การเปรียบเทียบวิธีการคัดเลือกข้อสอบขั้นแรกและลำดับข้อสอบที่มีต่อคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์

นางสาวทัศน์ศิรินทร์ สว่างบุญ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาครุศาสตรดุษฎีบัณฑิต
สาขาวิชาการวัดและประเมินการศึกษา ภาควิชาวิจัยและจิตวิทยาการศึกษา
คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2554
ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

บทคัดย่อและแฟ้มข้อมูลฉบับเต็มของวิทยานิพนธ์ตั้งแต่ปีการศึกษา 2554 ที่ให้บริการในคลังปัญญาจุฬาฯ (CUIR)
เป็นแฟ้มข้อมูลของนิสิตเจ้าของวิทยานิพนธ์ที่ส่งผ่านทางบัณฑิตวิทยาลัย
The abstract and full text of theses from the academic year 2011 in Chulalongkorn University Intellectual Repository(CUIR)
are the thesis authors' files submitted through the Graduate School.

THE COMPARISON OF ITEM SELECTION PROCEDURES AT THE EARLY STAGES AND ITEM ORDERS ON THE QUALITY OF MULTIDIMENSIONAL COMPUTERIZED ADAPTIVE TESTING

Miss Tatsirin Sawangboon

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 2011

Copyright of Chulalongkorn University

หัวข้อวิทยานิพนธ์	การเปรียบเทียบวิธีการคัดเลือกข้อสอบขั้นแรกและลำดับข้อสอบที่มีต่อคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์
โดย	นางสาวทัศน์ศิรินทร์ สว่างบุญ
สาขาวิชา	การวัดและประเมินผลการศึกษา
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	รองศาสตราจารย์ ดร.ศิริเดช สุชีวะ
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	ศาสตราจารย์ ดร.ศิริชัย กาญจนวาสี
	ศาสตราจารย์ ดร.เออิจิ มูราคิ

คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้วิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาดุษฎีบัณฑิต

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

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(รองศาสตราจารย์ ดร.โชติกา ภาษีผล)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(รองศาสตราจารย์ ดร.ศิริเดช สุชีวะ)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม
(ศาสตราจารย์ ดร.ศิริชัย กาญจนวาสี)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ณัฐฏภรณ์ หลาวทอง)

..... กรรมการภายนอกมหาวิทยาลัย
(ดร.รังสรรค์ มณีเล็ก)

ทัศน์ศิรินทร์ สว่างบุญ : การเปรียบเทียบวิธีการคัดเลือกข้อสอบขั้นแรก และลำดับข้อสอบที่มีต่อคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์ (THE COMPARISON OF ITEM SELECTION PROCEDURES AT THE EARLY STAGES AND ITEM ORDERS ON THE QUALITY OF MULTIDIMENSIONAL COMPUTERIZED ADAPTIVE TESTING) อ. ที่ปรึกษาวิทยานิพนธ์หลัก: รศ. ดร.ศิริเดช สุชีวะ , อ.ที่ปรึกษาวิทยานิพนธ์ร่วม: ศ.ดร.ศิริชัย กาญจนวาสี , ศ.ดร.เออิจิ มูราคิ , 222 หน้า.

การทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์เป็นวิธีการที่ผสมผสานระหว่างทฤษฎีการตอบสนองข้อสอบแบบพหุมิติ และการทดสอบแบบปรับเหมาะซึ่งพัฒนาขึ้นมาเพื่อวัดความสามารถของผู้สอบมากกว่า1มิติผ่านการทดสอบเพียงครั้งเดียว การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อ 1) เปรียบเทียบวิธีการคัดเลือกข้อสอบขั้นแรก (Fisher information และ Kullback-leibler information) และลำดับข้อสอบ (จำแนกมิติ และผสมผสานมิติ) ที่มีต่อคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์ รวมถึงศึกษาปฏิสัมพันธ์ระหว่างตัวแปรทั้ง 2 และ 2) เพื่อศึกษาผลประเมินโปรแกรมการทดสอบปรับเหมาะแบบพหุมิติ งานวิจัยนี้ศึกษาภายใต้โมเดลการตอบสนองข้อสอบแบบพหุมิติ 3 พารามิเตอร์ โดยพิจารณาคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์ ด้วยค่าความเที่ยง ค่าระยะทางแบบยุคลิด สารสนเทศแบบสอบ และความยาวข้อสอบ

การวิจัยครั้งนี้ศึกษาจากแหล่งข้อมูล 3 แหล่ง คือ การจำลองข้อมูล ข้อมูลทุติยภูมิจากการสอบ PAT2 ของ สทศ. และข้อมูลปฐมภูมิจากการดำเนินการสอบจริงกับนักเรียนชั้น ม.5 โดยการจำลองข้อมูลศึกษา 2 มิติ ข้อมูลทุติยภูมิและปฐมภูมิศึกษา 3 มิติ ในวิชาชีววิทยา เคมี และฟิสิกส์ ผลการวิจัยมีดังต่อไปนี้

1. วิธีการคัดเลือกข้อสอบขั้นแรกด้วยวิธี Kullback-Leibler information มีคุณภาพการทดสอบสูงกว่าวิธี Fisher information

2. ลำดับข้อสอบแบบผสมผสานมิติมีคุณภาพการทดสอบสูงกว่าลำดับข้อสอบแบบจำแนกมิติ

3. เกิดปฏิสัมพันธ์ระหว่างการคัดเลือกข้อสอบขั้นแรกและ ลำดับข้อสอบที่ส่งผลต่อคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติ โดยที่ วิธี Kullback-Leibler information แบบผสมผสานมิติ มีคุณภาพสูง สุดในกรณีการทดสอบขั้นแรก

4. ผลประเมินการใช้โปรแกรมการทดสอบปรับเหมาะแบบพหุมิติ พบว่า ผู้สอบมีความพึงพอใจระดับมาก (M = 3.52, $S.D.$ = 0.49)และพบว่า ภูมิหลังของผู้สอบ (เกรดเฉลี่ยสะสมและทักษะการใช้คอมพิวเตอร์) ไม่ส่งผลต่อผลประเมินโปรแกรมการทดสอบปรับเหมาะแบบพหุมิติ

ภาควิชา วิจัยและจิตวิทยาการศึกษา ลายมือชื่อนิสิต.....

สาขาวิชา การวัดและประเมินผลการศึกษา ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก.....

ปีการศึกษา 2554 ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์ร่วม.....

5084216527: MAJOR EDUCATONAL MEASUREMENT AND EVALUATION
KEYWORDS: MULTIDIMENSIONAL IRT / COMPUTERIZED ADAPTIVE TESTING / ITEM SELECTION PROCEDURES/ ITEM ORDERS

TATSIRIN SAWANGBOON: THE COMPARISON OF ITEM SELECTION PROCEDURES AT THE EARLY STAGES AND ITEM ORDERS ON THE QUALITY OF MULTIDIMENSIONAL COMPUTERIZED ADAPTIVE TESTING. ADVISOR: ASST. PROF. SIRIDEJ SUJIVA, Ph.D. CO-ADVISOR: PROF. SIRICHAI KANJANAWASEE, Ph.D., PROF. EIJI MURAKI, Ph.D. 222 pp.

Multidimensional computerized adaptive testing (MCAT) is a combined procedure between multidimensional item response theory (MIRT) and computerized adaptive testing (CAT). MCAT has been developed in order to measure several latent traits of a single examination. The purposes of this research were 1) to compare the item selection procedures at the early stage (Fisher information and Kullback-leibler information) and item orders (mixed items and classified items) on the quality of MCAT as well as to study the interaction effect between two mentioned variables, 2) to study the result of MCAT evaluation. This research was studied within the concept of the multidimensional three-parameter item response model. The quality of MCAT was operationally characterized by reliability, Euclidean distance, test information, and test length.

The analysis of this research was based on the data from three sources, that is, (a) simulation data, (b) secondary data from PAT2 testing conducted by NIETS, and (c) primary data from a real testing program with Grade 11 students. The simulated data were studied in two dimensions whereas the secondary data and primary data were studied in three dimensions– Biology, Chemistry, and Physics. The findings revealed that:

1. According to the item selection procedures at the early stage, the Kullback-Leibler information Method provided the better quality of MCAT than the Fisher information Method.

2. According to the item orders, the mixed-item order provided the better quality of MCAT than the classified items order.

3. The significant of interaction effects between the item selection procedures at the early stage and the item orders were found. It revealed that Kullback-Leibler information with mixed-item order provided the highest quality of MCAT at the early stage.

4. The evaluation result of MCAT program showed that the examinees were satisfied at the high level (M =3.52, SD = 0.49). In addition, there was no significant effect of the examinee backgrounds (GPAX and computer skills) on the evaluation result of MCAT program.

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

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

Academic Year: 2011 Co-advisor's Signature…….....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้คงสำเร็จลุล่วงไปไม่ได้ หากขาดความช่วยเหลือ ความเมตตา จาก รองศาสตราจารย์ ดร .ศิริเดช สุชีวะ และ ศาสตราจารย์ ดร .ศิริชัย กาญจนวาสี อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ได้สละเวลาให้คำปรึกษาในการทำวิทยานิพนธ์ พร้อมทั้งให้ข้อเสนอแนะเพื่อปรับปรุงแก้ไขข้อบกพร่องต่างๆ

ขอกราบขอบพระคุณ Prof.Dr. Eiji Muraki ที่ให้ความรู้และให้การดูแลตลอดระยะเวลาที่ศึกษาวิจัยอยู่ที่มหาวิทยาลัยโทโฮกุ ประเทศญี่ปุ่น รวมถึงคณาจารย์ภาควิชาวิจัยและจิตวิทยาการศึ กษาทุก ท่านที่ได้ประสิทธิ์ประสาทวิชาความรู้อันเป็นพื้นฐานสำคัญในการทำวิทยานิพนธ์ รวมทั้ง รองศาสตราจารย์ ดร .โชติกา ภาษีผล ผู้ช่วย ศาสตราจารย์ ดร.ณัฐฏภรณ์ หลาวทอง และ อาจารย์ ดร.รังสรรค์ มณีเล็ก ที่ให้ความกรุณาเป็นกรรมการสอบวิทยานิพนธ์

ขอกราบขอบพระคุณผู้ทรงคุณ วุฒิทุกท่าน ที่ได้พัฒนาข้อสอบ และตรวจสอบคุณภาพข้อสอบเพื่อใช้ในการวิจัยครั้งนี้ รวมถึงข้อมูลการสอบความถนัดทางวิชาการและวิชาชีพด้านวิทยาศาสตร์ จากสำนักทดสอบทางการศึกษาแห่งชาติ (สทศ.) ทำให้วิทยานิพนธ์ฉบับนี้มีความน่าเชื่อถือมากขึ้น

โครงการเครือข่ายเชิงกลยุทธ์เพื่อการผลิตและพัฒนาอาจารย์ในสถาบันอุดมศึกษา สำนักงานคณะกรรมการการอุดมศึกษา ที่มอบทุนการศึกษาและทุนวิจัย รวมถึง บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัยที่มอบ “ทุน 90 ปี จุฬาลงกรณ์มหาวิทยาลัย ” จากกองทุนรัชดาภิเษกสมโภช จุฬาลงกรณ์มหาวิทยาลัย ทำให้ผู้วิจัยสามารถพัฒนาวิทยานิพนธ์ได้อย่างเต็มที่

คุณสิวะโชติ ศรีสุทธิยากร และคุณชัยวัฒน์ รุจิเมธาภาส ที่ให้คำปรึกษา และเป็นผู้เชี่ยวชาญ เขียนสูตรทางสถิติ รวมถึง พัฒนาโปรแกรมคอมพิวเตอร์ ซึ่งเป็นเครื่องมือสำคัญในวิทยานิพนธ์ฉบับนี้ พี่น้องภาควิชาวิจัยและจิตวิทยาการศึกษาทุกๆ คน ที่ให้คำปรึกษา เป็นกำลังใจตลอดจนคอยช่วยเหลืออย่างเป็นกัลยาณมิตร

ท้ายสุดนี้ขอกราบขอบพระคุณบุคคลอันเป็นที่รักและสำคั ญอย่างยิ่ง คุณพ่อปัญญา คุณแม่ไพฑูรย์ คุณพฤตินัย สว่างบุญ และคุณอภิรัชช์ ธิติธรรมการ ที่ให้ความรัก ความห่วงใยและเป็นกำลังใจที่ดีเสมอมา

สารบัญ

หน้า

บทคัดย่อภาษาไทย..... ง
บทคัดย่อภาษาอังกฤษ..... จ
กิตติกรรมประกาศ..... ฉ
สารบัญ..... ช
สารบัญตาราง..... ฌ
สารบัญภาพ..... ฎ
บทที่
1 บทนำ..... 1
ความเป็นมาและความสำคัญของปัญหา..... 1
คำถามวิจัย..... 10
วัตถุประสงค์ของการวิจัย..... 10
สมมติฐานการวิจัย..... 11
ขอบเขตของการวิจัย..... 12
ข้อตกลงเบื้องต้น..... 14
ข้อจำกัดในการวิจัย..... 15
คำจำกัดความที่ใช้ในการวิจัย..... 16
กรอบแนวคิดในการวิจัย..... 19
ประโยชน์ที่คาดว่าจะได้รับ..... 20
2 เอกสารและงานวิจัยที่เกี่ยวข้อง..... 21
ตอนที่ 1 มโนทัศน์เกี่ยวกับการทดสอบแบบปรับเหมาะ..... 21
ตอนที่ 2 มโนทัศน์เกี่ยวกับโมเดลการตอบสนองข้อสอบแบบพหุมิติ..... 48
ตอนที่ 3 มโนทัศน์เกี่ยวกับการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์
แบบพหุมิติ..... 63
ตอนที่ 4 งานวิจัยที่เกี่ยวข้อง..... 80
3 วิธีดำเนินการวิจัย..... 95
ระยะที่ 1 การศึกษาจากการจำลองข้อมูล..... 95
ระยะที่ 2 การศึกษาจากข้อมูลทุติยภูมิ..... 102

บทที่		หน้า
	ระยะที่ 3 การศึกษาจากข้อมูลปฐมภูมิ (การทดลองใช้โปรแกรม).....	107
	สรุปวิธีการดำเนินการวิจัยจาก 3 แหล่งข้อมูล.....	116
	การวิเคราะห์ข้อมูลเพื่อตอบวัตถุประสงค์การวิจัย.....	117
4	ผลการวิเคราะห์ข้อมูล.....	119
	ตอนที่ 1 การวิเคราะห์ข้อมูลกรณีการจำลองข้อมูล.....	120
	ตอนที่ 2 การวิเคราะห์ข้อมูลกรณีการใช้ข้อมูลทุติยภูมิ.....	129
	ตอนที่ 3 การวิเคราะห์ข้อมูลกรณีการใช้ข้อมูลปฐมภูมิ.....	138
	ตอนที่ 4 สรุปผลการวิเคราะห์ข้อมูลจาก 3 แหล่งข้อมูล.....	145
5	สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ	148
	สรุปผลการวิจัย.....	148
	อภิปรายผลการวิจัย.....	153
	ข้อเสนอแนะในการนำไปใช้.....	155
	ข้อเสนอแนะในการวิจัยครั้งต่อไป.....	157
รายการอ้างอิง.....		159
ภาคผนวก.....		166
	ภาคผนวก ก รายชื่อผู้สร้างข้อสอบและผู้ตรวจสอบคุณภาพข้อสอบ…….....	167
	ภาคผนวก ข โครงสร้างข้อสอบวิทยาศาสตร์.....	168
	ภาคผนวก ค ค่าพารามิเตอร์ข้อสอบ.....	172
	ภาคผนวก ง ผลการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์.....	181
	ภาคผนวก จ เครื่องมือที่ใช้ในการวิจัย.....	186
	ภาคผนวก ฉ คำสั่งที่ใช้ในการวิเคราะห์…….	208
ประวัติผู้เขียนวิทยานิพนธ์..... .		222

สารบัญตาราง

ตารางที่		หน้า
1.1	เนื้อหาและศักยภาพที่วัดในการทดสอบความถนัดทางวิชาชีพและวิชาการด้านวิทยาศาสตร์	13
2.1	ความแตกต่างระหว่างการทดสอบแบบประเพณีนิยมและการทดสอบแบบปรับเหมาะ	23
2.2	สรุปความแตกต่างของการวิเคราะห์องค์ประกอบ (factor analysis) และ MIRT	52
2.3	สรุปประเด็นในงานวิจัยที่เกี่ยวข้องกับการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์	84
2.4	สรุปประเด็นในงานวิจัยที่เกี่ยวข้องกับการทดสอบปรับเหมาะด้วยคอมพิวเตอร์แบบพหุมิติ	92
3.1	เงื่อนไขและจำนวนกลุ่มตัวอย่างจากการจำลองข้อมูล	96
3.2	เงื่อนไขและจำนวนกลุ่มตัวอย่างจากข้อมูลทุติยภูมิ	102
3.3	โครงสร้างแบบทดสอบความถนัดทางวิชาชีพและวิชาการ PAT (professional and academic aptitude test) ศักยภาพทางวิทยาศาสตร์ (PAT2)	104
3.4	ค่าสถิติพื้นฐานของคะแนนสอบและผลการวิเคราะห์คุณภาพแบบทดสอบแบบเอกมิติ	105
3.5	ค่าพารามิเตอร์ข้อสอบ PAT2 จากการวิเคราะห์แบบพหุมิติ	107
3.6	เนื้อหาข้อสอบและจำนวนข้อสอบที่ผู้พัฒนาข้อสอบแต่ละคนสร้างขึ้น	110
3.7	โครงสร้างข้อสอบวิชาชีววิทยา	111
3.8	โครงสร้างข้อสอบวิชาเคมี	111
3.9	โครงสร้างข้อสอบวิชาฟิสิกส์	112
3.10	ค่าพารามิเตอร์ข้อสอบของแบบทดสอบความสามารถด้านวิทยาศาสตร์	113
3.11	สรุปจำนวนข้อสอบที่ได้รับสูงสุดและต่ำสุดในแต่ละวิชา	116
3.12	สรุปวิธีการดำเนินการวิจัยจาก 3 แหล่งข้อมูล	117
3.13	สรุปสถิติวิเคราะห์เพื่อตอบวัตถุประสงค์การวิจัย	118
4.1	ค่าสถิติพื้นฐานของพารามิเตอร์ข้อสอบและความสามารถผู้สอบที่จำลองขึ้น	121

ตารางที่ หน้า

4.2 ความเที่ยง (r^2) ของการทดสอบปรับเหมาะแบบพหุมิติจำแนกตามความยาวข้อสอบ (กรณีการจำลองข้อมูล)..... 122
4.3 ความเที่ยงเฉลี่ย (r^2) ของการทดสอบจำแนกตามวิธีการคัดเลือกข้อสอบและลำดับข้อสอบที่แตกต่างกัน (กรณีการจำลองข้อมูล)..... 124
4.4 การทดสอบความแตกต่างของค่าสัมประสิทธิ์สหสัมพันธ์ (กรณีการจำลองข้อมูล)..... 125
4.5 ระยะทางแบบยุคลิด (ED) ของการทดสอบปรับเหมาะแบบพหุมิติจำแนกตามความยาวข้อสอบ (กรณีการจำลองข้อมูล)..... 126
4.6 ค่าเฉลี่ยระยะทางแบบยุคลิดจำแนกตามวิธีการคัดเลือกข้อสอบขั้นแรกและลำดับข้อสอบ (กรณีการจำลองข้อมูล)..... 127
4.7 การทดสอบความแตกต่างของค่าเฉลี่ยระยะทางแบบยุคลิดด้วยการวิเคราะห์ความแปรปรวนแบบสองทางจำแนกตามวิธีการคัดเลือกข้อสอบขั้นแรก และลำดับข้อสอบ..... 128
4.8 ค่าพารามิเตอร์ข้อสอบและความสามารถของผู้สอบความถนัดทางวิชาชีพและวิชาการด้านวิทยาศาสตร์ (PAT2)..... 130
4.9 ความเที่ยง (r^2) ของการทดสอบปรับเหมาะแบบพหุมิติจำแนกตามความยาวข้อสอบ (กรณีการใช้ข้อมูลทุติยภูมิ)..... 131
4.10 ความเที่ยง (r^2) ของการทดสอบปรับเหมาะแบบพหุมิติจำแนกตามวิธีการคัดเลือก ข้อสอบและลำดับข้อสอบ (กรณีการใช้ข้อมูลทุติยภูมิ)..... 133
4.11 การทดสอบความแตกต่างของค่าสัมประสิทธิ์สหสัมพันธ์ (กรณีการใช้ข้อมูลทุติยภูมิ)..... 134
4.12 ระยะทางแบบยุคลิด (ED) ของการทดสอบปรับเหมาะแบบพหุมิติจำแนกตามความยาวข้อสอบ (กรณีการใช้ข้อมูลทุติยภูมิ)..... 135
4.13 ค่าเฉลี่ยระยะทางแบบยุคลิดจำแนกตามวิธีการคัดเลือกข้อสอบขั้นแรกและลำดับข้อสอบ (กรณีการใช้ข้อมูลทุติยภูมิ)..... 136
4.14 การทดสอบความแตกต่างของค่าเฉลี่ยระยะทางแบบยุคลิดด้วยการวิเคราะห์ความแปรปรวนแบบสองทางจำแนกตามวิธีการคัดเลือกข้อสอบขั้นแรกและลำดับข้อสอบ (กรณีการใช้ข้อมูลทุติยภูมิ)..... 137

ตารางที่		หน้า
4.15	ค่าสถิติพื้นฐานของผู้สอบความสามารถด้านวิทยาศาสตร์	138
4.16	ค่าสถิติพื้นฐานความสามารถของผู้สอบ ความคลาดเคลื่อนมาตรฐานและสารสนเทศแบบสอบ	139
4.17	ค่าสถิติพื้นฐานของคุณภาพของการทดสอบจำแนกตามลำดับข้อสอบ	140
4.18	การวิเคราะห์ความแปรปรวนตัวแปรพหุนามของคุณภาพการทดสอบปรับเหมาะแบบพหุมิติทั้ง 3 ตัวแปร ที่มีลำดับข้อสอบต่างกัน	141
4.19	การทดสอบอิทธิผลระหว่างตัวแปร (test of between-subjects effects)	142
4.20	ค่าสถิติพื้นฐานของภูมิหลังของนักเรียน	142
4.21	ผลประเมินการใช้โปรแกรมการทดสอบปรับเหมาะแบบพหุมิติ	143
4.22	ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างผลประเมินโปรแกรมเกรดเฉลี่ยสะสมและทักษะคอมพิวเตอร์	144
4.23	ผลการวิเคราะห์ความแปรปรวนที่มีผลต่อผลประเมินโปรแกรม	145
4.24	สรุปผลการเปรียบเทียบคุณภาพการทดสอบปรับเหมาะแบบพหุมิติจากข้อมูล 3 แหล่งข้อมูล	147
5.1	สรุปผลการวิเคราะห์ข้อมูลจำแนกตามวัตถุประสงค์การวิจัย	152
6.1	โครงสร้างข้อสอบวิชาชีววิทยาจากผู้พัฒนาข้อสอบ 3 คน	168
6.2	โครงสร้างข้อสอบวิชาเคมีจากผู้พัฒนาข้อสอบ 3 คน	169
6.3	โครงสร้างข้อสอบวิชาฟิสิกส์จากผู้พัฒนาข้อสอบ 3 คน	170
6.4	ค่าพารามิเตอร์ข้อสอบจากแบบวัดความถนัดทางวิชาการและวิชาชีพด้านวิทยาศาสตร์ (PAT2)	172
6.5	ค่าพารามิเตอร์ข้อสอบจากแบบวัดความสามารถวิทยาศาตร์กรณีสอบจริง	175
6.6	ค่าความสามารถของนักเรียนแต่ละคนแยกตามวิชาและคะแนนรวม	181
6.7	การคัดเลือกข้อสอบและการประมาณค่าความสามารถของผู้สอบ 1 คน	182
6.8	ค่าความสามารถของนักเรียนแต่ละคนแยกตามวิชาและคะแนนรวม (กรณีลำดับข้อสอบแบบผสมผสานมิติ)	183
6.9	ค่าความสามารถของนักเรียนแต่ละคนแยกตามวิชาและคะแนนรวม (กรณีลำดับข้อสอบแบบจำแนกมิติ)	184

สารบัญภาพ

ภาพที่		หน้า
1.1	กรอบแนวคิดในการวิจัย.....	19
2.1	ขั้นตอนการทดสอบแบบปรับเหมาะ.....	24
2.2	ประเภทของการทดสอบแบบปรับเหมาะจำแนกตามยุทธวิธีที่ใช้ในการทดสอบ..... ..	28
2.3	ลักษณะของ between-Item multidimensionality และ within-Item multidimensionality.....	65
2.4	กราฟแสดงวิธีการของ unidimensional และmultidimensional.....	66
2.5	กรอบแนวคิดในการวิจัยจากการทบทวนงานวิจัยที่เกี่ยวข้อง.....	94
3.1	ผังแสดงขั้นตอนการทำงานของโปรแกรมการทดสอบแบบปรับเหมาะแบบพหุมิติ (ลำดับข้อสอบแบบจำแนกมิติ).....	98
3.2	ผังแสดงขั้นตอนการทำงานของโปรแกรมการทดสอบแบบปรับเหมาะแบบพหุมิติ (ลำดับข้อสอบแบบผสมผสานมิติ).....	99
4.1	ความเที่ยงของการคัดเลือกข้อสอบขั้นแรกด้วยวิธี Fisher information (กรณีการจำลองข้อมูล).....	123
4.2	ความเที่ยงของการคัดเลือกข้อสอบขั้นแรกด้วยวิธี Kullback-leibler information (กรณีการจำลองข้อมูล).....	123
4.3	ความเที่ยงเฉลี่ยของวิธีการคัดเลือกข้อสอบและลำดับข้อสอบ (กรณีการจำลองข้อมูล).....	124
4.4	ระยะทางแบบยุคลิดเฉลี่ยของวิธีการคัดเลือกข้อสอบขั้นแรกทั้ง 2 วิธี (กรณีการจำลองข้อมูล).....	127
4.5	ปฏิสัมพันธ์ระหว่างการคัดเลือกข้อสอบและลำดับข้อสอบกรณีการทดสอบ (การจำลองข้อมูล).....	124
4.6	ความเที่ยงของการคัดเลือกข้อสอบขั้นแรกด้วยวิธี Fisher information (กรณีการใช้ข้อมูลทุติยภูมิ)..... .	132
4.7	ความเที่ยงของการคัดเลือกข้อสอบขั้นแรกด้วยวิธี kullback-leibler information (กรณีการใช้ข้อมูลทุติยภูมิ).....	132

ภาพที่ หน้า

4.8 ความเที่ยงเฉลี่ยของวิธีการคัดเลือกข้อสอบและลำดับข้อสอบ (กรณีการใช้ข้อมูลทุติยภูมิ)..... 133

4.9 ระยะทางแบบยุคลิดเฉลี่ยของวิธีการคัดเลือกข้อสอบขั้นแรกทั้ง 2 วิธี (กรณีการใช้ข้อมูลทุติยภูมิ)..... 135

4.10 ปฏิสัมพันธ์ระหว่างการคัดเลือกข้อสอบและลำดับข้อสอบกรณีการทดสอบ (การใช้ข้อมูลทุติยภูมิ)..... 137

บทที่ 1

บทนำ

ความเป็นมาและความสำคัญของปัญหา

การวัดผลทางการศึกษานั้นสิ่งที่เราต้องการจะวัดและตรวจสอบมักเป็นคุณลักษณะต่าง ๆ ซึ่งเป็นคุณลักษณะภายในที่ไม่สามารถสังเกตได้โดยตรง (latent traits) เช่น ความสามารถในการคิดคำนวณ ทักษะในการจับใจความ เป็นต้น ข้อสอบจะเป็นสิ่งเร้าอย่างหนึ่งที่กระตุ้นให้บุคคลแสดงพฤติกรรมภายนอกสนองตอบข้อสอบ (item response) เหล่านั้น แล้วก็พิจารณาว่าบุคคลนั้นทำข้อสอบได้ถูกต้องมากน้อยเพียงใด จากนั้นก็อนุมานพฤติกรรมการตอบซึ่งใช้จำนวนข้อสอบที่ทำถูกหรือผิดโยงไปหาความสามารถที่แท้จริง ซึ่งเป็นสมบัติภายในของเขาเอง และในการทดสอบแต่ละครั้งผู้สอบไม่ต้องการเพียงเพื่อให้รู้ผลคะแนนที่ผู้สอบทำได้เท่านั้น แต่ลึกลงไปกว่านั้นเขาต้องการรู้ด้วยว่าความสามารถที่แท้จริงของผู้สอบมีอยู่เท่าไร ดังนั้นแบบทดสอบที่ใช้วัดจึงจำเป็นต้องมีคุณภาพเชื่อถือได้ว่าคะแนนที่ได้นั้นใกล้เคียงกับความสามารถที่แท้จริงหรือมีความคลาดเคลื่อนน้อยที่สุด (สงบ ลักษณะ, 2525; วสันต์ ทองไทย, 2539)

นับเป็นเวลาอันยาวนานที่การทดสอบในด้านการศึกษาเน้นหนักไปที่การทดสอบข้อเขียน (paper and pencil) หรือที่เรียกว่าการทดสอบแบบดั้งเดิม (conventional testing) และประเมินความสามารถของผู้สอบจากวิธีการดังกล่าว ซึ่งในการทดสอบแบบดั้งเดิมการวัดจะมีความถูกต้องแม่นยำที่สุด เมื่อผู้สอบมีความสามารถปานกลาง และความคลาดเคลื่อนของการวัดจะเพิ่มขึ้นเมื่อความสามารถของผู้สอบเบี่ยงเบนจากความสามารถปานกลาง การทดสอบแบบนี้จะให้ผู้สอบทุกคนทำข้อสอบทุกๆ ข้อที่มีอยู่ในแบบทดสอบ โดยไม่คำนึงถึงว่า ข้อสอบเหล่านั้นจะเหมาะสมกับความสามารถของผู้สอบหรือไม่ เมื่อพิจารณาถึงผลกระทบในเชิงจิตวิทยาพบว่า ถ้าผู้สอบคนใดได้ตอบข้อสอบที่ยากเกินความสามารถของเขาแล้ว จะทำให้ผู้สอบเกิดความท้อถอยในการทำข้อสอบและมีแนวโน้มที่จะตอบข้อสอบโดยการเดาสูง ส่วนข้อสอบที่ง่ายเกินไปก็จะไม่ท้าทายให้ผู้สอบได้ใช้ความสามารถอย่างเต็มที่ ทำให้ผู้สอบเกิดความเบื่อหน่าย ผลกระทบเหล่านี้จะส่งผลทำให้ความถูกต้องแม่นยำของคะแนนที่ได้ลดต่ำลงและทำให้การประมาณค่าความสามารถของผู้สอบคลาดเคลื่อนไปด้วย (Weiss, 1974; van der Linden and Glas, 2000)

จากข้อจำกัดของการทดสอบแบบดั้งเดิมดังกล่าว จึงทำให้มีการคิดค้นวิธีการที่จะกำจัดความคลาดเคลื่อนในการวัดให้มีค่าน้อยที่สุด Lord and Novick (1968) ได้เสนอแนะว่าแบบทดสอบที่ใช้ในการสอบควรมีความเหมาะสมกับความสามารถของผู้สอบ โดยลักษณะเด่นของการทดสอบนั้นมุ่งที่จะจัดและคัดเลือกข้อสอบที่มีระดับความยากให้เหมาะสมกับระดับความสามารถของผู้สอบแต่ละคน โดยมีหลักการที่ว่าถ้าผู้สอบทำข้อสอบที่ยากๆ ได้ถูกต้องแล้วข้อง่ายๆก็น่าจะทำถูกด้วย จึงไม่จำเป็นต้องทำข้อสอบเหล่านั้น หรือถ้าทำข้อสอบที่ง่ายๆ ผิดแล้วข้อที่ยาก ๆ ก็น่าจะทำผิดด้วยจึงไม่ต้องทำข้อสอบข้อนั้นเช่นกัน ดังนั้นในการทดสอบครั้งหนึ่งๆ ผู้สอบแต่ละคนไม่จำเป็นต้องทำข้อสอบเหมือนกันทุกข้อ แต่จะขึ้นอยู่กับผลการตอบข้อสอบในแบบทดสอบของแต่ละคน โดยหลักทั่วไปจะให้ผู้สอบเริ่มทำข้อสอบที่มีความยากระดับปานกลางแล้วพิจารณาจากคำตอบของผู้สอบ ถ้าผู้สอบตอบถูกก็จะได้ตอบข้อสอบที่มีความยากเพิ่มขึ้น แต่ถ้าผู้สอบตอบผิดจะได้ข้อสอบที่มีความยากลดลง และในการทำข้อสอบข้อถัดไปจะปฏิบัติเช่นเดียวกันจนสิ้นสุดการสอบ (Green and others, 1984) การทดสอบดังกล่าวมีชื่อและวิธีการแตกต่างกันไป เช่น adaptive testing, branched testing, individualized testing, response contingent testing, tailor testing, computerized testing, programmed testing, computerized automate testing, sequential item testing (Weiss, 1988; Lord, 1980) คำแรกที่ใช้เรียกการทดสอบลักษณะนี้ คือ การทดสอบเทเลอร์ (tailor testing) เป็นคำที่ William W. Turnbull เป็นผู้คิดขึ้นเมื่อปี 1951 (Lord, 1980) สำหรับในประเทศไทยมีผู้นำแนวคิดนี้มาใช้ในการทดสอบโดยชื่อที่ใช้เรียกการทดสอบนี้ที่นิยมคือ การทดสอบแบบปรับเหมาะ (adaptive testing) และ การทดสอบเทเลอร์(tailor testing) ทั้งนี้แนวคิดการคัดเลือกข้อสอบให้เหมาะสมกับผู้สอบนั้นไม่ใช่เรื่องใหม่ เพราะในแบบทดสอบเชาวน์ปัญญาของบิเนท์ (Binet-Simon, 1905) ข้อคำถามได้ถูกแยกตามระดับความสามารถของสมอง คลังข้อสอบสำหรับแบบวัดเชาว์ปัญญาของบิเนท์จะใช้ระดับอายุเป็นโครงสร้างในการดำเนินการสอบนั้น เริ่มต้นโดยใช้ข้อสอบที่มีความเหมาะสมกับระดับอายุจริงของผู้สอบก่อน เมื่อตอบเสร็จก็ตรวจให้คะแนนทันที ถ้าผู้สอบตอบได้ถูกต้องข้อสอบข้อต่อไปจะเป็นข้อที่เหมาะกับระดับอายุที่สูงขึ้น แต่ถ้าตอบผิด ข้อสอบข้อต่อไปจะเหมาะกับผู้สอบที่มีระดับอายุ ต่ำกว่า การทดสอบจะมีลักษณะเช่นนี้ไปเรื่อยๆ จนกระทั่งถึง ระดับอายุที่ผู้สอบถูกระบุว่ามีความสามารถ อยู่ถึงระดับนั้น ความเป็นจริงแล้วการทดสอบ แบบปรับเหมาะมีความเก่าแก่เช่นเดียวกับการฝึกการทดสอบแบบปากเปล่า (oral examination) ผู้ดำเนินการทดสอบ แบบปากเปล่าที่ดี จะต้องรู้จัก ผู้สอบและเลือกคำถามที่มีความเหมาะสมที่มีผลต่อระดับความรู้ของผู้สอบ (van der Linden and Glas, 2000)

นับตั้งแต่ปี 1980 เป็นต้นมา การทดสอบที่มี สเกลขนาดใหญ่และมีผลกระทบสูง (high-stake testing) ได้พัฒนาไปสู่การใช้คอมพิวเตอร์ส่วนบุคคล (personal computers: PCs) ทำให้รูปแบบของการทดสอบถูกเลือกอย่างเหมาะสมและมีความรวดเร็วมากขึ้น การใช้คอมพิวเตอร์เป็นฐานในการส่งผ่านข้อสอบนั้นมีข้อดีจำนวนมากยกตัวอย่างเช่น การทดสอบนี้อนุญาตให้ทำการทดสอบได้ตามต้องการ นั่นคือ เมื่อไหร่ หรือที่ไหนก็ได้ ที่ผู้สอบมีความพร้อมที่จะทำแบบสอบ นอกจากนั้นคอมพิวเตอร์ยังสามารถใช้เพิ่มความแม่นยำในทางสถิติในการให้คะแนนการทดสอบ ซึ่งรวมถึงพัฒนาการของการทดสอบแบบปรับเหมาะในปัจจุบัน ที่นำคอมพิ วเตอร์และทฤษฎีการตอบสนองข้อสอบ (Item response theory) เข้ามาช่วยในการวิเคราะห์ข้อมูลและดำเนินการสอบ และเรียกการทดสอบนี้ว่า การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (computerized adaptive testing: CAT) โดยแทนที่จะให้ผู้สอบทุกคนใช้ข้อสอบชุดเดียวกัน แต่การเลือกข้อ สอบของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (CAT) จะปรับเปลี่ยนไปตามระดับความสามารถของผู้สอบแต่ละคน หลังจากที่ผู้สอบตอบข้อสอบในแต่ละข้อ วิธีการทดสอบแบบปรับเหมาะจะประมาณค่าความสามารถของผู้สอบ โดยปรับเปลี่ยนไปตามการตอบข้อสอบแต่ละข้อ และลำดับของการคั ดเลือกข้อสอบจะคัดเลือกจากข้อสอบที่มีคุณสมบัติเหมาะสมที่สุด ที่จะใช้ในการประมาณค่าความสามารถครั้งใหม่ตามทฤษฎีการตอบสนองข้อสอบ ซึ่งวิธีการดังกล่าวจำเป็นอย่างยิ่งที่ต้องใช้คอมพิวเตอร์เข้ามาช่วยในการประมวลผลในแต่ละขั้นตอน (van der Linden and Glas, 2000)

การพั ฒนาที่เพิ่มขึ้นและการปรับให้สอดคล้องกับเทคนิคการวัดทางจิตวิทยาได้ถูกนำมาใช้ในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ในช่วง 10 ปีที่ผ่านมา ทั้งนี้เนื่องมาจากความสามารถของคอมพิวเตอร์ในช่วงแรกที่ทำงานช้าและไม่อนุญาตให้ประมาณค่าความสามารถในสภาพความเป็นจริง ดังนั้นการศึกษาวิจัยก่อนหน้านี้จึงหาวิธีการประมาณค่าความสามารถหรือหารูปแบบใหม่ที่จะนำไปใช้กับการทดสอบแบบดั้งเดิม ตัวอย่างเช่น รูปแบบการทดสอบสองขั้นตอน (Cronbach & Gleser, 1965) การคัดเลือกข้อสอบตามวิธีของเบส์ ด้วยการประมาณค่า posterior distribution ของพารามิเตอร์ความสามารถ (Owen, 1969) การคัดเลือกข้อสอบด้วยวิธี up-and-down (Lord, 1970) วิธีการโรบิน-มอนโร (Lord, 1971) การทดสอบเฟลกซิเลเวล (flexilevel test) (Lord, 1971) การทดสอบปรับระดับชั้น (stradaptive test) (Weiss, 1973) และการทดสอบแบบปรับเหมาะแบบปิรามิด (Larkin & Weiss, 1975)

ในเวลาต่อมา เมื่อความสามารถของคอมพิวเตอร์เพิ่มมากขึ้นการประยุกต์ใช้การทดสอบแบบปรับเหมาะที่มีสเกลขนาดใหญ่และมีผลกระทบสูง (high-stake testing) จึงมีความเป็นไปได้และสะดวกมากขึ้น ผู้บุกเบิกวิธีการนี้ในการทดสอบขนาดใหญ่ คือ กองทัพบกของสหรัฐอเมริ กาที่ใช้ในการทดสอบ Armed Services Vocational Aptitude Battery (ASVAB) ทั้งนี้หลังจากที่ ได้มีการปรับปรุงการทดสอบ ASVAB ครั้งแรกในปี 1979 การทดสอบ ASVAB จึงได้ปรับเปลี่ยนมาอยู่ในรูปของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (CAT) ซึ่งเริ่มใช้งานได้จริงในช่วงกลา งของปี 1980-1989 อย่างไรก็ตามการเปลี่ยนจากการทดสอบด้วยแบบเขียนตอบบนกระดาษคำตอบ (paper and pencil) มาเป็นการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (CAT) ได้เริ่มต้นขึ้นอย่างแท้จริงเมื่อคณะกรรมการสภาพยาบาลแห่งชาติ (national council of state boards of nursing) ได้เริ่มดำเนินการทดสอบเพื่อให้การรับรอ งหรือให้ใบอนุญาตในรูปแบบข อง CAT หลังจากนั้นก็มีการทดสอบที่ปรับเปลี่ยนมาเป็นการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ตามมาอีกจำนวนมาก อาทิ Graduate record examination (GRE), Graduate Management Admission Test (GMAT) ซึ่งข้อดีหลั กๆของ CAT คือ ความสามารถในการประมาณค่าความสามารถของผู้สอบได้อย่างมีประสิทธิภาพและมีความแม่นยำมากกว่าทดสอบแบบประเพณีนิยม (Wainer, 1990 อ้างถึงใน Chen, Ankenmann, & Chang, 2000)นอกจากนั้นสถาบันบริการทดสอบทางการศึกษา(education testing service: ETS, 1996) ได้เสนอเหตุผลที่การทดสอบขนาดใหญ่ปรับเปลี่ยนมาใช้รูปแบบของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ไว้ดังนี้ (1) CAT มีความสะดวกต่อผู้สอบที่สามารถจัดตาราง การสอบได้ตามความต้องการ (2) การทดสอบนี้ดำเนินการได้ง่าย สะดวก และใช้ผู้บริหารการสอบจำนวนน้อยกว่ าการทดสอบแบบดั้งเดิม (3) กระบวนการทางอิเล็กทรอนิกส์สามารถรายงานผลคะแนนได้อย่างรวดเร็ว (4) ขอบเขต ของคำถามกว้างและเนื้อหาของข้อสอบยังสามารถนำใส่เข้าไปเพื่อนำมาใช้ต่อไปได้ การทดสอบแบบ CAT ในปัจจุบัน มีผู้ให้การยอมรับถึงข้อดีของวิธีการดังกล่าว และ ยังได้รับความชื่นชมจากผู้สอบ จะเห็นได้จากการเสนอการทดสอบแบบเลือกตอบในข้อสอบ ชุดเดียวกันระหว่างการทดสอบแบบดั้งเดิม และการทดสอบแบบปรับ เหมาะด้วยคอมพิวเตอร์ ผู้สอบส่วนใหญ่จะเลือกการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (van der Linden and Glas, 2000)

อย่างไรก็ตาม Segall (1996) ได้กล่าวว่า แม้การทดสอบแบบปรับเหมาะจะเป็นการทดสอบที่เพิ่มความแม่นยำในการวัด ลดระยะเวลาในการทดสอบ วิธีใช้ได้มาตรฐานและมีความยืดหยุ่นในการจัดตารางการสอบของผู้สอบ แต่อย่างไรก็ตามการทดสอบแบบปรับเหมาะส่วนใหญ่จะใช้การเลือกข้อคำถามและการคำนวณการให้ คะแนนโดยใช้พื้นฐานของทฤษฎีการตอบสนอง

ข้อสอบ (item response theory: IRT) ซึ่งในขณะนี้เทคนิคดังกล่าวใช้กับการทดสอบแบบ ปรับเหมาะภายใต้ ข้อตกลงเบื้องต้นที่การทดสอบจะต้องเป็นรูปแบบเอกมิติ (unidimensionality) ซึ่งการตอบสนองข้อสอบแบบเอกมิติในบางครั้งจะไม่ เหมาะสมกับการทดสอบจริง (Wang & Chen, 2004) เนื่องจาก การทดสอบส่วนใหญ่มีลักษณะเป็นพหุมิติ(multidimensional) การทบทวนรูปแบบใหม่ๆ ในการประเมิน และความเกี่ยวข้องในการให้คะแนน ได้กล่าวถึงความประทับใจเกี่ยวกับโมเดลพหุมิติว่า วิธีดังกล่าวได้อธิบายความสามารถของผู้ สอบได้อย่างมีประสิทธิภาพ (van der Linder & Hambleton, 1997) โมเดลการตอบสนองข้อสอบแบบพหุมิติเสนอไว้เพื่อศึกษาประเด็นการประเมินรูปแบบใหม่ที่มีความซับซ้อน เมื่อทฤษฎีการตอบสนองข้อสอบแบบพหุมิติเริ่มได้รับการยอมรับและการทดสอบแบบปรับเหมาะ กำลังได้รับความนิยมอย่างมากในการนำไปปฏิบัติ จึงเกิดการรวมกันของทั้งสองวิธีนี้ที่เรียกว่า การทดสอบปรับเหมาะแบบพหุมิติ (Multidimensional Adaptive testing: MAT) (Segall, 1996, 2000) ซึ่งถือเป็นทิศทางใหม่ที่ใช้เพื่อตรวจสอบ ดังนั้นปัจจุบันจึงมีผู้เริ่มศึกษาการทดสอบแบบปรับเห มาะในลักษณะของการตอบสนองแบบพหุมิติ (multidimensional IRT) มากขึ้น

มีนักวิจัย จำนวนมากที่ได้ศึกษาในประเด็นเกี่ยวกับการท ดสอบแบบเอกมิติในทฤษฎีการตอบสนองข้อสอบ งานวิจัยที่ผ่านมาส่วนใหญ่จ ะให้ความสำคัญที่การศึกษาผลลัพธ์ของการใช้โมเดลการตอบส นองข้อสอบแบบเอกมิติ (unidimensional IRT) ในข้อมูลที่มีลักษณะเป็นพหุมิติ (multidimensional) (Ackerman, 1989, 1991; Ansley & Forsyth, 1985; Drasgow & Parsons, 1993; Folk & Green, 1989: Harrison, 1988; Reckase, 1979; Reckase, Ackerman & Carlson, 1988; Way, Ansley & Forsyth, 1088; Yen, 1984 อ้างถึงใน Segall, 1996) ทั้งนี้มีผู้ให้การยอมรับในการพัฒนาและประเมินวิธีการประมาณค่าความสามารถในการทดสอบแบบปรับเหมาะแบบพหุมิติ (multidimensional adaptive testing: MAT) จากการศึกษาของ Bloxom and Vale (1987) และ Tam (1992) โดยกระบวนการพัฒนาของ Bloxom และ Vale เป็นการขยาย Multivariate ในวิธีการ sequential Bayesian adaptive updating ของ Owen (1975) และการศึกษาของ Tam (1992) ได้เปรียบเทียบ ตัวประมาณค่าที่ใช้เกณฑ์ที่หลากหลายรวมถึงความแม่นยำ สารสนเทศของแบบสอบ และเวลาในการคำนวณ ซึ่ งการศึกษาของเขาจะเน้นไปที่ความสามารถของวิธีการประมาณค่าความสามารถ และวิธีการคัดเลือกข้อสอบ ต่อมา Segall (1996) ได้ศึกษาตามรูปแบบของการทดสอบ Armed Services Vocational Aptitude Battery (ASVAB) โดยการจำลองข้อมูล ผลการวิจัยให้หลักฐานที่ชัดเจนว่า หลักการของ MCAT (Multidimensional computerized adaptive testing) นั้นมีประโยชน์มาก นั่นคือ การเพิ่ม

ประสิทธิภาพการวัด ซึ่งสามารถเป็นที่ประจักษ์ต่อบุคคลอื่นๆ เกี่ยวกับการลดความยาวของข้อสอบ หรือการวัดผลที่มีความแม่นยำ โดยผลลัพธ์ชี้ว่า MCAT ให้ความแม่นยำที่เทียบเท่าหรือสูงกว่า UCAT (unidimensional computerized adaptive testing) เมื่อลดจำนวนข้อสอบลง 1 ใน 3 และผลลัพธ์ยังชี้ว่า เมื่อให้จำนวนข้อสอบเป็นค่าคงที่พบว่า MCAT สามารถเพิ่มความเที่ยงได้อย่างมาก

นอกจากนั้น Segall (2000) ได้ยกตัวอย่างการนำแนวคิดการตอบสนองข้อสอบแบบพหุมิติมาใช้ในสถานการณ์จริงที่แตกต่างจากข้างต้น โดยเขาได้กล่าวว่าแบบทดสอบที่ใช้วัดความแตกต่างระหว่างบุคคล ถูกออกแบบขึ้นเพื่อให้ได้สารสนเทศที่กว้างขวางตามมิติในด้านต่างๆ ไม่ว่าจะเป็นความรู้ ทักษะ และความสามารถ ยกตัวอย่างเช่น การสอบเข้ามหาวิทยาลัย อาจมี การให้คะแนนแยกกันระหว่างมิติด้านคณิตศาสตร์และด้านภาษา และบางมหาวิทยาลัยอาจจะคัดเลือกโดยใช้คุณสมบัติอื่นๆเข้ามาทดแทน โดยที่คะแนนรวมของผู้สมัคร (คณิตศาสตร์รวมกับภาษา) จะต้องดีกว่าจุดตัดที่ชี้เฉพาะ ในกรณีนี้คะแนนคณิตศาสตร์และภาษาของแต่ละบุคคลจะให้ผลสะท้อนกลับที่มีประโยชน์ต่อนักเรียนและโรงเรียนเกี่ยวกับจุดแข็งและจุดอ่อนของความถนัดและหลักสูตร ในกรณีอื่นๆ เช่น มหาวิทยาลัยอาจจะคัดเลือกโดยใช้คุณสมบัติของผู้สอบ ซึ่งคะแนนของผู้สอบบนองค์ประกอบที่ถูกเลือกจะดีกว่าการแยกจุดตัดในแต่ละมิติ ยกตัวอย่างเช่น มหาวิทยาลัยอาจจะเลือกคุณสมบัติที่เป็นมาตรฐาน 1 ประเด็น สำหรับความรู้ด้านคณิตศาสตร์และมาตรฐานอื่นๆ สำหรับประสิทธิภาพด้านภาษา ผู้สมัครอาจจะถูกบังคับ 1 มาตรฐานหรือทั้ง 2 มาตรฐาน ที่จะมีคุณสมบัติในการเข้ามหาวิทยาลัย ซึ่งตัวอย่างทั้งหมดนั้นมีประโยชน์และสำคัญ สำหรับคะแนนองค์ประกอบ ของแต่ละบุคคลที่จะได้มาซึ่งคุณสมบัติการวัดประสิทธิภาพที่เพียงพอ รวมถึงความแม่นยำและความตรง เมื่อมิติต่าง ๆ ที่ถูกวัดโดยแบบสอบหรือชุดของแบบสอบมีความสัมพันธ์กัน การตอบสนองการวัดข้อคำถามใน 1 มิติจะให้หลักฐานเกี่ยวกับจุดยืนหรือความสามารถของผู้สอบ จะเห็นได้ว่าผู้ สอบที่มีความสามารถเกี่ยวกับคำศัพท์ในระดับสูงมีความเป็นไปได้ที่จะมีความสามารถในการอ่านสูง ความเข้าใจเกี่ยวกับขนาดความสัมพันธ์ระหว่างมิติในประชากรที่สนใจ นอกจากนั้นระดับความสามารถของแต่ละบุคคลสามารถเพิ่มแหล่งสารสนเทศที่มีลักษณะเฉพาะ และถ้าใช้วิธีที่มีความเหมาะสมก็สามารถจะนำไปสู่การประเมินค่าความสามารถที่มีความถูกต้องแม่นยำมากขึ้น นอกจากนั้นยังพบว่า สารสนเทศข้ามกลุ่ม (cross-information) ถูกละเลยจากวิธีการให้คะแนน แบบประเพณีนิยม และจากการคัดเลือกข้อสอบแบบเอกมิติ รวมถึงวิธีการให้คะแนนที่ใช้ในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ สารสนเทศข้ามกลุ่ม (cross-information) จะถูกเก็บรวบรวมมาจากข้อสอบที่มีมิติสัมพันธ์กันที่สามารถดำเนินการตามทฤษฎี

การตอบสนองข้อสอบแบบพหุมิติ (multidimensional item response theory: MIRT) ได้อย่างมีประสิทธิภาพ ในกรณีของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ สารสนเทศนี้สามารถช่วยวัดผลใน 2 วิธี วิธีแรก สารสนเทศช่วยในการคัดเลือกข้อสอบที่จะนำไปสู่ทางเลือกของข้อสอบที่มีสารสนเทศมาก วิธีที่สอง สารสนเทศสามารถช่วยประมาณค่าความสามารถที่จะนำไปสู่คะแนนสอบที่มีความแม่นยำเพิ่มขึ้น

สำหรับการเพิ่มประสิทธิภาพการวัด MCAT สามารถเกิดขึ้นเมื่อมิติที่ถูกวัดโดยชุดข้อสอบมีความสัมพันธ์กันไม่เป็นศูนย์ ในกรอบแนวคิดการประมาณค่าของเบส์ (Bayesian) การตอบสนองข้อสอบของการทดสอบหนึ่งสามารถให้สารสนเทศเกี่ยวกับระดับประสิทธิภาพการวัด ด้วยแบบทดสอบอื่นๆในชุดข้อสอบ เมื่อให้ p มิติมีความสัมพันธ์กัน นอกจากนั้นสิ่งที่เหนือกว่าการประมาณค่าความสามารถของมิติ p แบบแยก นั่นคือ วิธี MCAT ให้เวคเตอร์การประมาณค่าความสามารถของมิติ p สำหรับผู้สอบแต่ละคน หลังจากทำข้อสอบแต่ละข้อก็จะมีการปรับปรุงเวคเตอร์ของมิติ p ยกตัวอย่างเช่น ชุดข้อสอบที่ประกอบด้วยแบบสอบย่อย ที่ใช้วัดความสามารถในการอ่านและคำศัพท์ โดยจะพบว่าการตอบคำถามเกี่ยวกับคำศัพท์ได้ถูกต้องก็จะเพิ่มความสามารถที่ประมาณค่าได้แบบชั่วคราวในมิติของคำศัพท์และความสามารถในการอ่าน เมื่อมิติทั้งสองมีความสัมพันธ์กัน การตอบข้อสอบเกี่ยวกับคำศัพท์ไม่เพียงแต่ให้สารสนเทศในมิติเกี่ยวกับคำศัพท์เท่านั้นแต่ยังให้สารสนเทศเกี่ยวกับระดับความสามารถในการอ่าน รวมถึงระดับความสามารถในมิติอื่นๆ ที่มีความสัมพันธ์กับมิติคำศัพท์ ประโยชน์ของ MCAT เป็นการเพิ่มสารสนเทศที่ได้จากข้อสอบที่มีมิติสัมพันธ์กัน ซึ่งสามารถนำไปสู่ประสิทธิภาพการวัด เป็นที่ปรากฏชัดเกี่ยวกับความแม่นยำ หรือการลดความยาวของข้อสอบ นอกจากความสัมพันธ์ของมิติจะมีผลต่อประสิทธิภาพการวัดแล้ว วิธีการคัดเลือกข้อสอบ ก็มีผลต่อความแม่นยำของการประมาณค่าความสามารถของผู้สอบ โดยเฉพาะการคัดเลือกข้อสอบในขั้นแรก ซึ่งในการทดสอบปรับเหมาะแบบเอกมิติ (UCAT) มีการศึกษาวิธีการที่เหมาะสม (Chen, Ankerman and Chang, 2000; Chen and Ankenmann, 2004) แต่ยังไม่มีความชัดเจน และผลการวิจัยยังไม่มีความสอดคล้องกัน

สำหรับสารสนเทศของการคัดเลือกข้อสอบ Chang และ Ying (1996) ได้แบ่งสารสนเทศของข้อสอบออกเป็น 2 ประเภทคือ Local information และ Global information โดยสารสนเทศที่จัดอยู่ใน Local information คือ Fisher information ซึ่งสารสนเทศจะอยู่รอบๆ θ ที่แท้จริงหรือมีลักษณะเป็นสารสนเทศแบบจุด (The item information at point) ควรใช้ในกรณีที่มี n ขนาดใหญ่ สำหรับสารสนเทศที่จัดอยู่ใน Global information คือ Kullback-Leiber information ซึ่งสารสนเทศจะอยู่ภายนอกขอบเขตของ θ ที่แท้จริงหรือมีลักษณะเป็นเกณฑ์สารสนเทศแบบช่วง (Interval

information criterion) เหมาะสำหรับใช้ในกรณีที่มีn ขนาดเล็ก นั่นคือ เหมาะกับข้อสอบจำนวนน้อย อย่างไรก็ตามการศึกษาดังกล่าวอยู่ภายใต้ทฤษฎีของการทดสอบปรับเหมาะแบบเอกมิติและในช่วง 10 ปี ที่ผ่านมานี้ มีการพัฒนาวิธีการที่ใช้ในการทดสอบปรับเหมาะแบบเอกมิติมาใช้ในการทดสอบปรับเหมาะแบบพหุมิติ ซึ่งเป็นการพัฒนามาจาก 2 แนวคิดหลัก คือ Fisher information และ Kullback-Leibler information เช่น Fisher interval information (Berger & Veerkamp, 1997) Fisher information with a posterior distribution (van der Linden, 1998) และ Kullback-Leibler information with a posterior distribution (Chang & Ying, 1996) และเพื่อให้ได้คำตอบที่ชัดเจนเกี่ยวกับวิธีการคัดเลือกข้อสอบขั้นแรกสำหรับการทดสอบปรับเหมา ะแบบพหุมิติ ผู้วิจัยจึงสนใจที่จะศึกษาการคัดเลือกข้อสอบจาก 2 แนวคิดหลัก คือ Fisher information และ Kullback-Leibler information เพื่อเป็นตัวแทนของวิธีการคัดเลือกข้อสอบวิธีอื่นๆ

อย่างไรก็ตามการนำ MCAT มาประยุกต์ใช้ก็สามารถให้ประโยชน์ เพิ่มเติม สิ่งนั้นรวมถึงการใช้ MCAT ในการควบคุมความสมดุลของเนื้อหาในข้อ สอบระหว่างผู้สอบแต่ละคน โดย Segall (1996) ได้กล่าวว่า ข้อดี MCAT ที่ไม่สามารถโต้แย้งได้ นั่นคือ MCAT เป็นวิธีที่มีประสิทธิภาพ โดยทำให้มั่นใจว่าข้อสอบที่ได้รับจากการทดสอบแบบปรับเหมาะนั้นมีเนื้อหาครอบคลุม ยกตัวอย่าง เช่น ในการสอบวิทยาศาสตร์ทั่วไป การสอบนี้ย่ อมต้องการให้ผู้สอบ แต่ละคนได้รับจำนวนข้อสอบที่เพียงพอในศาสตร์ชีววิทยา ฟิสิกส์ และเคมี ซึ่งการทดสอบนี้มีแนวโน้มที่จะแยกมิติของความเชี่ยวชาญด้านวิทยาศาสตร์(แต่ละมิติมีความสัมพันธ์กัน) วิธีการทั่วไปในการทดสอบแบบปรับเหมาะด้วยวิธีเอกมิติ จะมีวิธีการคัดเลือกข้อ สอบให้กับผู้สอบโดยจำกัดจำนวนข้อสอบในแต่ละขอบเขตเนื้อหา อย่างไรก็ตาม วิธีการนี้อาจเกิดปัญหาขึ้นได้ถ้าเนื้อหานั้นมีความไม่ชัดเจนเรื่องความยากของข้อสอบ ตัวอย่างเช่นถ้าข้อสอบเคมียากกว่าข้อสอบชีววิทยาหรือฟิสิกส์ การดำเนินการสอบหากให้ข้อสอบเคมีกับผู้สอบความสามาร ถต่ำ จะทำให้ได้รับสารสนเทศเกี่ยวกับระดับความเชี่ยวชาญในด้านวิทยาศาสตร์น้อย และนำไปสู่การลดประสิทธิภาพในการวัด วิธีที่ดีกว่าการคัดเลือกข้อสอบให้กับผู้สอบโดยจำกัดจำนวนข้อสอบในแต่ละขอบเขตเนื้อหา คือการใช้แบบทดสอบปรับเหมาะแบบพหุมิติ โดยจะดำเนินการกับ 3 มิตินี้ (ชีววิทยา ฟิสิกส์ และเคมี) แบบแยก ทั้งนี้แต่ละมิติต้องมีความสัมพันธ์กันสูง โดยการรวมกันของสารสนเทศจากหลายแหล่งในทุกมิติไปพร้อมๆกัน รวมถึงความเชี่ยวชาญของผู้สอบ สารสนเทศของผู้สอบ และ การแจกแจงร่วมกันก่อนหน้าของความสามารถ (prior joint-distribution of ability) โดย Segall (1996) ได้กล่าวไว้ว่า สิ่งที่เป็นกังวลอย่างหนึ่งเกี่ยวกับการใช้ MCAT คือ ผลกระทบที่เกิดจากบริบทข้อสอบหรือ ผลกระทบของลำดับข้อสอบ (item context effects or affect of item-order) โมเดลที่นำเสนอนี้เป็นไปตาม ข้อตกลงเบื้องต้นเกี่ยวกับ local independent เนื่องจากลำดับการนำเสนอข้อสอบไม่มีผลต่อค่า อำนาจจำแนกและ

ความยากของข้อสอบ แต่ความยาก ของข้อสอบหรืออำนาจจำแนก มากน้อยจะขึ้นกับเนื้อหาของข้อสอบที่ผู้สอบได้รับ ดังนั้น ลำดับของข้อสอบจึงมีผลต่อการตอบข้อสอบข้อถัดไป ผลกระทบนี้ส่งผลต่อปฏิกิริยาตอบสนองของผู้สอบ รวมถึงความแม่นยำของการประมาณค่าความสามารถ ซึ่งปัญหานี้จะไม่เกิดกับการทดสอบแบบเอกมิติ เนื่องจากการทดสอบนั้นวัดเพียงมิติเดียว อย่างไรก็ตามยังไม่มีผู้ศึกษาอย่างชัดเจนว่า ลำดับข้อสอบจะมีผลกระทบต่อการประมาณค่าความสามารถของผู้สอบหรือไม่ และลำดับข้อสอบแบบใดจะส่งผลต่อคุณภาพการทดสอบที่สูงกว่า

จากที่กล่าวมาข้างต้นทำให้ผู้วิจัยสนใจศึกษาคุณภาพของการทดสอบแบบปรับเหมา ะด้วยคอมพิวเตอร์แบบพหุมิติ โดยศึกษาเปรียบเทียบวิธีการคัดเลือกข้อสอบขั้นแรกที่ยังไม่มีข้อสรุปว่าวิธีการใดที่มีความเหมาะสมกับการทดสอบปรับเหมาะแบบพหุมิติ รวมถึงศึกษา ลำดับข้อสอบที่มีต่อคุณภาพของการทดสอบแบบปรับเหมาะ ซึ่งความแตกต่างของ ลำดับข้อสอบจะเกิดขึ้นเฉพาะกับการทดสอบแบบพหุมิติเท่านั้น เนื่องจากการทดสอบแบบเอกมิติจะวัด เพียงเนื้ อหาเดียว โดยการศึกษาครั้งนี้ใช้แหล่งข้อมูล 3 แหล่ง คือ การจำลองข้อมูล ข้อมูลทุติยภูมิ และข้อมูลปฐมภูมิ โดยที่ข้อมูลทุติยภูมิได้จากการทดสอบระดับชาติของนักเรียนที่จะเข้าศึกษาต่อในระดับอุดมศึกษา โดยเลือกใช้การทดสอบความถนัดทาง วิชาชีพและวิชาการ(professional and academic aptitude test: PAT) ด้านวิทยาศาสตร์ (PAT2) ซึ่งแบบทดสอบมี ลักษณะเป็นแบบพหุ มิติ เนื่องจากแบบทดสอบวัดเนื้อหาวิทยาศาสตร์แยกเป็น 4 ด้าน ได้แก่ 1) ชีววิทยา สิ่งแวดล้อม 2) เคมี 3) ฟิสิกส์ และ 4) โลก ดาราศาสตร์ อวกาศ และ วัดศักยภาพ วิทยาศาสตร์ เกี่ยวกับ การคิดแบบนักวิทยาศาสตร์ การแก้ปัญหาแบบนักวิทยาศาสต ร์ และทักษะการอ่านบทความแบบนักวิทยาศาสตร์ อย่างไรก็ตามการศึกษาครั้งนี้ พิจารณาเพียง 3 มิติเท่านั้น เพื่อความชัดเจนของการตีความและง่ายต่อความเข้าใจ นอกจากนี้ เพื่อให้ผลการทดสอบมีความชัดเจนมากขึ้นและเห็นแนวทางในการนำมาปฏิบัติจริง ผู้วิจัยจึงพัฒนาให้มีการท ดสอบ ในสถานการณ์จริง (ข้อมูลปฐมภูมิ) โดยสร้างคลังข้อสอบขึ้นมาใหม่เพื่อการทดสอบครั้งนี้ และเก็บรวบรวมข้อมูลจากนักเรียนชั้นมัธยมศึกษาปีที่ 5 ซึ่งการศึกษาครั้งนี้จ ะเป็นแนวทางในการนำการทดสอบแบบปรับเหมาะมาใช้กับการทดสอบขนาดใหญ่ ที่มีผลกระทบสูงในประเทศไทยต่อไป

คำถามวิจัย

1. วิธีการคัดเลือกข้อสอบขั้นแรกวิธีใด ระหว่างวิธี Fisher information และ Kullback-Leibler information ที่มีคุณภาพสูงที่สุด เมื่อพิจารณาจากความเที่ยง และระยะทางแบบยุค ลิด สำหรับการทดสอบปรับเหมาะด้วยคอมพิวเตอร์แบบพหุมิติ

2. ลำดับข้อสอบที่มีรูปแบบแตกต่างกัน ระหว่างลำดับข้อสอบแบบจำแนกมิติ และลำดับข้อสอบแบบผสมผสานมิติ จะมีผลต่อคุณภาพของการทดสอบหรือไม่อย่างไร เมื่อพิจารณาคุณภาพการทดสอบปรับเหมาะแบบพหุมิติจาก ความเที่ยง ระยะทางแบบยุคลิด สารสนเทศแบบสอบ และความยาวของข้อสอบ

3. เกิดปฏิสัมพันธ์ระหว่างวิ ธีการคัดเลือกข้อสอบขั้นแรกและ ลำดับข้อสอบที่ส่งผลต่อคุณภาพของการทดสอบหรือไม่

4. นักเรียนที่ทำการทดสอบผ่านโปรแกรมการทดสอบปรับเหมาะแบบพหุมิติจะมีความพึงพอใจต่อการทดสอบมากน้อยเพียงใด และภูมิหลังของผู้สอบ จะส่งผลต่อผลการประเมินโปรแกรมการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์ หรือไม่

วัตถุประสงค์การวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์ภายใต้วิธีการคัดเลือกข้อสอบขั้นแรกแล ะลำดับข้อสอบ ที่แตกต่างกัน โดยศึกษาจาก การจำลองข้อมูล ข้อมูลทุติยภูมิและ ข้อมูลปฐมภูมิ รวมถึงศึกษาผลการประเมินโปรแกรมการทดสอบการทดสอบปรับเหมาะแบบพหุมิติจากนักเรียน ซึ่งมีวัตถุประสงค์ย่อยดังนี้

1. เพื่อเปรียบเทียบคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์ภายใต้วิธีการคัดเลือกข้อสอบขั้นแรกและลำดับข้อสอบที่แตกต่างกัน

1.1 เปรียบเทียบวิธีการคัดเลือกข้อสอบขั้นแรกที่แตกต่างกัน 2 วิธี ได้แก่

1) Fisher information (FI)

2) Kullback-Leibler information (KL)

1.2 เปรียบเทียบลำดับข้อสอบที่แตกต่างกัน 2 รูปแบบ ได้แก่

1) ลำดับข้อสอบแบบจำแนกมิติ

2) ลำดับข้อสอบแบบผสมผสานมิติ

1.3 ศึกษาปฏิสัมพันธ์ระหว่างวิธีก ารคัดเลือกข้อสอบขั้นแรกและ ลำดับข้อสอบ ที่ส่งผลต่อคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์

2. เพื่อศึกษาผลการประเมิน โปรแกรม การท ดสอบความสามารถ ด้านวิทยาศาสตร์ด้วยวิธีการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์

สมมติฐานการวิจัย

1. Chang และ Ying (1996) ได้แบ่งสารสนเทศของข้อสอบออกเป็น 2 ประเภทคือ Local information และ Global information โดยสารสนเทศที่จัดอยู่ใน Local information คือ Fisher information ซึ่งสารสนเทศจะอยู่รอบๆ θ ที่แท้จริงหรือมีลักษณะเป็นสารสนเทศแบบจุด (The item information at point) ควรใช้ในกรณีที่มี n ขนาดใหญ่ สำหรับสารสนเทศที่จัดอยู่ใน Global information คือ Kullback-Leiber information ซึ่งสารสนเทศจะอยู่ภายนอกขอบเขตของ θ ที่แท้จริงหรือมีลักษณะเป็นเกณฑ์สารสนเทศแบบช่วง (Interval information criterion) เหมาะสำหรับใช้ในกรณีที่มี n ขนาดเล็ก นั่นคือ เหมาะกับข้อสอบจำนวนน้อย ดังนั้นการคัดเลือกข้อสอบขั้นแรกในบริบทของการทดสอบแบบเอกมิติ วิธี Kullback-Leiber information จึงมีความเหมาะสมกว่า วิธี Fisher information แต่เนื่องจากที่ผ่านมายังไม่มีผู้ศึกษาวิธีการคัดเลือกข้อสอบขั้นแรก ในการทดสอบปรับเหมาะ แบบพหุมิติ ผู้วิจัยจึงใช้ข้อมูลจากการศึกษาแบบเอกมิติเพื่อตั้งสมมติฐานการวิจัยครั้งนี้ ดังนั้นผู้วิจัยจึงตั้งสมมติฐานข้อที่ 1 ดังนี้

สมมติฐานข้อ 1 วิธีการคัดเลือกข้อสอบขั้นแรกด้วยวิธี Kullback-Leiber information น่าจะมีคุณภาพการทดสอบที่สูงกว่าวิธีการคัดเลือกข้อสอบขั้นแรกด้วยวิธี Fisher information

2. Segall (1996) ได้กล่าวไว้ว่า สิ่งที่เป็นกังวลอย่างหนึ่งเกี่ยวกับการใช้ MCAT คือ ผลกระทบที่เกิดจาก ลำดับข้อสอบ (item order effects) โมเดลที่นำเสนอนี้ เป็นไปตามข้อตกลงเบื้องต้นเกี่ยวกับ local independent นั่นคือ ข้อสอบทุกข้อมีความเป็นอิสระจากกัน เนื่องจากลำดับการนำเสนอข้อสอบไม่มีผลต่อค่าอำนาจจำแนกและความยากของข้อสอบ แต่ความยากของข้อสอบหรืออำนาจจำแนกมากน้อยจะขึ้นกับเนื้อหาของข้อสอบที่ผู้สอบได้รับ ดังนั้น ลำดับของข้อสอบจึงมีผลต่อการ ตอบ ข้อสอบข้อถัดไป นอกจากนั้นจะเห็นว่าลำดับข้อสอบแบบผสมผสานมิติผู้สอบอาจจะได้ทำข้อสอบไม่ครอบคลุมในทุกมิติหรือทุกเนื้อหา ซึ่งอาจจะส่งผลต่อความแม่นยำในการประมาณค่าความสามารถในแต่ละมิติ แต่ในกรณีของลำดับข้อสอบแบบจำแนกมิติ ผู้วิจัยสาม ารถมั่นใจได้ว่าผู้สอบจะได้ทำข้อสอบครอบคลุมทุกมิติอย่างสมดุลดังนั้นผู้วิจัยจึงตั้งสมมติฐานข้อที่2ดังนี้

สมมติฐานข้อ 2 ลำดับข้อสอบแบบจำแนกมิติน่าจะมีคุณภาพการทดสอบที่สูงกว่า ลำดับข้อสอบแบบผสมผสานมิติ

3. จากการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องไม่พบการศึกษาปฏิสัมพันธ์ระหว่างตัวแปรวิธีการคัดเลือกข้อสอบขั้นแรกและลำดับข้อสอบ อย่างไรก็ตามผู้วิจัยมีความคิดเห็นว่าตัวแปรที่ศึกษาทั้งสองตัวแปรมีความเป็นอิสระจากกันและสามารถส่งผลต่อคุณภาพของการทดสอบปรับเหมาะได้อย่างอิสระ โดยไม่จำเป็นต้องทำหน้าที่ร่วมกัน ดังนั้นผู้วิจัยจึงตั้งสมมติฐานข้อที่ 3 ดังนี้

สมมติฐานข้อ 3 วิธีการคัดเลือกข้อสอบขั้นแรกและ ลำดับข้อสอบ ไม่น่าจะเกิดปฏิสัมพันธ์ที่ส่งผลต่อคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์

4. พงษ์พิชิต พรมสิทธิ์ (2544) ได้ศึกษาการดำเนินการสอบแบบทดสอบเทเลอร์ตามยุทธวิธีของเบส์ในวิชาคณิตศาสตร์ ค 014 โดยเว็บเพจ โดยได้สอบถามความคิดเห็นของผู้สอบด้วยการทดสอบแบบเทเลอร์ (การทดสอบปรับ เหมาะ) โดยนักเรียนเห็นด้วยอย่างยิ่งว่าการดำเนินการสอบด้วยแบบทดสอบเทเลอร์ ตามยุทธวิธีของเบย์ โดยเว็บเพจ มีการรายงานผลการสอบ ที่รวดเร็วทันใจ การป้อนข้อมูลผู้ทำการทดสอบทำได้ง่าย และการตอบข้อสอบง่ายและสะดวก นอกจากนั้นวิธีดังกล่าวยังทำให้มีความกระตือรือร้นและความน่าสนใจในการสอบมากกว่าการสอบด้วยการตอบในกระดาษ และไม่รู้สึกกลัวหรือกังวลใจที่ต้องดำเนินการสอบเช่นนี้ ดังนั้นผู้วิจัยจึงตั้งสมมติฐานข้อที่ 4 ดังนี้

สมมติฐานข้อ 4 ผลการประเมินโปรแกรมการทดสอบความสามารถด้านวิทยาศาสตร์ด้วยวิธีการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์ของนักเรียนกลุ่มตัวอย่างน่าจะมีความพึงพอใจอยู่ในระดับมากขึ้นไป

ขอบเขตของการวิจัย

การวิจัยครั้งนี้ใช้ข้อมูลจาก 3 แหล่ง คือ การจำลองข้อมูล ข้อมูลทุติยภูมิ และข้อมูลปฐมภูมิ โดยมีรายละเอียดดังนี้

1. การจำลองข้อมูล พัฒนาโดยใช้โปรแกรม R โดยกำหนดเป็นข้อสอบที่วัดความสามารถ 2 มิติ ซึ่งมี ลักษณะเป็นข้อสอบ พหุมิติระหว่างข้อ (between-item multidimensionality) นั่นคือ ข้อสอบหนึ่งข้อสอบวัดความสามารถ 1 มิติ แต่มิติทั้งสองมีความสัมพันธ์กัน โดยกำหนดค่าพารามิเตอร์ อำนาจจำแนกทั้ง 2 มิติ ให้มีลักษณะการแจกแจงเป็นโค้งปกติตัดปลาย โดยมีค่าเฉลี่ยเป็น 0.5 และค่าส่วนเบี่ยงเบนมาตรฐานเป็น 0.1 เนื่องจากหากอำนาจจำแนกมีค่าติดลบจะทำให้การประมาณค่าความสามารถไม่มีความแม่นยำ พารามิเตอร์ความยากอยู่ในช่วง -3 ถึง 3 มีการแจกแจงแบบเอกรูป และพารามิเตอร์โอกาสการเดามีค่าสูงสุด ไม่เกิน 0.1 ที่มีการแจกแจงแบบเอกรูป

2. ข้อมูลทุติยภูมิ ได้จากการสอบความถนัดทางวิชาชีพและวิชาการด้านวิทยาศาสตร์จากการสอบครั้งที่ 1/2552 ซึ่งมีเนื้อหาและศักยภาพที่วัดดังตาราง ที่ 1.1 อย่างไรก็ตามสำหรับข้อมูลที่นำมาศึกษาใช้เพียงเนื้อหาด้าน ชีววิทยา เคมี และฟิสิกส์ เท่านั้น

ตารางที่ 1.1 เนื้อหาและศักยภาพที่วัดในการทดสอบความถนัดทางวิชาชีพและวิชาการด้านวิทยาศาสตร์

สิ่งที่วัด	
เนื้อหา	**ศักยภาพ**
1) ชีววิทยา สิ่งแวดล้อม	1) คิดแบบนักวิทยาศาสตร์
2) เคมี	2) แก้ปัญหาแบบนักวิทยาศาสตร์
3) ฟิสิกส์	3) ทักษะการอ่านบทความแบบนักวิทยาศาสตร์
4) โลก ดาราศาสตร์ อวกาศ	

3. สำหรับข้อมูลปฐมภูมิ ผู้วิจัย ได้สร้างข้อสอบขึ้นใหม่ เพื่อใช้ทดสอบกับนักเรียนชั้นมัธยมศึกษาปีที่ 5 โดยพัฒนาข้อสอบวิทยาศาสตร์ จำแนกออกเป็น 3 ด้าน คือ ชีววิทยา เคมี และฟิสิกส์ โดยมีผู้พัฒนาข้อสอบคือ อาจารย์ที่สอนวิชาวิทยาศาสตร์ในระดับมัธยมศึกษาตอนปลาย ซึ่งการทดสอบจริงนี้เป็นการทดลองใช้โปรแกรมกับกลุ่มตัวอย่างขนาดเล็กเท่านั้น

4. ตัวแปรที่ใช้ในการศึกษาครั้งนี้ ประกอบด้วย

4.1 ตัวแปรอิสระ ได้แก่

4.1.1 วิธีการคัดเลือกข้อสอบขั้นแรก

1) Fisher information

2) Kullback-Leibler information

4.1.2 ลำดับข้อสอบ

1) ลำดับข้อสอบแบบจำแนกมิติ

2) ลำดับข้อสอบแบบผสมผสานมิติ

4.1.3 ภูมิหลังของผู้สอบ

1) เกรดเฉลี่ยสะสม (GPAX)

2) ทักษะการใช้คอมพิวเตอร์

4.2 ตัวแปรตาม ได้แก่

1) คุณภาพของการทดสอบ ปรับเหมาะแบบพหุมิติ วัดได้จาก ความเที่ยง ระยะทางแบบยุคลิด สารสนเทศแบบสอบ และความยาวของข้อสอบ

2) ผลประเมินการใช้โปรแกรมการทดสอบปรับเหมาะแบบพหุมิติ พิจารณาจาก ความยากในการใช้ ความเป็นไปได้ ในการนำไปใช้ และความพึงพอใจที่มีต่อการใช้

5. การยุติการทดสอบแบบปรับเหมาะ พิจารณาจาก 2 เกณฑ์ คือ การกำหนดจำนวนข้อสอบ (fixed test length) จำนวน 30 ข้อ กรณีการจำลองข้อมูลและการใช้ข้อมูลทุติยภูมิ พร้อมทั้งพิจารณาจากค่าความคลาดเคลื่อนมาตรฐานน้อยกว่าหรือเท่ากับ 0.3 (SEE ≤ 0.3) กรณีการใช้ข้อมูลปฐมภูมิ

ข้อตกลงเบื้องต้น

1. สำหรับการเปรียบเทียบวิธีการคัดเลือกข้อสอบขั้นแรกนั้น พิจารณาจากค่าความเที่ยงและค่าระยะทางแบบยุคลิด เท่านั้น เนื่องจากวิธีการคัดเลือกข้อสอบในแต่ละวิธีมีสเกลการวัดที่ต่างกัน จึงไม่สามารถนำสารสนเทศแบบสอบ หรือจำนวนข้อสอบมาเปรียบเทียบกันได้ส่วนในกรณีของการ สอบจริงซึ่งเลือกใช้วิธีการคัดเลือก ข้อสอบเพียงแบบ เดียว คือ สารสนเทศของฟิชเชอร์ (Fisher information) ผู้วิจัยพิจารณาจากค่าสารสนเทศแบบสอบ และความยาวข้อสอบ

2. การศึกษาจากข้อมูลทุ ติยภูมิ ใช้ข้อมูลจากการทดสอบความถนัดทางวิชาชีพและวิชาการด้านวิทยาศาสตร์ที่พัฒนาข้อสอบโดยสถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน) ซึ่งถือว่าเป็นข้อสอบที่ มีคุณภาพได้มาตรฐาน โดยผู้วิจัย ใช้ผลการทดสอบของนักเรียนมาคำนวณหา ค่าพารามิเตอร์ของข้อสอบแต่ละข้อ โดยการวิจัยครั้งนี้ ใช้ข้อมูลจากการสอบเนื้อหาชีววิทยา เคมี และฟิสิกส์ เท่านั้น เนื่องจากเนื้อหาทั้ง3 วิชา มีระดับความสำคัญไม่แตกต่างกัน และการทดสอบครั้งนี้ไม่ได้ตัดข้อสอบข้อใดออกจากคลังข้อสอบ แม้ผลการวิเคราะห์ค่าพารามิเตอร์จะมีค่าเท่าใดก็ตาม เพื่อคงสภาพข้อสอบให้คล้ายคลึงกับการทดสอบเดิมมากที่สุด

3. สำหรับการทดสอบจริง (ข้อมูลปฐมภูมิ) ข้อสอบที่ใช้ในการทดสอบพัฒนาโดยผู้ที่มีความรู้ความสามารถในวิชาวิทยาศาสตร์ นั่นคือ อาจารย์ที่สอนวิชาวิทยาศ าสตร์ในโรงเรียนมัธยมศึกษา โดยพัฒนาข้อสอบวิชา ชีววิทยา เคมี และฟิสิกส์ เท่านั้น

4. ข้อสอบที่นำมาศึกษาครั้งนี้มีรูปแบบความเป็นพหุมิติระหว่างข้อ (between-item multidimensionality) นั่นคือ ข้อสอบแต่ละข้อวัดเพียง 1 มิติ โดยแต่ละมิติต้องมีความสัมพันธ์กัน เนื่องจากการศึกษาลำดับข้อสอบนั้น ต้องจัดชุดข้อสอบเพื่อใช้ในการทดสอบ ดังนั้นข้อสอบแต่ละข้อจึงต้องมีความชัดเจนว่าวัดในมิติใด

5. การวิจัยครั้งนี้ใช้การกำหนดจำนวนข้อสอบ (Fixed test length) เนื่องจาก การวิจัยนี้เน้นการเปรียบเทียบวิธีวิทยาของการทดสอบปรับเหมาะแบบพหุมิติ ซึ่งงานวิจัยที่ผ่านมาส่วนใหญ่ใช้การกำหนดจำนวนข้อสอบทั้งสิ้น ดังนั้นผู้วิจัยจึงกำหนดให้ใช้จำนวนข้อสอบเป็นเกณฑ์ในการยุติการทดสอบซึ่งวิธีการดังกล่าวไม่ส่งผลต่อความแม่นยำในการประมาณค่าความสามารถอย่างไรก็ตาม การทดสอบปรับเหมาะแบบพหุมิติที่จะนำไปประยุกต์ใช้ในสถานการณ์จริง ควรกำหนดค่าความ

คลาดเคลื่อนมาตรฐานในการประมาณค่าเป็นเกณฑ์ในการยุติการทดสอบ เนื่องจากผู้วิจัยไม่ทราบค่าความสามารถที่แท้จริงของผู้สอบ ดังนั้น การจะพิจารณาว่าการทดสอบนั้นมีความเหมาะสมและค่าความสามารถที่ประมาณค่าได้มีความน่าเชื่อถือ จึงจำเป็นต้องพิจารณาค่าความคลาดเคลื่อนมาประกอบ ซึ่งในส่วนของการใช้ข้อมูลปฐมภูมิในงานวิจัยนี้ ได้นำค่าความคลาดเคลื่อนมาตรฐานในการประมาณค่ามาร่วมพิจารณาคุณภาพของการทดสอบด้วย

เหตุผลอีกประการหนึ่งที่งานวิจัยนี้ใช้การกำหนดจำนวนข้อสอบ เนื่องจาก การศึกษารูปแบบของลำดับข้อสอบกรณีลำดับข้อสอบแบบจำแนกมิติ ยังไม่มีเกณฑ์ที่เหมาะสมว่าจะเปลี่ยนมิติเมื่อใด ดังนั้นผู้วิจัยจึงใช้จำนวนข้อสอบเป็นตัวกำหนด และเหตุผลสำคัญที่งานวิจัยนี้ต้องใช้การกำหนดจำนวนข้อสอบ สืบเนื่องจากวิธีการคัดเลือกข้อสอบทั้ง 2 วิธี ที่มีวิธีการคิดและสเกลการวัดที่แตกต่างกัน กล่าวคือ สารสนเทศที่ได้จากวิธี Fisher information ไม่สามารถนำมาเปรียบเทียบกับสารสนเทศที่ได้จากวิธี Kullback-Leibler information ดังนั้นเมื่อแปลงค่าสารสนเทศแบบสอบมาเป็นค่าความคลาดเคลื่อนมาตรฐานในการประมาณค่า จึงไม่สามารถนำมาเปรียบเทียบกันได้ ซึ่งเป็นเหตุผลว่า ทำไมงานวิจัยครั้งนี้จึงใช้การกำหนดจำนวนข้อสอบแทนการพิจารณาค่าความคลาดเคลื่อนมาตรฐานในการประมาณค่า

ข้อจำกัดในการวิจัย

1. การใช้ข้อมูลปฐมภูมิ (การทดสอบจริง) เป็นการทดลองใช้โปรแกรมกับกลุ่มตัวอย่างขนาดเล็กเท่านั้น ดังนั้น ผลการวิจัยที่ได้จึงไม่สามารถสรุปอ้างอิงไปยังกลุ่มประชากรได้

2. การทดสอบจริง ใช้วิธีคัดเลือกข้อสอบด้วยวิธี Fisher information เท่านั้น เนื่องจากวิธี Kullback-Leibler information มีการคำนวณที่ซับซ้อนซึ่งเป็นไปได้ยากในทางปฏิบัติ ที่ต้องคัดเลือกข้อสอบทันทีหลังจากที่ผู้สอบตอบข้อสอบข้อก่อนหน้า ดังนั้น จึงไม่มีการเปรียบเทียบวิธีการคัดเลือกข้อสอบขั้นแรกในกรณีการทดสอบจริง

3. งานวิจัยนี้มีข้อจำ กัดที่จำนวนมิติที่ศึกษาจากแหล่งข้อมูลทั้ง 3 แหล่ง เนื่องจากผู้วิจัยศึกษาจากการจำลองข้อมูลเพียง 2 มิติ เพื่อตรวจสอบความถูกต้องของกระบวนการทดสอบปรับเหมาะแบบพหุมิติ ก่อนนำไปประยุกต์ใช้กับข้อมูลทุติยภูมิและข้อมูลปฐมภูมิที่ศึกษา 3 มิติในวิชาชีววิทยา เคมี และฟิสิกส์ ซึ่งผลวิจัยที่ได้เป็นผลที่เกิดจากการศึกษา 2 มิติ และ 3 มิติ

4. ในกรณีการทดสอบปรับเหมาะแบบพหุมิติที่มีลำดับข้อสอบแบบผสมผสานมิติ ผู้สอบบางคนอาจได้รับข้อสอบไม่สมดุลกันในแต่ละมิติ หรือไม่สอดคล้องกับโครงสร้างข้อสอบ เนื่องจากรูปแบบของการคัดเลือกข้อสอบที่มีลำดับข้อสอบแบบผสมผสานมิติ จะไม่คำนึงถึงเนื้อหาที่มุ่งวัด

คำจำกัดความที่ใช้ในการวิจัย

การทดสอบแบบปรับเหมาะ หมายถึง การทดสอบที่ผู้สอบแต่ละคนใช้แบบสอบต่างชุดกัน โดยปรับเปลี่ยนไปตามความสามารถของผู้สอบซึ่งมีการคัดเลือกข้ อสอบที่มีความยากเหมาะสมกับระดับความสามารถของผู้สอบ

โมเดลการตอบสนองข้อสอบแบบพหุมิติ หมายถึง การขยายโมเดลให้สามารถรองรับความสามารถ (θ) หลายองค์ประกอบ โดยจะถือว่าคุณลักษณะแฝงของบุคคลมากกว่า 1 องค์ประกอบ ส่งผลต่อการตอบข้อสอบ ดังนั้นพารามิเต อร์ความสามารถของผู้สอบจึงมีตั้งแต่ 2 พารามิเตอร์ขึ้นไป ซึ่งโมเดลการตอบสนองข้อสอบแบบพหุมิตินี้ เป็นการผ่อนคลายข้อตกลงเบื้องต้นของทฤษฎีการตอบสนองข้อสอบเกี่ยวกับความเป็นเอกมิติ

การทดสอบปรับเหมาะแบบพหุมิติ หมายถึง การผสมผสานแนวคิดของโมเดลการตอบสนองข้อสอบแบบพหุมิติ (multidimensional item response theory: MIRT) กับการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (computerized adaptive testing: CAT) ซึ่งเป็นการทดสอบที่ลดความยาวของข้อสอบ และเพิ่มความแม่นยำในการประมาณค่าความสามารถของผู้สอบ ซึ่งหลักการคล้ายกับการทดสอบปรับเหมาะแบบเอกมิติ โดยมีความแตกต่างเพียงค่าพารามิเตอร์อำนาจจำแนกจะมีหลายค่าตามมิติที่มุ่งวัด ดังนั้นการคำนวณค่าต่าง ๆ จะใช้ข้อมูลในลักษณะของเวคเตอร์

วิธีการคัดเลือกข้อสอบขั้นแรก หมายถึง วิธีการคัดเลือกข้อสอบ ซึ่งส่งผลต่อความแม่นยำในการประมาณค่าความสามารถของผู้สอบ โดยเฉพาะอย่างยิ่งเมื่อจำนวนข้อสอบที่ใช้ในการ ทดสอบมีน้อย (ระยะแรกของการทดสอบแบบปรั บเหมาะ) ซึ่งในการวิจัยครั้งนี้ พิจารณาที่การทำข้อสอบจำนวน 5 ข้อ โดยศึกษา วิธีการคัดเลือกข้อสอบขั้นแรก 2 วิธี คือ Fisher information และ Kullback-Leibler information

ลำดับข้อสอบ หมายถึง รูปแบบของข้อสอบที่ถูกคัดเลือกให้กับผู้สอบ ซึ่งแบ่งเป็น 2 รูปแบบ คือ ลำดับข้อสอบแบบจำแนกมิติ และลำดับข้อสอบแบบผสมผสานมิติ

ลำดับข้อสอบ แบบจำแนกมิติ หมายถึง การทดสอบ ที่ผู้สอบได้รับข้อสอบทีละ ชุดตามมิติหรือเนื้อหา ในจำนวนที่เท่ากัน โดยใช้จำนวนข้อเป็นตัวกำหนดการเปลี่ยนมิติในการทดสอบ ซึ่งการวิจัยครั้งนี้ ในกรณีการจำลองข้อมูล2 มิติ กำหนดให้ทำข้อสอบมิติแรก15 ข้อ และทำข้อสอบมิติที่สอง 15 ข้อ กรณีการทดสอบจริง กำหนดให้ ผู้สอบได้รับข้อสอบชุดแรกในมิติชีววิทยา 10 ข้อ ตามด้วย มิติเคมี 10 ข้อ และมิติฟิสิกส์ 10 ข้อ

ลำดับข้อสอบ แบบผสมผสานมิติ หมายถึง การทดสอบที่ผู้สอบ ได้รับข้อสอบสลับกันไปในทุกมิติ หรือทุกเนื้อหา เช่น ผู้สอบ ได้รับข้อสอบข้อแรกในมิติเคมี ตามด้วยข้อสอบในมิติชีววิทยา และมิติอื่นๆ สลับกันไป ขึ้นกับค่าสารสนเทศข้อสอบที่มีค่าสูงสุดในแต่ละรอบการคัดเลือก โดยไม่คำนึงว่าข้อสอบข้อนั้นจะวัดในมิติใด ซึ่งทำให้ผู้สอบได้รับข้อสอบจำนวนไม่เท่ากันในแต่ละมิติ

คุณภาพของการทดสอบแบบปรับเหมาะ หมายถึง ความสามารถของการวัดที่ไ ด้จากการทดสอบแบบปรับเหมาะ ซึ่งพิจารณาจากความเที่ยง ระยะทางแบบยุคลิด สารสนเทศข้อสอบ และความยาวของข้อสอบ ซึ่งมีรายละเอียดของวิธีการวัด ดังนี้

ความเที่ยง หมายถึง กำลังสองของความสัมพันธ์ระหว่างความสามารถที่แท้จริงกับค่าความสามารถที่ประมาณค่าได้ (Segall, 1996; Wang & Chen, 2004) สำหรับการจำลองข้อมูลพิจารณาจากความสัมพันธ์ของค่าความสามารถที่ประมาณค่าได้จากการทดสอบปรับเหมาะแบบพหุมิติ กับค่าความสามารถที่แท้จริง กรณีของการใช้ข้อมูล ทุติยภูมิ พิจารณาจากความสัมพันธ์ของค่าความสามารถที่ประมาณค่าได้จากการทดสอบปรับเหมาะแบบพหุมิติ กับค่าความสามารถที่ได้จากการทำข้อสอบทุกข้อ ทั้งนี้ ค่าความ เที่ยง ใช้พิจารณา สำหรับ การจำลองข้อมูลและ ข้อมูลทุติยภูมิเท่านั้น

ระยะทางแบบยุคลิด หมายถึง การหาระยะห่างระหว่างค่าความสามารถของผู้สอบที่ได้จากการประมาณค่าจากการทดสอบแบบปรับเหมาะ ($\hat{\theta}$) กับค่าความสามารถของผู้สอบที่แท้จริง (θ) โดยที่ ค่าระยะทางน้อย หมายถึง ค่าความสามารถจากการประมาณค่าและค่าความสามารถที่แท้จริงมีความใกล้เคียงกัน นั่นคือ วิธีการทดสอบปรับเหมาะแบบพหุมิติมีคุณภาพสูง ทั้งนี้ ค่าระยะทางแบบยุคลิดใช้พิจารณาสำหรับการจำลองข้อมูลและข้อมูลทุติยภูมิเท่านั้น

สารสนเทศขอ งแบบสอบ หมายถึง การประเมินคุณภาพของแบบสอบ โดยใช้ค่าฟังก์ชันสารสนเทศของแบบสอบ (test information: $I(\theta)$) ซึ่งเกิดจากผลรวมเชิงพีชคณิตของค่าฟังก์ชันสารสนเทศของข้อสอบแต่ละข้อรวมเข้าด้วยกันทั้งฉบับ ณ ตำแหน่ง θ เดียวกัน สำหรับค่าฟังก์ชันสารสนเทศขอ งแบบสอบในการทดสอบแบบปรับเหมาะ ใช้ค่าเฉลี่ยของ $I(\theta)$ ณ ตำแหน่งที่ผู้สอบได้รับข้อสอบจำนวนเท่ากัน เพื่อตรวจสอบคุณภาพที่มีต่อผู้สอบทุกความสามารถ ซึ่งในงานวิจัยนี้ ใช้ใน กรณีการคัดเลือกข้อสอบโดยวิธีของฟิชเ ชอร์ จากการเก็บข้อมูลจริงเท่านั้น

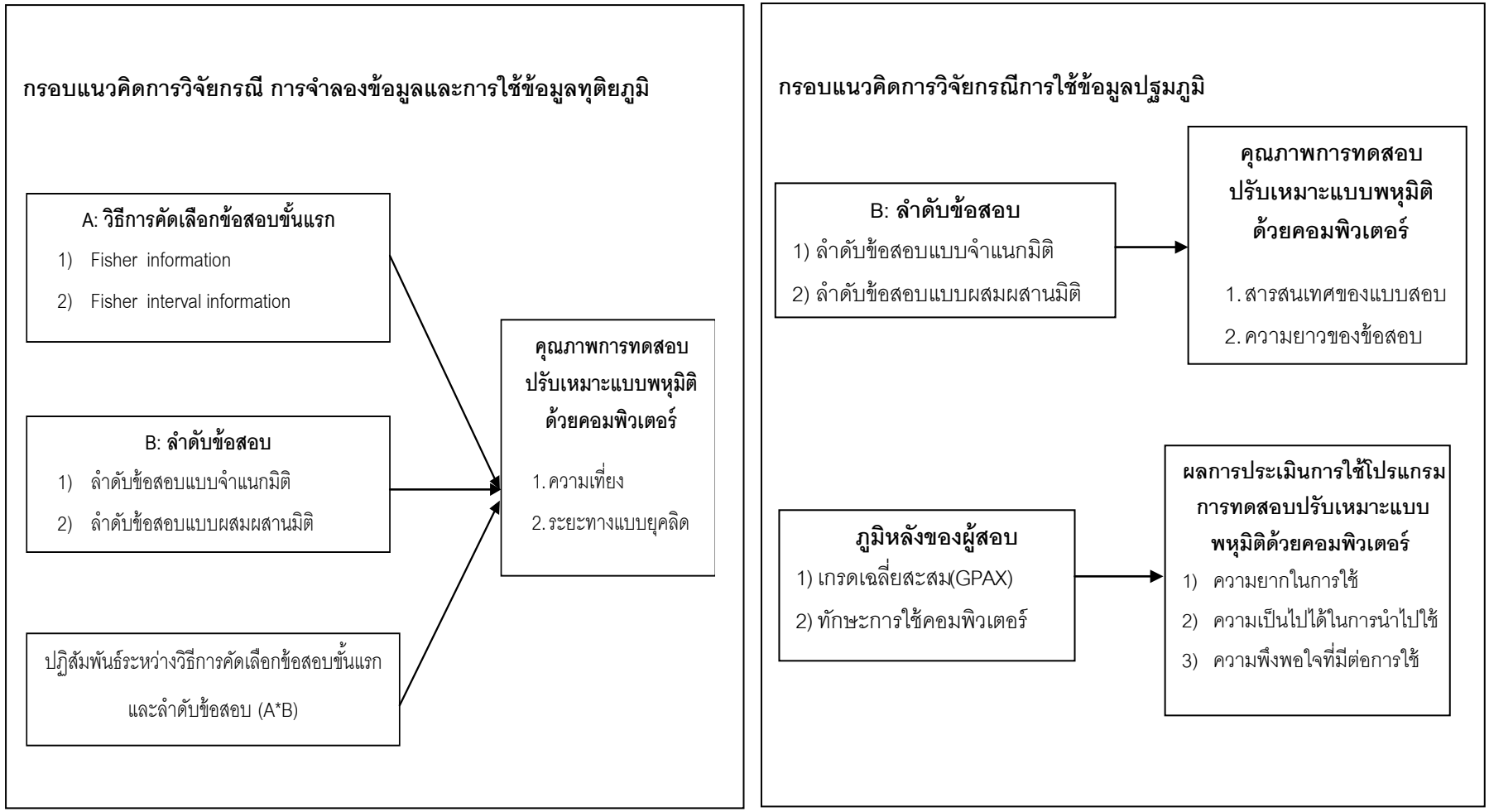
ความยาวของข้อสอบ หมายถึง จำนวนข้อสอบที่ผู้สอบแต่ละคนได้รับ เนื่องจากการทดสอบแบบปรับเหมาะ ผู้สอบแต่ละคน ได้ทำข้อสอบจำนวนไม่เท่ากัน ขึ้นกับเงื่อนไขของการยุติการ ทดสอบ ซึ่งสำหรับการวิจัยครั้งนี้ ยุติการทดสอบเมื่อความคลาดเคลื่อนมาตรฐานน้อย กว่าหรือเท่ากับ 0.3 สำหรับการพิจารณา คุณภาพของการทดสอบแบบปรับเหมาะ กล่าวได้ว่าการที่ความยาวของข้อสอบลดลง หมายถึง การทดสอบนั้นสามารถประมาณค่าความ สามารถผู้สอบได้อย่างแม่นยำ แม้ทำข้อสอบเพียงไม่กี่ข้อ ซึ่งในงานวิจัยนี้ ใช้ใน กรณีการคัดเลือกข้อสอบโดยวิธีของฟิชเชอร์ จากการเก็บข้อมูลจริงเท่านั้น

แบบทดสอบความถนัดทางวิชาชีพและวิชาการด้านวิทยาศาสตร์ หมายถึง แบบทดสอบความถนัดทางวิชาชีพและวิชาการ (professional and academic aptitude test: PAT) ด้านวิทยาศาสตร์ (PAT2) ที่พัฒนาขึ้นโด ยสถาบันทดสอบทางการศึกษาแห่งชาติ(องค์การมหาชน) ในการทดสอบครั้งที่ 1/2552 ซึ่งใช้เฉพาะผลการตอบข้อสอบของนักเรียน เท่านั้นและในกรณีของการทดสอบจริง ใช้แบบทดสอบ ความสามารถ ด้านวิทยาศาสตร์ ที่พัฒนาโดยอาจารย์ที่สอนวิชาวิทยาศาสตร์ระดับมัธยมศึกษาตอนปลายในโรงเรียนมัธยมศึกษา

ภูมิหลังของผู้สอบ หมายถึง ข้อมูลเบื้ องต้นของผู้สอบ ซึ่งงานวิจัยนี้ ศึกษาข้อมูลภูมิหลังของผู้สอบ คือ ผลสัมฤทธิ์ทางการเรียนของผู้สอบ (GPAX) และทักษะการใช้คอมพิวเตอร์ โดยใช้แบบวัดความสามารถคอมพิวเตอร์ในการใช้โปรแกรมประเมินผลคำ (micrsoft word) และโปรแกรมนำเสนอผลงาน (microsoft PowerPoint) ที่พัฒนาโดย ชุติวัฒน์ สุวัตถิพงศ์ (2551)

ผลประเมินการใช้โปรแกรมการทดสอบปรับเหมาะแบบพหุมิติ หมายถึง ผลประเมินที่ได้จากการประเมินของนักเรียนชั้นมัธยมศึกษาปีที่ 5 ที่ดำเนินการทดสอบความ สามารถ ด้านวิทยาศาสตร์ผ่านโปรแกรมการทดสอบปรั บเหมาะแบบพหุมิติ ซึ่งแบบประเมินสร้างขึ้นโดยผู้วิจัย โดยมีประเด็นการประเมินคือ ความยากในการใช้ ความเป็นไปได้ในการนำไปใช้และความพึงพอใจที่มีต่อการใช้

กรอบแนวคิดในการวิจัย



แผนภาพที่ 1.1 กรอบแนวคิดในการวิจัย

ประโยชน์ที่คาดว่าจะได้รับ

1. ได้วิธีการวัดความ ความสามารถ ด้านวิทยาศาสตร์และโปรแกรมคอมพิวเตอร์สำหรับการทดสอบปรับเหมาะแบบพหุมิติที่สามารถนำไปใช้ในการวัดความ สามารถผู้สอบ ได้อย่างมีประสิทธิภาพ

2. ได้สารสนเทศเกี่ยวกับวิธีการคัดเลือกข้อสอบขั้นแรกที่มีคุณภาพสำหรับการทดสอบปรับเหมาะแบบพหุมิติ เพื่อเลือกใช้วิธีที่มีความเหมาะสม และมีความแม่นยำในการประมาณค่าความสามารถของผู้สอบ

3. ได้สารสนเทศเกี่ยวกับลำดับข้อสอบที่มีผลต่อคุณภาพของการทดสอบแบบปรับเหมาะโดยเฉพาะการประมาณค่าความสามารถของผู้สอบ

4. ได้สารสนเทศเกี่ยวกับภูมิหลังของผู้สอบที่ส่งผลต่อผลการประเมินการทดสอบความสามารถด้านวิทยาศาสตร์ด้วยวิธีการทดสอบปรับเหมาะแบบพหุมิติ

5. ได้แนวทางในการปรับปรุง แก้ไข การทดสอบแบบปรับเหมาะให้มีความเหมาะสมสอดคล้องกับความต้องการของผู้สอบ และสามารถนำมาใช้กับการทดสอบที่มีสเกลขนาดใหญ่และมีผลกระทบสูง (high-stake testing) ได้อย่างมีประสิทธิภาพ เช่น การประเมินคุณภาพการศึกษาระดับชาติ (national test: NT) การทดสอบทางการศึกษาแห่งชาติขั้นพื้นฐาน (ordinary national educational Test: O-NET) และการทดสอบความถนัดทางวิชาชีพและวิชาการ (professional and academic aptitude test: PAT)

6. ได้แนวทางในการนำทฤษฎีการตอบสนองข้อสอบแบบพหุมิติและการทดสอบแบบปรับเหมาะมาใช้ในการทดสอบจริง ซึ่งเป็นการขยายความรู้ในเชิงทฤษฎีสู่การปฏิบัติ เพื่อให้ผู้ที่สนใจสามารถนำไปประยุกต์ใช้ในการทดสอบอื่นๆ ต่อไป

บทที่ 2

เอกสารและงานวิจัยที่เกี่ยวข้อง

ในการวิจัยครั้งนี้ผู้วิจัยแบ่งการนำเสนอเอกสารงานวิจัยที่เกี่ยวข้องออกเป็น 4 ตอน ซึ่งมีรายละเอียดดังนี้ ตอนที่ 1 มโนทัศน์เกี่ยวกับการทดสอบแบบปรับเหมาะ ตอนที่ 2 มโนทัศน์เกี่ยวกับโมเดลการตอบสนองข้อสอบแบบพ หุมิติ ตอนที่ 3 มโนทัศน์เกี่ยวกับการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบพหุมิติ และ ตอนที่ 4 งานวิจัยที่เกี่ยวข้อง ซึ่งมีรายละเอียดในแต่ละตอนดังนี้

ตอนที่ 1 มโนทัศน์เกี่ยวกับการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์

จากการศึกษาเอกสารและงานวิจัยที่ เกี่ยวข้องเกี่ยวกับการทดสอบแบบปรับเหมาะ ด้วยคอมพิวเตอร์ ซึ่งได้แบ่งหัวข้อการนำเสนอไว้ดังนี้ ความหมายของการทดสอบแบบปรับเหมาะ ประวัติของการทดสอบแบบปรับเหมาะ หลักการของการทดสอบแบบปรับเหมาะ ประเภทของการทดสอบแบบปรับเหมาะ องค์ประกอบของการทดสอบแบบปรับเหมาะ ประโยชน์และข้อจำกัดของการทดสอบแบบปรับเหมาะ เกณฑ์ที่ใช้ในการประเมินการทดสอบแบบปรับเหมาะ และประเด็นปัญหาของการใช้การทดสอบแบบปรับเหมาะ ซึ่งมีรายละเอียดแต่ละหัวข้อดังนี้

1.1 ความหมายของการทดสอบแบบปรับเหมาะ

แบบสอบปรับเหมาะ (adaptive test) หมายถึง แบบสอบที่สามารถป รับเปลี่ยนความยากของข้อสอบแต่ละข้อให้เหมาะสมกับระดับความสามารถของผู้สอบแต่ละบุคคล (Lord, 1971)

แบบสอบปรับเหมาะกับความสามารถของผู้สอบ (adaptive test) หมายถึง แบบสอบที่ประกอบด้วยชุดของข้อสอบแตกต่างกันที่จัดให้สำหรับผู้สอบต่างกัน โดยใช้การคัดเลือกข้อสอบสำหรับผู้สอบแต่ละคนตามระดับความสามารถที่มุ่งวัดของบุคคลนั้น (Weiss, 1988)

การทดสอบ แบบปรับเหมาะกับความสามารถของผู้สอบ (adaptive testing) หมายถึง การทดสอบที่ใช้แบบสอบต่างชุดกันสำหรับผู้สอบต่างกัน โดยมีการคัดเลือกข้อสอบที่มีความยากเหมาะสมกับระดับความสามารถของผู้สอบ (ศิริชัย กาญจนวาสี, 2550)

โดยสรุป การทดสอบแบบปรับเหมาะ (adaptive testing) หมายถึง การทดสอบที่ผู้สอบแต่ละคนใช้แบบสอบต่างชุดกัน โดยปรับเปลี่ยนไปตามความสามารถของผู้สอบซึ่งมีการคัดเลือกข้อสอบที่มีความยากเหมาะสมกับระดับความสามารถของผู้สอบ

1.2 ประวัติของการทดสอบแบบปรับเหมาะ

แบบทดสอบแบบปรับเหมาะฉบับแรก ได้แก่ แบบทดสอบเชาว์ปัญญาของบิเนท์ (Binet intelligence test) พัฒนาขึ้นโดย Alfred Binet นักจิตวิทยาชาวฝรั่งเศสเมื่อต้นปี ค .ศ. 1909 แบบทดสอบเชาว์ปัญญาของบิเนท์ใช้การเลือกชุดของข้อสอบเบื้องต้นตามระดับอายุ (ความยาก) ข้อสอบจะถูกให้คะแนนและถูกเลือกให้ผู้สอบทำตามทางเลือกที่กำหนดไว้คงที่ (fixed branching) โดยผู้บริหารการสอบ เมื่อผู้สอบทำข้อสอบในระดับอายุจริงได้ไม่ถูกต้องก็จะถูกให้ทำข้อสอบในระดับอายุที่ต่ำกว่า แต่ถ้าทำได้ถูกต้องก็จะถูกให้ทำข้อสอบในระดับอายุที่สู งกว่า จนกระทั่งถึงระดับเพดาน (ceiling level) เมื่อผู้สอบทำข้อสอบทุกข้อไม่ถูกต้อง หรือระดับพื้นฐาน (basal level) เมื่อผู้สอบทำข้อสอบทุกข้อได้ถูกต้อง เกณฑ์ที่ใช้ในการยุติการสอบจะแตกต่างกันตามระดับความสามารถของแต่ละบุคคลเมื่อสามารถระบุระดับเพดานและระดับพื้น ฐานของผู้สอบได้ (Weiss, 1983; Weiss, 1988,1990) แม้ว่าแบบทดสอบเชาว์ปัญญาของบิเนท์จะถูกนำไปใช้อย่างแพร่หลายและเป็นต้นฉบับในการกำหนดหลักการของการทดสอบแบบปรับเหมาะในปัจจุบัน แต่แบบทดสอบเชาว์ปัญญาของบิเนท์ก็มีข้อจำกัดที่ผู้สอบอาจแสดงผลการตอบไม่คงเส้นคงวาระหว่างการทดสอบ ผู้สอบบางคนอาจทำข้อสอบบางข้อได้ดีแตกต่างกันเนื่องจากอิทธิพลของเชื้อชาติ เพศ วัฒนธรรม รวมทั้งยังมีค่าใช้จ่ายในการทดสอบสูง และใช้เวลาในการสอบมาก (ศิริชัย กาญจนวาสี , 2538, 2550) แบบทดสอบแบบปรับเหมาะอื่น ๆ เริ่มพัฒนาขึ้นปี ค.ศ. 1950 โดยเริ่มตั้งแต่การ สอบเขียนตอบ (paper-and-pencil) จนถึงการใช้คอมพิวเตอร์ช่วยในการทดสอบ (computer assisted testing) ซึ่งมีการพัฒนาวิธีเลือกข้อสอบเริ่มต้นและกำหนดเกณฑ์ในการยุติการสอบที่แตกต่างกัน แต่ยังคงใช้หลักการที่เน้นความเหมาะสมกับความสามารถของผู้สอบเช่นเดียวกับแบบทดสอบเชาว์ปัญญาของบิเนท์ (Weiss, 1988, 1990) ทั้งนี้ได้นำเสนอหลักการของการทดสอบแบบปรับเหมาะและเสนอความแตกต่างระหว่างการทดสอบแบบดั้งเดิมกับการทดสอบแบบปรับเหมาะ ไว้ในหัวข้อถัดไป

1.3 หลักการของการทดสอบแบบปรับเหมาะ

การทดสอบแบบปรับเหมาะ (adaptive testing or tailored testing) เป็นการทดสอบที่มุ่งวัดคุณลักษณะที่ต้องการ โดยการคัดเลือกข้อสอบตามความสามารถของผู้สอบในระหว่างกระบวนการทดสอบ การคัดเลือกข้อสอบจะพิจารณาจากผลการตอบข้อสอบหรือข้อคำถามที่มีมาก่อน ถ้าตอบข้อสอบข้อนั้นได้ถูกต้อง ข้อสอบข้อต่อไปก็จะยา กขึ้นแต่ถ้าตอบข้อสอบข้อนั้นผิดข้อสอบข้อต่อไปก็จะง่ายลง (Weiss,1983; Weiss and Schleisman, 1999) แบบทดสอบที่ใช้ในการทดสอบแบบปรับเหมาะจึงเป็นแบบทดสอบที่ประกอบไปด้วยชุดของข้อสอบที่แตกต่างกัน

สำหรับการสอบของแต่ละบุคคล ซึ่งเป็นลักษณะที่ตรงกันข้ามกับแบบทดสอบแบ บประเพณีนิยม (conventional tests) ที่ผู้สอบทุกคนจะถูกสอบด้วยแบบทดสอบชุดเดียวกันWeiss (1988,1990) ทั้งนี้ Jian-quan, and others (2007) และ Straetmans and Eggen (1998) ได้สรุปความแตกต่างระหว่างการทดสอบแบบประเพณีนิยมและการทดสอบแบบปรับเหมาะไว้ ดังตารางที่ 2.1

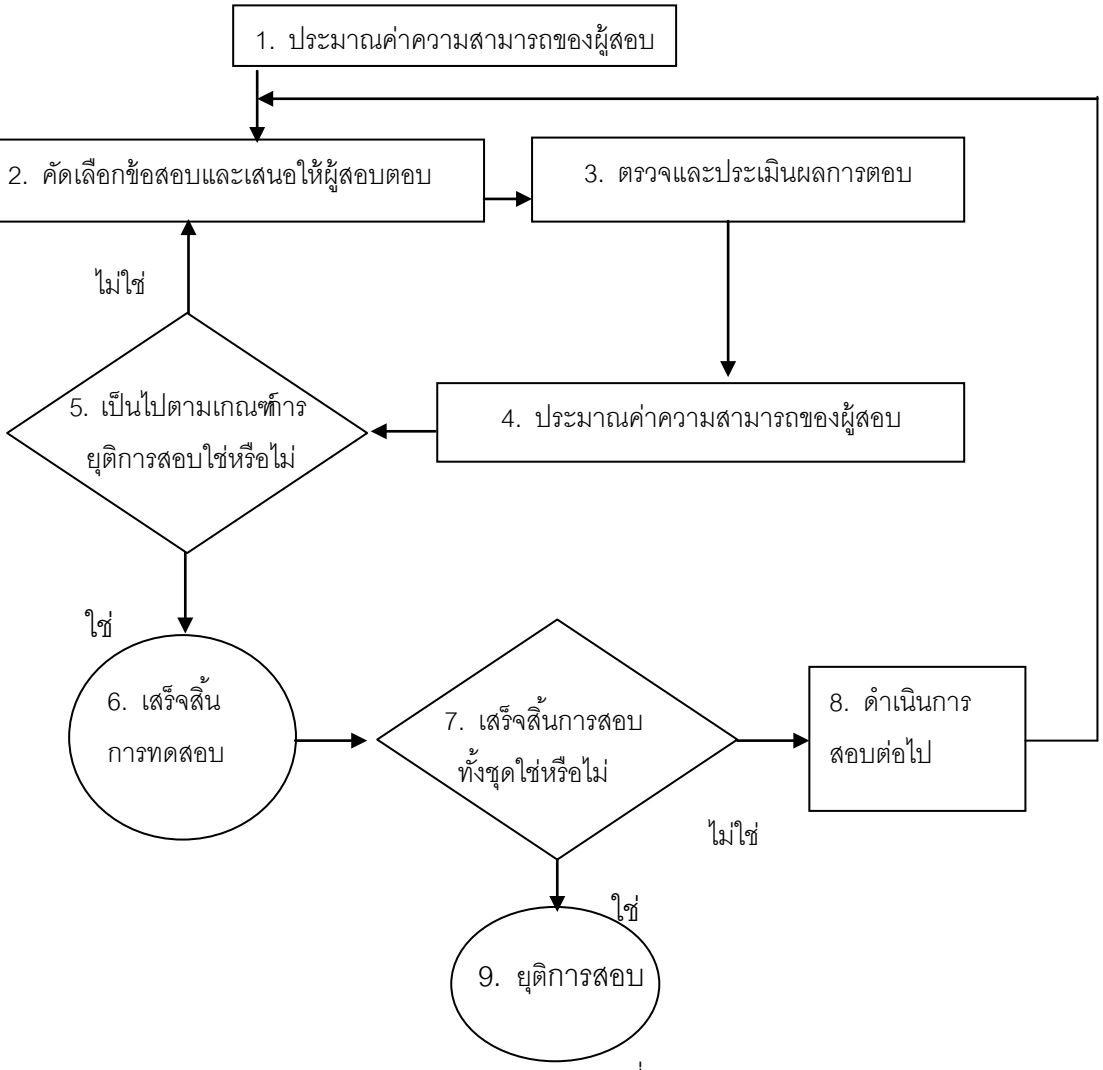
ตารางที่ 2.1 ความแตกต่างระหว่างการทดสอบแบบประเพณีนิยมและการทดสอบแบบปรับเหมาะ

รายการ	แบบทดสอบแบบประเพณีนิยม	แบบทดสอบแบบปรับเหมาะ
องค์ประกอบของแบบทดสอบ	ผู้สอบแต่ละคนใช้แบบทดสอบชุดเดียวกัน	ผู้สอบแต่ละคนใช้ข้อสอบที่แตกต่างกัน
ความยากของแบบทดสอบ	ใช้ความยากเฉลี่ยที่เหมาะกับผู้สอบ	ความยากจะขึ้นกับผู้สอบแต่ละคน
ความยาวของแบบทดสอบ	ผู้สอบใช้แบบทดสอบที่มีความยาวเท่ากันและมีความยาวมากกว่าแบบทดสอบแบบปรับเหมาะ	ผู้สอบใช้แบบทดสอบที่มีความยาวแตกต่างกันและ แบบทดสอบ จะสั้นกว่าแบบทดสอบแบบประเพณีนิยม
โอกาสในการสอบ	เจาะจงเวลาในการทดสอบให้กับผู้สอบทุกคน	ผู้สอบเลือกเวลาสอบได้ตามความเหมาะสม
การบริหารการสอบ	ใช้เวลาในการทดสอบมาก	ใช้เวลาในการทดสอบน้อย
การให้ผลคะแนน	ไม่สามารถให้ผลคะแนนได้ทันที	สามารถให้ผลคะแนนได้ทันที
ความปลอดภัย	มีปัญหาเรื่องความปลอดภัยเมื่อผู้เข้าสอบไม่ได้สอบในเวลาเดียวกัน	ป้องกันความปลอดภัยได้ดีกว่าเพราะผู้เข้าสอบจะใช้แบบทดสอบที่ต่างกัน
ความสมดุลของเนื้อหา	ไม่มีปัญหา	ไม่ได้อยู่ภายใต้การควบคุมโดยปราศจากการวัดผลเป็นพิเศษ

ที่มา : Jian-quan, and others (2007) ; Straetmans and Eggen (1998)

การดำเนินการทดสอบแบบปรับเหมาะจะเกี่ยวข้องกับคำถามสำคัญ 3 คำถาม ได้แก่ จะเลือกข้อสอบที่ใช้ในการเริ่มต้นการสอบได้อย่างไร (how to start) จะเลือกข้อสอบข้อถัดไปได้อย่างไร (how to continue) และรู้ได้อย่างไรว่าจะยุติการสอบ (how to stop) (Thissen and Mislevy, 1990; Wainer, 1990) คำถามดังกล่าวจะไม่พบในการทดสอบแบบประเพณีนิยม เนื่องจากในการเริ่มต้นทำแบบทดสอบมักจะเริ่มจากการทำข้อสอบข้อที่ 1 ข้อต่อ ๆ ไป จนถึงข้อสุดท้าย แต่ในการทดสอบแบบปรับเหมาะการเลือกข้อสอบที่ใช้ในการประมาณค่าความสามารถเบื้องต้นหรือเป็นขั้นตอนของการเริ่มก ารทดสอบเป็นที่สำคัญและมีผลต่อการประมาณค่าความสามารถของผู้สอบเพื่อการเลือกข้อสอบข้อต่อไป รวมถึงการยุติการสอบและการประมาณค่าความสามารถสุดท้ายของผู้สอบ การเริ่มต้นการสอบ การดำเนินการสอบและการยุติการสอบ

จึงเป็นสิ่งที่ต้องพิจารณาเพื่อให้ การทดสอบแบบปรับเหมาะมี ประสิทธิภาพสูงสุดและมักจะเป็นประเด็นที่ใช้ในการศึกษาวิจัยพัฒนาเกี่ยวกับการทดสอบแบบปรับเหมาะ โดยทั่วไปการทดสอบแบบปรับเหมาะมักดำเนินการโดยเริ่มจากการประมาณความสามารถของผู้สอบ คัดเลือกข้อสอบ ที่เหมาะสมกับความสามารถของผู้สอบและเสนอให้ผู้สอบทำ หลังจากนั้นจึงประม าณค่าความสามารถของผู้สอบอีกครั้งหนึ่ง ตรวจและประเมินคำตอบ พิจารณาว่าเป็นไปตามเกณฑ์ที่ใช้ในการยุติการสอบใช่หรือไม่ ถ้าไม่ใช่ก็ดำเนินการคัดเลือกข้อสอบข้อต่อไป ถ้าใช่ก็สิ้นสุดการสอบสำหรับแบบทดสอบฉบับนั้น และพิจารณาว่าเสร็จสิ้น การทดสอบทุกฉบับ ในแบบทดสอบชุดนั้นแล้วใช่หรือไม่ ถ้าไม่ใช่ก็ดำเนินการสอบแบบทดสอบฉบับต่อไป ถ้าใช่ก็ยุติการสอบ (Thissen and Mislevy, 1990) ดังแสดงในแผนภาพที่ 2.1



ที่มา: Thissen and Mislevy (1990)

แผนภาพที่ 2.1 ขั้นตอนการทดสอบแบบปรับเหมาะ

1.4 ประเภทของการทดสอบแบบปรับเหมาะ

การจำแนกประเภทของการทดสอบแบบปรับเหมาะพิจารณาได้จากเครื่องมือที่ใช้ในการทดสอบและยุทธวิธีที่ใช้ในการทดสอบ (ศิริชัย กาญจนวาสี, 2538, 2550) มีรายละเอียดดังนี้

1) การทดสอบแบบปรับเหมาะจำแนกตามเครื่องมือที่ใช้ ได้แก่ การทดสอบแบบปรับเหมาะที่ไม่ใช้คอมพิวเตอร์ ประกอบด้วย แบบทดสอบเชาวน์ปัญญาของบิเนท์ แบบทดสอบเลือกคำตอบเป็นกลุ่มแบบสองขั้นตอน (two-stage test)แบบทดสอบปรับระดับความสามารถที่ยืดหยุ่น (flexilevel Test) และแบบทดสอบกำหนดตามการแยกทาง (branching test) และ การทดสอบแบบปรับเหมาะโดยใช้คอมพิวเตอร์ซึ่งเป็นการนำคอมพิวเตอร์มาใช้เพื่อความสะดวกในการคัดเลือกข้อสอบและประมาณความสามารถของผู้สอบซึ่งนิยมใช้การกำหนดทางแยกแบบแปรผัน และใช้ทฤษฎีการตอบสนองข้อสอบเป็นพื้นฐานในการประมาณค่า

2) การทดสอบแบบปรับเหมาะจำแนกตามยุทธวิธีที่ใช้ในการทดสอบ ได้แก่ ยุทธวิธีสองขั้นตอน(two-stage strategies) และยุทธวิธีหลายขั้นตอน(multi-Strategies Strategies) ประกอบด้วยยุทธวิธี หลายขั้นตอนแบบทางแยกคงที่(fixed-Branching) และยุทธวิธีหลายขั้นตอนแบบทางแยกแปรผัน (variable branching) ซึ่งมีรายละเอียดของแต่ละยุทธวิธีดังนี้

2.1 <u>ยุทธวิธีสองขั้นตอน (two-stage strategies)</u>

ในปี ค.ศ.1970 และปีค.ศ.1980 นักวิจัยได้มี ความพยายามในการพัฒนาแบบทดสอบแบบปรับเหมาะโดยเน้นไปที่ขั้นตอนการคัดเลือกข้อสอบ เพื่อให้ได้จำนวนข้อสอบที่น้อยที่สุดในการประมาณค่าความสามารถของผู้ตอบได้อย่างถูกต้ อง ยุทธวิธีสองขั้นตอนจึงได้รับการพัฒนาขึ้นเพื่อตอบสนองวัตถุประสงค์ดังกล่าว โดยใช้แบบทดสอบสองฉบับในการทดสอบสองขั้นตอน แบบทดสอบแต่ละฉบับในแต่ละขั้นตอนจะประกอบไปด้วยกลุ่มของข้อสอบที่มุ่งวัดเนื้อหาใดเนื้อหาหนึ่งซึ่งกำหนดทิศทางการตอบสำหรับผู้สอบไว้ เรียกว่า testlets (Thissen and Mislevy, 1990) แบบทดสอบในแต่ละขั้นตอนมีรายละเอียด ดังนี้

2.1.1 แบบทดสอบขั้นตอนแรกหรือแบบทดสอบเพื่อกำหนดทิศทาง (first-stage test or routing test) เป็นแบบทดสอบที่ค่อนข้างสั้นประกอบด้วยข้อสอบที่มีระดับความยากกระจายในช่วงกว้างหรือมีระดับความยากปานกลาง เพื่อให้ผู้สอบทำในขั้นตอนแรกและตรวจข้อสอบทันทีเพื่อใช้ผลการทดสอบเป็นเกณฑ์ในการคัดเลือกแบบทดสอบในขั้นที่สองโดยพิจารณาตามค่าความยากของแบบทดสอบที่เหมาะสมกับระดับความสามารถของ ผู้สอบซึ่งประมาณได้จากการทดสอบแบบทดสอบฉบับแรก (นันทิยา พึ่งคำ , 2531; ศิริชัย กาญจนวาสี , 2538, 2550)

2.1.2 แบบทดสอบขั้นตอนที่สองหรือแบบทดสอบวัดผล (second-stage test, main test or measurement test) เป็นแบบทดสอบที่ถูกคัดเลือกให้เหมาะกับความสามารถของผู้สอบซึ่งประมาณได้จากการทดสอบในขั้นแรกมักเป็นแบบทดสอบที่ยาวกว่าแบบทดส อบในขั้นแรก และประกอบด้วยแบบทดสอบย่อย ประมาณ 3-5 ฉบับ แยกตามระดับความยาก ในแต่ละฉบับ จะมีข้อสอบจำนวน20-30 ข้อ คะแนนที่ได้จากแบบทดสอบแต่ละฉบับไม่สามาร ถเปรียบเทียบกันได้โดยตรง เนื่องจากข้อสอบแต่ละฉบับมีความยากต่างกัน วิธีนำคะแนนมาเปรียบเทียบกันอาจทำได้โดยการนำคะแนนของแต่ละฉบับมาถ่วงด้วยค่าความยากเฉลี่ยของแต่ละฉบับเพื่อปรับค่าความยากให้เท่ากัน (ศิริชัย กาญจนวาสี, 2538, 2550)

การคัดเลือกข้อสอบและจัดชุดของข้อสอบในการทดสอบทั้งสองขั้นตอนดังกล่าว ศิริชัย กาญจนวาสี (2538, 2550) ได้เสนอไว้ว่าสามารถนำทฤ ษฎีการตอบสนองข้อสอบ (item response theory) มาประยุกต์ใช้ได้ เพื่อให้การประมาณค่าความสามารถของผู้สอบไม่ขึ้นกับระดับความยากของแบบทดสอบ

สำหรับข้อดีและข้อจำกัดของการทดสอบโดยใช้ยุทธวิธีสองขั้นตอน ศิริชัย กาญจนวาสี (2538, 2550) ได้กล่าวไว้ว่า มีข้อดีที่ช่วยลดจำนวนข้อสอบที่แต่ละคนจะต้องทำ โดยไม่ทำให้ความแม่นยำในการวัดลดลง แต่มีข้อจำกัดที่ในการทดสอบกับคนจำนวนมากในเวลาเดียวกัน การตรวจข้อสอบกำหนดทิศทางจะต้องรีบเร่งตรวจและใช้คนจำนวนมาก รวมทั้งการปรับเหมาะของกระบวนการทดสอบเกิดขึ้นเพียง ครั้งเดียวหลังจากการทดสอบในขั้นแรก ซึ่งถ้าเกิดความคลาดเคลื่อนในขั้นนี้อาจทำให้เกิดความคลาดเคลื่อนในการทดสอบในขั้นที่สองได้ ความคลาดเคลื่อนในการจำแนกกลุ่มความสามารถของผู้สอบอาจเกิดขึ้นได้ประมาณร้อยละ 20 ตลอดจนแบบทดสอบที่มีช่วงค่าความยากที่แคบจะทำให้ได้สารส นเทศค่อนข้างน้อย แต่ถ้าสร้างแบบทดสอบวัดผลในขั้นที่สองให้มีช่วงความยากที่ค่อนข้างกว้างก็จะขัดแย้งกั บหลักการของการ ทดสอบแบบปรับเหมาะ ทั้งนี้ได้เสนอแนะการแก้ปัญหาดังกล่าวไว้ว่าสามารถดำเนินการได้ โดยใช้คอมพิวเตอร์ในการทดสอบแบบปรับเหมาะ

2.2 <u>ยุทธวิธีหลายขั้นตอน (multi-stage strategies)</u>

การทดสอบแบบปรับเหมาะโดยใช้ยุทธวิธีหลายขั้นตอนเป็นการใช้การทดสอบที่มากกว่า 2 ขั้นตอน ในการกำหนดทางแยก (branch) ในการทดสอบ จำแนกได้เป็น 2 รูปแบบการทดสอบแบบทางแยกคงที่และรูปแบบการทดสอบแบบทางแยกแปรผัน (Hambleton and Swaminathan, 1985) มีรายละเอียดดังนี้

2.2.1 รูปแบบทางแยกคงที่ (fixed-branching model)

รูปแบบทางแยกคงที่ที่ใช้ในยุทธวิธีหลายขั้นตอนเหมือนยุทธวิธีสองขั้นตอนในส่วนของการกำหนดทางแยกคงที่ แต กต่างกันในส่วนของการตัดสินใจแยกทาง ซึ่งในการใช้ยุทธวิธีสองขั้นตอนกำหนดการตัดสินใจเพียงครั้ง เดียวจากการทำแบบทดสอบกำหนดทิศทางในขั้นแรกไปการทำแบบทดสอบวัดผลในขั้นที่สอง ส่วนรูปแบบทางแยกคงที่ในการใช้ยุทธวิธีหลายขั้นตอนจะกำหนดการตัดสินใจหลังจากที่ผู้สอบทำข้อสอบในแต่ละข้อตามโครงสร้างของแบบทดสอบที่กำหนดไว้ล่วงหน้า โดยคำนวณค่าพารามิเตอร์ของข้อสอบไว้แล้ว ผู้สอบแต่ละคนจะได้ทำแบบทดสอบที่มีโครงสร้างของแบบทดสอบเดียวกันแต่จะได้ทำข้อสอบในข้อที่แตกต่างกันตามวิธีการคัดเลือกข้อสอบที่กำหนดไว้ รูปแบบในการกำหนดโครงสร้างของแบบทดสอบสรุปได้ดังนี้

1) รูปแบบปิรามิดหรือรูปแบบโครงสร้างต้นไม้ (pyramidal model)

รูปแบบปิรามิดมีหลายลักษณะ ได้แก่ ปิรามิดขนาดขั้นคงที่ ปิรามิดขนาดขั้นแปรผัน ปิรามิดชนิดรอบิน- มอนโร ปิรามิดชนิดข้างตัด ปิรามิดชนิดข้างตัดสกัดการสะท้อนกลับ ปิรามิดชนิดมีหลายข้อในแต่ละขั้น ปิรามิดชนิดให้น้ำหนักตัวเลือก

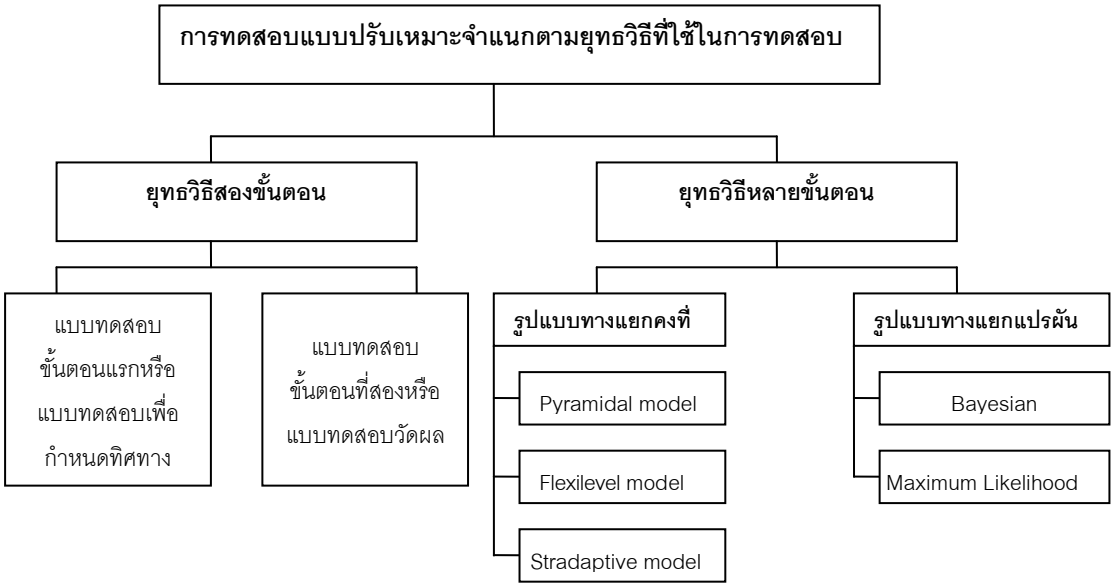
2) รูปแบบปรับระดับแบบยืดหยุ่น(flexilevel model)

3) รูปแบบปรับระดับชั้น (stradaptive model)

2.2.2 รูปแบบทางแยกแปรผัน (variable-branching model)

รูปแบบทางแยกแปรผันเป็นรูปแบบการทดสอบที่ไม่ได้กำหนดโครงสร้างและไม่ได้กำหนดข้อสอบไว้ล่วงหน้า ว่าถ้าผู้สอบตอบถูกจะต้องไปทำข้อสอบข้อใดหรือถ้าตอบผิดจะต้องไปทำข้อสอบข้ อใดแต่จะดำเนินการโดยใช้ข้อสอบที่กำหนดค่าฟังก์ชั่นสารสนเทศของข้อสอบไว้แล้ว การคัดเลือกข้อสอบเพื่อให้เหมาะสมกับความสามารถของผู้สอบจะดำเนินการโดยใช้วิธีการทางสถิติในการประมาณค่าความสามารถ ซึ่งวิธีการประมาณค่าความสามารถที่นิยมใช้มี 2 วิธี ได้แก่ การประมาณค่าแบบเบส์ (Bayesian estimation) และการประมาณค่าแบบความเป็นไปได้สูงสุด (maximum likelihood estimation)

จากการ จำแนกการทดสอบแบบปรับเหมาะตามยุทธวิธีที่ใช้ในการทดสอบสามารถสรุปยุทธวิธีที่ใช้ในการทดสอบได้ดังแผนภาพที่ 2.2



แผนภาพที่ 2.2 ประเภทของ การทดสอบแบบปรับเหมาะจำแนกตามยุทธวิธีที่ใช้ในการทดสอบ

อย่างไรก็ตามการทดสอบแบบปรับเหมาะกับความสามารถของผู้สอบโดยอาศัยเครื่องคอมพิวเตอร์ในการดำเนินการทดสอบนั้นนิยมใช้ยุทธวิธีหลายขั้นตอนซึ่งใช้ทางแยกแปรผันโดยใช้กลวิธีของเบส์และกลวิธีความเป็นไ ปได้สูงสูดเท่านั้น ส่วนยุทธวิธีหลายขั้นตอนซึ่งใช้ทางแยกคงที่ไม่เป็นที่นิยมในการนำรูปแบบนี้มาประยุกต์ใช้กับคอมพิวเตอร์

1.5 องค์ประกอบของการทดสอบแบบปรับเหมาะโดยใช้คอมพิวเตอร์

องค์ประกอบของการทดสอบแบบปรับเหมาะโดยใช้คอมพิวเตอร์ (Weiss & Kingsbury, 1984; รังสรรค์ มณีเล็ก, 2540) มีรายละเอียดดังนี้

1.5.1 โมเดลการตอบสนองข้อสอบ (item response model)

ในการประมาณค่าพารามิเตอร์ในการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์นั้นจะใช้วิธีการประมาณค่าตามทฤษฎีการตอบสนองข้อสอบ ซึ่งอาจเป็นโมเดลโลจิสติก (logistic model) หรือนอร์มอลโอไจฟ์ (normal ogive model) ชนิด 1 2 หรือ 3 พารามิเตอร์ก็ได้ ซึ่งการจะเลือกโมเดลใดมาใช้ในการวิเคราะห์ข้อมูลนั้นก็จำเป็นต้องคำนึงถึงธรรมชาติของข้อมูลว่ามีลักษณะอย่างไร มีความสอดคล้องกับโมเดลใด โมเดลการตอบสนองข้อสอบนับว่ามีความสำคัญต่อการทดสอบในลักษณะนี้ม าก เพราะการประมาณค่าความสามารถของผู้สอบนั้นจะเป็นอิสระจากข้อสอบที่สอบ ในทำนองเดียวกันการประมาณค่าพารามิเตอร์ของข้อสอบก็จะเป็นอิสระจาก

ผู้สอบ ดังนั้น จึงสามารถเปรียบเทียบความสามารถของผู้สอบได้ถึงแม้ว่าผู้สอบแต่ละคนจะได้รับข้อสอบคนละชุดที่มีค่าความยากต่างกันก็ตาม (Hambleton and Swaminathan, 1985)

1.5.2 **คลังข้อสอบที่มีความเหมาะสม** (calibrated item pool)

ความตรงของแบบทดสอบขึ้นอยู่กับความเกี่ยวข้องและความเหมาะสมของข้อสอบที่ ใช้ แต่สำหรับการทดสอบแบบปรับเหมาะโดยใช้คอมพิวเตอร์ความตรงของแบบทดสอบจะขึ้นอยู่กับทั้งจำน วนและเนื้อหาของข้อสอบในคลังข้อสอบ นอกจากข้อสอบจะต้องมีจำนวนมากเพียงพอ ในการประมาณค่าระดับความสามารถของผู้สอบซึ่งมีความแตกต่างหลากหลายกันได้อย่าง ถูกต้อง เหมาะสมแล้ว ข้อควรคำนึงที่สำคัญประการหนึ่ง ก็คือ เนื้อหาของข้อสอบ เนื่องจากการมีข้อสอบในคลังจำนวนมากทำให้ มีโอกาสที่จะมีข้อสอบที่ไม่ดีได้มาก ซึ่งข้อสอบเหล่านั้นจะมี ผลกระทบอย่างมากต่อการให้คะแนนสุดท้ายในการทดสอบแบบปรับเหมาะ เพราะในการทดสอบ แบบปรับเหมาะมีการตรวจให้คะแนนในทันที ข้อสอบที่ถูกเลือกให้ผู้สอบทำจึงมีผลต่อความ ถูกต้องในการประมาณค่าความสามารถของผู้สอบและก ารให้คะแนนสุดท้ายโดยไม่มีการแก้ไขหรือย้อนกลับได้ (Hambleton, zaal และ Pieters 2000 อ้างถึงใน Latu and Chapman, 2002) การทดสอบแบบปรับเหมาะจะมีประสิทธิภาพสูงขึ้น ถ้ามีคลังข้อสอบขนาดใหญ่ ข้อสอบดังกล่าวได้รับการประมาณค่าพารามิเตอร์โดยใช้ ทฤษฎีการตอบสนองข้อสอบ ข้อสอบมีค่าอำนาจจำแนกสูง มีค่าความยากครอบคลุมช่วงความสามารถของผู้สอบ และมีค่าการเดาต่ำ Weiss (1988) ได้เสนอไว้ว่า จำนวนข้อสอบในคลังข้อสอบควรมีอย่างน้อย 100-200 ข้อ Allen and Yen (1979) กล่าวว่า ถ้าจะให้ดีต้องมีจำนวนข้อสอบจำนวน 2^n ข้อ เมื่อ n = จำนวนข้อสอบที่ใช้ในการทดสอบ แต่ถ้าการทดสอบมีการกำหนดจำนวนข้อที่จะใช้คงที่เหมือนกันทุกคน จำนวนข้อสอบควรมีอย่างน้อย n(n+1)/2 โดยข้อสอบแต่ละข้อควรมีอำนาจจำแนกสูง ($a > 0.80$) มีค่าความยากที่ครอบคลุมระดับช่วงต่างๆอย่างเหมาะสม $[-2.0 < \theta < +2.0]$ ค่าสัมประสิทธิ์การเดา ($c < 0.30$) และควรมีข้อสอบเท่าๆกันในแต่ละระดับความยาก (Urry, 1977)

1.5.3 **จุดเริ่มต้นการทดสอบ** (starting point or entry level)

การทดสอบเริ่มต้นด้วยการคัดเลือกข้อสอบข้อแรก (initial item) การคัดเลือกข้อสอบข้อแรกนิยมใช้ข้อสอบที่มีความยากปา นกลางซึ่งสามารถแบ่งวิธีการคัดเลือกเป็น 2 กรณี (ศิริชัย กาญจนวาสี, 2550)

1) กรณีเมื่อประชากรผู้สอบมีความสามารถค่อนข้างใกล้เคียงกัน (homogeneous) หรือไม่มีข้อมูลสำหรับผลสัมฤทธิ์ที่ผ่านมาควรเริ่มต้นด้วยข้อสอบที่มีค่าความ

ยากปานกลางเท่ากันสำ หรับทุกคน เช่น นักศึกษาปีที่ 1 ระดับปริญญาตรี สาขาวิทยาศาสตร์ ควรเริ่มต้นด้วยข้อสอบข้อที่มีค่าความยากปานกลางสำหรับประชากรกลุ่มนี้

2) กรณีเมื่อประชากรผู้สอบมีความสามารถค่อนข้างแตกต่างกัน (heterogeneous) โดยพิจารณาจากข้อมูลที่มีอยู่ เช่น กลุ่มอายุ ระดับชั้นที่ศึกษา ซึ่งควรเริ่มต้นด้วยข้อสอบที่มีค่าความยากปานกลางสำหรับแต่ละระดับการศึกษาของผู้สอบ

อาจจะกล่าวได้ว่าระดับในการเริ่มต้นในการทดสอบแบบปรับเหมาะ ผู้สอบแต่ละคนจะได้ข้อสอบที่มีระดับความยากต่างกัน การเริ่มทำข้อสอบที่คลาดเคลื่อนไปจากความสามารถของผู้สอบไม่มีผลกระทบต่อการประมาณค่าความสามารถเท่าใดนัก แต่ถ้าให้ผู้สอบเริ่มทำข้อสอบที่ตรงกับความสามารถของตนก็จะทำให้ลดจำนวนข้อที่จะใช้ในการทดสอบลงได้ (พิมพ์สิริ เธียรนรเศรษฐ์, 2549) ซึ่งสอดคล้องกับ นันทิยา พึ่งคำ (2531) ที่กล่าวว่า การทดสอบควรเริ่มต้นด้ วยข้อสอบที่ตรงกับความสามารถของผู้สอบหรือใกล้เคียงกับความสามารถของผู้สอบมากที่สุด ซึ่งจะช่วยให้จำนวนข้อสอบที่ใช้ในการทดสอบสำหรับแต่ละบุคคลมีจำนวนลดลง และโดยทั่วไปนิยมเลือกข้อสอบที่มีความยากปานกลาง (Wainer, 1990; Hambleton, 1991; ศิริชัย กาญจนวาสี , 2550) ในขณะที่ Gershon (1992) แนะนำว่า ข้อสอบข้อแรกควรเป็นข้อสอบที่ค่อนข้างง่ายเพื่อให้ผู้สอบมีความรู้สึกถึงการบรรลุความสำเร็จ แต่ต้องอยู่ในสถานการณ์ที่ท้าทาย

1.5.4 **การคัดเลือกข้อสอบ** (item selection)

การคัดเลือกข้อสอบและการประมาณค่าความสามารถของผู้สอบมีประสิทธิภาพมากขึ้นเมื่อนำทฤษฎีการตอบสนองข้อสอบ (item response theory) มาประยุกต์ใช้ โดยทั่วไปนิยมใช้วิธีที่สอดคล้องกัน เพื่อความสะดวกรวดเร็วในการคำนวณ (Hulin, Drasgow & Parson, 1983 อ้างถึงใน ศิริชัย กาญจนวาสี, 2550)

<u>กรณีที่ใช้การประมาณค่าความสามารถด้วยวิธีของ maximum likelihood</u>

การประมาณค่าความสามารถด้วยวิธีของ maximum likelihood estimation นิยมใช้กับวิธีการคัดเลือกข้อสอบวิธีใดวิธีหนึ่งดังต่อไปนี้

1) คัดเลือกข้อสอบที่มีความยาก (b_i) สอดคล้องกับระดับความสามารถที่ประมาณค่าได้ (θ) [Match b_i to θ] โดยการตรวจสอบข้อสอบทุกข้อเพื่อเลือกข้อสอบข้อที่มีค่า b_i ใกล้เคียงกับ θ ที่เพิ่งประมาณได้ และใช้ข้อสอบข้อนั้นเป็นข้อถัดไป วิธีนี้ใช้กันมาตั้งแต่ตอนต้นยุคของ Tailored Testing เนื่องจากคำนวณง่ายและประหยัด (Urry, 1970; Reckase, 1973, 1974; Weiss, 1974)

2) คัดเลือกข้อสอบที่มีตำแหน่งสารสนเทศสูงสุด (m_i) สอดคล้องกับระดับความสามารถที่ประมาณได้ (θ) [Match m_i to θ] ข้อสอบมักมีปัจจัยการเดาเข้ามาเกี่ยวข้องการคัดเลือกข้อสอบที่มีค่า b_i ใกล้เคียงที่สุดกับ θ จึงมักมีแนวโน้มที่จะได้ข้อที่ยากเกินไปสำหรับผู้สอบ เนื่องจากข้อสอบแต่ละข้อให้สารสนเทศสูงสุดที่ตำแหน่ง $\theta = m_i$ เมื่อ m_i สำหรับโมเดล 3 พารามิเตอร์ คำนวณได้ดังนี้

$$m_i = b_i + \frac{1}{Da_1} \ln\left[\frac{1+\sqrt{1+8c_i}}{2}\right] \qquad (1)$$

เมื่อ D = ค่าคงที่

ln = natural logarithm

a_i = ค่าอำนาจจำแนกของข้อสอบ

b_i = ค่าความยากของข้อสอบ

c_i = ค่าโอกาสการเดาข้อสอบถูก

m_i เป็นค่า θ ณ ตำแหน่งที่ข้อสอบข้อนั้นให้สารสนเทศสูงสุด ซึ่ง $m_i > b_i$ ถ้า $c_i > 0$ แสดงว่าข้อสอบที่มีค่าความยาก b_i เหมาะที่จะใช้ประมาณค่าθ ของผู้สอบที่มีความสามารถสูงกว่า b_i เล็กน้อย ดังนั้นการคัดเลือกข้อสอบจากค่า b_i จึงสามารถปรับปรุงให้ดีขึ้น เมื่อพิจารณาจากค่าm_i

3) คัดเลือกข้อสอบที่ให้สารสนเทศสูงสุดตรงตำแหน่ง $\hat{\theta}$ (maximum Information Item information) โดยการตรวจสอบข้อสอบทุกข้อที่ยังไม่ได้นำมาใช้สอบ ที่สามารถให้สารสนเทศสูงสุด ณ ตำแหน่ง $\hat{\theta}$ การคัดเลือกข้อสอบตามวิธีนี้ จึงต้องคำนวณสารสนเทศของข้อสอบ (item information) ตรงตำแหน่ง $\hat{\theta}$ [$I(\hat{\theta}, u_i)$]

<u>กรณีที่ใช้การประมาณค่าความสามารถด้วยวิธี Bayesian</u>

การคัดเลือกข้อสอบจาก Bayes's theorem

$$f(\theta/U) = k\, L(U/\theta)\, f(\theta) \qquad (2)$$

เมื่อ $f(\theta/U)$ = posterior distribution ของ θ

$L(U/\theta)$ = likelihood function ของ vector U (item responses)

$f(\theta)$ = prior distribution ของ θ

K = constant

วิธีการคัดเลือกข้อสอบที่ใช้กันมากที่สุดควบคู่ไปกับการประมาณค่า θ ด้วยวิธี Bayesian คือ การเลือกข้อสอบข้อที่ยังไม่ได้นำมาใช้ในการทดสอบนั้นที่จะให้ค่าความแปรปรวน

ของค่าความสามารถที่คาดหวังมีค่าต่ำสุด (smallest posterior variance) นั่นคือ เป็นการเลือกข้อที่คาดว่าจะลดความไม่แน่นอนลงได้มากที่สุดในการประมาณค่า θ

1.5.5 การประมาณค่าความสามารถ (ability estimate)

Weiss (1974 อ้างถึงใน สุพัฒน์ สุกมลสันต์ , 2536) ได้กล่าวถึงวิธีการให้คะแนนการทดสอบแบบปรับเหมาะหรือการประมาณค่าความสามารถไว้หลายวิธี สรุปได้ดังนี้

1. ให้คะแนนตามค่าความยากของข้อสอบที่ยากที่สุดที่ตอบถูก (the most difficult of items answered correctly) เป็นการให้คะแนนโดยดูจากข้อสอบทั้งหมดที่ผู้สอบสามารถตอบได้ถูกต้อง หากข้อใดที่มีค่าความยากสูงสุดที่ผู้สอบตอบถูก ถือว่าเป็นค่าความสามารถของผู้สอบคนนั้น ค่าความยากดังกล่าวควรเป็นค่าความยากที่ได้จากการวิเคราะห์ข้อทดสอบตามทฤษฎีการตอบสนองข้อทดสอบ

2.ให้คะแนนตามค่าเฉลี่ยความยากของข้อที่ทำถูก (average difficulty of all items answered correctly) เป็นการให้คะแนนโดยการนำค่าความยากของข้อสอบทุกข้อที่ผู้สอบตอบถูกมาหาค่าเฉลี่ย ค่าเฉลี่ยความยากนี้ถือเป็นค่าความสามารถของผู้สอบคนนั้น

3. ให้คะแนนตามค่าเฉลี่ยความยากของทุกข้อที่ทำ (average difficulty) วิธีการให้คะแนนแบบนี้ Lord เป็นผู้เสนอ ในปี ค .ศ.1970 โดยการนำค่าความยากของข้อสอบทุกข้อที่ผู้สอบได้ตอบโดยไม่คำนึงว่าจะตอบถูกหรือไม่มาหาค่าเฉลี่ย ค่าเฉลี่ยนี้ถือเป็นค่าความสามารถของผู้สอบ

4. ให้คะแนนตามค่าความยากของข้อสุดท้ายที่ตอบ (difficulty of the final item) นั่นคือไม่คำนึงว่าขั้นสุดท้ายจะตอบถูกหรือไม่ก็ตาม จะ ถือว่าค่าความยากของข้อสอบข้อสุดท้ายที่ผู้สอบตอบคือค่าความสามารถของผู้สอบผู้นั้น

5.ให้คะแนนตามค่าความยากของข้อสอบที่ต่อจากขั้นสุดท้ายที่ตอบ (difficulty of $(N+1)^{th}$ item) ซึ่งมีลักษณะคล้ายกับวิธีที่ 4 แต่เป็นการตรวจสอบความสามารถต่อไปอีกขั้นหนึ่ง กล่า วคือ ถ้าขั้นสุดท้ายที่ผู้ตอบสามารถตอบถูก คะแนนในขั้นต่อไปก็จะเพิ่มขึ้น แต่ถ้าขั้นสุดท้ายตอบผิดคะแนนขั้นต่อไปก็จะลดลง

6. การให้คะแนนตามวิธีของเบส์ มีหลายวิธี แต่วิธีที่นิยมใช้กันมาก คือ วิธีของเบส์ที่ปรับปรุงให้ทันสมัย (Bayesian updating) ซึ่งสามารถประมาณค่าค วามสามารถของผู้สอบได้ค่อนข้างคงที่ ซึ่งมีสูตรดังนี้ (Owen, 1975 อ้างถึงใน สุพัฒน์ สุกมลสันต์, 2539; รังสรรค์ มณีเล็ก , 2540; ศิริชัย กาญจนวาสี, 2550)

กรณีที่ตอบถูก

$$\theta_{m+1} = \theta_m + (1-c)\left(\frac{\sigma_m^2}{\sqrt{\frac{1}{a^2}+\sigma_m^2}}\right)\left(\frac{O(D)}{c+(1-c)A(-D)}\right) \quad (3)$$

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

เมื่อ $D = \dfrac{b-\theta_m}{\sqrt{\frac{1}{a^2}+\sigma_m^2}}$

และ B = c+(1-c) x A(-D)

กรณีที่ตอบผิด

$$\theta_{m+1} = \theta_m - \left(\frac{\sigma_m^2}{\sqrt{\frac{1}{a^2}+\sigma_m^2}}\right)\left(\frac{O(D)}{A(D)}\right) \quad (4)$$

$$\sigma_{m+1}^2 = \sigma_m^2\left(1-\left(\frac{O(D)}{1+\frac{1}{a^2\sigma_m^2}}\right)\left(\frac{O(D)}{A(D)}+D\right)\div A(D)\right)$$

เมื่อ θ_m แทน ความสามารถของผู้สอบที่ประมาณค่าได้ก่อนตอบข้อสอบข้อที่ m+1 ซึ่งตามปกติแล้ว ถ้าไม่ทราบค่าความสามา รถเบื้องต้นของผู้สอบมักกำหนดให้เท่ากับ 0

σ_m^2 แทน ความแปรปรวนในการประมาณค่าความสามารถของผู้สอบก่อนตอบข้อที่ m+1 ถ้าไม่ทราบค่าความแปรปรวนดังกล่าวมาก่อนมักกำหนดให้เท่ากับ 1

θ_{m+1} แทน ค่าความสามารถของผู้สอบโดยประมาณหลังจากที่ตอบข้อที่ m+1

σ_{m+1}^2 แทน ค่าความแปรปรวนในการประมาณค่าความสามารถของผู้สอบเมื่อตอบข้อสอบข้อที่ m+1

a แทน พารามิเตอร์ค่าอำนาจจำแนกข้อสอบข้อที่ m+1

b แทน พารามิเตอร์ค่าความยากของข้อสอบข้อที่ m+1

c แทน พารามิเตอร์ระดับโอกาสการเดาข้อสอบได้ถูกข้อที่ m+1

D แทน จุดบนแกน X

$O(D)$ แทน ค่าออร์ดิเนท (ordinate) ของโค้งปกติที่จุด D

A (D) แทน พื้นที่ใต้โค้งปกติจากค่า D จนถึงจุด D

7. การให้คะแนนโดยวิธีการประมาณค่าความเป็นไปได้สูงสุด (maximum likelihood) มีหลายวิธี แต่วิธีที่ นิยมใช้มาก คือ วิธีความเป็นไปได้สูงสุดแบบมีเงื่อนไข (conditional maximum likelihood) การประมาณค่าโดยวิธีนี้มีข้อจำกัด คือ ถ้าผู้สอบตอบข้อสอบถูกหมดหรือผิดหมด จะไม่สามารถประมาณค่าได้ ขั้นตอนในการให้คะแนนมีดังนี้ (Hambleton and Swanminathan, 1985 อ้างถึงใน สุพัฒน์ สุกมลสันต์, 2539; รังสรรค์ มณีเล็ก, 2540; ศิริชัย กาญจนวาสี, 2550)

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

$$\theta_0 = \ln\left(\frac{r_a}{k - r_a}\right) \quad (5)$$

เมื่อ r_a = $\sum a_i U_l$

U_l = 1 เมื่อตอบข้อสอบถูก

U_l = 0 เมื่อตอบข้อสอบผิด

a_i คือ อำนาจจำแนกขอข้อทดสอบข้อที่ i

k คือ จำนวนข้อทดสอบทั้งหมดที่ผู้สอบตอบ

ขั้นที่ 2 หาค่า $P_1(\theta_m)$ และ $Q_1(\theta_m)$ โดยใช้สูตร ดังนี้

$$\mathrm{P_1(\theta_m) = c_1 + (1 - c_1)\frac{e^{Da_1(\theta_m - b_1)}}{Da_i(\theta_m - b_i)}} \quad (6)$$

$$Q_t(\theta_m) = 1 - P_1(\theta_m)$$

$$= \frac{1 - c_i}{1 + e^{Da_1(\theta_m - b_1)}}$$

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

$$h_m = \frac{D\left[r_m - \sum Pi(\theta_m)\right]}{-D^2 \sum Pi(\theta_m)Qi(\theta_m)} \quad (7)$$

ขั้นที่ 4 ประมาณค่าความสามารถของผู้สอบใหม่ $(\theta_m + 1)$ โดยใช้สูตร

$$\theta_m + 1 = \theta_m - h_m \quad (8)$$

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

จากวิธีการให้คะแนนหรือการประมาณค่าความสามารถทั้ง 7 วิธี วิธีที่นิยมนำมาใช้ในการประมาณค่าความสามารถมี 2 วิธี คือ วิธีการประมาณค่าความเป็นไปได้สูงสุด (maximum likelihood estimate) และการประมาณค่าแบบเบส์ (Bayesian estimate) วิธีการคัดเลือกข้อสอบที่ใช้สำหรับการประมาณค่าความเป็นไปได้สูงสุด ประกอบด้วยการคัดเลือกข้ อสอบที่มีค่าความยากสอดคล้องกับค่าความสามารถที่ประมาณได้ การคัดเลือกข้อสอบที่มีตำแหน่งสารสนเทศ (ค่าความยาก ค่าอำนาจจำแนกและค่าการเดา) สูงสุดสอดคล้องกับความสามารถที่ประมาณได้และการคัดเลือกข้อสอบที่ให้สารสนเทศสูงสุดที่ตำแหน่งความสามารถนั้น ส่วนวิธีการคัดเลือกข้อสอบที่ใช้สำหรับการประมาณค่าแบบเบส์ใช้การคัดเลือกข้อสอบข้อที่ยังไม่ได้นำมาใช้ในการทดสอบนั้นที่จะให้ค่าความแปรปรวนของค่าความสามารถที่คาดหวังมีค่าต่ำสุด (ศิริชัย กาญจนวาสี, 2550)

1.5.6 **เกณฑ์การยุติการสอบ** (termination criteria or Stopping criteria)

ปัจจุบันมีการพัฒนาวิธีที่ใช้ในการยุติการสอบหลายวิธี แต่วิธีที่นิยมใช้มี 2 วิธี วิธีแรก คือ การทดสอบจนกระทั่งได้ระดับความคงที่ในการวัด (level of measurement stability or consistency) โดยการกำหนดระดับความคลาดเคลื่อนที่ยอมรับได้ วิธีที่สอง คือ การกำ หนดจำนวนข้อสอบให้คงที่ วิธีการกำหนดความยาวของแบบสอบให้คงที่ช่วยให้แน่ใจได้ว่าผู้สอบได้ทำข้อสอบในจำนวนข้อสอบที่เท่ากันและช่วยให้แบบทดสอบไม่ยาวเกินไป แต่ ในทางปฏิบัติการกำหนดให้ทุกคนทำข้อสอบจำนวนเท่ากันอาจ ให้คุณภาพของการวัดผลที่แตกต่างกัน เนื่องจากมีระดับความยากของข้อสอบที่แตกต่างกัน การใช้การยุติการสอบโดยพิจารณาจากค่าความคลาดเคลื่อนในการวัดจึงเป็นวิธีที่ดีกว่า (Segall,1996; ศิริชัย กาญจนวาสี, 2550)

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}} \quad (9)$$

เมื่อ $SE(\theta)$ = ความคลาดเคลื่อนมาตรฐานของการประมาณค่า θ

$I(\theta)$ = สารสนเทศของแบบสอบที่ให้สำหรับผู้ที่มีความสามารถ θ

1.6 ประโยชน์และข้อจำกัดของการทดสอบแบบปรับเหมาะโดยใช้คอมพิวเตอร์

การทดสอบแบบปรับ เหมาะโดยใช้คอมพิวเตอร์ มีประโยชน์ต่อการทดสอบทางการศึกษาเป็นอย่างมาก แต่อย่างไรก็ตามการทดสอบดังกล่าวก็มีข้อจำกัดบางประการในการนำไปปฏิบัติ ซึ่งสามารถสรุปประโยชน์และข้อจำกัดได้ดังนี้

1.6.1 ประโยชน์ของการทดสอบแบบปรับเหมาะโดยใช้คอมพิวเตอร์

Green (1983 อ้างถึงใน Wainer, 1990) และ Latu และ Chapman (2002) ได้กล่าวถึงประโยชน์ของการทดสอบแบบปรับเหมาะโดยใช้คอมพิวเตอร์ไว้ สรุปได้ดังนี้

1. ช่วยเพิ่มประสิทธิภาพในการทดสอบ โดยเฉพาะภาระงานการทดสอบต่าง ๆ ได้แก่ การดำเนินการทดสอบ การกำหนดเกณฑ์และการให้คะแนน

2. แบบทดสอบที่เก็บไว้ในคลังข้อสอบในคอมพิวเตอร์จะมีความปลอดภัย (security) ต่อการคัดลอกมากกว่า

3. ไม่มีข้อจำกัดเรื่องเวลา ผู้สอบสามารถทำแบบทดสอบในเวลาที่แตกต่างกันตามความพร้อมของแต่ละบุคคลภายในช่วงเวลาที่กำหนด

4. ผู้สอบไม่เกิดความท้อแท้ในการทำข้อสอบ เนื่องจากเป็นการคัดเลือกข้อสอบที่เหมาะสมกับความสามารถของแต่ละบุคคล

5. การตรวจข้อสอบมีความชัดเจนขึ้น เนื่องจากไม่มีปัญหาที่เกิดจากกระดาษคำตอบ เช่น การทำเครื่องหมายไม่ชัดเจน ความไม่ชัดเจนในข้อที่เลือกเนื่องจากมีการลบเมื่อต้องการเปลี่ยนคำตอบ เป็นต้น

6. ข้อสอบสามารถตรวจ ให้คะแนนและให้ผลย้อนกลับ (feedback) กับผู้สอบได้ทันที

7. สามารถคัดเลือกข้อสอบจากการตอบข้อสอบเบื้องต้นได้อย่างรวดเร็ว

8. ผู้สอบได้ทำข้อสอบที่มีความยาวเหมาะสมกับระดับความสามารถ

9. มีการ เสนอข้อสอบให้ผู้สอบตอบในรูปแบบที่หลากหลาย ทั้งภาพนิ่ง ภาพเคลื่อนไหวและเสียงประกอบ

1.6.2 ข้อจำกัดของการทดสอบแบบปรับเหมาะโดยใช้คอมพิวเตอร์

Latu และ Chapman (2002) กล่าวถึงข้อจำกัดของการทดสอบแบบปรับเหมาะโดยใช้คอมพิวเตอร์ไว้ ดังนี้

1. ความปลอดภัยของแบบทดสอบ(test security) แม้ว่าจะมีนักวิจัยหลาย ๆคน แสดงให้เห็นว่าการทดสอบแบบปรับเหมาะโดย ใช้คอมพิวเตอร์มีความปลอดภัยของแบบทดสอบมากกว่าการทำแบบทดสอบบนโต๊ะ (desk drawer) แต่ก็มีนักวิจัยบางท่านโต้แย้งว่าปัจจุบันมีวิธีการขโมยข้อสอบได้หลายวิธี เช่น pager กล้องขนาดเล็ก (miniature cameras) กล้องวีดีโอ (video transmitters) เครื่องบันทึกวีดีโอขนาดเล็ก (micro video recorders) นอกจากนี้ผู้สอบสามารถคัดลอกข้อสอบจากเครื่องคอมพิวเตอร์ได้ง่ายและรวดเร็วกว่าการคัดลอกข้อสอบจากกระดาษคำตอบอีกด้วย

2. ปัญหาเกี่ยวกับผู้สอบ (examinee issues) แม้ว่าการทดสอบแบบปรับเหมาะจะมีประโยชน์ต่อผู้สอบมาก แต่ยังคงมีข้อจำกัดเกี่ยวกับผู้ สอบ เช่น ผู้สอบหลายคนไม่เคยชินกับการใช้คอมพิวเตอร์ ทำให้เกิดความวิตกกังวล ผู้สอบบางคนแสดงให้เห็นว่าการนำเสนอโดยใช้คอมพิวเตอร์เป็นสิ่งที่ยุ่งยาก (difficult or fatiguing) โดยเฉพาะการตอบโดยใช้เมาส์หรือคีย์บ์อร์ด

3. ค่าใช้จ่ายในการดำเนินการ (financial realities) การทดสอบแบบปรับเหมาะมีค่าใช้จ่ายในการดำเนินการสร้างแบบทดสอบ การเตรียมการและการดำเนินการสอบที่ค่อนข้างสูง เช่น Educational Testing Service (ETS) ศึกษาค่าใช้จ่ายในการทดสอบแบบปรับเหมาะโดยใช้คอมพิวเตอร์พบว่ามีค่าใช้จ่ายที่ต้องจ่ายจริงมากกว่าวิธีการทดสอบแบ บประเพณีนิยม ดังจะเห็นได้จาก ในสหรัฐอเมริกาผู้สอบจะต้องเสียค่าใช้จ่ายมากกว่า 100 ดอลลาร์ เพื่อสอบ TOFEL ในระบบการทดสอบแบบปรับเหมาะโดยใช้คอมพิวเตอร์ ค่าใช้จ่ายที่แพงมากดังกล่าวมาจากการพัฒนาคลังข้อสอบ การบริหารการทดสอบและการพัฒนาโปรแกรมสำเร็จรูป ซึ่งมีค่าใช้จ่ายในการพัฒนาและการนำไปใช้ในระบบที่สมบูรณ์ระหว่าง 3,000 - 5,000 ดอลลาร์

1.7 ทัศนคติต่อการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์

ถึงแม้การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์จะมีข้อดีสำหรับผู้สอบ แต่ก็มีข้อด้อยเช่นกัน ตัวอย่างเช่น ผู้สอบจำนวนมากรู้สึกไม่คุ้นเคยต่อสิ่งแวดล้อมของการใช้คอมพิวเตอร์ในการทดสอบ ซึ่งจะทำให้เพิ่มความวิตกกังวลในการสอบ(Latu & Chapman, 2002) นอกจากนั้น Wise (1997) กล่าวว่าผู้สอบบางคนมีความเห็นว่าการใช้คอมพิวเตอร์ในการนำเสนอข้อสอบนั้นมีความยากและเกิดความเมื่อยล้าจากการใช้เมาส์ และคีย์บอร์ดในการสอบ ซึ่งสามารถเรียก

พฤติกรรมนั้นว่า ความวิตกกังวลเกี่ยวกับคอมพิวเตอร์ (computer anxiety) ซึ่งเป็นการตอบสนองที่เกิดกับผู้ที่มีความกังวลเกี่ยวกับความอันตรายที่เกิดจากคอมพิวเตอร์ จากการวิจัยพบว่า การทดสอบด้วยคอมพิวเตอร์อาจจะทำให้ผู้สอบบางคนเ กิดความกลัว ความวิตกกังวล (Abdelhamid, 2002; Igbaria and Chakrabarti, 1990; Torkzadeh and Angulo, 1992 cited in Wiechmann and Ryan, 2003) นอกจากนั้นยังพบหลักฐานว่า ความวิตกกังวลเกี่ยวกับคอมพิวเตอร์ มีความสัมพันธ์ทางลบต่อทัศนคติที่มีต่อคอมพิวเตอร์ (Igbaria and Chakrabarti, 1990 cited in Wiechmann and Ryan,2003)

จากการศึกษาที่ผ่านมา พบว่า ผู้ที่มีประสบการณ์เกี่ยวกับคอมพิวเตอร์มากจะมีทัศนคติในด้านบวกต่อการทดสอบแบบปรับเหมาะ โดย Lever, Sherrod, Bransford (1989 cited in Sutton, 1993) ได้ศึกษากับนักเรียนในโรงเรี ยนระดับประถมศึกษา Arenz & Lee, 1990; Loyd & Loyd, 1988 (cited in Sutton, 1993) ศึกษากับนักเรียนในระดับมัธยมศึกษา และ Loyd & Loyd, 1988; Wu & Mogan, 1989 (cited in Sutton, 1993) ศึกษากับนักศึกษาระดับวิทยาลัย นอกจากนั้นงานวิจัยที่ ผ่านมาจำนวนมาก พบว่า ประสบการณ์เกี่ยวกับคอมพิวเตอร์มีความสัมพันธ์ทางบวกกับการรับรู้หลังจากการทดสอบผ่านคอมพิวเตอร์ และมีความสัมพันธ์กับความเต็มใจที่จะใช้คอมพิวเตอร์ (Zolton and Chapanis, 1982 cited in Wiechmann and Ryan, 2003) การเชื่อมโยงเกี่ยวกับการคำนวณ ทัศนคติโดยทั่วไป ต่อคอมพิวเตอร์ (Igbaria and Chakrabarti, 1990; Popovich, Hyde, Zakrajsek and Blumer, 1987 cited in Wiechmann and Ryan, 2003) และจะพบว่าผู้ที่มีประสบการณ์เกี่ยวกับคอมพิวเตอร์น้อยจะเห็นว่าการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ไม่คุ้นเคย ไม่ยุติธรรม นอกจากนั้น Bandura (1997) กล่าวว่า การรับรู้ความสามารถตนเองด้านคอมพิวเตอร์ (computer self-efficacy) แยกออกมาจากโครงสร้างของการรับรู้ความสามารถตนเอง (self-efficacy) นั่นคือ มีความสัมพันธ์ทางบวกกับทัศนคติ ความตั้งใจ และพฤติกรรมที่มีต่อคอมพิวเตอร์ เช่นเดียวกับtest self-efficacy ผู้สอบที่มีความสามารถคอมพิวเตอร์ต่ำย่อมมีอารมณ์ในด้านลบที่จะส่งผลต่อการรับรู้หลังจากการทดสอบ

จากการศึกษาที่ผ่านมาสรุปได้ว่าปัจจัยที่สำคัญที่จะส่งผลให้ผู้สอบมีทัศนคติทางบวกต่อการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์คือ ความสามารถหรือทักษะด้านคอมพิวเตอ ร์ของผู้สอบ ซึ่งจะทำให้ลดความวิตกกังวลของผู้สอบด้วย

1.8 ประเด็นปัญหาของการใช้การทดสอบแบบปรับเหมาะ

แม้ว่าการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์จะมีประสิทธิภาพในการทดสอบสูง แต่การทดสอบนี้ก็ยังมีประเด็นปัญหาที่ต้องการการแก้ไข ปรับปรุงเพื่อให้การทดสอบมีความ

น่าเชื่อถือมากขึ้น ซึ่งปัญหาที่พบต่าง ๆมักจะเป็นปัญหาเกี่ยวกับวิธีการในการดำเนินการทดสอบแบบปรับเหมาะ อาทิ การคัดเลือกข้อสอบข้อแรก การให้คะแนน การยุติการสอบ ซึ่งบางประเด็นได้มีผู้ศึกษาไว้แล้วจำนวนมาก ในที่นี้จึงขอยกประเด็นที่ได้รับความสนใจอยู่ในขณะนี้ และยัง มีผู้ศึกษาในประเด็นดังกล่าวไม่มากนัก ประเด็นที่กำลังได้รับความสนใจ ได้แก่ วิธีการคัดเลือกข้อสอบข้อแรก การควบคุมการใช้ข้อสอบซ้ำ และ การสร้างความสมดุลของเนื้อหา

1.8.1 วิธีการคัดเลือกข้อสอบขั้นแรก

เมื่อ 2 ทศวรรษที่ผ่านมา วิธีการทั่วไปที่ใช้มากที่สุ ดในการคัดเลือกข้อสอบจะพิจารณาที่ค่าสารสนเทศสูงสุด โดยเฉพาะอย่างยิ่งข้อสอบที่ให้สารสนเทศสูงสุด ณ จุดที่ประมาณค่าระดับความสามารถขณะนั้น นั่นคือ ถูกประมาณค่าจากการตอบสนองข้อสอบที่ผ่านมา ณ เวลานั้น สำหรับวิธีการอื่นนอกเหนือจากวิธีสารสนเทศสูงสุด คือ วิธีของเบย์ (Bayesian) โดยแทนที่จะใช้สารสนเทศของข้อสอบที่จุด $\hat{\theta}$ วิธีของเบย์จะใช้ความแปรปรวนภายหลัง (posterior variance) เป็นเกณฑ์สำหรับคัดเลือกข้อสอบ การแจกแจงภายหลัง (posterior distribution) ณ จุดเริ่มต้น จะขึ้นอยู่กับทางเ ลือกของการแจกแจงก่อนหน้าสำหรับ θ เป็นสำคัญ แต่จะลดน้อยลงในขั้นต่อไป(Chang and Ying, 1996)

นอกจากนั้นจากการศึกษาของChang และ stout (1993) พบว่า วิธีความแปรปรวนภายหลังจะให้ค่าที่ตรงข้ามกับสารสนเทศของแบบสอบ เมื่อจำนวนข้อสอบมากขึ้น สำหรั บสารสนเทศของข้อสอบนั้น มีการนิยามไว้เช่นเดียวกับสารสนเทศของฟิชเชอร์ ซึ่งจะมีการผันแปรจากผู้สอบคนหนึ่งไปยังผู้สอบคนหนึ่ง และสิ่งนั้นคือหน้าที่ของ θ ค่าสารสนเทศของฟิชเชอร์ ณ จุดที่ผู้มีความสามารถ θ แสดงโดย θ_0 ซึ่งเป็นสิ่งที่ชี้ให้เห็นถึงความมีประสิทธิภาพของข้อสอบในการประมาณค่า θ อย่างไรก็ตาม ค่าที่ได้ในตำแหน่งความสามารถ θ จะอยู่ห่างจาก θ_0 ซึ่งอาจจะไม่เป็นตัวบ่งชี้ที่ดี ที่จะทำให้เกิดประสิทธิภาพมากขึ้น เพราะการใช้สารสนเทศของฟิชเชอร์ในตำแหน่งความสามารถปัจจุบัน ($\hat{\theta}$) เกณฑ์ของสารสนเทศอาจจะไม่มีประสิทธิภาพ ถ้า $\hat{\theta}$ ไม่เข้าใกล้ θ_0 สิ่งนี้อาจจะเกิดขึ้นในขั้น เริ่มต้นของการทดสอบแบบปรับเหมาะ เมื่อการสอบเริ่มใช้ข้อสอบไปเพียงไม่กี่ข้อ (ซึ่งจะให้สารสนเทศน้อย)

การคัดเลือกข้อสอบขั้นแรก จึงอาจจะไม่มีประสิทธิภาพ ด้วยเหตุนี้ประเด็นการคัดเลือกข้อสอบที่ดีที่สุดในขั้นแรกจึงได้รับความสนใจมากขึ้น วิธีการใหม่นี้เป็น การศึกษาตามทฤษฎีและข้อมูลเชิงประจักษ์เพื่อที่จำนำเสนอการประมาณค่า θ ที่มีประสิทธิภาพและมีความแม่นยำ โดย Veerkamp และ Berger (1994) ได้เสนอวิธี “เกณฑ์สารสนเทศแบบช่วง (interval information criterion)” แทนการหาสารสนเทศของข้อสอบแบบจุด แต่ วิธีคัดเลือกข้อสอบของพวก

เขา จะพิจารณาจากค่าเฉลี่ยสูงสุดของฟังก์ชั่นสารสนเทศในช่วงความเชื่อมั่น อย่างไรก็ตาม Stocking (1993) โต้แย้งว่า นอกจากสารสนเทศของข้อสอบแล้ว การคัดเลือกข้อสอบควรเป็นการรวมเกณฑ์ที่เสริมกัน เช่น เงื่อนไขและอัตราการใช้ข้อสอบซ้ำ คลังข้อสอ บใหม่หรือมีการแทนที่รายละเอียดของการทดสอบ และลำดับ ประเภทของข้อสอบ นอกจากนั้น Chang และ Ying (1996) ได้เสนอมโนทัศน์เกี่ยวกับสารสนเทศ ที่เรียกว่า "global information" ซึ่งจะให้สารสนเทศเมื่อตัวประมาณค่าไม่เข้าใกล้ค่าพารามิเตอร์ที่แท้จริง รวมถึงวิธีการคัดเลือกข้อสอบโดยใช้ค่าเฉลี่ยของ global information ซึ่งการคัดเลือกข้อสอบรูปแบบนี้คือ Kullback-Leiber information (KL) ซึ่ง Chang และ Ying กล่าวว่าเป็นวิธีการที่ให้สารสนเทศมากกว่าวิธีการแบบดั้งเดิมหรือ Fisher information (FI)

Chang และ Ying (1996) ได้แบ่งสารสนเทศของข้อสอบออกเป็น 2 ประเภทคือ local information และ global information โดยสารสนเทศที่จัดอยู่ใน local information คือ Fisher information ซึ่งสารสนเทศจะอยู่รอบๆ θ ที่แท้จริงหรือมีลักษณะเป็นสารสนเทศแบบจุด (the item information at point) ควรใช้ในกรณีที่มี n ขนาดใหญ่ สำหรับสารสนเทศที่จัดอยู่ใน global information คือ Kullback-Leiber information ซึ่งสารสนเทศจะอยู่ภายนอกขอบเขตของ θ ที่แท้จริงหรือมีลักษณะเป็นเกณฑ์สารสนเทศแบบช่วง (interval information criterion) เหมาะสำหรับ ใช้ในกรณีที่มี n ขนาดเล็ก ซึ่งต่อมาได้มีผู้ศึกษาประเด็นการคัดเลือกข้อสอบขั้นแรกเพิ่มมากขึ้น แต่ผลลัพธ์จากการวิจัยยังให้ผลที่ไม่สอดคล้องกัน รายละเอียดได้เสนอไว้ในงานวิจัยที่เกี่ยวข้อง

การคัดเลือกข้อสอบมีความสัมพันธ์กับคุณภาพของการประมาณค่าคุณลักษณะ $(\hat{\theta})$ เป็นอย่างมาก การประมาณค่าคุณลักษณะยิ่งมีความถูกต้องแม่นยำมากยิ่งขึ้น ถ้าข้อสอบที่ถูกคัดเลือกมีความเหมาะสมกับความสามารถของผู้สอบ การประมาณค่าคุณลักษณะของผู้สอบด้วยข้อสอบที่เป็นขั้นแรกของการทดสอบแบบปรับเหมาะกับความสามารถของผู้สอบด้วยคอมพิวเตอร์ มีความคลาดเคลื่อนมากกว่าข้อสอบขั้นที่เป็นข้อต่อมา จนถึงข้อสอบขั้นที่เป็นข้อสุดท้ายของการทดสอบ ดังนั้นข้อสอบที่ถูกคัดเลือกที่ขั้นแรกมีแนวโน้มที่จะเหมาะสมน้อยกว่าข้อสอบที่ถูกคัดเลือกในขั้นต่อมาเพื่อปรับแก้ให้การคัดเลือกข้อสอบขั้นแรกของการทดสอบแบบปรับเหมาะกั บความสามารถของผู้สอบด้วยคอมพิวเตอร์มีประสิทธิภาพ ดังนั้นจึงได้มีผู้เสนอเกณฑ์การคัดเลือกข้อสอบที่ได้รับการปรับแก้แล้วหลายเกณฑ์ ได้แก่ สารสนเทศฟิชเชอร์แบบถ่วงน้ำหนัก (FII; Veekamp and Berger, 1997) สารสนเทศของฟิชเชอร์ที่มีการแจกแจงภายหลัง (FIP; van der Linden, 1995) สารสนเทศของคูลเบค- ไลเบลอร์ (KL; Chan and Ying, 1996)

และสารสนเทศของคูลเบค-ไลเบลอร์ที่มีการแจกแจงภายหลัง (KLP; Chan and Ying, 1996) ซึ่งมีรายละเอียดดังนี้

1) วิธีสารสนเทศของฟิชเชอร์

ให้ $P_j(\theta)$ แทน ฟังก์ชันการตอบสนองข้อสอบข้อที่ j

$Q_j(\theta)$ แทน 1 - $P_j(\theta)$

ซึ่งฟังก์ชันสารสนเทศมีดังนี้

$$I_j(\theta) = \frac{\left[\frac{\partial P_j(\theta)}{\partial \theta}\right]}{P_j(\theta)Q_j(\theta)} = \frac{\left[P_j'(\theta)\right]^2}{P_j(\theta)Q_j(\theta)} \quad (10)$$

สมการที่ 10 เป็นการอธิบายเกณฑ์การคัดเลือกสารสนเทศของฟิชเชอร์ โดยกำหนดว่าข้อสอบที่ถูกเลือกจะให้ค่าสารสนเทศสูงสุด $I_j(\hat{\theta}_n)$ สำหรับการประมาณคุณลักษณะชั่วคราว $(\hat{\theta}_n)$ ตามการใช้ข้อสอบ n ข้อ สำหรับโมเดลการตอบสนองข้อสอบโลจิสติก 3 พารามิเตอร์แบบทวิภาค (dichotomous) สมการที่10 ทำให้เข้าใจง่าย ดังนี้(Hambleton, Swaminathan and Rogers, 1991)

$$\mathrm{I_j(\theta)} = \frac{2.89\mathrm{a_j^2(1-c_j)}}{\left[\mathrm{c_j} + \exp\left[1.7\mathrm{a_j(\theta - b_j)}\right]\left[1 + \exp\left[-1.7\mathrm{a_j(\theta - b_j)}\right]\right]\right]^2} \quad (11)$$

ซึ่ง a_j เป็น พารามิเตอร์ค่าอำนาจจำแนกข้อสอบ

b_j เป็น พารามิเตอร์ค่ายากของข้อสอบ

c_j เป็น พารามิเตอร์ระดับโอกาสการเดาข้อสอบได้ถูก

สมการที่ 11 แสดงให้เห็นว่าสารสนเทศที่เพิ่มขึ้น คือ b_j เข้าใกล้ (θ), a_j เพิ่มขึ้น และ c_j เข้าใกล้ 0 (Hambleton et al., 1991) เกณฑ์การคัดเลือกข้อสอบสารสนเทศของฟิชเชอร์ที่ใช้ในการทดสอบแบบ ปรับเหมาะด้วยคอมพิวเตอร์ มีข้อตกลงว่าเมื่อการประมาณค่าคุณลักษณะความสามารถของผู้สอบ $(\hat{\theta})$ ตามข้อสอบที่ถูกใช้ n ข้อซึ่งข้อสอบที่ถูกคัดเลือกเป็นข้อสอบที่ให้ค่าสารสนเทศสูงสุด $I_j(\hat{\theta}_n)$

2) วิธีสารสนเทศของฟิชเชอร์ที่มีการถ่วงน้ำหนัก (FII)

Veerkamp and Berger (1997) ได้เสนอเกณฑ์การคัดเลือกข้อสอบโดยทั่วไป ซึ่งเรียกว่า เกณฑ์สารสนเทศถ่วงน้ำหนักโดยทั่วไป (general weighted information criterion: GWIC)

$$GWIC_j(\theta) = \int_{-\infty}^{\infty} W(\theta) I_j \, d\theta \qquad (12)$$

ซึ่ง $W(\theta)$ เป็น ฟังก์ชันถ่วงน้ำหนัก

$I_j(\theta)$ เป็น สารสนเทศของฟิชเชอร์

บนฐานของ GWI ค่าฟังก์ชันสารสนเทศของฟิชเชอร์ที่มากเกินช่วงระดับคุณลักษณะจะถูกรวมให้เป็นค่าเดียว คิดเป็นค่าเฉลี่ยถ่วงน้ำหนัก (weighted average) เมื่อใช้เป็นเกณฑ์การคัดเลือกข้อสอบแล้ว ข้อสอบที่มีค่าเฉลี่ยถ่วงน้ำหนักสูงที่สุดจะเป็นทางเลือกหนึ่งสำหรับผู้สอบ

การสร้างเกณฑ์การคัดเลือกข้อสอบที่แตกต่างกันนั้น สามารถพัฒนาเป็นเกณฑ์คัดเลือกที่เหมาะสม โดย การปรับฟังก์ชันถ่วงน้ำหนัก เช่น เมื่อการ ประมาณคุณลักษณะเข้าใกล้ระดับคุณลักษณะที่แท้จริง ฟังก์ชั นสารสนเทศของฟิชเชอร์ที่สูงที่สุดเป็นการคัดเลือกข้อสอบที่เหมาะสมและฟังก์ชันถ่วงน้ำหนักในสมการที่ 12 สามารถแสดงได้ดังนี้

$$W(\theta) = \begin{cases} 1, \theta = \hat{\theta}_n \\ 0, otherwise \end{cases} \qquad (13)$$

เช่น การใช้การประมาณค่าคุณลักษณะแบบจุดเท่านั้น และไม่สนใจระดับคุณลักษณะที่เป็นไปได้อย่างอื่น เป็นที่น่าสังเกตว่า เมื่อสมการที่ 13 ถูกใช้เป็นฟังก์ชันถ่วงน้ำหนักใน GWIC สมการที่ 12 ซึ่งได้เป็นเกณฑ์การคัดเลือกสารสนเทศของฟิชเชอร์ (F-rule)

3) สารสนเทศของฟิชเชอร์ที่มีการแจกแจงภายหลัง (FIP)

จากวิธีสารสนเทศของฟิชเชอร์ที่มีการถ่วงน้ำหนัก (FII) เมื่อมีการอธิบายถึงความไม่แน่นอนของการประมาณค่าคุณลักษณะในขั้นแรกของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ วิธีแก้ปัญหา คือ การกำหนดการแจกแจงภายหลังเป็นฟังก์ชันถ่วงน้ำหนักอาจจะเป็นฟังก์ชันที่เหมาะสมสำหรับการพิจารณาระดับคุณลักษณะทั้งหมด ในกรณีเกณฑ์ก ารคัดเลือกข้อสอบ สามารถแสดงได้ดังนี้

$$FIP_j(X_n) = GWIC_j(X_n) = \int_{-\infty}^{\infty} P(\theta | X_n) I_j(\theta) d\theta \qquad (14)$$

ให้ $I_j(\theta)$ เป็น ฟังก์ชันสารสนเทศของฟิชเชอร์ สำหรับข้อที่ j

$P(\theta | X_n)$ เป็น การแจกแจงคุณลักษณะภายหลัง หลังจากที่ข้อสอบ n ข้อถูกใช้สอบ

4) สารสนเทศคูลเบค-ไลเบเลอร์ (KL)

Chang and Ying (1996) ได้แสดงให้เห็นฟังก์ชันคูลเบค-ไลเบเลอร์ดังนี้

$$KL_j(\theta\|\theta_0) = P_j(\theta_0)\log\left[\frac{P_j(\theta_0)}{P(\theta)}\right] + \left[1-P_j(\theta_0)\right]\log\left[\frac{1-P_j(\theta_0)}{1-P_j(\theta)}\right] \quad (15)$$

ให้ $P_j(\theta)$ แทน ฟังก์ชันการตอบสนองข้อสอบสำหรับข้อที่ j ซึ่ง θ และ θ_n แทน 2 ระดับคุณลักษณะ โดย Chang and Ying สังเกตว่ามีคุณลักษณะที่สำคัญหลายประการของฟังก์ชันสารสนเทศคูลเบค-ไลเบเลอร์ (KL) ได้แก่

1) $\mathrm{KL}_j(\theta\|\theta_0) \neq \mathrm{KL}(\theta_0\|\theta)$

2) $\mathrm{KL}_j(\theta\|\theta_0) \geq 0$ และ $\mathrm{KL}_j(\theta_0\|\theta_0) = 0$

3) คล้ายกับผลรวมของค่าฟังก์ชันสารสนเทศของ ฟิชเชอร์ กล่าวคือ สารสนเทศระดับคะแนนแบบสอบเป็นผลรวมของสารสนเทศระดับข้อสอบ ดังนี้

$$\mathrm{KL}^n(\theta\|\theta_0) = \sum_{j=1}^{n}\mathrm{KL}_j(\theta\|\theta_0) \quad (16)$$

ในขณะที่ GWIC คือ ฟังก์ชันถ่วงน้ำหนักที่สามารถประยุกต์ใช้กับ $\mathrm{KL}_j(\theta\|\theta_0)$ เพื่อที่จะหาเกณฑ์การคัดเลือกข้อสอบ ซึ่งฟังก์ชันถ่วงน้ำหนักทั่วไป คือ

$$W(\theta,\theta_0) = \begin{cases} 1, \theta \in (\hat{\theta}-\delta_1, \hat{\theta}+\delta_2) \text{and} \theta_0 \in (\hat{\theta}-\delta_3, \hat{\theta}+\delta_4) \\ 0, \text{otherwise} \end{cases} \quad (17)$$

ซึ่งการประมาณค่าความสามารถ $(\hat{\theta})$ เป็นค่าความสามารถที่แท้จริง (θ) ตามข้อสอบ n ข้อ และพื้นที่ที่ใช้เป็นเกณฑ์ในการคัดเลือกข้อสอบ กล่าวคือ ข้อสอบที่มีพื้นที่สูงสุด ซึ่งเท่ากับสารสนเทศ $\mathrm{KL}_j(\theta\|\theta_0)$ ที่เฉลี่ยสูงสุดจะถูกคัดเลือก โดยมีสูตรดังนี้

$$\mathrm{KL}_j(\hat{\theta}) = \int_{\hat{\theta}_l}^{\hat{\theta}_u}\mathrm{KL}_j(\theta\|\hat{\theta})d\theta \quad (18)$$

เมื่อ $(\hat{\theta}_l, \hat{\theta}_u) = \left(\hat{\theta}-\frac{z}{\sqrt{n}}, \hat{\theta}+\frac{z}{\sqrt{n}}\right)$

$\frac{z}{\sqrt{n}}$ เป็น ลิมิตความเชื่อมั่น (confidence limit)

5) สารสนเทศคูลเบค-ไลเบลอร์ ที่มีการแจกแจงภายหลัง (KLP)

Chang and Ying (1996) ได้แทนฟังก์ชันสารสนเทศฟิชเชอร์ $\left(I_j(\theta)\right)$ ในสมการที่ 14 ด้วยสารสนเทศคูลเบค-ไลเบลอร์ $\left(\mathrm{KL}\left(\theta\middle\|\hat{\theta}_n\right)\right)$ ซึ่งได้เกณฑ์การคัดเลือกข้อสอบดังนี้

$$\mathrm{KLP}_j(\hat{\theta}_n, U_n) = \int_{-\infty}^{\infty} P(\theta|U_n)\mathrm{KL}_j\left(\theta\middle\|\hat{\theta}_n\right)d\theta \qquad (19)$$

1.8.2 การควบคุมการใช้ข้อสอบซ้ำ

แม้ว่าการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์จะเจริญเติบโตขึ้นอย่างรวดเร็ว แต่การแก้ปัญหาบางปัญหายังไม่สามารถหาวิธีมาแก้ไขได้ ปัญหาหนึ่งก็คือ การควบคุมอัตราการใช้ข้อสอบซ้ำ เพราะวิธีการแบบปรับเหมาะถูกออกแบบมาให้เลือกข้อสอบ ที่เหมาะสมโดยจะเลือกข้อสอบที่มีค่าอำนาจจำแนกสูงสุด ดังนั้นข้อสอบบางข้อจึงอาจจะถูกเลือกมากกว่าข้ออื่น (Lee, IP, and Fuh, 2008)

การควบคุมการใช้ข้อสอบซ้ำมีผู้เสนอวิธีการควบคุมการใช้ข้อสอบซ้ำต่าง ๆ กันหลายวิธีเพื่อจะควบคุมการใช้ข้อสอบซ้ำในการทดสอบแบบปรับเหมา ะด้วยคอมพิวเตอร์ (Chang and Ying, 1999; Davey and Parshall, 1995; McBride and Martin, 1983; Revuelte and ponsoda, 1998; Stocking and Lewis, 1998, 2000; Sympson and Hetter, 1985; cited in Chen and Lei, 2005; Chang and Ansley, 2003) ถึงแม้ว่าวิธีการส่วนใหญ่นี้ สามารถที่จะควบคุมการใช้ข้อสอบซ้ำได้เป็นอย่างดีในลักษณะที่ว่าข้อสอบทั้งหลายที่ถูกใช้แล้วมีอัตราการใช้ข้อสอบซ้ำน้อยกว่าค่าที่ได้กำหนดไว้ล่วงหน้า ซึ่งมีวิธีการควบคุมการใช้ข้อสอบซ้ำดังนี้

1) วิธีการของ McBride และ Martin (1983) (MM)

เทคนิค 5-4-3-2-1 ของ McBride และ Martin เลือกข้อคำถามแบบสุ่มสำหรับการบริหารการสอบจากกลุ่มของข้อคำถาม ซึ่งมีส่วนทำให้ค่าสารสนเทศในการประมาณค่าความสามารถในปัจจุบันมีค่าโดยประมาณเท่ากัน สำหรับผู้สอบที่มีการประมาณค่าความสามารถในระยะเริ่มต้น ข้อสอบข้อแรกที่ถูกเลือกมาจะเกิดจากการสุ่มจากข้อ คำถาม 5 ข้อดีที่สุดที่ถูกกระบุไว้ว่ามีความเหมาะสมที่สุด สำหรับข้อสอบ ข้อที่สองจะถูกสุ่มมาจากข้อคำถาม 4 ข้อดีที่สุดที่มีปรับการประมาณค่าความสามารถใหม่ ต่อจากนั้นข้อสอบข้อที่สามจะถูกสุ่มมากจากข้อคำถาม 3 ข้อดีที่สุด ข้อสอบข้อที่สี่จะถูกสุ่มมากจากข้อคำถาม 2 ข้อดีที่สุด และข้อสอบข้อที่ห้า และข้อสอบข้ออื่นๆที่ตามมา จะถูกเลือกให้มีความเหมาะสม ข้อสอบนั้นจะถูกเลือก แต่จะไม่นำกลับเข้าไปในคลังข้อสอบอีกเพื่อที่จะถูกเลือกอีกครั้ง

2) วิธีการของ Sympson และ Hetter (1985) (SH)

การใช้ข้อสอบซ้ำสามารถที่จะควบคุมได้เป็นอย่างดีภายใต้วิธีการของ Sympson และ Hetter (1985) ซึ่งจะเป็นพื้นฐานสำหรับวิธีการในการควบคุมการใช้ข้อสอบซ้ำที่มีประสิทธิภาพมากยิ่งขึ้น วิธี SH ได้ถูกนำเสนอเพื่อที่จะควบคุมการใช้ข้อสอบซ้ำโดยตรงในลักษณะของความเป็นไปได้ เช่น ข้อสอบที่ถูกใช้มากที่สุดด้วยความถี่ที่น้อยกว่า อัตราการใช้ข้อสอบซ้ำที่สูงสุดที่ได้มีการกำหนดไว้ล่วงหน้า (r_{max}) ซึ่งมีความน่าจะเป็น 3 อย่างดังต่อไปนี้

P(S) คือ ความน่าจะเป็นที่ข้อสอบที่ถูกเลือกเป็นข้อสอบที่ดีที่สุดตามวิธีการคำนวณ

P(A) คือ ความน่าจะเป็นที่ได้ถูกใช้จริงกับผู้เข้าสอบ

P(A/S) คือ ความน่าจะเป็นแบบมีเงื่อนไขที่ข้อสอบ1 ข้อถูกใช้ถ้าข้อสอบนั้นได้ถูกเลือกเป็นข้อสอบที่ดีที่สุดซึ่งได้ถูก เรียกว่าเป็น “พารามิเตอร์การใช้ข้อสอบซ้ำ” ด้วยการที่จะเข้าถึงความต้องการที่ไม่มีข้อสอบที่ถูกใช้มากกว่า r_{max} ความสัมพันธ์ ระหว่างความน่าจะเป็นทั้ง 3 อย่างนี้สามารถอธิบายได้ดังนี้ คือ $P(A) = P(A/S) \times P(A) \leq r_{max}$

พารามิเตอร์การใช้ข้อสอบซ้ำคือ P(A/S) ซึ่งมีบทบาทสำคัญในการควบคุมอัตราการใช้ข้อสอบซ้ำในวิธีการของ SH จุดประสงค์ของ P(A/S) เพื่อปรับ P(S) ในลักษณะที่สามารถทำให้ P(A) น้อยกว่าหรือเท่ากับ r_{max} ถึงแม้จะเป็นการง่ายที่จะทำให้ทราบค่าของ P(A/S) แต่ในการกำหนด P(S) ก็ไม่ใช่เรื่องง่าย ยิ่งไปกว่านั้นเมื่อ P(A/S) ได้ถูกกำหนดขึ้นสำหรับข้อสอบข้อใดข้อหนึ่งแล้วค่าของ P(S) สำหรับข้อส่วนที่เหลือในคลังข้อสอบจะถูกเปลี่ยนแปลงด้วยค่าของ P(A/S) ที่มีส่วนเกี่ยวข้องกับข้อสอบเหล่านั้น ดังนั้นการจำลองข้อมูลซ้ำที่ต่อเนื่องกันเป็นชุดมีความจำเป็นที่ต้องทำเพื่อที่จะทำให้ค่าของ P(S) และ P(A/S) คงที่ ซึ่งอยู่ภายใต้เงื่อนไขที่ค่าทั้งหมดของ P(A) น้อยกว่าหรือเท่ากับ r_{max}

3) วิธีการของ Davey และ Parshall (1995) (DP)

Stocking และ Lewis (1998) ได้เสนอแนะให้ใช้วิธีการควบคุมการใช้ข้อสอบซ้ำที่ถูกกำหนดให้ที่ระดับคุณลักษณะ (trait levels) แต่ Davey และ Parshall (1995) ได้เสนอวิธีการควบคุมการใช้ข้อสอบซ้ำที่ถูกกำหนดที่ข้อสอบ (items) ที่ได้ใช้ในการทดสอบแล้วที่ Davey และ Parshall ใช้วิธีนี้เพื่อที่จะหลีกเลี่ยงข้อสอบที่จะปรากฏเป็นกลุ่มย่อยมากเกินไปสำหรับผู้เข้าสอบ ตามวิธีของ Davey และ Parshall การควบคุมการใช้ข้อสอบซ้ำสามารถที่จะปรับปรุงได้ที่ระดับแบบสอบและระดับข้อสอบแต่ละข้อ อย่างไรก็ตามวิธีนี้ไม่สามารถรับประกันได้ว่าอัตราคาบเกี่ยวกันระหว่างแบบสอบที่ได้กำหนดไว้ล่วงหน้าสามารถที่จะตอบสนองได้อย่างแน่ชัด เพื่อที่จะให้มีความปลอดภัยในแบบสอบ (test security) การจำลองข้อมูลซ้ำสำหรับ P(A/S) ที่คงที่เป็นการใช้

เวลามากด้วยเหมือนกัน โดยเฉพาะอย่างยิ่งเมื่อมีคลังข้อสอบจำนวนมาก ดังนั้นวิธี DP จึงเป็นการยากที่จะนำมาใช้ในการปฏิบัติ Stocking และ Lewis (1998)

4) วิธีการแบบไม่มีเงื่อนไขของ Stocking และLewis (1998) (SL)

วิธีการแบบไม่มีเงื่อนไขของ Stocking และLewis ได้มาจากการปรับวิธีการของ SH วิธีนี้ได้พัฒนาพารามิเตอร์การควบคุมการใช้ข้อสอบซ้ำสำหรับข้อสอบแต่ละข้อตามวิธีการคำนวณของวิธี SH แต่แตกต่างในเรื่องการคัดเลือกข้อสอบ มากกว่าการใช้การคัดเลือกข้อสอบที่เหมาะสม Stocking และLewis ได้ใช้โมเดล multinomial ในการเลือกข้อสอบข้อถัดไป โดยที่การกระจาย Multinomial แบบสะสมถูกสร้างขึ้นจากความสำเร็จของการบวกเพิ่ มเข้าไปของความน่าจะเป็นที่จะทำให้เกิดผล Stocking และLewis (1995) กล่าวว่า วิธีนี้ยังคงรักษาข้อดีของยุทธวิธี SH ไว้ นั่นคือ วิธีนี้สามารถควบคุมความน่าจะเป็นของการบริหารข้อสอบได้โดยตรง เมื่อการทดสอบแบบปรับเหมาะถูกนำเสนอต่อผู้สอบที่เป็นกลุ่มเป้าหมาย แต่ลักษณะเช่นนี้จะพบกับปัญหาน้อยกว่าวิธีการของ SH ระหว่างการทำซ้ำของการปรับการจำลองข้อมูล

5) วิธีการแบบมีเงื่อนไขของ Stocking และLewis (1998) (SLC)

ถึงแม้ว่าการใช้ข้อสอบซ้ำสามารถที่จะควบคุมได้ดีสำหรับข้อมูลทั้งหมดที่น่าสนใจตามวิธีของ SH ซึ่งไม่สามารถรับประกัน ได้ว่าอัตราการใช้ข้อสอบซ้ำที่ระดับคุณลักษณะที่กำหนดให้ได้ถูกควบคุมเป็นอย่างดีด้วยข้อสอบ 1 ข้อ อาจจะถูกใช้ซ้ำกับผู้สอบเข้าสอบทั้งหมดที่ระดับคุณลักษณะเฉพาะ ถึงแม้ว่าอัตราการใช้ข้อสอบโดยทั่วๆ ไป จะมีอัตราต่ำสำหรับผู้เข้าสอบในระดับคุณลักษณะทั้งหมด ดังนั้น การค วบคุมอัตราการใช้ข้อสอบซ้ำที่ระดับคุณลักษณะแต่ละระดับดูเหมือนจะจำเป็นในทางปฏิบัติ

Stocking และ Lewis (1998) ได้เสนอวิธีการ Multinomial แบบมีเงื่อนไขเพื่อที่จะควบคุมอัตราการใช้ข้อสอบซ้ำที่ระดับคุณลักษณะตามวิธีการของ Stocking และLewis (1998) การจำลองข้อมูลทวนซ้ำสำหรับการหาค่า P(A/S) คงที่ได้ถูกดำเนินการภายใต้ระดับ คุณลักษณะเฉพาะมากกว่าข้อมูลทั้งหมดที่น่าสนใจ โดยอาศัยเมตริกซ์ของค่า P(A/S) ที่คงที่อัตราการใช้ข้อสอบซ้ำสามารถควบคุมได้อย่างมีประสิทธิภาพที่ระดับคุณลักษณะเฉพาะ อย่างไรก็ตาม การจำลองข้อมูลทวนซ้ำสำหรับวิธี SLC เป็นการใช้เวลามาก

1.8.3 การสร้างความสมดุลของเนื้อหา

การสร้างความสมดุลในเนื้อหา (content balancing) มิได้เป็นปัญหาของการวัดทางจิต ซึ่งไม่เกี่ยวกับโมเดลการวัด IRT/TRT ซึ่งได้คัดเลือกสำหรับการทดสอบ เพื่อที่จะให้มีการวัด

ที่ถูกต้องและเป็นธรรมสำหรับเนื้อหาทั้งหมด การสร้างความสมดุลของเนื้อหาโดยระบบของ CAT ย่อมเป็นสิ่งที่จำเป็น (Aimee, Michelle, Boyd, 2003 อ้างถึงใน สิริลักษณ์ เกษรปทุมานันท์ , 2549) การสร้างสมดุลของเนื้อหาในระบบ CAT มีวิธีการดังนี้ วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่มีเงื่อนไขบังคับ (constrained CAT: CCAT), วิธีโมเดลมัลติโนเมลที่ได้แก้ไขปรับปรุงแล้ว (modified multinomial model: MMM), วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่มีเงื่อนไขบังคับที่ได้มีการแก้ไขปรับปรุงแล้ว (modified constrained CAT: MCCAT) โดยมีหลายละเอียดในแต่ละวิธีดังนี้

1) วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่มีเงื่อนไขบังคับ (constrained CAT: CCAT) ปรัชญาพื้นฐานของ CCAT คือ ข้อสอบควรจะถูกคัดเลือกจากขอบเขตของเนื้อหา (content area) ที่มีอัตราการใช้ข้อสอบซ้ำในปัจจุบันน้อยที่สุด คิดจากอัตราการใช้ข้อสอบซ้ำตามเป้าหมายที่กำหนดไว้ล่วงหน้า

2) วิธีโมเดลมัลติโนเมลที่ได้แก้ไขปรับปรุงแล้ว (modified multinomial model) ทำการแจกแจงสะสม (cumulative distribution) ขึ้นก่อนโดยเทียบอัตราการใช้ข้อสอบซ้ำตามเป้าหมายที่ได้กำหนดไว้ล่วงหน้าของขอบเขตเนื้อหา (content areas) ทั้งหมดซึ่งรวมเป็นหน่วยเดียวกันแล้วตามด้วยการแจกแจงแบบ multinomial โดยเปรียบเทียบเปอร์เซ็นต์เป้าหมายของขอบเขตเนื้อหาซึ่งมีผลรวมเป็น 1.0 เสร็จแล้วให้ใช้ตัวเลขสุ่มจากการแจกแจงแบบสะสมซึ่งการแจกแจงสะสมดังกล่าวให้คัดเลือกข้อสอบที่เหมาะสมที่สุด อ ย่างไรก็ดีเนื่องจากความคลาดเคลื่อนในการสุ่ม เปอร์เซ็นต์ตามเป้าหมายของขอบเขตเนื้อหาอาจจะไม่เป็นไปตามที่คาดไว้ในการทดสอบจริง ดังนั้นเมื่อใดก็ตามที่ขอบเขตเนื้อหาเข้าถึงเปอร์เซ็นต์ที่ตั้งเป้าไว้ก็ต้องทำการปรับการแจกแจงแบบ Multinomial ใหม่ โดยการใช้เปอร์เซ็นต์ข องการที่บรรลุเป้าหมายของขอบเขตของเนื้อหานั้นยังเหลืออยู่ ในขณะที่กลไกการสุ่มได้รวบรวมอยู่ในวิธีนี้แล้วลำดับของขอบเขตเนื้อหาต่างกัน

3) วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่มีเงื่อนไขบังคับที่ได้มีการแก้ไขปรับปรุงแล้ว (modified constrained CAT: MCCAT) วิธีนี้เป็นวิธีที่ได้ปรับปรุงมาจาก CCAT แทนที่จะพยายามคัดเลือกข้อสอบจากขอบเขตเนื้อหาซึ่งอยู่ต่ำสุดจากอัตราการใช้ข้อสอบซ้ำที่ได้กำหนดไว้ แต่ก็จะคัดเลือกเอาข้อสอบที่เหมาะสมจากขอบข่ายของเนื้อหาที่ยังไม่ได้ใช้ด้วยวิธีนี้ผลที่ไม่พึงประสงค์จากการจัดลำดับโดยวิธี CCAT ก็จะสามารถหลีกเลี่ยงได้

โดยสรุป การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ เป็นการทดสอบที่ใช้แบบสอบต่างชุดกัน โดยปรับเปลี่ยนไปตามความสามารถของผู้สอบซึ่งมีการคัดเลือกข้อสอบที่มีความยากเหมาะสมกับระดับความสามารถของผู้สอบ ซึ่งการทดสอบแบบปรับเหมาะด้วยคอมพิวเต อร์ที่นิยมกันในปัจจุบันเป็นการทดสอบแบบปรับเหมาะหลายขั้นตอนแบบทางแยกแปรผัน ซึ่งเป็นวิธีการที่ต้องใช้คอมพิวเตอร์มาใช้ในการคำนวณและดำเนินการทดสอบ โดยขั้นตอนที่สำคัญของการทดสอบแบบปรับเหมาะจะเริ่มจากการเริ่มต้นการสอบ การคัดเลือกข้อสอบ การประมาณค่าความสามารถ และ ยุติการสอบ ซึ่งแต่ละขั้นตอนก็มีวิธีการที่เหมาะสมเพื่อให้การประมาณค่าความสามารถมีความแม่นยำ และการทดสอบมีประสิทธิภาพมากที่สุดในแต่ละกรณีไป อย่างไรก็ตาม แม้การทดสอบแบบปรับเหมาะนี้จะมีข้อดีจำนวนมาก แต่การทดสอบนี้ก็มีจุดอ่อนเช่นกัน เช่น ความปลอดภัยในการทดสอบ หรือ อเกิดความลำเอียงกับผู้สอบที่ไม่มีความชำนาญในการใช้คอมพิวเตอร์ เป็นต้น และจุดอ่อนอย่างหนึ่งของการทดสอบแบบปรับเหมาะที่ได้กล่าวไปทั้งหมด คือ การทดสอบที่ผ่านมาใช้ทฤษฎี การตอบสนองข้อสอบ ซึ่งมีข้อตกลงเบื้องต้นเกี่ยวกับความเป็นเอกมิติของข้อสอบ ทำให้ไม่สอดคล้องกั บการทดสอบบางประเภทที่มีลักษณะเป็นการตอบสนองข้อสอบแบบพหุมิติ ด้วยเหตุนี้จึงทำให้เกิดแนวคิดการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์สำหรับโมเดลการตอบสนองข้อสอบแบบพหุมิติขึ้นมา ซึ่งจะกล่าวถึงรายละเอียดต่อไป และเพื่อให้เกิดความเข้าใจเกี่ยวกับโมเดลการตอบสนองข้อสอบแบบ พหุมิติ (multidimensional item response theory model) ผู้นำเสนอจึงได้เสนอมโนทัศน์เบื้องต้นเกี่ยวกับ โมเดลการตอบสนองข้อสอบแบบพหุมิติ ซึ่งมีรายละเอียดที่กล่าวไว้ในตอนต่อไป

ตอนที่ 2 มโนทัศน์เกี่ยวกับโมเดลการตอบสนองข้อสอบแบบพหุมิติ(Multidimensional Item Response Theory Model)

โมเดลการตอบสนองข้อสอบแบบพหุมิติ (multidimensional item response theory model; MIRT model) ถือได้ว่าเป็นแนวคิดที่แผ่ขยาย มาจากโมเดลการตอบสนองข้อสอบแบบพหุมิติ (unidimensional item response theory model; UIRT model) เพื่อแก้ไขจุดอ่อนของข้อตกลงเบื้องต้นเกี่ยวกับความเป็นเอกมิติ โดยจะแตกต่างจากแนวคิดแบบเดิมในส่วนของระเบียบวิธีของโมเดล MIRT ที่นำมาประยุกต์และการแปลความหมายของผลที่ได้ โดยมีรายละเอียดดังต่อไปนี้ (พัชรี จันทร์เพ็ง, 2550)

2.1 ที่มาและความสำคัญของโมเดล MIRT

แนวคิดของทฤษฎีการ ตอบสนองข้อสอบแบบพหุมิติ (MIRT) ได้มาจากแนวคิดพื้นฐานที่ สำคัญ จำแนกออกเป็น 2 กลุ่มแนวคิด โดย กลุ่มแรกเป็นการศึกษาโดย Spearman (1927) & Thurstone (1947) ทำการศึกษาและพบว่าโมเดล MIRT มีที่มาจากการวิเคราะห์องค์ประกอบ (Factor Analysis) เนื่องจากลักษณะของการวิเคร าะห์องค์ประกอบจะ มีความสัมพันธ์โดยตรงกับการวิเคราะห์MIRT ส่วนอีกกลุ่มหนึ่ง เป็นการศึกษาของ Lazarsfeld (1950) กล่าวว่า MIRT ได้รับอิทธิพลมาจาก UIRT เช่นเดียวกับ การวิเคราะห์องค์ประกอบแต่แนวคิดเกี่ยวกับ UIRT ในช่วงแรกยังไม่เป็นที่นิยมเหมือนเช่นปัจจุบัน จากทั้งสองแนวคิดอาจกล่าวว่า MIRT เป็นแนวคิดที่แผ่ขยายมาจาก UIRT หรือ เป็นกรณีเฉพาะของการวิเคราะห์องค์ประกอบหรือโมเดลสมการเชิงโครงสร้าง (structural equation modeling) นั่นเอง เพื่อให้เห็นถึงความเหมือนและความต่างของการวิเคราะห์องค์ประกอบซึ่งเป็นรากฐานของโมเดล MIRT จึงได้นำเสนอวิธีการวิเคราะห์องค์ประกอบและทฤษฎีการตอบสนองข้อสอบ (item response theory; IRT) ที่มีความสัมพันธ์กับรากฐานของแนวคิด MIRT พร้อมทั้งเปรียบเทียบวิธีการของการวิเคราะห์องค์ประกอบ และ MIRT โดยมีรายละเอียดต่อไปนี้

2.1.1 การวิเคราะห์องค์ประกอบ (Factor Analysis)

จากการศึกษาของ Reckase (1997) พบว่า มีนักวิชาหลายท่านได้มีการศึกษาและเห็นตรงกันว่าถึงแม้กระบวนการทางคณิตศาสตร์ของการวิเคราะห์องค์ประกอบและ MIRT จะมีลักษณะระเบียบวิธีทางด้านตัวเลขที่คล้ายกัน ซึ่งทั้งสองวิธีพ ยายามที่จะระบุสเกลสมมติฐา น เพื่อถอดแบบข้อมูลเพื่อนำไปสู่การวิเคราะห์ข้อมูล โดยสเกลของการวิเคราะห์องค์ประกอบและ MIRT จะต้องมีการกำหนดจุดกำเนิด (origin) และหน่วยของการวัด (unit of measurement) เหมือนกัน แต่สิ่งที่ MIRT แตกต่างจากการวิเคราะห์องค์ประกอบ เมื่อพิจารณาถึงคุณลักษณะของข้อสอบหรือตัวแปรตั้งต้น (input variable) พบว่าการวิเคราะห์องค์ประกอบจะพิจารณาถึงความแตกต่างของ ลักษณะตัวแปรตั้งต้น เช่น ความแตกต่างของค่าเฉลี่ย($\bar{x}$) และส่วนเบี่ยงเบนมาตรฐาน (SD) เป็นต้น รวมถึงความเที่ยงของการวิเคราะห์ข้อมูล เช่น การกำจัดตัวแปรแทรกซ้อนหรือตัวแปรที่ไม่จำเป็นออกจากโมเดลการวิเคราะห์ โดยการปรับค่าสถิติให้เป็นมาตรฐาน จะเห็นได้ว่าการวิเคราะห์องค์ประกอบจะให้ความสำคัญกับเมตริกซ์ความสัมพันธ์แต่ไม่ให้ความสำคัญกับความแตกต่างของค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของตัวแปรที่ทำการศึก ษา เช่น จากการศึกษาของ Herman (1976) ได้อธิบายลักษณะของการวิเคราะห์องค์ประกอบว่า “หลักการเบื้องต้นของการวิเคราะห์องค์ประกอบคือ เป็นกระบวนการการจัดกระทำกับข้อมูลหรือ

ตัวแปรเพื่อจัดกลุ่มตัวแปรในการจำแนกประเภทหรือองค์ประกอบ ให้ได้จำนวนน้อยที่สุด วิธี การดังกล่าวสามารถทำได้โดยการศึกษาความสัมพันธ์ระหว่างตัวแปรซึ่งเมตริกซ์ความสัมพันธ์ ถือเป็นข้อมูลหลักในการวิเคราะห์ โดยไม่ได้ให้ความสนใจกับค่าพารามิเตอร์ของข้อสอบที่เป็นคุณลักษณะที่มีความสัมพันธ์กับข้อสอบหรือตัวแปรตั้งต้นโดยตรง เช่น ค่าความยากของข้อสอบ อำนาจจำแนกของข้อสอบ หรือโอกาสในการเดาของข้อสอบ เป็นต้น แต่จะพิจารณาตัวแปรแทรกซ้อนอื่นๆ ที่จะตัดออกจากโมเดลการวิเคราะห์เพื่อให้ความสัมพันธ์มีความถูกต้องมากขึ้น (Carroll, 1945)

2.1.2 **ทฤษฎีการตอบสนองข้อสอบ** (Item Response Theory; IRT)

เมื่อพิจารณา แนวคิดของ IRT และการวิเคราะห์องค์ประกอบจะเห็นได้ว่ามีจุดมุ่งหมายในการวิเคราะห์ที่แตกต่างกัน โดยการวิเคราะห์องค์ประกอบจะมุ่งไปที่การกำหนดองค์ประกอบให้น้อยที่สุดที่ถอดแบบข้อมูลในรูปของเมตริกซ์การตอบข้อสอบ ส่วนโมเดล IRT จะศึกษาความสัมพันธ์ระหว่างผู้สอบและข้อสอบ

จากการศึกษา แนวคิดของ IRT ในช่วงแรกยัง อยู่บนข้อตกลงเบื้องต้นของการอธิบายค่าพารามิเตอร์ของผู้สอบในมิติความสามารถเพียงมิติเดียว (Lord & Novick, 1968; Rasch, 1960) ซึ่งง่ายต่อการฝ่าฝืนข้อตกลงเบื้องต้นและส่งผลต่อการประมาณค่าพารามิเตอร์ไม่ถูกต้อง (Camilli, Wang & Fesq, 1995) ดังนั้นการศึกษาในช่วงแรกจึงพิจารณาลักษณะการวัดแบบพหุมิติในมุมมองของ IRT ดังที่จะนำเสนอต่อไปนี้

Rasch (1960) ในช่วงแรก Rasch ได้ศึกษาเฉพาะโมเดล UIRT ต่อมาในปี ค.ศ. 1962 ได้นำเสนอโมเดลที่ยอมรับโดยทั่วไปรวมถึงความเป็นไปไ ด้ของความสามารถของผู้สอบโดยนำเสนอในรูปของเวคเตอร์มากกว่าค่าสเกลา ซึ่งสามารถแสดงได้ดังสมการที่ 20

$$P\left(x\middle|\theta_j,\sigma_i\right)=\frac{1}{\gamma\left(\theta_j,\sigma_i\right)}\exp\left[\phi(x)'\theta_j+\psi(x)'\sigma_s+\theta_j'\chi(x)\sigma_i+\rho(x)\right] \qquad (20)$$

เมื่อ ϕ,ψ,χ และ ρ แทน ฟังก์ชันของคะแนนในแต่ละข้อ

X แทน ค่าสถิติที่ได้จากการสังเกตได้จากค่าพารามิเตอร์ของผู้สอบและแบบสอบ

σ แทน เวคเตอร์ค่าพารามิเตอร์ของข้อสอบ

$\gamma(\ )$ แทน ฟังก์ชัน Normalizing ที่จะสามารถรับประกันได้ว่าค่าที่อยู่ในฟังก์ชันของ IRT อยู่ในช่วง 0 และ 1

จากการพิจารณาโมเดลข้างต้นจะเห็นว่ามีการระบุระดับความสามารถของผู้สอบในรูปของเวคเตอร์ค่าพารามิเตอร์ โดยการประมาณค่าพารามิเตอร์ของผู้สอบและข้อสอบมีความเป็นอิสระจากกัน เมื่อพิจารณาส่วนประกอบอื่นๆ เช่น ϕ,ψ,χ ซึ่งเป็นเวคเตอร์ของกา รให้คะแนน

จะต้องทราบข้อมูลในส่วนนี้ก่อนทำการวิเคราะห์ด้วยโมเดลราสช์ (Rasch model) ไม่สามารถประมาณค่าได้จากข้อมูลในการตอบของผู้สอบ สำหรับในกรณีของข้อมูลที่มีลักษณะเป็นเอกมิติการให้คะแนนจะมี 2 ค่า คือ 0 หรือ 1 โดยไม่ได้คำนึงถึงลักษณะข้อสอบ ดังนั้นการพิจารณาเฉพาะจำนวนที่ตอบข้อสอบได้ถูกต้อง ถือเป็นค่าสถิติที่เพียงพอสำหรับการพิจารณาค่า θ

ถ้าหากมีการตอบสนองข้อสอบในมิติที่ต่างกัน 2 ด้าน โดยคะแนนการตอบข้อสอบในแต่ละข้อของมิติที่ 1 และ 2 คือ μ_1 และ μ_2 ดังนั้นค่าสถิติที่เพียงพอสำหรับประมาณค่า θ_1 และ θ_2 คือ $n\mu_1$ และ $n\mu_2$ เมื่อ n คือ จำนวนข้อที่ตอบข้อสอบได้ถูกต้อง อย่างไรก็ตามเนื่องจาก μ_1 และ μ_2 เป็นค่าคงที่ข้อสอบ ดังนั้นคะแนนที่ได้แต่ละมิติเป็นฟังก์ชันของ n เท่านั้น ดังนั้นค่า θ_1 และ θ_2 ที่ประมาณได้ จึงมีความสัมพันธ์กันอย่างสมบูรณ์ นั่นคือโมเดลที่ได้ยังคงเป็นโมเดลแบบเอกมิตินั่นเอง จากปัญหาดังกล่าวนี้ มีนักวิจัยหลายท่าน พยายามที่จะแบ่งคะแนนให้มากกว่า2 ค่า เช่น จัดกระทำกับข้อสอบจากลักษณะ*"Dichotomous Items"* แปลงให้เป็น *"Single Polytomous Item"* Rackase (1972) เพื่อที่จะนำไปสู่การวิเคราะห์แบบ *"Polytomous Items*" (Kelderman, 1994) ดังนั้นฟังก์ชันการให้คะแนนจึงมีความซับซ้อนมากขึ้นหรือกำหนดให้ฟังก์ชันการให้คะแนนแตกต่างในแต่ละข้อก่อนที่จะวิเคราะห์ด้วยโมเดลราสช์เพื่อทำให้การวิเคราะห์มีความเป็นเหตุเป็นผลมากขึ้นสำหรับการวิเคราะห์คุณลักษณะของข้อสอบ (Glass, 1992) จะเห็นได้ว่าไม่นิยมนำโมเดลราสช์มาใช้ในการวิเคราะห์แบบพหุมิติ เนื่องจากมีความซับซ้อนของกระบวนการและความถูกต้องเกี่ยวกับฟังก์ชันการให้คะแนนที่มีความเฉพาะเจาะจง

ต่อมา Lord & Novick (1968) ได้เสนอโมเดลพื้นฐานของ MIRT ที่บ่งชี้ถึงนิยามของมิติคุณลักษณะแฝงได้อย่างสมบูรณ์และข้อตกลงเบื้องต้นของความเป็นอิสระ ซึ่งความเป็นอิสระในที่นี้หมายถึงคุณลักษณะของกลุ่มผู้สอบต่างๆที่มีค่าเหมือนกันของมิติความสามารถ $\theta_1, \theta_2, ..., \theta_k$ เป็นการกระจายของคะแนนผลการตอบข้อนั้นของผู้สอบแต่ละคนไม่สัมพันธ์กัน เมื่อ k คือจำนวนของมิติ ส่วน θ ในที่นี้เป็นเวคเตอร์มิติคุณลักษณะที่สมบูรณ์ นอกจากนี้ Lord & Novick (1968) แสดงให้เห็นถึงความสัมพันธ์ระหว่างโมเดลปกติสะสมตามทฤษฎีตอบสนองข้อสอบแบบเอกมิติและโมเดลองค์ประกอบร่วมอย่างชัดเจนมากยิ่งขึ้น

นอกจากนี้ Samejima (1974) ได้เสนอสูตรที่ใช้โดยทั่วไป โดยมีข้อตกลงเบื้องต้นว่าการตอบที่มีลักษณะการตรวจให้คะแนน 2 ค่าหรือมากกว่า 2 ค่า เป็นผลมาจากการแบ่งเป็นส่วนย่อยๆ จากการตอบของตัวแปรต่อเนื่อง ซึ่ง Samejima (1974) ได้พัฒนาโมเดล MIRT สำหรับข้อสอบซึ่งมีลักษณะเป็นการตอบแบบต่อเนื่อง แสดงได้ดังสมการที่ 21

$$P_{z_i}(\theta) = 2\pi^{-1/2} \int_{-\alpha}^{a_i(\theta - b_i)} \exp\left(-\frac{u^2}{2}\right) du \qquad (21)$$

แม้ว่า โมเดลของ Samejima (1974) เป็นโมเดลแรกที่ทำให้แนวความคิดเกี่ยวกับโมเดล MIRT มีความชัดเจน แต่จากการศึกษาของ Bejar (1977) พบว่าโมเดลดังกล่าวไม่สามารถนำไปประยุกต์ใช้ได้จริ งในทางปฏิบัติ ซึ่งสาเหตุอาจเกิดจากการตอบข้อสอบ มีการพิจารณาเป็นตัวแปรต่อเนื่องซึ่งไม่สอดคล้องกับบริบทของแบบสอบทั้งทางด้านการศึกษาและจิตวิทยา แต่อย่างไรก็ตามถ้ามีการประเมินเกี่ยวกับทักษะการปฏิบัติ โมเดลดังกล่าวก็จะเหมาะในการประยุกต์ใช้ได้มากขึ้น (cited in Reskase, 1997)

2.1.3 เปรียบเทียบวิธีการของการวิเคราะห์องค์ประกอบ และ MIRT (Comparison of the Factor Analysis and MIRT Approach)

จากการศึกษาถึงวิธีการวิเคราะห์องค์ประกอบและ MIRT ของ Bock & Aitkin (1981), Samejima (1974) และ McDonald (1967) พบว่า สูตรทางสถิติของทั้งสองวิธีจะมีลักษณะเป็นเมตริกซ์ของการตอบข้อสอบที่ตรวจให้คะแนนแบบ 2 ค่า ซึ่งความ เป็นจริง โปรแกรมคอมพิวเตอร์ที่พัฒนาโดย Bock et. al ในปี ค.ศ. 1988 สามารถนำมาประยุกต์ใช้ทั้งการวิเคราะห์องค์ประกอบและ MIRT ซึ่งต่อมา McDonald (1989) พัฒนาโปรแกรมสำหรับวิเคราะห์ตามโมเดล MIRTรวมถึง การวิเคราะห์องค์ประกอบเชิงยืนยัน เมื่อพิจารณาถึงแนวความคิดของทั้ง factor analysis และ MIRT พบว่ามีทั้งส่วนที่คล้ายคลึงกันและส่วนที่ต่างกัน ซึ่งพัชรี จันทร์เพ็ง (2550) ได้สรุปความแตกต่างของทั้ง 2 วิธี ดังแสดงในตาราง 2.2

ตารางที่ 2.2 สรุปความแตกต่างของการวิเคราะห์องค์ประกอบ (factor analysis) และ MIRT

Methodology	Factor Analysis	MIRT
1. Focus	มุ่งไปที่เทคนิคการสกัดข้อมูลและสังเคราะห์ข้อมูลให้มีขนาดเล็กลง (data reduction technique) เพื่อให้ได้มิติที่น้อยที่สุดของข้อมูลและง่ายต่อความเข้าใจสำหรับการบ่งชี้คุณลักษณะต่างๆ	ไม่ได้มุ่งไปที่เทคนิคการสกัดข้อมูลและสังเคราะห์ข้อมูลให้มีขนาดเล็กลงแต่มุ่งไปที่ปฏิสัมพันธ์ระหว่างผู้สอบและข้อสอบ
2. Input Variable	สนใจ correlation หรือ variance/covariance matrices มากกว่าค่าเฉลี่ยและความสัมพันธ์ระหว่างตัวแปร ซึ่ง $\bar{x}$ และ S.D. มีผลน้อยมากหรืออาจจะไม่มีผลต่อผลการวิเคราะห์ที่ได้	ค่า $\bar{x}$ และ S.D. ของการตอบข้อสอบแทนด้วยค่าพารามิเตอร์ความยากและอำนาจจำแนกของข้อสอบ ซึ่งการประมาณค่าและการตีความหมายของค่าพารามิเตอร์เหล่านี้ถือเป็นส่วนประกอบสำคัญในการวิเคราะห์
3. Data Analysis	เมตริกซ์ที่ใช้เป็นข้อมูลในการวิเคราะห์ คือ Z-Score matrix นั่นคือ เป็นเมตริกซ์ความสัมพันธ์ (correlation matrices)	เป็นเมตริกซ์ของฟังก์ชันทางคณิตศาสตร์ในการตอบสนองข้อสอบ ที่จะระบุถึงคุณลักษณะในมิติของ θ

ตารางที่ 2.2 (ต่อ)

Methodology	Factor Analysis	MIRT
	ที่มีการแปลงคะแนนแบบเส้นตรงของคะแนนที่สังเกตได้ไปสู่ Z-Score ส่วนค่าความแปรปรวนร่วม (covariance matrices) คือ เมตริกซ์ของการแปลงคะแนนเชิงเส้นตรงของจำนวนคะแนนที่ตอบข้อสอบได้ถูกต้อง โดยมีค่าเฉลี่ยเท่ากับ 0	
4. Goodness of Fit Under Hypothesis Model	โมเดลสมมติฐานมุ่งไปที่เมตริกซ์ความแปรปรวนและเมตริกซ์ความแปรปรวนร่วมโดยภาพรวมของกลุ่มผู้สอบ มากกว่าตัวแปรใดตัวแปรหนึ่งหรือการเลือกเฉพาะกลุ่มย่อยใดกลุ่มย่อยหนึ่ง มุ่งไปที่การวัดโดยภาพรวมมากกว่าเงื่อนไขการวัดความสอดคล้องของข้อมูล	โมเดลที่แสดงปฏิสัมพันธ์ระหว่างผู้สอบและข้อสอบ มีเป้าหมายของการระบุถึงความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องแม่นยำของแต่ละบุคคลในแต่ละมิติของ θ โดยให้ความสำคัญกับข้อสอบที่ไม่สอดคล้องกับโมเดล หรือไม่สอดคล้องกับการทำนายความน่าเป็นโดยเฉพาะในช่วงพิสัยของความสามารถ ซึ่งเงื่อนไขของการวัดที่สอดคล้องกับแบบสอบชุดหนึ่งๆ จะต้องอยู่ภายใต้เงื่อนไขความสอดคล้องของผู้สอบ (Liou & Chang, 1992) และความเหมาะสมของการวัด (Drasgow, Levine & McLaughlin, 1991)
5. Analysis & Application	แม้ว่าวิธีการวิเคราะห์จะมุ่งไปที่การหมุนแกนตัวประกอบ (procrustes rotation) และสัมประสิทธิ์ความสอดคล้อง เพื่อที่จะแสวงหาองค์ประกอบร่วมกันแต่ยังมีลักษณะที่เน้นย้ำน้อยกว่า MIRT	วิธีการวิเคราะห์จะมีการศึกษาข้ามแบบสอบและกลุ่มผู้สอบ จุดมุ่งหมายของการวิเคราะห์เพื่อให้อยู่ในสเกลร่วมกัน โดยมีประโยชน์ในการจัดเก็บเป็นคลังข้อสอบ หรือการจัดชุดของแบบสอบเพื่อวัดโครงสร้างหรือคุณลักษณะที่มุ่งวัด (fixed-form test construction) หรือ การสร้างแบบสอบให้เหมาะสมกับความสามารถของผู้สอบ (adaptive testing)

2.2 ประเภทของโมเดล MIRT

โมเดล MIRT ประกอบด้วยค่าพารามิเตอร์ของผู้สอบตั้งแต่ 2 ค่าขึ้นไป โดยการศึกษาในหลายมิติ จะทำให้โมเดลมีความสอดคล้องกับข้อมูลมากขึ้น สำหรับผู้สอบที่มีความแตกต่างกันอย่างเป็นระบบ เนื่องจากความยากของชุดข้อสอบ โมเดลหลายโมเดลที่เป็นแบบพหุมิติค่าอำนาจจำแนกของข้อสอบจะส่งผลกระทบต่อมิติในแต่ละชุดข้อสอบ โดยมีสัญลักษณ์ที่แสดงในโมเดล ดังนี้

X_{is} = ผลการตอบของผู้สอบที่ s ในข้อที่ i

θ_{sm} = ระดับความสามารถ(Level Trait) สำหรับบุคคลที่s ในมิติ (dimension) ที่ m

β_i = ค่าความยากของข้อที่ i

δ_i = Easiness Intercept สำหรับข้อที่ i

α_{im} = ค่าอำนาจจำแนกของข้อที่ I ในมิติ (dimension) ที่ m

γ_i = ค่าโอกาสการเดาข้อสอบของข้อที่ i

การนำเสนอโมเดลแบบพหุมิติ Embretson & Reise (2000) ได้เสนอโมเดล MIRT เป็น 2 ประเภท คือ โมเดลการตอบสนองข้อสอบแบบพหุมิติเชิงสำรวจ (exploratory MIRT model) และโมเดลการตอบสนองข้อสอบแบบพหุมิติเชิงยืนยัน (confirmatory MIRT model) ซึ่งมีลักษณะคล้ายคลึงกับการวิเคราะห์องค์ประกอบ โดย exploratory MIRT model เป็นโมเดลที่เกี่ยวข้องกับการประมาณค่าพารามิเตอร์ของข้อสอบและผู้สอบที่มากกว่า1 มิติ เพื่อให้โมเดลมีความสอดคล้องกับข้อมูลโดยไม่ได้กำหนดจำนวนขององค์ประกอบไว้ล่วงหน้า ซึ่งตรงข้ามกับ confirmatory MIRT model เป็นโมเดลที่เกี่ยวข้องกับการประมาณค่าพารามิเตอร์ในมิติที่มีความเฉพาะเจาะจง ซึ่งคล้ายคลึงกับการวิเคราะห์องค์ประกอบ โดยการวิเคราะห์จะเกี่ยวข้องกับการระบุความสัมพันธ์ของข้อสอบไปยังมิติต่างๆ โดยมีรายละเอียดในแต่ละโมเดลดังต่อไปนี้

2.2.1 โมเดลการตอบสนองข้อสอบแบบพหุมิติเชิงสำรวจ (exploratory MIRT model)

ในการวิเคราะห์องค์ประกอบซึ่งมีลักษณะคล้ายกับโมเดล MIRT ดังที่กล่าวมาแล้วข้างต้น ความจริงแล้วเมื่อพิจารณาตามข้อตกลงเบื้องต้น สามารถพิสูจน์ได้ว่าเป็นโมเดลเดียวกัน (Takane & de Leeuw, 1988) ซึ่ง McDonald (1967) ได้เสนอโมเดลการวิเคราะห์องค์ประกอบแบบไม่ใช่เส้นตรง โดยสร้างจากการรวมพื้นฐานของการวิเคราะห์องค์ประกอบ CTT และ IRT เข้าด้วยกัน ซึ่งปัจจุบันจากการศึกษาของ Reckase (1997) ได้ระบุว่าขอบเขตของ exploratory

factor analysis จะนิยมใช้กับข้อมูลแบบแบ่ง 2 คล้ายกับโมเดลของ IRT จากการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องกับ exploratory MIRT model พบว่ามีความสัมพันธ์กันในระดับสูงกับ การวิเคราะห์องค์ประกอบโดยศักยภาพของผู้สอบพิจารณาจากการตอบข้อสอบ คือ การให้น้ำหนักรวมกันภายใต้มิติของคุณลักษณะที่ต้องการศึกษา

ปัจจุบันนี้โมเดลของ MIRT ส่วนใหญ่จะช่วยในการพิจารณาด้านโครงสร้างของการพัฒนาแบบสอบ นั่นคือการพิจารณาโครงสร้างที่ต้องการวัดในข้อสอบจำเป็นที่จะต้องอาศัยหลักการของโมเดล IRT เพื่อให้สอดคล้องกับข้อมูลโดยจำนวนของคุณลักษณะที่ต้องการวัดจะต้องมีความสอดคล้องกับข้อมูลเชิงประจักษ์ และเป็นส่วนย่อยของมิติ ต่างๆ เป็นการสร้างตามหลักการหรือทฤษฎีในชุดข้อสอบ ผลที่ได้จากการประมาณค่าพารามิเตอร์ของข้อสอบจะเป็นประโยชน์ในการเลือกข้อสอบเพื่อให้เหมาะสมกับคุณลักษณะแฝง นั่นคือ แม้ว่าลักษณะที่วัดจะมีความเกี่ยวข้องกันบางส่วนก็ตาม แต่สามารถเลือกศึกษาเพียงคุณลักษณะใดลักษณะหนึ่งได้

อย่างไรก็ตามเราสามารถประมาณค่าพารามิเตอร์ของผู้สอบที่มีหลายคุณลักษณะแฝงไปพร้อมๆ กันได้ในครั้งเดียวโดยใช้โมเดล MIRT ซึ่งข้อมูลที่ได้จากการตอบของผู้สอบถือว่าเป็นข้อมูลที่สำคัญในการวัดตั้งแต่ 2 คุณลักษณะขึ้นไป โดยในที่นี้ได้จำแนก exploratory MIRT model ออกเป็น ประเภทย่อย ได้ 2 ประเภทคือ 1) multidimensional logistic model และ 2) multidimensional normal ogive model ซึ่งมีรายละเอียดดังนี้

(1) โมเดลโลจิสติกแบบพหุมิติ (multidimensional logistic model)

โมเดลการตอบสนองข้อสอบ พหุมิติที่มี 3 พารา มิเตอร์ : โมเดลโลจิส (multidimensional three – parameter logistic model)

Embretson และ Reise (2000 อ้างถึงใน ศิริชัย กาญจนวาสี , 2550) ได้ขยายโมเดลโลจิสสามพารามิเตอร์แบบเอกมิติ เป็นโมเดลโลจิสสามพารามิเตอร์แบบหลายมิติ

$$P_i(\underset{\sim}{\theta}_p) = c_i + \frac{(1-c_i)}{1+e^{-(\sum_m a_{im}\theta_{pm}+\delta_i)}} \qquad (22)$$

เมื่อ $p_i(\underset{\sim p}{\theta})$ = ความน่าจะเป็นที่ผู้ตอบซึ่งมีความสามารถ θ_{pm} ($m \geq 2$)

จะตอบข้อสอบ i ได้ถูกต้อง

สมมติในกรณีที่ผู้สอบมีคุณลักษณะแฝง 2 องค์ประกอบ (m=2) ได้แก่ θ_{p1} และ θ_{p2} ความน่าจะเป็นของการตอบข้อสอบข้อ i ได้ถูกต้องของผู้สอบ p $[p_i(\underset{\sim p}{\theta})]$ ขึ้นอยู่กับคุณลักษณะแฝงของอำนาจจำแนกของข้อสอบข้อ i บนมิติ θ_1 และ θ_2 ได้แก่ a_{i1} และ a_{i2}

เมื่อทำการพล๊อตความน่าจะเป็นของการตอบข้อสอบข้อ i ได้ถูกต้องบนภาพแกน 3 มิติ จะทำให้เห็นถึงแบบแผนความสัมพันธ์ของ θ_1 และ θ_2 ต่อ $p_i(\underset{\sim p}{\theta})$ ซึ่งสามารถใช้เปรียบเทียบความสำคัญระหว่าง θ_1 และ θ_2 ได้ ถ้า a_{im} มีค่าสูงบน θ_m มิติใดก็ตาม แสดงว่า θ_m นั้นมีอิทธิพลต่อ $p_i(\underset{\sim p}{\theta})$ สูงกว่ามิติอื่น

(2) โมเดลปกติสะสมแบบพหุมิติ (multidimensional normal ogive model)

โมเดลการตอบสนองข้อสอบ พหุมิติที่มี 3 พารามิเตอร์ : โมเดลโค้งปกติสะสม (multidimensional three – parameter normal ogive model)

Bock, Gibbons และ Muraki (1988 อ้างถึงใน ศิริชัย กาญจนวาสี , 2550) ได้นำแนวคิดของการวิเคราะห์องค์ประกอบชนิดใช้สาร สนเทศทั้งหมด (factor analysis with full information) มาประยุกต์ โดยขยายโมเดลโค้งปกติสะสมสามพารามิเตอร์แบบเอกมิติ เป็นโมเดลโค้งปกติสะสม 3 พารามิเตอร์แบบหลายมิติ

กำหนด Z_{ip} เป็นศักยภาพของผู้สอบ p ในการตอบข้อสอบข้อ i โดย Z_{ip} เป็นค่าผลรวมถ่วงน้ำหนักของคุณลักษณะแฝง m องค์ประกอบตามคุณลักษณะของข้อสอบ i

$$Z_{ip} = \sum_m a_{im}\theta_{pm} + \delta_i$$

โมเดลโค้งปกติสะสมสามพารามิเตอร์แบบหลายมิติ มีสมการเป็นดังนี้

$$p_i(\underset{\sim p}{\theta}) = c_i + (1 - c_i)\int_{-Z_{ip}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-Z^2/2}\, dz \qquad (23)$$

โดย λ_{im} = น้ำหนักองค์ประกอบของข้อ i บนมิติ m

$$= \frac{a_{im}}{g_i}$$

$$b_i = \frac{\delta_i}{g_i}$$

และ $g_i = \sqrt{1 + \sum_m a_{im}^2}$

สำหรับโมเดลโค้งปกติสะสมสามพารามิเตอร์ค่า c_i สามารถกำหนดให้เป็นค่าอิสระ หรือคงที่ของทุกข้อหรือคงที่สำหรับชุดของชุดของข้อสอบย่อยก็ได้

2.2.2 **โมเดลการตอบสนองข้อสอบแบบพหุมิติเชิงยืนยัน** (confirmatory multidimensional IRT model)

สำหรับโมเดลการตอบสนองข้อสอบแบบพหุมิติเชิงยืนยัน พบว่าข้อสอบจะประกอบไปด้วยการออกแบบตามโครงสร้างและเป็นโมเดลในรูปฟังก์ชันเชิงคณิตศาสตร์ที่มีการเชื่อมโยงกับคุณลักษณะที่มีความเฉพาะเจาะจงในหลายลักษณะ ซึ่งโมเดลนี้จะมีความเหมาะสมสำหรับข้อสอบหรือชิ้นงานในลักษณะของทักษะการปฏิบัติ ที่มีการกำหนดหรือสร้างทฤษฎีไว้ล่วงหน้า ซึ่งจะได้นำเสนอตัวอย่างในลำดับต่อไป ซึ่ง Embretson & Reise (2000) ได้แบ่งออกเป็นโมเดลย่อยทั้งหมด 4 โมเดล คือ 1) models for noncompensatory dimensional, 2) models for learning and change, 3) models with specified trait level structures และ 4) models for distinct classes of persons โดยมีรายละเอียดดังต่อไปนี้

(1) **โมเดลสำหรับการวัดมิติที่ไม่สามารถทดแทนได้** (models for noncompensatory dimensional)

ในปี ค.ศ. 1980 Whitely ได้เสนอโมเดล multicomponent latent trait model (MLTM) เพื่อวัดกระบวนการในหลายด้าน ภายใต้ข้อสอบที่มีลักษณะของ การพิสูจน์หรือแก้ปัญหา โดย MLTM จะมีการรวมโมเดลทางคณิตศาสตร์ของกระบวนการในการ ตอบข้อสอบร่วมกับโมเดล IRT ซึ่งศึกษาทั้งในส่วนของ Trait Level และการประมาณค่าความยากของข้อสอบอยู่ภายใต้ข้อตกลงเบื้องต้นความถูกต้องในการประมาณค่าพารามิเตอร์ ขึ้นอยู่กับการป ระมาณค่าองค์ประกอบทั้งหลาย ถ้าองค์ประกอบไม่สามารถประมาณค่าได้ ข้อสอบจะไม่สามารถประมาณค่าได้เช่นกัน ดังนั้นความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องของบุคคลที่ s สำหรับข้อสอบข้อที่ i โดยรวมทุกมิติ แทนด้วย X_{isT} เกิดจากการรวมของความน่าจะเป็นในการตอบข้อ สอบได้ถูกต้องของบุคคลที่ s สำหรับข้อสอบข้อที่ i ในมิติที่ m แทนด้วย X_{ism} ซึ่งโมเดล MLTM สามารถแสดงได้ดังสมการ

$$P\left(x_{isT}=1\middle|\underline{\theta}_s,\underline{\beta}_i\right)=\prod_m \frac{\exp\left(\theta_{sm}-\beta_{im}\right)}{1+\exp\left(\theta_{sm}-\beta_{im}\right)} \qquad (24)$$

เมื่อ $\underline{\theta}_s$ แทน ระดับความสามารถของผู้สอบที่ j ของ M ส่วนประกอบ (M-component)

$\underline{\beta}_s$ แทน ค่าความยากของข้อสอบข้อที่ i ของ M ส่วนประกอบ (M-component)

θ_{sm} แทน ระดับความสามารถของผู้สอบที่ j ของส่วนประกอบที่ m

β_{im} แทน ค่าความยากของข้อสอบข้อที่ i ของส่วนประกอบที่ m

เป็นที่น่าสังเกตว่าสมการทางขวามือ คือโมเดลแบบราสช์ ซึ่งเป็นความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องขององค์ประกอบที่ m ซึ่งผลคูณความน่าจะเป็นในการตอบข้อสอบได้ถูกต้อง ของแต่ละองค์ประกอบคือความน่าจะเป็นทั้งหมดในการตอบข้ อสอบได้ถูกต้องสำหรับข้อสอบข้อที่ i

เมื่อพิจารณา General component latent trait model (GLTM) ซึ่งเป็นแนวคิดของ Embretson (1984) เป็นโมเดลที่มีการรวมค่าความยากของข้อสอบ มีการถ่วงน้ำหนักโดยรวมองค์ประกอบต่างๆ ของสิ่งเร้าเข้าด้วยกัน แทนด้วยสัญลักษณ์ q_{ikm} โมเด ลดังกล่าวแสดงได้ดังสมการ

$$P\left(x_{isT}=1\middle|\underline{\theta}_s,\underline{\beta}_i\right)=\prod_m \frac{\exp\left(\theta_{sm}-\sum_m \tau_{km}q_{ikm}\right)}{1+\exp\left(\theta_{sm}-\sum_m \tau_{km}q_{ikm}\right)} \qquad (25)$$

เมื่อ τ_{ikm} แทน น้ำหนักองค์ประกอบของสิ่งเร้าที่ k ในส่วนประกอบที่ m

q_{ikm} แทน คะแนนองค์ประกอบของสิ่งเร้าที่ k ในส่วนประกอบที่ m สำหรับข้อสอบข้อที่ i

โดยทั่วไป โมเดลMLTM และ GLTM มีการกำหนดส่วนประกอบของการตอบเหมือนกับการตอบข้อสอบโดยรวมของแต่ละข้อ เพื่อใช้ในการประมาณค่าพารามิเตอร์ของส่วนประกอบการตอบข้อสอบ ซึ่งเป็นสิ่งที่สามารถสังเกตได้ เช่น การดำเนินการสอบของแบบส อบย่อย หรือชิ้นงานย่อย (subtasks) โดยพิจารณาจากการตอบข้อสอบโดยรวมของแต่ละข้อ อย่างไรก็ตามในปัจจุบันพบว่า Marris (1995) ได้พัฒนาการประมาณค่าส่วนประกอบโดยการนำโมเดลGLTM มาประยุกต์ใช้เพื่อให้เป็นที่ยอมรับกันโดยทั่วไป โดยไม่จำเป็นต้องมีชิ้นงานย่อย การวัดโดยใช้ข้อสอบหรือใช้ชิ้นงาน ถือว่าทำให้ได้โมเดลทางคณิตศาสตร์ที่มีความเหมาะสมเพื่อใช้ในการทำนายค่าพารามิเตอร์ความยากของข้อสอบหรือค่าคงที่ เพื่อใช้ในการประมาณค่าพารามิเตอร์ขององค์ประกอบ

(2) โมเดลสำหรับการวัดการเรียนรู้และการเปลี่ยนแปลง (models for learning and change)

ในปี ค .ศ. 1991 Embretson ได้พัฒนาโมเดลที่เรียกว่า Multidimensional Rasch Model for Learning and Change (MRCML) เพื่อใช้ในการแก้ปัญหาทางด้านการวัดทางจิตวิทยาเพื่อการวัดความเปลี่ยนแปลง ซึ่งการออกแบบที่เหมาะสมสำหรับโมเดล MRCML คือผู้สอบจะถูกสังเกตภา ยใต้เงื่อนไขหรือสถานการณ์ตั้งแต่ 2 เงื่อนไขขึ้นไป เช่น สถานการณ์การทดสอบ Pre-Test และ Post-Test เป็นต้น ในการพิจารณาทักษะการปฏิบัติ โดยมีการแยกเงื่อนไขหรือสถานการณ์ออกจากกัน และนำเอากระบวนการทางโครงสร้างของ Wiener มาใช้ใน

การระบุมิติความสามารถ แทนด้วยสัญลักษณ์ θ_m โดยมีเงื่อนไขเฉพาะ c_k ในที่นี้ผู้วิจัยได้นำเสนอการออกแบบ สำหรับ 4 เงื่อนไขของการวัด ตามโครงสร้างที่กำหนดไว้

การออกแบบโครงสร้างที่ต้องการวัดจะยึดโมเดลราสช์ ซึ่งสามารถแสดงได้ดังสมการ

$$P\left(x_{i(k)s}=1|\underline{\theta}_s,\beta_i\right)=\frac{\exp\left(\sum_m \theta_{sm}+\beta_i\right)}{1+\exp\left(\sum_m^k \theta_{sm}-\beta_i\right)} \qquad (26)$$

เมื่อ θ_{sm} แทน ระดับความสามารถของผู้สอบที่ j ของลักษณะ (Trait) ที่ m

β_i แทน ค่าพารามิเตอร์ความยากของข้อสอบข้อที่ i

จากสมการที่ 26 เห็นได้ว่า สัญลักษณ์ $\sum_m^k$ เป็นการควบคุมเพื่อให้เกิดความเหมาะสม สำหรับ $\theta_2,\ldots,\theta_M$ ซึ่งเป็น trait level ที่ดัดแปลงขึ้นใหม่ภายใต้เงื่อนไขที่ k ซึ่งการนำโมเดล MRCML มาใช้ในการประมาณค่าสำหรับโมเดลที่มีลักษณะพหุมิติ จะเกิดความคลา ดเคลื่อนมาตรฐานทั้งความสามารถตั้งต้นและ trait level ที่ได้ดัดแปลงขึ้นใหม่

ภายใต้เงื่อนไขที่มีความชัดเจน มีความเป็นไปได้ว่า การแผ่ขยายแนวคิดของโมเดล UIRT สำหรับใช้ในการวัดการเปลี่ยนแปลงในมิติที่มุ่งวัด เช่น Fischer (1997) ได้ขยายโมเดล linear logistic latent trait model (LLTM) ไปใช้ในกรณีแบบพหุมิติโดยการระบุเทคนิคของผู้สอบที่(technical persons) โดยผู้สอบกลายเป็นเทคนิคอย่างหนึ่งที่มีความแตกต่างกันตั้งแต่ 2 คุณลักษณะขึ้นไปในแต่ละเงื่อนไข อย่างไรก็ตาม LLTM ไม่ใช่โมเดลของ MIRT และค่าความคลาดเคลื่อนมาตรฐานในการเปลี่ยนแปลงของ trait level สำหรับแต่ละบุคคลไม่ได้มีการเปลี่ยนแปลงเป็นปกติ

(3) โมเดลในการระบุโครงสร้างของระดับคุณลักษณะ (models with specified trait level structures)

เมื่อแบบสอบที่นำมาใช้อยู่ภายใต้สถานการณ์หรือเงื่อนไขที่มีความหลากหลาย จะมีความเกี่ยวข้ องกันอย่างเป็นระบบกับ trait level โดยการออกแบบโครงสร้างมีความเกี่ยวข้องภายในกับโครงสร้าง trait level ตามโมเดล IRT ในการระบุลักษณะเงื่อนไขที่มีลักษณะเฉพาะ ซึ่งจากมุมมองต่างๆ ของการออกแบบโครงสร้างปกติ จะมีการกำหนด trait level ออกเป็นส่วนย่อยๆ เพื่อ นำผลที่ได้มาเปรียบเทียบกับเงื่อนไขโดยรวมเพื่อวัดทักษะการปฏิบัติ โดยปัจจุบันพบว่า โมเดล IRT หลายๆ โมเดลมีจุดมุ่งหมายของการออกแบบโครงสร้างที่ต่างกันไป เช่น Embretson (1995, 1997) ได้นำเสนอ Structured Latent Traits Model (SLTM) โดยมีจุดมุ่งหมายในกา รออกแบบโครงสร้างของความสามารถด้านต่างๆ ไม่เฉพาะแต่ค่าอำนาจจำแนกของข้อสอบ แต่รวมถึงความสัมพันธ์ภายในของข้อสอบภายใต้เงื่อนไขการวัดที่แตกต่างกัน

ในปีเดียวกัน Wang, Wilson & Adams ได้เสนอโมเดลเชิงโครงสร้างในแต่ละมิติที่คล้ายคลึงกับโมเดลราสช์ ซึ่งเป็นโมเดลที่สามารถอ้างอิงไปยังสถานการณ์โดยทั่วไปและสามารถนำไปใช้ร่วมกับการออกแบบโครงสร้างอื่นๆ เช่น ร่วมกับโมเดล SLTM นอกจากนี้ Wang, Wilson & Adams (1995) ได้เสนอโปรแกรมทางคอมพิวเตอร์ MATS มาใช้ในการประมาณค่าพารามิเตอร์ ซึ่งคล้ายกับการศึกษาของ DiBello, Stout & Roussos (1995) ที่ได้เสนอ unified model เป็นโมเดลที่มีการรวมความสามารถและการออกแบบโครงสร้างของข้อสอบเข้าด้วยกัน ถือได้ว่าเป็นแนวคิดที่นิยมใช้โดยทั่วไป นอกจากนี้พบว่าไม่มีผลกระทบสำหรับข้อจำกัดของการประยุกต์ใช้โมเดลราสช์ที่กำหนดให้ค่าอำนาจจำแนกของข้อสอบเป็นค่าคงที่

การแสดงโมเดลการออกแบบโครงสร้าง จะพิจารณาจากโมเดล SLTM ซึ่งเป็นส่วนประกอบหนึ่งของสมการ multidimensional logistic model สามารถแสดงได้ดังสมการ

$$P\left(x_{i(k)s}=1\middle|\underline{\theta}_s,\beta_i,\lambda_{i(k)m}\right)=\frac{\exp\left(\sum_m \lambda_{i(k)m}\theta_{sm}+\sum_k \beta_{ik}\right)}{1+\exp\left(\sum_m^k \lambda_{i(k)m}\theta_{sm}-\sum_k \beta_{ik}\right)} \qquad (27)$$

เมื่อ θ_{sm} แทน ระดับความสามารถของผู้สอบที่ s ของลักษณะ (Trait) ที่ m

β_{ik} แทน การแจกแจงค่าพารามิเตอร์ความยากของเงื่อนไขที่ k สำหรับข้อสอบที่ i

$\lambda_{i(k)m}$ แทน ค่าน้ำหนักที่มีความเฉพาะเจาะจงของลักษณะ (Trait) ที่ m ของข้อสอบข้อที่ i ภายใต้เงื่อนไขที่ k

ความสำคัญของโมเดล SLTM จะแตกต่างจาก multidimensional logistic model ในสมการที่ 27 โดยการให้น้ำหนัก จะมุ่งไปที่การกำหนดค่าเป็น 0 ดังนั้น โมเดล SLTM จึงคล้ายกับ confirmatory factor analysis model มากกว่า exploratory factor analysis model เช่น การออกแบบ MRMLC เป็นกรณีเฉพาะของ โมเดล SLTM แม้ว่าโมเดลดังกล่าวจะคล้ายคลึงกับรูปแบบโมเดล 2 พารามิเตอร์ ซึ่งจะแตกต่างจาก exploratory multidimensional IRT model โดย SLTM กำหนดให้ค่าอำนาจจำแนกข้อสอบเป็นค่าคงที่ค่าคงที่ของค่าอำนาจจำแนกข้อสอบจะมี 2 ชนิด สำหรับการออกแบบ MRMLC โดยรูปแบบแรกเป็นค่าที่กำหนดให้เท่ากับ 0 เมื่อเป็นค่าอำนาจจำแนกในมิติเดียวกัน ส่วนรูปแบบที่สองจะมีค่าไม่เท่ากับ 0 เมื่อศึกษาข้ามมิติ

(4) โมเดลสำหรับการจำแนกกลุ่มบุคคล (models for distinct classes of persons)

โมเดล IRT เป็นโมเดลที่ต้องการศึกษากลุ่มคุณลักษณะแฝงที่ไม่สามารถสังเกตได้โดยตรง จากการศึกษาตัวแปรภายนอก โดยการแบ่งกลุ่มของลักษณะต่างๆ จะเป็นกระบวนการอย่างเป็นระบบที่แตกต่างกันของรูปแบบการตอบข้อสอบ ซึ่งโมเดลจะประกอบไปด้วย

class และ trait parameter เพื่อใช้ในการทำนายลักษณะของการตอบข้อสอบ แม้ว่าเทคนิควิธีของโมเดล IRT จะเป็นการศึกษาเพียง 1 Trait Level ของผู้สอบแต่ละคน แต่สามารถนำมาใช้ในการจำแนกเชิงพหุมิติได้ เช่น โมเดล SALTUS ที่พัฒนาโดย Wilson (1985) มีความเหมาะสมกับการจำแนกความแตกต่างของความรอบรู้กับไม่รอบรู้ การ เพิ่มขึ้นของความน่าจะเป็นในการตอบข้อสอบได้ถูกมักจะได้รับอิทธิพล มาจากขั้นของการวัดที่มีความเฉพาะเจาะจง แสดงได้ดังสมการ

$$P\left(x_{is}=1|\underline{\theta}_s,\beta_i\right)=\prod_m \frac{\exp\left(\theta_s-\beta_i+\xi_{h(s)k(i)}\right)}{1+\exp\left(\theta_s-\beta_i+\xi_{h(s)k(i)}\right)} \qquad (28)$$

เมื่อ $\xi_{h(s)k(i)}$ แทน การเพิ่มขึ้นของโอกาสในการตอบข้อสอบได้ถูกสำหรับข้อสอบในชนิดที่ k ในขั้นที่ h

สิ่งสำคัญประการหนึ่งของการเลือกใช้โมเดล SALTUS คือ ICC จะขึ้นอยู่กับระดับขั้นของการพัฒนา ซึ่งต่อ Rost (1990) ได้นำเสนอโมเดล Mixed Population Rasch Model (MIRA) โดยเหมาะกับข้อมูลในการแบ่งกลุ่ มความแตกต่างของบุคคลอย่างเป็นระบบ ซึ่งเป็นพื้นฐานของการตอบข้อสอบ การมีโครงสร้างหรือลำดับขั้นการเรียนรู้ที่ต่างกัน จะสามารถจำแนกรูปแบบของค่าความยากของข้อสอบในแต่ละกลุ่ม ถือได้ว่าเป็นโมเดลที่ใช้ในการประมาณค่า Trait Level และการจำแนกกลุ่มผู้สอบในแต่ละบุคคล

2.3 โปรแกรมการวิเคราะห์ข้อมูลสำหรับการทดสอบแบบพหุมิติ

สำหรับโปรแกรมการวิเคราะห์ข้อมูลสำหรับการทดสอบแบบพหุมิตินั้น มีผู้พัฒนาขึ้นจำนวนมาก และโปรแกรมเหล่านั้นก็มีวิธีการพัฒนาที่หลากหลายโดยได้รวมคุณสมบัติต่างๆไว้มากมาย ซึ่งในที่นี้จะกล่าวถึงความสามารถในการวิเคร าะห์ข้อมูลสำหรับการทดสอบแบบพหุมิติเท่านั้น อย่างไรก็ตามการพัฒนาโปรแกรมการวิเคราะห์นั้นมีการพัฒนาขึ้นใหม่อย่างรวดเร็ว และมีคุณสมบัติที่ดีขึ้น โดยจะขอยกตัวอย่างบางโปรแกรมที่ได้รับความนิยม อาทิ TESTFACT (Bock, Gibons, Schilling, Muraki, Wilson, and Wood, 2003) NOHARM (Fraser, 1998) ConQuest (Wu, Adams, and Wilson, 1997) และ BMIRT (Yao, 2003) ซึ่งแต่ละโปรแกรมก็มีคุณสมบัติและข้อจำกัดในการวิเคราะห์ที่แตกต่างกันไป ซึ่ง Reckase (2009) ได้สรุปไว้ดังนี้

TESTFACT ผู้พัฒนาโปรแกรมได้กล่าวว่า “วิธีการนี้ถือเป็นวิธีการให ม่ของการวิเคราะห์ปัจจัยของข้อสอบที่อยู่ภายใต้ทฤษฎีการตอบสนองข้อสอบ ” ซึ่งวิธีการใหม่ที่กล่าวนั้นหมายถึงการประมาณค่า พารามิเตอร์ข้อสอบ และพารามิเตอร์ผู้สอบ สำหรับการทดสอบแบบพหุมิติจาก 2 พารามิเตอร์ ให้สามารถวิเคราะห์ได้ทั้ง 3 พารามิเตอร์ (three-parameter normal ogive model) แต่ทั้งนี้โปรแกรม TESTFACT ไม่สามารถประมาณค่าโอกาสการเดาข้อสอบได้ ผู้พัฒนาโปรแกรม

ได้เสนอว่าให้ใช้โปรแกรมอื่นมาช่วยวิเคราะห์ เช่น BILOG-MG โดยโปรแกรมนี้จะประมาณค่าพารามิเตอร์ข้อสอบโดยใช้วิธีความแปรปรวนของ maximum likelihood ซึ่งเรียกว่า “marginal maximum likelihood” สำหรับการประมาณค่าพารามิเตอร์ผู้สอบหรือค่าความสามารถของผู้สอบจะใช้วิธีการของ Bayesian โดยมีข้อจำกัดคือ วิเคราะห์ได้เฉพาะข้อสอบที่ให้คะแนนแบบ dichotomous และมีข้อจำกัดเรื่องจำนวนมิติของข้อสอบ

NOHARM (Nomal-Ogive Harmonic Analysis Robust method) โปรแกรมนี้ประมาณค่าพารามิเตอร์ข้อสอบในรูปแบบ normal ogive model เช่นเดียวกับการวิเคราะห์โดย TESTFACT โดยสามารถวิเคราะห์ข้อสอบแบบพหุมิติได้มากถึง 50 มิติ อย่างไรก็ตามโปรแกรมนี้จะไม่สามารถประมาณค่าพารามิเตอร์ผู้สอบ และวิเคราะห์ได้เฉพาะข้อสอบที่ให้คะแนนแบบ dichotomous

ConQuest โปรแกรมนี้มีแนวคิดอยู่ภายใต้แนวคิด MIRT Rasch model ซึ่งจะศึกษาเพียง one-parameter นั่นคือค่าความยาก (b) โดยโปรแกรมนี้สามารถประมาณค่าพารามิเตอร์ข้อสอบและพารามิเตอร์ผู้สอบได้ ลักษณะที่สำคัญสำหรับโมเดลนี้คือ ผู้ใช้จะต้องระบุ เมตริกซ์การให้คะแนน (A) และออกแบบเมตริกซ์ (B) เมื่อต้องการใช้โปรแกรมนี้ประมาณค่าพารามิเตอร์

BMIRT (Bayesian Multivariate Item Respond Model) โปรแกรมนี้จะใช้วิธีการของ Markov chain Monte Carlo (MCMC) ซึ่งสามารถวิเคราะห์ข้อสอบที่ให้คะแนนทั้งแบบ dichotomous และ polytomous โดยสามารถประมาณค่าพารามิเตอร์ข้อสอบ พารามิเตอร์ผู้สอบ และพารามิเตอร์การกระจายของประชากร

กล่าวโดยสรุป โมเดล MIRT สามารถแบ่งได้เป็น 2 ประเภท คือ exploratory MIRT model และ confirmatory MIRT model ซึ่งมีลักษณะคล้ายคลึงกับการวิเคราะห์องค์ประกอบ โดย exploratory MIRT model เกี่ยวข้องกับการประมาณค่าพารามิเตอร์ข้อสอบและผู้สอบที่มากกว่า 1 มิติ เพื่อให้โมเดลมีความสอดคล้องกับข้อมูล โดยไม่ได้กำหนดจำนวนองค์ประกอบล่วงหน้า ประกอบด้วย 2 ประเด็นย่อย คือ 1) multidimensional logistic model และ 2) multidimensional normal ogive Model โดยโมเดลย่อยทั้ง 2 มีสูตรในการประมาณค่าคล้ายคลึงกับ logistic model และ normal ogive model ในโมเดล UIRT แต่จะต่างกันในจำนวนมิติที่ศึกษา ซึ่งตรงข้ามกับ confirmatory MIRT Model เกี่ยวข้องกับการประมาณค่าพารามิเตอร์ในมิติที่มีความเฉพาะเจาะจง โดยคล้ายกับการวิเคราะห์ องค์ประกอบที่เกี่ยวข้องกับการระบุความสัมพันธ์ของข้อสอบไปยังมิติต่างๆ โดยพบว่า ข้อสอบ จะประกอบไปด้วยการออกแบบตามโครงสร้างและเป็นโมเดลในรูปฟังก์ชันเชิงคณิตศาสตร์ที่มีการเชื่อมโยงกับคุณ ลักษณะที่มีความเฉพาะเจาะจงในหลายลักษณะ โมเดลนี้เหมาะสมสำหรับข้อสอบหรือชิ้นงานของทักษะการปฏิบัติ

ที่ได้มีกำหนดหรือสร้างทฤษฎีไว้ล่วงหน้า ซึ่ง Embretson & Reise (2000) แบ่งออกเป็นโมเดลย่อยทั้งหมด 4 โมเดล คือ 1) models for noncompensatory dimensional, 2) models for learning and change, 3) models with specified trait level structures และ 4) models for distinct classes of persons โดย confirmatory MIRT model เหมาะสำหรับการนำไปประยุกต์ใช้ได้หลากหลาย เช่น แก้ปัญหาทางด้านการวัดทางจิตวิทยาในการวัดความ เปลี่ยนแปลง (measurement of change) ออกแบบโครงสร้างของความสามารถด้านต่างๆ เป็นต้น(พัชรี จันทร์เพ็ง , 2550) สำหรับโปรแกรมการวิเคราะห์การทดสอบแบบพหุมิตินั้นมีคุณสมบัติที่เหมือนและแตกต่างกันไป จึงควรพิจารณาข้อจำกัดของแต่ละโปรแกรมก่อนที่จะนำมาใช้ประมาณค่าพารามิเตอร์ข้อสอบ หรือผู้สอบ เพื่อให้เกิดประสิทธิภาพ และมีความแม่นยำมากที่สุด

ตอนที่ 3 มโนทัศน์เกี่ยวกับ การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบพหุมิติ (multidimensional computerized adaptive testing)

การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบพหุมิติ (multidimensional computerized adaptive testing) เป็นการผสมผสานแนวคิดของ MIRT (multidimensional Item response theory) กับ CAT (computerized adaptive testing) ซึ่งเป็นการทดสอบที่มีประสิทธิภาพการวัดสูง มีความแม่นยำ และลดจำนวนข้อสอบ โดยการนำเสนอได้แบ่งหัวข้อออกเป็น ความเป็นมาและหลักการ ประเภทของ multidimensionality โมเดลการตอบสนองข้อสอบแบบพหุมิติที่ใช้ในการทดสอบแบบปรับเหมาะ การคัดเลือกข้อสอบและการประมาณค่าความสามารถในการทดสอบแบบ พหุมิติ การควบคุมการใช้ข้อสอบซ้ำ และการสร้างความสมดุลของเนื้อหา ซึ่งมีรายละเอียดดังนี้

3.1 ความเป็นมาและหลักก ารของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบพหุมิติ (multidimensional computerized adaptive testing)

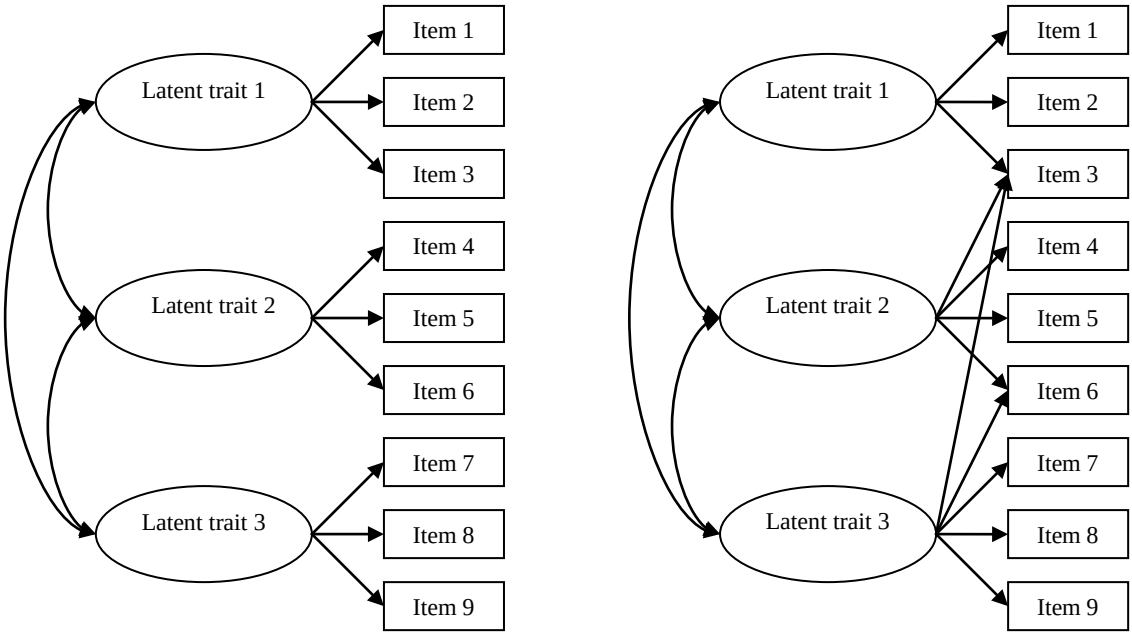
เป็นเวลามากกว่า 10 ปี ที่การทดสอบแบบ ปรับเหมาะด้วยคอมพิวเตอร์ (computerized adaptive testing) หรือที่เรียกว่า (CAT) ได้รับความนิยม และกล่าวถึง ในความสำเร็จ ของวิธีการทดสอบนี้ ซึ่งการทดสอบแบบปรับเหมาะนั้นมีความสามารถมากกว่าการทดสอบแบบดั้งเดิมที่เขียนลงบนกระดาษคำตอบ (paper-and-pencil tests) ซึ่งเป็นการเพิ่มความแม่นยำในการวัด ลดระยะเวลาในการทดสอบ วิธีใช้ได้มาตรฐานและมีความยืดหยุ่นในการจัดตารางการสอบของผู้ สอบ การทดสอบแบบปรับเหมาะส่วนใหญ่จะใช้การเลือกข้อคำถามและ

การคำนวณการให้คะแนนโดยใช้พื้นฐานของทฤษฎีการตอบสนองข้อสอบ (item response theory: IRT) ซึ่งในขณะนี้เทคนิคดังกล่าวใช้กับการทดสอบแบบปรับเหมาะภายใต้ข้อตกลงเบื้องต้นที่การทดสอบจะต้องเป็นรูปแบบเอกมิติ(unidimensionality) (Segall, 1996) และเนื่องจากการตอบสนองข้อสอบแบบเอกมิติ ในบางครั้งจะไม่เหมาะสมกับการทดสอบจริง ดังนั้นจึงมีการพัฒนาแนวคิดเกี่ยวกับ ทฤษฎีการตอบสนองข้อสอบ แบบพหุมิติ (multidimensional item response theory: MIRT) ซึ่งมีแนวคิดมาจาก 2 กลุ่มคือ Spearman (1927) and Thurstone (1947) และ Lazarsfeld (1950) ซึ่งในขณะที่ทฤษฎีการตอบสนอง ข้อสอบแบบพหุมิติเริ่มได้รับการยอมรับและการทดสอบแบบปรับเหมาะกำลังได้รับความนิยมอย่างมากในการนำไปปฏิบัติ จึงเกิดการรวมกันของทั้งสองวิธีนี้เรียกว่า การทดสอบปรับเหมาะแบบพหุมิติ (multidimensional adaptive testing: MAT) (Segall, 1996; 2000) และเมื่อนำคอมพิวเตอร์มาใช้ในกระบวนการทดสอบ จึงเรียกการทดสอบนี้ว่า การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบพหุมิติ (Multidimensional Computerized Adaptive Testing: MCAT) ซึ่งวิธีการทดสอบดังกล่าวนี้เป็นการทดสอบที่มีประสิทธิภาพในการวัด สูง ไม่ว่าจะเป็นความแม่นยำในการวัด การลดความยาวของข้อสอบ และเป็นการวัดที่ตรงกับสภาพจริงของลักษณะข้อสอบ

3.2 ประเภทของ multidimensionality

ประเภทของ multidimensionality แบ่งออกเป็นการทดสอบพหุมิติระหว่างข้อ (between-item multidimensionality) และ การทดสอบพหุมิติภายในข้อ (within-item multidimensionality) (Adam et al, 1997; Wang, Chen & Cheng, 2004; Wang, Wilson & Adams, 1997) โดยที่ข้อสอบจะถูกเรียกว่าเอกมิติ (unidimensional) เมื่อข้อสอบนั้นมุ่งที่จะวัดคุณลักษณะแฝงเพียงคุณลักษณะเดียว และจะเรียกว่าพหุมิติ (multidimensional) เมื่อข้อสอบนั้นมุ่งที่จะวัดคุณลักษณะหลายคุณลักษณะ และการทดสอบนั้นจะเป็นการทดสอบพหุมิติระหว่างข้อ (between-Item multidimensionality) เมื่อข้อสอบแต่ละข้อไม่ได้วัดหลายคุณลักษณะ และการทดสอบนั้นจะเป็นการทดสอบพหุมิติภายในข้อ (within-Item Multidimensionality) เมื่อข้อสอบแต่ละข้อวัดหลายคุณลักษณะ ดังแผนภาพที่ 2.3 แสดงลักษณะของการทดสอบพหุมิติระหว่างข้อ และการทดสอบพหุมิติภายในข้อ สำหรับลูกศรสองทางระหว่างคุณลักษณะแฝงบ่งชี้ความสัมพันธ์ของคุณลักษณะแฝงที่ไม่เป็น 0 และจะเห็นได้ว่า การทดสอบพหุมิติระหว่างข้อเป็นลักษณะพิเศษของการทดสอบพหุมิติภายในข้อ โดยที่ แต่ละข้อถูกกำหนดให้วัดเพียงคุณลักษณะเดียว การทดสอบพหุมิติระหว่างข้อ และ การทดสอบพหุมิติภายในข้อมีลักษณะคล้ายกับโครงสร้างปัจจัย “simple” และ “complex” ในการวิเคราะห์ปัจจัย (factor analysis) ขั้นตอนการทดสอบแบบปรับเหมาะถูก

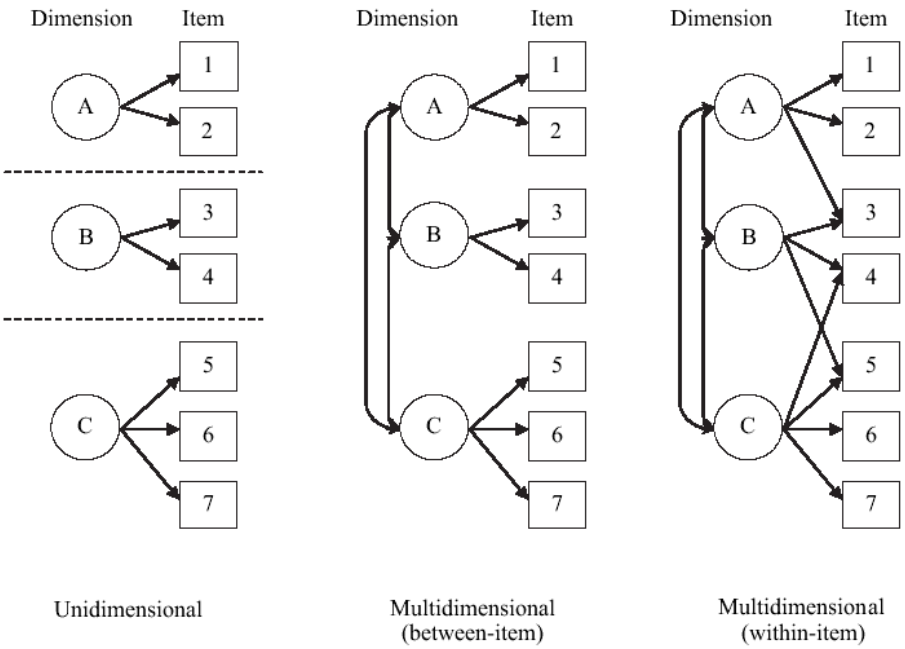
พัฒนามาจากโมเดลการตอบสนองข้อสอบพหุมิติระหว่างข้อ และภายในข้อจะถูกกล่าวถึงในชื่อ between-Item MAT และ Within-Item MAT ซึ่งการศึกษาของ Luecht (1996) และ Segall (1996) เป็นตัวอย่างการศึกษาการทดสอบแบ บปรับเหมาะแบบ between-Item MAT ในลักษณะของการให้คะแนนสองค่า (Wang, W. and Chen, P., 2004) ซึ่ง Li (2006) เรียกรูปแบบนี้ว่าโครงสร้างอย่างง่าย (simple structure)



between-Item Multidimensionality within-Item Multidimensionality

แผนภาพที่ 2.3 ลักษณะของ between-Item multidimensionality และ within-Item multidimensionality

จากแผนภาพ 2.3 จะเห็นว่าลักษณะของ between-item multidimensionality นั้น มีความคล้ายคลึงกับการทดสอบแบบเอกมิติ เพราะข้อสอบแต่ละข้อวัดเพียงคุณลักษณะเดียว อย่างไรก็ตาม ความแตกต่างของ between-item multidimensionality และแบบทดสอบเอกมิติอยู่ที่ความสัมพันธ์ของคุณลักษณะที่นำมาใช้ในการทดสอบ เนื่องจากการทดสอบแบบเอกมิติ คุณลักษณะที่นำมาใช้ในการทดสอบจะไม่มีความสัมพันธ์กัน แต่การทดสอบแบบพหุมิติแบบ between-item multidimensionality คุณลักษณะแต่ละคุณลักษณะจะมีความสัมพันธ์กัน ซึ่งแสดงให้เห็นชัดเจนในแผนภาพที่ 2.4



ที่มา Cheng, Y., and others. (2008)

แผนภาพที่ 2.4 กราฟแสดงวิธีการของ unidimensional และ multidimensional

3.3 โมเดลการตอบสนองข้อสอบแบบพหุมิติที่ใช้ในการทดสอบแบบปรับเหมาะ

Segall (1996) ได้เริ่มต้นโดยการแสดงเซตของคุณลักษณะแฝง p คุณลักษณะด้วยเวคเตอร์ $\theta = \{\theta_1, \theta_2, ..., \theta_p\}$ โดยสมมติว่า แต่ละ p คุณลักษณะมาจากข้อสอบ 1 ข้อหรือมากกว่า Hattie (1981 cited in Segall, 1996) ได้นิยามฟังก์ชันการตอบสนองข้อสอบข้อที่ i โดย

$$P_i(\theta) \equiv P(U_i = 1|\theta) = c_i + \frac{1 - c_i}{1 + \exp\left[-Da_i^{'}(\theta - b_i 1)\right]} \quad (29)$$

โดยที่ U_i = ตัวแปรสุ่มสองค่า ประกอบด้วยการตอบสนองข้อสอบข้อที่ i (U_i = 1 เมื่อตอบข้อสอบข้อที่ i ถูกต้อง และ U_i = 0 เมื่อตอบข้อสอบข้อที่ i ไม่ถูกต้อง)

c_i = โอกาสการเดาข้อที่ i ได้ถูกต้อง

b_i = พารามิเตอร์ความยากของข้อสอบข้อที่ i

a_i = เวคเตอร์ 1 x p ของพารามิเตอร์อำนาจจำแนกของข้อสอบข้อที่ i

1 = เวคเตอร์ 1 x p ของ 1's

D = ค่าคงที่มีค่าเท่ากับ 1.7

โดยที่เลขชี้กำลังของตัวหารในสมการที่ (29) สามารถอธิบายในรูปของสเกลาดังนี้

$$-Da_i'\left(\theta - b_i 1\right) = -D\sum_{k=1}^{p} a_{ki}\left(\theta_k - b_i\right) \qquad (30)$$

Segall (1996) สังเกตว่า สิ่งที่ง่ายขึ้นสำหรับโมเดลนี้เกิดขึ้น เมื่อ p=1 จากด้านขวามือของสมการที่ 30 จะสังเกตเห็นว่าสมการที่ 29 ถูกลดมาเป็น

$$P\left(U_i = 1|\theta\right) = c_i + \frac{1 - c_i}{1 + \exp\left[-Da_i\left(\theta - b_i\right)\right]}$$

สูตรนี้ใช้สำหรับโมเดลโลจิสติก 3 พารามิเตอร์ (3PLM; Birnbaum, 1968)

3.4 องค์ประกอบของการทดสอบปรับเหมาะแบบพหุมิติ

องค์ประกอบของการทดสอบปรับเหมาะแบบพหุมิติ มีขั้นตอนและองค์ประกอบเหมือนกับการทดสอบแบบปรับเหมาะแบบดั้งเดิม (เอกมิติ) ซึ่งมีรายละเอียดดังนี้

3.4.1 คลังข้อสอบสำหรับการทดสอบแบบพหุมิติ

แม้ว่าวิธีดำเนินการของการทดสอบปรับเหมาะจะดีเพียงใดหากคลังข้อสอบที่พัฒนาขึ้นไม่มีคุณภาพ การทดสอบครั้งนี้ก็ถือว่าไม่มีคุณภาพเช่นกัน ดังนั้นการออกแบบและพัฒนาคลังข้อสอบสำหรับการทดสอบแบบปรับเหมาะจึงเป็นสิ่งที่สำคัญที่ควรคำนึงถึง (Reckase, 2009)

จากการศึกษาที่ผ่านมา ไม่มีงานวิจัยใดที่ระบุเกี่ยวกับการพัฒนาคลังข้อสอบสำหรับการทดสอบปรับเหมาะแบบพหุมิติ โดยงานวิจัยที่ศึกษา MCAT เหล่านั้น ได้ศึกษาจากงานวิจัยต่างๆที่ผ่านมา ซึ่งข้อสอบที่ใช้ในคลังข้อสอบจะมาจากการจำลองข้อมูล และนำมาจากข้อสอบที่มีอยู่แล้วในลักษณะ paper-and-pencil ยกตัวอย่างเช่น Veldkamp และ van der Linden (2002) ใช้คลังข้อสอบจำนวน 176 ข้อ จากข้อสอบเดิมที่มีลักษณะ paper-and-pencil ในการทดสอบ ACT Assessment program นอกจากนั้น Lee, Ip และ Fuh (2008) ใช้ข้อสอบจำนวน 480 ข้อ จากการจำลองคลังข้อสอบขึ้นมา โดยแยกอำนาจจำแนกออกเป็น 4 ระดับ ซึ่งผลของการศึกษาเหล่านี้ก็ขึ้นอยู่กับคุณภาพของคลังข้อสอบที่นำมาใช้

3.4.2 การประมาณค่าความสามารถในการทดสอบปรับเหมาะแบบพหุมิติ

การประมาณค่าความสาม ารถในการทดสอบ แบบเอกมิติจะใช้การประมาณค่าแบบ Maximum likelihood (ML) และ Bayesian ซึ่งการทดสอบปรับเหมาะแบบพหุมิติ(MCAT) ก็ใช้วิธีการเดียวกัน(Segall, 2000) ซึ่งมีรายละเอียดของแต่ละวิธีดังนี้

1) maximum likelihood approach

การประมาณค่าความสามารถด้วยวิธี Maximum Likelihood เริ่มต้นจากการระบุฟังก์ชันความเป็นไปได้ ความเป็นไปได้ของเวคเตอร์ของการตอบสนองข้อสอบที่สังเกตได้ u ให้ค่าความสามารถ θ โดย

$$L(u|\theta) \equiv L(u_{v1}, u_{v2}, ...|\theta) = \prod_{i \in v} P_i(\theta)^{ui} Q(\theta)^{1-u_i} \quad (31)$$

โดยที่ $Q_i(\theta) = 1 - P_i(\theta)$ และ v แทนองค์ประกอบของ เวคเตอร์ที่เป็นตัวบ่งชี้ข้อสอบแบบปรับเหมาะ เวคเตอร์มีค่า $\{\hat{\theta}_1, \hat{\theta}_2, ...\hat{\theta}_p\}$ ฟังก์ชันความเป็นไปได้สูงสุด (maximum likelihood: ML)แสดงในสมการที่ 31 เพื่อประมาณค่าความสามารถ θ Tam (1992) ได้เสนอโดยใช้แนวคิดของโมเดล normal ogive เพื่อให้ได้การประมาณค่า ML ในสองมิติของคุณลักษณะแฝง การประมาณค่าความสามารถด้วยวิธี ML เป็นการแก้ปัญหาของเซตของสมการ simultaneous p ค่า แสดงโดย

$$\frac{\partial}{\partial \theta} \ln L(u|\theta) = 0 \quad (32)$$

โดยที่

$$\frac{\partial}{\partial \theta} \ln L(u|\theta) = \begin{bmatrix} \frac{\partial}{\partial \theta_1} \ln L(u|\theta) \\ \frac{\partial}{\partial \theta_2} \ln L(u|\theta) \\ \vdots \\ \frac{\partial}{\partial \theta_p} \ln L(u|\theta) \end{bmatrix}$$

2) Bayesian approach

การประมาณค่าความสามารถแบบ Bayesian สามารถนิยามได้ว่าเป็นค่า mode ของ posterior distribution $f(\theta|u)$ซึ่งเป็นสัดส่วนของโปรดักส์ระหว่าง likelihood และ prior (Segall, 1999)

$$f(\theta|u) \quad \alpha \quad L(u|\theta) f(\theta)$$

สำหรับการประมาณค่าแสดงด้วย $\hat{\theta}$ ซึ่งเป็นค่าของ θ ที่ต้องการสำหรับเซตของสมการ $\partial \ln f(\theta|u) / \partial \theta = 0$ ที่ดำเนินไปพร้อมกัน p ครั้ง โดยที่

$$\frac{\partial}{\partial \theta} \ln f(\theta|u) = D \sum_{i=1}^{n} v_i a_i - \theta$$

และ $$v_i = \frac{[P_i(\theta) - c_i][u_i - P_i(\theta)]}{(1 - c_i) P_i(\theta)}$$

และเนื่องจากเซตของสมการนี้ไม่มีการแก้ปัญหารูปแบบปิด ดังนั้นวิธีการวนซ้ำจึงถูกนำมาใช้ ซึ่งในที่นี้จะใช้วิธีของ Newton-Raphson โดยกำหนดให้ $\theta^{(m)}$ แทนการประมาณลำดับที่ m ของค่า θ ที่สูงที่สุดใน $f(\theta|u)$ ดังนั้นวิธีประมาณค่าที่ดีกว่าคือ ทำให้เป็นรูปทั่วไป โดยให้

$$\theta^{(m+1)} = \theta^{(m)} - \delta^{(m)} \tag{33}$$

โดยที่ $\delta^{(m)}$ คือเวคเตอร์ 1xp

$$\delta^{(m)} = \left[J\left(\theta^{(m)}\right)\right]^{-1} x \frac{\partial}{\partial\theta} \ln f\left(\theta^{(m)}|u\right) \tag{34}$$

เมตริกซ์ J(θ) คือ เมตรกซ์ของอนุพันธ์บางส่วนอันดับสองซึ่งถูกประเมินที่ตำแหน่ง $\theta = \theta^{(m)}$

$$J(\theta) = D^2 \sum a_i a_i' w_i - I$$

โดยที่ $$w_i = \frac{Q_i(\theta)\left[P_i(\theta) - c_i\right]\left[c_i u_i - P_i^2(\theta)\right]}{P_i^2(\theta)(1-c_i)^2}$$

การประมาณค่าจะดำเนินการด้วยวิธีในสมการที่ (33) และ (34) จนกระทั่งค่าลู่เข้าหากัน

3.4.3 วิธีการคัดเลือกข้อสอบในการทดสอบปรับเหมาะแบบพหุมิติ

วิธีการคัดเลือกข้อสอบ จะพิจารณาที่ค่าสารสนเทศสูงสุด โดย Chang และ Ying (1996) ได้แบ่งสารสนเทศของข้อสอบออกเป็น 2 ประเภทคือ Local information และ Global information โดยสารสนเทศที่จัดอยู่ใน Local information คือ Fisher information ซึ่งสารสนเทศจะอยู่รอบๆ θ ที่แท้จริงหรือมีลักษณะเป็นสารสนเทศแบบจุด (The item information at point) ควรใช้ในกรณีที่มี n ขนาดใหญ่ สำหรับสารสนเทศที่จัดอยู่ใน Global information คือ Kullback-Leiber information ซึ่งสารสนเทศจะอยู่ภายนอกขอบเขตของ θ ที่แท้จริงหรือมีลักษณะเป็นเกณฑ์สารสนเทศแบบช่วง (Interval information criterion) เหมาะสำหรับใช้ในกรณีที่มีn ขนาดเล็กซึ่งมีรายละเอียดของวิธีทั้งสอง ดังนี้

1) Fisher Information

สารสนเทศของฟิชเชอร์สำหรับข้อสอบข้อที่ i นิยามดังนี้

$$I_i(\theta) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta'} \log f\left(x_i|\theta\right)|\theta\right] \tag{35}$$

จากสมการที่35 ถ้าใช้โมเดลการตอบสนองข้อสอบแบบพหุมิติ 3 พารามิเตอร์ตามโมเดลในสมการที่ 29 สารสนเทศของฟิชเชอร์ จะได้ดังสมการที่ 36

$$I_i(\theta) = \frac{Q_i(\theta)\left[P_j(\theta) - c_i\right]^2}{P_j(\theta)(1-c_i)} \begin{bmatrix} a_{i1}^2 & a_{i1}a_{i2} & \cdots & a_{i1}a_{ip} \\ a_{i1}a_{i2} & a_{i2}^2 & \cdots & a_{i2}a_{ip} \\ \vdots & & & \\ a_{i1}a_{ip} & a_{i2}a_{ip} & \cdots & a_{ip}^2 \end{bmatrix} \tag{36}$$

ให้ $P_j(\theta)$ แทน ฟังก์ชันการตอบสนองข้อสอบข้อที่ j

$Q_j(\theta)$ แทน 1 - $P_j(\theta)$

สำหรับการทดสอบปรับเหมาะแบบเอกมิติเป็นที่ทราบกันดีว่าความแปรปรวนของ $\hat{\theta}^{mle}$ คือส่วนกลับที่มีความสัมพันธ์กับสารสนเทศของฟิชเชอร์ที่มีความยาวข้อสอบไม่สิ้นสุด ดังนั้นหากต้องการลดความแปรปรวนของ $\hat{\theta}^{mle}$ ให้ต่ำที่สุด จะต้องทำให้ค่าสารสนเทศของฟิชเชอร์ ณ ตำแหน่งความสามารถที่ประมาณค่าได้ปัจจุบันมีค่าสูงสุ ด วิธีการนี้คือ แนวความคิดเกี่ยวกับ maximum Fisher information criterion โดย Lord (1980) สำหรับการขยายแนวคิดไปสู่การทดสอบปรับเหมาะแบบพหุมิติ (MCAT) จำเป็นต้องกำหนดความสัมพันธ์ระหว่าง $\hat{\theta}^{mle}$ และเมตริกซ์สารสนเทศของฟิชเชอร์ (Fisher information matrix) สำหรับการวิเคราะห์หลายตัวแปร (multivariate analysis) Anderson (1984 cited in Wang and Chang, 2011) แสดงให้เห็นว่าตามทีกลุ่มตัวอย่าง (ความยาวข้อสอบ) มีค่าไม่สิ้นสุด ช่วงความเชื่อมั่นรอบๆค่า $\hat{\theta}^{mle}$ จะกลายเป็น ellipsoid และปริมาณของ ellipsoid นี้คือ สัดส่วนของค่าดีเทอร์มิแนนต์ ของส่วนกลับของเมตริกซ์สารสนเทศของฟิชเชอร์ ซึ่งจะถูกประเมินที่ค่าความสามารถที่แท้จริง นั่นคือ $I^{-1}(\theta_0)$ โดยที่ Segall (1996) เป็นคนแรกที่ประยุกต์ข้อสรุปนี้มาใช้ในบริบทของ MCAT และเสนอกฎการคัดเลือก ก ข้อสอบเพื่อให้ค่าดีเทอร์มิแนนต์ของเมตริกซ์สารสนเทศของฟิชเชอร์มีค่าสูงสุด ซึ่งวิธีการนี้สามารถให้ปริมาณความเชื่อมั่นรอบๆ $\hat{\theta}^{mle}$ สูงสุด และเพื่อให้วิธีการนี้เป็นแบบแผน จึงบ่งชี้ให้ข้อสอบที่มีโอเกาสถูกคัดเลือกทั้งหมด แทนด้วย j = 1, 2, …,J โดยที่ S_{k-1} = {j_1, j_2…, j_{k-1}} ของข้อสอบ k-1 ข้อ และ R_k: {1,…, N}/S_{k-1} คือ ข้อสอบที่เหลืออยู่ในคลังข้อสอบ เกณฑ์การคัดเลือกข้อสอบนี้ จะเลือกข้อสอบที่เหมาะสมตามกฎดังต่อไปนี้

$$j \equiv \arg\max_j \left\{ \det\left(I_{S_{k-1}}\left(\hat{\theta}^{mle}\right) + I_{jk}\left(\hat{\theta}^{mle}\right) \right), j \in R_k \right\}$$

โดยที่ $I_{S_{k-1}}(\hat{\theta})$ คือ สารสนเทศแบบสอบ สำหรับข้อสอบ k-1 ข้อ วิธีการนี้ต่อมาได้ใช้คำว่า D-optimality (Mulder and van der Linden, 2009)

Segall (1996) ได้นำเสนอ Bayesian ในรูปแบบของ D-optimality ซึ่งจะประเมินค่าดีเทอร์มิแนนต์ของ posterior variance-covariance matrix ณ ตำแหน่ง posterior mode ของค่าความสามารถ และ จากสมมติฐานของ multivariatr normal prior วิธีการคัดเลือกข้อสอบนี้จะกลายเป็น

$$j \equiv \arg\max_j \left\{ \det\left(I_{S_{k-1}}\left(\hat{\theta}^{MAP}\right) + I_{jk}\left(\hat{\theta}^{MAP}\right) + \sum\nolimits_0^{-1} \right), j \in R_k \right\}$$

โดยที่ $\sum_0$ คือ prior variance-covariance matrix

2) Kullback – Leibler information index

Chang และ Ying (1996) ได้เสนอสารสนเทศใหม่ที่ใช้พื้นฐานของ Kullback – Leibler divergence ซึ่งมีนิยามว่า KL divergence ระหว่างการแจกแจงความน่าจะเป็นของ f และ g ถูกเสนอไว้ดังนี้

$$KL(f\|g) = E_f\left[\log\frac{f(\theta)}{g(\theta)}\right]$$

สำหรับข้อสอบข้อที่ j , KL สามารถคำนวณได้ดังนี้

$$KL_j(\theta_0\|\hat{\theta}) = \sum_{x=0}^{1} f(x_j|\theta_0)\log\left[\frac{f(x_j|\theta_0)}{f(x_j\hat{\theta})}\right]$$

โดยที่ X_j คือ การตอบข้อสอบแบบให้คะแนนสองค่า

$f(x_j|\theta_0)$ คือ ฟังก์ชันการตอบสนองข้อสอบณ ตำแหน่งความสามารถที่แท้จริง θ_0

Chang และ Ying (1996) ได้ประยุกต์ใช้ KL information เพื่อใช้ในการคัดเลือกข้อสอบ โดยได้เสนอ ดัชนีของ KL (KL index: KI) ไว้ดังนี้

$$KI_j(\theta_0) = \int_{\theta_0-\frac{r}{\sqrt{k}}}^{\theta_0+\frac{r}{\sqrt{k}}} KL_j(\theta_0\|\theta)d\theta$$

K คือ จำนวนข้อสอบที่ถูกทำไปแล้ว

r คือ ค่าคงที่ ซึ่งโดยปกติระบุเท่ากับ 3

Veldkamp และ van der Linder (2002) ได้เสนอ KI ในรูปแบบของ multivarite ซึ่งเป็นค่าที่ประยุกต์มาใช้โดยตรงจากรูปแบบเอกมิติ ซึ่ง multivariate KI นิยามดังนี้

$$KI_j(\theta_0) = \int_{\theta_0-\frac{r}{\sqrt{k}}}^{\theta_0+\frac{r}{\sqrt{k}}} \cdots \int_{\theta_{p0}-\frac{r}{\sqrt{k}}}^{\theta_{p0}+\frac{r}{\sqrt{k}}} KL_j(\theta_0\|\theta)d\theta$$

ในที่นี้ KI คือ integration ของ KL ของ p มิติ เมื่อ θ_0 ไม่ทราบค่าในทางปฏิบัติ ซึ่งมักจะถูกแทนที่ด้วย $\hat{\theta}^{mle}, \hat{\theta}^{EAP}$ หรือ $\hat{\theta}^{MAP}$ และกฎการคัดเลือกข้อสอบจะอยู่ในรูปแบบที่เป็นทางการดังนี้

$$j \equiv \arg\max_j\left\{KI_j(\hat{\theta}), j \in R_k\right\}$$

เมื่อ ข้อสอบข้อที่ k ถูกคัดเลือก และด้วยเหตุนี้หนึ่งใน R_k จะให้ค่าสูงสุดสำหรับ $KI_j(\hat{\theta})$

3.4.4 กฎการยุติการทดสอบแบบพหุมิติ

จากที่กล่าวมาแล้วในการทดสอบปรับเหมาะแบบเอกมิติ (UCAT) จะเห็นว่าวิธีการโดยทั่วไปสำหรับยุติการทดสอบจะมี 2 วิธี นั่นคือ จำกัดจำนวนข้อสอบ(fixed length) และ จำนวนข้อสอบผันแปร (variable length) หรือการพิจารณาจากค่าความคลาดเคลื่อนมาตรฐานในการประมาณค่า (SEE: Standard error of estimation) สำหรับงานวิจัย MCAT ที่ผ่านนั้นส่วนใหญ่จะใช้การกำหนดจำนวนข้อสอบ ซึ่งยังไม่มีการศึกษาว่าจำนวน ข้อสอบเท่าใดจึงจะเหมาะสมสำหรับการทดสอบปรับเหมาะแบบพหุมิติ

3.4.5 การควบคุมการใช้ข้อสอบซ้ำ

การเสนอวิธีการควบคุมการใช้ข้อสอบซ้ำที่ผ่านมาเป็นการเสนอวิธีสำหรับการทดสอบแบบ 1 มิติ Lee, IP, and Fuh (2008) ได้เสนอวิธีการควบคุมการใช้ข้อสอบซ้ำสำหรับการทดสอบแบบพหุมิติ (multidimensional tests) โดยเป็นวิธีที่ยึดหลักการแบ่งชั้นให้สอดคล้องกับหน้าที่ของเวคเตอร์พารามิเตอร์อำนาจจำแนก ที่สามารถนำมาใช้กับการคำนวณที่มีค่าน้อยที่สุด ซึ่งวิธีการดังกล่าวเป็นการ ขยายงานของ Chang และ Ying (1999) การควบคุมการใช้ข้อสอบซ้ำ ซึ่งมีรายละเอียดดังนี้

1) maximum likelihood estimation and D-optimality

วิธีคัดเลือกข้อสอบสำหรับการทดสอบแบบเอกมิติจะยึดหลักเกณฑ์สารสนเทศสูงสุดของฟิชเชอร์ (maximum Fisher information criterion) เมื่อการทดสอบเป็นแบบพหุมิติและมีลักษณะเป็น 2 มิติ (d = 2) วิธีการคัดเลือกข้อสอบสามารถสร้างขึ้นได้ดังนี้ โดยสมมติให้ $\theta = (\theta_1, \theta_2), a = (a_1, a_2)$ หลังจากที่ผู้สอบทำข้อสอบแล้วจำนวน k ข้อ สารสนเทศฟิชเชอร์ที่ความสามารถ θ หาได้จาก

$$\mathrm{I_k}(\theta) = \mathrm{D}^2 \begin{bmatrix} \sum_{i=1}^{k} a_{1i}^2 p_i(\theta) q_i(\theta) & \sum_{i=1}^{k} a_{1i} a_{2i} p_i(\theta) q_i(\theta) \\ \sum_{i=1}^{k} a_{2i} a_{1i} p_i(\theta) q_i(\theta) & \sum_{i=1}^{k} a_{2i}^2 p_i(\theta) q_i(\theta) \end{bmatrix} \quad (37)$$

โดยที่ $q_i(\theta) = 1 - p_i(\theta)$

การประมาณค่าความเป็นไปได้สูงสุด (maximum likelihood estimate: MLE) ได้มาเพื่อแทนที่ปริมาณความสามารถ θ ที่ไม่ทราบค่าในการคำนวณค่าสารสนเทศ ขณะที่ข้อถัดไปหรือข้อที่ k +1 จะต้องถูกเลือก ดังนั้นสารสนเทศที่มากที่สุดสามารถให้ที่ความสามารถ $\hat{\theta}$

ประโยชน์ของวิธี MLE ภายใต้เงื่อนไขกฎเกณฑ์ โดยวิธีนี้จะมีแนวโน้มไปสู่การแจกแจงแบบปกติของตัวแปรพหุ ด้วยการกระจายของเมตริกซ์ $I_{k+1}^{-1}(\theta)$ ซึ่งเป็นส่วนกลับของเมตริกซ์สารสนเทศของฟิชเชอร์ ความเชื่อมั่นสำหรับความสามารถ θ มาจากรูปแบบดังนี้

$$\left\{\theta : (\theta - \hat{\theta})^T I_{k+1}(\hat{\theta})(\theta - \hat{\theta}) \leq cons\tan t\right\}$$

เพราะปริมาณของ พื้นที่ความเชื่อมั่น (ellipsoid) เป็นปริมาณที่เป็นสัดส่วนกันกับ $\left|I_{k+1}(\hat{\theta})\right|^{-1/2}$ (นั่นคือ $|\cdot|$ แทน determinant) โดยที่

$$\left|I_{k+1}(\hat{\theta})\right| = D^4\left\{\sum_{i=1}^{k+1} a_{1i}^2 p_i(\hat{\theta})q_i(\hat{\theta}) \cdot \sum_{i=1}^{k+1} a_{2i}^2 p_i(\hat{\theta})q_i(\hat{\theta}) - \left(\sum_{i=1}^{k+1} a_{1i}a_{2i} p_i(\hat{\theta})q_i(\hat{\theta})\right)^2\right\} \quad (38)$$

ธรรมชาติของการออกแบบเกณฑ์คือ การทำให้เกิดพื้นที่ความเชื่อมั่น (ellipsoid) น้อยที่สุดเท่าที่จะเป็นไปได้ หรือกล่าวได้ว่า การเลือกข้อสอบข้อที่ (K+1) ที่ให้ค่า $\left|I_{k+1}(\hat{\theta})\right|$ สูงสุด ซึ่งเป็นเกณฑ์ที่ได้มาตรฐานของ D-optimality (Atkinson & Donev, 1992; Kiefer, 1959; Silvery, 1980) หลังจากทำข้อสอบข้อ ที่ (K+1) และการตอบข้อสอบนั้นทำให้เกิดการจูงใจจากผู้สอบ ผลการตอบจะถูกให้คะแนนและความสามารถ $\hat{\theta}$ จะถูกประมาณค่าใหม่ กระบวนการนี้จะถูกทำซ้ำจนกระทั้งการทดสอบยุติและความสามารถ $\hat{\theta}$ ในขั้นสุดท้าย ถูกประมาณค่า

2) **วิธีการคัดเลือกข้อสอบด้วยวิธี** a-stratified multistage

การเสนอวิธีการควบคุมการใช้ข้อสอบซ้ำนี้จะใช้ การทำหน้าที่ของ a ในการแบ่งลำดับขั้นคลังข้อสอบ นอกเหนือจากกลไกการควบคุมการใช้ข้อสอบซ้ำ วิธีการคัดเลือกข้อสอบจะขึ้นอยู่กับ D-optimality เพียงอย่างเดียว มีแนวโน้มที่จะเลือกข้อสอบที่มี a_1 และ a_2 ใกล้กับขอบเขตบน เมื่อ d=1 ซึ่งหมายความว่าข้อสอบที่มีค่าพารามิเตอร์ a สูง จะมีความเหมาะสมที่จะถูกเลือก ปรากฏการณ์นี้ถูกตรวจสอบ โดย Chang and Ying (1999) เมื่อ d > 1 แนวคิดที่จะแบ่งลำดับขั้นของแบบแผนจะขึ้นอยู่กับ ค่าพารามิเตอร์ a เช่นเดียวกับ d=1 ซึ่งดูได้จากสมการที่ 37 ซึ่งผลของ b ถูกกำหนดใน $p_i(\hat{\theta})$ และ $q_i(\hat{\theta})$ เมื่อแบบทดสอบมีความก้าวหน้า $\hat{\theta}$ จะมีค่าคงที่ และ $|I|$ คือหน้าที่ของเวคเตอร์ a ดังนั้นความท้าทายของการคัดเลือกที่สมเหตุสมผล ในการเลือกฟังก์ชันที่เหมาะสมของ a ซึ่งทำให้แน่ใจได้ว่าลักษณะของการดำเนินการ a-stratified สำหรับคลังข้อสอบแบบพหุมิติ (multidimensional item pool) มีความคล้ายคลึงกับคลังข้อสอบแบบเอกมิติ (unidimensional item pool)

Chang and Ying (1999) ได้เสนอกระบวนการของ a-Stratified multistage (ASTR) ไว้ดังนี้

1) การแบ่งแยกข้อสอบในคลังข้อสอบเป็น M ระดับ ตามค่าของพารามิเตอร์ a ของข้อสอบ ชั้นของข้อสอบข้อแรก (first item stratum) จะกำหนดด้วย ค่าที่น้อยที่สุดของ $|a_1 - a_2|$
2) แบ่งแยกแต่ละระดับยกเว้นชั้นแรก เป็น 2 ส่วนย่อย ส่วนแรก ประกอบด้วยข้อสอบ $(a_1 > a_2)$ ส่วนที่สอง ประกอบด้วยข้อสอบ $(a_1 < a_2)$
3) แบ่งแยกข้อสอบเป็น M ขั้น
4) ขั้นแรก เลือกจากชั้นของข้อสอบข้อแรก (first item stratum) จำนวน n_1 ข้อ ตามเกณฑ์นี้ สำหรับการคัดเลือกแต่ละจุด จะเลือกข้อสอบที่มีค่าความยากพารามิเตอร์ b ที่มีค่าใกล้กับค่าตามสูตร

$$b_i = \frac{a_1\hat{\theta}_1 + a_2\hat{\theta}_2}{a_1 + a_2} \qquad (39)$$

5) ในขั้นที่ j เมื่อ j > 1 เลือกข้อสอบข้อที่ n_{j1} และ n_{j2} แต่ละส่วนย่อยของชั้นที่ j เช่น $n_j = n_{j1} + n_{j2}$ ถูกแบ่งชั้น จากนั้นจับคู่ความยากของข้อคำถามพารามิเตอร์ b กับค่าเฉลี่ยถ่วงน้ำหนักของ $\hat{\theta}$ เสนอไว้ในสมการที่ 42
6) ดำเนินการซ้ำในข้อที่ 4 สำหรับ j = 2, …, M. $\sum_{j=2}^{M} n_{j1} = \sum_{j=2}^{M} n_{j2}$

ค่าของ M, n_j, j = 1, …, M ถูกกำหนดในแบบทดสอบไว้ล่วงหน้า แบบทดสอบสามารถออกแบบไปยังการสิ้นสุดการสอบ เมื่อ $\hat{\theta}$ ไปถึงระดับที่ระบุได้อย่างแม่นยำ นอกจากนั้นเมื่อจำนวน ข้อสอบที่ $(a_1 > a_2)$ ถูกปรับเทียบกับจำนวนข้อสอบที่ $(a_1 < a_2)$ แบบสอบจะมีความสมดุลในเรื่องความสามารถในการจำแนกในทั้ง 2 มิติ หรืออาจกล่าวได้ว่า θ_1 และ θ_2 ควรได้รับการวัดที่มีความแม่นยำเท่าเทียมกัน

3.4.6 **ความสมดุลของเนื้อหา** (Content Balancing)

นอกเหนือจากการให้ประสิทธิภาพการวัดแล้ว MCAT ยังให้กลไกในประเด็นความสมดุลของเนื้อหาใน CAT ซึ่งสิ่งที่เป็นกังวลอย่างหนึ่ง สำหรับ CAT คือ การที่ผู้สอบได้รับจำนวนข้อสอบไม่เท่ากัน ผลลัพธ์ของการประมาณค่าความสาม ารถขึ้นกับความแตกต่างของการผสมผสานเนื้อหา เช่น การสอบวิชาวิทยาศาสตร์ทั่วไป อาจมีองค์ประกอบของข้อคำถามจาก 3 เนื้อหา ได้แก่ ชีววิทยา ฟิสิกส์ และ เคมี วิธีการหนึ่งในการสร้างสมดุลเนื้อหาใน CAT คือ การจัดสรรการนำเสนอ

ข้อสอบระหว่างเนื้อหาในทางที่มั่นใจว่า ผู้สอบ แต่ละคนจะได้รับจำนวนข้อสอบที่เฉพาะเจาะจงในแต่ละเนื้อหา ตัวอย่างเช่น อาจใช้วิธีการชี้เฉพาะว่าผู้สอบแต่ละคนจะได้รับข้อสอบชีววิทยา 5 ข้อ ฟิสิกส์ 5 ข้อ และ เคมี 5 ข้อ ซึ่งวิธีดังกล่าวเป็นวิธีที่เราไม่ต้องการ การบริหารข้อสอบจาก 1 เนื้อหาหรือหลายเนื้อหาอาจจะไม่ให้สารสนเทศเกี่ยวกับระดับทั่วไปของประสิทธิภาพสำหรับผู้สอบ นั่นคือ บางข้อคำถามอาจมีระดับความยากที่ไม่เหมาะสม เช่น ถ้าวิชาเคมีมีความยากมากกว่าเนื้อหาวิชาอื่นๆ ดังนั้นการบริหารข้อสอบข้อนี้กับผู้สอบ ที่มีความสามารถปานกลาง อาจจะให้สารสนเทศในเรื่องความรู้ในด้านวิทยาศาสตร์น้อยกว่าความเป็นจริง ทางเลือกหนึ่งของการสร้างสมดุลของเนื้อหาคือ การใช้ MAT จัดการในแต่ละเนื้อหาโดยแยกออกเป็นมิติดีกว่าที่จะกำหนดจำนวนข้อสอบในแต่ละเนื้อหา โดยข้อสอบจะถูกเลือกโดยใช้พื้นฐานของเกณฑ์การเลือกข้อสอบของเบส์ (multivariate Bayesian) การใช้เกณฑ์นี้ ระดับของประสิทธิภาพใน 1 มิติ จะช่วยเลือกสารสนเทศของข้อคำถามจาก 2 มิติอื่น ๆ ในแต่ละขั้นตอนของกระบวนการเลือกข้อสอบ (Segall, 1996) สิ่งนี้ชี้ให้เห็นว่าตั้งแต่ MCAT ให้การประมาณค่าความสามารถแยกออกในแต่ละมิติ ค่านี้จะถูกรวมเป็นรูปแบบการวัดองค์ประกอบเดียว ในการทดสอบแบบ UCAT เกี่ยวกับขอบเขตเนื้อหาที่มีความหลากหลาย จะมีการให้น้ำหนักเนื้อหาในแต่ละมิติที่แน่นอนเพื่อใช้ในการคัดเลือกข้อสอบสำหรับการทดสอบ MCAT การให้น้ำหนักจะต้องชัดเจนและชี้เฉพาะ ค่านี้สามารถใช้รูปแบบองค์ประกอบเชิงเส้นตรงแบบถ่วงน้ำหนักของการประมาณค่าความสามารถแบบแยกใน แต่ละมิติ (Segall, 1996)

3.4.7 **ลำดับข้อสอบหรือบริบทข้อสอบ** (item-order or item context)

สิ่งที่เป็นกังวลอย่างหนึ่งเกี่ยวกับการใช้ MCAT ซึ่งเกิดขึ้นจากลำดับของข้อสอบหรือบริบทข้อสอบ ตัวอย่างเช่น ข้อสอบเกี่ยวกับภาษาอาจจะทำหน้าที่ต่างกัน เมื่อผู้สอบได้ทำข้อสอบข้อก่อนหน้า เป็นเนื้อหาคณิตศาสตร์ ซึ่งจะส่งผลมากกว่าการทำข้อสอบข้อก่อนหน้าเป็นข้อสอบเกี่ยวกับภาษาแบบอื่นๆ สำหรับข้อคำถามแบบพหุมิติจากหนึ่งเนื้อหาสามารถทำให้กระจายไปยังข้อสอบในเนื้อหาอื่นๆ กล่าวคือ ข้อสอบหนึ่งข้ออาจจะวัดคว ามสามารถในหลายเนื้อหา โมเดลที่นำเสนอนี้เป็นไปตามข้อตกลงเบื้องต้นเกี่ยวกับ local independent นั่นคือ ข้อสอบทุกข้อมีความเป็นอิสระจากกัน เนื่องจากลำดับการนำเสนอข้อสอบไม่มีผลต่อค่าอำนาจจำแนกและความยากของข้อสอบ แต่ความยากของข้อสอบหรืออำนาจจำแนกมากน้อย จะขึ้นกับ เนื้อหาของข้อสอบที่ผู้สอบได้รับ แม้ว่าการผสมผสานเนื้อหาในข้อสอบจะไม่มีผลกระทบต่อการทำหน้าที่ของข้อสอบ แต่ ในทางปฏิบัติแล้ว ลำดับข้อสอบที่ผสมผสานเนื้อหาหรือมิติการวัด จะส่งผลต่อระดับความไม่สบายใจของผู้สอบ ซึ่งสิ่งนี้เกิดมาจากความยากของข้อสอบที่ ผู้สอบได้รับ เพราะข้อสอบที่มีเนื้อหาหรือมิติ

ต่างกัน ย่อมจะมีความยากที่ต่างกัน อย่างไรก็ตาม สำหรับข้อสอบที่มีความยาวปานกลาง ถึงยาวมาก การทดสอบด้วยรูปแบบผสมผสานเนื้อหา จะช่วยให้ผู้สอบมีแรงจูงใจในการทำข้อสอบ ดังนั้น ลำดับของข้อสอบจึงมีผลต่อการตอบข้อสอบข้อถัดไป ผลกระทบนี้ส่งผลต่อปฏิกิริยา ตอบสนองของผู้สอบ รวมถึงความแม่นยำของการประมาณค่าความสามารถ (Segall, 1996)

นอกจากนี้ Li และ Schafer (2004) ได้นำเสนอโครงร่างวิจัยต่อประชุมประจำปีของ American education research association เกี่ยวกับการศึกษาผลกระทบของลำดับข้อสอบหรือบริบทข้อสอบในการทดสอบปรับเหมาะแบบพหุมิติ พร้อมทั้งศึกษาอัตรการใช้ข้อสอบซ้ำ โดย Li และ Schafer (2004) กล่าวไว้ว่า เมื่อนำการทดสอบปรับเหมาะแบบพหุมิติมาใช้เก็บข้อมูลจากการทดสอบจริงจะพบปัญหาเกี่ยวกับลำดับข้อสอบ ในกรณีที่ต้องการวัดความสามารถของผู้สอบเกี่ยวกับการอ่านและคณิตศาสตร์ไปพร้อมกัน เราจะดำเนินการสอบนี้อย่างไร จะสอบโดยทำชุดข้อสอบการอ่านก่อนแล้วตามด้วยชุดข้อสอบคณิตศาสตร์ หรือ ทำชุดข้อสอบคณิตศาสตร์ก่อนแล้วตามด้วยชุดข้อสอบการอ่าน หรือผสามผสานข้อสอบ นั่นคือ ข้อสอบการอ่านตามด้วยข้อสอบคณิตศาสตร์ โดยพวกเขาตั้งคำถามว่า การดำเนินการสอบที่มีรูปแบบลำดับข้อสอบที่ต่างกัน จะส่งผลต่อการประมาณค่าความสามารถและอัตราการใช้ข้อสอบซ้ำอย่างมีนัยสำคัญหรือไม่ โดย Li และ Schafer ศึกษาโดยวิธีการจำลองข้อมูล โดยกำหนดความยาวข้อสอบไว้ 51 ข้อ แบ่งเป็นข้อสอบการอ่าน 25 ข้อ และข้อสอบคณิตศาสตร์ 26 ข้อ อย่างไรก็ตามผลวิจัยจากงานวิจัยชิ้นนี้ ยังไม่พบการตีพิมพ์เผยแพร่ในวารสารทางวิชาการฉบับใด จึงไม่สามารถนำผลวิจัยมานำเสนอได้

3.5 คุณภาพของการทดสอบปรับเหมาะแบบพหุมิติ

การศึกษากระบวนการการทดสอบปรับเหมาะในแต่ละวิธี ย่อมมีจุดมุ่งหมายที่จะเปรียบเทียบคุณภาพ หรือประสิทธิภาพของการทดสอบว่าวิธีใดเหมาะสมที่สุด ซึ่งการพิจารณาคุณภาพของการทดสอบปรับเหมาะแบบพหุมิตินั้น สามารถพิจารณาได้จากหลายค่า ซึ่งในที่นี้จะแบ่งคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติออกเป็น 2 กรณี คือ กรณีทราบค่าความสามารถที่แท้จริง และกรณีที่ไม่ทราบค่าความสามารถที่แท้จริง ซึ่งมีรายละเอียดดังนี้

3.5.1 **กรณีทราบค่าความสามารถที่แท้จริง** วิธีการนี้จะใช้กับการจำลองข้อมูล ซึ่งผู้วิจัยกำหนดค่าความสามารถที่แท้จริงขึ้นมา และการใช้ข้อมูลทุติยภูมิจากการทดสอบต่างๆ ที่ผ่านมา โดยค่าความสามารถที่แท้จริงจากการทดสอบทุติยภูมินี้ จะใช้ค่าที่ได้จากการสอบทำข้อสอบทุกข้อในคลังข้อสอบ ซึ่งการทดสอบที่ผ่านนั้นผู้สอบอาจทำการทดสอบแบบดั้งเดิม จำนวน 150 ข้อ แต่เมื่อนำข้อมูลดังกล่าวมาใช้กับการทดสอบแบบปรับเหมาะ โปรแกรมการทดสอบจะคัดเลือกข้อสอบขึ้นมาเพียงไม่กี่ข้อ ซึ่งในส่วนนี้ผู้วิจัยจึงต้องการหาความสัมพันธ์ของค่าที่ประมาณค่าได้จาก

ทดสอบแบบปรับเหมาะกับค่าความสามารถที่แท้จริง หรือเป็นการค่าความใกล้เคียงกันของค่าทั้งสอง ซึ่งสามารถพิจารณาคุณภาพข้อสอบได้2 วิธี ดังนี้

1) **ค่าความเที่ยง** (reliability)

ค่าความเที่ยงคือ กำลังสองของความสัมพันธ์ระหว่างค่าความสามารถของผู้สอบที่ได้จากการประมาณค่าจากการทดสอบแบบปรับเหมาะ กับค่าความสามารถที่แท้จริงของผู้สอบ โดยที่ ค่าความเที่ยงสูง หมายถึง วิธีการทดสอบปรับเหมาะแบบพหุมิติมีคุณภาพสูง ซึ่งค่าความเที่ยงหาได้จากสูตร ดังนี้

$$r^2 = r^2_{xy}$$

$$r_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left(n\sum x^2 - \left(\sum x\right)^2\right)\left(n\sum y^2 - \left(\sum y^2\right)\right)}}$$

เมื่อ r^2 คือ ความเที่ยงของการวัด

r_{xy} คือ ความสัมพันธ์ระหว่างค่าความสามารถ

n คือ จำนวนผู้สอบ

x คือ ค่าความสามารถของผู้สอบที่ได้จากการประมาณค่า

y คือ ค่าความสามารถที่แท้จริงของผู้สอบ

2) **ค่าระยะทางแบบยุคลิด** (Euclidean distance)

ค่าระยะทางแบบยุคลิด คือ การหาระยะห่างระหว่าง ค่าความสามารถของผู้สอบที่ได้จากการประมาณค่า จากการทดสอบแบบปรับเหมาะ ($\hat{\theta}$) กับค่าความสามารถของผู้สอบที่แท้จริง (θ) สำหรับกรณีการทดสอบสั้นๆ ก็สามารถนำระยะทางแบบยุคลิด มาเป็นดัชนีวัดความแม่นยำทางจิตมิติ (Finkelman et al., 2009) โดยที่ ค่าระยะทางน้อย หมายถึง ค่าความสามารถจากการประมาณค่าและค่าความสามารถที่แท้จริงมีความใกล้เคียงกัน นั่นคือ วิธีการทดสอบปรับเหมาะแบบพหุมิติมีคุณภาพสูง ซึ่งค่าระยะทางแบบยุคลิด หาได้จากสูตร ดังนี้

$$ED_i = \sqrt{\left(\theta_{i1} - \hat{\theta}_{i1}\right)^2 + \left(\theta_{i2} - \hat{\theta}_{i2}\right)^2 + \ldots \left(\theta_{ij} - \hat{\theta}_{ij}\right)^2}$$

ED_i คือ ค่าระยะทางแบบยุคลิดของผู้สอบคนที่ i

i คือ ผู้สอบคนที่ i

j คือ มิติความสามารถ ที่ j

$\hat{\theta}_{ij}$ คือ ค่าความสามารถที่ประมาณค่าได้ของผู้สอบคนที่ i มิติที่ j

θ_{ij} คือ ค่าความสามารถที่แท้จริงของผู้สอบคนที่ i มิติที่ j

3.5.1 **กรณีไม่ทราบค่าความสามารถที่แท้จริง** วิธีการนี้จะใช้กับกรณีการทดสอบจริงที่ผู้วิจัยไม่ทราบค่าความสามารถที่แท้จริงของผู้สอบ สิ่งที่สามารถพิจารณาได้คือ ค่าสารสนเทศของแบบสอบ ค่าความคลาดเคลื่อนมาตรฐานในการประมาณค่า และความยากของข้อสอบ

1) สารสนเทศของแบบสอบ

ค่าฟังก์ชันสารสนเทศของแบบสอบ (test information: $I(\theta)$) เกิดจากผลรวมเชิงพีชคณิตของค่าฟังก์ชันสารสนเทศของข้อสอบแต่ละข้อรวมเข้าด้วยกันทั้งฉบับ ณ ตำแหน่ง θ เดียวกัน สารสนเทศแบบสอบสูงจะมีคุณภาพการทดสอบสูง

$$I(\theta)=\sum_{i=1}^{k} I_i(\theta) \quad , \quad i = 1, 2, \ldots, k$$

$$I_i(\theta)=\frac{\left[P_i'(\theta)\right]^2}{P_i(\theta)Q_i(\theta)} \quad , \; i = 1, 2, \ldots, k$$

เมื่อ $I(\theta)$ = ค่าฟังก์ชันสารสนเทศ หรือค่าสารสนเทศที่ได้จากแบบสอบถามสำหรับผู้ตอบที่มีความสามารถ θ

$I_i(\theta)$ = ค่าฟังก์ชันสารสนเทศ หรือค่าสารสนเทศที่ได้จากข้อสอบข้อที่ i สำหรับผู้ตอบที่มีความสามารถ θ

$P_i'(\theta)$ = P_i' = ความชันของฟังก์ชันการตอบสนองข้อสอบข้อที่ i ณ ตำแหน่งความสามารถ θ

$P_i(\theta)$ = P_i = ความน่าจะเป็นของผู้ตอบที่มีความสามารถ θ จะตอบข้อสอบข้อที่ i ได้ถูกต้อง

$Q_i(\theta)$ = Q_i = 1- $P_i(\theta)$

2) ค่าความคลาดเคลื่อนมาตรฐานในการประมาณค่า

ในกรณีที่กำหนดจำนวนข้อสอบให้เท่ากัน ผู้วิจัยสามารถพิจารณาคุณภาพการทดสอบจากค่าความคลาดเคลื่อนมาตรฐานในการประมาณค่านั่นคือ การทดสอบที่มีค่าความคลาดเคลื่อนมาตรฐานในการประมาณค่าต่ำ แสดงว่ามีคุณภาพในการทดสอบสูง

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

เมื่อ $SE(\theta)$ = ความคลาดเคลื่อนมาตรฐานของการประมาณค่า θ

$I(\theta)$ = สารสนเทศของแบบสอบที่ให้สำหรับผู้ที่มีความสามารถ θ

3) ความยาวข้อสอบ

การพิจารณาจำนวนข้อสอบจะใช้ในกรณีที่การทดสอบปรับเหมาะ กำหนดเกณฑ์การยุติข้อสอบโดยพิจารณาค่า ความคลาดเคลื่อนมาตรฐานในการประมาณค่า ซึ่งโดยทั่วไปจะกำหนดไว้ที่ตำแหน่งความคลาดเคลื่อนมีค่าน้อยกว่าหรือเท่ากับ 0.3 (SE $\leq$0.3) โดยการทดสอบแบบปรับเหมาะที่มีคุณภาพนั้น ผู้สอบจะได้รับข้อสอบจำนวนน้อยแต่สามารถประมาณค่าความสามารถได้อย่างถูกต้องและแม่นยำ ดังนั้นวิธีการทดสอบรูปแบบใดที่ให้จำนวนข้อสอบน้อยกว่าจะมีคุณภาพการทดสอบปรับเหมาะสูงกว่า

3.6 ทิศทางในอนาคตของการทดสอบปรับเหมาะแบบพหุมิติ

Reckase (2009) กล่าวไว้ว่า แม้การพัฒนา MIRT จะเกิดขึ้นมาหลายปีแล้ว แต่การนำมาใช้ในทางปฏิบัตินั้นยังมีน้อยมาก สำหรับ MIRT-CAT ยังเป็นเรื่องใหม่ในการศึกษาวิจัยทั้งการนำไปประยุกต์ใช้และการนำไปใช้ในทางปฏิบัติซึ่งจะมีมากขึ้นในอนาคต สำหรับประเด็นที่น่าสนใจในอนาคตนั้นสามารถระบุได้ง่ายมาก เนื่องจากงานวิจัยที่ตีพิมพ์ออกมานั้นจะศึกษาเฉพาะกรณีที่เป็น 2 มิติ ซึ่งอาจจะไม่ตรงกับสภาพความเป็นจริงที่จะนำไปใช้ในทางปฏิบัติได้ ดังนั้นจึงควรจะมีการศึกษากรณีที่มี 3 มิติ หรือมากกว่า เพื่อจะหาข้อสรุปที่มีความเหมาะสมกับการทดสอบที่มีความหลากหลายมากขึ้น นอกจากนั้น งานวิจัยที่ศึกษาในปัจจุบันจะใช้การจำกัดจำนวนข้อสอบในการยุติการทดสอบ ซึ่งอาจจะได้สารสนเทศน้อย หรือหากจะใช้วิธีการดังกล่าวควรจะมีการศึกษาจำนวนความยาวข้อสอบที่เหมาะสมสำหรับการทดสอบปรับเหมาะแบบพหุมิติ และควรมีการพัฒนาแนวทางสำหรับการสร้างคลังข้อสอบ MCAT ให้มีความเหมาะสม โดยศึกษาจำนวนข้อสอบและลักษณะข้อสอบ โดยคำนึงถึงจำนวนมิติของการทดสอบว่ามีผลต่อจำนวนข้อสอบในคลังข้อสอบอย่างไร รวมถึงความเหมาะสมของความยาก อำนาจจำแนก

โดยสรุปการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์สำหรับโมเดลการต อบสนองข้อสอบแบบพหุมิติ (MCAT) เป็นการทดสอบแบบใหม่ที่เป็นการผสมผสานวิธีการของ การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (CAT) และ โมเดลการตอบสนองข้อสอบแบบพหุมิติ (MIRT) ซึ่งเป็นวิธีการที่มีความแม่นยำในการประมาณค่าความสามารถและมีประสิทธิภาพมากกว่าวิธีการทดสอบปรับเหมาะแบบเอกมิติ สำหรับขั้นตอนของการทดสอบก็มีความคล้ายคลึงกับการทดสอบแบบปรับเหมาะแบบ เอกมิติ จะมีความแตกต่างในรายละเอียดซึ่งมีผู้เริ่มศึกษาในเรื่องนี้มากขึ้น ซึ่งได้นำเสนอไว้ในตอนที่ 4 ซึ่งจะกล่าวถึงต่อไป

ตอนที่ 4 งานวิจัยที่เกี่ยวข้อง

การนำเสนองานวิจัยที่เกี่ยวข้องกับการศึกษาครั้งนี้ ได้แบ่งออกเป็นง านวิจัยที่เกี่ยวข้องกับการทดสอบแบบปรับเหมาะ และ งานวิจัยที่เกี่ยวข้องกับการทดสอบแบบปรับเหมาะสำหรับโมเดลการตอบสนองข้อสอบแบบพหุมิติ โดยมีรายละเอียดดังนี้

4.1 งานวิจัยที่เกี่ยวข้องกับการทดสอบแบบปรับเหมาะ

การนำเสนอ งานวิจัยที่เกี่ยวข้องกับการทดสอบแบบปรับเหมาะ ได้แบ่งตามองค์ประกอบของการทดสอบแบบปรับเหมาะ และประเด็นอื่นๆ ที่น่าสนใจ ได้แก่

4.1.1 คลังข้อสอบที่มีความเหมาะสม (calibrated item pool)

สำหรับส่วนที่เกี่ยวข้องกับคลังข้อสอบพบว่า ขนาดของคลังข้อสอบส่งผลต่ อการประมาณค่าความสามารถของผู้สอบน้อยมาก (Ho, 1989) และขนาดของคลังข้อสอบนั้นควรมีประมาณ 100-200 ข้อ โดยที่จำนวนคลังข้อสอบ 116-150 ข้อ จะให้ผลดีที่สุด (Weiss, 1988) ซึ่งสอดคล้องกับ Urry (1977) ที่พบว่า คลังข้อสอบที่ใช้ควรมีอย่างน้อย 100 ข้อ ส่วนข้อสอบในคลังข้อสอบนั้น ควรมีอำนาจจำแนกมากกว่า 0.8 ค่าความยากระหว่าง -2.0 ถึง 2.0 และมีโอกาสการเดาน้อยกว่า 0.3

4.1.2 จุดเริ่มต้นการทดสอบ (starting point or entry level) และวิธีการคัดเลือกขั้นสอบข้อแรก (item selection procedures at the early stages)

ความสำคัญของจุดเริ่มต้นยังคงเป็นที่โต้แย้งกันในกลุ่มนักวิจัย บางคนกล่าวว่าจุดเริ่มต้นไม่มีผลกระทบต่อการสอบถ้ามี จำนวนข้อสอบมาก เช่น Lord (1977 อ้างถึง ในศิริชัย กาญจนวาสี, 2550) ได้ทำการทดลองด้วยสถานการณ์จำลอง โดยเริ่มต้นด้วยข้อสอบที่มีค่าความยากระดับต่าง ๆ พบว่าเกือบไม่มีผลต่อความถูกต้องแม่นยำในการประมาณค่าความสามารถของผู้สอบเลย โดยเฉพาะกรณีที่มีข้อสอบในคลังข้อสอบจำนวนมาก ซึ่งสอดคล้องกับ รังสรรค์ มณีเล็ก (2539) ซึ่งกล่าวว่า การคัดเลือกข้อสอบข้อแรกไม่ค่อยมีความสำคัญต่อการทดสอบแบ บปรับเหมาะกับความสามารถของผู้สอบเท่าใดนัก โดยพบว่า เมื่อใช้ข้อสอบที่มีความยากต่างกัน 3 ระดับ คือ ง่าย ปานกลาง และยาก เป็นข้อแรกในการทดสอบแบบปรับเหมาะกับความสามารถของผู้สอบด้วยคอมพิวเตอร์ ไม่ทำให้ค่าความตรงเชิงสภาพของการทดสอบต่างกัน

Chen, Ankenmann, and Chang (2000) ได้ศึกษาเปรียบเทียบวิธีการเลือกข้อสอบขั้นแรกในการทดสอบแบบปรับเทียบคะแนนด้วยคอมพิวเตอร์ โดยได้เปรียบเทียบทั้งหมด 5 วิธี คือ วิธีสารสนเทศของฟิชเชอร์ (Fisher information: FI) วิธีสารสนเทศแบบช่วงของฟิชเชอร์ (Fisher interval information: FII) วิธีสาร สนเทศการแจกแจงภายหลังของฟิชเชอร์ (Fisher

information with a posterior distribution: FIP) วิธีสารสนเทศของคลูแบค- ไลเบอร์ (Kullback-Leibler information: KL) และ วิธีสารสนเทศการกระจาย ภาย หลังของคลูแบค- ไลเบอร์ (Kullback-Leibler information with a posterior distribution: KLP) ผลการศึกษาพบว่า ขั้นแรกของ CAT วิธีการ FI ดีกว่าวิธีอื่นๆสำหรับ θ มีค่าใกล้ 0 อย่างไรก็ตามวิธีการเลือกข้อสอบขั้นแรกวิธีอื่นๆ ดีกว่าวิธีการ FI เมื่อ θ มีค่าไกลจาก 0 และเมื่อความยาวของข้อสอบพบว่า ทุก วิธีให้ผลลัพธ์ที่คล้ายคลึงกัน ทั้งนี้ผลการวิจัยดังกล่าวมีความแตกต่างจากงานวิจัยของ Chen and Ankenmann (2004) ได้ศึกษาเพิ่มเติมโดยเปรียบเทียบวิธีการคัดเลือกข้อสอบข้อแรกใน 4 วิธี คือ วิธีสารสนเทศของฟิชเชอร์ (Fisher information: FI) วิธีสารสนเทศการแจกแจงภายหลั งของฟิชเชอร์ (Fisher information with a posterior distribution: FIP) วิธีสารสนเทศการกระจายภายหลังของคลูแบค-ไลเบอร์ (Kullback-Leibler information with a posterior distribution: KLP) และ วิธีการคัดเลือกแบบสุ่มสมบูรณ์ (completely randomized item selection: RN) โดยเปรียบเทียบในประเด็นของความแม่นยำในการประมาณค่าคุณลักษณะแฝง ซึ่งเปรียบเทียบใน 3 เงื่อนไข คือ 1) การใช้เพียงฟังก์ชันสารสนเทศของข้อสอบ เป็นเกณฑ์ในการคัดเลือก 2) การใช้ฟังก์ชันสารสนเทศของข้อสอบ และความสมดุลของเนื้อหาเป็ นเกณฑ์ในการคัดเลือก และ 3) การใช้ฟังก์ชันสารสนเทศของข้อสอบ ความสมดุลของเนื้อหาและการควบคุมการใช้ข้อสอบซ้ำเป็นเกณฑ์ในการคัดเลือก เมื่อข้อสอบมีความยาวน้อยกว่า 10 ข้อ พบว่า FIP และKLP มีคุณสมบัติที่ เหนือกว่า FI ที่ระดับความสามารถสุดโต่งในเงื่อนไขที่1 อย่างไรก็ตามในสถานการณ์จริง ไม่สามารถสรุปได้ว่า FIP และKLP มีคุณสมบัติที่เหนือกว่า FI โดยเฉพาะอย่างยิ่งเมื่อมีการควบคุมการใช้ข้อสอบซ้ำ ในกรณีที่ความยาวข้อสอบมากกว่า 10 ข้อ พบว่า 3 วิธีที่ไม่ได้ใช้การสุ่มมีผลการปฏิบัติคล้ายคลึงกันไม่ว่าจะเป็นเงื่อนไขใดก็ตาม ในข ณะที่วิธี สารสนเทศของ ฟิชเชอร์ (Fisher information: FI) มีการใช้ข้อสอบสูงกว่าวิธีอื่นเล็กน้อย

4.1.3 **การคัดเลือกข้อสอบ** (item selection)

Hulin and other (1983) พบว่า การใช้เกณฑ์การคัดเลือกข้อสอบโดยการจับคู่ระหว่างค่าความสามารถของผู้สอบกับค่า $\theta_{\max}$ เพื่อแก้ปัญหาการใช้ข้อสอบมากเกินไปนั้น พบว่ามีการสูญเสียสารสนเทศมากกว่าการคัดเลือกข้อสอบโดยวิธีสารสนเทศสูงสุด $(I_{\max})$ หรือเรียกอีกอย่างว่า วิธีสารสนเทศของฟิชเชอร์ ส่วน รังสรรค์ มณีเล็ก (2540) ได้ศึกษาวิธีเลือก 3 วิธี คือ การเลือกค่าความยากของข้อสอบ (b) ให้ใกล้เคียงกับความสามารถของผู้สอบ การเลือกค่าอำนาจจำแนก ค่าความยาก ค่าการเดา ให้ใกล้เคียงกับความสามารถของผู้สอบ และเลือกค่าสารสนเทศของข้อสอบที่สูงสุด จากการศึกษาพบว่า การคัดเลือกข้อสอบข้อต่อไปโดยวิธีเลือกค่า

สารสนเทศของข้อส อบที่สูงที่สุดจะมีค่าความตรงเชิงสภาพสูงสุด นอกจากนั้น Segall (2000) ได้ศึกษาการเลือกข้อสอบด้วยวิธี Maximum likelihood (ML) และวิธี Bayesian ซึ่งพบว่า การเลือกข้อสอบด้วยวิธี ML ไม่ได้พิจารณาความรู้ก่อนหน้าเกี่ยวกับการแจกแจงร่วมของความสามารถ ซึ่งข้อบกพร่องนี้ แก้ไขได้ด้วยวิธีของเบส์ van der Linden (2005) ได้ศึกษาเปรียบเทียบวิธีการคัดเลือกข้อสอบสำหรับการทดสอบแบบปรับเหมาะที่มีการกำหนดเนื้อหา ในองค์ประกอบของแบบสอบความแตกต่างพื้นฐานที่มีอยู่ระหว่างขั้นตอนการคำนวณ คือ การเลือกตามลำดับแบบสอบหรือการเกิดขึ้นพร้อมกัน การสอบตามลำดับนั้น ยอมให้ดำเนินการกับฟังก์ชันเป้าประสงค์ ในการประมาณค่าความสามารถของผู้สอบ ให้เหมาะสม เช่น ฟังก์ชันสารสนเทศของแบบสอบใน CAT แต่ก็นำไปสู่ปัญหาที่สำคัญของวิธีการทำให้เกิดความตระหนักถึงชุดของเนื้อหาที่กำหนดไว้ ในแบบสอบ ปัญหานี้มีวิธีการแก้ปัญหาที่เป็นธรรมชาติโดยใช้วิธีการคัดเลือกข้อสอบพร้อมกัน วิธีการคัดเลือกข้อสอบในการทดสอบแฟปรับเหมาะหลัก ๆ มี 3 วิธี ที่สามารถแก้ไขปัญหาในสถานการณ์นี้ วิธี spiraling ดำเนินการคัดเลือกข้อสอบตามหมวดของข้อสอบในคลังข้อสอบ โดยใช้สัดส่วนที่เหมาะสมกับจำนวนความต้องกา ร วิธี weighted-deviations method (WDM) และ วิธี shadow test approach (STA) มีฐานคิดมาจากการฉายภาพของผลลัพธ์ในอนาคตในการเลือกข้อสอบ สองวิธีนี้แตกต่างกันในการคำนวณการฉายภาพก่อนของผลรวมของน้ำหนักคุณลักษณะที่เกิดขึ้นขั้นสุดท้ายของแบบทดสอบ และการฉายภาพภายหลังของแบบทดสอบด้วยตนเอง การเปรียบเทียบเชิงประจักษ์ระหว่างวิธี WDM และ STA ถูกดำเนินการในการทดสอบ Law School Admission Test (LSAT) ในเวอร์ชันที่เป็นการทดสอบแบบปรับเหมาะ ด้วยอัตราการใช้ข้อสอบที่เท่าเทียมกัน แต่วิธี WDM ฝ่าฝืนข้อบังคับบางข้อและมีความลำเอียงมาก และการประมาณค่าความสามารถไม่มีความแม่นยำ

4.1.4 การประมาณค่าความสามารถ (ability estimate)

Skaggs and Stevenson (1989) ได้ศึกษาการประมาณค่าความสามารถ ด้วยวิธีของเบส์และการประมาณค่าความสามารถของผู้สอบด้วยวิธีความเป็นไปได้สูงสุด พบว่า หากผู้สอบมากก ว่า 2,000 คนแล้ว ผลที่ได้จากการประมาณค่าทั้งสองวิธีไม่แตกต่างกัน ซึ่งสอดคล้องกับ ต่าย เชี่ยงฉี (2534) ที่พบว่า วิธีการประมาณค่าความสามารถตามวิธีของเบส์ที่ปรับปรุงใหม่ (Bayesian updating) และวิธีความเป็นไปได้สูงสุดแบบมีเงื่อนไข (conditional maximum likelihood) ให้ผลไม่แตกต่างกัน ในขณะที่ รังสรรค์ มณีเล็ก (2540) ได้ศึกษาการประมาณค่าความสามารถ พบว่า การประมาณค่าความสามารถของผู้สอบ จากการศึกษาโดยการเปรียบเทียบวิธีการประมาณค่าความสามารถของผู้สอบด้วยวิธีของเบส์และการประมาณค่า

ความสามารถของผู้สอบด้วยวิธีความเป็นไปได้สูงสุด พบว่า การประมาณค่าความสามารถด้วยวิธีของเบส์ มีความตรงเชิงสภาพสูงกว่าการประมาณค่าความสามารถด้วยวิธีความเป็นไปได้สูงสุด

4.1.5 เกณฑ์การยุติการสอบ (termination criteria or Stopping criteria)

จากการศึกษาของ Dodd and Koch (1993) โดยใช้ Partial credit model พบว่า การยุติการทดสอบโดยดูจากค่าความคลาดเคลื่อนมาตรฐานจะให้ผลดีกว่าการยุติการทดสอบโดยการกำหนดจำนวนข้อสอบ ซึ่งรังสรรค์ มณีเล็ก (2540) พบว่า ถ้ากำหนดความคลาดเคลื่อนในการประมาณค่าความสามารถของผู้สอบให้น้อยกว่าหรือเท่ากับ 0.3 จะให้ค่าความตรงเชิงสภาพสูงสุด ทั้งนี้ สิริลักษณ์ เกษรปทุมานันท์ (2549) ได้ศึกษาการกำหนดความคลาดเคลื่อนในการประมาณค่าความสามารถของผู้สอบ ซึ่งพบว่า การกำหนดความคลาดเคลื่อนในการประมาณค่า SEE $\leq$ 0.3 และ SEE $\leq$ 0.45 ให้ผลไม่แตกต่างกัน

4.1.6 การควบคุมการใช้ข้อสอบซ้ำและการเหลื่อมล้ำของแบบสอบ (controlling item exposure and test overlap)

Chang (1998) พบว่า วิธีการควบคุมอัตราการใช้ข้อสอบซ้ำด้วยวิธีของ Sympson and Hetter มีประสิทธิภาพมาก และวิธีของ Stocking and Lewis ทำให้มีความปลอดภัยของแบบสอบ เนื่องมาจากการใช้ข้อสอบซ้ำ แต่ละความถูกต้องของการทดสอบโดยที่การสอบแต่ละครั้งควรมีอัตราการใช้ข้อสอบซ้ำไม่เกิน 20 เปอร์เซ็นต์ ของจำนวนข้อสอบที่ผู้สอบได้รับผ่านมา ส่วน Kalohn and Sparay (1998) พบว่า วิธีการคัดเลือกข้อสอบโดยมีอัตราการใช้ข้อสอบซ้ำ 10 และ 20 เปอร์เซ็นต์ มีความถูกต้องเหมือนกันและผู้สอบที่มีความสามารถระดับสูงควรมีอัตราการใช้ข้อสอบซ้ำไม่เกิน 10 เปอร์เซ็นต์ ของจำนวนข้อสอบที่ผู้สอบได้รับในการทดสอบครั้งที่ผ่านมา นอกจากนั้น Revuelta and Ponsoda (1998) ได้ทำการเปรียบเทียบวิธีการควบคุมอัตราการใช้ข้อสอบซ้ำ ผลการศึกษาพบว่า Restricted method ใช้ในการลดอัตราการใช้ข้อสอบซ้ำได้สูงสุด และ Progressive method ลดจำนวนข้อสอบภายในคลังข้อสอบที่ยังไม่ได้ใช้ ซึ่งการควบคุมอัตราการใช้ข้อสอบซ้ำมีจุดมุ่งหมายที่สำคัญ 2 ประการ คือ เพื่อป้องกันอัตราการใช้ข้อสอบซ้ำที่มากเกินไป และเพิ่มอัตราการใช้ข้อสอบที่ยังไม่ได้ใช้ภายในคลังข้อสอบ Chen and Lei (2005) ได้ศึกษาเกี่ยวกับการควบคุมการใช้ข้อสอบซ้ำและการควบคุมการเหลื่อมล้ำของแบบทดสอบ ซึ่งผลการศึกษาชี้ว่า การควบคุมการใช้ข้อสอบซ้ำและการควบคุมการเหลื่อมล้ำของแบบทดสอบ จะต้อง ดำเนินการไปพร้อมกัน เพื่อนำไปใช้ปรับปรุงกระบวนการการทดสอบแบบปรับเหมาะ การควบคุมการใช้ข้อสอบซ้ำจะได้รับการปรับปรุงและความแม่นยำของการประมาณค่าความสามารถจะลดลงเมื่ออัตราการเหลื่อมล้ำระหว่างแบบสอบที่สูงสุด ที่ได้กำหนดไว้ล่วงหน้า

มีความเข้มมากขึ้น สำหรับการศึกษ าในประเทศไทย สิริลักษณ์ เกษรปทุมานันท์ (2549) ได้ศึกษาอัตราการใช้ข้อสอบซ้ำ 10 20 และ 30 เปอร์เซ็นต์ ผลการวิจัยพบว่า ความตรงตามสภาพไม่แตกต่างกัน

4.1.7 การสร้างความสมดุลของเนื้อหา (content balancing)

Leung, Chang and Hua (2003) ได้ศึกษาโดยการ สร้างสถานการณ์จำลองทำการเปรียบเทียบวิธีการสร้างความสมดุลของเนื้อหา 3 วิธี ประกอบด้วย วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่มีเงื่อนไข บังคับ (constrained CAT: CCAT) วิธีโมเดลมัลติโนเมลที่ได้แก้ไขปรับปรุงแล้ว (modified multinomial model: MMM) วิธีการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ที่มีเงื่อนไขบังคับที่ได้ปรับปรุงแก้ไขแล้ว (modified constrained CAT: MCCAT) ภายใต้เงื่อนไขข้อสอบและความยาวแบบสอบ (test length) และอัตราการใช้ข้อสอบซ้ำสูงสุดที่กำหนดไว้ (target maximum exposure rate) ที่แตกต่างกัน ผล ของการจำลองข้อมูลหลายชุดชี้ให้เห็นว่า ไม่มีผลที่เป็นระบบ (systematic) ของวิธีการสร้างสมดุลเนื้อหา ในประสิทธิภาพการวัดและการใช้คลังข้อสอบ อย่างไรก็ดี ระหว่าง 3 วิธีนี้ วิธี MMM มีการใช้ข้อสอบน้อยกว่าวิธีอื่นๆ

จากการศึกษางานวิจัยที่เกี่ยวข้องกับการทดสอบแบบป รับเหมาะด้วยคอมพิวเตอร์ สามารถสรุปดังตารางที่ 2.3 ดังนี้

ตารางที่ 2.3 สรุปประเด็นในงานวิจัยที่เกี่ยวข้องกับการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์

ผู้วิจัย	ประเด็นที่ศึกษา						
	item pool	Starting point	Item selection	Ability estimate	termination criteria	Exposure	content balancing
Weiss(1988)	100-200 ข้อ						
Urry (1977)	อย่างน้อย 100ข้อ $a > 0.8$ $-0.2 \le b \le 0.2$ $c < 3$						
Chen, Ankenmann, and Chang (2000)		FI* FII FIP KL KLP					

ตารางที่ 2.3 (ต่อ)

ผู้วิจัย	ประเด็นที่ศึกษา						
	item pool	Starting point	Item selection	Ability estimate	termination criteria	Exposure	content balancing
Chen and Ankenmann (2004)		FI FIP * KLP * RN					
Hulin and other (1983)			วิธีฟิชเชอร์* วิธีจับคู่ค่าความสามารถของผู้สอบกับค่า θ_{max}				
Segall (2000)			ML Bayesian *				
van der Linden (2005)			STA * WDM spiraling				
Skaggs and Stevenson (1989)			ผู้สอบมากกว่า 2000 คน วิธี ML และ Bayesian ไม่แตกต่าง				
Dodd and Koch (1993)					SEE * กำหนดจำนวนข้อสอบ		
Chang (1998)						Sympson and Hetter * Stocking and Lewis ไม่เกิน 20%	
Kalohn and Sparay (1998)						10% 20% ไม่แตกต่าง กรณีผู้สอบมีความสามารถสูงไม่ควรเกิน 10%	

ตารางที่ 2.3 (ต่อ)

ผู้วิจัย	ประเด็นที่ศึกษา						
	item pool	Starting point	Item selection	Ability estimate	termination criteria	Exposure	content balancing
Revuelta and Ponsoda (1998)						Restricted method * Progressive method	
Leung, Chang and Hua (2003)							CCAT MMM MCCAT ไม่แตกต่าง แต่ MMM ใช้ข้อสอบน้อยสุด
สิริลักษณ์ (2549)		KL* FI		Bayesian	SEE ≤ 0.3 SEE ≤ 0.45 ไม่แตกต่าง	10% 20% 30%	
สายชล (2539)		คัดเลือกข้อสอบความยากปานกลาง	เลือกสารสนเทศข้อที่สูงที่สุด	Bayesian	SEE ≤ 0.3		
กนกวรรณ (2544)		คัดเลือกข้อสอบความยากปานกลาง	เลือกสารสนเทศข้อที่สูงที่สุด	Bayesian	SEE ≤ 0.3	10% 20% 30%	
รังสรรค์ (2540)		ง่าย ปานกลาง ยาก ไม่แตกต่าง	1.วิธีจับคู่ระหว่างความสามารถของผู้สอบกับค่าความยากของข้อสอบ 2. วิธีจับคู่ระหว่างความสามารถของผู้สอบกับค่า	Maximum Likelihood Bayesian*	SEE ≤ 0.3* SEE ≤ 0.4 SEE ≤ 0.5		

ตารางที่ 2.3 (ต่อ)

ผู้วิจัย	ประเด็นที่ศึกษา						
	item pool	Starting point	Item selection	Ability estimate	termination criteria	Exposure	content balancing
			ความสามารถที่ทำให้ข้อสอบมีสานสนเทศสูงสุด 3. วิธีใช้ค่าสารสนเทศของข้อสอบที่สูงที่สุด				
ต่าย (2534)				Conditional Maximum Likelihood และ Bayesian updating ไม่ต่างกัน			

* คือ วิธีการที่ทดสอบแล้วมีประสิทธิภาพมากกว่า

จากตารางที่ 3 จะเห็นว่าในแต่ละประเด็นได้มีนักวัดผลสนใจศึกษาเป็นจำนวนมาก และส่วนใหญ่จะได้ข้อสรุปที่ชัดเจนแล้ว ว่าควรจะเลือกวิธีใดมาใช้ในแต่ละขั้นตอน ซึ่งงานวิจัยที่นำเสนอนี้เป็นการทดสอบปรับเหมาะแบบเอกมิติ อย่างไรก็ ตามสำหรับการทดสอบปรับเหมาะแบบพหุมิติมีผู้ศึกษาในประเด็นต่างๆ ไม่มากนัก ส่วนใหญ่จะเน้นที่การเปรียบเทียบประสิทธิภาพของวิธีการเอกมิติและพหุมิติ ซึ่งจะเสนอรายละเอียดในหัวข้อต่อไป

4.2 งานวิจัยที่เกี่ยวข้องกับการทดสอบแบบปรับเหมาะสำหรับโมเดลการตอบสนองข้อสอบแบบพหุมิติ

สำหรับการทดสอบแบบปรับเหมาะสำหรับโมเดลการตอบสนองข้อสอบแบบพหุมิติ (Multidimensional computerized adaptive test: MCAT) มีหลักการที่คล้ายคลึงกับทดสอบแบบปรับเหมาะสำหรับโมเดลการตอบสนองข้อสอบแบบเอกมิติ (Unidimensional computerized adaptive test: UCAT) โดยมีองค์ประกอบของการทดสอบที่คล้ายคลึงกัน ดังนั้นงานวิจัยที่มีผู้ศึกษาในปัจจุบัน จึงเกี่ยวกับการเปรียบเทียบประสิทธิภาพของทั้งสองวิธี และการเปรียบเทียบวิธีการคัดเลือกข้อสอบและการประมาณความสามารถ รวมถึงอัตราการใช้ข้อสอบซ้ำ ซึ่งมีรายละเอียดดังนี้

4.2.1 การเปรียบเทียบ MCAT และ UCAT

Bloxom และ Vale (1987) ได้ทำการศึกษาอย่างเป็นทางการครั้งแรก และได้ขยายวิธีการของการทดสอบแบบปรับเหมาะที่มีลักษณะเอกมิติไปสู่ลักษณะพหุมิติ โดยเขาได้กล่าวไว้ว่า การสรุปอ้างอิงหลายตัวแปร (Multivariate) ของวิธีการให้คะแนนเอกมิติของโมเดลกา รตอบสนองข้อสอบ (unidimensional IRT) ให้ผลที่ดีกว่าในเรื่องเวลาในการคำนวณค่าต่าง ๆ และเพื่อหลีกเลี่ยงการคำนวณที่เพิ่มมากขึ้น ซึ่งสัมพันธ์กับวิธีการคำนวณและการรวมกันของตัวเลข เขาได้เสนอวิธีการให้คะแนนบนฐานของการขยายออกของพหุตัวแปร (multivariate) จากแนวคิดการจัดลำดับของ Owen ตลอดจนชุดของการประมาณค่าแบบปกติ การขยายออกของพหุตัวแปรจะแสดงให้เห็นรูปแบบที่ใกล้ชิดสำหรับการประมาณค่าความสามารถ นอกจากนี้ Haley (2006) ได้ศึกษาเกี่ยวกับ ความแม่นยำและประสิทธิภาพการวัดในการทดสอบปรับเหมาะแบบพหุมิติของการทำหน้าที่ทางกายภาพขอ งการประเมินเด็กที่ไร้ความสามารถ โดยวัตถุประสงค์ของการศึกษาครั้งนี้ คือ การเปรียบเทียบความแม่นยำและประสิทธิภาพการวัดในการทดสอบปรับเหมาะแบบพหุมิติ (MCAT) และเอกมิติ (UCAT) โดยใช้คลังข้อสอบจาก 2 ทักษะ คือ การเคลื่อนที่ได้ง่าย และการดูแลตนเอง ผลการเปรียบเทียบ พบว่า MCAT มีระดับความแม่นยำและมีประสิทธิภาพสูงกว่าการแยกทักษะในวิธีการแบบ UCAT โดยเฉพาะอย่างยิ่ง เมื่อโดเมนย่อยมีความสัมพันธ์กันสูง ซึ่งงานวิจัยดังกล่าวมีความสัมพันธ์กับงานของ Wang and Chen (2004) ที่ศึกษาเกี่ยวกับการวัดประสิทธิภาพการทดสอบปรับ เหมาะแบบพหุมิติและการนำไปปฏิบัติ เขาได้ศึกษาโดยการจำลองข้อมูลเพื่อเปรียบเทียบประสิทธิภาพการวัดของการทดสอบปรับเหมาะแบบพหุมิติ (unidimensional adaptive testing) ด้วยการทดสอบปรับเหมาะแบบเอกมิติ (unidimensional adaptive testing) และการบริหารการสอบแบบสุ่ม (random administration) ผลลัพธ์แสดงให้เห็นว่าการทดสอบปรับเหมาะแบบพหุมิติมีประสิทธิภาพสูงกว่าอีกสองวิธี ในขณะที่ Petersen, Groenvold, and Aaronson (2006) ได้ใช้การทดสอ บปรับเหมาะโดยใช้คอมพิวเตอร์กับการทดสอบแบบหลายมิติ โดยศึกษาเกี่ยวกับความสัมพันธ์ของสุขภาพและคุณภาพชีวิต health-related quality of life (HRQOL) โดยข้อคำถามจะถูกปรับไปตามผู้ป่วยแต่ละคน และในขณะเดียวกันก็จะมีการเปรียบเทียบระหว่างผู้ป่วยแต่ละคน และจะให้ผลที่ดียิ่งขึ้นถ้าใช้ computerized adaptive testing (CAT) แต่โดยปกติ CAT จะวิเคราะห์ในแต่ละครั้งเพียงมิติเดียวแต่สำหรับ HRQOL มีหลายมิติที่มีความสัมพันธ์กันสูง ดังนั้นการใช้แบบสอบแบบปรับเหมาะหลายมิติด้วยคอมพิวเตอร์ (multidimensional CAT)อาจจะมีประโยชน์ในการปรับปรุงประสิทธิภาพการวัด ดังนั้นเขาจึงได้มีการ ประเมิน Multidimensional CAT ใน 3 สเกล จาก EORTC QLQ-C30 ได้แก่ ปัจจัย ด้านกายภาพ ปัจจัยด้านอารมณ์ และสเกล

ความอ่อนล้า ผลการวิจัยพบว่า Multidimensional CAT ให้คะแนนใน 3 มิติ ด้วยข้อคำถาม 5-7 ข้อ ซึ่งมีความใกล้เคียงกับการใช้ข้อคำถาม 12 ข้อ และไม่สูญเสียความแม่นยำในการวัด ซึ่งข้อค้นพบนี้สรุปได้ว่า Multidimensional CAT มีความสำคัญในการปรับปรุงการวัดให้มีความแม่นยำและมีประสิทธิภาพ ทั้งนี้วิธีดังกล่าว เป็นการประยุกต์จากแนวคิดของ Segall (1996) โดยปรับให้เป็นการให้คะแนนแบบหลายค่า (polytomous ordinal items) จากการใช้แบบสอบถามที่มี 4 สเกล นอกจากนั้น Luecht (1996) ได้ศึกษาประโยชน์ของการประยุกต์ใช้การคัดเลือกข้อสอบแบบพหุมิติและเทคนิคการให้คะแนน ในบริบทของการให้ใบอนุญาตหรือให้การรับรอง โดยที่ความซับซ้อนของเนื้อหามีเงื่อนไขบังคับ เขาได้เปรียบเทียบความเที่ยงของวิธี ML ใน multidimensional (Segall, 1996) กับ unidimensional CAT ผลลัพธ์แสดงให้เห็นว่า การทดสอบ MCAT ที่มีจำนวนข้อสอบน้อยกว่าและมีการกำหนดเนื้อหาให้ผลสัมฤทธิ์เกี่ยวกับความเที่ยงของคะแนนย่อยเทียบเท่ากับ UCAT ที่มีจำนวนข้อสอบมากกว่าซึ่งการประมาณค่าความสามารถมีความสะดวกและไม่เสียเวลาเนื่องจากความยาวของข้อสอบสั้นลงซึ่งสอดคล้องกับข้อค้นพบของ Segall ที่ความยาวข้อสอบสั้นลง 25% -40%

4.2.2 การคัดเลือกข้อสอบและการประมาณค่าความสามารถ

Tam (1992) ได้พัฒนาวิธี Maximum likelihood แบบทำซ้ำในการประมาณค่าความสามารถสำหรับโมเดลโค้งปกติสะสมแบบ 2 มิติ ซึ่ง Tam ได้ประเมินวิธีการนี้กับการใช้ในรูปแบบต่างๆ เช่น เกณฑ์ที่มีความแม่นยำ สารสนเทศของแบบสอบ และเวลาในการคำนวณ วิธีการเลือกข้อคำถามของ Tam ทำหน้าที่เป็นคลังข้อสอบที่สมบูรณ์ซึ่งความยากของข้อสอบจะถูกจับคู่กับระดับความสามารถในปัจจุบันของผู้สอบ นอกจากนี้ Segall (1996) ได้ศึกษาเพิ่มเติมจากงานวิจัยก่อน หน้านี้ของ Bloxom, Vale และ Tam โดยกล่าวถึงวิธีการที่ใช้ทฤษฎีเป็นฐาน สำหรับการเลือกข้อคำถาม ซึ่งเป็นการรวมความรู้ที่ผ่านมาของการแจกแจงร่วมของความสามารถ Segall ได้เสนอวิธี maximum likelihood และ Bayesian สำหรับการคัดเลือกข้อสอบและการให้คะแนนสำหรับแบบสอบแบบปรับเหมาะหลายมิติ (multidimensional adaptive test) ความสามารถของคอมพิวเตอร์สามารถสนับสนุนการคำนวณซึ่งสัมพันธ์กับวิธีการของตัวเลขที่ซ้ำบ่อยๆ การจัดทำวิธีการให้คะแนนแบบประมาณค่าโดยคำแนะนำของ Bloxom และ Vale (1987) ประโยชน์ของวิธี Bayesian จะประเมินโดยการจำลองข้อมูลบนพื้นฐานของการทดสอบที่มีสเกลใหญ่และมีผลกระทบสูง โดยเขาใช้การทดสอบ Armed Services Vocational Aptitude Battery ในเวอร์ชั่นของของการทดสอบแบบปรับเหมาะโดยใช้คอมพิวเตอร์ นอกจากนี้ Segall ได้แสดงให้เห็นว่า คลังข้อสอบที่สร้างขึ้นจริง แบบสอบ แบบปรับเหมาะหลายมิติ สามารถให้ความถูกต้องแม่นยำที่เท่ากับหรือสูงกว่า โดยใช้ข้อคำถามประมาณ 1 ใน 3 ซึ่งน้อยกว่า ความต้องการจากการทดสอบปรับเหมาะแบบ

เอกมิติ นอกจากนี้ van der Linden (1999) ได้เสนอวิธี MCAT โดยตั้งใจที่จะทำให้องค์ประกอบการวัดมีความแม่นยำ และเหมาะสม โดยที่องค์ประกอบที่น่าสนใจคือ ฟังก์ชันเชิงเส้นของความสามารถแฝง ซึ่งน้ำหนักองค์ประกอบมีความสัมพันธ์กับแต่ละมิติถูกกำหนดไว้ ก่อน โดยยึดเกณฑ์ภายนอกที่ผู้พัฒนาได้กำหนด การดำเนินการประมาณค่าความสามารถใช้วิธี ML และการคัดเลือกข้อสอบใช้วิธีโดยยึดเกณฑ์ความแปรปรวนของความคลาดเคลื่อนน้อยสุด (minimum error variance) ความแปรปรวนของความคลาดเคลื่อน (หรือการสุ่ม) ขององค์ประกอบการวัดได้มาจากผลรวมเชิงเส้นขององค์ประกอบจากส่วนกลับของเมทริกซ์สารสนเทศของ Fisher (Fisher-information matrix) van der Linden ได้แสดงให้เห็นว่า สำหรับคลังข้อสอบที่มีข้อสอบแบบ 2 มิติ นั้นจำนวนข้อสอบแบบปรับเหมาะจำนวน50 ข้อ ให้ความแม่นยำของการวัดที่ใกล้เคียงกันข้ามช่องว่างของความสามารถสำหรับข้อสอบสั้นๆ (10 และ 30 ข้อ) การประมาณค่าแบบML มีแนวโน้มที่จะเกิดความคลาดเคลื่อนและไม่มีประสิทธิภาพโดยเขาได้เปรียบเทียบวิธีการประมาณค่าด้วยวิธี Bayesian กับวิธี Maximum likelihood ซึ่งพบว่า วิธี Bayesian ให้สารสนเทศมากกว่าการใช้วิธี Maximum likelihood ซึ่งสิ่งนี้จะถูกต้องเมื่อ มิติต่างๆ อยู่ภายใต้พื้นฐานของการ ตอบสนองข้อสอบที่มี ความสัมพันธ์กันและการแจกแจงร่วมของความสามารถแฝงนั้นทราบค่า หรือประมาณค่าได้อย่างยุติธรรม

Wang และ Chang (2011) ได้ศึกษาเปรียบเทียบวิธีการคัดเลือกข้อสอบสำหรับการทดสอบปรับเหมาะแบบพหุมิติ โดยศึกษาจากการจำลองข้อมูลใน 2 มิติ ในลักษณะของ within-item multidimensionality โดยเขาได้เปรียบเทียบวิธีการคัดเลือกข้อสอบระหว่าง D-optimality (fisher information), KL information index (KI), continuous entropy method (CEM), และ mutual information (MI) ซึ่งผลวิจัยพบว่า mutual information ไม่เพียงแต่จะปรับปรุงความแม่นยำในการประมาณค่าแต่ยังให้ค่าคว ามคลาดเคลื่อนต่ำที่สุด Diao, Q., and Reckase, M. (2009) ศึกษาเปรียบเทียบวิธีการคัดเลือกข้อสอบสำหรับการทดสอบปรับเหมาะแบบพหุมิติ ด้วยวิธี D-optimality, A-optimality และ KL ผลวิจัยพบว่า ทั้งสามวิธีให้ผลไม่แตกต่างกัน

นอกจากนั้น ยังมีผู้พัฒนาวิธีการคัดเลือกข้ อสอบให้สามารถปรับใช้กับการทดสอบปรับเหมาะแบบ พหุมิติจำนวนมากขึ้น โดยเป็นการประยุกต์วิธีการในการทดสอบปรับเหมาะแบบเอกมิติ แต่มีความยุ่งยากซับซ้อนมากกว่า อันเนื่องมาจากจำนวนมิติที่มากขึ้น การคำนวณส่วนใหญ่จะอยู่ในรูปเวคเตอร์ และเมทริกซ์ ซึ่งตัวอย่างวิ ธีการคัดเลือกข้อสอบ ได้แก่ Fisher information matrix (Segall, 1996; Mulder & van der Linden, 2009) maximum Kullback-Leibler (KL) information index (Veldkamp & van der Linden, 2002) minimum trace of the inverse Fisher information matrix (van der Linden, 1999), maximum KL

distance between two subsequent posteriors (Mulder & van der Linden, 2010), maximum mutual information method (Mulder & van der Linden, 2010) และ KL information with Bayesian update method (Wang & Chang, 2010) วัตถุประสงค์ของการคัดเลือกข้อส อบแบบ MCAT คือ การเลือกข้อสอบอย่างเป็นลำดับและเป็นการประมาณค่าเวคเตอร์ความสามารถ p มิติ $\theta_i = (\theta_{i1},\ldots, \theta_{ip})^T$ แบบเวียนซ้ำ กฎการคัดเลือกข้อสอบโดยทั่วไปจะมีแนวคิดเกี่ยวกับการให้สารสนเทศสูงสุดเกี่ยวกับตำแหน่งความสามารถของผู้สอบบนระบบพิกัด θ หรือ การลดความคลาดเคลื่อนในการประมาณค่าความสามารถให้น้อยที่สุด (Reckase, 2009)

4.2.3 อัตราการใช้ข้อสอบซ้ำ

Lee, IP and Fuh (2008) ได้ศึกษายุทธวิธีในการควบคุมการใช้ข้อสอบซ้ำในการทดสอบปรับเหมาะแบบพหุมิติ โดยวิธีการที่เขาเสนอใช้พื้นฐานของการแบ่งชั้นในความสอดคล้องด้วยฟังก์ชันของเวคเตอร์พารามิเตอร์อำนาจจำแนก ซึ่งสามารถนำไปใช้ในการคำนวณ การศึกษาครั้งนี้เขาเปรียบเทียบอัตราการใช้ข้อสอบซ้ำโดยวิธี ASTR และ D-optimality โดยใช้เกณฑ์การคัดเลือกข้อสอบด้วยวิธี Match-b เงื่อนไขการเปรียบเทียบประกอบด้วย 1) ความยาวของข้อสอบ 30 ข้อ และ 60 ข้อ 2) วิธีการ คือ ASTR, D-optimality และ Match-b 3) ความแตกต่างของค่าความสัมพันธ์ระหว่าง θ_1 และ θ_2 ประกอบด้วย ρ = 0, .3 และ .6 ซึ่งพบว่า การเปรียบเทียบด้วยการคัดเลือกข้อสอบวิธี Match-b ผลการวิจัยชี้ว่า วิธีการ ASTR ค่อนข้างมีประสิทธิภาพ และข้อสอบที่มีความยาวมากกว่าก็จะมีประสิทธิภาพมากกว่า โดยพิจารณาจากค่า Mean square error (MSE) ซึ่งมีค่าน้อยกว่า เพื่อพิจารณาที่ความสัมพันธ์ของ θ_1 และ θ_2 เมื่อไม่สนใจวิธีการพบว่า ความสัมพันธ์ในระดับสูงจะมีประสิทธิภาพน้อยกว่า ในทางตรงข้าม ค่า ρ ไม่มีผลต่อการประมาณค่าความสามารถ θ เมื่อแบบสอบมีความยาวมากกว่า 60 ข้อ

จากการศึกษาเอกสารงานวิจัยที่เกี่ยวข้องทั้งหมดสามารถสรุปได้ดังตารางที่ 2.4 ดังนี้

ตารางที่ 2.4 สรุปประเด็นในงานวิจัยที่เกี่ยวข้องกับการทดสอบปรับเหมาะด้วยคอมพิวเตอร์แบบพหุมิติ

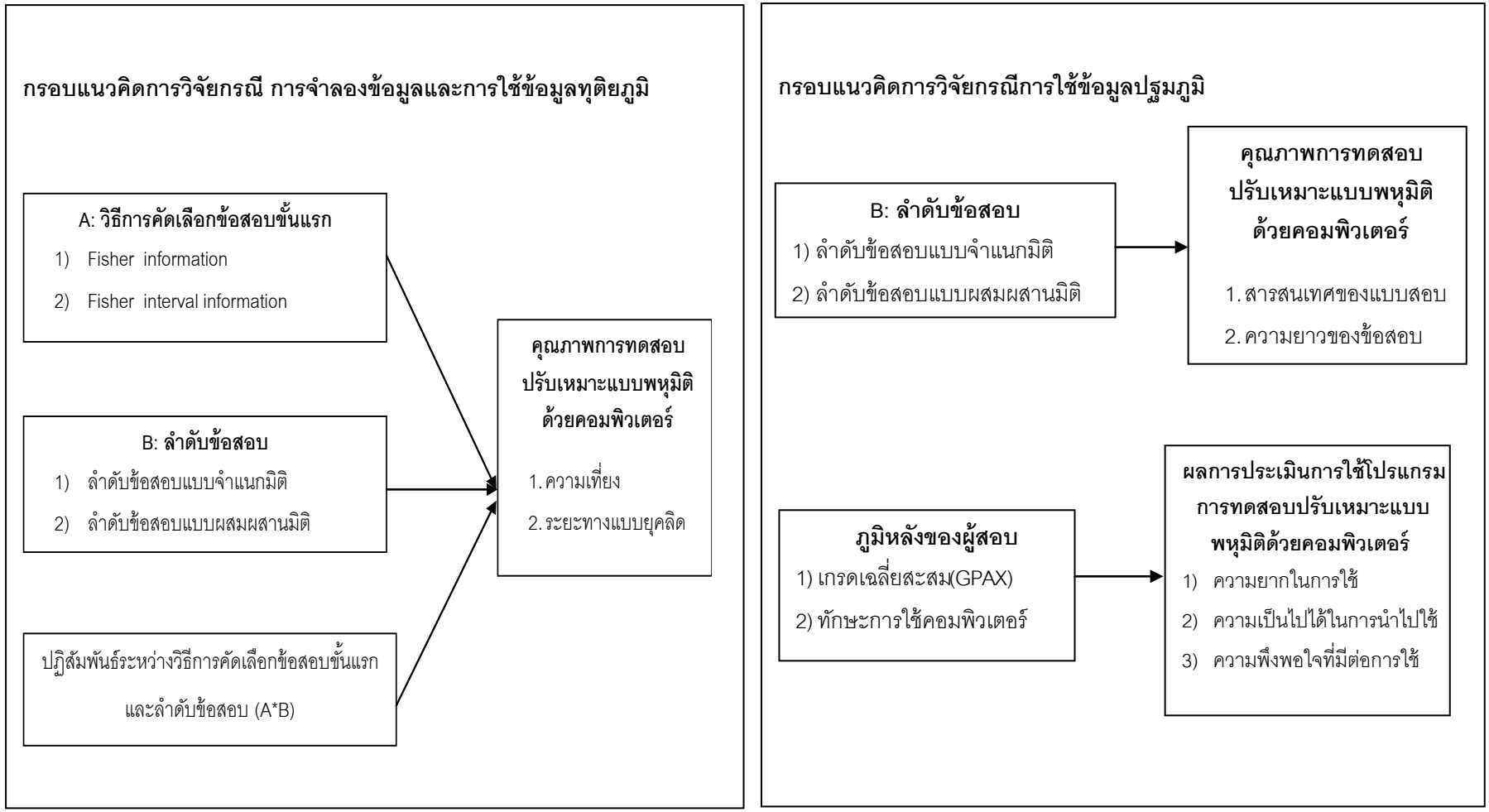
ผู้วิจัย	ประเด็นที่ศึกษา					
	Compare UCAT, MCAT	Starting point	Item selection	Ability estimate	termination criteria	Exposure
Bloxom and Vale (1987) ; Haley (2006); Wang and Chen (2004); Petersen, Groenvold, and Aaronson (2006) ; Luecht (1996)	MCAT มีประสิทธิภาพมากกว่า UCAT					
Tam (1992)			Maximum likelihood แบบทำซ้ำ			
Segall (1996)			Maximum likelihood และ Bayesian*			
van der Linden (1999)			minimum error variance	Maximum likelihood และ Bayesian*		
Lee, IP and Fuh (2008)			Match-b			ASTR * D-optimality
Diao, Q., and Reckase, M. (2009)			D-optimality A-optimality KL	MLE Bayesian*		
Wang & Chang (2011)			D-optimality KI CEM MI*			

* คือ วิธีการที่ทดสอบแล้วมีประสิทธิภาพมากกว่า

จากตารางที่ 2.4 จะเห็นว่าสำหรับการตอบสนองข้อสอบแบบพหุมิติยังมีหลายประเด็นที่ยังไม่มีความชัดเจน และยังไม่มีผู้ศึกษาวิธีการที่มีลักษณะเฉพาะสำหรับการทดสอบปรับเหมาะแบบพหุมิติ ไม่ว่าจะเป็นการเริ่มต้นการทดสอบ ก ารยุติการสอบ ทั้งนี้รวมถึง ลำดับ ข้อสอบ (item order) ซึ่งควรจะมีการศึกษาวิธีการที่มีความเหมาะสมกับการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ต่อไป นอกจากนั้น จะเห็นว่า ยังไม่มีงานวิจัยใดที่ศึกษาเปรียบเทียบวิธีการคัดเลือกข้อสอบขั้นแรก จาก 2 แนวคิดหลัก Local information และ Global information หรือ สารสนเทศของ Fisher และสารสนเทศของ Kullback-Leibler ดังนั้นเพื่อให้มีความชัดเจนในบริบทของการทดสอบปรับเหมาะแบบพหุมิติ ผู้วิจัยจึงสนใจเปรียบเทียบวิธีการคัดเลือกข้อสอบขั้นแรกด้วยวิธี Fisher information และ Kullback-Leiber information นอกจากนั้น ผู้วิจัยสนใจที่จะศึกษาเกี่ยวกับ ความแตกต่างของลำดับข้อสอบ ระหว่างลำดับข้อสอบ แบบจำแนกมิติ และ ลำดับข้อสอบแบบผสมผสานมิติ รวมถึงการนำทฤษฎีเหล่านี้ไปประยุกต์ใช้ในทางปฏิบัติ และศึกษาตัวแปรต่างๆ เกี่ยวกับผู้สอบ

จากการทบทวน เอกสารและ งานวิจัยที่ผ่าน มา ทำให้ผู้วิจัยสามารถสรุปกรอบแนวคิดในการวิจัยได้ดังตารางที่ 2.5

กรอบแนวคิดในการวิจัยจากการทบทวนงานวิจัยที่เกี่ยวข้อง



แผนภาพที่ 2.5 กรอบแนวคิดในการวิจัยจากการทบทวนงานวิจัยที่เกี่ยวข้อง

บทที่ 3

วิธีดำเนินการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการคัดเลือกข้อสอบขั้นแรกและ ลำดับข้อสอบที่มีต่อคุณภาพของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบพหุมิติ รวมถึงศึกษาปฏิสัมพันธ์ระหว่างวิธีการคัดเลือกข้อสอบขั้นแรกและลำดับข้อสอบที่มีต่อคุณภาพของการทดสอบ และเพื่อศึกษาผลการประเมิน โปรแกรม การทดสอบความสามารถด้านวิทยาศาสตร์ด้วยวิธีการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์ โดยการศึกษาครั้งนี้ แบ่งออกเป็น 3 ระยะ ตามข้อมูลที่นำมาศึกษา คือ ระยะที่ 1 การศึกษาจากการจำลองข้อมูล ระยะ ที่ 2 การศึกษาจากข้อมูลทุติยภูมิ จากการทดสอบความถนัดทางวิชาชีพและวิชาการ (professional and academic aptitude test: PAT) ด้านวิทยาศาสตร์ (PAT2) ของสถาบันทดสอบทางการศึกษาแห่งชาติ(องค์การมหาชน) และ ระยะที่ 3 การศึกษาจากข้อมูลปฐมภูมิ ซึ่งเป็นการทดลองใช้โปรแกรมจริงกับนักเรียนชั้นมัธยมศึกษาปีที่ 5 เพื่อศึกษาผลการประเมินการใช้โปรแกรมการทดสอบแบบปรับเหมาะการศึกษาทั้ง 3 ระยะ จากแหล่งข้อมูลที่แตกต่างกันจะช่วยยืนยันผลการทดลองได้ดียิ่งขึ้น และยังช่วยลดข้อจำกัดที่เกิดขึ้น โดยมีรายละเอียดของวิธีการดำเนินการวิจัยในแต่ละระยะ ดังนี้

ระยะที่ 1 การศึกษาจากการจำลองข้อมูล

1.1 การจำลองข้อมูล

จากการศึกษาที่ผ่านมาพบว่ากลุ่มผู้สอบประมาณ 2,000 คนขึ้นไปเหมาะสม สำหรับการศึกษาโดยโมเดล MIRT (Ackerman, 1994; Reckase, 1995; Yon, 2006) สำหรับการศึกษาครั้งนี้ จะใช้การจำลองข้อมูลการตอบสนองข้อสอบจากผู้สอบ 10,000 คน และคลังข้อสอบจำนวน 300 ข้อ โดยแบ่งเป็นการทดลองออกเป็น 4 โมเดลจำแนกตามวิธีการคัดเลือกข้อสอบ ขั้นแรก 2 วิธี และลำดับข้อสอบ 2 รูปแบบ สำหรับรูปแบบข้อสอบที่นำมาใช้ในการวิจัยครั้งนี้เป็นข้อสอบที่วัดความสามารถ 2 มิติ ที่มีความสัมพันธ์กันปานกลาง ซึ่งมีลักษณะพหุมิติระหว่างข้อสอบ (between-item multidimensionality) นั่นคือ ข้อสอบหนึ่งข้อ วัดความสามารถ 1 มิติ ทั้งนี้ข้อสอบแต่ละข้อต้องมีความสัมพันธ์กัน โดยกำหนดพารามิเตอร์อำนาจจำแนกทั้ง 2 มิติ ให้มีลักษณะเป็นโค้งปกติตัดปลาย (truncated normal distribution) โดยมีค่าเฉลี่ยเป็น 1 และมีค่าส่วนเบี่ยงเบนมาตรฐานเป็น 0.1 โดยตัดส่วนที่ติดลบออก เนื่องจากหากอำนาจจำแนกมีค่าติดลบจะทำให้การประมาณค่าความสามารถไม่มีความแม่นยำ พร้อมทั้งกำหนดให้พารามิเตอร์ความยากมีการแจกแจงแบบเอกรูป

(uniform distribution) อยู่ในช่วง -3 ถึง 3 และพารามิเตอร์โอกาสการเดามีการแจกแจงแบบเอกรูปเช่นกัน แต่จำกัดขอบเขตอยู่ระหว่าง 0 ถึง 0.1

1.2 เงื่อนไขที่ใช้ในการจำลองข้อมูล

เงื่อนไขที่ใช้ในการจำลองข้อมูลจะพิจารณาตาม ตัวแปรที่ใช้ในการวิจัย ซึ่งได้แก่ วิธีการคัดเลือกข้อสอบขั้นแรก 2 วิธี และลำดับข้อสอบ 2 รูปแบบ โดยผู้วิจัยวางแผนการวิจัยแบบแฟคตอเรียล 2 ปัจจัย 2 ระดับ 2 ระดับ (2 x 2 factorial designs) ซึ่งมีเงื่อนไขการทดลองทั้งหมด 4 เงื่อนไขดังนี้ รายละเอียดดังตารางที่ 3.1

ตารางที่ 3.1 เงื่อนไขและจำนวนกลุ่มตัวอย่างจากการจำลองข้อมูล

เงื่อนไข	จำนวนกลุ่มตัวอย่าง	วิธีการคัดเลือกข้อสอบขั้นแรก	ลำดับข้อสอบ
1	10,000 คน	FI	จำแนกมิติ
2	10,000 คน	FI	ผสมผสานมิติ
3	10,000 คน	KL	จำแนกมิติ
4	10,000 คน	KL	ผสมผสานมิติ

1.3 ขั้นตอนการทดสอบปรับเหมาะแบบพหุมิติกรณีการจำลองข้อมูล

1) วิธีการคัดเลือกข้อสอบขั้นแรก ใช้ 2 วิธี (ศึกษารายละเอียดเพิ่มเติมได้จากเอกสารและงานวิจัยที่เกี่ยวข้อง หน้า 69-71)

1.1) Fisher information (FI)

1.2) Kullback-Leibler information (KL)

จากการศึกษาของ Chang และ Ying (1996) ได้พยายามหาจุดตัดของจำนวนข้อสอบว่าจำนวนเท่าใดน้อย และจำนวนเท่าใดมาก เพื่อจะได้ใช้วิธีการคัดเลือกข้อสอบที่มีความเหมาะสม โดย Chang และ Ying ได้ใช้จำนวนข้อสอบ 5 ข้อ เพื่อศึกษาว่า KL information เหมาะสมกับจำนวนข้อสอบน้อย ดังนั้นการวิจัยครั้งนี้จึง พิจารณาที่ตำแหน่งการสอบข้อที่ 5 เพื่อเปรียบเทียบวิธีการคัดเลือกข้อสอบขั้นแรก

2) การประมาณค่าความสามารถ ใช้วิธีของ Bayesian โดยกำหนดค่าเริ่มต้น (prior) ให้มีค่าเฉลี่ยเป็น (0,0) เนื่องจาก ผู้วิจัยไม่มีข้อมูลใดๆ ของผู้สอบมาก่อน และกำหนดให้ co-variance matrix เป็น $\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} x100$ เพื่อให้ค่า prior ไม่มีผลต่อการคัดเลือกข้อสอบ เนื่องจากการคำนวณจะนำอินเวอร์สของprior มาใช้ในสมการ และใช้การวิเคราะห์วนซ้ำด้วยวิธีของ Newton-Raphson

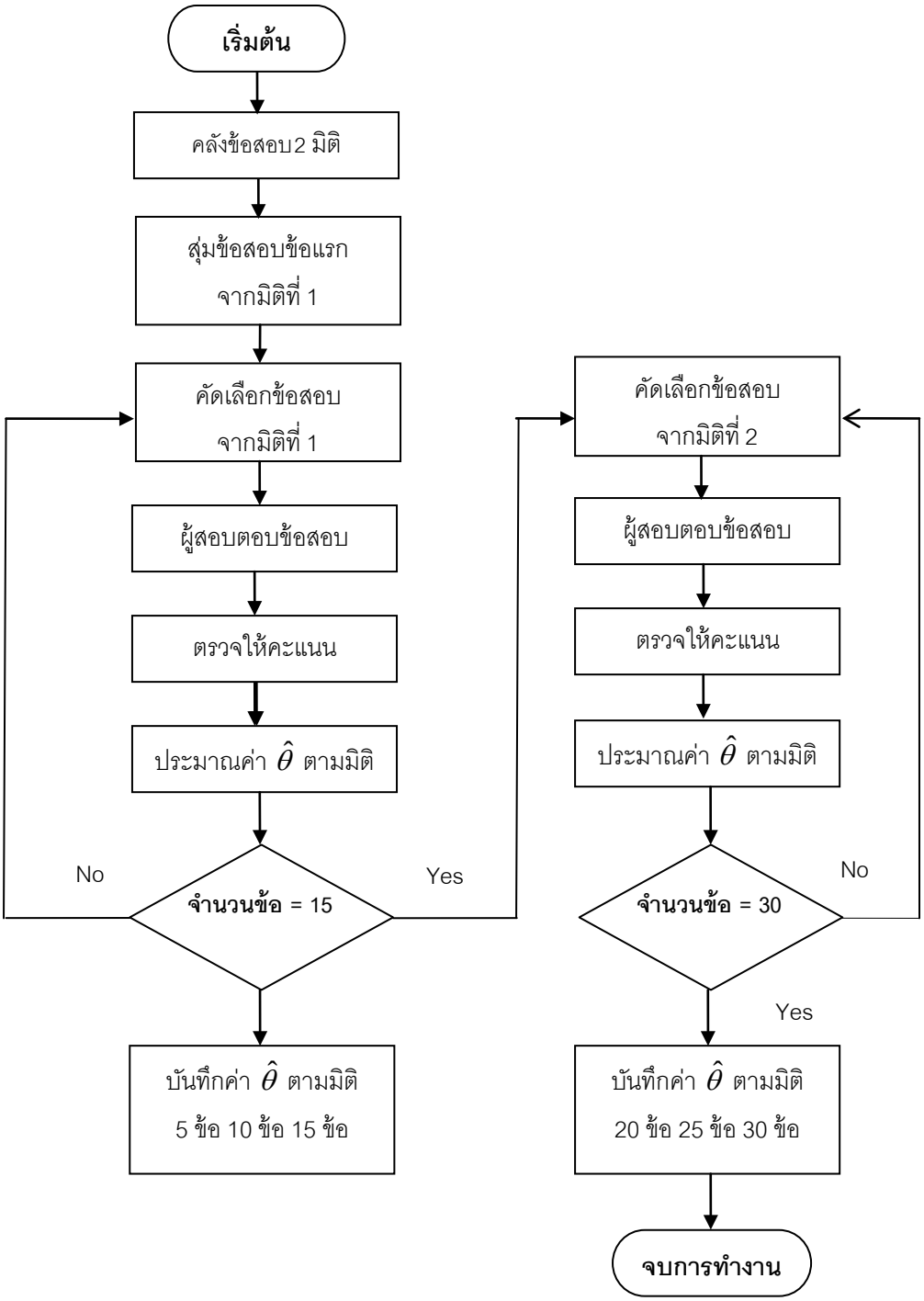
3) การยุติการสอบ ใช้การกำหนดจำนวนข้อสอบ 30 ข้อ

4) รูปแบบของลำดับข้อสอบ แบ่งเป็น 2 รูปแบบ ดังนี้

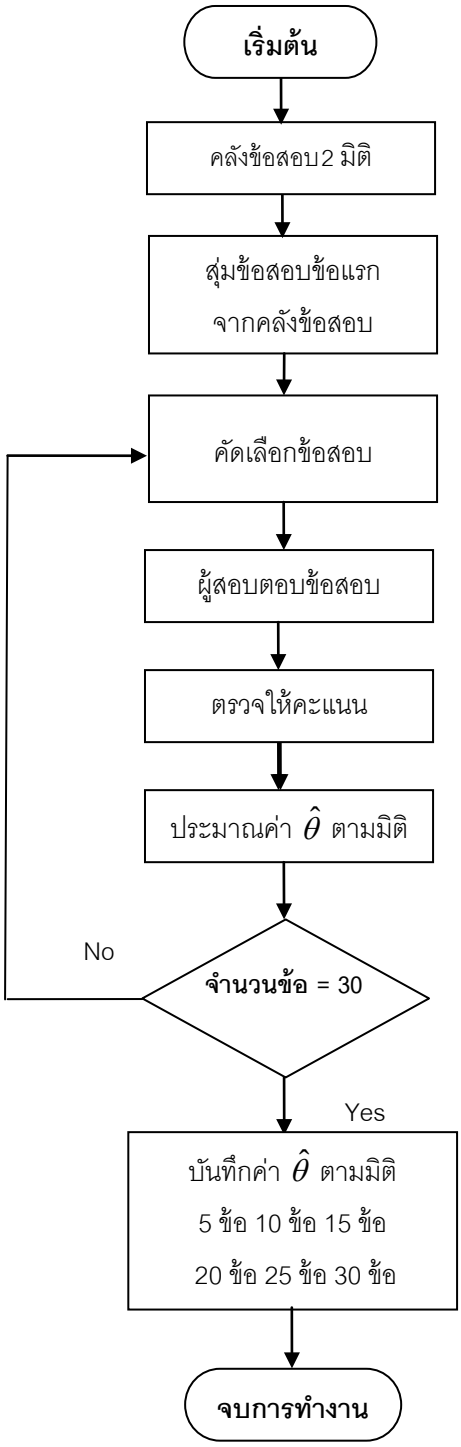
4.1) ลำดับข้อสอบ แบบจำแนกมิติ คือ การทดสอบที่คัดเลือกข้อสอบจำแนกตามมิติข้อสอบก่อนหลัง โดยในที่นี้จะจำลองให้คัด เลือกข้อสอบมิติที่ 1 จำนวน 15 ข้อ ตามด้วยคัดเลือกข้อสอบมิติที่ 2 จำนวน 15 ข้อ

4.2) ลำดับข้อสอบแบบผสมผสานมิติ คือ การทดสอบที่คัดเลือกข้อสอบโดยไม่คำนึงถึง มิติของข้อสอบ ข้อสอบข้อใดที่มีสารสนเทศสูงสุดจะถูกคัดเลือกขึ้นมาก่อน

5) ขั้นตอนการทดสอบปรับเหมาะแบบพหุมิติ กรณี ลำดับข้อสอบ แบบจำแนกมิติ ผู้วิจัยจะเริ่มจากการสุ่มข้อสอบข้อแรกที่มีความยากปานกลางจากคลังข้อสอบมิติที่ 1 และคัดเลือกข้อสอบข้อถัดไป โดยเลือกเฉพาะข้อสอบในมิติที่ 1 ที่มีสารสนเทศข้อสอบสูงสุด จากนั้นตรวจให้คะแนนและประมาณค่าความสามารถของผู้สอบ จนกระทั้งผู้สอบทำข้ อสอบครบ 15 ข้อ โปรแกรมจะคัดเลือกข้อสอบจากมิติที่ 2 โดยเลือกข้อสอบที่ให้สารสนเทศข้อสอบสูงสุดจากนั้นตรวจให้คะแนนและประมาณค่าความสามารถของผู้สอบ จนกระทั้งผู้สอบทำข้อสอบครบ30 ข้อ จะหยุดการทดสอบ ทั้งนี้จะบันทึกค่าความสามารถของผู้สอบในตำแหน่งที่ผู้สอบทำข้อสอบ 5 ข้อ 10 ข้อ 15 ข้อ 20 ข้อ 25 ข้อ และ 30 ข้อ รายละเอียดดังแสดงไว้ในแผนภาพที่ 3.1 สำหรับกรณีลำดับข้อสอบ แบบผสมผสานมิติ จะมีขั้นตอนต่างๆ คล้ายคลึงกับรูปแบบแรก แต่จะไม่กำหนดให้คัดเลือกข้อสอบในมิติใดมิติหนึ่ง การคัดเลือกข้อสอบจะเป็นอิสระผสมผสานกันทั้ง 2 มิติ โดยจะพิจารณาคัดเลือกข้อสอบที่มีค่าสารสนเทศข้อสอบสูงสุดไม่ว่าข้อสอบข้อนั้นจะอยู่ในมิติใด จากนั้นตรวจให้คะแนนและประมาณค่าความสามารถของผู้สอบ จนกระทั้งผู้สอบทำข้อสอบครบ 30 ข้อ ในกรณีนี้ข้อสอบที่คัดเลือกมาอาจจะมีจำนวนข้อสอบไม่เท่ากันใน 2 มิติ รายละเอียดดังแสดงไว้ ในแผนภาพที่ 3.2



แผนภาพที่ 3.1 ผังแสดงขั้นตอนการทำงานของโปรแกรมการทดสอบแบบปรับเหมาะแบบพหุมิติ (ลำดับข้อสอบแบบจำแนกมิติ)



แผนภาพที่ 3.2 ผังแสดงขั้นตอนการทำงานของโปรแกรมการทดสอบแบบปรับเหมาะแบบพหุมิติ (ลำดับข้อสอบแบบผสมผสานมิติ)

1.4 การวิเคราะห์คุณภาพการทดสอบปรับเหมาะแบบพหุมิติจากการจำลองข้อมูล

การศึกษาคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติ จะพิจารณาจากค่าความเที่ยง และค่าระยะทางแบบยุคลิด ซึ่งมีรายละเอียดดังนี้

1) ค่าความเที่ยง (reliability)

ค่าความเที่ยงคือ กำลังสองของความสัมพันธ์ระหว่างค่าความสามารถของผู้สอบที่ได้จากการประมาณค่าจากการทดสอบแบบปรับเหมาะ กับค่าความสามารถ ที่แท้จริงของผู้สอบ โดยที่ ค่าความเที่ยงสูง หมายถึง วิธีการทดสอบปรับเหมาะแบบพหุมิติมีคุณภาพสูง ซึ่งค่าความเที่ยงหาได้จากสูตร ดังนี้

$$r^2 = r^2_{xy}$$

$$r_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left(n\sum x^2 - \left(\sum x\right)^2\right)\left(n\sum y^2 - \left(\sum y^2\right)\right)}}$$

เมื่อ r^2 คือ ความเที่ยงของการวัด

r_{xy} คือ ความสัมพันธ์ระหว่างค่าความสามารถ

n คือ จำนวนผู้สอบ

x คือ ค่าความสามารถของผู้สอบที่ได้จากการประมาณค่า

y คือ ค่าความสามารถที่แท้จริงของผู้สอบ

การวิจัยครั้งนี้จะเ ปรียบเทียบความแตกต่างของค่าความเที่ยงโดยพิจารณาที่ค่าสัมประสิทธิ์สหสัมพันธ์ โดยการทดสอบความแตกต่างของค่าสัมประสิทธิ์สหสัมพันธ์ จะต้อง แปลงค่าสัมประสิทธิ์สหสัมพันธ์ (r_{xy}) ให้เป็นค่าคะแนนมาตรฐาน (Z) เพื่อให้ข้อมูลมีการแจกแจงเข้าใกล้โค้งปกติ มากที่สุด เป็นไปตามข้อตกลงเบื้องต้นของสถิติที่ใช้ทดสอบ โดยใช้ตารางการแปลง Fisher's Z transformation

$Z_r = 0.5[ln(1+r)-ln(1-r)]$

เมื่อ Z_r คือ ค่าคะแนนฟิชเชอร์ซี

r คือ ค่าสัมประสิทธิ์สหสัมพันธ์

เปรียบเทียบความแตกต่างรายคู่จากสูตร

$$Z = \frac{Z_{r1} - Z_{r2}}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

เมื่อ Z คือ สถิติทดสอบซี (Z test)

Z_{r1}, Z_{r2} คือ ค่าฟิเชอร์ซี ซึ่งแปลงมาจากค่าสหสัมพันธ์ r_1, r_2

n คือ ขนาดของกลุ่มตัวอย่าง

2) ค่าระยะทางแบบยุคลิด (Euclidean distance)

ค่าระยะทางแบบยุคลิด คือ การหาระยะห่างระหว่าง ค่าความสามารถของผู้สอบที่ได้จากการประมาณค่า จากการทดสอบแบบปรับเหมาะ ($\hat{\theta}$) กับค่าความสามารถของผู้สอบที่แท้จริง (θ) สำหรับกรณีการทดสอบสั้นๆ ก็สามารถนำระยะทางแบบยุคลิดมาเป็นดัชนีวัดความแม่นยำทางจิตมิติ (Finkelman et al., 2009) โดยที่ ค่าระยะทางน้อย หมายถึง ค่าควา มสามารถจากการประมาณค่าและค่าความสามารถที่แท้จริงมีความใกล้เคียงกัน นั่นคือ วิธีการทดสอบปรับเหมาะแบบพหุมิติมีคุณภาพสูง ซึ่งค่าระยะทางแบบยุคลิด หาได้จากสูตร ดังนี้

$$ED_i = \sqrt{\left(\theta_{i1} - \hat{\theta}_{i1}\right)^2 + \left(\theta_{i2} - \hat{\theta}_{i2}\right)^2 + \ldots \left(\theta_{ij} - \hat{\theta}_{ij}\right)^2}$$

ED_i คือ ค่าระยะทางแบบยุคลิดของผู้สอบคนที่ i

i คือ ผู้สอบคนที่ i

j คือ มิติความสามารถ ที่ j

$\hat{\theta}_{ij}$ คือ ค่าความสามารถที่ประมาณค่าได้ของผู้สอบคนที่ i มิติที่ j

θ_{ij} คือ ค่าความสามารถที่แท้จริงของผู้สอบคนที่ i มิติที่ j

1.5 การตรวจสอบความถูกต้องของโปรแกรม

ผู้วิจัยตรวจสอบความถูกต้องของโปร แกรม โดยใช้ค่าที่ได้จากตัวอย่างของ Segall (2000) มาใช้ในการทดลอง ซึ่งตัวอย่างดังกล่าวใช้การคัดเลือกข้อสอบด้วยวิธีของ Fisher ตัวอย่างนี้ศึกษาจากผู้สอบ 1 คน ดำเนินการสอบแบบกำหนดความยาวข้อสอบ (fixed-length) 4 ข้อ จากคลังข้อสอบ 8 ข้อ พร้อมกำหนดค่าพารา มิเตอร์ข้อสอบ และการตอบข้อสอบทั้ง 8 ข้อ โดยศึกษา 2 มิติ ที่มีความสัมพันธ์กัน 0.6 ผลการตรวจสอบความถูกต้องพบว่า ค่าการคำนวณการคัดเลือกข้อสอบถูกต้องตามตัวอย่างทุกขั้นตอน นอกจากนั้นผู้วิจัยจะพิจารณาจากข้อสอบที่ถูกคัดเลือกขึ้นมา ว่ามีความยากใกล้เคียงกับค่าควา มสามารถที่แท้จริงหรือไม่ และการประมาณค่าความสามารถที่ได้ควรมีแนวโน้มใกล้เคียงกับความสามารถที่แท้จริงมากขึ้น ไม่ว่าจะใช้วิธีการคัดเลือกข้อสอบวิธีใด

ระยะที่ 2 การศึกษาจากข้อมูลทุติยภูมิ

การศึกษาจาก ข้อมูลทุติยภูมิ จะใช้ ข้อมูลการตอบข้อสอบของนักเรียนจากการทดสอ บความถนัดทางวิชาชีพและวิชาการ (professional and academic aptitude test: PAT) ด้านวิทยาศาสตร์ (PAT2) ของสถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน)

2.1 ประชากรและกลุ่มตัวอย่าง

ประชากร

ประชากรจาก การวิจัยครั้งนี้ คือ ผลการตอบข้อสอบของผู้เข้าสอบความถนัดทาง วิชาชีพและวิชาการด้านวิทยาศาสตร์ (PAT2) ครั้งที่ 1/2552 สอบวันที่ 8 มีนาคม 2552 จำนวนผู้สอบ 123,264 คน โดยได้รับการสนับสนุนข้อมูลทุติยภูมิจากสถาบันการทดสอบทางการศึกษาแห่งชาติ

กลุ่มตัวอย่างและการสุ่มตัวอย่าง

กลุ่มตัวอย่างจากประชากรผู้เข้าสอบความถนัดทางวิชา ชีพและวิชาการด้านวิทยาศาสตร์ ครั้งที่ 1/2552 ผู้วิจัยจะใช้วิธีการสุ่มอย่างง่ายมา 5,000 คนและสุ่มผู้สอบมาอีกร้อยละ 50 ของผู้สอบทั้งหมด นั่นคือ 62,028 คน สำหรับประมาณค่าพารามิเตอร์ข้อสอบ และทดลองใช้โปรแกรม จากนั้นแบ่งเงื่อนไขการทดลอง ออกเป็น 4 กลุ่ม เพื่อให้ไ ด้รับวิธีการคัดเลือกข้อสอบขั้นแรกและรูปแบบของลำดับข้อสอบที่แตกต่างกันทั้งนี้กลุ่มตัวอย่าง 5,000 คน คือ ผู้สอบชุดเดียวกันทุกเงื่อนไขการทดลอง รายละเอียดดังตารางที่ 3.2

ตารางที่ 3.2 เงื่อนไขและจำนวนกลุ่มตัวอย่างจากข้อมูลทุติยภูมิ

เงื่อนไข	จำนวนกลุ่มตัวอย่าง	วิธีการคัดเลือกข้อสอบขั้นแรก	ลำดับข้อสอบ
1	5,000 คน	FI	จำแนกมิติ
2	5,000 คน	FI	ผสมผสานมิติ
3	5,000 คน	KL	จำแนกมิติ
4	5,000 คน	KL	ผสมผสานมิติ

2.2 ขั้นตอนการดำเนินการวิจัย

ขั้นตอน ที่ 1 การวิเคราะห์ข้อสอบความถนัดทางวิชาชีพและวิชาการด้านวิทยาศาสตร์(PAT2) มีรายละเอียดดังนี้

1) ขอความอนุเคราะห์ข้อมูลการตอบข้อสอบรายข้อของนักเรียนทั้งหมดที่สอบวิชา PAT2 ครั้งที่ 1/2552 พร้อมเฉลยข้อสอบจากสถาบันทดสอบทางการศึกษาแห่งชาติ

2) นำ ผลการทดสอบที่ ได้มา ตรวจสอบเชิงยืนยันว่าข้อสอบแต่ละข้อวัดความสามารถตามที่ ข้อสอบระบุไว้หรือไม่ ทั้งนี้ข้อสอบที่นำมาใช้เป็นคลังข้อสอบนั้น มีลักษณะเป็นข้อสอบที่วัดความสามารถหลายมิติ ซึ่ง โครงสร้างของข้อสอบมีลักษณะเป็นข้อสอบพหุมิติระหว่างข้อ (between-item multidimensionality) ซึ่งผู้วิจัยจะ ศึกษาจากข้อสอบ 3 วิชาเท่านั้น ได้แก่ ชีววิทยา เคมี และฟิสิกส์ โดยจะแบ่งมิติที่จะศึกษาออกเป็น 3 มิติ ตามวิชาข้างต้น

3) ประมาณค่าพารามิเตอร์ของข้อสอบ โดยใช้โปรแกรม TESTFACT ประมาณค่าพารามิเตอร์ a และ b แบบพหุมิติ เนื่องจาก โปรแกรม TESTFACT เป็นโปรแกรมคอมพิวเตอร์ที่สามารถประมาณค่าพารามิเตอร์ของข้อสอบแบบพหุมิติได้ และยังสามารถประมาณค่าพารามิเตอร์ข้อผู้สอบหรือความสามารถของผู้สอบได้อีกด้วย และประมาณค่าพารามิเตอร์ c ด้วยโปรแกรม BILOG

ขั้นตอนที่ 2 การนำข้อมูลเข้าสู่กระบวนการทดสอบปรับเหมาะแบบพหุมิติผ่านโปรแกรม R ที่พัฒนาขึ้นสำหรับการจำลองข้อมู ล ในระยะที่ 1 โดยใช้รูปแบบการทดสอบปรับเหมาะแบบพหุมิติลักษณะเดียวกันกับการจำลองข้อมูลในระยะที่ 1 ยกเว้นลำดับข้อสอบแบบจำแนกมิติ ที่จะจำแนกมิติตามวิชา เรียงลำดับดังนี้ ชีววิทยา 10 ข้อ ตามด้วยเคมี 10 ข้อ และฟิสิกส์ 10 ข้อ

ขั้นตอนที่ 3 บันทึกค่าที่ได้จากการทดสอบ ได้แก่ ค่าความสามารถที่ประมาณได้จากการทดสอบแบบปรับเหมาะ และค่าความสามารถที่ประมาณค่าได้จากการทำข้อสอบทุกข้อ ซึ่งค่านี้จะเป็นค่าความสามารถที่แท้จริงของผู้สอบ

2.3 เครื่องมือที่ใช้ในการวิจัย

สำหรับเครื่องมือที่ ใช้สำหรับการศึกษาข้อมูลทุติยภูมิ คือ แบบ บันทึกข้อมูลทุติย ภูมิ การทดสอบความถนัดทางวิชาชีพและวิชาการ PAT (professional and Academic Aptitude test) ด้านวิทยาศาสตร์ (PAT2) จากการสอบครั้งที่ 1/2552 สอบวันที่ 8 มีนาคม 2552 แบบทดสอบนี้สร้างโดยสถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน) ซึ่งมีจำนวนข้อส อบ 123 ข้อ

โดยข้อสอบมีลักษณะเป็นแบบเลือกตอบ 4 ตัวเลือก ข้อสอบแต่ละข้อมีน้ำหนักคะแนนไม่เท่ากัน ซึ่งมีรายละเอียดข้อสอบ ดังตารางที่ 3.3 โดยการวิจัยครั้งนี้ ใช้ข้อสอบใน 3 เนื้อหาเท่านั้น นั่นคือ ชีววิทยา เคมี และฟิสิกส์ รวมทั้งสิ้น 100 ข้อ

ตารางที่ 3.3 โครงสร้างแบบทดสอบความถนัดทางวิชาชีพและวิชาการ PAT (professional and academic aptitude test) ศักยภาพทางวิทยาศาสตร์ (PAT2)

สิ่งที่วัด	จำนวนข้อ	ข้อที่	การให้คะแนน
เนื้อหา			
1) ชีววิทยา สิ่งแวดล้อม	40	1-40	ข้อละ 2.0 คะแนน รวม 80 คะแนน
2) เคมี	32	41-72	ข้อละ 2.5 คะแนน รวม 80 คะแนน
3) ฟิสิกส์	28	73-100	ข้อละ 3.0 คะแนน รวม 84 คะแนน
4) โลก ดาราศาสตร์ อวกาศ	13	101-113	ข้อละ 2.0 คะแนน รวม 26 คะแนน
รวม	113		270 คะแนน
ศักยภาพ	10	114-123	ข้อละ 3.0 คะแนน รวม 30 คะแนน
1) คิดแบบนักวิทยาศาสตร์			
2) แก้ปัญหาแบบนักวิทยาศาสตร์			
3) ทักษะการอ่านบทความแบบนักวิทยาศาสตร์			
รวม	10		30 คะแนน
รวม	**123**		**300 คะแนน**

* งานวิจัยครั้งนี้ ศึกษาจากข้อสอบชีววิทยา เคมี และฟิสิกส์ เท่านั้น

2.3.1 ผลการวิเคราะห์คุณภาพแบบทดสอบความถนัดทางวิชาชีพและวิชาการด้านวิทยาศาสตร์แบบเอกมิติ

จากการทดสอบความถนัดทางวิชาชีพและวิชาการด้านวิทยาศาสตร์ (PAT2) ครั้งที่ 1/2552 วันที่ 8 มีนาคม พ.ศ. 2552 ผู้เข้าสอบ 123,263 คน พบว่า คะแนนสอบเต็ม 300 คะแนน ผู้สอบมีค่าเฉลี่ยของคะแนนอยู่ที่ 90.47 คะแนน ต่ำสุด 0 คะแนน และสูงสุด 254.5 คะแนน เมื่อศึกษาตามทฤษฎี IRT แบบเอกมิติ พบว่า แบบทดสอบมีค่าอำนาจจำแน ก (a) เฉลี่ย 1.087 ค่าความยาก (b) เฉลี่ย 2.365 และค่าโอกาสการเดา (c) เฉลี่ย 0.228 โดยมีค่าความเที่ยง (rtt) เท่ากับ 0.685 และมีค่าความคลาดเคลื่อนมาตรฐานของการวัด(SEM) เท่ากับ 10.119 ดังตารางที่ 3.4 (สถาบันทดสอบทางการศึกษาแห่งชาติ, 2553)

อย่างไรก็ตามข้อมูลดังกล่าว เป็นการศึกษาตาม ทฤษฎี IRT แบบเอกมิติ แต่ไม่มีข้อมูลชี้ชัดว่า แบบทดสอบดังกล่าวฝ่าฝืนข้อตกลงเบื้องต้นเกี่ยวกับความเป็นเอกมิติหรือไม่

ตารางที่ 3.4 ค่าสถิติพื้นฐานของคะแนนสอบและผลการวิเคราะห์คุณภาพแบบทดสอบแบบเอกมิติ

ค่าสถิติพื้นฐานของคะแนนสอบ		ผลการวิเคราะห์คุณภาพแบบทดสอบ	
จำนวนผู้สอบ (N)	123,263	ค่าอำนาจจำแนกเฉลี่ย (a)	1.087
ค่าเฉลี่ย (Mean)	90.47	ค่าความยากเฉลี่ย (b)	2.365
มัธยฐาน (median)	88.5	ค่าโอกาสการเดาเฉลี่ย (c)	0.228
ฐานนิยม (mode)	86.5	ค่าความเที่ยง(rtt)	0.685
ส่วนเบี่ยงเบนมาตรฐาน (SD.)	18.03	ค่าความคลาดเคลื่อน -	
ความเบ้ (Sk)	1.33	มาตรฐานของการวัด (SEM)	10.119
ความโด่ง (Ku)	4.51		
ค่าต่ำสุด (Min)	0		
ค่าสูงสุด (Max)	254.5		

2.3.2 ผลการวิเคราะห์องค์ประกอบเชิงยืนยันแบบทดสอบความถนัดทางวิชาชีพและวิชาการด้านวิทยาศาสตร์

สำหรับการวิจัยครั้งนี้ใช้การวิเคราะห์ข้อสอบภายใต้ทฤษฎีการตอบสนองข้อสอบแบบพหุมิติ โดยวิเคราะห์จากแบบทดสอบความถนัดทางวิชาชีพและวิชาการด้านวิทยาศาสตร์ ซึ่งนำมาศึกษาเฉพาะเนื้อหา ชีววิทยา เคมี และฟิสิกส์ เท่านั้น ทั้งนี้ผู้วิจัยมีสมมติฐานว่า แบบทดสอบจำแนกเป็น 3 องค์ประกอบตามเนื้อหาวิชาที่มุ่งวัด จึงดำเนินการทดสอบองค์ประกอบเชิงยืนยัน (confirmatory factor analysis) ด้วยโปรแกรม TESTFACT ซึ่งวิเคราะห์ในลักษณะของ confirmatory item factor analysis ที่เรียกว่า bi-factor analysis นั้นคือ ข้อสอบทุกข้ อวัดในองค์ประกอบทั่วไป (general factor) และองค์ประกอบแบบกลุ่ม (group factors) ทั้งนี้ข้อสอบทุกข้อมีน้ำหนักองค์ประกอบทั่วไป และ มีน้ำหนักองค์ประกอบแบบกลุ่ม ตามที่ผู้วิจัยกำหนดในแต่ละข้อ ในลักษณะที่ไม่เหลื่อมล้ำ ไปในองค์ประกอบอื่นๆ นั่นคือ มีลักษณะตรงกับข้อสอบแบบพหุมิติระหว่างข้อสอบ (between-item multidimensionality) การวิเคราะห์ครั้งนี้ เป็นการตรวจสอบว่าข้อสอบข้อที่1- 40 วัดองค์ประกอบที่ 1 (ชีววิทยา) ข้อสอบข้อที่ 41-72 วัดองค์ประกอบที่ 2 (เคมี) และ ข้อสอบข้อที่ 73 -100 วัดองค์ประกอบที่ 3 (ฟิสิกส์) ผลการวิเคราะห์จากข้อมูลผู้เข้าสอบทั้งสิ้น 62,028 คน พบว่า ข้อสอบแต่ละข้อ อยู่ในองค์ประกอบที่กำหนดไว้ข้างต้น อย่างมีนัยสำคัญที่ระดับ .05 โดยมีค่าไคสแควร์ เท่ากับ 5,549,277.44, DF = 61,727

2.3.3 การประมาณค่าพารามิเตอร์ข้อสอบแบบพหุมิติ

การประมาณค่าพารามิเตอร์ ตามโมเดลการตอบสนองข้อสอบแบบพหุมิติที่ใช้ในการวิจัยครั้งนี้ ใช้โปรแกรม BILOG-MG (Zimowski, Muraki, Mislevy & Bock, 2003) ในการประมาณค่าพารามิเตอร์โอกาสการเดาข้อสอบ (c) และใช้โปรแกรม TESTFACT ที่พัฒนาโดย Wilson, Wood & Gibbons (1984) เพื่อประมาณค่าพารามิเตอร์ความยาก (b) และอำนาจจำแนก(a)

TESTFACT เป็นโปรแกรมคอมพิวเตอร์ ที่ออกแบบขึ้นเพื่อใช้กับข้อมูลที่มีการตรวจให้คะแนน 2 ค่า คือ 1 และ 0 โดยมีข้อตกลงเบื้องต้นว่ากลุ่มตัวอย่างได้มาจากการสุ่มของประชากร และระดับความสามารถของคุณลักษณะแฝง (latent trait) ของผู้สอบมาจากการแจกแจงปกติแบบพหุ(MVN (0,1)) (Mislevy, 1986) นั่นคือมีค่าเฉลี่ยเป็น 0 และมีส่วนเบี่ยงเบนมาตรฐานเป็น 1 (Knol & Berger, 1988 อ้างถึงใน พัชรี จันทร์เพ็ง, 2550) ซึ่งโปรแกรมที่ใช้ในการประมาณค่าโมเดล MIRT ส่วนใหญ่รวมถึงโปรแกรม TESTFACT ได้มีการแสดงเพื่อระบุปัญหา (หรือการสร้างวิธีการที่เป็นแบบแผนเดียวกัน) คือเหมาะสมกับลักษณะของข้อมูลที่มีการแจกแจงปกติแบบพหุมิติ สำหรับการวิจัยครั้งนี้มีลักษณะพหุมิติระหว่างข้อ (between-item multidimensionality) ผู้วิจัยจึงใช้การวิเคราะห์แบบ bi-factor ซึ่งจะให้ค่าพารามิเตอร์อำนาจจำแนกตามมิติที่มุ่งวัด และอำนาจจำแนกรวม ดังนั้นข้อสอบทุกข้อจึงมีค่าอำนาจจำแนก 2 ค่า นั่นคือ ค่าอำนาจจำแนกในวิชาที่วัดและค่าอำนาจจำแนกรวม สำหรับอำนาจจำแนกในมิติอื่นๆ ที่ไม่ได้มุ่งวัดจะมีค่าเป็น 0

2.3.4 ผลการวิเคราะห์คุณภาพแบบทดสอบความถนัดทางวิชาชีพและวิชาการด้านวิทยาศาสตร์แบบพหุมิติ

การวิเคราะห์คุณภาพแบบทดสอบความถนัดทางวิชาชีพและวิชาการด้านวิทยาศาสตร์แบบพหุมิติจากข้อสอบ 100 ข้อ จะพิจารณา 3 พารามิเตอร์ คือ พารามิเตอร์อำนาจจำแนก (a) พารามิเตอร์ความยาก (b) และพารามิเตอร์โอกาสการเดา (c) ซึ่งพบว่า อำนาจจำแนกมีค่าอยู่ระหว่าง -0.19 ถึง 0.73 สำหรับอำนาจจำแนกที่ติดลบนั้น ถือว่ายังมีค่าน้อย ดังนั้นการวิจัยครั้งนี้จึงไม่ได้ตัดข้อสอบข้อใดออกจากการวิจัย สำหรับความยาก พบว่า ข้อสอบมีความยากระหว่าง -0.75 ถึง 1.21 ซึ่งอยู่ในระดับปานกลาง และโอกาสการเดามีค่าอยู่ระหว่าง 0.02 ถึง 0.20 ซึ่งถือว่าส่วนใหญ่มีค่าต่ำ รายละเอียดดังตารางที่ 3.5

ตารางที่ 3.5 ค่าพารามิเตอร์ข้อสอบ PAT2 จากการวิเคราะห์แบบพหุมิติ

พารามิเตอร์ข้อสอบ	N	Min	Max	Mean	SD.	Sk	Ku
อำนาจจำแนกภาพรวม (a)	100	-0.19	0.73	0.25	0.19	0.01	-0.17
อำนาจจำแนกชีววิทยา (a1)	40	-0.05	0.34	0.13	0.10	-0.07	-0.91
อำนาจจำแนกเคมี (a2)	32	0.01	0.38	0.14	0.10	0.65	-0.72
อำนาจจำแนกฟิสิกส์ (a3)	28	-0.08	0.36	0.11	0.12	0.38	-0.83
ความยาก (b)	100	-0.75	1.21	0.41	0.39	-0.20	-0.12
โอกาสการเดา (c)	100	0.02	0.20	0.17	0.04	-1.50	2.45

2.4 การวิเคราะห์คุณภาพของการทดสอบปรับเหมาะแบบพหุมิติกรณีการใช้ข้อมูลทุติยภูมิ

การศึกษาคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติ พิจารณาจากค่าความเที่ยงและค่าระยะทางแบบยุคลิด เช่นเดียวกับการจำลอง ข้อมูล (ศึกษารายละเอียดได้จากการจำลองข้อมูลในระยะที่ 1) อย่างไรก็ตามค่าความสามารถที่แท้ จริงของผู้สอบความถนัดทางวิชาชีพและวิชาการ ที่นำมาใช้เปรียบเทียบความ แม่นยำของการประมาณค่า ผู้วิจัย ใช้ผลจากการทำข้อสอบทั้ง 100 ข้อ เป็นค่าความสามารถที่แท้จริงของผู้สอบ เนื่องจากจุดเด่นของข้อมูลทุติยภูมิ คือ ผู้วิจัยมีข้อมูลการตอบสนองข้อสอบรายข้อ ครบทั้ง 100 ข้อที่มีในคลังข้อสอบ แต่การทดสอบปรับเหมาะจะคัดเลือกข้อสอบเพียงไม่กี่ข้อให้กับผู้สอบ แต่สามารถประมาณค่าความสามารถได้อย่างแม่นยำสอดคล้องกับการทำข้อสอบครบทุกข้อ

ระยะที่ 3 การศึกษาจากข้อมูลปฐมภูมิ (การทดลองใช้โปรแกรม)

การศึกษาจากข้อมูลปฐมภูมิ คือ การพัฒนาโปรแกรมการทดสอบปรับเหมาะแบบพหุมิติเพื่อใช้ทดสอบกับกลุ่มตัวอย่าง ซึ่งการวิจัยครั้งนี้ได้พัฒนาการทดสอบความสามารถด้านวิทยาศาสตร์ในระดับมัธยมศึกษาตอนปลาย ทั้งนี้ผู้วิจัยได้เลือก วิธีการคัดเลือกข้อสอบเพียงวิธีเดียวที่ ใช้กันอย่างแพร่หลายมากที่สุด เนื่องจากมีความซับซ้อนน้อยกว่าวิธีการอื่นๆ นั่นคือ การคัดเลือกข้อสอบโดยใช้วิธีสารสนเทศของฟิชเชอร์ (Fisher information) โดยศึกษากับกลุ่มตัวอย่างเพียง 1 กลุ่ม พร้อมทั้งประเมินผลการใช้โปรแกร มจากนักเรียนกลุ่มตัวอย่างพร้อมศึกษาภูมิหลังของนักเรียนที่ส่งผลต่อผลการประเมินการใช้โปรแกรม ซึ่งมีรายละเอียดดังนี้

3.1 ประชากรและกลุ่มตัวอย่าง

ประชากร คือ นักเรียนชั้นมัธยมศึกษาตอนปลาย ในโรงเรียนมัธยมศึกษาขนาดใหญ่พิเศษ

กลุ่มตัวอย่าง พิจารณาจากแนวคิด ของ Hair (1998) ที่เสนอว่าจำนวนหน่วยตัวอย่างที่นำมาศึกษาควรมีประมาณ 20 เท่าของตัวแปรอิสระ ซึ่งการศึกษาจากข้อมูลปฐมภูมิศึกษาเพียงตัวแปรลำดับ ข้อสอบเท่านั้น ดังนั้นจึงศึกษาตัวแปรอิสระเพียง 1 ตัวแปร ที่มี 2 กลุ่ม คือ ลำดับข้อสอบ แบบจำแนกมิติ และ ลำดับข้อสอบ แบบผสมผสานมิติ ขนาดของกลุ่มตัวอย่างที่นำมาใช้แต่ละกลุ่มควรมีอย่างน้อย 20 คน ซึ่งจะทำให้มีกลุ่มตัวอย่างขั้นต่ำ 2x20 = 40 คน อย่างไรก็ตาม การศึกษาในระยะที่3 เป็นเพียงการนำโปรแกรมการทดสอบปรับเหมาะแบบพหุมิติมาทดลองใช้เท่านั้น ผู้วิจัยจึงได้เลือกกลุ่มตัว อย่างมา 1 กลุ่ม คือ นักเรียนชั้นมัธยมศึกษาปีที่ 5 โรงเรียนอุตรดิตถ์ดรุณี อ.เมือง จ.อุตรดิตถ์ จำนวน 69 คน และนักเรียนชั้นมัธยมศึกษาตอนปลายจากโรงเรียนต่าง ๆ ที่เข้าระบบทำข้อสอบบนเว็บไซต์ที่ผู้วิจัยเปิดให้นักเรียนทั่วไปลงทะเบียนทำข้อสอบได้ รวม 74 คน

3.2 เครื่องมือที่ใช้ในการวิจัย

เครื่องมือสำหรับการทดลองใช้โปรแกรม ได้แก่ โปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบพหุมิติ แบบทดสอบความ สามารถ ด้านวิทยาศาสตร์ ระดับมัธยมศึกษาตอนปลาย และ แบบประเมินการใช้โปรแกรมการทดสอบแบบปรับเหมาะ ซึ่งมีรายละเอียดดังนี้

3.2.1 โปรแกรมการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์

โปรแกรมการทดสอบ ปรับเหมาะแบบพหุมิติ ใช้วิธีการคัดเลือกข้อสอบด้วยวิธี Fisher information โดยมีเงื่อนไขของโปรแกรมแยกออกเป็น 2 แบบ ตามรูปแบบของ ลำดับข้อสอบ ที่แตกต่างกัน ซึ่งมีรายละเอียดของโปรแกรมดังนี้ (คู่มือโปรแกรมเสนอไว้ดังภาคผนวก ง)

การพัฒนาโปรแกรมครั้งนี้ พัฒนาด้วย ASP.NET (active server pages) ซึ่งเป็นโปรแกรมพัฒนาผ่านเว็บเพจ (web application) ซึ่งเป็นการตัดปัญหาเรื่องการติดตั้ง และการแก้ไขบัค หรือไวรัสที่เครื่องลูกข่ายนอกจากนั้นยังสามารถรายงานข้อมูลเป็นกราฟ ได้ โดยไม่ต้องใช้เครื่องมือภายนอก สำหรับคุณลักษณะของระบบการทดสอบ เป็นระบบทดสอบแบบการสอบผ่านเว็บบราวเซอร์ ข้อสอบถูกรวบรวมและจัดเตรียมไว้ในฐานข้อมูล แล้วเลือกมาใช้ในการสอบตามเงื่อนไขที่ตั้งไว้ สามารถกำหนดเวลาที่ใช้สอบโดยมีนาฬิกาแจ้งตลอดเวลาทำการทดสอบ

ความสามารถของโปรแกรม

1) สำหรับวิธีการสอบเป็นแบบการสอบผ่านเว็บ บราวเซอร์ และแสดงผลทันทีหลังสอบเสร็จ ภายหลังการสอบเสร็จสามารถดูผ่านเว็บบราวเซอร์ หรือพิมพ์ออกทางเครื่องพิมพ์ เพื่อแจ้งผลการสอบได้

2) ภายหลังการสอบเสร็จสามารถแสดงผลคะแนนเป็นเปอร์เซ็นต์และกราฟได้

3) สามารถแสดงรายงานข้อมูลผลการทำข้อสอบได้ทุกที่ผ่านเว็บบราวเซอร์

5) รองรับภาษาไทยในการจัดเก็บข้อมูล รวมถึงการแสดงผล และการจัดเรียงลำดับภาษาไทย

ความต้องการขั้นต่ำของระบบ

1) เครื่องคอมพิวเตอร์ตั้งโต๊ะ หรือคอมพิวเตอร์พกพา ความเร็ว 1GHz, หน่วยความจำ1GB, ฮาร์ดดิสก์ 100GB

2) ระบบปฏิบัติการ Windows XP Pro หรือสูงกว่า

3) แพลตฟอร์มสำหรับพัฒนาซอฟแวร์ .Net Framework 2.0 หรือสูงกว่า

4) เว็บเซิร์ฟเวอร์ IIS5 หรือสูงกว่า

5) ฐานข้อมูล Microsoft SQL Server 2005 Express หรือสูงกว่า

6) เครื่องมือในการพัฒนา Visual Web Developer 2008 Express

* ซอฟแวร์ที่ต้องใช้ตั้งแต่ข้อที่ 3-6 เป็นซอฟแวร์ฟรี สามารถดาวน์โหลดมาตั้งได้อย่างถูกลิขสิทธิ์

การทดลองใช้โปรแกรม

สำหรับการทดสอบครั้งนี้จะดำเนินการสอบผ่านเว็บไซต์ www.mcatesting.com และเพื่อลดความคลาดเคลื่อนของโปรแกรมผู้วิจัยจึงได้ทดลองใช้โปรแกรมกับนักเรียนชั้น ม . 5 โรงเรียนนางรอง อำเภอนางรอง จังหวัดบุรีรัมย์ จำนวน 20 คน ผลการทดลองใช้โปรแกรมพบว่า โปรแกรมมีการคำนวณค่าต่างๆ จำนวนมาก อีกทั้งคำสั่งที่มีการหมุนวนซ้ำหลายรอบ ทำให้การคัดเลือกข้อสอบแต่ละข้อใช้เวลานานมาก โดยเฉพาะเมื่อผู้สอบทำข้อสอบหลายข้อมากขึ้น โปรแกรมจะนำค่าจากการตอบข้อก่อนหน้ามาคำนวณค่าวนซ้ำ จึงทำให้เกิดปัญหาที่ผู้สอบหยุดทำการทดสอบก่อนที่จะทำข้อสอบเสร็จสิ้น

สำหรับการแก้ปัญหา ผู้วิจัยได้ปรับสูตรการคำนวณให้มีความกระชับมากขึ้น และดึงค่าที่ ได้จากการคำนวณมาเก็บไว้ที่เครื่องคอมพิวเตอร์ แทนการคำนวณผ่าน เซิร์ฟเวอร์ และได้เพิ่มเว็บไซต์ขึ้นอีก 2 เว็บไซต์ ในกรณีที่ผู้ส อบลงทะเบียนเข้าสอบพร้อมกัน อาจ ทำให้ระบบล่มได้ นั่นคือ ww2.mcatesting.com และ ww3.mcatesting.com

3.2.2 แบบทดสอบความสามารถด้านวิทยาศาสตร์สำหรับเก็บข้อมูลจริง

แบบทดสอบ ความสามารถ ด้านวิทยาศาสตร์นี้พัฒนาขึ้นสำหรับการเก็บข้อ มูลจริง โดยมีผู้พัฒนาข้อสอบคือ อาจารย์ผู้สอนวิชาวิทยาศาสตร์ในโรงเรียนมัธยมศึกษา จำนวน 9 คน ซึ่งจะแยกตามความถนัดในแต่ละวิชา วิชาละ 3 คน สำหรับจำนวนข้อสอบขั้นต่ำที่ต้องการในการทดสอบแบบปรับเหมาะคือ เท่ากับจำนวนข้อสอบในการทดสอบความถนัดทางวิชาชีพและวิชาการด้านวิทยาศาสตร์ (PAT2)ในแต่ละครั้ง แต่เนื่องจากการสร้างข้อสอบต้องมีการคัดเลือกข้อสอบที่มีความเหมาะสมก่อนการนำไปทดสอบจริง ดังนั้นผู้วิจัยจึงวางแผนให้มีการสร้างข้อสอบขึ้น 3 ชุด แต่ละชุดจะมีจำนวน 108 ข้อ รวมข้อสอบที่ต้องพัฒนาขึ้นจำนวน 324 ข้อ ซึ่งผู้พัฒนาข้อสอบสร้างข้อสอบชีววิทยาคนละ43 ข้อ เคมี 35 ข้อ และฟิสิกส์ 30 ข้อ รายละเอียดดังตารางที่ 3.6 และมีรายละเอียดโครงสร้างข้อสอบในแต่ละวิชา จำแนกตามผู้สอบ แสดงไว้ในภาคผนวก ข

ตารางที่ 3.6 เนื้อหาข้อสอบและจำนวนข้อสอบที่ผู้พัฒนาข้อสอบแต่ละคนสร้างขึ้น

เนื้อหาข้อสอบ (จำนวนผู้พัฒนาข้อสอบ)	**จำนวนข้อสอบที่สร้างขึ้นแต่ละคน (ข้อ)**	**รวมข้อสอบ 3 ชุด (ข้อ)**
1) ชีววิทยา สิ่งแวดล้อม (3 คน)	*43*	129
2) เคมี (3 คน)	*35*	105
3) ฟิสิกส์ (3 คน)	*30*	90
รวม (9 คน)	***108***	**324**

การตรวจสอบคุณภาพของข้อสอบที่สร้างขึ้น ผู้วิจัยนำข้อสอบที่อาจารย์โรงเรียนมัธยมศึกษาสร้างขึ้นให้กับผู้ทรงคุณวุฒิทั้ง 3 สาขาวิชาตรวจสอบความถูกต้องเชิงเนื้อหา และปรับแก้ตามข้อเสนอแนะ (รายชื่อผู้พัฒนาข้อสอบและผู้ทรงคุณวุฒิตรวจสอบ เครื่องมือ แสดงไว้ในภาคผนวก ก) จากนั้นนำข้อสอบที่ผ่านการปรับแก้ไปทดลองกับนักเรียนมัธยมศึกษาปีที่ 6 ที่ไม่ใช่กลุ่มตัวอย่าง (เก็บข้อมูลระหว่างวันที่ 15 - 20 กันยายน 2554) ซึ่งการวิจัยครั้งนี้ใช้นักเรี ยนจำนวน 273 คน จากโรงเรียนสารคามพิทยาคม และโรงเรียนมหาวชิรานุกูล จังหวัดมหาสารคาม เป็นกลุ่มสำหรับการตรวจสอบคุณภาพข้อสอบ โดยแบ่งเป็นข้อสอบชุดที่ 1 จำนวน 95 คน ข้อสอบชุดที่ 2 จำนวน 96 คน และข้อสอบชุดที่ 3 จำนวน 82 คน และนำผลการทดสอบของนักเรียนมาวิเคราะห์พารามิเตอร์ข้อสอบและคัดเลือกข้อสอบที่มีความเหมาะสมโดยพิจารณาข้อสอบที่มีอำนาจจำแนก(a) เป็นบวก ค่าความยาก(b) อยู่ระหว่าง-2.50 ถึง +2.50 และค่าโอกาสการเดา(c) มีค่าไม่เกิน 0.3

ผลการวิเคราะห์คุณภาพข้อสอบพบว่า มีข้อสอบที่ตรงตามเงื่อนไข จำนวน 213 ข้อ ทั้งนี้เนื่ องจากปัญหาของโปรแกรมที่ต้องคำนวณค่า สำหรับใช้คัดเลื อกข้อสอบจำนวนมาก

จึงจำเป็นต้องคัดข้อสอบออก โดยพิจารณาข้อสอบที่มีค่าอำน าจจำแนกต่ำ ทำให้เหลือข้อสอบที่นำไปเก็บข้อมูล จำนวน 170 ข้อ แบ่งเป็นชีววิทยา 77 ข้อ เคมี 57 ข้อ และฟิสิกส์ 36 ข้อ โดยมีโครงสร้างของข้อสอบที่นำมาใช้ในการทดสอบในแต่ละวิชาดังตารางที่ 3.7 ถึง 3.9

ตารางที่ 3.7 โครงสร้างข้อสอบวิชาชีววิทยา

ลำดับที่	สาระการเรียนรู้	ข้อสอบข้อที่	จำนวน (ข้อ)
1.	การศึกษาชีววิทยา	32, 56-57, 77	4
2.	เคมีที่เป็นพื้นฐานของสิ่งมีชีวิต	10, 18, 59-60	4
3.	เซลล์ของสิ่งมีชีวิต	12-13, 17, 19, 33-34, 58, 61-62	9
4.	ระบบย่อยอาหารและการสลายอาหารเพื่อให้ได้พลังงาน	36, 63	2
5.	การรักษาดุลยภาพในร่างกาย	14, 21-26, 35, 64	9
6.	การเคลื่อนที่ของสิ่งมีชีวิต	65	1
7.	การรับรู้และการตอบสนอง	11, 37, 66, 70	4
8.	ระบบต่อมไร้ท่อ	67	1
9.	พฤติกรรมของสัตว์	3-6, 38	5
10.	โครงสร้างและหน้าที่ของพืชมีดอก	39, 69	2
11.	การสังเคราะห์ด้วยแสง	40, 52	2
12.	การสืบพันธุ์ของพืชดอก	41, 53, 55, 68	4
13.	การถ่ายทอดทางพันธุกรรม	27-28, 42, 46, 54	5
14.	ยีนและโครโมโซม	15-16, 20, 43, 48, 51	6
15.	พันธุศาสตร์และเทคโนโลยีทางDNA	49-50, 71-75	7
16.	ความหลากหลายทางชีวภาพ	29-31, 44, 47	5
17.	ระบบนิเวศ	1-2, 7-8, 45, 76	6
18.	ประชากร	9	1
	รวม		**77**

ตารางที่ 3.8 โครงสร้างข้อสอบวิชาเคมี

ลำดับที่	สาระการเรียนรู้	ข้อสอบข้อที่	จำนวน (ข้อ)
1.	โครงสร้างอะตอมและตารางธาตุ	83, 120, 122, 125	4
2.	สมบัติของธาตุ และสารประกอบ	78, 82, 95-96, 111	5
3.	ปริมาณสารสัมพันธ์ I	126-127	2
4.	พันธะเคมี	84-87, 103, 123-124	7

ตารางที่ 3.8 (ต่อ)

ลำดับที่	สาระการเรียนรู้	ข้อสอบข้อที่	จำนวน (ข้อ)
5.	ปริมาณสัมพันธ์ II	79-81, 94	4
6.	ของแข็ง ของเหลว แก๊ส	102, 107, 115, 128-129	5
7.	อัตราการเกิดปฏิกิริยาเคมี	88, 106, 116, 130	4
8.	สมดุลเคมี	89-90, 108-109, 121	5
9.	กรด-เบส	91-93, 131	4
10.	เคมีอินทรีย์	97-98, 104	3
11.	เชื้อเพลิงซากดึกดำบรรพ์ และผลิตภัณฑ์	100, 117-119	4
12.	สารชีวโมเลกุล	99, 110	2
13.	ไฟฟ้าเคมี	101, 105, 112-114, 132-133	7
14.	ธาตุ และสารประกอบในอุตสาหกรรม	134	1
	รวม		**57**

ตารางที่ 3.9 โครงสร้างข้อสอบวิชาฟิสิกส์

ลำดับที่	สาระการเรียนรู้	ข้อสอบข้อที่	จำนวน (ข้อ)
1.	การเคลื่อนที่ในหนึ่งและสองมิติ	137, 152- 153, 161, 163	5
2.	แรง มวล กฎการเคลื่อนที่	135-136	2
3.	การเคลื่อนที่แบบต่าง ๆ	154	1
4.	งานและพลังงาน	138-139, 155, 162	4
5.	โมเมนตัมและการดล	140	1
6.	การเคลื่อนที่แบบหมุน	141	1
7.	สภาพสมดุลและสภาพยืดหยุ่น	160	1
8.	ของไหล	142, 156	2
9.	ความร้อน	143-144	2
10.	คลื่นกล	146	1
11.	แสงและทัศนอุปกรณ์	145	1
12.	ไฟฟ้าสถิต	147-148, 164	3
13.	ไฟฟ้าและแม่เหล็ก	149-150, 157- 159, 168	6
14.	คลื่นแม่เหล็กไฟฟ้า	165-167	3
15.	ฟิสิกส์อะตอม	151	1
16.	ฟิสิกส์นิวเคลียร์	169-170	2
	รวม		**36**

สำหรับการวิเคราะห์ค่าพารามิเตอร์ข้อสอบ แบบพหุมิติ ผู้วิจัยวิเคราะห์ด้วยโปรแกรม TESTFACT โดยใช้การวิเคราะห์แบบ bi-factor ซึ่งให้ค่าพารามิเตอร์อำนาจจำแนกตามมิติที่มุ่งวัดและอำนาจจำแนกรวม ดังนั้นข้อสอบทุกข้อ จึงมีค่าอำนาจจำแนก 2 ค่า นั่นคือ ค่าอำนาจจำแนกในวิชาที่วัดและค่าอำนาจจำแนกรวม สำหรับอำนาจจำแ นกในมิติอื่นๆ ที่ไม่ได้มุ่งวัด มีค่าเป็น 0 โดยมีรายละเอียดค่าพารามิเตอร์ข้อสอบดังตารางที่ 3.10

ตารางที่ 3.10 ค่าพารามิเตอร์ข้อสอบของแบบทดสอบความสามารถด้านวิทยาศาสตร์

วิชา	พารามิเตอร์ข้อสอบ	N	Mean	SD.	Min	Max
1. ชีววิทยา	อำนาจจำแนกภาพรวม (a)	77	0.66	0.19	0.12	0.91
	อำนาจจำแนกชีววิทยา (a1)	77	0.36	0.18	0.01	0.84
	ความยาก (b)	77	0.32	0.39	-0.65	1.72
	โอกาสการเดา (c)	77	0.17	0.04	0.02	0.20
2. เคมี	อำนาจจำแนกภาพรวม (a)	57	0.49	0.26	0.05	0.90
	อำนาจจำแนกเคมี (a2)	57	0.35	0.18	0.03	0.76
	ความยาก (b)	57	0.10	0.40	-1.25	1.08
	โอกาสการเดา (c)	57	0.15	0.05	0.02	0.20
3. ฟิสิกส์	อำนาจจำแนกภาพรวม (a)	36	0.53	0.24	0.03	0.84
	อำนาจจำแนกฟิสิกส์ (a3)	36	0.38	0.19	0.02	0.79
	ความยาก (b)	36	-0.03	0.51	-0.55	1.47
	โอกาสการเดา (c)	36	0.15	0.04	0.03	0.20
รวม	อำนาจจำแนกภาพรวม (a)	170	0.58	0.24	0.03	0.91
	อำนาจจำแนกชีววิทยา (a1)	77	0.36	0.18	0.01	0.84
	อำนาจจำแนกเคมี (a2)	57	0.35	0.18	0.03	0.76
	อำนาจจำแนกฟิสิกส์ (a3)	36	0.38	0.19	0.02	0.79
	ความยาก (b)	170	0.17	0.44	-1.25	1.72
	โอกาสการเดา (c)	170	0.13	0.04	0.02	0.20

3.3.3 แบบประเมินการใช้โปรแกรมการทดสอบแบบปรับเหมาะ

แบบประเมินการใช้โปรแกรมการทดสอบแบบปรับเหมาะแบ่งออกเป็น 3 ด้าน คือ ด้านความยากในการใช้ ความเป็นไปได้ในการนำไปใช้ และความพึงพอใจที่มีต่อการใช้ ซึ่งเป็นแบบประเมิน ที่ผู้วิจัยพัฒนา เพิ่มเติมจากการศึกษาของ สิริลักษณ์ เกษรปทุ มมานันท์ (2549) โดยนักเรียน ตอบแบบประเมิน ผ่านทางโปรแกรมคอมพิวเตอร์หลังจากทำการทดสอบแบบปรับเหมาะเสร็จสิ้น แบบประเมินมีลักษณะเป็นมาตรประมาณค่า (rating scale) 5 ระดับ ตัวอย่างแบบประเมิน ดังแสดงไว้ในภาคผนวก จ

3.3.4 แบบวัดทักษะความสามารถด้านคอมพิวเตอร์

แบบวัดทักษะความสามารถด้าน คอมพิวเตอร์ในการวิจัยครั้งนี้ ใช้แบบทดสอบที่พัฒนาขึ้นโดย ชุติวัฒน์ สุวัตถิพงศ์ (2551) จำนวน 35 ข้อ ซึ่งเป็นแบบวัดความสามารถด้านคอมพิวเตอร์ที่สร้างขึ้นให้สอดคล้องกับหลักสูตรการศึกษาขั้นพื้นฐานในช่วงชั้นที่ 3-4 พ.ศ. 2551 โดยวัดความสามารถเกี่ยวกับการใช้โปรแกร มประมวล ผลคำ (microsoft word) และโปรแกรมนำเสนอผลงาน (microsoft PowerPoint) ซึ่งแบบวัดดังกล่าวผ่ านการตรวจสอบคุณภาพเครื่องมือเรียบร้อยแล้ว โดยมีค่าความยากเฉลี่ย 0.67 ค่าอำนาจจำแนกเฉลี่ย 0.34 มีค่าความเที่ยงของแบบทดสอบทั้งฉบับ 0.71 และมีค่าดัชนี IOC มากกว่า 0.5 ดังแสดงไว้ในภาคผนวก จ

3.4 การเก็บรวบรวมข้อมูล

สำหรับการเก็บรวบรวมข้อมูลจากแหล่งข้อมูลปฐมภูมิ จะดำเนินการสอบผ่านระบบอินเตอร์เน็ต โดยนักเรียนกลุ่มตัวอย่างต้องลงทะเบียนเข้าระบบและกรอกข้อมูลส่วนบุคคล ก่อนจะเริ่มสอบ โดยมีคุณ ครูคอยแนะนำขั้นตอนการสอบ และเนื่องจากการเข้าระบบพร้อมกัน อาจทำให้เว็บไซด์มีความผิดพลาด ผู้วิจัยจึงเพิ่ม URL เพื่อลดความผิดพลาดที่เกิดขึ้น ประกอบด้วย www.mcatesting.com, ww2.mcatesting.com และ ww3.mcatesting.com นอกจากนี้ ผู้วิจัยได้เปิดโอกาสให้นักเรียนระดับมัธยมศึกษาตอนปลาย ทั่วไป สามารถเข้ามาทำแบบทดสอบได้ เพื่อให้ได้ข้อมูลที่หลากหลายมากขึ้น

3.5 การวิเคราะห์คุณภาพของการทดสอบปรับเหมาะแบบพหุมิติที่ได้จากข้อมูลปฐมภูมิ (การทดลองใช้โปรแกรม)

สำหรับคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติที่ได้จากข้อมูลปฐมภูมิ พิจารณาจากค่าสารสนเทศของข้อสอบ และความยาวข้อสอบ ทั้งนี้ ไม่สามารถ พิจารณาค่าความเที่ยง และระยะทางแบบยุคลิดได้ เนื่องจาก การทดสอบจริงผู้วิจัย ไม่ทราบค่า ความสามารถที่

แท้จริงของผู้สอบ ซึ่งเป็นข้อมูลหลักที่ใช้สำหรับหาค่าความเที่ยง และระยะทางแบบยุคลิด สำหรับค่าสารสนเทศแบบสอบ และความยาวข้อสอบมีรายละเอียดในการพิจารณาดังนี้

3.5.4 สารสนเทศของแบบสอบ

ค่าฟังก์ชันสารสนเทศของแบบสอบ (test information: $I(\theta)$) เกิดจากผลรวมเชิงพีชคณิตของค่าฟังก์ชันสารสนเทศของข้อสอบแต่ละข้อรวมเข้าด้วยกันทั้งฉบับ ณ ตำแหน่ง θ เดียวกัน

$$I(\theta)=\sum_{i=1}^{k} I_i(\theta) \quad , \quad i = 1, 2, \ldots,k$$

$$I_i(\theta)=\frac{\left[P_i'(\theta)\right]^2}{P_i(\theta)Q_i(\theta)} \quad , \ i = 1, 2, \ldots,k$$

เมื่อ $I(\theta)$ = ค่าฟังก์ชันสารสนเทศ หรือค่าสารสนเทศที่ได้จากแบบสอบถาม สำหรับผู้ตอบที่มีความสามารถ θ

$I_i(\theta)$ = ค่าฟังก์ชันสารสนเทศ หรือค่าสารสนเทศที่ได้จากข้อสอบข้อที่ i สำหรับผู้ตอบที่มีความสามารถ θ

$P_i'(\theta)$ = P_i' = ความชันของฟังก์ชันการตอบสนองข้อสอบข้อที่ i ณ ตำแหน่งความสามารถ θ

$P_i(\theta)$ = P_i = ความน่าจะเป็นของผู้ตอบที่มีความสามารถ θ จะตอบข้อสอบข้อที่ i ได้ถูกต้อง

$Q_i(\theta)$ = Q_i = 1- $P_i(\theta)$

3.5.5 ความยาวข้อสอบ

ความยาวข้อสอบหาได้จากการหาค่าเฉลี่ยจำนวนข้อสอบที่ใช้ในการทดสอบแบบปรับเหมาะ ซึ่ง พิจารณาที่ตำแหน่งความคลาดเคลื่อนมีค่าน้อยกว่าหรือเท่ากับ 0.3 (SE $\leq$0.3) ซึ่งกำหนดให้เป็นเกณฑ์ยุติการทดสอบ โดยการทดสอบแบบปรับเหมาะที่ มีคุณภาพนั้น ผู้สอบจะได้รับข้อสอบจำนวนน้อยแต่สามารถประมาณค่าความสามารถได้อย่างถูกต้อง และแม่นยำ อย่างไรก็ตามการทดสอบครั้งนี้ ได้กำหนดให้ดำเนินการสอบทั้งสิ้นจำนวน 30 ข้อ เนื่องจากหากหยุดการทดสอบที่ SE $\leq$0.3 ผู้สอบจะได้ทำข้อสอบไม่ครบทุกมิติ ดังนั้นผู้วิจัยจึงจะพิจารณาเกณฑ์การยุติข้อสอบทั้งสองรูปแบบไปพร้อมกัน

3.6 สัดส่วนของจำนวนข้อสอบที่ผู้สอบได้รับในแต่ละวิชา

รูปแบบที่ต่างกันของลำดับข้อสอบ ระหว่างลำดับข้อสอบแบบจำแนกมิติ และลำดับข้อสอบแบบผสมผสานมิติ มีผลต่อจำนวนข้อสอบที่ผู้สอบได้รับในแต่ละวิชา โดยงานวิจัยนี้ กำหนดความยาวข้อสอบไว้ที่ 30 ข้อ ในกรณีของการทดสอบด้วยลำดับข้อสอบแบบจำแนกมิติ ผู้สอบแต่ละคนได้รับข้อสอบเท่ากันในทุกวิชา วิชาละ 10 ข้อ สำหรับในกรณีของการทดสอบด้วยลำดับข้อสอบแบบผสมผสานมิติ ผู้สอบแต่ละคนได้รับข้อสอบไม่เท่ากันในแต่ละวิชา เช่น ผู้สอบ คนที่ 1 ได้รับข้อสอบชีววิทยา เคมี และฟิสิกส์ เท่ากับ 11 ข้อ 6 ข้อ และ 13 ข้อ ตามลำดับ ผู้สอบคนที่ 2 ได้รับข้อสอบชีววิทยา เคมี และฟิสิกส์ เท่ากับ 10 ข้อ 8 ข้อ และ 12 ข้อ ตามลำดับ รายละเอียดของจำนวนข้อสอบที่ผู้สอบทั้ง 74 คน ได้รับ ผู้วิจัยเสนอไว้ดังภาคผนวก ง โดยพบว่า ข้อสอบที่ผู้สอบได้รับกระจายครอบคลุมทั้ง 3 วิชา ซึ่งสามารถสรุปจำนวนข้อสอบที่ได้รับสูงสุดและต่ำสุดในแต่ละวิชา ดังนี้ วิชาชีววิทยาจำนวนข้อสอบที่ได้รับสูงสุด 14 ข้อ ต่ำสุด 6 ข้อ วิชาเคมีจำนวนข้อสอบที่ได้รับสูงสุด 13 ข้อ ต่ำสุด 5 ข้อ และวิชาฟิสิกส์จำนวนข้อสอบที่ได้รับสูงสุด 17 ข้อ ต่ำสุด 8 ข้อ ดังตารางที่ 3.11

ตารางที่ 3.11 สรุปจำนวนข้อสอบที่ได้รับสูงสุดและต่ำสุดในแต่ละวิชา

	จำนวนข้อสอบที่ได้รับ (ข้อ)		
	ชีววิทยา	**เคมี**	**ฟิสิกส์**
สูงสุด	14	13	17
ต่ำสุด	6	5	8

สรุปวิธีการดำเนินการวิจัยจาก 3 แหล่งข้อมูล

การวิจัยครั้งนี้ศึกษาจาก 3 แหล่งข้อมูลที่มีความแตกต่างกัน และการพิจารณาในประเด็นต่าง ๆ ก็มีความแตกต่างกันตามบริบทของข้อมูล ดังนั้นเพื่อให้เกิดความเข้าใจ ในการวิจัยครั้งนี้มากขึ้น ผู้วิจัยได้สรุปประเด็นต่าง ๆ ที่มีความสำคัญเพื่อให้ผู้อ่านเข้าใจ งานวิจัยนี้มาก ขึ้น รายละเอียดดังตารางที่ 3.12

ตารางที่ 3.12 สรุปวิธีการดำเนินการวิจัยจาก 3 แหล่งข้อมูล

ประเด็นพิจารณา	แหล่งข้อมูล		
	การจำลองข้อมูล	**ข้อมูลทุติยภูมิ**	**ข้อมูลปฐมภูมิ**
1. การคัดเลือกข้อสอบ	FI, KL	FI, KL	FI
2. ลำดับข้อสอบ	mixed, classified	mixed, classified	mixed, classified
3. เกณฑ์การหยุด	fixed test length	fixed test length	fixed test length, SE $\leq$ 0.3
4. ความสามารถที่แท้จริง	ได้จากการจำลองข้อมูล	ได้จากการทำข้อสอบทุกข้อในคลังข้อสอบ	N/A
5. จำนวนมิติ	2 มิติ	3 มิติ (ชีวะ, เคมี, ฟิสิกส์)	3 มิติ (ชีวะ, เคมี, ฟิสิกส์)
6. กลุ่มตัวอย่าง	จำลองข้อมูลผู้สอบ 10,000 คน	ข้อมูลการสอบ PAT2 ของ สทศ. 5,000 คน	นักเรียนชั้น ม.5 ที่ทำข้อสอบผ่านโปรแกรม 74 คน
7. คุณภาพของ MCAT	R^2, Euclidean distance	R^2, Euclidean distance	test information test length
8. ผลประเมินโปรแกรม	N/A	N/A	ศึกษาภูมิหลังของผู้สอบที่มีต่อผลประเมินโปรแกรม

การวิเคราะห์ข้อมูลเพื่อตอบวัตถุประสงค์การวิจัย

การวิเคราะห์ข้อมูลเพื่อตอบวัตถุประสงค์การวิจัยมื่อพิจารณาจากคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติ เนื่องจากงานวิจัยครั้งนี้ พิจารณาคุณภาพการทดสอบหลายค่าที่มีลักษณะที่แตกต่างกัน ส่งผลให้วิธีการเปรียบเทียบคุณภาพการทดสอบดังกล่าวมีความแตกต่างกัน นั้นคือกรณีของการใช้ข้อมูลปฐมภูมิ ซึ่งผู้วิจัยไม่ทราบค่าความสามารถที่แท้จริงของผู้สอบ จะพิจารณาคุณภาพการทดสอบด้วยสารสนเทศแบบสอบและความยาวข้อสอบ เนื่องจาก ข้อมูลปฐมภูมินี้ใช้การคัดเลือกข้อสอบเพียงวิธีเดียวคือ Fisher information และจะใช้ข้อมูลนี้เปรียบเทียบ ลำดับข้อสอบที่แตกต่างกันโดยใช้สถิติ one-way MANOVA สำหรับกรณีการจำลองข้อมูล และการใช้ข้อมูลทุติยภูมิจะพิจารณาค่าความเที่ยงและระยะทางแบบยุคลิด ซึ่งค่าความเที่ยงนั้น จะใช้ค่าสัมประสิทธิ์สหสัมพันธ์มาช่วยทดสอบความแตกต่าง เนื่องจากความเที่ยงเป็นการหาความสัมพันธ์

กำลังสองจากผู้สอบทุกคน ดังนั้นจึงใช้การเปรียบเทียบ ค่าสัมประสิทธิ์สหสัมพันธ์โดยการแปลงค่าสัมประสิทธิ์สัมพันธ์เป็นคะแนนมาตรฐาน Fisher's Z และเปรียบเทียบราย คู่ด้วยสถิติ Z-test สำหรับระยะทางแบบยุคลิดจะให้ค่าเป็นรายบุคคล ดังนั้นการวิเคราะห์จึงใช้สถิติ two-way ANOVA เพื่อศึกษาเปรียบเทียบการคัดเลือกข้อสอบขั้นแรก ลำดับข้อสอบ ที่ส่งผลต่อคุณภาพการ ทดสอบปรับเหมาะแบบพหุมิติรวมถึงปฏิสัมพันธ์ระหว่างสองตัวแปร ซึ่งสามารถสรุปได้ดังตารางที่ 3.13

ตารางที่ 3.13 สรุปสถิติวิเคราะห์เพื่อตอบวัตถุประสงค์การวิจัย

<table>
<tr><th rowspan="2">วัตถุประสงค์การวิจัย</th><th colspan="4">คุณภาพของการทดสอบปรับเหมาะแบบพหุมิติ</th></tr>
<tr><th>สารสนเทศแบบสอบและความยาวข้อสอบ</th><th>ความเที่ยง</th><th>ระยะทางแบบยุคลิด</th></tr>
<tr><td>1. เพื่อเปรียบเทียบคุณภาพของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบพหุมิติภายใต้วิธีการคัดเลือกข้อสอบขั้นแรกและลำดับข้อสอบที่แตกต่างกัน</td><td></td><td></td><td></td></tr>
<tr><td>1.1 เพื่อเปรียบเทียบวิธีการคัดเลือกข้อสอบขั้นแรกที่แตกต่างกัน 2 วิธี</td><td rowspan="2">One-way MANOVA</td><td rowspan="2">แปลงค่า r เป็น Fisher'Z ทดสอบด้วย Z-test</td><td rowspan="3">Two- way ANOVA</td></tr>
<tr><td>1.2 เพื่อเปรียบเทียบลำดับข้อสอบที่แตกต่างกัน 2 รูปแบบ</td></tr>
<tr><td>1.3 เพื่อศึกษาปฏิสัมพันธ์ระหว่างวิธีการคัดเลือกข้อสอบขั้นแรก และลำดับข้อสอบ</td><td></td><td></td></tr>
<tr><th></th><th>ตัวแปรต้น</th><th>ตัวแปรตาม</th><th>สถิติวิเคราะห์</th></tr>
<tr><td>2. เพื่อศึกษาผลการประเมินโปรแกรมการทดสอบความสามารถด้านวิทยาศาสตร์ด้วยวิธีการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์</td><td>ภูมิหลังของนักเรียน
- GPAX
- ทักษะคอมพิวเตอร์</td><td>ผลการประเมิน
(rating scale 5 ระดับ)</td><td>การวิเคราะห์ถดถอยพหุคูณ
(multiple regression)</td></tr>
</table>

บทที่ 4

ผลการวิเคราะห์ข้อมูล

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อ ศึกษาคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์ภายใต้วิธีการคัดเลือกข้อสอบขั้นแรกและลำดับข้อสอบที่แตกต่างกัน รวมถึงศึกษาปฏิสัมพันธ์ระหว่างวิธีการคัดเลือกข้อสอบขั้นแรกและลำดับข้อสอบที่มีต่อคุณภาพของการทดสอบ และเพื่อศึกษาผลการประเมินการทดสอบความสามารถด้านวิทยาศาสตร์ด้วยวิธีการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์โดยศึกษาจากการจำลองข้อมูล ข้อมูลทุติยภูมิ และข้อมูลปฐมภูมิ ผู้วิจัยได้แบ่งการนำเสนอผลการวิเคราะห์ข้อมูลตามแหล่งข้อมูลที่ศึกษาออกเป็น 4 ตอน ดังนี้

ตอนที่ 1 การวิเคราะห์ข้อมูลกรณีการจำลองข้อมูล

1.1 ค่าสถิติพื้นฐานของพารามิเตอร์ข้อสอบและความสามารถผู้สอบที่จำลองขึ้น

1.2 ผลการวิเคราะห์และเปรียบเทียบคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติกรณีการจำลองข้อมูล

ตอนที่ 2 การวิเคราะห์ข้อมูลกรณีการใช้ข้อมูลทุติยภูมิ

2.1 ค่าสถิติพื้นฐานของข้อสอบและผู้สอบความถนัดทางวิชาชีพและวิชาการด้านวิทยาศาสตร์ (PAT2)

2.2 ผลการวิเคราะห์และเปรียบเทียบคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติกรณีการใช้ข้อมูลทุติยภูมิ

ตอนที่ 3 การวิเคราะห์ข้อมูลกรณีการใช้ข้อมูลปฐมภูมิ

3.1 ค่าสถิติพื้นฐานของผู้สอบวัดความสามารถด้านวิทยาศาสตร์

3.2 ผลการวิเคราะห์คุณภาพของการทดสอบปรับเหมาะแบบพหุมิติกรณีการใช้ข้อมูลปฐมภูมิ

3.3 ผลการเปรียบเทียบคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติกรณีการใช้ข้อมูลปฐมภูมิ จำแนกตามลำดับข้อสอบ

3.4 ค่าสถิติพื้นฐานภูมิหลังของนักเรียนและผลประเมินการใช้โปรแกรมการทดสอบปรับเหมาะแบบพหุมิติ

3.5 ผลการศึกษาภูมิหลังของนักเรียนที่ส่งผลต่อผลการประเมินโปรแกรมการทดสอบปรับเหมาะแบบพหุมิติ

ตอนที่ 4 สรุปผลการวิเคราะห์ข้อมูลจาก 3 แหล่งข้อมูล

เพื่อให้ง่ายต่อการนำเสนอผลการวิเคราะห์ข้อ มูล ผู้วิจัยจึงกำหนดสัญลักษณ์ที่ใช้แทนค่าตัวแปร และค่าสถิติต่าง ๆ ไว้ดังนี้

M	หมายถึง	ค่ามัชฌิมเลขคณิต หรือค่าเฉลี่ย (mean)
SD	หมายถึง	ส่วนเบี่ยงเบนมาตรฐาน (standard deviation)
r	หมายถึง	ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างความสามารถที่ประมาณค่าได้และค่าความสามารถที่แท้จริง
r^2	หมายถึง	ค่าความเที่ยง (reliability)
ED	หมายถึง	ค่าระยะทางแบบยุคลิด (Euclidean distance)
Z_r	หมายถึง	คะแนน Fisher's Z
Z	หมายถึง	สถิติทดสอบ Z-test
p	หมายถึง	ระดับนัยสำคัญทางสถิติ
$I(\theta)$	หมายถึง	ค่าสารสนเทศแบบสอบ (test information)
SE	หมายถึง	ความคลาดเคลื่อนมาตรฐานในการประมาณค่า
FI	หมายถึง	วิธีการคัดเลือกข้อสอบขั้นแรกแบบ Fisher information
KL	หมายถึง	วิธีการคัดเลือกข้อสอบขั้นแรกแบบ Kullback-Leibler information
Mixed	หมายถึง	ลำดับข้อสอบแบบผสมผสานมิติ
Classified	หมายถึง	ลำดับข้อสอบแบบจำแนกมิติ

ตอนที่ 1 การวิเคราะห์ข้อมูลกรณีการจำลองข้อมูล

การศึกษาด้วยการจำลองข้อมูลมีจุดเด่นอยู่ที่ผู้วิจัยทราบค่าความสามารถที่แท้จริงของผู้สอบ ซึ่งทำให้สามารถวิเคราะห์ความแม่นยำของค่าที่เกิดจากการประมาณค่าได้ อีกทั้งยังสามารถจัดกระทำกับข้อมูล ที่ไม่สามารถกระทำได้ กับการทดสอบจริง ซึ่งรายละเอียดของการวิเคราะห์ข้อมูลจากการจำลองข้อมูลมีดังนี้

1.1 ค่าสถิติพื้นฐานของพารามิเตอร์ข้อสอบและความสามารถผู้สอบที่จำลองขึ้น

สำหรับการจำลองข้อมูลครั้งนี้ ผู้วิจัยได้กำหนดความสามารถเป็น 2 มิติ โดยค่าพารามิเตอร์อำนาจจำแนกกำหนดให้ มีค่าเฉลี่ยเป็น 0.5 และส่วนเบี่ยงเบนมาตรฐานเป็น0.1 โดยมีการแจกแจงเป็นโค้งปกติตัดปลาย ค่าความยากกำหนดให้มีความยากอยู่ระหว่าง -3 ถึง 3 มีการแจกแจงแบบเอกรูป และโอกาสการเดาอยู่ระหว่าง0 ถึง 0.1 และมีการแจกแจงแบบเอกรูปสำหรับความสามารถของผู้สอบใน 2 มิติ ได้กำหนดให้มีค่าเป็นโค้งปกติโดยมิติที่ 1 มีความสามารถ อยู่ในช่วง -3.37 ถึง 3.44 และมิติที่ 2 มีความสามารถอยู่ในช่วง -3.89 ถึง 3.59 รายละเอียดดังตารางที่ 4.1

ตารางที่ 4.1 ค่าสถิติพื้นฐานของพารามิเตอร์ข้อสอบและความสามารถผู้สอบที่จำลองขึ้น

ค่าพารามิเตอร์	N	Min	Max	*M*	*SD*	Sk	Ku
อำนาจจำแนกมิติที่ 1 (a1)	150	0.16	0.78	0.50	0.10	-0.05	0.02
อำนาจจำแนกมิติที่ 2 (a2)	150	0.19	0.82	0.50	0.10	-0.14	0.26
ความยาก (b)	300	-3.00	3.00	0.00	1.74	0.00	-1.20
โอกาสการเดา (c)	300	0.00	0.10	0.05	0.03	0.03	-1.22
ความสามารถผู้สอบ							
ความสามารถมิติที่ 1	10000	-3.37	3.44	0.00	1.00	-0.02	-0.03
ความสามารถมิติที่ 2	10000	-3.89	3.59	0.00	1.00	-0.01	-0.06

1.2 ผลการวิเคราะห์และเปรียบเทียบคุณภาพของการ ทดสอบปรับเหมาะแบบพหุมิติ กรณีการจำลองข้อมูล

การวิเคราะห์คุณภาพของการทดสอบปรับเหมาะแบบพหุมิติกรณีการจำ ลองข้อมูล พิจารณาค่าสถิติ 2 ค่า ได้แก่ ค่าความเที่ยง และค่าระยะทาง แบบยุคลิด ซึ่งเป็นการวิเคราะห์ความแม่นยำของการประมาณค่าความสามารถ โดยผู้วิจัยได้นำเสนอจำแนกตาม คุณภาพของการทดสอบทั้ง 2 ค่า โดยนำเสนอค่าความเที่ยงและค่าระยะทางแบบยุคลิด พร้อมการเปรียบเทียบคุณภาพการทดสอบจำแนกตามวิธีการคัดเลือกข้อสอบขั้นแรกและ ลำดับข้อสอบ ซึ่งมีรายละเอียดดังนี้

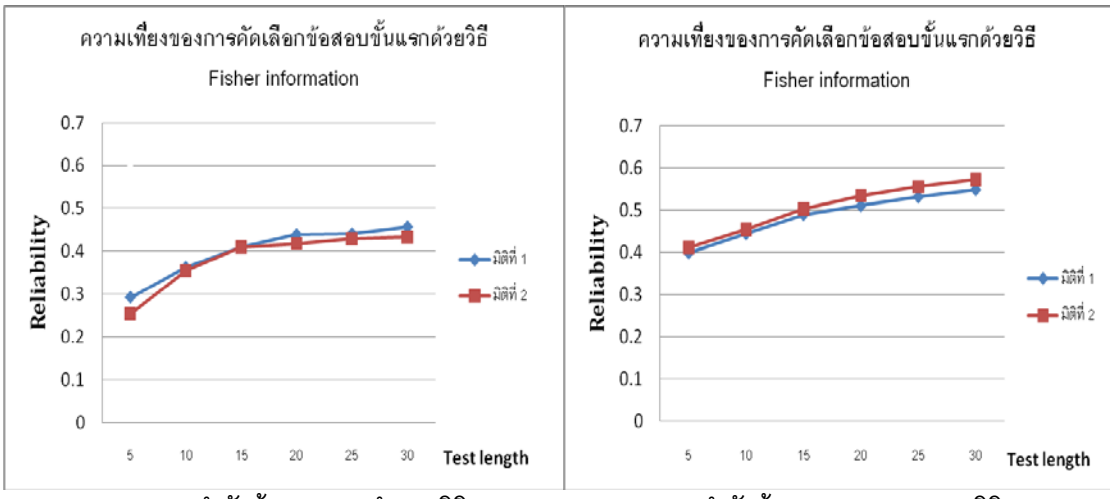
1.2.1 ผลการวิเคราะห์คุณภาพของการทดสอบเมื่อพิจารณาจากค่าความเที่ยง (กรณีการจำลองข้อมูล)

กรณีพิจารณาคุณภาพการทดสอบปรับเหมาะแบบพหุมิติด้วยค่าความเที่ยง พบว่า เมื่อพิจารณาค่าเฉลี่ยจากการทำข้อสอบ 5 ข้อ ซึ่งถือเป็นขั้นแรกของการ ทดสอบ พบว่า วิธี Fisher information มีค่าเฉลี่ยสำหรับลำดับข้อสอบแบบจำแนกมิติและผสมผสานมิติ เท่ากับ 0.273 และ 0.406 ตามลำดับ สำหรับวิธี Kullback-leibler information มีค่าเฉลี่ยเท่ากับ 0.322 และ 0.348 ตามลำดับ เมื่อพิจารณาค่าเฉลี่ยความเที่ยงสำหรับการทดสอบ 30 ข้อ วิธี Fisher information มีค่าเฉลี่ยสำหรับ ลำดับข้อสอบ แบบจำแนกมิติและผสมผสานมิติเท่ากับ 0.445 และ 0.560 ตามลำดับ สำหรับวิธี Kullback-leibler information มีค่าเฉลี่ยเท่ากับ 0.756 และ 0.730 ตามลำดับ อย่างไรก็ตาม เมื่อพิจารณาค่าความเที่ยงแต่ละช่วงความยาวข้อสอบจะพบว่า ความเที่ยงมีแนวโน้มสูงขึ้นเมื่อทำข้อสอบมากขึ้น ซึ่งสอดคล้องกันทุกเงื่อนไขการทดสอบ รายละเอียดดังตารางที่ 4.2

ตารางที่ 4.2 ความเที่ยง (r^2) ของการทดสอบปรับเหมาะ แบบพหุมิติจำแนกตามความยาวข้อสอบ (กรณีการจำลองข้อมูล)

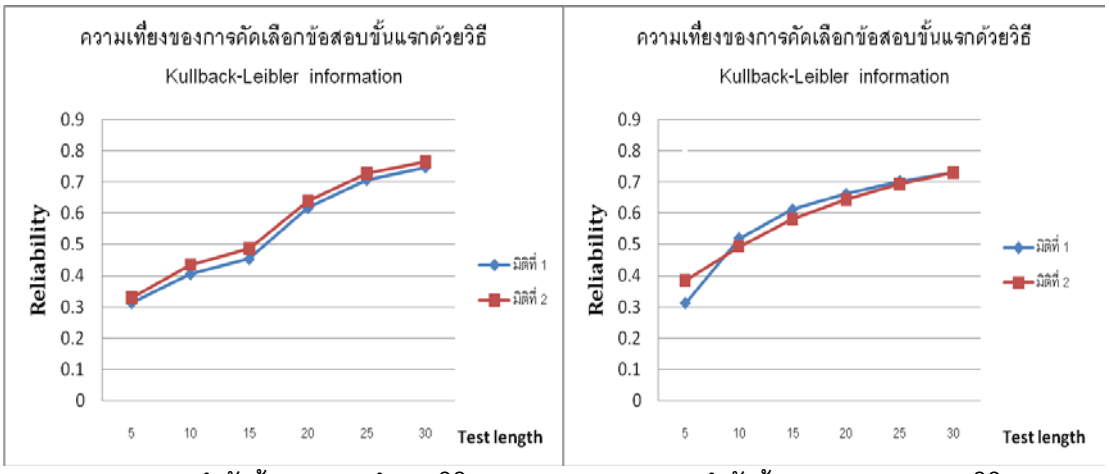
วิธีการคัดเลือกข้อสอบ	ลำดับข้อสอบแบบจำแนกมิติ			ลำดับข้อสอบแบบผสมผสานมิติ		
	r^2 มิติ 1	r^2 มิติ 2	r^2 เฉลี่ย	r^2 มิติ 1	r^2 มิติ 2	r^2 เฉลี่ย
1. Fisher information						
5 ข้อ	0.292	0.254	**0.273**	0.399	0.411	**0.406**
10 ข้อ	0.362	0.354	**0.358**	0.444	0.454	**0.449**
15 ข้อ	0.410	0.409	**0.410**	0.489	0.503	**0.496**
20 ข้อ	0.439	0.418	**0.429**	0.510	0.534	**0.523**
25 ข้อ	0.441	0.429	**0.435**	0.531	0.555	**0.543**
30 ข้อ	0.457	0.433	**0.445**	0.548	0.572	**0.560**
2. Kullback-leibler information						
5 ข้อ	0.314	0.330	**0.322**	0.312	0.384	**0.348**
10 ข้อ	0.406	0.435	**0.420**	0.519	0.493	**0.506**
15 ข้อ	0.454	0.486	**0.470**	0.611	0.581	**0.596**
20 ข้อ	0.617	0.639	**0.628**	0.661	0.644	**0.652**
25 ข้อ	0.707	0.728	**0.718**	0.702	0.693	**0.698**
30 ข้อ	0.747	0.765	**0.756**	0.730	0.730	**0.730**

เมื่อพิจารณาจากภาพที่ 4.1 และ 4.2 จะเห็นได้ว่าความเที่ยงของความสามารถทั้ง 2 มิติ มีแนวโน้มไปในทิศทางเดียวกัน และการทดสอบปรับเหมาะแบบพหุมิติที่มี ลำดับข้อสอบแบบผสมผสานมิติมีความเที่ยงสูงกว่า ลำดับข้อสอบ แบบจำแนกมิติ และหากเปรียบเทียบวิธีการคัดเลือกข้อสอบจะเห็น ได้ว่า วิธี Kullback-leibler information มีความเที่ยงสูงกว่า วิธี Fisher information



a. ลำดับข้อสอบแบบจำแนกมิติ b. ลำดับข้อสอบแบบผสมผสานมิติ

ภาพที่ 4.1 ความเที่ยงของการคัดเลือกข้อสอบขั้นแรกด้วยวิธี Fisher information (กรณีการจำลองข้อมูล)



a. ลำดับข้อสอบแบบจำแนกมิติ b. ลำดับข้อสอบแบบผสมผสานมิติ

ภาพที่ 4.2 ความเที่ยงของการคัดเลือกข้อสอบขั้นแรกด้วยวิธี Kullback-leibler information (กรณีการจำลองข้อมูล)

1.2.2 ผลการเปรียบเทียบค่าความเที่ยง (กรณีการจำลองข้อมูล)

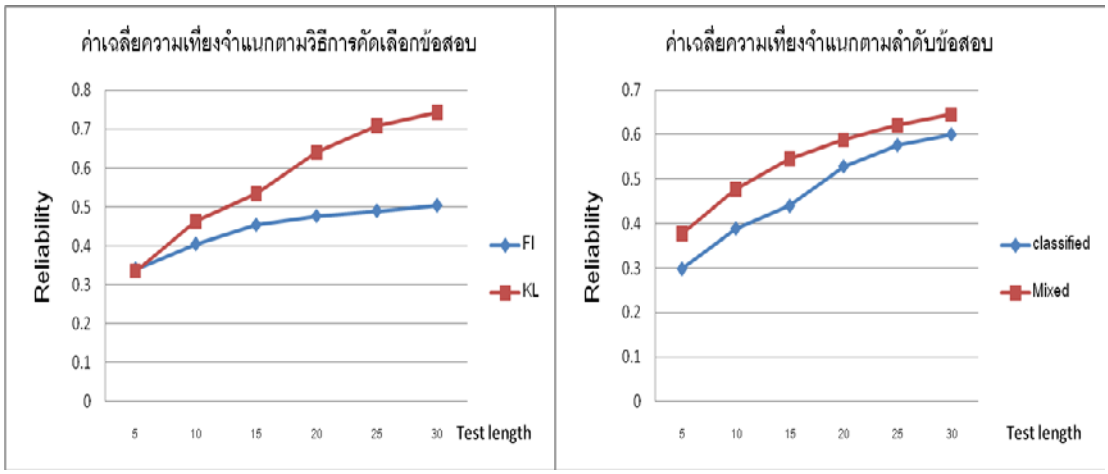
การเปรียบเทียบ คุณภาพของการทดสอบปรับเหมาะแบบพหุ มิติเมื่อพิจารณาจากค่าความเที่ยง ผู้วิจัยจะใช้การเปรียบเทียบค่าสัมประสิทธิ์สหสัมพันธ์ เนื่องจากค่าความเที่ยงเป็นการหากำลังสองของความสัมพันธ์ระหว่างค่าความสามารถที่ประมาณได้ และค่าความ สามารถที่แท้จริง ซึ่งวิเคราะห์จากผู้สอบทั้งหมดในกา รทดสอบ ทำ ให้ความเที่ยงจะมีเพียงหนึ่ง ค่าต่อหนึ่งการทดสอบ จึงไม่สามารถเปรียบเทียบด้วยสถิติวิเคราะห์อื่นๆ ได้ ดังนั้นผู้วิจัยจึงใช้การทดสอบ

ความแตกต่างของค่าสัมประสิทธิ์สหสัมพันธ์แทนการทดสอบค่าความเที่ยง ซึ่งจะกล่าวถึงรายละเอียดในลำดับต่อไป

ผลการศึกษาค่าเฉลี่ยความเที่ยงพบว่า วิธีการคัดเลือกข้อสอบด้วยวิธี kullback-leibler information มีค่าเฉลี่ยความเที่ยงโดยรวมสูงกว่าวิธี Fisher information เมื่อพิจารณาลำดับข้อสอบ พบว่า โดยภาพรวม ลำดับข้อสอบแบบผสมผสานมีความเที่ยงสูงกว่าลำดับข้อสอบแบบจำแนก รายละเอียดดังตารางที่ 4.3 และเพื่อให้เห็นภาพชัดเจนผู้วิจัยได้นำเสนอไว้ดังภาพที่ 4.3

ตารางที่ 4.3 ความเที่ยงเฉลี่ย (r^2) ของการทดสอบจำแนกตามวิธีการคัดเลือกข้อสอบและลำดับข้อสอบที่แตกต่างกัน (กรณีการจำลองข้อมูล)

วิธีการคัดเลือกข้อสอบ และลำดับข้อสอบ	ค่าความเที่ยงเฉลี่ยทั้ง 2 มิติ					
	5 ข้อ	10 ข้อ	15 ข้อ	20 ข้อ	25 ข้อ	30 ข้อ
1. วิธีการคัดเลือกข้อสอบ						
Fisher information	0.340	0.404	0.453	0.476	0.489	0.503
Kullback-Leibler information	0.335	0.463	0.533	0.640	0.708	0.743
เฉลี่ย	0.338	0.434	0.493	0.559	0.599	0.623
2. ลำดับข้อสอบ						
ลำดับข้อสอบแบบจำแนกมิติ	0.298	0.389	0.440	0.529	0.577	0.601
ลำดับข้อสอบแบบผสมผสานมิติ	0.377	0.478	0.546	0.588	0.621	0.645
เฉลี่ย	0.338	0.434	0.493	0.559	0.599	0.623



a. วิธีการคัดเลือกข้อสอบ b. ลำดับข้อสอบ

ภาพที่ 4.3 ความเที่ยงเฉลี่ยของวิธีการคัดเลือกข้อสอบและลำดับข้อสอบ (กรณีการจำลองข้อมูล)

อย่างไรก็ตาม เพื่อให้ผลการวิจัยมีความน่าเชื่อถือ มากขึ้น ผู้วิจัยจึงใช้การทดสอบความแตกต่างของค่าสัมประสิทธิ์สหสัมพันธ์ มาทดสอบความแตกต่างของค่าความเที่ยง เนื่องจากค่าความเที่ยงเป็นค่ากำลังสองของสัมประสิ ทธิ์สหสัมพันธ์ โดยจะใช้การแปลงค่าสัมประสิทธิ์สหสัมพันธ์ เป็นคะแนนมาตรฐาน Z_r เพื่อให้ข้อมูลมีการแจกแจงเข้าใกล้โค้งปกติมากที่สุด โดยใช้ตารางการแปลงค่าของ Fisher's Z transformation (ทรงศักดิ์ ภูสีอ่อน , 2552) และทดสอบความแตกต่างด้วยสถิติ Z-test ซึ่งผลการทดสอบ ค่าสัมประสิทธิ์สหสัมพันธ์ จะสรุปอ้างอิงไปถึงค่าความเที่ยงของการทดสอบ ผลการทดสอบความแตกต่างของค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างวิธีการคัดเลือกข้อสอบ พบว่า วิธี Fisher information (r = 0.583) และ Kullback-Leibler information(r= 0.579) มีค่าสัมประสิทธิ์สหสัมพั นธ์ไม่แตกต่างกัน สำหรับ ผลการทดสอบความแตกต่างของค่าสัมประสิทธิ์สหสัมพันธ์ระหว่าง ลำดับข้อสอบ พบว่า ลำดับข้อสอบ แบบผสมผสานมิติ (r=0.614) มีสัมประสิทธิ์สหสัมพันธ์สูง กว่าลำดับข้อสอบแบบจำแนกมิติ (r=0.545) อย่างมีนัยสำคัญ ที่ระดับ .01 รายละเอียดดังตารางที่ 4.4

ตารางที่ 4.4 การทดสอบความแตกต่างของค่าสัมประสิทธิ์สหสัมพันธ์ (กรณีการจำลองข้อมูล)

วิธีการคัดเลือกข้อสอบ และลำดับข้อสอบ	r^2	r	Z_r	Z
1. **วิธีการคัดเลือกข้อสอบ** (N=10000)				
Fisher information	0.340	0.583	0.667	0.424
Kullback-Leibler information	0.335	0.579	0.661	
2. **ลำดับข้อสอบ** (N=10000)				
ลำดับข้อสอบแบบจำแนกมิติ	0.298	0.545	0.611	-7.353**
ลำดับข้อสอบแบบผสมผสานมิติ	0.377	0.614	0.715	

*p< .05 **p<.01

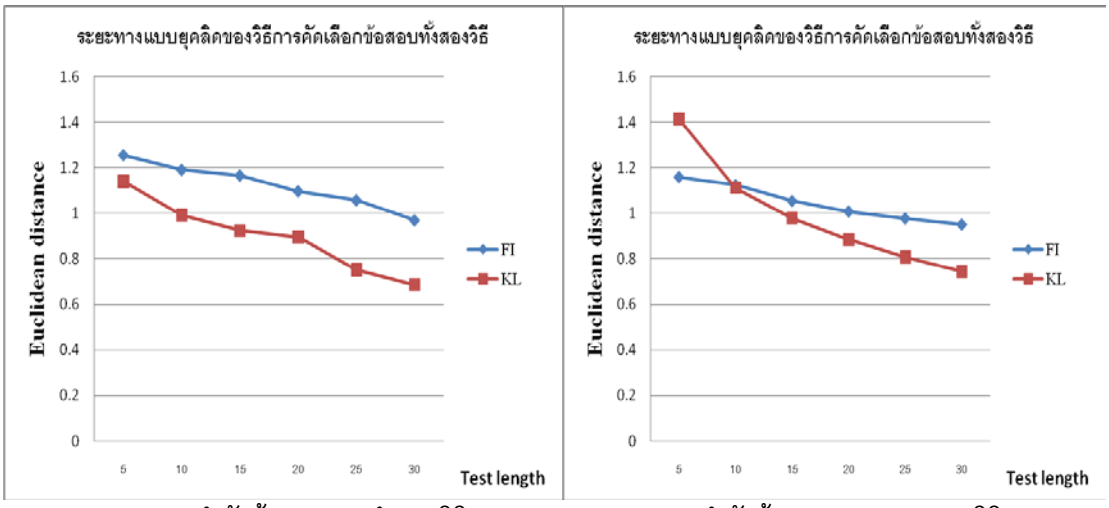
1.2.3 ผลการวิเคราะห์คุณภาพของการทดสอบมื่อพิจารณาจากระยะทางแบบยุคลิด (กรณีการจำลองข้อมูล)

กรณีพิจารณาคุณภาพการทดสอบปรับเหมาะแบบพหุมิติด้วยค่าระยะทางแบบยุคลิด พบว่า เมื่อพิจารณาระยะทางแบบยุคลิดสำหรับการทดสอบ 5 ข้อ ซึ่งถือเป็นระยะแรกของการทดสอบ การคัดเลือกข้อสอบขั้นแรกด้วยวิธี Fisher information จำแนกตามลำดับข้อสอบแบบจำแนกมิติและผสมผสานมิติ มีค่าเฉลี่ยระยะทางแบบยุคลิดเท่ากับ 1.256 และ1.160 ตามลำดับ สำหรับวิธี Kullback-leibler information มีค่าเฉลี่ยระยะทางแบบยุคลิดเท่ากับ 1.141 และ 1.413 ตามลำดับ เมื่อพิจารณา ระยะทางแบบยุคลิดสำหรับการทดสอบ 30 ข้อ หรือจบการทดสอบ

การคัดเลือกข้อสอบขั้นแรกด้วยวิธี Fisher information จำแนกตามลำดับข้อสอบแบบจำแนกมิติและผสมผสานมิติ มีค่าเฉลี่ยระยะทางแบบยุคลิดเท่ากับ 0.969 และ0.950 ตามลำดับ สำหรับวิธี Kullback-leibler information มีค่าเฉลี่ยระยะทางแบบยุคลิดเท่ากับ 0.687 และ0.745 ตามลำดับ ซึ่งการพิจารณาคุณภาพการทดสอบจากค่าระยะทางแบบยุคลิดนั้น สามารถอธิบายได้ว่า ค่าระยะทางแบบยุคลิดต่ำ หมายถึง การประมาณค่าความสามารถมีความแม่นยำสูงหรือมีคุณภาพการทดสอบสูง และ ค่าระยะทางแบบยุคลิดสูง หมายถึง การประมาณค่าความสามารถมีความแม่นยำต่ำหรือมีคุณ ภาพการทดสอบต่ำ รายละเอียดดังตารางที่ 4.5 นอกจากนี้ ผู้วิจัยได้นำเสนอภาพเพื่อแสดงระยะทางแบบยุคลิดจำแนกตามช่วงความยาวข้อสอบเพื่อให้เห็นแนวโน้มของการทดสอบ ซึ่งจะเห็นว่า เมื่อความยาวข้อสอบมากขึ้นระยะทางแบบยุคลิดจะลดต่ำลง และพบว่าการคัดเลือกข้อสอบด้วยวิธี Kullback-leibler information มีค่าระยะทางแบบยุคลิดต่ำกว่า วิธี Fisher information แต่ความแตกต่างของลำดับข้อสอบยังไม่ชัดเจน ดังภาพที่ 4.4

ตารางที่ 4.5 ระยะทางแบบยุคลิด (ED) ของการทดสอบปรับเหมาะแบบพหุมิติจำแนกตามความยาวข้อสอบ (กรณีการจำลองข้อมูล)

วิธีการคัดเลือกข้อสอบขั้นแรก	ลำดับข้อสอบแบบจำแนกมิติ				ลำดับข้อสอบแบบผสมผสานมิติ			
	M	*SD*	Min	Max	*M*	*SD*	Min	Max
1. Fisher information (N=10000)								
5 ข้อ	1.256	0.716	0.020	5.370	1.160	0.629	0.016	4.198
10 ข้อ	1.191	0.614	0.010	5.820	1.124	0.670	0.018	4.822
15 ข้อ	1.166	0.567	0.010	4.540	1.054	0.654	0.011	4.510
20 ข้อ	1.097	0.720	0.010	4.720	1.008	0.646	0.004	5.072
25 ข้อ	1.057	0.599	0.020	4.590	0.977	0.632	0.011	4.838
30 ข้อ	0.969	0.550	0.010	3.860	0.950	0.623	0.007	4.900
2. Kullback-leibler information(N=10000)								
5 ข้อ	1.141	0.652	0.008	5.375	1.413	0.775	0.019	5.900
10 ข้อ	0.992	0.560	0.011	4.107	1.112	0.637	0.005	5.648
15 ข้อ	0.924	0.529	0.003	3.830	0.979	0.579	0.002	4.958
20 ข้อ	0.895	0.513	0.006	3.781	0.885	0.537	0.005	4.324
25 ข้อ	0.752	0.407	0.010	3.087	0.807	0.494	0.007	4.241
30 ข้อ	0.687	0.367	0.005	2.965	0.745	0.464	0.005	4.205



a. ลำดับข้อสอบแบบจำแนกมิติ b. ลำดับข้อสอบแบบผสมผสานมิติ

ภาพที่ 4.4 ระยะทางแบบยุคลิดเฉลี่ยของวิธีการคัดเลือกข้อสอบขั้นแรกทั้ง 2 วิธี (กรณีการจำลองข้อมูล)

1.2.4 ผลการเปรียบเทียบค่าระยะทางแบบยุคลิด (กรณีการจำลองข้อมูล)

ผลการเปรียบเทียบคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติ เมื่อพิจารณาจากค่าระยะทางแบบยุคลิด ซึ่งค่าระยะทางต่ำหมายถึง การทดสอบมีคุณภาพสูง และค่าระยะทางสูงหมายถึง การทดสอบมีคุณภาพต่ำโดยค่าเฉลี่ยระยะทางแบบยุคลิด สำหรับการทดสอบขั้นแรก5 ข้อ การคัดเลือกข้อสอบด้วยวิธี Fisher information และ Kullback -Leibler information มีค่าเฉลี่ยเท่ากับ 1.208 และ 1.074 ตามลำดับ สำหรับ ลำดับข้อสอบ แบบจำแนกมิติและผสมผสานมิติ มีค่าเฉลี่ยเท่ากับ 1.199 และ 1.083 ตามลำดับ รายละเอียดดังตารางที่ 4.6

ตารางที่ 4.6 ค่าเฉลี่ยระยะทาง แบบยุคลิดจำแนกตามวิธีการคัดเลือกข้อสอบ ขั้นแรกและ ลำดับข้อสอบ (กรณีการจำลองข้อมูล)

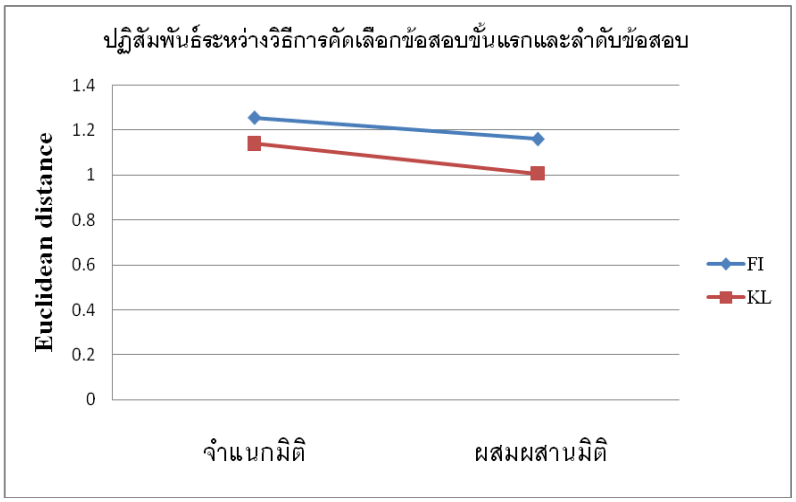
การคัดเลือกข้อสอบ	ลำดับข้อสอบ	N	*M*	*SD*
1. Fisher information	ลำดับข้อสอบแบบจำแนกมิติ	10000	1.256	0.716
	ลำดับข้อสอบแบบผสมผสานมิติ	10000	1.160	0.629
	รวม	**20000**	**1.208**	**0.676**
2. Kullback -Leibler information	ลำดับข้อสอบแบบจำแนกมิติ	10000	1.141	0.652
	ลำดับข้อสอบแบบผสมผสานมิติ	10000	1.007	0.548
	รวม	**20000**	**1.074**	**0.606**
รวม	ลำดับข้อสอบแบบจำแนกมิติ	20000	1.199	0.688
	ลำดับข้อสอบแบบผสมผสานมิติ	20000	1.083	0.595
	รวม	**40000**	**1.141**	**0.646**

สำหรับผลการเปรียบเทียบ คุณภาพของการทดสอบปรับเหมาะแบบพหุมิติ ผู้วิจัยใช้การวิเคราะห์ความแปรปรวนสองทาง (two-way ANOVA) เพื่อทดสอบความแตกต่างของค่าเฉลี่ยระยะทางแบบยุคลิด จำแนกตามวิธีการคัดเลือกข้อสอบขั้นแรกและลำดับข้อสอบ เมื่อพิจารณาการทดสอบขั้นแรก 5 ข้อ พบว่า ปฏิสัมพันธ์ระหว่างการคัดเลือกข้อสอบและลำดับข้อสอบร่วมกันส่งผลต่อค่าเฉลี่ยระยะทางแบบยุคลิด อย่างมีนัยสำคัญที่ระดับ .01 ($F=9.044$, $p = 0.003$) ดังตารางที่ 4.7 โดยพบว่า การคัดเลือกข้อสอบด้วยวิธี Kullback -Leibler information แบบผสมผสานมิติมีค่า ระยะทางแบบยุคลิดต่ำที่สุด (ED=1.007, *SD*=0.548) นั้นคือ มีคุณภาพสูงสุด และการคัดเลือกข้อสอบด้วยวิธี Fisher information แบบจำแนกมิติมีค่าระยะทางแบบยุคลิดสูงที่สุด (ED=1.256, *SD*=0.716) นั่นคือ มีคุณภาพต่ำสุด นอกจากนี้ผู้วิจัยได้นำเสนอ ภาพปฏิสัมพันธ์ของการคัดเลือกข้อสอบและลำดับข้อสอบไว้ดังภาพที่ 4.5

ตารางที่ 4.7 การทดสอบ ความแตกต่างของค่าเฉลี่ยระยะทางแบบยุคลิด ด้วยการวิเคราะห์ความแปรปรวนแบบสองทางจำแนกตามวิธีการคัดเลือกข้อสอบขั้นแรก และลำดับข้อสอบ

แหล่งความแปรปรวน	levene's test (p-value)	df	MS	F	P
Corrected Model	178.675	3	105.725	258.525	0.000
Intercept	(0.00)	1	52088.041	127368.725	0.000
การคัดเลือกข้อสอบ		1	180.241	440.736**	0.000
ลำดับข้อสอบ		1	133.236	325.795**	0.000
การคัดเลือกข้อสอบ * ลำดับข้อสอบ		1	3.699	9.044**	0.003
Error		39996	0.409		
Total		40000			
Corrected Total		39999			

**p < .01



ภาพที่ 4.5 ปฏิสัมพันธ์ระหว่างการคัดเลือกข้อสอบและลำดับข้อสอบ (กรณีการจำลองข้อมูล)

<u>สรุปผลการวิเคราะห์ข้อมูลจากการจำลองข้อมูล</u>

ผลการศึกษาคุณภาพการทดสอบจากค่าความเที่ยง พบว่า การคัดเลือกข้อสอบขั้นแรก5 ข้อ วิธี Fisher information (r^2= 0.340) และ Kullback-Leibler information(r^2= 0.335) มีค่าความเที่ยงไม่แตกต่างกัน สำหรับลำดับข้อสอบ พบว่า ลำดับข้อสอบแบบจำแนก มิติ (r^2= 0.298) และลำดับข้อสอบแบบผสมผสานมิติ(r^2= 0.377) มีค่าความเที่ยงไม่แตกต่างกัน

ผลการศึกษาคุณภาพการทดสอบจาก ระยะทางแบบยุคลิด พบว่า มีปฏิสัมพันธ์ระหว่างการคัดเลือกข้อสอบและ ลำดับข้อสอบ โดยที่วิธี Kullback-Leibler information แบบผสมผสานมิติ (ED=1.007) มีคุณภาพสูงสุด

ตอนที่ 2 การวิเคราะห์ข้อมูลกรณีการใช้ข้อมูลทุติยภูมิ

กรณีการใช้ข้อมูลทุติยภูมิ ใช้ผลการสอบ จากการทด สอบความถนัดทางวิชา ชีพและวิชาการด้านวิทยาศาสตร์ (PAT2) ของสถาบันทดสอบทางการศึกษาแห่งชาติ (สทศ.)ซึ่งการวิจัยครั้งนี้ศึกษาเฉพาะวิชาชีววิทยา เคมีและฟิสิกส์ เท่านั้น ผลวิเคราะห์ข้อมูลมีรายละเอียดดังนี้

2.1 ค่าสถิติพื้นฐานของข้อสอบและผู้สอบความถนัดทางวิชาชีพและวิชาการด้านวิทยาศาสตร์ (PAT2)

ค่าพารามิเตอร์ข้อสอบ ความถนัดทางวิชาชีพและวิชาการด้านวิทยาศาสตร์ (PAT2) ครั้งที่ 1/2552 ซึ่งคำนวณจากโปรแกรม TESTFACT และ BILOG มีค่าอำนาจจำแนก ค่อนข้างต่ำ ทั้งนี้เนื่องจาก โปรแกรม TESTFACT ให้ค่าพารามิเตอร์ที่ค่อนข้างต่ำ โดยมีค่าเฉลี่ยอำนาจจำแนก ชีววิทยา เคมี และฟิสิกส์ เท่ากับ 0.13, 0.14 และ 0.11 ตามลำดับ ค่าความยากมีค่าเฉลี่ยเท่ากับ 0.41 และโอกาสการเดามีค่าเฉลี่ยเท่ากับ 0.17

กลุ่มตัวอย่างของการวิ จัยครั้งนี้ คือ นักเรียนที่สอบ ความ ถนัดทางวิชาชีพและวิชาการด้านวิทยาศาสตร์ (PAT2) ครั้งที่ 1/2552 จำนวน 5,000 คน โดยความสามารถของผู้สอบจำแนกตามวิชาชีววิทยา เคมี และฟิสิกส์ มีค่าเฉลี่ยอยู่ที่ -4.807 -4.944 และ -4.601 ตามลำดับ ซึ่งถือว่ามีความสามารถระดับต่ำ อย่างไรก็ตาม การทดสอบถนัดทางวิชาชีพและวิชาการด้านวิทยาศาสตร์เป็นข้อสอบแข่งขันระดับชาติ ที่มีความยากสูง ทำให้ผู้สอบที่ถูกคัดเลือกมา 5,000 คนนั้นมีความสามารถระดับต่ำ รายละเอียดดังตารางที่ 4.8

ตารางที่ 4.8 ค่าพารามิเตอร์ข้อสอบและความสามา รถของผู้สอบความถนัดทางวิชาชีพและวิชาการด้านวิทยาศาสตร์ (PAT2)

ค่าพารามิเตอร์	N	Min	Max	*M*	*SD*	Sk	Ku
อำนาจจำแนกชีววิทยา (a1)	40	-0.05	0.34	0.13	0.10	-0.07	-0.91
อำนาจจำแนกเคมี (a2)	32	0.01	0.38	0.14	0.10	0.65	-0.72
อำนาจจำแนกฟิสิกส์ (a3)	28	-0.08	0.36	0.11	0.12	0.38	-0.83
ความยาก (b)	100	-0.75	1.21	0.41	0.39	-0.20	-0.12
โอกาสการเดา (c)	100	0.02	0.20	0.17	0.04	-1.50	2.45
ความสามารถผู้สอบ							
ความสามารถชีววิทยา	5000	-9.106	9.577	-4.807	1.395	0.804	4.944
ความสามารถเคมี	5000	-8.463	8.539	-4.944	1.300	0.922	5.131
ความสามารถฟิสิกส์	5000	-8.177	8.050	-4.601	1.370	0.692	2.956

2.2 ผลการวิเคราะห์ และเปรียบเที ยบคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติกรณีการใช้ข้อมูลทุติยภูมิ

การวิเคราะห์คุณภาพของการทดสอบปรับเหมาะแบบพหุมิติกรณีการใช้ข้อมูลทุติยภูมิ พิจารณาค่าสถิติ 2 ค่า ได้แก่ ค่าความเที่ยง และค่าระยะทางแบบยุคลิด ซึ่งเป็นการวิเคราะห์ความแม่นยำของการประมาณค่าความสามารถ โดยผู้วิจัยได้นำเสนอจำแนกตามคุณภาพของการทดสอบทั้ง 2 ค่า พร้อมทั้งเปรียบเทียบคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติ ซึ่งมีรายละเอียดดังนี้

2.2.1 ผลการวิเคราะห์คุณภาพของการทดสอบเมื่อพิจารณาจากค่าความเที่ยง (กรณีการใช้ข้อมูลทุติยภูมิ)

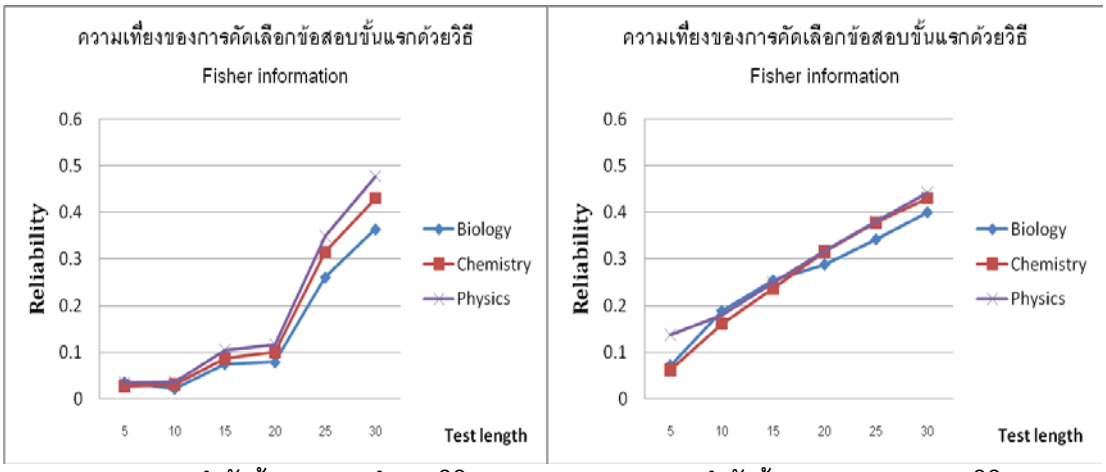
กรณีพิจารณาคุณภาพการทดสอบปรับเหมาะแบบพหุมิติด้วยค่าความเที่ยง พบว่า เมื่อพิจารณาค่าเฉลี่ยจากการทำข้อสอบ 5 ข้อ ซึ่งถือเป็นขั้นแรกของการทดสอบ พบว่า

วิธี Fisher information มีค่าเฉลี่ยสำหรับ ลำดับข้อสอบ แบบจำแนกมิติและผสมผสานมิติเท่ากับ 0.033 และ 0.090 ตามลำดับ สำหรับวิธี Kullback-leibler information มีค่าเฉลี่ยเท่ากับ 0.038 และ 0.088 ตามลำดับ เมื่อพิจารณาค่าเฉลี่ยความเที่ยงสำหรับการทดสอบ 30 ข้อด้วยวิธี Fisher information มีค่าเฉลี่ยสำหรับ ลำดับข้อสอบ แบบจำแนกมิติและผสมผสานมิติเท่ากับ 0.424 และ 0.424 ตามลำดับ สำหรับวิธี Kullback-leibler information มีค่าเฉลี่ยเท่ากับ 0.411 และ 0.364 ตามลำดับ อย่างไรก็ตาม เมื่อพิจารณาค่าความเที่ยงแต่ละช่วงความยาวข้อสอบจะพบว่า ความเที่ยงมีแนวโน้มสูงขึ้นเมื่อทำข้อสอบมากขึ้น ซึ่งสอดคล้องกันทุกเงื่อนไขการทดสอบ เพื่อให้เห็นภาพชัดเจนในแต่ละช่วงความยาวข้อสอบ รายละเอียดดังตารางที่4.9

เมื่อพิจารณาจาก ภาพที่ 4.6 และ 4.7 จะเห็นได้ว่าความเที่ยงของความสามารถทั้ง 3 มิติ (ชีววิทยา เคมี และฟิสิกส์) มีแนวโน้มไปในทิศทางเดียวกัน คือ ความเที่ยงสูงขึ้นเมื่อความยาวข้อสอบเพิ่มขึ้น อย่างไรก็ตามจาก ภาพยังไม่สามารถชี้ชัดได้ว่า วิธีการคัดเลือกข้อสอบ หรือ ลำดับข้อสอบ แบบใดที่มีค่าความเที่ยงสูงกว่า ซึ่งผู้วิจัยจะนำเสนอการทดสอบความแตกต่างของความเที่ยงในลำดับต่อไป

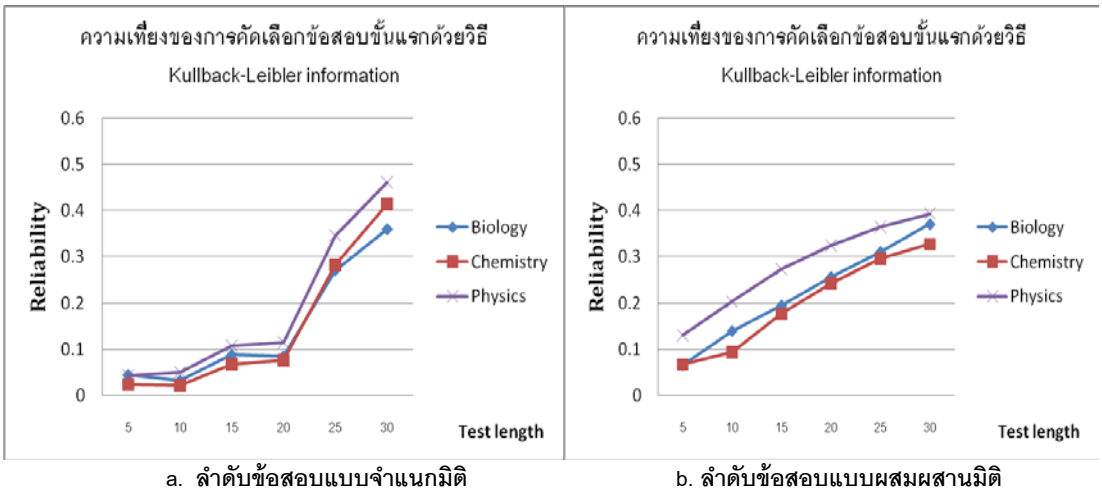
ตารางที่ 4.9 ความเที่ยง (r^2) ของการทดสอบปรับเหมาะแบบพหุมิติจำแนกตามความยาวข้อสอบ (กรณีการใช้ข้อมูลทุติยภูมิ)

วิธีการคัดเลือกข้อสอบ	ลำดับข้อสอบแบบจำแนกมิติ				ลำดับข้อสอบแบบผสมผสานมิติ			
	r^2 ชีววิทยา	r^2 เคมี	r^2 ฟิสิกส์	r^2 รวม	r^2 ชีววิทยา	r^2 เคมี	r^2 ฟิสิกส์	r^2 รวม
1. Fisher information								
5 ข้อ	0.036	0.027	0.035	**0.033**	0.072	0.062	0.137	**0.090**
10 ข้อ	0.022	0.031	0.036	**0.030**	0.188	0.161	0.179	**0.176**
15 ข้อ	0.075	0.086	0.105	**0.089**	0.254	0.237	0.250	**0.247**
20 ข้อ	0.079	0.101	0.117	**0.099**	0.288	0.316	0.317	**0.307**
25 ข้อ	0.261	0.315	0.349	**0.308**	0.342	0.378	0.380	**0.367**
30 ข้อ	0.364	0.430	0.477	**0.424**	0.400	0.430	0.442	**0.424**
2. Kullback-leibler information								
5 ข้อ	0.045	0.024	0.044	**0.038**	0.066	0.067	0.131	**0.088**
10 ข้อ	0.033	0.022	0.051	**0.035**	0.140	0.094	0.203	**0.146**
15 ข้อ	0.088	0.068	0.108	**0.088**	0.195	0.177	0.273	**0.215**
20 ข้อ	0.085	0.076	0.115	**0.092**	0.257	0.242	0.324	**0.274**
25 ข้อ	0.271	0.282	0.346	**0.300**	0.311	0.296	0.365	**0.324**
30 ข้อ	0.359	0.413	0.461	**0.411**	0.371	0.327	0.393	**0.364**



a. ลำดับข้อสอบแบบจำแนกมิติ b. ลำดับข้อสอบแบบผสมผสานมิติ

ภาพที่ 4.6 ความเที่ยงของการคัดเลือกข้อสอบขั้นแรกด้วยวิธี Fisher information (กรณีการใช้ข้อมูลทุติยภูมิ)



a. ลำดับข้อสอบแบบจำแนกมิติ b. ลำดับข้อสอบแบบผสมผสานมิติ

ภาพที่ 4.7 ความเที่ยงของการคัดเลือกข้อสอบขั้นแรกด้วยวิธี kullback-leibler information (กรณีการใช้ข้อมูลทุติยภูมิ)

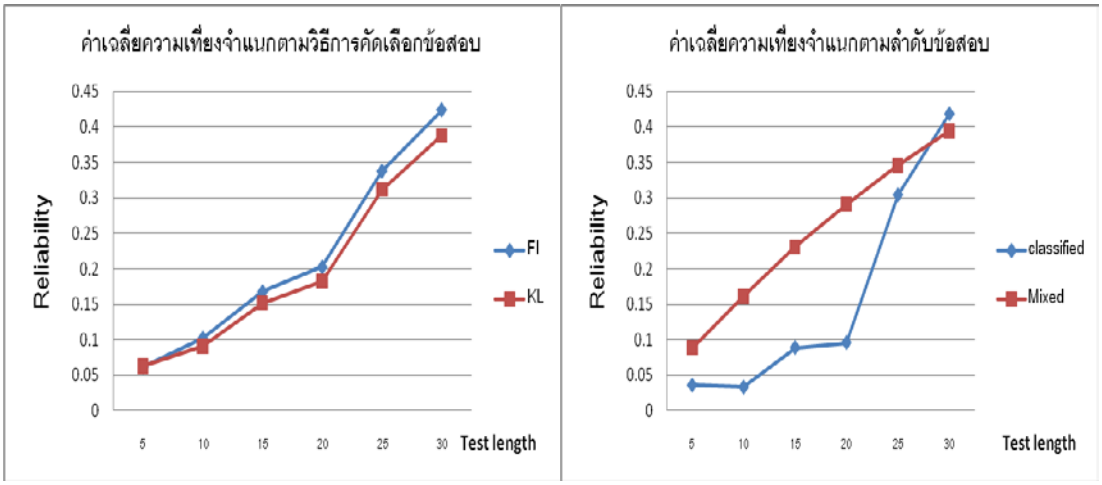
2.2.2 ผลการเปรียบเทียบค่าความเที่ยง (กรณีการใช้ข้อมูลทุติยภูมิ)

การเปรียบเทียบ คุณภาพของการทดสอบปรับเหมาะแบบพหุมิติ เมื่อพิจารณาจากค่าความเที่ยง ผู้วิจัย ใช้การ ทดสอบความแตกต่างของค่าสัมประสิทธิ์สหสัมพันธ์ เช่นเดียวกับการศึกษาจากการจำลองข้อมูล ซึ่งผลการทดสอบจะอธิบายในลำดับต่อไป

สำหรับผลการศึกษาค่าเฉลี่ยความเที่ยงพบว่า วิธีการคัดเลือกข้อสอบด้วยวิธี Fisher information มีค่าเฉลี่ยความเที่ยงโดยรวมสูงกว่าวิธี Kullback-leibler information เมื่อพิจารณาลำดับข้อสอบ พบว่า โดยภาพรวม ลำดับข้อสอบแบบผสมผสานมีความเที่ยงสูงกว่า ลำดับข้อสอบแบบจำแนก รายละเอียดดังตารางที่ 4.10 และภาพที่ 4.8

ตารางที่ 4.10 ความเที่ยง (r^2) ของการทดสอบปรับเหมาะแบบพหุมิติจำแนกตามวิธีการคัดเลือกข้อสอบและลำดับข้อสอบ (กรณีการใช้ข้อมูลทุติยภูมิ)

วิธีการคัดเลือกข้อสอบ และลำดับข้อสอบ	ค่าความเที่ยงเฉลี่ยทั้ง 3 วิชา					
	5 ข้อ	10 ข้อ	15 ข้อ	20 ข้อ	25 ข้อ	30 ข้อ
1. วิธีการคัดเลือกข้อสอบ						
Fisher information	0.062	0.103	0.168	0.203	0.338	0.424
Kullback-Leibler information	0.063	0.091	0.152	0.183	0.312	0.388
เฉลี่ย	0.063	0.097	0.160	0.193	0.325	0.406
2. ลำดับข้อสอบ						
ลำดับข้อสอบแบบจำแนกมิติ	0.036	0.033	0.089	0.096	0.304	0.418
ลำดับข้อสอบแบบผสมผสานมิติ	0.089	0.161	0.231	0.291	0.346	0.394
เฉลี่ย	0.063	0.097	0.160	0.193	0.325	0.406



a. วิธีการคัดเลือกข้อสอบ b. ลำดับข้อสอบ

ภาพที่ 4.8 ความเที่ยงเฉลี่ยของวิธีการคัดเลือกข้อสอบและลำดับข้อสอบ (กรณีการใช้ข้อมูลทุติยภูมิ)

อย่างไรก็ตาม เพื่อให้ผลการวิจัยมีความน่าเชื่อถือมากขึ้น ผู้วิจัยจึงใช้การทดสอบความแตกต่างของค่าสัมประสิทธิ์สหสัมพันธ์ มาทดสอบความแตกต่างของค่าความเที่ยง เนื่องจากค่าความเที่ยงเป็นค่ากำลังสองของสัมประสิทธิ์สหสัมพันธ์ ซึ่งผลการทดสอบจะอ้างอิงไปถึงค่าความเที่ยงของการทดสอบ ผลการทดสอบความแตกต่างของค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างวิธีการคัดเลือกข้อสอบ พบว่า วิธี Fisher information (r = 0.245) และ Kullback-Leibler information(r= 0.251) มีค่าสัมประสิทธิ์สหสัมพันธ์ไม่แตกต่างกัน สำหรับผลการทดสอบความแตกต่างของค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างลำดับข้อสอบ พบว่า ลำดับข้อสอบ แบบ

ผสมผสานมิติ (r=0.298) มีสัมประสิทธิ์สหสัมพันธ์สูงกว่า ลำดับข้อสอบแบบจำแนกมิติ (r=0.188) อย่างมีนัยสำคัญที่ระดับ .01 รายละเอียดดังตารางที่ 4.11

ตารางที่ 4.11 การทดสอบความแตกต่างของค่าสัมประสิทธิ์สหสัมพันธ์กรณีการใช้ข้อมูลทุติยภูมิ)

วิธีการคัดเลือกข้อสอบ และลำดับข้อสอบ	r^2	r	Z_r	Z
1. **วิธีการคัดเลือกข้อสอบ** (N=5000)				
Fisher information	0.062	0.249	0.254	-0.150
Kullback-Leibler information	0.063	0.251	0.257	
2. **ลำดับข้อสอบ** (N=5000)				
ลำดับข้อสอบแบบจำแนกมิติ	0.036	0.188	0.190	-5.850**
ลำดับข้อสอบแบบผสมผสานมิติ	0.089	0.298	0.307	

*p<.05 **p<.01

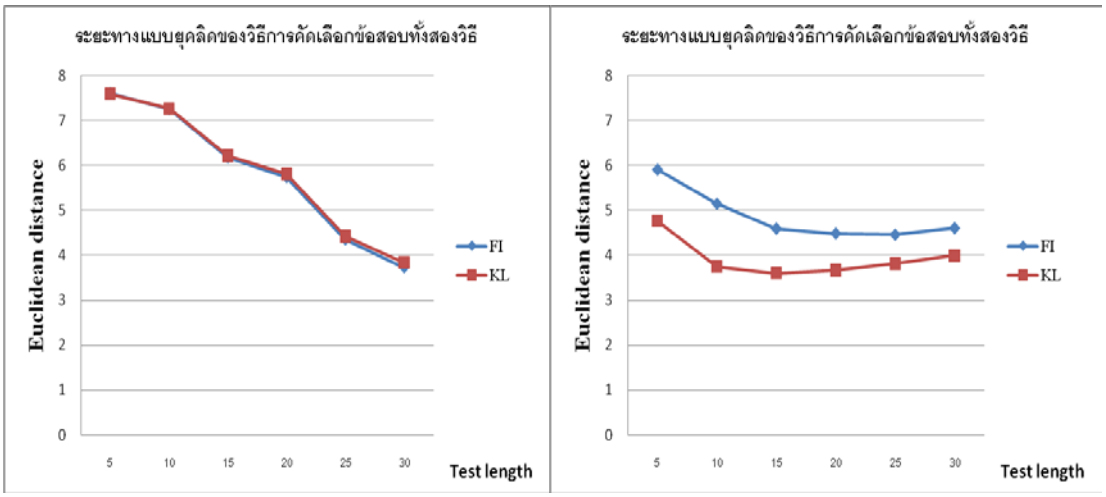
2.2.3 ผลการวิเคราะห์คุณภาพของการทดสอบมื่อพิจารณาจากระยะทางแบบยุคลิด (กรณีการใช้ข้อมูลทุติยภูมิ)

กรณีพิจารณาคุณภาพ การทดสอบปรับเหมาะแบบพหุมิติด้วยค่าระยะทางแบบยุคลิด จากตารางที่ 4.12 พบว่า เมื่อพิจารณาระยะทางแบบยุคลิดสำหรับการทดสอบ 5 ข้อ ซึ่งถือเป็นระยะแรกของการทดสอบการคัดเลือกข้อสอบด้วยวิธี Fisher information จำแนกตามลำดับข้อสอบแบบจำแนกมิติและผสมผสานมิติ มีค่าเฉลี่ยระยะทางแบบยุค ลิดเท่ากับ 7.606 และ 5.898 ตามลำดับ สำหรับวิธี Kullback-leibler information มีค่าเฉลี่ยระยะทางแบบยุคลิดเท่ากับ 7.592 และ 4.765 ตามลำดับ เมื่อพิจารณาระยะทางแบบยุคลิดสำหรับการทดสอบ 30 ข้อ หรือจบการทดสอบ การคัดเลือกข้อสอบขั้นแรกด้วยวิธี Fisher information จำแนกตามลำดับข้อสอบแบบจำแนกมิติและผสมผสานมิติ มีค่าเฉลี่ยระยะทางแบบยุคลิดเท่ากับ 3.730 และ 4.606 ตามลำดับ สำหรับวิธี Kullback-leibler information มีค่าเฉลี่ยระยะทางแบบยุคลิดเท่ากับ 3.842 และ 3.992 ตามลำดับ โดยการทดสอบจะให้ค่าระยะทางแบบยุคลิดต่ำลงเมื่อความยาวข้อสอบมากขึ้น นั่นคือคุณภาพการทดสอบจะสูงขึ้นเมื่อทำข้อสอบมากขึ้น

จากภาพที่ 4.9 แสดงระยะทางแบบยุคลิดจำแนกตามช่วงความยาวข้อสอบเพื่อให้เห็นแนวโน้มของการทดสอบ ซึ่งจะเห็นว่า เมื่อความยาวข้อสอบมากขึ้นระยะทางแบบยุคลิดจะลดต่ำลง และพบว่า กรณี ลำดับข้อสอบแบบ ผสมผสานมิติ การคัดเลือกข้อสอบด้วยวิธี Kullback-leibler information มีค่าระยะทางแบบยุคลิดต่ำกว่า วิ ธี Fisher information เมื่อพิจารณาลำดับข้อสอบแบบจำแนกมิติ พบว่า การคัดเลือกข้อสอบทั้งสองวิธีไม่มีความแตกต่างกัน

ตารางที่ 4.12 ระยะทางแบบยุคลิด (ED) ของการทดสอบปรับเหมาะ แบบพหุมิติจำแนกตามความยาวข้อสอบ (กรณีการใช้ข้อมูลทุติยภูมิ)

วิธีการคัดเลือกข้อสอบ	ลำดับข้อสอบแบบจำแนกมิติ				ลำดับข้อสอบแบบผสมผสานมิติ			
	M	*SD*	Min	Max	*M*	*SD*	Min	Max
1. Fisher information (N=5000)								
5 ข้อ	7.606	1.347	0.676	13.582	5.898	3.513	0.204	19.082
10 ข้อ	7.252	1.325	0.618	13.044	5.145	2.083	0.383	18.248
15 ข้อ	6.175	1.295	0.941	11.028	4.595	1.791	0.368	14.591
20 ข้อ	5.739	1.264	0.284	10.345	4.488	1.789	0.312	12.659
25 ข้อ	4.350	1.154	0.479	9.034	4.460	1.916	0.288	11.816
30 ข้อ	3.730	1.032	0.203	8.277	4.606	2.059	0.257	12.199
2. Kullback-leibler information (N=5000)								
5 ข้อ	7.592	1.350	0.676	13.700	4.765	3.414	0.148	18.603
10 ข้อ	7.264	1.339	0.604	13.242	3.746	2.067	0.149	15.584
15 ข้อ	6.222	1.309	0.858	11.412	3.599	1.647	0.192	14.854
20 ข้อ	5.809	1.272	0.991	10.773	3.670	1.602	0.300	15.329
25 ข้อ	4.423	1.161	0.436	8.656	3.819	1.639	0.220	13.312
30 ข้อ	3.842	1.037	0.284	8.048	3.992	1.719	0.152	12.441



a. ลำดับข้อสอบแบบจำแนกมิติ b. ลำดับข้อสอบแบบผสมผสานมิติ

ภาพที่ 4.9 ระยะทางแบบยุคลิดเฉลี่ยของวิธีการคัดเลือกข้อสอบขั้นแรกทั้ง 2 วิธี (กรณีการใช้ข้อมูลทุติยภูมิ)

2.2.4 ผลการเปรียบเทียบค่าระยะทางแบบยุคลิด (กรณีการใช้ข้อมูลทุติยภูมิ)

การเปรียบเทียบคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติเมื่อพิจารณาจากค่าระยะทางแบบยุคลิด สำหรับการทดสอบขั้นแรก5 ข้อ พบว่า การคัดเลือกข้อสอบด้วยวิธี Fisher information และ Kullback -Leibler information มีค่าเฉลี่ยเท่ากับ 6.752 และ 6.179 ตามลำดับ สำหรับ ลำดับข้อสอบ แบบจำแนกมิติและผสมผสานมิติ มีค่าเฉลี่ยเท่ากับ 7.599 และ 5.331 ตามลำดับ รายละเอียดดังตารางที่ 4.13

ตารางที่ 4.13 ค่าเฉลี่ยระยะทาง แบบยุคลิดจำแนกตามวิธีการคัดเลือกข้อสอบขั้นแรกและลำดับข้อสอบ (กรณีการใช้ข้อมูลทุติยภูมิ)

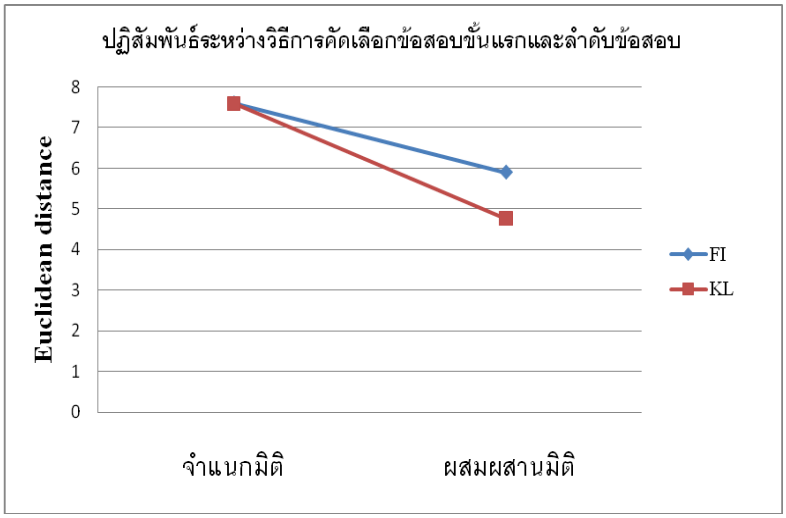
การคัดเลือกข้อสอบ	ลำดับข้อสอบ	N	*M*	*SD*
1. Fisher information	ลำดับข้อสอบแบบจำแนกมิติ	5000	7.606	1.347
	ลำดับข้อสอบแบบผสมผสานมิติ	5000	5.898	3.513
	รวม	**10000**	**6.752**	**2.794**
2. Kullback -Leibler information	ลำดับข้อสอบแบบจำแนกมิติ	5000	7.592	1.350
	ลำดับข้อสอบแบบผสมผสานมิติ	5000	4.765	3.414
	รวม	**10000**	**6.179**	**2.956**
รวม	ลำดับข้อสอบแบบจำแนกมิติ	10000	7.599	1.348
	ลำดับข้อสอบแบบผสมผสานมิติ	10000	5.331	3.510
	รวม	**20000**	**6.465**	**2.890**

สำหรับผลการเปรียบเทียบ คุณภาพของการทดสอบปรับเหมาะ แบบพหุมิติ ผู้วิจัยใช้การวิเคราะห์ความแปรปรวนสองทาง (two-way ANOVA) เพื่อทดสอบความแตกต่างของค่าเฉลี่ยระยะทาง แบบยุคลิด จำแนกตามวิธีการคัดเลือกข้อสอบขั้นแรกและ ลำดับข้อสอบ เมื่อพิจารณาการทดสอบขั้นแรก 5 ข้อ พบว่า ปฏิสัมพันธ์ระหว่างการคัดเลือกข้อสอบและ ลำดับข้อสอบร่วมกันส่งผลต่อค่าเฉลี่ยระยะทางแบบยุคลิด อย่างมีนัยสำคัญที่ระดับ .01 ($F=226.736$, $p=0.000$) ดังตารางที่ 4.14 โดยพบว่า การคัดเลือกข้อสอบด้วยวิธี Kullback -Leibler information แบบผสมผสานมิติมีค่าระยะทางแบบยุคลิดต่ำที่สุด (ED=4.765, *SD*=3.414) นั้นคือ มีคุณภาพสูงสุด และ การคัดเลือกข้อสอบด้วยวิธี Fisher information แบบจำแนกมิติมีค่าระยะทางแบบยุคลิดสูงที่สุด (ED=7.606, *SD*=1.347) นั้นคือ มีคุณภาพต่ำสุด นอกจากนี้ผู้วิจัยได้นำเสนอ ภาพปฏิสัมพันธ์ของการคัดเลือกข้อสอบและลำดับข้อสอบไว้ดังภาพที่ 4.10

ตารางที่ 4.14 การทดสอบ ความแตกต่างของค่าเฉลี่ย ระยะทางแบบยุคลิด ด้วยการวิเคราะห์ความแปรปรวนแบบสองทาง จำแนกตามวิธีการคัดเลือกข้อสอบขั้นแรกและลำดับข้อสอบ (กรณีการใช้ข้อมูลทุติยภูมิ)

แหล่งความแปรปรวน	levene's test (p-value)	df	MS	F	P
Corrected Model	2017.217	3	9639.675	1395.368	0.000
Intercept	(0.00)	1	835980.00	121010.28	0.000
การคัดเลือกข้อสอบ		1	1642.277	237.724**	0.000
ลำดับข้อสอบ		1	25710.383	3721.645**	0.000
การคัดเลือกข้อสอบ * ลำดับข้อสอบ		1	1566.366	226.736**	0.000
Error		19996	6.908		
Total		20000			
Corrected Total		19999			

**p < .01



ภาพที่ 4.10 ปฏิสัมพันธ์ระหว่างการคัดเลือกข้อสอบและลำดับข้อสอบกรณีการใช้ข้อมูลทุติยภูมิ

<u>สรุปผลการวิเคราะห์ข้อมูลจากการใช้ข้อมูลทุติยภูมิ</u>

ผลการศึกษาคุณภาพการทดสอบจากค่าความเที่ยง พบว่าการคัดเลือกข้อสอบขั้นแรก 5 ข้อ วิธี Fisher information (r^2= 0.062) และ Kullback-Leibler information (r^2= 0.063) ให้ผลไม่แตกต่างกัน สำหรับลำดับข้อสอบพบว่า ลำดับข้อสอบแบบผสมผสานมิติ (r^2= 0.089) มีคุณภาพสูงกว่าลำดับข้อสอบแบบจำแนกมิติ (r^2= 0.036)

ผลการศึกษาคุณภาพการทดสอบจากระยะท างแบบยุคลิด พบว่า มีปฏิสัมพันธ์ระหว่างการคัดเลือกข้อสอบและลำดับข้อสอบ โดยที่ วิธี Kullback-Leibler information แบบผสมผสานมิติ (ED=4.765) มีคุณภาพสูงสุด

ตอนที่ 3 การวิเคราะห์ข้อมูลกรณีการใช้ข้อมูลปฐมภูมิ

ผลการวิเคราะห์ข้อมูลกรณีการใช้ข้อมูลปฐมภู มิ จากการทดสอบความสามารถด้านวิทยาศาสตร์ผ่านโปรแกรมการทดสอบปรับเหมาะแบบพหุมิติ มีรายละเอียดดังนี้

3.1 ค่าสถิติพื้นฐานของผู้สอบวัดความสามารถด้านวิทยาศาสตร์

กลุ่มตัวอย่างนักเรียนที่ทดลองใช้โปรแกรมในครั้งนี้คือ นักเรียนโรงเรียน อุตรดิตถ์ดรุณี จำนวน 66 คน ซึ่งเป็นเพศหญิงทั้งหมด และนักเรียนโรงเรียนอื่น ๆ จำนวน 8 คน โดยนักเรียนส่วนใหญ่ใช้เวลาในการสอบอยู่ระหว่าง 11 -20 นาที จากเวลาที่กำหนดไว้ 1 ชั่วโมง (ข้อสอบ 30 ข้อ) นอกจากนี้ นักเรียนกลุ่มตัวอย่างมีค่าเฉลี่ยเกรดเฉลี่ยสะสมอยู่ที่3.21 และมีทักษะด้านคอมพิวเตอร์เท่ากับ 19.51 คะแนน จากคะแนนเต็ม35 คะแนน รายละเอียดดังตารางที่4.15

ตารางที่ 4.15 ค่าสถิติพื้นฐานของผู้สอบความสามารถด้านวิทยาศาสตร์

ตัวแปร		ความถี่	ร้อยละ
1. กลุ่มตัวอย่าง	โรงเรียนอุตรดิตถ์ดรุณี	66	89.19
	โรงเรียนอื่นๆ	8	10.81
	รวม	74	100.00
2. เพศ	หญิง	69	93.24
	ชาย	5	6.76
	รวม	74	100.00
3. เวลาที่ใช้ในการสอบ	5-10 นาที	18	24.32
	11-20 นาที	25	33.78
	21-30 นาที	24	32.43
	31 นาทีขึ้นไป	9	12.17
	รวม	74	100.00

ภูมิหลังของนักเรียน	N	Min	Max	*M*	*SD*	Sk	Ku
4. เกรดเฉลี่ยสะสม	73	2.01	4	3.12	0.48	0.22	-0.41
5. ทักษะคอมพิวเตอร์	74	7	30	19.51	4.88	-0.14	-0.21

นอกจากนี้พบว่า ความสามารถของผู้สอบวัดความสามารถด้านวิทยาศาสตร์ในภาพรวมอยู่ในระดับต่ำ (-1.15) ส่วนความสามารถในแต่ละวิชาชีววิทยา เคมี และฟิสิกส์ อยู่ในระดับปานกลาง (0.15, -0.93 และ 0.74 ตามลำดับ) และมีค่าความคลาดเคลื่อนมาตรฐาน โดยเฉลี่ยเท่ากับ 0.05 และมีสารสนเทศของแบบสอบ เท่ากับ 424.74 นอกจากนี้พบว่า ร้อยละการตอบถูก และความสามารถที่ประมาณค่าได้มีความสอดคล้องกัน รายละเอียดดังตารางที่ 4.16

ตารางที่ 4.16 ค่าสถิติพื้นฐานความสามารถของผู้สอบ ความคลาดเคลื่อนมาตรฐานและสารสนเทศแบบสอบ

วิชา	N	Min	Max	*M*	*SD*	Sk	ku
1. วิชาชีววิทยา	74	-0.98	1.53	0.15	0.48	0.29	0.43
		(10.00)	(72.73)	(41.85)	(14.93)	(-0.01)	(-0.62)
2. วิชาเคมี	74	-1.73	0.15	-0.93	0.44	0.47	-0.24
		(0.00)	(60.00)	(24.23)	(14.72)	(-0.18)	(-0.74)
3. วิชาฟิสิกส์	74	-1.60	1.05	-0.74	0.47	0.85	1.56
		(0.00)	(80.00)	(26.88)	(13.53)	(0.81)	(2.41)
4. รวม	74	-2.09	0.02	-1.15	0.51	0.16	-0.88
		(13.33)	(46.67)	(32.07)	(7.86)	(-0.24)	(-0.49)
ความคลาดเคลื่อนมาตรฐาน	74	0.04	0.09	0.05	0.01	0.91	0.91
สารสนเทศแบบสอบ	74	126.68	779.53	424.74	162.23	0.38	-0.70

* หมายเหตุ ตัวเลขในวงเล็บคือ ร้อยละการตอบถูก

3.2 ผลการวิเคราะห์คุณภาพของการทดสอบปรับเหมาะแบบพหุมิติกรณีการใช้ข้อมูลปฐมภูมิ

คุณภาพของการทดสอบปรับเหมาะแบบพหุมิติกรณีการใช้ข้อมูลปฐมภูมิพิจารณาจากความยาวข้อสอบ ณ จุดตัด SE $\leq$ 0.3 และสารสนเทศแบบสอบ ซึ่งผู้วิจัยพิจารณาสารสนเทศจากสองช่วงของการทดสอบ นั่นคือ สารสนเทศแบบสอบ ณ จุดตัดSE $\leq$ 0.3 และสารสนเทศแบบสอบเมื่อทำข้อสอบครบ 30 ข้อ พร้อมทั้งเสนอค่าความคลาดเคลื่อนมาตรฐานในการประมาณค่า (SE) ซึ่งเป็นค่าที่มีความสัมพันธ์ในทางลบกับค่าสารสนเทศแบบสอบ เมื่อพิจารณาความยาวข้อสอบพบว่า ณ จุดตัด SE $\leq$ 0.3 ผู้สอบส่วนใหญ่ได้รับข้อสอบจำนวน 5 ข้อ โดยมีสารสนเทศแบบสอบเท่ากับ 12.799 และเมื่อพิจารณาสารสนเทศแบบสอบจากการสอบ 30 ข้อ พบว่า มีค่าสูงขึ้นเท่ากับ 424.74 ส่วนความคลาดเคลื่อนมาตรฐานใน การประมาณค่าจะมีค่าต่ำลงเมื่อทำข้อสอบมากขึ้น รายละเอียดดังตารางที่ 4.17

ตารางที่ 4.17 ค่าสถิติพื้นฐานของคุณภาพของการทดสอบจำแนกตามลำดับข้อสอบ

คุณภาพของการทดสอบ	ลำดับข้อสอบ	N	*M*	*SD*	Min	Max
1. ความยาวข้อสอบ	จำแนกมิติ	37	5.108	0.315	5.000	6.000
(ณ จุดตัด SE ≤ 0.3)	ผสมผสานมิติ	37	5.000	0.000	5.000	5.000
	รวม	**74**	**5.054**	**0.228**	**5.000**	**6.000**
2. สารสนเทศแบบสอบ	จำแนกมิติ	37	12.193	0.753	11.175	13.982
(ณ จุดตัด SE ≤ 0.3)	ผสมผสานมิติ	37	13.405	0.970	12.184	14.699
	รวม	**74**	**12.799**	**1.056**	**11.175**	**14.699**
3. ความคลาดเคลื่อนมาตรฐาน (ณ จุดตัด SE ≤ 0.3)	จำแนกมิติ	37	0.287	0.009	0.267	0.299
	ผสมผสานมิติ	37	0.274	0.010	0.261	0.286
	รวม	**74**	**0.280**	**0.011**	**0.261**	**0.299**
4. สารสนเทศแบบสอบ (ข้อสอบ 30 ข้อ)	จำแนกมิติ	37	424.607	156.557	126.678	724.696
	ผสมผสานมิติ	37	424.872	169.873	179.706	779.533
	รวม	**74**	**424.740**	**162.228**	**126.678**	**779.533**
5. ความคลาดเคลื่อนมาตรฐาน (ข้อสอบ 30 ข้อ)	จำแนกมิติ	37	0.052	0.012	0.037	0.089
	ผสมผสานมิติ	37	0.051	0.010	0.036	0.075
	รวม	**74**	**0.052**	**0.011**	**0.036**	**0.089**

3.3 ผลการเปรียบเทียบ คุณภาพของการทดสอบปรับเหมาะแบบพหุมิติ กรณีใช้ข้อมูลปฐมภูมิ จำแนกตามลำดับข้อสอบ

การเปรียบเทียบคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติ จำแนกตามลำดับข้อสอบ ผู้วิจัยจะใช้การวิเคราะห์ความแปรปรวน ตัวแปรพหุนามแบบจำแนกทางเดียว (one-way MANOVA) โดยตัวแปรตามที่นำมาศึกษาคือ ความยาว ข้อสอบ สารสนเทศแบบสอบ ณ จุดตัด SE ≤ 0.3 และสารสนเทศแบบสอบทั้งฉบับ 30 ข้อ โดยไม่นำความคลาดเคลื่อนมาตรฐานในการประมาณค่ามาพิจารณา เนื่องจากสารสนเทศแบบสอบและความคลาดเคลื่อนมาตรฐานในการประมาณค่ามีความสัมพันธ์กันสูง

ผลการตรวจสอบข้อตกลงเบื้องต้นเกี่ยวกับความเท่ ากันของเมตริกซ์ความแปรปรวน-ความแปรปรวนร่วมของประชากรในการวิจัยครั้งนี้ไม่สามารถวิเคราะห์ Box's test ได้ เนื่องจากตัวแปรความยาวข้อสอบของ ลำดับข้อสอบ ผสมผสานมิติมีเพียงแค่ค่าเดียว คือ นักเรียนทุกคนได้รับข้อสอบ 5 ข้อ เท่ากันทุกคน อย่างไรก็ตาม ทรงศักดิ์ ภูสีอ่อน (2552) กล่าวว่า ถ้าการวิเคราะห์นั้นๆข้อมูลมีปัญหา เช่น กลุ่มตัวอย่างมีขนาดเล็ก หรือฝ่าฝืนข้อตกลงเบื้องต้น สถิติ Pillai'sTrace

จะมีความแกร่งมากกว่า ดังนั้นการวิเคราะห์ multivariate test เพื่อพิจารณาความแตกต่างระหว่างกลุ่ม ผู้วิจัยจึงพิจารณาที่ค่าของ Pillai'sTrace ซึ่งพบว่า ค่า p-value มีนัยสำคัญทางสถิติ นั่นคือ ค่าเฉลี่ยของตัวแปรตาม (พิจารณาร่วมกันทั้ง 3 ตัวแปร) มีความแตกต่างกันตาม รูปแบบของลำดับข้อสอบ นอกจากนั้นจากการวิเคราห์ Bartlett's test of sphericity เพื่อตรวจสอบความสัมพันธ์ระหว่าง ตัวแปรตาม พบว่า ความยาวข้อสอบ สารสนเทศแบบสอบ ณ จุดตัด SE $\leq$ 0.3 และสารสนเทศแบบสอบทั้งฉบับ 30 ข้อ มีความสัมพันธ์กัน ดังนั้นจึงสามารถวิเคราะห์ความแปรปรวนตัวแปรพหุนามได้ รายละเอียดดังตารางที่ 4.18

ตารางที่ 4.18 การวิเคราะห์ความแปรปรวนตัวแปรพหุนามของคุณภาพการทดสอบปรับเหมาะแบบพหุมิติทั้ง 3 ตัวแปร ที่มีลำดับข้อสอบต่างกัน

Effect		Value	F	Hypothesis df	Error df	p
Intercept	Pillai's Trace	0.998627	16969.67	3	70	0.00
	Wilks' Lambda	0.001373	16969.67	3	70	0.00
	Hotelling's Trace	727.2717	16969.67	3	70	0.00
	Roy's Largest Root	727.2717	16969.67	3	70	0.00
ลำดับ	Pillai's Trace	0.403879	15.80864	3	70	0.00
ข้อสอบ	Wilks' Lambda	0.596121	15.80864	3	70	0.00
	Hotelling's Trace	0.677513	15.80864	3	70	0.00
	Roy's Largest Root	0.677513	15.80864	3	70	0.00

ผลการทดสอบอิทธิผลระหว่างตัวแปร พบว่า ลำดับข้อสอบ มีอิทธิพลต่อความยากข้อสอบ (F=4.36, p=0.04) และสารสนเทศแบบสอบ ณ จุดตัด SE $\leq$ 0.3 (F=36.08, p=0.00) อย่างมีนัยสำคัญทางสถิติ ที่ระดับ .05 แต่ไม่มีอิทธิพลต่อสารสนเทศแบบสอบเมื่อสอบครบ 30 ข้อ และเมื่อ เปรียบเทียบค่าเฉลี่ย พบว่า ลำดับข้อสอบ แบบผสมผสานมิติ มีค่าเฉลี่ยความยากข้อสอบ (M=5.00, SD=0.00) น้อยกว่าลำดับข้อสอบ แบบจำแนกมิติ (M =5.11, SD=0.315) และลำดับข้อสอบแบบผสมผสานมิติมีค่าเฉลี่ยสารสนเทศแบบสอบ ณ จุดตัด SE $\leq$ 0.3 (M =13.41, SD=0.97) สูงกว่าลำดับข้อสอบแบบจำแนกมิติ (M =12.19, SD=0.75) รายละเอียดดังตารางที่ 4.19

ตารางที่ 4.19 การทดสอบอิทธิผลระหว่างตัวแปร (test of between-subjects effects)

แหล่งความแปรปรวน	ตัวแปรตาม	Type III Sum of Squares	df	Mean Square	F	p
Corrected Model	Test length	0.22	1	0.22	4.36	0.04
	Test Info (SE ≤ 0.3)	27.18	1	27.18	36.08	0.00
	Test Info (30 items)	1.30	1	1.30	0.00	0.99
Intercept	Test length	1890.22	1	1890.22	38148.00	0.00
	Test Info (SE ≤ 0.3)	12121.80	1	12121.80	16088.07	0.00
	Test Info (30 items)	13349872	1	13349872	500.30	0.00
ลำดับข้อสอบ	Test length	0.22	1	0.22	4.36*	0.04
	Test Info (SE ≤ 0.3)	27.18	1	27.18	36.08*	0.00
	Test Info (30 items)	1.30	1	1.30	0.00	0.99
Error	Test length	3.57	72	0.05		
	Test Info (SE ≤ 0.3)	54.25	72	0.75		
	Test Info (30 items)	1921216.70	72	26683.57		
Total	Test length	1894.00	74			
	Test Info (SE ≤ 0.3)	12203.23	74			
	Test Info (30 items)	15271090.3	74			
Corrected Total	Test length	3.78	73			
	Test Info (SE ≤ 0.3)	81.43	73			
	Test Info (30 items)	1921217.99	73			

*p < .05

3.4 ค่าสถิติพื้นฐาน ภูมิหลังของนักเรียนและผล ประเมินการใช้โปรแกรมการทดสอบปรับเหมาะแบบพหุมิติ

ค่าสถิติพื้นฐานของภูมิหลังของนักเรียนชั้น ม .5 จำนวน 74 คน มีค่าเกรดเฉลี่ยสะสม (GPAX) อยู่ระหว่าง 2.01 ถึง 4.00 โดยมีค่าเฉลี่ยของเกรดเฉลี่ยสะสม เท่ากับ 3.12 สำหรับทักษะคอมพิวเตอร์มีคะแนนเต็ม 35 คะแนน นักเรียนมีทักษะคอมพิวเตอร์อยู่ระหว่าง 7 – 30 คะแนน และมีค่าเฉลี่ยทักษะคอมพิวเตอร์ เท่ากับ 19.51 รายละเอียดดังตารางที่ 4.20

ตารางที่ 4.20 ค่าสถิติพื้นฐานของภูมิหลังของนักเรียน

ภูมิหลังของนักเรียน	N	Min	Max	*M*	*SD*	Sk	Ku
เกรดเฉลี่ยสะสม	73	2.01	4	3.12	0.48	0.22	-0.41
ทักษะคอมพิวเตอร์	74	7	30	19.51	4.88	-0.14	-0.21

ผลประเมินการใช้โปรแกรมการทดสอบปรับเหมาะแบบพหุมิติของนักเรียนชั้น ม. 5 ที่เข้าระบบเพื่อทำการทดสอบผ่านคอมพิวเตอร์ พบว่า ในภาพรวมมีความพึงพอใจมาก (M = 3.52, SD = 0.49) เมื่อพิจารณาแต่ละด้านพบว่า ด้านความยากในการใช้มีความพึงพอใจปานกลาง (M = 3.38, SD = 0.45) นักเรียนบางส่วนยังเห็นว่า โปรแกรมมีความยุ่งยากซับซ้อน และไม่คุ้นเคยกับการสอบผ่านคอมพิวเตอร์ ส่วนด้านความเป็นไปได้ในการนำไปใช้ พบว่า มีความพึงพอใจปานกลาง (M = 3.39, SD = 0.76) โดยประเด็นที่นักเรียนเห็นด้วยค่าเฉลี่ยต่ำสุดคือ การนำวิธีการทดสอบผ่านคอมพิวเตอร์มาใช้สำหรับการสอบเข้ามหาวิทยาลัย ส่วนด้านความพึงพอใจ ในการใช้งานผู้สอบมีความพึงพอใจมาก (M =3.83, SD = 0.67) รายละเอียดดังตารางที่ 4.21 นอกจากนั้นนักเรียนกลุ่มตัวอย่างยังให้ข้อเสนอแนะเพิ่มเติมว่า ควรจะออกแบบโปรแกรมให้สามารถย้อนกลับไปแก้ไขคำตอบได้ และควรคำนึงถึงการป้องกันการทุจริตการสอบ และเห็นว่ามีความเหมาะสมที่มีนาฬิกาจับเวลาในการทดสอบ

เกณฑ์การประเมินโปรแกรมการทดสอบปรับเหมาะแบบพหุมิติ

4.51 – 5.00	หมายถึง	เห็นด้วยหรือมีความพึงพอใจระดับมากที่สุด
3.51 – 4.50	หมายถึง	เห็นด้วยหรือมีความพึงพอใจระดับมาก
2.51 – 3.50	หมายถึง	เห็นด้วยหรือมีความพึงพอใจระดับปานกลาง
1.51 – 2.50	หมายถึง	เห็นด้วยหรือมีความพึงพอใจระดับน้อย
1.00 – 1.50	หมายถึง	เห็นด้วยหรือมีความพึงพอใจระดับน้อยที่สุด

ตารางที่ 4.21 ผลประเมินการใช้โปรแกรมการทดสอบปรับเหมาะแบบพหุมิติ

ประเด็นพิจารณา	N	Min	Max	*M*	SD	แปลผล
1. ด้านความยากในการใช้	**74**	**2**	**4.6**	**3.38**	**0.45**	**ปานกลาง**
1.1 โปรแกรมมีความเหมาะสมและให้ความสะดวกในการทดสอบ	74	2	5	3.93	0.91	มาก
1.2 การใช้โปรแกรมมีความยุ่งยาก ซับซ้อน	74	1	5	2.82	1.10	ปานกลาง
1.3 ผู้สอบไม่คุ้นเคยกับการสอบผ่านคอมพิวเตอร์	74	1	5	2.93	1.16	ปานกลาง
1.4 เมนูคำสั่งต่างๆ ทำให้ใช้โปรแกรมได้อย่างถูกต้อง	74	1	5	3.61	1.10	มาก
1.5 ลำดับข้อสอบมีความเหมาะสม	74	1	5	3.59	1.03	มาก
2. ความเป็นไปได้ในการนำไปใช้	**74**	**1.75**	**4.5**	**3.39**	**0.76**	**ปานกลาง**
2.1 การทดสอบผ่านคอมพิวเตอร์มีความน่าเชื่อถือและสามารถนำไปใช้ในการทดสอบได้จริง	74	2	5	3.66	0.82	มาก
2.2 ควรนำวิธีการทดสอบผ่านคอมพิวเตอร์มาใช้ในชั้นเรียน	74	2	5	3.41	0.81	ปานกลาง

ตารางที่ 4.21 (ต่อ)

ประเด็นพิจารณา	N	Min	Max	*M*	SD	แปลผล
2.3 ควรนำวิธีการทดสอบผ่านคอมพิวเตอร์มาใช้สำหรับการสอบเข้ามหาวิทยาลัย	74	1	5	2.82	1.04	ปานกลาง
2.4 มีความเป็นไปได้ที่จะนำวิธีการทดสอบผ่านคอมพิวเตอร์มาใช้ในอนาคต	74	2	5	3.66	1.04	มาก
3. ความพึงพอใจที่มีต่อการใช้	**74**	**2.25**	**5**	**3.83**	**0.67**	**มาก**
3.1 ผู้ใช้สามารถสั่งแสดงผลการทดสอบทั้งทางจอภาพ ทางเครื่องพิมพ์และอีเมล์	74	2	5	3.69	0.88	มาก
3.2 โปรแกรมการทดสอบมีความทันสมัย	74	1	5	3.86	1.14	มาก
3.3 การทดสอบโดยใช้คอมพิวเตอร์ให้คะแนนที่เชื่อถือได้	74	2	5	3.84	0.91	มาก
3.4 ระยะเวลาในการทดสอบมีความเหมาะสม	74	3	5	3.93	0.83	มาก
ภาพรวม	**74**	**2.31**	**4.31**	**3.52**	**0.49**	**มาก**

3.5 ผลการศึกษาภูมิหลังของนักเรียนที่ส่งผลต่อผลการประเมิน โปรแกรมการทดสอบปรับเหมาะแบบพหุมิติ

ผลการศึกษาภูมิหลังของนักเรียนที่ส่งผลต่อผลการประเมินโปรแกรมการทดสอบปรับเหมาะแบบพหุมิติ ผู้วิจัยใช้การวิเคราะห์ถดถอยเชิงซ้อน ซึ่งจากการศึกษาความสัมพันธ์ระหว่างตัวแปรเกรดเฉลี่ยสะสม ทักษะคอมพิวเตอร์ และผลการประเมินโปรแกรม พบว่า ตัวแปรอิสระทั้ง 2 ไม่มีความสัมพันธ์ต่อตัวแปรตาม และพบว่าตัวแปรเกรดเฉลี่ยสะสมและทักษะคอมพิวเตอร์มีความสัมพันธ์กันระดับระดับสูง(r=0.85) อย่างมีนัยสำคัญที่ระดับ.01 รายละเอียดดังตารางที่ 4.22

ตารางที่ 4.22 ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างผลประเมินโปรแกรมเกรดเฉลี่ยสะสมและทักษะคอมพิวเตอร์

ตัวแปร	*M*	*SD*	ผลประเมินโปรแกรม	เกรดเฉลี่ยสะสม	ทักษะคอมพิวเตอร์
ผลประเมินโปรแกรม	3.52	0.49	1.00		
เกรดเฉลี่ยสะสม	3.12	0.48	-0.10	1.00	
ทักษะคอมพิวเตอร์	19.51	4.88	0.05	0.85**	1.00

**p < .01

ผลการวิเคราะห์การถดถอย พบว่า ตัวแปรอิสระทั้ง 2 ประกอบด้วย เกรดเฉลี่ยสะสม และทักษะคอมพิวเตอร์ มีความสัมพันธ์กับผลการประเมินโปรแกรมการทดสอบปรับเหมาะแบบพหุมิติในระดับต่ำ คือ 0.273 ตัวแปรทั้งสองสามารถอธิ บายความผันแปรที่เกิดขึ้นในผลประเมินเพียง 7.4% และจากผลการวิเคราะห์ความแปรปรวน พบว่า ไม่มีตัวแปรใดที่ส่งต่อผลการประเมินโปรแกรมการทดสอบปรับเหมาะแบบพหุมิติ อย่างมีนัยสำคัญ (F=2.81, p = 0.07) ดังตารางที่ 4.23 ดังนั้นผู้วิจัยจึงไม่พิจารณาค่าสัมประสิทธิ์การถดถอยในขั้นตอนต่อไป

ตารางที่ 4.23 ผลการวิเคราะห์ความแปรปรวนที่มีผลต่อผลประเมินโปรแกรม

แหล่งความแปรปรวน	Sum of Squares	df	Mean Square	F	p
Regression	1.25	2	0.63	2.81	0.07
Residual	15.63	70	0.22		
รวม	16.88	72			
	R = .273 R^2 = .074 adjusted R^2 =.048				

<u>สรุปผลการวิเคราะห์ข้อมูลจากการใช้ข้อมูลปฐมภูมิ</u>

การใช้ข้อมูลปฐมภูมิจะดำเนินการสอบจริงโดยใช้วิธีคัดเลือกข้อสอบเพียงวิธีเดียว คือ Fisher information ดังนั้นจะเปรียบเทียบลำดับข้อสอบเท่านั้น ผลการศึกษาคุณภาพการทดสอบพิจารณาจากความยาวข้อสอบ สารสนเทศแบบสอบ ณ ตำแ หน่ง SE $\leq$ 0.3 และสารสนเทศแบบสอบเมื่อสอบครบ 30 ข้อ พบว่า ลำดับข้อสอบ ส่งผลให้เกิดความแตกต่างของความยาวข้อสอบและสารสนเทศแบบสอบ ณ ตำแหน่ง SE $\leq$ 0.3 โดยพบว่า ลำดับข้อสอบ แบบผสมผสานมิติให้คุณภาพการทดสอบสูงกว่า ลำดับข้อสอบ แบบจำแนกมิติ สำหรับการศึกษาภูมิหลังของผู้สอ บ พบว่า เกรดเฉลี่ยสะสมและทักษะคอมพิวเตอร์ไม่ส่งผลต่อผลประเมินโปรแกรมการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์

ตอนที่ 4 สรุปผลการวิเคราะห์ข้อมูลจาก 3 แหล่งข้อมูล

ผลการวิเคราะห์ข้อมูลจาก 3 แหล่งข้อมูล ได้ถูกจำแนกย่อย พิจารณาในหลายประเด็น ทำให้ไม่เห็นภาพรวมของการวิเคราะห์ข้อมูล และเพื่อให้สามารถเข้าใจผลการวิเคราะห์ข้อมูลง่ายขึ้น ผู้วิจัยจึงได้สรุปผลการวิเคราะห์ข้อมูลจากทั้ง 3 แหล่ง โดยเริ่มพิจารณาจากการจำลองข้อมูลและข้อมูลทุติภูมิ เนื่องจาก แหล่งข้อมูลทั้ง 2 พิจารณาคุณภาพการทดสอบด้วยค่าความเที่ยง แล ะระยะทางแบบยุคลิดเช่นเดียวกัน ซึ่งสามารถสรุปได้ว่า จากการจำลองข้อมูล พบว่า การคัดเลือกข้อสอบขั้นแรกวิธี Fisher information (r^2 = 0.340) และ Kullback-Leibler information (r^2 = 0.335)

มีค่าความเที่ยง ไม่แตกต่างกัน เช่นเดียวกับข้อมูลทุติยภูมิที่พบว่า การคัดเลือกข้อสอบขั้นแรกวิธี Fisher information (r^2 = 0.062) และ Kullback-Leibler information (r^2 = 0.063) มีค่าความเที่ยงไม่แตกต่างกัน สำหรับรูปแบบของลำดับข้อสอบ กรณีการจำลองข้อมูล พบว่า ลำดับข้อสอบ แบบผสมผสานมิติ (r^2 = 0.377) มีความเที่ยงสูงกว่า ลำดับข้อสอบ แบบจำแนกมิติ (r^2 = 0.298) ซึ่งสอดคล้องกับข้อมูลทุติยภูมิที่ พบว่า ลำดับข้อสอบแบบผสมผสานมิติ (r^2 = 0.089) มีความเที่ยงสูงกว่าลำดับข้อสอบแบบจำแนกมิติ (r^2 = 0.036) เมื่อพิจารณาปฏิสัมพันธ์ระหว่างการคัดเลือกข้อสอบและลำดับข้อสอบจากค่าระยะทางแบบยุคลิดจากการจำลองข้อมูล พบว่า มีปฏิสัมพันธ์ระหว่างการคัดเลือกข้อสอบและลำดับข้อสอบที่ร่วมกันส่งผลต่อค่าระยะทางแบบยุคลิดอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 (F=9.044, p=.003) โดยพบว่า การคัดเลือกข้อสอบด้วยวิธี Kullback-Leibler information แบบผสมผสานมิติ (ED=1.007, *S.D.*=0.548) มีค่าระยะทางแบบยุคลิดต่ำที่สุด และสอดคล้องกับข้อมูลทุติยภูมิ ซึ่งพบว่า มีปฏิสัมพันธ์ระหว่างการคัดเลือกข้อสอบและลำดับข้อสอบที่ร่วมกันส่งผลต่อค่าระยะทางแบบยุคลิด อย่างมีนัยสำคัญ ที่ระดับ .01 (F=226.736, p=.003) โดยที่ วิธี Kullback-Leibler information แบบผสมผสานมิติ (ED=4.765, *S.D.*=3.414) มีค่าระยะทางแบบยุคลิดต่ำที่สุด

เมื่อพิจารณาจากข้อมูลปฐมภูมิที่ใช้การคัดเลือกข้อสอบเพียงวิธีเดียวคือ วิธี Fisher information จะพิจารณาคุณภาพการทดสอบจากความยาวข้อสอบและสารสนเทศแบบสอบ ผลการวิเคราะห์ข้อมูลพบว่า ลำดับข้อสอบ แบบผสมผสานมิติ มีค่าเฉลี่ยความยาวข้อสอบต่ำกว่าลำดับข้อสอบ แบบจำแนกมิติ และลำดับข้อสอบแบบผสมผสานมิติมีค่าเฉลี่ยสารสนเทศแบบสอบณ ตำแหน่ง SE $\leq$ 0.3 สูงกว่าลำดับข้อสอบแบบจำแนกมิติ แต่ไม่พบความแตกต่างของสารสนเทศแบบสอบ เมื่อทำข้อสอบครบทั้ง 30 ข้อ นอกจากนั้น พบว่า ภูมิหลังนักเรียนเกี่ยวกับค่าเฉลี่ยสะสมและทักษะการใช้คอมพิวเตอร์ไม่ส่งผลต่อผลประเมินการใช้โปรแกรมการทดสอบปรับเหมาะแบบพหุมิติ ซึ่งรายละเอียดสามารถสรุปได้ดังตารางที่ 4.24

ตารางที่ 4.24 สรุปผลการเปรียบเทียบคุณภาพการทดสอบปรับเหมาะแบบพหุมิติจากข้อมูล 3 แหล่ง

ประเด็นที่ศึกษา	การจำลองข้อมูล		ข้อมูลทุติยภูมิ		ข้อมูลปฐมภูมิ			
	r^2	ED	r^2	ED	Test length	$I(\theta)$ SE≤0.3	$I(\theta)$ 30 items	*M*
1. วิธีการคัดเลือกข้อสอบขั้นแรก					ใช้วิธี Fisher information เพียงวิธีเดียว			
1.1 Fisher information	○		○					
1.2 Kullback-Leibler information	○	✓	○	✓				
2. ลำดับข้อสอบ								
2.1 ลำดับแบบจำแนกมิติ							○	
2.2 ลำดับแบบผสมผสานมิติ	✓	✓	✓	✓	✓	✓	○	
3. ปฏิสัมพันธ์ระหว่างวิธีการคัดเลือกข้อสอบขั้นแรกและลำดับข้อสอบ	N/A	✓	N/A	✓				
4. ภูมิหลังของนักเรียน								
4.1 เกรดเฉลี่ยสะสม								○
4.2 ทักษะการใช้คอมพิวเตอร์								○

***หมายเหตุ** r^2 คือ ความเที่ยง (reliability)

ED คือ ระยะทางแบบยุคลิด (Euclidean distance)

Test length คือ ความยาวข้อสอบ

$I(\theta)$ คือ สารสนเทศแบบสอบ

SE คือ ความคลาดเคลื่อนมาตรฐานในการประมาณค่า

M คือ ค่าเฉลี่ยผลประเมินโปรแกรมการทดสอบปรับเหมาะแบบพหุมิติ

✓ คือ การทดสอบที่มีคุณภาพการทดสอบสูงกว่าหรือมีปฏิสัมพันธ์อย่างมีนัยสำคัญ

○ คือ ผลการทดสอบไม่มีความแตกต่างกัน

N/A คือ ไม่มีข้อมูล

บทที่ 5

สรุปผลการวิจัย อภิปรายผลและข้อเสนอแนะ

การวิจัยครั้งนี้ มีวัตถุประสงค์ เพื่อเปรียบเทียบความแตกต่างของการคัดเลือกข้อสอบขั้นแรกและลำดับข้อสอบที่ส่งผลต่อคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติ รวมถึงศึกษาปฏิสัมพันธ์ระหว่างตัวแปรทั้งสองที่มีต่อ คุณภาพของการทดสอบ นอกจากนั้น การวิจัยครั้งนี้ได้ศึกษาผลการประเมิน โปรแกรม การทดสอบความสามารถ ด้านวิทยาศาสตร์ด้วยวิธีการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์ โดยที่การทดสอบปรับเหมาะแบบพหุมิติ(multidimensional computerized adaptive testing: MCAT) เป็นการผสมผสานแนวคิดของ การทดสอบแบบปรับเหมาะ และทฤษฎีการตอบสนองข้อสอบแบบพหุมิติ (multidimensional item response theory: MIRT) ซึ่งเป็นการทดสอบแนวใหม่ที่มีความแม่นยำในการประมาณค่าความสามารถ โดยคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติจะ พิจารณาจากค่าความเที่ยง ค่าระยะทางแบบยุคลิด สารสนเทศแบบสอบ และ ความยาวข้อสอบ ซึ่งการ พิจารณาค่าต่าง ๆ เหล่านี้จะแตกต่างไปตามแหล่งข้อมูลที่นำมาศึกษา โดยผู้วิจัยได้ ศึกษาจากแหล่งข้อมูล 3 แหล่ง ได้แก่ การจำลองข้อมูล ข้อมูลทุติยภูมิ และข้อมูลปฐมภูมิ ซึ่งสามารถสรุปผลการวิจัยแยกตามแหล่งข้อมูลได้ดังนี้

สรุปผลการวิจัย

1. สรุปผลการวิจัยกรณีการจำลองข้อมูล

สรุปผลการวิจัยกรณีการจำลองข้อมูลใช้การจำลองข้อมูลการตอบสนองข้อสอบจากผู้สอบ 10,000 คน และคลังข้อสอบจำนวน 300 ข้อ โดยแบ่งการทดลองออกเป็น 4 โมเดล จำแนกตามวิธีการคัดเลือกข้อสอบ ขั้นแรก และลำดับข้อสอบ รูปแบบข้อสอบที่นำมาใช้ในการวิจัยครั้งนี้เป็นข้อสอบที่วัดความสามารถ 2 มิติ ที่มีความสัมพันธ์กันระดับปานกลาง ซึ่งมีลักษณะพหุมิติระหว่างข้อสอบ (between-item multidimensionality) นั่นคือ ข้อสอบหนึ่งข้อวัดความสามารถ 1 มิติ ทั้งนี้ข้อสอบแต่ละข้อต้องมีความสัมพันธ์กั น โดยกำหนดพารามิเตอร์อำนาจจำแนก ให้มีลักษณะเป็นโค้งปกติ ตัดปลาย โดยมีค่าเฉลี่ยเป็น 0.5 และค่าส่วนเบี่ยงเบนมาตรฐาน เป็น 0.1 และตัดค่าที่ติดลบออก เนื่องจากหากอำนาจจำแนกมีค่าติดลบจะทำให้การประมาณค่าความสามารถไม่มีความแม่นยำ พารามิเตอร์ความยากอยู่ในช่วง -3 ถึง 3 มีการจำแนกแบบเอกรูป และพารามิเตอร์โอกาสการเดามีการจำแนกแบบเอกรูปโดยมีค่าสูงสุด ไม่เกิน 0.1การศึกษาคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติจะพิจารณาจากค่าความเที่ยงและค่าระยะทางแบบยุคลิด

สรุปผลการเปรียบเทียบ คุณภาพการทดสอบปรับเหมาะแบบพหุมิติ เมื่อพิจารณาคุณภาพจากค่าความเที่ยง พบว่า การคัดเลือกข้อสอบขั้นแรก 5 ข้อ วิธี Fisher information (r^2 = 0.340) และ วิธี Kullback-Leibler information (r^2 = 0.335) ให้คุณภาพการทดสอบไม่แตกต่างกัน อย่างไรก็ตาม เมื่อพิจารณา ค่าความเที่ยงแต่ละ ช่วงความยาวข้อสอบจะ พบว่า ความเที่ยงมีแนวโน้มสูงขึ้นเมื่อ ทำข้อสอบมากขึ้น ซึ่งสอดคล้องกัน ทุกเงื่อนไขการทดสอบ สำหรับลำดับข้อสอบพบว่า ลำดับข้อสอบแบบผสมผสานมิติ (r^2 = 0.377) มีคุณภาพสูงกว่าลำดับข้อสอบแบบจำแนกมิติ (r^2 = 0.298) และเมื่อพิจารณา คุณภาพการทดสอบจากค่าระยะทางแบบยุคลิด ซึ่งทดสอบด้วยความแปรปรวนสองทาง พบว่า มีปฏิสัมพันธ์ระหว่างการคัดเลือกข้อสอบและ ลำดับข้อสอบที่ร่วมกันส่งผลต่อคุณภาพการทดสอบ อย่างมีนัยสำคัญที่ระดับ .01 (F=9.044, p=.003) โดยที่วิธี Kullback-Leibler information แบบผสมผสานมิติ(ED=1.007, *S.D.*=0.548) มีคุณภาพสูงสุด

2. สรุปผลการวิจัยกรณีการใช้ข้อมูลทุติยภูมิ

สรุปผลการวิจัยกรณีการใช้ข้อมูลทุติยภูมิจาก ข้อมูลการตอบข้อสอบของนักเรียนจากการทดสอบความถนัดทางวิชาชีพและวิชาการ (professional and academic aptitude test: PAT) ด้านวิทยาศาสตร์ (PAT2) ของสถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน) โดยการวิจัยครั้งนี้จะศึกษาจาก3 วิชา คือ ชีววิทยา เคมี และฟิสิกส์ จำนวน100 ข้อ โดยมีค่าพารามิเตอร์อำนาจจำแนกมีค่าอยู่ระหว่าง -0.19 ถึง 0.73 สำหรับอำนาจจำแนกที่ติดลบนั้น ถือว่ายังมีค่าน้อย ดังนั้นการวิจัยครั้งนี้จึงไม่ได้ตัดข้อสอบข้อใดออกจากการวิจัย สำหรับความยาก พบว่า ข้อสอบมีความยากระหว่าง -0.75 ถึง 1.21 ซึ่งอยู่ในระดับปานกลาง และโอกาสการเดามีค่าอยู่ระหว่าง 0.02 ถึง 0.20 ซึ่งถือว่าส่วนใหญ่มีค่าต่ำ สำหรับการศึกษาคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติจะพิจารณาจากค่าความเที่ยงและค่าระยะทางแบบยุคลิดเช่นเดียวกับการจำลองข้อมูล และจะใช้กลุ่มตัวอย่างผู้สอบ PAT2 จำนวน 5,000 คน ได้จากการสุ่มอย่างง่าย

สรุปผลการเปรียบเทียบคุณภาพการทดสอบปรับเหมาะแบบพหุมิติ เมื่อพิจารณาจากค่าความเที่ยง พบว่า การคัดเลือกข้อสอบขั้นแรกด้วยวิธี Fisher information (r^2 = 0.062) และ Kullback-Leibler information (r^2 = 0.063) ให้คุณภาพการทดสอบไม่แตกต่างกัน สำหรับลำดับข้อสอบพบว่า ลำดับข้อสอบแบบผสมผสานมิติ (r^2 = 0.089) มีคุณภาพสูงกว่าลำดับข้อสอบแบบจำแนกมิติ (r^2 = 0.036) และเมื่อศึกษาคุณภาพการทดสอบด้วยค่าระยะทางแบบยุคลิด พบว่า จากการทดสอบด้วยความแปรปรวน สองทาง เกิดปฏิสัมพันธ์ระหว่างการ คัดเลือกข้อสอบและ ลำดับข้อสอบที่ร่วมกันส่งผลต่อคุณภาพการทดสอบ อย่างมีนัยสำคัญที่ระดับ .01 (F=226.736, p=.003) โดยที่ วิธี Kullback-Leibler information แบบผสมผสานมิติ(ED=4.765, *S.D.*=3.414) มีคุณภาพสูงสุด

3. สรุปผลการวิจัยกรณีการใช้ข้อมูลปฐมภูมิ

สรุปผลการวิจัยกรณีการใช้ข้อมูลปฐมภูมิ หรือการทดลองใช้โปรแกรมการทดสอบปรับเหมาะแบบพหุมิติกับนักเรียนชั้น ม5 เป็นการพัฒนาโปรแกรมการทดสอบปรับเหมาะแบบพหุมิติ โดยใช้วิธีการคัดเลือกข้อสอบด้วยวิธีของ Fisher information และประมาณค่าความสามารถด้วยวิธีของ Bayesian โดยขั้นตอนนี้ศึกษาความแตกต่างของ ลำดับข้อสอบ ที่แตกต่างกัน คือ ลำดับข้อสอบ แบบจำแนกมิติ และ ลำดับข้อสอบ แบบผสมผสานมิติ การพัฒนาข้อสอบสำหรับการทดสอบครั้งนี้ พัฒนาโดยอาจารย์ที่สอนวิทยาศาสตร์ระดับ ม .ปลาย จำนวน 9 คน และผู้เชี่ยวชาญตรวจสอบเครื่องมือ 3 ท่าน โดยข้อสอบมีข้อสอบคัดเลือกมาใช้ในการทดสอบครั้งนี้ จำนวน 170 ข้อ โดยมีค่าอำนาจจำแนกอยู่ระหว่าง 0.01-0.84 ค่าความยากอยู่ระหว่าง -1.25 ถึง 1.72 และค่าโอกาสการเดา อยู่ระหว่าง 0.02 – 0.20 สำหรับคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติสำหรับการทดสอบจริง พิจารณาจากสารสนเทศแบบสอบ และความยาวข้อสอบ ซึ่งพบว่าหากพิจารณาจากความคลาดเคลื่อนมาตรฐานที่ต่ำกว่าหรือเท่ากับ 0.3 (SE $\leq$ 0.3) ผู้สอบส่วนใหญ่จะได้รับข้อสอบจำนวน 5 ข้อ อย่างไรก็ตามการทำข้อสอบเพียง 5 ข้อนั้น ยังไม่สามารถวัดความสามารถได้ครบทุกมิติ ดังนั้นผู้วิจัยจึงกำหนดให้ผู้สอบดำเนินการสอบเท่ากันทุกคนคือ 30 ข้อ และเมื่อพิจารณาค่า สารสนเทศแบบสอบ และความคลาดเคลื่อนมาตรฐานในการประมาณค่า พบว่า กรณีลำดับข้อสอบแบบจำแนกมิติมีค่า ($I(\theta)$=12.193, SE=0.287) กรณีลำดับข้อสอบแบบผสมผสานมิติมีค่า ($I(\theta)$=13.405 , SE=0.274) เมื่อเปรียบเทียบลำดับข้อสอบที่มีอิทธิพลต่อคุณภาพการทดสอบ ซึ่งพิจารณาจากความยาวข้อสอบ สารสนเทศแบบสอบ ณ จุดตัด SE $\leq$ 0.3 และสารสนเทศแบบ สอบเมื่อสอบค รบ 30 ข้อ โดยใช้การวิเคราะห์ ความแปรปรวนหลายตัวแปร (one way-MANOVA) พบว่า ลำดับข้อสอบมีอิทธิพลต่อความยาวข้อสอบ (F=4.36, p=0.04) และสารสนเทศแบบสอบ ณ จุดตัด SE $\leq$ 0.3 (F=36.08, p=0.00) อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แต่ไม่มีอิทธิพลต่อสารสนเทศ แบบสอบเมื่อสอบครบ 30 ข้อ และเมื่อ เปรียบเทียบค่าเฉลี่ย พบว่า ลำดับข้อสอบแบบผสมผสานมิติมีค่าเฉลี่ยความยาวข้อสอบ (*M*=5.00, *S.D.*=0.00) น้อยกว่าลำดับข้อสอบแบบจำแนกมิติ (*M* =5.11, *S.D.*=0.315) และลำดับข้อสอบแบบผสมผสานมิติมีค่าเฉลี่ยสารสนเทศแบบสอบ ณ จุดตัดSE $\leq$ 0.3 (*M* =13.41, *S.D.*=0.97) สูงกว่าลำดับข้อสอบแบบจำแนกมิติ (*M* =12.19, *S.D=0.75*)

นอกจากนี้สำหรับการทดสอบจริง ผู้วิจัยได้ศึกษาผลการประเมินโปรแกรมการทดสอบปรับเหมาะแบบพหุมิติ พร้อมทั้งศึกษา ภูมิหลังของนักเรียนที่ส่งผลต่อผลการ ประเมิน โดยศึกษาจากภูมิหลัง 2 ค่า คือ เกรดเฉลี่ยสะสมและทักษะคอมพิวเตอร์ โดยที่นักเรียนชั้นม.5 จำนวน 74 คน

มีค่าเกรดเฉลี่ยสะสม (GPAX) อยู่ระหว่าง 2.01 ถึง 4.00 (M = 3.12, $S.D.$=0.48) สำหรับทักษะคอมพิวเตอร์มีคะแนนเต็ม 35 คะแนน นักเรียนมีทักษะคอมพิวเตอร์อยู่ระหว่าง 7 – 30 คะแนน (M =19.51, $S.D.$=4.88) ผลประเมินการใช้โปรแกรมการทดสอบปรับเหมาะ แบบพหุมิติ พบว่าในภาพรวมมีความพึงพอใจมาก (M = 3.52, $S.D.$ = 0.49) เมื่อพิจารณาแต่ละด้านพบว่า ด้านความยากในการใช้มีความพึงพอใจปานกลาง (M = 3.38, $S.D.$ = 0.45) นักเรียนบางส่วนยังเห็นว่าโปรแกรมมีความยุ่งยากซับซ้อนและไม่คุ้นเคยกับการสอบผ่านคอมพิวเตอร์ ส่วนด้านความเป็นไปได้ในการนำไปใช้ พบว่า มีความพึงพอใจปานกลาง (M = 3.39, $S.D.$ = 0.76) โดยประเด็นที่นักเรียนเห็นด้วยค่าเฉลี่ยต่ำสุดคือ การนำวิธีการทดสอบผ่านคอมพิวเตอร์มาใช้สำหรับการสอบเข้ามหาวิทยาลัย ส่วนด้านความพึงพอใจในการใช้งาน ผู้สอบมีความพึงพอใจมาก (M =3.83, $S.D.$ = 0.67) รายละเอียดดังตารางที่ 4.25 นอกจากนั้น นักเรียนกลุ่มตัวอย่างยังให้ข้อเสนอแนะเพิ่มเติมว่า ควรจะออกแบบโปรแกรมให้สามารถย้อนกลับไปแก้ไขคำตอบได้ และควรคำนึงถึงการป้องกันการทุจริตการสอบ และเห็นว่ามีความเหมาะสมที่มีนาฬิกาจับเวลาในการทดสอบ

สำหรับผลการศึกษา ภูมิหลังของนักเรียนที่ส่งผลต่อผลการประเมินโปรแกรมการทดสอบปรับเหมาะแบบพหุมิติ ด้วยการวิเคราะห์ถดถอยพหุคูณ (multiple regression) พบว่าตัวแปรอิสระทั้ง 2 ประกอบด้วย เกรดเฉลี่ยสะสม และ ทักษะคอมพิวเตอร์ มีความสัมพันธ์กับผลการประเมินโปรแกรมการทดสอบปรับเหมาะแบบพหุมิติ ในระดับต่ำ คือ 0.273 ตัวแปรทั้งสองสามารถอธิบายความผันแปรที่เกิดขึ้นในผลประเมินเพียง 7.4% และจากผลการวิเคราะห์ความแปรปรวน พบว่า ไม่มีตัวแปรใดที่ส่งต่อผลการประเมินโปรแกรมการทดสอบปรับเหมาะแบบพหุมิติอย่างมีนัยสำคัญ (F=2.81, p = 0.07)

4. สรุปผลการวิจัยรวมทั้ง 3 แหล่ง

สรุปผลการวิจัยจากการจำลองข้อมูล การใช้ข้อมูลทุติยภูมิ และการใช้ข้อมูลปฐมภูมิเพื่อตอบวัตถุประสงค์การวิจัยสามารถสรุปได้ดังนี้

1) คุณภาพการทดสอบปรับเหมาะแบบพหุมิติระหว่างวิธีการคัดเลือกข้อสอบขั้นแรก โดยภาพรวม พบว่า วิธี Kullback-Leibler information มีคุณภาพสูงกว่าวิธี Fisher information

2) สำหรับลำดับข้อสอบ ที่มีคุณภาพการทดสอบปรับเหมาะแบบพหุมิติสูงสุด พบว่าลำดับข้อสอบแบบผสมผสานมิติมีคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติสูงกว่า ลำดับข้อสอบแบบจำแนกมิติ

3) จากการศึกษาปฏิสัมพันธ์ระหว่างวิธีการคัดเลือกข้อสอบขั้นแรกและลำดับข้อสอบ พบว่า มีปฏิสัมพันธ์ระหว่าง วิธีการคัดเลือกข้อสอบขั้นแรก และลำดับข้อสอบที่ร่วมกันส่งผลต่อคุณภาพการทดสอบ โดยที่วิธี Kullback-Leibler information แบบผสมผสานมิติ มีคุณภาพสูงสุด

4) สำหรับผลกา รศึกษาผลการประเมิน โปรแกรม การทดสอบความสามารถด้านวิทยาศาสตร์ด้วยวิธีการทดสอบปรับเหมาะแบบพหุมิติ จากการทดสอบจริง พบว่า ผู้สอบมีความพึงพอใจอยู่ในระดับมาก โดยที่ ภูมิหลังของนักเรีย นเกี่ยวกับเกรดเฉลี่ยสะสม และทักษะคอมพิวเตอร์ ไม่ส่งผลต่อผลการประเมินโปรแกรมการทดสอบปรับเหมาะแบบพหุมิติ

สรุปผลการวิจัยจำแนกตามวัตถุประสงค์การวิจัย ได้นำเสนอไว้ดังตารางที่ 5.1

ตารางที่ 5.1 สรุปผลการวิจัยจำแนกตามวัตถุประสงค์การวิจัย

วัตถุประสงค์การวิจัย	ผลการวิจัย
1. เพื่อเปรียบเทียบคุณภาพของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบพหุมิติภายใต้วิธีการคัดเลือกข้อสอบขั้นแรกและลำดับข้อสอบที่แตกต่างกัน	
1.1 เพื่อเปรียบเทียบวิธีการคัดเลือกข้อสอบขั้นแรกที่แตกต่างกัน 2 วิธี	Kullback-leibler information มีคุณภาพการทดสอบสูงกว่า Fisher information
1.2 เพื่อเปรียบเทียบลำดับข้อสอบที่แตกต่างกัน 2 รูปแบบ	ลำดับข้อสอบแบบผสมผสานมิติมีคุณภาพสูงกว่าลำดับข้อสอบแบบจำแนกมิติ
1.3 เพื่อศึกษาปฏิสัมพันธ์ระหว่างวิธีการคัดเลือกข้อสอบขั้นแรกและลำดับข้อสอบ	เกิดปฏิสัมพันธ์ระหว่างวิธีการคัดเลือกข้อสอบขั้นแรกและลำดับข้อสอบ โดบพบว่า KL- Mixed มีคุณภาพสูงสุด
2. เพื่อศึกษาผลการประเมินโปรแกรม การทดสอบความสามารถด้านวิทยาศาสตร์ด้วยวิธีการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์	ผู้สอบมีความพึงพอใจระดับมาก และพบว่าภูมิหลัง ของนักเรียน (เกรดเฉลี่ยสะสม และทักษะคอมพิวเตอร์) ไม่ส่งผลต่อผลการประเมินโปรแกรมการทดสอบปรับเหมาะแบบพหุมิติ

อภิปรายผล

สำหรับการอภิปรายผลผู้วิจัยจะขอเสนอใน 2 รูปแบบ คือ การอภิปรายผลตามสมมติฐานการวิจัย และการอภิปรายผลในประเด็นอื่น ๆ ซึ่งค้นพบจากงานวิจัย ซึ่งมีรายละเอียดดังนี้

1. สมมติฐานข้อที่ 1 วิธีการคัดเลือกข้อสอบขั้นแรกด้วยวิธี Kullback-Leiber information น่าจะมีคุณภาพการทดสอบที่สูงกว่าวิธีการคัดเลือกข้อสอบขั้นแรกด้วยวิธี และ Fisher information

สรุปผลการวิจัยพบว่า วิธีการคัดเลือก ข้อสอบขั้นแรกด้วยวิธี Kullback-Leiber information มีคุณภาพการทดสอบที่ สูงกว่าวิธีการ คัดเลือกข้อสอบขั้นแรกด้วยวิธี Fisher information แต่การศึกษาในงานวิจัยที่ผ่านมายังไม่พบการศึกษาเปรียบเทียบ วิธีการคัดเลือกข้อสอบขั้นแรก ในบริบทของการทดสอบปรับเหมาะแบบพหุมิติ มีเพียงการศึกษาในบริบทของการทดสอบปรับเหมาะแบบเอกมิติเท่านั้น อย่างไรก็ ตามผลการวิจัยครั้งนี้ ให้ผลสอดคล้องกับผลที่ได้จากการศึกษาในบริบทเอกมิติ ของ Chang และ Ying (1996) ซึ่งกล่าวว่า Kullback-Leiber information เหมาะสำหรับใช้ในกรณีที่มี n ขนาดเล็ก นั่นคือ เหมาะกับข้อสอบจำนวนน้อย และจากงานวิจัยของสิริลักษณ์ เกษรปทุมานันท์ (2549) ที่ศึกษาจากการทดสอบแบบเอกมิติ พบว่าเกณฑ์การคัดเลือกข้อสอบขั้นแรกด้วยวิธี Kullback-Leiber information มีค่าความตรงตามสภาพของการทดสอบสูงกว่าวิธี Fisher information

2. สมมติฐานข้อ 2 ลำดับข้อสอบ แบบจำแนกมิติน่าจะมีคุณภาพการทดสอบที่สูงกว่าลำดับข้อสอบแบบผสมผสานมิติ

Segall (1996) ได้กล่าวว่า แม้ว่าการผสมผสานเนื้อหาข้อสอบจะไม่มีผลกระทบต่อการทำหน้าที่ของข้อสอบ แต่จะส่งผลต่อระดับความวิตกกังวลหรือความไม่สบายใจของผู้สอบ ซึ่งรูปแบบการทดสอบดังกล่าวไม่เป็นที่คุ้นเคยของผู้สอบ จึงอาจจะส่งผลต่อความแม่นยำในการประมาณค่าความสามารถผู้สอบ อย่างไรก็ตามผลการวิจัยพบว่า ลำดับข้อสอบแบบผสมผสานมิติมีคุณภาพการทดสอบสูงกว่าลำดับข้อสอบแบบจำแนกมิติ แม้จะศึกษาจากการทดสอบจริง ที่มีปัจจัยของความวิตก กังวลร่วมเข้ามาในการทดสอบซึ่งแตกต่างจากการจำลองข้อมูล และข้อมูลทุติยภูมิ ผลการวิจัยยังชี้ว่า ลำดับข้อสอบแบบผสมผสานมิติมีคุณภาพการทดสอบสูงกว่า ทั้งนี้เนื่องมาจากลำดับข้อสอบ แบบผสมผสานมิติ คัดเลือกข้อสอบในคลังข้อสอบที่เหมาะสมกับผู้สอบมากที่สุด โดยไม่คำนึงว่าวัดในมิติใด แต่ลำดับข้อสอบ แบบจำแนกมิติ การคัดเลือกข้อสอบจะถูกบังคับให้เลือกข้อสอบเฉพาะมิติที่กำหนด ทำให้ข้อสอบที่ได้รับอาจมีสารสนเทศไม่สูงเท่าที่ควร

3. สมมติฐานข้อ 3 วิธีการคัดเลือกข้อสอบขั้นแรกและลำดับข้อสอบไม่น่าจะเกิดปฏิสัมพันธ์ที่ส่งผลต่อคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์

สรุปผลการวิจัยจากการศึกษาครั้งนี้ พบว่า มีปฏิสัมพันธ์ระหว่างวิธีการคัดเลือกข้อสอบขั้นแรกและ ลำดับข้อสอบ ที่ร่วมกันส่งผลต่อคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติ โดยพบว่า ผลวิจัยจากการจำลองข้อมูลและข้อมูลทุติยภูมิ พบปฏิสัมพันธ์ระหว่างตัวแปรทั้งสอง โดยพบว่า วิธี Kullback-Leibler information แบบผสมผสานมิติ มีคุณภาพสูงสุด ซึ่งสอดคล้องกับการอภิปรายผลในข้อ 1 และ2 วิธี Kullback-Leibler information และลำดับข้อสอบ แบบผสมผสานมิติมีคุณภาพการทดสอบที่สูงกว่า อย่างไรก็ตาม ยังไม่มีงานวิจัยใดที่ศึกษาปฏิสัมพันธ์ที่เกิดจาก วิธีการคัดเลือกข้อสอบ ขั้นแรกและ ลำดับข้อสอบ จึงไม่มีงานวิจัยอื่นๆ มาสนับสนุนผลการวิจัยครั้งนี้

4. สมมติฐานข้อ 4 ผลการประเมินโปรแกรมการทดสอบความสามารถด้านวิทยาศาสตร์ด้วยวิธีการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์ของนักเรียนกลุ่มตัวอย่างน่าจะมีความพึงพอใจอยู่ในระดับมากขึ้นไป

ผลการวิจัย พบว่า นักเรียนกลุ่มตัวอย่างมีความพึงพอใจต่อการใช้ โปรแกรมการทดสอบความสามารถด้านวิทยาศาสตร์ ด้วยวิธีการทดสอบปรับเหมาะแบบพหุมิติ อยู่ในระดับ มาก ทั้งนี้อาจเนื่องมาจากความไม่ซับซ้อนของโปรแกรม และนักเรียนมีความคุ้นเคยในการใช้คอมพิวเตอร์ นอกจากนั้น พบว่าภูมิหลังของ นักเรียนเกี่ยวกับผลสัมฤทธิ์ทางการเรียน (เกรดเฉลี่ยสะสม) และทักษะการใช้คอมพิวเตอร์ ไม่ส่งผลต่อ ผลการประเมิน โปรแกรม การทดสอบความสามารถด้านวิทยาศาสตร์ด้วยวิธีการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์ ทั้งนี้เนื่องมาจากการออกแบบโปรแกรมการทดสอบปรับเ หมาะให้ง่ายต่อการใช้งาน ทำให้ความ แตกต่างของผลสัมฤทธิ์ทางการเรียนไม่ส่งผลต่อการประเมินโปรแกรม รวมถึง การเข้าถึงคอมพิวเตอร์และอินเตอร์เน็ตของนักเรียนในยุคปัจจุบันไม่ใช่เรื่องที่ไกลตัว ทำให้นักเรียนเกิดความชำนาญในการใช้คอมพิวเตอร์ ดังนั้นทักษะการใช้คอมพิวเตอร์ ซึ่งวัดทักษะเกี่ยวกับการประมวลผลคำ (microsoft word) และการนำเสนอผลงาน (microsoft PowerPoint) จึงไม่ส่งผลต่อผลประเมินการใช้โปรแกรม เพราะการทดสอบผ่านโปรแกรมที่ผู้วิจัยพัฒนาขึ้น ไม่ได้ใช้ทักษะคอมพิวเตอร์ที่ซับซ้อน สำหรับผลการประเมินโดยภารวม พบว่า ผู้สอบมีความพึงพอใจมาก (M = 3.52)

5. จากการวิจัยครั้งนี้ พบว่า คุณภาพของการทดสอบปรับเหมาะแบบพหุมิติมีค่าไม่สูงเท่าที่ควร ทั้งนี้เนื่องมาจาก ค่าอำนาจจำแนกของข้อสอบที่มีค่าต่ำ และความสัมพันธ์ระหว่างมิติของข้อสอบที่นำมาศึกษาก็มีความสัมพันธ์ในระดับต่ำจึงทำให้คุณภาพไม่สูงเท่าที่ควร ซึ่งสอดคล้องกับ Chang และ Wang (2011) ซึ่งกล่าวว่า ค่าอำนาจจำแน กมีผลต่อการคัดเลือกข้อสอบที่

เหมาะสม และจากการศึกษาของ Wang และ Chen (2004) พบว่า ความสัมพันธ์ระหว่างมิติระดับสูงจะมีประสิทธิภาพการวัดที่สูงกว่า นอกจากนั้น รูปแบบข้อสอบพหุมิติภายในข้ อ (within-item multidimensionality) จะมีประสิทธิภาพการวัดที่สูงกว่ารูปแบบข้อสอบพหุมิติระหว่างข้อ (between-item multidimensionality) ด้วยเหตุผลดังกล่าวจึงทำให้คุณภาพของการทดสอบปรับเหมาะแบบพหุมิติครั้งนี้มีค่าค่อนข้างต่ำ

6. งานวิจัยนี้มีข้อจำกัดที่ จำนวนมิติที่ศึกษา จากแหล่งข้อมูลทั้ง 3 แหล่ง เนื่องจากผู้วิจัยศึกษาจากการจำลองข้อมูลเพียง 2 มิติ เพื่อ ตรวจสอบความถูกต้องของกระบวนการทดสอบปรับเหมาะแบบพหุมิติ ก่อนนำไปประยุกต์ใช้กับ ข้อมูลทุติยภูมิและข้อมูลปฐมภูมิ ที่ศึกษา 3 มิติ ในวิชาชีววิทยา เคมี และฟิสิกส์ ซึ่งผลวิจัยทีได้เป็นผลที่เกิดจากการศึกษา 2 มิติ และ 3 มิติ แม้ผลการศึกษาจากการทดสอบปรับเหมาะขั้นแรก 5 ข้อ จะให้ผลที่สอดคล้องกัน แต่หากพิจารณาการทดสอบจนจบกระบวนการ 30 ข้อ จะให้ผลการวิจัยที่แตกต่างกัน อย่างไรก็ตาม ผลลัพธ์ที่ไม่สอดคล้องกันนั้น อาจจะมาจากหลายปัจจัย โดยเฉพาะค่าพารามิเตอร์ที่นำมาใช้ในแต่ละแหล่งข้อมูล หรือแม้แต่ความตั้งใจในการ ทำข้อ สอบในกรณีการเก็บข้อมูลจริง โดยพบว่า นักเรียน ส่วนใหญ่ทำข้อสอบจำนวน 30 ข้อ ใช้เวลาเพียง 11 -20 นาที จากเวลาที่กำหนดไว้ 1 ชั่วโมง ซึ่งอาจส่งผลต่อความแม่นยำในการประมาณค่าความสามารถของผู้สอบ

ข้อเสนอแนะการนำไปใช้

1. ข้อเสนอแนะในการนำวิธีการคัดเลือกข้อสอบไปใช้ในการทดสอบปรับเหมาะแบบพหุมิติ คือ ควรเลือกใช้วิธี Kullback-Leibler information สำหรับการคัดเลือกข้อสอบขั้นแรก หรือ 5 ข้อแรก ซึ่งจะทำให้การประมาณค่าความสามารถมีความแม่นยำมากขึ้น แม้จะทำข้อสอบเพียงไม่กี่ข้อ สำหรับการคัดเลือกข้อสอบในขั้นต่อไปในทางปฏิบัติสามารถที่จะเปลี่ยนวิธีการคัดเลือกข้อสอบได้ โดยที่จะไม่กระทบต่อความแม่นยำในการประมาณค่าความสามารถของผู้สอบ แต่มีความส ะดวกในทางปฏิบัติ เนื่องจากความซับซ้อนของการคำนวณค่าของวิธี Kullback-Leibler information โดยเฉพาะเมื่อทำข้อสอบจำนวนมากขึ้น

2. ลำดับข้อสอบ ที่เหมาะสมที่ควรเลือกใช้สำหรับการทดสอบปรับเหมาะแบบพหุมิติ ควรจะใช้ ลำดับข้อสอบ แบบผสมผสานมิติเพื่อให้ประมาณค่าความส ามารถได้อย่างแม่นยำ อย่างไรก็ตามควรจะมีการควบคุมเนื้อหาของข้อสอบให้มีความสมดุลกันในทุกมิติ เพื่อลดความลำเอียงที่จะเกิดกับผู้สอบแต่ละคน พร้อมทั้งชี้แจงลักษณะข้อสอบให้ผู้สอบเข้าใจ หรือแสดงตัวอย่างการทดสอบในรูปแบบนี้ให้ผู้สอบคุ้นเคยก่อนการทดสอบ

3. ในกรณีที่ต้อง การเลือกวิธีการคัดเลือกข้อสอบขั้นแรกและลำดับข้อสอบที่เหมาะสม งานวิจัยนี้ พบว่า ควรเลือกใช้วิธีการคัดเลือกข้อสอบขั้นแรกด้วย วิธี Kullback-Leibler information โดยกำหนดให้มีลำดับข้อสอบแบบผสมผสานมิติ เนื่องจากผลจากการศึกษาปฏิสัมพันธ์วิธีการดังกล่าวมีคุณภาพการทดสอบสูงที่สุด

4. การวิจัยครั้งนี้ พบว่า การทดสอบที่มีจำนวนมิติมากขึ้นเท่าใด การประมวลผลสำหรับการทดสอบปรับเหมาะแบบพหุมิติ ยิ่งมีความซับซ้อนมากขึ้นเท่านั้น และยังส่งผลถึงการคัดเลือกข้อสอบแต่ละข้อล่าช้า ซึ่งเป็นจะปัญหาเมื่อนำไปใช้กับการทดสอบจริง ดังนั้นการคัดเลือกวิธีการมาใช้ในทางปฏิบัติควรจะเลือกวิธีที่มีความซับซ้อนน้อย แต่ให้ผลการประมาณค่าที่มีความแม่นยำ ซึ่งในที่นี้คือ วิธีของ Fisher information และจะเห็นว่างานวิจัยที่ผ่านมาส่วนใหญ่จะศึกษาด้วยวิธีการจำลองข้อมูล เพื่อหาคำตอบทางทฤษฎีเท่านั้น จึงไม่พบปัญหาที่เกิดขึ้นในทางปฏิบัติ

5. สำหรับการประมาณค่าพารามิเตอร์ข้อสอบแบบพหุมิติด้วยโปรแกรม TESTFACT จะเห็นว่าค่าพารามิเตอร์ที่ได้จะ ค่อนข้างต่ำ (under estimate) กว่าการประมาณค่าแบบเอกมิติด้วยโปรแกรมการวิเคราะห์อื่นๆ ดังนั้ นจึงควรระมัดระวังการตีความค่าพารามิเตอร์ที่ประมาณค่าได้ นอกจากนี้ การเลือกวิธีการประมาณค่า ควรจะเลือกวิธีที่เหมาะสมกับลักษณะความเป็นพหุมิติของข้อสอบ นั่นคือ หากข้อสอบที่นำมาวิเคราะห์มีลักษณะพหุมิติระหว่างข้อ (between-item multidimensionality) และทราบ ว่า ข้อสอบแต่ละข้อวัดความสามารถใน มิติใด การวิเคราะห์ที่เลือกใช้ควรเป็นลักษณะของ bi-factor ซึ่งเป็นการ วิเคราะห์ เชิงยืนยัน (confirmatory factor analysis) ว่าข้อสอบข้อนั้นๆ วัดความสามารถในมิติที่เรากำหนดไว้ และค่าพารามิเตอร์ที่ได้จะแยกตามมิติที่เรากำหนดไว้ตอนต้น ในกรณีที่ข้อสอบมีลักษณะพหุมิติภายในข้อ (within-item multidimensionality) การวิเคราะห์ที่เลือกใช้ควรเป็นลักษณะของ full-information item factor analysis ซึ่งการวิเคราะห์ในลักษณะนี้จะเป็นการสำรวจว่าข้อสอบแต่ละข้อวัดความสามารถในมิติใด โดยพิจารณาที่ค่าน้ำหนักองค์ประกอบ

6. สำหรับการนำแนวคิดของ Bayesian มาใช้ในการคัดเลือกข้อสอบและประมาณค่าความสามารถของผู้สอบนั้น ผู้ใช้ต้องกำหนดค่าความรู้ก่อนหน้า (prior) ซึ่งค่านี้จะทำให้โปรแกรมการทดสอบปรับเหมาะคัดเลือกข้อสอบที่มีความเหมาะสมกับผู้สอบ และสามารถประมาณค่าความสามารถได้อย่างแม่นยำ ขึ้น อย่างไรก็ตาม เราสามารถกำหน ดให้ค่าความรู้ก่อนหน้าไม่มีผลต่อการทดสอบได้ ในกรณีที่ไม่มีข้อมูลใดๆ เกี่ยวกับผู้สอบ

ข้อเสนอแนะในการวิจัยครั้งต่อไป

1. การวิจัยครั้งนี้เปรียบเทียบวิธีการคัดเลือกข้อสอบขั้นแรกและ ลำดับข้อสอบที่ส่งผลต่อคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติ ภายใต้คุณลักษณะข้อสอบ พหุมิติระหว่างข้อ (between-item multidimensionality) เท่านั้น ดังนั้นจึงควรมีการศึกษารูปแบบของข้อสอบที่เป็นลักษณะพหุมิติภายในข้อ (within-item multidimensionality) นั่นคือ ข้อสอบ 1 ข้อ สามารถวัดความสามารถได้หลายมิติ ซึ่งข้อสอบในลักษณะนี้ น่าจะให้คุณภาพการวั ดที่สูงกว่า เนื่องจากมีความสัมพันธ์ระหว่างมิติสูง และควรจะมีการเปรียบเทียบวิธีการคัดเลือกข้อสอบและลำดับข้อสอบทั้งรูปแบบพหุมิติระหว่างข้อ และพหุมิติภายในข้อ เพื่อศึกษาปฏิสัมพันธ์ระหว่างตัวแปร ซึ่งผลลัพธ์น่าจะให้สารสนเทศที่สามารถนำไปประยุกต์ใช้ได้อย่างเหมาะสม

2. ผลการวิจัยครั้งนี้จะเห็นว่า การทดสอบปรับเหมาะแบบพหุมิติในกรณีของ ลำดับข้อสอบแบบผสมผสานมิตินั้น แม้จะให้คุณภาพการวัดสูงกว่า แต่ อาจทำให้ผู้สอบเกิดความวิตกกังวลว่าข้อสอบข้อถัดไปจะ เป็นวิชาใด (แม้ว่าการวิจัยครั้งนี้จะไม่ได้ชี้ให้เห็นชัดเจนว่าผู้สอบมีความวิตกกังวลกับการสอบรูปแบบดังกล่าว) รวมถึงข้อสอบที่ได้รับยังไม่มีความสมดุลของเนื้อหา นั่นคือผู้สอบอาจจะได้รับข้อสอบชีววิทยามากกว่าวิชาอื่นๆ ซึ่งอาจจะส่งผลให้การประมาณค่าความสามารถได้ไม่แม่นยำ ดังนั้นการวิจัยครั้งต่อไปควรจะมีการควบคุมสมดุลของข้อสอบ (content balancing) เพื่อให้ผู้สอบได้ทำข้อสอบครอบคลุม สมดุลไปทุกมิติที่มุ่งวัด

3. สำหรับการนำการทดสอบปรับเหมาะไปใช้กับการทดสอบที่มีผลกระทบสูง เช่น การสอบคัดเลือกเข้าในระดับอุดมศึกษา นั้น ควรจะมีการศึกษาเกี่ยวกับอัตราการได้รับข้อสอบซ้ำ (item exposure) ซึ่งจะเกี่ ยวโยงเกี่ยวกับความปลอดภัยของการทดสอบ ซึ่งข้อสอบซ้ำในที่นี้หมายถึงข้อสอบที่ถูกคัดเลือกให้กับผู้สอบแต่ละคน เป็นข้อสอบข้อเดียวกันจำนวนหลายข้อ ซึ่งอาจจะทำให้เกิดการทุจริตง่ายขึ้น ดังนั้นจึงควรมีการควบคุมอัตราการสอบซ้ำ ซึ่งยังไม่มีงานวิจัยใดศึกษาในของการทดสอบปรับเหมาะแบบพหุมิติ

4. ควรมีการศึกษาความคล้ายคลึงของวิธีการคัดเลือกข้อสอบ โดย ศึกษาอัตราเหลื่อมล้ำของข้อสอบ (overlap rate) ระหว่างการคัดเลือกข้อสอบแต่ละวิธี ซึ่งเป็นสัดส่วนของข้อสอบที่ใช้ร่วมกันระหว่างสองวิธี (Chen et al., 2000) เพื่อให้ทราบความแตกต่ างที่ชัดเจนของแต่ละวิธีและหากวิธีการเหล่านั้น เลือกข้อสอบไม่แตกต่างกันมากนัก ผู้วิจัยควรจะเลือกวิธีการที่มีความซับซ้อนของการคำนวณน้อยกว่า เพราะการคำนวณที่ซับซ้อนจะส่งผลถึงระยะเวลาของการประมวลผลค่าต่างๆ ล่าช้า การดำเนินการสอบอาจจะเกิดการชะงักขึ้นกลางคันได้

5. การวิจัยครั้งนี้ศึกษาเพียงการคัดเลือกข้อสอบขั้นแรก และ ลำดับข้อสอบ ที่ส่งผลต่อคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติเท่านั้น ยังคงมีประเด็นอื่น ๆ ที่เป็นกระบวนการของการทดสอบปรับเหมาะ เช่น คลังข้อสอบ วิธีการประมาณค่าความสามารถ และเกณฑ์การหยุด ซึ่งเป็นตัวแปรที่น่าสนใจในการศึกษาในบริบทของการทดสอบแบบพหุมิติ อีกทั้งประเด็นต่าง ๆ ที่เป็นจุดอ่อนของการทดสอบแบบปรับเหมาะ เช่น การที่ผู้สอบ ไม่สามารถย้อนกลับไปแก้ไขข้อสอบข้อก่อนหน้า หรือผู้สอบไม่สามารถทบทวนคำตอบที่ได้ตอบไปแล้ว

6. การวิจัยครั้งนี้ ไม่ได้เจาะจงศึกษาผลที่เกิดขึ้นกับผู้สอบในแต่ละช่วงความสามารถ จึงไม่สามารถระบุได้ว่า วิธีการคัดเลือกข้อสอบขั้นแรก หรือ ลำดับข้อสอบ ให้ผลลัพธ์ที่แตกต่างกันหรือไม่ เมื่อพิจารณาในแต่ละช่วงความสามารถ ซึ่งสามารถศึกษาได้จากการจำลองข้อมูลจำแนกตามช่วงความสามารถจาก -3 ถึง 3 โดยจำลองข้อมูลซ้ำในแต่ละช่วงความสามารถ และพิจารณาค่า BIAS และค่า RMSE ซึ่งจะแตกต่างจากผลการวิจัยครั้งนี้

รายการอ้างอิง

ภาษาไทย

ชุติวัฒน์ สุวัตถิพงศ์. (2551). *โมเดลเชิงสาเหตุและผลของการรับรู้ความสามารถด้านคอมพิวเตอร์ของตนเองที่มีต่อความสามารถทางคอมพิวเตอร์.* วิทยานิพนธ์ปริญญามหาบัณฑิต. ภาควิชาวิจัยการศึกษา. คณะครุศาสตร์. จุฬาลงกรณ์มหาวิทยาลัย.

ต่าย เซี่ยงฉี. (2534). *การศึกษาเปรียบเทียบประสิทธิภาพในการประมาณค่าความสามารถของผู้สอบจากการทดสอบเทเลอร์รูปปิรามิดที่มี รูปแบบ จำนวนขั้น และวิธีการให้คะแนนที่แตกต่างกัน โดยใช้วิธีมอนติคาร์โล* วิทยานิพนธ์ปริญญาดุษฎีบัณฑิต สาขาวิชาวิจัยและพัฒนาหลักสูตร คณะศึกษศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ

ทรงศักดิ์ ภูสีอ่อน. (2552). *การประยุกต์ใช้ SPSS วิเคราะห์ข้อมูลงานวิจัย.* พิมพ์ครั้งที่ 3. มหาสารคาม: มหาวิทยาลัยมหาสารคาม.

นันทิยา พึ่งคำ. (2531). *การเปรียบเทียบคุณภาพการทดสอบแบบ ซี เอ ทีและแบบประเพณีนิยมในการวัดความสามารถด้านคำศัพท์ภาษาอังกฤษของนักเรียนชั้นมัธยมศึกษาปีที่3.* วิทยานิพนธ์ปริญญามหาบัณฑิต. ภาควิชาวิจัยการศึกษา. บัณฑิตวิทยาลัย. จุฬาลงกรณ์มหาวิทยาลัย.

พงษ์พิชิต พรมสิทธิ์. (2544). *การดำเนินการสอบแบบทดสอบเทเลอร์ ตามยุทธวิธีของเบส์ในวิชาคณิตศาสตร์ ค 014 โดยเว็บเพจ.* วิทยานิพนธ์ปริญญามหาบัณฑิต. คณะศึกษาศาสตร์. มหาวิทยาลัยขอนแก่น.

พัชรี จันทร์เพ็ง. (2550). *การเปรียบเทียบคุณภาพของวิธีการเชื่อมโยงคะแนนตามทฤษฎีการตอบสนองข้อสอบแบบพหุมิติภายใต้การหมุนแกน โครงสร้างเชิงมิติและระดับความสัมพันธ์ที่แตกต่างกัน.* วิทยานิพนธ์ปริญญาดุษฎีบัณฑิต. ภาควิชาวิจัยการศึกษา. คณะครุศาสตร์. จุฬาลงกรณ์มหาวิทยาลัย.

พิมพ์สิริ เธียรนรเศรษฐ์. (2549). *การเปรียบเทียบค่าความสามารถ ลักษณะของการเปลี่ยนคำตอบ และเวลาที่ใช้ในการทำข้อสอบ ของผู้สอบในการทดสอบแบบปรับเหมาะกับความสามารถของผู้สอบด้วยคอมพิวเตอร์ เมื่อเงื่อนไขการทดสอบและระดับความสามารถของผู้สอบแตกต่างกัน.* วิทยานิพนธ์ปริญญามหาบัณฑิต. สาขาวิชาการวัดและประเมินผลการศึกษา. คณะครุศาสตร์. จุฬาลงกรณ์มหาวิทยาลัย.

รังสรรค์ มณีเล็ก. (2540). *ผลของตัวแปรบางตัวต่อความเที่ยงตรงเชิงสภาพและจำนวนข้อสอบที่ใช้ในการทดสอบแบบปรับเหมาะกับความสามารถของผู้สอบด้วยคอมพิวเตอร์.* วิทยานิพนธ์ปริญญาดุษฎีบัณฑิต. สาขาการทดสอบและวัดผลการศึกษา คณะศึกษาศาสตร์. มหาวิทยาลัยศรีนครินทรวิโรฒ.

วสันต์ ทองไทย. (2539). การเปรียบเทียบผลการประมาณค่าความสามารถด้วยวิธีการทดสอบแบบเทเลอร์รูปพีระมิดขนาดขั้นคงที่และรูปพีระมิดข้างตัด: การทดสอบทักษะกระบวนการทางวิทยาศาสตร์. วิทยานิพนธ์ปริญญามหาบัณฑิต. สาขาวิจัยและประเมินผลการศึกษา. คณะศึกษาศาสตร์. มหาวิทยาลัยเกษตรศาสตร์.

ศิริชัย กาญจนวาสี. (2538). *การทดสอบแบบปรับเหมาะกับความสามารถของผู้สอบ.* กรุงเทพฯ: ศูนย์ตำราและเอกสารทางวิชาการ คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.

ศิริชัย กาญจนวาสี. (2550). *ทฤษฎีการทดสอบแนวใหม่.* พิมพ์ครั้งที่ 3. กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

สงบ ลักษณะ. (2525). ความสามารถแท้กับการตอบข้อสอบ. วารสารวัดผลการศึกษา. 1(พ.ค.-ส.ค. 2525): 47-54

สิริลักษณ์ เกษรปทุมานันท์. (2549). การเปรียบเทียบความตรงตามสภาพในการประมาณค่าความสามารถของผู้สอบจากการทดสอบแบบปรับเหมาะโดยใช้คอมพิวเตอร์ ที่ใช้เกณฑ์การคัดเลือกข้อสอบขั้นแรกอัตราการใช้ข้อสอบซ้ำ และเกณฑ์ยุติการทดสอบที่ต่างกัน. วิทยานิพนธ์ปริญญามหาบัณฑิต. สาขาวิชาการวัดและประเมินผลการศึกษา. คณะครุศาสตร์. จุฬาลงกรณ์มหาวิทยาลัย.

สุพัฒน์ สุกมลสันต์. (2538). *การวิเคราะห์ข้อทดสอบแนวใหม่ด้วยคอมพิวเตอร์.* กรุงเทพฯ: อักษรการพิมพ์

สุพัฒน์ สุกมลสันต์. (2539). *ธนาคารข้อสอบและการทดสอบปรับเปลี่ยนด้วยคอมพิวเตอร์.* กรุงเทพฯ: วิทยพัฒน์.

สุวิมล ติรการนันท์. (2548). *การตรวจสอบความเป็นเอกมิติ (Examining The Unidimensionality) (ปรับปรุงใหม่).* กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย

ภาษาอังกฤษ

Ackerman, T. A. (1991). The Use of Unidimensional Parameter Estimates of Multidimensional Item in Adaptive Testing. *Applied Psychological Measurement* 15: 12-24.

Ackerman, T. A. (1994). Using Multidimensional Item Response Theory to Understand What Items and Tests are Measuring. *Applied Psychological Measurement* 18: 255-278.

Berger, M. P. F., & Veerkamp, W. J. J. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics 22*: 203-226.

Bloxom, B. M., and Vale, C. D. (1987). *Multidimensional adaptive testing: a procedure for sequential estimation of the posterior centroid and dispersion of theta.* Paper presented at the meeting of Psychometric society, Montreal.

Chang, S. and Ansley, T. N. (2003). A comparative study of Item exposure control methods in computerized adaptive testing. *Journal of Education Measurement* 40(1): 77-103.

Chang, H. H., and Ying, Z. (1999). a-Stratified multistage computerized adaptive .testing with b blocking. *Applied psychological measurement* 23(3): 211-222.

Chang, H. H., and Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied psychological measurement* 20(3): 213-229.

Chen, S., Ankenmann,R.D. (2004). Effect of practical constraints on item selection rules at the early stages of computerized adaptive testing. J*ournal of education measurement* 41(2): 149-174.

Chen, S., Ankenmann,R.D. and Chang.H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement* 24(3): 241-255.

Chen, S. and Lei, P. (2005). Controlling item exposure and test overlap in computerized adaptive testing. *Applied Psychological Measurement* 29(3): 204-217.

Cheng, Y., and others. (2008). Multidimensional rasch analysis of a psychological test with multiple subtests. *Educational and psychological measurement* Sep: 1-20.

Diao, Q., and Reckase, M. (2009). Comparison of ability estimation and item selection methods in multidimensional computerized adaptive testing. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.*

Finkelman, M., Nering, M.L., & Roussos, L.A. (2009). A conditional exposure control method for multidimensional adaptive testing. *Journal of Educational Measurement 46*(1): 84–103.

Green and others. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of education measurement* 21: 347-360.

Haley, S. M., and others. (2006). Measurement precision and efficiency of multidimensional computer adaptive testing of physical functioning using the pediatric evaluation of disability inventory. *Arch Phys Med Rehabil* 87: 1223-1229.

Hambleton, R.K. and Swaminathan, H. (1985). *Item response theory principles and applications*. Boston : Kluwer Nijhoff.

Jian-quan, T., and others. (2007). An introduction to the computerized adaptive testing. *US-China Education Review* 4(1): 72-81.

Latu & Chapman. (2002). Computerized adaptive testing. *British Journal of Educational Technology* 33(5): 619 – 622.

Lee, Y., IP, E. H., and Fuh, C. (2008). A strategy for controlling item exposure in multidimensional computerized adaptive testing. *Education and Psychological Measurement* 68(2): 215-232.

Lord, F. M. (1980). *Application of Item Response Theory to Practical Testing Problem.* New Jersey: Hillssdale Erlbaun.

Li,Y.H. & Schafer, W.D. (2004). The context effects of multidimensional CAT on the accuracy of multidimensional abilities and the item exposure rates. *The paper presented at the 2004 annual meeting of the America Educational Research Association,* San Diego CA.

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading: Addison-Welsley Publishing Company.

Latu, E. and Chapman, E. (2002). Computerized adaptive testing . *British Journal of Educational Technology* 33: 619-622.

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied psychological measurement* 20(4): 389-404.

Mulder, J., & van der Linden, W.J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika 74*(2): 273–296.

Mulder, J., & van der Linden, W.J. (2010). Multidimensional adaptive testing with Kullback–Leibler information item selection. In W.J. van der Linden, & C.A.W. Glas (Eds.), *Elements of adaptive testing* (pp. 77–101). New York: Springer.

Petersen,M.A. ; Groenvold.M and Aaronson.N. (2006). Multidimensional computerized adaptive testing of the EORTC QLQ-C30: Basic developments and evaluations. *Quality of Life Research* 15: 315–329.

Rackase, M. D. (1997). The Past and Future of Multidimensional Item Response Theory. *Applied Psychological Measurement* 21(1): 25-36.

Rackase, M. D. (2009). *Multidimensional Item Response Theory*. New York: Springer.

Revuelta, J and Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of educational measurement* 35: 311-327.

Segall, D.O. (1996). Multidimensional adaptive testing. *Psychometrika* 61(2): 331-354.

Segall, D.O. (1999). General ability measurement: an application of Multidimensional adaptive testing. *Paper presented at the annual meeting of the national council on measurement in education.* Montreal, Quebec, Canada. April 20-22

Segall, D.O. (2000). Principles of multidimensional adaptive testing. In W.J. van der Linden and C.A.W. Glas (eds.). *Computerized adaptive testing theory and Practice (pp .53-73).* Dordrecht :Kluwer academic publishers.

Sutton, R . E. (1993). Equity issues in high stakes computerized testing. *Paper presented at the 1993 Annual meeting of the American educational research association,* Atlanta, GA.

Tam, S. S. (1992). A comparison of methods for adaptive estimation of a multidimensional trait. *Dissertation abstracts international.* 53(03), 1646B. (UMI No. 9221219)

Thissen, D and Mislevy, R.J. (1990). Testing algorithms. In Wainer, H. (eds.), *Computerized Adaptive Testing: A Primer*, pp. 103-135. New Jersey: Lawrence Erlbaum Associates.

van der Linden, W.J. (1999). Multidimensional adaptive testing with a minimum error-variance criterion. *Journal of education and behavioral statistics* 24: 398-412.

van der Linden, W.J. (2000). Constrained adaptive testing with shadow tests. In W.J. van der Linden and C.A.W. Glas (eds.). *Computerized adaptive testing theory and Practice (pp .53-73).* Dordrecht :Kluwer academic publishers.

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.

van der Linden, W.J. and Glas C.A.W. (2000). *Computerized adaptive testing theory and Practice.* Dordrecht : Kluwer academic publishers.

Vassar College. (2008). *Calculator for statistical table entries. Fisher r-to-Z transformation*. http://vassarstats.net/tabs.html#fisher. (9 April 2012)

Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika* 67: 575–588.

Wainer, H. (1990). Introduction and history. In Wainer, H. (eds.), *Computerized Adaptive Testing :* A Primer, pp. 1-21. New Jersey: Lawrence Erlbaum Associate.

Wang, C. and Chang, H. (2011). Item selection in multidimensional computerized adaptive testing – gaining information from different angles. *Psychometrika* 76(3): 363-384.

Wang, C. and Chang, H. (2011). Kullback-leibler information and its applications in muti-dimensional adaptive testing. *Psychometrika* 76(1): 13-39.

Wang, W. and Chen, P. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement* 28(5): 295-316.

Weiss, D.J. (1983). Introduction. In Weiss,D.J. (eds.), *New horizons in testing : latent trait test theory and computerized adaptive testing*, pp. 1-7. New York: ACADEMIC.

Weiss, D.J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development* 37: 70-84.

Weiss, D.J. and Schleisman, J.L. (1999). Adaptive testing. In Masters, G.N. and Keeves, J.P. (eds.), *Advances in Measurement in Educational Research and Measurement*. pp138-150. Amsterdam: PERGAMON.

Wiechmann and Ryan. (2003). Reactions to computerized testing in selection contexts. *International journal of selection and assessment* 11 (2/3): 215-229.

ภาคผนวก

ภาคผนวก ก

รายชื่อผู้สร้างข้อสอบและผู้ตรวจสอบคุณภาพข้อสอบ

รายชื่อผู้สร้างข้อสอบ

วิชาชีววิทยา

1. อ.คำตัน วิชัยคำจร โรงเรียนกสิณธร (เซนต์ปีเตอร์) อ.บางบัวทอง จ.นนทบุรี
2. อ.เจตนิพิฐ แท่นทอง โรงเรียนสวายจีกพิทยาคม อ.เมือง จ.บุรีรัมย์
3. อ.ภาณุวัฒน์ เปรมปรี โรงเรียนประเทียบวิทยาทาน อ.วิหารแดง จ.สระบุรี

วิชาเคมี

1. ดร.พีราวุฏฐ์ พิมพ์รอด โรงเรียนโยธินบูรณะ กรุงเทพมหานคร
2. อ.พิพัฒน์พงษ์ สาจันทร์ โรงเรียนสุวรรณคูหาพิทยาสรรค์ อ.สุวรรณคูหา จ.หนองบัวลำภู
3. อ.สกลศักดิ์ มหาพรหม โรงเรียนสายปัญญารังสิต อ.ธัญบุรี จ.ปทุมธานี

วิชาฟิสิกส์

1. ผศ.ดร.แสงเดือน เจริญฉิม โรงเรียนสาธิตแห่งมหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตกำแพงแสน อ.กำแพงแสน จ.นครปฐม
2. อ.วิจิตตา อำไพจิตต์ โรงเรียนวรราชาทินัดดามาตุวิทยาอ.ลาดหลุมแก้ว จ.ปทุมธานี
3. อ.ศรีนวล วรกิจวัฒนา โรงเรียนห้วยแถลงพิทยาคม อ.ห้วยแถลง จ.นครราชสีมา

รายชื่อผู้ตรวจสอบคุณภาพข้อสอบ

วิชาชีววิทยา

อาจารย์ ดร.นันทรัตน์ เครืออินทร์ คณะศึกษาศาสตร์และพัฒนศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตกำแพงแสน

วิชาเคมี

อาจารย์ ดร.มังกร ศรีสะอาด คณะวิทยาศาสตร์ มหาวิทยาลัยมหาสารคาม

วิชาฟิสิกส์

อาจารย์ ดร.ทัศตริน เครือทอง คณะศึกษาศาสตร์และพัฒนศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตกำแพงแสน

ภาคผนวก ข

โครงสร้างข้อสอบวิทยาศาสตร์

ตารางที่ 6.1 โครงสร้างข้อสอบวิชาชีววิทยาจากผู้พัฒนาข้อสอบ 3 คน

สาระการเรียนรู้	ข้อสอบข้อที่ และ จำนวนข้อสอบ			
	คนที่ 1	คนที่ 2	คนที่ 3	รวม
1. การศึกษาชีววิทยา		7, 41-42 (3 ข้อ)	1-2, 34, 43 (4 ข้อ)	7 ข้อ
2. เคมีที่เป็นพื้นฐานของสิ่งมีชีวิต	12-13, 15, 25 (4 ข้อ)	9, 26-27, 40 (4 ข้อ)	5-6 (2 ข้อ)	10 ข้อ
3. เซลล์ของสิ่งมีชีวิต	16-18, 23, 26 (5 ข้อ)	1-2, 23, 43 (4 ข้อ)	3-4 (2 ข้อ)	11 ข้อ
4. ระบบย่อยอาหารและการสลายอาหารเพื่อให้ได้พลังงาน	3 (1 ข้อ)		9-10 (2 ข้อ)	3 ข้อ
5. การรักษาดุลยภาพในร่างกาย	19-20, 28-36 (11 ข้อ)	8, 10, 39 (3 ข้อ)	7-8, 11-12 (4 ข้อ)	18 ข้อ
6. การเคลื่อนที่ของสิ่งมีชีวิต		3-6 (2 ข้อ)	13-14 (2 ข้อ)	4 ข้อ
7. การรับรู้และการตอบสนอง	14 (1 ข้อ)	11 (1 ข้อ)	15-16, 28-30 (5 ข้อ)	7 ข้อ
8. ระบบต่อมไร้ท่อ		12 (1 ข้อ)	17-18 (2 ข้อ)	3 ข้อ
9. พฤติกรรมของสัตว์	4-7 (4 ข้อ)	13, 24-25 (3 ข้อ)	19-21 (3 ข้อ)	10 ข้อ
10. โครงสร้างและหน้าที่ของพืชมีดอก	24 (1 ข้อ)	15, 17 (2 ข้อ)	22-23 (2 ข้อ)	5 ข้อ
11. การสังเคราะห์ด้วยแสง		16, 37 (2 ข้อ)	24-25 (2 ข้อ)	4 ข้อ
12. การสืบพันธุ์ของพืชดอก		19-20, 38 (3 ข้อ)	26-27 (2 ข้อ)	5 ข้อ
13. การถ่ายทอดทางพันธุกรรม	37-39 (3 ข้อ)	14, 21, 31, 33 (4 ข้อ)		7 ข้อ

สาระการเรียนรู้	ข้อสอบข้อที่ และ จำนวนข้อสอบ			
	คนที่ 1	คนที่ 2	คนที่ 3	รวม
14. ยีนและโครโมโซม	21-22, 27 (3 ข้อ)	35 (1 ข้อ)	37-38 (2 ข้อ)	6 ข้อ
15. พันธุศาสตร์และเทคโนโลยีทางDNA		32, 34, 36 (3 ข้อ)	31-33, 35-36 (5 ข้อ)	8 ข้อ
16. ความหลากหลายทางชีวภาพ	40-43 (4 ข้อ)	18, 22, 28- 30 (5 ข้อ)	41-43 (3 ข้อ)	12 ข้อ
17. ระบบนิเวศ	1-2, 8-9 (4 ข้อ)	24-25 (2 ข้อ)	39-40 (2 ข้อ)	8 ข้อ
18. ประชากร	10-11 (2 ข้อ)			2 ข้อ
รวม	**43 ข้อ**	**43 ข้อ**	**43 ข้อ**	**129 ข้อ**

ตารางที่ 6.2 โครงสร้างข้อสอบวิชาเคมีจากผู้พัฒนาข้อสอบ 3 คน

สาระการเรียนรู้	ข้อสอบข้อที่ และ จำนวนข้อสอบ			
	คนที่ 1	คนที่ 2	คนที่ 3	รวม
1. โครงสร้างอะตอมและตารางธาตุ	12 (1 ข้อ)	1, 31 (2 ข้อ)	1-2, 6-7 (4 ข้อ)	7 ข้อ
2. สมบัติของธาตุ และสารประกอบ	1-2, 10-11 (4 ข้อ)	2 (1 ข้อ)		5 ข้อ
3. ปริมาณสารสัมพันธ์ I		11 (1 ข้อ)	8-10 (3 ข้อ)	4 ข้อ
4. พันธะเคมี	13-16 (4 ข้อ)		3-5 (3 ข้อ)	7 ข้อ
5. ปริมาณสัมพันธ์ II	3-7, 9 (6 ข้อ)	4,6,8, 27-28, 35 (6 ข้อ)		12 ข้อ
6. ของแข็ง ของเหลว แก๊ส		3, 16, 23, 33-34 (5 ข้อ)	11-14 (4 ข้อ)	9 ข้อ
7. อัตราการเกิดปฏิกิริยาเคมี	8, 17-18 (3 ข้อ)	7, 24 (2 ข้อ)	15-17 (3 ข้อ)	8 ข้อ

สาระการเรียนรู้	ข้อสอบข้อที่ และ จำนวนข้อสอบ			
	คนที่ 1	คนที่ 2	คนที่ 3	รวม
8. สมดุลเคมี	19-20 (2 ข้อ)	12-15 (4 ข้อ)	18-20 (3 ข้อ)	9 ข้อ
9. กรด-เบส	21-24 (4 ข้อ)	9 (1 ข้อ)	21-25 (5 ข้อ)	10 ข้อ
10. เคมีอินทรีย์	28-30 (3 ข้อ)	5, 10, 29-30 (4 ข้อ)		7 ข้อ
11. เชื้อเพลิงซากดึกดำบรรพ์ และผลิตภัณฑ์	32-34 (3 ข้อ)	22, 25-26 (3 ข้อ)		6 ข้อ
12. สารชีวโมเลกุล	31 (1 ข้อ)	17, 19 (2 ข้อ)		3 ข้อ
13. ไฟฟ้าเคมี	25-26, 35 (3 ข้อ)	20-21, 32 (3 ข้อ)	26-31 (6 ข้อ)	12 ข้อ
14. ธาตุ และสารประกอบในอุตสาหกรรม	27 (1 ข้อ)	18 (1 ข้อ)	32-35 (4 ข้อ)	6 ข้อ
รวม	**35 ข้อ**	**35 ข้อ**	**35 ข้อ**	**105 ข้อ**

ตารางที่ 6.3 โครงสร้างข้อสอบวิชาฟิสิกส์จากผู้พัฒนาข้อสอบ 3 คน

สาระการเรียนรู้	ข้อสอบข้อที่ และ จำนวนข้อสอบ			
	คนที่ 1	คนที่ 2	คนที่ 3	รวม
1. การเคลื่อนที่ในหนึ่งและสองมิติ	6, 30 (2 ข้อ)	2-6 (5 ข้อ)	1-3, 6-7, 10 (6 ข้อ)	13 ข้อ
2. แรง มวล กฎการเคลื่อนที่	3-5, 11 (3 ข้อ)	7 (1 ข้อ)	30 (1 ข้อ)	5 ข้อ
3. การเคลื่อนที่แบบต่าง ๆ	2 (1 ข้อ)	8, 14, 29 (3 ข้อ)		4 ข้อ
4. งานและพลังงาน	8-9 (2 ข้อ)	9-11 (3 ข้อ)	4, 8-9 (3 ข้อ)	8 ข้อ
5. โมเมนตัมและการดล	10, 13 (2 ข้อ)	12-13 (2 ข้อ)		4 ข้อ
6. การเคลื่อนที่แบบหมุน	7, 12 (2 ข้อ)	30 (1 ข้อ)	5 (1 ข้อ)	3 ข้อ

สาระการเรียนรู้	ข้อสอบข้อที่ และ จำนวนข้อสอบ			
	คนที่ 1	คนที่ 2	คนที่ 3	รวม
7. สภาพสมดุลและสภาพยืดหยุ่น		15-16 (2 ข้อ)		2 ข้อ
8. ของไหล	14-15 (2 ข้อ)	17-18 (2 ข้อ)		4 ข้อ
9. ความร้อน	16-17 (2 ข้อ)	19 (1 ข้อ)		3 ข้อ
10. คลื่นกล	18, 20 (2 ข้อ)	20 (1 ข้อ)		3 ข้อ
11. แสงและทัศนอุปกรณ์	19, 21 (2 ข้อ)	21 (1 ข้อ)	29 (1 ข้อ)	4 ข้อ
12. ไฟฟ้าสถิต	22 (1 ข้อ)	22-23 (2 ข้อ)	11, 15 (2 ข้อ)	5 ข้อ
13. ไฟฟ้าและแม่เหล็ก	23-26 (4 ข้อ)	24-27 (4 ข้อ)	14, 16-26 (12 ข้อ)	20 ข้อ
14. คลื่นแม่เหล็กไฟฟ้า	27 (1 ข้อ)		12-13 (2 ข้อ)	3 ข้อ
15. ฟิสิกส์อะตอม	29 (1 ข้อ)		27 (1 ข้อ)	2 ข้อ
16. ฟิสิกส์นิวเคลียร์	28 (1 ข้อ)	28 (1 ข้อ)	28, 30 (2 ข้อ)	4 ข้อ
17. วัสดุอิเล็กทรอนิกส์และการใช้งานวิทยาศาสตร์	1, 31 (2 ข้อ)	1 (1 ข้อ)		3 ข้อ
รวม	**30 ข้อ**	**30 ข้อ**	**30 ข้อ**	**90 ข้อ**

ภาคผนวก ค

ค่าพารามิเตอร์ข้อสอบ

ตารางที่ 6.4 ค่าพารามิเตอร์ข้อสอบจากแบบวัดความถนัดทางวิชาการและวิชาชีพด้านวิทยาศาสตร์ (PAT2)

No.	ITEM	Discrimination				Difficulty	Guessing
		Total	Bio	Chem	Phy		
1	ITEM1	-0.192	0.075	0	0	0.2535	0.017
2	ITEM2	0.137	-0.008	0	0	-0.1627	0.137
3	ITEM3	0.071	0.006	0	0	0.9547	0.147
4	ITEM4	0.041	0.008	0	0	0.7604	0.150
5	ITEM5	-0.059	0.091	0	0	1.1683	0.201
6	ITEM6	0.142	0.177	0	0	1.0047	0.201
7	ITEM7	0.362	0.025	0	0	-0.006	0.202
8	ITEM8	0.306	0.012	0	0	0.1867	0.139
9	ITEM9	0.133	0.055	0	0	-0.2288	0.135
10	ITEM10	0.095	-0.033	0	0	-0.3845	0.147
11	ITEM11	0.387	0.204	0	0	0.4614	0.083
12	ITEM12	0.46	0.151	0	0	-0.0649	0.134
13	ITEM13	0.294	0.251	0	0	0.691	0.199
14	ITEM14	0.147	0.187	0	0	1.0935	0.201
15	ITEM15	0.352	0.13	0	0	0.1696	0.202
16	ITEM16	0.328	0.122	0	0	0.0138	0.202
17	ITEM17	0.038	0.057	0	0	0.3523	0.118
18	ITEM18	0.109	0.192	0	0	0.9478	0.163
19	ITEM19	0.329	0.243	0	0	0.6653	0.202
20	ITEM20	0.303	0.101	0	0	0.0841	0.202
21	ITEM21	0.282	0.228	0	0	0.1825	0.115
22	ITEM22	0.568	0.217	0	0	-0.0675	0.176
23	ITEM23	0.418	0.338	0	0	0.8009	0.201
24	ITEM24	0.188	0.222	0	0	1.0988	0.100
25	ITEM25	0.304	0.19	0	0	0.0892	0.179
26	ITEM26	0.187	0.246	0	0	0.5672	0.197
27	ITEM27	0.253	0.164	0	0	0.4338	0.201

No.	ITEM	Discrimination				Difficulty	Guessing
		Total	Bio	Chem	Phy		
28	ITEM28	-0.03	-0.007	0	0	0.5034	0.201
29	ITEM29	0.172	0.235	0	0	0.2158	0.199
30	ITEM30	0.391	0.136	0	0	0.2537	0.121
31	ITEM31	0.235	0.258	0	0	0.953	0.119
32	ITEM32	0.351	0.154	0	0	0.3312	0.201
33	ITEM33	0.214	0.082	0	0	0.5178	0.201
34	ITEM34	0.124	0.189	0	0	0.3922	0.146
35	ITEM35	0.265	0.148	0	0	0.5706	0.173
36	ITEM36	-0.009	0.089	0	0	0.6837	0.132
37	ITEM37	0.285	0.157	0	0	0.28	0.201
38	ITEM38	0.116	-0.047	0	0	-0.0751	0.130
39	ITEM39	0.651	-0.011	0	0	-0.7473	0.184
40	ITEM40	-0.144	0.007	0	0	0.687	0.153
41	ITEM41	-0.133	0	0.04	0	1.1681	0.201
42	ITEM42	0.57	0	0.161	0	-0.4234	0.184
43	ITEM43	0.388	0	0.209	0	0.3495	0.179
44	ITEM44	0.167	0	0.12	0	0.1108	0.201
45	ITEM45	0.726	0	0.076	0	-0.0749	0.202
46	ITEM46	0.212	0	0.26	0	0.5008	0.198
47	ITEM47	0.567	0	0.23	0	0.0869	0.191
48	ITEM48	0.316	0	0.284	0	0.6723	0.201
49	ITEM49	0.281	0	0.057	0	0.1916	0.201
50	ITEM50	0.575	0	0.097	0	-0.2075	0.164
51	ITEM51	0.528	0	0.122	0	0.6344	0.165
52	ITEM52	0.394	0	0.09	0	0.2295	0.201
53	ITEM53	0.132	0	0.065	0	0.4429	0.180
54	ITEM54	0.289	0	0.313	0	0.8583	0.184
55	ITEM55	0.295	0	0.277	0	0.2693	0.201
56	ITEM56	0.101	0	0.226	0	0.7227	0.119
57	ITEM57	0.36	0	0.251	0	0.599	0.156
58	ITEM58	0.334	0	0.376	0	0.6559	0.202
59	ITEM59	0.381	0	0.281	0	0.1977	0.202
60	ITEM60	0.111	0	0.007	0	0.4812	0.148

No.	ITEM	Discrimination				Difficulty	Guessing
		Total	Bio	Chem	Phy		
61	ITEM61	0.246	0	0.07	0	0.3035	0.200
62	ITEM62	0.387	0	0.152	0	0.1013	0.150
63	ITEM63	0.19	0	0.187	0	0.8896	0.201
64	ITEM64	0.021	0	0.08	0	0.7072	0.173
65	ITEM65	-0.151	0	0.015	0	1.0079	0.159
66	ITEM66	0.093	0	0.019	0	0.4376	0.183
67	ITEM67	0.155	0	0.079	0	0.5676	0.202
68	ITEM68	0.452	0	0.071	0	0.6628	0.156
69	ITEM69	0.235	0	0.014	0	0.0184	0.202
70	ITEM70	-0.129	0	0.05	0	1.2085	0.130
71	ITEM71	0.199	0	0.079	0	-0.3942	0.103
72	ITEM72	0.319	0	0.079	0	0.1562	0.171
73	ITEM73	0.257	0	0	0.095	0.0201	0.201
74	ITEM74	0.604	0	0	0.285	0.4199	0.092
75	ITEM75	0.58	0	0	0.364	0.3334	0.151
76	ITEM76	0.447	0	0	0.276	0.4528	0.196
77	ITEM77	0.526	0	0	0.177	0.0117	0.202
78	ITEM78	0.46	0	0	0.252	0.045	0.030
79	ITEM79	0.232	0	0	0.179	0.8764	0.200
80	ITEM80	0.395	0	0	0.32	1.0013	0.203
81	ITEM81	0.135	0	0	-0.006	-0.0165	0.201
82	ITEM82	0.068	0	0	0.012	0.4923	0.160
83	ITEM83	0.151	0	0	0.044	0.5659	0.094
84	ITEM84	0.368	0	0	0.16	0.4642	0.181
85	ITEM85	0.077	0	0	-0.046	0.9248	0.140
86	ITEM86	0.27	0	0	0.133	0.3341	0.201
87	ITEM87	0.327	0	0	-0.001	-0.1001	0.201
88	ITEM88	0.451	0	0	-0.072	-0.1446	0.141
89	ITEM89	0.077	0	0	0.041	0.612	0.201
90	ITEM90	0.068	0	0	0.057	0.8447	0.202
91	ITEM91	0.407	0	0	0.235	0.3845	0.192
92	ITEM92	0.21	0	0	0.065	0.4007	0.148
93	ITEM93	0.355	0	0	0.137	0.6052	0.202

No.	ITEM	Discrimination				Difficulty	Guessing
		Total	Bio	Chem	Phy		
94	ITEM94	0.379	0	0	0.194	0.3993	0.135
95	ITEM95	0	0	0	0.003	0.6127	0.108
96	ITEM96	0.096	0	0	-0.079	0.5021	0.143
97	ITEM97	0.185	0	0	0.166	0.4542	0.201
98	ITEM98	0.015	0	0	0.018	0.5053	0.201
99	ITEM99	0.277	0	0	0.051	0.3378	0.024
100	ITEM100	-0.084	0	0	-0.026	0.8675	0.193

ตารางที่ 6.5 ค่าพารามิเตอร์ข้อสอบจากแบบวัดความสามารถวิทยาศาตร์กรณีสอบจริง

No.	ITEM	Discrimination				Difficulty	Guessing
		Total	Bio	Chem	Phy		
1	bio 1	0.8	0.26	0	0	0.09	0.02
2	bio 2	0.37	0.38	0	0	-0.22	0.14
3	bio 3	0.73	0.57	0	0	-0.65	0.15
4	bio 4	0.67	0.46	0	0	0.52	0.2
5	bio 5	0.73	0.45	0	0	-0.02	0.2
6	bio 6	0.27	0.17	0	0	0.14	0.2
7	bio 7	0.59	0.22	0	0	0.23	0.14
8	bio 8	0.74	0.26	0	0	-0.28	0.14
9	bio 9	0.51	0.27	0	0	0.16	0.15
10	bio 10	0.74	0.43	0	0	0.12	0.13
11	bio 11	0.62	0.6	0	0	-0.03	0.2
12	bio 12	0.66	0.44	0	0	-0.16	0.2
13	bio 13	0.81	0.05	0	0	0.17	0.12
14	bio 14	0.82	0.22	0	0	-0.41	0.2
15	bio 15	0.8	0.44	0	0	-0.33	0.12
16	bio 16	0.83	0.45	0	0	-0.16	0.18
17	bio 17	0.79	0.36	0	0	0.02	0.2
18	bio 18	0.63	0.64	0	0	-0.12	0.18
19	bio 19	0.45	0.12	0	0	0.37	0.2
20	bio 20	0.59	0.51	0	0	0.23	0.2
21	bio 21	0.73	0.28	0	0	0.02	0.2
22	bio 22	0.61	0.31	0	0	0.08	0.2

No.	ITEM	Discrimination				Difficulty	Guessing
		Total	Bio	Chem	Phy		
23	bio 23	0.62	0.33	0	0	-0.08	0.12
24	bio 24	0.75	0.55	0	0	-0.01	0.2
25	bio 25	0.74	0.39	0	0	-0.04	0.2
26	bio 26	0.4	0.39	0	0	-0.18	0.13
27	bio 27	0.19	0.16	0	0	0.79	0.2
28	bio 28	0.22	0.46	0	0	0.36	0.18
29	bio 29	0.42	0.09	0	0	0.33	0.15
30	bio 30	0.6	0.28	0	0	-0.01	0.18
31	bio 31	0.52	0.71	0	0	0.12	0.18
32	bio 32	0.66	0.58	0	0	0.95	0.2
33	bio 33	0.34	0.12	0	0	0.42	0.2
34	bio 34	0.66	0.47	0	0	0.4	0.2
35	bio 35	0.89	0.16	0	0	0.99	0.2
36	bio 36	0.89	0.25	0	0	0.93	0.16
37	bio 37	0.6	0.01	0	0	0.82	0.2
38	bio 38	0.68	0.61	0	0	0.87	0.18
39	bio 39	0.75	0.07	0	0	0.41	0.18
40	bio 40	0.72	0.58	0	0	0.03	0.2
41	bio 41	0.79	0.17	0	0	-0.02	0.16
42	bio 42	0.66	0.71	0	0	0.52	0.2
43	bio 43	0.51	0.57	0	0	0.33	0.2
44	bio 44	0.26	0.38	0	0	0.02	0.15
45	bio 45	0.82	0.49	0	0	0.26	0.16
46	bio 46	0.31	0.51	0	0	0.52	0.18
47	bio 47	0.73	0.07	0	0	0.35	0.2
48	bio 48	0.12	0.34	0	0	1.72	0.16
49	bio 49	0.84	0.49	0	0	0.53	0.2
50	bio 50	0.82	0.55	0	0	0.25	0.13
51	bio 51	0.7	0.59	0	0	0.54	0.1
52	bio 52	0.32	0.19	0	0	0.43	0.17
53	bio 53	0.48	0.01	0	0	0.16	0.2
54	bio 54	0.86	0.37	0	0	0.29	0.09
55	bio 55	0.34	0.84	0	0	0.68	0.15

No.	ITEM	Discrimination				Difficulty	Guessing
		Total	Bio	Chem	Phy		
56	bio 56	0.75	0.06	0	0	0.5	0.2
57	bio 57	0.71	0.32	0	0	0.74	0.03
58	bio 58	0.74	0.46	0	0	0.04	0.2
59	bio 59	0.75	0.42	0	0	0.83	0.2
60	bio 60	0.83	0.35	0	0	0.09	0.2
61	bio 61	0.82	0.36	0	0	0.78	0.16
62	bio 62	0.64	0.31	0	0	0.38	0.09
63	bio 63	0.82	0.33	0	0	0.64	0.18
64	bio 64	0.79	0.42	0	0	0.27	0.2
65	bio 65	0.86	0.34	0	0	0.4	0.2
66	bio 66	0.77	0.34	0	0	0.26	0.2
67	bio 67	0.91	0.25	0	0	1.23	0.2
68	bio 68	0.84	0.17	0	0	0.54	0.19
69	bio 69	0.84	0.28	0	0	0.53	0.15
70	bio 70	0.72	0.25	0	0	0.58	0.2
71	bio 71	0.81	0.39	0	0	0.79	0.14
72	bio 72	0.74	0.51	0	0	0.39	0.11
73	bio 73	0.81	0.54	0	0	0.29	0.14
74	bio 74	0.81	0.2	0	0	0.18	0.2
75	bio 75	0.79	0.56	0	0	0.6	0.02
76	bio 76	0.74	0.24	0	0	0.3	0.11
77	bio 77	0.85	0.29	0	0	0.85	0.1
78	chem 1	0.23	0	0.41	0	0.48	0.14
79	chem 2	0.37	0	0.18	0	1.08	0.13
80	chem 3	0.06	0	0.35	0	0.18	0.03
81	chem 4	0.36	0	0.37	0	0.08	0.2
82	chem 5	0.48	0	0.63	0	-0.39	0.2
83	chem 6	0.6	0	0.09	0	0.75	0.2
84	chem 7	0.52	0	0.03	0	0.21	0.16
85	chem 8	0.47	0	0.33	0	-0.26	0.2
86	chem 9	0.26	0	0.42	0	-0.35	0.2
87	chem 10	0.2	0	0.14	0	-0.08	0.19
88	chem 11	0.61	0	0.23	0	0.47	0.15

No.	ITEM	Discrimination				Difficulty	Guessing
		Total	Bio	Chem	Phy		
89	chem 12	0.76	0	0.17	0	-0.07	0.2
90	chem 13	0.06	0	0.25	0	0.3	0.14
91	chem 14	0.46	0	0.18	0	0.35	0.11
92	chem 15	0.62	0	0.37	0	0.28	0.14
93	chem 16	0.26	0	0.45	0	0.64	0.2
94	chem 17	0.53	0	0.3	0	0.43	0.11
95	chem 18	0.06	0	0.27	0	-0.31	0.1
96	chem 19	0.56	0	0.29	0	0.35	0.14
97	chem 20	0.32	0	0.5	0	-0.49	0.13
98	chem 21	0.06	0	0.19	0	-0.33	0.08
99	chem 22	0.24	0	0.39	0	-0.09	0.14
100	chem 23	0.32	0	0.36	0	-0.12	0.11
101	chem 24	0.62	0	0.27	0	0.44	0.14
102	chem 25	0.9	0	0.25	0	-0.08	0.2
103	chem 26	0.87	0	0.33	0	0.35	0.2
104	chem 27	0.07	0	0.34	0	-0.22	0.02
105	chem 28	0.87	0	0.22	0	-0.16	0.09
106	chem 29	0.63	0	0.56	0	-0.08	0.15
107	chem 30	0.79	0	0.14	0	-0.01	0.2
108	chem 31	0.62	0	0.39	0	0.29	0.2
109	chem 32	0.18	0	0.21	0	0.41	0.03
110	chem 33	0.81	0	0.24	0	0.13	0.2
111	chem 34	0.88	0	0.25	0	0.57	0.2
112	chem 35	0.73	0	0.5	0	-0.16	0.2
113	chem 36	0.66	0	0.47	0	0.03	0.16
114	chem 37	0.69	0	0.42	0	-0.29	0.1
115	chem 38	0.7	0	0.52	0	-0.26	0.1
116	chem 39	0.61	0	0.54	0	-0.12	0.15
117	chem 40	0.59	0	0.47	0	0.25	0.2
118	chem 41	0.76	0	0.27	0	0.16	0.17
119	chem 42	0.59	0	0.61	0	-0.15	0.16
120	chem 43	0.69	0	0.17	0	0.05	0.18
121	chem 44	0.63	0	0.57	0	0	0.2

No.	ITEM	Discrimination				Difficulty	Guessing
		Total	Bio	Chem	Phy		
122	chem 45	0.73	0	0.15	0	0.13	0.16
123	chem 46	0.79	0	0.13	0	0.56	0.2
124	chem 47	0.26	0	0.63	0	-0.13	0.13
125	chem 48	0.63	0	0.2	0	1.07	0.1
126	chem 49	0.7	0	0.57	0	0.83	0.17
127	chem 50	0.16	0	0.4	0	0.2	0.2
128	chem 51	0.34	0	0.6	0	-0.2	0.09
129	chem 52	0.05	0	0.52	0	0.46	0.15
130	chem 53	0.44	0	0.65	0	0.27	0.2
131	chem 54	0.1	0	0.71	0	-0.06	0.2
132	chem 55	0.84	0	0.14	0	-0.12	0.03
133	chem 56	0.44	0	0.76	0	-1.25	0.2
134	chem 57	0.21	0	0.1	0	-0.1	0.2
135	phy 1	0.65	0	0	0.3	-0.23	0.2
136	phy 2	0.4	0	0	0.26	-0.22	0.16
137	phy 3	0.49	0	0	0.54	-0.23	0.09
138	phy 4	0.1	0	0	0.5	-0.55	0.18
139	phy 5	0.1	0	0	0.5	-0.55	0.14
140	phy 6	0.41	0	0	0.5	-0.35	0.2
141	phy 7	0.46	0	0	0.79	-0.27	0.03
142	phy 8	0.34	0	0	0.02	-0.39	0.14
143	phy 9	0.84	0	0	0.21	-0.54	0.15
144	phy 10	0.39	0	0	0.37	-0.16	0.2
145	phy 11	0.81	0	0	0.3	-0.28	0.2
146	phy 12	0.22	0	0	0.38	-0.39	0.2
147	phy 13	0.03	0	0	0.42	-0.41	0.14
148	phy 14	0.67	0	0	0.43	-0.23	0.14
149	phy 15	0.67	0	0	0.43	-0.24	0.15
150	phy 16	0.14	0	0	0.42	-0.32	0.13
151	phy 17	0.38	0	0	0.08	0.65	0.2
152	phy 18	0.76	0	0	0.55	-0.12	0.1
153	phy 19	0.17	0	0	0.52	1	0.2
154	phy 20	0.68	0	0	0.44	0.79	0.2

No.	ITEM	Discrimination				Difficulty	Guessing
		Total	Bio	Chem	Phy		
155	phy 21	0.39	0	0	0.5	1.47	0.14
156	phy 22	0.64	0	0	0.7	0.64	0.14
157	phy 23	0.84	0	0	0.48	1.03	0.15
158	phy 24	0.33	0	0	0.69	0.52	0.1
159	phy 25	0.3	0	0	0.57	0.39	0.2
160	phy 26	0.67	0	0	0.61	0.39	0.2
161	phy 27	0.75	0	0	0.3	-0.26	0.14
162	phy 28	0.62	0	0	0.2	-0.19	0.14
163	phy 29	0.72	0	0	0.26	-0.2	0.11
164	phy 30	0.79	0	0	0.11	-0.46	0.1
165	phy 31	0.63	0	0	0.06	-0.05	0.14
166	phy 32	0.78	0	0	0.33	-0.25	0.13
167	phy 33	0.63	0	0	0.18	-0.2	0.08
168	phy 34	0.8	0	0	0.24	-0.23	0.1
169	phy 35	0.72	0	0	0.4	-0.26	0.11
170	phy 36	0.78	0	0	0.09	-0.4	0.1

ภาคผนวก ง

ผลการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์

ตัวอย่างผลการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์ เสนอไว้ดังตาราง ที่ 6.6 โดยนำเสนอผลสอบของนักเรียน 5 คน คะแนนดิบและจำนวนข้อสอบที่ได้รับ รวมถึงผลการประมาณค่าความสามารถแยกตามมิติความสามารถและค่าความสามารถรวม แ ละค่าความคลาดเคลื่อนมาตรฐาน (SE)

ตารางที่ 6.6 ค่าความสามารถของนักเรียนแต่ละคนแยกตามวิชาและคะแนนรวม

ผู้สอบ	คะแนนที่ได้ และ (จำนวนข้อสอบ)				ชีวะ (θ)	เคมี (θ)	ฟิสิกส์ (θ)	รวม(θ)	SE
	ชีวะ	เคมี	ฟิสิกส์	รวม					
คนที่ 1	7 (10)	2 (8)	5 (12)	14 (30)	0.994	-1.329	-0.717	-0.202	0.037
คนที่ 2	7 (10)	3 (10)	3 (10)	13 (30)	0.884	-0.991	-0.256	-0.456	0.040
คนที่ 3	6 (10)	1 (8)	3 (12)	10 (30)	0.845	-1.173	-0.445	-1.011	0.041
คนที่ 4	2 (10)	1 (8)	3 (12)	6 (30)	-0.424	-1.193	-1.000	-1.673	0.064
คนที่ 5	4 (9)	1 (8)	6 (13)	11 (30)	-0.049	-1.270	-0.445	-0.666	0.044

จากตาราง 6.6 ข้างต้น จะนำผู้สอบคนที่ 1 มาเสนอให้เห็นถึงลักษณะการทำงานของโปรแกรมและค่า ต่างๆ ที่เปลี่ยนแปลงไปในระหว่างการทดสอบ โดย item ID คือ ข้อสอบที่ถูกคัดเลือกขึ้นมาในโปรแกร ม FI คือค่า Fisher information ที่ใช้พิจารณาคัดเลือกข้อสอบขึ้นมา โดยในที่นี่จะแสดงเฉพาะข้อสอบที่ถูกคัดเลือกขึ้นมาแล้ว ซึ่งมีค่าสูงสุดเมื่อเทียบกับข้อสอบที่เหลือในคลังข้อสอบ สำหรับ I(θ) สะสม คือ ค่า Fisher information สะสม และจะกลาย เป็นค่าสารสนเทศของแบบสอบที่ จะบวกสะสมไปเรื่อยๆ เมื่อทำข้อสอบข้อถัดไป ค่า SE คือ ความคลาดเคลื่อนมาตรฐานในการวัด สำหรับค่า θ ในแต่ละวิชา คือ ความสามารถของผู้สอบแต่ละวิชา ซึ่งจะมีค่าปรับเปลี่ยนไปตามการตอบของผู้สอบ รายละเอียดดังตารางที่ 6.7

ตารางที่ 6.7 การคัดเลือกข้อสอบและการประมาณค่าความสามารถของผู้สอบ 1 คน

Item ID	วิชา	เฉลย	การตอบ	คะแนน	FI	I (θ) สะสม	SE	ชีววิทยา (θ)	เคมี (θ)	ฟิสิกส์ (θ)	รวม(θ)
152	ฟิสิกส์	3	3	1	1.519	1.519	0.811	0	0	0	0
50	ชีววิทยา	2	3	0	2.112	3.631	0.525	0	0	0.236	0.323
119	เคมี	1	2	0	2.714	6.345	0.397	-0.280	0	0.297	-0.010
141	ฟิสิกส์	2	3	0	3.353	9.698	0.321	-0.224	-0.416	0.345	-0.268
15	ชีววิทยา	2	2	1	4.161	13.859	0.269	-0.193	-0.384	-0.207	-0.430
115	เคมี	2	2	1	5.664	19.522	0.226	-0.029	-0.428	-0.268	-0.204
18	ชีววิทยา	3	3	1	7.124	26.646	0.194	-0.091	-0.216	-0.317	-0.028
131	เคมี	4	1	0	8.717	35.363	0.168	0.119	-0.244	-0.342	0.057
102	เคมี	1	3	0	10.21	45.570	0.148	0.105	-0.561	-0.352	0.090
31	ชีววิทยา	1	1	1	11.51	57.079	0.132	0.216	-0.592	-0.278	-0.165
55	ชีววิทยา	2	2	1	13.59	70.666	0.119	0.462	-0.607	-0.290	-0.124
82	เคมี	4	3	0	15.95	86.612	0.107	0.813	-0.602	-0.286	-0.138
145	ฟิสิกส์	3	4	0	17.14	103.75	0.098	0.845	-0.840	-0.267	-0.207
133	เคมี	2	3	0	18.57	122.32	0.090	0.941	-0.762	-0.357	-0.411
143	ฟิสิกส์	2	3	0	18.85	141.17	0.084	0.973	-1.209	-0.334	-0.480
1	ชีววิทยา	1	1	1	19.16	160.36	0.079	1.076	-1.136	-0.370	-0.698
75	ชีววิทยา	3	2	0	23.06	183.40	0.074	1.109	-1.204	-0.448	-0.496
137	ฟิสิกส์	3	2	0	24.94	208.34	0.069	1.025	-1.184	-0.425	-0.553
73	ชีววิทยา	4	4	1	25.52	233.86	0.065	1.053	-1.168	-0.610	-0.601
42	ชีววิทยา	2	2	1	29.45	263.31	0.062	1.173	-1.200	-0.650	-0.506
166	ฟิสิกส์	2	1	0	32.76	296.07	0.058	1.333	-1.212	-0.665	-0.473
138	ฟิสิกส์	4	4	1	33.37	32944	0.055	1.396	-1.178	-0.735	-0.569
164	ฟิสิกส์	2	2	1	37.74	367.0	0.052	1.414	-1.170	-0.525	-0.595
51	ชีววิทยา	4	2	0	43.13	410.31	0.049	1.331	-1.213	-0.547	-0.469
139	ฟิสิกส์	4	3	0	45.59	455.90	0.047	1.213	-1.198	-0.523	-0.513
170	ฟิสิกส์	1	1	1	47.20	50305	0.045	1.193	-1.208	-0.743	-0.485
105	เคมี	3	3	1	53.25	556.35	0.042	1.111	-1.250	-0.771	-0.368
112	เคมี	4	3	0	60.53	616.88	0.040	1.016	-1.212	-0.840	-0.234
169	ฟิสิกส์	3	3	1	62.52	679.05	0.038	1.048	-1.294	-0.817	-0.278
140	ฟิสิกส์	3	2	0	70.15	749.55	0.037	0.994	-1.329	-0.717	-0.202

ค่าความสามารถของนักเรียนแต่ละคนแยกตามวิชาและคะแนนรวม

ค่าความสามารถของนักเรียนที่เป็นกลุ่มตัวอย่าง 74 คน ผู้วิจัยนำเสนอให้เห็นจำนวนข้อสอบที่ผู้สอบได้รับในแต่ละวิชาพร้อมคะแนนดิบรายวิชาและภาพรวม รวมถึงค่าความสามารถที่ประมาณค่าได้จากโปรแกรมการทดสอบปรับเหมาะแบบพหุมิติ โดยตารางที่ 6.8 แสดงค่าความสามารถ ของนักเรียนที่ทำข้อสอบแบบผสมผสานมิติจำนวน 37 คน สำหรับตารางที่ 6.9 แสดงค่าความสามารถ ของนักเรียนที่ทำข้อสอบแบบจำแนกมิติจำนวน 37 คน ซึ่งมีรายละเอียดดังนี้

ตารางที่ 6.8 ค่าความสามารถของนักเรียนแต่ละคนแยกตามวิชาและคะแนนรวม

(กรณีลำดับข้อสอบแบบผสมผสานมิติ)

ผู้สอบคนที่	คะแนนที่ได้ และ (จำนวนข้อสอบ)				ชีวะ (θ)	เคมี (θ)	ฟิสิกส์ (θ)	รวม(θ)	SE
	ชีวะ	เคมี	ฟิสิกส์	รวม					
1	6 (11)	0 (6)	3 (13)	9 (30)	0.6594	-1.306	-0.6178	-1.5624	0.0574
2	6 (10)	1 (8)	3 (12)	10 (30)	0.8452	-1.1733	-0.4455	-1.0109	0.0412
3	2 (10)	1 (8)	3 (12)	6 (30)	-0.4235	-1.1934	-1.0001	-1.6731	0.0639
4	4 (9)	1 (8)	6 (13)	11 (30)	-0.0491	-1.2697	-0.4449	-0.6664	0.0442
5	3 (8)	3 (10)	5 (12)	11 (30)	-0.6716	-1.3465	-1.1838	-0.1913	0.0403
6	7 (10)	2 (8)	5 (12)	14 (30)	0.9941	-1.3294	-0.7172	-0.2025	0.0365
7	5 (9)	3 (11)	3 (10)	11 (30)	0.1772	-0.5193	-0.6433	-1.3427	0.0445
8	5 (9)	5 (11)	3 (10)	13 (30)	0.4154	-0.2203	-0.8159	-0.6247	0.036
9	6 (12)	0 (5)	3 (13)	9 (30)	0.0828	-1.4182	-0.9648	-1.2343	0.0579
10	2 (8)	0 (5)	9 (17)	11 (30)	-0.2953	-1.4239	-0.0457	-1.2572	0.046
11	4 (8)	6 (13)	0 (9)	10 (30)	0.1757	0.1075	-1.3998	-1.6385	0.0539
12	7 (13)	1 (7)	1 (10)	9 (30)	0.8774	-1.103	-1.3616	-1.1455	0.0517
13	1 (6)	4 (11)	6 (13)	11 (30)	-0.6714	-0.5517	-0.1946	-0.7665	0.0454
14	1 (6)	4 (12)	3 (12)	8 (30)	-0.4999	-0.4849	-0.6204	-1.6074	0.0485
15	6 (14)	0 (5)	1 (11)	7 (30)	0.5248	-1.3032	-1.2142	-1.7777	0.0657
16	3 (9)	2 (7)	4 (14)	9 (30)	0.043	-0.6846	-0.6092	-1.8238	0.0575
17	5 (9)	2 (8)	5 (13)	12 (30)	0.256	-1.1092	-0.0112	-0.8217	0.0393
18	4 (9)	6 (12)	3 (9)	13 (30)	0.1034	0.1451	-0.5875	-0.615	0.0358
19	2 (7)	3 (9)	6 (14)	11 (30)	-0.2658	-0.5044	-0.1454	-0.894	0.0436
20	3 (10)	0 (5)	4 (15)	7 (30)	-0.2306	-1.3802	-0.5068	-1.5047	0.054

ผู้สอบ คนที่	คะแนนที่ได้ และ (จำนวนข้อสอบ)				ชีวะ (θ)	เคมี (θ)	ฟิสิกส์ (θ)	รวม(θ)	SE
	ชีวะ	เคมี	ฟิสิกส์	รวม					
21	1 (6)	1 (7)	7 (17)	9 (30)	-0.4864	-1.1612	0.1211	-1.5302	0.0474
22	4 (11)	0 (6)	4 (13)	8 (30)	-0.0879	-1.5523	-0.7645	-1.0707	0.0491
23	4 (10)	1 (7)	3 (13)	8 (30)	0.0832	-1.1297	-0.7181	-1.5509	0.0484
24	4 (9)	3 (11)	1 (10)	8 (30)	0.2956	-0.7883	-1.038	-1.3254	0.0441
25	1 (8)	2 (9)	3 (13)	6 (30)	-0.5852	-0.8679	-0.8854	-1.5983	0.0611
26	4 (11)	3 (8)	1 (11)	8 (30)	0.2515	-0.8038	-1.1899	-1.7841	0.0598
27	7 (11)	0 (6)	4 (13)	11 (30)	0.7962	-1.4143	-0.439	-0.678	0.0399
28	4 (11)	2 (8)	3 (11)	9 (30)	-0.0164	-0.7594	-1.1297	-1.546	0.0561
29	2 (7)	4 (11)	3 (12)	9 (30)	-0.0619	-0.0885	-0.5257	-1.9524	0.0557
30	5 (11)	0 (6)	3 (13)	8 (30)	0.1143	-1.4914	-1.0617	-1.1793	0.051
31	2 (10)	3 (9)	1 (11)	6 (30)	-0.4007	-0.9741	-1.2396	-1.9779	0.0743
32	3 (11)	0 (7)	1 (12)	4 (30)	-0.0861	-1.464	-1.2668	-2.0252	0.0746
33	4 (12)	0 (6)	1 (12)	5 (30)	0.0764	-1.4322	-1.4233	-1.6573	0.0586
34	3 (10)	0 (7)	3 (13)	6 (30)	-0.1066	-1.4171	-0.9043	-1.9742	0.0699
35	7 (13)	1 (7)	1 (10)	9 (30)	0.3529	-1.2241	-1.2134	-1.351	0.0612
36	6 (11)	1 (10)	3 (9)	10 (30)	0.3208	-0.6817	-1.0973	-1.1319	0.0526
37	8 (11)	4 (11)	2 (8)	14 (30)	0.8849	-0.8024	-0.4481	-0.7322	0.0376

ตารางที่ 6.9 ค่าความสามารถของนักเรียนแต่ละคนแยกตามวิชาและคะแนนรวม (กรณีลำดับข้อสอบแบบจำแนกมิติ)

ผู้สอบ คนที่	คะแนนที่ได้ และ (จำนวนข้อสอบ)				ชีวะ (θ)	เคมี (θ)	ฟิสิกส์ (θ)	รวม(θ)	SE
	ชีวะ	เคมี	ฟิสิกส์	รวม					
38	6 (10)	2 (10)	3 (10)	11 (30)	0.4538	-0.8398	-1.2015	-0.6204	0.0421
39	3 (10)	3 (10)	2 (10)	8 (30)	0.2532	-1.1072	-0.7675	-1.6534	0.0679
40	6 (10)	4 (10)	1 (10)	11 (30)	1.5263	-0.6312	-0.6711	-1.0018	0.0473
41	5 (10)	4 (10)	1 (10)	10 (30)	0.1713	-1.337	-0.9783	-0.6802	0.0431
42	7 (10)	3 (10)	3 (10)	13 (30)	0.884	-0.9914	-0.2562	-0.4556	0.0396
43	4 (10)	3 (10)	4 (10)	11 (30)	-0.2896	-1.1264	-0.249	-0.8289	0.0462
44	6 (10)	1 (10)	4 (10)	11 (30)	0.3697	-1.4675	-0.8578	-0.3316	0.0409
45	4 (10)	2 (10)	3 (10)	9 (30)	0.0423	-1.1159	-0.5696	-1.1872	0.0514
46	5 (10)	4 (10)	4 (10)	13 (30)	0.3781	-0.3499	-0.3679	-0.7497	0.0428

ผู้สอบคนที่	คะแนนที่ได้ และ (จำนวนข้อสอบ)				ชีวะ (θ)	เคมี (θ)	ฟิสิกส์ (θ)	รวม(θ)	SE
	ชีวะ	เคมี	ฟิสิกส์	รวม					
47	5 (10)	3 (10)	4 (10)	12 (30)	0.6375	-0.6743	0.2015	-0.6177	0.0383
48	4 (10)	1 (10)	2 (10)	7 (30)	0.219	-1.6111	-1.2279	-0.9894	0.0547
49	4 (10)	4 (10)	2 (10)	10 (30)	0.2265	-0.4472	-1.0392	-1.0561	0.0512
50	4 (10)	3 (10)	2 (10)	9 (30)	0.4296	-0.5543	-0.9477	-1.4774	0.0543
51	3 (10)	1 (10)	3 (10)	7 (30)	0.0299	-1.7292	-0.2262	-1.7673	0.0665
52	4 (10)	4 (10)	3 (10)	11 (30)	0.0437	-0.757	-0.5274	-1.1554	0.0481
53	4 (10)	4 (10)	2 (10)	10 (30)	0.2661	-0.5811	-0.6081	-1.2682	0.0586
54	3 (10)	3 (10)	3 (10)	9 (30)	-0.2023	-0.634	-0.6087	-1.5618	0.0628
55	6 (10)	3 (10)	2 (10)	11 (30)	0.8462	-1.0494	-1.2915	-0.9646	0.045
56	4 (10)	3 (10)	1 (10)	8 (30)	0.3222	-0.9355	-1.3205	-1.061	0.0599
57	5 (10)	4 (10)	3 (10)	12 (30)	0.293	-0.4992	-0.9569	-0.8491	0.044
58	2 (10)	2 (10)	2 (10)	6 (30)	-0.4646	-1.0064	-1.1405	-1.6916	0.0788
59	7 (10)	1 (10)	3 (10)	11 (30)	1.333	-1.6642	-0.2622	-0.6925	0.0392
60	4 (10)	2 (10)	0 (10)	6 (30)	0.5365	-1.066	-1.5977	-1.4543	0.0644
61	1 (10)	4 (10)	8 (10)	13 (30)	-0.9811	-0.512	1.0526	-0.7098	0.0479
62	5 (10)	4 (10)	3 (10)	12 (30)	0.0123	-0.3252	-0.8079	-0.3455	0.0404
63	4 (10)	6 (10)	3 (10)	13 (30)	0.1446	0.08	-0.2451	-1.5743	0.0577
64	2 (10)	2 (10)	1 (10)	5 (30)	-0.2782	-0.8573	-1.1138	-2.0875	0.0888
65	6 (10)	4 (10)	4 (10)	14 (30)	0.3646	-1.0623	-0.4967	0.023	0.0371
66	4 (10)	4 (10)	2 (10)	10 (30)	-0.0567	-0.867	-1.089	-0.9231	0.0519
67	6 (10)	3 (10)	2 (10)	11 (30)	0.1825	-0.9109	-1.2685	-0.4893	0.0459
68	3 (10)	3 (10)	3 (10)	9 (30)	-0.1073	-0.6383	-0.4222	-1.3814	0.0538
69	3 (10)	4 (10)	3 (10)	10 (30)	-0.5666	-0.1361	-0.5934	-1.3013	0.0572
70	3 (10)	1 (10)	2 (10)	6 (30)	-0.0553	-1.44	-0.7262	-1.4528	0.0603
71	3 (10)	3 (10)	6 (10)	12 (30)	-0.6582	-1.1281	-0.4109	-0.3956	0.0459
72	4 (10)	2 (10)	4 (10)	10 (30)	0.6533	-0.8917	0.1115	-1.7517	0.0506
73	6 (10)	3 (10)	2 (10)	11 (30)	0.5897	-0.8979	-1.3259	-0.471	0.0436
74	5 (10)	2 (10)	3 (10)	10 (30)	0.0868	-0.9167	-1.1694	-0.5115	0.0386

ภาคผนวก จ

เครื่องมือที่ใช้ในการวิจัย

1. แบบทดสอบวัดความสามารถด้านวิทยาศาสตร์ระดับมัธยมศึกษาตอนปลาย
2. โปรแกรมการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิเตอร์ (คู่มือ)
3. แบบประเมินโปรแกรมการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์
4. แบบวัดความสามารถด้านคอมพิวเตอร์ พัฒนาโดย ชุติวัฒน์ สุวัตถิพงศ์ (2551)

แบบทดสอบวัดความสามารถด้านวิทยาศาสตร์ระดับมัธยมศึกษาตอนปลาย

ตัวอย่างข้อสอบวิชาชีววิทยา

1. สิ่งมีชีวิตข้อใดแสดงพฤติกรรมทางสังคม (social behavior) ถูกต้องมากที่สุด

   A. ลักษณะการเต้นของแมลงวันตัวผู้ก่อนเข้าผสมพันธุ์กับตัวเมีย
   B. ฝูงแมลงหวี่ที่มารุมตอมผลไม้ที่กำลังสุกงอม
   C. ฝูงผึ้งงานออกไปหาอาหารหลังจากรับรู้การเต้นของผึ้งงาน
   D. ฝูงแมลงดานาบินเข้าหาแสงไฟที่นำมาล่อเอาไว้
   F. ฝูงหิ่งห้อยกระพริบแสงเรียกเพศตรงข้ามให้มาผสมพันธุ์

   1. A B และ C
   2. A C และ F
   3. B C และ F
   4. C D และ F

2. ข้อใดกล่าวถึงความสัมพันธ์ของสิ่งมีชีวิตที่อยู่ในระบบนิเวศได้ถูกต้องที่สุด

	ชนิดของสิ่งมีชีวิต	ความสัมพันธ์ในระบบนิเวศ
1.	รากับสาหร่าย	การได้ประโยชน์ร่วมกัน (Protocooperation: +,+)
2.	นกทำรังบนต้นไม้	ภาวะพึ่งพา (Mutualism: +,+)
3.	ฉลามกับเหาฉลาม	ภาวะอิงอาศัย (Commensalism: +,0)
4.	เหยี่ยวกับงู	ภาวะปรสิต (Parasitism : +,-)

3. ความสัมพันธ์ของสิ่งมีชีวิตในข้อใดมีลักษณะเช่นเดียวกับปลวกและโพโทซัว

   1. ฉลามกับเหาฉลาม สาหร่ายและปะการัง
   2. ปลาเล็กกับฟองน้ำ พืชขนาดเล็กกับต้นไม้ใหญ่
   3. พืชตระกูลถั่วและแบคทีเรียในปนรากถั่ว ราและสาหร่าย
   4. นกทำรังบนต้นไม้ แมลงกับดอกไม้

ตัวอย่างข้อสอบวิชาเคมี

4. เกลือ A เมื่อเผาสลายตัวเป็นก๊าซหมด เกลือ B เมื่อเผาเปลี่ยนจากสีฟ้าเป็นสีขาว ไอที่ออกมาเป็นกลางต่อกระดาษลิตมัส เกลือ C เมื่อเผาสลายตัวให้ก๊าซที่เปลี่ยนสีกระดาษลิตมัสจากน้ำเงินเป็นแดง เกลือ D เมื่อเผาสลายตัวให้ก๊าซกลิ่นฉุน เกลือทั้ง 4 เรียงลำดับตามข้อใด

   1. $(NH_4)_2CO_3$, $CuSO_4 \cdot 5H_2O$, $CaSO_3$, $CaSO_4$
   2. NH_4NO_3 , $CuSO_4 \cdot 5H_2O$, $CaSO_3$, $CaCO_3$
   3. NH_4Cl , $CuSO_4 \cdot 5H_2O$, $CaCO_3$, FeS
   4. NH_4Cl , $CuSO_4 \cdot 5H_2O$, $CaCO_3$, $CaSO_3$

5. สารคู่ใดต่อไปนี้มีจำนวนโมเลกุลเท่ากัน

   1. C_2H_5OH 23 กรัม กับ CCl_4 77 กรัม
   2. CH_3COOH 30 กรัม กับ H_2O 18 กรัม
   3. CaO 56 กรัม กับ C_2H_5OH 23 กรัม
   4. CCl_4 77 กรัม กับ H_2O 18 กรัม

6. ในการศึกษาสารประกอบชนิดหนึ่ง เมื่อนำสารหนัก 1.05 กรัม มาเผาไหม้จนสมบูรณ์พบว่าเกิดก๊าซ CO_2 1.32 กรัม และ H_2O 0.63 กรัม และพบว่าสารนี้หนัก 0.90 กรัม มีไนโตรเจน 0.12 กรัม จงหาสูตรอย่างง่ายของสารประกอบชนิดนี้

   1. C_3H_7N
   2. $C_3H_7NO_3$
   3. $C_4H_7NO_5$
   4. $C_6H_{14}O_6N_2$

ตัวอย่างข้อสอบวิชาฟิสิกส์

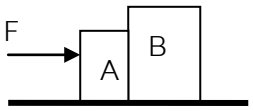
7. นักเรียนคนหนึ่งวัดความสูง และความยาวฐานของสามเหลี่ยมรูปหนึ่งได้ 13.07 เซนติเมตร และ 4.3 เซนติเมตรตามลำดับ เขาควรบันทึกพื้นที่ของสามเหลี่ยมรูปนี้เป็นกี่ตารางเซนติเมตร

   1. 28.1
   2. 28.10
   3. 28.101
   4. 28.1005

8. ชายคนหนึ่งขับรถไปทางทิศเหนือระยะทาง 6 กิโลเมตร ใช้เวลา 10 นาที แล้วหยุดพักเป็นเวลา 20 นาทีจากนั้นขับรถต่อไปทางทิศตะวันออกอีก 8 กิโลเมตร ใช้เวลา 15นาที ความเร็วเฉลี่ยในการขับรถของชายคนนี้เป็นเท่าไร

   1. 13.33 กิโลเมตรต่อชั่วโมง ไปทางทิศตะวันออกเฉียงเหนือ
   2. 18.66 กิโลเมตรต่อชั่วโมง ไปทางทิศตะวันออกเฉียงเหนือ
   3. 23.81 กิโลเมตรต่อชั่วโมง ไปทางทิศตะวันออก
   4. 33.33 กิโลเมตรต่อชั่วโมง ไปทางทิศตะวันออก

9. ออกแรง F ขนานกับพื้นราบ กระทำกับกล่อง A และ B ที่วางติดกัน โดยที่ $m_A = \frac{1}{2}m_B$ ดังรูป



ข้อใดกล่าวถูกต้อง

   1. ความเร่งของ B เป็น 2 เท่าของ A
   2. แรงที่กล่อง A กระทำกับ B มีขนาดเท่ากับ แรงที่ B กระทำกับ A
   3. แรงที่กล่อง A กระทำกับ B มีขนาดเป็น 2 เท่าของแรงที่ B กระทำกับ A
   4. แรงลัพธ์ที่กระทำกับกล่อง A มีขนาดเท่ากับแรงลัพธ์ที่กระทำกับกล่อง B

คู่มือการใช้โปรแกรมการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์

โดย

นางสาวทัศน์ศิรินทร์ สว่างบุญ

รองศาสตราจารย์ ดร.ศิริเดช สุชีวะ	อาจารย์ที่ปรึกษา
ศาสตราจารย์ ดร.ศิริชัย กาญจนวาสี	อาจารย์ที่ปรึกษาร่วม
ศาสตราจารย์ ดร.เออิจิ มูราคิ	อาจารย์ที่ปรึกษาร่วม

โปรแกรมการทดสอบนี้เป็นส่วนหนึ่งของวิทยานิพนธ์หลักสูตรปริญญาครุศาสตรดุษฎีบัณฑิต

สาขาวิชาการวัดและประเมินการศึกษา ภาควิชาวิจัยและจิตวิทยาการศึกษา

คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

คู่มือการใช้โปรแกรมการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์

คำนำ

โปรแกรมการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์ เพื่อวัดความสามารถด้านวิทยาศาสตร์ของนักเรียนชั้นมัธยมศึกษาตอนปลายครั้งนี้ เป็นส่วนหนึ่งของวิทยานิพนธ์ระดับปริญญาเอก สาขาการวัดและประเมินผลการศึกษา ภาควิชาวิจัยและจิตวิทยาการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย เรื่อง "การเปรียบเทียบวิธีการคัดเลือก ข้อสอบขั้นแรกและลำดับข้อสอบที่มีต่อคุณภาพของการทดสอบปรับเหมาะแบบพหุมิติ ด้วยคอมพิวเตอร์" โดยมีนิสิตผู้ทำวิจัยคือ นางสาวทัศน์ศิรินทร์ สว่างบุญ ภายใต้การควบคุมดูแลของอาจารย์ที่ปรึกษาคือ รองศาสตราจารย์ ดร.ศิริเดช สุชีวะ และอาจารย์ที่ปรึกษาร่วม คือ ศาสตราจารย์ ดร.ศิริชัย กาญจนวาสี และ Prof.Dr. Eiji Muraki

การพัฒนาโปรแกรมขึ้นมาครั้งนี้ มีวัตถุประสงค์ เพื่อทดลองใช้โปรแกรมกับนักเรียนและศึกษาประสิทธิภาพการวัดของการทดสอบปรับเหมาะแบบพหุมิติ รวมถึงให้นักเรียนผู้สอบได้แสดงความคิดเห็นและข้อเสนอแนะที่มีต่อการทดสอบรูปแบบใหม่ นี้ เพื่อผู้วิจัยจะได้นำข้อเสนอแนะเหล่านั้น มาปรับปรุงพัฒนาโปรแกรมการทดสอบให้มีความเหมาะสมมากยิ่งขึ้น นอกจากนี้นักเรียนยังได้ทราบถึงระดับความสามารถของตนในด้านวิทยาศาสตร์ และสามารถนำผลการทดสอบไปพัฒนาทักษะความสามารถในรายวิชาที่มีคะแนนต่ำ เพื่อใช้เป็นแนวทางในการเตรียมตัวสอบเข้าเรียนในระดับที่สูงขึ้น

ผู้วิจัยหวังเป็นอย่างยิ่งว่า นักเรียนผู้ใช้โปรแกรมจะได้รับข้อมูลผลการทดสอบ ที่เป็นแนวทางในการพัฒนาตนเองไม่มากก็น้อย และหวังเป็นอย่างยิ่งว่า จะนำข้อเสนอะแนะที่ได้รับจากนักเรียนทุกคนมาปรับปรุงพัฒนาโปรแกรมให้มีความเหมาะสมยิ่งๆ ขึ้นไป

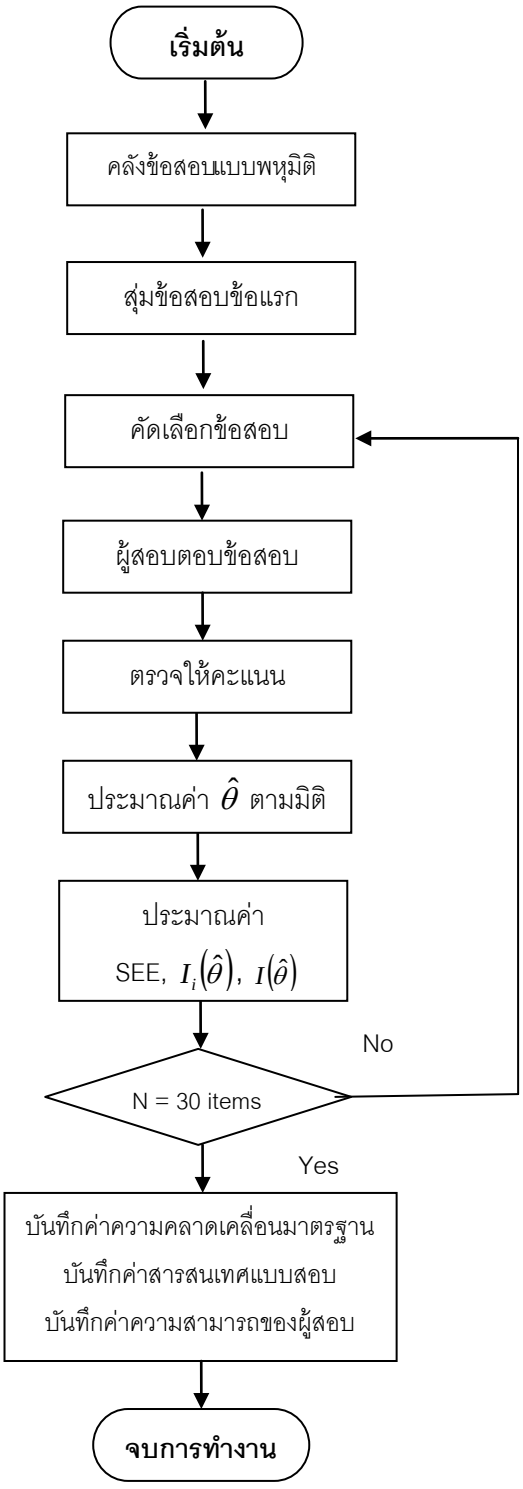
นางสาวทัศน์ศิรินทร์ สว่างบุญ
ผู้วิจัย

1. การทดสอบแบบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์และการพัฒนาข้อสอบ

หลักการของการทดสอบแบบปรับเหมาะ (Computerized Adaptive Testing: CAT) คือ การทดสอบที่คัดเลือกข้อสอบให้กับผู้สอบตามระดับความสามารถของผู้สอบ และพิจารณาข้อสอบข้อถัดไปให้เหมาะสมกับผู้ตอบ โดยนำผลการตอบข้อก่อนหน้ามาพิจารณาว่าจะเลือกข้อสอบที่ยากขึ้นและง่ายลง เพื่อให้ผู้สอบได้รับข้อสอบที่เหมาะสมกับความสามารถและสามารถวัดระดับความสามารถของผู้สอบได้อย่างแม่นยำ แม้ผู้สอบจะไม่ได้ทำข้อสอบทุกข้อในคลังข้อสอบ วิธีการดังกล่าวเป็นวิธีการทดสอบแนวใหม่ ที่เข้ามาแทนที่การสอบแบบเขียนตอบ (paper-and-pencil) โดยเป็นการเพิ่มความแม่นยำในการวัด ลดระยะเวลาในการทดสอบ วิธีใช้ได้มาตรฐาน และมีความยืดหยุ่นในการจัดตารางการสอบ การทดสอบแบบปรับเหมาะส่วนใหญ่จะ ใช้การเลือกข้อคำถามและการคำนวณการให้คะแนนโดยใช้พื้นฐานของทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT)

การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์สำหรับโมเดลการตอบสนองข้อสอบแบบพหุมิติ (Multidimensional Computerized Adaptive Testing :MCAT) เป็นการทดสอบแบบใหม่ที่เป็นการผสมผสานวิธีการของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (CAT) และ โมเดลการตอบสนองข้อสอบแบบพหุมิติ(MIRT) ซึ่งเป็นวิธีการที่มีความแม่นยำในการประมาณค่าความสามารถและมีประสิทธิภาพมากกว่าวิธีการทดสอบปรับเหมาะแบบเอกมิติ สำหรับขั้นตอนของการทดสอบก็มีความคล้ายคลึงกับการทดสอบแบบปรับเหมาะแบบเอกมิติ จะมีความแตกต่างในรายละเอียดของการคัดเลือกข้อสอบและการประมาณค่าความสามารถของผู้สอบ

ข้อสอบที่นำมาใช้ในการทดสอบความสามารถด้านวิทยาศาสตร์สำหรับนักเรียนชั้นมัธยมศึกษาตอนปลายครั้งนี้ พัฒนาโดยคุณครูที่สอนวิทยาศาสตร์ (ชีววิทยา เคมี และฟิสิกส์) ระดับชั้นมัธยมศึกษาตอนปลาย จำนวน 9 คน และผู้ทรงคุณวุฒิตรวจสอบความถูกต้อง 3 ท่าน โดยพัฒนาข้อสอบขึ้นมาทั้งสิ้น 324 ข้อ ผลการวิเคราะห์คุณภาพข้อสอบพบว่า มีข้อสอบที่ตรงตามเงื่อนไข จำนวน 213 ข้อ ทั้งนี้เนื่องจากปัญหาของโปรแกรมที่ต้องคำนวณค่า สำหรับใช้คัดเลือกข้อสอบจำนวนมาก จึงจำเป็นต้องคัดข้อสอบออก โดยพิจารณาข้อสอบที่มีค่าอำนาจจำแนกต่ำ ทำให้เหลือข้อสอบที่จะนำไปเก็บข้อมูล จำนวน 170 ข้อ แบ่งเป็นชีววิทยา 77 ข้อ เคมี 57 ข้อ และฟิสิกส์ 36 ข้อ โดยมีค่าอำนาจจำแนกอยู่ระหว่าง 0.01 ถึง 0.91 ค่าความยากอยู่ระหว่าง -1.25 ถึง 1.72 และค่าโอกาสการเดาอยู่ระหว่าง 0.02 ถึง 0.20 โดยวิเคราะห์ด้วยโปรแกรม TESTFACT และ BILOG-MG อย่างไรก็ตาม ผลการวิเคราะห์ด้วยโปรแกรม TESTFACT แบบพหุมิติให้ค่าพารามิเตอร์อำนาจจำแนกค่อนข้างต่ำ จึงควรระมัดระวังในการตีความ



ผังแสดงขั้นตอนการทำงานของโปรแกรมการทดสอบปรับเหมาะแบบพหุมิติ

2. ขั้นตอนการเตรียมข้อมูลสำหรับการทดสอบ

2.1 การเตรียมข้อสอบ

สำหรับการเตรียมข้อสอบเพื่อนำมาใช้ในการทดสอบปรับเหมาะแบบพหุมิตินั้น ผู้ใช้ต้องวางแผนการพัฒนาข้อสอบ เพื่อให้ข้อสอบสามารถวัดความสาม ารถได้แบบตามมิติที่กำหนดไว้ พิจารณารูปแบบของข้อสอบว่าเป็นข้อสอบพหุมิติในลักษณะใด รูปแบบพหุมิติภายในข้อ (within-item multidimensionality) คือ ข้อสอบ 1 ข้อสามารถวัดความสามารถได้มากกว่า 1 มิติ หรือรูปแบบพหุมิติระหว่างข้อ (between-item multidimensionality) คือข้อสอบ 1 ข้อวัดความสามารถเพียงมิติเดียว แต่การทดสอบครั้งเดียววัดความสามารถหลายมิติไปพร้อมกัน และมิติที่มุ่งวัดนั้นต้องมีความสัมพันธ์กัน และหากผู้ใช้ไม่มั่นใจว่าข้อสอบนั้นมีรูปแบบตรงตามทฤษฎีการตอบสนองข้อสอบแบบพหุมิติหรือไม่ หรือเป็นข้อสอบพหุมิติแบบใด ควรจะ นำข้อสอบไปวิเคราะห์ความเป็นเอกมิติ รวมถึงวิเคราะห์เชิงสำรวจว่าข้อสอบแต่ละข้อวัดในมิติใด สามารถวิเคราะห์ได้ด้วยโปรแกรม TESTFACT และโปรแกรมการวิเคราะห์อื่นๆ

2.2 การประมาณค่าพารามิเตอร์ข้อสอบ

เมื่อได้ข้อสอบที่จะนำมาใช้สำหรับการทดสอบปรับเหมาะแบบพหุมิติแล้ว ควรนำข้อสอบเหล่านั้น ไปทดลองสอบกับกลุ่มตัวอย่างที่คล้ายคลึงกับกลุ่มผู้สอบที่จะมาใช้โปรแกรม เพื่อนำผลการตอบสนองข้อสอบมาวิเคราะห์ค่าพารามิเตอร์ข้อสอบ ซึ่งจะเป็นการวิเคราะห์ตามทฤษฎีการตอบสนองข้อสอบแบบพหุมิติ (multidimensional item response theory: MIRT) หลังจากนั้นให้พิจารณาค่าพารามิเตอร์แต่ละข้อว่ามีความเหมาะสมหรือไม่ สำหรับข้อสอบที่เหมาะสมควรจะมีค่าพารามิเตอร์อำนาจจำแนกเป็นบวก (ทั้งนี้การประมาณค่าจากโปรแกรม TESTFACT ให้ค่าค่อนข้างต่ำ) ค่าพารามิเตอร์ควรมีความยากครอบคลุมตั้งแต่ -3 ถึง 3 และควรมีค่าพารามิ เตอร์โอกาสการเดาต่ำ น้อยกว่า 0.3 จากนั้นให้นำค่าพารามิเตอร์ข้อสอบที่ผ่านการคัดเลือกแล้วกรอกลงในโปรแกรมการทดสอบปรับเหมาะแบบพหุมิติ พร้อมทั้งเฉลยข้อสอบ ดังรูป

พารามิเตอร์ข้อสอบ

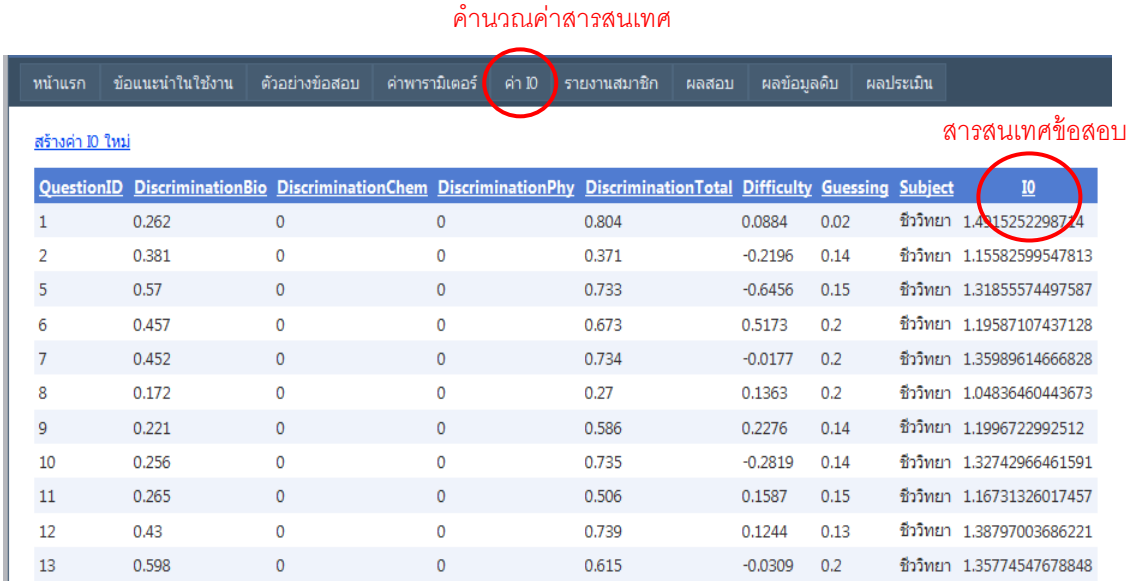
คำตอบถูก/เฉลย

เพิ่มข้อมูล

		ID	Diagram	DTotal	DBio	DChem	DPhy	Diff	Guessing	A	B	C	D	CorrectAnswer	Subject	QuestionType
Edit	Delete	1	B1_01	0.804	0.262	0	0	0.8884	0.02	1	2	3	4	1	ชีววิทยา	B
Edit	Delete	2	B1_02	0.371	0.381	0	0	-0.2196	0.14	1	2	3	4	2	ชีววิทยา	B
Edit	Delete	5	B1_04	0.733	0.57	0	0	-0.6456	0.15	1	2	3	4	1	ชีววิทยา	B
Edit	Delete	6	B1_05	0.673	0.457	0	0	0.5173	0.2	1	2	3	4	4	ชีววิทยา	B
Edit	Delete	7	B1_06	0.734	0.452	0	0	-0.0177	0.2	1	2	3	4	4	ชีววิทยา	B
Edit	Delete	8	B1_07	0.27	0.172	0	0	0.1363	0.2	1	2	3	4	2	ชีววิทยา	B
Edit	Delete	9	B1_08	0.586	0.221	0	0	0.2276	0.14	1	2	3	4	3	ชีววิทยา	B
Edit	Delete	10	B1_09	0.735	0.256	0	0	-0.2819	0.14	1	2	3	4	3	ชีววิทยา	B
Edit	Delete	11	B1_10	0.506	0.265	0	0	0.1587	0.15	1	2	3	4	2	ชีววิทยา	B
Edit	Delete	12	B1_12	0.739	0.43	0	0	0.1244	0.13	1	2	3	4	1	ชีววิทยา	B

2.3 การคำนวณค่าสารสนเทศข้อสอบ

เมื่อกรอกรายละเอียดเกี่ยวกับพารามิเตอร์ข้อสอบเรียบร้อยแล้ว ให้คลิ๊กที่ “ค่า I0” ซึ่งเป็นการคำนวณค่า Fisher information หรือสารสนเทศข้อสอบสำหรับใช้คัดเลือกข้อสอบ หลังจากรอให้โปรแกรมประมวลผล จะได้ค่าสารสนเทศข้อสอบ ดังรูป

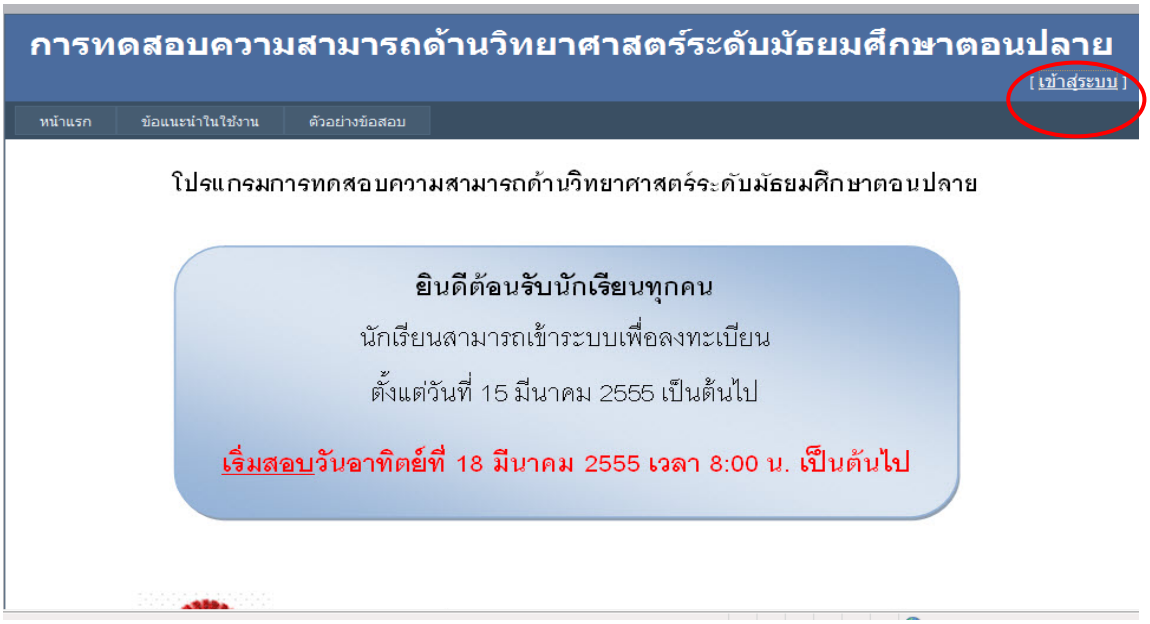


QuestionID	DiscriminationBio	DiscriminationChem	DiscriminationPhy	DiscriminationTotal	Difficulty	Guessing	Subject	I0
1	0.262	0	0	0.804	0.0884	0.02	ชีววิทยา	1.4315252298724
2	0.381	0	0	0.371	-0.2196	0.14	ชีววิทยา	1.15582599547813
5	0.57	0	0	0.733	-0.6456	0.15	ชีววิทยา	1.31855574497587
6	0.457	0	0	0.673	0.5173	0.2	ชีววิทยา	1.19587107437128
7	0.452	0	0	0.734	-0.0177	0.2	ชีววิทยา	1.35989614666828
8	0.172	0	0	0.27	0.1363	0.2	ชีววิทยา	1.04836460443673
9	0.221	0	0	0.586	0.2276	0.14	ชีววิทยา	1.1996722992512
10	0.256	0	0	0.735	-0.2819	0.14	ชีววิทยา	1.32742966461591
11	0.265	0	0	0.506	0.1587	0.15	ชีววิทยา	1.16731326017457
12	0.43	0	0	0.739	0.1244	0.13	ชีววิทยา	1.38797003686221
13	0.598	0	0	0.615	-0.0309	0.2	ชีววิทยา	1.35774547678848

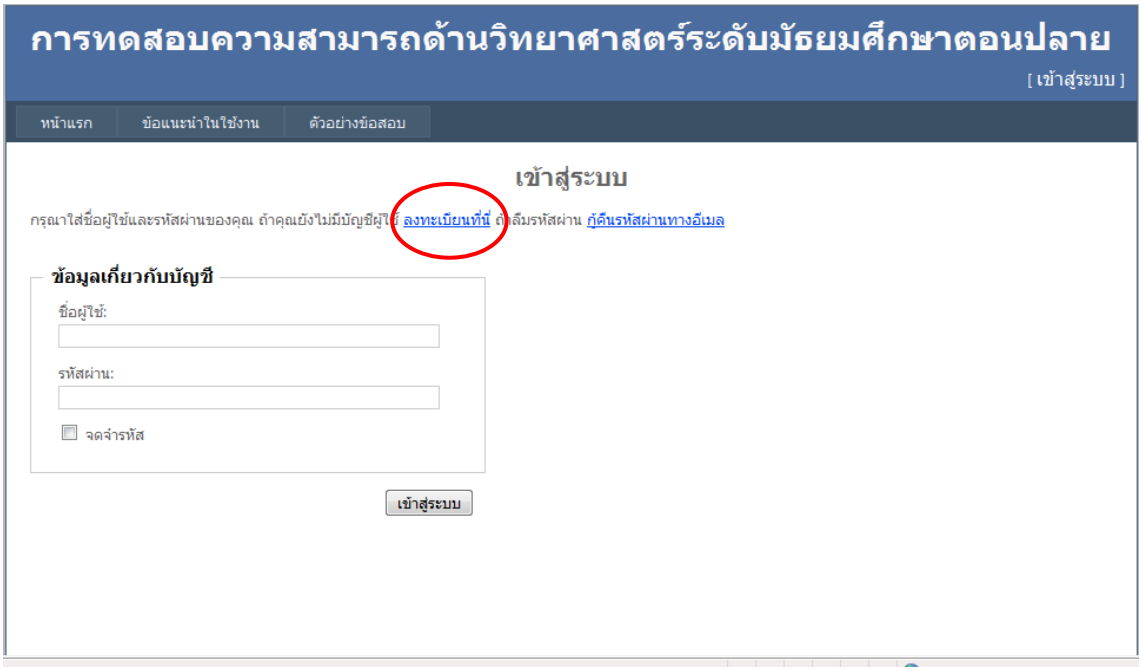
3. ขั้นตอนการใช้โปรแกรม

3.1 เข้าสู่ระบบ

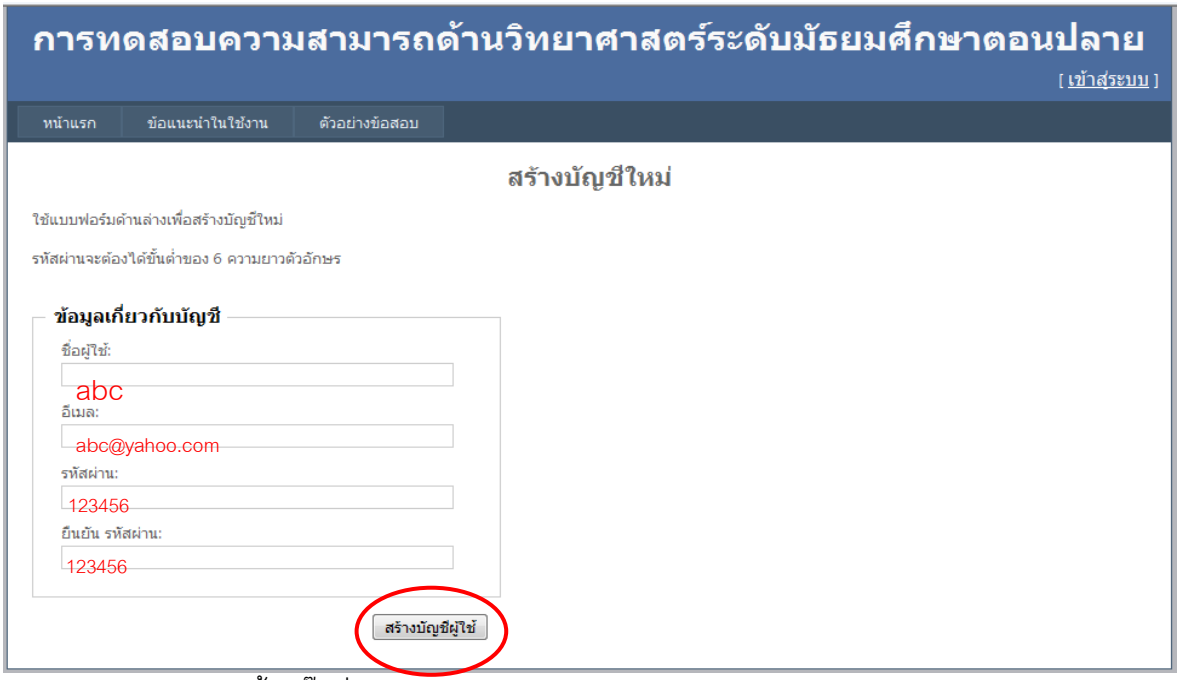
3.1.1 นักเรียนสามารถเริ่มต้นเข้าสู่ระบบการทดสอบโดยคลิ๊กที่ “เข้าสู่ระบบ’ ดังรูป



3.1.2 สำหรับนักเรียนที่เข้าระบบครั้งแรกให้ลงทะเบียนก่อน โดยคลิ๊กที่ “ลงทะเบียนที่นี่” ดังรูป

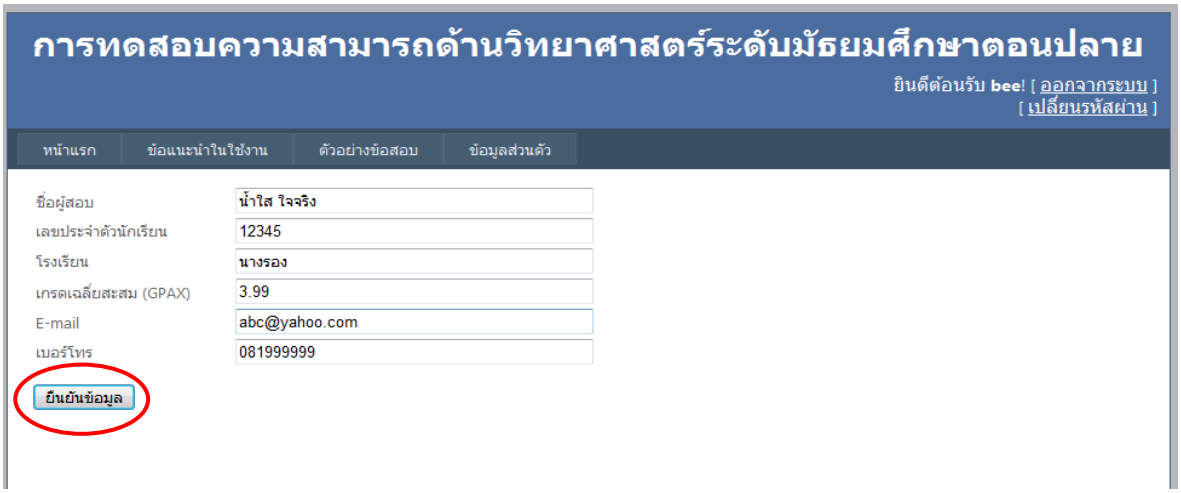


3.1.3 กรอกข้อมูลเกี่ยวกับบัญชีให้ครบถ้วน กำหนดชื่อผู้ใช้ อีเมล์ รหัสผ่าน และยืนยันรหัสผ่าน โดยกำหนดรหัสผ่านให้มีขั้นต่ำ 6 ความยาวตัวอักษร ตัวอย่างเช่น



จากนั้นคลิ๊กที่ “สร้างบัญชีผู้ใช้”

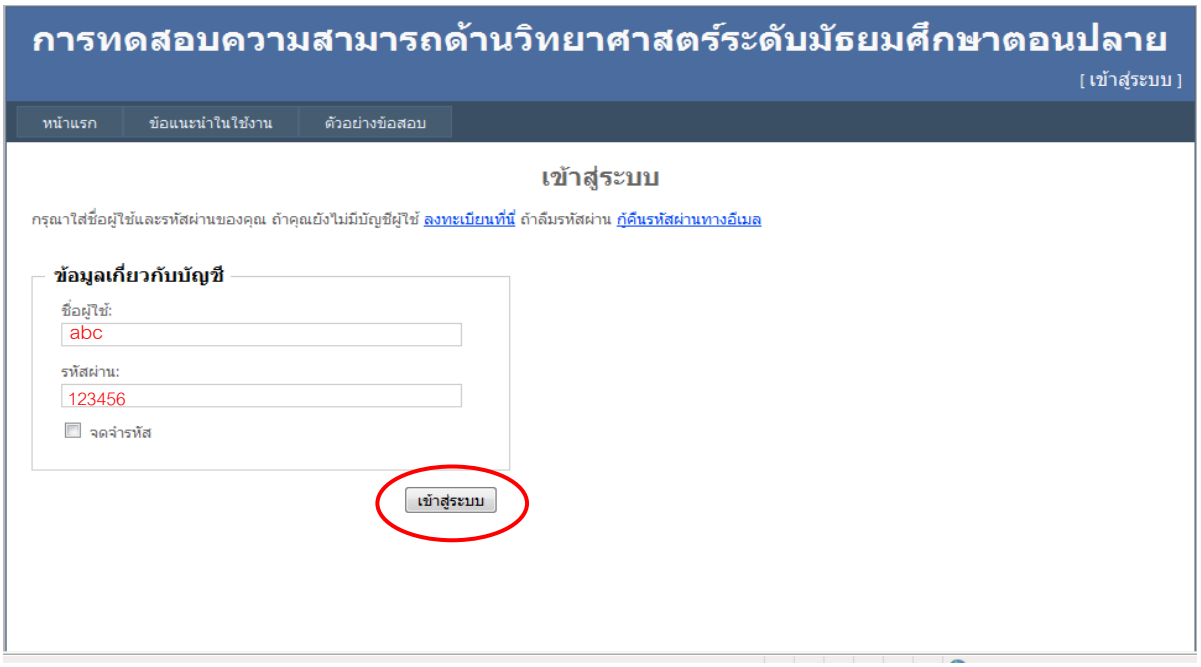
3.1.4 เมื่อปรากฏหน้าต่างข้อมูลผู้สอบ ให้นักเรียนกรอกข้อมูลตามความเป็นจริงให้ครบถ้วนทุกช่อง หากไม่ระบุจะไม่สามารถทำการสอบได้ ตรวจสอบข้อมูลให้ถูกต้อง จากนั้นคลิ๊ก “ยืนยันข้อมูล”



ในกรณีที่ต้องการเปลี่ยนแปลงข้อมูล ให้คลิ๊กตรงตำแหน่งที่ต้องการปรับแก้ และสามารถลบข้อมูล และกรอกข้อมูลใหม่อีกครั้ง จากนั้นคลิ๊กที่ “ยืนยันข้อมูล” ระบบจะบันทึกข้อมูลล่าสุด

3.2 การเข้าสู่ระบบการทดสอบ

ในกรณีที่นักเรียนเคยลงทะเบียนแล้วให้กรอก ชื่อผู้ใช้และรหัสผ่าน จากนั้นคลิ๊กที่ “เข้าสู่ระบบ”



3.3 การเข้าสู่เมนูต่าง ๆ

2.3.1 ข้อแนะนำในการใช้งาน นักเรียนสามารถศึกษาขั้นตอนการลงทะเบียน การเริ่มการทดสอบ และข้อแนะนำต่างๆ ได้จากเมนูนี้

2.3.2 ตัวอย่างข้อสอบ นักเรียนสามารถทดลองดูลักษณะข้อสอบที่ใช้ในกา รทดสอบครั้งนี้ รวมถึงรูปแบบของข้อสอบที่จะแสดงให้นักเรียนเห็นในแต่ละข้อ การเลือกคำตอบการยืนยันคำตอบ เพื่อให้เกอดความคุ้นเคยกับการทดสอบผ่านคอมพิวเตอร์ ก่อนที่นักเรียนจะทำการทดสอบจริง

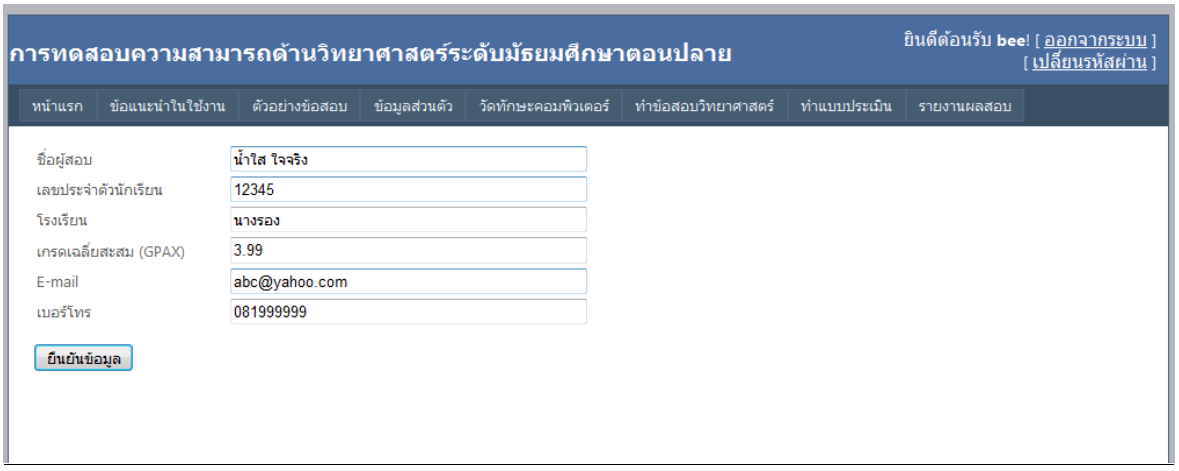
2.3.3 ข้อมูลส่วนตัว นักเรียนสามารถตรวจสอบความถูกต้องของข้อมูลส่ วนตัวของนักเรียนเอง และหากพบข้อผิดพลาด ให้คลิ๊กตรงตำแหน่งที่ต้องการปรับแก้ ลบข้อมูลเดิม และกรอกข้อมูลใหม่อีกครั้ง จากนั้นคลิ๊กที่ “ยืนยันข้อมูล” ระบบจะบันทึกข้อมูลล่าสุด

2.3.4 วัดทักษะคอมพิวเตอร์ ก่อนที่นักเรียนจะเริ่มสอบวัดความสามารถวิทยาศาสตร์ ให้นักเรียนวัดทักษะคอมพิวเตอร์ก่อน โดยคลิ๊กที่เมนู “วัดทักษะคอมพิวเตอร์”

2.3.5 ทำข้อสอบวิทยาศาสตร์ นักเรียนเริ่มการทดสอบโดยคลิ๊กที่เมนูนี้ ระบบจะคัดเลือกข้อสอบขึ้นมาทีละข้อ โดยระบุว่า ข้อสอบข้อนี้เป็นวิชาอะไร รวมถึงแสดงเวลาที่คงเหลือในการทดสอบ (นักเรียนควรศึกษาลักษณะการทดสอบจา ก เมนู “ตัวอย่างข้อสอบ”) เมื่อเริ่มการทดสอบแล้วจะไม่สามารถเข้าสู่เมนูอื่นๆ ได้ นักเรียนต้องดำเนินการจนเสร็จสิ้นกระบวนการทดสอบ

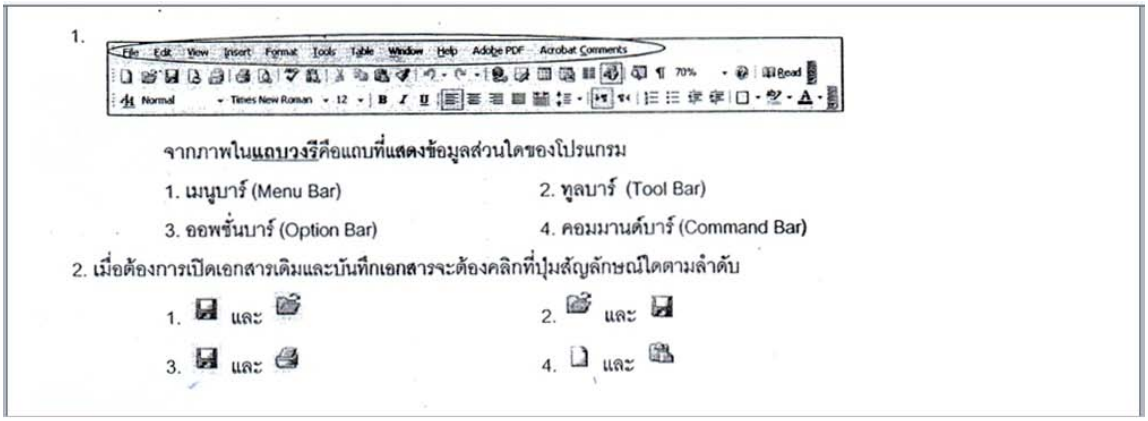
2.3.6 ทำแบบประเมิน นักเรียนทุกคนที่ทำข้อสอบเรียบร้อยแล้ว ให้คลิ๊กที่เมนู “ทำแบบประเมิน” เพื่อประเมินการทดสอบความสามารถด้านวิทยาศาสตร์ผ่านคอมพิวเตอร์

2.3.7 รายงานผลสอบ เมื่อนักเรียนทำข้อสอบเรียบร้อย สามารถดูผลการ สอบจากเมนู “รายงานผลสอบ’

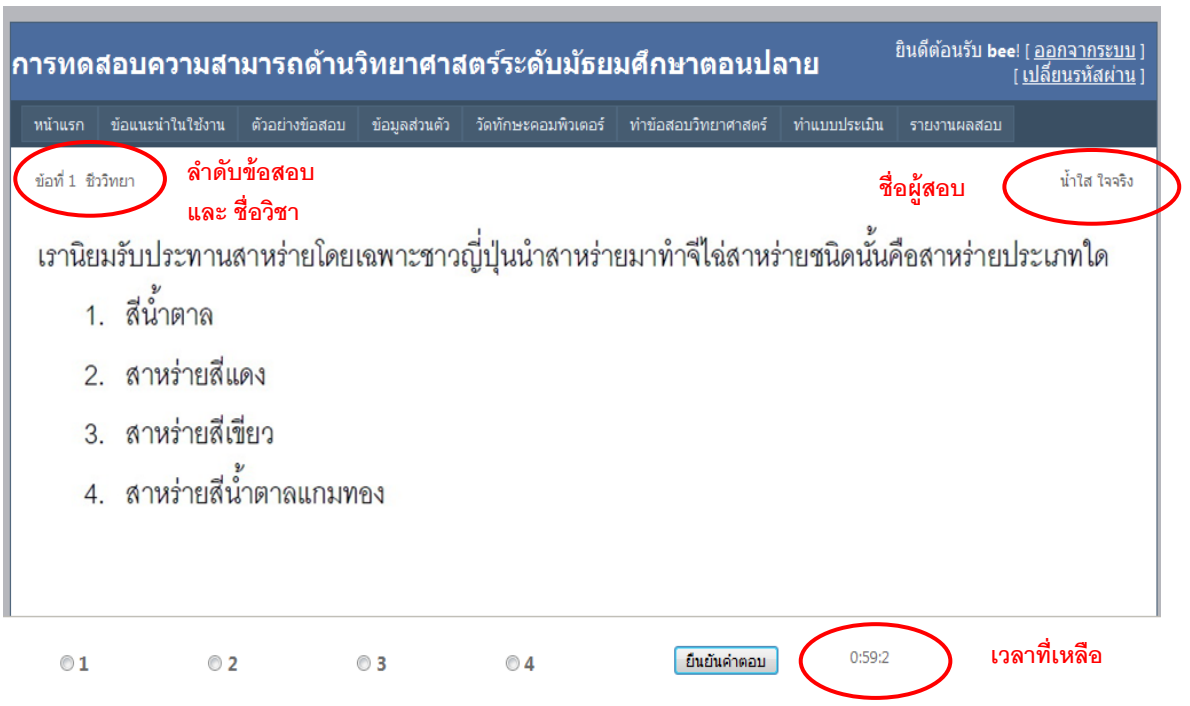


2.4 เริ่มทำการทดสอบ

หลังจากที่นักเรียนกรอกข้อมูลส่วนตัวเรียบร้อย แล้วให้นักเรียนเลือกวัดทักษะคอมพิวเตอร์ก่อน โดยคลิ๊กที่ “วัดทักษะคอมพิวเตอร์ ” ซึ่งเป็นการทดสอบที่นักเรียนจะเห็นข้อสอบหลายข้อพร้อม ๆ กัน โดยมีข้อสอบทั้งสิ้น 35 ข้อ ใช้เวลาทำข้อสอบ 30 นาที ดังนี้



เมื่อวัดทักษะคอมพิวเตอร์เรียบร้อยแล้ว ให้นักเรียนเริ่มทำข้อสอบวิทยาศาสตร์ โดยคลิ๊กที่เมนู “ทำข้อสอบวิทยาศาสตร์ ” ซึ่งระบบจะคัดเลือกข้อสอบขึ้นม าทีละข้อ ใน 3 วิชา คือ ชีววิทยา เคมี และฟิสิกส์ รวมทั้งสิ้น 30 ข้อ ใช้เวลา 1 ชั่วโมง และจะมีนาฬิกาจับเวลาถอยหลังแสดงไว้ที่มุมล่างขวา เมื่อนักเรียนเลือกคำตอบที่ถูกต้องแล้วให้ คลิ๊ก “ยืนยันคำตอบ” รายละเอียดดังรูป



2.4 **ผลการสอบ**

นักเรียนสามารถตรวจสอบ ผลการสอบได้ทันทีหลังจาก “จบการทดสอบ ” และสามารถคลิ๊กที่ “แสดงผลการสอบ ” เพื่อให้ระบบคำนวณผลการสอบในแต่ละวิชารวมถึงคะแนนรวม ซึ่งนักเรียนจะเห็นผลการสอบ ตัวอย่างดังนี้

ผลการทดสอบวัดทักษะคอมพิวเตอร์

คะแนน 51.4285714285714 %

ผลการทดสอบความสามารถทางวิทยาศาสตร์

ชื่อ น้ำใส ใจจริง เลขประจำตัว 12345 โรงเรียน นางรอง

3/23/2012 12:11:39 PM

เวลาที่ใช้ในการทดสอบ 00:09:23

วิชา	ชีววิทยา	เคมี	ฟิสิกส์	รวม
จำนวนข้อสอบที่ได้รับ	8	9	13	30
จำนวนการตอบถูก	1	2	2	5
เปอร์เซนต์การตอบถูก	12.5 %	22.2222222222222 %	15.3846153846154 %	16.6666666666667 %
ระดับความสามารถ	ควรปรับปรุง	ควรปรับปรุง	ควรปรับปรุง	ควรปรับปรุง

* หมายเหตุ เว็บไซด์การทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์เพื่อวัดความสามารถด้านวิทยาศาสตร์ ระดับมัธยมศึกษาตอนปลาย คือ www.mcatesting.com

แบบประเมินโปรแกรมการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์

คำชี้แจง ให้นักเรียนทำเครื่องหมาย ✓ลงในช่องที่ตรงกับระดับความคิดเห็นที่มีต่อการใช้โปรแกรมการทดสอบปรับเหมาะแบบพหุมิติด้วยคอมพิวเตอร์

ลำดับ	ประเด็นคำถาม	ระดับความคิดเห็น				
		มากที่สุด	มาก	ปานกลาง	น้อย	น้อยที่สุด
		5	4	3	2	1
1.	**ด้านความยากในการใช้**					
1.1	โปรแกรมมีความเหมาะสมและให้ความสะดวกในการทดสอบ					
1.2	*การใช้โปรแกรมมีความยุ่งยาก ซับซ้อน*					
1.3	*ผู้สอบไม่คุ้นเคยกับการสอบผ่านคอมพิวเตอร์*					
1.4	เมนูคำสั่งต่างๆ ทำให้ใช้โปรแกรมได้อย่างถูกต้อง					
1.5	ลำดับข้อสอบมีความเหมาะสม					
2.	**ความเป็นไปได้ในการนำไปใช้**					
2.1	การทดสอบผ่านคอมพิวเตอร์มีความน่าเชื่อถือและสามารถนำไปใช้ในการทดสอบได้จริง					
2.2	ควรนำวิธีการทดสอบผ่านคอมพิวเตอร์มาใช้ในชั้นเรียน					
2.3	ควรนำวิธีการทดสอบผ่านคอมพิวเตอร์มาใช้สำหรับการสอบเข้ามหาวิทยาลัย					
2.4	มีความเป็นไปได้ที่จะนำวิธีการทดสอบผ่านคอมพิวเตอร์มาใช้ในอนาคต					
3.	**ความพึงพอใจที่มีต่อการใช้**					
3.1	ผู้ใช้สามารถสั่งแสดงผลการทดสอบทั้งทางจอภาพ ทางเครื่องพิมพ์และอีเมล์					
3.2	โปรแกรมการทดสอบมีความทันสมัย					
3.3	การทดสอบโดยใช้คอมพิวเตอร์ให้คะแนนที่เชื่อถือได้					
3.4	ระยะเวลาในการทดสอบมีความเหมาะสม					

ข้อเสนอแนะเพิ่มเติม

.....

.....

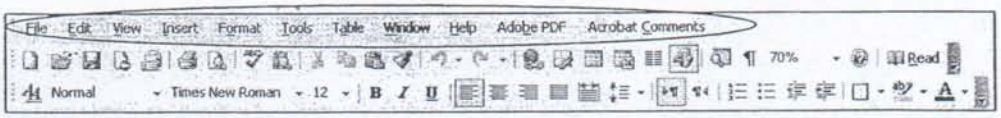
แบบทดสอบความสามารถทางคอมพิวเตอร์เบื้องต้น

คำชี้แจง

1. แบบทดสอบฉบับนี้มีทั้งหมด 35 ข้อ ใช้เวลาทั้งสิ้น 30 นาที
2. ให้เขียนชื่อ – สกุล และกรอกรายละเอียดอื่นๆ ในกระดาษคำตอบให้ชัดเจน
3. ให้นักเรียนเลือกคำตอบที่ถูกที่สุดเพียงข้อเดียว และทำเครื่องหมาย x ลงในกระดาษคำตอบ ห้ามขีดเขียนใดๆ ลงในแบบทดสอบ
4. แบบทดสอบแบ่งออกเป็น 2 ส่วน คือ ส่วนที่ 1 คือ แบบทดสอบความสามารถคอมพิวเตอร์ในการใช้โปรแกรมประมวลผลคำ (MS Word) และส่วนที่ 2 แบบทดสอบความสามารถคอมพิวเตอร์ในการใช้โปรแกรมนำเสนอผลงาน (MS PowerPoint)
5. ข้อมูลจากการทดสอบของนักเรียนจะถือเป็นความลับและไม่มีผลกระทบใดๆ ต่อนักเรียนทั้งสิ้น

ส่วนที่ 1 คือ แบบทดสอบความสามารถทางคอมพิวเตอร์ในการใช้โปรแกรมประมวลผลคำ

1.



จากภาพในแถบวงรีคือแถบที่แสดงข้อมูลส่วนใดของโปรแกรม

1. เมนูบาร์ (Menu Bar)
2. ทูลบาร์ (Tool Bar)
3. ออพชั่นบาร์ (Option Bar)
4. คอมมานด์บาร์ (Command Bar)

2. เมื่อต้องการเปิดเอกสารเดิมและบันทึกเอกสารจะต้องคลิกที่ปุ่มสัญลักษณ์ใดตามลำดับ

1. และ
2. และ
3. และ
4. และ

3. ถ้านักเรียนต้องการกำหนดขนาดกระดาษในการทำงานเป็น**ขนาด Legal** นักเรียนต้องทำอย่างไร

1. Edit > Page Setup > Paper > Paper Margin = A4
2. Edit > Page Setup > Paper > Paper Size = Legal
3. File > Page Setup > Paper > Paper Margin = A4
4. File > Page Setup > Paper > Paper Size = Legal

4. **ข้อใดกล่าวผิด**เกี่ยวกับสัญลักษณ์ในการกำหนดตำแหน่งแท็บ

1. ⌞ ใช้กำหนดให้ข้อความจัดชิดแท็บทางด้านซ้าย
2. ⊥ ใช้กำหนดให้จุดทศนิยม อยู่ในแนวเดียวกับตำแหน่ง ⊥ ในไม้บรรทัดแนวนอน เหมาะกับการเรียงตัวเลขให้จุดทศนิยมตรงกัน
3. ı ใช้กำหนดเส้นในแนวนอน ให้อยู่ในแนวเดียวกับตำแหน่งไม้บรรทัดในแนวตั้ง
4. ⊥ ใช้กำหนดให้ข้อความจัดกึ่งกลางแท็บ

5. เมื่อต้องการคัดลอกข้อความที่ได้เลือกเอาไว้ ต้องคลิกที่ปุ่มใด

1.
2.
3.
4.

6. การจัดข้อความทั้งย่อหน้าให้อยู่กึ่งกลางหน้ากระดาษต้องคลิกที่ปุ่มใด หรือใช้คำสั่งใดในการจัดย่อหน้า

1. คลิกที่ปุ่ม หรือใช้คำสั่ง Format > Paragraph > Alignment = Left
2. คลิกที่ปุ่ม หรือใช้คำสั่ง Format > Paragraph > Alignment = Centered
3. คลิกที่ปุ่ม หรือใช้คำสั่ง Format > Paragraph > Alignment = Right
4. คลิกที่ปุ่ม หรือใช้คำสั่ง Format > Paragraph > Alignment = Jusitified

7. เมื่อต้องการปรับแต่งปุ่มเครื่องหมายหน้าข้อความให้เป็นเครื่องหมายต่างๆสามารถทำได้อย่างไร

1. Format > Bullet and Numbering หรือคลิกที่ปุ่ม
2. Insert > Bullet and Numbering หรือคลิกที่ปุ่ม
3. Format > Bullet and Numbering หรือคลิกที่ปุ่ม
4. Insert > Bullet and Numbering หรือคลิกที่ปุ่ม

8. เมื่อเลือกข้อความ "คอมพิวเตอร์" แล้วคลิกปุ่ม และปุ่ม ข้อความจะเป็นอย่างไร

1. คอมพิวเตอร์
2. **<u>คอมพิวเตอร์</u>**
3. ***คอมพิวเตอร์***
4. *<u>คอมพิวเตอร์</u>*

9. ข้อใดกล่าวถูกต้องเกี่ยวกับการจัดย่อหน้า

ก. การปรับระยะห่างระหว่างบรรทัดต้องเลือกเมนู Edit และเลือกคำสั่ง Ruler

ข. การปรับระยะห่างระหว่างย่อหน้าทำได้โดยการคลิกเมาส์ปุ่มขวาในย่อหน้าที่ต้องการแล้วเลือกคำสั่ง Paragraph

1. ข้อ ก ถูกต้อง
2. ข้อ ข ถูกต้อง
3. ข้อ ก และ ข ถูกต้อง
4. ไม่มีข้อใดถูกต้อง

10. ขั้นตอนในการแทรกสัญลักษณ์พิเศษหรืออักขระพิเศษลงในเอกสารสามารถทำได้อย่างไร

ก. จากกล่องโต้ตอบ Symbol ในช่อง Font ให้เลือกรายการของรูปแบบที่ต้องการ

ข. คลิกปุ่ม Insert เพื่อแทรกสัญลักษณ์ที่จุดแทรกข้อความในเอกสาร

ค. คลิกเลือกสัญลักษณ์ที่ต้องการแทรกเข้าไปในเอกสาร

ง. คลิกเลือกเมนู Insert แล้วเลือกคำสั่ง Symbol

1. ง – ก – ค – ข
2. ง – ข – ค – ก
3. ข – ก – ค –ง
4. ก – ง – ค – ข

11. ถ้าต้องการกำหนดรูปแบบของกรอบข้อความให้เป็นเส้นประ จะต้องกำหนดที่หัวข้อใด ในแท็บ Borders ของกล่องโต้ตอบ Borders and Shading

1. Setting
2. Style
3. Color
4. Width

12. การเรียกดูเอกสารก่อนพิมพ์ครั้งละหลายๆหน้า ต้องคลิกที่ปุ่มใด

1.
2.
3.
4.

13. เมื่อต้องการพิมพ์เอกสารหน้า 14 ถึง หน้า 20 โดยพิมพ์เฉพาะหน้าคู่ต้องป้อนตัวเลขในช่อง Pages อย่างไร

1. 14 – 20
2. 14 – 16 , 18 – 20
3. 20 – 14
4. 14, 16 , 18 , 20

14. ข้อใดเป็นวิธีการเลือกเซลล์ทั้งตาราง

1. คลิกตัวชี้เมาส์บนคอลัมน์ที่ต้องการ แล้วเลือกเมนู Table > Select Column
2. คลิกตัวชี้เมาส์บนแถวที่ต้องการ แล้วเลือกเมนู Table > Select Row
3. คลิกตัวชี้เมาส์บนเซลล์ที่ต้องการ แล้วเลือกเมนู Table > Select Cell
4. คลิกตัวชี้เมาส์บนเซลล์ในตาราง แล้วเลือกเมนู Table > Select Table

15. ก่อนการคลิกเลือกเมนู Table > Insert Columns เพื่อแทรกคอลัมน์ จะต้องทำกระบวนการใดก่อน

1. คลิกเลือกเซลล์
2. คลิกเลือกกลุ่มเซลล์
3. คลิกเลือกคอลัมน์
4. ถูกทุกข้อ

16. หากต้องการสร้างตัวอักษรศิลป์ (Word art) ลงในเอกสารจะต้องคลิกที่ปุ่มใด

1.
2.
3.
4.

17. ปุ่มใดต่อไปนี้ ใช้สำหรับการสลับส่วนหัวกระดาษ และท้ายกระดาษของเอกสาร

1.
2.
3.
4.

ส่วนที่ 2 คือ แบบทดสอบความสามารถทางคอมพิวเตอร์ในการใช้โปรแกรมนำเสนอผลงาน

18. ข้อใดไม่ใช่ความสามารถของโปรแกรม PowerPoint

1. ทำงานงานคำนวณข้อมูลทางบัญชีโดยเฉพาะ
2. แทรกแผนภูมิและภาพเคลื่อนไหวได้
3. สามารถแทรกเสียงประกอบได้
4. สามารถวาดรูปขณะนำเสนอผลงาน

19. วิธีการปิดงานนำเสนอที่ได้สร้างไว้อย่างถูกต้อง คือข้อใด

1. เลื่อนเมาส์ไปคลิกที่ปุ่ม Minimize
2. ใช้คำสั่ง File > Close
3. เลื่อนเมาส์ไปคลิกที่ปุ่ม
4. กดคีย์ < Ctrl + C >

20. ลักษณะของการสร้างงานนำเสนอ โดยเลือกใช้ Blank Presentation คือข้อใด

1. แผ่นสไลด์เปล่า
2. แผ่นสไลด์ที่มีองค์ประกอบทางกราฟฟิก
3. แผ่นสไลด์ที่มีการกำหนดหัวข้อและโครงเนื้อหา
4. แผ่นสไลด์ที่มีกล่องข้อความและองค์ประกอบทางกราฟฟิก

21. หากนักเรียนต้องการดูสไลด์ทั้งหมด และต้องการเรียงลำดับการนำเสนอใหม่ควรคลิกที่ปุ่ม หรือเรียกใช้มุมมองในโปรแกรมใด

1. ปุ่ม หรือ มุมมอง Normal View
2. ปุ่ม หรือ มุมมอง Slide Show
3. ปุ่ม หรือ มุมมอง Slide Sorter View
4. ปุ่ม หรือ มุมมอง Slide Sorter View

22. ถ้านักเรียนต้องการเพิ่มสไลด์ จะต้องคลิกที่ปุ่มใดตามลำดับ

1. คลิกที่เมนู Insert -> New Slide หรือ กดคีย์ Ctrl + M
2. คลิกที่เมนู Slide Show -> New Slide หรือ กดคีย์ Ctrl + N
3. คลิกที่เมนู Insert -> New Slide หรือ กดคีย์ Ctrl + M
4. คลิกที่เมนู Slide Show -> เ New Slide หรือ กดคีย์ Ctrl + N

23. การเปลี่ยนขนาดตัวอักษร โดยการกดปุ่ม ผลลัพธ์จะเป็นอย่างไร

1. ตัวอักษรจะมีขนาดใหญ่กว่าขนาดเดิม
2. ตัวอักษรจะมีขนาดเล็กลงกว่าเดิม
3. ตัวอักษรจะมีตำแหน่งตัวยกขึ้นจากตำแหน่งเดิม
4. ตัวอักษรจะมีตำแหน่งตัวห้อยจากตำแหน่งเดิม

24. การกำหนดตัวอักษรหรือข้อความให้เป็นตัวหนาและใส่เงาให้กับตัวอักษรและข้อความต้องใช้เครื่องมือใดตามลำดับ

1. เครื่องมือ และ
2. เครื่องมือ และ
3. เครื่องมือ และ
4. เครื่องมือ และ

25. ถ้าเราต้องการเปลี่ยนสัญลักษณ์หรือตัวเลขที่ใช้เป็นหัวข้อสามารถเข้าปรับได้ที่เมนูใด

1. Tools > Bullet and Numbering
2. Insert > Bullet and Numbering
3. Format > Bullet and Numbering
4. Insert > Numbering and Bullet

26. หากเราต้องการวาดรูปทรงกระบอก และรูปหัวลูกศรแบบต่างๆด้วย AutoShape จะต้องใช้คำสั่งใด ตามลำดับ

1. AutoShape > Line , AutoShape > Connector
2. AutoShape > Callouts , AutoShape > Star and Banners
3. AutoShape > Basic shape , AutoShape > Block Arrows
4. AutoShape > Connector , AutoShape > Block Arrows

27. ข้อใดกล่าวถูกต้องเกี่ยวกับภาพแบบเวคเตอร์ (Vector) และภาพแบบราสเตอร์ (Raster)

ก. ภาพแบบเวคเตอร์ (Vector) เป็นภาพที่เกิดจากการวาดเส้นตรง เส้นโค้ง รูปทรงเรขาคณิต โดยภาพชนิดนี้จะถูกจัดเก็บในลักษณะการประมวลผลเป็นสูตรคณิตศาสตร์ ได้แก่ ไฟล์ฟอร์แมต .wmf

ข. ภาพแบบราสเตอร์ (Raster) เป็นภาพที่เกิดจากการเครื่องสแกนหรือภาพถ่ายจากกล้องดิจิตอลจนประกอบไปด้วยจุดเล็กๆ ได้แก่ ไฟล์ฟอร์แมต .bmp .cgm .gif .png .jpg

1. ข้อ ก ถูกต้อง
2. ข้อ ข ถูกต้อง
3. ข้อ ก และ ข ถูกต้อง
4. ไม่มีข้อใดถูก

28. ปุ่มคำสั่งใดที่ใช้ปรับแต่งโหมดสีของภาพ

1.
2.
3.
4.

29. ถ้าเราต้องการสร้างผังองค์กร จะต้องคลิกที่ปุ่มคำสั่งใด และโดยปกติการสร้างแผนผังขององค์กรจะเริ่มต้นที่บุคลากรกี่ตำแหน่ง

1. ปุ่ม , เริ่มต้นที่ 1 ตำแหน่ง
2. ปุ่ม เริ่มต้นที่ 2 ตำแหน่ง
3. ปุ่ม , เริ่มต้นที่ 3 ตำแหน่ง
4. ปุ่ม เริ่มต้นที่ 4 ตำแหน่ง

30. หากเราต้องการให้พื้นหลังมีลักษณะเป็นลาย "หินอ่อน" ต้องเลือกรูปแบบใด

1. Texture
2. Picture
3. Pattern
4. Gradient

31. ประโยชน์ของการสร้าง Slide Master คือข้อใด

1. ได้รูปแบบของสไลด์ที่มีความหลากหลาย
2. การแก้ไขข้อมูลบนสไลด์ไม่ขึ้นอยู่กับสไลด์อื่น
3. ได้สไลด์ที่ตรงกับความต้องการใช้งานมากที่สุด
4. อำนวยความสะดวกต่อการทำสไลด์ที่มีลักษณะเหมือนๆกัน

32. จงพิจารณาข้อความต่อไปนี้เกี่ยวกับการใส่เอฟเฟคและการใส่ทรานซิชัน (Transition)

ก. การกำหนดให้ข้อความวิ่งจากทางขวามาทางซ้ายมือของจอภาพ คือ คำสั่ง Fly From Left

ข. การกำหนดเอฟเฟคสามารถกำหนดให้ได้กับตัวหนังสือเท่านั้นไม่สามารถ กำหนดให้กับกราฟและรูปภาพได้

ค. การใส่ทรานซิชัน (Transition) เป็นเทคนิคอย่างหนึ่งที่ใช้ในการเปลี่ยนสไลด์

1. ข้อ ก ถูกต้องเพียงข้อเดียว
2. ข้อ ก และ ข ถูกต้อง
3. ข้อ ค ถูกต้องเพียงข้อเดียว
4. ข้อ ก และ ค ถูกต้อง

33. ไฟล์เสียงและไฟล์วิดีโอนามสกุลใดที่นักเรียนสามารถที่เราสามารถแทรกลงในการนำเสนอผลงาน Microsoft PowerPoint ได้ตามลำดับ

1. ไฟล์เสียง ได้แก่ .mid .wav .wma และ ไฟล์วิดีโอ ได้แก่ .mp4 .mov
2. ไฟล์เสียง ได้แก่ .mp3 .ogg .wma และ ไฟล์วิดีโอ ได้แก่ .wmv .mov
3. ไฟล์เสียง ได้แก่ .mid .wav .wma และ ไฟล์วิดีโอ ได้แก่ . wmv
4. ไฟล์เสียง ได้แก่ . mp3 .ogg .wma และ ไฟล์วิดีโอ ได้แก่ .mp4 . mov

34. ถ้าเราต้องการกำหนดการเชื่อมโยงสไลด์เพื่อความสะดวกในการนำเสนองาน เราสามารถสร้างปุ่มปฏิบัติการได้โดยขั้นตอนใด

1. ที่เมนูคลิกเลือก Slide Show > Setup Show...
2. ที่เมนูคลิกเลือก Slide Show > Action Buttons
3. ที่เมนูคลิกเลือก Slide Show > Setup Buttons
4. ที่เมนูคลิกเลือก Slide Show > Custom Buttons

35. ในการสั่ง Print What จะมีคำสั่งทั้งหมด 4 ประเภท คือ Slides, Handouts , Notepages, Outline View ดังนั้นคำสั่ง Print What หมายถึงอะไร

1. การกำหนดลำดับของสไลด์ที่ต้องการจะพิมพ์ออกทางเครื่องพิมพ์ที่ตั้งต่าไว้
2. การกำหนดจำนวนชุดของเอกสารที่ต้องการจะพิมพ์ออกทางเครื่องพิมพ์ที่ยังไม่ได้ตั้งค่า
3. การกำหนดมาตรฐานของการพิมพ์ของเครื่องพิมพ์ชนิดต่างๆ ที่ยังไม่ได้ตั้งค่า
4. การเลือกสิ่งที่ต้องการพิมพ์และกำหนดคุณสมบัติของเอกสารประกอบการบรรยาย

☺ ขอขอบคุณที่กรุณาตอบแบบสอบถาม ☺

ภาคผนวก ฉ

คำสั่งที่ใช้ในการวิเคราะห์

1. คำสั่งที่ใช้ในการวิเคราะห์ค่าพารามิเตอร์การเดา (c) ด้วยโปรแกรม BILOG
2. คำสั่งที่ใช้ในการวิเคราะห์ค่าพารามิเตอร์การความยาก (b) และอำนาจจำแนก (a) ด้วยโปรแกรม TESTFACT
3. คำสั่งที่ใช้ในการวิเคราะห์การทดสอบปรับเหมาะแบบพหุมิติ ด้วยโปรแกรม R

คำสั่งที่ใช้ในการวิเคราะห์ค่าพารามิเตอร์การเดา (c) ด้วยโปรแกรม BILOG

```
EXAMPL01.BLM - TRADITIONAL IRT ANALYSIS
>GLOBAL DFName = 'test1.DAT',
        NPArm = 3,
        SAVe;
>SAVE PARm = 'PAT.PAR',
      SCOre = 'PAT.SCO';
>LENGTH NITems = (100);
>INPUT NTOtal = 100,
       NIDchar = 2,
       KFName = 'test1.KEY',
       OFName = 'test1.OMT';
>ITEMS INAmes = (ITEM01(1)ITEM100);
>TEST1 TNAme = 'PRETEST',
       INUmber = (1(1)100);
(2A1,100A1)
>CALIB NQPt = 25,
       CYCles = 25,
       NEWton = 10,
       CRIt = 0.0010,
       PLOt = 1.0000,
       ACCel = 0.0000,
       CHIsquare = (100, 9);
>SCORE RSCtype = 4,
       INFo = 2,
       NOPrint,
       POP;
```

คำสั่งที่ใช้ในการวิเคราะห์ค่าพารามิเตอร์การความยาก (b) และอำนาจจำแนก (a) ด้วยโปรแกรม TESTFACT

```
>TITLE
Bifactor.TSF - LSAT DATA NON-ADAPTIVE FULL-INFORMATION ITEM
FACTOR ANALYSIS COUNTED RESPONSE PATTERNS

>PROBLEM       NITEMS=100, RESPONSE=5;

>NAMES         ITEM1, ITEM2, ITEM3, ITEM4, ITEM5, ITEM6, ITEM7,
ITEM8,  ITEM9, ITEM10, ITEM11, ITEM12, ITEM13, ITEM14, ITEM15,
ITEM16,  ITEM17,  ITEM18,  ITEM19,  ITEM20,  ITEM21,  ITEM22,
ITEM23,  ITEM24,  ITEM25,  ITEM26,  ITEM27,  ITEM28,  ITEM29,
ITEM30,  ITEM31,  ITEM32,  ITEM33,  ITEM34,  ITEM35,  ITEM36,
ITEM37,  ITEM38,  ITEM39,  ITEM40,  ITEM41,  ITEM42,  ITEM43,
ITEM44,  ITEM45,  ITEM46,  ITEM47,  ITEM48,  ITEM49,  ITEM50,
ITEM51,  ITEM52,  ITEM53,  ITEM54,  ITEM55,  ITEM56,  ITEM57,
ITEM58,  ITEM59,  ITEM60,  ITEM61,  ITEM62,  ITEM63,  ITEM64,
ITEM65,  ITEM66,  ITEM67,  ITEM68,  ITEM69,  ITEM70,  ITEM71,
ITEM72,  ITEM73,  ITEM74,  ITEM75,  ITEM76,  ITEM77,  ITEM78,
ITEM79,  ITEM80,  ITEM81,  ITEM82,  ITEM83,  ITEM84,  ITEM85,
ITEM86,  ITEM87,  ITEM88,  ITEM89,  ITEM90,  ITEM91,  ITEM92,
ITEM93,  ITEM94,  ITEM95,  ITEM96,  ITEM97,  ITEM98,    ITEM99,
ITEM100;

 >RESPONSE      '9','1','2','3','4';

 >KEY           42442321214213122313314223334344412244231342442

32343214144334243323134423342334322144221323312413224;

 >BIFACTOR      NIGROUP=3, LIST=5, CYCLES=10, QUAD=9,

IGROUPS=(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,

2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,3,3,3,3,3,3,3,
3,3,3,3,3,3,3,3,3,3,3,3,3,3);

CPARMS=(0.017, 0.137, 0.147, 0.15, 0.201, 0.201, 0.202, 0.139,
0.135, 0.147, 0.083, 0.134, 0.199, 0.201, 0.202, 0.202, 0.118,
0.163, 0.202, 0.202, 0.115, 0.176, 0.201, 0.1, 0.179, 0.197,
0.201, 0.201, 0.199, 0.121, 0.119, 0.201, 0.201, 0.146, 0.173,
0.132, 0.201, 0.13, 0.184, 0.153, 0.201, 0.184, 0.179, 0.201,
0.202, 0.198, 0.191, 0.201, 0.201, 0.164, 0.165, 0.201, 0.18,
0.184, 0.201, 0.119, 0.156, 0.202, 0.202, 0.148, 0.2, 0.15,
0.201, 0.173, 0.159, 0.183, 0.202, 0.156, 0.202, 0.13, 0.103,
0.171, 0.201, 0.092, 0.151, 0.196, 0.202, 0.03, 0.2, 0.203,
0.201, 0.16, 0.094, 0.181, 0.14, 0.201, 0.201, 0.141, 0.201,
0.202, 0.192, 0.148, 0.202, 0.135, 0.108, 0.143, 0.201, 0.201,
0.024, 0.193);

 >SCORE         LIST=10, CHANCE;

 >SAVE          PARM, FSCORES;

 >INPUT         NIDCHAR=5, SCORES, FILE='PAT 2.DAT';

(5A1, 100A1)
```

คำสั่งที่ใช้ในการจำลองข้อมูลและดำเนินการทดสอบปรับเหมาะแบบพหุมิติ ด้วยโปรแกรม R

1. คำสั่งการสร้างข้อมูลการสอบ (พารามิเตอร์ข้อสอบ และความสามารถผู้สอบ)

```
library("MASS")
library("msm")

#function to generate item parameters
item.par.mixed<-function(H,I,dis.mean,dis.sd,max.c)
{
#generate discrimination parameter
a<-matrix(nrow=I,ncol=H)
for (i in 1:H)
{
a[,i]<-rtnorm(I,mean=dis.mean[i],sd=dis.sd[i],lower=0,upper=3)
}
#generate difficulty parameters
#item<-1:I
#hi=3 #difficulty of the most item difficulties
#range<-2*hi #range of item difficulties
#b<-(item-mean(item))*range/(I-1)
b<-runif(I,-3,3)
#generate guessing parameter
c<-runif(I,0,max.c)

out<-list(a,b,c)
names(out)<-c("a","b","c")
return(out)
}

#function to generate true ability function
gen.theta<-function(H,N,theta.mean,cov.theta)
{
theta.true<-mvrnorm(N,mu=theta.mean,Sigma=cov.theta)
return(theta.true)
}

#function to generate response pattern
gen.data<-function(H,I,a,b,c,theta.true)
{
prob<-matrix(nrow=I,ncol=1)
y<-matrix(nrow=I,ncol=1)
D<-1.7
for (i in 1:I)
```

```
{
prob[i,1]<-c[i]+(1-c[i])/(1+exp(-D*t(a[i,])%*%(theta.true-b[i]*rep(1,H))))
y[i,1]<-ifelse(runif(1,0,1)<=prob[i,1],1,0)
}
out<-list(prob,y)
names(out)<-c("prob","response")
return(out)
}
```

2. คำสั่งของ Fisher information แบบผสมผสานมิติ

```
#FISHER
fisher.mixed<-function(H,I,a,b,c,y,mu0,phi0,theta0,testlength)
{
D<-1.7
round<-0
theta<-theta0
theta.iter<-matrix(nrow=testlength,ncol=H)
SE<-1
SE.iter<-matrix(nrow=testlength,ncol=1)
fisher<-matrix(nrow=testlength,ncol=1)
Ii<-matrix(nrow=I,ncol=1) #matrix of item selection criterion
Sn.dummy<-matrix(0,nrow=I,ncol=1)
Sn<-0
for (s in 1:testlength)
{

#item selection step
round<-round+1
if(round<2){

for (i in 1:I)
{
p<-c[i]+(1-c[i])/(1+exp(-D*t(a[i,])%*%(theta-b[i]*rep(1,H))))
q<-1-p
wi.s<-as.numeric((q/p)*((p-c[i])/(1-c[i]))^2)
Wi<-D^2*a[i,]%*%t(a[i,])*wi.s
Ii[i]<-det(solve(phi0)+Wi)
}

#choose item
k<-1
while(Ii[k]!=max(Ii))
{k<-k+1}
Sn<-k
```

```
Sn.dummy[k]<-1
I.final<-Ii*(Sn.dummy)}else{
W.ad<-0
for (i in 1:length(Sn))
{
p<-c[Sn[i]]+(1-c[Sn[i]])/(1+exp(-D*t(a[Sn[i],])%*%(theta-b[Sn[i]]*rep(1,H))))
q<-1-p
wi.s<-as.numeric((q/p)*((p-c[Sn[i]])/(1-c[Sn[i]]))^2)
W.ad<-W.ad+D^2*a[Sn[i],]%*%t(a[Sn[i],])*wi.s
}
for (i in 1:I)
{
p<-c[i]+(1-c[i])/(1+exp(-D*t(a[i,])%*%(theta-b[i]*rep(1,H))))
q<-1-p
wi.s<-as.numeric((q/p)*((p-c[i])/(1-c[i]))^2)
Wi<-D^2*a[i,]%*%t(a[i,])*wi.s
Ii[i]<-det(solve(phi0)+Wi+W.ad)
}
I.final<-Ii*Sn.dummy
Ii<-Ii*(1-Sn.dummy)

#choose item
k<-1
while(Ii[k]!=max(Ii))
{k<-k+1}
Sn<-cbind(Sn,k)
Sn.dummy[k]<-1
}


#ability estimation step
diff<-c(1,1)
while(abs(diff[1])>10^-5 & abs(diff[2])>10^-5)
{
sum1<-0
sum2<-0
for (i in 1:length(Sn))
{
p<-c[Sn[i]]+(1-c[Sn[i]])/(1+exp(-D*t(a[Sn[i],])%*%(theta-b[Sn[i]]*rep(1,H))))
q<-1-p

vi<-as.numeric((p-c[Sn[i]])*(y[Sn[i]]-p)/((1-c[Sn[i]])*p))
wi<-as.numeric((q*(p-c[Sn[i]])*(c[Sn[i]]*y[Sn[i]]-p^2))/(p*(1-c[Sn[i]]))^2)
sum1<-sum1+vi*a[Sn[i],]
sum2<-sum2+a[Sn[i],]%*%t(a[Sn[i],])*wi
}
```

```
sum1<-matrix(sum1,nrow=H,ncol=1)
#first-order partial derivative of log-posterior
dlog.p<-D*sum1-solve(phi0)%*%(theta-mu0)
#second-order partial derivative
M.theta<-D^2*sum2-solve(phi0)

delta<-solve(M.theta)%*%dlog.p
theta.new<-theta-delta
diff<-theta.new-theta
theta<-theta.new
}
}
out<-list(theta.iter,fisher,SE.iter,testlength)
names(out)<-c("theta.est","fisher","SE","testlength")
return(out)

}
```

3. คำสั่ง Fisher information แบบจำแนกมิติ

```
#fisher seperate
fisher.sep<-function(H,I,a,b,c,y,mu0,phi0,theta0,testlength)
{
item<-matrix(nrow=testlength/2,ncol=H)
b.sep<-matrix(nrow=I/2,ncol=H)
c.sep<-matrix(nrow=I/2,ncol=H)

theta.iter<-matrix(nrow=testlength,ncol=H)
fisher<-matrix(nrow=I,ncol=testlength)
SE.iter<-matrix(nrow=testlength,ncol=1)
I.final<-0

for (s in 1:H)
{
b.sep[,s]<-b[(1+(s-1)*(I/H)):(s*I/H)]
c.sep[,s]<-c[(1+(s-1)*(I/H)):(s*I/H)]
}

D<-1.7
theta<-theta0
SE<-1
Ii<-matrix(0,nrow=I/H,ncol=H)
Ii.choose<-matrix(0,nrow=I/H,ncol=H)
Sn<-0
Sn.dummy.sep<-matrix(0,nrow=I/H,ncol=H)
```

```
seq<-sample(subjects<-c(1:H),H,replace=FALSE)

for (s in 1:H)
{
Sn.sep<-0
a.sep<-a[(1+(seq[s]-1)*(I/H)):(seq[s]*I/H),]

for (t in 1:(testlength/2))
{
#item selection step

if(t<2){

for (i in 1:(I/2))
{
p<-c.sep[i,seq[s]]+(1-c.sep[i,seq[s]])/(1+exp(-D*t(a.sep[i,])%*%(theta-b.sep[i,seq[s]]*rep(1,H))))
q<-1-p
wi.s<-as.numeric((q/p)*((p-c.sep[i,seq[s]])/(1-c.sep[i,seq[s]]))^2)
Wi<-D^2*a.sep[i,]%*%t(a.sep[i,])*wi.s
Ii[i,seq[s]]<-det(solve(phi0)+Wi)
}

#choose item
k<-1
while(Ii[k,seq[s]]!=max(Ii[,seq[s]]))
{k<-k+1}
Sn.sep<-k
Sn<-k+(seq[s]-1)*(I/H)
Sn.dummy.sep[k,seq[s]]<-1

I.final<-matrix(Ii,nrow=I,ncol=1)*matrix(Sn.dummy.sep,nrow=I,ncol=1)}else{
W.ad<-0
for (i in 1:length(Sn))
{
p<-c.sep[Sn.sep[i],seq[s]]+(1-c.sep[Sn.sep[i],seq[s]])/(1+exp(-D*t(a.sep[Sn.sep[i],])%*%(theta-
b.sep[Sn.sep[i],seq[s]]*rep(1,H))))
q<-1-p
wi.s<-as.numeric((q/p)*((p-c.sep[Sn.sep[i],seq[s]])/(1-c.sep[Sn.sep[i],seq[s]]))^2)
W.ad<-W.ad+D^2*a.sep[Sn.sep[i],]%*%t(a.sep[Sn.sep[i],])*wi.s
}
for (i in 1:(I/H))
{
p<-c.sep[i,seq[s]]+(1-c.sep[i,seq[s]])/(1+exp(-D*t(a.sep[i,])%*%(theta-b.sep[i]*rep(1,H))))
q<-1-p
wi.s<-as.numeric((q/p)*((p-c.sep[i,seq[s]])/(1-c.sep[i,seq[s]]))^2)
```

```
Wi<-D^2*a.sep[i,]%*%t(a.sep[i,])*wi.s
Ii[i,seq[s]]<-det(solve(phi0)+Wi+W.ad)
}
Ii.choose[,seq[s]]<-Ii[,seq[s]]*(1-Sn.dummy.sep[,seq[s]])

#choose item
k<-1
while(Ii.choose[k,seq[s]]!=max(Ii.choose[,seq[s]]))
{k<-k+1}
Sn.sep<-cbind(Sn.sep,k)
Sn<-cbind(Sn,k+(seq[s]-1)*(I/H))
Sn.dummy.sep[k,seq[s]]<-1

I.final<-matrix(Ii,nrow=I,ncol=1)*matrix(Sn.dummy.sep,nrow=I,ncol=1)
}

#ability estimation step
diff<-c(1,1)
while(abs(diff[1])>10^-5 & abs(diff[2])>10^-5)
{
sum1<-0
sum2<-0
for (i in 1:length(Sn))
{
p<-c[Sn[i]]+(1-c[Sn[i]])/(1+exp(-D*t(a[Sn[i],])%*%(theta-b[Sn[i]]*rep(1,H))))
q<-1-p

vi<-as.numeric((p-c[Sn[i]])*(y[Sn[i]]-p)/((1-c[Sn[i]])*p))
wi<-as.numeric((q*(p-c[Sn[i]])*(c[Sn[i]]*y[Sn[i]]-p^2))/(p*(1-c[Sn[i]]))^2)
sum1<-sum1+vi*a[Sn[i],]
sum2<-sum2+a[Sn[i],]%*%t(a[Sn[i],])*wi
}
sum1<-matrix(sum1,nrow=H,ncol=1)
#first-order partial derivative of log-posterior
dlog.p<-D*sum1-solve(phi0)%*%(theta-mu0)
#second-order partial derivative
M.theta<-D^2*sum2-solve(phi0)

delta<-solve(M.theta)%*%dlog.p
theta.new<-theta-delta
diff<-theta.new-theta
theta<-theta.new
}

fisher[,t+(s-1)*(testlength/2)]<-I.final
I.test.s<-sum(I.final)
SE<-1/sqrt(I.test.s)
```

```
SE.iter[t+(s-1)*(testlength/2),]<-SE
theta.iter[t+(s-1)*(testlength/2),]<-theta
#cat("theta",theta,"\n")
}
}

out<-list(theta.iter,fisher,SE.iter,testlength)
names(out)<-c("theta.est","fisher","SE","testlength")
return(out)


}
```

4. คำสั่ง Kullback-Leibler information แบบผสมผสานมิติ

```
KL.mixed.fixed<-function(H,I,a,b,c,y,mu0,phi0,theta0,testlength)
{
D<-1.7
theta<-theta0
SKLi<-matrix(nrow=I,ncol=1)
Ii<-matrix(nrow=I,ncol=1)
Sn.dummy<-matrix(0,nrow=I,ncol=1)

SE<-1
round<-0

KL<-matrix(nrow=I,ncol=testlength)
SE.iter<-matrix(nrow=testlength,ncol=1)
theta.iter<-matrix(nrow=testlength,ncol=H)


for (s in 1:testlength)
{
round<-round+1
cat("round",round,"\n")

#calculate fisher information criterion
if(round<2){

for (i in 1:I)
{
p<-c[i]+(1-c[i])/(1+exp(-D*t(a[i,])%*%(theta-b[i]*rep(1,H))))
q<-1-p
wi.s<-as.numeric((q/p)*((p-c[i])/(1-c[i]))^2)
Wi<-D^2*a[i,]%*%t(a[i,])*wi.s
Ii[i]<-det(solve(phi0)+Wi)
```

```
}
}else{
W.ad<-0
for (i in 1:length(Sn))
{
p<-c[Sn[i]]+(1-c[Sn[i]])/(1+exp(-D*t(a[Sn[i],])%*%(theta-b[Sn[i]]*rep(1,H))))
q<-1-p
wi.s<-as.numeric((q/p)*((p-c[Sn[i]])/(1-c[Sn[i]]))^2)
W.ad<-W.ad+D^2*a[Sn[i],]%*%t(a[Sn[i],])*wi.s
}
for (i in 1:I)
{
p<-c[i]+(1-c[i])/(1+exp(-D*t(a[i,])%*%(theta-b[i]*rep(1,H))))
q<-1-p
wi.s<-as.numeric((q/p)*((p-c[i])/(1-c[i]))^2)
Wi<-D^2*a[i,]%*%t(a[i,])*wi.s
Ii[i]<-det(solve(phi0)+Wi+W.ad)
}
}

#calculate KL information index
n<-ifelse(round<=2,n<-1,n<-n+1)
for (i in 1:I)
{
p<-c[i]+(1-c[i])/(1+exp(-D*t(a[i,])%*%(theta-b[i]*rep(1,H))))
q<-1-p
SKLi[i]<-(2^(H-1)/3)*(3/sqrt(n))^(H+2)*((q*(p-c[i])^2)/(p*(1-c[i])))*t(a[i,])%*%a[i,]
#SKLi[i]<-(t(a[i,]%*%a[i,]))/as.numeric(abs(t(a[i,])%*%theta-b[i]))
}

SKLi.choose<-SKLi*(1-Sn.dummy)
#choose item
k<-1
while(SKLi.choose[k]!=max(SKLi.choose))
{k<-k+1}
ifelse(round<2,Sn<-k,Sn<-cbind(Sn,k))
Sn.dummy[k]<-1


#ability estimation step
diff<-c(1,1)
while(abs(diff[1])>10^-5 & abs(diff[2])>10^-5)
{
sum1<-0
sum2<-0
for (i in 1:length(Sn))
```

```
{
p<-c[Sn[i]]+(1-c[Sn[i]])/(1+exp(-D*t(a[Sn[i],])%*%(theta-b[Sn[i]]*rep(1,H))))
q<-1-p

vi<-as.numeric((p-c[Sn[i]])*(y[Sn[i]]-p)/((1-c[Sn[i]])*p))
wi<-as.numeric((q*(p-c[Sn[i]])*(c[Sn[i]]*y[Sn[i]]-p^2))/(p*(1-c[Sn[i]]))^2)
sum1<-sum1+vi*a[Sn[i],]
sum2<-sum2+a[Sn[i],]%*%t(a[Sn[i],])*wi
}
sum1<-matrix(sum1,nrow=H,ncol=1)
#first-order partial derivative of log-posterior
dlog.p<-D*sum1-solve(phi0)%*%(theta-mu0)
#second-order partial derivative
M.theta<-D^2*sum2-solve(phi0)

delta<-solve(M.theta)%*%dlog.p
theta.new<-theta-delta
diff<-theta.new-theta
theta<-theta.new
}
SKL.final<-SKLi*(Sn.dummy)
fisher.final<-Ii*(Sn.dummy)


KL[,s]<-SKL.final
I.test.KL<-sum(SKL.final)
I.test.fisher<-sum(fisher.final)
SE.fisher<-1/sqrt(I.test.fisher)
SE<-SE.fisher
SE.iter[s,]<-SE
theta.iter[s,]<-theta
KL.test<-I.test.KL
}
out<-list(theta.iter,KL.test,SE.iter,testlength)
names(out)<-c("theta.est","KL.test","SE","testlength")
return(out)
}
```

5. คำสั่ง Kullback-Leibler information แบบจำแนกมิติ

```
#KL seperate
KL.sep<-function(H,I,a,b,c,y,mu0,phi0,theta0,testlength)
{
b.sep<-matrix(nrow=I/2,ncol=H)
c.sep<-matrix(nrow=I/2,ncol=H)

theta.iter<-matrix(nrow=testlength,ncol=H)
```

```
KL<-matrix(nrow=I,ncol=testlength)
I.final<-0

for (s in 1:H)
{
b.sep[,s]<-b[(1+(s-1)*(I/H)):(s*I/H)]
c.sep[,s]<-c[(1+(s-1)*(I/H)):(s*I/H)]
}

D<-1.7
theta<-theta0
SE<-1
Ii<-matrix(0,nrow=I/H,ncol=H)
Ii.choose<-matrix(0,nrow=I/H,ncol=H)
Sn<-0
Sn.dummy.sep<-matrix(0,nrow=I/H,ncol=H)

seq<-sample(subjects<-c(1:H),H,replace=FALSE)
n<-0
for (s in 1:H)
{
Sn.sep<-0
a.sep<-a[(1+(seq[s]-1)*(I/H)):(seq[s]*I/H),]

for (t in 1:(testlength/2))
{
n<-n+1
for (i in 1:(I/H))
{
#item selection step
p<-c.sep[i,seq[s]]+(1-c.sep[i,seq[s]])/(1+exp(-D*t(a.sep[i,])%*%(theta-b.sep[i,seq[s]]*rep(1,H))))
q<-1-p
Ii[i,seq[s]]<-(2^(H-1)/3)*(3/sqrt(n))^(H+2)*((q*(p-c.sep[i,seq[s]])^2)/(p*(1-
c.sep[i,seq[s]])))*t(a.sep[i,])%*%a.sep[i,]
}

Ii.choose[,seq[s]]<-Ii[,seq[s]]*(1-Sn.dummy.sep[,seq[s]])
#choose item
k<-1
while(Ii.choose[k,seq[s]]!=max(Ii.choose[,seq[s]]))
{k<-k+1}
ifelse(t==1,Sn.sep<-k,Sn.sep<-cbind(Sn.sep,k))
ifelse(n==1,Sn<-k+(seq[s]-1)*(I/H),Sn<-cbind(Sn,k+(seq[s]-1)*(I/H)))
Sn.dummy.sep[k,seq[s]]<-1

I.final<-matrix(Ii,nrow=I,ncol=1)*matrix(Sn.dummy.sep,nrow=I,ncol=1)
```

```
#ability estimation step
diff<-c(1,1)
while(abs(diff[1])>10^-5 & abs(diff[2])>10^-5)
{
sum1<-0
sum2<-0
for (i in 1:length(Sn))
{
p<-c[Sn[i]]+(1-c[Sn[i]])/(1+exp(-D*t(a[Sn[i],])%*%(theta-b[Sn[i]]*rep(1,H))))
q<-1-p

vi<-as.numeric((p-c[Sn[i]])*(y[Sn[i]]-p)/((1-c[Sn[i]])*p))
wi<-as.numeric((q*(p-c[Sn[i]])*(c[Sn[i]]*y[Sn[i]]-p^2))/(p*(1-c[Sn[i]]))^2)
sum1<-sum1+vi*a[Sn[i],]
sum2<-sum2+a[Sn[i],]%*%t(a[Sn[i],])*wi
}
sum1<-matrix(sum1,nrow=H,ncol=1)
#first-order partial derivative of log-posterior
dlog.p<-D*sum1-solve(phi0)%*%(theta-mu0)
#second-order partial derivative
M.theta<-D^2*sum2-solve(phi0)

delta<-solve(M.theta)%*%dlog.p
theta.new<-theta-delta
diff<-theta.new-theta
theta<-theta.new
}
KL[,t+(s-1)*(testlength/2)]<-I.final
I.test.s<-sum(I.final)
theta.iter[t+(s-1)*(testlength/2),]<-theta
}
}

out<-list(theta.iter,I.test.s)
names(out)<-c("theta.est","KL")
return(out)
}
```

*** หมายเหตุ** คำสั่งที่นำเสนอเป็นเพียงตัวอย่างในส่วนของการจำลองข้อมูลเท่านั้น สำหรับการใช้ข้อมูลทุติยภูมิ จะมีลักษณะคล้ายคลึงกัน แต่เปลี่ยนจำนวนมิติที่ศึกษา และปรับการดึงข้อมูลจากข้อมูล PAT2 มาใช้แทน

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวทัศน์ศิรินทร์ สว่างบุญ เกิดวันที่ 13 สิงหาคม พ .ศ. 2523 ที่อำเภอเมือง จังหวัดร้อยเอ็ด จบการศึกษาครุศาสตรบัณฑิต ภาควิชามัธยมศึกษา สาขาวิชาคณิตศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ปีการศึกษา 2545 จบการศึกษาครุศาสตรมหาบัณฑิต ภาควิชาวิจัยและจิตวิทยาการศึกษา สาขาวิชาวิจัยการศึกษา ในปีการศึกษา 2548 และเข้าศึกษาต่อในหลักสูตรครุศาสตรดุษฎี บัณฑิต ภาควิชาวิจัยและจิตวิทยาการศึกษา สาขาวิชาการวัดและประเมินผลการศึกษา ในปีการศึกษา 2550 พร้อมทั้ง ไปศึกษาวิจัย ณ มหาวิทยาลัยโทโฮกุ เมืองเซ็ นได ประเทศญี่ปุ่น ในปีการศึกษา 2553 ด้วยทุนของ โครงการ เครือข่ายเชิงกลยุทธ์เพื่อการผลิตและพัฒนาอาจารย์ในสถาบันอุดมศึกษา สำนักงานคณะกรรมการการอุดมศึกษา และได้ทุนสนับสนุนการวิจัย “ทุน 90 ปี จุฬาลงกรณ์มหาวิทยาลัย ” จากกองทุนรัชดาภิเษกสมโภช จุฬาลงกรณ์มหาวิทยาลัย