



รายงานการวิจัย

เรื่อง

การสร้างแบบทดสอบที่มีความเที่ยงและความตรง
ตามทฤษฎีการตอบข้อสอบจากคลังข้อสอบ

THE RELIABILITY AND VALIDITY OF TEST CONSTRUCTION
FROM ITEM BANK BASED ON ITEM RESPONSE THEORY

ผู้ช่วยศาสตราจารย์ ดร.สุวิมล กฤษณกุลหาสน์

ภาควิชาการประเมินและการวิจัย คณะศึกษาศาสตร์ มหาวิทยาลัยรามคำแหง

ได้รับเงินทุนอุดหนุนโครงการวิจัยของคณะศึกษาศาสตร์ ประจำปีงบประมาณ พ.ศ. 2559

สารบัญ

บทที่	หน้า
คำนำ	ก
บทคัดย่อไทย	ข
บทคัดย่ออังกฤษ	ค
1 บทนำ.....	1
ความเป็นมาและความสำคัญของปัญหา.....	1
วัตถุประสงค์การวิจัย.....	3
ขอบเขตการวิจัย.....	4
นิยามศัพท์เฉพาะ.....	5
ประโยชน์ที่รับจากการวิจัย.....	8
กรอบแนวคิดการวิจัย.....	9
2 เอกสารและงานวิจัยที่เกี่ยวข้อง.....	11
ตอนที่ 1 แนวคิดและหลักการของโมเดลเชิงเส้นสำหรับการออกแบบการสร้าง	
แบบทดสอบและงานวิจัยที่เกี่ยวข้อง	12
ตอนที่ 2 แนวคิดและหลักการการตรวจสอบอัตราการใช้ข้อสอบซ้ำระหว่าง	
แบบทดสอบและงานวิจัยที่เกี่ยวข้อง.....	38
ตอนที่ 3 การใช้โปรแกรม R.....	49
ตอนที่ 4 ผลการสังเคราะห์เอกสารและงานวิจัยที่เกี่ยวข้องกับการสร้าง	
แบบทดสอบคู่ขนานมากกว่า 1 ฉบับตามทฤษฎีการตอบข้อสอบ.....	61
3 วิธีดำเนินการวิจัย.....	64
ขั้นตอนที่ 1 การสร้างและตรวจสอบคุณภาพของแบบทดสอบฉบับอ้างอิง.....	65
ขั้นตอนที่ 2 การสร้างคลังข้อสอบตามคุณสมบัติของแบบทดสอบฉบับอ้างอิง	
ด้วยการจำลองข้อมูลการสอบด้วยโปรแกรม WINGEN.....	74
สถิติที่ใช้ในการวิเคราะห์ข้อมูล.....	80
4 ผลการวิจัย.....	81
ตอนที่ 4.1 โปรแกรมการสร้างแบบทดสอบคู่ขนาน และผลการคัดเลือกข้อสอบ	
จากคลังข้อสอบด้วยโปรแกรมเพื่อสร้างแบบทดสอบคู่ขนานมี	
คุณภาพด้านความเที่ยง และความตรงตามทฤษฎีการตอบสนอง	
ข้อสอบ.....	82

บทที่		หน้า
	ตอนที่ 4.2 ผลการตรวจสอบความถูกต้องของโปรแกรมการคัดเลือกข้อสอบ ด้านคุณลักษณะเฉพาะของความเป็นคู่ขนานของแบบสอบตาม ระดับค่าฟังก์ชันคุณลักษณะของแบบทดสอบฉบับอ้างอิง.....	98
	ตอนที่ 4.3 ผลการตรวจสอบความถูกต้องของโปรแกรมการคัดเลือกด้านการใช้ ข้อสอบชำระระหว่างแบบทดสอบคู่ขนานตามตำแหน่งความสามารถ ของผู้สอบของแบบทดสอบฉบับอ้างอิง.....	101
5	สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ.....	105
	สรุปผลการวิจัย.....	107
	อภิปรายผลการวิจัย.....	114
	ข้อเสนอแนะ.....	117
	รายการอ้างอิง.....	118
	ภาคผนวก.....	119

คำนำ

รายงานการวิจัยเรื่อง “การสร้างแบบทดสอบที่มีความเที่ยงและความตรงตามทฤษฎีการตอบข้อสอบจากคลังข้อสอบ” จัดทำขึ้นเนื่องจากผู้วิจัยตระหนักถึงความสำคัญ และประโยชน์ของการสร้างแบบทดสอบคู่ขนานจากคลังข้อสอบ มหาวิทยาลัยรามคำแหงเป็นมหาวิทยาลัยในระบบตลาดวิชา การจัดการทดสอบเป็นงานที่มีความสำคัญ ทั้งนี้เพราะนักศึกษาของมหาวิทยาลัยรามคำแหงมีจำนวนมาก ลักษณะการผลิตจึงต้องมีความเข้มงวดมิเช่นนั้นจะได้บัณฑิตจำนวนมากแต่ไม่มีคุณภาพ ระบบการทดสอบของมหาวิทยาลัยรามคำแหงมีการสอบประจำภาคและการสอบซ่อมจึงมีการสอบถึงปีละ 5 ครั้ง แต่ด้วยจำนวนนักศึกษาที่มีมากทำให้การจัดการสอบมีความเข้มงวด โดยเฉพาะวิชาพื้นฐานที่มีนักศึกษาลงทะเบียนเป็นหมื่นคน ในห้องสอบห้องหนึ่งจึงต้องใช้แบบทดสอบ 2 รูปแบบสลับกันระหว่างแถว ซึ่งทำให้ต้องผลิตข้อสอบถึง 10 ชุดในแต่ละปีการศึกษา การใช้แบบทดสอบชุดเดียวกันทุกครั้งจะทำให้ข้อสอบรั่วเท่ากับเป็นการสนับสนุนให้นักศึกษาซื้อข้อสอบเก่ามาอ่านมากกว่าการอ่านตำราเรียน ในความเป็นจริงแล้วลักษณะของแบบทดสอบทุกฟอร์มควรเป็นแบบทดสอบที่คู่ขนานกัน นั่นคือมีเนื้อหาการวัดเดียวกัน มีความยากง่าย อำนาจจำแนก ความเที่ยง (reliability) และความตรง (validity) ใกล้เคียงกัน รวมทั้งต้องมีความปลอดภัยทั้งในด้านการจัดเก็บและอัตราการใช้ข้อสอบซ้ำระหว่างฉบับ (test overlap rate) อยู่ในระดับต่ำ จึงจะสามารถวัดผลได้อย่างแม่นยำ ตลอดจนขึ้นต้นถึงความยุติธรรมที่นักศึกษาจะได้รับ อีกทั้งคลังข้อสอบที่ดีนอกจากจะแสดงถึงความก้าวหน้าทางการบริหารวิชาการแล้ว ยังแสดงให้เห็นได้อีกด้วยว่ามีระบบบริหารงานวิชาการที่ดี และมีการดำเนินการเรียนการสอนที่ยึดหลักเกณฑ์หรือหลักการในศาสตร์ทางการศึกษาโดยตรงมาน้อยเพียงใด ผู้วิจัยจึงสนใจที่จะศึกษาวิธีการคัดเลือกข้อสอบในคลังข้อสอบเพื่อสร้างเป็น โปรแกรมการคัดเลือกข้อสอบที่จะทำได้แบบทดสอบคู่ขนานมากกว่า 1 ฉบับที่มีทั้งความเที่ยง ความตรงตามทฤษฎีการตอบข้อสอบ ประโยชน์ของการวิเคราะห์คุณภาพข้อสอบตามทฤษฎีการตอบข้อสอบ (Item response theory) ทำให้ได้ค่าพารามิเตอร์ไม่แปรเปลี่ยนไปตามกลุ่มผู้สอบ จึงทำให้ผลจากการวิจัยครั้งนี้จะนำไปสู่วิธีการคัดเลือกข้อสอบจากคลังข้อสอบด้วย โปรแกรมคอมพิวเตอร์ ทำให้ได้แบบทดสอบคู่ขนานตามทฤษฎีการตอบข้อสอบ (Item Response Theory: IRT) ที่มีประสิทธิภาพต่อไป และที่สำคัญยังเป็นพื้นฐานของการจัดการสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (Computer Adaptive Testing: CAT) อีกด้วย

รายงานการวิจัยเล่มนี้สำเร็จไปด้วยดีด้วยความอนุเคราะห์จากคณะศึกษาศาสตร์ มหาวิทยาลัยรามคำแหง ผู้วิจัยขอขอบคุณคณะศึกษาศาสตร์ มหาวิทยาลัยรามคำแหงที่ให้ทุนสนับสนุนจากเงินรายได้ผู้วิจัยหวังเป็นอย่างยิ่งว่าผลที่ได้จากการวิจัยครั้งนี้จะเป็นแนวทางในการนำโปรแกรมที่สร้างขึ้นไปใช้ในการสร้างและพัฒนาแบบทดสอบคู่ขนานต่อไป

สุวิมล กฤษณกุลหาสน์

การสร้างแบบทดสอบที่มีความเที่ยงและความตรงตามทฤษฎีการตอบข้อสอบจากคลังข้อสอบ

สุวิมล กฤษศยาสาสน์

บทคัดย่อ

การศึกษานี้มีวัตถุประสงค์หลักเพื่อสร้างแบบทดสอบคู่ขนานจากคลังข้อสอบที่มีคุณภาพด้านความเที่ยงและความตรงตามทฤษฎีการตอบข้อสอบ โดยมีวัตถุประสงค์ย่อย คือ 1) เพื่อสร้างโปรแกรมการคัดเลือกข้อสอบจากคลังข้อสอบ 2) เพื่อตรวจสอบความถูกต้องของโปรแกรมการคัดเลือกข้อสอบด้านคุณลักษณะเฉพาะของความเป็นคู่ขนานของแบบทดสอบตามตำแหน่งความสามารถของผู้สอบของแบบทดสอบฉบับอ้างอิง และ 3) เพื่อตรวจสอบความถูกต้องของโปรแกรมการคัดเลือกข้อสอบด้านการใช้ข้อสอบชำระระหว่างแบบทดสอบคู่ขนานตามตำแหน่งความสามารถของผู้สอบของแบบทดสอบฉบับอ้างอิง การวิจัยครั้งนี้ใช้วิธีการศึกษาโดยการจำลองข้อมูลการสอบที่มีคำตอบแบบ 0 และ 1 บรรจุในคลังข้อสอบ โดยข้อมูลการสอบที่จำลองขึ้นอยู่บนพื้นฐานของผลการทดสอบจากรายวิชา MER 2003 ความรู้เบื้องต้นเกี่ยวกับสถิติและวิจัยทางการศึกษา จากจำนวนผู้สอบ 2,800 คน สถิติที่ใช้คือ การวิเคราะห์ความเป็นเอกมิติ โมเดลการตอบสนองข้อสอบแบบ 3 พารามิเตอร์ การประเมินคุณลักษณะเฉพาะของความเป็นคู่ขนาน และการวิเคราะห์อัตราการใช้ข้อสอบชำระระหว่างแบบทดสอบเฉลี่ย ผลการศึกษาพบว่า 1) โปรแกรมการคัดเลือกข้อสอบสามารถสร้างแบบทดสอบ จำนวน 5 ฉบับที่มีความถูกต้องด้านความตรงตามเนื้อหา 2) โปรแกรมการคัดเลือกข้อสอบสามารถสร้างแบบทดสอบ จำนวน 4 ฉบับที่มีความถูกต้องด้านความเที่ยงตามทฤษฎีการตอบข้อสอบที่มีความเที่ยงใกล้เคียงกับแบบทดสอบฉบับอ้างอิง 3) โปรแกรมการคัดเลือกข้อสอบสามารถสร้างแบบทดสอบคู่ขนาน จำนวน 2 ฉบับที่มีความถูกต้องด้านคุณลักษณะเฉพาะของความเป็นคู่ขนาน ณ ตำแหน่งความสามารถของผู้สอบ 2 ตำแหน่ง 4) โปรแกรมการคัดเลือกข้อสอบสามารถสร้างแบบทดสอบคู่ขนาน จำนวน 1 ฉบับ ที่มีความถูกต้องด้านคุณลักษณะเฉพาะของความเป็นคู่ขนานและด้านอัตราการใช้ข้อสอบชำระระหว่างแบบทดสอบเฉลี่ยต่ำกว่าร้อยละ 25 ณ ตำแหน่งความสามารถของผู้สอบ 2 ตำแหน่ง 5) โปรแกรมการคัดเลือกข้อสอบสามารถสร้างแบบทดสอบคู่ขนาน จำนวน 1 ฉบับที่มีความถูกต้องด้านคุณลักษณะเฉพาะของความเป็นคู่ขนานและด้านอัตราการใช้ข้อสอบชำระระหว่างแบบทดสอบเฉลี่ยต่ำกว่าร้อยละ 25 ณ ตำแหน่งความสามารถของผู้สอบ 3 ตำแหน่ง และ 6) โปรแกรมการคัดเลือกข้อสอบสามารถสร้างแบบทดสอบคู่ขนาน จำนวน 1 ฉบับที่มีความถูกต้องด้านคุณลักษณะเฉพาะของความเป็นคู่ขนานและด้านอัตราการใช้ข้อสอบชำระระหว่างแบบทดสอบเฉลี่ยต่ำกว่าร้อยละ 25 ณ ตำแหน่งความสามารถของผู้สอบ 4 ตำแหน่ง

คำสำคัญ คลังข้อสอบ แบบทดสอบคู่ขนาน คุณลักษณะเฉพาะของความเป็นคู่ขนาน โมเดลการตอบสนองข้อสอบแบบ 3 พารามิเตอร์ การรวบรวมข้อสอบแบบอัตโนมัติ อัตราการใช้ข้อสอบชำระ

THE RELIABILITY AND VALIDITY OF TEST CONSTRUCTION FROM ITEM BANK BASED ON ITEM RESPONSE THEORY

SUWIMON KRITKHARUEHART

Abstract

The main purpose of this research was to construct parallel test forms from item bank based on Item Response Theory. The minor purposes were to programming for item selection form item bank with reliability and validity based on Item Response Theory, to verify the accuracy of item selection program with the specific characteristics of the test parallelism at points of examinee ability, and to verify the accuracy of item selection program with the specific characteristics of low test overlap rate at points of examinee ability. This research was conducted by simulating binary data in item bank, the simulated test data is based on real testing conducted with 2,800 students who enrolled in the course "Introduction to Statistics and Research in Education". The statistics use for data analysis included dimensionality, 3-parameters logistic item response model, test parallelism, and test overlap rate.

The research results revealed that the program could construct as follow: 1) 5 tests with content validity 2) 4 tests with reliability in terms of item response theory that are close to the target test 3) 2 tests with parallelism test at 2 examinee ability points 4) 1 test with parallelism test and low test overlap rate at 2 examinee ability points 5) 1 test with parallelism test and low test overlap rate at 3 examinee ability points and 6) 1 test with parallelism test and low test overlap rate at 4 examinee ability points.

Keywords: item bank, parallel forms, parallelism test, 3 parameters IRT, automated test assembly

บทที่ 1

บทนำ

ความเป็นมาและความสำคัญของปัญหา

ในระยะสิบปีที่ผ่านมาสถาบันอุดมศึกษาในประเทศไทยมีการเปลี่ยนแปลงอย่างเห็นได้ชัด นับตั้งแต่การมีสถาบันอุดมศึกษาหรือมหาวิทยาลัยเกิดขึ้นหลายแห่ง มีรูปแบบการบริหารจัดการที่เปลี่ยนไป มีทั้งแบบที่อยู่ในการกำกับของรัฐและแบบที่ยังเป็นหน่วยงานราชการ มีการจัดการเรียนการสอนที่แตกต่างจากแบบดั้งเดิมที่มีการจัดการศึกษาอยู่ในที่ดั่ง แล้วยังมีการจัดการศึกษานอกที่ดั่ง มีการใช้เทคโนโลยีสารสนเทศ และการสื่อสารเข้ามาจัดการเรียนการสอน มีหลักสูตรและการสอนและการตรวจสอบ มีระบบการประกันคุณภาพ มีเงื่อนไขของการจัดการเรียนการสอนที่มากขึ้น การจัดทำหลักสูตรเกี่ยวข้องกับองค์วิชาชีพ และสถานภาพของอาจารย์และบุคลากรทางการศึกษาที่เปลี่ยนไป เป็นต้น (กฤษมันต์ วัฒนารงค์, 2556)

เมื่อพิจารณาถึงลักษณะของมหาวิทยาลัยในประเทศไทยจะพบว่ามี 3 ระบบ คือ ระบบปิด ระบบเปิด และตลาดวิชา ซึ่งตลาดวิชาเกิดขึ้นแห่งแรกในประเทศไทย คือ มหาวิทยาลัยวิชาธรรมศาสตร์และการเมือง ก่อตั้งขึ้นเมื่อ พ.ศ. 2477 จัดการเรียนการสอนแบบตลาดวิชา คือเดินเข้าสอบออก ซึ่งต่างจากมหาวิทยาลัยเปิด คือ ไม่มีชั้นเรียน และสอนโดยใช้สื่อเป็นหลัก ขณะที่มหาวิทยาลัยปิดก็บังคับให้มาเรียน ตลาดวิชาจึงนำทั้งสองแบบมาบูรณาการเข้าด้วยกันคือ ไม่บังคับการเข้าชั้นเรียนสามารถเรียนและสอนด้วยตัวเองได้ แต่นักศึกษาที่ต้องการมาเรียนมหาวิทยาลัยก็จัดห้องเรียนให้เหมือนมหาวิทยาลัยปิดทุกประการ (กิม ไชยแสนสุข, 2553, หน้า 3)

มหาวิทยาลัยรามคำแหงเป็นมหาวิทยาลัยตลาดวิชาที่เปิดโอกาสทางการศึกษาให้แก่ผู้ที่ใฝ่รู้ใฝ่เรียนโดยเปิดการเรียนการสอนใน 2 ระบบคือ ระบบเปิดและระบบปิด รับนักศึกษาไม่จำกัดจำนวน ซึ่งปัจจุบันได้ปรับปรุงหลักสูตรให้มีคุณภาพและมาตรฐานมากยิ่งขึ้น พร้อมทั้งมีสื่อการเรียนการสอนที่ทันสมัยรองรับการเปลี่ยนแปลงของโลกยุคโลกาภิวัตน์ และเตรียมความพร้อมเข้าสู่ประชาคมอาเซียนในปี 2558 ซึ่งเน้นให้นักศึกษามีความรู้ความสามารถทั้งด้านวิชาการ ประสบการณ์การดำรงชีวิตระหว่างเรียนในมหาวิทยาลัยและมีวินัย เพื่อให้นักศึกษาจบออกไปประกอบอาชีพและอยู่ร่วมกันกับผู้อื่นในสังคมได้อย่างมีประสิทธิภาพสมเป็นบัณฑิตผู้มีความรู้คู่คุณธรรม (วุฒิสักดิ์ ลากเจริญทรัพย์, 2556, หน้า 11-12)

การวัดและประเมินผลการศึกษาเป็นกระบวนการหนึ่งที่มีส่วนช่วยให้การจัดการศึกษาในระดับอุดมศึกษาประสบความสำเร็จ ทำให้สามารถผลิตบัณฑิตที่มีคุณภาพออกสู่ตลาดแรงงาน โดยการวัดและประเมินผล การเรียนรู้ มุ่งเน้นกระบวนการเก็บรวบรวม ตรวจสอบ ตีความผลการเรียนรู้ และพัฒนาการด้านต่าง ๆ ของผู้เรียนตามมาตรฐานการเรียนรู้หรือตัวชี้วัดของหลักสูตรนำผลไปปรับปรุงพัฒนาการจัดการเรียนรู้และใช้เป็น

ข้อมูลสำหรับการตัดสินผลการเรียน สถานศึกษาต้องมีกระบวนการจัดการที่เป็นระบบ เพื่อให้การดำเนินการวัดและประเมินผลการเรียนรู้เป็นไปอย่างมีคุณภาพและประสิทธิภาพ และให้ผลการประเมินที่ตรงตามความรู้ความสามารถที่แท้จริงของผู้เรียน ถูกต้องตามหลักการวัดและประเมินผลการเรียนรู้ รวมทั้งสามารถรองรับการประเมินภายในและการประเมินภายนอกตามระบบประกันคุณภาพการศึกษาได้ (กระทรวงศึกษาธิการ, 2554, หน้า 10)

องค์ประกอบที่สำคัญของการวัดและประเมินผลการเรียนรู้ คือ ข้อสอบและแบบทดสอบที่ใช้ในการวัด ในส่วนของมหาวิทยาลัยรามคำแหงซึ่งจัดเป็นมหาวิทยาลัยในระบบตลาดวิชา การจัดการสอบจัดเป็นงานที่มีความสำคัญ ทั้งนี้เพราะนักศึกษาของมหาวิทยาลัยรามคำแหงมีจำนวนมาก ลักษณะการผลิตจึงต้องมีความเข้มงวดไม่เช่นนั้นจะได้บัณฑิตจำนวนมากแต่ไม่มีคุณภาพ ระบบการสอบของมหาวิทยาลัยรามคำแหงมีการสอบประจำภาคและการสอบซ่อมจึงมีการสอบถึงปีละ 5 ครั้ง แต่ด้วยจำนวนนักศึกษาที่มีมากทำให้การจัดการสอบมีความเข้มงวด โดยเฉพาะวิชาพื้นฐานที่มีนักศึกษาลงทะเบียนเป็นหมื่นคน ในห้องสอบห้องหนึ่งจึงต้องใช้แบบทดสอบ 2 ฟอรัมสลับกันระหว่างแถว ซึ่งทำให้ต้องผลิตข้อสอบถึง 10 ชุดในแต่ละปี การใช้แบบทดสอบชุดเดียวกันทุกครั้งจะทำให้ข้อสอบซ้ำเท่ากับเป็นการสนับสนุนให้นักศึกษาซื้อข้อสอบเก่ามาอ่านมากกว่าการอ่านตำราเรียน ทำให้เกิดข้อสงสัยเกี่ยวกับคุณภาพของบัณฑิตที่ผลิตออกไป ในความเป็นจริงแล้วลักษณะของแบบทดสอบทุกชุดควรเป็นแบบทดสอบที่คู่ขนานกัน นั่นคือมีเนื้อหาการวัดเดียวกัน มีความยากง่าย อำนาจจำแนก ความเที่ยง (reliability) และความตรง (validity) ใกล้เคียงกัน จึงจะสามารถวัดผลได้อย่างแม่นยำ ตลอดจนยืนยันถึงความยุติธรรมที่นักศึกษาจะได้รับ

การสร้างแบบทดสอบคู่ขนานมากกว่า เหนือจำเป็นจะต้องมีการสร้างคลังข้อสอบขึ้น และมีการพัฒนาระบบการจัดเก็บในคลังข้อสอบเพื่อให้เกิดความตรงตามเนื้อหา (content validity) รวมทั้งต้องมีความปลอดภัยทั้งในด้านการจัดเก็บและอัตราการใช้ข้อสอบซ้ำระหว่างฉบับ (test overlap rate) อยู่ในระดับต่ำ (Chen and Lei, 2010; Lin, 2010; Chang and Zhang, 2002) นอกจากนี้การพัฒนาคลังข้อสอบ (item bank) ที่ดียังแสดงถึงความก้าวหน้าทางการบริหารวิชาการของมหาวิทยาลัยแล้ว และยังแสดงให้เห็นได้อีกด้วยว่ามหาวิทยาลัยมีระบบบริหารงานวิชาการที่ดี และมีการดำเนินการเรียนการสอนที่ยึดหลักเกณฑ์หรือหลักการในศาสตร์ทางการศึกษาโดยตรงมากขึ้นเพียงใด คลังข้อสอบที่ดีจะทำงานสัมพันธ์กับงานวัดผลของมหาวิทยาลัยโดยตรง เมื่อคลังข้อสอบสามารถพัฒนาแบบทดสอบของมหาวิทยาลัยขึ้นเป็นแบบทดสอบมาตรฐานของตนขึ้นใช้ได้ จุดเน้นของการสร้างคลังข้อสอบอยู่ที่การวัดผลการเรียนการสอน ส่วนที่เป็นผลพลอยได้คือ การพัฒนาการทดสอบ การปรับปรุงระบบการสอบ และการเพิ่มจำนวนชุดแบบทดสอบเพื่อลดจำนวนข้อสอบที่ซ้ำกันในการสอบแต่ละครั้ง การดำเนินงานของคลังข้อสอบจะเกิดประโยชน์สูงสุดต่อเมื่อมีการใช้คลังข้อสอบนั้นอย่างมีประสิทธิภาพ

ดังนั้นคุณภาพที่สำคัญของแบบทดสอบคู่ขนานที่สร้างขึ้นจากคลังข้อสอบ ก็จะต้องมีทั้งความเที่ยง ความตรง และอัตราการใช้ข้อสอบซ้ำเป็นไปตามเกณฑ์ที่ยอมรับได้ โดยความเที่ยงจะบอกถึงความน่าเชื่อถือของการประมาณค่าความสามารถของผู้สอบ ความตรงในกรณีการวัดผลสัมฤทธิ์ทางการเรียนจะเน้นที่ความตรงเชิงเนื้อหา ส่วนอัตราการใช้ข้อสอบซ้ำแสดงถึงความปลอดภัยในการจัดเก็บข้อสอบและการปกป้องการเปิดเผยของข้อสอบที่ถูกใช้ซ้ำเกินค่าที่ควรจะเป็นของข้อสอบในคลังข้อสอบ การสุ่มข้อสอบจากคลังข้อสอบด้วยการสุ่มอย่างง่าย (simple random sampling) เพื่อนำมาสร้างแบบทดสอบมีข้อบกพร่องที่สำคัญ คือ ทำให้มีโอกาสได้ข้อสอบที่ไม่กระจายไปครอบคลุมทุกเนื้อหา ทำให้ขาดความตรงตามเนื้อหาได้ ขณะเดียวกันก็ไม่สามารถยืนยันความเที่ยงที่จะเกิดขึ้นว่าเป็นไปตามที่ต้องการหรือไม่ เพราะคุณภาพของแบบทดสอบที่อาศัยทฤษฎีแบบดั้งเดิมทำให้ค่าพารามิเตอร์ไม่ว่าจะเป็นค่าความยาก ค่าอำนาจจำแนก และสัมประสิทธิ์ความเที่ยงเปลี่ยนแปลงไปตามกลุ่มผู้สอบในแต่ละครั้ง

ด้วยสาเหตุที่กล่าวมาข้างต้น ผู้วิจัยจึงสนใจที่จะศึกษาถึงการจากระบบการจัดเก็บในคลังข้อสอบ สนใจวิธีการคัดเลือกข้อสอบในคลังข้อสอบที่จะทำให้แบบทดสอบคู่ขนานมากกว่า 1 ฉบับที่ได้มีทั้งความเที่ยงและความตรง ตลอดจนการวิเคราะห์คุณภาพของข้อสอบตามทฤษฎีการตอบข้อสอบ (Item response theory) ที่ให้ค่าพารามิเตอร์ไม่แปรเปลี่ยนไปตามกลุ่มตัวอย่าง การวิจัยครั้งนี้จะนำไปสู่การสร้างคลังข้อสอบ การทำแบบทดสอบคู่ขนานตามทฤษฎีการตอบข้อสอบ (Item Response Theory: IRT) ที่มีประสิทธิภาพต่อไป และที่สำคัญยังเป็นพื้นฐานของการจัดการสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (Computer Adaptive Testing: CAT) อีกด้วย

วัตถุประสงค์การวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์หลักเพื่อสร้างแบบทดสอบคู่ขนานจากคลังข้อสอบที่มีคุณภาพด้านความเที่ยงและความตรงตามทฤษฎีการตอบข้อสอบ โดยมีวัตถุประสงค์ย่อย คือ

- 1) เพื่อสร้างโปรแกรมการคัดเลือกข้อสอบจากคลังข้อสอบเพื่อสร้างแบบทดสอบคู่ขนานที่มีคุณภาพด้านความเที่ยง (reliability) และความตรง (validity) ตามทฤษฎีการตอบข้อสอบ
- 2) เพื่อตรวจสอบความถูกต้องของ โปรแกรมการคัดเลือกข้อสอบด้านคุณลักษณะเฉพาะของความเป็นคู่ขนาน (parallelism specification) ของแบบทดสอบตามตำแหน่งความสามารถของผู้สอบของแบบทดสอบฉบับอ้างอิง
- 3) เพื่อตรวจสอบความถูกต้องของ โปรแกรมการคัดเลือกข้อสอบด้านการใช้ข้อสอบซ้ำระหว่างแบบทดสอบคู่ขนาน (test overlap) ตามตำแหน่งความสามารถของผู้สอบของแบบทดสอบฉบับอ้างอิง

ขอบเขตการวิจัย

1. การวิจัยครั้งนี้ใช้วิธีการศึกษาโดยการจำลองข้อมูลการสอบที่มีคำตอบแบบ 0 และ 1 บรรจุคลังข้อสอบ โดยข้อมูลการสอบที่จำลองขึ้นอยู่บนพื้นฐานของผลการทดสอบจากรายวิชา MER 2003 ความรู้เบื้องต้นเกี่ยวกับสถิติและวิจัยทางการศึกษา จำนวนผู้สอบ 2,800 คน

2. ความตรงตามเนื้อหาได้จาก

- 1) การจัดทำตารางกำหนดลักษณะเฉพาะรายวิชา (table of specification)
- 2) จำลองข้อสอบคู่ขนานในแต่ละเนื้อหา

3. กำหนดให้เป็นแบบทดสอบปรนัยแบบเลือกตอบ จำนวน 4 ตัวเลือก

4. ตัวแปรในการวิจัย

ตัวแปรอิสระ คือ โปรแกรมการคัดเลือกข้อสอบที่มีการกำหนดค่าฟังก์ชันสารสนเทศตามตำแหน่งความสามารถของผู้สอบ (θ_p) ของแบบทดสอบฉบับอ้างอิง (Referenced Test Information Function: RTIF) จำแนกเป็น

- 1) ค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง 2 ตำแหน่ง คือ
 - 1.1) $RTIF_{\theta_1=0.00}$ เท่ากับ 16.80
 - 1.2) $RTIF_{\theta_2=1.50}$ เท่ากับ 10.00
- 2) ค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง 3 ตำแหน่ง คือ
 - 2.1) $RTIF_{\theta_1=-1.50}$ เท่ากับ 10.50
 - 2.2) $RTIF_{\theta_2=0.00}$ เท่ากับ 16.80
 - 2.3) $RTIF_{\theta_3=1.50}$ เท่ากับ 10.00
- 3) ค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง 4 ตำแหน่ง คือ
 - 3.1) $RTIF_{\theta_1=-1.50}$ เท่ากับ 10.50
 - 3.2) $RTIF_{\theta_2=0.00}$ เท่ากับ 16.08
 - 3.3) $RTIF_{\theta_3=1.50}$ เท่ากับ 10.00
 - 3.4) $RTIF_{\theta_4=2.50}$ เท่ากับ 5.00

ตัวแปรตาม คือ ความถูกต้องของโปรแกรมการคัดเลือกข้อสอบ จำแนกเป็น

- 1) ความตรงตามเนื้อหา
- 2) ความเที่ยง
- 3) ค่าฟังก์ชันสารสนเทศของแบบทดสอบ

- 4) คุณลักษณะเฉพาะของความเป็นคู่ขนาน
- 5) อัตราการใช้ข้อสอบชำระระหว่างแบบทดสอบ

นิยามศัพท์เฉพาะ

1. **คลังข้อสอบ (Item Bank)** หมายถึง การรวบรวมข้อสอบด้วยระบบคอมพิวเตอร์ โดยที่ข้อสอบในคลังวัดตรงตามเนื้อหารายวิชา MER 2003 ความรู้เบื้องต้นเกี่ยวกับสถิติและวิจัยทางการศึกษา คลังข้อสอบมีการกำหนดรหัสข้อสอบ รหัสเนื้อหาย่อยที่ใช้วัดผลการเรียนรู้ และระบุค่าพารามิเตอร์ของข้อสอบแบบ 3 พารามิเตอร์ คือ ค่าอำนาจจำแนก (a) ค่าความยาก (b) ค่าการเดา (c) สำหรับในการวิจัยครั้งนี้ดำเนินการสร้างคลังข้อสอบด้วยการจำลองข้อสอบเพิ่มเติมจากข้อสอบรายวิชา MER 2003 ความรู้เบื้องต้นเกี่ยวกับสถิติและวิจัยทางการศึกษา ซึ่งการจำลองข้อมูลการสอบได้จากโปรแกรม WINGEN ที่มีการอิงค่าพารามิเตอร์ของข้อสอบจากผลการทดสอบรายวิชา MER 2003 จำนวนผู้สอบ 2,800 คน

2. **โปรแกรมการคัดเลือกข้อสอบ (Item Selection Program)** หมายถึง กระบวนการทางคอมพิวเตอร์ที่ประยุกต์ใช้จากชุดคำสั่ง Ip_Solve Version 5.5 ในโปรแกรม R Studio เพื่อคัดเลือกข้อสอบจากคลังข้อสอบที่สร้างขึ้นจากงานวิจัยครั้งนี้ โดยข้อสอบที่ผ่านการคัดเลือกจะถูกนำไปสร้างเป็นแบบทดสอบคู่ขนานมากกว่า 1 ฉบับที่มีคุณสมบัติด้านความเป็นคุณลักษณะเฉพาะของความเป็นคู่ขนาน (test parallelism) ด้านความเที่ยงและด้านความตรงที่ใกล้เคียงหรือตรงกับแบบทดสอบฉบับอ้างอิงมากที่สุด รวมทั้งแบบทดสอบคู่ขนานมากกว่า 1 ฉบับมีอัตราการใช้ข้อสอบชำระระหว่างแบบทดสอบอยู่ในระดับต่ำ

3. **แบบทดสอบฉบับอ้างอิง (Referenced Test)** หมายถึง แบบทดสอบฉบับที่ใช้ในการอ้างอิงคุณสมบัติด้านความเป็นคุณลักษณะเฉพาะของความเป็นคู่ขนาน (test parallelism) ด้านความเที่ยงของการประมาณค่าความสามารถ และด้านความตรงตามเนื้อหา กับแบบทดสอบคู่ขนานที่สร้างขึ้นจากโปรแกรมการคัดเลือกข้อสอบ สำหรับการวิจัยครั้งนี้ใช้แบบทดสอบฉบับอ้างอิง คือ แบบทดสอบรายวิชา MER 2003 ความรู้เบื้องต้นเกี่ยวกับสถิติและวิจัยทางการศึกษา

4. **ความถูกต้องของโปรแกรมการคัดเลือกข้อสอบ (Accuracy of Item Selection Program)** หมายถึง ความแม่นยำของการคัดเลือกข้อสอบจากคลังข้อสอบเพื่อนำมาสร้างเป็นแบบทดสอบคู่ขนานมากกว่า 1 ฉบับ สำหรับการวิจัยครั้งนี้พิจารณาจาก 5 ด้าน คือ 1) ความตรงตามเนื้อหา 2) ความเที่ยงของการประมาณค่าความสามารถ 3) ค่าฟังก์ชันสารสนเทศของแบบทดสอบ 4) คุณลักษณะเฉพาะของความเป็นคู่ขนาน และ 5) อัตราการใช้ข้อสอบชำระระหว่างฉบับ

5. คุณลักษณะเฉพาะของความเป็นคู่ขนาน (Test Parallelism) หมายถึง แบบทดสอบคู่ขนานมากกว่า 1 ฉบับที่สร้างขึ้นจากโปรแกรมการคัดเลือกข้อสอบมีคุณลักษณะเฉพาะที่เหมือนกับแบบทดสอบฉบับอ้างอิง ซึ่งการวิจัยครั้งนี้ยึดคุณลักษณะเฉพาะของความเป็นคู่ขนานตามแนวคิดของ McDonald (2009, p. 347-348) กล่าวคือ พิจารณาจากค่าฟังก์ชันสารสนเทศของแบบทดสอบ (Test Information Function: TIF) หรือค่าฟังก์ชันคุณลักษณะของแบบทดสอบ (Test Characteristic Function: TCF) เพื่อเป็นประโยชน์ในการปรับเทียบคะแนนระหว่างแบบทดสอบ (test equating) สำหรับการวิจัยครั้งนี้กำหนดค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิงตามตำแหน่งความสามารถของผู้สอบ จำแนกเป็น 3 กรณี คือ กรณีที่ 1 ความสามารถของผู้สอบ 2 ตำแหน่ง คือ 0.00 และ 1.50 กรณีที่ 2 ความสามารถของผู้สอบ 3 ตำแหน่ง คือ (-1.50), 0.00 และ 1.50 กรณีที่ 3 ความสามารถของผู้สอบ 4 ตำแหน่ง คือ (-1.50), 0.00, 1.50 และ 2.50 ตามลำดับ ส่วนเกณฑ์การประเมินคุณลักษณะเฉพาะของความเป็นคู่ขนาน คือ พิจารณาจากค่าเฉลี่ยยกกำลังสองค่าเบี่ยงเบนของค่าฟังก์ชันสารสนเทศของแบบทดสอบ (Mean Square Deviation of Test Information Function) ที่มีค่าน้อยกว่า 0.05 (Luecht. 1998; Chen , P.-H., Chang, H-H. & Wu, H. 2012)

$$\frac{\sum_{i=1}^n [(TI_{obs}(\theta_i) - TI_{tgt}(\theta_i))^2]}{n}$$

เมื่อ $TI_{obs}(\theta_i)$ หมายถึง ค่าฟังก์ชันสารสนเทศของแบบทดสอบ ณ ตำแหน่งความสามารถ $i = 1, \dots, n$

$TI_{tgt}(\theta_i)$ หมายถึง ค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง

ณ ระดับความสามารถที่ $i = 1, \dots, n$

6. ค่าฟังก์ชันสารสนเทศของแบบทดสอบ (Test Information Function: TIF) หมายถึง ผลรวมเชิงพีชคณิตของค่าฟังก์ชันสารสนเทศของข้อสอบแต่ละข้อที่ได้รับการคัดเลือกเพื่อสร้างเป็นแบบทดสอบคู่ขนาน ซึ่งฟังก์ชันสารสนเทศของข้อสอบเป็นดัชนีผสมที่สร้างจากดัชนีคุณลักษณะของข้อสอบหลายลักษณะ ประกอบด้วยค่าพารามิเตอร์ความยาก อำนาจจำแนก ค่าการเดา และค่าความแปรปรวนของคะแนนรายข้อเพื่อใช้บ่งชี้ถึงความเที่ยง (reliability) ของแบบทดสอบตามทฤษฎีการตอบข้อสอบแบบ 3 พารามิเตอร์

7. ค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง (Referenced Test Information Function: TTIF) หมายถึง ผลรวมเชิงพีชคณิตของค่าฟังก์ชันสารสนเทศของข้อสอบฉบับอ้างอิง คือข้อสอบรายวิชา MER 2003 ซึ่งฟังก์ชันสารสนเทศของข้อสอบเป็นดัชนีผสมที่สร้างจากดัชนีคุณลักษณะของข้อสอบหลายลักษณะ ประกอบด้วยค่าพารามิเตอร์ความยาก อำนาจจำแนก ค่าการเดาและค่าความแปรปรวนของคะแนนรายข้อเพื่อใช้บ่งชี้ถึงความเที่ยง (reliability) ของแบบทดสอบตามทฤษฎีการตอบสนองข้อสอบ และเป็นค่าที่ใช้ในการแสดงความสัมพันธ์ระหว่างคะแนนจริงกับระดับความสามารถของผู้สอบตามทฤษฎีการตอบข้อสอบแบบ 3 พารามิเตอร์ การวิจัยครั้งนี้กำหนดค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิงที่ตรงกับตำแหน่ง

ความสามารถของผู้สอบ จำนวน 2, 3 และ 4 ตำแหน่ง โดยเป็นค่าที่อ้างอิงจากผลการสอบรายวิชา MER 2003 ความรู้เบื้องต้นเกี่ยวกับสถิติและวิจัยทางการศึกษา

8. อัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบ (Test Overlap Rate) หมายถึง สัดส่วนของข้อสอบที่มีการใช้ร่วมกันภายในคลังข้อสอบ โดยพิจารณาจากทุกสถานการณ์การทดสอบว่ามีจำนวนของข้อสอบใดที่ผู้สอบแต่ละคนได้ทำข้อนั้นซ้ำในปริมาณเท่าใด หลักการคำนวณเป็นการคิดจากสัดส่วนของข้อสอบที่ถูกใช้ซ้ำระหว่างแบบทดสอบคู่ขนาน 2 ฉบับ ของทุกคู่ของ 2 แบบทดสอบใดๆ ที่ได้รับการสร้างแบบทดสอบคู่ขนานจากคลังข้อสอบและนำมาหาค่าเฉลี่ยของทุกคู่จากสถานการณ์การทดสอบทั้งหมด สำหรับการวิจัยครั้งนี้กำหนดอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบเฉลี่ย (average test overlap rate: $\hat{T}$) ระดับต่ำ คือ ร้อยละ 25 ตามแนวคิดของ Chen, Ankenmann and Spray (2003, p.129) ที่ทำการควบคุมอัตราการใช้ข้อสอบที่พิจารณาจาก 2 มิติ คือ 1) การควบคุมระดับข้อสอบ คืออัตราของข้อสอบซ้ำ (item exposure rate: Γ_j) และ 2) การควบคุมระดับแบบทดสอบ คืออัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบ (item overlap rate) สามารถแสดงดังสูตร

$$\hat{T} = \frac{1}{\bar{r}} \cdot S_r^2 + \bar{r}$$

เมื่อ $\hat{T}$ หมายถึง อัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบเฉลี่ย
 S_r^2 หมายถึง ความแปรปรวนของอัตราของข้อสอบซ้ำ
 $\bar{r}$ หมายถึง ค่าเฉลี่ยของอัตราของข้อสอบซ้ำ

9. ความคลาดเคลื่อนมาตรฐานของการประมาณค่า (Standard Error of Estimation) หมายถึง ค่าเบี่ยงเบนมาตรฐานของการแจกแจงความน่าจะเป็นของค่าประมาณความสามารถที่แท้จริง (θ) ซึ่งเป็นค่าสัดส่วน ผกผันกับความถูกต้องแม่นยำของการประมาณค่าความสามารถหรือค่าฟังก์ชันสารสนเทศของแบบทดสอบ ดังนั้นค่าฟังก์ชันสารสนเทศของแบบทดสอบมีค่าสูงความคลาดเคลื่อนมาตรฐานของการประมาณค่าความสามารถจะต่ำ นั่นคือถ้าค่าฟังก์ชันสารสนเทศของแบบทดสอบมีค่าสูง ณ ตำแหน่งความสามารถของผู้สอบ θ ใด (ฟังก์ชันการตอบสนองข้อสอบแต่ละข้อมีความชันสูงและความแปรปรวนต่ำ ณ ตำแหน่ง θ นั้น) ความคลาดเคลื่อนมาตรฐานของการประมาณค่าความสามารถของผู้สอบ ณ ตำแหน่งความสามารถนั้นจะต่ำ หรือมีความแม่นยำสูงในการประมาณค่าความสามารถของผู้สอบ θ นั้น สำหรับค่าที่เหมาะสมสำหรับการคัดเลือกข้อสอบคือควรมีค่าความคลาดเคลื่อนมาตรฐานของการประมาณค่าไม่เกิน 0.50 (Hambleton and Swaminathan, 1991, p.103)

10. ความเป็นเอกมิติของแบบทดสอบ (Unidimensionality) หมายถึงข้อสอบทุกข้อในแบบทดสอบมุ่งวัดลักษณะเด่นเดียวหรือความสามารถเด่นเดียว การตรวจสอบความเป็นเอกมิติของแบบทดสอบสามารถทำได้โดยใช้เทคนิคการวิเคราะห์ทางสถิติ ได้แก่ การวิเคราะห์ตัวประกอบ (factor analysis) เพื่อคำนวณค่าไอเกน (eigen value) สำหรับในการวิจัยครั้งนี้ใช้วิธีการตรวจสอบด้วย eigen value test และพิจารณาอัตราส่วนของค่า eigen ที่ 1 กับค่า eigen ที่ 2 มากกว่า 3 ถือว่าแบบทดสอบนั้นมีความเป็นเอกมิติซึ่งถือว่าเป็นข้อตกลงเบื้องต้นที่สำคัญของทฤษฎีการตอบข้อสอบ

11. การประมาณค่าพารามิเตอร์ของข้อสอบ (estimating item parameter) หมายถึง การใช้แบบทดสอบที่วิเคราะห์ภายใต้ทฤษฎีการตอบข้อสอบแบบ 3 พารามิเตอร์เพื่อประมาณค่าพารามิเตอร์แบบทดสอบ โดยใช้ผลการประมาณค่าพารามิเตอร์ของข้อสอบที่ได้ แล้วนำเสนอสารสนเทศสำหรับคุณลักษณะทางเทคนิคของข้อสอบในแบบทดสอบ

ประโยชน์ที่ได้รับจากการวิจัย

ประโยชน์ทางด้านวิชาการ

1. ได้ขยายองค์ความรู้เกี่ยวกับการสร้างคลังข้อสอบและกระบวนการคัดเลือกข้อสอบด้วยโปรแกรมคอมพิวเตอร์เพื่อสร้างแบบทดสอบคู่ขนานมากกว่า 1 ฉบับตามทฤษฎีการตอบข้อสอบ ซึ่งเป็นทฤษฎีการวัดผลที่ค่าพารามิเตอร์ไม่ขึ้นอยู่กับกลุ่มผู้สอบ
2. สามารถสร้างแบบทดสอบคู่ขนานที่มีคุณภาพตามทฤษฎีการตอบสนองข้อสอบจากโปรแกรมที่สร้างขึ้นและเป็นแบบทดสอบที่มีความตรงตามเนื้อหา ความเป็นคู่ขนานของแบบทดสอบอันเป็นส่วนสำคัญของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียน และสามารถเทียบเคียงผลการสอบในแต่ละภาคการศึกษา
3. ผลการวิจัยก่อให้เกิดความยุติธรรมและความถูกต้องแม่นยำในการประมาณค่าความสามารถของนักศึกษา

ประโยชน์ด้านการนำไปปฏิบัติ

1. ผลการวิจัยสามารถนำไปใช้สร้างแบบทดสอบคู่ขนานมากกว่า 1 ฉบับสำหรับวิชาที่มีนักศึกษาลงทะเบียนจำนวนมากๆ ได้อย่างมีประสิทธิภาพ เช่น รายวิชา MER 2003 สถิติและการวิจัยเบื้องต้นทางการศึกษา (Introduction to Statistics and Research in Education)
2. การมีคลังข้อสอบและวิธีการคัดเลือกข้อสอบเพื่อสร้างแบบทดสอบคู่ขนานมากกว่า 1 ฉบับได้ในคราวเดียวกันจะช่วยลดปริมาณข้อสอบที่ซ้ำกันในแต่ละฉบับได้ และยังเป็นการลดเวลาในการทำงาน ตลอดจนเป็นแนวทางในการจัดการสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (computer adaptive testing)

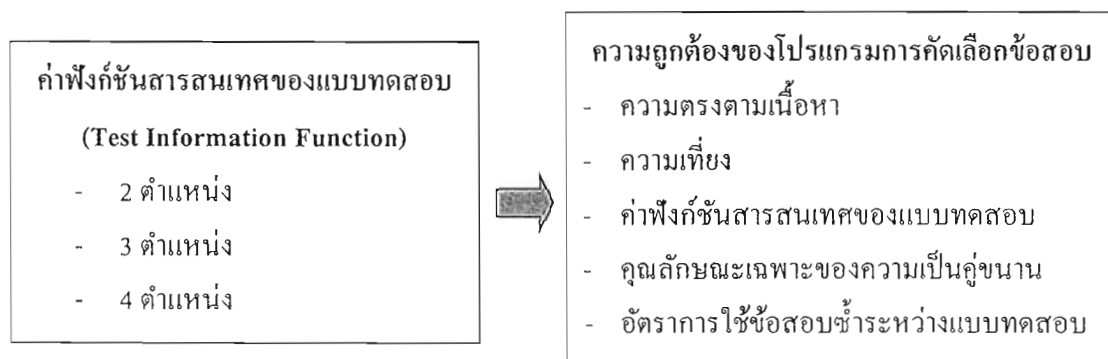
กรอบแนวคิดการวิจัย

คลังข้อสอบนิยมนำมาใช้ประโยชน์กับการวัดประเมินทางการศึกษานาขนาดใหญ่ (large scale assessment) กล่าวคือเป็นการทดสอบกับผู้สอบที่มีจำนวนมาก นอกจากนี้ครูผู้สอนสามารถใช้ประโยชน์จากคลังข้อสอบได้ด้วยการนำคลังข้อสอบที่ถูกจำแนกตามเนื้อหาการวัดมาสร้างเป็นแบบทดสอบคู่ขนานมากกว่า 1 ฉบับที่มีคุณสมบัติทั้งด้านความตรงตามเนื้อหา เป็นแบบทดสอบที่เหมาะสมกับระดับความสามารถของผู้สอบ รวมทั้งสามารถเทียบเคียงผลการสอบในแต่ละภาคการศึกษาจากการสอบด้วยแบบทดสอบต่างฉบับกัน คุณสมบัติของแบบทดสอบข้างต้นเป็นการสร้างแบบทดสอบคู่ขนานตามทฤษฎีการตอบสนองข้อสอบ อีกทั้งในปัจจุบันโปรแกรมคอมพิวเตอร์ช่วยให้การสร้างแบบทดสอบคู่ขนานจากคลังข้อสอบโดยอัตโนมัติ จากการทบทวนเอกสารที่เกี่ยวข้องพบว่าโปรแกรมการคัดเลือกข้อสอบคู่ขนานโดยอัตโนมัติมักเป็นโปรแกรมทางการค้า ดังนั้นผู้วิจัยจึงสนใจสร้างโปรแกรมการคัดเลือกข้อสอบเพื่อนำมาสร้างเป็นแบบทดสอบคู่ขนานจากคลังข้อสอบด้วยการประยุกต์ใช้คำสั่งจากโปรแกรม R Studio ซึ่งเป็นโปรแกรมสำเร็จรูปที่เปิดโอกาสให้บุคคลทั่วไปดาวน์โหลด ติดตั้งและใช้ได้ฟรี โดยไม่ถือว่าเป็นการละเมิดลิขสิทธิ์ รวมทั้งผู้ใช้สามารถเขียนฟังก์ชันต่างๆ เพิ่มเติมสำหรับการวิเคราะห์ข้อมูลทางสถิติตามเงื่อนไขที่ผู้ใช้ต้องการ แต่อย่างไรก็ตามแบบทดสอบคู่ขนานที่ถูกสร้างขึ้นจากโปรแกรมการคัดเลือกข้อสอบตามข้อกำหนดของการวิจัยต้องแสดงถึงความถูกต้องของการคัดเลือกข้อสอบตามเงื่อนไขที่กำหนด

ความถูกต้องของโปรแกรมการคัดเลือกข้อสอบที่สำคัญที่สุด คือ แบบทดสอบคู่ขนานที่ถูกสร้างขึ้นมากกว่า 1 ฉบับนั้นต้องมีคุณลักษณะเฉพาะของความเป็นคู่ขนานตามทฤษฎีการตอบสนองข้อสอบ โดยพิจารณาจากความสอดคล้องระหว่างค่าฟังก์ชันสารสนเทศของแบบทดสอบ กล่าวคือค่าฟังก์ชันสารสนเทศของแบบทดสอบที่ได้จากโปรแกรมการคัดเลือกข้อสอบกับค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิงมีค่าที่ตรงกันหรือใกล้เคียงกัน เพื่อเป็นประโยชน์ในการเปรียบเทียบคะแนนระหว่างแบบทดสอบ (test equating) ซึ่งการวิจัยครั้งนี้ยึดตามแนวคิดของ McDonald (2009, p. 347-348) สำหรับอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบก็เป็นสิ่งสำคัญเช่นกัน ซึ่งเป็นการป้องกันมิให้เกิดอัตราการใช้ข้อสอบซ้ำของข้อสอบที่ถูกคัดเลือกเพื่อนำไปสร้างแบบทดสอบคู่ขนานมากกว่า 1 ฉบับ นอกจากนี้ความตรงตามเนื้อหาและความเที่ยงของแบบทดสอบก็เป็นสิ่งที่สะท้อนถึงแบบทดสอบคู่ขนานนั้นวัดได้ตรงตามเนื้อหาที่ต้องการวัดและยังเป็นค่าที่บ่งบอกถึงความน่าเชื่อถือของการประมาณค่าความสามารถของผู้สอบ (θ) (Weiss. 2011; van de Linden.2005)

ดังนั้นผู้วิจัยจึงต้องการศึกษาความถูกต้องของโปรแกรมคัดเลือกข้อสอบที่สร้างขึ้นว่าหากมีการกำหนดระดับของค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิงตามตำแหน่งความสามารถของผู้สอบ (Referenced Test Information Function: RTIF) ที่แตกต่างกัน 3 กรณี คือ 2, 3 และ 4 ตำแหน่งแล้ว โปรแกรมการคัดเลือกข้อสอบที่สร้างขึ้นยังมีความถูกต้องของโปรแกรมที่แตกต่างกันหรือไม่ อย่างไร

จากแนวคิดเกี่ยวกับการสร้างแบบทดสอบคู่ขนานจากคลังข้อสอบตามทฤษฎีการตอบข้อสอบข้างต้น ผู้วิจัยจึงสร้างเป็นกรอบแนวคิดการวิจัยดังแผนภาพ 1



แผนภาพ 1 กรอบแนวคิดการวิจัย

บทที่ 2

เอกสารและงานวิจัยที่เกี่ยวข้อง

การศึกษานี้เป็นการวิจัยเพื่อสร้างแบบทดสอบคู่ขนานจากคลังข้อสอบที่มีคุณภาพด้านความเที่ยงและความตรงตามทฤษฎีการตอบข้อสอบ โดยมีวิธีการดำเนินการวิจัย คือ ระยะเวลาที่ 1 การสร้างและตรวจสอบคุณภาพของแบบทดสอบฉบับอ้างอิงตามทฤษฎีการทดสอบแบบดั้งเดิมและทฤษฎีการตอบข้อสอบ ระยะเวลาที่ 2 การสร้าง ตรวจสอบ และคัดเลือกข้อสอบเพื่อบรรจุในคลังข้อสอบตามคุณสมบัติของแบบทดสอบฉบับอ้างอิงด้วยการจำลองข้อมูลการสอบด้วยโปรแกรม WINGEN และระยะเวลาที่ 3 การตรวจสอบความถูกต้องของโปรแกรมการสร้างแบบทดสอบคู่ขนานมากกว่า 1 ฉบับจากคลังข้อสอบตามด้านคุณลักษณะเฉพาะของความเป็นคู่ขนาน (parallelism specification) และด้านอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบ ผู้วิจัยได้ทำการทบทวนเอกสารและงานวิจัยที่เกี่ยวข้อง ดังนี้

ตอนที่ 1 แนวคิดและหลักการของโมเดลเชิงเส้นสำหรับการออกแบบการสร้างแบบทดสอบ (Linear Models for Optimal Test Design) และงานวิจัยที่เกี่ยวข้อง

- 1.1 การออกแบบการสร้างแบบทดสอบแบบดั้งเดิม (Classical Test Theory)
- 1.2 การออกแบบการสร้างแบบทดสอบแนวใหม่ (Modern Test Theory)
- 1.3 การออกแบบการสร้างแบบทดสอบตามแนวคิดของ Birnbaum
- 1.4 การออกแบบการสร้างแบบทดสอบตามแนวคิดของ van der Linden
- 1.5 งานวิจัยที่เกี่ยวข้องกับโมเดลเชิงเส้นสำหรับการออกแบบการสร้างแบบทดสอบ

ตอนที่ 2 แนวคิดและหลักการการตรวจสอบอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบ (Test Overlap Rate) และงานวิจัยที่เกี่ยวข้อง

- 2.1 แนวคิดและหลักการการตรวจสอบอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบ
- 2.2 งานวิจัยที่เกี่ยวข้องกับการตรวจสอบอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบ

ตอนที่ 3 การใช้โปรแกรม R

- 3.1 ภาษา R
- 3.2 คำสั่งในการสร้างแบบทดสอบคู่ขนาน 2 ฉบับ

ตอนที่ 4 ผลการสังเคราะห์เอกสารและงานวิจัยที่เกี่ยวข้องกับการสร้างแบบทดสอบคู่ขนานจากคลังข้อสอบตามทฤษฎีการตอบข้อสอบ (Item Response Theory)

ดังรายละเอียด

ตอนที่ 1 แนวคิดและหลักการของโมเดลเชิงเส้นสำหรับการออกแบบการสร้างแบบทดสอบ (Linear Models for Optimal Test Design) และงานวิจัยที่เกี่ยวข้อง

โมเดลเชิงเส้นสำหรับการออกแบบการสร้างแบบทดสอบ (Linear Models for Optimal Test Design)

แบบทดสอบเป็นเครื่องมือที่ใช้ในการตรวจสอบระดับความรู้ความคิดของผู้เรียน นักวิจัยส่วนใหญ่คาดหวังที่จะได้ข้อมูลตรงกับความเป็นจริงให้มากที่สุด โมเดลเชิงเส้นสำหรับการออกแบบการสร้างแบบทดสอบจึงมีความสำคัญเพื่อให้ได้แบบทดสอบที่มีคุณภาพ ซึ่งวิวัฒนาการของโมเดลเชิงเส้นสำหรับการออกแบบการสร้างแบบทดสอบสามารถจำแนกได้เป็น 2 แนวคิด คือ 1) การออกแบบการสร้างแบบทดสอบมาตรฐานโดยใช้ทฤษฎีการทดสอบแบบดั้งเดิม และ 2) การออกแบบการสร้างแบบทดสอบสำหรับการทดสอบเป็นรายบุคคล โดยใช้ทฤษฎีการทดสอบแนวใหม่ สามารถกล่าวโดยสรุป คือ

การออกแบบการสร้างแบบทดสอบมาตรฐานโดยใช้ทฤษฎีการทดสอบแบบดั้งเดิมที่นิยมใช้ คือ แบบทดสอบมาตรฐาน (Standardized test) เริ่มใช้มาตั้งแต่การสอบคัดเลือกจอหงวนในประเทศจีน ต่อมา ค.ศ. 1905 Binet และ Simon สร้างแบบวัดเชาวน์ปัญญา ซึ่งจัดเป็นการทดสอบแบบกลุ่ม (Group Testing) และในปี ค.ศ. 1970 เป็นยุคของการวัดผลเพื่อวัดความรอบรู้ (Mastery Learning) มีการทดสอบแบบให้ผลสะท้อนกลับไปสู่ผู้เรียนเป็นรายบุคคล จัดเป็นการบูรณาการการสอนและการสอบเข้าด้วยกัน ส่งผลให้เกิดความต้องการคลังข้อสอบ และการจัดชุดแบบทดสอบ (Test Assembly) ตามศักยภาพของผู้สอบเป็นรายบุคคลเพื่อนำผลการสอบมาออกแบบการจัดการเรียนรู้ให้กับผู้เรียน ซึ่งแบบทดสอบมาตรฐานตามช่วงเวลาดังกล่าวใช้ทฤษฎีการทดสอบแบบดั้งเดิม (Classical Theory) เป็นทฤษฎีที่สนับสนุนการสร้างแบบทดสอบมาตรฐาน

ส่วนการออกแบบการสร้างแบบทดสอบสำหรับการทดสอบเป็นรายบุคคล โดยใช้ทฤษฎีการทดสอบแนวใหม่นั้น เริ่มตั้งแต่การพัฒนาทฤษฎีการทดสอบแนวใหม่ กล่าวคือ ในปี ค.ศ. 1950 ทฤษฎีการทดสอบแนวใหม่ (Item Response Theory) ถูกพัฒนาขึ้น ต่อมา ค.ศ. 1990 ทฤษฎีการทดสอบแนวใหม่ได้รับความนิยมอย่างแพร่หลาย เนื่องจากคอมพิวเตอร์ส่วนบุคคลเริ่มมีราคาถูกลงจึงทำให้นักวัดผลพัฒนาคคลังข้อสอบ (Item Banking) และดำเนินการสอบเป็นรายบุคคลมากขึ้น จึงทำให้การทดสอบแบบปรับเหมาะ (Adaptive Testing) ได้รับความนิยม (van der Linden, 2005, p.1)

ประเภทของการออกแบบการสร้างแบบทดสอบ

van der Linden (2005, p.1) จำแนกประเภทของการออกแบบการสร้างแบบทดสอบตามหลักการสำคัญของทฤษฎีการทดสอบ คือ 1) การออกแบบการสร้างแบบทดสอบแบบดั้งเดิม 2) การออกแบบการสร้างแบบทดสอบแนวใหม่

1.1 การออกแบบการสร้างแบบทดสอบแบบดั้งเดิม (Classical Test Design)

เป็นการออกแบบการสร้างแบบทดสอบที่ใช้สำหรับแบบทดสอบมาตรฐานที่วัดคุณลักษณะทางจิตวิทยาเช่น แบบวัดบุคลิกภาพ แบบวัดความสามารถพิเศษ เป็นต้น โดยกระบวนการสร้างผู้สร้างเป็นผู้เชี่ยวชาญเฉพาะด้าน มีการทดลองใช้แบบวัด และรายงานผลคุณภาพของแบบวัด โดยจุดมุ่งหมายของการทดลองใช้มี 3 ประการ คือ 1) คัดเลือกข้อสอบที่มีคุณภาพตามค่าสถิติที่กำหนดไว้ 2) ตรวจสอบความตรงตามโครงสร้าง (construct validity) และ 3) สร้างคะแนนเกณฑ์ปกติ (norm)

สิ่งที่ควรคำนึงถึงสำหรับการออกแบบการสร้างแบบทดสอบแบบดั้งเดิม คือ 1) ทฤษฎีการทดสอบแบบดั้งเดิม และ 2) ค่าพารามิเตอร์ของแบบทดสอบและข้อสอบ

ทฤษฎีการทดสอบแบบดั้งเดิม (Classical Theory)

van der Linden (2005, pp. 2-5) กล่าวว่า ทฤษฎีการทดสอบแบบดั้งเดิมจำแนกผลการวัดหรือคะแนนสังเกตได้ (X) ออกเป็น 2 ส่วน คือ ความสามารถที่แท้จริงของบุคคล หรือคะแนนจริง (T) และคะแนนคลาดเคลื่อน (E) สำคัญดังกล่าวดังกล่าวอธิบายได้ด้วย 2 หลักการ คือ

1) การกำหนดให้กลุ่มผู้สอบคงที่ (fixed person)

หลักการนี้คำนึงถึงการประมาณค่าคะแนนจริงด้วยการดำเนินการสอบด้วยการทดสอบซ้ำกับกลุ่มผู้สอบกลุ่มเดิม แต่ยังคงมีข้อตกลงเบื้องต้นสำคัญ คือ คะแนนสังเกตได้ (X_{jt}) เป็นตัวแปรสุ่ม จึงทำให้ไม่สามารถทราบถึงลักษณะของการแจกแจง ดังแสดงตามสมการ 1.1 และ 1.2

$$T_{jt} = \sum X_{jt} \quad \text{สมการ 1.1}$$

$$E_{jt} = X_{jt} - T_{jt} \quad \text{สมการ 1.2}$$

เมื่อ X_{jt} หมายถึง คะแนนสังเกตได้ที่เกิดจากการผู้สอบคนที่ j ทำแบบทดสอบฉบับที่ t

T_{jt} หมายถึง คะแนนจริงของผู้สอบคนที่ j ทำแบบทดสอบฉบับที่ t โดยเป็นค่าคาดหวัง หรือค่าเฉลี่ยของการกระจายของคะแนนสังเกตได้ (observed-score distribution)

E_{jt} หมายถึง คะแนนคลาดเคลื่อนระหว่างคะแนนสังเกตได้กับคะแนนจริง

แต่อย่างไรก็ตามสมการที่ 1.1 และ 1.2 จะสามารถเกิดขึ้นได้ในทางปฏิบัติ หากการแจกแจงของ X_{jt} ซึ่งเป็นค่าที่ได้จากการทดสอบซ้ำ และมีการแจกแจงแบบปกติ จึงทำให้ค่าเฉลี่ยของ X_{jt} จะเป็นค่าคาดหวังที่ไม่ลำเอียงของ T_{jt} จึงทำให้เกิดสมการ 1.3 คือ

$$X_{jt} = T_{jt} + E_{jt} \quad \text{สมการ 1.3}$$

จากสมการที่ 1.1-1.2 ได้อธิบายถึงคะแนนสังเกตได้ (X_{jt}) ด้วยทฤษฎีการทดสอบแบบดั้งเดิม กล่าวว่า “คะแนนสังเกตได้เป็นตัวแปรสุ่ม จึงไม่ทราบถึงลักษณะของการแจกแจงของคะแนน” ส่วนการประมาณค่าคะแนนจริงจากคะแนนสังเกตได้ในทางปฏิบัติจากสมการที่ 1.3 กล่าวว่า “หากจะประมาณค่าคะแนนจริงด้วยคะแนนสังเกตได้ต้องทำการทดสอบซ้ำ และคะแนนต้องมีการแจกแจงแบบปกติ ดังนั้นค่าเฉลี่ยของการ

กระจายของคะแนนสังเกตได้จึงจะเป็นค่าคาดหวังของคะแนนจริงที่ไม่ลำเอียง” จากข้อความข้างต้นจะสังเกตเห็นได้ว่าข้อตกลงเบื้องต้นเกี่ยวกับคะแนนสังเกตได้เป็นตัวแปรสุ่มไม่สอดคล้องในทางปฏิบัติ ส่วนสมการ 1.3 ช่วยทำให้คะแนนสังเกตได้ (X) ที่วัดได้ในแต่ละครั้งสามารถสรุปอ้างอิงกลับไปยังการแจกแจงของคะแนนจริง (T) ได้นั่นเอง

2) การสุ่มกลุ่มผู้สอบ (random person)

หากการทดสอบซ้ำแต่ละครั้งมีการสุ่มกลุ่มตัวอย่างผู้สอบจากประชากรผู้สอบแล้วคะแนนจริงของคะแนนสังเกตได้ และคะแนนคลาดเคลื่อน จะเป็นตัวแปรสุ่ม ดังแสดงในสมการ 1.4

$$X_{jt} = \tau_{jt} + E_{jt} \quad \text{สมการ 1.4}$$

เมื่อ X_{jt} หมายถึง คะแนนสังเกตได้ของตัวอย่างผู้สอบด้วยการสุ่มจากประชากรผู้สอบ j ทำแบบทดสอบฉบับที่ t

τ_{jt} หมายถึง คะแนนจริงของตัวอย่างผู้สอบด้วยการสุ่มจากประชากรผู้สอบ j ทำแบบทดสอบฉบับที่ t

E_{jt} หมายถึง คะแนนคลาดเคลื่อนระหว่างคะแนนสังเกตได้กับคะแนนจริง

ข้อสังเกตของหลักการประมาณค่าคะแนนจริงด้วยการสุ่มกลุ่มผู้สอบนั้น จะพบว่าสมการ 1.4 เป็นการรวมตัวของคะแนน 3 ส่วนซึ่งล้วนเป็นตัวแปรสุ่ม ดังนั้นจึงไม่มีข้อตกลงเบื้องต้นเกี่ยวกับการรวมตัวในเชิงเส้นตรง (linearity)

ค่าพารามิเตอร์ของข้อสอบและแบบทดสอบ (Item and Test Parameters)

1) ค่าพารามิเตอร์ของแบบทดสอบ (test parameters) ประกอบด้วยค่าสัมประสิทธิ์ความเที่ยง (reliability coefficient: ρ_{XT}^2) และค่าสัมประสิทธิ์ความตรงเชิงทำนาย (predictive validity: ρ_{XY})

สามารถประมาณค่าจากการยกกำลังสองค่าความสัมพันธ์ (ρ_{XT}^2) ของคะแนนจริง (T) กับคะแนนสังเกตได้ (X) หากสถานการณ์ทดสอบใดๆ ทดสอบจากประชากรผู้สอบแล้ว $X = T$ ทำให้ $\rho_{XT} = 1.00$ การนำ ρ_{XT} มายกกำลังสองเพื่อสะดวกในการแปลความหมายซึ่งหมายถึงสัดส่วนของการอธิบายความแปรปรวนของ X ต่อ T แสดงดังสมการ 1.5

$$\rho_{XT}^2 = \frac{\text{Var}(T_{jt})}{\text{Var}(X_{jt})} \quad \text{สมการ 1.5}$$

ถ้าเราสนใจศึกษาความสัมพันธ์ระหว่างคะแนนสังเกตได้ (X) กับคะแนนจากแบบทดสอบฉบับเกณฑ์ (Y) จึงทำให้เกิดความหมายใหม่ของค่าพารามิเตอร์ของแบบทดสอบ คือ ค่าสัมประสิทธิ์ความตรง (validity coefficient: ρ_{XY}) ซึ่งค่าสัมประสิทธิ์นี้มีได้จากการยกกำลังสองด้วยเหตุผล คือ กฎของค่าความแปรปรวน-ความแปรปรวนร่วม ดังแสดงสมการ 1.6

$$\rho_{XT}^2 = \rho_{XX}/ \quad \text{สมการ 1.6}$$

ความหมายของสมการที่ 1.6 คือค่ายกกำลังสองของคะแนนสังเกตได้กับคะแนนจริงจะเท่ากับความสัมพันธ์ระหว่างคะแนนที่ได้จากการทดสอบซ้ำ (X, X') จึงทำให้ค่าสัมประสิทธิ์ความตรงเชิงทำนาย (predictive validity) จะมีค่าน้อยกว่าหรือเท่ากับความสัมพันธ์ระหว่างคะแนนสังเกตได้กับคะแนนจริง ดังแสดงในสมการ 1.7

$$\rho_{XT} \geq \rho_{XY} \quad \text{สมการ 1.7}$$

2) ค่าพารามิเตอร์ของข้อสอบ (item parameters) ประกอบด้วยค่าความยาก อำนาจจำแนกและค่าความตรงของข้อสอบ

2.1) ค่าความยาก (item difficulty: π_i)

เป็นค่าคาดหวังของค่าเฉลี่ยของการตอบข้อสอบถูกของประชากรผู้สอบ ดังสมการ 1.8

$$\pi_i = \sum U_i \quad \text{สมการ 1.8}$$

เมื่อ U_i หมายถึง คะแนนของข้อสอบข้อที่ i โดย $U_i = 1$ เมื่อผู้สอบตอบข้อที่ i ถูก และ $U_i = 0$ เมื่อผู้สอบตอบข้อที่ i ผิด

2.2) ค่าอำนาจจำแนก (item discrimination: ρ_{iX})

เป็นค่าที่แสดงถึงความสัมพันธ์ระหว่างคะแนนรายข้อกับคะแนนรวม (X) ดังสมการ 1.9

$$\rho_{iX} = \text{Cor}(U_i, X) = \frac{\sigma_{iX}}{\sigma_i \sigma_X} \quad \text{สมการ 1.9}$$

เมื่อ σ_{iX} หมายถึง ความแปรปรวนร่วมระหว่างคะแนนของข้อสอบข้อที่ i กับคะแนนรวม (X)
 σ_i หมายถึง ค่าเบี่ยงเบนมาตรฐานของคะแนนของข้อสอบข้อที่ i
 σ_X หมายถึง ค่าเบี่ยงเบนมาตรฐานของคะแนนรวม (X)

1.3) ค่าความตรงของข้อสอบ (item validity: ρ_{iY})

เป็นค่าอำนาจจำแนกของข้อสอบอีกลักษณะหนึ่งซึ่งแสดงถึงการจำแนกระหว่างผู้สอบที่มีความสามารถระดับสูงกับระดับต่ำกับคะแนนจากแบบทดสอบที่เป็นฉบับเกณฑ์ (Y) ใช้ในการศึกษาเพื่อหาคุณภาพด้านความตรงเชิงทำนาย ดังสมการ 1.10

$$\rho_{iY} = \text{Cor}(U_i, Y) = \frac{\sigma_{iY}}{\sigma_i \sigma_Y} \quad \text{สมการ 1.10}$$

ข้อสังเกตของการประมาณค่าพารามิเตอร์ของแบบทดสอบตามทฤษฎีการทดสอบแบบดั้งเดิมจะพบว่าเป็นค่าที่ต้องคำนวณจากกลุ่มประชากรผู้สอบ หรือคำนวณจากตัวอย่างผู้สอบที่สุ่มจากประชากรผู้สอบที่มีคุณสมบัติที่สมมูลกัน (equivalents) จึงจะเป็นการคำนวณค่าสัมประสิทธิ์ความเที่ยงที่สอดคล้องกับสมการที่ 1.5 (ρ_{XT}^2) แต่ในทางปฏิบัตินักวัดผลนำผลการวัดที่ได้จากการทดสอบซ้ำมาประมาณค่าความเที่ยงดังสมการที่ 1.6 แต่ข้อจำกัดของการทดสอบซ้ำ คือ ผู้สอบเกิดการเรียนรู้ระหว่างการทดสอบ

ความหมายของสมการที่ 1.6 คือค่ายกกำลังสองของคะแนนสังเกตได้กับคะแนนจริงจะเท่ากับความสัมพันธ์ระหว่างคะแนนที่ได้จากการทดสอบซ้ำ (X, X') จึงทำให้ค่าสัมประสิทธิ์ความตรงเชิงทำนาย (predictive validity) จะมีค่าน้อยกว่าหรือเท่ากับความสัมพันธ์ระหว่างคะแนนสังเกตได้กับคะแนนจริง ดังแสดงในสมการ 1.7

$$\rho_{XT} \geq \rho_{XY} \quad \text{สมการ 1.7}$$

2) ค่าพารามิเตอร์ของข้อสอบ (item parameters) ประกอบด้วยค่าความยาก อำนาจจำแนกและค่าความตรงของข้อสอบ

2.1) ค่าความยาก (item difficulty: π_i)

เป็นค่าคาดหวังของค่าเฉลี่ยของการตอบข้อสอบถูกของประชากรผู้สอบ ดังสมการ 1.8

$$\pi_i = \sum U_i \quad \text{สมการ 1.8}$$

เมื่อ U_i หมายถึง คะแนนของข้อสอบข้อที่ i โดย $U_i = 1$ เมื่อผู้สอบตอบข้อที่ i ถูก และ $U_i = 0$ เมื่อผู้สอบตอบข้อที่ i ผิด

2.2) ค่าอำนาจจำแนก (item discrimination: ρ_{iX})

เป็นค่าที่แสดงถึงความสัมพันธ์ระหว่างคะแนนรายข้อกับคะแนนรวม (X) ดังสมการ 1.9

$$\rho_{iX} = \text{Cor}(U_i, X) = \frac{\sigma_{iX}}{\sigma_i \sigma_X} \quad \text{สมการ 1.9}$$

เมื่อ σ_{iX} หมายถึง ความแปรปรวนร่วมระหว่างคะแนนของข้อสอบข้อที่ i กับคะแนนรวม (X)

σ_i หมายถึง ค่าเบี่ยงเบนมาตรฐานของคะแนนของข้อสอบข้อที่ i

σ_X หมายถึง ค่าเบี่ยงเบนมาตรฐานของคะแนนรวม (X)

1.3) ค่าความตรงของข้อสอบ (item validity: ρ_{iY})

เป็นค่าอำนาจจำแนกของข้อสอบอีกลักษณะหนึ่งซึ่งแสดงถึงการจำแนกระหว่างผู้สอบที่มีความสามารถระดับสูงกับระดับต่ำกับคะแนนจากแบบทดสอบที่เป็นฉบับเกณฑ์ (Y) ใช้ในการศึกษาเพื่อหาคุณภาพด้านความตรงเชิงทำนาย ดังสมการ 1.10

$$\rho_{iY} = \text{Cor}(U_i, Y) = \frac{\sigma_{iY}}{\sigma_i \sigma_Y} \quad \text{สมการ 1.10}$$

ข้อสังเกตของการประมาณค่าพารามิเตอร์ของแบบทดสอบตามทฤษฎีการทดสอบแบบดั้งเดิมจะพบว่าเป็นค่าที่ต้องคำนวณจากกลุ่มประชากรผู้สอบ หรือคำนวณจากตัวอย่างผู้สอบที่สุ่มจากประชากรผู้สอบที่มีคุณสมบัติที่สมมูลกัน (equivalents) จึงจะเป็นการคำนวณค่าสัมประสิทธิ์ความเที่ยงที่สอดคล้องกับสมการที่ 1.5 (ρ_{XT}^2) แต่ในทางปฏิบัตินักวัดผลนำผลการวัดที่ได้จากการทดสอบซ้ำมาประมาณค่าความเที่ยงดังสมการที่ 1.6 แต่ข้อจำกัดของการทดสอบซ้ำ คือ ผู้สอบเกิดการเรียนรู้ระหว่างการทดสอบ

ความสามารถของผู้สอบไม่คงที่ จากข้อจำกัดดังกล่าวจึงเกิดการประมาณค่าสัมประสิทธิ์ความเที่ยงดัง
สมการ 1.11 และ 1.12

$$\rho_{XT}^2 \geq \alpha \quad \text{สมการ 1.11}$$

$$\alpha = \left[\frac{n}{n-1} \right] \left[1 - \frac{\sum_{i=1}^n \sigma_i^2}{\sigma_X^2} \right] \quad \text{สมการ 1.12}$$

เมื่อ n หมายถึง จำนวนข้อสอบ

α หมายถึง ค่าสัมประสิทธิ์ความเที่ยงแบบความสอดคล้องภายใน

จากสมการที่ 1.11 และ 1.12 ค่าสัมประสิทธิ์ความเที่ยงแบบสอดคล้องภายใน (α) จะน้อยกว่า
ค่าสัมประสิทธิ์ความเที่ยงที่คำนวณจากค่ากำลังสองของความสัมพันธ์ระหว่าง X กับ T (ρ_{XT}^2) หาก
แบบทดสอบมีความเป็นเอกมิติ (unidimension) แล้วความคลาดเคลื่อนระหว่าง ρ_{XT}^2 กับ α จะน้อยมาก
ทั้งนี้ค่าสัมประสิทธิ์ α สามารถประมาณค่าได้จากการทดสอบเพียงครั้งเดียว

เมื่อพิจารณาค่าเบี่ยงเบนมาตรฐานของคะแนนสังเกตได้ดังสมการ 1.13

$$\sigma_X = \sum_{i=1}^n \sigma_i \rho_{iX} \quad \text{สมการ 1.13}$$

หากนำสมการที่ 1.13 ไปแทนค่าในสมการที่ 1.12 ทำให้เกิดสมการที่ 1.14

$$\alpha = \frac{n}{n-1} \left[1 - \frac{\sum_{i=1}^n \sigma_i^2}{\left(\sum_{i=1}^n \sigma_i \rho_{iX} \right)^2} \right] \quad \text{สมการ 1.14}$$

จากสมการที่ 1.14 นั้นเองทำให้สามารถคำนวณได้ว่าควรตัดข้อสอบหรือเพิ่มข้อสอบข้อใดเพื่อทำ
ให้ค่าสัมประสิทธิ์ความเที่ยงเพิ่มขึ้น

สำหรับค่าสัมประสิทธิ์ความตรงเป็นค่าที่แสดงถึงความตรงเชิงทำนายที่เกิดจากการรวมค่าความ
แปรปรวนรายข้อ (item variance) ค่าอำนาจจำแนก (item-discriminant) และค่าความตรงของข้อสอบ (item
validity) ดังสมการ 1.15

$$\rho_{XY} = \frac{\sum_{i=1}^n \sigma_i \rho_{iY}}{\sum_{i=1}^n \sigma_i \rho_{iX}} \quad \text{สมการ 1.15}$$

1.2 การออกแบบการสร้างแบบทดสอบแนวใหม่ (Modern Test Design)

สถานการณ์ของการทดสอบกับผู้สอบขนาดใหญ่ (large-scale) มักใช้กับการทดสอบเพื่อคัดเลือก
ผู้เรียนเข้าศึกษาต่อระดับอุดมศึกษา ดังนั้นแนวคิดของการกำหนดให้กลุ่มประชากรผู้สอบคงที่ (fixed
person) ยังคงใช้ได้ดีในการออกแบบการสร้างแบบทดสอบแนวใหม่ อีกทั้งในปัจจุบันมีการจัดการเรียนรู้
แบบเป็นรายบุคคลมากขึ้นจึงทำให้สถานศึกษาจำเป็นต้องสร้างคลังข้อสอบ (item banking) เพื่อให้ครูผู้สอน
สามารถนำผลจากการวัดประเมิน (assessment) มาออกแบบการจัดการเรียนรู้เป็นรายบุคคล และสามารถวัด
ประเมินพัฒนาการของผู้เรียนได้ สำหรับบทบาทด้านการทดสอบที่สำคัญของทฤษฎีการตอบข้อสอบ (Item
Response Theory: IRT) คือ 1) สามารถรวบรวมข้อสอบ และคัดเลือกข้อสอบที่เหมาะสมกับระดับความสามารถ

ของผู้สอบ 2) สามารถวัดประเมินเป็นรายบุคคลโดยใช้สารสนเทศของคลังข้อสอบ 3) สามารถปรับเทียบข้อสอบระหว่างแบบทดสอบต่างฉบับ ณ ระดับความสามารถของผู้สอบระดับเดียวกัน

เมื่อพิจารณาถึงระบบการจัดเก็บและการรวบรวมข้อสอบแบบอัตโนมัติเพื่อจัดเป็นชุดแบบทดสอบ ความคุ้มค่าของการสอบและตามตำแหน่งความสามารถของผู้สอบ (Automated Test Assembly: ATA) นั้น การออกแบบการสร้างแบบทดสอบแนวใหม่โดยใช้ทฤษฎีของ IRT จะมีประสิทธิภาพมากกว่าการออกแบบการสร้างแบบทดสอบมาตรฐานตามทฤษฎีการทดสอบแบบดั้งเดิม

จุดเน้นของทฤษฎีการตอบข้อสอบ (Item Response Theory: IRT) คือสนใจผลการตอบข้อสอบของผู้สอบแต่ละคนที่ทำแบบทดสอบที่เหมาะสมกับความสามารถของตน ดังนั้นโมเดลการตอบข้อสอบจึงคำนวณตามหลักการสุ่มโดยพิจารณาค่าความน่าจะเป็นในการตอบถูกที่แปรเปลี่ยนไปตามค่าพารามิเตอร์ของข้อสอบ

โมเดลการตอบข้อสอบ (Item Response Model)

ศิริชัย กาญจนวาสี (2555, หน้า 52-78) กล่าวว่า โมเดลการตอบข้อสอบ (item response model) เป็นระบบความสัมพันธ์ระหว่างตัวแปรอิสระที่รวมกันสำหรับทำนายตัวแปรตาม โดยที่ตัวแปรอิสระประกอบด้วย ตัวแปรแฝง คือ ความสามารถที่แท้จริงของผู้สอบ (θ) และคุณลักษณะของข้อสอบหรือค่าพารามิเตอร์ของข้อสอบ (a, b, c) ส่วนตัวแปรตามเป็นตัวแปรสังเกตได้ คือ โอกาสในการตอบข้อสอบได้ถูกต้อง

คุณลักษณะของโมเดลการตอบสนองข้อสอบเป็นระบบความสัมพันธ์ระหว่างโอกาสตอบข้อสอบถูก (P_i) กับความสามารถที่มีอยู่ในตัวผู้สอบ (θ) ในรูปของโค้งลักษณะข้อสอบ (Item Characteristic Curve: ICC) ซึ่งมีลักษณะเป็นฟังก์ชันโลจิสต์ (logistic function) หรือฟังก์ชันปกติสะสม (normal ogive function) บางครั้งเรียกว่าโมเดลโลจิสต์ หรือโมเดลปกติสะสม

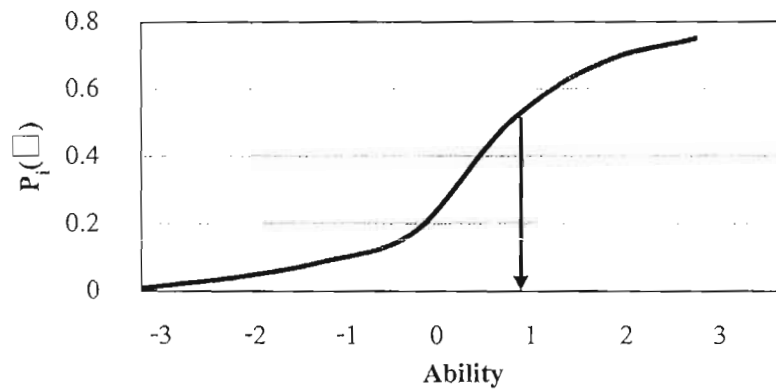
โมเดลปกติสะสมใช้ฟังก์ชันปกติสะสมแสดงความสัมพันธ์ระหว่างผลการตอบข้อสอบกับความสามารถของผู้สอบ ส่วนโมเดลโลจิสต์ใช้ฟังก์ชันโลจิสต์แสดงความสัมพันธ์ระหว่างผลการตอบกับความสามารถดังกล่าว ซึ่งฟังก์ชันทั้งสองให้ผลลัพธ์ของการประมาณค่าใกล้เคียงกันมาก แต่ฟังก์ชันโลจิสต์มีลักษณะของสูตรทางคณิตศาสตร์และวิธีการคำนวณง่ายและสะดวกกว่า นอกจากนี้โมเดลโลจิสต์ยังมีความทนทานต่อความคลาดเคลื่อนที่เกิดขึ้นกับผู้สอบที่มีความสามารถสูงจะตอบข้อสอบได้ดีกว่าจึงทำให้โมเดลโลจิสต์เป็นที่นิยมในการนำไปใช้จริง โมเดลการตอบข้อสอบ จำแนกเป็น 3 โมเดล คือ

1) โมเดลการตอบข้อสอบแบบ 1 พารามิเตอร์ (One-Parameter Model)

โมเดลการตอบข้อสอบแบบ 1 พารามิเตอร์ หรือ Rasch model มีโค้งลักษณะข้อสอบที่เขียนด้วยฟังก์ชันโลจิสต์ ดังสมการ

$$P_i(\theta) = \frac{1}{1 + e^{-(\theta - b_i)}}$$

- เมื่อ $P_i(\theta)$ แทนค่า ความน่าจะเป็นที่ผู้ตอบซึ่งมีความสามารถ θ จะตอบข้อสอบข้อที่ i ได้ถูกต้อง
 b_i แทนค่า ค่าพารามิเตอร์ความยากของข้อสอบข้อที่ i ซึ่งเป็นค่าแสดงตำแหน่งของ ICC
 ณ จุด θ ที่มีโอกาสตอบข้อสอบถูก 0.50
 e แทนค่า 2.718



ภาพ 2.1 โค้งลักษณะข้อสอบแบบ 1 พารามิเตอร์ของข้อสอบที่มีค่าพารามิเตอร์ความยาก (b) เท่ากับ 1.00
 ที่มา. จาก *The Basics of Item Response Theory* (p.27), by Baker, 2001, United States of America:
 ERIC Clearinghouse on Assessment and Evaluation.

โค้งลักษณะข้อสอบ (ICC) สำหรับโมเดลการตอบข้อสอบแบบ 1 พารามิเตอร์มีลักษณะที่สำคัญดังนี้

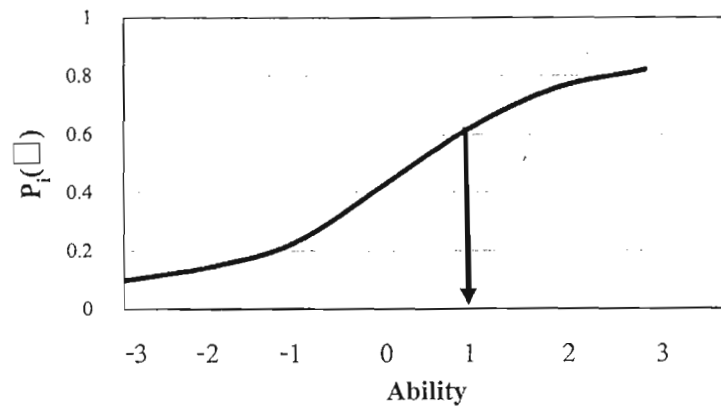
- b_i = มีค่าแปรเปลี่ยนตามลักษณะของข้อสอบแต่ละข้อ
 a_i = มีค่าคงที่
 c_i = 0

2) โมเดลการตอบข้อสอบแบบ 2 พารามิเตอร์ (Two-Parameter Model)

โมเดลการตอบข้อสอบแบบ 2 พารามิเตอร์ มีโค้งลักษณะข้อสอบที่เขียนด้วยฟังก์ชัน
 โลจิสต์ ดังสมการ

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

- เมื่อ $P_i(\theta)$ แทนค่า ความน่าจะเป็นที่ผู้ตอบซึ่งมีความสามารถ θ จะตอบข้อสอบข้อที่ i ได้ถูกต้อง
 b_i แทนค่า ค่าพารามิเตอร์ความยากของข้อสอบข้อที่ i ซึ่งเป็นค่าแสดงตำแหน่งของ ICC
 ณ จุด θ ที่มีโอกาสตอบข้อสอบถูก 0.50
 a_i แทนค่า ค่าพารามิเตอร์อำนาจจำแนกของข้อสอบข้อที่ i ซึ่งเป็นค่าความชันของ ICC
 ณ ตำแหน่ง b_i
 e แทนค่า 2.718
 D แทนค่า 1.70



ภาพ 2.2 โค้งลักษณะข้อสอบแบบ 2 พารามิเตอร์ของข้อสอบ

ที่มีค่าพารามิเตอร์ความยาก (b) เท่ากับ 1.00 ค่าพารามิเตอร์อำนาจจำแนก (a) เท่ากับ 1.50

ที่มา. จาก *The Basics of Item Response Theory* (p.25), by Baker, 2001, United States of America: ERIC Clearinghouse on Assessment and Evaluation.

โค้งลักษณะข้อสอบ (ICC) สำหรับ โมเดลการตอบข้อสอบแบบ 2 พารามิเตอร์มีลักษณะที่สำคัญดังนี้

- b_i = มีค่าแปรเปลี่ยนตามลักษณะของข้อสอบแต่ละข้อ
- a_i = มีค่าแปรเปลี่ยนตามลักษณะของข้อสอบแต่ละข้อ
- c_i = 0

3) โมเดลการตอบข้อสอบแบบ 3 พารามิเตอร์ (Three-Parameter Model)

โมเดลการตอบข้อสอบแบบ 3 พารามิเตอร์ มีโค้งลักษณะข้อสอบที่เขียนด้วยฟังก์ชัน

โลจิสต์ ดังสมการ

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

เมื่อ $P_i(\theta)$ แทนค่า ความน่าจะเป็นที่ผู้ตอบซึ่งมีความสามารถ θ จะตอบข้อสอบข้อที่ i ได้ถูกต้อง

b_i แทนค่า ค่าพารามิเตอร์ความยากของข้อสอบข้อที่ i ซึ่งเป็นค่าแสดงตำแหน่งของ ICC

ณ จุด θ ที่มีโอกาสตอบข้อสอบถูก $\frac{1+c_i}{2}$

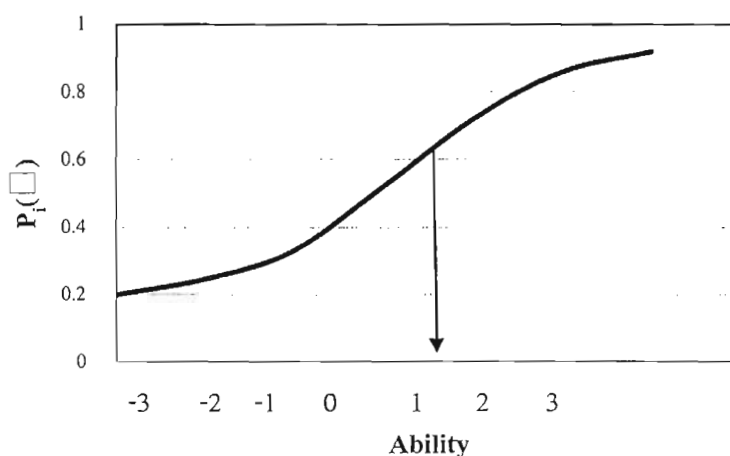
a_i แทนค่า ค่าพารามิเตอร์อำนาจจำแนกของข้อสอบข้อที่ i ซึ่งเป็นค่าความชันของ ICC

ณ ตำแหน่ง b_i

c_i แทนค่า ค่าพารามิเตอร์โอกาสเดาข้อสอบได้ถูก

e แทนค่า 2.718

D แทนค่า 1.70



ภาพ 2.3 โค้งลักษณะข้อสอบแบบ 3 พารามิเตอร์ของข้อสอบที่มี
ค่าพารามิเตอร์ความยาก (b) เท่ากับ 1.50 ค่าพารามิเตอร์อำนาจจำแนก (a) เท่ากับ 1.30
ค่าพารามิเตอร์โอกาสเดาข้อสอบได้ถูก (c) เท่ากับ 0.20

ที่มา. จาก *The Basics of Item Response Theory* (p.31), by Baker, 2001, United States of America:

ERIC Clearinghouse on Assessment and Evaluation.

โค้งลักษณะข้อสอบ (ICC) สำหรับ โมเดลการตอบข้อสอบแบบ 3 พารามิเตอร์มีลักษณะที่สำคัญดังนี้

- b_i = มีค่าแปรเปลี่ยนตามลักษณะของข้อสอบแต่ละข้อ
- a_i = มีค่าแปรเปลี่ยนตามลักษณะของข้อสอบแต่ละข้อ
- c_i = มีค่าแปรเปลี่ยนตามลักษณะของข้อสอบแต่ละข้อ

ข้อตกลงเบื้องต้นและการตรวจสอบ

ฟังก์ชันทางคณิตศาสตร์ที่ใช้ใน IRT กำหนดไว้ว่าความน่าจะเป็นของการตอบข้อสอบได้ถูกต้องขึ้นอยู่กับความสามารถของผู้สอบและคุณลักษณะของข้อสอบ โมเดลการตอบข้อสอบจึงตั้งอยู่บนฐานความเชื่อหรือข้อตกลงเบื้องต้นหลายประการเกี่ยวกับลักษณะข้อมูลที่จะทำให้โมเดลสามารถนำไปใช้ได้เหมาะสม ถึงแม้ข้อตกลงเบื้องต้นบางประการจะตรวจสอบไม่ได้โดยตรง แต่เราสามารถเก็บรวบรวมข้อมูลหรือหลักฐานทางอ้อมช่วยสนับสนุนยืนยันได้ ข้อตกลงเบื้องต้นที่สำคัญของ IRT มีดังนี้

1) ความเป็นเอกมิติ (Unidimensionality)

ข้อตกลงเบื้องต้นที่ใช้กันทั่วไปสำหรับ IRT คือข้อคำถามหรือข้อสอบทุกข้อในเครื่องมือหรือ

แบบทดสอบมุ่งวัดเพียงคุณลักษณะเดียว หรือ ความสามารถเดียว (one ability) ซึ่งเรียกว่าความเป็นเอกมิติ (unidimensionality)

การตรวจสอบความเป็นเอกมิติของเครื่องมือ หรือแบบทดสอบสามารถกระทำได้โดยใช้เทคนิคการวิเคราะห์ทางสถิติ ได้แก่ การวิเคราะห์ตัวประกอบ (Factor Analysis) เพื่อคำนวณค่าไอเกน (eigen value) สำหรับศึกษาอัตราส่วนที่สูงแสดงถึงเครื่องมือหรือแบบทดสอบวัดคุณลักษณะเด่นเดียว (single dominant factor) หรือทำการวิเคราะห์ให้เกิดความมั่นใจยิ่งขึ้นด้วยการวิเคราะห์หองค์ประกอบเชิงยืนยัน (confirmatory factor analysis) เพื่อตรวจสอบยืนยันว่าเครื่องมือหรือแบบทดสอบมุ่งวัดเพียงคุณลักษณะเดียวหรือความสามารถเดียว

2) ความเป็นอิสระ (Local Independence)

แนวคิดเกี่ยวกับความเป็นอิสระระหว่างข้อสอบและผู้สอบมีความเกี่ยวข้องและเชื่อมโยงจากความเป็นเอกมิติของแบบทดสอบ ความเป็นอิสระระหว่างข้อสอบและผู้สอบ หมายถึงเมื่อมีการควบคุมความสามารถ θ ที่ส่งผลต่อการตอบข้อสอบหรือให้ θ คงที่แล้ว ผลการตอบข้อสอบแต่ละข้อจะต้องเป็นอิสระจากกัน หรือกล่าวอีกนัยหนึ่งได้ว่าเมื่อควบคุมอิทธิพลของ θ แล้วผลการตอบข้อสอบรายข้อไม่มีความสัมพันธ์กัน นั่นคือ โมเดลการตอบสนองข้อสอบมีเพียง θ ปัจจัยเดียวเท่านั้นที่มีอิทธิพลต่อการตอบรายข้อ ความเป็นอิสระสามารถจำแนกการพิจารณาเป็นความเป็นอิสระระหว่างข้อสอบและความอิสระระหว่างผู้สอบ ดังต่อไปนี้

2.1) ความเป็นอิสระระหว่างข้อสอบ

เมื่อกลุ่มผู้สอบซึ่งมีความสามารถ θ ขึ้นมา 1 คน ในการตอบข้อสอบ k ข้อ ให้ U_j เป็นผลการตอบหรือคะแนนข้อสอบข้อที่ j หลังจากควบคุม θ ของผู้สอบแล้ว คะแนนผลการตอบของผู้สอบนั้นในแต่ละข้อไม่สัมพันธ์กัน

$$\begin{aligned} \text{ดังนั้น } P(U_1, U_2, \dots, U_k / \theta) &= P(U_1 / \theta)P(U_2 / \theta) \dots P(U_k / \theta) \\ &= \pi_{j=1}^k P(U_j / \theta) \end{aligned}$$

ถ้าผลการตอบข้อสอบรายข้อของผู้สอบคนเดียวกันเป็นอิสระจากกันความน่าจะเป็นของแบบแผนการตอบข้อสอบ k ข้อของผู้สอบที่มีความสามารถ θ จะเท่ากับผลคูณระหว่างความน่าจะเป็นของผลการตอบข้อสอบแต่ละข้อ

2.2) ความเป็นอิสระระหว่างผู้สอบ

เมื่อกลุ่มข้อสอบขึ้นมา 1 ข้อ ในการตอบข้อสอบของผู้สอบ n คน ให้ U_i เป็นผลการตอบหรือคะแนนข้อสอบของผู้สอบคนที่ i หลังจากควบคุม θ ของผู้สอบแต่ละคนแล้ว คะแนนผลการตอบของผู้สอบแต่ละคนไม่สัมพันธ์กัน

$$\begin{aligned} \text{ดังนั้น } P(U_1, U_2, \dots, U_n / \theta) &= P(U_1 / \theta)P(U_2 / \theta) \dots P(U_n / \theta) \\ &= \pi_{i=1}^n P(U_i / \theta) \end{aligned}$$

ถ้าผลการตอบข้อสอบข้อเดียวกันของผู้สอบแต่ละคนเป็นอิสระจากกันความน่าจะเป็นของแบบแผนการตอบข้อสอบของผู้สอบ n คน จะเท่ากับผลคูณระหว่างความน่าจะเป็นของผลการตอบข้อนั้นของผู้สอบแต่ละคน

การตรวจสอบความเป็นอิสระระหว่างข้อสอบและผู้สอบสามารถทำได้โดยการพิจารณาเมตริกซ์ ความแปรปรวนและความแปรปรวนร่วม (variance-covariance matrix) หรือเมตริกซ์สหสัมพันธ์ (correlation matrix) ของคะแนนคำตอบรายข้อ สำหรับกลุ่มผู้สอบที่มีช่วงความสามารถเท่ากัน โดยค่านอกแนวทแยงควรมีค่าต่ำหรือเข้าใกล้ 0

3) โมเดลการตอบข้อสอบ (Item Response Models)

IRT อยู่บนฐานความเชื่อว่าฟังก์ชันลักษณะข้อสอบ หรือ โก้งลักษณะข้อสอบ (ICC) สามารถสะท้อนความสัมพันธ์จริงระหว่างความสามารถของผู้สอบกับลักษณะของข้อสอบและผลการตอบข้อสอบ โมเดลการตอบข้อสอบเสนอ ICC ซึ่งเป็นฟังก์ชันโลจิสต์ด้วยสัญลักษณ์ที่แตกต่างตามจำนวนพารามิเตอร์ที่ใช้บรรยายลักษณะของข้อสอบ โมเดลการตอบข้อสอบแบบตรวจให้คะแนน 2 ค่า (dichotomous item response model) ที่ใช้กันแพร่หลาย ได้แก่ โมเดลการตอบข้อสอบแบบ 1 พารามิเตอร์ 2 พารามิเตอร์ และ 3 พารามิเตอร์ การเลือกใช้จึงขึ้นอยู่กับจุดมุ่งหมายของงานและธรรมชาติของข้อมูล

3.1) โมเดลการตอบข้อสอบแบบ 1 พารามิเตอร์ มีข้อตกลงเบื้องต้นว่าข้อสอบแต่ละข้อมีพารามิเตอร์ $c = 0$ และค่าพารามิเตอร์ a เท่ากัน แต่มีความแตกต่างกันเฉพาะพารามิเตอร์ b เท่านั้น โมเดลนี้จึงเหมาะสำหรับใช้กับข้อสอบอิงเกณฑ์ที่ไม่สลับซับซ้อน ข้อสอบที่ค่อนข้างเรียบง่ายสำหรับพัฒนาคลังข้อสอบที่มีความเป็นเอกพันธ์

3.2) โมเดลการตอบข้อสอบแบบ 2 พารามิเตอร์ มีข้อตกลงเบื้องต้นว่าข้อสอบแต่ละข้อมีพารามิเตอร์ $c = 0$ มีความแตกต่างกันของค่าพารามิเตอร์ a และ b โมเดลนี้จึงเหมาะสำหรับใช้กับข้อสอบที่ต้องเติมคำตอบ หรือข้อสอบแบบเลือกตอบที่ไม่ยากมากนักและกลุ่มผู้สอบมีความพร้อมในการตอบ

3.3) โมเดลการตอบข้อสอบแบบ 3 พารามิเตอร์ มีข้อตกลงเบื้องต้นว่าข้อสอบแต่ละข้อมีความแตกต่างกันของพารามิเตอร์ค่าพารามิเตอร์ a , b และ c โมเดลนี้จึงเหมาะสำหรับใช้กับข้อสอบแบบเลือกตอบทั่วไป ข้อสอบหลายตัวเลือก เนื่องจากผู้สอบสามารถเดาคำตอบได้

สำหรับการตรวจสอบความเหมาะสมของโมเดลการตอบข้อสอบกับข้อมูลไม่ว่าจะเลือกใช้โมเดลใดก็ตาม โมเดลการตอบสนองข้อสอบต้องสอดคล้องกับข้อมูลเพื่อให้ผลการวิเคราะห์มีความถูกต้องและน่าเชื่อถือ ในการประมาณค่าพารามิเตอร์โดยใช้หลักการความเป็นไปได้สูงสุด (maximum likelihood) ค่า $-2 \log \text{likelihood}$ ของข้อมูลแสดงถึงระดับความไม่สอดคล้องของข้อมูลเชิงประจักษ์กับโมเดลจึงเป็นค่าที่มีประโยชน์สำหรับประเมินความเหมาะสมของโมเดล ค่า $-2 \log \text{likelihood}$ ที่แตกต่างระหว่างโมเดลคำนวณด้วยสถิติทดสอบ χ^2 เพื่อประเมินว่าการปรับปรุงโมเดลที่ซับซ้อนขึ้นจะมีคุณภาพหรือความ

เหมาะสมเพิ่มขึ้นเพียงใด เมื่อ df สำหรับการทดสอบ χ^2 คือจำนวนพารามิเตอร์ที่เพิ่มขึ้นของโมเดลที่ซับซ้อนขึ้น เช่นการเปรียบเทียบความเหมาะสมระหว่างโมเดล 1, 2 และ 3 พารามิเตอร์ สำหรับแบบทดสอบการใช้เหตุผลเชิงนามธรรม จำนวน 30 ข้อ (ศิริชัย กาญจนวาสี. 2555, หน้า 61)

การเปรียบเทียบความเหมาะสมระหว่างโมเดล 1 พารามิเตอร์ กับ 2 พารามิเตอร์

$$\begin{aligned} \chi^2_{I-II} &= 25,924.52 - 25,659.39 \\ &= 265.13 \quad (df = 30) \end{aligned}$$

การทดสอบมีนัยสำคัญทางสถิติที่ $\alpha = 0.01$ ($\chi^2_{0.01,30} = 50.89$)

แสดงว่าโมเดลการตอบข้อสอบแบบ 2 พารามิเตอร์สอดคล้องกับข้อมูลเชิงประจักษ์มากกว่าโมเดล 1 พารามิเตอร์

การเปรียบเทียบความเหมาะสมระหว่างโมเดล 2 พารามิเตอร์ กับ 3 พารามิเตอร์

$$\begin{aligned} \chi^2_{II-III} &= 25,659.39 - 25,650.60 \\ &= 8.79 \quad (df = 30) \end{aligned}$$

การทดสอบมีนัยสำคัญทางสถิติที่ $\alpha = 0.05$ ($\chi^2_{0.05,30} = 43.77$)

แสดงว่าโมเดลการตอบข้อสอบแบบ 3 พารามิเตอร์สอดคล้องกับข้อมูลเชิงประจักษ์ไม่แตกต่างจากโมเดล 2 พารามิเตอร์

4) การสอบที่ไม่แข่งขันด้านเวลา (Nonspeeded Test Administration)

IRT ถือว่าความสามารถ θ เป็นปัจจัยสำคัญต่อผลการสอบ ความเร็วในการตอบจะต้องไม่มีอิทธิพลต่อผลการสอบ การจัดการสอบจึงต้องไม่อยู่ในสถานการณ์ที่สอบแข่งขันกันด้วยเวลาการสอบ จะต้องอยู่ในลักษณะที่ผู้สอบซึ่งมีความสามารถมีเวลาเพียงพอในการทำข้อสอบ

การตรวจสอบถึงความเหมาะสมของมิติด้านเวลา สำหรับการดำเนินการสอบสามารถพิจารณาได้จาก สัดส่วนหรือร้อยละของจำนวนผู้สอบที่ทำข้อสอบได้ครบทุกข้อ โดยผู้สอบส่วนใหญ่ (ร้อยละ 80 เป็นต้น) สามารถตอบข้อสอบได้ครบหรือเกือบครบทุกข้อ นอกจากนี้ควรพิจารณาเปรียบเทียบระหว่างความแปรปรวนของจำนวนข้อที่เว้นกับความแปรปรวนของจำนวนข้อที่ตอบผิด ถ้าสัดส่วนของความแปรปรวนเข้าใกล้ 0 แสดงว่าการจัดการสอบเป็นไปตามข้อตกลงเบื้องต้น

แนวทางการคัดเลือกข้อสอบจากค่าพารามิเตอร์ของข้อสอบ

1) ค่าพารามิเตอร์อำนาจจำแนก (a)

ศิริชัย กาญจนวาสี (2555, หน้า 55) กล่าวว่า ในทางทฤษฎีมีค่าพารามิเตอร์อำนาจจำแนก (a) อยู่

ระหว่าง $(-\infty, +\infty)$ ควรมีค่าเป็นบวก ในทางปฏิบัตินิยมใช้ข้อสอบที่มีค่าพารามิเตอร์อำนาจจำแนกอยู่ระหว่าง 0.50 ถึง 2.50

Baker (2001, p.34) กำหนดการแปลความหมายค่าพารามิเตอร์อำนาจจำแนก (a) เป็น 5 ระดับ คือ

0.01 – 0.34	ระดับค่อนข้างต่ำ	0.35 – 0.64	ระดับต่ำ
0.65 – 1.34	ระดับปานกลาง	1.35 – 1.69	ระดับค่อนข้างมาก
มากกว่า 1.70	ระดับมาก		

2) ค่าพารามิเตอร์ความยาก (b) ในทางปฏิบัตินิยมใช้ข้อสอบที่มีค่าพารามิเตอร์ความยากอยู่

ระหว่าง (-3.00) ถึง 3.00 (Baker, 2001, p.22; De Ayala, 2009, p.101)

ศิริชัย กาญจนวาสี (2555, หน้า 55) กล่าวว่า ในทางทฤษฎีมีค่าพารามิเตอร์ความยาก (b) ระหว่าง $(-\infty, +\infty)$ ควรมีค่าเป็นบวก ในทางปฏิบัตินิยมใช้ข้อสอบที่มีค่าพารามิเตอร์อำนาจจำแนกอยู่ระหว่าง (-2.50) ถึง 2.50

3) ค่าพารามิเตอร์โอกาสในการเดาข้อสอบได้ถูก (guessing parameter)

ศิริชัย กาญจนวาสี (2555, หน้า 55) กล่าวว่า ในทางทฤษฎีมีค่าพารามิเตอร์โอกาสในการเดาข้อสอบได้ถูก(guessing parameter) ระหว่าง 0.00 ถึง 1.00 โดยทั่วไปนิยมใช้ข้อสอบที่มีค่าพารามิเตอร์โอกาสในการเดาข้อสอบได้ถูก ไม่เกิน 0.30

Baker (2001, p.28) กล่าวว่า ในทางทฤษฎีมีค่าพารามิเตอร์โอกาสในการเดาข้อสอบได้ถูก (guessing parameter) ระหว่าง 0.00 ถึง 1.00 โดยทั่วไปนิยมใช้ข้อสอบที่มีค่าพารามิเตอร์โอกาสในการเดาข้อสอบได้ถูก ไม่เกิน 0.35

คุณสมบัติความไม่แปรเปลี่ยนของค่าพารามิเตอร์ (invariance)

เมื่อ IRT model สอดคล้อง (fit) กับข้อมูลจะทำให้เกิดความไม่แปรเปลี่ยนของพารามิเตอร์ของข้อสอบ (item parameter) และพารามิเตอร์ความสามารถของผู้สอบ (ability parameter) ซึ่งเป็นคุณสมบัติของ IRT (ศิริชัย กาญจนวาสี. 2555, หน้า 62-63)

1) ความไม่แปรเปลี่ยนของพารามิเตอร์ข้อสอบ (item invariance)

- ค่าพารามิเตอร์ของข้อสอบไม่แปรเปลี่ยนไปตามกลุ่มผู้สอบ
- ICC ลักษณะเดียวกัน (a, b, c) สำหรับทุกกลุ่มความสามารถของผู้สอบ แสดงว่า ICC มีความคงที่ข้ามกลุ่มผู้สอบ

2) ความไม่แปรเปลี่ยนของพารามิเตอร์ความสามารถของผู้สอบ (ability invariance)

- ค่าพารามิเตอร์ของผู้สอบไม่แปรเปลี่ยนไปตามชุดของผู้สอบ
- เมื่อนำข้อสอบต่างชุด (ทุกข้อมุ่งวัดสิ่งเดียวกัน) เช่น ข้อสอบชุดค่อนข้างง่ายกับข้อสอบชุดค่อนข้างยากมาสอบวัดผู้สอบกลุ่มเดียวกัน ค่า θ ที่ประมาณได้จากข้อสอบทั้ง 2 ชุดมีความแตกต่างกันไม่เกิน SEE แสดงว่าการประมาณค่าความสามารถมีความคงที่ข้ามชุดของข้อสอบ

ฟังก์ชันสารสนเทศของข้อสอบและแบบทดสอบ (Item and Test Information Functions)

1) ฟังก์ชันสารสนเทศของข้อสอบ (Item Information Function)

ฟังก์ชันสารสนเทศของข้อสอบเป็นดัชนีผสมที่สร้างจากดัชนีคุณลักษณะของข้อสอบหลายลักษณะ ประกอบด้วย ค่าพารามิเตอร์ความยาก ค่าพารามิเตอร์อำนาจจำแนก และค่าความแปรปรวนของคะแนนรายข้อ เพื่อใช้บ่งชี้คุณภาพของข้อสอบ

$$I_i(\theta) = \frac{[\hat{P}_i(\theta)]^2}{P_i(\theta)Q_i(\theta)} \quad i = 1, 2, \dots, k$$

เมื่อ $I_i(\theta)$ แทนค่า ค่าฟังก์ชันสารสนเทศหรือค่าสารสนเทศที่ได้รับจากข้อสอบข้อที่ i สำหรับผู้ตอบที่มีความสามารถ θ

$\hat{P}_i(\theta) = \hat{P}_i$ แทนค่า ความชันของฟังก์ชันการตอบสนองข้อสอบข้อที่ i ณ ตำแหน่งความสามารถ θ

$P_i(\theta) = P_i$ แทนค่า ความน่าจะเป็นที่ผู้ตอบซึ่งมีความสามารถ θ จะตอบข้อสอบข้อที่ i ได้ถูกต้อง

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

ฟังก์ชันสารสนเทศของข้อสอบ เมื่อพิจารณาเทียบกับค่าพารามิเตอร์ a , b และ c ของข้อสอบ พอสรุปเป็นแนวทางโดยทั่วไปได้ ดังนี้

- (1) ค่าสารสนเทศของข้อสอบจะสูงขึ้น สำหรับผู้สอบมีความสามารถ θ ใกล้กับค่าพารามิเตอร์ b ของข้อสอบ และค่าสารสนเทศของข้อสอบจะลดลง สำหรับผู้สอบที่มีความสามารถ θ ไกลจากค่าพารามิเตอร์ b ของข้อสอบ
- (2) ค่าสารสนเทศของข้อสอบโดยทั่วไปจะมีค่าสูงขึ้น ถ้าค่าพารามิเตอร์ a ของข้อสอบมีค่ามากขึ้น
- (3) ค่าสารสนเทศของข้อสอบมีค่าสูงขึ้น เมื่อค่าพารามิเตอร์ c ของข้อสอบเข้าใกล้ 0
- (4) $I_i(\theta)$ จะมีค่าสูงสุด ณ ตำแหน่ง $\theta_{\max}$ ถ้า $c_i = 0$, $\theta_{\max} = b$ แต่ถ้า $c_i > 0$, $\theta_{\max} > b$

2) ฟังก์ชันสารสนเทศของแบบทดสอบ (Test Information Function)

การวิเคราะห์ตามทฤษฎี IRT จะใช้แบบแผนการตอบสนองแบบทดสอบเป็นรายข้อในการประมาณค่าความสามารถของผู้สอบ ดังนั้นการประเมินคุณภาพของแบบทดสอบ จึงสามารถพิจารณาจากความถูกต้องแม่นยำในการประมาณความสามารถของผู้ตอบ โดยใช้ดัชนีตัวหนึ่งเรียกว่า สารสนเทศของแบบทดสอบ [test information: $I(\theta)$] ซึ่งเป็นค่าฟังก์ชันสารสนเทศของแบบทดสอบ อันเกิดจากผลรวมเชิงพีชคณิตของค่าฟังก์ชันสารสนเทศของข้อสอบแต่ละข้อรวมเข้าด้วยกันทั้งฉบับ ณ ตำแหน่ง θ เดียวกัน ดังสูตร

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

เมื่อ $I(\theta)$ แทนค่า ค่าฟังก์ชันสารสนเทศ หรือค่าสารสนเทศที่ได้รับจากแบบทดสอบสำหรับผู้ตอบที่มีความสามารถ θ

ค่าฟังก์ชันสารสนเทศของข้อสอบแต่ละข้อจึงมีส่วนอย่างเป็นอิสระจากกันต่อค่าสารสนเทศของแบบทดสอบ ลักษณะเช่นนี้ไม่ได้เกิดกับทฤษฎีการทดสอบแบบดั้งเดิม เนื่องจากค่าความยากและค่าอำนาจจำแนกของข้อสอบแต่ละข้อต่างก็ส่งผลกระทบต่อความเชื่อมั่นของแบบทดสอบทั้งฉบับ แต่ไม่สามารถคำนวณค่าของแต่ละข้อได้เป็นอิสระจากกัน ดังนั้นคะแนนที่ได้จึงขึ้นอยู่กับลักษณะเฉพาะของกลุ่มข้อสอบและแบบทดสอบเฉพาะฉบับที่เลือกมาใช้

เนื่องจากค่าสารสนเทศมีความสัมพันธ์ผกผันกับความคลาดเคลื่อนมาตรฐานของการประมาณค่า ดังนั้นถ้าค่าสารสนเทศของแบบทดสอบมีค่าสูงในช่วง θ ใดก็必将มีความถูกต้องแม่นยำสูง ในการประมาณค่าความสามารถของผู้ตอบในช่วง θ นั้นๆ โดยมีความคลาดเคลื่อนมาตรฐานของการประมาณค่าต่ำ

คะแนนความสามารถ และความคลาดเคลื่อนมาตรฐานของการประมาณค่า

(Ability score and standard error of estimation)

สเกลความสามารถของผู้ตอบ (Ability scales)

สเกลความสามารถของผู้ตอบจากการทำข้อสอบแบบ 0,1 จำนวน k ข้อ มีสัญลักษณ์และการประมาณค่า ดังนี้

จาก CTT : $X = T + E$

เมื่อ $X =$ คะแนนที่สังเกตได้ (observed score)

- คะแนนรวมของข้อที่ตอบถูก
- มีค่าอยู่ระหว่าง $[0, k]$

$T =$ คะแนนจริง (true score)

- มีค่าอยู่ระหว่าง $[0, k]$

$\pi =$ คะแนนปริเขต (domain score)

- สัดส่วนของการตอบถูก
- มีค่าอยู่ระหว่าง $[0\%, 100\%]$

จาก IRT : $\theta =$ คะแนนความสามารถจริง (ability score)

- คะแนนมาตรฐาน (standard score) นิยมคำนวณให้มีค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานเป็น 0 และ 1 ตามลำดับ
- มีค่าอยู่ระหว่าง $[-\infty, +\infty]$

$T = \sum_{i=1}^k P_i(\theta)$

- คะแนนจริง
- มีค่าอยู่ระหว่าง $[\sum C_i, k]$

$\pi = \frac{1}{k} \sum_{i=1}^k P_i(\theta)$

- คะแนนปริเขตคิดเป็นเปอร์เซ็นต์
- มีค่าอยู่ระหว่าง $[\frac{1}{100} (\sum c_i) \%, 100\%]$

ความคลาดเคลื่อนมาตรฐานของการประมาณค่า (standard error of estimation)

ความคลาดเคลื่อนมาตรฐานของการประมาณค่า $[SE(\theta)]$ เป็นค่าส่วนเบี่ยงเบนมาตรฐานของการแจกแจงความน่าจะเป็นของค่าประมาณความสามารถที่แท้จริง (θ) ซึ่งเป็นค่าสัดส่วนผกผันกับความถูกต้องแม่นยำของการประมาณค่าความสามารถหรือค่าฟังก์ชันสารสนเทศของแบบทดสอบ ดังสูตร

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

เมื่อ $SE(\theta)$ = ความคลาดเคลื่อนมาตรฐานของการประมาณค่า สำหรับผู้ตอบที่มีความสามารถ θ

$I(\theta)$ = สารสนเทศที่ได้จากแบบทดสอบ สำหรับผู้ตอบที่มีความสามารถ θ

การสอบทุกครั้งย่อมมีความคลาดเคลื่อนในการประมาณค่าความสามารถ θ ความคลาดเคลื่อนมาตรฐานของการประมาณค่าความสามารถ มีค่าเท่ากับส่วนกลับของรากที่สองของค่าฟังก์ชันสารสนเทศของแบบทดสอบ ดังนั้นค่าฟังก์ชันสารสนเทศของแบบทดสอบมีค่าสูง ความคลาดเคลื่อนมาตรฐานของการประมาณค่าความสามารถจะต่ำ นั่นคือ ถ้าค่าฟังก์ชันสารสนเทศของแบบทดสอบมีค่าสูง ณ ผู้สอบระดับความสามารถ θ ใด (ฟังก์ชันการตอบสนองข้อสอบแต่ละข้อมีความชันสูงและความแปรปรวนต่ำ ณ ระดับ θ นั้น) ความคลาดเคลื่อนมาตรฐานของการประมาณค่าความสามารถของผู้สอบ ณ ระดับ θ นั้นจะต่ำหรือมีความแม่นยำสูงในการประมาณค่าความสามารถของผู้สอบ θ นั้น

$SE(\theta)$ เป็นค่าที่คล้ายกับความคลาดเคลื่อนมาตรฐานของการวัด (standard error of measurement: SEM) ใน CTT แตกต่างกันตรงที่ $SE(\theta)$ มีค่าผันแปรไปตามตำแหน่ง θ ในขณะที่ SEM เป็นค่าคงที่สำหรับผู้ตอบทุกคน

ขนาดของ $SE(\theta)$ มีลักษณะโดยทั่วไปดังนี้

- (1) แบบทดสอบที่มีจำนวนข้อสอบมากขึ้น ขนาดของ $SE(\theta)$ จะลดลง
- (2) แบบทดสอบที่มีข้อสอบที่ a สูง, c ต่ำ ขนาดของ $SE(\theta)$ จะลดลง
- (3) แบบทดสอบที่ประกอบด้วยข้อสอบที่ b ใกล้เคียงกับ θ ของผู้สอบ (ไม่ง่ายหรือไม่ยากเกินไป)

ขนาดของ $SE(\theta)$ จะลดลง

- (4) $I(\theta)$ ที่มีค่าเกิน 25 จะทำให้ขนาดของ $SE(\theta)$ ค่อนข้างคงที่และมีผลน้อยต่อความคลาดเคลื่อนในการประมาณค่า θ

1.3 การออกแบบการสร้างแบบทดสอบตามแนวคิดของ Birnbaum

ค.ศ. 1968 Birnbaum เสนอวิธีการออกแบบการสร้างแบบทดสอบที่ใช้ประโยชน์ของความสัมพันธ์เชิงบวกของฟังก์ชันสารสนเทศของข้อสอบ $[I_i(\theta)]$ กับแบบทดสอบ $[I(\theta)]$ ซึ่งแนวคิดนี้ได้ถูกนำมา

ประยุกต์ใช้ในการจัดชุดแบบทดสอบ (Automated Test Assembly: ATA) (van der Linden, 2005, p.21-22) ประกอบด้วย 3 ขั้นตอน คือ

1) กำหนดเป้าหมายของการจัดชุดแบบทดสอบ จำแนกเป็น

- 1.1) การจัดชุดแบบทดสอบเพื่อนำไปจัดตำแหน่งผู้เรียน
- 1.2) การจัดชุดแบบทดสอบเพื่อนำไปวินิจฉัยระดับเชาวน์ปัญญา
- 1.3) การจัดชุดแบบทดสอบเพื่อนำไปคัดเลือกบุคคล
- 1.4) การจัดชุดแบบทดสอบเพื่อให้ได้แบบทดสอบฉบับใหม่ที่มีคุณสมบัติของความเป็น
คู่ขนาน

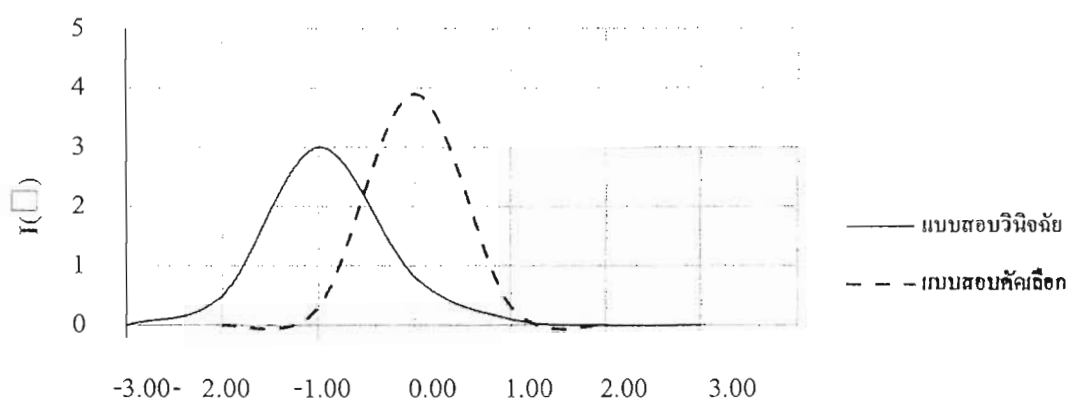
2) ดำเนินการแปลงเป้าหมายของการจัดชุดแบบทดสอบเป็นฟังก์ชันสารสนเทศของแบบทดสอบ (TIF) เช่น

2.1) กรณีกำหนดเป้าหมายของการจัดชุดแบบทดสอบเพื่อนำไปวินิจฉัยระดับเชาวน์ปัญญา

ผู้วิจัยต้องกำหนดระดับค่าความสามารถของผู้สอบ (θ) ที่จะวินิจฉัยว่ามีความเข้าใจที่คลาดเคลื่อนจากโค้ง TIF

2.2) กรณีกำหนดเป้าหมายของการจัดชุดแบบทดสอบเพื่อการคัดเลือกบุคคล ผู้วิจัยต้องกำหนดคะแนนจุดตัดเพื่อนำไปใช้ในการตัดสินใจเพื่อการคัดเลือกบุคคลจากโค้ง TIF

การแปลงเป้าหมายของการจัดชุดแบบทดสอบเป็น TIF สามารถแสดงดังภาพ



ภาพ 2.4 การกำหนดเป้าหมายของการจัดชุดแบบทดสอบวินิจฉัยและแบบทดสอบคัดเลือก
ที่มา. จาก *Linear Models for Optimal Test Design* (p.22), by van der Linden, 2005, United States of
America: Springer.

3) รวบรวมข้อสอบจากคลังข้อสอบ โดยข้อสอบจะถูกคัดเลือกจากฟังก์ชันสารสนเทศของข้อสอบ

(Item Information Function: IIFs) เพื่อให้ได้ฟังก์ชันสารสนเทศของแบบทดสอบ (Test Information Function: TIF) ตามที่เป้าหมายของการสร้างแบบทดสอบ

van der Linden (2005, p.22-23) กล่าวว่า การออกแบบการสร้างแบบทดสอบตามแนวคิดของ Bimbaum มีข้อจำกัดในทางปฏิบัติ ดังนี้

(1) กระบวนการคัดเลือกข้อสอบที่ต้องทำการสุ่มข้อสอบจากคลังข้อสอบและนำคำนวณค่า TIF เพื่อให้ได้ค่า TIF ที่ต้องการนั้น การสุ่มข้อสอบหากใช้หลักการของการจัดหมู่ (combination) คือ $\frac{N!}{n!(N-n)!}$ เมื่อ N คือ ข้อสอบจากคลังข้อสอบ และ n คือ ความยาวของแบบทดสอบ ซึ่งในทางปฏิบัติจะได้วิธีการจัดชุดแบบทดสอบจำนวนมาก แต่เป็นการทดลองเลือกข้อสอบแบบลองผิดลองถูก

(2) หากต้องการจัดชุดแบบทดสอบที่มีเงื่อนไขที่ซับซ้อน เช่น แบบทดสอบฉบับนั้นประกอบด้วยเนื้อหาการวัด 7 เนื้อหา แต่ละเนื้อหามีการกระจายประเภทของข้อสอบ ซึ่งการจัดชุดแบบทดสอบที่มีการลองผิดลองถูกนั้นจะไม่เหมาะสมในทางปฏิบัติ

1.4 การออกแบบการสร้างแบบทดสอบตามแนวคิดของ van der Linden

van der Linden กล่าวว่า การออกแบบการสร้างแบบทดสอบด้วยการจัดชุดแบบทดสอบโดยอาศัยโปรแกรมคอมพิวเตอร์ ควรประกอบด้วย 4 ขั้นตอน คือ

1) กำหนดชุดของคุณลักษณะเฉพาะของแบบทดสอบที่ต้องการสร้างขึ้น ประกอบด้วย โครงสร้างเนื้อหาในการวัด ค่าพารามิเตอร์ของข้อสอบและแบบทดสอบ และการกำหนดระดับความซ้ำซ้อนของการใช้ข้อสอบซ้ำระหว่างแบบทดสอบต่างฉบับ

2) ออกแบบการสร้างคลังข้อสอบตามคุณลักษณะของแบบทดสอบที่กำหนดขึ้น

3) พัฒนาคลังข้อสอบ ซึ่งประกอบด้วยขั้นตอน คือ การเขียนข้อสอบ การทบทวนและปรับปรุงข้อสอบ การทดลองใช้ข้อสอบ การเปรียบเทียบคะแนน และวิเคราะห์คุณภาพของข้อสอบและแบบทดสอบตามโมเดล 1 พารามิเตอร์ 2 พารามิเตอร์ หรือ 3 พารามิเตอร์ที่สอดคล้องกับลักษณะของข้อสอบ

4) จัดชุดแบบทดสอบตามคุณลักษณะของแบบทดสอบที่กำหนดขึ้น

วิธีการจัดชุดแบบทดสอบ (Modes of Test Assembly)

การจัดชุดแบบทดสอบจากคลังข้อสอบ สามารถจำแนกเป็น 4 วิธีการ โดยวิธีที่ 1 และ 2 ได้รับความนิยมในการจัดชุดแบบทดสอบเพื่อทดสอบเป็นรายบุคคล แต่กระบวนการจัดชุดผู้สร้างข้อสอบต้องจัดชุดด้วยตนเอง ส่วนวิธีที่ 3 และ 4 ได้รับความนิยมเมื่อกระบวนการจัดชุดนั้นผู้สร้างใช้คอมพิวเตอร์เข้ามาช่วย ดังรายละเอียด

วิธีการที่ 1 การจัดชุดแบบทดสอบด้วยการสุ่มข้อสอบ (Random Sampling of Tests)

วิธีการนี้ใช้เมื่อคลังข้อสอบมีจำนวนข้อสอบที่ตายตัว และทำการสุ่มข้อสอบให้สอดคล้องกับคุณลักษณะของผู้สอบเป็นรายบุคคล ทฤษฎีการทดสอบที่ใช้เพื่อนำมาควบคุมแหล่งความคลาดเคลื่อนจากการวัด คือ การทดสอบแบบดั้งเดิม (CTT) ด้วยการกำหนดให้กลุ่มผู้สอบคงที่ ภายใต้ข้อตกลงเบื้องต้น คือ

คะแนนจริง (T_{ij}) เท่ากับค่าสัดส่วนของการตอบถูก (π_j) หากมีคลังข้อสอบขนาดใหญ่ค่าพารามิเตอร์นี้จะแสดงถึง “ระดับความรู้เป็นรายบุคคล” ข้อจำกัดของวิธีการที่ 1 คือ ค่าพารามิเตอร์ (π_j) ผันแปรตามกลุ่มผู้สอบ

วิธีการที่ 2 การจัดชุดแบบทดสอบด้วยการสุ่มข้อสอบแบบลำดับ (Sequential Sampling of Tests)

วิธีการนี้ใช้ทฤษฎีการทดสอบแบบดั้งเดิม (CTT) โดยใช้หลักการของการประมาณค่าความคลาดเคลื่อนของค่า π_j จาก $\hat{\pi}_j$ โดยที่ค่า $\hat{\pi}_j$ ควรเข้าใกล้ค่า π_j ที่ระดับ 0.50 จากหลักการดังกล่าวจึงทำให้กระบวนการสุ่มข้อสอบเป็นไปตามลำดับขั้นตอนของการคัดเลือกข้อสอบที่มีค่าเข้าใกล้ 0.50 ดังกล่าวกล่าวคือในการสุ่มข้อสอบแต่ละครั้งจะมีการประมาณค่า $\hat{\pi}_j$ ทุกครั้งจะหยุดสุ่มข้อสอบเมื่อค่า $\hat{\pi}_j$ ไม่เข้าใกล้ค่า π_j ที่ระดับ 0.50 วิธีการนี้ถูกนำไปใช้ในการทดสอบวัดความรู้

วิธีการที่ 3 การจัดชุดเพื่อให้ได้แบบทดสอบที่ดีที่สุด (Optimal Test Assembly)

วิธีการนี้ใช้แนวคิดของ Bimbaum โดยมีขั้นตอนคือต้องค้นหาวิธีการรวบรวมข้อสอบจากคลังข้อสอบที่ดีที่สุดเพื่อให้เป็นแบบทดสอบต่างฉบับที่มีค่า TIF ตามเป้าหมายที่ผู้สร้างต้องการ โดยวิธีการรวบรวมข้อสอบนั้นจะประยุกต์ใช้ปัญหาโปรแกรมทางคณิตศาสตร์เพื่อนำไปสู่การตัดสินใจคัดเลือกข้อสอบนั้นๆ ที่มีความคลาดเคลื่อนจากค่าเป้าหมายที่กำหนดให้น้อยที่สุด

วิธีการที่ 4 การจัดชุดแบบทดสอบแบบปรับเหมาะ (Adaptive Test Assembly)

วิธีการนี้ใช้ทฤษฎีการตอบข้อสอบ (IRT) ในการสุ่มข้อสอบที่เหมาะสมกับความสามารถจริงของผู้สอบ (θ) และในการสุ่มข้อสอบแต่ละครั้งจะคำนึงถึงค่าสารสนเทศสูงสุดของผู้สอบแต่ละบุคคล

การประยุกต์ใช้โปรแกรมแก้ปัญหาทางคณิตศาสตร์เพื่อการจัดชุดแบบทดสอบ

(Integer Programming to Test Assembly)

กรณีมีคลังข้อสอบจำนวน 100 ข้อ ข้อสอบแต่ละข้อผ่านการปรับเทียบคะแนนแล้ว ข้อสอบทั้งหมดวัด 2 เนื้อหา คือ เนื้อหาที่ 1 ความรู้เกี่ยวกับข้อเท็จจริง และหลักการ จำนวน 50 ข้อ เนื้อหาที่ 2 การประยุกต์ใช้ความรู้ในชีวิตประจำวัน จำนวน 50 ข้อ และฟังก์ชันสารสนเทศของแบบทดสอบ คือ $\tau(\theta)$ หากต้องการจัดชุดแบบทดสอบ ประกอบด้วยข้อสอบ จำนวน 10 ข้อต่อเนื้อหา จากเงื่อนไขดังกล่าวสามารถใช้โปรแกรมปัญหาทางคณิตศาสตร์ภายใต้หลักการของความคลาดเคลื่อนน้อยที่สุดสามารถแปลงเป็นเงื่อนไขของการสุ่มข้อสอบเพื่อจัดเป็นฉบับ 3 ประการ คือ (1) ค่า TIF ณ ทุกระดับของ θ มีค่าเท่ากับหรือสูงกว่าค่าเป้าหมายที่กำหนด (2) จำนวนข้อสอบวัดความรู้มีจำนวนมากกว่าหรือเท่ากับ 10 ข้อ (3) จำนวนข้อสอบวัดการประยุกต์ใช้มีจำนวนมากกว่าหรือเท่ากับ 10 ข้อ

ถ้าผู้สร้างต้องการใช้วิธีการจัดชุดแบบทดสอบแบบ Optimal Test-Assembly ในการแก้ปัญหาทางคณิตศาสตร์ดังกล่าวจึงต้องใช้หลักการการตัดสินใจเลือกตัวแปรทางคณิตศาสตร์ โดยกฎการคัดเลือกตัวแปรเพื่อจัดชุดแบบทดสอบ มีดังนี้

(1) กฎการตัดสินใจเลือกตัวแปรพหุนาม

		$\sum_{i=1}^{100} x_i$	กฎการตัดสินใจที่ 1
เมื่อ	$i = 1, \dots, 50$	หมายถึง	ข้อสอบวัดความรู้
	$i = 51, \dots, 100$	หมายถึง	ข้อสอบวัดการประยุกต์ใช้
	$x_i = 1$	หมายถึง	ข้อสอบที่ได้รับการคัดเลือกเข้าฉบับ
	$x_i = 0$	หมายถึง	ข้อสอบที่ไม่ได้รับการคัดเลือกเข้าฉบับ

(2) กฎการรวมข้อสอบแบบถ่วงน้ำหนัก

	$\sum_{i=1}^{100} I_i(\theta)x_i$	กฎการตัดสินใจที่ 2
เมื่อ	$I_i(\theta)$	หมายถึง ฟังก์ชันสารสนเทศของข้อสอบข้อที่ i (Item Information Function: IIF)

จากกฎการตัดสินใจข้อที่ 2 หากผู้สร้างรวมข้อสอบทุกข้อที่ได้รับการคัดเลือก เราจะเรียกว่าฟังก์ชันสารสนเทศของแบบทดสอบ (Test Information Function: TIF) และค่า TIF ของ k ระดับความสามารถของผู้สอบ (θ_k) เมื่อ $k = 1, \dots, K$ จึงบ่งบอกถึงค่าสารสนเทศของแบบทดสอบเป้าหมาย ณ ระดับความสามารถของผู้สอบที่ k สามารถเขียนแทนด้วยสัญลักษณ์ (T_k) จึงต้องแปลงเป็นปัญหาทางคณิตศาสตร์ตามหลักการของความคลาดเคลื่อนน้อยที่สุด

(3) กฎของการคัดเลือกจำนวนข้อสอบเข้าฉบับเพื่อให้ได้ค่า TIF ตามเป้าหมาย

$$\text{minimize } \sum_{i=1}^I x_i \quad \text{กฎการตัดสินใจที่ 3}$$

(4) กฎของค่า TIF ขั้นต่ำ

$$\sum_{i=1}^{100} I_i(\theta_k)x_i \geq T_k \quad \text{สำหรับทุกระดับ } \theta \quad \text{กฎการตัดสินใจที่ 4}$$

จากกฎการตัดสินใจที่ 4 แสดงให้เห็นว่า T_k จากแบบทดสอบที่จัดชุดนั้นต้องมีค่าใกล้เคียงกับค่า TIF เป้าหมายให้มากที่สุด

กระบวนการของการกำหนดปัญหาทางคณิตศาสตร์สำหรับการจัดชุดแบบทดสอบ

การจัดชุดแบบทดสอบที่ดำเนินการด้วยการระบุปัญหาทางคณิตศาสตร์ที่เขียนในโปรแกรมคอมพิวเตอร์นั้นมีกระบวนการดำเนินงาน 3 ขั้นตอน คือ

- (1) การระบุตัวแปรที่ใช้ในการตัดสินใจ
- (2) การสร้างเงื่อนไขของการจัดชุดแบบทดสอบ
- (3) การสร้างวัตถุประสงค์ของการจัดชุดแบบทดสอบ ดังรายละเอียด
- (1) การระบุตัวแปรที่ใช้ในการตัดสินใจ (Identifying Decision Variables)

เป็นการระบุการคัดเลือกข้อสอบจากคลัง เช่น การกำหนดคุณลักษณะของข้อสอบรายข้อ ตัวเลือก

และชุดของข้อคำถาม ซึ่งมีการกำหนดค่าที่เหมาะสมและสอดคล้องกับจุดมุ่งหมายของการจัดชุด
แบบทดสอบแทนด้วย $i = 1, \dots, I$

กรณีที่มีได้กำหนดเงื่อนไขใดๆเพิ่มเติม จำนวนแบบทดสอบต่างฉบับที่สามารถจัดชุดจากคลัง
ข้อสอบ คือ 2^I ซึ่งจำนวนแบบทดสอบที่ได้ไม่สอดคล้องในทางปฏิบัติ

การกำหนดปัญหาทางคณิตศาสตร์ในรูปของตัวแปรจึงเป็นทางเลือกในการตัดสินใจเพื่อคัดเลือกตัว
แปรที่ต้องการ สามารถเขียนได้ดังนี้ $y = 1, \dots, 2^I$ โดยการกำหนดค่าของการตัดสินใจ คือ X_i ซึ่งค่านี้สามารถ
แปรค่าได้ 2 ค่า คือ 0 และ 1

$$X_i = \begin{cases} 1 & \text{เมื่อข้อสอบข้อที่ } i \text{ ถูกคัดเลือกเข้าแบบสอบฉบับนั้น} \\ 0 & \text{เมื่อข้อสอบข้อที่ } i \text{ ไม่ได้รับการคัดเลือกเข้าแบบสอบฉบับนั้น} \end{cases}$$

ตัวอย่างการกำหนดตัวแปรที่ใช้ในการตัดสินใจ

ข้อสอบ	1	2	...	i	...	I-1	I
ตัวแปร	X_1	X_2	...	X_i	...	X_{I-1}	X_I
แบบทดสอบ ฉบับ 1	0	0	...	...	...	0	0
แบบทดสอบ ฉบับ 2	1	0	...	...	...	0	0
แบบทดสอบ ฉบับ 3	1	1	...	...	...	0	0
...			...	...	...		
แบบทดสอบ ฉบับ 2^I	1	1	...	...	...	1	1

ตัวอย่างการกำหนดตัวแปรการตัดสินใจเลือกข้อสอบ

- 1) กำหนดความยาวของแบบทดสอบ

$$\sum_{i=1}^I X_i = n \quad \text{เมื่อ } n \text{ หมายถึงจำนวนข้อสอบในแบบทดสอบฉบับนั้น}$$

- 2) กำหนดข้อสอบที่ไม่ต้องการให้อยู่ในแบบทดสอบฉบับนั้น

$$X_{11} + X_{14} \leq 1$$

หมายความว่า ข้อสอบข้อที่ 11 ข้อสอบข้อที่ 14 สามารถบรรจุในแบบทดสอบต่างฉบับได้ แต่
ข้อสอบข้อที่ 11 และข้อสอบข้อที่ 14 จะไม่ถูกจัดชุดไว้ในฉบับเดียวกัน

(2) การสร้างเงื่อนไขของการจัดชุดแบบทดสอบ (Modeling the constraints)

การสร้างเงื่อนไขของการจัดชุดแบบทดสอบ จำแนกเป็น 3 ชนิด คือ (2.1) เงื่อนไขเชิงปริมาณ (quantitative constraints) (2.2) เงื่อนไขแบบกลุ่ม (categorical constraints) (2.3) เงื่อนไขเชิงเหตุผล (logical constraint)

(2.1) เงื่อนไขเชิงปริมาณ (quantitative constraints)

(2.1.1) เงื่อนไขเชิงปริมาณระดับแบบทดสอบ

$$\sum_{i=1}^I q_i x_i \begin{matrix} > \\ < \end{matrix} b_q$$

เมื่อ q_i หมายถึง สัญลักษณ์ที่เกี่ยวข้องกับแบบทดสอบ

กรณีแบบทดสอบกำหนดเวลาในการทดสอบ คือ 60 นาที (กำหนดจากค่าเปอร์เซ็นต์ 90 ของเวลาสอบที่ผู้สอบทำข้อสอบทันเวลา) เงื่อนไข คือ

$$\sum_{i=1}^I t_i x_i \leq 60$$

กรณีแบบทดสอบกำหนดจำนวนบรรทัดอยู่ระหว่างหน้า 145 – 150 เงื่อนไข คือ

$$\sum_{i=1}^I l_i x_i \leq 150$$

$$\sum_{i=1}^I l_i x_i \leq 145$$

กรณีกำหนดฟังก์ชันสารสนเทศของข้อสอบ ณ ระดับความสามารถที่ต้องการ เช่น ที่ระดับความสามารถ θ_c มีค่าฟังก์ชันสารสนเทศของข้อสอบ เท่ากับ 8 เงื่อนไข คือ

$$\sum_{i=1}^I I_i(\theta_c) x_i \geq 8$$

(2.1.2) เงื่อนไขเชิงปริมาณระดับชุดข้อสอบ

กรณีแบบทดสอบมีชุดข้อสอบที่มีการใช้ตัวเลือกร่วมกันโดยกำหนดว่าทั้งฉบับมีการใช้ตัวเลือกร่วมกันตั้งแต่ 2 – 3 ตัวเลือก เงื่อนไข คือ

$$\sum_{i \in V_S} \pi_i x_i \leq 3$$

$$\sum_{i \in V_S} \pi_i x_i \geq 2$$

เมื่อ V_s หมายถึง ชุดตัวเลือกที่ใช้ในชุดข้อสอบ
กรณีแบบทดสอบข้อที่ 23-35 มีการใช้ชุดตัวเลือกร่วมกัน เงื่อนไข คือ

$$\sum_{i=23}^{35} \pi_i x_i$$

(2.1.3) เงื่อนไขเชิงปริมาณระดับข้อสอบ

กรณีแบบทดสอบกำหนดค่าความยากของข้อสอบที่มีค่ามากกว่าหรือเท่ากับ 0.40 เงื่อนไข คือ

$$.40x_i \leq \pi_i \text{ สำหรับข้อสอบทุกข้อ}$$

(2.2) เงื่อนไขแบบกลุ่ม (categorical constraints)

(2.2.1) เงื่อนไขแบบกลุ่มระดับแบบทดสอบ

กรณีแบบทดสอบต้องการกำหนดสัดส่วนของเนื้อหาของแบบทดสอบ

$$\begin{aligned} \sum x_i &\leq n_c^{\max} \\ \sum_{i \in V_c} &\geq n_c^{\min} \end{aligned}$$

จากเงื่อนไขหมายความว่าเนื้อหา C จะมีจำนวนข้อสอบขั้นต่ำ คือ $n_c^{\min}$ และจำนวนข้อสอบสูงสุด คือ $n_c^{\max}$

กรณีแบบทดสอบต้องการกำหนดเงื่อนไขร่วมกันระหว่างเนื้อหากับประเภทของข้อสอบ

$$\begin{aligned} \sum_{i \in V_c \cap V_f} x_i &\leq n_{cf}^{\max} \\ \sum_{i \in V_c \cap V_f} x_i &\leq n_{cf}^{\min} \end{aligned}$$

หมายความว่าแบบทดสอบเนื้อหา C และข้อสอบประเภท f มีจำนวนข้อต่ำสุด คือ $n_{cf}^{\min}$ และจำนวนข้อสูงสุด คือ $n_{cf}^{\max}$

(2.2.2) เงื่อนไขแบบกลุ่มระดับชุดข้อสอบ

$$\begin{aligned} \sum_{i \in V_s \cap V_c} x_i &\leq n_{sc}^{\max} \\ \sum_{i \in V_s \cap V_c} x_i &\leq n_{sc}^{\min} \end{aligned}$$

หมายถึงแบบทดสอบเนื้อหา C และชุดข้อสอบนี้มีการใช้ตัวเลือกร่วมกัน คือ V_s มีจำนวนข้อต่ำสุด คือ $n_{sc}^{\min}$ และจำนวนข้อสูงสุด คือ $n_{sc}^{\max}$

(2.2.3) เงื่อนไขแบบกลุ่มระดับข้อสอบ

$$\sum_{i \in V_1} x_i = n_1$$

$$\sum_{i \in V_0} x_i = 0$$

หมายถึงข้อสอบชุด V_1 มีจำนวนข้อ เท่ากับ n_1

(2.3) เงื่อนไขเชิงเหตุผล (logical constraint)

เงื่อนไขประเภทนี้จะมีลักษณะการเขียนแบบ ถ้า....ดังนั้น....

กรณีข้อสอบข้อนั้นไม่สามารถจัดชุดข้อสอบร่วมกับข้ออื่นได้ สัญลักษณ์ที่ใช้แทนค่า คือ V_e (set of enemies)

$$\sum_{i \in V_e} x_i \leq 1$$

(3) การสร้างวัตถุประสงค์ของการจัดชุดแบบทดสอบ (Modeling the objective)

การสร้างวัตถุประสงค์ของการจัดชุดแบบทดสอบ จำแนกเป็น

(3.1) วัตถุประสงค์เชิงปริมาณ

การสร้างวัตถุประสงค์ของการจัดชุดแบบทดสอบเชิงปริมาณยึดหลักการของการจัดชุด เพื่อให้เกิดแบบทดสอบที่ตรงกับคุณลักษณะที่กำหนดมากที่สุด คือ

$$\text{optimize} \sum_{i=1}^I q_i x_i$$

เมื่อ q_i หมายถึง คุณลักษณะที่กำหนดให้เหมาะสมกับวัตถุประสงค์ของการทดสอบ เช่น

คะแนนจุดตัด เวลาที่ใช้ในการทดสอบ ค่าอำนาจจำแนกของข้อสอบ เป็นต้น

กรณีแบบทดสอบคัดเลือกต้องกำหนดคะแนนจุดตัดด้วยค่าฟังก์ชันสารสนเทศของแบบทดสอบ สามารถเขียนข้อกำหนด ดังสมการ

$$\text{maximum} \sum_{i=1}^I I_i(\theta_C) x_i$$

(3.2) วัตถุประสงค์แบบจัดประเภท

กรณีการจัดชุดแบบทดสอบมีการกำหนดจำนวนเนื้อหาการวัด เช่น จำนวน 3 เนื้อหา ดัง

สมการ $\text{minimize} \sum_{i \in V_1} x_i + \sum_{i \in V_2} x_i + \sum_{i \in V_3} x_i$ หรือ $\text{minimize} \sum_{i \in V_1 \cup V_2 \cup V_3} x_i$

(3.3) วัตถุประสงค์แบบกำหนดค่าเป้าหมาย

กรณีต้องการกำหนดค่าความยากของข้อสอบในแบบทดสอบที่ต้องการจัดชุด เช่น ต้องการให้แบบทดสอบมีค่าความยากเฉลี่ย เท่ากับ 0.50 ดังสมการ

$$\sum_{i=1}^I \pi_i x_i \leq 0.50 = y$$

$$\sum_{i=1}^I \pi_i x_i \geq 0.50 = y$$

กระบวนการของการกำหนดปัญหาทางคณิตศาสตร์สำหรับการจัดชุดแบบทดสอบมากกว่า 1 ฉบับ

ในคราวเดียวกัน (Simultaneous Assembling Multiple Tests)

การจัดชุดแบบทดสอบมากกว่า 1 ฉบับในคราวเดียวกันมีฟังก์ชันในการดำเนินการ ดังนี้

ฟังก์ชันที่ 1 ตัวแปรเพื่อการตัดสินใจเลือกข้อสอบเข้าฉบับ คือ

$$x_{it} = \begin{cases} 1 & \text{เมื่อข้อสอบที่ } i \text{ ถูกจัดอยู่ในแบบสอบ } t \\ 0 & \text{เมื่อข้อสอบที่ } i \text{ ไม่ถูกจัดอยู่ในแบบสอบ } t \end{cases}$$

ฟังก์ชันที่ 2 การกำหนดการใช้ข้อสอบซ้ำระหว่างฉบับ

$$\sum_{t=1}^T x_{it} \leq 1, \text{ for all } i$$

ฟังก์ชันที่ 3 การกำหนดความยาวของแบบทดสอบ

$$\sum_{i=1}^I x_{it} \geq n_t, \text{ for all } t$$

ฟังก์ชันที่ 4 การกำหนดจำนวนข้อสอบต่อเนื้อหาใดๆ

$$\sum_{i \in v_1} x_{it} = n_1 t, \text{ for all } t$$

ฟังก์ชันที่ 5 การกำหนดข้อสอบที่ไม่สามารถจัดชุดร่วมกันในแบบทดสอบฉบับเดียวกันได้

$$\sum_{i \in v_e} x_{it} \leq 1, \text{ for all } e \text{ and } t$$

ฟังก์ชันที่ 6 การกำหนดค่าเป้าหมายของแบบทดสอบ

$$\sum_{i=1}^I I_i(\theta_{kt}) x_{it} \leq T_{kt} + w_{ty}, \text{ for all } k \in v_t \text{ and } t$$

$$\sum_{i=1}^I I_i(\theta_{kt}) x_{it} \geq T_{kt} - w_{ty}, \text{ for all } k \in v_t \text{ and } t$$

การเขียนข้อกำหนดการใช้ข้อสอบซ้ำระหว่างแบบทดสอบ

การเขียนข้อกำหนดการใช้ข้อสอบซ้ำระหว่างแบบทดสอบ จำแนกเป็น 2 ประเภท คือ

1) กำหนดจำนวนฉบับที่อนุญาตให้ใช้ข้อสอบซ้ำ

$$\sum_{t=1}^T x_{it} \geq n_0 \max$$

$$n_0^{\max} > 1$$

2) กำหนดแบบทดสอบบางฉบับที่ให้ใช้ข้อสอบซ้ำ

ฟังก์ชันที่ 1 การกำหนดตัวแปรการใช้ข้อสอบซ้ำระหว่างฉบับ

$$z_{it} = \begin{cases} 1 & \text{เมื่อข้อสอบที่ } i \text{ ถูกจัดแบบสอบ } t \text{ และ } t/ \\ 0 & \text{เมื่อข้อสอบที่ } i \text{ ไม่ถูกจัดแบบสอบ } t \text{ และ } t/ \end{cases}$$

ฟังก์ชันที่ 2 การกำหนดจำนวนข้อซ้ำระหว่างฉบับ

$$\sum_{i=1}^I z_{it} \leq n_{tt'}^{\max}$$

$$\sum_{i=1}^I z_{it} \geq n_{tt'}^{\min}$$

$$2z_{tt'} \leq x_{it} + x_{it'}, \text{ for all } i$$

เมื่อ $z_{itt'} = 1$ ดังนั้น $x_{it} = 1, x_{it'} = 1$

$$z_{itt'} \geq x_{it} + x_{it'} - 1, \text{ for all } i$$

เมื่อ $x_{it} = x_{it'} = 1$ ดังนั้น $z_{itt'} = 1$

1.5 งานวิจัยที่เกี่ยวข้องกับโมเดลเชิงเส้นสำหรับการออกแบบการสร้างแบบทดสอบ

Lin (2008, pp. 1-39) ทำการศึกษา และเปรียบเทียบผลการจัดชุดแบบทดสอบอัตโนมัติโดยใช้โมเดลเชิงเส้นแบบถ่วงน้ำหนักค่าเบี่ยงเบน (weighted deviations model: WDM) เพื่อนำแบบทดสอบไปใช้ในการทดสอบแบบปรับเหมาะ การทดสอบบนคอมพิวเตอร์ วัตถุประสงค์การวิจัยเพื่อตรวจสอบระดับความเป็นคู่ขนานของแบบทดสอบที่ถูกจัดชุดด้วย WDM ระหว่างทฤษฎีการทดสอบแบบดั้งเดิมกับทฤษฎีการตอบข้อสอบ วิธีดำเนินการวิจัยด้วยจำลองข้อมูลการตอบจากผู้สอบ 20,000 คน ตามค่าพารามิเตอร์อำนาจจำแนก ค่าความยาก และค่าการเดาของคลังข้อสอบวิชาคณิตศาสตร์ 10 ฉบับ เนื้อหาการวัด 4 เนื้อหา จำแนกเป็น 152, 127, 147 และ 174 ข้อตามลำดับ ตามทฤษฎีการทดสอบแบบดั้งเดิมและทฤษฎีการตอบข้อสอบ แล้วทำการสุ่มข้อสอบ จำนวน 60 ข้อมาทำการจัดชุดแบบทดสอบคู่ขนาน ตัวแปรที่ศึกษา คือ ความถูกต้องตรงตามเงื่อนไขที่กำหนดของแบบทดสอบคู่ขนาน อัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบ ความเป็นคู่ขนานของเนื้อหาการวัด ค่าฟังก์ชันคุณลักษณะของแบบทดสอบตามตำแหน่งความสามารถของผู้สอบ จำแนกเป็น

2 ตำแหน่งความสามารถ และทุกตำแหน่งความสามารถ ผลการศึกษาในภาพรวมพบว่าแบบทดสอบกลุ่มงาน 6 ฉบับที่ถูกจัดชุดแบบทดสอบด้วยโมเดลเชิงเส้น WDM ระหว่างทฤษฎีการทดสอบแบบดั้งเดิมกับทฤษฎีการตอบข้อสอบไม่มีความแตกต่างกันทั้งด้านความถูกต้องตรงตามเงื่อนไขที่กำหนดของแบบทดสอบกลุ่มงาน ด้านอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบ และด้านความเป็นกลุ่มงานของเนื้อหาการวัด เมื่อพิจารณาถึงค่าฟังก์ชันคุณลักษณะของแบบทดสอบ พบว่าโมเดลเชิงเส้น WDM สามารถจัดชุดแบบทดสอบตามทฤษฎีการตอบข้อสอบที่มีคุณลักษณะของความเป็นกลุ่มงานตามค่าฟังก์ชันคุณลักษณะของแบบทดสอบที่ 2 ตำแหน่งความสามารถได้ถูกต้องมากกว่าการจัดชุดแบบทดสอบที่มีคุณลักษณะของความเป็นกลุ่มงานตามค่าฟังก์ชันคุณลักษณะของแบบทดสอบที่ทุกตำแหน่งความสามารถ สำหรับอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบ พบว่าเมื่อโมเดลเชิงเส้น WDM กำหนดค่าฟังก์ชันสารสนเทศของแบบทดสอบที่ทุกระดับความสามารถจะเกิดอัตราการใช้ข้อสอบซ้ำเกินเกณฑ์ที่กำหนด

Davey and Hendrickson (2010, pp.1-43) ทำการศึกษา และเปรียบเทียบผลการจัดชุดแบบทดสอบอัตโนมัติโดยใช้โมเดลเชิงเส้นแบบถ่วงน้ำหนักค่าเบี่ยงเบน (weighted deviations model: WDM) วัตถุประสงค์การวิจัยเพื่อตรวจสอบคุณลักษณะของแบบทดสอบที่ถูกจัดชุดด้วย WDM ระหว่างทฤษฎีการทดสอบแบบดั้งเดิมกับทฤษฎีการตอบข้อสอบ วิจัยดำเนินการวิจัยด้วยจำลองข้อมูลการตอบจากผู้สอบ 5,000 คน คลังข้อสอบ 240 ข้อ เนื้อหาการวัด 3 เนื้อหา โดยมีจำนวนข้อสอบต่อเนื้อหา คือ 96, 72 และ 72 ข้อ ตามลำดับ ตัวแปรที่ศึกษา คือ ความเป็นกลุ่มงานของเนื้อหาการวัด การกระจายของคุณลักษณะของคะแนนสอบ ค่าฟังก์ชันสารสนเทศของแบบทดสอบ ผลการศึกษาพบว่าแบบทดสอบที่ถูกจัดชุดแบบทดสอบด้วยโมเดลเชิงเส้น WDM ระหว่างทฤษฎีการทดสอบแบบดั้งเดิมกับทฤษฎีการตอบข้อสอบไม่มีความแตกต่างกัน

ตอนที่ 2 แนวคิดและหลักการการตรวจสอบอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบและงานวิจัยที่เกี่ยวข้อง

การตรวจสอบอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบเริ่มเป็นที่สนใจตั้งแต่ ค.ศ. 1976 ถึงปัจจุบัน โดยที่ Chang & Zhang (2002, p.387-398) ได้เสนอสูตรการคำนวณและค่าขีดจำกัดล่างที่แสดงถึงการใช้ข้อสอบซ้ำระหว่างแบบทดสอบ คำนวณจากสูตร $r = (k \div n)$ เมื่อ k หมายถึง ความยาวของแบบทดสอบ n หมายถึง จำนวนข้อสอบในคลังข้อสอบ โดยมีข้อตกลงเบื้องต้น คือ ข้อสอบต้องถูกสุ่มแบบสุ่มสมบูรณ์ (completely randomized item selection) แต่ข้อตกลงเบื้องต้นดังกล่าวไม่สอดคล้องกับแนวทางปฏิบัติ ต่อมา Chen, Ackermann & Spray (2003, p. 129-145) ได้ทำการศึกษาแนวคิดของ Chang & Zhang (2002) ต่อด้วยการเสนอแนวทางการนำไปใช้และให้ข้อคิดเห็นเพิ่มเติมเกี่ยวกับสูตรการควบคุมการใช้ข้อสอบซ้ำระหว่างแบบทดสอบให้ต่ำที่สุด $r = (k \div n)$ พบว่ากรณีวัตถุประสงค์ของการจัดชุดแบบทดสอบจากคลังข้อสอบคำนึงถึงคุณสมบัติด้านจิตมิติ (psychometric properties) ของแบบทดสอบ ค่าคำนวณอัตรา

การใช้ข้อสอบชำระระหว่างแบบทดสอบจาก $(k \div n)$ จะเป็นค่าประมาณที่สูงเกินความเป็นจริง ดังนั้นจึงมีข้อเสนอแนะในการนำสูตรคำนวณนี้ไปใช้ คือ 1) สามารถใช้สูตรคำนวณนี้เพื่อการคัดกรองการใช้ข้อสอบซ้ำขั้นต้นสำหรับการจัดชุดแบบทดสอบกลุ่มงานคราวละ 2 ฉบับ และ 2) อัตราการใช้ข้อสอบชำระระหว่างแบบทดสอบจากสูตรนี้จะแสดงถึงค่าขีดจำกัดบนของใช้ข้อสอบซ้ำเฉลี่ยก็ต่อเมื่อมีวัตถุประสงค์ของการจัดชุดแบบทดสอบตั้งแต่ 2 ฉบับในคราวเดียวกัน และมีการกำหนดเงื่อนไขเพิ่มเติมในการจัดชุดแบบทดสอบด้วย เช่น ความสอดคล้องกันของเนื้อหาในการวัดระหว่างแบบทดสอบมากกว่า 1 ฉบับ ค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง ความสอดคล้องกันระหว่างระดับค่าความสามารถของผู้สอบของแบบทดสอบฉบับอ้างอิงกับแบบทดสอบที่ถูกจัดชุดแบบทดสอบจากคลังข้อสอบ เป็นต้น

2.1 แนวคิดและหลักการการตรวจสอบอัตราการใช้ข้อสอบชำระระหว่างแบบทดสอบ

Chang & Zhang (2002, p.387-398) กล่าวว่า การทดสอบแบบปรับเหมาะ (Computerized Adaptive Testing: CAT) เป็นวิธีการวัดประเมินทางการศึกษา (assessment in education) ที่ได้รับความนิยมใช้กับการทดสอบขนาดใหญ่ (large scale testing) ซึ่งส่วนใหญ่เป็นการทดสอบระดับชาติที่ต้องดำเนินการสอบกับผู้สอบจำนวนมาก เช่น Graduate Record Examinations (GRE), Graduate Management Admission Test (GMAT) การพัฒนาระบบ และการดำเนินการทดสอบแบบปรับเหมาะจำเป็นต้องมีคลังข้อสอบที่ปลอดภัยเพื่อนำมาใช้ในการจัดชุดแบบทดสอบมากกว่า 1 ฉบับที่ตรงกับค่าความสามารถของผู้สอบนั้นๆ โดยขั้นตอนแรกของการทดสอบผู้สอบจะได้รับข้อสอบข้อที่มีความยากง่ายระดับปานกลางเพื่อนำผลการตอบไปใช้ในการคัดเลือกข้อต่อไปที่มีค่าพารามิเตอร์ของข้อสอบที่เหมาะสมกับความสามารถของตน ดังนั้นชุดข้อสอบที่ใช้เป็นข้อสอบร่วมที่นำไปใช้ในการปรับเทียบคะแนนระหว่างแบบทดสอบต่างฉบับ (common items) ที่มีความยากง่ายระดับปานกลางจึงมีโอกาสมากที่จะถูกสุ่มจากคลังข้อสอบบ่อยครั้งกว่าชุดข้อสอบอื่นๆ จึงส่งผลให้ชุดข้อสอบร่วม (common items) ดังกล่าวมีอัตราการใช้ข้อสอบซ้ำ (item exposure rates) ส่วนข้อสอบใดที่มีค่าอำนาจจำแนกสูงจะมีโอกาสเกิดอัตราการใช้ข้อสอบซ้ำ (item overlap rates) สูงกว่าข้อสอบที่มีอำนาจจำแนกต่ำกว่า เนื่องจากข้อสอบที่มีค่าอำนาจจำแนกสูงจะเป็นข้อสอบที่ทำให้การประมาณค่าฟังก์ชันสารสนเทศของข้อสอบ (Item Characteristic Curve: ICC) แม่นยำและสอดคล้องกับระดับความสามารถของผู้สอบ (θ_p) ส่งผลให้ค่าฟังก์ชันสารสนเทศของแบบทดสอบ (Test Information Function: TIF) สูงขึ้นตามมา (Chang and Ying, 1996 as cited in Chang & Zhang, 2002, p.389) จากปรากฏการณ์ดังกล่าวการควบคุมอัตราการใช้ข้อสอบซ้ำ (item exposure control) กลุ่มนักวัดผลจึงหยิบยกมาเป็นประเด็นสำคัญเพื่อใช้ในการประเมิน CAT ที่สามารถบ่งชี้ด้านความปลอดภัยของคลังข้อสอบและแบบทดสอบ Way (1998 as cited in Chang & Zhang, 2002, p.387) กล่าวว่า การควบคุมอัตราการใช้ข้อสอบซ้ำของข้อสอบนั้นควรพิจารณาทั้งอัตราของข้อสอบซ้ำ และการกำหนดอายุการใช้งานของคลังข้อสอบของ CAT นอกจากนี้เขาได้ให้ข้อเสนอแนะว่าข้อสอบที่ถูกสุ่มมาควรถูกใช้ซ้ำมีค่าน้อยกว่า 20 % ต่อมา Chen, Ankenmann and Spray (2003, p.129) ได้ทำการตรวจสอบ

ความสัมพันธ์ระหว่างอัตราของข้อสอบรั่ว (item exposure rates) การใช้ข้อสอบซ้ำระหว่างแบบทดสอบ (test overlap) กรณีการทดสอบแบบปรับเหมาะที่มีความยาวของแบบทดสอบเท่ากัน พบว่า 1) การใช้ข้อสอบซ้ำระหว่างแบบทดสอบจะมีค่าลดลงเมื่อความแปรปรวนของอัตราของข้อสอบรั่วลดลง และ 2) การใช้ข้อสอบซ้ำระหว่างแบบทดสอบจะมีค่าต่ำสุดเมื่ออัตราของข้อสอบรั่วมีค่าเท่ากันในทุกกรณีการทดสอบ แสดงดังสมการ

สูตรคำนวณการใช้ข้อสอบซ้ำระหว่างแบบทดสอบเฉลี่ย (Average Between-Test Overlap)

ใช้สำหรับการทดสอบกับผู้สอบจำนวนมาก

$$\begin{aligned}\bar{T} &= \frac{p \sum_{i=1}^n r_i^2}{k(p-1)} - \frac{1}{p-1} \xrightarrow{p \rightarrow \infty} \frac{\sum_{i=1}^n r_i^2}{k} = \hat{\bar{T}} \\ \hat{\bar{T}} &= \frac{\sum_{i=1}^n \left[\left(r - \frac{k}{n} \right)^2 + 2r \frac{k}{n} - \frac{k^2}{n^2} \right]}{k} \\ \hat{\bar{T}} &= \frac{\sum_{i=1}^n \left(r_i - \frac{k}{n} \right)^2 + \frac{k^2}{n}}{k} = \frac{\frac{\sum_{i=1}^n \left(r_i - \frac{k}{n} \right)^2}{n} + \left(\frac{k}{n} \right)^2}{\frac{k}{n}} \\ &\quad \text{เมื่อแทนค่า } \frac{k}{n} \text{ ด้วย } \bar{r} \\ \hat{\bar{T}} &= \frac{\frac{\sum_{i=1}^n (r_i - \bar{r})^2}{n} + \bar{r}^2}{\bar{r}} = \frac{S_r^2 + \bar{r}^2}{\bar{r}}\end{aligned}$$

เมื่อ $\bar{T}$ หมายถึง การใช้ข้อสอบซ้ำระหว่างแบบทดสอบเฉลี่ย

r_i หมายถึง อัตราของข้อสอบรั่ว เท่ากับ $\frac{m_i}{p}$ โดยที่ m_i หมายถึง จำนวนครั้งของข้อสอบ i ถูกสุ่ม

เพื่อจัดลงในแบบทดสอบ p หมายถึง จำนวนข้อสอบที่คงที่ของการทดสอบแบบปรับเหมาะแต่ละครั้ง

p หมายถึง จำนวนข้อสอบของการทดสอบแบบปรับเหมาะแต่ละครั้ง

k หมายถึง ผลรวมของอัตราของข้อสอบรั่ว

S_r^2 หมายถึง ความแปรปรวนของอัตราของข้อสอบรั่ว

$\bar{r}$ หมายถึง อัตราของข้อสอบรั่วเฉลี่ย

ความคลาดเคลื่อนของการประมาณค่า $\hat{\bar{T}}$ จาก $\bar{T}$ ดังสมการ

$$\epsilon(\hat{\bar{T}}) = \bar{T} - \hat{\bar{T}}$$

ตัวอย่างการคำนวณ การใช้ข้อสอบซ้ำระหว่างแบบทดสอบเฉลี่ย

หากผู้สร้างข้อสอบต้องการสร้างแบบทดสอบแบบปรับเหมาะ จำนวน 4 ฉบับ ฉบับละ 5 ข้อ (p) จากคลังข้อสอบที่มีจำนวนข้อสอบทั้งหมด 10 ข้อ แสดงดังตาราง

ฉบับที่	ข้อสอบข้อที่ i ที่ถูกเลือกลงในแบบทดสอบ				
1	2	5	4	7	8
2	1	3	4	6	2
3	7	9	2	5	8
4	8	7	1	3	2

จากข้อมูลข้างต้นผู้สร้างแบบทดสอบแบบปรับเหมาะต้องทำการคำนวณการใช้ข้อสอบซ้ำระหว่างแบบทดสอบทุกคู่ของจำนวนแบบทดสอบที่ต้องการสร้าง ดังรายละเอียด

ลำดับ	การเปรียบเทียบรายคู่	m_i	p	$r_i = \frac{m_i}{p}$
1	$P_1 P_2$	2	5	0.4
2	$P_1 P_3$	4	5	0.8
3	$P_1 P_4$	3	5	0.6
4	$P_2 P_3$	1	5	0.2
5	$P_2 P_4$	3	5	0.6
6	$P_3 P_4$	3	5	0.6
Σ		16	30	3.2
			$\bar{T} = \frac{16}{30} = 0.53$	$\bar{r} = \frac{3.2}{6} = 0.53$

$$S_r^2 = 0.04268$$

$$\hat{T} = \frac{S_r^2 + \bar{r}^2}{\bar{r}}$$

$$\hat{T} = \frac{0.042 + 0.1024}{0.32} = 0.45$$

Chang & Zhang ชี้ให้เห็นว่าการควบคุมอัตราการใช้ข้อสอบของ CAT ควรพิจารณาร่วมกันทั้ง 2 มิติ คือ 1) การควบคุมระดับข้อสอบ คืออัตราของข้อสอบซ้ำ (item exposure rates) และ 2) การควบคุมระดับแบบทดสอบ คืออัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบ (item overlap rates) และให้ข้อสังเกตสูตรของ Chen และคณะว่าเป็นการพิจารณาระหว่างผู้สอบ 2 คนเท่านั้น ในทางปฏิบัติเป็นการทดสอบแบบปรับเหมาะกับกลุ่มผู้สอบจึงได้เสนอสูตรการคำนวณค่าคาดหวังของข้อสอบที่ถูกใช้ซ้ำของกลุ่มผู้สอบ $\hat{E}[Y_\alpha]$ คือ

$$\hat{E}[Y_\alpha] = \frac{\sum_{i=1}^N \sum_{1 \leq i_j < \dots < i_\alpha \leq N} \#(A_i \cap \{U_{j=1}^\alpha A_{ij}\})}{N * \binom{N-1}{\alpha}}$$

เมื่อ $A_i \cap \{U_{j=1}^\alpha A_{ij}\}$ หมายถึง จำนวนข้อสอบซ้ำระหว่างผู้สอบคนที่ i ถึงคนที่ j
 N หมายถึง จำนวนผู้สอบทั้งหมด

หลักการควบคุมอัตราการรั่วไหลของข้อสอบ

หลักการควบคุมอัตราการรั่วไหลของข้อสอบ (item exposure control) เป็นการควบคุมทั้งระดับข้อสอบ คืออัตราของข้อสอบรั่ว (item exposure rates) และระดับแบบทดสอบ คืออัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบ (item overlap rates) โดยการพิจารณาจากค่าผลต่างค่าสุทธาระหว่างอัตราค่าความคาดหว้งของการใช้ข้อสอบซ้ำระหว่างแบบทดสอบ (Expected Baseline Test Overlap Rate: E(BTOR)) กับอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบเฉลี่ย (Average Test Overlap Rate: ATOR)

Lin (2008) ได้ให้ข้อเสนอแนะในการใช้ค่า E(BTOR) เพื่อการควบคุมอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบ เนื่องจากหลักการคำนวณของค่าดัชนี E(BTOR) ได้คำนึงถึงความเป็นตัวแทนของเนื้อหาการวัดซึ่งเป็นการสะท้อนคุณภาพของแบบทดสอบด้านความตรงตามเนื้อหา (content validity) จุดเด่นของค่า E(BTOR) คือ เป็นการผ่อนคลายข้อตกลงเบื้องต้นบางประการจากสูตรของ Chen, et al. คือ $r = (k \div n)$ ส่วนข้อจำกัด คือ หากการจัดชุดแบบทดสอบมากกว่า 1 ฉบับมีเงื่อนไขที่ซับซ้อน และคลังข้อสอบขนาดเล็กการประมาณค่าของดัชนี E(BTOR) จะมีความแม่นยำลดลง ดังนั้น Lin (2010) จึงได้ทำการศึกษาเพิ่มเติมด้วยการสร้างกำหนดการควบคุมอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบด้วยการพิจารณาทั้ง E(BTOR) และ ATOR เป็นการนำค่า E(BTOR) และ ATOR มาหาผลต่างโดยมีเกณฑ์การแปลความหมาย คือ หากมีความแตกต่างกันระดับสูงเป็นการบ่งบอกถึงความไม่ปลอดภัยของแบบทดสอบ

หากการคำนวณการใช้ข้อสอบร่วมกันเป็นการพิจารณาจากการนับจำนวนข้อสอบที่ใช้ซ้ำกันแล้วทำให้เราได้สูตรการคำนวณ คือ สัดส่วนของจำนวนข้อสอบข้อเดิมที่ถูกใช้ซ้ำในแบบทดสอบ 2 ฉบับ (Y) กับจำนวนข้อสอบทั้งหมดที่ถูกจัดลงชุดแบบทดสอบทั้งหมด (m) โดยค่าที่เป็นไปได้ทั้งหมดของ $Y = (0, 1, \dots, m)$ ถ้า $Y = 0$ หมายความว่าไม่มีการใช้ข้อสอบซ้ำ $Y = m$ หมายความว่ามีการใช้ข้อสอบซ้ำ

แต่ถ้ามีการพิจารณาการสุ่มข้อสอบจากจำนวนข้อสอบทั้งหมดในคลังข้อสอบตามเนื้อหาการวัด (m_j) จากจำนวนข้อสอบทั้งหมดของแต่ละเนื้อหาการวัดในคลังข้อสอบ จำนวน J เนื้อหา (C_j) แล้วนั้นหลักการคำนวณจึงต้องใช้หลักการของตัวแปรสุ่มมาประยุกต์สู่การสุ่มข้อสอบจากทุกเนื้อหาการวัดจากคลังข้อสอบ ดังแสดง

$$\begin{pmatrix} C_j \\ m_j \end{pmatrix}$$

เมื่อ C_j หมายถึง จำนวนข้อสอบทั้งหมดของแต่ละเนื้อหาการวัดในคลังข้อสอบ จำนวน J เนื้อหา
($j=1, 2, \dots, J$)

m_j หมายถึง จำนวนข้อสอบที่ต้องการใช้ในการจัดชุดแบบทดสอบของแต่ละเนื้อหาการวัด
จำนวน J เนื้อหา ($m_j = m_1, m_2, \dots, m_J$)

เมื่อมีการสุ่มข้อสอบจากคลังข้อสอบมากกว่า 1 ครั้งร่วมกับการพิจารณาเป็นรายเนื้อหาการวัด จาก
ค่าโอกาสความน่าจะเป็นของการเกิดตัวแปรสุ่ม x จาก X_j คือ

$$\text{prob}(X_j = x) = \frac{\binom{m_j}{x} \binom{C_j - m_j}{m_j - x}}{\binom{C_j}{m_j}}$$

เมื่อ m_j หมายถึง ข้อสอบที่ถูกเลือกจากแต่ละเนื้อหาของ C_j

X_j หมายถึง จำนวนข้อสอบที่ถูกใช้ร่วมกันระหว่างแบบทดสอบ 2 ฉบับ ของเนื้อหา j
จากสูตรข้างต้นแสดงให้เห็นถึงความน่าจะเป็นของ X_j เท่ากับ x เมื่อ $x = 0, 1, \dots, m_j$

ดังนั้นการประมาณค่าความคาดหวังของจำนวนข้อสอบข้อเดิมที่ถูกสุ่มจากแต่ละเนื้อหาการวัดลงใน
แบบทดสอบ 2 ฉบับ $E(X_j)$ สามารถแสดงดังสูตร

$$E(X_j) = \sum_{x=0}^{m_j} \frac{\binom{m_j}{x} \binom{C_j - m_j}{m_j - x}}{\binom{C_j}{m_j}} x = \frac{m_j^2}{C_j}$$

การประมาณค่าความคาดหวังของจำนวนข้อสอบข้อเดิมที่ถูกสุ่มจากทุกเนื้อหาการวัดลงใน
แบบทดสอบ 2 ฉบับ $E(Y)$ สามารถแสดงดังสูตร

$$E(Y) = \sum_{j=1}^J E(X_j) = \sum_{j=1}^J \frac{m_j^2}{C_j}$$

การประมาณค่าที่เป็นเส้นฐานของค่าความคาดหวังของจำนวนข้อสอบข้อเดิมที่ถูกสุ่มจากทุกเนื้อหา
การวัดลงในแบบทดสอบ 2 ฉบับ $E\left(\frac{Y}{m}\right)$ (Baseline test overlap rate) สามารถแสดงดังสูตร

$$E\frac{Y}{m} = \frac{1}{m} \times E(Y) = \frac{1}{m} \sum_{j=1}^J \frac{m_j^2}{C_j}$$

เมื่อ j หมายถึง จำนวนเนื้อหาการวัดทั้งหมดในคลังข้อสอบ จำนวน J เนื้อหา ($j = 1, 2, \dots, J$)

n_j หมายถึง จำนวนข้อสอบในคลังข้อสอบที่จำแนกตามเนื้อหาการวัด J เนื้อหา
($n_1, n_2, \dots, n_J$)

m_j หมายถึง จำนวนข้อสอบที่ถูกสุ่มเพื่อนำมาจัดชุดแบบทดสอบที่ครอบคลุมเนื้อหาการวัด
 J เนื้อหา

การควบคุมอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบแบบสุ่มข้อสอบจากคลังข้อสอบเป็นลำดับขั้น

(Controlling Test Overlap-Ordered Item Pooling)

เป็นการควบคุมอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบ ประกอบด้วย 3 ขั้นตอนหลัก คือ

ขั้นตอนที่ 1 การควบคุมอัตราการใช้ข้อสอบซ้ำด้วยการคำนวณ ATOR ประกอบด้วย 2 ขั้นตอนย่อย

1.1 การคำนวณอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบ 2 ฉบับเฉลี่ย ($\bar{\Omega}_{\alpha,t}$)

1.2 การควบคุม $\bar{\Omega}_{\alpha,t}$ ด้วยค่าเฉลี่ยสูงสุดของอัตราการใช้ข้อสอบซ้ำ ($\bar{\Omega}_{\max}$)

ขั้นตอนที่ 2 การควบคุมอัตราการใช้ข้อสอบซ้ำด้วยการพิจารณาความแตกต่างระหว่าง ATOR กับ

E(BTOR)

ขั้นตอนที่ 3 การกำหนดเกณฑ์การแปลความหมายอัตราการใช้ข้อสอบร่วมกันระหว่างแบบทดสอบ

คลังรายละเอียด

ขั้นตอนที่ 1 การควบคุมอัตราการใช้ข้อสอบซ้ำด้วยการคำนวณ ATOR

1.1 การคำนวณอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบ 2 ฉบับเฉลี่ย ($\bar{\Omega}_{\alpha,t}$)

$$\text{สมการที่ 1} \quad \bar{\Omega}_{\alpha,t} = \left(\frac{t-\alpha-1}{t} \right) (\bar{\Omega}_{\alpha,(t-1)}) + \frac{\alpha+1}{t} - \frac{\sum_{i=1}^n [m_{it} - m_{i(t-1)}] \binom{t-m_{it}}{\alpha}}{m_{(\alpha+1)}^t}$$

เมื่อ $\bar{\Omega}_{\alpha,t}$ หมายถึง อัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบเฉลี่ย กรณีการสุ่มข้อสอบจากคลังข้อสอบเป็นลำดับขั้นของแบบทดสอบฉบับ t

t หมายถึง ลำดับที่ของแบบทดสอบที่ถูกจัดชุด

α หมายถึง กลุ่มของแบบทดสอบที่ถูกจัดชุดแต่ละครั้ง เมื่อนำมาคำนวณกับฉบับก่อน ($t-1$)

เมื่อ $\alpha = 1$, $\bar{\Omega}_{\alpha,t}$ หมายถึง อัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบ 2 ฉบับ

$\bar{\Omega}_{\alpha,(t-1)}$ หมายถึง อัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบเฉลี่ย กรณีการสุ่มข้อสอบจากคลังข้อสอบเป็นลำดับขั้นของแบบทดสอบฉบับก่อน ($t-1$)

m_{it} หมายถึง จำนวนข้อสอบข้อที่ i ที่ถูกเลือกลงในแบบทดสอบทั้งหมด t ฉบับ

$m_{i(t-1)}$ หมายถึง จำนวนข้อสอบข้อที่ i ที่ถูกเลือกลงในแบบทดสอบฉบับก่อน $t (t-1)$

m หมายถึง จำนวนของแบบทดสอบทั้งหมดที่ต้องการจัดชุดแบบทดสอบในสถานการณ์การทดสอบนั้นๆ

n หมายถึง จำนวนข้อสอบทั้งหมดในคลังข้อสอบ

$$\lambda = \sum_{i=1}^n [m_{it} - m_{i(t-1)}] \binom{t-m_{it}}{\alpha}$$

เทอมของ $m_{it} - m_{i(t-1)} = 1$ เมื่อข้อสอบข้อที่ i ถูกเลือกลงในแบบทดสอบฉบับ t

เทอมของ $m_{it} - m_{i(t-1)} = 0$ เมื่อข้อสอบข้อที่ i ไม่ถูกเลือกลงในแบบทดสอบฉบับ t

ตัวอย่างการคำนวณ อัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบเฉลี่ย ($\bar{\Omega}_{\alpha,i}$)

ผู้สร้างข้อสอบต้องการจัดชุดแบบทดสอบทั้งหมด จำนวน 5 ฉบับ ฉบับละ 5 ข้อ จากคลังข้อสอบที่มีจำนวนข้อสอบทั้งหมด 10 ข้อ แสดงดังตาราง

ฉบับที่	ข้อสอบข้อที่ i ที่ถูกเลือกลงในแบบทดสอบ				
1	3	4	7	6	9
2	9	5	7	8	10
3	3	6	9	2	4
4	9	8	10	4	3
5	7	4	5	10	9

จากตารางแสดงให้เห็นว่าข้อสอบข้อที่ 1 ไม่ถูกเลือกลงในแบบสอบฉบับใดๆเลย ข้อสอบข้อที่ 2 ถูกเลือกลงในแบบทดสอบฉบับที่ 3 จำนวน 1 ครั้ง ส่วนข้อสอบข้อที่ 9 ถูกเลือกลงในแบบทดสอบทุกฉบับ นั่นคือข้อสอบข้อที่ 9 ถูกใช้ซ้ำทั้งหมด จำนวน 5 ครั้ง สามารถแสดงในตารางถัดไป

ฉบับที่ (m)	จำนวนการใช้ข้อสอบซ้ำในทุกฉบับ ($m_{i(t-1)}, m_{it}$)									
	1	2	3	4	5	6	7	8	9	10
1	0	0	1	1	0	1	1	0	1	0
2	0	0	1	1	1	1	2	1	2	1
3	0	1	2	1	1	2	2	1	3	1
4	0	1	3	1	1	2	2	2	4	2
5	0	1	3	2	2	2	2	2	5	3

จากตัวอย่างข้างต้นต้องทำการคำนวณอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบ 2 ฉบับเฉลี่ย ($\bar{\Omega}_{\alpha,i}$) จำนวน 4 ค่า คือ $\bar{\Omega}_{1,2}$, $\bar{\Omega}_{1,3}$, $\bar{\Omega}_{1,4}$ และ $\bar{\Omega}_{1,5}$

$$\text{จากสมการ } \bar{\Omega}_{\alpha,i} = \left(\frac{t-\alpha-1}{t} \right) (\bar{\Omega}_{\alpha,(i-1)}) + \frac{\alpha+1}{t} - \frac{\sum_{i=1}^n [m_{it}-m_{i(t-1)}] \binom{t-m_{it}}{\alpha}}{m(\alpha+1)}$$

$$\bar{\Omega}_{1,2} = \left(\frac{2-1-1}{2} \right) (\bar{\Omega}_{\alpha,(i-1)}) + \frac{1+1}{2} - \frac{\binom{2-2}{1} + \binom{2-1}{1} + \binom{2-2}{1} + \binom{2-1}{1} + \binom{2-1}{1}}{5 \binom{2}{1+1}} = \left(\frac{2}{5} \right)$$

$$\bar{\Omega}_{1,3} = \left(\frac{3-1-1}{3} \right) \left(\frac{2}{5} \right) + \frac{1+1}{3} - \frac{\binom{3-1}{1} + \binom{3-2}{1} + \binom{3-2}{1} + \binom{3-2}{1} + \binom{3-3}{1}}{5 \binom{3}{1+1}} = \left(\frac{7}{15} \right)$$

$$\bar{\Omega}_{1,4} = \left(\frac{4-1-1}{4} \right) \left(\frac{7}{15} \right) + \frac{1+1}{4} - \frac{\binom{4-3}{1} + \binom{4-3}{1} + \binom{4-2}{1} + \binom{4-4}{1} + \binom{4-2}{1}}{5 \binom{4}{1+1}} = \left(\frac{8}{15} \right)$$

$$\bar{\Omega}_{1,5} = \left(\frac{5-1-1}{5}\right) \left(\frac{8}{15}\right) + \frac{1+1}{5} - \frac{\binom{5-4}{1} + \binom{5-2}{1} + \binom{5-3}{1} + \binom{5-5}{1} + \binom{5-3}{1}}{5\binom{5}{1+1}} = \left(\frac{14}{25}\right)$$

1.2 การควบคุม $\bar{\Omega}_{\alpha,t}$ ด้วยค่าเฉลี่ยสูงสุดของอัตราการใช้ข้อสอบซ้ำ ($\bar{\Omega}_{\max}$)

$$\text{สมการที่ 2} \quad \bar{\Omega}_{\alpha,t} = \left(\frac{t-\alpha-1}{t}\right) (\bar{\Omega}_{\alpha,(t-1)}) + \frac{\alpha+1}{t} - \frac{\sum_{i=1}^n [m_{it} - m_{i(t-1)}] \binom{t-m_{it}}{\alpha}}{m_{(\alpha+1)}} \leq \bar{\Omega}_{\max}$$

$$\bar{\Omega}_{\max} = \frac{k}{n}$$

เมื่อ $\bar{\Omega}_{\max}$ หมายถึง ค่าสูงสุดของอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบเฉลี่ย

k หมายถึง จำนวนข้อสอบทั้งหมดในแบบทดสอบที่ต้องการจัดชุด

n หมายถึง จำนวนข้อสอบทั้งหมดในคลังข้อสอบ

ขั้นตอนที่ 2 การควบคุมอัตราการใช้ข้อสอบซ้ำด้วยการพิจารณาความแตกต่างระหว่าง ATOR กับ

E(BTOR)

สมการที่ 3

$$\lambda = \sum_{i=1}^n [m_{it} - m_{i(t-1)}] \binom{t-m_{it}}{\alpha} \geq m_{(\alpha+1)} \left[\left(\frac{t-\alpha-1}{t}\right) (\bar{\Omega}_{\alpha,(t-1)}) + \frac{\alpha+1}{t} - \bar{\Omega}_{\max} \right] = D$$

เมื่อ λ หมายถึง ค่าที่ใช้การควบคุมอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบเฉลี่ย ($\bar{\Omega}_{\max}$)

D หมายถึง การควบคุมอัตราการใช้ข้อสอบซ้ำด้วยการพิจารณาความแตกต่างระหว่าง ATOR กับ E(BTOR)

จากสมการที่ 3 ค่า λ เป็นค่าบ่งชี้ความสามารถในการควบคุมการใช้ข้อสอบซ้ำระหว่างแบบทดสอบทุกฉบับให้มีค่าไม่เกินค่าเฉลี่ยสูงสุดของอัตราการใช้ข้อสอบซ้ำ ($\bar{\Omega}_{\max}$) โดยพิจารณาจากค่า λ ต้องมีค่าไม่น้อยกว่าค่า D

แต่หากต้องการควบคุมการใช้ข้อสอบซ้ำเป็นรายข้อนั้นต้องพิจารณาจากค่า λ ต้องมีค่าไม่น้อยกว่า $\frac{D}{m}$

จากตัวอย่างการคำนวณข้างต้นหากกำหนดค่า $\bar{\Omega}_{\max} = 0.2$ ดังนั้นเมื่อต้องการจัดชุดแบบทดสอบทั้งหมด 5 ฉบับ (t) จากคลังข้อสอบ โดยมีการควบคุมอัตราการใช้ข้อสอบซ้ำเป็นลำดับขั้น นั่นคือ อัตราการใช้ข้อสอบซ้ำของการจัดชุดแบบทดสอบ 4 ฉบับ ($\bar{\Omega}_{\alpha,(t-1)}$) เท่ากับ $\frac{8}{15}$ และ α เท่ากับ 1 การคำนวณค่า D ดังสมการ

$$\lambda = \sum_{i=1}^n [m_{it} - m_{i(t-1)}] \binom{t-m_{it}}{\alpha} \geq \left[\left(\frac{5-1-1}{5}\right) \left(\frac{8}{15}\right) + \frac{1+1}{5} \right] - 0.2 = 26 = D$$

เมื่อได้ค่า D จึงนำค่า D เมื่อนำมาคำนวณเป็นรายข้อเพื่อนำมาใช้ในการควบคุมอัตราการใช้ข้อสอบซ้ำด้วยการนำมารวมกับจำนวนแบบทดสอบทั้งหมดที่ต้องการจัดชุด (m) = $\frac{D}{m}$

หลังจากนั้นนำค่า $\frac{D}{m}$ มากำหนดเกณฑ์การแปลความหมายอัตราการใช้ข้อสอบชำระระหว่างแบบทดสอบ

ขั้นตอนที่ 3 การกำหนดเกณฑ์การแปลความหมายอัตราการใช้ข้อสอบร่วมกันระหว่างแบบทดสอบ

จากตัวอย่างการคำนวณมีการจัดชุดแบบทดสอบทั้งหมด (m) 5 ฉบับ โดยกำหนดอัตราค่าเฉลี่ยสูงสุดของอัตราการใช้ข้อสอบชำระ ($\bar{\Omega}_{\max}$) เท่ากับ 0.2 และมีค่า $\frac{D}{m} = \frac{26}{5}$ สามารถสร้างเกณฑ์การแปลความหมายอัตราการใช้ข้อสอบชำระร่วมกันระหว่างแบบทดสอบ ดังนี้

1. เปรียบเทียบค่า λ กับ ค่า $\frac{D}{m}$ ว่า ค่า λ มีค่ามากกว่าหรือเท่ากับ ค่า $\frac{D}{m}$ หรือไม่
2. สรุปผลการควบคุมอัตราการใช้ข้อสอบชำระระหว่างแบบทดสอบ
 - 2.1 การจัดชุดแบบทดสอบครั้งนี้มีอัตราการใช้ข้อสอบชำระไม่เกินอัตราค่าเฉลี่ยสูงสุดของอัตราการใช้ข้อสอบชำระ ($\bar{\Omega}_{\max}$) คือมีค่า $\lambda \geq \frac{D}{m}$
 - 2.2 การจัดชุดแบบทดสอบครั้งนี้มีอัตราการใช้ข้อสอบชำระเกินอัตราค่าเฉลี่ยสูงสุดของอัตราการใช้ข้อสอบชำระ ($\bar{\Omega}_{\max}$) คือมีค่า $\lambda < \frac{D}{m}$

2.2 งานวิจัยที่เกี่ยวข้องกับอัตราการใช้ข้อสอบชำระระหว่างแบบทดสอบ

Chen and Lei (2010, pp. 205-226) ทำการศึกษา อธิบายและแสดงความสัมพันธ์ระหว่างอัตราข้อสอบรั่วไหลกับอัตราการใช้ข้อสอบชำระระหว่างแบบทดสอบด้วยการจำลองข้อมูลร่วมกับการวิเคราะห์ผลการสอบของวิชาคณิตศาสตร์ จำนวน 400 ข้อตามทฤษฎีการตอบข้อสอบแบบ 3 พารามิเตอร์ โดยมีข้อกำหนดในการจำลองข้อมูล คือ 1) ค่าอำนาจจำแนกของข้อสอบมีการแจกแจงแบบปกติ มีค่าเฉลี่ยเท่ากับ 0 ค่าเบี่ยงเบนมาตรฐาน เท่ากับ 1 2) ค่าความสามารถของผู้สอบมีการแจกแจงแบบปกติ มีค่าเฉลี่ยเท่ากับ 0 ค่าเบี่ยงเบนมาตรฐาน เท่ากับ 1 3) จำนวนผู้สอบ 10,000 คน 4) การควบคุมการรั่วไหลของข้อสอบระดับสูงสุด คือ ค่าสูงสุดของอัตราข้อสอบรั่วไหล เท่ากับ 0.20 และ ค่าสูงสุดอัตราการใช้ข้อสอบชำระระหว่างแบบทดสอบเฉลี่ย เท่ากับ 0.10 5) การควบคุมการรั่วไหลของข้อสอบระดับต่ำสุด คือ ค่าสูงสุดของอัตราข้อสอบรั่วไหล เท่ากับ 0.30 และ ค่าสูงสุดอัตราการใช้ข้อสอบชำระระหว่างแบบทดสอบเฉลี่ย เท่ากับ 0.20 ส่วนผลการสอบของคลังข้อสอบวิชาคณิตศาสตร์ จำนวน 400 ข้อ ประกอบด้วยเนื้อหาการวัด 6 เนื้อหา คือ ความรู้เบื้องต้นทางพีชคณิต 84 ข้อ พีชคณิตพื้นฐาน 60 ข้อ พีชคณิตระดับกลาง 54 ข้อ เรขาคณิตวิเคราะห์ 54 ข้อ รูปทรงเรขาคณิต 84 ข้อ และตรีโกณมิติ 64 ข้อ ตัวแปรที่ใช้ในการศึกษา 4 ตัวแปร คือ ตัวแปรที่ 1 ความยาวของแบบทดสอบที่เหมาะสม 2 ระดับ จำแนกเป็น 20 และ 40 ข้อ ตัวแปรที่ 2 วิธีการควบคุมอัตราการใช้ข้อสอบชำระ จำแนกเป็น 1) ไม่มีการควบคุมการรั่วไหลของข้อสอบ 2) การควบคุมการรั่วของข้อสอบระดับน้อยกว่า 0.30 3) การควบคุมการรั่วของข้อสอบระดับ 0.30 และอัตราการใช้ข้อสอบชำระเฉลี่ยระหว่างแบบทดสอบ 0.20 4) การควบคุมการรั่วของข้อสอบระดับน้อยกว่า 0.20 5) การควบคุมการรั่วของข้อสอบ

ระดับ 0.20 และอัตราการใช้ข้อสอบซ้ำเฉลี่ยระหว่างแบบทดสอบ 0.10 และ 6) การควบคุมการรั่วไหลของข้อสอบแบบสุ่มตามแนวคิดของ Chen และ Zhang ตัวแปรที่ 3 วิธีการวิเคราะห์การควบคุมการรั่วไหลจำแนกเป็น แบบไม่เป็นลำดับขั้น และแบบเป็นลำดับขั้น ตัวแปรที่ 4 คือ จำนวนครั้งของการทดสอบแบบปรับเหมาะ คือ 5 ครั้ง ผลการศึกษาพบว่า (1) เมื่อเพิ่มจำนวนครั้งของการทดสอบแบบปรับเหมาะส่งผลให้อัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบสูงขึ้นทั้งกรณีที่ใช้วิธีการวิเคราะห์การควบคุมการรั่วไหลแบบไม่เป็นลำดับขั้น และแบบเป็นลำดับขั้น (2) วิธีการควบคุมอัตราการใช้ข้อสอบแบบการควบคุมการรั่วไหลของข้อสอบระดับ 0.20 และอัตราการใช้ข้อสอบซ้ำเฉลี่ยระหว่างแบบทดสอบ 0.10 สามารถควบคุมอัตราการใช้ข้อสอบรั่วไหลได้ดีกว่าวิธีอื่น ทั้งกรณีที่ใช้วิธีการวิเคราะห์การควบคุมการรั่วไหลแบบไม่เป็นลำดับขั้น และแบบเป็นลำดับขั้น (3) เมื่อพิจารณาวิธีการวิเคราะห์การควบคุมการรั่วไหลระหว่างวิธีแบบไม่เป็นลำดับขั้น และแบบเป็นลำดับขั้นที่ดำเนินการควบคุมอัตราการใช้ข้อสอบแบบการควบคุมการรั่วไหลของข้อสอบระดับ 0.20 และอัตราการใช้ข้อสอบซ้ำเฉลี่ยระหว่างแบบทดสอบ 0.10 พบว่าวิธีการวิเคราะห์การควบคุมการรั่วไหลแบบเป็นลำดับขั้นสามารถควบคุมอัตราการใช้ข้อสอบรั่วไหลได้ดีกว่าวิธีการวิเคราะห์การควบคุมการรั่วไหลแบบไม่เป็นลำดับขั้น ดังนั้นผู้วิจัยจึงเสนอแนวทางการประยุกต์ใช้ผลการวิจัยว่ากรณีการทดสอบแบบปรับเหมาะดำเนินการกับผู้สอบขนาดใหญ่ควรใช้วิธีการวิเคราะห์การควบคุมการรั่วไหลแบบเป็นลำดับขั้นจะสามารถควบคุมอัตราการใช้ข้อสอบรั่วไหลได้ดีกว่าแบบไม่เป็นลำดับขั้น

Lin (2010, pp. 1-43) ทำการศึกษาการจัดชุดแบบทดสอบจากคลังข้อสอบที่มีการควบคุมอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบคู่ขนานมากกว่า 1 ฉบับ วัตถุประสงค์การวิจัยเพื่อตรวจสอบประสิทธิภาพของการจัดชุดแบบทดสอบด้วยโมเดลการถ่วงน้ำหนักความเบี่ยงเบน (Weighted Deviation Model: WDM) ที่มีการควบคุมอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบที่ใช้ในการจัดชุดแบบทดสอบคู่ขนาน 2 ฉบับ ใดๆ โดยวิธีการควบคุมการใช้ข้อสอบซ้ำ เรียกว่า test overlap rate-ordered item pooling ซึ่งเป็นการคำนวณอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบด้วยค่าสัดส่วนของข้อสอบที่มีการใช้ร่วมกันภายในคลังข้อสอบ โดยพิจารณาจากทุกสถานการณ์การทดสอบว่ามีจำนวนของข้อสอบใดที่ผู้สอบแต่ละคนได้ทำข้อนั้นซ้ำในปริมาณเท่าใด หลักการคำนวณเป็นการคิดจากการใช้ข้อสอบซ้ำระหว่างแบบทดสอบคู่ขนาน 2 ฉบับ ของทุกๆ คู่ของ 2 แบบทดสอบใดๆ ที่ได้รับการสร้างแบบทดสอบคู่ขนานจากคลังข้อสอบและนำมาหาค่าเฉลี่ยของทุกคู่จากสถานการณ์การทดสอบทั้งหมด สำหรับตัวแปรอิสระ คือ อัตราส่วนของจำนวนข้อของแบบทดสอบต่อจำนวนคลังข้อสอบ จำแนกเป็น 0.10 ($60 \div 600$), 0.15 ($90 \div 600$), 0.20 ($120 \div 600$) และเนื้อหาการวัดของแบบทดสอบ จำแนกเป็น 1) การจัดเนื้อหาสอดคล้องกับคลังข้อสอบ (balanced content specification) และ 2) การจัดเนื้อหาไม่สอดคล้องกับคลังข้อสอบ (unbalanced content specification) ตัวแปรตาม คือ ประสิทธิภาพของการจัดชุดแบบทดสอบด้วยโมเดลการถ่วงน้ำหนักความเบี่ยงเบน คือ 1) แบบทดสอบมีข้อกำหนดที่ตรงกับเกณฑ์ที่กำหนด 2) แบบทดสอบคู่ขนานมากกว่า 1 ฉบับมีคุณสมบัติด้านจิตมิติ (psychometric properties) 3) อัตราการใช้ข้อสอบซ้ำเฉลี่ย และ 4) การแจกแจงของอัตราของ

ข้อสอบรั่ว (item exposure rate) วิธีดำเนินการวิจัยใช้คลังข้อสอบรายวิชาคณิตศาสตร์ที่วิเคราะห์ค่าพารามิเตอร์ของข้อสอบตามโมเดลการตอบสนองข้อสอบแบบ 3 พารามิเตอร์ด้วยโปรแกรม BILOG-MG จำนวน 600 ข้อ ประกอบด้วย 4 เนื้อหา เนื้อหาที่ 1 – 4 คือ 152, 127, 147 และ 174 ข้อตามลำดับ ค่าอำนาจจำแนกเฉลี่ย ค่าความยากเฉลี่ย และค่าการเดาเฉลี่ย คือ 0.965, 0.127 และ 0.178 ตามลำดับ ค่าฟังก์ชันสารสนเทศของคลังข้อสอบสูงสุด ณ ระดับความสามารถของผู้สอบ $\theta = 1$ ซึ่งถือว่าเป็นคลังข้อสอบที่เหมาะสมกับผู้สอบที่มีความสามารถระดับปานกลางถึงระดับค่อนข้างสูงผลการศึกษา พบว่าจัดชุดแบบทดสอบด้วยโมเดลการถ่วงน้ำหนักความเบี่ยงเบน (Weighted Deviation Model: WDM) จากคลังข้อสอบสามารถควบคุมอัตราการใช้ข้อสอบซ้ำเฉลี่ยระหว่างแบบทดสอบคู่ขนานมากกว่า 1 ฉบับในทุกเงื่อนไขของการศึกษา คือ อัตราส่วนของจำนวนข้อของแบบทดสอบต่อจำนวนคลังข้อสอบ และ การจัดเนื้อหาการวัดของแบบทดสอบ อีกทั้งผลการวิจัยครั้งนี้สามารถยืนยันผลการใช้อัตราการใช้ข้อสอบซ้ำเฉลี่ยระหว่างแบบทดสอบเพื่อนำไปใช้ในการจัดชุดแบบทดสอบให้เกิดคุณสมบัติของความเสมอภาคระหว่างผู้สอบ

ตอนที่ 3 การใช้โปรแกรม R

นราทิพย์ จันสกุล (2552, หน้า 1-6) กล่าวว่า R เป็นโปรแกรมสำเร็จรูปที่เปิดโอกาสให้บุคคลทั่วไปดาวน์โหลด ติดตั้งและใช้ได้ฟรี โดยไม่ถือว่าเป็นการละเมิดลิขสิทธิ์ R มีความสามารถในการคำนวณและการวิเคราะห์ข้อมูลสถิติได้เทียบเท่ากับ โปรแกรมสำเร็จรูปทางสถิติอื่นๆ และยังสามารถจัดการกับข้อมูลในลักษณะต่างๆ รวมทั้งการนำเสนอข้อมูลโดยเฉพาะการวาดกราฟ หรือแผนภาพได้อย่างมีประสิทธิภาพ ภาษาที่ใช้สำหรับ R เป็นเครื่องมือหรือเป็นส่วนหนึ่งของภาษา S นอกจากนี้ R ยังมีรากฐานมาจากการพัฒนาระบบของโปรแกรมสำเร็จรูป S-PLUS ซึ่งเป็นโปรแกรมเพื่อการคำนวณทางสถิติที่มีศักยภาพมากแต่มีลิขสิทธิ์และราคาแพง R ได้รับการพัฒนาอย่างต่อเนื่องโดยมูลนิธิชื่อ R Development Core Team และเผยแพร่ผ่าน <http://www.r-project.org> หรือจากคู่สำเนา (mirror) ในประเทศต่างๆ ปัจจุบันในประเทศไทยมีคู่สำเนาที่ได้มีการเสนอให้ดาวน์โหลด R คือ <http://mirror.kapook.com/cran/> R ถูกพัฒนาอย่างต่อเนื่อง และมีรุ่นใหม่ออกมาเกือบทุกๆ 6 เดือน เนื่องจากมีนักสถิติเป็นจำนวนมากได้เขียนฟังก์ชันต่างๆ สำหรับการวิเคราะห์ข้อมูลสถิติ แล้วส่งมาใส่ไว้ใน R ตลอดเวลา แต่การพัฒนาดังกล่าวไม่มีผลกระทบต่อฟังก์ชันที่มีอยู่ใน R รุ่นเก่า นอกจากนี้ วิโรจน์ อรุณมานะกุล (2559, หน้า 7) กล่าวว่าสำหรับการใช้งาน R ผ่าน R console เป็นการส่งงานผ่านทาง command line ซึ่งหลายคนอาจไม่คุ้นเคย หากต้องการใช้งาน R ในลักษณะที่มีหน้าจอต่างๆ ปรากฏในคราวเดียวกันสำหรับการแก้ไขชุดคำสั่งต่างๆ หน้าจอแสดงผล หน้าจอแสดงตัวแปรต่างๆ สามารถติดตั้ง R Studio เพิ่มเติม

3.1 ประเภทข้อมูลใน R

วิโรจน์ อรุณมานะกุล (2559, หน้า 7-8) กล่าวว่าข้อมูลพื้นฐานที่ใช้ใน R คือ vector ซึ่งเป็นข้อมูลมิติเดียว มีข้อมูลตัวเดียวหรือหลายตัวก็ได้ แต่จะเป็น object ประเภทเดียวกัน object แบบ atomic ที่มีใช้ใน R ได้แก่ character, numeric, integer, complex number, และ logical (true/false) vector เป็นการมองข้อมูลแบบมีลำดับ เช่น `c(1,3,4)` จะต่างจาก `c(3,4,1)` เราสามารถป้อนข้อมูลเข้าโดยตรงเป็น vector ใน R ได้โดยพิมพ์ `c` และตามด้วยวงเล็บบอกลำดับข้อมูล (`c` ย่อมาจาก concatenate เป็นฟังก์ชันใน Rให้นำข้อมูลมาต่อกันเป็น vector) ข้อมูลที่อยู่ใน vector ต้องเป็นข้อมูลแบบเดียวกัน เช่น เป็นตัวเลขทั้งหมดหรืออักษรทั้งหมด เช่น `>x=c(10,5,23,18)` เป็นการสร้างเวกเตอร์ที่มีค่า 4 ค่าตามลำดับคือ 10,5,23,18 ไปเก็บไว้ที่ตัวแปร `x` เราสามารถอ้างถึงค่าแต่ละตัวใน vector โดยอ้างถึงลำดับที่ เช่น `x[2]` จะมีค่าเป็น 5 หากต้องการป้อนค่าเข้าไปทางคีย์บอร์ดให้ใช้คำสั่ง `scan()` เมื่อใช้คำสั่ง `x=scan()` แล้วพิมพ์ตัวเลขเข้าไป 20 ตัวโดยเว้นวรรคให้ตัวเลขแต่ละตัว เมื่อบรรทัดต่อไปเคาะ enter ก็จะหยุดรับข้อมูล แล้วรายงานว่ามีข้อมูลอ่านเข้าใน vector ที่ตั้งชื่อว่า `x` 20 ตัว

```
> x=scan( )
```

```
1: 1 3 2 3 3 3 4 2 1 2 2 1 1 4 3 3 2 1 2 3
```

```
21:
```

```
Read 20 items
```

หากข้อมูลที่นำเข้าเป็นข้อมูลต่างประเภทกัน เช่น เป็นตัวเลขและข้อความ สามารถนำเข้าได้โดยใช้คำสั่ง `list` และอ้างถึงข้อมูลเฉพาะตำแหน่งที่ต้องการได้

```
> l <- list('male',34, 'single', 67.8)
```

```
> l[3]
```

```
[[1]]
```

```
[1] "single"
```

ฟังก์ชัน `c()` เพื่อสร้าง vector หรือ list

รูปแบบการใช้งาน : ชื่อตัวแปร <- `c(1,2)`

ฟังก์ชัน `scan()` เป็นการรับค่าข้อมูล

รูปแบบการใช้งาน : ชื่อตัวแปร <- `scan()` จะปรากฏส่วนที่ให้ทำการกำหนดค่า

ฟังก์ชัน `edit()` เป็นการแก้ไขข้อมูล

รูปแบบการใช้งาน : ชื่อตัวแปร<-`edit(ชื่อตัวแปร)` จะปรากฏหน้าต่างข้อมูลของตัวแปรนั้น
ซึ่งสามารถทำการแก้ไขค่าได้

ฟังก์ชัน `plot()` เป็นการ plot กราฟแบบจุด

รูปแบบการใช้งาน : `plot(x,y)` โดย `x` เป็นค่าตามแนวนอนและ `y` เป็นค่าตามแนว `x`

ฟังก์ชัน `lines()` เป็นการลากเส้นเชื่อมจุด

รูปแบบการใช้งาน : `lines(x,y)` โดย `x` เป็นค่าตามแนวนอนและ `y` เป็นค่าตามแนว `x`

ฟังก์ชัน `barplot()` เป็นการ plot กราฟแบบกราฟแท่ง

รูปแบบการใช้งาน : `barplot(x,y)` โดย `x` เป็นค่าตามแนวนอนและ `y` เป็นค่าตามแนว `x`

ฟังก์ชัน `pie()` เป็นการ plot กราฟแบบ `piechart`

รูปแบบการใช้งาน : `pie(x,y)` โดย `x` เป็นค่าตามแนวนอนและ `y` เป็นค่าตามแนว `x`

ฟังก์ชัน `cbind()` เป็นการนำคอลัมน์มารวมกัน

รูปแบบการใช้งาน : `cbind(ชื่อตัวแปร)`

ฟังก์ชัน `mnorm()` เป็นการสร้างค่าแบบสุ่ม

รูปแบบการใช้งาน : `mnorm(มิติ)` เราสามารถกำหนดขนาดของมิติได้ภายในวงเล็บ

ฟังก์ชัน `data.entry()` เป็นการกำหนดค่าแบบแสดงเป็นตาราง โดยต้องมีการกำหนดค่าของตัวแปรนั้น

รูปแบบการใช้งาน : `data.entry(ชื่อตัวแปร)` จะปรากฏตารางเก็บค่าข้อมูลขึ้นมาให้กำหนดค่า

ข้อมูลเวกเตอร์และ `list` จึงเป็นข้อมูลแบบมิติเดียว ต่างกันที่เวกเตอร์ไม่มีการปนกันของข้อมูลภายใน ในขณะที่ข้อมูลที่เราใช้ในการคำนวณสถิติมีตัวแปรหลายตัว ข้อมูลจะมีลักษณะเป็นตารางโดยที่ข้อมูลแต่ละคอลัมน์แทนตัวแปรแต่ละตัวได้ ข้อมูลหลายมิตินี้เป็นข้อมูลแบบ `matrix` หรือ `data frame` จะเป็น `matrix` ถ้าข้อมูลทุกตัวในตารางต้องเป็น `object` แบบเดียวกัน ส่วน `data frame` เป็นตารางที่ข้อมูลที่อยู่ต่างคอลัมน์เป็นข้อมูลต่างชนิดกันได้ ความต่างของ `matrix` และ `data frame` จึงเหมือนความต่างระหว่างเวกเตอร์กับ `list` เพียงแต่ว่า `matrix` และ `data frame` เป็นข้อมูลหลายมิติ

3.2 กฎเกณฑ์เบื้องต้นของการเขียนหรือใช้ฟังก์ชันใน R

การเขียนฟังก์ชันใน R มีกฎเกณฑ์ คือ นราทิพย์ จันสกุล (2552, หน้า 10-20)

1. ภาษาอังกฤษตัวพิมพ์ใหญ่และพิมพ์เล็ก เช่น `A` และ `a` เป็นสัญลักษณ์ที่แตกต่างกัน ถ้าให้ `A` และ `a` เป็นตัวแปรก็ถือว่าให้ `A` และ `a` เป็นตัวแปร 2 ตัวที่แตกต่างกัน

2. ความยาวชื่อของวัตถุหรือตัวแปรมีได้ไม่จำกัด และสามารถจำแนกได้ด้วยจุด (`.`) หรือตัวเลข ภาษาอังกฤษตัวพิมพ์ใหญ่หรือตัวพิมพ์เล็ก

3. ฟังก์ชันพื้นฐานประกอบด้วย นิพจน์ (`expression`) หรือการกำหนด (`assignment`) ถ้านิพจน์ทำหน้าที่เป็นฟังก์ชันด้วย R จะทำการคำนวณและผลลัพธ์ปรากฏที่หน้าจอโดยอัตโนมัติ แต่ค่าหรือผลลัพธ์นั้นไม่สามารถเรียกหรือนำกลับมาใช้ได้อีก ส่วนการกำหนดหนึ่งๆ นอกจากมีการคำนวณตามนิพจน์ที่ต้องการแล้วยังเก็บผลลัพธ์ไว้ในตัวแปรหรือวัตถุที่เรากำหนดไว้ผลลัพธ์ที่ได้จะไม่ปรากฏบนหน้าจอแต่สามารถเรียกกลับมาดูหรือกลับมาใช้ในการคำนวณในขั้นตอนใดขั้นตอนหนึ่งได้ ถ้าต้องการดูผลลัพธ์ที่ได้ก็ทำได้ง่ายๆ โดยการพิมพ์ชื่อตัวแปรนั้นแล้วเคาะแป้น `enter`

ตัวอย่าง

ฟังก์ชันใน R	คำอธิบาย
<code>>2+3</code> <code>[1] 5</code>	ต้องการหาผลบวกของ 2 และ 3 ผลลัพธ์จะปรากฏบนหน้าต่าง R console แต่ไม่สามารถเรียกกลับมาใช้ได้
<code>> x <- 2+3</code>	ต้องการหาผลบวกของ 2 และ 3 และเก็บผลลัพธ์ไว้ในตัวแปรชื่อ x
<code>></code>	ผลที่ปรากฏบนหน้าต่าง R console
<code>> x</code> <code>[1] 5</code>	ต้องการดูผลลัพธ์ที่เก็บไว้ใน x ให้พิมพ์ x ผลลัพธ์ที่ปรากฏบนหน้าต่าง R console

4. ฟังก์ชันแต่ละฟังก์ชันสามารถพิมพ์ไว้ในบรรทัดเดียวกันได้ แต่ต้องคั่นด้วยเครื่องหมาย “;” เช่น

```
> mean(x); var(x)
```

5. ฟังก์ชันย่อยแต่ละฟังก์ชันสามารถรวมเป็นหนึ่งโปรแกรมฟังก์ชันได้โดยเชื่อมด้วยสัญลักษณ์ทางคณิตศาสตร์ และล้อมรอบด้วย {} เช่น

```
tTest <-function(x, mu=0)
{
  n <- length(x)
  t <- sqrt(n)*(mean(x) - mu)/sd(x)
  p.val <- pt(abs(t), n-1, lower.tail=F)
  ttest <- c(t, p.val)
  return(ttest)
}
```

6. สามารถเขียนคำอธิบายได้เกือบทุกที่บนหน้าต่าง R console เพียงแต่ให้ขึ้นต้นด้วย # และคำอธิบายต้องเป็นภาษาอังกฤษเท่านั้น

7. ในการใช้ฟังก์ชันต่างๆ ถ้าไม่สามารถเขียนให้จบภายในหนึ่งบรรทัด R จะแสดงเครื่องหมาย “+” ซึ่งเป็นตัวพร้อมใหม่แทน “>” ที่บอกให้ทราบว่าฟังก์ชันนั้นยังไม่เสร็จสมบูรณ์ ถ้าต้องการออกจากตัวพร้อม “+” ให้เกาะแป้น Esc

8. การนำข้อมูลเข้ามาใช้ในโปรแกรมต้องกำหนด directory ให้โปรแกรม R เข้าถึงได้เรียกว่า “working directory” (wd) ฟังก์ชันต่อไปนี้จะใช้จัดการกับ directory `getwd()` พิมพ์โดยไม่ต้องมี argument จะได้ working directory ปัจจุบัน `setwd(“dir”)` กำหนดให้ working directory เป็น dir

3.3 ภาษา R

คำและความหมายที่ใช้ใน R

3.3.1 วัตถุ

R จะให้หน่วยความจำเก็บรวบรวมข้อมูลที่มีโครงสร้างหลายประเภทและเรียกสิ่งหรือเครื่องมือที่ใช้เก็บรวบรวมข้อมูลที่มีโครงสร้างหลากหลายนั้นว่า วัตถุ

3.3.2 ประเภทของวัตถุใน R

1) เวกเตอร์

เวกเตอร์มีความหมายทางคณิตศาสตร์ คือ ปริมาณที่มีขนาดและทิศทางแต่ใน R จะหมายถึงสิ่งที่ใช้เก็บข้อมูลที่เป็นค่าของตัวแปรหนึ่งตัว เวกเตอร์ใน R มี 5 ประเภท ได้แก่ ตรรกะ (logical) จำนวนเต็ม (integer) จำนวนจริง (real) จำนวนเชิงซ้อน (complex) และตัวอักษร (character หรือ string) นั่นคือสมาชิกของเวกเตอร์อาจเป็น ค่าตรรกะตัวเลขที่เป็นจำนวนเต็มหรือทศนิยม หรือแม้แต่ข้อความอย่างใดอย่างหนึ่งก็ได้ เช่น อายุของตัวอย่าง ชั่วโมงทำงานต่อวัน เป็นต้น จำนวนสมาชิกในเวกเตอร์ก็คือขนาดของเวกเตอร์ ตัวอย่าง 1.3 ฟังก์ชันที่ใช้ในการสร้างเวกเตอร์ใน R ตัวอย่างเช่น `c()` และ `sqr()`

ฟังก์ชันใน R	คำอธิบาย
<pre>> x <- c(1, 5, 10, 2, 8, -3, 0, 7) > x [1] 1.0 5.0 10.2 8.0 -3.0 0.0 7.0</pre>	สร้างเวกเตอร์ชื่อ x ที่มีสมาชิก 7 ตัว ประกอบด้วย 1, 5, 10, 2, 8, -3, 0, 7
<pre>> y <- seq(2, 12, by=2) > y [1] 2 4 6 8 10 12</pre>	สร้างเวกเตอร์ ชื่อ y ที่มีลักษณะเป็นอันดับที่มีส่วนเพิ่ม เท่ากับ 2

2) เมทริกซ์

เมทริกซ์เป็นการนำเวกเตอร์มาเขียนในลักษณะที่กำหนดมิติ (dimension)

ฟังก์ชันใน R	คำอธิบาย
<pre>> y <- seq(2, 12, by=2) > y [1] 2 4 6 8 10 12</pre>	สร้างเวกเตอร์ ชื่อ y ที่มีลักษณะเป็นอันดับที่มีส่วนเพิ่ม เท่ากับ 2
<pre>> z <- matrix(y, nrow = 2) > z [,1] [,2] [,3] [1,] 2 6 10</pre>	สร้างเมทริกซ์ z ขนาด 2 x3 โดยใช้สมาชิกของ y

ฟังก์ชันใน R	คำอธิบาย
[2,] 4 8 12	

3) รายการ

รายการ (list) เป็นวัตถุอีกชนิดหนึ่งที่ใช้เก็บรวบรวมข้อมูล โดยข้อมูลนั้นอาจอยู่ในรูปของเวกเตอร์ เมทริกซ์ ฯลฯ นั่นคือรายการสามารถรวบรวมวัตถุต่างชนิดกันไว้ด้วยกัน

ฟังก์ชันใน R	คำอธิบาย
<pre>> xyz.list <-(x=x, y=y, z=z, sex=sex) > xyz.list \$x [1] 1.0 5.0 10.2 8.0 -3.0 0.0 7.0 \$y [1] 2 4 6 8 10 12 \$z [,1] [,2] [,3] [1,] 2 6 10 [2,] 4 8 12 \$sex [1] "male" "female" "female" "male" "female" "male" "female"</pre>	xyz.list เป็นวัตถุที่เป็นรายการและรวบรวม x, y, z ซึ่งมีคุณสมบัติต่างกันไว้ด้วยกัน ลักษณะของการเก็บวัตถุต่างๆ ในรายการที่ชื่อ xyz.list

4) ว่าง

ว่างเป็นวัตถุชนิดพิเศษที่บอกให้ทราบว่าวัตถุนั้นไม่มีตัวตน ว่างมักจะถูกกำหนดให้เป็นค่าเริ่มต้นของการคำนวณ ฟังก์ชันใน R ที่กำหนดให้วัตถุนั้นว่าง คือ NULL

ฟังก์ชันใน R	คำอธิบาย
<pre>> null Error: Object "null" not found > null <-NULL > null NULL > null.15 <-c(null, 15)</pre>	มี error เนื่องจากไม่มีวัตถุชื่อ null ในสารบบที่ทำงาน (work directory) null เป็นเวกเตอร์ที่ไม่มีสมาชิกเลข

> null.15 [1] 15	null.15 เป็นเวกเตอร์ที่มีสมาชิกเพียงตัวเดียว คือ 15
---------------------	---

5) แฟกเตอร์

ใน R แฟกเตอร์ คือ เวกเตอร์ที่มีสมาชิกเป็นค่าที่มีระดับการวัดมาตรฐานบัญญัติ หรืออันดับ ในการวิเคราะห์จำเป็นต้องแปลงมาตรนั้นให้เป็นตัวเลข หรือรหัส โดยรหัสที่จะให้มีคุณสมบัติเป็นระดับ แต่ไม่สามารถใช้ในการคำนวณได้นอกจากใช้ในการจำแนกประเภทเท่านั้น

ตัวอย่าง

ฟังก์ชันใน R	คำอธิบาย
> sex <- c("male", "female", "female", "male", "female", "male", "female")	Sex เป็นเวกเตอร์ที่มีสมาชิกเป็นตัวอักษร
> fsex <- factor(sex)	เปลี่ยนเวกเตอร์ sex เป็นแฟกเตอร์ขนาดเดียวกัน และเก็บไว้ในชื่อ sex
> fsex [1] male female female male female [5] male female Levels: female male	เรียก fsex เพื่อดูสมาชิกทั้งหมด เป็นตัวที่บอกให้รู้ว่าสมาชิกตัวแรกของ fsex เริ่มที่ male เป็นตัวที่บอกให้รู้ว่า fsex เป็นแฟกเตอร์ที่มี 2 ระดับ คือ female และ male

6) กรอบข้อมูล

กรอบข้อมูล คือ ชุดข้อมูลใน R กรอบข้อมูลมีลักษณะคล้ายหรือเกือบเหมือนกับชุดข้อมูล ซึ่งจะเก็บไว้ในหน่วยความจำในลักษณะเมทริกซ์ กรอบข้อมูลชุดหนึ่งอาจประกอบด้วย เวกเตอร์ หรือแฟกเตอร์ที่มีขนาด (หรือจำนวนแถว ถ้ามองในลักษณะของเมทริกซ์) เท่ากัน โดยทั่วไปสตมภ์ของกรอบข้อมูลจะเป็นชื่อตัวแปร ส่วนแถวจะแทนลำดับที่ของข้อมูลแต่ละหน่วย

ตัวอย่าง

ฟังก์ชันใน R	คำอธิบาย
> examp.dat <- data.frame (x,fsex) > examp.dat x fsex	กรอบข้อมูลชื่อ examp.dat สร้างจากเวกเตอร์ชื่อ x และ fsex และเก็บไว้ใน R โดยฟังก์ชัน data.frame()

ฟังก์ชันใน R			คำอธิบาย
1	1.0	male	
2	5.0	female	
3	10.2	female	
4	8.0	male	
5	-3.0	female	
6	0.0	male	
7	7.0	female	
<pre>> xz.dat <- data.frame(x, sex, fsex,z)</pre> Error in data.frame (x, sex, fsex, z) : arguments imply differing number of row: 7, 2			มี error ในการสร้างกรอบข้อมูล เนื่องจาก z เป็น เมทริกซ์ที่มีจำนวนแถวเท่ากับ 2 ส่วน x, sex และ fsex มีขนาดเท่ากับ 7

7) ลักษณะประจำ

วัตถุใน R ทุกประเภทจะมีลักษณะประจำ (attribute) อย่างน้อยหนึ่งชนิดเป็นส่วนประกอบอยู่ ยกเว้นว่า โดยส่วนประกอบจะถูกเก็บไว้ในวัตถุในลักษณะรายการตามชื่อของส่วนประกอบนั้น ฟังก์ชันที่ใช้ตรวจสอบลักษณะประจำของวัตถุ คือ attribute()

ตัวอย่าง

ฟังก์ชันใน R	คำอธิบาย
<pre>> attributes (examp.dat)</pre> \$names [1] "x" "fsex" \$row.names [1] "1" "2" "3" "4" "5" "6" "7" \$class [1] "data.frame" <pre>> examp.dat\$x</pre> [1] 1.0 5.0 10.2 8.0 -3.0 0.0 7.0 <pre>> examp.dat\$fsex</pre> [1] male female female male female male female Levels: female male <pre>> attributes (xyz.list)</pre>	examp.dat ประกอบด้วย -เวกเตอร์ ชื่อ x และ fsex -row, name คือ 1, 2, ..., 7 - คลาสของ exam.dat คือ กรอบข้อมูล -ต้องการเรียกใช้ x ใน examp.dat -ต้องการเรียกใช้ fsex ใน examp.dat xyz.list ประกอบด้วยวัตถุชื่อ x, y, z และsex

ฟังก์ชันใน R	คำอธิบาย
\$names [1] "x" "y" "z" "sex"	

ฟังก์ชันพื้นฐานใน R สำหรับนักคณิตศาสตร์และนักสถิติ

ฟังก์ชัน	คำอธิบาย	ตัวอย่าง
e()	สร้างเวกเตอร์	x<- c(15, 25, 8, 6 12) เวกเตอร์ x ที่มีสมาชิกเป็น 15, 25, 8, 6 และ 12
seq()	สร้างเวกเตอร์ที่มีคุณสมบัติเป็นอันดับ	y<-seq(2, 14, 2) เวกเตอร์ y ที่มีสมาชิกเป็น 2, 4,..., 14
sum()	คำนวณผลบวกของวัตถุใน ()	sum(x) ผลลัพธ์เหมือนกับ 15+25+8+6+12 คือ 66
sqrt()	รากที่สองของทุกสมาชิกของวัตถุใน ()	sqrt(x)
abs()	ค่าสัมบูรณ์ของทุกสมาชิกของวัตถุใน ()	abs(x)
median()	ค่ามัธยฐานของวัตถุใน ()	median(x)
range()	ค่าพิสัยของวัตถุใน ()	range(x)
mean()	คำนวณค่าเฉลี่ยของวัตถุใน ()	mean(x) ผลลัพธ์เหมือนกับ $\frac{15+25+8+6+12}{5}$
var()	คำนวณความแปรปรวนของวัตถุใน ()	var(x)
sd()	คำนวณค่าเบี่ยงเบนมาตรฐานของวัตถุใน ()	sd(x)
log()	รากที่สองของทุกสมาชิกของวัตถุใน ()	log(x)
exp()	รากที่สองของทุกสมาชิกของวัตถุใน ()	> exp(x)

สัญลักษณ์และการดำเนินการทางคณิตศาสตร์

สัญลักษณ์	คำอธิบาย	ตัวอย่าง
-	เครื่องหมายลบ	-2 หรือ 7 - 3 ให้ผลลัพธ์คือ 4
+	เครื่องหมายบวก	+2 หรือ 7 + 3 ให้ผลลัพธ์คือ 10
*	เครื่องหมายคูณ	7 * 3 ให้ผลลัพธ์คือ 21
/	เครื่องหมายหาร	9/3 ให้ผลลัพธ์คือ 3
^	การยกกำลัง	3^2 ให้ผลลัพธ์คือ 9
!	นิเสธ	> x <-c (2, 3, 6, 4)

สัญลักษณ์	คำอธิบาย	ตัวอย่าง
		$> (x \neq 3)$ [1] TRUE FALSE TRUE FALSE
$>$		มากกว่า
$<$		น้อยกว่า
$\geq$		มากกว่าหรือเท่ากับ
$\leq$		น้อยกว่าหรือเท่ากับ
$t(x)$	X^T	เมทริกซ์สลับเปลี่ยนของ X
$t(x)\%*\%X$	$X^T X$	

3.4 คำสั่งในการสร้างแบบทดสอบคู่ขนาน 2 ฉบับ

Diao & van der Linden (2011) แนะนำพร้อมทั้งยกตัวอย่างการใช้โปรแกรม R ชุดคำสั่ง lp_solve version 5.5 เพื่อใช้ในการจัดชุดแบบทดสอบ 3 กรณี คือ 1) การจัดชุดแบบทดสอบมากกว่า 1 ฉบับที่มีค่าฟังก์ชันสารสนเทศตรงกับค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับเป้าหมาย 2) การจัดชุดแบบทดสอบเพื่อการทดสอบแบบปรับเหมาะ และ 3) การจัดชุดแบบทดสอบมากกว่า 1 ขั้นตอน สำหรับคำสั่งที่ใช้ในการจัดชุดแบบทดสอบคู่ขนาน 2 ฉบับที่มีค่าฟังก์ชันสารสนเทศตรงกับค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับเป้าหมาย คือ

R Code for Assembly of Two Parallel Test Forms

```
library(lpSolveAPI)
```

```
S= read.csv("C:/BIB/single test/Science9data.csv,"header=T)
```

```
ID = S$Item
```

```
A = S$a
```

```
B = S$b
```

```
C = S$c
```

```
Content = S$Content
```

```
I = nrow(S)
```

```
F = 2
```

```
J = 3
```

```
N = 55
```

```
Vc = list()
```

```
for(k in 1:6){
```

```

Vc[[k]] = c(1:I)[Content==k]
}
theta = c(-1.5,0,1.5)
J = length(theta)
Info = array(0,c(I,J))
for(j in 1:J){
P = C+(1-C)/(1+exp(-1.7)*A*(theta[j]-B)))
Q = 1-P
Info[,j] = (1.7^2)*(A^2)*((P-C)/(1-C))^2*Q/P
}
M = I*J+1
# Create the model
lprec=make.lp(0,M)
# Set control parameters: minimization problem; integer tolerance is set to 0.1;
# absolute MIP gap is set to 0.1; relative MIP gap is set to 0.05
lp.control(lprec,sense="min",epsint = 0.1, mip.gap=c(0.1,0.05));
# Constraints (8) and (9)
set.type(lprec,columns=c(1:(2*I)),type="binary")
set.type(lprec,columns=M,type="real")
set.bounds(lprec,lower=rep(0,M),upper=rep(1,M))
# Constraint (5)
for (k in 1:I){
add.constraint(lprec,rep(1,2),"<=",1,indices=c(k,I+k))
}
# Constraint (6)
Nc = c(9,9,7,9,9,11)
# Form 1
for(k in 1:6){
add.constraint(lprec,rep(1,length(Vc[[k]])),">=",Nc[k],indices=Vc[[k]])
}
# Form 2

```

```

for(k in 1:6){
add.constraint(lprec,rep(1,length(Vc[[k]])),>=,Nc[k],indices=I+Vc[[k]])
}

# Constraint (7)
add.constraint(lprec, rep(1,I), "=", N, indices = 1:I)
add.constraint(lprec, rep(1,I), "=", N, indices = (I+1):(2*I))

# Constraints (3) and (4)
d_theta=c(5.4, 10, 5.4)
for(k in 1:3){
add.constraint(lprec,c(Info[,k],-1),<=,d_theta[k],indices=c(1:I,M))
add.constraint(lprec,c(Info[,k],-1),<=,d_theta[k],indices=c((I+1):(2*I),M))
add.constraint(lprec,c(Info[,k],1),>=,d_theta[k],indices=c(1:I,M))
add.constraint(lprec,c(Info[,k],1),>=,d_theta[k],indices=c((I+1):(2*I),M))
}

# Objective function
set.objfn(lprec,1,indices=M)

# Solve the model
res_flag = solve(lprec)

# The integer value containing the status code, for example, 0: "optimal solution found"
res_flag

# Retrieve the values of the decision variables
x_opt = get.variables(lprec)

```

ตอนที่ 4 ผลการสังเคราะห์เอกสารและงานวิจัยที่เกี่ยวข้องกับการสร้างแบบทดสอบคู่ขนานมากกว่า 1 ฉบับ

ตามทฤษฎีการตอบข้อสอบ

จากการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องกับการสร้างแบบทดสอบคู่ขนานมากกว่า 1 ฉบับตามทฤษฎีการตอบข้อสอบ ผู้วิจัยได้นำเสนอเป็น 3 ตอน คือ ตอนที่ 1 แนวคิดและหลักการของโมเดลเชิงเส้นสำหรับการออกแบบการสร้างแบบทดสอบ และงานวิจัยที่เกี่ยวข้อง ตอนที่ 2 แนวคิดและหลักการการตรวจสอบอัตราการใช้ข้อสอบระหว่างแบบทดสอบและงานวิจัยที่เกี่ยวข้อง และตอนที่ 3 การใช้โปรแกรม R ผู้วิจัยสามารถสังเคราะห์ประเด็นสำคัญ ดังนี้

1) การออกแบบการสร้างแบบทดสอบตามทฤษฎีการตอบข้อสอบ Birmmaum (1970) เป็นผู้เสนอขั้นตอนการออกแบบการสร้างแบบทดสอบ 3 ขั้นตอน คือ (1) กำหนดเป้าหมายของการใช้แบบทดสอบ เช่น การจัดตำแหน่งผู้เรียน การวินิจฉัยผู้เรียน การสร้างแบบทดสอบคู่ขนานมากกว่า 1 ฉบับ เป็นต้น (2) การแปลงเป้าหมายของการสร้างแบบทดสอบเป็นค่าฟังก์ชันสารสนเทศตามแบบทดสอบฉบับเป้าหมาย และ (3) การจัดชุดแบบทดสอบจากคลังข้อสอบ สำหรับข้อจำกัดการออกแบบการสร้างแบบทดสอบตามแนวคิดของ Birmmaum คือ ขาดการเสนอแนะแนวคิดลงสู่การปฏิบัติอย่างชัดเจน ส่วนแนวคิดของ van der Linden (2005) ได้เสนอขั้นตอนการออกแบบการสร้างแบบทดสอบ 3 ขั้นตอน คือ (1) ระบุคุณลักษณะเฉพาะของการออกแบบการสร้างแบบทดสอบ เช่น ค่าพารามิเตอร์ของข้อสอบ จำนวนคำทั้งหมดที่ใช้ในแบบทดสอบ และกำหนดระดับขั้นต่ำของการใช้ข้อสอบระหว่างแบบทดสอบ (2) สร้างตารางกำหนดคุณลักษณะของแบบทดสอบที่จำแนกตามเนื้อหาการวัดและระดับพฤติกรรมการเรียนรู้ และ (3) พัฒนากลังข้อสอบประกอบด้วย (3.1) การเขียนข้อสอบ ตรวจสอบความเหมาะสมของภาษาที่ใช้ (3.2) ตรวจสอบความตรงตามเนื้อหา (3.3) ทดลองใช้ข้อสอบ (3.4) วิเคราะห์คุณภาพข้อสอบและคัดเลือกข้อสอบที่มีคุณภาพ (3.5) วิเคราะห์ความสอดคล้องของผลการสอบกับโมเดลการวิเคราะห์ (3.6) ดำเนินการจัดชุดแบบทดสอบ

สำหรับในการวิจัยครั้งนี้ผู้วิจัยดำเนินการออกแบบการสร้างแบบทดสอบตามแนวคิดของ van der Linden กล่าวคือมีขั้นตอนการออกแบบการสร้างแบบทดสอบ 3 ขั้นตอน คือ (1) ระบุคุณลักษณะเฉพาะของการออกแบบการสร้างแบบทดสอบ ประกอบด้วย ระบุค่าพารามิเตอร์ของข้อสอบ ค่าฟังก์ชันสารสนเทศของแบบทดสอบ ค่าฟังก์ชันคุณลักษณะของแบบทดสอบตามแบบทดสอบฉบับเป้าหมาย และกำหนดอัตราการใช้ข้อสอบระหว่างแบบทดสอบ (2) สร้างตารางกำหนดคุณลักษณะของแบบทดสอบฉบับเป้าหมายที่จำแนกตามเนื้อหาการวัดรายวิชา MER 2003 และระดับพฤติกรรมการเรียนรู้ และ (3) พัฒนากลังข้อสอบประกอบด้วย (3.1) การเขียนข้อสอบ (3.2) ตรวจสอบความเหมาะสมของภาษาที่ใช้ ตรวจสอบความตรงตามเนื้อหา (3.3) ทดลองใช้ข้อสอบ (3.4) วิเคราะห์คุณภาพข้อสอบและคัดเลือกข้อสอบที่มีคุณภาพ (3.5) วิเคราะห์ความสอดคล้องของผลการสอบกับโมเดลการวิเคราะห์ตามทฤษฎีการตอบข้อสอบ (3.6) ดำเนินการสร้างแบบทดสอบคู่ขนาน

2) การจัดชุดแบบทดสอบ สามารถจัดการได้ 4 ลักษณะ คือ

2.1) การสุ่มข้อสอบเพื่อนำมาจัดชุดแบบทดสอบ (random sampling of tests) ข้อจำกัดของวิธีนี้คือ ค่อนข้างสิ้นเปลืองเวลา งบประมาณและแบบทดสอบที่ถูกจัดชุดมีคุณลักษณะเฉพาะที่ผู้สร้างต้องการค่อนข้างยาก

2.2) การสุ่มแบบทดสอบแบบลำดับขั้น (sequential sampling) วิธีนี้นิยมใช้กับแบบทดสอบเพื่อใช้ในการทดสอบแบบอิงเกณฑ์ เพื่อวัดความรอบรู้ของผู้สอบ

2.3) การจัดชุดแบบทดสอบให้ตรงกับคุณลักษณะเฉพาะที่ดีที่สุด (optimal test assembly) เป็นการจับหมวดหมู่ข้อสอบด้วยวิธี combination และใช้วิธีการแก้ปัญหาทางคณิตศาสตร์ด้วยโปรแกรมสำเร็จรูป (heuristic methods) เพื่อให้ได้แบบทดสอบที่ตรงตามคุณลักษณะเฉพาะที่ดีที่สุด

2.4) การจัดแบบทดสอบเพื่อการทดสอบแบบปรับเหมาะ

สำหรับการวิจัยครั้งนี้ใช้การจัดชุดแบบทดสอบเพื่อนำมาสร้างแบบทดสอบกลุ่มนานมากกว่า 1 ฉบับ ด้วยวิธีการจัดชุดแบบทดสอบให้ตรงกับคุณลักษณะเฉพาะที่ดีที่สุด (optimal test assembly) เป็นการจับหมวดหมู่ข้อสอบจากคลังข้อสอบด้วยวิธี combination และใช้วิธีการแก้ปัญหาทางคณิตศาสตร์ด้วยโปรแกรม R เพื่อให้ได้แบบทดสอบที่มีคุณลักษณะความเป็นกลุ่มนานตามทฤษฎีการตอบข้อสอบแบบ 3 พารามิเตอร์

3) อัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบ Chang & Zhang ได้เสนอสูตรการคำนวณและค่าขีดจำกัดล่างที่แสดงถึงการใช้ข้อสอบซ้ำระหว่างแบบทดสอบ คำนวณจากสูตร $r = (k \div n)$ เมื่อ k หมายถึงความยาวของแบบทดสอบ n หมายถึง จำนวนข้อสอบในคลังข้อสอบ โดยมีข้อตกลงเบื้องต้น คือ ข้อสอบต้องถูกสุ่มแบบสุ่มสมบูรณ์ (completely randomized item selection) แต่ข้อตกลงเบื้องต้นดังกล่าวไม่สอดคล้องกับแนวทางปฏิบัติ ต่อมา Chen, Ackermann & Spray ได้ทำการศึกษาต่อด้วยการเสนอแนวทางการนำไปใช้และให้ข้อคิดเห็นเพิ่มเติมเกี่ยวกับสูตรการควบคุมการใช้ข้อสอบซ้ำระหว่างแบบทดสอบให้ต่ำที่สุด $r = (k \div n)$ พบว่ากรณีวัตถุประสงค์ของการจัดชุดแบบทดสอบจากคลังข้อสอบคำนึงถึงคุณสมบัติด้านจิตมิติ (psychometric properties) ของแบบทดสอบ ค่าคำนวณอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบจาก $(k \div n)$ จะเป็นค่าประมาณที่สูงเกินความเป็นจริง และ Lin ได้ให้ข้อเสนอแนะในการใช้ค่า E(BTOR) เพื่อการควบคุมอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบ เนื่องจากหลักการคำนวณของค่าดัชนี E(BTOR) ได้คำนึงถึงความเป็นตัวแทนของเนื้อหาการวัดซึ่งเป็นการสะท้อนคุณภาพของแบบทดสอบด้านความตรงตามเนื้อหา (content validity) จุดเด่นของค่า E(BTOR) คือ เป็นการผ่อนคลायข้อตกลงเบื้องต้นบางประการจากสูตรของ Chen และคณะ

สำหรับการวิจัยครั้งนี้ผู้วิจัยใช้แนวคิดการคำนวณอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบตามแนวคิดของ ของ Chen, Ankenmann and Spray (2003, p.129) ที่ทำการควบคุมอัตราการใช้ข้อสอบซ้ำของข้อสอบที่พิจารณาจาก 2 มิติ คือ 1) การควบคุมระดับข้อสอบ คืออัตราของข้อสอบซ้ำ (item exposure rate: Γ_i) และ 2) การควบคุมระดับแบบทดสอบ คืออัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบ (item overlap rate)

จากผลการสังเคราะห์เอกสารและงานวิจัยที่เกี่ยวข้องจะเห็นได้ว่าการสร้างแบบทดสอบคู่ขนานจากคลังข้อสอบตามทฤษฎีการตอบข้อสอบนั้นควรคำนึงถึงความตรงตามเนื้อหา ความเที่ยงในการประมาณค่าความสามารถของผู้สอบ และคุณลักษณะเฉพาะของแบบทดสอบคู่ขนานเพื่อให้ได้แบบทดสอบคู่ขนานมากกว่า 1 ฉบับ ความเป็นคู่ขนานของแบบทดสอบเป็นส่วนสำคัญของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียน และสามารถเทียบเคียงผลการสอบอย่างยุติธรรมและเกิดความถูกต้องแม่นยำในการประมาณค่าความสามารถของนักศึกษา

บทที่ 3

วิธีดำเนินการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์หลักเพื่อสร้างแบบทดสอบคู่ขนานจากคลังข้อสอบที่มีคุณภาพด้านความเที่ยงและความตรงตามทฤษฎีการตอบข้อสอบ โดยมีวัตถุประสงค์ย่อย คือ 1) เพื่อสร้างโปรแกรมการคัดเลือกข้อสอบจากคลังข้อสอบเพื่อสร้างแบบทดสอบคู่ขนานที่มีคุณภาพด้านความเที่ยง (reliability) และความตรง (validity) ตามทฤษฎีการตอบข้อสอบ 2) เพื่อตรวจสอบความถูกต้องของโปรแกรมการคัดเลือกข้อสอบด้านคุณลักษณะเฉพาะของความเป็นคู่ขนาน (parallelism specification) ของแบบทดสอบตามตำแหน่งความสามารถของผู้สอบของแบบทดสอบฉบับอ้างอิง และ 3) เพื่อตรวจสอบความถูกต้องของโปรแกรมการคัดเลือกข้อสอบด้านการใช้ข้อสอบซ้ำระหว่างแบบทดสอบคู่ขนาน (test overlap) ตามตามตำแหน่งความสามารถของผู้สอบของแบบทดสอบฉบับอ้างอิง ผู้วิจัยจึงดำเนินการวิจัยเพื่อสร้างคลังข้อสอบตามทฤษฎีการตอบข้อสอบที่มีจุดมุ่งหมายเพื่อนำมาสร้างแบบทดสอบคู่ขนานมากกว่า 1 ฉบับ โดยมีขั้นตอนการวิจัย ดังนี้

ขั้นตอนที่ 1 การสร้างและตรวจสอบคุณภาพของแบบทดสอบฉบับอ้างอิง

1.1 การสร้าง ตรวจสอบคุณภาพ และคัดเลือกข้อสอบของแบบทดสอบฉบับอ้างอิงตามทฤษฎีการทดสอบแบบดั้งเดิม

1.2 การตรวจสอบคุณภาพของข้อสอบที่ผ่านการคัดเลือกของแบบทดสอบฉบับอ้างอิงตามทฤษฎีการตอบข้อสอบ

ขั้นตอนที่ 2 การสร้างคลังข้อสอบตามคุณสมบัติของแบบทดสอบฉบับอ้างอิงด้วยการจำลองข้อมูลการสอบด้วยโปรแกรม WINGEN

2.1 กำหนดสัดส่วนของคลังข้อสอบตามเนื้อหาการวัดตามแบบทดสอบฉบับอ้างอิง

2.2 การจำลองข้อมูลการสอบด้วยโปรแกรม WINGEN

2.3 การตรวจสอบและคัดเลือกข้อสอบเพื่อบรรจุในคลังข้อสอบที่มีคุณภาพตามทฤษฎีการตอบข้อสอบ ดังรายละเอียด

ขั้นตอนที่ 1 การสร้างและตรวจสอบคุณภาพของแบบทดสอบฉบับอ้างอิง

1.1 การสร้าง ตรวจสอบคุณภาพ และคัดเลือกข้อสอบของแบบทดสอบฉบับอ้างอิงตามทฤษฎีการทดสอบแบบดั้งเดิม

แบบทดสอบฉบับอ้างอิงของการวิจัยครั้งนี้ คือแบบทดสอบรายวิชา MER 2003 ความรู้เบื้องต้นเกี่ยวกับสถิติและวิจัยทางการศึกษา ซึ่งเป็นรายวิชาที่มีจำนวนนักศึกษาหลักสูตรศึกษาศาสตรบัณฑิตลงทะเบียนเรียนเป็นจำนวนมากในแต่ละภาคเรียน ผู้วิจัยจึงดำเนินการวิจัยในขั้นตอนการสร้างและตรวจสอบคุณภาพของแบบทดสอบรายวิชา MER 2003 ดังนี้

1.1.1 สร้างตารางกำหนดลักษณะเฉพาะของแบบทดสอบ (table of specification) รายวิชา MER 2003 ความรู้เบื้องต้นเกี่ยวกับสถิติและวิจัยทางการศึกษา เพื่อใช้ในการเขียนข้อสอบสำหรับการวัดและประเมินผลปลายภาคเรียน โดยกลุ่มอาจารย์ผู้สอนรายวิชา MER 2003 จำนวน 3 ท่านร่วมกันสร้างตารางกำหนดคุณลักษณะของแบบทดสอบ คือ 1) รองศาสตราจารย์ ดร.สุวิมล ติรภานันท์ 2) ผู้ช่วยศาสตราจารย์ ดร.ทวิภา ตั้งประภา และ 3) ผู้ช่วยศาสตราจารย์ ดร.สุวิมล กฤษกุลหาสน์ ดังตาราง 3.1

ตาราง 3.1 การกำหนดลักษณะเฉพาะของแบบทดสอบ (table of specification) รายวิชา MER 2003 ความรู้เบื้องต้นเกี่ยวกับสถิติและวิจัยทางการศึกษา เพื่อใช้ในการเขียนข้อสอบสำหรับการวัดและประเมินผลปลายภาคเรียน

ที่	เนื้อหา	จำนวนข้อสอบ
1	การค้นหาคำรู้และวิธีการทางวิทยาศาสตร์	9
2	ประเภทของการวิจัยทางสังคมศาสตร์	11
3	การกำหนดประเด็นปัญหาและวัตถุประสงค์ในการวิจัย การทบทวนวรรณกรรม	9
4	สมมติฐานการวิจัยและตัวแปรในการวิจัย	25
5	การออกแบบการวิจัย	11
6	เครื่องมือเก็บรวบรวมข้อมูลและการสร้างแบบทดสอบถาม การตรวจสอบคุณภาพเครื่องมือ	10
7	การสุ่มตัวอย่าง การใช้สถิติในการวิจัย	19
8	การเขียนข้อเสนอโครงการวิจัย และรายงานการวิจัย	6
รวม		100

1.1.2 สร้างข้อสอบปรนัยแบบเลือกตอบ จำนวน 4 ตัวเลือกตามตารางกำหนดลักษณะเฉพาะรายวิชา (table of specification) จำนวน 100 ข้อ

1.1.3 นำแบบทดสอบที่สร้างขึ้นไปตรวจสอบคุณภาพด้านความตรงตามเนื้อหา (content validity) โดยการเสนอต่อผู้เชี่ยวชาญ จำนวน 5 ท่านเพื่อทำการพิจารณาถึงความสอดคล้องของข้อสอบกับวัตถุประสงค์ในการวัดตามระดับความรู้และกระบวนการทางพุทธิปัญญา โดยผู้เชี่ยวชาญมีคุณสมบัติ คือ มีความเชี่ยวชาญทั้งด้านการสอน การวัดและประเมินผล มีตำแหน่งทางวิชาการและผลงานทางวิชาการ มีประสบการณ์การจัดการเรียนรู้วิชาการวิจัยทางการศึกษา ตั้งแต่ 5 ปีขึ้นไป

1.1.4 นำผลการพิจารณาไปคำนวณดัชนีความสอดคล้องระหว่างข้อสอบกับวัตถุประสงค์ในการวัด (Index of Item-Objective Congruence: IOC) ด้วยสูตรของ Hambleton (as cited in Turner & Carlson, 2003, p. 167) เกณฑ์ที่ใช้ในการคัดเลือกข้อสอบ คือ ข้อสอบทุกข้อต้องมีค่า IOC มากกว่า 0.50

$$I_{ik} = \frac{(N - 1) \sum_{j=1}^n X_{ijk} + N \sum_{j=1}^n X_{ijk} - \sum_{j=1}^n X_{ijk}}{2(N - 1)n}$$

เมื่อ	I_{ik}	หมายถึง	ดัชนีความสอดคล้องระหว่างข้อคำถาม (k) กับวัตถุประสงค์ในการวัด (i) (Index of Item-Objective Congruence)
	N	หมายถึง	จำนวนวัตถุประสงค์ในการวัด ($i = 1, 2, \dots, N$)
	n	หมายถึง	จำนวนผู้เชี่ยวชาญด้านเนื้อหา ($j = 1, 2, \dots, n$)
	X_{ijk}	หมายถึง	ระดับคะแนนจากการประเมินจากผู้เชี่ยวชาญของข้อคำถาม (k) จาก วัตถุประสงค์ในการวัด (i) และของผู้เชี่ยวชาญ (j) มีค่าตั้งแต่ 1, 0, (-1)
โดยที่	1	หมายถึง	ข้อคำถามนั้นสอดคล้องกับวัตถุประสงค์ในการวัดในระดับสูง
	0	หมายถึง	ข้อคำถามนั้นสอดคล้องกับวัตถุประสงค์ในการวัดในระดับปานกลาง
	(-1)	หมายถึง	ข้อคำถามนั้นสอดคล้องกับวัตถุประสงค์ในการวัดในระดับต่ำหรือไม่สอดคล้อง

1.1.5 ทำการคัดเลือกข้อสอบที่มีค่า IOC มากกว่า 0.50 จัดเป็นแบบทดสอบฉบับอ้างอิง ในการวิจัยครั้งนี้พบว่าข้อสอบรายวิชา MER 2003 มีค่า IOC มากกว่า 0.50 จำนวน 100 ข้อ กล่าวคือมีค่า IOC อยู่ระหว่าง 0.80 ถึง 1.00

1.1.6 นำข้อสอบที่ผ่านการคัดเลือกจัดพิมพ์เป็นแบบทดสอบ

1.1.7 นำแบบทดสอบไปทดสอบกับนักศึกษาที่ลงทะเบียนรายวิชา MER 2003 จำนวน 2,800 คน

1.1.8 ดำเนินการตรวจสอบและคัดเลือกคุณภาพของข้อสอบและแบบทดสอบตามทฤษฎีการทดสอบแบบดั้งเดิม พบว่าข้อสอบมีค่าความยาก (p) อยู่ระหว่าง 0.20-0.80 และมีค่าอำนาจจำแนกที่ประมาณค่า

แบบไบซีเรียล (Biserial Correlation) ตั้งแต่ 0.20 ขึ้นไปมีจำนวน 64 ข้อ และมีค่าความเที่ยงด้วยวิธีการประมาณค่าแบบ KR-20 เท่ากับ 0.841 ดังตาราง 3.2

ตาราง 3.2 สรุปจำนวนข้อสอบที่สร้างขึ้นและผ่านเกณฑ์จำแนกตามเนื้อหาในการวัด

ที่	เนื้อหา	จำนวนข้อสอบ	
		สร้างขึ้น	ผ่านเกณฑ์
1	การค้นหาคำรู้และวิธีการทางวิทยาศาสตร์	9	5
2	ประเภทของการวิจัยทางสังคมศาสตร์	11	9
3	การกำหนดประเด็นปัญหาและวัตถุประสงค์ในการวิจัย การทบทวนวรรณกรรม	9	5
4	สมมติฐานการวิจัยและตัวแปรในการวิจัย	25	21
5	การออกแบบการวิจัย	11	7
6	เครื่องมือเก็บรวบรวมข้อมูลและการสร้างแบบทดสอบถาม การตรวจสอบคุณภาพเครื่องมือ	10	4
7	การสุ่มตัวอย่าง การใช้สถิติในการวิจัย	19	10
8	การเขียนข้อเสนอโครงการวิจัย และรายงานการวิจัย	6	3
รวม		100	64

1.2 การตรวจสอบคุณภาพของข้อสอบที่ผ่านการคัดเลือกของแบบทดสอบฉบับอ้างอิงตามทฤษฎีการตอบข้อสอบ

ผู้วิจัยนำข้อสอบที่มีคุณภาพตามทฤษฎีการทดสอบแบบดั้งเดิม จำนวน 64 ข้อมาดำเนินการตรวจสอบคุณภาพของข้อสอบและแบบทดสอบด้วยทฤษฎีการตอบข้อสอบ แบบ 3 พารามิเตอร์ ตามขั้นตอนดังนี้

1.2.1 การตรวจสอบข้อตกลงเบื้องต้นของทฤษฎีการตอบข้อสอบ

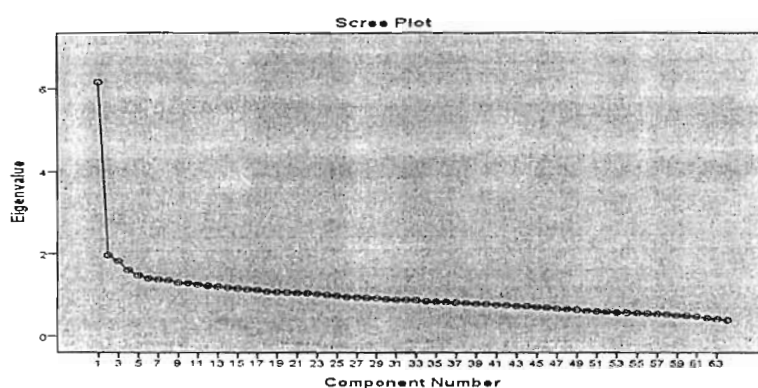
1) การตรวจสอบของแบบทดสอบ (unidimensionality)

ความเป็นเอกมิติของแบบทดสอบเป็นการพิจารณาถึงข้อสอบในแบบทดสอบมุ่งวัดคุณลักษณะเด่นเดียว (single dominant factor) ในการวิจัยครั้งนี้คณะผู้วิจัยใช้วิธีการพิจารณาความเป็นเอกมิติของแบบทดสอบด้วยการวิเคราะห์องค์ประกอบเชิงสำรวจที่อาศัยค่าไอเกนจากการวิเคราะห์องค์ประกอบเป็นหลักด้วยวิธีการทดสอบ eigen value test ซึ่งวิธีนี้เป็นการพิจารณาจากค่า eigen ที่ 1 ต่างจากที่เหลือมากๆ และค่า eigen ที่เหลือมีลักษณะเหมือนกันจะจัดว่าแบบทดสอบนั้นมีความเป็นเอกมิติ นอกจากนี้ Gorsuch (as cited

in Slocum. 2005, p.33) ได้เสนอว่าถ้าอัตราส่วนของค่า eigen ที่ 1 กับค่า eigen ที่ 2 มากกว่า 3 ถือว่าแบบทดสอบนั้นมีความเป็นเอกมิติ สำหรับผลการทดสอบแสดงดังตาราง 3.3 และภาพ 3.1

ตาราง 3.3 ค่าไอเกนของ 10 องค์ประกอบแรกของแบบทดสอบรายวิชา MER 2003 ความรู้เบื้องต้นเกี่ยวกับสถิติ และวิจัยทางการศึกษา

องค์ประกอบ	ค่าไอเกน		
	Total	% of Variance	Cumulative %
1	6.164	9.631	9.631
2	1.969	3.076	12.708
3	1.830	2.859	15.567
4	1.611	2.518	18.084
5	1.485	2.320	20.404
6	1.408	2.199	22.603
7	1.385	2.164	24.768
8	1.366	2.135	26.903
9	1.307	2.042	28.945
10	1.298	2.028	30.973



ภาพ 3.1 ค่าไอเกนจากการวิเคราะห์องค์ประกอบของผลการตอบข้อสอบรายวิชา MER 2003 จำนวน 64 ข้อ

จากตาราง 3.3 และภาพ 3.1 เมื่อพิจารณาจากค่า eigen ที่ 1 ต่างจากที่เหลือมากๆ และค่า eigen ที่เหลือมีลักษณะเหมือนๆ กัน รวมทั้งอัตราส่วนระหว่างค่า eigen ที่ 1 กับค่า eigen ที่ 2 เท่ากับ 3.131 ซึ่งอัตราส่วนดังกล่าวมีค่ามากกว่า 3 จากข้อมูลข้างต้นแสดงให้เห็นว่าแบบทดสอบ MER 2003 จำนวน 64 ข้อ วัดคุณลักษณะเด่นเดียว (single dominant factor) ซึ่งเป็นตามเกณฑ์ของ Gorsuch (as cited in Slocum, 2005, p. 33) ที่เสนอว่าหากอัตราส่วนของค่า eigen ที่ 1 กับค่า eigen ที่ 2 มากกว่า 3.000 ถือว่าแบบทดสอบนั้นมีความเป็นเอกมิติ

2) ความเป็นอิสระในการตอบข้อสอบ (local independence) หมายถึง โอกาสในการตอบข้อสอบได้ถูกต้องเป็นอิสระจากกัน นั่นคือ การตอบข้อสอบข้อใดข้อหนึ่งถูกหรือผิดจะไม่มีผลต่อการตอบข้ออื่นๆ และ Hambleton and Swaminathan (1996. pp.22-23) มีความเห็นตรงกันว่าถ้ามีความเป็นเอกมิติแล้วความเป็นอิสระในการตอบข้อสอบก็จะมีเช่นกัน

3) ตรวจสอบความสอดคล้องของข้อมูลกับโมเดล (model data fit assessment)

การประเมินความเหมาะสมของโมเดลการตอบข้อสอบเพื่อใช้ในการประมาณค่าค่าพารามิเตอร์ของข้อสอบและแบบทดสอบในการวิจัยครั้งนี้ใช้หลักการของการประมาณค่าพารามิเตอร์โดยใช้หลักการความเป็นไปได้สูงสุด (maximum likelihood) ค่า $-2 \log \text{likelihood}$ ของข้อมูลจึงแสดงถึงระดับความไม่สอดคล้องของข้อมูลเชิงประจักษ์กับโมเดล ซึ่งเป็นค่าที่มีประโยชน์สำหรับประเมินความเหมาะสมของโมเดล ค่า $-2 \log \text{likelihood}$ ที่แตกต่างระหว่างโมเดลคำนวณด้วยสถิติทดสอบ χ^2 เพื่อประเมินว่าการปรับปรุงโมเดลที่ซับซ้อนขึ้นจะมีคุณภาพหรือความเหมาะสมเพิ่มขึ้นเพียงใด เมื่อ df สำหรับการทดสอบ χ^2 คือจำนวนพารามิเตอร์ที่เพิ่มขึ้นของโมเดลที่ซับซ้อนขึ้น ผลการตรวจสอบความสอดคล้องของข้อมูลกับโมเดล พบว่าการประมาณค่าด้วยโมเดลการตอบข้อสอบแบบ 2 พารามิเตอร์ มีค่า $-2 \log \text{likelihood}$ เท่ากับ 171,870.5842 ส่วนโมเดลแบบ 3 พารามิเตอร์ค่า $-2 \log \text{likelihood}$ เท่ากับ 171,273.6639 ดังนั้นผลการเปรียบเทียบความเหมาะสมระหว่างโมเดล 2 พารามิเตอร์ กับ 3 พารามิเตอร์เพื่อใช้ในการประมาณค่าพารามิเตอร์ของข้อสอบและฟังก์ชันสารสนเทศของแบบทดสอบ MER 2003 จำนวน 64 ข้อ คือ $\chi^2_{II-III} = 171,870.5842 - 171,273.6639$ เท่ากับ 596.9203 และค่า χ^2 ที่ระดับนัยสำคัญทางสถิติ $\alpha = 0.01$ ($\chi^2_{0.01,64}$) เท่ากับ 95.6260 จากผลการทดสอบนัยสำคัญทางสถิติที่ระดับ 0.01 เพื่อประเมินความสอดคล้องของข้อมูลเชิงประจักษ์กับโมเดลการตอบข้อสอบระหว่าง 2 พารามิเตอร์กับ 3 พารามิเตอร์แสดงให้เห็นว่าโมเดลการตอบข้อสอบแบบ 3 พารามิเตอร์สอดคล้องกับข้อมูลเชิงประจักษ์มากกว่าโมเดลการตอบข้อสอบแบบ 2 พารามิเตอร์ ดังนั้นผู้วิจัยจึงใช้โมเดลการตอบข้อสอบแบบ 3 พารามิเตอร์ในการประมาณค่าพารามิเตอร์ของข้อสอบและแบบทดสอบนับอ้างอิงต่อไป แสดงดังตาราง 3.4

ตาราง 3.4 ค่า -2 log likelihood ระหว่างโมเดลการตอบข้อสอบแบบ 2 และ 3 พารามิเตอร์

โมเดลการตอบสนองข้อสอบ	ค่า -2 log likelihood	χ^2_{II-III}
2 พารามิเตอร์	171,870.5842	596.9203**
3 พารามิเตอร์	171,273.6639	

** ระดับนัยสำคัญที่ 0.01

1.2.2 วิเคราะห์ค่าพารามิเตอร์ของข้อสอบและแบบทดสอบฉบับอ้างอิง

ผลการวิเคราะห์ค่าพารามิเตอร์ของข้อสอบและค่าฟังก์ชันสารสนเทศของแบบทดสอบ MER 2003 พบว่าข้อสอบที่มีค่าอำนาจจำแนก ค่าความยาก และค่าการเดาผ่านเกณฑ์ จำนวน 64 ข้อ ซึ่งเกณฑ์การคัดเลือกค่าพารามิเตอร์ข้อสอบ คือ

1) ค่าพารามิเตอร์อำนาจจำแนก (a) ใช้เกณฑ์การแปลความหมายของ Baker (2001, p.34) คือ

0.01 – 0.34	ระดับค่อนข้างต่ำ	0.35 – 0.64	ระดับต่ำ
0.65 – 1.34	ระดับปานกลาง	1.35 – 1.69	ระดับค่อนข้างมาก
มากกว่า 1.70	ระดับมาก		

2) ค่าพารามิเตอร์ความยาก (b) ในทางปฏิบัตินิยมใช้ข้อสอบที่มีค่าพารามิเตอร์ความยากอยู่ระหว่าง (-3.00) ถึง 3.00 (Baker, 2001, p.22; De Ayala, 2009, p.101)

3) ค่าการเดา (c) ควรมีค่าต่ำกว่า 0.315 (Baker, 2001, p.28)

ผลการตรวจสอบคุณภาพของข้อสอบรายวิชา MER 2003 จำนวน 64 ข้อพบว่า มีค่าอำนาจจำแนก (a) อยู่ระหว่าง 0.206 ถึง 3.272 ค่าความยาก (b) อยู่ระหว่าง (-2.088) ถึง 2.843 และค่าการเดา (c) อยู่ระหว่าง 0.001 ถึง 0.315 โดยข้อสอบที่วัดเนื้อหาที่ 2 มีค่าความยากเฉลี่ยสูงสุด ข้อสอบที่วัดเนื้อหาที่ 6 มีความยากเฉลี่ยรองลงมา ส่วนข้อสอบที่วัดเนื้อหาที่ 8 มีค่าความยากเฉลี่ยต่ำสุด สำหรับค่าอำนาจจำแนก (a) เฉลี่ย ข้อสอบที่วัดเนื้อหาที่ 4 มีค่าเฉลี่ยของอำนาจจำแนกสูงสุด ส่วนข้อสอบที่วัดเนื้อหาที่ 1-3 และ 5-8 มีค่าเฉลี่ยของอำนาจจำแนกใกล้เคียงกัน ดังตาราง 3.5 และ 3.6

ตาราง 3.5 ค่าต่ำสุด ค่าสูงสุด ค่าเฉลี่ย และค่าเบี่ยงเบนมาตรฐานของค่าพารามิเตอร์ a, b และ c ของข้อสอบ
MER 2003 จำแนกตามเนื้อหาการวัด

ที่	เนื้อหา	จำนวนข้อ	$\bar{a}$ S.D.	$\bar{b}$ S.D.	$\bar{c}$ S.D.
1	การค้นหาคำรู้และวิธีการทางวิทยาศาสตร์	5	0.478 0.175	0.564 1.745	0.047 0.061
2	ประเภทของการวิจัยทางสังคมศาสตร์	9	0.551 0.237	1.041 1.299	0.100 0.078
3	การกำหนดประเด็นปัญหาและวัตถุประสงค์ ในการวิจัย การทบทวนวรรณกรรม	5	0.434 0.226	-0.059 1.198	0.019 0.037
4	สมมติฐานการวิจัยและตัวแปรในการวิจัย	21	1.005 0.798	0.776 0.878	0.143 0.126
5	การออกแบบการวิจัย	7	0.538 0.288	0.539 1.397	0.073 0.115
6	เครื่องมือเก็บรวบรวมข้อมูลและการสร้าง แบบทดสอบถาม การตรวจสอบคุณภาพ เครื่องมือ	4	0.600 0.206	0.901 1.374	0.128 0.149
7	การสุ่มตัวอย่าง การใช้สถิติในการวิจัย	6	0.538 0.215	0.889 1.234	0.069 0.075
8	การเขียนข้อเสนอโครงการวิจัย และรายงาน การวิจัย	7	0.566 0.226	-0.399 1.098	0.059 0.099
รวม		64			
ค่าต่ำสุด			0.206	-2.088	0.001
ค่าสูงสุด			3.272	2.843	0.315
ค่าเฉลี่ย			0.687	.595	0.095
ค่าเบี่ยงเบนมาตรฐาน			0.533	1.205	0.107

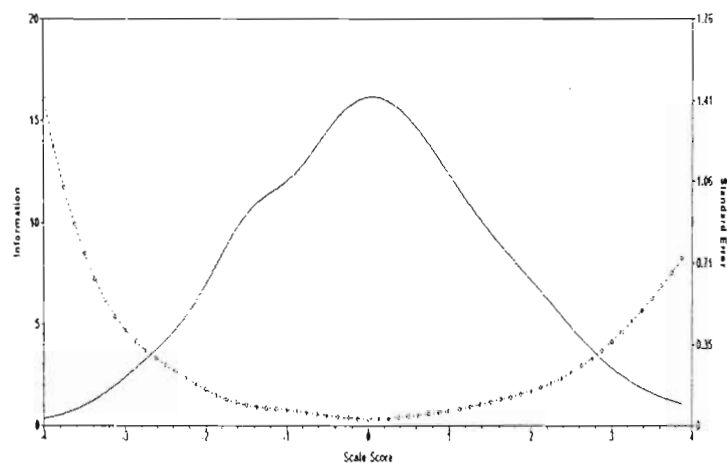
ตาราง 3.6 ค่าต่ำสุด ค่าสูงสุด ของค่าพารามิเตอร์ a, b และ c ของข้อสอบ MER 2003 ที่ผ่านการคัดเลือกจำแนกตามเนื้อหาการวัด

ที่	เนื้อหา	a		b		c	
		ค่าต่ำสุด	ค่าสูงสุด	ค่าต่ำสุด	ค่าสูงสุด	ค่าต่ำสุด	ค่าสูงสุด
1	การค้นหาคำรู้และวิธีการทางวิทยาศาสตร์	0.309	0.929	-2.088	2.270	0.001	0.225
2	ประเภทของการวิจัยทางสังคมศาสตร์	0.220	0.887	-1.840	2.843	0.001	0.223
3	การกำหนดประเด็นปัญหาและวัตถุประสงค์ในการวิจัย การทบทวนวรรณกรรม	0.234	1.578	-1.705	1.567	0.001	0.266
4	สมมติฐานการวิจัยและตัวแปรในการวิจัย	0.206	3.272	-1.883	2.660	0.001	0.315
5	การออกแบบการวิจัย	0.231	1.045	-1.218	2.276	0.001	0.183
6	เครื่องมือเก็บรวบรวมข้อมูลและการสร้างแบบทดสอบถาม การตรวจสอบคุณภาพเครื่องมือ	0.333	0.975	-1.159	1.901	0.001	0.298
7	การสุ่มตัวอย่าง การใช้สถิติในการวิจัย	0.217	1.445	-0.976	2.334	0.001	0.209
8	การเขียนข้อเสนอโครงการวิจัย และรายงานการวิจัย	0.239	1.209	-1.340	2.104	0.001	0.231

4) วิเคราะห์ค่าฟังก์ชันสารสนเทศของแบบทดสอบ

การวิเคราะห์คุณภาพของแบบทดสอบตามทฤษฎีการตอบข้อสอบแบบ 3 พารามิเตอร์จะใช้แบบแผนการตอบสนองแบบทดสอบเป็นรายชื่อในการประมาณค่าความสามารถของผู้สอบ ดังนั้นการประเมินคุณภาพของแบบทดสอบ จึงสามารถพิจารณาจากความถูกต้องแม่นยำในการประมาณค่าความสามารถของผู้ตอบ โดยใช้ดัชนีตัวหนึ่งเรียกว่า ค่าฟังก์ชันสารสนเทศของแบบสอบ [Test Information Function: TIF $I(\theta)$] สำหรับความคลาดเคลื่อนมาตรฐานของการประมาณค่า $[SE(\theta)]$ เป็นค่าเบี่ยงเบนมาตรฐานของการแจกแจงความน่าจะเป็นของค่าประมาณความสามารถที่แท้จริง (θ) ซึ่งเป็นค่าสัดส่วนผกผันกับความถูกต้องแม่นยำของการ

ประมาณค่าความสามารถหรือค่าฟังก์ชันสารสนเทศของแบบทดสอบ ดังนั้นถ้าค่าสารสนเทศของแบบทดสอบมีค่าสูงในช่วง θ ใดก็จะมีค่าความถูกต้องแม่นยำสูง ในการประมาณค่าความสามารถของผู้สอบในช่วง θ นั้นๆ โดยมีความคลาดเคลื่อนมาตรฐานของการประมาณค่าสำหรับค่าที่เหมาะสมสำหรับการคัดเลือกข้อสอบคือควรมีค่าความคลาดเคลื่อนมาตรฐานของการประมาณค่าไม่เกิน 0.50 (Hambleton and Swaminathan, 1991, p.103) นอกจากนี้ค่าความเที่ยงของการประมาณค่าความสามารถ (empirical reliability) เป็นค่าที่บ่งบอกถึงความน่าเชื่อถือของการประมาณค่าความสามารถของผู้สอบ (θ) โดยคำนวณได้จากอัตราส่วนของความแปรปรวนของคะแนนความสามารถที่ประมาณค่าได้กับผลรวมของความแปรปรวนของความแปรปรวนของคะแนนความสามารถที่ประมาณค่าได้กับความแปรปรวนของความคลาดเคลื่อน และหากนำค่าความเที่ยงของการประมาณค่ามายกกำลังสองและคูณด้วยร้อยแล้วเป็นเป็นค่าที่บ่งบอกถึงประสิทธิภาพของการประมาณค่าความสามารถของผู้สอบ (De Ayala, 2009, p.108; Toit, 2003, p.34) ผลการวิเคราะห์พบว่า แบบทดสอบ MER 2003 มีค่าฟังก์ชันสารสนเทศของแบบทดสอบสูงสุด เท่ากับ 16.80 ณ ตำแหน่งความสามารถผู้สอบ (θ) 0.00 นั่นคือแบบทดสอบฉบับนี้มีความแม่นยำในการประมาณค่าความสามารถของผู้สอบที่มีความสามารถในระดับปานกลาง ดังภาพ 3.2 รวมทั้งแบบทดสอบฉบับอ้างอิงมีค่าความเที่ยงของการประมาณค่าความสามารถ (empirical reliability) เท่ากับ 0.880 แสดงว่าแบบทดสอบฉบับอ้างอิงมีประสิทธิภาพในการประมาณค่าความสามารถของผู้สอบมีความสามารถในระดับปานกลางร้อยละ 77.44



ภาพ 3.2 ค่าฟังก์ชันสารสนเทศของแบบทดสอบ MER 2003

ขั้นตอนที่ 2 การสร้างคลังข้อสอบตามคุณสมบัติของแบบทดสอบฉบับอ้างอิงด้วยการจำลองข้อมูลการสอบด้วยโปรแกรม WINGEN

van der Linden (2005, p.22-23) กล่าวว่า การสร้างคลังข้อสอบตามทฤษฎีการตอบข้อสอบที่มีจุดมุ่งหมายเพื่อนำมาสร้างแบบทดสอบคู่ขนานมากกว่า 1 ฉบับที่ดีนั้นควรมีจำนวนข้อสอบมากกว่าหรือเท่ากับ 10 ข้อต่อ 1 เนื้อหา เพื่อให้ได้แบบทดสอบคู่ขนานมากกว่า 1 ฉบับที่มีค่าฟังก์ชันสารสนเทศของแบบทดสอบตรงกับแบบทดสอบฉบับอ้างอิง และมีความคลาดเคลื่อนในการประมาณค่าความสามารถของผู้สอบต่ำที่สุด ผู้วิจัยจึงดำเนินการสร้างคลังข้อสอบ ดังนี้

2.1 กำหนดสัดส่วนของคลังข้อสอบตามเนื้อหาการวัดโดยอิงจากแบบทดสอบฉบับอ้างอิง

การกำหนดสัดส่วนของคลังข้อสอบเป็นไปตามตารางการกำหนดลักษณะเฉพาะของแบบทดสอบ (table of specification) ของแบบทดสอบฉบับอ้างอิงร่วมกับเกณฑ์การสร้างจำนวนข้อสอบที่เหมาะสมกับการสร้างคลังข้อสอบตามทฤษฎีการตอบข้อสอบ คือ คลังข้อสอบในแต่ละเนื้อหาควรมีจำนวนข้อสอบอย่างน้อย 10 เท่าของจำนวนข้อสอบที่ใช้จริง ดังตาราง 3.7

ตาราง 3.7 สัดส่วนของคลังข้อสอบตามตารางการกำหนดลักษณะเฉพาะของแบบทดสอบฉบับอ้างอิง

ที่	เนื้อหา	จำนวนข้อสอบของ แบบทดสอบฉบับอ้างอิง	จำนวนข้อสอบของ คลังข้อสอบจากการจำลองข้อมูล
1	การค้นหาคำรู้และวิธีการทางวิทยาศาสตร์	5	110
2	ประเภทของการวิจัยทางสังคมศาสตร์	9	178
3	การกำหนดประเด็นปัญหาและวัตถุประสงค์ ในการวิจัย การทบทวนวรรณกรรม	5	167
4	สมมติฐานการวิจัยและตัวแปรในการวิจัย	21	1,053
5	การออกแบบการวิจัย	7	199
6	เครื่องมือเก็บรวบรวมข้อมูลและการสร้าง แบบทดสอบถาม การตรวจสอบคุณภาพ เครื่องมือ	4	292
7	การสุ่มตัวอย่าง การใช้สถิติในการวิจัย	6	167
8	การเขียนข้อเสนอโครงการวิจัย และ รายงานการวิจัย	7	323
	รวม	64	2,489

2.2 การจำลองข้อสอบด้วยโปรแกรม WINGEN

ในการวิจัยครั้งนี้ดำเนินการจำลองข้อสอบเพื่อนำมาสร้างคลังข้อสอบด้วยโปรแกรม WINGEN (Han, 2007) เพื่อให้ได้ช่วงของค่าพารามิเตอร์ข้อสอบ คือ ค่าอำนาจจำแนก (a parameter) ค่าความยาก (b parameter) และ ค่าการเดา (c parameter) สอดคล้องกับเนื้อหาการวัด 8 เนื้อหาของแบบทดสอบฉบับอ้างอิง โดยมีขั้นตอนการจำลองข้อมูลคลังข้อสอบตามทฤษฎีการตอบข้อสอบ ดังนี้

2.2.1 สร้างข้อกำหนดสำหรับการจำลองคลังข้อสอบเพื่อให้ได้ข้อสอบที่มีคุณลักษณะเฉพาะตามเนื้อหาการวัดตามแบบทดสอบเป้าหมายด้วยการกำหนดค่าพารามิเตอร์ a, b และ c ของคลังข้อสอบตามเนื้อหาการวัดโดยยึดตามค่าพารามิเตอร์ a, b และ c ของแบบทดสอบฉบับอ้างอิง ดังตาราง 3.8

ตาราง 3.8 ค่าต่ำสุด ค่าสูงสุดของค่าพารามิเตอร์ a, b และ c ของคลังข้อสอบตามเนื้อหาการวัดของแบบทดสอบฉบับอ้างอิง

ที่	เนื้อหา	a		b		c	
		min	max	min	Max	min	max
1	การค้นหาคำรู้และวิธีการทางวิทยาศาสตร์	0.309	0.929	-2.088	2.270	0.001	0.225
2	ประเภทของการวิจัยทางสังคมศาสตร์	0.220	0.887	-1.840	2.843	0.001	0.223
3	การกำหนดประเด็นปัญหาและวัตถุประสงค์ในการวิจัย การทบทวนวรรณกรรม	0.234	1.578	-1.705	1.567	0.001	0.266
4	สมมติฐานการวิจัยและตัวแปรในการวิจัย	0.206	3.272	-1.883	2.660	0.001	0.415
5	การออกแบบการวิจัย	0.231	1.045	-1.218	2.276	0.001	0.183
6	เครื่องมือเก็บรวบรวมข้อมูล และการสร้างแบบทดสอบถาม การตรวจสอบคุณภาพเครื่องมือ	0.333	0.975	-1.159	1.901	0.001	0.298
7	การสุ่มตัวอย่าง การใช้สถิติในการวิจัย	0.217	1.445	-0.976	2.334	0.001	0.209

ที่	เนื้อหา	a		b		c	
		min	max	min	Max	min	max
8	การเขียนข้อเสนอ โครงการวิจัย และรายงาน การวิจัย	0.239	1.209	-1.340	2.104	0.000	0.231

2.2.2 จำลองข้อมูลการสอบที่มีคำตอบแบบ 0 และ 1 ตามข้อกำหนดเพื่อนำมาสร้างคลังข้อสอบด้วยโปรแกรม WINGEN จำนวน 2,489 ข้อ จากจำนวนผู้สอบ 3,500 คน เพื่อให้ได้ช่วงของค่าพารามิเตอร์ข้อสอบ คือ ค่าอำนาจจำแนก (a parameter) ค่าความยาก (b parameter) และ ค่าการเดา (c parameter) ที่สอดคล้องกับเนื้อหาการวัด 8 เนื้อหาของแบบทดสอบฉบับอ้างอิง

2.3 การตรวจสอบและคัดเลือกข้อสอบเพื่อบรรจุในคลังข้อสอบที่มีคุณภาพตามทฤษฎีการตอบข้อสอบ

2.3.1 ตรวจสอบค่าพารามิเตอร์ a, b และ c ของคลังข้อสอบตามเนื้อหา โดยยึดตามค่าพารามิเตอร์ a, b และ c ของแบบทดสอบฉบับอ้างอิง

2.3.2 คัดเลือกข้อสอบทั้งหมดเพื่อนำไปบรรจุในคลังข้อสอบ จากคลังข้อสอบจำลองทั้งหมด 2,489 ข้อ พบว่า จำนวนคลังข้อสอบที่มีค่าพารามิเตอร์ a, b, c สอดคล้องกับแบบทดสอบฉบับอ้างอิง มีจำนวนทั้งสิ้น 2,089 ข้อ ซึ่งจำนวนคลังข้อสอบที่ใช้ได้ทั้ง 8 เนื้อหามีจำนวนข้อสอบมากกว่า 10 เท่าของข้อสอบที่ใช้จริง ดังตาราง 3.9

ตาราง 3.9 ความสอดคล้องระหว่างค่าพารามิเตอร์ของคลังข้อสอบกับแบบทดสอบฉบับอ้างอิง

ที่	เนื้อหา	ค่าพารามิเตอร์			ความ สอดคล้อง	จำนวน คลังข้อสอบ
			ฉบับอ้างอิง	คลังข้อสอบ		
1	การค้นหาคำรู้และวิธีการทาง วิทยาศาสตร์	a	min = 0.309 max = 0.929	min = 0.545 max = 0.929	✓	60
		b	min = -2.088 max = 2.270	min = -2.011 max = 2.267	✓	
		c	min = 0.001 max = 0.225	min = 0.092 max = 0.225	✓	
2	ประเภทของการวิจัยทาง สังคมศาสตร์	a	min = 0.220 max = 0.887	min = 0.426 max = 0.887	✓	128

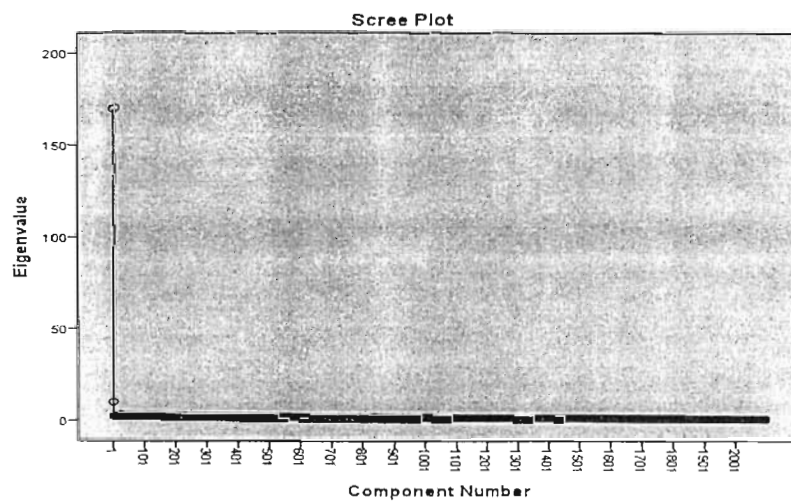
ที่	เนื้อหา	ค่าพารามิเตอร์			ความ สอดคล้อง	จำนวน คลังข้อสอบ
			ฉบับอ้างอิง	คลังข้อสอบ		
		b	min = -1.840 max = 2.843	min = -1.625 max = 2.630	✓	
		c	min = 0.001 max = 0.223	min = 0.099 max = 0.223	✓	
3	การกำหนดประเด็นปัญหาและ วัตถุประสงค์ในการวิจัย การทบทวน วรรณกรรม	a	min = 0.234 max = 1.578	min = 0.358 max = 1.578	✓	117
		b	min = -1.705 max = 1.567	min = -1.705 max = 1.333	✓	
		c	min = 0.001 max = 0.266	min = 0.073 max = 0.266	✓	
4	สมมติฐานการวิจัยและตัวแปรใน การวิจัย	a	min = 0.206 max = 3.272	min = 0.232 max = 1.903	✓	1,003
		b	min = -1.883 max = 2.660	min = -1.883 max = 2.660	✓	
		c	min = 0.001 max = 0.415	min = 0.095 max = 0.415	✓	
5	การออกแบบการวิจัย	a	min = 0.231 max = 1.045	min = 0.779 max = 1.045	✓	149
		b	min = -1.218 max = 2.276	min = -1.242 max = 2.201	✓	
		c	min = 0.001 max = 0.183	min = 0.072 max = 0.183	✓	
6	เครื่องมือเก็บรวบรวมข้อมูลและการ สร้างแบบทดสอบถาม การตรวจสอบคุณภาพเครื่องมือ	a	min = 0.333 max = 0.975	min = 0.239 max = 0.975	✓	242
		b	min = -1.159 max = 1.901	min = -1.159 max = 0.232	✓	

ที่	เนื้อหา	ค่าพารามิเตอร์			ความสอดคล้อง	จำนวนคลังข้อสอบ
			ฉบับอ้างอิง	คลังข้อสอบ		
6	เครื่องมือเก็บรวบรวมข้อมูลและการสร้างแบบทดสอบถามการตรวจสอบคุณภาพเครื่องมือ	c	min = 0.001 max = 0.298	min = 0.232 max = 0.298	✓	
7	การสุ่มตัวอย่าง การใช้สถิติในการวิจัย	a	min = 0.217 max = 1.445	min = 0.471 max = 1.445	✓	117
		b	min = -0.976 max = 2.334	min = -1.407 max = 2.334	✓	
		c	min = 0.001 max = 0.209	min = 0.095 max = 0.209	✓	
8	การเขียนข้อเสนอโครงการวิจัย และรายงานการวิจัย	a	min = 0.239 max = 1.209	min = 1.209 max = 1.209	✓	273
8	การเขียนข้อเสนอโครงการวิจัย และรายงานการวิจัย	b	min = -1.340 max = 2.104	min = -1.340 max = 2.104	✓	
		c	min = 0.000 max = 0.231	min = 0.163 max = 0.231	✓	

2.3.3 ตรวจสอบความเป็นเอกมิติของคลังข้อสอบที่ผ่านการคัดเลือก พบว่าข้อสอบทั้งหมด 2,089 ข้อมีความเป็นเอกมิติโดยพิจารณาจากค่า eigen ที่ 1 ต่างจากที่เหลือมากๆ และค่า eigen ที่เหลือมีลักษณะเหมือนกัน รวมทั้งอัตราส่วนระหว่างค่า eigen ที่ 1 กับค่า eigen ที่ 2 เท่ากับ 16.45 ซึ่งอัตราส่วนดังกล่าวมีค่ามากกว่า 3.00 จากข้อมูลข้างต้นแสดงให้เห็นว่าคลังข้อสอบที่ผ่านการคัดเลือก จำนวน 2,089 ข้อ วัตถุคุณลักษณะเด่นเดี่ยว (single dominant factor) ดังตาราง 3.10 และภาพ 3.3

ตาราง 3.10 ค่าไอเกน 10 องค์ประกอบแรกของคลังข้อสอบที่ผ่านการคัดเลือก 2,089 ข้อ

องค์ประกอบ	ค่าไอเกน		
	Total	% of Variance	Cumulative %
1	169.953	8.136	8.136
2	10.328	0.494	8.630
3	2.594	0.124	8.754
4	2.581	0.124	8.878
5	2.571	0.123	9.001
6	2.568	0.123	9.124
7	2.556	0.122	9.246
8	2.551	0.122	9.368
9	2.546	0.122	9.490
10	2.537	0.121	9.611



ภาพ 3.3 ค่าไอเกนของคลังข้อสอบที่ผ่านการคัดเลือก

สถิติที่ใช้ในการวิเคราะห์ข้อมูล

ในการวิจัยครั้งนี้ใช้สถิติสำหรับการวิเคราะห์ข้อมูล จำแนกเป็น 3 กลุ่ม ดังนี้

1. การวิเคราะห์คุณภาพของข้อสอบและแบบทดสอบตามทฤษฎีการทดสอบแบบดั้งเดิม

1.1 วิเคราะห์ดัชนีความสอดคล้องระหว่างข้อสอบกับวัตถุประสงค์ในการวัด (Index of Item-

Objective Congruence: IOC) ด้วยสูตรของ Hambleton (as cited in Turner & Carlson, 2003, p.167)

1.2 วิเคราะห์ค่าความยากของข้อสอบใช้ค่าสัดส่วนในการตอบข้อสอบแต่ละข้อถูก

1.3 วิเคราะห์อำนาจจำแนกของข้อสอบด้วยค่าสหสัมพันธ์แบบไบซีเรียล (Biserial Correlation)

1.4 วิเคราะห์ค่าสัมประสิทธิ์ความเที่ยงแบบความสอดคล้องในด้วยวิธีการของ Kuder Richardson

Formula 20

2. การวิเคราะห์ค่าพารามิเตอร์ของข้อสอบและแบบทดสอบตามทฤษฎีการตอบข้อสอบ

2.1 ตรวจสอบความเป็นเอกมิติแบบเด่นเดี่ยวของแบบทดสอบด้วยการวิเคราะห์องค์ประกอบเชิงสำรวจ (Exploratory Factor Analysis)

2.2 ประเมินความเหมาะสมของโมเดลการตอบสนองข้อสอบกับข้อมูลเชิงประจักษ์ด้วยการประมาณค่า -2 likelihood และทดสอบความเหมาะสมของโมเดลด้วย χ^2 - test

2.2 วิเคราะห์ค่าพารามิเตอร์ของข้อสอบและแบบทดสอบตามทฤษฎีการตอบข้อสอบแบบ

3 พารามิเตอร์

3. การประเมินคุณลักษณะเฉพาะของความเป็นคู่ขนาน (Test Parallelism) จากการพิจารณา Mean Square Deviation of Test Information Function ระหว่างแบบทดสอบคู่ขนานกับแบบทดสอบฉบับอ้างอิง ควรมีค่าน้อยกว่าหรือเท่ากับ 0.05 ของ Luecht (1998) และ Chen , P.-H., Chang, H-H. & Wu, H. (2012)

4. การวิเคราะห์อัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบเฉลี่ย (Test Overlap rate) ของ Chen, Ankenmaun & Spray (2003)

บทที่ 4

ผลการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์หลักเพื่อสร้างแบบทดสอบคู่ขนานจากคลังข้อสอบที่มีคุณภาพด้านความเที่ยงและความตรงตามทฤษฎีการตอบข้อสอบ โดยมีวัตถุประสงค์ย่อย คือ 1) เพื่อสร้างโปรแกรมการคัดเลือกข้อสอบจากคลังข้อสอบเพื่อสร้างแบบทดสอบคู่ขนานที่มีคุณภาพด้านความเที่ยง (reliability) และความตรง (validity) ตามทฤษฎีการตอบข้อสอบ 2) เพื่อตรวจสอบความถูกต้องของโปรแกรมการคัดเลือกข้อสอบด้านคุณลักษณะเฉพาะของความเป็นคู่ขนาน (parallelism specification) ของแบบทดสอบตามตำแหน่งความสามารถของผู้สอบของแบบทดสอบฉบับอ้างอิง และ 3) เพื่อตรวจสอบความถูกต้องของโปรแกรมการคัดเลือกข้อสอบด้านการใช้ข้อสอบซ้ำระหว่างแบบทดสอบคู่ขนาน (test overlap) ตามตำแหน่งความสามารถของผู้สอบของแบบทดสอบฉบับอ้างอิง ผู้วิจัยขอเสนอผลการวิเคราะห์ข้อมูล ดังนี้

ตอนที่ 4.1 โปรแกรมการสร้างแบบทดสอบคู่ขนาน และผลการคัดเลือกข้อสอบจากคลังข้อสอบด้วยโปรแกรมเพื่อสร้างแบบทดสอบคู่ขนานมีคุณภาพด้านความเที่ยง (reliability) และความตรง (validity) ตามทฤษฎีการตอบข้อสอบ

ตอนที่ 4.2 ผลการตรวจสอบความถูกต้องของโปรแกรมการคัดเลือกข้อสอบด้านคุณลักษณะเฉพาะของความเป็นคู่ขนาน (parallelism specification) ของแบบทดสอบตามตำแหน่งความสามารถของผู้สอบของแบบทดสอบฉบับอ้างอิง

ตอนที่ 4.3 ผลตรวจสอบความถูกต้องของโปรแกรมการคัดเลือกข้อสอบด้านการใช้ข้อสอบซ้ำระหว่างแบบทดสอบคู่ขนาน (test overlap) ตามตำแหน่งความสามารถของผู้สอบของแบบทดสอบฉบับอ้างอิง ดังรายละเอียด

ตอนที่ 4.1 โปรแกรมการสร้างแบบทดสอบคู่ขนาน และผลการคัดเลือกข้อสอบจากคลังข้อสอบด้วย
โปรแกรมการคัดเลือกข้อสอบเพื่อสร้างแบบทดสอบคู่ขนานที่มีคุณภาพด้านความเที่ยง
(reliability) และความตรง (validity) ตามทฤษฎีการตอบข้อสอบ

4.1.1 โปรแกรมการสร้างแบบทดสอบคู่ขนาน

โปรแกรมการสร้างข้อสอบคู่ขนานผู้วิจัยได้ประยุกต์จากชุดคำสั่ง Ip_Solve Version 5.5 ในโปรแกรม
R Studio เพื่อคัดเลือกข้อสอบจากคลังข้อสอบที่สร้างขึ้นจากงานวิจัยครั้งนี้ โดยข้อสอบที่ผ่าน
การคัดเลือกจะถูกนำไปสร้างเป็นแบบทดสอบคู่ขนานมากกว่า 1 ฉบับที่มีคุณสมบัติด้านความเป็น
คุณลักษณะเฉพาะของความเป็นคู่ขนาน (test parallelism) ด้านความเที่ยง และด้านความตรงตามเนื้อหาที่
ใกล้เคียงหรือตรงกับแบบทดสอบฉบับอ้างอิงมากที่สุด รวมทั้งแบบทดสอบคู่ขนานมากกว่า 1 ฉบับมีอัตรา
การใช้ข้อสอบซ้ำระหว่างแบบทดสอบอยู่ในระดับต่ำ

ส่วนประกอบของโปรแกรมการคัดเลือกข้อสอบ คือ

ส่วนที่ 1 การกำหนดปัจจัยนำเข้า ประกอบด้วย

1.1 การกำหนดคุณลักษณะเฉพาะของแบบทดสอบคู่ขนานตามคุณลักษณะเฉพาะของแบบทดสอบ
ฉบับอ้างอิง ประกอบด้วย โครงสร้างเนื้อหาในการวัดของแบบทดสอบ จำนวนแบบทดสอบคู่ขนานที่
ต้องการสร้าง ความยาวของแบบทดสอบ

1.2 การกำหนดค่าฟังก์ชันสารสนเทศของแบบทดสอบคู่ขนานตามค่าฟังก์ชันสารสนเทศของ
แบบทดสอบฉบับอ้างอิง ณ ตำแหน่งความสามารถของผู้สอบที่ใช้ในการวิจัยครั้งนี้

ส่วนที่ 2 ชุดคำสั่งของขั้นตอนวิธีและเงื่อนไขบังคับ ประกอบด้วย

2.1 ชุดคำสั่งการตัดสินใจคัดเลือกข้อสอบจากคลังข้อสอบเพื่อสร้างเป็นแบบทดสอบคู่ขนานที่แสดงใน
ลักษณะตัวแปรสุ่มที่มีการแจกแจงแบบทวินาม คือ

$$X_{it} = \begin{cases} 1 & \text{เมื่อข้อสอบที่ } i \text{ ถูกจัดอยู่ในแบบสอบ } t \\ 0 & \text{เมื่อข้อสอบที่ } i \text{ ไม่ถูกจัดอยู่ในแบบสอบ } t \end{cases}$$

2.2 ชุดคำสั่งเงื่อนไขบังคับของชุดคำสั่ง IpSolveAPI เพื่อสร้างเป็นแบบทดสอบคู่ขนานจากคลังข้อสอบ
ที่มีค่าฟังก์ชันสารสนเทศของแบบทดสอบตามตำแหน่งความสามารถของผู้สอบ (Test Information
Function: TIF) ที่มีค่าใกล้เคียงหรือมีความแตกต่างน้อยที่สุดจากค่าฟังก์ชันสารสนเทศของแบบทดสอบตาม
ตำแหน่งความสามารถของผู้สอบของแบบทดสอบฉบับอ้างอิง

2.3 ชุดคำสั่งเงื่อนไขบังคับการสุ่มข้อสอบที่ไม่ให้มีการใช้ข้อสอบซ้ำระหว่างฉบับ

2.4 ชุดคำสั่งขั้นตอนวิธีการสุ่มข้อสอบเพื่อสร้างเป็นแบบทดสอบคู่ขนานมากกว่า 1 ฉบับตาม
โครงสร้างเนื้อหาการวัด และค่าฟังก์ชันสารสนเทศของแบบทดสอบที่กำหนด

2.5 ชุดคำสั่งขั้นตอนวิธีการแสดงผลการคัดเลือกข้อสอบในรูปแบบตาราง

- 1) ค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง 2 ตำแหน่ง คือ
 - 1.1) $RTIF_{\theta_1} = 0.00$ เท่ากับ 16.80
 - 1.2) $RTIF_{\theta_2} = 1.50$ เท่ากับ 10.00
- 2) ค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง 3 ตำแหน่ง คือ
 - 2.1) $RTIF_{\theta_1} = -1.50$ เท่ากับ 10.50
 - 2.2) $RTIF_{\theta_2} = 0.00$ เท่ากับ 16.80
 - 2.3) $RTIF_{\theta_3} = 1.50$ เท่ากับ 10.00
- 3) ค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง 4 ตำแหน่ง คือ
 - 3.1) $RTIF_{\theta_1} = -1.50$ เท่ากับ 10.50
 - 3.2) $RTIF_{\theta_2} = 0.00$ เท่ากับ 16.08
 - 3.3) $RTIF_{\theta_3} = 1.50$ เท่ากับ 10.00
 - 3.4) $RTIF_{\theta_4} = 2.50$ เท่ากับ 5.00

ส่วนที่ 2 ขั้นตอนวิธีและเงื่อนไขบังคับ

2.1 ชุดคำสั่งการตัดสินใจคัดเลือกข้อสอบจากคลังข้อสอบเพื่อสร้างเป็นแบบทดสอบคู่ขนานที่แสดงในลักษณะตัวแปรสุ่มที่มีการแจกแจงแบบทวินาม คือ

$$x_{it} = \begin{cases} 1 & \text{เมื่อข้อสอบที่ } i \text{ ถูกจัดอยู่ในแบบสอบ } t \\ 0 & \text{เมื่อข้อสอบที่ } i \text{ ไม่ถูกจัดอยู่ในแบบสอบ } t \end{cases}$$

2.2 ชุดคำสั่งเงื่อนไขบังคับของชุดคำสั่ง IpSolveAPI สำหรับการคัดเลือกข้อสอบจากคลังข้อสอบแล้วทำการคำนวณค่าฟังก์ชันสารสนเทศของข้อสอบ (Item Information Function: IIF) หลังจากนั้นคำนวณผลรวมของค่า IIF ที่ถูกคัดเลือกซึ่งแสดงถึงค่าฟังก์ชันสารสนเทศของแบบทดสอบ (Test Information Function: TIF) และค่า TIF ของ k ตำแหน่งความสามารถของผู้สอบ (θ_k) เมื่อ $k = 1, \dots, K$ เพื่อนำมาสร้างเป็นแบบทดสอบคู่ขนานให้ได้ค่าฟังก์ชันสารสนเทศของแบบทดสอบตามตำแหน่งความสามารถของผู้สอบ (Test Information Function: TIF) ที่มีค่าใกล้เคียงหรือมีความแตกต่างน้อยที่สุดจากค่าฟังก์ชันสารสนเทศของแบบทดสอบตามตำแหน่งความสามารถของผู้สอบของแบบทดสอบฉบับอ้างอิง (Referenced Test Information Function: RTIF)

Set control parameters: minimization problem; integer tolerance is set to 0.1;

absolute MIP gap is set to 0.1; relative MIP gap is set to 0.05

lp.control(lpprec,sense="min",epsint=0.1,mip.gap=c(0.1,0.05));

2.3 ชุดคำสั่งเงื่อนไขบังคับการสุ่มข้อสอบที่ไม่ให้มีการใช้ข้อสอบซ้ำระหว่างฉบับ

for (k in 1:I){

```
add.constraint(lprec,rep(1,2), "<=", 1, indices=c(k,I+k))
```

2.4 ชุดคำสั่งขั้นตอนวิธีการสุ่มข้อสอบเพื่อสร้างเป็นแบบทดสอบอยู่ช้านานมากกว่า 1 ฉบับตามโครงสร้างเนื้อหาการวัด

```
Nc = c(5,9,5,21,7,4,10,3)
```

```
# Form 1
```

```
for(k in 1:8){
```

```
  add.constraint(lprec,rep(1,length(Vc[[k]])), ">=", Nc[k], indices=Vc[[k]])
```

```
}
```

```
# Form 2 to 30 (value 1)
```

```
for (f in 1:Fnew-1) {
```

```
  for(k in 1:8){
```

```
    add.constraint(lprec,rep(1,length(Vc[[k]])), ">=", Nc[k], indices=(I*f)+Vc[[k]])
```

```
  }
```

```
}
```

```
add.constraint(lprec, rep(1,I), "=", N, indices = 1:I)
```

```
# Form 2 to 30 (0)
```

```
for(f in 1:Fnew-1) {
```

```
  add.constraint(lprec, rep(1,I), "=", N, indices = ((I*f)+1):((f+1)*I))
```

```
}
```

```
d_theta=c(18.5718, 5.782)
```

```
for(k in 1:2)
```

```
{
```

```
  add.constraint(lprec,c(Info[,k],-1), "<=", d_theta[k], indices=c(1:I,M))
```

```
  add.constraint(lprec,c(Info[,k],-1), "<=", d_theta[k], indices=c((I+1):(2*I),M))
```

```
  add.constraint(lprec,c(Info[,k],1), ">=", d_theta[k], indices=c(1:I,M))
```

```
  add.constraint(lprec,c(Info[,k],1), ">=", d_theta[k], indices=c((I+1):(2*I),M))
```

```
}
```

```
for(f in 2:Fnew-1)
```

```
{
```

```
  for(k in 1:2)
```

```
{
```



```

#add.constraint(lprec,c(Info[,k],(f-1)),"<=",d_theta[k],indices=c(((I*(f-1))+1):(I*f),M))
add.constraint(lprec,c(Info[,k],(f-1)),"<=",d_theta[k],indices=c(((I*f)+1):(((f+1)*I)),M))
#add.constraint(lprec,c(Info[,k],f),">=",d_theta[k],indices=c(((I*(f-1))+1):(I*f),M))
add.constraint(lprec,c(Info[,k],f),">=",d_theta[k],indices=c(((I*f)+1):(((f+1)*I)),M))
}
}

# Objective function
set.objfn(lprec,l,indices=M)

# Solve the model
res_flag = solve(lprec)

# The integer value containing the status code, for example, 0: "optimal solution found"
res_flag

# Retrieve the values of the decision variables
x_opt = get.variables(lprec)

2.5 ชุดคำสั่งขั้นตอนวิธีการแสดงผลการคัดเลือกข้อสอบในรูปแบบตาราง
write.csv(x_opt,file="c://2560_suwimon/Folder_2T3F/N2089_opt_2T3F.csv")
#theta <- seq(-4.00, 4.00, by=0.25)
write.csv(Info,file="c://2560_suwimon/Folder_2T3F/N2089_opt_2T3F_ATA.csv")
# Sum V1 V2 V3 3.175, 5.864, 18,837
x_opt_New = array(0,c(I,Fnew))
i_opt = 0
for(j in 1:Fnew){
  for(i in 1:I){
    i_opt <- i_opt+1
    x_opt_New[i,j] = x_opt[i_opt]
  }
}
write.csv(x_opt_New,file="c://2560_suwimon/Folder_2T3F/N2089_opt_2T3F_Select.csv")
#TIF.theta <- seq(1, 33, by=1)
TIF.theta = array(0,c(1,J))
for(i in 1:J){

```

```

TIF.theta[i] = sum(Info[i], na.rm = FALSE)
}

plot(TIF.theta[,1:J],lty=2,lwd=2,col="red")

TIF.theta1 = array(0,c(F,J))

for(f in 1:Fnew)
{
  Info_Theta = array(0,c(I,J))
  for(i in 1:I)
  {
    if(x_opt_New[i,f]==1)
    {
      Info_Theta[i,1] = Info[i,1]
      Info_Theta[i,2] = Info[i,2]
      #Info_Theta[i,3] = Info[i,3]
      #Info_Theta[i,4] = Info[i,4]
    }
  }
  for(j in 1:J){
    TIF.theta1[f,j] = sum(Info_Theta[,j], na.rm = FALSE)
  }
}

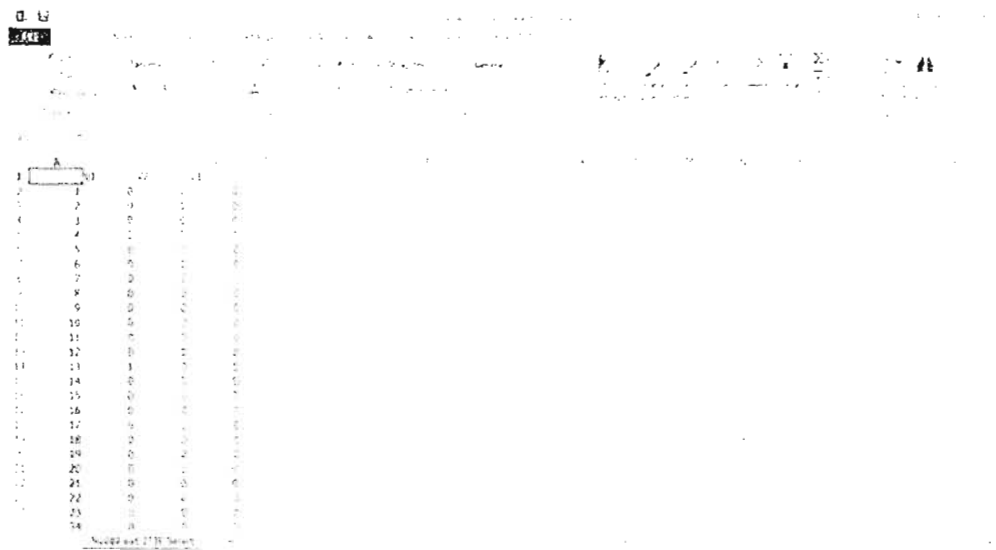
write.csv(TIF.theta1,file="c://2560_suwimon/Folder_2T3F/N2089_opt_2T3F_Theta.csv")

```

ส่วนที่ 3 ผลลัพธ์ของขั้นตอนวิธีและเงื่อนไขบังคับ

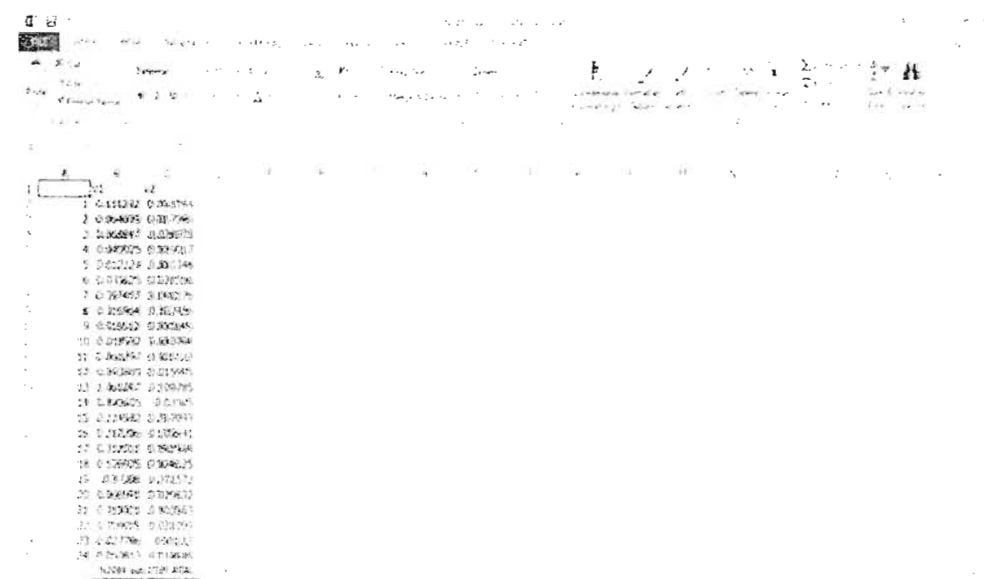
ผลลัพธ์ของขั้นตอนวิธีและเงื่อนไขบังคับที่อยู่ในรูปตาราง ประกอบด้วย

3.1 ผลลัพธ์การคัดเลือกข้อสอบจากคลังข้อสอบเพื่อสร้างเป็นแบบทดสอบคู่ขนานมากกว่า 1 ฉบับ โดยที่ค่าแสดง เท่ากับ 0 หมายถึง เป็นข้อสอบไม่ถูกคัดเลือกจากคลังข้อสอบตามเงื่อนไขบังคับ ส่วนค่าแสดง เท่ากับ 1 หมายถึง เป็นข้อสอบที่ถูกคัดเลือกจากคลังข้อสอบตามเงื่อนไขบังคับ แสดงดังภาพ 4.2



ภาพ 4.2 ผลลัพธ์การคัดเลือกข้อสอบจากคลังข้อสอบเพื่อสร้างเป็นแบบทดสอบคู่ขนานมากกว่า 1 ฉบับ

3.2 ผลลัพธ์ของแบบทดสอบคู่ขนานที่มีค่าฟังก์ชันสารสนเทศของแบบทดสอบตามตำแหน่งความสามารถของผู้สอบ (Test Information Function: TIF) ที่มีค่าใกล้เคียงหรือมีความแตกต่างกันน้อยที่สุดจากค่าฟังก์ชันสารสนเทศของแบบทดสอบตามตำแหน่งความสามารถของผู้สอบของแบบทดสอบฉบับอ้างอิง (Referenced Test Information Function: RTIF) แสดงดังภาพ 4.3



ภาพ 4.3 ผลลัพธ์ของค่าฟังก์ชันสารสนเทศของแบบทดสอบตามตำแหน่งความสามารถของผู้สอบ

4.1.2 ผลการคัดเลือกข้อสอบจากคลังข้อสอบด้วยโปรแกรมเพื่อสร้างแบบทดสอบคู่ขนานที่มีคุณภาพด้านความเที่ยง (reliability) และความตรง (validity) ตามทฤษฎีการตอบข้อสอบ

โปรแกรมการสร้างแบบทดสอบคู่ขนานในการวิจัยครั้งนี้ได้ดำเนินการสร้างแบบทดสอบที่มีคุณภาพด้านความเที่ยงและความตรงตามทฤษฎีการตอบข้อสอบด้วยการกำหนดเงื่อนไขบังคับ 2 เงื่อนไข คือ (1) จำนวนแบบทดสอบที่ต้องการสร้างเริ่มต้นตั้งแต่ 1 ฉบับ ถึง 5 ฉบับ และ (2) ค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง เริ่มต้นตั้งแต่ 2, 3 และ 4 ตำแหน่งความสามารถ ดังนั้นผู้วิจัยขอเสนอผลการคัดเลือกข้อสอบจากคลังข้อสอบ ดังนี้

1) ผลการคัดเลือกข้อสอบด้วยโปรแกรมเพื่อสร้างแบบทดสอบที่มีค่าความเที่ยงตามทฤษฎีการตอบข้อสอบ

สำหรับผลการคัดเลือกข้อสอบด้วยโปรแกรมเพื่อสร้างเป็นแบบทดสอบที่มีค่าความเที่ยงตามทฤษฎีการตอบข้อสอบนั้นเป็นการพิจารณาจากค่าฟังก์ชันสารสนเทศของแบบทดสอบที่ได้จากโปรแกรมมีค่าฟังก์ชันสารสนเทศสูงสุด ณ ตำแหน่งความสามารถที่ใกล้เคียงหรือตรงกับค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง ผลการวิจัยในส่วนนี้สามารถนำเสนอเป็น 3 กรณี คือ

กรณีที่ 1 กำหนดเงื่อนไขค่าฟังก์ชันสารสนเทศของแบบทดสอบของแบบทดสอบฉบับอ้างอิง ณ ตำแหน่งความสามารถของผู้สอบ 2 ตำแหน่ง พบว่ามีแบบทดสอบ จำนวน 4 ฉบับที่มีค่าฟังก์ชันสารสนเทศของแบบทดสอบสูงสุด ณ ตำแหน่งความสามารถ ($\theta = 0.00$) ที่ใกล้เคียงกับฉบับอ้างอิง

กรณีที่ 2 กำหนดเงื่อนไขค่าฟังก์ชันสารสนเทศของแบบทดสอบของแบบทดสอบฉบับอ้างอิง ณ ตำแหน่งความสามารถของผู้สอบ 3 ตำแหน่ง พบว่ามีแบบทดสอบ จำนวน 2 ฉบับที่มีค่าฟังก์ชันสารสนเทศสูงสุด ณ ตำแหน่งความสามารถ ($\theta = 0.00$) ที่ใกล้เคียงกับฉบับอ้างอิง

กรณีที่ 3 กำหนดเงื่อนไขค่าฟังก์ชันสารสนเทศของแบบทดสอบของแบบทดสอบฉบับอ้างอิง ณ ตำแหน่งความสามารถของผู้สอบ 4 ตำแหน่ง พบว่ามีแบบทดสอบ จำนวน 1 ฉบับที่มีค่าฟังก์ชันสารสนเทศสูงสุด ณ ตำแหน่งความสามารถ ($\theta = 0.00$) ที่ใกล้เคียงกับฉบับอ้างอิง

ดังรายละเอียด

กรณีที่ 1 กำหนดเงื่อนไขค่าฟังก์ชันสารสนเทศของแบบทดสอบของแบบทดสอบฉบับอ้างอิง ณ ตำแหน่งความสามารถของผู้สอบ 2 ตำแหน่ง

เมื่อโปรแกรมการคัดเลือกข้อสอบเพื่อสร้างเป็นแบบทดสอบกำหนดจำนวนแบบทดสอบที่ต้องการสร้างทั้งหมด เท่ากับ 5 ฉบับ และกำหนดเงื่อนไขของค่าฟังก์ชันสารสนเทศของแบบทดสอบตามฉบับอ้างอิง ณ 2 ตำแหน่งความสามารถของผู้สอบ ผลการศึกษาพบว่า มีแบบทดสอบ จำนวน 4 ฉบับเท่านั้นที่มีค่าฟังก์ชันสารสนเทศสูงสุด ณ ตำแหน่งความสามารถที่ใกล้เคียงกับฉบับอ้างอิง โดยที่แบบทดสอบฉบับอ้างอิงมีค่าฟังก์ชันสารสนเทศของแบบทดสอบสูงสุด เท่ากับ 16.80 ณ ตำแหน่งความสามารถผู้สอบ 0.00 มีค่าความคลาดเคลื่อนมาตรฐานของการประมาณค่า 0.10 นั่นคือแบบทดสอบฉบับนี้มีความแม่นยำในการ

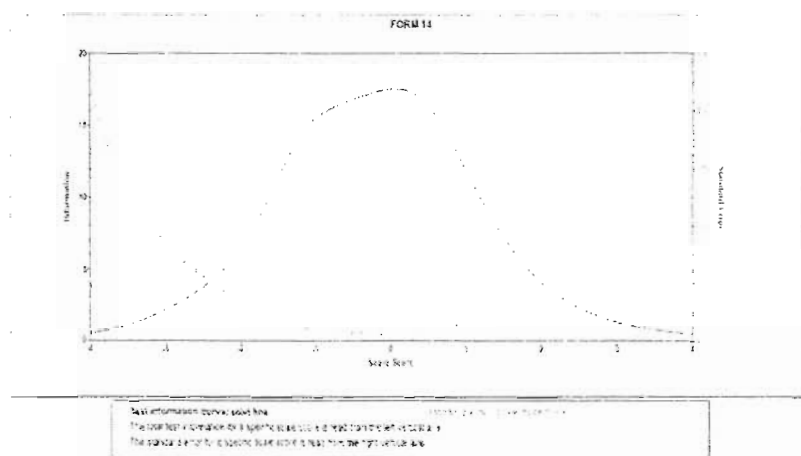
ประมาณค่าความสามารถของผู้สอบที่มีความสามารถในระดับปานกลาง สำหรับแบบทดสอบฉบับที่ 1 มีค่าฟังก์ชันสารสนเทศของแบบทดสอบสูงสุด เท่ากับ 16.85 ณ ตำแหน่งความสามารถผู้สอบ 0.00 มีค่าความคลาดเคลื่อนมาตรฐานของการประมาณค่า 0.10 แบบทดสอบฉบับที่ 2 มีค่าฟังก์ชันสารสนเทศของแบบทดสอบสูงสุด เท่ากับ 16.80 ณ ตำแหน่งความสามารถผู้สอบ 0.00 มีค่าความคลาดเคลื่อนมาตรฐานของการประมาณค่า 0.10 แบบทดสอบฉบับที่ 3 มีค่าฟังก์ชันสารสนเทศของแบบทดสอบสูงสุด เท่ากับ 17.00 ณ ตำแหน่งความสามารถผู้สอบ 0.00 มีค่าความคลาดเคลื่อนมาตรฐานของการประมาณค่า 0.10 และแบบทดสอบฉบับที่ 4 มีค่าฟังก์ชันสารสนเทศของแบบทดสอบสูงสุด เท่ากับ 17.03 ณ ตำแหน่งความสามารถผู้สอบ 0.00 มีค่าความคลาดเคลื่อนมาตรฐานของการประมาณค่า 0.10

จากการวิเคราะห์ค่าฟังก์ชันสารสนเทศของแบบทดสอบ จำนวน 4 ฉบับที่มีค่าใกล้เคียงหรือตรงกับแบบทดสอบฉบับอ้างอิง แสดงดังตาราง 4.2 และภาพ 4.5 – 4.6

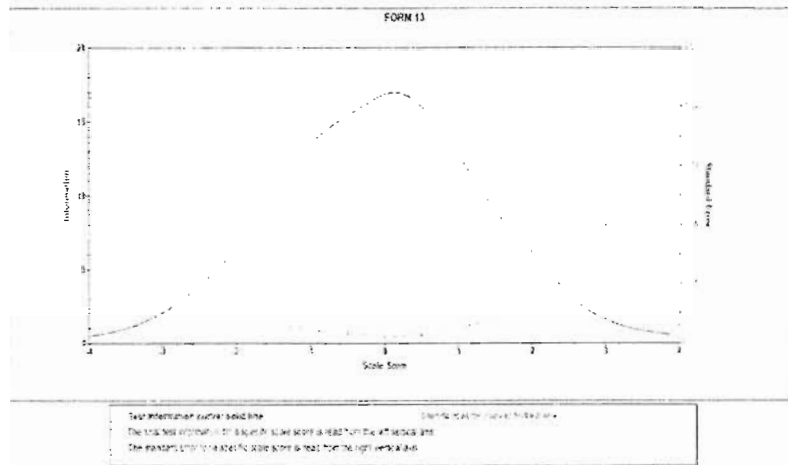
ตาราง 4.2 ค่าฟังก์ชันสารสนเทศของแบบทดสอบและค่าความคลาดเคลื่อนมาตรฐานของการประมาณค่าของแบบทดสอบฉบับอ้างอิงและแบบทดสอบ 4 ฉบับที่มีค่าใกล้เคียงหรือตรงกับแบบทดสอบฉบับอ้างอิง เมื่อกำหนดเงื่อนไขค่าฟังก์ชันสารสนเทศของแบบทดสอบของแบบทดสอบฉบับอ้างอิง ณ 2 ตำแหน่งความสามารถของผู้สอบ

ค่าฟังก์ชันสารสนเทศ ของแบบทดสอบฉบับอ้างอิง	แบบทดสอบฉบับ ที่	ค่าฟังก์ชันสารสนเทศ ของแบบทดสอบ	ค่าความคลาดเคลื่อน มาตรฐาน
16.80	1	16.85	0.10
	2	16.80	0.10
	3	17.00	0.10
	4	17.03	0.10

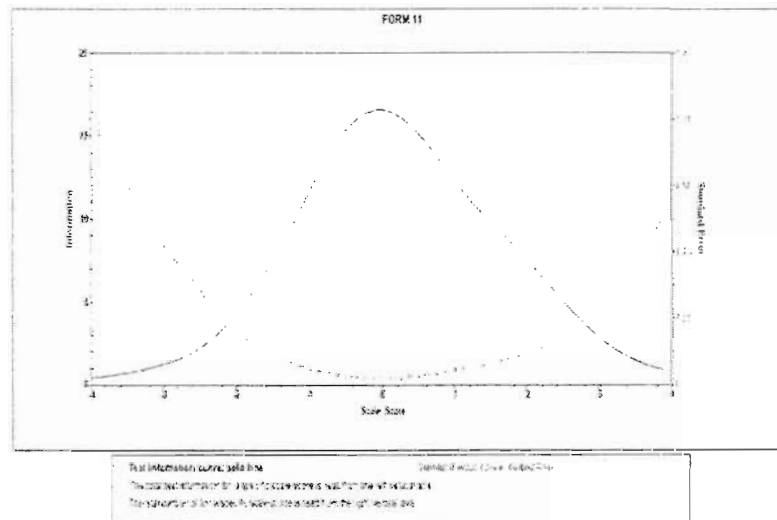
แบบทดสอบฉบับอ้างอิง



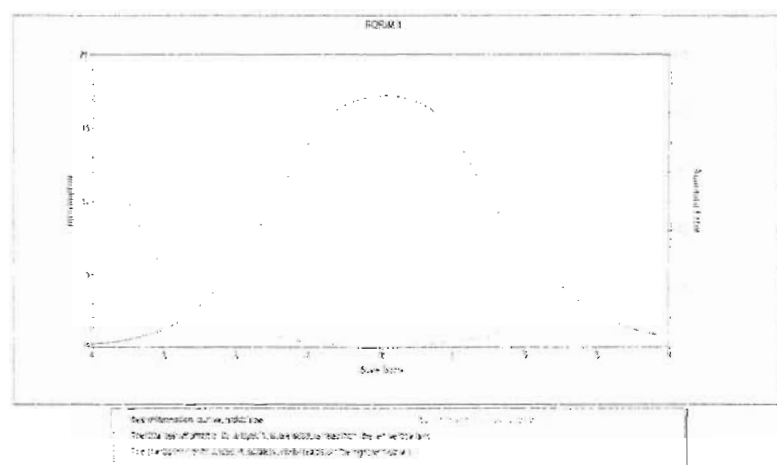
แบบทดสอบฉบับที่ 1

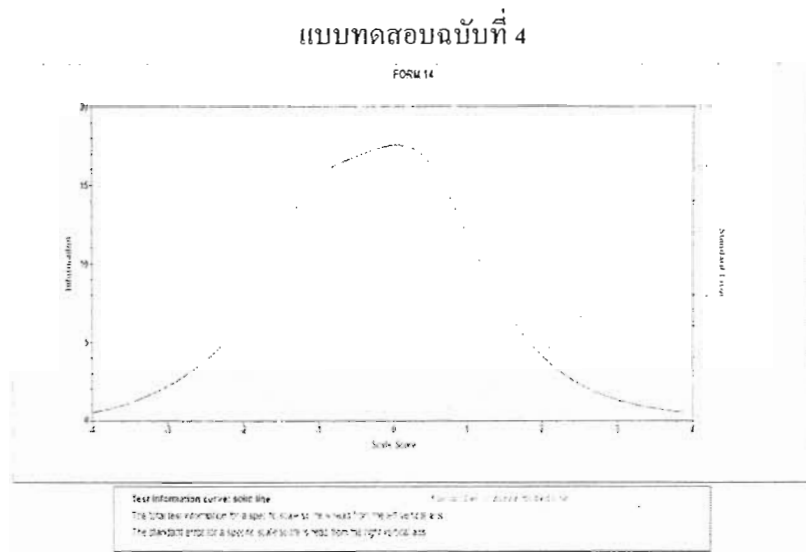


แบบทดสอบฉบับที่ 2



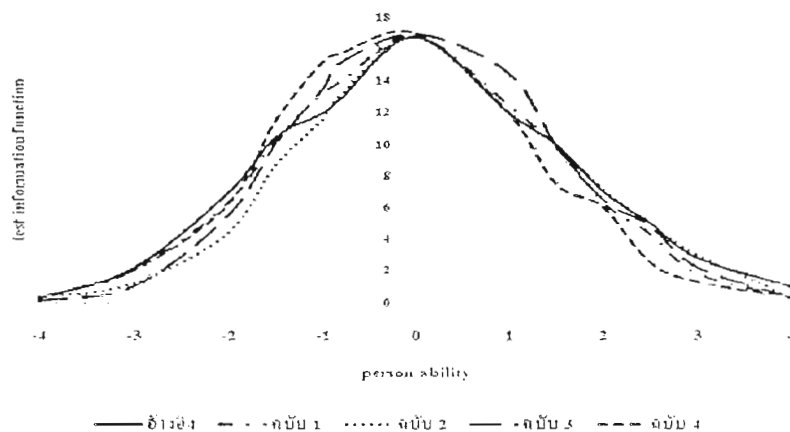
แบบทดสอบฉบับที่ 3





ภาพ 4.5 ค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง และแบบทดสอบที่สร้างจากโปรแกรม
การคัดเลือกข้อสอบ จำนวน 4 ฉบับ เมื่อกำหนดกำหนดเงื่อนไข

ค่าฟังก์ชันสารสนเทศของแบบทดสอบของแบบทดสอบฉบับอ้างอิง ณ 2 ตำแหน่งความสามารถ



ภาพ 4.6 ค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง และแบบทดสอบที่สร้างจากโปรแกรม
การคัดเลือกข้อสอบ จำนวน 4 ฉบับ เมื่อกำหนดกำหนดเงื่อนไข

ค่าฟังก์ชันสารสนเทศของแบบทดสอบของแบบทดสอบฉบับอ้างอิง ณ 2 ตำแหน่งความสามารถ

กรณีที่ 2 กำหนดเงื่อนไขค่าฟังก์ชันสารสนเทศของแบบทดสอบของแบบทดสอบฉบับอ้างอิง ณ ตำแหน่ง
ความสามารถของผู้สอบ 3 ตำแหน่ง

เมื่อโปรแกรมการคัดเลือกข้อสอบเพื่อสร้างเป็นแบบทดสอบกำหนดจำนวนแบบทดสอบที่ต้องการ
สร้างทั้งหมด เท่ากับ 5 ฉบับ และกำหนดเงื่อนไขของค่าฟังก์ชันสารสนเทศของแบบทดสอบตามฉบับอ้างอิง

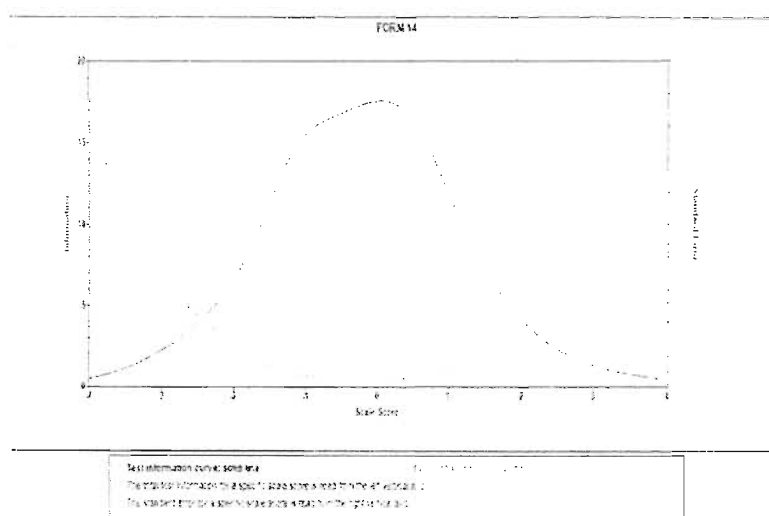
ณ 3 ตำแหน่งความสามารถ ผลการศึกษาพบว่า แบบทดสอบ จำนวน 2 ฉบับเท่านั้นที่มีค่าฟังก์ชันสารสนเทศสูงสุด ณ ตำแหน่งความสามารถที่ใกล้เคียงกับฉบับอ้างอิง โดยที่แบบทดสอบฉบับอ้างอิงมีค่าฟังก์ชันสารสนเทศของแบบทดสอบสูงสุด เท่ากับ 16.80 ณ ตำแหน่งความสามารถผู้สอบ 0.00 มีค่าความคลาดเคลื่อนมาตรฐานของการประมาณค่า 0.10 นั่นคือแบบทดสอบฉบับนี้มีความแม่นยำในการประมาณค่าความสามารถของผู้สอบที่มีความสามารถในระดับปานกลาง สำหรับแบบทดสอบฉบับที่ 1 มีค่าฟังก์ชันสารสนเทศของแบบทดสอบสูงสุด เท่ากับ 16.85 ณ ตำแหน่งความสามารถผู้สอบ 0.00 มีค่าความคลาดเคลื่อนมาตรฐานของการประมาณค่า 0.12 แบบทดสอบฉบับที่ 2 มีค่าฟังก์ชันสารสนเทศของแบบทดสอบสูงสุด เท่ากับ 16.80 ณ ตำแหน่งความสามารถผู้สอบ 0.00 มีค่าความคลาดเคลื่อนมาตรฐานของการประมาณค่า 0.12

จากการวิเคราะห์ค่าฟังก์ชันสารสนเทศของแบบทดสอบ จำนวน 2 ฉบับที่มีค่าใกล้เคียงหรือตรงกับแบบทดสอบฉบับอ้างอิง แสดงดังตาราง 4.3 และภาพ 4.7 – 4.8

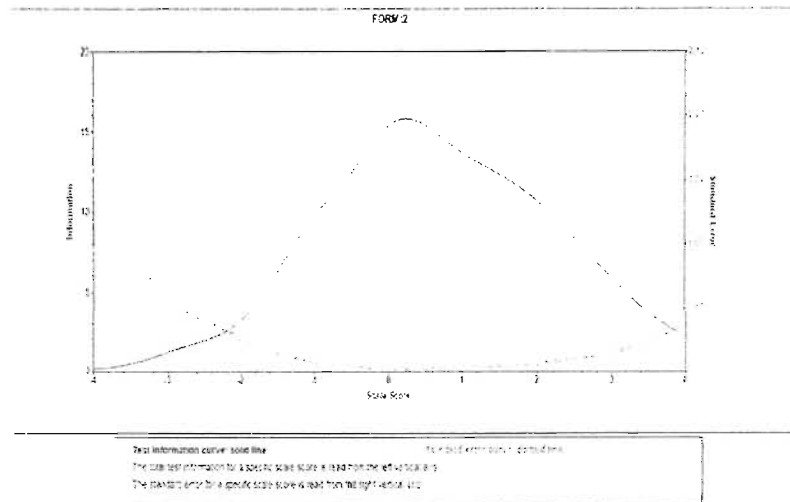
ตาราง 4.3 ค่าฟังก์ชันสารสนเทศของแบบทดสอบและค่าความคลาดเคลื่อนมาตรฐานของการประมาณค่าของแบบทดสอบฉบับอ้างอิงและแบบทดสอบ 2 ฉบับที่มีค่าใกล้เคียงหรือตรงกับแบบทดสอบฉบับอ้างอิง เมื่อกำหนดเงื่อนไขค่าฟังก์ชันสารสนเทศของแบบทดสอบของแบบทดสอบฉบับอ้างอิง ณ 3 ตำแหน่งความสามารถของผู้สอบ

ค่าฟังก์ชันสารสนเทศ ของแบบทดสอบฉบับอ้างอิง	แบบทดสอบฉบับ ที่	ค่าฟังก์ชันสารสนเทศ ของแบบทดสอบ	ค่าความคลาดเคลื่อน มาตรฐาน
16.80	1	16.85	0.12
	2	16.80	0.12

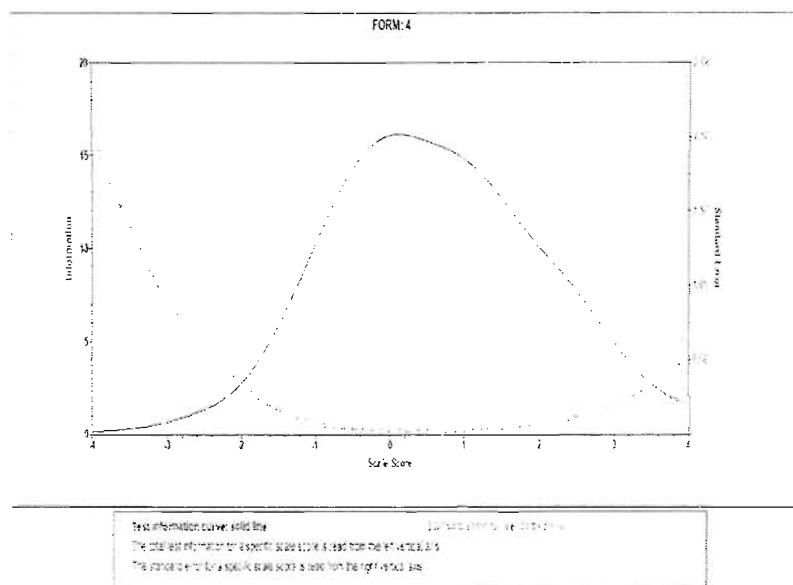
แบบทดสอบฉบับอ้างอิง



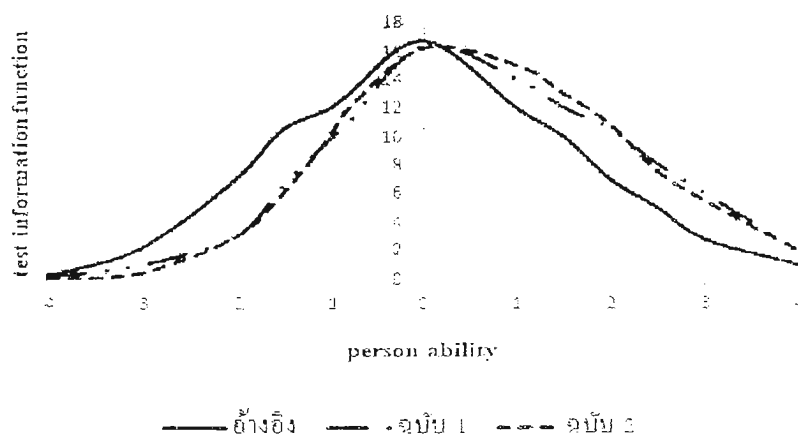
แบบทดสอบฉบับที่ 1



แบบทดสอบฉบับที่ 2



ภาพ 4.7 ค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง และแบบทดสอบที่สร้างจากโปรแกรม
การคัดเลือกข้อสอบ จำนวน 2 ฉบับ เมื่อกำหนดกำหนดเงื่อนไข
ค่าฟังก์ชันสารสนเทศของแบบทดสอบของแบบทดสอบฉบับอ้างอิง ณ 3 ตำแหน่งความสามารถ



ภาพ 4.8 ค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง และแบบทดสอบที่สร้างจากโปรแกรม
การคัดเลือกข้อสอบ จำนวน 2 ฉบับ เมื่อกำหนดค่าหนดเงื่อนไข

ค่าฟังก์ชันสารสนเทศของแบบทดสอบของแบบทดสอบฉบับอ้างอิง ณ 3 ตำแหน่งความสามารถ

กรณีที่ 3 กำหนดเงื่อนไขค่าฟังก์ชันสารสนเทศของแบบทดสอบของแบบทดสอบฉบับอ้างอิง ณ
ตำแหน่งความสามารถของผู้สอบ 4 ตำแหน่ง

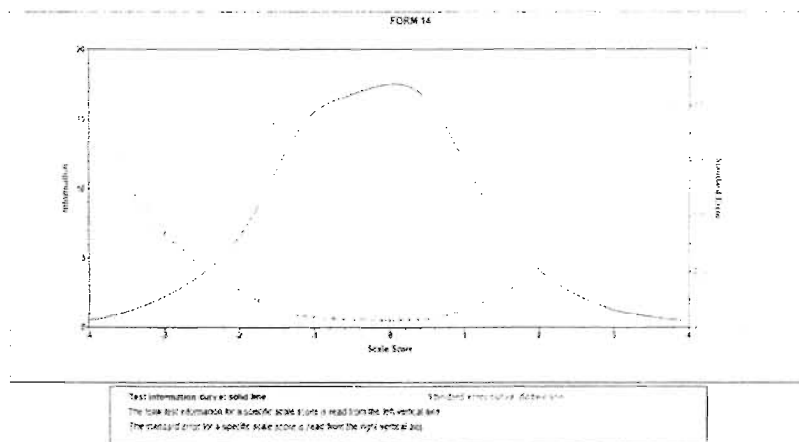
เมื่อโปรแกรมการคัดเลือกข้อสอบเพื่อสร้างเป็นแบบทดสอบกำหนดจำนวนแบบทดสอบที่ต้องการ
สร้างทั้งหมด เท่ากับ 5 ฉบับ และกำหนดเงื่อนไขของค่าฟังก์ชันสารสนเทศของแบบทดสอบตามฉบับอ้างอิง
ณ 4 ตำแหน่งความสามารถ ผลการศึกษาพบว่า มีแบบทดสอบ จำนวน 1 ฉบับเท่านั้นที่มีค่าฟังก์ชัน
สารสนเทศสูงสุด ณ ตำแหน่งความสามารถที่ใกล้เคียงกับฉบับอ้างอิง โดยที่แบบทดสอบฉบับอ้างอิงมีค่า
ฟังก์ชันสารสนเทศของแบบทดสอบสูงสุด เท่ากับ 16.80 ณ ตำแหน่งความสามารถผู้สอบ 0.00 มีค่าความ
คลาดเคลื่อนมาตรฐานของการประมาณค่า 0.10 นั่นคือแบบทดสอบฉบับนี้มีความแม่นยำในการประมาณค่า
ความสามารถของผู้สอบที่มีความสามารถในระดับปานกลาง สำหรับแบบทดสอบฉบับที่ 1 มีค่าฟังก์ชัน
สารสนเทศของแบบทดสอบสูงสุด เท่ากับ 16.85 ณ ตำแหน่งความสามารถผู้สอบ 0.00 มีค่าความ
คลาดเคลื่อนมาตรฐานของการประมาณค่า 0.10

จากการวิเคราะห์ค่าฟังก์ชันสารสนเทศของแบบทดสอบ จำนวน 1 ฉบับที่มีค่าใกล้เคียงหรือตรงกับ
แบบทดสอบฉบับอ้างอิง แสดงดังตาราง 4.4 และภาพ 4.9 – 4.10

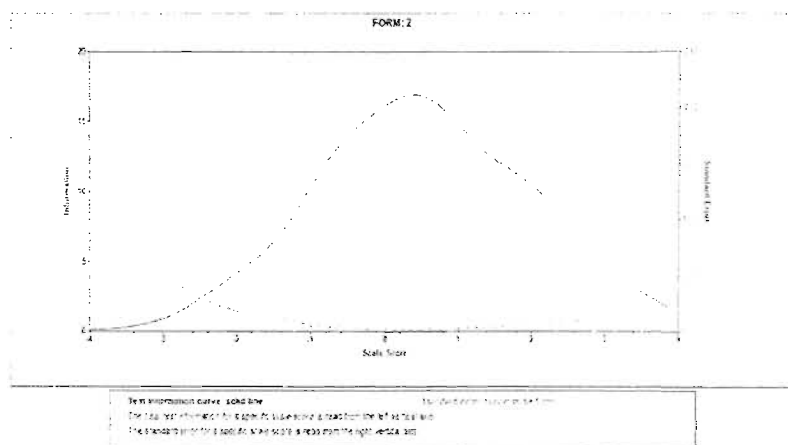
ตาราง 4.4 ค่าฟังก์ชันสารสนเทศของแบบทดสอบและค่าความคลาดเคลื่อนมาตรฐานของการประมาณค่าของแบบทดสอบฉบับอ้างอิงและแบบทดสอบ 1 ฉบับมีค่าใกล้เคียงหรือตรงกับแบบทดสอบฉบับอ้างอิง เมื่อกำหนดเงื่อนไขค่าฟังก์ชันสารสนเทศของแบบทดสอบของแบบทดสอบฉบับอ้างอิง ณ 4 ตำแหน่งความสามารถของผู้สอบ

ค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง	แบบทดสอบฉบับที่	ค่าฟังก์ชันสารสนเทศของแบบทดสอบ	ค่าความคลาดเคลื่อนมาตรฐาน
16.80	1	16.85	0.10

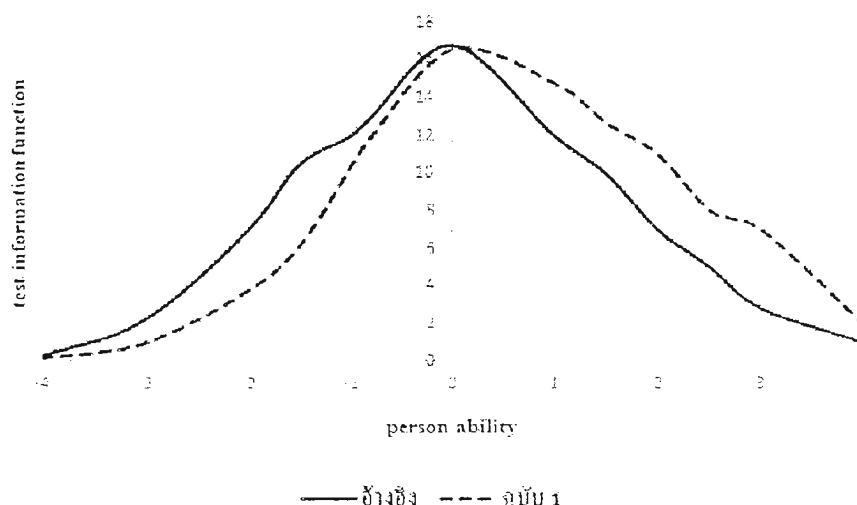
แบบทดสอบฉบับอ้างอิง



แบบทดสอบฉบับที่ 1



ภาพ 4.9 ค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง และแบบทดสอบที่สร้างจากโปรแกรมการคัดเลือกข้อสอบ จำนวน 1 ฉบับ เมื่อกำหนดค่าพิกัดเงื่อนไขค่าฟังก์ชันสารสนเทศของแบบทดสอบของแบบทดสอบฉบับอ้างอิง ณ 4 ตำแหน่งความสามารถ



ภาพ 4.10 ค่าฟังก์ชันสารสนเทศของแบบทดสอบของแบบทดสอบฉบับอ้างอิง และแบบทดสอบที่สร้างจากโปรแกรมการคัดเลือกข้อสอบ จำนวน 1 ฉบับ เมื่อกำหนดกำหนดเงื่อนไขค่าฟังก์ชันสารสนเทศของแบบทดสอบของแบบทดสอบฉบับอ้างอิง ณ 4 ตำแหน่งความสามารถ

2) ผลการคัดเลือกข้อสอบด้วยโปรแกรมเพื่อสร้างแบบทดสอบที่มีค่าความตรงตามทฤษฎีการตอบข้อสอบ

ส่วนผลการคัดเลือกข้อสอบจากคลังข้อสอบด้วยโปรแกรมเพื่อสร้างแบบทดสอบกลุ่มงานที่มีคุณภาพด้านความตรงตามเนื้อหา (content validity) เป็นการแสดงถึงข้อสอบที่ถูกคัดเลือกนั้นวัดเนื้อหาโดยตรงตามตารางกำหนดลักษณะเฉพาะรายวิชา (table of specification) และจำนวนข้อสอบในแต่ละเนื้อหาการวัดตรงตามแบบทดสอบฉบับอ้างอิง พบว่าเมื่อกำหนดเงื่อนไขค่าฟังก์ชันสารสนเทศของแบบทดสอบของแบบทดสอบฉบับอ้างอิง ณ ตำแหน่งความสามารถของผู้สอบ 2,3 และ 4 ตำแหน่ง โปรแกรมการคัดเลือกข้อสอบสามารถสร้างแบบทดสอบ จำนวน 5 ฉบับที่มีจำนวนข้อสอบในแต่ละเนื้อหาการวัดตรงกับแบบทดสอบฉบับอ้างอิง ดังตาราง 4.3

ตาราง 4.3 โครงสร้างของเนื้อหาการวัดของแบบทดสอบฉบับอ้างอิงและแบบทดสอบ 5 ฉบับ เมื่อกำหนด
เงื่อนไขค่าฟังก์ชันสารสนเทศของแบบทดสอบของแบบทดสอบฉบับอ้างอิง ณ ตำแหน่ง
ความสามารถของผู้สอบ 2,3 และ 4 ตำแหน่ง

เนื้อหาการวัด	จำนวนข้อสอบในแต่ละเนื้อหาการวัดของแบบทดสอบฉบับ					
	อ้างอิง	1	2	3	4	5
1	5	5	5	5	5	5
2	9	9	9	9	9	9
3	5	5	5	5	5	5
4	21	21	21	21	21	21
5	7	7	7	7	7	7
6	4	4	4	4	4	4
7	10	10	10	10	10	10
8	3	3	3	3	3	3

ตอนที่ 4.2 ผลการตรวจสอบความถูกต้องของโปรแกรมการคัดเลือกข้อสอบด้านคุณลักษณะเฉพาะของ

ความเป็นคู่ขนาน (parallelism specification) ของแบบทดสอบตามตำแหน่งความสามารถของ
ผู้สอบของแบบทดสอบฉบับอ้างอิง

จากผลการวิจัยตอนที่ 4.1 ผลการวิเคราะห์ด้านค่าความเที่ยงตามทฤษฎีการตอบข้อสอบที่พิจารณา
จากความสอดคล้องของค่าฟังก์ชันสารสนเทศของแบบทดสอบที่สร้างขึ้นจากโปรแกรมการคัดเลือกข้อสอบ
กับแบบทดสอบฉบับอ้างอิงนั้นเป็นการพิจารณาจากค่าฟังก์ชันสารสนเทศของแบบทดสอบ ณ ตำแหน่ง
ความสามารถผู้สอบเพียงตำแหน่งเดียวเท่านั้น แต่ McDonald (2009, p. 347-348) กล่าวว่าคุณลักษณะเฉพาะ
ของความเป็นคู่ขนาน (test parallelism) ควรพิจารณาจากค่าฟังก์ชันสารสนเทศของแบบทดสอบ (Test
Information Function: TIF) หรือค่าฟังก์ชันคุณลักษณะของแบบทดสอบ (Test Characteristic Function:
TCF) ตามตำแหน่งความสามารถของผู้สอบมากกว่า 1 ตำแหน่ง สำหรับการวิจัยครั้งนี้กำหนดค่าฟังก์ชัน
สารสนเทศของแบบทดสอบของแบบทดสอบฉบับอ้างอิงตามตำแหน่งความสามารถของผู้สอบ จำแนกเป็น
3 กรณี คือ กรณีที่ 1 ตำแหน่งความสามารถของผู้สอบ 2 ตำแหน่ง คือ 0.00 และ 1.50 กรณีที่ 2 ตำแหน่ง
ความสามารถของผู้สอบ 3 ตำแหน่ง คือ (-1.50), 0.00 และ 1.50 กรณีที่ 3 ตำแหน่งความสามารถของผู้สอบ 4
ตำแหน่ง คือ (-1.50), 0.00, 1.50 และ 2.50 ตามลำดับ และเกณฑ์การประเมินคุณลักษณะเฉพาะของความเป็น
คู่ขนาน คือ พิจารณาจากค่าเฉลี่ยยกกำลังสองค่าเบี่ยงเบนของค่าฟังก์ชันสารสนเทศของแบบทดสอบ (Mean

Square Deviation of Test Information Function: MSD of TIF) ที่มีค่าน้อยกว่า 0.05 (Luecht, 1998; Chen, P.-H., Chang, H.-H. & Wu, H. 2012)

ดังนั้นผู้วิจัยจึงนำเสนอผลการตรวจสอบความถูกต้องของโปรแกรมการคัดเลือกข้อสอบด้านคุณลักษณะเฉพาะของความเป็นคู่ขนาน (test parallelism) ของแบบทดสอบตามตำแหน่งความสามารถของผู้สอบของแบบทดสอบฉบับอ้างอิง ดังนี้

กรณีที่ 1 กำหนดค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง ณ ตำแหน่งความสามารถของผู้สอบ 2 ตำแหน่ง คือ 0.00 และ 1.50

เมื่อโปรแกรมการคัดเลือกข้อสอบเพื่อสร้างเป็นแบบทดสอบกำหนดจำนวนแบบทดสอบที่ต้องการสร้างทั้งหมด เท่ากับ 5 ฉบับ และกำหนดเงื่อนไขของค่าฟังก์ชันสารสนเทศของแบบทดสอบตามฉบับอ้างอิง ณ ตำแหน่งความสามารถของผู้สอบ 2 ตำแหน่ง คือ 0.00 และ 1.50 อีกทั้งจากผลการคัดเลือกข้อสอบเพื่อสร้างเป็นแบบทดสอบที่มีค่าความเที่ยงใกล้เคียงกับฉบับอ้างอิง พบว่า มีแบบทดสอบ 4 ฉบับที่มีคุณภาพด้านความเที่ยง แต่หากพิจารณาถึงคุณลักษณะเฉพาะด้านความเป็นคู่ขนานเพิ่มเติม พบว่า มีแบบทดสอบ 2 ฉบับเท่านั้นที่มีคุณลักษณะเฉพาะของความเป็นคู่ขนาน โดยที่แบบทดสอบคู่ขนานฉบับที่ 1 มีค่าเฉลี่ยยกกำลังสองค่าเบี่ยงเบนของค่าฟังก์ชันสารสนเทศของแบบทดสอบ (MSD of TIF) เท่ากับ 0.02 และแบบทดสอบคู่ขนานฉบับที่ 2 มีค่าเฉลี่ยยกกำลังสองค่าเบี่ยงเบนของค่าฟังก์ชันสารสนเทศของแบบทดสอบ แบบทดสอบ (MSD of TIF) เท่ากับ 0.00 ตามลำดับ ดังตาราง 4.4

ตาราง 4.4 ค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิงและแบบทดสอบคู่ขนาน 4 ฉบับ ณ ตำแหน่งความสามารถ 2 ตำแหน่ง และค่าเฉลี่ยยกกำลังสองค่าเบี่ยงเบนของค่าฟังก์ชันสารสนเทศของแบบทดสอบ

ตำแหน่ง ความสามารถ ของผู้สอบ	แบบทดสอบ ฉบับอ้างอิง	แบบทดสอบ ฉบับ 1	แบบทดสอบ ฉบับ 2	แบบทดสอบ ฉบับ 3	แบบทดสอบ ฉบับ 4
	$TI_{tgt}(\theta_i)$	$TI_{obs}(\theta_i)$			
0.00	16.80	16.85	16.80	17.00	17.03
1.50	10.00	9.8	10.00	10.30	7.50
MSD of TIF	-	0.02	0.00	0.06	3.37
คุณสมบัติ ความเป็น คู่ขนาน	-	มีคุณสมบัติ ความเป็น คู่ขนาน	มีคุณสมบัติ ความเป็น คู่ขนาน	ไม่มีคุณสมบัติ ความเป็น คู่ขนาน	ไม่มีคุณสมบัติ ความเป็น คู่ขนาน

กรณีที่ 2 ค่าฟังก์ชันสารสนเทศของแบบทดสอบ ณ ตำแหน่งความสามารถของผู้สอบ 3 ตำแหน่ง คือ (-1.50), 0.00 และ 1.50

เมื่อโปรแกรมการคัดเลือกข้อสอบเพื่อสร้างเป็นแบบทดสอบกำหนดจำนวนแบบทดสอบที่ต้องการสร้างทั้งหมด เท่ากับ 5 ฉบับ และกำหนดเงื่อนไขของค่าฟังก์ชันสารสนเทศของแบบทดสอบตามฉบับอ้างอิง ณ ตำแหน่งความสามารถของผู้สอบ 3 ตำแหน่ง คือ (-1.50), 0.00 และ 1.50 อีกทั้งจากการคัดเลือกข้อสอบเพื่อสร้างเป็นแบบทดสอบที่มีค่าความเที่ยงใกล้เคียงกับฉบับอ้างอิง พบว่า มีแบบทดสอบ 2 ฉบับที่มีคุณภาพด้านความเที่ยง แต่หากพิจารณาถึงคุณลักษณะเฉพาะด้านความเป็นคู่ขนานเพิ่มเติม พบว่า ไม่มีแบบทดสอบใดที่มีคุณสมบัติของคุณลักษณะเฉพาะของความเป็นคู่ขนาน ดังตาราง 4.5

ตาราง 4.5 ค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิงและแบบทดสอบคู่ขนาน 2 ฉบับ ณ ตำแหน่งความสามารถ 3 ตำแหน่ง และค่าเฉลี่ยยกกำลังสองค่าเบี่ยงเบนของค่าฟังก์ชันสารสนเทศของแบบทดสอบ

ตำแหน่งความสามารถของผู้สอบ	แบบทดสอบฉบับอ้างอิง	แบบทดสอบฉบับ 1	แบบทดสอบฉบับ 2
	$TI_{tgt}(\theta_i)$	$TI_{obs}(\theta_i)$	
-1.50	10.50	6.00	6.00
0.00	16.80	16.85	16.80
1.50	10.00	12.80	13.20
MSD of TIF	-	9.36	10.16
คุณสมบัติ ความเป็นคู่ขนาน	-	ไม่มีคุณสมบัติ ความเป็นคู่ขนาน	ไม่มีคุณสมบัติ ความเป็นคู่ขนาน

กรณีที่ 3 ค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง ณ ตำแหน่งความสามารถของผู้สอบ 4 ตำแหน่ง คือ (-1.50), 0.00, 1.50 และ 2.50

เมื่อโปรแกรมการคัดเลือกข้อสอบเพื่อสร้างเป็นแบบทดสอบกำหนดจำนวนแบบทดสอบที่ต้องการสร้างทั้งหมด เท่ากับ 5 ฉบับ และกำหนดเงื่อนไขของค่าฟังก์ชันสารสนเทศของแบบทดสอบตามฉบับอ้างอิง ณ ตำแหน่งความสามารถของผู้สอบ 4 ตำแหน่ง คือ (-1.50), 0.00, 1.50 และ 2.50 อีกทั้งจากการคัดเลือกข้อสอบเพื่อสร้างเป็นแบบทดสอบที่มีค่าความเที่ยงใกล้เคียงกับฉบับอ้างอิง พบว่า มีแบบทดสอบ 1 ฉบับที่มีคุณภาพด้านความเที่ยง แต่หากพิจารณาถึงคุณลักษณะเฉพาะด้านความเป็นคู่ขนานเพิ่มเติม พบว่า ไม่มีแบบทดสอบฉบับใดที่มีคุณสมบัติของคุณลักษณะเฉพาะของความเป็นคู่ขนาน ดังตาราง 4.6

ตาราง 4.6 ค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิงและแบบทดสอบคู่ขนาน 1 ฉบับตาม
ตำแหน่งความสามารถ 4 ตำแหน่ง และค่าเฉลี่ยยกกำลังสองค่าเบี่ยงเบนของค่าฟังก์ชันสารสนเทศ
ของแบบทดสอบ

ตำแหน่งความสามารถของผู้สอบ	แบบทดสอบฉบับอ้างอิง	แบบทดสอบฉบับ 1
	$TI_{tgt}(\theta_i)$	$TI_{obs}(\theta_i)$
-1.50	10.50	6.00
0.00	16.80	16.85
1.50	10.00	12.20
2.50	5.00	8.00
MSD of TIF	-	8.37
คุณสมบัติความเป็นคู่ขนาน	-	ไม่มีคุณสมบัติความเป็นคู่ขนาน

ตอนที่ 4.3 ผลตรวจสอบความถูกต้องของโปรแกรมการคัดเลือกข้อสอบด้านการใช้ข้อสอบชำระระหว่าง
แบบทดสอบคู่ขนาน (test overlap) ตามตำแหน่งความสามารถของผู้สอบของแบบทดสอบ
ฉบับอ้างอิง

ผลตรวจสอบความถูกต้องของโปรแกรมการคัดเลือกข้อสอบด้านการใช้ข้อสอบชำระระหว่าง
แบบทดสอบคู่ขนาน (test overlap) ตามตำแหน่งความสามารถของผู้สอบของแบบทดสอบฉบับอ้างอิง
(Referenced Test Information Function: RTIF) ณ ตำแหน่งความสามารถ 2, 3 และ 4 ตำแหน่ง
ดังรายละเอียด

กรณีที่ 1 กำหนดค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง ณ ตำแหน่งความสามารถของ
ผู้สอบ 2 ตำแหน่ง คือ 0.00 และ 1.50 พบว่าแบบทดสอบ 4 ฉบับที่มีค่าความเที่ยงใกล้เคียงกับแบบทดสอบ
ฉบับอ้างอิง แต่มีแบบทดสอบ จำนวน 1 ฉบับเท่านั้นที่มีอัตราการใช้ข้อสอบชำระระหว่างแบบทดสอบเฉลี่ย
(average test overlap rate: $\bar{T}$) เท่ากับ 0.070 ซึ่งเป็นอัตราการใช้ข้อสอบเฉลี่ยที่ต่ำกว่าเกณฑ์ที่กำหนด คือ
0.250 นั่นคือแบบทดสอบ จำนวน 1 ฉบับที่สร้างขึ้นจากโปรแกรมการคัดเลือกข้อสอบมีส่วนของการใช้
ร่วมกันข้อสอบไม่เกินเกณฑ์ที่กำหนด ส่วนแบบทดสอบที่เหลือ จำนวน 3 ฉบับมีอัตราการใช้ข้อสอบชำระ
ระหว่างแบบทดสอบเฉลี่ยสูงกว่าเกณฑ์ที่กำหนด โดยมีอัตราการใช้ข้อสอบชำระระหว่างแบบทดสอบเฉลี่ย
เท่ากับ 0.320, 0.369 และ 0.411 ตามลำดับ

กรณีที่ 2 กำหนดค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง ณ ตำแหน่งความสามารถของ
ผู้สอบ 3 ตำแหน่ง คือ (-1.50), 0.00 และ 1.50 พบว่าแบบทดสอบ 2 ฉบับที่มีค่าความเที่ยงใกล้เคียงกับ
แบบทดสอบฉบับอ้างอิง แต่แบบทดสอบ จำนวน 1 ฉบับเท่านั้นที่มีอัตราการใช้ข้อสอบชำระระหว่าง

แบบทดสอบเฉลี่ย (average test overlap rate: $\bar{T}$) เท่ากับ 0.070 ซึ่งเป็นอัตราการใช้ข้อสอบเฉลี่ยที่ต่ำกว่าเกณฑ์ที่กำหนด คือ 0.25 นั่นคือแบบทดสอบ จำนวน 1 ฉบับที่สร้างขึ้นจากโปรแกรมการคัดเลือกข้อสอบมีส่วนของการใช้ร่วมกันข้อสอบไม่เกินเกณฑ์ที่กำหนด ส่วนแบบทดสอบที่เหลืออีก 1 ฉบับมีอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบเฉลี่ยสูงกว่าเกณฑ์ที่กำหนด โดยมีอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบเฉลี่ย เท่ากับ 0.430

กรณีที่ 3 กำหนดค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง ณ ตำแหน่งความสามารถของผู้สอบ 4 ตำแหน่ง คือ (-1.50), 0.00, 1.50 และ 2.50 พบว่าแบบทดสอบ 1 ฉบับที่มีค่าความเที่ยงใกล้เคียงกับแบบทดสอบฉบับอ้างอิง และมีอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบเฉลี่ย (average test overlap rate: $\bar{T}$) เท่ากับ 0.070 ซึ่งเป็นอัตราการใช้ข้อสอบเฉลี่ยที่ต่ำกว่าเกณฑ์ที่กำหนด คือ 0.25 นั่นคือแบบทดสอบ 1 ฉบับที่สร้างขึ้นจากโปรแกรมการคัดเลือกข้อสอบมีส่วนของการใช้ร่วมกันข้อสอบไม่เกินเกณฑ์ที่กำหนด

ผลตรวจสอบความถูกต้องของ โปรแกรมการคัดเลือกข้อสอบด้านการใช้ข้อสอบซ้ำระหว่างแบบทดสอบคู่ขนาน (test overlap) ตามตำแหน่งความสามารถของผู้สอบของแบบทดสอบฉบับอ้างอิง (Referenced Test Information Function: RTIF) ดังตาราง 4.7

ตาราง 4.7 ค่าเฉลี่ยของอัตราข้อสอบซ้ำ ความแปรปรวนของข้อสอบซ้ำ และอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบเฉลี่ย

ตำแหน่ง ความ สามารถ ของ ผู้สอบ	จำนวน แบบทดสอบ ผ่านเกณฑ์ ความเที่ยง	ฉบับ ที่	ค่าเฉลี่ยของ อัตรา ข้อสอบซ้ำ	ความ แปรปรวนของ อัตราของ ข้อสอบซ้ำ	อัตราการใช้ ข้อสอบซ้ำ ระหว่าง แบบทดสอบ เฉลี่ย	ผ่านเกณฑ์ การใช้ ข้อสอบซ้ำ
2	4	1	0.050	0.001	0.070	ผ่าน
		2	0.250	0.020	0.320	ไม่ผ่าน
		3	0.280	0.025	0.369	ไม่ผ่าน
		4	0.330	0.027	0.411	ไม่ผ่าน
3	2	1	0.050	0.001	0.070	ผ่าน
		2	0.350	0.030	0.430	ไม่ผ่าน
4	1	1	0.050	0.001	0.070	ผ่าน

จากผลการวิจัยข้างต้นสามารถสรุปผลตรวจสอบความถูกต้องของโปรแกรมการคัดเลือกข้อสอบเพื่อสร้างเป็นแบบทดสอบคู่ขนานที่มีคุณภาพด้านความตรง ความเที่ยงตามทฤษฎีการตอบข้อสอบ และความถูกต้องของโปรแกรมการคัดเลือกข้อสอบด้านคุณลักษณะเฉพาะของความเป็นคู่ขนาน (parallelism specification) และด้านการใช้ข้อสอบซ้ำระหว่างแบบทดสอบคู่ขนาน (test overlap) ตามตำแหน่งความสามารถของผู้สอบของแบบทดสอบฉบับอ้างอิง (Referenced Test Information Function: RTIF) ดังตาราง 4.8

ตาราง 4.8 สรุปผลตรวจสอบความตรง ความเที่ยงของแบบทดสอบ และความถูกต้องของโปรแกรมการคัดเลือกข้อสอบด้านคุณลักษณะเฉพาะของความเป็นคู่ขนาน (parallelism specification) และด้านการใช้ข้อสอบซ้ำระหว่างแบบทดสอบคู่ขนาน (test overlap) ตามตำแหน่งความสามารถของผู้สอบของแบบทดสอบฉบับอ้างอิง (Referenced Test Information Function: RTIF)

เงื่อนไขตำแหน่ง ความสามารถ ของโปรแกรม	2 ตำแหน่ง		3 ตำแหน่ง		4 ตำแหน่ง	
	จำนวนแบบทดสอบ					
	สร้างขึ้น	ผ่านเกณฑ์	สร้างขึ้น	ผ่านเกณฑ์	สร้างขึ้น	ผ่านเกณฑ์
ความตรง	5	5	5	5	5	5
ความเที่ยง	5	4	5	2	5	1
ความเป็นคู่ขนาน	5	2	5	-	5	-
การใช้ข้อสอบซ้ำ	5	1	5	1	5	1

จากตาราง 4.8 แสดงให้เห็นถึงสรุปผลการตรวจสอบความถูกต้องของโปรแกรมการคัดเลือกข้อสอบทั้งด้านคุณลักษณะเฉพาะความเป็นคู่ขนานและด้านการใช้ข้อสอบซ้ำตามตำแหน่งค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง พบว่ากรณีที่ 1 กำหนดค่าฟังก์ชันสารสนเทศของแบบทดสอบตามตำแหน่งความสามารถของผู้สอบ 2 ตำแหน่ง คือ 0.00 และ 1.50 พบว่าโปรแกรมการคัดเลือกข้อสอบสามารถรวบรวมเป็นแบบทดสอบที่มีความตรงตามเนื้อหา จำนวน 5 ฉบับ รวบรวมเป็นแบบทดสอบที่มีความเที่ยง จำนวน 4 ฉบับ รวบรวมเป็นแบบทดสอบที่มีคุณลักษณะความเป็นคู่ขนาน จำนวน 2 ฉบับ และรวบรวมเป็นแบบทดสอบที่มีอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบเฉลี่ยระดับต่ำกว่าเกณฑ์ที่กำหนด คือ 0.25 จำนวน 1 ฉบับ ส่วนกรณีที่ 2 กำหนดค่าฟังก์ชันสารสนเทศของแบบทดสอบตามตำแหน่งความสามารถของผู้สอบ 3 ตำแหน่ง คือ -1.50, 0.00 และ 1.50 พบว่าโปรแกรมการคัดเลือกข้อสอบสามารถรวบรวมเป็นแบบทดสอบที่มีความตรงตามเนื้อหา จำนวน 5 ฉบับ รวบรวมเป็นแบบทดสอบที่มีความเที่ยง จำนวน 2 ฉบับ โปรแกรมไม่สามารถรวบรวมเป็นแบบทดสอบที่มีคุณลักษณะความเป็นคู่ขนานได้ และ

รวบรวมเป็นแบบทดสอบที่มีอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบเฉลี่ยระดับต่ำกว่าเกณฑ์ที่กำหนด คือ 0.25 จำนวน 1 ฉบับ และกรณีที่ 3 กำหนดค่าฟังก์ชันสารสนเทศของแบบทดสอบตามตำแหน่งความสามารถของผู้สอบ 4 ตำแหน่ง คือ -1.50, 0.00, 1.50 และ 2.50 พบว่าโปรแกรมการคัดเลือกข้อสอบสามารถรวบรวมเป็นแบบทดสอบที่มีความตรงตามเนื้อหา จำนวน 5 ฉบับ รวบรวมเป็นแบบทดสอบที่มีความเที่ยง จำนวน 1 ฉบับ โปรแกรมไม่สามารถรวบรวมเป็นแบบทดสอบที่มีคุณลักษณะความเป็นคู่ขนานได้ และรวบรวมเป็นแบบทดสอบที่มีอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบเฉลี่ยระดับต่ำกว่าเกณฑ์ที่กำหนด คือ 0.25 จำนวน 1 ฉบับ

บทที่ 5

สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

การวิจัยครั้งนี้มีวัตถุประสงค์หลักเพื่อสร้างแบบทดสอบคู่ขนานจากคลังข้อสอบที่มีคุณภาพด้านความเที่ยงและความตรงตามทฤษฎีการตอบข้อสอบ โดยมีวัตถุประสงค์ย่อย คือ

- 1) เพื่อสร้าง โปรแกรมการคัดเลือกข้อสอบจากคลังข้อสอบเพื่อสร้างแบบทดสอบคู่ขนานที่มีคุณภาพด้านความเที่ยง (reliability) และความตรง (validity) ตามทฤษฎีการตอบข้อสอบ
- 2) เพื่อตรวจสอบความถูกต้องของ โปรแกรมการคัดเลือกข้อสอบด้านคุณลักษณะเฉพาะของความเป็นคู่ขนาน (parallelism specification) ของแบบทดสอบตามตำแหน่งความสามารถของผู้สอบของแบบทดสอบฉบับอ้างอิง
- 3) เพื่อตรวจสอบความถูกต้องของ โปรแกรมการคัดเลือกข้อสอบด้านการใช้ข้อสอบชำระระหว่างแบบทดสอบคู่ขนาน (test overlap) ตามความสามารถของผู้สอบของแบบทดสอบฉบับอ้างอิง

ขอบเขตการวิจัย

1. การวิจัยครั้งนี้ใช้วิธีการศึกษาโดยการจำลองข้อมูลการสอบที่มีคำตอบแบบ 0 และ 1 บรรจุคลังข้อสอบ โดยข้อมูลการสอบที่จำลองขึ้นอยู่บนพื้นฐานของผลการทดสอบจากรายวิชา MER 2003 ความรู้เบื้องต้นเกี่ยวกับสถิติและวิจัยทางการศึกษา จำนวนผู้สอบ 2,800 คน

2. ความตรงตามเนื้อหาได้จาก

- 1) การจัดทำตารางกำหนดลักษณะเฉพาะรายวิชา (table of specification)
- 2) จำลองข้อสอบคู่ขนานในแต่ละเนื้อหา
3. กำหนดให้เป็นแบบทดสอบปรนัยแบบเลือกตอบ จำนวน 4 ตัวเลือก
4. ตัวแปรในการวิจัย

ตัวแปรอิสระ คือ โปรแกรมการคัดเลือกข้อสอบที่มีการกำหนดค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิงตามตำแหน่งความสามารถของผู้สอบของแบบทดสอบฉบับอ้างอิง จำแนกเป็น

- 1) ค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง 2 ตำแหน่ง คือ
 - 1) $RTIF_{\theta_1}=0.00$ เท่ากับ 16.80
 - 2) $RTIF_{\theta_2}=1.50$ เท่ากับ 10.00
- 2) ค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง 3 ตำแหน่ง คือ
 - 2.1) $RTIF_{\theta_1}=-1.50$ เท่ากับ 10.50
 - 2.2) $RTIF_{\theta_2}=0.00$ เท่ากับ 16.80
 - 2.3) $RTIF_{\theta_3}=1.50$ เท่ากับ 10.00

3) ค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง 4 ตำแหน่ง คือ

$$3.1) RTIF_{\theta_1} = -1.50 \quad \text{เท่ากับ } 10.50$$

$$3.2) RTIF_{\theta_2} = 0.00 \quad \text{เท่ากับ } 16.08$$

$$3.3) RTIF_{\theta_3} = 1.50 \quad \text{เท่ากับ } 10.00$$

$$3.4) RTIF_{\theta_4} = 2.50 \quad \text{เท่ากับ } 5.00$$

ตัวแปรตาม คือ ความถูกต้องของโปรแกรมการคัดเลือกข้อสอบ จำแนกเป็น

- 1) ความตรงตามเนื้อหา
- 2) ความเที่ยง
- 3) ค่าฟังก์ชันสารสนเทศของแบบทดสอบ
- 4) คุณลักษณะเฉพาะของความเป็นคู่ขนาน
- 5) อัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบ

ผู้วิจัยการจัดกระทำข้อมูลและการเก็บรวบรวมข้อมูล เพื่อสร้างคลังข้อสอบตามทฤษฎีการตอบข้อสอบ ที่มีจุดมุ่งหมายเพื่อนำมาสร้างแบบทดสอบคู่ขนานมากกว่า 1 ฉบับ โดยมีขั้นตอนการวิจัย ดังนี้

ขั้นตอนที่ 1 การสร้างและตรวจสอบคุณภาพของแบบทดสอบฉบับอ้างอิง

1.1 การสร้าง ตรวจสอบคุณภาพ และคัดเลือกข้อสอบของแบบทดสอบฉบับอ้างอิงตามทฤษฎีการทดสอบแบบดั้งเดิม

1.2 การตรวจสอบคุณภาพของข้อสอบที่ผ่านการคัดเลือกของแบบทดสอบฉบับอ้างอิงตามทฤษฎีการตอบข้อสอบ

ขั้นตอนที่ 2 การสร้างคลังข้อสอบตามคุณสมบัติของแบบทดสอบฉบับอ้างอิงด้วยการจำลองข้อมูลการสอบด้วยโปรแกรม WINGEN

2.1 กำหนดสัดส่วนของคลังข้อสอบตามเนื้อหาการวัดตามแบบทดสอบฉบับอ้างอิง

2.2 การจำลองข้อมูลการสอบด้วยโปรแกรม WINGEN

2.3 การตรวจสอบและคัดเลือกข้อสอบเพื่อบรรจุในคลังข้อสอบที่มีคุณภาพตามทฤษฎีการตอบข้อสอบ

การวิเคราะห์ข้อมูล ดำเนินการดังนี้

1. การวิเคราะห์คุณภาพของข้อสอบและแบบทดสอบตามทฤษฎีการทดสอบแบบดั้งเดิม

1.1 วิเคราะห์ดัชนีความสอดคล้องระหว่างข้อสอบกับวัตถุประสงค์ในการวัด (Index of Item-Objective Congruence: IOC) ด้วยสูตรของ Hambleton (as cited in Turner & Carlson, 2003, p.167)

1.2 วิเคราะห์ค่าความยากของข้อสอบใช้ค่าสัดส่วนในการตอบข้อสอบแต่ละข้อถูก

1.3 วิเคราะห์อำนาจจำแนกของข้อสอบด้วยค่าสหสัมพันธ์แบบไบซีเรียล (Biserial Correlation)

1.4 วิเคราะห์ค่าสัมประสิทธิ์ความเที่ยงแบบความสอดคล้องในด้วยวิธีการของ Kuder

Richardson Formula 20

2. การวิเคราะห์ค่าพารามิเตอร์ของข้อสอบและแบบทดสอบตามทฤษฎีการตอบข้อสอบ

2.1 ตรวจสอบความเป็นเอกมิติแบบเด่นเดี่ยวของแบบทดสอบด้วยการวิเคราะห์องค์ประกอบ

เชิงสำรวจ (Exploratory Factor Analysis)

2.2 ประเมินความเหมาะสมของโมเดลการตอบข้อสอบกับข้อมูลเชิงประจักษ์ด้วยการประมาณค่า -2 likelihood และทดสอบความเหมาะสมของโมเดลด้วย χ^2 - test

2.2 วิเคราะห์ค่าพารามิเตอร์ของข้อสอบและแบบทดสอบตามทฤษฎีการตอบข้อสอบ แบบ 3 พารามิเตอร์

3. การประเมินคุณลักษณะเฉพาะของความเป็นคู่ขนาน (Test Parallelism) จากการพิจารณา Mean Square Deviation of Test Information Function ระหว่างแบบทดสอบคู่ขนานกับแบบทดสอบฉบับอ้างอิง ควรมีค่าน้อยกว่าหรือเท่ากับ 0.05 ของ Luecht (1998) และ Chen, P.-H., Chang, H.-H. & Wu, H. (2012)

4. การวิเคราะห์อัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบเฉลี่ย (Test Overlap rate) ของ Chen, Ankenmann & Spray (2003)

สรุปผลการวิจัย

การวิจัยครั้งนี้สามารถสรุปผลการวิจัยแบ่งเป็น 3 ตอน คือ

ตอนที่ 1 โปรแกรมการสร้างแบบทดสอบคู่ขนาน และผลการคัดเลือกข้อสอบจากคลังข้อสอบด้วยโปรแกรมเพื่อสร้างแบบทดสอบคู่ขนานมีคุณภาพด้านความเที่ยง (reliability) และความตรง (validity) ตามทฤษฎีการตอบข้อสอบ

1.1 โปรแกรมการสร้างแบบทดสอบคู่ขนาน

โปรแกรมการสร้างข้อสอบคู่ขนานผู้วิจัยได้ประยุกต์จากชุดคำสั่ง Ip_Solve Version 5.5 ในโปรแกรม R Studio เพื่อคัดเลือกข้อสอบจากคลังข้อสอบที่สร้างขึ้นจากงานวิจัยครั้งนี้ โดยข้อสอบที่ผ่านการคัดเลือกจะถูกนำไปสร้างเป็นแบบทดสอบคู่ขนานมากกว่า 1 ฉบับที่มีคุณสมบัติด้านความเป็นคุณลักษณะเฉพาะของความเป็นคู่ขนาน (test parallelism) ด้านความเที่ยง และด้านความตรงตามเนื้อหาที่ใกล้เคียงหรือตรงกับแบบทดสอบฉบับอ้างอิงมากที่สุด รวมทั้งแบบทดสอบคู่ขนานมากกว่า 1 ฉบับมีอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบอยู่ในระดับต่ำ

ส่วนประกอบของโปรแกรมการคัดเลือกข้อสอบ คือ

ส่วนที่ 1 การกำหนดปัจจัยนำเข้า ประกอบด้วย

1.1 การกำหนดคุณลักษณะเฉพาะของแบบทดสอบคู่ขนานตามคุณลักษณะเฉพาะของแบบทดสอบฉบับอ้างอิง ประกอบด้วย โครงสร้างเนื้อหาในการวัดของแบบทดสอบ จำนวนแบบทดสอบคู่ขนานที่ต้องการสร้าง ความยาวของแบบทดสอบ

1.1.1 โครงสร้างของเนื้อหาการวัดของแบบทดสอบฉบับอ้างอิง คือ 8 เนื้อหาการวัด โดยมีจำนวนข้อสอบตามเนื้อหาการวัด คือ 5, 9, 5, 21, 7, 4, 10 และ 3 ข้อ ตามลำดับ

1.1.2 จำนวนแบบทดสอบคู่ขนานที่ต้องการสร้าง คือ 1 ถึง 5 ฉบับ

1.1.3 ความยาวของแบบทดสอบ คือ 64 ข้อ

1.2 การกำหนดค่าฟังก์ชันสารสนเทศของของแบบทดสอบคู่ขนานตามแบบทดสอบฉบับอ้างอิง ณ ตำแหน่งความสามารถของผู้สอบที่ต้องการ (RTIF) แบ่งเป็น 3 กรณี

1.2.1 ค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง 2 ตำแหน่ง คือ

1) $RTIF_{\theta_1=0.00}$ เท่ากับ 16.80

2) $RTIF_{\theta_2=1.50}$ เท่ากับ 10.00

1.2.2 ค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง 3 ตำแหน่ง คือ

1) $RTIF_{\theta_1=-1.50}$ เท่ากับ 10.50

2) $RTIF_{\theta_2=0.00}$ เท่ากับ 16.80

3) $RTIF_{\theta_3=1.50}$ เท่ากับ 10.00

1.2.3 ค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง 4 ตำแหน่ง คือ

1) $RTIF_{\theta_1=-1.50}$ เท่ากับ 10.50

2) $RTIF_{\theta_2=0.00}$ เท่ากับ 16.80

3) $RTIF_{\theta_3=1.50}$ เท่ากับ 10.00

4) $RTIF_{\theta_4=2.50}$ เท่ากับ 5.00

1.2 การกำหนดค่าฟังก์ชันสารสนเทศของแบบทดสอบคู่ขนานตามค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง ณ ตำแหน่งความสามารถของผู้สอบที่ใช้ในการวิจัยครั้งนี้

ส่วนที่ 2 ชุดคำสั่งของขั้นตอนวิธีและเงื่อนไขบังคับ ประกอบด้วย

2.1 ชุดคำสั่งการตัดสินใจคัดเลือกข้อสอบจากคลังข้อสอบเพื่อสร้างเป็นแบบทดสอบคู่ขนานที่แสดงในลักษณะตัวแปรสุ่มที่มีการแจกแจงแบบทวินาม คือ

$$X_{it} = \begin{cases} 1 & \text{เมื่อข้อสอบที่ } i \text{ ถูกจัดอยู่ในแบบสอบ } t \\ 0 & \text{เมื่อข้อสอบที่ } i \text{ ไม่ถูกจัดอยู่ในแบบสอบ } t \end{cases}$$

2.2 ชุดคำสั่งเงื่อนไขบังคับของชุดคำสั่ง IpSolveAPI เพื่อสร้างเป็นแบบทดสอบคู่ขนานจากคลังข้อสอบที่มีค่าฟังก์ชันสารสนเทศของแบบทดสอบตามตำแหน่งความสามารถของผู้สอบ (Test Information Function: TIF) ที่มีค่าใกล้เคียงหรือมีความแตกต่างน้อยที่สุดจากค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิงตามตำแหน่งความสามารถของผู้สอบ (Referenced Test Information Function: RTIF)

2.3 ชุดคำสั่งเงื่อนไขบังคับการสุ่มข้อสอบที่ไม่ให้มีการใช้ข้อสอบซ้ำระหว่างฉบับ

2.4 ชุดคำสั่งขั้นตอนวิธีการสุ่มข้อสอบเพื่อสร้างเป็นแบบทดสอบคู่ขนานมากกว่า 1 ฉบับตามโครงสร้างเนื้อหาการวัด และค่าฟังก์ชันสารสนเทศของแบบทดสอบที่กำหนด

2.5 ชุดคำสั่งขั้นตอนวิธีการแสดงผลการคัดเลือกข้อสอบในรูปแบบตาราง

ส่วนที่ 3 ผลลัพธ์ของขั้นตอนวิธีและเงื่อนไขบังคับ

ผลลัพธ์ของขั้นตอนวิธีและเงื่อนไขบังคับที่อยู่ในรูปตาราง ประกอบด้วย

3.1 ผลลัพธ์การคัดเลือกข้อสอบจากคลังข้อสอบเพื่อสร้างเป็นแบบทดสอบคู่ขนานมากกว่า 1 ฉบับ โดยที่ค่าแสดง เท่ากับ 0 หมายถึง เป็นข้อสอบไม่ถูกคัดเลือกจากคลังข้อสอบตามเงื่อนไขบังคับ ส่วนค่าแสดง เท่ากับ 1 หมายถึง เป็นข้อสอบที่ถูกคัดเลือกจากคลังข้อสอบตามเงื่อนไขบังคับ

3.2 ผลลัพธ์ของแบบทดสอบคู่ขนานที่มีค่าฟังก์ชันสารสนเทศของแบบทดสอบตามตำแหน่งความสามารถของผู้สอบ (Test Information Function: TIF) ที่มีค่าใกล้เคียงหรือมีความแตกต่างน้อยที่สุดจากค่าฟังก์ชันสารสนเทศของแบบทดสอบตามตำแหน่งความสามารถของผู้สอบของแบบทดสอบฉบับอ้างอิง (Referenced Test Information Function: RTIF)

1.2 ผลการคัดเลือกข้อสอบด้วยโปรแกรมเพื่อสร้างแบบทดสอบที่มีค่าความเที่ยงตามทฤษฎีการตอบสนองข้อสอบ

สำหรับผลการคัดเลือกข้อสอบด้วยโปรแกรมเพื่อสร้างเป็นแบบทดสอบที่มีค่าความเที่ยงตามทฤษฎีการตอบสนองข้อสอบ นั้นเป็นการพิจารณาจากค่าฟังก์ชันสารสนเทศของแบบทดสอบที่ได้จากโปรแกรมมีค่าฟังก์ชันสารสนเทศสูงสุด ณ ตำแหน่งความสามารถที่ใกล้เคียงหรือตรงกับค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง พบว่า

กรณีที่ 1 กำหนดเงื่อนไขค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง ณ 2 ตำแหน่งความสามารถของผู้สอบ

เมื่อโปรแกรมการคัดเลือกข้อสอบเพื่อสร้างเป็นแบบทดสอบกำหนดจำนวนแบบทดสอบที่ต้องการสร้างทั้งหมด เท่ากับ 5 ฉบับ และกำหนดเงื่อนไขของค่าฟังก์ชันสารสนเทศของแบบทดสอบตามฉบับอ้างอิง ณ 2 ตำแหน่งความสามารถของผู้สอบ ผลการศึกษาพบว่า มีแบบทดสอบ จำนวน 4 ฉบับเท่านั้นที่มีค่าฟังก์ชันสารสนเทศสูงสุด ณ ตำแหน่งความสามารถของผู้สอบที่ใกล้เคียงกับฉบับอ้างอิง โดยที่แบบทดสอบฉบับอ้างอิงมีค่าฟังก์ชันสารสนเทศของแบบทดสอบสูงสุด เท่ากับ 16.80 ณ ตำแหน่ง

แม่นยำในการประมาณค่าความสามารถของผู้สอบที่มีความสามารถในระดับปานกลาง สำหรับแบบทดสอบฉบับที่ 1 มีค่าฟังก์ชันสารสนเทศของแบบทดสอบสูงสุดเท่ากับ 16.85 ณ ตำแหน่งความสามารถของผู้สอบที่ 0.00 มีค่าความคลาดเคลื่อนมาตรฐานของการประมาณค่า 0.10

1.3 ผลการคัดเลือกข้อสอบด้วยโปรแกรมเพื่อสร้างแบบทดสอบที่มีค่าความตรงตามทฤษฎีการตอบสนองข้อสอบ

ส่วนผลการคัดเลือกข้อสอบจากคลังข้อสอบด้วยโปรแกรมเพื่อสร้างแบบทดสอบคู่ขนานที่มีคุณภาพด้านความตรงตามเนื้อหา (content validity) เป็นการแสดงถึงข้อสอบที่ถูกคัดเลือกนั้นวัดเนื้อหาข้อสอบตรงตามตารางกำหนดลักษณะเฉพาะรายวิชา (table of specification) และจำนวนข้อสอบในแต่ละเนื้อหาการวัดตรงตามแบบทดสอบฉบับอ้างอิง พบว่าเมื่อกำหนดเงื่อนไขค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง ณ 2, 3 และ 4 ตำแหน่งความสามารถของผู้สอบ โปรแกรมการคัดเลือกข้อสอบสามารถสร้างแบบทดสอบจำนวน 5 ฉบับที่มีจำนวนข้อสอบในแต่ละเนื้อหาการวัดตรงกับแบบทดสอบฉบับอ้างอิง

ตอนที่ 2 ผลการตรวจสอบความถูกต้องของโปรแกรมการคัดเลือกข้อสอบด้านคุณลักษณะเฉพาะของความเป็นคู่ขนาน (parallelism specification) ของแบบทดสอบตามตำแหน่งความสามารถของผู้สอบของแบบทดสอบฉบับอ้างอิง

จากผลการวิจัยด้านค่าความเที่ยงตามทฤษฎีการตอบข้อสอบ ที่พิจารณาจากความสอดคล้องของค่าฟังก์ชันสารสนเทศของแบบทดสอบกับแบบทดสอบฉบับอ้างอิงนั้นเป็นการพิจารณาจากค่าฟังก์ชันสารสนเทศของแบบทดสอบ ณ ตำแหน่งความสามารถของผู้สอบเพียงตำแหน่งเดียวเท่านั้น แต่คุณลักษณะเฉพาะของความเป็นคู่ขนาน (test parallelism) ควรพิจารณาจากค่าฟังก์ชันสารสนเทศของแบบทดสอบ (Test Information Function: TIF) หรือค่าฟังก์ชันคุณลักษณะของแบบทดสอบ (Test Characteristic Function: TCF) ตามตำแหน่งความสามารถของผู้สอบ สำหรับในการวิจัยครั้งนี้กำหนดตำแหน่งความสามารถของผู้สอบเป็น 3 กรณี คือ กรณีที่ 1 ความสามารถของผู้สอบ 2 ตำแหน่ง คือ 0.00 และ 1.50 กรณีที่ 2 ความสามารถของผู้สอบ 3 ตำแหน่ง คือ (-1.50), 0.00 และ 1.50 กรณีที่ 3 ความสามารถของผู้สอบ 4 ตำแหน่ง คือ (-1.50), 0.00, 1.50 และ 2.50 ตามลำดับ และเกณฑ์การประเมินคุณลักษณะเฉพาะของความเป็นคู่ขนาน คือ พิจารณาจากค่าเฉลี่ยยกกำลังสองค่าเบี่ยงเบนของค่าฟังก์ชันสารสนเทศของแบบทดสอบ (Mean Square Deviation of Test Information Function: MSD of TIF) ที่มีค่าน้อยกว่า 0.05 ผลการวิจัยพบว่า

กรณีที่ 1 กำหนดค่าฟังก์ชันสารสนเทศของแบบทดสอบ ณ ตำแหน่งความสามารถของผู้สอบ 2 ตำแหน่ง คือ 0.00 และ 1.50

เมื่อโปรแกรมการคัดเลือกข้อสอบเพื่อสร้างเป็นแบบทดสอบกำหนดจำนวนแบบทดสอบที่ต้องการสร้างทั้งหมด เท่ากับ 5 ฉบับ และกำหนดเงื่อนไขของค่าฟังก์ชันสารสนเทศของแบบทดสอบตามฉบับอ้างอิง ณ ตำแหน่งความสามารถของผู้สอบ 2 ตำแหน่ง คือ 0.00 และ 1.50 อีกทั้งจากผลการคัดเลือกข้อสอบเพื่อ

สร้างเป็นแบบทดสอบที่มีค่าความเที่ยงใกล้เคียงกับฉบับอ้างอิง พบว่า มีแบบทดสอบ 4 ฉบับที่มีคุณภาพด้านความเที่ยง แต่หากพิจารณาถึงคุณลักษณะเฉพาะด้านความเป็นคู่ขนานเพิ่มเติม พบว่า มีแบบทดสอบ 2 ฉบับเท่านั้นที่มีคุณลักษณะเฉพาะของความเป็นคู่ขนาน โดยที่แบบทดสอบคู่ขนานฉบับที่ 1 มีค่าเฉลี่ยยกกำลังสองค่าเบี่ยงเบนของค่าฟังก์ชันสารสนเทศของแบบทดสอบ เท่ากับ 0.02 และแบบทดสอบคู่ขนานฉบับที่ 2 มีค่าเฉลี่ยยกกำลังสองค่าเบี่ยงเบนของค่าฟังก์ชันสารสนเทศของแบบทดสอบ เท่ากับ 0.00 ตามลำดับ

กรณีที่ 2 ค่าฟังก์ชันสารสนเทศของแบบทดสอบ ณ ตำแหน่งความสามารถของผู้สอบ 3 ตำแหน่ง คือ (-1.50), 0.00 และ 1.50

เมื่อโปรแกรมการคัดเลือกข้อสอบเพื่อสร้างเป็นแบบทดสอบกำหนดจำนวนแบบทดสอบที่ต้องการสร้างทั้งหมด เท่ากับ 5 ฉบับ และกำหนดเงื่อนไขของค่าฟังก์ชันสารสนเทศของแบบทดสอบตามฉบับอ้างอิง ณ ตำแหน่งความสามารถของผู้สอบ 3 ตำแหน่ง คือ (-1.50), 0.00 และ 1.50 อีกทั้งจากการคัดเลือกข้อสอบเพื่อสร้างเป็นแบบทดสอบที่มีค่าความเที่ยงใกล้เคียงกับฉบับอ้างอิง พบว่า มีแบบทดสอบ 2 ฉบับที่มีคุณภาพด้านความเที่ยง แต่หากพิจารณาถึงคุณลักษณะเฉพาะด้านความเป็นคู่ขนานเพิ่มเติม พบว่า ไม่มีแบบทดสอบใดที่มีคุณสมบัติของคุณลักษณะเฉพาะของความเป็นคู่ขนาน

กรณีที่ 3 ค่าฟังก์ชันสารสนเทศของแบบทดสอบ ณ ตำแหน่งความสามารถของผู้สอบ 4 ตำแหน่ง คือ (-1.50), 0.00, 1.50 และ 2.50

เมื่อโปรแกรมการคัดเลือกข้อสอบเพื่อสร้างเป็นแบบทดสอบกำหนดจำนวนแบบทดสอบที่ต้องการสร้างทั้งหมด เท่ากับ 5 ฉบับ และกำหนดเงื่อนไขของค่าฟังก์ชันสารสนเทศของแบบทดสอบตามฉบับอ้างอิง ณ ตำแหน่งความสามารถของผู้สอบ 4 ตำแหน่ง คือ (-1.50), 0.00, 1.50 และ 2.50 อีกทั้งจากการคัดเลือกข้อสอบเพื่อสร้างเป็นแบบทดสอบที่มีค่าความเที่ยงใกล้เคียงกับฉบับอ้างอิง พบว่า มีแบบทดสอบ 1 ฉบับที่มีคุณภาพด้านความเที่ยง แต่หากพิจารณาถึงคุณลักษณะเฉพาะด้านความเป็นคู่ขนานเพิ่มเติม พบว่า ไม่มีแบบทดสอบฉบับใดที่มีคุณสมบัติของคุณลักษณะเฉพาะของความเป็นคู่ขนาน

ตอนที่ 3 ผลตรวจสอบความถูกต้องของโปรแกรมการคัดเลือกข้อสอบด้านการใช้ข้อสอบซ้ำระหว่างแบบทดสอบคู่ขนาน (test overlap) ตามตำแหน่งความสามารถของผู้สอบของแบบทดสอบฉบับอ้างอิง

ผลตรวจสอบความถูกต้องของโปรแกรมการคัดเลือกข้อสอบด้านการใช้ข้อสอบซ้ำระหว่างแบบทดสอบคู่ขนาน (test overlap) ตามค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง ณ ตำแหน่งความสามารถ 2, 3 และ 4 ตำแหน่ง พบว่า

กรณีที่ 1 กำหนดค่าฟังก์ชันสารสนเทศของแบบทดสอบ ณ ตำแหน่งความสามารถของผู้สอบ 2 ตำแหน่ง คือ 0.00 และ 1.50 พบว่าแบบทดสอบ 4 ฉบับที่มีค่าความเที่ยงใกล้เคียงกับแบบทดสอบฉบับอ้างอิง แต่มีแบบทดสอบ จำนวน 1 ฉบับเท่านั้นที่มีอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบเฉลี่ย (average test overlap rate: $\bar{T}$) เท่ากับ 0.070 ซึ่งเป็นอัตราการใช้ข้อสอบเฉลี่ยที่ต่ำกว่าเกณฑ์ที่กำหนด คือ 0.250 นั่นคือ

แบบทดสอบ จำนวน 1 ฉบับที่สร้างขึ้นจากโปรแกรมการคัดเลือกข้อสอบมีสัดส่วนของการใช้ร่วมกัน ข้อสอบไม่เกินเกณฑ์ที่กำหนด ส่วนแบบทดสอบที่เหลือ จำนวน 3 ฉบับมีอัตราการใช้ข้อสอบซ้ำระหว่าง แบบทดสอบเฉลี่ยสูงกว่าเกณฑ์ที่กำหนด โดยมีอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบเฉลี่ย เท่ากับ 0.320, 0.369 และ 0.411 ตามลำดับ

กรณีที่ 2 กำหนดค่าฟังก์ชันสารสนเทศของแบบทดสอบ ณ ตำแหน่งความสามารถของผู้สอบ 3 ตำแหน่ง คือ -1.50, 0.00 และ 1.50 พบว่าแบบทดสอบ 2 ฉบับที่มีค่าความเที่ยงใกล้เคียงกับแบบทดสอบฉบับ อ้างอิง แต่แบบทดสอบ จำนวน 1 ฉบับเท่านั้นที่มีอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบเฉลี่ย (average test overlap rate: $\bar{T}$) เท่ากับ 0.070 ซึ่งเป็นอัตราการใช้ข้อสอบเฉลี่ยที่ต่ำกว่าเกณฑ์ที่กำหนด คือ 0.25 นั่นคือ แบบทดสอบ จำนวน 1 ฉบับที่สร้างขึ้นจากโปรแกรมการคัดเลือกข้อสอบมีสัดส่วนของการใช้ร่วมกัน ข้อสอบไม่เกินเกณฑ์ที่กำหนด ส่วนแบบทดสอบที่เหลืออีก 1 ฉบับมีอัตราการใช้ข้อสอบซ้ำระหว่าง แบบทดสอบเฉลี่ยสูงกว่าเกณฑ์ที่กำหนด โดยมีอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบเฉลี่ย เท่ากับ 0.430

กรณีที่ 3 กำหนดค่าฟังก์ชันสารสนเทศของแบบทดสอบ ณ ตำแหน่งความสามารถของผู้สอบ 4 ตำแหน่ง คือ -1.50, 0.00, 1.50 และ 2.50 พบว่าแบบทดสอบ 1 ฉบับที่มีค่าความเที่ยงใกล้เคียงกับแบบทดสอบ ฉบับอ้างอิง และมีอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบเฉลี่ย (average test overlap rate: $\bar{T}$) เท่ากับ 0.070 ซึ่งเป็นอัตราการใช้ข้อสอบเฉลี่ยที่ต่ำกว่าเกณฑ์ที่กำหนด คือ 0.25 นั่นคือแบบทดสอบ 1 ฉบับที่สร้างขึ้นจากโปรแกรมการคัดเลือกข้อสอบมีสัดส่วนของการใช้ร่วมกันข้อสอบไม่เกินเกณฑ์ที่กำหนด

จากผลการวิจัยข้างต้นสามารถสรุปผลโดยภาพรวมที่แสดงถึงผลตรวจสอบความตรง ความเที่ยง และความถูกต้องของ โปรแกรมการคัดเลือกข้อสอบด้านคุณลักษณะเฉพาะของความเป็นคู่ขนาน (parallelism specification) และด้านการใช้ข้อสอบซ้ำระหว่างแบบทดสอบคู่ขนาน (test overlap) ตาม ตำแหน่งความสามารถของผู้สอบของแบบทดสอบฉบับอ้างอิง ดังตาราง 5.1

ตาราง 5.1 สรุปผลตรวจสอบความตรง ความเที่ยงของแบบทดสอบ และความถูกต้องของโปรแกรมการคัดเลือกข้อสอบด้านคุณลักษณะเฉพาะของความเป็นคู่ขนาน (parallelism specification) และด้านการใช้ข้อสอบซ้ำระหว่างแบบทดสอบคู่ขนาน (test overlap) ตามตำแหน่งความสามารถของผู้สอบของแบบทดสอบฉบับอ้างอิง

ตำแหน่ง ความสามารถ ของผู้สอบ ของโปรแกรม	2 ตำแหน่ง		3 ตำแหน่ง		4 ตำแหน่ง	
	จำนวนแบบทดสอบ					
	สร้างขึ้น	ผ่านเกณฑ์	สร้างขึ้น	ผ่านเกณฑ์	สร้างขึ้น	ผ่านเกณฑ์
ความตรง	5	5	5	5	5	5
ความเที่ยง	5	4	5	2	5	1
ความเป็นคู่ขนาน	5	2	5	-	5	-
อัตราการใช้ข้อสอบซ้ำ ระหว่างแบบทดสอบเฉลี่ย	5	1	5	1	5	1

อภิปรายผลการวิจัย

การศึกษาในครั้งนี้มีวัตถุประสงค์หลักเพื่อสร้างแบบทดสอบคู่ขนานจากคลังข้อสอบที่มีคุณภาพด้านความเที่ยงและความตรงตามทฤษฎีการตอบข้อสอบ จากผลการศึกษาครั้งนี้สามารถนำมาอภิปรายผลเกี่ยวกับความถูกต้องของโปรแกรมการคัดเลือกข้อสอบที่ประยุกต์ใช้จากชุดคำสั่ง Ip_Solve Version 5.5 ในโปรแกรม R Studio เพื่อคัดเลือกข้อสอบจากคลังข้อสอบ ตามลำดับดังนี้

1. โปรแกรมการคัดเลือกข้อสอบ และความตรงตามเนื้อหา

ผลของโปรแกรมคัดเลือกข้อสอบจากคลังข้อสอบสำหรับการวิจัยครั้งนี้ พบว่าสามารถสร้างแบบทดสอบทั้งหมดจำนวน 15 ฉบับ ซึ่งจำนวนแบบทดสอบทั้งหมดนั้นมาจากโปรแกรมการคัดเลือกข้อสอบที่มีการกำหนดจำนวนแบบทดสอบที่ต้องการตั้งแต่ 1 - 5 ฉบับ ในแต่ละเงื่อนไขของค่าฟังก์ชันสารสนเทศของแบบทดสอบตามฉบับอ้างอิง ณ ตำแหน่งความสามารถของผู้สอบ 2 - 4 ตำแหน่ง ผลการวิจัยพบว่าแบบทดสอบที่สร้างขึ้นมีจำนวนข้อสอบในแต่ละเนื้อหาการวัดตรงกับแบบทดสอบฉบับอ้างอิง นั่นคือแบบทดสอบที่สร้างขึ้นมีคุณภาพด้านความตรงตามเนื้อหา (content validity) ซึ่งเป็นการแสดงถึงข้อสอบที่ถูกคัดเลือกนั้นวัดเนื้อหาโดยตรงตามตารางกำหนดลักษณะเฉพาะรายวิชา (table of specification) และจำนวนข้อสอบในแต่ละเนื้อหาการวัดตรงตามแบบทดสอบฉบับอ้างอิง จากผลการวิจัยแสดงให้เห็นว่าโปรแกรมการคัดเลือกข้อสอบที่ใช้ชุดคำสั่ง Ip_Solve Version 5.5 ของโปรแกรม R Studio มีประสิทธิภาพ

ในการคัดเลือกข้อสอบ เนื่องจากเป็นโปรแกรมที่ใช้กระบวนการแก้ปัญหาทางคณิตศาสตร์ภายใต้หลักการของการตัดสินใจเลือกตัวแปรทางคณิตศาสตร์ หลักการความคลาดเคลื่อนน้อยที่สุด และนำมาแปลงเป็นเงื่อนไขของการสุ่มข้อสอบเพื่อสร้างเป็นแบบทดสอบตามเงื่อนไขเชิงปริมาณระดับชุดข้อสอบด้วยการกำหนดสัดส่วนของจำนวนข้อสอบระดับต่ำสุดและระดับสูงสุดในแต่ละเนื้อหาการวัด ดังสมการ $\sum x_i \leq n_c^{\max}$, $\sum i \in v_c \geq n_c^{\min}$ ซึ่งผลการวิจัยครั้งนี้สนับสนุนแนวคิดของ van der Linden (2005) ที่กล่าวว่าการออกแบบการสร้างแบบทดสอบแนวใหม่สถานศึกษาต้องมีคลังข้อสอบ (item bank) และใช้ประโยชน์จากคลังข้อสอบให้เกิดประสิทธิภาพ คือ 1) รวบรวมข้อสอบเพื่อสร้างเป็นแบบทดสอบจำนวนมาก 2) ควบคุมข้อสอบให้ตรงกับฉบับอ้างอิง 3) คัดเลือกข้อสอบที่เหมาะสมกับระดับความสามารถของผู้สอบ 4) วัดประเมินเป็นรายบุคคลโดยใช้สารสนเทศของคลังข้อสอบ และ 4) ปรับเทียบคะแนนสอบระหว่างแบบทดสอบต่างฉบับ ณ ตำแหน่งความสามารถของผู้สอบระดับเดียวกัน รวมทั้งสอดคล้องกับ Lin (2008) ทำการศึกษาผลการจัดแบบทดสอบอัตโนมัติโดยใช้โมเดลเชิงเส้นแบบถ่วงน้ำหนักค่าเบี่ยงเบน (weighted deviations model: WDM) พบว่าโมเดลเชิงเส้นแบบ WDM สามารถสร้างแบบทดสอบ จำนวน 10 ฉบับ ที่มีเนื้อหาการวัด 4 เนื้อหา และมีจำนวนข้อสอบตรงตามแบบทดสอบฉบับอ้างอิงนั้นคือแบบทดสอบที่สร้างขึ้นมีความตรงตามเนื้อหาและสะท้อนถึงความเป็นคู่ขนานของเนื้อหาการวัด

2. โปรแกรมการคัดเลือกข้อสอบ และความเที่ยงของแบบทดสอบตามทฤษฎีการตอบข้อสอบ

ผลการคัดเลือกข้อสอบด้วยโปรแกรมเพื่อสร้างเป็นแบบทดสอบที่มีค่าความเที่ยงตามทฤษฎีการตอบข้อสอบ นั้นเป็นการพิจารณาจากค่าฟังก์ชันสารสนเทศของแบบทดสอบที่ได้จากโปรแกรมมีค่าฟังก์ชันสารสนเทศสูงสุด ณ ตำแหน่งความสามารถที่ใกล้เคียงหรือตรงกับค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง โดยแบบทดสอบฉบับอ้างอิง คือแบบทดสอบรายวิชา MER 2003 ที่มีค่าฟังก์ชันสารสนเทศของแบบทดสอบสูงสุด เท่ากับ 16.80 ณ ตำแหน่งความสามารถผู้สอบที่ 0.00 นั่นคือแบบทดสอบฉบับนี้มีความแม่นยำในการประมาณค่าความสามารถของผู้สอบที่มีความสามารถในระดับปานกลาง ซึ่งผลการวิจัยพบว่า 1) เมื่อโปรแกรมกำหนดเงื่อนไขค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง ณ 2 ตำแหน่งความสามารถของผู้สอบ พบว่ามีแบบทดสอบ จำนวน 4 ฉบับที่มีค่าฟังก์ชันสารสนเทศสูงสุด ณ ตำแหน่งความสามารถที่ใกล้เคียงกับฉบับอ้างอิง 2) เมื่อโปรแกรมกำหนดเงื่อนไขค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง ณ 3 ตำแหน่งความสามารถของผู้สอบ พบว่ามีแบบทดสอบ จำนวน 2 ฉบับที่มีค่าฟังก์ชันสารสนเทศสูงสุด ณ ตำแหน่งความสามารถที่ใกล้เคียงกับฉบับอ้างอิง และ 3) เมื่อโปรแกรมกำหนดเงื่อนไขค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง ณ 4 ตำแหน่งความสามารถของผู้สอบ พบว่ามีแบบทดสอบ จำนวน 1 ฉบับที่มีค่าฟังก์ชันสารสนเทศสูงสุด ณ ตำแหน่งความสามารถที่ใกล้เคียงกับฉบับอ้างอิง จากผลการวิจัยแสดงให้เห็นว่าโปรแกรมการคัดเลือกข้อสอบที่กำหนดเงื่อนไขค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง ณ 2 ตำแหน่งความสามารถของผู้สอบทำให้สร้างแบบทดสอบสูงสุดถึง 4 ฉบับที่มีค่าความเที่ยงตามทฤษฎีการตอบข้อสอบ แต่หากเพิ่มตำแหน่งความสามารถของผู้สอบเป็น 3 และ

4 ตำแหน่งทำให้สร้างแบบทดสอบที่มีค่าความเที่ยงตามทฤษฎีการตอบข้อสอบ ได้ลดลง คือ จำนวน 2 และ 1 ฉบับตามลำดับ ซึ่งอาจเนื่องจากคุณภาพด้านความเที่ยงของแบบทดสอบเป็นการพิจารณาจากค่าฟังก์ชันสารสนเทศที่สูงสุด ณ ตำแหน่งความสามารถเพียง 1 ตำแหน่งเท่านั้นจึงทำให้โปรแกรมการคัดเลือกข้อสอบที่กำหนดเงื่อนไขค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง ณ 2 ตำแหน่งความสามารถของผู้สอบ ทำให้สร้างแบบทดสอบที่มีคุณภาพด้านความเที่ยงได้มากกว่า 3 และ 4 ตำแหน่งความสามารถ และผลการวิจัยสอดคล้องกับหลักการประมาณค่าพารามิเตอร์ของข้อสอบและแบบทดสอบตามทฤษฎีการตอบข้อสอบ ที่มีจุดมุ่งหมายเพื่อรวบรวมข้อสอบที่มีค่าพารามิเตอร์ข้อสอบด้านความยาก อำนาจจำแนก และการเดาที่เหมาะสมกับความสามารถของผู้สอบจึงทำให้ได้แบบทดสอบที่มีความถูกต้องแม่นยำสูงในการประมาณค่าความสามารถของผู้สอบในช่วงระดับความสามารถนั้นๆ (ศิริชัย กาญจนวาสี. 2555; De Ayala .2009; McDonald. 1999)

3. ความถูกต้องของโปรแกรมการคัดเลือกข้อสอบด้านคุณลักษณะเฉพาะของความเป็นคู่ขนาน

ผลการวิจัยเกี่ยวกับคุณลักษณะเฉพาะของความเป็นคู่ขนานของแบบทดสอบตามตำแหน่งค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง พบว่า เมื่อโปรแกรมการคัดเลือกข้อสอบเพื่อสร้างเป็นแบบทดสอบกำหนดจำนวนแบบทดสอบที่ต้องการสร้างทั้งหมด เท่ากับ 5 ฉบับ และกำหนดเงื่อนไขของค่าฟังก์ชันสารสนเทศของแบบทดสอบตามฉบับอ้างอิง ณ ตำแหน่งความสามารถของผู้สอบ 2 ตำแหน่ง พบว่ามีแบบทดสอบ 2 ฉบับเท่านั้นที่มีคุณลักษณะเฉพาะของความเป็นคู่ขนาน และเมื่อกำหนดเงื่อนไขของค่าฟังก์ชันสารสนเทศของแบบทดสอบตามฉบับอ้างอิง ณ ตำแหน่งความสามารถของผู้สอบ 3 และ 4 ตำแหน่ง พบว่า ไม่มีแบบทดสอบฉบับใดที่มีคุณสมบัติของคุณลักษณะเฉพาะของความเป็นคู่ขนาน จากผลการการวิจัยแสดงให้เห็นว่าโปรแกรมการคัดเลือกข้อสอบที่กำหนดเงื่อนไขค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง ณ 2 ตำแหน่งความสามารถของผู้สอบทำให้สร้างแบบทดสอบได้ 2 ฉบับที่มีคุณลักษณะเฉพาะของความเป็นคู่ขนาน แต่หากเพิ่มตำแหน่งความสามารถของผู้สอบเป็น 3 และ 4 ตำแหน่งทำให้โปรแกรมไม่สามารถสร้างแบบทดสอบที่มีคุณลักษณะเฉพาะของความเป็นคู่ขนาน ซึ่งสอดคล้องกับ Lin (2008) ทำการศึกษาผลการจัดแบบทดสอบอัตโนมัติโดยใช้โมเดลเชิงเส้นแบบถ่วงน้ำหนักค่าเบี่ยงเบน (weighted deviations model: WDM) พบว่าโมเดลเชิงเส้น WDM สามารถจัดชุดแบบทดสอบตามทฤษฎีการตอบข้อสอบ ที่มีคุณลักษณะของความเป็นคู่ขนานตามค่าฟังก์ชันสารสนเทศของแบบทดสอบที่ 2 ตำแหน่งความสามารถของผู้สอบได้ถูกต้องมากกว่าการจัดชุดแบบทดสอบที่มีคุณลักษณะของความเป็นคู่ขนานตามค่าฟังก์ชันสารสนเทศของแบบทดสอบที่ทุกตำแหน่งความสามารถ

4. ความถูกต้องของโปรแกรมการคัดเลือกข้อสอบด้านอัตราการใช้ข้อสอบชำระระหว่างแบบทดสอบ

ผลตรวจสอบความถูกต้องของโปรแกรมการคัดเลือกข้อสอบด้านการใช้ข้อสอบชำระระหว่างแบบทดสอบคู่ขนานตามค่าฟังก์ชันสารสนเทศของแบบทดสอบฉบับอ้างอิง ณ ตำแหน่งความสามารถ 2, 3 และ 4 ตำแหน่ง พบว่า 1) เมื่อโปรแกรมกำหนดค่าฟังก์ชันสารสนเทศของแบบทดสอบ ณ ตำแหน่ง

ความสามารถของผู้สอบ ณ 2 ตำแหน่ง พบว่ามีแบบทดสอบ 1 ฉบับเท่านั้นที่มีอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบเฉลี่ยต่ำกว่าเกณฑ์ที่กำหนด 2) เมื่อโปรแกรมกำหนดค่าฟังก์ชันสารสนเทศของแบบทดสอบ ณ ตำแหน่งความสามารถของผู้สอบ 3 ตำแหน่ง พบว่ามีแบบทดสอบ จำนวน 1 ฉบับเท่านั้นที่มีอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบเฉลี่ยต่ำกว่าเกณฑ์ที่กำหนด และ 3) เมื่อโปรแกรมกำหนดค่าฟังก์ชันสารสนเทศของแบบทดสอบ ณ ตำแหน่งความสามารถของผู้สอบ 4 ตำแหน่ง พบว่ามีแบบทดสอบจำนวน 1 ฉบับเท่านั้นที่มีอัตราการใช้ข้อสอบซ้ำระหว่างแบบทดสอบเฉลี่ยต่ำกว่าเกณฑ์ที่กำหนดซึ่งสอดคล้องกับ Lin (2008) ทำการศึกษาผลการจัดแบบทดสอบอัตโนมัติโดยใช้โมเดลเชิงเส้นแบบถ่วงน้ำหนักค่าเบี่ยงเบน (weighted deviations model: WDM) พบว่าโมเดลเชิงเส้น WDM สามารถจัดชุดแบบทดสอบตามทฤษฎีการตอบข้อสอบ ที่มีอัตราการใช้ข้อสอบซ้ำไม่เกินเกณฑ์ที่กำหนดตามค่าฟังก์ชันสารสนเทศของแบบทดสอบที่ 2 ตำแหน่งความสามารถของผู้สอบ ส่วนเมื่อโมเดลเชิงเส้น WDM กำหนดค่าฟังก์ชันสารสนเทศของแบบทดสอบที่ทุกตำแหน่งความสามารถจะเกิดอัตราการใช้ข้อสอบซ้ำเกินเกณฑ์ที่กำหนด

ข้อเสนอแนะ

ข้อเสนอแนะในการนำผลการวิจัยไปใช้

1. โปรแกรมการคัดเลือกข้อสอบที่สร้างขึ้นเพื่อนำไปสร้างเป็นแบบทดสอบมากกว่า 1 ฉบับที่มีคุณลักษณะเฉพาะความเป็นคู่ขนาน แบบทดสอบที่สร้างขึ้นจากโปรแกรมการคัดเลือกข้อสอบจะมีความแม่นยำในการประมาณค่าความสามารถผู้สอบที่ตำแหน่งความสามารถ 2 ตำแหน่ง คือผู้สอบที่มีความสามารถระดับปานกลาง และค่อนข้างเก่ง
2. โปรแกรมการคัดเลือกข้อสอบจะมีความถูกต้องด้านอัตราการใช้ข้อสอบซ้ำต่ำกว่าเกณฑ์ เมื่อคลังข้อสอบมีจำนวนข้อสอบที่มีค่าอำนาจจำแนกระดับต่ำ (a parameter น้อยกว่า 1.00) จำนวนมาก

ข้อเสนอแนะในการทำวิจัยครั้งต่อไป

1. ควรสร้างโปรแกรมการคัดเลือกข้อสอบโดยใช้กระบวนการแก้ปัญหาทางคณิตศาสตร์ภายใต้หลักการของการตัดสินใจเลือกตัวแปรทางคณิตศาสตร์ที่เพิ่มเงื่อนไขเกี่ยวกับการกำหนดข้อสอบที่ไม่สามารถจัดชุดร่วมกันในแบบทดสอบฉบับเดียวกัน
2. คลังข้อสอบจากการวิจัยครั้งนี้เป็นคลังข้อสอบที่ได้การจำลองข้อมูลการสอบที่มีคำตอบแบบ 0 และ 1 โดยข้อมูลการสอบที่จำลองขึ้นอยู่บนพื้นฐานของผลการทดสอบจากรายวิชา MER 2003 ความรู้เบื้องต้นเกี่ยวกับสถิติและวิจัยทางการศึกษา ควรทำการวิจัยเพื่อยืนยันความถูกต้องของโปรแกรมการคัดเลือกข้อสอบจากคลังข้อสอบจริง



รายการอ้างอิง

ภาษาไทย

- กระทรวงศึกษาธิการ. (2554). *แนวปฏิบัติการวัดและประเมินผลการเรียนรู้ ตามหลักสูตรแกนกลาง การศึกษาขั้นพื้นฐาน พุทธศักราช 2551*. กรุงเทพมหานคร: โรงพิมพ์ชุมนุมสหกรณ์การเกษตร แห่งประเทศไทยจำกัด
- กฤษมันต์ วัฒนารงค์. (2556). กฎหมายการอุดมศึกษา ตอนที่ 1 นสพ. ไทยรัฐ ฉบับวันจันทร์ที่ 16 ตุลาคม พ.ศ. 2556. ค้นเมื่อ 20 เมษายน 2558 จาก <http://www.thairath.co.th/content/edu/376425>
- กิม ไชยแสนสุข. (2553). ข่าวรวมคำแหง ปีที่ 40 ฉบับที่ 4 วันที่ 10-16 พฤษภาคม 2553.
- นราทิพย์ จันสกุล. (2552). *แผนแบบการทดลองและการวิเคราะห์ข้อมูลโดยใช้โปรแกรม R*. สำนักพิมพ์ แห่งจุฬาลงกรณ์มหาวิทยาลัย.
- วิโรจน์ อรุณมานะกุล. (2559). สถิติและการใช้โปรแกรม R. ค้นเมื่อ 25 กันยายน 2559, จาก <http://pioneer.chula.ac.th/~awirote/courses/res-tech-ling/statistics-and-r.pdf>
- วุฒิสกดิ์ ลาภเจริญทรัพย์. (2556). ข่าวรวมคำแหง ฉบับที่ 31. วันที่ 11-17 พฤศจิกายน 2556.
- ศิริชัย กาญจนวาสี. (2555). *ทฤษฎีการทดสอบแนวใหม่*. พิมพ์ครั้งที่ 4. กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์ มหาวิทยาลัย.
- สมศักดิ์ ลิลา. (2539). *การพัฒนาระบบคลังข้อสอบเพื่อการเรียนการสอนด้วยไมโครคอมพิวเตอร์*. ปรินญา นิพนธ์ปริญญาการศึกษาดุสิตบัณฑิต สาขาการวิจัยและพัฒนาหลักสูตร บัณฑิตวิทยาลัย มหาวิทยาลัย ศรีนครินทรวิโรฒ.
- อุทัย บุญประเสริฐ. (2540). *หลักและแนวทางในการสร้างคลังข้อสอบสำหรับโรงเรียน*. กรุงเทพมหานคร: คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.

ภาษาอังกฤษ

- American Institutes for Research. (2011). Delaware Comprehensive Assessment System 2010–2011: Computer Adaptive Testing Volume 2 Test Development. State of Delaware: American Institutes for Research. Retrieved May 1, 2014 from https://www.doe.k12.de.us/dcas/files/Vol2_DCAS_TestDev.pdf
- Baker, Frank, B. (2001). *The Basics of Item Response Theory*. United States of America: ERIC Clearinghouse on Assessment and Evaluation.

- Chang, H.-H., and Zhang, J. (2002). Hypergeometric family and item overlap rates in computerized adaptive Testing. *Psychometrika*, 67(3), 387-398. Retrieved April 30, 2016 from http://library.lib.ru.ac.th/search~S0*thx?/Xjournal+of+educational+measurement
- Chen, P.-H., Chang, H.-H. & Wu, H. (2012). Item Selection for the Development of Parallel Forms from an IRT-Based Seed Test Using a Sampling and Classification Approach. *Educational and Psychological Measurement*. May 16, 933-953, , Retrieved April 30, 2017 from <http://epm.sagepub.com> .DOI:10.1177/0013164412443688
- Chen, S., Ankenmann, R.D. & Lei, P. (2003). The Relation Between Item Exposure and Test Overlap in Computerized Adaptive Testing. Adaptive Testing. *Journal of Educational Measurement*, 40(2), 129-145.
- Chen, S., & Lei, P. (2009). Investigating the relationship between item exposure and test overlap: Item sharing And item pooling. *British Journal of Mathematical and Statistical Psychology*, 63, 205-226. Retrieved August 30, 2016 from <http://www.bpsjournals.co.uk>. DOI: 10.1348/000711009X430906
- Chen, S., Ankenmann, R.D., and Spray, J.A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Psychometrika*, 67, 387-398. Retrieved April 30, 2016 from <https://link.springer.com/article/10.1007%2F02294991>
- Chuesathuchon, Chaowprapha. (2008). *Item banking with Rasch measurement: an example for primary mathematics in Thailand*. Pert: Cowan University.
- Cohen, J. C., & Swerdlik, D.P. (2010). *Psychological testing and Assessment: An Introduction to Test and Measurement*. Boston: McGraw-Hill.
- Davey, T. & Hendrickson, A. (2010). *Classical versus IRT Statistical Test Specifications for Building Test Forms*. Paper presented at the Annual Conference of the National Council for Measurement in Education, Denver, Colorado. Retrieved April 30, 2016 from <https://research.collegeboard.org>.
- De Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York: The Guilford Press.
- Diol, Q. and van der Linden. (2013). Automated Test Assembly Using Ip_Solve Version 5.5 in R. *Applied Psychological Measurement*, 35(5): 398-409. Retrieved June 9, 2016 from <https://apm.sagepub.com>. DOI: 10.1170/01046621610392211

- Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement*, 31(5), 457-459.
- Han, K. T., & Hambleton, R. K. (2007). User's Manual: WinGen (*Center for Educational Assessment Report No. 642*). Amherst, MA: University of Massachusetts, School of Education.
- Hambleton, R. K. (1984). Validating the test scores. In *A Guide to Criterion-Referenced Test Construction*. Berk, R.A. (Ed). p. 199-230. Baltimore: The Johns Hopkins University Press.
- Hambleton, Ronald K. and Swaminathan, H.. (1996). *ITEM RESPONSE THEORY: Principles and Applications*. Norwell: Kluwer Academic Publishers.
- He, Q. and Tymms, P. (2005). A computer-assisted test design and diagnosis system for use by classroom teachers. *Journal of Computer Assisted Learning*. 21, 419–429. Retrieved June 9, 2014 from <http://web.a.ebscohost.com/ehost/pdfviewer/pdfviewer?vid=7&sid=d56a7abf-9339-4e41-95f0-febcd256e25e%40sessionmgr4005&hid=4204>.
- Kuhn, Jörg-Tobias and Kiefer, Thomas. (2013). Optimal Test Assembly in Practice: The design of the Australian Educational Standards Assessment in Mathematics. *Zeitschrift für Psychologie*, 221, 190-200.
- Lin, C-J. (2010). Controlling Test Overlap Rate in Automated Test Assembly of Multiple Equivalent Test Forms. *Journal of Technology, Learning, and Assessment*, 8(3). Retrieved June 9, 2016 from <http://www.jtla.org>.
- Lin, Chuan-Ju. (2008). Comparisons between Classical Test Theory and Item response Theory in Automated Test Assembly of Parallel Forms. *Journal of Technology, Learning, and Assessment*. 6(8), 1–42.
- Luecht, R. M. (1998). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement*, 22, 224-236
- McAlpine, M. (2002). *Design Requirements of a Databank*. England: Computer Assisted Assessment Centre.
- McDonald, Roderick. (2009). *Test theory: a unified treatment*. New Jersey: Lawrence Erlbaum Associates.
- Reckase, M.D..(2010). Designing item pools to optimize the functioning of a computerized adaptive test. *Psychological test and assessment modeling*. 52, 2: 127-141.
- Turner, R and Carlson, L. Indexes of Item-Objective Congruence for Multidimensional Items. *International Journal of Testing*, 3(2): 163-171.

- van der Linden, W.J.. (2005). *Linear Models for Optimal Test Design*. New York. Springer Science + Business Media.
- Veldkamp, B.P. and Van der Linden, W.J. (2010). *Designing Item Pools for Adaptive Testing*. New York: Springer.
- Ward, A.W. (1994). Guidelines for the Development of Item Banks. *Educational Measurement: Issues and Practice*. Spring : 34-39.
- Weiss, D.J. (2011). *Item Banking, Test Development, and Test Delivery*. Washington DC: American Psychological Association.
- Wright, B. D. and Stone, M. (1999). *Measurement Essentials*. Wilmington: Wide Range.
- Ying and Lei. (2010). Investigating the relationship between item exposure and test overlap: Item Sharing and item pooling. *British Journal of Mathematical and Statistical Psychology*. 63: 205-226.
DOI:10.1348/00071009X430906. Retrieved from www.bpsjournals.co.uk 23 June 2017
-

ภาคผนวก



RESEARCH ETHICS COMMITTEE

Ramkhamhaeng University, 2086, Ramkhamhaeng Road,
Hua-Mak, Bangkok, 10240, Thailand Tel 662-310-8000 ext 8018

Approval of Documents related to Study Protocol

The Research Ethics Committee of Ramkhamhaeng University, Thailand, has approved the following study which is to be carried out in compliance with the International guidelines for human research protection as Declaration of Helsinki, The Belmont Report, CIOMS Guideline and International Conference on Harmonization in Good Clinical Practice (ICH-GCP)

Study Title : The Reliability and Validity of Test Construction from Item Bank Based on Item Response Theory

Study Code : xd-003/61

Principal Investigator : Assistant Professor Suwimon Kritkharuehart, Ph.D.

Study Center : Ramkhamhaeng University

Document Approval :

1. Protocol version 1.0/020461
2. Information sheet version 1.1/060461 based on version 1.0/020461
3. Consent form version 1.1/060461 based on version 1.0/020461
4. Question version 1.0/020461

(Assistant Professor Boonchan Thongprayoon, Ph.D.)
Chairman of the Research Ethics Committee of
Ramkhamhaeng University, Thailand

(Assistant Professor Tirata Bhasavanija, Ph.D.)
Secretary of the Research Ethics Committee of
Ramkhamhaeng University, Thailand

Approval code: RU-REC/xd-003/61

Approval Date: 07/04/2018

Expiry Date: 07/04/2019

Approval granted is subject to the following conditions: (see back page of this Certificate)

แบบทดสอบกระบวนวิชา MER 2003
ความรู้เบื้องต้นเกี่ยวกับสถิติและวิจัยทางการศึกษา



คำชี้แจงการทำข้อสอบ

- 1) เป็นแบบทดสอบปรนัย 4 ตัวเลือก จำนวน 100 ข้อ 15 หน้า
- 2) เลือกคำตอบถูกต้องเพียงข้อเดียว

1. ข้อใดไม่ใช่จุดมุ่งหมายของวิธีการทางวิทยาศาสตร์เพื่อการวิจัย

1. บรรยายเหตุการณ์หรือปรากฏการณ์
2. ควบคุมเหตุการณ์หรือปรากฏการณ์
3. ศึกษาความสัมพันธ์ร่วมกันของเหตุการณ์
4. ทำนายเหตุการณ์หรือปรากฏการณ์ที่ศึกษา

คำสั่ง ให้ใช้ข้อมูลต่อไปนี้ตอบคำถามข้อ 2. -- 3.

ปัญหาหลักของนักเรียนในชนบท คือ สภาพและสิ่งแวดล้อมไม่เอื้อต่อการส่งเสริมความเข้าใจในการแก้โจทย์คณิตศาสตร์ ครูนาราจึงทำการทดสอบพบว่า นักเรียนกลุ่มที่สอนไม่สามารถแก้โจทย์ปัญหาได้ ครูนาราจึงทำการสรุปว่าสภาพและสิ่งแวดล้อมไม่เอื้อต่อการส่งเสริมความเข้าใจในการแก้โจทย์คณิตศาสตร์

2. จากข้อมูลข้างต้น ข้อใดคือวิธีค้นหาความรู้ของครูนารา

- | | |
|--------------------------------|-----------------------------------|
| 1. การใช้เหตุผลแบบอุปนัย | 2. การใช้เหตุผลแบบอุปมาน |
| 3. การใช้เหตุผลจากความสอดคล้อง | 4. การใช้เหตุผลแบบอาศัยประสบการณ์ |

3. ข้อใดคือข้อผิดพลาดของการสรุปของครูนารา

- | | |
|-----------------------------|---------------------------------------|
| 1. ขาดข้อเท็จจริงหลัก | 2. เลือกข้อเท็จจริงหลักไม่ถูกต้อง |
| 3. ข้อเท็จจริงรองไม่เพียงพอ | 4. เจาะจงวิเคราะห์เฉพาะข้อเท็จจริงรอง |

4. ข้อใดกล่าวถึงความหมายของ “การวิจัยทางการศึกษา” ได้ถูกต้อง

1. การบรรยายเหตุการณ์ที่เกิดขึ้นอย่างสมเหตุสมผล
2. วิธีการที่เป็นระบบเพื่อการแสวงหาความจริงของเหตุการณ์ที่เกิดขึ้น
3. การใช้ความเข้าใจของนักวิจัยอธิบายเหตุการณ์ที่เกิดขึ้นอย่างถูกต้อง
4. กระบวนการสรุปอ้างอิงเหตุการณ์ที่เกิดขึ้นจากกลุ่มตัวอย่างไปยังประชากร

5. ข้อใดแสดงถึงการใช้กระบวนการวิจัยเพื่อตอบปัญหาการวิจัย

1. สามารถปรับปรุงผลการวิจัยได้ 2. สามารถเก็บข้อมูลจากเหตุผลของนักวิจัย
3. สามารถตรวจสอบความถูกต้องได้ 4. สามารถกำหนดสมมติฐานตามข้อมูลที่ได้

6. การกระทำใดแสดงถึงผู้มีจรรยาบรรณนักวิจัย

1. เผยแพร่ผลงานวิจัย โดยเห็นแก่ประโยชน์ของนักวิจัยเป็นที่ตั้ง
2. ปรับแก้ข้อมูลหากพบว่าข้อมูลนั้นกระทบกับผู้ให้ข้อมูล
3. การประหยัดเวลาเมื่อต้องเก็บข้อมูลจากการสังเกต
4. เก็บรักษาคำตอบของผู้ให้ข้อมูลเป็นความลับ

7. ข้อใดไม่ใช่ประโยชน์ของการวิจัยทางสังคมศาสตร์

1. ทราบถึงการดำเนินงานตามนโยบายทางการศึกษา
2. ช่วยเตรียมผู้เรียนให้พร้อมกับการพัฒนา
3. ผู้เรียนไม่มั่งายเชื่อในสิ่งที่พิสูจน์ไม่ได้
4. ผู้วิจัยได้พัฒนาความคิดอย่างเป็นระบบ

คำสั่ง ให้ใช้ตัวเลือกที่กำหนดให้ สำหรับการตอบคำถามข้อ 8. - 10.

1. การอธิบาย 2. การทำนาย 3. การบรรยาย 4. การควบคุม

8. ครูประจักษ์นำข้อค้นพบของรูปแบบการจัดกิจกรรมการเรียนรู้มาป้องกันและส่งเสริมความรับผิดชอบของนักเรียน ป. 2 ข้อใดคือจุดมุ่งหมายของการวิจัยของครูประจักษ์

9. ครูหน้อยต้องการให้รายละเอียดของคุณลักษณะของผู้สำเร็จการศึกษาระดับ ปวส.ตามความต้องการของผู้ประกอบการ ข้อใดคือจุดมุ่งหมายของการวิจัยของครูหน้อย

10. ครูต้องต้องการศึกษาถึงสาเหตุและผลของการให้เหตุผลเชิงวิทยาศาสตร์ต่อตัวแทนความคิด เรื่อง การเกิดน้ำขึ้นน้ำลง ข้อใดคือจุดมุ่งหมายของการวิจัยของครูต้อง

คำสั่ง ให้ใช้ตัวเลือกที่กำหนดให้ สำหรับการตอบคำถามข้อ 11. - 13.

1. การวิจัยคุณภาพ 2. การวิจัยปริมาณ 3. การวิจัยพื้นฐาน 4. การวิจัยเชิงประเมิน

11. การศึกษาพฤติกรรมเกมการตัดสินใจของนักเรียน ป. 3 โดยใช้การบรรยายข้อมูลจากการสังเกต งานวิจัยนี้จัดเป็นการวิจัยประเภทใด

12. การพัฒนารูปแบบการบริหารงานวิชาการของสถานศึกษาขั้นพื้นฐานสู่การเตรียมพร้อม
ผู้เรียนสู่ความป็นพลเมืองโลก งานวิจัยนี้จัดเป็นการวิจัยประเภทใด

13. การติดตามโครงการการจัดการเรียนรู้ด้วยการใช้นวัตกรรมเพื่อสร้างสรรค์ด้วยปัญญา
งานวิจัยนี้จัดเป็นการวิจัยประเภทใด

คำสั่ง ให้ใช้ตัวเลือกที่กำหนดให้ สำหรับการตอบคำถามข้อ 14. – 16.

- | | |
|----------------------|-------------------------|
| 1. การวิจัยประยุกต์ | 2. การวิจัยเชิงบุกเบิก |
| 3. การศึกษาเฉพาะกรณี | 4. การศึกษาแบบต่อเนื่อง |

14. นวัตกรรมสร้างสรรค์สำหรับบัณฑิตพันธุ์ใหม่สู่ความป็นพลเมืองโลก จัดเป็น
การวิจัยประเภทใด

15. การศึกษาผลกระทบของการใช้โปรแกรมไลน์ (Line) ต่อการเรียนรู้ของนักเรียนชั้น

ป. 6 ในห้องเรียนแบบเรียนร่วม โรงเรียนมีสุขวิทยา จัดเป็นการวิจัยประเภทใด

16. การศึกษากลยุทธ์การเรียนรู้ภาษาจีนของนักเรียนชั้น ม. 3 ประจำปีการศึกษา 2559 โดยผู้วิจัย
ทำการสรุปผลการศึกษาเป็นรายภาคการศึกษา จัดเป็นการวิจัยประเภทใด

คำสั่ง ให้ตัวเลือกที่กำหนดให้ สำหรับการตอบคำถามข้อ 17. – 19.

1. การวิจัยปริมาณ 2. การวิจัยคุณภาพ 3. การวิจัยเชิงทดลอง 4. การวิจัยปฏิบัติการ

17. ครูชินทร์ต้องการนำข้อค้นพบจากการวิจัยไปแก้ปัญหาในห้องเรียนของตนเอง
ครูชินทร์ควรทำวิจัยประเภทใด

18. ครูน้อยเน้นการเก็บข้อมูลด้านพฤติกรรมที่วัดได้ และกลุ่มตัวอย่างได้มาจากการ
สุ่มตัวอย่างโดยอาศัยความน่าจะเป็น ครูน้อยควรทำวิจัยประเภทใด

19. การวิจัยประเภทใดเน้นการใช้การวิเคราะห์โดยใช้ความเป็นเหตุเป็นผล (logical analysis)
สำหรับการให้ความหมายของข้อมูล

คำสั่ง ให้ใช้ตัวเลือกที่กำหนดให้ สำหรับการตอบคำถามข้อ 20. – 24.

$$1. \quad x \quad O_2$$

$$2. \quad O_1 \times O_2$$

$$3. \quad O_1 \times O_2 \\ O_{C1} \cdot O_{C2}$$

$$4. \quad O_1 O_2 O_3 \times O_4 O_5 O_6$$

20. ผู้วิจัยต้องการเปรียบเทียบความสามารถในการฟังและพูดภาษาอังกฤษระหว่าง การสอนแบบปฏิสัมพันธ์ กับการสอนแบบบรรยาย ผู้วิจัยควรใช้รูปแบบการวิจัยใด
21. ผู้วิจัยต้องการเปรียบเทียบความคิดรวบยอดในรายวิชาเคมี ระหว่างก่อนและหลังทดลอง สอนแบบมุ่งปฏิบัติ ผู้วิจัยควรใช้รูปแบบการวิจัยใด
22. ผู้วิจัยต้องการเปรียบเทียบความคิดเชิงระบบหลังสอนด้วยชุดการเรียนรู้กับเกณฑ์
23. ผู้วิจัยต้องการศึกษาพัฒนาการทักษะการทดลอง ผู้วิจัยควรใช้รูปแบบการวิจัยใด
24. ผู้วิจัยควรเลือกใช้การวิจัยเชิงทดลองประเภทใดที่สามารถยืนยันความเป็นเหตุเป็นผล ได้ดี ที่สุด
25. งานวิจัยใดมีการจัดกระทำตัวแปร (manipulation) ในการวิจัยเชิงทดลอง
1. การศึกษาโปรแกรมฝึกที่มีต่อทักษะการเข้าสังคมของนักเรียนชั้น ม.1
 2. การศึกษาความต้องการเข้าร่วมกิจกรรมพัฒนาผู้เรียนของนักเรียนชั้น ม.1
 3. การศึกษาความคิดเห็นต่อการบริการห้องสมุดออนไลน์ของนักเรียนชั้น ม. 1
 4. การศึกษาสิ่งสนับสนุนกิจกรรมประชาธิปไตยในโรงเรียนของนักเรียนชั้น ม.1
26. ข้อใดกล่าวถูกต้องเกี่ยวกับการสุ่ม (randomization) ของการวิจัยเชิงทดลอง
1. สุ่มวิธีทดลองให้กับนักเรียนชั้น ป. 4 โรงเรียนครุณี
 2. สุ่มตัวอย่างนักเรียน 1 ห้องเรียน จากนักเรียนชั้น ป. 4 โรงเรียนครุณี
 3. สุ่มตัวอย่างจากนักเรียนชั้น ป. 4 จากโรงเรียนในเขตกรุงเทพมหานคร เขต 1
 4. สุ่มตัวอย่างนักเรียนชั้น ป. 4 จากโรงเรียนครุณีและสุ่มกลุ่มตัวอย่างเข้าวิธีทดลอง
27. วิธีการใดช่วยให้ผู้วิจัยสามารถกำหนดปัญหาการวิจัยได้
1. พิสูจน์การอธิบายของทฤษฎีในสภาพปัจจุบัน
 2. การเลือกจากเหตุการณ์ที่เกิดขึ้น
 3. การเลือกเหตุการณ์จากข่าว
 4. การเลียนแบบงานวิจัยที่ได้รางวัล

28. ข้อใดคือประเด็นปัญหาในการวิจัย

1. โจทย์ปัญหาลักษณะใดที่นักเรียนเข้าใจคลาดเคลื่อน
2. นักเรียนปัจจุบันไม่ค่อยกำหนดความมุ่งหวังในชีวิต
3. การศึกษาความสามารถในการเขียนสะกดคำยาก
4. ความพึงพอใจต่อแบบฝึกสะกดคำ

29. หัวข้อวิจัย คือ การใช้เหตุผลเชิงจริยธรรม สิ่งใดที่ช่วยทำให้หัวข้อวิจัยชัดเจนขึ้น

1. ระบุวิธีการสุ่มตัวอย่าง
2. กำหนดประเภทของการวิจัย
3. ระบุประโยชน์ของการวิจัย
4. กำหนดตัวแปรในการวิจัย

30. ข้อใดคือวัตถุประสงค์การวิจัยที่ดีของการวิจัยเรื่อง พฤติกรรมการออมของนักเรียนชั้น ป.3

1. เพื่อศึกษาสินค้าที่นักเรียนชั้น ป.3 ใช้
2. เพื่อศึกษาช่วงเวลาของการใช้จ่าย
3. เพื่อศึกษาลักษณะการใช้จ่ายของนักเรียนชั้น ป.3
4. เพื่อศึกษาสถานที่ที่มีการใช้จ่าย

คำสั่ง ให้ใช้สถานการณ์ และตัวเลือกที่กำหนดให้ สำหรับการตอบคำถามข้อ 31. – 35.

สถานการณ์ ผู้วิจัยต้องการพัฒนาความคิดสร้างสรรค์ของเด็กปฐมวัยด้วยการใช้นิทานเล่มเล็ก เขาจึงทำการค้นคว้าทฤษฎี แนวคิด และผลงานวิจัยที่เกี่ยวข้อง

- ตัวเลือก
1. ข้อความที่ 1) และ 2) ถูกต้องทั้งสองข้อความ
 2. ข้อความที่ 1) และ 2) ผิดทั้งสองข้อความ
 3. ข้อความที่ 1) ถูก ส่วนข้อความที่ 2) ผิด
 4. ข้อความที่ 1) ผิด ส่วนข้อความที่ 2) ถูก

31. ข้อความ 1) การกำหนดความหมาย “ความคิดสร้างสรรค์” คือการดำเนินการวิจัยด้วยการอ้างอิงเชิงทฤษฎี

ข้อความ 2) การศึกษาวิธีการสร้างนิทานเล่มเล็ก คือการดำเนินการวิจัยด้วยการอ้างอิงเชิงประจักษ์

32. ข้อความ 1) การอภิปรายผลการวิจัย คือการดำเนินการวิจัยด้วยการอ้างอิงเชิงทฤษฎี

ข้อความ 2) การเก็บรวบรวมข้อมูล คือการดำเนินการวิจัยด้วยการอ้างอิงเชิงประจักษ์

33. ข้อความ 1) หากอยากทราบว่าปัจจุบันเด็กปฐมวัยมีระดับความคิดสร้างสรรค์

เป็นอย่างไร ผู้วิจัยสามารถค้นหาจากพจนานุกรม

ข้อความ 2) คำสำคัญ (keyword) ในการค้นเอกสารของงานวิจัยนี้ คือ การพัฒนา

34. ข้อความ (1) การทบทวนวรรณกรรมทำให้กำหนดขอบเขตวิจัยชัดเจน

ข้อความ (2) การทบทวนวรรณกรรมทำให้มีความรู้ในการวัดความคิดสร้างสรรค์

35. ข้อความ 1) กรอบความคิดในการวิจัยคือการแสดงถึงขั้นตอนการวิจัย

ข้อความ 2) ตัวแปรที่เขียนในกรอบความคิด คือ นิทานเล่มเล็ก และความคิด
สร้างสรรค์

36. “วุฒิภาวะทางอาชีพของนักเรียนแตกต่างกันตามระดับชั้น” จัดเป็นสมมติฐานในลักษณะใด

1. เปรียบเทียบตัวแปรในประชากร 2 กลุ่มขึ้นไป 2. แสดงอิทธิพลของตัวแปร

3. แสดงความสัมพันธ์ของตัวแปร 2 ตัว 4. ทำนายปรากฏการณ์ในอนาคต

37. สมมติฐานการวิจัยใดที่ไม่กำหนดทิศทางของความแตกต่าง

1. การให้เหตุผลทางคณิตศาสตร์น่าจะไม่เท่ากัน

2. การเขียนเรียงความหลังใช้ชุดการเรียนรู้จะสูงกว่าก่อนใช้

3. ทักษะการจัดการก่อนและหลังสอนด้วยชุดฝึกน่าจะแตกต่างกัน

4. ทักษะการแก้โจทย์ปัญหาก่อนใช้แบบฝึกจะต่ำกว่าหลังใช้แบบฝึก

คำสั่ง จากข้อ 38.-41.ให้นักศึกษาพิจารณาว่าตัวเลือกที่เป็น ตัวหนาตัวใดที่ไม่ใช่ตัวแปร

38. ผลสัมฤทธิ์ทางการเรียนและความคิดเห็นหลังใช้บทเรียนคอมพิวเตอร์มัลติมีเดีย

1.

2.

รูปแบบการรู้ตนเคลื่อนไหว เรื่อง รามเกียรติ์ โดยวิธีการเรียนแบบร่วมมือ

3.

4.

39. รูปแบบการสอน STAD สามารถนำมาใช้ในการเรียนการสอนได้ เป็นรูปแบบที่

1.

พัฒนาทักษะทางสังคม และสามารถคิดแปลงได้ทุกวิชาและทุกระดับชั้น

2.

3.

4.

40. ทักษะการแก้ปัญหาและพฤติกรรมการเรียนรู้ด้วยชุดการสอนตามหลักไตรสิกขา

1. 2. 3.

โรงเรียนอนุบาลนครนายก

- 4.

41. การพัฒนา การใช้เทคโนโลยี และความคงทนในการเรียน เรื่อง การออกแบบภาพ

1. 2. 3.

สองมิติ โดยใช้วิธีการนำเสนอแบบจุดเดียวและแบบหลายจอ

- 4.

42. ตัวแปรใดไม่ใช่ตัวแปรคุณลักษณะ (attribute variable)

1. แผนการเรียน 2. การใช้งานคอมพิวเตอร์
3. สังกัดของสถานศึกษา 4. ความถนัดทางตัวเลข

43. ข้อใดคือตัวแปรต่อเนื่อง (continuous variable)

1. ลำดับการเกิด 2. ประเภทวรรณคดี 3. แหล่งที่ตั้งของบ้าน 4. รายได้ของครอบครัว

คำสั่ง ให้ใช้ข้อมูลต่อไปนี้สำหรับการตอบคำถาม 44. – 50.

การวิจัยเรื่อง การพัฒนารูปแบบการเสริมสร้างทักษะการอ่านคำศัพท์ภาษาอังกฤษ สำหรับนักเรียนชั้นประถมศึกษาปีที่ 2 – 3 ที่มีภาวะการอ่านบกพร่อง

44. ประเด็นปัญหาการวิจัยคือข้อใด

1. สภาพของนักเรียนชั้นประถมศึกษาที่มีภาวะบกพร่อง
2. ศักยภาพของนักเรียนชั้นประถมศึกษาที่มีภาวะบกพร่อง
3. ประสิทธิภาพของรูปแบบการเสริมสร้างทักษะการอ่าน
4. รูปแบบการเสริมสร้างทักษะการอ่านของนักเรียนชั้นประถมศึกษา

45. ตัวแปรอิสระ คือ

1. รูปแบบ 2. รูปแบบนักเรียน 3. รูปแบบการเสริมสร้าง 4. การพัฒนารูปแบบ

46. ตัวแปรตาม คือ

1. ทักษะการอ่านคำศัพท์ 2. การอ่าน 3. สภาพการอ่าน 4. ปัญหาการอ่าน

47. ตัวแปรแทรกซ้อน คือ

1. ระดับชั้น
2. วิธีการสอน
3. สถานที่เรียน
4. ความบกพร่อง

คำสั่ง จากข้อมูลข้างต้น ให้ใช้ตัวเลือกต่อไปนี้สำหรับการตอบคำถาม 48. – 49.

1. ทักษะการอ่านเป็นผลจากการใช้รูปแบบที่ผู้วิจัยพัฒนาขึ้น
2. ความเหมาะสมของรูปแบบการเสริมสร้างทักษะการอ่าน
3. การทบทวนวรรณกรรมที่เหมาะสมกับนักเรียนชั้นประถม
4. การสรุปผลไปยังนักเรียนชั้นป. 2 – 3 ที่มีภาวะการอ่านบกพร่อง

48. ข้อใดคือสิ่งที่บ่งบอกถึงความตรงภายใน (internal validity) ของงานวิจัย

49. ข้อใดคือสิ่งที่บ่งบอกถึงความตรงภายนอก (external validity) ของงานวิจัย

50. การวิจัยเรื่องนี้เป็น การวิจัยประเภทใด

1. การวิจัยเชิงความสัมพันธ์
2. การวิจัยเชิงวินิจฉัย
3. การวิจัยเชิงประเมิน
4. การศึกษาเฉพาะกรณี

คำสั่ง ให้ใช้ข้อมูลต่อไปนี้สำหรับการตอบคำถาม 51. – 54.

การวิจัยเรื่อง ผลการใช้ชุดกิจกรรมแบบ STEM ที่มีต่อการแก้ปัญหาทางวิทยาศาสตร์ของนักเรียนชั้นประถมศึกษาปีที่ 4

51. สมมติฐานการวิจัย คือ

1. การแก้ปัญหาทางวิทยาศาสตร์มีความสัมพันธ์กับชุดกิจกรรม
2. การแก้ปัญหาทางวิทยาศาสตร์ของผู้เรียนแต่ละคนแตกต่างกัน
3. การแก้ปัญหาทางวิทยาศาสตร์แตกต่างกันระหว่างสอนด้วยชุดกิจกรรมกับแบบปกติ
4. การแก้ปัญหาทางวิทยาศาสตร์ของผู้เรียนแต่ละคนสัมพันธ์กับการสอนด้วยชุดกิจกรรม

52. ข้อใดเป็นตัวแปรอิสระ

1. ผลการสอน
2. ชุดการเรียนรู้
3. นักเรียนชั้น ป. 4
4. ปัญหาวิทยาศาสตร์

53. ข้อใดเป็นตัวแปรตาม

1. วิชาวิทยาศาสตร์
2. ทักษะการทดลอง
3. ทักษะกระบวนการ
4. การแก้ปัญหาทางวิทยาศาสตร์

54. ข้อใดเป็นตัวแปรแทรกซ้อน

- | | |
|--------------------------------------|------------------------------|
| 1. การคิดอย่างมีวิจารณญาณของนักเรียน | 2. ระดับการศึกษาของผู้ปกครอง |
| 3. นิัยการเรียนของครู | 4. ระดับชั้น |

คำสั่ง ให้ให้ตัวเลือกต่อไปนี้สำหรับการตอบคำถาม 55. – 59.

1. มาตรฐานบัญญัติ 2. มาตรฐานระดับ 3. มาตรฐานตรภาค 4. มาตรฐานอัตราส่วน

55. ประเภทของคอมพิวเตอร์ อยู่ในมาตรใด

56. อันดับของผลการสอบ อยู่ในมาตรใด

57. เวลาที่ใช้ในการแก้โจทย์ปัญหา อยู่ในมาตรใด

58. ระดับการศึกษา อยู่ในมาตรใด

59. รายได้ของครอบครัว อยู่ในมาตรใด

60. ข้อใดกล่าวไม่ถูกต้องเกี่ยวกับการให้ความหมายของตัวแปรในการวิจัย

1. ความคิดเชิงกลยุทธ์ เป็นตัวแปรที่ต้องเขียนนิยามศัพท์
2. มหาวิทยาลัยเอกชน เป็นตัวแปรที่ต้องเขียนนิยามศัพท์
3. กลุ่มสาระการเรียนรู้ เป็นตัวแปรที่ไม่ต้องเขียนนิยามศัพท์
4. ความเป็นผู้บริหารมืออาชีพ เป็นตัวแปรที่ต้องเขียนนิยามศัพท์

61. ความสัมพันธ์ระหว่างการควบคุมอารมณ์กับการตระหนักรู้ทางสังคมของนิสิตปริญญาตรี

คณะศึกษาศาสตร์ มหาวิทยาลัยรามคำแหง ชื่อเรื่องข้างต้นนักวิจัยควรให้ความหมายคำใด

1. ปริญญาตรี 2. ความสัมพันธ์ 3. คณะศึกษาศาสตร์ 4. การควบคุมอารมณ์

62. การศึกษาการปรับตัวของนักศึกษาในระดับปริญญาตรีชั้นปีที่ 1 ในภาคการศึกษาแรก

ชื่อเรื่องดังกล่าว ผู้วิจัยต้องดำเนินการออกแบบการวิจัยใดเป็นขั้นตอนแรก

- | | |
|------------------------------|--------------------------------|
| 1. กำหนดขอบเขตการวิจัย | 2. กำหนดตัวแปรที่เกี่ยวข้อง |
| 3. พิจารณาข้อจำกัดในการวิจัย | 4. การออกแบบวิธีดำเนินการวิจัย |

คำสั่ง ให้ใช้ข้อมูลต่อไปนี้ และตัวเลือกที่กำหนดให้สำหรับตอบคำถาม 63. - 65.

การพัฒนาสื่อการออกแบบและสร้างหุ่นยนต์ที่มีต่อผลการเรียนรู้วิทยาศาสตร์และเทคโนโลยีตามทฤษฎีการเรียนรู้เพื่อสร้างสรรค์ปัญญา ของนักเรียนชั้น ม.3

1. กำหนดจำนวนของกลุ่มตัวอย่างของนักเรียนชั้น ม.3
2. กำหนดวิธีการเก็บข้อมูลผลการเรียนรู้วิทยาศาสตร์และเทคโนโลยี
3. ข้อมูลที่เป็นข้อเสนอแนะของการพัฒนาสื่อเลือกใช้สถิติอนุมานวิเคราะห์
4. กำหนดการวิเคราะห์เพื่อเปรียบเทียบผลการเรียนรู้วิทยาศาสตร์และเทคโนโลยี

63. ข้อใดแสดงถึงการออกแบบวิธีดำเนินการวิจัยด้วยการออกแบบการสุ่มตัวอย่าง

64. ข้อใดแสดงถึงการออกแบบวิธีดำเนินการวิจัยด้วยการออกแบบการวัดตัวแปร

65. ข้อใดแสดงถึงการออกแบบวิธีดำเนินการวิจัยด้วยการออกแบบการวิเคราะห์ข้อมูล

66. ข้อใดเป็นเครื่องมือที่ใช้วัดตัวแปรที่เกี่ยวกับความรู้ ความคิดของผู้เรียน

1. แบบวัดการกำกับตนเอง
2. แบบบันทึกความตั้งใจเรียน
3. แบบสอบถามแรงจูงใจใฝ่สัมฤทธิ์
4. แบบทดสอบวินิจัยคณิตศาสตร์

67. ข้อใดเป็นเครื่องมือที่ใช้วัดทางตัวแปรจิตวิทยา

1. แบบสังเกตการปฏิบัติการทดลอง
2. แบบสอบถามเจตคติต่อวิชาชีพครู
3. แบบทดสอบการเขียนเชิงสร้างสรรค์
4. แบบทดสอบกิจกรรมการเรียนรู้การสอน

68. เครื่องมือใดเหมาะกับการวัด “พฤติกรรมผู้นำ” ของนักเรียน ม. 6 จำนวน 10 ห้องเรียน

1. แบบสำรวจ
2. แบบทดสอบ
3. แบบสอบถามปลายเปิด
4. แบบสอบถามแบบมาตราส่วนประเมินค่า

69. ข้อใดเป็นลักษณะเครื่องมือวิจัยที่ดี

1. ลำดับของข้อคำถามเหมาะสม
2. เขียนเป็นประโยคปฏิเสธซ้อนปฏิเสธ
3. ใช้คำศัพท์ที่ดีความได้หลายความหมาย
4. ผลการให้คะแนนจากเครื่องมือไม่ตรงกัน

70. การเก็บรวบรวมข้อมูล “ความสามารถในการรับส่งบอล” ด้วยวิธีใดเหมาะสมที่สุด

1. การสังเกต
2. การสอบถาม
3. การสัมภาษณ์
4. การทดสอบปฏิบัติ

71. ข้อใดกล่าวถูกต้องเกี่ยวกับตัวแปร “ความสามารถในการคำนวณ”

- | | |
|------------------------------------|------------------------------|
| 1. วัดได้จากการกระตุ้นด้วยสิ่งเร้า | 2. วัดได้จากการสังเกตโดยตรง |
| 3. เป็นการวัดความรู้สึก | 4. เป็นการแสดงออกอย่างชัดเจน |

72. การนำแบบสอบวิชาภาษาไทยที่สร้างขึ้นไปตรวจสอบว่าข้อสอบในแบบสอบฉบับนั้นวัดเนื้อหาเดียวกันหรือไม่ แสดงถึงการหาคุณภาพเครื่องมือแบบใด

1. ความชัดเจน 2. ความง่าย 3. ความเที่ยง 4. อำนาจจำแนก

73. ผู้วิจัยนำผลการสังเกตจากผู้ประเมิน จำนวน 3 คนมาพิจารณาถึงความสอดคล้องของการบันทึกรายการการสังเกตที่ตรงกัน แสดงถึงการหาคุณภาพเครื่องมือแบบใด

1. ความเที่ยง 2. ความชัดเจน 3. ความเหมาะสม 4. การนำไปใช้ประโยชน์

74. หากผู้วิจัยพบว่า คะแนนที่ได้จากแบบสอบถามความมีวินัยไม่สอดคล้องกับพฤติกรรมการเข้าแถวซื้ออาหารของนักเรียน แสดงว่าเครื่องมือขาดคุณภาพแบบใด

- | | |
|----------------------------|-----------------------|
| 1. ความตรงตามภาวะสันนิษฐาน | 2. ความตรงเชิงเนื้อหา |
| 3. ความตรงเชิงทำนาย | 4. ความตรงร่วมสมัย |

คำสั่ง จงใช้ตัวเลือกต่อไปนี้ตอบคำถาม 75. – 76.

- | | |
|---------------------------------|-----------------------------|
| 1. สูตรของเพียร์สัน | 2. สูตรของครอนบาค |
| 3. สูตรของคูเคอร์-ริชาร์ดสัน 20 | 4. สูตรของสเปียร์แมน-บราวน์ |

75. สูตรใดใช้ในการประมาณค่าความเที่ยงของแบบทดสอบแบบปรนัยวิชาสังคม

76. สูตรใดใช้ในการประมาณค่าความเที่ยงของแบบวัดบุคลิกภาพด้วยการวัดซ้ำ

77. ข้อใดเป็นการสุ่มตัวอย่างแบบไม่อาศัยความน่าจะเป็น

1. ครูกึ่งเก็บข้อมูลกับนักเรียนที่ได้คะแนนสูง
2. ครูกึ่งดำเนินการทดลองกับนักเรียนที่จับสลากได้
3. ครูใหม่เก็บข้อมูลกับนักเรียนตามสัดส่วนของประชากรแต่ละกลุ่ม
4. ครูเก็บข้อมูลกับนักเรียนจากโรงเรียนขนาดใหญ่ กลาง และเล็ก

คำสั่ง จงใช้ตัวเลือกต่อไปนี้ตอบคำถาม 78. – 80.

1. การสุ่มแบบง่าย (simple random sampling)
2. การสุ่มแบบกลุ่ม (cluster random sampling)
3. การสุ่มแบบแบ่งชั้น (stratified random sampling)
4. การสุ่มแบบเป็นระบบ (systematic random sampling)

78. ผู้วิจัยต้องการสุ่มตัวอย่างนักศึกษาปีที่ 4 คณะรัฐศาสตร์ ด้วยการนำจำนวนประชากรหารด้วยจำนวนตัวอย่างที่ต้องการ ผู้วิจัยควรสุ่มแบบใด

79. ผู้วิจัยต้องการสุ่มตัวอย่างผู้ปกครองเพื่อศึกษาความคิดเห็นต่อการจัดการเรียนรู้ ซึ่งผู้ปกครองมีระดับการศึกษาตั้งแต่ประถมศึกษาถึงปริญญาตรี ผู้วิจัยควรสุ่มแบบใด

80. ผู้วิจัยต้องการสุ่มตัวอย่างนักเรียน ป.5 จำนวน 4 ห้องเรียนและนักเรียนแต่ละห้องมีระดับความสามารถที่แตกต่างกัน ผู้วิจัยควรสุ่มแบบใด

81. ข้อใดกล่าวถูกต้อง หากครูต้องการศึกษาความคิดเห็นต่อการจัดการเรียนรู้ของผู้ปกครอง ครูจึงเปิดตารางสุ่มตัวอย่างสำเร็จรูปเพื่อให้ได้จำนวนตัวอย่างที่เหมาะสม

1. ชีตหลักการใช้ตารางสำเร็จรูป
2. ชีตหลักการสุ่มเพื่อให้ได้ตัวอย่างที่เป็นตัวแทน
3. คำนึงถึงเฉพาะขนาดของตัวอย่างที่เหมาะสม
4. คำนึงถึงเฉพาะวิธีการสุ่มที่ถูกต้อง

คำสั่ง จงใช้ตัวเลือกต่อไปนี้ตอบคำถาม 82. – 84.

1. การวัดการกระจาย
2. การแจกแจงความถี่
3. การจัดตำแหน่งเปรียบเทียบ
4. การวัดแนวโน้มเข้าสู่ส่วนกลาง

82. สถิติกลุ่มใดที่ใช้ในการบรรยายว่าคะแนนแต่ละค่ามีค่าแตกต่างกันมากน้อยเพียงใด

83. การนำเสนอข้อมูลด้วยคำร้อยละ คะแนนมาตรฐาน แสดงว่าผู้วิจัยใช้สถิติกลุ่มใด

84. ต้องการนำเสนอจำนวนนักเรียนจำแนกตามกิจกรรมพัฒนาผู้เรียนที่สนใจ ผู้วิจัยควรใช้สถิติใด

85. ผลการสำรวจความคิดเห็นต่อหลักสูตรสถานศึกษา พบว่าผู้ปกครองส่วนใหญ่มีระดับความคิดเห็นระดับดีมาก ข้อมูลดังกล่าวมีการแจกแจงแบบใด

1. แบบเบ้ขวา
2. แบบเบ้ซ้าย
3. แบบโค้งมาก
4. แบบสมมาตร

86. ข้อใดคือความคลาดเคลื่อนในการทดสอบสมมติฐานทางสถิติ แบบที่ 1 (α)

1. การปฏิเสธ H_0 ที่ผิด
2. นิยมกำหนดเป็นค่า β
3. การปฏิเสธ H_0 ที่ถูกต้อง
4. ค่าวิกฤตของการทดสอบ

87. ผู้วิจัยต้องการทดสอบว่าความเป็นผู้นำระหว่างนักเรียนที่ได้รับการให้คำปรึกษาแบบ A ต่ำกว่าการให้คำปรึกษาแบบ B ข้อใดคือสมมติฐานทางสถิติ

1. $H_0: \mu_A < \mu_B$
2. $H_0: \mu_A > \mu_B$
- $H_1: \mu_A \neq \mu_B$
- $H_1: \mu_A > \mu_B$
3. $H_0: \mu_A = \mu_B$
4. $H_0: \mu_A > \mu_B$
- $H_1: \mu_A < \mu_B$
- $H_1: \mu_A < \mu_B$

คำสั่ง จงใช้ข้อมูลและตัวเลือกต่อไปนี้ตอบคำถาม 88. – 90.

จากการสำรวจจะเนนความเป็นผู้นำของนักเรียนชั้น ป. 2 จำนวน 20 คน ดังตาราง

คะแนนความเป็นผู้นำ	ความถี่
2	8
3	1
4	2
5	3
6	5
8	1

88. ข้อใดคือฐานนิยมของข้อมูลชุดนี้

1. 2.00
2. 4.00
3. 5.00
4. 6.00

89. ข้อใดคือมัธยฐานของข้อมูลชุดนี้

1. 3.50
2. 4.00
3. 4.50
4. 5.00

90. ข้อใดคือค่าเฉลี่ยของข้อมูลชุดนี้

1. 3.50 2. 4.00 3. 4.25 4. 4.50

91. ผู้วิจัยต้องการเปรียบเทียบความเป็นผู้นำของนักเรียนชั้น ป.2 ระหว่างก่อนและหลังทดลองสอนแบบบทบาทสมมติ

1. t - independent test 2. t - dependent test 3. F - test 4. χ^2 - test

คำสั่ง จงใช้ตัวเลือกที่กำหนดให้ ตอบคำถามข้อ 92. - 94.

1. χ^2 - test 2. z - test 3. F - test 4. t - test

92. ผู้วิจัยต้องการเปรียบเทียบสัดส่วนการศึกษาคือของนักศึกษาปริญญาตรีระหว่างเพศชายและหญิงว่าเท่ากันหรือไม่ ผู้วิจัยจะเลือกใช้สถิติทดสอบใด

93. ผู้วิจัยต้องการเปรียบเทียบทักษะชีวิตระหว่างนักเรียนแผนการเรียนวิทย์-คณิตกับศิลป์-คำนวณ โดยทราบความแปรปรวนของประชากร ผู้วิจัยจะเลือกใช้สถิติทดสอบใด

94. ผู้วิจัยต้องการเปรียบเทียบความแปรปรวนของทักษะการเขียนระหว่างนักเรียนที่ถนัดเขียนด้วยมือซ้ายกับถนัดเขียนด้วยมือขวา ผู้วิจัยจะเลือกใช้สถิติทดสอบใด

คำสั่ง ให้ใช้สถานการณ์ และตัวเลือกที่กำหนดให้ สำหรับการตอบคำถามข้อ 95. - 97.

สถานการณ์ ผู้วิจัยต้องการพัฒนาความคิดสร้างสรรค์ของเด็กปฐมวัยด้วยการใช้นิทาน

- ตัวเลือก
1. ข้อความที่ 1) และ 2) ถูกต้องทั้งสองข้อความ
 2. ข้อความที่ 1) และ 2) ผิดทั้งสองข้อความ
 3. ข้อความที่ 1) ถูก ส่วนข้อความที่ 2) ผิด
 4. ข้อความที่ 1) ผิด ส่วนข้อความที่ 2) ถูก

95. ข้อความ 1) หลักการเขียนบทคัดย่อ คือ ส่วนที่แสดงถึงผลงานที่ผู้วิจัย

ได้ดำเนินการมาทั้งหมด ความยาวไม่เกิน 4 หน้า

ข้อความ 2) ประโยชน์ของบทคัดย่อ คือ ช่วยให้ผู้วิจัยตัดสินใจว่าเป็นเรื่องที่สนใจที่ท้าวิจัยหรือไม่

96. ข้อความ 1) ขั้นตอนของการสุ่มตัวอย่างเด็กปฐมวัยต้องระบุในวิธีดำเนินการวิจัย

ข้อความ 2) การเขียนวิธีดำเนินการวิจัยเป็นการแสดงขั้นตอนของการวิเคราะห์

ความคิดสร้างสรรค์ของเด็กปฐมวัยที่ได้รับการทดลองใช้นิทาน

97. ข้อความ 1) จุดมุ่งหมายของการอภิปรายผลการวิจัย คือเพื่อตีความผลการวิเคราะห์

ที่เกี่ยวกับความคิดสร้างสรรค์ของเด็กปฐมวัยด้วยการใช้นิทาน

ข้อความ 2) หลักการเขียนการอภิปราย คือผู้วิจัยแสดงให้เห็นว่าผลการวิจัยครั้งนี้

แตกต่างจากทฤษฎี และผลงานวิจัยที่ผ่านมาอย่างไร

คำสั่ง ให้ใช้ตัวเลือกที่กำหนดให้ สำหรับการตอบคำถามข้อ 98. – 100.

1. ชื่อเรื่องวิจัย

2. ประเด็นปัญหาการวิจัย

3. วัตถุประสงค์การวิจัย

4. สมมติฐานการวิจัย

98. การศึกษาด้วยชุดการเรียนรู้ทำให้ทักษะการคิดแบบมีเหตุผลแตกต่างจากการเรียนรู้ด้วยวิธีการ

สอนแบบ 4 MAT ข้อความดังกล่าวตรงกับตัวเลือกใด

99. การพัฒนารูปแบบการสอนเพื่อเตรียมความพร้อมทางคณิตศาสตร์ของนักเรียน

ชั้นอนุบาล ข้อความดังกล่าวตรงกับตัวเลือกใด

100. ประสิทธิภาพการสอนของครู โรงเรียนมัธยมศึกษาที่สอนในโรงเรียนขนาดเล็ก ขนาดกลาง

และขนาดใหญ่ ข้อความดังกล่าวตรงกับตัวเลือกใด

