

การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ
โดยใช้รูปแบบและวิธีการต่างกัน

ปริญญาานิพนธ์
ของ
อารี วัชรโสคติกุล

เสนอต่อบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ เพื่อเป็นส่วนหนึ่งของการศึกษา
ตามหลักสูตรปริญญาการศึกษามหาบัณฑิต วิชาเอกการวัดผลการศึกษา
พฤษภาคม 2543
ลิขสิทธิ์เป็นของ มหาวิทยาลัยศรีนครินทรวิโรฒ

371.26013

ด บ59 ๗

๗.3

การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ
โดยใช้รูปแบบและวิธีการต่างกัน

บทคัดย่อ

ของ

อารี วัชรโสติดิกุล

26 ก.ค. 2543

เสนอต่อบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ เพื่อเป็นส่วนหนึ่งของการศึกษา
ตามหลักสูตรปริญญาการศึกษามหาบัณฑิต วิชาเอกการวัดผลการศึกษา
พฤษภาคม 2543

77

อารี วัชรโสติดิกุล. (2543). การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้
รูปแบบและวิธีการต่างกัน. ปรินญาณพนธ์ กศ.ม. (การวัดผลการศึกษา). กรุงเทพฯ :
บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ. คณะกรรมการควบคุม :
รองศาสตราจารย์ ดร.บุญเชิด ภิญโญนนตพงษ์, รองศาสตราจารย์ นิภา ศรีไพโรจน์.

การศึกษาค้นคว้าครั้งนี้มีจุดมุ่งหมาย เพื่อเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของ
ข้อสอบ โดยใช้รูปแบบต่างกัน คือ รูปแบบคะแนนรวมทั้งฉบับ แยกตามเนื้อหา และแยกตาม
ระดับพฤติกรรม ด้วยวิธีการตรวจสอบต่างกัน คือ วิธีชิปทดสอบ และวิธีถดถอยโลจิสติก แล้วทำการคัด
ข้อสอบที่ทำหน้าที่ต่างกันออกจากแบบทดสอบ เพื่อเปรียบเทียบค่าความเชื่อมั่น กลุ่มตัวอย่างเป็น
นักเรียนระดับชั้นมัธยมศึกษาปีที่ 2 ของโรงเรียนมัธยมศึกษา สังกัดกรมสามัญศึกษา กรุงเทพมหานคร
ส่วนกลาง กลุ่มที่ 4 จำนวน 994 คน จำแนกเป็นนักเรียนชาย 480 คน และนักเรียนหญิง 514 คน
ที่ได้จากการสุ่มแบบแบ่งชั้น (Stratified Random Sampling) เครื่องมือที่ใช้ในการศึกษาเป็น
แบบทดสอบผลสัมฤทธิ์วิชาคณิตศาสตร์ ชนิดเลือกตอบ 5 ตัวเลือก จำนวน 50 ข้อ ผลการศึกษาพบว่า

1. จำนวนข้อสอบที่ทำหน้าที่ต่างกันโดยใช้รูปแบบการตรวจสอบต่างกัน แตกต่างกัน
อย่างไม่มีนัยสำคัญทางสถิติ ไม่ว่าจะตรวจสอบด้วยวิธีชิปทดสอบหรือวิธีถดถอยโลจิสติก
2. จำนวนข้อสอบที่ทำหน้าที่ต่างกันโดยใช้วิธีการตรวจสอบต่างกัน แตกต่างกัน
อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ในรูปแบบรวมทั้งฉบับ และแตกต่างกันอย่างไม่มีนัยสำคัญ
ทางสถิติในรูปแบบแยกตามเนื้อหา และรูปแบบแยกตามระดับพฤติกรรม
3. ค่าความเชื่อมั่นของแบบทดสอบหลังจากคัดข้อสอบที่ทำหน้าที่ต่างกันออกโดยใช้
รูปแบบการตรวจสอบต่างกัน แล้วทำการปรับขยายค่าความเชื่อมั่นของแบบทดสอบให้มีจำนวนข้อ
เท่ากันด้วยสูตรเปียร์แมน-บราวน์แล้ว เมื่อตรวจสอบด้วยวิธีชิปทดสอบ พบว่า ค่าความเชื่อมั่น
แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 จึงทดสอบค่าความเชื่อมั่นเป็นรายคู่ระหว่างรูปแบบ
พบว่า ค่าความเชื่อมั่นแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ระหว่างรูปแบบแยกตาม
เนื้อหาและรูปแบบแยกตามระดับพฤติกรรม เมื่อตรวจสอบด้วยวิธีถดถอยโลจิสติก พบว่า
ค่าความเชื่อมั่นแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ
4. ค่าความเชื่อมั่นของแบบทดสอบหลังจากคัดข้อสอบที่ทำหน้าที่ต่างกันออกโดยใช้
วิธีการตรวจสอบต่างกัน แล้วทำการปรับขยายค่าความเชื่อมั่นของแบบทดสอบให้มีจำนวนข้อเท่ากัน
ด้วยสูตรเปียร์แมน-บราวน์แล้ว เมื่อตรวจสอบโดยใช้รูปแบบรวมทั้งฉบับ พบว่า ค่าความเชื่อมั่น
แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 เมื่อตรวจสอบโดยใช้รูปแบบแยกตามเนื้อหา พบว่า
ค่าความเชื่อมั่นแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 และเมื่อตรวจสอบโดยใช้รูปแบบแยก
ตามระดับพฤติกรรม พบว่า ค่าความเชื่อมั่นแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

A COMPARISON OF DIFFERENTIAL ITEM FUNCTIONING DETECTION
USING DIFFERENT FORMATS AND METHODS

AN ABSTRACT
BY
AREE WATCHARASOTTIKUL

Presented in partial fulfillment of the requirements
for the Master of Education degree in Educational Measurement
at Srinakharinwirot University
May 2000

Aree Watcharasottikul. (2000). *A Comparison of Differential Item Functioning Detection Using Different Formats and Methods*. Master Thesis, M.Ed. (Educational Measurement).
Bangkok : Graduate School, Srinakharinwirot University. Advisor Committee:
Assoc. Prof. Dr. Boonchird Pinyoanuntapong, Assoc. Prof. Nipa Sripairot.

The main purpose of this research was to compare differential item functioning (DIF) detection using different formats and methods. Different formats were total test score, subtest content score and subtest cognitive score. Different methods were the SIBTEST method and the Logistic Regression method. Another purpose was to test the reliability of the no DIF items. The sample group consisted of 994 Mathayomsuksa II students, 480 boys and 514 girls, from the secondary schools under the Department of General Education, which were selected by stratified random sampling. A 50-item test in Mathematics was used.

The results of the study were as follows.

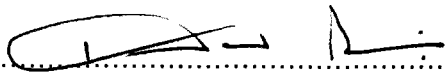
1. The number of DIF items detected by different formats was not statistically different in both methods.
2. The number of DIF items detected by different methods was statistically different at .05 level in the total test score whereas the subtest content score and the subtest cognitive score were not statistically different.
3. For three different formats with no DIF items, the reliability of each format was adjusted to equal length using the Spearman-Brown's formula. Detected by SIBTEST method, the reliability of each test format was statistically different at .05 level. When compared by paired-test, the subtest content score format and the subtest cognitive score format were statistically different at .05 level. Detected by Logistic Regression method, the reliability of each test format was not statistically different.
4. For two different methods with no DIF items, the reliability of each method was adjusted to equal length using the Spearman-Brown's formula. Detected by the total test score format, the reliability of each test method was statistically different at .05 level. When detected by the subtest content score format, the reliability of each test method was statistically different at .05 level. When detected by the subtest cognitive score format, the reliability of each test method was not statistically different.

ปริญญานิพนธ์
เรื่อง

การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ
โดยใช้รูปแบบและวิธีการต่างกัน

ของ
นางสาวอารี วัชรโสติกุล

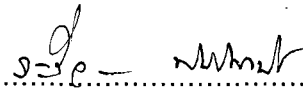
ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาการศึกษามหาบัณฑิต วิชาเอกการวัดผลการศึกษา
ของมหาวิทยาลัยศรีนครินทรวิโรฒ

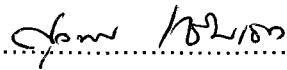
 คณบดีบัณฑิตวิทยาลัย
(ศาสตราจารย์ ดร.เสริมศักดิ์ วิศาลาภรณ์)
วันที่ 12 เดือน พฤษภาคม พ.ศ. 2543

คณะกรรมการสอบปริญญานิพนธ์

 ประธาน
(รองศาสตราจารย์ ดร.บุญเชิด ภิญโญนนตพงษ์)

 กรรมการ
(รองศาสตราจารย์ นิภา ศรีไพโรจน์)

 กรรมการที่แต่งตั้งเพิ่มเติม
(ผู้ช่วยศาสตราจารย์ ระวีวรรณ พันธุ์พานิช)

 กรรมการที่แต่งตั้งเพิ่มเติม
(อาจารย์ ดร.สุพร เข้มเฮง)

ประกาศคุณูปการ

ปริญญานิพนธ์ฉบับนี้ สำเร็จลุล่วงได้ด้วยความอนุเคราะห์และความกรุณาอย่างสูงจาก รองศาสตราจารย์ ดร.บุญเชิด ภิญโญอนันตพงษ์ และรองศาสตราจารย์ นิภา ศรีไพโรจน์ ที่ได้กรุณา ให้คำปรึกษา แนะนำ ชี้แนะแนวทาง ตรวจ แก้ไขข้อบกพร่องต่างๆ ทำให้ปริญญานิพนธ์ฉบับนี้สำเร็จ ลุล่วงไปด้วยดี ผู้วิจัยขอกราบขอบพระคุณเป็นอย่างสูงมา ณ โอกาสนี้

กราบขอบพระคุณผู้ช่วยศาสตราจารย์ ระวีวรรณ พันธุ์พานิช และ อาจารย์ ดร.สุพร เข้มแข็ง ที่กรุณาเป็นกรรมการในการสอบปริญญานิพนธ์ และตรวจสอบเครื่องมือวิจัย ตลอดจน ให้คำแนะนำและข้อคิดเห็นต่างๆ ในการแก้ไขเครื่องมือวิจัย และชี้แนะแนวทางในการทำวิจัย

กราบขอบพระคุณคณาจารย์ ในภาควิชาการวัดผลและวิจัยการศึกษาทุกท่าน ที่ได้อบรม สั่งสอนวิชาความรู้ในศาสตร์เกี่ยวกับการวัดผลการศึกษา นอกจากนี้ผู้วิจัยขอกราบขอบพระคุณ อาจารย์ชวลิต รวยอาจิน อาจารย์ที่ปรึกษา ที่ให้คำแนะนำปรึกษาที่เป็นประโยชน์มาโดยตลอด

กราบขอบพระคุณผู้ช่วยศาสตราจารย์ ดร.วีระ สุภากิจ ที่กรุณาให้ความช่วยเหลือ ในการตรวจ แก้ไขข้อบกพร่องต่างๆ ของบทคัดย่อภาษาอังกฤษ

ขอขอบคุณอาจารย์อุทัยวรรณ สายพัฒนา อาจารย์สายใจ เชียงอินทร์ และอาจารย์เสมอใจ สิงห์ชัย ที่กรุณาตรวจสอบเครื่องมือวิจัย ให้คำแนะนำและข้อคิดเห็นต่างๆ ในการแก้ไขและปรับปรุง เครื่องมือที่ใช้ในการวิจัยครั้งนี้

ขอขอบคุณผู้บริหาร คณะครู-อาจารย์ทุกท่านของโรงเรียนที่เป็นกลุ่มตัวอย่าง และขอขอบใจ นักเรียนทุกคนที่ให้ความร่วมมือในการเก็บรวบรวมข้อมูลเป็นอย่างดี

ขอขอบคุณคณะผู้บริหาร โรงเรียนพระหฤทัยคอนแวนต์ ที่ให้โอกาสแก่ผู้วิจัยในการทำ ปริญญานิพนธ์ฉบับนี้ และขอขอบคุณคณะครู โดยเฉพาะคณะครูหมวดวิชาคณิตศาสตร์ทุกคน ที่ช่วยเหลือและเป็นกำลังใจในการทำวิจัยครั้งนี้

ขอขอบคุณ พี่ๆ เพื่อนๆ และน้องๆ นิสิตระดับปริญญาโท วิชาเอกการวัดผลการศึกษา ภาควิชา ก รุ่นที่ 34 กับ แผน ข รุ่นที่ 2 และนิสิตระดับปริญญาเอก สาขาการทดสอบและวัดผล การศึกษา รุ่นที่ 4 รวมทั้งกัลยาณมิตรทุกคนที่ช่วยเหลือและเป็นกำลังใจในการทำวิจัยครั้งนี้

ขอขอบคุณ คุณฉัตรศิริ ปิยะพิมลสิทธิ์ ในฐานะผู้เอื้อเฟื้อโปรแกรม SIBTEST ที่นำมาใช้ในการวิเคราะห์ข้อมูลในการวิจัยครั้งนี้

กราบขอบพระคุณ บิดา มารดา และขอขอบคุณ พี่น้อง ที่เป็นกำลังใจ สนับสนุน ช่วยเหลือ ทำให้สามารถทำปริญญานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้

คุณค่าและประโยชน์อันพึงมีจากปริญญานิพนธ์ฉบับนี้ ผู้วิจัยขอมอบเป็นเครื่องบูชาพระคุณ บิดา มารดา และครูอาจารย์ ตลอดจนผู้มีพระคุณที่ได้ให้การอบรมสั่งสอนตลอดมา

อารี วัชรโสติกุล

สารบัญ

บทที่	หน้า
1 บทนำ	1
ภูมิหลัง	1
ความมุ่งหมายของการศึกษาค้นคว้า	4
ความสำคัญของการศึกษาค้นคว้า	4
ขอบเขตของการศึกษาค้นคว้า	5
การกำหนดประชากรและกลุ่มตัวอย่าง	5
ตัวแปรที่ศึกษา	5
นิยามศัพท์เฉพาะ	6
สมมุติฐานของการศึกษาค้นคว้า	7
2 เอกสารและงานวิจัยที่เกี่ยวข้อง	8
เอกสารที่เกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบ (DIF)	8
เอกสารเกี่ยวกับระดับพฤติกรรมและจุดประสงค์เชิงพฤติกรรม	17
งานวิจัยที่เกี่ยวข้อง	24
งานวิจัยในต่างประเทศ	24
งานวิจัยในประเทศ	29
3 วิธีดำเนินการศึกษาค้นคว้า	34
การกำหนดประชากรและเลือกกลุ่มตัวอย่าง	34
การสร้างเครื่องมือที่ใช้ในการศึกษาค้นคว้า	35
ลักษณะของแบบทดสอบ	37
การเก็บรวบรวมข้อมูล	39
ขั้นตอนในการวิเคราะห์ข้อมูล	40
สถิติที่ใช้ในการวิเคราะห์ข้อมูล	41
4 ผลการวิเคราะห์ข้อมูล	45
สัญลักษณ์และอักษรย่อที่ใช้ในการวิเคราะห์ข้อมูล	45
การเสนอผลการวิเคราะห์ข้อมูล	46
ผลการวิเคราะห์ข้อมูล	46

สารบัญ(ต่อ)

บทที่		หน้า
5	สรุปผล อภิปราย และข้อเสนอแนะ.....	55
	ความมุ่งหมายของการศึกษาค้นคว้า.....	55
	กลุ่มตัวอย่าง.....	55
	เครื่องมือที่ใช้ในการศึกษาค้นคว้า.....	55
	วิธีดำเนินการเก็บรวบรวมข้อมูล.....	55
	การเสนอผลการวิเคราะห์ข้อมูล.....	56
	สรุปผลการวิเคราะห์ข้อมูล.....	56
	อภิปรายผล.....	57
	ข้อเสนอแนะ.....	59
	บรรณานุกรม.....	60
	ภาคผนวก.....	67
	ภาคผนวก ก.....	68
	ภาคผนวก ข.....	70
	ภาคผนวก ค.....	72
	ภาคผนวก ง.....	79
	ประวัติย่อผู้วิจัย.....	84

บัญชีตาราง

ตาราง	หน้า
1 ผลบวกของ CDIF, NCDIF, หมวด CDIF, หมวด NCDIF และ DTF	16
2 จำนวนนักเรียนในกลุ่มตัวอย่างที่ใช้ในการวิจัย จำแนกตามขนาดโรงเรียน.....	35
3 โครงสร้างของแบบทดสอบผลสัมฤทธิ์วิชาคณิตศาสตร์ ค 203 จำแนกตามเนื้อหา และระดับพฤติกรรม	37
4 ค่าสถิติพื้นฐานของแบบทดสอบผลสัมฤทธิ์วิชาคณิตศาสตร์ จำแนกตามเนื้อหาและ ระดับพฤติกรรม	46
5 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจากแบบทดสอบผลสัมฤทธิ์ วิชาคณิตศาสตร์โดยรูปแบบและวิธีการต่างกัน	48
6 การเปรียบเทียบจำนวนข้อสอบที่ทำหน้าที่ต่างกันโดยใช้รูปแบบและวิธีการต่างกัน	50
7 ค่าความเชื่อมั่นของแบบทดสอบหลังการปรับขยายให้แบบทดสอบมีจำนวนข้อสอบ เท่ากัน.....	52
8 การเปรียบเทียบค่าความเชื่อมั่นของแบบทดสอบหลังจากตัดข้อสอบที่ทำหน้าที่- ต่างกันโดยใช้รูปแบบและวิธีการต่างกัน.....	53
9 การเปรียบเทียบความแตกต่างของค่าความเชื่อมั่นของแบบทดสอบระหว่างรูปแบบ รายคู่ โดยใช้วิธีการตรวจสอบด้วยวิธีชิปทดสอบด้วย UX_1	54
10 ค่าความยากง่าย (p) และค่าอำนาจจำแนก (r) ของแบบทดสอบผลสัมฤทธิ์ วิชาคณิตศาสตร์.....	71
11 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบรวมทั้งฉบับด้วยวิธีชิปทดสอบ.....	73
12 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบรวมทั้งฉบับด้วยวิธีถดถอยโลจิสติก	74
13 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแยกตามเนื้อหาด้วยวิธีชิปทดสอบ	75
14 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแยกตามเนื้อหาด้วย วิธีการถดถอยโลจิสติก	76
15 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแยกตามระดับพฤติกรรมด้วย วิธีชิปทดสอบ	77
16 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแยกตามระดับพฤติกรรมด้วย วิธีถดถอยโลจิสติก	78

บัญชีภาพประกอบ

ภาพประกอบ	หน้า
1 ขั้นตอนการวัดผลและการประเมินผล	17

บทที่ 1

บทนำ

ภูมิหลัง

แบบทดสอบเป็นเครื่องมือวัดผลและประเมินผล ที่ได้รับความนิยมมาก เนื่องจากสะดวกในการนำมาใช้ แบบทดสอบเป็นชุดของข้อสอบที่สร้างขึ้นมาอย่างมีระบบ เพื่อใช้วัดตัวอย่างพฤติกรรมของแต่ละบุคคล คุณสมบัติของแบบทดสอบที่ดีประกอบด้วย ความเที่ยงตรง ความเชื่อมั่น ง่ายลึกลับ ยั่วให้ยากตอบ มีความเป็นปรนัย และมีความยุติธรรม มีความจำเพาะเจาะจง มีประสิทธิภาพ มีความยากง่ายพอเหมาะ และมีอำนาจจำแนก (ขวาล แพรัตกุล. 2509 : 10-12)

โดยทั่วไปการสร้างแบบทดสอบ จะมุ่งวัดความสามารถหรือคุณลักษณะด้านใดด้านหนึ่ง ผู้สร้างจะเขียนข้อสอบวัดคุณลักษณะเฉพาะด้านนั้นๆ กล่าวคือเป็นการพิจารณาหรือคำนึงถึงความเป็นมิติเดียวกันของแบบทดสอบ (Unidimensionality) อย่างไรก็ตาม เมื่อนำแบบทดสอบไปสอบวัดกับกลุ่มตัวอย่างที่มีความสามารถเท่ากัน แต่มีความแตกต่างกันในเรื่อง เพศ เชื้อชาติ ศาสนา ภูมิฐานะ และระดับสติปัญญา ฯลฯ แล้วโอกาสในการตอบข้อสอบข้อนั้นได้ถูกต้องไม่เท่าเทียมกัน เรียกว่า แบบทดสอบนั้นมีความลำเอียง นั่นคือ เกิดความไม่ยุติธรรมในการสอบ หรือเอื้อประโยชน์ต่อคนบางกลุ่ม ทำให้ขาดความน่าเชื่อถือและเกิดความไม่เสมอภาคในการตัดสิน

การศึกษาความลำเอียงของข้อสอบและแบบทดสอบ ได้รับความสนใจจากนักวิจัยและนักวัดผลการศึกษาอย่างต่อเนื่อง ในต่างประเทศมีการศึกษาจากแบบทดสอบมาตรฐาน วัดผลสัมฤทธิ์ทางการเรียน และวัดเชาวน์ปัญญา เพื่อนำไปใช้ตัดสินจำแนกผู้สอบ (Wiersma, W. & Jurs, Stephen G. 1990 : 333) แต่คำว่า “ความลำเอียงของข้อสอบ (Item Bias)” เป็นคำที่ใช้กล่าวถึงตัวข้อสอบและให้ความหมายที่คลุมเครือในการใช้เกณฑ์ตัดสินเรื่องความลำเอียง ดังนั้นนักวิจัยและนักวัดผลการศึกษาจึงใช้สารสนเทศทางสถิติมาเป็นเกณฑ์ในการตัดสิน และใช้คำว่า “การทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning เรียกว่า DIF)” น่าจะเหมาะสมและมีความเป็นกลางมากกว่า (Holland & Wainer. 1993) การที่นักวิจัยและนักวัดผลการศึกษาจะเลือกใช้ คำว่า “การทำหน้าที่ต่างกันของข้อสอบ (DIF)” หรือ “ความลำเอียงของข้อสอบ (Item bias)” นั้นควรพิจารณาถึงความเหมาะสม และวัตถุประสงค์ของการนำไปใช้ (สมศักดิ์ จันผ่อง. 2541 : 90)

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบนั้น ในขั้นแรกจะให้ผู้เชี่ยวชาญตรวจสอบความเที่ยงตรงของแบบทดสอบ (Validity) แล้วใช้เทคนิคทางสถิติเข้ามาช่วย โดยนำแบบทดสอบไปใช้ แล้วนำคะแนนที่ได้จากผลการตอบมาวิเคราะห์ ดูลักษณะการตอบของกลุ่มผู้สอบที่อยู่ต่างกลุ่มกันว่า ข้อสอบหรือแบบทดสอบทำหน้าที่ต่างกันหรือไม่ ซึ่งสามารถใช้สารสนเทศทางสถิติสองแนวทาง (Camilli & Shepard. 1994 : 9) คือ ใช้เกณฑ์ภายนอก (External Method) หรือเกณฑ์ภายใน (Internal Method) ตัวอย่างการใช้เกณฑ์ภายนอก เช่น การศึกษาการทำหน้าที่ต่างกันในการสอบคัดเลือก เกณฑ์คือ ผลการเรียนรู้ ส่วนเกณฑ์ภายในจะพิจารณาคะแนนที่ได้จากผลการตอบของผู้สอบต่างกลุ่มว่าวัดในคุณลักษณะที่ต้องการวัดตามโครงสร้างเช่นเดียวกันหรือไม่ (ผจงจิต อินทสุวรรณ.

2530 : 54-55) การใช้เกณฑ์ภายนอกมีปัญหาในการสร้างเกณฑ์ที่เป็นมาตรฐาน ในปัจจุบันจึงนิยมใช้เกณฑ์ภายในมากกว่า และได้มีการพัฒนาเทคนิควิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบให้มีประสิทธิภาพขึ้น คือ ตรวจสอบได้ถูกต้อง ไม่สิ้นเปลืองค่าใช้จ่ายมากนัก และสะดวกในการนำมาใช้จำแนกได้สองวิธี ได้แก่ วิธีที่ใช้ทฤษฎีการตอบข้อสอบ (Item Response Theory เรียกว่า IRT) และวิธีที่ไม่ใช้ทฤษฎีการตอบข้อสอบ (Non Item Response Theory เรียกว่า NonIRT) ที่มีการให้คะแนนแบบสองค่า (Dichotomous) และการให้คะแนนแบบหลายค่า (Polytomous) โดยใช้คะแนนสองกลุ่ม คือ กลุ่มที่ใช้คะแนนที่สอบวัดได้โดยตรง (Observe Score) และกลุ่มที่ใช้คะแนนที่สอบวัดโดยตรงไม่ได้หรือตัวแปรแฝง (Latent Variable)

โปเทนซ่า และโดแรนส์ (Potenza & Dorans. 1995 : 25) ได้สรุปวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ที่มีการให้คะแนนแบบสองค่า (Dichotomous) ดังนี้

1. กลุ่มที่ใช้คะแนนที่สอบวัดได้โดยตรง (Observe Score) แบ่งได้ 2 วิธี คือ

1.1 วิธีการทดสอบสถิติแบบพาราเมตริกซ์ (Parametric) ได้แก่ วิธีถดถอยโลจิสติก (Logistic Regression เรียกว่า LRDI) (Swaminathan & Rogers. 1990) ซึ่งเป็นวิธีที่สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ดีทั้งแบบสม่ำเสมอ (Uniform DIF) และแบบไม่สม่ำเสมอ (Nonuniform DIF) สำหรับนิยามของการทำหน้าที่ต่างกันแบบสม่ำเสมอ (Uniform DIF) และแบบไม่สม่ำเสมอ (Nonuniform DIF) เมลเลนเบิร์ก (Mellenbergh. 1989 : 127-143) ได้นิยามการทำหน้าที่ต่างกันแบบสม่ำเสมอของข้อสอบ คือ ข้อสอบที่ไม่มีปฏิสัมพันธ์ระหว่างระดับความสามารถของผู้สอบกับกลุ่มผู้สอบ ซึ่งจะมีผลให้โอกาสในการตอบข้อสอบถูกจะสูงสำหรับกลุ่มหนึ่งมากกว่ากลุ่มอื่นในทุกระดับความสามารถของผู้สอบ และการทำหน้าที่ต่างกันแบบไม่สม่ำเสมอของข้อสอบ คือ ข้อสอบที่มีปฏิสัมพันธ์ระหว่างระดับความสามารถของผู้สอบกับกลุ่มผู้สอบ ทำให้ความแตกต่างของโอกาสในการตอบข้อสอบถูกสำหรับสองกลุ่มต่างกัน ในทุกระดับความสามารถของผู้สอบ

1.2 วิธีการทดสอบสถิติแบบนอนพาราเมตริกซ์ (Nonparametric) ได้แก่ วิธีแมนเทล-แฮนส์เซล (Mantel-Haenszel เรียกว่า MH) (Holland & Thayer. 1988) และวิธีการทำให้เป็นมาตรฐาน (Standardization Method) (Dorans & Holland. 1993) ซึ่งวิธีแมนเทล-แฮนส์เซลเป็นวิธีที่ได้รับการยอมรับ และเป็นวิธีมาตรฐานที่ใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในโครงการทดสอบของหน่วยงานมากที่สุด (Dorans & Holland. 1993)

2. กลุ่มที่ใช้คะแนนสอบวัดโดยตรงไม่ได้หรือตัวแปรแฝง (Latent Variable) แบ่งได้ 2 วิธี คือ

2.1 วิธีการทดสอบสถิติแบบพาราเมตริกซ์ (Parametric) เป็นวิธีที่มีพื้นฐานจากทฤษฎีการตอบข้อสอบ ที่ได้รับการพัฒนาและแยกย่อยออกไปหลายวิธี ได้แก่ วิธี General IRTLR (Tissen, Steinberg & Wainer. 1988) วิธี Limited information IRTLR วิธีล็อกลิเนียร์ (Loglinear IRTLR) (Kelderman, H. 1989) วิธี IRT-D² (Bock & Lieberman. 1970) และวิธีลอร์ด ไค-กำลังสอง (Lord's χ^2) (McLaughlin & Drasgow. 1987) ซึ่งกระบวนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในกลุ่มนี้ค่อนข้างยุ่งยากซับซ้อน เสียค่าใช้จ่ายสูง กลุ่มตัวอย่างต้องมีขนาดใหญ่ การนำไปใช้

อาจเกิดปัญหา เพราะข้อมูลจริงมักจะไม่เป็นไปตามแบบจำลองทฤษฎีการตอบข้อสอบ ทำให้การประมาณค่าของพารามิเตอร์ต่างๆ ทำไม่ได้ (Camilli & Shepard. 1994 : 50)

2.2 วิธีการทดสอบสถิติแบบนอนพาราเมตริกซ์ (Nonparametric) ได้แก่ วิธีชิปเทสต์ (SIBTEST) (Shealy & Stout. 1993) เป็นวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบและแบบทดสอบชนิดพหุมิติ (Multidimensional) สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้หลายข้อในคราวเดียวกัน และพบว่า ส่วนใหญ่จะเป็นการตรวจสอบการทำหน้าที่ต่างกันแบบสม่ำเสมอของข้อสอบ และเป็นวิธีที่สะดวกในคำนวณ

สำหรับงานวิจัยที่ศึกษาการทำหน้าที่ต่างกันของข้อสอบหรือแบบทดสอบนั้น ในต่างประเทศจะศึกษาเทคนิควิธีการตรวจสอบเพื่อเปรียบเทียบค่าพารามิเตอร์ของข้อสอบระหว่างสองกลุ่มย่อยว่ามีความแตกต่างกันหรือไม่ (Kim & Cohen. 1991 ; Rudner, Getson & Knight. 1980) ศึกษาปัจจัยที่ส่งผลต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เช่น ขนาดของกลุ่มตัวอย่าง ความยาวของแบบทดสอบ เป็นต้น (Oshima & Miller. 1992 ; Dorans & Holland. 1993 ; Mellenbergh. 1989 ; Scheuneman. 1979 ; Shepard, Camilli & Averill. 1981 ; Swaminathan & Rogers. 1990)

สำหรับงานวิจัยในประเทศไทยที่ศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ หรือการวิเคราะห์ความลำเอียงของข้อสอบ มีการศึกษาโดยใช้หลักของทฤษฎีการวัดแบบมาตรฐาน (Classical Test Theory เรียกว่า CTT) เช่น วิธีแปลงค่าความยาก (Transformed Item Difficulty) (พัชรีย์ ปิยพันธ์. 2531 ; เจน จันท์แสง. 2537 ; นิรมล ชัยสวัสดิ. 2537 ; จิตสุตา ธารพร. 2539 ; พรรณี จินตมาศ. 2540) วิธีวิเคราะห์ความแปรปรวน (Analysis of Variance) (เจน จันท์แสง. 2537 ; นิรมล ชัยสวัสดิ. 2537) วิธีไค-กำลังสอง (Chi-Square) (พัชรีย์ ปิยพันธ์. 2531 ; เจน จันท์แสง. 2537 ; นิรมล ชัยสวัสดิ. 2537) วิธีอำนาจจำแนก (คัมศักดิ์ ชื่นชม. 2539) วิธีวิเคราะห์องค์ประกอบ (Factor Analysis) (คัมศักดิ์ ชื่นชม. 2539) วิธีวิเคราะห์การถดถอย (Regression) (คัมศักดิ์ ชื่นชม. 2539) และการศึกษาโดยใช้หลักของทฤษฎีการตอบข้อสอบ (Item Response Theory เรียกว่า IRT) เช่น วิธีโค้งลักษณะข้อสอบ (Item Characteristic Curve) (พัชรีย์ ปิยพันธ์. 2531 ; เจน จันท์แสง. 2537 ; จิตสุตา ธารพร. 2539) วิธีลอร์ด ไค-กำลังสอง (Lord's χ^2) วิธี Raju's Area Measures วิธี Closed Interval Area (ญาณภัทร สีหะมงคล. 2540) และมีการศึกษาโดยใช้วิธีการทดสอบสถิติแบบนอนพาราเมตริกซ์ เช่น วิธีแมนเทิล-แฮนส์เซล (อังคณา สายยศ. 2539 ; เรวดี อินทสระ. 2539 ; พรรณี จินตมาศ. 2540 ; จิตสุตา ธารพร. 2539) และวิธีชิปเทสต์ (เรวดี อินทสระ. 2539 ; พรรณี จินตมาศ. 2540)

นอกจากนักวิจัยและนักวัดผลการศึกษาทั้งในต่างประเทศและในประเทศ ให้ความสนใจค้นคว้าพัฒนาเทคนิควิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบอย่างต่อเนื่อง เพื่อให้ได้ผลที่ถูกต้องที่สุดแล้ว จากการศึกษาการทำหน้าที่ต่างกันของข้อสอบ (Differential item Functioning เรียกว่า DIF) มาศึกษาการทำหน้าที่ต่างกันของหมวดข้อสอบ (Differential Bundle Functioning เรียกว่า DBF) ดังงานวิจัยของโอชิมา, ราจู, เฟาเวอร์ส และสไลนส์ (Oshima, Raju, Flowers &

Slinde. 1998 : 353-369) ที่ได้ศึกษาการรอบความคิดการทำหน้าที่ต่างกันของข้อสอบและแบบทดสอบ (Differential Functioning of Items and Tests เรียกว่า DFIT) ตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบจากแบบทดสอบความเข้าใจในการอ่านภาษาอังกฤษที่แบ่งสองลักษณะ คือ แบ่งระดับพฤติกรรม (Taxonomy) และแบ่งตามเนื้อหา จะเป็นการตรวจสอบการทำหน้าที่ต่างกันของระดับพฤติกรรม และเนื้อหา เป็นการพัฒนาเทคนิคการตรวจสอบการทำหน้าที่ต่างกัน

จากเหตุผลดังกล่าวข้างต้น การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีซิปเทสต์ (SIBTEST) และวิธีถดถอยโลจิสติก (LRDIF) เป็นวิธีที่สะดวกในการนำมาใช้ และมีประสิทธิภาพในการตรวจสอบ ทำให้ผู้วิจัยสนใจศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจากแบบทดสอบวิชาคณิตศาสตร์ โดยศึกษาในกรณีที่มีรูปแบบและวิธีการต่างกัน รูปแบบที่ศึกษา ได้แก่ รวมทั้งฉบับแยกเนื้อหา และแยกตามระดับพฤติกรรม ส่วนวิธีการตรวจสอบที่ศึกษา ได้แก่ วิธีซิปเทสต์ และวิธีถดถอยโลจิสติก ในระดับชั้นมัธยมศึกษาปีที่ 2 ศึกษาโดยให้กลุ่มผู้สอบเพศชายเป็นกลุ่มอ้างอิง (Reference Group) และกลุ่มผู้สอบเพศหญิงเป็นกลุ่มเปรียบเทียบ (Focal Group) เนื่องจากโอชิมะ, ราจู, เฟาเวอร์ส และ สไลนด์ (Oshima, Raju, Flowers & Slinde. 1998 : 367) ได้กล่าวถึงในงานวิจัยว่า เพศชายมีความสามารถทางคณิตศาสตร์สูงกว่าเพศหญิง

ความมุ่งหมายของการศึกษาค้นคว้า

การศึกษาค้นคว้าครั้งนี้มีจุดมุ่งหมายสำคัญ เพื่อศึกษาผลการตรวจสอบการทำหน้าที่ต่างกันโดยใช้รูปแบบต่างกันสามรูปแบบ คือ คະແນວรวมทั้งฉบับ แยกตามเนื้อหา และแยกตามระดับพฤติกรรม ด้วยวิธีการตรวจสอบต่างกันสองวิธี คือ วิธีซิปเทสต์ และวิธีถดถอยโลจิสติก เป็นจุดมุ่งหมายเฉพาะของการศึกษาค้นคว้าต่อไปนี้

1. เพื่อเปรียบเทียบจำนวนข้อสอบที่ทำหน้าที่ต่างกันโดยใช้รูปแบบการตรวจสอบต่างกันในแต่ละวิธีการตรวจสอบ
2. เพื่อเปรียบเทียบจำนวนข้อสอบที่ทำหน้าที่ต่างกันโดยใช้วิธีการตรวจสอบต่างกันในแต่ละรูปแบบการตรวจสอบ
3. เพื่อเปรียบเทียบค่าความเชื่อมั่นของแบบทดสอบ หลังจากคัดข้อสอบที่ทำหน้าที่ต่างกันออก โดยใช้รูปแบบการตรวจสอบต่างกันในแต่ละวิธีการตรวจสอบ
4. เพื่อเปรียบเทียบค่าความเชื่อมั่นของแบบทดสอบ หลังจากคัดข้อสอบที่ทำหน้าที่ต่างกันออก โดยใช้วิธีการตรวจสอบต่างกันในแต่ละรูปแบบการตรวจสอบ

ความสำคัญของการศึกษาค้นคว้า

ผลจากการศึกษาในครั้งนี้จะเป็นข้อมูลสำหรับเทคนิคการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ตรวจสอบรวมทั้งฉบับ แยกตามเนื้อหา และแยกตามระดับพฤติกรรม ตลอดจนเป็นแนวทางในการศึกษาค้นคว้าวิจัยเกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบต่อไป

ขอบเขตของการศึกษาค้นคว้า

การกำหนดประชากรและการเลือกกลุ่มตัวอย่าง

ประชากร

ประชากรที่ใช้ในการศึกษาค้นคว้าครั้งนี้ เป็นนักเรียนชั้นมัธยมศึกษาปีที่ 2 ภาคเรียนที่ 2 ปีการศึกษา 2542 ของโรงเรียนมัธยมศึกษา สังกัดกรมสามัญศึกษา กรุงเทพมหานคร ส่วนกลาง กลุ่มที่ 4 จำนวน 14 โรงเรียน จำนวน 187 ห้องเรียน รวมนักเรียนทั้งสิ้น 7,888 คน จำแนกเป็นนักเรียนชาย 3,928 คน นักเรียนหญิง 3,960 คน

กลุ่มตัวอย่าง

กลุ่มตัวอย่างที่ใช้ในการศึกษาค้นคว้าครั้งนี้ เป็นนักเรียนชั้นมัธยมศึกษาปีที่ 2 ภาคเรียนที่ 2 ปีการศึกษา 2542 ของโรงเรียนมัธยมศึกษา สังกัดกรมสามัญศึกษา กรุงเทพมหานคร ส่วนกลาง กลุ่มที่ 4 ซึ่งได้มาโดยการสุ่มแบบแบ่งชั้น (Stratified Random Sampling) มีขนาดโรงเรียนเป็นชั้น (Strata) และมีห้องเรียนเป็นหน่วยการสุ่ม (Sampling Unit) จำนวนโรงเรียนทั้งหมด 6 โรงเรียน จำนวนห้องเรียนทั้งหมด 23 ห้องเรียน รวมนักเรียนทั้งสิ้น 1,051 คน จำแนกเป็นนักเรียนชาย 522 คน และ นักเรียนหญิง 529 คน

ตัวแปรที่ศึกษา

1. ตัวแปรอิสระ ได้แก่

1.1 วิธีตรวจสอบการทำหน้าที่ต่างกัน ได้แก่

1.1.1 วิธีชิปทดสอบ

1.1.2 วิธีถดถอยโลจิสติก

1.2 รูปแบบการตรวจสอบการทำหน้าที่ต่างกัน แบ่งเป็นสามรูปแบบ ได้แก่

1.2.1 การตรวจสอบรวมทุกข้อ

1.2.2 การตรวจสอบแยกตามเนื้อหา

1.2.3 การตรวจสอบแยกตามระดับพฤติกรรม

2. ตัวแปรตาม ได้แก่

2.1 ข้อสอบที่ทำหน้าที่ต่างกัน

2.2 ความเชื่อมั่นของแบบทดสอบที่คัดข้อสอบที่ทำหน้าที่ต่างกันออก

นิยามศัพท์เฉพาะ

1. การทำหน้าที่ต่างกันของข้อสอบ หมายถึง ข้อสอบที่ผู้สอบซึ่งมีความสามารถเท่ากันในเรื่องที่ต้องการวัด มีโอกาสตอบข้อสอบข้อนั้นได้ถูกต้องไม่เท่ากัน เนื่องจากอยู่ในกลุ่มผู้สอบย่อยต่างกัน คือ กลุ่มผู้สอบเพศหญิง กับ กลุ่มผู้สอบเพศชาย

2. วิธีการตรวจสอบการทำหน้าที่ต่างกัน หมายถึง การวิเคราะห์หาดัชนีการทำหน้าที่ต่างกันของข้อสอบ ซึ่งในการศึกษาค้างนี้ ศึกษาสองวิธี คือ

2.1 วิธีซิปเทสต์ (SIBTEST) เป็นวิธีพิจารณาค่าการทำหน้าที่ต่างกันของข้อสอบที่ได้จากการทดสอบความแตกต่างของการตอบถูกของกลุ่มผู้สอบแต่ละกลุ่ม ข้อสอบที่มีค่าการทำหน้าที่ต่างกันของข้อสอบ จะมีค่า β_U แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05

2.2 วิธีถดถอยโลจิสติก (LRDIF) เป็นวิธีพิจารณาค่าการทำหน้าที่ต่างกันของข้อสอบที่ได้จากการทดสอบความแตกต่างของการตอบถูกของกลุ่มผู้สอบแต่ละกลุ่ม ข้อสอบที่มีค่าการทำหน้าที่ต่างกันของข้อสอบ จะมีค่าฟังก์ชันถดถอยโลจิสติกของผู้สอบสองกลุ่มต่างกัน

3. ดัชนีการทำหน้าที่ต่างกันของข้อสอบ หมายถึง ค่าสถิติที่ใช้แสดงความแตกต่างของผลการตอบข้อสอบระหว่างกลุ่มผู้สอบที่มีเพศต่างกันจากวิธีการตรวจสอบการทำหน้าที่ต่างกันสองวิธี

3.1 ดัชนีการทำหน้าที่ต่างกันด้วยวิธีซิปเทสต์ หมายถึง ค่าที่ได้จากการทดสอบความแตกต่างของการตอบถูกของกลุ่มผู้สอบแต่ละกลุ่ม คือ β_U

3.2 ดัชนีการทำหน้าที่ต่างกันด้วยวิธีถดถอยโลจิสติก หมายถึง ค่าความแตกต่างของฟังก์ชันถดถอยโลจิสติกระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ คือ B

4. กลุ่มอ้างอิง (Reference Group) หมายถึง กลุ่มผู้สอบที่คาดว่าจะจะเป็นกลุ่มที่ได้เปรียบจากการทำหน้าที่ต่างกันของข้อสอบ กล่าวคือ เป็นกลุ่มผู้สอบที่มีโอกาสตอบข้อสอบทำหน้าที่ต่างกันได้ถูกต้องสูงกว่ากลุ่มเปรียบเทียบ ในการศึกษาค้างนี้ใช้กลุ่มเพศชายเป็นกลุ่มอ้างอิง

5. กลุ่มเปรียบเทียบ (Focal Group) หมายถึง กลุ่มผู้สอบที่คาดว่าจะจะเป็นกลุ่มที่เสียเปรียบจากการทำหน้าที่ต่างกันของข้อสอบ กล่าวคือ เป็นกลุ่มผู้สอบที่มีโอกาสตอบข้อสอบทำหน้าที่ต่างกันได้ถูกต้องต่ำกว่ากลุ่มอ้างอิง ในการศึกษาค้างนี้ใช้กลุ่มเพศหญิงเป็นกลุ่มเปรียบเทียบ

6. แบบทดสอบ หมายถึง แบบทดสอบผลสัมฤทธิ์วิชาคณิตศาสตร์ (ค 203) ชั้นมัธยมศึกษาปีที่ 2 เป็นแบบทดสอบเลือกตอบชนิด 5 ตัวเลือก จำนวน 50 ข้อ ที่ผู้วิจัยสร้างขึ้นเอง

7. รูปแบบการตรวจสอบการทำหน้าที่ต่างกัน หมายถึง การตรวจสอบการทำหน้าที่ต่างกันโดยจัดการตรวจสอบข้อสอบรวมทั้งฉบับ แยกตามเนื้อหา และแยกตามระดับพฤติกรรม

7.1 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบรวมทั้งฉบับ เป็นการตรวจสอบข้อสอบเป็นรายข้อชุดเดียวกันทั้งฉบับ

7.2 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแยกตามเนื้อหา เป็นการตรวจสอบข้อสอบเป็นรายข้อหลายชุดแยกตามกลุ่มเนื้อหา คือ ระบบจำนวนเต็ม ความเท่ากันทุกประการ เศษส่วนและทศนิยม และสมบัติของรูปสามเหลี่ยมมุมฉาก

7.3 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแยกตามระดับพฤติกรรม
เป็นการตรวจสอบข้อสอบเป็นรายข้อหลายชุดแยกตามระดับพฤติกรรมด้านสติปัญญา (Cognitive Domain) ในวิชาคณิตศาสตร์ตามแนวคิดของวิลสัน (Wilson, 1971) แบ่งเป็น 4 ระดับ ได้แก่
ความรู้หรือทักษะการคิดคำนวณ ความเข้าใจ การนำไปใช้ และการวิเคราะห์

8. ความเชื่อมั่นของแบบทดสอบ หมายถึง สัดส่วนของความแปรปรวนของคะแนนจริงต่อความแปรปรวนของคะแนนที่สอบได้ โดยความแปรปรวนคะแนนจริงคำนวณจากความแปรปรวนร่วมระหว่างคะแนนส่วนย่อยของแบบทดสอบ ในที่นี้พิจารณาจากความคงที่แน่นอนของคะแนนที่นักเรียนแต่ละคนทำได้จากแบบทดสอบผลสัมฤทธิ์วิชาคณิตศาสตร์ ที่ตรวจให้คะแนนแบบ (0,1) ประมาณค่าความเชื่อมั่นด้วย สูตรสัมประสิทธิ์ r_B

9. ผู้เชี่ยวชาญ หมายถึง ครูผู้สอนคณิตศาสตร์ ระดับมัธยมศึกษาที่มีประสบการณ์อย่างน้อย 5 ปี จำนวน 2 ท่านกับนักวัดผลการศึกษา จำนวน 3 ท่าน

สมมุติฐานของการศึกษาค้นคว้า

1. เมื่อใช้การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ตามรูปแบบการตรวจสอบรวมทั้งฉบับ แยกตามเนื้อหา และแยกตามระดับพฤติกรรม จะได้จำนวนข้อแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ .05
2. เมื่อใช้การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีชิปทดสอบและวิธีถดถอยโลจิสติก จะได้จำนวนข้อแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ .05
3. ความเชื่อมั่นของแบบทดสอบหลังจากคัดข้อสอบที่ทำหน้าที่ต่างกันออก ที่ตรวจสอบตามรูปแบบรวมทั้งฉบับ แยกตามเนื้อหา และแยกตามระดับพฤติกรรม จะมีค่าต่างกัน
4. ความเชื่อมั่นของแบบทดสอบหลังจากคัดข้อสอบที่ทำหน้าที่ต่างกันออก ที่ตรวจสอบด้วยวิธีชิปทดสอบและวิธีถดถอยโลจิสติก จะมีค่าต่างกัน

บทที่ 2

เอกสารและงานวิจัยที่เกี่ยวข้อง

เอกสารและงานวิจัยที่เกี่ยวข้องกับการศึกษาค้นคว้า ผู้วิจัยได้แบ่งออกเป็นหัวข้อดังนี้

1. เอกสารเกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบ (DIF)
2. เอกสารเกี่ยวกับระดับพฤติกรรมและจุดประสงค์เชิงพฤติกรรม
3. งานวิจัยที่เกี่ยวข้อง

1. เอกสารเกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบ (DIF)

1.1 ความหมายของการทำหน้าที่ต่างกันของข้อสอบ (DIF)

การทำหน้าที่ต่างกันของข้อสอบและแบบทดสอบ ภาษาอังกฤษใช้คำว่า “Differential Item Functioning เรียกว่า DIF” หรือ “Differential Test Functioning เรียกว่า DTF” มีผู้ให้ความหมายไว้หลายท่าน ดังนี้

อลเลนและเยน (Allen & Yen. 1979 : 129) ให้ความหมายของการทำหน้าที่ต่างกันของข้อสอบว่า เป็นข้อสอบที่วัดความสามารถหรือคุณลักษณะทางจิตวิทยาของผู้สอบแต่ละกลุ่มไม่ตรงกัน

คามิลลีและเชฟเพ็ด (Camilli & Shepard. 1994 : 8) ให้ข้อสรุปเกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบ (DIF) ตามหลักในการวิเคราะห์นั้นมิได้ตรวจสอบหาความลำเอียงโดยตรง แต่เป็นการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างกลุ่มผู้สอบตั้งแต่สองกลุ่มขึ้นไปที่มีความสามารถหลัก (Primary Abilities) ที่มุ่งวัดเท่ากัน สถิติ DIF ได้มาจากการตรวจสอบความเป็นพหุมิติของข้อสอบโดยจะแสดงว่าข้อสอบทำหน้าที่ต่างกันระหว่างกลุ่ม เมื่อมีการแจกแจงของความสามารถรอง (Secondary Abilities) แตกต่างกันในกลุ่มผู้สอบที่มีความสามารถหลักเท่ากัน

ลอร์ด (Lord. 1980 . 212) กล่าวว่า “ในแบบจำลองของทฤษฎีการตอบข้อสอบ (IRT) มีการควบคุมความสามารถ โดยหมายถึงค่าพารามิเตอร์ของบุคคลในแบบจำลองข้อสอบจะแสดงการทำหน้าที่ต่างกันของข้อสอบ (DIF) ถ้าค่าพารามิเตอร์ของข้อสอบมีการแปรเปลี่ยนข้ามกลุ่มผู้สอบ”

เคลดอร์แมนและเกียร์ดีย์ (Kelderman & Georerdy. 1990) กล่าวว่า การทำหน้าที่ต่างกันของข้อสอบ หมายถึง ความแตกต่างในการทำหน้าที่ของข้อสอบหลังจากกลุ่มของผู้สอบได้ถูกจับคู่ตามความสามารถหรือคุณลักษณะที่ข้อสอบนั้นวัด

มิลสาร์พและเอเวอร์สัน (Millsap & Everson. 1993 : 297) กล่าวว่า การทำหน้าที่ต่างกันของข้อสอบ หมายถึง ความแตกต่างในการทำหน้าที่ของข้อสอบหลังจากกลุ่มของผู้สอบได้ถูกจับคู่ตามความสามารถหรือคุณลักษณะที่ข้อสอบนั้นวัด

สรุปได้ว่า การทำหน้าที่ต่างกันของข้อสอบ หมายถึง ความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องของผู้สอบที่มีความสามารถระดับเดียวกันแต่มาจากต่างกลุ่มกัน มีความแตกต่างกัน

1.2 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (Detect DIF)

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (Detect DIF) หรือนักวิจัยบางท่านเรียกว่า วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ (Analysis DIF) ซึ่งแต่เดิมใช้คำว่า การตรวจสอบความลำเอียงของข้อสอบ (item bias) โดยเริ่มจากการพิจารณาทางภาษาเป็นเกณฑ์ ในการศึกษาความลำเอียงของข้อสอบและแบบทดสอบ จะทำการตรวจสอบก่อนและหลังจากการทดลองใช้ โดยใช้การตรวจสอบโดยผู้เชี่ยวชาญ โดยพิจารณาถึงรูปแบบของข้อสอบ เนื้อหาคำที่ใช้ และอื่นๆ เพื่อไม่ให้เกิดความลำเอียง ต่อมาในระยะหลังเกิดความคลุมเครือในการใช้เกณฑ์ในการตัดสินเรื่องความลำเอียง นักวิจัยและนักวัดผลการศึกษาจึงใช้สารสนเทศทางสถิติมาเป็นเกณฑ์ในการตัดสิน สำหรับสารสนเทศทางสถิติได้จากผลการตอบของผู้สอบต่างกลุ่มที่มีความสามารถเท่ากัน แต่มีโอกาสดำเนินการตอบข้อสอบได้ถูกต้องไม่เท่ากัน

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ จำแนกตามทฤษฎีการตรวจสอบได้สองวิธี ได้แก่ วิธีที่ใช้ทฤษฎีการตอบข้อสอบ (IRT) และวิธีที่ไม่ใช้ทฤษฎีการตอบข้อสอบ (NonIRT) ที่มีการให้คะแนนแบบสองค่า (Dichotomous) และหลายค่า (Polytomous) โดยใช้คะแนนสองกลุ่มคือ กลุ่มที่ใช้คะแนนที่สอบวัดได้โดยตรง (Observe Score) และกลุ่มที่ใช้คะแนนที่สอบวัดโดยตรงไม่ได้หรือตัวแปรแฝง (Latent Variable)

โปเทนซ่า และโดแรนส์ (Potenza & Dorans, 1995 : 25) ได้สรุปวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ที่มีการให้คะแนนแบบสองค่า (Dichotomous) ดังนี้

1. กลุ่มที่ใช้คะแนนที่สอบวัดได้โดยตรง (Observe Score) แบ่งได้ 2 วิธี คือ

1.1 วิธีการทดสอบสถิติแบบพารามेटริกซ์ (Parametric) ได้แก่ วิธีถดถอยโลจิสติก

(Logistic Regression เรียกว่า LRDIF)

1.2 วิธีการทดสอบสถิติแบบนอนพารามेटริกซ์ (Nonparametric) ได้แก่ วิธีแมนเทิล-แฮนส์เซล

(Mantel-Haenszel เรียกว่า MH) และวิธีการทำให้เป็นมาตรฐาน (Standardization Method)

2. กลุ่มที่ใช้คะแนนสอบวัดโดยตรงไม่ได้หรือตัวแปรแฝง (Latent Variable) แบ่งได้ 2 วิธี คือ

2.1 วิธีการทดสอบสถิติแบบพารามेटริกซ์ (Parametric) ได้แก่ วิธี General IRTLR

วิธี Limited information IRTLR วิธีล็อกลิเนียร์ (Loglinear IRTLR) วิธี IRT-D² และวิธีลอร์ดไค-กกำลังสอง (Lord's χ^2)

2.2 วิธีการทดสอบสถิติแบบนอนพารามेटริกซ์ (Nonparametric) ได้แก่ วิธีซิปเทสต์

(SIBTEST)

ในการศึกษาครั้งนี้ ผู้วิจัยใช้วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีซิปเทสต์ (SIBTEST) และวิธีถดถอยโลจิสติก (LRDIF) ซึ่งทั้งสองวิธี เป็นวิธีที่มีผู้นำมาใช้ในทางปฏิบัติ เนื่องจากมีประสิทธิภาพในการตรวจสอบ และสามารถวิเคราะห์ด้วยคอมพิวเตอร์ส่วนบุคคลได้ มีรายละเอียดดังนี้

1.2.1 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีของชิปเทสต์ (SIBTEST)

เชลลีและสตาต์ (Shealy & Stout. 1993 : 159-194) เสนอวิธีชิปเทสต์ (SIBTEST) สำหรับตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ มีข้อตกลงเบื้องต้นว่า มีกลุ่มตัวอย่างสองกลุ่ม คือ กลุ่มอ้างอิง (Reference Group : R) และกลุ่มเปรียบเทียบ (Focal Group : F) โดยที่ข้อคำถามจะวัดความสามารถที่แท้จริงโดยใช้ θ หรือเรียกว่า ความสามารถเป้าหมาย (Target Ability) และความสามารถที่ไม่ตั้งใจที่มีอิทธิพลต่อการตอบข้อสอบแทนด้วย η หรือเรียกว่า ความสามารถสอดแทรก (Nuisance Ability) ตัวอย่างเช่น แบบทดสอบเหตุผลทางคณิตศาสตร์ ข้อสอบบางข้อมีเนื้อหาเกี่ยวกับเบสบอลและหมากรุก ซึ่งผู้สอบบางคนคุ้นเคยกับเรื่องดังกล่าว ในตัวอย่างนี้ มีความสามารถสอดแทรก 2 อย่าง คือ ความรู้ด้านเบสบอล (η_1) และความรู้ด้านหมากรุก (η_2) ส่วนความรู้ทางคณิตศาสตร์เป็นความสามารถเป้าหมาย (θ) (Li, Nandakumar & Stout. 1995 : 4-6)

สมมุติฐานเพื่อการทดสอบการทำหน้าที่ต่างกันของข้อสอบ

$$H_0 : \beta_U = 0$$

$$H_1 : \beta_U > 0$$

เมื่อ β_U เป็นพารามิเตอร์ที่แสดงถึงการทำหน้าที่ต่างกัน (DIF) ทิศทางเดียว โดยที่ $\beta_U = \int_0 B(\theta) f_F(\theta) d\theta$ และ $B(\theta) = T_{SR}(\theta) - T_{SF}(\theta)$ เป็นความแตกต่างในการศึกษาฟังก์ชันการตอบข้อสอบสำหรับกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ

ให้ N แทน ผลรวมของจำนวนข้อสอบ ให้ข้อสอบข้อที่ $1 \dots n$ แทน ข้อสอบที่มีความเที่ยงตรง และข้อสอบข้อที่ $n+1 \dots N$ แทน ข้อสอบที่ต้องการศึกษา ให้ U_i แทน ผลการตอบข้อสอบข้อที่ i ที่มีค่า 0 ถ้าตอบข้อสอบข้อ i ผิด และเป็น 1 เมื่อตอบถูก เมื่อ $X = \sum_{i=1}^n U_i$ แทน คะแนนรวมของแบบทดสอบที่มีความเที่ยงตรง และ $Y = \sum_{i=n+1}^N U_i$ แทน คะแนนรวมของแบบทดสอบที่ต้องการศึกษา สำหรับค่าสถิติที่ประมาณค่าการทำหน้าที่ต่างกันของข้อสอบ คำนวณได้จาก

$$\hat{\beta}_U = \sum_{k=0}^K \hat{p}_k (\bar{Y}_{Rk}^* - \bar{Y}_{Fk}^*)$$

เมื่อ $\hat{p}_k$ แทน สัดส่วนของผู้สอบกลุ่มเปรียบเทียบที่ตอบแบบทดสอบย่อย k ข้อ โดยที่ $\bar{Y}_{Rk}^*$ และ $\bar{Y}_{Fk}^*$ แทน การปรับค่าเฉลี่ยของแบบทดสอบย่อยที่ศึกษา สำหรับผู้สอบในกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ที่ตอบแบบทดสอบย่อย k ข้อ ค่า $\hat{\beta}_U$ เป็นการประมาณค่าการทำหน้าที่ต่างกันของข้อสอบ สถิติสำหรับการทดสอบสมมุติฐานไร้ภัย (null hypothesis) ของข้อสอบและแบบทดสอบที่ทำหน้าที่ไม่ต่างกัน โดยกำหนดให้สถิติทดสอบของวิธีชิปเทสต์ คือ

$$B = \hat{\beta}_U / \hat{\sigma}(\hat{\beta}_U)$$

เมื่อ $\hat{\sigma}(\hat{\beta}_U)$ เป็นการประมาณค่าความคลาดเคลื่อนมาตรฐานของ $\hat{\beta}_U$ ที่กำหนดให้

$$\hat{\sigma}(\hat{\beta}_U) = \left\{ \sum_{k=1}^K \hat{p}_k^2 \left[\frac{1}{J_{Rk}} \hat{\sigma}^2(Y/k, R) + \frac{1}{J_{Fk}} \hat{\sigma}^2(Y/k, F) \right] \right\}^{1/2}$$

J_{Rk}, J_{Fk} เป็นจำนวนผู้สอบในกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่ทำคะแนนได้เท่ากันแบบทดสอบย่อยที่เที่ยงตรง ($X=k$)

สมมุติฐานไรรันจะปฏิเสธอัตราความคลาดเคลื่อน α ถ้า β มีค่าเกินกว่าเปอร์เซ็นต์ไทล์ที่ $100(1-\alpha)$ ของการกระจายปกติมาตรฐาน

1.2.2 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีถดถอยโลจิสติก (Logistic Regression)

สวามินาธาน และ โรเจอร์ส (Swaminathan & Rogers. 1990) เสนอวิธีถดถอยโลจิสติก (Logistic Regression เรียกว่า LR DIF) เพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนสองค่า (Dichotomous) ดัดแปลงจากวิธีล็อกลิเนียร์ของเมลเลนเบอร์ก (Mellenbergh. 1982) และวิธีแมนเทิล-แฮนส์เซลของฮอลแลนด์และเทเยอร์ (Holland & Thayer. 1988) วิธีถดถอยโลจิสติก สามารถคำนวณได้ง่ายโดยใช้แบบจำลองถดถอยโลจิสติก ซึ่งแบบจำลองนี้ใช้ตัวแปรความสามารถแบบต่อเนื่อง มีความไวในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้งแบบสม่ำเสมอ (Uniform DIF) และแบบไม่สม่ำเสมอ (Nonuniform DIF) นอกจากนี้ยังเป็นวิธีที่มีความยืดหยุ่น เนื่องจากนำไปประยุกต์กับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีผู้สอบหลายกลุ่ม และการให้คะแนนข้อสอบแบบหลายค่า (Polytomous) (Miller & Spray. 1993) การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ จะใช้สมการมาตรฐานของแบบจำลองถดถอยโลจิสติกคำนวณผลการตอบข้อสอบถูก ดังนี้ (Swaminathan & Rogers. 1990)

$$P(u_{ij} = 1 | \theta_{ij}) = \frac{e^{(\beta_{0j} + \beta_{1j}\theta_{ij})}}{1 + e^{(\beta_{0j} + \beta_{1j}\theta_{ij})}}, i = 1, 2, \dots, n_j, j = 1, 2.$$

เมื่อ	u_{ij}	แทน	ผลการตอบข้อสอบของผู้สอบคนที่ i ในกลุ่ม j
	β_{0j}	แทน	พารามิเตอร์จุดตัด
	β_{1j}	แทน	พารามิเตอร์ความชันสำหรับกลุ่ม j
	θ_{ij}	แทน	ค่าความสามารถที่สังเกตได้ของผู้สอบคนที่ i ในกลุ่ม j

จากแบบจำลองดังกล่าว ถ้า $\beta_{01} = \beta_{02}$ และ $\beta_{11} = \beta_{12}$ แล้วฟังก์ชันถดถอยโลจิสติกของผู้สอบสองกลุ่มเหมือนกัน แสดงว่าข้อสอบทำหน้าที่ไม่ต่างกัน (No DIF) ถ้า $\beta_{11} = \beta_{12}$ และ $\beta_{01} \neq \beta_{02}$ แล้วฟังก์ชันถดถอยโลจิสติกของผู้สอบสองกลุ่มขนานกันแต่ไม่ทับกัน แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบสม่ำเสมอ (Uniform DIF) และถ้า $\beta_{01} = \beta_{02}$ และ $\beta_{11} \neq \beta_{12}$ แล้วฟังก์ชันถดถอยโลจิสติกของผู้สอบไม่ขนานกัน แสดงว่า ข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ

(Nonuniform DIF) จากแบบจำลองถดถอยโลจิสติกดังกล่าว สามารถเปลี่ยนเป็นแบบจำลองในรูปแบบการตรวจสอบการทำหน้าที่ต่างกันแบบสม่ำเสมอและแบบไม่สม่ำเสมอดังนี้

$$P(u = 1) = \frac{e^Z}{1 + e^Z}$$

$$\text{โดยที่ } Z = \tau_0 + \tau_1\theta + \tau_2g + \tau_3(\theta g)$$

โดยกำหนดให้ $g = 1$ ถ้าผู้สอบอยู่ในกลุ่มแรก และ $g = 0$ ถ้าผู้สอบอยู่ในกลุ่มที่สอง

เมื่อ θ แทน ระดับความสามารถของผู้สอบ (คะแนนรวม)

τ_0 แทน พารามิเตอร์จุดตัด

τ_1 แทน สัมประสิทธิ์ของความสามารถ

τ_2 แทน ความแตกต่างในกลุ่มของผลการตอบข้อสอบ

τ_3 แทน ปฏิสัมพันธ์ระหว่างกลุ่มและระดับความสามารถ

แบบจำลองถดถอยโลจิสติกข้างต้น สามารถเปลี่ยนเป็นแบบจำลองเชิงเส้นเมทริกโลจิท (logit metric) ซึ่งจะอยู่ในรูป log ของอัตราส่วนของโอกาสในการตอบข้อสอบถูกต้อง โอกาสในการตอบข้อสอบผิด ดังนี้

$$\log \left[\frac{P}{1-P} \right] = \tau_0 + \tau_1\theta_{pj} + \tau_2g_j + \tau_3(\theta_{pj}g_j)$$

จากแบบจำลองดังกล่าว θ_{pj} และ g เป็นตัวแปรอิสระ ส่วนพารามิเตอร์ τ_2 และ τ_3 สอดคล้องกับพารามิเตอร์ในสมการมาตรฐานของแบบจำลองถดถอยโลจิสติก ดังนี้

$$\tau_2 = \beta_{01} - \beta_{02}$$

$$\tau_3 = \beta_{11} - \beta_{12}$$

ในการตัดสินใจว่าข้อสอบทำหน้าที่ต่างกันแบบสม่ำเสมอหรือแบบไม่สม่ำเสมอจะพิจารณาพารามิเตอร์ τ_2 และ τ_3 กล่าวคือ ถ้า $\tau_2 \neq 0$ และ $\tau_3 = 0$ แสดงว่า ข้อสอบทำหน้าที่ต่างกันแบบสม่ำเสมอ และถ้า $\tau_3 \neq 0$ ซึ่ง $\tau_2 = 0$ หรือไม่ได้แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ สำหรับพารามิเตอร์แต่ละข้อในแบบจำลอง Z_{pj} สามารถประมาณค่าโดยวิธีไลค์ลิฮูดสูงสุด (Maximum Likelihood Estimation เรียกว่า MLE) ซึ่งกำหนดให้รูปฟังก์ชันดังนี้

$$L(\text{Data}|\tau) = \prod_{i=1}^N P(u_i)^{u_{pj}} [1 - P(u_{pj})]^{1-u_{pj}}$$

เมื่อ N แทน ขนาดกลุ่มตัวอย่าง ; $N = n_1 + n_2$

n แทน ความยาวของแบบสอบ

ค่าประมาณของพารามิเตอร์ใช้วิธีไลค์ลียูตสูงสุดมีการแจกแจงแบบปกติของตัวแปรพหุในรูปเชิงเส้นกำกับ (Asymptotically Multivariate Normal) ซึ่งมีค่าเฉลี่ยของเวกเตอร์ τ และเมทริกซ์ ความแปรปรวน-ความแปรปรวนร่วมในรูป Σ ในขณะที่ Σ^{-1} เป็นเมทริกซ์สารสนเทศกำหนดดังนี้

$$\Sigma^{-1} = -E \left[\frac{\partial^2}{\partial \tau_r \partial \tau_s} \ln L \right] ; r, s = 0, 1, 2, 3.$$

เมื่อ E แทน ค่าคาดหวังของเมทริกซ์

$\ln L$ แทน ลอการิทึมของฟังก์ชันไลค์ลียูต

ดังนั้นการแจกแจงของการประมาณค่าพารามิเตอร์ด้วยวิธี MLE จะอยู่ในรูป

$$\hat{\tau} \sim N(\tau, \Sigma)$$

โดยที่ $\hat{\tau} = [\tau_0, \tau_1, \tau_2, \tau_3]$ ส่วนความคลาดเคลื่อนมาตรฐานเชิงเส้นกำกับของค่าประมาณของ τ_s ($S=0,1,2,3$) เมื่อ S เป็นสมาชิกแนวเส้นทแยงมุมของ Σ สามารถคำนวณได้จากสูตร

$$SE(\hat{\tau}_s) = \sqrt{\Sigma_{ss}}$$

การทดสอบสมมติฐานของการทำหน้าที่ต่างกันของข้อสอบ จะทดสอบสมาชิกของ τ_s ซึ่งสมมติฐานที่สนใจคือ

$$H_0 : C_\tau = 0$$

$$H_1 : C_\tau \neq 0$$

โดยที่

$$C = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

ส่วนการทดสอบนัยสำคัญของสมมติฐาน จะใช้สถิติไค-กำลังสอง ที่ระดับชั้นความเป็นอิสระเท่ากับ 2 (df=2) ดังนี้

$$\chi^2 = \hat{\tau}' C' (C \Sigma C')^{-1} C \hat{\tau}$$

ถ้า χ^2 มีค่ามากกว่า $\chi^2_{\alpha,2}$ แสดงว่าปฏิเสธสมมติฐานของข้อสอบที่ทำหน้าที่ไม่ต่างกัน (No DIF) นั่นคือ ข้อสอบทำหน้าที่ต่างกัน (DIF) นั่นเอง

1.3. กรอบความคิดการทำหน้าที่ต่างกันของข้อสอบและแบบทดสอบ (THE DFIT FRAMEWORK)

ราชู, แวนเดอร์ลินเดน และเฟียร์ (Raju, Van der Linden & Fleer. 1995) ได้แนะนำกรอบความคิดการทำหน้าที่ต่างกันของแบบทดสอบ (DFIT) โดยใช้ทฤษฎีการตอบข้อสอบ (IRT) ซึ่งใช้ได้กับแบบทดสอบทั้งมิติเดียวและพหุมิติ และมีการให้คะแนนแบบสองค่า (Dichotomous) หรือหลายค่า (Polytomous) นักวิจัยหลายท่านได้ศึกษาการตรวจสอบการทำหน้าที่ต่างกันแบบนอนพาราเมตริกซ์ (Nonparatric) โดยใช้โปรแกรม BILOG3

ดักลาส รุสโซ และสตาต์ (Douglas, Roussos & Stout. 1996) ได้เสนอแนะกรอบความคิดการทำหน้าที่ต่างกันของหมวดข้อสอบ (Differential Bundle Functioning เรียกว่า DBF) โดยใช้โปรแกรม SIBTEST (Shealy & Stout. 1993) วิเคราะห์แหล่งของการทำหน้าที่ต่างกันโดยเปรียบเทียบเพศ แล้วสังเกตหมวดข้อสอบในแบบทดสอบ โดยใช้กรอบความคิดการทำหน้าที่ต่างกันของข้อสอบและแบบทดสอบ (DFIT) ของแบบทดสอบวัดผลสัมฤทธิ์ ที่สร้างจากตารางกำหนดข้อสอบที่แยกระดับพฤติกรรมการศึกษาตามแนวคิดของบลูม (Bloom. 1956) ทางด้านสติปัญญา (ความรู้ความจำ, ความเข้าใจ, การนำไปใช้, การวิเคราะห์, การสังเคราะห์, การประเมินค่า) เพื่ออธิบายการทำหน้าที่ต่างกันของหมวดข้อสอบ (DBF)

กรอบความคิดการทำหน้าที่ต่างกันของข้อสอบและแบบทดสอบ (DFIT)

คะแนนจริงของแบบทดสอบที่มีข้อสอบ k ข้อที่วัดผู้สอบด้วยแบบทดสอบที่มีมิติเดียว และตรวจให้คะแนนแบบสองค่า (Dichotomous) ตามทฤษฎีการตอบข้อสอบ (IRT) สามารถเขียนเป็นสมการได้ดังนี้

$$T_s = \sum_{i=1}^k P_i(\theta_s)$$

โดยที่ $P_i(\theta_s)$ แทนความน่าจะเป็นที่ผู้สอบคนที่ s ที่มีความสามารถ θ_s จะทำข้อสอบข้อที่ i ถูก ในกรอบความคิดการทำหน้าที่ต่างกันของข้อสอบและแบบทดสอบ (DFIT) ผู้สอบในกลุ่มเปรียบเทียบ (Focal Group) มีคะแนนจริง (T_{SF}) และผู้สอบในกลุ่มอ้างอิง (Reference Group) มีคะแนนจริง (T_{SR}) ถ้าคะแนนของผู้สอบมี T_{SR} เท่ากับ T_{SF} แสดงว่าคะแนนจริงของผู้สอบคนนั้นเป็นอิสระจากการเป็นสมาชิกของกลุ่ม ถ้าคะแนน T_{SR} และ T_{SF} ยังมีความแตกต่างกัน แสดงว่าแบบทดสอบยังทำหน้าที่ต่างกัน และการวัดหน้าที่ต่างกันของแบบทดสอบจากผู้สอบคนหนึ่งสามารถนิยามได้จาก $(T_{SF} - T_{SR})^2$ ดังนั้น การทำหน้าที่ต่างกันของแบบทดสอบของผู้สอบทั้งหมดนิยามได้จาก

$$DTF = \epsilon_F(DTF_s) = \epsilon_F(T_{SF} - T_{SR})^2$$

ราชู, แวนเดอร์ลินเดน และเฟียร์ (Raju, Van der Linden & Fleer. 1995) นิยาม การทำหน้าที่ต่างกันไว้สองชนิด คือ การทำหน้าที่ต่างกันของข้อสอบแบบทดแทน (Compensatory Differential Item Functioning เรียกว่า CDIF) และการทำหน้าที่ต่างกันของข้อสอบแบบไม่ทดแทน (Non Compensatory Differential Item Functioning เรียกว่า NCDIF) โดยที่ผลบวกของดัชนี CDIF มีค่าเท่ากับค่าการทำหน้าที่ต่างกันของข้อสอบ ดัชนี NCDIF พิจารณากรณีเฉพาะของดัชนี CDIF ซึ่งมี ข้อตกลงว่าข้อสอบทุกข้อจะปลอดจาก DIF ซึ่งสามารถทดสอบนัยสำคัญของดัชนี DTF, CDIF และ NCDIF ได้ด้วยการทดสอบสอดคล้องด้วย ไค-กำลังสอง ที่ df เท่ากับ $N_F - 1$ โดยที่ N_F คือ จำนวนคนที่ อยู่ในกลุ่มเปรียบเทียบ (Focal Group) ทดสอบนัยสำคัญของ DTF ทำนองเดียวกัน NCDIF สามารถ ทดสอบด้วยไค-กำลังสองที่ df เท่ากับ $N_F - 1$ อย่างไรก็ตามพบว่า การทดสอบไค-กำลังสองสำหรับ ดัชนี NCDIF มีผลต่อขนาดกลุ่มตัวอย่างที่มีขนาดใหญ่ จากการศึกษาแบบมอนติคาร์โล (Monte Carlo) ที่จุดตัด .006 ทดสอบนัยสำคัญ NCDIF, CDIF ข้อสอบที่ CDIF มีนัยสำคัญจะถูกตัดออกจาก ฉบับเพื่อให้เป็นไปตาม DTF ที่ไม่มีนัยสำคัญ

การทำหน้าที่ต่างกันของหมวดข้อสอบ (DBF)

กรอบความคิดการทำหน้าที่ต่างกันของข้อสอบและแบบทดสอบ (DFIT) เริ่มจาก การปรับข้อสอบทั้งหมดในแบบทดสอบ โดยใช้แบบจำลองทฤษฎีการตอบข้อสอบ (IRT) สำหรับกลุ่มที่ ศึกษาสองกลุ่ม หลังจากคำนวณค่าพารามิเตอร์ข้อสอบและกำหนดลงบนมาตราส่วน (common scale) แล้ววัดกลุ่มข้อสอบเหล่านั้นให้สอดคล้องตามหมวดหมู่ เช่น จำแนกตามกลุ่มทักษะที่ต่างกัน ตามกลุ่มที่เอื้อกับเชื้อชาติ หรือเพศ หรือจำแนกตามระดับพฤติกรรมเมื่อกำหนดหมวดข้อสอบแล้ว จึงนิยาม DBF สองชนิดคือ หมวด CDIF และหมวด DTF ตาม กรอบความคิด DFIT ซึ่งสรุปผล ได้ดังแสดงในตาราง 1 โดยที่ค่า DTF นิยามตามความหมายของ CDIF ซึ่งคล้ายกับ DTF ที่นิยามตาม ความหมายของ NCDIF

หมวด CDIF

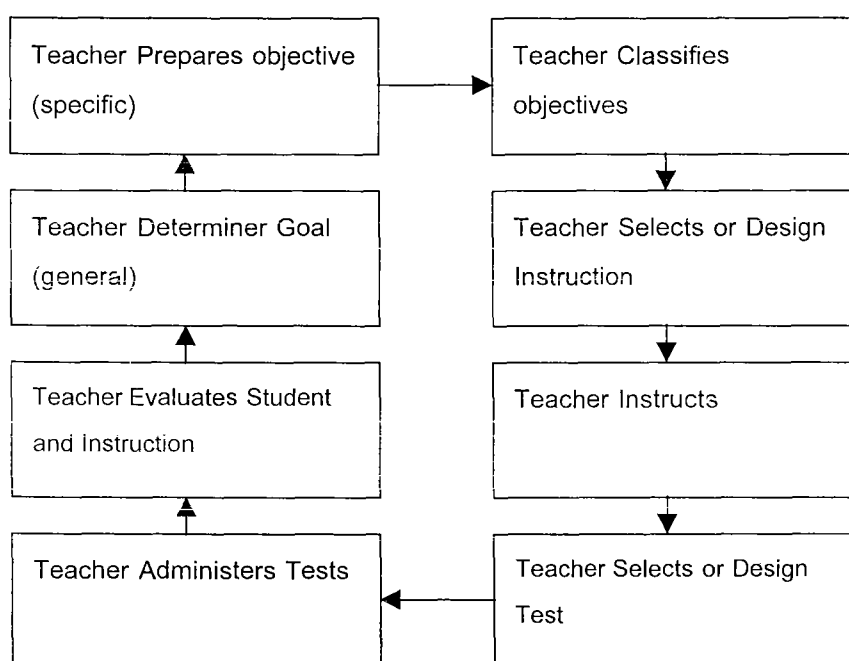
หมวด CDIF คือการขยาย CDIF การหาค่า CDIF ทำได้โดยการหาผลบวกข้อสอบ ของหมวด CDIF เนื่องจากการนำลักษณะของ CDIF มารวมกัน ดังนั้นผลรวมของค่าหมวด CDIF จึงเป็นค่าของ DTF ของทั้งฉบับ DTF สำหรับผลรวมของแบบทดสอบจะเป็นประโยชน์ต่อการพัฒนา แบบทดสอบที่ตรวจตราหมวด

หมวด NCDIF

หมวด NCDIF เป็นการขยาย NCDIF ออก อย่างไรก็ตามวิธีการจะไม่เหมือนกับ CDIF เพราะถ้าหมวด NCDIF ไม่ใช่ผลบวกของค่า NCDIF การวิเคราะห์ DFIT นั้นจะทำการวิเคราะห์ แยกกันในแต่ละหมวด และได้ผลลัพธ์เฉพาะของแต่ละหมวด สำหรับหมวด CDIF ของหมวดอื่นๆ จึงนำเข้ามาใช้ในการคำนวณ DTF ของหมวดที่ศึกษา แต่หมวด NCDIF จะไม่นำเข้ามาคำนวณด้วย ประเด็นที่น่าสนใจอีกข้อหนึ่ง คือ การประมาณค่าพารามิเตอร์ความสามารถ θ_s ที่คำนวณมาจาก ข้อสอบทุกข้อและจะคำนวณได้เพียงครั้งเดียว เพราะฉะนั้นสิ่งที่ต้องการศึกษาในวิธีนี้คือ หมวดของ ข้อสอบดังกล่าว จะมีผลการตอบข้อสอบของกลุ่มผู้สอบสองกลุ่มที่มีความสามารถโดยรวมจากข้อสอบ เท่ากันนั้นจะมีความแตกต่างกันหรือไม่ เมื่อแต่ละหมวดประกอบด้วยจำนวนข้อสอบไม่เท่ากันแล้ว

2. เอกสารที่เกี่ยวกับระดับพฤติกรรมและจุดประสงค์เชิงพฤติกรรม

หลักสูตรทุกสาขาวิชาประกอบด้วยอย่างน้อย 2 อย่าง คือ จุดหมายกับเนื้อหาวิชา จุดหมายเป็นตัวบอกทิศทางว่า แต่ละเนื้อหานั้นต้องการสอนให้ผู้เรียนเกิดพฤติกรรมใด (ล้วน สายยศ และ อังคณา สายยศ. 2539 : 42) เมื่อทราบจุดหมายแล้วจึงดำเนินการขั้นตอนต่อไป คือ การเรียนการสอน แล้วจึงวัดผลและประเมินผลการศึกษา ซึ่งขั้นตอนในการวัดผลและการประเมินผลการศึกษา ทักแมน (Tuckman, Bruce W. 1975 : 44) ได้เสนอเป็น 8 ขั้นตอน ดังนี้



ภาพประกอบ 1 ขั้นตอนการวัดผลและการประเมินผล

ขั้นตอนการวัดผลและการประเมินผลการศึกษานั้น เป็นกระบวนการที่ต่อเนื่องกัน แล้วแต่ว่าจะแบ่งอย่างไร กล่าวโดยสรุปได้ดังนี้

1. ตั้งวัตถุประสงค์ เพื่อจะได้ทราบว่าต้องการวัดอะไร วัดแค่ไหน และวัดเพื่ออะไร
2. เลือกสถานการณ์ ปัญหา และกิจกรรมในการสอบ
3. ร่างและเขียนข้อสอบ เพื่อกระตุ้นให้ผู้สอบแสดงพฤติกรรมที่ต้องการ
4. ดำเนินการสอบ จัดเตรียมข้อสอบ กำหนดเวลาสอบ สถานการณ์ และสิ่งแวดล้อม
5. กำหนดหลักเกณฑ์ในการให้คะแนน
6. ตรวจข้อสอบและวิเคราะห์ข้อสอบ
7. อภิปรายข้อสอบ เพื่อให้นักเรียนทราบว่าตอบผิด ถูก อย่างไร
8. ปรับปรุงข้อสอบ เพื่อให้ข้อสอบมีคุณภาพดีและเหมาะสมในการนำไปใช้

ขั้นตอนการเลือกสถานการณ์ ปัญหา และกิจกรรมที่จะใช้สอนนั้น จะต้องเลือกให้ครอบคลุมเนื้อหาทั้งหมดที่ต้องการสอบหรือกำหนดเป็นวัตถุประสงค์ไว้ก่อนสอน และจะต้องกำหนดพฤติกรรมที่วัดว่ามีอะไรบ้างตั้งข้อเสนอบนของบลูมและคณะ (ลัวัน สายยศ และ อังคณา สายยศ. 2539 : 41 ; อ้างอิงจาก Bloom, et al. 1956) จำแนกพฤติกรรม 3 ด้าน คือ พฤติกรรมด้านสติปัญญา (Cognitive Domain) พฤติกรรมด้านความรู้สึก (Affective Domain) และพฤติกรรมด้านทักษะกลไก (Psychomotor Domain)

สำหรับพฤติกรรมด้านสติปัญญา (Cognitive Domain) ในวิชาคณิตศาสตร์ ตามแนวคิดของ วิลสัน (Wilson. 1971 : 643-685) แบ่งเป็น 4 ระดับ ได้แก่

1. ความรู้หรือทักษะการคิดคำนวณ (Computation) ประกอบด้วยความรู้ความจำเกี่ยวกับข้อเท็จจริง ความรู้ความจำเกี่ยวกับศัพท์และนิยาม และความสามารถในการใช้กระบวนการคิดคำนวณตามลำดับขั้นตอนที่เคยเรียนรู้มาแล้ว

2. ความเข้าใจ (Comprehension) ประกอบด้วยความเข้าใจเกี่ยวกับความคิดรวบยอด ความเข้าใจเกี่ยวกับหลักการ กฎ นัยทั่วไป โครงสร้างทางคณิตศาสตร์ ความสามารถในการแปลงปัญหาจากรูปแบบหนึ่งไปเป็นอีกรูปแบบหนึ่ง ความสามารถในการคิดตามแนวของเหตุผล และความสามารถในการอ่านและตีความปัญหาทางคณิตศาสตร์

3. การนำไปใช้ (Application) ประกอบด้วยความสามารถในการแก้ปัญหาที่คุ้นเคย ความสามารถในการเปรียบเทียบ ความสามารถในการวิเคราะห์ข้อมูล และความสามารถในการมองเห็นรูปแบบ ลักษณะโครงสร้างที่มีความเหมือนและความสามารถกัน

4. การวิเคราะห์ (Analysis) ประกอบด้วยความสามารถในการแก้ปัญหาที่ไม่คุ้นเคย ความสามารถในการค้นพบความสัมพันธ์ ความสามารถในการพิสูจน์ ความสามารถในการวิพากษ์วิจารณ์การพิสูจน์ และความสามารถในการสร้างและทดสอบนัยทั่วไปที่สร้างขึ้น

เมื่อทราบถึงพฤติกรรมด้านสติปัญญาแล้ว จะต้องระบุพฤติกรรมของผู้เรียนให้ชัดเจนว่าต้องการให้ ผู้เรียนสามารถกระทำหรือปฏิบัติสิ่งใดได้หลังจากเรียนแล้ว คือ การกำหนดจุดมุ่งหมายเชิงพฤติกรรม ดังที่ บุญเชิด ภิญโญอนันตพงษ์ (ม.ป.ป. : 71-93) กล่าวไว้ดังนี้

ความหมายของจุดประสงค์เชิงพฤติกรรม

จุดประสงค์เชิงพฤติกรรม หมายถึง จุดประสงค์รายวิชาที่ระบุพฤติกรรมการแสดงออกไว้อย่างชัดเจน พฤติกรรมนั้นต้องสังเกตและวัดได้ และต้องเป็นจุดประสงค์ปลายทางที่ปรารถนาจะให้นักเรียนแสดงออกเมื่อสิ้นสุดการสอน

องค์ประกอบของจุดประสงค์เชิงพฤติกรรม แบ่งเป็น 3 ส่วน คือ

1. พฤติกรรมคาดหวัง (Expected behavior) เป็นพฤติกรรมที่ต้องการปลูกฝังให้นักเรียนเกิดขึ้นหลังจากมีการจัดการเรียนการสอน คำที่ใช้จึงมักเป็นคำที่แสดงการกระทำ (Action word) ที่สามารถสังเกตเห็นได้ชัด เช่น เขียน อ่าน วาด ชี้ บอก อธิบาย ฯลฯ

2. สถานการณ์หรือเงื่อนไข (Condition or Situation) หมายถึง สภาพของสิ่งที่ต้องการเรียนรู้เพื่อให้เกิดพฤติกรรมที่คาดหวัง คำที่ใช้มักจะมีคำว่า ถ้า...แล้ว, กำหนดให้..., หลังจาก...แล้ว, เมื่อ...แล้ว เป็นต้น

3. เกณฑ์หรือมาตรฐาน (Criterion or Performance Standard) จะอาศัยหลักการเหตุผลของผู้เชี่ยวชาญ หรือโดยวิธีการทางคณิตศาสตร์ก็ได้ แล้วแต่ความเหมาะสม การกำหนดเกณฑ์จึงมักเป็นเปอร์เซ็นต์หรือบอกจำนวนการกระทำสิ่งนั้นๆ ว่าได้เท่าไรจึงผ่าน

ประโยชน์ของการใช้จุดประสงค์เชิงพฤติกรรม จำแนกเป็น 3 ด้าน คือ

1. ช่วยชี้แนะแนวทางการสอนและการพัฒนาหลักสูตร
2. ช่วยชี้แนวทางในการประเมินผล
3. ช่วยให้การเรียนรู้เป็นไปโดยง่ายขึ้น

ข้อเสนอแนะในการเขียนจุดประสงค์เชิงพฤติกรรม

1. สามารถนำไปสร้างข้อสอบได้หลาย ๆ ข้อ
2. ประกอบด้วยสถานการณ์ พฤติกรรมที่คาดหวังและเกณฑ์
3. ครอบคลุมพฤติกรรมเชิงหลักการ หรือพฤติกรรมทั่วไป
4. ครอบคลุมเนื้อหาบทเรียนที่ต้องการวัดผลทั้งหมด
5. เน้นพฤติกรรมปลายทางของบทเรียน
6. สอดคล้องกับจุดประสงค์ของรายวิชาและตรงกับจุดประสงค์ของบทเรียน

ข้อสังเกต ข้อ 1-2 เป็นสมบัติเฉพาะข้อของจุดประสงค์เชิงพฤติกรรม ส่วนข้อ 3-6

เป็นสมบัติรวมของทุกข้อของจุดประสงค์เชิงพฤติกรรม

สรุป การทดสอบต้องสัมพันธ์กับจุดประสงค์ จุดประสงค์ที่เขียนนั้นมีความชัดเจนและระบุพฤติกรรมที่สังเกตเห็นได้โดยตรง ซึ่งมีองค์ประกอบอยู่ 3 ส่วน คือ สถานการณ์ พฤติกรรมที่คาดหวัง และเกณฑ์ การสร้างแบบทดสอบจะยึดจุดประสงค์เชิงพฤติกรรมเป็นแนวทางในการเขียนข้อสอบ เพื่อให้สามารถวัดพฤติกรรมตรงตามที่ระบุไว้ในจุดประสงค์

ในการวิจัยครั้งนี้จะวัดระดับพฤติกรรม แบ่งเป็น 4 ระดับ ได้แก่ ความรู้หรือทักษะการคิดคำนวณ ความเข้าใจ การนำไปใช้ และการวิเคราะห์ ซึ่งแบบทดสอบที่นำมาใช้ในการวิจัยต้องมีความเชื่อมั่น ในการวิจัยครั้งนี้จะใช้แบบทดสอบวิชาคณิตศาสตร์ที่มีการตรวจให้คะแนนแบบ (0,1) ประเมินค่าความเชื่อมั่นด้วยสัมประสิทธิ์ r_B

การแสดงผลที่มาของสัมประสิทธิ์ r_B

บุญเชิด ภิญโญนนตพงษ์ (2542 : 15-19) กล่าวถึงสัมประสิทธิ์ r_B ในการประมาณค่าความเชื่อมั่นของแบบทดสอบแบบคะแนนจริงสัมพันธ์แบบมาตรฐานเดิม ที่พัฒนาจากสัมประสิทธิ์ r_{L2} ของเลียว (บุญเชิด ภิญโญนนตพงษ์. 2542 : 15-19 ; อ้างอิงจาก Liou. 1989) ที่มี K ข้อ ให้คะแนนแบบ (0,1) หรือตอบถูกให้ 1 ตอบผิดหรือไม่ตอบให้ 0 โดยแสดงผลที่มาของสัมประสิทธิ์ r_B ดังนี้

กำหนดให้แบบทดสอบประกอบด้วยข้อสอบ K ข้อ คะแนนสอบ (Observed Scores) คะแนนจริง และคะแนนคลาดเคลื่อนของแต่ละข้อ และของทั้งฉบับแทนด้วยสัญลักษณ์ ดังนี้

คะแนนสอบแต่ละข้อและทั้งฉบับแทนด้วย $X_1, X_2, X_3, \dots, X_K$ และ X

คะแนนจริงแต่ละข้อและทั้งฉบับแทนด้วย $T_1, T_2, T_3, \dots, T_K$ และ T

คะแนนคลาดเคลื่อนแต่ละข้อและทั้งฉบับแทนด้วย $E_1, E_2, E_3, \dots, E_K$ และ E

ดังนั้น (1) $X = X_1 + X_2 + X_3 + \dots + X_K$

(2) $T = T_1 + T_2 + T_3 + \dots + T_K$

(3) $E = E_1 + E_2 + E_3 + \dots + E_K$

จากแบบจำลองคะแนนของทฤษฎีการทดสอบแบบมาตรฐานเดิม (Classical Test Theory Model) ได้คะแนนแต่ละข้อดังนี้

(4) $X_i = T_i + E_i \quad i = 1 \text{ ถึง } K$

และกำหนดข้อตกลงว่า แบบทดสอบประกอบด้วยข้อสอบที่วัดคุณลักษณะเดียวกัน และเป็นไปตามข้อตกลงแบบคะแนนจริงสัมพันธ์ คือ คะแนนจริงของแต่ละข้อมีความสัมพันธ์เป็นเส้นตรงกับคะแนนจริงทั้งฉบับ

(5) $T_i = \lambda_i T + C_i \quad i = 1 \text{ ถึง } K$

เมื่อ λ_i เป็นลักษณะประจำของข้อที่ i ซึ่งกำหนดให้มีค่าเป็นจำนวนจริงบวก

C_i เป็นค่าคงที่ของข้อสอบ i

โดยกำหนดเงื่อนไขว่า $\sum \lambda_i = 1$ และ $\sum C_i = 0$ นำสมการ (5) แทนลงในสมการ (4) ได้

(6) $X_i = \lambda_i T + C_i + E_i \quad ; i = 1 \text{ ถึง } K$

ถ้านำแบบทดสอบไปสอบกับกลุ่มตัวอย่างจำนวนหนึ่ง และหาค่าความแปรปรวนของข้อสอบแต่ละข้อ (σ_i^2) และหาค่าความแปรปรวนของข้อสอบแต่ละข้อ (σ_{Ei}^2) ได้ดังนี้

(7) $\sigma_i^2 = \lambda_i^2 \sigma_T^2 + \sigma_{Ei}^2$

(8) $\sigma_{ji} = \lambda_j \lambda_i \sigma_T^2 \quad (j \neq i)$

ผลรวมของความแปรปรวนของคะแนนสอบทุกข้อคือ

(9) $\sum \sigma_i^2 = \sum \lambda_i^2 \sigma_T^2 + \sum \sigma_{Ei}^2$

นำ σ_X^2 ไปหารสมการ (9) และนิยามความเชื่อมั่นของแบบทดสอบ $\rho_{XX'} = \sigma_T^2 / \sigma_X^2$ จะได้สูตรความเชื่อมั่นซึ่ง ราชู (บุญเชิด ภิญโญนนตพงษ์. 2542 : 15-19 ; อ้างอิงจาก Raju. 1977) เป็นผู้เสนอไว้เป็นคนแรก ดังนี้

$$(10) \rho_{XX'} = \frac{1}{1 - \sum \lambda_i^2} \left\{ 1 - \frac{\sum \sigma_i^2}{\sigma_X^2} \right\}$$

เลียว (Liou) ได้อาศัยวิธีการของ เฟลด์ท์ (Feldt) ที่จะประมาณค่า λ_i โดยพิจารณาว่าข้อสอบแต่ละข้อคือแบบทดสอบที่ลดจำนวนข้อลงจนเหลือข้อสอบเพียงข้อเดียว ดังนั้นข้อสอบก็คือแบบทดสอบฟอร์มหนึ่งนั่นเอง ซึ่งตามทฤษฎีการทดสอบแบบมาตรฐานเดิม ความแปรปรวนของคะแนนคลาดเคลื่อนแต่ละข้อ และทั้งฉบับจะมีความสัมพันธ์กันดังนี้ $\sigma_{Ei}^2 = \lambda_i \sigma_E^2$ และถ้านำคะแนนสอบของแต่ละข้อมาสร้างเป็นเมตริกซ์ความแปรปรวน-ความแปรปรวนร่วมแล้ว จะสามารถหาค่า λ_i ของแต่ละข้อจากแต่ละแถว (Row) ของเมตริกซ์ดังกล่าวได้ โดยนำค่า σ_X^2 ไปหารผลรวมของความแปรปรวน-ความแปรปรวนร่วมในแต่ละแถว ดังนี้

$$(11) \begin{cases} \lambda_1 = \sum \sigma_{1i} / \sigma_X^2 \\ \lambda_2 = \sum \sigma_{2i} / \sigma_X^2 \\ \dots\dots\dots \\ \lambda_K = \sum \sigma_{Ki} / \sigma_X^2 \end{cases}$$

จากสมการ (11) นำมาเขียนใหม่โดยแยกความแปรปรวนร่วมของคะแนนตัวมันเองออกมา ซึ่งก็คือ ค่าความแปรปรวนของข้อนั้น เช่น

$$(12) \lambda = \frac{\sigma_1^2 + \sum \sigma_{1i}}{\sigma_X^2} \quad (i \neq 1)$$

ถ้าสามารถหาค่าความแปรปรวนของคะแนนข้อ 1 กับข้ออื่นๆ ทั้งหมดได้ ก็สามารถหาค่า λ_i ได้ การหาค่าความแปรปรวนร่วมของข้อสอบแต่ละข้อจากสูตรโดยตรง มีความยุ่งยากและใช้เวลานาน วิธีที่ง่ายและรวดเร็วกว่า คืออาศัยคำนิยามของสัมประสิทธิ์สหสัมพันธ์ $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$ ดังนั้นความแปรปรวนร่วมของข้อสอบข้อ 1 และข้อ 2 ได้จากการนำค่าส่วนเบี่ยงเบนมาตรฐานคูณค่าสัมประสิทธิ์สหสัมพันธ์คือ $\sigma_{12} = \sigma_1 \sigma_2 \rho_{12}$ แต่เนื่องจากข้อสอบแต่ละข้อให้คะแนนในระบบ (0,1) ดังนั้น

$$(13) \sigma_1 = \sqrt{P_1(1-P_1)} \quad \text{และ} \quad \sigma_2 = \sqrt{P_2(1-P_2)}$$

เมื่อ P เป็นสัดส่วนของผู้ตอบถูก หรือความยากง่ายของข้อสอบ

สำหรับสัมประสิทธิ์สหสัมพันธ์จะต้องใช้สูตรกรณีเฉพาะของสัมประสิทธิ์สหสัมพันธ์ของ Pearson คือ สัมประสิทธิ์ ϕ ซึ่งมีสูตรดังนี้

$$(14) \phi = \frac{BC - AD}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

เมื่อ A, B, C และ D เป็นสัดส่วนจำนวนผู้ตอบข้อสอบสองข้อ ถูกหรือผิดดังแสดงในตาราง

		ข้อ 2		
		0	1	
ข้อ 1	1	A	B	A+B
	0	C	D	C+D
		A+C	B+D	1.0

จากตารางจะได้ว่า $A + B = P_1$,

$$B + D = P_2,$$

$$C + D = 1 - P_1,$$

$$A + C = 1 - P_2$$

และจัดค่าต่างๆ อยู่ในพจน์ของ P_1, P_2 และ B จะได้

$$A = P_1 - B,$$

$$D = P_2 - B,$$

$$C = 1 - P_1 - P_2 + B$$

นำค่าดังกล่าวแทนในสมการ (14) ได้

$$\phi_{12} = \frac{B_{12}(1 - P_1 - P_2 + B_{12}) - (P_1 - B_{12})(P_2 - B_{12})}{\sqrt{P_1(1 - P_1)(1 - P_2)P_2}}$$

$$(15) \quad \phi_{12} = \frac{B_{12} - P_1 P_2}{\sqrt{P_1(1 - P_1)}\sqrt{P_2(1 - P_2)}}$$

ดังนั้นจากสมการ (13) และ (15) ได้ความแปรปรวนร่วมดังนี้

$$\sigma_{12} = \sqrt{P_1(1 - P_1)}\sqrt{P_2(1 - P_2)} \frac{B_{12} - P_1 P_2}{\sqrt{P_1(1 - P_1)}\sqrt{P_2(1 - P_2)}}$$

$$\sigma_{12} = B_{12} - P_1 P_2$$

ทำนองเดียวกัน

$$\sigma_{13} = B_{13} - P_1 P_3$$

$$\sigma_{14} = B_{14} - P_1 P_4$$

หรือเขียนในรูปทั่วไปดังนี้ $\sigma_{1i} = B_{1i} - P_1 P_i$

$$\sigma_{2i} = B_{2i} - P_2 P_i$$

.....

$$\sigma_{Ki} = B_{Ki} - P_K P_i$$

ดังนั้นจากสมการ (11) และ (12) นำมาเขียนใหม่ได้

$$\begin{aligned}\lambda_i &= \frac{P_1(1-P_1) + \sum(B_{1i} - P_1P_i)}{\sigma_X^2} && ; i \neq 1 \\ &= \frac{P_1(1-P_1) + \sum B_{1i} - P_1 \sum P_i}{\sigma_X^2} && ; i \neq 1 \\ &= \frac{P_1(1-P_1 - \sum P_i) + \sum B_{1i}}{\sigma_X^2} && ; i \neq 1 \\ &= \frac{P_1(1 - \bar{X}) + \sum B_{1i}}{\sigma_X^2} && ; i \neq 1 \\ \lambda_1 &= \frac{\sum B_{1i} - P_1(\bar{X} - 1)}{\sigma_X^2} && ; i \neq 1\end{aligned}$$

ทำนองเดียวกัน

$$\begin{aligned}\lambda_2 &= \frac{\sum B_{2i} - P_2(\bar{X} - 1)}{\sigma_X^2} && ; i \neq 2 \\ &\dots\dots\dots \\ \lambda_K &= \frac{\sum B_{Ki} - P_K(\bar{X} - 1)}{\sigma_X^2} && ; i \neq K\end{aligned}$$

ดังนั้นการประมาณค่าความเชื่อมั่นแบบคะแนนจริงสัมพันธ์แบบมาตรฐานเดิม (Classical-Congeneric) จากแบบทดสอบที่ให้คะแนนระบบ (0,1) จำนวน K ข้อ โดยใช้สัมประสิทธิ์ r_B ที่สร้างขึ้นใหม่มีสูตรดังนี้

$$(15) \quad r_B = \frac{1}{1 - \sum \lambda_i^2} \left[1 - \frac{\sum P_i(1 - P_i)}{S_x^2} \right]$$

เมื่อ P_i แทน ค่าความยากง่ายของข้อสอบแต่ละข้อ

S_x^2 แทน ความแปรปรวนของคะแนนทั้งฉบับ

$$\lambda_1 = [\sum B_{1i} - P_1(\bar{X} - 1)] / S_x^2$$

$$\lambda_2 = [\sum B_{2i} - P_2(\bar{X} - 1)] / S_x^2$$

.....

$$\lambda_K = [\sum B_{Ki} - P_K(\bar{X} - 1)] / S_x^2$$

$$\bar{X} = \sum P_i$$

B_{ji} แทน สัดส่วนจำนวนผู้ตอบข้อสอบสองข้อใดๆ ถูกทั้งคู่

3. งานวิจัยที่เกี่ยวข้อง

3.1 งานวิจัยในต่างประเทศ

สำหรับงานวิจัยในต่างประเทศที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบนั้น ผู้วิจัยได้รวบรวมและสรุปไว้ดังนี้

ผจงจิต อินทสุวรรณ (Intasuwarn. 1979 : 92) ศึกษาวิธีการตรวจสอบการทำหน้าที่ต่างกันสามวิธี คือ วิธีโค้งลักษณะข้อสอบ-สามพารามิเตอร์ วิธีราซซ และวิธีไค-กำลังสองของแบบทดสอบ IEA (The International Association for The Evaluation of Education Achivement) กลุ่มตัวอย่างที่ใช้ในการศึกษาเป็นนักเรียนอายุ 14 ปี ชาวอังกฤษ อเมริกัน และนิวซีแลนด์ จำนวน 8,601 คน ผลการศึกษาพบว่า สหสัมพันธ์ระหว่างวิธีทั้งสามมีค่าอยู่ระหว่าง .51-.98 และวิธีไค-กำลังสองกับวิธีราซซมีสหสัมพันธ์กันสูงถึง .98

รัดเนอร์ เก็ตสัน และไนท์ (Rudner, Getson & Knight. 1980 : 1-10) ศึกษาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเจ็ดวิธี คือ วิธีแปลงค่าความยากของข้อสอบสองวิธี วิธีไค-กำลังสองสองวิธี และวิธีโค้งลักษณะข้อสอบสองวิธีกับข้อมูลที่จำลองขึ้น (Simulate Data) ผลการศึกษาพบว่า วิธีโค้งลักษณะข้อสอบ-สามพารามิเตอร์ เป็นวิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ดีที่สุด รองลงมาได้แก่ หนึ่งในสองของวิธีวิเคราะห์ไค-กำลังสอง ส่วนวิธีแปลงค่าความยากของข้อสอบ ผู้วิจัยแนะนำให้ใช้ได้ทางปฏิบัติ

สับโคเวียต และคณะ (Subkoviak, et al. 1984 : 49-59) ศึกษาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสามวิธี คือ วิธีโค้งลักษณะข้อสอบ-สามพารามิเตอร์ วิธีไค-กำลังสอง และวิธีแปลงค่าความยากของแบบทดสอบ วัดความรู้เกี่ยวกับคำศัพท์ 50 ข้อ เป็นคำศัพท์-ภาษาอังกฤษมาตรฐาน จำนวน 40 ข้อ ศัพท์แสงชาวผิวดำจำนวน 10 ข้อ กลุ่มตัวอย่างที่ใช้ในการศึกษาเป็นนักเรียนผิวขาว 1,021 คน นักเรียนผิวดำจำนวน 1,008 คน ผลการศึกษาพบว่า วิธีโค้งลักษณะข้อสอบ-สามพารามิเตอร์ เป็นวิธีที่มีประสิทธิภาพมากที่สุด รองลงมาคือวิธีไค-กำลังสอง ส่วนวิธีแปลงค่าความยากเป็นวิธีที่นำไปใช้ในทางปฏิบัติได้ ในกรณีที่ไม่มีเครื่องคอมพิวเตอร์

คูลิทเทิล และเคลียร์ (Doolittle & Cleary. 1987 : 157-166) ศึกษาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจากแบบทดสอบ ACT (Assessment Mathematics Test) ฉบับการคำนวณทางคณิตศาสตร์ เรขาคณิต และคณิตศาสตร์การใช้เหตุผล จำนวน 8 ฉบับ กลุ่มตัวอย่างที่ใช้ในการศึกษาเป็นเพศชายและหญิง จำนวน 1,300-1,400 คน โดยเป็นเพศหญิงประมาณร้อยละ 55 วิเคราะห์ข้อมูลด้วยดัชนีของ ลินน์ และฮาร์นิช (Linn & Harnisch. 1981) ซึ่งเป็นโมเดลโลจิสติกแบบสามพารามิเตอร์ ถ้าดัชนี Z ที่ได้เป็นบวกแสดงว่าง่ายสำหรับกลุ่มเปรียบเทียบ ถ้าเป็นลบ แสดงว่า ยากสำหรับกลุ่มเปรียบเทียบ ผลการศึกษาพบว่า ค่าดัชนี Z มีค่าเป็นลบสำหรับข้อเสนอที่วัดด้านเรขาคณิต การใช้เหตุผลเชิงพีชคณิต และเรขาคณิตในทุกฉบับ แสดงว่าง่ายสำหรับเพศชาย

สวามินาธาน และโรเจอร์ (Swaminathan & Rogers. 1990 : 361-370) ศึกษาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีถดถอยโลจิสติก (LRDIF) กับวิธีแมนเทิล-แฮนส์เซล โดยศึกษาผลการตรวจสอบการทำหน้าที่ต่างกันแบบสม่ำเสมอ (Uniform DIF) กับแบบไม่สม่ำเสมอ (Nonuniform DIF) จากกลุ่มตัวอย่างที่จำลองผลการตอบข้อสอบด้วยโปรแกรม DATAGEN แบบสามพารามิเตอร์ กลุ่มตัวอย่างที่ใช้มีขนาด 250 และ 500 คนต่อกลุ่ม ความยาวแบบทดสอบมี 3 ขนาด คือ 40 ข้อ 60 ข้อ และ 80 ข้อ โดยแต่ละฉบับจะมีข้อสอบที่ทำหน้าที่ต่างกันฉบับละ 20% เป็นข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอ (Uniform DIF) จำนวน 10% และแบบไม่สม่ำเสมอ (Nonuniform DIF) อีก 10% ผลการศึกษาพบว่า ทั้งสองวิธีให้ผลใกล้เคียงกัน โดยวิธีแมนเทิล-แฮนส์เซลจะดีกว่าเล็กน้อยในขนาดกลุ่มตัวอย่าง 250 คน ตรวจสอบได้ถูกต้องร้อยละ 75 และในขนาดกลุ่มตัวอย่าง 500 คน ตรวจสอบได้ถูกต้องร้อยละ 100 ในทุกความยาวของแบบทดสอบเมื่อข้อสอบมีการทำหน้าที่ต่างกันแบบสม่ำเสมอ แต่ไม่สามารถตรวจสอบได้ในข้อสอบที่มีการทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ สำหรับวิธีถดถอยโลจิสติกจะตรวจสอบได้อย่างถูกต้องประมาณร้อยละ 50 เมื่อกลุ่มตัวอย่างขนาดเล็ก และความยาวของแบบทดสอบสั้น แต่จะตรวจสอบได้ถูกต้องร้อยละ 75 ในแบบทดสอบที่ยาวและกลุ่มตัวอย่างมีขนาดใหญ่ ในด้านความคลาดเคลื่อนแบบที่ 1 (การระบุว่าข้อสอบทำหน้าที่ต่างกันทั้ง ๆ ที่ทำหน้าที่ไม่ต่างกัน) วิธีแมนเทิล-แฮนส์เซลจะตรวจสอบผิดประมาณร้อยละ 1 ส่วนวิธีถดถอยโลจิสติกจะตรวจสอบผิดประมาณร้อยละ 1 ถึง ร้อยละ 6

สำหรับนิยามของการทำหน้าที่ต่างกันแบบสม่ำเสมอ (Uniform DIF) และ การทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ (Nonuniform DIF) เมลเลนเบิร์ก (Mellenbergh. 1989 : 127-143) ได้นิยามข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอ คือ ข้อสอบที่ไม่มีปฏิสัมพันธ์ระหว่างระดับความสามารถของผู้สอบกับกลุ่มผู้สอบซึ่งจะมีผลให้โอกาสในการตอบข้อสอบถูกจะสูงสำหรับกลุ่มหนึ่งมากกว่ากลุ่มอื่นในทุกระดับความสามารถ และข้อสอบที่มีการทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ คือ ข้อสอบที่มีปฏิสัมพันธ์ระหว่างระดับความสามารถของผู้สอบกับกลุ่มผู้สอบ ทำให้ความแตกต่างของโอกาสในการตอบข้อสอบถูกสำหรับสองกลุ่มต่างกันในทุกระดับความสามารถ

ฟรีเดิล และคอสติน (Freedle & Kostin. 1990 : 329-343) ศึกษาวิธีแปลงค่าความยากของข้อสอบเป็นดัชนีบ่งการทำหน้าที่ต่างกันต่อสีผิวของแบบทดสอบทางภาษาของ Graduate Record Examination (GRE) จำนวน 13 ฉบับ 988 ข้อ และแบบทดสอบ Scholastic Aptitude Test (SAT) จำนวน 11 ฉบับ 935 ข้อ ซึ่งเป็นแบบทดสอบวัดความถนัดด้านภาษา 4 ด้าน คือ ด้านอุปมาอุปมัย ด้านคำตรงข้าม การเติมคำในประโยคให้สมบูรณ์ และด้านความเข้าใจในการอ่าน กลุ่มตัวอย่างที่ใช้ในการศึกษาเป็นกลุ่มผู้สอบผิวดำที่ตอบว่าใช้ภาษาอังกฤษเป็นภาษาหลักในแบบทดสอบ GRE และแบบทดสอบ SAT จำนวน 1,688 คน และ 6,332 คน ตามลำดับ เป็นกลุ่มเปรียบเทียบ กลุ่มผู้สอบผิวขาวที่ตอบว่าใช้ภาษาอังกฤษเป็นภาษาหลักในแบบทดสอบทั้ง 2 ฉบับ จำนวน 23,375 คน และ 34,208 คน ตามลำดับ เป็นกลุ่มอ้างอิง ตรวจสอบการทำหน้าที่ต่างกันด้วยวิธีแปลงค่าความยาก (เปลี่ยนค่า p ให้เป็นค่า z และแปลงเป็นค่าเดลต้า) และวิธี Standardization Approach ที่เสนอโดย ดูแรนและคูริค (Freedle & Kostin. 1990 : 329 ; citing Dorans & Kulick. 1986) เพื่อหาขนาดของการทำหน้าที่ต่างกันและใช้การปรับแก้ค่าร้อยละของการตอบถูกด้วยวิธีต่าง ๆ สามแบบ จากนั้น

นำค่าการทำหน้าที่ต่างกันจากทั้งสองวิธีหาความสัมพันธ์เฉลี่ยในแต่ละฉบับของแบบทดสอบ GRE และ SAT แล้วทำการทดสอบนัยสำคัญของค่าสหสัมพันธ์เฉลี่ยที่คำนวณได้ ผลการศึกษาพบว่า ค่าความยากของข้อสอบสามารถนำมาใช้เป็นดัชนีชี้บ่งขนาดการทำหน้าที่ต่างกันได้ในแบบทดสอบ GRE และ SAT กลุ่มผู้สอบผิวดำจะทำข้อสอบที่มีความยากสูงได้ดีกว่ากลุ่มผู้สอบผิวขาวแต่จะทำข้อสอบที่ง่ายได้ดีน้อยกว่าในระดับความสามารถเดียวกัน เนื่องจากในข้อสอบที่ง่ายมักใช้คำที่มีความหมายหลายอย่าง จึงมีผลต่อค่าความยากและเป็นแหล่งที่ก่อให้เกิดการทำหน้าที่ต่างกันของข้อสอบ สำหรับการใช่วิธีปรับแก้สามแบบ ในการคำนวณหาขนาดการทำหน้าที่ต่างกันกับการคำนวณหาขนาดการทำหน้าที่ต่างกันโดยไม่ใช้วิธีปรับแก้ ให้ผลไม่แตกต่างกัน

เมเซอร์ คลอสเซอร์ และแฮมเบิลตัน (Mazor, Clauser & Hambleton. 1992 : 443-451) ศึกษาผลกระทบของขนาดกลุ่มตัวอย่างที่มีต่อการตรวจสอบการทำหน้าที่ต่างกันด้วยวิธีแมนเทิล-แฮนส์เซล โดยใช้ข้อมูลที่จำลองด้วยโปรแกรม DATAGEN แบบสามพารามิเตอร์ ขนาดกลุ่มตัวอย่างที่ศึกษามี 5 ขนาด คือ 2,000 คน 1,000 คน 500 คน 200 คน และ 100 คน ความยาวแบบทดสอบชุดละ 75 ข้อ ผลการศึกษาพบว่า เมื่อใช้ขนาดกลุ่มตัวอย่าง 2,000 คน วิธีแมนเทิล-แฮนส์เซลจะตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ผิดพลาดร้อยละ 30 และเมื่อใช้จำนวนผู้สอบน้อยกว่า 500 คน พบว่า ตรวจสอบการทำหน้าที่ต่างกันผิดพลาดร้อยละ 50 และข้อสอบที่ไม่สามารถตรวจสอบการทำหน้าที่ต่างกันได้ เป็นข้อสอบที่ยากมาก ข้อสอบที่มีค่าอำนาจจำแนกต่ำ และข้อสอบที่มีค่าความยากต่างกันเล็กน้อยสำหรับสองกลุ่ม

มิลเลอร์ และโอชิม่า (Miller & Oshima. 1992 : 381-388) ศึกษาผลของขนาดกลุ่มตัวอย่าง จำนวนข้อสอบที่ทำหน้าที่ต่างกัน และขนาดของการทำหน้าที่ต่างกันที่ประมาณค่าการทำหน้าที่ต่างกันด้วยวิธีสองขั้นตอน ข้อมูลที่ใช้เป็นที่จำลองขึ้นแบบสองพารามิเตอร์ จำนวน 40 ข้อ มีค่าอำนาจจำแนกเฉลี่ยเป็น 1.13 ซึ่งมีค่าระหว่าง .41 ถึง 3.82 ส่วนเบี่ยงเบนมาตรฐานเป็น .63 ค่าความยากเฉลี่ยเป็น .25 มีค่าระหว่าง -1.99 ถึง 2.10 ส่วนเบี่ยงเบนมาตรฐานเป็น 1.05 เงื่อนไขการทำหน้าที่ต่างกันที่ศึกษามี 24 เงื่อนไข ดังนี้ พารามิเตอร์ของข้อสอบแบบสองพารามิเตอร์ ขนาดการทำหน้าที่ต่างกันของข้อสอบ 3 ขนาด คือ น้อย ปานกลาง และแบบผสม โดยที่เงื่อนไขขนาดการทำหน้าที่ต่างกันแบบน้อยจะมีค่าความยากเพิ่มขึ้น .20 ในขนาดการทำหน้าที่ต่างกันแบบปานกลางค่าความยากเพิ่มขึ้น .35 และขนาดการทำหน้าที่ต่างกันของข้อสอบแบบผสมเป็นแบบผสมระหว่างสองแบบ และเงื่อนไขจำนวนข้อที่ทำหน้าที่ต่างกันสี่แบบ คือ ร้อยละ 5 ร้อยละ 10 ร้อยละ 20 และ ร้อยละ 40 ของแบบทดสอบ ($2 \times 3 \times 4$) ดัชนีการทำหน้าที่ต่างกันคำนวณจากวิธีที่ใช้ทฤษฎีการตอบข้อสอบห้าวิธี คือ วิธี SA, UA, SSOS, USOS, WSSOS และวิธีแมนเทิล-แฮนส์เซล ขั้นที่ 2 ของวิธีการเป็นการหาจำนวนข้อสอบที่ผิดพลาดทางบวก (ตรวจสอบข้อสอบที่ไม่ทำหน้าที่ต่างกันว่าเป็นข้อสอบที่ทำหน้าที่ต่างกัน) ในขั้นตอนนี้จะมีประโยชน์เมื่อขนาดการทำหน้าที่ต่างกันมีมากและจำนวนข้อสอบที่ทำหน้าที่ต่างกันมีมาก (ร้อยละ 20 หรือ ร้อยละ 40 ของแบบทดสอบ) วิธีแมนเทิล-แฮนส์เซลจะตรวจสอบ ความผิดพลาดทางบวกได้น้อย และวิธีแมนเทิล-แฮนส์เซลจะตรวจสอบการทำหน้าที่ต่างกันได้ดีพอ ๆ กับวิธีทฤษฎีการตอบข้อสอบ

โรเจอร์ และสวามินาธาน (Rogers & Swaminathan. 1993 : 105-116) ศึกษาเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีถดถอยโลจิสติก (LRDIF) กับวิธีแมนเทิล-แฮนส์เซล โดยใช้ข้อมูลที่จำลองขึ้น เงื่อนไขการศึกษาเป็น 4 แบบ คือ ขนาดกลุ่มตัวอย่างสองกลุ่ม ขนาด 250 คน และ 500 คน ต่อกลุ่ม ระดับความเหมาะสมของโมเดล ข้อมูลเป็น 2 ระดับ คือ ระดับที่เหมาะสมใช้รูปแบบสองพารามิเตอร์ ระดับที่ไม่เหมาะสมใช้รูปแบบสามพารามิเตอร์โดย ค่าการเดา (c) จะถูกกำหนดที่ .2 เครื่องมือที่ใช้เป็นข้อสอบจำนวน 40 ข้อ ที่จำลองขึ้นมา 7 แบบ ดังนี้ แบบที่ 1 ค่าความยากและค่าอำนาจจำแนกต่ำ แบบที่ 2 ค่าความยากสูง ค่าอำนาจจำแนกต่ำ แบบที่ 3 ค่าความยากปานกลางค่าอำนาจจำแนกต่ำ แบบที่ 4 ค่าความยากปานกลางค่าอำนาจจำแนกสูง แบบที่ 5 ค่าความยากสูงค่าอำนาจจำแนกต่ำ แบบที่ 6 ค่าความยากและค่าอำนาจจำแนกสูง แบบที่ 7 เป็นแบบผสม คือ ค่าอำนาจจำแนกและค่าความยากต่างกันสำหรับสองกลุ่ม โดยวิเคราะห์ร้อยละของการทำหน้าที่ต่างกันของข้อสอบที่ตรวจสอบแบบสม่ำเสมอและไม่สม่ำเสมอ ผลการศึกษาพบว่า ร้อยละของการทำหน้าที่ต่างกันของข้อสอบที่ตรวจสอบแบบสม่ำเสมอสูงกว่าแบบไม่สม่ำเสมอทั้งสองวิธี และในทุกเงื่อนไข โดยวิธีถดถอยโลจิสติกจะตรวจสอบร้อยละของการทำหน้าที่ต่างกันของข้อสอบสูงที่ระดับความยากปานกลางค่าอำนาจจำแนกสูง สำหรับการทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ พบว่า วิธีถดถอยโลจิสติกจะตรวจสอบการทำหน้าที่ต่างกันของข้อสอบลดลงร้อยละ 14 ในโมเดลข้อมูลที่ไม่เหมาะสม ส่วนวิธีแมนเทิล-แฮนส์เซลจะตรวจสอบพบลดลงร้อยละ 7 และในขนาดกลุ่มตัวอย่างที่เพิ่มขึ้น วิธีถดถอยโลจิสติกตรวจสอบการทำหน้าที่ต่างกันได้เพิ่มขึ้นร้อยละ 19 ส่วนวิธีแมนเทิล-แฮนส์เซลจะตรวจสอบเพิ่มขึ้นร้อยละ 11 และที่ความยากต่ำ วิธีแมนเทิล-แฮนส์เซลตรวจสอบพบได้น้อยกว่าวิธีถดถอยโลจิสติกแต่ที่ความยากสูง ทั้งสองวิธีให้ผลใกล้เคียงกัน แต่วิธีแมนเทิล-แฮนส์เซลไม่สามารถตรวจสอบได้ในระดับความยากปานกลาง และวิธีถดถอยโลจิสติกจะตรวจสอบการทำหน้าที่ต่างกันได้ดีกว่าวิธีแมนเทิล-แฮนส์เซล แต่ค่าใช้จ่ายที่ใช้ในการตรวจสอบด้วยวิธีถดถอยโลจิสติกสูงกว่าวิธีแมนเทิล-แฮนส์เซล

รูสโซ และสเตาต์ (Roussos & Stout. 1996 : 215-230) ศึกษาผลการใช้ขนาดกลุ่มตัวอย่างขนาดเล็ก ความแตกต่างของค่าพารามิเตอร์ข้อสอบ รวมทั้งความแตกต่างในค่าเฉลี่ยของความสามารถของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบเพื่อศึกษาผลที่มีต่อความคลาดเคลื่อนแบบที่ 1 (Type I error) (การยอมรับว่าข้อสอบทำหน้าที่ต่างกันทั้ง ๆ ที่ทำหน้าที่ไม่ต่างกัน) ด้วยวิธีแมนเทิล-แฮนส์เซลและวิธีซิปเทสต์ ใช้ข้อมูลจำลอง โดยศึกษา 2 ครั้ง ครั้งแรกใช้ขนาดกลุ่มตัวอย่างกลุ่มละ 100, 200, 500 และ 1,000 คน จากแบบทดสอบ ASVAB จำนวน 25 ข้อ ความแตกต่างระหว่างค่าเฉลี่ยของการแจกแจงระหว่างสองกลุ่ม ถูกจำลองให้มีค่า 0.0, 0.5 และ 1.0 ผลการศึกษาพบว่า อัตราการเกิดความคลาดเคลื่อนแบบที่ 1 ของวิธีทั้งสองแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ การศึกษาที่ 2 ใช้แบบทดสอบชุดเดิม ทำการจำลองการแจกแจงความสามารถให้เป็น 2 แบบคือ 0.0 และ 1.0 นอกนั้นจำลองทุกอย่างเช่นเดียวกับการศึกษาที่ 1 แต่ใช้จำนวนผู้สอบขนาดใหญ่คือ 500, 1,000 และ 3,000 คน ผลการศึกษาพบว่า เมื่อค่าอำนาจจำแนกสูงขึ้นและค่าความยากต่ำ อัตราการเกิดความคลาดเคลื่อนประเภทที่ 1 จะสูงทั้ง 2 วิธีที่ระดับนัยสำคัญ .05 สรุปได้ว่า เมื่อใช้กลุ่มตัวอย่างขนาดเล็ก ไม่เกิน 1,000 คน วิธีทั้งสองให้ผลด้านความคลาดเคลื่อนประเภทที่ 1

ใกล้เคียงกัน แต่เมื่อใช้ขนาดกลุ่มตัวอย่างมากกว่า 1,000 คน และค่าเฉลี่ยการแจกแจงความสามารถของทั้งสองกลุ่มแตกต่างกัน รวมทั้งข้อสอบมีค่าความยากต่ำ ค่าอำนาจจำแนกสูง วิธีชิปเทสต์มีอัตราการเกิดความคลาดเคลื่อนประเภทที่ 1 ต่ำกว่าวิธีแมนเทิล-แฮนส์เซล

นารายานาน และสวามินาทาน (Narayanan & Swaminathan. 1994 : 315-328) ศึกษาเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันแบบสมำเสมอด้วยวิธีแมนเทิล-แฮนส์เซล และวิธีชิปเทสต์ โดยใช้ข้อมูลที่จำลองภายใต้เงื่อนไข 1,296 เงื่อนไข ($9 \times 3 \times 2 \times 4 \times 6$) แล้วคำนวณซ้ำ 100 ครั้งในแต่ละเงื่อนไข เงื่อนไขที่ทำการศึกษาประกอบด้วย ขนาดกลุ่มตัวอย่างที่ต่างกัน 3 ขนาด คือ กลุ่มเปรียบเทียบใช้ขนาด 100, 200 และ 300 คน กลุ่มอ้างอิงใช้ขนาด 300, 500 และ 1,000 คน วางไขว้กันได้กลุ่มตัวอย่างทั้งหมด 9 กลุ่ม ความแตกต่างในการแจกแจงความสามารถ 3 ระดับ ระดับที่ 1 ให้ทั้งสองกลุ่มมีค่าเฉลี่ยการแจกแจงความสามารถอยู่ที่ 0.0 ระดับที่ 2 ให้ค่าเฉลี่ยของความสามารถในกลุ่มอ้างอิงและกลุ่มเปรียบเทียบเป็น 0.0 และ 0.5 ตามลำดับ และระดับที่ 3 ค่าเฉลี่ยของความสามารถในกลุ่มอ้างอิงและกลุ่มเปรียบเทียบเป็น 0.0 และ -1.0 ตามลำดับ ส่วนเบี่ยงเบนมาตรฐานของทั้ง 3 ระดับ เป็น 1.0 ร้อยละของข้อสอบที่ทำหน้าที่ต่างกันมี 2 ขนาด คือ ร้อยละ 10 และร้อยละ 20 จากแบบทดสอบ จำนวน 40 ข้อ ขนาดของการทำหน้าที่ต่างกัน 4 ระดับ คือ 0.4, 0.6, 0.8 และ 1.0 ชนิดของแบบทดสอบ 6 ชนิด คือ ชนิดของแบบทดสอบที่ประกอบด้วย ค่าความยากและค่าอำนาจจำแนก 3 ระดับ คือ สูง ปานกลาง ต่ำ วางไขว้กัน แต่จัดแบบทดสอบที่ประกอบด้วย ค่าความยากและค่าอำนาจจำแนกสูง ปานกลาง ต่ำ ทั้ง 2 ค่าออก จำนวน 3 ฉบับ ผลการศึกษาพบว่า วิธีแมนเทิล-แฮนส์เซลและวิธีชิปเทสต์ สามารถตรวจสอบการทำหน้าที่ต่างกันได้ใกล้เคียงกัน โดยจะตรวจสอบได้ดีในขนาดกลุ่มตัวอย่าง 600 คน (กลุ่มอ้างอิงและกลุ่มเปรียบเทียบกลุ่มละ 300 คน) ขนาดของการทำหน้าที่ต่างกันทั้ง 4 ขนาด มีผลต่อการตรวจสอบการทำหน้าที่ต่างกันทั้งสองวิธีอย่างมีนัยสำคัญทางสถิติ ความแตกต่างในการแจกแจงความสามารถไม่มีผลต่อวิธีการวิเคราะห์การทำหน้าที่ต่างกันด้วยวิธีชิปเทสต์ แต่มีผลต่อวิธีแมนเทิล-แฮนส์เซล ที่เป็นเช่นนี้อาจเป็นเพราะวิธีชิปเทสต์ได้มีการปรับแก้การถดถอยขนาดของแบบทดสอบ และร้อยละของข้อสอบที่ทำหน้าที่ต่างกันไม่มีผลต่อวิธีทั้งสอง สำหรับการเกิดความคลาดเคลื่อนแบบที่ 1 จะมีอัตราการเกิดอยู่ในระดับปกติ แต่วิธีชิปเทสต์จะเพิ่มขึ้นเล็กน้อยในกลุ่มที่มีความสามารถเท่ากัน แต่ในกลุ่มที่มีความสามารถไม่เท่ากันเกิดความคลาดเคลื่อนแบบที่ 1 จะมีอัตราการเกิดสูงขึ้นทั้งสองวิธี ซึ่งการเพิ่มความยากของแบบทดสอบจะทำให้เกิดความเชื่อมั่นสูงขึ้น ซึ่งมีผลให้ความคลาดเคลื่อนแบบที่ 1 น่าจะมีอัตราการเกิดลดลง

โอชิม่า, ราจุ, เฟาเวอร์ส และสไลนด์ (Oshima, Raju, Flowers & Slinde. 1998 : 353-369) ศึกษาการรอบคอบคิดการทำหน้าที่ต่างกันของข้อสอบและแบบทดสอบ (DFIT) เพื่อขยายการทำหน้าที่ต่างกันของหมวดข้อสอบที่ระบุแหล่งที่เป็นไปได้ของการทำหน้าที่ต่างกัน ศึกษาจากกลุ่มตัวอย่าง 4,000 คน เป็นนักเรียนเกรด 4 จำแนกเป็นเพศชาย 1,000 คน เพศหญิง 1,000 คน กลุ่มที่มีฐานะทางเศรษฐกิจสูง 1,000 คน กลุ่มที่มีฐานะทางเศรษฐกิจต่ำ 1,000 คน โดยใช้แบบทดสอบความเข้าใจในการอ่านภาษาอังกฤษ 55 ข้อ เป็นการพัฒนาเทคนิคการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) เป็นการตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบ (DBF)

3.2 งานวิจัยในประเทศ

สำหรับงานวิจัยในประเทศไทยที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ หรือ การวิเคราะห์ความลำเอียงของข้อสอบ มีดังนี้

พัชรีย์ ปิยภักดิ์ (2531 : 64-69) ทำการวิจัยเรื่อง “การวิเคราะห์ความลำเอียงของข้อสอบจากแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียน” ศึกษาความลำเอียงของข้อสอบจากแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนคณิตศาสตร์ โดยมีตัวแปรเพศชายและหญิงกับตัวแปรนักเรียนในกรุงเทพมหานคร กับนักเรียนในจังหวัดสมุทรสาคร โดยใช้วิธีวิเคราะห์สามวิธี คือ วิธีแปลงค่าความยากของข้อสอบ วิธีไค-กำลังสอง และวิธีโค้งลักษณะข้อสอบแบบหนึ่งพารามิเตอร์กับกลุ่มตัวอย่างนักเรียนชั้นประถมศึกษาปีที่ 6 ผลการศึกษาพบว่า ทั้งสามวิธี จำนวนข้อสอบที่ลำเอียงต่างกัน วิธีโค้งลักษณะข้อสอบพบจำนวนข้อสอบที่ลำเอียงมากที่สุด และค่าความเชื่อมั่นของแบบทดสอบก่อนคัดเลือกข้อสอบที่ลำเอียงออกมีค่า 0.9277 เมื่อคัดเลือกข้อที่ลำเอียงออกระหว่างกลุ่มนักเรียนหญิงและนักเรียนชาย โดยวิธีแปลงค่าความยากของข้อสอบ วิธีไค-กำลังสอง และวิธีโค้งลักษณะข้อสอบมีค่า 0.9316, 0.9277 และ 0.9286 ตามลำดับ และค่าความเชื่อมั่นของแบบทดสอบหลังการคัดเลือกข้อสอบที่ลำเอียงออกระหว่างกลุ่มนักเรียนกรุงเทพมหานครกับกลุ่มนักเรียนสมุทรสาครมีค่า 0.9339, 0.9311 และ 0.9392 ตามลำดับ

เจน จันทรแสง (2537 : บทคัดย่อ) ทำการวิจัยเรื่อง “การวิเคราะห์ความลำเอียงของข้อสอบของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนคณิตศาสตร์ ชั้นมัธยมศึกษาปีที่ 3” ศึกษาเปรียบเทียบวิธีวิเคราะห์ความลำเอียงสี่วิธี คือ วิธีแปลงค่าความยากของข้อสอบ วิธีวิเคราะห์ความแปรปรวน วิธีไค-กำลังสอง และวิธีโค้งลักษณะข้อสอบ-หนึ่งพารามิเตอร์ ศึกษาสัมประสิทธิ์สหสัมพันธ์ระหว่างดัชนีความลำเอียงทั้งสี่วิธี และเปรียบเทียบความเชื่อมั่นของแบบทดสอบก่อนและหลังการคัดเลือกข้อสอบที่ลำเอียง ผลการศึกษาพบว่า วิธีโค้งลักษณะข้อสอบ-หนึ่งพารามิเตอร์ตรวจสอบความลำเอียงมากที่สุด รองลงมา คือวิธีวิเคราะห์ความแปรปรวน และวิธีไค-กำลังสอง และแบบทดสอบที่คัดเลือกข้อสอบที่ลำเอียงออกไปแล้วมีค่าความเชื่อมั่นลดลงจากเดิมอย่างไม่มีนัยสำคัญทางสถิติ เมื่อปรับค่าความเชื่อมั่นของแบบทดสอบหลังจากคัดเลือกข้อสอบที่ลำเอียงออกแล้ว ค่าความเชื่อมั่นแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

นิรมล ชัยชวลิต (2537 : 57-58) ทำการวิจัยเรื่อง “การเปรียบเทียบผลการวิเคราะห์ความลำเอียงของแบบทดสอบความเข้าใจในการอ่านภาษาไทย ตามทฤษฎีคลาสสิกอลที่ใช้วิธีวิเคราะห์ต่างกัน” ศึกษาเปรียบเทียบผลการวิเคราะห์ความลำเอียงระหว่างกลุ่มผู้สอบที่มีเพศต่างกัน ด้วยวิธีวิเคราะห์ความลำเอียงของข้อสอบตามทฤษฎีคลาสสิกอลสามวิธี คือ วิธีแปลงค่าความยาก วิธีไค-กำลังสอง และวิธีวิเคราะห์ความแปรปรวน ผลการศึกษาพบว่า การวิเคราะห์ความลำเอียงด้วยวิธีแปลงค่าความยากกับวิธีไค-กำลังสอง และ วิธีแปลงค่าความยากกับวิธีวิเคราะห์ความแปรปรวนมีจำนวนข้อที่มีความลำเอียงแตกต่างกันมากอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ส่วนวิธีไค-กำลังสองกับวิธีวิเคราะห์ความแปรปรวน พบจำนวนข้อที่มีความลำเอียงแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติที่ระดับ .05 สำหรับค่าความเชื่อมั่นของแบบทดสอบหลังคัดข้อที่

ลำเอียงออกแล้วที่คำนวณด้วยสูตร KR-20 และแบบแบ่งครึ่งฉบับที่วิเคราะห์ด้วยวิธีทั้ง 3 พบว่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติที่ระดับ .05

อังคณา สายยศ (2539 : บทคัดย่อ) ทำการวิจัยเรื่อง “การศึกษาผลการวิเคราะห์ความลำเอียงของแบบทดสอบความเข้าใจในการอ่านตามลักษณะการใช้ภาษาพูดในครอบครัว” ศึกษาผลการวิเคราะห์ความลำเอียงด้วยวิธีแมนเทิล-แฮนส์เซล กลุ่มตัวอย่างเป็นนักเรียนชั้นประถมศึกษาปีที่ 6 ในจังหวัดสุรินทร์ ได้มาโดยการสุ่มแบบแบ่งชั้น อำเภอเป็นชั้นและโรงเรียนเป็นหน่วยการสุ่ม ได้จำนวนโรงเรียนทั้งสิ้น 29 โรงเรียน นักเรียน 860 คน ผลการศึกษาพบว่า จำนวนข้อสอบที่ลำเอียงของแบบทดสอบความเข้าใจในการอ่านระหว่างกลุ่มผู้สอบที่ใช้ภาษาพูดเป็นภาษาลาว ภาษาเขมร และภาษาส่วย แตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ สำหรับความเชื่อมั่นของแบบทดสอบหลังคัดข้อลำเอียงออกแล้วพบว่า ค่าความเชื่อมั่นระหว่างกลุ่มผู้สอบที่ใช้ภาษาพูดเป็นภาษาลาว ภาษาเขมร และภาษาส่วยแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ ส่วนคะแนนที่ได้จากการสอบด้วยแบบทดสอบที่คัดข้อลำเอียงออกแล้วมีคะแนนระหว่างกลุ่มผู้สอบที่พูดภาษาลาวกับภาษาส่วย ภาษาเขมรกับภาษาส่วยมีค่าความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ ส่วนภาษาลาวกับภาษาเขมรมีคะแนนแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

คมศักดิ์ ชื่นชม (2539 : 108) ทำการวิจัยเรื่อง “การศึกษาผลการวิเคราะห์ความลำเอียงที่ใช้วิธีต่างกันของแบบทดสอบวัดจริยธรรมด้านความซื่อสัตย์” ศึกษาเปรียบเทียบผลการวิเคราะห์ความลำเอียงของแบบทดสอบตามทฤษฎีคลาสสิกอล ที่ใช้วิธีต่างกันสามวิธี คือ วิธีค่าอำนาจจำแนกของข้อสอบ วิธีวิเคราะห์องค์ประกอบ และวิธีวิเคราะห์การถดถอย โดยเปรียบเทียบจำนวนข้อที่มีความลำเอียงและค่าความเชื่อมั่นแบบสัมประสิทธิ์แอลฟา (α -Coefficient) หลังคัดเลือกข้อที่มีความลำเอียงออก กลุ่มตัวอย่างที่ใช้ในการศึกษาครั้งนี้เป็นนักเรียนชั้นประถมศึกษาปีที่ 6 ปีการศึกษา 2538 ในจังหวัดยะลา จำนวน 1,081 คน ซึ่งได้มาโดยการสุ่มแบบแบ่งชั้น ผลการศึกษาพบว่า การวิเคราะห์ความลำเอียงของข้อสอบด้วยวิธีค่าอำนาจจำแนกของข้อสอบ วิธีวิเคราะห์องค์ประกอบ และวิธีวิเคราะห์การถดถอย มีจำนวนข้อที่มีความลำเอียงน้อยและไม่แตกต่างกันที่ระดับนัยสำคัญทางสถิติ .05 ค่าความเชื่อมั่นของแบบทดสอบวัดจริยธรรมด้านความซื่อสัตย์ หลังคัดเลือกข้อที่มีความลำเอียงออก จากการวิเคราะห์ความลำเอียงของข้อสอบด้วยวิธีค่าอำนาจจำแนกของข้อสอบ วิธีวิเคราะห์องค์ประกอบและวิธีวิเคราะห์การถดถอยมีค่าไม่แตกต่างกัน ที่ระดับนัยสำคัญทางสถิติ .05

เรวดี อินทสระ (2539 : บทคัดย่อ) ทำการวิจัยเรื่อง “ผลการตรวจสอบความลำเอียงข้อสอบต่อการศึกษาความเที่ยงตรงเชิงพยากรณ์ของแบบทดสอบคัดเลือกที่คิดคะแนนต่างกัน” ศึกษาความเที่ยงตรงเชิงพยากรณ์ของแบบทดสอบคัดเลือกที่วิเคราะห์ความลำเอียงต่อเพศ ด้วยวิธีทฤษฎีการตอบข้อสอบ (IRT) วิธีแมนเทิล-แฮนส์เซล และวิธีชิปเทสต์ การตัดสินผลการสอบที่คิดคะแนนมาตรฐานที่ปกติ และคะแนนนำหน้าความสามารถและสาเหตุความลำเอียงของข้อสอบ โดยศึกษาความลำเอียงของข้อสอบคัดเลือกเข้าศึกษาในชั้นปีที่ 1 ประเภทรับตรง ปีการศึกษา 2538 ของมหาวิทยาลัยสงขลานครินทร์ ในวิชาภาษาไทย ก วิชาสังคมศึกษา ก วิชาภาษาอังกฤษ กข วิชาละ 8,127 คน (ชาย 2,722 คน หญิง 5,405 คน) วิชาภาษาไทย กข วิชาสังคมศึกษา กข และ วิชาภาษาอังกฤษ กขค วิชาละ 5,415 คน (ชาย 1,454 คน หญิง 3,961 คน) ความเที่ยงตรง

เชิงพยากรณ์ศึกษาจากคะแนนสอบคัดเลือกกับระดับผลการเรียนภาคเรียนที่ 1 ปีการศึกษา 2538 ของนักศึกษาที่ได้รับการคัดเลือกจากประเภทรับตรงสายวิทยาศาสตร์ 763 คน และสายศิลปศาสตร์ 281 คน และสาเหตุความลำเอียงของข้อสอบจากการระบุสาเหตุของนักวัดผลการศึกษาหรือ อาจารย์ผู้สอนจำนวน 50 คน และนิสิตนักศึกษาที่เรียนในสาขาวิชานั้นๆ วิชาละ 50 คน ผลการศึกษาพบว่า วิธีตรวจสอบความลำเอียงทั้งสามวิธี ดัดสินจำนวนข้อสอบที่ลำเอียงแตกต่างกันในวิชาภาษาไทย ก ฉบับที่ 2 และวิชาสังคมศึกษา ก ฉบับที่ 1 ที่ระดับนัยสำคัญทางสถิติที่ .05 นอกนั้นต่างกันที่ระดับนัยสำคัญทางสถิติที่ .01 โดยวิธีใช้ทฤษฎีการตอบข้อคำถามดัดสินจำนวนข้อสอบที่ลำเอียงได้มากที่สุด ความสัมพันธ์ของลำดับที่ของการสอบไม่ว่าจะคิดคะแนนมาตรฐานที่-ปกติ หรือ คิดคะแนนน้ำหนักความสามารถและใช้ข้อสอบจำนวนทั้งหมดหรือใช้เฉพาะข้อสอบที่ปราศจากความลำเอียงต่างมีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ประสิทธิภาพในการทำนายผลสัมฤทธิ์ทางการเรียนในสายวิทยาศาสตร์ การคิดคะแนนน้ำหนักความสามารถทั้งใช้ข้อสอบจำนวนทั้งหมด และ ใช้เฉพาะข้อสอบที่ปราศจากความลำเอียงมีประสิทธิภาพในการทำนายสูงกว่าการคิดคะแนนมาตรฐานที่-ปกติ ส่วนสายศิลปศาสตร์ การคิดคะแนนมาตรฐานที่-ปกติทั้งใช้ข้อสอบจำนวนทั้งหมด และใช้เฉพาะข้อสอบที่ปราศจากความลำเอียง มีประสิทธิภาพในการทำนายสูงกว่า การคิดคะแนนน้ำหนักความสามารถ และสาเหตุของ ความลำเอียงของข้อสอบต่อเพศทั้งชายและหญิง เกิดจากข้อสอบเป็นข้อคำถามที่ผู้สอบได้รับการฝึกฝนเฉพาะจะมีโอกาสตอบถูกมากกว่า เป็นเรื่องราวที่กลุ่มนั้น ๆ สนใจ และเป็นข้อสอบที่ถามความจำ

จิตสุตา ธารพร (2539 : บทคัดย่อ) ทำการวิจัยเรื่อง "การเปรียบเทียบผลการวิเคราะห์ความลำเอียงของข้อสอบจากแบบทดสอบโจทย์ปัญหาทางคณิตศาสตร์และแบบทดสอบวัดความเข้าใจทางภาษา" ศึกษาเปรียบเทียบผลการวิเคราะห์ความลำเอียงของข้อสอบสามวิธี คือ วิธีแปลงค่าความยาก วิธีโค้งลักษณะข้อสอบ และวิธีแมนเทิล-แฮนส์เซล กลุ่มตัวอย่างที่ใช้ในการศึกษาเป็นนักเรียนชั้นมัธยมศึกษาปีที่ 2 ในจังหวัดกระบี่ จำนวน 2,689 คน ผลการศึกษาพบว่า จำนวนข้อสอบที่มีความลำเอียงในแบบทดสอบโจทย์ปัญหาทางคณิตศาสตร์ การวิเคราะห์ความลำเอียงด้วยวิธีแปลงค่าความยากของข้อสอบกับวิธีแมนเทิล-แฮนส์เซล มีจำนวนข้อสอบที่มีความลำเอียงแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ส่วนการวิเคราะห์ความลำเอียงด้วยวิธีโค้งลักษณะข้อสอบกับวิธีแมนเทิล-แฮนส์เซล มีจำนวนข้อสอบที่มีความลำเอียงแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ ส่วนจำนวนข้อสอบที่มีความลำเอียงในแบบทดสอบความเข้าใจทางภาษา การวิเคราะห์ความลำเอียงด้วยวิธีแปลงค่าความยากของข้อสอบกับวิธีโค้งลักษณะข้อสอบ และการวิเคราะห์ความลำเอียงด้วยวิธีแปลงค่าความยากของข้อสอบกับวิธีแมนเทิล-แฮนส์เซล มีจำนวนข้อสอบที่มีความลำเอียงแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .001 ส่วนการวิเคราะห์ความลำเอียงด้วยวิธีโค้งลักษณะข้อสอบกับวิธีแมนเทิล-แฮนส์เซล มีจำนวนข้อสอบที่มีความลำเอียงแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ ค่าความเชื่อมั่นของแบบทดสอบโจทย์ปัญหาทางคณิตศาสตร์หลังคัดข้อสอบที่มีความลำเอียงออกด้วยวิธีแปลงค่าความยากของข้อสอบ วิธีโค้งลักษณะข้อสอบ และวิธีแมนเทิล-แฮนส์เซล แตกต่างกันอย่างไม่มีนัยสำคัญ ส่วนค่าความเชื่อมั่นของแบบทดสอบความเข้าใจทางภาษาหลังคัดข้อสอบที่มี

ความลำเอียงออกด้วยวิธีแปลงค่าความยากของข้อสอบ วิธีโค้งลักษณะข้อสอบ และวิธีแมนเทิล-แฮนส์เซล แตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

พรรณี จินตมาศ (2540 : บทคัดย่อ) ทำการวิจัยเรื่อง “การเปรียบเทียบผลการวิเคราะห์ความลำเอียงของข้อสอบ โดยใช้ขนาดกลุ่มผู้สอบและวิธีวิเคราะห์ต่างกัน” ศึกษาเปรียบเทียบการวิเคราะห์ความลำเอียงสามวิธี คือ วิธีแปลงค่าความยาก วิธีแมนเทิล-แฮนส์เซล และวิธีชิปเทสต์ มีขนาดกลุ่มผู้สอบ 500 คน และ 1,000 คน โดยเปรียบเทียบจำนวนข้อสอบที่มีความลำเอียงและเปรียบเทียบค่าความเชื่อมั่นแบบแบ่งครึ่งฉบับของแบบทดสอบหลังคัดเลือกข้อสอบที่มีความลำเอียงออกแล้ว ผลการศึกษาพบว่า เมื่อวิเคราะห์จากกลุ่มผู้สอบขนาด 500 คน วิธีชิปเทสต์พบข้อสอบที่มีความลำเอียงมากที่สุด และวิธีแปลงค่าความยากพบข้อสอบที่มีความลำเอียงน้อยที่สุด โดยจำนวนข้อสอบที่ลำเอียงแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติทุกวิธี เมื่อวิเคราะห์กลุ่มผู้สอบขนาด 1,000 คน วิธีแมนเทิล-แฮนส์เซลพบข้อสอบที่มีความลำเอียงมากที่สุด วิธีแปลงค่าความยากไม่พบข้อสอบที่ลำเอียง โดยจำนวนข้อสอบที่ลำเอียงจากการวิเคราะห์ด้วยวิธีแปลงค่าความยากกับวิธีแมนเทิล-แฮนส์เซล และวิธีแปลงค่าความยากกับวิธีชิปเทสต์ แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นอกนั้นมีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ สำหรับจำนวนข้อสอบที่ลำเอียงจากการวิเคราะห์ด้วยวิธีแปลงค่าความยากระหว่างกลุ่มผู้สอบขนาด 500 คน และกลุ่มผู้สอบขนาด 1,000 คน จะมีจำนวนข้อสอบแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นอกนั้นมีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ ค่าความเชื่อมั่นของแบบทดสอบหลังคัดเลือกข้อสอบที่มีความลำเอียงออกเมื่อวิเคราะห์จากกลุ่มผู้สอบขนาด 500 คน ด้วยวิธีแปลงค่าความยากกับวิธีชิปเทสต์ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 และเมื่อวิเคราะห์จากกลุ่มผู้สอบขนาด 1,000 คน ด้วยวิธีวิเคราะห์ความลำเอียงทั้งสามวิธี มีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

ญาณภัทร สีหะมงคล (2540 : บทคัดย่อ) ทำการวิจัยเรื่อง “การเปรียบเทียบความสอดคล้องของผลการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน ระหว่างวิธี Lord's χ^2 วิธี Raju's Area Measures และวิธี Closed Interval Area” ศึกษาเปรียบเทียบขนาดของกลุ่มตัวอย่าง ความยาวของแบบทดสอบ และ สัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันแบบทดสอบแตกต่างกัน ผลการศึกษาพบว่า จำนวนข้อสอบที่ทำหน้าที่ต่างกันที่ได้จากการตรวจสอบ ระหว่างวิธี Lord's χ^2 วิธี Raju's Area Measures และวิธี Closed Interval Area แตกต่างกันเมื่อขนาดของกลุ่มตัวอย่างและความยาวของแบบทดสอบต่างกัน ยกเว้นกรณีกลุ่มตัวอย่าง 500 คน แบบทดสอบยาว 20, 30 และ 40 ข้อ และกรณีกลุ่มตัวอย่าง 700 คน แบบทดสอบความยาว 40 ข้อ และอัตราความสอดคล้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อแบบทดสอบยาวมากขึ้น และสัดส่วนจำนวนข้อสอบที่ทำหน้าที่ต่างกันมากขึ้น อัตราความสอดคล้องชนิดที่ 1 ส่วนใหญ่ลดลง แต่อัตราความสอดคล้องชนิดที่ 2 ส่วนใหญ่เพิ่มขึ้น

จากการศึกษาเอกสารและงานวิจัยข้างต้น ทั้งในต่างประเทศและในประเทศ จะเห็นว่าได้มีผู้วิจัยหลายท่านให้ความสนใจในการศึกษาเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ หรือการวิเคราะห์ความลำเอียงของข้อสอบ โดยใช้เทคนิคการตรวจสอบหลายวิธี บางวิธีเป็นวิธีที่ใช้ทฤษฎีการวัดผลแบบดั้งเดิม (CTT) เช่น วิธีแปลงค่าความยาก วิธีวิเคราะห์ความแปรปรวน วิธีไค-กำลังสอง เป็นต้น บางวิธีเป็นวิธีที่ใช้ทฤษฎีการตอบข้อสอบ (IRT) เช่น วิธีโค้งลักษณะข้อสอบ วิธีแมนเทิล-แฮนส์เชล วิธีชิปเทสต์ ซึ่งเป็นวิธีที่ใช้สถิติแบบนอนพาราเมตริก จากวิธีดังกล่าว บางวิธีเป็นวิธีที่ให้ผลการวิเคราะห์ได้ถูกต้องแม่นยำ แต่ไม่เหมาะสมที่จะนำมาใช้ในทางปฏิบัติเพราะต้องใช้กลุ่มตัวอย่างขนาดใหญ่ การคำนวณและการแปลผลซับซ้อน ยุ่งยาก ต้องเสียค่าใช้จ่ายสูง ในขณะที่บางวิธีเป็นวิธีที่ใช้กลุ่มตัวอย่างไม่มาก การคำนวณและการแปลผลง่าย ไม่ซับซ้อน แต่ให้ผลการวิเคราะห์ที่มีความถูกต้องน้อยกว่าวิธีอื่น และบางวิธีใช้กลุ่มตัวอย่างไม่มาก การคำนวณและการแปลผลง่าย ให้ผลการวิเคราะห์ที่มีความถูกต้อง เหมาะที่จะนำมาใช้ในทางปฏิบัติได้ เช่น วิธีแมนเทิล-แฮนส์เชล วิธีชิปเทสต์ และวิธีแปลงค่าความยาก สำหรับตัวแปรการทำหน้าที่ต่างกันที่นำมาศึกษามีดังนี้ ภาษา เพศ วัฒนธรรม ศาสนา สภาพภูมิศาสตร์ เป็นต้น ทำให้ผู้วิจัยสนใจศึกษาเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้วิธีชิปเทสต์และวิธีถดถอย-โลจิสติก โดยใช้รูปแบบต่างกัน คือ รวมทั้งฉบับ แยกตามเนื้อหา และแยกตามระดับพฤติกรรม เป็นประโยชน์ต่อการศึกษาเทคนิคการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

บทที่ 3

วิธีดำเนินการศึกษาค้นคว้า

ในการศึกษาค้นคว้าครั้งนี้ ผู้วิจัยได้ดำเนินการตามลำดับดังนี้

1. การกำหนดประชากรและเลือกกลุ่มตัวอย่าง
2. การสร้างเครื่องมือที่ใช้ในการศึกษาค้นคว้า
3. การเก็บรวบรวมข้อมูล
4. การจัดกระทำและการวิเคราะห์ข้อมูล

การกำหนดประชากรและการเลือกกลุ่มตัวอย่าง

ประชากร

ประชากรที่ใช้ในการศึกษาค้นคว้าครั้งนี้ เป็นนักเรียนชั้นมัธยมศึกษาปีที่ 2 ภาคเรียนที่ 2 ปีการศึกษา 2542 ของโรงเรียนมัธยมศึกษา สังกัดกรมสามัญศึกษา กรุงเทพมหานคร ส่วนกลาง กลุ่มที่ 4 จำนวน 14 โรงเรียน จำนวน 187 ห้องเรียน รวมนักเรียนทั้งสิ้น 7,888 คน จำแนกเป็น นักเรียนชาย 3,928 คน นักเรียนหญิง 3,960 คน

กลุ่มตัวอย่าง

กลุ่มตัวอย่างที่ใช้ในการศึกษาค้นคว้าครั้งนี้ เป็นนักเรียนชั้นมัธยมศึกษาปีที่ 2 ภาคเรียนที่ 2 ปีการศึกษา 2542 ของโรงเรียนมัธยมศึกษา สังกัดกรมสามัญศึกษา กรุงเทพมหานคร ส่วนกลาง กลุ่มที่ 4 ซึ่งได้มาโดยการสุ่มแบบแบ่งชั้น (Stratified Random Sampling) มีขนาดโรงเรียนเป็นชั้น (Strata) และมีห้องเรียนเป็นหน่วยการสุ่ม (Sampling Unit) จำนวนโรงเรียนทั้งหมด 6 โรงเรียน จำนวนห้องเรียนทั้งหมด 23 ห้องเรียน รวมนักเรียนทั้งสิ้น 1,051 คน จำแนกเป็นนักเรียนชาย 522 คน และนักเรียนหญิง 529 คน โดยกำหนดขนาดกลุ่มตัวอย่างทั้งหมดด้วยความเชื่อมั่น 95% ($\alpha=0.05$) ซึ่งเทียบจากตารางขนาดกลุ่มตัวอย่างของยามานะ (ลวัน สายยศ และ อังคณา สายยศ. 2539 : 298 อ้างอิงจาก Yamane. 1967) ปรากฏว่าต้องใช้ขนาดกลุ่มตัวอย่างอย่างน้อย 364 คน ในการศึกษาครั้งนี้ มีลำดับขั้นตอนในการสุ่มดังนี้

ขั้นที่ 1 สุ่มโรงเรียนจากแต่ละขนาดของโรงเรียนด้วยวิธีการสุ่มแบบแบ่งชั้น (Stratified Random Sampling) มีขนาดของโรงเรียนเป็นชั้น (Strata) คือ โรงเรียนขนาดใหญ่พิเศษ (พิจารณาจากจำนวนนักเรียนทั้งหมดมากกว่า 2,500 คน) โรงเรียนขนาดใหญ่ (พิจารณาจากจำนวนนักเรียนทั้งหมดมากกว่า 1,500 คน แต่ไม่มากกว่า 2,500 คน) โรงเรียนขนาดกลาง (พิจารณาจากจำนวนนักเรียนทั้งหมดมากกว่า 500 คน แต่ไม่มากกว่า 1,500 คน) และมีโรงเรียนเป็นหน่วยการสุ่ม (Sampling Unit) ได้โรงเรียนขนาดใหญ่พิเศษ 3 โรงเรียน โรงเรียนขนาดใหญ่ 2 โรงเรียน และโรงเรียนขนาดกลาง 1 โรงเรียน

ขั้นที่ 2 สุ่มห้องเรียนจากโรงเรียนที่สุ่มได้ในแต่ละขนาด โดยวิธีสุ่มอย่างง่าย (Simple Random Sampling) มีห้องเรียนเป็นหน่วยการสุ่ม ได้จำนวนห้องเรียน 23 ห้องเรียน จำนวน 1,051 คน จำแนกเป็นนักเรียนชาย 522 คน นักเรียนหญิง 529 คน ดังตาราง 2

ตาราง 2 จำนวนนักเรียนในกลุ่มตัวอย่างที่ใช้ในการวิจัย จำแนกตามขนาดโรงเรียน

ขนาด	โรงเรียน	จำนวนนักเรียน (ห้องเรียน)		
		รวม(ห้อง)	ชาย	หญิง
ใหญ่พิเศษ	โรงเรียนเตรียมอุดมศึกษาพัฒนาการ	265(5)	125	140
	โรงเรียนพระโขนงพิทยาลัย	219(5)	110	109
	โรงเรียนวชิรธรรมสาธิต	201(5)	102	99
ใหญ่	โรงเรียนกุนนทีรุทธารามวิทยาคม	149(3)	76	73
	โรงเรียนมัธยมวัดธาตุทอง	136(3)	63	73
กลาง	โรงเรียนจันทร์หุ่นบำเพ็ญ	81(2)	46	35
รวม		1,051(23)	522	529

การสร้างเครื่องมือที่ใช้ในการศึกษาค้นคว้า

เครื่องมือที่ใช้ในการศึกษาค้นคว้าครั้งนี้ เป็นแบบทดสอบผลสัมฤทธิ์วิชาคณิตศาสตร์ (ค 203) ชั้นมัธยมศึกษาปีที่ 2 ตามหลักสูตรมัธยมศึกษาตอนต้น พุทธศักราช 2521 (ฉบับปรับปรุง พ.ศ. 2533) เป็นแบบเลือกตอบชนิด 5 ตัวเลือก จำนวน 50 ข้อ ที่ผู้วิจัยสร้างขึ้นเอง

วิธีดำเนินการสร้างเครื่องมือ

การสร้างเครื่องมือเพื่อใช้ในการเก็บรวบรวมข้อมูล ผู้วิจัยดำเนินการสร้างตามลำดับขั้นตอน ดังนี้

1. ศึกษาหลักสูตร คู่มือครู หนังสือแบบเรียนวิชาคณิตศาสตร์ (ค 203) ชั้นมัธยมศึกษาปีที่ 2 พุทธศักราช 2521 (ฉบับปรับปรุง พ.ศ. 2533) เพื่อกำหนดขอบเขตของเนื้อหาและพฤติกรรมที่ต้องการวัด

2. สร้างแบบทดสอบผลสัมฤทธิ์วิชาคณิตศาสตร์ แบบเลือกตอบชนิด 5 ตัวเลือก คณิตศาสตร์ (ค 203) ตามขอบเขตเนื้อหาที่กำหนดไว้ในข้อ 1 โดยสร้างตามตารางกำหนดข้อสอบ จำนวน 80 ข้อ

3. นำแบบทดสอบผลสัมฤทธิ์วิชาคณิตศาสตร์ที่สร้างขึ้นจำนวน 80 ข้อ ให้ผู้เชี่ยวชาญด้านการสอนคณิตศาสตร์ 2 คนและนักวัดผล 3 คน ตรวจสอบความเที่ยงตรงเชิงเนื้อหา โดยพิจารณาความสอดคล้องระหว่างข้อสอบแต่ละข้อกับเนื้อหาและระดับพฤติกรรม ข้อสอบที่วัดได้ตรงตามเนื้อหา และระดับพฤติกรรมจะต้องได้ความเห็นตรงกันจากผู้เชี่ยวชาญ 3 ใน 5 คน จึงถือว่า ข้อสอบข้อนั้นใช้ได้ ในการสร้างเครื่องมือที่ใช้ในการวิจัยครั้งนี้ หลังจากให้ผู้เชี่ยวชาญพิจารณาแล้วผู้วิจัยได้ปรับปรุงแก้ไขข้อคำถาม ข้อที่บกพร่องตามความคิดเห็นที่ผู้เชี่ยวชาญเสนอแนะ

4. นำแบบทดสอบที่ได้จากข้อ 3 ไปทดสอบครั้งที่ 1 กับกลุ่มตัวอย่างจำนวน 153 คน โดยสุ่มจากประชากรที่เหลือจากกลุ่มตัวอย่างที่ศึกษา โดยวิธีสุ่มอย่างง่าย (Simple Random Sampling) ใช้ห้องเรียนเป็นหน่วยการสุ่ม (Sampling Unit) เพื่อหาค่าความยากและค่าอำนาจจำแนก โดยใช้โปรแกรม CTIA ซึ่งพัฒนาโดย อาจารย์สุพัฒน์ สุขมลสันต์ (2538) ได้ค่าความยากระหว่าง .08 ถึง .80 และค่าอำนาจจำแนกระหว่าง -.20 ถึง .83 แล้วคัดเลือกข้อที่มีค่าความยากตั้งแต่ .20 ถึง .80 และค่าอำนาจจำแนกตั้งแต่ .20 ขึ้นไป ไว้จำนวน 50 ข้อ ปรากฏว่าได้ข้อสอบที่มีค่าความยากระหว่าง .20 ถึง .79 และค่าอำนาจจำแนกระหว่าง .20 ถึง .83 (รายละเอียดดังตารางใน ภาคผนวก ข)

5. นำแบบทดสอบที่คัดเลือกข้อสอบที่มีคุณภาพแล้วจำนวน 50 ข้อ ไปทดสอบครั้งที่ 2 กับนักเรียนจำนวน 200 คน โดยสุ่มจากประชากรที่เหลือจากกลุ่มตัวอย่างที่ศึกษา โดยวิธีสุ่มอย่างง่าย (Simple Random Sampling) ใช้ห้องเรียนเป็นหน่วยการสุ่ม (Sampling Unit) คำนวณค่าความเชื่อมั่นของแบบทดสอบด้วยสูตรสัมประสิทธิ์ r_B ได้เท่ากับ .82 ซึ่งมีค่าสูงเพียงพอที่นำมาใช้ในการวิจัยครั้งนี้

6. จัดพิมพ์แบบทดสอบเป็นฉบับสมบูรณ์ โดยจัดเรียงข้อสอบตามเนื้อหาวิชาคณิตศาสตร์ ได้แก่ ระบบจำนวนเต็ม ความเท่ากันทุกประการ เศษส่วนและทศนิยม และสมบัติของรูปสามเหลี่ยม มุมจาก ตามลำดับ ในแต่ละเนื้อหานั้นจัดเรียงข้อสอบตามระดับพฤติกรรมด้านสติปัญญา (Cognitive Domain) ของวิลสัน (Wilson. 1971) ได้แก่ ความรู้หรือทักษะการคิดคำนวณ ความเข้าใจ การนำไปใช้ และการวิเคราะห์ ตามลำดับ เพื่อนำไปใช้ในการวิจัยต่อไป

ลักษณะของแบบทดสอบ

แบบทดสอบที่ใช้ในการวิจัยครั้งนี้ เป็นแบบทดสอบเลือกตอบชนิด 5 ตัวเลือก ที่มีคำตอบถูกเพียงคำตอบเดียว เนื้อหาวิชาคณิตศาสตร์ ค 203 ชั้นมัธยมศึกษาปีที่ 2 ตามหลักสูตรมัธยมศึกษาตอนต้น พุทธศักราช 2521 (ฉบับปรับปรุง พ.ศ. 2533) ของกระทรวงศึกษาธิการ ที่จัดทำโดยสถาบันส่งเสริมการสอนวิทยาศาสตร์และเทคโนโลยี (สสวท.) จำแนกตามเนื้อหาและระดับพฤติกรรมด้านสติปัญญา (Cognitive Domain) ของวิลสัน (Wilson. 1971) ดังตาราง 3

ตาราง 3 โครงสร้างของแบบทดสอบผลสัมฤทธิ์วิชาคณิตศาสตร์ ค 203 จำแนกตามเนื้อหาและระดับพฤติกรรม

เนื้อหา \ ระดับพฤติกรรม	ความรู้หรือทักษะการคิดคำนวณ	ความเข้าใจ	การนำไปใช้	การวิเคราะห์	รวม
ระบบจำนวนเต็ม	ข้อ 1	ข้อ 2, 3, 4, 5, 6	ข้อ 7, 8, 9, 10, 11, 12	ข้อ 13, 14, 15	15
ความเท่ากันทุกประการ	ข้อ 16	ข้อ 17, 18	ข้อ 19, 20, 21	ข้อ 22	7
เศษส่วนและทศนิยม	ข้อ 23, 24, 25, 26	ข้อ 27, 28, 29, 30	ข้อ 31, 32, 33, 34, 35, 36	ข้อ 37	15
สมบัติของรูปสามเหลี่ยมมุมฉาก	ข้อ 38, 39	ข้อ 40, 41, 42	ข้อ 43, 44, 45, 46, 47, 48	ข้อ 49, 50	13
รวม	8	14	21	7	50

ตัวอย่างแบบทดสอบ
แบบทดสอบผลสัมฤทธิ์วิชาคณิตศาสตร์

คำชี้แจง ในข้อสอบแต่ละข้อให้เลือกคำตอบจาก ก ถึง จ ที่ถูกต้องที่สุดเพียงคำตอบเดียว
แล้วทำเครื่องหมายกากบาท (X) ในกระดาษคำตอบของแต่ละข้อ

เรื่อง เศษส่วนและทศนิยม

ข้อสอบวัดความรู้หรือทักษะการคิดคำนวณ

(0) -4.05 เปลี่ยนให้อยู่ในรูปเศษส่วนได้ตามข้อใด

ก. $-4\frac{5}{10}$

ข. $-4\frac{5}{100}$

ค. $-\frac{405}{10}$

ง. $-\frac{400}{10}$

จ. $-\frac{405}{1000}$

ตอบ ข้อ ข

เรื่อง เศษส่วนและทศนิยม

ข้อสอบวัดความเข้าใจ

(00) ผลบวกของ $\frac{a}{b}$ กับ $\frac{c}{d}$ เท่ากับเท่าไร

ก. $\frac{a+c}{b+d}$

ข. $\frac{a+c}{bd}$

ค. $\frac{ad+bc}{bd}$

ง. $\frac{ad+bc}{b+d}$

จ. $\frac{ac+bc}{bc+ad}$

ตอบ ข้อ ค

เรื่อง เศษส่วนและทศนิยม

ข้อสอบวัดการนำไปใช้

(000) ผลลัพธ์ของ $\frac{3}{4}$ ของ $\left(\frac{4}{9} + \frac{1}{3}\right) \div 3\frac{1}{2}$

เป็นเท่าไร

ก. $\frac{1}{12}$

ข. $\frac{1}{6}$

ค. $\frac{1}{4}$

ง. $\frac{1}{3}$

จ. $\frac{1}{2}$

ตอบ ข้อ ข

เรื่อง เศษส่วนและทศนิยม

ข้อสอบวัดการวิเคราะห์

(0000) ค่าของ $1 - \frac{1}{1 - \frac{1}{1 - \frac{1}{2}}}$ เท่ากับเท่าไร

ก. $\frac{1}{2}$

ข. $\frac{2}{3}$

ค. $\frac{3}{2}$

ง. 1

จ. 2

ตอบข้อ จ

การเก็บรวบรวมข้อมูล

การเก็บรวบรวมข้อมูลในครั้งนี้ ผู้วิจัยได้ดำเนินการตามลำดับขั้นตอนดังนี้

1. ติดต่อประสานงานกับโรงเรียนที่เป็นกลุ่มตัวอย่าง เพื่อกำหนดวัน เวลา ในการนำแบบทดสอบไปทดสอบกับนักเรียนที่เป็นกลุ่มตัวอย่าง
2. จัดเตรียมแบบทดสอบให้เพียงพอกับจำนวนนักเรียนที่จะทำการสอบในแต่ละครั้ง
3. ชี้แจงทำความเข้าใจกับผู้ประสานงานซึ่งอยู่ที่โรงเรียน ในเรื่องลักษณะของแบบทดสอบ และวิธีดำเนินการสอบ
4. นำแบบทดสอบที่มีคุณภาพจำนวน 50 ข้อ ไปทดสอบกับกลุ่มตัวอย่างที่มีจำนวน 1,051 คน เพื่อศึกษาผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้การตรวจสอบด้วยวิธีชิปทดสอบ และวิธีถดถอยโลจิสติก ในแต่ละรูปแบบที่ศึกษา และทำการทดสอบสมมุติฐานการศึกษาค้นคว้า ข้อ 1 และ ข้อ 2
5. นำผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามข้อ 4 คัดข้อที่ทำหน้าที่ต่างกันออก แล้วนำมาคำนวณหาค่าความเชื่อมั่นของแบบทดสอบ แล้วปรับขยายค่าความเชื่อมั่นของแบบทดสอบให้มีจำนวนข้อเท่ากัน เพื่อเปรียบเทียบค่าความเชื่อมั่นของแบบทดสอบหลังคัดข้อที่ทำหน้าที่ต่างกันออก และทำการทดสอบสมมุติฐานการศึกษาค้นคว้าข้อ 3 และ ข้อ 4

ขั้นตอนในการวิเคราะห์ข้อมูล

ผู้วิจัยนำคะแนนที่ได้จากการทดสอบมาวิเคราะห์ข้อมูล ตามลำดับขั้นตอนดังนี้

1. การวิเคราะห์ข้อมูลในการวิจัย ได้ดำเนินการดังนี้

1.1 ตรวจสอบให้คะแนนแบบทดสอบ พบแบบทดสอบที่ไม่สมบูรณ์จึงคัดออก แล้วตรวจสอบให้คะแนนเฉพาะแบบทดสอบที่สมบูรณ์ ได้กลุ่มตัวอย่างทั้งหมด 994 คน จำแนกเป็นนักเรียนชาย 480 คน นักเรียนหญิง 514 คน

1.2 หาค่าสถิติพื้นฐาน ได้แก่ ค่าเฉลี่ย ($\bar{X}$) และค่าความเบี่ยงเบนมาตรฐาน (S) ค่าความเบ้ (Sk) และความโด่ง (Ku)

1.3 หาค่าความเชื่อมั่นของแบบทดสอบด้วยสูตรสัมประสิทธิ์ r_B (บุญเชิด ภิญโญ-อนันตพงษ์. 2542 : 18-19) โดยใช้โปรแกรม SPSS FOR WINDOW

2. การทดสอบสมมติฐาน ได้ดำเนินการดังนี้

2.1 ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยใช้การตรวจสอบด้วยวิธีชิปเทสต์ (SIB) และวิธีถดถอยโลจิสติก (LRDIF) ในสมมติฐานข้อ 1 ที่ตรวจสอบรวมทั้งฉบับ แยกตามเนื้อหา และแยกตามระดับพฤติกรรม โดยใช้โปรแกรม SIBTEST และโปรแกรม SPSS FOR WINDOW ศึกษาวิธีการใช้จาก *SIBTEST USER MANUAL* (Stout, William F. & Roussos, Louis A. 1992), การวิเคราะห์ความลำเอียงด้วยโปรแกรมสำเร็จรูป *SIBTEST* (ฉัตรศิริ ปิยะพิมลสิทธิ์. 2542) และหลักการและวิธีใช้คอมพิวเตอร์ในงานสถิติเพื่อการวิจัย *SPSS FOR WINDOWS* (วัชรชัย งามสันติวงศ์. 2542)

2.2 คัดเลือกข้อสอบที่ทำหน้าที่ต่างกันออกในแต่ละวิธีการตรวจสอบและในแต่ละรูปแบบ

2.3 ทดสอบความแตกต่างของจำนวนข้อสอบที่ได้จากการตรวจสอบการทำหน้าที่ต่างกัน ในแต่ละวิธีการตรวจสอบและแต่ละรูปแบบในสมมติฐานข้อ 2 ด้วยสูตร Q-Test โดยใช้โปรแกรม Microsoft Excel

2.4 ปรับขยายค่าความเชื่อมั่นของแบบทดสอบหลังจากคัดข้อสอบที่ทำหน้าที่ต่างกันออก ในแต่ละวิธีการตรวจสอบและแต่ละรูปแบบ ด้วยสูตรสเปียร์แมนบราวน์ (Spearman-Brown) โดยใช้โปรแกรม Microsoft Excel

2.5 ทดสอบความแตกต่างของค่าความเชื่อมั่นของแบบทดสอบที่ปรับขยายให้จำนวนข้อสอบเท่ากัน ในแต่ละวิธีการตรวจสอบและแต่ละรูปแบบในสมมติฐานข้อ 3 ด้วยสูตร UX_1 โดยใช้โปรแกรม Microsoft Excel

สถิติที่ใช้ในการวิเคราะห์ข้อมูล

ผู้วิจัยนำคะแนนที่ได้จากการทดสอบมาวิเคราะห์หาค่าทางสถิติดังนี้

1. สถิติพื้นฐาน ได้แก่ ค่าเฉลี่ย ($\bar{X}$) และค่าความเบี่ยงเบนมาตรฐาน (S)
2. สถิติที่ใช้ในการหาคุณภาพเครื่องมือ

2.1 ค่าความยากและค่าอำนาจจำแนกของแบบทดสอบวิเคราะห์ด้วยโปรแกรม CTIA ซึ่งพัฒนาโดยอาจารย์สุวัฒน์ สุกมลสันต์ (2538)

2.2 ค่าความเชื่อมั่นของแบบทดสอบคำนวณด้วยสูตรสัมประสิทธิ์ r_B (บุญเชิด ภิญญอนันตพงษ์, 2542 : 18-19)

$$r_B = \frac{1}{1 - \sum \lambda_i^2} \left[1 - \frac{\sum P_i(1 - P_i)}{S_x^2} \right]$$

เมื่อ P_i แทน ค่าความยากง่ายของข้อสอบแต่ละข้อ

S_x^2 แทน ความแปรปรวนของคะแนนทั้งฉบับ

$$\lambda_1 = [\sum B_{1i} - P_1(\bar{X} - 1)] / S_x^2$$

$$\lambda_2 = [\sum B_{2i} - P_2(\bar{X} - 1)] / S_x^2$$

.....

$$\lambda_K = [\sum B_{Ki} - P_K(\bar{X} - 1)] / S_x^2$$

$$\bar{X} = \sum P_i$$

B_{ji} แทน สัดส่วนจำนวนผู้ตอบข้อสอบสองข้อใดๆ ถูกทั้งคู่

3. สถิติที่ใช้ในการทดสอบสมมติฐาน

3.1 สถิติที่ใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้วิธีชิปเทสต์ (SIBTEST) ด้วยสูตร (Li, H., Nandakumar, R. & Stout, W. 1995 : 5)

$$\hat{\beta}_u = \sum_{k=0}^n \hat{p}_k (\bar{Y}_{Rk} - \bar{Y}_{Fk})$$

เมื่อ $\hat{p}_k$ แทน สัดส่วนของผู้สอบกลุ่มเปรียบเทียบที่ตอบแบบทดสอบย่อยที่ไม่พบการทำหน้าที่ต่างกันของข้อสอบได้ถูก k ข้อ

$\bar{Y}_{Rk}$ แทน คะแนนเฉลี่ยของผู้สอบทุกคนในกลุ่มอ้างอิงที่ตอบแบบทดสอบย่อยที่ศึกษาได้ถูก k ข้อ

$\bar{Y}_{Fk}$ แทน คะแนนเฉลี่ยของผู้สอบทุกคนในกลุ่มเปรียบเทียบที่ตอบแบบทดสอบย่อยที่ศึกษาได้ถูก k ข้อ

3.1.1 สถิติที่ใช้ในการทดสอบค่า SIBTEST เมื่อค่า β_U หรือค่า Z_{SIB} ที่คำนวณจากสูตรแตกต่างจากค่า Z_α ที่ $N(0,1)$ อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 (Li, H., Nandakumar, R. & Stout, W. 1995 : 6)

$$\beta_U = \hat{\beta}_U / \hat{\sigma}(\hat{\beta}_U)$$

$$\hat{\sigma}(\hat{\beta}_U) = \left\{ \sum_{k=0}^n \hat{p}_k^2 \left[\frac{1}{J_{Rk}} \hat{\sigma}^2(Y/k, R) + \frac{1}{J_{Fk}} \hat{\sigma}^2(Y/k, F) \right] \right\}^{1/2}$$

เมื่อ $\hat{\sigma}(\hat{\beta}_U)$ แทน ค่าประมาณค่าความคลาดเคลื่อนมาตรฐานของ β_U

J_{Rk} แทน จำนวนผู้สอบในกลุ่มอ้างอิงในช่วงคะแนน k

J_{Fk} แทน จำนวนผู้สอบในกลุ่มเปรียบเทียบในช่วงคะแนน k

โดยที่กลุ่มอ้างอิง (Reference Group) หมายถึง กลุ่มผู้สอบที่คาดว่าจะจะเป็นกลุ่มที่ได้เปรียบจากการทำหน้าที่ต่างกันของข้อสอบ กล่าวคือ เป็นกลุ่มผู้สอบที่มีโอกาสตอบข้อสอบทำหน้าที่ต่างกันได้ถูกต้องสูงกว่ากลุ่มเปรียบเทียบ ในการวิจัยครั้งนี้ใช้กลุ่มเพศชาย เป็นกลุ่มอ้างอิง

กลุ่มเปรียบเทียบ (Focal Group) หมายถึง กลุ่มผู้สอบที่คาดว่าจะจะเป็นกลุ่มที่เสียเปรียบจากการทำหน้าที่ต่างกันของข้อสอบ กล่าวคือ เป็นกลุ่มผู้สอบที่มีโอกาสตอบข้อสอบทำหน้าที่ต่างกันได้ถูกต้องต่ำกว่ากลุ่มเปรียบเทียบ ในการวิจัยครั้งนี้ใช้กลุ่มเพศหญิง เป็นกลุ่มเปรียบเทียบ

3.2 สถิติที่ใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยใช้วิธีการถดถอยโลจิสติก (LRDIF) โดยใช้สูตร (Swaminathan & Rogers. 1990 : 361-370)

$$P(u=1) = \frac{e^Z}{1+e^Z}$$

$$\text{โดยที่ } Z = \tau_0 + \tau_1\theta + \tau_2g + \tau_3(\theta g)$$

โดยกำหนดให้ $g = 1$ ถ้าผู้สอบอยู่ในกลุ่ม 1 และ $g = 0$ ถ้าผู้สอบอยู่ในกลุ่ม 2

เมื่อ θ แทน ระดับความสามารถของผู้สอบ (คะแนนรวม)

τ_0 แทน พารามิเตอร์จุดตัด

τ_1 แทน สัมประสิทธิ์ของความสามารถ

τ_2 แทน ความแตกต่างในกลุ่มของผลการตอบข้อสอบ

τ_3 แทน ปฏิสัมพันธ์ระหว่างกลุ่มและระดับความสามารถ

การทดสอบสมมติฐานของการทำหน้าที่ต่างกันของข้อสอบ จะทดสอบสมาชิกของ τ_3 ซึ่งสมมติฐานที่สนใจคือ

$$H_0 : C_\tau = 0$$

$$H_1 : C_\tau \neq 0$$

$$\text{โดยที่ } C = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

3.2.1 สถิติที่ใช้ในการทดสอบนัยสำคัญใช้สถิติไค-กำลังสอง ทดสอบสมมติฐานการทำหน้าที่ต่างกัน ถ้า χ^2 มีค่ามากกว่า $\chi^2_{\alpha;2}$ แสดงว่า ปฏิเสธสมมติฐานของข้อสอบที่ทำหน้าที่ไม่ต่างกัน (No DIF) นั่นคือ ข้อสอบทำหน้าที่ต่างกัน (DIF)

$$\chi^2 = \hat{\tau}'C'(C\Sigma C')^{-1}C\hat{\tau}$$

3.3 สถิติทดสอบนัยสำคัญของความแตกต่างจำนวนข้อสอบ ที่ได้จากการตรวจสอบการทำหน้าที่ต่างกันด้วยวิธีต่างกัน โดยใช้สูตร Q-Test (บุญเชิด ภิญโญนนตพงษ์. 2527 : 100 ; อ้างอิงจาก Hambleton, et al. 1978 : 37)

$$Q = \frac{(n-1)(n \sum p^2 - T^2)}{nT - \sum X^2}$$

เมื่อ Q แทน การทดสอบแบบคิว

n แทน จำนวนวิธีวิเคราะห์ข้อสอบ

p แทน จำนวนข้อสอบที่ถูกคัดเลือกแต่ละวิธีวิเคราะห์ข้อสอบ

X แทน คะแนนของข้อสอบที่ถูกคัดเลือกในแต่ละข้อ

T แทน ผลรวมคะแนนของข้อสอบที่ถูกคัดเลือกทุกข้อ

3.4 ค่าความเชื่อมั่นของแบบทดสอบที่ปรับขยาย ด้วยสูตรของสเปียร์แมน-บราวน์ (Spearman-Brown) (บุญเชิด ภิญโญนนตพงษ์. ม.ป.ป. : 252 ; Stanley & Hopkins. 1972 : 125)

$$r_{tt} = \frac{nr_{xx}}{1 + (n-1)r_{xx}}$$

เมื่อ r_{tt} แทน ความเชื่อมั่นของแบบทดสอบที่ปรับขยาย

r_{xx} แทน ความเชื่อมั่นเดิม

n แทน จำนวนเท่าที่ขยายออก

3.5 สถิติทดสอบนัยสำคัญของความแตกต่างของค่าความเชื่อมั่นหลายค่า ซึ่งประมาณจากกลุ่มตัวอย่างกลุ่มเดียว ใช้สูตร UX_1 (บุญเชิด ภิญโญนนตพงษ์. 2537 : 57 ; อ้างอิงจาก Woodruff & Feldt. 1986 : 393-413)

$$UX_1 = \sum_{i=1}^m \frac{(u_i - \bar{u})^2}{S_u^2 - C_u}$$

$$\bar{u} = \sum_{i=1}^m \frac{u_i}{m}$$

$$u_i = \frac{1}{(1-r_i)^{1/3}}$$

$$S_u^2 = \frac{2}{9m(N_c - 1)} \sum_{i=1}^m u_i^2$$

$$C_u = \frac{4}{9m(m-1)(N_c-1)} \sum_{i=2}^m \sum_{j=1}^{i-1} r_{ij}^2 u_i u_j$$

$$N_c = N \left(\frac{\bar{n}_h - 1}{\bar{n}_h + 1} \right)$$

$$\bar{n}_h = \frac{m}{\sum_{i=1}^m \left(\frac{1}{n_i} \right)}$$

เมื่อ UX_i	แทน	สถิติทดสอบที่มีการแจกแจงแบบ χ^2 ; $df = m - 1$
r	แทน	ค่าความเชื่อมั่นที่คำนวณได้
m	แทน	จำนวนค่าสัมประสิทธิ์ความเชื่อมั่น
r_{ij}	แทน	ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปร
S_u^2	แทน	ค่าความแปรปรวนของ u_i
C_u	แทน	ค่าความแปรปรวนร่วมของ u_i
$\bar{n}_h$	แทน	ค่าเฉลี่ยฮาร์โมนิคของความยาวของแบบทดสอบ

บทที่ 4

ผลการวิเคราะห์ข้อมูล

สัญลักษณ์และอักษรย่อที่ใช้ในการวิเคราะห์ข้อมูล

การกำหนดสัญลักษณ์และอักษรย้อมีดังนี้

$\bar{X}$	แทน	ค่าเฉลี่ยของคะแนนผลการสอบ
S	แทน	ค่าความเบี่ยงเบนมาตรฐานของคะแนนผลการสอบ
CV	แทน	สัมประสิทธิ์การกระจายของคะแนนผลการสอบ
N	แทน	จำนวนนักเรียนในกลุ่มตัวอย่าง
K	แทน	จำนวนข้อของแบบทดสอบ
p	แทน	ค่าความยากง่ายของข้อสอบ
r	แทน	ค่าอำนาจจำแนกของข้อสอบ
Sk	แทน	ค่าความเบ้ของคะแนนผลการสอบ
Ku	แทน	ค่าความโด่งของคะแนนผลการสอบ
r_B	แทน	ค่าความเชื่อมั่นของแบบทดสอบตามแบบจำลอง คะแนนจริงสัมพัทธ์
DiF items	แทน	ข้อสอบที่ทำหน้าที่ต่างกัน
No DiF items	แทน	ข้อสอบที่ทำหน้าที่ไม่ต่างกัน
β_U	แทน	ค่าทดสอบการทำหน้าที่ต่างกันของข้อสอบอย่างมีนัยสำคัญ ของข้อสอบที่ได้จากการตรวจสอบด้วยวิธีชิปเทสต์
B	แทน	ค่าทดสอบการทำหน้าที่ต่างกันของข้อสอบอย่างมีนัยสำคัญ ของข้อสอบที่ได้จากการตรวจสอบด้วยวิธีถดถอยโลจิสติก
SIBTEST	แทน	การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีชิปเทสต์
LRDIF	แทน	การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วย วิธีถดถอยโลจิสติก
Q	แทน	ค่าสถิติทดสอบที่มีการแจกแจงแบบไค-กำลังสอง
r_{xx}	แทน	ค่าความเชื่อมั่นของแบบทดสอบหลังจากคัดข้อสอบที่ทำหน้าที่ ต่างกันออก
r_{tt}	แทน	ค่าความเชื่อมั่นของแบบทดสอบที่ปรับขยายด้วยสูตรของ สเปียร์แมน-บราวน์
UX_1	แทน	ค่าสถิติซึ่งมีการแจกแจงในรูปของ χ^2
df	แทน	ลำดับชั้นความเป็นอิสระ

การเสนอผลการวิเคราะห์ข้อมูล

การเสนอผลการวิเคราะห์ข้อมูลครั้งนี้ ผู้วิจัยได้นำเสนอตามลำดับดังนี้

1. ค่าสถิติพื้นฐานของแบบทดสอบผลสัมฤทธิ์วิชาคณิตศาสตร์ จำแนกตามเนื้อหาและระดับพฤติกรรม
2. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจากแบบทดสอบผลสัมฤทธิ์วิชาคณิตศาสตร์โดยใช้รูปแบบและวิธีการต่างกัน
3. การเปรียบเทียบจำนวนข้อสอบที่ทำหน้าที่ต่างกันโดยใช้รูปแบบและวิธีการต่างกัน
4. การเปรียบเทียบค่าความเชื่อมั่นของแบบทดสอบผลสัมฤทธิ์วิชาคณิตศาสตร์ หลังจากคัดข้อสอบที่ทำหน้าที่ต่างกันอย่างออก โดยใช้รูปแบบและวิธีการต่างกัน

ผลการวิเคราะห์ข้อมูล

1. ค่าสถิติพื้นฐานของแบบทดสอบผลสัมฤทธิ์วิชาคณิตศาสตร์ จำแนกตามเนื้อหาและระดับพฤติกรรม

การวิเคราะห์ครั้งนี้ ได้จากการนำคะแนนผลการสอบของแบบทดสอบ ที่ผู้วิจัยสร้างขึ้นจำนวน 50 ข้อ มาคำนวณหาค่าสถิติพื้นฐานของแบบทดสอบรวมทั้งฉบับ จำแนกตามเนื้อหา และจำแนกตามระดับพฤติกรรม ได้แก่ ค่าเฉลี่ย ($\bar{X}$) ค่าความเบี่ยงเบนมาตรฐาน (S) ค่าสัมประสิทธิ์ การกระจาย (CV) ค่าความเบ้ (Sk) และความโด่ง (Ku) ดังแสดงในตาราง 4

ตาราง 4 ค่าสถิติพื้นฐานของแบบทดสอบผลสัมฤทธิ์วิชาคณิตศาสตร์ จำแนกตามเนื้อหาและระดับพฤติกรรม

ค่าสถิติ	K	$\bar{X}$	S	CV	Sk	Ku
รวมทั้งฉบับ	50	20.39	7.21	35.36	.271	.788
แยกตามเนื้อหา						
ระบบจำนวนเต็ม	15	6.75	3.36	49.78	.090	-.083
ความเท่ากันทุกประการ	7	3.30	1.49	45.15	.233	-.100
เศษส่วนและทศนิยม	15	6.48	2.81	43.36	.105	-.180
สมบัติของรูปสามเหลี่ยมมุมฉาก	13	5.86	1.82	31.06	.808	1.249
แยกตามระดับพฤติกรรม						
ความรู้หรือทักษะการคิดคำนวณ	8	4.23	1.38	32.63	-.026	-.495
ความเข้าใจ	14	6.86	2.60	37.91	.341	.555
การนำไปใช้	21	10.81	3.59	33.21	-.033	-.160
การวิเคราะห์	7	3.49	1.36	38.97	.586	-.386

ผลการวิเคราะห์ตามตาราง 4 ปรากฏว่า แบบทดสอบทั้งฉบับมีค่าเฉลี่ย ($\bar{X}$) เท่ากับ 20.39 คะแนน ค่าความเบี่ยงเบนมาตรฐาน (S) เท่ากับ 7.21 เมื่อพิจารณาค่าความเบ้ (Sk) โดยเทียบค่าความเบ้ (Sk) กับโค้งปกติ (Normal Curve) ซึ่งมีค่าความเบ้เท่ากับ 0 (นิภา ศรีไพโรจน์. 2524 : 148-149) พบว่า มีค่าเท่ากับ .271 ซึ่งมีลักษณะเบ้ทางบวกเล็กน้อย (Positive Skewness) แสดงว่า คะแนนผลการสอบส่วนใหญ่มีค่าต่ำกว่าค่าเฉลี่ยเล็กน้อย เมื่อพิจารณาความโด่ง (Ku) โดยเทียบกับโค้งปกติ (Normal Curve) ที่มีค่าเท่ากับ 3 พบว่า มีค่าเท่ากับ .788 ซึ่งมีค่าน้อยกว่า 3 แสดงว่า มีการแจกแจงความถี่ในลักษณะโค้งลาด (Platykurtic)

เมื่อพิจารณาจำแนกตามเนื้อหา ซึ่งมีจำนวนข้อระหว่าง 7 ถึง 15 ข้อ ค่าเฉลี่ย ($\bar{X}$) มีค่าระหว่าง 3.30 ถึง 6.75 ค่าความเบี่ยงเบนมาตรฐาน (S) มีค่าระหว่าง 1.49 ถึง 3.36 เมื่อเปรียบเทียบค่าสัมประสิทธิ์การกระจายของคะแนนผลการสอบ มีค่าระหว่าง 31.06 ถึง 49.78 แสดงว่า คะแนนผลการสอบที่แยกตามเนื้อหา เรื่อง ระบบจำนวนเต็ม มีการกระจายของคะแนนมากที่สุด รองลงมาคือ ความเท่ากันทุกประการ เศษส่วนและทศนิยม และ สมบัติของรูปสามเหลี่ยมมุมฉาก ตามลำดับ เมื่อพิจารณาค่าความเบ้ (Sk) พบว่า มีค่าระหว่าง .090 ถึง .800 ซึ่งมีลักษณะเบ้ทางบวกเล็กน้อย (Positive Skewness) แสดงว่า คะแนนผลการสอบส่วนใหญ่มีค่าต่ำกว่าค่าเฉลี่ยเล็กน้อย เมื่อพิจารณาความโด่ง (Ku) พบว่า มีค่าระหว่าง -.180 ถึง 1.249 แสดงว่า มีการแจกแจงความถี่ในลักษณะโค้งลาด (Platykurtic)

เมื่อพิจารณาจำแนกตามระดับพฤติกรรม ซึ่งมีจำนวนข้อระหว่าง 7 ถึง 21 ข้อ ค่าเฉลี่ย ($\bar{X}$) มีค่าระหว่าง 3.49 ถึง 10.81 ค่าความเบี่ยงเบนมาตรฐาน (S) มีค่าระหว่าง 1.36 ถึง 3.59 เมื่อเปรียบเทียบค่าสัมประสิทธิ์การกระจายของคะแนนผลการสอบมีค่าระหว่าง 32.63 ถึง 38.97 แสดงว่า คะแนนผลการสอบที่แยกตามระดับพฤติกรรม วัดการวิเคราะห์ มีการกระจายของคะแนนมากที่สุด รองลงมาคือ วัดความเข้าใจ วัดการนำไปใช้ และวัดความรู้หรือทักษะการคิดคำนวณ ตามลำดับ เมื่อพิจารณาค่าความเบ้ (Sk) พบว่า มีค่าระหว่าง -.033 ถึง .586 ซึ่งมีลักษณะเบ้ทางบวกเล็กน้อย (Positive Skewness) ในระดับพฤติกรรมที่วัดความเข้าใจและการวิเคราะห์ แสดงว่า คะแนนผลการสอบส่วนใหญ่มีค่าต่ำกว่าค่าเฉลี่ยเล็กน้อย และมีลักษณะเบ้ทางลบเล็กน้อย (Negative Skewness) ในระดับพฤติกรรมที่วัดความรู้หรือทักษะการคิดคำนวณ และการนำไปใช้ แสดงว่า คะแนนผลการสอบส่วนใหญ่สูงกว่าค่าเฉลี่ยเล็กน้อย เมื่อพิจารณาความโด่ง (Ku) พบว่า มีค่าระหว่าง -.495 ถึง .555 แสดงว่า มีการแจกแจงความถี่ในลักษณะโค้งลาด (Platykurtic)

2. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจากแบบทดสอบผลสัมฤทธิ์
วิชาคณิตศาสตร์โดยใช้รูปแบบและวิธีการต่างกัน

การวิเคราะห์ในส่วนนี้ ผู้วิจัยนำคะแนนผลการสอบจากแบบทดสอบจำนวน 50 ข้อ ที่ตรวจให้คะแนนในแบบ (0, 1) มาตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยพิจารณาคะแนนรวมทั้งฉบับ แยกตามเนื้อหา และแยกตามระดับพฤติกรรม ด้วยวิธีซิปเทสต์โดยใช้โปรแกรม SIBTEST และวิธีถดถอยโลจิสติกโดยใช้โปรแกรม SPSS FOR WINDOW ดังแสดงในตาราง 5 (รายละเอียดดังตารางในภาคผนวก ค)

ตาราง 5 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจากแบบทดสอบผลสัมฤทธิ์วิชาคณิตศาสตร์
โดยใช้รูปแบบและวิธีการต่างกัน

รูปแบบ	วิธีการ ตรวจสอบ	ข้อที่	จำนวน	ร้อยละ
รวมทั้งฉบับ				
	SIBTEST	<u>1</u> , <u>18</u> , <u>21</u> , <u>26</u> , <u>27</u> , 28, <u>31</u> , <u>35</u> , <u>46</u>	9	18.00
รวม	LRDIF	<u>1</u> , 2, 3, 4, 8, <u>18</u> , <u>21</u> , <u>26</u> , <u>27</u> , 29, <u>31</u> , 32, 34, <u>35</u> , 36, 37, <u>46</u>	17	34.00
แยกตามเนื้อหา				
ระบบจำนวนเต็ม	SIBTEST	-	-	-
	LRDIF	4, 6	2	13.33
ความเท่ากันทุกประการ	SIBTEST	18	1	14.28
	LRDIF	20	1	14.28
เศษส่วนและทศนิยม	SIBTEST	27, 28, 29, 31, 35	5	33.33
	LRDIF	26, 27, 29, 31, 32, 34, 35, 36	8	53.33
สมบัติของ	SIBTEST	45, 46	2	15.38
รูปสามเหลี่ยมมุมฉาก	LRDIF	45, 46	2	15.38
รวม	SIBTEST	<u>18</u> , <u>27</u> , 28, <u>29</u> , <u>31</u> , <u>35</u> , <u>45</u> , <u>46</u>	8	16.00
	LRDIF	4, 6, 20, 26, <u>27</u> , <u>29</u> , <u>31</u> , 32, 34, <u>35</u> , 36, <u>45</u> , <u>46</u>	13	26.00
แยกตามระดับพฤติกรรม				
ความรู้หรือทักษะ	SIBTEST	24	1	12.50
การคิดคำนวณ	LRDIF	24, 25, 26	3	37.50
ความเข้าใจ	SIBTEST	3, 18, 27, 42	4	28.57
	LRDIF	2, 3, 17, 18, 27, 29, 42	7	50.00
การนำไปใช้	SIBTEST	21, 31, 35, 45	4	19.05
	LRDIF	8, 31, 35, 36, 46	5	23.81
การวิเคราะห์	SIBTEST	14	1	14.28
	LRDIF	13, 22	2	28.57
รวม	SIBTEST	<u>3</u> , 14, <u>18</u> , 21, <u>24</u> , <u>27</u> , <u>31</u> , <u>35</u> , <u>42</u> , 45,	10	20.00
	LRDIF	2, <u>3</u> , 8, 13, 17, <u>18</u> , 22, <u>24</u> , 25, 26, <u>27</u> , 29, <u>31</u> , <u>35</u> , 36, <u>42</u> , 46,	17	34.00

ผลการวิเคราะห์ตามตาราง 5 ให้ผลดังนี้

เมื่อพิจารณารูปแบบรวมทั้งฉบับ พบว่า ตรวจสอบด้วยวิธีชิปเทสต์มีจำนวนข้อสอบที่ทำหน้าที่ต่างกันรวม 9 ข้อ ได้แก่ ข้อ 1, 18, 21, 26, 27, 28, 31, 35 และ 46 คิดเป็นร้อยละ 18 และตรวจสอบด้วยวิธีถดถอยโลจิสติกมีจำนวนข้อสอบที่ทำหน้าที่ต่างกันรวม 17 ข้อ ได้แก่ ข้อ 1, 2, 3, 4, 8, 18, 21, 26, 27, 29, 31, 32, 34, 35, 36, 37 และ 46 คิดเป็นร้อยละ 34 โดยมีข้อสอบที่ตรวจสอบได้สอดคล้องกันทั้งสองวิธีรวม 8 ข้อ ได้แก่ ข้อ 1, 18, 21, 26, 27, 31, 35 และ 46

เมื่อพิจารณารูปแบบแยกตามเนื้อหา พบว่า ตรวจสอบด้วยวิธีชิปเทสต์มีจำนวนข้อสอบที่ทำหน้าที่ต่างกันรวม 8 ข้อ ได้แก่ ข้อ 18, 27, 28, 29, 31, 35, 45 และ 46 คิดเป็นร้อยละ 16 และตรวจสอบด้วยวิธีถดถอยโลจิสติกมีจำนวนข้อสอบที่ทำหน้าที่ต่างกันรวม 13 ข้อ ได้แก่ ข้อ 4, 6, 20, 26, 27, 29, 31, 32, 34, 35, 36, 45 และ 46 คิดเป็นร้อยละ 26 โดยมีข้อสอบที่ตรวจสอบได้สอดคล้องกันทั้งสองวิธีรวม 6 ข้อ ได้แก่ ข้อ 27, 29, 31, 35, 45 และ 46

เมื่อพิจารณารูปแบบแยกตามระดับพฤติกรรม พบว่า ตรวจสอบด้วยวิธีชิปเทสต์มีจำนวนข้อสอบที่ทำหน้าที่ต่างกันรวม 10 ข้อ ได้แก่ ข้อ 3, 14, 18, 21, 24, 27, 31, 35, 42 และ 45 คิดเป็นร้อยละ 20 และตรวจสอบด้วยวิธีถดถอยโลจิสติกมีจำนวนข้อสอบที่ทำหน้าที่ต่างกันรวม 17 ข้อ ได้แก่ ข้อ 2, 3, 8, 13, 17, 18, 22, 24, 25, 26, 27, 29, 31, 35, 36, 42 และ 46 คิดเป็นร้อยละ 34 โดยมีข้อสอบที่ตรวจสอบได้สอดคล้องกันทั้งสองวิธีรวม 7 ข้อ ได้แก่ ข้อ 3, 18, 24, 27, 31, 35, และ 42

3. การเปรียบเทียบจำนวนข้อสอบที่ทำหน้าที่ต่างกันโดยใช้รูปแบบและวิธีการต่างกัน

การวิเคราะห์ในส่วนนี้ ได้จากการนำผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีชิปเทสต์และวิธีถดถอยโลจิสติก ซึ่งพิจารณารูปแบบรวมทั้งฉบับ แยกตามเนื้อหา และแยกตามระดับพฤติกรรม มาทดสอบความแตกต่างจำนวนข้อที่ทำหน้าที่ต่างกันในแต่ละรูปแบบและในแต่ละวิธีการตรวจสอบด้วยสูตร Q-test โดยใช้โปรแกรม Microsoft Excel ดังแสดงในตาราง 6

ตาราง 6 การเปรียบเทียบจำนวนข้อสอบที่ทำหน้าที่ต่างกันโดยใช้รูปแบบและวิธีการต่างกัน

วิธีการตรวจสอบ รูปแบบ	K	SIBTEST		LRDIF		Q
		DIF items	No DIF items	DIF items	No DIF items	
รวมทั้งฉบับ	50	9	41	17	33	6.4000*
แยกตามเนื้อหา						
ระบบจำนวนเต็ม	15	-	15	2	13	2.0
ความเท่ากันทุกประการ	7	1	6	1	6	-
เศษส่วนและทศนิยม	15	5	10	8	7	1.8
สมบัติของรูปสามเหลี่ยมมุมฉาก	13	2	11	2	11	-
รวม	50	8	42	13	37	2.7778
แยกตามระดับพฤติกรรม						
ความรู้หรือทักษะการคิดคำนวณ	8	1	7	3	5	2.0
ความเข้าใจ	14	4	10	7	7	3.0
การนำไปใช้	21	4	17	5	16	0.2
การวิเคราะห์	7	1	6	2	5	0.3
รวม	50	10	40	17	33	3.7692
Q		0.5455		1.6842		

* มีนัยสำคัญทางสถิติที่ระดับ .05

ผลการวิเคราะห์ตามตาราง 6 มีดังนี้

การเปรียบเทียบจำนวนข้อสอบที่ทำหน้าที่ต่างกัน โดยใช้รูปแบบการตรวจสอบต่างกัน ในแต่ละวิธีการตรวจสอบ พบว่า

เมื่อใช้วิธีการตรวจสอบด้วยวิธีซิปเทสต์ ตรวจสอบโดยใช้รูปแบบรวมทั้งฉบับ พบข้อสอบที่ทำหน้าที่ต่างกันรวม 9 ข้อ ข้อสอบที่ทำหน้าที่ไม่ต่างกันรวม 41 ข้อ รูปแบบแยกตามเนื้อหา พบข้อสอบที่ทำหน้าที่ต่างกันรวม 8 ข้อ ข้อสอบที่ทำหน้าที่ไม่ต่างกันรวม 42 ข้อ และรูปแบบแยกตามระดับพฤติกรรม พบข้อสอบที่ทำหน้าที่ต่างกันรวม 10 ข้อ ข้อสอบที่ทำหน้าที่ไม่ต่างกันรวม 40 ข้อ เมื่อทดสอบความแตกต่างของจำนวนข้อที่ได้จากการตรวจสอบโดยใช้รูปแบบต่างกัน พบว่า จำนวนข้อสอบที่ทำหน้าที่ต่างกันแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ แสดงว่า รูปแบบรวมทั้งฉบับ แยกตามเนื้อหา และแยกตามระดับพฤติกรรม บ่งชี้การทำหน้าที่ต่างกันของข้อสอบได้จำนวนข้อไม่แตกต่างกัน

เมื่อใช้รูปแบบแยกตามระดับพฤติกรรม ตรวจสอบด้วยวิธีชิปทดสอบ พบข้อสอบที่ทำหน้าที่ต่างกันรวม 10 ข้อ ข้อสอบที่ทำหน้าที่ไม่ต่างกันรวม 40 ข้อ และตรวจสอบด้วยวิธีถดถอยโลจิสติก พบข้อสอบที่ทำหน้าที่ต่างกันรวม 17 ข้อ ข้อสอบที่ทำหน้าที่ไม่ต่างกันรวม 33 ข้อ เมื่อทดสอบความแตกต่างของจำนวนข้อที่ได้จากการตรวจสอบด้วยวิธีต่างกัน พบว่า จำนวนข้อสอบที่ทำหน้าที่ต่างกันแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ แสดงว่า วิธีชิปทดสอบกับวิธีถดถอยโลจิสติกบ่งชี้การทำหน้าที่ต่างกันของข้อสอบได้จำนวนข้อไม่แตกต่างกัน

4. การเปรียบเทียบค่าความเชื่อมั่นของแบบทดสอบผลสัมฤทธิ์วิชาคณิตศาสตร์ หลังจากคัดข้อสอบที่ทำหน้าที่ต่างกันออก โดยใช้รูปแบบและวิธีการต่างกัน

4.1 การวิเคราะห์ในส่วนนี้ ได้จากการนำแบบทดสอบหลังจากคัดข้อสอบที่ทำหน้าที่ต่างกันออก โดยใช้รูปแบบที่ศึกษา คือ รวมทั้งฉบับ แยกตามเนื้อหา และแยกตามระดับพฤติกรรม ที่ตรวจสอบด้วยวิธีชิปเทสต์และวิธีถดถอยโลจิสติก มาคำนวณค่าความเชื่อมั่นของแบบทดสอบด้วยสูตร r_B แล้วเปรียบเทียบกัน เนื่องจากการคัดข้อสอบที่ทำหน้าที่ต่างกันออก มีจำนวนข้อไม่เท่ากัน จึงปรับขยายค่าความเชื่อมั่นของแบบทดสอบให้มีจำนวนข้อเท่ากัน (แบบทดสอบทั้งฉบับมีจำนวน 50 ข้อ) ด้วยสูตรสเปียร์แมน-บราวน์ (Spearman-Brown) โดยใช้โปรแกรม Microsoft Excel แล้วเปรียบเทียบค่าความเชื่อมั่นของแบบทดสอบด้วยสูตร UX_1 โดยใช้โปรแกรม Microsoft Excel ปรากฏผลดังตาราง 7 - 8

ตาราง 7 ค่าความเชื่อมั่นของแบบทดสอบหลังการปรับขยายให้แบบทดสอบมีจำนวนข้อสอบเท่ากัน

วิธีการตรวจสอบ รูปแบบ	SIBTEST		LRDIF	
	r_{xx} (K)	r_{tt} (K)	r_{xx} (K)	r_{tt} (K)
รวมทั้งฉบับ	.7454 (41)	.7812 (50)	.6458 (33)	.7342 (50)
แยกตามเนื้อหา	.7592 (42)	.7896 (50)	.6768 (37)	.7389 (50)
แยกตามระดับพฤติกรรม	.7150 (40)	.7582 (50)	.6591 (33)	.7455 (50)

(ตัวเลขในวงเล็บ คือ จำนวนข้อของข้อสอบ)

ผลการวิเคราะห์ตามตาราง 7 พบว่า ค่าความเชื่อมั่นของแบบทดสอบที่ปรับขยายให้มีจำนวนข้อสอบเท่ากัน ในแต่ละรูปแบบมีผลดังนี้

เมื่อใช้รูปแบบทั้งฉบับ พบว่า ค่าความเชื่อมั่นของแบบทดสอบโดยใช้วิธีการตรวจสอบด้วยวิธีชิปเทสต์มีค่าเท่ากับ .7454 นำมาปรับขยายแล้วมีค่าเท่ากับ .7812 และตรวจสอบด้วยวิธีถดถอยโลจิสติกชิปเทสต์มีค่าเท่ากับ .6458 นำมาปรับขยายแล้วมีค่าเท่ากับ .7342

เมื่อใช้รูปแบบแยกตามเนื้อหา พบว่า ค่าความเชื่อมั่นของแบบทดสอบโดยใช้วิธีการตรวจสอบด้วยวิธีชิปเทสต์มีค่าเท่ากับ .7592 นำมาปรับขยายแล้วมีค่าเท่ากับ .7896 และตรวจสอบด้วยวิธีถดถอยโลจิสติกชิปเทสต์มีค่าเท่ากับ .6768 นำมาปรับขยายแล้วมีค่าเท่ากับ .7389

เมื่อใช้รูปแบบแยกตามระดับพฤติกรรม พบว่า ค่าความเชื่อมั่นของแบบทดสอบโดยใช้วิธีการตรวจสอบด้วยวิธีชิปเทสต์มีค่าเท่ากับ .7150 นำมาปรับขยายแล้วมีค่าเท่ากับ .7582 และตรวจสอบด้วยวิธีถดถอยโลจิสติกมีค่าเท่ากับ .6591 นำมาปรับขยายแล้วมีค่าเท่ากับ .7455

ตาราง 8 การเปรียบเทียบค่าความเชื่อมั่นของแบบทดสอบหลังจากคัดข้อสอบที่ทำหน้าที่ต่างกัน โดยใช้รูปแบบและวิธีการต่างกัน

วิธีการตรวจสอบ รูปแบบ	K	SIBTEST	LRDIF	UX ₁
		ค่าความเชื่อมั่น	ค่าความเชื่อมั่น	
รวมทั้งฉบับ	50	.7812	.7342	9.0148*
แยกตามเนื้อหา	50	.7896	.7389	11.0927*
แยกตามระดับพฤติกรรม	50	.7582	.7455	0.6249
UX ₁		6.8604*	0.4555	

* มีนัยสำคัญทางสถิติที่ระดับ .05

ผลการวิเคราะห์ตามตาราง 8 มีดังนี้

การเปรียบเทียบค่าความเชื่อมั่นของแบบทดสอบหลังจากคัดข้อสอบที่ทำหน้าที่ต่างกันออก โดยใช้รูปแบบการตรวจสอบต่างกัน ในแต่ละวิธีการตรวจสอบ ได้ผลดังนี้

เมื่อใช้วิธีการตรวจสอบด้วยวิธีชิปเทสต์ พบว่า ค่าความเชื่อมั่นของแบบทดสอบหลังจากคัดข้อสอบที่ทำหน้าที่ต่างกันออก โดยใช้รูปแบบรวมทั้งฉบับ แยกตามเนื้อหา และแยกตามระดับพฤติกรรมแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ($UX_1 = 6.8604$) แสดงว่า ค่าความเชื่อมั่นของแบบทดสอบแตกต่างกัน เมื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้รูปแบบต่างกัน

เมื่อใช้วิธีการตรวจสอบด้วยวิธีถดถอยโลจิสติก พบว่า ค่าความเชื่อมั่นของแบบทดสอบหลังจากคัดข้อสอบที่ทำหน้าที่ต่างกันออก โดยใช้รูปแบบรวมทั้งฉบับ แยกตามเนื้อหา และแยกตามระดับพฤติกรรมแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ แสดงว่า ค่าความเชื่อมั่นของแบบทดสอบ ไม่แตกต่างกัน เมื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้รูปแบบต่างกัน

การเปรียบเทียบค่าความเชื่อมั่นของแบบทดสอบหลังจากคัดข้อสอบที่ทำหน้าที่ต่างกันออก โดยใช้วิธีการตรวจสอบต่างกัน ในแต่ละรูปแบบการตรวจสอบ ได้ผลดังนี้

เมื่อใช้รูปแบบรวมทั้งฉบับ พบว่า ค่าความเชื่อมั่นของแบบทดสอบหลังจากคัดข้อสอบที่ทำหน้าที่ต่างกันออก โดยใช้การตรวจสอบด้วยวิธีชิปเทสต์และวิธีถดถอยโลจิสติกแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ($UX_1 = 9.0148$) แสดงว่า ค่าความเชื่อมั่นของแบบทดสอบแตกต่างกัน เมื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีต่างกัน

เมื่อใช้รูปแบบแยกตามเนื้อหา พบว่า ค่าความเชื่อมั่นของแบบทดสอบหลังจากคัดข้อสอบที่ทำหน้าที่ต่างกันออก โดยใช้การตรวจสอบด้วยวิธีชิปเทสต์และวิธีถดถอยโลจิสติกแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ($UX_1 = 11.0927$) แสดงว่า ค่าความเชื่อมั่นของแบบทดสอบแตกต่างกัน เมื่อตรวจสอบการทำหน้าที่ ต่างกันของข้อสอบด้วยวิธีต่างกัน

เมื่อใช้รูปแบบแยกตามระดับพฤติกรรม พบว่า ค่าความเชื่อมั่นของแบบทดสอบหลังจากตัดข้อสอบที่ทำหน้าที่ต่างกันออก โดยใช้การตรวจสอบด้วยวิธีชิปเทสต์และวิธีถดถอยโลจิสติกแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ แสดงว่า ค่าความเชื่อมั่นของแบบทดสอบไม่แตกต่างกัน เมื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีต่างกัน

4.2 ผลการวิเคราะห์ตามตาราง 8 พบว่า ค่าความเชื่อมั่นของแบบทดสอบโดยใช้รูปแบบต่างกัน ตรวจสอบด้วยวิธีชิปเทสต์มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แสดงว่า ค่าความเชื่อมั่นของแบบทดสอบอย่างน้อย 1 รูปแบบที่ใช้ตรวจสอบ มีค่าแตกต่างจากรูปแบบอื่น ดังนั้น จึงทดสอบความแตกต่างค่าความเชื่อมั่นรายคู่อีกครั้งหนึ่ง ปรากฏผลดังตาราง 9

ตาราง 9 การเปรียบเทียบความแตกต่างของค่าความเชื่อมั่นของแบบทดสอบระหว่างรูปแบบรายคู่ โดยใช้วิธีการตรวจสอบด้วยวิธีชิปเทสต์ ทดสอบด้วย UX_1

รูปแบบ	รวมทั้งฉบับ (.7812)	แยกตามเนื้อหา (.7896)	แยกตามระดับพฤติกรรม (.7582)
รวมทั้งฉบับ (.7812)		0.3655	2.3817
แยกตามเนื้อหา (.7896)	ไม่แตกต่าง		4.6107*
แยกตามระดับพฤติกรรม (.7582)	ไม่แตกต่าง	แตกต่าง	

* มีนัยสำคัญทางสถิติที่ระดับ .05

ผลการวิเคราะห์ตามตาราง 9 พบว่า รูปแบบแยกตามเนื้อหา มีค่าความเชื่อมั่นสูงสุด และรองลงมาตามลำดับ ได้แก่ รูปแบบรวมทั้งฉบับ และรูปแบบแยกตามระดับพฤติกรรม จากการเปรียบเทียบค่าความเชื่อมั่นเป็นรายคู่ พบว่า

รูปแบบรวมทั้งฉบับ มีค่าความเชื่อมั่นไม่แตกต่างจากรูปแบบแยกตามเนื้อหา และรูปแบบแยกตามระดับพฤติกรรม

รูปแบบแยกตามเนื้อหา มีค่าความเชื่อมั่นแตกต่างจากรูปแบบแยกตามระดับพฤติกรรม

บทที่ 5

สรุปผล อภิปรายผล และข้อเสนอแนะ

ความมุ่งหมายของการศึกษาค้นคว้า

การศึกษาค้นคว้าครั้งนี้มีจุดมุ่งหมายสำคัญ เพื่อศึกษาผลการตรวจสอบการทำหน้าที่ต่างกันโดยใช้รูปแบบต่างกันสามรูปแบบ คือ คะแนนรวมทั้งฉบับ แยกตามเนื้อหา และแยกตามระดับพฤติกรรม ด้วยวิธีการตรวจสอบต่างกันสองวิธี คือ วิธีชิปทดสอบ และวิธีถดถอยโลจิสติก เป็นจุดมุ่งหมายเฉพาะของการศึกษาค้นคว้าต่อไปนี้

1. เพื่อเปรียบเทียบจำนวนข้อสอบที่ทำหน้าที่ต่างกันโดยใช้รูปแบบการตรวจสอบต่างกันในแต่ละวิธีการตรวจสอบ
2. เพื่อเปรียบเทียบจำนวนข้อสอบที่ทำหน้าที่ต่างกันโดยใช้วิธีการตรวจสอบต่างกันในแต่ละรูปแบบการตรวจสอบ
3. เพื่อเปรียบเทียบค่าความเชื่อมั่นของแบบทดสอบ หลังจากคัดข้อสอบที่ทำหน้าที่ต่างกันออก โดยใช้รูปแบบการตรวจสอบต่างกันในแต่ละวิธีการตรวจสอบ
4. เพื่อเปรียบเทียบค่าความเชื่อมั่นของแบบทดสอบ หลังจากคัดข้อสอบที่ทำหน้าที่ต่างกันออก โดยใช้วิธีการตรวจสอบต่างกันในแต่ละรูปแบบการตรวจสอบ

กลุ่มตัวอย่าง

กลุ่มตัวอย่างที่ใช้ในการศึกษาค้นคว้าครั้งนี้ เป็นนักเรียนชั้นมัธยมศึกษาปีที่ 2 ภาคเรียนที่ 2 ปีการศึกษา 2542 ของโรงเรียนมัธยมศึกษา สังกัดกรมสามัญศึกษา กรุงเทพมหานคร ส่วนกลาง กลุ่มที่ 4 ซึ่งได้มาโดยการสุ่มแบบแบ่งชั้น (Stratified Random Sampling) มีขนาดโรงเรียนเป็นชั้น (Strata) และมีห้องเรียนเป็นหน่วยการสุ่ม (Sampling Unit) จำนวนโรงเรียนทั้งหมด 6 โรงเรียน จำนวนห้องเรียนทั้งหมด 23 ห้องเรียน รวมนักเรียนทั้งสิ้น 1,051 คน จำแนกเป็นนักเรียนชาย 522 คน และนักเรียนหญิง 529 คน

เครื่องมือที่ใช้ในการศึกษาค้นคว้า

เครื่องมือที่ใช้ในการศึกษาค้นคว้าครั้งนี้ เป็นแบบทดสอบผลสัมฤทธิ์วิชาคณิตศาสตร์ (ค 203) ชั้นมัธยมศึกษาปีที่ 2 ตามหลักสูตรมัธยมศึกษาตอนต้น พุทธศักราช 2521 (ฉบับปรับปรุง พ.ศ. 2533) เป็นแบบเลือกตอบชนิด 5 ตัวเลือก จำนวน 50 ข้อ ที่ผู้วิจัยสร้างขึ้นเอง 1 ฉบับ

วิธีดำเนินการเก็บรวบรวมข้อมูล

การเก็บรวบรวมข้อมูลครั้งนี้ ผู้วิจัยได้ดำเนินการตามลำดับขั้นตอน ดังนี้

1. ติดต่อประสานงานกับโรงเรียนที่เป็นกลุ่มตัวอย่าง เพื่อกำหนดวัน เวลา ในการนำแบบทดสอบไปทดสอบกับนักเรียนที่เป็นกลุ่มตัวอย่าง

2. จัดเตรียมแบบทดสอบให้เพียงพอกับจำนวนนักเรียนที่จะทำการสอบในแต่ละครั้ง
3. ชี้แจงทำความเข้าใจกับผู้ประสานงานซึ่งอยู่ที่โรงเรียน ในเรื่องลักษณะของแบบทดสอบ และวิธีดำเนินการสอบ
4. นำแบบทดสอบที่มีคุณภาพจำนวน 50 ข้อ ไปทดสอบกับกลุ่มตัวอย่างที่มีจำนวน 1,051 คน เพื่อศึกษาผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้การตรวจสอบด้วยวิธีชิปทดสอบ และวิธีถดถอยโลจิสติก ในแต่ละรูปแบบที่ศึกษา และทำการทดสอบสมมุติฐานการศึกษาค้นคว้า ข้อ 1 และ ข้อ 2
5. นำผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามข้อ 4 มาคัดข้อสอบที่ทำหน้าที่ต่างกันออก คำนวณหาค่าความเชื่อมั่นของแบบทดสอบ เพื่อเปรียบเทียบกับในแต่ละรูปแบบการตรวจสอบและในแต่ละวิธีการตรวจสอบ โดยปรับขยายค่าความเชื่อมั่นของแบบทดสอบให้มีจำนวนข้อเท่ากันก่อน และทำการทดสอบสมมุติฐานการศึกษาค้นคว้าข้อ 3 และ ข้อ 4

การเสนอผลการวิเคราะห์ข้อมูล

การเสนอผลการวิเคราะห์ข้อมูลครั้งนี้ ผู้วิจัยได้นำเสนอตามลำดับดังนี้

1. ค่าสถิติพื้นฐานของแบบทดสอบผลสัมฤทธิ์วิชาคณิตศาสตร์ จำแนกตามเนื้อหา และระดับพฤติกรรม
2. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจากแบบทดสอบผลสัมฤทธิ์วิชาคณิตศาสตร์โดยใช้รูปแบบและวิธีการต่างกัน
3. การเปรียบเทียบจำนวนข้อสอบที่ทำหน้าที่ต่างกันโดยใช้รูปแบบและวิธีการต่างกัน
4. การเปรียบเทียบค่าความเชื่อมั่นของแบบทดสอบผลสัมฤทธิ์วิชาคณิตศาสตร์ หลังจากคัดข้อสอบที่ทำหน้าที่ต่างกันออกโดยใช้รูปแบบและวิธีการต่างกัน

สรุปผลการวิเคราะห์ข้อมูล

ผลการวิเคราะห์ข้อมูลในการศึกษาค้นคว้า สรุปได้ดังนี้

1. จำนวนข้อสอบที่ทำหน้าที่ต่างกันโดยใช้รูปแบบการตรวจสอบต่างกันในแต่ละวิธีการตรวจสอบ
 - 1.1 ตรวจสอบด้วยวิธีชิปทดสอบ พบว่า จำนวนข้อสอบที่ทำหน้าที่ต่างกันโดยใช้รูปแบบการตรวจสอบรวมทั้งฉบับ แยกตามเนื้อหา และแยกตามระดับพฤติกรรม แตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ
 - 1.2 ตรวจสอบด้วยวิธีถดถอยโลจิสติก พบว่า จำนวนข้อสอบที่ทำหน้าที่ต่างกันโดยใช้รูปแบบการตรวจสอบรวมทั้งฉบับ แยกตามเนื้อหา และแยกตามระดับพฤติกรรม แตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

อภิปรายผล

จากการศึกษาครั้งนี้ สามารถอภิปรายผลได้ดังนี้

1. จำนวนข้อสอบที่ทำหน้าที่ต่างกันจากแบบทดสอบผลสัมฤทธิ์วิชาคณิตศาสตร์ โดยใช้รูปแบบการตรวจสอบรวมทั้งฉบับ แยกตามเนื้อหา และแยกตามระดับพฤติกรรม ในแต่ละวิธีการตรวจสอบมีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ ไม่ว่าจะตรวจสอบด้วยวิธีชิปเทสต์หรือวิธีถดถอยโลจิสติก ข้อค้นพบดังกล่าวไม่สอดคล้องกับสมมุติฐานการศึกษาค้นคว้าข้อ 1 ทั้งนี้อาจเนื่องมาจาก ทั้งสองวิธีการตรวจสอบพบข้อสอบที่ทำหน้าที่ต่างกันในแต่ละรูปแบบ มีจำนวนข้อใกล้เคียงกัน
2. จำนวนข้อสอบที่ทำหน้าที่ต่างกันจากแบบทดสอบผลสัมฤทธิ์วิชาคณิตศาสตร์ โดยใช้วิธีการตรวจสอบด้วยวิธีชิปเทสต์และวิธีถดถอยโลจิสติกนั้น ในแต่ละรูปแบบการตรวจสอบมีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ในรูปแบบรวมทั้งฉบับ ซึ่งข้อค้นพบดังกล่าวสอดคล้องกับสมมุติฐานการศึกษาค้นคว้า ข้อ 2 ส่วนในรูปแบบแยกตามเนื้อหา และแยกตามระดับพฤติกรรมนั้น พบมีจำนวนข้อสอบที่ทำหน้าที่ต่างกันไม่แตกต่างกัน ซึ่งไม่สอดคล้องกับสมมุติฐานการศึกษาค้นคว้า ข้อ 2 ทั้งนี้อาจเป็นเพราะ คะแนนผลการสอบในการศึกษาครั้งนี้ มีความแตกต่างระหว่างคะแนนเฉลี่ยและสัดส่วนการตอบข้อสอบระหว่างผู้สอบที่มีความสามารถระดับเดียวกันและต่างระดับกัน ทำให้ตรวจสอบด้วยวิธีถดถอยโลจิสติกพบข้อสอบที่ทำหน้าที่ต่างกันมากกว่าวิธีชิปเทสต์ ซึ่งข้อค้นพบดังกล่าว สอดคล้องกับผลการศึกษาของสวามินาธานและโรเจอร์ส (Swaminathan & Rogers, 1990) ที่พบว่า วิธีถดถอยโลจิสติกมีความไวในการตรวจการทำหน้าที่ต่างกันของข้อสอบ คือ พบข้อสอบที่ทำหน้าที่ต่างกันมาก และสามารถตรวจสอบการทำหน้าที่ต่างกันแบบสม่ำเสมอและแบบไม่สม่ำเสมอของข้อสอบได้ดี
3. ค่าความเชื่อมั่นของแบบทดสอบหลังการคัดข้อสอบที่ทำหน้าที่ต่างกันออกโดยใช้รูปแบบการตรวจสอบต่างกัน ในแต่ละวิธีการตรวจสอบ ทำการปรับขยายค่าความเชื่อมั่นของแบบทดสอบให้มีจำนวนข้อเท่ากันก่อน ด้วยสูตรสเปียร์แมน-บราวน์ (Spearman-Brown) แล้วทดสอบความแตกต่างของค่าความเชื่อมั่น พบว่า ตรวจสอบด้วยวิธีชิปเทสต์ มีค่าความเชื่อมั่นแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ข้อค้นพบดังกล่าว สอดคล้องกับสมมุติฐานการศึกษาค้นคว้าข้อ 3 จึงทดสอบความแตกต่างของค่าความเชื่อมั่นระหว่างรูปแบบรายคู่ พบว่า มีค่าความเชื่อมั่นแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ในรูปแบบที่แยกตามเนื้อหา กับแยกตามระดับพฤติกรรม โดยที่รูปแบบแยกตามเนื้อหาให้ค่าความเชื่อมั่นสูงสุด ทั้งนี้อาจเป็นเพราะ การคัดข้อออกในแต่ละรูปแบบเป็นข้อที่ไม่ซ้ำกันทุกข้อและมีจำนวนข้อไม่เท่ากัน ทำให้ค่าความเชื่อมั่นของแบบทดสอบต่างกัน ส่วนการตรวจสอบด้วยวิธีถดถอยโลจิสติกมีค่าความเชื่อมั่นแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ ทั้งนี้อาจเป็นเพราะการตรวจสอบด้วยวิธีถดถอยโลจิสติกเป็นข้อที่ซ้ำกันมากกว่าในแต่ละรูปแบบการตรวจสอบ ทำให้ค่าความเชื่อมั่นไม่แตกต่างกัน ดังที่บุญเชิด ภิญโญนนตพงษ์ (ม.ป.ป. : 277-278) กล่าวว่า iva จำนวนข้อของแบบทดสอบมีอิทธิพลต่อค่าความเชื่อมั่น แบบทดสอบใดมีจำนวนข้อน้อยจะมีค่าความเชื่อมั่นต่ำ ถ้ามีจำนวนข้อมากจะมีค่าความเชื่อมั่นสูง และข้อค้นพบดังกล่าวสอดคล้องกับผลการศึกษาของคลอเซอร์และคณะ (Clauser, et al. 1996) ที่พบว่า แบบทดสอบที่มีจุดมุ่งหมายใน

การวัดเพียงมิติเดียวนั้น การใช้คะแนนรวมทั้งฉบับในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมีความเหมาะสมที่สุด เพราะเป็นตัวแทนความสามารถที่มีความเที่ยงตรงในการวัดคุณลักษณะแฝงมากที่สุด ส่วนแบบทดสอบที่มีการวัดในหลายมิติ การใช้คะแนนรวมในการตรวจสอบทำให้เกิดความผิดพลาดของผลการตรวจสอบ เนื่องจากคะแนนผลการสอบที่ไม่สัมพันธ์กับคุณลักษณะแฝงที่ต้องการวัด เป็นข้อสอบที่ตรวจสอบว่าทำหน้าที่ต่างกัน แต่เมื่อใช้คะแนนผลการสอบย่อยจะให้ผลการตรวจสอบดีกว่า

4. ค่าความเชื่อมั่นของแบบทดสอบหลังการคัดข้อสอบที่ทำหน้าที่ต่างกันออกโดยใช้วิธีการตรวจสอบต่างกัน ในแต่ละรูปแบบการตรวจสอบ ทำการปรับขยายค่าความเชื่อมั่นของแบบทดสอบให้มีจำนวนข้อเท่ากันด้วยสูตรสเปียร์แมน-บราวน์ (Spearman-Brown) แล้วทดสอบความแตกต่างของค่าความเชื่อมั่น พบว่า แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ในรูปแบบรวมทั้งฉบับ และแยกตามเนื้อหา ข้อค้นพบดังกล่าว สอดคล้องกับสมมติฐานการศึกษาค้นคว้าข้อ 4 ส่วนในรูปแบบแยกตามระดับพฤติกรรม มีค่าความเชื่อมั่นแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติทั้งนี้อาจเป็นเพราะ การคัดข้อสอบที่ทำหน้าที่ต่างกันออก ไม่ว่าจะตรวจสอบด้วยวิธีชิปเทสต์หรือวิธีถดถอยโลจิสติก ในรูปแบบรวมทั้งฉบับและแยกตามเนื้อหา พบข้อสอบที่ทำหน้าที่ต่างกันเป็นข้อที่ไม่ซ้ำกันมากกว่ารูปแบบแยกตามระดับพฤติกรรม

ข้อเสนอแนะ

1. ข้อเสนอแนะในการนำผลการศึกษาไปใช้

1.1. จากผลการศึกษาครั้งนี้ พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้รูปแบบและวิธีการต่างกัน ควรใช้รูปแบบแยกตามเนื้อหาเพราะให้ค่าความเชื่อมั่นสูงสุด และใช้วิธีการตรวจสอบด้วยวิธีชิปเทสต์ ถึงแม้ว่าวิธีถดถอยโลจิสติกจะมีความไวในการตรวจสอบมากกว่า แต่วิธีชิปเทสต์มีประสิทธิภาพมากกว่า เพราะให้ค่าความเชื่อมั่นสูงกว่า และมีความเหมาะสมในการนำไปใช้ สามารถวิเคราะห์ได้พร้อมกันหลายข้อในคราวเดียวกัน

1.2. ในการสร้างแบบทดสอบ ควรกำหนดให้การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นขั้นตอนหนึ่งของการสร้างแบบทดสอบ ก่อนที่จะนำแบบทดสอบไปใช้ เพื่อจะทำให้แบบทดสอบมีคุณภาพดีขึ้น ได้ผลการทดสอบที่มีความยุติธรรมมากขึ้น

2. ข้อเสนอแนะในการทำวิจัย

2.1 ควรมีการศึกษาเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจากแบบทดสอบผลสัมฤทธิ์ เนื้อหาวิชาอื่นๆ และระดับชั้นอื่นๆ และเปรียบเทียบกับตัวแปรอื่นๆ เช่น การเรียนพิเศษของนักเรียน สถานภาพของครอบครัว เป็นต้น

2.2 ควรมีการศึกษาเทคนิคการตรวจสอบการทำหน้าที่ต่างกัน เช่น แบบทดสอบที่มีการให้คะแนนแบบหลายค่า (Polytomous) การตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบ (Differential Bundle Functioning หรือ DBF) จะเป็นประโยชน์ต่อการจัดการเรียนการสอนและการพัฒนาหลักสูตร

บรรณานุกรม

บรรณานุกรม

- กระทรวงศึกษาธิการ. กรมวิชาการ. (2535). หลักสูตรมัธยมศึกษาตอนต้น พุทธศักราช 2521 (ฉบับปรับปรุง พ.ศ. 2533). พิมพ์ครั้งที่ 2. กรุงเทพฯ : โรงพิมพ์คุรุสภาลาดพร้าว.
- กระทรวงศึกษาธิการ. สถาบันส่งเสริมการสอนวิทยาศาสตร์และเทคโนโลยี. (2536). คู่มือครู วิชาคณิตศาสตร์ ชั้นมัธยมศึกษาปีที่สอง. พิมพ์ครั้งที่ 1. กรุงเทพฯ : โรงพิมพ์คุรุสภาลาดพร้าว.
- คมศักดิ์ ชื่นชม. (2539). การศึกษาผลการวิเคราะห์ความลำเอียงที่ใช้วิธีต่างกันของแบบทดสอบวัดจริยธรรมด้านความซื่อสัตย์. วิทยานิพนธ์ กศ.ม. (การวัดผลการศึกษา). กรุงเทพฯ : บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ. ถ่ายเอกสาร.
- จิตสุตา ธารพร. (2539). การเปรียบเทียบผลการวิเคราะห์ความลำเอียงของข้อสอบจากแบบทดสอบโจทย์ปัญหาทางคณิตศาสตร์และแบบทดสอบความเข้าใจทางภาษา. วิทยานิพนธ์ กศ.ม. (การวัดผลการศึกษา). กรุงเทพฯ : บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ. ถ่ายเอกสาร.
- เจน จันท์แสง. (2537). การวิเคราะห์ความลำเอียงของข้อสอบของแบบสอบวัดผลสัมฤทธิ์ทางการเรียนคณิตศาสตร์ ชั้นมัธยมศึกษาปีที่ 3. วิทยานิพนธ์ กศ.ม. (การวัดผลการศึกษา). กรุงเทพฯ : บัณฑิตวิทยาลัย มหาวิทยาลัยนครสวรรค์. ถ่ายเอกสาร.
- ฉัตรศิริ ปิยะพิมลสิทธิ์. (2542). การวิเคราะห์ความลำเอียงด้วยโปรแกรมสำเร็จรูป SIBTEST. (Online). Available : measurement.hypermart.net; Directory : วารสารรายสัปดาห์. ขวาล แพร่ดกุล. (2509). เทคนิคการวัดผล. กรุงเทพฯ : วัฒนาพานิช.
- ญาณภัทร สีหะมงคล. (2540). การเปรียบเทียบความสอดคล้องของผลการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันระหว่าง วิธี Lord's χ^2 วิธี Raju's Area Measures และวิธี Closed Interval Area. วิทยานิพนธ์ กศ.ด. (การทดสอบและวัดผลการศึกษา). กรุงเทพฯ : บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ. ถ่ายเอกสาร.
- ธวัชชัย งามสันติวงศ์. (2542). หลักการและวิธีใช้คอมพิวเตอร์ในงานสถิติเพื่อการวิจัย SPSS FOR WINDOWS. พิมพ์ครั้งที่ 5. กรุงเทพฯ : จุฬาลงกรณ์มหาวิทยาลัย.
- นิภา ศรีไพโรจน์. (2524). สถิติประยุกต์ทางการศึกษา. กรุงเทพฯ : โรงพิมพ์รุ่งเรืองการพิมพ์.
- นิรมล ชัยขวลิต. (2537). การเปรียบเทียบผลการวิเคราะห์ความลำเอียงของแบบทดสอบความเข้าใจในการอ่านภาษาไทย ตามทฤษฎีคลาสสิกอลที่ใช้วิธีวิเคราะห์ต่างกัน. วิทยานิพนธ์ กศ.ม. (การวัดผลการศึกษา). กรุงเทพฯ : บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ. ถ่ายเอกสาร.

- บุญเชิด ภิบุญอนันตพงษ์. (ม.ป.ป.). การวัดและประเมินผลการศึกษา : ทฤษฎีและการประยุกต์. พิมพ์ครั้งที่ 2. กรุงเทพฯ : อักษรเจริญทัศน์.
- _____. (2527). การทดสอบแบบอิงเกณฑ์ : แนวคิดและวิธีการ. กรุงเทพฯ : ภาควิชาพื้นฐานการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร.
- _____. (2537). การประมาณค่าความเชื่อมั่นของแบบทดสอบที่แบ่งส่วนย่อยตามแบบจำลองคะแนนจริงสัมพันธ์. ปรินูญานิพนธ์ กศ.ด. (การทดสอบและวัดผลการศึกษา). กรุงเทพฯ : บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ. ถ่ายเอกสาร.
- _____. (2542). สัมประสิทธิ์ r_b : การประมาณค่าความเชื่อมั่นสำหรับแบบทดสอบผลสัมฤทธิ์ชนิดเลือกตอบที่ประกอบด้วยความยากรายข้อต่างกัน. กรุงเทพฯ : ภาควิชาการวัดผลและวิจัยการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ.
- ผจงจิต อินทสุวรรณ. (2530, พ.ศ.-ส.ศ.). “แนวทางการศึกษาความลำเอียงของข้อคำถาม,” *การวัดผลการศึกษา*. 9(25) : 54-52.
- พรรณี จินตมาศ. (2540). การเปรียบเทียบผลการวิเคราะห์ความลำเอียงของข้อสอบ โดยใช้ขนาดกลุ่มผู้สอบและวิธีวิเคราะห์ต่างกัน. ปรินูญานิพนธ์ กศ.ม. (การวัดผลการศึกษา). กรุงเทพฯ : บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ. ถ่ายเอกสาร.
- พัชร ปิยภักดิ์. (2531). การวิเคราะห์ความลำเอียงของข้อสอบจากแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียน ชั้นประถมศึกษาปีที่ 6. ปรินูญานิพนธ์ กศ.ม. (การวัดผลการศึกษา). กรุงเทพฯ : บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ. ถ่ายเอกสาร.
- เรวดี อินทสระ. (2539). ผลการตรวจสอบความลำเอียงข้อสอบต่อการศึกษาความเที่ยงตรงเชิงพยากรณ์ของแบบทดสอบคัดเลือกที่คิดคะแนนต่างกัน. ปรินูญานิพนธ์ กศ.ด. (การทดสอบและวัดผลการศึกษา). กรุงเทพฯ : บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ. ถ่ายเอกสาร.
- ล้วน สายยศ และอังคณา สายยศ. (2539). เทคนิคการวัดผลการเรียนรู้. กรุงเทพฯ : ชมรมเด็ก.
- สมศักดิ์ จันผ่อง. (2541, พ.ศ.-ส.ศ.). “จากความลำเอียงของข้อสอบ (Item Bias) มาเป็นการทำหน้าที่ต่างกันของข้อสอบ (DIF),” *ว.ศึกษาศาสตร์ มหาวิทยาลัยนเรศวร*. 3(1) : 81-93.
- สุพัฒน์ สุกมลสันต์. (2538). การวิเคราะห์ข้อสอบแนวใหม่ด้วยคอมพิวเตอร์. พิมพ์ครั้งที่ 1. กรุงเทพฯ : นำอักษรการพิมพ์.
- อังคณา สายยศ. (2539). การศึกษาผลการวิเคราะห์ความลำเอียงของแบบทดสอบความเข้าใจในการอ่านตามลักษณะการใช้ภาษาพูดในครอบครัว. กรุงเทพฯ : ภาควิชาการวัดผลและวิจัยการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร.

- Allen, M.J. & Yen, W.M. (1979). *Introduction to Measurement Theory*. Monterey : California Brooks/Cole.
- Bloom, B.S. (1956). *Taxonomy of Educational Objectives Handbock 1 : The Cognitive Domain*. New York : McKay.
- Bock, R.D. & Lieberman, M. (1970). "Fitting a Response Model for n Dichotomously Scored Items," *Psychometrika*. 35(2) : 179-197.
- Camilli, Gregory. & Shepard, Lorrie A. (1994). *Methods for Identifying Biased Test Items*. California : Sage Publications.
- Clauser, Brian E., et al. (1996). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional," *Journal of Educational Measurement*. 33(3) : 202-214.
- Doolittle, Allen E. & Cleary, Anne. T. (1987). "Gender-Based Differential Item Performance in Mathematics Achievement Items," *Journal of Educational Measurement*. 24(2) : 157-166.
- Dorans, N.J. & Holland, P.W. (1993). "DIF Detection and Description : Mantel-Haenszel and Standardization In P.W. Holland & H. Wainer (Eds.)," Hillsdale, N.J. : Erlbaum.
- Douglas, Jeffrey A., Roussos, Louis A. & Stout, William. (1996). "Item-Bundle DIF Hypothesis Testing : Identifying Suspect Bundles and Assessing Their Differential Functioning," *Journal of Educational Measurement*. 33(4) : 465-484.
- Freedle, Roy. & Kostin, Irene. (1990). "Item Difficulty of Four Verbal Item Types and an Index of Differential Item Functioning for Black and White Examinees," *Journal of Educational Measurement*. 27(4) : 329-343.
- Holland, P.W. & Thayer, D.T. (1988). *Differential Item Performance and the Mantel-Haenszel Procedure*. In Wainer, Howard and Brgun, Henty I Test Validity. Hillsdale, New Jersey. Lawkence Erlbaum Associates.
- Holland, P.W. & Wainer. H (1993). *Differential Item Functioning*. Hillsdale NJ : Erlbaum.
- Intasuwan, P. (1979). *A Comparison of Three Approaches for Determining Item Bias in Cross-National Testing*. Doctor's Thesis The University of Pittsburgh.
- Kelderman, H. & Georerdy, B. M. (1990). "The Use of Loginear Models for Assessing Differential item Functioning Across Manifest and Latent Examinee Groups," *Journal of Educational Measurement*. 27(4) : 370-427.

- Kelderman, H. (1989). "Item Bias Detection Using Loglinear IRT," *Psychometrika*. 54(5) : 681-697.
- Kim, Seock-Ho. & Cohen, Allan S. (1991). "A Comparison of Two Area Measures for Detecting Differential Item Functioning," *Applied Psychological Measurement*. 15(3) : 269-278.
- Li, Hsin-Hung., Nandakumar, R. & Stout, W. (1995, April). *Application of SIBTEST in Dealing With Issues of DIF in the Context of Multidimensional Data*. (Online). Available : FTP : stat.uiuc.edu; Directory: stoutlab/papers.html.
- Lord, F.M. (1980). *Application of Item Response Theory to Practical Testing Problems*. Hillsdale, N.J : Lawrence Erlbaum Associates.
- Mazor, K.M., Clauser, B.E. & Hambleton, R.K. (1992). "The effect of sample size on the functioning of the Mantel-Haenszel statistic," *Educational and Psychological Measurement*. 52(2) : 443-451.
- McLaughlin, M.E. & Drasgow, F. (1987). "Lord's Chi-square Test of Item Bias with Estimated and with Known Ability Parameter," *Applied Psychological Measurement*. 11(2) : 161-173.
- Mellenbergh, G. (1982). "Contingency Table Models for Assessing Item Bias," *Journal of Educational Statistics*. 7(2) : 105-118.
- Mellenbergh, G. (1989). "Item bias and item response theory," *International Journal of Educational Measurement*. 13(2) : 127-143.
- Miller, M.D. & Oshima, T.C. (1992). "Effect of Sample Size, Number of Biased Items, and Magnitude of Bias on a Two-Stage Item Bias Estimation Method," *Applied Psychological Measurement*. 16(4) : 381-388.
- Miller, T.R. & Spray, J.A. (1993). "Logistic Discriminant Function Analysis for DIF Identification of Polytomously Scored Items," *Journal of Educational Measurement*. 30(2) : 107-122.
- Millsap, R. E. & Everson, H.T. (1993). "Methodology Review : Statistical Approaches for Assessing Measurement Bias," *Applied Psychological Measurement*. 17(4) : 297-334.
- Narayanan, Pankaja. & Swaminathan, H. (1994). "Performance of the Mantel-Haenszel and Simultaneous Item Bias Procedures for Detection Differential Item Function," *Applied Psychological Measurement*. 18(4) : 315-328.

- Oshima, T.C. & Miller, M.D. (1992). "Multidimensionality and Item bias in Item Response Theory," *Applied Psychological Measurement*. 16(3) : 237-248.
- Oshima, T.C., Raju, Nambury S., Flowers, C.P. & Slinde, J.A. (1998). "Differential Bundle Functioning Using The DFIT Framework : Procedures for Identifying Possible Sources of Differential Functioning," *Applied Measurement In Education*. 11(4) : 353-369.
- Potenza, Maria T. & Dorans, Neil J. (1995). "DIF Assessment for Polytomously Scored Items : A Framework for Classification and Evaluation," *Applied Psychological Measurement*. 19(1) : 23-27.
- Raju, N.S., Van der Linden, W.J. & Fleer, P.F. (1995). "IRT-Based Internal Measures of Differential Functioning Items and Tests," *Applied Psychological Measurement*. 19(4) : 353-368.
- Rogers, H.J. & Swaminathan, H. (1993). "A Comparison of Logistic Regression and Mantel-Haenszel Procedures for Detecting Differential Item Functioning," *Applied Psychological Measurement*. 17(2) : 105-116.
- Roussos, Louis A. & stout, William F. (1996). "Simulation Studies of the Effect of Small Sample Size and Studied Item Parameter on SIBTEST and Mantel-Haenszel Type I Error Performance," *Journal of Educational Measurement*. 33(2) : 215-230.
- Rudner, Lawrence M., Getson, Pamela R. & Knight, David L. (1980). "A Monte Carlo Comparison of Seven Biased Item Detection Techniques," *Journal of Educational Measurement*. 17(1) : 1-10.
- Scheuneman, J.D. (1979). "A Method for Assessing Bias in Test Items," *Journal of Educational Measurement*. 16(2) : 143-152.
- Shealy, Robin. & Stout, William. (1993). "A Model-Based Standardization Approach that Separates True Bias/DIF from Group Ability Differences and Detects Test Bias/DTF as Well as Item Bias/DIF," *Psychometrika*. 58(2) : 159-194.
- Shepard, L.A. Camilli, G. & Averill, M. (1981). "Test Item Bias," *Journal of Educational Statistics*. 6(4) : 371-375.
- Stout, William F. & Roussos, Louis A. (1992). *SIBTEST USER MANUAL*. University of Illinois.

- Subkoviak, Micheal J., et al. (1984). "Empirical Comparison of Detected Item Bias Detection Procedure with Bias Manipulation," *Journal of Educational Measurement*. 21(10) : 49-59.
- Swaminathan, H. & Rogers, H.J. (1990). "Detecting differential item functioning using logistic regression procedures," *Journal of Educational Measurement*. 27(4) : 361-370.
- Thissen, D., Steinberg, L. & Wainer, H. (1988). *Use of Item Response Theory in the Study of Group Differences in Trace Lines*. In H. Wainer & H. Braun (Eds.), *Test validity*. Hillsdale NJ : Erlbaum.
- Tuckman, Bruce W. (1975). *Measuring Educational Outcomes : Fundamentals of Testing*. New York : Harcourt Brace Jovanovich, inc.
- Wiersma, W. & Jurs, Stephen G. (1990). *Educational Measurement and Testing*. 2nd ed The University of Toledo. United States.
- Wilson, Jame W. (1971). "Evaluation of Learning in Secondary School Mathematics," in *Handbook an Formative Evaluation of Student Learning*. Edit by Benjamin S. Bloom. U.S.A. : McGraw-Hill.

ภาคผนวก

ภาคผนวก ก

รายชื่อผู้เชี่ยวชาญตรวจสอบแนะนำแก้ไขเครื่องมือวิจัย

รายชื่อผู้เชี่ยวชาญ

ผู้ช่วยศาสตราจารย์ระวีวรรณ พันธุ์พานิช	ภาควิชาการวัดผลและวิจัยการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
อาจารย์ ดร.สุพร เข้มเฮง	ภาควิชาการวัดผลและวิจัยการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
อาจารย์อุทัยวรรณ สายพัฒนา	หมวดวิชาคณิตศาสตร์ โรงเรียนธรรมศาสตร์คลองหลวง เขตคลองหลวง จังหวัดปทุมธานี
อาจารย์สายใจ เชียงอินทร์	หมวดวิชาคณิตศาสตร์ โรงเรียนพระหฤทัยคอนแวนต์ เขตคลองเตย กรุงเทพมหานคร
อาจารย์เสมอใจ สิงห์ชัย	หมวดวิชาคณิตศาสตร์ โรงเรียนพระหฤทัยคอนแวนต์ เขตคลองเตย กรุงเทพมหานคร

ภาคผนวก ข

ตารางค่าความยากง่าย (p) และค่าอำนาจจำแนก (r)
ของแบบทดสอบผลสัมฤทธิ์วิชาคณิตศาสตร์

ตาราง 10 ค่าความยากง่าย (p) และค่าอำนาจจำแนก (r) ของแบบทดสอบผลสัมฤทธิ์วิชาคณิตศาสตร์

ข้อที่	p	r	ข้อที่	p	r
1	.21	.27	26	.74	.49
2	.35	.32	27	.44	.56
3	.52	.49	28	.65	.59
4	.60	.46	29	.56	.76
5	.35	.44	30	.61	.83
6	.20	.44	31	.60	.71
7	.79	.39	32	.67	.78
8	.64	.37	33	.49	.76
9	.33	.51	34	.66	.76
10	.67	.49	35	.71	.78
11	.54	.61	36	.43	.54
12	.65	.66	37	.32	.68
13	.54	.61	38	.68	.51
14	.36	.44	39	.35	.46
15	.41	.44	40	.34	.49
16	.33	.39	41	.53	.83
17	.65	.42	42	.76	.63
18	.51	.66	43	.46	.81
19	.47	.68	44	.39	.56
20	.67	.78	45	.49	.66
21	.55	.51	46	.56	.71
22	.37	.59	47	.52	.73
23	.35	.66	48	.61	.66
24	.71	.32	49	.48	.71
25	.57	.66	50	.21	.20

ภาคผนวก ค

ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจากแบบทดสอบผลสัมฤทธิ์วิชาคณิตศาสตร์
ที่มีรูปแบบต่างกัน ด้วยวิธีชิปทดสอบและวิธีถดถอยโลจิสติก

ตาราง 11 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบรวมทั้งฉบับด้วยวิธีซิปเทสต์

ข้อ	β_U	p-value	ข้อ	β_U	p-value
1	.054	.020*	26	.074	.043*
2	.035	.276	27	-.083	.003*
3	.063	.074	28	-.080	.031*
4	-.013	.704	29	.005	.877
5	.032	.238	30	-.029	.404
6	-.001	.978	31	-.114	.001*
7	.048	.176	32	.045	.212
8	.019	.594	33	-.029	.366
9	-.042	.117	34	.018	.600
10	.016	.595	35	-.096	.004*
11	.028	.377	36	.019	.380
12	.012	.720	37	-.007	.789
13	.047	.175	38	.031	.327
14	.037	.114	39	.001	.980
15	.030	.360	40	-.008	.780
16	-.002	.959	41	.007	.814
17	-.027	.451	42	.039	.244
18	-.089	.002*	43	.018	.572
19	-.040	.221	44	.035	.259
20	.043	.189	45	-.045	.108
21	-.083	.019*	46	.073	.008*
22	-.037	.232	47	-.033	.190
23	.016	.571	48	-.007	.825
24	-.022	.510	49	-.004	.883
25	-.056	.102	50	.056	.075

* $p < .05$

ตาราง 12 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบรวมทั้งฉบับด้วยวิธีถดถอยโลจิสติก

กลุ่มผู้สอบ (g)			ปฏิสัมพันธ์ระหว่าง กลุ่มกับความสามารถ (g by x)		
ข้อ	B	B	ข้อ	B	B
1	.5059*	-.0868*	26	.3453*	.0280
2	-.0329	.0738*	27	-.6338*	-.0199
3	.3668*	.0412	28	-.2118	.0046
4	-.1269	.0812*	29	.0573	.0570*
5	.0650	.0104	30	-.0998	-.0114
6	-.1142	-.0158	31	-.4462*	.0117
7	.1497	-.0146	32	.2412	-.0698*
8	.0220	.0643*	33	-.2006	-.0006
9	-.0290	-.0034	34	.1732	-.0661*
10	-.0216	-.0157	35	-.5438*	-.0697*
11	.1663	-.0112	36	-.0273	.0811*
12	.0436	.0409	37	.0582	-.0611*
13	-.0101	.0070	38	.0517	.0240
14	.3411	.0237	39	.1200	-.0114
15	.2388	-.0043	40	-.4081	.0110
16	-.0270	.0050	41	.1549	.0033
17	-.1067	.0413	42	.2265	.0111
18	-.4522*	-.0812*	43	.0276	.0391
19	-.0674	.0076	44	.1961	.0319
20	.2166	.0096	45	-.2326	-.0381
21	-.2747	-.0572*	46	.5082*	.3010
22	-.2215	.0003	47	-.2438	-.0277
23	.1836	.0229	48	.0282	-.0292
24	-.1026	.0169	49	-.0254	-.0445
25	-.2033	.0163	50	.2113	-.0125

* $p < .05$

ตาราง 13 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแยกตามเนื้อหาด้วยวิธีซิปเทสต์

เนื้อหา	ข้อ	β_U	p-value	เนื้อหา	ข้อ	β_U	p-value
ระบบจำนวนเต็ม	1	.3067	-.1414	เศษส่วนและ ทศนิยม	23	.007	.810
	2	-.1488	.0402		24	-.006	.863
	3	.2849	.1014		25	-.065	.059
	4	-.2682	.1915		26	.056	.123
	5	.0554	.0528		27	-.114	.000*
	6	-.2885	-.1906		28	-.085	.017*
	7	.0671	-.0582		29	.071	.030*
	8	-.0810	.0260		30	-.036	.291
	9	-.0518	-.0375		31	-.077	.023*
	10	-.1343	-.0821		32	.017	.628
	11	.0285	-.1039		33	-.044	.163
	12	-.0602	.0034		34	.037	.267
	13	-.0660	.0316		35	-.119	.000*
	14	.3153	.0210		36	-.005	.806
	15	.1575	.0286		37	-.014	.586
ความเท่ากัน ทุกประการ	16	-.010	.771	สมบัติของ รูปสามเหลี่ยม มุมฉาก	38	.027	.473
	17	.006	.894		39	.003	.929
	18	-.098	.002*		40	-.014	.701
	19	-.012	.770		41	.053	.176
	20	.045	.286		42	.019	.650
	21	-.079	.052		43	.025	.491
	22	-.037	.289		44	.065	.102
					45	-.090	.015*
					46	.078	.026*
					47	-.010	.743
					48	.012	.776
					49	-.007	.860
					50	.058	.147

* p < .05

ตาราง 14 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแยกตามเนื้อหาด้วยวิธีถดถอยโลจิสติก

เนื้อหา		กลุ่มผู้สอบ (g)		ปฏิสัมพันธ์ระหว่าง กลุ่มกับความสามารถ (g by x)		เนื้อหา		กลุ่มผู้สอบ (g)		ปฏิสัมพันธ์ระหว่าง กลุ่มกับความสามารถ (g by x)	
	ข้อ	B	B		ข้อ	B	B		ข้อ	B	B
ระบบ จำนวนเต็ม	1	.3067	-.1414	เศษส่วนและ ทศนิยม	23	.2521	.0444		23	.2521	.0444
	2	-.1488	.0402		24	.0517	.0617		24	.0517	.0617
	3	.2849	.1014		25	-.0951	.0889		25	-.0951	.0889
	4	-.2682	.1915*		26	.4715*	-.0139		26	.4715*	-.0139
	5	.0554	.0528		27	-.5718*	-.0822		27	-.5718*	-.0822
	6	-.2885	-.1906*		28	-.1597	.0015		28	-.1597	.0015
	7	.0671	-.0582		29	.1739	.1680*		29	.1739	.1680*
	8	-.0810	.0260		30	-.0109	.0148		30	-.0109	.0148
	9	-.0518	-.0375		31	-.3688*	.0625		31	-.3688*	.0625
	10	-.1343	-.0821		32	.4051*	-.1824*		32	.4051*	-.1824*
	11	.0285	-.1039		33	-.1453	-.0027		33	-.1453	-.0027
	12	-.0602	.0034		34	.3047*	.0042		34	.3047*	.0042
	13	-.0660	.0316		35	-.4441*	-.1809*		35	-.4441*	-.1809*
	14	.3153	.0210		36	.0070	.2067*		36	.0070	.2067*
	15	.1575	.0286		37	.1316	-.1152		37	.1316	-.1152
ความเท่ากัน ทุกประการ	16	.1187	.0921	สมบัติของ รูปสามเหลี่ยม มุมฉาก	38	.0230	.0323		38	.0230	.0323
	17	.0204	.2170		39	.0651	-.0126		39	.0651	-.0126
	18	-.3400	-.2349		40	-.1602	-.0384		40	-.1602	-.0384
	19	.0795	.2086		41	.0863	.0054		41	.0863	.0054
	20	.4494*	-.1366		42	.0869	-.1136		42	.0869	-.1136
	21	-.1544	-.0878		43	-.0131	.0842		43	-.0131	.0842
	22	-.0879	-.0724		44	.1425	.0796		44	.1425	.0796
					45	-.3539*	-.3132*		45	-.3539*	-.3132*
					46	.5258*	.0758		46	.5258*	.0758
					47	-.3041	.0468		47	-.3041	.0468
			48	-.0728	-.0544	48	-.0728	-.0544			
			49	-.0830	-.0176	49	-.0830	-.0176			
			50	.1538	.0384	50	.1538	.0384			

* p < .05

ตาราง 15 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแยกตามระดับพฤติกรรมด้วยวิธีซิปเทสต์

ระดับพฤติกรรม	ข้อ	β_U	p-value	ระดับพฤติกรรม	ข้อ	β_U	p-value
ความรู้หรือทักษะการคิดคำนวณ	1	.009	.727	การนำไปใช้	7	.049	.214
	16	-.031	.372		8	.036	.351
	23	.017	.578		9	-.022	.464
	24	-.118	.003*		10	-.013	.729
	25	-.067	.098		11	.044	.214
	26	.056	.186		12	.015	.690
	38	.037	.270		19	-.035	.347
	39	.009	.777		20	.067	.068
ความเข้าใจ	2	.012	.671		21	-.082	.031*
	3	.096	.001*		31	-.106	.007*
	4	.010	.738		32	.005	.895
	5	.011	.652		33	-.062	.067
	6	-.022	.406		34	.001	.970
	17	.002	.940		35	-.130	.000*
	18	-.057	.020*		36	.003	.891
	27	-.083	.001*		43	-.012	.719
	28	-.052	.099		44	.035	.289
	29	.018	.522		45	-.069	.031*
	30	-.008	.780		46	.053	.069
	40	-.014	.578		47	-.022	.434
	41	.030	.269		48	-.036	.346
	42	.059	.040*	การวิเคราะห์	13	-.032	.542
					14	.102	.004*
					15	.027	.583
					22	-.040	.354
					37	-.012	.749
					49	.014	.752
					50	.056	.217

* p < .05

ตาราง 16 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแยกตามระดับพฤติกรรมด้วยวิธีถดถอยโลจิสติก

ระดับ พฤติกรรม	กลุ่มผู้สอบ (g)	ปฏิสัมพันธ์ระหว่าง กลุ่มกับความสามารถ (g by x)	ระดับ พฤติกรรม	กลุ่มผู้สอบ (g)	ปฏิสัมพันธ์ระหว่าง กลุ่มกับความสามารถ (g by x)
	ข้อ	B		ข้อ	B
ความรู้ หรือทักษะ การคิด คำนวณ	1	.2373	การนำไปใช้	7	.1759
	16	-.1411		8	.0393
	23	.1536		9	-.0024
	24	-.2404		10	-.0107
	25	-.3546*		11	.2084
	26	.3347*		12	.0672
	38	.0404		19	-.0444
	39	.0843		20	.2259
ความเข้าใจ	2	.0079		21	-.2688
	3	.4835*		31	-.4499*
	4	-.0421		32	.2376
	5	.1161		33	-.1828
	6	-.0968		34	.2052
	17	-.0378		35	-.5516*
	18	-.4294*		36	-.0004
	27	-.6287*		43	.0476
	28	-.1816		44	.2091
	29	.1102		45	-.2188
	30	-.0482		46	.5161*
	40	-.0128		47	-.2309
	41	.2085		48	.0502
	42	.3154*	การวิเคราะห์	13	-.1098
				14	.3776
				15	.1455
				22	-.3360
				37	-.0291
				49	-.0732
				50	.2015

* p < .05

ภาคผนวก ง

ตัวอย่างผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีชิปทดสอบและวิธีถดถอยโลจิสติก

ตัวอย่างผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีชิปทดสอบ

***** SIMULTANEOUS ITEM BIAS ESTIMATION and HYPOTHESIS TESTING *****

R. Shealy and W. Stout, programming assistance by L. Roussos

```
name of input parameter file = inma_a4.dat
number of items on test = 13
name of file for Ref. grp. scores = mara4.dat
name of file for Focal grp. scores = mafa4.dat
minimum no. of examinees per statistic calculation cell = 2
estimate of guessing on the test = .20
number of runs for this data set = 13
number of examinees in Reference group = 480
number of examinees in Focal group = 514
```

p-value notation:

R denotes p-value for test of bias/DIF/DTF against Ref. group

F denotes p-value for test of bias/DIF/DTF against Foc. group

E denotes p-value for test of bias/DIF/DTF against either the
Ref. or Foc. group.

NOTE: M-H Chi-square p-value is restricted, by definition, to type E.

run no.				Mantel-Haenszel			
	Beta-uni	SIB-uni z-statistic	SIB-uni p-value	Chi sqr.	p value	Delta (D-DIF)	MH alpha
1	.027	.717	.473 E	.01	.928 E	.06	.97
2	.003	.089	.929 E	.01	.942 E	-.06	1.03
3	-.014	-.385	.701 E	.53	.466 E	.34	.86
4	.053	1.353	.176 E	.23	.628 E	-.22	1.10
5	.019	.453	.650 E	.19	.664 E	-.18	1.08
6	.025	.688	.491 E	.03	.853 E	.10	.96
7	.065	1.634	.102 E	.61	.434 E	-.34	1.15
8	-.090	-2.430	.015 E	3.54	.060 E	.79	.71
9	.078	2.228	.026 E	7.23	.007 E	-1.12	1.61
10	-.010	-.328	.743 E	1.99	.158 E	.69	.75
11	.012	.284	.776 E	.21	.646 E	.19	.92
12	-.007	-.176	.860 E	.26	.612 E	.22	.91
13	.058	1.449	.147 E	.48	.488 E	-.30	1.13

Program execution is completed.

Your output is stored on the file: outma_a4.dat

ตัวอย่างผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีถดถอยโลจิสติก

```
data list file = 'c:\maths.dat'/
      g 1 i1 to i50 3-52.
variable lable g 'group'.
value lable g 1 'focal' 2 'reference'.
count x = i1 to i50(1).
logistic regression i1 with g x g by x
      /method enter x
      /method enter g
      /method enter g by x.
```

Logistic Regression

```
Total number of cases:      994 (Unweighted)
Number of selected cases:    994
Number of unselected cases:  0

Number of selected cases:      994
Number rejected because of missing data:  0
Number of cases included in the analysis:  994
```

Dependent Variable Encoding:

Original Value	Internal Value
0	0
1	1

Interactions:

```
INT_1      G by X
```

```
Dependent Variable..  I1
```

```
Beginning Block Number  0.  Initial Log Likelihood Function
-2 Log Likelihood      716.27322
* Constant is included in the model.
```

```
Beginning Block Number  1.  Method: Enter
Variable(s) Entered on Step Number
1..      X
```

```
Estimation terminated at iteration number 5 because
Log Likelihood decreased by less than .01 percent.
```

-2 Log Likelihood	582.117
Goodness of Fit	1048.451
Cox & Snell - R ²	.126
Nagelkerke - R ²	.246

	Chi-Square	df	Significance
Model	134.156	1	.0000
Block	134.156	1	.0000
Step	134.156	1	.0000

Classification Table for I1
The Cut Value is .50

		Predicted		Percent Correct
		0	1	
Observed	0	868	10	98.86%
	1	96	20	17.24%
Overall				89.34%

----- Variables in the Equation -----

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
X	.1729	.0163	112.3993	1	.0000	.3926	1.1887
Constant	-4.9056	.3251	227.6219	1	.0000		

Beginning Block Number 2. Method: Enter
Variable(s) Entered on Step Number
1.. G group

Estimation terminated at iteration number 5 because
Log Likelihood decreased by less than .01 percent.

-2 Log Likelihood 576.861
Goodness of Fit 1031.191
Cox & Snell - R² .131
Nagelkerke - R² .255

	Chi-Square	df	Significance
Model	139.412	2	.0000
Block	5.255	1	.0219
Step	5.255	1	.0219

Classification Table for I1
The Cut Value is .50

		Predicted		Percent Correct
		0	1	
Observed	0	867	11	98.75%
	1	96	20	17.24%
Overall				89.24%

----- Variables in the Equation -----

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
X	.1780	.0167	113.9267	1	.0000	.4385	1.1948
G	.5059	.2224	5.1737	1	.0229	.0738	1.6584
Constant	-5.7601	.5144	125.3790	1	.0000		

Beginning Block Number 3. Method: Enter
 Variable(s) Entered on Step Number
 1.. G * X
 Estimation terminated at iteration number 5 because
 Log Likelihood decreased by less than .01 percent.

-2 Log Likelihood 570.299
 Goodness of Fit 1137.741
 Cox & Snell - R² .137
 Nagelkerke - R² .266

	Chi-Square	df	Significance
Model	145.974	3	.0000
Block	6.563	1	.0104
Step	6.563	1	.0104

Classification Table for I1
 The Cut Value is .50

		Predicted		Percent Correct
		0	1	
Observed	0	869	9	98.97%
	1	94	22	18.97%
Overall				89.64%

----- Variables in the Equation -----

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
X	.3102	.0569	29.7044	1	.0000	.2191	1.3637
G	2.1585	.7039	9.4027	1	.0022	.1133	8.6579
G by X	-.0868	.0344	6.3559	1	.0117	-.0869	.9168
Constant	-8.3452	1.2178	46.9608	1	.0000		

ประวัติย่อผู้วิจัย

ประวัติย่อผู้วิจัย

ชื่อ ชื่อสกุล	นางสาวอารี วัชรโสติกุล
วันเดือนปีเกิด	10 ตุลาคม 2505
สถานที่เกิด	อำเภอบางบ่อ จังหวัดสมุทรปราการ
สถานที่อยู่ปัจจุบัน	2420/2 ถนนเจริญกรุง แขวงบางค้อแหลม เขตบางค้อแหลม กรุงเทพมหานคร 10120
สถานที่ทำงานปัจจุบัน	โรงเรียนพระหฤทัยคอนแวนต์ เขตคลองเตย กรุงเทพมหานคร
ประวัติการศึกษา	
พ.ศ. 2523	มัธยมศึกษาตอนปลาย (วิทย์-คณิต) จาก โรงเรียนบางบ่อวิทยาคม
พ.ศ. 2527	ค.บ. (คณิตศาสตร์) จาก วิทยาลัยครูจันทระเกษม
พ.ศ. 2543	กศ.ม. (การวัดผลการศึกษา) จาก มหาวิทยาลัยศรีนครินทรวิโรฒ