ประสิทธิภาพของวิธีการคัดเลือกข้อสอบสองวิธีในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์สำหรับโมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย: การเปรียบเทียบระหว่างวิธีมอนติ คาร์โล ซีเอที และวิธีแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ

นายอนุสรณ์ เกิดศรี

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาครุศาสตรดุษฎีบัณฑิต
สาขาวิชาการวัดและประเมินผลการศึกษา ภาควิชาวิจัยและจิตวิทยาการศึกษา
คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2557

EFFICIENCY OF TWO ITEM SELECTION METHODS IN COMPUTERIZED ADAPTIVE TESTING FOR THE TESTLET RESPONSE MODEL: A COMPARISON BETWEEN THE MONTE CARLO CAT METHOD AND THE CONSTRAINT-WEIGHTED A-STRATIFICATION METHOD

Mr. Anusorn Koedsri

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 2014

หัวข้อวิทยานิพนธ์	ประสิทธิภาพของวิธีการคัดเลือกข้อสอบสองวิธีในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์สำหรับโมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย: การเปรียบเทียบระหว่างวิธีมอนติ คาร์โล ซีเอที และวิธีแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ
โดย	นายอนุสรณ์ เกิดศรี
สาขาวิชา	การวัดและประเมินผลการศึกษา
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร. ณัฏฐภรณ์ หลาวทอง
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	ผู้ช่วยศาสตราจารย์ ดร. สังวรณ์ งัดกระโทก

คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาดุษฎีบัณฑิต

.....คณบดีคณะครุศาสตร์

(รองศาสตราจารย์ ดร. บัญชา ชลาภิรมย์)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ

(ศาสตราจารย์ ดร. ศิริชัย กาญจนวาสี)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(ผู้ช่วยศาสตราจารย์ ดร. ณัฏฐภรณ์ หลาวทอง)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม

(ผู้ช่วยศาสตราจารย์ ดร. สังวรณ์ งัดกระโทก)

.....กรรมการ

(รองศาสตราจารย์ ดร. ศิริเดช สุชีวะ)

.....กรรมการ

(รองศาสตราจารย์ ดร. โชติกา ภาษีผล)

.....กรรมการภายนอกมหาวิทยาลัย

(ดร. รังสรรค์ มณีเล็ก)

อนุสรณ์ เกิดศรี : ประสิทธิภาพของวิธีการคัดเลือกข้อสอบสองวิธีในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์สำหรับโมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย: การเปรียบเทียบระหว่างวิธีมอนติ คาร์โล ซีเอที และวิธีแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (EFFICIENCY OF TWO ITEM SELECTION METHODS IN COMPUTERIZED ADAPTIVE TESTING FOR THE TESTLET RESPONSE MODEL: A COMPARISON BETWEEN THE MONTE CARLO CAT METHOD AND THE CONSTRAINT-WEIGHTED A-STRATIFICATION METHOD) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ผศ. ดร. ณัฏฐภรณ์ หลาวทอง, อ.ที่ปรึกษาวิทยานิพนธ์ร่วม: ผศ. ดร. สังวรณ์ งัดกระโทก, 245 หน้า.

การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อ 1) เปรียบเทียบประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถของผู้สอบระหว่างวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที (MCC) และวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (CWA) และ 2) เปรียบเทียบประสิทธิภาพด้านการใช้คลังข้อสอบระหว่างวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที (MCC) กับวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (CWA) การศึกษาครั้งนี้เปรียบเทียบประสิทธิภาพระหว่างวิธีการคัดเลือกแบบทดสอบย่อย 2 วิธี ภายใต้โมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย (testlet response model) ข้อมูลถูกจำลองขึ้นมาเพื่อศึกษาประสิทธิภาพภายใต้ตัวแปรอิสระ 2 ตัว คือ อัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (10%, 15%, 20%, 25%) และขนาดคลังข้อสอบ (600 และ 800 ข้อ) เกณฑ์ที่นำมาใช้เปรียบเทียบประสิทธิภาพในการประมาณค่าพารามิเตอร์ ได้แก่ ความลำเอียง ความแปรปรวนของความคลาดเคลื่อน สหสัมพันธ์ระหว่างค่าความสามารถจริงและความสามารถประมาณค่า และความยาวของแบบสอบ ประสิทธิภาพการใช้คลังข้อสอบพิจารณาจากเกณฑ์ 5 เกณฑ์ ได้แก่ อัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ อัตราการทับซ้อนของแบบสอบ ไคสแควร์ จำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ และจำนวนแบบทดสอบย่อยที่ถูกนำมาใช้มากเกินไป ซึ่งการทดลองถูกทำซ้ำ 10 ครั้ง สำหรับแต่ละเงื่อนไข ผลการวิจัยมีดังต่อไปนี้

1. การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที (MCC) มีประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถของผู้สอบสูงกว่าการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (CWA) ในทุกเงื่อนไขการทดลอง

2. การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที (MCC) มีประสิทธิภาพด้านการใช้คลังข้อสอบสูงกว่าการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (CWA) เมื่ออัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ต่ำกว่า 20% แต่เมื่ออัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด เพิ่มขึ้นเป็น 25% การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (CWA) มีประสิทธิภาพเหนือกว่าการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที (MCC) โดยพิจารณาได้จากจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ อัตราการแสดงแบบทดสอบย่อยที่สังเกตได้ และ ไคสแควร์ ที่มีค่าต่ำกว่า อย่างไรก็ตามการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที (MCC) มีประสิทธิภาพสูงกว่าเมื่อพิจารณาจากอัตราการทับซ้อนของแบบสอบที่ต่ำกว่า

ภาควิชา	วิจัยและจิตวิทยาการศึกษา	ลายมือชื่อนิสิต	.....
สาขาวิชา	การวัดและประเมินผลการศึกษา	ลายมือชื่อ อ.ที่ปรึกษาหลัก	.....
ปีการศึกษา	2557	ลายมือชื่อ อ.ที่ปรึกษาร่วม	.....

5284269027 : MAJOR EDUCATIONAL MEASUREMENT AND EVALUATION
KEYWORDS: COMPUTERIZED ADAPTIVE TESTING / TESTLET RESPONSE THEORY / CONSTRAINT-WEIGHTED A-STRATIFICATION METHOD / MONTE CARLO CAT METHOD / EFFICIENCY OF POOL UTILIZATION / EFFICIENCY OF MEASUREMENT PRECISION

ANUSORN KOEDSRI: EFFICIENCY OF TWO ITEM SELECTION METHODS IN COMPUTERIZED ADAPTIVE TESTING FOR THE TESTLET RESPONSE MODEL: A COMPARISON BETWEEN THE MONTE CARLO CAT METHOD AND THE CONSTRAINT-WEIGHTED A-STRATIFICATION METHOD. ADVISOR: ASST. PROF. NUTTAPORN LAWTHONG, Ph.D., CO-ADVISOR: ASST. PROF. SUNGWON NGUDGRATOKE, Ph.D., 245 pp.

The purposes of this research were 1) to compare the efficiency of measurement precision between the Monte Carlo CAT method (MCC) and the constraint-weighted a-stratification method (CWA), 2) to compare the efficiency of pool utilization between MCC and CWA. This research compared the efficiency between the two methods within the testlet response model. The data were generated to explore the efficiency under two independent variables: maximum testlet exposure rate (10%, 15%, 20%, 25%) and item pool size (600 and 800). The efficiency in parameter estimation was compared in term of bias, MSE, correlation between true ability and estimated ability as well as test length. The pool utilization efficiency was examined through five criteria: observed testlets exposure rate, test overlap rate, testlets exposure rate distribution (chi-square), number of never exposed testlets and number of overexposed testlets. The experiments were repeated 10 times for each condition. The findings of the research were as follows.

1. The MCC showed greater measurement precision than the CWA for all experimental conditions.

2. The MCC provided greater efficiency in pool utilization than the CWA when maximum testlet exposure rates of less than 20% were investigated, but when maximum testlet exposure rate increased to 25% the CWA was more efficient than the MCC which was evidenced by lower number of never exposed testlets, lower observed testlets exposure rate, and lower testlets exposure rate distribution (chi-square); however, the MCC was more efficient suggested by lower test overlap rate.

Department:	Educational Research and Psychology	Student's Signature .....
		Advisor's Signature .....
Field of Study:	Educational Measurement and Evaluation	Co-Advisor's Signature .....
Academic Year:	2014	

กิตติกรรมประกาศ

การทำวิทยานิพนธ์ฉบับนี้ ผู้วิจัยได้รับความรู้และคำแนะนำต่างๆ ด้วยความกรุณาอย่างยิ่งจากอาจารย์ที่ปรึกษาวิทยานิพนธ์ คือ ผู้ช่วยศาสตรจารย์ ดร. ณัฏฐภรณ์ หลาวทอง และอาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม คือ ผู้ช่วยศาสตราจารย์ ดร.สังวรณ์ งัดกระโทก ซึ่งอาจารย์ทั้งสองท่านได้ให้ความรู้ในหลักวิชามากมายอันทรงคุณค่ายิ่ง ให้ข้อเสนอแนะในการพัฒนาโปรแกรมคอมพิวเตอร์ ให้ข้อเสนอแนะที่เป็นประโยชน์ต่อการจำลองข้อมูลและการวิเคราะห์ข้อมูลจนงานวิจัยเสร็จสมบูรณ์ อีกทั้งเป็นผู้จุดประกายความคิดที่ทำให้ผู้วิจัยสนใจทำวิจัยเรื่องนี้ คอยแนะนำเอกสารงานวิจัยที่เป็นประโยชน์ พร้อมทั้งยังได้คอยดูแลเอาใจใส่คอยติดตามความก้าวหน้าของผู้วิจัยตลอดมา ผู้วิจัยรู้สึกซาบซึ้งและขอกราบขอบพระคุณเป็นอย่างยิ่งด้วยความเคารพมา ณ ที่นี้ด้วย

ขอกราบขอบพระคุณ ศาสตราจารย์ ดร.ศิริชัย กาญจนวาสี รองศาสตราจารย์ ดร. ศิริเดช สุชีวะ รองศาสตราจารย์ ดร. โชติกา ภาษีผล และอาจารย์ ดร.รังสรรค์ มณีเล็ก ซึ่งเป็นกรรมการสอบวิทยานิพนธ์ที่ได้กรุณาสละเวลาอันมีค่ามาร่วมสอบวิทยานิพนธ์ ตลอดจนให้ข้อเสนอแนะที่เป็นประโยชน์ต่อการทำวิทยานิพนธ์ฉบับนี้จนสมบูรณ์ และขอกราบขอบพระคุณคณาจารย์ภาควิชาวิจัยและจิตวิทยาการศึกษา จุฬาลงกรณ์มหาวิทยาลัย ทุกท่านที่คอยอบรมสั่งสอน คอยให้การสนับสนุนและเป็นกำลังใจด้วยดีตลอดมาจนผู้วิจัยสามารถประสบความสำเร็จด้านการศึกษาในวันนี้

ขอกราบขอบพระคุณ Dr.Qi Diao นักวิจัยผู้เชี่ยวชาญด้าน Computer-based assessment และ Automated test assembly จาก CTB/McGraw-Hill ประเทศสหรัฐอเมริกา ที่กรุณาอนุเคราะห์จัดส่งเอกสารที่มีคุณค่านำมาใช้เป็นพื้นฐานความรู้สำหรับการทำวิจัยครั้งนี้

ขอกราบขอบพระคุณ คุณวงษ์ สุขุประการ ผู้อำนวยการโรงเรียนวัดเขาพระยาสังฆารามที่สนับสนุนให้ผู้วิจัยมีโอกาสศึกษาต่อในระดับปริญญาเอก และขอขอบคุณคณะครูทุกท่านที่เสียสละปฏิบัติหน้าที่แทนผู้วิจัยตลอดช่วงเวลาของการลาศึกษาต่อ

ขอกราบขอบพระคุณบัณฑิตวิทยาลัยจุฬาลงกรณ์มหาวิทยาลัยที่มอบ “ทุน 90 ปี จุฬาลงกรณ์มหาวิทยาลัย ” จากกองทุนรัชดาภิเษกสมโภช จุฬาลงกรณ์มหาวิทยาลัย ทำให้ผู้วิจัยสามารถพัฒนาวิทยานิพนธ์ได้อย่างเต็มที่

ขอขอบคุณเป็นพิเศษสำหรับคุณเคียงขวัญ มหาโชคเลิสวัฒนา และพี่น้องภาควิชาวิจัยและจิตวิทยาการศึกษาทุกคนที่ให้คำปรึกษา และเป็นกำลังใจตลอดจนคอยช่วยเหลืออย่างกัลยาณมิตรด้วยดีเสมอมา

ท้ายสุดนี้ขอกราบขอบพระคุณบุคคลอันเป็นที่รักและสำคัญอย่างยิ่ง คุณพ่ออารี และคุณแม่สุภาภรณ์ เกิดศรี ตลอดจนญาติพี่น้องทุกคน ที่ได้ให้ความอบอุ่น ห่วงใยและสนับสนุนในทุกๆ ด้าน

สารบัญ

หน้า

หน้า

หน้า

หน้า

หน้า

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญภาพ

ภาพที่ **หน้า**

บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

การบริหารการทดสอบของโปรแกรมการทดสอบที่เป็นมาตรฐานในปัจจุบันสามารถจำแนกได้เป็น 2 รูปแบบ คือ 1) การบริหารที่เป็นการทดสอบแบบดั้งเดิมที่เป็นการทดสอบแบบข้อเขียน (paper & pencil test) เป็นการสร้างแบบสอบขึ้นมาชุดเดียวให้ครอบคลุมคุณลักษณะหรือเนื้อหาที่ต้องการวัด มีค่าความยากของข้อสอบที่หลากหลาย โดยไม่มีการกำหนดสัดส่วนของค่าความยากที่แน่นอน ไม่ว่าผู้สอบจะมีความสามารถระดับใด ผู้สอบทุกคนจะต้องทำข้อสอบฉบับเดียวกัน มีจำนวนข้อเท่ากันและเหมือนกันหมดทุกข้อ และ 2) การบริหารที่เป็นการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (Computerized Adaptive Testing: CAT) เป็นการทดสอบที่นำคอมพิวเตอร์เข้ามาช่วยในการดำเนินการทดสอบและใช้ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) เป็นพื้นฐานในการอธิบายความสัมพันธ์ระหว่างความสามารถที่มีอยู่ภายในตัวผู้สอบกับพฤติกรรมการตอบข้อสอบของตนเองโดยใช้ฟังก์ชันการตอบสนองข้อสอบที่มีความซับซ้อนทางสถิติ ผู้สอบแต่ละคนจะได้รับข้อสอบที่แตกต่างกันโดยพิจารณาจากความสามารถภายในของผู้สอบ หรือผู้สอบแต่ละคนได้รับข้อสอบที่มีพารามิเตอร์ข้อสอบเหมาะสมกับความสามารถของตน การทดสอบเริ่มต้นจากผู้สอบได้รับข้อสอบข้อแรกที่มีความยากปานกลาง ถ้าตอบข้อสอบถูกต้องข้อสอบข้อถัดไปจะมีความยากเพิ่มขึ้น ในทางกลับกันถ้าผู้สอบตอบผิด ข้อสอบข้อถัดไปจะมีความยากลดลง การทดสอบดำเนินต่อไปจนกระทั่งผลการทดสอบสามารถประมาณค่าความสามารถของผู้สอบใกล้เคียงและน่าเชื่อถือมากที่สุดการทดสอบจึงยุติ (ศิริชัย กาญจนวาสี, 2550)

การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ มีจุดเด่นหลายประการดังนี้ ช่วยลดจำนวนข้อและเวลาของการทดสอบ ลดความเหนื่อยล้าของผู้สอบจากการได้รับข้อสอบที่ไม่เหมาะสมกับระดับความสามารถของผู้สอบ เมื่อเปรียบเทียบจำนวนข้อสอบที่ผู้สอบจะได้รับกับการทดสอบแบบดั้งเดิม การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ สามารถที่จะลดจำนวนข้อสอบลงได้ประมาณครึ่งหนึ่งโดยไม่สูญเสียความแม่นยำในการวัด และสามารถรายงานผลการทดสอบได้ทันที (Chen, 2010; Chen, Ankenmann, & Spray, 2003; Chen & Lei, 2005, 2010; Wang & Vispoel, 1998)

องค์ประกอบสำคัญที่ถือว่าเป็นหัวใจของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ คือ การคัดเลือกข้อสอบ (Item Selection) (Weiss & Kingsbury, 1984) โดยข้อสอบจะถูกเลือกมาสร้างเป็นแบบสอบ ข้อสอบแต่ละข้อจะเหมาะสมกับความสามารถของผู้สอบแต่ละคน (Tailored Test)

วิธีการที่นิยมใช้ในการคัดเลือกข้อสอบ คือ เกณฑ์การคัดเลือกข้อสอบที่ให้สารสนเทศสูงสุดและสอดคล้องกับความสามารถของผู้สอบ (Maximum Information Criterion: MIC) ซึ่งวิธีคัดเลือกข้อสอบที่ให้สารสนเทศสูงสุดจะดำเนินการเปรียบเทียบค่าสารสนเทศของข้อสอบทุกข้อในคลังข้อสอบ (Item Pool) ข้อใดให้ค่าสารสนเทศสูงสุดข้อนั้นถูกเลือกให้กับผู้สอบ และการวัดสารสนเทศของข้อสอบส่วนใหญ่จะถูกกำหนดโดยพารามิเตอร์อำนาจจำแนกของข้อสอบ (Discrimination Parameter) ข้อที่มีอำนาจจำแนกสูงสุด ณ ตำแหน่งความสามารถนั้นๆ จะเป็นข้อที่ให้สารสนเทศสูงสุด และทำให้การประมาณค่าความสามารถมีความแม่นยำสูง (Precision) แต่การคัดเลือกข้อสอบด้วยวิธี MIC เพื่อให้ได้สารสนเทศสูงสุดจะทำให้ข้อสอบบางข้อที่มีอำนาจจำแนกสูงถูกเลือกมาใช้บ่อยเกินไป (Overexpose) ข้อสอบข้อนั้นมีอัตราการใช้ข้อสอบซ้ำสูงจะทำให้แบบสอบที่จัดให้ผู้สอบแต่ละคนมีอัตราการทับซ้อนของแบบสอบ (Test Overlap Rate) สูงขึ้น ซึ่งแบบสอบที่ดีควรจะมีการทับซ้อนกันต่ำที่สุด ข้อสอบที่มีอัตราการใช้ข้อสอบซ้ำสูงนำไปสู่ปัญหาด้านความปลอดภัยของแบบสอบ (Test security problem) เนื่องจากข้อสอบที่ถูกนำไปใช้บ่อยครั้งจนเป็นที่รู้จักกันแพร่หลายในกลุ่มผู้สอบ และข้อสอบส่วนที่ใหญ่มีจำนวนเกินกว่าครึ่งหนึ่งของจำนวนข้อทั้งหมดในคลังข้อสอบแทบไม่ถูกนำออกมาใช้ (Underexpose) จึงทำให้ขาดประสิทธิภาพในการใช้คลังข้อสอบ (Lack of Pool Utilization) ซึ่งไม่คุ้มค่ากับการลงทุนในการพัฒนาข้อสอบ (Cheng, Chang, Douglas, & Guo, 2009; Cheng, Chang, & Yi, 2007; Wim J. van der Linden, 2005)

ปัจจุบันการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ได้ถูกนำไปประยุกต์ใช้ในการทดสอบมาตรฐาน เช่น การสอบเพื่อเข้าศึกษาในสถาบันทางด้านกฎหมาย (Law School Admission Test: LSAT) การสอบภาษาอังกฤษเป็นภาษาต่างประเทศ (Test of English as a Foreign Language: TOEFL) ข้อสอบวัดเชาวน์ปัญญาทั่วไปก่อนเข้าศึกษาในระดับบัณฑิตศึกษา (Graduate Record Examinations: GRE) หรือข้อสอบที่ใช้วัดความสามารถของผู้ที่ต้องการเข้าศึกษาในระดับปริญญาโทและปริญญาเอกสาขาบริหารธุรกิจ (Graduate Management Admission Test: GMAT) ซึ่งเป็นโปรแกรมการทดสอบขนาดใหญ่มีผู้เข้าสอบครั้งละจำนวนมาก และมีผลสำคัญต่อผู้สอบ (High-Stakes Testing) ผลการทดสอบถูกนำไปใช้เป็นส่วนหนึ่งในการตัดสินใจในเรื่องสำคัญ เช่น คัดเลือกเข้าศึกษาต่อ เป็นต้น โดยโปรแกรมการทดสอบที่เป็นมาตรฐาน (Standardized Testing Program) ควรมีคุณลักษณะสำคัญ 2 ประการ ดังนี้ 1) มีการประมาณค่าความสามารถที่เชื่อถือได้ และ 2) แบบสอบที่ผู้สอบแต่ละคนได้รับต้องครอบคลุมมวลเนื้อเรื่องหรือประสบการณ์ที่ต้องการวัด หรือแบบสอบสามารถวัดตรงตามคุณลักษณะที่มุ่งวัด (Wim J. van der Linden, 2005) เพื่อตอบสนองความต้องการประการแรก โปรแกรมการทดสอบที่มีรูปแบบการบริหารการทดสอบด้วยคอมพิวเตอร์ ต้องสามารถประมาณค่าความสามารถของผู้สอบได้อย่างถูกต้องแม่นยำ และเลือกข้อสอบข้อถัดไปให้

เหมาะสมกับระดับความสามารถของผู้สอบและให้สารสนเทศเกี่ยวกับผู้สอบสูงสุดโดยตั้งอยู่บนพื้นฐานของคุณลักษณะประการที่ 2 เพื่อให้การทดสอบมีทั้งความเที่ยง (Reliability) และความตรง (Validity) ในขณะที่การตอบสนองคุณลักษณะประการสอง การทดสอบแบบดั้งเดิมสามารถดำเนินการได้ง่ายกว่าการบริหารการทดสอบด้วยคอมพิวเตอร์ โดยใช้ตารางกำหนดคุณลักษณะข้อสอบ (Table of Specification) เพื่อคัดเลือกข้อสอบรวมเป็นแบบสอบ (Assembling the Test) (ศิริชัย กาญจนวาสี, 2552) จากนั้นจึงนำแบบสอบไปใช้กับกลุ่มผู้สอบ การกระทำเช่นนี้สามารถรับประกันได้ว่า แบบสอบที่จะนำไปใช้มีเนื้อหาครอบคลุมจุดมุ่งหมายของการวัดและผู้สอบทุกคนจะได้รับแบบสอบที่มีความสมดุลของเนื้อหา (Content Balancing) ใกล้เคียงกัน ซึ่งตรงกันข้ามกับแนวคิดของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่ถูกพัฒนาขึ้นเพื่อจัดการกับปัญหาทางด้านสถิติ กระบวนการทำงานของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์เริ่มต้นจากการเลือกข้อสอบที่มีความยากระดับปานกลางเมื่อผู้สอบตอบคำถาม ข้อสอบข้อถัดไปจะถูกเลือกให้เหมาะสมกับระดับความสามารถของผู้สอบ สำหรับการประยุกต์ใช้การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์กับโปรแกรมการทดสอบขนาดใหญ่ (Large-scale testing program) การคัดเลือกข้อสอบเพื่อนำไปใช้กับผู้สอบไม่สามารถใช้เฉพาะค่าความยากของข้อสอบโดยลำพัง ดังนั้น การควบคุมเงื่อนไขบังคับ (Constraints Control) จึงจำเป็นต้องนำมาพิจารณาในการออกแบบการทดสอบแบบปรับเหมาะเพื่อควบคุมการคัดเลือกข้อสอบและทำให้เชื่อมั่นได้ว่าผู้สอบแต่ละคนจะได้รับแบบสอบที่มีสัดส่วนของเนื้อหาเหมือนกัน (Chen & Ankenman, 2004; Cheng et al., 2007; Eggen & Straetmans, 2000; Leung, Chang, & Hau, 2003; van der Linden, Ariel, & Veldkamp, 2006; van der Linden & Reese, 2001; Wang & Kolen, 2001)

การดำเนินงานโดยทั่วไปของโปรแกรมการทดสอบที่มีการบริหารการทดสอบโดยใช้คอมพิวเตอร์ ข้อสอบในคลังข้อสอบจะถูกนำกลับมาใช้ใหม่ในการทดสอบครั้งต่อไป ข้อสอบที่ถูกนำออกมาใช้บ่อยครั้งเกินไป (Overexposed Times) จนเป็นที่รู้จักทั่วไปในกลุ่มผู้สอบจะนำไปสู่ปัญหาในเรื่องความปลอดภัยของแบบสอบ (Test Security) ซึ่งเกิดจากการร่วมใช้ข้อสอบ (Item Sharing) ระหว่างผู้สอบ ผู้สอบสามารถตอบข้อสอบถูกต้องได้โดยไม่ได้ใช้ความสามารถที่แท้จริงของตนเอง เนื่องจากผู้สอบอาจได้รับข้อมูลเกี่ยวกับแบบสอบและล่วงรู้ข้อสอบก่อน (Item Pre-Knowledge) จากผู้ที่เข้ารับการทดสอบก่อนหน้า คะแนนที่สังเกตได้ของผู้สอบที่รู้ข้อสอบก่อนจะขาดความถูกต้องไม่สะท้อนความสามารถที่แท้จริงของผู้สอบ ในขณะที่การทดสอบแบบดั้งเดิม การร่วมใช้ข้อสอบไม่เป็นประเด็นปัญหาสำคัญ เนื่องจากการทดสอบแบบดั้งเดิมส่วนมากมีการจัดการทดสอบตามช่วงเวลาที่แน่นอนและข้อสอบที่ใช้ในการทดสอบแต่ละครั้งส่วนใหญ่ไม่ได้นำกลับมาใช้ใหม่ ซึ่งตรงกันข้ามกับการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (Chang & Zhang, 2002; Revuelta & Ponsoda,

1998; Way, 1998) เพื่อลดการคุกคามของการร่วมใช้ข้อสอบระหว่างผู้สอบซึ่งมีผลต่อความน่าเชื่อถือของการวัดโดยตรง ดังนั้น การควบคุมการใช้ข้อสอบซ้ำ (Item Exposure Control) ต้องถูกกำหนดรวมเข้าในการคัดเลือกข้อสอบ (Chen et al., 2003; Davis & Dodd, 2003; Eggen, 2001; Leung, Chang, & Hau, 2002; Revuelta & Ponsoda, 1998; Stocking, 1995; Stocking & Lewis, 1995; Stocking & Lewis, 1998; Stocking & Lewis, 2002; van der Linden, 2003)

ดังนั้นองค์ประกอบสำคัญในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ คือ วิธีการคัดเลือกข้อสอบ (Item Selection Method) วิธีที่นิยมใช้กันกว้างขวาง คือ วิธี MIC ด้วยวิธีดังกล่าวในทางทฤษฎีสามารถให้ประสิทธิภาพในการประมาณค่าความสามารถได้ดีที่สุด แต่โปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์จำเป็นต้องพิจารณาทั้งเงื่อนไขบังคับทางสถิติ (statistical constraints) เช่น ค่าสารสารเทศของข้อสอบ เป็นต้น และเงื่อนไขบังคับที่ไม่ใช่ทางสถิติ (non-statistical constraints) เช่น ขอบเขตเนื้อหา (content area) หรือที่เรียกว่าความสมดุลของเนื้อหา (content balancing) สัดส่วนของการกระจายของตัวเลือกที่ถูกต้องหรือที่เรียกว่า ความสมดุลของตัวเลือกที่ถูก (answer key balancing) ประเภทของข้อสอบ (item type) เช่น ข้อสอบเดี่ยว (discrete item) และชุดข้อสอบแบบ (item set) ที่ใช้สิ่งเร้าร่วมกัน หรือแบบทดสอบย่อย (testlet) และอัตราการใช้ข้อสอบซ้ำ (exposure rate) เป็นต้น เพื่อใช้ในการคัดเลือกข้อสอบให้เหมาะสมกับความสามารถของผู้สอบและมีคุณสมบัติตรงตามเงื่อนไขบังคับของการทดสอบ ดังนั้นการนำการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ไปใช้ในทางปฏิบัติจะไม่สามารถดำเนินการได้อย่างมีประสิทธิภาพโดยปราศจากการควบคุมเงื่อนไขบังคับที่ไม่ใช่ทางสถิติ เช่น การจัดสมดุลของเนื้อหา (Content Balancing) ในแบบสอบให้เป็นไปตามตารางกำหนดคุณลักษณะข้อสอบ และการควบคุมการใช้ข้อสอบซ้ำ (Item Exposure Control) ให้อยู่ในเกณฑ์ที่ยอมรับได้เพื่อป้องกันการใช้ข้อสอบที่มีอำนาจจำแนกสูงๆ บ่อยเกินไปจนกลายเป็นที่รู้จักกันแพร่หลายในกลุ่มผู้สอบโดยเฉพาะในการทดสอบที่มีผลสำคัญ (High-Stakes Testing) องค์กรการทดสอบจะกำหนดอัตราการแสดงสูงสุดอยู่ในช่วง 0.10 ถึง 0.30 นั่นหมายความว่าข้อสอบแต่ละข้อในคลังข้อสอบไม่ควรถูกนำมาใช้เกินช่วงความถี่ 10% ถึง 30% (Chang & Zhang, 2002; Chen & Doong, 2008; Chen & Lei, 2005; Chen, Lei, & Liao, 2008; Cheng & Liou, 2003)

จากการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง พบว่า วิธีการคัดเลือกข้อสอบทั้งสองวิธี คือ วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ แบบมอนติ คาร์โล (Monte Carlo CAT Method) ที่เสนอโดย Belov, Armstrong, และ Weissman (2008) และวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์วิธีแบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (constraint-weighted a-stratification method for CAT) ที่เสนอโดย Cheng และคณะ (2009) เป็นวิธีการที่มีความ

ยืดหยุ่นเหมาะสมกับการนำมาปรับใช้กับโมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย นอกจากนี้ยังสามารถประยุกต์ใช้กับการทดสอบที่มีการควบคุมเงื่อนไขบังคับที่หลากหลาย ซึ่งแต่ละวิธีมีรายละเอียดโดยสรุปดังนี้

Belov, Armstrong, และ Weissman (2008) เสนอ วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล (Monte Carlo CAT Method) ซึ่งเป็นวิธีการแบบใหม่ที่ประยุกต์ใช้แนวคิดการค้นหาแบบสโตแคสติก (Stochastic search) ร่วมกับวิธีการมอนติ คาร์โล (Monte Carlo method) เพื่อใช้ในการควบคุมเงื่อนไขบังคับด้านเนื้อหา (Content Constraints) และนำไปศึกษาเปรียบเทียบกับวิธี Shadow CAT ของ van der Linden (Li & Schafer, 2005; van der Linden, 2002, 2010; van der Linden & Chang, 2003; van der Linden & Veldkamp, 2004) ซึ่งถือว่าเป็นวิธีที่มีประสิทธิภาพสูงมีและความยืดหยุ่นในการนำไปใช้ ผลการศึกษาพบว่า วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล (Monte Carlo CAT Method) เป็นวิธีการที่มีประสิทธิภาพเหนือวิธีการ Shadow CAT คือ มีอัตราการใช้ข้อสอบซ้ำที่สังเกตได้สูงสุดต่ำกว่าและมีความสมดุลในการใช้คลังข้อสอบสูงกว่า เนื่องจากใช้วิธีการรวบแบบสอบแบบยูนิฟอร์ม มีการประมาณค่าความสามารถของผู้สอบที่แกร่งกว่า เนื่องจากการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล เป็นวิธีที่มีความไวต่อความคลาดเคลื่อนในการประมาณค่าต่ำกว่า เมื่อผู้สอบที่มีความสามารถสูงๆ หรือความสามารถต่ำๆ มีแบบแผนการตอบแบบคงที่ คือ ตอบถูกทั้งหมดหรือตอบผิดทั้งหมด ในช่วงเริ่มต้นการทดสอบ และใช้เวลาในการประมวลผลเร็วกว่าถึง 20 เท่า เนื่องจากมอนติ คาร์โล ซีเอที ใช้กระบวนการแบบสโตแคสติกเป็นส่วนประกอบในการคัดเลือกข้อสอบจึงไม่ต้องการทราบการแจกแจงความสามารถของประชากร และการแจกแจงของอัตราการใช้ข้อสอบซ้ำสังเกตได้เพื่อใช้ควบคุมการใช้ข้อสอบซ้ำ

Cheng และ Chang (2009) เสนอวิธีดัชนีลำดับความสำคัญสูงสุด (Maximum Priority Index: MPI) เพื่อใช้ควบคุมเงื่อนไขบังคับอย่างเข้มงวดในขั้นตอนการคัดเลือกข้อสอบของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ รวมทั้งประยุกต์ใช้เทคนิคการคัดเลือกข้อสอบแบบสองขั้น (two-phase item selection) ที่เสนอโดย Cheng และคณะ (2007) เพื่อทำให้วิธีการคัดเลือกข้อสอบแบบดัชนีลำดับความสำคัญสูงสุด สามารถจัดการกับเงื่อนไขบังคับได้อย่างยืดหยุ่น โดยสามารถกำหนดให้เลือกข้อสอบที่ตรงตามเงื่อนไขบังคับภายในช่วงขอบเขตบนและขอบเขตล่างแทนการกำหนดเป็นค่าคงที่ และนำวิธีการดังกล่าวมาเปรียบเทียบประสิทธิภาพกับวิธี Weighted Deviation Modeling (WDM) ซึ่งเป็นวิธีที่มีรูปแบบการคัดเลือกข้อสอบอย่างเป็นลำดับเหมือนกัน ผลการศึกษาพบว่า วิธีดัชนีลำดับความสำคัญสูงสุดมีประสิทธิภาพดีกว่า WDM อย่างเห็นได้ชัดในการจัดการเงื่อนไขข้อบังคับ อีกทั้งยังมีความถูกต้องแม่นยำของการประมาณค่าความสามารถของผู้สอบสูง และ

สามารถควบคุมอัตราการใช้ข้อสอบซ้ำให้อยู่ในระดับที่กำหนดไว้ได้ แต่มีสัดส่วนของข้อสอบที่ไม่ถูกนำมาใช้สูงถึงร้อยละ 50 ทำให้วิธี MPI มีจุดอ่อนในเรื่องการขาดความสมดุลในการใช้คลังข้อสอบ Cheng และ Chang จึงให้ข้อเสนอแนะว่าควรนำมาใช้ร่วมกับวิธีการแบ่งคลังข้อสอบออกเป็นชั้นๆ ตามค่าอำนาจจำแนกที่พัฒนาโดย Chang และ Ying (1999) เพื่อเพิ่มประสิทธิภาพของการใช้คลังข้อสอบ หลังจากนั้น Cheng และคณะ (2009) ศึกษาถึงประโยชน์ของวิธีการแบ่งคลังข้อสอบออกเป็นชั้นๆ ตามค่าอำนาจจำแนก โดยนำมาใช้ร่วมกับวิธี MPI เพื่อให้เกิดความยืดหยุ่นในการควบคุมเงื่อนไขบังคับที่ไม่ใช่ทางสถิติและสามารถใช้คลังข้อสอบได้อย่างมีประสิทธิภาพ ผลจากการศึกษาในสถานการณ์จำลอง พบว่า วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (constraint-weighted a-stratification CAT method) สามารถสร้างความสมดุลในการใช้คลังข้อสอบโดยสูญเสียความถูกความถูกต้องแม่นยำของการประมาณค่าความสามารถของผู้สอบน้อยที่สุด รวมทั้งยังควบคุมเงื่อนไขบังคับได้อย่างดีเยี่ยม โดยไม่มีการฝ่าฝืนเงื่อนไขบังคับเลย

เนื่องจากวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ทั้ง 2 วิธี คือ 1) วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติคาร์โล (Monte Carlo CAT Method) และ 2) วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (constraint-weighted a-stratification CAT method) พัฒนาขึ้นภายใต้บริบทของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่ใช้เกณฑ์การยุติการทดสอบโดยการกำหนดความยาวของแบบสอบไว้คงที่ (fixed test length termination rule) แต่คุณลักษณะสำคัญบางประการของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ คือ แบบสอบที่ผู้สอบแต่ละคนได้รับจะประกอบไปด้วยข้อสอบที่เหมาะสมกับความสามารถของตน ซึ่งในทางปฏิบัติการกำหนดให้ผู้สอบทุกคนทำข้อสอบจำนวนเท่ากันอาจมีคุณภาพของการวัดผลแตกต่างกัน ดังนั้นเกณฑ์ในอีกลักษณะหนึ่งจึงพิจารณาถึงความคลาดเคลื่อนมาตรฐานของการวัดเป็นสำคัญ ซึ่งความยาวของแบบสอบจะแปรผันไปตามระดับความคลาดเคลื่อนมาตรฐานในการประมาณค่าที่กำหนดเป็นเกณฑ์ยุติการทดสอบ ถ้ามีความคลาดคลื่นสูงกว่าเกณฑ์ที่กำหนดการทดสอบจะดำเนินต่อไปจนกระทั่งมีความคลาดเคลื่อนอยู่ในระดับที่ยอมรับได้จึงยุติการทดสอบ (ศิริชัย กาญจนวาสี, 2550; Thissen & Mislevy, 2000; Weiss & Kingsbury, 1984) ดังนั้นเพื่อขยายองค์ความรู้ใหม่ผู้วิจัยจึงสนใจนำวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ทั้ง 2 วิธี มาพัฒนาต่อในบริบทของการทดสอบที่แบบสอบมีความยาวแปรผันไปตามระดับความคลาดเคลื่อนมาตรฐานในการประมาณค่าที่กำหนดเป็นเกณฑ์ยุติการทดสอบ และเปรียบเทียบประสิทธิภาพของการทดสอบแบบปรับเหมาะทั้ง 2 วิธี ในด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถของผู้สอบ การใช้คลังข้อสอบ และการจัดการเงื่อนไขบังคับ โดยมี

เงื่อนไขในการทดสอบที่หลากหลาย และสารสนเทศที่ได้จากการวิจัยจะเป็นประโยชน์สำหรับองค์กรการทดสอบในการตัดสินใจเลือกวิธีการและกำหนดเงื่อนไขทางการทดสอบได้เหมาะสมกับสถานการณ์ที่นำไปใช้

เนื่องจากสถานการณ์ปัจจุบันการทดสอบทางการศึกษาที่เป็นมาตรฐาน (Standardized educational testing) จำนวนมากแบบสอบที่ใช้มีทั้งข้อสอบที่มีลักษณะเป็น ข้อสอบเดี่ยว (discrete item) และกลุ่มของข้อคำถามที่ใช้สิ่งเร้า (stimulus) ร่วมกัน เช่น การอ่านบทความ (reading passage) หรือ การแปลความรูปภาพและตาราง (Interpreting Figures and Tables) ตัวอย่างของสถานการณ์ที่นำแบบทดสอบย่อยมาใช้ ได้แก่ การสอบ TOEFL ในส่วนที่เป็นการอ่านเพื่อความเข้าใจ (reading comprehension) ซึ่งข้อสอบในส่วนนี้จะมีลักษณะเป็นกลุ่มของข้อสอบที่มีจำนวนประมาณ 6 ถึง 12 โดยใช้การอ่านบทความ (reading passage) เพื่อตอบข้อสอบทั้งหมดในกลุ่ม ซึ่งข้อสอบลักษณะนี้จะเรียกว่า แบบทดสอบย่อย (testlet) หรือ ชุดข้อสอบ (Item Set) เนื่องจากชุดข้อสอบมีความสัมพันธ์กับสิ่งเร้า โดยการใช้สิ่งเร้าร่วมกัน (common stimulus) ดังนั้นการตอบข้อสอบถูกหรือผิดในแบบทดสอบย่อยจึงไม่ขึ้นอยู่กับความสามารถของผู้สอบเพียงอย่างเดียวแต่ยังขึ้นอยู่กับความเข้าใจในการตีความ หรือแปลความจากสิ่งเร้านั้นเพื่อตอบคำถามด้วย ถ้าเข้าใจ แปลความ และตีความสิ่งเร้าได้ถูกจะสามารถตอบคำถามในแบบทดสอบย่อยได้ ถ้ามีความเข้าใจคลาดเคลื่อนอาจทำให้ตอบข้อสอบข้ออื่นที่อยู่ในแบบทดสอบย่อยผิดตามไปด้วย

จากสถานการณ์ดังกล่าวตามมุมมองของทฤษฎีการตอบสนองข้อสอบ ซึ่งอธิบายความสัมพันธ์ระหว่างความสามารถที่มีอยู่ภายในกับผลการตอบข้อสอบโดยใช้โค้งคุณลักษณะข้อสอบอันมีลักษณะเป็นฟังก์ชันทางคณิตศาสตร์ เรียกว่าฟังก์ชันโลจิสติก (Logistic function) หรือใกล้เคียงกับฟังก์ชันปกติสะสม (Normal ogive function) นักทดสอบจึงคิดวิธีการให้คะแนนข้อสอบที่มีลักษณะเป็นแบบทดสอบย่อย (testlet) โดยใช้ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory) เพื่ออธิบายความสัมพันธ์ โดยโมเดลที่ถูกเสนอขึ้นเพื่ออธิบายความสัมพันธ์ระหว่างความสามารถที่มีอยู่ภายในกับผลการตอบข้อสอบมีทั้งหมด 3 แนวคิด ดังนี้ 1) ข้อสอบในแบบทดสอบย่อยแต่ละข้อจะตรวจให้คะแนนแบบทวิวิภาคและวิเคราะห์ตามโมเดลการตอบสนองข้อสอบแบบโลจิส 3 พารามิเตอร์ (IRT 3 PL) 2) พิจารณาว่ากลุ่มของข้อสอบในแบบทดสอบย่อยเป็นข้อสอบข้อเดียวที่มีการให้คะแนนแบบหลายค่าและวิเคราะห์ตามทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค (Polytomous IRT) และ 3) ให้คะแนนและวิเคราะห์ตามโมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย (Testlet Respond Theory Model: TRT) (Wang, Bradlow, & Wainer, 2002) จากทั้ง 3 วิธีที่กล่าวมาแต่ละวิธีมีจุดเด่นและข้อจำกัดแตกต่างกันดังนี้

วิธีที่ 1) เป็นวิธีที่นิยมใช้กันมากเนื่องจากเป็นแนวคิดที่เข้าใจง่ายและไม่ซับซ้อนในการคำนวณ แต่ฝ่าฝืนข้อตกลงเบื้องต้นของ IRT เกี่ยวกับความเป็นอิสระ (local dependence) ที่กล่าวว่าผลการตอบข้อสอบรายข้อไม่มีความสัมพันธ์กันเนื่องจากโมเดลการตอบสนองข้อสอบมองว่าความสามารถผู้สอบ (θ) เป็นปัจจัยเดียวเท่านั้นที่มีอิทธิพลต่อผลการตอบรายข้อ เมื่อพิจารณาจะพบว่าผลการตอบข้อสอบในแบบทดสอบย่อยจะมีความสัมพันธ์กัน เนื่องจากกลุ่มของข้อสอบใช้สิ่งเร้าร่วมกัน ดังนั้นสิ่งเร้าจึงเป็นอีกปัจจัยหนึ่งที่มีอิทธิพลต่อผลการตอบข้อสอบที่นอกเหนือจากความสามารถของผู้สอบ (θ) เช่น ผู้สอบสามารถตอบข้อสอบแต่ละข้อในแบบทดสอบย่อย (testlet) ถูกหรือไม่นั้นขึ้นอยู่กับความเข้าใจในการอ่านบทความ (reading passage) ที่เป็นสิ่งเร้านั้น ซึ่งการละเลยข้อตกลงเบื้องต้นนี้ ถ้าสิ่งเร้ามีอิทธิพลทางบวกจะมีแนวโน้มที่ทำให้ความแม่นยำของการวัดที่ได้จากแบบทดสอบย่อยมีค่าสูงกว่าความเป็นจริง (overestimate) หรือ ถ้าสิ่งเร้ามีอิทธิพลทางลบจะมีแนวโน้มที่ทำให้ความแม่นยำของการวัดที่ได้จากแบบทดสอบย่อยมีค่าต่ำกว่าความเป็นจริง (underestimate)

วิธีที่ 2) เป็นวิธีที่ใช้ได้ดีในสถานการณ์ทั่วๆ ไป เมื่อแบบทดสอบย่อยเหมาะสม (fit) กับโมเดลการตอบสนองข้อสอบแบบพหุวิภาค (Polytomous IRT) คะแนนของแบบทดสอบย่อยจะแสดงเป็นจำนวนข้อที่ทำถูก สารสนเทศบางอย่าง เช่น สารสนเทศเกี่ยวกับแบบแผนการตอบ (response patterns) จะสูญหายไปเนื่องจากในการวิเคราะห์แบบทดสอบย่อยจะถูกมองว่าเป็นข้อสอบเดี่ยว (single item) ที่มีการให้คะแนนหลายแบบค่า นอกจากนี้ ในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ข้อสอบที่ถูกเลือกภายในแบบทดสอบย่อยจะเป็นการเลือกข้อสอบในลักษณะที่เรียกว่าอิสระอย่างมีเงื่อนไข (conditional independent) ซึ่งส่งผลต่อความถูกต้องในการให้คะแนน

วิธีที่ 3) เป็นการนำเอาโมเดลการตอบสนองข้อสอบ (IRT) มาปรับแก้ โดยเพิ่มเติมอิทธิพลสุ่ม (random effect) สำหรับข้อสอบที่อยู่ภายในแบบทดสอบย่อยชุดเดียวกัน จุดเด่นของการวิเคราะห์โดยใช้โมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย (Testlet Respond Theory Model: TRT) มี 2 ประการคือ 1) หน่วยของการวิเคราะห์ยังเป็นข้อสอบในแบบทดสอบย่อย และไม่มองแบบทดสอบย่อยเป็นข้อสอบข้อเดียวเหมือนในโมเดลการตอบสนองข้อสอบแบบพหุวิภาค (Polytomous IRT) ดังนั้น สารสนเทศเกี่ยวกับแบบแผนการตอบ (response patterns) ภายในแบบทดสอบย่อยจะไม่สูญหาย และ 2) แนวคิดของพารามิเตอร์ข้อสอบ เช่น อำนาจจำแนกและความยากของข้อสอบยังคงมีความสมเหตุสมผลและสามารถใช้งานได้ภายใต้โมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย

จากจุดเด่นและข้อจำกัดของแต่ละวิธีที่กล่าวมาผู้วิจัยจึงพิจารณาเลือกโมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย (TRT) มาใช้อธิบายความสัมพันธ์ระหว่างความสามารถที่มีอยู่ภายในกับ

ผลการตอบข้อสอบเพื่อช่วยในการประมาณค่าความสามารถของผู้สอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ทั้ง 2 วิธี คือ 1) วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติคาร์โล (Monte Carlo CAT Method) และ 2) วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (constraint-weighted a-stratification CAT method) ซึ่งนำไปสู่คำถามการวิจัยต่อไป

คำถามวิจัย

เมื่อนำแนวคิดวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติคาร์โล (Monte Carlo CAT Method) และวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (constraint-weighted a-stratification CAT method) ที่พัฒนาขึ้นภายใต้โมเดลการตอบสนองข้อสอบ (item respond model) มาพัฒนาต่อยอดเพื่อใช้กับโมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย (testlet respond model) ในบริบทของการทดสอบที่ผู้สอบต้องทำแบบสอบที่มีความยาวต่างกัน ภายใต้เงื่อนไขการออกแบบการทดสอบที่กำหนดอัตราการควบคุมอัตราการแสดงแบบทดสอบย่อยสูงสุดแตกต่างกัน และมีขนาดคลังข้อสอบแตกต่างกัน โดยผู้วิจัยตั้งประเด็นคำถามวิจัยไว้ดังนี้

1. วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบใด และในเงื่อนไขใด มีประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถของผู้สอบสูงกว่ากัน โดยพิจารณาจาก 1) ค่าความลำเอียงเฉลี่ย 2) ค่าความแปรปรวนของความคลาดเคลื่อน 3) ค่าสหสัมพันธ์ระหว่างค่าความสามารถจริงกับค่าประมาณความสามารถ 4) ความยาวเฉลี่ยของแบบสอบ (ความยาวของแบบสอบ: เมื่อ SEE มีขนาดใกล้เคียงกัน (Chang & Ansley, 2003; Revuelta & Ponsoda, 1998; Wainer, 1992)

2. วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบใด และในเงื่อนไขใด มีประสิทธิภาพด้านการใช้คลังข้อสอบสูงกว่ากัน โดยพิจารณาจาก 1) อัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ 2) อัตราการทับซ้อนของแบบสอบเฉลี่ย 3) ความสมดุลของการใช้คลังข้อสอบในภาพรวม (ค่าไคสแควร์) 4) จำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ และ 5) จำนวนแบบทดสอบย่อยที่ถูกนำมาใช้มากเกินไป

วัตถุประสงค์ของการวิจัย

เพื่อศึกษาประสิทธิภาพของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่ใช้สำหรับโมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย 2 วิธี คือ 1) วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที (Monte Carlo CAT Method: MCC) และ 2) วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ

(constraint-weighted a-stratification CAT method: CWA) เมื่อกำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด และขนาดคลังข้อสอบที่ต่างกัน โดยมีวัตถุประสงค์ย่อยดังนี้

1. เพื่อเปรียบเทียบประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถของผู้สอบ (Efficiency of Measurement Precision) ระหว่างวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที กับวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ เมื่อกำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด และขนาดคลังข้อสอบต่างกัน

2. เพื่อเปรียบเทียบประสิทธิภาพด้านการใช้คลังข้อสอบ (Efficiency of pool utilization) ระหว่างวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที กับวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ เมื่อกำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด และขนาดคลังข้อสอบที่ต่างกัน

สมมติฐานการวิจัย

จากการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง พบว่า ไม่มีการศึกษาวิจัยเพื่อเปรียบเทียบประสิทธิภาพด้าน ความถูกต้องแม่นยำของการประมาณค่า และสมดุลการใช้คลังข้อสอบ ของวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ ระหว่างวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที และวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับโดยตรง แต่มีการศึกษาภายใต้สถานการณ์การทดสอบที่ใช้โมเดลการตอบสนองข้อสอบและแบบสอบมีความยาวคงที่

ดังนั้นผู้วิจัยจึงตั้งสมมติฐานโดยอาศัยข้อมูลจากงานวิจัยของ Belov, Armstrong, และ Weissman (2008) ที่พัฒนาวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติคาร์โล จุดเด่นของวิธีการนี้คือให้ผลการประมาณค่าความสามารถผู้สอบที่มีความถูกต้องแม่นยำสูงพร้อมทั้งสร้างความสมดุลในการใช้คลังข้อสอบ สามารถควบคุมเงื่อนไขบังคับได้หลากหลายและยืดหยุ่น เนื่องจากอาศัยหลักการทางสโตคาสติก และเลือกข้อสอบที่ตรงตามเงื่อนไขมาสร้างเป็นลำดับเชิงสุ่มของข้อสอบ แล้วค้นหาข้อสอบที่ให้สารสนเทศสูงสุดจากลำดับเชิงสุ่มเพื่อนำไปใช้กับผู้สอบ และงานวิจัยของ Cheng et al. (2009) ที่ศึกษาประสิทธิภาพของวิธีการคัดเลือกข้อสอบแบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ ใช้วิธีการแบ่งคลังข้อสอบออกเป็นระดับชั้นตามค่าอำนาจจำแนก (a-Stratification) วิธีการที่กล่าวถึงนี้ใช้ดัชนีลำดับความสำคัญสูงสุด (maximum priority index) ช่วยถ่วงน้ำหนักเงื่อนไขบังคับโดยการให้น้ำหนักของเงื่อนไขบังคับแตกต่างกันตามระดับความสำคัญ (Cheng & Chang, 2009) และใช้วิธีการคัดเลือกข้อสอบสองขั้นตอน (Two-

Phase Item Selection Procedure) เพื่อช่วยให้เกิดความยืดหยุ่นในการควบคุมเงื่อนไขบังคับด้านเนื้อหา (Cheng et al., 2007) และใช้วิธีการแบ่งคลังข้อสอบออกเป็นระดับชั้นตามค่าอำนาจจำแนก (a-Stratification) เพื่อพยายามควบคุมการใช้ข้อสอบภายในคลังข้อสอบให้มีการกระจายอย่างสม่ำเสมอ โดยพิจารณาประสิทธิภาพของความสามารถในการประมาณค่าจากค่าเฉลี่ยของ Bias, MSE และเปรียบเทียบประสิทธิภาพสัมพัทธ์กับวิธีการคัดเลือกข้อสอบที่ใช้เกณฑ์สารสนเทศสูงสุด (maximum information criterion: MIC) และวิธีการคัดเลือกข้อสอบอย่างสุ่ม (Randomized) และพิจารณาประสิทธิภาพของการควบคุมการใช้ข้อสอบซ้ำจากค่า x^2 ผลการศึกษา พบว่าประสิทธิภาพของความสามารถในการประมาณค่า อยู่ที่ประมาณ 0.72 เมื่อเทียบกับ MIC ส่วนประสิทธิภาพของการควบคุมการใช้ข้อสอบซ้ำ มีค่า x^2 อยู่ที่ประมาณ 8.29 ซึ่งถือว่าใช้ข้อสอบในคลังได้ค่อนข้างอย่างสมดุล จากข้อค้นพบดังกล่าวจึงนำไปสู่การตั้งสมมติฐานเพื่อตอบคำถามวิจัยแต่ละข้อดังนี้

1. วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที น่าจะมีประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถสูงกว่าวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ

2. วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที น่าจะมีประสิทธิภาพด้านการใช้คลังข้อสอบสูงกว่าวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ

เหตุผลที่ผู้วิจัยคาดว่าวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที จะมีประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถ และด้านการใช้คลังข้อสอบ สูงกว่าวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ เนื่องจากวิธีแบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ ควบคุมการใช้ข้อสอบภายในคลังข้อสอบให้มีการกระจายอย่างสม่ำเสมอโดยการแบ่งคลังข้อสอบออกเป็นระดับชั้นตามค่าอำนาจจำแนก (a-Stratification) ซึ่งข้อสอบในแต่ละชั้นจำเป็นต้องมีจำนวนมากเพียงพอเพื่อทำให้ข้อสอบภายในชั้นนั้นๆ สามารถรองรับผู้สอบได้ครอบคลุมทุกช่วงความสามารถ และในการศึกษาครั้งนี้คลังข้อสอบมีขนาดเพียง 600 ข้อ และ 800 ข้อ โดยข้อสอบเหล่านี้ต้องต้องจัดกลุ่มกันตามชุดของแบบทดสอบย่อย ดังนั้นจำนวนแบบทดสอบย่อยที่มีอาจจัดเรียงเป็นระดับชั้นได้ไม่ครอบคลุมช่วงความสามารถ หรือแบบทดสอบย่อยภายในแต่ละชั้นมีค่าอำนาจจำแนกแตกต่างกันสูงมากจนทำให้แบบทดสอบย่อยที่มีค่าอำนาจจำแนกสูงๆ ภายในชั้น ได้รับการคัดเลือกนำมาใช้กับผู้สอบบ่อยครั้งเกินไป

ขอบเขตของการวิจัย

การวิจัยครั้งนี้ใช้ข้อมูลจากการจำลองข้อมูลมีรายละเอียดดังนี้

1. การจำลองข้อมูลเพื่อสร้างคลังสอบ ทำโดยกำหนดค่าพารามิเตอร์อำนาจจำแนกให้มีลักษณะการแจกแจงแบบ log normal มีค่าเฉลี่ยเป็น 0 และค่าส่วนเบี่ยงเบนมาตรฐานของ log เป็น .022 เพื่อไม่ให้ค่าอำนาจจำแนกติดลบ พารามิเตอร์ความยากมีลักษณะการแจกแจงแบบโค้งปกติ มีค่าเฉลี่ยเป็น 0 และค่าส่วนเบี่ยงเบนมาตรฐานเป็น 1 พารามิเตอร์โอกาสในการเดามีลักษณะการแจกแจงแบบเบต้า มีค่าแอลฟ่าเป็น 2 และ 10 เพื่อให้ค่าที่ได้มีการกระจายอยู่บนช่วง 0 ถึงประมาณ 0.2 ข้อสอบแต่ละข้อจะถูกเลือกเข้าไปในแต่ละแบบทดสอบย่อยอย่างสุ่ม ภายในคลังข้อสอบจะถูกแบ่งเป็น 3 ขอบเขตเนื้อหา โดยแบบทดสอบย่อยแต่ละชุดจะถูกเลือกอย่างสุ่มเข้าไปในแต่ละขอบเขตเนื้อหา โดยคลังข้อสอบของงานวิจัยนี้ถูกออกแบบให้มีลักษณะเป็นแบบ Unidimensional Models คือ มุ่งวัดความความสามารถของผู้สอบในคุณลักษณะเดียวคล้ายกับการโมเดลการตอบสนองข้อสอบ (IRT) และมีจุดความแตกต่างกัน คือ กลุ่มของข้อสอบ (4 ข้อ) มีความสัมพันธ์กันเนื่องจากใช้สิ่งเร้าร่วมกัน ผู้วิจัยจึงเรียกกลุ่มของข้อสอบที่ใช้สิ่งเร้าร่วมกันว่า “แบบทดสอบย่อย” คลังข้อสอบที่ใช้ในการวิจัยครั้งนี้มีจำนวน 2 คลังข้อสอบ คือ 1) คลังข้อสอบขนาด 600 ข้อ มีแบบทดสอบย่อยจำนวน 150 ชุด และ 2) คลังข้อสอบขนาด 800 ข้อ มีแบบทดสอบย่อยจำนวน 200 ชุด โดยที่ข้อสอบภายในแบบทดสอบย่อยนั้นจะวัดในขอบเขตเนื้อหาเดียวกันและมีค่าความยากใกล้เคียงกันมากที่สุด

2. การจำลองข้อมูลเพื่อสร้างผู้สอบ กำหนดค่าพารามิเตอร์ของผู้สอบ โดยกำหนดให้ค่าความสามารถจริงของผู้สอบ ที่สร้างขึ้นโดยสุ่มจากการแจกแจงโค้งปกติมาตรฐาน มีค่าเฉลี่ยเป็น 0 และค่าส่วนเบี่ยงเบนมาตรฐานเป็น 1 และ กำหนดพารามิเตอร์อิทธิพลของแบบทดสอบย่อยให้กับผู้สอบโดยเลือกอย่างสุ่มจากการแจกแจงปกติที่มีค่าเฉลี่ยเป็น 0 และความแปรปรวนเท่ากับ $\gamma_{id(j)}$

3. การจำลองแบบแผนการตอบข้อสอบ เนื่องจากภายในคลังข้อสอบประกอบด้วยข้อสอบที่มีลักษณะเป็นชุดแบบทดสอบย่อย โดยภายในแต่ละแบบทดสอบย่อยข้อสอบจะมีความสัมพันธ์กันเพราะใช้สิ่งเร้าร่วมกัน ดังนั้นจึงใช้โมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย แบบ 3 พารามิเตอร์ (3 Parameter Logistic Testlet Response Theory Model: 3PL TRT) มาใช้ในการจำลองข้อมูลเพื่ออธิบายความสัมพันธ์ระหว่างความสามารถที่มีอยู่ภายในกับผลการตอบแบบทดสอบย่อย โดยผลการตอบข้อสอบมีลักษณะเป็น 2 ค่า คือ ตอบถูก ได้คะแนนเท่ากับ 1 ตอบผิด ได้คะแนนเท่ากับ 0 (Dichotomous Scoring) ซึ่งการจำลองแบบแผนการตอบข้อสอบจะถูกทำซ้ำ 10 ชุด โดยแต่ละชุดจะสุ่มความสามารถของผู้สอบขึ้นมาใหม่ทุกครั้ง แบบแผนการตอบข้อสอบทั้ง 10 ชุด จะถูกนำไปใช้กับการทดสอบแบบปรับเหมาะในแต่ละเงื่อนไข

4. โปรแกรมคอมพิวเตอร์ที่ใช้ในการทดสอบแบบปรับเหมาะที่ใช้ในการวิจัยครั้งนี้พัฒนาขึ้นโดยใช้โปรแกรม R มีวิธีการคัดเลือกแบบทดสอบย่อย 4 วิธีหลักๆ ดังนี้ 1) วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที 2) วิธีการแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ 3) วิธีสารสนเทศสูงสุด และ 4) วิธีการเลือกแบบทดสอบย่อยแบบสุ่ม เมื่อดำเนินการทดสอบเมื่อแบบทดสอบย่อยจะถูกเลือกเพื่อนำไปใช้กับผู้สอบ ผู้สอบต้องทำข้อสอบทุกข้อในแบบทดสอบย่อย จากนั้นระบบจะจัดแบบทดสอบย่อยชุดต่อไปให้กับผู้สอบตามขั้นตอนวิธี (algorithm) ของวิธีการคัดเลือกแบบทดสอบย่อยแต่ละวิธี และการยุติการทดสอบพิจารณาจากค่าความคลาดเคลื่อนมาตรฐานในการประมาณค่าความสามารถของผู้สอบ (SEE) น้อยกว่าหรือเท่ากับ 0.3 เนื่องจากมีความตรงตามสภาพสูง (รังสรรค์ มณีเล็ก, 2540) หรือแบบสอบมีความยาวมากกว่า 60 ข้อ (15 ชุด)

5. ตัวแปรที่ใช้ในการศึกษาครั้งนี้ประกอบด้วย

5.1 ตัวแปรอิสระ ได้แก่

5.1.1 วิธีการเลือกแบบทดสอบย่อย

1) วิธีมอนติ คาร์โล ซีเอที

2) วิธีแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ

3) วิธีการเลือกแบบทดสอบย่อยที่ให้สารสนเทศสูงสุด ผู้วิจัยใช้เป็นฐานในการเปรียบเทียบเนื่องจากมีประสิทธิภาพในด้านแม่ยำของการวัดสูงที่สุด

4) วิธีการเลือกแบบทดสอบย่อยแบบสุ่ม ผู้วิจัยใช้เป็นฐานในการเปรียบเทียบเนื่องจากมีประสิทธิภาพในด้านการใช้คลังข้อสอบสูงที่สุด

5.1.2 ขนาดคลังข้อสอบ

1) จำนวน 600 ข้อ (แบบทดสอบย่อย 150 ชุด)

2) จำนวน 800 ข้อ (แบบทดสอบย่อย 200 ชุด)

5.1.3 อัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด

1) อัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด เท่ากับ 10 เปอร์เซ็นต์

2) อัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด เท่ากับ 15 เปอร์เซ็นต์

3) อัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด เท่ากับ 20 เปอร์เซ็นต์

4) อัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด เท่ากับ 25 เปอร์เซ็นต์

5.2 ตัวแปรตาม ได้แก่

5.2.1 ประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถของผู้สอบ พิจารณาได้จากเกณฑ์ ต่อไปนี้

1) ค่าเฉลี่ยสหสัมพันธ์ระหว่างค่าประมาณความสามารถกับค่าความสามารถจริง

2) ค่าเฉลี่ยความลำเอียง (Bias)

3) ค่าเฉลี่ยความแปรปรวนของความคลาดเคลื่อน (MSE)

4) ความยาวเฉลี่ยของแบบสอบและค่าเฉลี่ยคามคลาดเคลื่อนมาตรฐานในการประมาณค่า (SEE) มีขนาดใกล้เคียงกัน (Chang & Ansley, 2003; Revuelta & Ponsoda, 1998; Wainer, 1992)

5.2.2 ประสิทธิภาพด้านสมดุลการใช้แบบทดสอบย่อยในคลังข้อสอบ พิจารณาได้จากเกณฑ์ ต่อไปนี้

1) ค่าเฉลี่ยอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้

2) ค่าเฉลี่ยอัตราการทับซ้อนของแบบสอบ

3) ค่าเฉลี่ยของไคสแควร์

4) ค่าเฉลี่ยจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้

5) ค่าเฉลี่ยจำนวนทดสอบย่อยที่ถูกนำมาใช้มากเกินไป

ข้อจำกัดของการวิจัย

เนื่องจากการวิจัยครั้งนี้ศึกษาภายใต้โมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย (TRT) แบบ 3 พารามิเตอร์ และทำการศึกษาทั้งหมด 16 เงื่อนไข โดยแต่ละเงื่อนไขใช้ตัวอย่างจำนวน 1,000 ตัวอย่าง ด้วยโมเดลที่มีความซับซ้อนทำให้การประมาณค่าความสามารถผู้สอบใช้เวลาในการประมวลผลมาก ผู้วิจัยจึงกำหนดจำนวนครั้งในการทดลองซ้ำในแต่ละเงื่อนไขโดยการศึกษาจากเอกสารและงานวิจัยที่ทำการจำลองข้อมูลภายใต้โมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย พบว่า Boyd (2003) และ Keng (2008) ก็ทราบถึงปัญหาดังกล่าวจึงกำหนดจำนวนครั้งในการทดลองซ้ำ 10 ครั้ง ดังนั้นผู้วิจัยจึงทำการทดลองซ้ำ 10 ครั้ง ตามวิธีดำเนินการวิจัยของ Boyd (2003) และ

Keng (2008) แต่เมื่อพิจารณาเปรียบเทียบกับการทดลองที่ใช้โมเดล IRT แบบ 1 พารามิเตอร์ ในงานวิจัยของ (ต่าย เซี่ยงฉี, 2534) ทำการทดลองซ้ำในแต่ละเงื่อนไข 40 ครั้ง เนื่องจากโมเดล IRT แบบ 1 พารามิเตอร์ มีความซับซ้อนในการคำนวณน้อยกว่าโมเดล TRT แบบ 3 พารามิเตอร์

คำจำกัดความที่ใช้ในการวิจัย

1. การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที หมายถึง การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่คัดเลือกแบบทดสอบย่อย โดยการใช้วิธีการ Shadow tests เพื่อสร้างแบบสอบเงาที่มีคุณสมบัติตรงตามเงื่อนไขบังคับที่กำหนด นำแบบทดสอบย่อยภายในแบบสอบเงาทั้งหมดมาเรียงต่อกันเป็นลำดับ และสุ่มแบบทดสอบย่อยจำนวน k ชุด จากลำดับ จากนั้นจึงค้นหาแบบทดสอบย่อยที่ให้ค่าสารสนเทศสูงสุดที่ระดับค่าประมาณความสามารถของผู้สอบเพื่อนำไปใช้เป็นแบบทดสอบย่อยชุดถัดไป วิธีการนี้ประยุกต์มาจากแนวคิดของ Belov et al. (2008)

2. การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ หมายถึง การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่คัดเลือกแบบทดสอบย่อยโดยการแบ่งคลังข้อสอบออกเป็นชั้นตามลำดับค่าของอำนาจจำแนกเฉลี่ยของแบบทดสอบย่อย และเลือกแบบทดสอบย่อยจากชั้นที่มีค่าอำนาจจำแนกต่ำไปหาชั้นที่มีค่าอำนาจจำแนกสูง การคัดเลือกแบบทดสอบย่อยจะพิจารณาจากผลคูณของค่าสารสนเทศของแบบทดสอบย่อยกับดัชนีลำดับความสำคัญของเงื่อนไขบังคับ วิธีการนี้ประยุกต์มาจากแนวคิดของ Cheng et al. (2009)

3. อิทธิพลของแบบทดสอบย่อย ($\gamma_{id(j)}$) หมายถึง อิทธิพลสุ่มข้อสอบข้อที่ i ซึ่งอยู่ในแบบทดสอบย่อยฉบับ d ที่มีต่อผู้สอบคนที่ j

4. ขนาดคลังข้อสอบ (Item Pool Size) หมายถึง จำนวนข้อสอบทั้งหมดที่บรรจุในคลังข้อสอบ การวิจัยครั้งนี้กำหนดขนาดคลังข้อสอบไว้ 2 ขนาด คือ คลังข้อสอบจำนวน 600 ข้อ (แบบทดสอบย่อย 150 ชุด) และคือ คลังข้อสอบจำนวน 800 ข้อ (แบบทดสอบย่อย 200 ชุด)

5. เกณฑ์การยุติการทดสอบ (Stopping Rule) หมายถึง ข้อกำหนดเพื่อใช้ในการสิ้นสุดการทดสอบ ซึ่งกำหนดโดยใช้ระดับความคลาดเคลื่อนมาตรฐานของการประมาณค่าน้อยกว่าหรือเท่ากับ 0.3 หรือความยาวของแบบสอบเกิน 15 แบบทดสอบย่อย (60 ข้อ)

6. อัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (maximum testlet exposure rate: $r_{\max}$) หมายถึง ตัวเลขที่แสดงถึง ค่าเป้าหมายที่ตั้งขึ้นเพื่อนำไปใช้วางแผนป้องกันการใช้แบบทดสอบย่อยแต่ละฉบับในคลังข้อสอบบ่อยครั้งเกินไป โดยการวิจัยครั้งนี้กำหนดอัตราการแสดงแบบทดสอบสูงสุดไว้ 4 ระดับ คือ 10, 15, 20 และ 25 เปอร์เซ็นต์

7. ประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถของผู้สอบ (efficiency of measurement precision) หมายถึง ความสามารถของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ในการประมาณค่าความสามารถของผู้สอบให้เท่ากับหรือใกล้เคียงค่าความสามารถจริงของผู้สอบมากที่สุดและใช้ข้อสอบจำนวนน้อยกว่าขณะที่มีความแม่นยำใกล้เคียงกัน (Wainer, 1992) พิจารณาจาก 1) ค่าความลำเอียงเฉลี่ย 2) ค่าความแปรปรวนของความคลาดเคลื่อน 3) ค่าสหสัมพันธ์ระหว่างค่าความสามารถจริงกับค่าประมาณความสามารถ 4) ความยาวเฉลี่ยของแบบสอบและค่าเฉลี่ยคามคลาดเคลื่อนมาตรฐานในการประมาณค่า

7.1 ค่าความลำเอียง (Bias) หมายถึง ค่าความคลาดเคลื่อนที่บอกความแตกต่างระหว่างค่าความสามารถจริงของผู้สอบกับค่าความสามารถที่ประมาณค่าจากการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ ซึ่งค่าความลำเอียงสะท้อนให้เห็นถึงความถูกต้อง (Accuracy) ของการประมาณค่าความสามารถของผู้สอบ โดยมีหลักเกณฑ์ในการพิจารณาจากการเปรียบเทียบค่าความลำเอียงที่ได้ดังนี้ ถ้าค่าความลำเอียงเข้าใกล้ศูนย์มากกว่าสะท้อนให้เห็นถึงการประมาณค่าความสามารถที่ได้ใกล้เคียงกับค่าที่แท้จริงมากกว่า โดยเครื่องหมายที่ได้จะเป็นตัวสะท้อนถึงความคลาดเคลื่อนจากการประมาณค่าที่แตกต่างจากค่าความสามารถจริงในลักษณะใด (Dekking, Kraaikamp, Lopuhaa, & Meester, 2005b) นั่นคือ ถ้าดัชนี Bias ติดลบ (-) แสดงว่า มีการประมาณค่าความสามารถของผู้สอบต่ำกว่าความเป็นจริง (underestimate) ถ้าเป็นบวก (+) แสดงว่ามีการประมาณค่าความสามารถของผู้สอบสูงกว่าความเป็นจริง (overestimate) ซึ่งสามารถแสดงได้ดังสมการที่ 1.1

7.2 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error: MSE) หมายถึง ค่าความแปรปรวนของความคลาดเคลื่อนในการประมาณค่าความสามารถของผู้สอบ เป็นค่าที่บ่งชี้ว่าความคลาดเคลื่อนที่เกิดขึ้นจากการประมาณค่ามีการกระจายมากน้อยเพียงใด และค่า MSE สามารถนำมาใช้สำหรับเปรียบเทียบประสิทธิภาพของตัวประมาณค่าได้ มีหลักเกณฑ์ในการพิจารณา คือ ตัวประมาณค่าตัวใดมีความแปรปรวนน้อยกว่าแสดงว่าตัวประมาณค่าตัวนั้นมีประสิทธิภาพสูงกว่า ซึ่งตัวประมาณค่าในงานวิจัยครั้งนี้ คือ วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ และประสิทธิภาพของตัวประมาณค่า คือ ความสามารถในการประมาณค่าที่ทำให้ได้ค่าประมาณพารามิเตอร์เท่ากับค่าพารามิเตอร์ที่แท้จริง หรือมีความคลาดเคลื่อนน้อยที่สุด (Dekking, Kraaikamp, Lopuhaa, & Meester, 2005a) สามารถแสดงได้ดังสมการที่ 1.2

8. ประสิทธิภาพของการใช้คลังข้อสอบ (efficiency of pool utilization หรือ efficiency of item bank usage) หมายถึง ความสามารถของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ในการใช้คลังข้อสอบได้อย่างสมดุล คือ ข้อสอบแต่ละข้อมีโอกาสถูกนำออกมาใช้เท่าๆ กัน สามารถ

พิจารณาได้จาก 1) อัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ 2) อัตราการทับซ้อนของแบบสอบเฉลี่ย 3) ความสมดุลของการใช้คลังข้อสอบในภาพรวม 4) จำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ และ 5) จำนวนแบบทดสอบย่อยที่ถูกนำมาใช้มากเกินไป

8.1 อัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ (testlet exposure rate observed: r_{obs}) หมายถึง สัดส่วนระหว่างจำนวนครั้งของแบบทดสอบย่อยที่ได้รับการคัดเลือกแล้วนำไปใช้แก่ผู้สอบกับจำนวนผู้สอบทั้งหมดที่ทำการทดสอบ เช่น แบบทดสอบย่อยชุดที่ 1 ถูกนำไปใช้กับผู้สอบ 200 ครั้ง ในการทดสอบนั้นมีผู้สอบทั้งหมด 1,000 คน เพราะฉะนั้น แบบทดสอบย่อยชุดที่ 1 จะมีอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ เท่ากับ 0.2 หรือ 20 เปอร์เซ็นต์

8.2 อัตราการทับซ้อนของแบบสอบเฉลี่ย (average test overlap หรือ Average Between-Test Overlap: $\bar{t}$) หมายถึง ค่าเฉลี่ยเลขคณิตของสัดส่วนของข้อสอบในแบบสอบฉบับหนึ่งซึ่งปรากฏอยู่ในแบบสอบฉบับอื่นในทุกคู่ของแบบสอบที่เป็นไปได้ควรมีอัตราการนำไปใช้เท่าๆ กัน

8.3 ความสมดุลของการใช้คลังข้อสอบในภาพรวม (overall of pool usage balance) หมายถึง การตรวจสอบความแตกต่างระหว่างการแจกแจงของอัตราการใช้ข้อสอบซ้ำสังเกตได้และอัตราการใช้ข้อสอบซ้ำในอุดมคติที่ข้อสอบทุกข้อมีโอกาสถูกนำมาใช้เท่าๆ กัน สามารถทดสอบได้โดยใช้ สถิติไคสแควร์ (x^2) ถ้า ไคสแควร์ (x^2) มีค่าต่ำแสดงว่าข้อสอบส่วนใหญ่ในคลังข้อสอบถูกนำไปใช้งานในอัตราที่เท่าๆ กัน

8.4 จำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ (Numbers of testlets are neverexposed: $N_{neverexposed}$) หมายถึง จำนวนข้อสอบในคลังข้อสอบที่ไม่ถูกนำออกมาใช้เลย

8.5 จำนวนแบบทดสอบย่อยที่ถูกนำมาใช้มากเกินไป (Numbers of testlets are overexposed: $N_{overexposed}$) หมายถึง ชุดของข้อสอบที่มีอัตราการนำมาใช้ถี่เกินกว่าที่กำหนด

ประโยชน์ที่ได้รับ

1. ประโยชน์ทางด้านวิชาการ

1.1 ได้ขยายองค์ความรู้ในด้านการควบคุมเงื่อนไขบังคับในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์

1.2 ผลการวิจัยให้ข้อมูลสารสนเทศเกี่ยวกับผลการเปรียบเทียบประสิทธิภาพของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่มีวิธีการแตกต่างกัน 2 วิธี ระหว่างวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที (Monte Carlo CAT Method) และวิธีการทดสอบแบบ

ปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (constraint-weighted a-stratification CAT method) ในเงื่อนไขของการศึกษา คือ ขนาดของคลังข้อสอบ จำนวนขอบเขตเนื้อหาในคลังข้อสอบ เกณฑ์การยุติการทดสอบ และอัตราการใช้ข้อสอบซ้ำสูงสุด ที่แตกต่างกัน ทำให้ได้ผลสรุปที่ชัดเจนขึ้นเกี่ยวกับวิธีการเลือกใช้รูปแบบของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ในสถานการณ์การศึกษาที่แตกต่างกัน และเป็นพื้นฐานที่สำคัญในการวิจัยเกี่ยวกับการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์

2. ประโยชน์ทางด้านการนำไปใช้

2.1 ผลการวิจัยสามารถนำไปใช้ประกอบการตัดสินใจในการเลือกใช้วิธีการในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่เหมาะสมกับสถานการณ์ที่ต้องการใช้มากที่สุด

2.2 ผลการวิจัยเป็นแนวทางในการนำไปใช้ในการบริหารจัดการ การทดสอบให้มีประสิทธิภาพมากยิ่งขึ้น ในด้านต้นทุนในการเลือกใช้ขนาดคลังข้อสอบ การควบคุมเงื่อนไขบังคับในการทดสอบ และความถูกต้องแม่นยำของการประมาณค่า

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

จากการศึกษาเอกสาร งานวิจัยและตำราต่าง ๆ ในเรื่องที่เกี่ยวข้องกับการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ ผู้วิจัยได้นำเสนอเนื้อหาออกเป็น 6 ตอน คือ ตอนที่ 1 การทดสอบแบบปรับเหมาะกับความสามารถของผู้สอบ (Adaptive testing) ตอนที่ 2 โมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย (Testlet Response Theory Model: TRT) ตอนที่ 3 วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ แบบมอนติ คาร์โล ซีเอที (Monte Carlo CAT Method) ตอนที่ 4 วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (constraint-weighted a-stratification CAT method) ตอนที่ 5 การโปรแกรมเชิงคณิตศาสตร์ (Mathematical Programming) ตอนที่ 6 งานวิจัยที่เกี่ยวข้องกับการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ โดยแต่ละตอนมีรายละเอียดดังนี้

ตอนที่ 1 การทดสอบแบบปรับเหมาะกับความสามารถของผู้สอบ

1.1 หลักการ จุดเด่น และ การใช้งานการทดสอบแบบปรับเหมาะในปัจจุบัน

การทดสอบแบบปรับเหมาะกับความสามารถของผู้สอบ คือ การทดสอบที่ผู้สอบแต่ละคนจะได้รับข้อสอบที่ต่างกัน ข้อสอบที่ผู้สอบแต่ละคนได้รับจะมีความยากเหมาะสมกับความสามารถของตัวผู้สอบเอง การคัดเลือกข้อสอบสำหรับผู้สอบแต่ละคนอยู่บนพื้นฐานของผลการตอบข้อสอบที่ผู้สอบได้ทำไปแล้วก่อนหน้า เมื่อผู้สอบทำข้อสอบข้อแรกซึ่งได้จากคลังข้อสอบแล้ว จะมีการวิเคราะห์ระดับความสามารถหรือประเมินความสามารถของผู้สอบเบื้องต้น เพื่อคัดเลือกข้อสอบข้อต่อไปที่มีค่าความยาก และค่าอำนาจจำแนกเหมาะสมที่จะใช้วัดระดับความสามารถของผู้สอบและทำการประมาณค่าระดับความสามารถของผู้สอบใหม่ จากนั้นก็จะเลือกข้อที่เหมาะสมข้อต่อไป โดยใช้หลักการที่ว่าถ้าการทำข้อที่ผ่านมาถูกข้อถัดไปจะยากขึ้น แต่ถ้าการทำข้อที่ผ่านมาผิดข้อถัดไปจะง่ายลงกระบวนการนี้จะดำเนินต่อไปเรื่อยๆ จนสามารถประมาณค่าระดับความสามารถของผู้สอบได้อย่างเชื่อถือได้ตามเกณฑ์ที่กำหนดไว้ การทดสอบจึงจะยุติลง (ศิริชัย กาญจนวาสี, 2550)

การทดสอบแบบปรับเหมาะกับความสามารถของผู้สอบใช้เทคโนโลยีทางคอมพิวเตอร์เข้ามาช่วยในการบริหารการทดสอบ การทดสอบลักษณะนี้จึงมักเรียกกันว่า การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (Computerized Adaptive Testing: CAT) จากคุณลักษณะเด่นของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่กล่าวมาทำให้การทดสอบลักษณะนี้มีประสิทธิภาพสูงกว่าการทดสอบแบบดั้งเดิมที่ใช้กระดาษและดินสอ (Paper-and-Pencil: P&P) หลายประการคือ 1) ช่วยลด

จำนวนข้อและเวลาที่ใช้การทดสอบให้น้อยลงเนื่องจากผู้สอบจะได้รับข้อสอบที่เหมาะสมกับความสามารถของตนเอง ผู้สอบที่มีความสามารถสูงไม่จำเป็นต้องทำข้อสอบข้อที่ง่าย ในขณะเดียวกันผู้สอบที่มีความสามารถต่ำก็ไม่จำเป็นต้องทำข้อสอบที่ยากมากๆ ดังนั้นผู้สอบแต่ละคนไม่จำเป็นต้องได้รับข้อสอบที่เหมือนกันและจำนวนเท่ากัน 2) จากจุดเด่นในข้อแรกทำให้การประมาณค่าความสามารถของผู้สอบแต่ละคนได้อย่างถูกต้องแม่นยำสูง เนื่องจากข้อสอบที่ไม่เหมาะกับความสามารถจะให้สารสนเทศสำหรับผู้สอบน้อยมาก 3) สามารถบันทึกและรายงานผลการทดสอบได้ทันทีเนื่องจากใช้คอมพิวเตอร์เข้ามาช่วยดำเนินการทดสอบ และ 4) ด้วยประโยชน์ของเทคโนโลยีทางคอมพิวเตอร์และระบบเครือข่ายจึงมีความยืดหยุ่นในการจัดตารางการทดสอบผู้สอบสามารถสอบได้ต่างเวลากันตามความพร้อมของแต่ละคน

1.2 ประเภทของการทดสอบแบบปรับเหมาะกับความสามารถของผู้สอบ

ประเภทของการทดสอบแบบปรับเหมาะกับความสามารถของผู้สอบ (ต่าย เซี่ยงฉี, 2534; รังสรรค์ มณีเล็ก, 2540; สิริลักษณ์ เกษรปทุมานันท์ & ณัฏฐภรณ์ หลาวทอง, 2007) ได้แบ่งวิธีการคัดเลือกข้อสอบที่มีระดับความยากเหมาะสมกับความสามารถของผู้สอบเป็น 2 วิธี คือ 1) ยุทธวิธีสองขั้นตอน (Two–stage strategies) และ 2) ยุทธวิธีหลายขั้นตอน (Multi–stage strategies) ประกอบด้วย 2 รูปแบบหลักๆ คือ (1) รูปแบบแยกทางคงที่ (Fixed branching model) และ (2) รูปแบบแยกทางแปรผัน (Variable branching model) รายละเอียดของการทดสอบแบบต่าง ๆ สรุปได้ดังนี้

1.2.1 ยุทธวิธีสองขั้นตอน

การทดสอบแบบปรับเหมาะกับความสามารถของผู้สอบโดยใช้ยุทธวิธีสองขั้นตอน โดยมีการดำเนินการสอบเป็น 2 ขั้นตอน ขั้นตอนแรกทำการทดสอบด้วยแบบสอบกำหนดทิศทาง (Routing system) ซึ่งเป็นแบบสอบที่มีจำนวนข้อสอบประมาณ 10 ข้อ ประกอบด้วยข้อสอบที่มีค่าเฉลี่ยความยากอยู่ระดับปานกลาง ผู้สอบทุกคนต้องตอบข้อสอบเหมือนกัน เพื่อกำหนดทิศทางของผู้สอบแต่ละคนว่าจะต้องทำการทดสอบในขั้นตอนที่สองด้วยแบบสอบที่มีความยากระดับใดหลังจากทราบทิศทางหรือระดับความสามารถเบื้องต้นของผู้สอบแล้วก็ทำการทดสอบในขั้นตอนที่สอง ด้วยแบบสอบวัดผล (Measurement test) ซึ่งจะแบ่งเป็นแบบสอบชุดย่อย ๆ หลายชุดตามระดับความยากโดยปกติจะมีชุดละประมาณ 20 – 30 ข้อ ผู้สอบแต่ละคนจะได้รับแบบสอบชุดที่สองไม่เหมือนกัน ผู้สอบที่ได้รับการประเมินจากการทดสอบในขั้นตอนแรกว่ามีความสามารถสูงก็จะได้รับแบบสอบชุดที่สองที่ยาก ผู้สอบที่ความสามารถปานกลางจะได้รับแบบสอบชุดที่สองที่มีระดับความยากปานกลางและผู้สอบที่มี

ความสามารถต่ำก็จะได้รับแบบสอบชุดที่สองที่ง่ายโครงสร้างของการทดสอบแบบปรับเหมาะกับความสามารถของผู้สอบโดยใช้ยุทธวิธีสองขั้นตอน

1.2.2 ยุทธวิธีหลายขั้นตอน

การทดสอบแบบหลายขั้นตอน เป็นการทดสอบมากกว่า 2 ขั้นตอน โดยมีการจัดโครงสร้างของขั้นตอนการทดสอบ มีการจัดเรียงข้อสอบตามความยาก ในแต่ละขั้นตอนของการทดสอบมีวิธีดำเนินการแตกต่างกัน หลักการโดยทั่วไปเริ่มต้นจากข้อสอบที่มีความยากปานกลางถ้าตอบถูกข้อต่อไปจะยากขึ้น แต่ถ้าตอบผิดข้อต่อไปจะง่ายลง การทดสอบจะดำเนินเช่นนี้ไปเรื่อยและจะยุติตามเกณฑ์ที่กำหนดไว้ การทดสอบแบบหลายขั้นตอนนี้พอจะจำแนกเป็นแบบต่าง ๆ กันได้ดังนี้ (ศิริชัย กาญจนวาสี, 2538)

1. แบบทางแยกคงที่ เป็นการทดสอบที่กำหนดทางแยกของการเลือกข้อสอบไว้คงที่ตายตัว มีรูปแบบของการดำเนินการทดสอบแบบทางแยกคงที่นี้หลายรูปแบบดังเช่น 1) รูปแบบปิรามิด ได้แก่ รูปแบบปิรามิดขนาดขั้นคงที่ รูปแบบปิรามิดขนาดขั้นแปรผัน รูปแบบปิรามิดข้างตัด รูปแบบปิรามิดแบบมีหลายข้อในแต่ละขั้น และรูปแบบปิรามิดชนิดให้น้ำหนักตัวเลือกของข้อสอบเพื่อแยกทาง

2. รูปแบบเฟล็กซิเลเวล การทดสอบแบบนี้เป็นแบบที่ลอร์ดคิดขึ้นมาในครั้งแรกเพื่อใช้กับการทดสอบที่ใช้กระดาษกับดินสอ (Paper and pencil test) ซึ่งประกอบด้วยชุดของข้อสอบจำนวนหนึ่ง ข้อสอบแต่ละข้อมีช่วงห่างของค่าความยากเท่ากัน โดยทำการเรียงจากข้อที่ง่ายที่สุดไปยังข้อที่ยากที่สุด การทดสอบเริ่มต้นด้วยการให้ผู้ตอบทำข้อสอบข้อที่มีความยากปานกลาง ถ้าตอบถูกก็จะไปตอบข้อที่ยากขึ้น แต่ถ้าตอบผิดข้อต่อไปก็จะง่ายลง

3. รูปแบบปรับระดับชั้น แบบสอบปรับระดับแบบแบ่งชั้นพัฒนามาจากแบบสอบปรับระดับแบบเฟล็กซิเลเวล และมีลักษณะคล้ายกับลักษณะการใช้แบบสอบวัดเชาว์ปัญญาของ Binet คือ มีการดำเนินการทดสอบโดยใช้ข้อสอบจากคลังข้อสอบที่มีการจัดแบ่งความยากของข้อสอบเป็นระดับชั้น (Stratified) แต่ละระดับชั้นประกอบด้วยชุดของข้อสอบที่มีค่าความยากอยู่ในช่วงที่กำหนดเดียวกัน เช่น ค่าความยากอยู่ระหว่าง .20 ถึง .30, .31 ถึง .40 เป็นต้น การใช้แบบสอบปรับระดับชั้นนี้เริ่มต้นที่ข้อสอบระดับความยากใดก็ได้ ถ้าทำข้อนั้นถูก ข้อต่อไปจะเป็นข้อที่มีอำนาจจำแนกสูงสุดที่มีความยากถัดขึ้น แต่ถ้าทำข้อนั้นผิด ข้อต่อไปจะเป็นข้อที่มีอำนาจจำแนกสูงสุดที่มีความง่ายถัดลงไป การทดสอบจะดำเนินเช่นนี้ไปเรื่อยจนถึงความยากระดับเพดาน (Ceiling stratum) ซึ่งผู้สอบไม่สามารถทำข้อสอบชุดนั้นได้เลยสักข้อ หรือทำคะแนนได้ไม่เกินคะแนนที่ได้จากการเดา (Chance score)

4.รูปแบบแยกทางแปรผัน เป็นรูปแบบการตอบข้อสอบหลายขั้นตอนที่ไม่ได้กำหนดข้อสอบและเส้นทางในการตอบข้อสอบไว้ล่วงหน้าว่าถ้าผู้สอบตอบถูกจะต้องไปทำข้อสอบข้อใดหรือถ้าตอบผิดจะต้องไปทำข้อใดต่อไป ในแบบทางแยกแปรผันนี้จะไม่มีขนาดของขั้น แต่จะดำเนินการทดสอบจากกลุ่มข้อสอบที่คำนวณระดับความยากและค่าอำนาจจำแนกของข้อสอบไว้แล้ว กฎการคัดเลือกข้อสอบถัดไปที่ผู้สอบจะได้รับให้เหมาะสมกับความสามารถของผู้สอบนั้น จะใช้โมเดลทางคณิตศาสตร์เข้ามาช่วย ได้แก่ การใช้กลวิธีของเบส์และกลวิธีความเป็นไปได้สูงสุด

1.3 องค์ประกอบของระบบการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์

ระบบการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่ใช้ในโปรแกรมการประเมินทางการศึกษาขนาดใหญ่ที่กล่าวมาแล้ว ประกอบด้วย 4 องค์ประกอบสำคัญ คือ 1) คลังข้อสอบ (Item Pool หรือ Item Bank) 2) วิธีการคัดเลือกข้อสอบ (Item Selection Procedures) 3) วิธีการประมาณค่าความสามารถของผู้สอบ (Ability Estimation) และ 4) เกณฑ์ยุติการทดสอบ (Stopping Rules หรือ Termination criterion) (ศิริชัย กาญจนวาสี, 2550; Reckase, 1989 cited in Lee & Dodd, 2011) นอกจากนี้ Boyd (2003) เสนอว่า ในทางปฏิบัติการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ควรพิจารณาในเรื่องการควบคุมการใช้ข้อสอบซ้ำ (Exposure Control) และการสร้างความสมดุลของเนื้อหา (Content Balancing) ร่วมด้วย เพื่อให้มีความครอบคลุมรายละเอียดของการควบคุมการใช้ข้อสอบซ้ำ และ การสร้างความสมดุลของเนื้อหาจะได้รับการนำเสนอในตอนต่อไป

1.3.1 คลังข้อสอบ (Item Pool)

คลังข้อสอบ (Item Pool) คือ แหล่งเก็บรวบรวมข้อสอบหรือข้อคำถามขนาดใหญ่ ในการทดสอบแบบดั้งเดิมแบบสอบจะสร้างขึ้นจากคลังข้อสอบโดยข้อสอบจะถูกนำมารวมกันเป็นแบบสอบแล้วนำไปใช้ผู้สอบโดยผู้สอบทุกคนจะได้รับแบบสอบที่เหมือนกัน ซึ่งต่างจากการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ผู้สอบแต่ละคนจะได้รับแบบสอบที่แตกต่างกันขึ้นอยู่กับระดับความสามารถภายในของผู้สอบ ดังนั้นคุณภาพของคลังข้อสอบจึงมีอิทธิพลอย่างมากต่อประสิทธิภาพของขั้นตอนวิธี (Algorithm) การปรับเหมาะในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (Flaugher, 2000) ในการพิจารณาคุณภาพของคลังข้อสอบนั้นควรพิจารณาทั้งในเรื่องของขนาดของคลังข้อสอบความครอบคลุมทั้งในด้านความกว้างและความลึกของเนื้อหา และคุณลักษณะทางจิตมิติ (Parshall, Spray, Kalohn, & Davey, 2002) ข้อสอบที่บรรจุอยู่ในคลังข้อสอบจะต้องมีจำนวนมากเพียงพอที่จะทำให้ค่าพารามิเตอร์ของข้อสอบทั้ง พารามิเตอร์ความยากและพารามิเตอร์อำนาจจำแนกมีการกระจายอยู่ในทุกระดับความสามารถของผู้สอบและทุกระดับของเนื้อหาทั้งในด้านความกว้างและความลึก และค่าอำนาจจำแนกควรจะมีค่าสูงในเกณฑ์ที่ยอมรับได้ (ควรมีค่าสูงกว่า 0.5) เพื่อให้เกิด

ประสิทธิภาพสูงสุดในการทดสอบ และค่าของพารามิเตอร์การเดาควรเข้าใกล้หรือเท่ากับศูนย์ ดังนั้น การสร้างคลังข้อสอบสำหรับการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์จะใช้ข้อสอบจำนวนมากกว่าการสร้างคลังข้อสอบสำสำหรับการทดสอบแบบดั้งเดิม

การพัฒนาคลังข้อสอบสำหรับการทดสอบแบบปรับเหมาะนั้นคล้ายคลึงกันกับคลังข้อสอบสำหรับการทดสอบแบบดั้งเดิม ข้อสอบจำเป็นต้องเขียนตามตารางการกำหนดเนื้อหาและควรมีการตรวจสอบทั้งในเรื่องของคุณภาพของเนื้อหาและความไวของแบบสอบ (Test Sensitivity) และการทดลองใช้เพื่อศึกษาคุณลักษณะทางจิตมิติ (Flaugher, 2000) วิธีการที่ใช้สำหรับการตรวจสอบคุณภาพของข้อสอบสามารถใช้วิธีการทางสถิติของทฤษฎีการทดสอบแบบดั้งเดิม เช่น สัดส่วนของผู้ที่ตอบถูก (Proportion Correct) หรือ สหสัมพันธ์แบบไบซีเรียล (Biserial Correlation) ร่วมกับทฤษฎีการทดสอบแนวใหม่ เช่น พารามิเตอร์ของข้อสอบ (Item Parameters) และ ค่าสารสนเทศของข้อสอบ (Item Information) (Wainer, 1989) และสิ่งที่สำคัญอีกประการหนึ่ง คือ คุณลักษณะของข้อสอบ (Item Characteristic) เพื่อตรวจสอบว่าข้อสอบแต่ละข้อเหมาะสมกับโมเดลการวัด

ขั้นตอนการสร้างคลังข้อสอบสำหรับใช้ในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์มีขั้นตอนดังนี้ 1) สร้างข้อสอบใหม่ให้มีจำนวนเพียงพอในแต่ละเนื้อหาเน้นที่ครอบคลุมเนื้อหาหลักที่ครอบคลุมทั้งในด้านความกว้างและความลึกของเนื้อหา 2) ตรวจทานคุณภาพของข้อสอบเพื่อให้ข้อสอบมีคุณภาพสูงสุดเท่าที่เป็นไปได้ 3) ทำการทดสอบเบื้องต้นเพื่อตรวจสอบข้อบกพร่องของข้อสอบใหม่ที่สร้างขึ้น 4) ทำการเลือกชุดย่อยของข้อสอบใหม่ ตอบข้อสอบ และวิเคราะห์ผลสอบตามทฤษฎีการทดสอบแบบดั้งเดิมและทฤษฎีการตอบสนองข้อสอบ ทำการปรับปรุงข้อสอบใหม่เพื่อลดจำนวนข้อสอบที่ไม่เหมาะสม 5) เปรียบเทียบความสมดุลของเนื้อหาของคลังข้อสอบด้วยแบบทดสอบตามทฤษฎีการทดสอบแบบดั้งเดิม เพื่อสรุปอ้างอิงไปยังความสมดุลของเนื้อหาของคลังข้อสอบในการทดสอบแบบปรับเหมาะ ประเมินระบบการทดสอบด้วยการศึกษาในสถานการณ์การจำลองเพื่อศึกษาพฤติกรรมการตอบข้อสอบของผู้สอบในแต่ละระดับความสามารถ คะแนนจากการทดสอบของผู้สอบแต่ละคนต้องสามารถเปรียบเทียบกันได้ และ 6) นำข้อสอบไปรวมเป็นแบบสอบเพื่อใช้ในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (Flaugher, 2000)

เนื่องจาก ทฤษฎีการตอบสนองข้อสอบ (IRT) มีข้อตกลงเบื้องต้น เกี่ยวกับ ความเป็นเอกมิติ (Unidimension) ข้อสอบแต่ละข้อในแบบสอบนั้นควรวัดความสามารถหรือคุณลักษณะเดียวกัน (ศิริชัย กาญจนวาสี, 2550) แต่ในทางปฏิบัติแบบสอบมักประกอบไปด้วยขอบเขตเนื้อหาหลายด้านหรือหลายมิติ เพื่อจัดการกับปัญหานี้อาจใช้วิธีการแบ่งคลังข้อสอบใหญ่ เป็นคลังข้อสอบย่อย โดยแต่ละคลังแทนขอบเขตเนื้อหาที่แตกต่างกัน คะแนนแต่ละส่วนมาจากแต่ละเนื้อหา (Flaugher, 2000)

และการสร้างคลังข้อสอบมากกว่าหนึ่งคลังข้อสอบหรือแบ่งคลังข้อสอบใหญ่เป็นคลังข้อสอบย่อย สามารถสุ่มเลือกคลังข้อสอบและหมุนเวียน การใช้ชุดแบบทดสอบไปตามคลังข้อสอบต่างๆ นอกจากนี้สิ่งที่ควรพิจารณาเพิ่มเติมในการการสร้างคลังข้อสอบจำนวนมาก คือ ต้องทำการศึกษาเปรียบเทียบระหว่างคลังข้อสอบในประเด็นต่อไปนี้ 1) ขนาดของคลังข้อสอบ 2) จำนวนคลังข้อสอบเมื่อใช้พร้อมๆ กัน 3) การทับซ้อนของคลังข้อสอบ 4) วิธีเข้าสู่แต่ละคลังข้อสอบ 5) การหมุนเวียนใช้คลังข้อสอบและเกณฑ์การออกจากคลังข้อสอบ 6) กฎการนำข้อสอบกลับมาใช้ใหม่ 7) การทดสอบข้อสอบก่อนนำมาใช้ วิธีการคำนวณข้อสอบ และวิธีการควบคุมช่วงความสามารถบนเส้นต่อเนื่อง และ 8) วิธีการเลือกข้อสอบเพื่อให้ครอบคลุมแต่ละเนื้อหาวิชา ควบคุมอัตราการใช้ข้อสอบซ้ำ อัตราการทับซ้อนของข้อสอบและแบบสอบ (Wang & Kolen, 2001) ก่อนนำคลังข้อสอบนั้นไปใช้จริง

สำหรับวิธีการรวมข้อสอบเป็นแบบสอบมีการศึกษาพัฒนากันอย่างต่อเนื่อง วิธีการส่วนใหญ่จะใช้เทคนิคการโปรแกรมเชิงเส้นแบบไบนารี (0-1 Linear Programming) หรือเทคนิคการโปรแกรมแบบผสม (Mixed integer Programming) เพื่อสร้างแบบสอบคู่ขนานเพื่อออกแบบสอบที่เหมาะที่สุด (Optimal Test Design) เพื่อแก้ปัญหาในการรวมแบบสอบ และเพื่อสร้างแบบสอบอัตโนมัติ (Ariel, Veldkamp, & van der Linden, 2004; Boekkooi-Timminga, 1990; Chang & Shiu, 2011; van der Linden, 1998b; W. J. van der Linden, 2005a, 2005b; van der Linden & Adema, 1998; van der Linden & Diao, 2011; van der Linden & Reese, 2001; van der Linden & Veldkamp, 2004) รายละเอียดของการโปรแกรมเชิงเส้นแบบไบนารีจะนำเสนอต่อไป

1.3.2 วิธีการคัดเลือกข้อสอบ (Item Selection Procedures)

วิธีการคัดเลือกข้อสอบ (Item Selection Rule) มีวิธีการเลือกข้อสอบจากคลังข้อสอบที่มีประสิทธิภาพ 2 วิธีการ คือ วิธีการคัดเลือกข้อสอบที่มีค่าสารสนเทศสูงสุด (Maximum Information Item Selection) และวิธีการคัดเลือกข้อสอบแบบเบส์ (Bayesian Item Selection) ซึ่งวิธีแรกจะให้ค่าสารสนเทศสูงสุด ส่วนวิธีหลังจะให้ค่าความแปรปรวนที่คาดหวังของการประมาณค่าความสามารถต่ำสุด ทั้ง 2 วิธีจะคล้ายคลึงกันเนื่องจากว่าค่าสารสนเทศสูงสุดและความแปรปรวนที่คาดหวังของการประมาณค่ามีความสัมพันธ์กัน (ศิริชัย กาญจนวาสี, 2550)

1) วิธีการคัดเลือกข้อสอบที่มีค่าสารสนเทศสูงสุด (Maximum Information Item selection) จะคัดเลือกข้อสอบที่ให้สารสนเทศสูงสุด ตามค่าประมาณความสามารถของผู้สอบครั้งหลังสุด เมื่อสารสนเทศแบบทดสอบสูงขึ้น ความคลาดเคลื่อนมาตรฐานของการประมาณค่าความสามารถจะลดลง ทำให้ค่าประมาณความสามารถเข้าใกล้ค่าความสามารถจริงของผู้สอบมากที่สุด (ศิริชัย กาญจนวาสี, 2550) วิธีการนี้ได้พัฒนาขึ้นโดยใช้วิธีการประมาณค่าความเป็นไปได้สูงสุด

ตามทฤษฎีการตอบสนองข้อสอบ ซึ่งมีวิธีการประมาณค่าความสามารถของผู้สอบที่คล้ายคลึงกับกระบวนการตามทฤษฎีบทของเบส์ แม้ว่าเหตุผลทางคณิตศาสตร์ต่างกัน หลังจากที่ผู้สอบตอบข้อสอบเพียง 1 ข้อก็จะนำผลการทดสอบไปแก้สมการความเป็นไปได้สูงสุด และจะได้ค่าประมาณค่าความสามารถและความคลาดเคลื่อนมาตรฐาน ข้อสอบข้อถัดไปที่เลือกมาใช้ทดสอบจะเป็นข้อสอบจากกลุ่มข้อสอบและเป็นข้อสอบที่มีระดับความยากเหมาะสมกับความสามารถของผู้สอบมากที่สุด เมื่อผู้สอบตอบข้อสอบข้อนั้นแล้วก็จะมีการประเมินค่าความสามารถของผู้สอบทันทีจากข้อมูลการตอบทั้งหมดของผู้สอบ ซึ่งรวมถึงข้อสอบข้อสุดท้ายที่ผู้สอบตอบ จากนั้นก็ประมาณค่าความสามารถและความคลาดเคลื่อนมาตรฐานใหม่โดยใช้ฟังก์ชันความเป็นไปได้จนกว่าการทดสอบจะสิ้นสุดตามเกณฑ์ที่กำหนดไว้ ซึ่งความสัมพันธ์ระหว่างการตอบข้อสอบและฟังก์ชันความเป็นไปได้มีรายละเอียดดังนี้ (van der Linden & Pashley, 2002)

$$L\left(\theta|u_{i_1}\dots u_{i_{k-1}}\right) \equiv \prod_{j=1}^{k-1} \frac{\left\{exp\left[a_{i_j}\left(\theta - b_{i_j}\right)\right]\right\}^{u_{i_j}}}{1 + exp\left[a_{i_j}\left(\theta - b_{i_j}\right)\right]}$$

เมื่อ

$L\left(\theta|U_{i_1}\right)$ คือ ฟังก์ชันความเป็นไปได้ที่แสดงความสัมพันธ์ระหว่างความสามารถของผู้สอบ θ กับผลการตอบข้อสอบตั้งแต่ข้อแรกจนถึงข้อที่ $k-1$ ในแบบสอบปรับเหมาะ

i คือ ข้อสอบในคลังข้อสอบ $i = 1, \dots, I$

k คือ ช่วงพิสัยข้อสอบในแบบสอบปรับเหมาะ $k = 1, \dots, K$

i_k คือ ดัชนีที่บอกว่าข้อสอบในคลังข้อสอบถูกนำไปใช้กับผู้สอบเป็นจำนวน k ข้อในแบบสอบ

$u_{i_{k-1}}$ คือ ผลการตอบข้อสอบในแบบสอบปรับเหมาะข้อที่ $k-1$ ซึ่งเป็นผลการตอบก่อนหน้าโดย $u_{i_{k-1}} = U_{i_1} = u_{i_1}, \dots, U_{i_{k-1}}$

θ คือ ค่าความสามารถของผู้สอบ

a_{i_j} คือ ค่าพารามิเตอร์อำนาจจำแนกของข้อสอบในแบบสอบปรับเหมาะข้อที่ j

b_{i_j} คือ ค่าพารามิเตอร์ความยากของข้อสอบในแบบสอบปรับเหมาะข้อที่ j

2. วิธีการคัดเลือกข้อสอบแบบเบส์ (Bayesian Item selection) วิธีการนี้ประยุกต์ทฤษฎีของเบส์มาใช้ในกระบวนการประมาณค่าพารามิเตอร์ตอบข้อสอบและอาศัยเครื่องคอมพิวเตอร์ในการทดสอบ จึงมักเรียกว่าการทดสอบแบบปรับเหมาะกับความสามารถของผู้สอบ โดยทำการประมาณค่าความสามารถของผู้สอบและความคลาดเคลื่อนมาตรฐานในการประมาณค่าจากข้อมูลต่างๆ ที่มีอยู่เกี่ยวกับผู้สอบและข้อสอบ จากนั้นคัดเลือกข้อสอบจากกลุ่มข้อสอบที่ได้คำนวณค่าพารามิเตอร์ของข้อสอบไว้แล้ว กลุ่มข้อสอบทุกข้อที่ยังไม่ได้นำมาให้ผู้สอบคนนั้นๆ จะถือว่าเป็นข้อสอบที่มีโอกาสนำมาใช้สอบได้ กระบวนการนี้แสดงให้เห็นว่าข้อสอบข้อใดๆ ในกลุ่มข้อสอบที่นำมาใช้ในการทดสอบกับผู้สอบคนใดๆ จะเป็นข้อสอบที่มีระดับความยากใกล้เคียงกับระดับความสามารถของผู้สอบมากที่สุด หลังจากดำเนินการสอบโดยใช้ข้อสอบที่คัดเลือกไว้ก็จะประมาณความสามารถของผู้สอบไว้ก่อน ผลการตอบถูกหรือการตอบผิดจะนำมารวมกันเพื่อคำนวณโดยใช้ทฤษฎีบทของเบส์ แล้วหาค่าความสามารถในภายหลัง การประมาณภายหลังนี้เป็นการประมาณค่าที่ได้รับการปรับจากค่าที่ได้คำนวณไว้แต่เดิม กระบวนการนี้จะสิ้นสุดลงเมื่อความคลาดเคลื่อนในการประมาณค่าความสามารถมีค่าน้อยกว่าที่กำหนดไว้ (ศิริชัย กาญจนวาสี, 2550)

$$f(\theta \mid U) = KL(\theta \mid U) f(\theta)$$

เมื่อ

$f(\theta \mid U)$	คือ	การแจกแจงภายหลัง (Posterior Distribution) ของ θ
$L(U \mid \theta)$	คือ	ฟังก์ชันความเป็นไปได้ (Likelihood Function) ของเวคเตอร์ U (ผลการตอบข้อสอบ)
$f(\theta)$	คือ	การแจกแจงก่อนหน้า (Prior Distribution) ของ θ
K	คือ	ค่าคงที่ (Constant)

วิธีการคัดเลือกข้อสอบที่ใช้กันมากที่สุดควบคู่ไปกับการประมาณค่าความสามารถของผู้สอบ θ ด้วยวิธี Bayesian คือ การเลือกข้อสอบข้อที่ยังไม่ได้นำมาใช้ในการทดสอบนั้นที่จะให้ค่าความแปรปรวนของค่าความสามารถที่คาดหวังมีค่าต่ำสุด (Smallest Posterior Variance) นั้นเป็นการคัดเลือกข้อสอบที่คาดว่าจะลดความไม่แน่นอนลงได้มากที่สุดในความสามารถของผู้สอบ θ

การคัดเลือกข้อสอบทั้งวิธีสารสนเทศสูงสุดและวิธีการเลือกข้อสอบวิธีเบเซี่ยน สามารถลดความคลาดเคลื่อนกำลังสองเฉลี่ยได้ ทั้ง 2 วิธี และให้ผลการคัดเลือกข้อสอบใกล้เคียงกันเมื่อความยาวของแบบสอบมากเพียงพอ การทดสอบแบบปรับเหมาะสามารถใช้วิธีสารสนเทศสูงสุดและวิธีเบเซี่ยนในการทดสอบเดียวกัน พบในผลงานวิจัยของ van der Linden ในปี (1998a) นอกจากนี้ พบว่ามี

การศึกษาเปรียบเทียบวิธีการเลือกข้อสอบยังสามารถพบได้ในงานวิจัยต่างๆ เช่น วิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (Thissen & Mislevy, 2000) การเปรียบเทียบเกณฑ์การคัดเลือกข้อสอบในช่วงต้นการทดสอบ (Chen, Ankenman, & Chang, 2000) การเปรียบเทียบวิธีการคัดเลือกข้อสอบและการประมาณค่าความสามารถของผู้สอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ในช่วงต้น ช่วงกลาง และช่วงท้ายของการทดสอบ (van der Linden & Pashley, 2002) และการคัดเลือกข้อสอบในการทดสอบแบบหลายขั้น (Multistage Testing) (Hendrickson, 2007)

ปัจจุบันการคัดเลือกข้อสอบวิธีสารสนเทศสูงสุดประยุกต์ใช้กับการเลือกข้อสอบวิธีอื่นๆ เพื่อเพิ่มประสิทธิภาพในการเลือกข้อสอบ ได้แก่ การเลือกข้อสอบโดยใช้ค่าสารสนเทศโดยรวมเฉลี่ย (Kullback-Leibler Information: KLI) หรือเรียกอีกอย่างหนึ่งว่า Global Information) (Chang & Ying, 1996) การเลือกข้อสอบด้วยวิธีเกณฑ์ข้างเคียงใกล้ที่สุด (Nearest-Neighbors Criterion) (Cheng & Liou, 2003) วิธีเกณฑ์ข้างเคียงใกล้ที่สุดเป็นการรวมวิธีสารสนเทศสูงสุดของฟิชเชอร์และวิธีค่าความยากข้อสอบเหมาะสม (Optimal Item Difficulty) เข้าไว้ด้วยกัน นอกจากนี้ การศึกษาวิธีเลือกข้อสอบจากสารสนเทศเฉลี่ยสูงสุด นำไปใช้ผ่อนคลายค่าประมาณความสามารถระหว่างการเลือกข้อสอบ เช่น เกณฑ์สารสนเทศตามน้ำหนักทั่วไป (General Weight Information Criterion) เกณฑ์การพิจารณาสารสนเทศแบบช่วง (Interval Information Criterion)

1.3.3 วิธีการประมาณค่าความสามารถของผู้สอบ (Ability Estimation)

การทดสอบแบบปรับเหมาะ ผลการตอบข้อสอบและพารามิเตอร์ข้อสอบจะถูกนำไปคำนวณหาค่าประมาณความสามารถทุกครั้งหลังจากการตอบเพื่อนำค่าประมาณความสามารถของผู้สอบไปใช้ในการคัดเลือกข้อสอบข้อถัดไป วิธีการประมาณค่าความสามารถแบ่งเป็น 2 วิธี ใหญ่ๆ (ศิริชัย กาญจนวาสี, 2550) คือ

1. วิธีประมาณค่าความเป็นไปได้สูงสุด (Maximum Likelihood Estimation: MLE) วิธีประมาณค่าความเป็นไปได้สูงสุด สามารถให้พารามิเตอร์ข้อสอบค่อนข้างเป็นรูปสนิทกับรูปแบบตามทฤษฎีการตอบสนองข้อสอบ และให้ค่าประมาณความสามารถใกล้เคียงค่าความสามารถจริงของผู้สอบมากที่สุด วิธีประมาณค่าความเป็นไปได้สูงสุดมีความมั่งคงและประสิทธิภาพ (Hambleton & Swaminathan, 1985) แต่สามารถพบข้อจำกัดบางประการ (Meijer & Nering, 1999) ได้แก่ โค้งของฟังก์ชันความเป็นไปได้ (Likelihood) จะไม่เป็นจริงเมื่อรูปแบบคะแนนเป็นการตอบถูกทั้งหมดหรือตอบผิดทั้งหมด และค่าประมาณความสามารถจะมากเกินจริงเมื่อค่าความสามารถจริงเป็นค่าทางบวก และค่าประมาณความสามารถต่ำเกินจริงเมื่อค่าความสามารถจริงเป็นค่าทางลบ

วิธีการประมาณค่าความเป็นไปได้สูงสุด (Maximum Likelihood) ซึ่งมีหลายวิธีแต่ที่นิยมใช้มาก คือ วิธีความเป็นไปได้สูงสุดแบบมีเงื่อนไข (Conditional Maximum Likelihood) การประมาณค่าวิธีนี้มีข้อจำกัดกรณีที่ผู้สอบตอบข้อสอบถูกหมดหรือผิดหมดจะไม่สามารถประมาณค่าได้ ขั้นตอนวิธีการประมาณค่ามีดังนี้ (Hambleton & Swaminathan, 1985)

ขั้นที่ 1 ประมาณค่าความสามารถเริ่มต้น $(\theta_m; m = 0)$ โดยใช้สูตร ดังนี้

$$\theta_0 = \ln [r_a/(K - r_a)$$

เมื่อ $r_a = \sum a_i U_i r_a$

u_i = 1 เมื่อตอบถูก

u_i = 0 เมื่อตอบผิด

a_i คือ อำนาจจำแนกของข้อสอบข้อที่ i

k คือ จำนวนข้อทดสอบทั้งหมดที่ผู้สอบตอบ

ขั้นที่ 2 หาค่า $P_i(\theta_m)$ และ $Q_i(\theta_m)$ โดยใช้สูตร ดังนี้

$$P_i(\theta_m) = c_i + (1 - c_i)\frac{e^{Da_i(\theta_m - b_i)}}{Da_i(\theta_m - b_i)}$$

$$Q_i(\theta_m) = 1 - P_i(\theta_m) = \frac{1 - c_i}{1 + e^{Da_i(\theta_m - b_i)}}$$

เมื่อ

θ_m คือ ความสามารถของผู้สอบที่ประมาณค่าได้ในครั้งที่ m

a_1 คือ ค่าอำนาจจำแนกของข้อทดสอบข้อที่ i

b_1 คือ ค่าความยากของข้อสอบข้อที่ i

c_1 คือ ค่าการเดาของข้อสอบข้อที่ t

D คือ ค่าคงที่ ซึ่งมีค่าเท่ากับ 1.7

e คือ ค่าคงที่ ซึ่งมีค่าเท่ากับ 2.7182

ขั้นที่ 3 หาค่าปรับแก้ (h_m) โดยใช้สูตร

$$h_m = \frac{D\left[r_m - \sum Pi(\theta_m)\right]}{-D^2 \sum Pi(\theta_m)Qi(\theta_m)}$$

ขั้นที่ 4 ประมาณค่าความสามารถของผู้สอบใหม่ $(\theta_m + 1)$ โดยใช้สูตร $\theta_m + 1 = \theta_m - h_m$

ขั้นที่ 5 ทำการคำนวณซ้ำในขั้นที่ 2, 3 และ 4 จนกระทั่ง h_m เข้าใกล้ศูนย์ ($h_m < 0.001$)

2. วิธีประมาณค่าเบเซี่ยน (Bayesian Estimation) วิธีประมาณค่าของเบส์สำหรับใช้ในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ ได้แก่ การประมาณค่าภายหลังคาดหวัง (Expected A Posteriori: EAP) และ รูปแบบเบส์ (Bayes Model: BM) หรือการประมาณค่าภายหลังสูงสุด (Maximum A Posteriori: MAP) หลักการวิธีประมาณค่าของเบส์ใช้สารสนเทศจากรูปแบบการตอบข้อสอบหรือฟังก์ชันความเป็นไปได้ (Likelihood Function) และใช้สารสนเทศเดิมของผู้สอบ (Prior Information) นำไปคำนวณหาค่าประมาณความสามารถของผู้สอบ ถ้าพบว่าการประมาณค่าความเป็นไปได้ และค่าเฉลี่ยการแจกแจงค่าความสามารถเริ่มต้นมีความต่างกันมาก ค่าประมาณความสามารถจะถอยกลับไปยังค่าเฉลี่ยเบื้องต้น (Hambleton & Swaminathan, 1985)

วิธีการประมาณค่าความสามารถของผู้สอบตามวิธีของเบส์ (Bayesian) ได้ถูกนำไปประยุกต์เป็นวิธีต่างๆ อีกหลายวิธี แต่วิธีที่ได้รับความนิยมมากที่สุดคือ วิธีของเบส์ที่ปรับใหม่ ซึ่งเสนอโดย Owen (1975) โดยมี รายละเอียดและวิธีการประมาณค่าดังนี้

ในกรณีที่ตอบถูก

$$\theta_{m+1} = \theta_m + (1-c)\left[\frac{\sigma_m^2}{\sqrt{\frac{1}{a^2}+\sigma_m^2}}\right]\left[\frac{\Theta(D)}{c+(1-c)A(-D)}\right]$$

$$\sigma_{m+1}^2 = \sigma_m^2\left[1-\left\{\frac{1-c}{1+\frac{1}{a^2\sigma_m^2}}\right\}\left\{\frac{\Theta(D)}{B}\right\}\left\{\frac{(1-c)\Theta(D)}{B-D}\right\}\right]$$

เมื่อ $B = c(1-c)\times A(-D)$

ในกรณีที่ตอบผิด

$$\theta_{m+1} = \theta_m - \left[\frac{\sigma_m^2}{\sqrt{\frac{1}{a^2}+\sigma_m^2}}\right]\left[\frac{\Theta(D)}{A(D)}\right]$$

$$\sigma_{m+1}^2 = \sigma_m^2\left[1-\left[\frac{\Theta(D)}{1+\frac{1}{a^2\sigma_m^2}}\right]\left[\frac{\Theta(D)}{A(D)}+D\right]/[\Theta(D)]\right]$$

เมื่อ θ_m แทน ความสามารถของผู้สอบที่ประมาณค่าได้ก่อนตอบข้อสอบข้อที่ m+1 ซึ่งตามปกติแล้ว ถ้าไม่ทราบค่าความสามารถเบื้องต้นของผู้สอบก็มักกำหนดให้เท่ากับ 0

σ_m^2 แทน ความแปรปรวนในการประมาณค่าความสามารถของผู้สอบก่อนตอบ ข้อที่ m+1 ถ้าไม่ทราบค่าความแปรปรวนดังกล่าวมาก่อนก็มักกำหนดให้เท่ากับ 1

θ_{m+1} แทน ค่าความสามารถของผู้สอบโดยประมาณหลังจากที่ตอบข้อที่ m+1

σ_{m+1}^2 แทน ค่าความแปรปรวนในการประมาณค่าความสามารถของผู้สอบ เมื่อตอบข้อสอบข้อที่ m+1

a แทน พารามิเตอร์ค่าอำนาจจำแนกข้อสอบข้อที่ m+1

b แทน พารามิเตอร์ค่าความยากของข้อสอบข้อที่ m+1

c แทน พารามิเตอร์ระดับโอกาสการเดาข้อสอบได้ถูกข้อที่ m+1

D แทน จุดบนแกน X

$\Theta(D)$ แทน ค่าออร์ดิเนท (Ordinate) ของโค้งปกติที่จุด D

$A(D)$ แทน พื้นที่ใต้โค้งปกติจากค่า D จนถึงจุด D

ในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ การประมาณค่าความสามารถของผู้สอบเพื่อเริ่มการเลือกข้อสอบ ถ้าใช้วิธีการประมาณค่าความเป็นไปได้สูงสุด (Maximum Likelihood Estimation: MLE) จะให้การประมาณค่าความสามารถไม่แน่นอนถ้าตอบข้อสอบถูกทั้งหมดหรือผิดทั้งหมด (ศิริชัย กาญจนวาสี, 2550) จึงไม่ใช้วิธีนี้ในช่วงเริ่มต้นการทดสอบแต่จะเลือกใช้วิธีอื่นที่ให้ค่าประมาณความสามารถที่แน่นอนกว่าการใช้วิธีของเบส์ เช่น การประมาณค่าภายหลังคาดหวัง จำเป็นต้องใช้ข้อมูลการตอบข้อสอบที่มีอยู่ก่อนหน้านี้หรือข้อมูลพื้นฐานของผู้สอบ นำไปอ้างเป็นค่าประมาณความสามารถเริ่มต้น ซึ่งแต่ละวิธีมีข้อจำกัดบางประการทั้งสิ้น

1.3.4 เกณฑ์ยุติการทดสอบ (Stopping Rule)

ลักษณะสำคัญของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ คือ การทดสอบจะดำเนินไปเรื่อยจนกระทั่งถึงเกณฑ์ที่ระบุไว้ให้ยุติการทดสอบ การดำเนินการสอบก็จะยุติลง เกณฑ์ที่ใช้กันอยู่ทั่วไปมีอยู่ 2 แบบ (ศิริชัย กาญจนวาสี, 2550) คือ

1. ความยาวคงที่ (Fixed-Length) เป็นการกำหนดจำนวนข้อสอบที่ใช้ในการทดสอบให้คงที่สำหรับทุกคน เช่น กำหนดให้การทดสอบแบบปรับเหมาะใช้ข้อสอบจำนวน 25 ข้อ (n = 25) เมื่อ

ผู้สอบทำข้อสอบได้ครบ 25 ข้อ การทดสอบก็จะยุติลง เกณฑ์นี้ค่อนข้างเป็นประโยชน์ในการศึกษาภายใต้สถานการณ์จำลองแบบมอนติ คาร์โล (Monte Carlo Simulation) เนื่องจากจำนวนข้อสอบเท่ากันทำให้สามารถเปรียบเทียบสารสนเทศของแบบสอบได้โดยตรง ในทางปฏิบัติการกำหนดให้ทุกคนทำข้อสอบจำนวนเท่ากัน อาจมีคุณภาพของการวัดผลได้แตกต่างกัน (ศิริชัย กาญจนวาสี, 2550) เกณฑ์ความยาวคงที่นี้ใช้งานง่าย สามารถทำนายอัตราการใช้ข้อสอบซ้ำได้ แต่ต้องระวังการใช้กฎความยาวคงที่อาจให้ฟังก์ชันสารสนเทศเป็นโค้งลักษณะแบน ดังนั้น กฎความยาวคงที่อาจไม่ให้ความแม่นยำเท่าเทียมกันที่ระดับความสามารถต่างๆ (Thissen & Mislevy, 2000)

2. ความยาวแปรผัน (Variable-Length) เป็นการกำหนดความยาวของข้อสอบแบบไม่คงที่สำหรับผู้สอบแต่ละคน โดยจะผันแปรไปตามความคลาดเคลื่อนมาตรฐานของการประมาณค่าความสามารถ (Standard Error: $SEE(\theta)$) การทดสอบจะดำเนินไปจนกว่าการประมาณค่าความสามารถ (θ) ของผู้สอบมีความคลาดเคลื่อนมาตรฐานลดต่ำลงจนถึงระดับที่ยอมรับได้การทดสอบจึงยุติลง (ศิริชัย กาญจนวาสี, 2550) ดังนั้นการใช้เกณฑ์ยุติการทดสอบจากค่าความคลาดเคลื่อนมาตรฐานจะให้ผลดี คือ มีความแม่นยำในการวัดสูงสำหรับทุกระดับความสามารถของผู้สอบ (Thissen & Mislevy, 2000)

$$SEE(\theta) = \frac{1}{\sqrt{I(\theta)}}$$

เมื่อ $SEE(\theta)$ คือ ความคลาดเคลื่อนมาตรฐานของการประมาณค่า θ

$I(\theta)$ คือ ค่าฟังก์ชันสารสนเทศ หรือ ค่าสารสนเทศที่ได้รับจากแบบสอบสำหรับผู้มีความสามารถ θ

1.3.5 การควบคุมการใช้ข้อสอบซ้ำ (Exposure Control)

การใช้ข้อสอบซ้ำ (Item Exposure) คือ การนำข้อสอบจากคลังข้อสอบไปใช้กับผู้สอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ ผู้สอบแต่ละคนจะได้รับข้อสอบที่ถูกคัดเลือกให้เหมาะสมกับความสามารถของตนเอง ข้อสอบที่นำไปใช้นั้นจะให้สารสนเทศสูงสุด ณ ระดับความสามารถของผู้สอบที่ถูกประมาณค่า เมื่อพิจารณาจากโมเดลการตอบสนองข้อสอบแบบ 3 พารามิเตอร์ $P_i(\theta_m) = c_i + (1 - c_i)\frac{e^{Da_i(\theta_m - b_i)}}{Da_i(\theta_m - b_i)}$ และฟังก์ชันสารสนเทศของข้อสอบ

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)}$$

เมื่อ $I_i(\theta)$ คือ สารสนเทศได้จากข้อสอบข้อที่ i สำหรับผู้สอบความสามารถ θ

$P_i(\theta)$ คือ ความน่าจะเป็นที่ผู้สอบความสามารถ θ จะตอบข้อสอบที่ i ได้ถูกต้อง

$P_i'(\theta)$ คือ อนุพันธ์ของ $P_i(\theta)$ หรือ ความชันของฟังก์ชันการตอบสนองข้อสอบข้อที่ i เมื่อพิจารณาจากตำแหน่งความสามารถ θ

$Q_i(\theta)$ คือ ความน่าจะเป็นที่ผู้สอบความสามารถ θ จะตอบข้อสอบที่ i ผิด โดยที่ $Q_i(\theta) = 1 - P_i(\theta)$

ค่าสารสนเทศของข้อสอบจะมีค่าสูงขึ้นได้นั้น มี 2 กรณี คือ 1) เมื่อค่าพารามิเตอร์อำนาจจำแนกของข้อสอบมีค่ามากขึ้นแล้วค่าสารสนเทศของข้อสอบจะมีค่าเพิ่มขึ้น และ 2) เมื่อค่าความสามารถของผู้สอบ θ เข้าใกล้ค่าพารามิเตอร์ความยากของข้อสอบแล้วค่าสารสนเทศของข้อสอบจะมีค่าสูงขึ้น จากข้อค้นพบดังกล่าวทำให้อนุมานได้ว่า วิธีการคัดเลือกข้อสอบที่มีค่าสารสนเทศสูงสุด (MI) ซึ่งเป็นวิธีที่ได้รับความนิยมมากในการทดสอบแบบปรับเหมาะเนื่องจากให้ค่าสารสนเทศสูงสุดมักคัดเลือกข้อสอบที่มีค่าพารามิเตอร์อำนาจจำแนกสูงๆ จึงทำให้ข้อสอบบางส่วนในคลังข้อสอบถูกนำมาใช้บ่อย หรือ ถูกเปิดเผยกับผู้สอบถี่เกินไป (Overexposure) ขณะที่ข้อสอบส่วนที่เหลือแทบไม่ถูกนำออกมาใช้เลย (Underexposure or Neverexposure) และการใช้ข้อสอบอย่างไม่สมดุลคือเลือกใช้เฉพาะข้อสอบที่มีค่าพารามิเตอร์อำนาจจำแนกสูง แบบสอบที่ผู้สอบแต่ละคนได้รับจะถูกละเลยในเรื่องของความสมดุลของเนื้อหา (Balancing) ซึ่งได้รับกำหนดไว้ในผังแบบสอบ (test blueprint) หรือ ตารางกำหนดคุณลักษณะของข้อสอบ (Table of Content Specification)

ผลของการใช้ข้อสอบซ้ำถี่เกินไปก่อให้เกิดปัญหาในการวัดและประเมิน เช่น กลุ่มของข้อสอบที่มีค่าพารามิเตอร์อำนาจจำแนกสูงๆ เมื่อนำออกมาใช้บ่อยจนรู้จักแพร่หลายในกลุ่มผู้สอบ จะนำไปสู่ปัญหาในเรื่องความปลอดภัยของแบบสอบ (Test Security) ซึ่งเกิดจากการร่วมใช้ข้อสอบ (Item Sharing) ระหว่างผู้สอบ ผู้สอบจะได้รับข้อมูลของแบบสอบจากผู้ที่เคยสอบก่อนหน้าและผู้สามารถสอบตอบคำถามข้อนั้นถูกอย่างง่ายดายโดยการล่วงรู้ข้อสอบก่อน (Item Pre-Knowledge) โดยไม่ต้องใช้ความสามารถของตนเอง คะแนนที่สังเกตได้จะขาดความถูกต้อง ผู้สอบสามารถทำคะแนนได้มากเกินความสามารถจริงของตนเอง และจะสูญเสียคุณสมบัติทางจิตมิติขาดความตรงตามสภาพ เนื่องจากแบบสอบไม่สามารถวัดได้ตรงตามสภาพของความสามารถจริงของผู้สอบ ถ้าปัญหาดังกล่าวไม่ได้รับการแก้ไข จะส่งผลกระทบต่อความปลอดภัยของข้อสอบและประสิทธิภาพการใช้ข้อสอบในคลังข้อสอบ (Pool Utilization) ตามมา นักทดสอบทั้งทางด้านจิตวิทยาและทางด้านการศึกษาหลายท่านจึงให้ข้อเสนอว่า เพื่อให้เกิดความเหมาะสมในทางปฏิบัติ การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ควรให้ความสำคัญในเรื่องการควบคุมการใช้ข้อสอบซ้ำ (Item Exposure Control) และการสร้างความสมดุลของเนื้อหา (Content Balancing) (Chang & Zhang, 2002; Chen et al., 2003; Chen & Doong, 2008; Chen & Lei, 2005; Chen et al., 2008; Cheng & Liou, 2003; Meijer & Nering, 1999)

จากการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องพบว่ามีผู้ศึกษาเกี่ยวกับการควบคุมการใช้ข้อสอบซ้ำเป็นจำนวนมาก และสามารถจำแนกได้เป็น 4 กลุ่ม ตามกลยุทธ์ที่ใช้งาน คือ 1) กลุ่มที่ใช้กระบวนการสุ่มเข้ามาช่วยควบคุมการใช้ข้อสอบซ้ำในขั้นตอนวิธีการคัดเลือกข้อสอบ (Randomization) 2) กลุ่มที่ใช้ขั้นตอนวิธีการคัดเลือกข้อสอบอย่างมีเงื่อนไขเพื่อควบคุมโอกาสที่ข้อสอบจะถูกนำออกมาแสดงมากเกินไป (Conditional Selection) 3) กลุ่มที่ใช้กลวิธีในการแบ่งคลังข้อสอบออกเป็นชั้นๆ 4) กลุ่มที่ใช้กลวิธีแบบผสม (Combined Strategies) ซึ่งนำจุดเด่นของวิธีการในกลุ่มทั้ง 3 กลุ่มมาผสมเข้าด้วยกัน รายชื่อของวิธีการในแต่ละกลุ่มที่มีผู้พัฒนาขึ้น นำเสนอในตาราง 2.1 และจุดเด่นและข้อจำกัดของกลยุทธ์แต่ละกลุ่มในตารางที่ 2.2

ตารางที่ 2.1 จำแนกวิธีการการควบคุมการใช้ข้อสอบซ้ำตามกลยุทธ์ที่ใช้งาน

ประเภทกลยุทธ์	แหล่งอ้างอิง
Randomization	
5-4-3-2-1 Procedure	(Davis & Dodd, 2003)
Randomesque strategy	(Kingsbury & Zara, 1989)
Choose One of Three	(Davis & Dodd, 2003)
Within .10 Logits Procedure	(Davis & Dodd, 2003)
Progressive strategy	(Revuelta & Ponsoda, 1998)
Conditional Selection	
Sympson-Hetter (SH) strategy	(Stocking, 1993)
Extended SH strategy	(Stocking, 1993)
Stocking and Lewis Multinomial strategy	(Stocking & Lewis, 1995)
Restricted Maximum Information strategy	(Revuelta & Ponsoda, 1998)
Stocking and Lewis Conditioning on Estimated Ability	(Stocking & Lewis, 2002)
Chen and Lei strategy	(Chen & Lei, 2005)
Shadow Test approach	van der Linden and Veldkamp, 2005
Stratified Strategies	
a-Stratified strategy (*a*-STR)	(Chang & Ying, 1999)
a-STR with *b*-Blocking	(Chang, Qian, & Ying, 2001)
0-1 Stratification strategy	(Chang & van der Linden, 2003)
a-STR with Content Balancing	(Leung et al., 2003)

ตารางที่ 2.1 จำแนกวิธีการการควบคุมการใช้ข้อสอบซ้ำตามกลยุทธ์ที่ใช้งาน (ต่อ)

ประเภทกลยุทธ์	แหล่งอ้างอิง
Combined Strategies	
Progressive Restricted strategy	(Revuelta & Ponsoda, 1998)
a-Stratified Design with the Sympson-Hetter Algorithm	(Leung et al., 2002)
Incorporation of the SH into *a*-STR with Content Blocking	(Leung et al., 2002)
Constraints CAT using Shadow Test	(Wim J. van der Linden, 2005)
Computerized Adaptive Sequential Testing	(Luecht & Nungester, 1998)
Adaptive Multi-stage Item Bundles	Leucht, 2003
Multiple Forms Structures	(Armstrong, Jones, Koppel, & Pashley, 2004)
Testlet-Based Adaptive Mastery Testing	(Vos & Glas, 2002)
Two-Phase ISP for Flexible Content Balancing	(Cheng et al., 2007)

ตารางที่ 2.2 จุดเด่นและข้อจำกัดของกลยุทธ์ของแต่ละกลุ่ม

ประเภทกลยุทธ์	จุดเด่น	ข้อจำกัด
Randomization	เป็นวิธีที่ง่ายในการนำไปใช้	เนื่องจากเป็นการใช้กระบวนการสุ่มจึงไม่รับประกันในเรื่องของอัตราการใช้ข้อสอบซ้ำที่เกิดขึ้น และสารสนเทศได้ต่ำเพราะเป็นเลือกข้อสอบอย่างสุ่ม
Conditional Selection	1. ควบคุมความน่าจะเป็นของข้อสอบที่ได้รับการเลือกแล้วให้มีการใช้ข้อสอบเป็นไปตามกำหนด 2. วิธีนี้รับรองว่าอัตราการใช้ข้อสอบสูงสุดเป็นไปตามกำหนดล่วงหน้า	ขั้นตอนมีความซับซ้อน และต้องศึกษาในสถานการณ์จำลองโดยการทำซ้ำจำนวนมาก และในการใช้งานจริง เมื่อองค์ประกอบต่างๆ ในการทดสอบเปลี่ยนแปลงไป จำเป็นต้องศึกษาในสถานการณ์จำลองใหม่ทุกครั้งเพื่อกำหนดพารามิเตอร์ควบคุมอัตราการใช้ข้อสอบซ้ำ

ตารางที่ 2.2 จุดเด่นและข้อจำกัดของกลยุทธ์ของแต่ละกลุ่ม (ต่อ)

ประเภทกลยุทธ์	จุดเด่น	ข้อจำกัด
Stratified Strategies	เพิ่มอัตราการใช้ข้อสอบซ้ำในข้อสอบที่มีค่าอำนาจจำแนกต่ำ และทำให้อัตราการใช้ข้อสอบซ้ำของข้อสอบทั้งคลังข้อสอบมีความเท่าเทียมกัน	ข้อสอบที่บรรจุภายในแต่ละชั้นควรมีค่าอำนาจจำแนกใกล้เคียงกัน และมีช่วงของค่าความยากของข้อสอบครอบคลุมช่วงความสามารถของผู้สอบ
Combined Strategies	เป็นการนำเอาลักษณะเด่นของแต่ละวิธีมารวมกัน ทำให้มีความสามารถมากกว่ากลยุทธ์เดี่ยว จากงานวิจัยในอดีตพบว่ามีการนำเทคนิคมาบูรณาการเข้าด้วยกันเป็นจำนวนมาก	ข้อจำกัดที่มีจะได้รับสืบทอดมาจากวิธีการต้นแบบที่นำมาใช้

1. กลุ่มที่ใช้กระบวนการสุ่มเข้ามาช่วยควบคุมการใช้ข้อสอบซ้ำในขั้นตอนวิธีการคัดเลือกข้อสอบ (Randomization) วิธีนี้การเลือกกลุ่มข้อสอบที่เหมาะสมที่สุดจากชุดข้อสอบ แทนการเลือกข้อสอบสารสนเทศสูงสุดเพียงข้อเดียว และสุ่มเลือกข้อสอบข้อหนึ่งจากกลุ่มข้อสอบจัดให้กับผู้สอบ วิธีการเลือกอย่างสุ่มไม่รับรองว่าการใช้ข้อสอบจะควบคุมได้ตามเกณฑ์ที่ต้องการ ตัวอย่างวิธีนี้ เช่น

วิธี 5-4-3-2-1 (5-4-3-2-1 Procedure) เสนอโดย McBride และ Martin (1983, cited in Georgiadou, Triantafillou, & Economides, 2007) วิธีนี้เน้นการเลือกข้อสอบในระยะเริ่มต้นของการทดสอบ เมื่อผู้สอบมีแนวโน้มว่าเริ่มต้นการทดสอบที่ระดับความสามารถเท่าเทียมกัน (Stocking & Lewis, 2000) เริ่มจากทำการเลือกข้อสอบสารสนเทศสูงสุดมา 5 ข้อ ข้อสอบข้อแรกที่จัดให้สุ่มเลือกจากข้อสอบสารสนเทศสูงสุด 5 อันดับ ข้อสอบข้อที่ 2 สุ่มเลือกจากข้อสอบสารสนเทศสูงสุดที่เหลืออีก 4 อันดับ และดำเนินการไปเรื่อยๆ จนถึงข้อที่ 5 ข้อดีของวิธีนี้คือทำได้ง่าย แต่มีข้อจำกัดบางประการ คือ อัตราการใช้ข้อสอบซ้ำสูงโดยเฉพาะกลุ่มข้อสอบสารสนเทศสูงและไม่รับรองว่าการใช้ข้อสอบจะควบคุมได้ตามเกณฑ์ที่ต้องการ

วิธีสุ่มเลือกจากกลุ่มข้อสอบ (Randomesque Procedure) เสนอโดย Kingsbury และ Zara (1989, cited in Georgiadou, Triantafillou, & Economides, 2007) วิธีนี้จะสุ่มเลือกข้อสอบเพียงข้อเดียวจากกลุ่มข้อสอบสารสนเทศสูงสุดจำนวนเท่าๆ กัน

วิธีสุ่มเลือกข้อสอบหนึ่งในสาม (Choose One of Three) เสนอโดย Thomasson และ Drasgow (1990, cited in Georgiadou, Triantafillou, & Economides, 2007) ใช้วิธีสุ่มเลือกข้อสอบจากข้อสอบสารสนเทศสูงสุด 3 ข้อ ข้อสอบอีก 2 ข้อที่ไม่ได้รับเลือกจะนำกลับไปจัดชุดใหม่เพื่อทำการเลือกครั้งต่อไป

วิธีสุ่มเลือกโลจิทส์ (Within .10 Logits Procedure) เสนอโดย Lunz และ Stahl (1988, cited in Georgiadou, Triantafillou, & Economides, 2007) วิธีนี้ข้อสอบได้รับเลือกจากการสุ่มในกลุ่มข้อสอบความยากเดียวกันและจับคู่ระหว่างค่าความสามารถของผู้สอบกับค่าความยากของข้อสอบ ถ้าไม่ได้ข้อสอบที่มีความยากตามที่ต้องการ จะสุ่มเลือกข้อสอบความยากใกล้เคียงค่าความยากเป้าหมายแทน

วิธีสุ่มเลือกแบบก้าวหน้า (Progressive Procedure) เสนอโดย Revuelta และ Ponsoda (1998) เป็นวิธีรวมการสุ่มเลือกและการเลือกจากสารสนเทศสูงสุดเข้าไว้ด้วยกันในระยะแรกใช้วิธีการสุ่มเลือกมากกว่าการเลือกจากสารสนเทศสูงสุด แต่เมื่อการทดสอบดำเนินต่อไปสารสนเทศจะมีผลต่อการเลือกข้อสอบมากกว่าวิธีการสุ่มเลือกเพื่อควบคุมโอกาสที่ข้อสอบจะถูกนำออกมาแสดงมากเกินไป (Conditional Selection)

2. กลุ่มที่ใช้ขั้นตอนวิธีการคัดเลือกข้อสอบอย่างมีเงื่อนไข วิธีการเลือกอย่างมีเงื่อนไขเป็นวิธีควบคุมความน่าจะเป็นของข้อสอบที่ได้รับการเลือกแล้วให้มีการใช้ข้อสอบเป็นไปตามกำหนด วิธีนี้รับรองว่าอัตราการใช้ข้อสอบสูงสุดเป็นไปตามกำหนด แต่ขั้นตอนมีความซับซ้อน และต้องศึกษาในสถานการณ์จำลองในด้านจำนวนครั้งสำหรับการใช้ปฏิบัติ เมื่อองค์ประกอบต่างๆในการทดสอบเปลี่ยนแปลงไป จำเป็นต้องศึกษาในสถานการณ์จำลองก่อนใช้ในสถานการณ์จริง มีวิธีต่างๆ ได้แก่

วิธี Sympson-Hetter: SH เสนอโดย Sympson และ Hetter (1985) วิธีนี้เป็นที่รู้จักกันแพร่หลาย ดำเนินการโดยใช้ความถี่ของข้อสอบที่จัดให้กับกลุ่มผู้สอบจำนวนมากในสถานการณ์จำลองเปรียบเทียบกับอัตราการใช้ข้อสอบเป้าหมายเพื่อหาพารามิเตอร์การใช้ข้อสอบ ขบวนการนี้ทำซ้ำๆ กันจนพารามิเตอร์การใช้ข้อสอบแต่ละข้อมีค่าระหว่าง 0 ถึง 1 พารามิเตอร์นี้ใช้สำหรับการเลือกข้อสอบในการทดสอบสถานการณ์จริงโดยสร้างตัวเลขสุ่มจากการแจกแจงเหมือนกัน (0 ถึง 1) เปรียบเทียบกับพารามิเตอร์การใช้ข้อสอบซ้ำ ถ้าพารามิเตอร์การใช้ข้อสอบซ้ำมากกว่าตัวเลขสุ่มจะจัดข้อสอบให้ แต่ถ้าพารามิเตอร์การใช้ข้อสอบซ้ำน้อยกว่าตัวเลขสุ่มจะทำการเลือกข้อสอบข้อใหม่แทน

ข้อดีของวิธีนี้คือ ยอมให้มีการกำหนดอัตราการใช้ข้อสอบซ้ำล่วงหน้า เพื่อยืนยันว่าการแจกแจงอัตราการใช้ข้อสอบซ้ำควรมีลักษณะเดียวกัน วิธีนี้ยังเป็นวิธีพื้นฐานประยุกต์ไปสู่วิธีการเลือกอย่างมีเงื่อนไขอื่นๆ อีกด้วย (Stocking & Lewis, 2000)

ข้อควรระวังสำหรับการใช้วิธีควบคุมแสดงข้อสอบแบบ Sympson-Hetter ได้แก่ 1) ถ้าการแจกแจงค่าความสามารถของผู้สอบไม่เหมาะสม นำไปสู่การเคร่งครัดหรือผ่อนปรนพารามิเตอร์การใช้ข้อสอบซ้ำมากเกินไป ดังนั้น เมื่อมีการแจกแจงค่าความสามารถของผู้สอบใหม่ จะต้องคำนวณพารามิเตอร์การใช้ข้อสอบซ้ำใหม่ซ้ำทุกครั้ง 2) พารามิเตอร์การใช้ข้อสอบซ้ำวิธี Sympson-Hetter เป็นค่าโดยรวมของการควบคุมการใช้ข้อสอบซ้ำ จะไม่เหมาะสมสำหรับค่าความสามารถของผู้สอบบริเวณส่วนปลายของโค้งการแจกแจงค่าความสามารถของผู้สอบ และ 3) วิธี Sympson-Hetter ให้ผลในการลดข้อสอบที่มีอัตราการใช้ข้อสอบซ้ำสูงเกินไป แต่ไม่มีผลต่อข้อสอบที่มีอัตราการใช้ข้อสอบต่ำ

วิธี Sympson-Hetter แบบมีเงื่อนไข (Conditional Sympson-Hetter) เป็นวิธีที่พัฒนาจากวิธี Sympson-Hetter ที่เสนอโดย Stocking และ Lewis (1998) พารามิเตอร์การใช้ข้อสอบซ้ำได้จากการประมาณระดับค่าความสามารถแทนการแจกแจงค่าความสามารถของผู้สอบ ระหว่างขั้นตอนสถานการณ์จำลอง จะมีการสร้าง n × m เมทริกซ์ เมื่อแถวนอน n คือจำนวนข้อสอบในคลังข้อสอบ และแถวตั้ง m คือ จำนวนจุดค่าความสามารถที่กระจายบนช่วงการแจกแจงค่าความสามารถ ดังนั้น ข้อสอบแต่ละข้อจะมีพารามิเตอร์การใช้ข้อสอบซ้ำที่แต่ละจุดค่าความสามารถของผู้สอบ

3. กลุ่มที่ใช้กลวิธีในการแบ่งคลังข้อสอบออกเป็นชั้นๆ วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ตามระดับชั้นของค่าอำนาจจำแนก (a-Stratified Method: a-STR) วิธีนี้เสนอโดย (Chang & Ying, 1996) ประยุกต์มาจากรูปแบบปรับระดับชั้นของ Weiss (1985) เป็นวิธีการเลือกข้อสอบและมีประสิทธิภาพในการควบคุมการใช้ข้อสอบซ้ำโดยอัตโนมัติ เริ่มจากการแบ่งชั้นคลังข้อสอบตามค่าอำนาจจำแนกของข้อสอบ และแบ่งแบบทดสอบเข้าในแต่ละชั้นคลังข้อสอบ ชั้นแรกบรรจุข้อสอบอำนาจจำแนกต่ำสุดชั้นต่อไปบรรจุข้อสอบอำนาจจำแนกสูงกว่าชั้นแรก ทำเช่นนี้เรื่อยๆ จนกระทั่งชั้นสุดท้ายบรรจุข้อสอบค่าอำนาจจำแนกสูงสุด การทดสอบจะดำเนินจากชั้นแรกจนถึงชั้นสุดท้าย วิธีนี้กระจายการใช้ข้อสอบซ้ำในคลังข้อสอบให้เท่าเทียมกัน

1.3.6 การสร้างความสมดุลของเนื้อหา (Content Balancing)

ความสมดุลของเนื้อหา (Content Balancing) ตามการกำหนดคุณลักษณะของข้อสอบ (Content Specification) เป็นสิ่งสำคัญโดยเฉพาะอย่างยิ่งในการวัดผลสัมฤทธิ์ทางการศึกษา ในการทดสอบแบบดั้งเดิมที่เป็นมาตรฐาน แบบสอบที่มีรูปแบบคู่ขนานส่วนมากถูกสร้างตามรายละเอียดของตารางการกำหนดคุณลักษณะของข้อสอบ (Table of Content Specification) และในการออกแบบการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ควรคำนึงถึงข้อจำกัดในทางปฏิบัติ (Practical Constraints) เช่น การกำหนดคุณลักษณะของข้อสอบ และความสมดุลของเนื้อหา (Content Balancing) เพื่อรับประกันว่าข้อสอบที่ผู้สอบแต่ละคนได้รับเป็นตัวแทนตามสัดส่วนของแต่ละขอบเขตเนื้อหา (Content Area) และเพื่อให้แน่ใจว่าผู้ที่เข้าสอบได้รับการทดสอบในเนื้อหาที่เท่ากันและเหมาะสม ในระหว่างการใช้ CAT วิธีการคัดเลือกข้อสอบควรจะเลือกข้อสอบที่ให้สารสนเทศมากที่สุดในการประมาณค่าความสามารถของผู้ที่เข้าสอบในช่วงเวลานั้น แต่ไม่บรรลุถึงการกำหนดแบบแผนของเนื้อหา (Content Specifications) ของข้อสอบทั้งหมด ตัวอย่างเช่น การทดสอบทางด้านการคำนวณเพื่อที่จะให้ทราบถึงความสามารถ ของผู้เข้าสอบในเรื่อง การบวก การลบ การคูณ และการหาร ทั้งหมด แต่ใน CAT อาจจะคัดเลือกให้ทำข้อสอบเฉพาะเรื่องบางเรื่องเท่านั้น เช่น เรื่องการลบ และการคูณ ซึ่งผู้เข้าสอบจะไม่ได้รับการสอบในเรื่องการบวก และการหาร ผู้สอบที่ไม่มีความรู้เกี่ยวกับการบวก และการหาร ก็จะถูกประมาณค่าความสามารถสูงเกินจริง สำหรับผู้เข้าสอบที่มีความรู้เกี่ยวกับเรื่องบวก และหาร จะได้รับการประมาณค่าความสามารถที่แท้จริงของต่ำเกินไป ดังนั้น การสร้างความสมดุลของเนื้อหาจึงเป็นสิ่งจำเป็นในการออกแบบระบบ CAT (Boyd, 2003)

ตอนที่ 2 โมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย (Testlet Response Theory)

การพัฒนาแบบสอบมีวิธีการทางเทคนิคและทางการปฏิบัติที่หลากหลายในการสร้างแบบสอบที่เป็นข้อสอบแบบเลือกตอบ (multiple choice test items) ซึ่งสามารถตรวจให้คะแนนได้ง่ายและมีความเป็นปรนัยสูง ด้วยเหตุนี้ข้อสอบแบบเลือกตอบจึงถูกใช้อย่างกว้างขวางในการทดสอบทางการศึกษา ในปัจจุบันมีการประยุกต์ใช้ข้อสอบแบบเลือกตอบเพื่อช่วยในการวัดทักษะความสามารถที่มีความซับซ้อน เช่น ทักษะการให้เหตุผลเชิงวิเคราะห์ และทักษะการให้เหตุผลเชิงตรรกะ ทำให้ข้อสอบเดี่ยว (single item) ไม่สามารถวัดทักษะที่มีความซับซ้อนได้อย่างครอบคลุม จึงมีการพัฒนาเทคนิคใหม่ๆ ที่ใช้ในการสร้างข้อสอบเพื่อให้สามารถวัดทักษะที่ซับซ้อนได้อย่างครอบคลุม จึงเกิดแนวคิดของการจัดกลุ่มข้อสอบโดยใช้สิ่งเร้าร่วมกัน ข้อสอบลักษณะนี้จึงถูกมองว่าเป็นแบบสอบฉบับย่อย (testlet) ซึ่งรวมกันอยู่ในแบบสอบฉบับเต็ม ตัวอย่างสถานการณ์ที่นำแบบทดสอบย่อยมาใช้

ได้แก่ การสอบ TOEFL ในส่วนที่เป็นการอ่านเพื่อความเข้าใจ (reading comprehension) ข้อสอบในส่วนนี้จะมีลักษณะเป็นแบบทดสอบย่อย (testlet) โดยกลุ่มข้อสอบประมาณ 6 ถึง 12 ข้อจะใช้บทความ (passage) ที่มีความยาวประมาณ 250 ถึง 400 คำ เป็นสิ่งเร้าร่วมกันเพื่อให้ผู้สอบตอบคำถาม ด้วยเหตุนี้ การตอบข้อคำถามภายในแบบทดสอบย่อยของผู้สอบจึงขึ้นอยู่กับสิ่งเร้าที่ข้อสอบใช้ร่วมกัน (common stimulus) ดังนั้น การตอบคำถามถูกหรือผิดในแบบทดสอบย่อยไม่ได้ขึ้นอยู่กับความสามารถของผู้สอบเพียงอย่างเดียวแต่ยังขึ้นอยู่กับความเข้าใจในการตีความหรือแปลความจากสิ่งเร้า นอกจากนั้นผู้สอบอาจเกิดการเรียนรู้จากการตอบข้อสอบข้ออื่นในแบบสอบฉบับย่อย การตอบคำถามก่อนหน้าจึงอาจส่งผลต่อการตอบในข้อถัดไปด้วย ดังนั้นถ้าเข้าใจแล้วแปลความหรือตีความสิ่งเร้าได้ถูกต้อง ก็จะเกิดการเรียนรู้ในสิ่งที่ถูกต้องทำให้สามารถตอบคำถามข้ออื่นในแบบทดสอบย่อยได้ในทางตรงข้ามหากเกิดความเข้าใจผิดพลาด อาจทำให้ผลการตอบคำถามข้ออื่นๆ ในแบบทดสอบย่อยผิดตามไปด้วย

จากสถานการณ์ดังกล่าวจึงมีการคิดวิธีการให้คะแนนข้อสอบที่มีลักษณะเป็นแบบทดสอบย่อย (testlet) โดยมุมมองของทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) มีวิธีการให้คะแนนทั้งหมด 3 วิธีดังนี้ 1) ข้อสอบในแบบทดสอบย่อยแต่ละข้อจะตรวจให้คะแนนแบบทวิวิภาคและวิเคราะห์ตามโมเดลการตอบสนองข้อสอบแบบ โลจิส 3 พารามิเตอร์ 2) พิจารณาว่ากลุ่มของข้อสอบในแบบทดสอบย่อยเป็นข้อสอบข้อเดียวที่มีการให้คะแนนแบบหลายค่าและวิเคราะห์ตามทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค (Polytomous IRT) และ 3) ให้คะแนนและวิเคราะห์ตามทฤษฎีการตอบสนองแบบทดสอบย่อย (Testlet Response Theory: TRT) (Wainer, Bradlow, & Du, 2002) จากทั้ง 3 วิธีที่กล่าวมาแต่ละวิธีมีจุดเด่นและข้อจำกัดแตกต่างกันดังนี้

วิธีที่ 1) เป็นวิธีที่นิยมใช้กันมากเนื่องจากเป็นแนวคิดที่เข้าใจง่ายและไม่ซับซ้อน แต่ฝ่าฝืนข้อตกลงเบื้องต้นของ IRT เกี่ยวกับความเป็นอิสระ (local dependence) ที่กล่าวว่าผลการตอบข้อสอบรายข้อไม่มีความสัมพันธ์กันเนื่องจากโมเดลการตอบสนองข้อสอบมองว่าความสามารถผู้สอบเป็นปัจจัยเดียวเท่านั้นที่มีอิทธิพลต่อผลการตอบรายข้อ แต่เมื่อพิจารณาแล้วจะพบว่า ผลการตอบข้อสอบในแบบทดสอบย่อยจะมีความสัมพันธ์กัน เนื่องจาก กลุ่มของข้อสอบใช้สิ่งเร้าร่วมกัน สิ่งเร้าจึงเป็นอีกปัจจัยหนึ่งที่มีอิทธิพลต่อผลการตอบข้อสอบที่นอกเหนือจากความสามารถของผู้สอบ เช่น ผู้สอบสามารถตอบข้อสอบแต่ละข้อในแบบทดสอบย่อย (testlet) ถูกหรือไม่นั้นขึ้นอยู่กับความเข้าใจในการอ่านบทความ (reading passage) ที่เป็นสิ่งเร้านั้น ซึ่งการละเลยข้อตกลงเบื้องต้นนี้ ถ้าสิ่งเร้ามีอิทธิพลทางบวกจะมีแนวโน้มที่ทำให้ความแม่นยำของการวัดที่ได้จากแบบทดสอบย่อยมีค่าสูงกว่าความเป็นจริง (overestimate) แต่ถ้าสิ่งเร้ามีอิทธิพลทางลบจะมีแนวโน้มที่ทำให้ความแม่นยำของการวัดที่ได้จากแบบทดสอบย่อยมีค่าต่ำกว่าความเป็นจริง (underestimate)

วิธีที่ 2) เป็นวิธีที่ใช้ได้ดีในสถานการณ์ทั่วๆ ไป เมื่อแบบทดสอบย่อยเหมาะสม (fit) กับโมเดลการตอบสนองข้อสอบแบบพหุวิภาค (Polytomous IRT) คะแนนของแบบทดสอบย่อยจะแสดงเป็นจำนวนข้อที่ทำถูก สารสนเทศบางอย่าง เช่น สารสนเทศเกี่ยวกับแบบแผนการตอบ (response patterns) จะสูญหายไปเนื่องจากในการวิเคราะห์แบบทดสอบย่อยจะถูกมองว่าเป็นข้อสอบเดี่ยว (single item) ที่มีการให้คะแนนหลายแบบค่า นอกจากนี้ ในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ข้อสอบที่ถูกเลือกภายในแบบทดสอบย่อยจะเป็นการเลือกข้อสอบในลักษณะที่เรียกว่าอิสระอย่างมีเงื่อนไข (conditional independent) ซึ่งส่งผลต่อความถูกต้องในการให้คะแนน

วิธีที่ 3) เป็นการนำเอาโมเดลการตอบสนองข้อสอบ (IRT) มาปรับแก้ โดยเพิ่มเติมอิทธิพลสุ่ม (random effect) สำหรับข้อสอบที่อยู่ภายในแบบทดสอบย่อยเดียวกัน จุดเด่นของการวิเคราะห์โดยใช้ทฤษฎีการตอบสนองแบบทดสอบย่อย (Testlet Response Theory: TRT) มี 2 ประการคือ 1) หน่วยของการวิเคราะห์ยังเป็นข้อสอบในแบบทดสอบย่อย และไม่มองแบบทดสอบย่อยเป็นข้อสอบข้อเดียวเหมือนในโมเดลการตอบสนองข้อสอบแบบพหุวิภาค (Polytomous IRT) ดังนั้น สารสนเทศเกี่ยวกับแบบแผนการตอบ (response patterns) ภายในแบบทดสอบย่อยจะไม่สูญหาย และ 2) แนวคิดของพารามิเตอร์ข้อสอบ เช่น อำนาจจำแนกและความยากของข้อสอบยังคงมีความสมเหตุสมผลและสามารถใช้งานได้ภายใต้โมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย

เมื่อพิจารณาข้อสอบที่รวมกันอยู่เป็นแบบทดสอบย่อย จากการศึกษาถึงอิทธิพลของการละเลยปฏิสัมพันธ์ของข้อสอบที่เกิดขึ้นภายใต้โมเดล 2 พารามิเตอร์แบบโลจิส ของ Tuerlinckx and De Boeck (2001) พบว่า ขนาดของการประมาณค่าพารามิเตอร์อำนาจจำแนกมีความลำเอียงโดยขึ้นอยู่กับปฏิสัมพันธ์ของพารามิเตอร์ความยาก ในกรณีที่มีปฏิสัมพันธ์เชิงบวกการประมาณค่าพารามิเตอร์อำนาจจำแนกจะสูงเกินจริง (Overestimating) ทำให้สารสนเทศของข้อสอบสูงขึ้นผิดปกติ (inflated) และความคลาดเคลื่อนมาตรฐานก็ถูกกดให้ต่ำลง (deflated) แต่ในทางตรงกันข้าม ถ้าเกิดปฏิสัมพันธ์เชิงลบการประมาณค่าพารามิเตอร์อำนาจจำแนกจะต่ำเกินจริง (Underestimating) ทำให้สารสนเทศของข้อสอบถูกทำให้ต่ำลงอย่างผิดปกติ (deflated) และความคลาดเคลื่อนมาตรฐานของข้อสอบสูงขึ้น (inflated) ซึ่งในทางปฏิบัติคุณภาพของข้อสอบอธิบายได้จากขนาดของพารามิเตอร์อำนาจจำแนก (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985) ดังตัวอย่าง ฟังก์ชันสารสนเทศของข้อสอบสำหรับ ข้อสอบข้อที่ i ภายใต้โมเดล 2 พารามิเตอร์แบบโลจิส ซึ่งถูกกำหนดไว้ดังนี้

$$I_i(\theta_n) = a_i^2 \Pr(X_i = 1 \mid \theta_n)[1 - \Pr(X_i = 1 \mid \theta_n)]$$

เมื่อ $\Pr(X_i = 1 | \theta_n)$ คือ ความน่าจะเป็นในการตอบข้อสอบข้อที่ i ถูก สำหรับผู้สอบคนที่ n จากสูตรจะพบว่า ค่าสารสนเทศของข้อสอบมีความสัมพันธ์เชิงบวกกับค่าพารามิเตอร์อำนาจจำแนกยกกำลังสอง พารามิเตอร์อำนาจจำแนกที่มีค่ามากจะมีอำนาจจำแนกที่แข็งแกร่งในการจำแนกระหว่างผู้สอบที่อยู่ในตำแหน่งที่ใกล้เคียงกันของพารามิเตอร์ความยาก ดังนั้น ข้อสอบที่มีค่าอำนาจจำแนกสูงจึงได้รับการพิจารณาว่าเป็นข้อสอบที่มีคุณภาพ เนื่องจาก ค่าพารามิเตอร์อำนาจจำแนกมักถูกใช้เป็นดัชนีในการบ่งชี้คุณภาพของข้อสอบ ดังนั้นจึงควรทำให้มั่นใจว่าการประมาณค่าพารามิเตอร์อำนาจจำแนกจะไม่เกิดความลำเอียง โดยการไม่ละเลยประเด็นเรื่องความเป็นอิสระของข้อสอบที่จะทำให้ข้อสอบภายในแบบทดสอบย่อยเกิดปฏิสัมพันธ์กัน

เนื่องจากทฤษฎีการตอบสนองข้อสอบ (IRT) เป็นโมเดลที่มีข้อตกลงเบื้องต้นที่เคร่งครัด การฝ่าฝืนข้อตกลงเบื้องต้นในเรื่องของความเป็นอิสระในการตอบข้อสอบ (Local Independence) ที่กล่าวว่า เมื่อค่าความสามารถของผู้สอบเป็นค่าแน่นอน การตอบข้อสอบแต่ละข้อของผู้สอบคนหนึ่งจะมีความเป็นอิสระจากกัน หรืออาจกล่าวได้ว่าการตอบข้อสอบข้อใดๆ ของผู้สอบจะไม่มีผลต่อข้อสอบข้ออื่นๆ สิ่งที่ส่งผลต่อการตอบข้อสอบแต่ละข้อเป็นผลมาจากความสามารถของผู้สอบเท่านั้น (Hambleton & Swaminathan, 1985) ความเป็นอิสระในการตอบข้อสอบทำให้พารามิเตอร์ข้อสอบยังเป็นค่าคงที่ ไม่ว่าข้อสอบข้อนั้นอยู่ตำแหน่งใดๆ บนแบบทดสอบเกิดขึ้นเมื่อการตอบสนองของผู้สอบต่อข้อสอบในแบบสอบเป็นความสัมพันธ์กันอย่างมีเงื่อนไข ทั้งเงื่อนไขทางบวก และทางลบ ซึ่งรูปแบบความน่าจะเป็นของการตอบข้อสอบจะเปลี่ยนแปลงไม่คงที่ตามรูปแบบของทฤษฎีการตอบข้อสอบที่ได้กำหนดไว้

ดังนั้นทฤษฎีการตอบสนองแบบทดสอบย่อย (Testlet Response Theory: TRT) จึงเป็นแนวทางการแก้ปัญหาในการคิดคะแนนข้อสอบที่มีลักษณะเป็นแบบทดสอบย่อยที่ดูสมเหตุสมผลที่สุด เนื่องจาก จุดเด่นของการวิเคราะห์โดยใช้ทฤษฎีการตอบสนองแบบทดสอบย่อย (Testlet Response Theory: TRT) มี 2 ประการคือ 1) หน่วยของการวิเคราะห์ยังเป็นข้อสอบในแบบทดสอบย่อย และไม่มองแบบทดสอบย่อยเป็นข้อสอบข้อเดียวเหมือนในโมเดลการตอบสนองข้อสอบแบบพหุวิภาค (Polytomous IRT) ดังนั้น สารสนเทศเกี่ยวกับแบบแผนการตอบ (response patterns) ภายในแบบทดสอบย่อยจะไม่สูญหาย และ 2) แนวคิดของพารามิเตอร์ข้อสอบ เช่น อำนาจจำแนกและความยากของข้อสอบยังคงมีความสมเหตุสมผลและสามารถใช้งานได้ภายใต้โมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย

โมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อยจึงเป็นส่วนขยายของโมเดลการตอบสนองข้อสอบ (IRT) ซึ่งถูกปรับแก้โดยการเพิ่มอิทธิพลสุ่ม (random effect) ไปยังสมการโลจิท (logit of

equation) โดยที่โมเดลการตอบสนองข้อสอบแบบ 3 พารามิเตอร์ ที่เสนอโดย Birnbaum (1968) มีโค้งลักษณะข้อสอบที่เขียนด้วยฟังก์ชันโลจิส ดังสมการ

$$P(y_{ij}=1)=c_j+(1-c_j)\frac{\exp[a_j(\theta_i-b_j)]}{1+\exp[a_j(\theta_i-b_j)]}$$

เมื่อ y_{ij} คือ คะแนนสำหรับข้อสอบข้อที่ j ที่ผู้สอบคนที่ 1 ได้รับ

$P(y_{ij}=1)$ คือ ความน่าจะเป็นที่ผู้สอบคนที่ i ตอบข้อสอบข้อที่ j ถูก

θ_i คือ ความสามารถของผู้สอบที่ตอบข้อสอบข้อที่ i

a_j คือ พารามิเตอร์อำนาจจำแนกของข้อสอบข้อที่ j

b_j คือ พารามิเตอร์ความยากของข้อสอบข้อที่ j

c_j คือ พารามิเตอร์การเดาของข้อสอบข้อที่ j

การอธิบายรายละเอียดเกี่ยวกับที่มาของสมการสามารถศึกษาเพิ่มเติมได้จาก Hambleton & Swaminathan (1985)

โมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อยแบบ 3 พารามิเตอร์

โมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อยแบบ 3 พารามิเตอร์ (Wainer & Wang, 2000) ที่ได้รับการปรับแก้โดยเพิ่มอิทธิพลสุ่ม (random effect) คือ ปฏิสัมพันธ์ของผู้สอบคนที่ i กับแบบทดสอบย่อยที่ d(j) ซึ่งมีข้อสอบ j ข้อ ไปยังสมการโลจิท (logit of equation) ในโมเดลการตอบสนองข้อสอบ (IRT) โมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อยแบบ 3 พารามิเตอร์ มีโค้งลักษณะข้อสอบที่เขียนด้วยฟังก์ชันโลจิสติก ดังสมการ

$$P(y_{ij}=1)=c_j+(1-c_j)\frac{\exp[a_j(\theta_i-b_j-\gamma_{id(j)})]}{1+\exp[a_j(\theta_i-b_j-\gamma_{id(j)})]}$$

เมื่อ $\gamma_{id(j)}$ คือ อิทธิพลสุ่มข้อสอบข้อที่ i ซึ่งอยู่ในแบบทดสอบย่อยฉบับ d ที่มีต่อผู้สอบคนที่ j ในกรณีที่ข้อสอบข้อที่ i เป็นข้อสอบที่อิสระจากข้ออื่น $\gamma_{id(j)}=0$

ฟังก์ชันสารสนเทศของข้อสอบ (Item Information Function)

ฟังก์ชันสารสนเทศของข้อสอบในโมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อยแบบ 3 พารามิเตอร์ เป็นดัชนีผสมที่สร้างจากคุณลักษณะของข้อสอบหลายลักษณะ ประกอบด้วย ค่าพารามิเตอร์ความยาก ค่าพารามิเตอร์อำนาจจำแนก ค่าพารามิเตอร์การเดา ความแปรปรวนของ

คะแนนรายข้อ และอิทธิพลสุ่มในแบบทดสอบย่อย เพื่อบ่งชี้คุณภาพของข้อสอบ สามารถเขียนในรูปของสมการทางคณิตศาสตร์ได้ดังนี้ (Wainer & Wang, 2000)

$$I(\theta_i) = E[-\partial / \partial \theta_i^2 \log(p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}})]$$

$$= E[a_j^2 (\frac{\exp(t_{ij})}{(1+\exp(t_{ij}))^2} - y_{ij} c_j \frac{\exp(t_{ij})}{(c_j + \exp(t_{ij}))^2})]$$

$$= a_j^2 \left(\frac{\exp(t_{ij})}{(1+\exp(t_{ij})} \right)^2 \left(\frac{1-c_j}{c_j + \exp(t_{ij})} \right)$$

เมื่อ $I(\theta_i)$ คือ ค่าสารสนเทศที่ได้รับจากข้อสอบข้อที่ I สำหรับผู้สอบที่มีความสามารถ θ

t_{ij} คือ ตัวทำนายเชิงเส้นแบบโลจิท โดยที่ $t_{ij} = a_j(\theta_j - b_j - \gamma_{id(j)})$ สำหรับแบบทดสอบย่อยฉบับที่ d

E คือ ค่าคาดหวัง โดยที่ $\exp(x) = e^x$

ตอนที่ 3 วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ แบบมอนติ คาร์โล

วิธีมอนติ คาร์โล (Monte Carlo methods) ถือกำเนิดขึ้นช่วงทศวรรษที่ 1940 โดยนักคณิตศาสตร์และนักฟิสิกส์ในโครงการสร้างระเบิดปรมาณู (Manhattan Project) ระหว่างการศึกษานิวตรอน และได้รับการตั้งชื่อตามคาสิโนแห่งหนึ่งในประเทศโมนาโก สาเหตุของการใช้ชื่อดังกล่าวนั้นเนื่องจากวิธีมอนติ คาร์โล เป็นกระบวนการหรือขั้นตอนวิธีการคำนวณผลลัพธ์ เพื่อแก้ปัญหาทางคณิตศาสตร์ที่ไม่สามารถหาผลลัพธ์หรือคำตอบแม่นตรงได้ ซึ่งใช้เลขสุ่มที่สะท้อนถึงความไม่แน่นอนเป็นองค์ประกอบหลักดังนั้นวิธีการมอนติ คาร์โล จึงถูกนำมาใช้อย่างแพร่หลายในการศึกษาสถานการณ์จำลองที่มีความไม่แน่นอน หรือที่เรียกว่าการจำลองแบบสโตแคสติก (stochastic simulation) (Spall, 2003)

แต่ในงานวิจัยครั้งนี้นำวิธีมอนติ คาร์โล มาใช้ใน 2 ลักษณะ คือ 1) นำวิธีมอนติ คาร์โล มาใช้สร้างเวคเตอร์สุ่ม คือ ค่าความสามารถจริงของกลุ่มตัวอย่าง และ พารามิเตอร์ของข้อสอบซึ่งเป็นคุณลักษณะของคลังข้อสอบ และ 2) นำวิธีมอนติ คาร์โล มาใช้ในกระบวนการรวมข้อสอบเป็นแบบสอบที่มีคุณสมบัติตรงตามเงื่อนไขบังคับ และสร้างเป็นลำดับเชิงสุ่มของข้อสอบซึ่งจะถูกคัดเลือกและนำไปใช้กับผู้สอบ ขั้นตอนดังกล่าวถือเป็นส่วนหนึ่งในกระบวนการทำงานของระบบการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล (Monte Carlo CAT) ซึ่งพัฒนาโดย Belov, Armstrong, และ Weissman (2008)

แนวคิดของ Monte Carlo CAT ได้รับการพัฒนาขึ้นเพื่อตอบปัญหาพื้นฐานของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (Fundamental Problem of CAT: FPCAT) ที่ทำให้เกิดผลที่ไม่พึงปรารถนาจากการเลือกข้อสอบที่ให้สารสนเทศสูงสุดมีดังนี้ 1) เมื่อค่าประมาณความสามารถของผู้สอบช่วงเริ่มต้นถูกกำหนดให้คงที่ 2) ข้อสอบที่มีค่าสารสนเทศสูงสุดที่ระดับความสามารถของผู้สอบในบริเวณใกล้เคียงกับค่าประมาณความสามารถของผู้สอบจะถูกนำไปใช้ถี่กว่าข้อสอบข้ออื่น สิ่งเหล่านี้จะมีผลทำให้การใช้คลังข้อสอบมีประสิทธิภาพต่ำ (Low Utilization of Pool) และ 3) ถ้าผู้สอบที่มีความสามารถสูงและทำข้อสอบผิดบ่อยครั้งในช่วงเริ่มต้นการทดสอบ จะมีผลทำให้การค่าประมาณความสามารถของผู้สอบจะมีความแม่นยำต่ำกว่าที่ควรเป็น

โดยมีแนวคิดหลัก ดังนี้

ขั้นที่ 1 สร้างลำดับเชิงสุ่มของข้อสอบ (Generate a random sequence of items)

ขั้นที่ 2 ตรวจสอบข้อสอบที่สร้างขึ้นว่ามีคุณสมบัติตรงตามเงื่อนไขบังคับทั้งหมดหรือไม่ ถ้าตรงให้นำลำดับของข้อสอบเชิงสุ่มบันทึกในแบบสอบใหม่ (test form) ถ้าไม่ ให้ทำซ้ำในขั้นที่ 1

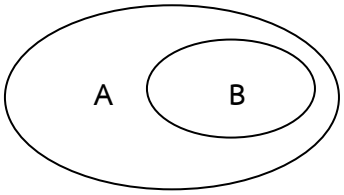
เนื่องจากการนำวิธีมอนติ คาร์โล มาประยุกต์ใช้ในการสร้างลำดับเชิงสุ่มของข้อสอบเพื่อรวมเป็นแบบสอบ แต่การใช้วิธีมอนติ คาร์โล เพียงอย่างเดียวในการสร้างลำดับของข้อสอบเชิงสุ่ม มีความเป็นไปได้ที่จะเกิดลำดับของข้อสอบเชิงสุ่มจำนวนมากที่ไม่มีประโยชน์ ดังนั้น Belov & Armstrong (2005) จึงใช้กลวิธีที่หลากหลายในการย่อขอบเขตการค้นหา (shrinking search region) เพื่อแก้ปัญหาการสร้างลำดับของข้อสอบเชิงสุ่มที่ไม่มีประโยชน์โดยการนำคุณสมบัติของเงื่อนไขบังคับของแบบสอบ (test constraints) ที่มีคุณลักษณะเป็นตัวแปรจัดประเภทมาใช้ประโยชน์โดยอาศัย 1) เทคนิคแบ่งเพื่อเอาชนะ (divide and conquer) 2) วิธีการค้นหาแบบตาบู (Tabu search) และ 3) การจัดลำดับความสำคัญของเงื่อนไขบังคับ (prioritization of the test constraints) จากนั้นจึงใช้วิธีการมอนติ คาร์โล เพื่อสร้างลำดับเชิงสุ่มของข้อสอบ โดยแนวคิดของแต่ละวิธีที่นำมาใช้มีดังนี้

1. เทคนิคแบ่งเพื่อเอาชนะ (divide and conquer)

เทคนิคแบ่งเพื่อเอาชนะ (divide and conquer) เป็นขั้นตอนวิธี (Algorithm) ที่ใช้ในการแก้ปัญหาโดยนำปัญหาหลักที่มีมาทำการแยกออกเป็นปัญหาย่อย ๆ แล้วนำคำตอบที่ได้จากปัญหาย่อยต่างๆ มารวมเข้าด้วยกัน เพื่อหาคำตอบของปัญหาใหญ่

จากแนวคิดดังกล่าวจึงได้นำมาประยุกต์ใช้ในการลดขอบเขตของการค้นหาข้อสอบในขั้นแรก โดยแบ่งขอบเขตการค้นหาเป็นขอบเขตการค้นหาย่อยๆ ในที่นี้ขอยกตัวอย่างให้ เซต A คือ ขอบเขตการค้นหาใหญ่ ซึ่งประกอบไปด้วยการรวมกันของข้อสอบทั้งหมดที่เป็นไปได้ (all possible

combinations of items) และ เซตย่อย B คือ การรวมกันของข้อสอบทั้งหมด (all combinations of items) จากแบบสอบที่ตรงตามเงื่อนไขบังคับ (feasible test form) ซึ่งเป็นสับเซตของเซต A ดังแสดงใน แผนภาพที่ 2.1

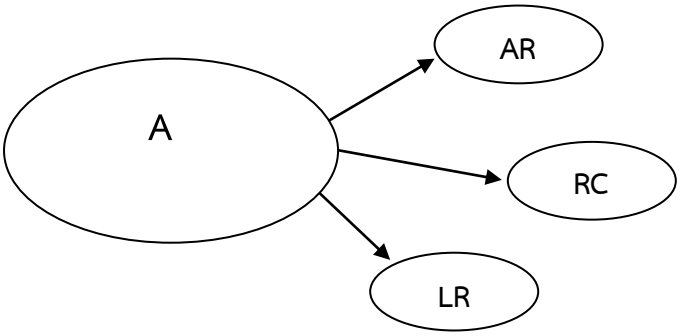


ภาพที่ 2.1 การรวมกันของข้อสอบทั้งหมดที่เป็นไปได้และการรวมกันของข้อสอบทั้งหมดจากแบบสอบที่ตรงตามเงื่อนไขบังคับ

จากภาพที่ 2.1 การค้นหาอย่างสุ่มที่แท้จริง (pure random search) อาศัยการแจกแจงแบบยูนิฟอร์มและลู่เข้า (Converge) ไปยังแบบสอบด้วยความน่าจะเป็นเท่ากับ $P = \frac{|B|}{|A|}$ ดังนั้น ถ้าย่อขนาดของเซต A โดยปราศจากการสูญเสียการรวมกันของข้อสอบ (losing combinations) จากเซต B แล้วความน่าจะเป็น (P) จะเพิ่มขึ้น และผลที่ตามมาคือ ความเร็วของการรวมแบบสอบโดยใช้วิธี มอนติ คาร์โล จะเพิ่มขึ้น

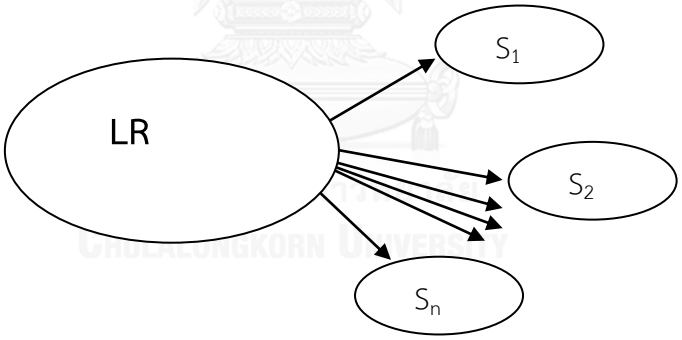
เนื่องจากแบบสอบ ประกอบด้วย หลายตอน (Sections) แต่ละตอน ประกอบด้วย หลายบทความ (Passage) แต่ละบทความนำไปสู่กลุ่มของข้อสอบหรือแบบทดสอบย่อย (testlet) ซึ่งข้อสอบเหล่านี้สามารถถูกจัดเป็นกลุ่มได้โดยการใช้โครงสร้างลำดับลดหลั่น (hierarchical structure) ของแบบสอบ (test form) ที่สอดคล้องกับเงื่อนไขบังคับของแบบสอบ (test constraints) และหลักของการแบ่งเพื่อเอาชนะ (divide and conquer) ซึ่งวิธีการดังกล่าวสามารถลดขนาดของขอบเขตการค้นหา (search region) ได้อย่างมีนัยสำคัญ

ในแผนภาพที่ 2.2 และ 2.3 เป็นการแสดงตัวอย่างการลดขนาดของขอบเขตการค้นหา (search region) ไปสู่ขอบเขตการค้นหาย่อยภายใต้เงื่อนไขบังคับ 2 ประเภท คือ ตอน (Section) และทักษะทางความคิด (cognitive skills) ภายในเนื้อหาที่เกี่ยวกับการให้เหตุผลเชิงตรรกะ (Logical Reasoning: LR)



ภาพที่ 2.2 การลดขอบเขตการค้นหา A เข้าไปในขอบเขตการค้นหาย่อยตามประเภทของข้อสอบ

จากแผนภาพ 2.2 แสดง การลดขนาดขอบเขตการค้นหา A เข้าไปในขอบเขตการค้นหาย่อยแต่ละประเภทของข้อสอบซึ่งแบ่งตามระดับของทักษะทางความคิด ได้แก่ การให้เหตุผลเชิงวิเคราะห์ (Analytical Reasoning: AR) ความเข้าใจในการอ่าน (Reading Comprehension: RC) และการให้เหตุผลเชิงผลเชิงตรรกะ (Logical Reasoning: LR) ตามลำดับ และขอบเขตการค้นหาย่อย การให้เหตุผลเชิงวิเคราะห์ (AR) ประกอบด้วย การรวมกันทั้งหมดของข้อสอบความเข้าใจในการอ่าน (RC) และข้อสอบการให้เชิงผลเชิงตรรกะ (LR) เมื่อแต่ละถูกกำหนดจำนวนข้อสอบที่แน่นอน โดยจำนวนของข้อสอบในขอบเขตการค้นหา A จะเป็น |A| >> |AR| + |RC| + |LR|

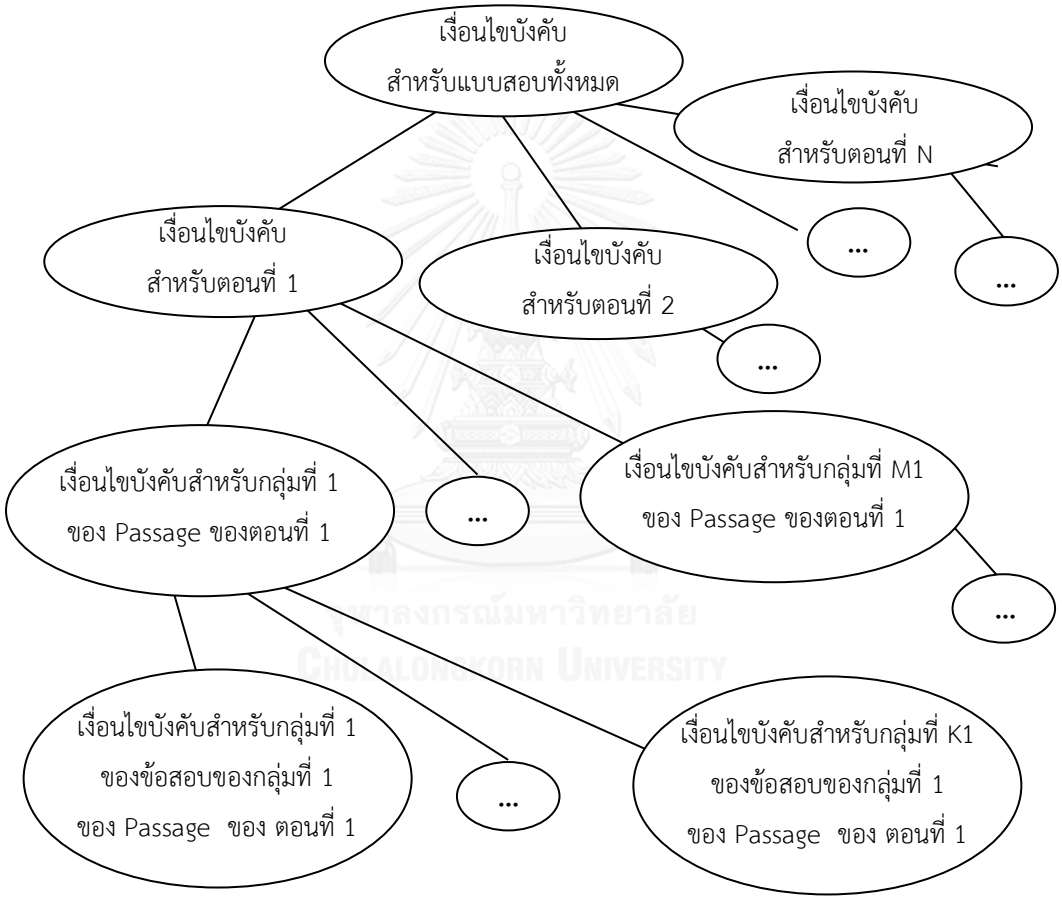


ภาพที่ 2.3 การลดขอบเขตการค้นหา LR เข้าไปในขอบเขตการค้นหาย่อยสำหรับกลุ่มของข้อสอบตามเงื่อนไขบังคับที่สอดคล้องกับเงื่อนไขบังคับทักษะทางความคิด (cognitive skills)

จากแผนภาพ 2.3 ขอบเขตการค้นหาย่อย S_i มีการรวมกันทั้งหมด (all combination) ของข้อสอบการให้เหตุผลเชิงตรรกะ (logical reasoning items) สำหรับทักษะทางความคิดที่ i เมื่อข้อสอบแต่ละชุดรวมกันจำนวนขอบเขตต่ำสุดของข้อสอบทั้งหมดภายในตอน LR จะหาได้จาก

$$|\text{LR}| >> \sum_{i-1}^{n} |S_i|$$

การกำหนดรูปแบบของต้นไม้ใน แผนภาพ 5 เมื่อแต่ละบัพ (Node) มีเงื่อนไขบังคับสำหรับส่วนต่างๆ ของแบบสอบโดยเฉพาะ โดยส่วนที่เป็นราก (Root) จะมีเงื่อนไขบังคับสำหรับแบบสอบทั้งหมด เช่น คะแนนของแบบสอบที่คาดหวัง (expected test score) จำนวนข้อที่ยอมให้มีในแบบสอบ (allowed number of items) ซึ่งส่วนที่ไม่ใช่บัพสุดท้าย (Terminal node) จะเป็นบัพที่มีเงื่อนไขบังคับสำหรับส่วน (parts) หรือ ตอน (section) ต่างๆ ในแบบสอบ ดังนั้นแต่ละเงื่อนไขบังคับจะเป็นของบัพใดบัพหนึ่งเพียงบัพเดียว และผลของการเชื่อมกันของชุดข้อสอบที่ถูกรวมในบัพสุดท้ายจะเป็นแบบสอบทั้งฉบับ



ภาพที่ 2.4 ชุดของข้อสอบทั้งหมดที่เป็นไปได้และชุดของข้อสอบทั้งหมดจากแบบสอบที่เป็นไปได้

จากแผนภาพ 2.4 การรวมแบบสอบเริ่มต้นจากบัพสุดท้าย โดยที่ข้อสอบที่สอดคล้องกับแต่ละบัพสุดท้ายจะถูกเลือก ในส่วนที่ไม่ใช่บัพสุดท้ายชิ้นส่วนจากบัพลูก (Children node) ถูกเชื่อมต่อเข้าด้วยกันแล้วถูกตรวจสอบโดยการเปรียบเทียบกับเงื่อนไขบังคับภายในบัพ ผลลัพธ์ที่ได้จะบอกว่าสำเร็จหรือล้มเหลว ในกรณีที่สำเร็จ ผลการเชื่อมต่อจะถูกส่งขึ้นไปยังบัพพ่อแม่ (Parent node) ที่สูงขึ้นไป (going up) แต่ในกรณีที่ล้มเหลว นั้นหมายความว่า ในบัพลูกจะมีการฝ่าฝืนเงื่อนไขบังคับ

(violating constraints) ดังนั้นจะมีการบังคับให้รวมส่วนของบัพนี้ใหม่โดยการส่งไปยังบัพที่ต่ำกว่า (going down) กระบวนการจะถูกทำซ้ำทั้งขึ้นและลงจนกระทั่งให้ผลสำเร็จที่บัพราก จึงจะได้ผลลัพธ์เป็นแบบสอบที่ตรงตามเงื่อนไขบังคับ

2. การค้นหาแบบตาบู (Tabu Search: TS)

การค้นหาแบบตาบู เป็นขั้นตอนวิธีที่ถูกนำมาประยุกต์ใช้ในแก้ปัญหาระบบที่เรียกว่าปัญหาการหาค่าเหมาะที่สุดเชิงผสมผสาน (combinatorial optimization) ได้อย่างมีประสิทธิภาพ โดย Glover เป็นผู้ริเริ่มเสนอแนวคิดวิธีการค้นหาแบบตาบูไว้เมื่อปี ค.ศ. 1977 ซึ่งได้อธิบายรายละเอียดไว้ใน Bland & Dawson (1991) การค้นหาแบบตาบูเป็นที่นิยมอย่างแพร่หลาย เนื่องจากวิธีการดังกล่าวสามารถหลุดพ้นจากคำตอบเหมาะที่สุดแบบวงแคบเฉพาะถิ่น (local optimum avoidance) และหลีกเลี่ยงเส้นทางการค้นหาคำตอบที่ทำให้เกิดการวนรอบอยู่กับที่ (cycle avoidance)

ด้วยความสามารถใหม่ที่กล่าวมาข้างต้น การค้นหาแบบตาบูได้มีการใช้ข้อมูลของการค้นหาคำตอบในอดีตมาช่วย ตัดสินการเดินว่าควรจะไปในทิศทางใด องค์ประกอบใหม่ซึ่งเป็นส่วนสำคัญในโครงสร้างของการค้นหาแบบตาบูที่ทำให้การค้นหา คำตอบมีประสิทธิภาพสูงขึ้นนี้ได้แก่

1) เงื่อนไขของความคงอยู่ล่าสุด (Recency condition) การใช้เงื่อนไขของความคงอยู่เป็นการติดตามการค้นหาคำตอบในช่วงเวลาที่ผ่านมา เมื่อคำตอบหนึ่งถูก ค้นพบแล้ว การเดินที่นำไปสู่คำตอบนั้นจะถูกตั้งเป็นสถานะต้องห้าม คำตอบที่ถูกค้นพบด้วยการเดินจะถือเป็นคำตอบล่าสุดและจะไม่ถูกค้นอีกภายในระยะเวลาหนึ่ง (เนื่องจากการเดินที่นำไปสู่คำตอบนี้ถูกห้ามใช้อีก ในขณะที่ยังมีสถานะต้องห้ามอยู่) หลังจากระยะเวลาที่กำหนดผ่านไป สถานการณ์เดินนี้ก็จะถูกตั้งค่ากลับ สู่สภาวะปกติ ดังนั้นแล้วภายหลังการเดินไปยังคำตอบหนึ่งๆ การค้นหาแบบตาบูจะบังคับให้การค้นหาคำตอบทำการเดินไป ยังคำตอบใหม่ โดยที่คำตอบเดิมจะไม่ถูกค้นอีก กลไกนี้ทำให้การค้นหาแบบตาบูสามารถหลุดออกจากคำตอบเหมาะที่สุดแบบวงแคบเฉพาะถิ่น และทำการค้นหาคำตอบที่ดีขึ้นไปเรื่อยๆ ได้ (ถึงแม้ว่าในบางครั้ง คำตอบใหม่ไม่ดีไปกว่าคำตอบที่มีอยู่ก็ตาม)

2) เงื่อนไขของความซ้ำซาก (frequency condition) ในลักษณะเดียวกันเราสามารถบันทึกจำนวนครั้งที่การเดินหนึ่งๆ ถูกเรียกใช้ได้ การค้นหาแบบตาบูถือว่าถ้ารูปแบบการเดินใดถูกเรียกใช้เป็นจำนวนมากครั้งเกินไป (เกินจำนวนที่ตั้งเอาไว้) การเดินนั้นควรจะถูกต้องห้ามหรือถูกตั้ง เป็นสถานะต้องห้าม เพื่อหลีกเลี่ยงเส้นทางการค้นหาคำตอบที่ทำให้เกิดการวนรอบอยู่กับที่ ทำให้สามารถหลุดพ้นจากคำตอบเหมาะที่สุดแบบวงแคบเฉพาะถิ่นได้ เงื่อนไขทั้งสองจะถูกใช้ร่วมกันเสมอ เนื่องมาจากเงื่อนไขเพียงอย่างใดอย่างหนึ่งไม่เพียงพอ เราสามารถกล่าวได้ ว่าเงื่อนไขทั้งสองเป็นส่วน

เติมเต็ม หรือ complimentary ซึ่งกันและกัน กล่าวคือการเดินใดถูกตั้งค่าสถานะใหม่ ต้องห้ามด้วยเงื่อนไขของความซ้ำซาก และได้คงสถานะต้องห้ามนานเกินระยะเวลาที่กำหนดไว้ การเดินนั้นจะสามารถถูกตั้งค่ากลับสู่สถานะปกติได้ด้วยเงื่อนไขของความคงอยู่ล่าสุด การค้นหาแบบตาบูยังมีอีก 2 องค์ประกอบอื่นๆอันเป็นกลไกสำคัญที่ทำให้การค้นหาคำตอบมีประสิทธิภาพเพิ่มมากขึ้น ได้แก่ กลไกการเน้น (intensification) และ กลไกการแปรเปลี่ยน (diversification)

กลไกการเน้น คือการค้นหาคำตอบที่เน้นไปยังกลุ่มคำตอบที่ได้ค้นพบแล้วว่าเป็นคำตอบที่ดี โดยใช้ข้อมูล ที่ได้บันทึกจากการค้นหาคำตอบที่ผ่านมาในอดีต กลยุทธ์นี้ทำให้การค้นหาแบบตาบูกลับไปค้นหาคำตอบในย่านที่เคยเจอ คำตอบที่ดี และทำการค้นหาในย่านนั้นอย่างละเอียดขึ้น ส่วนประกอบของคำตอบที่น่าจะเป็นประโยชน์หรือ เส้นทางที่นำไปสู่คำตอบนั้น จะถูกใช้เป็นข้อมูลโดยกลไกการเน้นในการค้นหาคำตอบใหม่ได้

กลไกการแปรเปลี่ยน ในทางตรงกันข้ามเป็นกลยุทธ์ที่ส่งเสริมให้ การค้นหาแบบตาบูไปทำการสำรวจย่านที่ยังไม่เคยถูก สำรวจมาก่อน ซึ่งอาจจะทำให้ได้คำตอบที่มีความแตกต่างไปจากกลุ่มคำตอบที่ได้ถูกสำรวจมาก่อนที่จะทำ กลยุทธ์นี้ ในบางครั้งการเลือกเส้นทางอื่นที่ยังไม่เคยสำรวจและแตกต่างไปจากแนวทางของเส้นทางเดิม อาจจะทำให้มีโอกาสเจอคำตอบที่ดีกว่าได้เช่นกัน

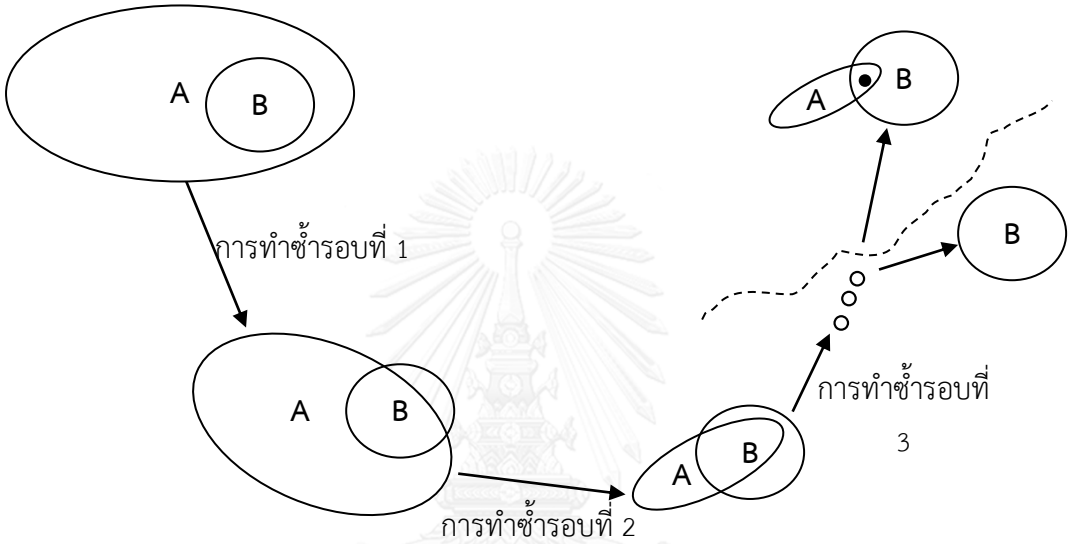
นอกเหนือจากองค์ประกอบต่างๆที่ได้กล่าวมาแล้วข้างต้น การค้นหาแบบตาบูยังมีองค์ประกอบ อื่นที่เป็นส่วนสำคัญในการออกแบบการค้นหาแบบตาบูนั้นก็คือ เกณฑ์ความทะเยอทะยานหรือเกณฑ์เชิงละโมบ ซึ่งเป็นเงื่อนไขที่สามารถทำให้เกิดการเดินไปในทิศทางที่ต้องการ ถึงแม้ว่าการเดินนั้นจะมีสถานะ ต้องห้าม โดยที่การเดินดังกล่าวจะได้รับอนุญาตก็ต่อเมื่อคำตอบที่ได้ดีกว่าทุกคำตอบที่เคยค้นพบมา การค้นหาแบบตาบูจะต้องทำการเก็บบันทึกผลการค้นหาคำตอบที่ดีที่สุดเอาไว้เพื่อใช้ในการตรวจสอบเงื่อนไข

ขั้นตอนวิธีเชิงละโมบ (Greedy algorithm) เป็นขั้นตอนวิธีการแก้ปัญหาที่คิดแบบง่าย ๆ และตรงไปตรงมา มีหลักการว่า ในขณะใดๆ ก็ตาม เราจะหาคำตอบที่ดีที่สุดของปัญหา โดยการเลือกเอาคำตอบที่ดีที่สุด ในสถานะนั้นออกมาก่อน โดยพิจารณาว่าข้อมูลที่มีอยู่ในขณะนั้นมีทางเลือกใดที่ให้คำตอบที่ดีที่สุด ขั้นตอนวิธีจะหาทางเลือกที่ดูดีที่สุดในขณะนั้นซึ่งถ้ามีข้อมูลเพียงพอจะทำให้สามารถสรุปคำตอบที่ดีที่สุดได้

ดังนั้นจึงสามารถสรุปได้ว่า องค์ประกอบของวิธีการค้นหาแบบตาบูที่แตกต่างจากวิธีการค้นหาแบบอื่นๆ คือ มีเกณฑ์ความเป็นตาบู (tabu list criteria) และ มีเกณฑ์ความปรารถนา (aspiration criteria) ซึ่ง “เกณฑ์ความเป็นตาบู” เป็นส่วนที่คอยเก็บข้อมูลของคำตอบในอดีตของกระบวนการค้นหานั้นๆ เพื่อเป็นตัวกำหนดการค้นหาคำตอบว่าจะมีทิศทางไปทางใด หลักการ

ออกแบบเกณฑ์ความเป็นตาบู จะมีลักษณะแตกต่างกันออกไป ขึ้นอยู่กับปัญหาแต่ละประเภท และ “เกณฑ์ความปรารถนา” เป็นเงื่อนไขที่จะใช้ในบางครั้งที่จำเป็นจะต้องเลือกคำตอบที่อยู่ในเกณฑ์ความเป็นตาบู งานบางชนิดที่ปัญหาไม่ซับซ้อนไม่จำเป็นต้องพึ่งส่วนนี้ก็ได้ เกณฑ์ความเป็นตาบูก็เพียงพอที่จะค้นหาคำตอบที่ดีที่สุดได้

จากความสามารถดังกล่าวการค้นหาตาบูและขั้นตอนวิธีเชิงละโมบจึงถูกนำมาใช้ในการลดขนาดขอบเขตการค้นหา ซึ่งแสดงให้เห็นใน แผนภาพ 2.5



ภาพที่ 2.5 การใช้ขั้นตอนวิธีเชิงละโมบ (Greedy algorithm) ในวิธีการค้นหาตาบูเพื่อย่อขอบเขตการค้นหา A

จากแผนภาพ 2.5 แสดงให้เห็นถึงการใช้ขั้นตอนวิธีเชิงละโมบและการค้นหาแบบตาบูในการย่อขอบเขตการค้นหา A เพื่อหาข้อสอบที่รวมกันอยู่ในเซต B จนกระทั่งขนาดของเซต A ถูกลดลงเรื่อยๆ จนหมดหรือว่างเปล่า ซึ่งมีเกณฑ์ในการพิจารณาดังนี้ ถ้าข้อสอบที่รวมกันเชิงสุ่มไม่ตรงตามเงื่อนไขบังคับผลการรวมกันนั้นจะถูกย้ายจากคลังข้อสอบยังขอบเขตตาบูหรือตาบูลิสต์ ในทางตรงกันข้ามถ้าพบข้อสอบที่ตรงตามเงื่อนไขบังคับ ข้อสอบนั้นจะถูกย้ายไปยังเซตของการเดินที่อนุญาต (ไม่เป็นตาบู) เมื่อพบข้อสอบที่ตรงตามเงื่อนไขบังคับหรือข้อสอบในคลังข้อสอบถูกเลือกใช้จนหมดข้อสอบจากตาบูลิสต์จะถูกนำกลับไปยังคลังข้อสอบ

3. การจัดลำดับความสำคัญของเงื่อนไขบังคับ (prioritization of the test constraints)

การจัดลำดับความสำคัญของเงื่อนไขบังคับที่นำมาใช้ในการลดขนาดขอบเขตการค้นหาอาศัยหลักการแจกนับเงื่อนไขบังคับ โดยมีข้อตอลงเบื้องต้นว่า $R = \{r_1, r_2, r_3, ..., r_m\}$ โดยที่ r_j

แต่ละตัวเป็นสมาชิกของจำนวนจริง และ r_j มีพิสัยเท่ากับขอบเขตบนและขอบเขตล่าง $[L_j, U_j]$ แล้วนำไปสู่การใช้ฟังก์ชันการแจงนับ Enumeration(R,s) เพื่อสร้างเวคเตอร์ E ของลำดับ E_i โดยที่ $E_i = \{e_1, e_2, e_3, ..., e_m\}$

จากฟังก์ชันการแจงนับ Enumeration(R,s) จะเห็นว่า R คือ เซตของคู่ลำดับที่แสดงขอบเขตบนและขอบเขตล่างของจำนวนข้อสอบในแต่ละตอนที่มีในแบบสอบ ตัวอย่างเช่น กำหนดเงื่อนไขบังคับให้เป็นดังนี้

แบบสอบมีจำนวนข้อสอบ 10 ข้อ NTest = 10

ในแบบสอบประกอบด้วย 2 ตอน M = 2

จำนวนของข้อสอบที่ยอมให้มีในตอนที่ 1 r_1 = [4,8]

จำนวนของข้อสอบที่ยอมให้มีในตอนที่ 2 r_2 = [3,7]

การคำนวณฟังก์ชัน E จะพิจารณาจากการแจงนับคู่ลำดับ r_1 และ r_2 ซึ่งผลรวมของคู่ลำดับจะต้องมีค่าเท่ากับความยาวของแบบสอบและแต่ละองค์ประกอบของ E จะต้องมีขอบเขตบนและขอบเขตล่างตามที่กำหนดไว้ในเงื่อนไขบังคับ

Enumeration(R,s) = Enumeration$(\{r1, r2\}, NTest)$

= E$(\{4,6\}, \{5,5\}, \{6,4\}, \{7,3\})$

จากนั้นจึงเลือกองค์ประกอบ $\{e_1, e_2\}$ อย่างสุ่มจาก E ซึ่งเป็นการลดขอบเขตค้นหา ผลจากการสุ่มจะทำให้แบบสอบในตอนที่หนึ่งมีจำนวนข้อเป็น e_1 ข้อ และแบบสอบตอนที่สองมีจำนวนข้อเป็น e_2 ข้อ และแบบสอบทั้งสองตอนตรงตามเงื่อนไขบังคับ

ขั้นตอนวิธีของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ (Algorithm of Monte Carlo CAT)

เมื่อพิจารณาในมุมมองของขั้นตอนวิธี (Algorithm) จะทำให้ทราบว่าการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล (Monte Carlo CAT) มีปัจจัยนำเข้า (Input) ประกอบด้วย 1) ค่าประมาณความสามารถปัจจุบันของผู้สอบ 2) เซตของแบบสอบ (เซต S) ประกอบด้วย แบบสอบจำนวน r แบบสอบที่มีคุณสมบัติตรงตามเงื่อนไขบังคับ (feasible test form) 3) เซตของข้อสอบ ประกอบด้วย ข้อสอบจำนวน l ข้อ ที่ได้จากการสร้างลำดับเชิงสุ่มของข้อสอบด้วยวิธีมอนติ คาร์โล โดยจำนวนข้อสอบในเซตจะมากกว่าหรือเท่ากับศูนย์และน้อยกว่าความยาวของแบบสอบทั้งฉบับ $(0 \le l < n)$ 4) พารามิเตอร์ m ที่ใช้กำหนดจำนวนเซตของแบบสอบ โดยที่

$m = \max(\frac{k}{n-1}, 1)$ และ พารามิเตอร์ k ที่ใช้ขนาดของลำดับเชิงสุ่มของแบบสอบ ที่พร้อมจะนำไปใช้กับผู้สอบ โดยที่ $k = 2l + 1$ และผลลัพธ์ (Output) ของขั้นตอนวิธี คือ ข้อสอบข้อถัดไป (φ) ที่จะนำไปใช้กับผู้สอบ และเซต S ที่ได้รับการปรับให้เป็นปัจจุบัน (update)

รายละเอียดของขั้นตอนวิธี Monte Carlo CAT มีดังนี้

ขั้นที่ 1 รวมแบบข้อสอบเป็นแบบสอบจำนวน (m-r) แบบสอบ อย่างเท่าเทียมกัน (Uniform) และเพิ่มแบบสอบชุดนี้ไปยัง เซต S เมื่อ แต่ละแบบสอบ (Form) ตอบสนองต่อทุกข้อจำกัดของเนื้อหา (Content Constraints) และมีข้อสอบจำนวน (l) ข้อ ซึ่งพร้อมที่จะนำไปใช้กับผู้สอบ; โดยเป็นการรวมข้อสอบเป็นแบบสอบอย่างเท่าเทียมกัน (Uniform assembly)

ขั้นที่ 2 แต่ละแบบสอบในเซต S มีจำนวนข้อที่ยังไม่ได้นำไปใช้กับผู้สอบ (n-l) ข้อ จากนั้นเพิ่มข้อสอบเหล่านี้ไปในลำดับ (sequence)

ผลลัพธ์ของลำดับจะมีจำนวนข้อเท่ากับ m(n-l) ข้อ ; แบบสอบในเซต S สามารถทับซ้อนกันกับแบบสอบอื่น ดั้งนั้น ข้อที่ซ้ำกันสามารถพบในลำดับ (sequence) ได้หลายครั้ง

ขั้นที่ 3 สุ่มตัวอย่าง k ข้อ จากลำดับ (sequence)

ขั้นที่ 4 ในกลุ่มตัวอย่างจากการสุ่มในขั้นตอนที่ 3 ค้นหาข้อสอบที่ให้สารสนเทศสูงสุดที่ระดับค่าประมาณความสามารถของผู้สอบในกลุ่มตัวอย่างที่สุ่มมา k ข้อ และเป็นข้อสอบข้อถัดไป (φ) ที่จะนำไปใช้กับผู้สอบ

ขั้นที่ 5 ลบแบบสอบ (Form) ทั้งหมดที่ไม่มี φ ออกจาก เซต *S*

ในช่วงเริ่มต้น สำหรับผู้สอบแต่ละคนจะมี $(l = 0)$ และ เซตของแบบสอบ (เซต *S*) เป็นเซตว่าง ขั้นตอนทั้งหมดด้านบนจะถูกทำซ้ำโดย (l) จะถูกเพิ่มขึ้น 1 ข้อ เมื่อทำซ้ำ 1 ครั้ง จนถึงจำนวน n-1 ครั้ง การทดสอบจะดำเนินการเสร็จสมบูรณ์ และพารามิเตอร์ *k* คือ พารามิเตอร์สำคัญของขั้นตอนวิธีนี้ ถ้าพารามิเตอร์ *k* มีค่ามาก การใช้ข้อสอบซ้ำจะสูงและความคลาดเคลื่อนจากการวัดจะมีค่าต่ำ เนื่องจาก มีข้อสอบจำนวนมากกว่าในกลุ่มตัวอย่างในขั้นตอนที่ 4 ถ้าพารามิเตอร์ *k* มีค่าน้อย การใช้ข้อสอบซ้ำจะต่ำและความคลาดเคลื่อนจากการวัดจะมีค่าสูง เนื่องจาก มีข้อสอบจำนวนน้อยกว่าในกลุ่มตัวอย่างในขั้นตอนที่ 4

ตอนที่ 4 วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ

Cheng และ Chang (2009) เสนอวิธีดัชนีลำดับความสำคัญสูงสุด (Maximum Priority Index: MPI) สำหรับการคัดเลือกข้อสอบที่มีการควบคุมเงื่อนไขบังคับอย่างเข้มงวดในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ ซึ่งนำเทคนิคการคัดเลือกข้อสอบแบบสองขั้น (two-phase item selection) ที่เสนอโดย Cheng และคณะ (2007) มาประยุกต์ใช้ร่วมด้วยเพื่อทำให้วิธี MIP สามารถจัดการกับเงื่อนไขบังคับได้อย่างยืดหยุ่นโดยสามารถกำหนดให้เลือกข้อสอบที่ตรงตามเงื่อนไขบังคับภายในช่วงขอบเขตบนและขอบเขตล่างแทนที่การกำหนดเป็นค่าคงที่ และนำมาเปรียบเทียบกับวิธี Weighted Deviation Modeling (WDM) ซึ่งเป็นวิธีที่มีรูปแบบการคัดเลือกข้อสอบอย่างเป็นลำดับเหมือนกัน ผลการศึกษาพบว่า วิธี MPI มีประสิทธิภาพดีกว่า WDM อย่างเห็นได้ชัดในการจัดการเงื่อนไขข้อบังคับ อีกทั้งยังมีความถูกต้องแม่นยำของการประมาณค่าความสามารถของผู้สอบสูง และสามารถควบคุมอัตราการใช้ข้อสอบซ้ำให้อยู่ในระดับที่กำหนดไว้ได้ แต่มีสัดส่วนของข้อสอบที่ไม่ถูกนำมาใช้สูงถึงร้อยละ 50 ทำให้วิธี MPI มีจุดอ่อนในเรื่องการขาดความสมดุลในการใช้คลังข้อสอบ

Cheng และ Chang จึงให้ข้อเสนอแนะว่าควรนำมาใช้ร่วมกับวิธีการแบ่งคลังข้อสอบออกเป็นชั้นๆ ตามค่าอำนาจจำแนกที่พัฒนาโดย Chang และ Ying (1999) เพื่อเพิ่มประสิทธิภาพของการใช้คลังข้อสอบ หลังจากนั้น Cheng และคณะ (2009) ศึกษาถึงประโยชน์ของวิธีการแบ่งคลังข้อสอบออกเป็นชั้นๆ ตามค่าอำนาจจำแนก โดยนำมาใช้ร่วมกับวิธี MPI เพื่อให้เกิดความยืดหยุ่นในการควบคุมเงื่อนไขบังคับที่ไม่ใช่ทางสถิติและสามารถใช้คลังข้อสอบได้อย่างมีประสิทธิภาพ ผลจากการศึกษาในสถานการณ์จำลอง พบว่า วิธีแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (constraint-weighted a-stratification method) สามารถสร้างความสมดุลในการใช้คลังข้อสอบโดยสูญเสียความถูกต้องแม่นยำของการประมาณค่าความสามารถของผู้สอบน้อยที่สุด รวมทั้งยังควบคุมเงื่อนไขบังคับได้อย่างดีเยี่ยม โดยไม่มีการฝ่าฝืนเงื่อนไขบังคับเลย

วิธีดัชนีลำดับความสำคัญจะพิจารณาความแปรปรวนของค่าสารสนเทศโดยการนำดัชนีไปเป็นตัวคูณให้กับค่าสารสนเทศของข้อสอบ ดังนั้นในการคัดเลือกข้อสอบแทนที่จะพิจารณาจากค่าสารสนเทศเพียงอย่างเดียววิธีการนี้จะนำลำดับความสำคัญซึ่งกำหนดโดยผู้เชี่ยวชาญที่พิจารณาร่วมด้วย การคัดเลือกข้อสอบจะพิจารณาจากผลคูณของค่าสารสนเทศของข้อสอบกับดัชนีลำดับความสำคัญ ข้อใดให้ผลคูณมากที่สุด ข้อนั้นจะได้รับการคัดเลือกให้นำไปใช้กับผู้สอบ ดังนั้น วิธี PI จึงเกี่ยวข้องกับเมตริกซ์เงื่อนไขบังคับ ซึ่งจะบอกว่าข้อสอบแต่ละข้อประกอบด้วยเงื่อนไขบังคับใดบ้าง

โดยกำหนดให้เมตริกซ์ดังกล่าว คือ เมตริกซ์ C.C เมื่อ c_{jk}= 1 แสดงว่า ข้อสอบข้อที่ j เกี่ยวข้องกับเงื่อนไขบังคับที่ K

เมื่อ C.C คือ เมตริกซ์ ที่เกี่ยวข้องกับข้อบังคับ และมีขนาด J x K

j คือ จำนวนข้อสอบในคลังข้อสอบ

K คือ จำนวนเงื่อนไขบังคับทั้งหมด

k คือ ข้อบังคับเนื้อหาจะเกี่ยวข้องกับค่าน้ำหนัก (weight) w_k

c_j = 1 บ่งชี้ว่า ข้อบังคับ k เกี่ยวข้องกับข้อสอบที่ j

c_j = 0 บ่งชี้ว่า ข้อบังคับ k ไม่เกี่ยวข้องกับข้อสอบที่ j

โดยทั่วไป Matrix C.C จะถูกระบุก่อนการเลือกข้อสอบโดยผู้เชี่ยวชาญด้านเนื้อหาหรือผู้เชี่ยวชาญทางด้านจิตมิติ โดยที่แต่ละข้อบังคับ k จะเกี่ยวข้องกับค่าน้ำหนัก (weight) w_k ซึ่งการทดสอบส่วนใหญ่จะใส่ค่าน้ำหนักขนาดใหญ่ให้กับข้อบังคับที่สำคัญและค่าน้ำหนักที่น้อยกว่ากับข้อบังคับอื่นๆ ดังนั้น Priority Index ของข้อสอบข้อที่ j สามารถคำนวณได้ดังนี้

$$PI_j = I_j \prod_{k=1}^{K} (w_k f_k)^{c_{jk}} \quad \text{.....(1)}$$

เมื่อ PI แทน Fisher information ของข้อสอบข้อที่ J ที่ถูกประมาณค่าที่ระดับความสามารถปัจจุบัน

X_k แทน ข้อสอบจากขอบเขตเนื้อหาที่แน่นอน

x_k แทน ข้อสอบที่จะถูกเลือก

f_{jk} แทน โควตาที่มีอยู่ของข้อบังคับ k

และผลของโควตาที่เหลืออยู่จะคำนวณได้จาก

$$f_k = \frac{(X_k - x_k)}{X_k} \quad \text{.....(2)}$$

เมื่อ C_{jk} = 0 หมายความว่า ข้อสอบข้อที่ j ไม่ถูกควบคุมด้วยข้อบังคับ k

สมมติว่า ต้องการข้อบังคับ k' ดังนั้นอัตราการแสดงของแต่ละข้อเป็นค่าที่ต่ำกว่าหรือเท่ากับ r และระหว่างผู้สอบทั้งหมด N คนผู้ซึ่งทำการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ มีผู้สอบ n คนที่ได้เห็นข้อสอบ j แล้ว $f_{jk'}$ สามารถคำนวณได้จาก

$$f_{k'} = \frac{(r-(n/N))}{r} \quad \text{.....(3)}$$

จากสมการที่ 3 สามารถคำนวณดัชนีลำดับความสำคัญ (Maximum Priority Index: MPI) สำหรับข้อสอบทุกข้อที่อยู่ในคลังข้อสอบ จากนั้นจึงเลือกใช้ข้อสอบที่มีค่าดัชนีลำดับความสำคัญสูงสุดแทนที่การข้อสอบข้อที่มีค่าสารสนเทศของฟิชเชอร์สูงสุด (Maximum Fisher information)

วิธี MPI ในรูปแบบปัจจุบันถูกจำกัดข้อบังคับในรูปของขอบเขตบน อย่างไรก็ตาม โปรแกรมการทดสอบมักจะมีข้อบังคับที่เกี่ยวข้องกับขอบเขตล่าง ดังตัวอย่าง ผังข้อสอบสำหรับแบบสอบคณิตศาสตร์อาจจะระบุไว้เพียงไม่เกิน 15 ข้อ (upper bound) แต่ก็ไม่น้อยกว่า 10 ข้อ (lower bound) ที่เป็นเนื้อหาทางด้านพีชคณิต

ดังนั้นข้อบังคับจะเรียกว่า ข้อบังคับความสมดุลทางเนื้อหาที่ยืดหยุ่น (Flexible content balancing constraint) (Cheng, Chang, & Yi, 2007) เมื่อข้อบังคับความสมดุลทางเนื้อหาที่ยืดหยุ่นได้ถูกนำเสนอวิธี MPI ข้างต้นจึงได้รับการปรับปรุงแก้ไขโดยใช้ร่วมกันกับวิธี two-phase item selection strategy (Cheng et al., 2007) เพื่อจัดการกับข้อบังคับความสมดุลทางเนื้อหาที่ยืดหยุ่น

กรอบแนวคิดการใช้ MPI ร่วมกับ two-phase item selection

สำหรับแต่ละข้อบังคับความสมดุลของเนื้อหาที่ยืดหยุ่นซึ่งเกี่ยวข้องกับ Lower bound และ Upper bound สามารถอธิบายได้ดังต่อไปนี้

$l_k \leq \mu_k \leq u_k$

และ

$$\sum_{k=1}^{K} = \mu_k = L \quad \text{.....(4)}$$

เมื่อ μ_k แทน จำนวนของข้อสอบที่ถูกเลือกจากขอบเขตเนื้อหา k

l_k แทน ขอบเขตล่างของข้อบังคับเนื้อหา (Lower bound)

u_k แทน ขอบเขตบนของข้อบังคับเนื้อหา (Upper bound)

K แทน จำนวนของขอบเขตเนื้อหาทั้งหมด

L แทน ความยาวของแบบสอบ

แนวคิดของการคัดเลือกข้อสอบแบบ two-phase เป็นการจัดการกับ lower bounds ในช่วง ระยะที่ 1 และจัดการกับ upper bounds ระยะที่ 2 ดังนั้นในระยะที่ 1 ของการคัดเลือกจะเกี่ยวข้องกับข้อสอบจำนวน L_1 เมื่อ $L_1 = \sum_{k=1}^{K} l_k$ และในระยะที่ 2 ของการคัดเลือกเมื่อ $L_2 = L - L_1$ (ในระยะที่ 1 lower bounds ถูกทำให้เป็น upper bounds)

ในทั้ง 2 ระยะ การคำนวณ Priority index ของแต่ละข้อยังใช้สูตรการคำนวณตามสมการที่ 1 อย่างไรก็ตาม f_k ถูกคำนวณอย่างแตกต่างจากกัน ซึ่งใน ระยะที่ 1 สามารถคำนวณได้จากสมการดังต่อไปนี้

$$f_k = \frac{(l_k - x_k)}{l_k} \quad \text{.....(5)}$$

สมการที่ 5 แตกต่างจากสมการที่ 2 เพียงการแทนตำแหน่งของ X_k ด้วย l_k และเมื่อข้อบังคับ k ไปถึงขอบเขตล่างของมัน แล้ว f_k เป็น 0 และ priority index ที่เกี่ยวข้องเปลี่ยนไปเป็น 0 ดังนั้น priority index จึงได้รับการจัดระดับความสำคัญต่ำกว่าข้อสอบข้ออื่นๆ ในคลังข้อสอบซึ่งมีค่าดัชนีความสำคัญเป็นบวก ตัวอย่าง เช่น สมมติให้แบบสอบมีเพียง 2 ข้อบังคับ และข้อบังคับแรกไปถึงขอบเขตล่างเรียบร้อยแล้วขอบเขตเนื้อหานั้นจะหยุดนิ่ง และไม่มีข้อสอบที่จะสามารถถูกเลือกจากขอบเขตเนื้อหานั้นจนกระทั่งขอบเขตเนื้อหาอื่นมาถึง เนื่องจาก ขอบเขตล่างทั้งหมดจะพบกันที่จุดสุดท้ายของระยะที่ 1

ในระยะที่ *2* f_k สามารถคำนวณได้จากสมการดังต่อไปนี้

$$f_k = \frac{(u_k - x_k)}{u_k} \quad \text{.....(6)}$$

เมื่อข้อบังคับ k มาถึงขอบเขตบน u_k, f_k จะเป็น 0 และไม่มีข้อสอบที่สามารถเลือกเพื่อนำไปใช้กับผู้สอบได้จากในขอบเขตเนื้อหานั้น ดังนั้นข้อสอบจะสามารถเลือกออกมาใช้ได้ก็ต่อเมื่อมีขอบเขตเนื้อหาอื่นที่ยังไม่ถูกนำมาใช้จนเต็มขอบเขตบน

วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ตามระดับชั้นของค่าอำนาจจำแนก (a-Stratified Method: a-STR) ดำเนินการตามขั้นตอนดังนี้คือ

1. แบ่งคลังข้อสอบเป็นชั้นตามค่าอำนาจจำแนกข้อสอบ (พารามิเตอร์ a) ชั้นแรกบรรจุข้อสอบอำนาจจำแนกต่ำสุด ชั้น 2 บรรจุข้อสอบอำนาจจำแนกสูงกว่าชั้นแรก และเพิ่มขึ้นเรื่อยๆ ตามลำดับชั้น จนกระทั่งชั้นสุดท้ายบรรจุข้อสอบอำนาจจำแนกสูงสุด

2. แบ่งขั้นการทดสอบหรือความยาวแบบทดสอบตามการแบ่งชั้นของคลังข้อสอบ

3. ค่าความสามารถของผู้สอบ ณ ตำแหน่งจุดเริ่มต้นการทดสอบ ข้อสอบที่มีค่าความยากใกล้เคียงกับค่าความสามารถของผู้สอบมากที่สุดจะถูกเลือกและนำไปใช้กับผู้สอบ เมื่อผู้สอบตอบ ผลการตอบข้อสอบจะถูกนำไปคำนวณหาค่าประมาณความสามารถ จากนั้นจึงนำค่าความสามารถของผู้สอบ ณ ตำแหน่งปัจจุบันไปใช้เลือกข้อสอบข้อต่อไป เมื่อทำการทดสอบจนครบตามความยาวแบบทดสอบย่อยในขั้นนั้น จึงเลื่อนการทดสอบขึ้นไปยังขั้นและชั้นคลังข้อสอบต่อไป

4. ทำการทดสอบจนครบทุกขั้นการทดสอบและทุกชั้นคลังข้อสอบ การทดสอบจึงยุติ

ตอนที่ 5 การโปรแกรมเชิงคณิตศาสตร์ (Mathematical Programming)

การศึกษาครั้งนี้ได้นำวิธีการโปรแกรมเชิงคณิตศาสตร์ (Mathematical Programming) เรียกอีกอย่างว่า การโปรแกรมเชิงคณิตศาสตร์ ความเป็นมาและแนวคิดของวิธีการโปรแกรมเชิงคณิตศาสตร์มีดังนี้

การโปรแกรมเชิงคณิตศาสตร์ (Mathematical Programming) เป็นเทคนิคการวิเคราะห์ที่เกี่ยวข้องกับกระบวนการตัดสินใจเพื่อหาแนวทางการแก้ปัญหาที่เหมาะสม (Optimal Solution) หรือการหาแนวทางปฏิบัติที่ดีที่สุด ที่เกิดขึ้นในศาสตร์ทางด้านการวิจัยดำเนินงาน (Operation Research) ซึ่งถูกนำไปใช้ในการแก้ปัญหาเกี่ยวกับการจัดสรรทรัพยากรในการดำเนินงานต่างๆ ขององค์กร เพื่อให้บรรลุเป้าหมายโดยอาศัยการวางแผนและการตัดสินใจที่เหมาะสม ในกิจกรรมขององค์กร เช่น กำหนดการผลิตสินค้า การมอบหมายงาน (Job Assignment) การควบคุมสินค้าคงคลัง (Inventory Control) และการวางแผนทางการตลาด โดยที่ทรัพยากรทุกอย่างไม่ว่าจะเป็นทรัพยากรบุคคล เครื่องจักร วัตถุดิบ และเงินทุน ต้องถูกนำไปใช้อย่างมีประสิทธิภาพสูงสุด ด้วยจุดเด่นที่สามารถช่วยในการตัดสินใจเพื่อหาแนวทางการแก้ปัญหาที่เหมาะสม เมื่อพิจารณาถึงโครงสร้างที่สอดคล้องกับรูปแบบของปัญหา เช่น การมอบหมายงานกับการคัดเลือกข้อสอบ และการควบคุมสินค้าคงคลังและการออกแบบคลังข้อสอบ ซึ่งลักษณะของปัญหานั้นมีความคล้ายคลึงกัน ดังนั้นการโปรแกรมเชิงคณิตศาสตร์สามารถนำไปประยุกต์ใช้ในศาสตร์ของการวัดผล ในประเด็นต่างๆ ดังนี้ การสร้างแบบสอบและการรวมข้อสอบเป็นแบบสอบแบบอัตโนมัติ (Adema, Boekkooi-Timminga, & van der Linden, 1991; Belov, 2008; Boekkooi-Timminga, 1990; Chang & Shiu, 2011; Luecht, 1998; Theunissen, 1985; van der Linden & Adema, 1998; van der Linden et al., 2006) การออกแบบคลังข้อสอบ (van der Linden, Veldkamp, & Reese, 2000) การประยุกต์ใช้กับทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัด (Sanders, 1992; Sanders, Theunissen, & Baas, 1989, 1991) และการนำไปใช้ในการทดสอบแบบปรับเหมาะ (Hendrickson, 2007; Wim J. van der Linden, 2005; van der Linden, 2010; van der

Linden & Chang, 2003; van der Linden & Diao, 2011) จากงานวิจัยที่กล่าวถึงจะพบว่าการโปรแกรมเชิงคณิตศาสตร์นั้นถูกใช้จัดการกับปัญหาที่มีความความซับซ้อนมาก ซึ่งวิธีการเชิงรูปนัย (Formal Approach) ส่วนใหญ่เกี่ยวข้องกับการสร้างตัวแบบทางคณิตศาสตร์ (Mathematical Modeling) และการใช้ขั้นตอนวิธีทางคอมพิวเตอร์ (Computer Algorithms) และถูกนำไปใช้ในซอฟต์แวร์ระบบเพื่อช่วยผู้ใช้ในการคำนวณแนวทางแก้ปัญหา (Solution) ภายใต้เงื่อนไขหรือข้อบังคับ (Constraints) ที่หลากหลาย ซอฟต์แวร์ที่นำมาใช้นั้นจะเรียกว่า Solver หรือ Optimizer รูปแบบโดยทั่วไปของปัญหาการโปรแกรมเชิงคณิตศาสตร์ (Mathematical Programming Problem) ประกอบด้วย 3 ส่วนดังนี้ 1) ฟังก์ชันเป้าหมาย (Objective Function) เป็นส่วนที่แสดงถึง วัตถุประสงค์ของการโปรแกรมว่าต้องการค่าสูงสุดหรือค่าต่ำสุด 2) ข้อจำกัดหรือเงื่อนไข (Constraints) แสดงถึง ขีดจำกัดของปัจจัยซึ่งอาจอยู่ในรูปสมการ และ/หรือ อสมการ และ 3) ตัวแปรการตัดสินใจ (Decision Variables) เป็นตัวแปรซึ่งเป็นผลเฉลยของกำหนดการว่าประกอบด้วยตัวแปรใดบ้าง โดยทั้งหมดจะถูกเขียนอยู่ในรูปของฟังก์ชันหรือความสัมพันธ์ของฟังก์ชันของตัวแปรที่แสดงเป็นตัวแบบทางคณิตศาสตร์

จากปัญหาโปรแกรมเชิงเส้นที่มีตัวแปรที่ควบคุมได้ n ตัว(n controlled variables) ภายใต้ข้อจำกัด m ข้อ (m constrains) ซึ่งมีตัวแบบดังนี้

หาค่าสูงสุด/ต่ำสุด $P = c_1x_1 + c_2x_2 + ... + c_nx_n$ .....(1)

โดยมีฟังก์ชันเงื่อนไขหรือข้อจำกัด

$$a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n \le b_1$$
$$a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n \le b_2$$
$$.$$
$$.$$
$$.$$
$$a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n \le b_m$$

.....(2)

และ $x_j \ge 0; j = 1,2,...,n$

เมื่อเขียนให้อยู่ในรูปมาตรฐานจะได้ตัวแบบดังนี้

หาค่าสูงสุด/ต่ำสุด $P = c_1x_1 + c_2x_2 + ... + c_nx_n$ .....(3)

โดยมีฟังก์ชันเงื่อนไขหรือข้อจำกัด

$$a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n + S_1 = b_1$$
$$a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n + S_2 = b_2$$
.
.
.
$$a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n + S_m = b_m$$ .....(4)

และ $x_j \geq 0; j = 1,2,...,n$ และ $S_i \geq 0; i = 1,2,...,m$ .....(5)

หรือเขียนแบบย่อได้ดังนี้ หาค่าสูงสุดของ $P = \overline{C}'\overline{X}$

โดยมีฟังก์ชันเงื่อนไขหรือข้อจำกัด $A\overline{X} = B$ และ $\overline{X} \geq 0$

เมื่อ

$$\overline{C} = \begin{bmatrix} c_1 \\ . \\ . \\ . \\ c_n \\ 0 \\ . \\ . \\ . \\ 0 \end{bmatrix} \qquad \overline{X} = \begin{bmatrix} x_1 \\ . \\ . \\ . \\ x_n \\ S_1 \\ . \\ . \\ . \\ S_m \end{bmatrix} \qquad B = \begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ . \\ b_m \end{bmatrix} \qquad A = (A_{m\times n} I_m)$$

จะเรียก $S_1, S_2, ...S_m$ ว่าตัวแปรอยู่เฉยหรือตัวแปรส่วนขาด (Slack Variables)

นิยาม 1 คำตอบที่เป็นไปได้ (Feasible Solution) ของปัญหาโปรแกรมเชิงเส้นตรง ก็คือค่าของ $X = (x_1, x_2, .., x_{n+m})$ ที่สอดคล้องกับเงื่อนไข (4) และ (5)

นิยาม 2.1 คำตอบฐาน (Basic Solution) คือคำตอบที่ได้จากการแก้สมการใน (4) โดยการกำหนดตัวแปร n ตัวให้เท่ากับ 0 เสียก่อน แล้วจึงแก้สมการหาค่าตัวแปร m ตัวที่เหลือ กำหนดว่า Determinant ของตัวแปร m ตัวที่เหลือนี้จะต้องไม่มีค่าเป็น 0 ตัวแปร m มีชื่อว่าตัวแปรฐาน (Basic Variable)

นิยาม 2.2 คำตอบที่เป็นไปได้ขั้นฐาน (Basic Feasible Solution) ก็คือคำตอบที่สอดคล้องกับเงื่อนไข (5) นั้นคือ ตัวแปรฐานทุกตัวต้องไม่มีค่าเป็นลบ

นิยาม 3 Non-degenerate basic feasible solution ก็คือคำตอบที่เป็นไปได้ขั้นฐานที่มีค่าของ x_j เป็นบวกเพียง m ตัวเท่านั้นก็คือ ตัวแปรฐานทุกตัวเป็นบวก

นิยาม 4 คำตอบที่เป็นไปได้สูงสุด (Maximum feasible solution) คือคำตอบที่เป็นไปได้ที่ทำให้ฟังก์ชันเป้าหมาย (3) มีค่ามากที่สุด

เมื่อนำไปประยุกต์ใช้กับการทดสอบแบบปรับเหมาะโดยมีฟังก์ชันเป้าหมายเป็นการคัดเลือกข้อสอบที่ให้สารสนเทศสูงสุด และมีฟังก์ชันเงื่อนไขหรือข้อจำกัด รูปแบบและคุณลักษณะของข้อสอบสามารถเขียนเป็นตัวแบบ ได้ดังนี้

ฟังก์ชันเป้าหมาย

$$\text{Maximize} \sum_{i \in 0_k} I_i\left(\hat{\theta}_{k-1}\right) x_i \quad \text{(สารสนเทศสูงสุด)}$$

ฟังก์ชันเงื่อนไขบังคับ

$$\sum_{i=1}^{I} x_i = n \qquad \text{(ความยาวของแบบสอบ)}$$

$$\sum_{s=1}^{S} z_s = m \qquad \text{(จำนวนของสิ่งกระตุ้น stimulus)}$$

$$\sum_{i \in S_{k-1}} x_i = k-1 \qquad \text{(ข้อสอบที่พร้อมนำไปใช้กับผู้สอบ)}$$

$$\sum_{i \in V_s} x_i \leq\geq n_s z_s, s = 1,...,S \qquad \text{(จำนวนข้อต่อสิ่งกระตุ้น stimulus)}$$

$$\sum_{i \in V_c^{item}} x_i \leq\geq n_c^{item}, c = 1,...,C \qquad \text{(คุณสมบัติข้อสอบเชิงจัดประเภท)}$$

$$\sum_{i=1}^{I} q_i z_i \leq\geq b_q^{item} \qquad \text{(คุณสมบัติข้อสอบเชิงจำนวน)}$$

$$\sum_{i \in V_c^{item}} z_s \leq\geq n_c^{stim}, c = 1,...,C \qquad \text{(คุณสมบัติสิ่งกระตุ้นเชิงจัดประเภท)}$$

$$\sum_{i=1}^{I} q_s z_s \leq\geq b_q^{stim} \qquad \text{(คุณสมบัติสิ่งกระตุ้นเชิงจำนวน)}$$

$$\sum_{i \in V_e^{item}} x_i \leq 1, e = 1,...,E \qquad \text{(ข้อสอบที่ไม่เกี่ยวข้องกัน)}$$

$$\sum_{s \in V_e^{item}} z_s \leq 1, e = 1,...,E \qquad \text{(สิ่งกระตุ้นที่ไม่เกี่ยวข้องกัน)}$$

$$x_i \in \{0,1\}, i = 1,...,I \qquad \text{(ปริเขตของตัวแปร)}$$

$$z_S \in \{0,1\}, s = 1,...,S \qquad \text{(ปริเขตของตัวแปร)}$$

จากการทบทวนเอกสารและงานวิจัยที่เกี่ยวข้องพบว่า ซอฟต์แวร์ที่เรียกว่า Solver หรือ Optimizer ที่มีการนำมาใช้กันอย่างแพร่หลายเพื่อจัดการกับปัญหาที่ต้องการทราบจุดหรือคำตอบที่เหมาะสม คือ 1) ซอฟต์แวร์ ILOG CPLEX เป็นโปแกรมสำหรับแก้ปัญหาด้าน Optimization ในการ

ตัดสินใจทางธุรกิจที่มีประสิทธิภาพสูงและมีความยืดหยุ่นในใช้งานสามารถเลือกใช้ภาษาที่หลากหลายในการเขียนคำสั่งได้ เช่น C, Visual Basic, FORTRAN และ Java เป็นต้น CPLEX เป็นซอฟต์แวร์ที่มีราคาสูงเป็นที่นิยมใช้ในองค์กรขนาดใหญ่ จากการศึกษาเอกสารและงานวิจัยทางการทดสอบพบว่ามีงานวิจัยจำนวนมากที่ใช้โปรแกรม CLPEX ในการแก้ปัญหาด้าน Optimization เพื่อควบคุมเงื่อนไขบังคับ (Ariel, Veldkamp, & Breithaupt, 2006; Belov et al., 2008; Chang & van der Linden, 2003; Cheng et al., 2007; van der Linden, 2002; Wim J. van der Linden, 2005; van der Linden, 2010; van der Linden & Chang, 2003; van der Linden & Veldkamp, 2004; Veldkamp, 2010) 2) โปแกรม Solver เป็นโปรแกรมเสริมหรือเป็นส่วนหนึ่งของชุดของคำสั่งในของโปรแกรมไมโครซอฟต์เอ็กเซล โปรแกรม Solver สามารถค้นหาค่าที่เหมาะสม (ค่าสูงสุดหรือต่ำสุด) สำหรับเซลล์เป้าหมาย ซึ่งขึ้นกับเงื่อนไขหรือขีดจำกัดของค่าในเซลล์สูตรอื่นๆ ในแผ่นงาน Solver ทำงานกับกลุ่มเซลล์ที่เรียกว่าเซลล์ตัวแปรการตัดสิน เนื่องจากโปรแกรม Solver ต้องใช้งานผ่านโปรแกรมเอ็กเซลจึงมีข้อจำกัดในเรื่องของการวิเคราะห์ข้อมูลที่มีข้อมูลรับเข้าจำนวนมากๆ และ 3) โปรแกรม lp_solve (Berkelaar, 2007; Berkelaar, Eikland, & Notebaert, 2004) เป็นฟรีโปรแกรมที่พัฒนาขึ้นครั้งแรกโดย Michel Berkelaar และได้รับการพัฒนาต่อยอดมาจนถึงปัจจุบันโดยกลุ่มนักวิชาการด้าน Optimization โปรแกรม lp_solve ใช้วิธี revised simplex และวิธี Branch-and-bound เพื่อหาค่าที่เหมาะสมสำหรับฟงก์ชั่นเป้าหมาย นอกจากนี้ จุดเด่นอีกประการหนึ่งของโปรแกรม lp_solve คือ ไม่จำกัดขนาดของโมเดล และตัวแปรเงื่อนไขบังคับ และยังสามารถเรียนใช้ไลบารีจากโปรแกรมภาษาอื่น เช่น C, VB, .NET และ Delphi ในปัจจุบันโปรแกรม lp_solve เป็นแพ็คเกจหนึ่งของโปรแกรม R จากการศึกษาเอกสาร พบว่า โปรแกรม lp_solve ถูกนำมาใช้ในการรวมข้อสอบเป็นแบบสอบให้มีคุณลักษณะตรงตามเงื่อนไขบังคับที่กำหนด (Diao & van der Linden, 2011) ดังนั้นผู้วิจัยจึงเลือกใช้ โปรแกรม lp_solve เพื่อใช้ในการหาค่าที่เหมาะสมสำหรับการรวมแบบทดสอบย่อยให้เป็นแบบสอบเสมือนในการทดสอบแบบปรับเหมาะที่ใช้ในการศึกษาครั้งนี้ เนื่องจาก lp_solve อยู่ในสภาพแวดล้อมของ R ผู้วิจัยจึงสามารถเรียกใช้คำสั่ง ส่งผ่านข้อมูล กำหนดเงื่อนไขบังคับต่างๆ และบันทึกผลลัพธ์ที่อยู่ภายใน workspace เดียวกันได้รวดเร็วโดยไม่ต้องส่งผ่านข้อมูลออกไปประมวลผลด้วยโปรแกรมอื่น

ตอนที่ 6 งานวิจัยที่เกี่ยวข้องกับการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์

จากศึกษางานวิจัยในอดีตที่เกี่ยวข้องกับการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์การควบคุมเงื่อนไขบังคับที่ไม่ใช่ทางสถิติในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ผู้วิจัยสามารถสรุปเป็นประเด็นต่างๆ ได้ดังนี้ 1) การเลือกใช้โมเดลการตอบสนองข้อสอบ 2) การทำให้คลังข้อสอบมีความเป็นมาตรฐาน (calibrated item pool) 3) วิธีการคัดเลือกข้อสอบและประมาณค่าความสามารถของผู้สอบ 4) วิธีการควบคุมเงื่อนไขบังคับในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ 5) การพิจารณาประสิทธิภาพของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์

6.1. การเลือกใช้โมเดลการตอบสนองข้อสอบ

Wainer et al. (2002) กล่าวว่าการเลือกใช้โมเดลการตอบสนองข้อสอบควรพิจารณาให้สอดคล้องกับลักษณะของข้อสอบ เนื่องจากโมเดลแต่ละแบบมีข้อตกลงเบื้องต้นแตกต่างกัน จากการศึกษาถึงอิทธิพลของการละเลยปฏิสัมพันธ์ของข้อสอบที่เกิดขึ้นภายใต้โมเดล 2 พารามิเตอร์แบบโลจิสติก ของ Tuerlinckx and De Boeck (2001) พบว่า ขนาดของการประมาณค่าพารามิเตอร์อำนาจจำแนกมีความลำเอียงโดยขึ้นอยู่กับปฏิสัมพันธ์ของพารามิเตอร์ความยาก ในกรณีที่มีปฏิสัมพันธ์เชิงบวกการประมาณค่าพารามิเตอร์อำนาจจำแนกจะสูงเกินจริง (Overestimating) ทำให้สารสนเทศของข้อสอบสูงขึ้นผิดปกติ (inflated) และความคลาดเคลื่อนมาตรฐานก็ถูกกดให้ต่ำลง (deflated) แต่ในทางตรงกันข้าม ถ้าเกิดปฏิสัมพันธ์เชิงลบการประมาณค่าพารามิเตอร์อำนาจจำแนกจะต่ำเกินจริง (Underestimating) ทำให้สารสนเทศของข้อสอบถูกทำให้ต่ำลงอย่างผิดปกติ (deflated) และความคลาดเคลื่อนมาตรฐานของข้อสอบสูงขึ้น (inflated) ซึ่งในทางปฏิบัติคุณภาพของข้อสอบอธิบายได้จากขนาดของพารามิเตอร์อำนาจจำแนก ดังนั้น ค่าสารสนเทศของข้อสอบมีความสัมพันธ์เชิงบวกกับค่าพารามิเตอร์อำนาจจำแนกยกกำลังสอง และทฤษฎีการตอบสนองข้อสอบ (IRT) เป็นโมเดลที่มีข้อตกลงเบื้องต้นที่เคร่งครัด การฝ่าฝืนข้อตกลงเบื้องต้นในเรื่องของความเป็นอิสระในการตอบข้อสอบ (Local Independence) เกิดขึ้นเมื่อผู้สอบตอบข้อสอบ เมื่อกลุ่มข้อสอบใช้สิ่งเร้าร่วมกัน เช่น การอ่านบทความ (reading passage) การดูกราฟหรือตาราง ดังนั้นการตอบสนองของผู้สอบต่อข้อสอบในแบบสอบจะมีความสัมพันธ์กันอย่างมีเงื่อนไขเพราะกลุ่มของข้อสอบใช้สิ่งเร้าร่วมกัน สิ่งเร้าจึงเป็นอีกปัจจัยหนึ่งที่มีอิทธิพลต่อผลการตอบข้อสอบที่นอกเหนือจากความสามารถของผู้สอบ การละเลยข้อตกลงเบื้องต้นนี้ ถ้าสิ่งเร้ามีอิทธิพลทางบวกจะมีแนวโน้มที่ทำให้ความแม่นยำของการวัดที่ได้จากแบบทดสอบย่อยมีค่าสูงกว่าความเป็นจริง (overestimate) แต่ถ้าสิ่งเร้ามีอิทธิพลทางลบจะมีแนวโน้มที่ทำให้ความแม่นยำของการวัดที่ได้จากแบบทดสอบย่อยมีค่าต่ำกว่าความเป็นจริง (underestimate)

เนื่องจากการศึกษาครั้งนี้ศึกษาใช้ข้อสอบที่มีลักษณะเป็นกลุ่มของข้อสอบที่ใช้สิ่งเร้าร่วมกัน ดังนั้นผู้วิจัยจึงเลือกใช้โมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย (Testlet Response Model) สำหรับการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์

6.2 การทำให้คลังข้อสอบมีความเป็นมาตรฐาน

คลังข้อสอบถือว่าเป็นส่วนประกอบหลักที่สำคัญที่สุดส่วนหนึ่งของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (Flaugher, 2000) ดังนั้นคุณภาพของคลังข้อสอบจึงมีอิทธิพลอย่างมากต่อประสิทธิภาพของ การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ Weiss (1985) เสนอว่าเพื่อให้การทดสอบแบบปรับเหมาะ ด้วยคอมพิวเตอร์มีประสิทธิภาพจำเป็นจะต้องใช้คลังข้อสอบที่มีขนาดใหญ่และข้อสอบมีค่าอำนาจจำแนกสูง ครอบคลุมตลอดช่วงความสามารถของผู้สอบ ซึ่งคลังข้อสอบควรจะมีข้อสอบขั้นต่ำอย่างน้อย 100 ข้อ อย่างไรก็ตามก็ขึ้นอยู่กับโครงสร้างของคลังข้อสอบด้วยและถ้าเพิ่มขนาดคลังข้อสอบให้อยู่ในช่วง 150 - 200 ขึ้นจะให้ผลลัพธ์ที่ดีขึ้นกว่าเดิม นอกจากนี้ Reckase (2010) ได้ทำการศึกษาเกี่ยวกับการออกแบบคลังข้อสอบ ให้เหมาะสมกับการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ด้วยเทคนิค p-optimal และ r-optimal พบว่า การออกแบบคลังข้อสอบนั้นส่งผลต่อประสิทธิภาพของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ แต่ผล จากการศึกษายังไม่สามารถให้ข้อสรุปที่ชัดเจนได้ว่าคลังข้อสอบของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ควรจะมีขนาดใหญ่เท่าใด Reckase (2010) ให้แนวทางการพิจารณาไว้ คือ ขนาดของคลังข้อสอบและการแจกแจงของค่าความยากของข้อสอบขึ้นอยู่กับลักษณะการแจกแจงของความสามารถของกลุ่มผู้สอบ และการออกแบบการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (fixed test หรือ variable test) ผู้บริหารการทดสอบจำเป็นจะต้องทราบลักษณะการแจกแจงของความสามารถของผู้สอบเพื่อที่จะสามารถจัดคลังข้อสอบให้ค่าความยากของข้อสอบมีลักษณะการแจกแจงครอบคลุมช่วงความสามารถของผู้สอบ

ดังนั้นในการศึกษาครั้งนี้ ผู้วิจัยใช้ตัวอย่างที่มีการแจกแจงปกติดังนั้น การแจกแจงของค่าพารามิเตอร์ความยากจึงต้องมีความสอดคล้องกับลักษณะการแจกแจงของตัวอย่างที่ศึกษาเพื่อให้สอดคล้องกับคำแนะนำของ Reckase (2010) และผู้วิจัยสร้างค่าพารามิเตอร์อำนาจจำแนกโดยกำหนดลักษณะการแจกแจงตามงานวิจัยของ Ngudgratoke and Yon (2006) เพื่อให้ค่าพารามิเตอร์มีคุณสมบัติตามคำแนะนำของ Weiss (1985) และเนื่องจากข้อสอบที่ใช้ในการศึกษาครั้งนี้ถูกจัดเป็นกลุ่มและใช้สิ่งเร้าร่วมกัน ผู้วิจัยจึง calibrated คลังข้อสอบ โดยใช้โมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย (Testlet Response Model) และใช้เทคนิค mcmc ในการ

ประมาณค่าพารามิเตอร์ข้อสอบโดยใช้ฟังก์ชั่น 3PNO Testlet Model ในแพ็กเกจ (Supplementary Item Response Theory Models: SIRT) จากโปรแกรม R

6.3 การคัดเลือกข้อสอบและการประมาณค่าความสามารถ

Murphy, Dodd, and Vaughn (2010) ศึกษาเปรียบเทียบประสิทธิภาพของเทคนิคการคัดเลือกข้อสอบ 3 วิธี คือ 1) maximum Fisher's information (MFI) 2) maximum posterior weighted information (MPWI) และ 3) minimum expected posterior variance (MEPV) ในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่ข้อสอบถูกจัดกลุ่มเป็นแบบทดสอบย่อย (testlet) ภายใต้โมเดล IRT และ TRT ดำเนินการวิจัยผ่านการศึกษาในสถานการจำลองโดยใช้วิธีมอนติ คาร์โล เพื่อสร้างชุดข้อมูลแบบแผนการตอบจำนวน 10 กลุ่ม แต่ละกลุ่มมีผู้สอบจำนวน 1,000 คน ข้อมูลทั้ง 10 กลุ่ม สำหรับทำการทดลองซ้ำ 10 ครั้ง ในแต่ละเงื่อนไขการศึกษา ผลการศึกษาพบว่า เมื่อใช้โมเดล IRT กับข้อมูลที่ผ่าผืนข้อตกลงเบื้องต้นเรื่อง local item dependence (LID) ผลการประมาณค่าความสามารถมีแนวโน้มที่จะมีค่ามากกว่าค่าที่เป็นจริง (overestimate) และเมื่อเปรียบเทียบประสิทธิภาพของการคัดเลือกข้อสอบทั้งสามวิธี พบว่า ไม่แตกต่างกัน เนื่องจากวิธีการคัดเลือกข้อสอบแบบเบเซี่ยนจะให้ประสิทธิภาพสูงกว่าวิธีการคัดเลือกข้อสอบแบบสารสนเทศสูงสุดเมื่อใช้ข้อสอบจำนวนน้อยข้อ แต่การศึกษาครั้งนี้ใช้การเลือกข้อสอบเป็นชุดของแบบทดสอบย่อย แม้ว่าผู้สอบแต่ละคนจะได้แบบทดสอบย่อยจำนวนไม่มาก แต่เมื่อคิดรวมเป็นจำนวนข้อผู้สอบได้ทำข้อสอบคนละประมาณ 50 ข้อ ซึ่ง van der Linden (1998a) ได้อธิบายไว้ว่า เมื่อจำนวนข้อสอบในแบบสอบที่ผู้สอบได้รับมีความยาวเพิ่มขึ้น (เข้าใกล้ค่าอนันต์) การแจกแจงของการประมาณค่าความสามารถจะลู่เข้าระดับค่าความสามารถจริง

เนื่องจากการวิจับครั้งนี้ผู้วิจัยกำหนดเกณฑ์การยุติการทดสอบโดยพิจารณาจากค่า SEE ดังนั้นผู้สอบแต่ละคนอาจจะได้รับแบบสอบที่มีความยาวแตกต่างกัน เพื่อให้การประมาณค่าความสามารถมีความถูกต้องแม่นยำ ครอบคลุมกรณีที่ผู้สอบได้รับแบบสอบขนาดสั้น วิจัยผู้วิจัยเลือกใช้วิธีการประมาณค่าความสามารถของผู้สอบเป็นแบบเบเซี่ยน โดยใช้วิธีการแบบ expected a posterior variance (EPV)

6.4 วิธีการควบคุมเงื่อนไขบังคับในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์

Kingsbury & Zara (1991 cited in Wang & Kolen, 2001) เสนอว่าในการเลือกข้อสอบควรให้ความสำคัญกับการกำหนดคุณลักษณะของข้อสอบเป็นอันดับแรก ขั้นตอนวิธี (Algorithm) ของ Kingsbury และ Zara จะแบ่งคลังข้อสอบออกเป็นพาร์ทิชันละเท่าๆ กัน ตามขอบเขตเนื้อหาที่มีความแตกต่างกันและเลือกข้อสอบที่มีค่าสารสนเทศสูงสุดจากแต่ละพาร์ทิชัน ประโยชน์จากขั้นตอน

วิธีนี้ช่วยให้มั่นใจว่าการคัดเลือกข้อสอบถูกจำกัดตามตารางการกำหนดคุณลักษณะของข้อสอบอย่างเข้มงวด ข้อเสียคือต้องยอมลดความแม่นยำของการประมาณค่าเนื่องจากการสร้างสมดุลของเนื้อหาจะถูกกำหนดในขั้นตอนของการเลือกข้อสอบทุกข้อจึงเป็นการลดตัวเลือกในการเลือกข้อสอบที่ให้สารสนเทศสูงบริเวณตำแหน่งความสามารถของผู้สอบ

van der Linden และ Reese (1998) ได้ประยุกต์ใช้เทคนิคการโปรแกรมเชิงเส้น (Linear Programming) ในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ ขั้นตอนวิธีนี้สามารถให้ผลลัพธ์ที่เป็นไปได้ (Feasible Solution) ในการนำไปใช้กับการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ และผลลัพธ์ที่เป็นไปได้จะเป็นแบบสอบที่มีคุณลักษณะตามเงื่อนไขบังคับที่กำหนดในตารางคุณลักษณะของข้อสอบ (Content Specifications) และให้ค่าที่เหมาะที่สุดสำหรับฟังก์ชันเป้าหมาย

Stocking & Swanson (1991 cited in Wang & Kolen, 2001) เสนอขั้นตอนวิธีแบบฮิริสติก (Heuristic Algorithm) การเลือกข้อสอบให้ผู้สอบจะใช้วิธีการหาผลรวมถ่วงน้ำหนักของสารสนเทศของข้อสอบกับปริมาณของข้อกำหนดที่รองรับในแต่ละเนื้อหา การกำหนดน้ำหนักจะพิจารณาจากความสำคัญของเนื้อหาหรือข้อจำกัดนั้น เนื่องจากค่าสารสนเทศและค่าเบี่ยงเบนของเนื้อหา (Content Deviation) ไม่อยู่ในมาตรเดียวกัน การกำหนดน้ำหนักไปยังค่าสารสนเทศนั้นจำเป็นต้องพิจารณาผ่านสถานการณ์จำลอง จุดเด่นประการแรกของขั้นตอนวิธีค่าเบี่ยงเบนถ่วงน้ำหนัก คือ ข้อจำกัด (Constraints) จำนวนมากนอกเหนือจากการกำหนดคุณลักษณะของข้อสอบ (Content Specification) สามารถรวมเข้าในขั้นตอนวิธีค่าเบี่ยงเบนถ่วงน้ำหนักได้ ประการที่สองน้ำหนักที่กำหนดลำดับความสำคัญที่อยู่บนข้อกำหนดเกี่ยวกับเนื้อหาและความแม่นยำการวัดสามารถกำหนดได้อย่างยืดหยุ่น และข้อจำกัด คือ มีความไม่แน่นอนในเรื่องความสมดุลของการกำหนดคุณลักษณะของข้อสอบ (Content Specifications)

Chang และ Ying (1999) เสนอการเลือกข้อสอบวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ตามระดับชั้นของค่าอำนาจจำแนก (a-Stratified Computerized Adaptive Testing: a-STR) ทำการศึกษาในสถานการณ์จำลองเพื่อเปรียบเทียบวิธี a-STR กับวิธี Maximum Fisher Information ร่วมกับการควบคุมการใช้ข้อสอบซ้ำของ Sympson และ Hetter (FSH) และวิธีการเลือกข้อสอบแบบเบเซี่ยน (EAP) ร่วมกับการควบคุมการใช้ข้อสอบซ้ำ Sympson และ Hetter (BSH) ผลการศึกษาพบว่า

ผลการศึกษาที่ 1 สำหรับความยาวแบบทดสอบ 40 ข้อพบว่า วิธีการเลือกข้อสอบทั้ง 3 วิธีให้ความลำเอียงเฉลี่ยและความคลาดเคลื่อนกำลังสองเฉลี่ยใกล้เคียงกัน อัตราการทับซ้อนข้อสอบของวิธี a-STR มีค่าน้อยอัตราทับซ้อนข้อสอบของวิธี FSH ประมาณ 0.5 เท่า และวิธี BSH กับวิธี a-STR

ช่วยเพิ่มอัตราการใช้ข้อสอบที่ถูกนำมาใช้น้อยครั้ง วิธี FSH และวิธี BSH ใช้คลังข้อสอบได้สมดุลมากกว่า เมื่อพิจารณาการแจกแจงของอัตราการใช้ข้อสอบซ้ำด้วยสถิติไคสแควร์พบว่าวิธี a-STR สัมพันธ์กับวิธี FSH และ BSH ส่วนแบบสอบที่มีความยาว 60 ข้อ วิธีการที่ศึกษาทั้ง 3 วิธีให้ค่าความลำเอียงเฉลี่ยแตกต่างกันเล็กน้อยเมื่อเทียบกับแบบทดสอบที่มีความยาว 40 ข้อ วิธี a-STR ให้อัตราการทับซ้อนข้อสอบน้อยกว่าวิธี FSH และวิธี BSH วิธี a-STR ให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุดขณะที่วิธี BSH ให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยมากที่สุด วิธี a-STR ลดความเบ้ได้มากกว่า 50 เปอร์เซ็นต์เมื่อเทียบกับวิธี FSH และวิธี BSH นอกจากนี้วิธี a-STR ให้สมดุลการแจกแจงอัตราการใช้ข้อสอบซ้ำมากกว่าและยังเพิ่มการใช้ข้อสอบซ้ำอำนาจจำแนกต่ำ

ผลการศึกษาที่ 2 แสดงถึงความคลาดเคลื่อนกำลังสองเฉลี่ยและความลำเอียงเฉลี่ยของทั้ง 3 วิธีไม่แตกต่างกัน แต่วิธี a-STR มีแนวโน้มให้ความคลาดเคลื่อนกำลังสองเฉลี่ยค่อยๆ เพิ่มขึ้น ลดอัตราการทับซ้อนข้อสอบ ลดจำนวนข้อสอบที่มีการใช้ข้อสอบซ้ำน้อย และลดความเบ้ของการแจกแจงอัตราการใช้ข้อสอบซ้ำเมื่อเปรียบเทียบกับวิธี FSH และวิธี BSH

Chang, Qian และ Ying (2001) ศึกษาประสิทธิภาพการเลือกข้อสอบวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ตามระดับชั้นของค่าอำนาจจำแนก (a-Stratified Method: a-STR) เปรียบเทียบกับวิธีตามระดับชั้นของค่าอำนาจจำแนกและค่าความยาก (a-Stratified Method with b Blocking: ab-STR) ในด้านประสิทธิภาพการประมาณค่าความสามารถของผู้สอบ การใช้ข้อสอบในคลังข้อสอบ และความปลอดภัยของข้อสอบ ผลการศึกษาพบว่า วิธี ab-STR ให้ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างความสามารถจริงและค่าประมาณความสามารถสูงเมื่อเทียบกับวิธี a-STR วิธี ab-STR และมีประสิทธิภาพสูงกว่ากว่าวิธี a-STR ในด้านดังต่อไปนี้ มีประสิทธิภาพการใช้คลังข้อสอบสูงกว่า ลดความลำเอียง ลดความคลาดเคลื่อนกำลังสองเฉลี่ย และลดการอัตราการทับซ้อนของข้อสอบในส่วนของการแจกแจงอัตราการใช้ข้อสอบซ้ำของวิธี ab-STR มีประมาณ 1 ใน 5 ของวิธี a-STR และความเบ้ของการแจกแจงอัตราการใช้ข้อสอบซ้ำในวิธี ab-STR ถูกทำให้ลดลงเมื่อเทียบกับวิธี a-STR ประมาณร้อยละ 74

Hau และ Chang (2001) ศึกษาเปรียบเทียบวิธีการคัดเลือกข้อสอบแบบวิธี Descending a-Stratified: a-DSTR และวิธี Non Systematic Stratified: N-STR เปรียบเทียบกับวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ตามระดับชั้นของค่าอำนาจจำแนก (a-STR) เมื่อ a-DSTR คือ วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ตามระดับชั้นของค่าอำนาจจำแนก (a-STR) เรียงระดับค่าอำนาจจำแนก (a) ในคลังข้อสอบชั้นที่ 1 บรรจุข้อสอบอำนาจจำแนกสูง และค่าอำนาจจำแนกจะลดลงตามชั้นของคลังข้อสอบจนถึงชั้นสุดท้าย N-STR คือ วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ตามระดับชั้นของค่าอำนาจจำแนก (a-STR) ชั้นแรกบรรจุด้วยข้อสอบ

อำนาจจำแนกปานกลาง ชั้นท้ายๆ บรรจุด้วยข้อสอบอำนาจจำแนกสูงและต่ำ ผลการศึกษาพบว่า วิธี a-STR และ วิธี a-DSTR ให้ความคลาดเคลื่อนกำลังสองเฉลี่ยและความลำเอียงเฉลี่ยมีแนวโน้มลดลงเหมือนกันเมื่อความยาวแบบทดสอบเพิ่มขึ้น แต่ความคลาดเคลื่อนกำลังสองเฉลี่ยในวิธี a-STR ต่ำกว่า วิธี a-DSTR ซึ่งสนับสนุนงานวิจัยของ Chang และ Ying (1999) ที่กล่าวว่าข้อสอบอำนาจจำแนกต่ำจะให้สารสนเทศโดยรวมได้มากกว่าในระยะเริ่มต้นของการทดสอบ เมื่อค่าประมาณความสามารถไกลจากค่าความสามารถจริง ความลำเอียงให้ผลด้านบวก ด้านลบ และเข้าใกล้ 0 เมื่อความยาวแบบทดสอบเพิ่มขึ้น วิธี a-STR ดีกว่าวิธี a-DSTR เพราะให้สมดุลการใช้ข้อสอบดีกว่าส่วนวิธี N-STR ให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยกึ่งกลางระหว่างวิธี a-STR และวิธี a-DSTR

Leung, Chang และ Hau (2002) ศึกษาเปรียบเทียบวิธีการเลือกข้อสอบ 3 วิธี คือ วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ตามระดับชั้นของค่าอำนาจจำแนก (a-Stratified Method: a-STR) ร่วมกับการควบคุมการใช้ข้อสอบซ้ำของ Sympson-Hetter (a-Stratified Method with Sympson-Hetter Procedure: a-STR-SH) และวิธีสารสนเทศสูงสุดร่วมกับการควบคุมการใช้ข้อสอบซ้ำของ Sympson-Hetter (Maximum Information with Sympson-Hetter procedure: Max-I-SH) ประเมินผลโดยค่าความเชื่อถือ (Fidelity) ความลำเอียงเฉลี่ย ความคลาดเคลื่อนกำลังสองเฉลี่ย จำนวนข้อสอบที่มีการแสดงมากเกินไป จำนวนข้อสอบที่มีการใช้ต่ำ การแจกแจงอัตราการใช้ข้อสอบซ้ำด้วยสถิติไคสแควร์ อัตราการทับซ้อนข้อสอบ เวลาในการคำนวณ (Computing Time) การศึกษาทำในสถานการณ์จำลอง 4 สถานการณ์ 2 สถานการณ์แรก การทดสอบจำลองในกลุ่มตัวอย่างค่าความสามารถ 3,000 ค่าซึ่งสุ่มจากโค้งการแจกแจงมาตรฐาน N(0,1) 2 สถานการณ์หลังใช้ผู้สอบค่าความสามารถ 10,000 ค่า (จำแนกระดับความสามารถ -3.0,..., 3.0)

ผลการศึกษาที่ 1 พบว่า สัมประสิทธิ์ความเชื่อถือ (Fidelity Coefficients) ทั้ง 3 วิธีมีความสัมพันธ์ของการประมาณค่าความสามารถสูง ความลำเอียงเฉลี่ยเข้าใกล้ 0 ความคลาดเคลื่อนกำลังสองเฉลี่ยสำหรับความยาวแบบทดสอบ 40 ข้อของวิธี Max-I-SH น้อยกว่าวิธี a-STR และวิธี a-STR-SH แต่เมื่อความยาวแบบทดสอบ 60 ข้อ พบว่า วิธี a-STR ให้ผลน้อยกว่า เมื่อเปรียบเทียบวิธี a-STR-SH กับวิธี Max-I-SH วิธี a-STR-SH มีอัตราการใช้ข้อสอบซ้ำจำนวนน้อยที่สุดเมื่อเทียบกับอีกวิธี การทับซ้อนข้อสอบของวิธี Max-I-SH สูงกว่าอีก 2 วิธีเมื่อใช้ความยาวแบบทดสอบ 40 ข้อ แต่เมื่อจำนวนข้อสอบเพิ่มขึ้นความแตกต่างจะลดลงวิธี a-STR-SH จะให้ค่าการทับซ้อนข้อสอบเล็กที่สุด การสร้างพารามิเตอร์ควบคุมการใช้ข้อสอบซ้ำทั้งความยาวแบบทดสอบ 40 และ 60 ข้อ พบว่า วิธี a-STR-SH ใช้เวลาน้อยกว่าวิธี Max-I-SH ผลการศึกษาที่ 2 พบว่า สัมประสิทธิ์ความเชื่อถือทั้ง 3 วิธีลดลงเมื่อความยาวแบบทดสอบสั้นลงและขนาดคลังข้อสอบเล็กลงแต่ยังเปรียบเทียบกันได้ ความลำเอียงเฉลี่ยทั้ง 3 วิธีมี ขนาดลดลงและวิธี a-STR และวิธี a-STR-SH มีความสมดุลการใช้คลังข้อสอบ

และอัตราการทับซ้อนข้อสอบของดีกว่าวิธี Max-I-SH ผลการศึกษาที่ 3 และ 4 พบว่า ทั้ง 3 วิธีให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยใกล้เคียงกันวิธี a-STR และวิธี a-STR-SH ส่วนวิธี MI-SH จะให้ประสิทธิภาพดีเมื่อความยาวแบบทดสอบสั้น

Chang และ Ansley (2003) ศึกษาเปรียบเทียบคุณสมบัติของวิธีการควบคุมการใช้ข้อสอบซ้ำ 5 วิธี ดังนี้ 1) เทคนิค 5-4-3-2-1 ของ McBride and Martin (MM) 2) วิธีการ Sympson-Hetter (SH) 3) วิธีการ Davey-Parshall (DP) 4) วิธีการ Stocking and Lewis Unconditional Multinomial (SL) และ 5) วิธีการ Stocking and Lewis Conditional Multinomial (SLC) โดยใช้คลังข้อสอบขนาด 360 และ 720 ข้อ ความยาวของแบบสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ถูกกำหนดไว้ที่ 30 ข้อ และอัตราการใช้ข้อสอบซ้ำสูงสุดที่พึงปรารถนาถูกกำหนดไว้ที่ 0.1 และ 0.2 ยกเว้นวิธี MM เพราะไม่มีการควบคุมโดยตรง มีการกำหนดให้ความสามารถของผู้สอบเป็น 0 ในช่วงเริ่มต้นการทดสอบ วิธีการประมาณค่าความสามารถของผู้สอบใช้วิธี Maximum Likelihood Estimation (MLE)

ผลการศึกษาแสดงให้เห็น 1) ข้อจุดเด่นและข้อจำกัดของแต่ละวิธี มีดังนี้ วิธี MM ไม่รับรองความปลอดภัยของแบบสอบ ส่วนวิธี SH และ SL ให้ผลที่คล้ายกันมากยกเว้นเมื่อใช้กับคลังข้อสอบขนาดเล็กภายใต้เงื่อนไขของอัตราการใช้ข้อสอบซ้ำที่ 0.1 วิธี SH และ SL ไม่มีความแตกต่างกันในเรื่องของประสิทธิภาพในการควบคุมการใช้ข้อสอบซ้ำ แต่แตกต่างกันในเรื่องวิธีการเลือกข้อสอบ และเพิ่มพารามิเตอร์การควบคุมการใช้ข้อสอบซ้ำวิธี SH มีประสิทธิภาพกว่าวิธีอื่นๆ วิธี SLC ให้ผลที่พึงพอใจในเรื่องของอัตราการใช้ข้อสอบซ้ำสูงสุด และ อัตราการทับซ้อนของแบบสอบ อย่างไรก็ตาม เมื่อมีการควบคุมอัตราการใช้ข้อสอบซ้ำอย่างเข้มงวด โดยเฉพาะอย่างยิ่งเมื่อระดับความสามารถของผู้สอบมีค่าสูงมาก ๆ ความคลาดเคลื่อนมาตรฐานอย่างมีเงื่อนไขของการวัดจะเพิ่มขึ้น 2) ผลของขนาดคลังข้อสอบ เมื่อคลังข้อสอบมีขนาดใหญ่ไม่เพียงพอ ซึ่งวิธีการโดยสุ่ม (Randomization) ไม่รับประกันความปลอดภัยของแบบสอบ ส่วนวิธีอื่นๆ อัตราการใช้ข้อสอบซ้ำสูงสุดโดยเฉลี่ยอยู่ในระดับยอมรับได้ และเมื่อขนาดของคลังข้อสอบที่ใหญ่ขึ้นเป็น 2 เท่า ไม่เกิดผลดีกับ วิธี SH และ SL ในการใช้คลังข้อสอบที่มีขนาดใหญ่เพื่อลดอัตราการทับซ้อนของแบบสอบให้ต่ำที่สุด ส่วน วิธี DP และ SLC มีข้อได้เปรียบเมื่อขนาดของคลังข้อสอบใหญ่ขึ้นสามารถจัดการกับปัญหาการสูญเสียความถูกต้องแม่นยำในการวัดดีกว่าวิธีอื่นๆ และ 3) ผลของอัตราการใช้ข้อสอบซ้ำสูงสุดโดยจำกัดอัตราการใช้ข้อสอบซ้ำอยู่ที่ระดับ 0.1 พบว่า ประสิทธิภาพของวิธีการที่นำมาศึกษาให้ผลไม่เป็นที่น่าพอใจ และเมื่อผ่อนคลายข้อจำกัดของอัตราการใช้ข้อสอบซ้ำสูงสุดเป็น 0.2 ผลของการผ่อนคลายข้อจำกัดทำให้มีผลต่อผลของการทับซ้อนของแบบสอบ วิธี SH และ SL มีแนวโน้มที่จะเสียเปรียบมากกว่าเมื่อผ่อนคลายอัตราการใช้ข้อสอบซ้ำสูงสุด วิธี DP

Leung, Chang และ Hau (2003) ศึกษาเปรียบเทียบการจัดสมดุลเนื้อหา 3 วิธี 1) การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ร่วมกับการควบคุมสมดุลเนื้อหา (Constrained Computerized Adaptive Testing: CCAT) 2) รูปแบบมัลติโนเมียลที่ถูกปรับปรุง (Modified Multinomial Model: MMM) และ 3) การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ร่วมกับการควบคุมสมดุลเนื้อหาประยุกต์ (Modified Constrained Computerized Adaptive Testing: MCCAT) โดยการนำมารวมกับกลยุทธ์ในการแบ่งชั้น (Stratification Strategy) จำนวน 3 วิธี ได้แก่ 1) การแบ่งคลังข้อสอบเป็นชั้นตามค่าอำนาจจำแนก(multistage a-Stratified design: ASTR) 2) การแบ่งคลังข้อสอบเป็นชั้นตามค่าความยากและอำนาจจำแนก (a-Stratified with b-blocking: BASTR) 3) การแบ่งคลังข้อสอบเป็นชั้นตามค่าความยาก อำนาจจำแนก และขอบเขตเนื้อหา (multiple Stratification: CBASTR) ผลการศึกษาพบว่า วิธีการคัดเลือกข้อสอบทั้งหมดให้ผลในเรื่องของความแม่นยำในการวัดและการประมาณค่า MSE ใกล้เคียงกัน ซึ่งบ่งชี้ว่าวิธีการทั้งหมดสามารถเปรียบเทียบกันได้ในเรื่องของประสิทธิภาพในการวัด และพบว่า วิธีการจัดสมดุลของเนื้อหามีอิทธิพลอย่างมากต่อการใช้คลังข้อสอบ (η^2 =.87) และเมื่อนำเทคนิคการจัดสมดุลเนื้อหาและการแบ่งชั้นคลังข้อสอบมาทำงานร่วมกัน พบว่าว่าวิธี CBASTR เมื่อใช้งานร่วมกับวิธี MMM ให้ผลการทดสอบดีที่สุดในเรื่องของประสิทธิภาพในการวัด และการใช้คลังข้อสอบ (Pool Utilization) รวมทั้งเพิ่มการใช้ข้อสอบซ้ำในข้อที่ถูกเรียกใช้น้อยครั้งได้ดีอีกด้วย

Belov และ Armstrong (2005) พบว่าการโปรแกรมเชิงตัวเลขแบบผสม (mixed-integer Programming: MIP) ในการรวบรวมข้อสอบเป็นแบบสอบ (Test Assembly) มีข้อจำกัดคือแบบสอบที่ได้มีความลำเอียงตามฟังก์ชันเป้าหมาย (Objective Function) ดังนั้นจึงไม่สามารถที่จะสร้างแบบสอบที่มีความเท่าเทียมกันได้ (ในที่นี้หมายถึงการกระจายของแบบสอบที่มีลักษณะเป็น Uniform) ด้วยเหตุนี้ Belov และ Armstrong จึงเสนอขั้นตอนวิธี (Algorithm) ใหม่ในการรวมแบบสอบ (Test Assembly) ขั้นตอนวิธีนี้ใช้พื้นฐานของการค้นหาอย่างสุ่ม (Random Search) แบบมอนติ คาร์โล จึงถูกเรียกว่า การรวมแบบสอบแบบมอนติ คาร์โล เพื่อรวมแบบสอบมาตรฐาน (Standardized Tests) โดยใช้คุณสมบัติของข้อบังคับ (Constraint) ที่กำหนดขึ้นเป็นแนวทางในการค้นหา วิธีนี้มีจุดเด่นเหนือวิธีอื่น คือ 1) ดำเนินการสุ่มอย่างเท่าเทียมกัน (Uniform Sampling) จากคลังข้อสอบ 2) ขั้นตอนวิธีนี้สามารถนำไปประยุกต์ใช้กับการทดสอบที่มีความซับซ้อนมากโดยไม่ยุ่งยาก 3) ขณะดำเนินการค้นหาแนวทางการแก้ปัญหา หรือ ผลลัพธ์ที่เป็นไปได้ วิธีการมอนติ คาร์โลจะสร้างข้อมูลความถี่ที่สามารถใช้ประโยชน์ในการประเมินความสามารถในการใช้งานของข้อสอบและความยากของข้อจำกัดในการรวมแบบสอบ องค์กรการทดสอบสามารถใช้วิธีการนี้เพื่อประเมิน

คลังข้อสอบและประเมินความต้องการจำเป็นเพื่อเพิ่มประสิทธิภาพของการใช้งานคลังข้อสอบและลดจำนวนของข้อสอบใหม่ที่จะต้องพัฒนาในอนาคต

Belov, Armstrong, และ Weissman (2008) ได้เสนอวิธีการใหม่สำหรับการควบคุมเนื้อหา (Content Constraints) โดยทำการเปรียบเทียบวิธี Shadow CAT กับวิธี Monte Carlo CAT ที่ใช้ควบคุมเนื้อหา การศึกษาในครั้งนี้ใช้คลังข้อสอบจากการสอบเข้าศึกษาต่อในด้านกฎหมาย (Law School Admission Test: LSAT) ผลการศึกษาพบว่า 1) Monte Carlo CAT มีอัตราการใช้ข้อสอบซ้ำสูงสุดต่ำกว่า และมีการใช้ประโยชน์ของคลังข้อสอบดีกว่า โดยเฉพาะอย่างยิ่งอัตราการใช้ข้อสอบซ้ำสูงสุดสำหรับ Monte Carlo CAT นั้นมีค่าต่ำกว่าวิธี Shadow CAT 2) เมื่อมีการเดาเกิดขึ้นทั้ง Monte Carlo CAT และ Shadow CAT มี Bias เพิ่มขึ้นทั้งคู่ และ MSE ของ Monte Carlo CAT แทบจะไม่มีการเปลี่ยนแปลง ขณะที่ของ Shadow CAT เพิ่มขึ้น อย่างไรก็ตามการเปลี่ยนแปลงเหล่านี้ไม่ได้เพิ่มขึ้นมาก และ3) Monte Carlo CAT สามารถรองรับผู้สอบทั้งหมด 2,723 คนที่ทำข้อสอบพร้อมๆ กัน ไม่เกิน 5 วินาทีในการนำข้อสอบออกมาใช้ ขณะที่ Shadow CAT ใช้เวลามากกว่า 20 นาที กับผู้สอบเพียง 134 คน โดยสรุปแล้ว Monte Carlo CAT ให้ผลดีกว่าอย่างเห็นได้ชัดทั้งในเรื่อง อัตราการใช้ข้อสอบซ้ำ Bias และ MSE ดังนั้น Monte Carlo CAT จึงเหมาะสมกับการนำไปใช้ในทางปฏิบัติเพื่อบริหารการสอบ

Cheng และ Chang (2009) เสนอวิธีดัชนีลำดับความสำคัญสูงสุด (Maximum Priority Index) สำหรับการคัดเลือกข้อสอบที่มีข้อบังคับอย่างเข้มงวด การวิจัยครั้งนี้การศึกษาผ่านการจำลองข้อมูลเพื่อเปรียบเทียบวิธีการจัดการข้อบังคับ 2 วิธี ได้แก่ 1)วิธีดัชนีลำดับความสำคัญสูงสุด (Maximum Priority Index: MPI) 2) วิธี Weighted Deviation Modeling (WMD) โดยใช้วิธีการคัดเลือกข้อสอบแบบ Maximum Information และวิธี Randomized เป็นฐานในการเปรียบเทียบ โดยใช้ข้อมูลจากคลังข้อสอบของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ของการสอบเพื่อจัดตำแหน่งในมหาวิทยาลัย (College Placement CAT) ในวิชาพื้นฐานพีชคณิต คลังข้อสอบประกอบด้วยข้อสอบจาก 3 กลุ่มเนื้อหาวิชา แต่ละกลุ่มเนื้อหาวิชาแบ่งออกเป็น 4-10 ขอบเขตเนื้อหา ผลการศึกษาพบว่า 1) วิธี MPI และวิธี Weighted Deviation Modeling (WDM) ลดการฝ่าฝืนข้อบังคับในการเปรียบเทียบกับวิธี MI และ Randomized และผลการทดสอบแสดงให้เห็นอย่างชัดเจนว่า MPI มีประสิทธิภาพดีกว่า WDM ทุกด้าน โดยมีมีอัตราการใช้ข้อสอบซ้ำสูงสุด การใช้ข้อสอบซ้ำเกิน และอัตราการทับซ้อนของแบบสอบต่ำกว่าวิธี WDM นอกจากนี้ยังพบว่าทุกวิธียกเว้นวิธี Randomized มีสัดส่วนของ Never Exposed สูงมากกว่า 50% แสดงว่าข้อสอบในคลังข้อสอบถูกใช้อย่างไม่มีประสิทธิภาพ ซึ่งปัญหานี้สามารถทำให้ลดลงได้โดยการนำไปรวมเข้ากับวิธี a-Stratified Design ของ (Chang & Ying, 1999) จากข้อค้นพบของการศึกษาครั้งนี้ วิธี MPI นี้

สามารถปรับให้เข้ากับข้อบังคับที่ไม่ใช่ทางสถิติที่หลากหลายและสามารถควบคุมได้หลายอย่างพร้อมๆ กัน เช่น ความสมดุลของเนื้อหา การควบคุมการใช้ข้อสอบซ้ำ ความสมดุลของคำตอบ โดยที่วิธีการนี้สามารถนำไปปรับใช้กับโปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่มีอยู่ในปัจจุบันได้สะดวก โดยไม่จำเป็นต้องปรับค่าน้ำหนักความสัมพันธ์ระหว่างข้อบังคับและสารสนเทศ

Cheng และคณะ (2009) นำวิธี MPI มาใช้ร่วมกับการแบ่งชั้นพารามิเตอร์อำนาจจำแนก (a-Stratification) จึงถูกเรียกชื่อใหม่ว่าวิธีการคัดเลือกข้อสอบแบบแบ่งค่าอำนาจจำแนกที่มีการถ่วงน้ำหนักเงื่อนไขบังคับ (Constraint-Weighted a-Stratification: CWASRT) ผลการศึกษาพบว่า วิธี CWASRT มีประสิทธิภาพมากในเรื่องของความสมดุลของเนื้อหาแต่มีอัตราความแปรปรวนของความคลาดเคลื่อนในการประมาณค่าความสามารถมากกว่าวิธีการคัดเลือกข้อสอบที่ให้สารสนเทศสูงสุด แต่วิธีการคัดเลือกข้อสอบที่ให้สารสนเทศสูงสุดและวิธีการคัดเลือกอย่างสุ่มมีการฝ่าฝืนในเรื่องของข้อบังคับด้านความสมดุลของเนื้อหา นอกจากนี้ยังพบว่า การจัดเรียงข้อสอบในแต่ละชั้น ตามวิธีการ Ascending-a ดีกว่า วิธีการ Descending-a ในด้านประสิทธิภาพในการประมาณค่าความสามารถแต่ด้านประสิทธิภาพของการควบคุมการใช้ข้อสอบซ้ำมีประสิทธิภาพดีใกล้เคียงกัน ซึ่งสนับสนุนผลการวิจัยของ Hau และ Chang (2001) และ Chang และ Ying (1999) ที่การศึกษาพบว่า วิธี a-STR และ วิธี a-DSTR ให้ความคลาดเคลื่อนกำลังสองเฉลี่ยและความลำเอียงเฉลี่ยมีแนวโน้มลดลงเหมือนกันเมื่อความยาวแบบทดสอบเพิ่มขึ้น แต่ความคลาดเคลื่อนกำลังสองเฉลี่ยในวิธี a-STR ต่ำกว่าวิธี a-DSTR

โดยสรุปแล้วแนวโน้มของการทดสอบแบบปรับเหมาะกับความสามารถของผู้สอบด้วยคอมพิวเตอร์ในยุคปัจจุบันให้ความสำคัญเรื่องการควบคุมเงื่อนไขบังคับอื่นๆ เช่น การสร้างความสุดดุลของเนื้อหา การควบคุมการใช้ข้อสอบซ้ำ นอกเหนือ เป็นต้น นักวิชาการด้านการทดสอบพยายามคิดค้นและพัฒนาวิธีการคัดเลือกข้อสอบให้สามารถนำไปใช้ในสถานการณ์จริงได้อย่างเหมาะสม โดยการศึกษาผ่านสถานการณ์จำลอง เพื่อให้เห็นจุดเด่นของวิธีการคัดเลือกแบบต่าง ๆ เมื่อนำไปใช้ในสถานการณ์ที่แตกต่างกัน และจากการศึกษางานวิจัยในอดีตพบว่า วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ แบบมอนติ คาร์โล (Monte Carlo CAT Method) ที่เสนอโดย Belov, Armstrong, และ Weissman (2008) และวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์วิธีแบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (constraint-weighted a-stratification method for CAT) ที่เสนอโดย Cheng และคณะ (2009) เป็นวิธีการคัดเลือกข้อสอบที่มีจุดเด่นเหนือวิธีการอื่นในด้านของประสิทธิภาพในการประมาณค่าความสามารถ และประสิทธิภาพในการควบคุมเงื่อนไขบังคับที่ไม่ใช่เงื่อนไขบังคับทางสถิติ เช่น การควบคุมการใช้ข้อสอบซ้ำ และการจัดสมดุลของเนื้อหา แต่การศึกษาดังกล่าวทำการศึกษาภายใต้โมเดล IRT และผู้สอบทุกคนได้รับแบบสอบที่มีความ

ยาวเท่ากัน เพื่อขยายองค์ความรู้ด้านการทดสอบผู้วิจัยจึงนำวิธีการคัดเลือกข้อสอบทั้งสองวิธีมาศึกษาภายใต้เงื่อนไขที่ผู้สอบจะได้รับแบบสอบที่มีความยาวแตกต่างกันขึ้นอยู่กับความคลาดเคลื่อนในการประมาณค่า และศึกษาภายใต้โมเดล TRT ซึ่งผู้สอบแต่ละคนจะได้รับแบบทดสอบย่อย ที่ใช้สิ่งเร้าร่วมกันคล้ายกับ Multistage CAT แต่แตกต่างกันตรงที่การวิจัยครั้งนี้นำเอาจุดเด่นของการเลือกข้อสอบแบบแบบมอนติ คาร์โล และแบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับมาใช้ โดยผู้วิจัยไม่ได้กำหนดชุดของแบบทดสอบและเส้นทางในการตอบข้อสอบไว้ล่วงหน้าเป็นพาเนลตามระดับความง่าย-ยาก ของชุดข้อสอบเหมือนกับ Multistage CAT

6.5 การพิจารณาประสิทธิภาพของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์

จากการศึกษางานวิจัยในอดีตพบว่าการประเมินประสิทธิภาพของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ ส่วนใหญ่พิจารณาเป็นสองประเด็นหลักๆ คือ 1) ประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถของผู้สอบ (Measurement Precision) และ 2) ประสิทธิภาพของการใช้คลังข้อสอบ (Efficiency of Pool Utilization หรือ Efficiency of item bank usage) โดยแต่ละด้านมีการเลือกดัชนีที่ใช้พิจารณาประสิทธิภาพแตกต่างกันโดยมีรายละเอียดดังนี้

1. ประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถของผู้สอบ (Measurement Precision) จากการสังเคราะห์เอกสารและงานวิจัย พบว่า ตัวบ่งชี้ในการประเมินความถูกต้องแม่นยำของการประมาณค่าความสามารถของผู้สอบ คือ 1) ค่า Fidelity (สหสัมพันธ์ระหว่าง $\theta, \hat{\theta}$) 2) ความลำเอียง (BIAS) 3) ความคลาดเคลื่อนกำลังสองเฉลี่ยหรือความแปรปรวนของความคลาดเคลื่อน (MSE) 4) ความยาวเฉลี่ยของแบบสอบ (ใช้ในกรณีที่ผู้สอบแต่ละคนได้รับแบบสอบที่มีความยาวแตกต่างกันแต่มีความคลาดเคลื่อนมาตรฐานของการประมาณค่า (SEE) ใกล้เคียงกัน) 5) ความคลาดเคลื่อนมาตรฐานของการประมาณค่า (SEE) ใช้ในกรณีที่ผู้สอบได้รับแบบสอบที่มีความยาวเท่ากัน

2. ประสิทธิภาพของการใช้คลังข้อสอบ (Efficiency of Pool Utilization หรือ Efficiency of item bank usage) จากการสังเคราะห์เอกสารและงานวิจัย พบว่าเป็นความสามารถของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ในการใช้คลังข้อสอบได้อย่างสมดุล ข้อสอบแต่ละข้อมีโอกาสถูกนำออกมาใช้เท่าๆ กัน ตัวบ่งชี้ในการประเมินประสิทธิภาพของการใช้คลังข้อสอบจากงานวิจัยในอดีต ได้แก่ 1) อัตราการใช้ข้อสอบซ้ำที่สังเกตได้ 2) อัตราการทับซ้อนของแบบสอบสอบเฉลี่ย 3) การทดสอบความแตกต่างของการแจกแจงของอัตราการใช้ข้อสอบซ้ำที่สังเกตได้กับอัตราการใช้ข้อสอบซ้ำที่คาดหวัง (x^2) 4) จำนวนข้อสอบที่ไม่ถูกนำมาใช้ และ 5) จำนวนข้อสอบที่ถูกนำมาใช้มากเกินไป

ตารางที่ 2.3 การสังเคราะห์งานวิจัยที่เกี่ยวข้องกับวิธีการควบคุมเงื่อนไขบังคับในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์

ผู้วิจัย	สิ่งที่ศึกษา	ความยาวแบบสอบ	ขนาดคลังข้อสอบ	โมเดล	ตัวอย่าง	Estimation	เกณฑ์การพิจารณา
Revuelta and Ponsoda (1998)	เปรียบเทียบวิธีการคัดเลือกข้อสอบ 6 วิธี Maximum Information Method (MI) One Parameter Method (1P) McBride and Martin Method (MM) Randomesque Method (RA) Sympson and Hetter Method (SH) **Restricted Maximum Information Method (Rk;k=40)** **Progressive Method (PR)**	(a) fixed-test length (35 items) (b) variable-test length (stopping rule SE <= 0.22, or max length = 50 items)	Study 1 221 Study 2 500, 1000	IRT 3PL	2,000	MLE	BAIS, SE, ค่าเฉลี่ยของความยาวแบบสอบ, ร้อยละของจำนวนข้อสอบที่ถูกใช้, จำนวนข้อสอบที่ไม่ถูกใช้
Chang and Ying (1999)	เปรียบเทียบวิธีการควบคุมการแสดงข้อสอบ **a-Stratified Method** SH based on maximizing item Fisher information (FSH) SH based on Bayesian item selection (BSH)	40 และ 60 ข้อ	400 ข้อ แบ่งคลังข้อสอบเป็น 4 ชั้น	IRT 3PL	3,000	MLE	Bias, MSE, exposure rates (er), Test Overlap rate , Pearson's x^2

ตารางที่ 2.3 การสังเคราะห์งานวิจัยที่เกี่ยวข้องกับวิธีการควบคุมเงื่อนไขบังคับในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (ต่อ)

ผู้วิจัย	สิ่งที่ศึกษา	ความยาวแบบสอบ	ขนาดคลังข้อสอบ	โมเดล	ตัวอย่าง	Estimation	เกณฑ์การพิจารณา
Hau and Chang (2001)	เปรียบเทียบวิธีคัดเลือกข้อสอบ **Ascending a-STR (ASTR)** Descending a-STR (DSTR) SH based on maximizing item Fisher information (FSH)	40 ข้อ	400 ข้อ	IRT 3PL	3,000	MLE	Bias, MSE, จำนวนครั้งของข้อสอบที่ถูกเรียกใช้
Leung et al. (2002)	เปรียบเทียบวิธีการควบคุมการแสดงข้อสอบโดยปรับปรุงวิธี a-Stratified Method (STR) เพื่อนำมาใช้ร่วมกับ Sympson and Hetter Method (SH) วิธีการที่นำมาศึกษามีดังนี้ **Method 1 (STR)** **Method 2 (STR-SH)** Method 3 (Max-I-SH)	Study 1, 3 40, 60 ข้อ Study 2, 4 24 ข้อ	Simulate item 400 operational item 254	IRT 3PL	3,000	MLE	Fidelity (สหสัมพันธ์ระหว่าง $\theta, \hat{\theta}$), Bias, MSE, exposure rates (er), Test Overlap rate , Pearson's x^2, No. of underutilized items, No. of overexposed items, computing time

ตารางที่ 2.3 การสังเคราะห์งานวิจัยที่เกี่ยวข้องกับวิธีการควบคุมเงื่อนไขบังคับในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (ต่อ)

ผู้วิจัย	สิ่งที่ศึกษา	ความยาวแบบสอบ	ขนาดคลังข้อสอบ	โมเดล	ตัวอย่าง	Estimation	เกณฑ์การพิจารณา
Chang and Ansley (2003)	เปรียบเทียบวิธีการควบคุมการแสดงข้อสอบเมื่อขนาดคลังข้อสอบ และอัตราการแสดงข้อสอบสูงสุดต่างกัน 5-4-3-2-1 technique (MM), Sympson and Hetter (SH), Davey and Parshall (DP) ,Stocking and Lewis , unconditional multinomial (SL) **Stocking and Lewis conditional multinomial (SLC)**	30, 60 ข้อ	360, 720 ข้อ r_{max} = 0.1, 0.2	IRT 3PL	3,000 * 17 (point of theta)	MLE	SEE, exposure rates (er), Test Overlap rate , Scaled chi-square x^2, No. of underutilized items, No. of overexposed items, computing time
Leung et al. (2003)	นำวิธีการควบคุมความสมดุลของเนื้อหา constrained CAT (CCAT), modified multinomial model (MMM), modified CCAT (MCCAT) มาใช้ร่วมกับวิธี a-Stratified Method (STR)	35 ข้อ	700 ข้อ 4 เนื้อหา	IRT 3PL	5,000	MLE	Fidelity, MSE, Number of overexposed items, Number of underutilized items, Scaled chi-square x^2, Test Overlap rate

ตารางที่ 2.3 การสังเคราะห์งานวิจัยที่เกี่ยวข้องกับวิธีการควบคุมเงื่อนไขบังคับในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (ต่อ)

ผู้วิจัย	สิ่งที่ศึกษา	ความยาวแบบสอบ	ขนาดคลังข้อสอบ	โมเดล	ตัวอย่าง	Estimation	เกณฑ์การพิจารณา
Belov and Armstrong (2005)	เสนอขั้นตอนวิธีในการรวมแบบสอบโดยใช้เทคนิคการค้นหาแบบ stochastic และนับจำนวนแบบสอบที่ผิดเงื่อนไข TIF	ขึ้นอยู่กับ TIF	2,418 2,492	IRT 3PL	-	-	Number of Violations of Information Targets
Belov et al. (2008)	เปรียบเทียบวิธี Shadow CAT กับวิธี Monte Carlo CAT	50	500	IRT 3PL	9,000	EAP	Bias, MSE, exposure rates (er), computing time
Cheng. and Chang. (2009)	เสนอวิธีดัชนีลำดับความสำคัญสูงสุด (Maximum Priority Index)	12	642 3 เนื้อหา	IRT 3PL	3,000* 19 (point of theta)	ช่วงต้นการทดสอบใช้ EAP จากนั้นจึงใช้ MLE	Fidelity, Bias, MSE, Test Overlap rate
Cheng. et al. (2009)	เสนอวิธีการคัดเลือกข้อสอบแบบแบ่งค่าอำนาจจำแนกที่มีการถ่วงน้ำหนักเงื่อนไขบังคับ	40	1000 4 เนื้อหา	IRT 3PL	3,000	ช่วงต้นการทดสอบใช้ EAP จากนั้นจึงใช้ MLE	Bias, MSE, exposure rates (er), Scaled chi-square x^2

จากการทบทวนเอกสารและงานวิจัยในอดีต พบว่า ตัวแปรที่นักวิชาการด้านการทดสอบนำมาใช้เป็นเกณฑ์พิจารณาประสิทธิภาพของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ด้วยการจำลองข้อมูล ได้แก่ Bias, MSE, สหสัมพันธ์ระหว่างความสามารถประมาณค่ากับความสามารถจริง (Fidelity), ความยาวของแบบสอบ (เมื่อมีความแม่นยำในการวัดใกล้เคียงกัน), ความคลาดเคลื่อนมาตรฐานของการประมาณค่า (SEE), อัตราการใช้ข้อสอบซ้ำ (exposure rates: er), อัตราการทับซ้อนของแบบสอบ (Test Overlap rate), ค่าไคสแควร์ (Pearson's x^2), จำนวนของข้อสอบที่ไม่ถูกนำมาใช้ (No. of underutilized items), จำนวนของข้อสอบที่นำมาใช้มากเกินไป (No. of overexposed items) เมื่อพิจารณาจะพบว่าตัวแปรที่กล่าวถึงสามารถจำแนกออกเป็น 2กลุ่มใหญ่ๆ คือ 1) กลุ่มที่เกี่ยวข้องกับการประมาณค่าความสามารถของผู้สอบ 2) กลุ่มที่เกี่ยวข้องกับการใช้คลังข้อสอบ ดังเสนอในตารางที่ 2.4 ทำให้ผู้วิจัยสามารถสรุปกรอบแนวคิดในการวิจัยได้ดังภาพที่ 2.6

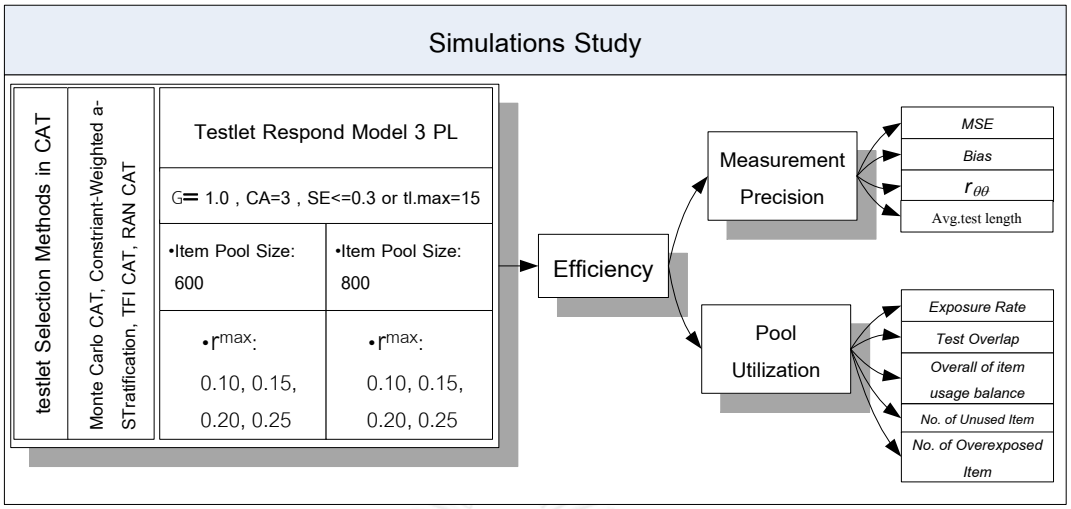
ตารางที่ 2.4 การจัดประเภทตัวแปรสำหรับพิจารณาประสิทธิภาพของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ด้วยการจำลองข้อมูล

ตัวแปรที่เกี่ยวข้องกับการประมาณค่าความสามารถของผู้สอบ	ตัวแปรที่เกี่ยวข้องกับการใช้คลังข้อสอบ
1. Bias (Belov et al., 2008; Chang & Ying, 1999; Cheng & Chang, 2009; Cheng et al., 2009; Hau & Chang, 2001; Leung et al., 2002, 2003; Revuelta & Ponsoda, 1998)	1. อัตราการใช้ข้อสอบซ้ำ (Belov et al., 2008; Chang & Ying, 1999; Chang & Ansley, 2003; Cheng et al., 2009; Hau & Chang, 2001; Leung et al., 2002, 2003)
2. MSE (Belov et al., 2008; Chang & Ying, 1999; Cheng & Chang, 2009; Cheng et al., 2009; Hau & Chang, 2001; Leung et al., 2002, 2003; Revuelta & Ponsoda, 1998)	2. อัตราการทับซ้อนของแบบสอบ (Chang & Ying, 1999; Chang & Ansley, 2003; Cheng & Chang, 2009; Leung et al., 2002, 2003)
3. Fidelity (Cheng & Chang, 2009; Leung et al., 2002, 2003)	3. ค่าไคสแควร์ (Belov et al., 2008; Chang & Ying, 1999; Chang & Ansley, 2003; Cheng et al., 2009; Hau & Chang, 2001; Leung et al., 2002, 2003)
4. ความยาวของแบบสอบ: เมื่อ SEE มีขนาดใกล้เคียงกัน (Chang & Ansley, 2003; Revuelta & Ponsoda, 1998; Wainer, 1992)	4. จำนวนของข้อสอบที่ไม่ถูกนำมาใช้ และ 5. จำนวนของข้อสอบที่นำมาใช้มากเกินไป (Chang & Ansley, 2003; Leung et al., 2002, 2003; Revuelta & Ponsoda, 1998)

กรอบแนวคิดที่ใช้ในการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาประสิทธิภาพของวิธีการคัดเลือกข้อสอบ 2 วิธีระหว่างวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล และวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ สำหรับการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่ใช้โมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย แบบ 3 พารามิเตอร์ (3 Parameter Logistic Testlet Response Theory Model: 3 PL TRT) โดยผู้วิจัยกำหนดความแปรปรวนของอิทธิพลของแบบทดสอบย่อย (testlet effect) ให้ค่าเท่ากับ 1 เพื่อสะท้อนว่าข้อสอบภายในแบบทดสอบย่อยมีความสัมพันธ์กับสิ่งเร้าสูงและใช้แบบทดสอบย่อยเป็นขนาดที่เล็ก คือ มีข้อสอบภายในแบบทดสอบย่อยจำนวน 4 ข้อ (Wainer, Bradlow, & Wang, 2007) และการศึกษาครั้งนี้ใช้วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่เลือกแบบทดสอบย่อยที่ให้สารสนเทศสูงสุด และวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่เลือกแบบทดสอบย่อยอย่างสุ่ม ผู้วิจัยใช้เป็นฐานในการเปรียบเทียบเนื่องจากมีประสิทธิภาพในด้านแม่ยำของการวัดสูงที่สุดและมีประสิทธิภาพในด้านสมดุลการใช้คลังข้อสอบสูงที่สุดตามลำดับ

โดยมีเงื่อนไขในการศึกษา คือ คลังข้อสอบ มี 2 ขนาด ประกอบด้วย คลังข้อสอบขนาด 600 ข้อ และ คลังข้อสอบขนาด 800 ข้อ อัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด 4 ระดับ คือ 10, 15, 20 และ 25 เปอร์เซ็นต์ และใช้ความคลาดเคลื่อนมาตรฐานในการประมาณค่าความสามารถ (SEE) ที่มีค่าน้อยกว่าหรือเท่ากับ 0.3 เนื่องจากให้ความตรงตามสภาพสูงเหมาะสำหรับการทดสอบที่มีการแข่งขันสูง (รังสรรค์ มณีเล็ก, 2540; สิริลักษณ์ เกษรปทุมานันท์ & ณัฏฐภรณ์ หลาวทอง, 2007) หรือ ความยาวของแบบสอบ มากกว่า 60 ข้อ เป็นเกณฑ์การยุติการทดสอบ โดย เกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ทั้ง 2 แบบ พิจารณาจากประสิทธิภาพด้านความแม่นยำของการวัด พิจารณาได้จากเกณฑ์ ต่อไปนี้ 1) ค่าความลำเอียง (bias) 2) ค่าความแปรปรวนของคลาดเคลื่อน (MSE) 3) ค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) ระหว่างค่าความสามารถจริงและค่าความสามารถที่ประมาณค่าของผู้สอบ และ 4) ความยาวของแบบสอบ (เมื่อ SEE มีขนาดใกล้เคียง) ประสิทธิภาพด้านการใช้คลังข้อสอบ พิจารณาได้จากเกณฑ์ ต่อไปนี้ 1) อัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ 2) อัตราการทับซ้อนของแบบสอบ 3) ความสมดุลในการใช้คลังข้อสอบ 4) จำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ และ 5) จำนวนแบบทดสอบย่อยที่ถูกนำมาใช้มากเกินไป โดยสามารถแสดงกรอบแนวคิดในการวิจัยดังแผนภาพ 2.6



ภาพที่ 2.6 กรอบแนวคิดในการวิจัย

บทที่ 3
วิธีดำเนินการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาประสิทธิภาพของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่ใช้สำหรับโมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย 2 วิธี คือ 1) วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที (Monte Carlo CAT Method) และ 2) วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (constraint-weighted a-stratification CAT method) เมื่อกำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด และขนาดคลังข้อสอบที่ต่างกัน การวิจัยครั้งนี้เป็นการวิจัยเชิงทดลองโดยใช้เทคนิควิธีมอนติ คาร์โล ซีเอที (Monte Carlo Method) คือ ใช้เครื่องคอมพิวเตอร์ในการสุ่มตัวอย่าง จำลองผลการตอบ และประมาณค่าความสามารถของผู้สอบจากการทดสอบแบบปรับเหมาะ ซึ่งมีขั้นตอนการดำเนินงานวิจัยดังต่อไปนี้

กลุ่มตัวอย่าง

กลุ่มตัวอย่างสำหรับการศึกษาในครั้งนี้ คือ พารามิเตอร์ผู้สอบ ได้แก่ 1) ค่าความสามารถจริง (θ_i) ที่สร้างขึ้นโดยสุ่มการแจกแจงโค้งปกติมาตรฐาน ซึ่งมีค่าเฉลี่ยเท่ากับ 0 และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 1 [θ ~N(0,1)] และกำหนดให้มีค่าตั้งแต่ -3.0 ถึง 3.0 และ 2) ค่าอิทธิพลของแบบทดสอบย่อยของผู้สอบแต่ละคนที่มีต่อข้อสอบภายแบบทดสอบย่อยแต่ละชุด สำหรับการทดสอบในแต่ละเงื่อนไขผู้วิจัยกำหนดผู้เข้าสอบไว้ที่ 1,000 คน ดังนั้นในการทดสอบหนึ่งครั้งจะต้องมีพารามิเตอร์ผู้สอบจำนวน 1,000 ชุด เมื่อเปลี่ยนเงื่อนไขการทดสอบ ระบบจะทำการสุ่มค่าพารามิเตอร์ผู้สอบใหม่ทุกครั้ง โดยในการทดลองแต่ละเงื่อนไขจะสุ่มตัวอย่างเพื่อนำมาทำซ้ำ 10 รอบ (Boyd, 2003; Keng, 2008) ดังนั้น ตัวอย่างที่ใช้ในการทดลองแต่ละเงื่อนไขมีจำนวน 10,000 คน และการศึกษาครั้งนี้มีเงื่อนไขที่ศึกษาทั้งหมด 20 เงื่อนไข ดังนั้นจึงใช้จำนวนตัวอย่างทั้งหมด 200,000 ตัวอย่าง รายละเอียดดังตารางที่ 3.1

ตารางที่ 3.1 เงื่อนไขในการทดสอบ

เงื่อนไข	วิธีการคัดเลือกแบบทดสอบย่อย	อัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด	ขนาดคลังข้อสอบ	จำนวนตัวอย่าง	การทำซ้ำ (รอบ)	รวมตัวอย่าง
1	TFI	-	600	1,000	10	10,000
2	RAN	-	600	1,000	10	10,000
3	TFI	-	800	1,000	10	10,000
4	RAN	-	800	1,000	10	10,000
5	MCC	10%	600	1,000	10	10,000
6	MCC	15%	600	1,000	10	10,000
7	MCC	20%	600	1,000	10	10,000
8	MCC	25%	600	1,000	10	10,000
9	CWA	10%	600	1,000	10	10,000
10	CWA	15%	600	1,000	10	10,000
11	CWA	20%	600	1,000	10	10,000
12	CWA	25%	600	1,000	10	10,000
13	MCC	10%	800	1,000	10	10,000
14	MCC	15%	800	1,000	10	10,000
15	MCC	20%	800	1,000	10	10,000
16	MCC	25%	800	1,000	10	10,000
17	CWA	10%	800	1,000	10	10,000
18	CWA	15%	800	1,000	10	10,000
19	CWA	20%	800	1,000	10	10,000
20	CWA	25%	800	1,000	10	10,000

ตัวแปรที่ใช้ในการศึกษา

เนื่องจากการวิจัยครั้งนี้มุ่งเปรียบเทียบประสิทธิภาพของวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ เงื่อนไขที่ทำการศึกษา คือ วิธีการคัดเลือกข้อสอบ 2 วิธี ขนาดของคลังข้อสอบที่แตกต่างกัน 2 ขนาด และอัตราการใช้แบบทดสอบย่อยซ้ำสอบสูงสุดที่แตกต่างกัน 4 ระดับ มีรายละเอียดดังต่อไปนี้

1. ตัวแปรอิสระ

1.1 วิธีการเลือกแบบทดสอบย่อยในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์

1) วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล (Monte Carlo CAT Method)

2) วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (constraint-weighted a-stratification method)

3) วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบค่าสารสนเทศสูงสุด (Maximum Fisher's information method) ผู้วิจัยใช้เป็นฐานในการเปรียบเทียบเนื่องจากมีประสิทธิภาพในด้านความถูกต้องแม่ยำของการวัดสูงที่สุด

4) วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบสุ่ม (Randomization method) ผู้วิจัยใช้เป็นฐานในการเปรียบเทียบเนื่องจากมีประสิทธิภาพในด้านการใช้คลังข้อสอบสูงที่สุด

1.2 ขนาดคลังข้อสอบ มี 2 ขนาด คือ

1) คลังข้อสอบขนาด 600 ข้อ

2) คลังข้อสอบขนาด 800 ข้อ

1.3 อัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดมี 4 ระดับ คือ

1) กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดไว้ที่ 10 เปอร์เซ็นต์

2) กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดไว้ที่ 15 เปอร์เซ็นต์

3) กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดไว้ที่ 20 เปอร์เซ็นต์

4) กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดไว้ที่ 25 เปอร์เซ็นต์

2. ตัวแปรตาม

2.1 ประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถของผู้สอบ (Efficiency of Measurement Precision) เกณฑ์ที่ใช้พิจารณาได้แก่ 1) ความลำเอียงหรือความคลาดเคลื่อน (Bias) 2) ความแปรปรวนของคลาดเคลื่อน (MSE) 3) สหสัมพันธ์ระหว่างค่าความสามารถจริงและค่าประมาณความสามารถของผู้สอบ และ 4) ความยาวเฉลี่ยของแบบสอบ (เมื่อมีค่าความคลาดเคลื่อนมาตรฐานของการประมาณค่าใกล้เคียงกัน)

2.2 ประสิทธิภาพด้านการใช้คลังข้อสอบ (Efficiency of Pool Utilization หรือ Efficiency of item bank usage) เกณฑ์ที่ใช้พิจารณาได้แก่ 1) อัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ 2) อัตราการทับซ้อนของแบบสอบ 3) ความสมดุลของการใช้คลังข้อสอบในภาพรวม 4) จำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ และ 5) จำนวนแบบทดสอบย่อยที่ถูกนำมาใช้มากเกินไป

ขั้นตอนในการทดลอง

การวิจัยครั้งนี้มีลำดับขั้นตอนในการทดลอง ดังนี้

1. การจำลองกลุ่มตัวอย่าง ความสามารถจริง (true ability: θ_i) และอิทธิพลที่มีต่อแบบทดสอบย่อย (testlet effect) ของตัวอย่าง

2. การจำลองคลังข้อสอบ และโครงสร้างของคลังข้อสอบ

3. การหาค่าความน่าจะเป็น (probability) ในการตอบข้อสอบถูกของกลุ่มตัวอย่างในข้อที่ 1 กับคลังข้อสอบในข้อที่ 2

4. การหาผลการตอบและการประมาณค่าพารามิเตอร์ผู้สอบ

5. การตรวจสอบความถูกต้องของการจำลองข้อมูล ว่าผลการตอบข้อสอบสามารถบ่งชี้ได้หรือไม่ว่าข้อสอบที่รวมกันอยู่ภายในแบบทดสอบย่อยไม่เป็นอิสระต่อกัน ซึ่งตรงกับเงื่อนไขในการใช้โมเดลการตอบสนองแบบทดสอยย่อยที่ผู้วิจัยต้องการศึกษา ตรวจสอบโดยใช้สถิติ Q_3 ของ Yen (1984) รายละเอียดขั้นตอนในการทดลองมีดังนี้

1. จำลองกลุ่มตัวอย่าง ความสามารถจริง (true ability) และอิทธิพลที่มีต่อแบบทดสอบย่อย (testlet effect) ของตัวอย่าง

กลุ่มตัวอย่างสำหรับการศึกษาในครั้งนี้ ได้แก่ ค่าความสามารถจริง (θ_i) ของกลุ่มตัวอย่างผู้สอบสร้างขึ้นโดยสุ่มการแจกแจงโค้งปกติมาตรฐาน ซึ่งมีค่าเฉลี่ยเท่ากับ 0 และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 1 [θ ~N(0,1)] และกำหนดให้มีค่าตั้งแต่ -3.0 ถึง 3.0 ค่าความสามารถจริงของกลุ่มตัวอย่างผู้สอบสำหรับการทดสอบแต่ละวิธีกำหนดมีจำนวนเท่ากับ 1,000 ค่า เมื่อเปลี่ยนวิธีกำหนดจะทำการสุ่มค่าความสามารถจริงของกลุ่มตัวอย่างผู้สอบใหม่ทุกครั้ง โดยในการทดลองแต่ละเงื่อนไขจะสุ่มตัวอย่างเพื่อนำมาทำซ้ำ 10 รอบ ดังนั้น ตัวอย่างที่ใช้ในการทดลอง แต่ละเงื่อนไขจะเป็น 10,000 ค่า และกำหนดพารามิเตอร์อิทธิพลของแบบทดสอบย่อยให้กับผู้สอบโดยเลือกอย่างสุ่มจากการแจกแจงปกติด้วยค่าเฉลี่ยเท่ากับศูนย์ และความแปรปรวนเท่ากับ $\gamma_{id(j)}$ [γ ~N(0, $\gamma_{id(j)}$)] (Keng, Ho, Chen, & Dodd, 2008) ในการวิจัยครั้งนี้กำหนดความแปรปรวนของ $\gamma_{id(j)}$ ให้มีค่าเป็น 1 เพื่อสะท้อนว่ามีอิทธิพลของแบบทดสอบย่อยสูง

ค่าสถิติพื้นฐานพารามิเตอร์ผู้สอบของกลุ่มตัวอย่างที่ใช้ทดลองแต่ละเงื่อนไขแสดงในตารางต่อไปนี้ ตารางที่ 3.2 และ 3.3 แสดงค่าสถิติพื้นฐาน เช่น ค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐาน ค่าต่ำสุดและค่าสูงสุด ของค่าความสามารถจริงของกลุ่มตัวอย่างบางส่วนที่ใช้ในการวิจัยที่เกิดจากการจำลองข้อมูล เช่น CWA10.1 หมายถึง กลุ่มตัวอย่างของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดไว้ เท่ากับ 10 เปอร์เซ็นต์ ในการทำซ้ำรอบที่ 1 ทำให้ผู้วิจัยทราบว่าข้อมูลที่จำลองขึ้นมีคุณสมบัติตรงตามที่กำหนดไว้ คือ ตัวอย่างแต่ละกลุ่มควรมีค่าเฉลี่ยของความสามารถจริงใกล้เคียงกับ 0 และมีส่วนเบี่ยงเบนมาตรฐานใกล้เคียงกับ 1 และกำหนดให้มีค่าต่ำสุดประมาณ -3.5 และค่าสูงสุดประมาณ 3.5 และมีการแจกแจงแบบโค้งปกติหรือไม่

ตารางที่ 3.4 และ 3.5 แสดงตัวอย่างบางส่วนจากการจำลองข้อมูลพารามิเตอร์อิทธิพลของแบบทดสอบย่อย และแต่ละแถวในตารางแสดงค่าสถิติของตัวอย่างแต่ละกลุ่ม ข้อมูลที่ได้จากการจำลองของแต่ละกลุ่มคือ ค่าพารามิเตอร์อิทธิพลของแบบทดสอบย่อยของผู้สอบซึ่งมีจำนวนเท่ากับจำนวนผู้สอบคูณด้วยจำนวนแบบทดสอบย่อยในคลังข้อสอบ จากตารางทำให้ผู้วิจัยทราบว่าข้อมูลที่จำลองขึ้นมีคุณสมบัติตรงตามที่กำหนดไว้ คือ ตัวอย่างแต่ละกลุ่มมีค่าเฉลี่ยของพารามิเตอร์อิทธิพลของแบบทดสอบย่อยใกล้เคียงกับ 0 และมีความแปรปรวนของ $\gamma_{id(j)}$ เท่ากับ 1 และมีการแจกแจงแบบโค้งปกติ

ตารางที่ 3.2 ตัวอย่างค่าต่ำสุด สูงสุด ค่าเฉลี่ย และค่าส่วนเบี่ยงเบนมาตรฐานของค่าความสามารถจริงของกลุ่มตัวอย่าง ที่ใช้กับคลังข้อสอบขนาด 600 ข้อ

กลุ่มตัวอย่าง		ความสามารถจริง (θ_i)			
	N	*M*	*SD*	*Min*	*Max*
CWA10.1	1,000	-0.02769	1.004796	-3.19549	3.031839
CWA10.2	1,000	-0.03547	1.020362	-3.43375	3.402699
CWA15.1	1,000	-0.05700	0.993673	-3.12198	3.065491
CWA15.2	1,000	-0.05296	0.998418	-3.53151	2.962610
CWA20.1	1,000	-0.00334	1.001314	-3.37719	2.883315
CWA20.2	1,000	-0.04154	1.038738	-3.43989	2.727260
CWA25.1	1,000	-0.01153	0.981983	-3.06492	2.927809
CWA25.2	1,000	-0.02929	0.987648	-3.15392	3.449513
MCC10.1	1,000	0.042368	0.959969	-3.27228	2.934557
MCC10.2	1,000	-0.01419	0.997236	-3.0372	3.276220
MCC15.1	1,000	-0.00669	0.995901	-3.10555	2.849567
MCC15.2	1,000	0.01523	0.977154	-2.8723	3.167309
MCC20.1	1,000	-0.04799	0.981171	-3.62399	3.218905

ตารางที่ 3.3 ตัวอย่างค่าต่ำสุด สูงสุด ค่าเฉลี่ย และค่าส่วนเบี่ยงเบนมาตรฐานของค่าความสามารถจริง ของกลุ่มตัวอย่าง ที่ใช้กับคลังข้อสอบขนาด 800 ข้อ

กลุ่มตัวอย่าง		ความสามารถจริง (θ_i)			
	N	*M*	*SD*	*Min*	*Max*
CWA10.1	1,000	-0.03274	0.99026	-3.26283	3.25702
CWA10.2	1,000	0.03084	0.98535	-3.02673	3.21127
CWA10.3	1,000	-0.01669	1.00600	-3.51609	3.56128
CWA15.1	1,000	0.00242	1.01332	-2.87905	3.45047
CWA15.2	1,000	0.04703	0.98297	-3.63821	3.02305
CWA15.3	1,000	0.00052	1.04132	-3.15796	2.79714
CWA20.1	1,000	-0.00501	1.01334	-3.38941	3.11029
CWA20.2	1,000	0.04488	0.98785	-3.69380	3.71375
CWA20.3	1,000	0.03171	1.01274	-3.16064	2.99171
CWA25.1	1,000	-0.02049	1.02682	-2.96446	3.27223
CWA25.2	1,000	0.00166	1.00873	-3.28816	3.41389
CWA25.3	1,000	-0.07313	1.00078	-3.83436	3.85517
MCC10.1	1,000	-0.03768	1.00758	-3.55259	3.27884
MCC10.2	1,000	-0.00433	1.05278	-3.35760	2.83381
MCC10.3	1,000	-0.02153	0.99863	-3.05302	3.14327
MCC15.1	1,000	-0.00573	0.99841	-3.10390	3.29723
MCC15.2	1,000	-0.03904	1.01287	-3.34147	2.85651
MCC15.3	1,000	-0.02232	1.00838	-2.72291	3.35096
MCC20.1	1,000	0.04978	0.96076	-3.61662	3.00316
MCC20.2	1,000	0.04580	1.01171	-3.77723	3.60987
MCC20.3	1,000	0.01086	0.99449	-2.91896	2.85321
MCC25.1	1,000	0.00304	1.00900	-3.44745	3.01446
MCC25.2	1,000	0.03298	1.00597	-3.34306	3.64978
MCC25.3	1,000	-0.02489	1.03647	-3.47169	3.22033

ตารางที่ 3.4 ตัวอย่างค่าต่ำสุด สูงสุด ค่าเฉลี่ย และค่าส่วนเบี่ยงเบนมาตรฐานของพารามิเตอร์อิทธิพลของแบบทดสอบย่อยของกลุ่มตัวอย่าง ที่ใช้กับคลังข้อสอบขนาด 600 ข้อ

กลุ่มตัวอย่าง	N x testlet	ความสามารถจริง (θ_i)			
		M	*SD*	*Min*	*Max*
CWA10.1	1,000x150	0.00194	1.00310	-4.4874	4.50004
CWA10.2	1,000x150	0.00208	1.00000	-4.4793	4.71791
CWA10.3	1,000x150	0.00103	0.99888	-4.1593	4.36196
CWA15.1	1,000x150	-0.00220	0.99950	-4.4100	4.53620
CWA15.2	1,000x150	0.00064	1.00295	-4.6111	4.97150
CWA15.3	1,000x150	-0.00060	0.99503	-4.3505	4.23575
CWA20.1	1,000x150	0.00241	0.99940	-4.3045	5.09418
CWA20.2	1,000x150	-0.00070	0.99789	-4.4322	4.20917
CWA20.3	1,000x150	0.00100	0.99751	-4.3966	4.13721
CWA25.1	1,000x150	-0.00130	1.00002	-4.6091	4.37648
CWA25.2	1,000x150	0.00244	0.99905	-4.3266	4.48451
CWA25.3	1,000x150	-0.00240	0.99937	-4.6778	4.43282
MCC10.1	1,000x150	0.00153	0.99940	-4.8709	4.19628
MCC10.2	1,000x150	0.00526	0.99872	-4.4086	4.31871
MCC10.3	1,000x150	-0.00350	0.99825	-4.2286	4.60080
MCC15.1	1,000x150	0.00580	1.00058	-4.3879	4.46961
MCC15.2	1,000x150	-0.00160	0.99896	-4.5266	4.33149
MCC15.3	1,000x150	0.00060	0.99824	-4.0448	4.22375
MCC20.1	1,000x150	-0.00230	1.00072	-4.2955	5.20761
MCC20.2	1,000x150	0.00081	1.00356	-4.0702	4.22597
MCC20.3	1,000x150	0.00471	0.99931	-5.3006	5.52680
MCC25.1	1,000x150	0.00240	0.99904	-4.6543	4.79903
MCC25.2	1,000x150	0.00616	0.99845	-4.2339	4.24767
MCC25.3	1,000x150	-0.00510	0.99795	-4.4451	4.49620

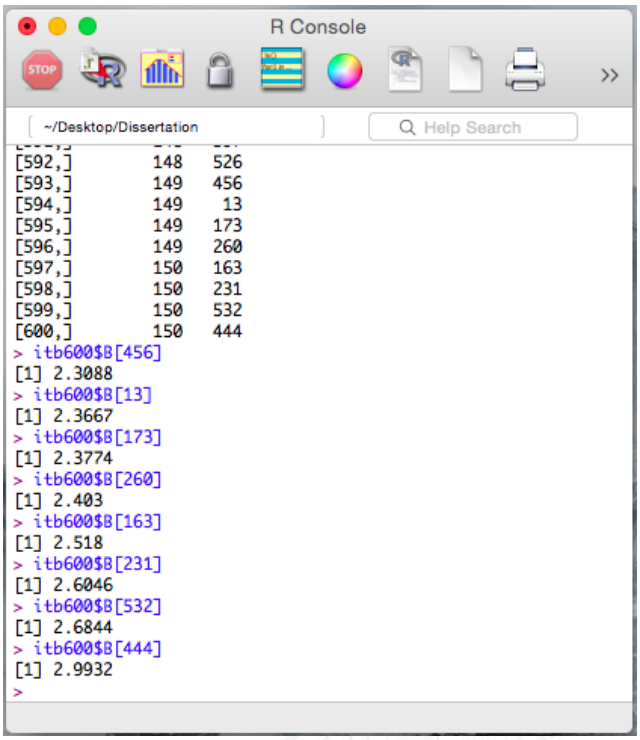
ตารางที่ 3.5 ตัวอย่างค่าต่ำสุด สูงสุด ค่าเฉลี่ย และค่าส่วนเบี่ยงเบนมาตรฐานของพารามิเตอร์อิทธิพลของแบบทดสอบย่อยของกลุ่มตัวอย่าง ที่ใช้กับคลังข้อสอบขนาด 800 ข้อ

กลุ่มตัวอย่าง	N x testlet	ความสามารถจริง (θ_i)			
		M	*SD*	*Min*	*Max*
CWA10.1	1,000x200	0.001424	1.003152	-4.48740	4.738351
CWA10.2	1,000x200	-0.000666	0.998417	-4.85824	4.731317
CWA10.3	1,000x200	-0.001858	0.998988	-4.14906	4.282637
CWA15.1	1,000x200	-0.003363	0.998987	-4.27378	4.574597
CWA15.2	1,000x200	0.000842	1.001688	-4.35731	4.418685
CWA15.3	1,000x200	-0.000669	1.000606	-5.04485	4.444578
CWA20.1	1,000x200	0.001386	0.997209	-4.88768	4.821648
CWA20.2	1,000x200	-0.002183	1.003332	-4.26489	4.707157
CWA20.3	1,000x200	0.001051	1.000240	-5.04485	4.733173
CWA25.1	1,000x200	0.001404	1.001026	-4.48740	4.500042
CWA25.2	1,000x200	-0.001479	0.997496	-4.30352	4.610346
CWA25.3	1,000x200	-0.000009	0.998572	-4.43288	4.282637
MCC10.1	1,000x200	-0.002623	1.001021	-5.21801	4.787140
MCC10.2	1,000x200	-0.000471	0.998389	-4.53231	4.810740
MCC10.3	1,000x200	0.005336	0.999529	-4.37326	4.558704
MCC15.1	1,000x200	0.001933	0.998183	-5.01329	4.287699
MCC15.2	1,000x200	-0.000615	0.997904	-4.16310	4.592001
MCC15.3	1,000x200	-0.003446	0.997012	-4.71234	4.423295
MCC20.1	1,000x200	0.005204	1.000339	-4.26383	4.843073
MCC20.2	1,000x200	0.000913	1.001794	-4.44908	4.738039
MCC20.3	1,000x200	0.000884	0.999171	-4.51555	4.829784
MCC25.1	1,000x200	-0.002268	0.996493	-4.34769	4.131303
MCC25.2	1,000x200	-0.002985	1.000402	-4.23705	4.614571
MCC25.3	1,000x200	0.002849	1.001576	-4.53493	4.833720

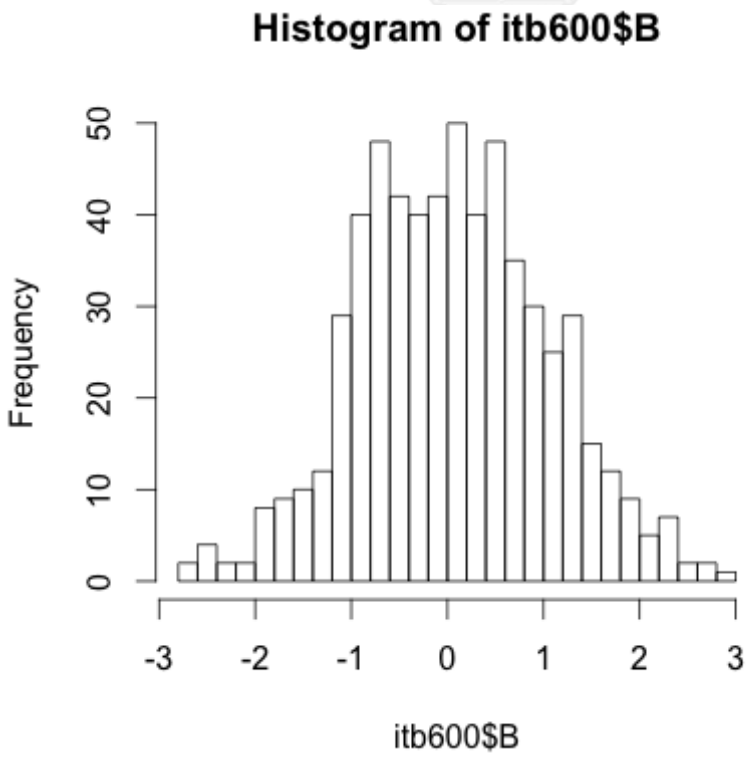
2. การจำลองคลังข้อสอบ และโครงสร้างของคลังข้อสอบ

จากการสังเคราะห์เอกสารและงานวิจัย พบว่า งานวิจัยในอดีตใช้ขนาดของคลังข้อสอบมีตั้งแต่ ขนาดประมาณ 300 ข้อ จนถึงขนาดมากกว่า 1,000 ข้อ จากองค์ความรู้ที่ค้นพบในอดีต ทำให้ทราบว่าเมื่อคลังข้อสอบมีขนาดใหญ่ขึ้นประสิทธิภาพของการทดสอบจะดีขึ้นตามไปด้วย แต่เนื่องจากการพัฒนาคลังข้อสอบขนาดใหญ่นั้นต้องใช้งบประมาณ และทรัพยากรจำนวนมาก ดังนั้นผู้วิจัยจึงสนใจที่จะศึกษาในคลังข้อสอบขนาด 600 ข้อ และ 800 ข้อ ซึ่งเป็นขนาดปานกลางและมีความเป็นไปได้ในการนำไปใช้ในสถานการณ์การทดสอบที่เป็นจริง

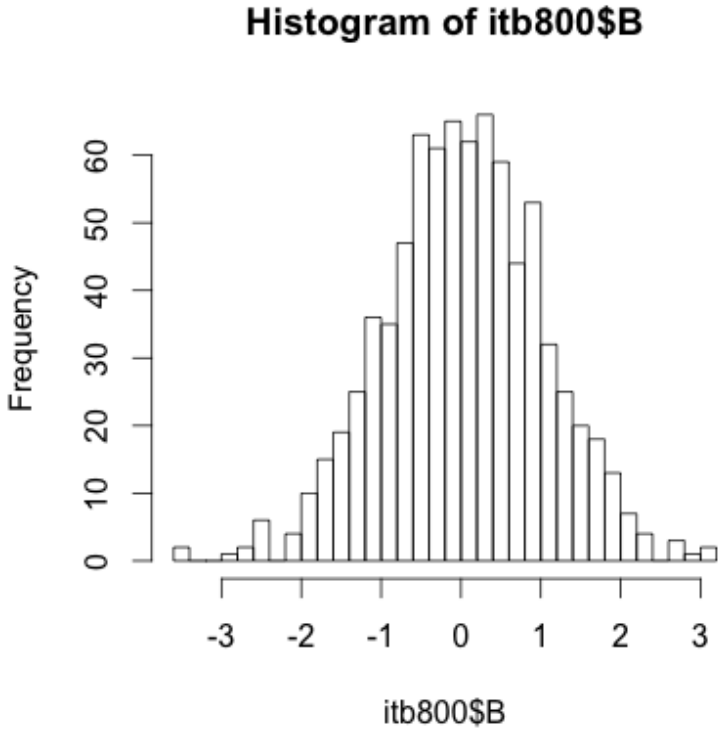
ค่าพารามิเตอร์ของข้อสอบจะถูกสร้างขึ้นจากการกำหนดค่าเฉลี่ย และความแปรปรวน ซึ่งพารามิเตอร์แต่ละตัวมีการกำหนดรายละเอียดดังนี้ พารามิเตอร์อำนาจจำแนก (a_j) ถูกสร้างขึ้นจากการแจกแจงแบบ Log-Normal [a ~LN(0.02,0.22^2)] เนื่องจากพารามิเตอร์อำนาจจำแนกมีค่าเป็นบวกและมากกว่า 0.5 พารามิเตอร์ความยาก (b_j) สร้างขึ้นจากการแจกแจงปกติมาตรฐาน [b ~N(0,1)] เนื่องจากข้อสอบที่มีความยากหรือง่ายมากๆ จะไม่ค่อยพบในแบบสอบ พารามิเตอร์การเดา (c_j) สร้างขึ้นจากการแจกแจงแบบ Beta [c ~BE(2,10)] เนื่องจากพารามิเตอร์การเดามีค่าเป็นบวกและค่าจะมีขนาดไม่เกิน 0.2 (Ngudgratoke & Yon, 2006) ซึ่งสอดคล้องกับเกณฑ์ในการคัดเลือกข้อสอบสำหรับคลังข้อสอบของ Urry (1977) และแบบทดสอบย่อยแต่ละฉบับจะถูกกำหนดให้มีข้อสอบจำนวน 4 ข้อ ข้อสอบภายในแบบทดสอบย่อยจะวัดในขอบเขตเนื้อหาเดียวกัน และมีค่าความยากใกล้เคียงกันมากที่สุด ดังตัวอย่างในภาพที่ 3.1 การแจกแจงของค่าพารามิเตอร์ข้อสอบ (b, a, c) มีรายละเอียดตามภาพที่ 3.2 ถึง 3.7 ค่าสถิติพื้นฐานของคลังข้อสอบ รายละเอียดดังตารางที่ 3.6 และขอบเขตเนื้อหาของรหัสของแบบทดสอบย่อยที่แต่ละฉบับ มีรายละเอียดตามตารางที่ 3.7



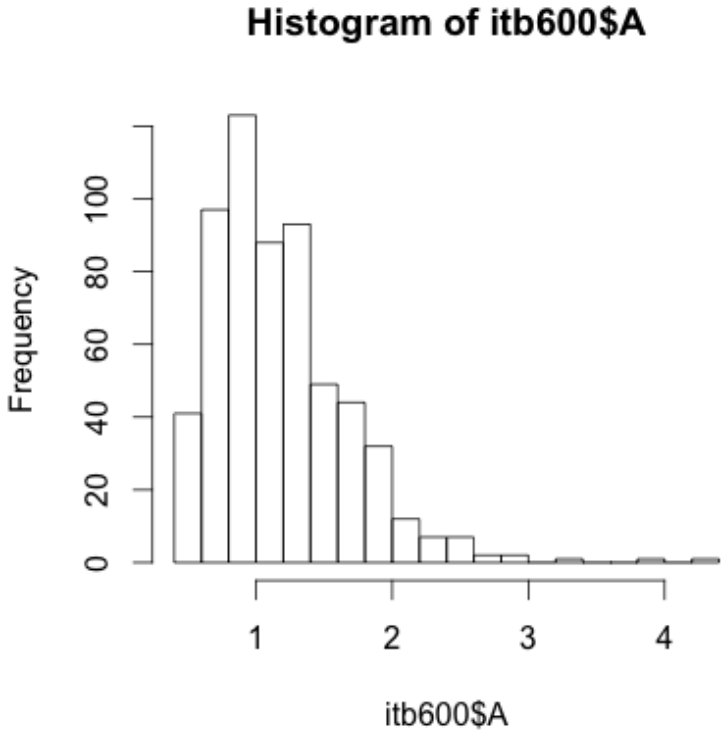
ภาพที่ 3.1 ตัวอย่างค่าพารามิเตอร์ความยากของข้อสอบที่อยู่ภายในแบบทดสอบย่อยเดียวกัน



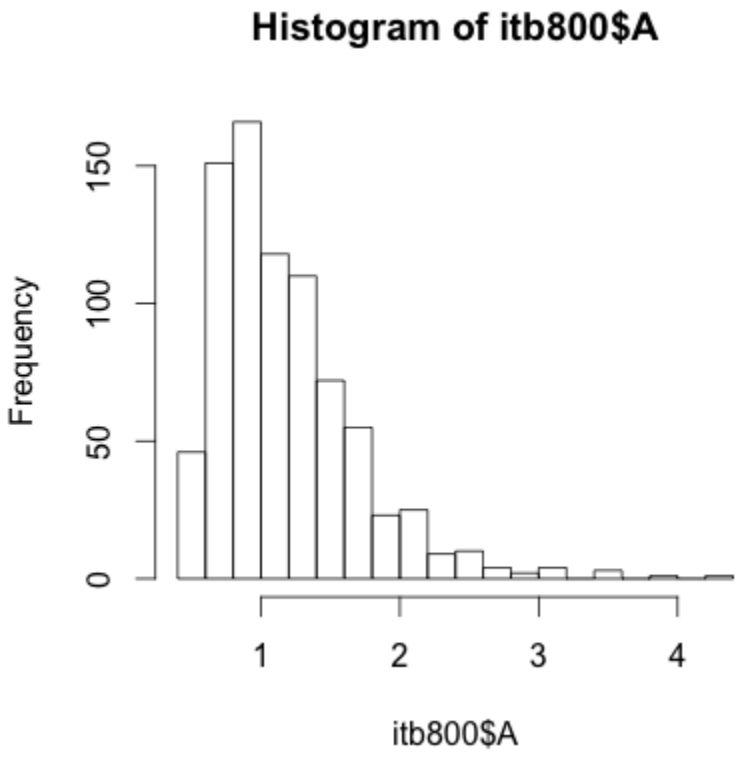
ภาพที่ 3.2 แสดงความถี่ค่าพารามิเตอร์ความยาก ของคลังข้อสอบขนาด 600 ข้อ



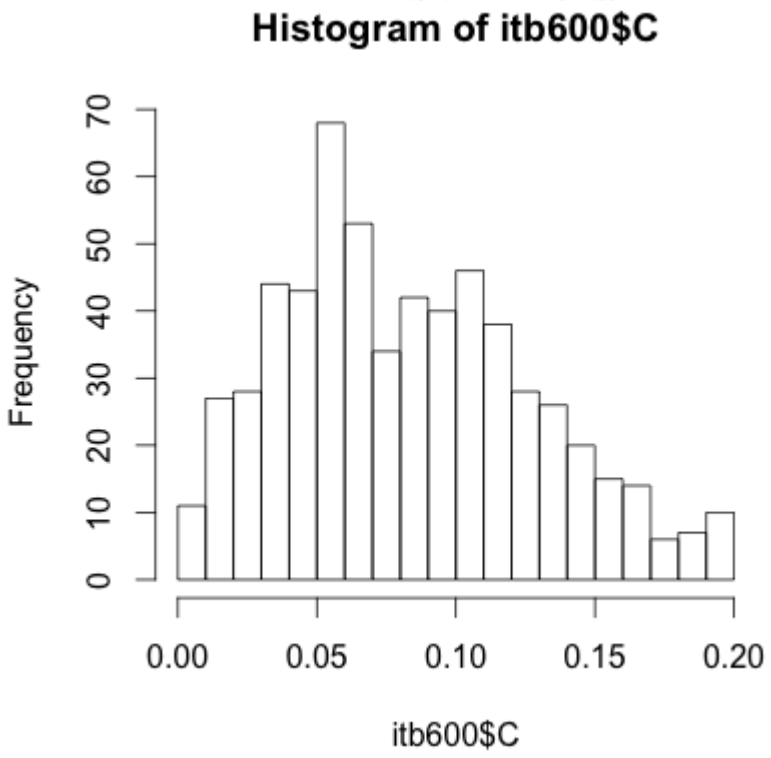
ภาพที่ 3.3 แสดงความถี่ค่าพารามิเตอร์ความยาก ของคลังข้อสอบขนาด 800 ข้อ



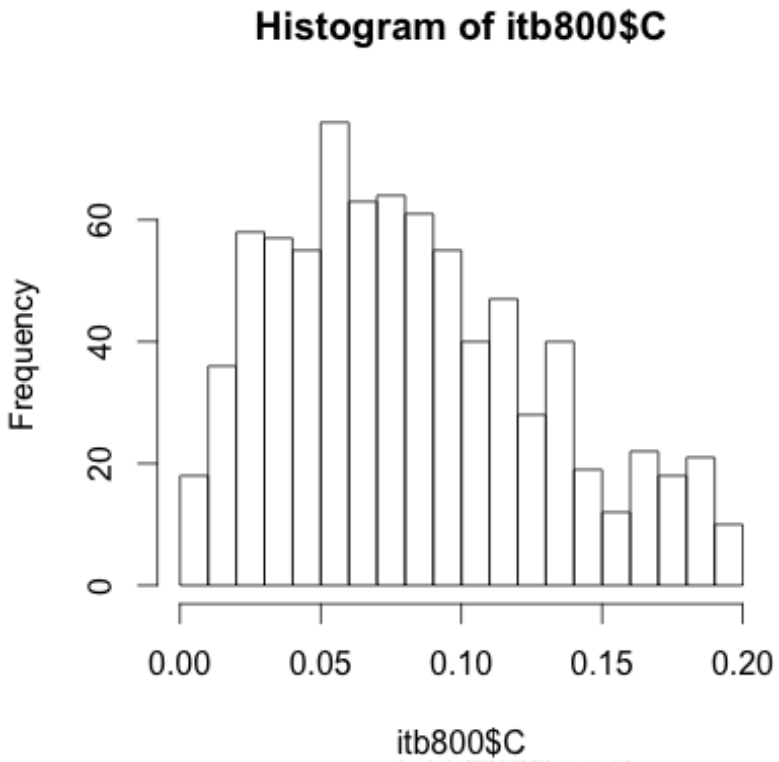
ภาพที่ 3.4 แสดงความถี่ค่าพารามิเตอร์อำนาจจำแนก ของคลังข้อสอบขนาด 600 ข้อ



ภาพที่ 3.5 แสดงความถี่ค่าพารามิเตอร์อำนาจจำแนก ของคลังข้อสอบขนาด 800 ข้อ



ภาพที่ 3.6 แสดงความถี่ค่าพารามิเตอร์การเดา ของคลังข้อสอบขนาด 600 ข้อ



ภาพที่ 3.7 แสดงความถี่ค่าพารามิเตอร์การเดา ของคลังข้อสอบขนาด 800 ข้อ

ตารางที่ 3.6 ค่าสูงสุด ค่าต่ำสุด ค่าเฉลี่ย และส่วนเบี่ยงเบนมาตรฐาน ของพารามิเตอร์ข้อสอบ

Item bank	600			800		
parameter	a	b	c	a	b	c
MAX	4.2388	2.9932	0.1984	4.2802	3.1295	0.1976
MIN	0.5029	-2.6280	0.0025	0.5028	-3.5584	0.0036
MEAN	1.1780	0.0610	0.0830	1.1794	0.0275	0.0822
SD	0.4909	0.9930	0.0448	0.5204	0.9960	0.0470
a>=3	3	-	-	9	-	-

จากตารางที่ 3.6 แสดงค่าสถิติพื้นฐานของพารามิเตอร์ข้อสอบของคลังข้อสอบขนาด 600 ข้อ และ 800 ข้อ ที่ได้จากการลำลองข้อมูลโดยกำหนดค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐาน และรูปแบบการแจกจงของข้อมูล ตามผลการศึกษาของ Ngudgratoke and Yon (2006) เมื่อผู้วิจัยวิเคราะห์มูลด้วยสถิติเบื้องต้นทำให้ผู้วิจัยทราบค่าเฉลี่ยของพารามิเตอร์อำนาจจำแนก (a) มีค่าเท่ากับ 1.178 และ

1.1794 ค่าเฉลี่ยของพารามิเตอร์ความยาก (b) มีค่าเท่ากับ 0.0610 และ 0.0275 และค่าเฉลี่ยของพารามิเตอร์โอกาสในการเดา (c) มีค่าเท่ากับ 0.0830 และ 0.0822 ว่าอยู่ในเกณฑ์ที่เหมาะสมและเมื่อพิจารณาค่าพารามิเตอร์เป็นรายข้อพบว่าข้อสอบส่วนใหญ่มีค่าพารามิเตอร์อยู่ในเกณฑ์ที่เหมาะสมยอมรับได้ตามงานวิจัยของ Mislevy (1991) แต่มีข้อสอบในคลังข้อสอบทั้งหมด 12 ข้อ จากทั้งหมด 1400 ข้อ แบ่งเป็น จากคลังข้อสอบขนาด 600 ข้อ จำนวน 3 ข้อ และจากคลังข้อสอบขนาด 800 ข้อ จำนวน 9 ข้อ ที่มีค่าพารามิเตอร์อำนาจจำแนกสูง (a) สูงกว่า 3.0 เมื่อพิจารณาข้อผิดพลาดที่เกิดขึ้นพบว่าคิดเป็นร้อยละ 0.86 ถือว่าเป็นความผิดพลาดที่เกิดอย่างสุ่มในขั้นตอนการจำลองข้อมูลพารามิเตอร์ข้อสอบที่ผู้วิจัยสามารถยอมรับได้

ตารางที่ 3.7 หมายเลขแบบทดสอบย่อยที่อยู่ในแต่ละขอบเขตเนื้อหา

คลังข้อสอบขนาด 600 ข้อ			คลังข้อสอบขนาด 800 ข้อ		
เนื้อหา 1	เนื้อหา 2	เนื้อหา 3	เนื้อหา 1	เนื้อหา 2	เนื้อหา 3
1, 4, 6, 7, 8, 9, 10, 12, 13, 16, 17, 21, 27, 28, 30, 32, 36, 42, 44, 49, 50, 51, 56, 58, 60, 61, 63, 64, 65, 67, 72, 75, 76, 82, 84, 85, 87, 91, 99, 100, 102, 107, 108, 112, 115, 116, 118, 119, 120, 125, 126, 131, 137, 140, 144, 146	2, 3, 5, 11, 14, 15, 19, 24, 25, 31, 33, 34, 35, 38, 40, 45, 46, 48, 52, 53, 54, 55, 57, 59, 66, 69, 79, 80, 83, 86, 88, 89, 93, 94, 95, 97, 98, 101, 103, 109, 110, 111, 113, 114, 123, 124, 128, 133, 139, 142, 143, 147, 148	18, 20, 22, 23, 26, 29, 37, 39, 41, 43, 47, 62, 68, 70, 71, 73, 74, 77, 78, 81, 90, 92, 96, 104, 105, 106, 117, 121, 122, 127, 129, 130, 132, 134, 135, 136, 138, 141, 145, 149, 150	6, 8, 10, 11, 12, 16, 18, 19, 20, 30, 31, 42, 48, 49, 51, 55, 56, 57, 59, 60, 71, 82, 83, 86, 87, 90, 92, 93, 96, 97, 104, 106, 107, 113, 118, 121, 125, 128, 132, 133, 142, 144, 145, 146, 147, 148, 149, 150, 151, 153, 159, 160, 164, 166, 169, 172, 173, 177, 180, 182, 186, 187, 190, 192, 195, 200	3, 4, 5, 7, 9, 13, 14, 17, 22, 27, 32, 35, 36, 37, 40, 44, 52, 63, 65, 67, 68, 73, 74, 75, 76, 77, 78, 80, 81, 84, 85, 88, 89, 91, 94, 98, 99, 100, 101, 102, 105, 108, 109, 111, 122, 124, 127, 130, 134, 135, 136, 140, 141, 143, 155, 156, 157, 158, 165, 168, 175, 176, 179, 184, 193	1, 2, 15, 21, 23, 24, 25, 26, 28, 29, 33, 34, 38, 39, 41, 43, 45, 46, 47, 50, 53, 54, 58, 61, 62, 64, 66, 69, 70, 72, 79, 95, 103, 110, 112, 114, 115, 116, 117, 119, 120, 123, 126, 129, 131, 137, 138, 139, 152, 154, 161, 162, 163, 167, 170, 171, 174, 178, 181, 183, 185, 188, 189, 191, 194, 196, 197, 198, 199
รวม 56	53	41	66	65	69
ร้อยละ 37.33	35.33	27.33	33.00	32.50	34.50

3. การคำนวณค่าความน่าจะเป็น (probability) ในการตอบข้อสอบถูก

การคำนวณค่าความน่าจะเป็นในการตอบข้อสอบถูกของกลุ่มตัวอย่างที่จำลองขึ้นในข้อที่ 1 กับคลังข้อสอบในข้อที่ 2 โดยใช้สูตร แบบ 3-พารามิเตอร์ Testlet response model ดังนี้

$$P_{ij} = c_j + (1-c_j)\frac{\exp\left[a_j(\theta_i - b_j - \gamma_{id(j)})\right]}{1+\exp\left[a_j(\theta_i - b_j - \gamma_{id(j)})\right]}$$

เมื่อ P_{ij} คือ โอกาสในการตอบข้อสอบถูกของคนที่ i ในข้อที่ j

θ_i คือ ความสามารถของผู้สอบคนที่ i

b_j คือ ค่าความยากของข้อสอบข้อที่ j

a_j คือ ค่าอำนาจจำแนกของข้อสอบข้อที่ j

c_j คือ ค่าโอกาสในการเดาของข้อสอบข้อที่ j

$\gamma_{id(j)}$ คือ อิทธิพลสุ่มข้อสอบข้อที่ i ซึ่งอยู่ในแบบทดสอบย่อยฉบับ d ที่มีต่อผู้สอบคนที่ j ในกรณีที่ข้อสอบข้อที่ i เป็นข้อสอบที่อิสระจากข้ออื่น $\gamma_{id(j)} = 0$

จากการใช้สูตรนี้คำนวณค่าโอกาสในการตอบข้อสอบถูก (P_{ij}) ของกลุ่มตัวอย่างในข้อที่ 1. กับข้อสอบที่จำลองขึ้นในข้อที่ 2. ในแต่ละครั้งของการทดลองจะได้ผลดังตาราง 3.8

ตารางที่ 3.8 ค่าโอกาสในการตอบข้อสอบถูกของผู้สอบแต่ละคนในแต่ละข้อเมื่อใช้สูตรแบบ 3-พารามิเตอร์ testlet response model

คนที่	T1							T150			
	Item 549	Item 473	Item 10	Item 584	...	...	...	Item 163	Item 231	Item 532	Item 444
1	P1,549	P1,473	P1,10	P1, 584				P1, 163	P1, 231	P1, 231	P1, 444
2	P2, 549	P2, 473	P2, 10	P2, 584				P2, 163	P2, 231	P2, 231	P2, 444
3	P3, 549	P3, 473	P3, 10	P3, 584				P3, 163	P3, 231	P3, 231	P3, 444
-	-	-	-	-				-	-	-	-
-	-	-	-	-				-	-	-	-
-	-	-	-	-				-	-	-	-
n	Pn, 549	Pn, 473	Pn, 10	Pn, 584				Pn, 163	Pn, 231	Pn, 231	Pn, 444

4. การคำนวณผลการตอบและการประมาณค่าพารามิเตอร์ผู้สอบ

การคำนวณผลการตอบสามารถทำได้โดยนำความน่าจะเป็นในการตอบข้อสอบถูกในข้อ 3.3 ไปคำนวณหาผลการตอบของผู้สอบ (U_{ij}) โดยการนำค่าความน่าจะเป็นในการตอบข้อสอบถูก (P_{ij}) จากข้อที่ 3.3 ไปพิจารณาผลการตอบข้อสอบ ตอบถูก=1 หรือ ตอบผิด=0 ของผู้สอบแต่ละคนในข้อสอบแต่ละข้อ โดยการเรียกเลขสุ่ม (X_{ij}) จากเครื่องคอมพิวเตอร์ที่มีการแจกแจงแบบยูนิฟอร์มมีค่าตั้งแต่ 0 ถึง 1 แล้วนำค่า X_{ij} ไปเปรียบเทียบกับ P_{ij} ถ้า X_{ij} <= P_{ij} แสดงว่าตกอยู่ในพื้นที่ยอมรับว่าตอบถูก จึงให้ U_{ij}=1 ถ้า X_{ij} > P_{ij} ก็จะปฏิเสธการตอบถูก นั่นคือ ตอบผิด ก็ทำให้ U_{ij} = 0 ทำอย่างนี้กับข้อสอบทุกข้อผู้สอบทุกคน ผู้วิจัยจึงได้ผลการตอบ (0, 1) เมื่อได้ผลการตอบแล้วสิ่งที่ต้องดำเนินการขั้นต่อไปคือ การประมาณค่าพารามิเตอร์ผู้สอบ (ความสามารถของผู้สอบ และอิทธิพลสุ่มที่มีต่อแบบทดสอบย่อย) และการประมาณค่าพารามิเตอร์ข้อสอบ (caribrate item) เพื่อหาค่าพารามิเตอร์ของข้อสอบภายใต้โมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย (Testlet Response Model)

วิธีการวิเคราะห์ข้อสอบที่ใช้โมเดลตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย ที่ได้ศึกษาเอกสารและวิจัยงานวิจัยที่เกี่ยวข้อง พบว่า ในปัจจุบันมีซอฟต์แวร์สำหรับอำนวยความสะดวกในการวิเคราะห์ข้อสอบที่ใช้โมเดลตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย ดังนี้ 1) โปรแกรม WinBug 2) โปรแกรม SCORIGHT และ 3) โปรแกรม R ซึ่งซอฟต์แวร์ทั้ง 3 ตัวเป็นซอฟต์แวร์ที่นำการวิเคราะห์แบบเบย์มาใช้สำหรับโมเดลทางคณิตศาสตร์ที่มีความซับซ้อนโดยใช้วิธีการ Markov chain Monte Carlo (MCMC) เข้ามาช่วยในการประมาณค่าพารามิเตอร์ ศึกษาเพิ่มเติมเกี่ยวกับเรื่อง MCMC ได้จาก Gilks (2005) จากศึกษาเอกสารคู่มือการใช้งานของซอฟต์แวร์ทั้ง 3 ตัว พบว่า โปรแกรม WinBug และโปรแกรม R เป็น ซอฟต์แวร์ที่มีความยืดหยุ่นในด้านการใช้งานสูง ผู้วิจัยสามารถเขียนคำสั่งเพื่อสร้างหรือเลือกโมเดลสำหรับวิเคราะห์ข้อมูลได้ รวมทั้งสามารถเตรียมข้อมูลในรูปของไฟล์เอกสาร ดังนั้นผู้วิจัยจึงเลือกใช้โปรแกรม R เพราะอยู่บนสภาพแวดล้อมเดียวกับระบบการทดสอบการทดสอบแบบปรับเมาะด้วยคอมพิวเตอร์ที่ใช้ในการศึกษาครั้งนี้ ทำให้สามารถส่งผ่านข้อมูลที่ถูกเก็บในรูปแบบเฟรมเวิร์กได้อย่างรวดเร็ว ขั้นตอนในการวิเคราะห์ข้อสอบที่ใช้โมเดลตอบสนองข้อสอบที่ใช้แบบทดสอบย่อยมีรายละเอียดดังนี้

1) บันทึกข้อมูลผลการตอบข้อสอบของผู้สอบ โดยกำหนดให้แถวบนสุดเป็นหมายเลขของข้อสอบ เช่น แบบทดสอบย่อยฉบับ A มีข้อสอบ จำนวน 4 ข้อ และกำหนดให้คอลัมน์แรกเป็นรหัสผู้สอบ รายละเอียดดังภาพที่ 3.8

2) นำข้อมูลเข้าสู่โปรแกรม R จากภาพที่ 3.8 ผู้วิจัยบันทึกข้อมูลอยู่ในรูปแบบไฟล์ csv ซึ่งข้อมูลของแต่ละคอลัมน์จะถูกแบ่งด้วยเครื่องหมายคอมมา (,) การนำข้อมูลเข้าสามารถทำได้โดยใช้คำสั่ง read.table เพื่ออ่านไฟล์ csv แล้วทำการเก็บค่าผลการตอบเข้าสู่วัตถุ (object) ที่ชื่อว่า dat ตัวอย่างคำสั่งคือ dat <- read.table("ชื่อไฟล์.csv", header=TRUE, sep=",", row.names="id")

	A1	A2	A3	A4	B1	B2	B3	B4	C1	C2	C3	C4
2	1	1	1	1	1	1	1	1	1	1	1	0
22	1	1	0	0	1	0	1	1	1	0	1	0
23	1	1	0	1	1	0	1	1	1	1	1	1
41	1	1	1	1	1	1	1	1	1	1	1	1
43	1	0	0	1	0	0	1	1	1	0	1	0
63	1	1	0	0	1	0	1	1	1	1	1	1
64	1	1	0	0	0	0	0	0	0	0	0	0
84	1	1	1	0	1	1	1	1	1	1	1	1

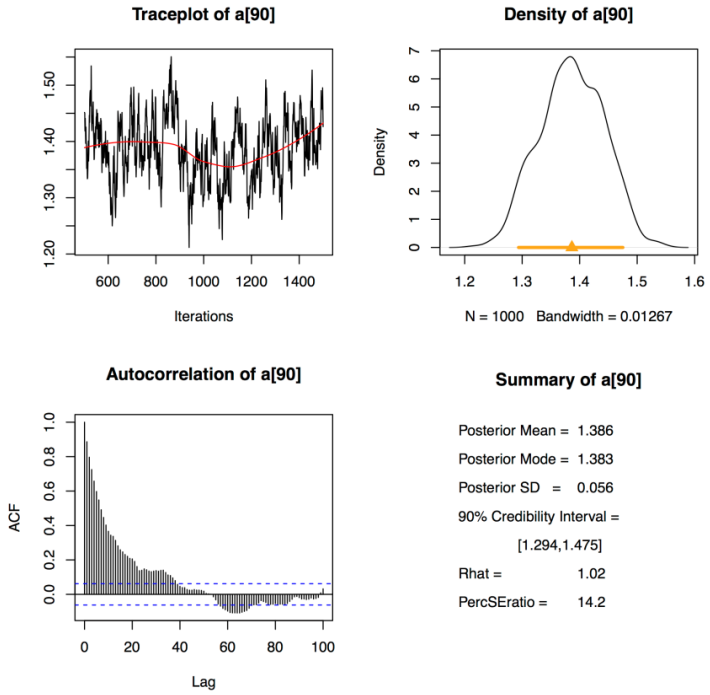
ภาพที่ 3.8 ตัวอย่างการคีย์ข้อมูลผลการตอบข้อสอบของผู้สอบ

3) เรียกใช้เพ็คเกจ “SIRT” โดยใช้คำสั่ง library("sirt") เพื่อคำสั่ง mcmc.3pno.testlet สำหรับการประมาณค่าพารามเตอร์ของข้อสอบและพารามิเตอร์ของผู้สอบด้วยโมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย (testlet response theory) การเรียกใช้คำสั่งมีอากิวเมนต์ สำคัญที่ต้องใส่ดังนี้ ข้อมูลผลการตอบ (dat) ชุดของตัวเลขหรือตัวอักษรที่ระบุความเป็นแบบทดสอบย่อย (testlets) การประมาณค่าความชัน (est.slope) การประมาณค่าการเดา (est.guess) ข้อมูลการแจกแจงโอกาในการเดา (guess.prior) ความแปรปรวนของอิทธิพลของแบบทดสอบย่อย (testlet.variance.prior) การกำหนดจำนวนรอบในการทำซ้ำ (iter) กำหนดจำนวนรอบในการเบิร์นอิน (burnin) ตัวอย่างการเรียกใช้คำสั่งมีดังนี้

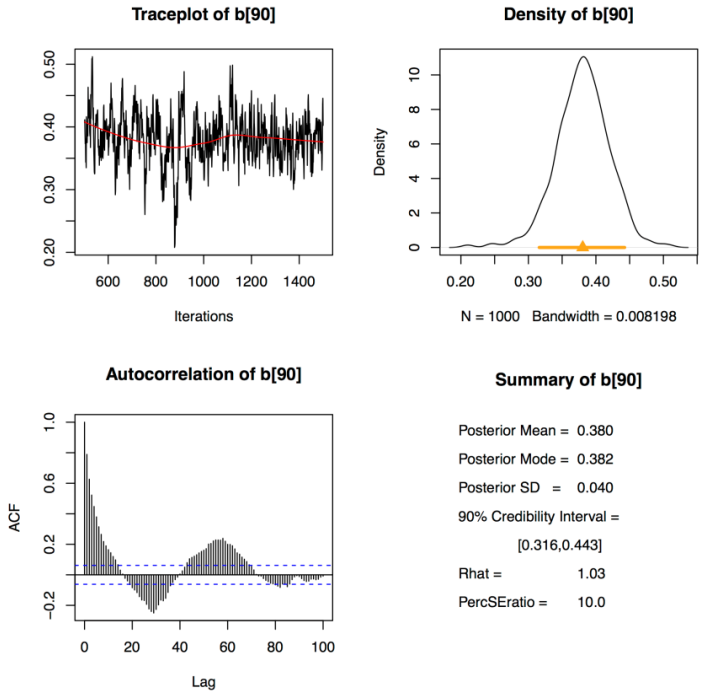
```
I <- ncol(dat);

burnin <- 7000;

iter <- 8000;

mod <- mcmc.3pno.testlet( dat , guess.prior=c(2,10) , burnin=burnin, iter=iter)

summary(mod)

plot(mod,ask=TRUE) # plot MCMC chains in coda style
```

จากการประมวลผลด้วยคำสั่งด้านบนนี้ ผลลัพธ์จะถูกเก็บในวัตถุที่ชื่อว่า “mod” การดึงข้อมูลผลลัพธ์มาใช้ต้องจัดกระทำให้อยู่ในรูปของค่าสถิติพื้นฐาน เช่น ค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐาน โดยการใช้คำสั่ง summary(ชื่อวัตถุ) และในการวิเคราะห์ ผู้วิเคราะห์สามารถขอดูกราฟ

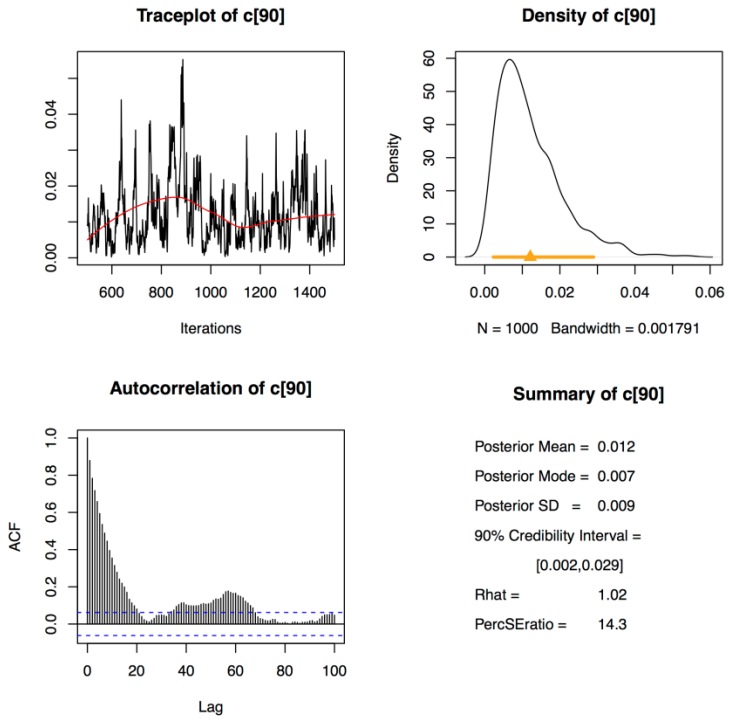
Traceplot, Autocorrelation และ Density ของค่าพารามิเตอร์ข้อสอบและพารามิเตอร์ผู้สอบได้ตัวอย่างของกราฟ แสดงดังภาพที่ 3.9 – 3.13



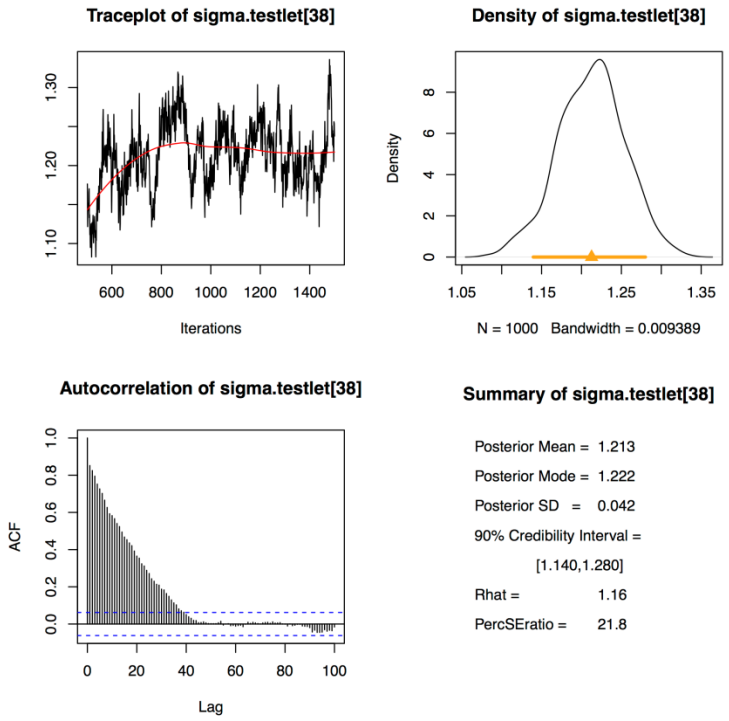
ภาพที่ 3.9 ตัวอย่างผลการประมาณค่าพารามิเตอร์อำนาจจำแนกโดยใช้วิธี mcmc ในโปรแกรม R



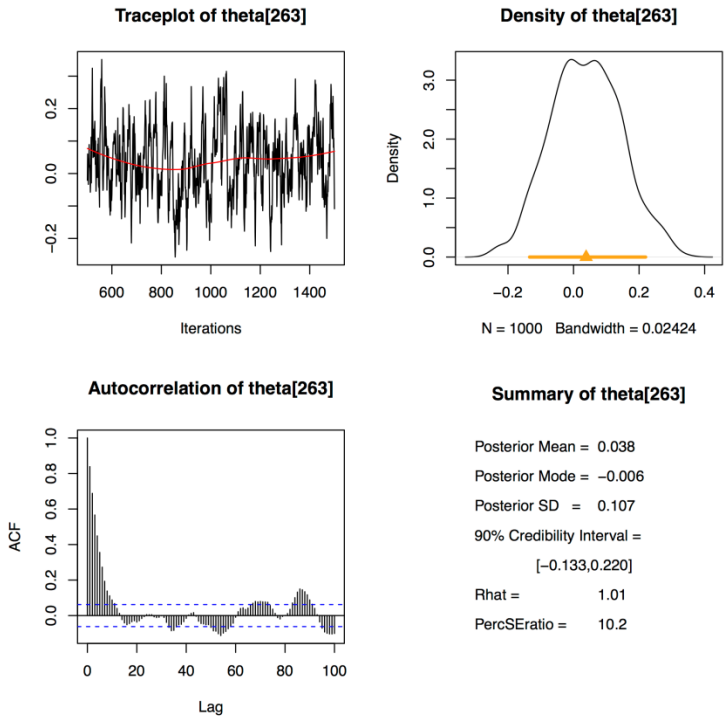
ภาพที่ 3.10 ตัวอย่างผลการประมาณค่าพารามิเตอร์ความยากโดยใช้วิธี mcmc ในโปรแกรม R



ภาพที่ 3.11 ตัวอย่างผลการประมาณค่าพารามิเตอร์โอกาสการเดาโดยใช้วิธี mcmc ในโปรแกรม R



ภาพที่ 3.12 ตัวอย่างผลการประมาณค่าพารามิเตอร์อิทธิพลของแบบทดสอบย่อยโดยใช้วิธี mcmc ในโปรแกรม R



ภาพที่ 3.13 ตัวอย่างผลการประมาณค่าความสามารถของผู้สอบโดยใช้วิธี mcmc ในโปรแกรม R

จากนั้นจึงนำผลการตอบไปคำนวณหาค่าพารามิเตอร์ของผู้สอบ (person parameter) โดยใช้ฟังก์ชั่น 3PNO Testlet Model ในแพ็กเกจ (Supplementary Item Response Theory Models: SIRT) จากโปรแกรม R ได้ผลดังตาราง 3.9

เมื่อได้ผลการตอบข้อสอบแล้วจึงมาใช้หาเส้นทางการตอบข้อสอบด้วยวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล (Monte Carlo CAT Method: MCC) และวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (constraint-weighted a-stratification CAT method: CWA) ซึ่งเป็นเครื่องมือในการวิจัยกับเงื่อนไขที่ต้องการศึกษา รายละเอียดดังตารางที่ 3.10 จากนั้นจะได้ค่าความสามารถของผู้สอบที่ผ่านการประมาณค่าจากการทดสอบปรับเหมาะ ($\hat{\theta}_j$) แล้วจึงนำผลการทดสอบไปวิเคราะห์ตามขั้นตอนการวิเคราะห์ข้อมูลในตอนที่ 4

ตารางที่ 3.9 ตัวอย่างผลการตอบข้อสอบและผลการประมาณค่าพารามิเตอร์ผู้สอบ 10 คน

คน	T1				T2				T150				Estimate person parameter			
ที่	549	473	10	584	527	185	120	386	163	231	532	444	theta	G1	G2	G150
1	0	0	0	0	0	1	0	0	1	0	0	0	-0.75	1.93	1.09	-0.06
2	0	0	1	1	1	1	0	1	1	0	0	1	0.37	-0.77	1.42	-0.91
3	1	1	1	0	0	1	1	1	1	0	1	0	0.12	0.06	-0.96	-0.37
4	1	0	1	1	0	1	1	0	1	1	1	1	-0.49	0.61	0.39	-0.98
5	1	1	0	0	1	1	0	1	1	1	0	1	0.36	0.25	0.66	-0.50
6	1	0	0	0	1	1	0	0	1	0	0	0	-0.63	-2.24	0.09	0.52
7	1	0	0	1	1	1	0	0	1	1	0	1	-0.64	-2.06	-1.21	-0.09
8	1	0	0	0	1	1	0	0	0	0	0	0	-0.64	0.32	0.45	0.30
9	1	1	1	1	1	1	0	0	1	0	0	0	0.68	0.33	-0.26	0.15
10	1	0	0	0	0	1	0	0	1	0	0	0	-0.75	1.66	-0.25	-0.06

* G หมายถึง อิทธิพลของแบบทดสอบย่อย (testlet effect)

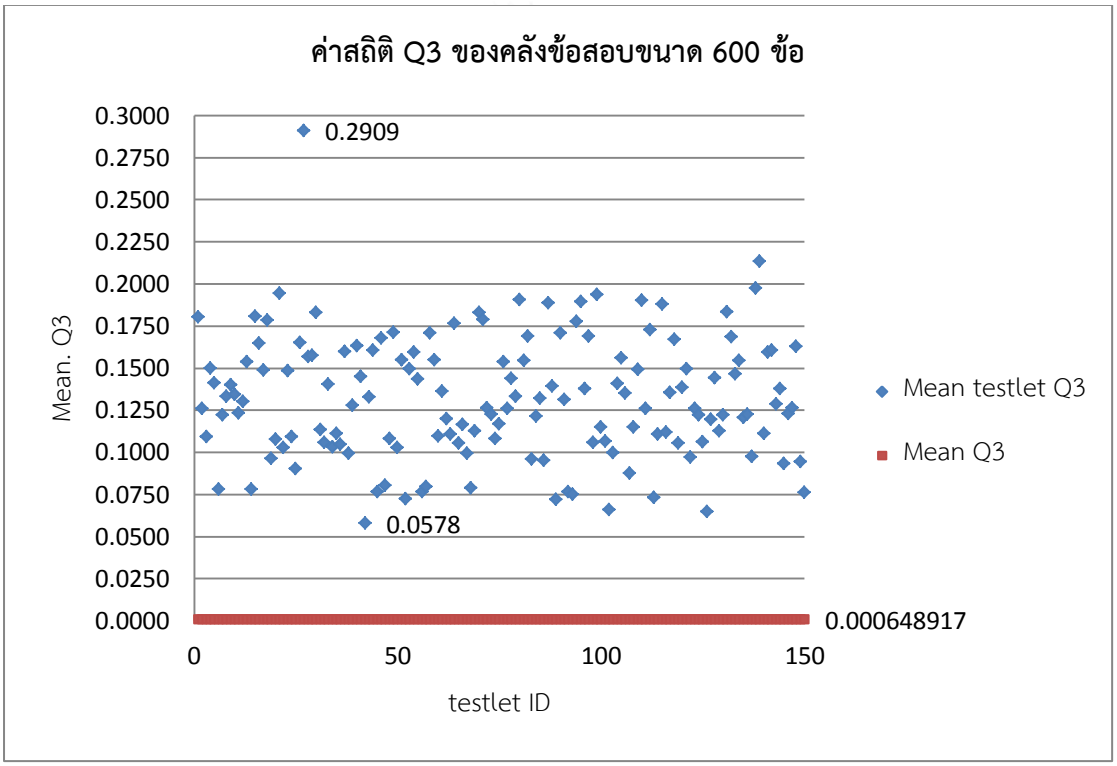
5. การตรวจสอบความถูกต้องของการจำลองข้อมูล

เพื่อให้การจำลองข้อมูลผลการตอบข้อสอบสอดคล้องกับจุดมุ่งหมายของการวิจัยที่ต้องการศึกษาภายใต้บริบทของการใช้โมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย (testlet response model) ผู้วิจัยตรวจสอบผลการจำลองข้อมูลโดยใช้ข้อตกลงเบื้องต้นเกี่ยวกับความไม่เป็นอิสระเฉพาะที่ของข้อสอบ (local item dependence: LID) เพื่อวินิจฉัยว่าชุดของข้อสอบที่รวมกันอยู่ภายในแบบทดสอบย่อยเป็นอิสระต่อกันหรือไม่ โดย Yen (1984) ได้เสนอสถิติ Q3 ซึ่งเป็นดัชนีสำหรับตรวจสอบความไม่เป็นอิสระเฉพาะที่ของข้อสอบ โดยการหาความสัมพันธ์ของส่วนที่เหลือ (residuals) ของคู่ของข้อสอบหลังจากตัด (partialling) ส่วนที่เป็นการประมาณค่าคุณลักษณะ (trait estimate) ออกไป การวินิจฉัยความไม่เป็นอิสระเฉพาะที่ของข้อสอบพิจารณาได้จากเกณฑ์ดังนี้ 1) ถ้าค่าสถิติ Q3 มีค่าเป็นลบต่ำๆ หรือมีค่าเข้าใกล้ 0 มากๆ แสดงว่าข้อสอบคู่นั้นมีหลักฐานพอเชื่อได้ว่าเป็นอิสระจากกัน 2) ถ้าค่าสถิติ Q3 มีค่าต่างจาก 0 เช่น มีค่าเป็นลบหรือเป็นบวกมากๆ แสดงว่าข้อสอบคู่นั้นมีหลักฐานพอเชื่อได้ว่าไม่เป็นอิสระจากกัน

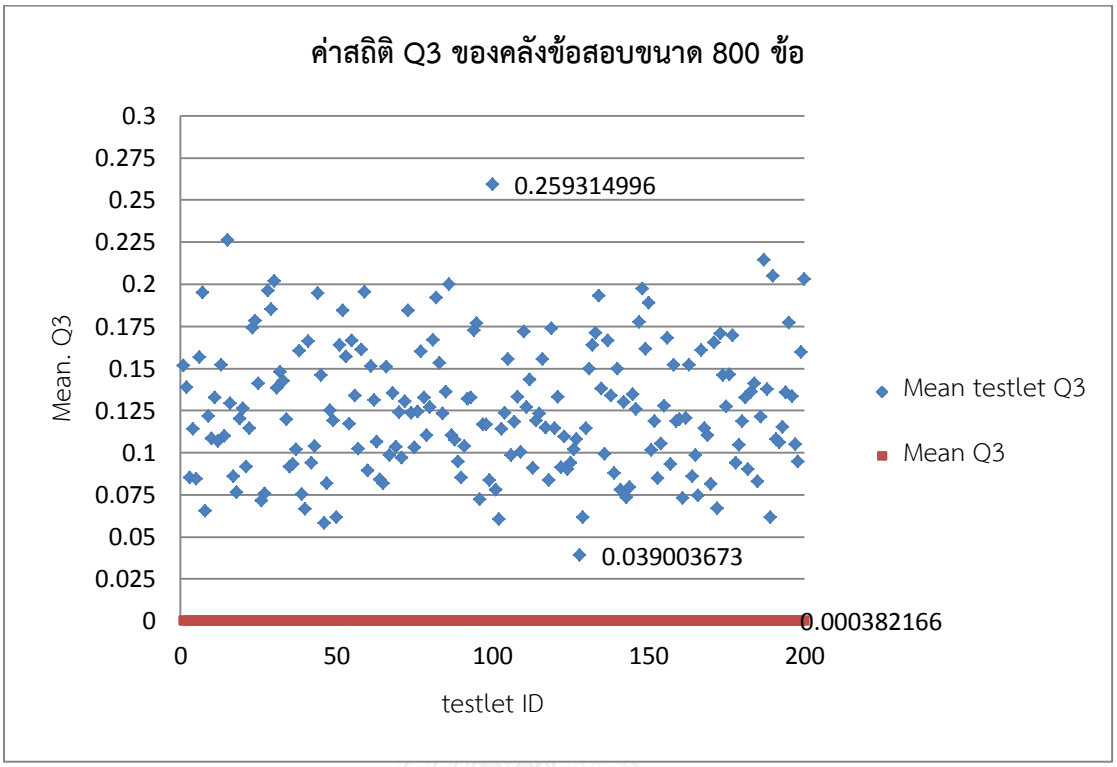
ผลการวิเคราะห์ค่าสถิติ Q3 พบว่า เมื่อวิเคราะห์สถิติ Q3 โดยใช้หน่วยการวิเคราะห์เป็นรายข้อจากคลังข้อสอบขนาด 600 ข้อ และ 800 ข้อ มีค่าเฉลี่ยของสถิติ Q3 เป็น 0.00065 และ 0.00038 ตามลำดับ แสดงว่าข้อสอบมีความเป็นอิสระเฉพาะที่ของข้อสอบ แต่เมื่อใช้หน่วยการวิเคราะห์ชุดของข้อสอบตามแบบทดสอบย่อย พบว่า ผลการวิเคราะห์ค่าเฉลี่ย และส่วนเบี่ยงเบนมาตรฐานของสถิติ Q3 แสดงในตารางที่ 3.10 และค่าเฉลี่ยของสถิติ Q3 เป็นรายแบบทดสอบย่อยสำหรับคลังข้อสอบขนาด 600 ข้อ และ 800 ข้อ แสดงในภาพที่ 3.14 และ 3.15

ตารางที่ 3.10 ผลการวิเคราะห์ค่าสถิติ Q3

Item pool	Unit of analysis	*M*	*SD*	*Min*	*Max*
600	Items (600)	0.00065	0.03835	-0.54324	0.45326
	Testlets (150)	0.13219	0.03639	0.05778	0.29094
800	Items (800)	0.00038	0.03726	-0.60373	0.46146
	Testlets (200)	0.12622	0.03818	0.03900	0.25931



ภาพที่ 3.14 ค่าเฉลี่ยของสถิติ Q3 เป็นรายแบบทดสอบย่อยสำหรับคลังข้อสอบขนาด 600 ข้อ



ภาพที่ 3.15 ค่าเฉลี่ยของสถิติ Q3 เป็นรายแบบทดสอบย่อยสำหรับคลังข้อสอบขนาด 800 ข้อ

เครื่องมือที่ใช้ในการวิจัย

การวิจัยครั้งนี้เป็นการศึกษาในสถานการณ์จำลอง ผู้วิจัยใช้โปรแกรมภาษา R (R Programming language) ซึ่งเป็นภาษาที่พัฒนามาเพื่อใช้ในการคำนวณทางสถิติและทางด้านภาพกราฟิก การสร้างเครื่องมือในการจำลองข้อมูล จัดกระทำข้อมูลในการทดสอบ โดยผ่านขั้นตอนดังนี้คือ ศึกษาความมุ่งหมายการวิจัยและรวบรวมข้อมูลที่เกี่ยวข้องเพื่อออกแบบและทดสอบโปรแกรมภาษา R ที่พัฒนาขึ้น ปรับปรุงโปรแกรมตามลำดับผังขั้นตอนการทดสอบ นำโปรแกรมไปทดสอบกับข้อมูลที่รู้ค่าผลลัพธ์และทดสอบซ้ำจนแน่ใจว่าการทำงานของโปรแกรมถูกต้องและผลการทดสอบแม่นยำเพียงพอจึงนำไปใช้ทดสอบ โดยโปรแกรมที่พัฒนาขึ้นประกอบด้วย 4 ส่วน คือ 1) ชุดคำสั่งสำหรับสร้างพารามิเตอร์ผู้สอบ 2) ชุดคำสั่งสำหรับสร้างพารามิเตอร์ข้อสอบ 3) ชุดคำสั่งสำหรับวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที และ 4) ชุดคำสั่งสำหรับวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ

ผู้วิจัยได้ออกแบบโปรแกรมคอมพิวเตอร์สร้างเป็นเครื่องมือสำหรับการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่ใช้สำหรับโมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย 2 วิธี คือ 1)

วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที (Monte Carlo CAT Method: MCC) และ 2) วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (constraint-weighted a-stratification CAT method: CWA)

ขั้นตอนที่ 1 ศึกษาความมุ่งหมายการวิจัย นำไปวิเคราะห์สิ่งที่ต้องการศึกษาและรวบรวมข้อมูลที่เกี่ยวข้อง กำหนดความชัดเจนของข้อมูล พิจารณาความสัมพันธ์ระหว่างตัวแปรอิสระตัวแปรตาม และเงื่อนไขการทดลอง

ขั้นตอนที่ 2 เขียนผังขั้นตอนการทดสอบ โดยกำหนดลักษณะหรือโครงร่างการทดสอบครอบคลุมสิ่งที่ต้องการศึกษา แสดงความสัมพันธ์ของขั้นตอนการทดสอบตั้งแต่เริ่มต้นจนสิ้นสุดการทดสอบ คาดการณ์ความเป็นไปได้ในการทดสอบและวิธีเก็บรวบรวมผลการทดสอบ

ขั้นตอนที่ 3 ออกแบบโปรแกรม R ซึ่งเป็นภาษาที่เหมาะสำหรับปฏิบัติการคำนวณชั้นสูงและใช้กับเครื่องคอมพิวเตอร์ระบบไมโครคอมพิวเตอร์ได้ โปรแกรมออกแบบขึ้นสำหรับจัดกระทำข้อมูลที่ใช้ในการวิจัย โดยเขียนคำสั่งต่างๆ ทีละขั้นตอนการทำงานเป็นไปตามลำดับและมีความถูกต้องครบถ้วนตามผังขั้นตอนการทดสอบ

ขั้นตอนที่ 4 ทดสอบและปรับปรุงโปรแกรม การตรวจสอบทำตามขั้นตอนการทำงานของโปรแกรมโดยติดตามการทำงานทีละคำสั่งจนกว่าจะจบคำสั่ง ทดสอบโดยการควบคุมเครื่องคอมพิวเตอร์ให้ทำงานตามคำสั่งด้วยข้อมูลที่รู้ค่าผลลัพธ์แล้วและตรวจสอบว่าได้ผลตรงกันหรือไม่ถ้าพบข้อบกพร่องจะทำการปรับปรุงแก้ไขและนำไปทดสอบซ้ำ เมื่อแน่ใจว่าการทำงานของโปรแกรมถูกต้องและผลการทดสอบแม่นยำเพียงพอสำหรับเก็บข้อมูลตัวแปรตามได้ครบถ้วน จึงนำไปใช้ทดสอบในงานวิจัยต่อไป

วิธีการเลือกข้อสอบที่ใช้ในการศึกษาครั้งนี้มีรายละเอียดดังต่อไปนี้

1. ขั้นตอนวิธีของวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ แบบมอนติ คาร์โล ซีเอที (Monte Carlo CAT Method: MCC)

ผู้วิจัยได้ประยุกต์ใช้วิธีการ Monte Carlo CAT ที่เสนอโดย Belov, Armstrong, และ Weissman (2008) เพื่อใช้ในการควบคุมเงื่อนไขบังคับที่ไม่ใช่ทางสถิติ เช่น เงื่อนไขบังคับด้านเนื้อหาและเงื่อนไขบังคับด้านการควบคุมการใช้ข้อสอบซ้ำ โดยผู้วิจัยนำมาประยุกต์ให้เหมาะสมกับการศึกษาครั้งนี้และใช้กับโมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย (TRT) Mao and Xin (2013) รายละเอียดมีดังนี้

กำหนดปัจจัยนำเข้า (Input) ของขั้นตอนวิธี มีดังนี้

1) ค่าประมาณความสามารถปัจจุบันของผู้สอบ ($\hat{\theta}_i$) ในช่วงเริ่มต้นการทดสอบกำหนดให้ผู้สอบแต่ละคนมี $\hat{\theta}_i$ = 0 และจะปรับปรุงค่า $\hat{\theta}_i$ ทุกครั้งหลังจากที่ผู้สอบทำข้อสอบแต่ละข้อแล้ว

2) กำหนดพารามิเตอร์ความยาวของแบบสอบ (n) ในการศึกษาครั้งนี้ ผู้สอบสามารถทำแบบทดสอบย่อยได้สูงสุด 15 ชุด หรือมี SEE ≤ 0.3

3) กำหนดพารามิเตอร์ l ทำหน้าที่นับจำนวนแบบทดสอบย่อยที่นำไปใช้กับผู้สอบแต่ละคนเรียบร้อยแล้ว (เมื่อเริ่มต้นการทดสอบ l=0 และมีค่าอยู่ในช่วง (0≤l<n)

4) กำหนดพารามิเตอร์ k=2l+1 โดยใช้สูตรจากงานวิจัยของ Belov et al. (2008) เพื่อใช้กำหนดจำนวนแบบทดสอบย่อยที่ต้องสุ่มเลือกในแต่ละครั้งของการเลือกแบบทดสอบย่อย

5) กำหนดตัวแปร m สำหรับเก็บแบบสอบเงา ที่เกิดจากการรวมแบบทดสอบย่อย (Assembly of Testlets) ภายในคลังข้อสอบอย่างสุ่ม ซึ่งแบบสอบเงาแต่ละฉบับจะมีความยาวเท่ากับ n-l แบบทดสอบย่อย

6) กำหนดตัวแปรเซต S เพื่อเก็บแบบสอบเงาที่มีเงื่อนไขบังคับครบถ้วนตามที่ผู้วิจัยกำหนด

7) กำหนดตัวแปร r เพื่อนับจำนวนแบบสอบเงาที่มีในเซต S

8) กำหนดตัวแปรลำดับ (sequence) เพื่อนำแบบทดสอบย่อยภายในเซต S เรียงต่อกัน

ขั้นตอนวิธี (Algorithm) มีดังนี้

ขั้นที่ 1 ผู้วิจัยใช้แพ็คเกจ lpSolve ของโปรแกรม R ทำกำหนดการเชิงเส้น (linear programming) เพื่อรวม testlet ให้เป็นแบบสอบอย่างอัตโนมัติ (automated test assembly) โดยแบบสอบที่รวมขึ้นจะมีเงื่อนไขบังคับครบถ้วนตามที่ผู้วิจัยกำหนด และแบบสอบทั้งหมดจะมีจำนวน r ชุด (form) ผู้วิจัยจะเรียกแบบสอบนี้ว่าแบบสอบเงา (shadow tests) สำหรับคำสั่งที่ใช้ผู้วิจัยประยุกต์ใช้ชุดคำสั่ง R จากงานวิจัยของ Diao and van der Linden (2011) เพื่อเขียนการโปรแกรมเชิงจำนวนเต็มแบบผสม (mixed integer programming) ซึ่งเรียกใช้งานผ่านแพ็คเกจ IP_solve ให้คำนวณหาค่าที่เหมาะสมที่สุด (optimal) สำหรับฟังก์ชั่นเป้าหมาย (objective

function) ที่ฟังก์ชั่นเงื่อนไขบังคับ (constrains function) ที่กำหนดขึ้นเพื่อใช้ในการรวมแบบทดสอบเงาจากแบบทดสอบย่อยภายในคลังข้อสอบ

เงื่อนไขบังคับสำหรับทำ automated test assembly ที่ผู้วิจัยกำหนดมีดังนี้

C1. ภายในแบบสอบเงาจะไม่มีแบบทดสอบย่อยซ้ำกัน

C2. แบบสอบเงาต้องรวมแบบทดสอบย่อยถูกที่นำไปใช้กับผู้สอบแล้ว

C3. จำนวนแบบทดสอบย่อยทั้งหมดในแบบสอบเงามีจำนวนเท่ากับ n ชุด

C4. การกระจายของเนื้อหาวิชาในแบบสอบเงาแต่ละฉบับต้องตรงตามที่ผู้วิจัยกำหนด คือ แบบสอบจะต้องมีเนื้อหาวิชา 3 วิชา กระจายเท่าๆ กัน

ขั้นที่ 2 แต่ละแบบสอบเงาในเซต S มีจำนวนแบบทดสอบย่อยที่ยังไม่ได้นำไปใช้กับผู้สอบจำนวน (n-l) ชุด จากนั้นเพิ่มแบบทดสอบย่อยทั้งหมดที่มีในตัวแปรเซต S ไปในลำดับ (sequence) ผลลัพธ์ของลำดับจะมีจำนวนแบบทดสอบ่อยเท่ากับ m(n-l) ชุด (หมายเหตุ แบบสอบเงาในเซต S สามารถทับซ้อนกันกับแบบสอบเงาอื่นได้ ดั้งนั้น แบบสอบย่อยที่ซ้ำกันสามารถพบในลำดับได้หลายครั้ง

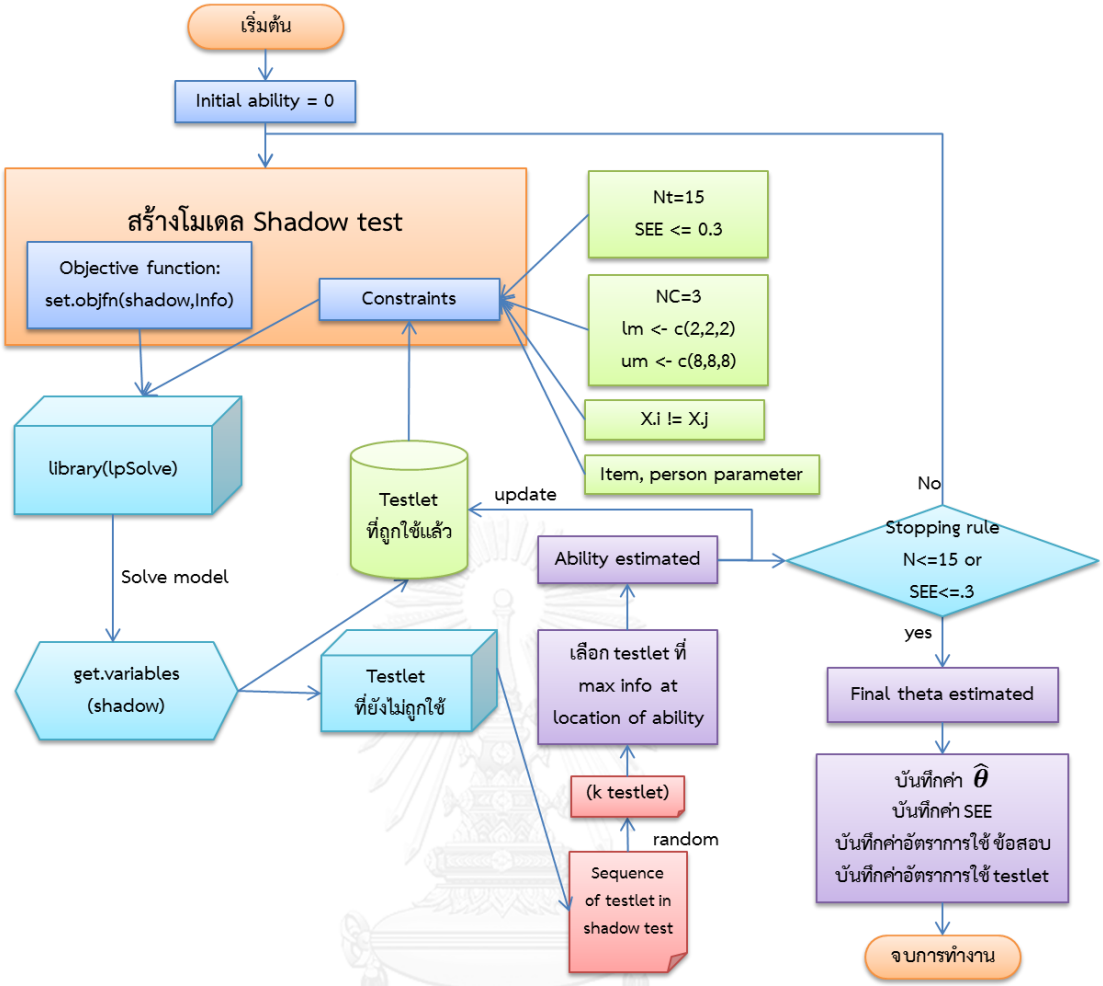
ขั้นที่ 3 สุ่มหยิบแบบทดสอบย่อย k ชุด จากลำดับในขั้นตอนที่ 2

ขั้นที่ 4 ค้นหาแบบทดสอบย่อยจากขั้นตอนที่ 3 ที่ให้สารสนเทศสูงสุด ณ ตำแหน่ง $\hat{\theta}$ และกำหนดให้แบบทดสอบย่อยชุดนั้นเป็นแบบทดสอบย่อยชุดถัดไป (φ) ที่จะนำไปใช้กับผู้สอบ

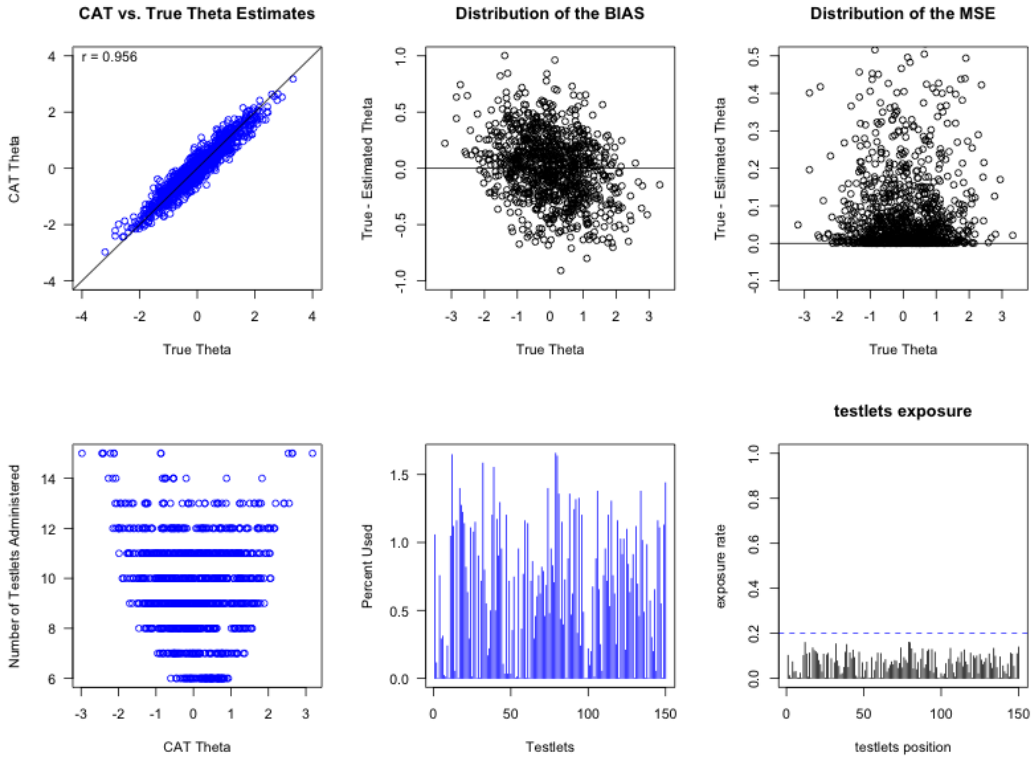
ขั้นที่ 5 ปรับปรุงสถานะของเซต S ให้เห็นปัจจุบัน โดยเก็บแบบทดสอบเงาทั้งหมดที่มี φ บรรจุอยู่ และลบแบบสอบเงาทั้งหมดที่ไม่มี φ ออกจาก เซต S

ผลลัพธ์ (Output) ของขั้นตอนวิธี คือ แบบทดสอบย่อยชุดถัดไป (φ) ที่จะนำไปใช้กับผู้สอบ และประมาณค่าความสามารถผู้สอบด้วยการประมาณค่าแบบ Expected A Posteriori (EAP)

รายละเอียดขั้นตอนวิธีของ Monte Carlo CAT แสดงดังภาพที่ 3.16 และผลลัพธ์จากการทดสอบดังภาพที่ 3.17



ภาพที่ 3.16 ผังแสดงขั้นตอนวิธีของ Monte Carlo CAT



ภาพที่ 3.17 ผลลัพธ์จากการทดสอบ 1 รอบ ของวิธี Monte Carlo CAT

2. ขั้นตอนวิธีของวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (constraint-weighted a-stratification CAT Method: CWA)

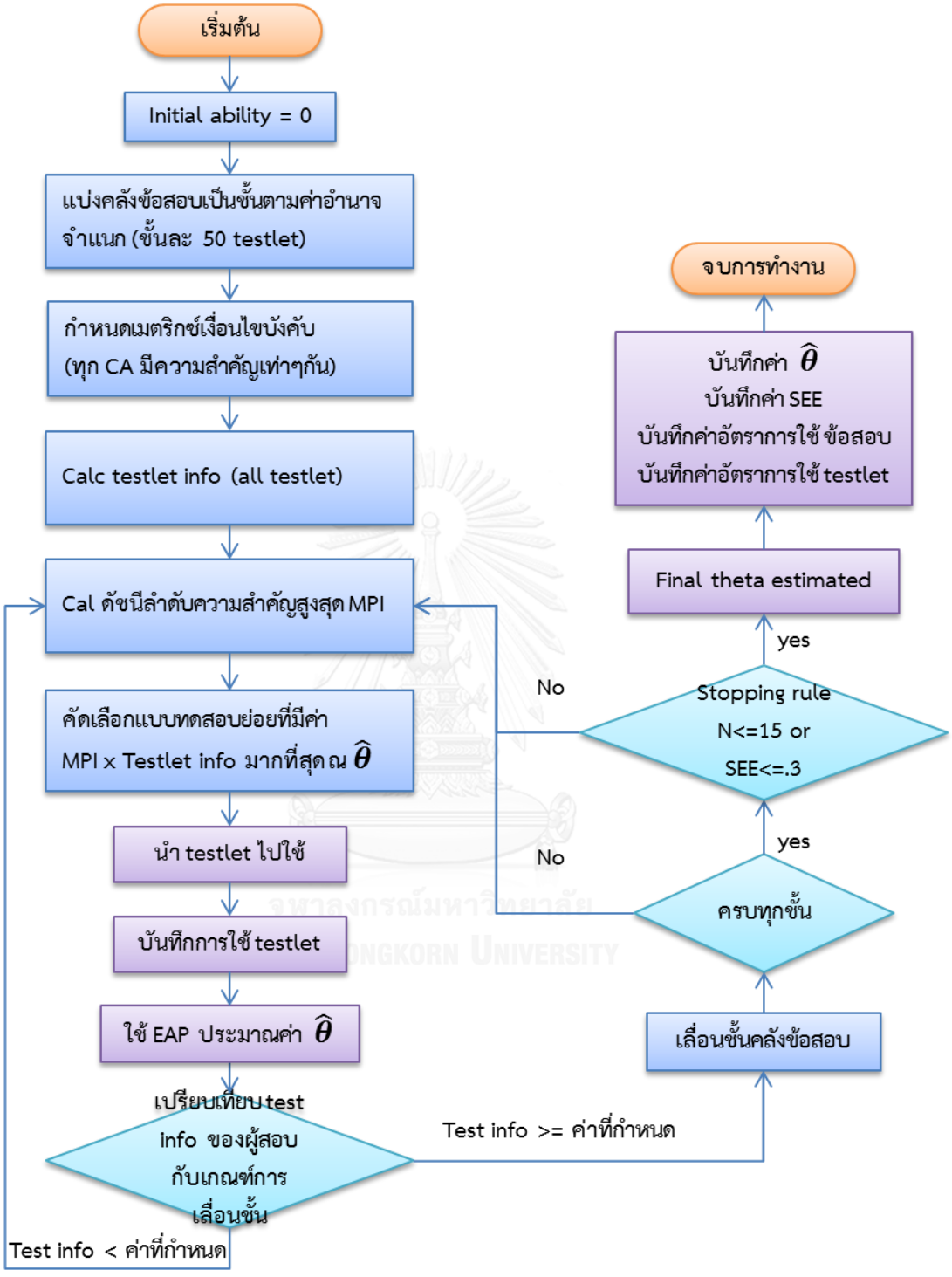
ขั้นที่ 1 แบ่งคลังข้อสอบเป็นชั้นตามค่าอำนาจจำแนกเฉลี่ยของแบบทดสอบย่อย ชั้นแรกบรรจุแบบทดสอบย่อยที่มีอำนาจจำแนกเฉลี่ยต่ำสุด ชั้น 2 บรรจุแบบทดสอบย่อยอำนาจจำแนกเฉลี่ยสูงกว่าชั้นแรก และเพิ่มขึ้นเรื่อยๆ ตามลำดับชั้น จนกระทั่งชั้นสุดท้ายบรรจุข้อสอบอำนาจจำแนกสูงสุด และกำหนดให้แต่ละชั้นของคลังข้อสอบมีจำนวนข้อสอบ เท่าๆ กัน โดยแต่ละชั้นมีแบบทดสอบย่อยทั้งหมด 50 ชุด คิดเป็นข้อสอบจำนวน 200 ข้อ ข้อสอบในแต่ละชั้นจะบรรจุด้วยแบบทดสอบย่อยที่มีค่าอำนาจจำแนกเฉลี่ยต่ำที่สุด ไปจนถึงค่าอำนาจสูงที่สุดที่อยู่ในชั้นท้าย และภายในแต่ละชั้นของคลังข้อสอบจะแบบทดสอบย่อยที่มีค่าความยากเฉลี่ยกระจายตัวแบบยูนิฟอร์ม

ขั้นที่ 2 กำหนดเมตริกซ์เงื่อนไขบังคับ ซึ่งจะบอกว่าข้อสอบแต่ละข้อประกอบด้วยเงื่อนไขบังคับใดบ้าง โดยกำหนดให้เมตริกซ์ดังกล่าว คือ เมตริกซ์ C.C เมื่อ c_{jk}= 1 แสดงว่า ข้อสอบข้อที่ j เกี่ยวข้องกับเงื่อนไขบังคับที่ K ตัวอย่างเมตริกซ์เงื่อนไขบังคับ ดังแสดงในภาคผนวก

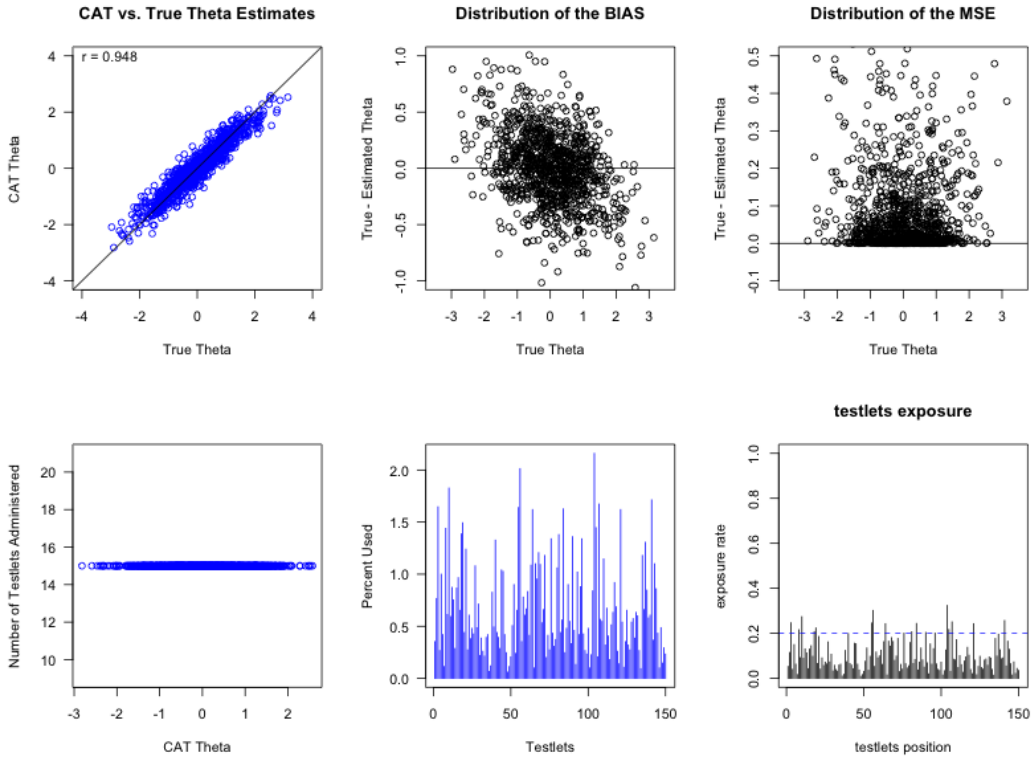
ขั้นที่ 3 คำนวณดัชนีลำดับความสำคัญสูงสุด (Maximum Priority Index: MPI) โดยนำค่าลำดับความสำคัญซึ่งกำหนดโดยผู้วิจัยมาพิจารณาร่วมในการคัดเลือกข้อสอบ ซึ่งพิจารณาจากผลคูณของค่าสารสนเทศของข้อสอบกับดัชนีลำดับความสำคัญ ข้อใดให้ผลคูณมากที่สุด ข้อนั้นจะได้รับการคัดเลือกให้นำไปใช้กับผู้สอบ

ขั้นที่ 4 คัดเลือกแบบทดสอบย่อยที่มีค่า MPI มากที่สุด ณ ตำแหน่งค่าความสามารถของผู้สอบเมื่อแบบทดสอบย่อยถูกเลือกและนำไปใช้กับผู้สอบ เมื่อผู้สอบตอบ ผลการตอบข้อสอบจะถูกนำไปคำนวณหาค่าประมาณความสามารถด้วยการประมาณค่าแบบ Expected A Posteriori (EAP) จากนั้นจึงนำค่าความสามารถของผู้สอบ ณ ตำแหน่งปัจจุบันไปใช้เลือกแบบทดสอบย่อยชุดต่อไป

ขั้นที่ 5 เปรียบเทียบสารสนเทศของแบบสอบกับเกณฑ์ที่ใช้พิจารณาการเลื่อนชั้นคลังข้อสอบ ถ้าค่าสารสนเทศของแบบสอบมีค่าสูงกว่าเกณฑ์ที่ตั้งไว้ แบบทดสอบย่อยจะถูกเลือกจากชั้นถัดไปของคลังข้อสอบ เกณฑ์ที่ใช้กำหนดการเลื่อนชั้นของคลังข้อสอบมีดังนี้ 1) คลังข้อสอบ 600 ข้อ กำหนดค่าสารสนเทศของแบบสอบไว้ที่ 2.7, 6.2 และ 11.112 2) คลังข้อสอบ 800 ข้อ กำหนดค่าสารสนเทศของแบบสอบไว้ที่ 1.112, 3.336, 6.672, และ 11.112 รายละเอียดขั้นตอนวิธีของ constraint-weighted a-stratification CAT ดังภาพที่ 3.18 และผลลัพธ์จากการทดสอบ ดังภาพที่ 3.19



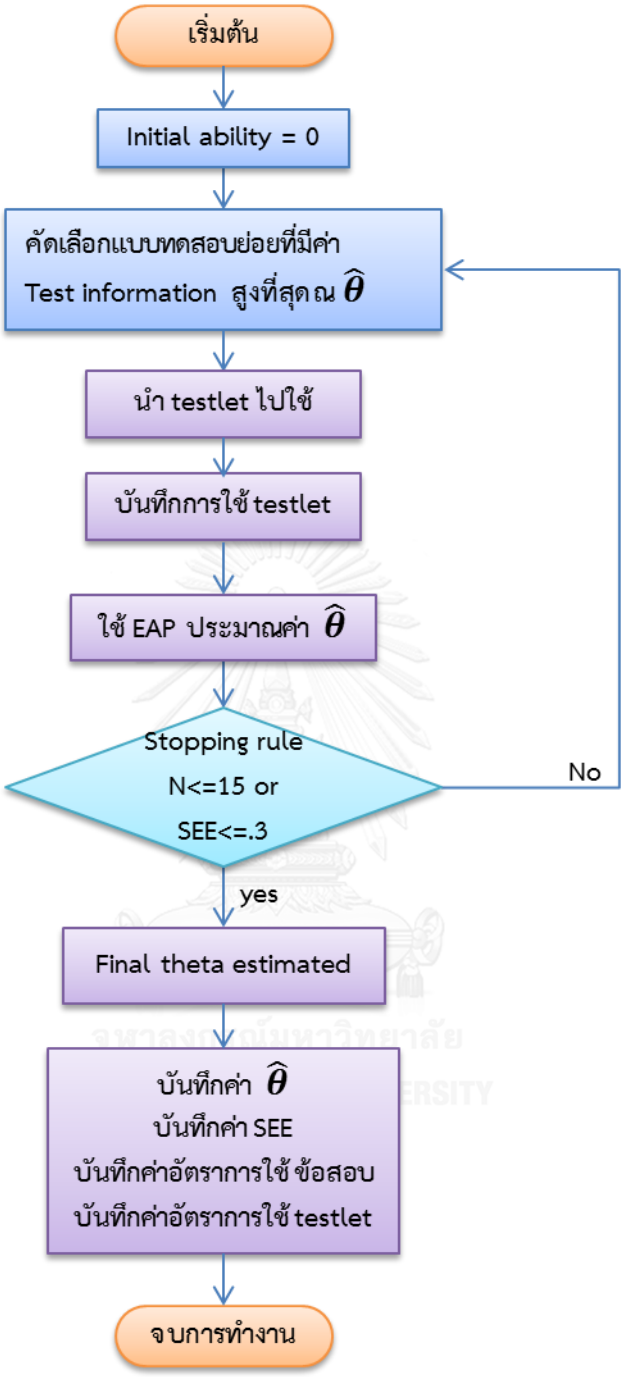
ภาพที่ 3.18 ผังแสดงขั้นตอนวิธีของ constraint-weighted a-stratification CAT



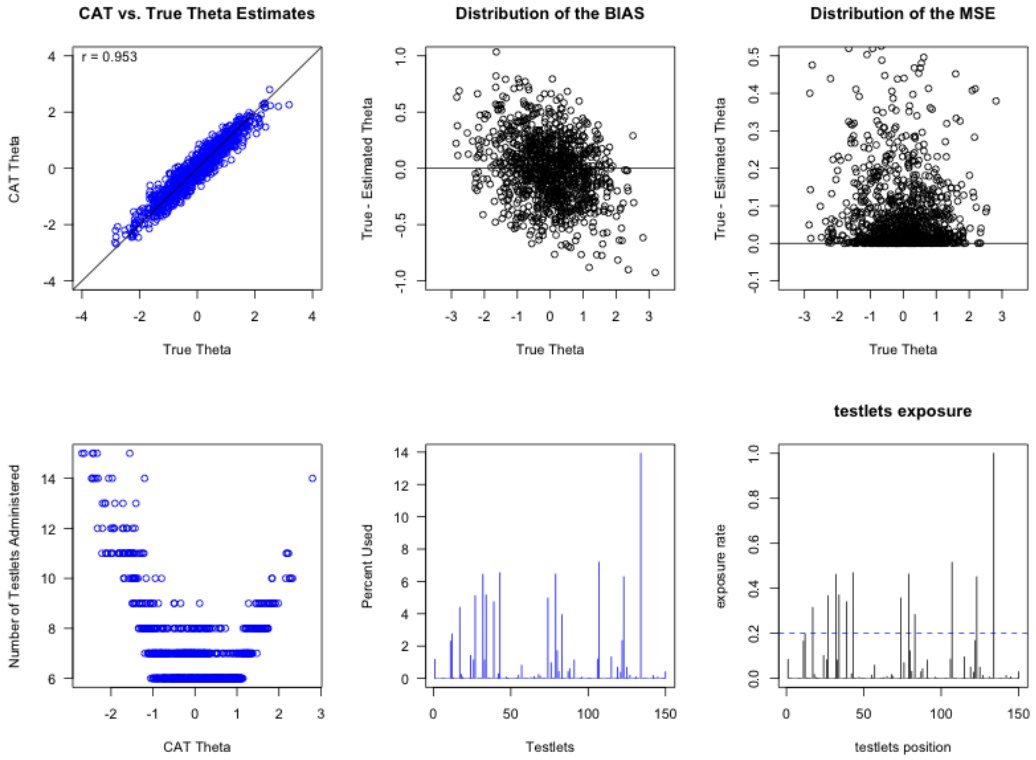
ภาพที่ 3.19 ผลลัพธ์จากการทดสอบ 1 รอบ ของวิธี constraint-weighted a-stratification CAT

3. ขั้นตอนวิธีของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบค่าสารสนเทศสูงสุด (Maximum Fisher's information method)

วิธีของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบค่าสารสนเทศสูงสุด ในขั้นเริ่มต้นการทดสอบผู้วิจัยกำหนดให้ผู้สอบทุกคนมีความสามารถอยู่ในระดับปานกลาง (θ=0) การเลือกแบบทดสอบย่อยจะเลือกแบบทดสอบย่อยที่มีค่าฟังก์ชันสารสนเทศของแบบสอบสูงที่สุด ณ ตำแหน่งความสามารถประมาณค่าของผู้สอบเมื่อผู้สอบตอบข้อสอบทุกข้อในแบบทดสอบย่อย โปรแกรมจะประมาณค่าความสามารถของผู้สอบด้วยการประมาณค่าแบบ Expected A Posteriori และจะทำซ้ำไปเรื่อยๆ จนกระทั่ง การประมาณความสามารถของผู้สอบมีความคลาดเคลื่อนน้อยกว่าหรือเท่ากับ 0.3 หรือได้รับแบบทดสอบย่อยจำนวน 15 ฉบับแล้วจึงหยุดการทดสอบ และผู้วิจัยใช้วิธีการนี้เป็นฐานในการเปรียบเทียบเนื่องจากมีประสิทธิภาพในด้านความถูกต้องแม่ยำของการวัดสูงที่สุด รายละเอียดขั้นตอนวิธีของของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบค่าสารสนเทศ แสดงดังภาพที่ 3.20 และผลลัพธ์จากการทดสอบ ดังภาพที่ 3.21



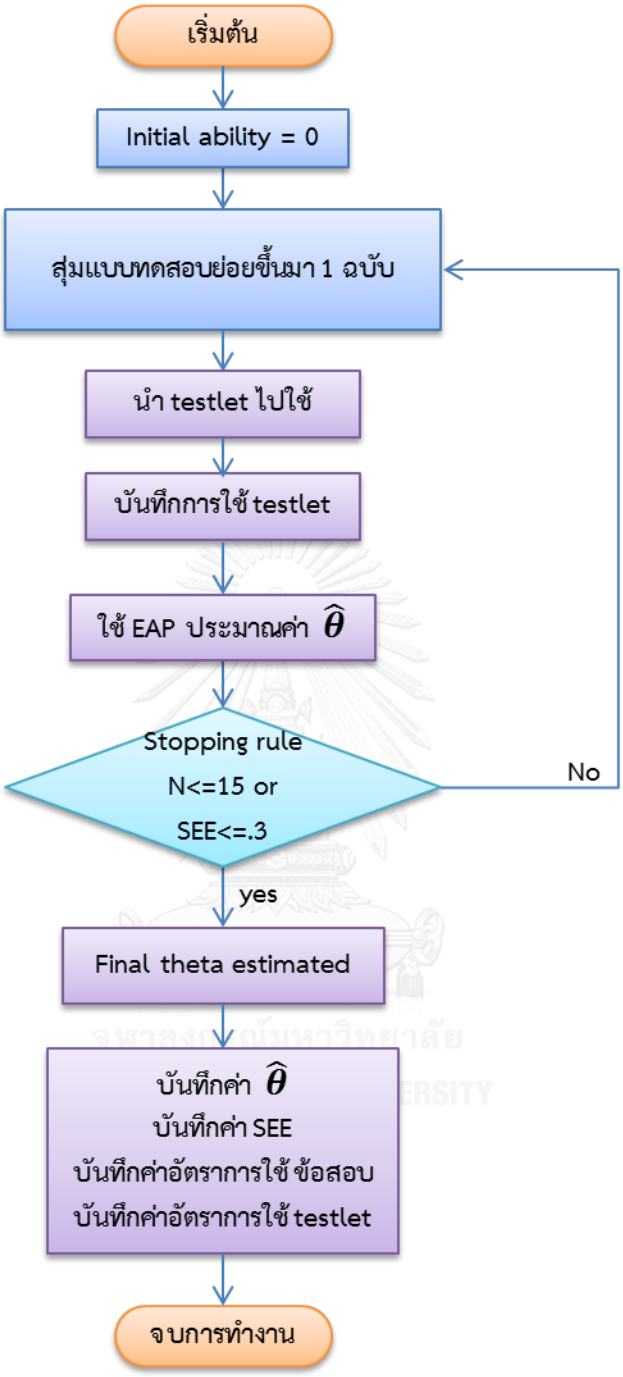
ภาพที่ 3.20 ผังแสดงขั้นตอนวิธีของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบค่าสารสนเทศสูงสุด



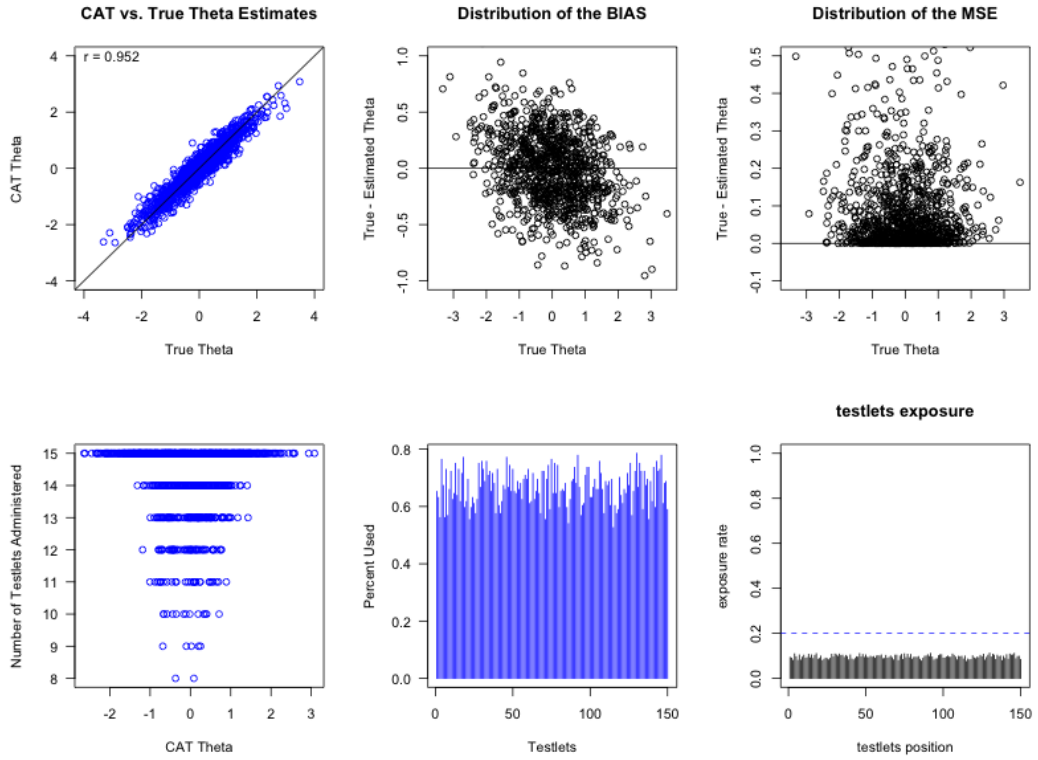
ภาพที่ 3.21 ผลลัพธ์จากการทดสอบ 1 รอบ ของวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบค่าสารสนเทศสูงสุด

4. ขั้นตอนวิธีของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบสุ่ม (Randomization method)

ในขั้นเริ่มต้นการทดสอบผู้วิจัยกำหนดให้ผู้สอบทุกคนมีความสามารถอยู่ในระดับปานกลาง (θ=0) การเลือกแบบทดสอบย่อยจะเลือกแบบทดสอบย่อยที่มีค่าฟังก์ชันสารสนเทศของแบบสอบสูงที่สุด ณ ตำแหน่งความสามารถประมาณค่าของผู้สอบเมื่อผู้สอบตอบข้อสอบทุกข้อในแบบทดสอบย่อย โปรแกรมจะประมาณค่าความสามารถของผู้สอบด้วยการประมาณค่าแบบ Expected A Posteriori และจะทำซ้ำไปเรื่อยๆ จนกระทั่ง การประมาณความสามารถของผู้สอบมีความคลาดเคลื่อนน้อยกว่าหรือเท่ากับ 0.3 หรือได้รับแบบทดสอบย่อยจำนวน 15 ฉบับแล้วจึงหยุดการทดสอบ และผู้วิจัยใช้เป็นฐานในการเปรียบเทียบเนื่องจากมีประสิทธิภาพในด้านการใช้คลังข้อสอบสูงที่สุด รายละเอียดขั้นตอนวิธีของของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบค่าสารสนเทศแสดงดังภาพที่ 3.22 และผลลัพธ์จากการทดสอบ ดังภาพที่ 3.23



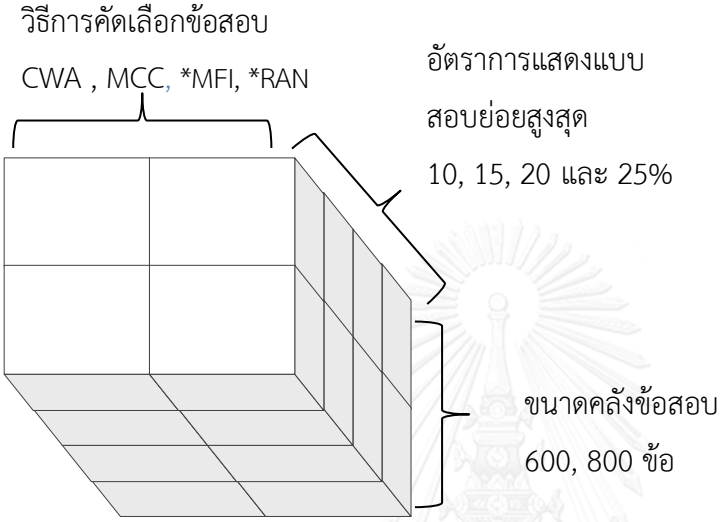
ภาพที่ 3.22 ผังแสดงขั้นตอนวิธีของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบสุ่ม



ภาพที่ 3.23 ผลลัพธ์จากการทดสอบ 1 รอบ ของวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบสุ่ม

การวิเคราะห์ข้อมูล

การวิจัยครั้งนี้ได้ออกแบบการวิจัยแบบคอมพลีทลี แรมดอมไมซ์ แฟคทอเรียล ดีไซน์ (Completely Randomized Factorial Design) โดยใช้ขนาดกลุ่มตัวอย่างในแต่ละเซลล์ 1,000 คน ทำการทดลอง ซ้ำ 10 ครั้ง และจะใช้ผลของการทดลองซ้ำแต่ละครั้งเป็นหน่วยในการวิเคราะห์เปรียบเทียบ รายละเอียดดัง ภาพที่ 3.24



ภาพที่ 3.24 แสดงรูปแบบการทดลอง

ขั้นตอนการวิเคราะห์ข้อมูล

1. ในแต่ละครั้งการทดลองจะ หาค่าต่างๆ ดังนี้

1.1 หาความลำเอียงเฉลี่ย หรือค่าความคลาดเคลื่อนในการประมาณค่าความสามารถของผู้สอบจากการทดสอบแบบปรับเหมาะ โดยใช้สูตร $Bias(\theta) = \frac{1}{m}\sum_{i=1}^{m}(\hat{\theta}_i - \theta_i)$

1.2 หาค่าความแปรปรวนของคลาดเคลื่อน โดยใช้สูตร

$$MSE(\theta) = \frac{1}{m}\sum_{i=1}^{m}(\hat{\theta}_i - \theta_i)^2$$

1.3 หาค่าเฉลี่ยของความสัมพันธ์ระหว่างค่าความสารถจริงกับค่าความสามารถประมาณค่า

1.4 หาค่าเฉลี่ยของความยาวแบบสอบและค่าเฉลี่ยความคลาดเคลื่อนมาตรฐานของการประมาณค่า

1.5 หาค่าอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ (testlet exposure rate observed) โดยใช้สูตร $r_{obs} = \sum_{i=1}^{m}(\frac{n.tl.used_i}{m})$ เมื่อ $n.tl.used$ คือจำนวนครั้งของแบบทดสอบย่อยที่ถูกใช้ และ m คือ จำนวนผู้สอบทั้งหมด (Chang et al., 2001)

1.6 หาค่าอัตราการทับซ้อนของแบบสอบเฉลี่ย (average test overlap) โดยใช้สูตร $\bar{t} = \frac{s^2 + u^2}{u}$ เมื่อ u และ s^2 คือ ค่าเฉลี่ย และความแปรปรวนของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ (Chen et al., 2003)

1.7 ความสมดุลของการใช้คลังข้อสอบในภาพรวม (overall of pool usage balance) โดยใช้สูตร $x^2 = \sum_{j=1}^{N} \frac{(er_j - L/N)^2}{L/N}$ เมื่อ er_j อัตราการใช้ข้อสอบซ้ำข้อที่ j ที่สังเกตได้ L คือความยาวของแบบสอบ และ N คือ จำนวนข้อสอบทั้งหมด

1.8 หาค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้

1.9 หาค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ถูกนำมาใช้มากเกินไป (มากกว่าค่า $r_{\max}$)

2. จากการทดลองซ้ำ 10 ครั้ง ในแต่ละเงื่อนไขการทดลองก็จะมีค่าที่กล่าวมา 1.1-1.9 จำนวนเงื่อนไขละ 10 ค่า

2.1 เขียนกราฟเปรียบเทียบค่าเฉลี่ยที่ได้ในข้อที่ 1

2.2 การเลือกใช้สถิติทดสอบ เนื่องจากจำนวนหน่วยวิเคราะห์ในแต่ละเซลล์มีเพียง 10 ตัวอย่าง ดังนั้นการใช้สถิติทดสอบแบบพาราเมตริก (Parametric statistics) มีความไม่เหมาะสมด้วยเหตุผลดังนี้ คือ จะทำให้การทดสอบนั้นมีอำนาจการทดสอบต่ำเนื่องจากหน่วยวิเคราะห์มีจำนวนน้อยและไม่แจกแจงเป็นโค้งปกติ ดังนั้นผู้วิจัยจึงเปรียบเทียบความแตกต่างของค่าเฉลี่ยในข้อที่ 1 โดยการวิเคราะห์ด้วยสถิติแบบนอนพาราเมตริก (Nonparametric statistics) สำหรับกลุ่มตัวอย่าง k กลุ่มที่มีความเป็นอิสระต่อกัน ด้วยสถิติ Kruskal-Wallis Test ถ้าผลการวิเคราะห์ในข้อ 2.2 มีนัยสำคัญทางสถิติ จะทำการเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ต่อไป

บทที่ 4
ผลการวิเคราะห์ข้อมูล

งานวิจัยนี้มีวัตถุประสงค์ เพื่อศึกษาเปรียบเทียบประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถของผู้สอบและประสิทธิภาพด้านการใช้คลังข้อสอบระหว่างวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที กับวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ เมื่อกำหนดอัตราการควบคุมการแสดงแบบทดสอบสูงสุด (r_{max}) และขนาดคลังข้อสอบต่างกัน โดยใช้วิธีจำลองข้อมูล ผลการวิจัยแบ่งการนำเสนอข้อมูลออกเป็น 3 ตอน ดังนี้

ตอนที่ 1 ผลการวิเคราะห์ข้อมูลเบื้องต้นในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์

1.1 สถิติพื้นฐานของประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถผู้สอบของวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์

1.2 สถิติพื้นฐานของประสิทธิภาพด้านการใช้คลังข้อสอบ

ตอนที่ 2 ผลการวิเคราะห์ประสิทธิภาพของวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์สำหรับโมเดลการตอบสนองแบบทดสอบย่อย

2.1 ประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถของผู้สอบ

2.2 ประสิทธิภาพด้านการใช้คลังข้อสอบ

ตอนที่ 3 ผลการเปรียบเทียบประสิทธิภาพของวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์

3.1 ผลการเปรียบเทียบประสิทธิภาพด้านความถูกต้องแม่นยำในการวัด

3.2 ผลการเปรียบเทียบประสิทธิภาพด้านการใช้คลังข้อสอบ

เพื่อให้ง่ายต่อการนำเสนอผลการวิเคราะห์ข้อมูล ผู้วิจัยจึงกำหนดสัญลักษณ์ที่ใช้แทนค่าตัวแปรและค่าสถิติต่าง ๆ ไว้ดังนี้

MCC10 หมายถึง การทดสอบแบบปรับเหมาะที่ใช้วิธีการคัดเลือกข้อสอบแบบมอนติ คาร์โล ซีเอที ที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดไว้ เท่ากับ 10 เปอร์เซ็นต์

MCC15 หมายถึง การทดสอบแบบปรับเหมาะที่ใช้วิธีการคัดเลือกข้อสอบแบบมอนติ คาร์โล ซีเอที ที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดไว้ เท่ากับ 15 เปอร์เซ็นต์

MCC20 หมายถึง การทดสอบแบบปรับเหมาะที่ใช้วิธีการคัดเลือกข้อสอบแบบมอนติ คาร์โล ซีเอที ที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดไว้ เท่ากับ 20 เปอร์เซ็นต์

MCC25 หมายถึง การทดสอบแบบปรับเหมาะที่ใช้วิธีการคัดเลือกข้อสอบแบบมอนติ คาร์โล ซีเอที ที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดไว้ เท่ากับ 25 เปอร์เซ็นต์

CWA10 หมายถึง การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดไว้เท่ากับ 10 เปอร์เซ็นต์

CWA15 หมายถึง การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดไว้เท่ากับ 15 เปอร์เซ็นต์

CWA20 หมายถึง การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดไว้เท่ากับ 20 เปอร์เซ็นต์

CWA25 หมายถึง การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดไว้เท่ากับ 25 เปอร์เซ็นต์

θ_i หมายถึง ค่าความสามารถของผู้สอบคนที่ i

$\hat{\theta}_i$ หมายถึง ค่าความสามารถที่ได้จากการประมาณค่าของผู้สอบคนที่ i

$r_{\theta\hat{\theta}}$ หมายถึง ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าความสามารถจริงกับความสามารถประมาณค่า

N หมายถึง จำนวนผู้สอบทั้งหมด

M หมายถึง ค่ามัชฌิมเลขคณิต หรือค่าเฉลี่ย (mean)

SD หมายถึง ส่วนเบี่ยงเบนมาตรฐาน (standard deviation)

Min หมายถึง ค่าต่ำสุด

Max หมายถึง ค่าสูงสุด

M_{tls} หมายถึง ค่าเฉลี่ยความยาวของแบบสอบที่ผู้สอบได้รับมีหน่วยการนับเป็นจำนวนแบบทดสอบย่อย

M_{it} หมายถึง ค่าเฉลี่ยของความยาวของแบบสอบที่ผู้สอบได้รับมีหน่วยการนับเป็นจำนวนข้อสอบ

M_{SEE} หมายถึง ค่าเฉลี่ยของความคลาดเคลื่อนมาตรฐานของการประมาณค่า

M_{Bias} หมายถึง ค่าเฉลี่ยของความลำเอียงเฉลี่ย

M_{MSE} หมายถึง ค่าเฉลี่ยความแปรปรวนของความคลาดเคลื่อน

r_{max} หมายถึง อัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด

r_{obs} หมายถึง อัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้

$\bar{t}$ หมายถึง อัตราการทับซ้อนกันของแบบสอบ

x^2 หมายถึง ค่าที่แสดงถึงสมดุลการใช้คลังข้อสอบในภาพรวม

$N_{neverexposed}$ หมายถึง จำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้

$N_{overexposed}$ หมายถึง จำนวนแบบทดสอบย่อยที่ถูกนำมาใช้มากเกินไป

p หมายถึง ระดับนัยสำคัญทางสถิติ

ตอนที่ 1 ผลการวิเคราะห์ข้อมูลเบื้องต้นในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์

1.1 สถิติพื้นฐานของประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถผู้สอบของวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์

การวิเคราะห์ในส่วนนี้ ผู้วิจัยนำเสนอผลการวิเคราะห์ประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถผู้สอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ด้วยวิธีการเลือกแบบทดสอบย่อยที่ให้ค่าสารสนเทศสูงสุด วิธีการสุ่ม วิธีแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ และวิธีมอนติ คาร์โล ซีเอที โดยจำแนกตาม ขนาดคลังข้อสอบ จากผลการวิเคราะห์ค่าสถิติพื้นฐาน พบว่า ค่าเฉลี่ยของความคลาดเคลื่อนมาตรฐานของการประมาณค่า (SEE) ของทุกเงื่อนไขการทดสอบ ที่ใช้กับคลังข้อสอบทั้งขนาด 600 และ 800 ข้อ มีค่า SEE ประมาณ 0.290 ถึง 0.30 ซึ่งถือว่าใกล้เคียงกันมาก ยกเว้นวิธีการคัดเลือกแบบทดสอบย่อยแบบสุ่มที่มีค่า SEE มากกว่า 0.31 ด้วยเหตุนี้ผู้วิจัยจึงนำความยาวของแบบสอบมาเป็นตัวชี้วัดตัวหนึ่งในการพิจารณาประสิทธิภาพของความถูกต้องแม่นยำในการวัดตามแนวคิดของ Wainer (1992) ที่กล่าวว่าประสิทธิภาพของความถูกต้องแม่นยำในการวัดนั้นเกี่ยวข้องกับการได้รับความแม่นยำของการวัดในระดับเดียวกันขณะที่ใช้จำนวนข้อสอบน้อยกว่า โดยผู้วิจัยนำความยาวของแบบสอบเฉลี่ยไปทดสอบทางสถิติเพื่อเปรียบเทียบให้เห็นความแตกต่าง รายละเอียดดังเสนอใน ตอนที่ 3 ของการรายงานผลการวิเคราะห์ข้อมูล

1) การทดลองกับคลังข้อสอบที่มีข้อสอบจำนวน 600 ข้อ ภายในคลังข้อสอบถูกจัดเป็นแบบทดสอบย่อยได้ 150 ชุด โดยแบบทดสอบย่อยแต่ละชุดมีข้อสอบไม่ซ้ำกัน ผลการทดลองมีค่าเฉลี่ยตัวชี้วัดที่ใช้พิจารณาประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถผู้สอบดังนี้ วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (CWA) พบว่า เมื่อเพิ่มขนาดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดจาก 10 เปอร์เซ็นต์ เป็น 15, 20 และ 25 เปอร์เซ็นต์ ตามลำดับแล้ว ปรากฏว่า ความยาวของแบบสอบเฉลี่ยมีทั้งเพิ่มขึ้นและลดลง คือ จากเดิมใช้ข้อสอบประมาณ 54 ข้อ ลดลงมาเป็นใช้ข้อสอบประมาณ 51 ข้อ แต่เมื่อพิจารณาจำนวนแบบทดสอบย่อยเฉลี่ยที่ใช้พบว่าจะมีค่าใกล้เคียงกันคือ ใช้ประมาณ 13 ชุด ขณะที่วิธีมอนติ คาร์โล ซีเอที (MCC) ความยาวของแบบสอบเฉลี่ยที่ใช้ และจำนวนแบบทดสอบย่อยเฉลี่ยที่ใช้มีแนวโน้มลงลดอย่างเห็นได้ชัด สาเหตุที่วิธี CWA มีจำนวนข้อที่ใช้ไม่ลดลงอย่างชัดเจน เป็นเพราะวิธีการดังกล่าวแบ่งคลังข้อสอบออกเป็นชั้นๆ ตามลำดับค่าพารามิเตอร์อำนาจจำแนกเฉลี่ยของแต่ละแบบทดสอบย่อย และคลังข้อสอบดังกล่าวมีขนาดเล็กเนื่องจากมีแบบทดสอบย่อยให้เลือกสำหรับนำไปใช้กับผู้สอบเพียง 150 ชุด เมื่อคลังข้อสอบถูกแบ่งเป็นชั้น ทำให้ชุดแบบทดสอบย่อยที่มี

ให้เลือกในแต่ละชั้นมีจำนวนน้อยลง คือ มีชั้นละ 50 ชุด ส่วนสาเหตุที่วิธี MCC มีจำนวนข้อที่ใช้ลดลงอย่างชัดเจน เป็นเพราะวิธีการดังกล่าวไม่แบ่งคลังข้อสอบออกเป็นชั้นๆ ตามลำดับค่าพารามิเตอร์อำนาจจำแนกเฉลี่ย เมื่อพิจารณาค่าเฉลี่ยของสหสัมพันธ์ระหว่างความสามารถจริง และความสามารถที่ถูกประมาณค่า จะพบว่ามีพิสัยอยู่ในช่วง 0.951 ถึง 0.957 เมื่อพิจารณาค่าเฉลี่ยของ Bias และ MSE จะพบว่ามีพิสัยอยู่ในช่วง -0.004 ถึง 0.009 และ 0.83 ถึง 0.97 ตามลำดับ จะเห็นได้ว่าวิธีการคัดเลือกแบบทดสอบย่อยที่มีค่า MSE ต่ำที่สุด คือ MCC ที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดที่ระดับ 20 เปอร์เซ็นต์ (MSE=0.83) และวิธีการคัดเลือกแบบทดสอบย่อยที่มีค่า MSE สูงที่สุด คือ วิธีการคัดเลือกแบบทดสอบย่อยที่ใช้วิธีการสุ่ม (MSE=0.97) รายละเอียดดังตารางที่ 4.1

ตารางที่ 4.1 ค่าเฉลี่ยดัชนีที่ใช้พิจารณาประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถผู้สอบของวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ ที่ใช้กับคลังข้อสอบขนาด 600 ข้อ

วิธีการ	r_{max}	ค่าเฉลี่ยของการทำซ้ำ 10 ครั้ง					
		ความยาวของแบบสอบ (ข้อ)	ความยาวของแบบสอบ (ชุด)	$\overline{SEE}$	$r_{\theta\hat{\theta}}$	Bias	MSE
TFI	-	29.19	7.30	0.289	0.957	0.002	0.084
RAN	-	57.34	14.33	0.313	0.951	-0.004	0.097
CWA	0.10	53.45	13.36	0.299	0.952	0.000	0.091
	0.15	54.03	13.51	0.299	0.955	0.009	0.090
	0.20	53.08	13.27	0.298	0.955	-0.002	0.088
	0.25	51.72	12.93	0.298	0.956	0.001	0.086
MCC	0.10	45.60	11.40	0.294	0.953	0.002	0.091
	0.15	41.51	10.38	0.292	0.956	0.003	0.085
	0.20	38.69	9.67	0.291	0.956	-0.001	0.083
	0.25	37.09	9.27	0.291	0.957	0.005	0.084

2) การทดลองกับคลังข้อสอบที่มีข้อสอบจำนวน 800 ข้อ ภายในคลังข้อสอบถูกจัดเป็นแบบทดสอบย่อยได้ 200 ชุด โดยแบบทดสอบย่อยแต่ละชุดมีข้อสอบไม่ซ้ำกัน ผลการทดลองมีค่าเฉลี่ยตัวชี้วัดที่ใช้พิจารณาประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถผู้สอบดังนี้ วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบ CWA พบว่าเมื่อเพิ่มขนาดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดจาก 10 เปอร์เซ็นต์ เป็น 15, 20 และ 25 เปอร์เซ็นต์ ตามลำดับแล้ว ผล

ปรากฏว่า ความยาวของแบบสอบเฉลี่ยมีแนวโน้มลดลง คือ จากเดิมที่ใช้ข้อสอบประมาณ 56 ข้อ ลดลงมาเป็นใช้ข้อสอบประมาณ 53 ข้อ แต่เมื่อพิจารณาจำนวนแบบทดสอบย่อยเฉลี่ยที่ใช้พบว่าจะมีค่าใกล้เคียงกันคือ ใช้ประมาณ 13-14 ชุด ขณะที่วิธี MCC ความยาวของแบบสอบเฉลี่ยที่ใช้ และจำนวนแบบทดสอบย่อยเฉลี่ยที่ใช้มีแนวโน้มลงลดอย่างเห็นได้ชัด คือ ลดลงจาก 41.45 เหลือ 32.80 และ 10.36 เหลือ 8.32 ตามลำดับ เมื่อพิจารณาค่าเฉลี่ยของสหสัมพันธ์ระหว่างความสามารถจริงและความสามารถที่ถูกประมาณค่า จะพบว่ามีพิสัยอยู่ในช่วง 0.950 ถึง 0.959 เมื่อพิจารณาค่าเฉลี่ยของ Bias และ MSE จะพบว่ามีพิสัยอยู่ในช่วง -0.004 ถึง 0.005 และ 0.81 ถึง 0.98 ตามลำดับ จะเห็นได้ว่าวิธีการคัดเลือกแบบทดสอบย่อยที่มีค่า MSE ต่ำที่สุดคือ วิธีการคัดเลือกแบบทดสอบย่อยที่ใช้ค่าสารสนเทศของฟิชเชอร์สูงสุด (MSE=0.81) และวิธีการคัดเลือกแบบทดสอบย่อยที่มีค่า MSE สูงที่สุดคือ วิธีการคัดเลือกแบบทดสอบย่อยที่ใช้วิธีการสุ่ม (MSE=0.98) รายละเอียดดังตารางที่ 4.2

ตารางที่ 4.2 ค่าเฉลี่ยตัวชี้วัดที่ใช้พิจารณาประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถผู้สอบของวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ ที่ใช้กับคลังข้อสอบขนาด 800 ข้อ

วิธีการ	r_{max}	ค่าเฉลี่ยของการทำซ้ำ 10 ครั้ง					
		ความยาวของแบบสอบ (ข้อ)	ความยาวของแบบสอบ (ชุด)	$\overline{SEE}$	$r_{\theta\hat{\theta}}$	Bias	MSE
TFI	-	27.17	6.79	0.283	0.959	-0.002	0.081
RAN	-	56.78	14.61	0.312	0.950	-0.003	0.098
CWA	10	55.87	13.97	0.299	0.953	0.002	0.093
	15	55.27	13.82	0.300	0.954	0.005	0.090
	20	54.60	13.65	0.299	0.955	-0.001	0.090
	25	53.47	13.37	0.298	0.954	-0.001	0.090
MCC	10	41.45	10.36	0.291	0.959	0.001	0.082
	15	37.51	9.38	0.290	0.957	-0.004	0.086
	20	34.51	8.63	0.289	0.958	-0.001	0.083
	25	32.80	8.20	0.288	0.958	0.000	0.083

1.2 สถิติพื้นฐานของประสิทธิภาพด้านการใช้คลังข้อสอบ

การวิเคราะห์ในส่วนนี้ ผู้วิจัยนำเสนอผลการวิเคราะห์ประสิทธิภาพด้านการใช้คลังข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ด้วยวิธีการเลือกแบบทดสอบย่อยที่ให้ค่าสารสนเทศสูงสุด

วิธีการสุ่ม วิธีแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (CWA) และวิธีมอนติ คาร์โล ซีเอที (MCC) โดยจำแนกตาม ขนาดคลังข้อสอบ และอัตราการควบคุมการใช้แบบทดสอบย่อยซ้ำ

1) การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แต่ละวิธีที่ใช้กับคลังข้อสอบขนาด 600 ข้อ พบว่า วิธีการที่มีค่าเฉลี่ยอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ต่ำที่สุด 3 วิธีแรก คือ วิธีการคัดเลือกแบบทดสอบย่อยแบบสุ่ม (RAN) และ วิธีมอนติ คาร์โล ซีเอที (MCC) ที่มีการควบคุมอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดไว้ที่ 10 และ 15 เปอร์เซ็นต์ ค่าเฉลี่ยอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้มีดังนี้ คือ 0.12 0.128 และ 0.132 ตามลำดับ เมื่อพิจารณาที่ค่าเฉลี่ยอัตราการทับซ้อนกันของแบบสอบที่สังเกตได้ และค่าไคสแควร์ พบว่า ทั้งสามวิธีที่กล่าวมาข้างต้น มีค่าต่ำที่สุด 3 อันดับแรก

2) การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แต่ละวิธีที่ใช้กับคลังข้อสอบขนาด 800 ข้อ พบว่า วิธีการที่มีค่าเฉลี่ยอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ต่ำที่สุด 3 วิธีแรก คือ RAN และ วิธี MCC ที่มีการควบคุมอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดไว้ที่ 10 และ 15 เปอร์เซ็นต์ เหมือนกับในการทดสอบด้วยคลังข้อสอบ ขนาด 600 ข้อ ค่าเฉลี่ยอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้มีดังนี้ คือ 0.12 0.128 และ 0.132 ตามลำดับ เมื่อพิจารณาที่ค่าเฉลี่ยอัตราการทับซ้อนกันของแบบสอบที่สังเกตได้ และค่าไคสแควร์ พบว่า ทั้งสามวิธีที่กล่าวมาข้างต้น มีค่าต่ำที่สุด 3 อันดับแรก ผลการวิเคราะห์ ดังตารางที่ 4.3

ตารางที่ 4.3 ค่าเฉลี่ยของเกณฑ์ที่ใช้พิจารณาประสิทธิภาพด้านการใช้คลังข้อสอบของวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ จำแนกตามขนาดคลังข้อสอบ

วิธีการ	r_{max}	ค่าเฉลี่ยของการทำซ้ำ 10 ครั้ง									
		คลังข้อสอบ 600					คลังข้อสอบ 800				
		r_{obs}	$\bar{t}$	x^2	N_{never}	N_{over}	r_{obs}	$\bar{t}$	x^2	N_{never}	N_{over}
TFI	-	1.000	0.204	30.983	96.30	17.5	1.000	0.182	37.864	138.4	23
RAN	-	0.120	0.091	0.158	0.00	0.00	0.087	0.067	0.131	0.00	0.00
CWA	10	0.344	0.112	5.213	6.20	24.80	0.408	0.109	9.051	34.70	22.00
	15	0.206	0.107	4.208	9.10	6.60	0.254	0.102	7.825	41.50	6.60
	20	0.177	0.109	4.895	12.70	0.00	0.184	0.104	8.656	48.10	0.00
	25	0.196	0.110	5.700	17.00	0.00	0.195	0.107	9.718	52.40	0.00
MCC	10	0.128	0.061	1.519	0.00	0.30	0.09	0.041	2.627	4.10	0.30
	15	0.132	0.060	3.310	7.60	0.00	0.126	0.045	5.368	27.40	0.00
	20	0.165	0.063	5.231	19.40	0.00	0.163	0.051	7.942	52.10	0.00
	25	0.200	0.069	6.814	30.30	0.00	0.191	0.056	9.986	71.80	0.00

ตอนที่ 2 ผลการวิเคราะห์ประสิทธิภาพของวิธีการคัดเลือกข้อสอบในการทดสอบแบบประเหมาะด้วยคอมพิวเตอร์สำหรับโมเดลการตอบสนองแบบทดสอบย่อย

การวิเคราะห์ตอนนี้ ผู้วิจัยทำการทดสอบประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถของผู้สอบ และประสิทธิภาพด้านการใช้คลังข้อสอบในการทดสอบแบบประเหมาะด้วยคอมพิวเตอร์สำหรับโมเดลการตอบสนองแบบทดสอบย่อย ผลการวิเคราะห์จะจำแนกประสิทธิภาพแต่ละด้านและตามวิธีการคัดเลือกข้อสอบ โดยมีรายละเอียดดังนี้

2.1 ประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถของผู้สอบ

2.1.1 วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (CWA)

จากการวิเคราะห์ข้อมูลประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถของผู้สอบของวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (CWA) พบว่า เมื่อใช้การควบคุมอัตราการใช้แบบทดสอบย่อยซ้ำเพิ่มขึ้นครั้งละ 5 % จาก 10% เป็น 15%, 20% และ 25% ตามลำดับ ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าความสามารถที่แท้จริงกับความสามารถที่ประมาณค่า ($r_{\theta\hat{\theta}}$) มีเปลี่ยนแปลงน้อยมากคือเปลี่ยนแปลงไม่เกิน 0.004 ส่วนค่าเฉลี่ยของความยาวของแบบสอบจะลดลงตามลำดับเมื่อเพิ่มอัตราการใช้แบบทดสอบย่อยซ้ำขึ้นครั้งละ 5 % คือ โดยเฉลี่ยแล้วผู้สอบแต่ละคนจะใช้แบบทดสอบย่อย (M_{tls}) จำนวน 13.63, 13.51, 13.27 และ 12.93 ชุดตามลำดับ นอกจากนี้เมื่อเพิ่มขนาดคลังข้อสอบจาก 600 ข้อ เป็น 800 ข้อ พบว่า ค่าเฉลี่ยของความยาวของแบบสอบก็เพิ่มขึ้นเช่นกัน โดยความค่าเฉลี่ยความยาวของแบบสอบที่เพิ่มขึ้นเมื่อขนาดคลังข้อสอบใหญ่ขึ้นมีค่าประมาณ 0.37 ชุด นอกจากนั้นเมื่อพิจารณาจากค่า ค่าเฉลี่ยของความลำเอียง (M_{Bias}) และค่าเฉลี่ยของความแปรปรวนของคลาดเคลื่อน (M_{MSE}) พบว่ามีอัตราการเพิ่มขึ้นหรือลดลงน้อยมากซึ่งสอดคล้องกับการเปลี่ยนแปลงของค่า $r_{\theta\hat{\theta}}$ นอกจากนี้ยังพบว่า ค่าเฉลี่ยของความคลาดเคลื่อนมาตรฐานของการประมาณค่า (SEE) ของทุกเงื่อนไขการทดสอบมีค่าประมาณ 0.290 ถึง 0.30 ซึ่งถือว่าใกล้เคียงกันมาก รายละเอียดดังตารางที่ 4.4

ตารางที่ 4.4 ค่าเฉลี่ยของ ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าความสามารถที่แท้จริงกับความสามารถที่ประมาณค่า ความยาวของแบบสอบ ความคลาดเคลื่อนมาตรฐานของการประมาณค่าความลำเอียงเฉลี่ย และความแปรปรวนของความคลาดเคลื่อน ของวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (CWA)

ผลการวิเคราะห์	ขนาดคลังข้อสอบ	อัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max})			
		10%	15%	20%	25%
M_{Bias}	600	0.000	0.009	-0.002	0.001
	800	0.002	0.005	-0.001	-0.001
M_{MSE}	600	0.091	0.090	0.088	0.086
	800	0.093	0.090	0.090	0.090
$r_{\theta\hat{\theta}}$	600	0.952	0.955	0.955	0.956
	800	0.953	0.954	0.955	0.954
M_{it}	600	53.45	54.03	53.08	51.72
	800	55.87	55.27	54.60	53.47
M_{tls}	600	13.63	13.51	13.27	12.93
	800	13.97	13.82	13.65	13.37
M_{SEE}	600	0.299	0.299	0.298	0.298
	800	0.299	0.300	0.299	0.298

2.1.2 วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที (MCC)

จากการวิเคราะห์ข้อมูลประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถของผู้สอบของวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที (MCC) พบว่า เมื่อใช้การควบคุมอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดเพิ่มขึ้นครั้งละ 5 % จาก 10% เป็น 15%, 20% และ 25% ตามลำดับ ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าความสามารถที่แท้จริงกับความสามารถที่ประมาณค่า ($r_{\theta\hat{\theta}}$) มีเปลี่ยนแปลงน้อยมากคือเปลี่ยนแปลงไม่เกิน 0.0025 ส่วนค่าเฉลี่ยของความยาวของแบบสอบจะลดลงตามลำดับเมื่อเพิ่มอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดขึ้นครั้งละ 5 % คือ โดยเฉลี่ยแล้วผู้สอบแต่ละคนจะใช้แบบทดสอบย่อย (M_{tls}) จำนวน 11.40, 10.38, 9.67 และ 9.27 ชุดตามลำดับ นอกจากนี้เมื่อเพิ่มขนาดคลังข้อสอบจาก 600 ข้อ เป็น 800 ข้อ พบว่าค่าเฉลี่ยของความยาวของแบบสอบมีค่าลดลง โดยความค่าเฉลี่ยความยาวของแบบสอบที่ลดลงเมื่อ

ขนาดคลังข้อสอบใหญ่ขึ้นมีค่าประมาณ 1.04 ชุด นอกจากนั้นเมื่อพิจารณาจากค่า ค่าเฉลี่ยของความลำเอียง (M_{Bias}) และค่าเฉลี่ยของความแปรปรวนของความคลาดเคลื่อน (M_{MSE}) พบว่ามีอัตราการเพิ่มขึ้นหรือลดลงน้อยมากซึ่งสอดคล้องกับการเปลี่ยนแปลงของค่า $r_{\theta\hat{\theta}}$ นอกจากนี้ยังพบว่า ค่าเฉลี่ยของความคลาดเคลื่อนมาตรฐานของการประมาณค่า (SEE) มีค่าน้อยกว่า 0.3 ในทุกเงื่อนไขการทดสอบ เนื่องจาก โดยกำหนดเกณฑ์การยุติการทดสอบเมื่อมี SEE <= 0.3 รายละเอียดดัง ดังตารางที่ 4.5

ตารางที่ 4.5 ค่าเฉลี่ยของ ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าความสามารถที่แท้จริงกับความสามารถที่ประมาณค่า ความยาวของแบบสอบ ความคลาดเคลื่อนมาตรฐานของการประมาณค่าความลำเอียงเฉลี่ย และความแปรปรวนของความคลาดเคลื่อน ของวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที (MCC)

ผลการวิเคราะห์	ขนาดคลังข้อสอบ	อัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max})			
		10%	15%	20%	25%
M_{Bias}	600	0.002	0.003	-0.001	0.005
	800	0.001	-0.004	-0.001	0.000
M_{MSE}	600	0.091	0.085	0.083	0.084
	800	0.082	0.086	0.083	0.083
$r_{\theta\hat{\theta}}$	600	0.953	0.956	0.956	0.957
	800	0.959	0.957	0.958	0.958
M_{it}	600	45.60	41.51	38.69	37.09
	800	41.45	37.51	34.51	32.80
M_{tl}	600	11.40	10.38	9.67	9.27
	800	10.36	9.38	8.63	8.20
M_{SEE}	600	0.294	0.292	0.291	0.291
	800	0.291	0.290	0.289	0.288

2.2 ประสิทธิภาพด้านการใช้คลังข้อสอบ

2.2.1 วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (CWA)

จากการวิเคราะห์ข้อมูลประสิทธิภาพด้านการใช้คลังข้อสอบของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่มีการคัดเลือกข้อสอบแบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (CWA) พบว่า เมื่อผ่อนคลายระดับการควบคุมอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) จากระดับ 10 เปอร์เซ็นต์ เป็น 15, 20 และ 25 เปอร์เซ็นต์ ค่าเฉลี่ยของอัตราการทับซ้อนกันของแบบสอบ ($\bar{t}$) มีค่าใกล้เคียงกันโดยมีพิสัยอยู่ระหว่าง .107 - .112 สำหรับคลังข้อสอบขนาด 600 ข้อ และมีพิสัยอยู่ระหว่าง .102 - .109 สำหรับคลังข้อสอบขนาด 800 ข้อ นอกจากนี้ยังพบว่าเมื่อเพิ่มขนาดคลังข้อสอบ จาก 600 ข้อ เป็น 800 ข้อ ค่าเฉลี่ยของอัตราการทับซ้อนกันของแบบสอบจะลดลงในทุกระดับเงื่อนไขของ $r_{\max}$

เมื่อพิจารณาค่าเฉลี่ยของไคสแควร์ (x^2) จะทำให้ทราบความสมดุลของการใช้คลังข้อสอบในภาพรวม และเมื่อผ่อนคลายระดับการควบคุมอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) จากระดับ 10 เปอร์เซ็นต์ เป็น 15, 20 และ 25 เปอร์เซ็นต์ พบว่า ค่าเฉลี่ยของไคสแควร์มีการแกว่งตัวลดลงและเพิ่มขึ้นเล็กน้อย โดยมีพิสัยอยู่ระหว่าง 4.208 ถึง 5.70 สำหรับคลังข้อสอบขนาด 600 ข้อ และมีพิสัยอยู่ระหว่าง 7.825 ถึง 9.781 สำหรับคลังข้อสอบขนาด 800 ข้อ และเมื่อเพิ่มขนาดคลังข้อสอบให้ใหญ่ขึ้นจาก 600 ข้อ เป็น 800 ข้อ ค่าเฉลี่ยของไคสแควร์มีการเพิ่มขึ้นในทุกระดับของ $r_{\max}$ แต่เป็นการเพิ่มขึ้นเพียงเล็กน้อย คือ มีพิสัยอยู่ระหว่าง 3.62 – 4.02 การเพิ่มขึ้นของค่าเฉลี่ยของไคสแควร์เพียงเล็กน้อยนี้สะท้อนให้เห็นว่า แบบทดสอบย่อยมีความถี่ในการถูกนำไปใช้ใกล้เคียงกับคลังข้อสอบขนาด 600 ข้อ และค่าเฉลี่ยของไคสแควร์ที่มีค่าต่ำจึงเป็นหลักฐานยืนยันว่าวิธีการคัดเลือกข้อสอบแบบ CWA สามารถสร้างความสมดุลของการใช้คลังข้อสอบได้ดี เพราะช่วยลดความถี่ในการเลือกใช้แบบทดสอบย่อยที่มีค่าอำนาจจำแนกสูงๆ และกระจายโอกาสในการเลือกแบบทดสอบย่อยไปยังแบบทดสอบย่อยที่มีค่าอำนาจจำแนกต่ำกว่า

เมื่อพิจารณาที่ค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ (r_{obs}) ทำให้ทราบว่าในแต่ละกลุ่มการทดลอง แบบทดสอบย่อยแต่ละชุดมีอัตราการนำไปใช้เฉลี่ยสูงสุดเท่าไหร่ จากผลการทดสอบพบว่า เมื่อควบคุมอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) ไว้ที่ 10 และ 15 เปอร์เซ็นต์ แบบทดสอบย่อยบางชุดถูกเลือกมาใช้มีค่าความถี่สูงสุดมากกว่าอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) ที่กำหนดขึ้น ดังนั้นจึงฝ่าฝืนข้อกำหนดที่ตั้งขึ้น แต่เมื่อปรับระดับอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) ไว้ที่ 20 และ 25 เปอร์เซ็นต์ ค่า r_{obs} ที่ได้มีค่าไม่เกิน 0.20 และ 0.25 ซึ่งหมายความ

ว่าแบบทดสอบย่อยแต่ละชุดในคลังข้อสอบถูกนำมาใช้โดยมีความถี่ไม่เกิน 20 และ 25 เปอร์เซ็นต์ ปรากฏการณ์ดังกล่าวเกิดขึ้นแบบเดียวกันทั้งกับคลังข้อสอบขนาด 600 ข้อ และ 800 ข้อ

เมื่อพิจารณาค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ ($N_{neverexposed}$) พบว่า ทั้งกับคลังข้อสอบขนาด 600 ข้อ และ 800 ข้อ $N_{neverexposed}$ มีแนวโน้มเพิ่มขึ้นเมื่อระดับการควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max}) เพิ่มขึ้น การเพิ่มขึ้นของ $N_{neverexposed}$ มีสาเหตุมาจากแบบทดสอบย่อยแต่ละชุดในคลังข้อสอบมีโอกาสถูกนำออกไปใช้ซ้ำได้เพิ่มมากขึ้นนั่นเอง โดยสังเกตได้จากความถี่ของการนำแบบทดสอบย่อยฉบับนั้นมาใช้ ดังแสดงในตารางที่ 4.8 และ 4.9

เมื่อพิจารณาค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ถูกนำมาใช้มากกว่าอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดที่กำหนดไว้ ($N_{overexposed}$) พบว่า เมื่อใช้วิธี CWA กับคลังข้อสอบขนาด 600 ข้อ และ 800 ข้อ โดยควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max}) ไว้ที่ระดับ 10 เปอร์เซ็นต์ ผลของ $N_{overexposed}$ มีจำนวนเฉลี่ยสูงถึง 24.8 ชุด และ 22 ชุด ตามลำดับ ดังนั้นจึงเป็นหลักฐานยืนยันว่าวิธีการคัดเลือกข้อสอบแบบ CWA ไม่เหมาะสมกับการกำหนด r_{max} ต่ำกว่า 15 เปอร์เซ็นต์ รายละเอียดดังแสดงในตารางที่ 4.6

ตารางที่ 4.6 ประสิทธิภาพด้านการใช้คลังข้อสอบของวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (CWA)

ผลการวิเคราะห์	ขนาดคลังข้อสอบ	อัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max})			
		10	15	20	25
$\bar{t}$	600	0.112	0.107	0.109	0.110
	800	0.109	0.102	0.104	0.107
x^2	600	5.213	4.208	4.895	5.700
	800	9.051	7.825	8.656	9.718
r_{obs}	600	0.344	0.206	0.177	0.196
	800	0.408	0.254	0.184	0.195
$N_{neverexposed}$	600	6.20	9.10	12.70	17.00
	800	34.70	41.50	48.10	52.40
$N_{overexposed}$	600	24.80	6.60	0.00	0.00
	800	22.00	6.60	0.00	0.00

2.2.2 วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที (MCC)

จากการวิเคราะห์ข้อมูลประสิทธิภาพด้านการใช้คลังข้อสอบของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่มีการคัดเลือกข้อสอบแบบมอนติ คาร์โล ซีเอที (MCC) พบว่า เมื่อผ่อนคลายระดับการควบคุมอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) จากระดับ 10 เปอร์เซ็นต์ เป็น 15, 20 และ 25 เปอร์เซ็นต์ ค่าเฉลี่ยของอัตราการทับซ้อนกันของแบบสอบ ($\bar{t}$) มีค่าใกล้เคียงกันโดยมีพิสัยอยู่ระหว่าง .060 - .069 สำหรับคลังข้อสอบขนาด 600 ข้อ และมีพิสัยอยู่ระหว่าง .041 - .056 สำหรับคลังข้อสอบขนาด 800 ข้อ นอกจากนี้ยังพบว่าเมื่อเพิ่มขนาดคลังข้อสอบ จาก 600 ข้อ เป็น 800 ข้อ ค่าเฉลี่ยของอัตราการทับซ้อนกันของแบบสอบจะลดลงในทุกระดับเงื่อนไขของ $r_{\max}$

เมื่อพิจารณาค่าเฉลี่ยของไคสแควร์ (x^2) ทำให้ทราบความสมดุลของการใช้คลังข้อสอบในภาพรวม เมื่อผ่อนคลายระดับการควบคุมอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) จากระดับ 10 เปอร์เซ็นต์ เป็น 15, 20 และ 25 เปอร์เซ็นต์ พบว่า ค่าเฉลี่ยของไคสแควร์มีการเพิ่มขึ้นเล็กน้อย โดยมีพิสัยอยู่ระหว่าง 1.519 ถึง 6.814 สำหรับคลังข้อสอบขนาด 600 ข้อ และมีพิสัยอยู่ระหว่าง 2.627 ถึง 9.986 สำหรับคลังข้อสอบขนาด 800 ข้อ และเมื่อเพิ่มขนาดคลังข้อสอบให้ใหญ่ขึ้นจาก 600 ข้อ เป็น 800 ข้อ ค่าเฉลี่ยของไคสแควร์มีการเพิ่มขึ้นในทุกระดับของ $r_{\max}$ แต่เป็นการเพิ่มขึ้นเพียงเล็กน้อย คือ มีพิสัยอยู่ระหว่าง 1.11 – 3.17 การเพิ่มขึ้นของค่าเฉลี่ยของไคสแควร์เพียงเล็กน้อยนี้สะท้อนให้เห็นว่าแบบทดสอบย่อยมีความถี่ในการถูกนำไปใช้ใกล้เคียงกับคลังข้อสอบขนาด 600 ข้อ และค่าเฉลี่ยของไคสแควร์ที่มีค่าค่อนข้างต่ำจึงเป็นหลักฐานยืนยันว่าวิธีการคัดเลือกข้อสอบแบบ MCC สามารถสร้างความสมดุลของการใช้คลังข้อสอบได้ดี เพราะช่วยกระจายโอกาสในการเลือกแบบทดสอบย่อยได้อย่างทั่วถึงตลอดทั้งคลังข้อสอบ

เมื่อพิจารณาที่ค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ (r_{obs}) ทำให้ทราบว่าในแต่ละกลุ่มการทดลอง แบบทดสอบย่อยแต่ละชุดมีอัตราการนำไปใช้เฉลี่ยสูงสุดเท่าไหร่ จากผลการทดสอบพบว่า เมื่อคลังข้อสอบมีขนาด 600 ข้อ และควบคุมระดับการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) ไว้ที่ 10 เปอร์เซ็นต์ แบบทดสอบย่อยบางชุดถูกเลือกนำมาใช้โดยมีความถี่มากกว่า $r_{\max}$ ที่กำหนดขึ้น ดังนั้นจึงฝ่าฝืนข้อกำหนดที่ตั้งขึ้น แต่เมื่อพิจารณาในเงื่อนไขทดลองอื่นๆ พบว่า r_{obs} มีค่าน้อยกว่า $r_{\max}$ ในทุกเงื่อนไข

เมื่อพิจารณาค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ ($N_{neverexposed}$) พบว่า ทั้งกับคลังข้อสอบขนาด 600 ข้อ และ 800 ข้อ $N_{neverexposed}$ มีแนวโน้มเพิ่มขึ้นเมื่อระดับการควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) เพิ่มขึ้น การเพิ่มขึ้น $N_{neverexposed}$ มีสาเหตุมาจากแบบทดสอบย่อยแต่ละชุดในคลังข้อสอบมีโอกาสถูกนำออกไปใช้ซ้ำได้เพิ่มมากขึ้นนั่นเอง ดังแสดงในตารางที่ 4.13

เมื่อพิจารณาค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ถูกนำมาใช้มากกว่ามากเกินไป ($N_{overexposed}$) พบว่า เมื่อวิธี MCC ที่ใช้คลังข้อสอบขนาด 600 ข้อ และควบคุมอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max}) ไว้ที่ระดับ 10 เปอร์เซ็นต์ ผลของ $N_{overexposed}$ มีค่าเฉลี่ยอยู่ที่ประมาณ 0.3 ชุด ซึ่งถือว่าค่อนข้างน้อยมากเมื่อเปรียบเทียบกับวิธี CWA ที่มีเงื่อนไขการทดลองแบบเดียวกัน รายละเอียดดังแสดงในตารางที่ 4.7, 4.8 และ 4.9

ตารางที่ 4.7 ประสิทธิภาพด้านการใช้คลังข้อสอบของวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที (MCC)

ผลการวิเคราะห์	ขนาดคลังข้อสอบ	อัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max})			
		0.10	0.15	0.20	0.25
$\bar{t}$	600	0.061	0.060	0.063	0.069
	800	0.041	0.045	0.051	0.056
x^2	600	1.519	3.31	5.231	6.814
	800	2.627	5.368	7.942	9.986
r_{obs}	600	0.128	0.132	0.165	0.200
	800	0.090	0.126	0.163	0.191
$N_{neverexposed}$	600	0.00	7.60	19.40	30.30
	800	4.10	27.40	52.10	71.80
$N_{overexposed}$	600	0.30	0.00	0.00	0.00
	800	0.00	0.00	0.00	0.00

ตารางที่ 4.8 ความถี่เฉลี่ยของอัตราการแสดงแบบทดสอบย่อยสำหรับคลังข้อสอบขนาด 600 ข้อ จากการทำซ้ำ 10 รอบ

Exposure Rate	Exposure Control Condition									
(r_{obs})	TMF	RAN	CWA10	CWA15	CWA20	CWA25	MCC10	MCC15	MCC20	MCC25
71-1.0	0.8	0	0	0	0	0	0	0	0	0
61-.70	0.3	0	0	0	0	0	0	0	0	0
51-.60	0.9	0	0	0	0	0	0	0	0	0
41-.50	2.5	0	0	0	0	0	0	0	0	0
36-.40	1.3	0	0.2	0	0	0	0	0	0	0
31-.35	1.6	0	1.3	0	0	0	0	0	0	0
26-.30	1.8	0	2.4	0	0	0	0	0	0	0
21-.25	2.2	0	4.2	0.5	0	0	0.1	0	0	0
16-.20	2.7	0	4.9	6.1	17.8	22.3	0.1	0	1.4	9.0
11-.15	3.4	9.6	11.8	50.1	44.8	36.1	0.1	17.6	30.0	25.8
06-.10	5.4	140.3	86.6	48.7	35.2	36.1	121.9	76.5	47.6	35.8
01-.05	30.8	0.1	32.4	35.5	39.5	38.5	27.8	48.3	51.6	49.1
.00	96.3	0.0	6.2	9.1	12.7	17.0	0.0	7.6	19.4	30.3
(Not admin)	(64.20%)	(0.00%)	(4.13%)	(6.07%)	(8.47%)	(11.33%)	(0.00%)	(5.07%)	(12.93%)	(20.20%)

ตารางที่ 4.9 ความถี่เฉลี่ยของอัตราการแสดงแบบทดสอบย่อยสำหรับคลังข้อสอบขนาด 800 ข้อ จากการทำซ้ำ 10 รอบ

Exposure Rate	Exposure Control Condition									
(r_{obs})	TMF	RAN	CWA10	CWA15	CWA20	CWA25	MCC10	MCC15	MCC20	MCC25
71-1.0	5.0	0	0	0	0	0	0	0	0	0
61-.70	1.4	0	0	0	0	0	0	0	0	0
51-.60	1.7	0	0	0	0	0	0	0	0	0
41-.50	1.9	0	0.7	0	0	0	0	0	0	0
36-.40	1.6	0	0.5	0	0	0	0	0	0	0
31-.35	1.1	0	0.8	0	0	0	0	0	0	0
26-.30	1.5	0	0.8	0.6	0	0	0	0	0	0
21-.25	2.6	0	1.1	1.9	0	0	0	0	0	0
16-.20	2.7	13.6	6.0	4.1	12.5	14.0	0	0	1.5	7.8
11-.15	3.6	185.9	12.1	43.3	42.3	44.3	0	9.6	18.9	18.8
06-.10	5.1	0.5	96.9	66.5	54.9	43.3	78.9	63.5	45.2	31.7
01-.05	33.4	0	46.4	42.1	42.2	46.0	117.0	99.5	82.3	69.9
.00	138.4	0	34.7	41.5	48.1	52.4	4.1	27.4	52.1	71.8
(Not admin)	(92.27%)	(0.00%)	(23.13%)	(27.67%)	(32.07%)	(34.93%)	(2.73%)	(18.27%)	(34.73%)	(47.87%)

ตอนที่ 3 ผลการเปรียบเทียบประสิทธิภาพของวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์

ผลการทดสอบความแตกต่างของประสิทธิภาพด้านความถูกต้องแม่นยำในการวัดและประสิทธิภาพด้านการใช้คลังข้อสอบของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โลซีเอที (MCC) กับแบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (CWA) เมื่อกำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max}) เป็น 4 ระดับ คือ 10%, 15%, 20% และ 25% และใช้กับคลังข้อสอบ 2 ขนาด คือ คลังข้อสอบขนาด 600 ข้อ และ 800 ข้อ ดังนั้นในการทดลองครั้งนี้จึงมี 16 เงื่อนไขการทดลอง (2 วิธีการคัดเลือกแบบทดสอบย่อย x 2 ขนาดคลังข้อสอบ x 4 อัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด) โดยผู้วิจัยตัดวิธีการคัดเลือกแบบทดสอบย่อยแบบใช้ค่าสารสนเทศสูงสุดและแบบสุ่ม ออกจากการทดสอบเปรียบเทียบความแตกต่าง เนื่องจากวิธีการทั้ง 2 วิธี (4 เงื่อนไข) นำมาใช้เพื่อแสดงให้เห็นค่าสุดโต่งของวิธีการที่ให้ประสิทธิภาพด้านความถูกต้องแม่นยำในการวัดและประสิทธิภาพด้านการใช้คลังข้อสอบ เมื่อใช้กับคลังข้อสอบแต่ละขนาดโดยไม่มีการกำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด

ผู้วิจัยทำการทดสอบเปรียบเทียบความแตกต่างของค่าสถิติที่ใช้เป็นดัชนีในการพิจารณาประสิทธิภาพด้านความถูกต้องแม่นยำในการวัดและประสิทธิภาพด้านการใช้คลังข้อสอบโดยจำแนกตามขนาดของคลังข้อสอบ โดยใช้การทดสอบของ Kruskal-Wallis เพื่อตอบคำถามวิจัยที่ว่า

1. วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบใด และในเงื่อนไขใด มีประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถของผู้สอบสูงกว่ากัน โดยพิจารณาจากเกณฑ์ ดังนี้ 1) ค่าความลำเอียงเฉลี่ย 2) ค่าความแปรปรวนของความคลาดเคลื่อน 3) ค่าความตรงหรือสหสัมพันธ์ระหว่างค่าความสามารถจริงกับค่าประมาณความสามารถ และ 4) ความยาวเฉลี่ยของแบบสอบ (เมื่อแต่ละเงื่อนไขของการทดลองมีค่าความคลาดเคลื่อนมากตรฐานของการประมาณค่าใกล้เคียงกัน)

2. วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบใด และในเงื่อนไขใด มีประสิทธิภาพด้านการใช้คลังข้อสอบสูงกว่ากัน โดยพิจารณาจาก 1) อัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ 2) อัตราการทับซ้อนของแบบสอบเฉลี่ย 3) ความสมดุลของการใช้คลังข้อสอบในภาพรวม 4) จำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ และ 5) จำนวนแบบทดสอบย่อยที่ถูกนำมาใช้มากเกินไป

จากผลการทดสอบความแตกต่างของเกณฑ์ที่ใช้พิจารณาประสิทธิภาพของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ด้านความถูกต้องแม่นยำในการวัดและด้านการใช้คลังข้อสอบ พบว่า ค่าเฉลี่ยของ ความยาวแบบสอบ ความคลาดเคลื่อนเฉลี่ยกำลังสอง สหสัมพันธ์ระหว่างความสามารถจริงและความสามารถประมาณค่า อัตราการทับซ้อนของแบบสอบ ไคสแควร์ อัตราการใช้

แบบทดสอบย่อยซ้ำสูงสุด จำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ และจำนวนแบบทดสอบย่อยที่ใช้มากเกินไป ของตัวอย่างทั้ง 8 กลุ่ม แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ยกเว้นค่าเฉลี่ยของ Bias เพียงอย่างเดียวที่มีค่าไม่แตกต่างกัน รายละเอียดดังตารางที่ 4.10 ดังนั้นผู้วิจัยจึงได้ทำการทดสอบค่าเฉลี่ยของอันดับเป็นรายคู่เพื่อเปรียบเทียบความแตกต่างค่าสถิติที่ใช้เป็นดัชนีในการพิจารณาประสิทธิภาพของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที (MCC) กับแบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (CWA) ที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) แตกต่างกัน ผลการเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่โดยใช้วิธี Kruskal-Wallis one-way ANOVA รายละเอียดผลการวิเคราะห์นำเสนอในหัวข้อถัดไป

ตารางที่ 4.10 ผลการวิเคราะห์ความแตกต่างของประสิทธิภาพด้านความถูกต้องแม่นยำในการวัดและประสิทธิภาพด้านการใช้คลังข้อสอบของวิธีการคัดเลือกข้อสอบและการกำหนดอัตราการควบคุมการใช้แบบทดสอบย่อยซ้ำที่แตกต่างกัน จำแนกตามขนาดคลังข้อสอบ โดยใช้ Kruskal-Wallis Test

ประสิทธิภาพด้าน	คลังข้อสอบขนาด 600 ข้อ			คลังข้อสอบขนาด 800 ข้อ		
ความถูกต้องแม่นยำในการวัด	x^2	*df*	*Asymp.Sig*	x^2	*df*	*Asymp.Sig*
M_{Bias}	9.616	7	.211	5.565	7	.591
M_{MSE}	27.007	7	.000	48.016	7	.000
$r_{\theta\hat{\theta}}$	19.369	7	.007	35.114	7	.000
M_{tl}	74.011	7	.000	75.544	7	.000
สมดุลการใช้คลังข้อสอบ						
r_{obs}	65.108	7	.000	77.244	7	.000
$\bar{t}$	68.999	7	.000	75.14	7	.000
x^2	74.651	7	.000	72.078	7	.000
$N_{neverexposed}$	73.445	7	.000	73.845	7	.000
$N_{overexposed}$	73.046	7	.000	78.518	7	.000

3.1 ผลการเปรียบเทียบประสิทธิภาพด้านความถูกต้องแม่นยำในการวัด

จากการเปรียบเทียบความแตกต่างของค่าสถิติที่ใช้เป็นดัชนีในการพิจารณาประสิทธิภาพด้านความถูกต้องแม่นยำในการวัดโดยใช้การทดสอบของ Kruskal-Wallis ทำให้ผู้วิจัยทราบว่า ดัชนีในการพิจารณาประสิทธิภาพด้านความถูกต้องแม่นยำในการวัดทุกตัวของตัวอย่างทั้ง 8 กลุ่ม แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ยกเว้น ค่าของ Bias ผู้วิจัยจึงเปรียบเทียบค่าเฉลี่ยของอันดับ

เป็นรายคู่โดยใช้วิธี Kruskal-Wallis one-way ANOVA กับค่าสถิติที่ใช้เป็นดัชนีในการพิจารณาประสิทธิภาพด้านความถูกต้องแม่นยำในการวัดทุกตัว ผลการเปรียบเทียบมีรายละเอียดดังต่อไปนี้

3.1.1 ผลการเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ยของความแปรปรวนของความคลาดเคลื่อน (MSE)

การเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ยของความแปรปรวนของความคลาดเคลื่อน (MSE) จากวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่ใช้กับคลังข้อสอบ 600 ข้อ ซึ่งมีกลุ่มตัวอย่างทั้งหมด 8 กลุ่ม จะเห็นว่า กลุ่มที่ให้ค่าเฉลี่ยของ MSE ต่ำสุดหรือมีประสิทธิภาพในการประมาณค่าความสามารถของผู้สอบสูงสุดและรองลงมาใน 3 อันดับแรก ได้แก่

อันดับที่ 1 ได้แก่ กลุ่ม MCC20 ซึ่งมีค่าเฉลี่ยของ MSE ต่ำที่สุด เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของ MSE ต่ำกว่ากลุ่ม CWA10 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และต่ำกว่ากลุ่ม CWA15 และ MCC10 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ส่วนกลุ่มอื่นๆ ที่เหลือมีค่าไม่แตกต่างกัน

อันดับที่ 2 ได้แก่ กลุ่ม MCC25 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของ MSE ไม่ต่างจากกลุ่มที่เหลือ

อันดับที่ 3 ได้แก่ กลุ่ม MCC15 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของ MSE ไม่ต่างจากกลุ่มที่เหลือ รายละเอียดดังตารางที่ 4.11

ตารางที่ 4.11 ผลการเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ย MSE ระหว่างวิธี MCC กับวิธี CWA ที่มีการควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุดแตกต่างกัน สำหรับคลังข้อสอบ 600 ข้อ

กลุ่ม	CWA10	CWA15	CWA20	CWA25	MCC10	MCC15	MCC20	MCC25
M_{MSE}	0.091	0.090	0.088	0.086	0.091	0.085	0.084	0.084
CWA10	-	2.60	14.20	20.80	3.40	26.40	37.50**	32.70*
CWA15		-	11.60	18.20	0.80	23.80	34.90*	30.10
CWA20			-	6.60	-10.80	12.20	23.30	18.50
CWA25				-	-17.40	5.60	16.70	11.90
MCC10					-	23.00	34.10*	29.30
MCC15						-	11.10	6.30
MCC20							-	-4.80
MCC25								-

* p<.05 , ** p< .01

การเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ยของความแปรปรวนของความคลาดเคลื่อน (MSE) จากวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่ใช้กับคลังข้อสอบ 800 ข้อ ซึ่งมีกลุ่มตัวอย่างทั้งหมด 8 กลุ่ม จะเห็นว่า กลุ่มที่ให้ค่าเฉลี่ยของ MSE ต่ำสุดหรือมีประสิทธิภาพในการประมาณค่าความสามารถของผู้สอบสูงสุดและรองลงมาใน 3 อันดับแรก ได้แก่

อันดับที่ 1 ได้แก่ กลุ่ม MCC10 ซึ่งมีค่าเฉลี่ยของ MSE ต่ำที่สุด เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของ MSE ต่ำกว่ากลุ่ม CWA10 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และต่ำกว่ากลุ่ม CWA25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ส่วนกลุ่มอื่นๆ ที่เหลือมีค่าไม่แตกต่างกัน

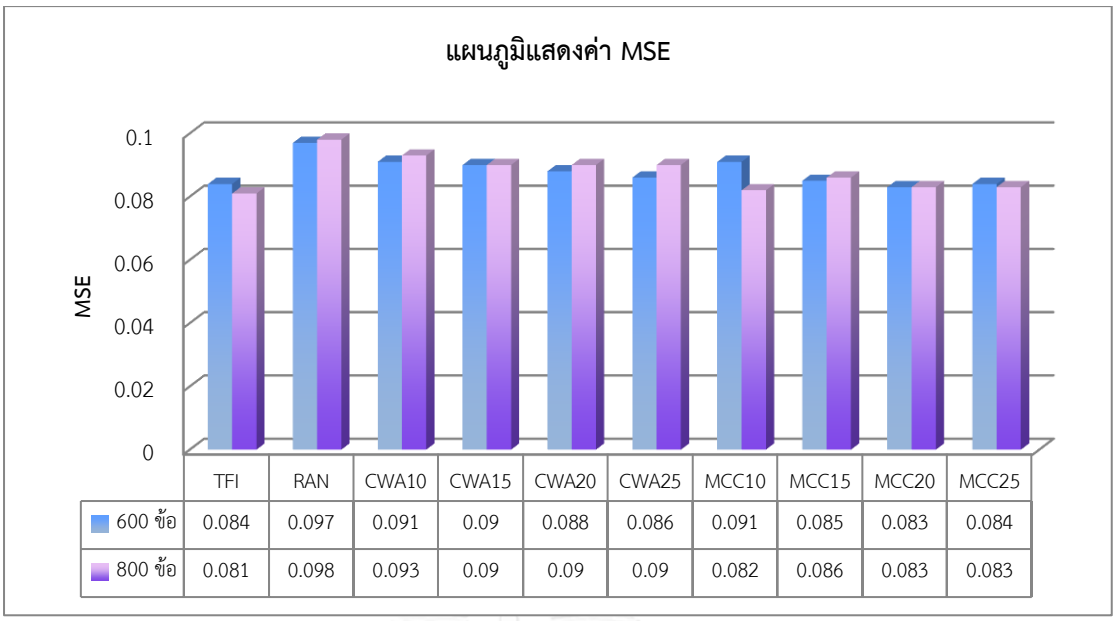
อันดับที่ 2 ได้แก่ กลุ่ม MCC20 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของ RMSE ต่ำกว่ากลุ่ม CWA10 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

อันดับที่ 3 ได้แก่ กลุ่ม MCC25 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของ MSE ต่ำกว่ากลุ่ม CWA10 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 รายละเอียดดังตารางที่ 4.12

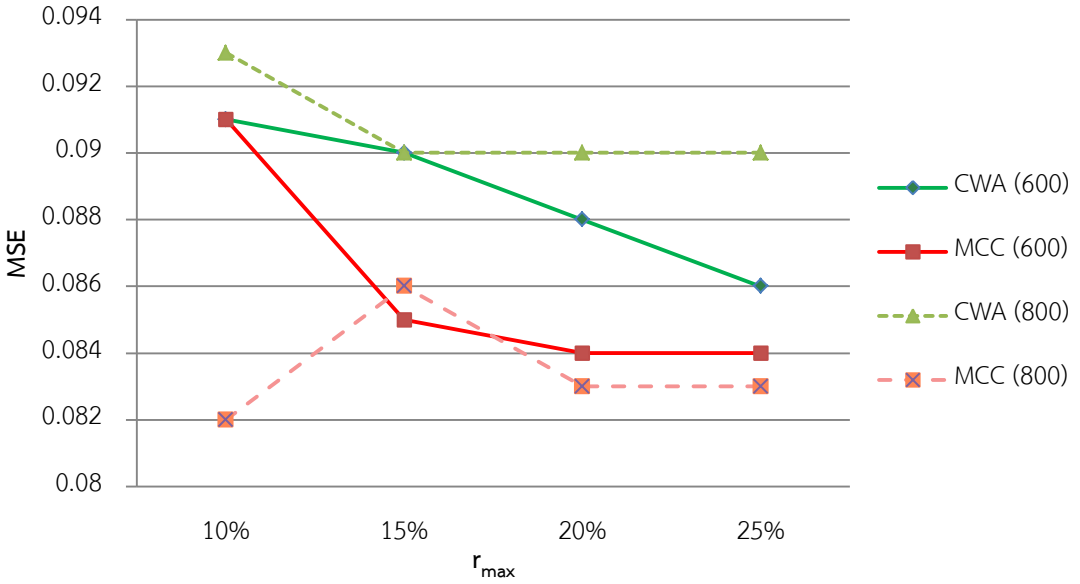
ตารางที่ 4.12 ผลการเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ย MSE ระหว่างวิธี MCC กับวิธี CWA ที่มีการควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุดแตกต่างกัน สำหรับคลังข้อสอบ 800 ข้อ

กลุ่ม	CWA10	CWA15	CWA20	CWA25	MCC10	MCC15	MCC20	MCC25
M.MSE	0.093	0.090	0.090	0.090	0.082	0.086	0.083	0.083
CWA10	-	-7.60	-15.60	-9.50	-45.90**	-30.10	-35.90*	-35.40*
CWA15		-	-8.00	-1.90	-38.30	-22.50	-28.30	-27.80
CWA20			-	6.10	-30.30	-14.50	-20.30	-19.80
CWA25				-	-36.40*	-20.60	-26.40	-25.90
MCC10					-	15.80	10.00	10.50
MCC15						-	-5.80	-5.30
MCC20							-	0.50
MCC25								-

* p<.05 , ** p< .01



ภาพที่ 4.1 กราฟค่าเฉลี่ยของ MSE ระหว่างวิธี MCC กับวิธี CWA ที่มีการควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุดแตกต่างกัน



ภาพที่ 4.2 กราฟแสดงความสัมพันธ์ระหว่างค่าเฉลี่ยของ MSE และ r_{max} ระหว่างวิธี MCC กับวิธี CWA ที่มีการควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุดแตกต่างกัน

จากภาพที่ 4.1 และ 4.2 แสดงกราฟเปรียบเทียบค่าเฉลี่ยความแปรปรวนของความคลาดเคลื่อน (MSE) จากการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่มีวิธีการคัดเลือกข้อสอบแตกต่างกัน ซึ่งค่าเฉลี่ยของ MSE ที่เกิดขึ้นจะบอกถึงขนาดของความแปรปรวนของความคลาดเคลื่อนที่เกิดจากการประมาณค่าความสามารถของผู้สอบของวิธีการคัดเลือกแบบทดสอบย่อย โดย

วิธีการคัดเลือกแบบทดสอบย่อยวิธีใดมีค่าเฉลี่ยของ MSE ต่ำกว่า แสดงว่าวิธีการนั้นจะมีประสิทธิภาพในการประมาณค่าความสามารถของผู้สอบได้ดีกว่า

จากแนวโน้มค่าเฉลี่ยของ MSE ที่เปลี่ยนไปเมื่ออัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) เพิ่มขึ้น เมื่อพิจารณาจะพบว่า เมื่อคลังข้อสอบที่มีขนาด 600 ข้อ และ 800 ข้อ เฉลี่ยของ MSE ของวิธี MCC มีค่าสูงกว่าวิธี CWA ในทุกระดับของอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) แสดงว่าวิธี MCC มีประสิทธิภาพในการประมาณค่าความสามารถของผู้สอบสูงกว่าวิธี CWA ในทุกเงื่อนไขการทดลอง

3.1.2 ผลการเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ยสัมประสิทธิ์สหสัมพันธ์ระหว่างความสามารถจริงและความสามารถประมาณค่า

การเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ยสัมประสิทธิ์สหสัมพันธ์ระหว่างความสามารถจริงและความสามารถประมาณค่า ($r_{\theta\hat{\theta}}$) จากวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่ใช้กับคลังข้อสอบ 600 ข้อ ซึ่งมีกลุ่มตัวอย่างทั้งหมด 8 กลุ่ม จะเห็นว่ากลุ่มที่ให้ค่าเฉลี่ยของ $r_{\theta\hat{\theta}}$ สูงที่สุดหรือมีประสิทธิภาพในการประมาณค่าความสามารถของผู้สอบสูงสุดและรองลงมาใน 3 อันดับแรก ได้แก่

อันดับที่ 1 ได้แก่ กลุ่ม MCC25 ซึ่งมีค่าเฉลี่ยของ $r_{\theta\hat{\theta}}$ สูงที่สุด เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของ $r_{\theta\hat{\theta}}$ ไม่แตกต่างกัน

อันดับที่ 2 ได้แก่ กลุ่ม MCC15, MCC20 และ CWA25 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของ $r_{\theta\hat{\theta}}$ ไม่ต่างจากกลุ่มที่เหลือ รายละเอียดดังตารางที่ 4.13

ตารางที่ 4.13 ผลการเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ยสหสัมพันธ์ระหว่างวิธี MCC กับวิธี CWA ที่มีการควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุดแตกต่างกัน สำหรับคลังข้อสอบ 600 ข้อ

กลุ่ม	CWA10	CWA15	CWA20	CWA25	MCC10	MCC15	MCC20	MCC25
M. $r_{\theta\hat{\theta}}$	0.952	0.955	0.955	0.956	0.953	0.956	0.956	0.957
CWA10	-	-19.80	-20.90	-24.30	-8.80	-29.60	-31.30	-37.30**
CWA15		-	-1.10	-4.50	11.00	-9.80	-11.50	-17.50
CWA20			-	-3.40	12.10	-8.70	-10.40	-16.40
CWA25				-	15.50	-5.30	-7.00	-13.00
MCC10					-	-20.80	-22.50	-28.50
MCC15						-	-1.70	-7.70
MCC20							-	-6.00
MCC25								-

* p<.05 , ** p< .01

การเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ยสัมประสิทธิ์สหสัมพันธ์ระหว่างความสามารถจริงและความสามารถประมาณค่า ($r_{\theta\hat{\theta}}$) จากวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่ใช้กับคลังข้อสอบ 800 ข้อ ซึ่งมีกลุ่มตัวอย่างทั้งหมด 8 กลุ่ม จะเห็นว่ากลุ่มที่ให้ค่าเฉลี่ยของ $r_{\theta\hat{\theta}}$ สูงที่สุดหรือมีประสิทธิภาพในการประมาณค่าความสามารถของผู้สอบสูงสุดและรองลงมาใน 3 อันดับแรก ได้แก่

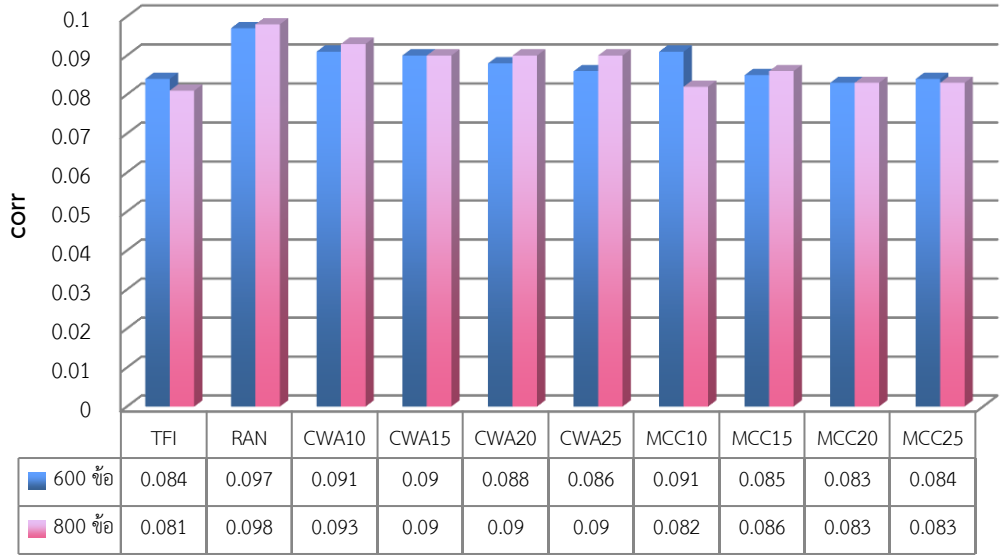
อันดับที่ 1 ได้แก่ กลุ่ม MCC10 ซึ่งมีค่าเฉลี่ยของ $r_{\theta\hat{\theta}}$ สูงที่สุด เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของ $r_{\theta\hat{\theta}}$ มีค่าสูงกว่ากลุ่ม CWA10, CWA15 และ CWA25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และสูงกว่ากลุ่ม CWA20 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

อันดับที่ 2 ได้แก่ กลุ่ม MCC20 และ MCC25 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ย $r_{\theta\hat{\theta}}$ ของกลุ่ม MCC20 มีค่าสูงกว่ากลุ่ม CWA10, CWA15 และ CWA25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และ ค่าเฉลี่ย $r_{\theta\hat{\theta}}$ ของกลุ่ม MCC25 มีค่าสูงกว่ากลุ่ม CWA10 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และสูงกว่ากลุ่ม CWA15 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 รายละเอียดดังตารางที่ 4.14

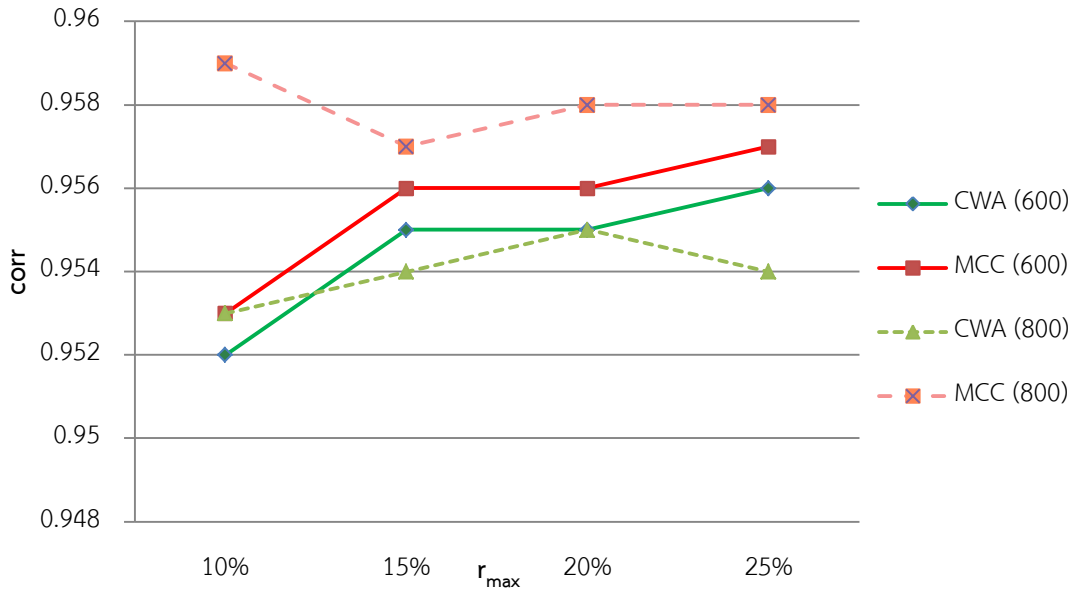
ตารางที่ 4.14 ผลการเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ยสหสัมพันธ์ระหว่างความสามารถประมาณค่ากับความสามารถจริง ระหว่างวิธี MCC กับวิธี CWA ที่มีการควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุดแตกต่างกัน สำหรับคลังข้อสอบ 800 ข้อ

กลุ่ม	CWA10	CWA15	CWA20	CWA25	MCC10	MCC15	MCC20	MCC25
M. $r_{\theta\hat{\theta}}$	0.953	0.954	0.955	0.954	0.959	0.957	0.958	0.958
CWA10	-	8.30	13.40	11.70	48.90**	30.80	45.40**	43.90**
CWA15		-	5.10	3.40	40.60**	22.50	37.10**	35.60*
CWA20			-	-1.70	35.50*	17.40	32.00	30.50
CWA25				-	37.20**	19.10	33.70**	32.20
MCC10					-	-18.10	-3.50	-5.00
MCC15						-	14.60	13.10
MCC20							-	-1.50
MCC25								-

* p<.05 , ** p< .01



ภาพที่ 4.3 กราฟค่าเฉลี่ยของสหสัมพันธ์ระหว่างค่าเฉลี่ยของสหสัมพันธ์ระหว่างความสามารถประมาณค่ากับความสามารถจริง ระหว่างวิธี MCC กับวิธี CWA ที่มีการควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุดแตกต่างกัน



ภาพที่ 4.4 กราฟแสดงความสัมพันธ์ระหว่างค่าเฉลี่ยสหสัมพันธ์ระหว่างความสามารถประมาณค่ากับความสามารถจริง และ r_{max} ระหว่างวิธี MCC กับวิธี CWA ที่มีการควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุดแตกต่างกัน

จากภาพที่ 4.3 และ 4.4 แสดงกราฟเปรียบเทียบค่าเฉลี่ยของค่าเฉลี่ยสัมประสิทธ์สหสัมพันธ์ระหว่างความสามารถประมาณค่ากับความสามารถจริง ($r_{\theta\hat{\theta}}$) จากการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่มีวิธีการคัดเลือกข้อสอบแตกต่างกัน ซึ่งสหสัมพันธ์ระหว่างความสามารถประมาณค่ากับความสามารถจริง ($r_{\theta\hat{\theta}}$) ที่เกิดขึ้นจะบอกถึงความตรงในการประมาณค่าความสามารถของผู้สอบของวิธีการคัดเลือกแบบทดสอบย่อย โดยวิธีการคัดเลือกแบบทดสอบย่อยวิธีใดมีค่าสหสัมพันธ์ระหว่างความสามารถประมาณค่ากับความสามารถจริงเฉลี่ยสูงกว่า แสดงว่าวิธีการนั้นจะมีประสิทธิภาพในการประมาณค่าความสามารถของผู้สอบได้ดีกว่า

จากแนวโน้มค่าเฉลี่ยของ $r_{\theta\hat{\theta}}$ ที่เปลี่ยนไปเมื่ออัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) เพิ่มขึ้น เมื่อพิจารณาจะพบว่า เมื่อคลังข้อสอบที่มีขนาด 600 ข้อ และ 800 ข้อ เฉลี่ยของ $r_{\theta\hat{\theta}}$ ของวิธี MCC มีค่าสูงกว่าวิธี CWA ในทุกระดับของอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) แสดงว่า วิธี MCC มีประสิทธิภาพในการประมาณค่าความสามารถของผู้สอบสูงกว่าวิธี CWA ในทุกเงื่อนไขการทดลอง

3.1.3 ผลการเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของความยาวแบบสอบเฉลี่ย

การเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของความยาวแบบสอบเฉลี่ยจากวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่ใช้กับคลังข้อสอบ 600 ข้อ ซึ่งมีกลุ่มตัวอย่างทั้งหมด 8 กลุ่ม จะเห็นว่า กลุ่มที่ให้ค่าเฉลี่ยของความยาวแบบสอบเฉลี่ยต่ำสุดหรือมีประสิทธิภาพในการประมาณค่าความสามารถของผู้สอบสูงสุดและรองลงมาใน 3 อันดับแรก ได้แก่

อันดับที่ 1 ได้แก่ กลุ่ม MCC25 ซึ่งมีค่าเฉลี่ยของความยาวแบบสอบเฉลี่ยต่ำที่สุด เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของความยาวแบบสอบเฉลี่ย ต่ำกว่ากลุ่ม CWA10, CWA15และ CWA20 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และต่ำกว่ากลุ่ม CWA25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ส่วนกลุ่มอื่นๆ ที่เหลือมีค่าไม่แตกต่างกัน

อันดับที่ 2 ได้แก่ กลุ่ม MCC20 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของความยาวแบบสอบเฉลี่ย ต่ำกว่ากลุ่ม CWA10, CWA15 และ CWA20 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และต่ำกว่ากลุ่ม CWA25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

อันดับที่ 3 ได้แก่ กลุ่ม MCC15 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของความยาวแบบสอบเฉลี่ย ต่ำกว่ากลุ่ม CWA10 และ CWA15 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และต่ำกว่ากลุ่ม CWA20 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 รายละเอียดดังตารางที่ 4.15

ตารางที่ 4.15 ผลการเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของความยาวแบบสอบเฉลี่ย ระหว่างวิธี MCC กับวิธี CWA ที่มีการควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุดแตกต่างกัน สำหรับคลังข้อสอบ 600 ข้อ

กลุ่ม	CWA10	CWA15	CWA20	CWA25	MCC10	MCC15	MCC20	MCC25
M.it	53.45	54.03	53.08	51.72	45.60	41.51	38.69	37.09
CWA10	-	-6.60	5.90	17.90	29.30	39.30**	51.00**	57.60**
CWA15		-	12.50	24.50	35.90*	45.90**	57.60**	64.20**
CWA20			-	12.00	23.40	33.40*	45.10**	51.70**
CWA25				-	11.40	21.40	33.10*	39.70*
MCC10					-	10.00	21.70	28.30
MCC15						-	11 .70	18.30
MCC20							-	6.60
MCC25								-

* p<.05 , ** p< .01

การเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของความยาวแบบสอบเฉลี่ยจากวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่ใช้กับคลังข้อสอบ 800 ข้อ ซึ่งมีกลุ่มตัวอย่างทั้งหมด 8 กลุ่ม จะเห็นว่า กลุ่มที่ให้ค่าเฉลี่ยของความยาวแบบสอบเฉลี่ยต่ำสุดหรือมีประสิทธิภาพในการประมาณค่าความสามารถของผู้สอบสูงสุดและรองลงมาใน 3 อันดับแรก ได้แก่

อันดับที่ 1 ได้แก่ กลุ่ม MCC25 ซึ่งมีค่าเฉลี่ยของความยาวแบบสอบเฉลี่ยต่ำที่สุด เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่า มีค่าเฉลี่ยของความยาวแบบสอบเฉลี่ยต่ำกว่ากลุ่ม CWA10, CWA15, CWA20 และ CWA25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ส่วนกลุ่ม MCC ไม่แตกต่างกัน

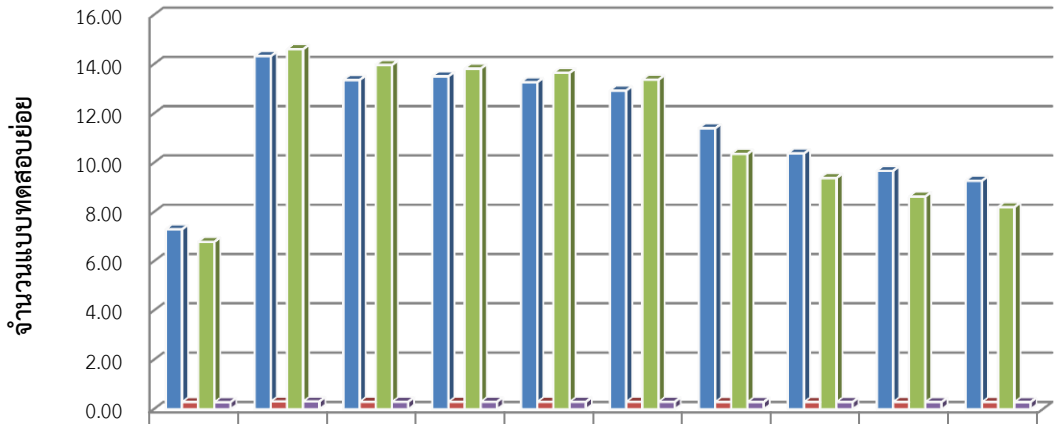
อันดับที่ 2 ได้แก่ กลุ่ม MCC20 ซึ่งมีค่าเฉลี่ยของความยาวแบบสอบเฉลี่ยต่ำรองลงมา เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่า มีค่าเฉลี่ยของความยาวแบบสอบเฉลี่ยต่ำกว่ากลุ่ม CWA10, CWA15 และ CWA20 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ส่วนกลุ่มอื่นๆ ไม่แตกต่างกัน

อันดับที่ 3 ได้แก่ กลุ่ม MCC15 ซึ่งมีค่าเฉลี่ยของความยาวแบบสอบเฉลี่ยต่ำรองลงมาเป็นอันดับที่ 3 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่า มีค่าเฉลี่ยของความยาวแบบสอบเฉลี่ยต่ำกว่ากลุ่ม CWA10 และ CWA15 CWA20 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ส่วนกลุ่มอื่นๆ ไม่แตกต่างกัน รายละเอียดดังตารางที่ 4.16

ตารางที่ 4.16 ผลการเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของความยาวแบบสอบเฉลี่ย ระหว่างวิธี MCC กับวิธี CWA ที่มีการควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุดแตกต่างกัน สำหรับคลังข้อสอบ 800 ข้อ

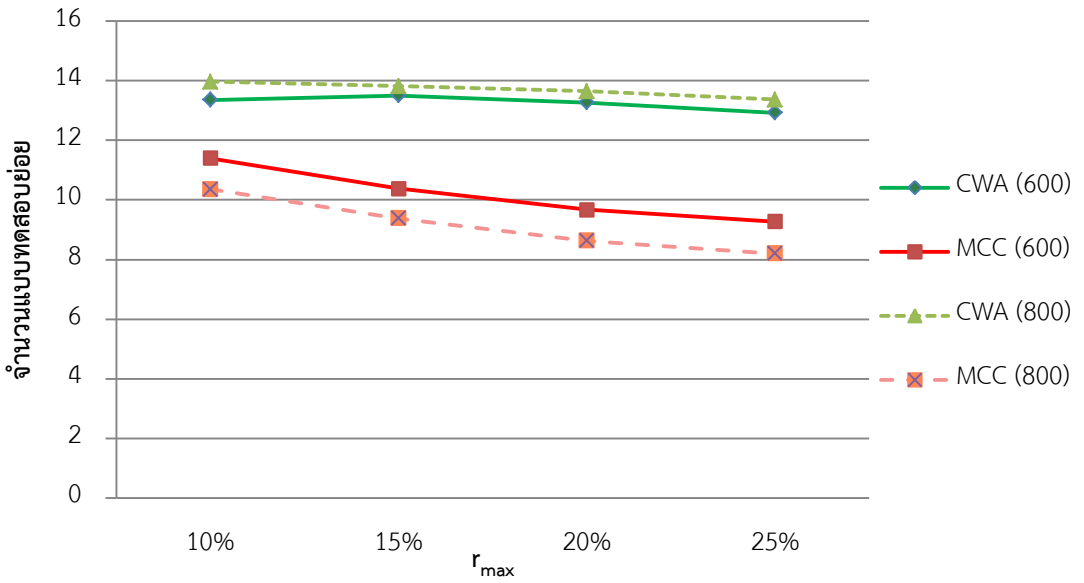
กลุ่ม	CWA10	CWA15	CWA20	CWA25	MCC10	MCC15	MCC20	MCC25
M.it	55.87	55.27	54.60	53.47	41.45	37.51	34.51	32.80
CWA10	-	8.05	16.65	26.90	37.90**	48.10**	58.60**	67.00**
CWA15		-	8.60	18.85	29.85	40.05**	50.55**	58.95**
CWA20			-	10.25	21.25	31.45	41.95**	50.35**
CWA25				-	11.00	21.20	31.70	40.10**
MCC10					-	10.20	20.70	29.10
MCC15						-	10.50	18.90
MCC20							-	8.40
MCC25								-

* p<.05 , ** p< .01



	TFI	RAN	CWA10	CWA15	CWA20	CWA25	MCC10	MCC15	MCC20	MCC25
600 ข้อ	7.30	14.33	13.36	13.51	13.27	12.93	11.40	10.38	9.67	9.27
SEE (600)	0.289	0.313	0.299	0.299	0.298	0.298	0.294	0.292	0.291	0.291
800 ข้อ	6.79	14.61	13.97	13.82	13.65	13.37	10.36	9.38	8.63	8.20
SEE (800)	0.283	0.312	0.299	0.300	0.299	0.298	0.291	0.290	0.289	0.288

ภาพที่ 4.5 กราฟค่าเฉลี่ยของความยาวแบบสอบ ระหว่างวิธี MCC กับวิธี CWA ที่มีการควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุดแตกต่างกัน



ภาพที่ 4.6 กราฟแสดงความสัมพันธ์ระหว่างเฉลี่ยของความยาวแบบสอบ และ r_{max} ระหว่างวิธี MCC กับวิธี CWA ที่มีการควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุดแตกต่างกัน

จากภาพที่ 4.5 และ 4.6 แสดงกราฟเปรียบเทียบค่าเฉลี่ยของความยาวแบบสอบเฉลี่ยจากการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่มีวิธีการคัดเลือกข้อสอบแตกต่างกัน ซึ่งค่าความยาวแบบสอบเฉลี่ยที่เกิดขึ้นจะบอกถึงประสิทธิภาพในการประมาณค่าความสามารถของผู้สอบของวิธีการคัดเลือกแบบทดสอบย่อย โดยวิธีการคัดเลือกแบบทดสอบย่อยวิธีใดใช้ความยาวของแบบสอบเฉลี่ยต่ำกว่าในการประมาณค่าความสามารถของผู้สอบให้มีค่าความคลาดเคลื่อนมาตรฐานในการประมาณค่า (SEE) น้อยกว่าหรือเท่ากับ 0.3 แสดงว่าวิธีการนั้นจะมีประสิทธิภาพในการประมาณค่าความสามารถของผู้สอบสูงกว่า

จากแนวโน้มความยาวของแบบสอบเฉลี่ยที่เปลี่ยนไปเมื่ออัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max}) เพิ่มขึ้น เมื่อพิจารณาจะพบว่า เมื่อคลังข้อสอบที่มีขนาด 600 ข้อ และ 800 ข้อ ความยาวของแบบสอบเฉลี่ย ของวิธี MCC มีค่าต่ำกว่าวิธีการคัดเลือกแบบทดสอบย่อยแบบ CWA ในทุกระดับของอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max}) แสดงว่า วิธี MCC มีประสิทธิภาพในการประมาณค่าความสามารถของผู้สอบสูงกว่าวิธี CWA ในทุกเงื่อนไขการทดลอง

3.2 ผลการเปรียบเทียบประสิทธิภาพด้านการใช้คลังข้อสอบ

จากการเปรียบเทียบความแตกต่างของค่าสถิติที่ใช้เป็นดัชนีในการพิจารณาประสิทธิภาพด้านการใช้คลังข้อสอบโดยใช้การทดสอบของ Kruskal-Wallis ทำให้ผู้วิจัยทราบว่า ดัชนีในการพิจารณาประสิทธิภาพด้านการใช้คลังข้อสอบทุกตัวของตัวอย่างทั้ง 8 กลุ่ม แตกต่างกันอย่างมีนัยสำคัญทาง

สถิติที่ระดับ .01 ยกเว้น ผู้วิจัยจึงเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่โดยใช้วิธี Kruskal-Wallis one-way ANOVA กับค่าสถิติที่ใช้เป็นดัชนีในการพิจารณาประสิทธิภาพด้านการใช้คลังข้อสอบทุกตัว ผลการเปรียบเทียบมีรายละเอียดดังต่อไปนี้

3.2.1 ผลการเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ยอัตราการทับซ้อนของแบบสอบ

การเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ยอัตราการทับซ้อนของแบบสอบจากวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่ใช้กับคลังข้อสอบ 600 ข้อ ซึ่งมีกลุ่มตัวอย่างทั้งหมด 8 กลุ่ม จะเห็นว่า กลุ่มที่ให้ค่าเฉลี่ยของอัตราการทับซ้อนของแบบสอบต่ำที่สุดหรือมีประสิทธิภาพด้านการใช้คลังข้อสอบสูงสุดและรองลงมาใน 3 อันดับแรก ได้แก่

อันดับที่ 1 และ 2 ได้แก่ กลุ่ม MCC15 และ กลุ่ม MCC10 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นปรากฏว่ามีค่าเฉลี่ยของอัตราการทับซ้อนของแบบสอบมีค่าต่ำกว่ากลุ่ม CWA10, CWA15 , CWA20 และ CWA25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01

อันดับที่ 3 ได้แก่ กลุ่ม MCC20 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของอัตราการทับซ้อนของแบบสอบมีค่าต่ำกว่ากลุ่ม CWA10 และ CWA25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และต่ำกว่ากลุ่ม CWA20 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 รายละเอียดดังตารางที่ 4.17

ตารางที่ 4.17 ผลการเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ยอัตราการทับซ้อนของแบบสอบ ระหว่างวิธี MCC กับวิธี CWA ที่มีการควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุดแตกต่างกัน สำหรับคลังข้อสอบ 600 ข้อ

กลุ่ม	CWA10	CWA15	CWA20	CWA25	MCC10	MCC15	MCC20	MCC25
M. $\bar{t}$	0.112	0.107	0.109	0.110	0.061	0.060	0.063	0.069
CWA10	-	16.20	11.80	8.00	54.40**	60.90**	46.00**	34.70*
CWA15		-	-4.40	-8.20	38.20**	44.70**	29.80	18.50
CWA20			-	-3.80	42.60**	49.10**	34.20*	22.90
CWA25				-	46.40**	52.90**	38.00**	26.70
MCC10					-	6.50	-8.40	-19.70
MCC15						-	-14.90	-26.20
MCC20							-	-11.30
MCC25								-

* p<.05 , ** p< .01

การเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ยอัตราการทับซ้อนของแบบสอบจากวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่ใช้กับคลังข้อสอบ 800 ข้อ

ซึ่งมีกลุ่มตัวอย่างทั้งหมด 8 กลุ่ม จะเห็นว่า กลุ่มที่ให้ค่าเฉลี่ยของอัตราการทับซ้อนของแบบสอบต่ำที่สุดหรือมีประสิทธิภาพด้านการใช้คลังข้อสอบสูงสุดและรองลงมาใน 3 อันดับแรก ได้แก่

อันดับที่ 1 ได้แก่ กลุ่ม MCC10 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของอัตราการทับซ้อนของแบบสอบมีค่าต่ำกว่ากลุ่ม CWA10, CWA15 , CWA20 และ CWA25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01

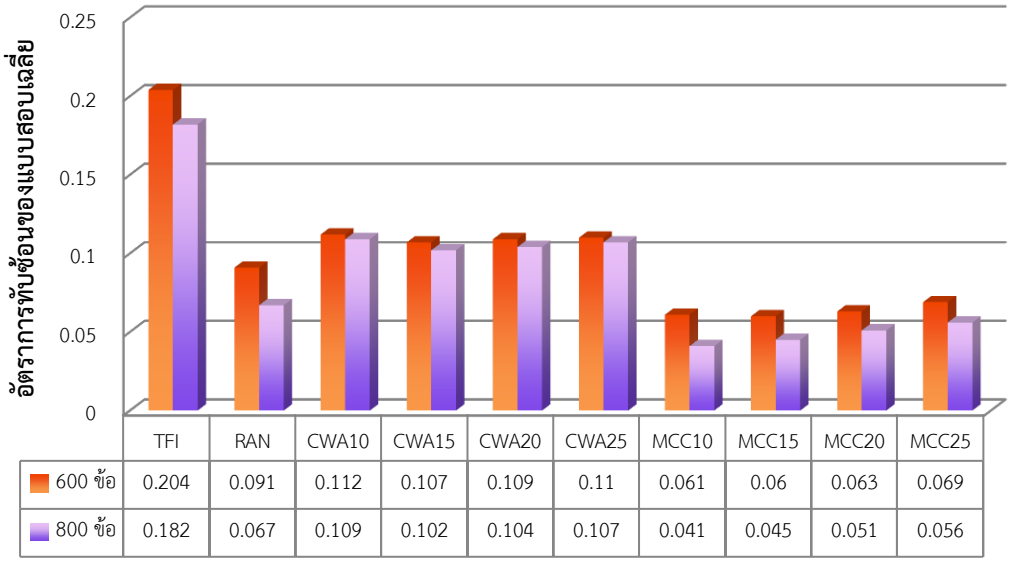
อันดับที่ 2 ได้แก่ กลุ่ม MCC15 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของอัตราการทับซ้อนของแบบสอบมีค่าต่ำกว่ากลุ่ม CWA10, CWA20 และ CWA25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01

อันดับที่ 3 ได้แก่ กลุ่ม MCC20 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของอัตราการทับซ้อนของแบบสอบมีค่าต่ำกว่ากลุ่ม CWA10 และ CWA25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 รายละเอียดดังตารางที่ 4.18

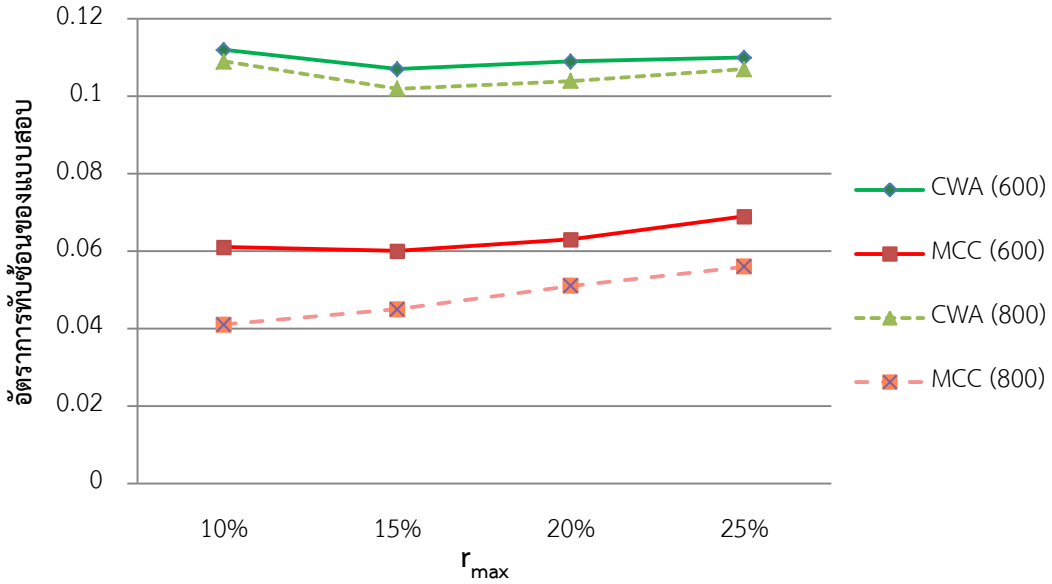
ตารางที่ 4.18 ผลการเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ยอัตราการทับซ้อนของแบบสอบ ระหว่างวิธี MCC กับวิธี CWA ที่มีการควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุดแตกต่างกัน สำหรับคลังข้อสอบ 800 ข้อ

กลุ่ม	CWA10	CWA15	CWA20	CWA25	MCC10	MCC15	MCC20	MCC25
M. $\bar{t}$	0.109	0.102	0.104	0.107	0.041	0.045	0.051	0.056
CWA10	-	24.80	16.10	5.90	66.60**	56.80**	46.60**	36.80*
CWA15		-	-8.70	-18.90	41.80**	32.00	21.80	12.00
CWA20			-	-10.20	50.50**	40.70**	30.50	20.70
CWA25				-	60.70**	50.90**	40.70**	30.90
MCC10					-	-9.80	-20.00	-29.80
MCC15						-	-10.20	-20.00
MCC20							-	-9.80
MCC25								-

* p<.05 , ** p< .01



ภาพที่ 4.7 กราฟค่าอัตราการทับซ้อนของแบบสอบเฉลี่ย ของวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่ใช้ในการวิจัยครั้งนี้



ภาพที่ 4.8 กราฟแสดงความสัมพันธ์ระหว่างอัตราการทับซ้อนของแบบสอบ เฉลี่ยและ r_{max} ของวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่ใช้ในการวิจัยครั้งนี้

จากภาพที่ 4.7 และ 4.8 แสดงกราฟเปรียบเทียบค่าเฉลี่ยของอัตราการทับซ้อนของแบบสอบ ($\bar{t}$) จากการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่มีวิธีการคัดเลือกข้อสอบแตกต่างกัน หลักเกณฑ์ในการพิจารณา คือ วิธีคัดเลือกแบบทดสอบย่อยที่มีค่าเฉลี่ยของอัตราการทับซ้อนของแบบสอบ ($\bar{t}$) ต่ำกว่าจะเป็นวิธีการคัดเลือกแบบทดสอบย่อยที่ควบคุมการใช้แบบทดสอบย่อยซ้ำได้ดีกว่า

จากแนวโน้มของอัตราการทับซ้อนของแบบสอบ ($\bar{t}$) ที่เปลี่ยนไปเมื่ออัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max}) เพิ่มขึ้น เมื่อพิจารณาจะพบว่า เมื่อใช้คลังข้อสอบที่มีขนาด 600 ข้อ และ 800 ข้อ ค่าเฉลี่ยของอัตราการทับซ้อนของแบบสอบ ($\bar{t}$) ของวิธี MCC มีค่าต่ำกว่าวิธี CWA ในทุกระดับของอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max}) แสดงว่า วิธี MCC สามารถควบคุมการใช้แบบทดสอบย่อยซ้ำได้ดีกว่าวิธี CWA ในทุกเงื่อนไขการทดลอง

3.2.2 ผลการเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ยไคสแควร์

การเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ยไคสแควร์ จากวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่ใช้กับคลังข้อสอบ 600 ข้อ ซึ่งมีกลุ่มตัวอย่างทั้งหมด 8 กลุ่ม จะเห็นว่า กลุ่มที่ให้ค่าเฉลี่ยไคสแควร์ ต่ำที่สุดหรือมีประสิทธิภาพด้านการใช้คลังข้อสอบสูงสุดและรองลงมาใน 3 อันดับแรก ได้แก่

อันดับที่ 1 ได้แก่ กลุ่ม MCC10 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของไคสแควร์ มีค่าต่ำกว่ากลุ่ม CWA10, CWA25 และ MCC25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และต่ำกว่ากลุ่ม CWA20 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ส่วนกลุ่มอื่นๆ ที่เหลือนั้นมีค่าไม่ต่างกัน

อันดับที่ 2 ได้แก่ กลุ่ม MCC15 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของไคสแควร์ มีค่าต่ำกว่ากลุ่ม CWA25 และ MCC25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และต่ำกว่ากลุ่ม CWA10 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ส่วนกลุ่มอื่นๆ ที่เหลือนั้นมีค่าไม่ต่างกัน

อันดับที่ 3 ได้แก่ กลุ่ม CWA15 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของไคสแควร์ มีค่าต่ำกว่ากลุ่ม CWA25 และ MCC25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 รายละเอียดดังตารางที่ 4.19

ตารางที่ 4.19 ผลการเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ยไคสแควร์ ระหว่างวิธี MCC กับวิธี CWA ที่มีการควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุดแตกต่างกัน สำหรับคลังข้อสอบขนาด 600 ข้อ

กลุ่ม	CWA10	CWA15	CWA20	CWA25	MCC10	MCC15	MCC20	MCC25
M. χ^2	5.213	4.208	4.895	5.700	1.519	3.310	5.231	6.814
CWA10	-	24.20	11.70	-14.00	44.20**	34.20*	-0.90	-25.80
CWA15		-	-12.50	-38.20**	20.00	10.00	-25.10	-50.00**
CWA20			-	-25.70	32.50*	22.50	-12.60	-37.50**
CWA25				-	58.20**	48.20**	13.10	-11.80
MCC10					-	-10.00	-45.10	-70.00**
MCC15						-	-35.10*	-60.00**
MCC20							-	-24.90
MCC25								-

* p<.05 , ** p< .01

การเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ยไคสแควร์ จากวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่ใช้กับคลังข้อสอบ 800 ข้อ ซึ่งมีกลุ่มตัวอย่างทั้งหมด 8 กลุ่ม จะเห็นว่า กลุ่มที่ให้ค่าเฉลี่ยของไคสแควร์ ต่ำที่สุดหรือมีประสิทธิภาพด้านการใช้คลังข้อสอบสูงสุดและรองลงมาใน 3 อันดับแรก ได้แก่

อันดับที่ 1 ได้แก่ กลุ่ม MCC10 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของไคสแควร์ มีค่าต่ำกว่ากลุ่ม CWA10, CWA20, CWA25 และ MCC25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ส่วนกลุ่มอื่นๆ ที่เหลือนั้นมีค่าไม่ต่างกัน

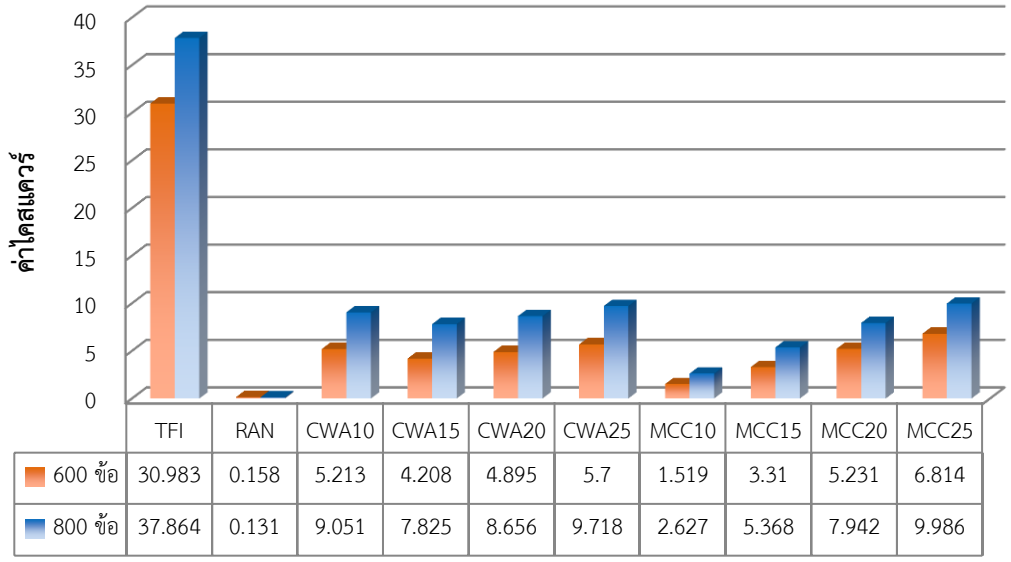
อันดับที่ 2 ได้แก่ กลุ่ม MCC15 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของไคสแควร์ มีค่าต่ำกว่ากลุ่ม CWA10, CWA25 และ MCC25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ส่วนกลุ่มอื่นๆ ที่เหลือนั้นมีค่าไม่ต่างกัน

อันดับที่ 3 ได้แก่ กลุ่ม MCC20 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของไคสแควร์ มีค่าต่ำกว่ากลุ่ม MCC25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และต่ำกว่ากลุ่ม CWA25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 รายละเอียดดังตารางที่ 4.20

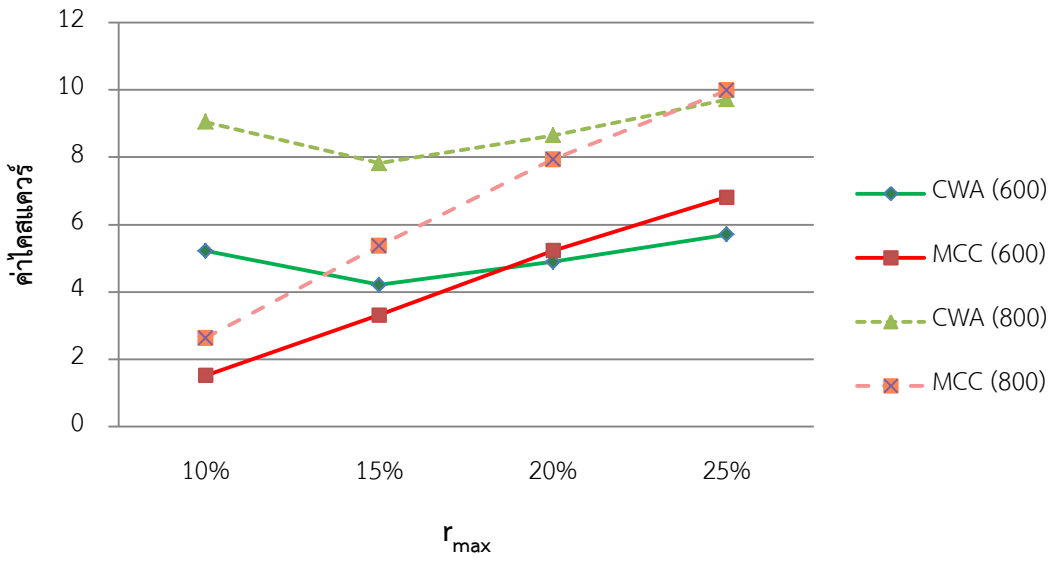
ตารางที่ 4.20 ผลการเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ยไคสแควร์ ระหว่างวิธี MCC กับวิธี CWA ที่มีการควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุดแตกต่างกัน สำหรับคลังข้อสอบขนาด 800 ข้อ

กลุ่ม	CWA10	CWA15	CWA20	CWA25	MCC10	MCC15	MCC20	MCC25
M. χ^2	9.051	7.825	8.656	9.718	2.627	5.368	7.942	9.986
CWA10	-	23.30	6.90	-12.80	48.40**	38.40**	20.90	-17.90
CWA15		-	-16.40	-36.10	25.10	15.10	-2.40	-41.20**
CWA20			-	-19.70	41.50**	31.50	14.00	-24.80
CWA25				-	61.20**	51.20**	33.70*	-5.10
MCC10					-	-10.00	-27.50	-66.30**
MCC15						-	-17.50	-56.30**
MCC20							-	-38.80**
MCC25								-

* p<.05 , ** p< .01



ภาพที่ 4.9 กราฟค่าเฉลี่ยของไคสแควร์ ของวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่ใช้ในการวิจัยครั้งนี้



ภาพที่ 4.10 กราฟแสดงความสัมพันธ์ระหว่างเฉลี่ยของไคสแควร์ และ r_{max} ของวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่ใช้ในการวิจัยครั้งนี้

จากภาพที่ 4.9 และ 4.10 แสดงกราฟเปรียบเทียบค่าเฉลี่ยของไคสแควร์ (x^2) จากการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่มีวิธีการคัดเลือกข้อสอบแตกต่างกัน ซึ่งค่าเฉลี่ยของไคสแควร์ (x^2) ที่เกิดขึ้นจะบอกถึงความแตกต่างระหว่างอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ (observed testlet exposure rates) กับอัตราการใช้แบบทดสอบย่อยซ้ำที่คาดหวัง (expected testlet exposure rates) ค่าเฉลี่ยของไคสแควร์ที่ต่ำกว่า แสดงว่า อัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้กับอัตราการใช้แบบทดสอบย่อยซ้ำที่คาดหวังแตกต่างกันน้อยกว่า นั่นหมายความว่าอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้มีลักษณะการแจกแจงแบบยูนิฟอร์ม คือ แบบทดสอบย่อยทั้งคลังข้อสอบถูกนำมาใช้กับผู้สอบด้วยความถี่ที่เท่าๆ กัน

จากแนวโน้มค่าเฉลี่ยของไคสแควร์ (x^2) ที่เปลี่ยนไปเมื่ออัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max}) เพิ่มขึ้น เมื่อพิจารณาจะพบว่า ถ้าใช้วิธีการคัดเลือกแบบทดสอบย่อยแบบ MCC กับคลังข้อสอบที่มีขนาด 600 ข้อ ค่าไคสแควร์ (x^2) ของวิธี MCC มีค่าต่ำกว่า CWA เมื่อกำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max}) ไว้ที่ 10 และ 15 เปอร์เซ็นต์ แสดงว่าอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ของวิธี MCC มีการแจกแจงใกล้เคียงแบบยูนิฟอร์มมากกว่าวิธี CWA

ในกรณีที่ใช้คลังข้อสอบมีขนาด 800 ข้อ และกำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max}) เท่ากับ 25 เปอร์เซ็นต์ ค่าไคสแควร์ (x^2) ของวิธี MCC (9.986) และวิธี CWA (9.781) มีค่า

ใกล้เคียงกันมาก แสดงว่าอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ของทั้ง 2 วิธี มีการแจกแจงแบบยูนิฟอร์มที่มีลักษณะคล้ายๆ กัน

3.2.3 ผลการเปรียบเทียบค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้

การเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ จากวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่ใช้กับคลังข้อสอบ 600 ข้อ ซึ่งมีกลุ่มตัวอย่างทั้งหมด 8 กลุ่ม จะเห็นว่า กลุ่มที่ให้ค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ต่ำที่สุดหรือมีประสิทธิภาพด้านการใช้คลังข้อสอบสูงสุดและรองลงมาใน 3 อันดับแรก ได้แก่

อันดับที่ 1 ได้แก่ กลุ่ม MCC10 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นปรากฏว่ามีค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้มีค่าต่ำกว่ากลุ่ม CWA10, CWA15 และ MCC25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และต่ำกว่ากลุ่ม CWA25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

อันดับที่ 2 ได้แก่ กลุ่ม MCC15 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นปรากฏว่ามีค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้มีค่าต่ำกว่ากลุ่ม CWA10, CWA15 และ MCC25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และต่ำกว่ากลุ่ม CWA25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

อันดับที่ 3 ได้แก่ กลุ่ม MCC20 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นปรากฏว่ามีค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้มีค่าต่ำกว่ากลุ่ม CWA10 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และต่ำกว่ากลุ่ม CWA15 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 รายละเอียดดังตารางที่ 4.21

ตารางที่ 4.21 ผลการเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ ระหว่างวิธี MCC กับวิธี CWA ที่มีการควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุดแตกต่างกัน สำหรับคลังข้อสอบ 600 ข้อ

กลุ่ม	CWA10	CWA15	CWA20	CWA25	MCC10	MCC15	MCC20	MCC25
M. r_{obs}	0.344	0.206	0.177	0.196	0.128	0.132	0.165	0.200
CWA10	-	16.70	42.00**	26.10	58.80**	63.00**	52.00**	21.40
CWA15		-	25.30	9.40	42.10**	46.30**	35.30*	4.70
CWA20			-	-15.90	16.80	21.00	10.00	-20.60
CWA25				-	32.70*	36.90*	25.90	-4.70
MCC10					-	4.20	-6.80	-37.40**
MCC15						-	-11.00	-41.60**
MCC20							-	-30.60
MCC25								-

* p<.05 , ** p< .01

การเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ จากวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่ใช้กับคลังข้อสอบ 800 ข้อ ซึ่งมีกลุ่มตัวอย่างทั้งหมด 8 กลุ่ม จะเห็นว่า กลุ่มที่ให้ค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ต่ำที่สุดหรือมีประสิทธิภาพด้านการใช้คลังข้อสอบสูงสุดและรองลงมาใน 3 อันดับแรก ได้แก่

อันดับที่ 1 ได้แก่ กลุ่ม MCC10 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นปรากฏว่ามีค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้มีค่าต่ำกว่ากลุ่ม CWA10, CWA15, CWA25 และ MCC25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01

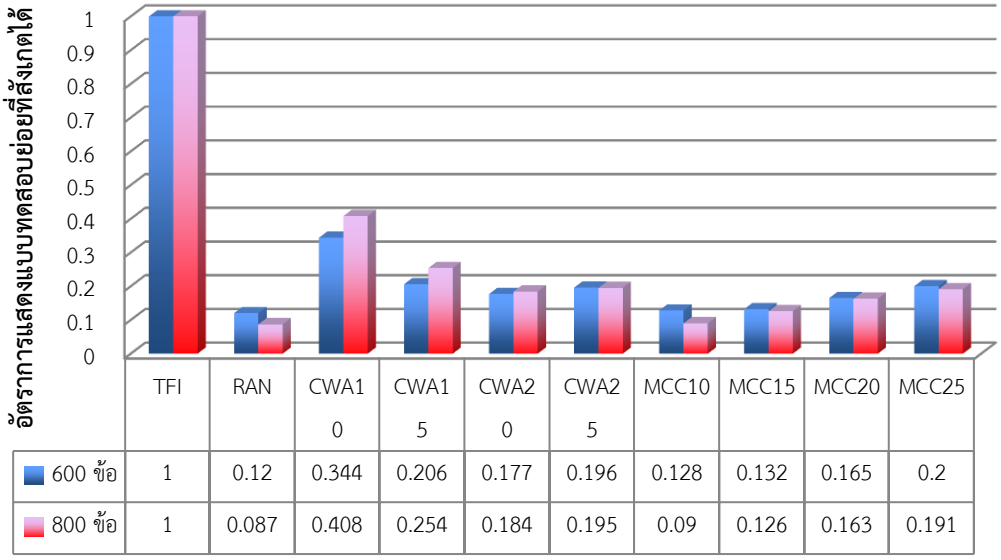
อันดับที่ 2 ได้แก่ กลุ่ม MCC15 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นปรากฏว่ามีค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้มีค่าต่ำกว่ากลุ่ม CWA10, CWA15 และ CWA25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01

อันดับที่ 3 ได้แก่ กลุ่ม MCC20 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นปรากฏว่ามีค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้มีค่าต่ำกว่ากลุ่ม CWA10 และ CWA15 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 รายละเอียดดังตารางที่ 4.22

ตารางที่ 4.22 ผลการเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ยอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ ระหว่างวิธี MCC กับวิธี CWA ที่มีการควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุดแตกต่างกัน สำหรับคลังข้อสอบ 800 ข้อ

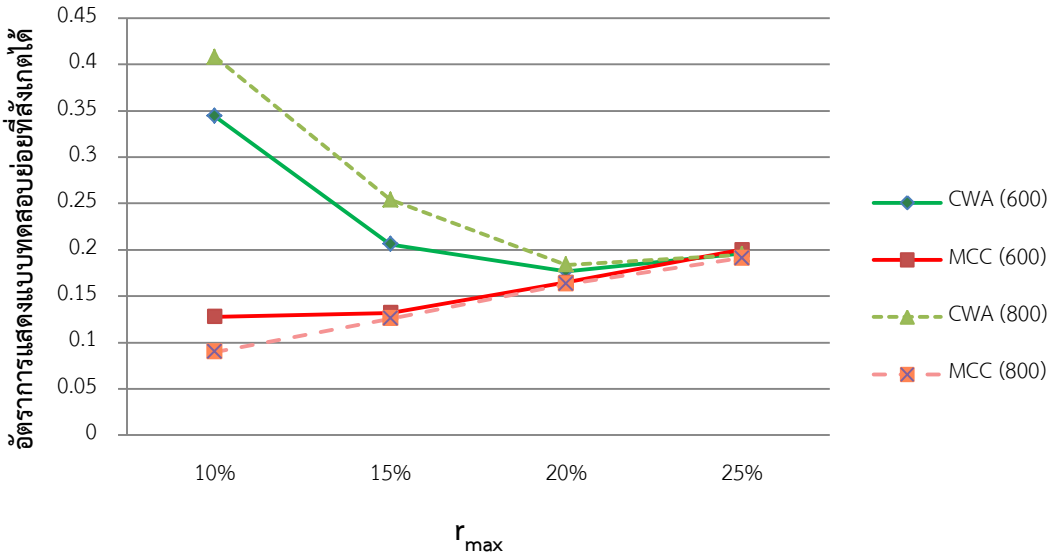
กลุ่ม	CWA10	CWA15	CWA20	CWA25	MCC10	MCC15	MCC20	MCC25
M. r_{obs}	0.408	0.254	0.184	0.195	0.090	0.126	0.163	0.191
CWA10	-	10.00	40.15**	22.05	70.00**	60.00**	49.85**	27.95
CWA15		-	30.15	12.05	60.00**	50.00**	39.85**	17.95
CWA20			-	-18.10	29.85	19.85	9.70	-12.20
CWA25				-	47.95**	37.95**	27.80	5.90
MCC10					-	-10.00	-20.15	-42.05**
MCC15						-	-10.15	-32.05
MCC20							-	-21.90
MCC25								-

* p<.05 , ** p< .01



	TFI	RAN	CWA10	CWA15	CWA20	CWA25	MCC10	MCC15	MCC20	MCC25
600 ข้อ	1	0.12	0.344	0.206	0.177	0.196	0.128	0.132	0.165	0.2
800 ข้อ	1	0.087	0.408	0.254	0.184	0.195	0.09	0.126	0.163	0.191

ภาพที่ 4.11 กราฟค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ ระหว่างวิธี MCC กับวิธี CWA ที่มีการควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุดแตกต่างกัน



ภาพที่ 4.12 กราฟแสดงความสัมพันธ์ระหว่างเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้และ r_{max} ระหว่างวิธี MCC กับวิธี CWA ที่มีการควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุดแตกต่างกัน

จากภาพที่ 4.11 และ 4.12 แสดงกราฟเปรียบเทียบค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ (r_{obs}) จากการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่มีวิธีการคัดเลือกข้อสอบ

แตกต่างกัน หลักเกณฑ์ในการพิจารณา คือ วิธีคัดเลือกแบบทดสอบย่อยวิธีใดที่มีค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ (r_{obs}) ต่ำกว่าอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) จะเป็นวิธีการคัดเลือกแบบทดสอบย่อยที่ควบคุมการใช้แบบทดสอบย่อยซ้ำได้ดีกว่า

จากแนวโน้มของค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ (r_{obs}) ที่เปลี่ยนไปเมื่ออัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) เพิ่มขึ้น เมื่อพิจารณาจะพบว่า เมื่อคลังข้อสอบที่มีขนาด 600 ข้อ และ 800 ข้อ ค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ (r_{obs}) ของวิธี MCC มีค่าต่ำกว่าหรือใกล้เคียงกับอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) ทุกระดับ ขณะที่วิธี CWA จะมีค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ (r_{obs}) ต่ำกว่าอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) เมื่อกำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) อยู่ที่ระดับ 20 และ 25 เปอร์เซ็นต์ แสดงว่า วิธี MCC สามารถควบคุมการใช้แบบทดสอบย่อยซ้ำได้ดีกว่าวิธี CWA ในทุกเงื่อนไขการทดลอง ขณะที่วิธี CWA สามารถควบคุมการใช้แบบทดสอบย่อยซ้ำได้ดีเมื่อกำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) อยู่ที่ระดับ 20 และ 25 เปอร์เซ็นต์

3.2.4 ผลการเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ยจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้

การการเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ยจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ จากวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่ใช้กับคลังข้อสอบ 600 ข้อ ซึ่งมีกลุ่มตัวอย่างทั้งหมด 8 กลุ่ม จะเห็นว่า กลุ่มที่ให้ค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ที่มีค่าต่ำที่สุดหรือมีประสิทธิภาพด้านการใช้คลังข้อสอบสูงสุดและรองลงมาใน 3 อันดับแรก ได้แก่

อันดับที่ 1 ได้แก่ กลุ่ม MCC10 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้มีค่าต่ำกว่ากลุ่ม CWA20, CWA25, MCC20 และ MCC25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01

อันดับที่ 2 ได้แก่ กลุ่ม CWA10 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ มีค่าต่ำกว่ากลุ่ม MCC20 และ MCC25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และต่ำกว่ากลุ่ม CWA25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

อันดับที่ 3 ได้แก่ กลุ่ม MCC15 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ มีค่าต่ำกว่ากลุ่ม MCC20 และ MCC25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 รายละเอียดดังตารางที่ 4.23

ตารางที่ 4.23 ผลการเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ยจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ ระหว่างวิธี MCC กับวิธี CWA ที่มีการควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุดแตกต่างกัน สำหรับคลังข้อสอบ 600 ข้อ

กลุ่ม	CWA10	CWA15	CWA20	CWA25	MCC10	MCC15	MCC20	MCC25
$M.N_{neverexposed}$	6.200	9.100	12.700	17.000	0.000	7.600	19.400	30.300
CWA10	-	-10.40	-23.45	-36.15*	15.45	-4.80	-42.50**	-54.55**
CWA15		-	-13.05	-25.75	25.85	5.60	-32.10	-44.15**
CWA20			-	-12.70	38.90**	18.65	-19.05	-31.10
CWA25				-	51.60**	31.35	-6.35	-18.40
MCC10					-	-20.25	-57.95**	-70.00**
MCC15						-	-37.70**	-49.75**
MCC20							-	-12.05
MCC25								-

* p<.05 , ** p< .01

การเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ยจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ จากวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่ใช้กับคลังข้อสอบ 800 ข้อ ซึ่งมีกลุ่มตัวอย่างทั้งหมด 8 กลุ่ม จะเห็นว่า กลุ่มที่ให้ค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ที่มีค่าต่ำที่สุดหรือมีประสิทธิภาพด้านการใช้คลังข้อสอบสูงสุดและรองลงมาใน 3 อันดับแรก ได้แก่

อันดับที่ 1 ได้แก่ กลุ่ม MCC10 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้มีค่าต่ำกว่ากลุ่ม CWA20, CWA25, MCC20 และ MCC25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01

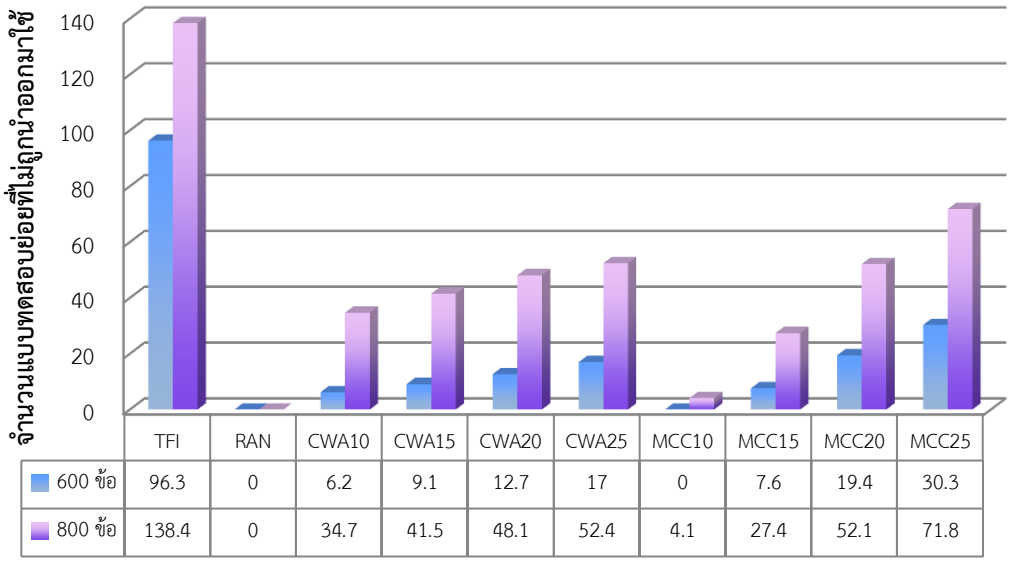
อันดับที่ 2 ได้แก่ กลุ่ม MCC15 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ มีค่าต่ำกว่ากลุ่ม CWA25, MCC20 และ MCC25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และต่ำกว่ากลุ่ม CWA20 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

อันดับที่ 3 ได้แก่ กลุ่ม CWA10 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ มีค่าต่ำกว่ากลุ่ม MCC25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และต่ำกว่ากลุ่ม CWA25 และ MCC20 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 รายละเอียดดังตารางที่ 4.24

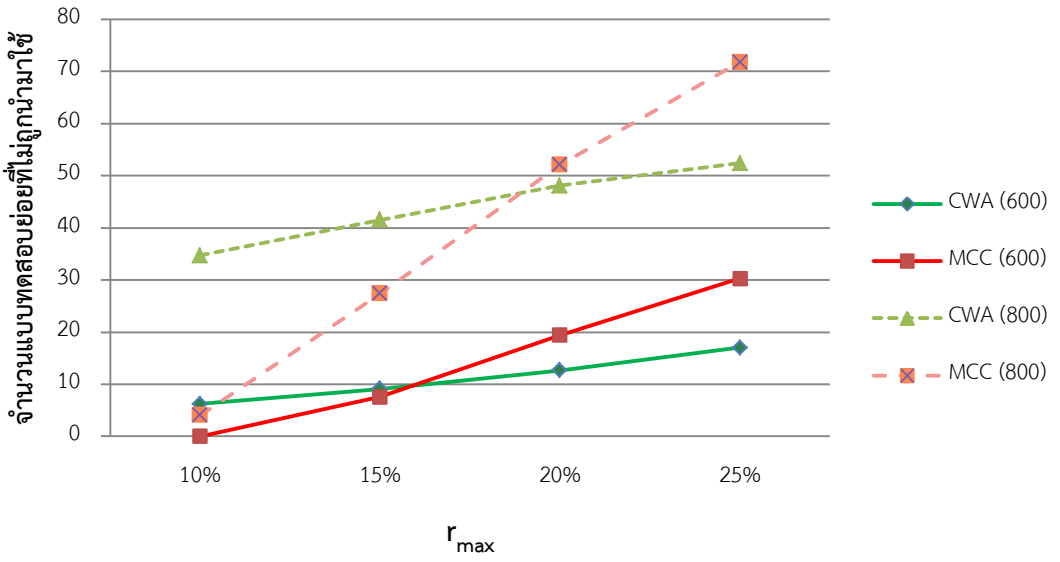
ตารางที่ 4.24 ผลการเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ยจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ ระหว่างวิธี MCC กับวิธี CWA ที่มีการควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุดแตกต่างกันสำหรับคลังข้อสอบ 800 ข้อ

กลุ่ม	CWA10	CWA15	CWA20	CWA25	MCC10	MCC15	MCC20	MCC25
$M.N_{neverexposed}$	34.70	41.50	48.10	52.40	4.10	27.40	52.10	71.80
CWA10	-	-11.85	-25.90	-34.60*	18.80	7.60	-32.75*	-50.90**
CWA15		-	-14.05	-22.75	30.65	19.45	-20.90	-39.05**
CWA20			-	-8.70	44.70**	33.50*	-6.85	-25.00
CWA25				-	53.40**	42.20**	1.85	-16.30
MCC10					-	-11.20	-51.55**	-69.70**
MCC15						-	-40.35**	-58.50**
MCC20							-	-18.15
MCC25								-

* p<.05 , ** p< .01



ภาพที่ 4.13 กราฟค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ ระหว่างวิธี MCC กับวิธี CWA ที่มีการควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุดแตกต่างกัน



ภาพที่ 4.14 กราฟแสดงความสัมพันธ์ระหว่างเฉลี่ยของจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ และ r_{max} ระหว่างวิธี MCC กับวิธี CWA ที่มีการควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุดแตกต่างกัน

จากภาพที่ 4.14 และ 4.15 แสดงกราฟเปรียบเทียบค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ ($N_{neverexposed}$) จากการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่มีวิธีการคัดเลือกข้อสอบแตกต่างกัน ซึ่งค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ที่เกิดขึ้นจะบอกถึง จำนวนของแบบทดสอบย่อยที่ไม่ถูกนำมาใช้กับผู้สอบคนใดเลยตลอดการทดสอบ นั่นหมายความว่าแบบทดสอบย่อยชุดนั้นไม่ได้ถูกนำมาใช้ประโยชน์ ถ้าวิธีการคัดเลือกแบบทดสอบย่อยวิธีใดมีค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้น้อยกว่า แสดงว่าสามารถใช้ประโยชน์จากคลังข้อสอบได้มีประสิทธิภาพดีกว่า

จากแนวโน้มค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ ($N_{neverexposed}$) ที่เปลี่ยนไปเมื่ออัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max}) เพิ่มขึ้น เมื่อพิจารณาจะพบว่า ถ้าใช้วิธีการคัดเลือกแบบทดสอบย่อยแบบ MCC กับคลังข้อสอบที่มีขนาด 600 ข้อ และ 800 ข้อ ค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ ($N_{neverexposed}$) ของวิธี MCC มีค่าต่ำกว่า CWA เมื่อกำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max}) ไว้ที่ 10 และ 15 เปอร์เซ็นต์ แสดงว่าวิธี MCC สามารถใช้ประโยชน์จากคลังข้อสอบได้มีประสิทธิภาพดีกว่าวิธี CWA

ในกรณีที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max}) ไว้ที่ 20 และ 25 เปอร์เซ็นต์ ค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ ($N_{neverexposed}$) ของวิธี MCC มีค่าสูงกว่าวิธี CWA แสดงว่าวิธี CWA สามารถใช้ประโยชน์จากคลังข้อสอบได้มีประสิทธิภาพดีกว่าวิธี MCC

3.2.5 ผลการเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ยจำนวนแบบทดสอบย่อยที่ใช้มากเกินไป

การเปรียบเทียบค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ถูกนำมาใช้ถี่เกินไปจากวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่ใช้กับคลังข้อสอบ 600 ข้อ ซึ่งมีกลุ่มตัวอย่างทั้งหมด 8 กลุ่ม จะเห็นว่า กลุ่มที่ให้ค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ถูกนำมาใช้ถี่เกินไปที่มีค่าสูงที่สุดหรือมีประสิทธิภาพด้านการใช้คลังข้อสอบต่ำที่สุดและรองลงมาใน 2 อันดับแรก ได้แก่

อันดับที่ 1 ได้แก่ กลุ่ม CWA10 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ถูกนำมาใช้ถี่เกินไปมีค่าสูงกว่าทุกกลุ่มยกเว้น CWA15 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01

อันดับที่ 2 ได้แก่ กลุ่ม CWA15 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ถูกนำมาใช้ถี่เกินไปมีค่าสูงกว่าทุกกลุ่มยกเว้น CWA10 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 รายละเอียดดังตารางที่ 4.25

ตารางที่ 4.25 ผลการเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ยจำนวนแบบทดสอบย่อยที่ถูกนำมาใช้มากเกินไป ระหว่างวิธี MCC กับวิธี CWA ที่มีการควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุดแตกต่างกัน สำหรับคลังข้อสอบ 600 ข้อ

กลุ่ม	CWA10	CWA15	CWA20	CWA25	MCC10	MCC15	MCC20	MCC25
$M.N_{overexposed}$	24.80	6.60	0.00	0.00	0.30	0.00	0.00	0.00
CWA10	-	10	46.5**	36.5**	37.5**	46.5**	46.5**	46.5**
CWA15		-	36.5**	46.5**	27.5*	36.5**	36.5**	36.5**
CWA20			-	0	-9	0	0	0
CWA25				-	-9	0	0	0
MCC10					-	9	9	9
MCC15						-	0	0
MCC20							-	0
MCC25								-

* p<.05 , ** p< .01

การเปรียบเทียบค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ถูกนำมาใช้มากเกินไปจากวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่ใช้กับคลังข้อสอบ 800 ข้อ ซึ่งมีกลุ่มตัวอย่างทั้งหมด 8 กลุ่ม จะเห็นว่า กลุ่มที่ให้ค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ถูกนำมาใช้ถี่เกินไปที่มีค่าสูงที่สุดหรือมีประสิทธิภาพด้านการใช้คลังข้อสอบต่ำที่สุดและรองลงมาใน 2 อันดับแรก ได้แก่

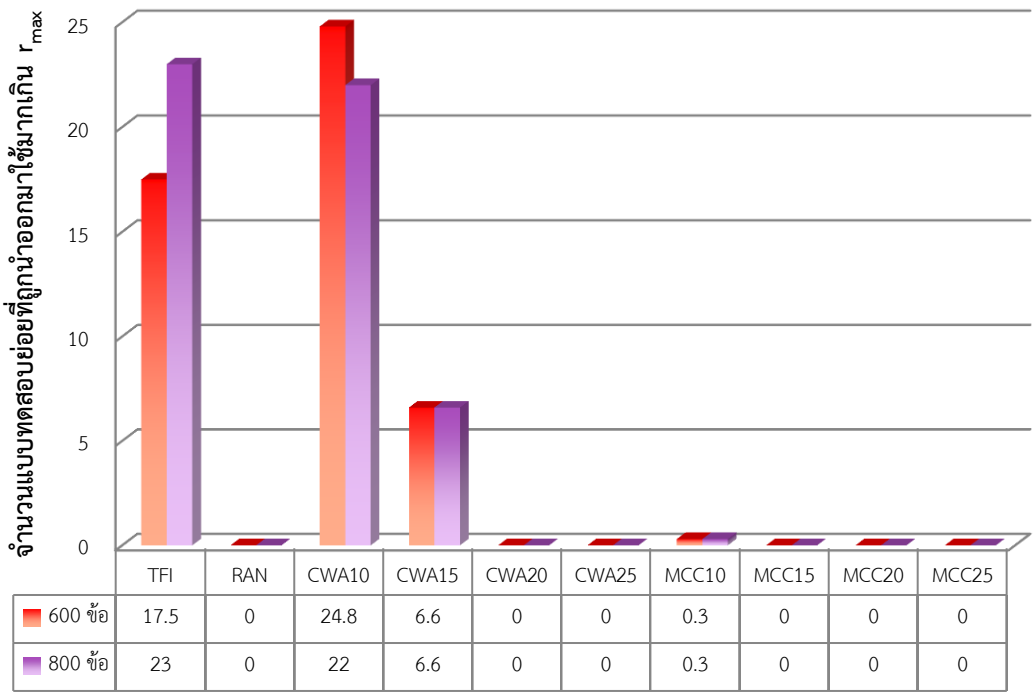
อันดับที่ 1 ได้แก่ กลุ่ม CWA10 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ถูกนำมาใช้ถี่เกินไปมีค่าสูงกว่าทุกกลุ่มยกเว้น CWA15 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01

อันดับที่ 2 ได้แก่ กลุ่ม CWA15 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ถูกนำมาใช้ถี่เกินไปมีค่าสูงกว่าทุกกลุ่มยกเว้น CWA10 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 รายละเอียดดังตารางที่ 4.26

ตารางที่ 4.26 ผลการเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ของค่าเฉลี่ยจำนวนแบบทดสอบย่อยที่ถูกนำมาใช้มากเกินไป ระหว่างวิธี MCC กับวิธี CWA ที่มีการควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุดแตกต่างกัน สำหรับคลังข้อสอบ 800 ข้อ

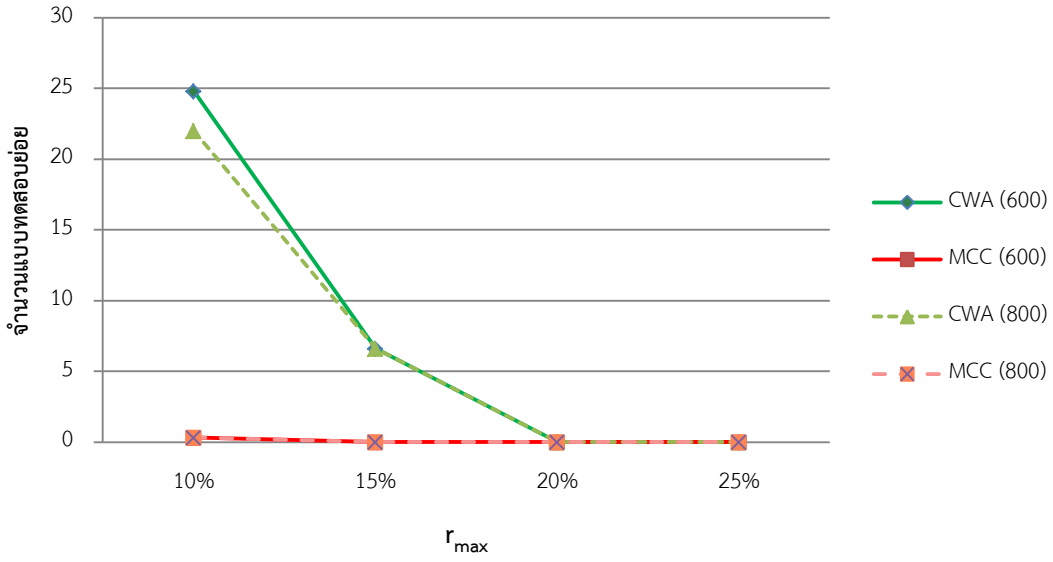
กลุ่ม	CWA10	CWA15	CWA20	CWA25	MCC10	MCC15	MCC20	MCC25
$M.N_{overexpos}$	22.00	6.60	0.00	0.00	0.00	0.00	0.00	0.00
CWA10	-	10	45**	45**	45**	45**	45**	45**
CWA15		-	35**	35**	35**	35**	35**	35**
CWA20			-	0	0	0	0	0
CWA25				-	0	0	0	0
MCC10					-	0	0	0
MCC15						-	0	0
MCC20							-	0
MCC25								-

* p<.05 , ** p< .01



	TFI	RAN	CWA10	CWA15	CWA20	CWA25	MCC10	MCC15	MCC20	MCC25
600 ข้อ	17.5	0	24.8	6.6	0	0	0.3	0	0	0
800 ข้อ	23	0	22	6.6	0	0	0.3	0	0	0

ภาพที่ 4.15 กราฟค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ถูกนำมาใช้มากเกินไป ระหว่างวิธี MCC กับวิธี CWA ที่มีการควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุดแตกต่างกัน



ภาพที่ 4.16 กราฟแสดงความสัมพันธ์ระหว่างเฉลี่ยของจำนวนแบบทดสอบย่อยที่ถูกนำมาใช้มากเกินไป และ r_{max} ระหว่างวิธี MCC กับวิธี CWA ที่มีการควบคุมการใช้แบบทดสอบย่อยซ้ำสูงสุดแตกต่างกัน

จากภาพที่ 4.16 และ 4.17 แสดงกราฟเปรียบเทียบค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ถูกนำมาใช้มากเกินไป ($N_{overexposed}$) จากการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่มีวิธีการคัดเลือกข้อสอบแตกต่างกัน ซึ่งค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ถูกนำมาใช้มากเกินไปจะบอกถึงจำนวนของแบบทดสอบย่อยที่ถูกนำมาใช้กับผู้สอบบ่อยครั้งจนกระทั่งมีความถี่สูงกว่าที่อัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max}) กำหนด นั่นหมายความว่าแบบทดสอบย่อยชุดนั้นถูกนำมาใช้บ่อยเกินไปอาจจะส่งผลต่ออาจจะส่งผลต่อความปลอดภัยของคลังข้อสอบเนื่องจากผู้สอบอาจจะจำแบบทดสอบย่อยชุดนั้นออกมาเผยแพร่ ถ้าวิธีการคัดเลือกแบบทดสอบย่อยวิธีใดมีค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ถูกนำมาใช้มากเกินไปต่ำกว่าอีกวิธี แสดงว่าวิธีการคัดเลือกแบบทดสอบย่อยวิธีนั้นสามารถควบคุมการใช้แบบทดสอบย่อยซ้ำได้ดีกว่า

จากแนวโน้มค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ถูกนำมาใช้มากเกินไป ($N_{overexposed}$) ที่เปลี่ยนไปเมื่ออัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max}) เพิ่มขึ้น เมื่อพิจารณาจะพบว่า ถ้าใช้วิธีการคัดเลือกแบบทดสอบย่อยแบบ MCC กับคลังข้อสอบที่มีขนาด 600 ข้อ และ 800 ข้อ ค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ถูกนำมาใช้มากเกินไป ($N_{overexposed}$) ของวิธี MCC มีค่าต่ำกว่า CWA เมื่อกำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max}) ไว้ที่ 10 และ 15 เปอร์เซ็นต์ แสดงว่าวิธี MCC สามารถควบคุมการใช้แบบทดสอบย่อยซ้ำได้มีประสิทธิภาพดีกว่าวิธี CWA เมื่อกำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max}) ไว้ที่ 10 และ 15 เปอร์เซ็นต์

ในกรณีที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max}) ไว้ที่ 20 และ 25 เปอร์เซ็นต์ ค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ถูกนำมาใช้มากเกินไป ($N_{overexposed}$) ของวิธี MCC มีค่าเท่ากันกับวิธี CWA แสดงว่าวิธีทั้งสองวิธีสามารถควบคุมการใช้แบบทดสอบย่อยซ้ำได้มีประสิทธิภาพดีพอๆ กัน เมื่อกำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max}) ไว้ที่ 20 และ 25 เปอร์เซ็นต์

สรุปผลการเปรียบเทียบประสิทธิภาพของวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ จากเกณฑ์ที่ใช้พิจารณาประสิทธิภาพ พบว่า วิธีมอนติ คาร์โล ซีเอที (MCC) มีประสิทธิภาพดีกว่าวิธีแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (CWA) ถึง 64 เซลล์ ของเกณฑ์ที่ใช้พิจารณาประสิทธิภาพจากเงื่อนไขการทดลองทั้งหมดที่มี 80 เซลล์ ขณะที่วิธี CWA มีประสิทธิภาพดีกว่าวิธี MCC 10 เซลล์ เมื่อพิจารณาในรายละเอียดจะพบว่า วิธี MCC มีข้อได้เปรียบในด้านประสิทธิภาพของความถูกต้องแม่นยำในการประมาณค่าความสามารถของผู้สอบและมีความแกร่งต่อการกำหนดอัตราการใช้แบบทดสอบย่อยซ้ำในระดับที่ต่ำ (10% และ 15%) ขณะที่วิธี CWA สามารถใช้แบบทดสอบย่อยในคลังข้อสอบได้อย่างกระจายใกล้เคียงการแจกแจงแบบเอกรูปมากกว่า

วิธี MCC เมื่อกำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดตั้งแต่ 20 เปอร์เซ็นต์ขึ้นไป ดังพิจารณาได้จากค่าอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ (r_{obs}) ค่าการเปรียบเทียบการแจกแจงของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ (ไคสแควร์) และจำนวนของแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ ($N_{neverexposured}$) ที่มีค่าต่ำกว่าค่าของวิธี MCC รายละเอียดดังตารางที่ 4.27

ตารางที่ 4.27 สรุปผลการเปรียบเทียบประสิทธิภาพของวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์

DV	item bank	r_{max}			
		10%	15%	20%	25%
Test length	600	MCC	MCC	MCC	MCC
	800	MCC	MCC	MCC	MCC
SEE	600	MCC	MCC	MCC	MCC
	800	MCC	MCC	MCC	MCC
$r_{\theta\hat{\theta}}$	600	MCC	MCC	MCC	MCC
	800	MCC	MCC	MCC	MCC
Bias	600	CWA	MCC	MCC	CWA
	800	MCC	MCC	=	MCC
MSE	600	=	MCC	MCC	MCC
	800	MCC	MCC	MCC	MCC
r_{obs}	600	MCC	MCC	MCC	CWA
	800	MCC	MCC	MCC	CWA
test overlap rate	600	MCC	MCC	MCC	MCC
	800	MCC	MCC	MCC	MCC
x^2	600	MCC	MCC	MCC	CWA
	800	MCC	MCC	MCC	CWA
$N_{neverexposured}$	600	MCC	MCC	CWA	CWA
	800	MCC	MCC	CWA	CWA
$N_{overexposured}$	600	MCC	MCC	=	=
	800	MCC	MCC	=	=

หมายเหตุ = หมายถึง ผลการวิเคราะห์ของ CWA และ MCC มีค่าเท่ากัน

บทที่ 5
สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถของผู้สอบและประสิทธิภาพด้านการใช้คลังข้อสอบของการทดสอบปรับเหมาะแบบด้วยคอมพิวเตอร์ที่ใช้สำหรับโมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย 2 วิธี คือ 1) วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที (Monte Carlo CAT Method) และ 2) วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (constraint-weighted a-stratification CAT method) เมื่อกำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด และขนาดคลังข้อสอบที่ต่างกัน

ตัวแปรที่ศึกษา ประกอบด้วย ตัวแปรอิสระ ได้แก่ 1. วิธีการเลือกแบบทดสอบย่อยมี 2 วิธี คือ 1) วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที และ 2) วิธีการแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ 2 ขนาดคลังข้อสอบ มี 2 ขนาด คือ 1) คลังข้อสอบขนาด 600 ข้อ และ2) คลังข้อสอบขนาด 800 ข้อ และ 3 อัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดมี 4 ระดับ คือ 10, 15, 20 และ 25 เปอร์เซ็นต์ ตัวแปรตาม ได้แก่ 1. ประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถของผู้สอบ พิจารณาได้จากเกณฑ์ดังนี้ 1) สหสัมพันธ์ระหว่างค่าประมาณความสามารถกับค่าความสามารถจริง 2) ค่าความลำเอียง 3) ค่าความแปรปรวนของความคลาดเคลื่อน และ 4) ความยาวเฉลี่ยของแบบสอบ (เมื่อมีค่าความคลาดเคลื่อนมาตรฐานของการประมาณค่าใกล้เคียงกัน) และ 2. ประสิทธิภาพด้านสมดุลการใช้แบบทดสอบย่อยในคลังข้อสอบ พิจารณาได้จากเกณฑ์ดังนี้ 1) อัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ 2) อัตราการทับซ้อนของแบบสอบ 3) ความสมดุลของการใช้คลังข้อสอบในภาพรวม 4) จำนวนข้อสอบที่ไม่ถูกนำมาใช้ และ 5) จำนวนข้อสอบที่ถูกนำมาใช้มากเกินไป

สมมติฐานในการวิจัยมีดังนี้ 1) วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที น่าจะมีประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถสูงกว่าวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ และ 2) วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที น่าจะมีประสิทธิภาพด้านการใช้คลังข้อสอบสูงกว่าวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ

เมื่อประสิทธิภาพในการประมาณค่าความสามารถของผู้สอบสูง หมายถึง 1) มีความตรงในการประมาณค่าสูง 2) มีความลำเอียงต่ำ 3) มีความแปรปรวนของความคลาดเคลื่อนต่ำ และ 4) มีความยาวเฉลี่ยของแบบสอบต่ำ

เมื่อประสิทธิภาพในสมดุลการใช้คลังข้อสอบสูง หมายถึง 1) มีอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ต่ำกว่าเกณฑ์ที่กำหนด 2) มีอัตราการทับซ้อนของแบบสอบต่ำ 3) มีความสมดุลของการใช้คลังข้อสอบในภาพรวมสูง (มีค่า x^2 ต่ำ) 4) มีจำนวนข้อสอบที่ไม่ถูกนำมาใช้ต่ำ และ 5) มีจำนวนข้อสอบที่ถูกนำมาใช้มากเกินไปต่ำ

ตัวอย่างที่ใช้ในการศึกษาเป็นค่าความสามารถจริงของ ผู้สอบที่สร้างขึ้นโดยสุ่มการแจกแจงโค้งปกติมาตรฐาน ซึ่งมีค่าเฉลี่ยเท่ากับ 0 และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 1 โดยกำหนดให้มีค่าอยู่ในช่วงประมาณ -3.5 ถึง 3.5 ค่าความสามารถจริงของตัวอย่างผู้สอบที่ใช้ในทดลองแต่ละรอบมีจำนวนเท่ากับ 1,000 ค่า และเมื่อทำการทดลองรอบใหม่ผู้วิจัยทำการสุ่มค่าความสามารถจริงของตัวอย่างผู้สอบใหม่ทุกครั้ง โดยในการทดลองแต่ละเงื่อนไขจะสุ่มตัวอย่างเพื่อนำมาทำซ้ำ 10 รอบ นั่นหมายความว่า ใช้ตัวอย่างที่เป็นค่าความสามารถของผู้สอบทั้งหมด 10,000 ค่า ต่อการทดลอง 1 เงื่อนไข

เครื่องมือใช้ในการวิจัย การวิจัยครั้งนี้เป็นการศึกษาในสถานการณ์จำลอง ผู้วิจัยใช้โปรแกรม R (R Programming language) ซึ่งเป็นภาษาคอมพิวเตอร์ภาษาหนึ่งที่พัฒนาโดยใช้ฐานของภาษา S ซึ่งเป็นซอฟต์แวร์ทางสถิติที่มีราคาสูง โปรแกรม R พัฒนาขึ้นโดย Robert Gentleman และ Ross Ihaka ที่คณะสถิติ มหาวิทยาลัยโอ๊คแลนด์ ประเทศนิวซีแลนด์ และตั้งแต่กลางปี 1997 เป็นต้นมาโปรแกรม R ได้รับการพัฒนาอย่างต่อเนื่องจากกลุ่มนักวิชาการที่ชื่อว่า “R Core Team” จนกระทั่งโปรแกรม R เป็นที่นิยมในวงวิชาการเนื่องจากมีฟังก์ชั่นภายใน (built-in function) จำนวนมาก รวมทั้งมีการเผยแพร่แบบ General Public License ตั้งแต่ปี 1995 ดังนั้นโปรแกรม R จึงเป็นฟรีซอฟต์แวร์ที่เปิดเผยโค้ดของตัวโปรแกรม (open source) เพื่อให้นักวิจัยทั่วโลกได้นำไปใช้และพัฒนาต่อยอด ดังนั้นผู้วิจัยจึงนำโปรแกรม R มาใช้เป็นเครื่องมือในการจำลองข้อมูล จัดกระทำข้อมูล วิเคราะห์ข้อมูล และแสดงข้อมูลในรูปแบบกราฟ โดยผ่านขั้นตอนต่างๆ ดังนี้คือ ศึกษาความมุ่งหมายการวิจัยและรวบรวมข้อมูลที่เกี่ยวข้องเพื่อออกแบบและทดสอบวิธีการทดสอบแบบปรับเหมาะแบบมอนติ คาร์โล ซีเอที และแบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ ที่ผู้วิจัยเขียนขึ้นด้วยภาษา R จากนั้นจึงปรับปรุงโปรแกรมเพื่อให้ทำงานตามลำดับขั้นตอนการทดสอบได้อย่างถูกต้อง โดยการนำโปรแกรมไปทดสอบกับข้อมูลที่รู้ค่าผลลัพธ์และทดสอบซ้ำจนแน่ใจว่าการทำงานของโปรแกรมถูกต้องและผลการทดสอบแม่นยำเพียงพอจึงนำไปใช้ทดสอบ โดยโปรแกรมที่พัฒนาขึ้นประกอบด้วย 4 ส่วน คือ 1) ชุดคำสั่งสำหรับสร้างพารามิเตอร์ผู้สอบ 2) ชุดคำสั่งสำหรับ สร้าง

พารามิเตอร์ข้อสอบ 3) ชุดคำสั่งสำหรับวิธีการทดสอบแบบปรับเหมาะแบบมอนติ คาร์โล ซีเอที และ 4) ชุดคำสั่งสำหรับวิธีการทดสอบแบบปรับเหมาะแบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ

รูปแบบการทดลอง การวิจัยครั้งนี้ได้ออกแบบการวิจัยแบบคอมพลีทลี แรมดอมไมซ์ แฟคทอเรียล ดีไซน์ (Completely Randomized Factorial Design: CRF) ซึ่งมีเงื่อนไขการทดลองทั้งหมด 20 เงื่อนไข โดยทำการสุ่มตัวอย่างเข้าไปในแต่ละเงื่อนไขการทดลอง กลุ่มละ 1000 คน และทำการทดลอง ซ้ำ 10 ครั้ง และจะใช้ผลของการทดลองซ้ำแต่ละครั้งเป็นหน่วยในการวิเคราะห์เปรียบเทียบ

การวิจัยครั้งนี้มีลำดับขั้นตอนในการทดลอง ดังนี้

1. จำลองตัวอย่าง ความสามารถจริง (true ability) และอิทธิพลที่มีต่อแบบทดสอบย่อย (testlet effect) ของตัวอย่าง การสร้างค่าความสามารถจริง (θ_i) ของตัวอย่างทำโดยสุ่มจากค่าที่มีการแจกแจงโค้งปกติมาตรฐาน ซึ่งมีค่าเฉลี่ยเท่ากับ 0 และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 1 และกำหนดให้มีค่าอยู่ในช่วง -3.5 ถึง 3.5 ค่าความสามารถจริงของกลุ่มตัวอย่างผู้สอบสำหรับการทดสอบแต่ละวิธีกำหนดมีจำนวนเท่ากับ 1,000 ค่า เมื่อเปลี่ยนวิธีกำหนดจะทำการสุ่มค่าความสามารถจริงของกลุ่มตัวอย่างผู้สอบใหม่ทุกครั้ง โดยในการทดลองแต่ละเงื่อนไขจะสุ่มตัวอย่างเพื่อนำมาทำซ้ำ 10 รอบ ดังนั้น ตัวอย่างที่ใช้ในการทดลอง แต่ละเงื่อนไขจะเป็น 10,000 ค่า และกำหนดพารามิเตอร์อิทธิพลของแบบทดสอบย่อยให้กับผู้สอบโดยเลือกอย่างสุ่มจากการแจกแจงปกติด้วยค่าเฉลี่ยเท่ากับศูนย์ และความแปรปรวนเท่ากับ $\gamma_{id(j)}$

2. จำลองคลังข้อสอบ และโครงสร้างของคลังข้อสอบ โดยการสร้างพารามิเตอร์ของข้อสอบจากการกำหนดค่าเฉลี่ย และความแปรปรวน ซึ่งพารามิเตอร์แต่ละตัวมีการกำหนดรายละเอียดดังนี้ พารามิเตอร์อำนาจจำแนก (a_j) ถูกสร้างขึ้นจากการแจกแจงแบบ Log-Normal [a ~LN(0.02,0.22^2)] เนื่องจากพารามิเตอร์อำนาจจำแนกมีค่าเป็นบวกและมากกว่า 0.8 พารามิเตอร์ความยาก (b_j) สร้างขึ้นจากการแจกแจงปกติมาตรฐาน [b ~N(0,1)] เนื่องจากข้อสอบที่มีความยากหรือง่ายมากๆ จะไม่ค่อยพบในแบบสอบ พารามิเตอร์การเดา (c_j) สร้างขึ้นจากการแจกแจงแบบ Beta [c ~BE(2,10)] เนื่องจากพารามิเตอร์การเดามีค่าเป็นบวกและค่าจะมีขนาดไม่เกิน 0.2

3. หาค่าความน่าจะเป็น (probability) ในการตอบข้อสอบถูกของคลังข้อสอบในข้อที่ 1กับกลุ่มตัวอย่างในข้อที่ 2 โดยใช้สูตร แบบ 3-พารามิเตอร์ Testlet Response Model

4. หาผลการตอบของผู้สอบ (U_{ij}) โดยนำค่าความน่าจะเป็นในการตอบข้อสอบถูก (P_{ij}) ที่ได้ในข้อ 3 มาหาผลการตอบข้อสอบ ตอบถูก=1 ตอบผิด=0 ของผู้สอบแต่ละคนในข้อสอบแต่ละข้อ โดยการเรียกเลขสุ่ม (X_{ij}) จากเครื่องคอมพิวเตอร์ที่มีการแจกแจงแบบยูนิฟอร์มมีค่าตั้งแต่ 0 ถึง 1

แล้วนำค่า X_{ij} ไปเปรียบเทียบกับ P_{ij} ถ้า X_{ij} <= P_{ij} แสดงว่าตกอยู่ในพื้นที่ยอมรับว่าตอบถูก จึงให้ U_{ij} =1 ถ้า X_{ij} > P_{ij} ก็จะปฏิเสธการตอบถูก นั่นคือ ตอบผิด ก็ทำให้ U_{ij} = 0 ทำอย่างนี้กับข้อสอบทุกข้อผู้สอบทุกคน ผู้วิจัยจึงได้ผลการตอบ (0, 1) ไปขั้นในการวิจัยขั้นต่อไป

5. นำผลการตอบในข้อที่ 4 มาใช้หาเส้นทางการตอบข้อสอบในการทดสอบแบบปรับเหมาะที่พัฒนาขึ้น และทำการทดลองซ้ำเงื่อนไขละ 10 รอบ ในข้อที่ 1, 3, 4 และ 5

การวิเคราะห์ข้อมูล หาค่าสถิติเบื้องต้น เปรียบเทียบความแตกต่างของค่าเฉลี่ยที่ได้ โดยการวิเคราะห์สถิติแบบนอนพาราเมตริก (Nonparametric statistics) สำหรับกลุ่มตัวอย่าง k กลุ่มที่มีความเป็นอิสระต่อกัน โดยใช้สถิติ Kruskal-Wallis Test ถ้าผลการวิเคราะห์มีนัยสำคัญทางสถิติจึงเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ต่อไป

สรุปผลการวิจัย

จากผลการวิเคราะห์ข้อมูล สามารถสรุปผลการวิจัยโดยจำแนกตามวัตถุประสงค์ได้ดังนี้

1. ผลจากการศึกษาเปรียบเทียบ ประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถของผู้สอบจากการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที กับการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ ที่มีขนาดคลังข้อสอบ และ อัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดที่แตกต่างกัน พบว่า

1.1 ผลการเปรียบเทียบค่าเฉลี่ยของอันดับของค่าเฉลี่ยความลำเอียง (Bias) จากวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ 2 วิธี ระหว่างวิธี CWA กับวิธี MCC ที่ใช้กับคลังข้อสอบ 600 ข้อ และ 800 ข้อ กับทุกระดับของอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด พบว่า ไม่แตกต่างกัน

1.2 การเปรียบเทียบค่าเฉลี่ยของอันดับของค่าเฉลี่ย MSE จากวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ 2 วิธี ระหว่างวิธี CWA กับวิธี MCC ที่ใช้กับคลังข้อสอบ 600 ข้อ และ 800 ข้อ พบว่าวิธี MCC มีค่าเฉลี่ยของ MSE ต่ำกว่าวิธี CWA ในทุกระดับที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (10%, 15%, 20% และ 25%) โดยวิธี MCC ใช้กับคลังข้อสอบขนาด 600 ข้อ มีค่า MSE ต่ำที่สุด 3 อับดับแรก คือ วิธี MCC ที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดไว้เท่ากับ 20%, 25% และ 15% ตามลำดับและวิธี MCC ใช้กับคลังข้อสอบขนาด 800 ข้อ ที่มีค่า MSE ต่ำที่สุด 3 อับดับแรก คือ วิธี MCC ที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดไว้เท่ากับ 10%, 20% และ 25% ตามลำดับ นอกจากนี้ ผลการวิเคราะห์บ่งชี้ว่า เมื่อคลังข้อสอบมีขนาดใหญ่ขึ้น ค่า MSE จะลดลง

เมื่อพิจารณาจากแนวโน้มค่าเฉลี่ยของ MSE ที่เปลี่ยนไปเมื่ออัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) เพิ่มขึ้น เมื่อพิจารณาจะพบว่า เมื่อคลังข้อสอบที่มีขนาด 600 ข้อ และ 800 ข้อ เฉลี่ยของ MSE ของวิธี MCC มีค่าสูงกว่าวิธี CWA ในทุกระดับของอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) แสดงว่า วิธี MCC มีประสิทธิภาพในการประมาณค่าความสามารถของผู้สอบสูงกว่าวิธี CWA ในทุกเงื่อนไขการทดลอง

1.3 การเปรียบเทียบค่าเฉลี่ยของ $r_{\theta\hat{\theta}}$ จากวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ 2 วิธี ระหว่างวิธี CWA กับวิธี MCC ที่ใช้กับคลังข้อสอบ 600 ข้อ และ 800 ข้อ พบว่าวิธี MCC มีค่าเฉลี่ยของ $r_{\theta\hat{\theta}}$ สูงกว่าวิธี CWA ในทุกระดับที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (10%, 15%, 20% และ 25%) ซึ่งวิธี MCC ที่ใช้กับคลังข้อสอบขนาด 600 ข้อ ที่มีค่า $r_{\theta\hat{\theta}}$ สูงที่สุดและมีค่าแตกต่างจากวิธีอื่นอย่างมีนัยสำคัญทางสถิติ คือ วิธี MCC ที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดไว้เท่ากับ 25% โดยมีค่า $r_{\theta\hat{\theta}}$ มากกว่าวิธี CWA ที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดไว้เท่ากับ 10% อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.1 และวิธี MCC ที่ใช้กับคลังข้อสอบขนาด 800 ข้อ ที่มีค่า $r_{\theta\hat{\theta}}$ สูงที่สุด และมีค่าแตกต่างจากวิธีอื่นอย่างมีนัยสำคัญทางสถิติ คือ วิธี MCC ที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดไว้เท่ากับ 10% และ 20% มีค่า $r_{\theta\hat{\theta}}$ มากกว่าวิธี CWA ที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดไว้เท่ากับ 10%, 15% และ 25% อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.1 ส่วนวิธี MCC ที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดไว้เท่ากับ 25% โดยมีค่า $r_{\theta\hat{\theta}}$ มากกว่าวิธี CWA ที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดไว้เท่ากับ 10% อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.5 นอกจากนี้ ผลการวิเคราะห์ยังบ่งชี้ว่า เมื่อคลังข้อสอบมีขนาดใหญ่ขึ้น ค่า $r_{\theta\hat{\theta}}$ จะเพิ่มขึ้น และเพิ่มขึ้นมากที่สุดเท่ากับ 0.006 เมื่อที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดเท่ากับ 10%

เมื่อพิจารณาจากแนวโน้มค่าเฉลี่ยของ $r_{\theta\hat{\theta}}$ ที่เปลี่ยนไปเมื่ออัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) เพิ่มขึ้น เมื่อพิจารณาจะพบว่า เมื่อคลังข้อสอบที่มีขนาด 600 ข้อ และ 800 ข้อ เฉลี่ยของ $r_{\theta\hat{\theta}}$ ของวิธี MCC มีค่าสูงกว่าวิธี CWA ในทุกระดับของอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) แสดงว่า วิธี MCC มีประสิทธิภาพในการประมาณค่าความสามารถของผู้สอบสูงกว่าวิธี CWA ในทุกเงื่อนไขการทดลอง

1.4 ผลการเปรียบเทียบความยาวของแบบสอบเฉลี่ยจากวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ 2 วิธี ระหว่างวิธี CWA กับวิธี MCC ที่ใช้กับคลังข้อสอบ 600 ข้อ และ 800 ข้อ พบว่า วิธีที่มีความยาวของแบบสอบเฉลี่ยต่ำที่สุดหรือมีประสิทธิภาพในการประมาณค่าความสามารถของผู้สอบสูงสุดและรองลงมาใน 3 อันดับแรก ได้แก่ วิธี MCC ที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดไว้เท่ากับ 25%, 20% และ 15% ซึ่งวิธี MCC ที่กำหนดอัตรา

การใช้แบบทดสอบย่อยซ้ำสูงสุดไว้ทั้ง 3 ระดับนั้น มีความยาวของเฉลี่ยแบบสอบต่ำกว่า วิธี CWA ที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดไว้เท่ากับ 10% และ 15% อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และต่ำกว่า วิธี CWA ที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดไว้เท่ากับ 20% อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นั่นหมายความว่า เมื่อกำหนดเกณฑ์การยุติการทดสอบโดยให้ค่าความคลาดเคลื่อนมาตรฐานในการประมาณค่าความสามารถของผู้สอบ (SEE) น้อยกว่าหรือเท่ากับ 0.3 วิธี MCC ใช้จำนวนข้อสอบน้อยกว่าวิธี CWA

เมื่อพิจารณาจากแนวโน้มความยาวของแบบสอบเฉลี่ยที่เปลี่ยนไปเมื่ออัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max}) เพิ่มขึ้น เมื่อพิจารณาจะพบว่า เมื่อคลังข้อสอบที่มีขนาด 600 ข้อ และ 800 ข้อ ความยาวของแบบสอบเฉลี่ย ของวิธี MCC มีค่าต่ำกว่าวิธีการคัดเลือกแบบทดสอบย่อยแบบ CWA ในทุกระดับของอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max}) แสดงว่า วิธี MCC มีประสิทธิภาพในการประมาณค่าความสามารถของผู้สอบสูงกว่าวิธี CWA ในทุกเงื่อนไขการทดลอง

2. ผลจากการศึกษาเปรียบเทียบ ประสิทธิภาพด้านการใช้คลังข้อสอบจากการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที กับการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ ที่มีขนาดคลังข้อสอบ และอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดที่แตกต่างกัน พบว่า

2.1 การเปรียบเทียบค่าเฉลี่ยของอัตราการทับซ้อนของแบบสอบ ($\bar{t}$) จากวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ 2 วิธี ระหว่างวิธี CWA กับวิธี MCC ที่ใช้กับคลังข้อสอบ 600 ข้อ และ 800 ข้อ พบว่าวิธี MCC มีค่าเฉลี่ยของอัตราการทับซ้อนของแบบสอบ ($\bar{t}$) ต่ำกว่าวิธี CWA ในทุกระดับที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (10%, 15%, 20% และ 25%) และเมื่อเปรียบเทียบค่าเฉลี่ยของอันดับเป็นรายคู่ พบว่า วิธีที่มีประสิทธิภาพด้านการใช้คลังข้อสอบสูงสุดคือ วิธีที่มีค่าเฉลี่ยของอัตราการทับซ้อนของแบบสอบ ($\bar{t}$) ต่ำที่สุด ใน 3 อันดับแรก คือ อันดับแรก คือ วิธี MCC10 ที่ใช้กับคลังข้อสอบทั้งขนาด 600 ข้อ และ 800 ข้อ มีค่าเฉลี่ยของ $\bar{t}$ ต่ำกว่าวิธี CWA ที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดทุกระดับ อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.01 อันดับที่ 2 คือ วิธี MCC15 ที่ใช้กับคลังข้อสอบขนาด 600 ข้อ มีค่าเฉลี่ยของ $\bar{t}$ ต่ำกว่าวิธี CWA ที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดทุกระดับ อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.01 แต่เมื่อใช้กับคลังข้อสอบขนาด 800 ข้อ ค่าเฉลี่ยของ $\bar{t}$ ต่ำกว่าวิธี CWA ที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดทุกระดับยกเว้นที่ระดับ 15% อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.01 และอันดับที่สามคือ วิธี MCC20 ที่ใช้กับคลังข้อสอบขนาด 800ข้อ มีค่าเฉลี่ยของ $\bar{t}$ ต่ำกว่าวิธี CWA ที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดทุกระดับยกเว้นที่ระดับ 15% อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.01 และเมื่อนำมาใช้กับคลังข้อสอบขนาด 600 ข้อ วิธี MCC20 มีค่าเฉลี่ยของ $\bar{t}$ ต่ำกว่าวิธี

CWA10 และ CWA25 อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.01 และมีค่าเฉลี่ยของ $\bar{t}$ ต่ำกว่าวิธี CWA20 อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05

เมื่อพิจารณาจากแนวโน้มของอัตราการทับซ้อนของแบบสอบ ($\bar{t}$) ที่เปลี่ยนไปเมื่ออัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) เพิ่มขึ้น เมื่อพิจารณาจะพบว่า เมื่อใช้คลังข้อสอบที่มีขนาด 600 ข้อ และ 800 ข้อ ค่าเฉลี่ยของอัตราการทับซ้อนของแบบสอบ ($\bar{t}$) ของวิธี MCC มีค่าต่ำกว่าวิธี CWA ในทุกระดับของอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) แสดงว่า วิธี MCC สามารถควบคุมการใช้แบบทดสอบย่อยซ้ำได้ดีกว่าวิธี CWA ในทุกเงื่อนไขการทดลอง

2.2 การเปรียบเทียบค่าเฉลี่ยของไคสแควร์ เพื่อพิจารณาความสมดุลของการใช้คลังข้อสอบในภาพรวม จากวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ 2 วิธี ระหว่างวิธี CWA กับวิธี MCC ที่ใช้กับคลังข้อสอบ 600 ข้อ พบว่า วิธีที่ให้ค่าเฉลี่ยของไคสแควร์ต่ำที่สุดหรือมีประสิทธิภาพด้านการใช้คลังข้อสอบสูงสุดและรองลงมาใน 3 อันดับแรก ได้แก่ อันดับที่ 1 ได้แก่ วิธี MCC10 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่น ปรากฏว่ามีค่าเฉลี่ยของไคสแควร์ มีค่าต่ำกว่าวิธี CWA10, CWA25 และ MCC25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และต่ำกว่ากลุ่ม CWA20 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 อันดับที่ 2 ได้แก่ วิธี MCC15 เมื่อทดสอบเปรียบเทียบกับวิธีอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของไคสแควร์ มีค่าต่ำกว่าวิธี CWA25 และ MCC25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และต่ำกว่าวิธี CWA10 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 และอันดับที่ 3 ได้แก่ วิธี CWA15 เมื่อทดสอบเปรียบเทียบกับวิธีอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของไคสแควร์ มีค่าต่ำกว่าวิธี CWA25 และ MCC25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01

เมื่อใช้กับคลังข้อสอบ 800 ข้อ อันดับที่ 1 ได้แก่ วิธี MCC10 เมื่อทดสอบเปรียบเทียบกับวิธีอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของไคสแควร์ต่ำกว่าวิธี CWA10, CWA20, CWA25 และ MCC25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 อันดับที่ 2 ได้แก่ วิธี MCC15 เมื่อทดสอบเปรียบเทียบกับวิธี อื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของไคสแควร์ต่ำกว่าวิธี CWA10, CWA25 และ MCC25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และอันดับที่ 3 ได้แก่ วิธี MCC20 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของไคสแควร์ต่ำกว่าวิธี MCC25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และต่ำกว่าวิธี CWA25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

เมื่อพิจารณาจากแนวโน้มค่าเฉลี่ยของไคสแควร์ที่เปลี่ยนไปเมื่ออัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) เพิ่มขึ้น เมื่อพิจารณาจะพบว่า ถ้าใช้วิธีการคัดเลือกแบบทดสอบย่อยแบบ MCC กับคลังข้อสอบที่มีขนาด 600 ข้อ ค่าไคสแควร์ของวิธี MCC มีค่าต่ำกว่า CWA เมื่อกำหนดอัตราการใช้

แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) ไว้ที่ 10 และ 15 เปอร์เซ็นต์ แสดงว่าอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ของวิธี MCC มีการแจกแจงใกล้เคียงแบบยูนิฟอร์มมากกว่าวิธี CWA

ในกรณีที่ใช้คลังข้อสอบมีขนาด 800 ข้อ และกำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) เท่ากับ 25 เปอร์เซ็นต์ ค่าไคสแควร์ของวิธี MCC (9.986) และวิธี CWA (9.781) มีค่าใกล้เคียงกันมาก แสดงว่าอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ของทั้ง 2 วิธี มีการแจกแจงแบบยูนิฟอร์มที่มีลักษณะคล้ายๆ กัน

2.3 การเปรียบเทียบค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ (r_{obs}) จากวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ 2 วิธี ระหว่างวิธี CWA กับวิธี MCC ที่ใช้กับคลังข้อสอบ 600 ข้อ พบว่า วิธีที่ให้ค่าเฉลี่ยของ r_{obs} ต่ำที่สุดหรือมีประสิทธิภาพด้านการใช้คลังข้อสอบสูงสุดและรองลงมาใน 3 อันดับแรก ได้แก่ กลุ่ม MCC10 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นปรากฏว่าค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้มีค่าต่ำกว่ากลุ่ม CWA10, CWA15 และ MCC25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และต่ำกว่ากลุ่ม CWA25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 อันดับที่ 2 ได้แก่ กลุ่ม MCC15 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นปรากฏว่าค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้มีค่าต่ำกว่ากลุ่ม CWA10, CWA15 และ MCC25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และต่ำกว่ากลุ่ม CWA25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 อันดับที่ 3 ได้แก่ กลุ่ม MCC20 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นปรากฏว่าค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้มีค่าต่ำกว่ากลุ่ม CWA10 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และต่ำกว่ากลุ่ม CWA15 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

เมื่อใช้กับคลังข้อสอบ 800 ข้อ ซึ่งมีกลุ่มตัวอย่างทั้งหมด 8 กลุ่ม จะเห็นว่า กลุ่มที่ให้ค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ต่ำที่สุดหรือมีประสิทธิภาพด้านการใช้คลังข้อสอบสูงสุดและรองลงมาใน 3 อันดับแรก ได้แก่ อันดับที่ 1 กลุ่ม MCC10 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นปรากฏว่ามีค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้มีค่าต่ำกว่ากลุ่ม CWA10, CWA15, CWA25 และ MCC25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 อันดับที่ 2 กลุ่ม MCC15 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นปรากฏว่ามีค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้มีค่าต่ำกว่ากลุ่ม CWA10, CWA15 และ CWA25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 อันดับที่ 3 กลุ่ม MCC20 เมื่อทดสอบเปรียบเทียบกับกลุ่มอื่นปรากฏว่ามีค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้มีค่าต่ำกว่ากลุ่ม CWA10 และ CWA15 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01

เมื่อพิจารณาจากแนวโน้มของค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ (r_{obs}) ที่เปลี่ยนไปเมื่ออัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) เพิ่มขึ้น เมื่อพิจารณาจะพบว่า เมื่อคลัง

ข้อสอบที่มีขนาด 600 ข้อ และ 800 ข้อ ค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ (r_{obs}) ของวิธี MCC มีค่าต่ำกว่าหรือใกล้เคียงกับอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) ทุกระดับ ขณะที่วิธี CWA จะมีค่าเฉลี่ยของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ (r_{obs}) ต่ำกว่าอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) เมื่อกำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) อยู่ที่ระดับ 20 และ 25 เปอร์เซ็นต์ แสดงว่า วิธี MCC สามารถควบคุมการใช้แบบทดสอบย่อยซ้ำได้ดีกว่าวิธี CWA ในทุกเงื่อนไขการทดลอง ขณะที่วิธี CWA สามารถควบคุมการใช้แบบทดสอบย่อยซ้ำได้ดีเมื่อกำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) อยู่ที่ระดับ 20 และ 25 เปอร์เซ็นต์

2.4 การเปรียบเทียบค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ ($N_{neverexposed}$) จากวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ 2 วิธี ระหว่างวิธี CWA กับวิธี MCC ที่ใช้กับคลังข้อสอบ 600 ข้อ พบว่าวิธีที่ให้ค่าเฉลี่ยของ $N_{neverexposed}$ ต่ำที่สุดหรือมีประสิทธิภาพด้านการใช้คลังข้อสอบสูงสุดและรองลงมาใน 3 อันดับแรก ได้แก่ อันดับที่ 1 ได้แก่ วิธี MCC10 เมื่อทดสอบเปรียบเทียบกับวิธีอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของ $N_{neverexposed}$ มีค่าต่ำกว่าวิธี CWA20, CWA25, MCC20 และ MCC25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 อันดับที่ 2 ได้แก่ วิธี MCC15 เมื่อทดสอบเปรียบเทียบกับวิธีอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของ $N_{neverexposed}$ มีค่าต่ำกว่าวิธี CWA25, MCC20 และ MCC25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และต่ำกว่าวิธี CWA20 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 และอันดับที่ 3 ได้แก่ วิธี CWA10 เมื่อทดสอบเปรียบเทียบกับวิธีอื่นๆ แล้วปรากฏว่ามีค่าเฉลี่ยของ $N_{neverexposed}$ มีค่าต่ำกว่าวิธี MCC25 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และต่ำกว่าวิธี CWA25 และ MCC20 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

เมื่อพิจารณาจากแนวโน้มค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ ($N_{neverexposed}$) ที่เปลี่ยนไปเมื่ออัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) เพิ่มขึ้น เมื่อพิจารณาจะพบว่า ถ้าใช้วิธีการคัดเลือกแบบทดสอบย่อยแบบ MCC กับคลังข้อสอบที่มีขนาด 600 ข้อ และ 800 ข้อ ค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ ($N_{neverexposed}$) ของวิธี MCC มีค่าต่ำกว่า CWA เมื่อกำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) ไว้ที่ 10 และ 15 เปอร์เซ็นต์ แสดงว่าวิธี MCC สามารถใช้ประโยชน์จากคลังข้อสอบได้มีประสิทธิภาพดีกว่าวิธี CWA

ในกรณีที่กำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) ไว้ที่ 20 และ 25 เปอร์เซ็นต์ ค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ ($N_{neverexposed}$) ของวิธี MCC มีค่าสูงกว่าวิธี CWA แสดงว่าวิธี CWA สามารถใช้ประโยชน์จากคลังข้อสอบได้มีประสิทธิภาพดีกว่าวิธี MCC

2.5 ค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ใช้มากเกินไป ($N_{overexposed}$) จากวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ 2 วิธี ระหว่างวิธี CWA กับวิธี MCC ที่ใช้กับ

คลังข้อสอบ 600 ข้อ และ 800 ข้อ พบว่าวิธีที่มีค่าเฉลี่ยของ $N_{overexposed}$ สูงที่สุดหรือมีประสิทธิภาพด้านการใช้คลังข้อสอบต่ำที่สุดและรองลงมา ได้แก่ อันดับที่ 1 คือ วิธี CWA10 เมื่อทดสอบเปรียบเทียบกับวิธีอื่นปรากฏว่ามีค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ถูกนำมาใช้ถี่เกินไปมีค่าสูงกว่าทุกวิธียกเว้น CWA15 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 อันดับที่ 2 คือ วิธี CWA15 เมื่อทดสอบเปรียบเทียบกับกลุ่มวิธีอื่น ปรากฏว่ามีค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ถูกนำมาใช้ถี่เกินไปมีค่าสูงกว่าทุกวิธียกเว้น CWA10 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01

เมื่อพิจารณาจากแนวโน้มค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ถูกนำมาใช้มากเกินไป ($N_{overexposed}$) ที่เปลี่ยนไปเมื่ออัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max}) เพิ่มขึ้น เมื่อพิจารณาจะพบว่า ถ้าใช้วิธีการคัดเลือกแบบทดสอบย่อยแบบ MCC กับคลังข้อสอบที่มีขนาด 600 ข้อ และ 800 ข้อ ค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ถูกนำมาใช้มากเกินไป ($N_{overexposed}$) ของวิธี MCC มีค่าต่ำกว่า CWA เมื่อกำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max}) ไว้ที่ 10 และ 15 เปอร์เซ็นต์ แสดงว่าวิธี MCC สามารถควบคุมการใช้แบบทดสอบย่อยซ้ำได้มีประสิทธิภาพดีกว่าวิธี CWA เมื่อกำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max}) ไว้ที่ 10 และ 15 เปอร์เซ็นต์

กล่าวโดยสรุป ผลการเปรียบเทียบประสิทธิภาพของวิธีการคัดเลือกข้อสอบในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ จากเกณฑ์ที่ใช้พิจารณาประสิทธิภาพ พบว่า วิธีมอนติ คาร์โล ซีเอที (MCC) มีข้อได้เปรียบในด้านประสิทธิภาพของความถูกต้องแม่นยำในการประมาณค่าความสามารถของผู้สอบและมีความแกร่งต่อการกำหนดอัตราการใช้แบบทดสอบย่อยซ้ำในระดับที่ต่ำ (10% และ 15%) ขณะที่วิธีแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (CWA) สามารถใช้แบบทดสอบย่อยในคลังข้อสอบได้อย่างกระจายใกล้เคียงการแจกแจงแบบเอกรูปมากกว่าวิธีมอนติ คาร์โล ซีเอที (MCC) เมื่อกำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุดตั้งแต่ 20 เปอร์เซ็นต์ขึ้นไป ดังพิจารณาได้จากค่าอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ (r_{obs}) ค่าการเปรียบเทียบการแจกแจงของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ (ไคสแควร์) และจำนวนของแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ ($N_{neverexposured}$) ที่มีค่าต่ำกว่าค่าของวิธีมอนติ คาร์โล ซีเอที (MCC)

อภิปรายผลการวิจัย

การอภิปรายผลในงานวิจัยนี้นำเสนอ 2 ประเด็นหลัก คือ การอภิปรายผลตามสมมติฐานการวิจัย และการอภิปรายผลจากการจำลองข้อมูล มีรายละเอียดดังต่อไปนี้

1. การอภิปรายผลตามสมมติฐานการวิจัย

1.1 การอภิปรายตามสมมติฐานการวิจัยข้อที่ 1 วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที น่าจะมีประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถสูงกว่าวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ

จากผลการวิจัย เมื่อทดสอบความแตกต่างของค่าสถิติที่ใช้เป็นดัชนีในการพิจารณาประสิทธิภาพด้านความถูกต้องแม่นยำในการวัด จำแนกตามขนาดของคลังข้อสอบ โดยใช้การทดสอบของ Kruskal-Wallis พบว่าค่าเฉลี่ยของ รากที่สองความคลาดเคลื่อนเฉลี่ยกำลังสอง สหสัมพันธ์ระหว่างความสามารถจริงและความสามารถประมาณค่า และความยาวแบบสอบ ของตัวอย่างทั้ง 8 กลุ่ม แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 นั่นหมายความว่า มีตัวอย่างอย่างน้อย 1 กลุ่มที่มีค่าเฉลี่ยของ ค่าเฉลี่ยของ รากที่สองความคลาดเคลื่อนเฉลี่ยกำลังสอง สหสัมพันธ์ระหว่างความสามารถจริงและความสามารถประมาณค่า และความยาวแบบสอบ ต่างจากกลุ่มอื่น ผู้วิจัยจึงทดสอบความแตกต่างค่าเฉลี่ยของอันดับเป็นรายคู่โดยใช้วิธีของ Kruskal-Wallis one-way ANOVA พบว่า วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที มีประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถสูงกว่าวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ สังเกตได้จากค่า MSE และค่าความยาวของแบบสอบเฉลี่ย ที่มีค่าต่ำกว่าในทุกระดับของ $r_{\max}$ นอกจากนั้นค่าเฉลี่ยสหสัมพันธ์ระหว่างความสามารถประมาณค่ากับความสามารถจริงก็สูงกว่าในทุกระดับของ $r_{\max}$ ด้วยเช่นกัน โดยเงื่อนไขที่ทำให้วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที มีประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถสูงกว่าวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 คือ เงื่อนไขที่กำหนดระดับของ $r_{\max}$ ไว้ที่ 20% และ 25% ซึ่งสามารถใช้ได้กับทั้งคลังข้อสอบขนาด 600 และ 800 ข้อ จากผลการวิจัยที่เกิดขึ้นค่อนข้างสอดคล้องกับสมมติฐานการวิจัยข้อที่ 1 เนื่องจากการนำเทคนิคแบ่งคลังข้อสอบออกเป็นชั้นตามลำดับค่าอำนาจจำแนกของ Chang and Ying (1999) มาประยุกต์ใช้กับการถ่วงน้ำหนักด้วยด้วยดัชนีลำดับความสำคัญสูงสุด (maximum priority index) ของ Cheng and Chang (2009) จำเป็นต้องมีคลัง

ข้อสอบขนาดใหญ่ เพื่อจัดข้อสอบเข้าสู่ชั้นได้อย่างครอบคลุมช่วงความสามารถของผู้สอบ (-3.5 ถึง 3.5) และภายในแต่ละชั้นควรเป็นข้อสอบที่มีค่าอำนาจจำแนกใกล้เคียงกัน ถึงแม้ว่าการวิจัยครั้งนี้จะใช้คลังข้อสอบมีขนาดถึง 600 ข้อ และ 800 ข้อ แต่ข้อสอบเหล่านี้ถูกจัดเป็นชุดแบบทดสอบย่อย ดังนั้นหน่วยในการเลือกข้อสอบเพื่อนำไปใช้กับผู้สอบจึงเป็นชุดแบบทดสอบย่อย โดยคลังข้อสอบที่มีขนาด 600 ข้อ และ 800 ข้อ เมื่อจัดเป็นชุดแบบทดสอบย่อย จะเหลือเพียง 150 และ 200 ชุดแบบทดสอบย่อย ดังนั้นการจัดเรียงแบบทดสอบย่อยในแต่ละชั้น ซึ่งผู้วิจัยแบ่งชั้นของคลังข้อสอบขนาด 600 ข้อ เป็น 3 ชั้น และ 800 ข้อ เป็น 4 ชั้น แต่ละชั้นมีแบบทดสอบย่อยเพียง 50 ชุด ซึ่งอาจจะไม่คลอบคลุมช่วงความสามารถของผู้สอบ และค่าอำนาจจำแนกเฉลี่ยของแบบทดสอบย่อยแต่ละชุดภายในแต่ละชั้นของคลังข้อสอบอาจจะมีค่าแตกต่างกันมากเนื่องจากผู้วิจัยแบ่งคลังข้อสอบเป็น 3 และ 4 ชั้นเท่านั้น นอกจากนี้เหตุผลสำคัญอีกประการหนึ่งที่ทำให้วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที มีประสิทธิภาพด้านความถูกต้องแม่นยำของการประมาณค่าความสามารถสูงกว่าวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับอย่างเห็นได้ชัดเนื่องจากจากการศึกษาในสถานการณ์จำลองของ Belov et al. (2008) พบว่า วิธีมอนติ คาร์โล ซีเอที เป็นการประมาณค่าที่มีความแกร่ง (robust estimations) เนื่องจากวิธีมอนติ คาร์โล ซีเอที มีความไวต่ำ (less sensitive) ต่อการประมาณค่าความสามารถที่มีความคลาดเคลื่อนในช่วงเริ่มต้นของการทดสอบ เมื่อพบผู้สอบมีความสามารถสูงหรือต่ำ มากๆ ทำข้อสอบถูกหรือผิดซ้ำๆ กันในช่วงเริ่มต้นการทดสอบ

1.2 การอภิปรายตามสมมติฐานการวิจัยข้อที่ 2 วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที น่าจะมีประสิทธิภาพด้านการใช้คลังข้อสอบสูงกว่าวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ

สรุปผลจากการศึกษาครั้งนี้ พบว่า วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที (MCC) มีประสิทธิภาพด้านการใช้คลังข้อสอบสูงกว่าวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (CWA) เมื่อกำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) อยู่ที่ระดับ 10% และ 15% โดยสังเกตได้จากค่าดัชนีต่างๆ ดังต่อไปนี้ ได้แก่ ค่าเฉลี่ยของ อัตราการทับซ้อนของแบบสอบ ($\bar{t}$) ไคสแควร์ (x^2) อัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ (r_{obs}) จำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ ($N_{neverexposed}$) และจำนวนแบบทดสอบย่อยที่ใช้มากเกินไป ($N_{overexposed}$) ค่าดัชนีทั้ง 5 ตัว ของวิธี มอนติ คาร์โล ซีเอที มีค่าต่ำกว่าของวิธีแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ

เมื่อกำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด ($r_{\max}$) อยู่ที่ระดับ 25% ผลการทดลองปรากฏว่าดัชนี $N_{neverexposed}$ ของวิธี มอนติ คาร์โล ซีเอที (71.8) มีค่าสูงกว่า วิธีแบ่งกลุ่มค่าอำนาจ

จำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (52.4) แสดงว่า วิธีแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับสามารถเพิ่มอัตราการใช้ข้อสอบที่มีการใช้ต่ำได้ดีกว่าวิธี มอนติ คาร์โล ซีเอที เนื่องจากคลังข้อสอบถูกแบ่งออกเป็นชั้นตามค่าอำนาจจำแนก (a-Stratification) ทำให้ข้อสอบที่มีค่าอำนาจจำแนกๆ ซึ่งบรรจุอยู่ในคลังข้อสอบชั้นต้นๆ มีโอกาสที่จะถูกเลือกใช้เพิ่มมากขึ้น (Chang et al., 2001; Chang & van der Linden, 2003; Chang & Ying, 1999; Leung et al., 2002, 2003)

แต่เมื่อพิจารณาประสิทธิภาพด้านการใช้คลังข้อสอบโดยรวม พบว่าการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบมอนติ คาร์โล ซีเอที มีประสิทธิภาพเหนือกว่าการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ ผลการวิเคราะห์เป็นเช่นนี้ เนื่องจากวิธีมอนติ คาร์โล ซีเอที ไม่ต้องแบ่งคลังข้อสอบออกเป็นชั้นตามค่าอำนาจจำแนก ทำให้ลดข้อจำกัดเกี่ยวกับจำนวนแบบทดสอบย่อยที่บรรจุอยู่ภายในแต่ละชั้น ทำให้มีแบบทดสอบย่อยจำนวนมากให้เลือก และวิธีการได้มาของแบบทดสอบย่อยแต่ละชุดที่จะนำไปใช้กับผู้สอบนั้นจะต้องผ่านการทำแบบสอบเงา (shadow tests) เพื่อให้มีคุณสมบัติตรงตามเงื่อนไขที่กำหนด จากนั้นจึงสุ่มแบบทดสอบย่อยออกมาจำนวนหนึ่งจากลำดับของแบบทดสอบย่อยทั้งหมดแล้วเลือกข้อสอบที่ให้สารสนเทศสูงสุด ณ ตำแหน่ง $\hat{\theta}$ (Belov et al., 2008; Mao & Xin, 2013) หลักการดังกล่าวทำให้สามารถลดอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ลงอย่างชัดเจน แต่พบว่าไม่สามารถลดจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ ($N_{neverexposed}$) สังเกตได้จากค่าแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ที่เพิ่มขึ้นอย่างต่อเนื่อง เมื่ออัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max}) เพิ่มขึ้น

2. การอภิปรายผลจากการจำลองข้อมูล

2.1 การสร้างค่าพารามิเตอร์ผู้สอบและพารามิเตอร์ข้อสอบผู้วิจัยสร้างพารามิเตอร์ข้อสอบโดยกำหนดค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐาน และรูปแบบการแจกแจงข้อมูลตามการศึกษาวิจัยของ Ngudgratoke and Yon (2006) โดยกำหนดให้พารามิเตอร์อำนาจจำแนก (Discrimination parameter) สร้างขึ้นจากการแจกแจงแบบ Log-Normal [a ~LN(0.02,0.222)] เนื่องจากพารามิเตอร์อำนาจจำแนกควรมีค่าเป็นบวกและมากกว่า 0.5 และมีค่าสูงสุดไม่เกิน 3 พารามิเตอร์ความยาก (Difficulty parameter) สร้างขึ้นจากการแจกแจงปกติมาตรฐาน [b ~N(0,1)] เนื่องจากผู้สอบที่ผู้วิจัยสร้างขึ้นมีความสามารถแจกแจงแบบโค้งปกติมีค่าเฉลี่ยเท่ากับ 0 และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 1 ผู้สอบที่มีความสามารถสูงมากๆ และต่ำมากๆ มักพบได้น้อย ดั้งนั้นการกำหนดความยากของข้อสอบจึงต้องมีความสอดคล้องกัน ค่าพารามิเตอร์ความยากของข้อสอบที่สร้างขึ้นจึงมีข้อสอบที่มีความยากมากๆ หรือง่ายมากๆ จำนวนน้อยข้อ และมีความยากระดับกลางๆ จำนวนมาก และพารามิเตอร์การเดา (Guessing parameter) สร้างขึ้นจากการแจกแจงแบบ Beta [c ~BE(2,10)] เนื่องจากพารามิเตอร์การเดามีค่าเป็นบวกและค่าจะมีค่าไม่เกิน 0.2 ผลสร้างพารามิเตอร์ข้อสอบจาก

การใช้องค์ความรู้ที่ได้จากการวิจัยของ Ngudgratoke and Yon (2006) ทำให้ผู้วิจัยได้ข้อสอบที่มีค่าพารามิเตอร์เหมาะกับการนำไปใช้ทดสอบโดยใช้เกณฑ์การพิจารณาของ Mislevy (1991) ดังนั้นจึงมีข้อสอบจำนวน 12 ข้อ จากข้อสอบทั้งหมด 1400 ข้อ ที่มีค่าพารามิเตอร์อำนาจจำแนกสูง (a) สูงกว่า 3.0 ถือเป็นความผิดพลาดในการจำลองข้อมูลที่เกิดขึ้นต่ำมาก คิดเป็นร้อยละ 0.86 เท่านั้น นอกจากนี้ผู้วิจัยสังเกตพบว่า การใช้คลังข้อสอบที่มีค่าพารามิเตอร์เหมาะสมกับผู้สอบ จะทำให้การประมาณค่าความสามารถของผู้สอบมีความตรงค่อนข้างสูง ซึ่งสังเกตได้จากค่าสหสัมพันธ์ระหว่างค่าความสามารถที่แท้จริงกับความสามารถที่ประมาณค่า ($r_{\theta\hat{\theta}}$) ของการทดลองในแต่ละเงื่อนไขมีค่าสูงกว่า 0.95 ในทุกเงื่อนไขการทดลองซึ่งถือว่าสูงมาก

2.2 การตรวจสอบความถูกต้องของข้อมูลที่จำลองได้ การวิจัยครั้งนี้ศึกษาภายใต้ทฤษฎีการตอบสนองข้อสอบที่เป็นแบบทดสอบย่อย (Testlet Response Theory) เพื่อให้ข้อมูลที่จำลองขึ้นสอดคล้องกับบริบทที่ผู้วิจัยศึกษา ดังนั้นผู้วิจัยจำเป็นจะต้องตรวจสอบความไม่เป็นอิสระเฉพาะที่ของข้อสอบ (local item dependence: LID) เพื่อให้มั่นใจว่าผลการตอบของผู้สอบที่เกิดขึ้นมีอิทธิพลของแบบทดสอบย่อยแฝงอยู่ด้วย ผู้วิจัยใช้สถิติ Q3 ของ Yen (1984) ตามคำแนะนำของ Wainer et al. (2007) และ Thissen, Steinberg, and Mooney (1989) เป็นเครื่องมือในการวินิจฉัยความไม่เป็นอิสระเฉพาะที่ของข้อสอบ เกณฑ์การพิจารณามีดังนี้ ถ้าค่าสถิติ Q3 มีค่าเป็นลบต่ำๆ หรือมีค่าเข้าใกล้ 0 มากๆ แสดงว่าข้อสอบคู่นั้นเป็นอิสระกัน ถ้าค่าสถิติ Q3 มีค่าเป็นบวกแสดงว่าข้อสอบคู่นั้นมีหลักฐานพอเชื่อได้ว่าไม่เป็นอิสระจากกัน ผลการวิเคราะห์ค่าสถิติ Q3 ของแบบทดสอบย่อยแต่ละชุดในคลังข้อสอบขนาด 600 ข้อ และ 800 ข้อ พบว่าค่าเฉลี่ยของสถิติ Q3 มีค่าเป็น 0.132 และ 0.126 ตามลำดับ และเมื่อพิจารณาเป็นรายแบบทดสอบย่อย พบว่า ค่าสถิติ Q3 ของแบบทดสอบย่อยในคลังข้อสอบขนาด 600 ข้อ มีค่าต่ำสุด (0.058) และ ค่าสูงสุด (0.291) และพบว่า ค่าสถิติ Q3 ของแบบทดสอบย่อยในคลังข้อสอบขนาด 800 ข้อ มีค่าต่ำสุด (0.039) และค่าสูงสุด (0.259) แสดงว่าข้อสอบภายในแบบทดสอบย่อยมีความสัมพันธ์กันอันเนื่องมีจากอิทธิพลของแบบทดสอบย่อยซึ่งเหมาะสมกับบริบทที่ต้องใช้โมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย (Testlet Response Model)

ข้อเสนอแนะ

ข้อเสนอแนะในการนำผลการวิจัยไปใช้

การนำผลการศึกษาไปประยุกต์ใช้สำหรับหน่วยงานหรือองค์กรการทดสอบที่ต้องการพัฒนาระบบการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ Thompson and Weiss (2011) เสนอขอบข่ายงานสำหรับการพัฒนาระบบการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์สำหรับองค์กรการทดสอบประกอบด้วย 5 ขั้นตอนหลักที่สำคัญ ดังนี้

ขั้นที่ 1) การศึกษาความเป็นไปได้ (feasibility) การประยุกต์ใช้ (applicability) และการวางแผน (planning) งานสำคัญในขั้นนี้คือ การศึกษาในสถานการณ์จำลองแบบมอนติ คาร์โล และการประเมินงบประมาณด้านความคุ้มค่าต่อต้นทุนที่ใช้ในการพัฒนาระบบการทดสอบแบบปรับเหมาะด้วยความพิวเตอร์ (business case evaluation) ขั้นที่ 2) การพัฒนาคลังข้อสอบหรือการใช้ประโยชน์จากคลังข้อสอบเดิมที่มีอยู่ งานสำคัญสำหรับขั้นนี้คือ การเขียนและทบทวนข้อสอบ ขั้นที่ 3) การทดสอบเบื้องต้น (pretest) และการทำคลังข้อสอบให้เป็นมาตรฐานตามโมเดลการวัดที่ใช้สำหรับการทดสอบ (calibrate itembank) งานสำคัญสำหรับขั้นตอนนี้คือ การทำแบบสอบก่อน (pretesting) และการวิเคราะห์ข้อสอบ (item analysis) ขั้นที่ 4) การพิจารณาข้อกำหนดจำเพาะสำหรับการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (CAT) ขั้นสุดท้าย งานสำคัญสำหรับขั้นนี้คือ การศึกษาในสถานการณ์จำลองแบบ post-hoc ที่เป็นการศึกษาในสถานการณ์จำลองโดยใช้ข้อมูลจริงที่ได้จากการทดสอบแบบข้อเขียนเพื่อตอบคำถามว่า “อะไรจะเกิดขึ้นถ้าผู้สอบดำเนินการทดสอบแบบ CAT โดยใช้ข้อสอบทั้งหมดนี้” เพื่อเป็นการทำนายผลลัพธ์ที่เกิดขึ้นเมื่อนำคลังข้อสอบที่ใช้กับแบบข้อเขียนไปใช้กับ CAT หรือ การศึกษาในสถานการณ์จำลองแบบ hybrid ที่เป็นการศึกษาในสถานการณ์จำลองโดยนำจุดเด่นของมอนติ คาร์โล และ post-hoc มาใช้ร่วมกันเพื่อแก้ปัญหาบางอย่างที่ไม่สามารถแก้ได้โดยการศึกษาในสถานการณ์จริง ซึ่งวิธีการจำลองแบบ hybrid เป็นกาอนุญาตให้ผู้วิจัยทำการทดลองแบบ post-hoc จากเมตริกซ์ข้อมูลผลการตอบของผู้สอบที่ไม่สมบูรณ์โดยวิธีการแบบ hybrid มีทางเลือกที่เป็นไปได้ในการดำเนินการ 2 กรณี คือ (1) ถ้าพารามิเตอร์ของข้อสอบที่ใช้ใน CAT ยังไม่เคยถูกประมาณค่ามาก่อน และ (2) เป็นการประเมินประสิทธิภาพของ CAT กับผู้สอบกลุ่มใหม่ (Nydick & Weiss, 2009) ขั้นที่ 5) การนำระบบ CAT ไปใช้กับผู้สอบจริง (live CAT) งานสำคัญสำหรับขั้นนี้คือ การพัฒนาซอฟต์แวร์ การเผยแพร่และการกระจายซอฟต์แวร์

ผลการศึกษาครั้งนี้เกิดจากการจำลองข้อมูลแบบมอนติ คาร์โล ซึ่งถือเป็นการศึกษาในขั้นแรกของการออกแบบและพัฒนาระบบการแบบปรับเหมาะด้วยคอมพิวเตอร์ โดยทำศึกษาในเชิงทฤษฎีเพื่อเปรียบเทียบประสิทธิภาพของวิธีมอนติ คาร์โล ซีเอที (MCC) (Belov et al., 2008) และวิธีการ

แบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับ (CWA) (Cheng et al., 2009) เมื่อนำวิธีการดังกล่าวมาใช้กับคลังข้อสอบที่ภายในบรรจุข้อสอบมีลักษณะเป็นแบบทดสอบย่อย กับเงื่อนไขที่มีการกำหนดขนาดคลังข้อสอบ และอัตราการใช้แบบทดสอบย่อยซ้ำแตกต่างกัน ผลการวิเคราะห์พบว่า วิธีมอนติ คาร์โล ซีเอที มีประสิทธิภาพเหนือกว่าทั้งด้านความแม่นยำในการวัดและความสมดุลของการใช้คลังข้อสอบในทุกเงื่อนไขของการทดลอง การนำผลการวิจัยไปใช้ในแต่ละขั้นของการออกแบบและพัฒนาระบบการแบบปรับเหมาะด้วยคอมพิวเตอร์สามารถทำได้ดังนี้

1. ผลการวิจัยบ่งชี้ว่าถ้าลดขนาดคลังข้อสอบลงจาก 800 ข้อ เป็น 600 ข้อ วิธี MCC จะใช้แบบทดสอบย่อยประมาณ 10 ฉบับ ซึ่งมากกว่า คลังข้อสอบขนาด 600 ข้อ เพียง 1 ฉบับ และมีความคลาดเคลื่อนมาตรฐานของการประมาณค่าต่ำมาก โดยมีพิสัยอยู่ระหว่าง 0.291 ถึง 0.294 ความคลาดเคลื่อนที่เกิดขึ้นมีค่าน้อยกว่า 0.3 จะให้ความตรงตามสภาพสูงสุด (รังสรรค์ มณีเล็ก, 2540) และจากการศึกษาของ สิริลักษณ์ เกษรปทุมานันท์ และ ณัฏฐภรณ์ หลาวทอง (2007) พบว่าถ้าผู้บริหารยอมให้มีความคลาดเคลื่อนมาตรฐานของการประมาณค่า มีค่าประมาณ 0.45 ผู้สอบจะใช้จำนวนข้อสอบลดลงประมาณร้อยละ 50 แต่ค่าฟังก์ชันสารสนเทศของแบบสอบโดยเฉลี่ยก็จะลดลงประมาณร้อยละ 50 เช่นกัน จากประเด็นนี้ จึงควรนำผลการวิจัยที่ค้นพบไปศึกษาเพิ่มเติมเพื่อให้เกิดความชัดเจน ว่าสามารถลดความยาวของแบบสอบลงได้ครึ่งหนึ่งหรือไม่และความตรงตามสภาพที่เกิดขึ้นมีความน่าเชื่อถือเพียงใด เพื่อนำผลการวิจัยไปวางแผนการตัดสินใจในการกำหนดเป้าหมายของการทดสอบ ว่าเป็นการทดสอบที่นำผลมาปรับปรุงการเรียนการสอน (Formative assessment) หรือเป็นการทดสอบเพื่อตัดสินผลการเรียนรู้ (Summative assessment) หรือเป็นการทดสอบเพื่อคัดเลือกคนเข้าศึกษาต่อ จากนั้นจึงดำเนินการวางแผนออกแบบและพัฒนาคลังข้อสอบต่อไป

2. ผลการวิจัยบ่งชี้ว่าการทดสอบแบบปรับเหมาะแบบมอนติ คาร์โล ซีเอที มีประสิทธิภาพเหนือกว่าการทดสอบแบบปรับเหมาะแบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับในทุกๆ ด้าน แต่มีข้อสังเกตบางประการจากผลการทดสอบที่รายงานว่าค่าเฉลี่ยของจำนวนแบบทดสอบย่อยที่ไม่ถูกนำมาใช้ ($N_{neverexposed}$) ของการทดสอบแบบปรับเหมาะแบบมอนติ คาร์โล ซีเอที มีแนวโน้นเพิ่มสูงขึ้นอย่างเห็นได้ชัด เมื่อยอมให้มีอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max}) เพิ่มขึ้น ขณะที่ดัชนีตัวอื่นๆ เช่น อัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้ อัตราการทับซ้อนของแบบสอบ และค่าไคสแควร์ที่เปรียบความแตกต่างของอัตราการใช้แบบทดสอบย่อยซ้ำที่สังเกตได้กับอัตราการใช้แบบทดสอบย่อยซ้ำที่คาดหวัง มีค่าเฉลี่ยออกมาออกมาอยู่ในเกณฑ์ที่ดีคือมีค่าต่ำกว่าค่าที่ได้จากการทดสอบแบบปรับเหมาะแบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับใน ค่าสถิติเหล่านี้บ่งชี้ให้เห็นว่าการทดสอบแบบปรับเหมาะแบบมอนติ คาร์โล ซีเอที สามารถควบคุมอัตราการใช้แบบทดสอบย่อยซ้ำได้ดีแต่มีจุดอ่อนในเรื่องการใช้ประโยชน์จากคลังข้อสอบเนื่องจากที่ไม่ถูก

นำมาใช้มีจำนวนเพิ่มขึ้นมากเมื่อยอมให้มีอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max}) เกิดขึ้นได้ 25 % ดังนั้นผู้ใช้ผลการควรพิจาณาในเรื่องของการกำหนดอัตราการใช้แบบทดสอบย่อยซ้ำสูงสุด (r_{max}) ให้เหมาะสมกับบริบทของการทดสอบไม่ให้มีค่ามากหรือน้อยจนเกินไปโดยพิจารณาว่าการทดสอบเป็นการทดสอบที่มีส่วนได้เสียสูงหรือไม่ หรือเป็นการทดสอบเพื่อนำข้อมูลมาพัฒนาผู้สอบ

3. การพัฒนาคลังข้อสอบสำหรับการแบบปรับเหมาะด้วยคอมพิวเตอร์โดยทั่วๆ ไป รวมถึงวิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่ใช้ในการศึกษาครั้งนี้ สิ่งที่ผู้ตัดสินใจเชิงนโยบายของการทดสอบต้องตั้งคำถาม คือ ผู้สอบทุกคนที่ทำการทดสอบไม่ว่าจะเป็นผู้สอบที่มีความสามารถสูงมากๆ หรือ ผู้สอบที่มีความสามารถต่ำมากๆ จำเป็นที่จะต้องมีความแม่นยำของการวัดเท่าๆ กันหรือไม่ ถ้าจำเป็นควรจะมีจำนวนข้อสอบที่มีระดับความง่ายและมากสูงๆ มากเพียงพอกับความต้องการด้วย เพื่อที่จะประมาณค่าความสามารถของผู้สอบโดยมีความคลาดเคลื่อนอยู่ในระดับเดียวกัน ดังภาพที่ 3.12 ที่แสดงให้เห็นว่า ผู้สอบที่มีความสามารถสูงหรือต่ำมากๆ จำเป็นที่จะต้องใช้แบบสอบที่มีความยาวค่อนข้างมากเพื่อที่จะทำให้ความคลาดเคลื่อนมาตรฐานของการประมาณค่าอยู่ในระดับเดียวกับผู้สอบที่มีความสามารถปานกลาง สำหรับการพัฒนาคลังข้อสอบสำหรับการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์นั้น มีแนวทางในการดำเนินการสองแบบคือ 1) เป็นคลังข้อสอบที่มีข้อสอบที่ถูกสร้างขึ้นใหม่ทั้งหมด และ 2) เป็นการนำคลังข้อสอบที่มีอยู่มาใช้ให้เกิดประโยชน์ ซึ่งไม่ว่าจะเป็นคลังข้อสอบที่ประกอบไปด้วย ข้อสอบใหม่ทั้งหมด หรือ เป็นคลังข้อสอบที่มีทั้งข้อสอบใหม่และข้อสอบเก่าผสมกัน สิ่งสำคัญที่ต้องพิจารณาเหมือนกันคือ ค่าทางสถิติของข้อสอบซึ่งเกี่ยวข้องกับขั้นที่ 3 ตามกรอบงานของ Thompson and Weiss (2011) โดยพารามิเตอร์ของข้อสอบจะต้องถูกประมาณค่าตามโมเดลที่ใช้ในการทดสอบ ซึ่งองค์ประกอบสำคัญที่จะทำให้เชื่อมั่นได้ค่าพารามิเตอร์ของข้อสอบทั้งหมดถูกทำให้เป็นมาตรฐานตามโมเดลที่ใช้ในการทดสอบ (calibrated) บนมาตรวัดที่ใช้ร่วมกัน (common scale) สิ่งนั้นเรียกว่าการทำ linking ซึ่งแนวทางสำหรับการทำ linking สามารถศึกษาเพิ่มเติมได้จาก (Kolen & Brennan, 2014)

ข้อเสนอแนะในการวิจัยครั้งต่อไป

1. การวิจัยครั้งนี้เปรียบเทียบประสิทธิภาพของการประมาณค่าความสามารถของผู้สอบและประสิทธิภาพความสมดุลการใช้คลังข้อสอบของวิธีการคัดเลือกแบบทดสอบย่อยภายใต้ทฤษฎีการตอบสนองข้อสอบที่เป็นแบบทดสอบย่อย (Testlet Response Theory) ผู้วิจัยออกแบบให้แบบทดสอบย่อย 1 ชุด มีจำนวน ข้อสอบ 4 ข้อ และให้ข้อสอบทั้ง 4 ข้อภายใน แบบทดสอบย่อยใช้สิ่งเร้าร่วมกัน ตลอดทั้งข้อสอบ และการเลือกคัดเลือกข้อสอบจะเป็นการเลือกข้อสอบทั้งหมดที่อยู่

ภายในแบบทดสอบย่อยมาใช้กับผู้สอบ การพิจารณาอัตราการใช้ข้อสอบซ้ำจึงพิจารณาตามแบบทดสอบย่อย ข้อสอบแต่ละภายในแบบทดสอบย่อยจะมีอัตราการใช้ข้อสอบซ้ำเท่าๆ กัน เพราะถูกนำมาใช้พร้อมกันทั้งมด ดังนั้นในการศึกษาครั้งต่อไป ควรกำหนดสิ่งเร้า และข้อสอบต้องใช้ร่วมกัน แยกออกจากัน ผู้สอบแต่ละคนที่ได้รับแบบทดสอบย่อย อาจจะได้รับแบบทดสอบย่อยเรื่องเดียวกัน แต่จำนวนข้อสอบไม่เท่ากัน เนื่องจากการกำหนดจำนวนข้อสอบ และสิ่งเร้าที่กำหนดขึ้นแบบตายตัวนี้ อาจจะไม่ยืดหยุ่นต่อการนำไปใช้งานในสถานการณ์จริง และแบบทดสอบย่อยบางเรื่องอาจจะจำเป็นต้องใช้ข้อสอบจำนวนมากกว่า 4 ข้อ และควรศึกษาอัตราการใช้ข้อสอบซ้ำในระดับข้อ เพื่อให้ทราบถึงการใช้ถึงการใช้ประโยชน์จากคลังข้อสอบที่เหมาะสมกับสถานการณ์จริง

2. งานวิจัยครั้งนี้กำหนดขอบเขตเนื้อหาของคลังข้อสอบไว้ 3 ขอบเขตเนื้อหา เพื่อสร้างเงื่อนไขบังคับในการเลือกข้อสอบ แต่ผู้วิจัยศึกษาเฉพาะประสิทธิภาพของการประมาณค่าความสามารถของผู้สอบและประสิทธิภาพความสมดุลการใช้คลังข้อสอบของวิธีการคัดเลือกแบบทดสอบย่อยแบบมอนติ คาร์โล ซีเอที กับแบบแบ่งกลุ่มค่าอำนาจจำแนกแบบถ่วงน้ำหนักที่มีการบังคับเท่านั้น ยังขาดองค์ความรู้ในเรื่องการศึกษาเปรียบเทียบการกระจายขอบเขตเนื้อหาของข้อสอบตามผังการทดสอบ งานวิจัยครั้งต่อไปควรศึกษาเกี่ยวกับการควบคุมความสมดุลของเนื้อหาที่ผู้สอบแต่ละคนได้รับ รวมถึงรวมถึงอัตราการฝ่าฝืนเงื่อนไขบังคับที่กำหนดขึ้นว่ามีโอกาสเกิดขึ้นแบบใดบ้าง และเกิดในสถานการณ์การทดสอบแบบใด เพื่อให้ผลการวิจัยสามารถรองรับการนำวิธีการทดสอบแบบปรับเหมาะทั้งสองวิธีไปใช้ในสถานการณ์จริงได้อย่างเหมาะสม

3. เนื่องจากศึกษาครั้งนี้ใช้วิธีมอนติ คาร์โล โดยใช้คลังข้อสอบเชิงสมมติฐานซึ่งกำหนดให้ข้อสอบจัดกลุ่มเป็นแบบทดสอบย่อยและใช้สิ่งเร้าร่วมกัน โดยแบบทดสอบย่อยแต่ละฉบับมีจำนวนข้อสอบทั้งหมด 4 ข้อ และข้อสอบในคลังข้อสอบทั้งหมดมุ่งวัดคุณลักษณะเดียว (unidimension) ซึ่งศึกษาภายใต้โมเดลการตอบสนองข้อสอบที่ใช้แบบทดสอบย่อย ดังนั้นจึงควรศึกษาเพิ่มเติมกับการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่ใช้โมเดลการตอบสนองข้อสอบแบบพหุมิติ กับคลังข้อสอบที่มีการใช้แบบทดสอบย่อย เนื่องจากสิ่งเร้าที่กลุ่มของข้อสอบที่ใช้ร่วมกันภายในแบบทดสอบย่อย เช่น บทความ กราฟ หรือ รูปภาพ ผู้สอบอาจต้องใช้ความสามารถมากกว่า 1 ด้านในการตอบข้อสอบภายในแบบทดสอบย่อยชุดนั้น หรือต้องใช้ความสามารถมากกว่า 1 ด้านในการทดสอบ 1 ครั้ง เช่น การสอบ PISA ที่ต้องใช้ความสามารถด้านการอ่านเพื่อความเข้าใน การให้เหตุผลเชิงวิเคราะห์ และการให้เหตุผลเชิงตรรกะ

รายการอ้างอิง

ภาษาไทย

ต่าย เซี่ยงฉี. (2534). การศึกษาเปรียบเทียบประสิทธิภาพในการประมาณค่าความสามารถของผู้สอบจากการทดสอบเทเลอร์รูปปิรามิดที่มีรูปแบบ จำนวนขั้นและวิธีการให้คะแนนที่แตกต่างกันโดยใช้วิธีมอนติคาร์โล. (ปริญญานิพนธ์การศึกษาศาสตรดุษฎีบัณฑิต), มหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร.

รังสรรค์ มณีเล็ก. (2540). ผลของตัวแปรบางตัวต่อความเที่ยงตรงเชิงสภาพและจำนวนข้อสอบที่ใช้ในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์. (ปริญญานิพนธ์การศึกษาศาสตรดุษฎีบัณฑิต), มหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร.

ศิริชัย กาญจนวาสี. (2538). การทดสอบแบบปรับเหมาะกับความสามารถของผู้สอบ = *Adaptive testing*. กรุงเทพมหานคร: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

ศิริชัย กาญจนวาสี. (2550). ทฤษฎีการทดสอบแนวใหม่ (*Modern test theories)* (พิมพ์ครั้งที่ 4 ed.). กรุงเทพมหานคร: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

ศิริชัย กาญจนวาสี. (2552). ทฤษฎีการทดสอบแบบดั้งเดิม (*Classical test theories)* (พิมพ์ครั้งที่ 6 ed.). กรุงเทพมหานคร: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

สิริลักษณ์ เกษรปทุมานันท์, & ณัฏฐภรณ์ หลาวทอง. (2007). การเปรียบเทียบความตรงตามสภาพในการประมาณค่าความสามารถของผู้สอบจากการทดสอบแบบปรับเหมาะโดยใช้คอมพิวเตอร์ที่ใช้เกณฑ์การคัดเลือกข้อสอบขั้นแรก อัตราการใช้ข้อสอบซ้ำ และเกณฑ์ยุติการทดสอบที่ต่างกัน. *OJED, 2*(1), 902-916.

ภาษาอังกฤษ

Adema, J. J., Boekkooi-Timminga, E., & van der Linden, W. J. (1991). Achievement test construction using 0–1 linear programming. *European Journal of Operational Research, 55*(1), 103-111. doi: 10.1016/0377-2217(91)90195-2

Ariel, A., Veldkamp, B. P., & Breithaupt, K. (2006). Optimal Testlet Pool Assembly for Multistage Testing Designs. *Applied Psychological Measurement, 30*(3), 204-215. doi: 10.1177/0146621605284350

Ariel, A., Veldkamp, B. P., & van der Linden, W. J. (2004). Constructing Rotating Item Pools for Constrained Adaptive Testing. *Journal of Educational Measurement, 41*(4), 345-359. doi: 10.1111/j.1745-3984.2004.tb01170.x

Armstrong, R. D., Jones, D. H., Koppel, N. B., & Pashley, P. J. (2004). Computerized Adaptive Testing With Multiple-Form Structures. *Applied Psychological Measurement, 28*(3), 147-164. doi: 10.1177/0146621604263652

Belov, D. I. (2008). Uniform Test Assembly. *Psychometrika, 73*(1), 21-38. doi: 10.1007/s11336-007-9025-0

Belov, D. I., & Armstrong, R. D. (2005). Monte Carlo Test Assembly for Item Pool Analysis and Extension. *Applied Psychological Measurement, 29*(4), 239-261. doi: 10.1177/0146621605275413

Belov, D. I., Armstrong, R. D., & Weissman, A. (2008). A Monte Carlo Approach for Adaptive Testing With Content Constraints. *Applied Psychological Measurement, 32*(6), 431-446. doi: 10.1177/0146621607309081

Berkelaar, M. (2007). lpSolve: Interface to Lp solve v. 5.5 to solve linear or integer programs. *R package version, 5*(8).

Berkelaar, M., Eikland, K., & Notebaert, P. (2004). lpsolve: Open source (mixed-integer) linear programming system. *Eindhoven U. of Technology*.

Bland, J. A., & Dawson, G. P. (1991). Tabu search and design optimization. *Computer-Aided Design, 23*(3), 195-201. doi: 10.1016/0010-4485(91)90089-f

Boekkooi-Timminga, E. (1990). The Construction of Parallel Tests from IRT-Based Item Banks. *Journal of Educational Statistics, 15*(2), 129-145.

Boyd, A. M. (2003). *Strategies for controlling testlet exposure rates in computerized adaptive testing systems.* (Doctoral Dissertation), The University of Texas at Austin.

Chang, H. H., Qian, J., & Ying, Z. (2001). a-Stratified Multistage Computerized Adaptive Testing with b Blocking. *Applied Psychological Measurement, 25*(4), 333-341. doi: 10.1177/01466210122032181

Chang, H. H., & van der Linden, W. J. (2003). Optimal Stratification of Item Pools in a-Stratified Computerized Adaptive Testing. *Applied Psychological Measurement, 27*(4), 262-274. doi: 10.1177/0146621603027004002

Chang, H. H., & Ying, Z. (1996). A Global Information Approach to Computerized Adaptive Testing. *Applied Psychological Measurement, 20*(3), 213-229. doi: 10.1177/014662169602000303

Chang, H. H., & Ying, Z. (1999). a-Stratified Multistage Computerized Adaptive Testing. *Applied Psychological Measurement, 23*(3), 211-222. doi: 10.1177/01466219922031338

Chang, H. H., & Zhang, J. (2002). Hypergeometric family and item overlap rates in computerized adaptive testing. *Psychometrika, 67*(3), 387-398. doi: 10.1007/bf02294991

Chang, S.-W., & Ansley, T. N. (2003). A Comparative Study of Item Exposure Control Methods in Computerized Adaptive Testing. *Journal of Educational Measurement, 40*(1), 71-103. doi: 10.1111/j.1745-3984.2003.tb01097.x

Chang, T.-Y., & Shiu, Y.-F. (2011). Simultaneously construct IRT-based parallel tests based on an adapted CLONALG algorithm. *Applied Intelligence*, 1-16. doi: 10.1007/s10489-011-0308-x

Chen, S.-Y. (2010). A Procedure for Controlling General Test Overlap in Computerized Adaptive Testing. *Applied Psychological Measurement, 34*(6), 393-409. doi: 10.1177/0146621610367788

Chen, S.-Y., & Ankenman, R. D. (2004). Effects of Practical Constraints on Item Selection Rules at the Early Stages of Computerized Adaptive Testing. *Journal of Educational Measurement, 41*(2), 149-174. doi: 10.1111/j.1745-3984.2004.tb01112.x

Chen, S.-Y., Ankenman, R. D., & Chang, H. H. (2000). A Comparison of Item Selection Rules at the Early Stages of Computerized Adaptive Testing. *Applied Psychological Measurement, 24*(3), 241-255. doi: 10.1177/01466210022031705

Chen, S.-Y., Ankenmann, R. D., & Spray, J. A. (2003). The Relationship Between Item Exposure and Test Overlap in Computerized Adaptive Testing. *Journal of Educational Measurement, 40*(2), 129-145. doi: 10.1111/j.1745-3984.2003.tb01100.x

Chen, S.-Y., & Doong, S. H. (2008). Predicting item exposure parameters in computerized adaptive testing. *Br J Math Stat Psychol, 61*(Pt 1), 75-91. doi: 10.1348/000711006X129553

Chen, S.-Y., & Lei, P. W. (2005). Controlling Item Exposure and Test Overlap in Computerized Adaptive Testing. *Applied Psychological Measurement, 29*(3), 204-217. doi: 10.1177/0146621604271495

Chen, S.-Y., & Lei, P. W. (2010). Investigating the relationship between item exposure and test overlap: item sharing and item pooling. *Br J Math Stat Psychol, 63*(Pt 1), 205-226. doi: 10.1348/000711009X430906

Chen, S.-Y., Lei, P. W., & Liao, W. H. (2008). Controlling item exposure and test overlap on the fly in computerized adaptive testing. *Br J Math Stat Psychol, 61*(Pt 2), 471-492. doi: 10.1348/000711007X227067

Cheng, P. E., & Liou, M. (2003). Computerized Adaptive Testing Using the Nearest-Neighbors Criterion. *Applied Psychological Measurement, 27*(3), 204-216. doi: 10.1177/0146621603027003002

Cheng, Y., & Chang, H. H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *Br J Math Stat Psychol, 62*(Pt 2), 369-383. doi: 10.1348/000711008X304376

Cheng, Y., Chang, H. H., Douglas, J., & Guo, F. (2009). Constraint-Weighted a-Stratification for Computerized Adaptive Testing With Nonstatistical Constraints. *Educational and Psychological Measurement, 69*(1), 35-49. doi: 10.1177/0013164408322030

Cheng, Y., Chang, H. H., & Yi, Q. (2007). Two-Phase Item Selection Procedure for Flexible Content Balancing in CAT. *Applied Psychological Measurement, 31*(6), 467-482. doi: 10.1177/0146621606292933

Davis, L. L., & Dodd, B. G. (2003). Item Exposure Constraints for Testlets in the Verbal Reasoning Section of the MCAT. *Applied Psychological Measurement, 27*(5), 335-356. doi: 10.1177/0146621603256804

Dekking, F. M., Kraaikamp, C., Lopuhaa, H. P., & Meester, L. E. (2005a). Efficiency and mean squared error. *A modern introduction to probability and statistics: Understanding why and how.* (1 ed., pp. 299-311). Verlag London: Springer.

Dekking, F. M., Kraaikamp, C., Lopuhaa, H. P., & Meester, L. E. (2005b). Unbiased estimators. *A modern introduction to probability and statistics: Understanding why and how.* (1 ed., pp. 285-297). Verlag London: Springer.

Diao, Q., & van der Linden, W. J. (2011). Automated Test Assembly Using lp_Solve Version 5.5 in R. *Applied Psychological Measurement, 35*(5), 398-409. doi: 10.1177/0146621610392211

Eggen, T. J. H. M. (2001). *Overexposure and underexposure of items in computerized adaptive testing*. (Measurement and Research Department Reports 2001-1), (Citogroep). Arnhem, The Netherlands.

Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized Adaptive Testing for Classifying Examinees into three Categories. *Educational and Psychological Measurement, 60*(5), 713-734. doi: 10.1177/00131640021970862

Elissavet, G., Evangelos, T., & Economides, A. A. (2007). Review of Item Exposure Control Strategies for Computerized Adaptive Testing Developed from 1983 to 2005. *The Journal of Technology, Learning, and Assessment, 5*(8).

Flaugher, R. (2000). Item Pools. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 37-60). Mahwah, NJ: Lawrence Erlbaum.

Gilks, W. R. (2005). Markov Chain Monte Carlo *Encyclopedia of Biostatistics*: John Wiley & Sons, Ltd.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: MA: Kluwer-Nijhoff Publishing.

Hau, K. T., & Chang, H. H. (2001). Item Selection in Computerized Adaptive Testing: Should More Discriminating Items be Used First? *Journal of Educational Measurement, 38*(3), 249-266. doi: 10.1111/j.1745-3984.2001.tb01126.x

Hendrickson, A. (2007). An NCME Instructional Module on Multistage Testing. *Educational Measurement: Issues and Practice, 26*(2), 44-52. doi: 10.1111/j.1745-3992.2007.00093.x

Keng, L. (2008). *A Comparison of the Performance of Testlet-Based Computer Adaptive Tests and Multistage Tests.* (Doctoral Dissertation), The University of Texas at Austin.

Keng, L., Ho, T.-H., Chen, T.-A. A., & Dodd, B. G. (2008). *A Comparison of Item and Testlet Selection Procedures in Computerized Adaptive Testing* Paper presented at the annual meeting of the National Council on Measurement in Education, NYC, NY.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for Selecting Items for Computerized Adaptive Tests. *Applied Measurement in Education, 2*(4), 359.

Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking: Methods and Practices*: Springer.

Lee, H., & Dodd, B. G. (2011). Comparison of Exposure Controls, Item Pool Characteristics, and Population Distributions for CAT Using the Partial Credit Model. *Educational and Psychological Measurement*. doi: 10.1177/0013164411411296

Leung, C. K., Chang, H. H., & Hau, K. T. (2002). Item Selection in Computerized Adaptive Testing: Improving the a-Stratified Design with the Sympson-Hetter Algorithm. *Applied Psychological Measurement, 26*(4), 376-392. doi: 10.1177/014662102237795

Leung, C. K., Chang, H. H., & Hau, K. T. (2003). Incorporation Of Content Balancing Requirements In Stratification Designs For Computerized Adaptive Testing. *Educational and Psychological Measurement, 63*(2), 257-270. doi: 10.1177/0013164403251326

Li, Y. H., & Schafer, W. D. (2005). Increasing the Homogeneity of CAT's Item-Exposure Rates by Minimizing or Maximizing Varied Target Functions While Assembling Shadow Tests. *Journal of Educational Measurement, 42*(3), 245-269. doi: 10.1111/j.1745-3984.2005.00013.x

Luecht, R. M. (1998). Computer-Assisted Test Assembly Using Optimization Heuristics. *Applied Psychological Measurement, 22*(3), 224-236. doi: 10.1177/01466216980223003

Luecht, R. M., & Nungester, R. J. (1998). Some Practical Examples of Computer-Adaptive Sequential Testing. *Journal of Educational Measurement, 35*(3), 229-249. doi: 10.1111/j.1745-3984.1998.tb00537.x

Mao, X., & Xin, T. (2013). The Application of the Monte Carlo Approach to Cognitive Diagnostic Computerized Adaptive Testing With Content Constraints. *Applied Psychological Measurement, 37*(6), 482-496. doi: 10.1177/0146621613486015

Meijer, R. R., & Nering, M. L. (1999). Computerized Adaptive Testing: Overview and Introduction. *Applied Psychological Measurement, 23*(3), 187-194. doi: 10.1177/01466219922031310

Mislevy, R. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*(2), 177-196. doi: 10.1007/BF02294457

Murphy, D. L., Dodd, B. G., & Vaughn, B. K. (2010). A Comparison of Item Selection Techniques for Testlets. *Applied Psychological Measurement, 34*(6), 424-437. doi: 10.1177/0146621609349804

Ngudgratoke, S., & Yon, H. (2006). *Vertical linking of tests composed of testlets: A comparison between 3-PL model, graded response model, and testlet response model*. Paper presented at the annual meeting of the National Council on Measurement in Education, SF, CA.

Nydick, S. W., & Weiss, D. J. (2009). *A hybrid simulation procedure for the development of CATs*. Paper presented at the Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing. www.psych.umn.edu/psylabs/CATCentral/

Owen, R. J. (1975). A Bayesian Sequential Procedure for Quantal Response in the Context of Adaptive Mental Testing. *Journal of the American Statistical Association, 70*(350), 351-356. doi: 10.1080/01621459.1975.10479871

Parshall, C. G., Spray, J. A., Kalohn, J., & Davey, T. (2002). *Practical considerations in computer-based testing*: Springer.

Reckase, M. D. (1989). Adaptive Testing: The Evolution of a Good Idea. *Educational Measurement: Issues and Practice, 8*(3), 11-15. doi: 10.1111/j.1745-3992.1989.tb00326.x

Reckase, M. D. (2010). Designing item pools to optimize the functioning of a computerized adaptive test. *Psychological Test and Assessment Modeling, 52*(2), 127-141.

Revuelta, J., & Ponsoda, V. (1998). A Comparison of Item Exposure Control Methods in Computerized Adaptive Testing. *Journal of Educational Measurement, 35*(4), 311-327. doi: 10.1111/j.1745-3984.1998.tb00541.x

Sanders, P. (1992). Alternative solutions for optimization problems in generalizability theory. *Psychometrika, 57*(3), 351-356. doi: 10.1007/bf02295423

Sanders, P., Theunissen, T., & Baas, S. (1989). Minimizing the number of observations: A generalization of the spearman-brown formula. *Psychometrika, 54*(4), 587-598. doi: 10.1007/bf02296398

Sanders, P., Theunissen, T., & Baas, S. (1991). Maximizing the coefficient of generalizability under the constraint of limited resources. *Psychometrika, 56*(1), 87-96. doi: 10.1007/bf02294588

Spall, J. C. (2003). *Introduction to stochastic search and optimization: estimation, simulation, and control*: Wiley-Interscience.

Stocking, M. L. (1993). Controlling item exposure rates in a realistic adaptive testing paradigm *(ETS Research Report No. 93-2)*. Princeton, NJ: Educational Testing Service.

Stocking, M. L. (1995). Controlling item exposure conditional on ability in computerized adaptive testing. *(Research Report No. 95-24)*. Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lewis, C. (1995). A new method of controlling item exposure in computerized adaptive testing *(Research Report No. 95-25)*. Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lewis, C. (1998). Controlling Item Exposure Conditional on Ability in Computerized Adaptive Testing. *Journal of Educational and Behavioral Statistics, 23*(1), 57-75.

Stocking, M. L., & Lewis, C. (2000). Methods of Controlling the Exposure of Items in CAT. In W. J. Linden & G. A. W. Glas (Eds.), *Computerized Adaptive Testing: Theory and Practice* (pp. 163-182): Springer Netherlands.

Stocking, M. L., & Lewis, C. (2002). Methods of Controlling the Exposure of Items in CAT. In W. J. Linden & G. A. W. Glas (Eds.), *Computerized Adaptive Testing: Theory and Practice* (pp. 163-182): Springer Netherlands.

Theunissen, T. J. J. M. (1985). Binary programming and test design. *Psychometrika, 50*(4), 411-420. doi: 10.1007/bf02296260

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 101-133). Mahwah, NJ: Lawrence Erlbaum.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace Lines for Testlets: A Use of Multiple-Categorical-Response Models. *Journal of Educational Measurement, 26*(3), 247-260. doi: 10.1111/j.1745-3984.1989.tb00331.x

Thompson, N. A., & Weiss, D. J. (2011). A Framework for the Development of Computerized Adaptive Tests. *Practical Assessment, Research & Evaluation, 16*(1).

Urry, V. W. (1977). TAILORED TESTING: A SUCCESSFUL APPLICATION OF LATENT TRAIT THEORY. *Journal of Educational Measurement, 14*(2), 181-196. doi: 10.1111/j.1745-3984.1977.tb00035.x

van der Linden, W. J. (1998a). Bayesian item selection criteria for adaptive testing. *Psychometrika, 63*(2), 201-216. doi: 10.1007/bf02294775

van der Linden, W. J. (1998b). Optimal Assembly of Psychological and Educational Tests. *Applied Psychological Measurement, 22*(3), 195-211. doi: 10.1177/01466216980223001

van der Linden, W. J. (2002). Constrained Adaptive Testing with Shadow Tests. In W. J. Linden & G. A. W. Glas (Eds.), *Computerized Adaptive Testing: Theory and Practice* (pp. 27-52): Springer Netherlands.

van der Linden, W. J. (2003). Some Alternatives to Sympson-Hetter Item-Exposure Control in Computerized Adaptive Testing. *Journal of Educational and Behavioral Statistics, 28*(3), 249-265. doi: 10.3102/10769986028003249

van der Linden, W. J. (2005). A Comparison of Item-Selection Methods for Adaptive Tests with Content Constraints. *Journal of Educational Measurement, 42*(3), 283-302. doi: 10.1111/j.1745-3984.2005.00015.x

van der Linden, W. J. (2005a). Models for Adaptive Test Assembly *Linear Models for Optimal Test Design* (pp. 211-263): Springer New York.

van der Linden, W. J. (2005b). Solving Test-Assembly Problems *Linear Models for Optimal Test Design* (pp. 77-104): Springer New York.

van der Linden, W. J. (2010). Constrained Adaptive Testing with Shadow Tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of Adaptive Testing* (pp. 31-55): Springer New York.

van der Linden, W. J., & Adema, J. J. (1998). Simultaneous Assembly of Multiple Test Forms. *Journal of Educational Measurement, 35*(3), 185-198. doi: 10.1111/j.1745-3984.1998.tb00533.x

van der Linden, W. J., Ariel, A., & Veldkamp, B. P. (2006). Assembling a Computerized Adaptive Testing Item Pool as a Set of Linear Tests. *Journal of Educational and Behavioral Statistics, 31*(1), 81-99. doi: 10.3102/10769986031001081

van der Linden, W. J., & Chang, H. H. (2003). Implementing Content Constraints in Alpha-Stratified Adaptive Testing Using a Shadow Test Approach. *Applied Psychological Measurement, 27*(2), 107-120. doi: 10.1177/0146621602250531

van der Linden, W. J., & Diao, Q. (2011). Automated Test-Form Generation. *Journal of Educational Measurement, 48*(2), 206-222. doi: 10.1111/j.1745-3984.2011.00140.x

van der Linden, W. J., & Pashley, P. J. (2002). Item Selection and Ability Estimation in Adaptive Testing. In W. J. Linden & G. A. W. Glas (Eds.), *Computerized Adaptive Testing: Theory and Practice* (pp. 1-25): Springer Netherlands.

van der Linden, W. J., & Reese, L. M. (1998). A Model for Optimal Constrained Adaptive Testing. *Applied Psychological Measurement, 22*(3), 259-270. doi: 10.1177/01466216980223006

van der Linden, W. J., & Reese, L. M. (2001). A Model for Optimal Constrained Adaptive Testing *(Computerized Testing Report 97-07)*. Newtown, PA: Law School Admission Council.

van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining Item Exposure in Computerized Adaptive Testing With Shadow Tests. *Journal of Educational and Behavioral Statistics, 29*(3), 273-291. doi: 10.3102/10769986029003273

van der Linden, W. J., Veldkamp, B. P., & Reese, L. M. (2000). An Integer Programming Approach to Item Bank Design. *Applied Psychological Measurement, 24*(2), 139-150. doi: 10.1177/01466210022031570

Veldkamp, B. P. (2010). Bayesian item selection in constrained adaptive testing using shadow tests. *Psicologica, 31*(1), 149-169.

Vos, H. J., & Glas, G. A. W. (2002). Testlet-Based Adaptive Mastery Testing. In W. J. Linden & G. A. W. Glas (Eds.), *Computerized Adaptive Testing: Theory and Practice* (pp. 289-309): Springer Netherlands.

Wainer, H. (1989). The Future of Item Analysis. *Journal of Educational Measurement, 26*(2), 191-208.

Wainer, H. (1992). SOME PRACTICAL CONSIDERATIONS WHEN CONVERTING A LINEARLY ADMINISTERED TEST TO AN ADAPTIVE FORMAT. *ETS Research Report Series, 1992*(1), i-11. doi: 10.1002/j.2333-8504.1992.tb01445.x

Wainer, H., Bradlow, E. T., & Du, Z. (2002). Testlet Response Theory: An Analog for the 3PL Model Useful in Testlet-Based Adaptive Testing. In W. Linden & G. W. Glas (Eds.), *Computerized Adaptive Testing: Theory and Practice* (pp. 245-269): Springer Netherlands.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*: Cambridge University Press.

Wainer, H., & Wang, X. (2000). Using a New Statistical Model for Testlets to Score TOEFL. *Journal of Educational Measurement, 37*(3), 203-220. doi: 10.1111/j.1745-3984.2000.tb01083.x

Wang, T., & Kolen, M. J. (2001). Evaluating Comparability in Computerized Adaptive Testing: Issues, Criteria and an Example. *Journal of Educational Measurement, 38*(1), 19-49. doi: 10.1111/j.1745-3984.2001.tb01115.x

Wang, T., & Vispoel, W. P. (1998). Properties of Ability Estimation Methods in Computerized Adaptive Testing. *Journal of Educational Measurement, 35*(2), 109-135. doi: 10.1111/j.1745-3984.1998.tb00530.x

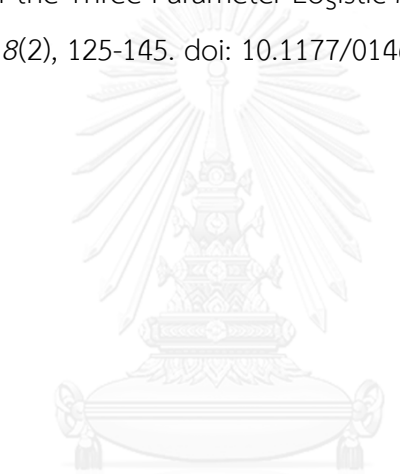
Wang, X., Bradlow, E. T., & Wainer, H. (2002). A GENERAL BAYESIAN MODEL FOR TESTLETS: THEORY AND APPLICATIONS. *ETS Research Report Series, 2002*(1), i-37. doi: 10.1002/j.2333-8504.2002.tb01869.x

Way, W. D. (1998). Protecting the Integrity of Computerized Testing Item Pools. *Educational Measurement: Issues and Practice, 17*(4), 17-27. doi: 10.1111/j.1745-3992.1998.tb00632.x

Weiss, D. J. (1985). Adaptive Testing by Computer. *Journal of Consulting and Clinical Psychology, 53*(6), 774-789.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of Computerized Adaptive Testing to Educational Problems. *Journal of Educational Measurement, 21*(4), 361-375. doi: 10.1111/j.1745-3984.1984.tb01040.x

Yen, W. M. (1984). Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model. *Applied Psychological Measurement, 8*(2), 125-145. doi: 10.1177/014662168400800201



ภาคผนวก

ภาคผนวก ก
ค่าพารามิเตอร์ของผู้สอบ

ตาราง ค่าต่ำสุด สูงสุด ค่าเฉลี่ย และค่าส่วนเบี่ยงเบนมาตรฐานของค่าความสามารถจริงของกลุ่มตัวอย่าง ที่ใช้กับคลังข้อสอบขนาด 600 ข้อ

กลุ่มตัวอย่าง	ความสามารถจริง (θ_i)				
	N	M	SD	Min	Max
t6tmf1	1000	0.058655	0.961625	-2.90634	3.097145
t6tmf2	1000	-0.01645	0.966518	-2.69676	3.387819
t6tmf3	1000	-0.00963	0.999124	-3.227	2.996641
t6tmf4	1000	-0.01579	1.019763	-3.19468	2.917513
t6tmf5	1000	0.015666	1.014175	-3.30617	3.765374
t6tmf6	1000	0.007038	0.988451	-3.06995	3.105933
t6tmf7	1000	-0.05344	1.009423	-3.73192	3.297128
t6tmf8	1000	-0.01964	1.017511	-3.24193	3.610747
t6tmf9	1000	0.000357	0.975448	-2.88338	2.9929
t6tmf10	1000	-0.01796	0.973265	-3.08038	3.302255
t6ran1	1000	0.004089	1.008899	-2.94335	3.365329
t6ran2	1000	0.019968	1.037545	-3.49871	3.128121
t6ran3	1000	0.004596	0.997797	-3.37897	3.039238
t6ran4	1000	-0.03075	1.010309	-3.95336	2.875124
t6ran5	1000	0.006769	1.020814	-3.24711	3.907841
t6ran6	1000	-0.00852	1.007125	-3.32089	3.042283
t6ran7	1000	-0.00191	1.009185	-3.33392	3.055734
t6ran8	1000	-0.05984	0.982787	-3.28169	2.882398
t6ran9	1000	0.010001	0.9936	-3.19866	3.307349
t6ran10	1000	0.076036	1.02025	-2.97053	3.452799
t6Cwa10.1	1000	-0.02769	1.004796	-3.19549	3.031839
t6Cwa10.2	1000	-0.03547	1.020362	-3.43375	3.402699
t6Cwa10.3	1000	0.009896	0.958362	-3.14874	3.310619
t6Cwa10.4	1000	0.026059	0.993672	-3.2111	3.322297
t6Cwa10.5	1000	0.001762	0.972315	-2.95491	3.635398
t6Cwa10.6	1000	-0.01603	0.975741	-2.8586	3.015234
t6Cwa10.7	1000	0.027897	0.996045	-3.59401	3.806045
t6Cwa10.8	1000	-0.01693	1.00866	-2.97246	3.074121
t6Cwa10.9	1000	0.012622	0.953871	-3.16168	3.439015
t6Cwa10.10	1000	0.035721	0.984941	-3.22667	3.054572

t6Cwa15.1	1000	-0.057	0.993673	-3.12198	3.065491
t6Cwa15.2	1000	-0.05296	0.998418	-3.53151	2.96261
t6Cwa15.3	1000	-0.04461	0.964651	-3.84544	2.587114
t6Cwa15.4	1000	-0.01339	0.96648	-2.86099	3.326844
t6Cwa15.5	1000	0.022806	1.038911	-2.88136	3.234539
t6Cwa15.6	1000	-0.00453	0.973871	-3.35597	2.614623
t6Cwa15.7	1000	-0.00311	1.009253	-3.25776	3.383438
t6Cwa15.8	1000	0.038842	1.041957	-3.25578	3.176711
t6Cwa15.9	1000	0.0296	1.042086	-3.45243	3.48968
t6Cwa15.10	1000	0.03973	1.040239	-3.56873	3.800983
t6Cwa20.1	1000	-0.00334	1.001314	-3.37719	2.883315
t6Cwa20.2	1000	-0.04154	1.038738	-3.43989	2.72726
t6Cwa20.3	1000	-0.03435	1.017918	-2.60793	3.534522
t6Cwa20.4	1000	0.017535	1.004942	-2.90163	3.46898
t6Cwa20.5	1000	-0.00767	0.965054	-3.25295	3.537395
t6Cwa20.6	1000	0.039202	1.021699	-3.06355	3.639147
t6Cwa20.7	1000	0.062305	0.989331	-3.04342	3.847813
t6Cwa20.8	1000	0.007606	1.000712	-3.48593	3.928846
t6Cwa20.9	1000	0.045123	0.972373	-2.78624	2.996036
t6Cwa20.10	1000	0.030932	0.989144	-2.46362	3.239538
t6Cwa25.1	1000	-0.01153	0.981983	-3.06492	2.927809
t6Cwa25.2	1000	-0.02929	0.987648	-3.15392	3.449513
t6Cwa25.3	1000	0.011842	1.016489	-3.70731	3.010666
t6Cwa25.4	1000	0.003237	0.992556	-3.86681	2.820192
t6Cwa25.5	1000	0.006901	0.988805	-2.93242	3.037529
t6Cwa25.6	1000	0.086369	0.999315	-2.92136	3.289384
t6Cwa25.7	1000	0.001294	0.99256	-3.13899	3.298267
t6Cwa25.8	1000	-0.02452	0.995167	-3.24387	2.820388
t6Cwa25.9	1000	-0.03605	1.02046	-3.92148	3.968082
t6Cwa25.10	1000	-0.04987	1.016992	-3.1836	3.09011
t6mcc10.1	1000	0.042368	0.959969	-3.27228	2.934557
t6mcc10.2	1000	-0.01419	0.997236	-3.0372	3.27622
t6mcc10.3	1000	-0.03159	0.985882	-2.98097	3.698512
t6mcc10.4	1000	-0.00369	0.963744	-3.25341	2.839842
t6mcc10.5	1000	-0.03497	1.011009	-3.32601	3.869984
t6mcc10.6	1000	-0.04642	0.98735	-3.01166	2.834364
t6mcc10.7	1000	0.018998	0.989927	-3.00258	3.557093
t6mcc10.8	1000	0.007211	1.041926	-3.16607	3.121243

t6mcc10.9	1000	-0.00882	0.992958	-3.20113	3.422663
t6mcc10.10	1000	0.034196	1.008714	-3.61537	2.978975
t6mcc15.1	1000	-0.00669	0.995901	-3.10555	2.849567
t6mcc15.2	1000	0.01523	0.977154	-2.8723	3.167309
t6mcc15.3	1000	0.02969	0.995786	-3.09101	3.279008
t6mcc15.4	1000	-0.00559	1.010855	-3.75002	3.554534
t6mcc15.5	1000	-0.01546	1.029028	-3.40598	3.237316
t6mcc15.6	1000	-0.00508	0.981739	-3.40121	3.264542
t6mcc15.7	1000	-0.01895	1.009926	-3.20095	3.275764
t6mcc15.8	1000	0.00016	0.982739	-3.60809	2.715813
t6mcc15.9	1000	-0.0406	1.019649	-3.36092	3.022366
t6mcc15.10	1000	-0.00657	0.984516	-2.62446	3.115895
t6mcc20.1	1000	-0.04799	0.981171	-3.62399	3.218905
t6mcc20.2	1000	0.024731	0.980032	-2.75904	3.076456
t6mcc20.3	1000	0.037073	1.024046	-3.5045	3.378422
t6mcc20.4	1000	-0.00666	0.950794	-3.1621	2.814078
t6mcc20.5	1000	-0.0141	0.99813	-3.18742	2.782905
t6mcc20.6	1000	0.011498	0.951262	-2.78576	2.708992
t6mcc20.7	1000	0.000554	0.99507	-3.45529	2.790523
t6mcc20.8	1000	-0.02099	0.960987	-3.06418	2.981012
t6mcc20.9	1000	0.014483	0.997127	-3.29176	3.371405
t6mcc20.10	1000	-0.02907	1.035599	-3.04073	3.364119
t6mcc25.1	1000	-0.01185	0.994672	-3.59589	2.681345
t6mcc25.2	1000	-0.03381	1.004864	-3.63341	3.491795
t6mcc25.3	1000	-0.03149	0.97466	-3.62402	3.209792
t6mcc25.4	1000	-0.00373	0.992734	-2.96539	2.596555
t6mcc25.5	1000	0.006977	0.992348	-2.88182	3.002983
t6mcc25.6	1000	0.000687	1.046722	-2.91123	3.294718
t6mcc25.7	1000	-0.00969	1.031128	-3.8678	3.170417
t6mcc25.8	1000	-0.01438	0.992695	-2.79903	3.185936
t6mcc25.9	1000	-0.02465	1.047287	-3.08247	3.379821
t6mcc25.10	1000	-0.06243	0.98011	-3.46344	2.626023

ตาราง แสดงค่าต่ำสุด สูงสุด ค่าเฉลี่ย และค่าส่วนเบี่ยงเบนมาตรฐานของค่าความสามารถจริงของกลุ่มตัวอย่าง ที่ใช้กับคลังข้อสอบขนาด 800 ข้อ

กลุ่มตัวอย่าง	ความสามารถจริง (θ_i)				
	N	M	SD	Min	Max
t8tmf1	1000	0.00827	0.97038	-3.26366	3.48311
t8tmf2	1000	0.00960	0.98022	-3.91093	3.78336
t8tmf3	1000	-0.00333	1.02815	-3.47224	3.14354
t8tmf4	1000	-0.01396	0.99138	-3.58375	3.36831
t8tmf5	1000	0.00580	1.00515	-3.53561	3.81383
t8tmf6	1000	0.00036	1.00078	-3.37048	3.34590
t8tmf7	1000	0.01222	1.00223	-3.95702	3.13438
t8tmf8	1000	-0.01775	1.00647	-3.71577	3.25775
t8tmf9	1000	0.00195	1.01783	-3.64218	2.86374
t8tmf10	1000	-0.01727	0.98557	-3.46280	3.59849
t8ran1	1000	-0.00439	0.98461	-3.37650	3.31977
t8ran2	1000	-0.04778	1.01566	-3.05027	3.28204
t8ran3	1000	0.04815	0.99711	-3.82028	3.59899
t8ran4	1000	-0.00742	0.96799	-2.96493	3.35050
t8ran5	1000	0.02386	0.98376	-3.05036	3.38661
t8ran6	1000	0.01861	0.97018	-2.88611	3.49343
t8ran7	1000	-0.04335	1.00186	-3.53333	3.42170
t8ran8	1000	-0.00766	1.02643	-3.32166	3.49866
t8ran9	1000	0.03894	1.02201	-3.00286	3.17674
t8ran10	1000	0.03135	1.02970	-3.32367	3.58819
t8Cwa10.1	1000	-0.03274	0.99026	-3.26283	3.25702
t8Cwa10.2	1000	0.03084	0.98535	-3.02673	3.21127
t8Cwa10.3	1000	-0.01669	1.00600	-3.51609	3.56128
t8Cwa10.4	1000	-0.02022	1.00505	-2.98985	3.36405
t8Cwa10.5	1000	-0.00755	1.03424	-3.24876	2.94514
t8Cwa10.6	1000	0.02982	1.03219	-3.33392	2.87025
t8Cwa10.7	1000	-0.01733	0.99499	-3.63374	3.31311
t8Cwa10.8	1000	-0.05928	1.01775	-3.19656	3.24463
t8Cwa10.9	1000	0.02510	0.97857	-3.73890	3.01804
t8Cwa10.10	1000	0.01272	1.02352	-3.16975	3.12391
t8Cwa15.1	1000	0.00242	1.01332	-2.87905	3.45047
t8Cwa15.2	1000	0.04703	0.98297	-3.63821	3.02305
t8Cwa15.3	1000	0.00052	1.04132	-3.15796	2.79714

t8Cwa15.4	1000	-0.01258	1.00135	-3.30848	3.07692
t8Cwa15.5	1000	0.01309	1.00577	-3.10085	3.13983
t8Cwa15.6	1000	-0.03207	0.98696	-3.46305	3.33948
t8Cwa15.7	1000	-0.02541	0.97831	-3.12828	2.76820
t8Cwa15.8	1000	-0.00706	1.00652	-3.10334	3.30262
t8Cwa15.9	1000	0.04922	1.03770	-3.16102	3.20785
t8Cwa15.10	1000	-0.05873	0.99704	-3.34331	3.23862
t8Cwa20.1	1000	-0.00501	1.01334	-3.38941	3.11029
t8Cwa20.2	1000	0.04488	0.98785	-3.69380	3.71375
t8Cwa20.3	1000	0.03171	1.01274	-3.16064	2.99171
t8Cwa20.4	1000	0.00550	1.01739	-2.95771	3.33233
t8Cwa20.5	1000	0.03557	1.03493	-3.55873	3.11403
t8Cwa20.6	1000	0.01764	0.99061	-3.32983	3.20790
t8Cwa20.7	1000	-0.00089	0.99866	-3.08770	2.79251
t8Cwa20.8	1000	0.00813	1.02452	-2.89401	3.13369
t8Cwa20.9	1000	-0.02710	1.00665	-3.73390	2.94243
t8Cwa20.10	1000	-0.00715	1.01635	-3.71136	3.35693
t8Cwa25.1	1000	-0.02049	1.02682	-2.96446	3.27223
t8Cwa25.2	1000	0.00166	1.00873	-3.28816	3.41389
t8Cwa25.3	1000	-0.07313	1.00078	-3.83436	3.85517
t8Cwa25.4	1000	0.01144	1.03078	-3.68136	3.52799
t8Cwa25.5	1000	-0.02781	0.98759	-3.25623	2.97731
t8Cwa25.6	1000	0.02019	1.00201	-3.04485	3.37249
t8Cwa25.7	1000	-0.01023	1.01094	-3.82117	3.11499
t8Cwa25.8	1000	-0.07375	0.97786	-2.83339	2.62925
t8Cwa25.9	1000	0.01231	1.03172	-2.75427	3.14027
t8Cwa25.10	1000	0.00718	0.94658	-2.99450	3.66701
t8mcc10.1	1000	-0.03768	1.00758	-3.55259	3.27884
t8mcc10.2	1000	-0.00433	1.05278	-3.35760	2.83381
t8mcc10.3	1000	-0.02153	0.99863	-3.05302	3.14327
t8mcc10.4	1000	-0.02695	1.03878	-3.45199	3.02581
t8mcc10.5	1000	0.00625	1.03450	-3.23311	3.01996
t8mcc10.6	1000	-0.01828	1.00274	-2.89516	2.92283
t8mcc10.7	1000	-0.03163	0.99922	-3.18021	2.86307
t8mcc10.8	1000	0.02134	1.02091	-2.82424	3.47494
t8mcc10.9	1000	0.01983	0.99826	-3.15999	3.14068
t8mcc10.10	1000	-0.03090	0.97686	-3.22989	3.35391
t8mcc15.1	1000	-0.00573	0.99841	-3.10390	3.29723

t8mcc15.2	1000	-0.03904	1.01287	-3.34147	2.85651
t8mcc15.3	1000	-0.02232	1.00838	-2.72291	3.35096
t8mcc15.4	1000	0.08580	1.02204	-3.00012	3.37460
t8mcc15.5	1000	0.03442	0.98576	-2.99971	3.32678
t8mcc15.6	1000	-0.06310	1.05009	-3.18104	3.19554
t8mcc15.7	1000	-0.01909	1.01684	-2.93009	2.88962
t8mcc15.8	1000	-0.00540	0.98380	-2.92601	3.21414
t8mcc15.9	1000	-0.02685	1.00427	-3.37163	3.15283
t8mcc15.10	1000	0.01756	1.00102	-2.68347	2.66458
t8mcc20.1	1000	0.04978	0.96076	-3.61662	3.00316
t8mcc20.2	1000	0.04580	1.01171	-3.77723	3.60987
t8mcc20.3	1000	0.01086	0.99449	-2.91896	2.85321
t8mcc20.4	1000	0.00264	0.96723	-2.71644	3.16439
t8mcc20.5	1000	-0.03237	0.99665	-3.38672	2.62031
t8mcc20.6	1000	0.00247	1.02394	-3.20127	3.30551
t8mcc20.7	1000	0.02832	0.98147	-3.24509	3.27873
t8mcc20.8	1000	0.01102	1.04728	-3.02703	3.21439
t8mcc20.9	1000	-0.01656	1.02650	-3.26565	3.02480
t8mcc20.10	1000	0.01027	1.03575	-3.22860	3.89727
t8mcc25.1	1000	0.00304	1.00900	-3.44745	3.01446
t8mcc25.2	1000	0.03298	1.00597	-3.34306	3.64978
t8mcc25.3	1000	-0.02489	1.03647	-3.47169	3.22033
t8mcc25.4	1000	0.02781	1.03799	-3.12788	3.51018
t8mcc25.5	1000	-0.02591	0.97286	-3.73707	3.50422
t8mcc25.6	1000	0.07812	1.00857	-3.43647	2.90955
t8mcc25.7	1000	0.01682	0.96688	-2.69357	3.12637
t8mcc25.8	1000	-0.01477	0.96540	-3.12206	3.16299
t8mcc25.9	1000	0.00523	1.01360	-3.93647	3.15113
t8mcc25.10	1000	-0.00294	0.98720	-2.82928	3.62457

ตาราง แสดงค่าต่ำสุด สูงสุด ค่าเฉลี่ย และค่าส่วนเบี่ยงเบนมาตรฐานของพารามิเตอร์อิทธิพลของแบบทดสอบย่อยของกลุ่มตัวอย่าง ที่ใช้กับคลังข้อสอบขนาด 600 ข้อ

กลุ่มตัวอย่าง	ความสามารถจริง (θ_i)				
	N	M	SD	Min	Max
t6tmf1	1000x150	-0.0005	0.99758	-4.5047	4.06700
t6tmf2	1000x150	0.00257	1.00162	-4.2760	4.18572
t6tmf3	1000x150	-0.0009	0.99898	-4.7323	4.29588
t6tmf4	1000x150	-0.0032	1.00051	-4.9187	4.19252
t6tmf5	1000x150	0.00223	0.99981	-4.1634	4.34865
t6tmf6	1000x150	0.00141	1.00267	-4.3154	4.5035
t6tmf7	1000x150	0.00435	1.00074	-4.5992	4.26729
t6tmf8	1000x150	0.00065	1.0015	-5.1407	4.38752
t6tmf9	1000x150	0.00043	1.00145	-4.5670	4.33433
t6tmf10	1000x150	-0.0007	0.99759	-4.8501	4.75292
t6ran1	1000x150	0.00205	1.00151	-4.6411	4.48408
t6ran2	1000x150	-0.0004	1.00125	-4.4047	4.37207
t6ran3	1000x150	6.83E-05	1.00008	-4.3273	4.0615
t6ran4	1000x150	-0.0002	1.00008	-4.8240	4.35718
t6ran5	1000x150	-0.001	1.00141	-4.4519	4.16712
t6ran6	1000x150	0.00414	0.99838	-4.6225	4.21001
t6ran7	1000x150	-0.0002	0.99628	-4.2976	4.24894
t6ran8	1000x150	-0.0011	0.99998	-4.3508	4.30116
t6ran9	1000x150	-0.0029	1.00624	-4.2474	4.25278
t6ran10	1000x150	0.00321	0.99699	-4.1536	4.46411
t6Cwa10.1	1000x150	0.00194	1.0031	-4.4874	4.50004
t6Cwa10.2	1000x150	0.00208	1	-4.4793	4.71791
t6Cwa10.3	1000x150	0.00103	0.99888	-4.1593	4.36196
t6Cwa10.4	1000x150	-0.0012	0.99904	-4.1491	4.12356
t6Cwa10.5	1000x150	0.00067	1.00104	-4.5133	4.65575
t6Cwa10.6	1000x150	0.00202	0.99932	-4.6870	4.29068
t6Cwa10.7	1000x150	-0.0017	1.00191	-4.3615	4.69063
t6Cwa10.8	1000x150	0.00022	1.00029	-4.2876	4.27452
t6Cwa10.9	1000x150	0.00163	1.00025	-4.5100	4.76222
t6Cwa10.10	1000x150	0.00583	0.99745	-4.5974	4.26506
t6Cwa15.1	1000x150	-0.0022	0.9995	-4.4100	4.5362
t6Cwa15.2	1000x150	0.00064	1.00295	-4.6111	4.9715
t6Cwa15.3	1000x150	-0.0006	0.99503	-4.3505	4.23575

t6Cwa15.4	1000x150	-0.0026	0.99831	-4.2738	4.5746
t6Cwa15.5	1000x150	-0.0034	1.00414	-4.2649	4.70716
t6Cwa15.6	1000x150	-0.0023	0.99848	-4.2588	4.54367
t6Cwa15.7	1000x150	-0.0009	1.00261	-4.2363	4.94954
t6Cwa15.8	1000x150	-0.0024	1.00012	-5.0406	4.21533
t6Cwa15.9	1000x150	-0.0023	1.00098	-4.3936	4.59914
t6Cwa15.10	1000x150	-0.0023	0.99896	-4.1059	4.95916
t6Cwa20.1	1000x150	0.00241	0.9994	-4.3045	5.09418
t6Cwa20.2	1000x150	-0.0007	0.99789	-4.4322	4.20917
t6Cwa20.3	1000x150	0.001	0.99751	-4.3966	4.13721
t6Cwa20.4	1000x150	0.00126	1.00116	-4.2807	4.28292
t6Cwa20.5	1000x150	-0.0048	1.00046	-4.6729	4.31448
t6Cwa20.6	1000x150	-0.003	0.99866	-4.1598	4.60172
t6Cwa20.7	1000x150	-0.0027	1.00029	-4.5211	4.51722
t6Cwa20.8	1000x150	0.00107	0.99923	-4.3346	4.44068
t6Cwa20.9	1000x150	0.00174	1.00032	-4.1873	4.17747
t6Cwa20.10	1000x150	0.00084	1.0014	-4.4196	4.55783
t6Cwa25.1	1000x150	-0.0013	1.00002	-4.6091	4.37648
t6Cwa25.2	1000x150	0.00244	0.99905	-4.3266	4.48451
t6Cwa25.3	1000x150	-0.0024	0.99937	-4.6778	4.43282
t6Cwa25.4	1000x150	0.00132	0.99704	-4.3537	4.84756
t6Cwa25.5	1000x150	-0.0031	1.00152	-4.734	4.24868
t6Cwa25.6	1000x150	0.00486	0.99686	-4.2102	4.22736
t6Cwa25.7	1000x150	-0.0051	0.99978	-4.4279	4.39508
t6Cwa25.8	1000x150	-0.0013	0.99837	-4.3196	4.67646
t6Cwa25.9	1000x150	-0.0015	1.00134	-4.4862	4.44122
t6Cwa25.10	1000x150	0.00162	0.99882	-4.4857	4.43378
t6mccCAT10.1	1000x150	0.00153	0.9994	-4.8709	4.19628
t6mccCAT10.2	1000x150	0.00526	0.99872	-4.4086	4.31871
t6mcc10.3	1000x150	-0.0035	0.99825	-4.2286	4.60080
t6mcc10.4	1000x150	0.00666	0.99935	-4.4947	4.46123
t6mcc10.5	1000x150	-0.0026	1.00274	-4.2874	4.57607
t6mcc10.6	1000x150	0.00025	1.00196	-5.1986	4.27314
t6mcc10.7	1000x150	-0.0029	1.00214	-3.9738	4.39201
t6mcc10.8	1000x150	0.00512	0.99968	-4.2022	4.46132
t6mcc10.9	1000x150	-0.0003	0.99855	-4.3591	4.48919
t6mcc10.10	1000x150	0.00266	1.00137	-4.5212	4.48255
t6mcc15.1	1000x150	0.0058	1.00058	-4.3879	4.46961

t6mcc15.2	1000x150	-0.0016	0.99896	-4.5266	4.33149
t6mcc15.3	1000x150	0.0006	0.99824	-4.0448	4.22375
t6mcc15.4	1000x150	-0.0039	1.00178	-4.6936	4.47361
t6mcc15.5	1000x150	-0.0065	0.99876	-4.4163	4.24847
t6mcc15.6	1000x150	0.00558	0.9979	-4.3113	4.38843
t6mcc15.7	1000x150	0.00081	0.99984	-4.2445	4.1515
t6mcc15.8	1000x150	0.00311	1.00001	-4.7511	4.39968
t6mcc15.9	1000x150	0.00472	0.99864	-4.7489	4.26523
t6mcc15.10	1000x150	0.00033	1.00127	-4.2315	4.30682
t6mcc20.1	1000x150	-0.0023	1.00072	-4.2955	5.20761
t6mcc20.2	1000x150	0.00081	1.00356	-4.0702	4.22597
t6mcc20.3	1000x150	0.00471	0.99931	-5.3006	5.526800
t6mcc20.4	1000x150	0.00076	1.00284	-4.4255	4.29962
t6mcc20.5	1000x150	-0.0041	0.99908	-4.5002	4.45724
t6mcc20.6	1000x150	-0.0034	0.99857	-4.1409	4.29406
t6mcc20.7	1000x150	0.00294	0.99673	-4.6921	4.29806
t6mcc20.8	1000x150	-0.0046	0.99792	-4.5521	4.24135
t6mcc20.9	1000x150	-0.0011	1.00201	-4.9238	4.04267
t6mcc20.10	1000x150	0.00289	1.00015	-4.4671	4.6466
t6mcc25.1	1000x150	0.0024	0.99904	-4.6543	4.79903
t6mcc25.2	1000x150	0.00616	0.99845	-4.2339	4.24767
t6mcc25.3	1000x150	-0.0051	0.99795	-4.4451	4.49620
t6mcc25.4	1000x150	0.00029	0.9971	-4.4657	4.86421
t6mcc25.5	1000x150	-0.0017	0.99778	-4.5661	5.07597
t6mcc25.6	1000x150	-0.0008	1.00296	-4.0561	4.51801
t6mcc25.7	1000x150	-0.0021	0.9963	-4.5565	4.47695
t6mcc25.8	1000x150	-0.0024	0.99851	-4.5972	4.64254
t6mcc25.9	1000x150	0.00015	1.00308	-4.2433	4.37818
t6mcc25.10	1000x150	-0.0017	0.99833	-4.3227	4.40411

ตารางที่ ตัวอย่างค่าต่ำสุด สูงสุด ค่าเฉลี่ย และค่าส่วนเบี่ยงเบนมาตรฐานของพารามิเตอร์อิทธิพลของแบบทดสอบย่อย (θ_i) ของกลุ่มตัวอย่าง ที่ใช้กับคลังข้อสอบขนาด 800 ข้อ

กลุ่มตัวอย่าง		ความสามารถจริง (θ_i)			
	N	M	SD	Min	Max
t8tmf1	1000x200	0.00097	0.999878	-4.6778	4.493154
t8tmf2	1000x200	0.002125	0.999143	-4.69703	4.352850
t8tmf3	1000x200	0.001255	1.000158	-4.79344	4.763630
t8tmf4	1000x200	0.001509	1.000115	-4.34220	5.065057
t8tmf5	1000x200	-0.000948	0.999866	-4.37160	4.467179
t8tmf6	1000x200	0.001356	1.000897	-4.43981	4.660737
t8tmf7	1000x200	0.002275	1.001981	-4.77543	5.011751
t8tmf8	1000x200	0.000325	0.999427	-4.59607	4.497998
t8tmf9	1000x200	-0.000579	1.000337	-4.76556	4.719067
t8tmf10	1000x200	-8.04E-05	1.001491	-4.91672	4.735025
t8ran1	1000x200	0.001548	0.998508	-4.25069	4.719352
t8ran2	1000x200	-0.00083	1.000486	-4.3615	4.690633
t8ran3	1000x200	0.002523	1.000806	-4.79618	4.577841
t8ran4	1000x200	-0.000111	0.998368	-4.59395	4.660628
t8ran5	1000x200	-0.000843	0.99959	-4.44071	4.543669
t8ran6	1000x200	0.00047	1.000357	-4.48735	4.599135
t8ran7	1000x200	-0.00128	0.99895	-4.61728	4.705622
t8ran8	1000x200	0.000192	1.001496	-4.72891	4.31821
t8ran9	1000x200	-0.00098	1.000257	-4.52114	4.57154
t8ran10	1000x200	0.002257	1.001149	-5.16505	4.557829
t8Cwa10.1	1000x200	0.001424	1.003152	-4.48740	4.738351
t8Cwa10.2	1000x200	-0.000666	0.998417	-4.85824	4.731317
t8Cwa10.3	1000x200	-0.001858	0.998988	-4.14906	4.282637
t8Cwa10.4	1000x200	0.002679	0.999656	-4.62405	4.789446
t8Cwa10.5	1000x200	0.002251	0.999973	-4.82990	4.239003
t8Cwa10.6	1000x200	-0.000559	1.000856	-4.45862	4.274521
t8Cwa10.7	1000x200	0.001205	1.001261	-4.37956	4.196546
t8Cwa10.8	1000x200	0.002835	1.001009	-4.56607	4.317035
t8Cwa10.9	1000x200	0.005145	0.998619	-4.59512	4.305565
t8Cwa10.10	1000x200	-1.54E-06	0.998797	-4.59395	4.660628
t8Cwa15.1	1000x200	-0.003363	0.998987	-4.27378	4.574597
t8Cwa15.2	1000x200	0.000842	1.001688	-4.35731	4.418685
t8Cwa15.3	1000x200	-0.000669	1.000606	-5.04485	4.444578

t8Cwa15.4	1000x200	-0.000741	1.000635	-5.04055	4.240995
t8Cwa15.5	1000x200	0.003864	0.99996	-4.48735	4.424205
t8Cwa15.6	1000x200	0.003186	0.999409	-4.63183	4.84592
t8Cwa15.7	1000x200	-0.00187	0.999575	-4.39991	4.705622
t8Cwa15.8	1000x200	0.003957	0.999816	-5.35509	4.384146
t8Cwa15.9	1000x200	-0.000838	0.999048	-4.70495	4.446981
t8Cwa15.10	1000x200	-0.002916	1.000669	-4.67288	4.314475
t8Cwa20.1	1000x200	0.001386	0.997209	-4.88768	4.821648
t8Cwa20.2	1000x200	-0.002183	1.003332	-4.26489	4.707157
t8Cwa20.3	1000x200	0.001051	1.00024	-5.04485	4.733173
t8Cwa20.4	1000x200	-0.000369	1.002639	-4.43081	4.531311
t8Cwa20.5	1000x200	-0.001766	1.000736	-4.3936	4.599135
t8Cwa20.6	1000x200	-0.000281	0.999226	-4.63183	4.978731
t8Cwa20.7	1000x200	-0.001323	0.997143	-4.3706	4.174271
t8Cwa20.8	1000x200	-0.001907	0.998845	-4.43217	4.720729
t8Cwa20.9	1000x200	0.000546	1.001504	-4.26890	4.287915
t8Cwa20.10	1000x200	0.002697	0.999777	-4.54966	4.198021
t8Cwa25.1	1000x200	0.001404	1.001026	-4.4874	4.500042
t8Cwa25.2	1000x200	-0.001479	0.997496	-4.30352	4.610346
t8Cwa25.3	1000x200	9.10E-05	0.998572	-4.43288	4.282637
t8Cwa25.4	1000x200	-0.001608	1.001228	-4.47326	4.655747
t8Cwa25.5	1000x200	0.004112	0.998539	-4.68705	4.220041
t8Cwa25.6	1000x200	-0.000482	0.998069	-4.27083	4.136346
t8Cwa25.7	1000x200	4.26E-05	1.00076	-4.50998	4.762215
t8Cwa25.8	1000x200	-0.000144	1.001246	-4.79618	4.317035
t8Cwa25.9	1000x200	0.000496	0.999649	-4.40997	4.536201
t8Cwa25.10	1000x200	7.88E-05	0.998681	-4.23554	4.660628
t8mcc10.1	1000x200	-0.002623	1.001021	-5.21801	4.78714
t8mcc10.2	1000x200	-0.000471	0.998389	-4.53231	4.81074
t8mcc10.3	1000x200	0.005336	0.999529	-4.37326	4.558704
t8mcc10.4	1000x200	0.000506	0.998263	-4.28428	4.648809
t8mcc10.5	1000x200	-0.003251	0.997571	-4.21311	4.480912
t8mcc10.6	1000x200	0.005206	1.000147	-4.74624	4.360413
t8mcc10.7	1000x200	0.000322	1.000466	-4.24633	4.651585
t8mcc10.8	1000x200	0.00568	0.999702	-4.72046	4.212393
t8mcc10.9	1000x200	-0.003255	1.00034	-4.76542	4.761172
t8mcc10.10	1000x200	0.001613	0.999357	-4.5419	4.70524
t8mcc15.1	1000x200	0.001933	0.998183	-5.01329	4.287699

t8mcc15.2	1000x200	-0.000615	0.997904	-4.16310	4.592001
t8mcc15.3	1000x200	-0.003446	0.997012	-4.71234	4.423295
t8mcc15.4	1000x200	0.009046	0.999507	-4.50066	4.186377
t8mcc15.5	1000x200	0.003095	1.001442	-4.70544	4.55581
t8mcc15.6	1000x200	-0.004392	0.998974	-4.25779	4.393707
t8mcc15.7	1000x200	0.004383	1.002376	-4.68861	4.695451
t8mcc15.8	1000x200	-0.002012	1.000928	-4.07509	4.377112
t8mcc15.9	1000x200	-0.001236	0.999858	-4.30036	4.171017
t8mcc15.10	1000x200	0.000176	0.997122	-4.80000	4.778189
t8mcc20.1	1000x200	0.005204	1.000339	-4.26383	4.843073
t8mcc20.2	1000x200	0.000913	1.001794	-4.44908	4.738039
t8mcc20.3	1000x200	0.000884	0.999171	-4.51555	4.829784
t8mcc20.4	1000x200	0.000654	1.004226	-4.63466	4.408685
t8mcc20.5	1000x200	0.002206	1.000378	-4.17392	4.602412
t8mcc20.6	1000x200	-0.000583	1.00189	-4.70468	5.061831
t8mcc20.7	1000x200	0.000209	0.998499	-4.61967	4.207201
t8mcc20.8	1000x200	0.000741	0.999664	-4.43543	5.199861
t8mcc20.9	1000x200	0.001163	0.99862	-4.21941	4.374613
t8mcc20.10	1000x200	-0.00069	0.999737	-4.94947	4.461089
t8mcc25.1	1000x200	-0.002268	0.996493	-4.34769	4.131303
t8mcc25.2	1000x200	-0.002985	1.000402	-4.23705	4.614571
t8mcc25.3	1000x200	0.002849	1.001576	-4.53493	4.83372
t8mcc25.4	1000x200	-0.000144	0.999872	-4.27969	4.393313
t8mcc25.5	1000x200	-0.0002	0.999078	-4.65562	4.944294
t8mcc25.6	1000x200	0.003436	0.999641	-4.22807	4.690055
t8mcc25.7	1000x200	0.001734	0.999869	-4.38036	4.821316
t8mcc25.8	1000x200	-0.002169	1.001155	-4.45237	4.324817
t8mcc25.9	1000x200	0.002167	0.999557	-5.10571	4.531911
t8mcc25.10	1000x200	-0.001977	1.00044	-4.46218	4.214635

ภาคผนวก ข

ผลการตรวจสอบการจำลองข้อมูลผลการตอบด้วยสถิติ Q3

ตาราง ค่าเฉลี่ยของสถิติ Q3 ของแบบสอบย่อยแต่ละชุดในคลังข้อสอบขนาด 600 และ 800 ข้อ

Testlet	No. of Items	600 items		800 items	
		Mean testlet Q3	Mean Q3	Mean testlet Q3	Mean Q3
t001.	4	0.180231	0.000649	0.151828	0.000382
t002.	4	0.125943	0.000649	0.138806	0.000382
t003.	4	0.109221	0.000649	0.085033	0.000382
t004.	4	0.149904	0.000649	0.113988	0.000382
t005.	4	0.141079	0.000649	0.084486	0.000382
t006.	4	0.077951	0.000649	0.156605	0.000382
t007.	4	0.122214	0.000649	0.194948	0.000382
t008.	4	0.132924	0.000649	0.065466	0.000382
t009.	4	0.140112	0.000649	0.121586	0.000382
t010.	4	0.134323	0.000649	0.108492	0.000382
t011.	4	0.123138	0.000649	0.132525	0.000382
t012.	4	0.129881	0.000649	0.106665	0.000382
t013.	4	0.153638	0.000649	0.152207	0.000382
t014.	4	0.077801	0.000649	0.109910	0.000382
t015.	4	0.180537	0.000649	0.226232	0.000382
t016.	4	0.164495	0.000649	0.129338	0.000382
t017.	4	0.148688	0.000649	0.085879	0.000382
t018.	4	0.178438	0.000649	0.076506	0.000382
t019.	4	0.096354	0.000649	0.120169	0.000382
t020.	4	0.107511	0.000649	0.126150	0.000382
t021.	4	0.194250	0.000649	0.091493	0.000382
t022.	4	0.102563	0.000649	0.114431	0.000382
t023.	4	0.148473	0.000649	0.174262	0.000382
t024.	4	0.109182	0.000649	0.178291	0.000382
t025.	4	0.090276	0.000649	0.141214	0.000382
t026.	4	0.164985	0.000649	0.071460	0.000382
t027.	4	0.290944	0.000649	0.075666	0.000382

Testlet	No. of Items	600 items		800 items	
		Mean testlet Q3	Mean Q3	Mean testlet Q3	Mean Q3
t028.	4	0.156609	0.000649	0.196190	0.000382
t029.	4	0.157339	0.000649	0.185293	0.000382
t030.	4	0.182885	0.000649	0.201864	0.000382
t031.	4	0.113441	0.000649	0.138226	0.000382
t032.	4	0.105621	0.000649	0.147761	0.000382
t033.	4	0.140316	0.000649	0.142603	0.000382
t034.	4	0.103128	0.000649	0.119812	0.000382
t035.	4	0.110913	0.000649	0.091459	0.000382
t036.	4	0.104444	0.000649	0.092980	0.000382
t037.	4	0.159692	0.000649	0.101914	0.000382
t038.	4	0.099240	0.000649	0.160436	0.000382
t039.	4	0.127818	0.000649	0.075149	0.000382
t040.	4	0.163220	0.000649	0.066562	0.000382
t041.	4	0.144769	0.000649	0.166177	0.000382
t042.	4	0.057776	0.000649	0.093879	0.000382
t043.	4	0.132602	0.000649	0.103676	0.000382
t044.	4	0.160620	0.000649	0.194655	0.000382
t045.	4	0.076268	0.000649	0.146097	0.000382
t046.	4	0.167773	0.000649	0.058342	0.000382
t047.	4	0.080109	0.000649	0.081620	0.000382
t048.	4	0.107954	0.000649	0.124959	0.000382
t049.	4	0.171106	0.000649	0.118963	0.000382
t050.	4	0.102580	0.000649	0.061519	0.000382
t051.	4	0.154828	0.000649	0.163901	0.000382
t052.	4	0.072160	0.000649	0.184487	0.000382
t053.	4	0.149577	0.000649	0.157000	0.000382
t054.	4	0.159149	0.000649	0.117263	0.000382
t055.	4	0.143289	0.000649	0.166549	0.000382
t056.	4	0.076532	0.000649	0.133923	0.000382
t057.	4	0.079629	0.000649	0.102461	0.000382
t058.	4	0.170778	0.000649	0.161119	0.000382

Testlet	No. of Items	600 items		800 items	
		Mean testlet Q3	Mean Q3	Mean testlet Q3	Mean Q3
t059.	4	0.154573	0.000649	0.195435	0.000382
t060.	4	0.109552	0.000649	0.089173	0.000382
t061.	4	0.136027	0.000649	0.151146	0.000382
t062.	4	0.119889	0.000649	0.131207	0.000382
t063.	4	0.110804	0.000649	0.106323	0.000382
t064.	4	0.176591	0.000649	0.083891	0.000382
t065.	4	0.105492	0.000649	0.081724	0.000382
t066.	4	0.116362	0.000649	0.151117	0.000382
t067.	4	0.099215	0.000649	0.098637	0.000382
t068.	4	0.078796	0.000649	0.135317	0.000382
t069.	4	0.112612	0.000649	0.103315	0.000382
t070.	4	0.183025	0.000649	0.123799	0.000382
t071.	4	0.178828	0.000649	0.096985	0.000382
t072.	4	0.126184	0.000649	0.130499	0.000382
t073.	4	0.122511	0.000649	0.184478	0.000382
t074.	4	0.107914	0.000649	0.123725	0.000382
t075.	4	0.116762	0.000649	0.103224	0.000382
t076.	4	0.153646	0.000649	0.124159	0.000382
t077.	4	0.125873	0.000649	0.160164	0.000382
t078.	4	0.143802	0.000649	0.132711	0.000382
t079.	4	0.133138	0.000649	0.110362	0.000382
t080.	4	0.190496	0.000649	0.126972	0.000382
t081.	4	0.154265	0.000649	0.166956	0.000382
t082.	4	0.168662	0.000649	0.192090	0.000382
t083.	4	0.095987	0.000649	0.153362	0.000382
t084.	4	0.121398	0.000649	0.123268	0.000382
t085.	4	0.131763	0.000649	0.136001	0.000382
t086.	4	0.094928	0.000649	0.199977	0.000382
t087.	4	0.188724	0.000649	0.110250	0.000382
t088.	4	0.139317	0.000649	0.107785	0.000382
t089.	4	0.071976	0.000649	0.094856	0.000382

Testlet	No. of Items	600 items		800 items	
		Mean testlet Q3	Mean Q3	Mean testlet Q3	Mean Q3
t090.	4	0.170645	0.000649	0.085186	0.000382
t091.	4	0.131142	0.000649	0.103755	0.000382
t092.	4	0.076400	0.000649	0.131910	0.000382
t093.	4	0.074956	0.000649	0.132755	0.000382
t094.	4	0.177647	0.000649	0.172776	0.000382
t095.	4	0.189177	0.000649	0.176885	0.000382
t096.	4	0.137538	0.000649	0.072312	0.000382
t097.	4	0.168832	0.000649	0.116863	0.000382
t098.	4	0.105879	0.000649	0.116671	0.000382
t099.	4	0.193380	0.000649	0.083530	0.000382
t100.	4	0.114658	0.000649	0.259315	0.000382
t101.	4	0.106433	0.000649	0.077887	0.000382
t102.	4	0.065806	0.000649	0.060296	0.000382
t103.	4	0.099759	0.000649	0.113914	0.000382
t104.	4	0.140619	0.000649	0.123472	0.000382
t105.	4	0.155770	0.000649	0.155609	0.000382
t106.	4	0.134867	0.000649	0.098361	0.000382
t107.	4	0.087397	0.000649	0.118418	0.000382
t108.	4	0.115014	0.000649	0.132991	0.000382
t109.	4	0.148916	0.000649	0.100377	0.000382
t110.	4	0.190052	0.000649	0.171770	0.000382
t111.	4	0.125711	0.000649	0.127033	0.000382
t112.	4	0.172674	0.000649	0.143509	0.000382
t113.	4	0.073126	0.000649	0.090900	0.000382
t114.	4	0.110687	0.000649	0.119109	0.000382
t115.	4	0.187893	0.000649	0.123342	0.000382
t116.	4	0.111870	0.000649	0.155548	0.000382
t117.	4	0.135322	0.000649	0.114773	0.000382
t118.	4	0.166999	0.000649	0.083565	0.000382
t119.	4	0.105147	0.000649	0.173645	0.000382
t120.	4	0.138288	0.000649	0.114579	0.000382

Testlet	No. of Items	600 items		800 items	
		Mean testlet Q3	Mean Q3	Mean testlet Q3	Mean Q3
t121.	4	0.149534	0.000649	0.133048	0.000382
t122.	4	0.096770	0.000649	0.091346	0.000382
t123.	4	0.126014	0.000649	0.109576	0.000382
t124.	4	0.122111	0.000649	0.089944	0.000382
t125.	4	0.105944	0.000649	0.094059	0.000382
t126.	4	0.064516	0.000649	0.101874	0.000382
t127.	4	0.119397	0.000649	0.107945	0.000382
t128.	4	0.144185	0.000649	0.039004	0.000382
t129.	4	0.112486	0.000649	0.061514	0.000382
t130.	4	0.122123	0.000649	0.114278	0.000382
t131.	4	0.183182	0.000649	0.149863	0.000382
t132.	4	0.168629	0.000649	0.163860	0.000382
t133.	4	0.146278	0.000649	0.171132	0.000382
t134.	4	0.154546	0.000649	0.193125	0.000382
t135.	4	0.120527	0.000649	0.138163	0.000382
t136.	4	0.122285	0.000649	0.099053	0.000382
t137.	4	0.097217	0.000649	0.166574	0.000382
t138.	4	0.197376	0.000649	0.133879	0.000382
t139.	4	0.213392	0.000649	0.087913	0.000382
t140.	4	0.111120	0.000649	0.149783	0.000382
t141.	4	0.159489	0.000649	0.078071	0.000382
t142.	4	0.160397	0.000649	0.130061	0.000382
t143.	4	0.128414	0.000649	0.073228	0.000382
t144.	4	0.137708	0.000649	0.079587	0.000382
t145.	4	0.093292	0.000649	0.134731	0.000382
t146.	4	0.122932	0.000649	0.125868	0.000382
t147.	4	0.126052	0.000649	0.177607	0.000382
t148.	4	0.162851	0.000649	0.197210	0.000382
t149.	4	0.094399	0.000649	0.161602	0.000382
t150.	4	0.075981	0.000649	0.188914	0.000382
t151.	4	-	-	0.101654	0.000382

Testlet	No. of Items	600 items		800 items	
		Mean testlet Q3	Mean Q3	Mean testlet Q3	Mean Q3
t152.	4	-	-	0.118699	0.000382
t153.	4	-	-	0.084852	0.000382
t154.	4	-	-	0.105311	0.000382
t155.	4	-	-	0.127617	0.000382
t156.	4	-	-	0.168176	0.000382
t157.	4	-	-	0.093171	0.000382
t158.	4	-	-	0.152171	0.000382
t159.	4	-	-	0.118531	0.000382
t160.	4	-	-	0.119864	0.000382
t161.	4	-	-	0.072998	0.000382
t162.	4	-	-	0.120490	0.000382
t163.	4	-	-	0.152059	0.000382
t164.	4	-	-	0.085952	0.000382
t165.	4	-	-	0.098326	0.000382
t166.	4	-	-	0.074659	0.000382
t167.	4	-	-	0.160659	0.000382
t168.	4	-	-	0.114491	0.000382
t169.	4	-	-	0.110131	0.000382
t170.	4	-	-	0.081543	0.000382
t171.	4	-	-	0.165469	0.000382
t172.	4	-	-	0.066760	0.000382
t173.	4	-	-	0.170771	0.000382
t174.	4	-	-	0.145842	0.000382
t175.	4	-	-	0.127435	0.000382
t176.	4	-	-	0.146489	0.000382
t177.	4	-	-	0.169641	0.000382
t178.	4	-	-	0.093934	0.000382
t179.	4	-	-	0.104632	0.000382
t180.	4	-	-	0.118677	0.000382
t181.	4	-	-	0.132861	0.000382
t182.	4	-	-	0.090035	0.000382

Testlet	No. of Items	600 items		800 items	
		Mean testlet Q3	Mean Q3	Mean testlet Q3	Mean Q3
t183.	4	-	-	0.135947	0.000382
t184.	4	-	-	0.141039	0.000382
t185.	4	-	-	0.082716	0.000382
t186.	4	-	-	0.121412	0.000382
t187.	4	-	-	0.214557	0.000382
t188.	4	-	-	0.137750	0.000382
t189.	4	-	-	0.061695	0.000382
t190.	4	-	-	0.205101	0.000382
t191.	4	-	-	0.107900	0.000382
t192.	4	-	-	0.106189	0.000382
t193.	4	-	-	0.115321	0.000382
t194.	4	-	-	0.135767	0.000382
t195.	4	-	-	0.177006	0.000382
t196.	4	-	-	0.133290	0.000382
t197.	4	-	-	0.104808	0.000382
t198.	4	-	-	0.094584	0.000382
t199.	4	-	-	0.159557	0.000382
t200.	4	-	-	0.203160	0.000382

ภาคผนวก ค
เครื่องมือที่ใช้ในการวิจัย

คำสั่งการสร้างพารามิเตอร์ข้อสอบ

```
#*****************************************************************************************
# generate true item parameters v.02
#*****************************************************************************************
gen.item.v.02 <-function(ni,ipt,str=4,K=4) {
    library("gdata")
    #library("truncnorm")
    #set.seed(seed);
    strata=str
    ntestlets=ni/ipt
    c1<-ifelse(runif(ntestlets)>.65,1,0)
    c2<-numeric(ntestlets)
    c3<-numeric(ntestlets)
    c4<-rep(1,ntestlets)
    for(i in 1:ntestlets){
        c2[i] <-ifelse(runif(1)>.5 && c1[i]==0,1,0)
        c3[i] <-ifelse(c1[i]==0 && c2[i]==0,1,0)
    }
    CC <-cbind(c1,c2,c3,c4)
    C.C <- matrix( CC, nrow=ntestlets , ncol=K)
    CA <- matrix(NA, nrow=ntestlets , ncol=1)
    CA[,1] <- C.C[,1]*1 + C.C[,2]*2 + C.C[,3]*3
    ncc <- length(unique(CA));
    b <- rnorm(ni);
    a <- rlnorm( ni , 0.02 , .22 )
    c <- rbeta( ni , 2 , 20 )
    i.id <- seq(1:ni);
    order.b <- order(b);
    i.index <- resample(order.b);
    b.str <- matrix(b, ni/strata, strata);
    order.b.str <- matrix(i.index, nrow = ni/strata, ncol = strata, byrow=F)
```

```
    sq.b.M <- matrix(NA, ni/strata, strata);
    sum.b.testlets <- matrix(0, ntestlets/strata, strata);
    for (s in 1:strata)
        for (i in 1:ni/strata)
            sq.b.M[i,s] <- b[order.b.str[i,s]]
    ### sum b-parameter each testlet
    for (ss in 1:strata){
        i=0
        for (tt in 1:(ntestlets/strata) )
            for (ii in 1:ipt)
            {
                i=i+1
                sum.b.testlets[tt,ss] <- sum.b.testlets[tt,ss] + sq.b.M[i,ss]
            }
    }
    mean.b.testlets <- sum.b.testlets/ipt
        testlet.id <- rep(1:ntestlets , each=ipt , len = ni)
        index.T <- matrix(c(testlet.id, i.index), nrow = ni, ncol = 2, byrow = F,
                          dimnames = list(c(),c("id.testlet", "items")) ) #items = order of b value
        content.cat <- CA
    res<- list( "A" = a , "B" = b , "C" = c , "testlets" = index.T, "mean.b" = mean.b.testlets, "C.C"=C.C,
"CA"=CA)
    return(res);
}
```


คำสั่งการสร้างแบบแผนการตอบของผู้สอบ

```
#*************************************************************************************************************
# calculate P_i(theta) for 3PL TRT model
#*************************************************************************************************************
.prob.3pltrt <- function( theta , gammaM , b , a , c){
    #set.seed(seed);
    l1  <- rep( 1, length(theta) )
    cM <- outer( l1 , c )
    aM <- outer( l1 , a )
```

```
    bM <- outer( l1 , b )
    thetaM <- outer( theta , rep(1,length(b) ) )
    #gammaM <- gamma[ , rep(1:ntestlets,each=ipt) ]
    pp<-cM + (1-cM) * plogis( aM*(thetaM - bM - gammaM) )
}

#*****************************************************************************************
# generate response data  (mod1)
#*****************************************************************************************
gen.Resp<-function( n , itbank ) {
    #set.seed(seed);
    A <- itbank$A;
    B <- itbank$B;
    C <- itbank$C;
    ntestlets <- length(unique(itbank$testlets[,1]))
    ni <- length(B);
    ipt <- ni/ntestlets;
    person.para <- gen.persons(n,sdt=1,ntestlets,ipt);
    gammaM <- person.para$gamma;
    theta <- person.para$true.theta;
    n <- length(theta);

    #name <- seq(1:ni);
    if(ipt > "1"){""
        prob <- .prob.3pltrt( theta , gammaM , B , A , C);
        resp <- (matrix( runif(n*ni) , n , ni) < prob )*1
        c.names <-paste("pa", itbank$testlets[,1],sep="");
        colnames(resp) <- paste( c.names, itbank$testlets[,2],sep="__i");
    } else {
        resp<-matrix(1,n,ni);
        pp<-array(0,c(n,ni,2));
        for (i in 1:ni) {
            pp[,i,2]<-C[i]+(1-C[i])/(1+exp(-D*A[i]*(theta-B[i]-gammaM[i])));
            pp[,i,1]<-1-pp[,i,2];
        }
```

```
        random<-matrix(runif(n*ni),n,ni);
        resp<-ifelse(pp[,,2]>random,1,0);
        colnames(resp) <- paste( "R" , 1:ni , sep="");
    }
    #colnames(resp) <- paste( "R" , 1:ni , sep="");
    row.names(resp) <- paste(1:n)
    res <- list( "resp" = resp , "theta" = theta , "gamma" = gammaM);
    return(res);
}
```

คำสั่งของการประมาณค่าความสามารถของผู้สอบ

```
#*****************************************************************************************
# prep.prob.info
#*****************************************************************************************
prep.prob.info<-function(){
    pp<-array(0,c(nq,ni,2));
    matrix.info<-matrix(0,nq,ni);
    if (ipt > "1"){
        for (i in 1:ni) {
            pp[,i,2]<-C[i]+(1-C[i])/(1+exp(-D*A[i]*(theta-B[i]-gammaM[i])));
            pp[,i,1]<-1-pp[,i,2];
            matrix.info[,i]<-D^2*A[i]^2*pp[,i,1]*(pp[,i,2]-C[i])^2/((1-C[i])^2*pp[,i,2]);
        }
        list(pp=pp , matrix.info=matrix.info , testlets.id=items$testlets);
    } else {
        for (i in 1:ni) {
            pp[,i,2]<-C[i]+(1-C[i])/(1+exp(-D*A[i]*(theta-B[i])));
            pp[,i,1]<-1-pp[,i,2];
            matrix.info[,i]<-D^2*A[i]^2*pp[,i,1]*(pp[,i,2]-C[i])^2/((1-C[i])^2*pp[,i,2]);
        }
        list(pp=pp,matrix.info=matrix.info);
    }
}
prep.prob.info.testlets<-function(){
        tmp.matrix.info <- prep.prob.info();
```

```
        matrix.info.testlets<-matrix(0,nq,ntestlets);
        tmp<-matrix(0,nq,ipt);
        sq.matrix.info <- matrix(0, nq , ni);
        for (item in 1:ni) sq.matrix.info[,item] <- tmp.matrix.info$matrix.info[ ,
tmp.matrix.info$testlets.id[item,2] ];
        ### sum testlet info
        i=0
        for (TT in 1:ntestlets){
            for (tt in 1:ipt){
                i=i+1
                matrix.info.testlets[,TT] <- matrix.info.testlets[,TT] + sq.matrix.info[,i]
            }
        }
        list(matrix.info=sq.matrix.info , matrix.info.testlets=matrix.info.testlets ,
testlets.id=tmp.matrix.info$testlets.id);
}

############################################################
# calcFullLengthEAP
############################################################
calcFullLengthEAP<-function() {
    posterior<-matrix(rep(prior,nExaminees),nExaminees,nq,byrow=T);
    for (i in 1:ni) {
        resp<-matrix(resp.data$resp[[i]],nExaminees,1)+1;
        prob<-t(PP[,i,resp]);
        prob[is.na(prob)]<-1.0
        posterior<-posterior*prob;
    }
    EAP<-posterior%*%theta/rowSums(posterior);
    SEM<-sqrt(rowSums(posterior*(matrix(theta,nExaminees,nq,byrow=T)-
matrix(EAP,nExaminees,nq))^2)/rowSums(posterior));
    return(list(theta=EAP,SE=SEM))
}
```

```
############################################################
# calcInfo
############################################################
calcInfo<-function(th) {
    info<-numeric(ni);
    available<-items.available;
    if (content.balancing) available<-items.available & (content.cat==next.content());
    if(ipt > 1){
        for (i in 1:ni) {
            if (available[i]==TRUE) {
                Ta<-(A[i]*(th-B[i]-gammaM[i]));
                info[i]<-A[i]^2*(exp(Ta)/(1+exp(Ta)))^2 * ((1-C[i])/(C[i]+exp(Ta)));
            } else info[i]<- 0
        }
        info<-list(info=info , tlst.id = items$testlets )
    } else{
        for (i in 1:ni) {
            if (available[i]==TRUE) {
                P<-C[i]+(1-C[i])/(1+exp(-D*A[i]*(th-B[i])));
                Q<-1-P;
                info[i]<-D^2*A[i]^2*Q*(P-C[i])^2/((1-C[i])^2*P);
            }
        }
        info<-list(info=info)
    }
    return(info);
}

############################################################
# calcInfo.Testlet
############################################################
calcInfo.Testlet<-function(th) {
    array.info.item<-calcInfo(th);
    tlst.info<-numeric(ntestlets);
    available<-testlets.available;
```

```
    if (content.balancing) available<-testlets.available & (content.cat==next.content());
    sq.array.info <- numeric(ntestlets);
    for (item in 1:ni) sq.array.info[item] <- array.info.item$info[array.info.item$tlst.id[item,2]];
    i=0
    for (TT in 1:ntestlets){
       for (tt in 1:ipt){
          i=i+1
          if (available[TT]==TRUE) tlst.info[TT] <- tlst.info[TT] + sq.array.info[i]
          else tlst.info[TT] <- 0
       }
    }
    res <- list(tlst.info=tlst.info)
    return(res);
}

#########################################################
# select.maxInfo
#########################################################
select.maxInfo<-function () {
    if (exposure.control) {
        rc<-runif(ni,max=max(array.info$info));
        rc[!items.available]<-0;
        rxx<-ifelse(ni.given==0,0,1-se.history[j,ni.given]^2);
        if (rxx<0) rxx<-0;
        array.info$info<-rxx*(1-exposure.rate/j)*array.info$info+(1-rxx)*rc;
    }
    info.index<-rev(order(array.info$info));
    if (ni.available>=topN) {
        item.selected<-info.index[sample(topN,1)];
    }
    else if (ni.available>0) {
        item.selected<-info.index[sample(ni.available,1)];
    }
    if (exposure.control) exposure.rate[item.selected]<<-exposure.rate[item.selected]+1;
    return (item.selected);
```

```
}

############################################################
# select.testlet.maxInfo
############################################################
select.testlet.maxInfo<-function () {
    info.index<-rev(order(array.info.testlet));
    if (ntestlets.available>0) {
        testlet.selected<-info.index[sample(ntestlets.available,1)];
    }
    return (testlet.selected);
}

############################################################
# calcEAP
############################################################
calcEAP<-function (examinee,ngiven) {
    lh<-rep(1,nq);
    for (i in 1:ngiven) {
        item<-items.used[examinee,i];
        resp<-resp.data$resp[examinee,item];
        prob<-PP[,item,resp+1];
        lh<-lh*prob;
    }
    posterior<-prior*lh;
    EAP<-sum(posterior*theta)/sum(posterior);
    if (se.method==1) {
        SEM<-sqrt(sum(posterior*(theta-EAP)^2)/sum(posterior));
    } else if (se.method==2) {
        SEM<-calcSE(examinee,ngiven,EAP);
    }
    return(list(THETA=EAP,SEM=SEM,LH=lh,posterior=posterior));
}
```

```
############################################################
# calcSE
############################################################
calcSE<-function(examinee,ngiven,th) {
    info<-matrix(0,1,ngiven)
    for (i in 1:ngiven) {
        item<-items.used[examinee,i];
        Ta<-A[i]*(th-B[i]-gammaM[i]);
        info[i]<-A[i]^2*(exp(Ta)/(1+exp(Ta)))^2 * ((1-C[i])/(C[i]+exp(Ta)));
    }
    SEM<-1/sqrt(sum(info));
    return(SEM);
}

############################################################
# calcMLE
############################################################
calcMLE<-function (examinee,ngiven) {
    EAP.estimates<-calcEAP(examinee,ngiven);
    total.raw<-sum(resp.data$resp[examinee,items.used[examinee,1:ngiven]]);
    if (total.raw==0 | total.raw==ngiven) {
        MLE<-EAP.estimates$THETA;
        SEM<-EAP.estimates$SEM;
    } else {
        maxIter<-20; crit<-0.01; maxStep<-0.5; change<-1000; nIter<-0;
        post.theta<-EAP.estimates$THETA;
        while (nIter<=maxIter && change>crit) {
            pre.theta<-theta.history[examinee,ngiven-1];
            deriv1<-0; deriv2<-0; info<-0;
            for (i in 1:ngiven) {
                item<-items.used[examinee,i];
                P<-C[item]+(1-C[item])/(1+exp(-D*A[item]*(pre.theta-B[item]-gammaM[item])));
                Q<-1-P;
                U<-resp.data$resp[examinee,item];
                deriv1<-deriv1+D*A[item]*(U-P)*(P-C[item])/(P*(1-C[item]));
```

```
            info<-info+D^2*A[item]^2*Q*(P-C[item])^2/((1-C[item])^2*P);
        }
        h<-deriv1/-info;
        if (h>maxStep | h< -maxStep) h<-sign(h)*maxStep;
        post.theta<-pre.theta-h;
        change<-abs(h);
        SEM<-calcSE(examinee,ngiven,post.theta);
        nIter<-nIter+1;
    }
    if (post.theta<minTheta) {
        MLE<-minTheta;
    } else if (post.theta>maxTheta) {
        MLE<-maxTheta;
    } else {
        MLE<-post.theta;
    }
  }
  return(list(THETA=MLE,SEM=SEM,LH=EAP.estimates$LH,posterior=EAP.estimates$posterior));
}
```

คำสั่งสำหรับคำนวณดัชนีลำดับความสำคัญสูงสุด (Maximum Priority Index: MPI)

```
############################################################
# MPI
############################################################
mpi <- function(K,w,lm,um,ca.cnt,ntestlet.given,C.C) {
  ########################################
  ## The maximum priority index method ##
  ########################################
  # m is total number of constraints (m=1,2,3,..,M)
  # Constraint relevancy matrix by C.C is a ni × M matrix
  # c1,c2,c3 as content area 1,2,3
  # and c4 is Exposure control (maximum item exposure rate is 15% or 0.15)
  # weight <- numeric(M); # weight of constraints
  # L is total test length
  # l_k <- numeric(M); # lower bound , M is total content areas
```

```
    # u_k <- numeric(M); # upper bound , M is total content areas
    # Mid_m as the midpoint between Upper_m and Lower_m;
    # Prevalence_m as the proportion of the items in the pool
    # having the property associated with constraint m.
    M <- K-1 # number of constraints
    #w.m <- matrix (NA, ntestlets, M)     #weight in content area (constraints)
    #for (i in 1:M) w.m[,i] <- rep(w[i], (ntestlets))
    length.test <- maxNtestlets     # total test testlets in test (15*4 = 60 items)
    x_m <- numeric(M)
    fx <- numeric(M)
    fm <- numeric(M)
    fxlow <- numeric(M)
    fxup <- numeric(M)
    ps <- matrix(NA, ntestlets , M)
    pri.score <- matrix(1, nrow=ntestlets ,1)
    x_m <- ca.cnt;
    for (m in 1:M) {
        fxlow[m] <- (lm[m]-x_m[m])/lm[m];
        if(fxlow[m]<=0) {
            fxup[m] <- (um[m]-x_m[m])/um[m];
            fx[m] <- fxup[m]
            if(fxup[m]<=0) fx[m] <- 0
        }
        if(fxlow[m]>0) {
            fx[m] <- fxlow[m]
        }
        fm[m] <- fx[m] * w[m];
            for (t in 1:ntestlets) ps[t,m] <- fm[m]^C.C[t,m];
        }
    pri.score[,1] <- pri.score[,1] * ps[,1]  * ps[,2] * ps[,3];
    res <- pri.score
    return(res);
}
```

```
#############################################################
# priority.score
#############################################################
priority.score <- function(K,w,lm,um,ca.cnt,ntestlet.given) {
   ########################################
   ## The maximum priority index method ##
   ########################################
   # m is total number of constraints (m=1,2,3,..,M)
   # Constraint relevancy matrix by C.C is a ni × M matrix
   # c1,c2,c3 as content area 1,2,3
   # and c4 is Exposure control (maximum item exposure rate is 15% or 0.15)
   # weight <- numeric(M); # weight of constraints
   # L is total test length
   # l_k <- numeric(M); # lower bound , M is total content areas
   # u_k <- numeric(M); # upper bound , M is total content areas
   # Mid_m as the midpoint between Upper_m and Lower_m;
   # Prevalence_m as the proportion of the items in the pool
   # having the property associated with constraint m.
   M <- K-1 # number of constraints
   #w.m <- matrix (NA, ntestlets, M)    #weight in content area (constraints)
   #for (i in 1:M) w.m[,i] <- rep(w[i], (ntestlets))
   length.test <- maxNtestlets    # total test testlets in test (15*4 = 60 items)
   x_m <- numeric(M)
   f1 <- numeric(M)
   f2 <- numeric(M)
   f1f2 <-  matrix(NA, 1 , ncol=M)
   #pri.sc <- numeric(M)
   ps <- matrix(NA, ntestlets , M)
   pri.score <- matrix(1, nrow=ntestlets ,1)
   x_m <- ca.cnt;
   for (m in 1:M) {
      f1[m] <- (um[m]-x_m[m]-1)/um[m];
      f2[m] <- ((length.test-lm[m]) - (ntestlet.given-x_m[m])) / (length.test-lm[m]);
      f1f2[m] <- f1[m]*f2[m]*w[m];
      for (t in 1:ntestlets) ps[t,m] <- f1f2[m]^C.C[t,m];
```

```
    }
    #for (m in 1:M)
    pri.score[,1] <- pri.score[,1] * ps[,1]  * ps[,2] * ps[,3] * f.er;
    res <- pri.score
    #res <- pri.score;
    return(res);
}
```

คำสั่งการคำนวณค่าสถิติที่ใช้พิจารณาประสิทธิภาพของการทดสอบแบบปรับเหมาะ

```
# average number of items and testlets used
avg.it<-sum(!is.na(items.used))/dim(resp.data)[1]
avg.tl<-sum(!is.na(testlets.used))/dim(resp.data)[1]
avg.sem<-mean(sem.CAT);
# correlation between thetas
cor.theta<-cor(ext.theta,theta.CAT)
# compute bias mse rmse aad sdm  thetas
delta.BIAS <- matrix(NA,1,nExaminees)
delta.MSE <- matrix(NA,1,nExaminees)
delta.AAD <- matrix(NA,1,nExaminees)
for (i in 1:nExaminees ){
    delta.BIAS[i] <- theta.CAT[i]-ext.theta[i]
    delta.MSE[i] <- (theta.CAT[i]-ext.theta[i])^2
    delta.AAD[i] <- abs(theta.CAT[i]-ext.theta[i])
}
bias <- sum(delta.BIAS)/nExaminees;
mse <- sum(delta.MSE)/nExaminees;
rmse <- sqrt(mse);
aad <- sum(delta.AAD)/nExaminees;
sdm <- (mean(theta.CAT) - mean(ext.theta)) / sqrt( (sd(theta.CAT)+sd(ext.theta))/2 )
# compute Test overlap rate
tl.used.vector <- as.vector(testlets.used)
n.tl.used <- matrix(0,nt,1)
n.tl.sum <- matrix(0,nt,1)
for (i in 1: nt ) {
```

```
    n.tl.used[i] <- length(which(tl.used.vector %in% i))
    n.tl.sum[i] <- (n.tl.used[i] * (n.tl.used[i]-1))
}
t.bar <- sum(n.tl.sum)/(maxNt * nExaminees*(nExaminees-1) )
# compute exposure rate
er <- matrix(0,nt,1)
for (j in 1: nt ) {
    er[j] <- n.tl.used[j]/nExaminees
}
# desirable uniform rate for all items
er_bar <- maxNt/nt
#Mimicking Pearson's X2 statistic in that it analyzes frequency data (Bishop, Fienberg, & Holland,
1975),
# compue X2 was designed to measure the similarity of the observed and desired exposure
rates:
x_s <- matrix(0,nt,1)
for (j in 1: nt ) {
    x_s[j] <-  ((er[j]-er_bar)^2)/er_bar
}
x_sqrt <- sum(x_s)
inv.se <-(sem.CAT)^2
test.info <- 1/inv.se
plot(er, main="testlets exposure", type="h" ,
     xlab="testlets position", ylab="exposure rate",
     xlim=c(1, nt), ylim=c(0, 1))
abline(h=c(r.max),lty=2,col="blue")
des<-cbind(avg.it, avg.tl, avg.sem, cor.theta, bias, mse, rmse, aad, sdm, max(er), t.bar, x_sqrt)
res <- list("selection.method"=selection.method, "r.max"=r.max, "resp.data"=resp.data,
"ext.theta"=ext.theta,"Yid"=Yid, "C.C"=C.C, "CA"=CA, "ca.rec"=ca.rec, "items.used"=items.used,
"testlets.used"=testlets.used,  "selected.item.resp"=selected.item.resp,
"ni.administered"=ni.administered, "nt.administered"=ntestlets.administered,
"theta.history"=theta.history, "se.history"=se.history, "theta.CAT"=theta.CAT,
"sem.CAT"=sem.CAT,"LH"=LH, "posterior.matrix"=posterior.matrix, "INFO"=INFO,
"TESTLETINFO"=TLSTINFO, "er"=er,
"des"=des ,"target.content.dist"=target.content.dist,"overall.content.freq"=overall.content.freq)
```

คำสั่งของ constraint-weighted a-stratification CAT method: CWA

```
############################################################
# constraint-weighted a-stratification CAT method: CWA
############################################################
strata<-T;
D<-1;          # logistic scale
simulateTheta<-T
nSimulee<-100        # number of Examinees
popMean<-0          # Mean of population
popSD<-1            # SD of population
eapFulllength<-T
ni=600              # Total Items
ipt = 4  # number of item per testlet
nt = ni/ipt # number of testlet
maxNt= 15;
minNt= 6;
maxNI= maxNt* ipt;          # Maximum item for administration
minNI= minNt* ipt;
itb <- itb600
A<-itb$A;
B<-itb$B;
C<-itb$C;
C.C<-itb$C.C;
CA<-itb$CA;
tlst.id <- as.vector(itb$testlets[,2])

if (strata==T){
    row.str <- 50
    str <- nt/row.str
    sq.a <- numeric(ni);
    for (i in 1:ni)  sq.a[i] <- A[itb$testlets[i,2]]
    t.a <- numeric(nt);
```

```
    i=0
    for (TT in 1:nt){
        for (tt in 1:ipt){
            i=i+1
            t.a[TT] <- t.a[TT] + sq.a[i]
        }
    }
    t.ma<-t.a/ipt # mean a of testlet
    a.str <- matrix(rev(order(t.ma)),row.str,str) #index of mean a of testlet sory by asc
    sq.b <- numeric(ni);
    for (i in 1:ni)  sq.b[i] <- B[itb$testlets[i,2]]
    t.b <- numeric(nt);
    i=0
    for (TT in 1:nt){
        for (tt in 1:ipt){
            i=i+1
            t.b[TT] <- t.b[TT] + sq.b[i]
        }
    }
    t.mb<-t.b/ipt # mean b of testlet
    if(str==3) t.used.str<-c(2.7,6.2,11.112) #20% 40% 40% t.used.str<-c(5,5,5);
    if(str==4) t.used.str<-c(1.112,3.336,6.672,11.112) #t.used.str<-c(3,3,4,5);
}
selection.method="CWaSTRi"
r.max=.10
content.balancing<-F;
if (ncc>1){
    #target.content.dist<-read.csv(filename.content.dist,header=F,skip = 0)[[1]];
    #content.cat<-read.csv(filename.content.cat,header=T,skip = 0)[[3]];
    content.cat<-CA;
    if (abs(sum(target.content.dist)-1)>.1) warning("ERROR: the sum of content proportions should
add up to 1.0\n:content balancing not used")
    else if (length(target.content.dist)!=ncc) warning("ERROR: the number of content categories
(ncc) does not match the number of target proportions in the content control file\n:content
balancing not used")
```

```
    else if (length(content.cat)!=nt) warning("ERROR: the number of records in the content control
file does not match the number of items in the bank\n:content balancing not used")
    else {
        overall.content.freq<-numeric(ncc);
        content.balancing<-T;
    }
}
Yid <- matrix(0, nrow=nSimulee, ncol=nt);
for (tt in 1:nt){
    Yid[,tt] <- rnorm( nSimulee , sd = LID );
}
Yid <- Yid[ , rep(1:nt,each=ipt) ]
resp.data<-genResp(nSimulee,Yid);
theta<-seq(minTheta,maxTheta,inc);
nq=length(theta);
start.theta<-prior.mean;
prior<-dnorm(theta);
nExaminees<-dim(resp.data)[1];
items.used<-matrix(NA,nExaminees,maxNI);
testlets.used<-matrix(NA,nExaminees,maxNt);
selected.item.resp<-matrix(NA,nExaminees,maxNI);
ni.administered<-numeric(nExaminees);
ntestlets.administered<-numeric(nExaminees);
theta.CAT<-rep(NA,nExaminees);
sem.CAT<-rep(NA,nExaminees);
theta.history<-matrix(NA,nExaminees,maxNI);
se.history<-matrix(NA,nExaminees,maxNI);
posterior.matrix<-matrix(NA,nExaminees,nq);
LH<-matrix(NA,nExaminees,nq);
matrix.prob.info<-prep.prob.info();
PP<-matrix.prob.info$pp;
INFO<-matrix.prob.info$matrix.info;
TLSTINFO<-matrix.prob.info$matrix.TLST.info;
ca.rec <- array(NA,c(3,maxNt,nExaminees))
f.er <- matrix(1, nrow=nt ,1) # initail exposure rate
```

```
n.tl.used <- matrix(0,nt,1)
n.tl.sum <- matrix(0,nt,1)
ext.theta<-resp.data$theta

#################################################
# Constraint-Weighted aSTR testlet Selection (i match) ver.1.0.5 10.5.57#
#################################################
if (selection.method=="CWaSTRi") {
    for (j in 1:nExaminees) {
        critMet<-FALSE;
        items.available<-rep(TRUE,ni);
        testlets.available<-rep(TRUE, nt);
        items.available[is.na(resp.data[j,paste("R",1:ni,sep="")])]<-FALSE;
        ni.given<-0
        nt.given<-0;
        theta.current<-start.theta;
        b.in.str<-matrix(t.mb,row.str,str);
        ca.count <- c(0,0,0);
        t.info <- 0;
        array.info.testlet.M <- numeric(row.str);
        zz.M <- numeric(row.str);
        f.er.M <- numeric(row.str);
        if (content.balancing) {
            current.content.dist<-numeric(ncc);
            current.content.freq<-numeric(ncc); }
        while (critMet==FALSE && sum(testlets.available)>0){
            for (s in 1:str) {
                while (t.info<=t.used.str[s]&& nt.given<maxNt){
                    array.info<-calcInfo(theta.current);
                    ni.available<-sum(array.info[,1]>0);
                    array.info.testlet<-calcInfo.Testlet(theta.current);
                    for(tt in 1:row.str) array.info.testlet.M[tt] <- array.info.testlet[a.str[tt,s]]
                    ntestlets.available<-sum(array.info.testlet.M>0);
                    zz<-mpi(w,lm,um,ca.count,nt.given,C.C)
                    for(tt in 1:row.str){
```

```
    zz.M[tt] <- zz[a.str[tt,s]]
    f.er.M[tt] <- f.er[a.str[tt,s]]
}
pri.sc <- array.info.testlet.M * f.er.M
max.PI <- max(pri.sc, na.rm = T);
if(max.PI==0) {#break #stop("max.PI==0")
    pri.sc <- array.info.testlet.M
    max.PI <- max(pri.sc, na.rm = T);
}
max.PI.index <- match(max.PI , pri.sc , nomatch = NA);
testlet.selected <- a.str[max.PI.index,s];
#if (nt.given==0) testlet.selected <-  sample(a.str[,1],1)
nt.given<-nt.given+1;
testlets.used[j,nt.given]<-testlet.selected;
testlets.available[testlet.selected]<-FALSE;
#show(max.PI)

if (content.balancing) update.content.dist();
show(j)
show(testlet.selected)
if (itb$CA[testlet.selected]==1){
    ca.count[1] <- ca.count[1]+1
    ca.rec[1,ca.count[1],j] <- testlet.selected }
if (itb$CA[testlet.selected]==2) {
    ca.count[2] <- ca.count[2]+1
    ca.rec[2,ca.count[2],j] <- testlet.selected }
if (itb$CA[testlet.selected]==3) {
    ca.count[3] <- ca.count[3]+1
    ca.rec[3,ca.count[3],j] <- testlet.selected }
tl.used.vector <- as.vector(testlets.used)
for (i in 1: nt) n.tl.used[i] <- length(which(tl.used.vector %in% i));
for (i in 1: nt) {
    f.er[i] <- (r.max-(n.tl.used[i]/nExaminees))/r.max;
    if (f.er[i]< 0) f.er[i] = 0; }
for (ii in 1:ni) {
```

```
                    if (itb$testlets[ii,1]==testlet.selected) {
                        item.selected <- itb$testlets[ii,2];
                        ni.given<-ni.given+1;
                        items.used[j,ni.given]<-item.selected;
                        resp<-resp.data[j,item.selected];
                        selected.item.resp[j,ni.given]<-resp;
                        items.available[item.selected]<-FALSE;
                        if (interim.Theta==1) estimates<-calcEAP(j,ni.given);
                        else if (interim.Theta==2) estimates<-calcMLE(j,ni.given);
                        theta.history[j,ni.given]<-estimates$THETA;
                        se.history[j,ni.given]<-estimates$SEM;
                        theta.current<-estimates$THETA;
}
                    }
                    t.info <- 1/(estimates$SEM)^2;
                }
            }
            if (nt.given>=maxNt || (estimates$SEM<=maxSE && nt.given>=minNt || max.PI==0)){
                critMet<-TRUE;
                theta.CAT[j]<-estimates$THETA;
                sem.CAT[j]<-estimates$SEM;
                LH[j,]<-estimates$LH;
                posterior.matrix[j,]<-estimates$posterior;
                ni.administered[j]<-ni.given;
                ntestlets.administered[j]<-nt.given;
            }
            #if(max.PI==0) break #stop("max.PI==0")
        }
    }
}
```

คำสั่งของ Monte Carlo CAT Method: MCC

```
#####################################################
# Monte Carlo CAT Method ver.1.0.5 10.5.57#
#####################################################
D<-1 #.702;          # Haleyìs Constant
ipt = 4  # number of item per testlet
nt = ni/ipt # number of testlet
# magnitudes of testlet effects have been previously studied by Wainer, Bradlow, & Du (2000)
LID <- sqrt(1);
minTheta=-4
maxTheta=4
inc=0.1
maxNt= 15;
minNt= 6;
maxNI= maxNt* ipt;            # Maximum item for administration
minNI= minNt* ipt;            # Minimum item for administration
maxSE=0.3          # SE of stopping rule
#target.content.dist<-read.csv(filename.content.dist,header=F,skip = 0)[[1]];
# set target content distribution
target.content.dist <-c(0.33, 0.33 , 0.34);
ncc=length(target.content.dist)              # Number of content
cb=T
topN=1
exposure.control=F
interim.Theta=1     # 1 calcEAP    ,    # 2 calcMLE
se.method=1         # 1 SE Postteriior # 2 SE Infomation
first.item.selection=1
first.at.theta=0
prior.dist=1
prior.mean=0
prior.sd=1
selection.method="mccCAT"
r.max=.25
library(lpSolveAPI)
library(gdata)
```

```
        itb <- itb800
        A<-itb$A;
        B<-itb$B;
        C.C<-itb$C.C;
        CA<-itb$CA;
        tlst.id <- as.vector(itb$testlets[,2])
         ID = S$Item
ni=length(B)
nTestlet=length(tlst.id)
Content = C.C
I = nrow(B)
FF= 3
N = nTestlet
Vc = list()
for(k in 1:5){
   Vc[[k]] = c(1:I)[Content == k] }
theta = c(-1.5,0,1.5)
J = length(theta)
Info = array(0,c(I,J))
for(j in 1:J){
   P = C+(1-C)/(1+exp(-1.7*A*(theta[j]-B)))
   Q = 1-P
   Info[,j] = (1.7^2)*(A^2)*((P-C)/(1-C))^2*Q/P }
M = I*FF+1
# Create the model
lprec = make.lp(0,M)
# Set control parameters: minimization problem; integer tolerance is set to 0.1;
# absolute MIP gap is set to 0.1;  relative MIP gap is set to 0.05;
lp.control(lprec,sense="min", epsint = 0.1, mip.gap = c(0.1,0.05));
# Constraints (8) and (9)
set.type(lprec,columns = c(1:(FF*I)),type = "binary")
set.type(lprec,columns = M,type = "real")
set.bounds(lprec,lower = rep(0,M),upper = rep(1,M))
# Constraint (5) guarantees no item overlap between the two forms
FF.Nc <- numeric(FF)
```

```
    for (i in 1:FF){
        if(i==1) FF.Nc[i] <-1
        if(i>1)  FF.Nc[i] <-1+(I*(i-1))
    }
    for (k in 1:I){
        add.constraint(lprec,rep(1,FF),"<=",1,indices = FF.Nc)
    }
#    for (k in 1:I){
#        add.constraint(lprec,rep(1,2),"<=",1,indices = c(k,I+k)) }

    # Constraint (6) The content category requirements
    Nc = c(10,5,5,5,4)
    # Form 1
    for(k in 1:5){
        add.constraint(lprec,rep(1,length(Vc[[k]])),">=",Nc[k],indices = Vc[[k]]) }
    # Form 2
    for(k in 1:5){
        add.constraint(lprec,rep(1,length(Vc[[k]])),">=",Nc[k],indices = I+Vc[[k]]) }
    # Form 3
    for(k in 1:5){
        add.constraint(lprec,rep(1,length(Vc[[k]])),">=",Nc[k],indices = (I*2)+Vc[[k]]) }
    # Constraint (7) The constraint on the test length for both forms.
    add.constraint(lprec, rep(1,I), "=", N, indices = 1:I)
    add.constraint(lprec, rep(1,I), "=", N, indices = (I+1):(2*I))
    add.constraint(lprec, rep(1,I), "=", N, indices = ((I*2)+1):(3*I))
    # Constraints (3) and (4) Ii  ≤ Ty +- y; for all three ys and both forms,
    # For each form and y value, the constraints in Equations 3 and 4 require
    # the distance between the TIF of the assembled test and the target value T to be no greater
than y
    d_theta = c(5.4, 10, 5.4)
    for(k in 1:3){
        add.constraint(lprec,c(Info[,k],-1),"<=",d_theta[k],indices = c(1:I,M))
        add.constraint(lprec,c(Info[,k],-1),"<=",d_theta[k],indices = c((I+1):(2*I),M))
        add.constraint(lprec,c(Info[,k],-1),"<=",d_theta[k],indices = c(((I*2)+1):(3*I),M))
        add.constraint(lprec,c(Info[,k],1),">=",d_theta[k],indices = c(1:I,M))
```

```
        add.constraint(lprec,c(Info[,k],1),">=",d_theta[k],indices = c((I+1):(2*I),M))
        add.constraint(lprec,c(Info[,k],1),">=",d_theta[k],indices = c(((I*2)+1):(3*I),M))
    }
    # Objective function
    set.objfn(lprec,1,indices = M)
    # Solve the model
    res_flag = solve(lprec)
    # The integer value containing the status code, for example, 0: ‘‘optimal solution found’’
res_flag
    # Retrieve the values of the decision variables
    x_opt = get.variables(lprec)
          set_Shadowstest = x_opt;

content.balancing<-F;
if (ncc>1){
    #target.content.dist<-read.csv(filename.content.dist,header=F,skip = 0)[[1]];
    #content.cat<-read.csv(filename.content.cat,header=T,skip = 0)[[3]];
    content.cat<-CA;
    if (abs(sum(target.content.dist)-1)>.1) warning("ERROR: the sum of content proportions should
add up to 1.0\n:content balancing not used")
    else if (length(target.content.dist)!=ncc) warning("ERROR: the number of content categories
(ncc) does not match the number of target proportions in the content control file\n:content
balancing not used")
    else if (length(content.cat)!=nt) warning("ERROR: the number of records in the content control
file does not match the number of items in the bank\n:content balancing not used")
    else {
        overall.content.freq<-numeric(ncc);
        content.balancing<-T;
    }
}
Yid <- matrix(0, nrow=nSimulee, ncol=nt);
for (tt in 1:nt){
    Yid[,tt] <- rnorm( nSimulee , sd = LID );
}
Yid <- Yid[ , rep(1:nt,each=ipt) ]
```

```
resp.data<-genResp(nSimulee,Yid);
theta<-seq(minTheta,maxTheta,inc);
nq=length(theta);
start.theta<-prior.mean;
prior<-dnorm(theta);
nExaminees<-dim(resp.data)[1];
items.used<-matrix(NA,nExaminees,maxNI);
testlets.used<-matrix(NA,nExaminees,maxNt);
selected.item.resp<-matrix(NA,nExaminees,maxNI);
ni.administered<-numeric(nExaminees);
ntestlets.administered<-numeric(nExaminees);
theta.CAT<-rep(NA,nExaminees);
sem.CAT<-rep(NA,nExaminees);
theta.history<-matrix(NA,nExaminees,maxNI);
se.history<-matrix(NA,nExaminees,maxNI);
posterior.matrix<-matrix(NA,nExaminees,nq);
LH<-matrix(NA,nExaminees,nq);
matrix.prob.info<-prep.prob.info();
PP<-matrix.prob.info$pp;
INFO<-matrix.prob.info$matrix.info;
TLSTINFO<-matrix.prob.info$matrix.TLST.info;
ca.rec <- array(NA,c(3,maxNt,nExaminees));
f.er <- matrix(1, nrow=nt ,1) # initail exposure rate;
n.tl.used <- matrix(0,nt,1); n.tl.sum <- matrix(0,nt,1); ext.theta<-resp.data$theta;
if (selection.method=="mccCAT") {
    for (j in 1:nExaminees) {
        critMet<-FALSE;
        items.available<-rep(TRUE,ni);
        testlets.available<-rep(TRUE, nt);
        items.available[is.na(resp.data[j,paste("R",1:ni,sep="")])]<-FALSE;
        ni.given<-0
        nt.given<-0;
        theta.current<-start.theta;
        ca.count <- c(0,0,0);
        if (content.balancing) {
```

```
    current.content.dist<-numeric(ncc);
    current.content.freq<-numeric(ncc);
}
l <- 0;
while (critMet==FALSE && sum(testlets.available)>0) {
    array.info<-calcInfo(theta.current);
    ni.available<-sum(array.info[,1]>0);
    array.info.testlet<-calcInfo.Testlet(theta.current);
    nt.available<-sum(array.info.testlet>0);
    sequence_testlet<-sample(set_Shadowstest, k=2l+1);
    l <- l+1;
    max.TI <- max(asequence_testlet, na.rm = T);
            max.TI.index <- match(max.TI , array.info.testlet , nomatch = NA);
            testlet.selected <- max.TI.index;
    nt.given<-nt.given+1;
    testlets.used[j,nt.given]<-testlet.selected;
    if (content.balancing) update.content.dist();
    show(j)
    show(testlet.selected)
    if (itb$CA[testlet.selected]==1){
        ca.count[1] <- ca.count[1]+1
        ca.rec[1,ca.count[1],j] <- testlet.selected }
    if (itb$CA[testlet.selected]==2) {
        ca.count[2] <- ca.count[2]+1
        ca.rec[2,ca.count[2],j] <- testlet.selected }
    if (itb$CA[testlet.selected]==3) {
        ca.count[3] <- ca.count[3]+1
        ca.rec[3,ca.count[3],j] <- testlet.selected }
    tl.used.vector <- as.vector(testlets.used)
    for (i in 1: nt) n.tl.used[i] <- length(which(tl.used.vector %in% i));
    for (i in 1: nt) {
        f.er[i] <- (r.max-(n.tl.used[i]/nExaminees))/r.max;
        if (f.er[i]< 0) f.er[i] = 0;
        update.Shadowstest();
    }
```

```
            for (ii in 1:ni) {
                if (itb$testlets[ii,1]==testlet.selected) {
                    item.selected <- itb$testlets[ii,2];
                    ni.given<-ni.given+1;
                    items.used[j,ni.given]<-item.selected;
                    resp<-resp.data[j,item.selected];
                    selected.item.resp[j,ni.given]<-resp;
                    items.available[item.selected]<-FALSE;
                    if (interim.Theta==1) {
                        estimates<-calcEAP(j,ni.given);
                    } else if (interim.Theta==2) {
                        estimates<-calcMLE(j,ni.given);
                    }
                    theta.history[j,ni.given]<-estimates$THETA;
                    se.history[j,ni.given]<-estimates$SEM;
                    theta.current<-estimates$THETA;
                }
            }
            testlets.available[testlet.selected]<-FALSE;
            #if (nt.given>=maxNt || (estimates$SEM<=maxSE && nt.given>=minNt) &&
((ca.count[1]>=lm[1])&&(ca.count[2]>=lm[2])&&(ca.count[3]>=lm[3]))) {
            if (nt.given>=maxNt || (estimates$SEM<=maxSE && nt.given>=minNt)){

                critMet<-TRUE;
                theta.CAT[j]<-estimates$THETA;
                sem.CAT[j]<-estimates$SEM;
                LH[j,]<-estimates$LH;
                posterior.matrix[j,]<-estimates$posterior;
                ni.administered[j]<-ni.given;
                ntestlets.administered[j]<-nt.given;
            }
        }
    }
}
```

คำสั่งของ Testlet Maximun Fisher Information: TFI

```
#####################################################
# Testlet Maximun Fisher Information
#####################################################
if (selection.method=="TFI") {
  for (j in 1:nExaminees) {
    critMet<-FALSE;    items.available<-rep(TRUE,ni);    testlets.available<-rep(TRUE, nt);
    items.available[is.na(resp.data[j,paste("R",1:ni,sep="")])]<-FALSE;
    ni.given<-0;        nt.given<-0;
    theta.current<-start.theta;
    ca.count <- c(0,0,0);
    if (content.balancing) {
      current.content.dist<-numeric(ncc);
      current.content.freq<-numeric(ncc);
    }
    while (critMet==FALSE && sum(testlets.available)>0) {
      array.info<-calcInfo(theta.current);
      ni.available<-sum(array.info[,1]>0);
      array.info.testlet<-calcInfo.Testlet(theta.current);
      ntestlets.available<-sum(array.info.testlet>0);
      max.TI <- max(array.info.testlet, na.rm = T);
      max.TI.index <- match(max.TI , array.info.testlet , nomatch = NA);
      testlet.selected <- max.TI.index;
      nt.given<-nt.given+1;
      testlets.used[j,nt.given]<-testlet.selected;
      if (content.balancing) update.content.dist();
      show(j)
      show(testlet.selected)
      if (itb$CA[testlet.selected]==1){
        ca.count[1] <- ca.count[1]+1
        ca.rec[1,ca.count[1],j] <- testlet.selected
      }
      if (itb$CA[testlet.selected]==2) {
        ca.count[2] <- ca.count[2]+1
        ca.rec[2,ca.count[2],j] <- testlet.selected
```

```
      }
      if (itb$CA[testlet.selected]==3) {
        ca.count[3] <- ca.count[3]+1
        ca.rec[3,ca.count[3],j] <- testlet.selected
      }
      for (ii in 1:ni) {
        if (itb$testlets[ii,1]==testlet.selected) {
          item.selected <- itb$testlets[ii,2];
          ni.given<-ni.given+1;
          items.used[j,ni.given]<-item.selected;
          resp<-resp.data[j,item.selected];
          selected.item.resp[j,ni.given]<-resp;
          items.available[item.selected]<-FALSE;
          if (interim.Theta==1) {
            estimates<-calcEAP(j,ni.given);
          } else if (interim.Theta==2) {
            estimates<-calcMLE(j,ni.given);
          }
          theta.history[j,ni.given]<-estimates$THETA;
          se.history[j,ni.given]<-estimates$SEM;
          theta.current<-estimates$THETA;
        }
      }
      testlets.available[testlet.selected]<-FALSE;
      if (nt.given>=maxNt || (estimates$SEM<=maxSE && nt.given>=minNt)) {
        critMet<-TRUE;
        theta.CAT[j]<-estimates$THETA;        sem.CAT[j]<-estimates$SEM;
        LH[j,]<-estimates$LH;
        posterior.matrix[j,]<-estimates$posterior;
        ni.administered[j]<-ni.given;
        ntestlets.administered[j]<-nt.given;
      }
    }
  }
}
```

คำสั่งของ Testlet Random Selection: RAN

```
#####################################################
# Testlet random selection
#####################################################
# random selection
if (selection.method=="random") {
  for (j in 1:nExaminees) {
    critMet<-FALSE;
    items.available<-rep(TRUE,ni);
    testlets.available<-rep(TRUE, nt);
    items.available[is.na(resp.data[j,1:ni])]<-FALSE;
    ni.given<-0
    nt.given<-0;
    theta.current<-start.theta;
    ca.count <- c(0,0,0);
    random<-runif(nt);
    is.na(random[!testlets.available])<-TRUE;
    testlet.order<-order(random);
    critMet<-FALSE;
    items.available<-rep(TRUE,ni);
    items.available[is.na(resp.data[j,1:ni])]<-FALSE;
    ni.given<-0;
    random<-runif(ni);
    is.na(random[!items.available])<-TRUE;
    item.order<-order(random);
    while (critMet==FALSE && sum(items.available)>0) {
      testlet.selected<-testlet.order[nt.given+1];
      nt.given<-nt.given+1;
      testlets.used[j,nt.given]<-testlet.selected;
      show(j)
      show(testlet.selected)
      if (itb$CA[testlet.selected]==1){
        ca.count[1] <- ca.count[1]+1;  ca.rec[1,ca.count[1],j] <- testlet.selected;
      }
      if (itb$CA[testlet.selected]==2) {
```

```
      ca.count[2] <- ca.count[2]+1;  ca.rec[2,ca.count[2],j] <- testlet.selected;
    }
    if (itb$CA[testlet.selected]==3) {
      ca.count[3] <- ca.count[3]+1;  ca.rec[3,ca.count[3],j] <- testlet.selected;
    }
    for (ii in 1:ni) {
      if (itb$testlets[ii,1]==testlet.selected) {
        item.selected <- itb$testlets[ii,2];
        ni.given<-ni.given+1;
        items.used[j,ni.given]<-item.selected;
        resp<-resp.data[j,item.selected];
        selected.item.resp[j,ni.given]<-resp;
        items.available[item.selected]<-FALSE;
        if (interim.Theta==1) {
          estimates<-calcEAP(j,ni.given);
        } else if (interim.Theta==2) {
          estimates<-calcMLE(j,ni.given);
        }
        theta.history[j,ni.given]<-estimates$THETA;  se.history[j,ni.given]<-estimates$SEM;
        theta.current<-estimates$THETA;
      }
    }
    testlets.available[testlet.selected]<-FALSE;
    #if (nt.given>=maxNt || (estimates$SEM<=maxSE && nt.given>=minNt)) {
    if  (estimates$SEM<=maxSE && nt.given>=minNt) {
     critMet<-TRUE;
      theta.CAT[j]<-estimates$THETA;  sem.CAT[j]<-estimates$SEM;
      LH[j,]<-estimates$LH;  posterior.matrix[j,]<-estimates$posterior;
      ni.administered[j]<-ni.given;  ntestlets.administered[j]<-nt.given;
    }
   }
  }
}
```

ประวัติผู้เขียนวิทยานิพนธ์

นายอนุสรณ์ เกิดศรี เกิดวันที่ 2 มีนาคม พ.ศ. 2526 ที่อำเภอเมือง จังหวัดอุทัยธานี สำเร็จการศึกษาปริญญาวิทยาศาสตรบัณฑิต (เกียรตินิยม อันดับ 2) สาขาวิทยาการคอมพิวเตอร์ จากมหาวิทยาลัยราชภัฏนครสวรรค์ ในปีการศึกษา 2548 สำเร็จการศึกษาประกาศนียบัตรบัณฑิตวิชาชีพครู จากมหาวิทยาลัยราชภัฏนครสวรรค์ ในปีการศึกษา 2549 โดยได้รับทุนโครงการส่งเสริมการผลิตครูที่มีความสามารถ พิเศษทางวิทยาศาสตร์และคณิตศาสตร์ (สควค.) สำเร็จการศึกษาปริญญาครุศาสตรมหาบัณฑิต สาขาวิชาวิจัยการศึกษา ภาควิชาวิจัยและจิตวิทยาการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2551 และเข้าศึกษาต่อในหลักสูตรครุศาสตรดุษฎีบัณฑิต สาขาวิชาการวัดและประเมินผลการศึกษา ภาควิชาวิจัยและจิตวิทยาการศึกษา ในปีการศึกษา 2552 พร้อมทั้งได้รับทุนนำเสนอผลงานวิจัย "ทุนสนับสนุนนิสิต ป.เอก ไปเสนอผลงานวิชาการในต่างประเทศ" จากบัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย และได้รับทุนสนับสนุนการวิจัย “ทุน 90 ปี จุฬาลงกรณ์มหาวิทยาลัย ” จากกองทุน รัชดาภิเษกสมโภช จุฬาลงกรณ์มหาวิทยาลัย ปัจจุบันเป็นข้าราชการครู สังกัดโรงเรียนวัดเขาพระยาสังฆาราม อำเภอลานสัก จังหวัดอุทัยธานี