษาในช่วงแรกซึ่งรวมถึง Huynh (1990) ใช้โมเดลของ Rasch (Rasch model) และ Wang, Kolen & Harris (2000 cited in Lee, 2010) ใช้โมเดล polytomous IRT (polytomous IRT model) ส่วน Lee (2010) ได้พัฒนาวิธีการที่นำไปใช้สำหรับโมเดล mixture IRT (mixture IRT model) โดยวิธีการต่างๆ ที่ใช้ IRT model ในการคำนวณค่าความน่าจะเป็นเกี่ยวกับเวคเตอร์ของการตอบสนองแบบมีเงื่อนไขบนความสามารถแฝง (latent ability: θ) และจากนั้นใช้โมเดลทวินามเชิงซ้อน (compound binomial model) หรือโมเดลอเนกนาม (multinomial model) ในการคำนวณการแจกแจงแบบมีเงื่อนไขของคะแนนดิบที่สังเกตได้บนสเกลความสามารถแฝง ซึ่งทำการรวมการแจกแจงของคะแนนดิบทั้งหมดของทุกความสามารถแฝง โดยทำตามข้อตกลงเบื้องต้นของการแจกแจงความสามารถแฝง หรือการใช้การประมาณค่าความสามารถแฝงโดยเฉพาะ

อีกวิธีการหนึ่งที่ต่างออกไปคือวิธีการที่พัฒนาโดย Rudner (2001, 2005) ซึ่งได้ทำการพัฒนาวิธีการสำหรับแบบสอบที่มีการให้คะแนนบนสเกลของความสามารถแฝง ซึ่งมีข้อตกลงเบื้องต้นว่าการแจกแจงแบบมีเงื่อนไขของการประมาณค่าความสามารถ $\hat{\theta}$ ตามการแจกแจงปกติที่มีค่าเฉลี่ยของความสามารถแฝงและส่วนเบี่ยงเบนมาตรฐานของ SE($\hat{\theta}$) และต่อมาในปี ค.ศ. 2006 Guo (2006) ได้ทำการขยายวิธีการประมาณค่าความถูกต้องของการจำแนกประเภทที่ Rudner พัฒนาไว้ไปยังการประมาณค่าความสอดคล้องของการจำแนกประเภท

4.2.2 วิธีการประมาณค่าการแจกแจงของคะแนนจริง

ข้อตกลงเบื้องต้นที่สร้างขึ้นมาสำหรับการแจกแจงของคะแนนจริง (true score distributions) สามารถแบ่งได้เป็น 2 ประเภท ประเภทแรกคือ วิธีการแจกแจง (distributional approach) และประเภทที่สองคือ วิธีการเฉพาะตัว (individual approach) (Wheadon & Stockford, 2010) โดยวิธีการแจกแจง (distributional approach) มีข้อตกลงเบื้องต้นเกี่ยวกับการแจกแจงสำหรับความสามารถจริง (true abilities) เช่น วิธีการของ Huynh (1976) และวิธีการของ Hanson & Brennan (1990) ส่วนวิธีการของ Livingston & Lewis (1995) มีข้อตกลงเบื้องต้นโดยใช้กลุ่มของการแจกแจงแบบเบต้า (beta distributions) สำหรับการแจกแจงของคะแนนจริง (true score distributions) ส่วนวิธีการเฉพาะตัว (individual approach) ทำการคำนวณดัชนีความสอดคล้องของการจำแนกประเภทสำหรับผู้สอบแต่ละคนเพียงครั้งเดียวและทำการหาค่าเฉลี่ยของผู้สอบทั้งหมดโดยปราศจากการสร้างข้อตกลงเบื้องต้นเกี่ยวกับความสามารถที่แท้จริง

ตัวอย่างเช่น วิธีการของ Subkoviak (1976), Lee (2007), Lee et al., (2009) และ Brennan และ Wan (2004)

4.2.3 การแจกแจงของคะแนนที่สังเกตได้

เนื่องจากความซับซ้อนในการคำนวณการแจกแจงเบต้าแบบทวินาม (beta-binomial distribution) นักวิจัยบางคนจึงได้เสนอวิธีการประมาณค่าปกติโดยมีข้อตกลงเบื้องต้นว่า การแจกแจงของคะแนนที่สังเกตได้ (observed score distribution) จากการบริหารจัดการการทดสอบสองสถานการณ์เป็นการแจกแจงแบบปกติสองทาง (bivariate normal distribution) ที่มีสหสัมพันธ์เท่ากับความเที่ยงของคะแนนสอบ นอกจากนี้ยังมีอีกหลายวิธีซึ่งแต่ละวิธีแตกต่างกันไปในแนวทางที่ใช้ในการคำนวณค่าความเที่ยง (reliability) เช่น Peng & Subkoviak (1980 cited in Lee, 2010) ใช้สัมประสิทธิ์ KR-21 เป็นความเที่ยงของแบบสอบ ส่วน Breyer & Lewis (1994 cited in Lee, 2010) ใช้วิธีแบ่งครึ่งข้อสอบแต่กำหนดคะแนนจุดตัดสำหรับแต่ละครึ่งของแบบสอบ และใช้สหสัมพันธ์ tetrachoric ในการคำนวณค่าความเที่ยง

4.2.4 ข้อตกลงเบื้องต้นของทฤษฎีการตอบสนองข้อสอบ (สำหรับ IRT-based method)

วิธีการที่ใช้แนวคิดของทฤษฎีการตอบสนองข้อสอบเป็นฐาน (IRT-based method) นั้นพัฒนาขึ้นพร้อมกับความนิยมที่เพิ่มขึ้นของการใช้งานทฤษฎีการตอบสนองข้อสอบในด้านต่างๆ ของการปฏิบัติเกี่ยวกับการทดสอบ โดยพื้นฐานแล้วทุกวิธีการที่พัฒนาโดยใช้ทฤษฎีการตอบสนองข้อสอบเป็นฐานนั้นล้วนตั้งอยู่บนพื้นฐานของการใช้ข้อตกลงเบื้องต้นเดียวกันกับการประยุกต์ใช้ทฤษฎีการตอบสนองข้อสอบในด้านต่างๆ รวมถึงความเป็นเอกมิติ (unidimensionality) ความเป็นอิสระระหว่างข้อสอบ (local item independency) และความสอดคล้องของโมเดล (model fit) (Lee, 2010) นอกจากนี้ขนาดของตัวอย่างก็ยังเป็นสิ่งที่จำเป็นสำหรับการประมาณค่าพารามิเตอร์ของข้อสอบที่ถูกต้องอีกด้วย ขณะเดียวกันก็ไม่สามารถทราบได้เลยว่าความคลาดเคลื่อนแบบสุ่ม ที่ได้มาจากการประมาณค่าพารามิเตอร์ของข้อสอบนั้นเกิดขึ้นเนื่องจากตัวอย่างขนาดเล็ก ซึ่งอาจจะกำลังดำเนินการอยู่ในการประมาณค่าความสอดคล้องและความถูกต้องของการจำแนกประเภทของการบริหารจัดการการทดสอบเดียวก็เป็นได้

4.2.5 ค่าน้ำหนักของข้อสอบที่แตกต่างกัน

ค่าน้ำหนักของคะแนนที่แตกต่างกันในบางครั้งมีความเกี่ยวข้องกับรูปแบบของข้อสอบที่แตกต่างกัน ดังนั้นคะแนนรวมที่ได้มาจากการรวมกันของข้อสอบแต่ละประเภทจึงควรจัดการอย่างมีประสิทธิภาพ เพราะค่าน้ำหนักนี้สามารถนำไปรวมเข้าด้วยกันในสูตรที่ซ้ำกับขั้นตอนเดิมเพื่อให้ได้การแจกแจงของคะแนนรวมต่อไป (summed-score distribution) (Wheadon & Stockford, 2010; Lee, 2010)

4.2.6 ประเภทของข้อสอบ

โดยพิจารณาว่าเป็นข้อสอบที่มีการให้คะแนนได้สองค่า (dichotomous item) ได้หลายค่า (polytomous item) หรือแบบซับซ้อนคือรวมการให้คะแนนทั้งสองอย่างเข้าด้วยกัน (complex item) (Wheadon & Stockford, 2010)

4.2.7 ประเภทของคะแนน

โดยพิจารณาว่าคะแนนที่ได้มีลักษณะเป็นคะแนนดิบ (raw scores) คะแนนมาตรฐาน (scale scores) หรือคะแนนรวมจากชุดข้อสอบที่หลากหลาย (composite scores) (Wheadon & Stockford, 2010)

4.2.8 ความเหมาะสมของซอฟแวร์ที่ใช้ในการดำเนินการ

ซอฟแวร์หรือโปรแกรมที่ใช้ในการประมาณค่าดัชนีการจำแนกประเภทมีความสอดคล้องเหมาะสมกับวิธีการประมาณค่าดัชนีการจำแนกประเภทใช้ และเป็นไปตามข้อตกลงเบื้องต้นของวิธีการประมาณค่านั้นๆ (Wheadon & Stockford, 2010)

4.3 ปัจจัยที่ส่งผลต่อดัชนีการจำแนกประเภท

สำหรับปัจจัยที่ส่งผลต่อดัชนีการจำแนกประเภทนั้น พบว่าในงานวิจัยเกี่ยวข้องมีการศึกษาถึงปัจจัยที่อาจจะส่งผลต่อดัชนีการจำแนกประเภทหลายปัจจัย ซึ่งจากงานวิจัยส่วนใหญ่ให้ผลที่สอดคล้องกันว่าปัจจัยที่นำมาศึกษาเกือบทั้งหมดส่งผลต่อดัชนีการจำแนกประเภทจะมีเพียงบางส่วนเท่านั้นที่ไม่ส่งผลต่อดัชนีการจำแนกประเภท ซึ่งประกอบด้วยปัจจัยดังต่อไปนี้

4.3.1 โมเดลการวัด (measurement model)

จากงานวิจัยที่ศึกษามีการใช้โมเดลการวัดที่หลากหลาย ทั้งโมเดลการวัดตามทฤษฎีการทดสอบแบบดั้งเดิม (CTT) และโมเดลการวัดตามทฤษฎีการตอบสนองข้อสอบ (IRT) ซึ่งผลการวิจัยพบว่าการเลือกใช้โมเดลการวัดที่สอดคล้องกับวิธีการประมาณค่าดัชนีการจำแนกประเภทจะส่งผลให้ได้ค่าการประมาณที่ดี

ดังงานวิจัยของ Lathrop และ Cheng (2013) ได้ทำการเปรียบเทียบดัชนีความถูกต้องของการจำแนกประเภท (classification accuracy) ที่ประมาณค่าได้จากวิธีการของ Lee (Lee approach) กับวิธีการของ Rudner (Rudner approach) ซึ่งเป็นวิธีการประมาณค่าที่อยู่บนพื้นฐานของทฤษฎีการตอบสนองข้อสอบทั้งคู่ ภายใต้โมเดลการวัดตามทฤษฎีการตอบสนองข้อสอบ (IRT model) ที่แตกต่างกันสี่โมเดล คือ โมเดลโลจิสติกแบบหนึ่งพารามิเตอร์ (one-parameter logistic model: 1PL) โมเดลโลจิสติกแบบสองพารามิเตอร์ (two-parameter logistic model: 2PL) โมเดลโลจิสติกแบบสามพารามิเตอร์ (three-parameter logistic model: 3PL) และ graded response model (GRM) ผลการวิจัยพบว่าวิธีการประมาณค่าทั้งสองวิธีการให้ค่าการประมาณที่แตกต่างกันเพียงเล็กน้อย เนื่องจาก

ทั้งสี่โมเดลที่นำมาใช้ในการวิจัยเป็นโมเดลการวัดตามทฤษฎีการตอบสนองข้อสอบ (IRT model) จึงมีความสอดคล้องกับวิธีการประมาณค่าทั้งสองวิธี

งานวิจัยของ Lee (2010) ได้ทำการพัฒนาวิธีการประมาณค่าความสอดคล้อง (classification consistency) และความถูกต้องของการจำแนกประเภท (classification accuracy) สำหรับการประเมินที่ซับซ้อนด้วยทฤษฎีการตอบสนองข้อสอบ (IRT) การศึกษาครั้งนี้ใช้ข้อมูลจริง (real data) จำนวนสองชุด ในแต่ละชุดประกอบด้วยคะแนนสอบจากข้อสอบที่มีการให้คะแนนแบบผสมระหว่างข้อสอบที่ให้คะแนนได้สองค่าและข้อสอบที่ให้คะแนนได้หลายค่า และทำการเปรียบเทียบวิธีการประมาณค่าความสอดคล้องและความถูกต้องของการจำแนกประเภทภายใต้การรวมโมเดลตามทฤษฎีการตอบสนองข้อสอบ (IRT model combinations) จำนวน 6 ชุด ดังนี้ 1) one-parameter logistic model + partial credit medel (1PL+PC) 2) two-parameter logistic model + generalized partial credit medel (2PL+GPC) 3) three-parameter logistic model + generalized partial credit medel (3PL+GPC) 4) one-parameter logistic model + graded response model (1PL+GR) 5) two-parameter logistic model + graded response medel (2PL+GR) และ 6) three-parameter logistic model + graded response medel (3PL+GR) โดยวิธีการที่นำมาเปรียบเทียบมี 3 วิธี ประกอบด้วย 1) วิธีการของ Lee ซึ่งอยู่บนแนวคิดของทฤษฎีการตอบสนองข้อสอบ 2) วิธีการของ Livingtons และ Lewis และ 3) วิธีการ compound multinomial โดยวิธีการที่ 2) และ 3) เป็นวิธีการที่ไม่ได้อยู่บนแนวคิดของทฤษฎีการตอบสนองข้อสอบ ผลการวิจัยพบว่า วิธีการของ Lee ให้ค่าการประมาณที่ดีกว่า เนื่องจากทั้งหกโมเดลที่นำมาใช้ในการวิจัยเป็นโมเดลการวัดตามทฤษฎีการตอบสนองข้อสอบ (IRT model) จึงมีความสอดคล้องกับวิธีการประมาณค่าที่อยู่บนแนวคิดทฤษฎีการตอบสนองข้อสอบซึ่งคือวิธีของ Lee นั่นเอง

4.3.2 ขนาดกลุ่มตัวอย่าง (sample size)

จากงานวิจัยที่ศึกษามีการใช้ขนาดของกลุ่มตัวอย่างที่หลากหลาย ซึ่งผลการวิจัยพบว่าการกำหนดขนาดกลุ่มตัวอย่างในปริมาณมากหรือกลุ่มตัวอย่างขนาดใหญ่จะส่งผลให้ได้ค่าการประมาณที่มีประสิทธิภาพดังงานวิจัยต่อไปนี้

Lathrop และ Cheng (2013) ได้ทำการเปรียบเทียบดัชนีความถูกต้องของการจำแนกประเภท (classification accuracy) ที่ประมาณค่าได้จากวิธีการของ Lee (Lee approach) กับวิธีการของ Rudner (Rudner approach) ภายใต้ขนาดของกลุ่มตัวอย่างที่แตกต่างกัน โดยทำการจำลองข้อมูลขนาดกลุ่มตัวอย่างออกเป็นจำนวน 250 คน 500 คน และ 1,000 คน ผลการวิจัยพบว่าขนาดกลุ่มตัวอย่างที่เพิ่มขึ้นจะส่งผลให้ได้ค่าการประมาณที่มีประสิทธิภาพ โดยพิจารณาจากค่าความคลาดเคลื่อนมาตรฐาน (standard error: SE) ที่ลดลง

Cui, Gierl และ Chang (2012) ทำการพัฒนาวิธีการประมาณค่าดัชนีความสอดคล้องของการจำแนกประเภทและดัชนีความถูกต้องของการจำแนกประเภทภายใต้กรอบแนวคิดของโมเดลวินิจฉัยทางพุทธิปัญญา (Cognitive Diagnostic Model: CDM) ซึ่งทำการศึกษาจำลองภายใต้เงื่อนไขขนาดกลุ่มตัวอย่าง โดยแบ่งขนาดของกลุ่มตัวอย่างออกเป็น 3 เงื่อนไข คือ ขนาดกลุ่มตัวอย่าง 100 คน 500 คน และ 1,000 คน ผลการวิจัยพบว่าขนาดกลุ่มตัวอย่างที่เพิ่มขึ้นจะส่งผลให้ได้ค่าการประมาณที่มีประสิทธิภาพ

แต่ในทางกลับกันมีงานวิจัยจำนวนหนึ่งที่พบว่าขนาดกลุ่มตัวอย่างไม่ส่งผลต่อค่าดัชนีการจำแนกประเภท ได้แก่ งานวิจัยของ Wyse และ Hao (2012) ได้ทำการตรวจสอบขั้นต้นเกี่ยวกับจำนวนกลุ่มตัวอย่างที่แตกต่างกันพบว่า การใช้กลุ่มตัวอย่างจำนวน 2,000 คน ให้ผลการประมาณค่าดัชนีการจำแนกประเภทที่ใกล้เคียงกับการใช้กลุ่มตัวอย่างจำนวน 10,000 หรือ 25,000 คน และงานวิจัยของ Suaklay, Lawthong & Kanjanawasee (2016) ที่พบว่าขนาดกลุ่มตัวอย่างที่แตกต่างกัน 4 เงื่อนไข ได้แก่ กลุ่มตัวอย่างจำนวน 100, 500, 1,000 และ 2,000 คน ไม่ส่งผลต่อค่าดัชนีความถูกต้องและดัชนีความสอดคล้องของการจำแนกประเภท ไม่ว่าจะประมาณค่าดัชนีด้วยวิธีการของ Rudner หรือ Lee ก็ตาม

4.3.3 ตำแหน่งของคะแนนจุดตัด (cut score location)

งานวิจัยส่วนใหญ่ศึกษาถึงความสัมพันธ์ระหว่างตำแหน่งของคะแนนจุดตัดกับตำแหน่งคะแนนจริง (θ) ของผู้สอบ ผลการวิจัยพบว่าเมื่อตำแหน่งของคะแนนจุดตัดอยู่ใกล้กับตำแหน่งของคะแนนจริงของผู้สอบจะส่งผลให้ค่าการประมาณที่ได้มีค่าต่ำ ดังงานวิจัยของ Lathrop และ Cheng (2013) ได้ทำการเปรียบเทียบดัชนีความถูกต้องของการจำแนกประเภท (classification accuracy) ที่ประมาณค่าได้จากวิธีการของ Lee (Lee approach) กับวิธีการของ Rudner (Rudner approach) ซึ่งเป็นวิธีการประมาณค่าที่อยู่บนพื้นฐานของทฤษฎีการตอบสนองข้อสอบทั้งคู่ ผลการวิจัยพบว่าเมื่อคะแนนจุดตัดอยู่ที่จุดสูงสุดของการแจกแจงความสามารถของผู้สอบจะส่งผลให้ค่าการประมาณที่ได้มีค่าต่ำ ซึ่งสอดคล้องกับผลการวิจัยของ Wyse (2011) ที่ทำการศึกษาถึงผลที่อาจเกิดขึ้นกับความถูกต้องของการจำแนกประเภทที่คาดหวัง (expected classification accuracy) อันเนื่องมาจากการที่ไม่สามารถสร้างแบบสอบคู่ขนานได้

4.3.4 จำนวนตำแหน่งของคะแนนจุดตัด (a number of cut score placement)

งานวิจัยของ Martineau (2007) ได้ทำการศึกษาถึงการนำการประมาณค่าความถูกต้องของการจำแนกประเภทที่คาดหวัง (expected classification accuracy) ไปใช้ในทางปฏิบัติ โดยทำการประมาณค่าความถูกต้องของการจำแนกประเภทด้วยวิธีการของ Rudner (Rudner approach) ซึ่งเป็นวิธีการประมาณค่าที่อยู่บนพื้นฐานของทฤษฎีการตอบสนองข้อสอบ ภายใต้เงื่อนไขของจำนวนตำแหน่งของคะแนนจุดตัดที่แตกต่างกัน โดยทำการจำลองข้อมูลจำนวนตำแหน่งของคะแนนจุดตัดออกเป็น 2 รูปแบบ

คือ การกำหนดกลุ่มที่ต้องการจำแนกออกเป็น 2 กลุ่ม คือมีจำนวนคะแนนจุดตัดเพียงตำแหน่งเดียว และการกำหนดกลุ่มที่ต้องการจำแนกออกเป็น 4 กลุ่ม คือมีจำนวนคะแนนจุดตัดสามตำแหน่ง ผลการวิจัยพบว่าการกำหนดคะแนนจุดตัดเพียงตำแหน่งเดียวให้ค่าความถูกต้องของการจำแนกประเภทที่คาดหวัง (expected classification accuracy) ที่ดีกว่าการกำหนดคะแนนจุดตัดหลายตำแหน่ง

4.3.4 คุณลักษณะทางจิตมิติของแบบสอบ (psychometric property of test)

คุณลักษณะทางจิตมิติของแบบสอบที่ใช้ศึกษาในงานวิจัยที่ผ่านมาประกอบด้วย ค่าสารสนเทศของแบบสอบ และค่าความเที่ยงของแบบสอบ ผลการวิจัยพบว่าแบบสอบที่มีค่าคุณลักษณะทางจิตมิติที่ดีจะส่งผลให้ได้ค่าการประมาณที่ดี ดังงานวิจัยของ MacCann และ Stanley (2010) ที่ทำการตรวจสอบความสอดคล้องของการจำแนกประเภท (classification consistency) ของคะแนนที่แปลงเป็นเกรด โดยเปรียบเทียบระหว่างคะแนนที่ได้จากการประเมินภายในโรงเรียนกับคะแนนที่ได้จากการทดสอบ ในการศึกษาครั้งนี้ทำการประมาณค่าความสอดคล้องของการจำแนกประเภทภายใต้กรอบแนวคิดของทฤษฎีการทดสอบแบบดั้งเดิม ในการศึกษาจำลองจึงต้องจัดกระทำให้แบบสอบสองฉบับเป็นแบบสอบที่คู่ขนานกัน โดยการกำหนดค่าความเที่ยงของคะแนนสอบจากแบบสอบทั้งสองฉบับให้มีค่าเท่ากัน ซึ่งการศึกษาจำลองในครั้งนี้ได้วางเงื่อนไขเกี่ยวกับค่าความเที่ยงที่แตกต่างกันจำนวน 4 ค่า คือ 0.75, 0.80, 0.85 และ 0.90 ผลการวิจัยพบว่าค่าความเที่ยงที่เพิ่มขึ้นส่งผลให้ได้ค่าการประมาณที่ดี โดยพิจารณาจากค่าร้อยละของความไม่เห็นพ้องต้องกันที่ลดลงเมื่อค่าความเที่ยงที่กำหนดให้มีค่าเพิ่มขึ้น

4.3.5 คุณลักษณะทางจิตมิติของข้อสอบ (psychometric property of item)

คุณลักษณะทางจิตมิติของข้อสอบที่กล่าวถึงนี้คือค่าอำนาจจำแนกของข้อสอบ ซึ่งพบในงานวิจัยของ Cui, Gierl และ Chang (2012) ที่ได้ทำการพัฒนาวิธีการประมาณค่าความสอดคล้อง (classification consistency) และความถูกต้องของการจำแนกประเภท (classification accuracy) สำหรับการประเมินวินิจฉัยทางพุทธิปัญญา โดยใช้การศึกษาจำลองเพื่อประมาณค่าความสอดคล้องและความถูกต้องของการจำแนกประเภทภายใต้เงื่อนไขของค่าอำนาจจำแนกของข้อสอบ 2 เงื่อนไข คือ 1) อำนาจจำแนกของข้อสอบมีค่าต่ำ และ 2) อำนาจจำแนกของข้อสอบมีค่าสูง ผลการวิจัยพบว่าข้อสอบที่มีค่าอำนาจจำแนกสูงจะส่งผลให้ได้ค่าการประมาณที่ดี

4.3.6 ความยาวของแบบสอบ (test lengths)

จากงานวิจัยของ Lathrop และ Cheng (2013) ที่ได้ทำการเปรียบเทียบดัชนีความถูกต้องของการจำแนกประเภท (classification accuracy) ที่ประมาณค่าได้จากวิธีการของ Lee (Lee approach) กับวิธีการของ Rudner (Rudner approach) ซึ่งเป็นวิธีการประมาณค่าที่อยู่บนพื้นฐานของทฤษฎีการตอบสนองข้อสอบทั้งคู่ ภายใต้ความยาวของแบบสอบที่แตกต่างกัน โดยทำการจำลองความยาวของแบบสอบออกเป็น 4 รูปแบบ คือ แบบสอบที่มีข้อสอบจำนวน 10 ข้อ 20 ข้อ 40 ข้อ และ 80 ข้อ

ผลการวิจัยพบว่าความยาวของแบบสอบที่เพิ่มขึ้นจะส่งผลให้ได้ค่าการประมาณที่มีประสิทธิภาพ โดยพิจารณาจากค่าความคลาดเคลื่อนมาตรฐาน (standard error: SE) ที่ลดลง

4.3.7 จำนวนคุณลักษณะที่ต้องการจะวัด (total number of attributes)

จากงานวิจัยของ Cui et al. (2012) ที่ได้ทำการพัฒนาวิธีการประมาณค่าความสอดคล้อง (classification consistency) และความถูกต้องของการจำแนกประเภท (classification accuracy) สำหรับการประเมินวินิจฉัยทางพุทธิปัญญา โดยใช้การศึกษาจำลองเพื่อประมาณค่าความสอดคล้องและความถูกต้องของการจำแนกประเภทภายใต้เงื่อนไขของจำนวนคุณลักษณะที่ต้องการจะวัด จำนวน 3 เงื่อนไข คือ คุณลักษณะที่ต้องการจะวัดมีจำนวน 3, 5 และ 8 คุณลักษณะ ผลการวิจัยพบว่าถ้ามีคุณลักษณะที่ต้องการจะวัดจำนวนน้อยจะส่งผลให้ได้ค่าการประมาณที่ดี

4.3.8 ความเป็นอิสระระหว่างคุณลักษณะที่ต้องการจะวัด (dependency among the attributes)

จากงานวิจัยของ Cui et al. (2012) ที่ทำการพัฒนาวิธีการประมาณค่าดัชนีความสอดคล้องของการจำแนกประเภทและดัชนีความถูกต้องของการจำแนกประเภทภายใต้กรอบแนวคิดของโมเดลวินิจฉัยทางพุทธิปัญญา (Cognitive Diagnostic Model: CDM) ซึ่งทำการศึกษาจำลองภายใต้เงื่อนไขความเป็นอิสระระหว่างคุณลักษณะที่ต้องการจะวัด โดยแบ่งคุณลักษณะที่ต้องการจะวัดออกเป็นสองกลุ่ม คือกลุ่มที่มีจำนวนคุณลักษณะที่ต้องการจะวัดแตกต่างกันเพียงเล็กน้อย หรือคุณลักษณะมีความสัมพันธ์กัน ซึ่งจะใช้การแจกแจงแบบ dichotomized multivariate normal distribution และกลุ่มที่มีจำนวนคุณลักษณะที่ต้องการจะวัดแตกต่างกันมาก หรือคุณลักษณะไม่มีความสัมพันธ์กัน ซึ่งจะใช้การแจกแจงแบบ uniform distribution ผลการวิจัยพบว่าคุณลักษณะที่ต้องการจะวัดที่มีความแตกต่างกันมากจะส่งผลให้ได้ค่าการประมาณที่ดี

4.3.9 รูปแบบการแจกแจงความสามารถของผู้สอบ (shape of the ability distribution)

โดย Hubregtse และ Eggen (2012) ได้ทำการศึกษาถึงอิทธิพลของชุดการสอบ (exam sets) ที่มีต่อความถูกต้องของการจำแนกประเภท (classification accuracy) กลุ่มตัวอย่างในการศึกษาครั้งนี้คือนักศึกษาศูนย์การศึกษาและฝึกอบรมทางวิชาชีพ การประมาณค่าความถูกต้องของการจำแนกประเภทคำนวณโดยการจำลองข้อมูลตามทฤษฎีการตอบสนองข้อสอบ (IRT) สำหรับชุดการสอบ (exam sets) ผลการวิจัยพบว่าความถูกต้องของการจำแนกประเภทจะมีค่าสูงเมื่อคะแนนสอบทั้งหมดที่ได้จากชุดการสอบหนึ่งมีการแจกแจงความสามารถที่เท่าเทียมกันและเป็นการแจกแจงมาตรฐาน

4.3.10 ความสูงของมาตรฐานในการจำแนกประเภท (height of the standards)

โดย Hubregtse และ Eggen (2012) ได้ทำการศึกษาถึงอิทธิพลของชุดการสอบ (exam sets) ที่มีต่อความถูกต้องของการจำแนกประเภท (classification accuracy) กลุ่มตัวอย่างในการศึกษาครั้งนี้คือนักศึกษาศูนย์การศึกษาและฝึกอบรมทางวิชาชีพ การประมาณค่าความถูกต้องของการจำแนกประเภทคำนวณโดยการจำลองข้อมูลตามทฤษฎีการตอบสนองข้อสอบ (IRT) สำหรับชุดการสอบ (exam sets) ผลการวิจัยพบว่าถ้ามีการกำหนดมาตรฐานที่ไม่สูงหรือต่ำจนเกินไปจะส่งผลให้ได้ค่าการประมาณที่ดี

4.3.11 สหสัมพันธ์ภายในแบบสอบ (inter-item correlation)

Suaklay, Lawthong & Kanjanawasee (2016) ได้ทำการศึกษาถึงอิทธิพลของสหสัมพันธ์ภายในแบบสอบที่มีต่อค่าดัชนีการจำแนกประเภททั้งดัชนีความถูกต้องและดัชนีความสอดคล้อง โดยใช้วิธีการประมาณค่าของ Rudner และ Lee โดยกำหนดเงื่อนไขของระดับสหสัมพันธ์ภายในแบบสอบที่แตกต่างกัน 2 เงื่อนไข ได้แก่ แบบสอบที่มีระดับสหสัมพันธ์ภายในแบบสอบต่ำคือมีค่าสหสัมพันธ์อยู่ในช่วง 0.1 ถึง 0.49 และแบบสอบที่มีระดับสหสัมพันธ์ภายในแบบสอบสูงคือมีค่าสหสัมพันธ์อยู่ในช่วง 0.5 ถึง 0.8 พบว่าภายใต้สถานการณ์ของแบบสอบที่มีระดับสหสัมพันธ์ภายในแบบสอบต่ำไม่ส่งผลต่อค่าดัชนีความถูกต้องและดัชนีความสอดคล้องของการจำแนกประเภท แต่ภายใต้สถานการณ์ของแบบสอบที่มีระดับสหสัมพันธ์ภายในแบบสอบสูงกลับส่งผลต่อค่าดัชนีความถูกต้องและดัชนีความสอดคล้องของการจำแนกประเภทอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ไม่ว่าจะประมาณค่าดัชนีด้วยวิธีการของ Rudner หรือ Lee ก็ตาม

4.3.12 ความเป็นไปได้ของการชดเชย (possibility for compensation)

โดย Hubregtse และ Eggen (2012) ได้ทำการศึกษาถึงอิทธิพลของชุดการสอบ (exam sets) ที่มีต่อความถูกต้องของการจำแนกประเภท (classification accuracy) กลุ่มตัวอย่างในการศึกษาครั้งนี้คือนักศึกษาศูนย์การศึกษาและฝึกอบรมทางวิชาชีพ การประมาณค่าความถูกต้องของการจำแนกประเภทคำนวณโดยการจำลองข้อมูลตามทฤษฎีการตอบสนองข้อสอบ (IRT) สำหรับชุดการสอบ (exam sets) ผลการวิจัยพบว่าการยอมให้มีการชดเชยได้จะส่งผลให้ได้ค่าการประมาณที่ดี

แต่ทั้งนี้ก็มีปัจจัยหนึ่งที่นำมาศึกษาแต่ไม่ส่งผลต่อค่าการประมาณที่ได้คือรูปแบบของการบริหารจัดการการทดสอบซึ่งทำการศึกษาโดย Kapoor และ Welch (2011)

ตอนที่ 5 กรอบแนวคิดในการวิจัย

จากการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องกับดัชนีการจำแนกประเภท พบว่ามีวิธีการประมาณค่าดัชนีการจำแนกประเภทที่ได้รับการพัฒนาขึ้นหลายวิธี และมีการนำมาประยุกต์ใช้จนถึงปัจจุบัน ซึ่งในการพัฒนาวิธีการประมาณค่านั้นก็มีพื้นฐานแนวคิดที่แตกต่างกัน ดังนั้นการวิเคราะห์และเปรียบเทียบค่าดัชนีที่ได้จากการประมาณค่าที่มีแนวคิดและทฤษฎีพื้นฐานที่แตกต่างกัน เพื่อตรวจสอบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทเหล่านั้นจึงเป็นสิ่งสำคัญ เพื่อให้ได้มาซึ่งสารสนเทศที่เป็นประโยชน์ต่อการนำไปประยุกต์ใช้ในสถานการณ์ต่างๆ ที่หลากหลายได้อย่างเหมาะสม โดยเฉพาะสำหรับการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET) ซึ่งเป็นการทดสอบที่มีผลได้ผลเสียสูง (high stakes test) และมีวิธีการประเมินที่ได้มาตรฐาน

นอกจากนี้การศึกษาแนวคิด ทฤษฎีพื้นฐาน วิธีการคำนวณ และข้อตกลงเบื้องต้นของแต่ละวิธีการประมาณค่าดัชนีการจำแนกประเภทแล้ว พบว่าวิธีการประมาณค่าที่มีแนวคิดพื้นฐานตามทฤษฎีการทดสอบแบบดั้งเดิม (CTT-based) นั้นยากต่อการนำไปประยุกต์ใช้ให้เหมาะสมกับสถานการณ์การทดสอบในปัจจุบันที่ส่วนใหญ่จัดขึ้นภายใต้แนวคิดของทฤษฎีการตอบสนองข้อสอบ ผู้วิจัยจึงไม่ได้สนใจที่จะศึกษาวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการทดสอบแบบดั้งเดิมเนื่องด้วยเหตุผลดังกล่าว และเนื่องจากวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบนั้นพัฒนาขึ้นพร้อมกับความนิยมที่เพิ่มขึ้นของการใช้งานทฤษฎีการตอบสนองข้อสอบในด้านต่างๆ ของการปฏิบัติเกี่ยวกับการทดสอบ โดยพื้นฐานแล้วทุกวิธีการที่พัฒนาโดยใช้ทฤษฎีการตอบสนองข้อสอบเป็นฐานนั้นล้วนตั้งอยู่บนพื้นฐานของการใช้ข้อตกลงเบื้องต้นเดียวกันกับการประยุกต์ใช้ทฤษฎี การตอบสนองข้อสอบในด้านต่างๆ รวมถึงความเป็นเอกมิติ (unidimensionality) ความเป็นอิสระระหว่างข้อสอบ (local item independency) และความเหมาะสมของโมเดลการตอบสนองข้อสอบกับข้อมูล (model fit) (Lee, 2010)

เมื่อพิจารณาแนวคิดและทฤษฎีพื้นฐานของวิธีการประมาณค่าดัชนีการจำแนกประเภททั้งหมด พบว่า วิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบ (IRT) มีจำนวน 3 วิธีการ คือ วิธีการของ Rudner (2005) วิธีการของ Guo (2006) และวิธีการของ Lee (2010) จากการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องพบว่า วิธีการประมาณค่าทั้งสามวิธีนี้มีข้อตกลงเบื้องต้นเกี่ยวกับโมเดลตามทฤษฎีการตอบสนองข้อสอบ (IRT model) เหมือนกัน แต่ต่างกันที่ข้อตกลงเบื้องต้นเกี่ยวกับการแจกแจงของคะแนนที่นำมาใช้ในการคำนวณความน่าจะเป็นในการจำแนกผู้สอบที่มีความสามารถ θ เข้าแต่ละระดับความสามารถ ลักษณะของข้อมูลที่ใช้ในการวิเคราะห์ และสูตรความน่าจะเป็นในการจำแนกผู้สอบเข้าสู่แต่ละระดับความสามารถ โดยข้อตกลงเบื้องต้นเกี่ยวกับการแจกแจงของคะแนนที่นำมาใช้ในการคำนวณความน่าจะเป็นในการจำแนกผู้สอบที่มี

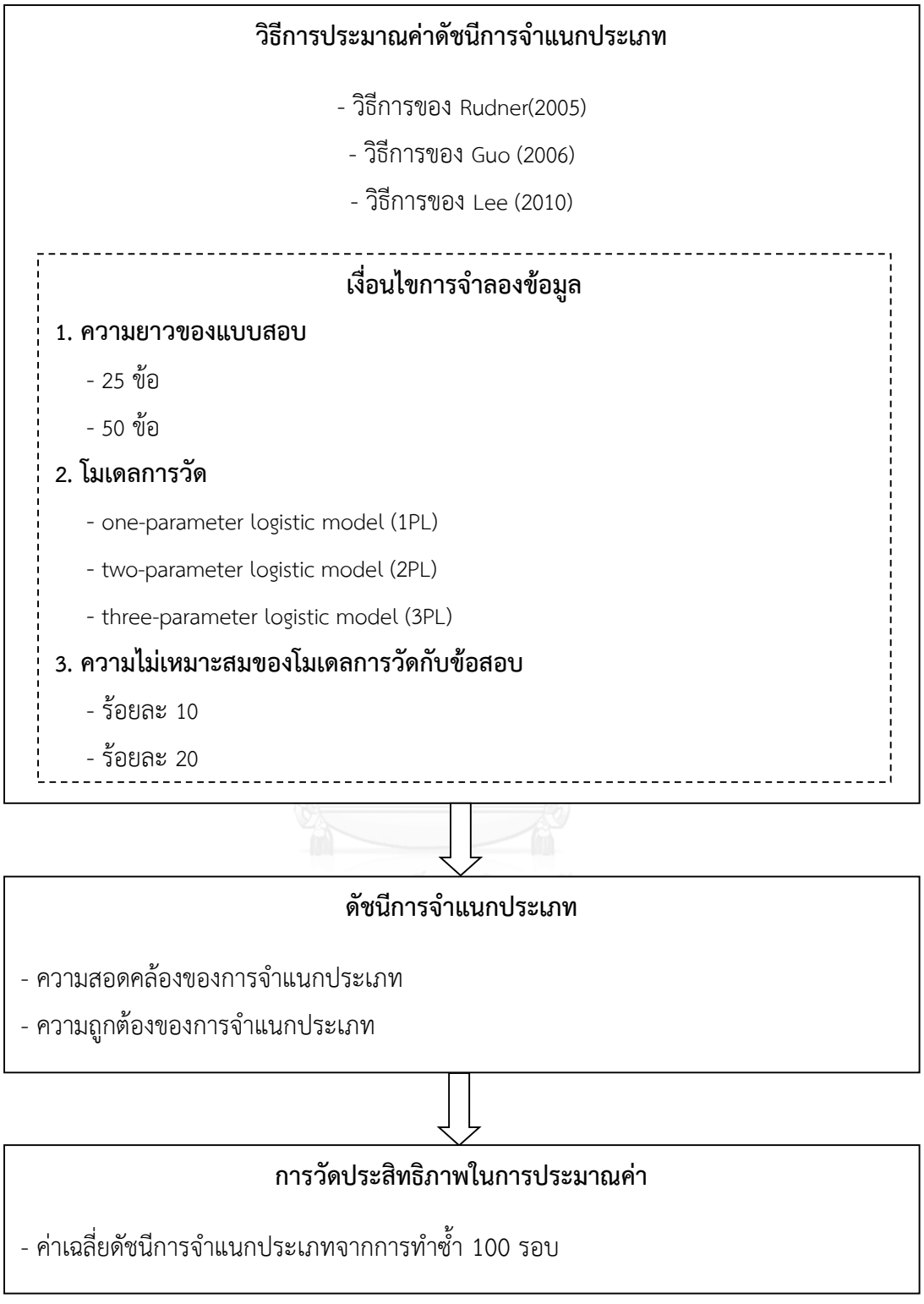
ความสามารถ θ เข้าแต่ละระดับความสามารถนั้น วิธีการของ Rudner (2005) มีข้อตกลงเบื้องต้นว่า การแจกแจงของความคลาดเคลื่อนมาตรฐานของการประมาณค่าความสามารถผู้สอบต้องเป็นการแจกแจงแบบปกติ (normal distribution) ส่วนวิธีการของ Guo (2006) มีข้อตกลงเบื้องต้นว่าการแจกแจงของฟังก์ชันความน่าจะเป็นในการตอบข้อสอบของผู้สอบต้องเป็นการแจกแจงแบบปกติ (normal distribution) และวิธีการของ Lee (2010) มีข้อตกลงเบื้องต้นว่าคะแนนจริงของผู้สอบต้องเป็นการแจกแจงแบบอเนกนามเชิงซ้อน (compound binomial distribution) ในด้านลักษณะของข้อมูลหรือคะแนนที่ใช้ในการประมาณค่านั้นมีความแตกต่างกันคือ วิธีการของ Rudner (2005) และวิธีการของ Guo (2006) มีการใช้ข้อมูลในการลักษณะเดียวกันคือใช้คะแนนที่อยู่บนสเกลของคะแนนความสามารถ (theta scale) ส่วนวิธีการของ Lee (2010) ใช้คะแนนที่อยู่บนสเกลของคะแนนรวม (summed score scale) สำหรับสูตรความน่าจะเป็นในการจำแนกผู้สอบเข้าสู่แต่ละระดับความสามารถนั้น แต่ละวิธีการมีความแตกต่างกันอันเนื่องมาจากข้อตกลงเบื้องต้นเกี่ยวกับการแจกแจงของคะแนนที่นำมาใช้ในการคำนวณความน่าจะเป็นในการจำแนกผู้สอบที่มีความสามารถ θ เข้าแต่ละระดับความสามารถตามที่กล่าวมาข้างต้น เนื่องด้วยความแตกต่างเหล่านี้ทำให้ผู้วิจัยเลือกที่จะศึกษาถึงประสิทธิภาพในการประมาณค่าดัชนีการจำแนกประเภทของวิธีการประมาณค่าตามทฤษฎีการตอบสนองข้อสอบทั้งสามวิธี

ด้านตัวแปรที่เป็นเงื่อนไขในการศึกษาประสิทธิภาพของการประมาณค่าดัชนีการจำแนกประเภท จากการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องพบว่าได้มีการศึกษาเกี่ยวกับตัวแปรที่น่าจะส่งผลต่อค่าดัชนีการจำแนกประเภทจำนวน 12 ตัวแปร คือ โมเดลการวัด ขนาดกลุ่มตัวอย่าง ตำแหน่งของคะแนนจุดตัด จำนวนจุดตัด คุณลักษณะทางจิตมิติของแบบสอบ คุณลักษณะทาง จิตมิติของข้อสอบ ความยาวของแบบสอบ จำนวนคุณลักษณะที่ต้องการจะวัด ความเป็นอิสระระหว่างคุณลักษณะที่ต้องการจะวัด รูปแบบการแจกแจงความสามารถของผู้สอบ ความสูงของมาตรฐานในการจำแนกประเภท สหสัมพันธ์ภายในข้อสอบ และความเป็นไปได้ของการชดเชย ซึ่งในงานวิจัยที่ผ่านมาได้ทำการศึกษาตัวแปรเหล่านี้ภายใต้เงื่อนไข สถานการณ์ และการใช้วิธีการประมาณค่าดัชนีการจำแนกประเภทที่แตกต่างกันไปดังรายละเอียดที่นำเสนอไปในตอนที่ 4 เกี่ยวกับปัจจัยที่ส่งผลต่อดัชนีการจำแนกประเภท โดยในการวิจัยครั้งนี้ผู้วิจัยเลือกตัวแปรที่เป็นเงื่อนไขในการศึกษาจำนวน 3 ตัวแปร คือ ความยาวของแบบสอบ โมเดลการวัด และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ ส่วนตัวแปรอื่นจะควบคุมให้อยู่ในสถานการณ์มาตรฐานทั่วไป เนื่องจากทั้งสามตัวแปรนี้เป็นปัจจัยที่มีความสำคัญต่อการประมาณค่าดัชนี ถึงแม้ว่าตัวแปรความยาวของแบบสอบนั้น พบว่ามีงานวิจัยของ Lathrop และ Cheng (2013) เพียงเรื่องเดียวที่ทำการศึกษาในเรื่องนี้ ซึ่งก็ยังไม่ได้ทำการศึกษาที่ครอบคลุมวิธีการประมาณค่าดัชนีการจำแนกประเภทครบทั้งสามวิธีที่ใช้ในการศึกษาครั้งนี้ และเนื่องจากในการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET) ซึ่งเป็นการทดสอบที่ได้มาตรฐาน มี

การกำหนดความยาวของแบบสอบที่ใช้ในแต่ละวิชาอยู่ในช่วง 25 ถึง 50 ข้อ ผู้วิจัยจึงนำจำนวนข้อที่น้อยที่สุดและมากที่สุดมากำหนดเป็นเงื่อนไขในการศึกษาจำลองครั้งนี้ โดยในการวิจัยครั้งนี้กำหนดความยาวของแบบสอบที่ใช้เป็นเงื่อนไขในการจำลองข้อมูลจำนวน 2 เงื่อนไข คือ แบบสอบที่มีความยาว 25 ข้อ และแบบสอบที่มีความยาว 50 ข้อ

สำหรับตัวแปรโมเดลการวัดนั้นจะมีงานวิจัยที่ทำการศึกษาไว้จำนวนหนึ่ง เช่นงานวิจัยของ Lathrop และ Cheng (2013), Lee (2010) และ Zhang (2008, 2010) แต่ก็ยังมีไม่มากนัก และงานวิจัยเหล่านี้ก็ยังไม่ได้ทำการศึกษาที่ครอบคลุมวิธีการประมาณค่าดัชนีการจำแนกประเภทครบทั้งสามวิธีที่ใช้ในการศึกษาครั้งนี้ และเนื่องจากโมเดลการวัดภายใต้โมเดลการตอบสนองข้อสอบแบบสองค่ามีทั้งสิ้น 3 โมเดล คือ โมเดลโลจิสติกแบบหนึ่ง สอง และสามพารามิเตอร์ โดยในการวิจัยครั้งนี้ได้ทำการกำหนดโมเดลการวัดที่ใช้เป็นเงื่อนไขในการจำลองข้อมูลจำนวน 3 โมเดล คือ โมเดลโลจิสติกแบบหนึ่งพารามิเตอร์ (one-parameter logistic model: 1PL) โมเดลโลจิสติกแบบสองพารามิเตอร์ (two-parameter logistic model: 2PL) และโมเดลโลจิสติกแบบสามพารามิเตอร์ (three-parameter logistic model: 3PL) ส่วนตัวแปรความไม่เหมาะสมของโมเดลการวัดกับข้อสอบนั้นพบว่ายังไม่มีงานวิจัยใดที่ทำการศึกษาเกี่ยวกับเรื่องนี้ แต่เนื่องจากในสถานการณ์การทดสอบทั่วไปนั้นมีโอกาสที่จะเกิดความไม่เหมาะสมของโมเดลการวัดกับข้อสอบได้สูง ดังนั้นจึงมีความจำเป็นที่ต้องพิจารณาถึงปัจจัยสำคัญนี้ด้วย และเนื่องจากในสถานการณ์การทดสอบที่ได้มาตรฐานนั้น ระดับความไม่เหมาะสมของโมเดลการวัดกับข้อสอบที่ยอมรับได้ไม่ควรเกินร้อยละ 20 ผู้วิจัยจึงนำระดับความไม่เหมาะสมของโมเดลการวัดกับข้อสอบมากำหนดเป็นเงื่อนไขในการศึกษาจำลองครั้งนี้ โดยในการวิจัยครั้งนี้กำหนดความไม่เหมาะสมของโมเดลการวัดกับข้อสอบที่ใช้เป็นเงื่อนไขในการจำลองข้อมูลจำนวน 2 เงื่อนไข คือ แบบสอบที่มีความไม่เหมาะสมของโมเดลการวัดกับข้อสอบร้อยละ 10 และแบบสอบที่มีความไม่เหมาะสมของโมเดลการวัดกับข้อสอบร้อยละ 20

ดังนั้น ผู้วิจัยจึงมีความสนใจที่จะศึกษาประสิทธิภาพของวิธีการประมาณตามทฤษฎีการตอบสนองข้อสอบ ซึ่งคำนวณได้จากการหาค่าเฉลี่ยของดัชนีการจำแนกประเภทจากการทำวนซ้ำ (replication) จำนวน 100 รอบ โดยมีกรอบแนวคิดในการวิจัยดังนี้



ภาพที่ 2.2 กรอบแนวคิดในการวิจัย

บทที่ 3
วิธีดำเนินการวิจัย

การศึกษาในครั้งนี้มีวัตถุประสงค์เพื่อประมาณค่าดัชนีการจำแนกประเภทโดยใช้วิธีการประมาณค่าตามทฤษฎีการตอบสนองข้อสอบสามวิธี คือวิธีการของ Rudner (2005) วิธีการของ Guo (2006) และวิธีการของ Lee (2010) จากการจำลองข้อมูลภายใต้เงื่อนไขของการศึกษา และเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภททั้งสามวิธีด้วยการพิจารณาจากค่าความคลาดเคลื่อนมาตรฐานของการประมาณค่า นอกจากนี้ยังได้ทำการประมาณค่าและเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบทั้งสามวิธี เมื่อนำไปใช้กับข้อมูลเชิงประจักษ์ ซึ่งผลที่ได้จากการวิจัยครั้งนี้จะเป็นประโยชน์เพื่อประกอบการตัดสินใจเลือกใช้วิธีการประมาณค่าดัชนีการจำแนกประเภท ทั้งดัชนีความสอดคล้องและดัชนีความถูกต้องที่เหมาะสมกับสถานการณ์การทดสอบจริงมากที่สุด

วิธีดำเนินการวิจัยเป็นแบบการศึกษาการจำลองข้อมูล (simulation study) โดยทำการจำลองข้อมูลด้วยโปรแกรม WinGen และทำการวิเคราะห์ข้อมูลภายใต้เงื่อนไขของการศึกษาด้วยโปรแกรม R เพื่อให้ได้ข้อสรุปเกี่ยวกับประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบทั้งสามวิธี ซึ่งทั้งสามวิธีมีความแตกต่างกันในเรื่องของข้อตกลงเบื้องต้นเกี่ยวกับลักษณะการแจกแจง ลักษณะของข้อมูลที่ใช้ในการประมาณค่า และความน่าจะเป็นที่คาดหวังในการจำแนกผู้สอบเข้าสู่แต่ละระดับความสามารถ จากนั้นนำวิธีการประมาณค่าที่มีค่าความคลาดเคลื่อนมาตรฐานของการประมาณค่าน้อยที่สุดไปใช้ในการประมาณค่าดัชนีการจำแนกประเภทกับข้อมูลจริง (real data) หรือข้อมูลเชิงประจักษ์ (empirical data) ต่อไป โดยดำเนินการวิจัยตาม 3 ขั้นตอนต่อไปนี้

ขั้นตอนที่ 1 การศึกษาผลการประมาณค่าดัชนีการจำแนกประเภทด้วยวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบ 3 วิธี คือ วิธีการของ Rudner (2005) วิธีการของ Guo (2006) และวิธีการของ Lee (2010)

ขั้นตอนที่ 2 การเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภททั้งสามวิธี

ขั้นตอนที่ 3 การศึกษาผลการประมาณค่าและเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบทั้งสามวิธี เมื่อใช้กับข้อมูลเชิงประจักษ์

ซึ่งสามารถสรุปขั้นตอนในการดำเนินการวิจัยได้ดังตารางที่ 3.1

ตารางที่ 3.1 ขั้นตอนในการดำเนินการวิจัย

วัตถุประสงค์	กิจกรรม	ผลลัพธ์
1. เพื่อประมาณค่าดัชนีการจำแนกประเภทโดยใช้วิธีการประมาณค่าตามทฤษฎีการตอบสนองข้อสอบสามวิธี คือ วิธีการของ Rudner (2005) วิธีการของ Guo (2006) และวิธีการของ Lee (2010) จากการจำลองข้อมูลภายใต้เงื่อนไขของการศึกษา	- การศึกษาจำลอง (simulation study) โดยทำการประมาณค่าดัชนีความสอดคล้องของการจำแนกประเภท ทั้งดัชนีความสอดคล้องของการจำแนกประเภทและดัชนีความถูกต้องของการจำแนกประเภท	- ค่าดัชนีความสอดคล้องของการจำแนกประเภท - ค่าดัชนีความถูกต้องของการจำแนกประเภท
2. เพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบสามวิธี คือ วิธีการของ Rudner (2005) วิธีการของ Guo (2006) และวิธีการของ Lee (2010) จากการจำลองข้อมูลภายใต้เงื่อนไขของการศึกษา	- การทำซ้ำ (replication) จำนวน 100 รอบ เพื่อทำการหาค่าเฉลี่ยดัชนีการจำแนกประเภท	- ค่าเฉลี่ยดัชนีการจำแนกประเภทของการประมาณค่า 100 รอบ จากสถานการณ์จำลอง - วิธีการประมาณค่าดัชนีการจำแนกประเภทที่มีประสิทธิภาพสูงสุดจากสถานการณ์จำลอง
3. เพื่อประมาณค่าและเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบทั้งสามวิธี เมื่อใช้กับข้อมูลเชิงประจักษ์	- การศึกษากับข้อมูลจริง (real data study) โดยทำการประมาณค่าดัชนีความสอดคล้องของการจำแนกประเภท ทั้งดัชนีความสอดคล้องของการจำแนกประเภทและดัชนีความถูกต้องของการจำแนกประเภท - การทำซ้ำ (replication) จำนวน 100 รอบ เพื่อทำการหาค่าเฉลี่ยดัชนีการจำแนกประเภท	- ค่าเฉลี่ยดัชนีการจำแนกประเภทของการประมาณค่า 100 รอบ จากสถานการณ์จริง - วิธีการประมาณค่าดัชนีการจำแนกประเภทที่มีประสิทธิภาพสูงสุดจากสถานการณ์จริง

ขั้นตอนที่ 1 การศึกษาผลการประมาณค่าดัชนีการจำแนกประเภทด้วยวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบ 3 วิธี คือ วิธีการของ Rudner (2005) วิธีการของ Guo (2006) และวิธีการของ Lee (2010)

การดำเนินการศึกษาในขั้นตอนนี้มีวัตถุประสงค์เพื่อประมาณค่าดัชนีการจำแนกประเภทโดยใช้วิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบสามวิธี คือ วิธีการของ Rudner (2005) วิธีการของ Guo (2006) และวิธีการของ Lee (2010) ด้วยการจำลองข้อมูลภายใต้เงื่อนไขที่แตกต่างกัน โดยมีรายละเอียดในการดำเนินการวิจัยดังนี้

1.1 เงื่อนไขในการจำลองข้อมูล

การจำลองข้อมูลในการศึกษาครั้งนี้อยู่ภายใต้เงื่อนไขของปัจจัยที่แตกต่างกัน 3 ปัจจัย คือ ความยาวของแบบสอบที่ใช้เป็นข้อมูลในการวิเคราะห์ จำนวน 2 เงื่อนไข โมเดลการวัดที่ใช้ในการประมาณค่าพารามิเตอร์ข้อสอบและความสามารถของผู้สอบ จำนวน 3 เงื่อนไข และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบที่ใช้ในการวิเคราะห์ จำนวน 2 เงื่อนไข จากปัจจัยดังกล่าวมีข้อมูลที่ศึกษาทั้งสิ้น 12 เงื่อนไข (2 เงื่อนไข × 3 เงื่อนไข × 2 เงื่อนไข) ส่วนตัวแปรอื่นจะควบคุมให้อยู่ในสถานการณ์มาตรฐานของการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET) โดยมีรายละเอียดของแต่ละปัจจัยดังนี้

1) ความยาวของแบบสอบ พบว่ามีงานวิจัยของ Lathrop และ Cheng (2013) เพียงเรื่องเดียวที่ทำการศึกษาในเรื่องนี้ ซึ่งก็ยังไม่ได้ทำการศึกษาที่ครอบคลุมวิธีการประมาณค่าดัชนีการจำแนกประเภทครบทั้งสามวิธีที่ใช้ในการศึกษาครั้งนี้ โดยในการวิจัยครั้งนี้ได้ทำการกำหนดความยาวของแบบสอบที่ใช้เป็นเงื่อนไขในการจำลองข้อมูลจำนวน 2 เงื่อนไข ได้แก่

2.1) แบบสอบที่มีความยาว 25 ข้อ

2.2) แบบสอบที่มีความยาว 50 ข้อ

2) โมเดลการวัด เนื่องจากตัวแปรโมเดลการวัดเป็นปัจจัยที่มีความสำคัญต่อการประมาณค่าดัชนี ถึงแม้ว่าตัวแปรโมเดลการวัดนั้นจะมีงานวิจัยที่ทำการศึกษาไว้จำนวนหนึ่ง เช่น งานวิจัยของ Lathrop และ Cheng (2013), Lee (2010) และ Zhang (2008, 2010) แต่งานวิจัยเหล่านี้ก็ยังไม่ได้ทำการศึกษาที่ครอบคลุมวิธีการประมาณค่าดัชนีการจำแนกประเภทครบทั้งสามวิธีที่ใช้ในการศึกษาครั้งนี้ โดยในการวิจัยครั้งนี้ได้ทำการกำหนดโมเดลการวัดที่ใช้เป็นเงื่อนไขในการจำลองข้อมูลจำนวน 3 เงื่อนไข ได้แก่

1.1) โมเดลโลจิสติกแบบหนึ่งพารามิเตอร์ (one-parameter logistic model: 1PL)

1.2) โมเดลโลจิสติกแบบสองพารามิเตอร์ (two-parameter logistic model: 2PL)

1.3) โมเดลโลจิสติกแบบสามพารามิเตอร์ (three-parameter logistic model: 3PL)

3) ความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ พบว่ายังไม่มีงานวิจัยใดที่ทำการศึกษาเกี่ยวกับเรื่องนี้ แต่เนื่องจากในสถานการณ์การทดสอบทั่วไปนั้นมีโอกาสที่จะเกิดความไม่เหมาะสมของโมเดลการวัดกับข้อสอบได้ ดังนั้นจึงมีความจำเป็นที่ต้องพิจารณาถึงปัจจัยสำคัญนี้ด้วย โดยในการวิจัยครั้งนี้ได้ทำการกำหนดความไม่เหมาะสมของโมเดลการวัดกับข้อสอบที่ใช้เป็นเงื่อนไขในการจำลองข้อมูลจำนวน 2 เงื่อนไข ได้แก่

3.1) แบบสอบที่มีความไม่เหมาะสมของโมเดลการวัดกับข้อสอบร้อยละ 10

3.2) แบบสอบที่มีความไม่เหมาะสมของโมเดลการวัดกับข้อสอบร้อยละ 20

1.2 เงื่อนไขในการวิเคราะห์ข้อมูล

ในการศึกษาครั้งนี้ได้ศึกษาวิธีการประมาณค่าดัชนีการจำแนกประเภท 3 วิธีการ ซึ่งเป็นวิธีการที่อยู่บนพื้นฐานของแนวคิดทฤษฎีการตอบสนองข้อสอบทั้งหมด และสามารถใช้ประมาณค่าได้ทั้งดัชนีความสอดคล้องของการจำแนกประเภทและดัชนีความถูกต้องของการจำแนกประเภท ได้แก่

1) วิธีการที่พัฒนาขึ้นโดย Rudner (2005) มีข้อตกลงเบื้องต้นเกี่ยวกับการแจกแจงปกติของความคลาดเคลื่อนมาตรฐานในการประมาณค่าคะแนนจริง เป็นวิธีการที่ใช้ได้กับทั้งข้อมูลที่มีลักษณะของข้อสอบที่มีการให้คะแนนได้สองค่า (dichotomous item) ข้อสอบที่มีการให้คะแนนได้มากกว่าสองค่า (polytomous item) และข้อสอบที่มีการให้คะแนนในรูปแบบคะแนนตามทฤษฎีการตอบสนองข้อสอบ (IRT pattern score) และใช้คะแนนความสามารถตามสเกลของ θ (test scored on theta scale) เป็นโมเดลในการประมาณค่าดัชนี

วิธีการของ Rudner มีสูตรในการหาค่าความน่าจะเป็นในการจำแนกผู้สอบที่มีความสามารถ θ เข้าสู่กลุ่มความสามารถ (Wyse & Hao, 2012) ดังนี้

$$\hat{p}_{iC} = \phi\left(\kappa_{C_i}, \kappa_{C_{i+1}}, \hat{\theta}_i, \hat{\sigma}_{\theta_i}\right)$$

เมื่อ $\hat{p}_{iC}$ คือ ความน่าจะเป็นในการจำแนกผู้สอบที่มีความสามารถ θ เข้าสู่กลุ่มความสามารถ

ϕ คือ พื้นที่ใต้โค้งปกติ

κ คือ คะแนนจุดตัด

c_i คือ ตำแหน่งของคะแนนจุดตัดที่ i

ϕ คือ พื้นที่ใต้โค้งปกติ

$\hat{\theta}_i$ คือ ความสามารถที่ประมาณค่าได้ของผู้สอบคนที่ i

$\hat{\sigma}_{\theta_i}$ คือ ความคลาดเคลื่อนในการประมาณค่าความสามารถของคนที่ i

สูตรในการคำนวณค่าดัชนีความสอดคล้องของการจำแนกประเภท (Wyse & Hao, 2012) คือ

$$\hat{\gamma} = \frac{\Sigma\left(\hat{\mathrm{P}} * \hat{\mathrm{P}}\right)}{\mathrm{N}_e}$$

เมื่อ $\hat{\gamma}$ คือ ดัชนีความสอดคล้องของการจำแนกประเภท

$\hat{P}$ คือ เมตริกซ์ $N_e \times C$ ของความน่าจะเป็นที่คาดหวัง

N_e คือ จำนวนผู้สอบ

สูตรในการคำนวณค่าดัชนีความถูกต้องของการจำแนกประเภท (Wyse & Hao, 2012) คือ

$$\hat{\tau} = \frac{\Sigma\left(\hat{\mathrm{P}} * W\right)}{\mathrm{N}_e}$$

เมื่อ $\hat{\tau}$ คือ ดัชนีความถูกต้องของการจำแนกประเภท

$\hat{P}$ คือ เมตริกซ์ $N_e \times C$ ของความน่าจะเป็นที่คาดหวัง

W คือ เมตริกซ์ $N_e \times W$ ของน้ำหนักซึ่งใช้ในการกำหนดกลุ่มระดับความสามารถที่ผู้สอบได้รับในการประเมิน

N_e คือ จำนวนผู้สอบ

2) วิธีการที่พัฒนาขึ้นโดย Guo (2006) มีข้อตกลงเบื้องต้นเกี่ยวกับการแจกแจงปกติของฟังก์ชันความน่าจะเป็น (likelihood functions) ในการตอบข้อสอบของผู้สอบ เป็นวิธีการที่ใช้ได้กับทั้งข้อมูลที่มีลักษณะของข้อสอบที่มีการให้คะแนนได้สองค่า (dichotomous item) ข้อสอบที่มีการให้คะแนนได้มากกว่าสองค่า (polytomous item) และข้อสอบที่มีการให้คะแนนในรูปแบบคะแนนตามทฤษฎีการตอบสนองข้อสอบ (IRT pattern score) และใช้ latent distribution เป็นโมเดลในการประมาณค่าดัชนี

วิธีการของ Guo มีสูตรในการหาค่าความน่าจะเป็นในการจำแนกผู้สอบที่มีความสามารถ θ เข้าสู่กลุ่มความสามารถ (Wyse & Hao, 2012) ดังนี้

$$\hat{p}_{ic} = \frac{\sum_{\theta=\kappa_{c_i}}^{\kappa_{c_{i+1}}} L(u_{1i}, u_{2i}, \ldots, u_{ni} \mid \theta)}{\sum_{h=1}^{C+1} \sum_{\theta=\kappa_h}^{\kappa_{h+1}} L(u_{1i}, u_{2i}, \ldots, u_{ni} \mid \theta)}$$

เมื่อ $\hat{p}_{iC}$ คือ ความน่าจะเป็นในการจำแนกผู้สอบที่มีความสามารถ θ เข้าสู่กลุ่มความสามารถ

$L(u_{1i}, u_{2i}, \ldots, u_{ni} \mid \theta)$ คือ ฟังก์ชั่นความน่าจะเป็นในการตอบข้อสอบของผู้สอบคนที่ i

κ คือ คะแนนจุดตัด

C_i คือ ตำแหน่งของคะแนนจุดตัดที่ i

ส่วนสูตรที่ใช้ในการคำนวณค่าดัชนีความสอดคล้องและความถูกต้องของการจำแนกประเภทนั้นใช้สูตรเดียวกับวิธีการของ Rudner ซึ่งพัฒนาขึ้นโดย Wyse และ Hao (2012)

3) วิธีการที่พัฒนาขึ้นโดย Lee (2010) มีข้อตกลงเบื้องต้นเกี่ยวกับการแจกแจงแบบทวินามเชิงซ้อน (compound binomial) ของคะแนนจริง เป็นวิธีการที่ใช้ได้กับทั้งข้อมูลที่มีลักษณะของข้อสอบที่มีการให้คะแนนได้สองค่า (dichotomous item) ข้อสอบที่มีการให้คะแนนได้มากกว่าสองค่า (polytomous item) และข้อสอบที่มีการให้คะแนนในรูปแบบคะแนนรวม (summed score) และใช้ mixture IRT model เป็นโมเดลในการประมาณค่าดัชนี

วิธีการของ Lee มีสูตรในการหาค่าความน่าจะเป็นในการจำแนกผู้สอบที่มีความสามารถ θ เข้าสู่กลุ่มความสามารถ (Wyse & Hao, 2012) ดังนี้

$$\hat{p}_{ic} = \sum_{x=\kappa_C}^{\kappa_{C+1}} fn(X = x \mid \hat{\theta})$$

เมื่อ $\hat{p}_{iC}$ คือ ความน่าจะเป็นในการจำแนกผู้สอบที่มีความสามารถ θ เข้าสู่กลุ่มความสามารถ

X คือ คะแนนสอบ

κ คือ คะแนนจุดตัด

C_i คือ ตำแหน่งของคะแนนจุดตัดที่ i

$\hat{\theta}$ คือ ความสามารถที่ประมาณค่าได้ของผู้สอบ

$\hat{\sigma}_{\theta_i}$ คือ ความคลาดเคลื่อนในการประมาณค่าความสามารถของคนที่ i

ส่วนสูตรที่ใช้ในการคำนวณค่าดัชนีความสอดคล้องและความถูกต้องของการจำแนกประเภทนั้นใช้สูตรเดียวกับวิธีการของ Rudner ซึ่งพัฒนาขึ้นโดย Wyse และ Hao (2012)

จากรายละเอียดของวิธีการประมาณค่าทั้งสามวิธีข้างต้นสามารถสรุปข้อตกลงเบื้องต้นและสูตรที่ใช้ในการประมาณค่าดัชนีการจำแนกประเภทของแต่ละวิธีได้ดังตารางต่อไปนี้

ตารางที่ 3.2 ข้อตกลงเบื้องต้นและสูตรที่ใช้ในการประมาณค่าดัชนีการจำแนกประเภทของวิธีการประมาณค่าตามทฤษฎีการตอบสนองข้อสอบ

รายการ	**วิธีการ**		
	Rudner	**Guo**	**Lee**
1. ลักษณะของคะแนนจุดตัด	คะแนนในสเกลของ θ	คะแนนในสเกลของ θ	คะแนนดิบ
2. ข้อตกลงเบื้องต้นเกี่ยวกับการแจกแจง	การแจกแจงแบบปกติของความคลาดเคลื่อนมาตรฐานในการประมาณค่าคะแนนจริง	การแจกแจงแบบปกติของฟังก์ชันความน่าจะเป็นในการตอบข้อสอบของผู้สอบ	การแจกแจงแบบทวินามของคะแนนจริง
3. สูตรความน่าจะเป็นในการจำแนกผู้สอบเข้าสู่แต่ละระดับความสามารถ	$\hat{p}_{iC}=\phi\left(\kappa_{C_i},\kappa_{C_{i+1}},\hat{\theta}_i,\hat{\sigma}_{\theta_i}\right)$	$\hat{p}_{ic}=\frac{\sum_{\theta=\kappa_{c_i}}^{\kappa_{c_{i+1}}} L(u_{1i},u_{2i},\ldots,u_{ni}\mid\theta)}{\sum_{h=1}^{C+1}\sum_{\theta=\kappa_h}^{\kappa_{h+1}} L(u_{1i},u_{2i},\ldots,u_{ni}\mid\theta)}$	$\hat{p}_{ic}=\sum_{x=\kappa_C}^{\kappa_{C+1}} f_n\left(X=x\mid\hat{\theta}\right)$
4. สูตรดัชนีความสอดคล้องของการจำแนกประเภท	$\hat{\gamma}=\frac{\sum\left(\hat{\mathbf{P}}*\hat{\mathbf{P}}\right)}{N_e}$	$\hat{\gamma}=\frac{\sum\left(\hat{\mathbf{P}}*\hat{\mathbf{P}}\right)}{N_e}$	$\hat{\gamma}=\frac{\sum\left(\hat{\mathbf{P}}*\hat{\mathbf{P}}\right)}{N_e}$
5. สูตรดัชนีความถูกต้องของการจำแนกประเภท	$\hat{\tau}=\frac{\sum\left(\hat{\mathbf{P}}*\mathbf{W}\right)}{N_e}$	$\hat{\tau}=\frac{\sum\left(\hat{\mathbf{P}}*\mathbf{W}\right)}{N_e}$	$\hat{\tau}=\frac{\sum\left(\hat{\mathbf{P}}*\mathbf{W}\right)}{N_e}$

1.3 การศึกษาการจำลองข้อมูล (Simulation Study)

การศึกษาในครั้งนี้มีเงื่อนไขที่ทำการศึกษาทั้งหมด 12 เงื่อนไข (3 เงื่อนไข × 2 เงื่อนไข × 2 เงื่อนไข) ซึ่งในการศึกษาการจำลองข้อมูลนี้มีการดำเนินการ 2 ขั้นตอน คือ ขั้นการเตรียมข้อมูลและขั้นดำเนินการประมาณค่าดัชนีการจำแนกประเภท โดยมีรายละเอียดของแต่ละขั้นตอนดังนี้

1) ขั้นเตรียมข้อมูล

ข้อมูลที่ต้องใช้ในการประมาณค่าดัชนีการจำแนกประเภทประกอบด้วย ข้อมูลการตอบข้อสอบของผู้สอบเพื่อนำไปใช้ในการประมาณค่าพารามิเตอร์ของข้อสอบและความสามารถของผู้สอบ และคะแนนจุดตัดเพื่อนำไปใช้ในการประมาณค่าดัชนีการจำแนกประเภท ซึ่งการเตรียมข้อมูลแต่ละส่วนนั้นมีรายละเอียดในการดำเนินการดังนี้

1.1) ข้อมูลการตอบข้อสอบ

ผู้วิจัยทำการจำลองข้อมูลชุดการตอบข้อสอบของผู้สอบด้วยโปรแกรม WinGen3 ซึ่งกำหนดให้มีรูปแบบการตอบข้อสอบเป็นการตอบข้อสอบแบบเลือกตอบ 4 ตัวเลือก 1 คำตอบ มีการให้คะแนนได้สองค่า (dichotomous item) ภายใต้เงื่อนไขความยาวของแบบสอบและเงื่อนไขของโมเดลการวัดตามที่กำหนดไว้ โดยในแต่ละชุดการตอบของแบบสอบที่มีความยาว 25 และ 50 ข้อนั้น กำหนดให้เป็นแบบสอบที่มีความไม่เหมาะสมของโมเดลการวัดกับข้อสอบร้อยละ 10 และร้อยละ 20 ตามลำดับ ภายใต้เงื่อนไขความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ มีขั้นตอนในการจำลองข้อมูลดังนี้

<u>ขั้นตอนที่ 1</u> จำลองข้อมูลของผู้สอบ (generating examinee data)

1) ระบุจำนวนผู้สอบ 2,000 คน ซึ่งในการจำลองรูปแบบการตอบใช้กลุ่มตัวอย่างจำนวน 2,000 คน เนื่องจากในงานวิจัยของ Wyse & Hao (2012) ได้ทำการตรวจสอบขั้นต้นเกี่ยวกับจำนวนกลุ่มตัวอย่างที่แตกต่างกันพบว่า การใช้กลุ่มตัวอย่างจำนวน 2,000 คน ให้ผลการประมาณค่าที่ใกล้เคียงกับการใช้กลุ่มตัวอย่างจำนวน 10,000 หรือ 25,000 คน

2) เลือกประเภทการแจกแจงของคะแนนเป็นการแจกแจงปกติ (normal distribution) และระบุค่า mean เป็น 0 และ SD เป็น 1

3) สั่งให้โปรแกรมจำลองข้อมูลของผู้สอบ จะได้ค่าพารามิเตอร์ความสามารถของผู้สอบ (theta) ตามจำนวนผู้สอบที่ระบุไว้

<u>ขั้นตอนที่ 2</u> จำลองข้อมูลของข้อสอบ (generating item data)

1) ระบุจำนวนข้อสอบที่ต้องการ ซึ่งในการจำลองครั้งนี้มีเงื่อนไขเกี่ยวกับความยาวของแบบสอบสองเงื่อนไข คือ แบบสอบยาว 25 ข้อ และ 50 ข้อตามลำดับ

2) ระบุจำนวนตัวเลือกเป็น 2 ค่า คือ 0=ตอบผิด และ 1=ตอบถูก

3) เลือกประเภทโมเดล IRT ซึ่งในการจำลองครั้งนี้มีเงื่อนไขเกี่ยวกับโมเดล IRT จำนวน 3 โมเดล คือ 1PL, 2PL และ 3PL และในการจำลองข้อมูลการตอบแต่ละครั้งนั้นมีการกำหนดความไม่สอดคล้องของคำตอบกับโมเดล IRT ที่ใช้ในการวิเคราะห์ด้วย ซึ่งสามารถดำเนินการได้ดังนี้

3.1) กำหนดประเภทการแจกแจงของค่าพารามิเตอร์ข้อสอบเพื่อให้ข้อมูลที่ได้มีลักษณะเป็นไปตามทฤษฎีที่ยอมรับได้ดังนี้ พารามิเตอร์อำนาจจำแนกเป็นการแจกแจงแบบ

lognormal โดยระบุค่า mean เป็น 0.2 และ SD เป็น 0.148 พารามิเตอร์ความยากเป็นการแจกแจงแบบ normal โดยระบุค่า mean เป็น 0 และ SD เป็น 1 และพารามิเตอร์โอกาสในการเดาข้อสอบถูกเป็นการแจกแจงแบบ beta โดยระบุค่า a เป็น 2 และ b เป็น 10 (อนุสรณ์ เกิดศรี, 2557)

3.2) สั่งให้โปรแกรมจำลองข้อมูลของข้อสอบ จะได้ค่าพารามิเตอร์ตามที่กำหนดไว้ในขั้นตอนของการเลือกประเภทโมเดล เช่น สถานการณ์เงื่อนไขของโมเดลการวัดแบบ 3PL ที่มีความไม่สอดคล้องของข้อสอบกับโมเดลร้อยละ 10 ของข้อสอบที่มีความยาว 50 ข้อ ในการจำลองข้อมูลจะต้องกำหนดข้อสอบที่สอดคล้องกับโมเดล 3PL จำนวน 45 ข้อ และข้อสอบที่ไม่สอดคล้องกับโมเดล 3PL จำนวน 5 ข้อ โดยในการวิจัยครั้งนี้กำหนดให้เป็นข้อสอบแบบ 2PL จำนวน 3 ข้อ และข้อสอบแบบ 1PL จำนวน 2 ข้อ ผลที่ได้คือค่าพารามิเตอร์ประจำข้อสอบแต่ละข้อตามโมเดลที่กำหนดไว้ข้างต้น

ขั้นตอนที่ 3 จำลองข้อมูลการตอบข้อสอบ (generating item response data)

1) ระบุ output file สำหรับจัดเก็บข้อมูลที่จำลองขึ้น

2) สั่งให้โปรแกรมจำลองข้อมูลการตอบข้อสอบ จะได้รูปแบบการตอบจำนวนข้อสอบและจำนวนผู้สอบตามที่กำหนดไว้ข้างต้น

1.2) คะแนนจุดตัด

ในการศึกษาการจำลองครั้งนี้ใช้คะแนนจุดตัดที่อ้างอิงจากคะแนนจุดตัดจริงของผลการประเมินระดับชาติขั้นพื้นฐานปีการศึกษา 2556 ที่จัดทดสอบโดยสถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน) ใน 2 รายวิชาหลัก คือ วิชาคณิตศาสตร์และภาษาไทย โดยมีขั้นตอนในการกำหนดคะแนนจุดตัดดังนี้ 1) กำหนดระดับคะแนนเป็น 8 ระดับ 2) กำหนดช่วงคะแนนในแต่ละระดับด้วยวิธี Normalized T-Score 3) กำหนดเกณฑ์คะแนนต่ำสุดระดับผ่านหรือระดับ 1 ที่ควรสูงกว่าคะแนนค่าของโอกาสการเดา เช่น แบบสอบปรนัยแบบ 4 ตัวเลือก คะแนนเต็ม 100 คะแนน เกณฑ์คะแนนต่ำสุดระดับผ่านควรสูงกว่า 25 คะแนน 4) กำหนดเกณฑ์คะแนนต่ำสุดที่ได้รับ 4 ควรมีคะแนนตั้งแต่ร้อยละ 80 และ 5) ช่วงคะแนนในแต่ละระดับแต่ละวิชาจะไม่กำหนดคงที่ โดยจะผันแปรไปตามการกระจายของคะแนนวิชานั้นๆ และระดับความยากง่ายของข้อสอบ (สถาบันทดสอบทางการศึกษาแห่งชาติ, 2556) ซึ่งมีรูปแบบของข้อสอบที่ใช้ในการทดสอบในแต่ละวิชามีไม่เกิน 2 รูปแบบ คือ รูปแบบปรนัย แบบเลือกตอบ 4 ตัวเลือก มีคำตอบที่ถูกที่สุด 1 คำตอบ และรูปแบบอื่นๆ เช่น รูปแบบปรนัยแบบเลือกตอบที่มีคำตอบถูกมากกว่า 1 คำตอบ แบบเลือกตอบจากแต่ละหมวดที่สัมพันธ์กัน และแบบระบายคำตอบเป็นค่าหรือตัวเลข เป็นต้น โดยรูปแบบข้อสอบสำหรับนักเรียนชั้นมัธยมศึกษาปีที่ 3 ที่ทำการทดสอบในปีการศึกษา 2556 แสดงได้ดังตารางที่ 3.3 และแนวการให้ระดับผลการประเมิน 8 ระดับ โดยมีการแสดงระดับเป็นตัวเลขและความหมายของแต่ละระดับดังตารางที่ 3.4

ตารางที่ 3.3 รูปแบบข้อสอบที่ใช้ในการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐานสำหรับนักเรียนชั้นมัธยมศึกษาปีที่ 3 ประจำปีการศึกษา 2556 จำแนกตามวิชา

ที่	รูปแบบ	วิชา			
		คณิตศาสตร์		ภาษาไทย	
		จำนวนข้อ	คะแนน	จำนวนข้อ	คะแนน
1	ปรนัย				
	1.1) 4 ตัวเลือก 1 คำตอบ	*25*	*80*	*50*	*80*
	1.2) 4 ตัวเลือก 2 คำตอบ	-	-	-	-
	1.3) 5 ตัวเลือก 1 คำตอบ	-	-	-	-
	1.4) 5 ตัวเลือก 2 คำตอบ	-	-	-	-
2	ปรนัย หลายตัวเลือก 1 คำตอบ	-	-	-	-
3	ปรนัย หลายตัวเลือก มากกว่า 1 คำตอบ	-	-	2	20
4	แบบเลือกคำตอบจากแต่ละหมวดที่สัมพันธ์กัน	-	-	-	-
5	แบบระบายคำตอบที่เป็นค่า/ตัวเลข	5	20	-	-
	รวม	**30**	**100**	**52**	**100**
	จำนวนเวลาที่ใช้สอบ	**90 นาที**		**90 นาที**	

ที่มา: สถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน), (2556)

ตารางที่ 3.4 ระดับการประเมินผลการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET) สำหรับนักเรียนชั้นมัธยมศึกษาปีที่ 3

ความหมาย	ระดับ	ช่วงคะแนน	
		คณิตศาสตร์	ภาษาไทย
ดีเยี่ยม	4	80-100	80-100
ดีมาก	3.5	66.4-79.99	73.3-79.99
ดี	3	53.6-66.39	64.8-73.29
ค่อนข้างดี	2.5	44-53.59	55.7-64.79
ปานกลาง	2	36.8-43.99	42.2-55.69
พอใช้	1.5	31.2-36.79	34.5-45.19
ควรปรับปรุง	1	25.01-31.19	25.01-34.49
ควรปรับปรุงอย่างยิ่ง	0	0-25	0-25

ที่มา: สถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน), (2556)

เนื่องด้วยการวิจัยครั้งนี้เลือกใช้เพียงรูปแบบการตอบข้อสอบแบบปรนัยแบบเลือกตอบ 4 ตัวเลือก 1 คำตอบเท่านั้น ดังนั้นในแต่ละวิชาที่ผู้วิจัยใช้ในการวิจัยครั้งนี้จึงประกอบด้วยจำนวนข้อสอบและคะแนนสอบที่แตกต่างกันดังนี้ วิชาคณิตศาสตร์ประกอบด้วยข้อสอบปรนัยแบบเลือกตอบ 4 ตัวเลือก จำนวน 25 ข้อ คิดเป็น 25 คะแนน และวิชาภาษาไทยประกอบด้วยข้อสอบปรนัยแบบเลือกตอบ 4 ตัวเลือกจำนวน 50 ข้อ คิดเป็น 50 คะแนน ซึ่งคะแนนจุดตัดที่ทางสำนักทดสอบทางการศึกษาแห่งชาติได้กำหนดไว้นั้นอยู่ในฐานคะแนนเต็ม 100 คะแนน ดังนั้นผู้วิจัยจึงทำการปรับเทียบคะแนนจุดตัดจากฐานคะแนนเต็ม 100 คะแนน ให้อยู่ในฐานของคะแนนเต็ม 25 คะแนนสำหรับวิชาคณิตศาสตร์ และคะแนนเต็ม 50 คะแนนสำหรับวิชาภาษาไทย โดยใช้วิธีการปรับเทียบคะแนนเชิงเส้นตรง (Linear Equating) (ศิริชัย กาญจนวาสี, 2555) ผลการปรับเทียบคะแนนเชิงเส้นตรงแสดงได้ดังตารางที่ 3.5-3.8

นอกจากนี้ในการดำเนินการประมาณค่าดัชนีการจำแนกประเภทนั้นต้องใช้ทั้งข้อมูลที่เป็นคะแนนสอบหรือคะแนนดิบ และคะแนนที่อยู่บนสเกลของ θ ดังนั้นจึงต้องมีการปรับเทียบคะแนนจุดตัดที่เป็นคะแนนดิบให้อยู่ในสเกลของ θ โดยใช้วิธีโค้งลักษณะแบบสอบ (Test Characteristic Curve: TCC) (Lathrop & Cheng, 2013) ทำให้ได้คะแนนจุดตัดที่ปรับเทียบแล้วของแต่ละรายวิชาดังตารางที่ 3.5-3.8

ตารางที่ 3.5 คะแนนจุดตัดวิชาคณิตศาสตร์ ชุด A ที่ปรับเทียบตามสเกลของ θ

ความหมาย	ระดับ	ช่วงคะแนน	คะแนนจุดตัด				
			เต็ม 100	เต็ม 25	TCC: θ		
					1PL	2PL	3PL
ดีเยี่ยม	4	80-100	80	21.96	6.9	7	3.7
ดีมาก	3.5	66.4-79.99	66.4	18.31	5.1	5.3	2.6
ดี	3	53.6-66.39	53.6	14.87	3.5	3.5	2.1
ค่อนข้างดี	2.5	44-53.59	44	12.29	2.1	2.1	1.7
ปานกลาง	2	36.8-43.99	36.8	10.35	1.4	1.3	1.4
พอใช้	1.5	31.2-36.79	31.2	8.85	0.7	0.7	1
ควรปรับปรุง	1	25.01-31.19	25.01	7.18	0	0	0.4
ควรปรับปรุงอย่างยิ่ง	0	0-25					

ตารางที่ 3.6 คะแนนจุดตัดวิชาคณิตศาสตร์ ชุด B ที่ปรับเทียบตามสเกลของ θ

ความหมาย	ระดับ	ช่วงคะแนน	คะแนนจุดตัด				
			เต็ม 100	เต็ม 25	TCC: θ		
					1PL	2PL	3PL
ดีเยี่ยม	4	80-100	80	21.82	7.1	7.3	4
ดีมาก	3.5	66.4-79.99	66.4	18.20	5.2	5.3	2.8
ดี	3	53.6-66.39	53.6	14.80	3.4	3.5	2.2
ค่อนข้างดี	2.5	44-53.59	44	12.25	2.3	2.2	1.8
ปานกลาง	2	36.8-43.99	36.8	10.33	1.4	1.4	1.4
พอใช้	1.5	31.2-36.79	31.2	8.84	0.8	0.8	1
ควรปรับปรุง	1	25.01-31.19	25.01	7.19	0	0	0.4
ควรปรับปรุงอย่างยิ่ง	0	0-25					

ตารางที่ 3.7 คะแนนจุดตัดวิชาภาษาไทย ชุด A ที่ปรับเทียบตามสเกลของ θ

ความหมาย	ระดับ	ช่วงคะแนน	คะแนนจุดตัด				
			เต็ม 100	เต็ม 50	TCC: θ		
					1PL	2PL	3PL
ดีเยี่ยม	4	80-100	80	38.00	3.7	4	3.7
ดีมาก	3.5	73.3-79.99	73.3	34.74	2.9	3.2	3
ดี	3	64.8-73.29	64.8	30.61	2	2.4	1.8
ค่อนข้างดี	2.5	55.7-64.79	55.7	26.19	1.1	1.1	1
ปานกลาง	2	42.2-55.69	42.2	19.62	-0.2	-0.2	-0.1
พอใช้	1.5	34.5-45.19	34.5	15.88	-1	-0.9	-0.9
ควรปรับปรุง	1	25.01-34.49	25.01	11.27	-2.1	-1.9	-2.8
ควรปรับปรุงอย่างยิ่ง	0	0-25					

ตารางที่ 3.8 คะแนนจุดตัดวิชาภาษาไทย ชุด B ที่ปรับเทียบตามสเกลของ θ

ความหมาย	ระดับ	ช่วงคะแนน	คะแนนจุดตัด				
			เต็ม 100	เต็ม 50	TCC: θ		
					1PL	2PL	3PL
ดีเยี่ยม	4	80-100	80	37.80	3.7	4	3.7
ดีมาก	3.5	73.3-79.99	73.3	34.57	2.9	3.2	3
ดี	3	64.8-73.29	64.8	30.48	2	2.5	1.8
ค่อนข้างดี	2.5	55.7-64.79	55.7	26.09	1.1	1.1	1
ปานกลาง	2	42.2-55.69	42.2	19.59	-0.2	-0.2	-0.1
พอใช้	1.5	34.5-45.19	34.5	15.88	-1	-0.9	-0.9
ควรปรับปรุง	1	25.01-34.49	25.01	11.30	-2	-1.9	-2.5
ควรปรับปรุงอย่างยิ่ง	0	0-25					

2) ขั้นดำเนินการประมาณค่าดัชนีการจำแนกประเภท

การประมาณค่าดัชนีการจำแนกประเภทดำเนินการโดยใช้โปรแกรม R โดยมีรายละเอียดในการดำเนินการดังนี้

2.1) นำเข้าข้อมูลรูปแบบการตอบข้อสอบของผู้สอบเพื่อประมาณค่าพารามิเตอร์ข้อสอบและพารามิเตอร์ความสามารถของผู้สอบตามทฤษฎีการตอบสนองข้อสอบ (IRT) โดยใช้โมเดลเงื่อนไขที่กำหนดไว้ในการประมาณค่าพารามิเตอร์สำหรับรูปแบบการตอบข้อสอบแบบเลือกตอบ 4 ตัวเลือก 1 คำตอบ มีการให้คะแนนได้สองค่า (dichotomous item) กับชุดการตอบที่จำลองขึ้นในขั้นตอนแรก โดยใช้โมเดล 1PL, 2PL และ 3PL ซึ่งเป็นโมเดลเงื่อนไขที่กำหนดไว้ในการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบและข้อสอบ

2.2) ทำการประมาณค่าดัชนีความสอดคล้องของการจำแนกประเภทโดยใช้วิธีการประมาณค่า 3 วิธี คือ วิธีการของ Rudner (2005) วิธีการของ Guo (2006) และวิธีการของ Lee (2010) โดยทำการประมาณค่าดัชนีภายใต้เงื่อนไขของจำนวนคะแนนจุดตัดเดียวกัน คือ จำนวน 7 ตำแหน่งตามจำนวนคะแนนจุดตัดที่สำนักทดสอบทางการศึกษาแห่งชาติกำหนดไว้สำหรับเป็นเกณฑ์ในการตัดสินผลการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET) และเนื่องด้วยข้อมูลของคะแนนที่จำลองขึ้นมีลักษณะเป็นคะแนนดิบ (raw score scale) ซึ่งสอดคล้องกับข้อตกลงเบื้องต้นของวิธีการของ Lee ที่ยอมรับข้อมูลที่อยู่ในสเกลของคะแนนดิบได้ จึงสามารถใช้ข้อมูลนี้ในการวิเคราะห์ได้เลย ส่วนวิธีการของ Rudner และ Guo นั้น ยอมรับข้อมูลที่อยู่ในสเกล theta (theta scale) จึงต้องมีการแปลงสเกลของคะแนนจุดตัดเป็น theta scale ให้เรียบร้อยก่อนนำไปประมาณค่าดัชนีการจำแนกประเภท รายละเอียดในการประมาณค่าดัชนีความสอดคล้องของการจำแนกประเภทมีดังนี้

2.2.1) นำรูปแบบการตอบข้อสอบมาประมาณค่าพารามิเตอร์ความสามารถของผู้สอบและทำการจำแนกผู้สอบเข้าสู่กลุ่มระดับสมรรถภาพจากคะแนนที่ได้ตามเงื่อนไขของคะแนนจุดตัดที่กำหนดไว้ สังเกตได้ว่าคะแนนที่ใช้ในขั้นตอนนี้คือคะแนนที่สังเกตได้ และผลลัพธ์ที่ได้จากขั้นตอนนี้คือเมตริกซ์ $\hat{P}$ หรือเมตริกซ์ของความน่าจะเป็นในการจำแนกผู้สอบที่มีความสามารถ θ เข้าสู่กลุ่มความสามารถ โดยเมตริกซ์ $\hat{P}$ มีลักษณะดังนี้

$$\hat{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1C} \\ p_{21} & p_{22} & \cdots & p_{2C} \\ \vdots & \vdots & \cdots & \vdots \\ p_{N_e1} & p_{N_e2} & \cdots & p_{N_eC} \end{bmatrix}$$

2.2.2) ทำการประมาณค่าดัชนีความสอดคล้องของการจำแนกประเภท (classification consistency index) โดยใช้วิธีการประมาณค่าความสอดคล้องของการจำแนกประเภททั้ง 3 วิธี ซึ่งคำนวณค่าดัชนีความสอดคล้องของการจำแนกประเภทได้ด้วยสูตรเดียวกันดังนี้

$$\hat{\gamma} = \frac{\Sigma\left(\hat{\mathrm{P}} * \hat{\mathrm{P}}\right)}{\mathrm{N}_e}$$

2.3) ทำการประมาณค่าดัชนีความถูกต้องของการจำแนกประเภทโดยใช้วิธีการประมาณค่า 3 วิธี คือ วิธีการของ Rudner (2005) วิธีการของ Guo (2006) และวิธีการของ Lee (2010) ซึ่งทำการประมาณค่าดัชนีภายใต้เงื่อนไขของจำนวนคะแนนจุดตัดเดียวกัน คือ จำนวน 7 ตำแหน่ง และดำเนินการแปลงสเกลของคะแนนตามที่ระบุไว้ในขั้นของการเตรียมข้อมูล นอกจากนี้จากนิยามของดัชนีความถูกต้องของการจำแนกที่นิยามว่าเป็นระดับของความเห็นพ้องต้องกันระหว่างการจำแนกโดยใช้ข้อมูลจริงหรือคะแนนที่สังเกตได้ (observed score) และการจำแนกโดยใช้คะแนนจริง (true score) โดยในความเป็นจริงนั้นจะไม่สามารถทราบค่าของคะแนนจริงได้ แต่จะทราบได้จากการประมาณค่าพารามิเตอร์ของข้อสอบและผู้สอบจากข้อมูลจำลอง รายละเอียดในการประมาณค่าดัชนีความถูกต้องของการจำแนกประเภทมีดังนี้

2.3.1) นำรูปแบบการตอบข้อสอบมาประมาณค่าพารามิเตอร์ความสามารถของผู้สอบและทำการจำแนกผู้สอบเข้าสู่กลุ่มระดับสมรรถภาพจากคะแนนที่ได้ตามเงื่อนไขของคะแนนจุดตัดที่กำหนดไว้ สังเกตได้ว่าคะแนนที่ใช้ในขั้นตอนนี้คือคะแนนที่สังเกตได้ และผลลัพธ์ที่ได้จากขั้นตอนนี้คือเมตริกซ์ $\hat{P}$ หรือเมตริกซ์ของความน่าจะเป็นในการจำแนกผู้สอบที่มีความสามารถ θ เข้าสู่กลุ่มความสามารถ โดยเมตริกซ์ $\hat{P}$ มีลักษณะดังนี้

$$\hat{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1C} \\ p_{21} & p_{22} & \cdots & p_{2C} \\ \vdots & \vdots & \cdots & \vdots \\ p_{N_e 1} & p_{N_e 2} & \cdots & p_{N_e C} \end{bmatrix}$$

2.3.2) นำคะแนนจริงที่ได้จากการประมาณค่าพารามิเตอร์ของข้อสอบและผู้สอบจากข้อมูลจำลองมาทำการจำแนกผู้สอบเข้าสู่กลุ่มระดับสมรรถภาพตามเงื่อนไขของคะแนนจุดตัดที่กำหนดไว้ ผลลัพธ์ที่ได้จากขั้นตอนนี้คือเมตริกซ์ W ที่มีลักษณะดังนี้

$$W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1C} \\ w_{21} & w_{22} & \cdots & w_{2C} \\ \vdots & \vdots & \cdots & \vdots \\ w_{N_e 1} & w_{N_e 2} & \cdots & w_{N_e C} \end{bmatrix}$$

ในการเขียนเมตริกซ์จะกำหนดให้น้ำหนัก W_{Ci} มีค่าเท่ากับ 1 เมื่อคะแนนของผู้สอบถูกจำแนกเข้าสู่กลุ่มระดับความสามารถ C และมีค่าเท่ากับ 0 เมื่อเป็นไปในทางตรงกันข้าม โดยในขั้นตอนนี้การใช้วิธีการประมาณค่าความถูกต้องของการจำแนกประเภททั้ง 3 วิธีมีการหาค่าเมตริกซ์ W ได้ด้วยวิธีการเดียวกัน

2.3.3) ทำการประมาณค่าดัชนีความถูกต้องของการจำแนกประเภท (classification accuracy index) โดยใช้วิธีการประมาณค่าความถูกต้องของการจำแนกประเภททั้ง 3 วิธี ซึ่งคำนวณค่าดัชนีความสอดคล้องของการจำแนกประเภทได้ด้วยสูตรเดียวกันดังนี้

$$\hat{\tau} = \frac{\Sigma\left(\hat{\mathrm{P}} * W\right)}{\mathrm{N}_e}$$

2.4) ตรวจสอบประสิทธิภาพของวิธีการประมาณค่าที่ใช้กับข้อมูลจำลองโดยคำนวณได้จากการหาค่าเฉลี่ยของค่าดัชนีการจำแนกประเภท ทั้งดัชนีความสอดคล้อง (consistency) และดัชนีความถูกต้อง (accuracy) ของการจำแนกประเภทจากการทำวนซ้ำ (replication) จำนวน 100 รอบ โดยหากวิธีการใดมีค่าเฉลี่ยดัชนีการจำแนกประเภทสูงจะหมายความว่าวิธีการประมาณค่านั้นมีประสิทธิภาพสูงที่สุด

1.4 โปรแกรมที่ใช้ในการจำลองข้อมูลและวิเคราะห์ข้อมูล

การวิเคราะห์ข้อมูลในขั้นตอนนี้ ใช้โปรแกรมคอมพิวเตอร์ดังนี้

1) โปรแกรม WinGen3 นำมาใช้ในการจำลองรูปแบบการตอบข้อสอบของผู้สอบ ซึ่งโปรแกรมนี้รองรับกับโมเดลตามทฤษฎีการตอบสนองข้อสอบต่างๆ เป็นอย่างดีไม่ว่าจะเป็นโมเดลการตอบสนองข้อสอบแบบตรวจให้คะแนนได้สองค่า (dichotomous item response models) หรือโมเดลการตอบสนองข้อสอบแบบตรวจให้คะแนนได้มากกว่าสองค่า (polytomous item response models) และสามารถสร้างชุดของค่าพารามิเตอร์ข้อสอบ และชุดของพารามิเตอร์ความสามารถของผู้สอบ เพื่อสร้างข้อมูลการตอบข้อสอบตามการแจกแจงได้หลายชนิด นอกจากนี้ยังง่ายต่อการใช้และสะดวกในการเข้าถึงโปรแกรมอีกด้วย ขั้นตอนและคำสั่งที่ใช้ในการประมาณค่าดัชนีการจำแนกประเภทอธิบายไว้ในภาคผนวก ก

2) โปรแกรม Excel นำมาใช้ในการปรับเทียบคะแนนเชิงเส้นตรง (Linear Equating)

3) โปรแกรม IRTPRO นำมาใช้ในการประมาณค่าพารามิเตอร์ของข้อสอบ (item parameters) และพารามิเตอร์ความสามารถของผู้สอบ (examinee ability parameters) ตามทฤษฎีการตอบสนองข้อสอบ (IRT) โดยใช้โมเดล 1PL, 2PL และ 3PL ในการประมาณค่าสำหรับรูปแบบการตอบข้อสอบแบบเลือกตอบ 4 ตัวเลือก 1 คำตอบ มีการให้คะแนนได้สองค่า (dichotomous item) เพื่อนำผลการวิเคราะห์ค่าพารามิเตอร์ของข้อสอบและค่าพารามิเตอร์ระดับความสามารถของผู้สอบมาใช้ในขั้นตอนการสร้างโค้งคุณลักษณะแบบสอบ (Test Characteristic Curve: TCC) เพื่อทำการปรับเทียบคะแนนจุดตัดจากสเกลของคะแนนดิบให้อยู่ในสเกลของ θ

4) โปรแกรม R นำมาใช้ในขั้นตอนการประมาณค่าดัชนีการจำแนกประเภททั้งดัชนีความสอดคล้องและดัชนีความถูกต้องของการจำแนก สำหรับทั้งสามวิธีการ คือ วิธีการของ Rudner (2005) วิธีการของ Guo (2006) และวิธีการของ Lee (2010) ใน การวิเคราะห์ด้วยโปรแกรม R นั้นมี Package ที่ใช้ในการวิเคราะห์และประมวลผลดังนี้

4.1) ltm ใช้ในการประมาณค่าพารามิเตอร์ข้อสอบและความสามารถของผู้สอบ

4.2) matrixStats ใช้ในการดำเนินการเกี่ยวกับเมตริกซ์

4.3) cacIRT (Lathrop, 2015) ใช้ในการประมาณค่าดัชนีการจำแนกประเภทด้วยวิธีการของ Rudner และวิธีการของ Lee ส่วนวิธีการของ Guo นั้นผู้วิจัยเป็นผู้เขียนคำสั่งที่ใช้ในการวิเคราะห์และประมวลผล ในการดำเนินการกับข้อมูลทั้ง 12 เงื่อนไขนั้นได้กำหนดให้มีการทำซ้ำ (replication) จำนวน 100 รอบสำหรับทั้งสามวิธีการ โดยขั้นตอนและคำสั่งที่ใช้ในการประมาณค่าดัชนีการจำแนกประเภทอธิบายไว้ในภาคผนวก ข

5) โปรแกรม SPSS นำมาใช้ในการวิเคราะห์ความแปรปรวนแบบสามทาง (Three-Way Analysis of Variance: 3-Way ANOVA)

ขั้นตอนที่ 2 การเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภททั้งสามวิธี

การดำเนินการศึกษาในขั้นตอนนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบสามวิธี คือ วิธีการของ Rudner (2005) วิธีการของ Guo (2006) และวิธีการของ Lee (2010) ภายใต้เงื่อนไขการจำลองข้อมูลในขั้นตอนที่ 1 มาทำการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภททั้งสามวิธี โดยประสิทธิภาพของการประมาณค่าดัชนีการจำแนกประเภทในการศึกษาครั้งนี้หาได้โดยการ

หาค่าเฉลี่ยของค่าดัชนีการจำแนกประเภท ทั้งดัชนีความสอดคล้อง (consistency) และดัชนีความถูกต้อง (accuracy) ของการจำแนกประเภทจากการทำวนซ้ำ (replication) จำนวน 100 รอบ หากวิธีการใดมีค่าเฉลี่ยดัชนีการจำแนกประเภทสูงจะหมายความว่าวิธีการประมาณค่านั้นมีประสิทธิภาพสูงที่สุด ซึ่งผู้วิจัยได้ทำการทดลองเกี่ยวกับจำนวนรอบเบื้องต้นพบว่า จำนวนการทำซ้ำ 100, 500 และ 1,000 รอบ ให้ผลการประมาณค่าที่ไม่แตกต่างกัน จึงเลือกใช้จำนวน 100 รอบ เพื่อลดระยะเวลาในการดำเนินการประมวลผลของโปรแกรม เมื่อทำซ้ำครบ 100 รอบแล้วนำค่าเฉลี่ยดัชนีการจำแนกประเภท ทั้งดัชนีความสอดคล้อง (consistency) และดัชนีความถูกต้อง (accuracy) ของการจำแนกประเภทที่ได้จากแต่ละวิธีการมาทำการทดสอบความแตกต่างของค่าเฉลี่ยโดยใช้การวิเคราะห์ความแปรปรวน (Analysis of Variance: ANOVA) ด้วยโปรแกรม SPSS เพื่อทำการเปรียบเทียบประสิทธิภาพของวิธีการต่างๆ ตามวัตถุประสงค์การวิจัยข้อที่ 2

ขั้นตอนที่ 3 การศึกษาผลการประมาณค่าและเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบทั้งสามวิธี เมื่อใช้กับข้อมูลเชิงประจักษ์

การดำเนินการศึกษาในขั้นตอนนี้มีวัตถุประสงค์เพื่อทำการประมาณค่าและเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบ เมื่อนำไปใช้กับข้อมูลเชิงประจักษ์ (empirical data) หรือข้อมูลการตอบข้อสอบจริง (real data) โดยมีรายละเอียดในการดำเนินการวิจัยดังนี้

3.1 ประชากรและกลุ่มตัวอย่าง

ประชากรที่ใช้ในการศึกษาครั้งนี้ คือ ชั้นมัธยมศึกษาปีที่ 3 ที่ทำการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET) ในปีการศึกษา 2556 จำนวน 2 รายวิชาหลัก ได้แก่ วิชาคณิตศาสตร์และภาษาไทย

กลุ่มตัวอย่างที่ใช้ในการศึกษาครั้งนี้คือ ชั้นมัธยมศึกษาปีที่ 3 ที่ทำการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET) ในปีการศึกษา 2556 ทั้ง 2 วิชา จำนวนรายวิชาละ 4,000 คน รวมเป็นกลุ่มตัวอย่างทั้งสิ้น 8,000 คน เนื่องจากในงานวิจัยของ Wyse และ Hao (2012) ได้ทำการตรวจสอบขั้นต้นเกี่ยวกับจำนวนกลุ่มตัวอย่างที่แตกต่างกันพบว่า การใช้กลุ่มตัวอย่างจำนวน 2,000 คน ให้ผลการประมาณค่าดัชนีการจำแนกประเภทที่ใกล้เคียงกับการใช้กลุ่มตัวอย่างจำนวน 10,000 หรือ 25,000 คน และเพื่อให้การประมาณค่ามีความลำเอียงเกิดขึ้นน้อยที่สุด ดังงานวิจัยของ Martineua (2007) ที่เสนอว่าความลำเอียงของการประมาณค่าจะมีค่าน้อยเมื่อใช้ตัวอย่างใน

การวิเคราะห์ข้อมูลอย่างน้อย 200 คน โดยใช้วิธีการสุ่มอย่างง่ายในการเลือกตัวอย่าง จำนวนประชากรและกลุ่มตัวอย่างจำแนกตามรายวิชาที่ใช้ในขั้นตอนนี้แสดงได้ดังตารางต่อไปนี้

ตารางที่ 3.9 จำนวนประชากรและกลุ่มตัวอย่างจำแนกตามรายวิชา

รายวิชา	ชุดข้อสอบ	ประชากร (คน)	กลุ่มตัวอย่าง (คน)
คณิตศาสตร์			
	A	340,084	2,000
	B	339,961	2,000
	รวม	**680,045**	**4,000**
ภาษาไทย			
	A	340,100	2,000
	B	340,552	2,000
	รวม	**680,652**	**4,000**

3.2 ข้อมูลที่ใช้ในการศึกษา

การศึกษาครั้งนี้ใช้ข้อมูลการตอบข้อสอบและคะแนนสอบของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ที่ทำการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน ในปีการศึกษา 2556 จำนวน 2 รายวิชาหลัก ได้แก่ วิชาคณิตศาสตร์และภาษาไทย และเกณฑ์การเทียบคะแนน Normalized T-score Norm จากผลการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET) โดยมีขั้นตอนในการกำหนดคะแนนจุดตัดดังนี้ 1) กำหนดระดับคะแนนเป็น 8 ระดับ 2) กำหนดช่วงคะแนนในแต่ละระดับด้วยวิธี Normalized T-Score 3) กำหนดเกณฑ์คะแนนต่ำสุดระดับผ่านหรือระดับ 1 ที่ควรสูงกว่าคะแนนค่าของโอกาสการเดา เช่น แบบสอบปรนัยแบบ 4 ตัวเลือก คะแนนเต็ม 100 คะแนน เกณฑ์คะแนนต่ำสุดระดับผ่านควรสูงกว่า 25 คะแนน 4) กำหนดเกณฑ์คะแนนต่ำสุดที่ได้รับ 4 ควรมีคะแนนตั้งแต่ร้อยละ 80 และ 5) ช่วงคะแนนในแต่ละระดับแต่ละวิชาจะไม่กำหนดคงที่ โดยจะผันแปรไปตามการกระจายของคะแนนวิชานั้นๆ และระดับความยากง่ายของข้อสอบ (สถาบันทดสอบทางการศึกษาแห่งชาติ, 2556) ซึ่งรายละเอียดเกี่ยวกับแบบสอบและคะแนนสอบได้กล่าวถึงไว้ในขั้นตอนที่ 1

3.3 เครื่องมือที่ใช้ในการวิจัย

เครื่องมือที่ใช้ในการศึกษาครั้งนี้เป็นแบบบันทึกคะแนนสอบและผลการตอบรายข้อของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ที่ทำการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET) ในปีการศึกษา 2556 จำนวน 2 รายวิชาหลัก ได้แก่ วิชาคณิตศาสตร์และภาษาไทย

3.4 การประมาณค่าดัชนีการจำแนกประเภท

การศึกษาในขั้นตอนนี้เป็นการนำวิธีการประมาณค่าที่มีประสิทธิภาพมากที่สุดภายใต้แต่ละเงื่อนไขที่กำหนดไว้ที่ได้จากการดำเนินการในขั้นตอนที่ 1 และ 2 มาทำการประมาณค่าดัชนีการจำแนกประเภท ทั้งดัชนีความสอดคล้องของการจำแนกและดัชนีความถูกต้องของการจำแนก โดยนำมาวิเคราะห์กับข้อมูลจากการทดสอบจริง มีรายละเอียดของของการศึกษาดังนี้

1) สุ่มรูปแบบการตอบข้อสอบของนักเรียนในแต่รายวิชาหลัก 2 วิชา คือวิชาคณิตศาสตร์และภาษาไทย จำนวนวิชาละ 4,000 คน ซึ่งมีรูปแบบการตอบข้อสอบเป็นแบบปรนัยเลือกตอบ 4 ตัวเลือก มีคำตอบที่ถูกที่สุด 1 คำตอบ โดยในแต่ละวิชาประกอบด้วยจำนวนข้อสอบและคะแนนสอบที่แตกต่างกันดังนี้ วิชาคณิตศาสตร์ประกอบด้วยข้อสอบปรนัยแบบเลือกตอบ 4 ตัวเลือก จำนวน 25 ข้อ ส่วนวิชาภาษาไทยประกอบด้วยข้อสอบปรนัยแบบเลือกตอบ 4 ตัวเลือกจำนวน 50 ข้อ โดยใช้โปรแกรม SPSS ในการสุ่มข้อมูล

2) ทำการตรวจสอบความสอดคล้องของโมเดลกับข้อมูลที่ใช้ในการประมาณค่า (model-data fit) โดยใช้โมเดลเงื่อนไขที่กำหนดไว้ในการประมาณค่าพารามิเตอร์ สำหรับรูปแบบการตอบข้อสอบแบบเลือกตอบ 4 ตัวเลือก 1 คำตอบ มีการให้คะแนนได้สองค่า (dichotomous item) ใช้โมเดล 1PL, 2PL และ 3PL ในการประมาณค่า พร้อมทั้งหาค่าพารามิเตอร์ความสามารถของผู้สอบและข้อสอบตามทฤษฎีการตอบสนองข้อสอบ (IRT) โดยใช้โปรแกรม IRTPRO ในการประมาณค่าพารามิเตอร์

3) ทำการประมาณค่าดัชนีการจำแนกประเภทดำเนินการโดยใช้โปรแกรม R โดยมีรายละเอียดในการดำเนินการดังนี้

3.1) นำเข้าข้อมูลรูปแบบการตอบข้อสอบของผู้สอบเพื่อประมาณค่าพารามิเตอร์ข้อสอบและพารามิเตอร์ความสามารถของผู้สอบตามทฤษฎีการตอบสนองข้อสอบ (IRT) โดยใช้โมเดลเงื่อนไขที่กำหนดไว้ในการประมาณค่าพารามิเตอร์สำหรับรูปแบบการตอบข้อสอบแบบเลือกตอบ 4 ตัวเลือก 1 คำตอบ มีการให้คะแนนได้สองค่า (dichotomous item) กับชุดการตอบข้อสอบจากแบบทดสอบทางการศึกษาระดับชาติ โดยใช้โมเดล 1PL, 2PL และ 3PL ซึ่งเป็นโมเดลเงื่อนไขที่กำหนดไว้ในการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบและข้อสอบ

3.2) ทำการประมาณค่าดัชนีความสอดคล้องของการจำแนกประเภทโดยใช้วิธีการประมาณค่า 3 วิธี คือ วิธีการของ Rudner (2005) วิธีการของ Guo (2006) และวิธีการของ Lee (2010) โดยทำการประมาณค่าดัชนีภายใต้เงื่อนไขของจำนวนคะแนนจุดตัดเดียวกัน คือ จำนวน 7 ตำแหน่งตามจำนวนคะแนนจุดตัดที่สำนักทดสอบทางการศึกษาแห่งชาติกำหนดไว้สำหรับเป็นเกณฑ์ในการตัดสินผลการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET) และเนื่องด้วยข้อมูลของคะแนนที่จำลองขึ้นมีลักษณะเป็นคะแนนดิบ (raw score scale) ซึ่งสอดคล้องกับข้อตกลง

เบื้องต้นของวิธีการของ Lee ที่ยอมรับข้อมูลที่อยู่ในสเกลของคะแนนดิบได้ จึงสามารถใช้ข้อมูลนี้ในการวิเคราะห์ได้เลย ส่วนวิธีการของ Rudner และ Guo นั้น ยอมรับข้อมูลที่อยู่ในสเกล theta (theta scale) จึงต้องมีการแปลงสเกลของคะแนนจุดตัดเป็น theta scale ให้เรียบร้อยก่อนนำไปประมาณค่าดัชนีการจำแนกประเภท รายละเอียดในการประมาณค่าดัชนีความสอดคล้องของการจำแนกประเภทมีดังนี้

3.2.1) นำรูปแบบการตอบข้อสอบมาประมาณค่าพารามิเตอร์ความสามารถของผู้สอบและทำการจำแนกผู้สอบเข้าสู่กลุ่มระดับสมรรถภาพจากคะแนนที่ได้ตามเงื่อนไขของคะแนนจุดตัดที่กำหนดไว้ สังเกตได้ว่าคะแนนที่ใช้ในขั้นตอนนี้คือคะแนนที่สังเกตได้ และผลลัพธ์ที่ได้จากขั้นตอนนี้คือเมตริกซ์ $\hat{P}$ หรือเมตริกซ์ของความน่าจะเป็นในการจำแนกผู้สอบที่มีความสามารถ θ เข้าสู่กลุ่มความสามารถ โดยเมตริกซ์ $\hat{P}$ มีลักษณะดังนี้

$$\hat{P}=\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1C} \\ p_{21} & p_{22} & \cdots & p_{2C} \\ \vdots & \vdots & \cdots & \vdots \\ p_{N_e1} & p_{N_e2} & \cdots & p_{N_eC} \end{bmatrix}$$

3.2.2) ทำการประมาณค่าดัชนีความสอดคล้องของการจำแนกประเภท (classification consistency index) โดยใช้วิธีการประมาณค่าความสอดคล้องของการจำแนกประเภททั้ง 3 วิธี ซึ่งคำนวณค่าดัชนีความสอดคล้องของการจำแนกประเภทได้ด้วยสูตรเดียวกันดังนี้

$$\hat{\gamma}=\frac{\Sigma\left(\hat{\mathrm{P}}*\hat{\mathrm{P}}\right)}{\mathrm{N}_e}$$

3.3) ทำการประมาณค่าดัชนีความถูกต้องของการจำแนกประเภทโดยใช้วิธีการประมาณค่า 3 วิธี คือ วิธีการของ Rudner (2005) วิธีการของ Guo (2006) และวิธีการของ Lee (2010) ซึ่งทำการประมาณค่าดัชนีภายใต้เงื่อนไขของจำนวนคะแนนจุดตัดเดียวกัน คือ จำนวน 7 ตำแหน่ง และดำเนินการแปลงสเกลของคะแนนตามที่ระบุไว้ในขั้นของการเตรียมข้อมูล นอกจากนี้จากนิยามของดัชนีความถูกต้องของการจำแนกที่นิยามว่าเป็นระดับของความเห็นพ้องต้องกันระหว่างการจำแนกโดยใช้ข้อมูลจริงหรือคะแนนที่สังเกตได้ (observed score) และการจำแนกโดยใช้คะแนนจริง (true score) โดยในความเป็นจริงนั้นจะไม่สามารถทราบค่าของคะแนนจริงได้ แต่จะทราบได้จากการประมาณค่าพารามิเตอร์ของข้อสอบและผู้สอบจากข้อมูลรูปแบบการตอบข้อสอบจาก

แบบทดสอบทางการศึกษาระดับชาติ รายละเอียดในการประมาณค่าดัชนีความถูกต้องของการจำแนกประเภทมีดังนี้

3.3.1) นำรูปแบบการตอบข้อสอบมาประมาณค่าพารามิเตอร์ความสามารถของผู้สอบและทำการจำแนกผู้สอบเข้าสู่กลุ่มระดับสมรรถภาพจากคะแนนที่ได้ตามเงื่อนไขของคะแนนจุดตัดที่กำหนดไว้ สังเกตได้ว่าคะแนนที่ใช้ในขั้นตอนนี้คือคะแนนที่สังเกตได้ และผลลัพธ์ที่ได้จากขั้นตอนนี้คือเมตริกซ์ $\hat{P}$ หรือเมตริกซ์ของความน่าจะเป็นในการจำแนกผู้สอบที่มีความสามารถ θ เข้าสู่กลุ่มความสามารถ โดยเมตริกซ์ $\hat{P}$ มีลักษณะดังนี้

$$\hat{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1C} \\ p_{21} & p_{22} & \cdots & p_{2C} \\ \vdots & \vdots & \cdots & \vdots \\ p_{N_e1} & p_{N_e2} & \cdots & p_{N_eC} \end{bmatrix}$$

3.3.2) นำคะแนนจริงที่ได้จากการประมาณค่าพารามิเตอร์ของข้อสอบและผู้สอบจากข้อมูลจำลองมาทำการจำแนกผู้สอบเข้าสู่กลุ่มระดับสมรรถภาพตามเงื่อนไขของคะแนนจุดตัดที่กำหนดไว้ ผลลัพธ์ที่ได้จากขั้นตอนนี้คือเมตริกซ์ W ที่มีลักษณะดังนี้

$$W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1C} \\ w_{21} & w_{22} & \cdots & w_{2C} \\ \vdots & \vdots & \cdots & \vdots \\ w_{N_e1} & w_{N_e2} & \cdots & w_{N_eC} \end{bmatrix}$$

ในการเขียนเมตริกซ์จะกำหนดให้น้ำหนัก W_{ci} มีค่าเท่ากับ 1 เมื่อคะแนนของผู้สอบถูกจำแนกเข้าสู่กลุ่มระดับความสามารถ C และมีค่าเท่ากับ 0 เมื่อเป็นไปในทางตรงกันข้าม โดยในขั้นตอนนี้การใช้วิธีการประมาณค่าความถูกต้องของการจำแนกประเภททั้ง 3 วิธีมีการหาค่าเมตริกซ์ W ได้ด้วยวิธีการเดียวกัน

3.3.3) ทำการประมาณค่าดัชนีความถูกต้องของการจำแนกประเภท (classification accuracy index) โดยใช้วิธีการประมาณค่าความถูกต้องของการจำแนกประเภททั้ง 3 วิธี ซึ่งคำนวณค่าดัชนีความสอดคล้องของการจำแนกประเภทได้ด้วยสูตรเดียวกันดังนี้

$$\hat{\tau} = \frac{\Sigma\left(\hat{\mathrm{P}} * W\right)}{\mathrm{N}_e}$$

3.4) ตรวจสอบประสิทธิภาพของวิธีการประมาณค่าที่ใช้กับข้อมูลจากการทดสอบทางการศึกษาระดับชาติ โดยคำนวณได้จากการหาค่าเฉลี่ยของค่าดัชนีการจำแนกประเภท ทั้งดัชนีความสอดคล้อง (consistency) และดัชนีความถูกต้อง (accuracy) ของการจำแนกประเภทจากการทำวนซ้ำ (replication) จำนวน 100 รอบ โดยหากวิธีการใดมีค่าเฉลี่ยดัชนีการจำแนกประเภทสูงจะหมายความว่าวิธีการประมาณค่านั้นมีประสิทธิภาพสูงที่สุด

4) ทำการจำลองข้อมูลจากรูปแบบการตอบข้อสอบจริงของนักเรียนสำหรับใช้ในการทำซ้ำ (replication) 100 รอบ เพื่อหาค่าเฉลี่ยดัชนีการจำแนกประเภททั้งดัชนีความถูกต้องและความสอดคล้องของการจำแนกประเภท การดำเนินการในขั้นตอนนี้เป็นการดำเนินการเพื่อพิสูจน์ว่าผลการประมาณค่าดัชนีที่ได้จากการใช้กับข้อมูลจริงจะตรงกับผลการประมาณค่าดัชนีที่ได้จากการศึกษาจำลองในขั้นตอนที่ 1 และ 2 หรือไม่ และมีลักษณะเป็นอย่างไร การประมาณค่าดัชนีการจำแนกประเภทในขั้นนี้ดำเนินการโดยใช้โปรแกรม R

5) นำค่าเฉลี่ยดัชนีการจำแนกประเภททั้งดัชนีความสอดคล้องและดัชนีความถูกต้องของการจำแนกประเภทที่ได้จากแต่ละวิธีการมาทำการทดสอบความแตกต่างของค่าเฉลี่ยโดยใช้การวิเคราะห์ความแปรปรวน (Analysis of Variance: ANOVA) ด้วยโปรแกรม SPSS เพื่อทำการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าตามทฤษฎีการตอบสนองข้อสอบทั้งสามวิธี

3.5 โปรแกรมที่ใช้ในการจำลองข้อมูลและวิเคราะห์ข้อมูล

การวิเคราะห์ข้อมูลในขั้นตอนนี้ ใช้โปรแกรมคอมพิวเตอร์ดังนี้

1) โปรแกรม SPSS ใช้ในการสุ่มตัวอย่างและการวิเคราะห์ค่าสถิติพื้นฐานของคะแนนสอบ พารามิเตอร์ความสามารถผู้สอบ ความคลาดเคลื่อนในการประมาณค่าความสามารถผู้สอบ และการวิเคราะห์ความแปรปรวน (Analysis of Variance: ANOVA)

2) โปรแกรม IRTPRO ใช้ในการประมาณค่าพารามิเตอร์ของข้อสอบ (item parameters) และพารามิเตอร์ความสามารถของผู้สอบ (examinee ability parameters) ตามทฤษฎีการตอบสนองข้อสอบ (IRT) โดยใช้โมเดล 1PL, 2PL และ 3PL ในการประมาณค่าสำหรับรูปแบบการตอบข้อสอบแบบเลือกตอบ 4 ตัวเลือก 1 คำตอบ มีการให้คะแนนได้สองค่า (dichotomous item) เพื่อนำผลการประมาณค่าพารามิเตอร์ของข้อสอบและค่าพารามิเตอร์ระดับความสามารถของผู้สอบไปใช้ในการประมาณค่าดัชนีการจำแนกประเภทต่อไป

3) โปรแกรม R ใช้ในขั้นตอนการวิเคราะห์หาค่าดัชนีการจำแนกประเภททั้งดัชนีความสอดคล้องและดัชนีความถูกต้องของการจำแนกประเภทกับทั้งสามวิธีการคือวิธีการของ Rudner (2005) วิธีการของ Guo (2006) และวิธีการของ Lee (2010) โดยในแต่ละวิธีนั้นได้กำหนดให้มีการทำซ้ำ (replication) จำนวน 100 รอบเพื่อนำมาหาประสิทธิภาพของวิธีการประมาณค่าต่อไป

บทที่ 4
ผลการวิเคราะห์ข้อมูล

ผลการวิเคราะห์ข้อมูลที่ได้จากการศึกษาครั้งนี้แบ่งการนำเสนอออกเป็น 3 ตอน โดยตอนแรกเป็นการนำเสนอผลการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบสามวิธีภายใต้การศึกษาการจำลองข้อมูล (simulation study) อันประกอบด้วยวิธีการของ Rudner (2005) วิธีการของ Guo (2006) และวิธีการของ Lee (2010) ตอนที่สองนำเสนอผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบทั้งสามวิธีภายใต้การศึกษาการจำลองข้อมูล (simulation study) และตอนสุดท้ายนำเสนอผลการประมาณค่าและผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบทั้งสามวิธี เมื่อนำไปใช้กับข้อมูลเชิงประจักษ์ (empirical data) หรือข้อมูลการตอบข้อสอบจริง (real data) ของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ในปีการศึกษา 2556 ซึ่งมีรายละเอียดในแต่ละตอนดังต่อไปนี้

ตอนที่ 1 ผลการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบสามวิธีภายใต้การศึกษาการจำลองข้อมูล (simulation study)

ตอนนี้เป็นการนำเสนอผลการวิเคราะห์ประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบสามวิธี ได้แก่ วิธีการของ Rudner (2005) วิธีการของ Guo (2006) และวิธีการของ Lee (2010) โดยการศึกษาการจำลองข้อมูล (simulation study) ภายใต้สถานการณ์เงื่อนไขด้วยโปรแกรม R ซึ่งประกอบด้วยสถานการณ์เงื่อนไขจำนวน 12 สถานการณ์ ดังนี้

สถานการณ์ที่ 1 ประกอบด้วยข้อสอบจำนวน 25 ข้อ ใช้โมเดลโลจิสติกแบบหนึ่งพารามิเตอร์ (one-parameter logistic model: 1PL) ในการประมาณค่าพารามิเตอร์ของข้อสอบและพารามิเตอร์ความสามารถของผู้สอบ และมีความไม่เหมาะสมของโมเดลการวัดกับข้อสอบร้อยละ 10 เขียนเป็นรหัสย่อได้ดังนี้ 251PL10

สถานการณ์ที่ 2 ประกอบด้วยข้อสอบจำนวน 25 ข้อ ใช้โมเดลโลจิสติกแบบหนึ่งพารามิเตอร์ (one-parameter logistic model: 1PL) ในการประมาณค่าพารามิเตอร์ของข้อสอบและพารามิเตอร์ความสามารถของผู้สอบ และมีความไม่เหมาะสมของโมเดลการวัดกับข้อสอบร้อยละ 20 เขียนเป็นรหัสย่อได้ดังนี้ 251PL20

สถานการณ์ที่ 3 ประกอบด้วยข้อสอบจำนวน 25 ข้อ ใช้โมเดลโลจิสติกแบบสองพารามิเตอร์ (two-parameter logistic model: 2PL) ในการประมาณค่าพารามิเตอร์ของข้อสอบและพารามิเตอร์

ความสามารถของผู้สอบ และมีความไม่เหมาะสมของโมเดลการวัดกับข้อสอบร้อยละ 10 เขียนเป็นรหัสย่อได้ดังนี้ 252PL10

สถานการณ์ที่ 4 ประกอบด้วยข้อสอบจำนวน 25 ข้อ ใช้โมเดลโลจิสติกแบบสองพารามิเตอร์ (two-parameter logistic model: 2PL) ในการประมาณค่าพารามิเตอร์ของข้อสอบและพารามิเตอร์ความสามารถของผู้สอบ และมีความไม่เหมาะสมของโมเดลการวัดกับข้อสอบร้อยละ 20 เขียนเป็นรหัสย่อได้ดังนี้ 252PL20

สถานการณ์ที่ 5 ประกอบด้วยข้อสอบจำนวน 25 ข้อ ใช้โมเดลโลจิสติกแบบสามพารามิเตอร์ (three-parameter logistic model: 3PL) ในการประมาณค่าพารามิเตอร์ของข้อสอบและพารามิเตอร์ความสามารถของผู้สอบ และมีความไม่เหมาะสมของโมเดลการวัดกับข้อสอบร้อยละ 10 เขียนเป็นรหัสย่อได้ดังนี้ 253PL10

สถานการณ์ที่ 6 ประกอบด้วยข้อสอบจำนวน 25 ข้อ ใช้โมเดลโลจิสติกแบบสามพารามิเตอร์ (three-parameter logistic model: 3PL) ในการประมาณค่าพารามิเตอร์ของข้อสอบและพารามิเตอร์ความสามารถของผู้สอบ และมีความไม่เหมาะสมของโมเดลการวัดกับข้อสอบร้อยละ 20 เขียนเป็นรหัสย่อได้ดังนี้ 253PL20

สถานการณ์ที่ 7 ประกอบด้วยข้อสอบจำนวน 50 ข้อ ใช้โมเดลโลจิสติกแบบหนึ่งพารามิเตอร์ (one-parameter logistic model: 1PL) ในการประมาณค่าพารามิเตอร์ของข้อสอบและพารามิเตอร์ความสามารถของผู้สอบ และมีความไม่เหมาะสมของโมเดลการวัดกับข้อสอบร้อยละ 10 เขียนเป็นรหัสย่อได้ดังนี้ 501PL10

สถานการณ์ที่ 8 ประกอบด้วยข้อสอบจำนวน 50 ข้อ ใช้โมเดลโลจิสติกแบบหนึ่งพารามิเตอร์ (one-parameter logistic model: 1PL) ในการประมาณค่าพารามิเตอร์ของข้อสอบและพารามิเตอร์ความสามารถของผู้สอบ และมีความไม่เหมาะสมของโมเดลการวัดกับข้อสอบร้อยละ 20 เขียนเป็นรหัสย่อได้ดังนี้ 501PL20

สถานการณ์ที่ 9 ประกอบด้วยข้อสอบจำนวน 50 ข้อ ใช้โมเดลโลจิสติกแบบสองพารามิเตอร์ (two-parameter logistic model: 2PL) ในการประมาณค่าพารามิเตอร์ของข้อสอบและพารามิเตอร์ความสามารถของผู้สอบ และมีความไม่เหมาะสมของโมเดลการวัดกับข้อสอบร้อยละ 10 เขียนเป็นรหัสย่อได้ดังนี้ 502PL10

สถานการณ์ที่ 10 ประกอบด้วยข้อสอบจำนวน 50 ข้อ ใช้โมเดลโลจิสติกแบบสองพารามิเตอร์ (two-parameter logistic model: 2PL) ในการประมาณค่าพารามิเตอร์ของข้อสอบและพารามิเตอร์ความสามารถของผู้สอบ และมีความไม่เหมาะสมของโมเดลการวัดกับข้อสอบร้อยละ 20 เขียนเป็นรหัสย่อได้ดังนี้ 502PL20

สถานการณ์ที่ 11 ประกอบด้วยข้อสอบจำนวน 50 ข้อ ใช้โมเดลโลจิสติกแบบสามพารามิเตอร์ (three-parameter logistic model: 3PL) ในการประมาณค่าพารามิเตอร์ของข้อสอบและพารามิเตอร์ความสามารถของผู้สอบ และมีความไม่เหมาะสมของโมเดลการวัดกับข้อสอบร้อยละ 10 เขียนเป็นรหัสย่อได้ดังนี้ 503PL10

สถานการณ์ที่ 12 ประกอบด้วยข้อสอบจำนวน 50 ข้อ ใช้โมเดลโลจิสติกแบบสามพารามิเตอร์ (three-parameter logistic model: 3PL) ในการประมาณค่าพารามิเตอร์ของข้อสอบและพารามิเตอร์ความสามารถของผู้สอบ และมีความไม่เหมาะสมของโมเดลการวัดกับข้อสอบร้อยละ 20 เขียนเป็นรหัสย่อได้ดังนี้ 503PL20

สถานการณ์เงื่อนไขทั้ง 12 สถานการณ์ข้างต้นสามารถสรุปได้ดังตารางต่อไปนี้

ตารางที่ 4.1 สถานการณ์เงื่อนไขในการจำลองข้อมูล

สถานการณ์ที่	รหัสย่อ	ความยาวของแบบสอบ		โมเดลการวัด			ความไม่เหมาะสมของโมเดลการตอบสนองข้อสอบกับข้อมูล	
		25 ข้อ	50 ข้อ	1PL	2PL	3PL	10%	20%
1	251PL10	✓		✓			✓	
2	251PL20	✓		✓				✓
3	252PL10	✓			✓		✓	
4	252PL20	✓			✓			✓
5	253PL10	✓				✓	✓	
6	253PL20	✓				✓		✓
7	501PL10		✓	✓			✓	
8	501PL20		✓	✓				✓
9	502PL10		✓		✓		✓	
10	502PL20		✓		✓			✓
11	503PL10		✓			✓	✓	
12	503PL20		✓			✓		✓

การนำเสนอผลการวิเคราะห์ในตอนนี้ประกอบด้วย ส่วนแรกเป็นผลการวิเคราะห์ข้อมูลเบื้องต้นของการทดสอบที่ได้จากการจำลองข้อมูล ได้แก่ ผลการวิเคราะห์ค่าสถิติพื้นฐานของค่าพารามิเตอร์ข้อสอบที่ได้จากการจำลองข้อมูล และผลการวิเคราะห์ค่าสถิติพื้นฐานของค่าพารามิเตอร์ความสามารถผู้สอบที่ได้จากการจำลองข้อมูล ส่วนที่สองเป็นผลการวิเคราะห์ค่าดัชนี

การจำแนกประเภทด้วยวิธีการประมาณค่าตามทฤษฎีการตอบสนองข้อสอบสามวิธีจากการจำลองข้อมูลภายใต้สถานการณ์เงื่อนไข และสุดท้ายเป็นการนำเสนอผลการทดสอบขนาดอิทธิพลของปัจจัยที่ส่งผลต่อค่าเฉลี่ยดัชนีการจำแนกประเภทของวิธีการประมาณค่าตามทฤษฎีการตอบสนองข้อสอบสามวิธีจากการศึกษาการจำลองข้อมูล โดยมีรายละเอียดของผลการวิเคราะห์ดังต่อไปนี้

1.1 ผลการวิเคราะห์ข้อมูลเบื้องต้นของการทดสอบที่ได้จากการจำลองข้อมูล

ส่วนนี้เป็นการนำเสนอผลการวิเคราะห์ค่าสถิติพื้นฐานของข้อมูลที่ได้จากการจำลองข้อมูลด้วยโปรแกรม WINGEN ซึ่งประกอบด้วยค่าพารามิเตอร์ของข้อสอบและค่าพารามิเตอร์ความสามารถของผู้สอบ ผลการวิเคราะห์มีรายละเอียดดังนี้

1) ผลการวิเคราะห์ค่าสถิติพื้นฐานของค่าพารามิเตอร์ข้อสอบที่ได้จากการจำลองข้อมูล

ส่วนนี้เป็นการนำเสนอผลการวิเคราะห์ค่าสถิติพื้นฐานของค่าพารามิเตอร์ข้อสอบที่ได้จากการจำลองข้อมูลด้วยโปรแกรม WINGEN ซึ่งประกอบด้วยค่าอำนาจจำแนก (a) ค่าความยาก (b) และโอกาสในการเดาข้อสอบถูก (c) โดยทำการจำลองข้อมูลพารามิเตอร์ข้อสอบขึ้นมาจำนวน 12 ชุดตามเงื่อนไขในการศึกษาที่กำหนดไว้ คือความยาวของแบบสอบ (25 และ 50 ข้อ) โมเดลการวัด (1PL, 2PL และ 3PL) และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ (ร้อยละ 10 และ ร้อยละ 20) ซึ่งผลการวิเคราะห์ในส่วนนี้ประกอบด้วยผลการจำลองข้อมูลพารามิเตอร์ข้อสอบตามสถานการณ์เงื่อนไขและผลการวิเคราะห์ค่าเฉลี่ยของพารามิเตอร์ข้อสอบที่ได้จากการจำลองข้อมูลแต่ละสถานการณ์ ผลการวิเคราะห์มีรายละเอียดดังนี้

ผลการจำลองข้อมูลของสถานการณ์ที่ 1 (251PL10) พบว่า ข้อสอบชุดนี้มีค่าอำนาจจำแนก (a) อยู่ในช่วง 0.969 ถึง 1.172 และมีค่าเฉลี่ยเท่ากับ 1.043 ค่าความยาก (b) อยู่ในช่วง -2.074 ถึง 1.684 และมีค่าเฉลี่ยเท่ากับ -0.095 และโอกาสในการเดาข้อสอบถูก (c) อยู่ในช่วง 0 ถึง 0.240 และมีค่าเฉลี่ยเท่ากับ 0.178 โดยมีข้อสอบที่ไม่สอดคล้องกับโมเดลจำนวน 3 ข้อ คือข้อ 23-25

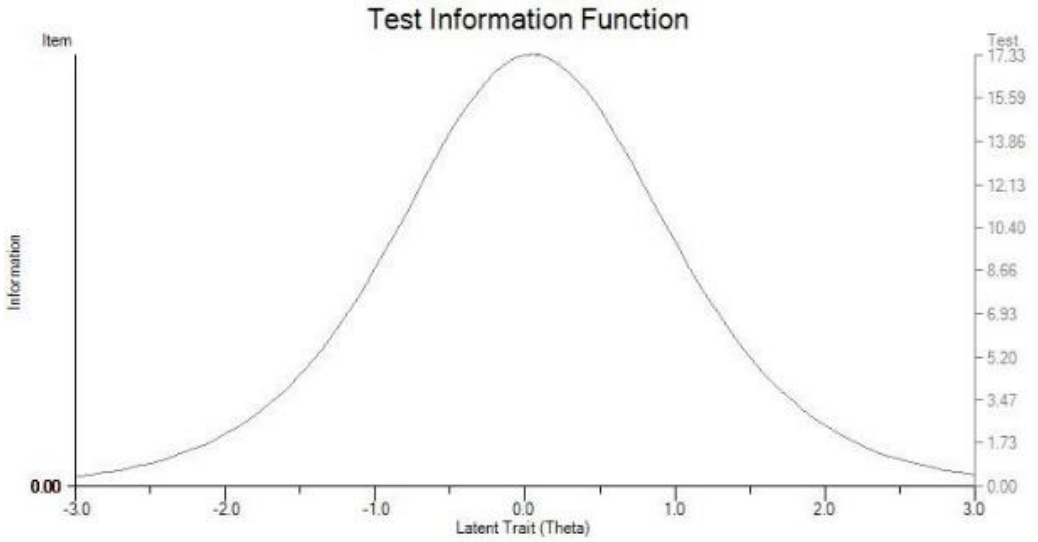
ผลการจำลองข้อมูลของสถานการณ์ที่ 2 (251PL20) พบว่า ข้อสอบชุดนี้มีค่าอำนาจจำแนก (a) อยู่ในช่วง 1.114 ถึง 1.554 และมีค่าเฉลี่ยเท่ากับ 1.298 ค่าความยาก (b) อยู่ในช่วง -2.168 ถึง 1.416 และมีค่าเฉลี่ยเท่ากับ -0.064 และโอกาสในการเดาข้อสอบถูก (c) อยู่ในช่วง 0 ถึง 0.404 และมีค่าเฉลี่ยเท่ากับ 0.264 โดยมีข้อสอบที่ไม่สอดคล้องกับโมเดลจำนวน 5 ข้อ คือข้อ 21-25

ผลการจำลองข้อมูลพารามิเตอร์ข้อสอบตามสถานการณ์เงื่อนไขและผลการวิเคราะห์ค่าเฉลี่ยของพารามิเตอร์ข้อสอบที่ได้จากการจำลองข้อมูลของสถานการณ์ที่ 1-2 แสดงได้ดังตารางต่อไปนี้

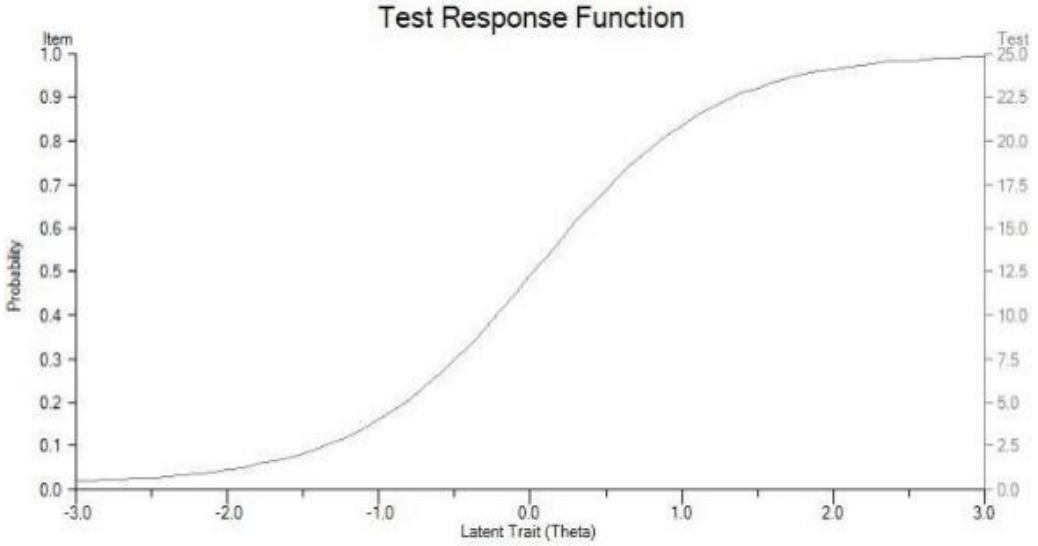
ตารางที่ 4.2 ค่าพารามิเตอร์ข้อสอบของข้อมูลจำลองสถานการณ์ที่ 1-2

ข้อสอบ	251PL10				251PL20			
	a	b	c	model	a	b	c	model
1	-	-2.074	0	1PL	-	-1.612	0	1PL
2	-	-1.283	0	1PL	-	0.871	0	1PL
3	-	-0.198	0	1PL	-	-0.323	0	1PL
4	-	1.684	0	1PL	-	-2.168	0	1PL
5	-	0.530	0	1PL	-	-1.038	0	1PL
6	-	-0.837	0	1PL	-	0.998	0	1PL
7	-	0.533	0	1PL	-	-0.186	0	1PL
8	-	-1.763	0	1PL	-	-0.686	0	1PL
9	-	-0.373	0	1PL	-	-0.503	0	1PL
10	-	0.424	0	1PL	-	-0.099	0	1PL
11	-	0.810	0	1PL	-	-1.822	0	1PL
12	-	-1.653	0	1PL	-	-0.722	0	1PL
13	-	0.402	0	1PL	-	1.179	0	1PL
14	-	-1.241	0	1PL	-	0.104	0	1PL
15	-	0.285	0	1PL	-	-0.422	0	1PL
16	-	-0.753	0	1PL	-	-0.320	0	1PL
17	-	1.311	0	1PL	-	1.416	0	1PL
18	-	0.882	0	1PL	-	-0.004	0	1PL
19	-	-1.273	0	1PL	-	1.057	0	1PL
20	-	-0.776	0	1PL	-	0.463	0	1PL
21	-	1.491	0	1PL	1.114	0.087	0.256	3PL*
22	-	0.409	0	1PL	1.162	0.413	0.404	3PL*
23	0.987	0.366	0.240	3PL*	1.554	1.085	0.133	3PL*
24	1.172	0.712	0.115	3PL*	1.469	0.931	0	2PL*
25	0.969	0.009	0	2PL*	1.192	-0.291	0	2PL*
ค่าเฉลี่ย	***1.043***	***-0.095***	***0.178***		***1.298***	***-0.064***	***0.264***	

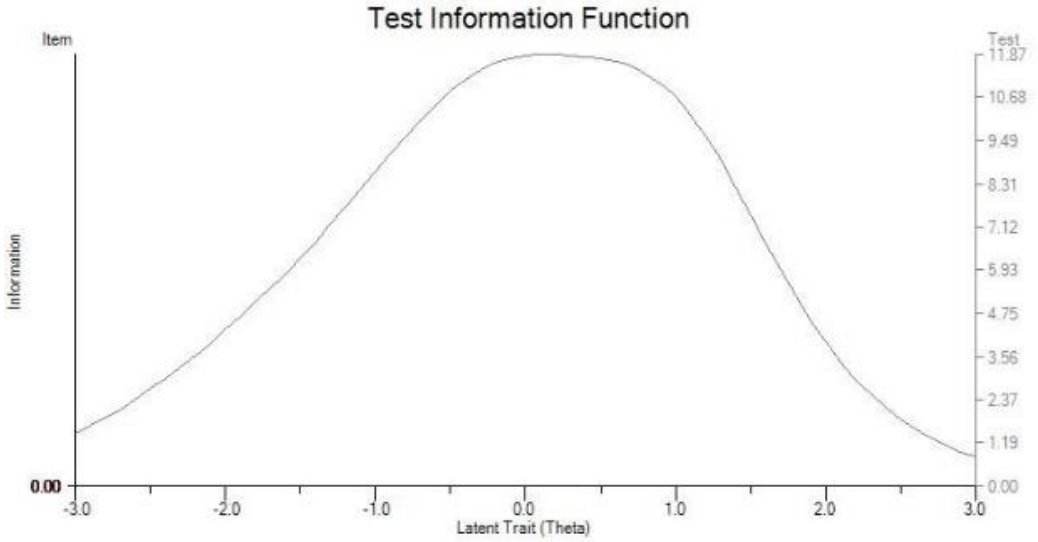
* ข้อสอบที่ไม่เหมาะสมกับโมเดลการวัด



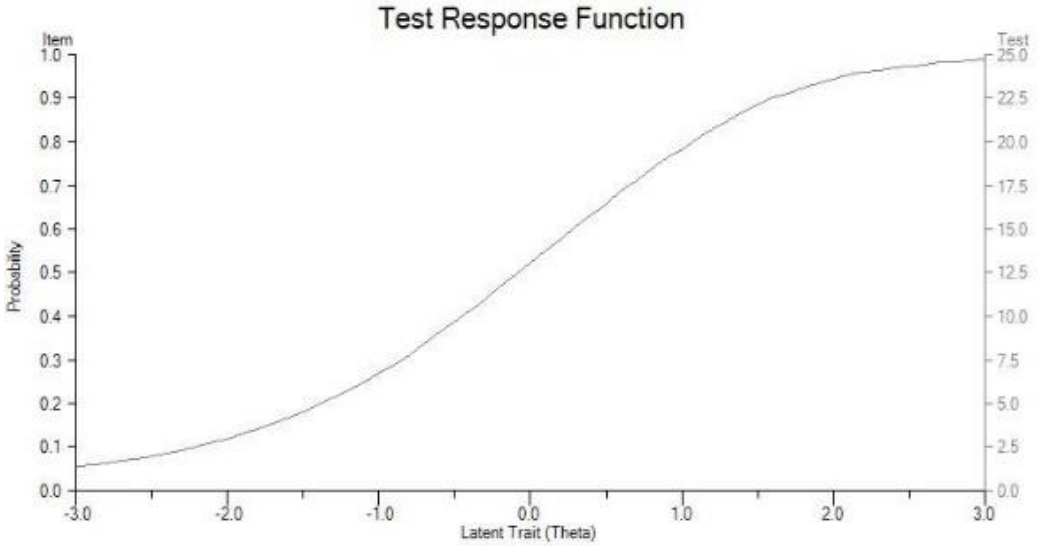
ภาพที่ 4.1 ฟังก์ชันสารสนเทศของแบบสอบของข้อมูลจำลองชุดที่ 1 (251PL10)



ภาพที่ 4.2 ฟังก์ชันการตอบสนองข้อสอบของข้อมูลจำลองชุดที่ 1 (251PL10)



ภาพที่ 4.3 ฟังก์ชันสารสนเทศของแบบสอบของข้อมูลจำลองชุดที่ 2 (251PL20)



ภาพที่ 4.4 ฟังก์ชันการตอบสนองข้อสอบของข้อมูลจำลองชุดที่ 2 (251PL20)

ผลการจำลองข้อมูลของสถานการณ์ที่ 3 (252PL10) พบว่า ข้อสอบชุดนี้มีค่าอำนาจจำแนก (a) อยู่ในช่วง 0.961 ถึง 1.727 และมีค่าเฉลี่ยเท่ากับ 1.259 ค่าความยาก (b) อยู่ในช่วง -2.794 ถึง 2.322 และมีค่าเฉลี่ยเท่ากับ -0.300 และโอกาสในการเดาข้อสอบถูก (c) อยู่ในช่วง 0 ถึง 0.326 และมีค่าเฉลี่ยเท่ากับ 0.246 โดยมีข้อสอบที่ไม่สอดคล้องกับโมเดลจำนวน 3 ข้อ คือข้อ 23-25

ผลการจำลองข้อมูลของสถานการณ์ที่ 4 (252PL20) พบว่า ข้อสอบชุดนี้มีค่าอำนาจจำแนก (a) อยู่ในช่วง 0.919 ถึง 1.559 และมีค่าเฉลี่ยเท่ากับ 1.264 ค่าความยาก (b) อยู่ในช่วง -1.818 ถึง 1.870 และมีค่าเฉลี่ยเท่ากับ 0.093 และโอกาสในการเดาข้อสอบถูก (c) อยู่ในช่วง 0 ถึง 0.150 และมีค่าเฉลี่ยเท่ากับ 0.118 โดยมีข้อสอบที่ไม่สอดคล้องกับโมเดลจำนวน 5 ข้อ คือข้อ 21-25

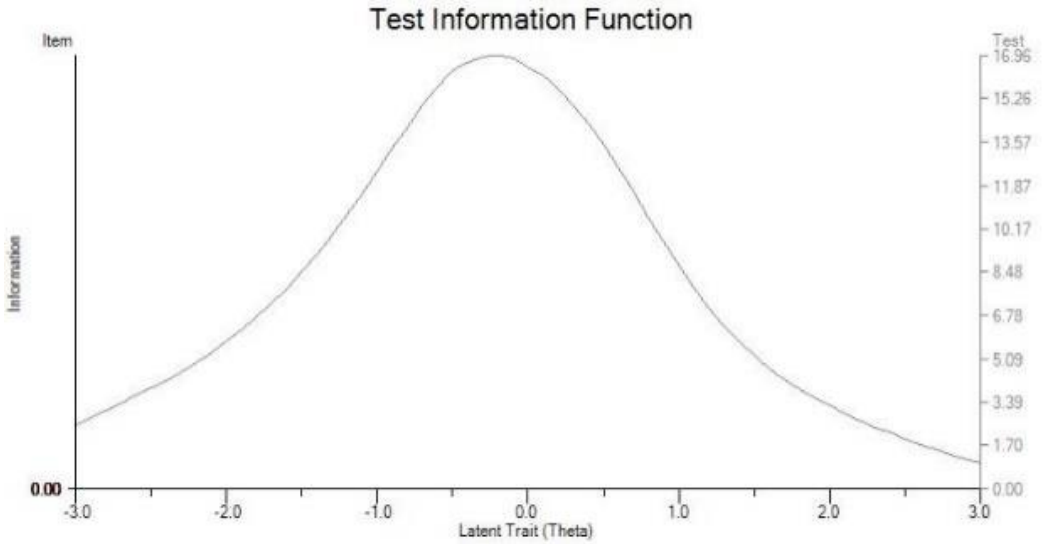
ผลการจำลองข้อมูลพารามิเตอร์ข้อสอบตามสถานการณ์เงื่อนไขและผลการวิเคราะห์ค่าเฉลี่ยของพารามิเตอร์ข้อสอบที่ได้จากการจำลองข้อมูลของสถานการณ์ที่ 3-4 แสดงได้ดังตารางต่อไปนี้

ตารางที่ 4.3 ค่าพารามิเตอร์ข้อสอบของข้อมูลจำลองสถานการณ์ที่ 3-4

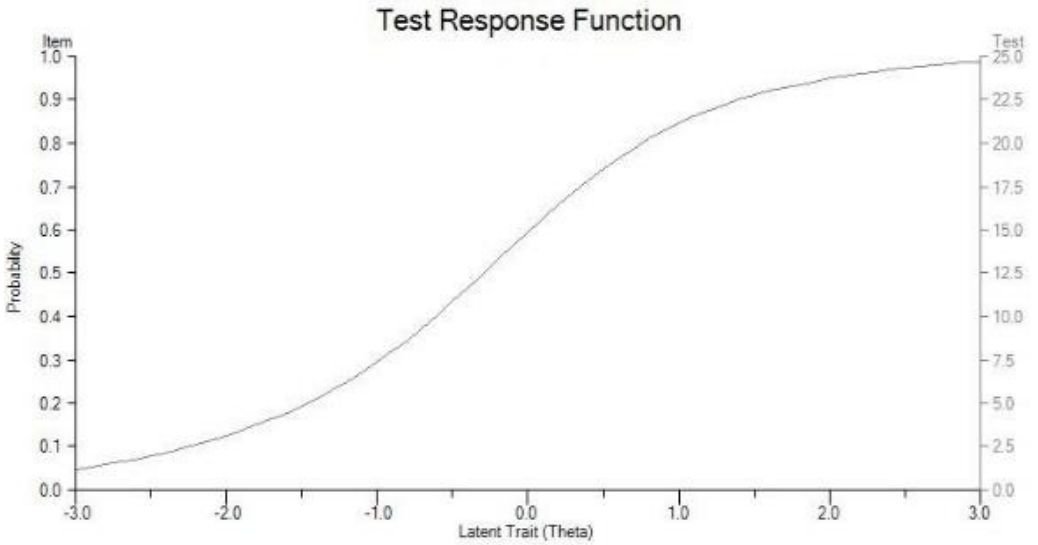
ข้อสอบ	252PL10				252PL20			
	a	b	c	model	a	b	c	model
1	1.138	-0.195	0	2PL	1.271	-0.238	0	2PL
2	1.429	0.587	0	2PL	1.214	0.219	0	2PL
3	1.144	-0.075	0	2PL	1.476	-1.093	0	2PL
4	1.452	-0.551	0	2PL	1.559	0.930	0	1PL
5	1.402	0.353	0	2PL	0.950	0.521	0	1PL
6	1.164	0.180	0	2PL	1.492	0.122	0	2PL
7	1.288	-1.993	0	2PL	1.327	-0.842	0	2PL
8	0.978	-0.595	0	2PL	1.462	1.678	0	2PL
9	1.186	2.322	0	2PL	1.183	-0.771	0	2PL
10	1.647	-0.549	0	2PL	1.404	0.095	0	2PL
11	1.287	0.518	0	2PL	1.040	-0.694	0	2PL
12	1.234	-1.057	0	2PL	1.381	1.746	0	2PL
13	1.035	-0.054	0	2PL	1.335	1.033	0	2PL
14	1.267	-1.392	0	2PL	1.273	-1.644	0	2PL
15	1.050	-1.779	0	2PL	1.194	-1.818	0	2PL
16	1.240	-1.068	0	2PL	1.503	0.974	0	2PL
17	1.727	-0.291	0	2PL	1.034	-0.135	0	2PL

ข้อสอบ	252PL10				252PL20			
	a	b	c	model	a	b	c	model
18	1.146	-0.609	0	2PL	1.348	-0.657	0	2PL
19	1.362	-2.794	0	2PL	1.271	1.870	0	2PL
20	1.095	0.003	0	2PL	0.930	-0.314	0	2PL
21	1.348	1.577	0	2PL	0.919	0.043	0.063	3PL*
22	0.961	0.059	0	2PL	1.510	0.510	0.142	3PL*
23	1.444	0.411	0.165	3PL*	1.002	0.270	0.150	3PL*
24	1.203	0.079	0.326	3PL*	-	0.696	0	1PL*
25	-	-0.575	0	1PL*	-	-0.187	0	1PL*
ค่าเฉลี่ย	***1.259***	***-0.300***	***0.246***		***1.264***	***0.093***	***0.118***	

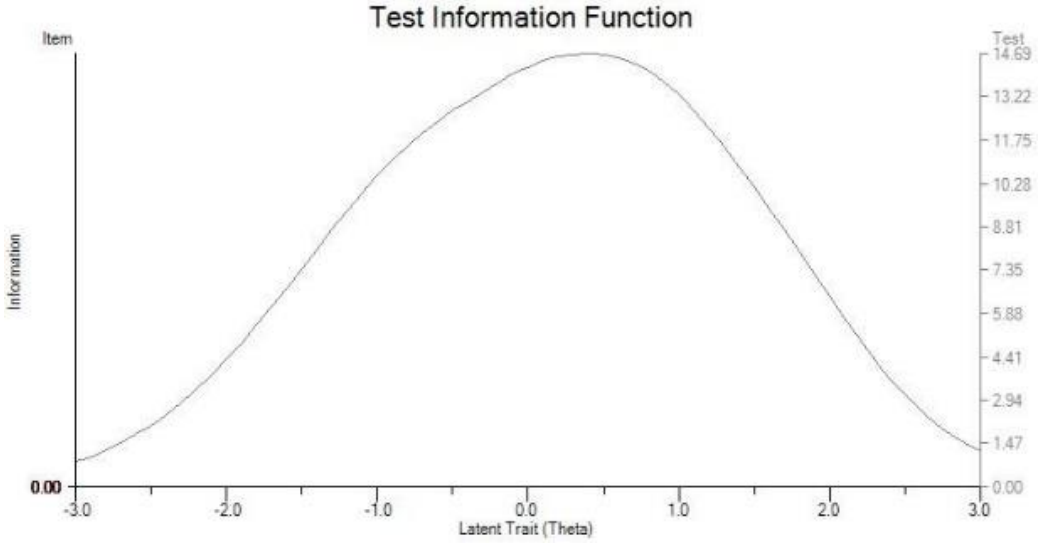
* ข้อสอบที่ไม่เหมาะสมกับโมเดลการวัด



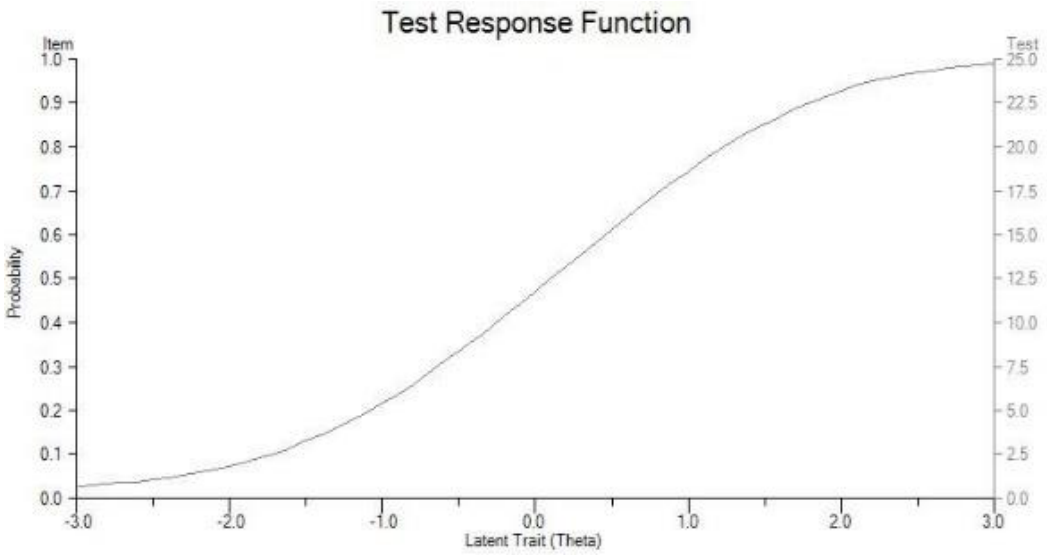
ภาพที่ 4.5 ฟังก์ชันสารสนเทศของแบบสอบของข้อมูลจำลองชุดที่ 3 (252PL10)



ภาพที่ 4.6 ฟังก์ชันการตอบสนองข้อสอบของข้อมูลจำลองชุดที่ 3 (252PL10)



ภาพที่ 4.7 ฟังก์ชันสารสนเทศของแบบสอบของข้อมูลจำลองชุดที่ 4 (252PL20)



ภาพที่ 4.8 ฟังก์ชันการตอบสนองข้อสอบของข้อมูลจำลองชุดที่ 4 (252PL20)

ผลการจำลองข้อมูลของสถานการณ์ที่ 5 (253PL10) พบว่า ข้อสอบชุดนี้มีค่าอำนาจจำแนก (a) อยู่ในช่วง 1.475 ถึง 0.841 และมีค่าเฉลี่ยเท่ากับ 1.219 ค่าความยาก (b) อยู่ในช่วง -1.553 ถึง 1.998 และมีค่าเฉลี่ยเท่ากับ 0.194 และโอกาสในการเดาข้อสอบถูก (c) อยู่ในช่วง 0 ถึง 0.373 และมีค่าเฉลี่ยเท่ากับ 0.154 โดยมีข้อสอบที่ไม่สอดคล้องกับโมเดลจำนวน 3 ข้อ คือข้อ 23-25

ผลการจำลองข้อมูลของสถานการณ์ที่ 6 (253PL20) พบว่า ข้อสอบชุดนี้มีค่าอำนาจจำแนก (a) อยู่ในช่วง 0.905 ถึง 1.760 และมีค่าเฉลี่ยเท่ากับ 1.249 ค่าความยาก (b) อยู่ในช่วง -1.716 ถึง 0.595 และมีค่าเฉลี่ยเท่ากับ -0.229 และโอกาสในการเดาข้อสอบถูก (c) อยู่ในช่วง 0 ถึง 0.473 และมีค่าเฉลี่ยเท่ากับ 0.166 โดยมีข้อสอบที่ไม่สอดคล้องกับโมเดลจำนวน 5 ข้อ คือข้อ 21-25

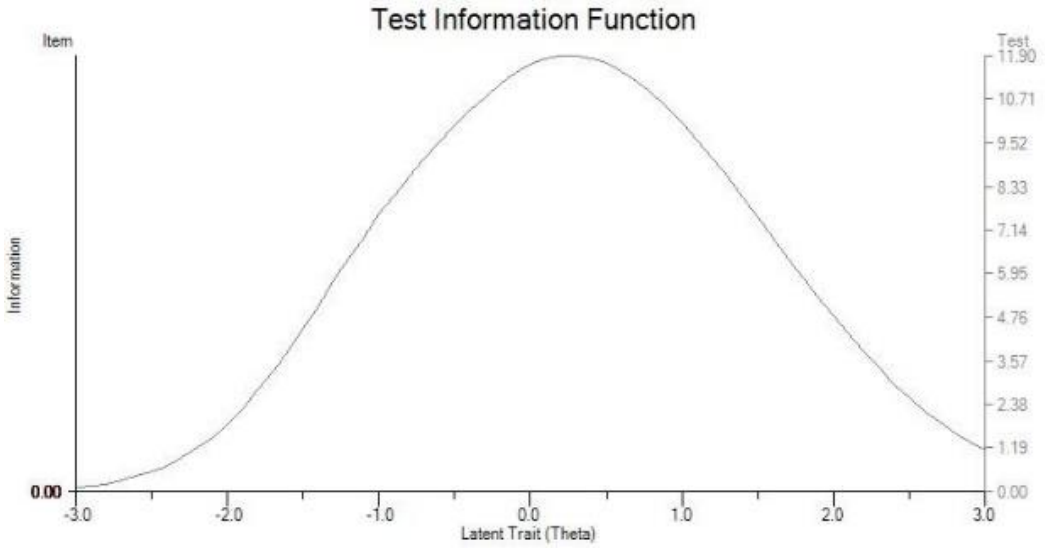
ผลการจำลองข้อมูลพารามิเตอร์ข้อสอบตามสถานการณ์เงื่อนไขและผลการวิเคราะห์ค่าเฉลี่ยของพารามิเตอร์ข้อสอบที่ได้จากการจำลองข้อมูลของสถานการณ์ที่ 5-6 แสดงได้ดังตารางต่อไปนี้

ตารางที่ 4.4 ค่าพารามิเตอร์ข้อสอบของข้อมูลจำลองสถานการณ์ที่ 5-6

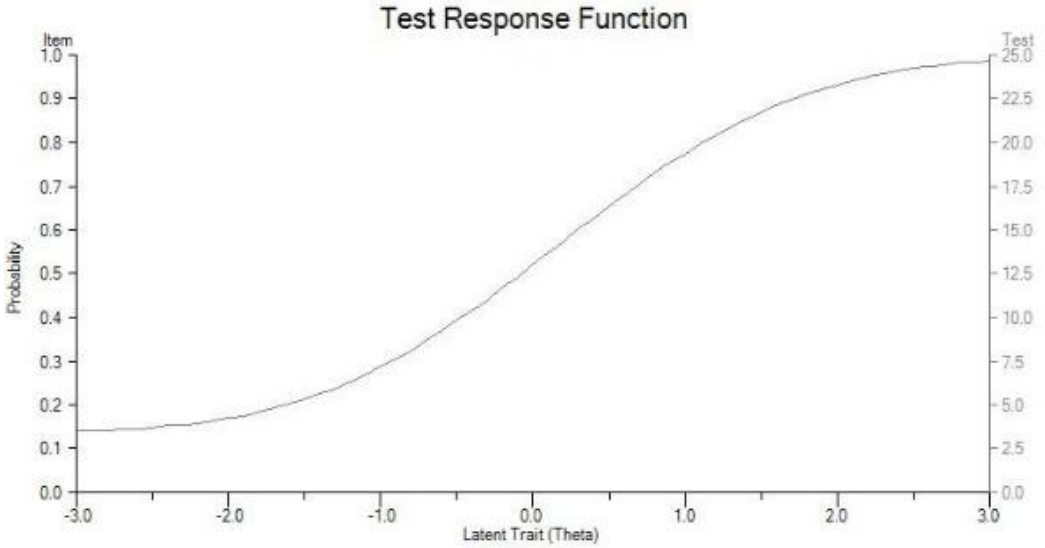
ข้อสอบ	253PL10				253PL20			
	a	b	c	model	a	b	c	model
1	1.102	0.444	0.062	3PL	0.905	-0.031	0.053	3PL
2	1.106	0.586	0.085	3PL	1.201	0.011	0.189	3PL
3	1.313	1.358	0.069	3PL	1.303	-0.226	0.104	3PL
4	1.239	-0.319	0.015	3PL	1.133	-0.711	0.095	3PL
5	1.205	-0.295	0.109	3PL	1.625	-0.539	0.140	3PL

ข้อสอบ	253PL10				253PL20			
	a	b	c	model	a	b	c	model
6	1.418	0.086	0.368	3PL	1.250	-1.716	0.064	3PL
7	1.104	-0.983	0.130	3PL	1.056	0.052	0.110	3PL
8	1.021	1.753	0.061	3PL	1.760	-0.075	0.144	3PL
9	1.084	0.966	0.241	3PL	1.010	-1.475	0.190	3PL
10	1.013	0.238	0.263	3PL	1.389	0.454	0.199	3PL
11	1.268	-1.026	0.190	3PL	1.267	0.273	0.102	3PL
12	1.258	1.175	0.087	3PL	1.553	-0.030	0.194	3PL
13	1.177	0.403	0.336	3PL	1.080	-0.875	0.317	3PL
14	1.302	0.859	0.096	3PL	1.408	0.390	0.210	3PL
15	1.301	1.998	0.116	3PL	1.221	-0.187	0.173	3PL
16	0.841	0.117	0.373	3PL	1.311	-0.435	0.126	3PL
17	1.446	0.405	0.180	3PL	1.414	0.271	0.124	3PL
18	1.250	-1.553	0.100	3PL	0.960	0.345	0.197	3PL
19	1.348	-1.104	0.033	3PL	1.021	0.187	0.116	3PL
20	1.253	0.804	0.261	3PL	1.274	-0.494	0.473	3PL
21	1.335	-0.842	0.097	3PL	1.238	0.343	0	2PL*
22	1.098	-0.508	0.109	3PL	1.296	0.595	0	2PL*
23	1.296	-0.094	0	2PL*	1.047	-0.538	0	2PL*
24	1.475	-0.071	0	2PL*	-	-0.789	0	1PL*
25	-	0.458	0	1PL*	-	-0.518	0	1PL*
ค่าเฉลี่ย	***1.219***	***0.194***	***0.154***		***1.249***	***-0.229***	***0.166***	

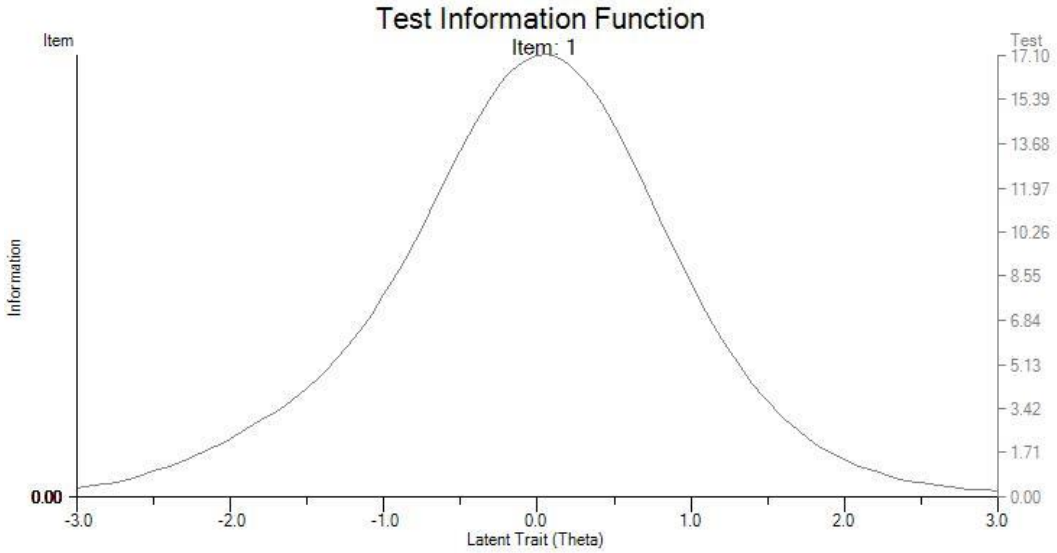
* ข้อสอบที่ไม่เหมาะสมกับโมเดลการวัด



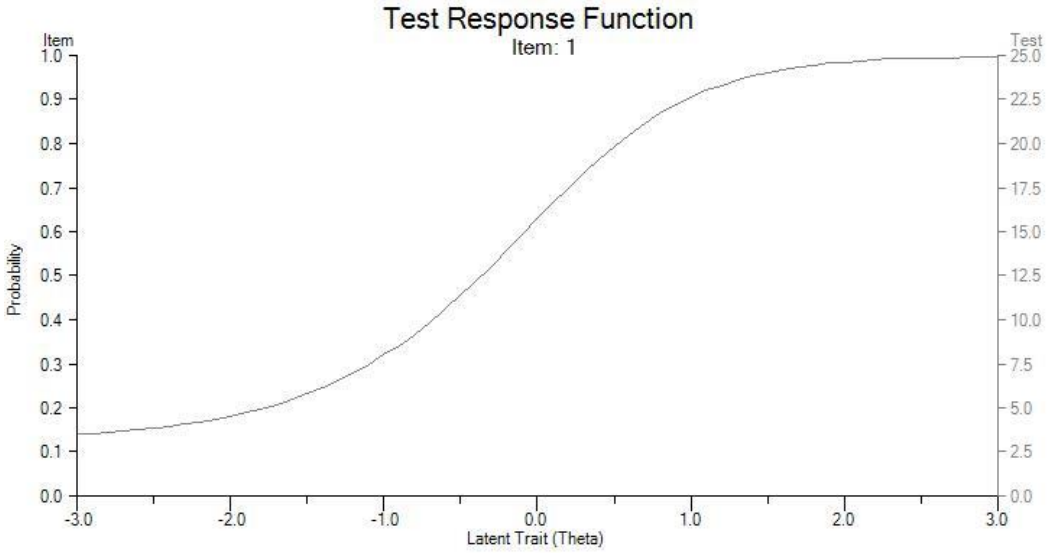
ภาพที่ 4.9 ฟังก์ชันสารสนเทศของแบบสอบของข้อมูลจำลองชุดที่ 5 (253PL10)



ภาพที่ 4.10 ฟังก์ชันสารสนเทศของแบบสอบและฟังก์ชันการตอบสนองข้อสอบของข้อมูลจำลองชุดที่ 5 (253PL10)



ภาพที่ 4.11 ฟังก์ชันสารสนเทศของแบบสอบของข้อมูลจำลองชุดที่ 6 (253PL20)



ภาพที่ 4.12 ฟังก์ชันการตอบสนองข้อสอบของข้อมูลจำลองชุดที่ 6 (253PL20)

ผลการจำลองข้อมูลของสถานการณ์ที่ 7 (501PL10) พบว่า ข้อสอบชุดนี้มีค่าอำนาจจำแนก (a) อยู่ในช่วง 1.156 ถึง 1.438 และมีค่าเฉลี่ยเท่ากับ 1.286 ค่าความยาก (b) อยู่ในช่วง -1.990 ถึง 2.407 และมีค่าเฉลี่ยเท่ากับ 0.011 และโอกาสในการเดาข้อสอบถูก (c) อยู่ในช่วง 0 ถึง 0.503 และมีค่าเฉลี่ยเท่ากับ 0.265 โดยมีข้อสอบที่ไม่สอดคล้องกับโมเดลจำนวน 5 ข้อ คือข้อ 46-50

ผลการจำลองข้อมูลของสถานการณ์ที่ 8 (501PL20) พบว่า ข้อสอบชุดนี้มีค่าอำนาจจำแนก (a) อยู่ในช่วง 0.858 ถึง 1.285 และมีค่าเฉลี่ยเท่ากับ 1.136 ค่าความยาก (b) อยู่ในช่วง -2.859 ถึง 2.905 และมีค่าเฉลี่ยเท่ากับ 0.035 และโอกาสในการเดาข้อสอบถูก (c) อยู่ในช่วง 0 ถึง 0.296 และมีค่าเฉลี่ยเท่ากับ 0.158 โดยมีข้อสอบที่ไม่สอดคล้องกับโมเดลจำนวน 10 ข้อ คือข้อ 41-50

ผลการจำลองข้อมูลพารามิเตอร์ข้อสอบตามสถานการณ์เงื่อนไขและผลการวิเคราะห์ค่าเฉลี่ยของพารามิเตอร์ข้อสอบที่ได้จากการจำลองข้อมูลของสถานการณ์ที่ 7-8 แสดงได้ดังตารางต่อไปนี้

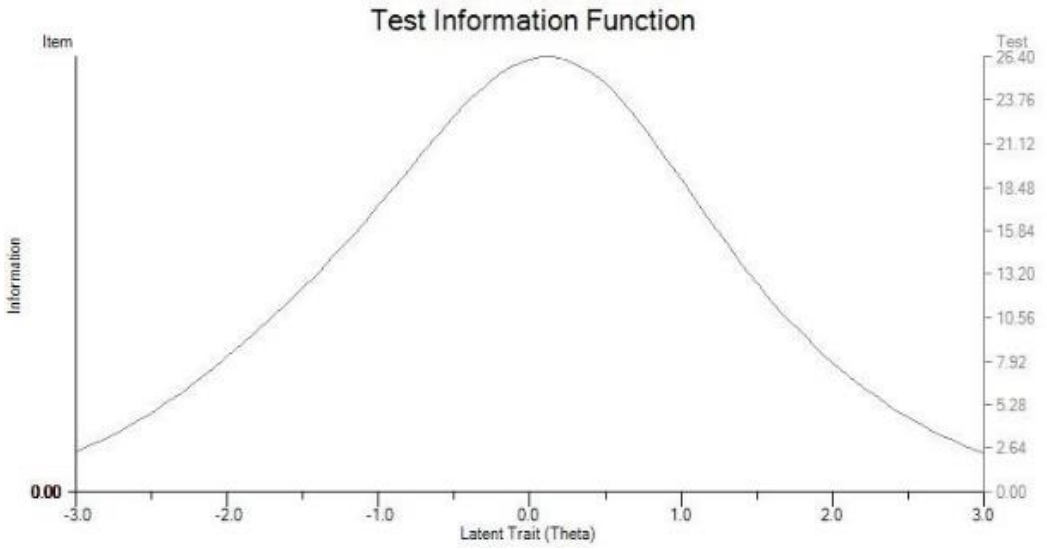
ตารางที่ 4.5 ค่าพารามิเตอร์ข้อสอบของข้อมูลจำลองสถานการณ์ที่ 7-8

ข้อสอบ	501PL10				501PL20			
	a	b	c	model	a	b	c	model
1	-	-0.036	0	1PL	-	-0.408	0	1PL
2	-	-1.334	0	1PL	-	0.594	0	1PL
3	-	-0.761	0	1PL	-	-0.557	0	1PL
4	-	-0.445	0	1PL	-	0.400	0	1PL
5	-	0.513	0	1PL	-	-0.444	0	1PL
6	-	0.847	0	1PL	-	-0.244	0	1PL
7	-	1.138	0	1PL	-	0.716	0	1PL
8	-	-1.190	0	1PL	-	-0.764	0	1PL
9	-	2.407	0	1PL	-	-0.923	0	1PL
10	-	-0.289	0	1PL	-	1.534	0	1PL
11	-	-1.821	0	1PL	-	-1.606	0	1PL
12	-	1.906	0	1PL	-	-0.770	0	1PL
13	-	-1.079	0	1PL	-	0.034	0	1PL
14	-	0.245	0	1PL	-	-1.010	0	1PL
15	-	0.412	0	1PL	-	2.905	0	1PL
16	-	1.554	0	1PL	-	0.545	0	1PL
17	-	0.307	0	1PL	-	0.610	0	1PL

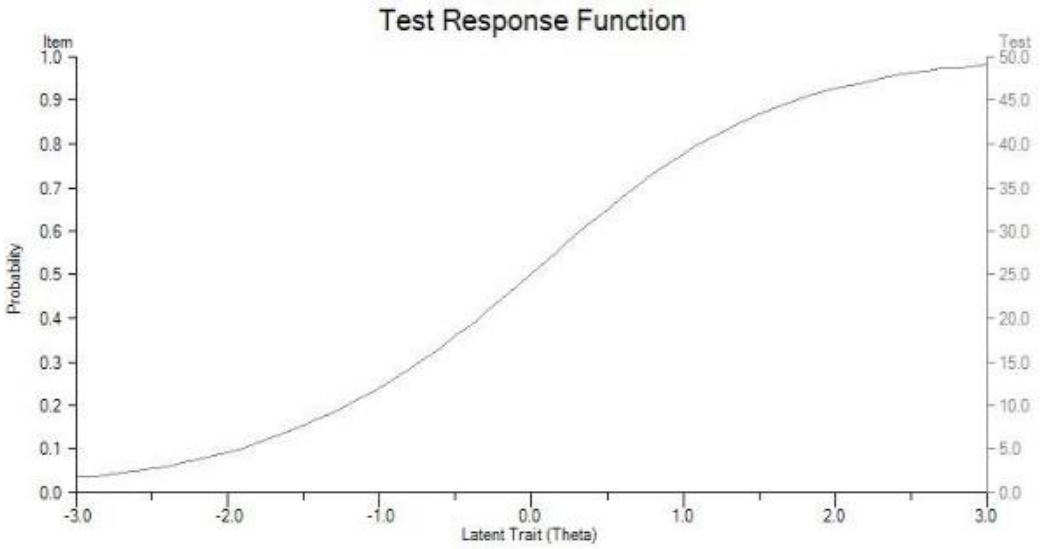
ข้อสอบ	501PL10				501PL20			
	a	b	c	model	a	b	c	model
18	-	-0.267	0	1PL	-	-1.437	0	1PL
19	-	0.570	0	1PL	-	0.024	0	1PL
20	-	0.177	0	1PL	-	-0.049	0	1PL
21	-	-0.943	0	1PL	-	1.977	0	1PL
22	-	1.527	0	1PL	-	0.644	0	1PL
23	-	0.147	0	1PL	-	1.121	0	1PL
24	-	0.064	0	1PL	-	1.169	0	1PL
25	-	-0.254	0	1PL	-	-0.079	0	1PL
26	-	-0.114	0	1PL	-	-0.113	0	1PL
27	-	-0.535	0	1PL	-	-0.487	0	1PL
28	-	0.239	0	1PL	-	-0.565	0	1PL
29	-	-0.641	0	1PL	-	0.803	0	1PL
30	-	-1.636	0	1PL	-	-1.738	0	1PL
31	-	0.181	0	1PL	-	-1.503	0	1PL
32	-	0.125	0	1PL	-	0.071	0	1PL
33	-	0.705	0	1PL	-	-0.175	0	1PL
34	-	0.723	0	1PL	-	0.419	0	1PL
35	-	0.445	0	1PL	-	-0.685	0	1PL
36	-	0.334	0	1PL	-	1.388	0	1PL
37	-	-1.830	0	1PL	-	-0.110	0	1PL
38	-	-1.488	0	1PL	-	0.780	0	1PL
39	-	0.207	0	1PL	-	1.079	0	1PL
40	-	-0.214	0	1PL	-	1.299	0	1PL
41	-	-1.990	0	1PL	1.093	0.574	0.106	3PL*
42	-	0.271	0	1PL	1.064	0.712	0.134	3PL*
43	-	-0.514	0	1PL	1.052	-0.582	0.110	3PL*
44	-	0.904	0	1PL	1.254	0.217	0.296	3PL*
45	-	0.258	0	1PL	1.254	0.064	0.144	3PL*

ข้อสอบ	501PL10				501PL20			
	a	b	c	model	a	b	c	model
46	1.325	-0.457	0.135	3PL*	1.244	0.449	0	2PL*
47	1.156	1.322	0.503	3PL*	1.193	-0.117	0	2PL*
48	1.438	0.230	0.158	3PL*	1.285	-0.190	0	2PL*
49	1.269	0.600	0	2PL*	0.858	-2.859	0	2PL*
50	1.244	0.035	0	2PL*	1.058	-0.953	0	2PL*
ค่าเฉลี่ย	***1.286***	***0.011***	***0.265***		***1.136***	***0.035***	***0.158***	

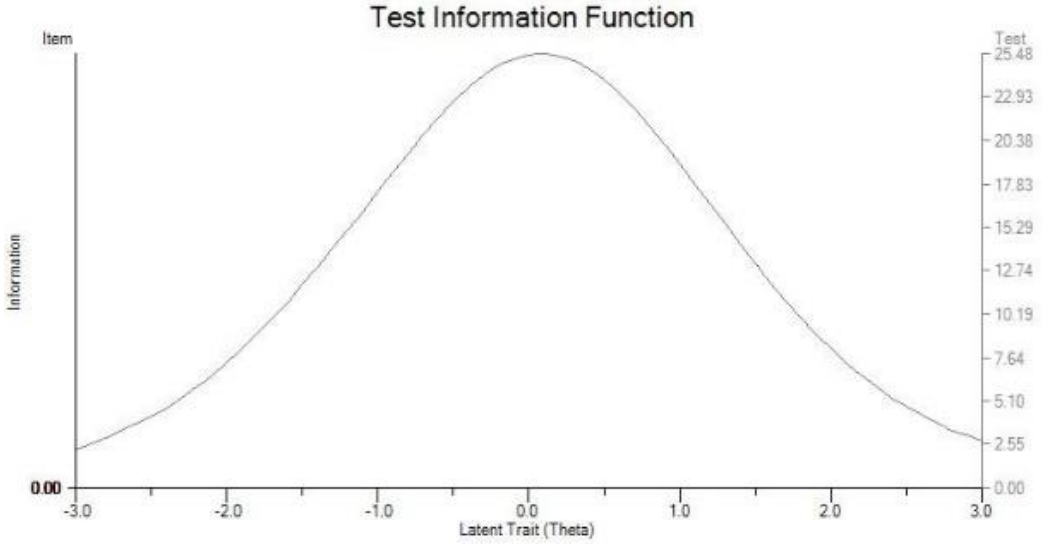
* ข้อสอบที่ไม่เหมาะสมกับโมเดลการวัด



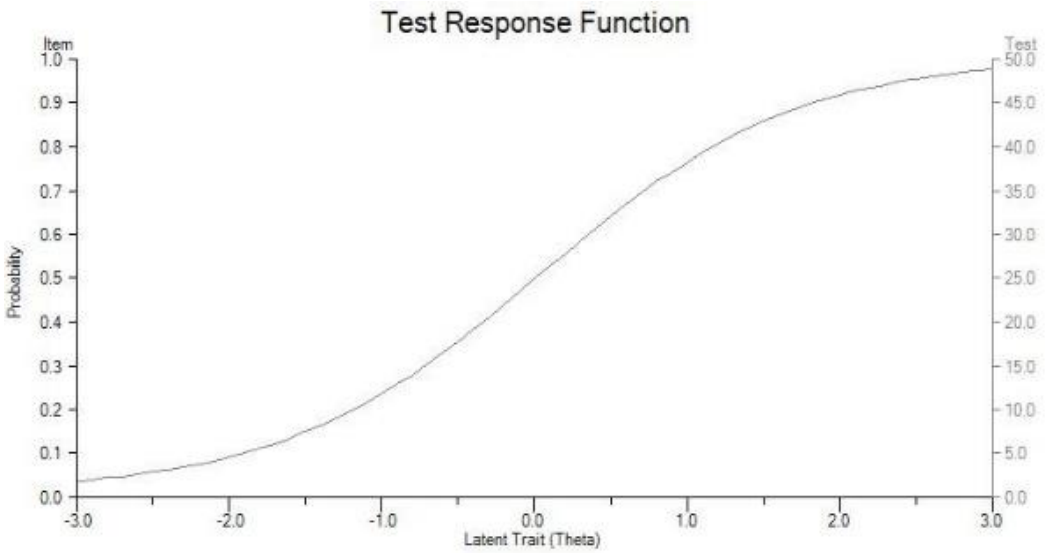
ภาพที่ 4.13 ฟังก์ชันสารสนเทศของแบบสอบของข้อมูลจำลองชุดที่ 7 (501PL10)



ภาพที่ 4.14 ฟังก์ชันการตอบสนองข้อสอบของข้อมูลจำลองชุดที่ 7 (501PL10)



ภาพที่ 4.15 ฟังก์ชันสารสนเทศของแบบสอบของข้อมูลจำลองชุดที่ 8 (501PL20)



ภาพที่ 4.16 ฟังก์ชันการตอบสนองข้อสอบของข้อมูลจำลองชุดที่ 8 (501PL20)

ผลการจำลองข้อมูลของสถานการณ์ที่ 9 (502PL10) พบว่า ข้อสอบชุดนี้มีค่าอำนาจจำแนก (a) อยู่ในช่วง 0.886 ถึง 1.716 และมีค่าเฉลี่ยเท่ากับ 1.221 ค่าความยาก (b) อยู่ในช่วง -2.282 ถึง 3.114 และมีค่าเฉลี่ยเท่ากับ 0.077 และโอกาสในการเดาข้อสอบถูก (c) อยู่ในช่วง 0 ถึง 0.308 และมีค่าเฉลี่ยเท่ากับ 0.169 โดยมีข้อสอบที่ไม่สอดคล้องกับโมเดลจำนวน 5 ข้อ คือข้อ 46-50

ผลการจำลองข้อมูลของสถานการณ์ที่ 10 (502PL20) พบว่า ข้อสอบชุดนี้มีค่าอำนาจจำแนก (a) อยู่ในช่วง 0.961 ถึง 1.560 และมีค่าเฉลี่ยเท่ากับ 1.234 ค่าความยาก (b) อยู่ในช่วง -1.833 ถึง 2.020 และมีค่าเฉลี่ยเท่ากับ 0.128 และโอกาสในการเดาข้อสอบถูก (c) อยู่ในช่วง 0 ถึง 0.219 และมีค่าเฉลี่ยเท่ากับ 0.173 โดยมีข้อสอบที่ไม่สอดคล้องกับโมเดลจำนวน 10 ข้อ คือข้อ 41-50

ผลการจำลองข้อมูลพารามิเตอร์ข้อสอบตามสถานการณ์เงื่อนไขและผลการวิเคราะห์ค่าเฉลี่ยของพารามิเตอร์ข้อสอบที่ได้จากการจำลองข้อมูลของสถานการณ์ที่ 9-10 แสดงได้ดังตารางต่อไปนี้

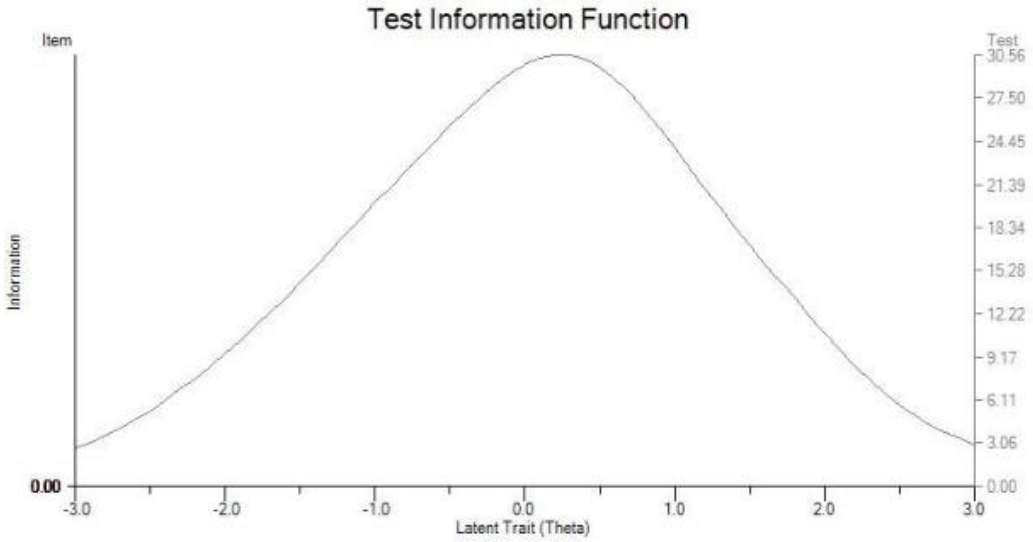
ตารางที่ 4.6 ค่าพารามิเตอร์ข้อสอบของข้อมูลจำลองสถานการณ์ที่ 9-10

ข้อสอบ	502PL10				502PL20			
	a	b	c	model	a	b	c	model
1	1.083	-2.282	0	2PL	1.229	0.989	0	2PL
2	1.085	0.870	0	2PL	1.025	0.023	0	2PL
3	1.253	-0.950	0	2PL	1.061	0.565	0	2PL
4	1.229	-0.437	0	2PL	1.320	1.452	0	2PL
5	1.192	0.833	0	2PL	1.335	-0.863	0	2PL

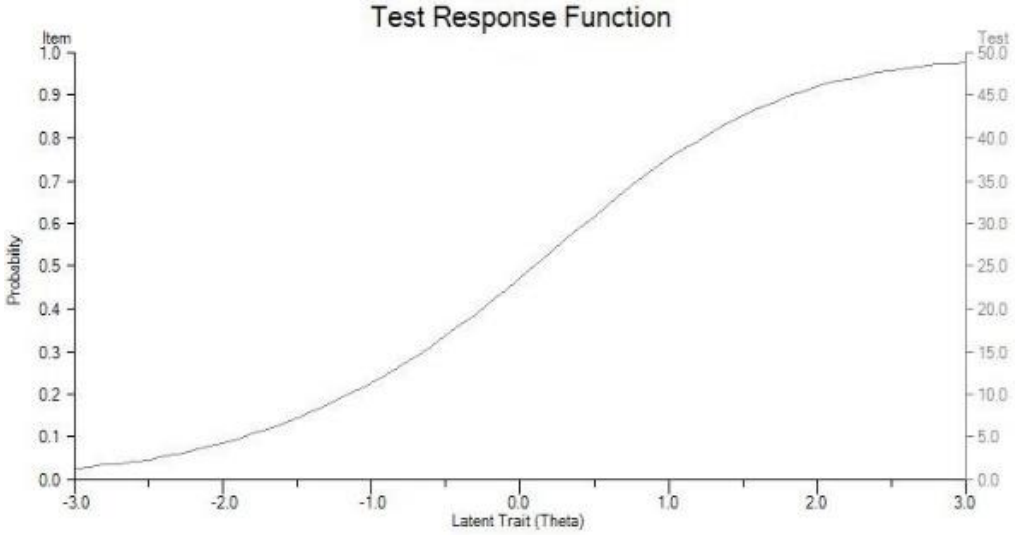
ข้อสอบ	502PL10				502PL20			
	a	b	c	model	a	b	c	model
6	1.297	-0.997	0	2PL	1.305	0.441	0	2PL
7	1.249	0.073	0	2PL	0.965	2.020	0	2PL
8	0.894	0.777	0	2PL	1.031	1.924	0	2PL
9	1.399	0.835	0	2PL	1.279	-0.292	0	2PL
10	1.200	0.483	0	2PL	1.542	0.144	0	2PL
11	0.965	0.846	0	2PL	1.189	-0.019	0	2PL
12	1.248	0.382	0	2PL	1.206	0.322	0	2PL
13	1.069	-0.464	0	2PL	1.326	0.070	0	2PL
14	1.119	-0.882	0	2PL	1.243	0.547	0	2PL
15	1.392	0.364	0	2PL	1.303	0.121	0	2PL
16	1.145	3.114	0	2PL	1.104	-0.532	0	2PL
17	1.411	0.069	0	2PL	1.227	0.885	0	2PL
18	1.207	-0.374	0	2PL	1.404	-1.833	0	2PL
19	1.175	0.200	0	2PL	1.119	-1.623	0	2PL
20	1.327	1.835	0	2PL	1.128	-0.962	0	2PL
21	1.056	-1.613	0	2PL	1.382	0.614	0	2PL
22	1.273	0.233	0	2PL	1.423	0.439	0	2PL
23	1.140	-0.834	0	2PL	1.090	1.751	0	2PL
24	1.065	0.915	0	2PL	1.200	-0.605	0	2PL
25	1.210	0.326	0	2PL	0.963	-1.536	0	2PL
26	1.284	0.244	0	2PL	1.467	-1.242	0	2PL
27	1.452	-0.330	0	2PL	1.438	-0.379	0	2PL
28	0.886	-1.841	0	2PL	1.174	0.385	0	2PL
29	1.305	0.043	0	2PL	1.179	0.616	0	2PL
30	1.009	0.426	0	2PL	1.328	-1.580	0	2PL
31	1.536	-1.503	0	2PL	0.961	0.330	0	2PL
32	1.716	-0.991	0	2PL	1.308	-0.675	0	2PL
33	1.369	1.003	0	2PL	1.124	0.763	0	2PL

ข้อสอบ	502PL10				502PL20			
	a	b	c	model	a	b	c	model
34	1.409	0.490	0	2PL	1.178	0.045	0	2PL
35	1.284	1.739	0	2PL	1.323	0.447	0	2PL
36	1.076	0.455	0	2PL	1.560	-0.226	0	2PL
37	1.194	-0.476	0	2PL	1.446	1.029	0	2PL
38	1.084	1.404	0	2PL	1.044	-0.609	0	2PL
39	1.151	0.649	0	2PL	1.111	0.137	0	2PL
40	1.449	-0.241	0	2PL	1.458	0.932	0	2PL
41	1.281	-1.904	0	2PL	0.977	1.311	0.219	3PL*
42	1.022	0.837	0	2PL	1.258	0.023	0.106	3PL*
43	1.317	0.344	0	2PL	1.277	-1.350	0.150	3PL*
44	1.623	1.804	0	2PL	1.359	-0.576	0.179	3PL*
45	1.100	-1.952	0	2PL	1.149	1.177	0.210	3PL*
46	1.083	1.219	0.058	3PL*	-	-0.062	0	1PL*
47	1.288	-0.581	0.142	3PL*	-	-0.607	0	1PL*
48	0.991	0.455	0.308	3PL*	-	-0.214	0	1PL*
49	-	0.926	0	1PL*	-	1.365	0	1PL*
50	-	-1.693	0	1PL*	-	1.316	0	1PL*
ค่าเฉลี่ย	***1.221***	***0.077***	***0.169***		***1.234***	***0.128***	***0.173***	

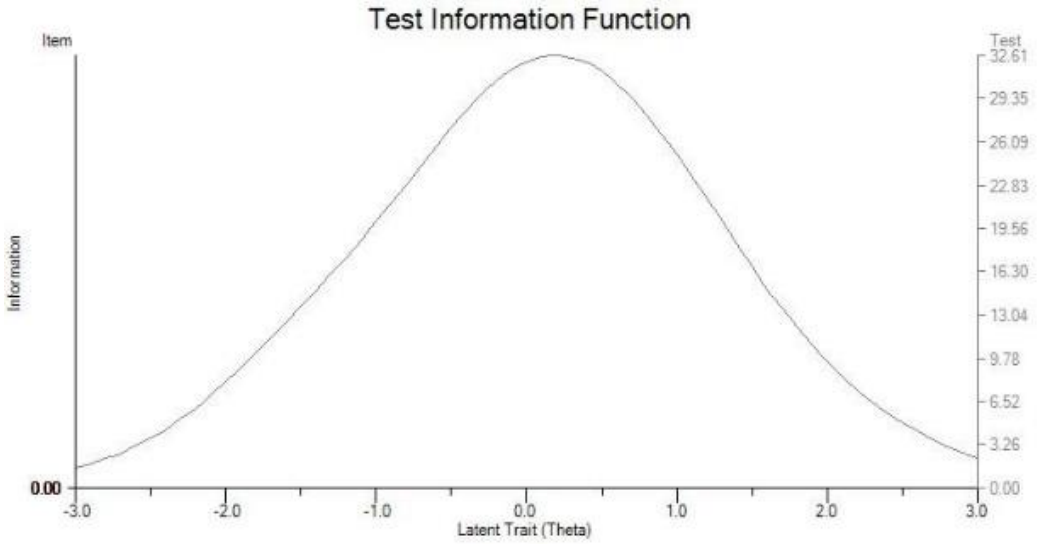
* ข้อสอบที่ไม่เหมาะสมกับโมเดลการวัด



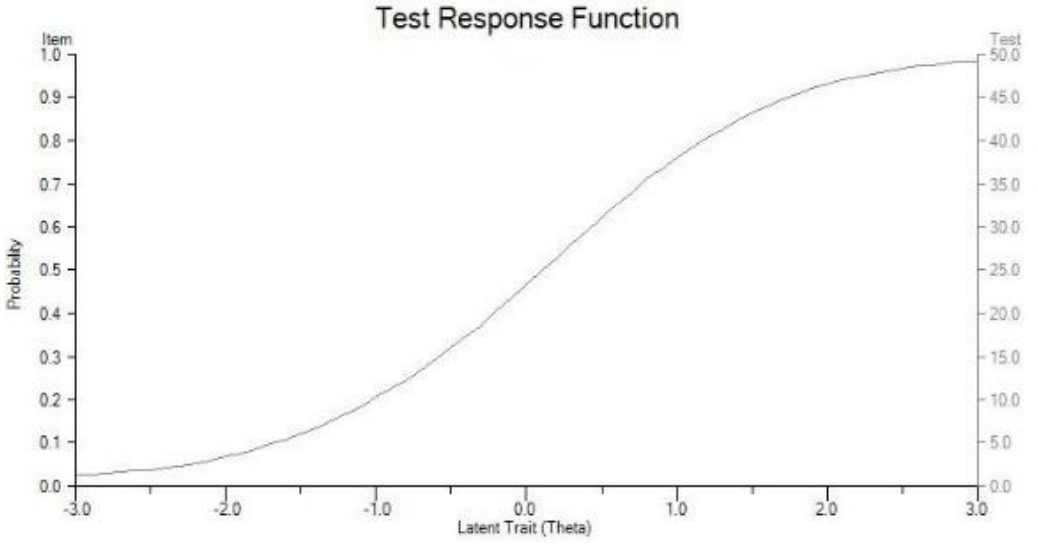
ภาพที่ 4.17 ฟังก์ชันสารสนเทศของแบบสอบของข้อมูลจำลองชุดที่ 9 (502PL10)



ภาพที่ 4.18 ฟังก์ชันการตอบสนองข้อสอบของข้อมูลจำลองชุดที่ 9 (502PL10)



ภาพที่ 4.19 ฟังก์ชันสารสนเทศของแบบสอบของข้อมูลจำลองชุดที่ 10 (502PL20)



ภาพที่ 4.20 ฟังก์ชันการตอบสนองข้อสอบของข้อมูลจำลองชุดที่ 10 (502PL20)

ผลการจำลองข้อมูลของสถานการณ์ที่ 11 (503PL10) พบว่า ข้อสอบชุดนี้มีค่าอำนาจจำแนก (a) อยู่ในช่วง 0.897 ถึง 1.815 และมีค่าเฉลี่ยเท่ากับ 1.237 ค่าความยาก (b) อยู่ในช่วง -1.98 ถึง 2.004 และมีค่าเฉลี่ยเท่ากับ -0.153 และโอกาสในการเดาข้อสอบถูก (c) อยู่ในช่วง 0 ถึง 0.421 และมีค่าเฉลี่ยเท่ากับ 0.199 โดยมีข้อสอบที่ไม่สอดคล้องกับโมเดลจำนวน 5 ข้อ คือข้อ 46-50

ผลการจำลองข้อมูลของสถานการณ์ที่ 12 (503PL20) พบว่า ข้อสอบชุดนี้มีค่าอำนาจจำแนก (a) อยู่ในช่วง 0.847 ถึง 1.763 และมีค่าเฉลี่ยเท่ากับ 1.236 ค่าความยาก (b) อยู่ในช่วง -1.702 ถึง 2.093 และมีค่าเฉลี่ยเท่ากับ 0.062 และโอกาสในการเดาข้อสอบถูก (c) อยู่ในช่วง 0 ถึง 0.46 และมีค่าเฉลี่ยเท่ากับ 0.168 โดยมีข้อสอบที่ไม่สอดคล้องกับโมเดลจำนวน 10 ข้อ คือข้อ 41-50

ผลการจำลองข้อมูลพารามิเตอร์ข้อสอบตามสถานการณ์เงื่อนไขและผลการวิเคราะห์ค่าเฉลี่ยของพารามิเตอร์ข้อสอบที่ได้จากการจำลองข้อมูลของสถานการณ์ที่ 11-12 แสดงได้ดังตารางต่อไปนี้

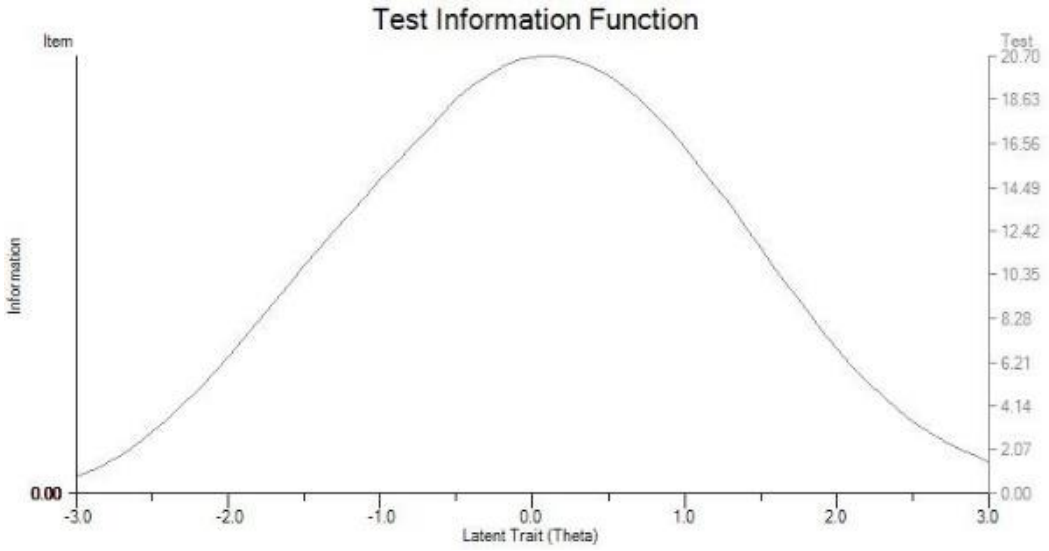
ตารางที่ 4.7 ค่าพารามิเตอร์ข้อสอบของข้อมูลจำลองสถานการณ์ที่ 11-12

ข้อสอบ	503PL10				503PL20			
	a	b	c	model	a	b	c	model
1	1.151	-0.813	0.065	3PL	1.161	1.080	0.307	3PL
2	1.815	-0.654	0.149	3PL	1.259	-0.218	0.087	3PL
3	1.128	-0.341	0.301	3PL	0.955	-0.898	0.226	3PL
4	1.600	0.357	0.174	3PL	1.183	1.364	0.208	3PL
5	1.203	-1.442	0.407	3PL	1.192	2.093	0.024	3PL
6	1.188	-1.004	0.049	3PL	1.305	0.507	0.303	3PL
7	1.149	-0.795	0.293	3PL	0.847	1.260	0.036	3PL
8	1.082	-0.401	0.041	3PL	1.138	0.507	0.121	3PL
9	1.409	2.004	0.279	3PL	0.963	0.047	0.231	3PL
10	1.158	-0.282	0.157	3PL	1.159	0.326	0.089	3PL
11	1.432	-0.433	0.359	3PL	1.006	-1.702	0.069	3PL
12	1.084	0.818	0.160	3PL	1.268	-0.253	0.082	3PL
13	1.507	-0.139	0.188	3PL	1.174	0.510	0.119	3PL
14	1.164	-0.244	0.312	3PL	1.419	-0.760	0.101	3PL
15	1.181	-0.273	0.261	3PL	1.124	-0.485	0.189	3PL
16	1.106	0.019	0.211	3PL	1.592	0.473	0.034	3PL
17	1.501	-0.493	0.252	3PL	1.216	-0.693	0.121	3PL

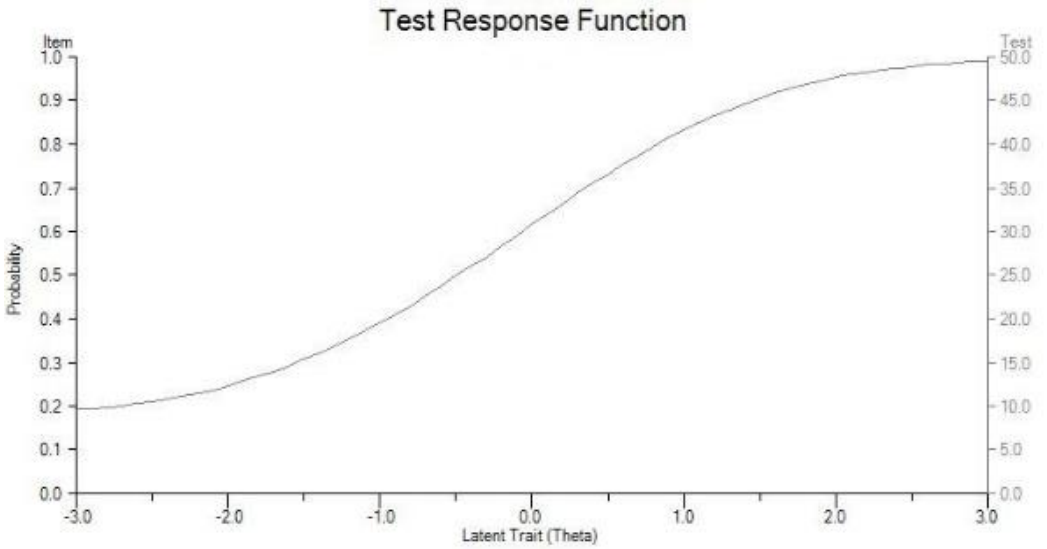
ข้อสอบ	503PL10				503PL20			
	a	b	c	model	a	b	c	model
18	1.087	-0.708	0.047	3PL	1.469	1.402	0.134	3PL
19	1.068	-0.526	0.080	3PL	1.289	0.245	0.209	3PL
20	1.083	1.289	0.073	3PL	1.198	1.084	0.337	3PL
21	0.912	-0.386	0.421	3PL	1.255	0.391	0.269	3PL
22	1.663	-1.363	0.224	3PL	1.382	-0.158	0.086	3PL
23	1.465	1.186	0.147	3PL	1.250	-1.516	0.418	3PL
24	1.326	-0.505	0.153	3PL	1.190	1.428	0.103	3PL
25	1.209	-0.906	0.213	3PL	1.281	-1.041	0.312	3PL
26	1.039	0.265	0.041	3PL	1.154	0.762	0.384	3PL
27	1.436	-0.086	0.152	3PL	1.023	0.347	0.060	3PL
28	1.306	0.371	0.213	3PL	1.185	0.715	0.460	3PL
29	1.066	-1.201	0.018	3PL	1.303	-0.816	0.052	3PL
30	1.530	-0.695	0.063	3PL	1.135	0.757	0.160	3PL
31	1.337	-1.225	0.171	3PL	1.002	0.651	0.420	3PL
32	1.258	1.134	0.112	3PL	1.612	0.628	0.101	3PL
33	0.997	0.050	0.105	3PL	1.489	-0.900	0.075	3PL
34	1.368	-1.955	0.076	3PL	1.348	-1.087	0.052	3PL
35	1.206	-0.796	0.338	3PL	1.309	-0.407	0.139	3PL
36	1.454	0.461	0.120	3PL	1.413	0.212	0.097	3PL
37	1.389	1.431	0.204	3PL	1.215	-1.368	0.137	3PL
38	0.981	-1.340	0.179	3PL	1.366	0.743	0.084	3PL
39	0.897	1.688	0.098	3PL	0.995	-1.052	0.175	3PL
40	1.026	-0.746	0.036	3PL	1.193	0.106	0.093	3PL
41	1.269	-0.937	0.194	3PL	1.469	-0.772	0	2PL*
42	1.331	-1.980	0.050	3PL	1.763	-1.683	0	2PL*
43	1.278	1.200	0.143	3PL	1.027	0.423	0	2PL*
44	1.317	-0.240	0.040	3PL	1.284	-1.101	0	2PL*
45	1.344	0.553	0.107	3PL	1.076	0.046	0	2PL*

ข้อสอบ	503PL10				503PL20			
	a	b	c	model	a	b	c	model
46	1.113	0.899	0	2PL*	-	-0.710	0	1PL*
47	1.206	0.656	0	2PL*	-	-0.663	0	1PL*
48	1.650	1.209	0	2PL*	-	0.548	0	1PL*
49	-	-0.586	0	1PL*	-	1.464	0	1PL*
50	-	-0.507	0	1PL*	-	1.278	0	1PL*
ค่าเฉลี่ย	***1.237***	***-0.153***	***0.199***		***1.236***	***0.062***	***0.168***	

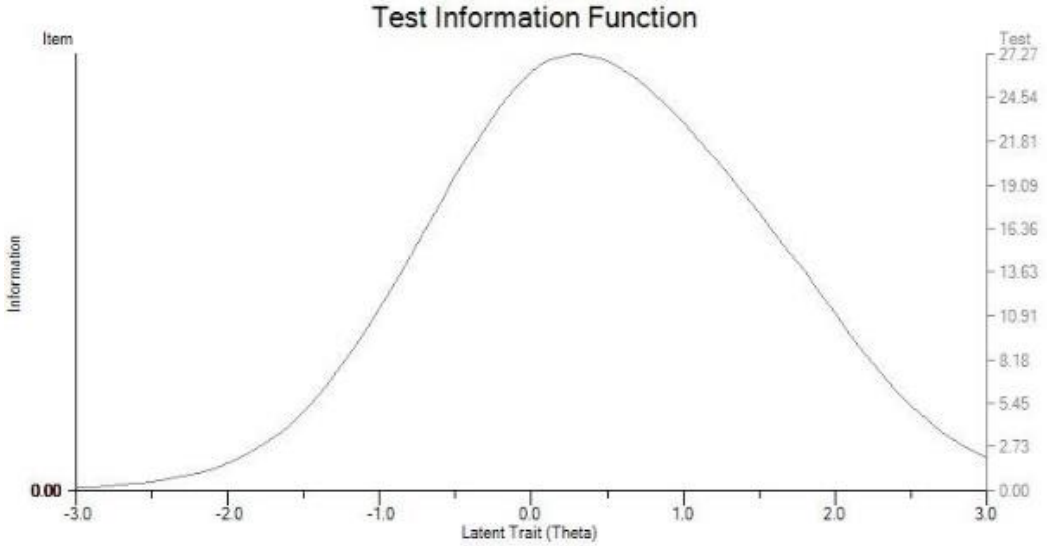
* ข้อสอบที่ไม่เหมาะสมกับโมเดลการวัด



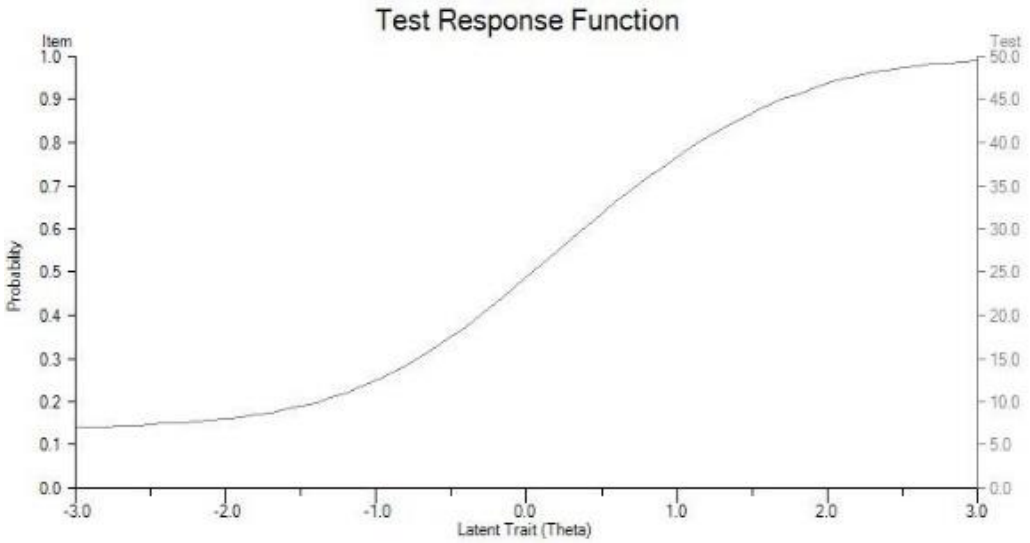
ภาพที่ 4.21 ฟังก์ชันสารสนเทศของแบบสอบของข้อมูลจำลองชุดที่ 11 (503PL10)



ภาพที่ 4.22 ฟังก์ชันการตอบสนองข้อสอบของข้อมูลจำลองชุดที่ 11 (503PL10)



ภาพที่ 4.23 ฟังก์ชันสารสนเทศของแบบสอบของข้อมูลจำลองชุดที่ 12 (503PL20)



ภาพที่ 4.24 ฟังก์ชันการตอบสนองข้อสอบของข้อมูลจำลองชุดที่ 12 (503PL20)

2) ผลการวิเคราะห์ค่าสถิติพื้นฐานของค่าพารามิเตอร์ความสามารถของผู้สอบที่ได้จากการจำลองข้อมูล

ส่วนนี้เป็นการนำเสนอผลการวิเคราะห์ค่าสถิติพื้นฐานของค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) ที่ได้จากการจำลองข้อมูลด้วยโปรแกรม WINGEN โดยทำการจำลองข้อมูลพารามิเตอร์ข้อสอบขึ้นมาจำนวน 12 ชุดตามเงื่อนไขในการศึกษาที่กำหนดไว้ ผลการวิเคราะห์ในส่วนนี้ประกอบด้วยผลการวิเคราะห์ค่าเฉลี่ยพารามิเตอร์ความสามารถของผู้สอบและลักษณะการแจกแจงของค่าความสามารถของผู้สอบที่ได้จากการจำลองข้อมูลแต่ละสถานการณ์เงื่อนไข ผลการวิเคราะห์มีรายละเอียดดังนี้

ผลการจำลองข้อมูลของสถานการณ์ที่ 1 (251PL10) พบว่า ผู้สอบมีค่าความสามารถเฉลี่ยเท่ากับ -0.029 และมีลักษณะการแจกแจงแบบปกติดังภาพที่ 4.13

ผลการจำลองข้อมูลของสถานการณ์ที่ 2 (251PL20) พบว่า ผู้สอบมีค่าความสามารถเฉลี่ยเท่ากับ 0.001 และมีลักษณะการแจกแจงแบบปกติดังภาพที่ 4.14

ผลการจำลองข้อมูลของสถานการณ์ที่ 3 (252PL10) พบว่า ผู้สอบมีค่าความสามารถเฉลี่ยเท่ากับ 0.015 และมีลักษณะการแจกแจงแบบปกติดังภาพที่ 4.15

ผลการจำลองข้อมูลของสถานการณ์ที่ 4 (252PL20) พบว่า ผู้สอบมีค่าความสามารถเฉลี่ยเท่ากับ 0.008 และมีลักษณะการแจกแจงแบบปกติดังภาพที่ 4.16

ผลการจำลองข้อมูลของสถานการณ์ที่ 5 (253PL10) พบว่า ผู้สอบมีค่าความสามารถเฉลี่ยเท่ากับ 0.035 และมีลักษณะการแจกแจงแบบปกติดังภาพที่ 4.17

ผลการจำลองข้อมูลของสถานการณ์ที่ 6 (253PL20) พบว่า ผู้สอบมีค่าความสามารถเฉลี่ยเท่ากับ 0.007 และมีลักษณะการแจกแจงแบบปกติดังภาพที่ 4.18

ผลการจำลองข้อมูลของสถานการณ์ที่ 7 (501PL10) พบว่า ผู้สอบมีค่าความสามารถเฉลี่ยเท่ากับ 0.031 และมีลักษณะการแจกแจงแบบปกติดังภาพที่ 4.19

ผลการจำลองข้อมูลของสถานการณ์ที่ 8 (501PL20) พบว่า ผู้สอบมีค่าความสามารถเฉลี่ยเท่ากับ 0.057 และมีลักษณะการแจกแจงแบบปกติดังภาพที่ 4.20

ผลการจำลองข้อมูลของสถานการณ์ที่ 9 (502PL10) พบว่า ผู้สอบมีค่าความสามารถเฉลี่ยเท่ากับ 0.005 และมีลักษณะการแจกแจงแบบปกติดังภาพที่ 4.21

ผลการจำลองข้อมูลของสถานการณ์ที่ 10 (502PL20) พบว่า ผู้สอบมีค่าความสามารถเฉลี่ยเท่ากับ -0.001 และมีลักษณะการแจกแจงแบบปกติดังภาพที่ 4.22

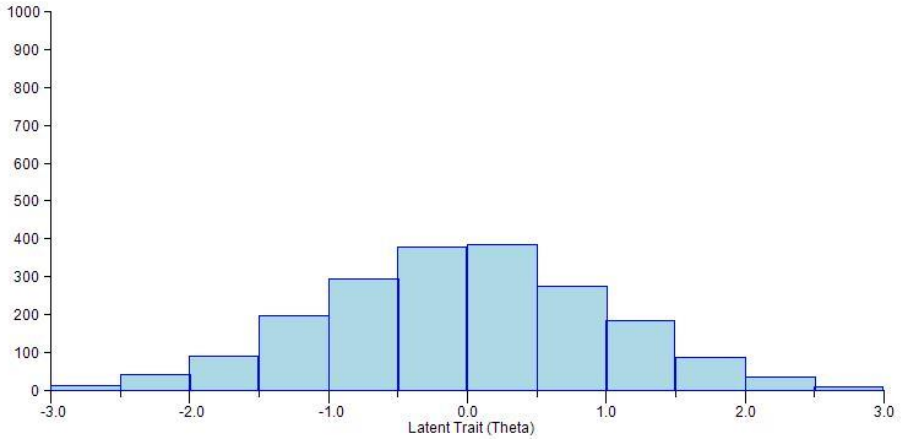
ผลการจำลองข้อมูลของสถานการณ์ที่ 11 (503PL10) พบว่า ผู้สอบมีค่าความสามารถเฉลี่ยเท่ากับ 0.025 และมีลักษณะการแจกแจงแบบปกติดังภาพที่ 4.23

ผลการจำลองข้อมูลของสถานการณ์ที่ 12 (503PL20) พบว่า ผู้สอบมีค่าความสามารถเฉลี่ยเท่ากับ 0.006 และมีลักษณะการแจกแจงแบบปกติดังภาพที่ 4.24

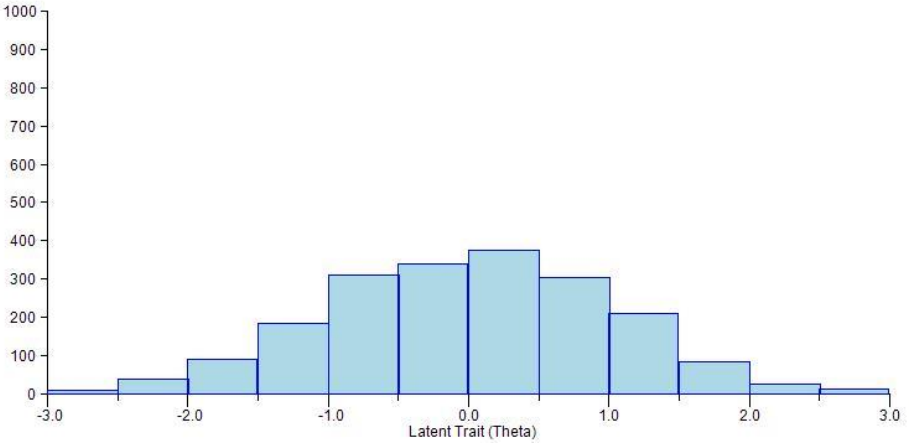
ผลการจำลองข้อมูลพารามิเตอร์ความสามารถของผู้สอบและลักษณะการแจกแจงของค่าความสามารถของผู้สอบที่ได้จากการจำลองข้อมูลของสถานการณ์ที่ 1-12 แสดงได้ดังตารางต่อไปนี้

ตารางที่ 4.8 ค่าสถิติพื้นฐานของค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) ที่ได้จากการจำลองข้อมูล

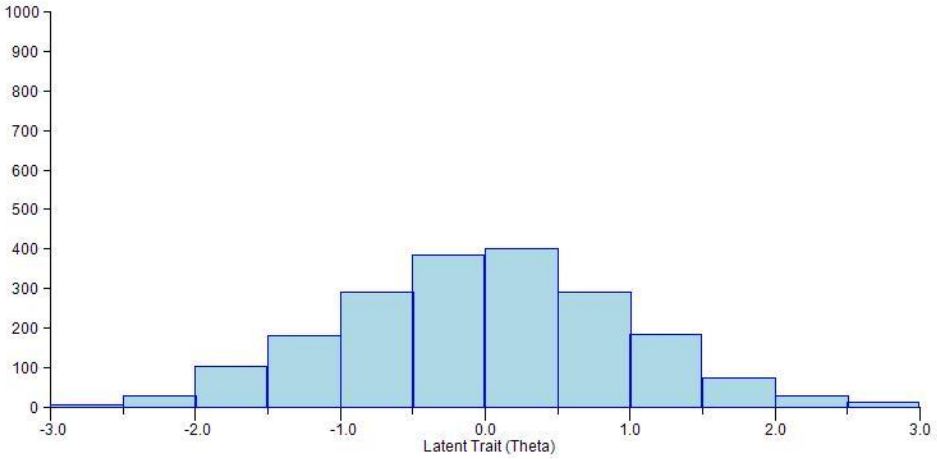
สถานการณ์	Max	Min	Mean	SD	Sk	Ku
251PL10	-3.464	3.437	-0.029	1.031	-0.006	0.032
251PL20	-3.688	3.444	0.001	1.029	-0.090	0.052
252PL10	-3.397	3.192	0.015	0.983	-0.012	0.009
252PL20	-3.344	4.047	0.008	0.992	0.000	0.095
253PL10	-3.449	3.186	0.035	1.013	0.027	-0.055
253PL20	-3.460	3.394	0.007	0.993	-0.032	0.210
501PL10	-4.607	3.329	0.031	0.977	0.005	0.031
501PL20	-3.187	3.123	0.057	1.006	-0.074	-0.249
502PL10	-4.255	3.619	0.005	1.028	-0.052	-0.053
502PL20	-3.442	3.138	-0.001	0.983	-0.012	0.097
503PL10	-4.462	3.632	0.025	1.017	-0.003	0.099
503PL20	-3.463	3.459	0.006	0.978	0.013	-0.067



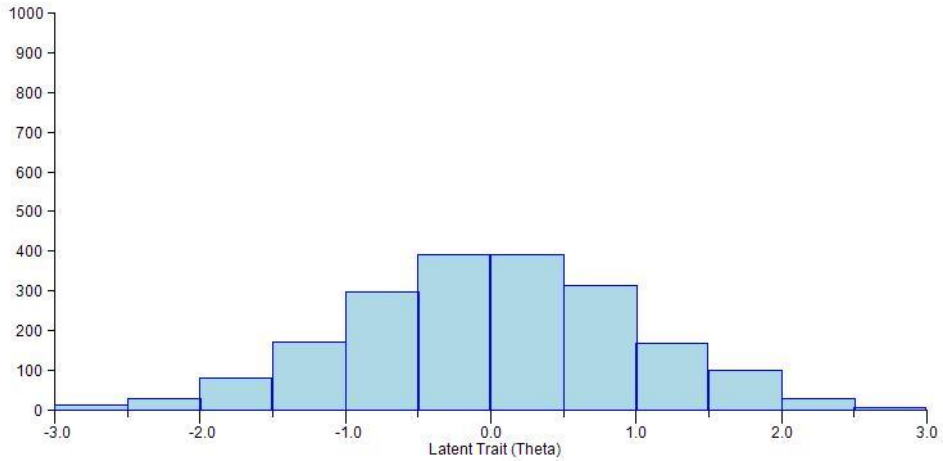
ภาพที่ 4.25 ลักษณะการแจกแจงค่าความสามารถผู้สอบของข้อมูลจำลองชุดที่ 1 (251PL10)



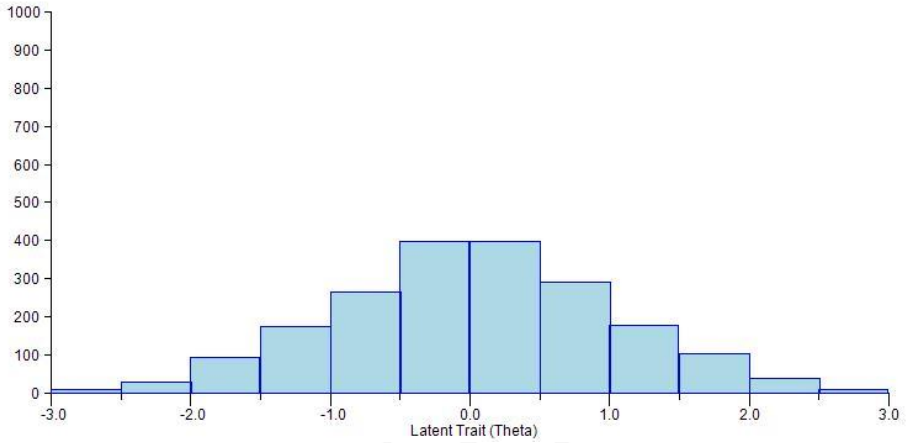
ภาพที่ 4.26 ลักษณะการแจกแจงค่าความสามารถผู้สอบของข้อมูลจำลองชุดที่ 2 (251PL20)



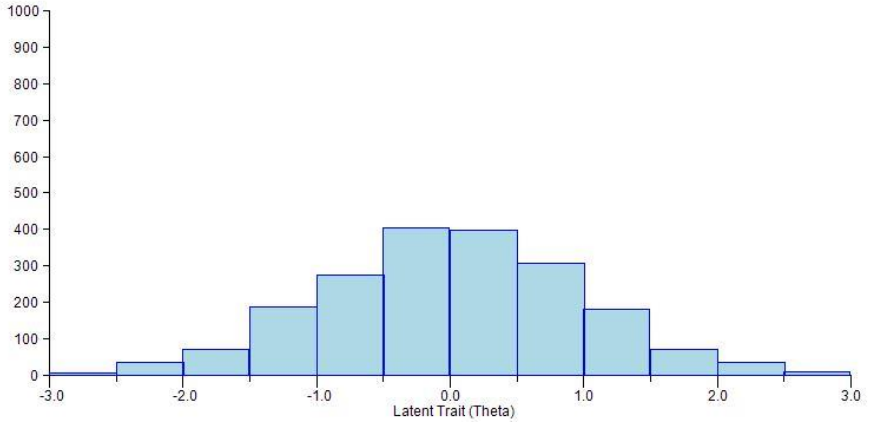
ภาพที่ 4.27 ลักษณะการแจกแจงค่าความสามารถผู้สอบของข้อมูลจำลองชุดที่ 3 (252PL10)



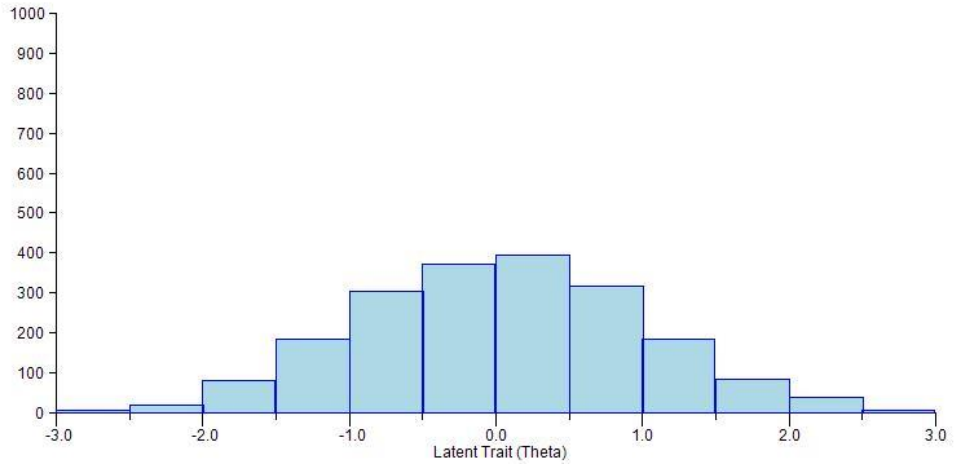
ภาพที่ 4.28 ลักษณะการแจกแจงค่าความสามารถผู้สอบของข้อมูลจำลองชุดที่ 4 (252PL20)



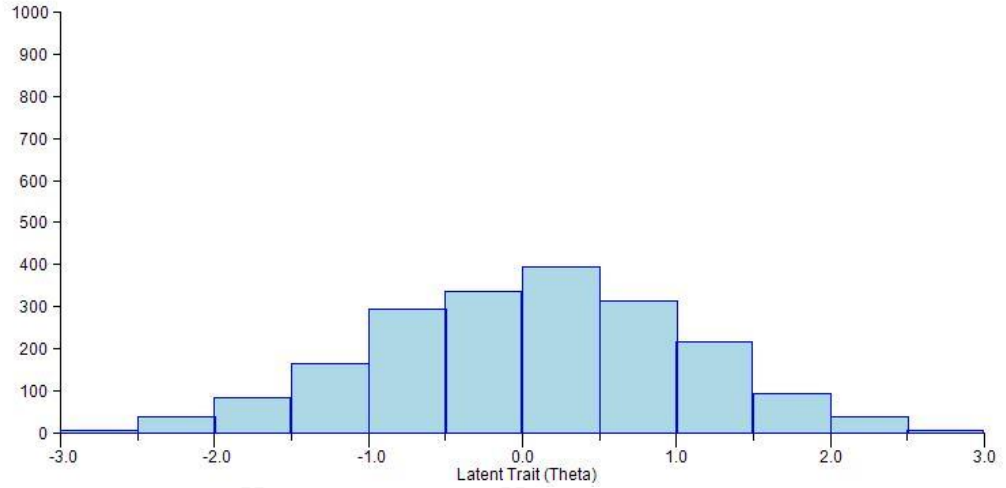
ภาพที่ 4.29 ลักษณะการแจกแจงค่าความสามารถผู้สอบของข้อมูลจำลองชุดที่ 5 (253PL10)



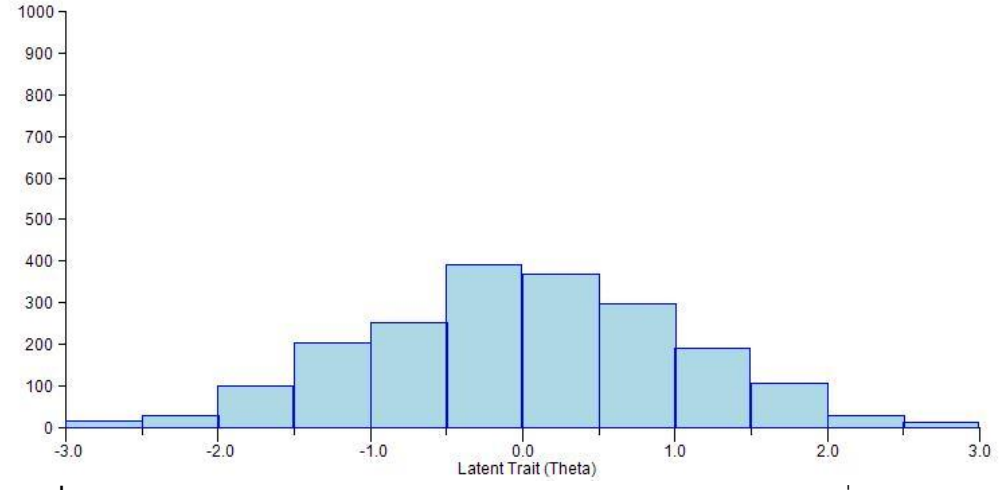
ภาพที่ 4.30 ลักษณะการแจกแจงค่าความสามารถผู้สอบของข้อมูลจำลองชุดที่ 6 (253PL20)



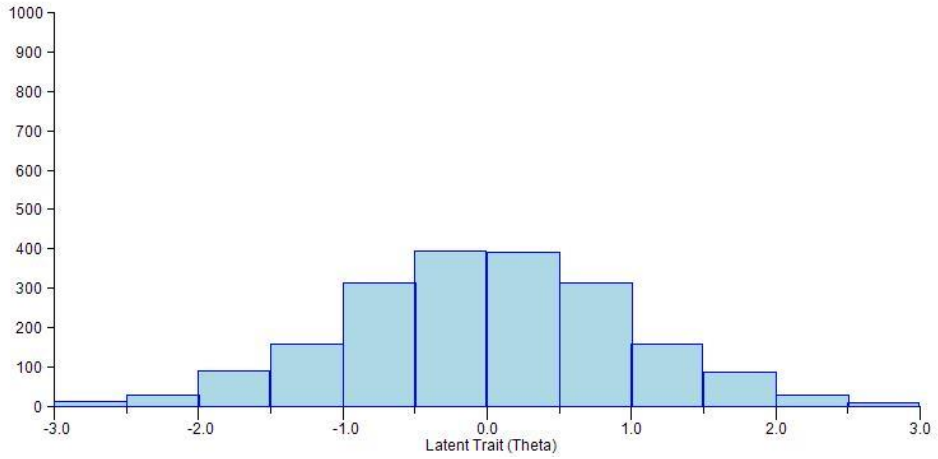
ภาพที่ 4.31 ลักษณะการแจกแจงค่าความสามารถผู้สอบของข้อมูลจำลองชุดที่ 7 (501PL10)



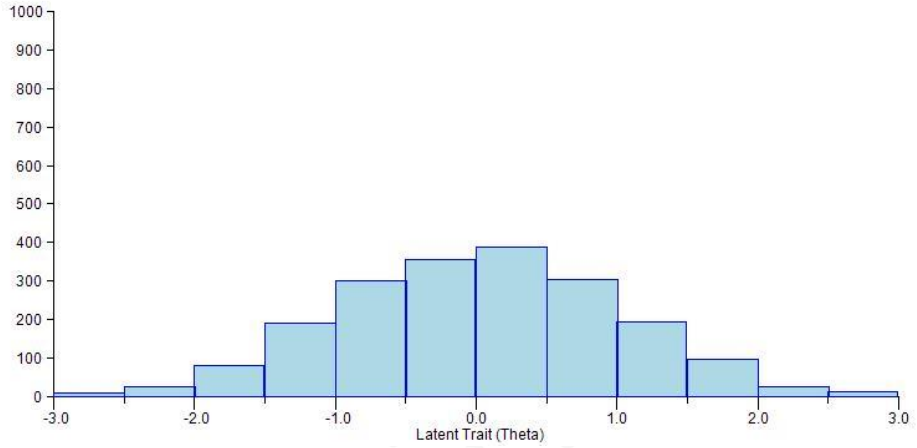
ภาพที่ 4.32 ลักษณะการแจกแจงค่าความสามารถผู้สอบของข้อมูลจำลองชุดที่ 8 (501PL20)



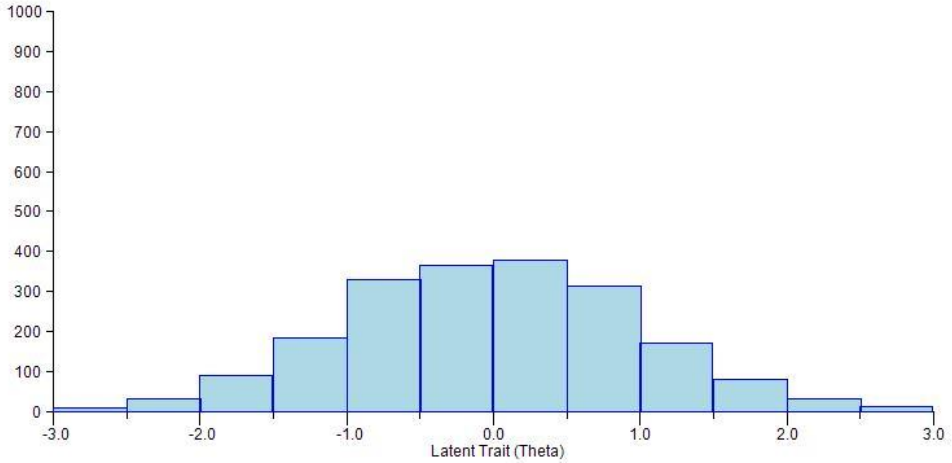
ภาพที่ 4.33 ลักษณะการแจกแจงค่าความสามารถผู้สอบของข้อมูลจำลองชุดที่ 9 (502PL10)



ภาพที่ 4.34 ลักษณะการแจกแจงค่าความสามารถผู้สอบของข้อมูลจำลองชุดที่ 10 (502PL20)



ภาพที่ 4.35 ลักษณะการแจกแจงค่าความสามารถผู้สอบของข้อมูลจำลองชุดที่ 11 (503PL10)



ภาพที่ 4.36 ลักษณะการแจกแจงค่าความสามารถผู้สอบของข้อมูลจำลองชุดที่ 12 (503PL20)

1.2 ผลการวิเคราะห์ค่าดัชนีการจำแนกประเภทด้วยวิธีการประมาณค่าตามทฤษฎีการตอบสนองข้อสอบสามวิธีจากการศึกษาการจำลองข้อมูล (simulation study) ภายใต้สถานการณ์เงื่อนไข

ผลการวิเคราะห์ในส่วนนี้ประกอบด้วยค่าดัชนีความถูกต้อง (accuracy) และค่าดัชนีความสอดคล้อง (consistency) ของการจำแนกประเภทเฉลี่ยจากการทำซ้ำจำนวน 100 รอบ โดยแบ่งการนำเสนอตามวิธีที่ใช้ในการประมาณค่า ผลการวิเคราะห์มีรายละเอียดดังนี้

วิธีการของ Rudner

ดัชนีความถูกต้องของการจำแนกประเภทในภาพรวม พบว่ามีค่าเฉลี่ยอยู่ในช่วง 0.8234-0.9086 โดยสถานการณ์ที่ 4 (252PL20) มีค่าดัชนีความถูกต้องต่ำสุดและสถานการณ์ที่ 6 (253PL20) มีค่าดัชนีความถูกต้องสูงสุด

ดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวม พบว่ามีค่าเฉลี่ยอยู่ในช่วง 0.7550-0.8749 โดยสถานการณ์ที่ 4 (252PL20) มีค่าดัชนีความสอดคล้องต่ำสุดและสถานการณ์ที่ 6 (253PL20) มีค่าดัชนีความสอดคล้องสูงสุด

เมื่อพิจารณาค่าดัชนีการจำแนกประเภททั้งสองพบว่าค่าดัชนีความถูกต้องมีค่าสูงกว่าค่าดัชนีความสอดคล้องในทุกสถานการณ์ โดยรายละเอียดของค่าเฉลี่ยดัชนีการจำแนกประเภทจากการทำซ้ำ 100 รอบ ที่ประมาณค่าได้จากวิธีการของ Rudner แสดงได้ดังตารางที่ 4.9 และภาพที่ 4.37

วิธีการของ Guo

ดัชนีความถูกต้องของการจำแนกประเภทในภาพรวม พบว่ามีค่าเฉลี่ยอยู่ในช่วง 0.9987-1.0000 โดยสถานการณ์ที่ 6 (253PL20) มีค่าดัชนีความถูกต้องต่ำสุดและสถานการณ์ที่ 7 (501PL10) มีค่าดัชนีความถูกต้องสูงสุด

ดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวม พบว่ามีค่าเฉลี่ยอยู่ในช่วง 0.9982-1.0000 โดยสถานการณ์ที่ 6 (253PL20) มีค่าดัชนีความสอดคล้องต่ำสุดและสถานการณ์ที่ 7 (501PL10) มีค่าดัชนีความสอดคล้องสูงสุด

เมื่อพิจารณาค่าดัชนีการจำแนกประเภททั้งสองพบว่าค่าดัชนีความถูกต้องมีค่าสูงกว่าค่าดัชนีความสอดคล้องในทุกสถานการณ์ โดยรายละเอียดของค่าเฉลี่ยดัชนีการจำแนกประเภทจากการทำซ้ำ 100 รอบ ที่ประมาณค่าได้จากวิธีการของ Guo แสดงได้ดังตารางที่ 4.9 และภาพที่ 4.38

วิธีการของ Lee

ดัชนีความถูกต้องของการจำแนกประเภทในภาพรวม พบว่ามีค่าเฉลี่ยอยู่ในช่วง 0.6285-0.7496 โดยสถานการณ์ที่ 12 (503PL20) มีค่าดัชนีความถูกต้องต่ำสุดและสถานการณ์ที่ 1 (251PL10) มีค่าดัชนีความถูกต้องสูงสุด

ดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวม พบว่ามีค่าเฉลี่ยอยู่ในช่วง 0.5372-0.6938 โดยสถานการณ์ที่ 12 (503PL20) มีค่าดัชนีความสอดคล้องต่ำสุดและสถานการณ์ที่ 1 (251PL10) มีค่าดัชนีความสอดคล้องสูงสุด

เมื่อพิจารณาค่าดัชนีการจำแนกประเภททั้งสองพบว่าค่าดัชนีความถูกต้องมีค่าสูงกว่าค่าดัชนีความสอดคล้องในทุกสถานการณ์ โดยรายละเอียดของค่าเฉลี่ยดัชนีการจำแนกประเภทจากการทำซ้ำ 100 รอบ ที่ประมาณค่าได้จากวิธีการของ Lee แสดงได้ดังตารางที่ 4.9 และภาพที่ 4.39

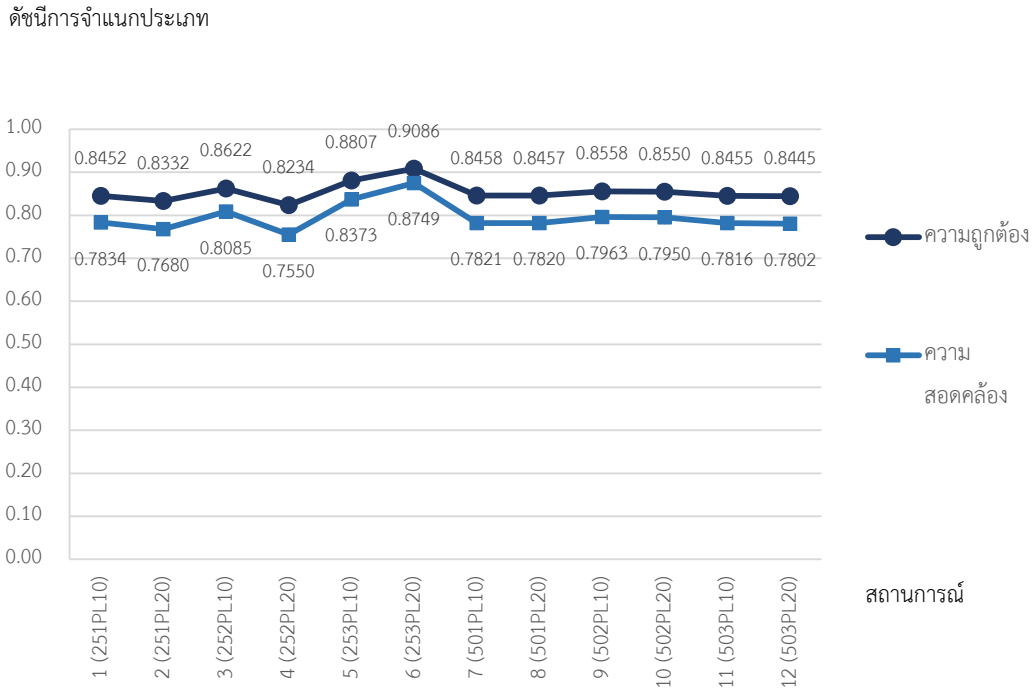
ตารางที่ 4.9 ค่าเฉลี่ยดัชนีความถูกต้องและความสอดคล้องของการจำแนกประเภทจากการทำซ้ำจำนวน 100 รอบ ในสถานการณ์จำลองจำแนกตามวิธีการประมาณค่า

สถานการณ์	ตำแหน่งคะแนนจุดตัด	ความถูกต้อง (accuracy)			ความสอดคล้อง (consistency)		
		Rudner	Guo	Lee	Rudner	Guo	Lee
251PL10	1	0.9274	1.0000	0.9543	0.8965	1.0000	0.9349
	2	0.9257	1.0000	0.9549	0.8978	1.0000	0.9357
	3	0.9921	0.9999	0.9555	0.9869	0.9998	0.9366
	4	0.9999	1.0000	0.9557	0.9998	1.0000	0.9369
	5	1.0000	1.0000	0.9554	1.0000	1.0000	0.9365
	6	1.0000	1.0000	0.9537	1.0000	1.0000	0.9340
	7	1.0000	1.0000	0.9519	1.0000	1.0000	0.9314
	ทั้งหมด	0.8452	0.9999	0.7496	0.7834	0.9998	0.6938
251PL20	1	0.9128	1.0000	0.9509	0.8772	1.0000	0.9297
	2	0.9415	1.0000	0.9489	0.9177	1.0000	0.9267
	3	0.9797	1.0000	0.9446	0.9715	1.0000	0.9208
	4	0.9990	0.9999	0.9411	0.9983	0.9999	0.9157
	5	1.0000	1.0000	0.9382	1.0000	1.0000	0.9116
	6	1.0000	1.0000	0.9432	1.0000	1.0000	0.9184
	7	1.0000	1.0000	0.9603	1.0000	1.0000	0.9425
	ทั้งหมด	0.8332	0.9999	0.6760	0.7680	0.9999	0.5892
252PL10	1	0.9106	1.0000	0.9586	0.8743	1.0000	0.9429
	2	0.9591	0.9998	0.9569	0.9427	0.9997	0.9404
	3	0.9932	0.9995	0.9546	0.9902	0.9994	0.9360
	4	0.9991	0.9997	0.9503	0.9985	0.9996	0.9302

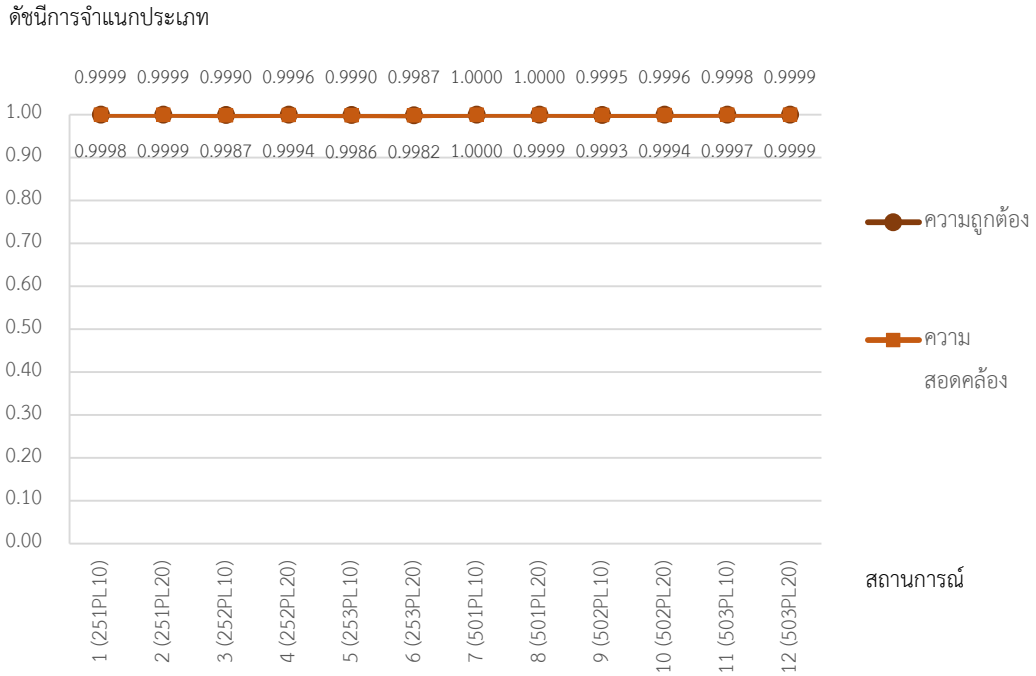
สถานการณ์	ตำแหน่งคะแนนจุดตัด	ความถูกต้อง (accuracy)			ความสอดคล้อง (consistency)		
		Rudner	Guo	Lee	Rudner	Guo	Lee
	5	1.0000	1.0000	0.9450	1.0000	1.0000	0.9241
	6	1.0000	1.0000	0.9426	1.0000	1.0000	0.9190
	7	1.0000	1.0000	0.9311	1.0000	1.0000	0.9209
	ทั้งหมด	0.8622	0.9990	0.6729	0.8085	0.9987	0.6119
252PL20	1	0.9176	1.0000	0.9429	0.8837	1.0000	0.9212
	2	0.9325	1.0000	0.9427	0.9056	1.0000	0.9202
	3	0.9768	0.9999	0.9425	0.9667	0.9998	0.9192
	4	0.9961	0.9997	0.9427	0.9943	0.9996	0.9195
	5	1.0000	1.0000	0.9426	1.0000	1.0000	0.9207
	6	1.0000	1.0000	0.9476	1.0000	1.0000	0.9265
	7	1.0000	1.0000	0.9485	1.0000	1.0000	0.9413
	ทั้งหมด	0.8234	0.9996	0.6633	0.7550	0.9994	0.6065
253PL10	1	0.9331	1.0000	0.9431	0.9054	1.0000	0.9215
	2	0.9641	0.9999	0.9427	0.9493	0.9998	0.9202
	3	0.9864	0.9996	0.9420	0.9804	0.9995	0.9186
	4	0.9949	0.9997	0.9421	0.9926	0.9996	0.9189
	5	0.9989	0.9998	0.9428	0.9983	0.9997	0.9207
	6	0.9999	0.9999	0.9477	0.9999	0.9999	0.9268
	7	1.0000	1.0000	0.9490	1.0000	1.0000	0.9417
	ทั้งหมด	0.8807	0.9990	0.6634	0.8373	0.9986	0.6067
253PL20	1	0.9366	0.9999	0.9428	0.9108	0.9999	0.9212
	2	0.9791	0.9997	0.9424	0.9700	0.9996	0.9199
	3	0.9928	0.9993	0.9423	0.9896	0.9990	0.9190
	4	0.9977	0.9998	0.9426	0.9964	0.9997	0.9195
	5	0.9996	0.9999	0.9427	0.9994	0.9998	0.9206
	6	1.0000	1.0000	0.9472	0.9999	1.0000	0.9260
	7	1.0000	1.0000	0.9489	1.0000	1.0000	0.9416
	ทั้งหมด	0.9086	0.9987	0.6630	0.8749	0.9982	0.6059

สถานการณ์	ตำแหน่งคะแนนจุดตัด	ความถูกต้อง (accuracy)			ความสอดคล้อง (consistency)		
		Rudner	Guo	Lee	Rudner	Guo	Lee
501PL10	1	0.9878	1.0000	0.9636	0.9827	1.0000	0.9481
	2	0.9585	1.0000	0.9580	0.9413	1.0000	0.9403
	3	0.9504	1.0000	0.9555	0.9297	1.0000	0.9367
	4	0.9601	1.0000	0.9554	0.9436	1.0000	0.9366
	5	0.9896	1.0000	0.9561	0.9851	1.0000	0.9377
	6	0.9994	1.0000	0.9578	0.9991	1.0000	0.9399
	7	1.0000	1.0000	0.9557	1.0000	1.0000	0.9429
	ทั้งหมด	0.8458	1.0000	0.7149	0.7821	1.0000	0.6367
501PL20	1	0.9916	1.0000	0.9557	0.9879	1.0000	0.9368
	2	0.9558	1.0000	0.9521	0.9376	1.0000	0.9319
	3	0.9475	1.0000	0.9528	0.9257	1.0000	0.9328
	4	0.9660	1.0000	0.9563	0.9519	1.0000	0.9377
	5	0.9857	1.0000	0.9600	0.9800	1.0000	0.9431
	6	0.9992	1.0000	0.9655	0.9986	0.9999	0.9509
	7	1.0000	1.0000	0.9700	1.0000	1.0000	0.9573
	ทั้งหมด	0.8457	1.0000	0.7241	0.7820	0.9999	0.6385
502PL10	1	0.9859	1.0000	0.9636	0.9801	1.0000	0.9493
	2	0.9582	1.0000	0.9550	0.9409	1.0000	0.9377
	3	0.9507	1.0000	0.9540	0.9302	1.0000	0.9351
	4	0.9659	1.0000	0.9582	0.9516	1.0000	0.9414
	5	0.9955	0.9997	0.9617	0.9935	0.9995	0.9459
	6	0.9997	0.9999	0.9648	0.9996	0.9999	0.9505
	7	1.0000	1.0000	0.9667	1.0000	1.0000	0.9545
	ทั้งหมด	0.8558	0.9995	0.7313	0.7963	0.9993	0.6492
502PL20	1	0.9861	0.9999	0.9597	0.9802	0.9999	0.9438
	2	0.9583	1.0000	0.9556	0.9411	1.0000	0.9379
	3	0.9516	1.0000	0.9547	0.9316	1.0000	0.9361
	4	0.9662	1.0000	0.9565	0.9520	1.0000	0.9390

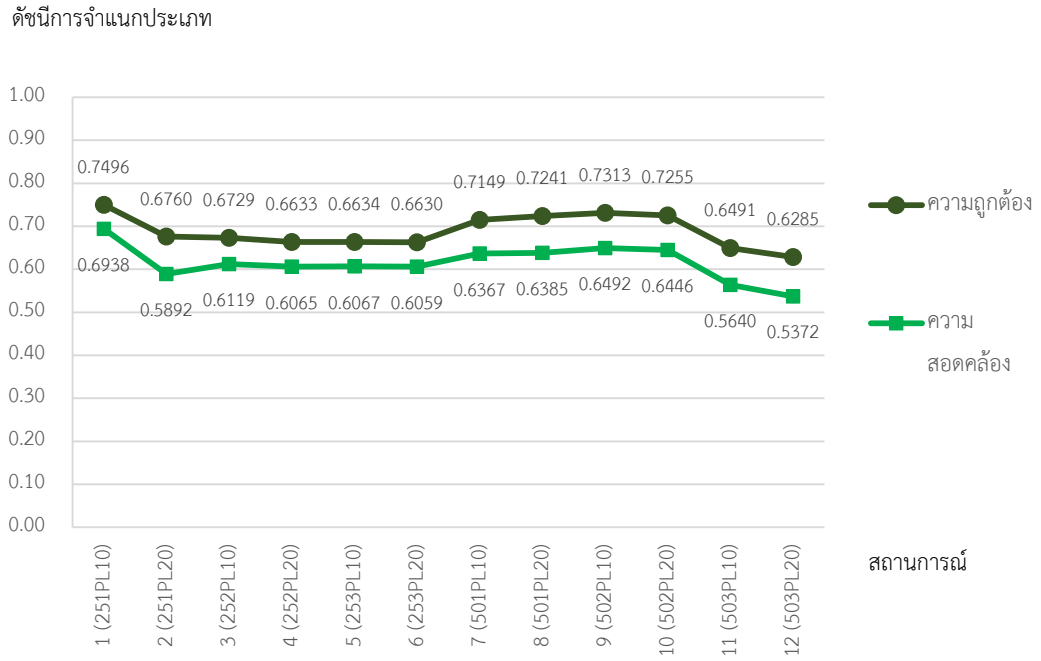
สถานการณ์	ตำแหน่งคะแนนจุดตัด	ความถูกต้อง (accuracy)			ความสอดคล้อง (consistency)		
		Rudner	Guo	Lee	Rudner	Guo	Lee
	5	0.9933	0.9997	0.9594	0.9904	0.9996	0.9427
	6	0.9995	1.0000	0.9635	0.9992	0.9999	0.9489
	7	1.0000	1.0000	0.9671	1.0000	1.0000	0.9549
	ทั้งหมด	0.8550	0.9996	0.7255	0.7950	0.9994	0.6446
503PL10	1	0.9970	1.0000	0.9706	0.9957	1.0000	0.9589
	2	0.9563	1.0000	0.9583	0.9381	1.0000	0.9416
	3	0.9431	1.0000	0.9504	0.9197	1.0000	0.9300
	4	0.9589	1.0000	0.9397	0.9418	1.0000	0.9152
	5	0.9903	0.9998	0.9348	0.9860	0.9997	0.9082
	6	1.0000	1.0000	0.9334	0.9999	1.0000	0.9066
	7	1.0000	1.0000	0.9357	1.0000	1.0000	0.9115
	ทั้งหมด	0.8455	0.9998	0.6491	0.7816	0.9997	0.5640
503PL20	1	0.9975	1.0000	0.9614	0.9961	1.0000	0.9460
	2	0.9579	1.0000	0.9476	0.9401	1.0000	0.9266
	3	0.9417	1.0000	0.9376	0.9182	1.0000	0.9122
	4	0.9632	1.0000	0.9300	0.9477	1.0000	0.9021
	5	0.9843	1.0000	0.9342	0.9780	1.0000	0.9072
	6	0.9999	1.0000	0.9417	0.9998	0.9999	0.9179
	7	1.0000	1.0000	0.9480	1.0000	1.0000	0.9282
	ทั้งหมด	0.8445	0.9999	0.6285	0.7802	0.9999	0.5372



ภาพที่ 4.37 ค่าเฉลี่ยดัชนีการจำแนกประเภทของข้อมูลจำลองจากการทำซ้ำ 100 รอบ โดยใช้วิธีการของ Rudner



ภาพที่ 4.38 ค่าเฉลี่ยดัชนีการจำแนกประเภทของข้อมูลจำลองจากการทำซ้ำ 100 รอบ โดยใช้วิธีการของ Guo



ภาพที่ 4.39 ค่าเฉลี่ยดัชนีการจำแนกประเภทของข้อมูลจำลองจากการทำซ้ำ 100 รอบ โดยใช้วิธีการของ Lee

1.3 ผลการทดสอบขนาดอิทธิพลของปัจจัยที่ส่งผลต่อค่าเฉลี่ยดัชนีการจำแนกประเภทของวิธีการประมาณค่าตามทฤษฎีการตอบสนองข้อสอบสามวิธีจากการศึกษาการจำลองข้อมูล (simulation study)

ผลการวิเคราะห์ในส่วนนี้เป็นผลการเปรียบเทียบค่าเฉลี่ยดัชนีการจำแนกประเภทของวิธีการประมาณค่าทั้งสามวิธีว่าแต่ละสถานการณ์แตกต่างกันเนื่องจากอิทธิพลของปัจจัยใด โดยใช้การวิเคราะห์ความแปรปรวนแบบสามทาง (3-WAY ANOVA) ซึ่งในการจำลองข้อมูลครั้งนี้ประกอบด้วย 3 ปัจจัย คือ ความยาวของแบบสอบ โมเดลการวัด และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ

ในการวิเคราะห์ความแปรปรวนนั้นค่าขนาดอิทธิพล (effect size) ใช้ค่า Eta square คำนวณได้จากผลรวมรากที่สองของค่าเฉลี่ยภายในกลุ่ม (the within-groups sum of squares) หารด้วยผลรวมรากที่สองของค่าเฉลี่ยทั้งหมด (the total sum of squares) โดยขนาดอิทธิพลมีค่าอยู่ระหว่าง 0 ถึง 1 และแปลความหมายของแต่ละช่วงคะแนนดังนี้ (Muijs, 2004)

ขนาดอิทธิพล (effect size)	การแปลความหมาย
0-0.10	อิทธิพลระดับน้อย (weak effect)
0.11-0.30	อิทธิพลระดับพอประมาณ (modest effect)
0.31-0.50	อิทธิพลระดับปานกลาง (moderate effect)
> 0.50	อิทธิพลระดับมาก (strong effect)

ผลการวิเคราะห์สามารถแบ่งการนำเสนอตามวิธีการประมาณค่าได้ดังนี้

วิธีการของ Rudner

เมื่อพิจารณาปัจจัยที่ส่งผลต่อดัชนีความถูกต้องของการจำแนกประเภท (classification accuracy) พบว่า ปัจจัยทั้งด้านความยาวของแบบสอบ โมเดลการวัด และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ มีอิทธิพลร่วมกันต่อค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภท ซึ่งทำให้ค่าเฉลี่ยของดัชนีความถูกต้องของการจำแนกประเภทแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 (sig=.000) โดยมีอิทธิพลอยู่ในระดับมากด้วยขนาดอิทธิพล .602

ในส่วนของดัชนีความสอดคล้องของการจำแนกประเภท (classification consistency) พบว่า ปัจจัยทั้งด้านความยาวของแบบสอบ โมเดลการวัด และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ มีอิทธิพลร่วมกันต่อค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภท ซึ่งทำให้ค่าเฉลี่ยของดัชนีความถูกต้องของการจำแนกประเภทแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 (sig=.000) โดยมีอิทธิพลอยู่ในระดับมากด้วยขนาดอิทธิพล .610

รายละเอียดของผลการวิเคราะห์แสดงได้ดังตารางที่ 4.10

ตารางที่ 4.10 ผลการวิเคราะห์ความแปรปรวนแบบสามทาง (3-WAY ANOVA) ของค่าเฉลี่ยดัชนีความถูกต้องและดัชนีความสอดคล้องของการจำแนกประเภทจากการทำซ้ำจำนวน 100 รอบ ในสถานการณ์จำลองโดยใช้วิธีการของ Rudner

แหล่งของความแปรปรวน	SS	df	Mean Square	F	p-value	Eta Squared	Post Hoc
ดัชนีความถูกต้อง							
Corrected Model	.551[a]	11	.050	1580.092*	.000	.936	
Intercept	874.773	1	874.773	27616547.045*	.000	1.000	
length	.031	1	.031	981.937*	.000	.453	25 > 50(.000)
model	.163	2	.081	2569.674*	.000	.812	3PL > 2PL(.000) 2PL > 1PL(.000)
misfit	.005	1	.005	162.736*	.000	.120	10% > 20%(.000)
length * model	.235	2	.118	3711.695*	.000	.862	
length * misfit	.004	1	.004	117.443*	.000	.090	

แหล่งของความแปรปรวน	SS	df	Mean Square	F	p-value	Eta Squared	Post Hoc
model * misfit	.056	2	.028	880.798*	.000	.597	
length * model * misfit	.057	2	.028	897.279*	.000	.602	
Error	.038	1188	.00003				
Total	875.361	1200					
Corrected Total	.588	1199					
a. R Squared = .936 (Adjusted R Squared = .935)							
ดัชนีความสอดคล้อง							
Corrected Model	1.156[a]	11	.105	1852.380*	.000	.945	
Intercept	759.139	1	759.139	13382627.423*	.000	1.000	
length	.101	1	.101	1777.070*	.000	.599	25 > 50(.000)
model	.341	2	.170	3004.300*	.000	.835	3PL > 2PL(.000) 2PL > 1PL(.000)
misfit	.010	1	.010	168.698*	.000	.124	10% > 20%(.000)
length * model	.488	2	.244	4304.848*	.000	.879	
length * misfit	.007	1	.007	120.300*	.000	.092	
model * misfit	.104	2	.052	918.672*	.000	.607	
length * model * misfit	.105	2	.053	927.236*	.000	.610	
Error	.067	1188	.00006				
Total	760.363	1200					
Corrected Total	1.223	1199					

a. R Squared = .945 (Adjusted R Squared = .944)

หมายเหตุ : *p < .05

วิธีการของ Guo

เมื่อพิจารณาปัจจัยที่ส่งผลต่อดัชนีความถูกต้องของการจำแนกประเภท (classification accuracy) พบว่า ปัจจัยทั้งด้านความยาวของแบบสอบ โมเดลการวัด และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ มีอิทธิพลร่วมกันต่อค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภท ซึ่งทำให้ค่าเฉลี่ยของดัชนีความถูกต้องของการจำแนกประเภทแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 (sig=.000) โดยมีอิทธิพลอยู่ในระดับพอประมาณด้วยขนาดอิทธิพล .129

ในส่วนของดัชนีความสอดคล้องของการจำแนกประเภท (classification consistency) พบว่า ปัจจัยทั้งด้านความยาวของแบบสอบ โมเดลการวัด และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ มีอิทธิพลร่วมกันต่อค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภท ซึ่งทำให้ค่าเฉลี่ย

ของดัชนีความถูกต้องของการจำแนกประเภทแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 (sig=.000) โดยมีอิทธิพลอยู่ในระดับน้อยด้วยขนาดอิทธิพล .100

รายละเอียดของผลการวิเคราะห์แสดงได้ดังตารางที่ 4.11

ตารางที่ 4.11 ผลการวิเคราะห์ความแปรปรวนแบบสามทาง (3-WAY ANOVA) ของค่าเฉลี่ยดัชนีความถูกต้องและดัชนีความสอดคล้องของการจำแนกประเภทจากการทำซ้ำจำนวน 100 รอบ ในสถานการณ์จำลองโดยใช้วิธีการของ Guo

แหล่งของความแปรปรวน	SS	df	Mean Square	F	p-value	Eta Squared	Post Hoc
ดัชนีความถูกต้อง							
Corrected Model	.000[a]	11	.00001	285.979*	.000	.726	
Intercept	1198.991	1	1198.991	17264939929.451*	.000	1.000	
length	.00006	1	.00006	887.230*	.000	.428	50 > 25(.000)
model	.00008	2	.00004	591.974*	.000	.499	1PL > 2PL(.000) 2PL > 3PL(.000)
misfit	.000001	1	.000001	22.708*	.000	.019	20% > 10%(.000)
length * model	.00005	2	.00002	369.946*	.000	.384	
length * misfit	.00000004	1	.00000004	.616	.433	.001	
model * misfit	.000009	2	.000004	67.836*	.000	.102	
length * model * misfit	.00001	2	.000006	87.853*	.000	.129	
Error	.000008	1188	.00000007				
Total	1198.991	1200					
Corrected Total	.000	1199					
a. R Squared = .726 (Adjusted R Squared = .723)							
ดัชนีความสอดคล้อง							
Corrected Model	.000[a]	11	.00004	275.959*	.000	.719	
Intercept	1198.573	1	1198.573	9201939902.436*	.000	1.000	
length	.000	1	.000	829.220*	.000	.411	50 > 25(.000)
model	.000	2	.00008	619.318*	.000	.510	1PL > 2PL(.000) 2PL > 3PL(.000)
misfit	.000003	1	.000003	21.110*	.000	.017	20% > 10%(.000)
length * model	.00009	2	.00005	360.637*	.000	.378	
length * misfit	.00000009	1	.00000009	.728	.394	.001	
model * misfit	.00001	2	.000006	46.105*	.000	.072	
length * model * misfit	.00002	2	.000009	66.186*	.000	.100	
Error	.000	1188	.0000001				

แหล่งของความแปรปรวน	SS	df	Mean Square	F	p-value	Eta Squared	Post Hoc
Total	1198.574	1200					
Corrected Total	.001	1199					

a. R Squared = .719 (Adjusted R Squared = .716)

หมายเหตุ : *p < .05

วิธีการของ Lee

เมื่อพิจารณาปัจจัยที่ส่งผลต่อดัชนีความถูกต้องของการจำแนกประเภท (classification accuracy) พบว่า ปัจจัยทั้งด้านความยาวของแบบสอบ โมเดลการวัด และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ มีอิทธิพลร่วมกันต่อค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภท ซึ่งทำให้ค่าเฉลี่ยของดัชนีความถูกต้องของการจำแนกประเภทแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 (sig=.000) โดยมีอิทธิพลอยู่ในระดับมากด้วยขนาดอิทธิพล .821

ในส่วนของดัชนีความสอดคล้องของการจำแนกประเภท (classification consistency) พบว่า ปัจจัยทั้งด้านความยาวของแบบสอบ โมเดลการวัด และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ มีอิทธิพลร่วมกันต่อค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภท ซึ่งทำให้ค่าเฉลี่ยของดัชนีความถูกต้องของการจำแนกประเภทแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 (sig=.000) โดยมีอิทธิพลอยู่ในระดับมากด้วยขนาดอิทธิพล .863

รายละเอียดของผลการวิเคราะห์แสดงได้ดังตารางที่ 4.12

ตารางที่ 4.12 ผลการวิเคราะห์ความแปรปรวนแบบสามทาง (3-WAY ANOVA) ของค่าเฉลี่ยดัชนีความถูกต้องและดัชนีความสอดคล้องของการจำแนกประเภทจากการทำซ้ำจำนวน 100 รอบ ในสถานการณ์จำลองโดยใช้วิธีการของ Lee

แหล่งของความแปรปรวน	SS	df	Mean Square	F	p-value	Eta Squared	Post Hoc
ดัชนีความถูกต้อง							
Corrected Model	1.636[a]	11	.149	5588.449*	.000	.981	
Intercept	568.800	1	568.800	21367125.383*	.000	1.000	
length	.061	1	.061	2277.139*	.000	.657	50 > 25(.000)
model	.906	2	.453	17007.779*	.000	.966	1PL > 2PL(.000) 2PL > 3PL(.000)
misfit	.085	1	.085	3193.052*	.000	.729	10% > 20%(.000)
length * model	.367	2	.184	6895.376*	.000	.921	
length * misfit	.037	1	.037	1380.116*	.000	.537	

แหล่งของความแปรปรวน	SS	df	Mean Square	F	p-value	Eta Squared	Post Hoc
model * misfit	.036	2	.018	675.741*	.000	.532	
length * model * misfit	.145	2	.073	2732.423*	.000	.821	
Error	.032	1188	.00002				
Total	570.468	1200					
Corrected Total	1.668	1199					
a. R Squared = .981 (Adjusted R Squared = .981)							
ดัชนีความสอดคล้อง							
Corrected Model	1.883[a]	11	.171	5238.811	.000	.980	
Intercept	454.379	1	454.379	13904487.154*	.000	1.000	
length	.016	1	.016	489.448	.000	.292	25 > 50(.000)
model	.843	2	.422	12904.831	.000	.956	1PL > 2PL(.000) 2PL > 3PL(.000)
misfit	.164	1	.164	5019.205	.000	.809	10% > 20%(.000)
length * model	.438	2	.219	6696.908	.000	.919	
length * misfit	.055	1	.055	1679.891	.000	.586	
model * misfit	.122	2	.061	1861.238	.000	.758	
length * model * misfit	.245	2	.123	3756.213	.000	.863	
Error	.039	1188	.00003				
Total	456.301	1200					
Corrected Total	1.922	1199					

a. R Squared = .980 (Adjusted R Squared = .980)

หมายเหตุ : *p < .05

เมื่อนำผลการทดสอบขนาดอิทธิพลของปัจจัยที่มีอิทธิพลต่อค่าเฉลี่ยดัชนีการจำแนกประเภทของวิธีการประมาณค่าตามทฤษฎีการตอบสนองข้อสอบสามวิธีจากการศึกษาการจำลองข้อมูล (simulation study) มาวาดกราฟเพื่อพิจารณาว่าในแต่ละปัจจัยนั้นเงื่อนไขใดมีค่าเฉลี่ยดัชนีการจำแนกประเภทสูงสุด ผลการวิเคราะห์ดังกล่าวสามารถแบ่งการนำเสนอตามวิธีการประมาณค่าได้ดังนี้

วิธีการของ Rudner จะมีค่าเฉลี่ยดัชนีความถูกต้องและความสอดคล้องของการจำแนกประเภทสูง เมื่อแบบสอบมีลักษณะเป็นแบบสอบสั้น (25 ข้อ) ภายใต้โมเดลการวัดแบบ 3PL และมีความไม่เหมาะสมของโมเดลการวัดกับข้อสอบร้อยละ 10

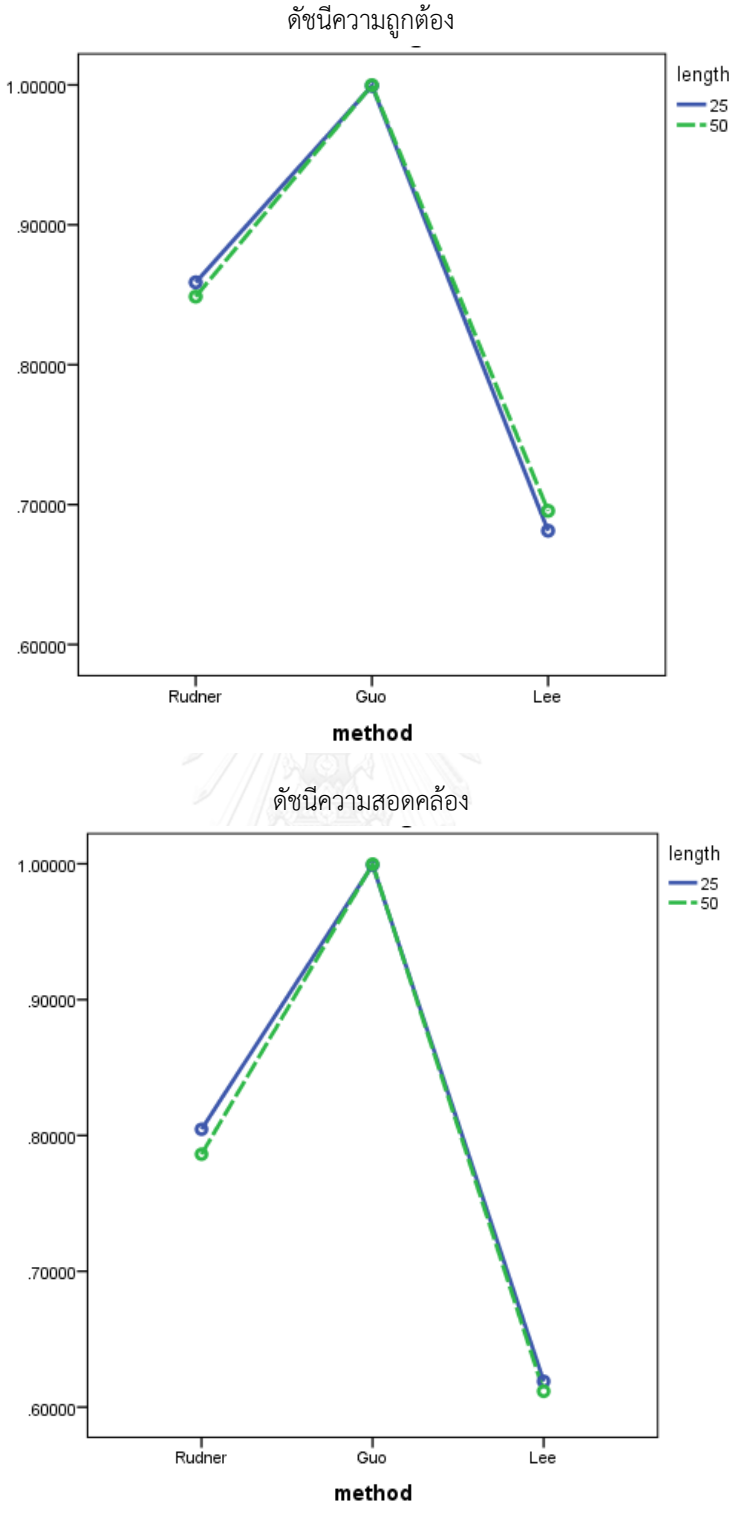
วิธีการของ Guo จะมีค่าเฉลี่ยดัชนีความถูกต้องและความสอดคล้องของการจำแนกประเภทสูง เมื่อแบบสอบมีลักษณะเป็นแบบสอบยาว (50 ข้อ) ภายใต้โมเดลการวัดแบบ 1PL และมีความไม่เหมาะสมของโมเดลการวัดกับข้อสอบร้อยละ 20

วิธีการของ Lee จะมีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทสูง เมื่อแบบสอบมีลักษณะเป็นแบบสอบยาว (50 ข้อ) ภายใต้โมเดลการวัดแบบ 1PL และมีความไม่เหมาะสมของโมเดลการวัดกับข้อสอบร้อยละ 10 และค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทสูง เมื่อแบบสอบมีลักษณะเป็นแบบสอบสั้น (25 ข้อ) ภายใต้โมเดลการวัดแบบ 1PL และมีความไม่เหมาะสมของโมเดลการวัดกับข้อสอบร้อยละ 10

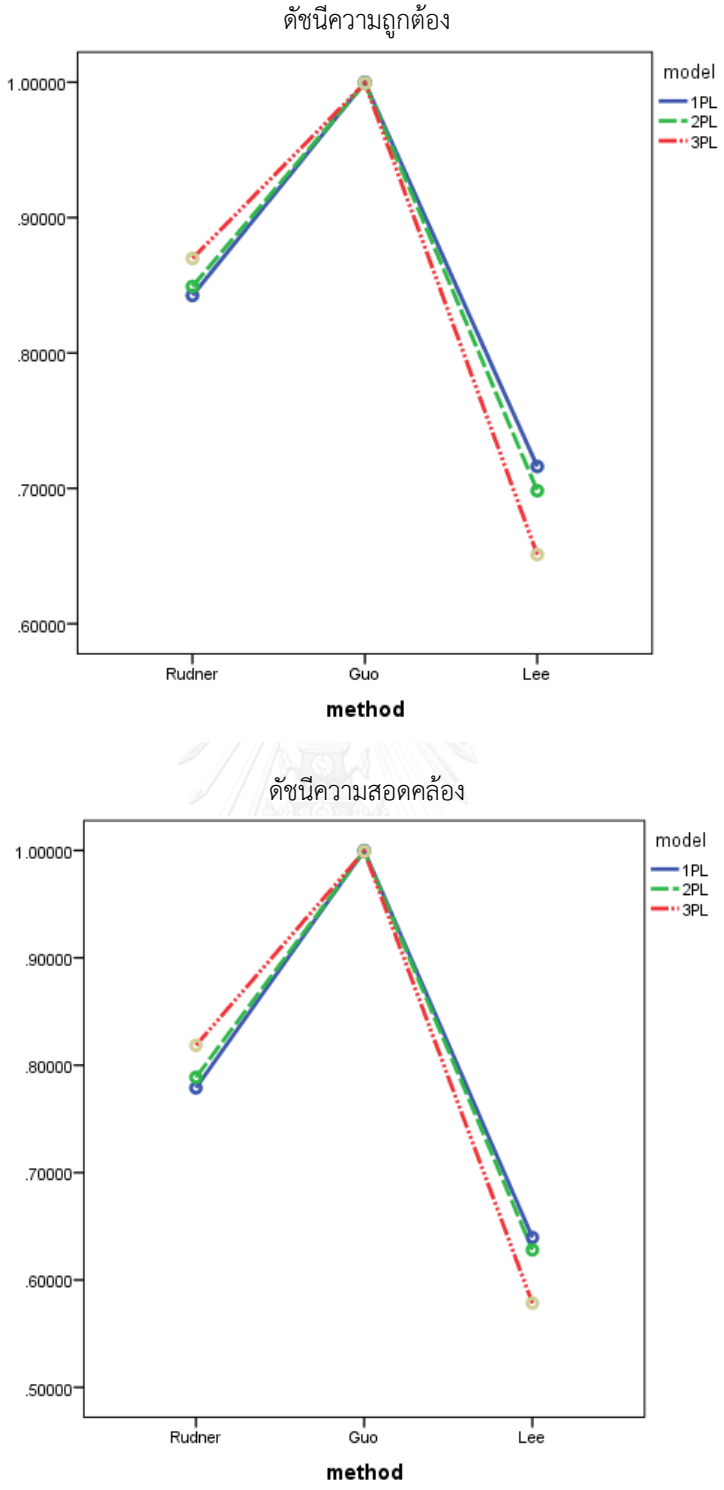
รายละเอียดของผลการวิเคราะห์แสดงได้ดังตารางที่ 4.13 และภาพที่ 4.40-4.42

ตารางที่ 4.13 ผลการเปรียบเทียบค่าเฉลี่ยดัชนีความถูกต้องและดัชนีความสอดคล้องของการจำแนกประเภทจากการทำซ้ำของปัจจัยที่มีอิทธิพลต่อค่าเฉลี่ยดัชนีการจำแนกประเภท จำแนกตามวิธีการประมาณค่า

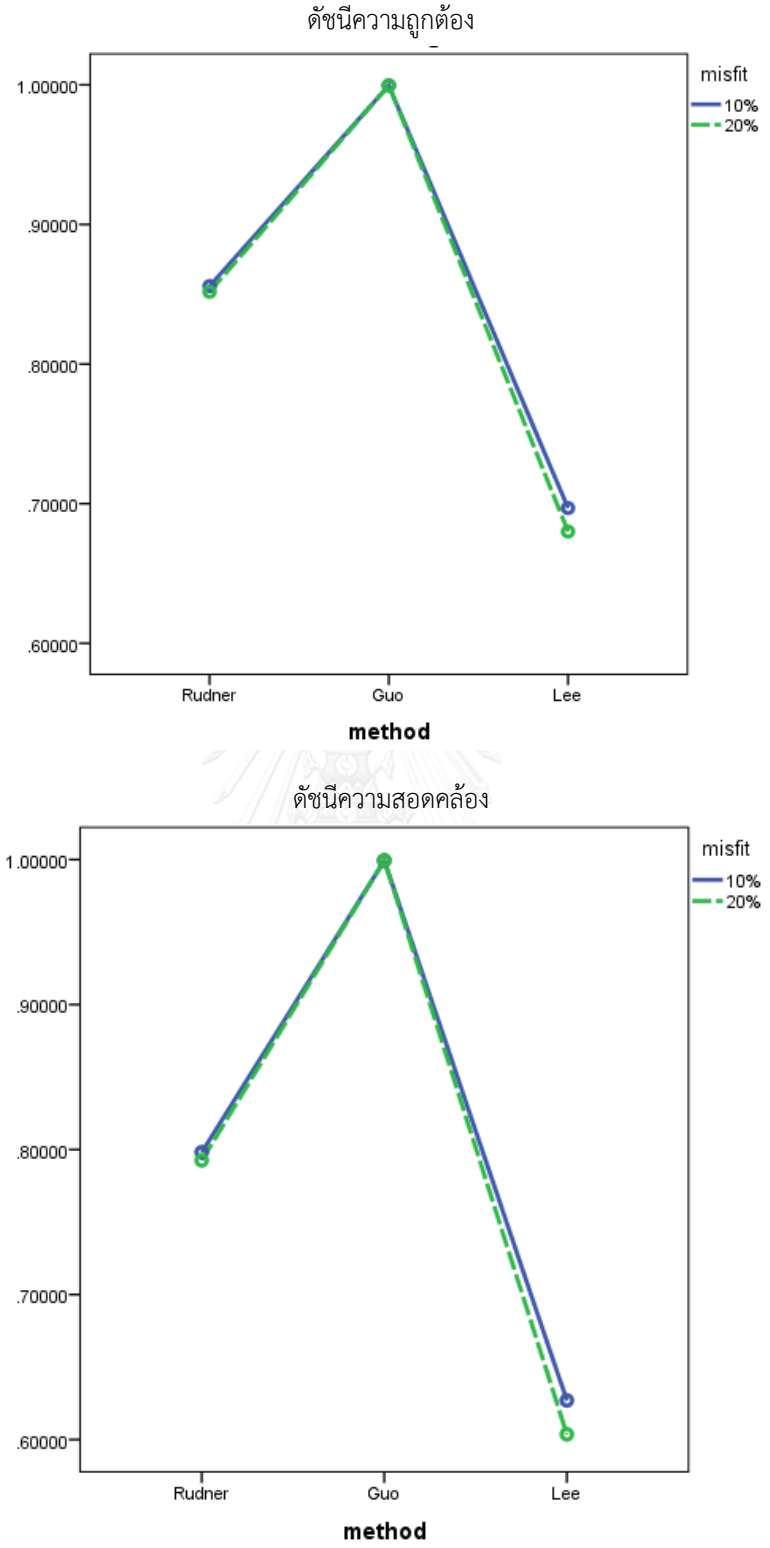
ปัจจัย	ค่าเฉลี่ยดัชนีความถูกต้อง			ค่าเฉลี่ยดัชนีความสอดคล้อง		
	Rudner	Guo	Lee	Rudner	Guo	Lee
1. ความยาวของแบบสอบ						
25 ข้อ	***0.859***	0.999	0.681	***0.805***	0.999	***0.619***
50 ข้อ	0.849	***1.000***	***0.696***	0.786	***1.000***	0.612
2. โมเดลการวัด						
1PL	0.842	***1.000***	***0.716***	0.779	***1.000***	***0.640***
2PL	0.849	0.999	0.698	0.789	0.999	0.628
3PL	***0.870***	0.999	0.651	***0.819***	0.999	0.578
3. ความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ						
ร้อยละ 10	***0.856***	0.999	***0.697***	***0.798***	0.999	***0.627***
ร้อยละ 20	0.852	***1.000***	0.680	0.793	***1.000***	0.604



ภาพที่ 4.40 กราฟเปรียบเทียบค่าเฉลี่ยดัชนีการจำแนกประเภทจากการทำซ้ำ 100 รอบ ในสถานการณ์จำลองจำแนกตามวิธีการประมาณค่า 3 วิธี เมื่อพิจารณาปัจจัยด้านความยาวของแบบสอบ



ภาพที่ 4.41 กราฟเปรียบเทียบค่าเฉลี่ยดัชนีการจำแนกประเภทจากการทำซ้ำ 100 รอบ ในสถานการณ์จำลองจำแนกตามวิธีการประมาณค่า 3 วิธี เมื่อพิจารณาปัจจัยด้านโมเดลการวัด



ภาพที่ 4.42 กราฟเปรียบเทียบค่าเฉลี่ยดัชนีการจำแนกประเภทจากการทำซ้ำ 100 รอบ ในสถานการณ์จำลองจำแนกตามวิธีการประมาณค่า 3 วิธี เมื่อพิจารณาปัจจัยด้านความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ

ตอนที่ 2 ผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทตามแนวคิดทฤษฎีการตอบสนองข้อสอบทั้งสามวิธีภายใต้การศึกษาการจำลองข้อมูล (simulation study)

ตอนนี้เป็นการนำเสนอผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทตามแนวคิดทฤษฎีการตอบสนองข้อสอบทั้งสามวิธี คือ วิธีการของ Rudner (2005) วิธีการของ Guo (2006) และวิธีการของ Lee (2010) ภายใต้การศึกษาการจำลองข้อมูล (simulation study) โดยใช้การวิเคราะห์ความแปรปรวน (ANOVA)

ผลการวิเคราะห์ความแตกต่างของค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทจากการทำซ้ำจำนวน 100 รอบ โดยใช้วิธีการประมาณค่าที่แตกต่างกันสามวิธีในการประมาณค่าดัชนีความถูกต้องของการจำแนกประเภท (classification accuracy) พบว่าค่าดัชนีความถูกต้องที่ได้จากการประมาณค่าด้วยสามวิธีการนั้น มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ทุกสถานการณ์เงื่อนไข โดยวิธีการที่มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทสูงที่สุดคือวิธีการของ Guo รองลงมาคือวิธีการของ Rudner และวิธีการของ Lee ตามลำดับ รายละเอียดของผลการวิเคราะห์แสดงได้ดังตารางที่ 4.14 และแผนภาพที่ 4.43

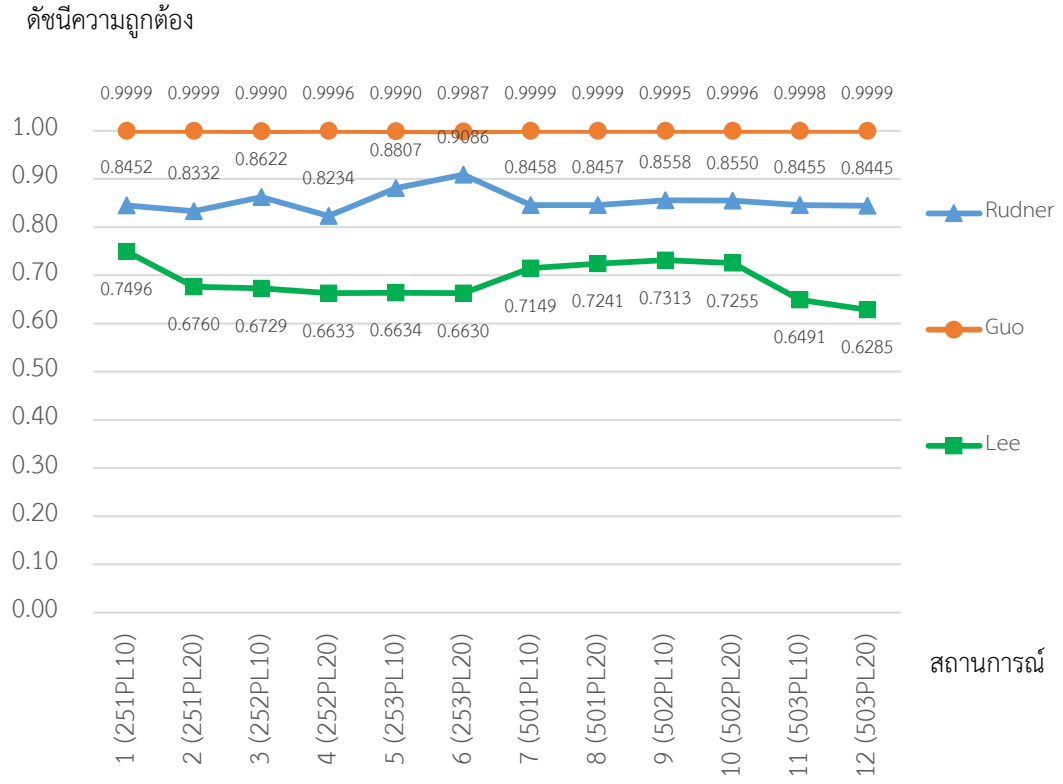
ในส่วนของค่าดัชนีความสอดคล้องที่ได้จากการประมาณค่าด้วยสามวิธีการนั้นก็เป็นไปในแนวทางเดียวกันคือมีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ทุกสถานการณ์เงื่อนไข โดยวิธีการที่มีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทสูงที่สุดคือวิธีการของ Guo รองลงมาคือวิธีการของ Rudner และวิธีการของ Lee ตามลำดับ รายละเอียดของผลการวิเคราะห์แสดงได้ดังตารางที่ 4.15 และแผนภาพที่ 4.44

ตารางที่ 4.14 ผลการวิเคราะห์ความแตกต่างของค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทจากการทำซ้ำจำนวน 100 รอบ ในสถานการณ์จำลองจำแนกตามวิธีการประมาณค่า

	Rudner		Guo		Lee				
สถานการณ์	Mean	SD	Mean	SD	Mean	SD	F	p-value	Post Hoc
251PL10	0.8452	0.0063	0.9999	0.0007	0.7496	0.0054	69,716.15*	.000	Guo>Rud(.000) Rud>Lee(.000)
251PL20	0.8332	0.0061	0.9999	0.0001	0.6760	0.0054	120,663.55*	.000	Guo>Rud(.000) Rud>Lee(.000)
252PL10	0.8622	0.0084	0.9990	0.0003	0.6729	0.0062	73,248.92*	.000	Guo>Rud(.000) Rud>Lee(.000)
252PL20	0.8234	0.0061	0.9996	0.0002	0.6633	0.0061	112,648.85*	.000	Guo>Rud(.000) Rud>Lee(.000)
253PL10	0.8807	0.0072	0.9990	0.0003	0.6634	0.0050	113,757.56*	.000	Guo>Rud(.000) Rud>Lee(.000)

สถานการณ์	Rudner		Guo		Lee		F	p-value	Post Hoc
	Mean	SD	Mean	SD	Mean	SD			
253PL20	0.9086	0.0062	0.9987	0.0003	0.6630	0.0052	135,966.17*	.000	Guo>Rud(.000) Rud>Lee(.000)
501PL10	0.8458	0.0041	0.9999	0.0000	0.7149	0.0043	169,732.85*	.000	Guo>Rud(.000) Rud>Lee(.000)
501PL20	0.8457	0.0051	0.9999	0.0001	0.7241	0.0043	131,036.75*	.000	Guo>Rud(.000) Rud>Lee(.000)
502PL10	0.8558	0.0038	0.9995	0.0002	0.7313	0.0051	132,099.53*	.000	Guo>Rud(.000) Rud>Lee(.000)
502PL20	0.8550	0.0039	0.9996	0.0002	0.7255	0.0045	156,448.81*	.000	Guo>Rud(.000) Rud>Lee(.000)
503PL10	0.8455	0.0039	0.9998	0.0002	0.6491	0.0050	228,875.89*	.000	Guo>Rud(.000) Rud>Lee(.000)
503PL20	0.8445	0.0041	0.9999	0.0001	0.6285	0.0048	258,629.21*	.000	Guo>Rud(.000) Rud>Lee(.000)

หมายเหตุ : *p < .05

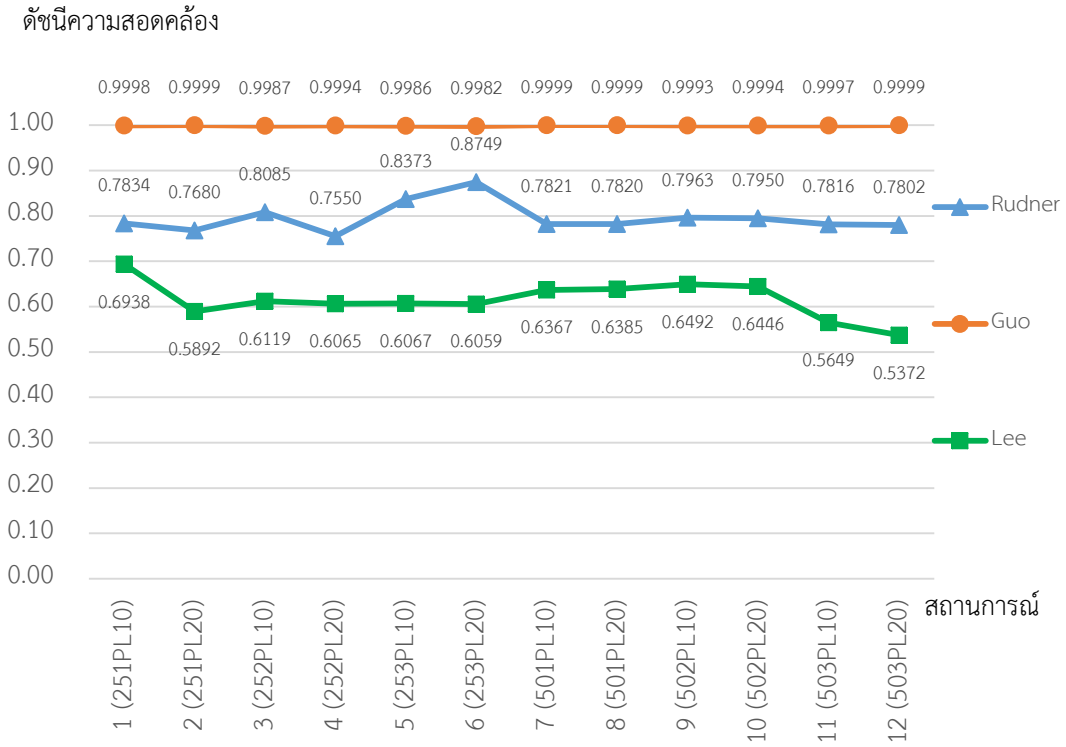


ภาพที่ 4.43 กราฟเปรียบเทียบค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทจากการทำซ้ำ 100 รอบ ในสถานการณ์จำลองด้วยวิธีการประมาณค่า 3 วิธี

ตารางที่ 4.15 ผลการวิเคราะห์ความแตกต่างของค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทจากการทำซ้ำจำนวน 100 รอบ ในสถานการณ์จำลองจำแนกตามวิธีการประมาณค่า

	Rudner		Guo		Lee				
สถานการณ์	Mean	SD	Mean	SD	Mean	SD	F	p-value	Post Hoc
251PL10	0.7834	0.0094	0.9998	0.0010	0.6938	0.0060	58,969.49*	.000	Guo>Rud(.000) Rud>Lee(.000)
251PL20	0.7680	0.0085	0.9999	0.0001	0.5892	0.0058	120,663.55*	.000	Guo>Rud(.000) Rud>Lee(.000)
252PL10	0.8085	0.0116	0.9987	0.0003	0.6119	0.0064	64,158.27*	.000	Guo>Rud(.000) Rud>Lee(.000)
252PL20	0.7550	0.0085	0.9994	0.0002	0.6065	0.0064	104,708.89*	.000	Guo>Rud(.000) Rud>Lee(.000)
253PL10	0.8373	0.0092	0.9986	0.0004	0.6067	0.0053	103,812.28*	.000	Guo>Rud(.000) Rud>Lee(.000)
253PL20	0.8749	0.0079	0.9982	0.0004	0.6059	0.0057	126,479.54*	.000	Guo>Rud(.000) Rud>Lee(.000)
501PL10	0.7821	0.0052	0.9999	0.0000	0.6367	0.0050	192,067.24*	.000	Guo>Rud(.000) Rud>Lee(.000)
501PL20	0.7820	0.0063	0.9999	0.0001	0.6385	0.0052	149,508.68*	.000	Guo>Rud(.000) Rud>Lee(.000)
502PL10	0.7963	0.0048	0.9993	0.0003	0.6492	0.0061	153,894.58*	.000	Guo>Rud(.000) Rud>Lee(.000)
502PL20	0.7950	0.0049	0.9994	0.0003	0.6446	0.0054	179,243.61*	.000	Guo>Rud(.000) Rud>Lee(.000)
503PL10	0.7816	0.0051	0.9997	0.0002	0.5649	0.0057	243,113.21*	.000	Guo>Rud(.000) Rud>Lee(.000)
503PL20	0.7802	0.0052	0.9999	0.0001	0.5372	0.0055	280,975.50*	.000	Guo>Rud(.000) Rud>Lee(.000)

หมายเหตุ : *p < .05



ภาพที่ 4.44 กราฟเปรียบเทียบค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทจากการทำซ้ำ 100 รอบ ในสถานการณ์จำลองด้วยวิธีการประมาณค่า 3 วิธี

เมื่อทำการเปรียบเทียบค่าเฉลี่ยดัชนีการจำแนกประเภทว่าแต่ละสถานการณ์แตกต่างกันเนื่องจากอิทธิพลของปัจจัยใด โดยใช้การวิเคราะห์ความแปรปรวนแบบสี่ทาง (4-WAY ANOVA) ซึ่งในการวิเคราะห์ครั้งนี้พิจารณาจาก 4 ปัจจัย คือ วิธีการประมาณค่า ความยาวของแบบสอบ โมเดลการวัด และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ มีรายละเอียดดังนี้

เมื่อพิจารณาปัจจัยที่ส่งผลต่อดัชนีความถูกต้องของการจำแนกประเภท (classification accuracy) พบว่า ปัจจัยทั้งด้านวิธีการประมาณค่า ความยาวของแบบสอบ โมเดลการวัด และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ มีอิทธิพลร่วมกันต่อค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภท ซึ่งทำให้ค่าเฉลี่ยของดัชนีความถูกต้องของการจำแนกประเภทแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 (sig=.000) โดยมีอิทธิพลอยู่ในระดับมากด้วยขนาดอิทธิพล .624

ในส่วนของดัชนีความสอดคล้องของการจำแนกประเภท (classification consistency) พบว่า ปัจจัยทั้งด้านวิธีการประมาณค่า ความยาวของแบบสอบ โมเดลการวัด และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ มีอิทธิพลร่วมกันต่อค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภท ซึ่งทำให้ค่าเฉลี่ยของดัชนีความถูกต้องของการจำแนกประเภทแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 (sig=.000) โดยมีอิทธิพลอยู่ในระดับมากด้วยขนาดอิทธิพล .656

ตารางที่ 4.16 ผลการวิเคราะห์ความแปรปรวนแบบสี่ทาง (4-WAY ANOVA) ของค่าเฉลี่ยดัชนีความถูกต้องและดัชนีความสอดคล้องของการจำแนกประเภทจากการทำซ้ำจำนวน 100 รอบ ในสถานการณ์จำลอง

แหล่งของความแปรปรวน	SS	df	Mean Square	F	p-value	Eta Squared	Post Hoc
ดัชนีความถูกต้อง							
Corrected Model	60.335[a]	35	1.724	88606.127*	.000	.999	
Intercept	2584.42	1	2584.42	2584.42*	.000	1.000	
method	58.147	2	29.074	1494395.787*	.000	.999	Guo>Rud(.000) Rud>Lee(.000)
length	.002	1	.002	103.427*	.000	.028	50 > 25(.000)
model	.103	2	.052	2648.706*	.000	.598	1PL > 2PL(.000) 2PL > 3PL(.000)
misfit	.044	1	.044	2246.333*	.000	.387	10% > 20%(.000)
method * length	.090	2	.045	2307.131*	.000	.564	
method * model	.965	4	.241	12404.401*	.000	.933	
method * misfit	.046	2	.023	1193.866*	.000	.401	
length * model	.360	2	.180	9243.024*	.000	.838	
length * misfit	.021	1	.021	1092.020*	.000	.235	
model * misfit	.030	2	.015	776.766*	.000	.304	
method * length * model	.243	4	.061	3118.168*	.000	.778	
method * length * misfit	.019	2	.010	493.799*	.000	.217	
method * model * misfit	.062	4	.015	791.074*	.000	.470	
length * model * misfit	.087	2	.044	2248.367*	.000	.558	
method * length * model * misfit	.115	4	.029	1475.797*	.000	.624	
Error	.069	3564	.00002				
Total	2644.82	3600					
Corrected Total	60.404	3599					
a. R Squared = .999 (Adjusted R Squared = .999)							
ดัชนีความสอดคล้อง							
Corrected Model	91.656[a]	35	2.619	87745.471*	.000	.999	
Intercept	2323.48	1	2323.48	77851771.724*	.000	1.000	
method	88.617	2	44.308	1484625.485*	.000	.999	Guo>Rud(.000) Rud>Lee(.000)
length	.063	1	.063	2099.610*	.000	.371	25 > 50(.000)
model	.040	2	.020	668.592*	.000	.273	1PL > 2PL(.000) 2PL > 3PL(.000)

แหล่งของความแปรปรวน	SS	df	Mean Square	F	p-value	Eta Squared	Post Hoc
misfit	.084	1	.084	2805.191*	.000	.440	10% > 20%(.000)
method * length	.054	2	.027	908.791*	.000	.338	
method * model	1.145	4	.286	9587.235*	.000	.915	
method * misfit	.090	2	.045	1505.657*	.000	.458	
length * model	.587	2	.293	9829.234*	.000	.847	
length * misfit	.033	1	.033	1119.521*	.000	.239	
model * misfit	.067	2	.034	1130.461*	.000	.388	
method * length * model	.339	4	.085	2843.640*	.000	.761	
method * length * misfit	.028	2	.014	474.265*	.000	.210	
method * model * misfit	.158	4	.040	1326.904*	.000	.598	
length * model * misfit	.148	2	.074	2474.307*	.000	.581	
method * length * model * misfit	.203	4	.051	1700.613*	.000	.656	
Error	.106	3564	.00003				
Total	2415.24	3600					
Corrected Total	91.763	3599					

a. R Squared = .999 (Adjusted R Squared = .999)

หมายเหตุ : *p < .05

เมื่อพิจารณาขนาดอิทธิพลของปัจจัยที่มีอิทธิพลต่อค่าเฉลี่ยดัชนีการจำแนกประเภท เพื่อเปรียบเทียบว่าปัจจัยที่กำหนดขึ้นมีอิทธิพลต่อค่าเฉลี่ยดัชนีการจำแนกประเภทที่ประมาณค่าได้จากวิธีการใดมากที่สุดและน้อยที่สุด อันจะนำไปสู่การอธิบายถึงความแข็งแกร่งของวิธีการประมาณค่าดัชนีการจำแนกประเภท โดยผลการวิเคราะห์ความแปรปรวนแบบสี่ทาง (4-WAY ANOVA) ซึ่งพิจารณาจาก 4 ปัจจัย คือ วิธีการประมาณค่า ความยาวของแบบสอบ โมเดลการวัด และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ พบว่า ปัจจัยทั้งสี่มีอิทธิพลร่วมกันต่อค่าเฉลี่ยดัชนีความถูกต้องและดัชนีความสอดคล้องของการจำแนกประเภทโดยมีอิทธิพลอยู่ในระดับมากด้วยขนาดอิทธิพล .624 และ .656 ตามลำดับ และผลการวิเคราะห์ความแปรปรวนแบบสามทาง (3-WAY ANOVA) ซึ่งพิจารณาจาก 3 ปัจจัย คือ ความยาวของแบบสอบ โมเดลการวัด และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ พบว่า ปัจจัยทั้งสามมีอิทธิพลร่วมกันต่อค่าเฉลี่ยดัชนีความถูกต้องและดัชนีความสอดคล้องของการจำแนกประเภทที่ประมาณค่าได้จากวิธีการของ Lee มากที่สุด ด้วยขนาดอิทธิพล .821 และ .863 ตามลำดับ รองลงมาคือวิธีการของ Rudner ด้วยขนาดอิทธิพล .602 และ .610 ตามลำดับ และมีอิทธิพลต่อวิธีการของ Guo น้อยที่สุดด้วยขนาดอิทธิพล .129 และ .100 ตามลำดับ

ซึ่งแสดงให้เห็นว่าวิธีการของ Guo มีความแข็งแกร่งมากที่สุด ส่วนวิธีการของ Lee มีความอ่อนไหวมากที่สุด รายละเอียดของผลการวิเคราะห์แสดงได้ดังตารางที่ 4.17

ตารางที่ 4.17 ผลการเปรียบเทียบขนาดอิทธิพลของปฏิสัมพันธ์ร่วมระหว่างปัจจัยที่มีต่อค่าเฉลี่ยดัชนีการจำแนกประเภทจำแนกตามวิธีการประมาณค่า

การวิเคราะห์ความแปรปรวน	ขนาดอิทธิพล	
	ความถูกต้อง	ความสอดคล้อง
1. การวิเคราะห์ความแปรปรวนแบบสี่ทาง (4-WAY ANOVA)		
ปฏิสัมพันธ์ร่วมระหว่างวิธีการประมาณค่า ความยาวของแบบสอบ โมเดลการวัด และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ (method * length * model * misfit)	.624	.656
2. การวิเคราะห์ความแปรปรวนแบบสามทาง (3-WAY ANOVA)		
ปฏิสัมพันธ์ร่วมระหว่างความยาวของแบบสอบ โมเดลการวัด และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ ต่อวิธีการของ Rudner (length * model * misfit)	.602	.610
ปฏิสัมพันธ์ร่วมระหว่างความยาวของแบบสอบ โมเดลการวัด และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ ต่อวิธีการของ Guo (length * model * misfit)	.129	.100
ปฏิสัมพันธ์ร่วมระหว่างความยาวของแบบสอบ โมเดลการวัด และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ ต่อวิธีการของ Lee (length * model * misfit)	.821	.863

ตอนที่ 3 ผลการประมาณค่าและผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบทั้งสามวิธี เมื่อนำไปใช้กับข้อมูลเชิงประจักษ์ (empirical data)

การนำเสนอผลการวิเคราะห์ประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนก ประเภทตามแนวคิดทฤษฎีการตอบสนองข้อสอบที่ได้จากการจำลองข้อมูล เมื่อนำไปใช้กับข้อมูลเชิงประจักษ์ (empirical data) หรือข้อมูลการตอบข้อสอบจริง (real data) ของนักเรียนชั้นมัธยมศึกษา ปีที่ 3 ในปีการศึกษา 2556 เป็นการนำเสนอผลการวิเคราะห์ข้อมูลเบื้องต้นในการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET) ของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ในปีการศึกษา 2556 ประกอบด้วยผลการวิเคราะห์ข้อมูล 2 ส่วน คือ ส่วนแรกเป็นผลการวิเคราะห์ค่าสถิติพื้นฐานของคะแนนผลการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET) (หรือคะแนนที่สังเกตได้) และส่วนที่สองเป็นผล

การวิเคราะห์ค่าสถิติพื้นฐานของค่าพารามิเตอร์ความสามารถของผู้สอบ (หรือคะแนนจริง) และค่าพารามิเตอร์ของข้อสอบ โดยในแต่ละส่วนมีรายละเอียดดังนี้

3.1 ผลการวิเคราะห์ค่าสถิติพื้นฐานของคะแนนผลการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET)

ผลการวิเคราะห์ค่าสถิติพื้นฐานที่นำเสนอในตอนนี้เป็นผลการวิเคราะห์ค่าสถิติพื้นฐานของคะแนนผลการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET) ของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ที่ทำการทดสอบในปีการศึกษา 2556 จำนวนสองรายวิชาหลัก คือวิชาคณิตศาสตร์ และวิชาภาษาไทย ซึ่งได้รับความอนุเคราะห์ข้อมูลจากสถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน) ทั้งนี้ชุดข้อสอบในแต่ละรายวิชาแบ่งออกเป็น 2 ชุด คือ ชุด A และ B ผลการวิเคราะห์ประกอบด้วย คะแนนมากที่สุด (Max) คะแนนน้อยที่สุด (Min) คะแนนเฉลี่ย (Mean) ส่วนเบี่ยงเบนมาตรฐาน (SD) ความเบ้ (Sk) และความโด่ง (Ku) ของคะแนนผลการทดสอบของนักเรียนซึ่งเป็นคะแนนที่สังเกตได้ โดยใช้โปรแกรม SPSS ในการวิเคราะห์ การนำเสนอในส่วนนี้เป็นการนำเสนอตามรายวิชาและชุดของข้อสอบมีรายละเอียดดังนี้

ผลการวิเคราะห์ค่าสถิติพื้นฐานของคะแนนผลการทดสอบรายวิชาคณิตศาสตร์ ชุด A จากจำนวนข้อสอบ 30 ข้อ คะแนนเต็ม 100 คะแนน พบว่า นักเรียนที่เข้าสอบจำนวน 340,084 คน มีเฉลี่ยเท่ากับ 25.27 คะแนน ส่วนเบี่ยงเบนมาตรฐานเท่ากับ 11.29 ผู้สอบสอบได้คะแนนสูงสุด 100 คะแนน และคะแนนต่ำสุด 0 คะแนน เมื่อพิจารณาค่าความเบ้ (Sk) พบว่า โดยภาพรวมมีลักษณะการแจกแจงในลักษณะเบ้ขวา (ค่าความเบ้มีค่าเป็นบวก) แสดงว่าคนส่วนใหญ่มีคะแนนสอบต่ำกว่าค่าเฉลี่ย และเมื่อพิจารณาค่าความโด่ง (Ku) พบว่า คะแนนมีการกระจายน้อย (ความโด่งมีค่าเป็นบวก)

คะแนนผลการทดสอบรายวิชาคณิตศาสตร์ ชุด B จากจำนวนข้อสอบ 30 ข้อ คะแนนเต็ม 100 คะแนน พบว่า นักเรียนที่เข้าสอบจำนวน 339,961 คน มีเฉลี่ยเท่ากับ 25.63 คะแนน ส่วนเบี่ยงเบนมาตรฐานเท่ากับ 11.21 ผู้สอบสอบได้คะแนนสูงสุด 100 คะแนน และคะแนนต่ำสุด 0 คะแนน เมื่อพิจารณาค่าความเบ้ (Sk) พบว่า โดยภาพรวมมีลักษณะการแจกแจงในลักษณะเบ้ขวา (ค่าความเบ้มีค่าเป็นบวก) แสดงว่าคนส่วนใหญ่มีคะแนนสอบต่ำกว่าค่าเฉลี่ย และเมื่อพิจารณาค่าความโด่ง (Ku) พบว่า คะแนนมีการกระจายน้อย (ความโด่งมีค่าเป็นบวก)

คะแนนผลการทดสอบรายวิชาภาษาไทย ชุด A จากจำนวนข้อสอบ 52 ข้อ คะแนนเต็ม 100 คะแนน พบว่า นักเรียนที่เข้าสอบจำนวน 340,100 คน มีเฉลี่ยเท่ากับ 41.19 คะแนน ส่วนเบี่ยงเบนมาตรฐานเท่ากับ 11.59 ผู้สอบสอบได้คะแนนสูงสุด 84 คะแนน และคะแนนต่ำสุด 0 คะแนน เมื่อพิจารณาค่าความเบ้ (Sk) พบว่า โดยภาพรวมมีลักษณะ การแจกแจงในลักษณะเบ้ขวา (ค่าความเบ้มี

ค่าเป็นบวก) แสดงว่าคนส่วนใหญ่มีคะแนนสอบต่ำกว่าค่าเฉลี่ย และเมื่อพิจารณาค่าความโด่ง (Ku) พบว่า คะแนนมีการกระจายมาก (ความโด่งมีค่าเป็นลบ)

คะแนนผลการทดสอบรายวิชาภาษาไทย ชุด B ผลการวิเคราะห์ค่าสถิติพื้นฐานของคะแนนผลการทดสอบรายวิชาภาษาไทย ชุด B จากจำนวนข้อสอบ 52 ข้อ คะแนนเต็ม 100 คะแนน พบว่า นักเรียนที่เข้าสอบจำนวน 340,552 คน มีเฉลี่ยเท่ากับ 41.22 คะแนน ส่วนเบี่ยงเบนมาตรฐานเท่ากับ 11.66 ผู้สอบสอบได้คะแนนสูงสุด 84 คะแนน และคะแนนต่ำสุด 2 คะแนน เมื่อพิจารณาค่าความเบ้ (Sk) พบว่า โดยภาพรวมมีลักษณะการแจกแจงในลักษณะเบ้ขวา (ค่าความเบ้มีค่าเป็นบวก) แสดงว่าคนส่วนใหญ่มีคะแนนสอบต่ำกว่าค่าเฉลี่ย และเมื่อพิจารณาค่าความโด่ง (Ku) พบว่า คะแนนมีการกระจายมาก (ความโด่งมีค่าเป็นลบ)

ซึ่งผลการวิเคราะห์ค่าสถิติพื้นฐานดังกล่าวมาข้างต้นแสดงได้ในตารางต่อไปนี้

ตารางที่ 4.18 ค่าสถิติพื้นฐานของคะแนนผลการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET)

รายวิชา	ชุดข้อสอบ	N	Max	Min	Mean	SD	Sk	Ku
คณิตศาสตร์	A	340,084	100	0	25.27	11.29	1.65	5.42
	B	339,961	100	0	25.63	11.21	1.63	5.44
ภาษาไทย	A	340,100	84	0	41.19	11.59	0.05	-0.53
	B	340,552	84	2	41.22	11.66	0.04	-0.53

3.2 ผลการวิเคราะห์ค่าสถิติพื้นฐานของค่าพารามิเตอร์ความสามารถของผู้สอบและค่าพารามิเตอร์ของข้อสอบ

1) ผลการวิเคราะห์ค่าสถิติพื้นฐานของค่าพารามิเตอร์ของข้อสอบ

ส่วนนี้เป็นการนำเสนอผลการวิเคราะห์ข้อสอบจากแบบทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET) ของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ที่ทำการทดสอบในปีการศึกษา 2556 จำนวน 8,000 คน สองรายวิชาหลักคือวิชาคณิตศาสตร์และภาษาไทย ซึ่งในแต่ละรายวิชามีชุดข้อสอบจำนวน 2 ชุด คือชุด A และชุด B ซึ่งผลการวิเคราะห์ในส่วนนี้ประกอบด้วยผลการวิเคราะห์ข้อสอบรายข้อและผลการประเมินความสอดคล้องของโมเดลที่ใช้ในการวิเคราะห์กับข้อมูลเชิงประจักษ์

เมื่อนำผลการตอบข้อสอบของรายวิชาภาษาไทยและคณิตศาสตร์ทั้งชุด A และชุด B มาทำการวิเคราะห์ด้วยโมเดลการทดสอบแบบดั้งเดิม (CTT) และโมเดลการตอบสนองข้อสอบ (IRT) โดยใช้โปรแกรม IRTPRO ในการวิเคราะห์ ซึ่งในการประมาณค่าพารามิเตอร์ด้วยโมเดลการตอบสนองข้อสอบนั้นใช้หลักการความเป็นไปได้สูงสุด (Maximum Likelihood) ผลการวิเคราะห์มีรายละเอียดดังนี้

แบบสอบวิชาคณิตศาสตร์ชุด A มีข้อสอบจำนวน 25 ข้อ เมื่อวิเคราะห์ด้วยโมเดลการทดสอบแบบดั้งเดิม (CTT) พบว่า มีค่าความยาก (p) อยู่ในช่วง 0.12 ถึง 0.59 ค่าอำนาจจำแนก (Item-total correlation) อยู่ในช่วง -0.04 ถึง 0.35 ค่าความเที่ยงเท่ากับ 0.51 เมื่อวิเคราะห์ด้วยโมเดลการตอบสนองข้อสอบ (IRT) พบว่า การประมาณค่าพารามิเตอร์ด้วยโมเดล 1PL พบว่า มีค่าความยาก (b) อยู่ในช่วง -0.87 ถึง 4 ค่าความเที่ยงเท่ากับ 0.48 มีข้อสอบที่ไม่เหมาะสมกับโมเดลจำนวน 19 ข้อ คิดเป็นร้อยละ 76 ส่วนการประมาณค่าพารามิเตอร์ด้วยโมเดล 2PL มีค่าความยาก (b) อยู่ในช่วง -4 ถึง 4 ค่าอำนาจจำแนก (a) อยู่ในช่วง -0.21 ถึง 1.60 ค่าความเที่ยงเท่ากับ 0.56 มีข้อสอบที่ไม่เหมาะสมกับโมเดลจำนวน 18 ข้อ คิดเป็นร้อยละ 72 และการประมาณค่าพารามิเตอร์ด้วยโมเดล 3PL มีค่าความยาก (b) อยู่ในช่วง -0.08 ถึง 4 ค่าอำนาจจำแนก (a) อยู่ในช่วง 0.37 ถึง 3.44 โอกาสในการเดาข้อสอบถูก (c) อยู่ในช่วง 0.09 ถึง 0.32 ค่าความเที่ยงเท่ากับ 0.45 มีข้อสอบที่ไม่เหมาะสมกับโมเดลจำนวน 5 ข้อ คิดเป็นร้อยละ 20

เมื่อตรวจสอบความเหมาะสมระหว่างโมเดล 1PL, 2PL และ 3PL สำหรับแบบสอบวิชาคณิตศาสตร์ชุด A โดยพิจารณาจากค่าเกณฑ์สารสนเทศเอไคคี (Akaike Information Criterion: AIC) และค่าเกณฑ์สารสนเทศเบเซียน (Bayesian Information Criterion: BIC) พบว่า โมเดล 1PL มีค่า AIC เท่ากับ 56,749.73 และมีค่า BIC เท่ากับ 56,895.36 โมเดล 2PL มีค่า AIC เท่ากับ 56,527.73 และมีค่า BIC เท่ากับ 56,807.78 ส่วนโมเดล 3PL มีค่า AIC เท่ากับ 56,255.48 และมีค่า BIC เท่ากับ 56,675.55 ซึ่งโมเดล 3PL มีค่า AIC และค่า BIC น้อยที่สุด จากการเปรียบเทียบโมเดลแสดงว่าโมเดล 3PL มีความเหมาะสมของโมเดลกับข้อมูลมากกว่าโมเดล 1PL และ 2PL ผลการวิเคราะห์แสดงได้ดังตารางที่ 4.19

เมื่อพิจารณาค่าฟังก์ชันสารสนเทศของแบบสอบวิชาคณิตศาสตร์ชุด A ตามโมเดล 3PL พบว่า สามารถวิเคราะห์ข้อสอบได้ดีในช่วง θ ระหว่าง 1 ถึง 4 แสดงว่าแบบสอบคณิตศาสตร์ชุด A สามารถใช้ได้ดีกับนักเรียนที่มีความสามารถทางคณิตศาสตร์ระดับสูง ดังภาพที่ 4.45

ตารางที่ 4.19 ผลการวิเคราะห์ค่าสถิติของข้อสอบ 25 ข้อ จากแบบสอบวิชาคณิตศาสตร์ ชุด A (n=2,000) จำแนกตามโมเดลการวัด

ข้อสอบ	โมเดลการวัด							
	CTT		1PL	2PL		3PL		
	p	Item-total	b	a	b	a	b	c
1	0.12	0.20	4.00*	0.68	3.19*	2.56	2.35	0.09
2	0.30	0.18	1.98	0.61	1.51	0.90	1.58	0.09
3	0.44	0.18	0.60	0.68	0.41*	1.37	1.13	0.26
4	0.27	-0.04	2.34*	-0.21	-4.00	0.37	4.00*	0.27
5	0.26	0.15	2.37*	0.43	2.49*	2.79	2.04	0.23
6	0.29	0.14	2.02*	0.43	2.13*	1.14	2.16*	0.20
7	0.46	0.08	0.40*	0.26	0.68*	0.49	1.61	0.20
8	0.19	0.19	3.32*	0.70	2.24*	2.49	1.94	0.14
9	0.26	0.06	2.38*	0.14	4.00	2.23	2.86	0.25
10	0.21	0.12	3.11*	0.36	3.88*	2.00	2.50	0.18
11	0.25	0.35	2.55*	1.60	0.97*	3.42	1.13	0.11
12	0.23	0.14	2.73*	0.46	2.68*	2.51	2.04	0.19
13	0.34	0.21	1.49*	0.64	1.09*	1.25	1.47	0.19
14	0.22	0.08	2.87*	0.24	4.00*	1.48	2.95	0.20
15	0.16	0.10	3.78	0.35	4.00	1.02	3.46	0.12
16	0.36	0.12	1.29*	0.38	1.54*	2.33	1.91*	0.32
17	0.31	0.19	1.86*	0.63	1.39*	3.29	1.59*	0.25
18	0.28	0.18	2.15*	0.51	1.91*	2.36	1.88*	0.23
19	0.28	0.12	2.16*	0.35	2.74*	1.33	2.41	0.22
20	0.20	0.09	3.27*	0.19	4.00*	2.94	2.56	0.18
21	0.49	0.15	0.11	0.47	0.10	0.65	0.89	0.19
22	0.47	0.18	0.31	0.55	0.25	0.81	0.80	0.17
23	0.59	0.13	-0.87	0.51	-0.79	0.73	-0.08	0.16
24	0.21	0.13	3.01*	0.41	3.30*	3.44	2.19	0.19
25	0.12	0.09	4.00*	0.28	4.00*	2.87	2.88	0.12

ข้อสอบ	โมเดลการวัด								
	CTT		1PL	2PL		3PL		
	p	Item-total	b	a	b	a	b	c
-2 Log Likelihood			56,697.73	> 56,427.73		> 56,105.48		
Akaike Information Criterion (AIC):			56,749.73	> 56,527.73		> 56,255.48		
Bayesian Information Criterion (BIC):			56,895.36	> 56,807.78		> 56,675.55		

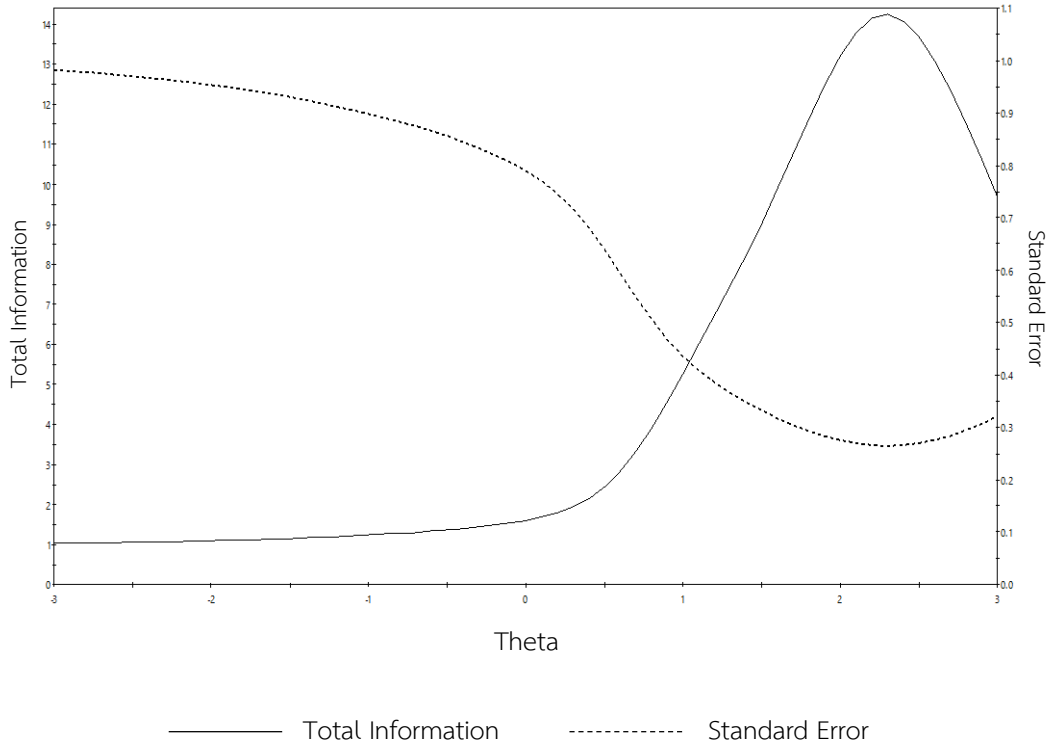
* ข้อสอบที่ไม่เหมาะสมกับโมเดลการวัด

หมายเหตุ : Coefficient alpha สำหรับโมเดล CTT = 0.51

marginal reliability สำหรับโมเดล 1PL = 0.48

marginal reliability สำหรับโมเดล 2PL = 0.56

marginal reliability สำหรับโมเดล 3PL = 0.45



ภาพที่ 4.45 โค้งสารสนเทศและความคลาดเคลื่อนมาตรฐานของแบบสอบวิชาคณิตศาสตร์ชุด A

แบบสอบวิชาคณิตศาสตร์ชุด B มีข้อสอบจำนวน 25 ข้อ เมื่อวิเคราะห์ด้วยโมเดลการทดสอบแบบดั้งเดิม (CTT) พบว่า มีค่าความยาก (p) อยู่ในช่วง 0.13 ถึง 0.59 ค่าอำนาจจำแนก (Item-total correlation) อยู่ในช่วง -0.08 ถึง 0.29 ค่าความเที่ยงเท่ากับ 0.46 เมื่อวิเคราะห์ด้วยโมเดลการตอบสนองข้อสอบ (IRT) พบว่า การประมาณค่าพารามิเตอร์ด้วยโมเดล 1PL พบว่า มีค่าความยาก (b) อยู่ในช่วง -0.96 ถึง 4 ค่าความเที่ยงเท่ากับ 0.44 มีข้อสอบที่ไม่เหมาะสมกับโมเดลจำนวน 16 ข้อ คิดเป็นร้อยละ 64 ส่วนการประมาณค่าพารามิเตอร์ด้วยโมเดล 2PL มีค่าความยาก (b) อยู่ในช่วง -1.02 ถึง 4 ค่าอำนาจจำแนก (a) อยู่ในช่วง -0.32 ถึง 1.24 ค่าความเที่ยงเท่ากับ 0.53 มีข้อสอบที่ไม่เหมาะสมกับโมเดลจำนวน 15 ข้อ คิดเป็นร้อยละ 60 และการประมาณค่าพารามิเตอร์ด้วยโมเดล 3PL มีค่าความยาก (b) อยู่ในช่วง 0 ถึง 4 ค่าอำนาจจำแนก (a) อยู่ในช่วง 0.54 ถึง 4 โอกาสในการเดาข้อสอบถูก (c) อยู่ในช่วง 0.10 ถึง 0.34 ค่าความเที่ยงเท่ากับ 0.42 มีข้อสอบที่ไม่เหมาะสมกับโมเดลจำนวน 1 ข้อ คิดเป็นร้อยละ 4

เมื่อตรวจสอบความเหมาะสมระหว่างโมเดล 1PL, 2PL และ 3PL สำหรับแบบสอบวิชาคณิตศาสตร์ชุด B โดยพิจารณาจากค่าเกณฑ์สารสนเทศเอไคคี (Akaike Information Criterion: AIC) และค่าเกณฑ์สารสนเทศเบเซียน (Bayesian Information Criterion: BIC) พบว่า โมเดล 1PL มีค่า AIC เท่ากับ 57,311.24 และมีค่า BIC เท่ากับ 57,456.87 โมเดล 2PL มีค่า AIC เท่ากับ 57,057.09 และมีค่า BIC เท่ากับ 57,337.13 ส่วนโมเดล 3PL มีค่า AIC เท่ากับ 56,842.31 และมีค่า BIC เท่ากับ 57,262.38 ซึ่งโมเดล 3PL มีค่า AIC และค่า BIC น้อยที่สุด จากการเปรียบเทียบโมเดลแสดงว่าโมเดล 3PL มีความเหมาะสมของโมเดลกับข้อมูลมากกว่าโมเดล 1PL และ 2PL ผลการวิเคราะห์แสดงได้ดังตารางที่ 4.20

เมื่อพิจารณาค่าฟังก์ชันสารสนเทศของแบบสอบวิชาคณิตศาสตร์ชุด B ตามโมเดล 3PL พบว่า สามารถวิเคราะห์ข้อสอบได้ดีในช่วง θ ระหว่าง 1 ถึง 4 แสดงว่าแบบสอบคณิตศาสตร์ชุด A สามารถใช้ได้ดีกับนักเรียนที่มีความสามารถทางคณิตศาสตร์ระดับสูง ดังภาพที่ 4.46

ตารางที่ 4.20 ผลการวิเคราะห์ค่าสถิติของข้อสอบ 25 ข้อ จากแบบสอบรายวิชาคณิตศาสตร์ ชุด B (n=2,000) จำแนกตามโมเดลการวัด

ข้อสอบ	โมเดลการวัด							
	CTT		1PL	2PL		3PL		
	p	Item-total	b	a	b	a	b	c
1	0.32	0.17	1.94	0.62	1.33	1.08	1.61	0.15
2	0.41	0.15	0.93*	0.59	0.67*	1.47	1.41	0.28
3	0.30	-0.08	2.11*	-0.32	-2.64*	4.00	2.71*	0.30
4	0.25	0.20	2.85*	0.58	2.07*	3.07	1.83	0.20
5	0.16	0.09	4.00*	0.24	4.00*	2.23	2.73	0.15
6	0.46	0.14	0.42	0.44	0.39	0.77	1.37	0.24
7	0.19	0.19	3.60*	0.67	2.33*	2.32	2.02	0.15
8	0.26	0.03	2.67*	0.10	4.00*	2.35	2.78	0.25
9	0.31	0.07	2.08*	0.23	3.70*	1.18	2.84	0.26
10	0.25	0.29	2.74*	1.24	1.11*	2.91	1.30	0.14
11	0.19	0.14	3.64	0.49	3.09	1.03	2.77	0.12
12	0.33	0.14	1.78	0.46	1.59	1.09	1.94	0.21
13	0.22	0.09	3.21*	0.23	4.00*	1.73	2.73	0.19
14	0.20	0.12	3.45	0.38	3.69	0.98	3.02	0.14
15	0.39	0.14	1.10*	0.39	1.14*	2.15	1.84	0.34
16	0.32	0.20	1.87*	0.68	1.18*	2.21	1.54	0.23
17	0.25	0.21	2.76*	0.69	1.75*	3.30	1.76	0.20
18	0.29	0.04	2.27*	0.11	4.00*	2.22	2.71	0.27
19	0.18	0.02	3.79*	0.14	4.00	0.54	4.00	0.16
20	0.46	0.16	0.46*	0.62	0.32*	1.37	1.11	0.28
21	0.46	0.18	0.38	0.60	0.27	0.85	0.73	0.15
22	0.59	0.11	-0.96	0.38	-1.02	0.58	0.00	0.19
23	0.22	0.12	3.26	0.35	3.82	1.78	2.61	0.19
24	0.13	0.09	4.00	0.37	4.00	1.20	3.39	0.10
25	0.21	0.06	3.30*	0.15	4.00*	2.30	2.94	0.20

ข้อสอบ	โมเดลการวัด								
	CTT		1PL	2PL		3PL			
	p	Item-total	b	a	b	a	b	c
-2 Log Likelihood			57,259.24	>	56,957.09	>	56,692.31		
Akaike Information Criterion (AIC):			57,311.24	>	57,057.09	>	56,842.31		
Bayesian Information Criterion (BIC):			57,456.87	>	57,337.13	>	57,262.38		

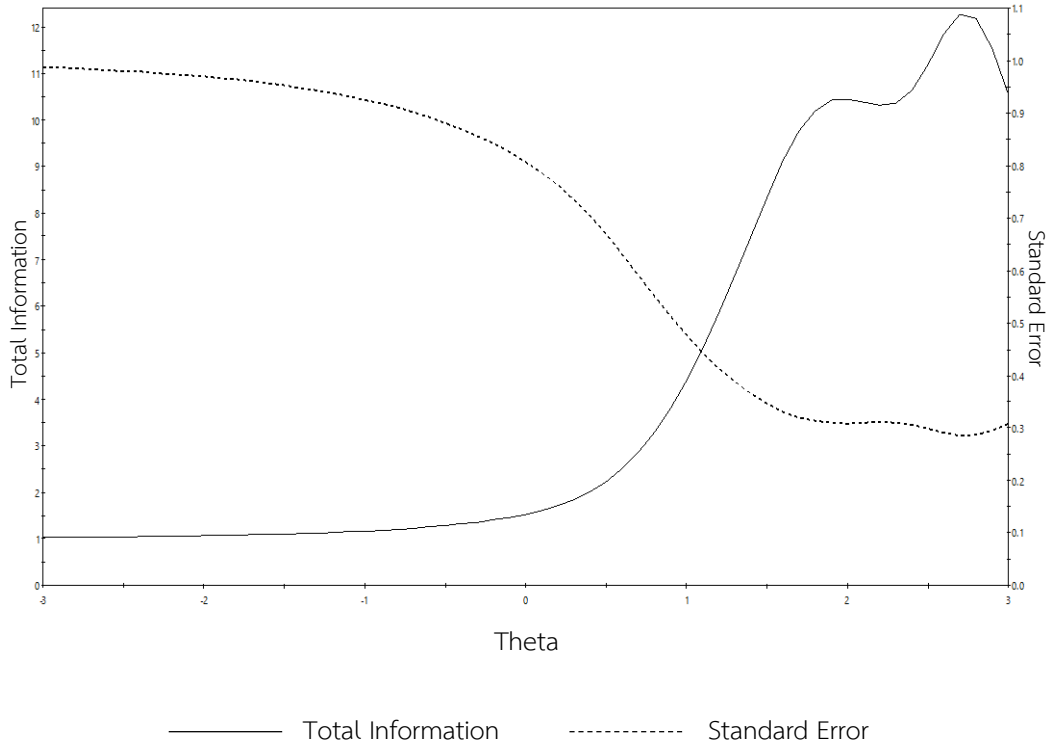
* ข้อสอบที่ไม่เหมาะสมกับโมเดลการวัด

หมายเหตุ : Coefficient alpha สำหรับโมเดล CTT = 0.46

marginal reliability สำหรับโมเดล 1PL = 0.44

marginal reliability สำหรับโมเดล 2PL = 0.53

marginal reliability สำหรับโมเดล 3PL = 0.42



ภาพที่ 4.46 โค้งสารสนเทศและความคลาดเคลื่อนมาตรฐานของแบบสอบวิชาคณิตศาสตร์ชุด B

แบบสอบวิชาภาษาไทยชุด A มีข้อสอบจำนวน 50 ข้อ เมื่อวิเคราะห์ด้วยโมเดลการทดสอบแบบดั้งเดิม (CTT) พบว่า มีค่าความยาก (p) อยู่ในช่วง 0.04 ถึง 0.81 ค่าอำนาจจำแนก (Item-total correlation) อยู่ในช่วง -0.17 ถึง 0.39 ค่าความเที่ยงเท่ากับ 0.70 เมื่อวิเคราะห์ด้วยโมเดลการตอบสนองข้อสอบ (IRT) พบว่า การประมาณค่าพารามิเตอร์ด้วยโมเดล 1PL พบว่า มีค่าความยาก (b) อยู่ในช่วง -3.17 ถึง 4 ค่าความเที่ยงเท่ากับ 0.69 มีข้อสอบที่ไม่เหมาะสมกับโมเดลจำนวน 40 ข้อ คิดเป็นร้อยละ 80 ส่วนการประมาณค่าพารามิเตอร์ด้วยโมเดล 2PL มีค่าความยาก (b) อยู่ในช่วง -4 ถึง 4 ค่าอำนาจจำแนก (a) อยู่ในช่วง -0.63 ถึง 1.78 ค่าความเที่ยงเท่ากับ 0.80 มีข้อสอบที่ไม่เหมาะสมกับโมเดลจำนวน 9 ข้อ คิดเป็นร้อยละ 18 และการประมาณค่าพารามิเตอร์ด้วยโมเดล 3PL มีค่าความยาก (b) อยู่ในช่วง -1.13 ถึง 4 ค่าอำนาจจำแนก (a) อยู่ในช่วง 0.02 ถึง 2.81 โอกาสในการเดาข้อสอบถูก (c) อยู่ในช่วง 0.08 ถึง 0.30 ค่าความเที่ยงเท่ากับ 0.80 มีข้อสอบที่ไม่เหมาะสมกับโมเดลจำนวน 6 ข้อ คิดเป็นร้อยละ 12

เมื่อตรวจสอบความเหมาะสมระหว่างโมเดล 1PL, 2PL และ 3PL สำหรับแบบสอบวิชาภาษาไทยชุด A โดยพิจารณาจากค่าเกณฑ์สารสนเทศเอไคคี (Akaike Information Criterion: AIC) และค่าเกณฑ์สารสนเทศเบเซียน (Bayesian Information Criterion: BIC) พบว่า โมเดล 1PL มีค่า AIC เท่ากับ 118,537.22 และมีค่า BIC เท่ากับ 118,822.87 โมเดล 2PL มีค่า AIC เท่ากับ 116,560.77 และมีค่า BIC เท่ากับ 117,120.86 ส่วนโมเดล 3PL มีค่า AIC เท่ากับ 116,673.55 และมีค่า BIC เท่ากับ 117,513.69 ซึ่งโมเดล 2PL มีค่า AIC และค่า BIC น้อยที่สุด จากการเปรียบเทียบโมเดลแสดงว่าโมเดล 2PL มีความเหมาะสมของโมเดลกับข้อมูลมากกว่าโมเดล 1PL และ 3PL ผลการวิเคราะห์แสดงได้ดังตารางที่ 4.21

เมื่อพิจารณาค่าฟังก์ชันสารสนเทศของแบบสอบวิชาภาษาไทยชุด A ตามโมเดล 2PL พบว่าสามารถวิเคราะห์ข้อสอบได้ดีในช่วง θ ระหว่าง -2.25 ถึง 1.30 แสดงว่าแบบสอบภาษาไทยชุด A สามารถใช้ได้ดีกับนักเรียนที่มีความสามารถทางภาษาไทยระดับต่ำ ปานกลาง และสูง ดังภาพที่ 4.47

ตารางที่ 4.21 ผลการวิเคราะห์ค่าสถิติของข้อสอบ 50 ข้อ จากแบบสอบรายวิชาภาษาไทย ชุด A (n=2,000) จำแนกตามโมเดลการวัด

ข้อสอบ	โมเดลการวัด							
	CTT		1PL	2PL		3PL		
	p	Item-total	b	a	b	a	b	c
1	0.04	-0.01	4.00*	-0.03	-4.00	0.02	4.00*	0.08
2	0.22	0.00	2.83*	0.01	4.00	0.36	4.00	0.22
3	0.05	-0.01	4.00*	-0.02	-4.00*	0.03	4.00*	0.08
4	0.25	-0.04	2.37*	-0.09	-4.00	0.37	4.00	0.25
5	0.50	0.21	0.04	0.52	0.04	0.62	0.58	0.13
6	0.49	0.28	0.09*	0.75	0.05	0.93	0.51	0.15
7	0.19	-0.05	3.19*	-0.16	-4.00	0.37	4.00	0.19
8	0.51	0.17	-0.09	0.42	-0.11	0.58	0.92	0.21
9	0.43	0.06	0.64*	0.15	1.88	0.24	3.73	0.19
10	0.41	0.20	0.82	0.53	0.75	0.89	1.37	0.20
11	0.30	0.07	1.91*	0.21	4.00	0.41	4.00	0.18
12	0.41	0.20	0.81	0.51	0.76	0.84	1.42	0.20
13	0.54	0.38	-0.35*	1.11	-0.19	1.42	0.18*	0.15
14	0.48	0.24	0.15	0.64	0.11	0.85	0.72	0.18
15	0.22	-0.01	2.75*	0.00	4.00*	0.49	4.00*	0.21
16	0.59	0.13	-0.82*	0.32	-1.19	0.38	-0.16	0.16
17	0.30	0.08	1.90*	0.15	4.00	0.33	4.00	0.18
18	0.61	0.35	-0.97*	1.02	-0.54	1.27	-0.07	0.19
19	0.48	0.22	0.22	0.56	0.19	0.67	0.73	0.14
20	0.42	0.20	0.72	0.55	0.63	0.70	1.09	0.13
21	0.53	0.19	-0.29	0.48	-0.29	0.55	0.32	0.14
22	0.72	0.33	-2.09*	1.12	-1.07	1.17	-0.80	0.14
23	0.25	-0.02	2.45*	-0.04	-4.00	0.37	4.00	0.25
24	0.47	0.31	0.27*	0.90	0.15	1.49	0.70	0.21
25	0.20	-0.02	3.02*	-0.03	-4.00	0.37	4.00	0.20

ข้อสอบ	โมเดลการวัด							
	CTT		1PL	2PL		3PL		
	p	Item-total	b	a	b	a	b	c
26	0.39	0.23	1.02	0.56	0.88	1.19	1.42	0.22
27	0.28	-0.02	2.04*	-0.04	-4.00	0.37	4.00	0.28
28	0.81	0.39	-3.17*	1.78	-1.23	2.03	-0.92	0.21
29	0.57	0.13	-0.61*	0.35	-0.82*	0.40	0.08*	0.15
30	0.57	0.24	-0.62*	0.64	-0.48	0.71	-0.03	0.13
31	0.44	0.23	0.51	0.58	0.43	1.55	1.22	0.29
32	0.49	0.30	0.08*	0.83	0.04*	1.26	0.60	0.20
33	0.45	0.24	0.42*	0.63	0.33*	1.23	1.03	0.25
34	0.56	0.36	-0.53*	1.15	-0.28*	2.81	0.41	0.30
35	0.23	0.17	2.62*	0.52	2.42*	1.76	1.89	0.16
36	0.79	0.32	-2.90*	1.17	-1.42	1.25	-1.13	0.16
37	0.56	0.36	-0.54*	1.09	-0.29	1.57	0.21	0.21
38	0.62	0.35	-1.04*	1.02	-0.57	1.28	-0.07	0.20
39	0.32	0.05	1.66*	0.14	4.00	1.49	2.82	0.29
40	0.62	0.30	-1.04*	0.87	-0.64	0.99	-0.18	0.17
41	0.31	0.22	1.75*	0.62	1.39*	1.71	1.50	0.19
42	0.39	0.08	1.02*	0.19	2.42	0.31	3.70	0.19
43	0.27	-0.01	2.19*	0.00	4.00	0.38	4.00	0.26
44	0.23	0.10	2.64*	0.21	4.00	1.46	2.77	0.20
45	0.14	-0.17	3.95*	-0.63	-3.09	0.02	4.00*	0.20
46	0.29	0.07	1.95*	0.17	4.00	0.36	5.10	0.18
47	0.66	0.37	-1.44*	1.17	-0.72	1.33	-0.42	0.14
48	0.21	0.21	2.86*	0.64	2.22*	2.61	1.68	0.14
49	0.18	0.01	3.27*	0.03	4.00	0.41	4.00	0.17
50	0.71	0.30	-2.00*	0.96	-1.14	1.01	-0.84	0.13

ข้อสอบ	โมเดลการวัด								
	CTT		1PL	2PL		3PL			
	p	Item-total	b	a	b	a	b	c	
-2 Log Likelihood			118,435.22	>	116,360.77	<	116,373.55		
Akaike Information Criterion (AIC):			118,537.22	>	116,560.77	<	116,673.55		
Bayesian Information Criterion (BIC):			118,822.87	>	117,120.86	<	117,513.69		

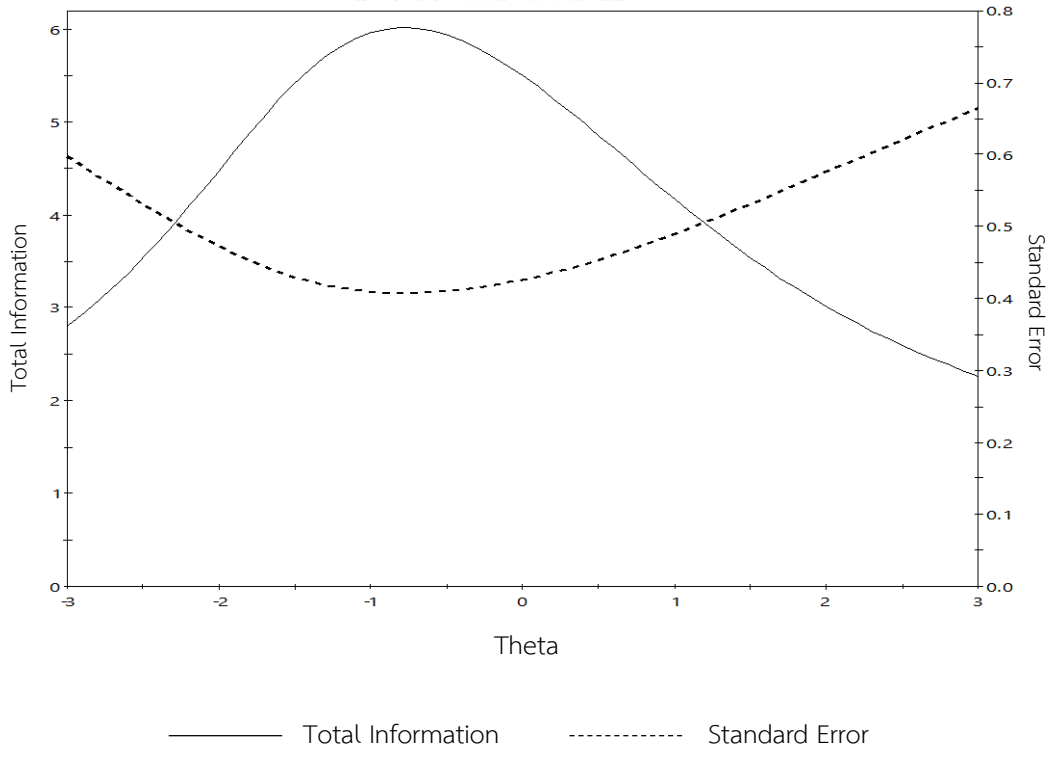
* ข้อสอบที่ไม่เหมาะสมกับโมเดลการวัด

หมายเหตุ : Coefficient alpha สำหรับโมเดล CTT = 0.70

marginal reliability สำหรับโมเดล 1PL = 0.69

marginal reliability สำหรับโมเดล 2PL = 0.80

marginal reliability สำหรับโมเดล 3PL = 0.80



ภาพที่ 4.47 โค้งสารสนเทศและความคลาดเคลื่อนมาตรฐานของแบบสอบวิชาภาษาไทยชุด A

แบบสอบวิชาภาษาไทยชุด B มีข้อสอบจำนวน 50 ข้อ เมื่อวิเคราะห์ด้วยโมเดลการทดสอบแบบดั้งเดิม (CTT) พบว่า มีค่าความยาก (p) อยู่ในช่วง 0.06 ถึง 0.83 ค่าอำนาจจำแนก (Item-total correlation) อยู่ในช่วง -0.18 ถึง 0.41 ค่าความเที่ยงเท่ากับ 0.70 เมื่อวิเคราะห์ด้วยโมเดลการตอบสนองข้อสอบ (IRT) พบว่า การประมาณค่าพารามิเตอร์ด้วยโมเดล 1PL พบว่า มีค่าความยาก (b) อยู่ในช่วง -3.44 ถึง 4 ค่าความเที่ยงเท่ากับ 0.70 มีข้อสอบที่ไม่เหมาะสมกับโมเดลจำนวน 41 ข้อ คิดเป็นร้อยละ 82 ส่วนการประมาณค่าพารามิเตอร์ด้วยโมเดล 2PL มีค่าความยาก (b) อยู่ในช่วง -4 ถึง 4 ค่าอำนาจจำแนก (a) อยู่ในช่วง -0.63 ถึง 2.09 ค่าความเที่ยงเท่ากับ 0.80 มีข้อสอบที่ไม่เหมาะสมกับโมเดลจำนวน 14 ข้อ คิดเป็นร้อยละ 28 และการประมาณค่าพารามิเตอร์ด้วยโมเดล 3PL มีค่าความยาก (b) อยู่ในช่วง -1.07 ถึง 4 ค่าอำนาจจำแนก (a) อยู่ในช่วง 0.03 ถึง 2.28 โอกาสในการเดาข้อสอบถูก (c) อยู่ในช่วง 0.07 ถึง 0.30 ค่าความเที่ยงเท่ากับ 0.80 มีข้อสอบที่ไม่เหมาะสมกับโมเดลจำนวน 14 ข้อ คิดเป็นร้อยละ 28

เมื่อตรวจสอบความเหมาะสมระหว่างโมเดล 1PL, 2PL และ 3PL สำหรับแบบสอบวิชาภาษาไทยชุด B โดยพิจารณาจากค่าเกณฑ์สารสนเทศเอไคคี (Akaike Information Criterion: AIC) และค่าเกณฑ์สารสนเทศเบเซียน (Bayesian Information Criterion: BIC) พบว่า โมเดล 1PL มีค่า AIC เท่ากับ 118,690.82 และมีค่า BIC เท่ากับ 118,976.47 โมเดล 2PL มีค่า AIC เท่ากับ 116,566.48 และมีค่า BIC เท่ากับ 117,126.57 ส่วนโมเดล 3PL มีค่า AIC เท่ากับ 116,666.68 และมีค่า BIC เท่ากับ 117,506.81 ซึ่งโมเดล 2PL มีค่า AIC และค่า BIC น้อยที่สุด จากการเปรียบเทียบโมเดลแสดงว่าโมเดล 2PL มีความเหมาะสมของโมเดลกับข้อมูลมากกว่าโมเดล 1PL และ 3PL ผลการวิเคราะห์แสดงได้ดังตารางที่ 4.22

เมื่อพิจารณาค่าฟังก์ชันสารสนเทศของแบบสอบวิชาภาษาไทยชุด B ตามโมเดล 2PL พบว่าสามารถวิเคราะห์ข้อสอบได้ดีในช่วง θ ระหว่าง -2.25 ถึง 1.20 แสดงว่าแบบสอบภาษาไทยชุด B สามารถใช้ได้ดีกับนักเรียนที่มีความสามารถทางภาษาไทยระดับต่ำ ปานกลาง และสูง ดังภาพที่ 4.48

ตารางที่ 4.22 ผลการวิเคราะห์ค่าสถิติของข้อสอบ 50 ข้อ จากแบบสอบรายวิชาภาษาไทย ชุด B (n=2,000) จำแนกตามโมเดลการวัด

ข้อสอบ	โมเดลการวัด							
	CTT		1PL	2PL		3PL		
	p	Item-total	b	a	b	a	b	c
1	0.54	0.19	-0.32	0.45	-0.34	0.52	0.36	0.15
2	0.43	0.07	0.65*	0.18	1.70	0.25	3.30	0.17
3	0.41	0.20	0.79	0.50	0.77	0.75	1.36	0.17
4	0.29	0.02	1.96*	0.02	4.00	0.34	4.00	0.27
5	0.43	0.20	0.63	0.48	0.62	0.83	1.38	0.22
6	0.56	0.35	-0.53*	1.05	-0.29	1.27	0.07	0.14
7	0.48	0.22	0.22*	0.56	0.18*	0.85	0.90*	0.20
8	0.23	0.00	2.63*	-0.02	-4.00	0.38	4.00	0.23
9	0.06	-0.02	4.00*	-0.16	-4.00*	0.03	4.00*	0.07
10	0.23	-0.02	2.69*	-0.06	-4.00	0.27	4.00	0.23
11	0.07	-0.05	4.00*	-0.28	-4.00*	0.03	4.00*	0.10
12	0.22	-0.02	2.73*	-0.03	-4.00	0.37	4.00	0.22
13	0.48	0.25	0.14*	0.70	0.09	0.78	0.42*	0.10
14	0.49	0.30	0.13*	0.85	0.07*	1.26	0.60	0.19
15	0.22	-0.04	2.72*	-0.14	-4.00	0.32	4.00	0.23
16	0.61	0.11	-0.98*	0.30	-1.52*	0.35	-0.30*	0.18
17	0.21	0.02	2.91*	0.07	4.00	0.53	4.00	0.18
18	0.41	0.19	0.79	0.45	0.84	1.01	1.57	0.25
19	0.26	-0.03	2.28*	-0.07	-4.00	0.37	4.00	0.26
20	0.83	0.40	-3.44*	2.09	-1.25*	2.28	-1.07	0.16
21	0.59	0.15	-0.82*	0.40	-0.99*	0.43	-0.27*	0.14
22	0.58	0.22	-0.69	0.54	-0.62	0.64	0.05	0.17
23	0.43	0.28	0.66*	0.76	0.45	1.35	0.94	0.20
24	0.27	0.10	2.16*	0.25	4.00	0.51	3.94	0.16
25	0.56	0.36	-0.53*	1.05	-0.29	1.34	0.14	0.17

ข้อสอบ	โมเดลการวัด							
	CTT		1PL	2PL		3PL		
	p	Item-total	b	a	b	a	b	c
26	0.47	0.19	0.27	0.45	0.29	0.56	0.90	0.14
27	0.38	0.21	1.07	0.58	0.91	0.70	1.27*	0.10
28	0.52	0.19	-0.17	0.45	-0.18	0.55	0.55	0.16
29	0.71	0.36	-1.99*	1.24	-0.96	1.36	-0.67	0.15
30	0.26	0.05	2.25*	0.15	4.00	1.47	2.74	0.23
31	0.45	0.26	0.44*	0.66	0.33	1.33	1.02	0.25
32	0.52	0.30	-0.14*	0.83	-0.09	1.05	0.34	0.15
33	0.44	0.25	0.55*	0.64	0.43*	1.04	0.99*	0.19
34	0.58	0.36	-0.69*	1.06	-0.37*	2.24	0.36	0.30
35	0.23	0.16	2.58	0.45	2.76	2.14	1.88	0.17
36	0.79	0.31	-2.86*	1.17	-1.42	1.28	-1.07	0.19
37	0.55	0.35	-0.45*	0.99	-0.26	1.43	0.26	0.20
38	0.59	0.37	-0.81*	1.07	-0.44	1.22	-0.13*	0.13
39	0.29	0.07	1.91*	0.17	4.00*	0.46	4.00*	0.20
40	0.35	0.12	1.37*	0.35	1.86	0.43	2.53*	0.12
41	0.69	0.41	-1.73*	1.49	-0.76	1.59	-0.60	0.09
42	0.37	0.07	1.16*	0.20	2.72*	0.28	4.00*	0.16
43	0.22	-0.01	2.74*	-0.04	-4.00	0.37	4.00	0.22
44	0.25	0.03	2.42*	0.02	4.00	0.42	4.00	0.23
45	0.14	-0.18	3.92*	-0.63	-3.08	0.03	4.00*	0.20
46	0.56	0.33	-0.49*	0.93	-0.29*	1.14	0.13*	0.16
47	0.29	0.21	1.98*	0.57	1.71*	2.17	1.52	0.19
48	0.70	0.31	-1.87*	0.95	-1.08	1.02	-0.73	0.15
49	0.19	0.03	3.13*	0.08	4.00	0.41	4.00	0.16
50	0.19	0.20	3.12*	0.59	2.61*	2.08	1.82	0.12

ข้อสอบ	โมเดลการวัด								
	CTT		1PL	2PL		3PL			
	p	Item-total	b	a	b	a	b	c	
-2 Log Likelihood			118,588.82	>	116,366.48	<	116,366.68		
Akaike Information Criterion (AIC):			118,690.82	>	116,566.48	<	116,666.68		
Bayesian Information Criterion (BIC):			118,976.47	>	117,126.57	<	117,506.81		

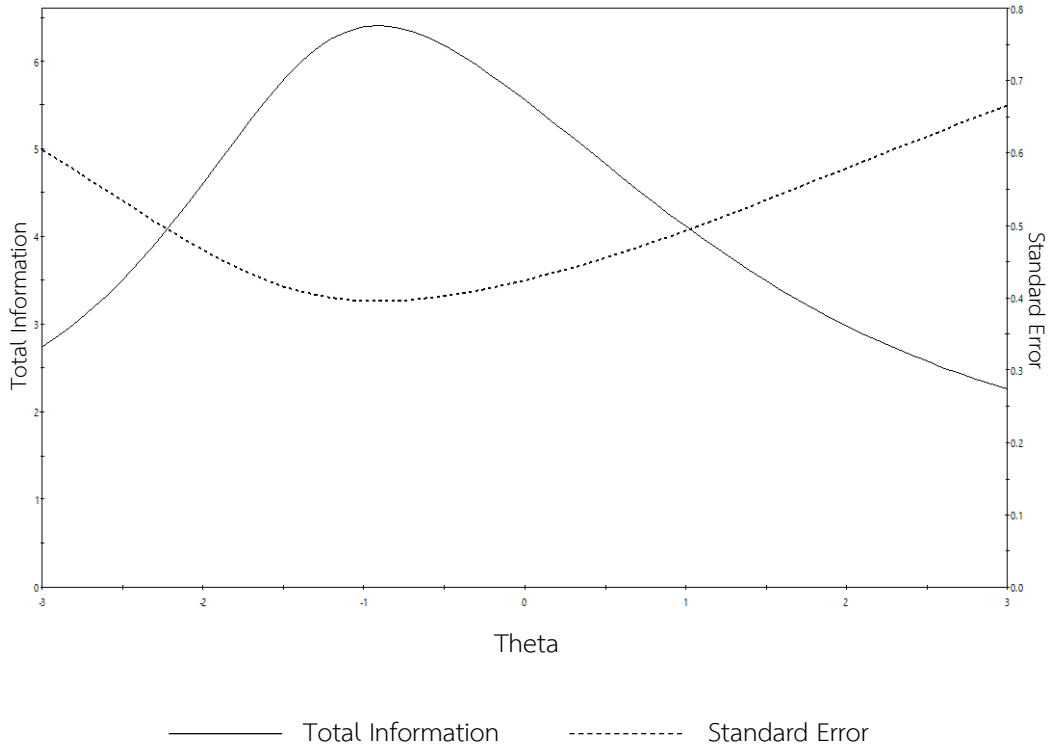
* ข้อสอบที่ไม่เหมาะสมกับโมเดลการวัด

หมายเหตุ : Coefficient alpha สำหรับโมเดล CTT = 0.70

marginal reliability สำหรับโมเดล 1PL = 0.70

marginal reliability สำหรับโมเดล 2PL = 0.80

marginal reliability สำหรับโมเดล 3PL = 0.80



ภาพที่ 4.48 โค้งสารสนเทศและความคลาดเคลื่อนมาตรฐานของแบบสอบวิชาภาษาไทยชุด B

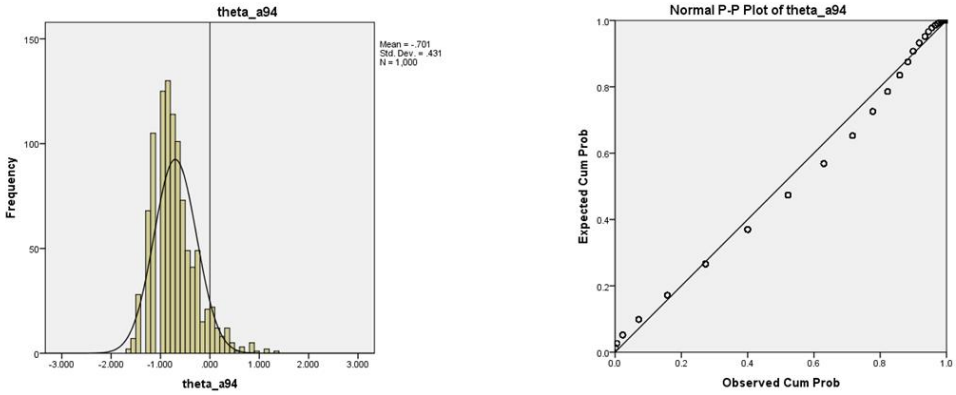
2) ผลการวิเคราะห์ค่าสถิติพื้นฐานของค่าพารามิเตอร์ความสามารถของผู้สอบ

การวิเคราะห์ข้อมูลในส่วนนี้เป็นการวิเคราะห์ด้วยค่าสถิติพื้นฐานเพื่อบรรยายลักษณะของค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) ที่เป็นตัวอย่างในการวิจัยจำนวน 8,000 คน และค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (se) โดยค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) และค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (se) นั้นใช้โปรแกรม IRTPRO ในการวิเคราะห์ ผลการวิเคราะห์ในส่วนนี้ประกอบด้วยผลการวิเคราะห์ค่าสถิติเชิงบรรยาย และผลการตรวจสอบการแจกแจงปกติของค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (se) เพื่อนำไปใช้ในการพิจารณาเลือกใช้วิธีการประมาณค่าดัชนีการจำแนกประเภทที่เหมาะสมกับรูปแบบของการแจกแจงของค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (se) โดยใช้โปรแกรม SPSS ในการวิเคราะห์ การนำเสนอในส่วนนี้เป็นการนำเสนอตามรายวิชาและชุดของข้อสอบมีรายละเอียดดังนี้

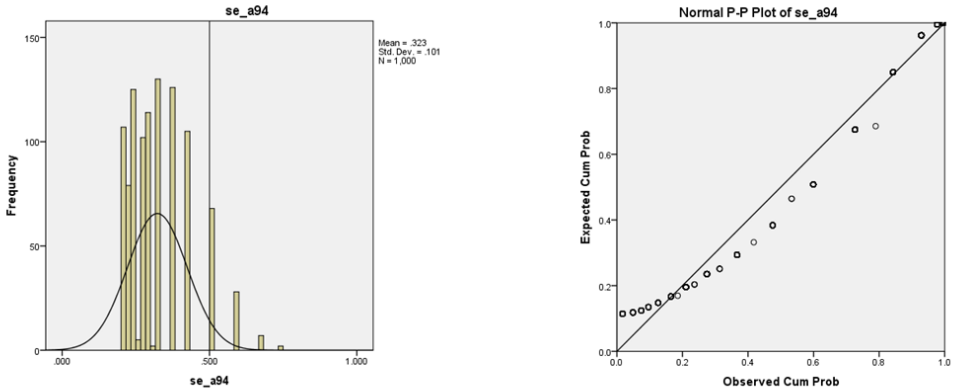
ผลการวิเคราะห์ค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) และค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (se) ของคะแนนผลการทดสอบรายวิชาคณิตศาสตร์ ชุด A พบว่าตัวอย่างผู้สอบจำนวน 2,000 คน มี θ เฉลี่ยเท่ากับ -0.70 คะแนน ส่วนเบี่ยงเบนมาตรฐานเท่ากับ 0.43 มีค่า θ อยู่ในช่วง -1.65-1.35 เมื่อทำการทดสอบการแจกแจงปกติของค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) และค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (se) โดยการวาดกราฟ P-P Plot พบว่าการแจกแจงของค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) เป็นการแจกแจงปกติ ส่วนการแจกแจงของค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ *(se)* ไม่เป็นการแจกแจงปกติ ผลการวิเคราะห์ค่าสถิติพื้นฐานและการตรวจสอบการแจกแจงปกติแสดงได้ในตารางและภาพต่อไปนี้

ตารางที่ 4.23 ค่าสถิติพื้นฐานของค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) และค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (se) ของคะแนนผลการทดสอบรายวิชาคณิตศาสตร์ ชุด A

ค่าพารามิเตอร์	Max	Min	Mean	SD	Sk	Ku
θ	1.35	-1.65	-0.70	0.43	0.96	1.67
se	0.74	0.20	0.32	0.10	1.08	0.90



ภาพที่ 4.49 ฮิสโตแกรมและ Normal P-P Plot ของค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) รายวิชาคณิตศาสตร์ ชุด A

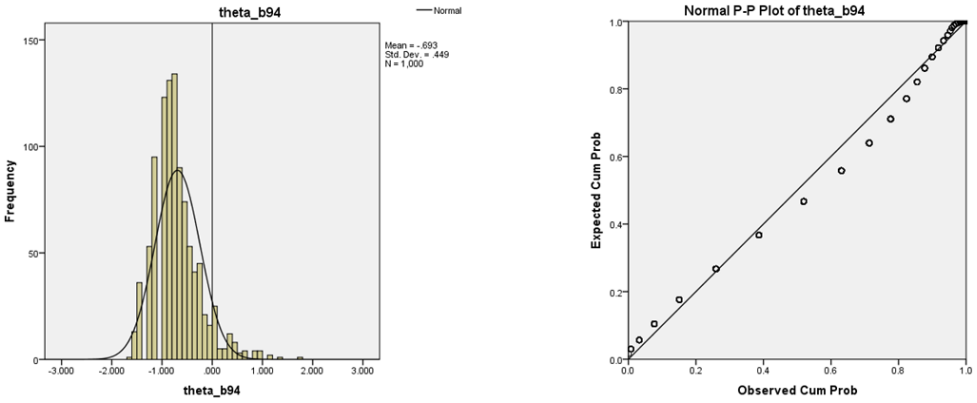


ภาพที่ 4.50 ฮิสโตแกรมและ Normal P-P Plot ของค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (se) รายวิชาคณิตศาสตร์ ชุด A

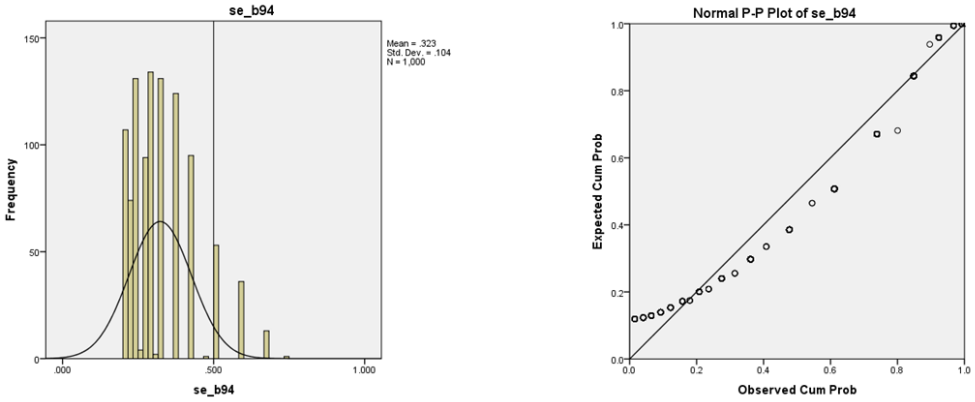
ผลการวิเคราะห์ค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) และค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (se) ของคะแนนผลการทดสอบรายวิชาคณิตศาสตร์ ชุด B พบว่าตัวอย่างผู้สอบจำนวน 2,000 คน มี θ เฉลี่ยเท่ากับ -0.69 คะแนน ส่วนเบี่ยงเบนมาตรฐานเท่ากับ 0.45 มีค่า θ อยู่ในช่วง -1.65-1.70 เมื่อทำการทดสอบการแจกแจงปกติของค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) และค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (se) โดยการวาดกราฟ P-P Plot พบว่าการแจกแจงของค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) เป็นการแจกแจงปกติ ส่วนการแจกแจงของค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (se) ไม่เป็นการแจกแจงปกติ ผลการวิเคราะห์ค่าสถิติพื้นฐานและการตรวจสอบการแจกแจงปกติแสดงได้ในตารางและภาพต่อไปนี้

ตารางที่ 4.24 ค่าสถิติพื้นฐานของค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) และค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (se) ของคะแนนผลการทดสอบรายวิชาคณิตศาสตร์ ชุด B

ค่าพารามิเตอร์	Max	Min	Mean	SD	Sk	Ku
θ	1.70	-1.65	-0.69	0.45	1.10	2.43
se	0.74	0.20	0.32	0.10	1.21	1.23



ภาพที่ 4.51 ฮิสโตแกรมและ Normal P-P Plot ของค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) รายวิชาคณิตศาสตร์ ชุด B

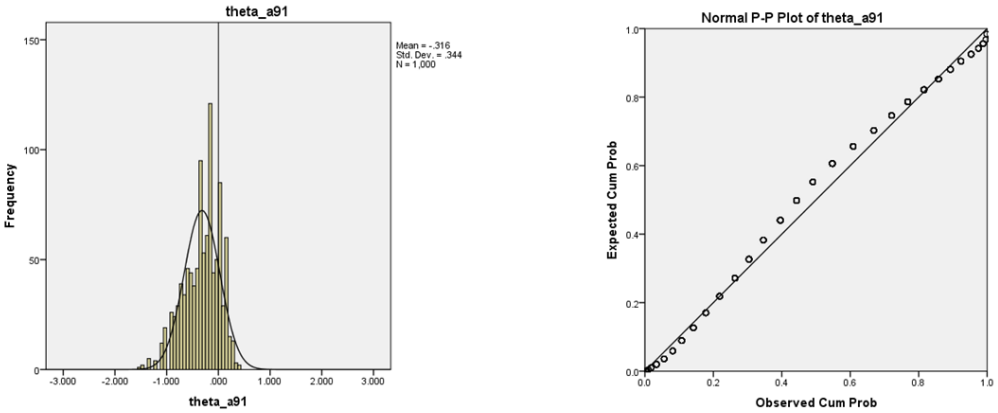


ภาพที่ 4.52 ฮิสโตแกรมและ Normal P-P Plot ของค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (se) รายวิชาคณิตศาสตร์ ชุด B

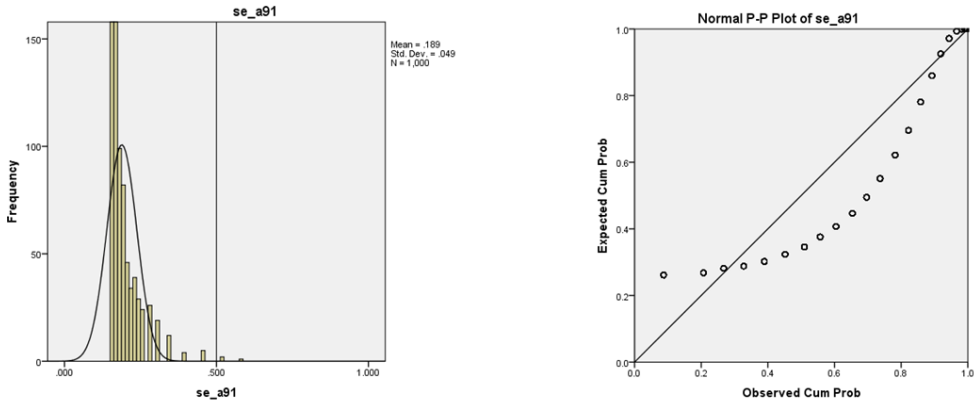
ผลการวิเคราะห์ค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) และค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (se) ของคะแนนผลการทดสอบรายวิชาภาษาไทย ชุด A พบว่า ตัวอย่างผู้สอบจำนวน 2,000 คน มี θ เฉลี่ยเท่ากับ -0.32 คะแนน ส่วนเบี่ยงเบนมาตรฐานเท่ากับ 0.34 มีค่า θ อยู่ในช่วง -1.55-0.42 เมื่อทำการทดสอบการแจกแจงปกติของค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) และค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (se) โดยการวาดกราฟ P-P Plot พบว่าการแจกแจงของค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) เป็นการแจกแจงปกติ ส่วนการแจกแจงของค่าความคลาดเคลื่อนในการประมาณค่าความสามารถ (se) ไม่เป็นการแจกแจงปกติ ผลการวิเคราะห์ค่าสถิติพื้นฐานและการตรวจสอบการแจกแจงปกติแสดงได้ในตารางและภาพต่อไปนี้

ตารางที่ 4.25 ค่าสถิติพื้นฐานของค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) และค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (se) ของคะแนนผลการทดสอบรายวิชาภาษาไทย ชุด A

ค่าพารามิเตอร์	Max	Min	Mean	SD	Sk	Ku
θ	0.42	-1.55	-0.32	0.34	-0.58	-0.02
se	0.58	0.16	0.19	0.05	3.02	12.51



ภาพที่ 4.53 ฮิสโตแกรมและ Normal P-P Plot ของค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) รายวิชาภาษาไทย ชุด A

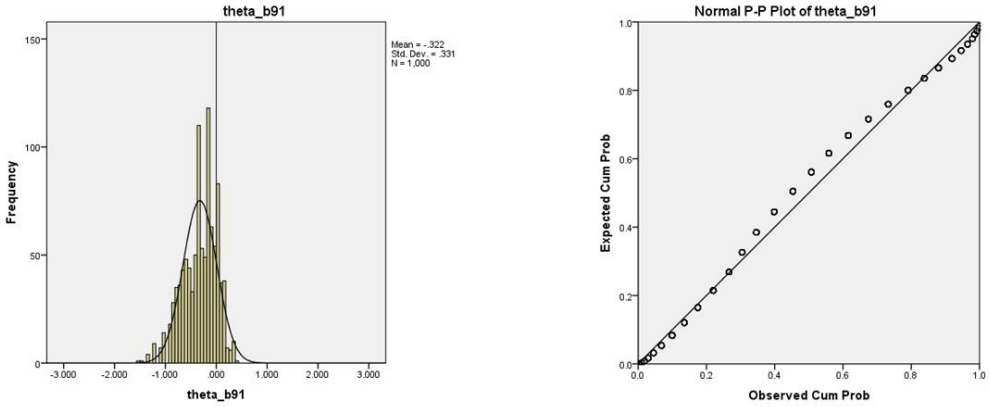


ภาพที่ 4.54 ฮิสโตแกรมและ Normal P-P Plot ของค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (se) รายวิชาภาษาไทย ชุด A

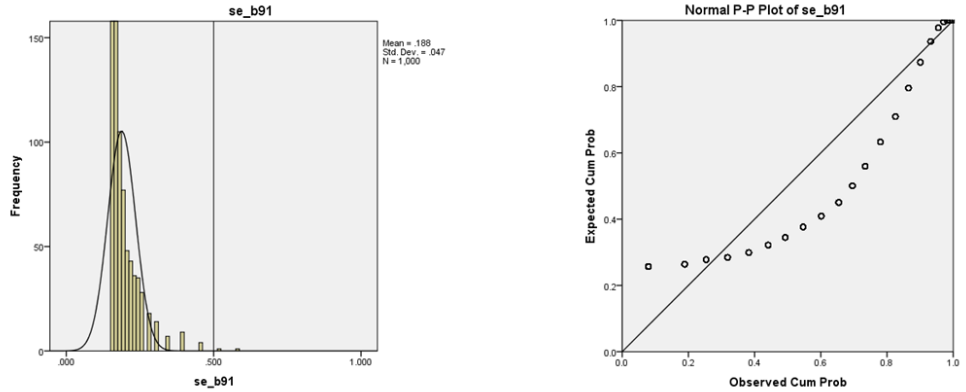
ผลการวิเคราะห์ค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) และค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (se) ของคะแนนผลการทดสอบรายวิชาภาษาไทย ชุด B พบว่า ตัวอย่างผู้สอบจำนวน 2,000 คน มี θ เฉลี่ยเท่ากับ -0.32 คะแนน ส่วนเบี่ยงเบนมาตรฐานเท่ากับ 0.33 มีค่า θ อยู่ในช่วง -1.55-0.42 เมื่อทำการทดสอบการแจกแจงปกติของค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) และค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (se) โดยการวาดกราฟ P-P Plot พบว่าการแจกแจงของค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) เป็นการแจกแจงปกติ ส่วนการแจกแจงของค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (se) ไม่เป็นการแจกแจงปกติ ผลการวิเคราะห์ค่าสถิติพื้นฐานและการตรวจสอบการแจกแจงปกติแสดงได้ในตารางและภาพต่อไปนี้

ตารางที่ 4.26 ค่าสถิติพื้นฐานของค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) และค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (se) ของคะแนนผลการทดสอบรายวิชาภาษาไทย ชุด B

ค่าพารามิเตอร์	Max	Min	Mean	SD	Sk	Ku
θ	0.42	-1.55	-0.32	0.33	-0.57	0.07
se	0.58	0.16	0.19	0.05	3.10	13.37



ภาพที่ 4.55 ฮิสโตแกรมและ Normal P-P Plot ของค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) รายวิชาภาษาไทย ชุด B



ภาพที่ 4.56 ฮิสโตแกรมและ Normal P-P Plot ของค่าความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (se) รายวิชาภาษาไทย ชุด B

3.4 ผลการวิเคราะห์ประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทตามแนวคิดทฤษฎีการตอบสนองข้อสอบเมื่อนำไปใช้กับข้อมูลเชิงประจักษ์ (empirical data)

ในส่วนนี้เป็นการนำเสนอผลการวิเคราะห์ประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทตามแนวคิดทฤษฎีการตอบสนองข้อสอบเมื่อนำไปใช้กับข้อมูลเชิงประจักษ์ (empirical data) หรือข้อมูลการตอบข้อสอบจริง (real data) ของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ในปีการศึกษา 2556 โดยจากการวิเคราะห์ค่าสถิติพื้นฐานของคะแนนสอบในส่วนที่ 3.1-3.3 ข้างต้น พบว่าคะแนนที่ได้จากแบบสอบทั้งสี่ชุดมีลักษณะดังนี้

แบบสอบคณิตศาสตร์ชุด A ประกอบด้วยข้อสอบจำนวน 25 ข้อ ข้อมูลมีความเหมาะสมกับโมเดล 3PL โดยมีความไม่เหมาะสมของข้อมูลกับโมเดลร้อยละ 20 การแจกแจงของคะแนนจริงเป็นการแจกแจงปกติ และการแจกแจงของความคลาดเคลื่อนมาตรฐานในการประมาณค่าความสามารถ (se) ไม่เป็นการแจกแจงปกติ

แบบสอบคณิตศาสตร์ชุด B ประกอบด้วยข้อสอบจำนวน 25 ข้อ ข้อมูลมีความเหมาะสมกับโมเดล 3PL โดยมีความไม่เหมาะสมของข้อมูลกับโมเดลร้อยละ 4 การแจกแจงของคะแนนจริงเป็นการแจกแจงปกติ และการแจกแจงของความคลาดเคลื่อนมาตรฐานในการประมาณค่าความสามารถ (se) ไม่เป็นการแจกแจงปกติ

แบบสอบภาษาไทยชุด A ประกอบด้วยข้อสอบจำนวน 50 ข้อ ข้อมูลมีความเหมาะสมกับโมเดล 2PL โดยมีความไม่เหมาะสมของข้อมูลกับโมเดลร้อยละ 18 การแจกแจงของคะแนนจริงเป็นการแจกแจงปกติ และการแจกแจงของความคลาดเคลื่อนมาตรฐานในการประมาณค่าความสามารถ (se) ไม่เป็นการแจกแจงปกติ

แบบสอบภาษาไทยชุด B ประกอบด้วยข้อสอบจำนวน 50 ข้อ ข้อมูลมีความเหมาะสมกับโมเดล 2PL โดยมีความไม่เหมาะสมของข้อมูลกับโมเดลร้อยละ 28 การแจกแจงของคะแนนจริงเป็นการแจกแจงปกติ และการแจกแจงของความคลาดเคลื่อนมาตรฐานในการประมาณค่าความสามารถ (se) ไม่เป็นการแจกแจงปกติ แสดงได้ดังตารางต่อไปนี้

ตารางที่ 4.27 ลักษณะของแบบสอบจำแนกตามข้อตกลงเบื้องต้น

ข้อตกลงเบื้องต้น	คณิตศาสตร์		ภาษาไทย	
	A	B	A	B
1. จำนวนข้อสอบ	25 ข้อ	25 ข้อ	50 ข้อ	50 ข้อ
2. โมเดลการวัด	3PL	3PL	2PL	2PL
3. ความไม่เหมาะสมของข้อมูลกับโมเดล	ร้อยละ 20	ร้อยละ 4	ร้อยละ 18	ร้อยละ 28
4. การแจกแจงของคะแนนจริง	การแจกแจงปกติ	การแจกแจงปกติ	การแจกแจงปกติ	การแจกแจงปกติ
5. การแจกแจงของความคลาดเคลื่อนมาตรฐานในการประมาณค่าความสามารถ (se)	ไม่เป็นการแจกแจงปกติ	ไม่เป็นการแจกแจงปกติ	ไม่เป็นการแจกแจงปกติ	ไม่เป็นการแจกแจงปกติ

เมื่อวิเคราะห์ข้อมูลพื้นฐานของแบบสอบทั้งหมดแล้วจึงทำการประมาณค่าดัชนีการจำแนกประเภทกับข้อมูลจริงโดยใช้ตัวอย่างจากแบบสอบแต่ละฉบับจำนวน 2,000 คน รวมเป็นกลุ่มตัวอย่างทั้งสิ้น 8,000 คน โดยมีผลการประมาณค่าดัชนีความถูกต้อง (accuracy) และค่าดัชนีความสอดคล้อง (consistency) ของการจำแนกประเภทของแบบสอบแต่ละฉบับดังนี้

วิธีการของ Rudner

แบบสอบคณิตศาสตร์ ชุด A มีค่าดัชนีความถูกต้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.6840 และค่าดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.5891 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับปานกลาง

แบบสอบคณิตศาสตร์ ชุด B มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.6404 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.5627 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับปานกลาง

แบบสอบภาษาไทย ชุด A มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.7504 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.6573 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับค่อนข้างสูง

แบบสอบภาษาไทย ชุด B มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.7677 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.6718 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับค่อนข้างสูง

วิธีการของ Guo

แบบสอบคณิตศาสตร์ ชุด A มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวมเท่ากับ 1 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 1 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับสูง

แบบสอบคณิตศาสตร์ ชุด B มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวมเท่ากับ 1 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 1 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับสูง

แบบสอบภาษาไทย ชุด A มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวมเท่ากับ 1 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 1 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับสูง

แบบสอบภาษาไทย ชุด B มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวมเท่ากับ 1 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 1 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับสูง

วิธีการของ Lee

แบบสอบคณิตศาสตร์ ชุด A มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.6404 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.5627 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับปานกลาง

แบบสอบคณิตศาสตร์ ชุด B มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.6424 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.5482 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับปานกลาง

แบบสอบภาษาไทย ชุด A มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.6183 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.5122 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับปานกลาง

แบบสอบภาษาไทย ชุด B มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.6267 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวม เท่ากับ 0.5232 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับปานกลาง

โดยรายละเอียดของดัชนีการจำแนกประเภทที่ประมาณค่าได้จากวิธีการของ Rudner, Guo และ Lee แสดงได้ดังตารางที่ 4.28

ตารางที่ 4.28 ค่าดัชนีความถูกต้องและความสอดคล้องของการจำแนกประเภทในสถานการณ์จริง จำแนกตามวิธีการประมาณค่า

แบบสอบ	ตำแหน่งคะแนนจุดตัด	ความถูกต้อง (accuracy)			ความสอดคล้อง (consistency)		
		Rudner	Guo	Lee	Rudner	Guo	Lee
คณิตศาสตร์ A	1	0.8070	1.0000	0.7728	0.7375	1.0000	0.6933
	2	0.9001	1.0000	0.8485	0.8485	1.0000	0.7742
	3	0.9291	1.0000	0.9372	0.8937	1.0000	0.9021
	4	0.9697	1.0000	0.9734	0.9512	1.0000	0.9597
	5	0.9875	1.0000	0.9884	0.9797	1.0000	0.9821
	6	0.9962	1.0000	0.9963	0.9942	1.0000	0.9941
	7	1.0000	1.0000	0.9957	0.9998	1.0000	0.9947
	ทั้งหมด	0.6840	1.0000	0.6406	0.5891	1.0000	0.5627
คณิตศาสตร์ B	1	0.8185	1.0000	0.7672	0.7574	1.0000	0.6741
	2	0.8979	1.0000	0.8444	0.8481	1.0000	0.7666
	3	0.9326	1.0000	0.9377	0.8960	1.0000	0.9032

แบบสอบ	ตำแหน่งคะแนนจุดตัด	ความถูกต้อง (accuracy)			ความสอดคล้อง (consistency)		
		Rudner	Guo	Lee	Rudner	Guo	Lee
	4	0.9703	1.0000	0.9726	0.9510	1.0000	0.9586
	5	0.9843	1.0000	0.9852	0.9755	1.0000	0.9786
	6	0.9971	1.0000	0.9970	0.9952	1.0000	0.9955
	7	0.9998	1.0000	0.9998	0.9996	1.0000	0.9996
	ทั้งหมด	0.6969	1.0000	0.6424	0.6061	1.0000	0.5482
ภาษาไทย A	1	0.9806	1.0000	0.9308	0.9718	1.0000	0.9031
	2	0.9230	1.0000	0.9066	0.8928	1.0000	0.8686
	3	0.8998	1.0000	0.8924	0.8596	1.0000	0.8496
	4	0.9526	1.0000	0.9156	0.9319	1.0000	0.8813
	5	0.9927	1.0000	0.9583	0.9894	1.0000	0.9407
	6	0.9997	1.0000	0.9945	0.9995	1.0000	0.9909
	7	1.0000	1.0000	0.9998	1.0000	1.0000	0.9996
	ทั้งหมด	0.7504	1.0000	0.6183	0.6573	1.0000	0.5122
ภาษาไทย B	1	0.9831	1.0000	0.9324	0.9761	1.0000	0.9067
	2	0.9438	1.0000	0.9107	0.9193	1.0000	0.8750
	3	0.9104	1.0000	0.8942	0.8713	1.0000	0.8517
	4	0.9380	1.0000	0.9136	0.9126	1.0000	0.8801
	5	0.9923	1.0000	0.9625	0.9884	1.0000	0.9452
	6	0.9993	1.0000	0.9938	0.9988	1.0000	0.9903
	7	1.0000	1.0000	0.9997	0.9999	1.0000	0.9995
	ทั้งหมด	0.7677	1.0000	0.6267	0.6718	1.0000	0.5232

จากนั้นทำการจำลองข้อมูลจากรูปแบบการตอบข้อสอบจริง เพื่อประมาณค่าดัชนีการจำแนกและหาค่าเฉลี่ยดัชนีการจำแนกประเภทประเภทจากการทำซ้ำจำนวน 100 รอบ โดยผลการวิเคราะห์ในส่วนนี้ประกอบด้วยค่าดัชนีความถูกต้อง (accuracy) และค่าดัชนีความสอดคล้อง (consistency) ของการจำแนกประเภทเฉลี่ยจากการทำซ้ำจำนวน 100 รอบ เมื่อนำไปใช้กับข้อมูลเชิงประจักษ์ (empirical data) หรือข้อมูลการตอบข้อสอบจริง (real data) โดยแบ่งการนำเสนอตามวิธีที่ใช้ในการประมาณค่า ผลการวิเคราะห์มีรายละเอียดดังนี้

วิธีการของ Rudner

แบบสอบคณิตศาสตร์ ชุด A มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.7012 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.5995 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับปานกลาง

แบบสอบคณิตศาสตร์ ชุด B มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.6676 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.5579 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับปานกลาง

แบบสอบภาษาไทย ชุด A มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.7508 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.6564 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับค่อนข้างสูง

แบบสอบภาษาไทย ชุด B มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.7581 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.6628 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับค่อนข้างสูง

วิธีการของ Guo

แบบสอบคณิตศาสตร์ ชุด A มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวมเท่ากับ 1 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 1 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับสูง

แบบสอบคณิตศาสตร์ ชุด B มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวมเท่ากับ 1 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 1 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับสูง

แบบสอบภาษาไทย ชุด A มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวมเท่ากับ 1 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 1 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับสูง

แบบสอบภาษาไทย ชุด B มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวมเท่ากับ 1 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 1 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับสูง

วิธีการของ Lee

แบบสอบคณิตศาสตร์ ชุด A มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.6368 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.5570 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับปานกลาง

แบบสอบคณิตศาสตร์ ชุด B มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.6445 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.5483 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับปานกลาง

แบบสอบภาษาไทย ชุด A มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.6223 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.5173 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับปานกลาง

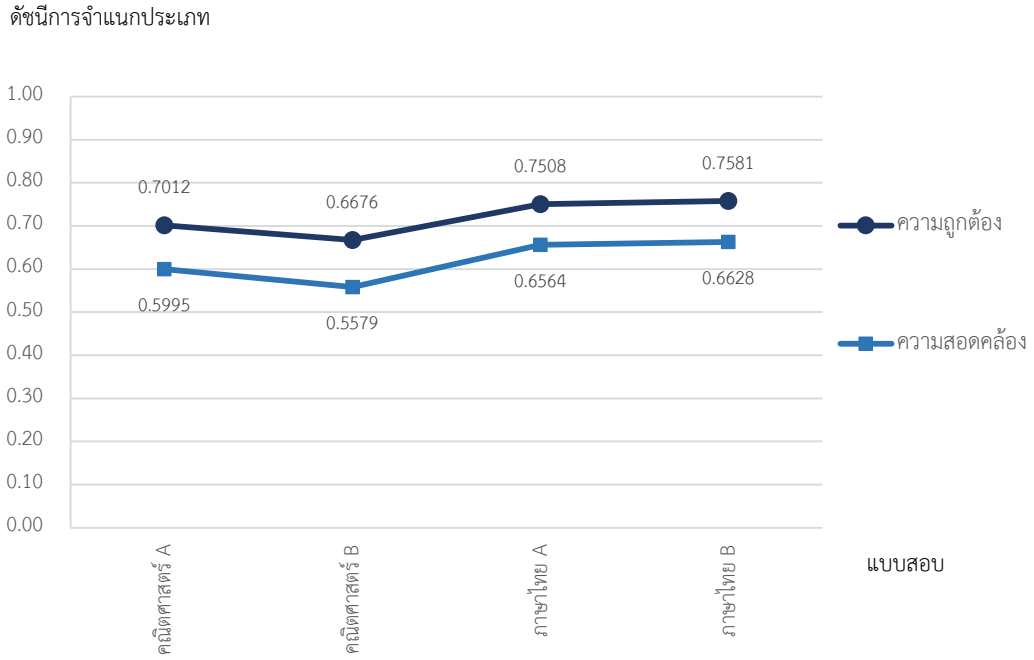
แบบสอบภาษาไทย ชุด B มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.6250 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวม เท่ากับ 0.5201 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับปานกลาง

เมื่อพิจารณาผลการประมาณค่าดัชนีการจำแนกประเภทที่ได้จากการวิเคราะห์ข้อมูลจริงจากแบบสอบทั้งสี่ฉบับ กลุ่มตัวอย่างจำนวนฉบับละ 2,000 คน รวมเป็นกลุ่มตัวอย่างทั้งสิ้น 8,000 คน และผลการประมาณค่าเฉลี่ยดัชนีการจำแนกประเภทที่ได้จากการวิเคราะห์ข้อมูลที่จำลองมาจากข้อมูลจริง โดยการทำซ้ำ 100 รอบ พบว่าค่าดัชนีการจำแนกประเภททั้งดัชนีความถูกต้องและดัชนีความสอดคล้องของการจำแนกประเภทมีค่าที่ใกล้เคียงกัน โดยรายละเอียดของค่าเฉลี่ยดัชนีการจำแนกประเภทจากการทำซ้ำ 100 รอบ ที่ประมาณค่าได้จากวิธีการตามทฤษฎีการตอบสนองข้อสอบทั้งสามวิธีแสดงได้ดังตารางที่ 4.29 และภาพที่ 4.57-4.59

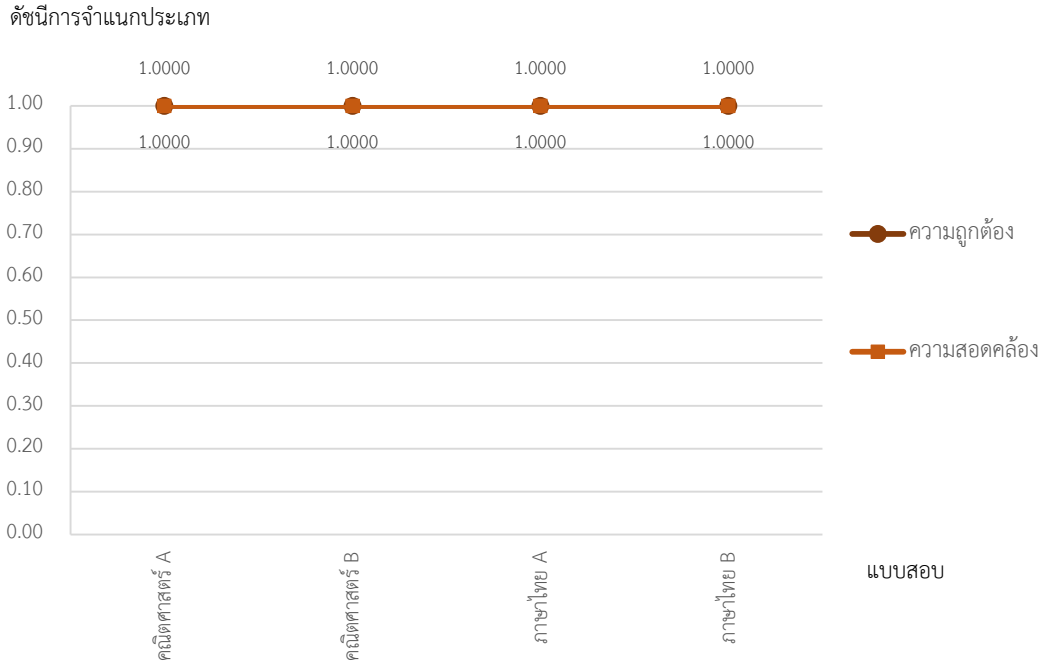
ตารางที่ 4.29 ค่าเฉลี่ยดัชนีความถูกต้องและความสอดคล้องของการจำแนกประเภทจากการทำซ้ำจำนวน 100 รอบ ในสถานการณ์จริงจำแนกตามวิธีการประมาณค่า

แบบสอบ	ตำแหน่งคะแนนจุดตัด	ความถูกต้อง (accuracy)			ความสอดคล้อง (consistency)		
		Rudner	Guo	Lee	Rudner	Guo	Lee
คณิตศาสตร์ A	1	0.8169	1.0000	0.7751	0.7473	1.0000	0.6941
	2	0.9019	1.0000	0.8448	0.8512	1.0000	0.7708
	3	0.9411	1.0000	0.9321	0.9080	1.0000	0.8960
	4	0.9752	1.0000	0.9717	0.9595	1.0000	0.9568
	5	0.9898	1.0000	0.9870	0.9838	1.0000	0.9807
	6	0.9969	1.0000	0.9957	0.9953	1.0000	0.9937
	7	0.9999	1.0000	0.9975	0.9998	1.0000	0.9966
	ทั้งหมด	0.7012	1.0000	0.6368	0.5995	1.0000	0.5570

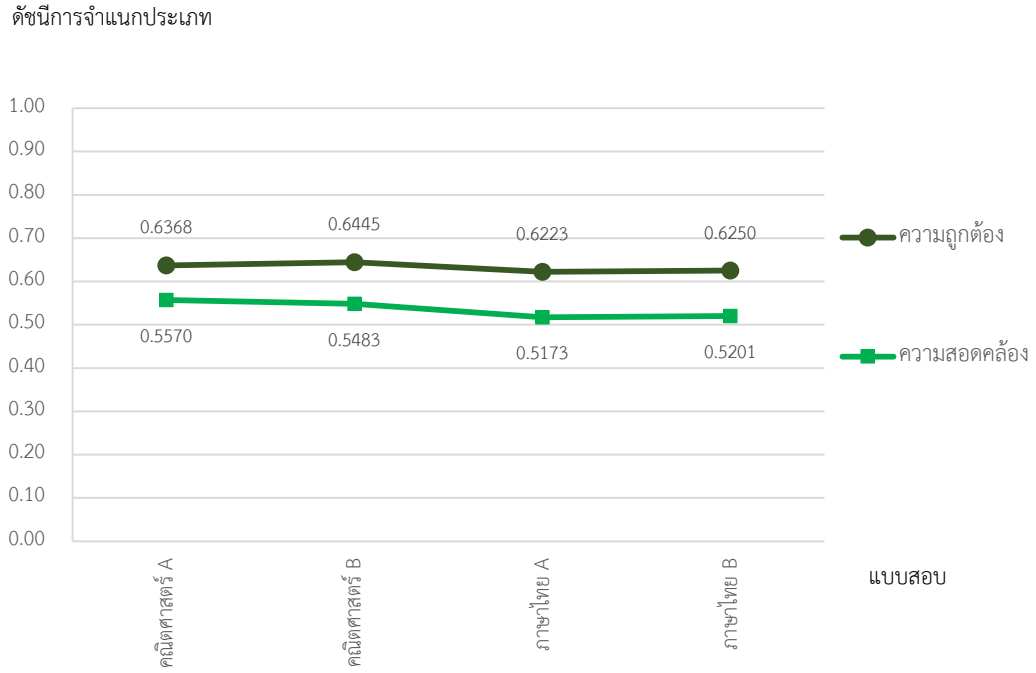
แบบสอบ	ตำแหน่งคะแนนจุดตัด	ความถูกต้อง (accuracy)			ความสอดคล้อง (consistency)		
		Rudner	Guo	Lee	Rudner	Guo	Lee
คณิตศาสตร์ B	1	0.7960	1.0000	0.7682	0.7213	1.0000	0.6739
	2	0.8801	1.0000	0.8438	0.8196	1.0000	0.7660
	3	0.9270	1.0000	0.9367	0.8847	1.0000	0.9021
	4	0.9652	1.0000	0.9734	0.9420	1.0000	0.9595
	5	0.9858	1.0000	0.9872	0.9769	1.0000	0.9811
	6	0.9963	1.0000	0.9968	0.9940	1.0000	0.9953
	7	0.9998	1.0000	0.9987	0.9996	1.0000	0.9982
	ทั้งหมด	0.6676	1.0000	0.6445	0.5579	1.0000	0.5483
ภาษาไทย A	1	0.9804	1.0000	0.9308	0.9722	1.0000	0.9029
	2	0.9303	1.0000	0.9064	0.9021	1.0000	0.8694
	3	0.9041	1.0000	0.8925	0.8645	1.0000	0.8494
	4	0.9412	1.0000	0.9152	0.9174	1.0000	0.8818
	5	0.9937	1.0000	0.9626	0.9906	1.0000	0.9456
	6	0.9996	1.0000	0.9946	0.9993	1.0000	0.9912
	7	1.0000	1.0000	0.9998	1.0000	1.0000	0.9996
	ทั้งหมด	0.7508	1.0000	0.6223	0.6564	1.0000	0.5173
ภาษาไทย B	1	0.9817	1.0000	0.9301	0.9746	1.0000	0.9020
	2	0.9418	1.0000	0.9057	0.9170	1.0000	0.8683
	3	0.9080	1.0000	0.8914	0.8698	1.0000	0.8478
	4	0.9328	1.0000	0.9173	0.9064	1.0000	0.8846
	5	0.9935	1.0000	0.9652	0.9900	1.0000	0.9492
	6	0.9995	1.0000	0.9954	0.9992	1.0000	0.9924
	7	1.0000	1.0000	0.9998	1.0000	1.0000	0.9996
	ทั้งหมด	0.7581	1.0000	0.6250	0.6628	1.0000	0.5201



ภาพที่ 4.57 ค่าเฉลี่ยดัชนีการจำแนกประเภทของข้อมูลจริงจากการทำซ้ำ 100 รอบ โดยใช้วิธีการของ Rudner



ภาพที่ 4.58 ค่าเฉลี่ยดัชนีการจำแนกประเภทของข้อมูลจริงจากการทำซ้ำ 100 รอบ โดยใช้วิธีการของ Guo



ภาพที่ 4.59 ค่าเฉลี่ยดัชนีการจำแนกประเภทของข้อมูลจริงจากการทำซ้ำ 100 รอบ โดยใช้วิธีการของ Lee

3.5 ผลการวิเคราะห์เปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทตามแนวคิดทฤษฎีการตอบสนองข้อสอบเมื่อนำไปใช้กับข้อมูลเชิงประจักษ์ (empirical data)

ตอนนี้เป็นการนำเสนอผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทตามแนวคิดทฤษฎีการตอบสนองข้อสอบทั้งสามวิธี คือ วิธีการของ Rudner (2005) วิธีการของ Guo (2006) และวิธีการของ Lee (2010) ภายใต้การศึกษากับข้อมูลเชิงประจักษ์ (empirical data) หรือข้อมูลจริง (real data) โดยใช้การวิเคราะห์ความแปรปรวน (ANOVA)

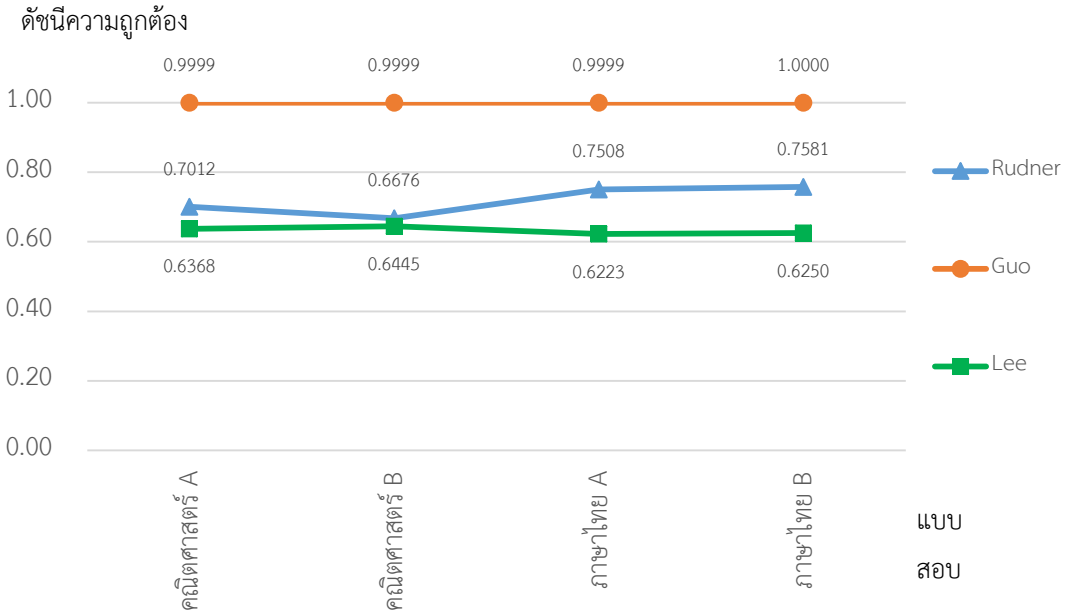
ผลการวิเคราะห์ความแตกต่างของค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทจากการทำซ้ำจำนวน 100 รอบ โดยใช้วิธีการประมาณค่าที่แตกต่างกันสามวิธีในการประมาณค่าดัชนีความถูกต้องของการจำแนกประเภท (classification accuracy) พบว่าแบบสอบทุกฉบับมีค่าดัชนีความถูกต้องที่ได้จากการประมาณค่าด้วยสามวิธีการนั้น มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 โดยวิธีการที่มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทสูงที่สุดคือวิธีการของ Guo รองลงมาคือวิธีการของ Rudner และวิธีการของ Lee ตามลำดับ รายละเอียดของผลการวิเคราะห์แสดงได้ดังตารางที่ 4.30 และแผนภาพที่ 4.60

ในส่วนของค่าดัชนีความสอดคล้องของการจำแนกประเภทนั้น (classification consistency) พบว่า แบบสอบทุกฉบับมีค่าดัชนีความสอดคล้องที่ได้จากการประมาณค่าด้วยสามวิธีการตามทฤษฎีการตอบสนองข้อสอบก็เป็นไปในแนวทางเดียวกัน คือมีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 โดยวิธีการที่มีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทสูงที่สุดคือวิธีการของ Guo รองลงมาคือวิธีการของ Rudner และวิธีการของ Lee ตามลำดับ รายละเอียดของผลการวิเคราะห์แสดงได้ดังตารางที่ 4.31 และแผนภาพที่ 4.61

ตารางที่ 4.30 ผลการวิเคราะห์ความแตกต่างของค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภท (classification accuracy) จากการทำซ้ำจำนวน 100 รอบ ในสถานการณ์จริงจำแนกตามวิธีการประมาณค่า

	Rudner		Guo		Lee				
แบบสอบ	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**F**	**p-value**	**Post Hoc**
คณิตศาสตร์ ชุด A	0.7012	0.0137	0.9999	0.0000	0.6368	0.0058	50,878.31*	.000	Guo>Rud(.000) Rud>Lee(.000)
คณิตศาสตร์ ชุด B	0.6676	0.0157	0.9999	0.0001	0.6445	0.0057	42,585.75*	.000	Guo>Rud(.000) Rud>Lee(.000)
ภาษาไทย ชุด A	0.7508	0.0069	0.9999	0.0000	0.6223	0.0034	185,116.56*	.000	Guo>Rud(.000) Rud>Lee(.000)
ภาษาไทย ชุด B	0.7581	0.0042	1.0000	0.0000	0.625	0.0038	337,389.74*	.000	Guo>Rud(.000) Rud>Lee(.000)

หมายเหตุ : *p < .05

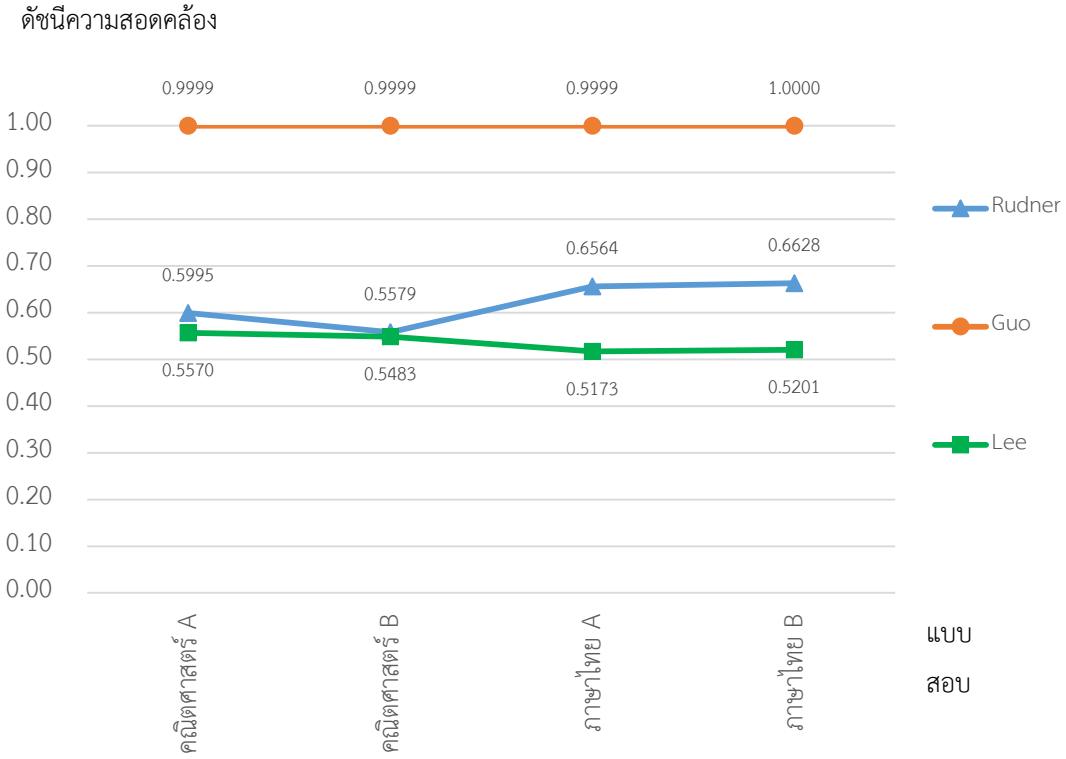


ภาพที่ 4.60 กราฟเปรียบเทียบค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทของข้อมูลจริงจากการทำซ้ำ 100 รอบโดยใช้วิธีการประมาณค่า 3 วิธี

ตารางที่ 4.31 ผลการวิเคราะห์ความแตกต่างของค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภท (classification consistency) จากการทำซ้ำจำนวน 100 รอบ ในสถานการณ์จริงจำแนกตามวิธีการประมาณค่า

	Rudner		Guo		Lee				
แบบสอบ	Mean	SD	Mean	SD	Mean	SD	F	p-value	Post Hoc
คณิตศาสตร์ ชุด A	0.5995	0.0135	0.9999	0.0000	0.5570	0.0043	89,195.85*	.000	Guo>Rud(.000) Rud>Lee(.000)
คณิตศาสตร์ ชุด B	0.5579	0.0209	0.9999	0.0001	0.5483	0.0039	44,246.06*	.000	Guo>Rud(.000) Rud>Lee(.000)
ภาษาไทย ชุด A	0.6564	0.0067	0.9999	0.0000	0.5173	0.0037	311,897.97*	.000	Guo>Rud(.000) Rud>Lee(.000)
ภาษาไทย ชุด B	0.6628	0.0048	1.0000	0.0000	0.5201	0.0038	476,765.78*	.000	Guo>Rud(.000) Rud>Lee(.000)

หมายเหตุ : *p < .05



ภาพที่ 4.61 กราฟเปรียบเทียบค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทของข้อมูลจริงจากการทำซ้ำ 100 รอบโดยใช้วิธีการประมาณค่า 3 วิธี

บทที่ 5
สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อประมาณค่าดัชนีการจำแนกประเภทโดยใช้วิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบสามวิธี ได้แก่ วิธีการของ Rudner (2005) วิธีการของ Guo (2006) และวิธีการของ Lee (2010) ด้วยการจำลองข้อมูลภายใต้เงื่อนไขของการศึกษา และทำการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภททั้งสามวิธี โดยพิจารณาจากค่าเฉลี่ยของดัชนีการจำแนกประเภทจากการทำซ้ำ 100 รอบ นอกจากนี้ยังทำการเพื่อประมาณค่าและเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบทั้งสามวิธี เมื่อใช้กับข้อมูลเชิงประจักษ์หรือข้อมูลจริง

วิธีดำเนินการวิจัยเป็นแบบการศึกษาการจำลองข้อมูล (simulation study) โดยทำการจำลองข้อมูลด้วยโปรแกรม R และทำการวิเคราะห์ข้อมูลภายใต้เงื่อนไขของการศึกษา เพื่อให้ได้ข้อสรุปเกี่ยวกับประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทตามแนวคิดทฤษฎีการตอบสนองข้อสอบทั้งสามวิธี การนำเสนอสรุปผลการวิจัยแบ่งออกเป็น 3 ตอน ตอนแรกเป็นการสรุปผลการประมาณค่าดัชนีการจำแนกประเภทโดยใช้วิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบสามวิธี คือ วิธีการของ Rudner (2005) วิธีการของ Guo (2006) และวิธีการของ Lee (2010) ด้วยการจำลองข้อมูลภายใต้เงื่อนไขของการศึกษา เพื่อตอบวัตถุประสงค์การวิจัยข้อที่ 1 ตอนที่ 2 นำเสนอสรุปผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบทั้งสามวิธีภายใต้เงื่อนไขของการศึกษา เพื่อตอบวัตถุประสงค์การวิจัยข้อที่ 2 และตอนสุดท้ายนำเสนอสรุปผลการประมาณค่าและผลการเปรียบเทียบประสิทธิภาพของดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบทั้งสามวิธี เมื่อนำไปใช้กับข้อมูลเชิงประจักษ์ (empirical data) หรือข้อมูลการตอบข้อสอบจริง (real data) ของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ในปีการศึกษา 2556 เพื่อตอบวัตถุประสงค์การวิจัยข้อที่ 3 ซึ่งสามารถสรุปผลการวิจัยในแต่ละตอนได้ดังนี้

สรุปผลการวิจัย

การสรุปผลการวิจัยในครั้งนี้เป็นการสรุปตามวัตถุประสงค์การวิจัย โดยมีรายละเอียดดังนี้

1. ผลการประมาณค่าดัชนีการจำแนกประเภทภายใต้การศึกษาจากการจำลองข้อมูล

1.1 ค่าดัชนีการจำแนกประเภทที่ได้จากวิธีการประมาณค่าตามทฤษฎีการตอบสนองข้อสอบสามวิธีด้วยการศึกษาการจำลองข้อมูลภายใต้สถานการณ์เงื่อนไขทั้งหมด สรุปได้ดังนี้

1.1.1 วิธีการของ Rudner (2005) พบว่า ดัชนีความถูกต้องของการจำแนกประเภทในภาพรวมพบว่า ค่าเฉลี่ยอยู่ในช่วง 0.8234-0.9086 โดยสถานการณ์ที่ 6 (253PL20) มีค่าดัชนีความถูกต้องสูงสุด และสถานการณ์ที่ 4 (252PL20) มีค่าดัชนีความถูกต้องต่ำสุด ดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมพบว่า ค่าเฉลี่ยอยู่ในช่วง 0.7550-0.8749 โดยสถานการณ์ที่ 6 (253PL20) มีค่าดัชนีความสอดคล้องสูงสุด และสถานการณ์ที่ 4 (252PL20) มีค่าดัชนีความสอดคล้องต่ำสุด

1.1.2 วิธีการของ Guo (2006) พบว่า ดัชนีความถูกต้องของการจำแนกประเภทในภาพรวมพบว่า ค่าเฉลี่ยอยู่ในช่วง 0.9987- 1.0000 โดยสถานการณ์ที่ 7 (501PL10) มีค่าดัชนีความถูกต้องสูงสุด และสถานการณ์ที่ 6 (253PL20) มีค่าดัชนีความถูกต้องต่ำสุด ดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมพบว่า ค่าเฉลี่ยอยู่ในช่วง 0.9982- 1.0000 โดยสถานการณ์ที่ 7 (501PL10) มีค่าดัชนีความสอดคล้องสูงสุด และสถานการณ์ที่ 6 (253PL20) มีค่าดัชนีความสอดคล้องต่ำสุด

1.1.3 วิธีการของ Lee (2010) พบว่า ดัชนีความถูกต้องของการจำแนกประเภทในภาพรวมพบว่า ค่าเฉลี่ยอยู่ในช่วง 0.6285- 0.7496 โดยสถานการณ์ที่ 1 (251PL10) มีค่าดัชนีความถูกต้องสูงสุด และสถานการณ์ที่ 12 (503PL20) มีค่าดัชนีความถูกต้องต่ำสุด ดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมพบว่า ค่าเฉลี่ยอยู่ในช่วง 0.5372- 0.6938 โดยสถานการณ์ที่ 1 (251PL10) มีค่าดัชนีความสอดคล้องสูงสุดและสถานการณ์ที่ 12 (503PL20) มีค่าดัชนีความสอดคล้องต่ำสุด

1.2 ปัจจัยที่มีอิทธิพลต่อค่าดัชนีการจำแนกประเภทที่ได้จากวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบสามวิธีด้วยการศึกษาการจำลองข้อมูลภายใต้สถานการณ์เงื่อนไขทั้งหมด สรุปได้ดังนี้

1.2.1 วิธีการของ Rudner (2005) เมื่อพิจารณาปัจจัยที่ส่งผลต่อดัชนีความถูกต้องของการจำแนกประเภท (classification accuracy) พบว่า ปัจจัยทั้งด้านความยาวของแบบสอบ โมเดลการวัด และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ มีอิทธิพลร่วมกันต่อค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภท ซึ่งทำให้ค่าเฉลี่ยของดัชนีความถูกต้องของการจำแนกประเภท

แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 (sig=.000) โดยมีอิทธิพลอยู่ในระดับมากด้วยขนาดอิทธิพล .602

ในส่วนของดัชนีความสอดคล้องของการจำแนกประเภท (classification consistency) พบว่า ปัจจัยทั้งด้านความยาวของแบบสอบ โมเดลการวัด และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ มีอิทธิพลร่วมกันต่อค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภท ซึ่งทำให้ค่าเฉลี่ยของดัชนีความถูกต้องของการจำแนกประเภทแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 (sig=.000) โดยมีอิทธิพลอยู่ในระดับมากด้วยขนาดอิทธิพล .610

1.2.2 วิธีการของ Guo (2006) เมื่อพิจารณาปัจจัยที่ส่งผลต่อดัชนีความถูกต้องของการจำแนกประเภท (classification accuracy) พบว่า ปัจจัยทั้งด้านความยาวของแบบสอบ โมเดลการวัด และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ มีอิทธิพลร่วมกันต่อค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภท ซึ่งทำให้ค่าเฉลี่ยของดัชนีความถูกต้องของการจำแนกประเภทแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 (sig=.000) โดยมีอิทธิพลอยู่ในระดับมากด้วยขนาดอิทธิพล .129

ในส่วนของดัชนีความสอดคล้องของการจำแนกประเภท (classification consistency) พบว่า ปัจจัยทั้งด้านความยาวของแบบสอบ โมเดลการวัด และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ มีอิทธิพลร่วมกันต่อค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภท ซึ่งทำให้ค่าเฉลี่ยของดัชนีความถูกต้องของการจำแนกประเภทแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 (sig=.000) โดยมีอิทธิพลอยู่ในระดับมากด้วยขนาดอิทธิพล .100

1.2.3 วิธีการของ Lee (2010) เมื่อพิจารณาปัจจัยที่ส่งผลต่อดัชนีความถูกต้องของการจำแนกประเภท (classification accuracy) พบว่า ปัจจัยทั้งด้านความยาวของแบบสอบ โมเดลการวัด และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ มีอิทธิพลร่วมกันต่อค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภท ซึ่งทำให้ค่าเฉลี่ยของดัชนีความถูกต้องของการจำแนกประเภทแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 (sig=.000) โดยมีอิทธิพลอยู่ในระดับมากด้วยขนาดอิทธิพล .821

ในส่วนของดัชนีความสอดคล้องของการจำแนกประเภท (classification consistency) พบว่า ปัจจัยทั้งด้านความยาวของแบบสอบ โมเดลการวัด และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ มีอิทธิพลร่วมกันต่อค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภท ซึ่งทำให้ค่าเฉลี่ยของดัชนีความถูกต้องของการจำแนกประเภทแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 (sig=.000) โดยมีอิทธิพลอยู่ในระดับมากด้วยขนาดอิทธิพล .863

1.3 เมื่อนำผลการทดสอบขนาดอิทธิพลของปัจจัยที่มีอิทธิพลต่อค่าเฉลี่ยดัชนีการจำแนกประเภทของวิธีการประมาณค่าตามทฤษฎีการตอบสนองข้อสอบสามวิธีจากการศึกษาการจำลองข้อมูล (simulation study) มาวาดกราฟเพื่อพิจารณาว่าในแต่ละปัจจัยนั้นเงื่อนไขใดมีค่าเฉลี่ยดัชนีการจำแนกประเภทสูงสุด โดยสามารถสรุปได้ดังนี้

1.3.1 วิธีการของ Rudner จะมีค่าเฉลี่ยดัชนีความถูกต้องและความสอดคล้องของการจำแนกประเภทสูง เมื่อแบบสอบมีลักษณะเป็นแบบสอบสั้น (25 ข้อ) ภายใต้โมเดลการวัดแบบ 3PL และมีความไม่เหมาะสมของโมเดลการวัดกับข้อสอบร้อยละ 10

1.3.2 วิธีการของ Guo จะมีค่าเฉลี่ยดัชนีความถูกต้องและความสอดคล้องของการจำแนกประเภทสูง เมื่อแบบสอบมีลักษณะเป็นแบบสอบยาว (50 ข้อ) ภายใต้โมเดลการวัดแบบ 1PL และมีความไม่เหมาะสมของโมเดลการวัดกับข้อสอบร้อยละ 20

1.3.3 วิธีการของ Lee จะมีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทสูง เมื่อแบบสอบมีลักษณะเป็นแบบสอบยาว (50 ข้อ) ภายใต้โมเดลการวัดแบบ 1PL และมีความไม่เหมาะสมของโมเดลการวัดกับข้อสอบร้อยละ 10 และค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทสูง เมื่อแบบสอบมีลักษณะเป็นแบบสอบสั้น (25 ข้อ) ภายใต้โมเดลการวัดแบบ 1PL และมีความไม่เหมาะสมของโมเดลการวัดกับข้อสอบร้อยละ 10

2. ผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบสามวิธี ภายใต้การศึกษาจากการจำลองข้อมูล

2.1 ผลการวิเคราะห์ความแตกต่างของค่าเฉลี่ยดัชนีการจำแนกประเภทจากการทำซ้ำจำนวน 100 รอบ เพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบสามวิธี ภายใต้การศึกษาจากการจำลองข้อมูลสรุปได้ดังนี้

เมื่อพิจารณาค่าดัชนีความถูกต้องของการจำแนกประเภท (classification accuracy) พบว่าค่าเฉลี่ยดัชนีความถูกต้องที่ได้จากการประมาณค่าด้วยทั้งสามวิธีการนั้นมีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ทุกสถานการณ์เงื่อนไข โดยวิธีการที่มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทสูงที่สุดคือวิธีการของ Guo รองลงมาคือวิธีการของ Rudner และวิธีการของ Lee ตามลำดับ

สำหรับค่าดัชนีความสอดคล้องของการจำแนกประเภท (classification consistency) ที่ได้จากการประมาณค่าด้วยทั้งสามวิธีการนั้นก็เป็นไปในแนวทางเดียวกันคือ ค่าเฉลี่ยดัชนีความสอดคล้องมีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ทุกสถานการณ์เงื่อนไข โดยวิธีการที่มีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทสูงที่สุดคือวิธีการของ Guo รองลงมาคือวิธีการของ Rudner และวิธีการของ Lee ตามลำดับ

2.2 ปัจจัยที่มีอิทธิพลต่อค่าเฉลี่ยดัชนีการจำแนกประเภทจากการศึกษาการจำลองข้อมูลภายใต้สถานการณ์เงื่อนไขทั้งหมด สรุปได้ดังนี้

เมื่อพิจารณาปัจจัยที่มีอิทธิพลต่อดัชนีความถูกต้องของการจำแนกประเภท (classification accuracy) พบว่า ปัจจัยทั้งด้านวิธีการประมาณค่า ความยาวของแบบสอบ โมเดลการวัด และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ มีอิทธิพลร่วมกันต่อค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภท ซึ่งทำให้ค่าเฉลี่ยของดัชนีความถูกต้องของการจำแนกประเภทแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 (sig=.000) โดยมีอิทธิพลอยู่ในระดับมากด้วยขนาดอิทธิพล .624

ในส่วนของดัชนีความสอดคล้องของการจำแนกประเภท (classification consistency) พบว่า ปัจจัยทั้งด้านวิธีการประมาณค่า ความยาวของแบบสอบ โมเดลการวัด และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ มีอิทธิพลร่วมกันต่อค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภท ซึ่งทำให้ค่าเฉลี่ยของดัชนีความถูกต้องของการจำแนกประเภทแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 (sig=.000) โดยมีอิทธิพลอยู่ในระดับมากด้วยขนาดอิทธิพล .656

เมื่อพิจารณาขนาดอิทธิพลของปัจจัยที่มีอิทธิพลต่อค่าเฉลี่ยดัชนีการจำแนกประเภท เพื่อเปรียบเทียบว่าปัจจัยที่กำหนดขึ้นมีอิทธิพลต่อค่าเฉลี่ยดัชนีการจำแนกประเภทที่ประมาณค่าได้จากวิธีการใดมากที่สุดและน้อยที่สุด อันจะนำไปสู่การอธิบายถึงความแข็งแกร่งของวิธีการประมาณค่าดัชนีการจำแนกประเภท โดยผลการวิเคราะห์ความแปรปรวนแบบสี่ทาง (4-WAY ANOVA) ซึ่งพิจารณาจาก 4 ปัจจัย คือ วิธีการประมาณค่า ความยาวของแบบสอบ โมเดลการวัด และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ พบว่า ปัจจัยทั้งสี่มีอิทธิพลร่วมกันต่อค่าเฉลี่ยดัชนีความถูกต้องและดัชนีความสอดคล้องของการจำแนกประเภทโดยมีอิทธิพลอยู่ในระดับมากด้วยขนาดอิทธิพล .624 และ .656 ตามลำดับ และผลการวิเคราะห์ความแปรปรวนแบบสามทาง (3-WAY ANOVA) ซึ่งพิจารณาจาก 3 ปัจจัย คือ ความยาวของแบบสอบ โมเดลการวัด และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ พบว่า ปัจจัยทั้งสามมีอิทธิพลร่วมกันต่อค่าเฉลี่ยดัชนีความถูกต้องและดัชนีความสอดคล้องของการจำแนกประเภทที่ประมาณค่าได้จากวิธีการของ Lee มากที่สุด ด้วยขนาดอิทธิพล .821 และ .863 ตามลำดับ รองลงมาคือวิธีการของ Rudner ด้วยขนาดอิทธิพล .602 และ .610 ตามลำดับ และมีอิทธิพลต่อวิธีการของ Guo น้อยที่สุดด้วยขนาดอิทธิพล .129 และ .100 ตามลำดับ ซึ่งแสดงให้เห็นว่าวิธีการของ Guo มีความแข็งแกร่งมากที่สุด

3. ผลการประมาณค่าและผลการเปรียบเทียบประสิทธิภาพของดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบทั้งสามวิธี เมื่อนำไปใช้กับข้อมูลเชิงประจักษ์ (empirical data) หรือข้อมูลการตอบข้อสอบจริง (real data)

3.1 ผลการประมาณค่าดัชนีการจำแนกประเภทที่ได้จากวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบสามวิธีภายใต้สถานการณ์จริงทั้งหมด สรุปได้ดังนี้

3.1.1 วิธีการของ Rudner (2005) พบว่า แบบสอบคณิตศาสตร์ ชุด A มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวม เท่ากับ 0.7012 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.5995 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับปานกลาง

แบบสอบคณิตศาสตร์ ชุด B มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวม เท่ากับ 0.6676 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.5579 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับปานกลาง

แบบสอบภาษาไทย ชุด A มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวม เท่ากับ 0.7508 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.6564 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับค่อนข้างสูง

แบบสอบภาษาไทย ชุด B มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวม เท่ากับ 0.7581 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.6628 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับค่อนข้างสูง

3.1.2 วิธีการของ Guo (2006) พบว่า แบบสอบคณิตศาสตร์ ชุด A มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวม เท่ากับ 1 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 1 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับสูง

แบบสอบคณิตศาสตร์ ชุด B มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวม เท่ากับ 1 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 1 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับสูง

แบบสอบภาษาไทย ชุด A มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวม เท่ากับ 1 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 1 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับสูง

แบบสอบภาษาไทย ชุด B มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวม เท่ากับ 1 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 1 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับสูง

3.1.3 วิธีการของ Lee (2010) พบว่า แบบสอบคณิตศาสตร์ ชุด A มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวม เท่ากับ 0.6368 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.5570 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับปานกลาง

แบบสอบคณิตศาสตร์ ชุด B มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวม เท่ากับ 0.6445 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.5483 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับปานกลาง

แบบสอบภาษาไทย ชุด A มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวม เท่ากับ 0.6223 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวมเท่ากับ 0.5173 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับปานกลาง

แบบสอบภาษาไทย ชุด B มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทในภาพรวม เท่ากับ 0.6250 และมีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทในภาพรวม เท่ากับ 0.5201 แสดงว่าแบบสอบฉบับนี้มีดัชนีความถูกต้องและดัชนีความสอดคล้องในระดับปานกลาง

3.2 ผลการเปรียบเทียบประสิทธิภาพของดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบทั้งสามวิธี เมื่อนำไปใช้กับข้อมูลเชิงประจักษ์ (empirical data) หรือข้อมูลการตอบข้อสอบจริง (real data) ของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ในปีการศึกษา 2556 พบว่า ค่าดัชนีความถูกต้องที่ได้จากการประมาณค่าด้วยสามวิธีการนั้น มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ทุกแบบสอบ โดยวิธีการที่มีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทสูงที่สุดคือวิธีการของ Guo รองลงมาคือวิธีการของ Rudner และวิธีการของ Lee ตามลำดับ ในส่วนของค่าดัชนีความสอดคล้องที่ได้จากการประมาณค่าด้วยสามวิธีการนั้นก็เป็นไปในแนวทางเดียวกันคือมีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ทุกแบบสอบ โดยวิธีการที่มีค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทสูงที่สุดคือวิธีการของ Guo รองลงมาคือวิธีการของ Rudner และวิธีการของ Lee ตามลำดับ แสดงให้เห็นว่าคะแนนสอบทั้งวิชาคณิตศาสตร์และภาษาไทยสามารถใช้จำแนกผู้สอบได้อย่างถูกต้องและสอดคล้องกับความสามารถที่แท้จริงของผู้สอบ

อภิปรายผลการวิจัย

การอภิปรายผลการวิจัยครั้งนี้มีประเด็นที่นำมาอภิปรายตามวัตถุประสงค์และสมมติฐานการวิจัยดังรายละเอียดต่อไปนี้

1. ผลการประมาณค่าดัชนีการจำแนกประเภทภายใต้การศึกษาการจำลอง

1.1 เมื่อพิจารณาวิธีการประมาณค่าทั้งสามวิธีตามทฤษฎีการตอบสนองข้อสอบภายใต้การศึกษาจากการจำลองข้อมูลพบว่า ค่าดัชนีความถูกต้องของการจำแนกประเภทมีค่าสูงกว่าค่าดัชนีความสอดคล้องของการจำแนกประเภทในทุกสถานการณ์ เนื่องจากค่าดัชนีความถูกต้องของการจำแนกประเภทเป็นค่าที่คำนวณได้จากการทดสอบเดียว แต่ค่าดัชนีความสอดคล้องของการจำแนกนั้นเป็นค่าที่คำนวณได้จากการทดสอบสองสถานการณ์ที่คู่ขนานกันหรือเป็นการยกกำลังสองของผลที่ได้จากการคำนวณจากการทดสอบเดียวเพื่อประมาณค่าความสอดคล้องกันของการจำแนกระหว่างชุดข้อสอบ (Wyse & Hao, 2012) ดังนั้นจึงทำให้ค่าดัชนีความถูกต้องของการจำแนกประเภทมีความเป็นไปได้ที่จะมีค่าเท่ากับหรือสูงกว่าค่าดัชนีความสอดคล้องของการจำแนกประเภทนั่นเอง

1.2 เมื่อพิจารณาถึงปัจจัยที่มีอิทธิพลต่อค่าดัชนีการจำแนกประเภทที่ได้จากวิธีการประมาณค่าดัชนีการจำแนกประเภทสามวิธีตามทฤษฎีการตอบสนองข้อสอบด้วยการศึกษาการจำลองข้อมูลภายใต้สถานการณ์เงื่อนไขทั้งหมดพบว่า ปัจจัยด้านความยาวของแบบสอบ โมเดลการวัด และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบมีอิทธิพลร่วมกันต่อค่าเฉลี่ยดัชนีความถูกต้องและดัชนีความสอดคล้องของการจำแนกประเภทที่ได้จากทุกวิธีการ โดยทำให้ค่าเฉลี่ยของดัชนีการจำแนกประเภทแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นั้นแสดงให้เห็นว่าทั้งสามปัจจัยเป็นสิ่งสำคัญที่ควรพิจารณาร่วมกันเกี่ยวกับการเลือกวิธีการประมาณค่าดัชนีการจำแนกประเภทให้เหมาะสมกับลักษณะของข้อมูลหรือแบบสอบที่จะนำมาใช้ในตรวจสอบหาค่าดัชนีการจำแนกประเภท ซึ่งสอดคล้องกับงานวิจัยของ Lathrop และ Cheng (2013) ที่ได้ทำการเปรียบเทียบดัชนีความถูกต้องของการจำแนกประเภท (classification accuracy) ที่ประมาณค่าได้จากวิธีการของ Lee กับวิธีการของ Rudner ซึ่งเป็นวิธีการประมาณค่าตามทฤษฎีการตอบสนองข้อสอบทั้งคู่ภายใต้ความยาวของแบบสอบที่แตกต่างกัน โดยทำการจำลองความยาวของแบบสอบออกเป็น 4 รูปแบบ ได้แก่ แบบสอบที่มีข้อสอบจำนวน 10, 20, 40 และ 80 ข้อ ผลการวิจัยพบว่า ความยาวของแบบสอบที่เพิ่มขึ้นจะส่งผลให้ได้ค่าการประมาณที่มีประสิทธิภาพ โดยพิจารณาจากค่าความคลาดเคลื่อนมาตรฐานของการประมาณค่า (standard error: SE) ที่ลดลง และนอกจากนี้ Lathrop และ Cheng (2013) ยังได้ทำการเปรียบเทียบดัชนีความถูกต้องของการจำแนกประเภท (classification accuracy) ภายใต้โมเดลการวัดตามทฤษฎีการตอบสนองข้อสอบ (IRT model) ที่แตกต่างกันสี่โมเดล คือ โมเดลโลจิสติกแบบหนึ่งพารามิเตอร์ (one-parameter logistic model: 1PL) โมเดลโลจิสติกแบบสองพารามิเตอร์ (two-

parameter logistic model: 2PL) โมเดลโลจิสติกแบบสามพารามิเตอร์ (three-parameter logistic model: 3PL) และ graded response model (GRM) ผลการวิจัยพบว่า วิธีการประมาณค่าทั้งสองวิธีการให้ค่าการประมาณที่แตกต่างกันเล็กน้อย ซึ่งอาจเนื่องจากโมเดลที่นำมาใช้ในการวิจัยเป็นโมเดลการวัดตามทฤษฎีการตอบสนองข้อสอบ (IRT model) และข้อมูลในการศึกษามีความสอดคล้องกับโมเดลการวิเคราะห์

1.3 เมื่อพิจารณาในแต่ละปัจจัยที่ศึกษาจากค่าเฉลี่ยดัชนีการจำแนกประเภทสูงสุดพบว่า แต่ละวิธีการจะให้ค่าเฉลี่ยดัชนีการจำแนกประเภทสูงสุดภายใต้เงื่อนไขที่แตกต่างกัน โดยวิธีการของ Rudner จะมีค่าเฉลี่ยดัชนีความถูกต้องและความสอดคล้องของการจำแนกประเภทสูง เมื่อแบบสอบมีลักษณะเป็นแบบสอบสั้น (25 ข้อ) ภายใต้โมเดลการวัดแบบ 3PL และมีจำนวนข้อสอบไม่เหมาะสมกับโมเดลการวัดร้อยละ 10 วิธีการของ Guo จะมีค่าเฉลี่ยดัชนีความถูกต้องและความสอดคล้องของการจำแนกประเภทสูง เมื่อแบบสอบมีลักษณะเป็นแบบสอบยาว (50 ข้อ) ภายใต้โมเดลการวัดแบบ 1PL และมีจำนวนข้อสอบไม่เหมาะสมกับโมเดลการวัดร้อยละ 20 ส่วนวิธีการของ Lee จะมีค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภทสูง เมื่อแบบสอบมีลักษณะเป็นแบบสอบยาว (50 ข้อ) ภายใต้โมเดลการวัดแบบ 1PL และมีจำนวนข้อสอบไม่เหมาะสมกับโมเดลการวัดร้อยละ 10 และค่าเฉลี่ยดัชนีความสอดคล้องของการจำแนกประเภทสูง เมื่อแบบสอบมีลักษณะเป็นแบบสอบสั้น (25 ข้อ) ภายใต้โมเดลการวัดแบบ 1PL และมีจำนวนข้อสอบไม่เหมาะสมกับโมเดลการวัดร้อยละ 10 ความแตกต่างที่เกิดขึ้นนี้อาจเนื่องมาจากข้อตกลงเบื้องต้นที่แตกต่างกันของวิธีการประมาณค่าแต่ละวิธี จึงทำให้แต่ละวิธีการมีความเหมาะสมกับลักษณะข้อมูลที่แตกต่างกัน

2. ผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบสามวิธี ภายใต้การศึกษาการจำลอง

2.1 เมื่อพิจารณาเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภทสามวิธีภายใต้การศึกษาจำลองพบว่า วิธีการของ Guo มีประสิทธิภาพในการประมาณค่าดัชนีการจำแนกประเภททั้งดัชนีความถูกต้องและดัชนีความสอดคล้องมากที่สุด รองลงมาคือวิธีการของ Rudner และวิธีการของ Lee ตามลำดับ ซึ่งเป็นไปตามสมมติฐานของการวิจัยที่ตั้งไว้ เนื่องมาจากวิธีการของ Guo และ Rudner ใช้วิธีการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบด้วยวิธีการประมาณค่าที่เป็นไปได้สูงสุด (Maximum Likelihood Estimate; MLE) ซึ่งเป็นวิธีการที่สามารถใช้ได้อย่างมีประสิทธิภาพกับกลุ่มตัวอย่างขนาดใหญ่เช่นเดียวกับขนาดตัวอย่างที่ใช้ในการศึกษาจำลองครั้งนี้ และยิ่งถ้าแบบสอบมีจำนวนข้อสอบที่ยาวขึ้น การประมาณค่าที่เป็นไปได้สูงสุดของค่าความสามารถของผู้สอบจะมีการแจกแจงปกติ จึงเป็นการประมาณค่าที่ไม่ลำเอียง (ศิริชัย กาญจนวาสี, 2555) ส่งผลให้ผลการประมาณค่าดัชนีความถูกต้องและความสอดคล้องของการจำแนก

ประเภทที่ได้มีประสิทธิภาพที่ดีกว่าวิธีการของ Lee ซึ่งใช้วิธีการอินทิเกรท (integrate) ในการประมาณค่าดัชนีการจำแนกประเภทนั่นเอง

นอกจากนี้การที่วิธีการของ Guo มีค่าดัชนีการจำแนกประเภทสูงกว่าวิธีการของ Rudner เนื่องมาจากวิธีการของ Guo มีการคำนวณจำนวนผู้สอบที่คาดหวังในแต่ละกลุ่มความสามารถด้วยการแจกแจงภายหลังของผู้สอบ (the normalized likelihood function) (Guo, 2006) เป็นผลให้ข้อตกลงเบื้องต้นเกี่ยวกับการแจกแจงปกติของความคลาดเคลื่อนมาตรฐานในการประมาณค่าความสามารถของผู้สอบสำหรับวิธีการของ Rudner ไม่มีความจำเป็น ถึงแม้จะจัดกระทำตัวแปรการแจกแจงของคะแนนสอบให้เป็นการแจกแจงปกติในการจำลองข้อมูลแล้วก็ตาม โดย Martineau (2007) ได้อธิบายประเด็นนี้ไว้ว่าวิธีการของ Rudner ใช้ข้อตกลงเบื้องต้นเกี่ยวกับการแจกแจงปกติของความคลาดเคลื่อนมาตรฐานในการประมาณค่าความสามารถของผู้สอบซึ่งมีความสัมพันธ์กับค่าสารสนเทศของแบบสอบ เมื่อโมเดลการวัดตามทฤษฎีการตอบสนองข้อสอบเปลี่ยนไปก็จะทำให้ค่าสารสนเทศของแบบสอบเปลี่ยนไปตามโมเดลนั้นๆ จึงทำให้วิธีการของ Guo เป็นวิธีการที่แข็งแกร่งกว่า ทั้งยังสอดคล้องกับงานวิจัยของ Lathrop และ Cheng (2013) ที่ได้ทำการเปรียบเทียบวิธีการประมาณค่าดัชนีความถูกต้องของการจำแนกประเภทสองวิธีตามทฤษฎีการตอบสนองข้อสอบ โดยพบว่าวิธีการของ Rudner มีประสิทธิภาพกว่าวิธีการของ Lee เนื่องจากการที่โมเดลมีความเหมาะสมกับข้อมูลนั้น ทำให้การจำแนกประเภทที่ได้จากการประมาณค่าคุณลักษณะแฝง (latent trait estimates) มีความถูกต้องพอกันหรือมากกว่าการจำแนกประเภทที่ได้จากการประมาณค่าด้วยคะแนนรวม (total score) ที่วิธีการของ Lee ใช้ในการคำนวณนั่นเอง

2.2 เมื่อพิจารณาดัชนีความถูกต้องและความสอดคล้องของการจำแนกประเภทพบว่า ปัจจัยทั้งสามด้าน ได้แก่ ความยาวของแบบสอบ โมเดลการวัด และความไม่เหมาะสมของโมเดลการวัดกับข้อสอบ ร่วมกันมีอิทธิพลต่อค่าเฉลี่ยดัชนีความถูกต้องของการจำแนกประเภท โดยทำให้ค่าเฉลี่ยของดัชนีความถูกต้องของการจำแนกประเภทแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 (sig=.000) ปัจจัยทั้งสามส่งผลต่อค่าดัชนีความถูกต้องและดัชนีความสอดคล้องของการจำแนกประเภทวิธีการของ Lee มากที่สุด ด้วยขนาดอิทธิพล .821 และ .863 ตามลำดับ คือวิธีการของ Rudner ด้วยขนาดอิทธิพล .602 และ .610 ตามลำดับ และส่งผลต่อวิธีการของ Guo น้อยที่สุด ด้วยขนาดอิทธิพล .129 และ .100 ตามลำดับ หรือกล่าวอีกนัยหนึ่งว่าปัจจัยทั้งสามมีอิทธิพลต่อค่าดัชนีการจำแนกประเภทที่ได้จากวิธีการของ Guo น้อยที่สุด จึงทำให้วิธีการของ Guo เป็นวิธีการที่แข็งแกร่งที่สุดเนื่องด้วยค่าดัชนีที่ได้มีการเปลี่ยนแปลงน้อยมากเมื่อปัจจัยเปลี่ยนแปลงไป ส่วนวิธีการของ Lee เป็นวิธีการที่อ่อนไหวที่สุด ซึ่งสอดคล้องกับงานวิจัยของ Wyse & Hao (2012) ที่พบว่า ไม่ว่าจะกำหนดปัจจัยให้แตกต่างกันในการศึกษาจำลองเป็นอย่างไร วิธีการของ Guo ก็ยังเป็นวิธีที่ให้ค่าดัชนีการจำแนกประเภทอยู่ในระดับสูง และงานวิจัยของ Lathrop & Cheng (2013) ที่ได้ทำการเปรียบเทียบดัชนี

ความถูกต้องของการจำแนกประเภท (classification accuracy) ที่ประมาณค่าได้จากวิธีการของ Lee กับวิธีการของ Rudner ภายใต้เงื่อนไขของปัจจัยด้านโมเดลการวัด ขนาดกลุ่มตัวอย่าง ความยาวของแบบสอบ และตำแหน่งของคะแนนจุดตัด ซึ่งพบว่าเมื่อกำหนดเงื่อนไขของแต่ละปัจจัยให้เปลี่ยนไปก็ทำให้ผลการประมาณค่าดัชนีความถูกต้องของการจำแนกประเภทที่ได้จากวิธีการของ Lee และ Rudner เปลี่ยนแปลงตามไปด้วย

3. ผลการประมาณค่าและหาประสิทธิภาพของดัชนีการจำแนกประเภทตามทฤษฎีการตอบสนองข้อสอบที่ได้จากการจำลองข้อมูล เมื่อนำไปใช้กับข้อมูลเชิงประจักษ์ (empirical data) หรือข้อมูลการตอบข้อสอบจริง (real data)

3.1 เมื่อพิจารณาถึงค่าดัชนีการจำแนกประเภททั้งดัชนีความถูกต้องและดัชนีความสอดคล้องที่ได้จากวิธีการของ Guo พบว่าแบบสอบทั้งวิชาคณิตศาสตร์และวิชาภาษาไทยมีค่าดัชนีการจำแนกประเภทอยู่ในระดับสูง จึงสรุปได้ว่าคะแนนจุดตัดที่ทางสำนักทดสอบทางการศึกษาแห่งชาติกำหนดขึ้นมานั้นสามารถนำมาใช้ในการจำแนกประเภทได้อย่างถูกต้องและสามารถใช้จำแนกผู้สอบได้สอดคล้องกับความสามารถที่แท้จริง อันเนื่องมาจากวิธีการที่ใช้ในการกำหนดคะแนนจุดตัดเป็นวิธีการที่มีประสิทธิภาพและรัดกุม โดยมีขั้นตอนในการกำหนดคะแนนจุดตัดดังนี้ 1) กำหนดระดับคะแนนเป็น 8 ระดับ 2) กำหนดช่วงคะแนนในแต่ละระดับด้วยวิธี Normalized T-Score 3) กำหนดเกณฑ์คะแนนต่ำสุดระดับผ่านหรือระดับ 1 ที่ควรสูงกว่าคะแนนค่าของโอกาสการเดา เช่น แบบสอบปรนัยแบบ 4 ตัวเลือก คะแนนเต็ม 100 คะแนน เกณฑ์คะแนนต่ำสุดระดับผ่านควรสูงกว่า 25 คะแนน 4) กำหนดเกณฑ์คะแนนต่ำสุดที่ได้รับ 4 ควรมีคะแนนตั้งแต่ร้อยละ 80 และ 5) ช่วงคะแนนในแต่ละระดับแต่ละวิชาจะไม่กำหนดคงที่ โดยจะผันแปรไปตามการกระจายของคะแนนวิชานั้นๆ และระดับความยากง่ายของข้อสอบ (สถาบันทดสอบทางการศึกษาแห่งชาติ, 2556) หรือกล่าวอีกนัยหนึ่งคือการกำหนดคะแนนจุดตัดด้วยวิธีการ Normalized T-Score น่าจะเป็นวิธีการกำหนดคะแนนจุดตัดที่เหมาะสมกับการตรวจสอบดัชนีการจำแนกประเภทด้วยวิธีการของ Guo

3.2 เมื่อพิจารณาผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าดัชนีการจำแนกประเภททั้งสามวิธีเมื่อนำมาใช้กับข้อมูลเชิงประจักษ์ (empirical data) หรือข้อมูลการตอบข้อสอบจริง (real data) ของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ในปีการศึกษา 2556 พบว่า วิธีการของ Guo มีประสิทธิภาพในการประมาณค่าทั้งดัชนีความถูกต้องและดัชนีความสอดคล้องของการจำแนกประเภทสูงที่สุด เนื่องมาจากในการประมาณค่าความน่าจะเป็นในการจำแนกผู้สอบเข้าสู่แต่ละระดับความสามารถนั้น วิธีการของ Guo ใช้ฟังก์ชันความน่าจะเป็นในการตอบข้อสอบของผู้สอบ โดยมีข้อตกลงเบื้องต้นว่าฟังก์ชันความน่าจะเป็นในการตอบข้อสอบของผู้สอบต้องมีการแจกแจงแบบปกติ ซึ่งจากการตรวจสอบการแจกแจงของคะแนนความสามารถที่แท้จริงของผู้สอบเบื้องต้นพบว่าเป็นการ

แจกแจงปกติ จึงไม่เป็นการฝ่าฝืนข้อตกลงเบื้องต้นนี้ตามวิธีการของ Guo แต่กลับเป็นการฝ่าฝืนข้อตกลงเบื้องต้นตามวิธีการของ Lee ที่มีข้อตกลงเบื้องต้นว่าคะแนนจริงต้องมีลักษณะเป็นการแจกแจงแบบทวินาม อีกทั้งผลการตรวจสอบการแจกแจงของความคลาดเคลื่อนมาตรฐานในการประมาณค่าความสามารถของผู้สอบเบื้องต้นพบว่าไม่เป็นการแจกแจงปกติ จึงเป็นการฝ่าฝืนข้อตกลงเบื้องต้นตามวิธีการของ Rudner (Rudner, 2005) นอกจากนี้การใช้ฟังก์ชันความน่าจะเป็นในการตอบข้อสอบของผู้สอบในการประมาณค่าความน่าจะเป็นที่คาดหวังนั้น เป็นการใช้ฟังก์ชันตัวเดียวกันกับที่ใช้ในการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบ ทำให้ค่าความน่าจะเป็นที่คาดหวังในการจำแนกผู้สอบเข้าสู่แต่ละระดับความสามารถมีค่าที่ใกล้เคียงจนแทบจะเป็นค่าเดียวกันกับความสามารถที่แท้จริงของผู้สอบ (Guo, 2006) ดังนั้นจึงทำให้ค่าดัชนีที่ได้จากวิธีการของ Guo มีค่าที่ค่อนข้างสูงกว่าวิธีการอื่น

ข้อเสนอแนะจากการวิจัย

1. ข้อเสนอแนะในการนำผลวิจัยไปใช้

1.1 การพิจารณาเลือกวิธีการประมาณค่าดัชนีการจำแนกประเภทที่เหมาะสมเพื่อนำไปประยุกต์ใช้ในทางปฏิบัตินั้น จะต้องคำนึงถึงข้อตกลงเบื้องต้นของแต่ละวิธีการด้วย โดยวิธีการของ Rudner ต้องคำนึงถึงการแจกแจงของความคลาดเคลื่อนในการประมาณค่าความสามารถผู้สอบว่าต้องเป็นการแจกแจงปกติ วิธีการของ Guo ต้องคำนึงถึงการแจกแจงของฟังก์ชันความน่าจะเป็นในการตอบข้อสอบของผู้สอบว่าต้องเป็นการแจกแจงปกติ และวิธีการของ Lee ต้องคำนึงถึงการแจกแจงของคะแนนสอบว่าต้องเป็นการแจกแจงแบบอเนกนาม ดังนั้นนอกจากที่จะพิจารณาถึงประสิทธิภาพของวิธีการต่างๆ ที่ได้จากการศึกษาครั้งนี้แล้วจึงควรพิจารณาถึงลักษณะของข้อมูลที่จะนำมาใช้ร่วมด้วยว่ามีลักษณะที่เป็นไปตามข้อตกลงเบื้องต้นของวิธีการใดมากที่สุด หรือมีการฝ่าฝืนข้อตกลงเบื้องต้นของวิธีการใดบ้าง และถ้าพบว่ามีการฝ่าฝืนข้อตกลงเบื้องต้นของวิธีการใดวิธีการหนึ่งก็ควรเลี่ยงที่จะไม่ใช้วิธีการนั้นในการประมาณค่าดัชนีการจำแนกประเภท

1.2 การพิจารณาเลือกวิธีการประมาณค่าดัชนีการจำแนกประเภทไปใช้ในทางปฏิบัตินั้นจะต้องพิจารณาถึงประสิทธิภาพของแต่ละวิธีการร่วมกับปัจจัยด้านความยาวของแบบสอบ โมเดลการวัด และระดับความไม่เหมาะสมของข้อมูลกับโมเดลด้วย เนื่องจากข้อค้นพบจากการวิจัยที่ว่าปัจจัยเหล่านี้มีปฏิสัมพันธ์ร่วมกันมีอิทธิพลต่อค่าดัชนีการจำแนกประเภท ทั้งดัชนีความสอดคล้องและดัชนีความถูกต้องของการจำแนกประเภท เพื่อทำให้คะแนนจุดตัดที่กำหนดขึ้นมานั้นสามารถจำแนกผู้สอบได้ถูกต้องและสอดคล้องกับความสามารถที่แท้จริง

1.3 เนื่องด้วยจากผลการวิเคราะห์ข้อมูลเบื้องต้นพบว่าข้อมูลที่นำมาใช้ในการวิจัยครั้งนี้มีการแจกแจงแบบปกติซึ่งเป็นไปตามข้อตกลงเบื้องต้นตามวิธีการของ Guo แต่ในทางกลับกันค่าความคลาดเคลื่อนมาตรฐานในการประมาณค่าคะแนนจริงมีการแจกแจงแบบไม่ปกติซึ่งทำให้เป็นการฝ่าฝืนข้อตกลงเบื้องต้นตามวิธีการของ Rudner จึงทำให้ผลการประมาณค่าดัชนีด้วยวิธีการของ Guo มีค่าสูงกว่าวิธีการอื่น ดังนั้นในการนำวิธีการประมาณค่าวิธีใดก็ตามไปใช้กับข้อมูลจริงในสถานการณ์อื่นจึงต้องคำนึงถึงข้อตกลงเบื้องต้นเกี่ยวกับลักษณะการแจกแจงของข้อมูลประกอบกับการเลือกใช้วิธีการประมาณค่าดัชนีการจำแนกประเภทด้วย

2. ข้อเสนอแนะในการวิจัยครั้งต่อไป

2.1 การศึกษาเชิงจำลองครั้งนี้เป็นการศึกษากับโมเดลการวัดตามทฤษฎีการตอบสนองข้อสอบแบบตรวจให้คะแนนสองค่า ซึ่งในเชิงทฤษฎียังมีโมเดลการวัดตามทฤษฎีการตอบสนองข้อสอบแบบตรวจให้คะแนนมากกว่าสองค่าด้วย ดังนั้นในการศึกษาครั้งต่อไปอาจทำการศึกษาถึงวิธีการประมาณค่าดัชนีการจำแนกประเภทที่ใช้กับโมเดลการตอบสนองข้อสอบแบบตรวจให้คะแนนมากกว่าสองค่า เพื่อเป็นประโยชน์สำหรับการศึกษาในเชิงทฤษฎีต่อไป

2.2 การศึกษาเชิงจำลองครั้งนี้ได้ข้อค้นพบเกี่ยวกับการเลือกใช้วิธีการประมาณค่าดัชนีการจำแนกประเภทโดยพิจารณาจากค่าเฉลี่ยดัชนีการจำแนกประเภทในการทำซ้ำ 100 รอบ ซึ่งสารสนเทศที่ได้คือวิธีการประมาณค่าที่ให้ค่าดัชนีการจำแนกประเภทสูงที่สุดเพื่อนำมาใช้ในการประมาณค่าดัชนีความถูกต้องและความสอดคล้องของการจำแนกประเภท ในการศึกษาครั้งต่อไปอาจทำการศึกษาเพิ่มเติมในส่วนของความลำเอียงในการประมาณค่าและความคลาดเคลื่อนมาตรฐานของการประมาณค่าดัชนีการจำแนกประเภท เพื่อให้ได้มาซึ่งสารสนเทศที่มากขึ้นสำหรับใช้ประกอบการตัดสินใจและมีประโยชน์ต่อการตัดสินใจเลือกใช้วิธีการประมาณค่าดัชนีการจำแนกประเภทต่อไป

2.3 การศึกษาครั้งนี้เป็นการประมาณค่าดัชนีการจำแนกประเภทของการตัดสินจากคะแนนสอบของนักเรียน โดยพิจารณาจากคะแนนการทดสอบมาตรฐานระดับชาติขั้นพื้นฐาน (o-net) ซึ่งเป็นผลการเรียนรู้ด้านพุทธิพิสัยของนักเรียนเพียงอย่างเดียว หรือเป็นโมเดลการตอบสนองข้อสอบแบบมิติเดียว (Unidimensional IRT Models) ดังนั้นการพัฒนาวิธีการประมาณค่าดัชนีการจำแนกประเภทในอนาคตอาจพิจารณาให้ครอบคลุมถึงโมเดลการตอบสนองข้อสอบแบบหลายมิติ (Multidimensional IRT Models) ก็จะเป็นประโยชน์อย่างมากในการประมาณค่าดัชนีการจำแนกประเภทต่อไป

2.4 การศึกษาครั้งนี้เป็นการศึกษาที่พิจารณาคะแนนจุดตัดที่สำนักทดสอบทางการศึกษาแห่งชาติกำหนดขึ้นมาด้วยวิธีการ Normalized T-Score ซึ่งพบว่าวิธีการประมาณค่าดัชนีการจำแนกประเภททั้งสามวิธีตามทฤษฎีการตอบสนองข้อสอบให้ค่าดัชนีการจำแนกประเภทในระดับค่อนข้างสูง

แต่ก็ยังไม่พบการศึกษาว่าเมื่อนำวิธีเหล่านี้ไปใช้ในการตรวจสอบคะแนนจุดตัดที่กำหนดขึ้นมาด้วยวิธีการอื่น เช่น วิธีแองกอฟ วิธีบุ๊คมาร์ค วิธีกลุ่มคาบเส้น วิธีกลุ่มตรงข้าม เป็นต้น วิธีการประมาณค่าดัชนีการจำแนกประเภททั้งสามวิธีนี้จะให้ผลการประมาณค่าเป็นอย่างไร ดังนั้นในการวิจัยครั้งต่อไปอาจนำวิธีการกำหนดคะแนนจุดตัดมาเป็นอีกปัจจัยหนึ่งในการศึกษา

2.5 เนื่องด้วยข้อมูลจริงที่ใช้ในการศึกษาครั้งนี้มีการแจกแจงแบบปกติ ซึ่งเป็นไปตามข้อตกลงเบื้องต้นตามวิธีการของ Guo จึงทำดัชนีการจำแนกประเภทมีค่าสูง ดังนั้นควรทำการศึกษาเพิ่มเติมในกรณีที่ข้อมูลที่นำมาศึกษามีการแจกแจงไม่ปกติว่าวิธีการของ Guo จะให้ค่าดัชนีการจำแนกประเภทเป็นอย่างไร

รายการอ้างอิง

ภาษาไทย

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

ศิริชัย กาญจนวาสี. (2556). *ทฤษฎีการทดสอบแบบดั้งเดิม*. พิมพ์ครั้งที่ 7. กรุงเทพฯ : โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

สถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน). (2556). *จำนวนข้อสอบ O-NET มัธยมศึกษาปีที่ 3 แยกตามรูปแบบของข้อสอบ ประจำปีการศึกษา 2556.* [online]. มาจาก: http://www.niets.or.th/[21 สิงหาคม 2556].

สถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน). (2556). *การดำเนินงานเกี่ยวกับการใช้ผล O-NET เป็นองค์ประกอบหนึ่งในการตัดสินผลการเรียนของผู้เรียน*. [online]. มาจาก: http://www.niets.or.th/[21 สิงหาคม 2556].

สถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน). (2557). *การใช้ผลการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐานเป็นองค์ประกอบหนึ่งในการตัดสินผลการเรียนของผู้เรียนที่จบการศึกษาขั้นพื้นฐาน พุทธศักราช 2551*. [online]. มาจาก: http://www.niets.or.th/[2 มิถุนายน 2557].

อนุสรณ์ เกิดศรี. (2557). *ประสิทธิภาพของวิธีการคัดเลือกข้อสอบสองวิธีในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ สำหรับโมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย: การเปรียบเทียบระหว่างวิธีมอนติ คาร์โล ซีเอที และวิธีแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ*. วิทยานิพนธ์ปริญญาครุศาสตรดุษฎีบัณฑิต สาขาวิชาการวัดและประเมินผลการศึกษา ภาควิชาวิจัยและจิตวิทยาการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.

ภาษาอังกฤษ

Brennan, R. L. & Lee, W. (2006). *Correcting for bias in single-administration decision consistency indexes (CASMA Research Report No. 18)*. Iowa City: Center for Advanced Studies in Measurement and Assessment, University of Iowa.

Brennan, R. L. & Wan, L. (2004). *A bootstrap procedure for estimating decision consistency for single-administration complex assessment (CASMA Research Report No. 7)*. Iowa City: Center for Advanced Studies in Measurement and Assessment, University of Iowa.

Cheng, Y. (2008). *Comparison of methods for constrained CAT item selection in classification accuracy and consistency.* University of Illinois at Urbana-Champaign, Deanna Morgan, College Board.

Clauser, B. E., Margolis, M. J., & Case, S. M. (2006). Testing for licensure and certification in the professions. In R. L. Brennan (ed). *Educational measurement (pp. 701-731)*. Westport, Ct.: Praeger.

Cui, Y., Gierl, M. J., & Chang, H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement*, *49,* 19-38.

Guo, F. (2006). Expected classification accuracy using the latent distribution. Practical assessment. *Research & Evaluation, 11(6)*, 1-9.

Hambleton, R. K. & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement, 10(3)*, 159-170.

Hanson, B. A. & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement, 27(4)*, 345-359.

Hubregtse, M. & Eggen, T. J. H. M. (2012). *Influences on classification accuracy of exam sets: An example from vocational education and training*. In Eggen, T. J. H. M. & Veldkamp, B. P. (Ed). Psychometrics in Practice at RCEC (pp. 107-123). Ipskamp Drukkers, Enschede: Netherlands.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement, 13(4)*, 253-264.

Huynh, H. (1990). Computation and statistical inference for decision consistency indexs based on the Rasch model. *Journal of Educational Statistics, 15(4)*, 353-368.

Kapoor, S. & Welch, C. (2011). *Comparability of paper and computer administrations in terms of proficiency interpretations*. A paper presented at the annual

meeting of the National Council on Measurement in Education New Orleans, LA.

Keller, L. A., Swaminathan, H. & Sireci, S. G. (2003). Evaluating scoring procedures for context-dependent item sets. *Applied Measurement in education, 16(3)*, 207-222.

Lathrop, Q. N., & Cheng, Y. (2013). Two Approaches to Estimation of Classification Accuracy Rate Under Item Response Theory. *Applied Psychological Measurement, 37*(3), 226-241. doi: 10.1177/0146621612471888

Lathrop, Q. N., & Cheng, Y. (2015). cacIRT: Classification Accuracy and Consistency under Item Response Theory. *R package version 1.4.* http://CRAN.R-project.org/package=cacIRT.

Lee, W. (2007). Multinomial and compound multinomial error models for tests with complex item scoring. *Applied Psychological Measurement, 31(4)*, 255-274.

Lee, W. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement, 47*, 1-17.

Lee, W., Brennan, R. L. & Wan, L. (2009). Classification consistency and accuracy for complex assessments under the compound multinomial model. *Applied Psychological Measurement, 33(5)*, 374-390.

Lee, W., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement, 26*, 412-432.

Lee, W., Brennan, R. L., & Kolen, M. J. (2000). Estimators of conditional scale-score standard errors of measurement: A simulation study. *Journal of Educational Measurement, 37(1)*, 1-20.

Livingston, S. A. & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32(2)*, 179-197.

Martineau, J. A. (2007). An expansion and practical evaluation of expected classification accuracy. *Applied Psychological Measurement, 31(3)*, 181-194.

Muijs, D. (2004). *Doing Quantitative Research in Education with SPSS*. London: Sage.

Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research & Evaluation, 7(14)*, 1-5.

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment, Research & Evaluation, 10(13)*, 1-4.

Suaklay, N., Lawthong, N. & Kanjanawasee, S. (2016). The influences of inter-item correlations and sample sizes on the classification indices under item response theory: the simulation study. *International Journal of Education and Psychology in the Community, 6(1 & 2)*, 106-120.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement, 13(4)*, 265-276.

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement, 25(1)*, 47-55.

Swaminathan, H., Hambleton, R. K. & Algina, J. (1974). Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement, 11(4)*, 263-267.

Wainer, H. & Kiely, G. L., (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24(3)*, 185–201.

Wan, L., Brennan, R. L., & Lee, W. (2007). *Estimating classification consistency for complex assessments (CASMA Research Report No. 22)*. Iowa City: Center for Advanced Studies in Measurement and Assessment, University of Iowa.

Wheadon, C. & Stockford, I. (2010). *Classification accuracy and consistency in GCSE and a level examinations offered by the Assessment and Qualifications Alliance (AQA) November 2008 to June 2009*. This report has been commissioned by the Office of Qualifications and Examinations Regulation.

Wyse, A. E. (2011). The potential impact of not being able to create parallel tests on expected classification accuracy. *Applied Psychological Measurement, 35*, 110-126.

Wyse, A. E., & Hao, S. (2012). An evaluation of item response theory classification accuracy and consistency indices. *Applied Psychological Measurement, 36*, 602-624.

Young, M. J. & Yoon, B. (1998). *Estimating the consistency and accuracy of classifications in a standards-referenced assessment*. CSE Technical Report 475.

Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. *Language Testing, 27*, 119-140.

Zhang, B. (2008). Investigating proficiency classification for the examination for the certificate of proficiency in English (ECPE). *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 6*, 57–75.

ภาคผนวก

ภาคผนวก ก

การจำลองข้อมูลการตอบข้อสอบด้วยโปรแกรม WINGEN3

การจำลองข้อมูลการตอบข้อสอบด้วยโปรแกรม WINGEN3

การจำลองข้อมูลการตอบข้อสอบของผู้สอบภายใต้สถานการณ์เงื่อนไขด้วยโปรแกรม WINGEN3 มีขั้นตอนในการจำลองข้อมูลดังนี้

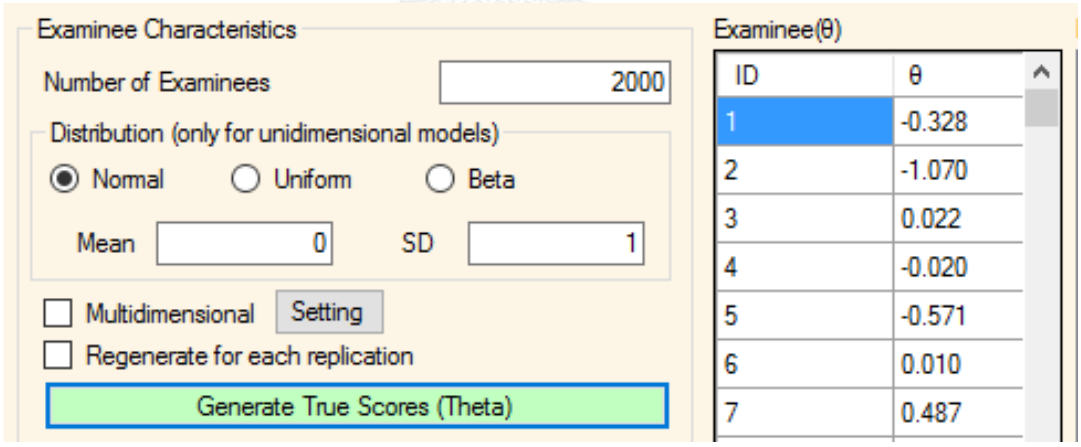
ขั้นตอนที่ 1 จำลองข้อมูลของผู้สอบ (generating examinee data)

1) ระบุจำนวนผู้สอบ 2,000 คน ซึ่งในการจำลองรูปแบบการตอบใช้กลุ่มตัวอย่างจำนวน 2,000 คน เนื่องจากในงานวิจัยของ Wyse & Hao (2012) ได้ทำการตรวจสอบขั้นต้นเกี่ยวกับจำนวนกลุ่มตัวอย่างที่แตกต่างกันพบว่า การใช้กลุ่มตัวอย่างจำนวน 2,000 คน ให้ผลการประมาณค่าที่ใกล้เคียงกับการใช้กลุ่มตัวอย่างจำนวน 10,000 หรือ 25,000 คน

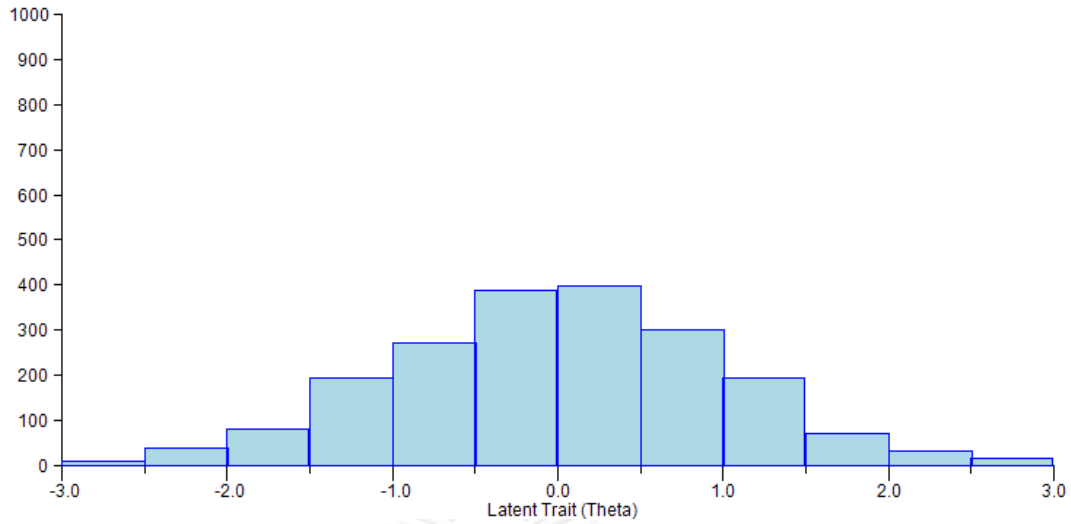
2) เลือกประเภทการแจกแจงของคะแนนเป็นการแจกแจงปกติ (normal distribution) และระบุค่า mean เป็น 0 และ SD เป็น 1

3) สั่งให้โปรแกรมจำลองข้อมูลของผู้สอบ จะได้ค่าพารามิเตอร์ความสามารถของผู้สอบ (theta) ตามจำนวนผู้สอบที่ระบุไว้

ตัวอย่างการกำหนดค่าในขั้นตอนที่ 1: สถานการณ์ 251PL10



ตัวอย่างการแจกแจงของความสามารถผู้สอบที่ได้จากการจำลองข้อมูล



ขั้นตอนที่ 2 จำลองข้อมูลของข้อสอบ (generating item data)

1) ระบุจำนวนข้อสอบที่ต้องการ ซึ่งในการจำลองครั้งนี้มีเงื่อนไขเกี่ยวกับความยาวของแบบสอบสองเงื่อนไข คือ แบบสอบยาว 25 ข้อ และ 50 ข้อตามลำดับ

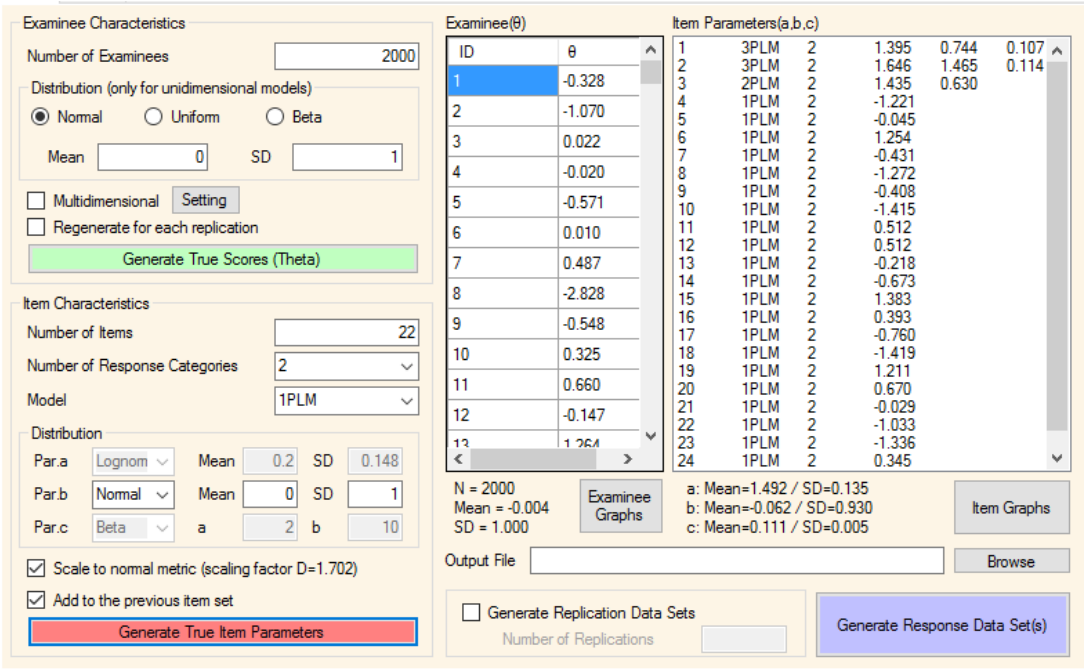
2) ระบุจำนวนตัวเลือกเป็น 2 ค่า คือ 0=ตอบผิด และ 1=ตอบถูก

3) เลือกประเภทโมเดล IRT ซึ่งในการจำลองครั้งนี้มีเงื่อนไขเกี่ยวกับโมเดล IRT จำนวน 3 โมเดล คือ 1PL, 2PL และ 3PL และในการจำลองข้อมูลการตอบแต่ละครั้งนั้นมีการกำหนดความไม่สอดคล้องของคำตอบกับโมเดล IRT ที่ใช้ในการวิเคราะห์ด้วย ซึ่งสามารถดำเนินการได้ดังนี้

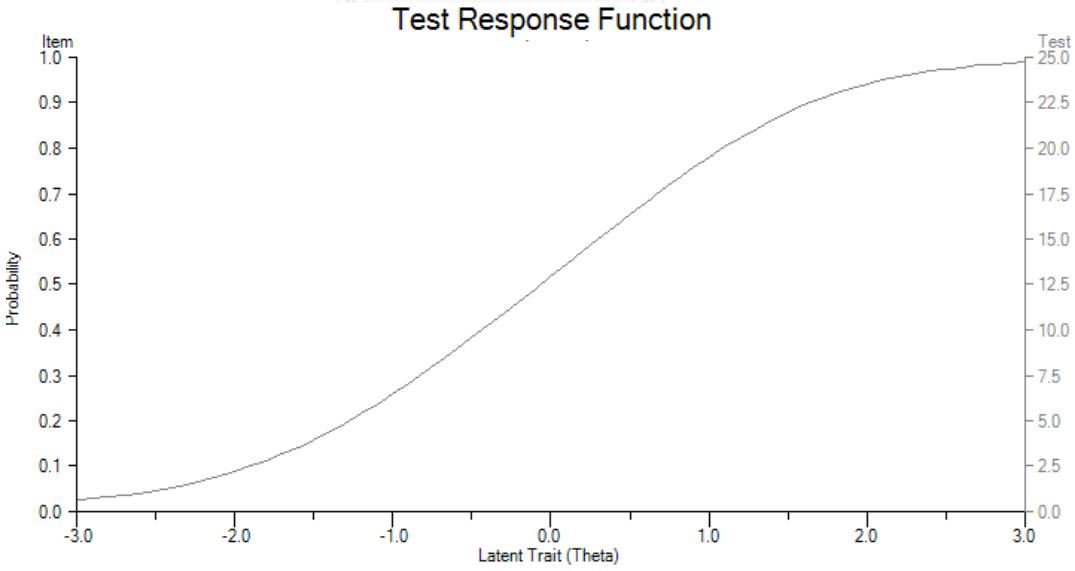
3.1) กำหนดประเภทการแจกแจงของค่าพารามิเตอร์ข้อสอบเพื่อให้ข้อมูลที่ได้มีลักษณะเป็นไปตามทฤษฎีที่ยอมรับได้ดังนี้ พารามิเตอร์อำนาจจำแนกเป็นการแจกแจงแบบ lognormal โดยระบุค่า mean เป็น 0.2 และ SD เป็น 0.148 พารามิเตอร์ความยากเป็นการแจกแจงแบบ normal โดยระบุค่า mean เป็น 0 และ SD เป็น 1 และพารามิเตอร์โอกาสในการเดาข้อสอบถูกเป็นการแจกแจงแบบ beta โดยระบุค่า a เป็น 2 และ b เป็น 10 (อนุสรณ์ เกิดศรี, 2557)

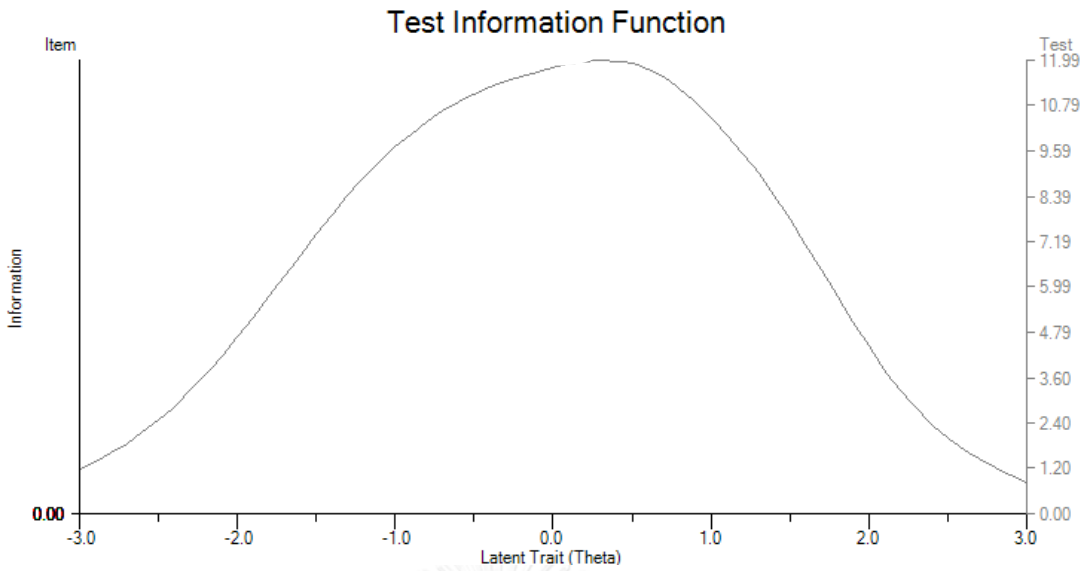
3.2) สั่งให้โปรแกรมจำลองข้อมูลของข้อสอบ จะได้ค่าพารามิเตอร์ตามที่กำหนดไว้ในขั้นตอนของการเลือกประเภทโมเดล เช่น สถานการณ์เงื่อนไขของโมเดลการวัดแบบ 3PL ที่มีความไม่สอดคล้องของข้อสอบกับโมเดลร้อยละ 10 ของข้อสอบที่มีความยาว 50 ข้อ ในการจำลองข้อมูลจะต้องกำหนดข้อสอบที่สอดคล้องกับโมเดล 3PL จำนวน 45 ข้อ และข้อสอบที่ไม่สอดคล้องกับโมเดล 3PL จำนวน 5 ข้อ โดยในการวิจัยครั้งนี้กำหนดให้เป็นข้อสอบแบบ 2PL จำนวน 3 ข้อ และข้อสอบแบบ 1PL จำนวน 2 ข้อ ผลที่ได้คือค่าพารามิเตอร์ประจำข้อสอบแต่ละข้อตามโมเดลที่กำหนดไว้ข้างต้น

ตัวอย่างการกำหนดค่าในขั้นตอนที่ 2: สถานการณ์ 251PL10



ตัวอย่างฟังก์ชันการตอบสนองข้อสอบของแบบสอบและฟังก์ชันสารสนเทศของแบบสอบที่ได้จากการจำลองข้อมูล



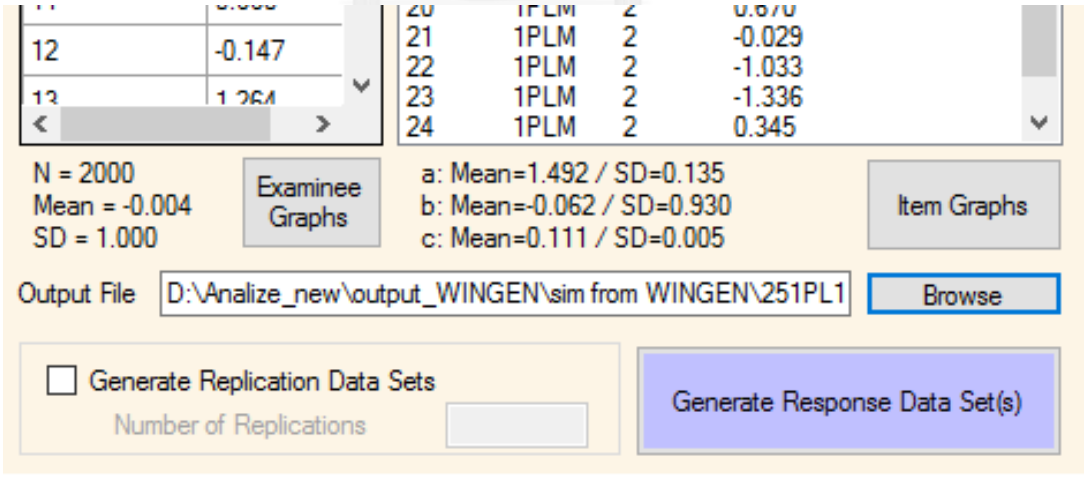


<u>ขั้นตอนที่ 3</u> จำลองข้อมูลการตอบข้อสอบ (generating item response data)

1) ระบุ output file สำหรับจัดเก็บข้อมูลที่จำลองขึ้น

2) สั่งให้โปรแกรมจำลองข้อมูลการตอบข้อสอบ จะได้รูปแบบการตอบจำนวนข้อสอบและจำนวนผู้สอบตามที่กำหนดไว้ข้างต้น

ตัวอย่างการกำหนดค่าในขั้นตอนที่ 3: สถานการณ์ 251PL10



ตัวอย่างรูปแบบการตอบข้อสอบของผู้สอบที่ได้จากการจำลองข้อมูล

resp_251PL10_r - Notepad

File Edit Format View Help

1	00000000101000000010000000
2	00000100010000000000000000
3	10101011110111111111011111
4	11110011111110101111011110
5	10111011101001101011110111
6	11111110011101111111111100
7	00111100101100000000110000

ภาคผนวก ข

คำสั่งที่ใช้ในโปรแกรม R

คำสั่งที่ใช้ในการประมาณค่าดัชนีการจำแนกประเภทด้วยโปรแกรม R (การศึกษาจำลอง)

1. วิธีการของ Rudner (2005)

1.1 คำสั่งสำหรับประมาณค่าดัชนีการจำแนกประเภท

```
> Rud.P
function (cutscore, theta, sem)
{
  os <- theta
  nn <- length(os)
  nc <- length(cutscore)
  if (nn != length(sem))
    stop("Ability and se of different length")
  esacc <- matrix(NA, length(cutscore), nn, dimnames = list(paste("cut at",
                                                round(cutscore, 3)), round(os, 3)))
  escon <- esacc
  for (j in 1:length(cutscore)) {
    cuts <- c(-Inf, cutscore[j], Inf)
    categ <- cut(os, cuts, labels = FALSE, right = FALSE)
    for (i in 1:nn) {
      esacc[j, i] <- (pnorm(cuts[categ[i] + 1], os[i],
                            sem[i]) - pnorm(cuts[categ[i]], os[i], sem[i]))
      escon[j, i] <- ((pnorm(cuts[2], os[i], sem[i]) -
                         pnorm(cuts[1], os[i], sem[i]))^2 + (pnorm(cuts[3],
                                                os[i], sem[i]) - pnorm(cuts[2], os[i], sem[i]))^2)
    }
  }
  if (nc == 1) {
    ans <- (list(Marginal = cbind(Accuracy = rowMeans(esacc),
                                  Consistency = rowMeans(escon)), Conditional = list(Accuracy = t(esacc),
                                                                                     Consistency = t(escon))))
    return(ans)
  }
```

```
  else {
    simul <- matrix(NA, nn, 2, dimnames = list(round(os,
                                              3), c("Accuracy", "Consistency")))
    cuts <- c(-Inf, cutscore, Inf)
    categ <- cut(os, cuts, labels = FALSE, right = FALSE)
    for (i in 1:nn) {
      simul[i, 1] <- (pnorm(cuts[categ[i] + 1], os[i],
                            sem[i]) - pnorm(cuts[categ[i]], os[i], sem[i]))
      sha <- 0
      for (j in 1:(nc + 1)) {
        sha <- sha + (pnorm(cuts[j + 1], os[i], sem[i]) -
                        pnorm(cuts[j], os[i], sem[i]))^2
      }
      simul[i, 2] <- sha
    }
    ans <- (list(Marginal = rbind(cbind(Accuracy = rowMeans(esacc),
                                        Consistency = rowMeans(escon)), Simultaneous = colMeans(simul)),
                 Conditional = list(Accuracy = cbind(t(esacc), Simultaneous = simul[,
                                                                                    1]), Consistency = cbind(t(escon),
Simultaneous = simul[,

2])))
    ans
  }
}
<environment: namespace:cacIRT>
```

1.2 ขั้นตอนและลำดับของคำสั่งในการวิเคราะห์

1) ดาวน์โหลดแพ็คเกจของคำสั่งที่ต้องใช้ในการวิเคราะห์ทั้งหมด

```
library("ltm")
library(matrixStats)
library(cacIRT)
```

2) นำเข้าข้อมูลการตอบข้อสอบที่จำลองมาจากโปรแกรม WINGEN

```
#input response data
setwd("d:\\Analize\\data\\")
##use response from WINGEN
resp<-read.csv("resp_25items_1pl_mis10.csv")
#resp<-read.csv("resp_25items_1pl_mis20.csv")
#resp<-read.csv("resp_25items_2pl_mis10.csv")
#resp<-read.csv("resp_25items_2pl_mis20.csv")
#resp<-read.csv("resp_25items_3pl_mis10.csv")
#resp<-read.csv("resp_25items_3pl_mis20.csv")
#resp<-read.csv("resp_50items_1pl_mis10.csv")
#resp<-read.csv("resp_50items_1pl_mis20.csv")
#resp<-read.csv("resp_50items_2pl_mis10.csv")
#resp<-read.csv("resp_50items_2pl_mis20.csv")
#resp<-read.csv("resp_50items_3pl_mis10.csv")
#resp<-read.csv("resp_50items_3pl_mis20.csv")
```

3) ประมาณค่าพารามิเตอร์ข้อสอบและความสามารถของผู้สอบ โดยในการประมาณค่าจะเปลี่ยนโมเดลการวัดตามเงื่อนไขที่กำหนด (1PL, 2PL, 3PL)

```
#estimate item parameter & ability
model1<-tpm(resp, type = c("latent.trait")) #3PL
#model1<-ltm(rdm~z1) #2PL
#model1<-rasch(rdm) #1PL
theta1<-factor.scores(model1)
item1<-theta1$coef #item parameter
ability1<-theta1$score.dat #ability&se
param<-matrix(c(item1[,3],item1[,2],item1[,1]),nrow=50,ncol=3,byrow=F) #50items for 3PL
```

```
#param<-matrix(c(item1[,3],item1[,2],item1[,1]),nrow=25,ncol=3,byrow=F) #25items for 3PL
#gues<-matrix(0,nrow=50,ncol=1,byrow=T) #50items for 1PL,2PL
#gues<-matrix(0,nrow=25,ncol=1,byrow=T) #25items for 1PL,2PL
#param<-matrix(c(item1[,2],item1[,1],gues[,1]),nrow=50,ncol=3,byrow=F) #50items for 1PL,2PL
#param<-matrix(c(item1[,2],item1[,1],gues[,1]),nrow=25,ncol=3,byrow=F) #25items for 1PL,2PL
#theta_se<-matrix(c(ability1[,53],ability1[,54]),ncol=2,byrow=F) #50items
#theta_se<-matrix(c(ability1[,28],ability1[,29]),ncol=2,byrow=F) #25items
theta<-matrix(ability1$z1)
se<-matrix(ability1$se.z1)
```

4) ประมาณดัชนีการจำแนกประเภท

4.1) กำหนดจำนวนรอบในการทำซ้ำ 100 รอบ

```
class<-matrix(nrow=100,ncol=16) #nrow=nloop
acc<-matrix(nrow=1,ncol=8)
cons<-matrix(nrow=1,ncol=8)
for(l in 1:101)
{#----start loop l
  class[l,1:8]<-acc
  class[l,9:16]<-cons
```

4.2) จำลองข้อมูลการตอบข้อสอบโดยใช้ค่าพารามิเตอร์ข้อสอบและความสามารถของผู้สอบที่ได้จากข้อ 3) เป็นตัวกำหนด

```
rdm<-sim(param,rnorm(2000))
```

4.3) ประมาณค่าพารามิเตอร์ความสามารถของผู้สอบโดยใช้โมเดลการวัดตามเงื่อนไขที่กำหนด (1PL, 2PL, 3PL) เพื่อนำค่าความสามารถของผู้สอบและค่าความคลาดเคลื่อนในการประมาณค่าความสามารถของผู้สอบไปใช้ในการประมาณค่าดัชนีการจำแนกประเภทต่อไป

```
  model<-rasch(rdm) #1PL
  #model<-ltm(rdm~z1) #2PL
  #model<-tpm(rdm, type = c("latent.trait")) #3PL
  theta<-factor.scores(model)
  item<-theta$coef #item parameter
  ability<-theta$score.dat #ability&se
  abilityj<-ability$z1
```

```
  se<-ability$se.z1
```

4.4) กำหนดคะแนนจุดตัด

```
  cutscore<-c(-2.8, -0.9, -0.1, 1, 1.8, 3, 3.7) #91_50items for 3PL
  #cutscore<-c(-1.9, -0.9, -0.2, 1.1, 2.4, 3.2, 4) #91_50items for 2PL
  #cutscore<-c(-2.1, -1, -0.2, 1.1, 2, 2.9, 3.7) #91_50items for 1PL
  #cutscore<-c(0.4, 1, 1.4, 1.7, 2.1, 2.6, 3.7) #94_25items for 3PL
  #cutscore<-c(0, 0.7, 1.3, 2.1, 3.5, 5.3, 7) #94_25items for 2PL
  #cutscore<-c(0, 0.7, 1.4, 2.1, 3.5, 5.1, 6.9) #94_25items for 1PL
```

4.5) ประมาณค่าดัชนีการจำแนกประเภทโดยใช้คำสั่ง Rud.P

```
  #crud<-Rud.P(cutscore,theta,sem)
  crud<-Rud.P(cutscore, abilityj, se)$Marginal
  acc<-crud[,1]
  cons<-crud[,2]

}
```

5) จัดเก็บค่าดัชนีการจำแนกประเภทที่ได้จากการประมาณค่าในขั้นที่ 4)

```
write.csv(class, file = "rudner_25items_1pl_mis10.csv")
#write.csv(class, file = "rudner_25items_1pl_mis20.csv")
#write.csv(class, file = "rudner_25items_2pl_mis10.csv")
#write.csv(class, file = "rudner_25items_2pl_mis20.csv")
#write.csv(class, file = "rudner_25items_3pl_mis10.csv")
#write.csv(class, file = "rudner_25items_3pl_mis20.csv")
#write.csv(class, file = "rudner_50items_1pl_mis10.csv")
#write.csv(class, file = "rudner_50items_1pl_mis20.csv")
#write.csv(class, file = "rudner_50items_2pl_mis10.csv")
#write.csv(class, file = "rudner_50items_2pl_mis20.csv")
#write.csv(class, file = "rudner_50items_3pl_mis10.csv")
#write.csv(class, file = "rudner_50items_3pl_mis20.csv")
```

2. วิธีการของ Guo (2006)

2.1 คำสั่งสำหรับประมาณค่าดัชนีการจำแนกประเภท

```
#function Guo
Guo.P<-function (cutscore, theta, like)
{
  os <- theta
  nn <- length(os)
  nc <- length(cutscore)
  if (nn != length(like))
    stop("theta and like of different length")
  esacc <- matrix(NA, length(cutscore), nn, dimnames = list(paste("cut at",
                                                  round(cutscore, 3)), round(os, 3)))
  escon <- esacc
  for (j in 1:length(cutscore)) {
    cuts <- c(-8, cutscore[j], 8)
    categ <- cut(os, cuts, labels = FALSE, right = FALSE)
    for (i in 1:nn) {
      esacc[j, i] <- (pnorm(cuts[categ[i] + 1], os[i],
                            like[i]) - pnorm(cuts[categ[i]], os[i], like[i]))
      escon[j, i] <- ((pnorm(cuts[2], os[i], like[i]) -
                         pnorm(cuts[1], os[i], like[i]))^2 + (pnorm(cuts[3],
                                                  os[i], like[i]) - pnorm(cuts[2], os[i], like[i]))^2)
    }
  }
  if (nc == 1) {
    ans <- (list(Marginal = cbind(Accuracy = rowMeans(esacc),
                                  Consistency = rowMeans(escon)), Conditional = list(Accuracy = t(esacc),
                                                                                     Consistency = t(escon))))
    return(ans)
  }
  else {
    simul <- matrix(NA, nn, 2, dimnames = list(round(os,
                                               3), c("Accuracy", "Consistency")))
```

```
    cuts <- c(-8, cutscore, 8)
    categ <- cut(os, cuts, labels = FALSE, right = FALSE)
    for (i in 1:nn) {
      simul[i, 1] <- (pnorm(cuts[categ[i] + 1], os[i],
                            like[i]) - pnorm(cuts[categ[i]], os[i], like[i]))
      sha <- 0
      for (j in 1:(nc + 1)) {
        sha <- sha + (pnorm(cuts[j + 1], os[i], like[i]) -
                        pnorm(cuts[j], os[i], like[i]))^2
      }
      simul[i, 2] <- sha
    }
    ans <- (list(Marginal = rbind(cbind(Accuracy = rowMeans(esacc),
                                        Consistency = rowMeans(escon)), Simultaneous = colMeans(simul)),
                 Conditional = list(Accuracy = cbind(t(esacc), Simultaneous = simul[,
                                                                                    1]), Consistency =
cbind(t(escon), Simultaneous = simul[,

2])))))
    ans
  }
}
```

2.2 ขั้นตอนและลำดับของคำสั่งในการวิเคราะห์

1) ดาวน์โหลดแพ็คเกจของคำสั่งที่ต้องใช้ในการวิเคราะห์ทั้งหมด

```
library("ltm")
library(matrixStats)
library(cacIRT)
```

2) นำเข้าข้อมูลการตอบข้อสอบที่จำลองมาจากโปรแกรม WINGEN

```
#input response data
setwd("d:\\Analize\\data\\")
##use response from WINGEN
resp<-read.csv("resp_25items_1pl_mis10.csv")
#resp<-read.csv("resp_25items_1pl_mis20.csv")
#resp<-read.csv("resp_25items_2pl_mis10.csv")
#resp<-read.csv("resp_25items_2pl_mis20.csv")
#resp<-read.csv("resp_25items_3pl_mis10.csv")
#resp<-read.csv("resp_25items_3pl_mis20.csv")
#resp<-read.csv("resp_50items_1pl_mis10.csv")
#resp<-read.csv("resp_50items_1pl_mis20.csv")
#resp<-read.csv("resp_50items_2pl_mis10.csv")
#resp<-read.csv("resp_50items_2pl_mis20.csv")
#resp<-read.csv("resp_50items_3pl_mis10.csv")
#resp<-read.csv("resp_50items_3pl_mis20.csv")
```

3) ประมาณค่าพารามิเตอร์ข้อสอบและความสามารถของผู้สอบ โดยในการประมาณค่าจะเปลี่ยนโมเดลการวัดตามเงื่อนไขที่กำหนด (1PL, 2PL, 3PL)

```
#estimate item parameter & ability
model1<-tpm(resp, type = c("latent.trait")) #3PL
#model1<-ltm(resp~z1) #2PL
#model1<-rasch(resp) #1PL
theta1<-factor.scores(model1)
item1<-theta1$coef #item parameter
ability1<-theta1$score.dat #ability&se
param<-matrix(c(item1[,3],item1[,2],item1[,1]),nrow=50,ncol=3,byrow=F) #50items for 3PL
```

```
#param<-matrix(c(item1[,3],item1[,2],item1[,1]),nrow=25,ncol=3,byrow=F) #25items for 3PL
#gues<-matrix(0,nrow=50,ncol=1,byrow=T) #50items for 1PL,2PL
#gues<-matrix(0,nrow=25,ncol=1,byrow=T) #25items for 1PL,2PL
#param<-matrix(c(item1[,2],item1[,1],gues[,1]),nrow=50,ncol=3,byrow=F) #50items for 1PL,2PL
#param<-matrix(c(item1[,2],item1[,1],gues[,1]),nrow=25,ncol=3,byrow=F) #25items for 1PL,2PL
theta<-matrix(ability1$z1)
se<-matrix(ability1$se.z1)
```

4) ประมาณดัชนีการจำแนกประเภท

4.1) กำหนดจำนวนรอบในการทำซ้ำ 100 รอบ

```
class<-matrix(nrow=100,ncol=16) #nrow=nloop
acc<-matrix(nrow=1,ncol=8)
cons<-matrix(nrow=1,ncol=8)
for(l in 1:101)
{#----start loop l
  class[l,1:8]<-acc
  class[l,9:16]<-cons
```

4.2) จำลองข้อมูลการตอบข้อสอบโดยใช้ค่าพารามิเตอร์ข้อสอบและความสามารถของผู้สอบที่ได้จากข้อ 3) เป็นตัวกำหนด

```
rdm<-sim(param,rnorm(2000))
```

4.3) ประมาณค่าพารามิเตอร์ความสามารถของผู้สอบโดยใช้โมเดลการวัดตามเงื่อนไขที่กำหนด (1PL, 2PL, 3PL) เพื่อนำค่าความสามารถของผู้สอบและค่าความคลาดเคลื่อนในการประมาณค่าความสามารถของผู้สอบไปใช้ในการประมาณค่าดัชนีการจำแนกประเภทต่อไป

```
model<-tpm(rdm, type = c("latent.trait")) #3PL
  #model<-ltm(rdm~z1) #2PL
  #model<-rasch(rdm) #1PL
  theta<-factor.scores(model)
  item<-theta$coef #item parameter
  ability<-theta$score.dat #ability&se
  #sort ability by z1
  abilityj<-ability[order(ability$z1),]
```

4.4) หาฟังก์ชันความน่าจะเป็นในการตอบข้อสอบของผู้สอบ (likelihood function) เพื่อนำไปใช้ในการประมาณค่าดัชนีการจำแนกประเภทต่อไป

```
#ni<-50 #nitems change to 50!!!
  ni<-25 #25items
  nj<-nrow(abilityj) #nexaminees
  ai<-t(matrix(matrix(param[,1]),nrow=ni,ncol=nj)) #discrimination of n items
  bi<-t(matrix(matrix(param[,2]),nrow=ni,ncol=nj)) #difficultly of n items
  ci<-t(matrix(matrix(param[,3]),nrow=ni,ncol=nj)) #guessing of n items
  D<-1.7
  thetaj<-matrix(ability$z1,nrow=nj,ncol=ni) #true
  pij<-1/(1+exp(-thetaj+bi)) #1PL
  #pij<-1/(1+exp(-D*ai*(thetaj-bi))) #2PL
  #pij<-ci+(1-ci)/(1+exp(-D*ai*(thetaj-bi))) #3PL

  qij<-1-pij
  #uij<-ability[,1:50] #nitems change to 50!!!
  uij<-ability[,1:25] #nitems change to 25!!!
  uij<-as.matrix(uij)
  Lij<-(pij^uij)*(qij^(1-uij))
  Lcat<-as.matrix(rowProds(Lij))
  thetag<-as.matrix(ability$z1)

  like<-apply(Lcat, 1, as.numeric)
  thetalike<-apply(thetag, 1, as.numeric)
```

4.5) กำหนดคะแนนจุดตัด

```
cutscore<-c(0.4, 1, 1.4, 1.7, 2.1, 2.6, 3.7) #25items for 3PL
  #cutscore<-c(0, 0.7, 1.3, 2.1, 3.5, 5.3, 7) #25items for 2PL
  #cutscore<-c(0, 0.7, 1.4, 2.1, 3.5, 5.1, 6.9) #25items for 1PL
  #cutscore<-c(-2.8, -0.9, -0.1, 1, 1.8, 3, 3.7) #50items for 3PL
  #cutscore<-c(-1.9, -0.9, -0.2, 1.1, 2.4, 3.2, 4) #50items for 2PL
  #cutscore<-c(-2.1, -1, -0.2, 1.1, 2, 2.9, 3.7) #50items for 1PL
```

4.6) ประมาณค่าดัชนีการจำแนกประเภทโดยใช้คำสั่ง Guo.P

```
#Guo.P<-function (cutscore, thetaj, Lcat)
cguo<-Guo.P(cutscore, thetalike, like)$Marginal

acc<-cguo[,1]
cons<-cguo[,2]

} #----end loop l
```

5) จัดเก็บค่าดัชนีการจำแนกประเภทที่ได้จากการประมาณค่าในขั้นที่ 4)

```
write.csv(class, file = "guo_25items_1pl_mis10.csv")
#write.csv(class, file = "guo_25items_1pl_mis20.csv")
#write.csv(class, file = "guo_25items_2pl_mis10.csv")
#write.csv(class, file = "guo_25items_2pl_mis20.csv")
#write.csv(class, file = "guo_25items_3pl_mis10.csv")
#write.csv(class, file = "guo_25items_3pl_mis20.csv")
#write.csv(class, file = "guo_50items_1pl_mis10.csv")
#write.csv(class, file = "guo_50items_1pl_mis20.csv")
#write.csv(class, file = "guo_50items_2pl_mis10.csv")
#write.csv(class, file = "guo_50items_2pl_mis20.csv")
#write.csv(class, file = "guo_50items_3pl_mis10.csv")
#write.csv(class, file = "guo_50items_3pl_mis20.csv")
```

3. วิธีการของ Lee (2010)

3.1 คำสั่งสำหรับประมาณค่าดัชนีการจำแนกประเภท

```
> Lee.P
function (cutscore, ip, theta, D = 1.7)
{
    ut <- theta
    ni <- dim(ip)[1]
    nn <- length(ut)
    sc <- ni + 1
    nc <- length(cutscore)
    exp.TS <- rowSums(sapply(1:ni, function(i) ip[i, 3] + (1 -
        ip[i, 3])/(1 + exp(-D * ip[i, 1] * (ut - ip[i, 2])))))
    rec.mat <- recursive.raw(ip, ut)
    esacc <- escon <- matrix(NA, nc, nn, dimnames = list(paste("cut at",
        round(cutscore, 3)), round(ut, 3)))
    for (j in 1:nc) {
        cuts <- c(0, cutscore[j], sc)
        categ <- cut(exp.TS, cuts, labels = FALSE, right = FALSE)
        bang <- ceiling(cuts)
        rec.s <- list(NA)
        for (i in 1:2) {
            rec.s[[i]] <- as.matrix(rec.mat[, (bang[i] + 1):bang[i +
                1]])
        }
        for (i in 1:nn) {
            esacc[j, i] <- sum(rec.s[[categ[i]]][i, ])
        }
        escon[j, ] <- rowSums(rec.s[[1]])^2 + rowSums(rec.s[[2]])^2
    }
    if (nc > 1) {
        simul <- matrix(NA, nn, 2, dimnames = list(round(ut,
            3), c("Accuracy", "Consistency")))
        cuts <- c(0, cutscore, sc)
        categ <- cut(exp.TS, cuts, labels = FALSE, right = FALSE)
        bang <- ceiling(cuts)
```

```
        rec.s <- list(NA)
        for (i in 1:(nc + 1)) {
            rec.s[[i]] <- as.matrix(rec.mat[, (bang[i] + 1):bang[i +
                1]])
        }
        for (i in 1:nn) {
            simul[i, 1] <- sum(rec.s[[categ[i]]][i, ])
        }
        sha <- matrix(0, nn, 1)
        for (i in 1:(nc + 1)) {
            sha <- sha + rowSums(rec.s[[i]])^2
        }
        simul[, 2] <- sha
        ans <- (list(Marginal = rbind(cbind(Accuracy = rowMeans(esacc),
            Consistency = rowMeans(escon)), Simultaneous = colMeans(simul)),
            Conditional = list(Accuracy = cbind(t(esacc), Simultaneous = simul[,
                1]), Consistency = cbind(t(escon), Simultaneous = simul[,
                2]))))
        return(ans)
    }
    else {
        ans <- (list(Marginal = cbind(Accuracy = rowMeans(esacc),
            Consistency = rowMeans(escon)), Conditional = list(Accuracy = t(esacc),
            Consistency = t(escon))))
        return(ans)
    }
}
<environment: namespace:cacIRT>
```

3.2 ขั้นตอนและลำดับของคำสั่งในการวิเคราะห์

2.2 ขั้นตอนและลำดับของคำสั่งในการวิเคราะห์

1) ดาวน์โหลดแพ็คเกจของคำสั่งที่ต้องใช้ในการวิเคราะห์ทั้งหมด

```
library("ltm")
library(matrixStats)
library(cacIRT)
```

2) นำเข้าข้อมูลการตอบข้อสอบที่จำลองมาจากโปรแกรม WINGEN

```
#input response data
setwd("d:\\Analize\\data\\")
##use response from WINGEN
resp<-read.csv("resp_25items_1pl_mis10.csv")
#resp<-read.csv("resp_25items_1pl_mis20.csv")
#resp<-read.csv("resp_25items_2pl_mis10.csv")
#resp<-read.csv("resp_25items_2pl_mis20.csv")
#resp<-read.csv("resp_25items_3pl_mis10.csv")
#resp<-read.csv("resp_25items_3pl_mis20.csv")
#resp<-read.csv("resp_50items_1pl_mis10.csv")
#resp<-read.csv("resp_50items_1pl_mis20.csv")
#resp<-read.csv("resp_50items_2pl_mis10.csv")
#resp<-read.csv("resp_50items_2pl_mis20.csv")
#resp<-read.csv("resp_50items_3pl_mis10.csv")
#resp<-read.csv("resp_50items_3pl_mis20.csv")
```

3) ประมาณค่าพารามิเตอร์ข้อสอบและความสามารถของผู้สอบ โดยในการประมาณค่าจะเปลี่ยนโมเดลการวัดตามเงื่อนไขที่กำหนด (1PL, 2PL, 3PL)

```
#estimate item parameter & ability
model1<-tpm(resp, type = c("latent.trait")) #3PL
#model1<-ltm(resp~z1) #2PL
#model1<-rasch(resp) #1PL
theta1<-factor.scores(model1)
item1<-theta1$coef #item parameter
```

```
ability1<-theta1$score.dat #ability&se
param<-matrix(c(item1[,3],item1[,2],item1[,1]),nrow=50,ncol=3,byrow=F) #50items for 3PL
#param<-matrix(c(item1[,3],item1[,2],item1[,1]),nrow=25,ncol=3,byrow=F) #25items for 3PL
#gues<-matrix(0,nrow=50,ncol=1,byrow=T) #50items for 1PL,2PL
#gues<-matrix(0,nrow=25,ncol=1,byrow=T) #25items for 1PL,2PL
#param<-matrix(c(item1[,2],item1[,1],gues[,1]),nrow=50,ncol=3,byrow=F) #50items for 1PL,2PL
#param<-matrix(c(item1[,2],item1[,1],gues[,1]),nrow=25,ncol=3,byrow=F) #25items for 1PL,2PL
theta<-matrix(ability1$z1)
se<-matrix(ability1$se.z1)
```

4) ประมาณดัชนีการจำแนกประเภท

4.1) กำหนดจำนวนรอบในการทำซ้ำ 100 รอบ

```
class<-matrix(nrow=100,ncol=16) #nrow=nloop
acc<-matrix(nrow=1,ncol=8)
cons<-matrix(nrow=1,ncol=8)
for(l in 1:101)
{#----start loop l
  class[l,1:8]<-acc
  class[l,9:16]<-cons
```

4.2) จำลองข้อมูลการตอบข้อสอบโดยใช้ค่าพารามิเตอร์ข้อสอบและความสามารถของผู้สอบที่ได้จากข้อ 3) เป็นตัวกำหนด

```
rdm<-sim(param,rnorm(2000))
```

4.3) ประมาณค่าพารามิเตอร์ข้อสอบโดยใช้โมเดลการวัดตามเงื่อนไขที่กำหนด (1PL, 2PL, 3PL) เพื่อนำค่าพารามิเตอร์ข้อสอบไปใช้ในการประมาณค่าดัชนีการจำแนกประเภทต่อไป

```
  model<-rasch(rdm) #1PL
  #model<-ltm(rdm~z1) #2PL
  #model<-tpm(rdm, type = c("latent.trait")) #3PL
  theta<-factor.scores(model)
  item<-theta$coef #item parameter
  ability<-theta$score.dat #ability&se
  gues<-matrix(0,nrow=25,ncol=1,byrow=T) #25items for 1PL,2PL
  #gues<-matrix(0,nrow=50,ncol=1,byrow=T) #50items for 1PL,2PL
```

```
ip<-matrix(c(item[,2],item[,1],gues[,1]),nrow=25,ncol=3,byrow=F) #25items for 1PL,2PL
#ip<-matrix(c(item[,2],item[,1],gues[,1]),nrow=50,ncol=3,byrow=F) #50items for 1PL,2PL
#ip<-matrix(c(item[,3],item[,2],item[,1]),nrow=25,ncol=3,byrow=F) #25items for 3PL
#ip<-matrix(c(item[,3],item[,2],item[,1]),nrow=50,ncol=3,byrow=F) #50items for 3PL
```

4.4) กำหนดคะแนนจุดตัด

```
cutscore<-c(7.18, 8.85, 10.35, 12.29, 14.87, 18.31, 21.96) #25items for 1PL, 2PL, 3PL
#cutscore<-c(11.27, 15.88, 19.62, 26.19, 30.61, 34.74, 38) #50items for 1PL, 2PL, 3PL
```

4.5) ประมาณค่าดัชนีการจำแนกประเภทโดยใช้คำสั่ง class.Lee

```
clee<-class.Lee(cutscore, ip, rdm=rdm)$Marginal

acc<-clee[,1]
cons<-clee[,2]

} #----end loop l
```

5) จัดเก็บค่าดัชนีการจำแนกประเภทที่ได้จากการประมาณค่าในขั้นที่ 4)

```
write.csv(class, file = "lee_25items_1pl_mis10.csv")
#write.csv(class, file = "lee_25items_1pl_mis20.csv")
#write.csv(class, file = "lee_25items_2pl_mis10.csv")
#write.csv(class, file = "lee_25items_2pl_mis20.csv")
#write.csv(class, file = "lee_25items_3pl_mis10.csv")
#write.csv(class, file = "lee_25items_3pl_mis20.csv")
#write.csv(class, file = "lee_50items_1pl_mis10.csv")
#write.csv(class, file = "lee_50items_1pl_mis20.csv")
#write.csv(class, file = "lee_50items_2pl_mis10.csv")
#write.csv(class, file = "lee_50items_2pl_mis20.csv")
#write.csv(class, file = "lee_50items_3pl_mis10.csv")
#write.csv(class, file = "lee_50items_3pl_mis20.csv")
```

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวนริศรา เสือคล้าย เกิดเมื่อวันที่ 26 มกราคม พ.ศ. 2526 ที่จังหวัดพิษณุโลก สำเร็จการศึกษาหลักสูตรครุศาสตรบัณฑิต สาขาวิชามัธยมศึกษา-วิทยาศาสตร์ (วิชาเอกวิทยาศาสตร์ทั่วไป-ฟิสิกส์) จากคณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย เมื่อปีการศึกษา 2548 ต่อมาสำเร็จการศึกษาหลักสูตรครุศาสตรมหาบัณฑิต สาขาวิชาวิจัยการศึกษา ภาควิชาวิจัยและจิตวิทยาการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย เมื่อปีการศึกษา 2550 และเข้าศึกษาต่อในหลักสูตรครุศาสตรดุษฎีบัณฑิต สาขาวิชาการวัดและประเมินผลการศึกษา ภาควิชาวิจัยและจิตวิทยาการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2555