

การศึกษาคูณลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์เมื่อจำนวนผู้ตรวจ
และรูปแบบการตรวจให้คะแนนต่างกัน โดยใช้โมเดลการสรุปอ้างอิง
และโมเดลหลายองค์ประกอบของราส์ช



เสนอต่อบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ เพื่อเป็นส่วนหนึ่งของการศึกษา
ตามหลักสูตรปริญญาการศึกษาดุริยบัณฑิต สาขาวิชาการทดสอบและวัดผลการศึกษา

พฤษภาคม 2554

การศึกษาคูณลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์เมื่อจำนวนผู้ตรวจ
และรูปแบบการตรวจให้คะแนนต่างกัน โดยใช้โมเดลการสรุปอ้างอิง
และโมเดลหลายองค์ประกอบของราส์ซ



เสนอต่อบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ เพื่อเป็นส่วนหนึ่งของการศึกษา
ตามหลักสูตรปริญญาการศึกษาดุฎบัณฑิต สาขาวิชาการทดสอบและวัดผลการศึกษา

พฤษภาคม 2554

ลิขสิทธิ์เป็นของมหาวิทยาลัยศรีนครินทรวิโรฒ

การศึกษาคูณลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์เมื่อจำนวนผู้ตรวจ
และรูปแบบการตรวจให้คะแนนต่างกัน โดยใช้โมเดลการสรุปอ้างอิง
และโมเดลหลายองค์ประกอบของราส์ช



เสนอต่อบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ เพื่อเป็นส่วนหนึ่งของการศึกษา
ตามหลักสูตรปริญญาการศึกษาดุขบัณฑิต สาขาวิชาการทดสอบและวัดผลการศึกษา
พฤษภาคม 2554

น้ำผึ้ง อินทะเนตร. (2554). การศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิดวิชา

คณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน โดยใช้

โมเดลการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราล์ซ. ปริญญาโท กศ.ด.

(การทดสอบและวัดผลการศึกษา). กรุงเทพฯ: บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ.

คณะกรรมการควบคุม: รองศาสตราจารย์ ดร.องอาจ นัยพัฒน์, รองศาสตราจารย์ ดร.

ผจงจิต อินทสุวรรณ, รองศาสตราจารย์ ดร.สุทธิวรรณ พิรศักดิ์โสภณ

การวิจัยในครั้งนี้เป็นการศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ โดยใช้โมเดลการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราล์ซ ภายใต้เงื่อนไขจำนวนผู้ตรวจต่างกันสามลักษณะ คือ 2 คน 3 คน และ 4 คน และรูปแบบการตรวจให้คะแนนต่างกันสามลักษณะ คือ รูปแบบที่ 1 ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน รูปแบบที่ 2 ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน และรูปแบบที่ 3 ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน คุณลักษณะของคะแนนพิจารณาจากขนาดขององค์ประกอบความแปรปรวน ค่าความเชื่อมั่นและค่าความเที่ยงตรงตามสภาพ

เครื่องมือที่ใช้ในการวิจัยครั้งนี้คือแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ ระดับมัธยมศึกษาตอนต้น ตามหลักสูตรการศึกษาขั้นพื้นฐาน พุทธศักราช 2551 จำนวน 12 ข้อ กลุ่มตัวอย่างที่ใช้คือ นักเรียนระดับชั้นมัธยมศึกษาปีที่ 4 จากโรงเรียนในสังกัดสำนักงานเขตพื้นที่การศึกษาจังหวัดน่าน ปีการศึกษา 2552 จำนวน 180 คน ซึ่งได้มาจากการสุ่มแบบสองขั้นตอน

ผลการวิจัยสรุปได้ว่า

1. เมื่อวิเคราะห์ด้วยโมเดลการสรุปอ้างอิง พบว่า เมื่อใช้รูปแบบการตรวจให้คะแนนเดียวกันในทุกเงื่อนไขจำนวนผู้ตรวจ ความแปรปรวนขององค์ประกอบเดียวกันมีค่าใกล้เคียงกัน และค่าสัมประสิทธิ์การสรุปอ้างอิงในรูปแบบการตรวจที่ 2 มีค่าสูงสุด รองลงมาคือรูปแบบการตรวจที่ 1 และรูปแบบการตรวจที่ 3 มีค่าต่ำสุด ค่าสัมประสิทธิ์การสรุปอ้างอิงในรูปแบบการตรวจที่ 1 มีค่าสูงขึ้นเมื่อจำนวนผู้ตรวจเพิ่มขึ้น คะแนนในทุกเงื่อนไขที่ต่างกันมีความเที่ยงตรงตามสภาพสูงและแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

2. เมื่อวิเคราะห์ด้วยโมเดลหลายองค์ประกอบของราล์ซ พบว่า ในทุกเงื่อนไขจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนน ความแปรปรวนของผู้สอบมีค่าสูงที่สุด รองลงมาคือ ความแปรปรวนของข้อสอบ และความแปรปรวนของผู้ตรวจมีค่าต่ำที่สุด และค่าความเชื่อมั่นแยกส่วนของผู้สอบของรูปแบบการตรวจที่ 1 มีค่าสูงสุดในทุกเงื่อนไขจำนวนผู้ตรวจ ค่าความเชื่อมั่นแยกส่วนของผู้สอบของรูปแบบการตรวจที่ 1 มีค่าสูงขึ้นเมื่อจำนวนผู้ตรวจเพิ่มขึ้น คะแนนความสามารถในทุกเงื่อนไขที่ต่างกันมีความเที่ยงตรงตามสภาพสูงและแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

CHARACTERISTICS OF THE OPEN-ENDED MATHEMATICS TEST SCORES FOR
DIFFERENT NUMBERS OF RATERS AND SCORING PATTERNS USING
GENERALIZABILITY MODEL AND MANY- FACET RASCH MODEL

AN ABSTRACT
BY
NAMPUENG INTANATE



Presented in Partial Fulfillment of the Requirements for the
Doctor of Education Degree in Testing and Measurement
at Srinakharinwirot University

May 2011

Nampueng Intanate. (2011). **Characteristics of the Open-Ended Mathematics Test Scores for Different Numbers of Raters and Scoring Patterns Using Generalizability Model and Many-Facet Rasch Model**. Dissertation, Ed.D. (Testing and Measurement). Bangkok: Graduate School, Srinakharinwirot University. Advisor Committee: Assoc. Prof. Dr. Ong-Art Naiyapatana, Assoc. Prof. Dr. Pachongchit Intasuwan, Assoc. Prof. Dr. Suthiwan Pirasaksopon.

The purpose of this research was to study the characteristics of the open-ended mathematics test scores analyzed by using Generalizability Model and Many-Facet Rasch Model under different conditions of three numbers of raters (2, 3 and 4 raters) and three scoring patterns. The scoring patterns were 1) the rater rated all items of all students, 2) the rater rated all items of some students and 3) the rater rated some items of all students. The score characteristics were considered from the magnitude of variance components, the reliability, and the concurrent validity.

The research tool was the 12-item open-ended mathematics test for lower secondary level, according to the 2008 Basic Education Curriculum. The sample consisted of 180 Mathayomsuksa 4 students in the schools attached to Nan Educational Service Area Office in 2009 academic year, and was selected by two-stage random sampling.

The results of the study were as follows:

1. Analyzed by Generalizability Model, when used the same scoring patterns in all conditions of numbers of raters, there were similar magnitudes of variances in the same components. The generalizability coefficient obtained from the 2nd pattern was maximum, followed by the 1st pattern, and the 3rd pattern was minimum. The generalizability coefficient obtained from the 1st pattern was higher when the numbers of raters increased. The values of concurrent validity of scores in all different conditions were high but not significantly different.

2. Analyzed by Many-Facet Rasch Model, in all conditions of numbers of raters and scoring patterns, the examinees variance had the maximum value, followed by items and the raters variance which had the minimum value. The person separation reliability of the 1st pattern had the maximum value in all conditions of numbers of raters. The person separation reliability of the 1st pattern was higher when the numbers of raters increased. The values of concurrent validity of measured scores in all different conditions were high but not significantly different.



งานวิจัยนี้ได้รับทุนอุดหนุนการวิจัย
จาก
มหาวิทยาลัยศรีนครินทรวิโรฒ
ประเภททุนงบประมาณแผ่นดิน ประจำปีงบประมาณ 2552

ปริญญานิพนธ์

เรื่อง

การศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์เมื่อจำนวนผู้ตรวจ
และรูปแบบการตรวจให้คะแนนต่างกัน โดยใช้โมเดลการสุปอ้างอิง
และโมเดลหลายองค์ประกอบของราล์ช

ของ

น้ำผึ้ง อินทะเนตร

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาการศึกษาดุสิตบัณฑิต สาขาวิชาการทดสอบและวัดผลการศึกษา
ของมหาวิทยาลัยศรีนครินทรวิโรฒ

.....คนบดีบัณฑิตวิทยาลัย

(รองศาสตราจารย์ ดร. สมชาย สันติวัฒนกุล)

วันที่ เดือน..... พ.ศ. 2554

คณะกรรมการควบคุมปริญญานิพนธ์

คณะกรรมการสอบปากเปล่า

..... ประธาน

..... ประธาน

(รองศาสตราจารย์ ดร.องอาจ นัยพัฒน์)

(ผู้ช่วยศาสตราจารย์ ดร.สุนันท์ ศลโกสม)

..... กรรมการ

..... กรรมการ

(รองศาสตราจารย์ ดร.ผจจจิต อินทสุวรรณ)

(รองศาสตราจารย์ ดร.องอาจ นัยพัฒน์)

..... กรรมการ

..... กรรมการ

(รองศาสตราจารย์ ดร.สุทธีวรรณ พิรศักดิ์โสภณ)

(รองศาสตราจารย์ ดร.ผจจจิต อินทสุวรรณ)

..... กรรมการ

(รองศาสตราจารย์ ดร.สุทธีวรรณ พิรศักดิ์โสภณ)

..... กรรมการ

(รองศาสตราจารย์ ดร.สมถวิล วิจิตรวรรณา)

ประกาศคุณประการ

ปริญญานิพนธ์นี้สำเร็จได้ด้วยความกรุณาอย่างสูงจากคณะกรรมการควบคุมปริญญานิพนธ์ รองศาสตราจารย์ ดร.องอาจ นัยพัฒน์ ประธานควบคุมปริญญานิพนธ์ รองศาสตราจารย์ ดร.ผจงจิต อินทสุวรรณ และรองศาสตราจารย์ ดร.สุทธีวรรณ พีรศักดิ์โสภณ กรรมการควบคุมปริญญานิพนธ์ที่ได้ให้คำปรึกษา ข้อชี้แนะในการจัดทำงานวิจัยนี้ทุกขั้นตอน และแก้ไขข้อบกพร่องต่าง ๆ ด้วยความเอาใจใส่ ตลอดจนให้กำลังใจเสมอมา ผู้วิจัยขอกราบขอบพระคุณในความกรุณาของท่านทั้งสามด้วยความเคารพอย่างสูง

ขอกราบขอบพระคุณกรรมการที่แต่งตั้งเพิ่มเติมทั้งสองท่านคือ ผู้ช่วยศาสตราจารย์ ดร.สุนันท์ ศลโกสุมและรองศาสตราจารย์ ดร. สมถวิล วิจิตรวรรณมา ที่ให้ความกรุณาเป็นกรรมการในการสอบปริญญานิพนธ์และให้ข้อเสนอแนะที่เป็นประโยชน์เพิ่มเติมทำให้ปริญญานิพนธ์ฉบับนี้สมบูรณ์ยิ่งขึ้น

ขอกราบขอบพระคุณคณาจารย์ภาควิชาการวัดผลและวิจัยการศึกษา มหาวิทยาลัยศรีนครินทรวิโรฒทุกท่านที่ประสิทธิ์ประสาทวิชาความรู้แก่ผู้วิจัย พร้อมทั้งให้คำปรึกษาและเอาใจใส่อย่างดีเยี่ยมตลอดมา

ขอกราบขอบพระคุณอธิการบดี คณบดีคณะศึกษาศาสตร์ มหาวิทยาลัยเชียงใหม่ที่สนับสนุนทุนโครงการพัฒนาอาจารย์ มหาวิทยาลัยเชียงใหม่ ตลอดจนคณาจารย์ สายวิชาประเมินผลและวิจัยการศึกษา สายวิชาคณิตศาสตร์ คณะศึกษาศาสตร์ มหาวิทยาลัยเชียงใหม่ทุกท่านที่ให้กำลังใจและสนับสนุนในการศึกษาต่ออย่างดียิ่ง

ขอขอบพระคุณผู้เชี่ยวชาญทุกท่านที่ให้ความช่วยเหลือตรวจสอบคุณภาพของเครื่องมือในการวิจัย ขอขอบพระคุณผู้บริหาร คณาจารย์ที่เสียสละเวลาในการตรวจให้คะแนนแบบทดสอบซึ่งเป็นส่วนสำคัญส่วนหนึ่งสำหรับการวิจัย และขอบใจนักเรียนในเขตพื้นที่การศึกษาจังหวัดน่าน ที่ให้ความร่วมมือในการเก็บรวบรวมข้อมูลงานวิจัยอย่างดียิ่ง

ท้ายสุดผู้วิจัยขอกราบขอบพระคุณบิดา มารดา ขอบคุณพี่สาว น้องสาวและเพื่อนๆ ทุกคนที่ให้กำลังใจที่ดีเยี่ยมตลอดระยะเวลาที่ศึกษาและทำวิจัย จนทำให้ผู้วิจัยประสบความสำเร็จในการศึกษา

น้ำผึ้ง อินทะเนตร

สารบัญ

บทที่	หน้า
1 บทนำ.....	1
ภูมิหลัง.....	1
ความมุ่งหมายของการวิจัย.....	9
ความสำคัญของการวิจัย.....	10
ขอบเขตของการวิจัย.....	11
นิยามศัพท์เฉพาะ.....	12
กรอบแนวคิดในการวิจัย.....	14
สมมติฐานของการวิจัย.....	19
2 เอกสารและงานวิจัยที่เกี่ยวข้อง.....	21
คำถามปลายเปิดวิชาคณิตศาสตร์.....	21
กฎเกณฑ์การให้คะแนน.....	32
ทฤษฎีการสรุปอ้างอิง.....	38
โมเดลหลายองค์ประกอบของราล์ฟ.....	60
การทดสอบทางการศึกษาระดับชาตินี้พื้นฐาน.....	75
งานวิจัยที่เกี่ยวข้อง.....	79
3 วิธีดำเนินการวิจัย.....	101
การกำหนดประชากรและการเลือกกลุ่มตัวอย่าง.....	101
การสร้างเครื่องมือที่ใช้ในงานวิจัย.....	102
การเก็บรวบรวมข้อมูล.....	110
การจัดกระทำข้อมูลและการวิเคราะห์ข้อมูล.....	111
4 ผลการวิเคราะห์ข้อมูล.....	120
การวิเคราะห์สถิติพื้นฐานของคะแนนสอบ	121
การศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิด โดยใช้โมเดลการสรุปอ้างอิง.....	127

สารบัญ (ต่อ)

บทที่	หน้า
4 (ต่อ)	
การศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิด โดยใช้โมเดลการหลายองค์ประกอบของราสส์.....	150
5 สรุป อภิปรายและข้อเสนอแนะ.....	176
สรุปผลการวิจัย.....	179
อภิปรายผล.....	182
ข้อเสนอแนะ.....	196
บรรณานุกรม	199
ภาคผนวก.....	212
ภาคผนวก ก	213
ภาคผนวก ข	226
ภาคผนวก ค	240
ประวัติย่อผู้วิจัย.....	261

บัญชีตาราง

ตาราง	หน้า
1 การเปลี่ยนคำถามปลายปิดให้เป็นคำถามปลายเปิด	27
2 ตัวอย่างกฎเกณฑ์การให้คะแนนแบบแยกองค์ประกอบ	34
3 ตัวอย่างกฎเกณฑ์การให้คะแนนแบบภาพรวม	36
4 ANOVA สำหรับการวัดที่มีองค์ประกอบเดียวแบบไขว้ $p \times i$	51
5 ANOVA สำหรับการวัดที่มีสององค์ประกอบแบบไขว้ $p \times i \times r$	52
6 จำนวนนักเรียนชั้นมัธยมศึกษาปีที่ 4 ในแต่ละโรงเรียนที่เป็นกลุ่มตัวอย่างการวิจัย..	102
7 การวิเคราะห์หลักสูตรกลุ่มสาระการเรียนรู้คณิตศาสตร์ ชั้นมัธยมศึกษาตอนต้น ...	215
8 ค่าสถิติพื้นฐานของคะแนนสอบ	122
9 ค่าพารามิเตอร์ขององค์ประกอบ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้ คะแนนต่างกัน.....	125
10 ค่าความแปรปรวนขององค์ประกอบ เมื่อมีผู้ตรวจจำนวน 4 คน และตรวจข้อสอบ ทุกข้อของผู้สอบทุกคน	128
11 ค่าความแปรปรวนขององค์ประกอบ เมื่อมีผู้ตรวจจำนวน 4 คน และตรวจข้อสอบ ทุกข้อของผู้สอบบางคน	129
12 ค่าความแปรปรวนขององค์ประกอบ เมื่อมีผู้ตรวจจำนวน 4 คนตรวจข้อสอบบาง ข้อของผู้สอบทุกคน	130
13 ค่าความแปรปรวนขององค์ประกอบ เมื่อมีผู้ตรวจจำนวน 3 คน และตรวจข้อสอบ ทุกข้อของผู้สอบทุกคน.....	131
14 ค่าความแปรปรวนขององค์ประกอบ เมื่อมีผู้ตรวจจำนวน 3 คน และตรวจข้อสอบ ทุกข้อของผู้สอบบางคน	132
15 ค่าความแปรปรวนขององค์ประกอบ เมื่อมีผู้ตรวจจำนวน 3 คน ตรวจข้อสอบบาง ข้อของผู้สอบทุกคน	133
16 ค่าความแปรปรวนขององค์ประกอบ เมื่อมีผู้ตรวจจำนวน 2 คน และตรวจข้อสอบ ทุกข้อของผู้สอบทุกคน.....	134
17 ค่าความแปรปรวนขององค์ประกอบ เมื่อมีผู้ตรวจจำนวน 2 คน และตรวจข้อสอบ ทุกข้อของผู้สอบบางคน	135

บัญชีตาราง (ต่อ)

ตาราง	หน้า
18 ค่าความแปรปรวนขององค์ประกอบ เมื่อมีผู้ตรวจจำนวน 2 คน และตรวจข้อสอบ บางข้อของผู้สอบทุกคน	136
19 ขนาดความแปรปรวนของแต่ละองค์ประกอบ เมื่อรูปแบบการตรวจให้คะแนน เหมือนกันและจำนวนผู้ตรวจต่างกัน	137
20 ค่าสัมประสิทธิ์การสรุปอ้างอิงเมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้ คะแนนต่างกัน.....	140
21 การทดสอบความแตกต่างรายคู่ของค่าสัมประสิทธิ์การสรุปอ้างอิงเมื่อจำนวน ผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน	143
22 ค่าสัมประสิทธิ์การสรุปอ้างอิงเมื่อรูปแบบการตรวจให้คะแนนเหมือนกันและ จำนวนผู้ตรวจต่างกัน	144
23 การทดสอบความแตกต่างรายคู่ของค่าสัมประสิทธิ์การสรุปอ้างอิงเมื่อรูปแบบ การตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจต่างกัน.....	146
24 ค่าสัมประสิทธิ์สหสัมพันธ์เมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนน ต่างกัน.....	148
25 ค่าสัมประสิทธิ์สหสัมพันธ์เมื่อรูปแบบการตรวจให้คะแนนเหมือนกันและจำนวน ผู้ตรวจต่างกัน	149
26 ขนาดความแปรปรวนของแต่ละองค์ประกอบ เมื่อจำนวนผู้ตรวจและรูปแบบ การตรวจให้คะแนนต่างกัน.....	150
27 ค่าความเชื่อมั่นแยกส่วนของผู้สอบเมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจ ให้คะแนนต่างกัน.....	154
28 การทดสอบความแตกต่างรายคู่ของค่าความเชื่อมั่นแยกส่วนของผู้สอบเมื่อจำนวน ผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน.....	156
29 ค่าความเชื่อมั่นแยกส่วนของผู้ตรวจเมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจ ให้คะแนนต่างกัน.....	157
30 การทดสอบความแตกต่างรายคู่ของค่าความเชื่อมั่นแยกส่วนของผู้ตรวจเมื่อ จำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน.....	159

บัญชีตาราง (ต่อ)

ตาราง	หน้า
31 ค่าความเชื่อมั่นแยกส่วนของข้อสอบเมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน	160
32 การทดสอบความแตกต่างรายคู่ของค่าความเชื่อมั่นแยกส่วนของข้อสอบเมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน.....	162
33 ค่าความเชื่อมั่นแยกส่วนของข้อสอบเมื่อรูปแบบการตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจต่างกัน.....	164
34 การทดสอบความแตกต่างรายคู่ของค่าความเชื่อมั่นแยกส่วนของข้อสอบเมื่อผู้ตรวจตรวจสอบข้อสอบทุกข้อของผู้สอบทุกคนและจำนวนผู้ตรวจต่างกัน.....	166
35 ค่าความเชื่อมั่นแยกส่วนของข้อสอบเมื่อรูปแบบการตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจต่างกัน.....	167
36 การทดสอบความแตกต่างรายคู่ของค่าความเชื่อมั่นแยกส่วนของข้อสอบเมื่อรูปแบบการตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจต่างกัน	169
37 ค่าความเชื่อมั่นแยกส่วนของข้อสอบเมื่อรูปแบบการตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจต่างกัน.....	170
38 การทดสอบความแตกต่างรายคู่ของค่าความเชื่อมั่นแยกส่วนของข้อสอบเมื่อผู้ตรวจตรวจสอบข้อสอบบางข้อของผู้สอบทุกคน และจำนวนผู้ตรวจต่างกัน.....	172
39 ค่าสัมประสิทธิ์สหสัมพันธ์เมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน.....	174
40 ค่าสัมประสิทธิ์สหสัมพันธ์เมื่อรูปแบบการตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจต่างกัน.....	175

บัญชีภาพประกอบ

ภาพประกอบ	หน้า
1 แหล่งของความแปรปรวนและองค์ประกอบของความแปรปรวน สำหรับหนึ่ง องค์ประกอบแบบไขว้ $p \times i$ (One Facet, $p \times i$ Design)	46
2 แหล่งของความแปรปรวนและองค์ประกอบของความแปรปรวนสำหรับสอง องค์ประกอบแบบไขว้ $p \times r \times o$ (Two – facet, $p \times r \times o$ Design)	47
3 แหล่งของความแปรปรวนและองค์ประกอบของความแปรปรวนสำหรับหนึ่ง องค์ประกอบแบบแฝง เมื่อ i แฝงใน p (One Facet, $i : p$ Design)	49
4 แหล่งของความแปรปรวนและองค์ประกอบของความแปรปรวนสำหรับสอง องค์ประกอบแบบแฝง กรณี $i : r : p$ (Two Facets, $i : r : p$ Design)	50
5 แหล่งของความแปรปรวนและองค์ประกอบของความแปรปรวนสำหรับสอง องค์ประกอบแบบผสมกรณี $(o : p) \times r$ (Two – Facet, $(o : p) \times r$ Design) ...	51
6 แหล่งความคลาดเคลื่อนแบบสัมบูรณ์และความคลาดเคลื่อนแบบสัมพัทธ์ สำหรับ $p \times i$ Design	54
7 แหล่งความคลาดเคลื่อนแบบสัมบูรณ์และความคลาดเคลื่อนแบบสัมพัทธ์ สำหรับ $p \times r \times o$ Design	55
8 แหล่งความคลาดเคลื่อนแบบสัมบูรณ์และความคลาดเคลื่อนแบบสัมพัทธ์ สำหรับ $i : p$ Design	56
9 แหล่งความคลาดเคลื่อนแบบสัมบูรณ์และความคลาดเคลื่อนแบบสัมพัทธ์ สำหรับ $(o : p) \times r$ Design	56
10 ขั้นตอนการสร้างแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ ชั้นมัธยมศึกษาตอนต้น	103
11 การจัดกระทำข้อมูลตามเงื่อนไขและระดับของเงื่อนไขที่ศึกษา	114
12 ร้อยละของความแปรปรวนของแต่ละองค์ประกอบเมื่อรูปแบบการตรวจให้คะแนน เหมือนกัน และจำนวนผู้ตรวจต่างกัน	138
13 ค่าสัมประสิทธิ์การสรุปอ้างอิง เมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจ ให้คะแนนต่างกัน	141
14 ค่าสัมประสิทธิ์การสรุปอ้างอิง เมื่อรูปแบบการตรวจให้คะแนนเหมือนกันและ จำนวนผู้ตรวจต่างกัน	145

บัญชีภาพประกอบ (ต่อ)

ภาพประกอบ	หน้า
15 ร้อยละของความแปรปรวนของแต่ละองค์ประกอบเมื่อรูปแบบการตรวจให้คะแนน ต่างกันและจำนวนผู้ตรวจเท่ากัน	151
16 ค่าความเชื่อมั่นแยกส่วนของผู้สอบ เมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบ การตรวจให้คะแนนต่างกัน.....	154
17 ค่าความเชื่อมั่นแยกส่วนของผู้ตรวจ เมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบ การตรวจให้คะแนนต่างกัน	158
18 ค่าความเชื่อมั่นแยกส่วนของข้อสอบ เมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบ การตรวจให้คะแนนต่างกัน	161
19 ค่าความเชื่อมั่นแยกส่วนของผู้สอบ เมื่อรูปแบบการตรวจให้คะแนนเหมือนกัน และจำนวนผู้ตรวจต่างกัน.....	165
20 ค่าความเชื่อมั่นแยกส่วนของผู้ตรวจ เมื่อรูปแบบการตรวจให้คะแนนเหมือนกัน และจำนวนผู้ตรวจต่างกัน.....	168
21 ค่าความเชื่อมั่นแยกส่วนของข้อสอบ เมื่อรูปแบบการตรวจให้คะแนนเหมือนกัน และจำนวนผู้ตรวจต่างกัน	171

บทที่ 1

บทนำ

ภูมิหลัง

แบบทดสอบปลายเปิด (Open-Ended Test) เป็นเครื่องมือที่ในปัจจุบันนิยมนำมาใช้วัดประเมิน (Assessment) เพิ่มมากขึ้น ทั้งนี้นอกจากจะเหมาะสมในการวัดความเข้าใจ กระบวนการคิดในระดับสูง ความสามารถในการเขียนและทักษะในการแก้ปัญหา (Hopkins; & Antes. 1990: 151) ได้เป็นอย่างดีแล้ว ยังเปิดโอกาสให้นักเรียนแสดงออกได้อย่างหลากหลายทั้งคำตอบและวิธีการในการแก้ปัญหา คำถามปลายเปิดจะกระตุ้นให้นักเรียนสนใจในสถานการณ์ที่กำหนดให้และส่งเสริมให้นักเรียนที่มีความสามารถต่างกันสามารถแก้ปัญหาได้ด้วยตนเอง โดยการพัฒนาวิธีการหาคำตอบและการสื่อสารวิธีการแก้ปัญหาของตนเอง (NCTM. 1995; citing Pandey. 1991) คำถามปลายเปิดจึงไม่ต้องการเพียงคำตอบที่ได้เท่านั้น แต่ยังให้นักเรียนแสดงวิธีการได้มาซึ่งคำตอบนั้นด้วย จึงเป็นการเปิดโอกาสให้นักเรียนได้แสดงทักษะการคิดคำนวณ การให้เหตุผล การสื่อสารและการแก้ปัญหาได้อย่างหลากหลาย

แม้ว่าแบบทดสอบปลายเปิด จะเป็นแบบทดสอบที่มีคุณค่ามากชนิดหนึ่ง แต่ยังมีปัญหาอยู่หลายประการ เช่น แบบทดสอบปลายเปิดมักจะมีค่าความเชื่อมั่นต่ำ สิ้นเปลืองเวลา แรงงานและค่าใช้จ่ายในการตรวจมาก (Coffman. 1971; Chui; & Wolfe. 2002: 322; Bock; & Ockey. 2003: online) มีจุดอ่อนสำคัญอยู่ที่การให้คะแนน (Striggins. 1994: 154-155; บุญเชิด ภิญญอนันตพงษ์. 2547: ค) การตรวจยังมีความเป็นอัตนัย (Subjective) คะแนนที่ได้จึงแปรเปลี่ยนไปตามลักษณะของผู้ตรวจ (Lunz; Wright; & Linacre. 1990: online; Bachman; Lynch; & Mason. 1993: online; Linacre. 1993: 2; Turner. 2003: online; Smith; & Kulikowich. 2004: 619) นั่นคือ มีความคลาดเคลื่อนอันเนื่องมาจากการตรวจ นอกจากนี้ยังมีอิทธิพลแทรกซ้อนอีกหลายปัจจัยที่มีผลต่อการตรวจให้คะแนน เช่น ลักษณะข้อสอบ จำนวนครั้งของการตรวจ ความรู้พื้นฐานของนักเรียนและปฏิสัมพันธ์ของปัจจัยต่าง ๆ เหล่านี้ (Linacre. 1993: 2; Lee; & Kantor. 2005: 1; Sudweeks; Reeve; & Bradshaw. 2005: 240) คะแนนที่ได้จึงไม่ใช่ค่าที่แท้จริงของสิ่งที่มุ่งวัด (Lord; & Novick. 1968) ทำให้มีข้อสงสัยในความเชื่อมั่นของคะแนนที่ได้จากการวัดประเมินด้วยคำถามปลายเปิดอยู่เสมอ (Swartz; et al. 1999: 1) ซึ่งส่งผลต่อความเที่ยงตรงของการแปลความหมายของคะแนนและความยุติธรรมในการตัดสินตามมา (Messick. 1989; Linacre. 1993: 3; Nitko. 1996)

นักวัดผลทางการศึกษาหลายท่านได้ให้ข้อเสนอแนะวิธีการปรับปรุงเพื่อให้ผลการวัดประเมินคำถามปลายเปิดมีความถูกต้องแม่นยำและน่าเชื่อถือมากขึ้นไว้ เช่น เมอเรนส์และลีแมน (Mehrens;

& Lehmann. 1972: 228) กล่าวว่า เพื่อให้การตรวจคำถามปลายเปิดมีความเชื่อมั่นเพิ่มมากขึ้น ผู้ตรวจควรใช้วิธีการตรวจที่เหมาะสม ใช้เกณฑ์การตรวจกับนักเรียนทุกคน มีการออกแบบกฎเกณฑ์การให้คะแนน (Scoring rubric) การฝึกหรืออบรมผู้ตรวจ (Lane; et al. 1996: 72; citing Linn. 1993) และควบคุมแหล่งความคลาดเคลื่อน เช่น ข้อสอบ ผู้ตรวจ วัสดุงานหน้า ใช้ผู้ตรวจหลาย ๆ คนในการตรวจข้อสอบแต่ละข้อแล้วจึงหาค่าเฉลี่ยของคะแนน (Ebel. 1951; Rudner. 1992: online) และไม่ควรให้ผู้ตรวจเพียงคนเดียวเพราะจะทำให้ขาดความเที่ยงตรงของคะแนน (Uebersax. 2003: online) จากผลการวิจัย พบว่า เมื่อเพิ่มจำนวนผู้ตรวจจะทำให้ค่าความเชื่อมั่นของคะแนนเพิ่มสูงขึ้น (Swartz; et al. 1999; Sudweeks; Reeve; & Bradshaw. 2005; Brennan Gao; & Colton. 1995) แต่จำนวนผู้ตรวจที่เหมาะสมนั้นยังไม่สามารถสรุปได้แน่นอนเนื่องจากผลการศึกษายังพบแตกต่างกัน จึงยังคงต้องการคำตอบว่าจำนวนผู้ตรวจควรเป็นเท่าใดจึงจะทำให้คะแนนที่ได้มีความน่าเชื่อถือ (Sudweeks; Reeve; & Bradshaw. 2005: 240) อย่างไรก็ตามควรคำนึงในทางปฏิบัติที่จะให้นักเรียนทำข้อสอบหลายข้อและได้รับการตรวจจากผู้ตรวจหลายคนด้วย (Nakamura. 2002: online; Sudweeks; Reeve; & Bradshaw. 2005: 240)

การตรวจให้คะแนนที่ยุติธรรมที่สุดคือให้ผู้ตรวจทุกคนตรวจข้อสอบทุกข้อของผู้สอบทุกคน แล้วคำนวณคะแนนเฉลี่ย แต่ต้องใช้เวลาในการตรวจค่อนข้างมากและมีค่าใช้จ่ายสูง หากสามารถวางแผนการตรวจหรือหาวิธีการตรวจเพื่อลดภาระ ลดเวลาในการตรวจได้โดยที่มีความเป็นไปได้ในทางปฏิบัติ (Practical) แต่ยังคงไว้ซึ่งความน่าเชื่อถือ (Reliable) ของคะแนนและให้ค่าทางสถิติที่ยอมรับได้ (Statistically well-founded) จะเป็นทางหนึ่งในการแก้ปัญหาที่เกิดขึ้นได้ (Linacre; & Wright. 2002: 490-493; Nakamura. 2002: online) ทั้งนี้ คิม (Kim. 2000: online) มายฟอร์ดและวูล์ฟ (Myford; & Wolfe. 2003: 392) ไอรามณีรัตน์และคณะ (Iramaneerat; et al. 2007: online) ได้ให้ข้อเสนอแนะในการวางแผนการตรวจไว้สอดคล้องกันว่าไม่จำเป็นที่ผู้ตรวจต้องตรวจให้คะแนนผู้สอบทุกคนทุกข้อ อาจให้ผู้ตรวจตรวจข้อสอบของผู้สอบบางคน หรือตรวจข้อสอบในบางคุณลักษณะหรือข้อสอบบางข้อ การวางแผนการตรวจอาจทำได้ทั้งแบบบางส่วน (Partially) หรือแบบไขว้สมบูรณ (Full crossed) ผู้ตรวจสามารถตรวจข้อสอบบางข้อได้โดยมีการเชื่อมโยง (Connected) หรือไขว้กัน (Crossed) ของผู้ตรวจ และข้อสอบก็เพียงพอแล้วในการประมาณค่า โดยที่ยังคงมีความน่าเชื่อถือในทางสถิติ (Engelhard. 1992: 188; Linacre. 1994; Linacre; et al. 1994; Brennan. 2001) นั่นคือ ควรมีการออกแบบรูปแบบการตรวจให้คะแนนที่เหมาะสม เสียค่าใช้จ่ายน้อยที่สุด พยายามให้ผู้ตรวจหรือเงื่อนไขการวัดให้น้อยที่สุด โดยที่ยังมีอำนาจ (Power) มากที่สุด (Kim. 2000: online) ซึ่งจากการศึกษาของ สัตวิวัฒน์ ธีรและแบรดชอว์ (Sudweeks; Reeve; & Bradshaw. 2005: 259) พบว่าการใช้แบบแผนแบบแฟก

(Nested) ที่ให้ผู้ตรวจแต่ละคนตรวจข้อสอบทุกข้อของนักเรียนที่ต่างกัน จะได้คะแนนที่มีค่าความเชื่อมั่นแตกต่างจากแบบแผนแบบไขว้สมบูรณ์ (Full crossed) เพียง 1.5 – 2.0 % เท่านั้น

นอกจากนี้แล้วยังมีวิธีการทางสถิติที่สามารถค้นหา ตรวจสอบแหล่งความคลาดเคลื่อนหรือความแปรปรวนจากแหล่งต่าง ๆ ที่เกิดขึ้นจากการวัดประเมินด้วยแบบทดสอบปลายเปิด (Engelhard. 1992: 172-173; Myford; & Wolfe. 2003: 396; Smith; & Kulikowich. 2004: 619; Iramaneerat; et al. 2007: online) โดยเฉพาะความคลาดเคลื่อนจากข้อสอบและผู้ตรวจ (Bachman; Lynch; & Mason.1993: online) พร้อมทั้งปรับแก้คะแนนที่เกิดจากความแตกต่างของลักษณะของผู้ตรวจหรือความยากของข้อสอบ เพื่อให้เกิดความยุติธรรม ความเชื่อมั่นของคะแนนและความเที่ยงตรงของการตัดสินใจ (Lunz; Wright; & Linacre. 1990: online) ทั้งนี้วิธีการทางสถิติที่มีผู้นิยมนำมาใช้ ได้แก่ โมเดลการสรุปอ้างอิงตามทฤษฎีการสรุปอ้างอิง (Generalizability Theory) (Baxter; et al. 1992; Swartz; et al. 1999; Chiu; & Wolfe. 2002: 322; Lee; & Kantor. 2005: online; Iramaneerat; et al. 2007: online) และโมเดลหลายองค์ประกอบของราสช์ (Many-Facet Rasch Model) ตามทฤษฎีการตอบข้อสอบ (Item Response Theory) (Lunz; Wright; & Linacre. 1990; Engelhard.1992,1994; Linacre; et al. 1994; Nakamura. 2002: online. Bock; & Ockey. 2003: online; Myford; & Wolfe. 2003: 396; Park. 2004: online; Iramaneerat; et al. 2007: online)

ทฤษฎีการสรุปอ้างอิง (Generalizability Theory: G theory) (Cronbach; et al. 1972) ขยายมาจากทฤษฎีการทดสอบแบบมาตรฐานเดิม (Classical Test Theory: CTT) จากเดิมที่มีพื้นฐานความเชื่อว่าการแปรปรวนของคะแนนสังเกต (σ_x^2) ประกอบด้วยความแปรปรวนของคะแนนจริง (σ_t^2) และความแปรปรวนของความคลาดเคลื่อน (σ_e^2) ซึ่งเป็นความคลาดเคลื่อนรวมทุกแหล่ง (Single error source) ไม่สามารถระบุหรือแบ่งแยกได้ ประมาณค่าความเชื่อมั่นโดยใช้ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างคะแนนจริงกับคะแนนสังเกต ภายใต้เงื่อนไขการวัดที่เท่าเทียมกันหรือการวัดนั้นต้องมีความเป็นคู่ขนานกัน ซึ่งในทางปฏิบัติโดยทั่วไปไม่สามารถจัดหาเครื่องมือที่มีคุณภาพที่เท่าเทียมกันอย่างสมบูรณ์ได้ ในขณะที่ทฤษฎีการสรุปอ้างอิงนี้ได้พิจารณาความคลาดเคลื่อนจากหลายแหล่ง (Multiple error sources) ยอมรับว่าความแปรปรวนของผลการวัดที่เกิดขึ้นเกิดจากองค์ประกอบต่าง ๆ หลายแหล่งด้วยกันและใช้วิธีวิเคราะห์ความแปรปรวน (Analysis of Variance: ANOVA) ในการวิเคราะห์ความแปรปรวนของคะแนนความคลาดเคลื่อนจากแหล่งต่าง ๆ อันเป็นสถานการณ์ของการวัดที่เป็นไปได้ ทำให้สามารถตรวจสอบถึงแหล่งความแปรปรวนที่ไม่พึงประสงค์ และสามารถควบคุมแหล่งความคลาดเคลื่อนได้ตรงประเด็น (Shavelson; & Webb. 1991: 6) อีกทั้งไม่ยึดข้อตกลงเกี่ยวกับคุณสมบัติความเท่าเทียมกันของข้อสอบ นอกจากนี้แล้วยังสามารถประมาณค่าสัมประสิทธิ์การสรุปอ้างอิง (Generalizability coefficient) ที่แสดงถึงระดับความเชื่อถือได้ของคะแนน (Level of

dependability) คล้ายคลึงกับสัมประสิทธิ์ความเชื่อมั่นในทฤษฎีการทดสอบแบบมาตรฐานเดิม (Shavelson; & Webb. 1991: 2) ทฤษฎีนี้จึงเหมาะกับการวัดประเมินการปฏิบัติของบุคคล (Brennan. 2001: 117) สามารถนำมาใช้กับกรณีของการวัดประเมินที่มีแหล่งความคลาดเคลื่อนหลายแหล่งได้ เช่น ความคลาดเคลื่อนจากผู้ตรวจให้คะแนนในการตรวจสอบความเรียงหรือแบบทดสอบปลายเปิด เป็นต้น (ศิริชัย กาญจนวาสี. 2550: 11)

จากคุณลักษณะของทฤษฎีการสรุปอ้างอิงที่สามารถวิเคราะห์แหล่งความแปรปรวนของความคลาดเคลื่อนได้หลายแหล่งและไม่ยึดข้อตกลงเกี่ยวกับคุณสมบัติเท่าเทียมกันของข้อสอบ ทำให้มีการนำโมเดลนี้มาใช้อย่างกว้างขวางในการวัดประเมินทางด้านการเขียนและการวัดทางการปฏิบัติ ซึ่งการวัดประเมินเหล่านี้ต่างมีแหล่งความคลาดเคลื่อนหลายแหล่ง โดยเฉพาะจากผู้ตรวจให้คะแนน (Brennan. 1992) เช่น งานวิจัยของไอบราฮิม (Ibrahim. 1984) โอเบรียนและโจนส์ (O'Brien; & Jones. 1986) ที่ศึกษาแหล่งความแปรปรวนที่ส่งผลต่อการวัดประเมินพบว่า ปัจจัยที่มีผลต่อความเชื่อมั่นของการวัดประเมินมากที่สุดคือ ผู้ตรวจ จำนวนผู้ตรวจ สอดคล้องกับการศึกษาของเบรนนอน เกาและคอลตัน (Brennan; Gao; & Colton. 1995) ที่พบว่าจำนวนผู้ตรวจมีอิทธิพลต่อค่าความเชื่อมั่นและความแม่นยำของคะแนนจากการทดสอบทักษะการฟังและการเขียน โดยหากต้องการความแม่นยำของการวัดต้องใช้ผู้ตรวจ 2 คนขึ้นไป ส่วนสวาร์ทซ์และคณะ (Swartz; et al. 1999) ได้ประมาณค่าความเชื่อมั่นของคะแนนจากงานเขียน (Writing scores) พบว่า เมื่อเพิ่มจำนวนผู้ตรวจจะทำให้สัมประสิทธิ์ความเชื่อมั่นเพิ่มสูงขึ้น หากต้องการให้ค่าความเชื่อมั่นเป็นไปตามเกณฑ์ที่ยอมรับได้ของนูนัลลี (Nunnally) คือ 0.90 (Swartz. 1999: 505; citing Nunnally. 1967) นั้นต้องใช้ผู้ตรวจ 4 คนขึ้นไป ในขณะที่ผลการศึกษาของเลนและคณะ (Lane; et al. 1996) พบว่า มีความคลาดเคลื่อนจากผู้ตรวจน้อยในการวัดประเมินคำถามปลายเปิด โดยได้ให้เหตุผลว่าเนื่องมาจากการให้รายละเอียดเกี่ยวกับกฎเกณฑ์การให้คะแนนและตัวอย่างการตอบของนักเรียน พร้อมทั้งฝึกฝนวิธีการตรวจให้คะแนนกับผู้ตรวจ และผลการศึกษาของสัลดวีส์ รีฟและแบรดชอว์ (Sudweeks; Reeve; & Bradshaw. 2005) พบว่า ค่าความเชื่อมั่นของคะแนนงานเขียนจะมีค่ามากกว่า 0.70 เมื่อใช้ข้อสอบอย่างน้อย 3 ข้อ และใช้ผู้ตรวจอย่างน้อย 3 คน หากเพิ่มจำนวนข้อสอบและจำนวนผู้ตรวจจะให้ค่าความเชื่อมั่นเพิ่มสูงขึ้น นอกจากนี้ชัยฤทธิ์ ศิลาเดช (2540) ได้ประมาณค่าความเชื่อมั่นของการประเมินคุณภาพของการใช้แฟ้มสะสมงานในการประเมินผลการเรียน พบว่า จำนวนผู้ประเมินที่เหมาะสมคือ 2-3 คน พร้อมทั้งให้ข้อเสนอแนะว่าไม่ควรให้ผู้ประเมินเพียงคนเดียว เพราะจะทำให้ความเชื่อมั่นของคะแนนมีค่าค่อนข้างต่ำและหากใช้ผู้ประเมินมากกว่า 4 คน จะทำให้ค่าความเชื่อมั่นเพิ่มขึ้นอีกเพียงเล็กน้อยเท่านั้น ดังนั้น จึงไม่จำเป็นที่ต้องใช้จำนวนผู้ประเมินมากเกินไป

ผลการวิจัยดังกล่าวข้างต้นต่างสอดคล้องกันในประเด็นที่พบว่า จำนวนผู้ตรวจส่งผลต่อค่าสัมประสิทธิ์การสรุปอ้างอิง กล่าวคือ การเพิ่มจำนวนผู้ตรวจจะทำให้ค่าความเชื่อมั่นเพิ่มสูงขึ้น แต่จำนวนผู้ตรวจที่เหมาะสมนั้นผลการศึกษายังพบแตกต่างกัน อย่างไรก็ตามทฤษฎีการสรุปอ้างอิงจะประมาณค่าความเชื่อมั่นโดยใช้คะแนนดิบที่ยังไม่ได้ปรับแก้คะแนนอันเกิดจากความแตกต่างของความเข้มงวดของผู้ตรวจหรือความยากของข้อสอบ (Smith; & Kulikowich. 2004: 627) และเป็นการศึกษาในระดับกลุ่ม (Group level) หรือโดยภาพรวมทั้งหมด ยังไม่สามารถระบุได้ถึงแต่ละสมาชิกในองค์ประกอบ (Individual level) ทำให้ยังไม่สามารถอธิบายรายละเอียดของแต่ละสมาชิกในแต่ละองค์ประกอบได้ (Linacre. 1993; Stemler. 2004: online; Lee; & Kantor. 2005: 40-41; Sudweeks; Reeve; & Bradshaw. 2005: 251; Iramaneerat; et al. 2007: online)

นอกจากโมเดลการสรุปอ้างอิงแล้วยังมีวิธีการทางสถิติอีกวิธีหนึ่งที่สามารถค้นหา ตรวจสอบ แหล่งความแปรปรวนได้หลายแหล่ง คือ โมเดลหลายองค์ประกอบของราส์ช (Many - Facet Rasch Model: MFRM) พัฒนาขึ้นในปี 1989 โดย ลินาเคอร์ (Linacre. 1994) ซึ่งได้ปรับขยายมาจากราส์ช โมเดล (Rasch Model) แบบ 1 พารามิเตอร์ ตามแนวคิดของทฤษฎีการตอบข้อสอบ (Item Response Theory: IRT) ทั้งนี้ทฤษฎีการตอบข้อสอบจะอธิบายกระบวนการวัดในรูปคุณลักษณะแฝง (Latent Trait) และมีข้อตกลงว่าการตอบข้อสอบเป็นฟังก์ชันของระดับคุณลักษณะของผู้สอบและข้อสอบ และแสดงถึงความน่าจะเป็นในการตอบข้อสอบได้ถูกต้อง โดยมีแนวคิดพื้นฐานว่า ถ้าความสามารถของผู้สอบอยู่ในระดับสูง โอกาสในการตอบข้อสอบได้ถูกต้องจะสูงตามไปด้วย ในระยะแรกนั้นมุ่งนำไปใช้วิเคราะห์กับแบบทดสอบที่ตรวจให้คะแนนแบบสองค่า ในเวลาต่อมาได้มีการพัฒนาโมเดลการตรวจให้คะแนนแบบหลายค่า (Hambleton; Swanminathan; & Roger. 1991) ซึ่งโมเดลเหล่านี้จะประมาณค่าพารามิเตอร์ได้ 2 ค่าคือ พารามิเตอร์ของผู้สอบและพารามิเตอร์ของข้อสอบ โมเดลหลายองค์ประกอบของราส์ชที่ลินาเคอร์ (Linacre. 1994) พัฒนาขึ้นมานั้น ได้ปรับขยายแนวคิดจากเดิมที่มุ่งเน้นไปที่สององค์ประกอบคือ ผู้สอบและข้อสอบ และให้ค่าพารามิเตอร์ 2 ค่าคือ พารามิเตอร์ของผู้สอบ (Examinee parameter: θ) แสดงถึงค่าความสามารถ (Ability or Competence) ของผู้สอบ และพารามิเตอร์ของข้อสอบ (Item parameter: b_i) ที่แสดงถึงระดับความยาก (Difficulty) ของข้อสอบ ไปสู่สถานการณ์ที่มีมากกว่า 2 องค์ประกอบ เช่น เพิ่มองค์ประกอบผู้ตรวจให้คะแนน (Rater facet) เข้าไปในโมเดล ทำให้สามารถประมาณค่าพารามิเตอร์ของผู้ตรวจให้คะแนน (Rater parameter) ที่แสดงถึงความเข้มงวดของผู้ตรวจ (Rater severity) ได้ หรืออาจเพิ่มองค์ประกอบอื่นที่คาดว่าจะมีอิทธิพลต่อคะแนน โดยการนำองค์ประกอบเหล่านี้เข้าไปศึกษาในโมเดลด้วย จึงทำให้สามารถศึกษาองค์ประกอบต่าง ๆ ที่มีอิทธิพลต่อการวัดประเมินได้หลายองค์ประกอบพร้อมกัน

ลินาเคอร์ (Linacre. 1994) มีแนวคิดว่าการที่ผู้สอบตอบคำถามปลายเปิดแล้วจะได้รับการประเมินในระดับคะแนนใดนั้นได้รับอิทธิพลจาก 4 องค์ประกอบ คือ ความสามารถของผู้สอบ ความเข้มงวดของผู้ตรวจ ความยากของข้อสอบและโครงสร้างของมาตราประเมินค่า โดย MFRM จะประมาณค่าความสามารถของผู้สอบ (θ) ที่เป็นอิสระจากกลุ่มผู้สอบ (Sample free) อิสระจากผู้ตรวจ (Rater free) และอิสระจากข้อสอบ (Item free) (Linacre; & Wright. 2002: 511; Smith; & Kulikowich. 2004: 621) ดังนั้นค่าความสามารถของผู้สอบ (θ) จึงมีค่าไม่แปรเปลี่ยน (Invariant) ไปตามผู้ตรวจและข้อสอบ เนื่องจากค่าความสามารถของผู้สอบ (θ) ที่ได้จาก MFRM ได้รับการปรับแก้จากความแตกต่างของความเข้มงวดของผู้ตรวจและความยากของข้อสอบ ซึ่งคุณสมบัติเหล่านี้มีประโยชน์อย่างมากสำหรับการนำมาใช้ในการสร้างและพัฒนาแบบทดสอบ (ผจจจิต อินทสุวรรณ. 2525: 66-67) ทำให้สามารถนำคะแนนที่ได้ไปเปรียบเทียบกันได้โดยตรง จึงเป็นวิธีหาคะแนนที่ยุติธรรมสำหรับการตรวจข้อสอบปลายเปิดแม้ว่าจะได้รับการตรวจจากผู้ตรวจที่มีความเข้มงวดต่างกันหรือจากข้อสอบที่มีความยากต่างกันก็ตาม โดยที่ผู้ตรวจทุกคนไม่จำเป็นต้องตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Sudweeks; Reeve; & Bradshaw. 2005: 251) การวิเคราะห์ด้วย MFRM มุ่งเน้นไปที่แต่ละสมาชิกในองค์ประกอบ (Individual level) ที่เป็นแหล่งของความคลาดเคลื่อน เช่น พฤติกรรมการตรวจให้คะแนนของผู้ตรวจแต่ละคนว่ามีความเข้มงวดมากน้อยเพียงใด (Stemler. 2004: online; Lee; & Kantor. 2005: 40-41; Iramaneerat; et al. 2007: online) ซึ่งผลการศึกษานี้สามารถใช้เป็นข้อมูลย้อนกลับ (Feedback) ให้กับผู้ที่เกี่ยวข้อง เพื่อการพัฒนาคุณภาพของแต่ละสมาชิกในระบบการวัดประเมิน (Myford; & Wolfe. 2000: online; Iramaneerat; et al. 2007: online) เพื่อทำให้การตรวจให้คะแนนมีความน่าเชื่อถือ ยุติธรรมและเป็นปรนัยมากขึ้น (Engelhard. 1992: 187; Linacre. 1994; Park. 2004: online)

นักวิจัยได้ศึกษาและประยุกต์ใช้ MFRM อย่างหลากหลายเพื่อตรวจสอบและสร้างความเข้าใจถึงแหล่งความแปรปรวนของคะแนนที่ได้ ทั้งจากการวัดประเมินทางภาษาในด้านการพูดหรือการเขียน การวัดประเมินแฟ้มสะสมงาน เช่น เอนเกลฮาร์ด (Engelhard. 1992; 1994) บอนคและออกเคย์ (Bonk; & Ockey. 2003: online) ได้ใช้โมเดลนี้ในการวัดความสามารถของผู้สอบและตรวจสอบความคลาดเคลื่อนของผู้ตรวจ พบว่า มีความแตกต่างอย่างมีนัยสำคัญของความเข้มงวดของผู้ตรวจ ในขณะที่มายฟอร์ดและวูล์ฟ (Myford; & Wolfe. 2000: online) ได้ศึกษาแหล่งความแปรปรวนพร้อมทั้งหาขนาดและขอบเขตของแหล่งความแปรปรวนจากการทดสอบภาษาอังกฤษในฐานะภาษาต่างประเทศ (The Test of English as a Foreign Language: TOEFL) ของหน่วยบริการการสอบทางการศึกษาของสหรัฐอเมริกา (Educational Testing Service: ETS) พบว่า ความคลาดเคลื่อนมาตรฐานของการวัดมีการกระจายตลอดช่วงการแจกแจง โดยเฉพาะในส่วนปลายของการแจกแจง

ข้อสอบมีความยากแตกต่างกันเล็กน้อยและผู้ตรวจจะมีระดับความเข้มงวดแตกต่างกัน เมื่อปรับแก้คะแนนจากความเข้มงวดของผู้ตรวจที่ต่างกัน ทำให้คะแนน 2 ใน 3 ของผู้สอบทั้งหมดมีค่าเฉลี่ยต่างจากค่าเฉลี่ยของคะแนนดิบประมาณ 0.5-3.6 คะแนน สอดคล้องกับงานวิจัยของเทอร์เนอร์ (Turner. 2003: online) ที่นำ MFRM ไปวิเคราะห์การให้คะแนนแฟ้มสะสมงานทางศิลปะ และปาร์ค (Park. 2004: online) ซึ่งได้ตรวจสอบอิทธิพลของแหล่งความแปรปรวนที่ส่งผลถึงการปฏิบัติของผู้สอบจากการทดสอบการเขียนภาษาอังกฤษในฐานะภาษาที่สอง (English as a Second Language: ESL) พบว่า ผู้ตรวจจะมีระดับความเข้มงวดแตกต่างกันอย่างมีนัยสำคัญ และเมื่อปรับแก้ค่าความสามารถของนักเรียนจากความเข้มงวดของผู้ตรวจที่ต่างกันแล้วทำให้มีการเปลี่ยนลำดับของนักเรียน ทั้งนี้เทอร์เนอร์ได้กล่าวสรุปไว้ว่า การวิเคราะห์ด้วย MFRM จะให้ค่าการวัดความสามารถของนักเรียนที่ปลอดภัยกว่าความเห็นส่วนตัวหรือความเป็นอัตนัย (Subjective) ของผู้ตรวจ ส่วนนากามูระ (Nakamura. 2002: online) ได้ศึกษาการใช้ระบบการตรวจเป็นคู่ว่าจะสามารถลดภาระงานของผู้ตรวจได้หรือไม่ โดยที่ยังมีความน่าเชื่อถือทางสถิติ พบว่า การใช้ระบบการตรวจเป็นคู่สามารถลดภาระงานของผู้ตรวจได้ แม้ว่าจะมีระดับของความเข้มงวดของผู้ตรวจแตกต่างกันแต่อยู่ในช่วงที่ยอมรับได้และไม่มีอิทธิพลอย่างมีนัยสำคัญ คะแนนที่ได้จึงไม่ขึ้นอยู่กับคู่ของผู้ตรวจ ดังนั้นผู้ตรวจจึงไม่จำเป็นต้องตรวจข้อสอบของนักเรียนทุกคน ในขณะที่นักวิจัยของไทยยังมีการนำ MFRM ไปประยุกต์ใช้ไม่มากนัก จากงานวิจัยที่กล่าวมาข้างต้นต่างให้ผลการวิจัยที่สอดคล้องกันว่าแหล่งความคลาดเคลื่อนที่สำคัญในการวัดประเมินคือ ผู้ตรวจ ผู้ตรวจมีความเข้มงวดที่แตกต่างซึ่งส่งผลต่อการประมาณค่าความสามารถของผู้สอบ และการใช้ MFRM วิเคราะห์คะแนนสามารถแสดงให้เห็นรายละเอียดของแต่ละสมาชิกในองค์ประกอบที่ศึกษา เช่น ความเข้มงวดของผู้ตรวจแต่ละคนและเมื่อนำเอาความแตกต่างของผู้ตรวจนี้เข้าไปปรับแก้ในโมเดลทำให้คะแนนมีความน่าเชื่อถือและยุติธรรมมากขึ้น โดยที่ผู้ตรวจทุกคนไม่จำเป็นต้องตรวจข้อสอบทุกข้อของผู้สอบทุกคน

นอกจากนี้ยังมีการใช้โมเดลหลายองค์ประกอบของราล์ซและโมเดลการสรุปอ้างอิงร่วมกันในการศึกษาวิจัยในการวัดประเมินทักษะทางการพูด (Bachman; Lynch; & Mason.1993: online; Lynch; & McNamara. 1998: online) การเขียน (Sudweeks; Reeve; & Bradshaw. 2005) ทักษะการแก้ปัญหาที่ซับซ้อน (Smith; & Kulikowich. 2004) ตลอดจนถึงการทดสอบในวงการแพทย์ (Iramaneerat; et al. 2007: online) ซึ่งนักวิจัยเหล่านี้ต่างให้ความเห็นที่สอดคล้องกันว่าโมเดลทั้งสองเป็นโมเดลที่เติมเต็ม (Complementary) ซึ่งกันและกัน การใช้ทั้งสองโมเดลร่วมกันเป็นการขยายขอบเขตของการศึกษาและต่างให้ผลการศึกษที่สนับสนุนซึ่งกันและกัน สามารถศึกษาได้ทั้งในระดับกลุ่มและระดับแต่ละสมาชิกในองค์ประกอบ จึงทำให้ได้สารสนเทศของการวัดประเมินที่สมบูรณ์มากขึ้น ก่อให้เกิดประโยชน์ทั้งผู้ใช้และผู้สร้างแบบทดสอบ อีกทั้งยังสามารถวิเคราะห์คุณลักษณะของ

คะแนนได้โดยที่ผู้ตรวจไม่จำเป็นต้องตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Bachman; Lynch; & Mason.1993: online; Lynch; & McNamara. 1998: online, Lee; & Kantor. 2005: 41; Iramaneerat; et al.2007: online) ส่วนงานวิจัยในประเทศไทยนั้น จากการศึกษาค้นคว้ายังไม่พบว่ามีกรนำโมเดลการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราล์สไปใช้ศึกษาวิจัยร่วมกันแต่อย่างใด

จากที่กล่าวมาข้างต้นจะเห็นว่าคะแนนที่ได้จากแบบทดสอบปลายเปิดนั้นมีแหล่งความคลาดเคลื่อนหลายแหล่งที่ส่งผลต่อความน่าเชื่อถือ ความเที่ยงตรงและความยุติธรรมของคะแนน โดยเฉพาะความคลาดเคลื่อนจากผู้ตรวจ หากสามารถตรวจสอบแหล่งความแปรปรวนเหล่านี้ได้โดยใช้วิธีการทางสถิติในการค้นหาและอธิบายแหล่งความคลาดเคลื่อนต่าง ๆ (Engelhard. 1992: 172-173; Myford; & Wolfe. 2003: 396; Smith; & Kulikowich. 2004: 619; Iramaneerat; et al. 2007: online) จะทำให้สามารถอธิบายคะแนนได้อย่างชัดเจน และมีความน่าเชื่อถือของคะแนนมากขึ้น นอกจากนี้แล้วในการตรวจให้คะแนนแบบทดสอบปลายเปิดนั้นต้องใช้เวลา แรงงานและค่าใช้จ่ายสูง ซึ่งผลการศึกษาวิจัย พบว่า การเพิ่มจำนวนผู้ตรวจทำให้ค่าความเชื่อมั่นเพิ่มสูงขึ้น (O'Brien; & Jones. 1986; Brennan; Gao; & Colton. 1995; Swartz; et al. 1999; Sudweeks; Reeve; & Bradshaw. 2005) แต่จำนวนผู้ตรวจที่เหมาะสมนั้นผลการศึกษายังพบแตกต่างกันและนักวิจัยเหล่านี้ยังได้ให้ข้อเสนอแนะในการศึกษารูปแบบการตรวจให้คะแนนที่ผู้ตรวจไม่จำเป็นต้องตรวจข้อสอบทุกข้อของผู้สอบทุกคน เพื่อประหยัดเวลา ลดความยุ่งยากและภาระในการตรวจ ทำให้ยังคงต้องการคำตอบว่าจำนวนผู้ตรวจควรเป็นเท่าใด (Sudweeks; Reeve; & Bradshaw. 2005) หรือควรมีรูปแบบการตรวจให้คะแนนอย่างไรเพื่อลดภาระและเวลาในการตรวจ โดยที่ยังคงไว้ซึ่งความเชื่อมั่นและความเที่ยงตรงของคะแนน (Linacre. 2002: 490-493) ทั้งนี้โมเดลการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราล์สสามารถตรวจสอบแหล่งความคลาดเคลื่อนได้หลายแหล่งในสถานการณ์การวัดผลได้พร้อมกันจากการวิเคราะห์เพียงครั้งเดียว สามารถประมาณค่าน้ำหนัก (Magnitude) ของแหล่งความแปรปรวนของความคลาดเคลื่อน (Bachman; Lynch; & Mason.1993: online) และประมาณค่าความเชื่อมั่นของคะแนนได้ด้วยเช่นกัน โดยไม่จำเป็นที่ผู้ตรวจต้องตรวจให้คะแนนทุกข้อของผู้สอบทุกคน จึงทำให้สามารถนำโมเดลทั้งสองมาใช้ศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิด เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกันได้ โดยโมเดลการสรุปอ้างอิงจะเป็นการศึกษาแหล่งความแปรปรวนในภาพรวม (Group level) ในขณะที่โมเดลหลายองค์ประกอบของราล์สนั้นจะอธิบายแหล่งความแปรปรวนต่าง ๆ ในระดับแต่ละสมาชิก (Individual level) ขององค์ประกอบ และนำเอาความแตกต่างของความเข้มงวดของผู้ตรวจและความยากของข้อสอบเข้าไปปรับแก้ในโมเดลด้วยโมเดลการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราล์สจึงเป็นวิธีการทางสถิติที่สามารถวัดประเมินคุณลักษณะของคะแนนได้ (Iramaneerat; et al. 2007: online) โมเดลทั้งสองเป็นโมเดลที่เติม

เติมซึ่งกันและกัน ทำให้ได้สารสนเทศของการวัดประเมินที่สมบูรณ์มากขึ้น ก่อให้เกิดประโยชน์ทั้งผู้ใช้และผู้สร้างแบบทดสอบ (Bachman; Lynch; & Mason. 1993: online; Lynch; & McNamara. 1998: online; Lee; & Kantor. 2005: 41; Iramaneerat; et al. 2007: online)

ด้วยเหตุผลดังกล่าวข้างต้น ผู้วิจัยจึงสนใจศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน โดยใช้โมเดลการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราล์ช ผู้วิจัยศึกษาความแปรปรวนขององค์ประกอบหลัก ได้แก่ ผู้สอบ ข้อสอบและผู้ตรวจ พิจารณาคุณลักษณะของคะแนนจากค่าสัมประสิทธิ์การสรุปอ้างอิง ค่าความเชื่อมั่นแยกส่วน ค่าความเที่ยงตรงตามสภาพ และเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิง ค่าความเชื่อมั่นแยกส่วน ค่าความเที่ยงตรงตามสภาพ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน ซึ่งเป็นการเปรียบเทียบภายในแต่ละโมเดล เพื่อเป็นแนวทางให้กับผู้ที่เกี่ยวข้องในการเลือกใช้จำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนแบบทดสอบปลายเปิดที่มีความเชื่อมั่นและความเที่ยงตรงของคะแนน โดยสามารถประหยัดเวลา แรงงานในการตรวจและสอดคล้องกับสภาพจริงในทางปฏิบัติ

ความมุ่งหมายของการวิจัย

การวิจัยในครั้งนี้มีความมุ่งหมายหลักเพื่อศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน โดยใช้โมเดลการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราล์ช มีความมุ่งหมายเฉพาะ ดังต่อไปนี้

1. เพื่อศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ที่มีจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน โดยใช้โมเดลการสรุปอ้างอิง แยกได้เป็น
 - 1.1 เพื่อศึกษาขนาดขององค์ประกอบความแปรปรวนของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน
 - 1.2 เพื่อประมาณค่าและเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน
 - 1.3 เพื่อประมาณค่าและเปรียบเทียบความเที่ยงตรงตามสภาพของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน
2. เพื่อศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ที่มีจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน โดยใช้โมเดลหลายองค์ประกอบของราล์ช แยกได้เป็น
 - 2.1 เพื่อศึกษาขนาดขององค์ประกอบความแปรปรวนของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน

2.2 เพื่อประมาณค่าและเปรียบเทียบค่าความเชื่อมั่นแยกส่วนของแต่ละองค์ประกอบ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน

2.3 เพื่อประมาณค่าและเปรียบเทียบความเที่ยงตรงตามสภาพของคะแนนแบบทดสอบ ปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน

ความสำคัญของการวิจัย

การวิจัยในครั้งนี้เป็นการศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน โดยใช้โมเดลการสรุปอ้างอิง และโมเดลหลายองค์ประกอบของราล์ช มีความสำคัญของการวิจัยทั้งทางทฤษฎีและทางปฏิบัติ ดังนี้

1. การวิจัยในครั้งนี้เป็นส่วนต่อขยายองค์ความรู้ทางการวัดประเมินที่มีการตรวจให้คะแนนแบบหลายค่า (Polytomous) โดยเฉพาะอย่างยิ่งด้านคุณลักษณะของคะแนนที่ได้จากการวัดประเมิน ได้แก่ ค่าความเชื่อมั่นและความเที่ยงตรง ด้วยโมเดลการสรุปอ้างอิงในทฤษฎีการสรุปอ้างอิง และโมเดลหลายองค์ประกอบของราล์ชในทฤษฎีการตอบข้อสอบให้กว้างขวางและหลากหลายมากขึ้น ในวงการการทดสอบทางการศึกษา โดยนำทั้งสองโมเดลมาประยุกต์ใช้ในการศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิด ซึ่งเป็นเครื่องมือในการวัดประเมินที่ยังมีปัญหามาในประเด็นความน่าเชื่อถือและความเที่ยงตรงของคะแนน ต้องใช้เวลาและแรงงานในการตรวจค่อนข้างมาก ผลการศึกษาในครั้งนี้ทำให้ได้ข้อมูลเชิงประจักษ์เกี่ยวกับจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนแบบทดสอบปลายเปิดที่ทำให้คะแนนมีความเชื่อมั่นและความเที่ยงตรง นอกจากนี้ยังทำให้ทราบขนาดความแปรปรวนของแหล่งต่าง ๆ ที่มีอิทธิพลต่อความน่าเชื่อถือของคะแนน ซึ่งเป็นประโยชน์ในแง่ของการตรวจสอบและใช้เป็นหลักฐานที่แสดงถึงแหล่งความคลาดเคลื่อนทั้งจากผู้สอบ ข้อสอบ และผู้ตรวจ เพื่อการอธิบายและแปลผลคะแนนได้อย่างถูกต้อง

2. สารสนเทศที่ได้จากการศึกษาในครั้งนี้ เป็นประโยชน์กับผู้ที่เกี่ยวข้องกับการวัดประเมินทางการศึกษาในหลายระดับ ดังนี้

2.1 ระดับชาติ หน่วยงานที่เกี่ยวข้องของการวัดและการประเมิน เช่น หน่วยงานทางการทดสอบระดับชาติที่ต้องการใช้แบบทดสอบปลายเปิดในการทดสอบนักเรียน เช่น การทดสอบความถนัดทั่วไป (General Aptitude Test: GAT) ความถนัดเฉพาะด้านวิชาการ (Professional Aptitude Test: PAT) การทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (Ordinary National Educational Test: O-NET) เป็นต้น สามารถเลือกใช้จำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนแบบทดสอบปลายเปิดที่สามารถลดภาระ เวลาในการตรวจ แต่มีความเชื่อมั่นและความเที่ยงตรงของคะแนน ตลอดจนสามารถอธิบายคะแนนที่ได้จากแบบทดสอบปลายเปิดว่ามีความคลาดเคลื่อนเกิดจากแหล่งใดบ้าง

และมีขนาดเท่าใด เพื่อนำไปปรับปรุง พัฒนาระบบการวัดประเมินและการบริหารการสอบให้มีคุณภาพมาตรฐานมากขึ้น ก่อให้เกิดความมั่นใจในการนำผลการวัดประเมินไปใช้

2.2 ระดับเขตพื้นที่การศึกษาและระดับสถานศึกษา ผู้ที่เกี่ยวข้องในกลุ่มงานวัดและประเมินผลการศึกษา ฝ่ายวัดประเมินของสถานศึกษา สามารถนำเสนอสารสนเทศที่ได้ไปกระตุ้นและส่งเสริมให้ครูผู้สอนใช้แบบทดสอบปลายเปิดในการวัดประเมินได้อย่างมั่นใจมากขึ้น ตลอดจนครูผู้สอนที่สามารถเลือกใช้รูปแบบการตรวจให้คะแนนและจำนวนผู้ตรวจที่ประหยัดเวลาและภาระงาน แต่ได้ผลการตรวจที่มีความน่าเชื่อถือ มีความเที่ยงตรงและสอดคล้องกับความเป็นจริงในทางปฏิบัติ และทำให้ทราบแหล่งความคลาดเคลื่อนที่เกิดขึ้น เพื่อให้ผู้ที่เกี่ยวข้องตระหนักถึงความสำคัญของแหล่งความคลาดเคลื่อนและนำมาพิจารณาเพื่อการประเมินผลหรือตัดสินผลการเรียนของนักเรียนได้ถูกต้อง อีกทั้งสามารถใช้เป็นข้อมูลย้อนกลับ (Feedback) ให้กับผู้ที่เกี่ยวข้องในการวางแผนเพื่อลดความคลาดเคลื่อนในการบริหารการสอบครั้งต่อไปเพื่อให้กระบวนการวัดประเมินมีคุณภาพมากยิ่งขึ้น

ขอบเขตของการวิจัย

การวิจัยครั้งนี้ผู้วิจัยกำหนดขอบเขตของการวิจัย ดังต่อไปนี้

ประชากรที่ใช้ในการวิจัย

ประชากรที่ใช้ในการวิจัยครั้งนี้คือ นักเรียนระดับชั้นมัธยมศึกษาปีที่ 4 สำนักงานเขตพื้นที่การศึกษาจังหวัดน่าน ปีการศึกษา 2552 จำนวน 3,476 คน

กลุ่มตัวอย่างที่ใช้ในการวิจัย

กลุ่มตัวอย่างที่ใช้ในการวิจัยครั้งนี้ คือ นักเรียนระดับชั้นมัธยมศึกษาปีที่ 4 สำนักงานเขตพื้นที่การศึกษาจังหวัดน่าน ปีการศึกษา 2552 จำนวน 180 คน โดยวิธีการสุ่มแบบสองขั้นตอน

ตัวแปรที่ศึกษา

1. ตัวแปรอิสระมี 2 ตัว ได้แก่

1.1 จำนวนผู้ตรวจ มี 3 ลักษณะ คือ

1.1.1 จำนวนผู้ตรวจ 2 คน

1.1.2 จำนวนผู้ตรวจ 3 คน

1.1.3 จำนวนผู้ตรวจ 4 คน

1.2 รูปแบบการตรวจให้คะแนน มี 3 ลักษณะ คือ

1.2.1 ตรวจข้อสอบทุกข้อของผู้สอบทุกคน

1.2.2 ตรวจข้อสอบทุกข้อของผู้สอบบางคน

- 1.2.3 ตรวจสอบข้อสอบบางข้อของผู้สอบทุกคน
2. ตัวแปรตาม จำแนกเป็น 2 กลุ่ม ตามโมเดลที่ใช้ในการวิเคราะห์ ดังนี้
 - 2.1 วิเคราะห์ด้วยโมเดลการสรุปอ้างอิง ได้แก่
 - 2.1.1 ขนาดขององค์ประกอบความแปรปรวน
 - 2.1.2 ค่าสัมประสิทธิ์การสรุปอ้างอิง
 - 2.1.3 ความเที่ยงตรงตามสภาพ
 - 2.2 วิเคราะห์ด้วยโมเดลหลายองค์ประกอบของราล์ช ได้แก่
 - 2.2.1 ขนาดขององค์ประกอบความแปรปรวน
 - 2.2.2 ค่าความเชื่อมั่นแยกส่วน
 - 2.2.3 ความเที่ยงตรงตามสภาพ

นิยามศัพท์เฉพาะ

1. **คุณลักษณะของคะแนน** หมายถึง คุณสมบัติเฉพาะของคะแนนแบบทดสอบปลายเปิด วิชาคณิตศาสตร์ ที่ได้จากการวิเคราะห์ด้วยโมเดลการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราล์ช ได้แก่ ค่าสัมประสิทธิ์การสรุปอ้างอิง ค่าความเชื่อมั่นแยกส่วนและค่าความเที่ยงตรงตามสภาพ
2. **โมเดลการสรุปอ้างอิง** หมายถึง โมเดลของทฤษฎีการสรุปอ้างอิง (Generalizability Theory: G-Theory) ที่จำแนกแหล่งความคลาดเคลื่อนได้หลายแหล่ง ซึ่งในงานวิจัยนี้ศึกษา 2 แหล่ง คือ ข้อสอบและผู้ตรวจ ประเมินค่าขนาดความแปรปรวนของผู้สอบ ข้อสอบและผู้ตรวจ และประมาณค่าสัมประสิทธิ์การสรุปอ้างอิง วิเคราะห์โดยใช้โปรแกรม GENOVA (Crick; & Brennan. 1983)
3. **โมเดลหลายองค์ประกอบของราล์ช** หมายถึง โมเดลที่ขยายมาจากโมเดลแบบ 1 พารามิเตอร์ของราล์ชตามทฤษฎีการตอบข้อสอบ (Item Response Theory: IRT) ใช้สำหรับแบบทดสอบที่ให้คะแนนแบบหลายค่า โดยเพิ่มองค์ประกอบของผู้ตรวจ สามารถวิเคราะห์ความแปรปรวนจากหลายแหล่งได้ ซึ่งในงานวิจัยนี้ศึกษา 3 แหล่ง คือ ผู้สอบ ข้อสอบและผู้ตรวจ ประเมินค่าขนาดความแปรปรวนของผู้สอบ ข้อสอบและผู้ตรวจ และค่าความเชื่อมั่นแยกส่วนของแต่ละองค์ประกอบ วิเคราะห์โดยใช้โปรแกรม FACETS (Linacre. 2009)
4. **ค่าสัมประสิทธิ์การสรุปอ้างอิง** หมายถึง ค่าความเชื่อมั่นของคะแนนที่ได้จากแบบทดสอบปลายเปิดในเงื่อนไขหรือสถานการณ์ที่กำหนด แสดงถึงระดับแม่นยำและความน่าเชื่อถือของการสรุปอ้างอิงคะแนนที่ได้จากแบบทดสอบไปยังคะแนนเอกภพของผู้สอบ ประมาณค่าจากอัตราส่วนของความแปรปรวนของคะแนนเอกภพกับความแปรปรวนของคะแนนสังเกต

5. ค่าความเชื่อมั่นแยกส่วน หมายถึง ค่าความเชื่อมั่นแยกส่วนของแต่ละองค์ประกอบที่ได้จากโมเดลหลายองค์ประกอบของราล์ฟ ซึ่งงานวิจัยในครั้งนี้ ได้แก่

5.1 ค่าความเชื่อมั่นแยกส่วนของผู้สอบ หมายถึง ดัชนีที่แสดงถึงความแม่นยำของการวัดและความเชื่อถือได้ของคะแนนความสามารถของผู้สอบที่ได้จากข้อสอบและผู้ตรวจ บ่งชี้ถึงความสามารถในการจำแนกความสามารถผู้สอบ ประมาณค่าจากอัตราส่วนของความแปรปรวนของคะแนนจริงกับความแปรปรวนของคะแนนที่สังเกตได้ของผู้สอบ

5.2 ค่าความเชื่อมั่นแยกส่วนของข้อสอบ หมายถึง ดัชนีที่แสดงถึงความแตกต่างของระดับความยากของข้อสอบ บ่งชี้ถึงความสามารถในการจำแนกความยากของข้อสอบ ประมาณค่าจากอัตราส่วนของความแปรปรวนของค่าความยากที่แท้จริงกับความแปรปรวนของค่าความยากที่สังเกตได้ของข้อสอบ

5.3 ค่าความเชื่อมั่นแยกส่วนของผู้ตรวจ หมายถึง ดัชนีที่แสดงถึงความแตกต่างหรือความแปรปรวนของความเข้มงวดในการให้คะแนนของผู้ตรวจ เพื่อบ่งชี้ว่าผู้ตรวจมีความเข้มงวดในการให้คะแนนแตกต่างกันหรือไม่ ประมาณค่าจากอัตราส่วนของความแปรปรวนของค่าความเข้มงวดที่แท้จริงกับความแปรปรวนของค่าความเข้มงวดที่สังเกตได้ของผู้ตรวจ

6. องค์ประกอบความแปรปรวน หมายถึง ความแปรปรวนของแหล่งต่างๆ ที่มีอิทธิพลต่อคะแนนที่ได้จากแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ ซึ่งในงานวิจัยนี้ แหล่งของความแปรปรวนหลักที่ผู้วิจัยศึกษา คือ ผู้สอบ ข้อสอบและผู้ตรวจ และมีหน่วยการวิเคราะห์ (Unit analysis) ความแปรปรวนของผู้สอบ ข้อสอบและผู้ตรวจ คือ กลุ่มผู้สอบ กลุ่มข้อสอบ และกลุ่มผู้ตรวจ ตามลำดับ

7. ความเที่ยงตรงตามสภาพ หมายถึง คุณภาพของแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ ระดับชั้นมัธยมศึกษาตอนต้นที่สามารถวัดได้ตรงตามสภาพที่แท้จริงในปัจจุบันของผู้สอบ พิจารณาจากระดับความสอดคล้องระหว่างคะแนนที่ได้จากแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ที่สร้างขึ้นกับเกณฑ์ในขณะปัจจุบัน โดยหาความสัมพันธ์ของคะแนนที่ได้จากแบบทดสอบปลายเปิด ได้แก่ คะแนนดิบและคะแนนความสามารถ กับคะแนนจากแบบทดสอบมาตรฐาน ซึ่งงานวิจัยนี้ใช้ผลการทดสอบทางการศึกษาระดับชาตินี้พื้นฐาน (Ordinary National Educational Test: O-NET) ในระดับชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2551

8. แบบทดสอบปลายเปิดวิชาคณิตศาสตร์ หมายถึง ชุดของข้อคำถามที่มีคำตอบถูกเพียงคำตอบเดียวแต่อาจมีหลากหลายวิธีการได้มาซึ่งคำตอบ โดยให้นักเรียนเขียนอธิบาย แสดงเหตุผลหรือแสดงวิธีการหาคำตอบที่มีวิธีการหรือขั้นตอนที่ชัดเจน เนื้อหาที่ใช้เป็นเนื้อหาตามสาระการเรียนรู้คณิตศาสตร์ ระดับชั้นมัธยมศึกษาตอนต้น (ชั้นมัธยมศึกษาปีที่ 1-3) ตามหลักสูตรการศึกษาขั้น

พื้นฐาน พุทธศักราช 2551 เพื่อแสดงถึงความสามารถทางคณิตศาสตร์ จำนวน 12 ข้อ ข้อละ 4 คะแนน ให้คะแนนตามกฎเกณฑ์การให้คะแนนที่กำหนดไว้ ซึ่งบ่งชี้สัมฤทธิ์ผลในแต่ละระดับของผู้สอบ

9. การตรวจให้คะแนนแบบหลายค่า หมายถึง การกำหนดคะแนนรายข้อให้กับคำตอบของผู้สอบให้มีมากกว่า 2 ค่า โดยกำหนดคะแนนเป็น 5 ระดับคะแนนคือ 0, 1, 2, 3 และ 4 ตามระดับของความถูกต้องเหมาะสมจากน้อยไปหามาก ซึ่งเป็นไปตามกฎเกณฑ์การให้คะแนนที่กำหนดไว้

10. ผู้ตรวจ หมายถึง ครูผู้ตรวจให้คะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ที่มีวุฒิทางการศึกษาตั้งแต่ระดับปริญญาตรีขึ้นไปทางด้านคณิตศาสตร์ มีประสบการณ์ในการสอนวิชาคณิตศาสตร์ในระดับชั้นมัธยมศึกษาปีที่ 3 ไม่น้อยกว่า 5 ปี และได้รับการอบรมชี้แจงวิธีการตรวจและกฎเกณฑ์การให้คะแนน จนมีความเข้าใจตรงกัน

11. จำนวนผู้ตรวจ หมายถึง จำนวนครูผู้ตรวจให้คะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ ในการวิจัยครั้งนี้แบ่งจำนวนผู้ตรวจเป็น 3 ลักษณะคือ 1) จำนวนผู้ตรวจ 2 คน 2) จำนวนผู้ตรวจ 3 คน และ 3) จำนวนผู้ตรวจ 4 คน

12. รูปแบบการตรวจให้คะแนน หมายถึง ลักษณะการตรวจให้คะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ แบ่งเป็น 3 ลักษณะ คือ 1) ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน 2) ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน และ 3) ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน

กรอบแนวคิดในการวิจัย

การวิจัยในครั้งนี้เป็นการศึกษาลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน โดยใช้โมเดลการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราล์ซ ผู้วิจัยได้กำหนดกรอบแนวคิดการวิจัยไว้ 3 ประเด็น ดังนี้

1. จำนวนผู้ตรวจ

จำนวนผู้ตรวจเป็นปัจจัยหนึ่งที่ส่งผลต่อคุณลักษณะของคะแนนแบบทดสอบปลายเปิด ซึ่งจากการศึกษาของสัดวีคส์ รีฟและแบรดชอว์ (Sudweeks; Reeve; & Bradshaw. 2005) พบว่า ในการตรวจข้อสอบแบบทดสอบปลายเปิดหนึ่งครั้งควรให้ผู้ตรวจอย่างน้อย 3 คน และใช้ข้อสอบอย่างน้อย 3 ข้อ เมื่อเพิ่มจำนวนผู้ตรวจจะทำให้ค่าความเชื่อมั่นเพิ่มขึ้น และได้ให้ข้อควรคำนึงถึงการนำคำถามปลายเปิดไปใช้อยู่สองประการคือ ความเป็นไปได้จริงในเชิงปฏิบัติที่จะให้นักเรียนทำข้อสอบหลายข้อ และแต่ละคนได้รับการตรวจจากผู้ตรวจจำนวนมาก ทำให้ยังคงต้องการคำตอบที่ว่าจำนวนผู้ตรวจควรเป็นเท่าใดที่จะนำไปหาคะแนนเฉลี่ยเพื่อให้คะแนนมีความน่าเชื่อถือมากขึ้น ในขณะที่เบรนนอน เกาและคอลตัน (Brennan; Gao; & Colton. 1995) ได้ศึกษาปัจจัยที่มีอิทธิพลต่อค่าความเชื่อมั่นและความแม่นยำของการให้คะแนนจากการทดสอบทักษะการฟังและการเขียน พบว่า จำนวน

ผู้ตรวจมีอิทธิพลต่อความแม่นยำของการวัด หากต้องการความแม่นยำของการวัดต้องใช้ผู้ตรวจตั้งแต่ 2 คนขึ้นไป ทั้งนี้ไม่ควรใช้ผู้ตรวจเพียงคนเดียวเพราะจะทำให้ขาดความเที่ยงตรงในการศึกษา (Uebersax. 2003: online) ส่วนผลการวิจัยของสวาร์ทซ์และคณะ (Swartz; et al. 1999) พบว่า หากเพิ่มจำนวนผู้ตรวจจะทำให้ค่าประสิทธิภาพความเชื่อมั่นเพิ่มสูงขึ้น และต้องใช้ผู้ตรวจตั้งแต่ 4 คนขึ้นไป จึงจะทำให้สัมประสิทธิ์ความเชื่อมั่นของคะแนนอยู่ในช่วงที่ยอมรับได้ตามเกณฑ์ของนันนัลลี (Nunnally's criteria) คือ 0.90 นอกจากนี้แล้ว ชัยฤทธิ์ ศิลาเดช (2540) ได้กล่าวถึงความเชื่อมั่นของการประเมินคุณภาพของการใช้แฟ้มสะสมงานในการประเมินผลการเรียนไว้ว่า ไม่ควรใช้ผู้ประเมินเพียงคนเดียว เพราะจะทำให้ความเชื่อมั่นของคะแนนมีค่าค่อนข้างต่ำ จำนวนผู้ประเมินที่เหมาะสมคือ 2-3 คน และหากใช้ผู้ประเมินมากกว่า 4 คนจะทำให้ค่าความเชื่อมั่นเพิ่มขึ้นอีกเพียงเล็กน้อยเท่านั้น ดังนั้นจึงไม่จำเป็นต้องใช้จำนวนผู้ประเมินมากเกินไป

จากที่กล่าวมาข้างต้น จะเห็นได้ว่าการเพิ่มจำนวนผู้ตรวจจะทำให้ค่าความเชื่อมั่นเพิ่มสูงขึ้น (O'Brien; & Jones. 1986; Brennan; Gao; & Colton. 1995; Swartz; et al. 1999; Sudweeks; Reeve; & Bradshaw. 2005) แต่จำนวนผู้ตรวจที่เหมาะสมนั้นผลการศึกษายังพบแตกต่างกัน จากสภาพการเรียนการสอนในปัจจุบันที่ระดับชั้นหนึ่งจะมีครูผู้สอนวิชาคณิตศาสตร์ประมาณ 1-3 คน แตกต่างกันไปตามขนาดของระดับชั้นเรียน ดังนั้น ผู้วิจัยจึงกำหนดจำนวนผู้ตรวจที่ใช้ในการศึกษาครั้งนี้เป็น 3 ลักษณะที่สอดคล้องกับสภาพความเป็นจริงของการเรียนการสอน คือผู้ตรวจจำนวน 2 คน 3 คน และ 4 คน เพื่อศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อมีจำนวนผู้ตรวจใน 3 ลักษณะ โดยใช้โมเดลการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราล์ซ ซึ่งทำให้ได้จำนวนผู้ตรวจที่เหมาะสมที่ทำให้คะแนนที่ได้จากการตรวจมีความเชื่อมั่นและความเที่ยงตรง

2. รูปแบบการตรวจให้คะแนน

รูปแบบการตรวจให้คะแนนที่ยุติธรรมที่สุดคือการให้ผู้ตรวจทุกคนตรวจสอบทุกข้อของผู้สอบทุกคน เช่น ผู้สอบ 15 คน ทำข้อสอบวิชาหนึ่งคนละ 3 ข้อ มีผู้ตรวจ 3 คน การหาคะแนนที่ยุติธรรมคือให้ผู้ตรวจทั้ง 3 คน ตรวจข้อสอบทั้ง 3 ข้อ ของผู้สอบทุกคน แล้วคำนวณคะแนนเฉลี่ย แต่การตรวจลักษณะนี้ต้องใช้เวลาในการตรวจและมีค่าใช้จ่ายสูง ดังนั้นหากสามารถวางแผนการตรวจที่สามารถลดภาระ ลดเวลาในการตรวจได้โดยมีความเป็นไปได้ในทางปฏิบัติ แต่ยังคงไว้ซึ่งความน่าเชื่อถือของคะแนนและมีค่าทางสถิติที่ยอมรับได้ จึงเป็นทางหนึ่งในการแก้ปัญหาได้ (Linacre. 2002: 490-493; Nakamura. 2002: online) นั่นคือ ในการวางแผนการตรวจผู้ตรวจสามารถตรวจข้อสอบบางข้อโดยที่มีการเชื่อมโยง (Connected) ไขว้กัน (Crossed) ของผู้ตรวจและข้อสอบได้ โดยที่ยังคงมีความน่าเชื่อถือทางสถิติ (Linacre. 1994; Linacre; et al. 1994) ซึ่ง คิม (Kim. 2000: online)

ได้ให้ข้อเสนอแนะในการวางแผนการตรวจไว้ว่าควรมีการออกแบบรูปแบบการตรวจให้มีความเหมาะสมและเสียค่าใช้จ่ายน้อยที่สุด โดยใช้ผู้ตรวจหรือเงื่อนไขการวัดให้น้อยที่สุด แต่ยังมีอำนาจการทดสอบ (Power) มากที่สุด ดังนั้น รูปแบบการตรวจให้คะแนนสามารถออกแบบเป็นแบบไขว้ (Crossed) แบบแฝง (Nested) หรือแบบหลายองค์ประกอบ (Multifacet) ได้ สอดคล้องกับข้อเสนอแนะของมายฟอร์ดและวูล์ฟ (Myford; & Wolfe. 2003: 392) ที่กล่าวว่าไม่จำเป็นที่ผู้ตรวจต้องตรวจให้คะแนนผู้สอบทุกคนทุกข้อ อาจให้ผู้ตรวจตรวจข้อสอบของผู้สอบบางคนหรือตรวจข้อสอบในบางคุณลักษณะ (Trait) หรือข้อสอบบางข้อ และข้อเสนอแนะของไรมานีรัตและคณะ (Iramaneerat; et al. 2007: online) ที่ว่าควรมีการออกแบบเพื่อวางแผนการตรวจทั้งแบบแบ่งส่วน (Partially) หรือ แบบไขว้สมบูรณ์ (Full crossed) เพื่อตรวจสอบอิทธิพลขององค์ประกอบต่าง ๆ ที่มีต่อคุณลักษณะของคะแนนให้ครอบคลุมมากยิ่งขึ้น ทั้งนี้จากการศึกษาของสัตวิคส์ รีฟและแบรดชอว์ (Sudweeks; Reeve; & Bradshaw. 2005: 259) พบว่า การใช้แบบแผนแบบแฝง (Nested) โดยให้ผู้ตรวจแต่ละคนตรวจข้อสอบทุกข้อของนักเรียนที่ต่างกันนั้นค่าความเชื่อมั่นของคะแนนแตกต่างจากรูปแบบแบบไขว้สมบูรณ์ (Full crossed) เพียง 1.5 – 2.0 % เท่านั้น

ดังนั้นการวิจัยในครั้งนี้ ผู้วิจัยจึงออกแบบรูปแบบการตรวจให้คะแนนเป็น 3 ลักษณะ คือ 1) ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน 2) ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน และ 3) ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน เพื่อศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อมีรูปแบบการตรวจให้คะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ที่ต่างกัน 3 ลักษณะ โดยใช้โมเดลการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของวาร์ช ซึ่งทำให้ได้รูปแบบการตรวจที่สามารถลดภาระ แรงงาน เวลาและค่าใช้จ่ายในการตรวจแต่ยังคงไว้ซึ่งความเชื่อมั่นและความเที่ยงตรงของคะแนน

3. โมเดลที่ใช้ในการวิเคราะห์

จากการศึกษาวิธีการทางสถิติและการวัดผล พบว่า โมเดลของทฤษฎีการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของวาร์ชตามทฤษฎีการตอบข้อสอบ เป็นโมเดลที่สามารถนำมาใช้เพื่อการศึกษา ตรวจสอบแหล่งความแปรปรวนของความคลาดเคลื่อนได้หลายแหล่งพร้อมกัน โดยทฤษฎีการสรุปอ้างอิง (Generalizability Theory: G Theory) (Cronbach; et al. 1972; Shavelson; & Webb. 1991; Brennan. 2001) เป็นทฤษฎีที่สามารถศึกษาอิทธิพลของแหล่งความคลาดเคลื่อนได้หลายแหล่งที่ส่งผลต่อค่าความเชื่อมั่นของการสอบได้พร้อมกันจากการวิเคราะห์เพียงครั้งเดียว และไม่ยึดติดกับคุณสมบัติความเท่าเทียมกันของข้อสอบ มีเป้าหมายเพื่อประมาณค่าองค์ประกอบความแปรปรวน (Variance components) ของแหล่งต่าง ๆ ที่อยู่ในบริบทของการวัดและประมาณค่า

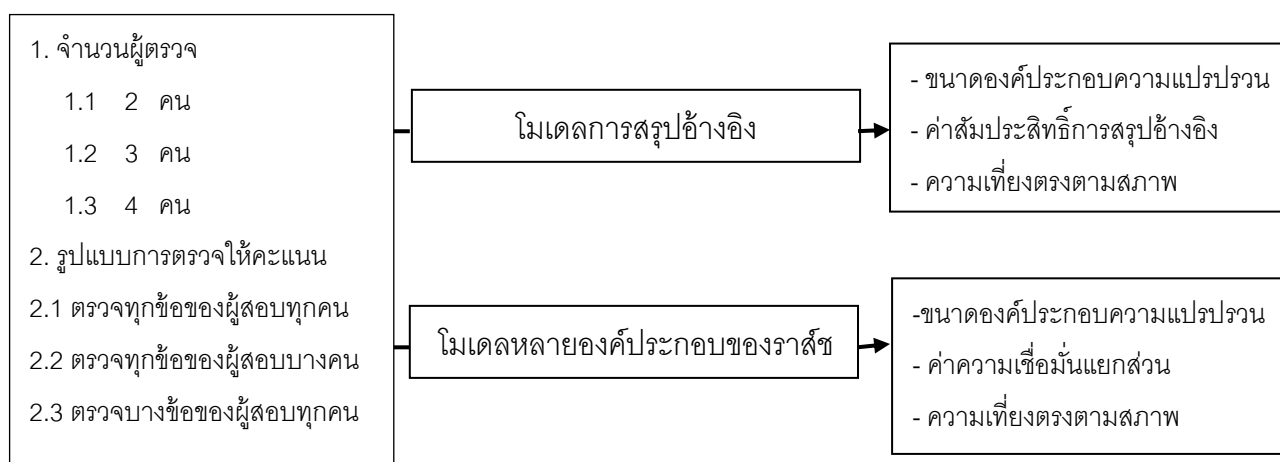
สัมประสิทธิ์การสรุปอ้างอิงในสถานการณ์ต่างๆ ของการวัดผล (Shavelson; & Webb. 1991; Bachman; Lynch; & Mason. 1993) ทฤษฎีการสรุปอ้างอิงจึงเป็นทฤษฎีที่มีประโยชน์ในการตรวจสอบคุณลักษณะของคะแนน โดยเฉพาะคะแนนจากการวัดประเมินที่มีความคลาดเคลื่อนหลายแหล่ง เช่น การวัดประเมินการปฏิบัติ แบบทดสอบปลายเปิด เป็นต้น โดยสามารถอธิบายองค์ประกอบต่าง ๆ ของการวัดประเมิน เช่น ผู้สอบ ข้อสอบและผู้ตรวจ ซึ่งเป็นแหล่งของความคลาดเคลื่อนหลักของการวัด ทฤษฎีนี้จึงเหมาะสมกับการวัดประเมินการปฏิบัติ (Brennan. 2001: 117) เช่น ทักษะการเขียน การพูด เป็นต้น จึงทำให้มีการนำทฤษฎีนี้ไปใช้ในการประมาณค่าองค์ประกอบความแปรปรวนและประมาณค่าสัมประสิทธิ์การสรุปอ้างอิงในสถานการณ์การวัดต่าง ๆ อย่างกว้างขวาง (Ibrahim. 1984; O'Brien; & Jones. 1986; Engelhard. 1992; 1994; Brennan; Gao; & Colton. 1995; Lane; et al. 1996; Swartz; et al. 1999) ผลการวิจัยเหล่านี้ต่างให้ผลที่สอดคล้องกันในประเด็นที่พบว่า ผู้ตรวจเป็นแหล่งความคลาดเคลื่อนขนาดใหญ่ของการวัดประเมิน จำนวนผู้ตรวจมีอิทธิพลต่อค่าสัมประสิทธิ์การสรุปอ้างอิง เมื่อเพิ่มจำนวนผู้ตรวจจะทำให้ค่าความเชื่อมั่นเพิ่มสูงขึ้น แต่จำนวนผู้ตรวจที่เหมาะสมนั้นผลการศึกษายังพบแตกต่างกัน อย่างไรก็ตามการศึกษาโดยใช้ทฤษฎีการสรุปอ้างอิงนั้นเป็นการศึกษาในระดับกลุ่มหรือโดยภาพรวม (Linacre. 1993; Stemler. 2004: online; Iramaneerat; et al. 2007: online) ไม่สามารถแสดงรายละเอียดได้ถึงระดับรายบุคคลหรือระดับสมาชิกในแต่ละองค์ประกอบได้

ในขณะที่โมเดลหลายองค์ประกอบของราสช์ (Many – Facet Rasch Model: MFRM) (Linacre. 1994) เป็นโมเดลที่ปรับขยายมาจากโมเดลของราสช์แบบ 1 พารามิเตอร์ จากเดิมที่สามารถประมาณค่าพารามิเตอร์ได้เพียง 2 ค่า คือ พารามิเตอร์ของผู้สอบและพารามิเตอร์ของข้อสอบ ทำให้สามารถประมาณค่าความเข้มงวดของผู้ตรวจ (พารามิเตอร์ของผู้ตรวจ) และองค์ประกอบอื่น ๆ ที่นำเข้าไปศึกษาในโมเดลได้ MFRM จะปรับแก้การวัดเชิงเส้นของคะแนนดิบของผู้สอบแต่ละคนจากผู้ตรวจและข้อสอบที่ใช้ พยายามให้การวัดของผู้สอบแต่ละคนเป็นอิสระจากผลของความแตกต่างของความเข้มงวดของผู้ตรวจและความยากของข้อสอบ (Linacre. 1994; Linacre & Wright. 2002) จุดเน้นของ MFRM คือ ศึกษาผู้สอบแต่ละคน ข้อสอบแต่ละข้อและผู้ตรวจแต่ละคน ข้อค้นพบของการวิเคราะห์ด้วย MFRM นั้นอยู่ที่แต่ละสมาชิกภายในแต่ละองค์ประกอบที่เป็นสาเหตุของความคลาดเคลื่อนในการวัดประเมิน (Iramaneerat; et al. 2007: online) หากสามารถอธิบายพฤติกรรมของผู้ตรวจและนำค่าความต่างนี้เข้าไปปรับแก้ค่าความสามารถของผู้สอบ จะทำให้คะแนนที่ได้มีความยุติธรรม (Fairness) และน่าเชื่อถือมากขึ้น (Linacre. 1992, 1994; Park. 2004: online) มีการศึกษาวิจัยและประยุกต์ใช้ MFRM อย่างหลากหลาย โดยเฉพาะการวัดประเมินที่ต้องใช้ผู้ประเมินหรือผู้ตรวจให้คะแนน ไม่ว่าจะเป็นการวัดประเมินทางด้านภาษา เช่น การพูด (Myford; & Wolfe. 2000: online; Bonk; & Ockey. 2003: online) การเขียน (Nakamura. 2002: online; Park. 2004:

online) การวัดประเมินแฟ้มสะสมงาน (Turner. 2003: online) เป็นต้น ผลการวิจัยเหล่านี้ต่างมีความสอดคล้องกันที่ว่า ผู้ตรวจมีความเข้มงวดแตกต่างกัน เมื่อปรับแก้คะแนนจากความแตกต่างของความเข้มงวดของผู้ตรวจแล้ว MFRM จะให้ค่าการวัดความสามารถของผู้สอบที่ปลอดจากความเห็นส่วนตัวของผู้ตรวจ ทำให้คะแนนที่ได้มีความน่าเชื่อถือ เทียบตรงและยุติธรรมมากขึ้น อีกทั้ง MFRM ยังสามารถวิเคราะห์ได้ถึงระดับแต่ละสมาชิกในองค์ประกอบ ทำให้สามารถใช้เป็นข้อมูลย้อนกลับ เพื่อการปรับปรุงพัฒนาระบบการวัดประเมินต่อไปได้

นอกจากนี้ยังมีการนำโมเดลการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราล์ชไปใช้ศึกษาร่วมกัน ในการวัดประเมินทักษะทางการพูด (Bachman; Lynch; & Mason.1993: online; Lynch; & McNamara. 1998: online) การเขียน (Sudweeks; Reeve; & Bradshaw. 2005) ทักษะการแก้ปัญหาที่ซับซ้อน (Smith; & Kulikowich. 2004) ตลอดจนถึงการทดสอบในวงการแพทย์ (Iramaneerat; et al. 2007: online) ซึ่งนักวิจัยเหล่านี้ต่างให้ความเห็นที่สอดคล้องกันว่า โมเดลทั้งสองเป็นโมเดลที่เติมเต็มซึ่งกันและกัน (Complementary) แสดงให้เห็นทั้งภาพรวม(ระดับกลุ่ม)และภาพย่อย (ระดับสมาชิก) จึงทำให้ได้สารสนเทศของการวัดประเมินที่สมบูรณ์มากขึ้น สามารถนำทั้งสองโมเดลมาวิเคราะห์คุณลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ที่ยังคงมีข้อสงสัยในคะแนนที่ได้และมีแหล่งความคลาดเคลื่อนหลายแหล่ง เพื่อให้สามารถอธิบายคะแนนได้อย่างชัดเจน ถูกต้องและน่าเชื่อถือมากขึ้น และเพื่อการพัฒนา ปรับปรุงกระบวนการวัดประเมินการปฏิบัติต่อไป ดังนั้น การวิจัยครั้งนี้ผู้วิจัยจึงเลือกโมเดลการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราล์ชมาใช้ในการศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน

จากที่กล่าวมาข้างต้นสามารถสรุปเป็นผังของกรอบแนวคิดในการทำวิจัยดังนี้



สมมติฐานของการวิจัย

การวิจัยในครั้งนี้ เป็นการศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน โดยใช้โมเดลการสรุปอ้างอิง และโมเดลหลายองค์ประกอบของราล์ซ ผู้วิจัยพิจารณาคุณลักษณะของคะแนนจากค่าสัมประสิทธิ์การสรุปอ้างอิง ค่าความเชื่อมั่นแยกส่วนและค่าความเที่ยงตรงตามสภาพ โดยเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิง ค่าความเชื่อมั่นแยกส่วน ค่าความเที่ยงตรงตามสภาพซึ่งเป็นค่าสหสัมพันธ์ของคะแนนที่ได้จากแบบทดสอบปลายเปิดกับคะแนนจากผลการทดสอบทางการศึกษาระดับชาตินัยพื้นฐาน (Ordinary National Educational Test: O-NET) เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน ซึ่งเป็นการเปรียบเทียบภายในแต่ละโมเดล ผู้วิจัยนำผลการศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้องมาเป็นแนวทางในการตั้งสมมติฐานของการวิจัย ผู้วิจัยตั้งสมมติฐานของการวิจัย จำแนกตามโมเดลที่ใช้ในการวิเคราะห์ ดังนี้

1. เมื่อวิเคราะห์คุณลักษณะของคะแนนแบบทดสอบปลายเปิด ที่มีจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน ด้วยโมเดลการสรุปอ้างอิง คาดว่า ค่าสัมประสิทธิ์การสรุปอ้างอิงและความเที่ยงตรงตามสภาพ จะมีค่าดังนี้

1.1 ค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนที่มีรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคนจะมีค่าสูงกว่ารูปแบบการตรวจให้คะแนนรูปแบบอื่น ในทุกจำนวนผู้ตรวจให้คะแนน

1.2 ค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนที่มีจำนวนผู้ตรวจที่มากกว่าจะมีค่าสูงกว่าจำนวนผู้ตรวจที่น้อยกว่า ในทุกรูปแบบการตรวจให้คะแนน

1.3 ค่าความเที่ยงตรงตามสภาพของคะแนนที่มีรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคนจะมีค่าสูงกว่ารูปแบบการตรวจให้คะแนนรูปแบบอื่น ในทุกจำนวนผู้ตรวจให้คะแนน

1.4 ค่าความเที่ยงตรงตามสภาพของคะแนนที่มีจำนวนผู้ตรวจที่มากกว่าจะมีค่าสูงกว่าจำนวนผู้ตรวจที่น้อยกว่า ในทุกรูปแบบการตรวจให้คะแนน

2. เมื่อวิเคราะห์คุณลักษณะของคะแนนแบบทดสอบปลายเปิด ที่มีจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน ด้วยโมเดลหลายองค์ประกอบของราล์ซ คาดว่า ค่าความเชื่อมั่นแยกส่วนของแต่ละองค์ประกอบ และความเที่ยงตรงตามสภาพ จะมีค่าดังนี้

2.1 ค่าความเชื่อมั่นแยกส่วนของผู้สอบและข้อสอบที่มีรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคนจะมีค่าสูงกว่ารูปแบบการตรวจให้คะแนนรูปแบบอื่น ส่วน

ค่าความเชื่อมั่นแยกส่วนของผู้ตรวจจะมีค่าต่ำกว่ารูปแบบการตรวจให้คะแนนรูปแบบอื่น ในทุกจำนวนผู้ตรวจให้คะแนน

2.2 ค่าความเชื่อมั่นแยกส่วนของผู้สอบและข้อสอบที่มีจำนวนผู้ตรวจที่มากกว่าจะมีค่าสูงกว่าจำนวนผู้ตรวจที่น้อยกว่า ส่วนค่าความเชื่อมั่นแยกส่วนของผู้ตรวจจะมีค่าต่ำกว่าจำนวนผู้ตรวจที่น้อยกว่า ในทุกรูปแบบการตรวจให้คะแนน

2.3 ค่าความเที่ยงตรงตามสภาพของคะแนนที่มีรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคนจะมีค่าสูงกว่ารูปแบบการตรวจให้คะแนนรูปแบบอื่น ในทุกจำนวนผู้ตรวจให้คะแนน

2.4 ค่าความเที่ยงตรงตามสภาพของคะแนนที่มีจำนวนผู้ตรวจที่มากกว่าจะมีค่าสูงกว่าจำนวนผู้ตรวจที่น้อยกว่า ในทุกรูปแบบการตรวจให้คะแนน



บทที่ 2

เอกสารและงานวิจัยที่เกี่ยวข้อง

การศึกษาวิจัยในครั้งนี้มีจุดมุ่งหมายเพื่อศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิด วิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน โดยใช้โมเดลการสรุปอ้างอิงตามทฤษฎีการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราล์ฟตามทฤษฎีการตอบข้อสอบ ผู้วิจัยได้ศึกษากรอบแนวคิด หลักการและทฤษฎีต่าง ๆ ตลอดจนงานวิจัยที่เกี่ยวข้อง เพื่อนำองค์ความรู้ที่ได้มาประยุกต์ใช้ในการศึกษาวิจัยครั้งนี้ แนวทางของการศึกษาค้นคว้า สามารถกำหนดเป็นกรอบในการศึกษาใน 6 ประเด็น และได้นำเสนอตามหัวข้อต่อไปนี้

1. คำถามปลายเปิดวิชาคณิตศาสตร์
2. กฎเกณฑ์การให้คะแนน
3. ทฤษฎีการสรุปอ้างอิง
4. โมเดลหลายองค์ประกอบของราล์ฟ
5. การทดสอบทางการศึกษาระดับชาตินิยมพื้นฐาน
6. งานวิจัยที่เกี่ยวข้อง

คำถามปลายเปิดวิชาคณิตศาสตร์

คำถามปลายเปิด (Becker; & Shimada. 1997: 1) เป็นข้อคำถามที่เปิดโอกาสให้มีคำตอบได้อย่างหลากหลายหรือมีคำตอบเดียวแต่มีวิธีคิดที่หลากหลายวิธี โดยจะไม่ถามผลลัพธ์หรือคำตอบที่ได้เพียงเท่านั้น แต่จะให้นักเรียนได้แสดงถึงวิธีการให้เหตุผลหรือการได้มาซึ่งคำตอบนั้นด้วย จึงเป็นการเปิดโอกาสให้นักเรียนได้แสดงทักษะการคิดคำนวณ การให้เหตุผล การสื่อสาร และการแก้ปัญหาในเชิงคณิตศาสตร์ได้อย่างหลากหลาย (Cai; et al. 1996) นอกจากนี้คำถามปลายเปิดยังช่วยให้นักเรียนสามารถแสดงความคิดและความเข้าใจในความคิดรวบยอดทางคณิตศาสตร์ได้ในหลายรูปแบบ ทั้งถ้อยคำหรือแผนภาพ (Foong. n.d.: online) เพื่อให้เกิดความเข้าใจในคำถามปลายเปิดมากยิ่งขึ้น ผู้วิจัยขอเสนอในประเด็นความหมายของคำถามปลายเปิด ประเภทของคำถาม ลักษณะของคำถาม การสร้างคำถามปลายเปิดและยกตัวอย่างคำถามปลายเปิด ข้อเสนอแนะในการนำคำถามปลายเปิดไปใช้และปัญหาของการใช้คำถามปลายเปิด มีรายละเอียดตามลำดับดังนี้

ความหมายของคำถามปลายเปิด

สเทนมาร์ค (Stenmark . 1991: 20) ได้กล่าวไว้ว่า คำถามปลายเปิดเป็นคำถามที่มีแนวทางเข้าสู่คำตอบได้อย่างหลากหลายและให้นักเรียนตอบได้อย่างหลากหลายวิธี

แฮนค็อก (Hancock. 1995) ได้กล่าวว่า คำถามปลายเปิด คือ คำถามที่มีวิธีการแก้ปัญหาที่ถูกต้องได้มากกว่าหนึ่งวิธี เปิดโอกาสให้นักเรียนใช้วิธีการที่หลากหลายของตนในการแก้ปัญหา

เบคเกอร์และชิมาดะ (Becker; & Shimada. 1997: 1) กล่าวว่า คำถามปลายเปิดเป็นคำถามที่สร้างขึ้นให้มีคำตอบที่ถูกต้องหลายคำตอบ โดยมีจุดมุ่งหมายเพื่อพัฒนาความหลากหลายของวิธีการหรือแนวทางการหาคำตอบ

คูนีและคณะ (Cooney; et al. n.d: online) ได้กล่าวว่า คำถามปลายเปิดคือ คำถามที่เปิดโอกาสให้มีคำตอบได้หลากหลาย เป็นไปตามความสามารถในการแก้ปัญหาที่แตกต่างกันตามกระบวนการคิดของแต่ละคน แสดงถึงความงดงามและการสร้างสรรค์ทางคณิตศาสตร์

ปรีชา เนาว์เย็นผล (2544: 27) ได้สรุปไว้ว่าคำถามปลายเปิด เป็นคำถามที่สร้างขึ้นให้มีคำตอบเปิดกว้าง มีคำตอบที่ถูกต้องหลายคำตอบหรือมีวิธีการหรือแนวทางการหาคำตอบได้หลายวิธี

จากความหมายของคำถามปลายเปิดที่มีนักการศึกษาได้ให้ไว้ สรุปได้ว่า คำถามปลายเปิดเป็นคำถามที่มีความหลากหลายของคำตอบหรือวิธีการได้มาซึ่งคำตอบ รวมถึงคำถามที่มีคำตอบเดียวแต่มีความหลากหลายในการให้เหตุผล โดยคำถามปลายเปิดที่ใช้ในการศึกษาวิจัยครั้งนี้ หมายถึงคำถามที่มีคำตอบที่ถูกต้องเพียงคำตอบเดียว แต่มีหลากหลายวิธีการได้มาซึ่งคำตอบ

ประเภทของคำถามปลายเปิด

โนตะ (ไมตรี อินทร์ประสิทธิ์. 2547: 6-8; อ้างอิงจาก Nohda. 1983) ได้ขยายแ่งมุมในการพิจารณาความเปิดเพิ่มขึ้น โดยได้แบ่งปัญหาปลายเปิดออกเป็น 3 ประเภท คือ

1. กระบวนการเปิด (Process is open) ปัญหาชนิดนี้มีแนวทางในการแก้ปัญหาซึ่งเป็นปัญหาด้นกำเนิดที่กำหนดได้อย่างหลากหลาย ซึ่งแน่นอนว่าปัญหาคณิตศาสตร์ทุกปัญหาต่างเป็นปัญหาปลายเปิดโดยนัยนี้ แต่ในโรงเรียนทั่วไปมักจะเน้นพิจารณาคำตอบเพียงคำตอบเดียวรวมทั้งไม่ได้เน้นแ่งมุมเชิงกระบวนการ ดังนั้นในปัญหาปลายเปิดชนิดนี้จึงมีการระบุคำตอบเพื่อให้นักเรียนได้พยายามหาแนวทางในการแก้ปัญหาให้ได้หลากหลาย เช่น จงหาคำตอบด้วยวิธีการที่แตกต่างกัน... วิธี เป็นต้น แนวทางการแก้ปัญหาที่หลากหลายนั้นทำให้นักเรียนดำเนินกิจกรรมไปได้ตามความสามารถและความสนใจและโดยอาศัยการอภิปรายกลุ่มจะทำให้นักเรียนมีกระบวนการแก้ปัญหาที่ดีกว่าเดิม

2. ผลลัพธ์เปิด (End product are open) ปัญหาปลายเปิดชนิดนี้มีคำตอบที่ถูกต้องหลากหลาย

3. แนวทางการพัฒนาปัญหาเปิด (Way to develop are open) หลังจากที่นักเรียนได้แก้ปัญหาไปแล้ว นักเรียนสามารถพัฒนาไปเป็นปัญหาใหม่ด้วยการเปลี่ยนแปลงเงื่อนไขหรือองค์ประกอบของปัญหาเดิม การเน้นแง่มุมนี้จะเรียกว่า “จากปัญหาสู่ปัญหา” ถือได้ว่าเป็นแนวทางการพัฒนาปัญหาปลายเปิด ด้วยแนวทางนี้ นักเรียนสามารถสนุกกับการตั้งปัญหาด้วยตนเอง ยิ่งไปกว่านั้นจากการเปรียบเทียบกับเพื่อนๆ นักเรียนสามารถอภิปรายเกี่ยวกับโครงสร้างทางคณิตศาสตร์ของปัญหาและความเป็นกรณีทั่วไปของแนวทางคำตอบที่นักเรียนคิดได้

ในขณะที่ยาทากาฮาชิ (Takahashi. 2000: online) ได้แบ่ง คำถามปลายเปิดเป็น 2 ประเภท คือ

1. คำถามปลายเปิดที่มีคำตอบเดียวแต่มีความหลากหลายในวิธีการให้เหตุผลหรือการได้มาซึ่งคำตอบ

2. คำถามปลายเปิดที่มีความหลากหลายทั้งคำตอบและวิธีการได้มาซึ่งคำตอบ ส่วนฟูง (Foong. 2008: online) ได้กล่าวถึงประเภทของคำถามปลายเปิดไว้ดังนี้

1. คำถามที่ไม่กำหนดวิธีการแก้ปัญหาตายตัว
2. คำถามที่ไม่กำหนดคำตอบตายตัว นั่นคือ มีคำตอบที่เป็นไปได้หลากหลาย
3. คำถามที่มีหลากหลายวิธีในการแก้ปัญหา ความสามารถของเด็กที่ต่างกันสามารถแก้ปัญหาด้วยวิธีการของตนเองได้
4. คำถามที่เปิดโอกาสให้เด็กได้ตัดสินใจแก้ปัญหาด้วยตนเอง ตามกระบวนการคิดทางคณิตศาสตร์ของตน
5. คำถามที่เปิดโอกาสให้มีการพัฒนาทักษะการให้เหตุผลและทักษะการสื่อสาร
6. คำถามที่เปิดโอกาสให้ใช้ความคิดสร้างสรรค์และจินตนาการที่เกี่ยวข้องกับบริบท หรือประสบการณ์ในชีวิตจริง

ลักษณะของคำถามปลายเปิด

มากอนและคณะ (Magone; et al. 1994: 317-340) ได้สร้างข้อคำถามปลายเปิดวิชาคณิตศาสตร์ โดยวัดใน 3 ด้าน คือ

1. ด้านความเข้าใจในความคิดรวบยอดทางคณิตศาสตร์ เช่น การวัดประเมิน ความเข้าใจในความคิดรวบยอดเกี่ยวกับค่าประจำตำแหน่งของทศนิยม โดยให้นักเรียนเลือกจำนวนที่มีค่ามากที่สุดระหว่าง .08 .8 .080 .008000 พร้อมอธิบายเหตุผลการประกอบการเลือกตอบ
2. ด้านการแก้โจทย์ปัญหา เช่น การประเมินความสามารถในการใช้ความรู้พื้นฐานทางจำนวนและตัวเลขในการแก้ปัญหา ซึ่งมีสถานการณ์คือ

“ยลดาบอกพี่ชายว่า วันนี้อาจารย์ให้แบ่งแท่งไม้ ซึ่งถ้าแบ่งแท่งไม้กลุ่มละ 2 แท่ง จะเหลือแท่งไม้ 1 แท่ง แบ่งกลุ่มละ 2 แท่งจะเหลือแท่งไม้ 1 แท่ง และถ้าแบ่งกลุ่มละ 4 แท่ง จะเหลือแท่งไม้ 1 แท่งเช่นกัน จงหาว่ายลดา มีแท่งไม้ทั้งหมดกี่แท่ง”

3. ด้านการให้เหตุผล เช่น การให้เหตุผลเพื่อบอกหลักเกณฑ์ในการสร้างรูปภาพ รูปถัดไป จากรูปแบบที่กำหนดให้

ส่วน เบคเกอร์และชิมาดา (Becker; & Shimada. 1997: 27) ได้แบ่งคำถามปลายเปิดออกเป็น 3 ลักษณะ ดังนี้

1. คำถามเกี่ยวกับการหาความสัมพันธ์ (Finding Relations) คำถามประเภทนี้จะให้นักเรียน ค้นหาหรือความสัมพันธ์ทางคณิตศาสตร์ เช่น “87 เป็นจำนวนที่อยู่ในพจน์ใดพจน์หนึ่งของลำดับ 3, 10, 17, 24, 31, ... หรือไม่ เพราะเหตุใด”

2. คำถามเกี่ยวกับการจำแนก (Classifying) เป็นคำถามที่ให้นักเรียนได้จำแนกหรือแยกแยะ สิ่งต่างๆ ที่มีลักษณะแตกต่างกันโดยใช้เกณฑ์ของนักเรียนเอง เพื่อนำไปสู่การสร้างมโนคติทาง คณิตศาสตร์ เช่น “จากสมการต่อไปนี้ $y = 3$, $y = x + 3$, $y = x - 1$, $y^2 + x^2 = 16$, $y^2 = 3x + 8$, $4y^2 + 9x^2 = 36$, $2x + 4 = 3$, $4y^2 - 25x^2 = 100$, $x = 0$ ให้นักเรียนจัดกลุ่มของสมการ พร้อม กับบอกเกณฑ์ที่ใช้”

3. คำถามที่เกี่ยวกับการวัด (Measuring) การประเมินหรือการประมาณ คำถามลักษณะนี้ จะให้นักเรียนได้กำหนดการวัดเชิงตัวเลขให้กับสถานการณ์ เป็นการฝึกให้นักเรียนได้ใช้คณิตศาสตร์ ประกอบการตัดสินใจ ซึ่งคำถามประเภทนี้คาดหวังให้นักเรียนประยุกต์ใช้ความรู้และทักษะทาง คณิตศาสตร์ที่ได้เรียนรู้มาใช้ในการแก้ปัญหา เช่น “ในการแข่งรถจักรยานยนต์ มีทีมเข้าร่วมการ แข่งขันทั้งหมด 4 ทีมคือทีม A, B, C และ D แต่ละทีมประกอบด้วยสมาชิก 5 คน ปราบกฏผลการแข่งขัน ดังตารางที่กำหนดให้ จงหาว่าทีมใดเป็นผู้ชนะ พร้อมให้เหตุผลประกอบการตัดสินใจ”

ลำดับที่	1	2	3	4	5	6	7	8	9	10
ทีม	A	C	D	C	A	B	B	D	D	C
ลำดับที่	11	12	13	14	15	16	17	18	19	20
ทีม	B	C	A	B	D	D	B	A	A	C

เกลทเซอร์และเกลทเซอร์ (Glatzer; & Glatzer. 2005: IX-X) ได้กล่าวถึงคำถามปลายเปิด ว่ามีลักษณะที่ถามในประเด็นต่อไปนี้

1. คำถามที่ให้แสดงกระบวนการหรือวิธีการ เช่น ให้หาพื้นที่พร้อมทั้งอธิบายวิธีการ/กระบวนการในการหาคำตอบ

2. คำถามที่ให้อธิบายเหตุผล ถึงวิธีการหรือคำตอบที่กำหนดมาให้ว่าถูกต้องหรือไม่ถูกต้อง เพราะเหตุใด เช่น ค่าเฉลี่ยของนักเรียนทั้งห้อง 20 คน เป็น 80 คะแนนในขณะที่อีกห้องมีนักเรียน 30 คนมีค่าเฉลี่ยเป็น 70 คะแนน มีขสสรุปว่าค่าเฉลี่ยของคะแนนนักเรียนทั้งหมด 50 คนเป็น 75 คะแนน โดยการนำ 80 รวมกับ 70 แล้วหารด้วย 2 วิธีการของมิชถูกต้องหรือไม่ จงอธิบาย

3. คำถามที่ต้องการคำตอบที่เป็นไปได้หรือสอดคล้องกับเงื่อนไขที่กำหนด เช่น ให้หาจำนวนหรือหาสมการต่าง ๆ ที่เป็นไปได้ เช่น ค่าเฉลี่ยของ 79, 80 และ 81 คือ 80 จงหาอีกสามจำนวนที่มีค่าเฉลี่ยเท่ากับ 80

4. คำถามที่ให้วาดภาพหรือไดอะแกรมที่สอดคล้องกับเงื่อนไข เช่น ให้วาดรูปสามเหลี่ยมสี่เหลี่ยมและห้าเหลี่ยม ที่มีพื้นที่เท่ากับ 12 ตารางนิ้ว

5. คำถามที่ให้อธิบายและ/หรือบอกหลักเกณฑ์ในการค้นหา สร้างรูปภาพ หรือรูปถัดไปจากรูปแบบที่กำหนดให้ เช่น รูปที่ 100 หรือจำนวนที่ 100 คือเท่าใด

6. คำถามที่ให้คาดเดาว่าจะเกิดสิ่งใดขึ้นหากมีการเปลี่ยนแปลงสถานการณ์ที่เป็นอยู่ เช่น ปริซึมทรงสี่เหลี่ยมมุมฉาก กว้าง 8 ซม. ยาว 4 ซม. และสูง 2 ซม. ถ้าเพิ่มขนาดของทั้งสามด้านเป็นสองเท่าแล้วปริมาตรของปริซึมนี้จะมีการเปลี่ยนแปลงอย่างไร

7. คำถามที่ให้วาดไดอะแกรมประกอบการอธิบาย

8. คำถามที่ให้แสดงกระบวนการวัด (Measure) ตามสิ่งที่กำหนด

ในขณะที่ คูนีและคณะ (Cooney; et al. n.d: online) ได้กล่าวถึงคำถามปลายเปิดว่าเป็นข้อคำถามที่มีลักษณะดังนี้

1. คำถามที่ให้นักเรียนคิดแก้ปัญหาหรือยกตัวอย่างที่สอดคล้องกับเงื่อนไขที่กำหนด เช่น ให้หาค่าของ a และ b ที่ทำให้สมการ $\sqrt{a} = 2\sqrt{b}$ นี้เป็นจริง เพราะเหตุใด จงอธิบาย หรือ จงหาว่ามีจำนวนตรรกยะที่ยกกำลังสองแล้วมีค่าน้อยกว่าตนเองหรือไม่ จงอธิบายและยกตัวอย่าง

2. คำถามที่ให้นักเรียนอธิบายว่าใครตอบถูกต้องและเพราะเหตุใด เช่น

- มีนักเรียนตอบคำถามเมื่อให้หาค่า 2^8 ดังนี้

มิเชลล์ ตอบว่า $2^8 = 2 + 2 + 2 + 2 + 2 + 2 + 2 + 2 = 16$

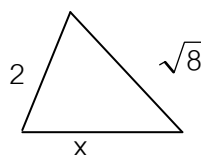
เดมอน ตอบว่า $2^8 = 2^2 \cdot 2^2 \cdot 2^2 \cdot 2^2 = 256$

ดาร์วิน ตอบว่า $2^8 = 2^2 \cdot 2^2 \cdot 2^2 = 64$

นักเรียนคนใดตอบถูกต้อง จงอธิบาย

- เคนท์ คำนวณหาค่า $\tan \theta$ และ $\sin \theta$ และสรุปว่า $\tan \theta < \sin \theta$ แต่वलล์กล่าวว่า เป็นไปไม่ได้ที่ $\tan \theta < \sin \theta$ จงหาว่าใครตอบถูกต้องและเพราะเหตุใด

3. คำถามที่ถามวิธีการแก้ปัญหาหรืออธิบายกระบวนการแก้ปัญหาที่มีตั้งแต่ 2 วิธีขึ้นไป เช่น ให้หาจำนวนจริง x สองจำนวนที่แตกต่างกันที่เป็นไปได้ในการสร้างรูปสามเหลี่ยมเมื่อกำหนด ความยาวให้ ดังรูป และอธิบายว่าทำไมค่าของ x ที่หาได้จึงทำให้เกิดรูปสามเหลี่ยมได้



ส่วน ฟุง (Foong, 2008: online) ได้กล่าวถึงคำถามปลายเปิดไว้ว่ามี 5 ลักษณะ ดังนี้

1. คำถามที่ให้แก้ปัญหาโดยให้ข้อมูลไม่สมบูรณ์ หรือขาดหาย
2. คำถามที่ให้อธิบายในทัศน์ กระบวนการ หรือความผิดพลาด
3. คำถามที่ใช้ปัญหาจากคำตอบหรือกระบวนการแก้ปัญหานักเรียนที่ตอบมา
4. คำถามที่เป็นปัญหาในชีวิตจริงหรือเป็นจริงในทางปฏิบัติ
5. คำถามที่ให้ค้นหา เปรียบเทียบ ได้แย้ง จำแนก ทดสอบสมมติฐานและการสรุป

คำถามปลายเปิดจึงเป็นข้อคำถามที่ให้นักเรียนได้อธิบายกระบวนการ แสดงเหตุผล สนับสนุนหรือโต้แย้ง ค้นหาหรือสูตร จำแนกหรือยกตัวอย่าง ซึ่งอาจหลากหลายทั้งคำตอบและ กระบวนการได้มาซึ่งคำตอบ ดังนั้นคำตอบหรือกระบวนการที่ได้จึงอาจแตกต่างกันตามกระบวนการ คิดของแต่ละคน อันแสดงถึงความรู้ความเข้าใจทางคณิตศาสตร์

การสร้างคำถามปลายเปิด

คำถามปลายเปิดเป็นคำถามที่เปิดโอกาสให้นักเรียนได้แสดงความคิดของตนเองได้อย่าง หลากหลายทั้งคำตอบและวิธีการได้มาซึ่งคำตอบ ซึ่งบางครั้งคำตอบที่ได้ อาจไม่ตรงกับสิ่งที่ต้องการ ประเมิน ซึ่ง เบคเกอร์และชิมาดา (Becker; & Shimada, 1997: 28-31) ได้กล่าวว่าเป็นสิ่งที่ยากใน การสร้างคำถามปลายเปิดที่ดีและเหมาะสมกับระดับความสามารถของนักเรียนที่แตกต่างกัน แต่ได้ให้ ข้อเสนอแนะในการสร้างคำถามปลายเปิดไว้ดังนี้

1. เตรียมสถานการณ์เชิงกายภาพที่เกี่ยวข้องกับตัวแปรเชิงปริมาณที่สามารถสังเกต ความสัมพันธ์ได้
2. เปลี่ยนคำถามจากเดิมที่ให้พิสูจน์ทฤษฎีบทในรูป “ถ้า P แล้ว Q” เปลี่ยนเป็น “ถ้า P แล้ว ความสัมพันธ์ของสิ่งต่าง ๆ ที่นักเรียนพบมีอะไรบ้าง” โดยจะต้องมีการกำหนดขอบเขตของ “สิ่งต่าง ๆ” ให้เฉพาะเจาะจงมากขึ้น

3. ในการสอนเกี่ยวกับทฤษฎีบท ควรเริ่มต้นด้วยการยกตัวอย่างหลาย ๆ ตัวอย่างก่อน เพื่อให้นักเรียนได้สร้างข้อสรุปด้วยตนเอง

4. แสดงรายการที่เป็นลำดับหรือตารางของข้อมูลต่าง ๆ เพื่อให้นักเรียนได้ค้นหาความสัมพันธ์หรือกฎทางคณิตศาสตร์

5. ใช้ตัวอย่างจริงเพื่อให้นักเรียนได้เห็นภาพ

6. แสดงคำถามที่มีลักษณะคล้ายกันหลาย ๆ คำถามเพื่อให้นักเรียนได้หาคำตอบและหาสมบัติที่คำถามเหล่านั้นมีร่วมกัน

7. จัดสถานการณ์กึ่งคณิตศาสตร์ (Quasi-Mathematics) ซึ่งเป็นสถานการณ์ที่นักเรียนสามารถนำคณิตศาสตร์ไปใช้ในการช่วยอธิบายได้

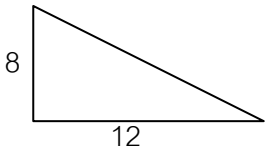
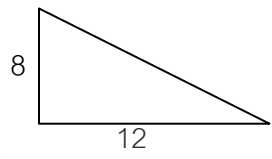
8. แสดงตัวอย่างที่ชัดเจนของโครงสร้างทางพีชคณิต โดยแสดงตัวอย่างที่เป็นข้อมูลเชิงตัวเลขที่ง่ายในการพิจารณาเพื่อให้นักเรียนได้ค้นหากฎทางคณิตศาสตร์

คูนียและคณะ (Cooney; et al. n.d.: online) ซานเชสและไอซ์ (Sanchez; & Ice. 2004: online) ฟุง (Foong. 2008: online) และปรีชา เนาว์เย็นผล (2544: 28) ได้ให้ข้อเสนอแนะไว้ว่าในการสร้างคำถามปลายเปิดสำหรับนำมาใช้ในกิจกรรมการเรียนการสอน แนวทางหนึ่งที่สามารถทำได้ คือ การปรับคำถามปลายปิดที่มีอยู่ให้เป็นคำถามปลายเปิด โดยใช้วิธีการบางอย่าง เช่น ตัดเงื่อนไขบางประการออกไป การย้ายคำถาม การเพิ่มข้อมูลที่ไม่จำเป็นเข้าไปในคำถาม เป็นต้น ซึ่งแสดงตัวอย่างการปรับคำถามไว้ดังตาราง 1

ตาราง 1 การเปลี่ยนคำถามปลายปิดให้เป็นคำถามปลายเปิด

คำถามปลายปิด	คำถามปลายเปิด
1. จำนวนใดต่อไปนี้ เป็นจำนวนเฉพาะ 7, 57, 67, 117	1. “เฟรดคิดว่า 57 กับ 67 เป็นจำนวนเฉพาะ เนื่องจากจำนวนทั้งสองลงท้ายด้วย 7 และจำนวนที่ลงท้ายด้วย 7 เป็นจำนวนเฉพาะ แต่ดิคบอกว่าที่เฟรดคิดนั้นไม่ถูกต้อง” ถามว่าใครกล่าวได้ถูกต้อง เพราะเหตุใด
2. จงหาจำนวนถัดไปอีก 3 จำนวนจาก ลำดับ 1, 4, 7, 10, 13, ...	2. 100 เป็นพจน์ใดพจน์หนึ่งของลำดับ 1, 4, 7, 10, 13, ... นี้หรือไม่ เพราะเหตุใด
3. จงหา ค.ร.น. ของ 18 และ 24	3. เพราะเหตุใด 48 จึงไม่เป็น ค.ร.น. ของ 18 และ 24
4. $(2 + 6) - 3 = [\quad]$	4. สร้างจำนวนใดได้บ้างจาก 2, 3 และ 6

ตาราง 1 (ต่อ)

คำถามปลายปิด	คำถามปลายเปิด
5. $3 \times 5 = []$ 6. จงหาพื้นที่ของรูปสามเหลี่ยม 	5. จงสร้างคำถามที่มีคำตอบเป็น 15 6. จงสร้างรูปสามเหลี่ยมให้มีพื้นที่เท่ากับพื้นที่ของรูปสามเหลี่ยมรูปนี้ 
7. เราเรียกรูปที่มีห้าด้านว่ารูปอะไร 8. จงเขียนกราฟของ $y = 3x + 5$, $y = 2x - 1$, $y = 7 - x$ 9. มีตุ๊กตา 12 ตัว จัดใส่ถุง ๆ ละ 3 ตัว จัดได้กี่ถุง 10. หมีโพลาร์มีน้ำหนักเป็น 20 เท่าของ อาลี ถ้าอาลีมีน้ำหนัก 25 กก. จงหา น้ำหนักของหมีโพลาร์ 11. จงหาเศษส่วนอย่างต่ำของ $\frac{58}{87}$	7. เราสามารถสร้างรูปเรขาคณิตอะไรได้บ้างจากส่วนของเส้นตรง 5 เส้น 8. จงศึกษากราฟของ $y = ax + b$ สำหรับค่า a และ b ต่าง ๆ 9. มีตุ๊กตา 12 ตัว จัดใส่ถุง ถุงละเท่า ๆ กัน ได้กี่ถุง ถุงละกี่ตัว 10. ถ้าหมีโพลาร์มีน้ำหนัก 500 กก. จะต้องใช้นักเรียนในห้องนี้กี่คนจึงจะมีน้ำหนักรวมเท่ากับน้ำหนักของหมีโพลาร์นี้ 11. จงหาเศษส่วนที่มีเศษส่วนอย่างต่ำเป็น $\frac{2}{3}$ ที่ต่างกันสามจำนวน

ที่มา: คูณีย์ และคณะ. (Cooney; et al. n.d.: online) ชันเชสและไอซ์. (Sanchez; & Ice. 2004: online) ฟุง. (Foong. 2008: online) และ ปรีชา เนาว์เย็นผล. (2544: 28)

เบคเกอร์และชิมาดะ (Becker; & Shimada .1997: 31–32) ได้ให้ข้อเสนอแนะว่าการนำคำถามปลายเปิดไปใช้ควรพิจารณาถึงประเด็นดังต่อไปนี้

1. คำถามนั้นควรมีคุณค่าทางคณิตศาสตร์ โดยคำถามที่ใช้นอกจากกระตุ้นให้นักเรียนได้คิดจากมุมมองที่แตกต่างกันแล้วควรมีคุณค่าในเชิงเนื้อหาคณิตศาสตร์ กล่าวคือ นักเรียนทั้งที่มี

ผลสัมฤทธิ์ทางการเรียนสูงหรือต่ำ สามารถแก้ปัญหาได้โดยใช้ความรู้ทางคณิตศาสตร์ที่ตนเองมีอยู่ ซึ่งอาจจะใช้วิธีการที่แตกต่างกันแต่ในแต่ละวิธีการนั้นยังคงมีคุณค่าทางคณิตศาสตร์

2. ระดับของความรู้คณิตศาสตร์ที่ต้องใช้ในการตอบคำถามควรเหมาะสมกับระดับความรู้ของนักเรียน ครูควรเลือกใช้คำถามที่เหมาะสมกับพื้นฐานความรู้ของนักเรียนเพราะเมื่อนักเรียนต้องตอบคำถามปลายเปิดนั้นเขาอาจต้องใช้ความรู้และทักษะทางคณิตศาสตร์ที่ได้เรียนมาแล้ว

3. คำถามนั้นเมื่อใช้แล้วสามารถนำไปสู่การพัฒนาเชิงคณิตศาสตร์ได้ กล่าวคือ คำตอบที่เป็นไปได้ของคำถามปลายเปิดนั้นควรมีบางคำตอบที่สามารถเชื่อมโยงหรือสัมพันธ์กับโมเดลทางคณิตศาสตร์ในระดับที่สูงขึ้น หรือสามารถพัฒนาระดับการคิดเชิงคณิตศาสตร์ที่สูงขึ้นได้

4. ควรเป็นคำถามที่นักเรียนสามารถตอบได้อย่างหลากหลายทั้งวิธีการและคำตอบ ทั้งนี้เนื่องจากนักเรียนแต่ละคนย่อมมีความคิดที่แตกต่างกัน และควรเปิดโอกาสให้นักเรียนได้สื่อสาร เพื่อแสดงถึงความเข้าใจในเรื่องนั้น ๆ ได้อย่างอิสระและเต็มความสามารถ

5. เป็นคำถามที่เปิดโอกาสให้นักเรียนได้สื่อสารความคิดของตนเอง เพราะเมื่อใดที่นักเรียนได้สื่อสารความคิดหรือเหตุผลของตนเองแล้ว ครูสามารถรับรู้ได้ว่านักเรียนมีความรู้ความเข้าใจและสามารถประยุกต์ใช้ความรู้ได้อย่างไรบ้าง

6. คำถามควรมีความชัดเจนว่าต้องการให้นักเรียนทำหรือแสดงอะไร เมื่ออ่านคำถามแล้วควรจะคาดเดาได้ว่าคำตอบลักษณะใดที่เป็นคำตอบที่เหมาะสมและตรงกับความต้องการของครู

ในขณะที่ คูนี และคณะ (Cooney; et al. online) ได้กล่าวว่า คำถามปลายเปิดต้องการให้นักเรียนได้สื่อสารกระบวนการคิดทางคณิตศาสตร์ ในการนำไปใช้ควรคำนึงถึงประเด็นต่อไปนี้

1. คำถามควรมีนัยสำคัญทางคณิตศาสตร์ (Involve significant mathematic) นั่นคือควรเป็นการถามถึงสิ่งที่มีคุณค่าและสิ่งที่สำคัญ เปิดโอกาสให้นักเรียนได้แสดงความรู้ ความเข้าใจ ผ่านประเด็นทางคณิตศาสตร์ และการใช้คณิตศาสตร์ในโลกของชีวิตจริง

2. ควรเป็นคำถามที่ดึงดูดการตอบสนองออกมา (Elicit a range of response) ข้อคำถามควรถามเพื่อให้นักเรียนได้อธิบายกระบวนการคิดของตน กระตุ้นให้เกิดการตอบสนองที่หลากหลายเพราะนักเรียนทุกคนไม่ได้คิดเหมือนกัน เช่น เมื่อถามว่า “สามเหลี่ยมด้านเท่าเป็นสามเหลี่ยมมุมฉากหรือไม่ เพราะเหตุใด” หากนักเรียนมุ่งความสนใจไปที่เรื่องมุม ก็จะตอบว่า “เป็นไปไม่ได้ เพราะมุมทุกมุมของสามเหลี่ยมด้านเท่าจะต้องเท่ากันและมุมภายในสามเหลี่ยมไม่สามารถเป็น 270 องศาได้” หากมีนักเรียนคนหนึ่งตอบว่า “หากเราสนใจที่ความยาวของด้านจะพบว่า เป็นไปไม่ได้เช่นกัน เพราะถ้าเป็นสามเหลี่ยมมุมฉากจะมีด้านหนึ่งที่ยาวที่สุด แต่สามเหลี่ยมด้านเท่าทุกด้านต้องยาวเท่ากัน ดังนั้นจึงเป็นไปไม่ได้” ซึ่งทำให้เกิดการแลกเปลี่ยนประสบการณ์ในชั้นเรียนที่นอกเหนือไปจากวิธีการคิดเพียงวิธีเดียวในการแก้ปัญหา

3. คำถามนั้นต้องการการสื่อสาร (Require communication) ในการใช้คำถามปลายเปิด ต้องเปิดโอกาสให้นักเรียนได้สื่อสารความคิด เพื่อแสดงถึงความเข้าใจในสิ่งที่นักเรียนรู้และนำไปใช้ในการแก้ปัญหา

4. คำถามต้องมีความชัดเจน (Be clear stated) และมีเป้าหมายที่แน่นอน แม้ว่าจะมีการตอบได้หลากหลายก็ตาม เพื่อให้นักเรียนทราบสิ่งที่ตนเองต้องทำและสิ่งที่ครูคาดหวังหรือความสมบูรณ์ในการตอบ ครูบางท่านอาจนำคำตอบทั้งหมดไปแสดงแก่นักเรียนแล้วให้นักเรียนเลือกคำตอบที่เหมาะสม เนื่องจากนักเรียนไม่ค่อยใช้การอธิบายความคิดโดยการเขียน จะสามารถช่วยพัฒนาทักษะการสื่อสารและความสามารถในการวิเคราะห์ว่าจะเขียนนำเสนอเหตุผลอย่างไร

5. มีกฎเกณฑ์การให้คะแนน (Lend itself to a scoring rubric) การวัดประเมินส่วนใหญ่ จะมี 2 เกณฑ์คือ ตอบถูกหรือตอบผิด แต่ในคำถามปลายเปิดจะให้นักเรียนมีโอกาสสื่อสารความเข้าใจบางอย่างมากกว่าการตอบถูกหรือผิด การให้กฎเกณฑ์การให้คะแนนที่เป็นบางส่วน (Partial credit) จึงทำให้มั่นใจว่ามีความยุติธรรม

ปัญหาของการใช้คำถามปลายเปิด

คำถามปลายเปิดมีคุณค่าในการวัดผลหลายประการ แต่อย่างไรก็ตาม ยังมีปัญหาในการใช้คำถามปลายเปิดอยู่หลายประการ ดังเช่น สจิมล ดิรกันนัท (2550: 118-119) เยาวดี ราชัยกุล วิบูลย์ศรี (2553: 244-245) ได้กล่าวถึงปัญหาของข้อสอบปลายเปิดหรือข้อสอบที่ให้ผู้ตอบเขียนตอบไว้สอดคล้องกันว่ามีปัญหาในประเด็นต่อไปนี้

1. ปัญหาในการตรวจให้คะแนน ได้แก่ ผลการตรวจข้อสอบของผู้สอบคนแรกมีผลต่อผู้สอบคนต่อ ๆ มา เช่น ผู้สอบคนแรกเขียนไม่เก่ง คนที่สองปานกลางแต่จะได้คะแนนมากเกินจริง หรือ คนแรกเขียนเก่ง คนที่สองปานกลางจะได้คะแนนน้อยกว่าความจริง และการให้คะแนนมักมีความเชื่อมั่นต่ำ (reliability) เพราะมีปัจจัยที่มีอิทธิพลต่อคะแนนได้มาก เช่น ความเห็นส่วนตัว อารมณ์ของผู้ตรวจ ความลำเอียง การใช้ภาษาเขียน ลำดับของข้อความ เป็นต้น

2. ใช้เวลาในการตรวจค่อนข้างมาก

3. นักเรียนที่ใช้ภาษาได้ดีกว่าจะได้เปรียบ โดยเด็กบางคนอาจมีความสามารถทางความคิด แต่ไม่สามารถแสดงให้เห็นในการเขียนคำตอบได้

4. ความเชื่อมั่น (reliability) ต่ำจากผู้สอบและผู้ตรวจ นั่นคือ การสอบแต่ละครั้งด้วยข้อสอบเดิมผู้สอบจะตอบแตกต่างกัน ขณะเดียวกันผู้ตรวจแต่ละคนจะให้คะแนนแตกต่างกันไป นั่นคือการใช้คะแนนเป็นอัตนัย จึงมีโอกาที่จะลำเอียงได้ง่ายในการตรวจ อีกทั้งการตรวจต้องใช้ผู้ที่มีความเชี่ยวชาญในเนื้อหาสาขาวิชาเฉพาะด้านซึ่งมีจำนวนจำกัด

นอกจากนี้ บุญเชิด ภิญโญนนท์พงษ์ (2547: 63-64) ได้กล่าวถึงปัญหาในการตรวจคะแนนข้อสอบปลายเปิดหรือข้อสอบอัตนัยนั้นมีปัญหาที่ทำให้คะแนนมีความคลาดเคลื่อนไปจากความเป็นจริง สรุปได้ 4 ประการ ได้แก่

1. ผลจากสิ่งอื่น ความคลาดเคลื่อนของการให้คะแนนอาจเป็นผลมาจากสิ่งอื่น เช่น เมื่อผู้ตรวจอ่านคำตอบของคำถามข้อแรกซึ่งนักเรียนผู้นี้ตอบได้ดีมาก และได้คะแนนสูง แต่ในข้อถัด ๆ มา คำตอบของนักเรียนไม่ดีเท่าที่ควรแต่คะแนนที่ได้ก็ยังคงได้คะแนนสูงอยู่อีก ทั้งนี้เพราะอิทธิพลของข้อแรกส่งผลต่อข้อถัดมา นอกจากนั้นความรู้จำมักคุ้นก็มีผลต่อการตรวจให้คะแนน เช่น ในการตรวจให้คะแนนคำตอบของนักเรียนคนหนึ่งอยู่ และถ้ารู้ว่าแผ่นคำตอบนั้นเป็นของนักเรียนที่เรียนดี ก็จะทำให้คำตอบนั้นสูงกว่าที่ควรจะเป็น แต่ในทางตรงกันข้ามสมมติว่า คำตอบแผ่นนั้นเป็นของนักเรียนที่เรียนไม่เก่ง ก็จะทำให้คะแนนคำตอบแผ่นนั้นต่ำกว่าความเป็นจริง และในทำนองเดียวกัน สมมติว่านักเรียนตอบคำถาม ข้อแรกไม่ค่อยดีก็จะได้คะแนนต่ำและทำให้คะแนนในข้อถัด ๆ มาต่ำลงด้วย

2. ความใจดี มีครูจำนวนมากที่มักให้คะแนนสูงกว่าที่ควรจะเป็นเสมอ หรือเรียกว่ามักให้คะแนนเฟ้ออยู่เสมอ ทั้ง ๆ ที่คำตอบของนักเรียนจะอยู่ในระดับปานกลางก็ตาม ความคลาดเคลื่อนชนิดนี้เป็นเพราะความใจดีของครู เมื่อเห็นคำตอบนั้นถูกบ้างก็จะให้คะแนนสูงมากทั้ง ๆ ที่ถ้าให้ครูคนอื่น ๆ ตรวจคำตอบนั้นจะอยู่ในระดับปานกลางเท่านั้น

3. การปรับโทษ ครูกดคะแนนหรือการตัดทอนคะแนนของนักเรียนที่ไม่สามารถตอบได้ตามที่คาดหวังไว้ ทั้งที่แท้จริงของนักเรียนอาจจะอยู่ในระดับที่ใช้ได้ แต่ครูได้ตั้งมาตรฐานและคาดหวังความสามารถของนักเรียนไว้สูงเกินไป เมื่อนักเรียนทำไม่ได้ดังที่ตนเองตั้งไว้จึงทำให้เกิดความขัดแย้งว่านักเรียนมีความสามารถต่ำจึงให้คะแนนน้อย

4. องค์ประกอบแทรกซ้อน อิทธิพลที่มีผลต่อการให้คะแนน อาจเกิดจากองค์ประกอบแทรกซ้อนอื่น ๆ ซึ่งยากต่อการควบคุม ทั้งนี้เพราะเป็นผลมาจากตัวนักเรียนเอง หรือ ลายมือ ตัวสะกด การันต์ หลักไวยากรณ์ การเรียบเรียงถ้อยคำ สิ่งเหล่านี้ย่อมมีอิทธิพลต่อการให้คะแนนอยู่เสมอ องค์ประกอบเหล่านี้ครูจำเป็นต้องแจ้งให้นักเรียนทราบว่าการตรวจให้คะแนนจะยึดถือหลักเกณฑ์ดังกล่าวด้วยหรือไม่ เพื่อให้ นักเรียนพยายามและระมัดระวังในการตอบมากขึ้น

ในขณะที่ มายฟอร์ดและวูล์ฟ (Myford & Wolfe. 2003: 393-401) ได้กล่าวถึง ความคลาดเคลื่อนหรืออิทธิพลจากผู้ตรวจให้คะแนน (Rater error/ Rater effect) สรุปได้ดังนี้

1. Halo Effect เป็นอิทธิพลจากความประทับใจที่เกิดจากการรู้จักนักเรียนหรือการตรวจข้อสอบข้อแรก แล้วทำให้ผู้ตรวจให้คะแนนสูงหรือต่ำตามความประทับใจ ผู้ตรวจมีแนวโน้มจะให้คะแนนผู้สอบเท่ากันในทุกข้อ

2. Centrality Effect/Restriction of range คือ การให้คะแนนส่วนใหญ่มีแนวโน้มอยู่ตรงกลางของมาตราประเมิน เนื่องจากผู้ตรวจไม่สามารถจำแนกมาตราประเมินได้ จึงเลือกตรงกลางไว้ก่อน หรืออาจเป็นเพราะไม่มีความรู้ ไม่ได้รับการอบรมเรื่องมาตราประเมินมาก่อนและหากผู้ประเมินอยู่เฉพาะมาตราใดมาตราหนึ่ง อาจเป็นส่วนปลายหรือกลางจะเป็นการจำกัดช่วง (Restriction of range) ส่งผลให้คะแนนผู้สอบสูงหรือต่ำกว่าความเป็นจริง อันเกิดจากพฤติกรรมผู้ตรวจ ไม่ใช่ผลจากผู้สอบ

3. Severity/Leniency Effect คือ ผู้ตรวจมีความเข้มงวด(Severity) ในการให้คะแนน นั่นคือ ผู้ตรวจมีแนวโน้มให้คะแนน (โดยเฉลี่ย) ต่ำกว่าที่ผู้ตรวจคนอื่น เป็นการให้คะแนนต่ำกว่าความเป็นจริง ตลอดช่วงความสามารถ (Continuum) ทำให้มีความผิดพลาดในการกำหนดความสามารถผู้สอบ ส่วนผู้ตรวจที่ใจดี (Leniency) คือผู้ตรวจมีแนวโน้มให้คะแนน (โดยเฉลี่ย) สูงกว่าที่ผู้ตรวจคนอื่น นั่นคือ ให้คะแนนสูงกว่าความเป็นจริง ตลอดช่วงความสามารถ (Continuum) ทำให้มีความผิดพลาดในการกำหนดความสามารถผู้สอบเช่นกัน

4. Random Effect คือ ผู้ตรวจให้คะแนนไม่สม่ำเสมอ หรือไม่สอดคล้องกันกับผู้ตรวจคนอื่น นั่นคือ มีการให้คะแนนในหลายลักษณะในขณะที่ผู้อื่นให้เหมือนกัน

คำถามปลายเปิดจึงเป็นคำถามที่เปิดโอกาสให้นักเรียนได้คิดแก้ปัญหาและสื่อสารวิธีการแก้ปัญหาของตนเองโดยการนำเสนอแนวคิดของตนเองอย่างอิสระ ทำให้นักเรียนได้แสดงทักษะการคิดคำนวณ การให้เหตุผลและการสื่อสารได้อย่างหลากหลาย สามารถใช้วัดความรู้ ความเข้าใจและกระบวนการคิดในระดับสูงได้ดี ทำให้การวัดประเมินมีความสมบูรณ์มากยิ่งขึ้น ซึ่งการวิจัยในครั้งนี้ ลักษณะคำถามปลายเปิดที่ผู้วิจัยใช้เป็นคำถามปลายเปิดที่มีคำตอบที่ถูกต้องเพียงคำตอบเดียว แต่มีความหลากหลายในวิธีการให้เหตุผลหรือการได้มาซึ่งคำตอบ แต่อย่างไรก็ตามการวัดประเมินด้วยคำถามปลายเปิดยังมีปัญหาในประเด็นของความน่าเชื่อถือของคะแนนที่ได้ที่ขึ้นอยู่กับผู้ตรวจเป็นสำคัญ เพื่อให้การตรวจให้คะแนนคำถามปลายเปิดมีความเชื่อมั่นเพิ่มขึ้น จะต้องมีการออกแบบกฎเกณฑ์การให้คะแนน (Scoring rubric) และใช้วิธีการตรวจที่เหมาะสม เพื่อให้เกิดความเป็นปรนัยในการตรวจมากขึ้น

กฎเกณฑ์การให้คะแนน (Scoring Rubric)

คำถามปลายเปิดเป็นคำถามที่นักเรียนสามารถตอบได้อย่างหลากหลายทั้งคำตอบและวิธีการได้มาซึ่งคำตอบ ผู้ตรวจต้องอ่านคำตอบแล้วประเมินค่าออกมาเป็นคะแนนโดยวิธีใดวิธีหนึ่ง ผู้ตรวจจึงต้องเป็นผู้มีความรู้ ความเชี่ยวชาญในสาขาวิชานั้น ซึ่งปัญหาของคำถามปลายเปิดคือ การให้คะแนนขึ้นอยู่กับผู้ตรวจให้คะแนนเป็นสำคัญ ดังนั้น จึงมีคำถามตามมาว่าจะมีการตรวจให้คะแนนอย่างไรจึงจะชัดเจนและยุติธรรม ซึ่งคำถามนี้มีนักการศึกษาหลาย ๆ ท่าน (Cai; et al. 1996; Cooney;

et al. n.d.: online; Hancock. 1995) ให้ความเห็นตรงกันว่าในการตรวจให้คะแนนควรมีการสร้างกฎเกณฑ์การให้คะแนน (Scoring rubric) สำหรับเป็นแนวทางในการตรวจให้คะแนน เพื่อช่วยให้เห็นจุดเน้นว่าอะไรคือสิ่งที่นักเรียนรู้และสามารถทำได้ ช่วยให้การให้คะแนนมีความคงที่ ป้องกันไม่ให้เกิดความลำเอียงหรือความสับสนในการตรวจ และช่วยให้นักเรียนเข้าใจการตอบที่ได้คะแนนในระดับสูงและระดับต่ำ

ในการนำเสนอกฎเกณฑ์การให้คะแนน ผู้วิจัยจึงขอนำเสนอในประเด็นต่อไปนี้ตามลำดับ คือ ความหมายของกฎเกณฑ์การให้คะแนน กฎเกณฑ์การให้คะแนนแบบรวม กฎเกณฑ์การให้คะแนนแบบย่อย มีรายละเอียดดังนี้

ความหมายของกฎเกณฑ์การให้คะแนน

ทอมสันและเซนค (Thompson; & Senk. 1998: 786) ให้ความหมายของกฎเกณฑ์การให้คะแนนว่าเป็นกลุ่มของแนวทางสำหรับการวัดการตอบสนองของนักเรียนที่มีต่อเครื่องมือใดเครื่องมือหนึ่ง

คูนีและคณะ (Cooney; et al. n.d.: online) กล่าวว่า กฎเกณฑ์การให้คะแนน เป็นรายการของตัวบ่งชี้ที่ช่วยให้ผู้ตรวจสามารถจัดระดับการตอบสนองของนักเรียนได้สะดวกขึ้น

บรูคฮาร์ท (Moskal. 2000: online; citing Brookhart. 1999) กล่าวว่า กฎเกณฑ์การให้คะแนน คือ รายการที่บรรยายการให้คะแนนเพื่อใช้เป็นแนวทางในการวิเคราะห์ผลงานและกระบวนการจากการแสดงออกของนักเรียน

เมอร์ทเลอร์ (Mertler. 2001: online) กล่าวว่า กฎเกณฑ์การให้คะแนน คือแนวทางในการให้คะแนน ซึ่งได้กำหนดเกณฑ์ในการแสดงออก เพื่อประเมินผลงานและการปฏิบัติของนักเรียน

ชัยฤทธิ์ ศิลาเดช (2540) กล่าวว่า กฎเกณฑ์การให้คะแนน คือ แนวทางการให้คะแนนที่ต้องกำหนดมาตรฐานวัดและรายการของคุณลักษณะที่บรรยายถึงความสามารถในการแสดงออกหรือคุณลักษณะแต่ละจุดในมาตรฐาน

จากความหมายดังกล่าวสรุปได้ว่ากฎเกณฑ์การให้คะแนน คือ กลุ่มหรือรายการของตัวบ่งชี้ที่แสดงถึงระดับความสามารถในการแสดงออกของนักเรียน ซึ่งกฎเกณฑ์การให้คะแนนเป็นเครื่องมือที่ช่วยทำให้ง่ายต่อการตรวจให้คะแนนและมีความยุติธรรมในการตรวจมากขึ้น

ฮอปกินส์และแอนเตส (Hopkind; & Antes. 1990: 237-238) สเตนมาร์ค (Stenmark. 1991: 20-24) Chicago Board of Education [CBE] (2000: online) เมิร์ตเลอร์ (Mertler. 2001: online) ได้แบ่งกฎเกณฑ์การให้คะแนนไว้สอดคล้องกัน คือแบ่งออกเป็น 2 แบบ ได้แก่ กฎเกณฑ์การให้คะแนนแบบแยกองค์ประกอบ (Analytic Scoring Rubrics) และกฎเกณฑ์การให้คะแนนแบบภาพรวม (Holistic Scoring Rubrics) มีรายละเอียดดังนี้

กฎเกณฑ์การให้คะแนนแบบแยกองค์ประกอบ (Analytic Scoring Rubrics)

กฎเกณฑ์การให้คะแนนแบบนี้เป็นการวิเคราะห์รายละเอียดออกเป็นองค์ประกอบย่อย เช่น การสื่อสาร ความถูกต้องทางคณิตศาสตร์หรือความสมบูรณ์ และลักษณะอื่น ๆ ที่คิดว่าจะมีความสำคัญ จากนั้นกำหนดคะแนนสำหรับแต่ละองค์ประกอบย่อยในแง่ของคุณภาพของคำตอบ โดยหลังจากที่ออกข้อสอบปลายเปิดแล้ว ครูจะดำเนินการเขียนคำตอบต้นแบบ (Model response) ขึ้นมา เพื่อให้มีโครงร่างขององค์ประกอบหลักซึ่งจะระบุไว้ในกระบวนการให้คะแนน ในการรวมคะแนนทั้งหมดควรจะลดบทบาทของข้อมูลที่ไม่เกี่ยวข้องในคำตอบที่จะมีต่อผลรวมของคะแนน ส่วนใหญ่แล้วพบว่าวิธีการให้คะแนนแบบนี้ใช้เวลานานและสร้างความเหนื่อยล้าถ้าหากข้อสอบมีจำนวนมาก

นิตโก (Nitko, 1996: 266) CBE (2000: online) ปรีชา เนาร์เย็นผล (2544: 54) ได้กล่าวถึงข้อดีของการให้คะแนนแบบแยกองค์ประกอบว่าทำให้เกิดการพิจารณาขั้นตอนของกระบวนการแก้ปัญหาที่ไม่ใช่พิจารณาเพียงคำตอบ เป็นวิธีการกำหนดคุณค่าของงานด้วยตัวเลขที่ชัดเจน ทำให้ได้รับข้อมูลที่มีรายละเอียดซึ่งแสดงถึงจุดเด่นและจุดด้อยของผู้เรียนในแต่ละองค์ประกอบ ช่วยให้ครูเน้นหรือปรับปรุงจุดที่นักเรียนต้องปรับปรุงแก้ไขได้ตรงกับความต้องการของนักเรียนมากขึ้น นำไปใช้ในการวางแผนและปรับปรุงการเรียนการสอน รวมถึงการให้ข้อมูลย้อนกลับแก่ผู้เรียนเพื่อพัฒนาหรือปรับปรุงตนเองได้อย่างชัดเจน ตัวอย่างเกณฑ์การให้คะแนนแบบแยกองค์ประกอบ แสดงดังตาราง 2

ตาราง 2 ตัวอย่างกฎเกณฑ์การให้คะแนนแบบแยกองค์ประกอบ

องค์ประกอบย่อย	เกณฑ์การให้คะแนน
● ความเข้าใจในปัญหา	1 ไม่เข้าใจปัญหา 2 เข้าใจผิดปัญหาบางส่วนหรือตีความผิด 3 เข้าใจปัญหาอย่างถูกต้อง ชัดเจน
● การวางแผนแก้ปัญหา	1 ไม่มีการวางแผนหรือวางแผนไม่ถูกต้อง 2 วางแผนถูกต้องบางส่วน 3 วางแผนแล้วนำไปสู่การแก้ปัญหาได้อย่างถูกต้อง
● การได้มาซึ่งคำตอบ	1 ไม่ตอบหรือได้คำตอบที่มาจากวางแผนที่ผิด 2 เขียนคำตอบผิด หรือ คำนวณผิดพลาด 3 คำตอบถูกต้อง

ที่มา: ชาร์ลส์และคณะ. (Stenmark . 1991: 21; citing Charles; et al. 1987)

กฎเกณฑ์การให้คะแนนแบบภาพรวม (Holistic Scoring Rubrics)

กฎเกณฑ์การให้คะแนนแบบนี้จะเป็นการให้คะแนนที่เน้นภาพรวมหรือองค์รวมของงานทั้งหมด โดยจะไม่กำหนดหรือระบุคะแนนแยกแยะไปถึงองค์ประกอบหรือประเด็นต่างๆ ที่จะต้องมีไว้ล่วงหน้าเป็นการเฉพาะเหมือนกับกฎเกณฑ์การให้คะแนนแบบแยกองค์ประกอบ แต่จะเป็นการเขียนบรรยายคุณภาพของงานในแต่ละระดับคะแนน วิธีการนี้จะตรวจคำตอบ โดยอาศัยความประทับใจโดยภาพรวม วิธีการนี้ต้องอาศัยการอ่านข้อสอบในครั้งแรกอย่างรวดเร็วเพื่อสร้างความประทับใจอย่างกว้าง ๆ ต่อคำตอบแต่ละคำตอบ ทดลองแบ่งข้อสอบออกเป็นกลุ่ม ๆ ตามคุณภาพของคำตอบที่แตกต่างกันไป จำนวนกลุ่มนั้นไม่แน่นอน ขึ้นอยู่กับความต้องการของผู้ให้คะแนนและระดับการแยกความแตกต่างที่ต้องการ โดยทั่วไปแล้วผู้ให้คะแนนจะอ่านจนกระทั่งมีคำตอบหนึ่งที่ได้รับการประเมินว่าดีกว่าอย่างแน่นอน และใช้คำตอบนี้เป็นคะแนนอ้างอิงในการให้คะแนนคำตอบอื่น ๆ การอ่านอย่างรอบคอบขึ้นในรอบที่สองสามารถทำให้เกิดการเปลี่ยนแปลงจากกลุ่มที่กำหนดไว้ได้ กล่าวคือ การจัดกลุ่มคำตอบของนักเรียนอาจจะได้รับการกลั่นกรองอีกครั้งถ้าผู้ตรวจรู้สึกว่าการกำหนดประเภทในขั้นแรกผิดพลาด ในขั้นนี้ข้อสอบที่แยกไว้ในแต่ละกลุ่มควรจะอยู่ในระดับเดียวกัน ถ้าข้อสอบฉบับใดมีความโดดเด่นกว่าฉบับอื่น ๆ ในกลุ่มก็ควรจะแยกไปอยู่ในกลุ่มที่ดีกว่า หรือถ้าแยกว่าก็แยกไปกลุ่มแย่กว่า การแยกกลุ่มก่อนที่จะให้คะแนนนี้จะลดความเป็นไปได้ที่ว่าข้อสอบที่อยู่ในระดับเดียวกันได้คะแนนขั้นสุดท้ายแตกต่างกัน หลังจากที่ผู้อ่านพอใจกับการแยกกลุ่มอย่างเหมาะสมแล้ว ก็จะมีการให้คะแนนขั้นสุดท้ายหรือการประเมินอื่น ๆ การตรวจแบบทดสอบปลายเปิดแบบภาพรวมนี้จะมีประสิทธิภาพสำหรับข้อสอบที่มีจำนวนมาก

คูนีและคณะ (Cooney; et al. n.d.: online) ได้กล่าวว่า การให้คะแนนแบบภาพรวมนี้เหมาะสมกับคำถามปลายเปิด เพราะเป็นกฎเกณฑ์การให้คะแนนที่มีประสิทธิภาพ (Effective) และเป็นมิตร (Friendly) กับผู้ใช้ ซึ่งปรีชา เนาว์เย็นผล (2544: 55) และนิตโก (Nitko. 1996: 266) ได้กล่าวถึงข้อดีของการให้คะแนนแบบภาพรวมไว้ว่าเป็นการเปิดโอกาสให้มีการพิจารณาการตอบของนักเรียนได้อย่างรวดเร็ว ใช้เวลาน้อยในการตรวจและง่ายในการนำไปใช้ เน้นการพิจารณากระบวนการที่ใช้ไม่ใช่พิจารณาเฉพาะคำตอบเท่านั้น ใช้คะแนนค่าเดียวซึ่งครอบคลุมภาพรวมของคำตอบ อย่างไรก็ตามแม้ว่าวิธีการนี้จะทำให้การให้คะแนนมีความเชื่อมั่นสูง แต่ก็ยากที่จะอธิบายเหตุผลในการให้คะแนนกับนักเรียนเมื่อคืนข้อสอบให้นักเรียนแล้ว คำตอบที่เป็นแบบอย่างควรจะนำไปอ่านให้นักเรียนในชั้นฟังเพื่อเป็นคะแนนอ้างอิงให้กับนักเรียนแต่ละคน ซึ่งตัวอย่างของกฎเกณฑ์การประเมินแบบภาพรวมในการแก้สมการทางคณิตศาสตร์ แสดงดังตาราง 3

ตาราง 3 ตัวอย่างกฎเกณฑ์การให้คะแนนแบบภาพรวม

ระดับคะแนน	รายละเอียดของแต่ละระดับคะแนน
4	แสดงถึงความเข้าใจเกี่ยวกับความคิดและหลักการทางคณิตศาสตร์ที่สมบูรณ์ คำนวณถูกต้อง มีขั้นตอนในการแก้สมการที่ถูกต้อง สมบูรณ์
3	แสดงถึงความเข้าใจเกี่ยวกับความคิดและหลักการทางคณิตศาสตร์เกือบสมบูรณ์ มีข้อผิดพลาดในการคำนวณเล็กน้อย มีขั้นตอนในการแก้ปัญหาถูกต้อง
2	แสดงถึงความเข้าใจเกี่ยวกับความคิดและหลักการทางคณิตศาสตร์ได้บางส่วน การคำนวณส่วนใหญ่ผิดพลาด มีขั้นตอนในการแก้ปัญหาที่ผิดพลาด
1	แสดงถึงความเข้าใจเพียงเล็กน้อย มีความเข้าใจที่คลาดเคลื่อนมากในความคิด ทางคณิตศาสตร์ ใช้กระบวนการแก้ปัญหาที่ผิด

ที่มา: The Online Teacher Resource (2009: online)

ส่วน ทักแมน (Kubiszyn; & Borich. 1996: 112; citing Tuckman. 1975) ได้ระบุถึงเกณฑ์การให้คะแนนสำหรับคำถามปลายเปิดไว้ว่าควรมีส่วนประกอบ 3 ส่วนที่สำคัญ คือ

1. เนื้อหา (Content) ถึงแม้ว่าเหตุผลหลักของการใช้ข้อสอบแบบปลายเปิดเพื่อใช้วัดกระบวนการคิดของผู้ตอบมากกว่าเพื่อวัดความรู้ตามความข้อเท็จจริงในเนื้อหาวิชา แต่ก็ยังคงต้องการให้ผู้ตอบได้แสดงข้อมูลต่าง ๆ เพื่อประกอบคำถาม การใช้เกณฑ์เนื้อหาในการให้คะแนน จะช่วยให้ถึงความถูกต้องและความเพียงพอของเนื้อหาความรู้ที่ผู้ตอบควรมีในขอบข่ายของคำถามที่ต้องการ เกณฑ์ข้อนี้มีความจำเป็นที่ต้องใช้เมื่อเห็นว่า ความรู้ด้านเนื้อหาสาระนั้นเป็นรากฐานในการพัฒนาการคิด

2. การเรียบเรียง (Organization) การเรียบเรียงนั้น โดยทั่วไปจะดูว่าคำตอบนั้นมีส่วนประกอบ 3 ส่วนคือบทนำ เนื้อความ และส่วนสรุปหรือไม่ ซึ่งผู้ตอบจะต้องเตรียมโครงเรื่องก่อนลงมือเขียนตอบ โดยโครงเรื่องจะเป็นเครื่องช่วยในการเรียบเรียงการนำเสนอความคิดออกมาในรูปแบบที่สามารถสื่อสารได้เข้าใจง่าย นอกจากนี้อาจจะพิจารณาตั้งเกณฑ์เฉพาะอื่น ๆ ขึ้นมา รวมถึงการสะกดคำและหลักไวยากรณ์ ซึ่งการใช้เกณฑ์นี้จะต้องมีการชี้แจงกฎเกณฑ์การให้คะแนนในการเรียบเรียงของคำตอบว่ามีผลต่อคะแนนให้ผู้ตอบทราบล่วงหน้า

3. กระบวนการ (Process) เกณฑ์ของกระบวนการนี้ พิจารณาได้จากคุณลักษณะ 3 ประการคือ 1) ความแม่นยำหรือความถูกต้องของเหตุผล หมายถึง การเสนอวิธีการแก้ปัญหาที่สามารถนำไปใช้ได้จริงโดยสังเกตได้จากการกำหนดแนวทางการวิเคราะห์ที่ถูกต้อง มีความเป็นไปได้ ซึ่งอาจจะเป็นการยากที่จะให้ได้คำตอบที่สมบูรณ์หรือถูกต้องอย่างมีความเป็นปรนัย เพราะการตัดสินใจขึ้นอยู่กับความเห็นผู้ให้คะแนน ดังนั้นครูอาจกำหนดวิธีการแก้ปัญหาที่ถูกต้องไว้ก่อน ในขณะเดียวกันก็ต้องเปิดใจให้กว้าง เมื่อพบคำตอบใหม่ที่ไม่ตรงกับที่คาดหวัง จะต้องพิจารณาถึงความเป็นไปได้ของคำตอบดังกล่าวอย่างรอบคอบด้วย 2) ความสมบูรณ์และความสอดคล้องของคำตอบ หมายถึงสิ่งที่ผู้ตอบได้นำมาเป็นเหตุผลในการสนับสนุนนั้นมีความเหมาะสมกับปัญหาเพียงใด และ 3) ความคิดริเริ่มสร้างสรรค์ของผู้ตอบในการแก้ปัญหา รวมทั้งข้อโต้แย้ง ซึ่งถือว่าเป็นองค์ประกอบที่สำคัญ

นอกจากนี้ CBE (2000: online) ได้ให้ข้อเสนอแนะไว้ว่า ครูสามารถใช้กฎเกณฑ์การให้คะแนนในการพัฒนาหรือกระตุ้นการเรียนรู้ของผู้เรียนได้โดยให้นักเรียนลองใช้กฎเกณฑ์การให้คะแนนในการตรวจงานของเพื่อน เพื่อจะได้แลกเปลี่ยนและเปรียบเทียบแนวคิดกันระหว่างกลุ่มเพื่อนทำให้เห็นมุมมองการตอบที่หลากหลายมากขึ้น รวมถึงสามารถส่งเสริมการตอบของตนเองในครั้งต่อไปได้ดียิ่งขึ้นด้วย

ในการเลือกใช้กฎเกณฑ์การให้คะแนนไม่ว่าจะเป็นกฎเกณฑ์การให้คะแนนแบบรวมหรือแบบย่อย ต่างมีข้อดี ข้อจำกัดแตกต่างกัน ดังนั้นควรเลือกใช้ให้เหมาะสมกับสิ่งที่ต้องการวัด เครื่องมือที่ใช้ในการวัด ลักษณะของผู้เรียนและการจัดกิจกรรมการเรียนการสอน กฎเกณฑ์การให้คะแนนสามารถปรับเปลี่ยนไปได้ตลอดเวลาตามความเหมาะสม ซึ่งงานวิจัยในครั้งนี้ ผู้วิจัยเลือกใช้กฎเกณฑ์การให้คะแนนแบบภาพรวม (Holistic Scoring Rubrics) ในการวัดประเมินแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ ทั้งนี้เนื่องจากเป็นกฎเกณฑ์การให้คะแนนที่มีประสิทธิภาพ (Effective) เมื่อข้อสอบมีจำนวนมาก เป็นมิตร (Friendly) กับผู้ใช้และมีความเหมาะสมกับคำถามปลายเปิดวิชาคณิตศาสตร์ (Cooney; et al. n.d: online) อีกทั้งเป็นกฎเกณฑ์การให้คะแนนที่ใช้อย่างประหยัดเวลาและค่าใช้จ่าย (Nitko. 1996: 266; Nakamura. 2004: online; Zimmaro. 2004: online) นั่นคือ ใช้ได้จริงในทางปฏิบัติ (Practicality) (Nakamura. 2004: online) มีกระบวนการที่เป็นธรรมชาติ (Nature process) สอดคล้องกับสภาพจริง (Authenticity) (Nakamura. 2002: online; citing White.1995. An apologia for the timed impromptu essay test. College composition and communication. 46: 30-45) และเหมาะสมเมื่อวัดประเมินเพื่อการสรุปผล (Summative) (Mertler. 2001: online)

ทฤษฎีการสรุปอ้างอิง (Generalizability Theory)

ทฤษฎีการสรุปอ้างอิง (Generalizability Theory: G-Theory) เริ่มต้นแนวคิดโดยครอนบาคและคณะ (Cronbach; et al. 1972) เป็นทฤษฎีสำหรับการวิเคราะห์ความเชื่อถือได้ของผลการวัดพฤติกรรม (The Dependability of behavioral measurement) ที่อ้างถึงประสิทธิภาพในการสรุปอ้างอิงจากคะแนนสังเกตที่ได้จากแบบทดสอบ ไปยังคะแนนเฉลี่ยที่ผู้สอบคนนั้นพึงได้ในสถานการณ์ต่าง ๆ ของการวัด เป็นการขยายแนวคิดของความเชื่อมั่น (Reliability) จากทฤษฎีการทดสอบแบบมาตรฐานเดิม (Classical Test Theory: CTT) สามารถประมาณค่าความเชื่อมั่นได้โดยไม่ยึดความเท่าเทียมกันของข้อสอบและสามารถประมาณค่าความคลาดเคลื่อนได้หลายแหล่ง (Multiple error sources) พร้อมกัน

เพื่อให้เกิดความเข้าใจในทฤษฎีการสรุปอ้างอิง ผู้วิจัยขอแนะนำเสนอในประเด็นต่อไปนี้ตามลำดับ คือ ความเป็นมาของทฤษฎี แนวคิดพื้นฐานของทฤษฎี คำศัพท์เฉพาะทางทฤษฎีที่จำเป็นต้องมีความรู้ความเข้าใจ การศึกษาแหล่งความแปรปรวนและองค์ประกอบความแปรปรวน จากนั้นนำเสนอวิธีการประมาณค่าองค์ประกอบความแปรปรวน ความแปรปรวนของคะแนนความคลาดเคลื่อน และในตอนสุดท้ายเป็นการประมาณค่าสัมประสิทธิ์การสรุปอ้างอิง ในแต่ละประเด็นมีรายละเอียดดังนี้

ความเป็นมาของทฤษฎีการสรุปอ้างอิง

การศึกษาความเชื่อมั่นและวิธีการประมาณค่าความเชื่อมั่นของแบบทดสอบส่วนใหญ่จะใช้ทฤษฎีการทดสอบแบบมาตรฐานเดิม (Classical Test Theory: CTT) มาอธิบายความแม่นยำในการวัดโดยยึดข้อตกลงคุณสมบัติคู่ขนานหรือความเท่าเทียมกันของแบบทดสอบเป็นสำคัญ และคะแนนที่ได้จากการสอบ (X) ประกอบด้วยคะแนนจริง (T) รวมกับคะแนนความคลาดเคลื่อน (E) ซึ่งสามารถเขียนเป็นสมการได้เป็น $X = T + E$ เมื่อ X คือ คะแนนสังเกตหรือคะแนนที่ได้จากการสอบ (Observed scores) T คือ คะแนนจริง (True scores) และ E คือ คะแนนความคลาดเคลื่อน (Error scores) โดยเทคนิคการศึกษาความเชื่อมั่นของแบบทดสอบดังกล่าว เป็นการประมาณค่าความเชื่อมั่นของแบบทดสอบที่ใช้เฉพาะกลุ่มบุคคลภายใต้เงื่อนไขของการทดสอบเฉพาะได้แก่การทดสอบที่ผู้สอบทุกคนต้องทำข้อสอบเหมือนกันทุกข้อ ผลการตอบได้รับการให้คะแนนจากผู้ตรวจคนเดียว แต่ถ้าสถานการณ์ของการทดสอบแตกต่างไปจากที่กล่าวมา เทคนิคการประมาณค่าความเชื่อมั่นตามทฤษฎีการทดสอบแบบมาตรฐานเดิมไม่สามารถนำมาประยุกต์ใช้ได้ เช่น กรณีของการทดสอบที่มีผู้ตรวจหลายคนให้คะแนนข้อสอบความเรียงของผู้สอบทุกคน เป็นต้น นอกจากนี้โมเดลการวัดของทฤษฎีการทดสอบแบบมาตรฐานเดิมนั้นคะแนนความคลาดเคลื่อนหรือความคลาดเคลื่อน

ของการวัดเป็นคะแนนรวมของความคลาดเคลื่อนทุกแหล่งเข้าด้วยกันและเป็นความคลาดเคลื่อนรวมอันเดียวที่ไม่สามารถแบ่งแยกได้ จึงเป็นข้อจำกัดทางทฤษฎีที่ไม่สามารถศึกษารายละเอียดของแหล่งความคลาดเคลื่อนของการวัดในสถานการณ์ หรือเงื่อนไขของการวัดต่าง ๆ ได้ จึงทำให้มีนักทฤษฎีการทดสอบหลายท่านพยายามศึกษาเชิงจำแนกแหล่งความคลาดเคลื่อนของการวัด ซึ่งวิธีการหนึ่งที่เป็นประโยชน์ต่อการศึกษาเชิงจำแนกแหล่งความคลาดเคลื่อนของการวัด คือ การวิเคราะห์ความแปรปรวน (Analysis of Variance: ANOVA) เป็นการนำเทคนิคการวิเคราะห์ความแปรปรวนซึ่งเป็นวิธีการทางสถิติมาประมาณค่าความเชื่อมั่นของแบบทดสอบ และในปี 1963 ครอนบาค เกลเซอร์และราชาร์ทน์ (Cronbach; Gleser; & Rajaratnam. 1963; citing Cronbach; et al. 1972) ได้เสนอทฤษฎีการสรุปอ้างอิง (Generalizability Theory: G-Theory) ทาง การทดสอบขึ้นเป็นครั้งแรก ซึ่งเป็นทฤษฎีที่สามารถวิเคราะห์ความแปรปรวนของคะแนนความคลาดเคลื่อนอย่างเป็นระบบจากแหล่งต่าง ๆ อันเป็นสถานการณ์ของการวัดที่เป็นไปได้ จึงทำให้ทราบและสามารถควบคุมแหล่งความคลาดเคลื่อนได้ และสามารถประมาณค่าความเชื่อมั่นของแบบทดสอบในสถานการณ์หรือบริบทการวัดต่าง ๆ ได้

แนวคิดพื้นฐานของทฤษฎีการสรุปอ้างอิง

แนวคิดพื้นฐานของทฤษฎีการสรุปอ้างอิงนั้น เกิดจากข้อจำกัดของทฤษฎีการทดสอบแบบมาตรฐานเดิมที่มีข้อตกลงเบื้องต้นว่า เครื่องมือวัดจะต้องมีความเท่าเทียมกันในลักษณะความเป็นคู่ขนาน คือ มีเนื้อหาเดียวกัน มีค่าเฉลี่ยเท่ากัน มีค่าความแปรปรวนเท่ากันและมีค่าสหสัมพันธ์ระหว่างกันเท่ากัน ซึ่งก่อให้เกิดปัญหาในทางปฏิบัติเพราะไม่สามารถหาเครื่องมือที่มีความเป็นคู่ขนานกันอย่างสมบูรณ์ได้ และในการประมาณค่าความเชื่อมั่นนั้นมีพื้นฐานความเชื่อว่า ความแปรปรวนของคะแนนสังเกต (σ_x^2) ประกอบด้วย 2 ส่วน คือ ความแปรปรวนของคะแนนจริง (σ_t^2) ซึ่งเป็นความแตกต่างที่แท้จริงระหว่างบุคคล และความแปรปรวนของคะแนนความคลาดเคลื่อน (σ_e^2) ซึ่งเป็นความคลาดเคลื่อนรวมทุกแหล่งที่ไม่สามารถระบุหรือแบ่งแยกได้ (Single error source) จึงไม่ได้ให้ความสนใจกับสถานการณ์หรือเงื่อนไขการวัด และปฏิสัมพันธ์ระหว่างผู้สอบกับเงื่อนไขของการวัดที่สามารถส่งผลต่อความน่าเชื่อถือของผลการวัด จึงทำให้ครอนบาคและคณะ (Cronbach; et al. 1972) ได้พัฒนาทฤษฎีความเชื่อถือได้ของการวัดหรือทฤษฎีการสรุปอ้างอิง (Generalizability Theory: G-Theory) ขึ้นในปี 1963 ที่สามารถประมาณค่าความเชื่อมั่นโดยที่ไม่ยึดข้อตกลงของความเท่าเทียมกัน สามารถวิเคราะห์ความเชื่อถือได้ของผลการวัดพฤติกรรมในสถานการณ์การวัดผลลักษณะต่าง ๆ เป็นการขยายแนวคิดของความเชื่อมั่น (Reliability) จากทฤษฎีการทดสอบแบบมาตรฐานเดิม ให้สามารถสรุปความเชื่อมั่นในสถานการณ์หรือเงื่อนไขการทดสอบในลักษณะต่าง ๆ ได้ และมีจุดแวง

กว่าทฤษฎีการทดสอบแบบมาตรฐานเดิม คือ สามารถประมาณความคลาดเคลื่อนของการวัดได้หลายแหล่ง (Multiple error sources) ประกอบด้วยแหล่งความคลาดเคลื่อนอย่างเป็นระบบ (Systematic source) และแหล่งความคลาดเคลื่อนแบบสุ่ม (Random source) ได้ในการวิเคราะห์เพียงครั้งเดียว

ทฤษฎีการสรุปอ้างอิงเป็นทฤษฎีทางสถิติ ใช้วิเคราะห์ความเชื่อถือได้ของผลการวัดทางพฤติกรรม (The Dependability of behavioral measurement) ทั้งนี้ความเชื่อถือได้นั้น หมายถึง ความถูกต้องของการสรุปอ้างอิงจากคะแนนที่สังเกตได้ของบุคคลจากการสอบไปเป็นคะแนนเฉลี่ยที่บุคคลนั้นพึงได้รับภายใต้เงื่อนไขการวัดที่ยอมรับที่เป็นไปได้ทั้งหมด (Shavelson; & Webb. 1991: 1) โดยมีข้อตกลงเบื้องต้นของการศึกษาความเชื่อถือได้ของผลการวัดตามแนวทฤษฎีการสรุปอ้างอิง ดังนี้ (Shavelson; & Webb. 1991: 1)

1. คุณลักษณะที่ต้องการวัดของบุคคล เช่น ความรู้ ทักษะ ทักษะ หรืออื่น ๆ ต้องมีลักษณะคงที่ (Steady state)

2. ผู้สอบแต่ละคนที่ได้คะแนนแตกต่างกันจากการสอบในแต่ละสถานการณ์นั้นเกิดจากความคลาดเคลื่อนอย่างน้อย 1 แหล่ง โดยที่วุฒิภาวะ (Maturation) หรือ การเรียนรู้ (Learning) ไม่ได้เป็นแหล่งความคลาดเคลื่อนของคะแนนที่ได้จากการวัด

ทฤษฎีการสรุปอ้างอิงได้เสนอวิธีการวิเคราะห์ความแปรปรวนของคะแนนความคลาดเคลื่อนอย่างเป็นระบบจากแหล่งต่าง ๆ อันเป็นสถานการณ์ของการวัดที่เป็นไปได้ทั้งหมด เช่น ความยาวของแบบทดสอบ จำนวนครั้งของการทดสอบ เป็นต้น รวมทั้งปฏิสัมพันธ์ระหว่างผู้สอบกับเงื่อนไขของการวัด โดยเป้าหมายหนึ่งของทฤษฎีนี้คือ วิเคราะห์หาแหล่งของความแปรปรวนหลัก เพื่อลดความแปรปรวนที่ไม่ต้องการให้น้อยลงในการเก็บข้อมูลครั้งต่อไป (Cronbach; et al. 1972: 1; Shavelson; & Webb. 1991: 6) จึงทำให้ทราบและสามารถควบคุมแหล่งความคลาดเคลื่อนได้นอกจากนี้ยังสามารถหาค่าสัมประสิทธิ์การสรุปอ้างอิง (Generalizability coefficient) ซึ่งเป็นสัมประสิทธิ์ที่แสดงถึงระดับความเชื่อถือได้ คล้ายกับสัมประสิทธิ์ความเชื่อมั่น (Reliability coefficient) ในทฤษฎีการทดสอบแบบมาตรฐานเดิม ทฤษฎีการสรุปอ้างอิงจึงเป็นทั้งวิธีประเมินความน่าเชื่อถือหรือความเชื่อมั่นของเครื่องมือวัด และช่วยผู้บริหารการสอบออกแบบการวัดเพื่อให้ได้ผลการวัดที่มีความน่าเชื่อถือตามที่ต้องการ (Shavelson; & Webb. 1991: 3) ทำให้มีความถูกต้อง (Accuracy) แม่นยำ (Precision) มากขึ้น สามารถนำไปใช้เป็นสารสนเทศที่เพียงพอสำหรับการตัดสินใจได้อย่างมีประสิทธิภาพ (ศิริชัย กาญจนวาสี. 2550: 13) นอกจากนี้ยังเป็นทฤษฎีที่ยินยอมให้ผู้ตัดสินใจเลือกใช้การแปลความที่แตกต่างกันได้ทั้งในลักษณะของความแปรปรวนของความคลาดเคลื่อนสัมพัทธ์ และความแปรปรวนของความคลาดเคลื่อนสัมบูรณ์ ซึ่งจะกล่าวรายละเอียดต่อไป

แต่ก่อนที่จะนำทฤษฎีการสรุปอ้างอิงไปใช้ในทางองค์ประกอบความแปรปรวนและประมาณค่าสัมประสิทธิ์การสรุปอ้างอิง มีความจำเป็นจะต้องมีความรู้และเข้าใจเกี่ยวกับคำศัพท์ของทฤษฎีการสรุปอ้างอิง เนื่องจากมีคำศัพท์เฉพาะที่เกี่ยวข้องของหลายคำที่จะต้องมีความเข้าใจที่ถูกต้องก่อน

คำศัพท์ที่เกี่ยวข้องกับทฤษฎีการสรุปอ้างอิง

การที่จะนำทฤษฎีการสรุปอ้างอิงไปใช้ในการทางองค์ประกอบความแปรปรวนหรือ ประมาณค่าสัมประสิทธิ์การสรุปอ้างอิงนั้น จะต้องมีความรู้ความเข้าใจในคำศัพท์เฉพาะของทฤษฎีการสรุปอ้างอิง ดังนี้ (Brennan. 1983: 2-6; Shavelson; & Webb. 1991; ศิริชัย กาญจนวาสี. 2550: 15-16)

1. ประชากร (Population) หมายถึง สิ่งที่ต้องการวัด (Object of measurement) หรือสิ่งที่มีวัดทั้งหมดในสถานการณ์ของการทดสอบทั่วไป สิ่งที่มีวัดมักได้แก่บุคคลหรือผู้สอบ

2. องค์ประกอบ (Facet) หมายถึง ชุดของเงื่อนไขการวัดที่คล้ายคลึงกัน (Similar conditions of measurement) หรือกลุ่มเงื่อนไขของการวัด ซึ่งเป็นองค์ประกอบที่คาดว่าจะมีผลต่อการวัดความคลาดเคลื่อน เช่น ความยาวของแบบทดสอบ รูปแบบของข้อสอบ จำนวนครั้งของการสอบ เป็นต้น โดยที่องค์ประกอบที่ต้องการศึกษาอาจเป็นองค์ประกอบสุ่มหรือองค์ประกอบแบบเจาะจงก็ได้

3. เงื่อนไขของการวัด (Condition of measurement) หมายถึง ระดับขององค์ประกอบที่ทำให้ได้ค่าสังเกตแต่ละค่าในการวัดครั้งหนึ่ง ๆ เช่น องค์ประกอบของความยาวของแบบทดสอบอาจกำหนดระดับความยาวของแบบทดสอบเป็น 10, 30, 50 ข้อ เป็นต้น

4. เอกภพ (Universe) หมายถึง กลุ่มเงื่อนไขของการวัดทั้งหมดของแต่ละองค์ประกอบ เช่น จำนวนข้อสอบทุกข้อ ซึ่งมีความหมายคล้ายกับประชากร (Population) แต่แตกต่างกันคือ เอกภพใช้กับองค์ประกอบที่เกี่ยวข้องกับการวัด ส่วนประชากรใช้กับองค์ประกอบที่ถูกวัด แยกออกมาเป็น

4.1 เอกภพของการสังเกตที่ยอมรับได้ (Universe of admissible observation) หรือเอกภพของค่าที่ได้จากการสังเกตทั้งหมด หมายถึง กลุ่มเงื่อนไขของการวัดที่สามารถวัดหรือสังเกตได้ในแต่ละองค์ประกอบ เช่น ประกอบด้วยองค์ประกอบข้อสอบและองค์ประกอบของผู้ตรวจ

4.2 เอกภพของการสรุปอ้างอิง (Universe of generalization) เป็นเงื่อนไขของการวัดทั้งหมดขององค์ประกอบที่ต้องการสรุปอ้างอิงผลการวัดจากกลุ่มเงื่อนไขของการวัดเหล่านี้ไปยังกลุ่มเงื่อนไขของการวัดทั้งหมดขององค์ประกอบนั้น ๆ ดังนั้นเงื่อนไขขององค์ประกอบในเอกภพการสรุปอ้างอิงจึงเป็นสับเซตในเอกภพของการสังเกตที่ยอมรับได้ เอกภพของการสรุปอ้างอิงนี้จะต้องระบุประเด็นต่าง ๆ ให้ชัดเจน ซึ่งจะนำไปใช้ในการออกแบบการวัดของการศึกษาเพื่อตัดสินใจ

5. คะแนนเอกภพ (Universe score) หมายถึง คะแนนเฉลี่ยของบุคคลที่ได้จากเงื่อนไขการวัดทั้งหมดของแต่ละองค์ประกอบ ซึ่งคล้ายกับคะแนนจริงในทฤษฎีทดสอบแบบมาตรฐานเดิม

6. การศึกษาเพื่อการสรุปอ้างอิง (การศึกษา G หรือ G Study) และการศึกษาเพื่อการตัดสินใจ (การศึกษา D หรือ D Study) ทั้งนี้ทฤษฎีการสรุปอ้างอิงมีขั้นตอนการศึกษาที่สำคัญสองขั้นตอน คือ การศึกษา G และ การศึกษา D

6.1 การศึกษา G เป็นการสรุปอ้างอิงผลจากการศึกษาตัวอย่างการวัดตามเงื่อนไขที่สนใจ ประมาณค่าความแปรปรวนของคะแนนจริงและความแปรปรวนของคะแนนความคลาดเคลื่อนจากแหล่งความคลาดเคลื่อนต่าง ๆ ที่สนใจ เพื่อสรุปอ้างอิงไปยังเอกภพของการวัด และใช้เป็นข้อมูลสำหรับวางแผนเพื่อตัดสินใจในการศึกษา D

6.2 การศึกษา D เป็นการใช้ข้อมูลจากการศึกษา G ที่สอดคล้องกับจุดประสงค์เฉพาะของการตัดสินใจ การเลือกใช้แบบทดสอบในสถานการณ์ต่าง ๆ ของการวัด ซึ่งเกี่ยวข้องกับค่าความเชื่อมั่นของแบบทดสอบในสถานการณ์ของการวัดต่าง ๆ

7. รูปแบบความสัมพันธ์ระหว่างองค์ประกอบ หมายถึง องค์ประกอบที่บ่งชี้แหล่งความแปรปรวนของรูปแบบการวัด ในแบบจำลองการวิเคราะห์ความแปรปรวน รูปแบบความสัมพันธ์ แบ่งออกเป็น 3 ลักษณะดังนี้

7.1 ความสัมพันธ์แบบไขว้ (Crossed) หมายถึง ความสัมพันธ์ในลักษณะที่แต่ละระดับของสิ่งที่ถูกวัด ถูกวัดภายใต้เงื่อนไขเดียวทั้งหมด สัญลักษณ์ คือ “x” อ่านว่า “crossed with” เช่น กำหนดรูปแบบของความสัมพันธ์ระหว่างองค์ประกอบเป็น $p \times i \times r$ หมายถึง นักเรียน (p) ทำข้อสอบ (i) ทุกข้อ และผู้ตรวจ (r) ตรวจข้อสอบของนักเรียนทุกคน ทุก ๆ ข้อ

7.2 ความสัมพันธ์แบบแฝง (Nested) หมายถึง ความสัมพันธ์ในลักษณะที่แต่ละระดับของสิ่งที่ถูกวัด ถูกวัดภายใต้เงื่อนไขที่แตกต่างกัน สัญลักษณ์ คือ “:” อ่านว่า “nested within” เช่น กำหนดรูปแบบของความสัมพันธ์ระหว่างองค์ประกอบเป็น $i:r$ หมายถึง ผู้ตรวจข้อสอบ (r) ตรวจข้อสอบ (i) ต่างข้อกัน

7.3 ความสัมพันธ์แบบผสม (Confounded) หมายถึง ความสัมพันธ์ที่มีทั้งความสัมพันธ์แบบไขว้และความสัมพันธ์แบบแฝงปนกันอยู่ เช่น $p \times (i:r)$ หมายถึง นักเรียน (p) ทำข้อสอบ (i) บางข้อ ที่มีผู้ตรวจข้อสอบ (r) ตรวจข้อสอบต่างข้อกัน

แหล่งความแปรปรวนและองค์ประกอบความแปรปรวน

คะแนนจริงของผู้สอบ (True score: T) ในทฤษฎีการทดสอบแบบมาตรฐานเดิม (Classical Test Theory) คือ ค่าเฉลี่ยของคะแนนจากการสอบซ้ำ ๆ ด้วยแบบทดสอบคู่ขนาน ความแปรปรวนของคะแนนจริงจึงเป็นความแปรปรวนของค่าเฉลี่ยของการสอบซ้ำ และความแปรปรวนของคะแนนสังเกต เป็นผลรวมของความแปรปรวนของคะแนนจริงกับความแปรปรวนของความคลาดเคลื่อน นั้น

คือ $\sigma_x^2 = \sigma_T^2 + \sigma_E^2$ แต่ในทฤษฎีการสุ่มอ้างอิงนั้นเป็นโมเดลทางสถิติที่แบ่งความแปรปรวนเป็นหลายส่วน โดยใช้หลักการวิเคราะห์ความแปรปรวน (ANOVA)

ตัวอย่างเช่น ในกรณีที่เป็นการศึกษาแบบองค์ประกอบเดียว สามารถแบ่งส่วนของคะแนนบุคคลได้เป็น อิทธิพลของบุคคล (ความแปรปรวนของคะแนนเอกภพ) อิทธิพลของข้อสอบ (ความแปรปรวนเนื่องจากความยากของข้อสอบ) และ ส่วนที่เหลือ ซึ่งเป็นปฏิสัมพันธ์ของบุคคลกับข้อสอบ ความคลาดเคลื่อนอย่างเป็นระบบอื่นนอกเหนือจากบุคคลกับข้อสอบและความคลาดเคลื่อนแบบสุ่ม (Shavelson; & Webb. 1991: 17-18) ทั้งนี้ชาร์เวลสันและเวบบ์ (Shavelson; & Webb. 1991: 18-21) ได้กล่าวถึงแต่ละส่วนไว้ดังนี้

คะแนนสังเกตของบุคคล (p) ที่ได้จากข้อสอบ (i) แทนด้วย X_{pi} และคะแนน X_{pi} นั้นเป็นผลรวมของ 3 พารามิเตอร์ คือ μ_p , μ_i และ μ ซึ่ง พารามิเตอร์ตัวแรกคือ คะแนนเอกภพของแต่ละคน (Universe Score : μ_p) เป็นคะแนนเฉลี่ยของบุคคลจากเอกภพของข้อสอบ กล่าวอีกอย่างได้ว่า คะแนนเอกภพของบุคคล (μ_p) เป็นค่าคาดหวัง(Expected value: E) ของตัวแปรสุ่มของคะแนนสังเกต (X_{pi}) จากข้อสอบทั้งหมด นั่นคือ

$$\mu_p \equiv E_i(X_{pi})$$

ซึ่งค่าคาดหวัง(E) ของตัวแปรสุ่ม (X_{pi}) นั้นเป็นแนวโน้มค่าเฉลี่ยจากข้อสอบทั้งหมด k ข้อ ซึ่งเป็นจำนวนอนันต์ (infinity) นั่นคือ ค่าคาดหวังของคะแนนสังเกต คือ ค่าลิมิตของค่าเฉลี่ยของข้อสอบ k ข้อ เมื่อ k เข้าใกล้อนันต์ ($k \rightarrow \infty$) กำหนดโดย

$$E_i(X_{pi}) = \lim_{k \rightarrow \infty} \frac{1}{k} \sum_{i=1}^k X_{pi}$$

สำหรับพารามิเตอร์ μ_i เป็นค่าเฉลี่ยของประชากรสำหรับข้อสอบข้อ i เป็นค่าคาดหวังของ X_{pi} จากบุคคลทั้งหมด กำหนดโดย

$$\mu_i \equiv E_p X_{pi}$$

ส่วนพารามิเตอร์ μ คือค่าเฉลี่ยของคะแนนทั้งหมดทั้งจากประชากรและเอกภพ (Grand mean) กำหนดโดย

$$\mu \equiv E_p E_i X_{pi}$$

พารามิเตอร์ μ_p , μ_i และ μ ไม่สามารถสังเกตเห็นได้ สามารถแสดงค่า X_{pi} ในเทอมของ μ_p , μ_i และ μ กล่าวคือ คะแนนสังเกตของแต่ละคน (X_{pi}) สามารถแบ่งเป็นส่วนย่อย ๆ ตามการออกแบบการวัด ดังนี้ (Shavelson; & Webb. 1991: 19-60)

1. กรณีออกแบบการวัดแบบไขว้ (Crossed design)

1.1 กรณีการวัดที่มีหนึ่งองค์ประกอบแบบไขว้ (One – facet, $p \times i$ Design)

คะแนนสังเกตของบุคคลในข้อสอบข้อหนึ่ง ได้รับอิทธิพลมาจากส่วนต่าง ๆ ดังนี้ (Shavelson; & Webb. 1991: 19)

$$\begin{aligned} X_{pi} = & \mu && [\text{ค่าคงที่หรือค่าเฉลี่ยของทั้งหมด}] \\ & + \mu_p - \mu && [\text{อิทธิพลของบุคคล}] \\ & + \mu_i - \mu && [\text{อิทธิพลของข้อสอบ}] \\ & + X_{pi} - \mu_p - \mu_i + \mu && [\text{ส่วนที่เหลือ หรือ } e_{pi}] \end{aligned}$$

เมื่อ μ คือ ค่าเฉลี่ยของคะแนนทั้งหมดจากประชากรในเอกภพ (Grand mean)

μ_p คือ คะแนนเอกภพของแต่ละคน

μ_i คือ คะแนนเฉลี่ยรายข้อ

จะพบว่า อิทธิพลอื่นที่ต่างไปจากค่าเฉลี่ยทั้งหมด (Grand mean) มีการกระจายเกิดขึ้น แต่ค่าเฉลี่ยทั้งหมดมีค่าคงที่ จึงมีค่าความแปรปรวนเป็นศูนย์

การแจกแจงที่มีค่าเฉลี่ยเป็นศูนย์ และมีความแปรปรวน (σ^2) เรียกว่าองค์ประกอบความแปรปรวน (Variance component) ประกอบด้วยส่วนแรกเป็นอิทธิพลของบุคคล ค่าเฉลี่ยของความแตกต่างของคะแนนระหว่างบุคคลเป็นศูนย์ คือ

$$E_p(\mu_p - \mu) = E_p(\mu_p) - E_p(\mu) = \mu - \mu = 0$$

ความแปรปรวนของอิทธิพลของบุคคล ใช้สัญลักษณ์เป็น σ_p^2 เรียกว่าองค์ประกอบความแปรปรวนของบุคคล หรือ ความแปรปรวนของคะแนนเอกภพ เขียนสมการได้เป็น

$$\sigma_p^2 = E_p(\mu_p - \mu)^2$$

ค่าเฉลี่ยและองค์ประกอบความแปรปรวนของข้อสอบ จะมีค่าเฉลี่ยรายข้อเป็นศูนย์ และองค์ประกอบความแปรปรวนของข้อสอบใช้สัญลักษณ์เป็น σ_i^2

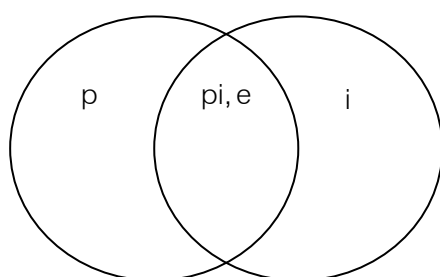
อิทธิพลส่วนสุดท้าย คือส่วนที่เหลือ มีค่าเฉลี่ยเป็นศูนย์และความแปรปรวนเป็น $\sigma_{pi,e}^2$

จะได้ว่าความแปรปรวนของคะแนนสังเกต เป็นผลรวมของ 3 องค์ประกอบความแปรปรวนข้างต้น ดังนี้

$$\sigma_{X_{pi}}^2 = \sigma_p^2 + \sigma_i^2 + \sigma_{pi,e}^2$$

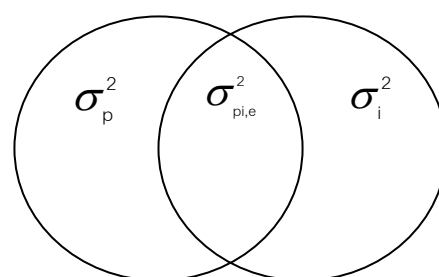
เมื่อ $\sigma_{X_{pi}}^2$ คือ ความแปรปรวนของคะแนนสังเกต (X_{pi})
 σ_p^2 คือ ความแปรปรวนของคะแนนเอกภพ (μ_p)
 σ_i^2 คือ ความแปรปรวนของคะแนนเฉลี่ยรายข้อ (μ_i)
 $\sigma_{pi,e}^2$ คือ ความแปรปรวนของคะแนนความคลาดเคลื่อน หรือ ส่วนที่เหลือ (e_{pi})

ความแปรปรวนของคะแนนที่สังเกตได้ในการวัดที่มีหนึ่งองค์ประกอบแบบไขว้ (One-facet crossed design) จึงขึ้นอยู่กับ 4 แหล่ง คือ 1) ความแตกต่างระหว่างบุคคลหรือสิ่งที่วัด (Person effect) 2) ความแตกต่างระหว่างความยากของข้อสอบ (Item effect) 3) ปฏิสัมพันธ์ระหว่างบุคคลกับข้อสอบ ($p \times i$ Interaction) และ 4) ความคลาดเคลื่อนที่เกิดขึ้นอย่างสุ่ม (Unsystematic or Random error) แต่ความแปรปรวนจากแหล่งที่ 3 และ ที่ 4 ไม่สามารถแยกออกจากกันได้ เรียกว่าเป็นส่วนที่เหลือ (Residual) จึงเป็นแหล่งความแปรปรวนร่วมกัน ดังภาพประกอบ 1



แหล่งของความแปรปรวน

(Source of variation)



องค์ประกอบของความแปรปรวน

(Variance component)

ภาพประกอบ 1 แหล่งของความแปรปรวนและองค์ประกอบของความแปรปรวน สำหรับหนึ่ง
องค์ประกอบแบบไขว้ $p \times i$ (One Facet, $p \times i$ Design)

1.2 กรณีการวัดที่มีสององค์ประกอบแบบไขว้ (Two – Facet, $p \times r \times o$ Design)

คะแนนสังเกตของบุคคลในข้อสอบข้อหนึ่ง ได้รับอิทธิพลมาจากส่วนต่าง ๆ ดังนี้

(Shavelson; & Webb. 1991: 22-23)

$$\begin{aligned}
 X_{pro} = & \mu && [\text{ค่าคงที่หรือค่าเฉลี่ยทั้งหมด}] \\
 & + \mu_p - \mu && [\text{อิทธิพลของบุคคล}] \\
 & + \mu_r - \mu && [\text{อิทธิพลของผู้ตรวจ}] \\
 & + \mu_o - \mu && [\text{อิทธิพลของสถานการณ์}] \\
 & + \mu_{pr} - \mu_p - \mu_r + \mu && [\text{ปฏิสัมพันธ์ระหว่างบุคคลกับผู้ตรวจ}] \\
 & + \mu_{po} - \mu_p - \mu_o + \mu && [\text{ปฏิสัมพันธ์ระหว่างบุคคลกับสถานการณ์}] \\
 & + \mu_{ro} - \mu_r - \mu_o + \mu && [\text{ปฏิสัมพันธ์ระหว่างผู้ตรวจกับสถานการณ์}] \\
 & + X_{pro} - \mu_{po} - \mu_{pr} - \mu_{ro} + \mu_p + \mu_r + \mu_o - \mu && [\text{อิทธิพลของส่วนที่}]
 \end{aligned}$$

เหลือ]

แต่ละองค์ประกอบยกเว้นค่าเฉลี่ยทั้งหมด (Grand mean) นั้นมีการกระจาย นั่นคือ พฤติกรรมของผู้สอบเปลี่ยนไป ผู้ตรวจให้คะแนนแตกต่างกัน ซึ่งแต่ละการกระจายมีค่าเฉลี่ยเป็นศูนย์ และมีความแปรปรวน ความแปรปรวนทั้งหมดของเอกภาพและประชากร จึงเป็นผลรวมของ องค์ประกอบความแปรปรวน 7 ส่วน ดังนี้

$$\sigma^2(X_{\text{pro}}) = \sigma_p^2 + \sigma_r^2 + \sigma_o^2 + \sigma_{\text{pr}}^2 + \sigma_{\text{po}}^2 + \sigma_{\text{ro}}^2 + \sigma_{\text{pro,e}}^2$$

เมื่อ σ_{pro}^2 คือ ความแปรปรวนของคะแนนสังเกต

σ_p^2 คือ ความแปรปรวนของคะแนนเอกภพของผู้สอบ (μ_p)

σ_r^2 คือ ความแปรปรวนของคะแนนเฉลี่ยจากผู้ตรวจ (μ_r)

σ_o^2 คือ ความแปรปรวนของคะแนนเฉลี่ยของสถานการณ์การสอบแต่ละครั้ง (μ_o)

σ_{pr}^2 คือ ความแปรปรวนของคะแนนปฏิสัมพันธ์ระหว่างผู้สอบกับผู้ตรวจ

σ_{po}^2 คือ ความแปรปรวนของคะแนนปฏิสัมพันธ์ระหว่างผู้สอบกับ

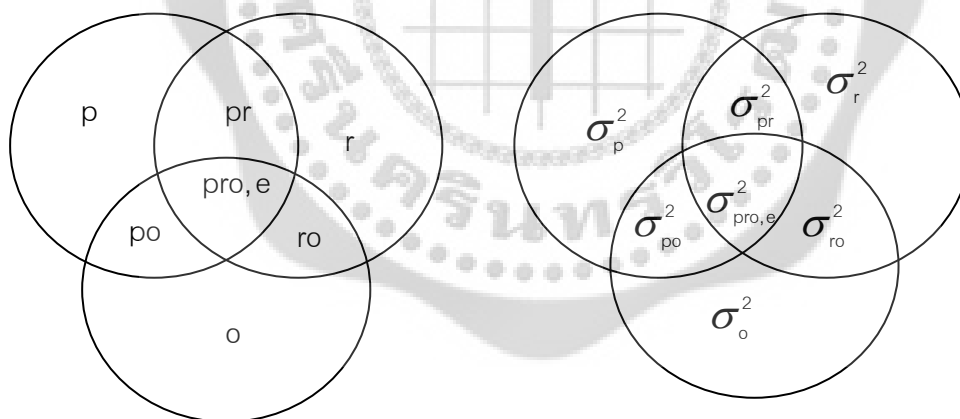
สถานการณ์ของการสอบแต่ละครั้ง

σ_{ro}^2 คือ ความแปรปรวนของคะแนนปฏิสัมพันธ์ระหว่างผู้ตรวจกับ

สถานการณ์ของการสอบแต่ละครั้ง

$\sigma_{\text{pro,e}}^2$ คือ ความแปรปรวนของคะแนนคลาดเคลื่อนหรือส่วนที่เหลือ (e_{pro})

เขียนแผนภาพเวนนีได้ดังนี้



แหล่งของความแปรปรวน

(Source of variation)

องค์ประกอบของความแปรปรวน

(Variance component)

ภาพประกอบ 2 แหล่งของความแปรปรวนและองค์ประกอบของความแปรปรวนสำหรับสอง

องค์ประกอบแบบไขว้ $p \times r \times o$ (Two – facet, $p \times r \times o$ Design)

2. กรณีออกแบบการวัดแบบแฝง (Nested design)

2.1 กรณีการวัดที่มีหนึ่งองค์ประกอบแบบแฝง เมื่อ i แฝงใน p (One facet; $i:p$ Design)

คะแนนสังเกตของบุคคลในข้อสอบข้อหนึ่ง ได้รับอิทธิพลมาจากส่วนต่าง ๆ ดังนี้ (Shavelson; & Webb. 1991: 49 -50)

$$\begin{aligned}
 X_{pi} &= \mu && \text{[ค่าเฉลี่ยทั้งหมด]} \\
 &+ \mu_p - \mu && \text{[อิทธิพลของบุคคล]} \\
 &+ X_{pi} - \mu_p && \text{[ส่วนที่เหลือ]}
 \end{aligned}$$

การออกแบบแบบแฝง (Nested $i:p$ Design) จะไม่ได้จำแนกเทอมออกมาเป็นอิทธิพลของข้อสอบ โดยอิทธิพลของข้อสอบเป็นส่วนหนึ่งของเทอมส่วนที่เหลือ เนื่องจากบุคคลทำข้อสอบต่างข้อกัน จึงไม่สามารถประมาณค่าอิทธิพลของข้อสอบได้อย่างอิสระจากปฏิสัมพันธ์ระหว่างบุคคลกับข้อสอบได้ สามารถเขียนอยู่ในอีกรูปได้เพื่อแสดงให้เห็นว่าอิทธิพลของข้อสอบเป็นส่วนหนึ่งของส่วนที่เหลือ ดังนี้

$$X_{pi} - \mu_p = (\mu_i - \mu) + (X_{pi} - \mu_i - \mu_p + \mu)$$

จะได้ว่า องค์ประกอบความแปรปรวนของบุคคลหรือความแปรปรวนของคะแนนเอกภพ นิยามเหมือนกับในการออกแบบแบบไขว้ นั่นคือ

$$\sigma_p^2 = E_p (\mu_p - \mu)^2$$

และองค์ประกอบความแปรปรวนของส่วนที่เหลือ มีค่าเป็น

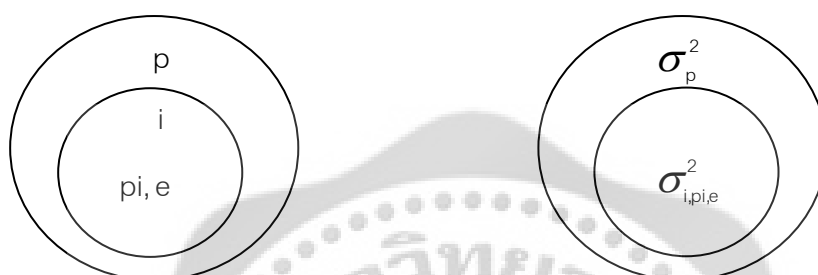
$$\sigma_{i,pi,e}^2 = E_p E_i (X_{pi} - \mu_p)^2$$

จากองค์ประกอบความแปรปรวนของส่วนที่เหลือ แสดงให้เห็นว่าอิทธิพลของข้อสอบนั้นรวมอยู่กับอิทธิพลของปฏิสัมพันธ์ระหว่างบุคคลและข้อสอบ ซึ่งมีแหล่งของความแปรปรวนที่ไม่เป็นระบบหรือไม่สามารถวัดได้รวมอยู่ด้วยเช่นกัน

ดังนั้น ความแปรปรวนของคะแนนสังเกตได้ (X_{pi}) ของบุคคลและข้อสอบทั้งหมดจึงเป็นผลรวมขององค์ประกอบความแปรปรวนสององค์ประกอบ ดังนี้

$$\sigma^2(X_{pi}) = \sigma_p^2 + \sigma_{i,pi,e}^2$$

เขียนแผนภาพเวนน์ได้ดังนี้



แหล่งของความแปรปรวน
(Source of variation)

องค์ประกอบของความแปรปรวน
(Variance component)

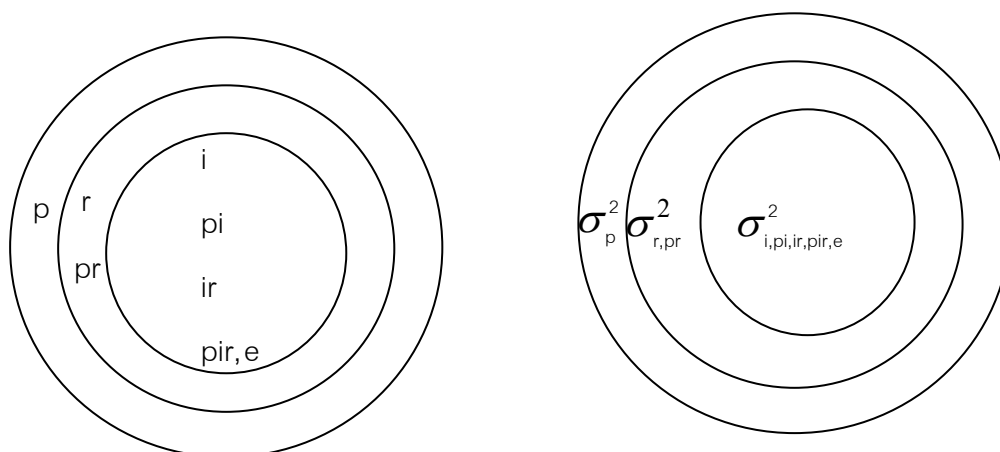
ภาพประกอบ 3 แหล่งของความแปรปรวนและองค์ประกอบของความแปรปรวนสำหรับหนึ่งองค์ประกอบแบบแฝง เมื่อ i แฝงใน p (One Facet, $i : p$ Design)

2.2 กรณีการวัดที่มีสององค์ประกอบแบบแฝง เช่น กรณี $i : r : p$ (Two Facets, $i : r : p$ Design)

คะแนนสังเกตของบุคคลในข้อสอบข้อหนึ่ง ได้รับอิทธิพลมาจากส่วนต่างๆ ดังนี้ (Shavelson; & Webb. 1991: 60)

$$\begin{aligned} X_{pir} &= \mu \\ &+ \mu_p - \mu \\ &+ \mu_{pr} - \mu_p \\ &+ X_{pir} - \mu_{pr} \\ \sigma_{X_{pir}}^2 &= \sigma_p^2 + \sigma_{r,pr}^2 + \sigma_{i,pi,ir,pir,e}^2 \end{aligned}$$

เขียนแผนภาพเวนน์ ได้ดังนี้



แหล่งของความแปรปรวน

(Source of variation)

องค์ประกอบของความแปรปรวน

(Variance component)

ภาพประกอบ 4 แหล่งของความแปรปรวนและองค์ประกอบของความแปรปรวนสำหรับสอง
องค์ประกอบแบบแฝง กรณี $i : r : p$ (Two Facets, $i : r : p$ Design)

3. กรณีออกแบบการวัดแบบผสม (Confound design) เช่น กรณี $(o : p) \times r$ Design

คะแนนสังเกตของบุคคลในข้อสอบข้อหนึ่ง ได้รับอิทธิพลมาจากส่วนต่าง ๆ ดังนี้

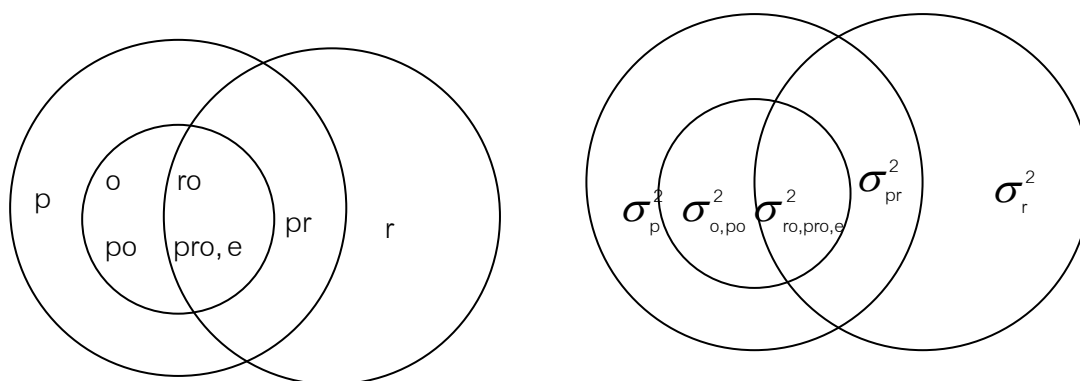
(Shavelson; & Webb. 1991: 52-53)

$$\begin{aligned}
 X_{\text{pro}} = & \mu \\
 & + \mu_p - \mu \\
 & + \mu_r - \mu \\
 & + \mu_{\text{po}} - \mu_p \\
 & + \mu_{\text{pr}} - \mu_p - \mu_r + \mu \\
 & + X_{\text{pro}} - \mu_{\text{pr}} - \mu_{\text{po}} + \mu_p
 \end{aligned}$$

และ

$$\sigma_{X_{\text{pro}}}^2 = \sigma_p^2 + \sigma_r^2 + \sigma_{\text{op}}^2 + \sigma_{\text{pr}}^2 + \sigma_{\text{ro,pro,e}}^2$$

เขียนแผนภาพเวนนได้ดังนี้



แหล่งของความแปรปรวน

(Source of variation)

องค์ประกอบของความแปรปรวน

(Variance component)

ภาพประกอบ 5 แหล่งของความแปรปรวนและองค์ประกอบของความแปรปรวนสำหรับสอง

องค์ประกอบแบบผสมกรณี (o : p) × r (Two – Facet, (o : p) × r Design)

การประมาณค่าองค์ประกอบความแปรปรวน

การประมาณองค์ประกอบความแปรปรวน สามารถนำการวิเคราะห์ความแปรปรวน (Analysis of Variance: ANOVA) มาประยุกต์ใช้กับการศึกษา G (G-Study) (Shavelson; & Webb. 1991: 28-29, 33; Brennan. 2001: 26-27) ดังนี้

1. กรณีการวัดที่มีองค์ประกอบเดียวแบบไขว้ p × i

ตาราง 4 ANOVA สำหรับการวัดที่มีองค์ประกอบเดียวแบบไขว้ p × i

แหล่งความแปรปรวน	Sum of Squares	df	Mean Squares	Expected Mean Squares
Person (p)	SS_p	$n_p - 1$	$MS_p = SS_p/df_p$	$E(MS_p) = \sigma_{pi,e}^2 + n_i \sigma_p^2$
Item (i)	SS_i	$n_i - 1$	$MS_i = SS_i/df_i$	$E(MS_i) = \sigma_{pi,e}^2 + n_p \sigma_i^2$
pi, e	$SS_{pi,e}$	$(n_p - 1)(n_i - 1)$	$MS_{pi,e} = SS_{pi,e}/df_{pi,e}$	$E(MS_{pi,e}) = \sigma_{pi,e}^2$

ที่มา: ชาร์เวลสันและเวบบ์ (Shavelson; & Webb. 1991: 28)

จากตาราง ในการประมาณค่า ค่าเฉลี่ยกำลังสองที่คาดหวัง (Expected mean squares: E(MS)) จะใช้ค่าเฉลี่ยกำลังสองของค่าสังเกต (Observed mean square: MS) เป็นตัวประมาณค่า และการประมาณค่า σ^2 ให้ $\hat{\sigma}^2$ เป็นตัวประมาณค่า จะได้

$$MS_p = \hat{\sigma}_{pi,e}^2 + n_i \hat{\sigma}_p^2$$

$$MS_i = \hat{\sigma}_{pi,e}^2 + n_p \hat{\sigma}_i^2$$

$$MS_{pi,e} = \hat{\sigma}_{pi,e}^2$$

เมื่อจัดสมการใหม่ ได้เป็น

$$\hat{\sigma}_{pi,e}^2 = MS_{pi,e}$$

$$\hat{\sigma}_i^2 = (MS_i - \hat{\sigma}_{pi,e}^2) / n_p = (MS_i - MS_{pi,e}) / n_p$$

$$\hat{\sigma}_p^2 = (MS_p - \hat{\sigma}_{pi,e}^2) / n_i = (MS_p - MS_{pi,e}) / n_i$$

2. กรณีการวัดที่มีสององค์ประกอบแบบไขว้ $p \times i \times r$

ตาราง 5 ANOVA สำหรับการวัดที่มีสององค์ประกอบแบบไขว้ $p \times i \times r$

แหล่งความแปรปรวน	Sum of Squares	df	Mean Squares	Expected Mean Square (EMS)
Persons(p)	ss_p	$n_p - 1$	ss_p / df_p	$\sigma_{pir,e}^2 + n_r \sigma_{pi}^2 + n_i \sigma_{pr}^2 + n_i n_r \sigma_p^2$
Items(i)	ss_i	$n_i - 1$	ss_i / df_i	$\sigma_{pir,e}^2 + n_p \sigma_{ir}^2 + n_r \sigma_{pi}^2 + n_p n_r \sigma_i^2$
Rater(r)	ss_r	$n_r - 1$	ss_r / df_r	$\sigma_{pir,e}^2 + n_p \sigma_{ir}^2 + n_i \sigma_{pr}^2 + n_p \sigma_r^2$
pi	ss_{pi}	$(n_p - 1)(n_i - 1)$	ss_{pi} / df_{pi}	$\sigma_{pir,e}^2 + n_r \sigma_{pi}^2$
pr	ss_{pr}	$(n_p - 1)(n_r - 1)$	ss_{pr} / df_{pr}	$\sigma_{pir,e}^2 + n_i \sigma_{pr}^2$
ir	ss_{ir}	$(n_i - 1)(n_r - 1)$	ss_{ir} / df_{ir}	$\sigma_{pir,e}^2 + n_p \sigma_{ir}^2$
pir,e	$ss_{pir,e}$	$(n_p - 1)(n_i - 1)(n_r - 1)$	$ss_{pir,e} / df_{pir,e}$	$\sigma_{pir,e}^2$

ที่มา: ชาร์เวลสันและเวบบ์ (Shavelson; & Webb. 1991: 33)

การประมาณค่าดังกล่าวในบางสถานการณ์ องค์ประกอบความแปรปรวนอาจจะมีค่าติดลบได้ เนื่องจากการกำหนดโมเดลการวัดผิดพลาดหรือเกิดจากความคลาดเคลื่อนจากการสุ่มตัวอย่าง หากเกิดเหตุการณ์แบบนี้ขึ้น ครอนบาคและคณะ (Cronbach; et al. 1972) ได้เสนอให้ค่าประมาณที่ติดลบมีค่าเป็นศูนย์ แล้วจึงคำนวณสมการอื่นต่อไป ทำให้ง่ายสำหรับการคำนวณสมการอื่น ๆ แต่จะเกิดอคติในการประมาณองค์ประกอบความแปรปรวนค่าอื่น ๆ อีกวิธีหนึ่งเสนอโดย เบรนนอน (Shavelson; & Webb. 1991; citing Brennan. 1983) ซึ่งเสนอให้ใช้ค่าศูนย์แทนค่าติดลบเช่นกัน แต่ในการประมาณค่าองค์ประกอบอื่น ๆ ให้ใช้ค่าเดิมในการคำนวณ เมื่อประมาณค่าหมดทุกสมการแล้วจึงรายงานค่าประมาณที่ติดลบเป็นศูนย์ วิธีการนี้จะยุ่งยากในการคำนวณแต่จะแก้ปัญหาคติในการประมาณค่าองค์ประกอบประกอบในสมการต่าง ๆ ได้

ความแปรปรวนของคะแนนความคลาดเคลื่อน

คะแนนเอกภพ (Universe score) คือ ค่าเฉลี่ยของการวัดซ้ำหลาย ๆ ครั้ง ตามเงื่อนไขของการวัดในเอกภพของการสุ่มอ้างอิง สำหรับความคลาดเคลื่อนของการวัด (E_{pi}) ถูกจำแนกออกมาเป็นความคลาดเคลื่อนจากองค์ประกอบหรือกลุ่มเงื่อนไขการวัด (E_i) และความคลาดเคลื่อนจากแหล่งที่เหลือนอื่น ๆ (e_{pi}) ซึ่งการวัดแต่ละครั้งไม่จำเป็นต้องเป็นข้อสอบคู่ขนาน ส่วนความแปรปรวนของค่าคาดหวังของคะแนนที่สังเกตได้เป็นผลรวมของความแปรปรวนของคะแนนเอกภพ ($\sigma_{\mu_p}^2$ หรือ σ_p^2) กับความแปรปรวนของความคลาดเคลื่อนจากแหล่งอื่น (σ_{ep}^2) ดังนี้

$$X_{pi} = T_{pi} + E_{pi} + e_{pi}$$

$$\sigma_{X_p}^2 = \sigma_{\mu_p}^2 + \sigma_{E_i}^2 + \sigma_{e_p}^2$$

ความแปรปรวนของคะแนนความคลาดเคลื่อนจากองค์ประกอบต่าง ๆ ของการวัดแบ่งออกเป็น 2 ประเภท ได้แก่ (Shavelson; & Webb. 1991: 84; ศิริชัย กาญจนวาสี. 2550: 17-18)

1. ความแปรปรวนของคะแนนความคลาดเคลื่อนแบบสัมบูรณ์ (Absolute error variance) ใช้สัญลักษณ์ σ_{Abs}^2 หรือ $\sigma^2(\Delta)$ คือ ความแปรปรวนของผลต่างระหว่างคะแนนสังเกตและคะแนนเอกภพ ซึ่งมีค่าเท่ากับ ผลรวมของความแปรปรวนที่ประมาณได้ทั้งหมด ยกเว้นความแปรปรวนจากแหล่งบุคคล (σ_p^2) หรือสิ่งที่ต้องการวัด ดังนั้นจึงคำนวณได้จาก ผลรวมของความแปรปรวนของคะแนนจากแหล่งต่าง ๆ ยกเว้น σ_p^2 ดังเช่น กรณี $p \times i \times r$ คำนวณ σ_{Abs}^2 ได้จาก

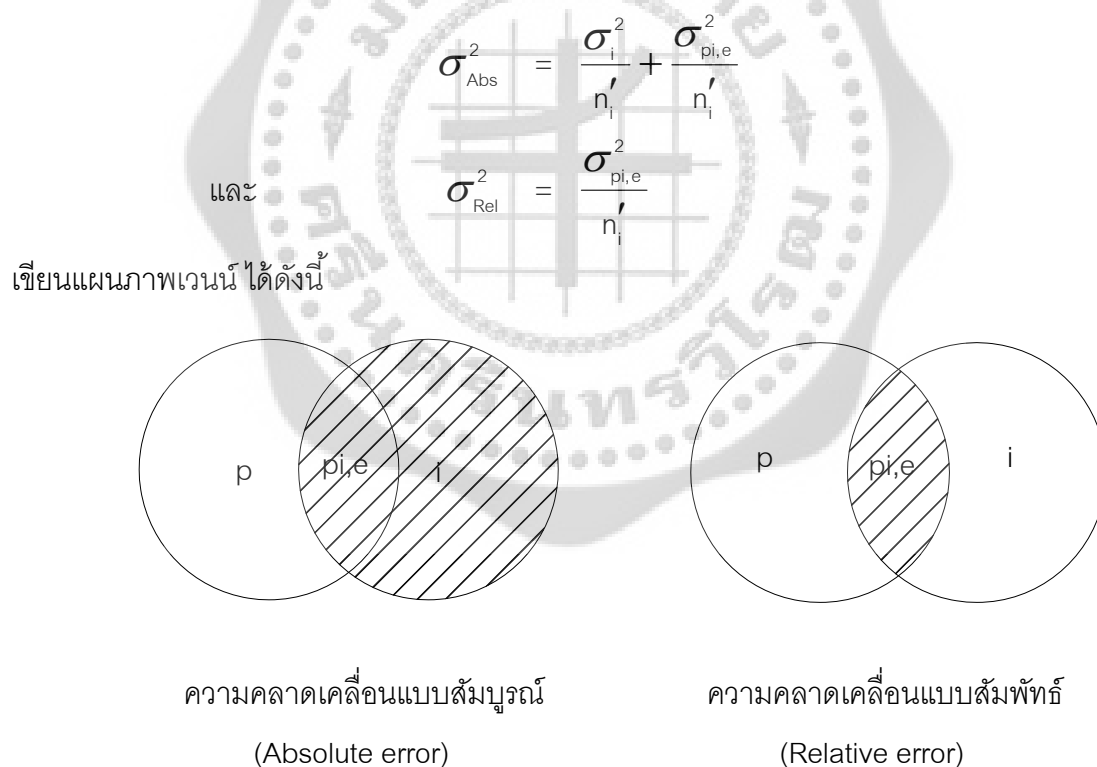
$$\sigma_{Abs}^2 = \sigma_i^2 + \sigma_r^2 + \sigma_{pi}^2 + \sigma_{pr}^2 + \sigma_{ri}^2 + \sigma_{pir,e}^2$$

2. ความแปรปรวนของคะแนนความคลาดเคลื่อนแบบสัมพัทธ์ (Relative error variance) ใช้สัญลักษณ์ σ_{Rel}^2 หรือ $\sigma^2(\delta)$ คือ ความแปรปรวนของผลต่างระหว่างส่วนเบี่ยงเบนคะแนนสังเกต(ความแตกต่างของคะแนนสังเกตจากคะแนนเฉลี่ยของประชากรของคะแนนสังเกต) กับส่วนเบี่ยงเบนคะแนนเอกภพ (ความแตกต่างของคะแนนเอกภพจากคะแนนเฉลี่ยของประชากรของคะแนนเอกภพ) จึงคำนวณได้จากผลรวมของความแปรปรวนของคะแนนจากแหล่งต่าง ๆ ที่มีปฏิสัมพันธ์กับผู้สอบ ดังเช่น กรณี $p \times i \times r$ คำนวณ σ_{Rel}^2 ได้จาก

$$\sigma_{\text{Rel}}^2 = \sigma_{\text{pi}}^2 + \sigma_{\text{pr}}^2 + \sigma_{\text{pir,e}}^2$$

แหล่งความคลาดเคลื่อนแบบสัมบูรณ์และแบบสัมพัทธ์ จะแตกต่างกันไปตามรูปแบบการวัดที่ได้ออกแบบไว้ มีรายละเอียดดังต่อไปนี้ (Shavelson; & Webb, 1991: 84-92)

1. กรณีการวัดที่มีองค์ประกอบเดียว แบบ $p \times i$ Design (One-facet crossed design)



ภาพประกอบ 6 แหล่งความคลาดเคลื่อนแบบสัมบูรณ์และความคลาดเคลื่อนแบบสัมพัทธ์

สำหรับ $p \times i$ Design

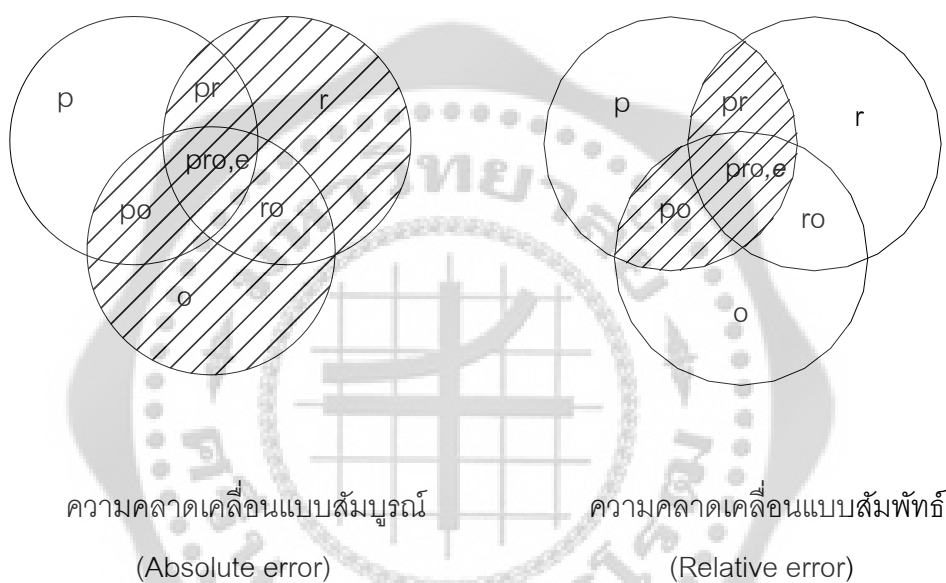
2. กรณีการวัดที่มีสององค์ประกอบแบบ $p \times r \times o$ Design

$$\sigma_{\text{Abs}}^2 = \frac{\sigma_r^2}{n'_r} + \frac{\sigma_o^2}{n'_o} + \frac{\sigma_{pr}^2}{n'_r} + \frac{\sigma_{po}^2}{n'_o} + \frac{\sigma_{ro}^2}{n'_r n'_o} + \frac{\sigma_{pro,e}^2}{n'_r n'_o}$$

และ

$$\sigma_{\text{Rel}}^2 = \frac{\sigma_{pr}^2}{n'_r} + \frac{\sigma_{po}^2}{n'_o} + \frac{\sigma_{pro,e}^2}{n'_r n'_o}$$

เขียนแผนภาพเวนน์ได้ดังนี้



ภาพประกอบ 7 แหล่งความคลาดเคลื่อนแบบสัมบูรณ์ และความคลาดเคลื่อนแบบสัมพัทธ์
สำหรับ $p \times r \times o$ Design

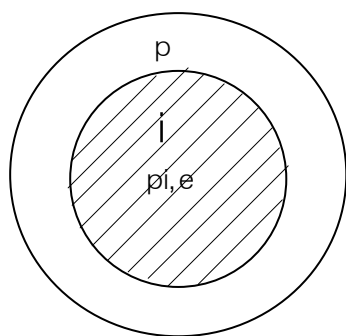
3. กรณีการวัดที่มีองค์ประกอบเดียว แบบ $i:p$ Design

$$\sigma_{\text{Abs}}^2 = \frac{\sigma_{i,pi,e}^2}{n'_i}$$

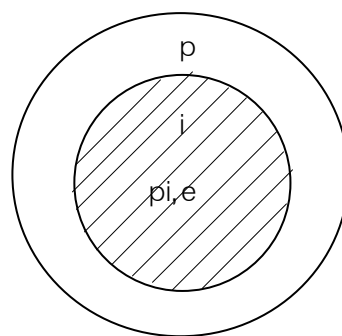
และ

$$\sigma_{\text{Rel}}^2 = \frac{\sigma_{i,pi,e}^2}{n'_i}$$

เขียนแผนภาพเวนน์ได้ดังนี้



ความคลาดเคลื่อนแบบสัมบูรณ์
(Absolute error)



ความคลาดเคลื่อนแบบสัมพัทธ์
(Relative error)

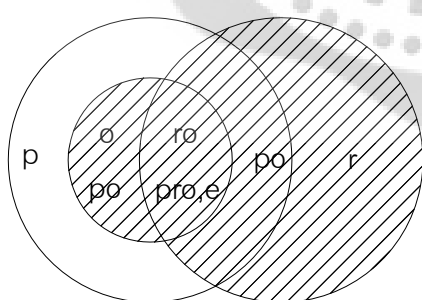
ภาพประกอบ 8 แหล่งความคลาดเคลื่อนแบบสัมบูรณ์ และความคลาดเคลื่อนแบบสัมพัทธ์
สำหรับ i:p Design

4. กรณีการวัดที่มีสององค์ประกอบแบบ (o:p) × r Design

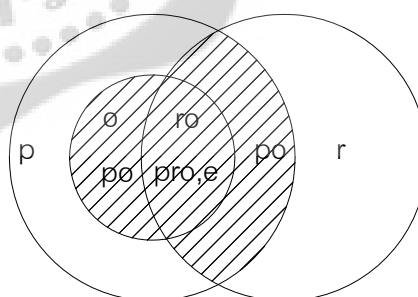
$$\sigma_{\text{Abs}}^2 = \frac{\sigma_r^2}{n'_r} + \frac{\sigma_{pr}^2}{n'_r} + \frac{\sigma_{o,po}^2}{n'_o} + \frac{\sigma_{ro,pro,e}^2}{n'_r n'_o}$$

$$\sigma_{\text{Rel}}^2 = \frac{\sigma_{pr}^2}{n'_r} + \frac{\sigma_{o,po}^2}{n'_o} + \frac{\sigma_{ro,pro,e}^2}{n'_r n'_o}$$

เขียนแผนภาพเวนน ได้ดังนี้



ความคลาดเคลื่อนแบบสัมบูรณ์
(Absolute error)



ความคลาดเคลื่อนแบบสัมพัทธ์
(Relative error)

ภาพประกอบ 9 แหล่งความคลาดเคลื่อนแบบสัมบูรณ์ และความคลาดเคลื่อนแบบสัมพัทธ์
สำหรับ (o:p) × r Design

ค่าสัมประสิทธิ์การสรุปอ้างอิง (Generalizability Coefficient)

ทฤษฎีการสรุปอ้างอิง นอกจากมุ่งให้ความสำคัญกับองค์ประกอบความแปรปรวนและความคลาดเคลื่อนของการวัดแล้วยังสามารถประมาณค่าสัมประสิทธิ์การสรุปอ้างอิง ซึ่งคล้ายคลึงกับค่าสัมประสิทธิ์ความเชื่อมั่นในทฤษฎีการทดสอบแบบมาตรฐานเดิม (Shavelson; & Webb. 1991: 92) ที่เป็นอัตราส่วนระหว่างความแปรปรวนของคะแนนจริง กับความแปรปรวนของคะแนนที่สังเกตได้

ค่าสัมประสิทธิ์การสรุปอ้างอิงเป็นอัตราส่วนระหว่างความแปรปรวนของคะแนนเอกภพและความแปรปรวนของคะแนนสังเกตที่คาดหวัง สามารถประมาณค่าได้จากกำลังสองของค่าสหสัมพันธ์ระหว่างคะแนนเอกภพกับคะแนนสังเกต (Cronbach; et al. 1972) ดังนั้น สัมประสิทธิ์การสรุปอ้างอิงจึงเป็นอัตราส่วนระหว่างความแปรปรวนของคะแนนเอกภพ กับความแปรปรวนของค่าคาดหวังของคะแนนสังเกต ดังนี้

$$G\text{-coefficient } (\rho^2) = \frac{\sigma_p^2}{\sigma_p^2 + \text{Error Variance}}$$

เนื่องจากความแปรปรวนของคะแนนความคลาดเคลื่อนจากองค์ประกอบต่าง ๆ ของการวัดมี 2 ประเภท จึงทำให้สัมประสิทธิ์การสรุปอ้างอิงมี 2 ประเภทดังนี้

1. สัมประสิทธิ์การสรุปอ้างอิงสำหรับการตัดสินใจเชิงสัมบูรณ์ (ρ_{Abs}^2 หรือ $\rho^2(\Delta)$)

เมื่อคะแนนความคลาดเคลื่อน เป็นความแปรปรวนของคะแนนความคลาดเคลื่อนสัมบูรณ์ สัมประสิทธิ์นี้บ่งบอกความเชื่อมั่นของแบบทดสอบ ในสถานการณ์ของการตัดสินใจที่ขึ้นกับคะแนนของผู้สอบตามลำดับ ไม่มีการเปรียบเทียบภายในกลุ่มหรือระหว่างกลุ่ม เช่น แบบทดสอบอิงเกณฑ์ ที่ใช้ตรวจสอบความสามารถของผู้สอบเทียบกับเกณฑ์ที่กำหนดไว้ล่วงหน้า

การประมาณค่าสัมประสิทธิ์การสรุปอ้างอิงเชิงสัมบูรณ์เทียบเท่ากับค่าความเชื่อมั่นในทฤษฎีการทดสอบแบบมาตรฐานเดิม เบรนนอน และเคน (Shavelson; & Webb. 1991: 93; citing Brennan; & Kane.1977. An index of dependability for mastery test. Journal of Educational Measurement. 14: 277-289) เรียกสัมประสิทธิ์นี้ว่า ดัชนีความเชื่อถือได้ (Index of dependability) และใช้สัญลักษณ์ ϕ (phi) หรือ ρ_{Abs}^2 มีสูตรดังนี้ (Shavelson; & Webb. 1991: 93)

$$\phi = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{\text{Abs}}^2}$$

หรือ (Brennan. 2001: 35)

$$\phi = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(\Delta)}$$

2. สัมประสิทธิ์การสรุปอ้างอิงสำหรับการตัดสินใจเชิงสัมพัทธ์ (ρ_{Rel}^2)

เมื่อคะแนนความคลาดเคลื่อนเป็นความแปรปรวนของคะแนนความคลาดเคลื่อนสัมพัทธ์ สัมประสิทธิ์นี้บ่งบอกความเชื่อมั่นของแบบทดสอบ ในสถานการณ์ของการตัดสินใจที่มีการเปรียบเทียบคะแนนระหว่างผู้สอบ เช่น แบบทดสอบอิงกลุ่มที่เปรียบเทียบคะแนนระหว่างผู้สอบด้วยการเข้าโค้งการแจกแจงคะแนนและตัดเกรดหรือให้ระดับผลการเรียน หรือการสอบแข่งขันเพื่อหาผู้ที่ได้คะแนนสูงเข้าศึกษา การประมาณค่าสัมประสิทธิ์การสรุปอ้างอิงเชิงสัมพัทธ์ มีสูตรดังนี้ (Shavelson; & Webb. 1991: 93)

$$E\rho_{\text{Rel}}^2 = \frac{\sigma_p^2}{E\sigma^2(X_{pi})} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{\text{Rel}}^2}$$

หรือ (Brennan. 2001: 35)

$$E\rho^2 = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(\delta)}$$

กล่าวโดยสรุปได้ว่า ทฤษฎีการสรุปอ้างอิงขยายมาจากทฤษฎีการทดสอบแบบมาตรฐานเดิมใน 4 ลักษณะ คือ 1) สามารถประมาณค่าขนาดของความแปรปรวนจากหลาย ๆ แหล่งในบริบทการวัดได้พร้อมกันจากการวิเคราะห์เพียงครั้งเดียว 2) สามารถประมาณค่าสัมประสิทธิ์ความเชื่อมั่น ที่เรียกว่า สัมประสิทธิ์การสรุปอ้างอิง (Generalizability coefficient) มุ่งเน้นไปที่องค์ประกอบความแปรปรวนของแต่ละแหล่งของความคลาดเคลื่อนที่ส่งผลต่อการวัด 3) สามารถแยกการตัดสินใจเป็น 2 ลักษณะ คือ การตัดสินใจเชิงสมบูรณ์ (Absolute decision) และการตัดสินใจเชิงสัมพัทธ์ (Relative decision) และ 4) มีการวิเคราะห์ 2 ขั้นตอนแยกกัน คือ การศึกษา G และการศึกษา D โดยการศึกษา G เป็นการประมาณค่าน้ำหนักของแหล่งความคลาดเคลื่อนของการวัดที่เป็นไปได้ทั้งหมด การศึกษา D ใช้ข้อมูลจากการศึกษา G ในการออกแบบการวัดที่ทำให้มีความคลาดเคลื่อนน้อยที่สุดตามเป้าหมายที่ต้องการ การประมาณค่าองค์ประกอบความแปรปรวนในการศึกษา G จะใช้ในการออกแบบกระบวนการวัดให้มีความคุ้มค่าทั้งเวลาและค่าใช้จ่าย สามารถนำไปใช้ได้หลายรูปแบบของ

การวัด ทั้งรูปแบบเป็นแบบไขว้ (Crossed) แบบแฝง (Nested) หรือแบบแฝงบางส่วน (Partially nested)

จากการศึกษางานวิจัย (Ibrahim. 1984; O'Brien; & Jones. 1986; Brennan; Gao; & Colton.1995; Lane; et al. 1996; Swartz, et al. 1999; ไพรัตน์ วงษ์นาม. 2533; ชัยฤทธิ์ คีลาเดช. 2540; ปวีณา ปิอาทิตย์. 2545; พรรณี เจียมสุบุตร. 2543; อุษณีย์ บัวศิริพันธ์. 2543) พบว่า มีการนำทฤษฎีนี้ไปใช้อย่างกว้างขวางในหลายลักษณะ เช่น การวัดประเมินการปฏิบัติ การวัดประเมินทางการพูด การเขียน และการแก้โจทย์ปัญหา เป็นต้น การวัดประเมินเหล่านี้ต่างมีแหล่งความคลาดเคลื่อนของคะแนนหลายแหล่ง และมีปัญหาในแง่ของความน่าเชื่อถือของคะแนน อันเนื่องมาจากผู้ตรวจให้คะแนน จึงมีการนำทฤษฎีการสรุปอ้างอิงไปใช้ในการประมาณค่าองค์ประกอบความแปรปรวนจากแหล่งต่าง ๆ เพื่อศึกษาว่าแหล่งใดเป็นแหล่งของความคลาดเคลื่อนหลักและประมาณค่าสัมประสิทธิ์การสรุปอ้างอิง ผลการวิจัยสอดคล้องกันในประเด็นที่ว่า ผู้ตรวจให้คะแนนเป็นแหล่งความคลาดเคลื่อนที่สำคัญ ส่งผลต่อความน่าเชื่อถือและความเที่ยงตรงของคะแนน และนักวิจัยเหล่านี้ต่างได้ให้ข้อเสนอแนะถึงจำนวนผู้ตรวจว่าควรใช้ผู้ตรวจมากกว่าหนึ่งคน แต่จำนวนผู้ตรวจที่เหมาะสมนั้นผลการศึกษายังพบแตกต่างกัน นอกจากนี้งานวิจัยส่วนใหญ่เป็นการศึกษาในกรณีที่ผู้ตรวจตรวจให้คะแนนข้อสอบทุกข้อของผู้สอบทุกคน ซึ่งค่อนข้างใช้เวลาในการตรวจมาก และในการปฏิบัติจริงของการวัดประเมินในระดับชั้นเรียนนั้นต่างมีจำนวนครูผู้สอนที่จำกัดและนักเรียนมีจำนวนมาก ทั้งนี้รูปแบบการเก็บรวบรวมข้อมูลหรือการตรวจให้คะแนนนั้นมีหลากหลายรูปแบบสามารถนำมาใช้วางแผนการตรวจให้คะแนนได้ เช่น แบบไขว้ แบบแฝง หรือแบบแฝงบางส่วน ซึ่งสามารถลดเวลาและภาระในการตรวจให้คะแนนได้

ผู้วิจัยจึงสนใจศึกษาลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน โดยใช้โมเดลการสรุปอ้างอิง ซึ่งผู้วิจัยได้ศึกษาองค์ประกอบความแปรปรวนในแต่ละแหล่ง พิจารณาลักษณะของคะแนนจากค่าสัมประสิทธิ์การสรุปอ้างอิงและความเที่ยงตรงตามสภาพ โดยทำการประมาณค่าและเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิง เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน ประมาณค่าและเปรียบเทียบความเที่ยงตรงตามสภาพของคะแนน เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจต่างกัน โดยนำคะแนนที่ได้ไปหาค่าสัมประสิทธิ์สหสัมพันธ์กับคะแนนที่ใช้เป็นเกณฑ์ซึ่งเป็นผลการทดสอบทางการศึกษาระดับชาติด้านพื้นฐาน (Ordinary National Education Test: O-NET) เพื่อตรวจสอบคุณลักษณะของคะแนนได้แก่ ค่าสัมประสิทธิ์การสรุปอ้างอิงและความเที่ยงตรงตามสภาพของคะแนนแบบทดสอบปลายเปิด เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน อันจะเป็นแนวทางในการนำไปใช้ต่อไป

อย่างไรก็ตามทฤษฎีการสรุปอ้างอิงนี้จะประมาณค่าความเชื่อมั่นโดยใช้คะแนนดิบ ยังไม่ได้ปรับแก้คะแนนดิบที่เกิดจากความแตกต่างของความเข้มงวดของผู้ตรวจหรือความยากของข้อสอบ ตลอดจนข้อค้นพบของโมเดลนี้เป็นการศึกษาในระดับกลุ่ม (Group level) หรือโดยภาพรวมทั้งหมด ยังไม่สามารถระบุได้ถึงระดับแต่ละสมาชิก (Individual level) ในองค์ประกอบได้ (Linacre. 1993; Stemler. 2004: online; Lee; & Kantor. 2005: 40-41; Iramaneerat; et al. 2007: online) การศึกษาถึงรายละเอียดในระดับแต่ละสมาชิก (Individual level) หรือโดยภาพย่อยนั้นมีการพัฒนาโมเดลหลายองค์ประกอบของราส์ซ (Many-Facet Rasch Model: MFRM) ตามทฤษฎีการตอบข้อสอบ (Item Response Theory: IRT) ขึ้น ซึ่งโมเดลนี้สามารถศึกษาแหล่งความแปรปรวนได้หลายแหล่งและศึกษาได้ถึงระดับแต่ละสมาชิกในแต่ละองค์ประกอบ โดยขอล่าวถึงรายละเอียดในลำดับถัดไป

โมเดลหลายองค์ประกอบของราส์ซ (Many-Facet Rasch Model: MFRM)

ในการนำเสนอโมเดลหลายองค์ประกอบของราส์ซ ผู้วิจัยขอนำเสนอประเด็นต่าง ๆ ต่อไปนี้ตามลำดับ คือ แนวคิดพื้นฐานของโมเดลหลายองค์ประกอบของราส์ซ โมเดลทางคณิตศาสตร์ ค่าสถิติต่าง ๆ ที่สำคัญจากผลการวิเคราะห์ ได้แก่ สถิติทดสอบความสอดคล้องของโมเดล (Fit statistics) ความเชื่อมั่นแยกส่วน (Reliability of separation index) สถิติแยกส่วน (Separation statistics) และสถิติทดสอบไค-สแควร์ (Fixed effect chi-square test) มีรายละเอียดในแต่ละประเด็น ดังนี้

แนวคิดพื้นฐานของโมเดลหลายองค์ประกอบของราส์ซ

แนวคิดโมเดลของราส์ซ (Rasch Model) ตามทฤษฎีการตอบข้อสอบนั้นจะให้ค่าพารามิเตอร์สองตัว คือ พารามิเตอร์ข้อคำถาม (Item parameter) คือ ค่าความยากของข้อสอบ (Item difficulty) กับพารามิเตอร์ของผู้สอบ (Examinee parameter) ที่แสดงถึง ความสามารถของผู้สอบ (Examinee ability) ซึ่งในระยะแรกนั้นจะเป็นการตรวจให้คะแนนแบบ 2 ค่า (Dichotomous) แต่การวัดทางพฤติกรรมศาสตร์นั้นไม่ได้จำกัดอยู่เพียงการให้คะแนนแบบ 2 ค่าเท่านั้น เช่น การวัดเจตคติ บุคลิกภาพ การวัดการปฏิบัติ นั้นมีการให้คะแนนแบบหลายค่า (Polytomous) ทำให้ในระยะเวลาดังกล่าวจะมีการพัฒนาโมเดลสำหรับการตรวจให้คะแนนแบบหลายค่า โดยรูปแบบของเครื่องมือที่นิยมใช้กัน เช่น มาตรฐานค่า (Rating scale) การตรวจให้คะแนนความรู้บางส่วน (Partial credit) การตรวจให้คะแนนตามลำดับขั้นของรายการหลายคำตอบ (Order-response categories) เป็นต้น ซึ่งโมเดลที่พัฒนาขึ้นมาจากโมเดลของราส์ซ เช่น เรทติงสเกลโมเดลของแอนดริช (Andrich. 1978) พาเชียลเครดิตโมเดลของมาสเตอร์ (Master. 1982)

ต่อมาในปี 1989 ลินาเคอร์ (Linacre.1994) ได้ปรับขยายโมเดลของราสช์ สำหรับการตอบข้อสอบแบบลำดับขั้น (Order categories) ข้อสอบที่มีการตรวจให้คะแนนแบบบางส่วน (Partial credit test items) ซึ่งจากเดิมโมเดลของราสช์จะมุ่งเน้นไปที่สององค์ประกอบคือ ข้อสอบและผู้สอบ แต่ในทางปฏิบัติของการตรวจให้คะแนนแบบหลายค่า นั้น มีอีกองค์ประกอบหนึ่งที่มีผลต่อการให้คะแนนคือ ผู้ตัดสิน (Judge) หรือ ผู้ตรวจ (Rater) ซึ่งเป็นผู้ให้คะแนนผู้สอบจากผลการปฏิบัติจากข้อสอบที่กำหนดให้ โดย ลินาเคอร์ (Linacre.1994) ได้ปรับขยายโมเดลไปสู่สถานการณ์ที่มากกว่าสององค์ประกอบ (Facets) โดยเพิ่มองค์ประกอบที่สนใจเข้าไปในโมเดล ทำให้สามารถศึกษาได้มากกว่า 2 องค์ประกอบ เช่น เพิ่มองค์ประกอบผู้ตรวจเข้าไปในโมเดล จึงทำให้สามารถประมาณค่าพารามิเตอร์ของผู้ตรวจ (Rater parameter) ได้ เป็นต้น โมเดลที่พัฒนาขึ้นนี้เรียกว่า โมเดลหลายองค์ประกอบของราสช์ (Many Facets Rasch Model: MFRM) ลินาเคอร์มีแนวคิดว่าการได้คะแนนในระดับใดระดับหนึ่งของผู้สอบได้รับอิทธิพลจาก 4 องค์ประกอบ คือ ความสามารถของผู้สอบ (Ability) ความยากของข้อสอบ (Difficulty) ความเข้มงวดของผู้ตรวจ (Rater severity) และโครงสร้างของมาตราประมาณค่า (Rating scale) (Linacre. 1994) ฟังก์ชันความสัมพันธ์ขององค์ประกอบทั้งหมดจะแสดงด้วยสมการทางคณิตศาสตร์ สามารถใช้ตรวจสอบคุณภาพของการตรวจให้คะแนนตอบที่เป็นลำดับขั้น (Order response categories) การตรวจให้คะแนนแบบบางส่วน (Partial credit) โดย MFRM สามารถบ่งชี้พฤติกรรมการตรวจให้คะแนนของผู้ตรวจว่ามีความเข้มงวดหรือใจดีมากเกินไปใด การวิเคราะห์ข้อมูลโดยใช้ MFRM จึงเป็นการศึกษาอิทธิพลของแหล่งความคลาดเคลื่อนต่าง ๆ ที่มีต่อคุณภาพของการวัดประเมิน โดยผู้วิจัยสามารถใช้โครงสร้างทางสถิติในการวิเคราะห์ผลการประเมินเพื่อการสรุปรูปแบบ (Pattern) ของผลหลักสำหรับผู้ตรวจ (Rater) ผู้สอบ (Ratee) และคุณลักษณะ (Trait) หรือข้อสอบ (Item) และตัวแปรอื่น ๆ ที่สนใจ ทั้งนี้การวิเคราะห์ด้วย MFRM สามารถแยกคำนวณในแต่ละองค์ประกอบ (Facet) อย่างเป็นอิสระจากองค์ประกอบอื่น (Myford; & Wolfe. 2003: 402) นอกจากนี้แล้วยังสามารถประเมินในระดับบุคคล (Individual – level effects) ได้แก่ ผู้ตรวจแต่ละคน ผู้สอบแต่ละคน คุณลักษณะหรือข้อสอบแต่ละข้อ ซึ่งเป็นการศึกษาในระดับสมาชิกภายในแต่ละองค์ประกอบว่ามีการลักษณะอย่างไร ซึ่งหากทราบข้อมูลเหล่านี้จะช่วยให้เกิดการปรับปรุงพัฒนาการวัดประเมินมากขึ้น ในอีกทางหนึ่งการวิเคราะห์ด้วย MFRM คล้ายคลึงกับการวิเคราะห์ความแปรปรวน (ANOVA) ที่สามารถศึกษาถึงอิทธิพลหลัก (Main effects) ปฏิสัมพันธ์ (Interactions) แต่ว่า MFRM นั้นสามารถตรวจสอบได้ถึงระดับแต่ละสมาชิก (Myford; & Wolfe. 2003: 404)

โมเดลทางคณิตศาสตร์ (The Mathematical Model)

โมเดลหลายองค์ประกอบของราส์มีลักษณะสำคัญคือ เป็นโมเดลแบบเส้นตรงเชิงบวก (Additive Linear Model) ที่ใช้การแปลงคะแนนสังเกตของผู้ตอบแต่ละคน แต่ละรายการของกลุ่มตัวอย่างให้อยู่ในรูปของคะแนนโลจิท (Logit Score - Log Odds score) และใช้วิธีการวิเคราะห์ข้อมูลตามแนวทางของการถดถอย (Regression) โดยมีอัตราส่วนของชั้นความน่าจะเป็นที่ต่อเนื่องกัน (Successive category probabilities; Log odds) เป็นตัวแปรตามและองค์ประกอบต่าง ๆ (Facets) เป็นตัวแปรอิสระ

หากวัดความสามารถของผู้สอบด้วยข้อสอบแบบเลือกตอบ (Multiple choice items) โมเดลทางคณิตศาสตร์ของราส์ที่สอดคล้องกับข้อมูลเชิงประจักษ์ที่มีการให้คะแนนแบบสองค่า (Dichotomous) สามารถเขียนเป็นสมการได้ดังนี้ (Engelhard. 1994: 95)

$$\ln[P_{ni1}/P_{ni0}] = \beta_n - \delta_i$$

เมื่อ P_{ni1} คือ ความน่าจะเป็นที่ผู้ตอบคนที่ n จะตอบข้อสอบข้อที่ i ได้ถูกต้อง ($x=1$)
 P_{ni0} คือ ความน่าจะเป็นที่ผู้ตอบคนที่ n จะตอบข้อสอบข้อที่ i ไม่ถูกต้องหรือตอบผิด ($x=0$)
 β_n คือ ระดับความสามารถของผู้ตอบคนที่ n
 δ_i คือ ค่าความยากของข้อสอบข้อที่ i

โมเดลนี้มี 2 องค์ประกอบ คือ องค์ประกอบความสามารถของนักเรียน (Student competence) และองค์ประกอบความยากของข้อสอบ (Item difficulty)

นอกจากนี้แล้วยังสามารถปรับขยายโมเดลให้เข้ากับข้อมูลมาตราประเมิน (Rating scale) การตรวจให้คะแนนความรู้บางส่วน (Partial credit) ซึ่งเป็นการให้คะแนนแบบหลายค่า (Polytomous) ที่มีหลายองค์ประกอบ ในกรณีโมเดลที่มี 3 องค์ประกอบ ที่ประกอบด้วย ความสามารถของนักเรียน ความยากของข้อสอบและความเข้มงวดของผู้ตรวจ โมเดลทางคณิตศาสตร์ของราส์ที่สอดคล้องกับข้อมูลในการให้คะแนนแบบหลายค่านี้ ซึ่งสามารถเขียนเป็นสมการได้ดังนี้ (Linacre. 1994: 1; Engelhard. 1994: 95)

$$\ln[P_{nij}/P_{nij-1}] = \beta_n - \delta_i - \lambda_j - \tau_k$$

เมื่อ P_{nijk} คือ ความน่าจะเป็นที่ผู้ตรวจ j จะประเมินผู้ตอบคนที่ n โดยให้ระดับ
คะแนน k จากการตอบข้อความรายการที่ i

P_{nijk-1} คือ ความน่าจะเป็นที่ผู้ตรวจ j จะประเมินผู้ตอบคนที่ n โดยให้ระดับ
คะแนน $k-1$ จากการตอบข้อความรายการที่ i

β_n คือ ระดับความสามารถของผู้ตอบคนที่ n

δ_i คือ ค่าความยากของข้อความรายการที่ i

λ_j คือ ค่าความเข้มงวดของผู้ตรวจ j

τ_k คือ ค่าความยากประจำขั้นที่ k สัมพันธ์กับขั้นที่ $k-1$

โดยที่ τ_k ถือว่าเป็นค่าพารามิเตอร์ ไม่นับว่าเป็นองค์ประกอบของโมเดล

จากสมการข้างต้น ตัวแปรตามคือคะแนนโลจิต ตัวแปรอิสระคือ องค์ประกอบต่างๆ (Facets)
อันได้แก่ ผู้สอบ คุณลักษณะ(ข้อสอบ)และผู้ตรวจ

โมเดลทางคณิตศาสตร์นี้แสดงถึงความน่าจะเป็นของผู้สอบที่จะได้รับการประเมิน k มากกว่า
 $k-1$ ในคุณลักษณะ i จากผู้ตรวจ j ขึ้นอยู่กับผลเชิงบวกของความสามารถของผู้สอบ (β_n) ความยาก
ของคุณลักษณะ (δ_i) ความเข้มงวดของผู้ตรวจ (λ_j) และความยากสัมพัทธ์ของรายการประเมิน k
เมื่อเปรียบเทียบกับรายการประเมิน $k-1$ ในทุกคุณลักษณะและผู้ตรวจทุกคน (τ_k) โมเดลนี้สามารถ
แสดงให้อยู่ในรูปของความน่าจะเป็นของการได้รับการประเมิน เมื่อให้พารามิเตอร์ของผู้สอบ
คุณลักษณะ ผู้ตรวจ และ Thresholds ของรายการมาตรวจประเมิน (Myford; & Wolfe. 2003: 404)
ดังนี้

$$P_{nijk} = \frac{\exp \sum_{c=0}^k (\beta_n - \delta_i - \lambda_j - \tau_k)}{\sum_{h=0}^K \exp \sum_{c=0}^k (\beta_n - \delta_i - \lambda_j - \tau_k)}$$

เมื่อ P_{nijk} คือ ความน่าจะเป็นของผู้สอบ n ได้รับการประเมิน x ในคุณลักษณะ i
จากผู้ตรวจ j

k คือ รายการของมาตรการประเมิน

c คือ ดัชนีแสดงจำนวนมีค่าตั้งแต่ 0 ถึง k

K คือ รายการมาตรการประเมินที่มากที่สุด

h คือ ดัชนีแสดงจำนวนมีค่าตั้งแต่ 0 ถึง K

$$F_0 \equiv 0$$

ในการวิเคราะห์ องค์ประกอบต่าง ๆ จะถูกวิเคราะห์พร้อมกันอย่างเป็นอิสระ และเมื่อปรับ (Calibrate) ให้อยู่บนมาตรเชิงเส้นเดียว (มาตรโลจิต) แล้วทำให้สามารถวัดความเข้มงวดของผู้ตรวจให้อยู่บนมาตรเดียวกันกับความสามารถของผู้สอบและความยากของข้อสอบได้

โมเดลหลายองค์ประกอบของราล์ชเป็นโมเดลที่เป็นเอกมิติ (Unidimension) ที่สามารถรวมค่าพารามิเตอร์ความสามารถของนักเรียนเข้ากับองค์ประกอบอื่น ๆ ได้ เช่น รวมเอาองค์ประกอบความยากของข้อคำถามและองค์ประกอบของผู้ตรวจ เป็นต้น โมเดลหลายองค์ประกอบของราล์ชมีความเหมาะสม หากผู้พัฒนาการประเมินมีจุดมุ่งหมาย เพื่อรวมผลการประเมินที่ได้มาจากโอกาสในการวัดขององค์ประกอบต่าง ๆ แยกกัน เพื่อที่จะรวมเป็นคะแนนรวม (Total score) โมเดลหลายองค์ประกอบของราล์ช มีข้อตกลงเบื้องต้นว่า ชุดของประชากรที่ได้รับการวัดและชุดของข้อคำถามที่ใช้วัดกลุ่มประชากรนั้นสามารถจัดเป็นลำดับเฉพาะ (Unique ordering) ในแต่ละเทอมตามลำดับของความสามารถและความยากของข้อคำถาม (Engelhard. 1994: 95; citing Choppin. 1987. The Rasch model for item analysis. In Alternative Approaches to the Assessment for Achievement. McArthur D.L. editor. pp. 99-127. Norwell, MA: Kluwer) ถ้าข้อมูลมีความสอดคล้องกับโมเดลและลำดับเฉพาะนี้เป็นจริงจะเป็นไปตามลักษณะความต้องการของการวัด ที่ประกอบด้วย ความสามารถในการแยกพารามิเตอร์กับค่าสถิติที่พอเพียงออกจากกันในการประมาณค่าพารามิเตอร์ ความไม่แปรเปลี่ยน (Invariance) ในการประมาณค่าความสามารถ (Ability or Competence) ความเข้มงวดของผู้ตรวจ (Rater severity) และความยากของข้อคำถาม (Item difficulty) (ซึ่งเป็นผลที่เกิดจากคุณสมบัติที่เรียกว่า Specific objectivity ของราล์ชโมเดล) ช่วงสเกลของการวัดที่เท่าเทียมกัน (Equal interval scale) โมเดลหลายองค์ประกอบของราล์ชมีส่วนคล้ายกับโมเดลการปรับเทียบ (Equating model) โดยเฉพาะกับผู้ตรวจ เทียบได้กับแบบทดสอบที่มีหลากหลายรูปแบบ เนื่องจากนักเรียนมีความสามารถแตกต่างกันและถูกตรวจด้วยผู้ตรวจที่ต่างกัน ดังนั้นจึงจำเป็นต้องมีการปรับสถิติเพื่อให้เหมาะสมกับผู้ตรวจในกลุ่มนั้น

จากโมเดลหลายองค์ประกอบของราล์ช ความน่าจะเป็นของนักเรียนคนที่ n ที่มีความสามารถ β_n ได้รับการให้คะแนน x ($x = 0, 1, 2, \dots, m$) จากรายการประเมินที่ δ_i โดยผู้ตรวจ λ_j ที่มีความยากประจำชิ้น τ_k เขียนเป็นสมการได้ดังนี้ (Engelhard. 1994: 96)

$$\pi_{ijk} = \frac{1}{1 + \sum_{k=1}^m \exp \left[k(\beta_n - \delta_i - \lambda_j) - \sum_{s=1}^k \tau_s \right]}$$

เมื่อ $x=0$ และ

$$\pi_{ijk} = \frac{\exp \left[x(\beta_n - \delta_i - \lambda_j) - \sum_{s=1}^k \tau_s \right]}{1 + \exp \left[k(\beta_n - \delta_i - \lambda_j) - \sum_{s=1}^k \tau_s \right]}$$

เมื่อ $x=1, \dots, m$

สถิติทดสอบความสอดคล้องของโมเดล (Fit Statistics)

สถิติทดสอบความสอดคล้องของโมเดล (Fit Statistics) แสดงถึงระดับความสอดคล้องของข้อมูลเชิงประจักษ์ (คะแนนจากการตรวจของผู้ตรวจ) กับค่าที่คาดหวังจากโมเดลหลายองค์ประกอบของราล์ซ (คะแนนจากการทำนายการให้คะแนนของผู้ตรวจของโมเดล)

การตรวจสอบความสอดคล้องของข้อมูลเชิงประจักษ์กับโมเดลนั้น สามารถตรวจสอบได้หลายแนวทาง แนวทางหนึ่งที่ยอมรับก็คือ การใช้ค่ามาตรฐานของส่วนที่เหลือ (Standardized Residual: Z_{nij}) เป็นดัชนีบ่งชี้คุณภาพของการให้คะแนน นอกจากนี้การตรวจสอบความสอดคล้องของข้อมูลกับโมเดลยังมีประโยชน์เกี่ยวกับการวิเคราะห์ค่าความเที่ยงตรง (Validity) และค่าความเชื่อมั่น (Reliability) ของแบบทดสอบอีกด้วย

ค่าส่วนที่เหลือของค่าการสังเกตกับค่าที่คาดหวังของผู้สอบ (n) ที่ได้รับการตรวจจากผู้ตรวจ (j) ในข้อสอบหรือคุณลักษณะ (i) หาค่าได้จากสมการ (Myford; & Wolfe. 2003: 407)

$$R_{nij} = E_{nij} - X_{nij}$$

เมื่อ E_{nij} คือ ค่าที่คาดหวังของการประเมินบนฐานของการประมาณค่าพารามิเตอร์ของราล์ซ

X_{nij} คือ ค่าที่สังเกตได้ของการประเมิน

และสูตรสำหรับคำนวณ E_{nij} (Engelhard. 1994: 96; Myford; & Wolfe. 2003: 407) เป็นดังนี้

$$E_{nij} = \sum_{k=0}^m k \pi_{ijk}$$

เมื่อ m คือ จำนวนของรายการในมาตรการประเมิน

k คือ ดัชนีจำนวนแสดงถึงค่าของแต่ละรายการของมาตรประเมิน

ค่ามาตรฐานส่วนที่เหลือคำนวณได้ดังนี้ (Engelhard. 1994: 96; Myford; & Wolfe. 2003: 407)

$$Z_{nij} = \frac{R_{nij}}{\sqrt{W_{nij}}}$$

เมื่อ

$$W_{nij} = \sum_{k=0}^m (k - E_{nij})^2 \pi_{nijk}$$

นั่นคือ

$$Z_{nij} = \frac{X_{nij} - E_{nij}}{\left[\sum_{k=0}^m (k - E_{nij})^2 \pi_{nijk} \right]^{1/2}}$$

เมื่อ Z_{nij} คือ ค่ามาตรฐานของส่วนที่เหลือ

X_{nij} คือ คะแนนจากผู้ตรวจคนที่ j ตรวจนักเรียนคนที่ n ในข้อ i

E_{nij} คือ ค่าความคาดหวังของการประเมินจากโมเดลการวัด

ค่ามาตรฐานของส่วนที่เหลือ (Standardized residuals) จะให้รายละเอียดเกี่ยวกับการตรวจสอบความสอดคล้องระหว่างคะแนนสังเกตกับรูปแบบการประเมินที่คาดหวัง (Expecting rating patterns) ของโมเดล ค่ามาตรฐานของส่วนที่เหลือสามารถสรุปความแตกต่างขององค์ประกอบ (Facets) และความแตกต่างของสมาชิก (Elements) ในองค์ประกอบ ค่ามาตรฐานของส่วนที่เหลือจึงเป็นดัชนีที่แสดงถึงความสอดคล้องของข้อมูลเชิงประจักษ์กับโมเดลทางทฤษฎี (Engelhard. 1994: 97)

เมื่อนำค่ามาตรฐานส่วนที่เหลือยกกำลังสองแล้วหาค่าเฉลี่ยทั้งผู้สอบและข้อสอบ เพื่อสรุปเป็นค่าสถิติของผู้ตรวจแต่ละคน แสดงถึง ค่าความไม่สอดคล้อง (Misfit) ทั้งหมดในการตรวจของผู้ตรวจ (Linacre. 1994) ซึ่งค่าเฉลี่ยนี้จะให้ค่าดัชนีใน 2 ลักษณะ ดังนี้

1. สถิติ OUTFIT เป็นดัชนีของค่าเฉลี่ยยกกำลังสองของส่วนที่เหลือแบบไม่ถ่วงน้ำหนัก (Unweighted mean square residual) ที่แสดงถึงความแตกต่างระหว่างคะแนนสังเกตกับค่าความ

คาดหวัง สถิติ OUTFIT, u_j สำหรับองค์ประกอบผู้ตรวจแสดงถึงค่าเฉลี่ยยกกำลังสองของค่ามาตรฐาน ส่วนที่เหลือจากผู้สอบและข้อสอบทั้งหมดของผู้ตรวจ สามารถคำนวณได้จากสมการ (Myford; & Wolfe. 2003: 408)

$$u_j = \frac{\sum_{n=1}^N \sum_{i=1}^I z_{nij}^2}{NI}$$

เมื่อ N คือ จำนวนผู้สอบ

I คือ จำนวนข้อสอบ

เนื่องจากสถิติ OUTFIT เป็นสถิติที่มีค่าค่อนข้างไว (Sensitivity) ต่อการระบุข้อความที่ไม่สอดคล้องกับโมเดล (Misfit) (Engelhard. 1994: 97; Myford; & Wolfe. 2000: online) ดังนั้น ค่าสถิตินี้จึงมีจุดอ่อนในบางลักษณะของการตอบของกลุ่มผู้ตอบ เช่น กรณีที่ผู้ตอบบางคนมีรูปแบบการตอบที่แตกต่างไปจากลักษณะการตอบคำถามของกลุ่มผู้ตอบโดยส่วนรวม เช่น คนที่มีคุณลักษณะ (Trait) ในระดับสูงตอบข้อคำถามในบางรายการในระดับต่ำ ๆ หรือในทางตรงกันข้าม ผู้ตอบที่มีคุณลักษณะระดับต่ำกลับตอบข้อความบางรายการในระดับสูง ๆ ซึ่งลักษณะการตอบดังกล่าวนี้ ถือว่าไม่สอดคล้องกับค่าความน่าจะเป็นที่เขาควรจะตอบ จึงทำให้ค่าสถิติปฏิเสธข้อความในบางรายการไปทั้งที่ความจริงแล้วข้อความดังกล่าวยังใช้ได้ เพียงแต่ผู้ตอบบางคนตอบข้อความรายการนั้นด้วยความเสแสร้งเท่านั้น (Wright; & Master. 1982: 99) ดังนั้น จึงมีสถิติอีกค่าหนึ่งซึ่งมีความละเอียดขึ้นโดยมี w_{nij} เป็นตัวถ่วงน้ำหนัก

2. สถิติ INFIT เป็นค่าดัชนีของค่าเฉลี่ยยกกำลังสองแบบถ่วงน้ำหนัก (Weighted mean squares) กับสัดส่วนความแปรปรวนส่วนที่เหลือยกกำลังสองแบบถ่วงน้ำหนัก (Weighted proportion to the variance of the square residuals) ที่สามารถคำนวณแต่ละสมาชิกขององค์ประกอบได้ สถิติ INFIT มีความไว (Sensitivity) ต่อผลการประเมินที่ไม่คาดหวังต่ำ (Outlying unexpected) สถิติ INFIT, v_j สำหรับองค์ประกอบผู้ตรวจ สามารถคำนวณได้ดังสมการ (Engelhard. 1994: 97)

$$v_j = \frac{\sum_{n=1}^N \sum_{i=1}^I w_{nij} z_{nij}^2}{\sum_{n=1}^N \sum_{i=1}^I w_{nij}}$$

$$\text{เมื่อ } W_{nij} = \sum_{k=0}^m (k - E_{nij})^2 \pi_{nijk}$$

สถิติทดสอบความสอดคล้องของโมเดลสำหรับองค์ประกอบอื่น ๆ และสมาชิกภายในองค์ประกอบนั้น นิยมคำนวณแล้วสรุปบนค่าที่เหมาะสม เมื่อโมเดลมีความสอดคล้องกับข้อมูลเชิงประจักษ์ ค่าคาดหวังของดัชนีค่าเฉลี่ยยกกำลังสอง INFIT และ OUTFIT มีค่าเท่ากับ 1 หากค่าทั้งสองมีค่าเข้าใกล้ 1 แสดงว่าคะแนนสังเกตกับค่าที่คาดหวังนั้นมีค่าใกล้กัน หากมีค่าน้อยกว่า 1 แสดงว่าค่าสังเกตมีค่าใกล้กับค่าที่คาดหวังมากกว่าที่โมเดลทำนายไว้ (Overfit) และหากมีค่ามากกว่า 1 แสดงว่าค่าที่สังเกตอยู่ห่างจากค่าที่คาดหวังมากกว่าที่โมเดลทำนายไว้ (Misfit) (Myford; & Wolfe. 2003: 408) ทั้งนี้ขอบเขตของการยอมรับได้ที่เหมาะสมของโมเดลที่สอดคล้องกับข้อมูลนั้น ลันซ์ ไรท์ และลินาเคอร์ (Lunz; Wright; & Linacre. 1990) แนะนำว่า ขอบเขตของการยอมรับได้ควรมีค่าอยู่ระหว่าง 0.6 – 1.5 ส่วนไรท์และมาสเตอร์ (Wright; & Master. 1982: 102) ได้เสนอขอบเขตที่ยอมรับได้ที่เหมาะสมของโมเดลที่สอดคล้องกับข้อมูลเมื่อคำนวณสถิติ INFIT และ OUTFIT ว่าควรมีค่าอยู่ระหว่าง -2 ถึง +2 ทั้งนี้ มายฟอร์ดและวูล์ฟ (Myford; & Wolfe. 2003: 409) ได้ให้ข้อเสนอแนะว่าขอบเขตที่เหมาะสมของสถิติ INFIT และ OUTFIT นั้นขึ้นอยู่กับความต้องการใช้สารสนเทศที่ได้จากการวิเคราะห์ MFRM เช่น หากนำไปใช้ตัดสินใจในกรณีที่มีผู้มีส่วนได้ส่วนเสียจำนวนมาก ขอบเขตอาจจะแคบลง โดยอาจอยู่ระหว่าง 0.8 – 1.2 แต่หากมีผู้มีส่วนได้ส่วนเสียไม่มาก อาจผ่อนปรนให้อยู่ระหว่าง 0.5 – 2.0

ความเชื่อมั่นแยกส่วน (Reliability of Separation index)

ความเชื่อมั่นแยกส่วน (Reliability of Separation index) เป็นดัชนีที่มีประโยชน์ในการตรวจสอบตัวแปรแฝง ดัชนีนี้ให้ค่าการวัดของสมาชิกต่าง ๆ ในองค์ประกอบว่ามีการกระจายหรือแยก (Separated) อยู่ตลอดช่วงความสามารถ (Continuum) หรือไม่ (Smith; & Kulkowich. 2004: 629) แสดงถึงปริมาณในการประมาณค่าการวัดของสมาชิกต่าง ๆ ในองค์ประกอบที่กำหนดไว้ ซึ่งมีความสัมพันธ์กับความแม่นยำของการประมาณค่าการวัด (Sudweeks; Reeve ; & Bradshaw. 2005) เป็นดัชนีที่ให้ข้อมูลเกี่ยวกับสมาชิกต่าง ๆ ในองค์ประกอบ มีความคล้ายคลึงกับดัชนีความเชื่อมั่นแบบมาตรฐานเดิม (Traditional indices of reliability) ที่เป็นการประมาณค่าความสอดคล้องภายใน (Internal consistency) (Smith. 2001: 284) เช่น สัมประสิทธิ์ของครอนบาคแอลฟา (Cronbach's coefficient alpha) และ KR-20 ในลักษณะที่เป็นการประมาณค่าจากอัตราส่วนของความแปรปรวนของคะแนนจริงกับความแปรปรวนของคะแนนการสังเกต (Wright; & Master. 1982; Engelhard. 1994: 97) ดังสมการ

$$R = \frac{SD^2 - MSE}{SD^2}$$

เมื่อ SD^2 คือ ความแปรปรวนของคะแนนสังเกตขององค์ประกอบที่อยู่บนสเกลของตัวแปรแฝง และ MSE คือ ค่าประมาณความคลาดเคลื่อนของค่าเฉลี่ยยกกำลังสอง เป็นค่าเฉลี่ยของการวัดความแปรปรวนของความคลาดเคลื่อนของแต่ละสมาชิกภายในองค์ประกอบ

ยกตัวอย่างเช่น ในกรณีการประมาณค่าความเชื่อมั่นแยกส่วนของบุคคลหรือผู้สอบ (Person separation reliability, Person reliability) ในอันดับแรกจะคำนวณความแปรปรวนของความคลาดเคลื่อนของค่าเฉลี่ยของบุคคล (Mean Square Measurement Error: MSE_p) (Smith, 2001: 285) ดังสูตร

$$MSE_p = \frac{\sum_{n=1}^N S_n^2}{N}$$

เมื่อ S_n คือ SE ของผู้สอบแต่ละคน

จากความแปรปรวนของค่าสังเกตระหว่างบุคคล SD^2 จะได้

$$SA_p^2 = SD_p^2 - MSE_p$$

เมื่อ SA_p^2 คือ ความแปรปรวนที่แท้จริงของผู้สอบ เป็นความแปรปรวนของผู้สอบที่ปรับแก้การวัดความคลาดเคลื่อน ทำให้สามารถคำนวณหาความเชื่อมั่นแยกส่วนของบุคคลได้เป็นอัตราส่วนของความแปรปรวนที่ปรับแก้แล้ว(ความแปรปรวนที่แท้จริง) กับความแปรปรวนของค่าสังเกต จึงเป็นสัดส่วนของความแปรปรวนที่ไม่ได้เกิดจากความคลาดเคลื่อน ดังสูตร

$$\begin{aligned}
 R_p &= \frac{SA_p^2}{SD_p^2} \\
 &= \frac{SD_p^2 - MSE_p}{SD_p^2} \\
 &= 1 - \frac{MSE_p}{SD_p^2}
 \end{aligned}$$

ในกรณีของการประมาณค่าความเชื่อมั่นแยกส่วนขององค์ประกอบอื่น สามารถประมาณค่าได้ในทำนองเดียวกัน

ในการแปลความของความเชื่อมั่นแยกส่วนขององค์ประกอบต่าง ๆ ที่อยู่ในโมเดลนั้นมีความแตกต่างกันเพียงเล็กน้อย (Myford ; & Wolfe, 2003: 411; Sudweek; Reeve; & Bradshaw. 2005) ดังนี้

1. ความเชื่อมั่นแยกส่วนของผู้สอบ (Person separation reliability; Person reliability) คล้ายกับสัมประสิทธิ์แอลฟา และ KR-20 (Engelhard. 2002; Myford; & Wolfe. 2003) ซึ่งเป็นอัตราส่วนของความแปรปรวนของคะแนนจริงกับความแปรปรวนของคะแนนสังเกตในการประมาณค่าความสามารถของผู้สอบ เช่น ความเชื่อมั่นแยกส่วนของผู้สอบเป็น 0.96 หมายถึง 96 % ของความแปรปรวนในการวัดความสามารถของนักเรียนเป็นความแปรปรวนของคะแนนจริง มีเพียง 4 % ที่เกิดจากความแปรปรวนของความคลาดเคลื่อนของการวัด (Paulukonis; Myford; & Heller. n.d.: online) ค่าความเชื่อมั่นแยกส่วนของผู้สอบที่มีค่าสูง แสดงว่ามีความแตกต่างระหว่างความสามารถของผู้สอบ แสดงถึงความสามารถในการจำแนกความสามารถของผู้สอบแต่ละคนออกจากกัน (Sudweek; Reeve; & Bradshaw. 2005) ดังนั้นในองค์ประกอบของผู้สอบ จึงต้องการค่าความเชื่อมั่นแยกส่วนที่มีค่าสูง เพื่อเป็นดัชนีที่แสดงถึงความเชื่อถือได้ของคะแนนความสามารถของผู้สอบที่ได้จากข้อสอบและผู้ตรวจ (Upshur; & Turner. 1999: 97)

2. ความเชื่อมั่นแยกส่วนของผู้ตรวจ (Rater separation reliability) แสดงถึง ความแปรปรวนของการให้คะแนนที่ไม่ต้องการให้เกิดขึ้นระหว่างผู้ตรวจในระดับของความเข้มงวด/ใจดี ซึ่งแมคนามารา (Sudweek; Reeve; & Bradshaw. 2005; citing McNamara. 1996. Measuring Second Language Performance) กล่าวว่าดัชนีนี้ไม่ได้บ่งชี้ถึงความสอดคล้องกันของผู้ตรวจ แต่บ่งชี้ว่าผู้ตรวจมีความเข้มงวดในการตรวจให้คะแนนแตกต่างกันหรือไม่ นั่นคือ เพื่อแสดงว่าความเข้มงวดของผู้ตรวจมีความแตกต่างกันอย่างไร (ซึ่งมีความหมายตรงข้ามกับความเชื่อมั่นภายในของ

ผู้ตรวจ (Interrater reliability) ที่แสดงถึงความคล้ายคลึงของความเข้มงวดของผู้ตรวจ) หากมีค่าสูง แสดงว่ามีความแตกต่างในการตรวจให้คะแนนของผู้ตรวจ กล่าวอีกนัยหนึ่งคือ ความเชื่อมั่นแยกส่วนของผู้ตรวจ บ่งชี้ถึงระดับความเชื่อมั่นของความแตกต่าง (Reliably different) มากกว่าระดับความเชื่อมั่นของความคล้ายคลึงกัน (Reliably similar)(Sudweek; Reeve; & Bradshaw. 2005; citing Linacre. 2004. Personal communication) ดังนั้น ถ้ามีเป้าหมายเพื่อดูว่าผู้ตรวจให้คะแนนแตกต่างกันหรือไม่ในการวัดคุณลักษณะเดียวกัน จึงต้องการความเชื่อมั่นแยกส่วนของผู้ตรวจที่มีค่าต่ำ

3. ความเชื่อมั่นแยกส่วนของคุณลักษณะ(Trait) หรือ ข้อสอบ (Item separation reliability) เป็นดัชนีที่แสดงถึงความสามารถในการจำแนกความยากของข้อสอบหรือคุณลักษณะว่ามีความแตกต่างกันหรือไม่ อย่างไร ซึ่งความเชื่อมั่นแยกส่วนของข้อสอบควรมีค่าสูงหากต้องการให้คุณลักษณะหรือข้อสอบนั้นกระจายอยู่ตลอดช่วงความสามารถ (Continuum)

สถิติแยกส่วน (Separation statistics)

สถิติแยกส่วนนี้แบ่งออกเป็น 2 ลักษณะคือ อัตราส่วนแยก (Separation ratio) และดัชนีแยกส่วน (Separation index) เนื่องจากการประมาณค่าความเชื่อมั่นได้รับผลจากค่าอิทธิพลเพดาน (Ceiling effects) ที่จำกัดอยู่ในช่วง 0 ถึง 1 จึงมีการใช้อัตราส่วนแยก (Separation ratio) หรือ G ขึ้นมาใช้แทน R ซึ่งมีช่วงตั้งแต่ 0 ถึง Infinity (Smith. 2001: 285) ทั้งนี้หากเป็นอัตราส่วนของผู้สอบ (Ratee separation ratio) นั้นจะเป็นการวัดความผันแปร (Spread) ของคะแนนผู้สอบที่สัมพันธ์กับความแม่นยำ (Precision) (Fisher. 1992: online) แสดงถึงสัดส่วนของส่วนเบี่ยงเบนมาตรฐานของคะแนนจริงของคะแนนผู้สอบ (ส่วนเบี่ยงเบนมาตรฐานที่ปรับแก้ค่าความคลาดเคลื่อน) กับค่าเฉลี่ยของความคลาดเคลื่อนมาตรฐานของผู้สอบ สามารถคำนวณได้ดังสมการ (Smith. 2001: 285; Myford; & Wolfe. 2003: 410)

$$G_p = \frac{SA_p}{SE_p}$$

$$= \left[\left(\frac{R_p}{1 - R_p} \right)^{1/2} \right]$$

เมื่อ SE_p คือ รากที่สองของค่าเฉลี่ยกำลังสองของความคลาดเคลื่อน (Root Mean Square Measurement Error: RMSE) ซึ่งมีค่าเท่ากับ $(MSE_p)^{1/2}$

อัตราส่วนแยกของบุคคล (Person separation ratio: G_p) อยู่บนมาตราอัตราส่วน (Ratio scale) เป็นการเปรียบเทียบความผันแปร (Spread) ที่แท้จริงของความสามารถของบุคคลกับค่าความคลาดเคลื่อน และชี้ถึงความผันแปรของการวัดบุคคลในหน่วยของความคลาดเคลื่อนมาตรฐาน (Fisher. 1992: online) ซึ่ง G_p ที่มีค่าสูงกว่าแสดงว่ามีความผันแปรออกไปจากบุคคลอื่น ๆ มากบนตัวแปรที่ต้องการวัด ถ้าเป้าหมายของการวัดเป็นแบบอิงกลุ่มนั้นต้องการค่า G_p ที่มีค่าสูง หากเปรียบเทียบการวิเคราะห์ที่หลากหลายในข้อมูลเดียวกันนั้นค่า R_p นั้นเหมาะสมกว่าค่า R_p ซึ่งค่า R_p นั้นมีค่าสูงสุดเป็น 1 ดังนั้นหากมีการวิเคราะห์ที่หลากหลายในข้อมูลเดียวกันนั้นค่า R_p จะไม่เปลี่ยนแปลง ในขณะที่ค่า G_p นั้นเพิ่มสูงขึ้น (Smith. 2001: 285)

เมื่อได้อัตราส่วนแยกของผู้สอบแล้ว สามารถคำนวณหาค่าดัชนีแยกของผู้สอบ (Ratee separation index) หรือ H ซึ่งเป็นจำนวนของระดับชั้น (Strata) ของความสามารถของผู้สอบในกลุ่มตัวอย่างของผู้สอบที่แตกต่างกัน (Wright; & Master. 1982) เช่น หากดัชนีแยกของผู้สอบมีค่าเป็น 3 แสดงว่า มีการแบ่งกลุ่มผู้สอบตามสถิติ ได้เป็น 3 กลุ่ม ซึ่งฟิชเชอร์ (Fisher. 1992: online) กล่าวว่า ช่วงของฟังก์ชันของความสามารถของผู้สอบโดยทั่วไปมีค่าประมาณ 4 เท่าของส่วนเบี่ยงเบนมาตรฐานที่แท้จริง ถ้ามีการกำหนดระดับการแบ่งทางสถิติของความสามารถของผู้สอบเป็นชั้น (Strata) ซึ่งแบ่งเป็นอย่างน้อย 3 เท่าของส่วนเบี่ยงเบนมาตรฐานแล้ว ดัชนีแยกของผู้สอบ (Ratee separation index: H) คำนวณได้ดังสมการ (Smith. 2001: 286; Myford; & Wolfe. 2003: 410)

$$H = \frac{4G + 1}{3}$$

ในกรณีคำนวณหาค่าอัตราส่วนแยกและดัชนีแยกขององค์ประกอบอื่นๆ เช่น ข้อสอบ หรือ ผู้ตรวจ สามารถใช้สูตรในทำนองเดียวกัน

สถิติทดสอบไค-สแควร์ (Fixed effect chi-square test)

สถิติไค-สแควร์ (Chi-Square Statistic) ที่ได้จากโปรแกรม FACETS เป็นสถิติที่ใช้ทดสอบความแตกต่างอย่างมีนัยสำคัญทางสถิติระหว่างสมาชิกต่าง ๆ ในองค์ประกอบ (Engelhard. 1996: 62) โดยมีสมมติฐานสำหรับการทดสอบองค์ประกอบของผู้ตรวจว่าผู้ตรวจทุกคนมีระดับความเข้มงวดอยู่ในระดับเดียวกันภายหลังจากอธิบายความคลาดเคลื่อนแล้ว ในการทดสอบองค์ประกอบผู้สอบมีสมมติฐานว่า ผู้สอบทุกคนมีระดับความสามารถเดียวกันหลังจากอธิบายความคลาดเคลื่อนแล้ว และในองค์ประกอบของข้อสอบมีสมมติฐานว่า ข้อสอบทุกข้อมีระดับความยากที่เท่ากันหลังจากอธิบายความคลาดเคลื่อนแล้ว หากค่าไค-สแควร์มีนัยสำคัญแสดงว่ามีสมาชิกอย่างน้อย 2 ตัวในองค์ประกอบ

ที่แตกต่างกันอย่างมีนัยสำคัญทางสถิติ แต่ค่าสถิตินี้เป็นเพียงค่าประมาณของการแจกแจงแบบไค-สแควร์เท่านั้น โดยมีสูตรทางสถิติสำหรับผลแบบกำหนดของข้อสอบเป็นดังนี้ (Myford; & Wolfe. 2003: 409; citing Linacre. 2001b. A user's guide to facets: Rasch measurement computer program [Computer program manual])

$$\chi^2 = \sum (w_i \times D_i^2) - \frac{(\sum w_i \times D_i)^2}{\sum w_i}, \text{ d.f.} = L - 1$$

เมื่อ D คือ ความยากของข้อสอบ

L คือ จำนวนของข้อสอบ และ

$$w_i = \frac{1}{SE_i^2} \text{ สำหรับ } i = 1, \dots, L$$

หากต้องการทดสอบผลแบบกำหนดของผู้ตรวจทั้งหมดว่ามีระดับความเข้มงวดเดียวกันหรือไม่ในการประเมินผู้สอบ จากสมการใช้ C_j แทน D_i และ d.f. = J - 1 แทน L - 1 ในทำนองเดียวกันหากต้องการทดสอบสมมติฐานผลแบบกำหนดของผู้สอบทุกคนว่าอยู่ในระดับมาตรฐานเดียวกันของการปฏิบัติหรือไม่ จากสมการใช้ B_n แทน D_i และ d.f. = N - 1 แทน L - 1 ทั้งนี้ หากมีนัยสำคัญทางสถิติของผู้สอบแสดงว่าผู้สอบนั้นมีความสามารถในการวัดแตกต่างกัน หากมีนัยสำคัญทางสถิติของผู้ตรวจแสดงว่าระดับความเข้มงวดของผู้ตรวจนั้นแตกต่างกัน ดังนั้นจึงไม่ต้องการให้มีนัยสำคัญทางสถิติขององค์ประกอบผู้ตรวจ ทั้งนี้เพื่อที่จะบ่งชี้ว่าผู้ตรวจมีความคล้ายคลึงกันในการประเมินหรือมีระดับความเข้มงวดที่เท่าเทียมกันนั่นเอง และในองค์ประกอบของข้อสอบ หากมีนัยสำคัญทางสถิติของข้อสอบ แสดงว่า ความยากของข้อสอบนั้นมีการกระจายตลอดช่วงความสามารถ (Continuum) (Smith; & Kulikowich. 2004: 629-636)

นอกจากนี้การวิเคราะห์โดยใช้ MFRM ยังสามารถวิเคราะห์ความลำเอียง (Bias analysis) (Myford; & Wolfe. 2003: 411) โดยจะให้ค่าดัชนีความลำเอียง (Bias index) หรือ ปฏิสัมพันธ์ (Interaction) ระหว่างองค์ประกอบต่าง ๆ ทำให้สามารถบ่งชี้ได้ว่าผู้ตรวจคนใดบ้างมีแนวโน้มที่จะประเมินแตกต่างไปจากผู้ตรวจคนอื่น บุคคลใดที่ทำงานชิ้นหนึ่ง (ข้อสอบข้อหนึ่ง) แตกต่างไปจากการทำงานชิ้นอื่น (ข้อสอบข้ออื่น) โดยให้ I_{jg} แทน ปฏิสัมพันธ์ลำเอียง (Bias interaction) แสดงถึงระดับ (Degree) ที่ผู้ตรวจ j ตรวจให้คะแนนผู้สอบกลุ่ม g แตกต่างไปจากค่าที่คาดหวังของโมเดล ซึ่งมีสมมติฐานศูนย์ว่า $I_{jg} = 0$ ทดสอบโดยใช้ z-statistic (Myford; & Wolfe. 2003: 411) ดังนี้

$$Z_{\text{bias}} = \frac{I_{\text{Jg}}}{SE_{I_{\text{Jg}}}}$$

หาก z_{bias} มีนัยสำคัญทางสถิติ แสดงว่า ผู้ตรวจมีความลำเอียงให้กับผู้สอบกลุ่มหนึ่ง เช่น ผู้ตรวจที่มีความเข้มงวดต่างกันได้ให้คะแนนสูงเกินความคาดหวังกับกลุ่มผู้สอบบางกลุ่ม

กล่าวโดยสรุปได้ว่าโมเดลหลายองค์ประกอบของราส์ซ (MFRM) นั้นปรับขยายมาจากโมเดลพื้นฐานของราส์ซ (Rasch Model) แบบ 1 พารามิเตอร์ จากเดิมที่สามารถประมาณค่าพารามิเตอร์ของผู้สอบที่แสดงถึงความยากของข้อสอบและพารามิเตอร์ของผู้สอบที่แสดงถึงความสามารถของผู้สอบ โดยสมารถที่จะประมาณค่าพารามิเตอร์ของผู้ตรวจ ซึ่งแสดงถึงความเข้มงวดของผู้ตรวจ (Rater severity) หรือเพิ่มองค์ประกอบอื่น ๆ ที่สนใจศึกษาเข้าไปในโมเดลได้ จึงทำให้สามารถตรวจสอบอิทธิพลของผู้ตรวจที่มีต่อการวัดประเมินได้ นอกจากนี้ MFRM จะปรับแก้คะแนนที่เกิดจากความแตกต่างของความเข้มงวดของผู้ตรวจและความยากของข้อสอบ ทำให้คะแนนที่ได้มีความน่าเชื่อถือเที่ยงตรงและยุติธรรมมากขึ้น ทั้งนี้ MFRM มีจุดเน้นที่ต้องการศึกษาถึงแต่ละสมาชิกในองค์ประกอบที่เป็นแหล่งของความคลาดเคลื่อน เช่น พฤติกรรมการตรวจให้คะแนนของผู้ตรวจแต่ละคนว่ามีความเข้มงวดมากน้อยเพียงใด เพื่อการพัฒนาคุณภาพของแต่ละสมาชิกในระบบการวัดประเมิน (Iramaneerat; et al. 2007: online) ทำให้การตรวจให้คะแนนมีความเป็นปรนัยมากขึ้น (Linacre. 1994) นั่นคือ การวิเคราะห์โดย MFRM ทำให้เห็นภาพส่วนต่าง ๆ อันเป็นสารสนเทศที่มีประโยชน์ ไม่เฉพาะภาพรวมบนเส้นความสามารถเดียวกันเท่านั้น แต่ยังมีรายละเอียดภาพย่อยของแต่ละองค์ประกอบ อีกทั้ง สามารถประมาณค่าพารามิเตอร์ของผู้ตรวจ ทำให้ผู้วิจัยได้เห็นรายละเอียดและอธิบายผลอย่างทั่วถึงของผลการวิเคราะห์คะแนนจากแบบทดสอบที่เกี่ยวข้องกับผู้ตรวจ

มีการนำโมเดลนี้ไปใช้วิเคราะห์การวัดประเมินอย่างหลากหลายทั้งการวัดประเมินทางด้านภาษา ทางด้านศิลปะ หรือทางด้านการแพทย์ เป็นต้น (Engelhard. 1992; 1994; Myford; & Wolfe: 2000: online; Nakamura. 2002: online; Bonk; & Ockey. 2003: online; Turner. 2003: online; ชาญวิทย์ จรัสสุทธิศิร. 2550) ซึ่งการวัดประเมินเหล่านี้ต่างมีแหล่งความคลาดเคลื่อนของการให้คะแนนหลายแหล่ง โดยเฉพาะจากผู้ตรวจให้คะแนน ผลการวิจัยเหล่านี้ต่างพบประเด็นที่สอดคล้องกันว่า ผู้ตรวจเป็นแหล่งความคลาดเคลื่อนที่สำคัญและผู้ตรวจมีความเข้มงวดที่แตกต่างกัน ส่งผลกับคะแนนที่ได้ ทั้งนี้เมื่อใช้โมเดลหลายองค์ประกอบของราส์ซในการวิเคราะห์คะแนน จะสามารถประมาณค่าพารามิเตอร์ของแต่ละองค์ประกอบ เช่น ผู้สอบ ข้อสอบและผู้ตรวจ ได้อย่างเป็นอิสระจากกัน ตลอดจนสามารถประมาณค่าความเชื่อมั่นแยกส่วนของแต่ละองค์ประกอบได้ ทำให้สามารถอธิบายได้อย่างชัดเจนว่าแต่ละองค์ประกอบมีอิทธิพลต่อคะแนนอย่างไร นอกจากนี้โมเดลนี้ยัง

สามารถวิเคราะห์ผลการตรวจให้คะแนนโดยที่ผู้ตรวจไม่จำเป็นต้องตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Bachman; Lynch; & Mason. 1993: online; Lynch; & McNamara. 1998: online; Iramaneerat; et al. 2007: online) จึงสามารถแก้ปัญหาการตรวจให้คะแนนที่ต้องใช้เวลา แรงงานและค่าใช้จ่ายค่อนข้างสูงได้

จากผลการวิจัยดังกล่าวผู้วิจัยจึงสนใจศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิด วิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจการให้คะแนนต่างกัน ด้วยโมเดลหลายองค์ประกอบของราล์ช ซึ่งผู้วิจัยได้ศึกษาขนาดขององค์ประกอบความแปรปรวนของแต่ละแหล่ง ได้แก่ ผู้สอบ ข้อสอบและผู้ตรวจ ประมาณค่าพารามิเตอร์ และพิจารณาคุณลักษณะคะแนนจากค่าความเชื่อมั่นแยกส่วน ค่าความเที่ยงตรงตามสภาพ โดยประมาณค่าและเปรียบเทียบความเชื่อมั่นแยกส่วนในแต่ละองค์ประกอบเมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน ประมาณค่าและเปรียบเทียบความเที่ยงตรงตามสภาพของคะแนนเมื่อจำนวนผู้ตรวจและรูปแบบการตรวจต่างกัน โดยการนำคะแนนความสามารถจากการวิเคราะห์ด้วยโมเดลหลายองค์ประกอบของราล์ชไปหาค่าสัมประสิทธิ์สหสัมพันธ์กับคะแนนที่ใช้เป็นเกณฑ์ ซึ่งในงานวิจัยนี้ใช้ คะแนนที่ได้จากการทดสอบทางการศึกษาระดับชาตินัยพื้นฐาน (Ordinary National Education Test: O-NET) ทั้งนี้เพื่อตรวจสอบคุณลักษณะของคะแนน อันได้แก่ ค่าความเชื่อมั่นแยกส่วนและความเที่ยงตรงตามสภาพ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน เพื่อให้ได้ผลการตรวจให้คะแนนที่มีความเชื่อมั่นและความเที่ยงตรงมากที่สุด อันจะเป็นประโยชน์ในการนำไปใช้ต่อไป

การทดสอบทางการศึกษาระดับชาตินัยพื้นฐาน (Ordinary National Education Test: O-NET)

สถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน) หรือ สทศ. (National Institute of Educational Testing Service: NIETS) เป็นหน่วยงานของรัฐจัดตั้งขึ้นตามพระราชบัญญัติการจัดตั้งสถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน) พ.ศ. 2548 เพื่อเป็นองค์กรกลางในการศึกษาวิจัย พัฒนา และให้บริการประเมินผลทางการศึกษาและทดสอบทางการศึกษา รวมทั้งเป็นศูนย์กลางความร่วมมือด้านการทดสอบทางการศึกษาในระดับชาติและนานาชาติ (สทศ. 2554: online)

ในหลักสูตรการศึกษาขั้นพื้นฐาน พ.ศ. 2544 ได้กำหนดให้มีการประเมินคุณภาพระดับชาติ สถานศึกษาที่ต้องจัดให้ผู้เรียนทุกคนที่เรียนในชั้นประถมศึกษาปีที่ 6 ชั้นมัธยมศึกษาปีที่ 3 และชั้นมัธยมศึกษาปีที่ 6 เข้ารับการประเมินคุณภาพระดับชาติในกลุ่มสาระการเรียนรู้ที่สำคัญ เพื่อประโยชน์ในการประกันคุณภาพการศึกษาและเป็นดัชนีบ่งชี้เพื่อใช้ในการปรับปรุงคุณภาพการเรียนการสอนของโรงเรียน สทศ. จึงเป็นสถาบันกลางที่จัดสอบวัดผลระดับชาติให้กับนักเรียนทุกคน และได้จัดการ

ทดสอบทางการศึกษาระดับชาตินี้ขั้นพื้นฐาน (Ordinary National Education Test: O-NET) เพื่อการประเมินผลทางการศึกษาและเพื่อให้การวัดผลสัมฤทธิ์ทางการศึกษาเป็นมาตรฐานเดียวกันทั่วประเทศ จึงต้องมีการจัดสอบด้วยข้อสอบเดียวกัน โดย สทศ. ทำหน้าที่ออกข้อสอบ จัดสอบและประเมินผลออกมาแล้วนำมาเทียบเคียงกันเพื่อให้เป็นมาตรฐานเดียวกันทั่วประเทศ

การทดสอบทางการศึกษาระดับชาตินี้ขั้นพื้นฐาน (O-NET) เป็นการทดสอบเพื่อวัดความรู้และความคิดของนักเรียนในระดับชั้นประถมศึกษาปีที่ 6 มัธยมศึกษาปีที่ 3 และมัธยมศึกษาปีที่ 6 ตามมาตรฐานการเรียนรู้ในหลักสูตรการศึกษาขั้นพื้นฐาน พ.ศ. 2544 ครอบคลุมเนื้อหาวิชา 8 กลุ่มสาระการเรียนรู้ ได้แก่ 1) ภาษาไทย 2) สังคมศึกษา ศาสนาและวัฒนธรรม 3) ภาษาอังกฤษ 4) คณิตศาสตร์ 5) วิทยาศาสตร์ 6) สุขศึกษาและพลศึกษา 7) การงานอาชีพและเทคโนโลยี และ 8) ศิลปะ

วัตถุประสงค์ของการสอบ O-NET

สัมพันธ์ พันธุ์พฤกษ์. (2553: online) ได้กล่าวถึงวัตถุประสงค์และหน่วยงานที่สอบ O-NET ไว้ดังนี้

1. เพื่อทดสอบความรู้และความคิดของนักเรียน ตามหลักสูตรการศึกษาขั้นพื้นฐาน พ.ศ. 2544
2. เพื่อนำผลการสอบไปใช้ในการปรับปรุงคุณภาพการเรียนการสอนในโรงเรียน ประเมินผล การเรียนรู้ของนักเรียนระดับชาติ และเป็นส่วนหนึ่งในการคัดเลือกเข้าศึกษาต่อในชั้นมัธยมศึกษาปีที่ 1 มัธยมศึกษาปีที่ 4 และระดับมหาวิทยาลัย

ทั้งนี้หน่วยงานต้นสังกัดที่สอบ O-NET ทั้งหมด 4 กระทรวง/เทียบเท่า รวม 9 หน่วยงาน ดังนี้

1. กระทรวงศึกษาธิการ ได้แก่ สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน สำนักงานคณะกรรมการส่งเสริมการศึกษาเอกชน (ยกเว้นจังหวัดสงขลา สตูล ยะลา ปัตตานี และนราธิวาส) สำนักงานคณะกรรมการการอุดมศึกษา (โรงเรียนสาธิตแห่งมหาวิทยาลัย)
2. กระทรวงมหาดไทย ได้แก่ กรมส่งเสริมการปกครองท้องถิ่น (โรงเรียนเทศบาล โรงเรียน อบต. โรงเรียน อบจ.) สำนักการศึกษากรุงเทพมหานคร สำนักการศึกษาเมืองพัทยา
3. กระทรวงการท่องเที่ยวและกีฬา ได้แก่ สถาบันการพลศึกษา (โรงเรียนกีฬา)
4. สำนักนายกรัฐมนตรี ได้แก่ กองบัญชาการตำรวจตระเวนชายแดน สังกัดสำนักงานตำรวจแห่งชาติ (โรงเรียนตำรวจตระเวนชายแดน) สำนักงานพระพุทธศาสนาแห่งชาติ (โรงเรียนพระปริยัติธรรม)

ระบบการพัฒนาข้อสอบของ สทศ.

สทศ. มีระบบการพัฒนาข้อสอบ สรุปได้ดังต่อไปนี้ (สทศ. 2553: online; สทศ. n.d: online)

1. คุณสมบัติของผู้ออกและกลั่นกรองข้อสอบ ผู้ออกและกลั่นกรองข้อสอบต้องมีความรู้ในเนื้อหาที่จะออกข้อสอบอย่างดี (Content specialist) มีความสามารถในการออกข้อสอบ เก็บความลับได้ ไม่กวดวิชาและมีเวลามาทำงานจนแล้วเสร็จ

2. ขั้นตอนการพัฒนาข้อสอบ ประกอบด้วย การออกข้อสอบ การกลั่นกรองข้อสอบและการพัฒนาผู้ออกข้อสอบ ดังนี้

2.1 การออกข้อสอบ

1. การทำผังการออกข้อสอบ (Test Blueprint) โดยอิงมาตรฐานการเรียนรู้ตามหลักสูตรการศึกษาขั้นพื้นฐาน พ.ศ. 2544 (ปรับปรุง พ.ศ. 2551)

2. การสังเคราะห์เนื้อหาพร้อมของแบบเรียนที่กระทรวงศึกษาธิการให้ความเห็นชอบ

3. วางระดับความยากง่ายของข้อสอบ O-NET คือ ง่ายมาก 10% ค่อนข้างง่าย 15% ปานกลาง 60% ยาก 10% ยากมาก 5%

4. กำหนดรูปแบบของข้อสอบมี 3 รูปแบบ คือ 1) ปรนัย 4 ตัวเลือก 2) มีบทความให้อ่านและให้เลือกคำตอบแบบ 4 ตัวเลือกหรือมากกว่า 4 ตัวเลือก และ 3) ให้ระบุค่า/คะแนนที่ตัวเลือก (กรณีวิชาวิทยาศาสตร์และคณิตศาสตร์)

5. ทำรายละเอียดของผังการออกข้อสอบเป็น 3 ระดับ คือ ประเด็นหลัก ประเด็นรอง และประเด็นย่อย

6. ทำผังการออกข้อสอบ 2 ที่ละเอียดกว่าผังการออกข้อสอบ 1

7. เชิญผู้ออกข้อสอบที่มีคุณสมบัติตามข้อ 1 มาออกข้อสอบ โดย O-NET ป.6, ม.3 ใช้ครูโรงเรียนมาออกข้อสอบวิชาละ 5- 10 คน และ O-NET ม.6 ใช้อาจารย์มหาวิทยาลัยเป็นผู้ออกข้อสอบ

8. จัดทำแบบคู่ขนานตามจำนวนที่ต้องการ

9. ทำผังการออกข้อสอบ 3 ที่ระบุเลขที่ข้อตามระดับความยากง่ายหรือความซับซ้อน

2.2 การกลั่นกรองข้อสอบ การกลั่นกรองข้อสอบมี 2 ระยะ ดังนี้

1. ระยะตรวจสอบความถูกต้องของภาษา ตัวสะกด ระยะห่างของข้อสอบ สทศ.มีผู้เชี่ยวชาญด้านภาษาไทยมาตรวจสอบให้

2. ระยะกลั่นกรองความถูกต้องของเนื้อหา สทศ. เชิญผู้เชี่ยวชาญภายนอกวิชาละ 3 - 5 คน มาตรวจสอบความถูกต้องของเนื้อหา หลังตรวจสอบได้เชิญประธานของกลุ่มออกข้อสอบมา

พิจารณาแก้ไข พร้อมเฉลยคำตอบด้วยการระบายในกระดาษคำตอบ และกำหนดเกณฑ์การให้คะแนน หลังจากกลั่นกรองทั้ง 2 ระยะแล้ว สทศ.จัดพิมพ์ต้นฉบับลงไข ตรวจสอบอีกครั้งก่อนส่งโรงพิมพ์

2.3 การพัฒนาผู้ออกข้อสอบ

เนื่องจาก สทศ.เปิดเผยข้อสอบให้สาธารณชนทราบหลังสอบ ข้อสอบที่ใช้สอบแล้วจึงไม่สามารถนำมาเข้าคลังข้อสอบได้ สทศ.จึงได้จัดให้มีการพัฒนาผู้ออกข้อสอบแทนในการพัฒนาผู้ออกข้อสอบ สทศ.ได้วิเคราะห์คะแนนและข้อสอบเพื่อให้เป็นข้อมูลแก่ผู้ออกข้อสอบ ดังนี้

- 1) ค่า Mean, Median, Mode, Minimum, Maximum, Standard deviation ระดับโรงเรียน สังเกต และประเทศ จำแนกตามกลุ่มสาระการเรียนรู้ สาระย่อย และมาตรฐานการเรียนรู้
- 2) จำนวนผู้เข้าสอบจำแนกตามช่วงคะแนน
- 3) จำนวนร้อยละผู้ตอบถูกรายข้อ
- 4) ค่าความยาก อำนาจจำแนก และค่าการเดา จากการวิเคราะห์ข้อสอบ (Item Analysis) ด้วยทฤษฎี CTT และ IRT
- 5) ค่าความเชื่อมั่น (Reliability) และความคลาดเคลื่อนในการวัด (Standard Error of Measurement: SEM) จากการวิเคราะห์แบบทดสอบ (Test Analysis) ในการพัฒนาผู้ออกข้อสอบ ใช้การสัมมนาโดยอธิบายผลวิเคราะห์ควบคู่กับ ข้อสอบเพื่อให้เห็นจุดอ่อน/ จุดแข็งรายข้อ เพื่อให้ผู้ออกข้อสอบนำไปปรับปรุงการออกข้อสอบในครั้งต่อไป

3. นวัตกรรมใหม่ของระบบพัฒนาข้อสอบของ สทศ.

3.1 รูปแบบข้อสอบ สทศ. นำผลวิเคราะห์ข้อสอบโดยเฉพาะพฤติกรรมคำตอบข้อสอบ เช่น การระบายแบบทั้งดิ่ง และ ค่า parameter c (ค่าการเดา) มาพิจารณาปรับรูปแบบข้อสอบ ได้เป็น 3 แบบ คือ 1) คงจำนวนข้อสอบปรนัย 4 ตัวเลือกไว้แต่ให้มีจำนวนน้อย 2) เพิ่มข้อสอบที่ให้อ่านบทความ แล้วถามด้วยคำถามแบบปรนัยหลายตัวเลือก และ 3) เพิ่มข้อสอบที่ให้คิดเลข ค่าคะแนนแล้วระบายคำตอบ

การหารูปแบบใหม่ๆ ทำให้เกิดความคิดสร้างสรรค์ในการผลิตข้อสอบ และนำไปสู่การวัดได้ค่าความสามารถที่แท้จริง (True Score) ของผู้สอบการที่มีรูปแบบของข้อสอบที่หลากหลาย สทศ.ได้วางระบบการตรวจให้คะแนนไว้ 2 ระยะ คือ

ระยะที่ 1 การอ่านกระดาษคำตอบด้วยเครื่อง OMR เพื่อบันทึกผลการอ่าน

ระยะที่ 2 การเขียนโปรแกรมหยอดคะแนนรายข้อ เพราะแต่ละข้อคะแนนไม่เท่ากัน

การดำเนินงานดังกล่าวมุ่งที่การวัดความสามารถจริง (True Score) ของผู้สอบซึ่งเป็นเป้าหมายหลักของการวัดผล

3.2 การพิมพ์ข้อมูลผู้สอบบนหัวกระดาษคำตอบ ได้แก่ ชื่อ-สกุล เลขที่บัตรประชาชน เลขที่นั่งสอบ วิชาสอบ วันสอบ ห้องสอบ สนามสอบ

3.3 การทำ Test Bank เนื่องจากในแต่ละปี สทศ. ต้องสอบมากกว่า 1 ครั้ง/วิชา จึงหาวิธีที่จะออกข้อสอบให้อยู่ในรูปแบบ/ชุด เพื่อการเรียกใช้ที่รวดเร็ว ดังนั้นจึงต้องออกข้อสอบให้เป็นฉบับ/ชุดที่คู่ขนานโดยจัดทำเข้าคลัง ๆ ละอย่างน้อย 5 ฉบับที่คู่ขนานกันในเรื่องเนื้อหา วัตถุประสงค์ในการวัดและระดับความยากง่ายการทำ Test Bank นี้ นอกจากจะสามารถเรียกใช้งานได้รวดเร็ว ยังช่วยให้ออกข้อสอบได้เร็วขึ้นด้วย เพราะผู้ออกข้อสอบจะดำเนินการไปพร้อมกันกับจำนวนฉบับที่ต้องการผลิต เช่น ถ้าต้องการผลิตข้อสอบเนื้อหา A จำนวน 10 ข้อ 5 ฉบับ ก็จะต้องผลิต 50 ข้อ โดยทำเป็นชุดๆ ละ 10 ข้อที่คู่ขนาน

งานวิจัยที่เกี่ยวข้อง

ในการศึกษางานวิจัยที่เกี่ยวข้องนี้ ผู้วิจัยแบ่งการนำเสนองานวิจัยออกเป็น 4 ประเด็น คือ งานวิจัยที่เกี่ยวข้องกับคำถามปลายเปิดวิชาคณิตศาสตร์ งานวิจัยที่เกี่ยวข้องกับทฤษฎีการสรุปอ้างอิง งานวิจัยที่เกี่ยวข้องกับโมเดลหลายองค์ประกอบของราล์ซ และงานวิจัยที่เกี่ยวข้องกับทฤษฎีการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราล์ซร่วมกัน โดยขอแนะนำตามลำดับ ดังนี้

งานวิจัยที่เกี่ยวข้องกับคำถามปลายเปิดวิชาคณิตศาสตร์

คำถามปลายเปิดเป็นคำถามที่มีคุณค่าในการวัดความรู้ ความเข้าใจและกระบวนการคิดของผู้สอบ ตลอดจนกระบวนการแก้ปัญหา ซึ่งคำถามปลายเปิดมีทั้งคำถามที่มีคำตอบถูกหลายคำตอบ หรือมีคำตอบถูกคำตอบเดียวแต่มีหลายวิธีการได้มาซึ่งคำตอบ จึงมีนักวิจัยได้ศึกษาหรือใช้แบบทดสอบปลายเปิดอย่างกว้างขวาง งานวิจัยที่น่าเสนอจะเน้นเฉพาะกรณีที่เป็นแบบทดสอบปลายเปิดวิชาคณิตศาสตร์

มากอนและคณะ (Magone; et al. 1994) ได้ศึกษาและประเมินการแสดงวิธีทำ การให้เหตุผลเชิงซับซ้อนในขอบเขตเนื้อหาวิชาคณิตศาสตร์ กลุ่มตัวอย่างนักเรียนเป็นนักเรียนเกรด 6 และเกรด 7 จำนวน 1,627 คน แบ่งเป็น 3 กลุ่มคือ กลุ่มที่ 1 สอบเกี่ยวกับความเข้าใจในความคิดรวบยอดทางคณิตศาสตร์ จำนวน 568 คน กลุ่มที่ 2 สอบเกี่ยวกับการแก้โจทย์ปัญหาทางคณิตศาสตร์ จำนวน 485 คนและกลุ่มที่ 3 สอบเกี่ยวกับการให้เหตุผลทางคณิตศาสตร์จำนวน 574 คน เครื่องมือที่ใช้เป็นแบบทดสอบปลายเปิดจำนวน 9 ข้อ ประกอบด้วยการวัดประเมินความเข้าใจในความคิดรวบยอดทางคณิตศาสตร์ ความสามารถในการแก้โจทย์ปัญหาและการให้เหตุผลทางคณิตศาสตร์ พบว่า

1. การวัดประเมินความเข้าใจในความคิดรวบยอดทางคณิตศาสตร์ในเรื่องทศนิยมนั้น นักเรียนเกรด 6 ตอบถูก 33 % และเพิ่มเป็น 59 % เมื่ออยู่เกรด 7 ครั้งหนึ่งของนักเรียนที่ตอบถูกให้เหตุผลได้ถูกต้องหรือผิดพลาดเพียงเล็กน้อย ส่วนใหญ่ใช้การเขียนบรรยายและให้เหตุผลในรูปของความสัมพันธ์ระหว่างตัวเลขและตำแหน่งของทศนิยมหรือจำนวนเต็มหรือใช้เศษส่วนที่เทียบเท่ากัน

2. การวัดประเมินความสามารถในการแก้โจทย์ปัญหาทฤษฎีจำนวน พบว่า ในการแสดงวิธีการแก้ปัญหานั้นนักเรียนใช้รูปภาพประมาณครึ่งหนึ่ง ใช้สัญลักษณ์ทางคณิตศาสตร์ประมาณ 1 ใน 5 ใช้ร่วมกันระหว่างการบรรยาย รูปภาพและสัญลักษณ์ประมาณ 1 ใน 4 และใช้การบรรยายประมาณ 5 % ทั้งนี้ นักเรียนใช้กลยุทธ์ในการแก้ปัญหามากมายไม่น้อยกว่า 5 วิธี แต่มีข้อบกพร่องคือ ความไม่เข้าใจในเงื่อนไขของปัญหา

3. การวัดประเมินการให้เหตุผลทางคณิตศาสตร์ โดยใช้แบบรูป (Pattern) พบว่า นักเรียนส่วนใหญ่เขียนรูปต่อไปจากรูปที่กำหนดได้และให้เหตุผลด้วยวิธีการบรรยายประมาณ 90 % นอกนั้นใช้รูปภาพหรือจุดในการแสดงคำตอบ ข้อบกพร่องที่พบคือ ประมาณ 40 % เขียนรูปต่อไปไม่ถูกต้องเนื่องจากไม่เข้าใจในความสัมพันธ์ของภาพ 2 มิติ ที่แสดงถึงจำนวนจุดและรูปร่างของภาพ และประมาณ 30% ของผู้สอบที่เขียนรูปถัดไปถูกต้องเพียงมิติใดมิติหนึ่ง เช่น รูปร่างของภาพไม่ถูกต้องแต่จำนวนจุดถูกต้องหรือจำนวนจุดไม่ถูกต้องแต่รูปร่างของภาพถูกต้อง และมี 16% ของผู้ที่ไม่ถูกต้องนั้นเกือบทำถูกต้องในทั้งสองมิติ ส่วนอีก 54% ของรูปไม่ถูกต้องนั้นทำผิดทั้งรูปร่างและจำนวนจุด

การศึกษาของมากอนและคณะทำให้ได้ข้อสรุปว่า นักเรียนมีการแสดงออกที่หลากหลาย นักเรียนแสดงวิธีแก้ปัญหาก็ใช้กลยุทธ์หลายรูปแบบที่แสดงถึงกระบวนการคิดที่ซับซ้อน สามารถตรวจสอบความเข้าใจผิดและข้อบกพร่องได้จากการให้เหตุผลและขั้นตอนการแก้ปัญห คำตอบของนักเรียนแสดงให้เห็นถึงความลุ่มลึกในขอบเขตของเนื้อหาวิชา และในระดับขั้นที่สูงขึ้นคำตอบของนักเรียนจะสมบูรณ์และซับซ้อนมากขึ้น

คอนเวย์ (Conway. 1999: online) ได้ศึกษาเกี่ยวกับการวัดความสามารถในการแก้ปัญหาลายเปิด เพื่อศึกษาความสามารถในการแก้ปัญหาลายเปิดจากการวัดลักษณะการคิด 3 ลักษณะ ได้แก่ ความคิดคล่อง ความคิดยืดหยุ่น และความคิดริเริ่ม รวมทั้งพิจารณาการแสดงหมวดหมู่ของคำตอบที่เป็นไปได้จากการแก้ปัญหาลายเปิด ซึ่งผู้วิจัยสามารถวัดความสามารถในการแก้ปัญหาลายเปิดได้จากกระบวนการแก้ปัญหของนักเรียน จากการวิจัยพบว่า สามารถวัดทักษะในการแก้ปัญหของนักเรียนได้จากลักษณะการคิดทั้งสามลักษณะ

โลว และเชน (Loe; & Chen. 2004: online) ได้ศึกษาผลของการใช้ปัญหาลายเปิดวิชาคณิตศาสตร์ในเรื่อง เรขาคณิตและพีชคณิต ในประเทศจีน โดยใช้ปัญหาลายเปิดในการเรียนการสอนตามแนวการปฏิรูปการศึกษาของจีน ใช้เวลาในการเก็บข้อมูลประมาณ 7 ปี และมีการ

เปรียบเทียบผลที่ได้ระหว่างนักเรียนที่มีระดับความสามารถสูง ปานกลาง และต่ำ ซึ่งพบว่าหลังจากนักเรียนได้เรียนโดยใช้ปัญหาปลายเปิดแล้ว นักเรียนที่มีความสามารถสูงจะสามารถทำคะแนนวิชาคณิตศาสตร์ได้เพิ่มสูงมากที่สุด ในขณะที่นักเรียนอีกสองกลุ่มจะมีระดับคะแนนเพิ่มขึ้นเพียงเล็กน้อยในปีแรก แต่เมื่อนักเรียนได้เรียนโดยใช้ปัญหาปลายเปิดต่อไปอย่างต่อเนื่อง พบว่าระดับคะแนนและความสามารถทางคณิตศาสตร์ของนักเรียนสองกลุ่มหลังเพิ่มขึ้นอย่างเป็นที่น่าพอใจ จากการสอบถามความพึงพอใจของนักเรียนจำนวน 194 คน พบว่า มีนักเรียน 85% ที่มีความพอใจในการเรียนวิชาคณิตและพีชคณิตโดยใช้ปัญหาปลายเปิด

ปรีชา เนาว์เย็นผล (2544) ได้พัฒนากิจกรรมการเรียนการสอนคณิตศาสตร์โดยใช้ปัญหาปลายเปิดสำหรับนักเรียนชั้นมัธยมศึกษาปีที่ 1 ผลการวิจัย พบว่า กิจกรรมการเรียนการสอนคณิตศาสตร์โดยใช้ปัญหาปลายเปิดมีประสิทธิภาพตามเกณฑ์ 70/70 ในช่วงแรกนักเรียนกลุ่มที่ได้รับการสอนโดยใช้ปัญหาปลายเปิดส่วนใหญ่มีความสามารถในการแก้ปัญหาค่อนข้างต่ำ ในระหว่างการเรียนรู้ครูต้องให้การถามกระตุ้นแนะแนวทางในการแก้ปัญหาคณิตศาสตร์ แต่ในช่วงสุดท้ายของการสอนพบว่า นักเรียนส่วนใหญ่มีความสามารถในการวางแผนแนวคิดในการแก้ปัญหาได้ด้วยตนเอง สำหรับด้านพฤติกรรมความคิดแก้ปัญหา ก่อนเรียนนักเรียนทุกคนในกลุ่มทดลองอยู่ในระดับต้องแก้ไข หลังเรียนนักเรียนส่วนใหญ่มีพฤติกรรมแก้ปัญหายุ่งในระดับดีถึงดีมาก และผลสัมฤทธิ์ทางการเรียนคณิตศาสตร์ของนักเรียนกลุ่มทดลองสูงกว่าคะแนนผลสัมฤทธิ์ทางการเรียนตามเกณฑ์ปกติของโรงเรียนที่ตั้งไว้ นอกจากนี้ยังพบว่านักเรียนในกลุ่มทดลองมีเจตคติที่ดีต่อวิชาคณิตศาสตร์

จิตติมา ขอบเอียด (2551) ได้ใช้ปัญหาปลายเปิดเพื่อส่งเสริมและเปรียบเทียบทักษะการให้เหตุผลและทักษะการสื่อสารทางคณิตศาสตร์ของนักเรียนชั้นมัธยมศึกษาปีที่ 1 ก่อนและหลังการใช้ปัญหาปลายเปิด โรงเรียนศรีบุญยานนท์ จังหวัดนนทบุรี จำนวนนักเรียน 45 คน เครื่องมือที่ใช้ในการวิจัยครั้งนี้ ได้แก่ แผนการจัดการเรียนรู้โดยใช้ปัญหาปลายเปิด เรื่อง การประยุกต์ 2 แบบทดสอบวัดทักษะการให้เหตุผลทางคณิตศาสตร์ และแบบสังเกตทักษะการสื่อสารทางคณิตศาสตร์ ดำเนินการสอนโดยใช้การจัดการเรียนรู้โดยใช้ปัญหาปลายเปิด จำนวน 15 คาบ ผลการศึกษาพบว่า ทักษะการให้เหตุผลทางคณิตศาสตร์ของนักเรียนหลังการใช้ปัญหาปลายเปิดสูงกว่าก่อนการใช้ปัญหาปลายเปิด อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และ ทักษะการสื่อสารทางคณิตศาสตร์ของนักเรียนหลังการใช้ปัญหาปลายเปิดสูงกว่าเกณฑ์ร้อยละ 70 อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 โดยมีค่าเฉลี่ยคิดเป็นร้อยละ 80.67

จากผลการศึกษาวิจัย สรุปได้ว่า แบบทดสอบปลายเปิดเป็นแบบทดสอบที่เปิดโอกาสให้นักเรียนแสดงวิธีแก้ปัญหาได้อย่างหลายรูปแบบซึ่งแสดงถึงกระบวนการคิดที่ซับซ้อน สามารถใช้ตรวจสอบความเข้าใจผิด ข้อบกพร่องได้จากการให้เหตุผลและขั้นตอนการแก้ปัญหา คำตอบของ

นักเรียนแสดงให้เห็นถึงความลุ่มลึกในเนื้อหาวิชา และการศึกษาวิจัยส่วนใหญ่เป็นการนำคำถามปลายเปิดไปใช้ในกิจกรรมการเรียนการสอนในห้องเรียน ซึ่งผลการวิจัยต่างพบว่า หลังจากนำคำถามปลายเปิดไปใช้อย่างต่อเนื่องนั้นทำให้ความสามารถ พฤติกรรมการแก้ปัญหา การให้เหตุผล และการสื่อสารของผู้เรียนเพิ่มสูงขึ้น ตลอดจนมีเจตคติที่ดีต่อวิชาคณิตศาสตร์เพิ่มขึ้นด้วย

งานวิจัยที่เกี่ยวข้องกับทฤษฎีการสรุปอ้างอิง

ครอนบาคและคณะ (Cronbach; et al. 1972) ได้เสนอทฤษฎีสำหรับวิเคราะห์ความน่าเชื่อถือของผลการวัดพฤติกรรม สำหรับสถานการณ์ของการวัดผลในลักษณะต่าง ๆ ซึ่งต่อมาเป็นที่รู้จักกันในชื่อของทฤษฎีการสรุปอ้างอิง (Generalizability Theory: G-Theory) โดยทฤษฎีนี้ได้ขยายแนวคิดของความเชื่อมั่น (Reliability) ตามทฤษฎีการทดสอบแบบมาตรฐานเดิม ให้สามารถประมาณค่าความเชื่อมั่นในสถานการณ์หรือเงื่อนไขการทดสอบในลักษณะต่าง ๆ ได้ ทำให้มีนักวิจัยได้นำทฤษฎีนี้ไปใช้อย่างกว้างขวาง โดยในปี 1984 ไอบราฮิม (Ibrahim. 1984) ได้ประยุกต์ทฤษฎีการสรุปอ้างอิงในการประมาณค่าความแปรปรวนของการประเมินเป้าหมายของการศึกษา (The Rating of Educational Goals) โดยสุ่มตัวอย่างครู 80 คน และนักศึกษา 80 คน องค์ประกอบที่ศึกษาคือ ผู้ตรวจ กลุ่มผู้ตรวจ จำนวนครั้งของการประเมิน ถิ่นที่อยู่ของผู้ตรวจ ชนิดของวัตถุประสงค์ สถานที่ทำงานของผู้ตรวจ และเพศของผู้ตรวจ พบว่าปัจจัยที่มีผลต่อการประเมินมากที่สุดคือ ผู้ตรวจและกลุ่มของผู้ตรวจ สถานที่ทำงานของผู้ตรวจมีผลเล็กน้อยเท่านั้น ส่วนองค์ประกอบอื่นที่เหลือไม่มีผลต่อการประเมิน

โอเบรียนและโจนส์ (O'Brien; & Jones. 1986) ได้ใช้ทฤษฎีการสรุปอ้างอิงในการประมาณค่าความเชื่อมั่นของตัวแปรในระดับโรงเรียน 16 ตัว ซึ่งเป็นค่าเฉลี่ยหรือร้อยละของกลุ่มตัวอย่างนักเรียน ตัวแปรแบ่งเป็น 5 กลุ่ม คือ ค่าเฉลี่ยของสถานภาพทางสังคมและเศรษฐกิจของครอบครัว ค่าเฉลี่ยของผลสัมฤทธิ์ทางการเรียน ร้อยละของนักเรียนที่มีบิดาและมารดาอาศัยอยู่ด้วยกัน ร้อยละของคุณภาพของห้องสมุดและการเรียนการสอนของโรงเรียน และร้อยละของนักเรียนที่ยอมรับกฎระเบียบของโรงเรียนกลุ่มตัวอย่าง กลุ่มตัวอย่างเป็นนักเรียนชั้นมัธยมศึกษาปีที่ 2 จาก 1,122 โรงเรียน โรงเรียนละ 36 คน ผลการวิจัยพบว่า ความเชื่อมั่นในการตอบแบบสอบถามเพิ่มขึ้นตามการเพิ่มขึ้นของจำนวนผู้ประเมิน จำนวนข้อคำถามและจำนวนโรงเรียน แสดงให้เห็นว่า จำนวนผู้ประเมิน จำนวนข้อคำถามและจำนวนสถานที่ทำงานของผู้ประเมินมีผลต่อค่าความเชื่อมั่นของการประเมิน

นอกจากนี้ยังมีการประยุกต์ใช้ในการทดสอบทางภาษา โดย เบรนนอน เกาและคอลตัน (Brennan; Gao; & Colton. 1995) ได้ใช้ทฤษฎีการสรุปอ้างอิงในการวิเคราะห์แบบทดสอบทักษะการฟังและการเขียนของผู้สอบ 50 คน ผู้ตรวจชุดละ 3 คน ผู้ตรวจทักษะการพูดและการเขียนเป็นคนละชุดกัน เพื่อศึกษาปัจจัยที่มีอิทธิพลต่อค่าความเชื่อมั่นและความแม่นยำของการวัด ใช้ข้อสอบ 12 ข้อ

ผู้ตรวจ 3 คน โดยให้ผู้สอบฟังเทปจำนวน 12 ข้อ ฟังทีละข้อและให้เขียนบันทึก ประเมินทั้งทักษะการฟังและทักษะการเขียน คะแนนการฟังพิจารณาจากความถูกต้อง (Accuracy) และความสมบูรณ์ของสารสนเทศที่ได้ ส่วนทักษะการเขียนใช้กฎเกณฑ์การให้คะแนนแบบรวม (Holistic scale) ใช้มาตรประเมินเป็น 0-5 คะแนน ผลการศึกษา พบว่า ความคลาดเคลื่อนมาตรฐาน (SE) ของทั้งทักษะการฟังและการเขียนมีค่าน้อยมากสัมพันธ์กับค่าประมาณขององค์ประกอบความแปรปรวน ในการศึกษา G นั้น องค์ประกอบที่มีความแปรปรวนมากที่สุดของทักษะการฟัง(0.324) และการเขียน (0.691) คือ $\hat{\sigma}^2(p)$ แสดงว่า ความแตกต่างของผู้สอบสัมพันธ์กับระดับความสามารถของผู้สอบ ค่าประมาณองค์ประกอบความแปรปรวน $\hat{\sigma}^2(pt)$ มีขนาดใหญ่ แสดงว่า อันดับความสามารถของผู้สอบที่ต่างกันเกิดจากระดับความยากของข้อสอบที่ต่างกัน (หรือมีลำดับของระดับความยากของข้อสอบต่างกันเมื่อกลุ่มผู้สอบต่างกัน) ทำให้เกิดความแปรปรวนของความคลาดเคลื่อน นั่นคือ จำนวนข้อสอบมีอิทธิพลต่อความแม่นยำของการวัด ความแตกต่างของการจัดอันดับผู้สอบขึ้นอยู่กับข้อสอบที่ใช้ หากใช้ข้อสอบน้อยจะทำให้มีความกำกวมของผลการวัด ส่งผลต่อความเชื่อมั่นและความคลาดเคลื่อนมาตรฐานของการวัด การเพิ่มจำนวนผู้ตรวจมากกว่า 1 คนทำให้มีความแม่นยำของการวัดเพิ่มมากขึ้นและใช้ข้อสอบน้อยกว่า 12 ข้อได้ ทั้งนี้ค่าประมาณองค์ประกอบความแปรปรวน $\hat{\sigma}^2(r)$ และ $\hat{\sigma}^2(pr)$ มีค่าน้อยมาก แสดงว่าผู้ตรวจมีความเข้มงวดโดยเฉลี่ยแล้วเท่า ๆ กัน และผู้ตรวจจัดลำดับของผู้สอบเหมือนกัน ถึงแม้ว่าความแปรปรวนของผู้ตรวจจะมีค่าน้อย แต่ผู้วิจัยไม่ควรเพิกเฉยต่ออิทธิพลของความผันแปรของผู้ตรวจที่มีความคลาดเคลื่อน ในการศึกษา D พบว่า เมื่อเพิ่มผู้ตรวจจาก 1 คนเป็น 2 คน ความแปรปรวนของความคลาดเคลื่อนสมบูรณ์และสัมพันธ์ (ความคลาดเคลื่อนมาตรฐานของการวัดสำหรับการตัดสินใจเชิงสมบูรณ์และเชิงสัมพันธ์) มีเปอร์เซ็นต์ลดลงมากกว่าการเพิ่มผู้ตรวจจาก 2 คนเป็น 3 คน ดังนั้น หากต้องการพัฒนาความแม่นยำของการวัด ต้องใช้ผู้ตรวจตั้งแต่ 2 คนขึ้นไปดีกว่าใช้ผู้ตรวจเพียงคนเดียว และเมื่อประมาณค่าสัมประสิทธิ์การสรุปอ้างอิง (Generalizability coefficient: $E\rho^2$) ทั้งการฟังและการเขียน พบว่าเมื่อเพิ่มผู้ตรวจจาก 1 คน เป็น 2 คน มีค่า $E\rho^2$ ตั้งแต่ 0.7 ขึ้นไป และมีสัดส่วนการเพิ่มขึ้นมากกว่าการเพิ่มผู้ตรวจจาก 2 คนเป็น 3 คน แต่การเพิ่มจำนวนข้อสอบทำให้ค่าสัมประสิทธิ์การสรุปอ้างอิงเพิ่มขึ้นไม่มากนัก และดัชนีความเชื่อถือได้ (Index of dependability: ϕ) เมื่อกำหนดจำนวนผู้ตรวจและข้อสอบคงที่ ค่าของ ϕ มีค่าน้อยกว่าหรือเท่ากับ $E\rho^2$ แต่มีความสอดคล้องกัน กล่าวคือ การเพิ่มขึ้นของดัชนีความเชื่อถือได้จากผู้ตรวจ 1 คน เป็น 2 คน มีสัดส่วนการเพิ่มขึ้นมากกว่าการเพิ่มจากผู้ตรวจ 2 คนเป็น 3 คน

ส่วน เลน และคณะ (Lane; et al. 1996) ได้ศึกษาค่าสัมประสิทธิ์การสรุปอ้างอิงและความเที่ยงตรงของการประเมินผลงานของนักเรียนในวิชาคณิตศาสตร์ที่พัฒนาจากเครื่องมือประเมินความรู้

แบบ QUASAR ชื่อว่า QCAI ซึ่งเป็นชุดของข้อคำถามปลายเปิด (Opened-ended) ที่เป็นการประเมินการแก้ปัญหาทางคณิตศาสตร์ การให้เหตุผลทางคณิตศาสตร์ และการสื่อสารทางคณิตศาสตร์ โดยเน้นที่วิธีการตรวจให้คะแนนแบบการประเมินรวม (Holistic) และมีกฎเกณฑ์การให้คะแนน (Scoring rubric) แล้วจึงใช้ทฤษฎีการสรุปอ้างอิงตรวจสอบความสอดคล้องภายในข้อสอบและภายในผู้ตรวจ โดยมีแบบแผน 3 รูปแบบ คือ 1) $p \times t$ ผู้สอบทำข้อสอบทุกข้อ 2) $p \times t \times r$ ผู้สอบทำข้อสอบทุกข้อและผู้ตรวจแต่ละคนตรวจข้อสอบทุกข้อ และ 3) $(p : s) \times t$ ผู้สอบที่แนบอยู่ในแต่ละโรงเรียนทำข้อสอบทุกข้อ วิเคราะห์ข้อมูลโดยใช้โปรแกรม GENOVA ผลการวิจัย พบว่า ความคลาดเคลื่อนเนื่องจากผู้ตรวจมีค่าน้อย เนื่องมาจากการให้รายละเอียดเกี่ยวกับกฎเกณฑ์การให้คะแนนและตัวอย่างการตอบของนักเรียน รวมทั้งการฝึกฝนวิธีการตรวจให้คะแนนกับผู้ตรวจถูกต้องเพียงพอ ค่าความเชื่อมั่นที่ได้จากการศึกษา D Study แบบสัมบูรณ์ พบว่าเมื่อจำนวนนักเรียนเท่ากับ 350 คน ค่าสัมประสิทธิ์อยู่ระหว่าง 0.80 – 0.97 ขึ้นอยู่กับชุดของข้อสอบและระดับชั้น

สวาร์ทซ์ และคณะ (Swartz; et al. 1999) ได้ใช้ทฤษฎีการสรุปอ้างอิงในการประมาณค่าความเชื่อมั่นของคะแนนงานเขียนที่ได้จากวิธีการตรวจแบบรวม (Holistic) และแบบย่อย (Analytic) ของแบบทดสอบมาตรฐาน (Standardized) และแบบทดสอบที่ไม่ใช่มาตรฐาน (Nonstandardized) ที่ครูใช้วัดประเมินในห้องเรียนตนเอง เพื่อดูอิทธิพลของจำนวนผู้ตรวจและชนิดของการตัดสินใจที่มีต่อความเชื่อมั่นของคะแนน พบว่า ความเชื่อมั่นมีค่าเพิ่มขึ้นเมื่อเพิ่มจำนวนผู้ตรวจ และต้องใช้ผู้ตรวจจำนวน 4 คนจึงจะทำให้ค่าความเชื่อมั่นอยู่ในเกณฑ์ที่ยอมรับได้ของนูนัลลี (Nunnally 's criterion. 1967; citing Swartz; et al. 1999: 505) คือ 0.90 เมื่อพิจารณาความเชื่อมั่นของการตัดสินใจเชิงสัมพัทธ์ (Relative decisions) ของผู้ตรวจ 3 คน ของวิธีการตรวจแบบรวมเป็น 0.94 และวิธีการตรวจแบบย่อยอยู่ในช่วง 0.71-0.97 การใช้ผู้ตรวจน้อยกว่า 3 คน จะทำให้ค่าความเชื่อมั่นต่ำกว่าเกณฑ์ที่ยอมรับได้ และค่าความเชื่อมั่นของการตัดสินใจเชิงสัมบูรณ์ (Absolute decisions) มีความสอดคล้องกับค่าความเชื่อมั่นของการตัดสินใจเชิงสัมพัทธ์

ชู และวูล์ฟ (Chiu; & Wolfe. 2008) ได้ศึกษาการออกแบบการวิเคราะห์ข้อมูลที่ขาดหาย (Missing) โดยการจัดแบ่งข้อมูลให้อยู่ในรูปข้อมูลเมตริกซ์ที่มีขนาดเล็กกว่า ได้แก่ การออกแบบไขว้ (Crossed), MBIB และแบบแฝง ใช้การวิเคราะห์การสรุปอ้างอิง โดยใช้การจำลองข้อมูลแบบมอนติคาร์โล (Monte Carlo Simulation) เพื่อประมาณค่าองค์ประกอบความแปรปรวน ค่าสัมประสิทธิ์ความเชื่อมั่น ผลการศึกษาพบว่า ความแปรปรวนขององค์ประกอบต่าง ๆ ในแต่ละรูปแบบการจัดข้อมูล มีค่าคล้ายคลึงกัน โดยในรูปแบบแฝง (Nested) พบว่าองค์ประกอบผู้สอบ และปฏิสัมพันธ์ระหว่างผู้สอบและผู้ตรวจในข้อสอบแต่ละข้อและความคลาดเคลื่อนที่ไม่อาจจะระบุแหล่งได้ มีค่าค่อนข้างน้อย ความแปรปรวนของข้อสอบและปฏิสัมพันธ์ของผู้สอบและข้อสอบมีค่ามากกว่าใน

รูปแบบไขว้ หรือ MBIB อย่างไรก็ตามโดยภาพรวมแล้วความแปรปรวนของผู้สอบ ความแปรปรวนของปฏิสัมพันธ์ระหว่างผู้สอบและข้อสอบ และความแปรปรวนของปฏิสัมพันธ์ระหว่างผู้สอบกับผู้ตรวจในข้อสอบแต่ละข้อและความคลาดเคลื่อนที่ไม่อาจระบุแหล่งได้เป็นองค์ประกอบที่มีค่าสูง

ในส่วนของการงานวิจัยในประเทศไทยนั้น มีการนำโมเดลนี้ไปประยุกต์ใช้ทั้งในการประมาณค่าองค์ประกอบความแปรปรวน ประมาณค่าความเชื่อมั่น โดยในปี 2531 แดง กลางท่าใต้ (2531) ได้ประยุกต์ทฤษฎีการสรุปอ้างอิงในการหาค่าความเชื่อมั่นของการประเมินความเที่ยงตรงเชิงเนื้อหา ให้ผู้เชี่ยวชาญทุกคนประเมินความตรงของข้อสอบทุกข้อว่าวัดตรงกับจุดประสงค์เพียงใด รูปแบบการศึกษา คือ $i \times (r : s)$ โดยผู้เชี่ยวชาญทุกคนประเมินข้อสอบทุกข้อ ผู้เชี่ยวชาญสุ่มมาจากโรงเรียนแต่ละโรงเรียนจำนวนเท่า ๆ กัน เครื่องมือที่ใช้ในการวิจัย คือ แบบทดสอบวิชาคณิตศาสตร์ เรื่อง เซต ใช้มาตราประเมิน 5 มาตรา ผลการวิจัย พบว่า แหล่งความแปรปรวนที่มีอิทธิพลต่อการประมาณค่าความเที่ยง ได้แก่ ความแปรปรวนของข้อสอบ ความแปรปรวนของผู้เชี่ยวชาญ และปฏิสัมพันธ์ระหว่างข้อสอบกับโรงเรียน ขนาดกลุ่มตัวอย่างที่จะเป็นสมาชิกทั้งหมดในเอกภพขององค์ประกอบ ได้แก่ ข้อสอบ 9 ข้อ และผู้เชี่ยวชาญไม่เกิน 9 คนต่อโรงเรียน ซึ่งสุ่มมาจากโรงเรียนอย่างน้อย 7 โรงเรียนจะได้ค่าสัมประสิทธิ์การสรุปอ้างอิงมีค่าน้อย 0.80

ไพรัตน์ วงษ์นาม (2533) ได้ศึกษาและเปรียบเทียบผลของวิธีตรวจ การชี้แจงและการเรียนรู้ของผู้ตอบที่มีต่อสัมประสิทธิ์การสรุปอ้างอิง (Generalizability coefficient) ของแบบทดสอบความเรียงที่วัดความสามารถในการแสดงความคิดเห็นเกี่ยวกับข่าวและเหตุการณ์ของนักเรียนชั้นประถมศึกษาปีที่ 6 จำนวน 30 คน ข้อสอบความเรียง 5 ข้อ และครูผู้ตรวจให้คะแนน 20 คน สุ่มแบ่งผู้ตรวจให้คะแนนเป็น 2 กลุ่ม กลุ่มละ 10 คน สำหรับการตรวจ 2 วิธี คือ วิธีประเมินรวมและวิธีวิเคราะห์ ในแต่ละวิธีทำการตรวจ 3 ครั้ง คือ ครั้งแรกตรวจโดยใช้ประสบการณ์เดิมของผู้ตรวจให้คะแนน ครั้งสองตรวจตามเกณฑ์ที่ได้รับการชี้แจงจากผู้วิจัยโดยไม่รู้ผลการเรียนของผู้ตอบ ครั้งที่สามตรวจตามเกณฑ์ที่ได้รับการชี้แจงจากผู้วิจัยเหมือนครั้งที่สองแต่รู้ผลการเรียนของผู้ตอบด้วย ผลการวิจัย พบว่า เมื่ออ้างอิงสรุปไปยังเอกภพของข้อสอบและผู้ตรวจให้คะแนนพร้อมกัน ค่าสัมประสิทธิ์การสรุปอ้างอิงของแบบทดสอบความเรียงที่ตรวจด้วยวิธีประเมินรวม มีค่าอยู่ระหว่าง 0.3328-0.4782 และวิธีวิเคราะห์มีค่าอยู่ระหว่าง 0.3348-0.5895 และใช้ข้อสอบ 5 ข้อ ผู้ตรวจให้คะแนนอย่างน้อย 5 คน ใช้วิธีการตรวจวิเคราะห์ตามที่ได้รับการชี้แจงและไม่รู้ผลการเรียนของผู้ตอบ จะให้ค่าสัมประสิทธิ์การสรุปอ้างอิงไม่น้อยกว่า 0.50 เมื่ออ้างอิงสรุปไปยังเอกภพของผู้ตรวจให้คะแนนอย่างเดียว ค่าสัมประสิทธิ์การสรุปอ้างอิงของแบบทดสอบความเรียงที่ตรวจด้วยวิธีประเมินรวม มีค่าอยู่ระหว่าง 0.4730-0.6865 และตรวจด้วยวิธีวิเคราะห์มีค่าอยู่ระหว่าง 0.5985-0.7761 วิธีการตรวจตามที่ได้รับการชี้แจงและการรู้ผลการเรียนของผู้ตอบต่างกันไม่มีผลต่อความแตกต่างของสัมประสิทธิ์

การสรุปอ้างอิงสำหรับเอกภพของข้อสอบและผู้ตรวจพร้อมกัน เมื่ออ้างอิงสรุปไปยังเอกภพของผู้ตรวจเพียงอย่างเดียว พบว่า ค่าสัมประสิทธิ์การสรุปอ้างอิงของการตรวจตามที่ได้รับการชี้แจงมีค่าสูงกว่าการตรวจโดยใช้ประสบการณ์เดิม เฉพาะกรณีที่ผู้ตรวจที่ได้รับการฝึกอบรมแต่ไม่รู้ผลการเรียนของผู้ตอบเท่านั้น ถ้ารู้ผลการเรียนของผู้ตอบค่าสัมประสิทธิ์จะไม่แตกต่างกัน ค่าสัมประสิทธิ์ของการตรวจที่ผู้ตรวจไม่รู้ผลการเรียนของผู้ตอบมีค่าสูงกว่าการตรวจที่รู้ผลการเรียนของผู้ตอบ วิธีตรวจวิเคราะห์ตามที่ได้รับการชี้แจงและผู้ตรวจไม่รู้ผลการเรียนของผู้ตอบ

ชัยฤทธิ์ ศิลาเดช (2540) ได้พัฒนาแฟ้มสะสมงานในการประเมินผลการเรียนวิชาภาษาอังกฤษ ระดับชั้นมัธยมศึกษาปีที่ 3 โดยได้ศึกษาคุณภาพของการใช้แฟ้มสะสมงานในการประเมินผลการเรียน ในด้านความเที่ยงตรงและความเชื่อมั่นของการประเมิน พบว่า ค่าสหสัมพันธ์ระหว่างอันดับที่จากคะแนนเฉลี่ยของผู้ประเมินสองคนกับการจัดอันดับความสามารถในวิชาภาษาอังกฤษของนักเรียนที่อาจารย์ที่ปรึกษาเป็นผู้จัดอันดับ เป็น 0.94 มีนัยสำคัญทางสถิติที่ระดับ .001 แสดงว่าการประเมินผลโดยใช้แฟ้มสะสมงานมีความเที่ยงตรงตามสภาพ ในด้านความเชื่อมั่นในการประเมินผลนั้น พบว่าถ้าใช้ผู้ประเมินคนเดียว จะทำให้ได้ความเชื่อมั่นค่อนข้างต่ำ และเมื่อเพิ่มจำนวนผู้ประเมินมากขึ้น คือตั้งแต่ 2 คนขึ้นไปจะทำให้ความเชื่อมั่นเพิ่มสูงขึ้นมาก แต่เมื่อใช้ผู้ประเมินมากกว่า 4 คน ทำให้ค่าความเชื่อมั่นเพิ่มขึ้นอีกเพียงเล็กน้อย ดังนั้น ไม่ควรใช้ผู้ประเมินเพียงคนเดียว เพราะจะทำให้ค่าความเชื่อมั่นของการให้คะแนนค่อนข้างต่ำ จำนวนผู้ประเมินที่เหมาะสมน่าจะประมาณ 2-3 คน และไม่จำเป็นต้องใช้ผู้ประเมินมากเกินไป

พรณี เจียมสุบุตร (2543) ได้ทำวิจัยเพื่อเปรียบเทียบความเชื่อมั่นของแบบทดสอบวัดความสามารถในการแก้โจทย์ปัญหาทางคณิตศาสตร์ที่มีจำนวนผู้ตรวจและวิธีการตรวจต่างกัน ผลการศึกษาพบว่า ค่าความเชื่อมั่นของแบบทดสอบวัดความสามารถในการแก้โจทย์ปัญหาทางคณิตศาสตร์ โดยการตรวจด้วยวิธีการวิเคราะห์ที่มีผู้ตรวจ 2 คนและ 3 คน มีค่าเท่ากับ 0.92 และ 0.96 ตามลำดับ ซึ่งแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ส่วนการตรวจให้คะแนนด้วยวิธีรวมที่มีผู้ตรวจ 2 คน และ 3 คน มีค่าเท่ากับ 0.97 และ 0.98 ตามลำดับ ซึ่งแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ ความเชื่อมั่นของแบบทดสอบวัดความสามารถในการแก้โจทย์ปัญหาทางคณิตศาสตร์ ที่ใช้วิธีการตรวจแบบวิธีรวมและวิธีวิเคราะห์ของผู้ตรวจให้คะแนน 2 คนมีค่าเท่ากับ 0.97 และ 0.92 ตามลำดับ แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ส่วนแบบทดสอบที่ใช้วิธีการตรวจให้คะแนนแบบวิธีรวมและวิธีวิเคราะห์ของผู้ตรวจ 3 คน มีค่าความเชื่อมั่นเท่ากับ 0.98 และ 0.96 ตามลำดับ ซึ่งแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05

อุษณีย์ บัวศิริพันธ์ (2543) ได้ทำการศึกษาเปรียบเทียบค่าสัมประสิทธิ์สรุปอ้างอิงของแบบทดสอบวิชาคณิตศาสตร์ที่มีวิธีการตรวจ จำนวนผู้ตรวจ และประสบการณ์ของผู้ตรวจต่างกัน

โดยมีจุดมุ่งหมายเพื่อเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิงของแบบทดสอบวิชาคณิตศาสตร์ที่มีวิธีการตรวจให้คะแนนต่างกัน คือ การตรวจให้คะแนนแบบประเมินรวมโดยใช้เกณฑ์การให้คะแนนของผู้ตรวจเองและการตรวจให้คะแนนแบบประเมินรวมโดยใช้เกณฑ์การให้คะแนนแบบรูบริกส์ และมีจำนวนผู้ตรวจต่างกัน คือ 3, 4, 5 และ 6 คน และผู้ตรวจมีประสบการณ์ต่างกัน คือ ผู้ตรวจที่มีประสบการณ์มากกว่า 5 ปี และผู้ตรวจที่มีประสบการณ์ไม่เกิน 5 ปี ผลการวิจัยพบว่า ค่าสัมประสิทธิ์การสรุปอ้างอิงของแบบทดสอบวิชาคณิตศาสตร์ที่มีวิธีการตรวจให้คะแนนต่างกัน เมื่อมีจำนวนผู้ตรวจเท่ากัน มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ค่าสัมประสิทธิ์การสรุปอ้างอิงของแบบทดสอบวิชาคณิตศาสตร์ที่มีวิธีการตรวจให้คะแนนต่างกัน เมื่อผู้ตรวจมีประสบการณ์เหมือนกัน มีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ ค่าสัมประสิทธิ์การสรุปอ้างอิงของแบบทดสอบวิชาคณิตศาสตร์ที่มีจำนวนผู้ตรวจต่างกัน เมื่อมีวิธีการตรวจให้คะแนนเหมือนกัน มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และค่าสัมประสิทธิ์การสรุปอ้างอิงของแบบทดสอบวิชาคณิตศาสตร์ที่ผู้ตรวจมีประสบการณ์ต่างกัน เมื่อมีวิธีการตรวจให้คะแนนเหมือนกัน มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01

ปวีณา ปิอาทิตย (2545) ได้ศึกษาจำนวนผู้ตรวจและจำนวนงานเขียนที่เหมาะสม เมื่อใช้เกณฑ์การให้คะแนนที่ต่างกัน โดยเปรียบเทียบความเที่ยงซึ่งแทนด้วยค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนการประเมินงานเขียนจากผู้ตรวจจำนวน 4 คน เมื่อใช้เกณฑ์การให้คะแนนแบบภาพรวม (Holistic scoring rubric) และผู้ตรวจจำนวน 4 คนที่ใช้เกณฑ์การให้คะแนนแบบแยกองค์ประกอบ (Analytic scoring rubric) เพื่อหาจำนวนผู้ตรวจและจำนวนงานเขียนที่เหมาะสมเมื่อใช้เกณฑ์การให้คะแนนแบบภาพรวม (Holistic scoring rubric) หรือ เกณฑ์การให้คะแนนแบบแยกองค์ประกอบ (Analytic scoring rubric) คำนวณค่าความเที่ยงจากค่าสัมประสิทธิ์การสรุปอ้างอิงด้วยโปรแกรมสำเร็จรูป GENOVA ผลการวิจัยพบว่า หากใช้จำนวนชิ้นงาน 117 ชิ้น ผู้ตรวจจำนวน 4 คน เมื่อใช้เกณฑ์การให้คะแนนแบบภาพรวมค่าสัมประสิทธิ์การสรุปอ้างอิงเท่ากับ 0.71484 และเมื่อใช้เกณฑ์การให้คะแนนแบบแยกองค์ประกอบค่าสัมประสิทธิ์การสรุปอ้างอิงเท่ากับ 0.81410 และถ้าต้องการใช้ผู้ตรวจ 1 คน กำหนดให้ค่าสัมประสิทธิ์การสรุปอ้างอิงของเกณฑ์การให้คะแนนแบบภาพรวมเป็น 0.2, 0.4, 0.6 และ 0.8 ใช้จำนวนชิ้นงานอย่างน้อย 4, 8, 16 และ 28 ชิ้น ตามลำดับ ส่วนเกณฑ์การให้คะแนนแบบแยกองค์ประกอบ เมื่อกำหนดค่าสัมประสิทธิ์การสรุปอ้างอิงเป็น 0.2 และ 0.4 ใช้จำนวนชิ้นงานน้อยกว่า 4 ชิ้น เมื่อกำหนดค่าสัมประสิทธิ์การสรุปอ้างอิงเป็น 0.6 ใช้จำนวนชิ้นงานอย่างน้อย 8 ชิ้น และเมื่อกำหนดค่าสัมประสิทธิ์การสรุปอ้างอิงเป็น 0.8 ใช้จำนวนชิ้นงานอย่างน้อย 12 ชิ้น ทั้งนี้ขึ้นอยู่กับผู้ปฏิบัติว่าจะเลือกใช้เกณฑ์การให้คะแนนให้เหมาะสมกับการเรียนการสอน

จากผลการศึกษางานวิจัยดังกล่าวข้างต้น (Ibrahim. 1984; O'Brien; & Jones. 1986; Brennan; Gao; & Colton. 1995; Lane; et al. 1996; Swartz; et al. 1999; ไพรัตน์ วงษ์นาม. 2533; ชัยฤทธิ์ ศิลาเดช. 2540; ปวีณา ปิอาทิตย. 2545; พรรณี เจียมสุบุตร. 2543; อุษณีย์ บัวศิริพันธ์. 2543) สรุปได้ว่า มีการนำทฤษฎีการสรุปอ้างอิงนี้ไปใช้อย่างกว้างขวางในหลายลักษณะ เช่น การวัดประเมินการปฏิบัติ การวัดประเมินทางด้านการพูด การเขียน และการแก้โจทย์ปัญหา เป็นต้น ซึ่งการวัดประเมินเหล่านี้ต่างมีแหล่งความคลาดเคลื่อนของคะแนนหลายแหล่ง และมีปัญหาในแง่ของความน่าเชื่อถือของคะแนน อันเนื่องมาจากผู้ตรวจให้คะแนน จึงมีการนำทฤษฎีนี้ไปใช้ทั้งในการประมาณค่าองค์ประกอบความแปรปรวนจากแหล่งต่าง ๆ ว่าแหล่งใดเป็นแหล่งของความคลาดเคลื่อนหลักและประมาณค่าสัมประสิทธิ์การสรุปอ้างอิง ผลการวิจัยสอดคล้องกันในประเทศนี้ว่า ผู้ตรวจให้คะแนนเป็นแหล่งความคลาดเคลื่อนที่สำคัญ ส่งผลต่อความน่าเชื่อถือและความเที่ยงตรงของคะแนน การเพิ่มจำนวนผู้ตรวจทำให้ค่าความเชื่อมั่นเพิ่มสูงขึ้น โดยนักวิจัยเหล่านี้ต่างได้ให้ข้อเสนอแนะถึงจำนวนผู้ตรวจว่าควรใช้ผู้ตรวจมากกว่าหนึ่งคน แต่จำนวนผู้ตรวจที่เหมาะสมนั้นผลการศึกษายังพบแตกต่างกัน และงานวิจัยส่วนใหญ่เป็นการศึกษาในกรณีที่ผู้ตรวจตรวจให้คะแนนข้อสอบทุกข้อของนักเรียนทุกคน ซึ่งค่อนข้างใช้เวลาในการตรวจมาก ทั้งนี้ทฤษฎีการสรุปอ้างอิงสามารถประมาณค่าความเชื่อมั่นได้โดยที่ผู้ตรวจไม่จำเป็นต้องตรวจให้คะแนนทุกข้อของผู้สอบทุกคน ซึ่งรูปแบบการเก็บรวบรวมข้อมูลหรือการตรวจให้คะแนนนั้นมีหลากหลายรูปแบบ ทั้งแบบไขว้ แบบแบ่ง หรือแบบแบ่งบางส่วน สามารถนำมาใช้วางแผนการตรวจให้คะแนนได้

งานวิจัยที่เกี่ยวข้องกับโมเดลหลายองค์ประกอบของราล์ซ

ลินาเคอร์ (Linacre. 1994) ได้พัฒนาโมเดลหลายองค์ประกอบของราล์ซ ขึ้นในปี 1989 ซึ่งเป็นการขยายแนวคิดทางทฤษฎีให้สามารถนำไปใช้อย่างได้กว้างขวางมากขึ้น ทั้งนี้มีการวัดประเมินจำนวนมากที่เกี่ยวข้องกับผู้ตรวจให้คะแนนหรือเกี่ยวข้องกับองค์ประกอบอื่น อีกทั้งปัจจุบันเน้นการวัดประเมินการปฏิบัติเพิ่มมากขึ้น จึงมีผู้นำโมเดลนี้ไปประยุกต์ใช้หรือทำการศึกษาวิจัยมากขึ้นโดยเฉพาะในต่างประเทศ ระยะแรกนั้น ผู้ที่โดดเด่นมากในการนำโมเดลนี้ไปใช้คือ เอนเกลฮาร์ด (Engelhard) โดยในปี 1992 เอนเกลฮาร์ด (Engelhard. 1992) ได้ศึกษาการใช้โมเดลหลายองค์ประกอบของราล์ซ ในการวัดความสามารถทางการเขียน โดยสุ่มกลุ่มตัวอย่างนักเรียนจำนวน 1,000 คน พบว่า ความเชื่อมั่นแยกส่วนของผู้ตรวจเป็น 0.87 ซึ่งมีความแตกต่างอย่างมีนัยสำคัญของความเข้มงวดของผู้ตรวจ แม้ว่าจะมีการอบรมผู้ตรวจอย่างเข้มข้นก็ตาม INFIT และ OUTFIT ของข้อสอบอยู่ระหว่าง 0.9 -1.1 และมีค่าความเชื่อมั่นแยกส่วนของข้อสอบเป็น 0.46 ซึ่งค่อนข้างต่ำแต่มีนัยสำคัญทางสถิติ แสดงว่าความยากของข้อสอบมีความแตกต่างกันเพียงเล็กน้อย คะแนนดิบมีค่าตั้งแต่ 10-40 คะแนน

(MD=27.6, SD=6.2) คะแนนโลจิตมีค่า 7.07 ถึง -6.84 โลจิต (MD=1.02, SD=2.46) ค่าความเชื่อมั่นแยกส่วนของผู้สอบเป็น 0.93 ซึ่งแตกต่างกันอย่างมีนัยสำคัญ และต่อมาอีก 2 ปี เอนเกลฮาร์ด (Engelhard. 1994) ได้ทำวิจัยเพื่อตรวจสอบความคลาดเคลื่อนของผู้ตรวจในการวัดประเมินงานเขียนโดยใช้โมเดลหลายองค์ประกอบของราสซัส ได้แก่ ความเข้มงวดของผู้ตรวจ (Rater severity) ผลจากฮาโล (Halo effect) แนวโน้มเข้าสู่ส่วนกลาง (Central tendency) และการจำกัดช่วง (Restriction of range) โดยสุ่มกลุ่มตัวอย่างจำนวน 264 คน ผู้ตรวจ 15 คน ใช้แบบทดสอบการเขียนของเกรด 8 ในการสอบปี 1990 ของจอร์เจีย ผลการศึกษาพบว่า มีความแตกต่างอย่างมีนัยสำคัญของความเข้มงวดของผู้ตรวจ มีผลของฮาโลในการประเมินแบบรวมมากกว่าแบบย่อยของผู้ตรวจ 2 คน มีการประเมินที่อยู่ในตรงค่ากลางของมาตรฐานประมาณ 80 % แสดงว่ามีการประเมินที่มีแนวโน้มเข้าสู่ส่วนกลาง หากยังไม่ได้ปรับแก้การกระจายของคะแนนดิบจะมีการจำกัดช่วงเกิดขึ้น นอกจากนี้เมื่อปรับแก้การประมาณค่าความสามารถในการเขียนแล้วพบว่ามีผลคลาดเคลื่อนของผู้ตรวจเพียงเล็กน้อยเท่านั้น

มายฟอร์ด และวูล์ฟ (Myford; & Wolfe. 2000: online) ได้นำโมเดลหลายองค์ประกอบของราสซัสไปใช้ในหน่วยงานการทดสอบทางการศึกษาของสหรัฐอเมริกา (Educational Testing Service : ETS) โดยศึกษาแหล่งของความผันแปรในระบบการวัดประเมินการพูดภาษาอังกฤษ (Test of Spoken English: TSE) หาช่วงของความผันแปรในแต่ละแหล่ง หาขอบเขตของแหล่งต่าง ๆ เหล่านี้ที่ส่งผลต่อการปฏิบัติของผู้สอบ เพื่อเป็นข้อเสนอแนะในการเปลี่ยนแปลงระบบการวัดประเมิน ข้อมูลที่ใช้วิเคราะห์เป็นคะแนน TSE ของเดือนกุมภาพันธ์และเมษายน ในปี 1997 ผลการวิจัย พบว่าแบบทดสอบสามารถจำแนกระดับความสามารถของผู้สอบตามสถิติได้เป็น 8 ระดับ ความสามารถของผู้สอบมีความเชื่อถือได้ในประเด็นของความแม่นยำและความคงที่ ค่าความเชื่อมั่นแยกส่วนของผู้สอบมีค่าเป็น 0.98 ความคลาดเคลื่อนมาตรฐานของการวัดมีการกระจายตลอดช่วงการแจกแจง โดยเฉพาะในส่วนปลายของการแจกแจง การตรวจข้อสอบข้อหนึ่งสัมพันธ์กับข้อสอบข้ออื่น ๆ ข้อสอบทั้ง 12 ข้อ มีค่า INFIT และ OUTFIT อยู่ในช่วงที่ยอมรับได้ (0.7-1.3) จึงมีความเหมาะสมที่จะสรุปถึงความสามารถของผู้สอบจากการวัดด้วยข้อสอบ 12 ข้อ ข้อสอบมีความยากแตกต่างกันเล็กน้อย รายการต่าง ๆ ของมาตรฐานวัดประเมิน TSE ที่มี 5 ระดับนั้นมีความชัดเจนในการแยกความแตกต่างในแต่ละระดับและสามารถใช้เกณฑ์เดียวกันได้กับข้อสอบทั้ง 12 ข้อ สำหรับผู้ตรวจนั้น มีระดับความเข้มงวดที่แตกต่างกัน หากปรับแก้ความแตกต่างของความเข้มงวดของผู้ตรวจจะทำให้คะแนน 2 ใน 3 ของผู้สอบทั้งหมดมีค่าเฉลี่ยแตกต่างจากคะแนนดิบเฉลี่ยประมาณ 0.5 – 3.6 คะแนน ซึ่งมีผลอย่างมากกับผู้สอบที่อยู่ใกล้คะแนนจุดตัด

ในปี 2000 นักการศึกษาชาวญี่ปุ่น นากามูระ (Nakamura. 2000: online) ได้ประยุกต์ใช้โมเดลหลายองค์ประกอบของราล์ฟในการวิเคราะห์ข้อมูลของการทดสอบการสื่อสารทางภาษา พบว่าสามารถตรวจสอบความสามารถนักเรียนแต่ละคนได้อย่างชัดเจน ค่าความเชื่อมั่นของนักเรียนเป็น 0.99 ข้อสอบจำนวน 11 ข้อจาก 15 ข้อ มีค่า INFIT และ OUTFIT อยู่ในช่วงที่ยอมรับได้ ค่าความเชื่อมั่นแยกส่วนของผู้ตรวจเป็น 0.96 แสดงว่าผู้ตรวจมีความเข้มงวดแตกต่างกัน แต่ยอมรับได้ในทางปฏิบัติ ค่า INFIT และ OUTFIT ของผู้ตรวจอยู่ในช่วงที่ยอมรับได้ ต่อมาในปี 2002 นากามูระ (Nakamura. 2002: online) ได้ประยุกต์ใช้โมเดลหลายองค์ประกอบของราล์ฟในการวิเคราะห์แบบทดสอบการเขียน โดยศึกษาว่าหากใช้ผู้ตรวจ 2 คน จะสามารถลดภาระงานของผู้ตรวจได้หรือไม่โดยที่ยังมีความน่าเชื่อถือทางสถิติ กลุ่มตัวอย่างที่ใช้เป็นนักศึกษาญี่ปุ่นในระดับปริญญาตรีจำนวน 32 คน กำหนดให้นักศึกษาเขียนเรียงความภายในเวลา 40 นาที ใช้ผู้ตรวจทั้งหมด 4 คน โดยให้ผู้ตรวจ 2 คนตรวจงานนักศึกษาแต่ละคน มีเกณฑ์การประเมินจำนวน 7 เกณฑ์ ใช้มาตรฐานประเมิน 4 ระดับ ผลการวิจัยสรุปได้ว่าการใช้ระบบการตรวจเป็นคู่สามารถลดภาระงานของผู้ตรวจได้มาตรฐานประเมินทั้ง 4 ระดับ (1 = ไม่ดี, 2, 3, 4 = ดีมาก) ของเกณฑ์การประเมินทั้งหมด มีสถิติ OUTFIT ที่อยู่ในช่วงที่ยอมรับได้ (0.8 – 1.1) และมีค่าการวัดเป็น -1.42, -.40, 1.37, 3.80 ช่วงความสามารถของผู้สอบค่อนข้างกว้าง (5 ถึง -1) มีนักศึกษา 5 คนที่ misfit และ 2 คนที่ overfit มีความแตกต่างอย่างมีนัยสำคัญในความเข้มงวดของผู้ตรวจ ค่า INFIT และ OUTFIT ของผู้ตรวจทุกคนอยู่ในช่วงที่ยอมรับได้ ความเชื่อมั่นระหว่างผู้ตรวจ (Inter – rater reliability) มีค่าสูง แม้ว่าจะมีระดับของความเข้มงวด/ใจดีแตกต่างกัน เกณฑ์ทั้ง 7 ที่ใช้ในการศึกษานั้น ด้านที่มีค่าความยากมากที่สุดคือ ไวยากรณ์ ส่วนด้านเนื้อหาและคำศัพท์นั้นง่ายที่สุดและมีความแตกต่างอย่างมีนัยสำคัญของค่าความยาก ค่า INFIT และ OUTFIT ของผู้ตรวจทั้ง 6 คู่อยู่ในช่วงที่ยอมรับได้และไม่มีอิทธิพลอย่างมีนัยสำคัญ ดังนั้น คะแนนที่ได้จึงไม่ขึ้นอยู่กับผู้ตรวจ นั่นคือ ผู้ตรวจทั้ง 4 คนไม่จำเป็นต้องตรวจข้อสอบนักเรียนทุกคน เนื่องจากความสอดคล้องของการตรวจอยู่ในช่วงที่ยอมรับได้

ชี (Chi. 2001) ได้ศึกษาเปรียบเทียบวิธีการตรวจให้คะแนนแบบรวม (Holistic method) กับการตรวจให้คะแนนแบบย่อย (Analytic method) ในการวัดประเมินการปฏิบัติโดยใช้โมเดลหลายองค์ประกอบของราล์ฟ กลุ่มตัวอย่างที่ใช้คือ นักเรียนจำนวน 43 คน ผู้ตรวจจำนวน 4 คน พบว่าสหสัมพันธ์ของคะแนนนักเรียนทั้ง 2 วิธีนั้นมีค่าสูง (สหสัมพันธ์แบบเพียร์สันมีค่าเป็น 0.89, สเปียร์แมนแรงค์ มีค่าเป็น 0.87) ค่าเฉลี่ยของคะแนนที่ได้จากการให้คะแนนแบบรวมสูงกว่าการให้คะแนนแบบย่อย เมื่อเปรียบเทียบเชิงสัมพัทธ์ไม่พบความแตกต่างอย่างมีนัยสำคัญของวิธีการให้คะแนนทั้งสองวิธี แต่มีความแตกต่างกันเมื่อเปรียบเทียบเชิงสมบูรณ์ การให้คะแนนแบบย่อยมีความสอดคล้องกันของความเข้มงวดของผู้ตรวจมากกว่าการให้คะแนนแบบรวม ช่วงของความเข้มงวดของผู้ตรวจที่ให้

คะแนนแบบรวมนักเรียนมากกว่าการให้คะแนนแบบย่อย และความเชื่อมั่นแยกส่วนของผู้ตรวจ (Rater separation reliability) ในการให้คะแนนแบบรวมนักเรียนมีค่าสูงกว่าการให้คะแนนแบบย่อย

บอนคและออกเคย์ (Bonk; & Ockey. 2003: online) ได้ใช้โมเดลหลายองค์ประกอบของราล์ซในการวิเคราะห์ข้อสอบแบบอภิปรายโดยใช้ปากเปล่าของนักศึกษาที่เรียนภาษาที่สอง ประเมินจากนักศึกษาวิชาเอกภาษาญี่ปุ่น ภาษาอังกฤษ โดยองค์ประกอบที่ศึกษาคือผู้สอบ ข้อสอบ ผู้ตรวจและมาตรประเมิน 5 ระดับ พบว่าข้อมูลนั้นเป็นเอกมิติ (Unidimension) และมีความสอดคล้องกับโมเดลองค์ประกอบที่สำคัญคือ ความสามารถของผู้สอบ รองลงไปคือความเข้มงวดของผู้ตรวจและข้อสอบ ความเข้มงวดของผู้ตรวจมีความแตกต่างกันค่อนข้างมาก และเมื่อเวลาเปลี่ยนไปลักษณะการประเมินยังไม่คงที่แต่มีแนวโน้มว่าจะมีความเข้มงวดและสอดคล้องกันมากขึ้น ส่วนในการวิเคราะห์มาตรประเมินนั้นมีความเที่ยงตรงของลำดับขั้น แต่ผู้ตรวจบางคนยังไม่เข้าใจในช่วงปลายของมาตรประเมินทักษะการออกเสียงและการสื่อสาร

เทอร์เนอร์ (Turner. 2003: online) ได้ใช้โมเดลหลายองค์ประกอบของราล์ซในการตรวจสอบการวัดประเมินแฟ้มสะสมงาน (Portfolio) ทางศิลปะ เพื่อวัดความสามารถทางศิลปะของนักเรียน ระบบการวัดประเมินแฟ้มสะสมงานทางศิลปะ ประกอบด้วย 1) การกำหนดให้นักเรียนทำชิ้นงานจำนวน 12 ชิ้นตามวัตถุประสงค์ของหลักสูตร 2) ให้นักเรียนทั้ง 15 คน นำเสนองานเมื่อสิ้นสุดโครงการ 3) คณะกรรมการด้านความเที่ยงตรงทำการตรวจสอบความเข้มงวด/ใจดีของผู้ตรวจ 4) ผู้เชี่ยวชาญด้านการออกแบบกราฟฟิค 10 คนประเมินโครงการ และ 5) ใช้สถิติในการวิเคราะห์การให้คะแนนโครงการ ทั้งนี้ผู้เชี่ยวชาญด้านการออกแบบกราฟฟิคจะประเมินชิ้นงานศิลปะของนักเรียนโดยใช้กฎเกณฑ์การให้คะแนนที่ประกอบด้วยเกณฑ์การออกแบบชิ้นงาน ทำการวัดประเมินใน 4 มิติตามลักษณะเฉพาะของชิ้นงานและใช้มาตรประเมิน 5 ระดับ สุ่มผู้ตรวจจำนวน 3 คนในการตรวจแฟ้มสะสมงานของนักเรียนแต่ละคน จากนั้นใช้โมเดลหลายองค์ประกอบของราล์ซ(MFRM) วิเคราะห์ความสามารถของนักเรียน ความยากของชิ้นงาน ความเข้มงวดของผู้ตรวจและมิติของความยากในการประเมิน ผลการศึกษาพบว่า มีนักเรียน 3 คนที่ประมาณค่าความสามารถทางศิลปะเกินจริง (Over estimated) และอีก 3 คนที่ประมาณค่าต่ำกว่าความเป็นจริง (Under estimated) บนฐานของค่าเฉลี่ยของคะแนนดิบ ทั้งนี้ MFRM ได้ปรับแก้ค่าความสามารถของนักเรียนจากความเข้มงวด/ใจดีของผู้ตรวจที่ต่างกัน ทำให้มีการเปลี่ยนลำดับของนักเรียน 6 คน โดยมีค่าความเชื่อมั่นแยกส่วนของผู้ตรวจเป็น 0.93 ซึ่งมีนัยสำคัญของความเข้มงวด/ใจดีของผู้ตรวจ ค่าความเชื่อมั่นแยกส่วนของมิติการประเมินเป็น 0.73 แสดงว่าผู้ตรวจไม่สามารถจำแนกความแตกต่างมิติของการประเมินได้ ผู้ตรวจเลือกรายการการประเมินเดียวกัน 43 % ความยากของชิ้นงานและมิติของความยากของการประเมินไม่มีผลต่อการจัดอันดับความสามารถของนักเรียน ส่วนความเที่ยงตรงเชิงเนื้อหาพิจารณาจากลักษณะทั้งหมดของ

ระบบการประเมินแฟ้มสะสมงานทางศิลป์ ความเที่ยงตรงเชิงโครงสร้างจะตรวจสอบจากการจัดลำดับของความยากของข้อสอบ ส่วนความเที่ยงตรงภายในนั้นพิจารณาจาก องค์ประกอบที่ให้พยานหลักฐานว่าข้อมูลนั้นสอดคล้องกับ MFRM เทอร์เนอร์ได้กล่าวสรุปว่า ระบบการวัดประเมินแฟ้มสะสมงานทางศิลป์ที่วิเคราะห์ด้วย MFRM จะให้ค่าสถิติที่เป็นค่าการวัดความสามารถทางศิลป์ของนักเรียนที่ปลอดจากความเห็นส่วนตัวหรือความเป็นอัตนัย (Subjective) ของผู้ตรวจ

ปาร์ค (Park. 2004: online) ได้ใช้โมเดลหลายองค์ประกอบของราล์ซในการตรวจสอบอิทธิพลของแหล่งความแปรปรวน 4 แหล่ง คือ ผู้สอบ ผู้ตรวจ ข้อสอบและมาตรฐานประเมิน ซึ่งทั้ง 4 แหล่งส่งผลถึงการปฏิบัติของผู้สอบในการทดสอบ ทั้งนี้เพื่อจัดลำดับผู้สอบจากการเขียนภาษาอังกฤษในฐานะภาษาที่สอง (English as a Second Language: ESL) (ESL Placement Test of Writing) พบว่า ผู้สอบมีความแปรปรวนของความสามารถของค่อนข้างกว้าง คือ ตั้งแต่ -8 โลจิทถึง 7 โลจิท ผู้สอบส่วนใหญ่ได้คะแนนสูง นั่นคือ ความสามารถของผู้สอบมีการแจกแจงเบ้ทางลบ และมี 4 % ของผู้สอบที่ไม่สอดคล้อง (Misfit) กับโมเดล จึงต้องมีการทบทวนการประเมินอีกครั้งก่อนการรายงานคะแนน โดยเฉพาะผู้สอบที่อยู่ใกล้เขตวิกฤตของการตัดสินใจ ความเชื่อมั่นแยกส่วนของผู้ตรวจมีค่าเป็น 0.97 แสดงว่าผู้ตรวจมีระดับความเข้มงวดที่แตกต่างกัน ค่าความสอดคล้อง (Fit value) ของผู้ตรวจทุกคนอยู่ในช่วงที่ยอมรับได้ ส่วนค่าความยากของข้อสอบทั้ง 4 ข้อมีค่าใกล้เคียง 0 ค่าความเชื่อมั่นแยกส่วนของข้อสอบมีค่าเป็น 0.64 แสดงว่าข้อสอบทั้ง 4 ข้อมีค่าความยากใกล้เคียงกัน และค่าความสอดคล้อง (Fit value) อยู่ในช่วงที่ยอมรับได้ และมาตรฐานประเมินทั้ง 4 ระดับมีความเหมาะสมในการจัดลำดับและมีความชัดเจนในการจำแนกแต่ละระดับ เพราะผู้สอบมีค่าเฉลี่ยของคะแนนสูงขึ้นตามน้ำหนักของรายการประเมินที่เพิ่มขึ้นในข้อสอบทุกข้อ

บัสเตอร์ก (Basturk. 2008) ได้ประยุกต์ใช้โมเดลหลายองค์ประกอบของราล์ซในการประเมินการนำเสนองานโดยใช้ Power Point ของนักศึกษาปีที่ 1 วิชาเอกวิทยาศาสตร์ ในช่วงฤดูใบไม้ร่วงปี 2005 – 2006 ประเทศตุรกี ในรายวิชา เทคนิคการสอนเบื้องต้น โดยการแบ่งนักศึกษา 30 คน เป็น 6 กลุ่มย่อย และแต่ละกลุ่มย่อยจะได้ประเด็นที่เกี่ยวข้องกับเนื้อหาวิชาที่เรียนและให้นำเสนอโดยใช้ Power Point ใช้ผู้ประเมิน 7 คน คือ ผู้สอนและนักศึกษาที่สุ่มในแต่ละกลุ่มอีก 6 คน พบว่า นักศึกษาผู้ประเมิน และประเด็นที่นำเสนอมีความสอดคล้องกับโมเดล การวัดประเมินผู้เรียนควรมีการประเมินระหว่างเรียน การใช้ Power Point ในการนำเสนอทำให้นักศึกษาพัฒนาความรู้และความสามารถในการสื่อสารเพิ่มขึ้น และการใช้โมเดลหลายองค์ประกอบของราล์ซ มีค่าความเชื่อมั่นของนักศึกษา ผู้ประเมิน และประเด็นนำเสนอ เป็น 0.86, 0.88 และ 0.82 ตามลำดับ แสดงว่า มีความยุติธรรมในการจำแนกนักศึกษา ผู้ประเมิน และข้อสอบ ให้แตกต่างกันได้ตามระดับคุณภาพ

สำหรับในประเทศไทยนั้นยังมีการศึกษาวิจัยการประยุกต์ใช้โมเดลหลายองค์ประกอบของราล์ชค่อนข้างน้อยและเอกสารภาษาไทยสำหรับการศึกษาโมเดลนี้ก็มีอยู่น้อยเช่นกัน จากการค้นคว้าพบว่า ในปี 2543 สุภมาส อังศุโชติ (2543) ได้ประยุกต์ใช้โมเดลหลายองค์ประกอบของราล์ชในการเปรียบเทียบคะแนนและเปรียบเทียบกับวิธีการเปรียบเทียบคะแนนอีก 4 วิธี โดยวิเคราะห์ผลการปรับคะแนนเฉลี่ยสะสมของนักเรียนในระดับมัธยมศึกษาตอนปลายด้วยวิธีปรับคะแนนเฉลี่ยสะสม 5 วิธี คือ วิธีปรับคะแนนเฉลี่ยสะสมเชิงเส้นตรงตาม design IV C-2 ของ Angoff (ANGOFF-4C2) วิธีปรับคะแนนเฉลี่ยสะสมโดยใช้ทฤษฎีการตอบข้อสอบภายใต้ Graded Response Model (IRT-GRM) วิธีปรับคะแนนเฉลี่ยสะสมด้วยการวิเคราะห์ตัวแบบเชิงเส้นทั่วไป (GLM-MODEL) วิธีปรับคะแนนเฉลี่ยสะสมด้วยการวิเคราะห์องค์ประกอบเชิงยืนยันโมเดลการวัดคอนเจนเนอริค 1 องค์ประกอบ (CON-CFA) และวิธีปรับคะแนนเฉลี่ยสะสมด้วยโมเดลหลายองค์ประกอบของราล์ช (RASCH-FACET) และเปรียบเทียบความตรงเชิงทำนายระหว่างวิธีปรับคะแนนเฉลี่ยสะสมทั้ง 5 วิธี โดยใช้คะแนนเฉลี่ยสะสมระดับมหาวิทยาลัยปี 1 และปี 2 เป็นเกณฑ์

ผลการวิจัยพบว่า เมื่อใช้คะแนนเฉลี่ยสะสมระดับมหาวิทยาลัยปี 1 และปี 2 ที่ปรับโดยใช้สมการถดถอย และที่ปรับด้วยโมเดลหลายองค์ประกอบของราล์ชเป็นเกณฑ์ให้ผลสอดคล้องกัน คือ ที่ระดับนัยสำคัญ .05 คะแนนเฉลี่ยสะสมระดับมัธยมศึกษาตอนปลายที่ปรับด้วยวิธีปรับ 3 วิธี คือ วิธี RASCH-FACET วิธี ANGOFF-4C2 และวิธี IRT-GRM ให้ความตรงเชิงทำนายสูงกว่าคะแนนเฉลี่ยสะสมระดับมัธยมศึกษาตอนปลายเดิม และวิธี CON-CFA ให้ความตรงเชิงทำนายต่ำกว่าคะแนนเฉลี่ยสะสมระดับมัธยมศึกษาตอนปลายเดิม ผลที่ต่างกันคือเมื่อใช้คะแนนเฉลี่ยสะสมระดับมหาวิทยาลัยปี 1 และปี 2 ที่ปรับโดยใช้สมการถดถอยเป็นเกณฑ์ วิธี GLM-MODEL ให้ความตรงเชิงทำนายสูงกว่าคะแนนเฉลี่ยสะสมระดับมัธยมศึกษาตอนปลายเดิมแต่จะไม่แตกต่างกันเมื่อใช้คะแนนเฉลี่ยสะสมระดับมหาวิทยาลัยปี 1 และปี 2 ที่ปรับด้วยโมเดลหลายองค์ประกอบของราล์ชเป็นเกณฑ์ ส่วนการเปรียบเทียบความตรงเชิงทำนายระหว่างวิธีพบว่า วิธี RASCH-FACET ให้ผลดีที่สุด รองลงมาคือวิธี ANGOFF-4C2 และวิธี IRT-GRM ซึ่งสองวิธีนี้ให้ความตรงเชิงทำนายไม่แตกต่างกัน อันดับที่ดีที่สุดคือ วิธี GLM-MODEL และอันดับสุดท้ายคือวิธี CON-CFA

ชาญวิทย์ จรัสสุทธิอิศวร (2550) ได้ประยุกต์ใช้โมเดลหลายองค์ประกอบของราล์ชในการแสดงหลักฐานความเที่ยงตรงเชิงโครงสร้างของคะแนนการประเมินทักษะกระบวนการทางวิทยาศาสตร์ โดยมีวัตถุประสงค์การวิจัยเพื่อแสดงหลักฐานความเที่ยงตรงเชิงโครงสร้างของการให้คะแนนทักษะกระบวนการทางวิทยาศาสตร์ขั้นบูรณาการ กลุ่มตัวอย่างที่ใช้ คือ นักเรียนชั้นมัธยมศึกษาปีที่ 4 สังกัดเขตพื้นที่การศึกษานครราชสีมา เขต 3 จำนวน 280 คน ที่สุ่มแบบสองขั้นตอน ตรวจให้คะแนนโดยผู้ตรวจจำนวน 14 คน ภายใต้เงื่อนไขการจับคู่ตรวจให้คะแนน การ

แสดงหลักฐานความเที่ยงตรงเชิงโครงสร้างของการให้คะแนนโดยการวิเคราะห์ด้วยโมเดลหลายองค์ประกอบของราล์ชที่มี 3 องค์ประกอบ ตามแนวคิดของทฤษฎีการตอบข้อสอบ พบว่าองค์ประกอบทั้งสามให้ค่าพารามิเตอร์อยู่ในเกณฑ์ที่ยอมรับได้ และจากการวิเคราะห์ค่าสถิติความสอดคล้องกับโมเดล พบว่า โมเดลการวิจัยมีความสอดคล้องกลมกลืนกับข้อมูลเชิงประจักษ์ทั้งสามองค์ประกอบ

จากผลการศึกษาทางนิเวศวิทยาข้างต้น สรุปได้ว่า มีการนำโมเดลหลายองค์ประกอบของราล์ชไปใช้อย่างกว้างขวางในหลายลักษณะ เช่น การวัดประเมินทางด้านภาษา ทางด้านศิลปะ เป็นต้น การวัดประเมินเหล่านี้ต่างมีแหล่งความคลาดเคลื่อนของคะแนนหลายแหล่ง และมีปัญหาในแง่ของความน่าเชื่อถือของคะแนนอันเนื่องมาจากผู้ตรวจให้คะแนน ผลการวิจัยต่างพบในประเด็นที่สอดคล้องกันว่าผู้ตรวจเป็นแหล่งความคลาดเคลื่อนที่สำคัญ และผู้ตรวจมีความเข้มงวดที่แตกต่างกัน ส่งผลต่อความน่าเชื่อถือและความเที่ยงตรงของคะแนน ทั้งนี้เมื่อใช้โมเดลหลายองค์ประกอบของราล์ชในการวิเคราะห์ จะสามารถประมาณค่าพารามิเตอร์ของแต่ละองค์ประกอบ เช่น ผู้สอบ ข้อสอบ และผู้ตรวจ ได้อย่างเป็นอิสระจากกัน สามารถศึกษาแหล่งความแปรปรวนขององค์ประกอบว่าแหล่งใดเป็นแหล่งของความคลาดเคลื่อนหลัก และประมาณค่าความเชื่อมั่นแยกส่วนในแต่ละองค์ประกอบได้โดยไม่จำเป็นที่ผู้ตรวจต้องตรวจให้คะแนนข้อสอบทุกข้อของผู้สอบทุกคน นอกจากนี้โมเดลหลายองค์ประกอบของราล์ชจะปรับแก้คะแนนอันเกิดจากความแตกต่างกันของผู้ตรวจและข้อสอบ ทำให้ได้คะแนนที่มีความน่าเชื่อถือ เที่ยงตรงและยุติธรรม

งานวิจัยที่เกี่ยวข้องกับทฤษฎีการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราล์ช

มีงานวิจัยที่ประยุกต์ใช้โมเดลการสรุปอ้างอิงของทฤษฎีการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราล์ชของทฤษฎีการตอบข้อสอบร่วมกัน เพื่อวิเคราะห์การวัดประเมินที่มีการให้คะแนนแบบหลายค่า และเกี่ยวข้องกับผู้ตรวจอย่างกว้างขวาง ทั้งในการวัดประเมินทางภาษา เช่น การพูด การเขียน การวัดประเมินทักษะการแก้ปัญหา ตลอดจนการวัดประเมินทางด้านการแพทย์ ซึ่งงานวิจัยส่วนใหญ่ยังคงเป็นงานวิจัยในต่างประเทศ

ในการวัดประเมินทางภาษานั้นในปี 1993 บาซแมน ไลนซ์ และมาสัน (Bachman; Lynch; & Mason.1993: online) ได้ศึกษาความแปรปรวนของข้อสอบและการตัดสินใจของผู้ตรวจในการทดสอบการพูดภาษาต่างประเทศ เพื่อนำผลไปพัฒนาระบบการวัดประเมินความสามารถทางภาษา (Language Ability Assessment System: LAAS) ให้มีความน่าเชื่อถือมากขึ้น ประมาณค่าความเชื่อมั่นโดยใช้ทฤษฎีการสรุปอ้างอิง (G-Theory) และโมเดลหลายองค์ประกอบของราล์ช (MFRM) ใช้ผู้สอบจำนวน 218 คน โดยให้อ่านบทความแล้วตอบคำถาม จากนั้นให้นำเสนอข้อสรุปโดยการสอบ

ปากเปล่า เมื่อวิเคราะห์ด้วยโมเดลการสรุปอ้างอิง พบว่า องค์ประกอบความแปรปรวนที่มีขนาดใหญ่ที่สุดคือ ผู้สอบ คิดเป็น 91 % แสดงว่า ความแปรปรวนของคะแนนเอกภพหรือคะแนนจริงมีความเชื่อถือได้ สัมพันธ์กับความแตกต่างของระดับความสามารถของผู้สอบ ส่วนผู้ตรวจและข้อสอบมีความแปรปรวนน้อย แสดงว่า ผู้ตรวจมีความเข้มงวดใกล้เคียงกัน และข้อสอบมีความยากเท่าๆ กัน ค่าสัมประสิทธิ์การสรุปอ้างอิงมีค่าเป็น 0.94 และเมื่อวิเคราะห์ด้วยโมเดลหลายองค์ประกอบของราล์ซ พบว่า ผู้ตรวจแต่ละคนและข้อสอบแต่ละข้อ ส่งผลต่อการประมาณค่าความสามารถทางภาษาของผู้สอบ และมีค่าความเชื่อมั่นแยกส่วนของผู้สอบเป็น 0.98 ความเชื่อมั่นแยกส่วนของผู้ตรวจเป็น 0.92 แสดงว่า ช่วงความเข้มงวดของผู้ตรวจมีขนาดใหญ่ เนื่องมาจากการให้คะแนนแบบสุดโต่ง (Extreme) ของผู้ตรวจ 2 คน แต่โดยเฉลี่ยแล้วความเข้มงวดของผู้ตรวจนั้นไม่มีความแตกต่างกัน และความเชื่อมั่นแยกส่วนของข้อสอบมีค่าเป็น 0.83 ส่วน ไลงซ์และแมคนามารา (Lynch; & McNamara. 1998: online) ได้ใช้ทฤษฎีการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราล์ซในการพัฒนาการวัดประเมินทักษะการพูดภาษาอังกฤษในฐานะภาษาที่สองของผู้ที่จะเดินทางไปประเทศออสเตรเลีย วิเคราะห์ด้วยโปรแกรม GENOVA (Crick; & Brennan. 1984) และ FACETS (Linacre; & Wright. 1993) พบว่า การใช้ทั้งสองโมเดลเป็นการขยายขอบเขตของการสนับสนุนซึ่งกันและกัน ทำให้ทราบช่วงขององค์ประกอบต่าง ๆ ที่มีอิทธิพลต่อความสำเร็จของผู้สอบ

สมิท และคูลิกอวิช (Smith; & Kulikowich. 2004) ได้นำทฤษฎีการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราล์ซไปใช้ในการวัดประเมินทักษะการแก้ปัญหาที่ซับซ้อน เพื่อศึกษาทฤษฎีทางสถิติที่ตอบสนองความต้องการของผู้วิจัย ครู นักจิตวิทยาคลินิกและการให้คำปรึกษา ในการวินิจฉัยและการตัดสินใจจากสารสนเทศที่ได้และพัฒนาเครื่องมือวัดโครงสร้างทางพุทธิพิสัย (Cognitive) คือ ทักษะกระบวนการแก้ปัญหาที่ซับซ้อน โดยวัดประเมินจากการปฏิบัติ (Performance – based assessment) ทั้งรายบุคคลและกลุ่ม เพื่อดูการวางแผนและการตัดสินใจจากสถานการณ์ที่กำหนดให้เมื่ออยู่นอกบริบทห้องเรียน กลุ่มตัวอย่างเป็นนักเรียนเกรด 4 จำนวน 44 คน ให้ข้อสอบจำนวน 5 ข้อ ผู้ตรวจจำนวน 2 คน และทดสอบ 2 ครั้ง สถานการณ์ที่กำหนดเป็นกิจกรรมคิกบอล (Kickball) โดยให้ผู้สอบจัดทีมที่เหมาะสมภายใต้ข้อมูลที่กำหนดให้ จากนั้นให้นำเสนองาน ผู้ตรวจจะตั้งคำถามในด้าน การสังเกต การเปรียบเทียบความเหมือน ความต่าง การจำแนกและการวางแผน เมื่อวิเคราะห์ข้อมูลด้วยโมเดลการสรุปอ้างอิง พบว่า แหล่งของความแปรปรวนที่มากที่สุดคือ ผู้สอบ (0.325) แสดงว่าผู้สอบมีความแตกต่างทางด้านความคิดซับซ้อน รองลงไปคือ ข้อสอบ (0.257) ส่วนความแปรปรวนของผู้ตรวจและจำนวนครั้งของการสอบมีน้อย (0.001 และ 0.004 ตามลำดับ) นั่นคือไม่มีความแตกต่างของผู้ตรวจ มีความแตกต่างเพียงเล็กน้อยในการสอบ 2 ครั้ง ในการศึกษา D พบว่า หากใช้ผู้ตรวจ 2 คนและตรวจ 2 ครั้ง จะมีค่าสัมประสิทธิ์การสรุปอ้างอิงเป็น 0.8012 และค่า

สัมประสิทธิ์ความเชื่อถือได้เป็น 0.6977 ส่วนการวิเคราะห์ด้วย MFRM พบว่า ความเชื่อมั่นแยกส่วนของผู้สอบมีค่าเป็น 0.90 มีค่าไค-สแควร์เป็น 371.2 มี df เป็น 43 และมีนัยสำคัญทางสถิติ ($p < .001$) แสดงว่าผู้สอบมีความแตกต่างกันทางด้านความคิดซับซ้อน ความเชื่อมั่นแยกส่วนของผู้ตรวจมีค่าเป็น 0.00 มีค่าไค-สแควร์เป็น 0.2 มี df เป็น 1 และไม่มีนัยสำคัญทางสถิติ ($p > .01$) แสดงว่าผู้ตรวจมีระดับความเข้มงวดคล้ายคลึงกัน ความเชื่อมั่นแยกส่วนของจำนวนครั้งการสอบ มีค่าเป็น 0.00 มีค่าไค-สแควร์เป็น 0.5 มี df เป็น 1 และไม่มีนัยสำคัญทางสถิติ ($p > .01$) แสดงว่าจำนวนครั้งของการสอบไม่มีผลต่อการประมาณค่า ความเชื่อมั่นแยกส่วนของข้อสอบ มีค่าเป็น 0.98 มีค่าไค-สแควร์เป็น 234.5 มี df เป็น 4 และมีนัยสำคัญทางสถิติ ($p < .001$) แสดงว่าความยากของข้อสอบนั้นมีการกระจายอยู่บนตลอดช่วงความสามารถ (Continuum) โดยการวางแผนมีค่าความยากมากที่สุด คือ มีคะแนนโลจิทเป็น 0.80 ส่วนอีก 3 ข้อ นั้นมีค่าความยากอยู่รอบ ๆ 0 (คะแนนโลจิท) ข้อสอบที่ง่ายที่สุด คือ การจำแนก ซึ่งมีค่าความยากของคะแนนโลจิทเป็น -1.0

สัปดาห์ รีฟ และแบรดชอว์ (Sudweeks; Reeve; & Bradshaw. 2005) ได้เปรียบเทียบการใช้ทฤษฎีการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราล์ฟในการวิเคราะห์งานเขียนในระดับมหาวิทยาลัย โดยมีวัตถุประสงค์เพื่อประมาณค่าแหล่งความคลาดเคลื่อนของการประเมิน ประมาณค่าความเชื่อมั่น และเพื่อให้ข้อเสนอแนะในการพัฒนากระบวนการประเมิน ใช้กลุ่มตัวอย่าง 24 คน จากนักศึกษาทั้งหมด 497 คน ที่ลงทะเบียนเรียนวิชาประวัติศาสตร์โลกในภาคเรียนฤดูร้อน ปี 2002 โดยมอบหมายให้เขียนเรียงความจำนวน 2 บทความในหัวข้อที่เกี่ยวข้องกับประวัติศาสตร์โลก ผู้ตรวจคือครูผู้สอนในภาควิชาภาษาอังกฤษจำนวน 9 คน ใช้มาตราประเมิน 9 ระดับ ใช้กฎเกณฑ์การให้คะแนนแบบรวม (Holistic rating scale) ใช้รูปแบบแบบไขว้สมบูรณ์ (Full crossed design) และแบบแฝง (Nested design) โดยศึกษาใน 4 องค์ประกอบคือ ผู้สอบ ข้อสอบ ผู้ตรวจและจำนวนครั้งในการตรวจ มีการอบรมผู้ตรวจก่อนการให้คะแนนจำนวน 2 ชั่วโมง และในการตรวจจะไม่ให้ผู้ตรวจทราบชื่อผู้สอบ ผลการวิจัยพบว่า ทั้งสองทฤษฎีต่างให้ผลการวิเคราะห์ที่สอดคล้องกัน โดยในทฤษฎีการสรุปอ้างอิง พบว่า ความแปรปรวนที่เกิดจากข้อสอบและปฏิสัมพันธ์ระหว่างผู้สอบกับข้อสอบนั้นมีค่าสูงและแตกต่างอย่างมีนัยสำคัญ ในขณะที่ความแปรปรวนของผู้ตรวจและจำนวนครั้งของการตรวจนั้นมีค่าต่ำ และมีความแปรปรวนที่ไม่สามารถอธิบายได้ประมาณ 22 % ค่าความเชื่อมั่นเมื่อใช้งานเขียน 2 ชิ้น ผู้ตรวจ 9 คนและตรวจ 2 ครั้งโดยใช้รูปแบบแบบไขว้สมบูรณ์มีค่าเป็น 0.75 ซึ่งมีค่าค่อนข้างต่ำ เนื่องจากมีแหล่งของความคลาดเคลื่อนหลายแหล่ง แต่ช่วงความเชื่อมั่นนั้นเป็นที่น่าพอใจในการอ้างอิงจากคะแนนสังเกตไปยังคะแนนเอกภพ ค่าสัมประสิทธิ์ความเชื่อมั่นมีค่าสูงขึ้นหากคำนวณจากค่าเฉลี่ยของการประเมินนักเรียนโดยใช้ข้อสอบอย่างน้อย 3 ข้อ จากผู้ตรวจอย่างน้อย 2 คน การใช้ข้อสอบหลายข้อและผู้ตรวจหลายคนจะให้ค่าความเชื่อมั่นเพิ่มสูงขึ้นมากกว่าการตรวจหลายครั้ง

การใช้แบบแผนแบบแฝงโดยให้ผู้ตรวจแต่ละคนตรวจข้อสอบทุกข้อของนักเรียนที่แตกต่างกันนั้นค่าความเชื่อมั่นแตกต่างจากรูปแบบแบบไขว้สมบูรณเพียง 1.5 – 2.0 % เท่านั้น ในขณะที่การวิเคราะห์ด้วยโมเดลหลายองค์ประกอบของราล์ช พบว่า ดัชนีแยก (Separation index) และความเชื่อมั่นแยกส่วน (Separation reliability) ของข้อสอบมีค่าสูงและแตกต่างกันอย่างมีนัยสำคัญของความยากของข้อสอบ และปฏิสัมพันธ์แบบสองทางระหว่างผู้สอบกับข้อสอบมีเปอร์เซ็นต์สูงสุดอย่างมีนัยสำคัญ ความแปรปรวนของผู้ตรวจและจำนวนครั้งของการตรวจนั้นมีค่าต่ำ ความเชื่อมั่นแยกส่วนของผู้สอบผู้ตรวจ และข้อสอบมีค่าสูง แต่ความเชื่อมั่นแยกส่วนของจำนวนครั้งของการตรวจมีค่าต่ำ แสดงว่ามีความสอดคล้องกันของผู้ตรวจจากการตรวจเพียงครั้งเดียว

นอกจากนี้ยังมีการนำโมเดลการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราล์ชไปใช้ในการทดสอบทางการแพทย์ โดย ไอรามณีรัตน์และคณะ (Iramaneerat; et al. 2007: online) ได้นำโมเดลทั้งสองไปประยุกต์ใช้ในการทดสอบในวงการแพทย์ ทั้งนี้เพื่อควบคุมคุณภาพของการทดสอบอิงโครงสร้างวัตถุประสงค์ทางคลินิก (An Objective Structured Clinical Examination: OSCE) ซึ่งคะแนนที่ได้จาก OSCE ยังคงมีความคลาดเคลื่อนของการวัดค่อนข้างมาก ทั้งจาก คนไข้ (ผู้ประเมิน) ข้อสอบ หรือแพทย์ผู้สอบ ซึ่งการให้ความสนใจกับแหล่งความคลาดเคลื่อนเหล่านี้มีความสำคัญมากในกลไกการควบคุมคุณภาพเพื่อความมั่นใจในความเที่ยงตรงของการแปลความคะแนนที่ได้ งานวิจัยนี้ใช้การทดสอบทักษะการสื่อสาร OSCE ของแพทย์จำนวน 79 คนจาก Midwestern University ในสหรัฐอเมริกา แพทย์แต่ละคนจะสนทนากับคนไข้ใน 6 ประเด็น และให้คนไข้ประเมินโดยใช้มาตรประเมิน 5 ระดับ จากการศึกษาโดยใช้ทฤษฎีการสรุปอ้างอิงแสดงให้เห็นว่าแหล่งของความแปรปรวนของความคลาดเคลื่อนที่ใหญ่ที่สุดรองมาจากความคลาดเคลื่อนของส่วนที่เหลือคือ ความแปรปรวนของคนไข้ (ผู้ประเมิน) แสดงว่าคนไข้ (ผู้ประเมิน) มีความเข้มงวดต่างกัน และเมื่อศึกษาโดยใช้โมเดลหลายองค์ประกอบของราล์ช ที่มองว่าคนไข้และข้อสอบเป็นความคลาดเคลื่อนของการวัด พบว่ามีความแตกต่างอย่างมีนัยสำคัญของคนไข้ (ผู้ประเมิน) ในระดับของความเข้มงวด คนไข้สองคนมีการประเมินที่ไม่สอดคล้องกัน แสดงถึง ปัญหาด้านการสื่อสารของนักเรียนแพทย์ ความเข้าใจในมาตรประเมินหรือเนื้อหาที่ทดสอบ มาตรประเมินของข้อสอบ 2 ข้อไม่ได้เป็นไปตามฟังก์ชันของมาตรประเมิน ดังนั้น การใช้ทั้งสองโมเดลในการวิเคราะห์คะแนนจาก OSCE นั้นต่างให้สารสนเทศที่เป็นประโยชน์เติมเต็มซึ่งกันและกัน ทำให้การตรวจสอบและพัฒนาคุณภาพของ OSCE สมบูรณ์มากขึ้น

ส่วนงานวิจัยที่เกี่ยวข้องกับการนำโมเดลการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราล์ชไปใช้ร่วมกันในประเทศไทยนั้น จากการศึกษาค้นคว้า พบว่ายังไม่มีการทำวิจัยหรือประยุกต์ใช้ทั้งสองโมเดลนี้ร่วมกันแต่อย่างใด

จากผลการวิจัยที่มีการนำโมเดลการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราล์ชไปประยุกต์ใช้ร่วมกัน (Lynch; & McNamara. 1998: online; Smith; & Kulikowich. 2004; Sudweeks; Reeve; & Bradshaw. 2005; Iramaneerat; et al. 2007: online) สรุปได้ว่า ทั้งสองโมเดลให้ผลการวิเคราะห์ที่สอดคล้องกัน โดยสามารถแสดงให้เห็นถึงขนาดของแหล่งความคลาดเคลื่อนขององค์ประกอบต่าง ๆ ในการวัดประเมิน อีกทั้งยังสามารถประมาณค่าความเชื่อมั่นในสถานการณ์ต่าง ๆ ของการวัดผลได้ ซึ่งผลการศึกษาแต่ละครั้งยังพบในประเด็นที่แตกต่างกัน คือ ความแปรปรวนของผู้ตรวจ ความเข้มงวดของผู้ตรวจ นั้นยังพบว่า มีขนาดแตกต่างกัน การเพิ่มจำนวนผู้ตรวจทำให้ค่าความเชื่อมั่นเพิ่มสูงขึ้น นอกจากนี้นักวิจัยยังได้ให้ข้อเสนอแนะในการศึกษารูปแบบการตรวจให้คะแนนรูปแบบอื่นนอกเหนือจากการตรวจข้อสอบทุกข้อของผู้สอบทุกคน เพื่อแก้ปัญหาในการตรวจให้คะแนนที่ต้องใช้เวลา แรงงานและค่าใช้จ่ายในการตรวจค่อนข้างมาก แต่ยังคงไว้ซึ่งความเชื่อมั่นและความเที่ยงตรงของคะแนน ทั้งนี้โมเดลการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราล์ชสามารถวิเคราะห์ผลการตรวจให้คะแนนโดยที่ผู้ตรวจไม่จำเป็นต้องตรวจข้อสอบทุกข้อของผู้สอบทุกคน โมเดลทั้งสองจึงเป็นโมเดลที่เติมเต็ม(Complementary)ซึ่งกันและกัน ทำให้ได้สารสนเทศที่สมบูรณ์มากยิ่งขึ้น (Bachman; Lynch; & Mason.1993: online; Lynch; & McNamara. 1998: online; Iramaneerat; et al. 2007: online)

นอกจากนี้ ยังมีนักวิจัยได้ประยุกต์ใช้ทฤษฎีการทดสอบแบบมาตรฐานเดิม (Classical Test Theory: CTT) และโมเดลหลายองค์ประกอบของราล์ช (Many Facet Rasch Measurement: MFRM) ร่วมกัน ดังเช่น กุลเลอร์และเกลบัล (Guler; & Gelbal. 2010) ได้ศึกษาการใช้ทฤษฎีการทดสอบแบบมาตรฐานเดิม และโมเดลหลายองค์ประกอบของราล์ช ในการประมาณค่าและเปรียบเทียบความเชื่อมั่นของคะแนนผลสัมฤทธิ์วิชาคณิตศาสตร์ ซึ่งเป็นแบบทดสอบแบบปลายเปิด (Open-ended questions) ผลการวิจัย พบว่า ค่าความสอดคล้องภายใน (Internal consistency) ของคะแนนเป็น 0.92 ค่าสัมประสิทธิ์ของเคนดอล (Kendall's concordance) เป็น 0.52 ค่าสัมประสิทธิ์สหสัมพันธ์ของผู้ตรวจ 4 คน มีค่า 0.90-0.97 ในขณะที่ผลการวิเคราะห์ด้วย MFRM ได้ค่าความเชื่อมั่นขององค์ประกอบผู้สอบเป็น 0.95 ค่าความเชื่อมั่นขององค์ประกอบผู้ตรวจเป็น 0.99 แสดงว่า ความเชื่อมั่นของคะแนนวิชาคณิตศาสตร์มีค่าสูง แม้ว่าคะแนนเฉลี่ยของผู้ตรวจทั้ง 4 คน จะมีค่าแตกต่างกัน ดังนั้น การเลือกทฤษฎีในการประมาณค่าความเชื่อมั่นของคะแนน ขึ้นอยู่กับจุดประสงค์ของการนำคะแนนไปใช้ หากต้องการหาความเชื่อมั่นของการวัดควรจะใช้ทฤษฎีในการประมาณค่าอย่างน้อย 2 ทฤษฎี ในการวิเคราะห์หาความเชื่อมั่น เพื่อให้มีความน่าเชื่อถือมากขึ้น

ไหหยาง (Haiyang. 2010) ได้ประยุกต์ใช้ทฤษฎีการทดสอบแบบมาตรฐานเดิม (CTT) และโมเดลหลายองค์ประกอบของราล์ช (MFRM) ในการวิเคราะห์ความเชื่อมั่นของแบบทดสอบวิชา

ภาษาอังกฤษสำหรับบัณฑิตที่ไม่ใช่วิชาเอกภาษาอังกฤษ โดยใช้แบบทดสอบอัตนัย 3 ข้อ คือ การแปลภาษาจีนเป็นอังกฤษ การแปลอังกฤษเป็นภาษาจีนและการเขียน ใช้ผู้ตรวจ 2 คน ผู้สอบ 56 คน พบว่าความเชื่อมั่นของ CTT ของคะแนนผู้สอบจากแบบทดสอบปรนัยมีความสัมพันธ์อย่างไม่มีนัยสำคัญกับคะแนนที่ได้แบบทดสอบอัตนัย ค่าความสอดคล้องภายในของข้อสอบทั้ง 3 ข้อยังมีค่าต่ำ และผลจากการวิเคราะห์ด้วย MFRM พบว่า องค์ประกอบผู้สอบมีค่าความเชื่อมั่นเป็น 0.77 ความเชื่อมั่นของข้อสอบเป็น 0.97 ส่วนความเชื่อมั่นของผู้ตรวจเป็น 0.91 แสดงว่า ผู้ตรวจทั้งสองคนมีความเข้มงวดในการให้คะแนนแตกต่างกัน ความยากของข้อสอบแตกต่างกัน มีความลำเอียงของปฏิสัมพันธ์ของผู้สอบบางคนกับข้อสอบที่ได้รับ ทำให้เกิดความแปรปรวนของคะแนนสูงสุด แสดงให้เห็นถึงความจำเป็นของการอบรมผู้ตรวจที่ไม่เฉพาะให้มีความสอดคล้องภายในตัวผู้ตรวจเองเท่านั้น แต่ยังคงมีความสอดคล้องกับผู้ตรวจคนอื่นด้วย และควรมีการศึกษาทฤษฎีการวัดอย่างเป็นระบบ เช่น การสร้างข้อสอบที่เป็นการเขียน การออกแบบข้อสอบ ให้เป็นส่วนหนึ่งของครูผู้สอน

กล่าวโดยสรุปได้ว่าการวัดประเมินทางด้านภาษา ศิลปะ การแก้ปัญหาที่ซับซ้อน การวัดประเมินทางการแพทย์ หรือแบบทดสอบปลายเปิด นั้นเป็นการวัดประเมินที่มีแหล่งความคลาดเคลื่อนของคะแนนหลายแหล่ง โดยเฉพาะจากผู้ตรวจ ซึ่งส่งผลต่อความเชื่อมั่นและความเที่ยงตรงของคะแนน ผลการวิจัยต่างพบในประเด็นที่สอดคล้องกันว่า ผู้ตรวจมีความเข้มงวดที่แตกต่างกัน ผู้ตรวจเป็นแหล่งความคลาดเคลื่อนที่สำคัญและส่งผลต่อคะแนนของผู้สอบ การเพิ่มจำนวนผู้ตรวจจะทำให้ความเชื่อมั่นเพิ่มสูงขึ้น แต่จำนวนผู้ตรวจที่เหมาะสมผลการศึกษายังพบแตกต่างกัน นอกจากนี้แล้วในการตรวจให้คะแนนนั้นต้องใช้เวลาค่อนข้างมาก จึงมีข้อเสนอแนะให้ทำการศึกษาการใช้รูปแบบการตรวจให้คะแนนแบบต่าง ๆ เพื่อลดเวลาและภาระในการตรวจ แต่ยังคงไว้ซึ่งความน่าเชื่อถือและความเที่ยงตรงของคะแนน ทั้งนี้วิธีการทางสถิติที่สามารถศึกษา วิเคราะห์และตรวจสอบแหล่งความคลาดเคลื่อนได้หลายแหล่งจากการวิเคราะห์เพียงครั้งเดียว และสามารถประมาณค่าความเชื่อมั่น เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกันได้ คือ การวิเคราะห์ด้วยโมเดลการสรุปอ้างอิงตามทฤษฎีการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราล์ชตามทฤษฎีการตอบข้อสอบ ซึ่งโมเดลทั้งสองเป็นโมเดลที่เติมเต็มซึ่งกันและกัน ทำให้ได้สารสนเทศที่สมบูรณ์มากยิ่งขึ้น

จากที่กล่าวมาแล้วข้างต้น ผู้วิจัยจึงสนใจศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจการให้คะแนนต่างกัน โดยใช้โมเดลการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราล์ช ซึ่งใช้โมเดลการสรุปอ้างอิงในการศึกษาองค์ประกอบความแปรปรวนจากแหล่งต่าง ๆ ได้แก่ ผู้สอบ ข้อสอบ และผู้ตรวจ ประมาณค่าและเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิง เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน และใช้โมเดลหลายองค์ประกอบของราล์ช ในการศึกษาองค์ประกอบความแปรปรวนของแต่ละแหล่ง

ประมาณค่าและเปรียบเทียบความเชื่อมั่นแยกส่วนในแต่ละองค์ประกอบ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน นอกจากนี้ยังได้ศึกษาความเที่ยงตรงตามสภาพของคะแนนเมื่อจำนวนผู้ตรวจและรูปแบบการตรวจต่างกัน โดยการนำคะแนนดิบและคะแนนความสามารถไปหาค่าสัมประสิทธิ์สหสัมพันธ์กับคะแนนที่ใช้เป็นเกณฑ์ ซึ่งในงานวิจัยนี้ใช้ ผลการทดสอบทางการศึกษาระดับชาติด้านพื้นฐาน (Ordinary National Education Test: O-NET) ทั้งนี้เพื่อตรวจสอบคุณลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ ได้แก่ ค่าสัมประสิทธิ์การสรุปอ้างอิง ค่าความเชื่อมั่นแยกส่วน และค่าความเที่ยงตรงตามสภาพ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน ซึ่งจะเป็นแนวทางในการเลือกใช้จำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนแบบทดสอบปลายเปิดที่สามารถลดภาระ เวลาและค่าใช้จ่าย โดยที่ยังมีความเชื่อมั่นและความเที่ยงตรงของคะแนนที่ยอมรับได้ในทางสถิติต่อไป



บทที่ 3

วิธีดำเนินการวิจัย

การศึกษาวิจัยในครั้งนี้มีจุดมุ่งหมายเพื่อศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน โดยใช้โมเดลการสรุปอ้างอิงตามทฤษฎีการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราล์ฟตามทฤษฎีการตอบข้อสอบ ซึ่งมีขั้นตอนการดำเนินการวิจัยดังนี้

1. การกำหนดประชากรและการเลือกกลุ่มตัวอย่าง
2. การสร้างเครื่องมือที่ใช้ในการวิจัย
3. การเก็บรวบรวมข้อมูล
4. การจัดกระทำและการวิเคราะห์ข้อมูล

การกำหนดประชากรและการเลือกกลุ่มตัวอย่าง

ประชากร

ประชากรที่ใช้ในการวิจัยครั้งนี้ คือ นักเรียนระดับชั้นมัธยมศึกษาปีที่ 4 สำนักงานเขตพื้นที่การศึกษาจังหวัดน่าน ปีการศึกษา 2552 จำนวน 3,476 คน

กลุ่มตัวอย่าง

กลุ่มตัวอย่างที่ใช้ในการวิจัยครั้งนี้ แบ่งเป็น 2 กลุ่ม ได้แก่

1. กลุ่มตัวอย่างที่ใช้ในการเก็บข้อมูล เป็นนักเรียนระดับชั้นมัธยมศึกษาปีที่ 4 สำนักงานเขตพื้นที่การศึกษาจังหวัดน่าน ปีการศึกษา 2552 จำนวน 180 คน จาก 6 โรงเรียน ซึ่งได้มาจากการสุ่มแบบสองขั้นตอน (Two-stage Random Sampling) มีขั้นตอนการสุ่มดังนี้

ขั้นที่ 1 จำแนกโรงเรียนที่เปิดสอนในระดับช่วงชั้นที่ 3 จำนวน 31 โรงเรียนออกเป็น 4 ขนาด คือ ขนาดเล็ก ขนาดกลาง ขนาดใหญ่และขนาดใหญ่พิเศษ พบว่า มีโรงเรียนขนาดเล็กจำนวน 1 โรงเรียน ขนาดกลาง 21 โรงเรียน ขนาดใหญ่ 4 โรงเรียน และขนาดใหญ่พิเศษจำนวน 5 โรงเรียน เพื่อจัดทำกรอบการสุ่มโดยอาศัยการแบ่งตามขนาดโรงเรียน

ขั้นที่ 2 สุ่มจำนวนโรงเรียนร้อยละ 20 ของจำนวนโรงเรียนในแต่ละขนาดโรงเรียน โดยใช้วิธีการสุ่มอย่างง่าย ได้จำนวนโรงเรียนทั้งหมด 6 โรงเรียน จำแนกเป็นโรงเรียนขนาดกลาง 4 โรงเรียน โรงเรียนขนาดใหญ่ 1 โรงเรียน และโรงเรียนขนาดใหญ่พิเศษ 1 โรงเรียน

ขั้นที่ 3 ทำการสุ่มแบบแบ่งชั้น (Stratified Random Sampling) ซึ่งมีขนาดโรงเรียนเป็นชั้น (Strata) นักเรียนเป็นหน่วยการสุ่ม (Sample Unit) โดยนำจำนวนกลุ่มตัวอย่างไปเทียบสัดส่วนกับจำนวนนักเรียนในแต่ละขนาดโรงเรียน แล้วสุ่มนักเรียนจำแนกตามขนาดโรงเรียน ได้เป็น โรงเรียนขนาดกลาง 4 โรงเรียน จำนวน 60 คน โรงเรียนขนาดใหญ่ 1 โรงเรียน จำนวน 33 คน โรงเรียนขนาดใหญ่พิเศษ 1 โรงเรียน จำนวน 87 คน ดังแสดงในตาราง 6

ตาราง 6 จำนวนนักเรียนชั้นมัธยมศึกษาปีที่ 4 ในแต่ละโรงเรียนที่เป็นกลุ่มตัวอย่างการวิจัย

ขนาดโรงเรียน	ชื่อโรงเรียน	ประชากร(คน)	กลุ่มตัวอย่าง(คน)
กลาง	ศรัทธาศิลาเพชรรังสรรค์	56	10
	สันติสุขพิทยาคม	80	14
	ทุ่งช้าง	105	18
	บ้านหลวง	107	18
ใหญ่	นาน้อย	194	33
ใหญ่พิเศษ	สตรีศรีน่าน	512	87
รวม		1054	180

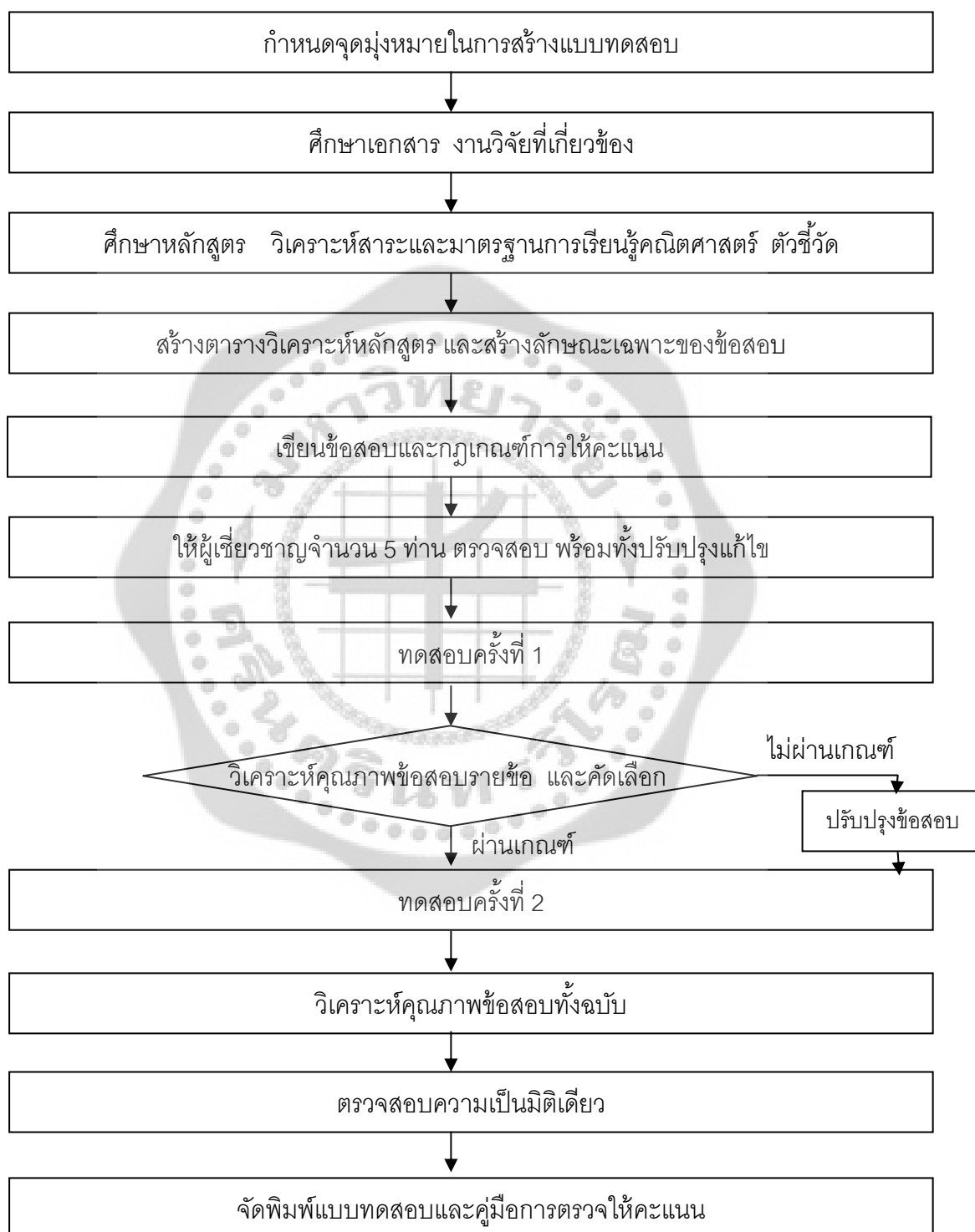
2. กลุ่มตัวอย่างที่เป็นผู้ตรวจให้คะแนน (Raters) แบบทดสอบปลายเปิดวิชาคณิตศาสตร์ โดยการคัดเลือกครูผู้สอนที่มีวุฒิทางการศึกษาตั้งแต่ระดับปริญญาตรีขึ้นไปทางด้านคณิตศาสตร์และมีประสบการณ์ในการสอนวิชาคณิตศาสตร์ระดับชั้นมัธยมศึกษาปีที่ 3 ไม่น้อยกว่า 5 ปี จำนวน 4 คน จากครูผู้สอนทั้งหมด 56 คน โดยเป็นครูที่เต็มใจให้ความร่วมมือ

การสร้างเครื่องมือที่ใช้ในการวิจัย

เครื่องมือที่ใช้ในการวิจัยครั้งนี้ คือ แบบทดสอบปลายเปิดวิชาคณิตศาสตร์ที่ผู้วิจัยสร้างขึ้น จำนวน 12 ข้อ ข้อละ 4 คะแนน เป็นการวัดความสามารถในวิชาคณิตศาสตร์ โดยเนื้อหาที่ใช้เป็นเนื้อหาตามสาระการเรียนรู้คณิตศาสตร์ ระดับชั้นมัธยมศึกษาตอนต้น (ชั้นมัธยมศึกษาปีที่ 1- 3) ตามหลักสูตรการศึกษาขั้นพื้นฐาน พุทธศักราช 2551

ขั้นตอนในการสร้างเครื่องมือ

เครื่องมือที่ใช้ในการศึกษาครั้งนี้เป็นแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ที่ผู้วิจัยสร้างขึ้น โดยมีขั้นตอนการสร้างและพัฒนาแบบทดสอบดังนี้



ภาพประกอบ 10 ขั้นตอนการสร้างแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ ชั้นมัธยมศึกษาตอนต้น

จากภาพประกอบ 10 ขั้นตอนการสร้างแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ ชั้นมัธยมศึกษาตอนต้น (ชั้นมัธยมศึกษาปีที่ 1-3) มีรายละเอียดดังนี้

1. กำหนดจุดมุ่งหมายในการสร้างแบบทดสอบ คือ เพื่อสร้างแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ ในการวัดความสามารถทางคณิตศาสตร์ ตามสาระการเรียนรู้คณิตศาสตร์ ระดับชั้นมัธยมศึกษาตอนต้น (ชั้นมัธยมศึกษาปีที่ 1-3) สำหรับใช้ในการวิจัย

2. ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องกับความหมาย ลักษณะและการสร้างคำถามปลายเปิดวิชาคณิตศาสตร์ ความหมายและลักษณะของกฎเกณฑ์การให้คะแนน เพื่อเป็นแนวทางในการสร้างแบบทดสอบปลายเปิดและกฎเกณฑ์การให้คะแนน

3. ศึกษาหลักสูตร วิเคราะห์สาระ มาตรฐานการเรียนรู้และตัวชี้วัด ศึกษาเนื้อหากลุ่มสาระการเรียนรู้คณิตศาสตร์ ระดับชั้นมัธยมศึกษาตอนต้น (ชั้นมัธยมศึกษาปีที่ 1-3) เพื่อกำหนดขอบเขตของเนื้อหาและพฤติกรรมที่ต้องการวัด โดยใช้แนวทางการแบ่งพฤติกรรมด้านพุทธิพิสัย (Cognitive domains) ตามแนวคิดของบลูม (Bloom, 1956)

4. สร้างตารางวิเคราะห์หลักสูตร โดยยึดมาตรฐานการเรียนรู้และตัวชี้วัด ตามหลักสูตรการศึกษาขั้นพื้นฐาน พุทธศักราช 2551 กลุ่มสาระการเรียนรู้คณิตศาสตร์ ระดับชั้นมัธยมศึกษาตอนต้น (ชั้นมัธยมศึกษาปีที่ 1-3) ได้แก่ สาระที่ 1 จำนวนและการดำเนินการ สาระที่ 2 การวัด สาระที่ 3 เรขาคณิต สาระที่ 4 พีชคณิต และสาระที่ 5 การวิเคราะห์ข้อมูลและความน่าจะเป็น ส่วนสาระที่ 6 ทักษะ/กระบวนการทางคณิตศาสตร์ ใช้เป็นแนวทางในการสร้างกฎเกณฑ์การให้คะแนน (รายละเอียดดังตาราง 7 ในภาคผนวก ก)

5. สร้างลักษณะเฉพาะของข้อสอบ (Item specification) เพื่อกำหนดกรอบลักษณะคำถามเขียนข้อสอบตามกรอบโครงสร้างที่กำหนดไว้ โดยสร้างข้อสอบตามสาระการเรียนรู้ทั้ง 5 สาระ สาระการเรียนรู้ละ 3 ข้อ รวมทั้งหมด 15 ข้อ

6. สร้างกฎเกณฑ์การให้คะแนน (Scoring rubrics) ของข้อสอบแต่ละข้อ เพื่อกำหนดกฎเกณฑ์การให้คะแนนตามลำดับขั้นความสำเร็จในการตอบ และใช้สาระที่ 6 คือ ทักษะ/กระบวนการทางคณิตศาสตร์ ได้แก่ ความสามารถในการแก้ปัญหา ความสามารถในการให้เหตุผล ความสามารถในการสื่อสาร การสื่อความหมายทางคณิตศาสตร์และการนำเสนอ ความสามารถในการเชื่อมโยงความรู้ต่าง ๆ ทางคณิตศาสตร์และเชื่อมโยงคณิตศาสตร์กับศาสตร์อื่น ๆ และการมีความคิดริเริ่มสร้างสรรค์ เป็นแนวทางในการสร้างกฎเกณฑ์การให้คะแนน ทั้งนี้เนื่องจากเนื้อหาและผลการเรียนรู้ที่คาดหวังในแต่ละสาระการเรียนรู้มีลักษณะแตกต่างกัน ผู้วิจัยใช้กฎเกณฑ์การให้คะแนนแบบรวม (Holistic scoring method) ที่มีลักษณะเป็นเกณฑ์เฉพาะ (Specific rubric) ของข้อสอบแต่ละข้อ ซึ่งเป็นการดูภาพรวมของแต่ละข้อว่ามีคุณลักษณะอย่างไร โดยการเขียนบรรยายความสำเร็จในการตอบ

ในแต่ละข้อ ตามระดับของความถูกต้องเหมาะสมจากน้อยไปหามาก โดยกำหนดเป็น 5 ระดับ คะแนน คือ 0, 1, 2, 3 และ 4 คะแนน

ตัวอย่างข้อสอบ

ข้อ 0. ราคาปกติของผู้เียนที่ขาย คือ 8,000 บาท ตะวันซื้อเงินสดทางร้านลดราคาให้ 25 % จากราคาปกติ จากนั้นตะวันนำไปขายให้เพื่อน โดยเพิ่มราคาขึ้นอีก 25 % จากราคาที่ซื้อมา อยากทราบว่าราคาผู้เียนที่ตะวันขายให้เพื่อนจะน้อยกว่า มากกว่า หรือ เท่ากับ 8,000 บาท จงอธิบาย

แนวการตอบ

ราคาผู้เียนที่ตะวันขายให้เพื่อนน้อยกว่า 8,000 บาท คำนวณได้ดังนี้

ราคาปกติของผู้เียน คือ 8,000 บาท ตะวันซื้อเงินสดทางร้านลดราคาให้ 25% จากราคาปกติ ลดราคาให้ 25 % จากราคาปกติ แสดงว่า

ตั้งราคาขายไว้ 100 บาท ขายจริง 75 บาท

ตั้งราคาขายไว้ 8,000 บาท ขายจริง $\frac{8,000 \times 75}{100} = 6,000$ บาท

(นักเรียนอาจหาราคาที่ลดก่อน แล้วนำมาลบออกจากราคาปกติที่ขาย จะเป็นราคาขายจริง)

แสดงว่าตะวันซื้อผู้เียนในราคา 6,000 บาท

จากนั้นตะวันนำไปขายให้เพื่อน โดยปรับราคาขึ้นอีก 25 % จากราคาที่ซื้อมา

ปรับราคาเพิ่มขึ้น 25 % จากราคาที่ซื้อมา แสดงว่า

ถ้าซื้อมาราคา 100 บาท ราคาขาย เป็น 125 บาท

ถ้าซื้อมาราคา 6,000 บาท ราคาขาย เป็น $\frac{6,000 \times 125}{100} = 7,500$ บาท

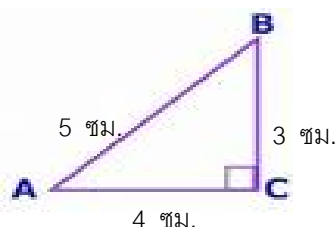
(นักเรียนอาจหาราคาที่เพิ่มขึ้นก่อน แล้วนำมาบวกกับราคาที่ซื้อมา จะเป็นราคาขาย)

ดังนั้น ตะวันขายผู้เียนในราคา 7,500 บาท ซึ่งมีค่าน้อยกว่า 8,000 บาท อยู่ 500 บาท

เกณฑ์การให้คะแนน

ระดับคะแนน	รายละเอียดของเกณฑ์
0	ไม่ตอบ หรือคำตอบไม่ถูกต้อง โดยไม่เขียนคำอธิบายหรือปรากฏร่องรอยการคิดใด ๆ ในการแก้ปัญหา
1	-คำตอบไม่ถูกต้อง แต่ปรากฏร่องรอยการคิดหรือเขียนคำอธิบาย โดยใช้หลักทางคณิตศาสตร์ที่ไม่ถูกต้อง เช่น เข้าใจว่าการลดราคา 25 % จากราคาปกติ มีค่าเท่ากับการเพิ่มราคาขึ้นอีก 25 % จากราคาที่ซื้อมา -แสดงวิธีคิดที่แสดงถึงความเข้าใจปัญหาได้เล็กน้อย แต่ไม่สามารถดำเนินการแก้ปัญหาได้ ไม่ได้คำตอบ
2	-คำตอบไม่ถูกต้อง แต่ได้แสดงวิธีแก้ปัญหาโดยใช้หลักการทางคณิตศาสตร์ได้ถูกต้องบางส่วน เช่น แสดงวิธีหารราคาซื้อมาได้ แต่หารราคาขายใหม่ไม่ถูกต้อง -แสดงวิธีแก้ปัญหาโดยใช้หลักการทางคณิตศาสตร์ได้ถูกต้อง แต่คำนวณผิดพลาด -คำตอบถูกต้อง แต่ไม่แสดงแนวคิดวิธีการแก้ปัญหา
3	แสดงวิธีการหารราคาซื้อและราคาใหม่ที่ขายโดยใช้หลักการทางคณิตศาสตร์ได้อย่างถูกต้อง คำนวณผิดพลาดเล็กน้อย
4	แสดงวิธีการหารราคาซื้อและราคาใหม่ที่ขายโดยใช้หลักการทางคณิตศาสตร์ได้อย่างถูกต้อง สมบูรณ์ คำนวณถูกต้องและคำตอบถูกต้อง

ข้อ 00. ครูให้นักเรียนวาดรูปสามเหลี่ยมมุมฉากที่มีด้านตรงข้ามมุมฉากยาว 5 ซม. ด.ช. ต้น วาดรูป โดยให้ด้านตรงข้ามมุมฉากยาว 5 ซม. และด้านประกอบมุมฉากยาว 3 ซม. และ 4 ซม. ดังรูป



แต่ ด.ช. ใจ บอกครูว่า เขาสามารถวาดรูปสามเหลี่ยมมุมฉากรูปอื่น ที่มีด้านตรงข้ามมุมฉากยาว 5 ซม. แต่ด้านประกอบมุมฉากไม่ได้ยาว 3 และ 4 ซม. ได้ นักเรียนเห็นด้วยกับ ด.ช. ใจ หรือไม่ จงอธิบายและยกตัวอย่างประกอบ

แนวการตอบ

ใจ้ กล่าวถูกต้อง เพราะรูปสามเหลี่ยมมุมฉากที่มีด้านตรงข้ามมุมฉากยาว 5 ซม. ไม่จำเป็นที่ด้านประกอบมุมฉากต้องยาว 3 และ 4 ซม. เสมอไป

ทั้งนี้สามารถกำหนดความยาวของด้านประกอบมุมฉาก 1 ค่า แล้วใช้ทฤษฎีพีทาโกรัส คำนวณหาความยาวของด้านประกอบมุมฉากอีกด้านได้ ตัวอย่างเช่น

ด้านตรงข้ามมุมฉากยาว 5 ซม. ให้ด้านประกอบมุมฉากด้านหนึ่งยาว 1 ซม. อีกด้านยาว a ซม.

$$\begin{aligned}\text{จะได้} \quad 5^2 &= 1^2 + a^2 \\ 25 &= 1 + a^2 \\ a &= \sqrt{24}\end{aligned}$$

ดังนั้น รูปสามเหลี่ยมมุมฉากที่มีด้านตรงข้ามมุมฉากยาว 5 ซม. ด้านประกอบมุมฉากยาว 1 ซม. และ $\sqrt{24}$ ซม.

หมายเหตุ ด้านประกอบมุมฉากอาจมีความยาวเป็น 2, $\sqrt{21}$ หรือ $\frac{5}{\sqrt{2}}, \frac{5}{\sqrt{2}}$ ซม. เป็นต้น

เกณฑ์การให้คะแนน

ระดับคะแนน	รายละเอียดของเกณฑ์
0	ไม่มีคำตอบ คำตอบไม่ถูกต้อง โดยไม่ได้แสดงเหตุผลหรือยกตัวอย่างประกอบ
1	คำตอบไม่ถูกต้อง โดยอธิบายแสดงเหตุผล หรือยกตัวอย่างประกอบ
2	คำตอบถูกต้อง แต่ - ไม่อธิบายวิธีคิดหรือแสดงเหตุผลว่าคำตอบได้มาอย่างไรและไม่ยกตัวอย่างประกอบ - อธิบายวิธีคิดหรือแสดงเหตุผลไม่ถูกต้อง หรือยกตัวอย่างประกอบไม่ถูกต้อง
3	คำตอบถูกต้อง แสดงเหตุผลได้ถูกต้อง แต่ไม่ได้ยกตัวอย่างประกอบ หรือยกตัวอย่างประกอบ แต่การให้เหตุผลไม่สมบูรณ์หรือชัดเจน
4	คำตอบถูกต้อง ใช้ความรู้มาให้เหตุผลได้อย่างชัดเจนสมบูรณ์ และยกตัวอย่างได้ถูกต้อง

7. ทำการตรวจสอบคุณภาพแบบทดสอบโดยพิจารณาความเที่ยงตรงเชิงเนื้อหาและความเป็นปรนัยของข้อคำถาม รวมทั้งกฎเกณฑ์การให้คะแนนจากผู้เชี่ยวชาญทางด้านการวัดประเมินในสาขาวิชาคณิตศาสตร์ จำนวน 5 ท่าน (ดังรายชื่อในภาคผนวก ก) โดยพิจารณาว่าข้อคำถามนั้นตรงตามตัวชี้วัดหรือไม่ พร้อมทั้งความเหมาะสมของภาษากับระดับนักเรียน โดยใช้ดัชนีความสอดคล้อง (Item-Objective Congruency Index: IOC) ของโรวินลลีและแฮมเบิลตัน (ล้วน สายยศ; และอังคณา สายยศ. 2539: 148-149 อ้างอิงจาก Rowinelli; & Hambleton. 1977. Journal for Education Research) ซึ่งเป็นค่าเฉลี่ยของคะแนนตัดสินของผู้เชี่ยวชาญในแต่ละข้อ คัดเลือกข้อสอบข้อที่มีค่าเฉลี่ยสูงกว่า 0.5 ซึ่งแสดงว่าข้อสอบนั้นสามารถวัดได้ตามที่กรอบโครงสร้างกำหนดไว้

8. นำผลการพิจารณาของผู้เชี่ยวชาญมาคำนวณหาค่า IOC ได้ค่าคะแนนเฉลี่ยจากผลการตัดสินของผู้เชี่ยวชาญในแต่ละข้ออยู่ระหว่าง 0.40-1.00 (ผลปรากฏดังภาคผนวก ก) นั่นคือข้อสอบและกฎเกณฑ์การให้คะแนนที่สร้างขึ้นมีค่าเฉลี่ยน้อยกว่า 0.50 จำนวน 1 ข้อ ผู้วิจัยทำการปรับปรุงข้อสอบและกฎเกณฑ์การให้คะแนนข้อดังกล่าว แล้วนำไปให้ผู้เชี่ยวชาญพิจารณาอีกครั้งหนึ่งก่อนนำไปทดลองใช้

9. ทดลองสอบครั้งที่ 1 โดยการนำแบบทดสอบที่ผ่านตรวจสอบในขั้นต้นแล้วไปทดลองสอบกับกลุ่มตัวอย่างที่ใช้ในการหาคุณภาพเครื่องมือ ซึ่งเป็นนักเรียนชั้นมัธยมศึกษาปีที่ 4 โรงเรียนสาธิตแห่งมหาวิทยาลัยราชภัฏรำไพพรรณี จังหวัดระยอง จำนวน 40 คน เพื่อตรวจสอบเวลาที่เหมาะสมในการทำข้อสอบ โดยพิจารณาจากจำนวนนักเรียนที่ทำเสร็จประมาณ 80 % ของนักเรียนทั้งหมด

10. นำผลการทดลองสอบครั้งที่ 1 มาตรวจให้คะแนนตามกฎเกณฑ์ที่กำหนด โดยใช้ผู้ตรวจเป็นครูผู้สอนวิชาคณิตศาสตร์ในระดับชั้นมัธยมศึกษาปีที่ 3 จำนวน 3 คน (ดังรายชื่อในภาคผนวก ก) ผู้วิจัยชี้แจงวิธีการตรวจ กฎเกณฑ์การให้คะแนน เพื่อสร้างความเข้าใจให้ตรงกัน และให้ผู้ตรวจตรวจข้อสอบทีละฉบับและตรวจทีละข้ออย่างเป็นอิสระต่อกัน

11. วิเคราะห์คุณภาพข้อสอบ ตามทฤษฎีการทดสอบแบบมาตรฐานเดิม (CTT) เพื่อหาค่าความยาก ค่าอำนาจจำแนกรายข้อ โดยพิจารณาข้อสอบที่มีค่าความยากตั้งแต่ 0.20 ถึง 0.80 มีค่าดัชนีอำนาจจำแนกตั้งแต่ 0.20 ขึ้นไป ซึ่งค่าความยาก (P) และค่าอำนาจจำแนก (r) ของแบบทดสอบปลายเปิด ใช้การแบ่งกลุ่มผู้สอบออกเป็นกลุ่มเก่งและกลุ่มอ่อน โดยใช้เทคนิค 25 % ของผู้สอบทั้งหมด คำนวณจากสูตรของวิทนีและซาเบอร์ส (ล้วน สายยศ; และอังคณา สายยศ. 2539: 199-200; อ้างอิงจาก Whitney; & Sabers. 1970) ดังนี้

$$P = \frac{S_U + S_L - (2N X_{\min})}{2N(X_{\max} - X_{\min})}$$

$$r = \frac{S_U - S_L}{N(X_{\max} - X_{\min})}$$

เมื่อ	P	คือ	ค่าความยาก
	r	คือ	ค่าอำนาจจำแนก
	S_U	คือ	ผลรวมของคะแนนกลุ่มเก่ง
	S_L	คือ	ผลรวมของคะแนนกลุ่มอ่อน
	N	คือ	จำนวนผู้เข้าสอบของกลุ่มเก่งหรือกลุ่มอ่อน
	$X_{\max}$	คือ	คะแนนที่นักเรียนทำได้สูงสุด
	$X_{\min}$	คือ	คะแนนที่นักเรียนทำได้ต่ำสุด

ผลการวิเคราะห์ค่าความยากของข้อสอบได้ค่าความยากของข้อสอบอยู่ระหว่าง 0.3000–0.7726 ค่าอำนาจจำแนกของข้อสอบอยู่ระหว่าง 0.1567–0.6753 มีข้อสอบที่มีคุณลักษณะทางสถิติที่ไม่อยู่ในเกณฑ์ คือ มีอำนาจจำแนกต่ำกว่า 0.20 จำนวน 1 ข้อ รวมทั้งพิจารณาเวลาที่พอเหมาะสำหรับการทำข้อสอบทั้งฉบับ ซึ่งเป็นเวลาที่นักเรียนส่วนใหญ่ประมาณร้อยละ 80 ทำข้อสอบเสร็จ คือ 2 ชั่วโมง แล้วไปทดลองใช้แบบทดสอบครั้งที่ 2

12. ทดลองใช้แบบทดสอบครั้งที่ 2 โดยนำแบบทดสอบที่ปรับปรุงแก้ไขแล้วจากการทดลองครั้งที่ 1 ไปทดสอบกับนักเรียนชั้นมัธยมศึกษาปีที่ 4 โรงเรียนสา อำเภอยางสงขลา จังหวัดน่าน จำนวน 40 คน วิเคราะห์ข้อสอบเช่นเดียวกับข้อ 11 เพื่อให้ได้แบบทดสอบที่มีความยากอยู่ระหว่าง 0.20–0.80 ข้อสอบแต่ละข้อมีอำนาจจำแนกตั้งแต่ 0.20 ขึ้นไป ซึ่งได้ค่าความยากอยู่ระหว่าง 0.2458–0.7684 ค่าอำนาจจำแนกของข้อสอบอยู่ระหว่าง 0.1918–0.6920 ได้พิจารณาคำตอบของนักเรียนประกอบเพื่อเป็นตัวแบบในการพิจารณาคุณภาพเกณฑ์การให้คะแนนเพิ่มเติม และค่าความเชื่อมั่นของแบบทดสอบ ใช้การวิเคราะห์ตามแบบคะแนนจริงสมมูล โดยใช้สูตรสัมประสิทธิ์แอลฟา (Coefficient Alpha) (บุญเชิด ภิญญอนันตพงษ์, 2547: 220) สูตรที่ใช้คือ

$$\alpha = \frac{K}{K-1} \left[1 - \frac{\sum S_i^2}{S_x^2} \right]$$

เมื่อ	α	คือ	ค่าความเชื่อมั่นของแบบทดสอบ
	K	คือ	จำนวนข้อคำถามของแบบทดสอบ

S_i^2 คือ ค่าความแปรปรวนของคะแนนแต่ละข้อ

S_x^2 คือ ค่าความแปรปรวนของคะแนนทั้งฉบับ

ผลการวิเคราะห์ค่าความเชื่อมั่นของแบบทดสอบ ได้ค่าความเชื่อมั่นเป็น 0.8826

13. คัดเลือกข้อสอบที่มีคุณภาพไว้ทั้งหมดจำนวน 12 ข้อ เพื่อใช้ในการศึกษาคั้งนี้โดยคัดเลือกข้อสอบให้ได้ครบทุกสาระการเรียนรู้ แล้ววิเคราะห์ค่าความยากของข้อสอบซ้ำ ซึ่งได้ค่าความยากมีค่าตั้งแต่ 0.2499–0.7729 ค่าอำนาจจำแนกของข้อสอบมีค่าตั้งแต่ 0.2015-0.6752 (ดังผลปรากฏในภาคผนวก ก) และค่าความเชื่อมั่นของแบบทดสอบทั้งฉบับเป็น 0.8771 จากนั้นผู้วิจัยตรวจสอบความเป็นมิติเดียว (Unidimensional) ของแบบทดสอบ เพื่อให้เป็นไปตามข้อตกลงเบื้องต้นของทฤษฎีการตอบข้อสอบ โดยใช้การวิเคราะห์องค์ประกอบเชิงยืนยัน (Confirmatory Factor Analysis: CFA) เพื่อทดสอบความสอดคล้องระหว่างโมเดลกับข้อมูลเชิงประจักษ์ที่กำหนดให้มีตัวแปรแฝงเพียงตัวเดียวในตัวแปรสังเกตทั้งหมด วิเคราะห์ด้วยโปรแกรม LISREL (ปรากฏผลดังภาคผนวก ก) ทำการวิเคราะห์โดยการแปลงตัวแปรคะแนนให้เป็นตัวแปรสุ่มแบบปกติ พบว่า แบบทดสอบมีค่าไคสแควร์ เท่ากับ 60.17 ค่า p เท่ากับ 0.20 มีชั้นแห่งความเป็นอิสระ(df) เท่ากับ 52 และเมื่อหารค่าไคสแควร์ ด้วยชั้นแห่งความเป็นอิสระ มีค่าเท่ากับ 1.1571 แสดงว่า โมเดลสอดคล้องกับข้อมูล และเมื่อพิจารณาค่า RMSEA, NNFI, GFI, AGFI และ CFI มีค่าเท่ากับ 0.03, 0.99, 0.95, 0.92, 0.99 ตามลำดับ ซึ่งค่าดังกล่าว อยู่ในเกณฑ์ที่ยอมรับได้ แสดงว่าแบบทดสอบฉบับนี้มีหนึ่งองค์ประกอบหลักจริงหรือแบบทดสอบนี้มีความเป็นมิติเดียว

15. นำข้อสอบที่คัดเลือกไว้จำนวน 12 ข้อ มาจัดพิมพ์เป็นรูปเล่มแบบทดสอบปลายเปิด วิชาคณิตศาสตร์ เพื่อนำไปทดสอบกับกลุ่มตัวอย่างที่ใช้ในการวิจัยต่อไปและจัดพิมพ์คู่มือการตรวจให้คะแนนสำหรับผู้ตรวจให้คะแนน (ปรากฏดังภาคผนวก ค)

การเก็บรวบรวมข้อมูล

การเก็บรวบรวมข้อมูลในครั้งนี้ ผู้วิจัยดำเนินการเก็บรวบรวมข้อมูลด้วยตนเอง โดยมีการดำเนินการเก็บรวบรวมข้อมูล ดังนี้

1. ทำหนังสือจากบัณฑิตวิทยาลัย เพื่อขอความอนุเคราะห์กับผู้บริหารสถานศึกษาที่เป็นกลุ่มตัวอย่าง
2. ประสานงานและชี้แจงผู้บริหารสถานศึกษาและครูผู้สอนของโรงเรียนที่เป็นกลุ่มตัวอย่างเกี่ยวกับวัตถุประสงค์การเก็บรวบรวมข้อมูล เพื่อกำหนดวัน เวลาและสถานที่สอบ
3. ดำเนินการสอบ โดยนำแบบทดสอบที่มีความยาว 12 ข้อ ไปทดสอบกับนักเรียนที่เป็นกลุ่มตัวอย่างจำนวน 180 คน ก่อนให้นักเรียนลงมือทำข้อสอบผู้วิจัยได้ชี้แจงให้นักเรียนเข้าใจ

วัตถุประสงค์ ประโยชน์ที่ได้รับและอธิบายวิธีการทำแบบทดสอบ โดยได้ดำเนินการสอบทั้ง 6 โรงเรียน ระหว่างวันที่ 24 สิงหาคม 2552 ถึงวันที่ 31 สิงหาคม 2552

4. นำกระดาษคำตอบของผู้เข้าสอบทั้งหมดมาถ่ายเอกสารจำนวน 4 ชุด จัดให้ผู้ตรวจจำนวน 4 คน โดยผู้วิจัยได้อบรมชี้แจงวิธีการตรวจและกฎเกณฑ์การให้คะแนนที่ผู้วิจัยได้สร้างขึ้นในคู่มือการตรวจให้คะแนนเพื่อสร้างความเข้าใจให้ตรงกัน แล้วจึงให้ผู้ตรวจตรวจให้คะแนนข้อสอบทุกข้อของผู้สอบทุกคน พร้อมทั้งจดบันทึกเวลาที่ใช้ในการตรวจให้คะแนน

5. นำผลการตรวจที่ได้มาทำการวิเคราะห์เพื่อหาค่าทางสถิติต่อไป

การจัดกระทำและการวิเคราะห์ข้อมูล

การวิจัยในครั้งนี้มีขั้นตอนการจัดกระทำและการวิเคราะห์ข้อมูล ดังนี้

1. นำกระดาษคำตอบของผู้เข้าสอบทั้งหมดมาถ่ายเอกสารจำนวน 4 ชุด จัดให้ผู้ตรวจจำนวน 4 คน ตรวจให้คะแนนตามกฎเกณฑ์การให้คะแนนที่ผู้วิจัยได้สร้างขึ้นในคู่มือการตรวจให้คะแนน ผู้วิจัยได้อบรมชี้แจงวิธีการตรวจและกฎเกณฑ์การให้คะแนนที่ผู้วิจัยได้สร้างขึ้นในคู่มือการตรวจให้คะแนนเพื่อสร้างความเข้าใจให้ตรงกัน เมื่อผู้ตรวจเข้าใจตรงกันแล้ว จึงให้อิสระผู้ตรวจแต่ละคนตรวจข้อสอบทีละข้อจนครบทุกข้อและครบตามจำนวนผู้สอบทุกคน พร้อมทั้งจดบันทึกเวลาทั้งหมดที่ใช้ในตรวจให้คะแนน

2. ผู้วิจัยนำผลการตรวจให้คะแนนข้อสอบจำนวน 12 ข้อจากผู้ตรวจทั้ง 4 คนของนักเรียนทั้งหมด 180 คนมาจัดกระทำข้อมูลตามเงื่อนไขและระดับของเงื่อนไขที่ต้องการศึกษาดังนี้

2.1 กรณีผู้ตรวจจำนวน 4 คน นำผลการตรวจให้คะแนนของผู้ตรวจทั้ง 4 คนมาดำเนินการจัดกระทำตามรูปแบบการตรวจให้คะแนน ดังนี้

2.1.1 ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน [$p \times i \times r$ (full-crossed design)] ใช้ผลการตรวจให้คะแนนของผู้ตรวจทั้ง 4 คนที่ตรวจข้อสอบทั้ง 12 ข้อของผู้สอบทั้งหมด 180 คน

2.1.2 ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน [$(p : r) \times i$ [(persons nested within raters) crossed with items]] โดยการนำผลการตอบของผู้สอบทั้ง 180 คน มาสุ่มด้วยการสุ่มอย่างง่าย เพื่อสุ่มผลการตอบของผู้สอบออกเป็น 4 กลุ่มๆ ละเท่า ๆ กัน ได้กลุ่มละ 45 คน จากนั้นสุ่มผลการตรวจให้คะแนนข้อสอบทั้ง 12 ข้อของผู้ตรวจ 1 คนต่อ 1 กลุ่มผู้สอบโดยไม่ซ้ำกัน ด้วยการสุ่มอย่างง่าย

2.1.3 ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน [$p \times (i : r)$ [persons crossed with (items nested within raters)]] โดยแบ่งข้อสอบเป็น 4 กลุ่ม กลุ่มละเท่า ๆ กัน จากข้อสอบทั้งหมด 12 ข้อ ได้กลุ่มละ 3 ข้อ ด้วยวิธีการสุ่มแบบแบ่งชั้นโดยมีสาระการเรียนรู้เป็นชั้นและข้อสอบ

เป็นหน่วยการสุ่ม ทำให้ได้ข้อสอบในแต่ละกลุ่มที่ประกอบด้วย 3 สารการเรียนรู้ จากนั้นจึงสุ่มผลการตรวจของผู้ตรวจ 1 คนต่อ 1 กลุ่มข้อสอบโดยไม่ซ้ำกัน ด้วยการสุ่มอย่างง่าย

2.2 กรณีผู้ตรวจจำนวน 3 คน และ 2 คน นั้นใช้การสุ่มจากผู้ตรวจ 4 คน โดยการสุ่มอย่างง่าย และนำผลการตรวจให้คะแนนของผู้ตรวจจำนวน 3 คน และ 2 คน มาจัดกระทำตามรูปแบบการตรวจให้คะแนนที่ต่างกัน เช่นเดียวกับข้อ 2.1.1-2.1.3 โดยจำนวนผู้สอบในแต่ละกลุ่ม (ข้อ 2.1.2) จะได้กลุ่มละ 60 คนและ 90 คน ตามลำดับ ส่วนจำนวนข้อสอบในแต่ละกลุ่ม (ข้อ 2.1.3) จะได้กลุ่มละ 4 ข้อและ 6 ข้อ ตามลำดับ ซึ่งแตกต่างกันไปจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนที่ต่างกัน สามารถแสดงดังภาพประกอบ 11

กรณีผู้ตรวจ 4 คน

1.ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน

[illegible]

2. ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน

p	r1				p	r2				p	r3				p	r4			
	i1	i2	...	i12		i1	i2	...	i12		i1	i2	...	i12		i1	i2	...	i12
1					46					91					136				
2					47					92					137				
...					...					...					...				
45					90					135					180				

3. ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน

[illegible]

กรณีผู้ตรวจ 3 คน

1. ผู้ตรวจตรวจสอบข้อสอบทุกข้อของผู้สอบทุกคน

p	r1				r2				r3			
	i1	i2	...	i12	i1	i2	...	i12	i1	i2	...	i12
1												
2												
...												
180												

2. ผู้ตรวจตรวจสอบข้อสอบทุกข้อของผู้สอบบางคน

p	r1				p	r2				p	r3			
	i1	i2	...	i12		i1	i2	...	i12		i1	i2	...	i12
1					61					121				
2					62					122				
...					...					...				
60					120					180				

3. ผู้ตรวจตรวจสอบข้อสอบบางข้อของผู้สอบทุกคน

p	r1				r2				r3			
	i1	i4	i7	i10	i2	i5	i8	i11	i3	i6	i9	i12
1												
2												
...												
180												

กรณีผู้ตรวจ 2 คน

1. ผู้ตรวจตรวจสอบข้อสอบทุกข้อของผู้สอบทุกคน

p	r1				r2			
	i1	i2	...	i12	i1	i2	...	i12
1								
2								
...								
180								

2. ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน

p	r1				p	r2			
	i1	i2	...	i12		i1	i2	...	i12
1					91				
2					92				
...					...				
60					180				

3. ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน

p	r1						r2					
	i1	i3	i5	i7	i9	i11	i2	i4	i6	i8	i10	i12
1												
2												
...												
180												

หมายเหตุ เมื่อ p แทน ผู้สอบคนที่... r1-r4 แทน ผู้ตรวจคนที่1-4 และ i1-i12 แทน ข้อสอบข้อที่ 1-12

ภาพประกอบ 11 การจัดกระทำข้อมูลตามเงื่อนไขและระดับของเงื่อนไขที่ศึกษา

ดังนั้นจึงสามารถจัดข้อมูลสำหรับการศึกษานี้ได้ทั้งหมด 9 ชุด (3x3) ดังนี้

รูปแบบการตรวจให้คะแนน				
เงื่อนไขที่ศึกษา				
	ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน	ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน	ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน	
จำนวน	4 คน	12 ข้อ (180 คน)	12 ข้อ (45 คน)	3 ข้อ (180 คน)
ผู้ตรวจ	3 คน	12 ข้อ (180 คน)	12 ข้อ (60 คน)	4 ข้อ (180 คน)
	2 คน	12 ข้อ (180 คน)	12 ข้อ (90 คน)	6 ข้อ (180 คน)

3. นำผลที่ได้จากการตรวจให้คะแนนในแต่ละเงื่อนไข มาวิเคราะห์ข้อมูล

3.1 วิเคราะห์ค่าสถิติพื้นฐานของคะแนนสอบ โดยคำนวณค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐาน คะแนนสูงสุด คะแนนต่ำสุด ความเบ้และความโด่ง เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกันโดยใช้โปรแกรมสำเร็จรูป และสถิติพื้นฐานของค่าพารามิเตอร์ของผู้สอบ

ผู้ตรวจและข้อสอบ ได้แก่ ค่าสูงสุด ค่าต่ำสุด ค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐาน ค่า INFIT และ OUTFIT ด้วยโปรแกรม FACETS (Linacre. 2009)

3.2 วิเคราะห์องค์ประกอบความแปรปรวน โดยใช้โมเดลการสรุปอ้างอิง เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน ซึ่งหน่วยการวิเคราะห์ (Unit analysis) ความแปรปรวนของผู้สอบ ข้อสอบ และผู้ตรวจ คือ กลุ่มผู้สอบ กลุ่มข้อสอบ และกลุ่มผู้ตรวจ ตามลำดับ วิเคราะห์ด้วยโปรแกรม GENOVA (Crick; & Brennan. 1983)

3.3 ประเมินค่าและเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิง ($E\rho^2$) เมื่อจำนวนผู้ตรวจเท่ากัน แต่มีรูปแบบการตรวจให้คะแนนต่างกัน ทดสอบความแตกต่างของค่าสัมประสิทธิ์การสรุปอ้างอิง โดยใช้สูตร UX_1 ของวูดรอฟฟ์และเฟลด์ท์ (Woodruff; & Feldt. 1986: 393–413)

3.4 ประเมินค่าและเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิง ($E\rho^2$) เมื่อรูปแบบการตรวจให้คะแนนเหมือนกัน แต่จำนวนผู้ตรวจต่างกัน ทดสอบความแตกต่างของค่าสัมประสิทธิ์การสรุปอ้างอิง โดยใช้สูตร UX_1

3.5 วิเคราะห์องค์ประกอบความแปรปรวน โดยใช้โมเดลหลายองค์ประกอบของราล์ช เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน ซึ่งหน่วยการวิเคราะห์ (Unit analysis) ความแปรปรวนของผู้สอบ ข้อสอบ และผู้ตรวจ คือ กลุ่มผู้สอบ กลุ่มข้อสอบ และกลุ่มผู้ตรวจ ตามลำดับ ด้วยโปรแกรม FACETS (Linacre. 2009)

3.6 ประเมินค่าและเปรียบเทียบค่าความเชื่อมั่นแยกส่วนของแต่ละองค์ประกอบ เมื่อจำนวนผู้ตรวจเท่ากัน แต่รูปแบบการตรวจให้คะแนนต่างกัน โดยการทดสอบความแตกต่างของค่าความเชื่อมั่นแยกส่วน ใช้สูตร UX_1

3.7 ประเมินค่าและเปรียบเทียบค่าความเชื่อมั่นแยกส่วนของแต่ละองค์ประกอบ เมื่อรูปแบบการตรวจให้คะแนนเหมือนกัน แต่จำนวนผู้ตรวจต่างกัน โดยการทดสอบความแตกต่างของค่าความเชื่อมั่นแยกส่วน โดยใช้สูตร UX_1

3.8 วิเคราะห์หาความเที่ยงตรงตามสภาพ (Concurrent validity) ของคะแนนที่ได้จากแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน โดยพิจารณาจากค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation coefficient) ของคะแนนดิบ คะแนนความสามารถกับคะแนนที่ได้จากผลการทดสอบทางการศึกษาระดับชาตินี้พื้นฐาน (Ordinary National Education Test: O-NET) ในช่วงชั้นที่ 3 ปีการศึกษา 2551 (มัธยมศึกษาปีที่ 3) ซึ่งผู้วิจัยใช้เป็นคะแนนเกณฑ์ในการพิจารณา ทั้งนี้การสอบ O-NET นั้นดำเนินการสอบโดยสถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน)(สทศ.) ซึ่งในวิชาคณิตศาสตร์ เป็นข้อสอบปรนัยแบบ 4 ตัวเลือก จำนวน 30 ข้อ คะแนนเต็ม 100 คะแนน โดยเนื้อหาในแบบทดสอบเป็นไปตามมาตรฐานการเรียนรู้ที่

ครอบคลุม 5 สาระการเรียนรู้ ส่วนสาระที่ 6 เป็นสาระที่แทรกอยู่ในสาระที่ 1-5 ซึ่งผลการสอบ O-NET ในปีการศึกษา 2551 พบว่า ค่าสถิติพื้นฐานผลการสอบ O-NET ในวิชาคณิตศาสตร์ มีคะแนนเฉลี่ย 32.66 คะแนน ส่วนเบี่ยงเบนมาตรฐาน 11.69 คะแนนสูงสุด 97.00 คะแนน คะแนนต่ำสุด 0.00 คะแนน มัธยฐาน 31.00 คะแนน และฐานนิยม 29.00 คะแนน (สทศ. 2552: online) โดยใช้ค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน (Pearson Product Moment Correlation)

3.9 เปรียบเทียบค่าความเที่ยงตรงตามสภาพของคะแนนแบบทดสอบปลายเปิด เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจคะแนนต่างกัน ซึ่งเป็นการเปรียบเทียบภายในแต่ละโมเดล โดยการนำค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันมาแปลงเป็นคะแนนมาตรฐาน (Z) ของฟิชเชอร์ แล้วนำมาทดสอบความแตกต่าง โดยใช้สถิติไค-สแควร์ (Chi-Square; χ^2) (Meng; Rosenthal; & Robin. 1992: 173) ถ้าพบว่าค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันที่นำมาทดสอบ มีความแตกต่างอย่างมีนัยสำคัญ จึงนำค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันมาทดสอบความแตกต่างรายคู่ โดยใช้สถิติ Z (Meng; Rosenthal; & Robin. 1992: 173) ต่อไป

4. สถิติที่ใช้ในการวิเคราะห์ข้อมูล

4.1 ค่าสัมประสิทธิ์การสรุปอ้างอิง (Generalizability Coefficient: $E\rho^2$) วิเคราะห์โดยใช้โปรแกรม GENOVA (Crick; & Brennan. 1983) คำนวณจากสูตร (Brennan. 2001: 35)

$$E\rho^2 = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(\delta)}$$

เมื่อ $E\rho^2$ คือ ค่าสัมประสิทธิ์การสรุปอ้างอิง

$\sigma^2(p)$ คือ ความแปรปรวนของคะแนนเอกภพ

$\sigma^2(\delta)$ คือ ความแปรปรวนของคะแนนความคลาดเคลื่อน

4.2 ค่าความเชื่อมั่นแยกส่วน (The reliability of separate index: R) วิเคราะห์โดยใช้โปรแกรม FACETS (Linacre. 2009) คำนวณจากสูตร (Engelhard. 1994: 97)

$$R = \frac{SD^2 - MSE}{SD^2}$$

เมื่อ R คือ ค่าความเชื่อมั่นแยกส่วน

SD คือ ส่วนเบี่ยงเบนมาตรฐานของค่าการสังเกต

MSE คือ ค่าเฉลี่ยของความคลาดเคลื่อนของการวัด

4.3 ทดสอบความแตกต่างของค่าสัมประสิทธิ์การสรุปอ้างอิง/ค่าความเชื่อมั่นแยกส่วน โดยใช้สูตร UX_1 ดังนี้ (Woodruff; & Feldt. 1986: 393-413)

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

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

$$u_i = \left[\frac{1}{1-r_i} \right]^{1/3}$$

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

$$C_u = \frac{4}{9m(m-1)(N_c - 1)} \sum_{i=2}^m \sum_{j=1}^{i-1} [r(x_1, x_2)]^2 u_i u_j$$

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

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

เมื่อ UX_1 คือ ค่าสถิติทดสอบที่มีการแจกแจงคล้าย χ^2 ที่ $df = m-1$

N คือ จำนวนนักเรียนที่สอบ

n คือ จำนวนค่าสังเกตหรือจำนวนข้อสอบ

m คือ จำนวนค่าสัมประสิทธิ์การสรุปอ้างอิง/ค่าความเชื่อมั่นแยกส่วน

r_i คือ ค่าสัมประสิทธิ์การสรุปอ้างอิง/ค่าความเชื่อมั่นแยกส่วน

- s_u^2 คือ ความแปรปรวนของ u_i
 $\bar{u}$ คือ ค่าเฉลี่ยของคะแนน u_i
 N_c คือ การปรับแก้สมการของความไม่เป็นอิสระ
 $\bar{n}_h$ คือ ค่าเฉลี่ยฮาโมนิคของความยาวของข้อสอบ
 C_u คือ ค่าที่ปรับแก้ความแปรปรวนของ u_i ในกรณีที่มีค่าสัมประสิทธิ์การ
 สรุปร่างอิงคำนวณจากกลุ่มตัวอย่างกลุ่มเดียวกัน
 $r(x_i, x_j)$ คือ ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างแบบทดสอบวิชาคณิตศาสตร์
 แต่ละรูปแบบที่จับคู่กัน

4.4 ความเที่ยงตรงตามสภาพ (Concurrent validity) ของคะแนนแบบทดสอบปลายเปิด โดยการหาค่าสัมประสิทธิ์สหสัมพันธ์ของคะแนนจากแบบทดสอบปลายเปิดที่ได้ในแต่ละเงื่อนไข กับ คะแนนที่ได้จากผลการทดสอบทางการศึกษาระดับชาตินขั้นพื้นฐาน (Ordinary National Education Test: O-NET) ใช้สูตรสหสัมพันธ์แบบเพียร์สัน ดังนี้ (บุญเชิด ภิญโญนนตพงษ์. 2547: 181)

$$r_{xy} = \frac{N \sum XY - \sum X \sum Y}{\sqrt{[N \sum X^2 - (\sum X)^2][N \sum Y^2 - (\sum Y)^2]}}$$

- เมื่อ r_{xy} คือ ค่าสัมประสิทธิ์สหสัมพันธ์
 X คือ คะแนนจากแบบทดสอบปลายเปิดวิชาคณิตศาสตร์
 Y คือ คะแนนที่ได้จากผลการทดสอบทางการศึกษาระดับชาตินขั้นพื้นฐาน
 (Ordinary National Education Test: O-NET)
 N คือ จำนวนนักเรียน

4.5 ทดสอบความแตกต่างของค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน โดยใช้สถิติไค-สแควร์ (Chi-Square; χ^2) มีสูตร ดังนี้ (Meng; Rosenthal; & Robin. 1992: 173)

$$\chi^2(k-1) = \frac{(N-3) \sum_i (Z_{ri} - \bar{Z}_r)^2}{(1-r_x)h}$$

โดยที่

$$h = 1 + \frac{\overline{r^2}}{1 - r^2} (1 - f)$$

$$f = \frac{1 - r_x}{2(1 - r^2)}$$

เมื่อ Z_r คือ คะแนนมาตรฐานที่แปลงมาจากค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันของพิชเชอร์

$\overline{Z_r}$ คือ ค่าเฉลี่ยของคะแนนมาตรฐานที่แปลงมาจากค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันของพิชเชอร์

r คือ ค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน

$\overline{r^2}$ คือ ค่าเฉลี่ยของกำลังสองของค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน

r_x คือ ค่ามัธยฐาน (Median) ของค่าสัมประสิทธิ์สหสัมพันธ์

N คือ จำนวนผู้สอบในกลุ่มตัวอย่าง

k คือ จำนวนค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันที่ต้องการเปรียบเทียบ

4.6 ทดสอบความแตกต่างรายคู่ของค่าสัมประสิทธิ์สหสัมพันธ์ โดยใช้สถิติ Z มีสูตรดังนี้

(Meng; Rosenthal; & Robin. 1992: 173)

$$Z = (Z_{r1} - Z_{r2}) \sqrt{\frac{N-3}{2(1-r_x)h}}$$

เมื่อ Z_r คือ คะแนนมาตรฐานที่แปลงมาจากค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันของพิชเชอร์

r คือ ค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน

r_x คือ ค่ามัธยฐาน (Median) ของค่าสัมประสิทธิ์สหสัมพันธ์

N คือ จำนวนผู้สอบในกลุ่มตัวอย่าง

k คือ จำนวนค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันที่ต้องการเปรียบเทียบ

บทที่ 4

ผลการวิเคราะห์ข้อมูล

การศึกษาวิจัยในครั้งนี้มีจุดมุ่งหมายเพื่อศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิด วิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน โดยใช้โมเดลการสรุปอ้างอิงตามทฤษฎีการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราล์ฟตามทฤษฎีการตอบข้อสอบ ผู้วิจัยศึกษาขนาดองค์ประกอบความแปรปรวนของแต่ละองค์ประกอบ พิจารณาคุณลักษณะของคะแนนจากค่าสัมประสิทธิ์การสรุปอ้างอิง ค่าความเชื่อมั่นแยกส่วน ค่าความเที่ยงตรงตามสภาพ และเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิง ค่าความเชื่อมั่นแยกส่วน ค่าความเที่ยงตรงตามสภาพ ซึ่งเป็นการเปรียบเทียบภายในแต่ละโมเดล ผู้วิจัยนำเสนอผลการวิเคราะห์ข้อมูลเป็น 3 ตอน ดังนี้

ตอนที่ 1 การวิเคราะห์ค่าสถิติพื้นฐานของคะแนนสอบ

ตอนที่ 2 การศึกษาคุณลักษณะของคะแนนโดยใช้โมเดลการสรุปอ้างอิง

ตอนที่ 3 การศึกษาคุณลักษณะของคะแนนโดยใช้โมเดลหลายองค์ประกอบของราล์ฟ

โดยในการนำเสนอผลการวิเคราะห์ข้อมูล เพื่อให้เกิดความเข้าใจในผลของการวิจัยตรงกัน ผู้วิจัยได้กำหนดสัญลักษณ์และอักษรย่อที่ใช้แทนค่าสถิติและตัวแปรต่าง ๆ ดังต่อไปนี้

N	แทน	ขนาดกลุ่มตัวอย่าง
n_r	แทน	จำนวนผู้ตรวจ
n_p	แทน	จำนวนผู้สอบ
n_i	แทน	จำนวนข้อสอบ
Max	แทน	คะแนนสูงสุด
Min	แทน	คะแนนต่ำสุด
SD	แทน	ส่วนเบี่ยงเบนมาตรฐาน
SE	แทน	ความคลาดเคลื่อนมาตรฐาน
S_k	แทน	ค่าความเบ้ของการแจกแจงความถี่ของคะแนน
K_u	แทน	ค่าความโด่งของการแจกแจงความถี่ของคะแนน
$\bar{X}$	แทน	ค่าเฉลี่ยเลขคณิตของคะแนนดิบ
P	แทน	ผู้สอบ
I	แทน	ข้อสอบ
R	แทน	ผู้ตรวจ

PI	แทน	ปฏิสัมพันธ์ระหว่างผู้สอบกับข้อสอบ
PR	แทน	ปฏิสัมพันธ์ระหว่างผู้สอบกับผู้ตรวจ
IR	แทน	ปฏิสัมพันธ์ระหว่างข้อสอบกับผู้ตรวจ
P:R	แทน	ผู้สอบในแต่ละผู้ตรวจ
I:R	แทน	ข้อสอบในแต่ละผู้ตรวจ
IP:R,e	แทน	ปฏิสัมพันธ์ระหว่างข้อสอบกับผู้สอบในแต่ละผู้ตรวจและความคลาดเคลื่อนเชิงสุ่มที่ไม่สามารถระบุแหล่งได้
PIR ,e	แทน	ปฏิสัมพันธ์ระหว่างผู้สอบ ข้อสอบกับผู้ตรวจและความคลาดเคลื่อนเชิงสุ่มที่ไม่สามารถระบุแหล่งได้
Pt	แทน	รูปแบบการตรวจให้คะแนน
Pt ₁	แทน	รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน
Pt ₂	แทน	รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน
Pt ₃	แทน	รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน
INFIT	แทน	ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนที่มีการถ่วงน้ำหนัก
OUTFIT	แทน	ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนที่ไม่มีการถ่วงน้ำหนัก

ตอนที่ 1 การวิเคราะห์ค่าสถิติพื้นฐานของคะแนนสอบ

ผู้วิจัยแบ่งการวิเคราะห์ในตอนนี้ออกเป็น 2 ประเด็นคือ 1) การวิเคราะห์ค่าสถิติพื้นฐานของคะแนนสอบ และ 2) การวิเคราะห์ค่าพารามิเตอร์ของผู้สอบ ข้อสอบและผู้ตรวจ ซึ่งการวิเคราะห์ดังกล่าวเป็นการอธิบายลักษณะของคะแนนดิบและคะแนนความสามารถ จากการตรวจให้คะแนนที่จำนวนผู้ตรวจไม่เท่ากัน คือ 4 คน 3 คน และ 2 คน ด้วยรูปแบบการตรวจให้คะแนนที่ต่างกัน คือ ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน มีรายละเอียดดังนี้

1. การวิเคราะห์ค่าสถิติพื้นฐานของคะแนนสอบ ผู้วิจัยตรวจสอบลักษณะโดยทั่วไปของข้อมูล ได้แก่ ค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐาน คะแนนสูงสุด คะแนนต่ำสุด ความเบ้และความโด่ง โดยได้จำแนกข้อมูลตามจำนวนผู้ตรวจให้คะแนน คือ 4 คน 3 คน และ 2 คน และรูปแบบการตรวจให้คะแนน คือ ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt₁) ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt₂) และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt₃) ปรากฏผลดังตาราง 8

ตาราง 8 ค่าสถิติพื้นฐานของคะแนนสอบ (N =180)

n_r	Pt	$\bar{X}$	SD	Max	Min	S_k	K_u
4	Pt ₁	27.588	11.086	46.750	9.000	0.037	-1.380
	Pt ₂	26.539	11.002	46.000	6.000	-0.055	-1.282
	Pt ₃	27.739	11.180	47.000	8.000	-0.012	-1.338
3	Pt ₁	27.506	11.374	47.000	8.333	0.012	-1.382
	Pt ₂	27.111	11.943	48.000	6.000	-0.052	-1.272
	Pt ₃	27.333	11.653	47.000	7.000	-0.031	-1.346
2	Pt ₁	27.211	11.399	46.500	8.500	0.010	-1.359
	Pt ₂	27.239	11.786	47.000	6.000	-0.065	-1.213
	Pt ₃	27.289	11.323	46.000	6.000	-0.017	-1.363

ผลการวิเคราะห์ดังตาราง 8 พบว่า

1.1 ผลการวิเคราะห์ค่าสถิติพื้นฐานของคะแนนสอบ เมื่อจำนวนผู้ตรวจ 4 คน โดยใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt₁) ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt₂) และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt₃) พบว่า คะแนนสอบเมื่อผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคนมีคะแนนเฉลี่ยสูงสุด คือ 27.739 คะแนน รองลงมาคือผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคนและผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน ซึ่งมีคะแนนเฉลี่ยเป็น 27.588 และ 26.539 คะแนน ตามลำดับ ส่วนเบี่ยงเบนมาตรฐานของคะแนนสอบพบว่า เมื่อผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคนมีค่าสูงสุด (Pt₃) รองลงมาคือ ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt₁) และผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt₂) ซึ่งมีค่าเป็น 11.180, 11.086 และ 11.002 ตามลำดับ แสดงว่า คะแนนเฉลี่ยของผู้สอบและส่วนเบี่ยงเบนมาตรฐานของคะแนนสอบในทุกรูปแบบการตรวจให้คะแนนเมื่อจำนวนผู้ตรวจ 4 คน มีค่าใกล้เคียงกัน

เมื่อพิจารณาลักษณะการแจกแจงของคะแนนสอบจากค่าความเบ้และความโด่ง พบว่ารูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt₁) มีลักษณะการแจกแจงแบบเบ้ขวา มีค่าเท่ากับ 0.037 ส่วนคะแนนสอบจากรูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt₂) และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt₃) มีการแจกแจงแบบเบ้ซ้าย มีค่าเท่ากับ -0.055 และ -0.012 ตามลำดับ ในส่วนของค่าความโด่ง พบว่า คะแนนจากรูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt₁) มีลักษณะการแจกแจงที่แบนมากที่สุด รองลงมาคือ ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt₃) และผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt₂)

ผู้สอบบางคน (Pt_2) มีค่าเท่ากับ -1.380, -1.338 และ -1.282 ตามลำดับ แสดงว่า คะแนนสอบที่ได้จากการตรวจให้คะแนนทั้งสามรูปแบบมีค่าความเบ้ต่างกันเพียงเล็กน้อยและมีลักษณะการแจกแจงที่คล้ายคลึงกัน คือ แบบกว่าโค้งปกติ

1.2 ผลการวิเคราะห์ค่าสถิติพื้นฐานของคะแนนสอบ เมื่อจำนวนผู้ตรวจ 3 คน โดยใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) พบว่า คะแนนสอบเมื่อผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) มีคะแนนเฉลี่ยสูงสุด คือ 27.506 คะแนน รองลงมาคือ ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) และผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) ซึ่งมีคะแนนเฉลี่ยเป็น 27.333, 27.111 คะแนน ตามลำดับ ส่วนเบี่ยงเบนมาตรฐานของคะแนนสอบ พบว่า เมื่อผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) มีค่าสูงสุด รองลงมาคือ ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) และผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) ซึ่งมีค่าเป็น 11.943, 11.653 และ 11.374 ตามลำดับ แสดงว่า คะแนนเฉลี่ยของผู้สอบและส่วนเบี่ยงเบนมาตรฐานของคะแนนสอบในทุกรูปแบบการตรวจให้คะแนนเมื่อจำนวนผู้ตรวจ 3 คน มีค่าใกล้เคียงกัน

เมื่อพิจารณาลักษณะการแจกแจงของคะแนนสอบจากค่าความเบ้และความโด่ง พบว่า รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) มีลักษณะการแจกแจงแบบเบ้ขวา มีค่าเท่ากับ 0.012 ส่วนคะแนนสอบจากการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) มีการแจกแจงแบบเบ้ซ้าย คือมีค่าเท่ากับ -0.052 และ -0.031 ตามลำดับ ในส่วนของค่าความโด่ง พบว่า คะแนนจากการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) มีลักษณะการแจกแจงที่แบนมากที่สุด รองลงมาคือ ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) และผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) ซึ่งมีค่าเป็น -1.382, -1.346 และ -1.272 ตามลำดับ แสดงว่า คะแนนจากการตรวจให้คะแนนทั้งสามรูปแบบมีลักษณะการแจกแจงที่คล้ายคลึงกันคือ แบบกว่าโค้งปกติ และมีค่าความเบ้ต่างกันเพียงเล็กน้อย

1.3 ผลการวิเคราะห์ค่าสถิติพื้นฐานของคะแนนสอบ เมื่อมีจำนวนผู้ตรวจ 2 คน โดยใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) พบว่า คะแนนสอบเมื่อผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) มีคะแนนเฉลี่ยสูงสุด คือ 27.289 คะแนน รองลงมาคือ ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) และผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) ซึ่งมีคะแนนเฉลี่ยเป็น 27.239, 27.211 คะแนน ตามลำดับ ส่วนเบี่ยงเบนมาตรฐาน

ของคะแนนสอบ พบว่า ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) มีค่าสูงสุด รองลงมาคือ ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) มีค่าเป็น 11.786, 11.399 และ 11.323 ตามลำดับ แสดงว่า คะแนนเฉลี่ยของผู้สอบและส่วนเบี่ยงเบนมาตรฐานของคะแนนสอบในทุกรูปแบบการตรวจให้คะแนนเมื่อจำนวนผู้ตรวจ 2 คน มีค่าใกล้เคียงกัน

เมื่อพิจารณาลักษณะการแจกแจงของคะแนนสอบจากค่าความเบ้และความโด่ง พบว่า รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) มีลักษณะการแจกแจงแบบเบ้ขวา มีค่าเท่ากับ 0.010 ส่วนคะแนนสอบจากการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) มีการแจกแจงแบบเบ้ซ้าย คือ มีค่าเท่ากับ -0.065 และ -0.017 ตามลำดับ ในส่วนของค่าความโด่ง พบว่า คะแนนจากการตรวจที่ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) มีลักษณะการแจกแจงที่แบนมากที่สุด รองลงมาคือ ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) และผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) ซึ่งมีค่าเท่ากับ -1.363, -1.359 และ -1.213 ตามลำดับ แสดงว่า คะแนนจากการตรวจให้คะแนนทั้งสามรูปแบบมีลักษณะการแจกแจงที่คล้ายคลึงกันคือ แบนกว่าโค้งปกติ และค่าความเบ้ต่างกันเล็กน้อย

เมื่อพิจารณาโดยภาพรวมแล้วพบว่า คะแนนเฉลี่ยของผู้สอบเมื่อมีจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกันมีค่าใกล้เคียงกันคือ มีค่าตั้งแต่ 26.539 ถึง 27.739 คะแนน ส่วนเบี่ยงเบนมาตรฐาน มีค่าตั้งแต่ 11.002 ถึง 11.934 แสดงว่า คะแนนเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานมีค่าใกล้เคียงกัน และค่าความเบ้ในรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) มีค่าเบ้ขวาในทุกจำนวนผู้ตรวจ ส่วนผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) นั้นมีลักษณะเบ้ซ้ายในทุกจำนวนผู้ตรวจ แต่อย่างไรก็ตามค่าความเบ้มีค่าต่างกันไม่มาก ในส่วนของค่าความโด่งนั้นทุกรูปแบบการตรวจให้คะแนนและจำนวนผู้ตรวจ มีลักษณะการแจกแจงของคะแนนที่คล้ายคลึงกันคือ มีการแจกแจงที่แบนกว่าปกติ

2. ผู้วิจัยนำคะแนนสอบที่ได้มาวิเคราะห์เพื่อหาค่าพารามิเตอร์ของผู้สอบ ผู้ตรวจและข้อสอบ ซึ่งแสดงถึง ค่าความสามารถของผู้สอบ ค่าความเข้มงวดของผู้ตรวจและค่าความยากของข้อสอบ ตามลำดับ ด้วยโมเดลหลายองค์ประกอบของราล์ซ โดยโปรแกรม FACETS (Linacre, 2009) จำแนกข้อมูลตามจำนวนผู้ตรวจให้คะแนน คือ 4 คน 3 คน และ 2 คน และรูปแบบการตรวจให้คะแนน คือ ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) ปรากฏผลดังตาราง 9

ตาราง 9 ค่าพารามิเตอร์ขององค์ประกอบ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน

องค์ประกอบ	η_r	4 คน			3 คน			2 คน		
	Pt	Pt ₁	Pt ₂	Pt ₃	Pt ₁	Pt ₂	Pt ₃	Pt ₁	Pt ₂	Pt ₃
ผู้สอบ	ค่าความสามารถ									
	Max	3.00	2.71	2.83	2.86	3.75	2.83	1.56	3.19	2.33
	Min	-1.20	-1.40	-1.44	-1.23	-1.29	-1.44	-1.22	-1.50	-1.27
	Mean	0.44	0.44	0.38	0.41	0.36	0.38	0.39	0.40	0.38
	SD	0.91	0.93	0.93	0.89	0.95	0.93	0.90	0.99	0.86
	INFIT									
	Mean	0.97	1.00	0.97	0.97	0.99	0.97	0.97	1.00	0.99
	SD	0.33	0.44	0.41	0.35	0.41	0.41	0.36	0.46	0.40
	OUTFIT									
	Mean	1.05	1.07	1.07	1.06	1.07	1.07	1.07	1.09	1.04
	SD	0.59	0.65	0.85	0.62	0.66	0.85	0.61	0.73	0.55
ผู้ตรวจ	ค่าความเข้มงวด									
	Max	0.16	0.38	0.47	0.14	0.26	0.47	0.47	0.22	0.06
	Min	-0.09	-0.28	-0.35	-0.09	-0.19	-0.35	-0.35	-0.22	-0.06
	Mean	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	SD	0.11	0.27	0.42	0.13	0.23	0.42	0.42	0.31	0.09
	INFIT									
	Mean	0.96	0.98	0.96	0.96	0.98	0.96	0.96	0.97	0.97
	SD	0.12	0.09	0.15	0.09	0.04	0.15	0.15	0.16	0.19
	OUTFIT									
	Mean	1.05	1.07	1.08	1.06	1.07	1.08	1.08	1.09	1.04
	SD	0.11	0.12	0.27	0.11	0.08	0.27	0.27	0.10	0.31
ข้อสอบ	ค่าความยาก									
	Max	0.85	0.86	0.79	0.84	0.85	0.57	0.81	0.86	1.05
	Min	-1.03	-1.04	-1.30	-1.00	-0.99	-0.84	-0.97	-1.03	-0.69
	Mean	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	SD	0.54	0.55	0.49	0.54	0.54	0.49	0.53	0.57	0.46
	INFIT									
	Mean	0.97	0.98	0.97	0.97	0.98	0.97	0.97	0.98	0.98
	SD	0.12	0.15	0.21	0.14	0.13	0.23	0.14	0.16	0.20
	OUTFIT									
	Mean	1.05	1.07	1.03	1.06	1.07	1.07	1.07	1.09	1.04
	SD	0.39	0.40	0.41	0.41	0.40	0.48	0.43	0.45	0.42

จากตาราง 9 พบว่า

1. **ค่าความสามารถของผู้สอบ (Examinee Ability)** เมื่อจำนวนผู้ตรวจ 4 คน 3 คน และ 2 คน ที่ใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) พบว่า ระดับความสามารถของผู้สอบมีค่าต่ำสุดอยู่ในช่วง -1.50 ถึง -1.20 และมีค่าสูงสุดอยู่ในช่วง 1.56 ถึง 3.75 ส่วนค่า Mean มีค่าตั้งแต่ 0.36 ถึง 0.44 และค่า SD มีค่าตั้งแต่ 0.86 ถึง 0.99 แสดงว่า ระดับความสามารถของผู้สอบแต่ละคนมีการกระจายแตกต่างกันไป ระดับความสามารถของผู้สอบโดยภาพรวมอยู่ในเกณฑ์ที่ยอมรับได้ คือ -3 ถึง +3 (ศิริชัย กาญจนวาสี. 2550: 55) และเมื่อพิจารณาเป็นรายบุคคล พบว่า มีผู้สอบที่มีค่าความสามารถเกินเกณฑ์ที่ยอมรับได้ คือ มีความสามารถเป็น 3.19 และ 3.75 เมื่อจำนวนผู้ตรวจ 2 คนและผู้ตรวจ 3 คน ที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน ตามลำดับ

การเปรียบเทียบความเหมาะสมของข้อมูลเชิงประจักษ์กับโมเดลหลายองค์ประกอบของราล์ช โดยใช้ Weighted mean square residual: INFIT และ Unweighted mean square residual: OUTFIT ในการทดสอบ ผลการวิเคราะห์ INFIT ของผู้สอบเมื่อพิจารณาโดยภาพรวม ค่า Mean อยู่ในช่วง 0.97 ถึง 1.00 (SD มีค่าตั้งแต่ 0.33 ถึง 0.46) และ OUTFIT ของผู้สอบมีค่า Mean อยู่ในช่วง 1.04 ถึง 1.07 (SD มีค่าตั้งแต่ 0.55 ถึง 0.85) ทั้งสองค่าอยู่ในเกณฑ์ที่ยอมรับได้ตามเกณฑ์ของลินซ์ ไวท์และลินาเคอร์ (Engelhard. 1994: 97; citing Lunz; Wright; & Linacre. 1990) ที่กำหนดขอบเขตของการยอมรับได้ของโมเดลที่เหมาะสมกับข้อมูลอยู่ระหว่าง 0.6 ถึง 1.5

2. **ค่าความเข้มงวดของผู้ตรวจ (Rater Severity)** เมื่อจำนวนผู้ตรวจ 4 คน 3 คนและ 2 คน และใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) พบว่า ค่าความเข้มงวดของผู้ตรวจในการตรวจให้คะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์มีค่าต่ำสุดอยู่ในช่วง -0.35 ถึง -0.06 ส่วนค่าความเข้มงวดสูงสุดอยู่ในช่วง 0.06 ถึง 0.47 โดยค่า Mean มีค่า 0.00 และค่า SD มีค่าตั้งแต่ 0.09 ถึง 0.42 แสดงว่า ค่าความเข้มงวดของผู้ตรวจมีการกระจายแตกต่างกัน แต่อย่างไรก็ตามเมื่อพิจารณาโดยภาพรวม ยังถือว่าผู้ตรวจแต่ละคนมีค่าความเข้มงวดใกล้เคียงกัน ไม่ต่างกันมากนัก

การเปรียบเทียบความเหมาะสมของข้อมูลเชิงประจักษ์กับโมเดลหลายองค์ประกอบของราล์ชโดยใช้ INFIT และ OUTFIT ในการทดสอบ ผลการวิเคราะห์ INFIT ของผู้ตรวจมีค่า Mean ตั้งแต่ 0.96 ถึง 0.98 (SD มีค่าตั้งแต่ 0.04 ถึง 0.19) และ OUTFIT ของผู้ตรวจมีค่า Mean ตั้งแต่ 1.04 ถึง 1.09 (SD มีค่าตั้งแต่ 0.08 ถึง 0.31) ทั้งสองค่าอยู่ในเกณฑ์ที่ยอมรับได้ตามเกณฑ์ของลินซ์ ไวท์และ

ลินาเคอร์ (Engelhard. 1994: 97; citing Lunz; Wright; & Linacre. 1990) ที่กำหนดขอบเขตการยอมรับได้ของโมเดลที่เหมาะสมกับข้อมูลอยู่ระหว่าง 0.6 ถึง 1.5 และผู้ตรวจทุกคนมีค่า INFIT และ OUTFIT เหมาะสมกับโมเดลทุกคน

2. ค่าความยากของข้อสอบ (Item Difficulty) เมื่อจำนวนผู้ตรวจ 4 คน 3 คน และ 2 คน และใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (P_{tj}) ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (P_{tj}) และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (P_{tj}) พบว่า ค่าความยากของข้อสอบมีค่าต่ำสุดอยู่ในช่วง -0.69 ถึง -1.30 และค่าความยากสูงสุดอยู่ในช่วง 0.57 ถึง 1.05 ส่วนค่า Mean มีค่า 0.00 และค่า SD มีค่าตั้งแต่ 0.46 ถึง 0.57 ทั้งนี้ในทุกเงื่อนไขจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนน ข้อสอบทุกข้อมีค่าความยากอยู่ในเกณฑ์ที่ยอมรับได้คืออยู่ระหว่าง -2.5 ถึง +2.5 (ศิริชัย กาญจนวาสี. 2550: 56)

การเปรียบเทียบความเหมาะสมของข้อมูลเชิงประจักษ์กับโมเดลหลายองค์ประกอบของราล์ชโดยใช้ INFIT และ OUTFIT ในการทดสอบ ผลการวิเคราะห์ INFIT ของข้อสอบมีค่า Mean ตั้งแต่ 0.97 ถึง 0.98 (SD มีค่า 0.12 ถึง 0.23) และ OUTFIT ของข้อสอบมีค่า Mean ตั้งแต่ 1.03 ถึง 1.09 (SD มีค่า 0.39 ถึง 0.48) ถ้าพิจารณาภาพรวมแล้ว ทั้งสองค่าอยู่ในเกณฑ์ที่ยอมรับได้ตามเกณฑ์ของลันซ์ ไรท์และลินาเคอร์ (Engelhard. 1994: 97; citing Lunz; Wright; & Linacre. 1990) ที่กำหนดขอบเขตของการยอมรับได้ของโมเดลที่เหมาะสมกับข้อมูลอยู่ระหว่าง 0.6 ถึง 1.5

ตอนที่ 2 การศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิด โดยใช้โมเดลการสรุปอ้างอิง

ผู้วิจัยนำผลคะแนนที่ได้มาวิเคราะห์โดยใช้โมเดลการสรุปอ้างอิง เพื่อศึกษาขนาดความแปรปรวนในแต่ละองค์ประกอบ ประมาณค่าและเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิง ประมาณค่าและเปรียบเทียบค่าความเที่ยงตรงตามสภาพ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน มีรายละเอียดดังนี้

1. การศึกษาขนาดความแปรปรวนในแต่ละองค์ประกอบ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน

ผู้วิจัยนำผลที่ได้มาวิเคราะห์ โดยใช้โปรแกรม GENOVA (Crick; & Brennan. 1983) เพื่อศึกษาขนาดความแปรปรวนในแต่ละองค์ประกอบ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน ผลการวิเคราะห์เป็นดังนี้

1.1 การประมาณค่าความแปรปรวนขององค์ประกอบ เมื่อมีผู้ตรวจจำนวน 4 คน และรูปแบบการตรวจให้คะแนนต่างกัน ผลเป็นดังนี้

1.1.1 การประมาณค่าความแปรปรวนขององค์ประกอบ เมื่อมีผู้ตรวจจำนวน 4 คน และตรวจข้อสอบทุกข้อของผู้สอบทุกคน

ตาราง 10 ค่าความแปรปรวนขององค์ประกอบ เมื่อมีผู้ตรวจจำนวน 4 คน และตรวจข้อสอบทุกข้อของผู้สอบทุกคน ($n_p=180$, $n_i=12$, $n_r=4$)

Source of Variation	DF	Sum of Square (SS)	Mean Square (MS)	Estimated Variance Component	% of Total Variance
P	179	53033.6875	40.9133	0.7709	36.5338
I	11	47853.5347	194.8486	0.2631	12.4686
R	3	45780.4625	23.4208	0.0010	0.0474
PI	1969	62621.2500	3.7807	0.9085	43.0548
PR	537	53252.7500	0.2771	0.0109	0.5166
IR	33	47982.6278	1.7828	0.0091	0.4313
PIR,e	5907	63765.0000	865.8568	0.1466	6.9475
Total	8639	374289.3125	-	2.1101	100.0000

จากตาราง 10 พบว่า ค่าความแปรปรวนของแหล่งต่าง ๆ เมื่อเทียบกับความแปรปรวนรวมเรียงจากมากไปหาน้อยได้แก่ ปฏิสัมพันธ์ระหว่างผู้สอบกับข้อสอบ (PI) มีค่า 0.9085 คิดเป็นร้อยละ 43.0548 ซึ่งแสดงว่า ผู้สอบแต่ละคนมีความไม่คงเส้นคงวาในการตอบข้อสอบแต่ละข้อหรือความแตกต่างของการตอบข้อสอบเปลี่ยนแปลงไปตามข้อสอบ รองลงมาคือ ความแปรปรวนของผู้สอบ (P) มีค่า 0.7709 ซึ่งแสดงถึง ความแตกต่างของคะแนนอันเนื่องมาจากความสามารถที่ต่างกันของผู้สอบแต่ละคน คิดเป็นร้อยละ 36.5338 และความแปรปรวนของข้อสอบ (I) มีค่า 0.2631 คิดเป็นร้อยละ 12.4686 ซึ่งแสดงถึง ความแตกต่างของค่าความยากของข้อสอบแต่ละข้อ ส่วนแหล่งความแปรปรวนที่มีค่าน้อยที่สุดคือ ความแปรปรวนของผู้ตรวจ (R) ซึ่งมีค่า 0.0010 คิดเป็นร้อยละ 0.0474 แสดงว่า ผู้ตรวจมีความแตกต่างในการตรวจให้คะแนนน้อย ดังนั้นแสดงว่า ความแตกต่างของคะแนนที่ได้เกือบครึ่งหนึ่งเนื่องมาจากปฏิสัมพันธ์ระหว่างผู้สอบกับข้อสอบ และเกิดจากผู้สอบแต่ละคนประมาณ 37 % ส่วนผู้ตรวจมีผลต่อคะแนนที่ได้น้อยมากหรือผู้ตรวจมีความแตกต่างกันในการตรวจให้คะแนนน้อยมากไม่ถึง 0.05 %

1.1.2 การประมาณค่าความแปรปรวนขององค์ประกอบ เมื่อมีผู้ตรวจจำนวน 4 คน และตรวจข้อสอบทุกข้อของผู้สอบบางคน

ตาราง 11 ค่าความแปรปรวนขององค์ประกอบ เมื่อมีผู้ตรวจจำนวน 4 คน และตรวจข้อสอบทุกข้อของผู้สอบบางคน ($n_p=180$, $n_i=12$, $n_r=4$)

Source of Variation	DF	Sum of Square (SS)	Mean Square (MS)	Estimated Variance Component	% of Total Variance
I	11	11875.4333	50.0008	0.2695	12.3545
R	3	11447.4333	40.6698	0.0553	2.5350
P:R	176	13276.5000	10.3924	0.7768	35.6102
IR	33	12046.8000	1.4957	0.0095	0.4355
IP:R,e	1936	15948.0000	1.0703	1.0703	49.0648
Total	2159	64594.1666	-	2.1814	100.0000

จากตาราง 11 พบว่า ค่าความแปรปรวนของแหล่งต่าง ๆ เมื่อเทียบกับความแปรปรวนรวมเรียงจากมากไปหาน้อยได้แก่ ปฏิสัมพันธ์ระหว่างข้อสอบกับผู้สอบในแต่ละผู้ตรวจและความคลาดเคลื่อนเชิงสุ่มที่ไม่สามารถระบุได้ (IP:R,e) มีค่า 1.0703 คิดเป็นร้อยละ 49.0648 รองลงมาคือความแปรปรวนผู้สอบในแต่ละผู้ตรวจ(P:R) มีค่า 0.7768 คิดเป็นร้อยละ 35.6102 แสดงว่า มีความแตกต่างของคะแนนอันเนื่องมาจากความสามารถที่ต่างกันของผู้สอบในแต่ละผู้ตรวจ และความแปรปรวนของข้อสอบ(I) มีค่า 0.2695 คิดเป็นร้อยละ 12.3545 แสดงถึง ความแตกต่างของค่าความยากของข้อสอบแต่ละข้อ ส่วนความแปรปรวนของผู้ตรวจ (R) มีค่า 0.0553 คิดเป็นร้อยละ 2.5350 แสดงว่า ผู้ตรวจมีความแตกต่างกันในการให้คะแนนค่อนข้างน้อย และความแปรปรวนของปฏิสัมพันธ์ระหว่างข้อสอบกับผู้ตรวจ(IR) มีค่าเป็น 0.0095 คิดเป็นร้อยละ 0.4355 แสดงให้เห็นถึง ความแตกต่างของคะแนนของข้อสอบแต่ละข้อที่ได้จากการตรวจให้คะแนนของผู้ตรวจแต่ละคน ซึ่งถือว่ามีความแตกต่างกันน้อยที่สุด นั่นคือ ความแตกต่างของคะแนนที่ได้เกือบครึ่งหนึ่งเนื่องมาจากปฏิสัมพันธ์ระหว่างข้อสอบกับผู้สอบในแต่ละผู้ตรวจและความคลาดเคลื่อนเชิงสุ่มที่ไม่สามารถระบุได้ และประมาณ 36 % เกิดจากผู้สอบในแต่ละผู้ตรวจ ส่วนปฏิสัมพันธ์ระหว่างข้อสอบกับผู้ตรวจและผู้ตรวจส่งผลต่อความแตกต่างของคะแนนน้อยมากรวมกันไม่ถึง 3 %

1.1.3 การประมาณค่าความแปรปรวนขององค์ประกอบ เมื่อมีผู้ตรวจจำนวน 4 คน และตรวจข้อสอบบางข้อของผู้สอบทุกคน

ตาราง 12 ค่าความแปรปรวนขององค์ประกอบ เมื่อมีผู้ตรวจจำนวน 4 คน และตรวจข้อสอบบางข้อของผู้สอบทุกคน ($n_p=180$, $n_i=12$, $n_r=4$)

Source of Variation	DF	Sum of Square (SS)	Mean Square (MS)	Estimated Variance Component	% of Total Variance
P	179	13405.9167	10.4147	0.7656	35.8209
R	3	11684.4793	47.5968	0.0000	0.0000
I:R	8	12073.0056	48.5657	0.2640	12.3520
PR	537	14207.6667	1.2271	0.0597	2.7933
PI:R,e	1432	16097.0000	1.0480	1.0480	49.0338
Total	2159	67468.0683	-	2.1373	100.000

จากตาราง 12 พบว่า ค่าความแปรปรวนของแหล่งต่าง ๆ เมื่อเทียบกับความแปรปรวนรวมเรียงจากมากไปหาน้อยได้แก่ ปฏิสัมพันธ์ระหว่างผู้สอบกับข้อสอบในแต่ละผู้ตรวจและความคลาดเคลื่อนเชิงสุ่มที่ไม่สามารถระบุได้ (PI:R,e) มีค่า 1.0480 คิดเป็นร้อยละ 49.0338 รองลงมาคือความแปรปรวนผู้สอบ (P) มีค่า 0.7656 คิดเป็นร้อยละ 35.8209 แสดงถึง ความแตกต่างของคะแนนอันเนื่องมาจากความสามารถที่ต่างกันของผู้สอบแต่ละคน และความแปรปรวนของข้อสอบในแต่ละผู้ตรวจ (I:R) มีค่า 0.2640 คิดเป็นร้อยละ 12.3520 แสดงถึง ความแตกต่างของความยากของข้อสอบในแต่ละผู้ตรวจ ส่วนแหล่งความแปรปรวนที่มีค่าน้อยที่สุดคือ ผู้ตรวจ (R) ซึ่งมีค่าเป็น 0.0000 แสดงว่า ผู้ตรวจไม่มีความแตกต่างกันในการตรวจให้คะแนน นั่นคือ ความแตกต่างของคะแนนที่ได้เกือบ 50 % เนื่องมาจากปฏิสัมพันธ์ระหว่างผู้สอบกับข้อสอบในแต่ละผู้ตรวจและความคลาดเคลื่อนเชิงสุ่มที่ไม่สามารถระบุได้ และประมาณ 36 % เกิดจากผู้สอบแต่ละคน ส่วนผู้ตรวจไม่มีความแตกต่างกันในการตรวจให้คะแนน

1.2. การประมาณค่าความแปรปรวนขององค์ประกอบ เมื่อมีผู้ตรวจจำนวน 3 คน และรูปแบบการตรวจให้คะแนนต่างกัน ดังนี้

1.2.1 การประมาณค่าความแปรปรวนขององค์ประกอบ เมื่อมีผู้ตรวจจำนวน 3 คน และตรวจข้อสอบทุกข้อของผู้สอบทุกคน

ตาราง 13 ค่าความแปรปรวนขององค์ประกอบ เมื่อมีผู้ตรวจจำนวน 3 คน และตรวจข้อสอบทุกข้อของผู้สอบทุกคน ($n_p=180$, $n_i=12$, $n_r=3$)

Source of Variation	DF	Sum of Square (SS)	Mean Square (MS)	Estimated Variance Component	% of Total Variance
P	179	39834.2500	32.3422	0.8111	36.2860
I	11	35785.1093	158.1916	0.2846	12.7321
R	2	34113.9607	34.4796	0.0152	0.6800
PI	1969	47615.6667	3.0682	0.9791	43.8017
PR	358	39977.0833	0.2064	0.0063	0.2818
IR	22	35889.4389	1.6077	0.0082	0.3668
PIR,e	3938	48309.0000	0.1308	0.1308	5.8516
Total	6479	281524.5089	-	2.2353	100.0000

จากตาราง 13 พบว่า ค่าความแปรปรวนของแหล่งต่าง ๆ เมื่อเทียบกับความแปรปรวนรวมเรียงจากมากไปหาน้อยได้แก่ ปฏิสัมพันธ์ระหว่างผู้สอบกับข้อสอบ (PI) มีค่า 0.9791 แสดงว่า ผู้สอบแต่ละคนมีความไม่คงเส้นคงวาในการตอบข้อสอบแต่ละข้อ คิดเป็นร้อยละ 43.8017 รองลงมาคือ ความแปรปรวนของผู้สอบ (P) มีค่า 0.8111 คิดเป็นร้อยละ 36.2860 แสดงถึง ความแตกต่างของคะแนนอันเนื่องมาจากความสามารถที่ต่างกันของผู้สอบแต่ละคน และความแปรปรวนของข้อสอบ (I) มีค่า 0.2846 คิดเป็นร้อยละ 12.7321 แสดงถึง ความแตกต่างของความยากของข้อสอบ ส่วนความแปรปรวนของผู้ตรวจ (R) ปฏิสัมพันธ์ระหว่างข้อสอบกับผู้ตรวจ (IR) ปฏิสัมพันธ์ระหว่างผู้สอบกับผู้ตรวจ (PR) มีค่าเป็น 0.0152, 0.0082 และ 0.0063 คิดเป็นร้อยละ 0.6800, 0.3668 และ 0.2818 ตามลำดับ แสดงว่า ความแตกต่างของการตรวจให้คะแนนของผู้ตรวจ ความแตกต่างของคะแนนของข้อสอบแต่ละข้อที่ได้จากการตรวจให้คะแนนของผู้ตรวจแต่ละคน และความแตกต่างของความสามารถของผู้สอบแต่ละคนที่ได้จากการตรวจให้คะแนนของผู้ตรวจแต่ละคน มีค่าค่อนข้างน้อย นั่นคือ ความแตกต่างของคะแนนประมาณ 44 % เนื่องมาจากปฏิสัมพันธ์ระหว่างผู้สอบกับข้อสอบ และประมาณ 36 % เกิดจากผู้สอบแต่ละคน ส่วนความแตกต่างของคะแนนที่เกิดจากผู้ตรวจ ปฏิสัมพันธ์ระหว่างข้อสอบกับผู้ตรวจและปฏิสัมพันธ์ระหว่างผู้สอบกับผู้ตรวจ มีค่าไม่มากนัก

1.2.2 การประมาณค่าความแปรปรวนขององค์ประกอบ เมื่อมีผู้ตรวจจำนวน 3 คน และตรวจข้อสอบทุกข้อของผู้สอบบางคน

ตาราง 14 ค่าความแปรปรวนขององค์ประกอบ เมื่อมีผู้ตรวจจำนวน 3 คน และตรวจ ข้อสอบทุกข้อของผู้สอบบางคน ($n_p=180$, $n_i=12$, $n_r=3$)

Source of Variation	DF	Sum of Square (SS)	Mean Square (MS)	Estimated Variance Component	% of Total Variance
I	11	11629.6556	54.9519	0.2978	12.6734
R	2	11123.3083	49.0616	0.0519	2.2087
P:R	177	13152.6667	11.4653	0.8608	36.6329
IR	22	11757.4667	1.3495	0.0036	0.1532
IP:R,e	1947	15998.0000	1.1357	1.1357	48.3318
Total	2159	63661.0973	-	2.3498	100.0000

จากตาราง 14 พบว่า ค่าความแปรปรวนของแหล่งต่าง ๆ เมื่อเทียบกับความแปรปรวนรวมเรียงจากมากไปหาน้อยได้แก่ ปฏิสัมพันธ์ระหว่างข้อสอบกับผู้สอบในแต่ละผู้ตรวจและความคลาดเคลื่อนเชิงสุ่มที่ไม่สามารถระบุได้ (IP:R,e) มีค่า 1.1357 คิดเป็นร้อยละ 48.3318 รองลงมาคือความแปรปรวนของผู้สอบในแต่ละผู้ตรวจ (P:R) มีค่า 0.8608 คิดเป็นร้อยละ 36.6329 แสดงว่ามีความแตกต่างของคะแนนอันเนื่องมาจากความสามารถที่ต่างกันของผู้สอบในแต่ละผู้ตรวจ และความแปรปรวนของข้อสอบ (I) มีค่า 0.2978 คิดเป็นร้อยละ 12.6734 แสดงถึงความแตกต่างของความยากของข้อสอบแต่ละข้อ ส่วนความแปรปรวนของผู้ตรวจ (R) และความแปรปรวนของปฏิสัมพันธ์ระหว่างข้อสอบกับผู้ตรวจ (IR) มีค่าค่อนข้างน้อย คือมีค่าเป็น 0.0519 และ 0.0036 คิดเป็นร้อยละ 2.2087 และ 0.1532 ตามลำดับ แสดงว่า ผู้ตรวจมีความแตกต่างกันน้อยในการตรวจให้คะแนนและความแตกต่างของคะแนนของข้อสอบแต่ละข้อที่ได้จากการตรวจให้คะแนนของผู้ตรวจแต่ละคนมีค่าค่อนข้างน้อย นั่นคือ ประมาณ 50 % ของความแตกต่างของคะแนน เนื่องมาจากปฏิสัมพันธ์ระหว่างข้อสอบกับผู้สอบในแต่ละผู้ตรวจและความคลาดเคลื่อนเชิงสุ่มที่ไม่สามารถระบุได้ และประมาณ 37% เกิดจากผู้สอบในแต่ละผู้ตรวจ ส่วนความแตกต่างของคะแนนที่เกิดจากผู้ตรวจ ปฏิสัมพันธ์ระหว่างข้อสอบกับผู้ตรวจมีค่าน้อยรวมกันไม่ถึง 3 %

1.2.3 การประมาณค่าความแปรปรวนขององค์ประกอบ เมื่อมีผู้ตรวจจำนวน 3 คน และตรวจข้อสอบบางข้อของผู้สอบทุกคน

ตาราง 15 ค่าความแปรปรวนขององค์ประกอบ เมื่อมีผู้ตรวจจำนวน 3 คน และตรวจข้อสอบบางข้อของผู้สอบทุกคน ($n_p=180$, $n_i=12$, $n_r=3$)

Source of Variation	DF	Sum of Square (SS)	Mean Square (MS)	Estimated Variance Component	% of Total Variance
P	179	13232.3333	11.3166	0.8492	36.5294
R	2	11424.1361	108.7347	0.0829	3.5661
I:R	9	11865.7889	49.0725	0.2664	11.4595
PR	358	13838.0000	1.0844	0.0000	0.0000
PI:R,e	1611	16094.0000	1.1262	1.1262	48.4450
Total	2159	66454.2583	-	2.3247	100.0000

จากตาราง 15 พบว่า ค่าความแปรปรวนของแหล่งต่าง ๆ เมื่อเทียบกับความแปรปรวนรวมเรียงจากมากไปหาน้อยได้แก่ ปฏิสัมพันธ์ระหว่างผู้สอบกับข้อสอบในแต่ละผู้ตรวจและความคลาดเคลื่อนเชิงสุ่มที่ไม่สามารถระบุได้ (PI:R,e) มีค่า 1.1262 คิดเป็นร้อยละ 48.4450 รองลงมาคือความแปรปรวนของผู้สอบ (P) มีค่า 0.8492 แสดงว่า ความแตกต่างของคะแนนอันเนื่องมาจากความสามารถที่ต่างกันของผู้สอบแต่ละคน คิดเป็นร้อยละ 36.5294 และความแปรปรวนของข้อสอบในแต่ละผู้ตรวจ (I:R) มีค่า 0.2664 คิดเป็นร้อยละ 11.4595 แสดงถึง ความแตกต่างของความยากของข้อสอบในแต่ละผู้ตรวจ ส่วนแหล่งความแปรปรวนที่มีค่าน้อย คือ ผู้ตรวจ (R) และปฏิสัมพันธ์ระหว่างผู้สอบกับผู้ตรวจ (PR) มีค่าเป็น 0.0829 และ 0.0000 คิดเป็นร้อยละ 3.5661 และ 0.0000 ตามลำดับ แสดงว่า ผู้ตรวจมีความแตกต่างกันในการให้คะแนนค่อนข้างน้อย และไม่มี ความแตกต่างกันของความสามารถของผู้สอบแต่ละคนที่ได้จากการตรวจให้คะแนนของผู้ตรวจแต่ละคน นั่นคือ ความแตกต่างของคะแนนที่ได้เกือบ 50 % เนื่องมาจากปฏิสัมพันธ์ระหว่างผู้สอบกับข้อสอบในแต่ละผู้ตรวจ และความคลาดเคลื่อนเชิงสุ่มที่ไม่สามารถระบุได้ และประมาณ 37 % เกิดจากผู้สอบแต่ละคน ส่วนผู้ตรวจส่งผลถึงความแตกต่างของคะแนนไม่ถึง 4 % และปฏิสัมพันธ์ระหว่างผู้สอบกับผู้ตรวจ ไม่ส่งผลต่อความแตกต่างของคะแนน

1.3 การประมาณค่าความแปรปรวนขององค์ประกอบ เมื่อมีผู้ตรวจจำนวน 2 คน และมีรูปแบบการตรวจให้คะแนนต่างกัน ดังนี้

1.3.1 การประมาณค่าความแปรปรวนขององค์ประกอบ เมื่อมีผู้ตรวจจำนวน 2 คน และตรวจข้อสอบทุกข้อของผู้สอบทุกคน

ตาราง 16 ค่าความแปรปรวนขององค์ประกอบ เมื่อมีผู้ตรวจจำนวน 2 คน และตรวจข้อสอบทุกข้อของผู้สอบทุกคน ($n_p=180$, $n_i=12$, $n_r=2$)

Source of Variation	DF	Sum of Square (SS)	Mean Square (MS)	Estimated Variance Component	% of Total Variance
P	179	26089.9167	21.6569	0.8128	36.5057
I	11	23324.0389	100.9729	0.2691	12.0862
R	1	22274.4935	61.1565	0.0273	1.2261
PI	1969	31257.0000	2.0601	0.9617	43.1934
PR	179	26191.6667	0.2268	0.0075	0.3369
IR	11	23409.0778	2.1711	0.0113	0.5075
PIR,e	1969	31652.0000	0.1368	0.1368	6.1442
Total	4319	184198.1936	-	2.2265	100.000

จากตาราง 16 พบว่า ค่าความแปรปรวนของแหล่งต่าง ๆ เมื่อเทียบกับความแปรปรวนรวมเรียงจากมากไปหาน้อยได้แก่ ปฏิสัมพันธ์ระหว่างผู้สอบกับข้อสอบ (PI) มีค่า 0.9617 คิดเป็นร้อยละ 43.1934 แสดงว่า ผู้สอบแต่ละคนมีความไม่คงเส้นคงวาในการตอบข้อสอบแต่ละข้อ รองลงมาคือ ความแปรปรวนของผู้สอบ (P) มีค่า 0.8128 คิดเป็นร้อยละ 36.5057 แสดงถึง ความแตกต่างของความสามารถของผู้สอบแต่ละคน และความแปรปรวนของข้อสอบ (I) มีค่า 0.2691 คิดเป็นร้อยละ 12.0862 แสดงถึง ความแตกต่างของค่าความยากของข้อสอบแต่ละข้อ ส่วนแหล่งความแปรปรวนที่มีค่าน้อย คือ ความแปรปรวนของผู้ตรวจ (R) ปฏิสัมพันธ์ระหว่างข้อสอบกับผู้ตรวจ (IR) และ ปฏิสัมพันธ์ระหว่างผู้สอบกับผู้ตรวจ (PR) โดยมีค่าเป็น 0.0273, 0.0113 และ 0.0075 คิดเป็นร้อยละ 1.2261, 0.5075 และ 0.3369 ตามลำดับ แสดงว่า ผู้ตรวจมีความแตกต่างกันในการตรวจให้คะแนนไม่มากนัก ความแตกต่างของคะแนนของข้อสอบแต่ละข้อที่ได้จากการตรวจให้คะแนนของผู้ตรวจแต่ละคนและความแตกต่างของความสามารถของผู้สอบแต่ละคนที่ได้จากผู้ตรวจแต่ละคนมีค่าน้อย นั่นคือ ความแตกต่างของคะแนนที่ได้เกือบครึ่งหนึ่งเนื่องมาจากปฏิสัมพันธ์ระหว่างผู้สอบกับข้อสอบ และ

ประมาณ 37 % เกิดจากผู้สอบแต่ละคน ส่วนความแตกต่างของคะแนนอันเนื่องมาจากผู้ตรวจแต่ละคน ปฏิสัมพันธ์ระหว่างข้อสอบกับผู้ตรวจและปฏิสัมพันธ์ระหว่างผู้สอบกับผู้ตรวจ รวมกันประมาณ 2 % ซึ่งถือว่าค่อนข้างน้อย

1.3.2 การประมาณค่าความแปรปรวนขององค์ประกอบ เมื่อมีผู้ตรวจจำนวน 2 คน และตรวจข้อสอบทุกข้อของผู้สอบบางคน

ตาราง 17 ค่าความแปรปรวนขององค์ประกอบ เมื่อมีผู้ตรวจจำนวน 2 คน และตรวจข้อสอบทุกข้อของผู้สอบบางคน ($n_p=180$, $n_i=12$, $n_r=2$)

Source of Variation	DF	Sum of Square (SS)	Mean Square (MS)	Estimated Variance Component	% of Total Variance
I	11	11699.9611	51.8732	0.2809	12.3925
R	1	11239.1565	109.8005	0.0912	4.0235
P:R	178	13201.2500	11.0230	0.8301	36.6215
IR	11	11824.2778	1.3197	0.0029	0.1279
IP:R,e	1958	15865.0000	1.0617	1.0616	46.8346
Total	2159	63829.6454	-	2.2667	100.0000

จากตาราง 17 พบว่า ค่าความแปรปรวนของแหล่งต่าง ๆ เมื่อเทียบกับความแปรปรวนรวมเรียงจากมากไปหาน้อยได้แก่ ปฏิสัมพันธ์ระหว่างข้อสอบกับผู้สอบในแต่ละผู้ตรวจและความคลาดเคลื่อนเชิงสุ่มที่สามารถระบุได้ (IP:R,e) มีค่า 1.0616 คิดเป็นร้อยละ 46.8346 รองลงมาคือความแปรปรวนของผู้สอบในแต่ละผู้ตรวจ (P:R) มีค่า 0.8301 คิดเป็นร้อยละ 36.6215 แสดงว่ามีความแตกต่างของคะแนนอันเนื่องมาจากความสามารถที่ต่างกันของผู้สอบในแต่ละผู้ตรวจ และความแปรปรวนของข้อสอบ (I) มีค่า 0.2809 คิดเป็นร้อยละ 12.3925 แสดงถึงความแตกต่างของความยากของข้อสอบแต่ละข้อ ส่วนแหล่งความแปรปรวนที่มีค่าน้อยคือ ความแปรปรวนของผู้ตรวจ (R) และความแปรปรวนของปฏิสัมพันธ์ระหว่างข้อสอบกับผู้ตรวจ (IR) มีค่าเป็น 0.0912, 0.0029 คิดเป็นร้อยละ 4.0235 และ 0.1279 ตามลำดับ แสดงว่า ผู้ตรวจมีความแตกต่างกันในการตรวจให้คะแนนไม่มากนัก และความแตกต่างของคะแนนของข้อสอบแต่ละข้อที่ได้จากการตรวจให้คะแนนของผู้ตรวจมีค่าน้อย นั่นคือ ความแตกต่างของคะแนนที่ได้ประมาณ 47 % เนื่องมาจากปฏิสัมพันธ์ระหว่างข้อสอบกับผู้สอบในแต่ละผู้ตรวจและความคลาดเคลื่อนเชิงสุ่มที่ไม่สามารถระบุได้ และประมาณ 37

% เกิดจากผู้สอบในแต่ละผู้ตรวจ และความแตกต่างของคะแนนอันเนื่องมาจากผู้ตรวจ ปฏิสัมพันธ์ระหว่างข้อสอบกับผู้ตรวจส่งผลรวมกันไม่ถึง 5 % ซึ่งถือว่าค่อนข้างน้อย

1.3.3 การประมาณค่าความแปรปรวนขององค์ประกอบ เมื่อมีผู้ตรวจจำนวน 2 คน และตรวจข้อสอบบางข้อของผู้สอบทุกคน

ตาราง 18 ค่าความแปรปรวนขององค์ประกอบ เมื่อมีผู้ตรวจจำนวน 2 คน และตรวจข้อสอบบางข้อของผู้สอบทุกคน ($n_p=180$, $n_i=12$, $n_r=2$)

Source of Variation	DF	Sum of Square (SS)	Mean Square (MS)	Estimated Variance Component	% of Total Variance
P	179	13082.8333	10.6848	0.7958	36.8921
R	1	11206.0315	35.7796	0.0000	0.0000
I:R	10	11623.6333	41.7602	0.2257	10.4631
PR	179	13313.6667	1.0897	0.0000	0.0000
PI:R,e	1790	15764.0000	1.1356	1.1356	52.6448
Total	2159	64990.1648	-	2.1571	100.0000

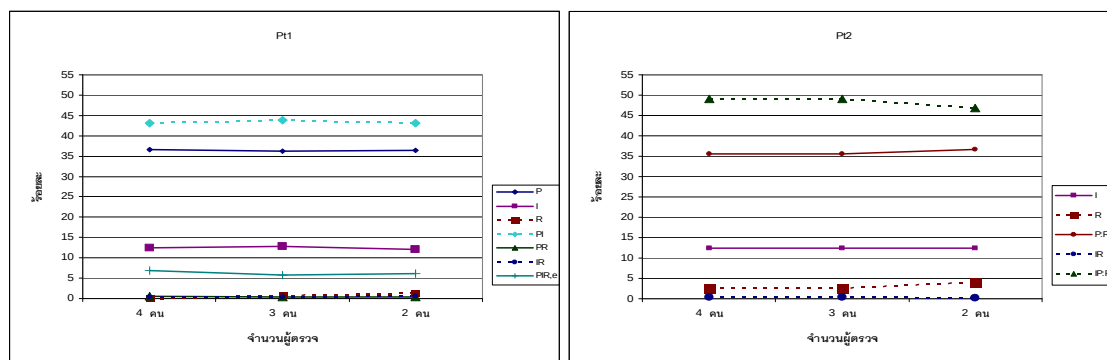
จากตาราง 18 พบว่า ค่าความแปรปรวนของแหล่งต่าง ๆ เมื่อเทียบกับความแปรปรวนรวมเรียงจากมากไปหาน้อยได้แก่ ปฏิสัมพันธ์ระหว่างผู้สอบกับข้อสอบในแต่ละผู้ตรวจและความคลาดเคลื่อนเชิงสุ่มที่สามารถระบุได้ (PI:R,e) มีค่า 1.1356 คิดเป็น 52.6448 รองลงมาคือ ความแปรปรวนของผู้สอบ (P) มีค่า 0.7958 คิดเป็นร้อยละ 36.8921 แสดงว่า ความแตกต่างของคะแนนอันเนื่องมาจากความสามารถที่ต่างกันของผู้สอบแต่ละคน และความแปรปรวนของข้อสอบในแต่ละผู้ตรวจ (I:R) มีค่า 0.2257 คิดเป็นร้อยละ 10.4631 แสดงถึง ความแตกต่างของความยากของข้อสอบในแต่ละผู้ตรวจ ส่วนความแปรปรวนของผู้ตรวจ (R) และความแปรปรวนของปฏิสัมพันธ์ระหว่างผู้สอบกับผู้ตรวจ (RR) มีค่า 0.0000 แสดงว่า ผู้ตรวจไม่มีความแตกต่างกันในการตรวจให้คะแนน และความสามารถของผู้สอบแต่ละคนที่ได้จากการตรวจให้คะแนนของผู้ตรวจแต่ละคนไม่แตกต่างกัน นั่นคือ ความแตกต่างของคะแนนที่ได้เนื่องมาจากปฏิสัมพันธ์ระหว่างผู้สอบกับข้อสอบในแต่ละผู้ตรวจและความคลาดเคลื่อนเชิงสุ่มที่ไม่สามารถระบุได้ประมาณ 53 % และเกิดจากผู้สอบแต่ละคนประมาณ 37 % ส่วนผู้ตรวจและปฏิสัมพันธ์ระหว่างผู้สอบกับผู้ตรวจ ไม่ส่งผลต่อความแตกต่างของคะแนน

ทั้งนี้ สามารถสรุปโดยภาพรวมของค่าความแปรปรวนในแต่ละองค์ประกอบ เมื่อรูปแบบการตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจต่างกัน ผลปรากฏดังตาราง 19

ตาราง 19 ขนาดความแปรปรวนของแต่ละองค์ประกอบ เมื่อรูปแบบการตรวจให้คะแนนเหมือนกัน และจำนวนผู้ตรวจต่างกัน

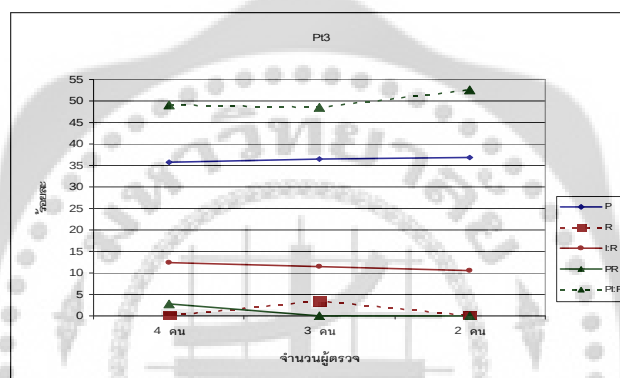
รูปแบบการตรวจ	แหล่งความแปรปรวน	จำนวนผู้ตรวจ					
		4 คน		3 คน		2 คน	
		ความแปรปรวน	% ความแปรปรวน	ความแปรปรวน	% ความแปรปรวน	ความแปรปรวน	% ความแปรปรวน
Pt ₁ (ตรวจข้อสอบทุกข้อของผู้สอบทุกคน)	P	0.7709	36.5338	0.8111	36.2860	0.8128	36.5057
	I	0.2631	12.4686	0.2846	12.7321	0.2691	12.0862
	R	0.0010	0.0474	0.0152	0.6800	0.0273	1.2261
	PI	0.9085	43.0548	0.9791	43.8017	0.9617	43.1934
	PR	0.0109	0.5166	0.0063	0.2818	0.0075	0.3369
	IR	0.0091	0.4313	0.0082	0.3668	0.0113	0.5075
	PIR,e	0.1466	6.9475	0.1308	5.8516	0.1368	6.1442
Total		2.1101	100.0000	2.2353	100.0000	2.2265	100.000
Pt ₂ (ตรวจข้อสอบทุกข้อของผู้สอบบางคน)	I	0.2695	12.3545	0.2978	12.3545	0.2809	12.3925
	R	0.0553	2.5350	0.0519	2.5350	0.0912	4.0235
	P:R	0.7768	35.6102	0.8608	35.6102	0.8301	36.6215
	IR	0.0095	0.4355	0.0036	0.4355	0.0029	0.1279
	IP:R,e	1.0703	49.0648	1.1357	49.0648	1.0616	46.8346
Total		2.1814	100.0000	2.3498	100.0000	2.2667	100.0000
Pt ₃ (ตรวจข้อสอบบางข้อของผู้สอบทุกคน)	P	0.7656	35.8209	0.8492	36.5294	0.7958	36.8921
	R	0.0000	0.0000	0.0829	3.5661	0.0000	0.0000
	I:R	0.2640	12.3520	0.2664	11.4595	0.2257	10.4631
	PR	0.0597	2.7933	0.0000	0.0000	0.0000	0.0000
	PI:R,e	1.0480	49.0338	1.1262	48.4450	1.1356	52.6448
Total		2.1373	100.000	2.3247	100.0000	2.1571	100.0000

และนำผลการวิเคราะห์ตามตาราง 19 มาเขียนแผนภูมิได้ดังภาพประกอบ 12(ก-ค) ดังนี้



12ก

12ข



12ค

ภาพประกอบ 12(ก-ค) ร้อยละของความแปรปรวนของแต่ละองค์ประกอบเมื่อรูปแบบการตรวจให้คะแนนเหมือนกัน (12ก) ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (12ข) ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน และ (12ค) ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน และจำนวนผู้ตรวจต่างกัน

จากตาราง 19 และภาพประกอบ 12(ก-ค) พบว่า

เมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) และมีจำนวนผู้ตรวจ 4 คน 3 คน และ 2 คน พบว่า องค์ประกอบที่มีความแปรปรวนมากที่สุดในทุกจำนวนผู้ตรวจเมื่อเทียบกับความแปรปรวนรวม คือ ปฏิสัมพันธ์ระหว่างผู้สอบกับข้อสอบ (PI) มีค่าตั้งแต่ 0.9085 ถึง 0.9791 คิดเป็นร้อยละ 43.0548 - 43.8017 รองลงมาคือ ความแปรปรวนของผู้สอบ (P) มีค่าตั้งแต่ 0.7709 ถึง 0.8111 คิดเป็นร้อยละ 36.2860-36.5057 และความแปรปรวนของข้อสอบ (I) มีค่าตั้งแต่ 0.0010 ถึง 0.0273 คิดเป็นร้อยละ 12.0862-12.7321 ส่วนแหล่งความแปรปรวนที่มีค่าน้อยคือ ความแปรปรวนของผู้ตรวจ (R) มีค่าตั้งแต่ 0.0010 - 0.0152 คิดเป็นร้อยละ 0.0474 - 1.2261 ความแปรปรวนของปฏิสัมพันธ์ระหว่างผู้สอบกับผู้ตรวจ (PR) มีค่าตั้งแต่ 0.0063 ถึง 0.0109

คิดเป็นร้อยละ 0.2818 – 0.5166 ความแปรปรวนของปฏิสัมพันธ์ระหว่างข้อสอบกับผู้ตรวจ (IR) มีค่าตั้งแต่ 0.0082 ถึง 0.0113 คิดเป็นร้อยละ 0.3668 – 0.5075 แสดงว่า ความแตกต่างของคะแนนสอบเมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) ในทุกจำนวนผู้ตรวจเกิดจากปฏิสัมพันธ์ระหว่างผู้สอบกับข้อสอบมากที่สุด รองลงมาคือ ผู้สอบและข้อสอบ ตามลำดับ ส่วนความแตกต่างของคะแนนที่เกิดจากผู้ตรวจ ปฏิสัมพันธ์ระหว่างผู้สอบกับผู้ตรวจ ปฏิสัมพันธ์ระหว่างข้อสอบกับผู้ตรวจมีค่าน้อย

เมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) และมีจำนวนผู้ตรวจ 4 คน 3 คน และ 2 คน พบว่า องค์ประกอบที่มีความแปรปรวนมากที่สุด ในทุกจำนวนผู้ตรวจ เมื่อเทียบกับความแปรปรวนรวม คือ ปฏิสัมพันธ์ระหว่างข้อสอบกับผู้สอบในแต่ละผู้ตรวจและความคลาดเคลื่อนเชิงสุ่มที่ไม่สามารถระบุได้ (IP:R,e) มีค่าตั้งแต่ 1.0616 ถึง 1.1357 คิดเป็นร้อยละ 46.8346-49.0648 รองลงมาคือ ความแปรปรวนของผู้สอบในแต่ละผู้ตรวจ (P:R) มีค่าตั้งแต่ 0.7768 ถึง 0.8608 คิดเป็นร้อยละ 35.6102-36.6329 และความแปรปรวนของข้อสอบ (I) มีค่าตั้งแต่ 0.2695 ถึง 0.2978 คิดเป็นร้อยละ 12.3545 -12.6734 ส่วนแหล่งความแปรปรวนที่มีค่าน้อยคือ ความแปรปรวนของผู้ตรวจ (R) มีค่าตั้งแต่ 0.0519 ถึง 0.0912 คิดเป็นร้อยละ 2.2087 – 4.0235 และความแปรปรวนของปฏิสัมพันธ์ระหว่างข้อสอบกับผู้ตรวจ (IR) มีค่าตั้งแต่ 0.0029 ถึง 0.0095 คิดเป็นร้อยละ 0.1279-0.4355 แสดงว่า ความแตกต่างของคะแนนสอบ เมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) ในทุกจำนวนผู้ตรวจ เกิดจากปฏิสัมพันธ์ระหว่างข้อสอบกับผู้สอบในแต่ละผู้ตรวจและความคลาดเคลื่อนเชิงสุ่มที่ไม่สามารถระบุได้มากที่สุด รองลงมาคือ ผู้สอบในแต่ละผู้ตรวจและข้อสอบ ตามลำดับ ส่วนความแตกต่างของคะแนนสอบเกิดจากผู้ตรวจ และปฏิสัมพันธ์ระหว่างข้อสอบกับผู้ตรวจค่อนข้างน้อย

เมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) และมีจำนวนผู้ตรวจ 4 คน 3 คน และ 2 คน พบว่า องค์ประกอบที่มีความแปรปรวนมากที่สุด ในทุกจำนวนผู้ตรวจ เมื่อเทียบกับความแปรปรวนรวม คือ ปฏิสัมพันธ์ระหว่างข้อสอบกับผู้สอบในแต่ละผู้ตรวจและความคลาดเคลื่อนเชิงสุ่มที่ไม่สามารถระบุได้ (IP:R,e) มีค่าตั้งแต่ 1.0480 ถึง 1.1356 คิดเป็นร้อยละ 48.4450-52.6448 รองลงมาคือ ความแปรปรวนของผู้สอบ (P) มีค่าตั้งแต่ 0.7656 ถึง 0.8492 คิดเป็นร้อยละ 35.8209-36.8921 และความแปรปรวนของข้อสอบในแต่ละผู้ตรวจ (I:R) มีค่าตั้งแต่ 0.2257 ถึง 0.2664 คิดเป็นร้อยละ 10.4631-12.3520 ส่วนแหล่งความแปรปรวนที่มีค่าน้อยคือ ความแปรปรวนของผู้ตรวจ (R) มีค่าตั้งแต่ 0.0000 ถึง 0.0829 คิดเป็นร้อยละ 0.0000-3.5661 และความแปรปรวนปฏิสัมพันธ์ระหว่างผู้สอบกับผู้ตรวจ (PR) มีค่าตั้งแต่ 0.0000 ถึง 0.0597 คิดเป็นร้อยละ 0.0000- 2.7933 แสดงว่า ความแตกต่างของคะแนนสอบ เมื่อใช้รูปแบบการตรวจให้

คะแนนที่ผู้ตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_j) ในทุกจำนวนผู้ตรวจ เกิดจากปฏิสัมพันธ์ระหว่างข้อสอบกับผู้สอบในแต่ละผู้ตรวจและความคลาดเคลื่อนเชิงสุ่มที่ไม่สามารถระบุได้มากที่สุด รองลงมา คือ ผู้สอบและข้อสอบในแต่ละผู้ตรวจ ตามลำดับ ส่วนความแตกต่างของคะแนนสอบเกิดจากผู้ตรวจและปฏิสัมพันธ์ระหว่างผู้สอบกับผู้ตรวจค่อนข้างน้อย

กล่าวโดยสรุปได้ว่าในทุกเงื่อนไขจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนที่ต่างกัน พบว่า เมื่อใช้รูปแบบการตรวจให้คะแนนเหมือนกัน ความแปรปรวนขององค์ประกอบเดียวกันมีค่าใกล้เคียงกันในทุกจำนวนผู้ตรวจ ความแปรปรวนของผู้สอบและข้อสอบมีค่าความแปรปรวนมากกว่าผู้ตรวจ ทั้งนี้ความแปรปรวนของผู้ตรวจมีค่าค่อนข้างน้อย ไม่เกิน 5% ในทุกรูปแบบการตรวจให้คะแนนและจำนวนผู้ตรวจ แสดงให้เห็นว่าผู้ตรวจมีความแตกต่างกันในการให้คะแนนค่อนข้างน้อย

2. การประมาณค่าและเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิง เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน

ผู้วิจัยประมาณค่าสัมประสิทธิ์การสรุปอ้างอิง โดยใช้โปรแกรม GENOVA (Crick; & Brennan. 1983) เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน จากนั้นนำค่าสัมประสิทธิ์การสรุปอ้างอิงมาทดสอบความแตกต่าง โดยใช้สูตร UX_1 (Woodruff; & Feldt. 1986: 393-413) ดังนี้

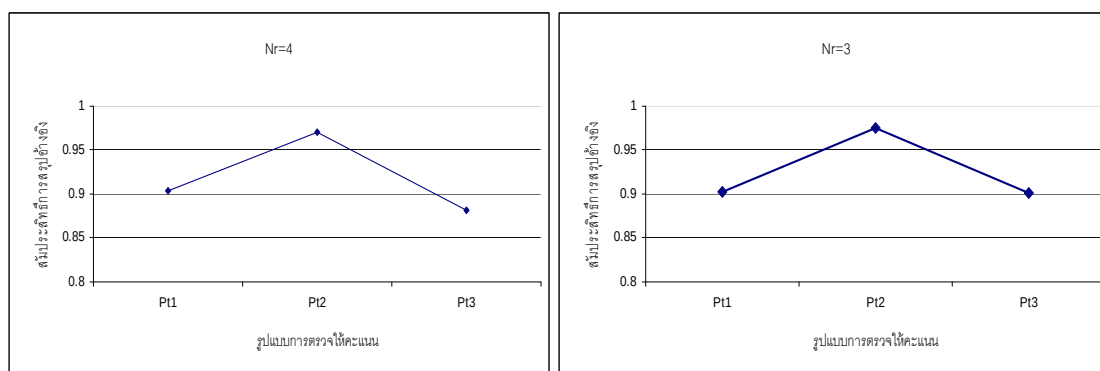
2.1 การประมาณค่าและเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิงเมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน ปรากฏผลดังตาราง 20

ตาราง 20 ค่าสัมประสิทธิ์การสรุปอ้างอิงเมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน

n_r	Pt	ค่าสัมประสิทธิ์การสรุปอ้างอิง	UX_1
4	Pt_1	0.904	914.3420**
	Pt_2	0.970	
	Pt_3	0.882	
3	Pt_1	0.903	1113.0348**
	Pt_2	0.975	
	Pt_3	0.901	
2	Pt_1	0.901	1085.8397**
	Pt_2	0.975	
	Pt_3	0.894	

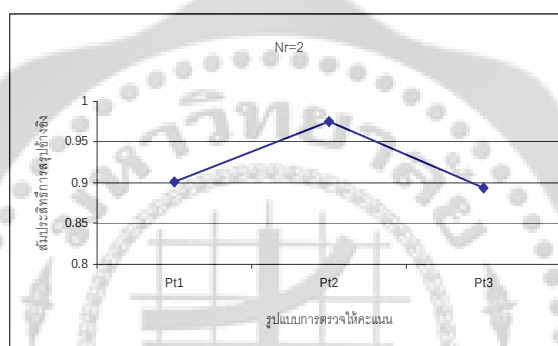
** $p < .01$

และนำผลการวิเคราะห์ตามตาราง 20 มาเขียนแผนภูมิได้ดังภาพประกอบ 13 (ก-ค) ดังนี้



13ก

13ข



13ค

ภาพประกอบ 13(ก-ค) ค่าสัมประสิทธิ์การสรุปอ้างอิง เมื่อจำนวนผู้ตรวจเท่ากัน (13ก) 4 คน (13ข) 3 คน และ (13ค) 2 คน และรูปแบบการตรวจให้คะแนนต่างกัน

จากตาราง 20 และภาพประกอบ 13 (ก-ค) พบว่า ค่าสัมประสิทธิ์การสรุปอ้างอิงเมื่อจำนวนผู้ตรวจเท่ากันและใช้รูปแบบการตรวจให้คะแนนต่างกัน เมื่อใช้ผู้ตรวจ 4 คน รูปแบบการตรวจให้คะแนนที่มีค่าสัมประสิทธิ์การอ้างอิงสูงสุด คือ ผู้ตรวจตรวจสอบข้อสอบทุกข้อของผู้สอบบางคน (P_2) รองลงมาคือ รูปแบบการตรวจที่ผู้ตรวจตรวจสอบข้อสอบทุกข้อของผู้สอบทุกคน (P_1) และผู้ตรวจตรวจสอบข้อสอบบางข้อของผู้สอบทุกคน (P_3) โดยมีค่าสัมประสิทธิ์การสรุปอ้างอิงเป็น 0.970, 0.904 และ 0.882 ตามลำดับ เมื่อใช้ผู้ตรวจ 3 คน รูปแบบการตรวจให้คะแนนที่มีค่าสัมประสิทธิ์การอ้างอิงสูงสุด คือ ผู้ตรวจตรวจสอบข้อสอบทุกข้อของผู้สอบบางคน (P_2) รองลงมาคือ รูปแบบการตรวจที่ผู้ตรวจตรวจสอบข้อสอบทุกข้อของผู้สอบทุกคน (P_1) และผู้ตรวจตรวจสอบข้อสอบบางข้อของผู้สอบทุกคน (P_3) โดยมีค่าสัมประสิทธิ์การสรุปอ้างอิงเป็น 0.975, 0.903 และ 0.901 ตามลำดับ และเมื่อใช้ผู้ตรวจ 2 คน รูปแบบการตรวจให้คะแนนที่มีค่าสัมประสิทธิ์การอ้างอิงสูงสุด คือ ผู้ตรวจตรวจสอบข้อสอบทุกข้อของผู้สอบบางคน (P_2) รองลงมาคือ รูปแบบการตรวจที่ผู้ตรวจตรวจสอบข้อสอบทุกข้อของผู้สอบทุกคน (P_1) และผู้ตรวจตรวจสอบข้อสอบบางข้อของผู้สอบทุกคน (P_3) โดยมีค่าสัมประสิทธิ์การสรุปอ้างอิงเป็น 0.975, 0.903 และ 0.901 ตามลำดับ และเมื่อใช้ผู้ตรวจ 2 คน รูปแบบการตรวจให้คะแนนที่มีค่าสัมประสิทธิ์การอ้างอิงสูงสุด คือ ผู้ตรวจตรวจสอบข้อสอบทุกข้อของผู้สอบบางคน (P_2) รองลงมาคือ รูปแบบการตรวจที่ผู้ตรวจตรวจสอบข้อสอบทุกข้อของผู้สอบทุกคน (P_1) และผู้ตรวจตรวจสอบข้อสอบบางข้อของผู้สอบทุกคน (P_3) โดยมีค่าสัมประสิทธิ์การสรุปอ้างอิงเป็น 0.975, 0.903 และ 0.901 ตามลำดับ

ผู้สอบบางคน (Pt_2) รองลงมาคือ รูปแบบการตรวจที่ผู้ตรวจตรวจสอบข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) และผู้ตรวจตรวจสอบข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) โดยมีค่าสัมประสิทธิ์การสรุปอ้างอิงเป็น 0.975, 0.901 และ 0.894 ตามลำดับ จะเห็นว่า ในทุกจำนวนผู้ตรวจ รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจสอบข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) ให้ค่าสัมประสิทธิ์การสรุปอ้างอิงสูงที่สุด รองลงมาคือ เมื่อผู้ตรวจตรวจสอบข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) และผู้ตรวจตรวจสอบข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) ตามลำดับ

เมื่อนำค่าสัมประสิทธิ์การสรุปอ้างอิงมาทดสอบในแต่ละเงื่อนไขจำนวนผู้ตรวจที่เท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน พบว่า มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 แสดงว่า ค่าสัมประสิทธิ์การสรุปอ้างอิงในแต่ละเงื่อนไขแตกต่างกันอย่างน้อย 1 คู่ และเมื่อพบว่าค่าสัมประสิทธิ์การสรุปอ้างอิงที่นำมาทดสอบ มีความแตกต่างกันอย่างมีนัยสำคัญ จึงนำค่าสัมประสิทธิ์การสรุปอ้างอิงมาทดสอบความแตกต่างรายคู่ โดยใช้สูตร UX_1 (Woodruff; & Feldt. 1986: 393-413) เพื่อดูว่า ค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนคู่ใดแตกต่างกัน เมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน ปรากฏผลดังตาราง 21

ตาราง 21 การทดสอบความแตกต่างรายคู่ของค่าสัมประสิทธิ์การสรุปอ้างอิงเมื่อจำนวนผู้ตรวจ
เท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน

n_r	Pt	ค่าสัมประสิทธิ์การสรุปอ้างอิง	UX_1
4	Pt_1	0.904	455.5802**
	Pt_2	0.970	
	Pt_1	0.904	130.5470**
	Pt_3	0.882	
	Pt_2	0.970	476.5076**
	Pt_3	0.882	
3	Pt_1	0.903	580.7111**
	Pt_2	0.975	
	Pt_1	0.903	5.8079
	Pt_3	0.901	
	Pt_2	0.975	550.1308**
	Pt_3	0.901	
2	Pt_1	0.901	565.3696**
	Pt_2	0.975	
	Pt_1	0.901	28.2490**
	Pt_3	0.894	
	Pt_2	0.975	536.7266**
	Pt_3	0.894	

** $p < .01$

จากตาราง 21 พบว่า ค่าสัมประสิทธิ์การสรุปอ้างอิงที่ประมาณค่ามาจากจำนวนผู้ตรวจที่เท่ากันและมีรูปแบบการตรวจให้คะแนนต่างกัน มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ทุกคู่ ยกเว้น ค่าสัมประสิทธิ์การสรุปอ้างอิง ในกรณีของผู้ตรวจจำนวน 3 คนที่ใช้รูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) กับผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) ที่มีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติที่ระดับ .01 และเมื่อพิจารณาตามสมมติฐานที่ตั้งไว้ในข้อ 1.1 ที่ว่า “ค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนที่มีรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน จะมีค่าสูงกว่า รูปแบบการตรวจให้คะแนนรูปแบบอื่น ในทุก

จำนวนผู้ตรวจให้คะแนน” จะเห็นว่าเป็นไปตามสมมติฐาน เฉพาะค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนในกรณีผู้ตรวจจำนวน 4 คน และ 2 คน ที่มีรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_i) มีค่าสูงกว่าผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_j) อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ส่วนในกรณีผู้ตรวจจำนวน 3 คน ค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนที่มีรูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_i) มีค่าสูงกว่าผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_j) อย่างไม่มีนัยสำคัญทางสถิติที่ระดับ .01 และค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนที่มีรูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_i) นั้นมีค่าต่ำกว่ารูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_j) ในทุกจำนวนผู้ตรวจ อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ซึ่งไม่เป็นไปตามสมมติฐานที่ตั้งไว้

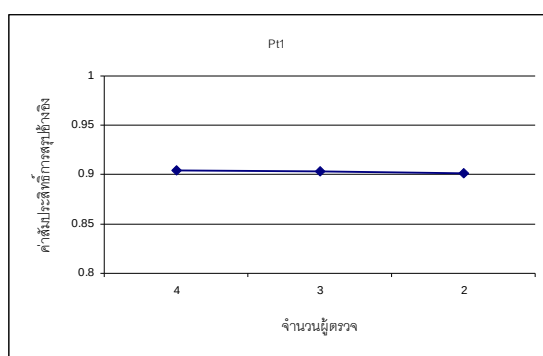
2.2 การประมาณค่าและทดสอบความแตกต่างของค่าสัมประสิทธิ์การสรุปอ้างอิงเมื่อรูปแบบการตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจต่างกัน ปราบกฎผลดังตาราง 22

ตาราง 22 ค่าสัมประสิทธิ์การสรุปอ้างอิงเมื่อรูปแบบการตรวจให้คะแนนเหมือนกัน และจำนวนผู้ตรวจต่างกัน

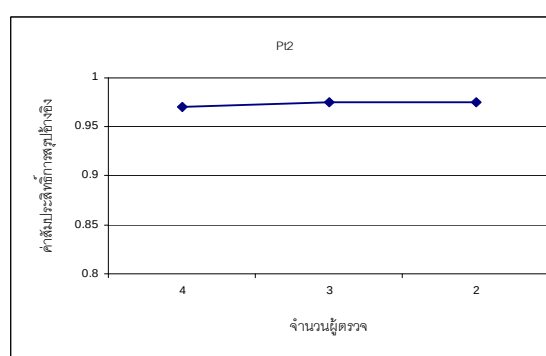
Pt	n_r	ค่าสัมประสิทธิ์การสรุปอ้างอิง	UX_i
Pt_1	4	0.904	11.0365**
	3	0.903	
	2	0.901	
Pt_2	4	0.970	48.3027**
	3	0.975	
	2	0.975	
Pt_3	4	0.882	62.7300**
	3	0.901	
	2	0.894	

** $p < .01$

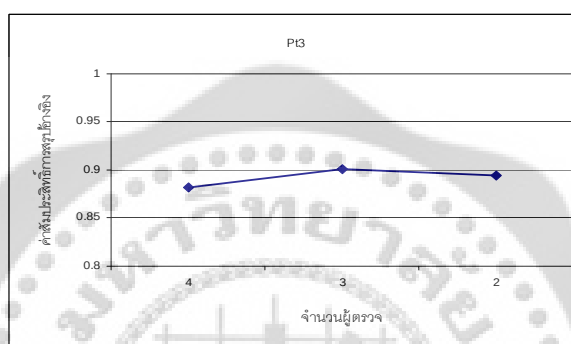
และนำผลการวิเคราะห์ตามตาราง 22 มาเขียนแผนภูมิได้ดังภาพประกอบ 14 (ก-ค) ดังนี้



14ก



14ข



14ค

ภาพประกอบ 14(ก-ค) ค่าสัมประสิทธิ์การสรุปอ้างอิงเมื่อรูปแบบการตรวจให้คะแนนเหมือนกัน

(14ก) ผู้ตรวจตรวจสอบข้อสอบทุกข้อของผู้สอบทุกคน (14ข) ผู้ตรวจตรวจสอบข้อสอบทุกข้อของผู้สอบบางคน และ(14ค) ผู้ตรวจตรวจสอบข้อสอบบางข้อของผู้สอบทุกคน และจำนวนผู้ตรวจต่างกัน

จากตาราง 22 และภาพประกอบ 14 (ก-ค) พบว่า ค่าสัมประสิทธิ์การสรุปอ้างอิงเมื่อรูปแบบการตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจต่างกัน เมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจสอบข้อสอบทุกข้อของผู้สอบทุกคน (P_{t1}) ค่าสัมประสิทธิ์การสรุปอ้างอิงมีค่าสูงสุดเมื่อใช้ผู้ตรวจจำนวน 4 คน รองลงมาคือ 3 คน และ 2 คน โดยมีค่าเป็น 0.904, 0.903 และ 0.901 ตามลำดับ ส่วนเมื่อใช้รูปแบบการตรวจที่ผู้ตรวจตรวจสอบข้อสอบทุกข้อของผู้สอบบางคน (P_{t2}) ค่าสัมประสิทธิ์การสรุปอ้างอิงมีค่าสูงสุดเมื่อใช้ผู้ตรวจจำนวน 3 และ 2 คน คือมีค่าเป็น 0.975 เท่ากัน รองลงมาคือผู้ตรวจ 4 คน มีค่าเป็น 0.970 และเมื่อใช้รูปแบบการตรวจที่ผู้ตรวจตรวจสอบข้อสอบบางข้อของผู้สอบทุกคน (P_{t3}) ค่าสัมประสิทธิ์การสรุปอ้างอิงมีค่าสูงสุดเมื่อใช้ผู้ตรวจจำนวน 3 คน รองลงมาคือ 2 คน และ 4 คน โดยมีค่าเป็น 0.901, 0.894 และ 0.882 ตามลำดับ จะเห็นว่า ในแต่ละรูปแบบการตรวจให้คะแนนเมื่อจำนวนผู้ตรวจต่างกันให้ค่าสัมประสิทธิ์ใกล้เคียงกัน

เมื่อนำค่าสัมประสิทธิ์การสรุปอ้างอิงมาทดสอบในแต่ละเงื่อนไขรูปแบบการตรวจให้คะแนนเหมือนกัน แต่จำนวนผู้ตรวจต่างกัน พบว่า มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 แสดงว่า ค่าสัมประสิทธิ์การสรุปอ้างอิงในแต่ละเงื่อนไขแตกต่างกันอย่างน้อย 1 คู่ และเมื่อพบว่าค่าสัมประสิทธิ์การสรุปอ้างอิงที่นำมาทดสอบมีความแตกต่างกันอย่างมีนัยสำคัญ จึงนำค่าสัมประสิทธิ์การสรุปอ้างอิงมาทดสอบความแตกต่างรายคู่ โดยใช้สูตร UX_1 (Woodruff; & Feldt. 1986: 393-413) เพื่อดูว่า ค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนคู่ใดแตกต่างกัน เมื่อรูปแบบการตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจต่างกัน ปรากฏผลดังตาราง 23

ตาราง 23 การทดสอบความแตกต่างรายคู่ของค่าสัมประสิทธิ์การสรุปอ้างอิง เมื่อรูปแบบการตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจต่างกัน

Pt	n_i	ค่าสัมประสิทธิ์การสรุปอ้างอิง	UX_1
Pt ₁	4	0.904	2.0260
	3	0.903	
	4	0.904	5.9267**
	2	0.901	
	3	0.903	7.7917**
	2	0.901	
Pt ₂	4	0.970	37.4441**
	3	0.975	
	4	0.970	26.5816**
	2	0.975	
	3	0.975	0.0000
	2	0.975	
Pt ₃	4	0.882	49.4353**
	3	0.901	
	4	0.882	26.2513**
	2	0.894	
	3	0.901	10.9023**
	2	0.894	

** $p < .01$

จากตาราง 23 พบว่า ค่าสัมประสิทธิ์การสรุปอ้างอิงที่ประมาณค่ามาจากรูปแบบการตรวจให้คะแนนเหมือนกันแต่จำนวนผู้ตรวจต่างกัน มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ยกเว้น ค่าสัมประสิทธิ์การสรุปอ้างอิงในกรณีที่ผู้ตรวจตรวจสอบข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) เมื่อใช้จำนวนผู้ตรวจ 4 คนและ 3 คน และผู้ตรวจตรวจสอบข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) เมื่อใช้จำนวนผู้ตรวจ 3 คน และ 2 คน ที่มีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติที่ระดับ .01 เมื่อพิจารณาสมมติฐานข้อ 1.2 ที่ว่า “ค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนที่มีจำนวนผู้ตรวจที่มากกว่า จะมีค่าสูงกว่า จำนวนผู้ตรวจที่น้อยกว่า ในทุกรูปแบบการตรวจให้คะแนน” พบว่า เป็นไปตามสมมติฐานที่ตั้งไว้ เมื่อใช้รูปแบบการตรวจที่ผู้ตรวจตรวจสอบข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) ที่ค่าสัมประสิทธิ์การสรุปอ้างอิงของผู้ตรวจจำนวน 4 คน มีค่าสูงกว่าผู้ตรวจจำนวน 2 คน และผู้ตรวจจำนวน 3 คน มีค่าสูงกว่าผู้ตรวจจำนวน 2 คน อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และเมื่อใช้รูปแบบผู้ตรวจตรวจสอบข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) ที่ผู้ตรวจจำนวน 3 คน ให้ค่าสัมประสิทธิ์การสรุปอ้างอิงสูงกว่าผู้ตรวจจำนวน 2 คน อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ส่วนเมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจสอบข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) ค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนเมื่อใช้ผู้ตรวจจำนวน 4 คนมีค่าสูงกว่าผู้ตรวจจำนวน 3 คน อย่างไม่มีนัยสำคัญทางสถิติที่ระดับ .01 และเมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจสอบข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) และรูปแบบการตรวจที่ผู้ตรวจตรวจสอบข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) ค่าสัมประสิทธิ์การสรุปอ้างอิงเมื่อใช้ผู้ตรวจ 4 คน มีค่าต่ำกว่าผู้ตรวจ 3 คน และเมื่อใช้ผู้ตรวจ 4 คน มีค่าต่ำกว่าผู้ตรวจ 2 คน อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ซึ่งไม่เป็นไปตามสมมติฐานที่ตั้งไว้

3. การประมาณค่าและเปรียบเทียบค่าความเที่ยงตรงตามสภาพ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน

ผู้วิจัยประมาณค่าความเที่ยงตรงตามสภาพของคะแนนสอบด้วยการวิเคราะห์หาค่าสัมประสิทธิ์สหสัมพันธ์ของคะแนนสอบกับคะแนนที่ได้จากผลการทดสอบทางการศึกษาระดับชาตินี้พื้นฐาน (Ordinary National Education Test: O-NET) และเปรียบเทียบค่าความเที่ยงตรงตามสภาพของคะแนนแบบทดสอบปลายเปิด เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจคะแนนต่างกัน โดยการนำค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันมาแปลงเป็นคะแนนมาตรฐาน (Z) ของฟิชเชอร์ แล้วนำมาทดสอบความแตกต่าง โดยใช้สถิติไค-สแควร์ (Chi-Square; χ^2) (Meng; Rosenthal; & Robin. 1992: 173) ปรากฏผลดังตาราง 24-25

ตาราง 24 ค่าสัมประสิทธิ์สหสัมพันธ์เมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน

n_r	Pt	r	Z	χ^2
4	Pt ₁	0.794**	1.082	3.1961
	Pt ₂	0.778**	1.040	
	Pt ₃	0.785**	1.058	
3	Pt ₁	0.794**	1.082	0.2511
	Pt ₂	0.793**	1.079	
	Pt ₃	0.796**	1.088	
2	Pt ₁	0.791**	1.074	0.3413
	Pt ₂	0.787**	1.064	
	Pt ₃	0.790**	1.071	

** $p_r < .01$

จากตาราง 24 พบว่า ค่าสัมประสิทธิ์สหสัมพันธ์ของคะแนนสอบกับคะแนน O-NET มีนัยสำคัญทางสถิติที่ระดับ .01 ในทุกเงื่อนไข ค่าสัมประสิทธิ์สหสัมพันธ์ของคะแนนในทุกจำนวนผู้ตรวจและทุกรูปแบบการตรวจให้คะแนนมีค่าใกล้เคียงกัน (r มีค่าตั้งแต่ 0.778 ถึง 0.796) โดยคะแนนจากจำนวนผู้ตรวจ 4 คน มีค่าสหสัมพันธ์สูงสุด เมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt₁) ซึ่งมีค่าเป็น 0.794 ส่วนคะแนนจากจำนวนผู้ตรวจ 3 คน มีค่าสหสัมพันธ์สูงสุด เมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt₃) โดยมีค่าเป็น 0.796 และคะแนนจากจำนวนผู้ตรวจ 2 คน จะให้ค่าสหสัมพันธ์สูงสุดเมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt₁) ซึ่งมีค่าเป็น 0.791

เมื่อนำค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันมาแปลงเป็นคะแนนมาตรฐาน (Z) ของฟิชเชอร์ แล้วนำมาทดสอบความแตกต่าง โดยใช้สถิติไค-สแควร์ (Chi-Square; χ^2) พบว่า ค่าสัมประสิทธิ์สหสัมพันธ์ของคะแนนสอบกับคะแนน O-NET ในแต่ละเงื่อนไขจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนที่ต่างกัน มีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติที่ระดับ .01 ซึ่งไม่เป็นไปตามสมมติฐานที่ตั้งไว้ในข้อ 1.3 ที่ว่า “ค่าความเที่ยงตรงตามสภาพของคะแนนที่มีรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน จะมีค่าสูงกว่า รูปแบบการตรวจให้คะแนนรูปแบบอื่น ในทุกจำนวนผู้ตรวจให้คะแนน” แสดงว่า คะแนนสอบเมื่อใช้จำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน มีความเที่ยงตรงตามสภาพไม่แตกต่างกัน

ตาราง 25 ค่าสัมประสิทธิ์สหสัมพันธ์เมื่อรูปแบบการตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจต่างกัน

Pt	n_r	r	Z	χ^2
Pt ₁	4	0.794**	1.082	2.8053
	3	0.794**	1.082	
	2	0.791**	1.074	
Pt ₂	4	0.778**	1.040	3.2638
	3	0.793**	1.079	
	2	0.787**	1.058	
Pt ₃	4	0.785**	1.074	1.37667
	3	0.796**	1.088	
	2	0.790**	1.071	

** $p_r < .01$

จากตาราง 25 พบว่า ค่าสัมประสิทธิ์สหสัมพันธ์ของคะแนนสอบกับคะแนน O-NET มีนัยสำคัญทางสถิติที่ระดับ .01 ในทุกเงื่อนไข โดยค่าสัมประสิทธิ์สหสัมพันธ์คะแนนในทูลจำนวนผู้ตรวจและทุกรูปแบบการตรวจให้คะแนนมีค่าใกล้เคียงกัน (r มีค่าตั้งแต่ 0.778 ถึง 0.796) โดยคะแนนจากรูปแบบการตรวจที่ผู้ตรวจตรวจสอบข้อสอบทุกข้อของผู้สอบทุกคน (Pt₁) มีค่าสหสัมพันธ์สูงสุดเมื่อใช้ผู้ตรวจ 3 และ 4 คน มีค่าเป็น 0.794 เท่ากัน ส่วนรูปแบบการตรวจที่ผู้ตรวจตรวจสอบข้อสอบทุกข้อของผู้สอบบาง (Pt₂) มีค่าสหสัมพันธ์สูงสุดเมื่อใช้ผู้ตรวจ 3 คน ซึ่งมีค่าเป็น 0.793 และรูปแบบการตรวจที่ผู้ตรวจตรวจสอบข้อสอบบางข้อของผู้สอบทุกคน (Pt₃) มีค่าสหสัมพันธ์สูงสุดเมื่อใช้ผู้ตรวจ 3 คน ซึ่งมีค่าเป็น 0.796

เมื่อนำค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันมาแปลงเป็นคะแนนมาตรฐาน (Z) ของฟิชเชอร์ แล้วนำมาทดสอบความแตกต่าง โดยใช้สถิติไค-สแควร์ (Chi-Square; χ^2) พบว่า ค่าสัมประสิทธิ์สหสัมพันธ์ของคะแนนสอบกับคะแนน O-NET ในแต่ละเงื่อนไขรูปแบบการตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจที่ต่างกัน มีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติที่ระดับ .01 ซึ่งไม่เป็นไปตามสมมติฐานที่ตั้งไว้ในข้อ 1.4 ที่ว่า “ค่าความเที่ยงตรงตามสภาพของคะแนนที่มีจำนวนผู้ตรวจที่มากกว่าจะมีค่าสูงกว่า จำนวนผู้ตรวจที่น้อยกว่า ในทุกรูปแบบการตรวจให้คะแนน” แสดงว่าคะแนนสอบเมื่อใช้รูปแบบการตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจต่างกันมีค่าความเที่ยงตรงตามสภาพไม่แตกต่างกัน

ตอนที่ 3 การศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิด โดยใช้โมเดลหลายองค์ประกอบของราล์ช

ผู้วิจัยนำคะแนนที่ได้มาวิเคราะห์โดยใช้โมเดลหลายองค์ประกอบของราล์ช เพื่อประมาณค่าขนาดความแปรปรวนของแต่ละองค์ประกอบ ประมาณค่าและเปรียบเทียบค่าความเชื่อมั่นแยกส่วนของแต่ละองค์ประกอบ เปรียบเทียบความเที่ยงตรงตามสภาพ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน มีรายละเอียดดังนี้

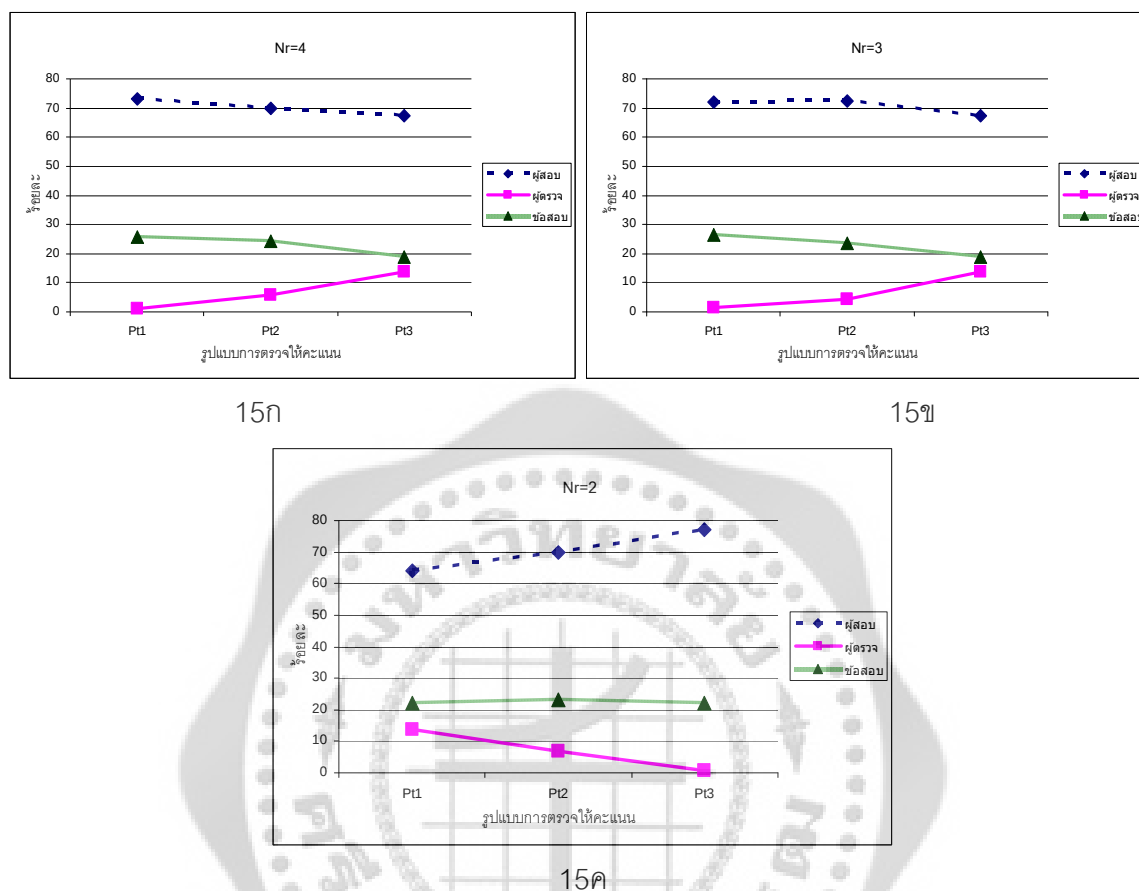
1. การศึกษาความแปรปรวนของแต่ละองค์ประกอบ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน

ผู้วิจัยประมาณค่าความแปรปรวนของแต่ละองค์ประกอบ ได้แก่ ความแปรปรวนของค่าความสามารถของผู้สอบ ค่าความเข้มงวดของผู้ตรวจและค่าความยากของข้อสอบ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน ผลปรากฏดังตาราง 26

ตาราง 26 ขนาดความแปรปรวนของแต่ละองค์ประกอบ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน

จำนวนผู้ตรวจ (คน)	องค์ประกอบ	รูปแบบการตรวจให้คะแนน					
		Pt ₁		Pt ₂		Pt ₃	
		ความแปรปรวน (in logits)	% ความแปรปรวน	ความแปรปรวน (in logits)	% ความแปรปรวน	ความแปรปรวน (in logits)	% ความแปรปรวน
4	ผู้สอบ	0.8281	73.1666	0.8649	69.7331	0.8649	67.4965
	ผู้ตรวจ	0.0121	1.0691	0.0729	5.8776	0.1764	13.7662
	ข้อสอบ	0.2916	25.7643	0.3025	24.3893	0.2401	18.7373
	รวม	1.1318	100.0000	1.2403	100.0000	1.2814	100.0000
3	ผู้สอบ	0.7921	71.9698	0.9025	72.3737	0.8649	67.4965
	ผู้ตรวจ	0.0169	1.5355	0.0529	4.2422	0.1764	13.7662
	ข้อสอบ	0.2916	26.4946	0.2916	23.3841	0.2401	18.7373
	รวม	1.1006	100.0000	1.2470	100.0000	1.2814	100.0000
2	ผู้สอบ	0.8100	63.9154	0.9801	69.9522	0.7396	77.0979
	ผู้ตรวจ	0.1764	13.9194	0.0961	6.8589	0.0081	0.8444
	ข้อสอบ	0.2809	22.1652	0.3249	23.1889	0.2116	22.0578
	รวม	1.2673	100.0000	1.4011	100.0000	0.9593	100.0000

และนำผลการวิเคราะห์ตามตาราง 26 มาเขียนแผนภูมิได้ดังภาพประกอบ 15(ก-ค) ดังนี้



ภาพประกอบ 15 (ก-ค) ร้อยละของความแปรปรวนของแต่ละองค์ประกอบเมื่อจำนวนผู้ตรวจเท่ากัน (15ก) 4 คน (15ข) 3 คน และ (15ค) 2 คน และรูปแบบการตรวจให้คะแนนต่างกัน

จากตาราง 26 และภาพประกอบ 15(ก-ค) พบว่า

1.1 เมื่อจำนวนผู้ตรวจ 4 คนและผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (P_{t1}) พบว่า ความแปรปรวนของผู้สอบเมื่อเทียบกับความแปรปรวนรวมมีค่ามากที่สุด คือ มีค่าเท่ากับ 0.8281 คิดเป็นร้อยละ 73.1666 รองลงมา คือ ความแปรปรวนของข้อสอบและผู้ตรวจ โดยมีค่าเท่ากับ 0.2916 และ 0.0121 คิดเป็นร้อยละ 25.7643 และ 1.0691 ตามลำดับ แสดงว่า ความสามารถของผู้สอบมีความแตกต่างกันมากกว่าความแตกต่างของค่าความยากของข้อสอบและค่าความเข้มงวดของผู้ตรวจ

เมื่อใช้รูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (P_{t2}) พบว่า ความแปรปรวนของผู้สอบเมื่อเทียบกับความแปรปรวนรวมมีค่ามากที่สุด คือ มีค่าเท่ากับ 0.8649 คิดเป็นร้อยละ 69.7331 รองลงมา คือ ความแปรปรวนของข้อสอบและผู้ตรวจ โดยมีค่าเท่ากับ 0.3025 และ

0.0961 คิดเป็นร้อยละ 23.1889 และ 6.8589 ตามลำดับ แสดงว่า ความสามารถของผู้สอบมีความแตกต่างกันมากกว่าความแตกต่างของค่าความยากของข้อสอบและค่าความเข้มงวดของผู้ตรวจ

เมื่อใช้รูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_j) พบว่า ความแปรปรวนของผู้สอบเมื่อเทียบกับความแปรปรวนรวมมีค่ามากที่สุด คือ มีค่าเท่ากับ 0.7396 คิดเป็นร้อยละ 77.0979 รองลงมา คือ ความแปรปรวนของข้อสอบและผู้ตรวจ โดยมีค่าเท่ากับ 0.2116 และ 0.0081 คิดเป็นร้อยละ 22.0578 และ 0.8444 ตามลำดับ แสดงว่า ความสามารถของผู้สอบมีความแตกต่างกันมากกว่าความแตกต่างของค่าความยากของข้อสอบและค่าความเข้มงวดของผู้ตรวจ

เมื่อพิจารณาโดยภาพรวมแล้ว พบว่า ในทุกเงื่อนไขจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนน ความแปรปรวนของผู้สอบมีค่ามากที่สุดเมื่อเทียบกับความแปรปรวนรวม คือ มีค่าร้อยละ 63.9154 - 77.0979 รองลงมาคือ ความแปรปรวนของข้อสอบ มีค่าร้อยละ 18.7373 - 26.4946 และ ความแปรปรวนขององค์ประกอบที่มีค่าน้อยที่สุดคือ ผู้ตรวจ ซึ่งมีค่าร้อยละ 0.8444 - 13.7662 แสดงว่า ความสามารถของผู้สอบมีความแตกต่างกันมากกว่า ความแตกต่างของค่าความยากของข้อสอบและค่าความเข้มงวดของผู้ตรวจ ในทุกเงื่อนไขจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนน

2. การประมาณค่าและเปรียบเทียบค่าความเชื่อมั่นแยกส่วน (Reliability of separation index) ของแต่ละองค์ประกอบ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน

ผู้วิจัยประมาณค่าความเชื่อมั่นแยกส่วนของแต่ละองค์ประกอบ โดยใช้โปรแกรม FACETS (Linacre. 2009) และนำค่าความเชื่อมั่นแยกส่วนของแต่ละองค์ประกอบ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน มาทดสอบความแตกต่างโดยใช้ สูตร UX_1 (Woodruff; & Feldt. 1986: 393-413) ดังนี้

2.1 การประมาณค่าและทดสอบความแตกต่างของค่าความเชื่อมั่นแยกส่วนของผู้สอบ ผู้ตรวจ และข้อสอบ เมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน ปรากฏผลดังนี้

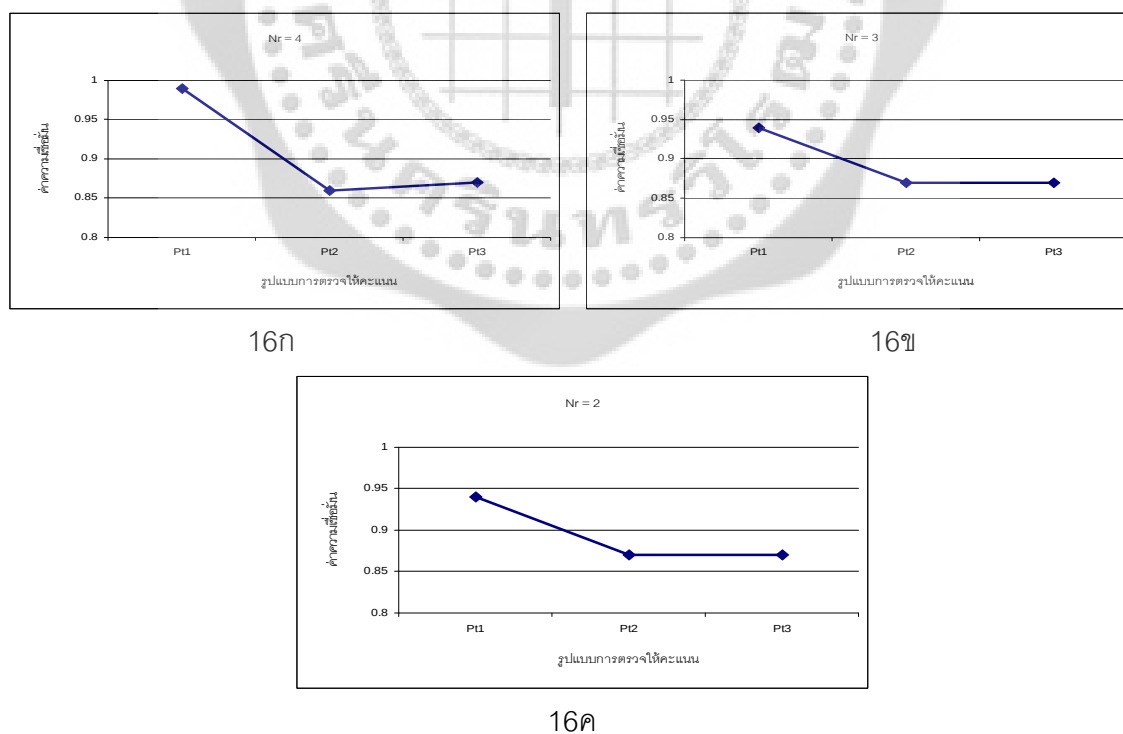
2.1.1 การประมาณค่าความเชื่อมั่นแยกส่วนของผู้สอบเมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน ปรากฏผลดังตาราง 27-28

ตาราง 27 ค่าความเชื่อมั่นแยกส่วนของผู้สอบเมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน

n_r	Pt	ค่าความเชื่อมั่นแยกส่วน	UX_1
4	Pt ₁	0.97	918.1284**
	Pt ₂	0.86	
	Pt ₃	0.87	
3	Pt ₁	0.96	838.9328**
	Pt ₂	0.87	
	Pt ₃	0.87	
2	Pt ₁	0.94	670.6218**
	Pt ₂	0.87	
	Pt ₃	0.87	

** $p < .01$

และนำผลการวิเคราะห์ตามตาราง 27 มาเขียนแผนภูมิได้ดังภาพประกอบ 16(ก-ค) ดังนี้



ภาพประกอบ 16 (ก-ค) ค่าความเชื่อมั่นแยกส่วนของผู้สอบเมื่อจำนวนผู้ตรวจเท่ากัน (16ก) 4 คน (16ข) 3 คน และ (16ค) 2 คน และรูปแบบการตรวจให้คะแนนต่างกัน

จากตาราง 27 และภาพประกอบ 16(ก-ค) พบว่า ค่าความเชื่อมั่นแยกส่วนของผู้สอบ เมื่อจำนวนผู้ตรวจ 4 คน มีค่าความเชื่อมั่นสูงสุดเมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน(Pt_1) คือ มีค่าเป็น 0.97 รองลงมาคือ ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_2) และผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_3) มีค่าเป็น 0.87 และ 0.86 ตามลำดับ เมื่อจำนวนผู้ตรวจ 3 คน มีค่าความเชื่อมั่นสูงสุดเป็น 0.96 เมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน(Pt_1) รองลงมาคือผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) มีค่าเป็น 0.87 เท่ากัน และเมื่อจำนวนผู้ตรวจ 2 คน มีค่าความเชื่อมั่นสูงสุดเมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) คือมีค่าเป็น 0.97 รองลงมาคือ ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) คือมีค่าเป็น 0.87 เท่ากัน จะเห็นว่าในทุกจำนวนผู้ตรวจมีแนวโน้มว่ารูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน(Pt_1) จะให้ค่าความเชื่อมั่นแยกส่วนของผู้สอบสูงที่สุด ส่วนในรูปแบบการตรวจอื่นมีค่าใกล้เคียงกัน

เมื่อนำค่าความเชื่อมั่นแยกส่วนของผู้สอบมาทดสอบความแตกต่างในแต่ละเงื่อนไขที่จำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน พบว่า มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 แสดงว่า ค่าความเชื่อมั่นแยกส่วนของผู้สอบในแต่ละเงื่อนไขแตกต่างกันอย่างน้อย 1 คู่ และเมื่อพบว่าค่าความเชื่อมั่นแยกส่วนของผู้สอบที่นำมาทดสอบ มีความแตกต่างอย่างมีนัยสำคัญ จึงนำค่าความเชื่อมั่นแยกส่วนมาทดสอบความแตกต่างรายคู่ โดยใช้สูตร UX_1 (Woodruff; & Feldt.1986: 393-413) เพื่อดูว่าค่าความเชื่อมั่นแยกส่วนของผู้สอบคู่ใดแตกต่างกัน เมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน ปรัชญาผลดังตาราง 28

ตาราง 28 การทดสอบความแตกต่างรายคู่ของค่าความเชื่อมั่นแยกส่วนของผู้สอบเมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน

n_r	Pt	ค่าความเชื่อมั่นแยกส่วน	UX_1
4	Pt ₁	0.97	493.0830**
	Pt ₂	0.86	
	Pt ₁	0.97	554.3837**
	Pt ₃	0.87	
	Pt ₂	0.86	2.7752
	Pt ₃	0.87	
3	Pt ₁	0.96	451.9521**
	Pt ₂	0.87	
	Pt ₁	0.96	531.9976**
	Pt ₃	0.87	
	Pt ₂	0.87	0.0000
	Pt ₃	0.87	
2	Pt ₁	0.94	359.9598**
	Pt ₂	0.87	
	Pt ₁	0.94	524.2886**
	Pt ₃	0.87	
	Pt ₂	0.87	0.0000
	Pt ₃	0.87	

** $p < .01$

จากตาราง 28 พบว่า ค่าความเชื่อมั่นแยกส่วนของผู้สอบที่ประมาณค่ามาจากจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ยกเว้น ค่าความเชื่อมั่นแยกส่วนของผู้สอบที่ใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt₂) และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt₃) มีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติที่ระดับ .01 ในทุกจำนวนผู้ตรวจ

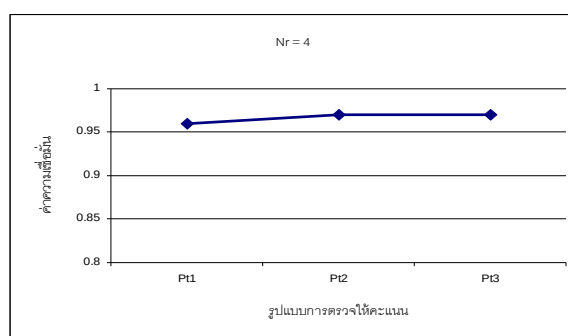
2.1.2 การประมาณค่าและทดสอบความแตกต่างของค่าความเชื่อมั่นแยกส่วนของผู้ตรวจเมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน ปรากฏผลดังตาราง 29-30

ตาราง 29 ค่าความเชื่อมั่นแยกส่วนของผู้ตรวจเมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน

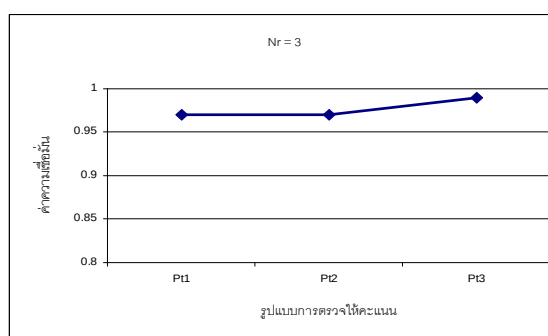
n_r	Pt	ค่าความเชื่อมั่นแยกส่วน	UX_1
4	Pt ₁	0.96	16.1384**
	Pt ₂	0.97	
	Pt ₃	0.97	
3	Pt ₁	0.97	425.0071**
	Pt ₂	0.97	
	Pt ₃	0.99	
2	Pt ₁	0.98	1361.7692**
	Pt ₂	0.99	
	Pt ₃	0.89	

** $p < .01$

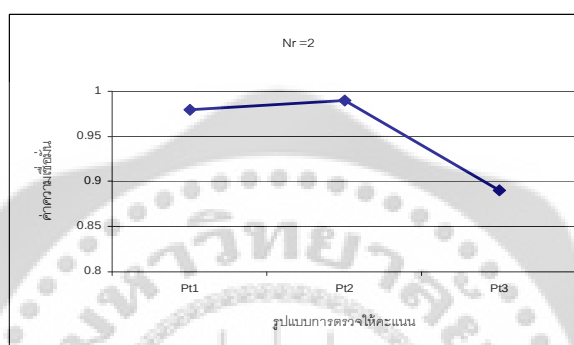
และนำผลการวิเคราะห์ตามตาราง 29 มาเขียนแผนภูมิได้ดังภาพประกอบ 17(ก-ค) ดังนี้



17ก



17ข



17ค

ภาพประกอบ 17 (ก-ค) ค่าความเชื่อมั่นแยกส่วนของผู้ตรวจเมื่อจำนวนผู้ตรวจเท่ากัน (17ก) 4 คน (17ข) 3 คน และ (17ค) 2 คน และรูปแบบการตรวจให้คะแนนต่างกัน

จากตาราง 29 และภาพประกอบ 17(ก-ค) พบว่า ค่าความเชื่อมั่นแยกส่วนของผู้ตรวจเมื่อจำนวนผู้ตรวจ 4 คน มีค่าความเชื่อมั่นสูงสุดเมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) คือมีค่าเป็น 0.97 เท่ากัน ส่วนผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) มีค่าความเชื่อมั่นเป็น 0.96 เมื่อจำนวนผู้ตรวจ 3 คนมีค่าความเชื่อมั่นสูงสุดเป็น 0.99 เมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) รองลงมาคือ ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) และผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) มีค่าเป็น 0.97 เท่ากัน และเมื่อจำนวนผู้ตรวจ 2 คนมีค่าความเชื่อมั่นสูงสุดเมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) คือ 0.99 รองลงมาคือ ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) คือมีค่าเป็น 0.98 และ 0.89 ตามลำดับ จะเห็นว่าในแต่ละจำนวนผู้ตรวจ ค่าความเชื่อมั่นแยกส่วนของผู้ตรวจมีค่าใกล้เคียงกันในทุกรูปแบบการตรวจให้คะแนน ยกเว้นกรณีที่ผู้ตรวจ 2 คนที่ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) ที่ให้ค่าความเชื่อมั่นต่ำกว่ากรณีอื่น

เมื่อนำค่าความเชื่อมั่นแยกส่วนของผู้ตรวจมาทดสอบความแตกต่างในแต่ละเงื่อนไขของจำนวนผู้ตรวจที่เท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน พบว่า มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ.01 แสดงว่าค่าความเชื่อมั่นแยกส่วนของผู้ตรวจในแต่ละเงื่อนไขแตกต่างกันอย่างน้อย 1 คู่ และเมื่อพบว่าค่าความเชื่อมั่นแยกส่วนของผู้ตรวจที่นำมาทดสอบมีความแตกต่างอย่างมีนัยสำคัญ จึงนำค่าความเชื่อมั่นแยกส่วนของผู้ตรวจมาทดสอบความแตกต่างรายคู่ โดยใช้สูตร UX_1 (Woodruff; & Feldt, 1986: 393-413) เพื่อดูว่าค่าความเชื่อมั่นแยกส่วนของผู้ตรวจคู่ใดแตกต่างกัน เมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน ปรากฏผลดังตาราง 30

ตาราง 30 การทดสอบความแตกต่างรายคู่ของค่าความเชื่อมั่นแยกส่วนของผู้ตรวจ เมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน

n_r	Pt	ค่าความเชื่อมั่นแยกส่วน	UX_1
4	Pt_1	0.96	8.9483**
	Pt_2	0.97	
	Pt_1	0.96	8.8262**
	Pt_3	0.97	
	Pt_2	0.97	0.0000
	Pt_3	0.97	
3	Pt_1	0.97	0.0000
	Pt_2	0.97	
	Pt_1	0.97	185.5760**
	Pt_3	0.99	
	Pt_2	0.97	680.8846**
	Pt_3	0.99	
2	Pt_1	0.98	680.8846**
	Pt_2	0.99	
	Pt_1	0.98	680.8846**
	Pt_3	0.89	
	Pt_2	0.99	680.8846**
	Pt_3	0.89	

** $p < .01$

จากตาราง 30 พบว่า ค่าความเชื่อมั่นแยกส่วนของผู้ตรวจในทุกจำนวนผู้ตรวจที่ใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ยกเว้น กรณีค่าความเชื่อมั่นแยกส่วนของผู้ตรวจ เมื่อจำนวนผู้ตรวจ 4 คน ที่ใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) และกรณีจำนวนผู้ตรวจ 3 คน ที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) และผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) มีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติที่ระดับ .01

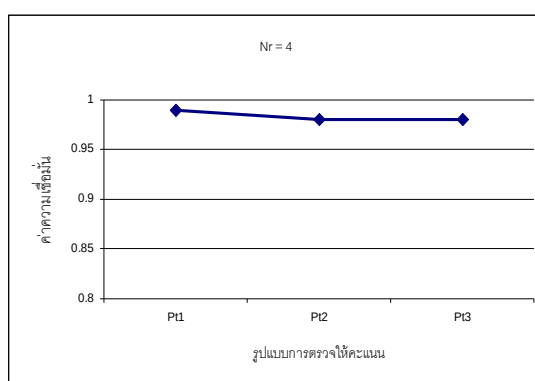
2.1.3 การประมาณค่าและทดสอบความแตกต่างของค่าความเชื่อมั่นแยกส่วนของข้อสอบเมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน ปรากฏผลดังตาราง 31- 32

ตาราง 31 ค่าความเชื่อมั่นแยกส่วนของข้อสอบเมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน

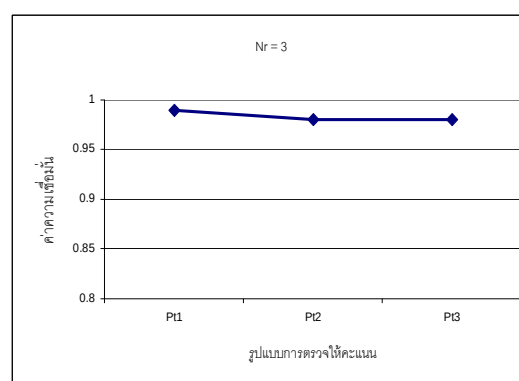
n_r	Pt	ค่าความเชื่อมั่นแยกส่วน	UX_1
4	Pt_1	0.99	123.3171**
	Pt_2	0.98	
	Pt_3	0.98	
3	Pt_1	0.99	219.8016**
	Pt_2	0.98	
	Pt_3	0.98	
2	Pt_1	0.99	113.8032**
	Pt_2	0.98	
	Pt_3	0.97	

**p < .01

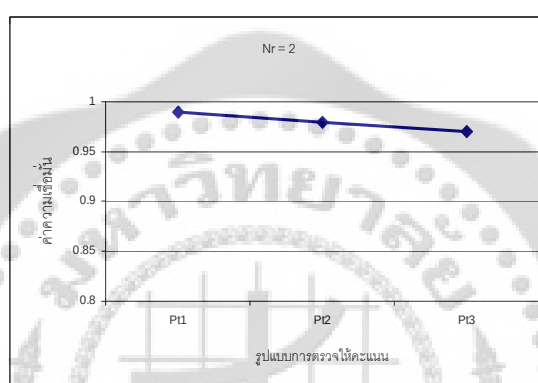
และนำผลการวิเคราะห์ตามตาราง 31 มาเขียนแผนภูมิได้ดังภาพประกอบ 18(ก-ค) ดังนี้



18ก



18ข



18ค

ภาพประกอบ 18 (ก-ค) ค่าความเชื่อมั่นแยกส่วนของข้อสอบเมื่อจำนวนผู้ตรวจเท่ากัน (18ก) 4 คน (18ข) 3 คน และ (18ค) 2 คน และรูปแบบการตรวจให้คะแนนต่างกัน

จากตาราง 31 และภาพประกอบ 18(ก-ค) พบว่า ค่าความเชื่อมั่นแยกส่วนของข้อสอบเมื่อจำนวนผู้ตรวจ 4 คนและจำนวนผู้ตรวจ 3 คน ให้ค่าความเชื่อมั่นของข้อสอบสูงสุดเมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) คือมีค่าเป็น 0.99 รองลงมาคือ ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน(Pt_2) และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) ซึ่งมีความเชื่อมั่นเท่ากัน คือมีค่าเป็น 0.98 และเมื่อจำนวนผู้ตรวจ 2 คน รูปแบบการตรวจให้คะแนนที่มีค่าความเชื่อมั่นสูงสุดคือ ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) มีค่าเป็น 0.99 รองลงมาคือ ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) มีค่าเป็น 0.98 และ 0.97 ตามลำดับ จะเห็นว่า ในทุกจำนวนผู้ตรวจมีแนวโน้มว่าค่าความเชื่อมั่นแยกส่วนของข้อสอบในรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) มีค่าสูงสุด ส่วนรูปแบบการตรวจให้คะแนนรูปแบบอื่นมีค่าใกล้เคียงกัน

เมื่อนำค่าความเชื่อมั่นแยกส่วนของข้อสอบมาทดสอบความแตกต่างในแต่ละเงื่อนไขของจำนวนผู้ตรวจที่เท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน พบว่า มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 แสดงว่า ค่าความเชื่อมั่นแยกส่วนของข้อสอบในแต่ละเงื่อนไขแตกต่างกันอย่างน้อย 1 คู่ และเมื่อพบว่าค่าความเชื่อมั่นแยกส่วนของข้อสอบที่นำมาทดสอบมีความแตกต่างกันอย่างมีนัยสำคัญ จึงนำค่าความเชื่อมั่นแยกส่วนมาทดสอบความแตกต่างรายคู่ โดยใช้สูตร UX_1 (Woodruff; & Feldt, 1986: 393-413) เพื่อดูว่า ค่าความเชื่อมั่นแยกส่วนของข้อสอบคู่ใดแตกต่างกัน เมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน ปรากฏผลดังตาราง 32

ตาราง 32 การทดสอบความแตกต่างรายคู่ของค่าความเชื่อมั่นแยกส่วนของข้อสอบ เมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน

n_r	Pt	ค่าความเชื่อมั่นแยกส่วน	UX_1
4	Pt ₁	0.99	556.5262**
	Pt ₂	0.98	
	Pt ₁	0.99	58.6800**
	Pt ₃	0.98	
	Pt ₂	0.98	0.0000
	Pt ₃	0.98	
3	Pt ₁	0.99	110.2180**
	Pt ₂	0.98	
	Pt ₁	0.99	63.5598**
	Pt ₃	0.98	
	Pt ₂	0.98	0.0000
	Pt ₃	0.98	
2	Pt ₁	0.99	633.6407**
	Pt ₂	0.98	
	Pt ₁	0.99	210.2694**
	Pt ₃	0.97	
	Pt ₂	0.98	37.4244**
	Pt ₃	0.97	

** $p < .01$

จากตาราง 32 พบว่า ค่าความเชื่อมั่นแยกส่วนของข้อสอบ เมื่อจำนวนผู้ตรวจ 4 คน และจำนวนผู้ตรวจ 3 คน ที่ใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) และผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ส่วนกรณีที่ใช้รูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) มีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติที่ระดับ .01 ส่วนเมื่อจำนวนผู้ตรวจ 2 คน ค่าความเชื่อมั่นแยกส่วนของข้อสอบ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ในทุกเงื่อนไขรูปแบบการตรวจให้คะแนน

จากผลการทดสอบความแตกต่างของค่าความเชื่อมั่นแยกส่วนของผู้สอบ ผู้ตรวจ และข้อสอบ ที่มีจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน (2.1.1-2.1.3) และเมื่อพิจารณาตามสมมติฐานที่ตั้งไว้ในข้อ 2.1 ที่ว่า “ค่าความเชื่อมั่นแยกส่วนของผู้สอบและข้อสอบ ที่มีรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน จะมีค่าสูงกว่า รูปแบบการตรวจให้คะแนนรูปแบบอื่น ส่วนค่าความเชื่อมั่นแยกส่วนของผู้ตรวจ จะมีค่าต่ำกว่า รูปแบบการตรวจให้คะแนนรูปแบบอื่น ในทุกจำนวนผู้ตรวจให้คะแนน” พบว่า เป็นไปตามสมมติฐานที่ตั้งไว้ เฉพาะค่าความเชื่อมั่นแยกส่วนของผู้สอบและข้อสอบในทุกเงื่อนไขจำนวนผู้ตรวจ ส่วนค่าความเชื่อมั่นแยกส่วนของผู้ตรวจนั้นเป็นไปตามสมมติฐานที่ตั้งไว้ ยกเว้น ค่าความเชื่อมั่นแยกส่วนของผู้ตรวจ ในกรณีผู้ตรวจจำนวน 2 คน เมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) มีค่าสูงกว่ารูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และค่าความเชื่อมั่นแยกส่วนของผู้ตรวจ ในกรณีผู้ตรวจจำนวน 3 คน เมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) มีค่าเท่ากับรูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) ซึ่งไม่เป็นไปตามสมมติฐานที่ตั้งไว้

2.2 การประมาณค่าและทดสอบความแตกต่างของค่าความเชื่อมั่นแยกส่วนของผู้สอบผู้ตรวจ และข้อสอบ เมื่อรูปแบบการตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจต่างกัน ปรากฏผลดังนี้

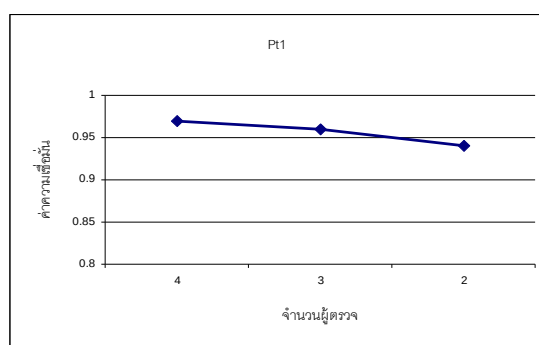
2.2.1 การประมาณค่าและทดสอบความแตกต่างของค่าความเชื่อมั่นแยกส่วนของผู้สอบ เมื่อรูปแบบการตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจต่างกัน ปรากฏผลดังตาราง 33-34

ตาราง 33 ค่าความเชื่อมั่นแยกส่วนของผู้สอบเมื่อรูปแบบการตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจต่างกัน

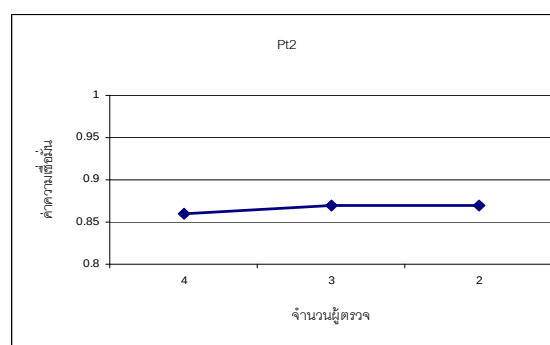
Pt	n_r	ค่าความเชื่อมั่นแยกส่วน	UX_1
Pt ₁	4	0.97	941.9428**
	3	0.96	
	2	0.94	
Pt ₂	4	0.86	5.3427
	3	0.87	
	2	0.87	
Pt ₃	4	0.87	0.0000
	3	0.87	
	2	0.87	

** $p < .01$

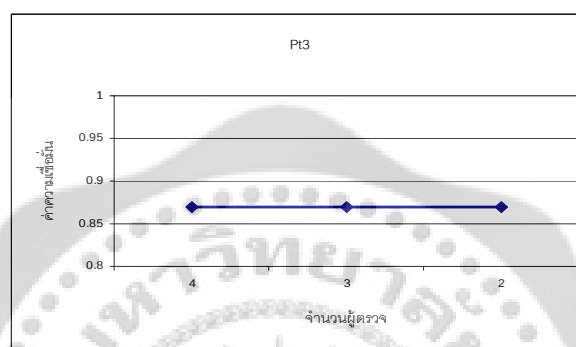
และนำผลการวิเคราะห์ตามตาราง 33 มาเขียนแผนภูมิได้ดังภาพประกอบ 19 (ก-ค) ดังนี้



19ก



19ข



19ค

ภาพประกอบ 19 (ก-ค) ค่าความเชื่อมั่นแยกส่วนของผู้สอบเมื่อรูปแบบการตรวจให้คะแนนเหมือนกัน (19ก) ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (19ข) เมื่อผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน และ (19ค) ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน และจำนวนผู้ตรวจต่างกัน

จากตาราง 33 และภาพประกอบ 19(ก-ค) พบว่า ค่าความเชื่อมั่นแยกส่วนของผู้สอบที่ใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) ให้ค่าความเชื่อมั่นสูงสุดเป็น 0.97 เมื่อใช้ผู้ตรวจจำนวน 4 คน และให้ค่าความเชื่อมั่นเป็น 0.96 และ 0.94 เมื่อใช้ผู้ตรวจจำนวน 3 คน และ 2 คน ตามลำดับ และเมื่อผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) ให้ค่าความเชื่อมั่นสูงสุดเป็น 0.87 เมื่อใช้ผู้ตรวจจำนวน 2 คนและ 3 คน ส่วนเมื่อใช้ผู้ตรวจจำนวน 4 คน ให้ค่าความเชื่อมั่นเป็น 0.86 และเมื่อผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) ในทุกจำนวนผู้ตรวจให้ค่าความเชื่อมั่นเป็น 0.87 เท่ากัน จะเห็นว่า รูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) มีแนวโน้มว่าจำนวนผู้ตรวจที่มากกว่ามีแนวโน้มให้ค่าความเชื่อมั่นแยกส่วนของผู้สอบสูงกว่าจำนวนผู้ตรวจที่น้อยกว่า ส่วนรูปแบบการตรวจให้คะแนนอื่นให้ค่าความเชื่อมั่นแยกส่วนของผู้สอบใกล้เคียงกันในทุกจำนวนผู้ตรวจ

เมื่อนำค่าความเชื่อมั่นแยกส่วนของผู้สอบมาทดสอบความแตกต่างในแต่ละเงื่อนไขของรูปแบบการตรวจให้คะแนนที่เหมือนกันและจำนวนผู้ตรวจต่างกัน พบว่า รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_i) มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ส่วนกรณีเมื่อใช้รูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) มีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติที่ระดับ .01 แสดงว่า ค่าความเชื่อมั่นแยกส่วนของผู้สอบในรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_i) เมื่อจำนวนผู้ตรวจต่างกัน มีค่าแตกต่างกันอย่างน้อย 1 คู่ และเมื่อพบว่าค่าความเชื่อมั่นแยกส่วนของผู้สอบที่นำมาทดสอบมีความแตกต่างอย่างมีนัยสำคัญ จึงนำค่าความเชื่อมั่นแยกส่วนมาทดสอบความแตกต่างรายคู่ โดยใช้สูตร UX_1 (Woodruff; & Feldt. 1986: 393-413) เพื่อดูว่าค่าความเชื่อมั่นแยกส่วนของผู้สอบคู่ใด เมื่อผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_i) และมีจำนวนผู้ตรวจต่างกันมีค่าแตกต่างกัน ปราบกฎผลดังตาราง 34

ตาราง 34 การทดสอบความแตกต่างรายคู่ของค่าความเชื่อมั่นแยกส่วนของผู้สอบเมื่อผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน และจำนวนผู้ตรวจต่างกัน

Pt	n_i	ค่าความเชื่อมั่นแยกส่วน	UX_1
Pt_i	4	0.97	364.4135**
	3	0.96	
	4	0.97	524.6491**
	2	0.94	
	3	0.96	411.4247**
	2	0.94	

** $p < .01$

จากตาราง 34 พบว่า ค่าความเชื่อมั่นแยกส่วนของผู้สอบที่ประมาณค่ามาจากรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_i) เมื่อมีจำนวนผู้ตรวจต่างกันมีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ทุกคู่

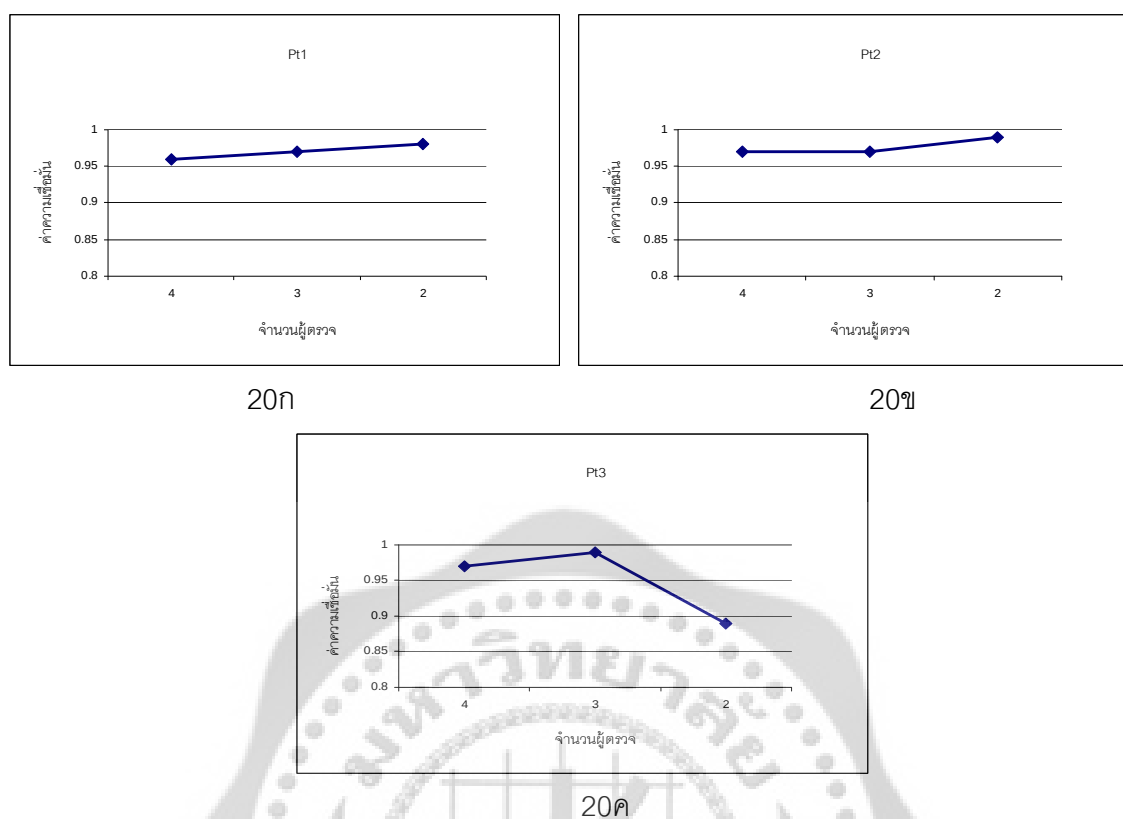
2.2.2 การประมาณค่าและทดสอบความแตกต่างของค่าความเชื่อมั่นแยกส่วนของผู้ตรวจ เมื่อรูปแบบการตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจต่างกัน ปรากฏผลดังตาราง 35-36

ตาราง 35 ค่าความเชื่อมั่นแยกส่วนของผู้ตรวจ เมื่อรูปแบบการตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจต่างกัน

Pt	n_r	ค่าความเชื่อมั่นแยกส่วน	UX_1
Pt ₁	4	0.96	1361.7692**
	3	0.97	
	2	0.98	
Pt ₂	4	0.97	1014.9089**
	3	0.97	
	2	0.99	
Pt ₃	4	0.97	1361.7692**
	3	0.99	
	2	0.89	

** p < .01

และนำผลการวิเคราะห์ตามตาราง 35 มาเขียนแผนภูมิได้ดังภาพประกอบ 20(ก-ค) ดังนี้



ภาพประกอบ 20 (ก-ค) ค่าความเชื่อมั่นแยกส่วนของผู้ตรวจ เมื่อรูปแบบการตรวจให้คะแนนเหมือนกัน (20ก) ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (20ข) ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน และ (20ค) ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน และจำนวนผู้ตรวจต่างกัน

จากตาราง 35 และภาพประกอบ 20(ก-ค) พบว่า ค่าความเชื่อมั่นแยกส่วนของผู้ตรวจ ที่ใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) ให้ค่าความเชื่อมั่นสูงสุดเป็น 0.98 เมื่อผู้ตรวจจำนวน 2 คน และให้ค่าความเชื่อมั่นเป็น 0.97 และ 0.96 เมื่อจำนวนผู้ตรวจ 3 คน และ 4 คน ตามลำดับ เมื่อผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) ให้ค่าความเชื่อมั่นสูงสุดเป็น 0.99 เมื่อจำนวนผู้ตรวจ 2 คน ส่วนจำนวนผู้ตรวจ 3 คนและ 4 คน ให้ค่าความเชื่อมั่นเท่ากันคือ 0.97 และเมื่อผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) ค่าความเชื่อมั่นสูงสุดเป็น 0.99 เมื่อจำนวนผู้ตรวจ 3 คน รองลงมาคือ จำนวนผู้ตรวจ 4 คน และ 2 คน ซึ่งมีค่าความเชื่อมั่นเป็น 0.97 และ 0.89 ตามลำดับ

เมื่อนำค่าความเชื่อมั่นแยกส่วนของผู้ตรวจมาทดสอบความแตกต่างในแต่ละเงื่อนไขรูปแบบการตรวจให้คะแนนที่เหมือนกันและจำนวนผู้ตรวจต่างกัน พบว่า มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 แสดงว่า ค่าความเชื่อมั่นแยกส่วนของผู้ตรวจ ในแต่ละเงื่อนไขแตกต่างกันอย่างน้อย

1 คู่และเมื่อพบว่าค่าความเชื่อมั่นแยกส่วนของผู้ตรวจที่นำมาทดสอบมีความแตกต่างกันอย่างมีนัยสำคัญ จึงนำค่าความเชื่อมั่นแยกส่วนมาทดสอบความแตกต่างรายคู่ โดยใช้สูตร UX_1 (Woodruff; & Feldt. 1986: 393-413) เพื่อดูว่าค่าความเชื่อมั่นแยกส่วนของผู้ตรวจคู่ใดแตกต่างกัน เมื่อรูปแบบการตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจต่างกัน ปรากฏผลดังตาราง 36

ตาราง 36 การทดสอบความแตกต่างรายคู่ของค่าความเชื่อมั่นแยกส่วนของผู้ตรวจเมื่อรูปแบบการตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจต่างกัน

Pt	n_r	ค่าความเชื่อมั่นแยกส่วน	UX_1
Pt ₁	4	0.96	680.8846**
	3	0.97	
	4	0.96	680.8846**
	2	0.98	
	3	0.97	680.8846**
	2	0.98	
Pt ₂	4	0.97	0.0000
	3	0.97	
	4	0.97	680.8846**
	2	0.99	
	3	0.97	680.8846**
	2	0.99	
Pt ₃	4	0.97	680.8846**
	3	0.99	
	4	0.97	680.8846**
	2	0.89	
	3	0.99	680.8846**
	2	0.89	

** $p < .01$

จากตาราง 36 พบว่า ค่าความเชื่อมั่นแยกส่วนของผู้ตรวจที่ใช้รูปแบบการตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจต่างกัน มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ในทุก

เงื่อนไขจำนวนผู้ตรวจ ยกเว้น ค่าความเชื่อมั่นแยกส่วนของผู้ตรวจที่ใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจสอบข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) เมื่อมีจำนวนผู้ตรวจ 4 คน และ 3 คน มีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติที่ระดับ .01

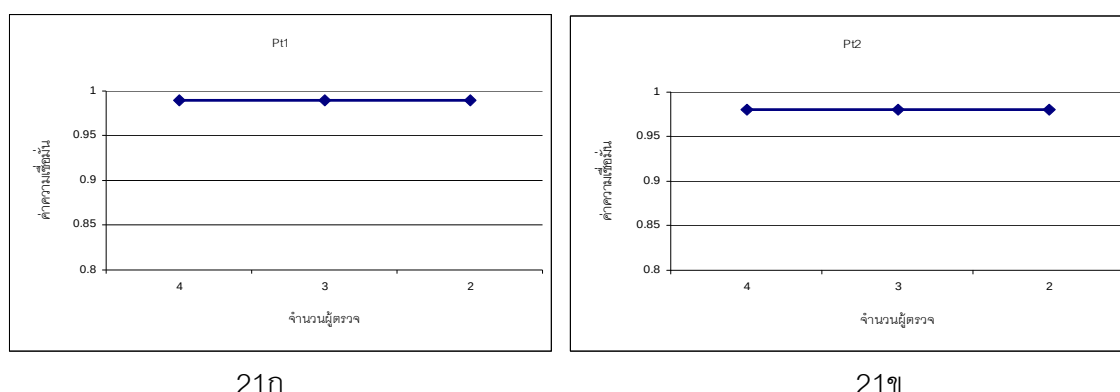
2.2.3 การประมาณค่าและทดสอบความแตกต่างของค่าความเชื่อมั่นแยกส่วนของข้อสอบเมื่อรูปแบบการตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจต่างกัน ปรากฏผลดังตาราง 37-38

ตาราง 37 ค่าความเชื่อมั่นแยกส่วนของข้อสอบเมื่อรูปแบบการตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจต่างกัน

Pt	n_i	ค่าความเชื่อมั่นแยกส่วน	UX_1
Pt_1	4	0.99	0.0000
	3	0.99	
	2	0.99	
Pt_2	4	0.98	0.0000
	3	0.98	
	2	0.98	
Pt_3	4	0.98	20.0687**
	3	0.98	
	2	0.97	

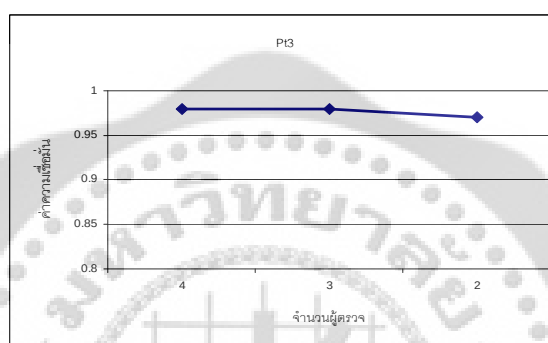
** $p < .01$

และนำผลการวิเคราะห์ตามตาราง 37 มาเขียนแผนภูมิได้ดังภาพประกอบ 21(ก-ค) ดังนี้



21ก

21ข



21ค

ภาพประกอบ 21(ก-ค) ค่าความเชื่อมั่นแยกส่วนของข้อสอบเมื่อรูปแบบการตรวจให้คะแนนเหมือนกัน (21ก) ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (21ข) ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน และ (21ค) ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน และจำนวนผู้ตรวจต่างกัน

จากตาราง 37 และภาพประกอบ 21(ก-ค) พบว่า ค่าความเชื่อมั่นแยกส่วนของข้อสอบเมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt_1) ในทุกจำนวนผู้ตรวจ ให้ค่าความเชื่อมั่นเท่ากัน คือมีค่าเป็น 0.99 และเมื่อผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt_2) มีค่าความเชื่อมั่นเป็น 0.98 ในทุกจำนวนผู้ตรวจเช่นกัน ส่วนเมื่อผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) มีค่าความเชื่อมั่นเท่ากับ 0.98 เมื่อมีผู้ตรวจจำนวน 4 คนและ 3 คน และมีค่าความเชื่อมั่นเป็น 0.97 เมื่อมีผู้ตรวจจำนวน 2 คน จะเห็นว่าในทุกรูปแบบการตรวจให้คะแนนมีแนวโน้มว่า จำนวนผู้ตรวจที่ต่างกันให้ค่าความเชื่อมั่นแยกส่วนของข้อสอบใกล้เคียงกัน

เมื่อนำค่าความเชื่อมั่นแยกส่วนของข้อสอบมาทดสอบความแตกต่างในแต่ละเงื่อนไขรูปแบบการตรวจให้คะแนนที่เหมือนกันและจำนวนผู้ตรวจต่างกัน พบว่า รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 แสดงว่า ค่าความเชื่อมั่นแยกส่วนของข้อสอบในแต่ละเงื่อนไขจำนวนผู้ตรวจมีค่าแตกต่างกันอย่างน้อย 1 คู่

ส่วนเมื่อใช้รูปแบบการตรวจให้คะแนนแบบอื่นมีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติที่ระดับ .01 และเมื่อพบว่า ค่าความเชื่อมั่นแยกส่วนของข้อสอบที่นำมาทดสอบมีความแตกต่างกันอย่างมีนัยสำคัญ จึงนำค่าความเชื่อมั่นแยกส่วนมาทดสอบความแตกต่างรายคู่ โดยใช้สูตร UX_1 (Woodruff; & Feldt. 1986: 393-413) เพื่อดูว่าค่าความเชื่อมั่นแยกส่วนของข้อสอบคู่ใดแตกต่างกัน เมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) แต่มีจำนวนผู้ตรวจต่างกัน ปรากฏผลดังตาราง 38

ตาราง 38 การทดสอบความแตกต่างรายคู่ของค่าความเชื่อมั่นแยกส่วนของข้อสอบ เมื่อผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน และจำนวนผู้ตรวจต่างกัน

Pt	n_r	ค่าความเชื่อมั่นแยกส่วน	UX_1
Pt_3	4	0.98	0.0000
	3	0.98	
	4	0.98	23.2550 **
	2	0.97	
	3	0.98	14.6921 **
	2	0.97	

** $p < .01$

จากตาราง 38 พบว่า ค่าความเชื่อมั่นแยกส่วนของข้อสอบที่ประมาณค่ามาจากรูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt_3) เมื่อจำนวนผู้ตรวจ 4 คน และ 2 คน จำนวนผู้ตรวจ 3 คน และ 2 คน มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ส่วนจำนวนผู้ตรวจ 4 คน และ 3 คน มีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติที่ระดับ .01

จากผลการทดสอบความแตกต่างของค่าความเชื่อมั่นแยกส่วนของข้อสอบ ผู้ตรวจ และข้อสอบ เมื่อรูปแบบการตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจต่างกัน (2.2.1-2.2.3) จะเห็นว่า เป็นไปตามสมมติฐานที่ตั้งไว้ในข้อ 2.2 ที่ว่า “ค่าความเชื่อมั่นแยกส่วนของข้อสอบและผู้สอบ ที่มีจำนวนผู้ตรวจที่มากกว่า จะมีค่าสูงกว่า จำนวนผู้ตรวจที่น้อยกว่า ส่วนค่าความเชื่อมั่นแยกส่วนของผู้ตรวจ จะมีค่าต่ำกว่า จำนวนผู้ตรวจที่น้อยกว่า ในทุกรูปแบบการตรวจให้คะแนน” เฉพาะความเชื่อมั่นแยกส่วนของข้อสอบเมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุก

คน (P_{t_1}) ที่จำนวนผู้ตรวจที่มากกว่าจะให้ค่าสูงกว่าจำนวนผู้ตรวจที่น้อยกว่า อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และค่าความเชื่อมั่นแยกส่วนของข้อสอบเมื่อใช้รูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (P_{t_2}) ที่จำนวนผู้ตรวจ 4 คน มีค่าสูงกว่าผู้ตรวจ 2 คน จำนวนผู้ตรวจ 3 คนมีค่าสูงกว่าผู้ตรวจ 2 คน อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และความเชื่อมั่นแยกส่วนของผู้ตรวจเมื่อใช้รูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (P_{t_1}) ที่จำนวนผู้ตรวจที่มากกว่าจะมีค่าต่ำกว่าจำนวนผู้ตรวจที่น้อยกว่าอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 เมื่อผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (P_{t_2}) จำนวนผู้ตรวจ 4 คน มีค่าต่ำกว่าผู้ตรวจ 2 คน และจำนวนผู้ตรวจ 3 คนมีค่าต่ำกว่าผู้ตรวจ 2 คน และเมื่อผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (P_{t_3}) จำนวนผู้ตรวจ 4 คนมีค่าต่ำกว่าผู้ตรวจ 3 คน อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ส่วนในกรณีอื่น ไม่เป็นไปตามสมมติฐานที่ตั้งไว้

3. การประมาณค่าและเปรียบเทียบความเที่ยงตรงตามสภาพของคะแนนความสามารถเมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน

ผู้วิจัยประมาณค่าความเที่ยงตรงตามสภาพของคะแนนความสามารถ ด้วยการวิเคราะห์หาค่าสัมประสิทธิ์สหสัมพันธ์ของคะแนนความสามารถกับคะแนน O-NET และเปรียบเทียบค่าความเที่ยงตรงตามสภาพของคะแนนแบบทดสอบปลายเปิด เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน โดยการนำค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันมาแปลงเป็นคะแนนมาตรฐาน (Z) ของพิชเชอร์ แล้วนำมาทดสอบความแตกต่าง โดยใช้สถิติไค-สแควร์ (Chi-Square; χ^2) (Meng; Rosenthal; & Robin. 1992: 173) ปรากฏผลดังตาราง 39-40

ตาราง 39 ค่าสัมประสิทธิ์สหสัมพันธ์เมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน

n_r	Pt	r	Z	χ^2
4	Pt ₁	0.789**	1.069	4.1687
	Pt ₂	0.769**	1.018	
	Pt ₃	0.772**	1.025	
3	Pt ₁	0.791**	1.074	2.1515
	Pt ₂	0.778**	1.040	
	Pt ₃	0.788**	1.066	
2	Pt ₁	0.786**	1.061	2.6822
	Pt ₂	0.774**	1.030	
	Pt ₃	0.785**	1.058	

**p < .01

จากตาราง 39 พบว่า ค่าสัมประสิทธิ์สหสัมพันธ์ของคะแนนความสามารถกับคะแนนที่ได้จากผลการทดสอบทางการศึกษาระดับชาตินิยมพื้นฐาน (O-NET) มีนัยสำคัญทางสถิติที่ระดับ .01 ในทุกเงื่อนไข ค่าสัมประสิทธิ์สหสัมพันธ์คะแนนในทุกรายการผู้ตรวจและทุกรูปแบบการตรวจให้คะแนนมีค่าใกล้เคียงกัน (r มีค่าตั้งแต่ 0.769 ถึง 0.791) โดยคะแนนจากจำนวนผู้ตรวจ 4 คน 3 คน และ 2 คน ที่ใช้รูปแบบการตรวจที่ผู้ตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt₁) มีค่าสหสัมพันธ์สูงสุด มีค่าเป็น 0.789, 0.791 และ 0.786 ตามลำดับ

เมื่อนำค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันมาแปลงเป็นคะแนนมาตรฐาน (Z) ของฟิชเชอร์ แล้วนำมาทดสอบความแตกต่าง โดยใช้สถิติไค-สแควร์ (Chi-Square; χ^2) พบว่า ค่าสัมประสิทธิ์สหสัมพันธ์ของคะแนนที่ได้จากผลการทดสอบทางการศึกษาระดับชาตินิยมพื้นฐาน (O-NET) ในแต่ละเงื่อนไขรูปแบบการตรวจให้คะแนนที่ต่างกันและจำนวนผู้ตรวจเท่ากันมีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติที่ระดับ .01 ซึ่งไม่เป็นไปตามสมมติฐานที่ตั้งไว้ในข้อ 2.3 ที่ว่า “ค่าความเที่ยงตรงตามสภาพของคะแนนที่มีรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน จะมีค่าสูงกว่า รูปแบบการตรวจให้คะแนนรูปแบบอื่น ในทุกรายการผู้ตรวจให้คะแนน” แสดงว่าคะแนนความสามารถที่ได้จากจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน มีความเที่ยงตรงตามสภาพไม่แตกต่างกัน

ตาราง 40 ค่าสัมประสิทธิ์สหสัมพันธ์เมื่อรูปแบบการตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจต่างกัน

Pt	n_r	r	Z	χ^2
Pt ₁	4	0.789**	1.069	1.9227
	3	0.791**	1.074	
	2	0.786**	1.061	
Pt ₂	4	0.769**	1.018	0.7361
	3	0.778**	1.040	
	2	0.774**	1.030	
Pt ₃	4	0.772**	1.025	3.3181
	3	0.788**	1.066	
	2	0.785**	1.058	

** p < .01

จากตาราง 40 พบว่า ค่าสัมประสิทธิ์สหสัมพันธ์ของคะแนนความสามารถกับคะแนนที่ได้จากผลทดสอบทางการศึกษาระดับชาตินิยมพื้นฐาน (O-NET) มีนัยสำคัญทางสถิติที่ระดับ .01 ในทุกเงื่อนไข ค่าสัมประสิทธิ์สหสัมพันธ์ของคะแนนในทุกจำนวนผู้ตรวจและทุกรูปแบบการตรวจให้คะแนนมีค่าใกล้เคียงกัน (r มีค่าตั้งแต่ 0.769 ถึง 0.791) โดยในทุกรูปแบบการตรวจให้คะแนน คือ ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Pt₁) ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (Pt₂) และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (Pt₃) มีค่าสหสัมพันธ์สูงสุด เมื่อใช้ผู้ตรวจ 3 คน โดยมีค่าเป็น 0.791, 0.778 และ 0.788 ตามลำดับ

เมื่อนำค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันมาแปลงเป็นคะแนนมาตรฐาน (Z) ของพิชเชอร์ แล้วนำมาทดสอบความแตกต่าง โดยใช้สถิติไค-สแควร์ (Chi-Square; χ^2) พบว่า ค่าสัมประสิทธิ์สหสัมพันธ์ของคะแนนความสามารถกับคะแนน O-NET ในแต่ละเงื่อนไขรูปแบบการตรวจคะแนนที่เหมือนกันและจำนวนผู้ตรวจที่ต่างกัน มีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติที่ระดับ .01 ซึ่งไม่เป็นไปตามสมมติฐานที่ตั้งไว้ในข้อ 2.4 ที่ว่า “ค่าความเที่ยงตรงตามสภาพของคะแนนที่มีจำนวนผู้ตรวจที่มากกว่า จะมีค่าสูงกว่า จำนวนผู้ตรวจที่น้อยกว่า ในทุกรูปแบบการตรวจให้คะแนน” แสดงว่า คะแนนความสามารถที่ได้จากรูปแบบการตรวจให้คะแนนที่เหมือนกันและจำนวนผู้ตรวจต่างกัน มีความเที่ยงตรงตามสภาพไม่แตกต่างกัน

บทที่ 5

สรุป อภิปรายผล และข้อเสนอแนะ

การศึกษาวิจัยในครั้งนี้เป็นการศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน โดยใช้โมเดลการสรุปอ้างอิงตามทฤษฎีการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราล์ชตามทฤษฎีการตอบข้อสอบ ผู้วิจัยศึกษาความแปรปรวนขององค์ประกอบหลัก ได้แก่ ผู้สอบ ข้อสอบและผู้ตรวจ ประเมินค่าและเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิง ค่าความเชื่อมั่นแยกส่วน และค่าความเที่ยงตรงตามสภาพ เมื่อจำนวนผู้ตรวจต่างกันคือ 4 คน 3 คน และ 2 คน และรูปแบบการตรวจให้คะแนนต่างกัน คือ ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน ซึ่งเป็นการเปรียบเทียบภายในแต่ละโมเดล มีความมุ่งหมายของการวิจัย สมมติฐานของการวิจัยและวิธีดำเนินการวิจัย ดังนี้

ความมุ่งหมายของการวิจัย

1. เพื่อศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ ที่มีจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน โดยใช้โมเดลการสรุปอ้างอิง แยกได้เป็น
 - 1.1 เพื่อศึกษาขนาดขององค์ประกอบความแปรปรวนของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจ และรูปแบบการตรวจให้คะแนนต่างกัน
 - 1.2 เพื่อประมาณค่าและเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจ และรูปแบบการตรวจให้คะแนนต่างกัน
 - 1.3 เพื่อประมาณค่าและเปรียบเทียบความเที่ยงตรงตามสภาพของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน
2. เพื่อศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ ที่มีจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน โดยใช้โมเดลหลายองค์ประกอบของราล์ช แยกได้เป็น
 - 2.1 เพื่อศึกษาขนาดขององค์ประกอบความแปรปรวนของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจ และรูปแบบการตรวจให้คะแนนต่างกัน
 - 2.2 เพื่อประมาณค่าและเปรียบเทียบค่าความเชื่อมั่นแยกส่วนของแต่ละองค์ประกอบเมื่อจำนวนผู้ตรวจ และรูปแบบการตรวจให้คะแนนต่างกัน
 - 2.3 เพื่อประมาณค่าและเปรียบเทียบความเที่ยงตรงตามสภาพของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน

สมมติฐานของการวิจัย

1. เมื่อวิเคราะห์คุณลักษณะของคะแนนแบบทดสอบปลายเปิด ที่มีจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน ด้วยโมเดลการสรุปอ้างอิง คาดว่า ค่าสัมประสิทธิ์การสรุปอ้างอิง และความเที่ยงตรงตามสภาพ จะมีค่าดังนี้

1.1 ค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนที่มีรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคนจะมีค่าสูงกว่ารูปแบบการตรวจให้คะแนนรูปแบบอื่น ในทุกจำนวนผู้ตรวจให้คะแนน

1.2 ค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนที่มีจำนวนผู้ตรวจที่มากกว่าจะมีค่าสูงกว่าจำนวนผู้ตรวจที่น้อยกว่า ในทุกรูปแบบการตรวจให้คะแนน

1.3 ค่าความเที่ยงตรงตามสภาพของคะแนนที่มีรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคนจะมีค่าสูงกว่ารูปแบบการตรวจให้คะแนนรูปแบบอื่น ในทุกจำนวนผู้ตรวจให้คะแนน

1.4 ค่าความเที่ยงตรงตามสภาพของคะแนนที่มีจำนวนผู้ตรวจที่มากกว่าจะมีค่าสูงกว่าจำนวนผู้ตรวจที่น้อยกว่า ในทุกรูปแบบการตรวจให้คะแนน

2. เมื่อวิเคราะห์คุณลักษณะของคะแนนแบบทดสอบปลายเปิดที่มีจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน ด้วยโมเดลหลายองค์ประกอบของราล์ซ คาดว่า ค่าความเชื่อมั่นแยกส่วนของแต่ละองค์ประกอบ และความเที่ยงตรงตามสภาพ จะมีค่าดังนี้

2.1 ค่าความเชื่อมั่นแยกส่วนของผู้สอบและข้อสอบที่มีรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคนจะมีค่าสูงกว่ารูปแบบการตรวจให้คะแนนรูปแบบอื่น ส่วนค่าความเชื่อมั่นแยกส่วนของผู้ตรวจจะมีค่าต่ำกว่ารูปแบบการตรวจให้คะแนนรูปแบบอื่น ในทุกจำนวนผู้ตรวจให้คะแนน

2.2 ค่าความเชื่อมั่นแยกส่วนของผู้สอบและข้อสอบ ที่มีจำนวนผู้ตรวจที่มากกว่าจะมีค่าสูงกว่าจำนวนผู้ตรวจที่น้อยกว่า ส่วนค่าความเชื่อมั่นแยกส่วนของผู้ตรวจจะมีค่าต่ำกว่าจำนวนผู้ตรวจที่น้อยกว่า ในทุกรูปแบบการตรวจให้คะแนน

2.3 ค่าความเที่ยงตรงตามสภาพของคะแนนที่มีรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคนจะมีค่าสูงกว่ารูปแบบการตรวจให้คะแนนรูปแบบอื่น ในทุกจำนวนผู้ตรวจให้คะแนน

2.4 ค่าความเที่ยงตรงตามสภาพของคะแนนที่มีจำนวนผู้ตรวจที่มากกว่าจะมีค่าสูงกว่าจำนวนผู้ตรวจที่น้อยกว่าในทุกรูปแบบการตรวจให้คะแนน

วิธีดำเนินการวิจัย

กลุ่มตัวอย่างที่ใช้ในการวิจัยครั้งนี้ คือ นักเรียนระดับชั้นมัธยมศึกษาปีที่ 4 สำนักงานเขตพื้นที่การศึกษาจังหวัดน่าน ปีการศึกษา 2552 จำนวน 180 คน โดยวิธีการสุ่มแบบสองขั้นตอน และผู้ตรวจให้คะแนนจำนวน 4 คน ซึ่งเต็มใจให้ความร่วมมือในการตรวจ โดยเป็นครูผู้สอนที่มีวุฒิการศึกษาทางด้านคณิตศาสตร์ตั้งแต่ระดับปริญญาตรีขึ้นไป และมีประสบการณ์ในการสอนวิชาคณิตศาสตร์ในระดับชั้นมัธยมศึกษาปีที่ 3 อย่างน้อย 5 ปี

เครื่องมือที่ใช้ในการวิจัยครั้งนี้ คือ แบบทดสอบปลายเปิดวิชาคณิตศาสตร์ที่ผู้วิจัยสร้างขึ้นจำนวน 12 ข้อ ข้อละ 4 คะแนน เป็นการวัดความสามารถในวิชาคณิตศาสตร์ โดยเนื้อหาที่ใช้เป็นเนื้อหาตามสาระการเรียนรู้คณิตศาสตร์ ระดับชั้นมัธยมศึกษาตอนต้น (ชั้นมัธยมศึกษาปีที่ 1-3) ตามหลักสูตรการศึกษาขั้นพื้นฐาน พุทธศักราช 2551 ผ่านการพิจารณาจากผู้เชี่ยวชาญและทดลองใช้สองครั้งเพื่อปรับปรุงข้อสอบ หากคุณลักษณะทางสถิติรายข้อและคัดเลือกข้อสอบ หากคุณภาพแบบทดสอบทั้งฉบับ ซึ่งข้อสอบมีค่าความยากอยู่ระหว่าง 0.2499–0.7729 ค่าอำนาจจำแนกอยู่ระหว่าง 0.2015–0.6752 และค่าความเชื่อมั่นของแบบทดสอบทั้งฉบับเป็น 0.8771 และตรวจสอบความเป็นมิติเดียว (Unidimensional) ของแบบทดสอบโดยใช้การวิเคราะห์องค์ประกอบเชิงยืนยัน

ในการเก็บรวบรวมข้อมูล ผู้วิจัยดำเนินการสอบด้วยตนเอง โดยนำแบบทดสอบที่มีความยาว 12 ข้อ ไปทดสอบกับนักเรียนที่เป็นกลุ่มตัวอย่าง จาก 6 โรงเรียน ระหว่างวันที่ 24 สิงหาคม 2552 ถึงวันที่ 31 สิงหาคม 2552 ผู้วิจัยนำกระดาษคำตอบของผู้สอบทั้งหมดมาถ่ายเอกสารจำนวน 4 ชุด จัดให้ผู้ตรวจจำนวน 4 คน ตรวจให้คะแนนตามกฎเกณฑ์การให้คะแนนที่ผู้วิจัยได้สร้างขึ้นในคู่มือการตรวจให้คะแนน ซึ่งผู้วิจัยได้อบรมชี้แจงวิธีการตรวจและกฎเกณฑ์การให้คะแนนในคู่มือการตรวจให้คะแนนให้ผู้ตรวจเข้าใจตรงกันแล้ว จึงให้อิสระผู้ตรวจแต่ละคนตรวจข้อสอบทีละข้อจนครบทุกข้อและครบตามจำนวนผู้สอบทุกคน

ผู้วิจัยนำผลการตรวจให้คะแนนข้อสอบทั้ง 12 ข้อ จากผู้ตรวจทั้งหมดมาจัดกระทำข้อมูลตามเงื่อนไขและระดับของเงื่อนไขที่ต้องการศึกษา คือ เมื่อจำนวนผู้ตรวจต่างกันคือ 2 คน 3 คน และ 4 คน และรูปแบบการตรวจให้คะแนนต่างกัน คือ ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน วิเคราะห์ค่าสถิติพื้นฐานของคะแนนสอบ และวิเคราะห์คะแนนที่ได้โดยใช้ 2 โมเดลคือ 1) โมเดลการสรุปอ้างอิงเพื่อหาขนาดองค์ประกอบความแปรปรวน ประมาณค่าและเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิง ประมาณค่าและเปรียบเทียบค่าความเที่ยงตรงตามสภาพ และ 2) โมเดลหลายองค์ประกอบของราล์ซเพื่อหาขนาดองค์ประกอบความแปรปรวน ประมาณค่าและเปรียบเทียบค่าความเชื่อมั่นแยกส่วนของแต่ละองค์ประกอบ ประมาณค่าและเปรียบเทียบค่าความเที่ยงตรงตามสภาพ ตามลำดับ

สรุปผลการวิจัย

การศึกษาวิจัยในครั้งนี้ สรุปผลการวิจัยได้ดังต่อไปนี้

1. ผลการศึกษาคูณลักษณะคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ที่วิเคราะห์ด้วยโมเดลการสรุปอ้างอิง เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน สรุปได้ดังนี้

1.1 ขนาดขององค์ประกอบความแปรปรวนของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน องค์ประกอบที่มีความแปรปรวนมากที่สุดในทุกจำนวนผู้ตรวจเมื่อเทียบกับความแปรปรวนรวม คือ ปฏิสัมพันธ์ระหว่างผู้สอบกับข้อสอบ รองลงมาคือ ความแปรปรวนของผู้สอบและข้อสอบ ส่วนความแปรปรวนของผู้ตรวจ ปฏิสัมพันธ์ระหว่างผู้สอบกับผู้ตรวจ ปฏิสัมพันธ์ระหว่างข้อสอบกับผู้ตรวจ มีค่าค่อนข้างน้อย

เมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน องค์ประกอบที่มีความแปรปรวนมากที่สุดในทุกจำนวนผู้ตรวจ เมื่อเทียบกับความแปรปรวนรวม คือ ปฏิสัมพันธ์ระหว่างข้อสอบกับผู้สอบในแต่ละผู้ตรวจและความคลาดเคลื่อนเชิงสุ่มที่ไม่อาจจะระบุได้ รองลงมาคือ ความแปรปรวนของผู้สอบในแต่ละผู้ตรวจและข้อสอบ ส่วนความแปรปรวนของผู้ตรวจ ปฏิสัมพันธ์ระหว่างข้อสอบกับผู้ตรวจ มีค่าค่อนข้างน้อย

เมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน องค์ประกอบที่มีความแปรปรวนมากที่สุดในทุกจำนวนผู้ตรวจ เมื่อเทียบกับความแปรปรวนรวม คือ ปฏิสัมพันธ์ระหว่างข้อสอบกับผู้สอบในแต่ละผู้ตรวจและความคลาดเคลื่อนเชิงสุ่มที่ไม่อาจจะระบุได้ รองลงมาคือ ความแปรปรวนของผู้สอบและข้อสอบในแต่ละผู้ตรวจ ส่วนความแปรปรวนของผู้ตรวจ ปฏิสัมพันธ์ระหว่างผู้สอบกับผู้ตรวจ มีค่าค่อนข้างน้อย

1.2 การประมาณค่าและเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ สรุปผลตามสมมติฐานของการวิจัย ดังนี้

1.2.1 เมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน พบว่า ในทุกจำนวนผู้ตรวจ รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคนให้ค่าสัมประสิทธิ์การสรุปอ้างอิงสูงที่สุด รองลงมาคือ ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคนและผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน ตามลำดับ และการเปรียบเทียบตามสมมติฐานข้อ 1.1 ที่ว่า “ค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนที่มีรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน จะมีค่าสูงกว่า รูปแบบการตรวจให้คะแนนรูปแบบอื่น ในทุกจำนวนผู้ตรวจให้คะแนน” สรุปได้ว่า ค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนในกรณีผู้ตรวจจำนวน 4 คน และ 2 คน เมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคนมีค่าสูงกว่าผู้ตรวจ

ตรวจข้อสอบบางข้อของผู้สอบทุกคนอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ส่วนในกรณีผู้ตรวจจำนวน 3 คนนั้นสูงกว่าอย่างไม่มีนัยสำคัญทางสถิติที่ระดับ .01 และค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคนมีค่าต่ำกว่ารูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคนในทุกจำนวนผู้ตรวจ อย่างมีนัยสำคัญทางสถิติที่ระดับ .01

1.2.2 เมื่อรูปแบบการตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจต่างกัน พบว่าเมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน จำนวนผู้ตรวจที่มากกว่ามีแนวโน้มว่าค่าสัมประสิทธิ์การสรุปอ้างอิงมีค่าสูงกว่าจำนวนผู้ตรวจที่น้อยกว่า ส่วนในรูปแบบการตรวจอื่นค่าสัมประสิทธิ์การสรุปอ้างอิงมีค่าสูงสุดเมื่อใช้ผู้ตรวจจำนวน 3 คน รองลงมาคือ 2 คนและ 4 คนตามลำดับ และการเปรียบเทียบตามสมมติฐานข้อ 1.2 ที่ว่า “ค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนที่มีจำนวนผู้ตรวจที่มากกว่า จะมีค่าสูงกว่า จำนวนผู้ตรวจที่น้อยกว่า ในทุกรูปแบบการตรวจให้คะแนน” สรุปได้ว่า ค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนในกรณีที่ใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน เมื่อใช้ผู้ตรวจ 4 คนมีค่าสูงกว่าผู้ตรวจ 2 คน และผู้ตรวจ 3 คนมีค่าสูงกว่าผู้ตรวจ 2 คน อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และในกรณีที่ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคนเมื่อใช้ผู้ตรวจ 3 คน มีค่าสูงกว่าผู้ตรวจ 2 คน อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ส่วนค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนในกรณีผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน เมื่อใช้ผู้ตรวจ 4 คนมีค่าสูงกว่าผู้ตรวจ 3 คน อย่างไม่มีนัยสำคัญทางสถิติที่ระดับ .01 และค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนในกรณีผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคนและผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน เมื่อใช้ผู้ตรวจ 4 คน มีค่าต่ำกว่าผู้ตรวจ 3 คน และผู้ตรวจ 4 คน มีค่าต่ำกว่าผู้ตรวจ 2 คน อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน เมื่อใช้ผู้ตรวจ 3 คนและ 2 คนมีค่าเท่ากัน

1.3 การประมาณค่าและเปรียบเทียบความเที่ยงตรงตามสภาพของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน พิจารณาจากค่าสัมประสิทธิ์สหสัมพันธ์ของคะแนนสอบกับคะแนนที่ได้จากผลการทดสอบทางการศึกษาระดับชาตินั้นพื้นฐาน (O-NET) พบว่า ค่าสัมประสิทธิ์สหสัมพันธ์ในทุกเงื่อนไขมีค่าใกล้เคียงกัน (r มีค่าตั้งแต่ 0.778 ถึง 0.796) และมีนัยสำคัญทางสถิติที่ระดับ .01 ในทุกเงื่อนไข เมื่อนำมาทดสอบความแตกต่างพบว่า ค่าสัมประสิทธิ์สหสัมพันธ์ของคะแนนสอบกับคะแนนที่ได้จากผลการทดสอบทางการศึกษาระดับชาตินั้นพื้นฐาน ในแต่ละเงื่อนไขจำนวนผู้ตรวจและรูปแบบการตรวจที่ต่างกัน มีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติที่ระดับ .01 แสดงว่า คะแนนที่ได้จากทุกเงื่อนไขจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนน มีความเที่ยงตรงตามสภาพไม่แตกต่างกัน ซึ่งไม่เป็นไปตามสมมติฐานข้อ

1.3 ที่ว่า “ค่าความเที่ยงตรงตามสภาพของคะแนนที่มีรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน จะมีค่าสูงกว่า รูปแบบการตรวจให้คะแนนรูปแบบอื่น ในทุกจำนวนผู้ตรวจให้คะแนน” และข้อ 1.4 ที่ว่า “ค่าความเที่ยงตรงตามสภาพของคะแนนที่มีจำนวนผู้ตรวจที่มากกว่า จะมีค่าสูงกว่า จำนวนผู้ตรวจที่น้อยกว่า ในทุกรูปแบบการตรวจให้คะแนน”

2. ผลการศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ โดยใช้โมเดลหลายองค์ประกอบของราล์ซ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน สรุปได้ดังนี้

2.1 ขนาดขององค์ประกอบความแปรปรวนของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ พบว่า ในทุกจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนน ความแปรปรวนของผู้สอบมีค่ามากที่สุด เมื่อเทียบกับความแปรปรวนรวม รองลงมาคือ ความแปรปรวนของข้อสอบและความแปรปรวนของผู้ตรวจ ตามลำดับ

2.2 การประมาณค่าและเปรียบเทียบค่าความเชื่อมั่นแยกส่วนของแต่ละองค์ประกอบเมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน พบว่า

2.2.1 เมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน และเปรียบเทียบตามสมมติฐานที่ตั้งไว้ในข้อ 2.1 ที่ว่า “ค่าความเชื่อมั่นแยกส่วนของผู้สอบและข้อสอบ ที่มีรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน จะมีค่าสูงกว่า รูปแบบการตรวจให้คะแนนรูปแบบอื่น ส่วนค่าความเชื่อมั่นแยกส่วนของผู้ตรวจ จะมีค่าต่ำกว่า รูปแบบการตรวจให้คะแนนรูปแบบอื่น ในทุกจำนวนผู้ตรวจให้คะแนน” สรุปได้ว่า ค่าความเชื่อมั่นแยกส่วนของผู้สอบและข้อสอบที่มีรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคนมีค่าสูงกว่า รูปแบบการตรวจให้คะแนนรูปแบบอื่นในทุกเงื่อนไขจำนวนผู้ตรวจ ส่วนค่าความเชื่อมั่นแยกส่วนของผู้ตรวจเป็นไปตามสมมติฐานทุกเงื่อนไข ยกเว้น ค่าความเชื่อมั่นแยกส่วนของผู้ตรวจ ในกรณีผู้ตรวจ 2 คน ที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคนมีค่าสูงกว่าผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคนอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และในกรณีผู้ตรวจ 3 คน ที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคนและผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคนมีค่าเท่ากัน

2.2.2 เมื่อรูปแบบการตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจต่างกัน เปรียบเทียบตามสมมติฐานที่ตั้งไว้ในข้อ 2.2 ที่ว่า “ค่าความเชื่อมั่นแยกส่วนของผู้สอบและข้อสอบ ที่มีจำนวนผู้ตรวจที่มากกว่า จะมีค่าสูงกว่า จำนวนผู้ตรวจที่น้อยกว่า ส่วนค่าความเชื่อมั่นแยกส่วนของผู้ตรวจ จะมีค่าต่ำกว่า จำนวนผู้ตรวจที่น้อยกว่า ในทุกรูปแบบการตรวจให้คะแนน” สรุปได้ว่า ความเชื่อมั่นแยกส่วนของผู้สอบที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคนที่มีจำนวนผู้ตรวจที่มากกว่ามีค่าสูงกว่าจำนวนผู้ตรวจที่น้อยกว่าอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ความเชื่อมั่นแยกส่วนของ

ข้อสอบที่ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคนที่ใช้ผู้ตรวจ 4 คน มีค่าสูงกว่าผู้ตรวจ 2 คน ผู้ตรวจจำนวน 3 คนมีค่าสูงกว่าผู้ตรวจ 2 คน อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และความเชื่อมั่นแยกส่วนของผู้ตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคนที่มีจำนวนผู้ตรวจที่มากกว่ามีค่าต่ำกว่าจำนวนผู้ตรวจที่น้อยกว่าอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และในกรณีผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคนที่ผู้ตรวจจำนวน 4 คน มีค่าต่ำกว่าผู้ตรวจ 2 คน และผู้ตรวจ 3 คนมีค่าต่ำกว่าผู้ตรวจ 2 คน และกรณีผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคนที่ผู้ตรวจ 4 คน มีค่าต่ำกว่าผู้ตรวจ 3 คน อย่างมีนัยสำคัญทางสถิติที่ระดับ .01

2.3 การประมาณค่าและเปรียบเทียบความเที่ยงตรงตามสภาพของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน พิจารณาจากค่าสัมประสิทธิ์สหสัมพันธ์ของคะแนนความสามารถกับคะแนนที่ได้จากผลการทดสอบทางการศึกษาระดับชาติด้านพื้นฐาน (O-NET) พบว่า ค่าสัมประสิทธิ์สหสัมพันธ์ในทุกเงื่อนไขมีค่าใกล้เคียงกัน (r มีค่าตั้งแต่ 0.769 ถึง 0.791) และมีนัยสำคัญทางสถิติที่ระดับ .01 ในทุกเงื่อนไข เมื่อนำมาทดสอบความแตกต่าง พบว่า ค่าสัมประสิทธิ์สหสัมพันธ์ของคะแนนความสามารถกับคะแนนที่ได้จากผลการทดสอบทางการศึกษาระดับชาติด้านพื้นฐาน ในแต่ละเงื่อนไขจำนวนผู้ตรวจและรูปแบบการตรวจที่ต่างกัน มีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติที่ระดับ .01 แสดงว่า คะแนนความสามารถที่ได้จากทุกเงื่อนไขจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนน มีความเที่ยงตรงตามสภาพไม่แตกต่างกัน ซึ่งไม่เป็นไปตามสมมติฐานข้อ 2.3 ที่ว่า “ค่าความเที่ยงตรงตามสภาพของคะแนนที่มีรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน จะมีค่าสูงกว่า รูปแบบการตรวจให้คะแนนรูปแบบอื่น ในทุกจำนวนผู้ตรวจให้คะแนน” และข้อ 2.4 ที่ว่า “ค่าความเที่ยงตรงตามสภาพของคะแนนที่มีจำนวนผู้ตรวจที่มากกว่า จะมีค่าสูงกว่า จำนวนผู้ตรวจที่น้อยกว่า ในทุกรูปแบบการตรวจให้คะแนน”

อภิปรายผล

การอภิปรายผลการวิจัยที่นำเสนอในตอนนี้อยู่จัดแบ่งเป็น 3 ประเด็น คือ 1) การศึกษาความแปรปรวนของแต่ละองค์ประกอบ 2) การศึกษาค่าความเชื่อมั่น และ 3) การศึกษาค่าความเที่ยงตรงตามสภาพ มีรายละเอียดดังต่อไปนี้

1. ความแปรปรวนของแต่ละองค์ประกอบ

ผู้วิจัยแบ่งการอภิปรายผลการศึกษาค่าความแปรปรวนของแต่ละองค์ประกอบออกเป็น 2 ประเด็นคือ 1) ความแปรปรวนของแต่ละองค์ประกอบที่ประมาณค่ามาจากโมเดลการสุ่มปัจจัย และ

2) ความแปรปรวนของแต่ละองค์ประกอบที่ประมาณค่ามาจากโมเดลหลายองค์ประกอบของราล์ชดังนี้

1.1 การศึกษาความแปรปรวนขององค์ประกอบที่ประมาณค่ามาจากโมเดลการสรุปอ้างอิง เมื่อใช้รูปแบบการตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจต่างกัน พบว่า ในรูปแบบการตรวจให้คะแนนที่เหมือนกัน ความแปรปรวนขององค์ประกอบเดียวกันในทุกจำนวนผู้ตรวจมีค่าใกล้เคียงกัน ทั้งนี้องค์ประกอบที่มีความแปรปรวนมากที่สุดในรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน คือ ปฏิสัมพันธ์ระหว่างผู้สอบกับข้อสอบ ส่วนในรูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคนและผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน องค์ประกอบที่มีความแปรปรวนมากที่สุดคือ ปฏิสัมพันธ์ระหว่างผู้สอบกับข้อสอบในแต่ละผู้ตรวจและความคลาดเคลื่อนเชิงสุ่มที่ไม่อาจระบุได้ แต่อย่างไรก็ตามเมื่อพิจารณาองค์ประกอบหลักแล้ว ผู้สอบและข้อสอบมีความแปรปรวนมากกว่าผู้ตรวจ โดยที่ความแปรปรวนของผู้ตรวจในทุกเงื่อนไขจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนมีค่าค่อนข้างน้อย ทั้งนี้การที่ความแปรปรวนของปฏิสัมพันธ์ระหว่างผู้สอบกับข้อสอบมีมาก แสดงว่า อันดับของนักเรียนมีความแตกต่างกันไปตามการตอบข้อสอบหรือมีการเปลี่ยนแปลงไปตามข้อสอบ หรือผู้สอบไม่มีความคงเส้นคงวาในการตอบข้อสอบแต่ละข้อ ซึ่งความสามารถของผู้สอบที่ไม่สอดคล้องกันตามการเปลี่ยนแปลงของข้อสอบ สอดคล้องกับผลการศึกษาของเบกเทอร์และคณะ (Baxter; et al. 1993) ชาเวลสันและคณะ (Shavelson; et al. 1993) เลนนีและคณะ (Lane; et al. 1996) ชูและวูล์ฟ (Chiu; & Wolfe. 2008) ที่พบว่า ความแปรปรวนของปฏิสัมพันธ์ระหว่างผู้สอบและข้อสอบมีค่าสูงที่สุด ทั้งนี้อาจเนื่องมาจากแบบทดสอบที่ผู้วิจัยสร้างขึ้นนี้ข้อสอบแต่ละข้อมีระดับความยากที่แตกต่างกันตั้งแต่ค่อนข้างง่ายจนถึงค่อนข้างยาก ประกอบกับเนื้อหาครอบคลุมสาระการเรียนรู้คณิตศาสตร์ทั้ง 6 สาระในหลักสูตรมัธยมศึกษาตอนต้นทั้งหมด จึงอาจทำให้มีเนื้อหาสาระที่หลากหลาย ผู้สอบบางคนอาจมีความเชี่ยวชาญหรือชำนาญในบางเนื้อหาต่างกัน จึงทำให้ความสามารถในการตอบข้อสอบแต่ละข้อไม่เท่ากัน สอดคล้องคำกล่าวของไหหยาง (Haiyang. 2010: 98) ที่ว่าความแปรปรวนของปฏิสัมพันธ์ของผู้สอบกับข้อสอบมีค่าสูงนั้นอาจเกิดจากความยากของข้อสอบที่แตกต่างกันในแต่ละข้อ และผู้สอบไม่ได้มีความรู้ความสามารถที่เท่าเทียมกันในการตอบข้อสอบแต่ละข้อ ประกอบกับความไม่สอดคล้องกันในการให้คะแนนของผู้ตรวจ

นอกจากนี้ในรูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน เป็นลักษณะการตรวจที่ผู้สอบแนบในผู้ตรวจและข้อสอบแนบในผู้ตรวจ ซึ่งองค์ประกอบที่มีความแปรปรวนมากที่สุดคือ ปฏิสัมพันธ์ระหว่างผู้สอบกับข้อสอบในแต่ละผู้ตรวจและความคลาดเคลื่อนเชิงสุ่มที่ไม่อาจระบุได้ แสดงว่า คะแนนสังเกตจากปฏิสัมพันธ์ของผู้สอบ

กับข้อสอบไม่ได้มีค่าเท่ากันตลอดตามผู้ตรวจที่ต่างกัน รวมทั้งแหล่งความคลาดเคลื่อนที่ไม่อาจระบุได้ ทั้งนี้อาจเนื่องมาจากการสุ่มกลุ่มผู้สอบให้กับผู้ตรวจแต่ละคน ซึ่งผู้วิจัยใช้การสุ่มอย่างง่าย หรือจากการสุ่มกลุ่มข้อสอบให้กับผู้ตรวจแต่ละคน ซึ่งผู้วิจัยใช้การสุ่มแบบแบ่งชั้น โดยมีสาระการเรียนรู้เป็นชั้น และข้อสอบเป็นหน่วยการสุ่ม ทำให้ได้กลุ่มข้อสอบที่ประกอบด้วยสาระการเรียนรู้ต่างกัน จึงอาจทำให้ผู้สอบในแต่ละกลุ่มมีความสามารถแตกต่างกันในแต่ละผู้ตรวจ เช่น ในรูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน ผู้ตรวจแต่ละคนจะได้ตรวจข้อสอบของผู้สอบที่แตกต่างกัน ซึ่งในกรณีผู้ตรวจ 4 คน นั้นผู้ตรวจคนที่ 1 ได้ตรวจข้อสอบของกลุ่มผู้สอบที่มีความสามารถค่อนข้างน้อย แต่ผู้ตรวจคนที่ 4 ได้ตรวจข้อสอบของกลุ่มผู้สอบที่มีความสามารถค่อนข้างสูง เป็นต้น หรือผู้ตรวจแต่ละคนได้ตรวจกลุ่มข้อสอบที่มีความยากแตกต่างกัน เช่น ในรูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน ในกรณีผู้ตรวจ 4 คน ผู้ตรวจคนที่ 1 สุ่มได้ตรวจข้อสอบข้อที่ค่อนข้างง่ายถึงง่าย คือข้อ 1, 5, 9 ส่วนผู้ตรวจคนที่ 4 ได้ตรวจข้อสอบที่ค่อนข้างยากถึงยาก คือข้อ 4, 8 และ 12 หรือในกรณีผู้ตรวจ 2 คน ผู้ตรวจคนที่ 1 สุ่มได้ข้อสอบที่อยู่ในกลุ่มข้อง่ายเป็นส่วนใหญ่คือข้อ 1, 3, 5, 7, 9 และ 11 ส่วนผู้ตรวจคนที่ 2 ได้ตรวจข้อสอบข้อที่เหลือซึ่งเป็นกลุ่มข้อยาก ประกอบกับผู้สอบอาจมีความเชี่ยวชาญในเนื้อหาต่างกัน ไม่ได้มีความรู้ที่เท่าเทียมกันในทุกเนื้อหา จึงทำให้ความสามารถในการตอบข้อสอบแต่ละข้อไม่เท่ากัน รวมกับความแปรปรวนที่ไม่อาจระบุได้ นั้นแสดงว่าหลังจากพยายามอธิบายความไม่สอดคล้องกันนี้ด้วยองค์ประกอบต่าง ๆ ทุกแหล่งแล้ว ก็ยังมีค่าความแปรปรวนอีกจำนวนหนึ่งที่ไม่สามารถอธิบายได้ ดังนั้นหากต้องการอ้างอิงไปยังความสามารถผู้สอบที่เชื่อถือได้มากขึ้น ความแปรปรวนแหล่งนี้ควรมีค่าน้อยหรือไม่มี ทั้งนี้เพราะเป้าหมายในการทดสอบคือต้องการให้ผู้สอบแสดงความสามารถให้มากที่สุดภายใต้ข้อสอบแต่ละข้อ (Smith; & Kulikowich. 2004: 619) ดังนั้น หากใช้รูปแบบการตรวจสองรูปแบบนี้ควรหาพยานหลักฐาน (Evidences) เพิ่มหรือเพิ่มองค์ประกอบในการศึกษาแหล่งความแปรปรวนให้มากขึ้น เช่น จำนวนครั้งในการตรวจ จำนวนระดับคะแนน เพื่อการอ้างอิงที่เชื่อถือได้ไปยังความสามารถของผู้สอบ

เมื่อพิจารณาความแปรปรวนของผู้สอบ ซึ่งมีค่าความแปรปรวนสูงรองลงมาในทุกเงื่อนไข จำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนน แสดงถึง ความสามารถในการแก้ปัญหาคณิตศาสตร์ของผู้สอบแต่ละคนที่แตกต่างกัน ซึ่งเป็นแหล่งความแปรปรวนที่ต้องการให้มีค่าสูง (Sudweeks; Reeve; & Bradshaw. 2005: 248) ซึ่งในทฤษฎีการสรุปอ้างอิงถือว่าองค์ประกอบผู้สอบเป็นสิ่งที่ต้องการวัด (Object of measurement) ไม่จัดว่าเป็นแหล่งความคลาดเคลื่อนของการวัด (Smith; & Kulikowich. 2004: 619) ทั้งนี้ความแปรปรวนของผู้สอบที่มีค่าสูง อาจเนื่องมาจากความสามารถในการแก้ปัญหาของผู้สอบที่ย่อมแตกต่างกันไปตามประสบการณ์พื้นฐาน (Sudweeks; Reeve; & Bradshaw. 2005: 248) ประกอบกับกลุ่มตัวอย่างของผู้วิจัยได้มาโดยการสุ่ม ทำให้เป็นตัวแทนที่ดีของ

ประชากร ที่มีความสามารถหลากหลายตั้งแต่น้อยถึงมาก จึงทำให้ได้ผู้สอบที่มีความสามารถทางคณิตศาสตร์ที่แตกต่างกัน สอดคล้องกับงานวิจัยของbachman โลงซ์และมาสัน (Bachman; Lynch; & Mason.1993: online) ที่พบว่า องค์ประกอบความแปรปรวนที่มีขนาดใหญ่ที่สุดคือ ผู้สอบ คิดเป็น 91 % ส่วนผู้ตรวจและข้อสอบมีความแปรปรวนน้อย แสดงว่า ผู้ตรวจมีความเข้มงวดใกล้เคียงกันและข้อสอบมีความยากเท่า ๆ กัน สอดคล้องกับงานวิจัยของsmithและคูลิกอวิช (Smith; & Kulikowich. 2004) ที่พบว่า แหล่งของความแปรปรวนที่มากที่สุดคือ ผู้สอบ รองลงไปคือ ข้อสอบ ส่วนความแปรปรวนของผู้ตรวจมีน้อย นั่นคือ ผู้ตรวจมีความแตกต่างในการให้คะแนนน้อย

ในส่วนของความแปรปรวนของข้อสอบที่มีค่าไม่มากนัก แสดงว่ามีความแตกต่างอย่างมีระบบของผู้สอบในการตอบข้อสอบไม่มากนัก ทำให้ค่าเฉลี่ยของคะแนนในแต่ละข้อสอบต่างกัน แสดงถึงความยากของข้อสอบที่ต่างกัน อาจเนื่องมาจากธรรมชาติของข้อสอบแต่ละข้อ ความแตกต่างของความรู้พื้นฐานของผู้สอบ หรือความสนใจในประเด็นที่ถามหรือเงื่อนไขในสิ่งที่โจทย์กำหนดไว้ (Sudweeks; Reeve; & Bradshaw. 2005: 248) ทำให้ข้อสอบบางข้อสะท้อนทักษะการแก้ปัญหาได้มากกว่าข้ออื่น ๆ (Smith; & Kulikowich. 2004: 619)

นอกจากนี้ผลการศึกษาความแปรปรวนในทุกเงื่อนไขจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนน พบว่า ความแปรปรวนของผู้ตรวจมีค่าค่อนข้างน้อยและน้อยกว่าองค์ประกอบอื่น แสดงให้เห็นว่า ผู้ตรวจมีความแตกต่างกันในการให้คะแนนน้อย หรือผู้ตรวจส่งผลต่อความแตกต่างของคะแนนค่อนข้างน้อย นั่นคือ ผู้ตรวจมีความแม่นยำในการนำเกณฑ์การให้คะแนนที่กำหนดไปตรวจให้คะแนนผลการตอบของผู้สอบ และในบางเงื่อนไข ได้แก่ เมื่อผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคนและใช้ผู้ตรวจ 4 คนและ 2 คน มีค่าความแปรปรวนของผู้ตรวจเป็นศูนย์ นั้นแสดงว่า สามารถจัดความแปรปรวนของผู้ตรวจได้อย่างสิ้นเชิง ทั้งนี้เนื่องจากผู้ตรวจกลุ่มนี้ได้รับการฝึกอบรมการตรวจให้คะแนนตามเกณฑ์ที่ผู้วิจัยสร้างขึ้นในคู่มือการตรวจ พร้อมกับมีแนวปฏิบัติการตรวจให้คะแนนที่ชัดเจน ซึ่งช่วยให้ผู้ตรวจมีความเห็นที่สอดคล้องกันมากขึ้น ประกอบกับผู้สอบได้รับการปิดชื่อโรงเรียน จึงทำให้อคติของผู้ตรวจถูกขจัดทิ้งไปและมีความเป็นปรนัยของการให้คะแนนมากขึ้น สอดคล้องกับผลการวิจัยของเบรนนอน เกาและคอลตัน (Brennan; Gao; & Colton. 1995) เลนนีและคณะ (Lane; et al. 1996: 87) เกรล (Gierl. 1998) smithและคูลิกอวิช (Smith; & ulikowich. 2004) ทินดัล โยวานอฟและเกลเลอร์(Tindal; Yovanoff; & Geller. 2010) ที่พบว่า ค่าความแปรปรวนของผู้ตรวจหรือความคลาดเคลื่อนจากผู้ตรวจมีน้อยมาก เนื่องมาจากการให้รายละเอียดของเกณฑ์การตรวจและมีการฝึกอบรมผู้ตรวจให้เข้าใจวิธีการตรวจและกฎเกณฑ์การตรวจ สอดคล้องกับคำกล่าวของเมอเรนส์และลีแมน (Mehrens; & Lehmann. 1972: 228) ดันบาร์ โคเรทซ์และฮูเวอร์ (Dunbar; Koretz; & Hoover. 1991) ที่ว่าความแปรปรวนของผู้ตรวจสามารถควบคุมได้โดยการออกแบบ

กฎเกณฑ์การให้คะแนน (Scoring rubric) ให้เหมาะสม ชัดเจนและมีการฝึกอบรมผู้ตรวจอย่างเหมาะสมเพียงพอ ใช้เกณฑ์การตรวจเดียวกันกับผู้สอบทุกคนเพื่อเพิ่มความน่าเชื่อถือของผลการตรวจ และคำกล่าวของอีเบล (Ebel. 1951) รูเนอร์ (Rudner. 1992: online) กุลเลอร์และเกิลบัล (Guler; & Gelbal. 2010) ที่ว่าแม้ว่าผู้ตรวจจะเป็นแหล่งความคลาดเคลื่อนที่สำคัญมากของการตรวจให้คะแนน คำถามปลายเปิด แต่ถ้าหากมีการควบคุมแหล่งความคลาดเคลื่อน เช่น ข้อสอบ ผู้ตรวจ วัสดุงาน จะทำให้ความคลาดเคลื่อนของการให้คะแนนจากผู้ตรวจลดลงได้

1.2 การศึกษาความแปรปรวนของแต่ละองค์ประกอบที่ประมาณค่ามาจากโมเดลหลายองค์ประกอบของราล์ช พบว่า องค์ประกอบที่มีความแปรปรวนมากที่สุดในทุกเงื่อนไขจำนวนผู้ตรวจ และรูปแบบการตรวจให้คะแนนเมื่อเทียบกับความแปรปรวนรวมคือ ผู้สอบ รองลงมาคือ ข้อสอบ และองค์ประกอบที่ให้ค่าความแปรปรวนน้อยที่สุดคือ ผู้ตรวจ แสดงว่า ความสามารถของผู้สอบมีความแตกต่างกันมากกว่าค่าความยากของข้อสอบและความเข้มงวดของผู้ตรวจ สอดคล้องกับผลการวิจัยของอัปเชอร์และเทอร์เนอร์ (Upshur; & Turner. 1999: 97) ซึ่งพบว่า องค์ประกอบที่มีความแปรปรวนมากที่สุดคือ ผู้สอบ รองลงมาคือ ข้อสอบและผู้ตรวจ ตามลำดับ และผลการวิจัยของบอนคและออกคีย์ (Bonk; & Ockey. 2003: 96-97) แมคมานัส ทอมป์สันและมุลลอน (McManus; Thompson; & Mollon. 2006: online) ที่พบว่าความแปรปรวนของผู้สอบมีค่าสูงสุด รองลงมาคือ ผู้ตรวจและข้อสอบ ตามลำดับ ทั้งนี้เหตุผลสามารถอธิบายได้ตาม 1.1

2. การศึกษาค่าความเชื่อมั่น

ผู้วิจัยแบ่งการอธิบายผลการศึกษาค่าความเชื่อมั่น ออกเป็น 2 ประเด็นคือ 1) การศึกษาค่าสัมประสิทธิ์การสรุปอ้างอิงจากทฤษฎีการสรุปอ้างอิง และ 2) การศึกษาค่าความเชื่อมั่นแยกส่วนของแต่ละองค์ประกอบจากโมเดลหลายองค์ประกอบของราล์ช ดังนี้

2.1 การศึกษาค่าสัมประสิทธิ์การสรุปอ้างอิง เมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน พบว่า รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคนให้ค่าสัมประสิทธิ์การสรุปอ้างอิงสูงที่สุด รองลงมาคือ เมื่อผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคนและผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน ตามลำดับ ในทุกจำนวนผู้ตรวจ และการเปรียบเทียบตามสมมติฐานข้อ 1.1 ที่ว่า “ค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนที่มีรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน จะมีค่าสูงกว่า รูปแบบการตรวจให้คะแนนรูปแบบอื่น ในทุกจำนวนผู้ตรวจให้คะแนน” พบว่า ค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนเมื่อใช้ผู้ตรวจจำนวน 4 คน และ 2 คน ที่มีรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคนมีค่าสูงกว่าผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน อย่างมีนัยสำคัญทางสถิติที่ระดับ

.01 ส่วนผู้ตรวจจำนวน 3 คนนั้นสูงกว่าอย่างไม่มีนัยสำคัญทางสถิติที่ระดับ .01 สอดคล้องกับผลการวิจัยของ สัดวิคส์ รีฟและแบรดชอว์ (Sudweeks; Reeve; & Bradshaw. 2005: 259) ที่พบว่า การใช้แบบแผนแบบแฝง (Nested) จะได้คะแนนที่มีค่าความเชื่อมั่นแตกต่างจากแบบแผนแบบไขว้สมบูรณ์ (Full crossed) เพียงเล็กน้อย (ประมาณ 1.5 – 2.0 %) เท่านั้น ทั้งนี้คะแนนสอบที่ได้จากรูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคนเกิดจากการนำผลตรวจของผู้ตรวจ 4 คน 3 คน และ 2 คน มาเฉลี่ยเป็นคะแนนประจำข้อของผู้สอบแต่ละคน ซึ่งเป็นรูปแบบการตรวจที่เป็นอุดมคติ (Ideal) เป็นรูปแบบการตรวจที่ยุติธรรมที่สุด (Linacre; & Wright. 2002: 491; Sudweeks; Reeve; & Bradshaw. 2005)

ส่วนค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคนนั้นมีค่าต่ำกว่ารูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคนในทุกจำนวนผู้ตรวจอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ซึ่งไม่เป็นไปตามสมมติฐานที่ตั้งไว้ ทั้งนี้อาจเนื่องมาจากการกระจายของคะแนนที่ต่างกัน กล่าวคือถ้ารูปแบบการตรวจใดมีการกระจายของคะแนนของผู้สอบมากจะทำให้ค่าสัมประสิทธิ์การสรุปอ้างอิงมีค่าสูง ถ้ามีการกระจายของคะแนนผู้สอบน้อย ค่าสัมประสิทธิ์การสรุปอ้างอิงจะมีค่าต่ำ ส่งผลให้ค่าสัมประสิทธิ์การสรุปอ้างอิงมีค่าแตกต่างกันในเชิงสถิติ (Statistic significance) ทั้งนี้ความเชื่อมั่นของแบบทดสอบขึ้นอยู่กับ การกระจายของคะแนน (อนันต์ ศรีโสภณ. 2525: 147) ซึ่งเมื่อพิจารณาคะแนนสอบ จะเห็นว่าการกระจายของคะแนนโดยเฉพาะคะแนนในรูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคนที่มีค่าพิสัยมากกว่ารูปแบบการตรวจอื่นในทุกจำนวนผู้ตรวจและมีค่าส่วนเบี่ยงเบนมาตรฐาน (SD) ของคะแนนมากกว่าผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคนโดยเฉพาะในกรณีผู้ตรวจ 3 คนและ 2 คน ซึ่งกรอนลันด์ (Gronlund. 1976: 117) ได้กล่าวไว้ว่า ถ้าความสามารถของผู้สอบในกลุ่มแตกต่างกันมากหรือมีการกระจายของคะแนนมากจะทำให้ความเชื่อมั่นสูงกว่าผู้สอบในกลุ่มที่มีความสามารถใกล้เคียงกันหรือมีการกระจายของคะแนนน้อย สอดคล้องกับคำกล่าวของคินนิงแฮม (Cunningham. 1986: 112-118) ที่ว่า หากความสามารถของผู้สอบมีหลากหลายหรือมีความแปรปรวนของคะแนนเพิ่มมากขึ้น จะส่งผลให้มีค่าความเชื่อมั่นสูงขึ้น ทั้งนี้การกระจายของคะแนนที่ต่างจากกันนี้ อาจเนื่องมาจากการสุ่มกลุ่มผู้สอบในแต่ละรูปแบบการตรวจให้คะแนนเพื่อให้ผู้ตรวจตรวจโดยไม่ซ้ำกัน ผู้วิจัยใช้วิธีการสุ่มอย่างง่าย จึงอาจทำให้สุ่มได้ผู้สอบที่มีความสามารถใกล้เคียงกันหรือแตกต่างกันมากอยู่ในกลุ่มเดียวกัน จึงทำให้ผลคะแนนของผู้สอบในผู้ตรวจบางคนมีค่าแตกต่างกันมาก นั่นคือคะแนนมีการกระจายมาก จึงทำให้ค่าความเชื่อมั่นสูงขึ้นได้ นอกจากนี้เมื่อพิจารณาคะแนนที่ได้จากรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคนมีคะแนนสุดโต่งเกิดขึ้น คือ มีผู้สอบได้คะแนนเต็ม ซึ่งอาจมีส่วนทำให้ค่าสัมประสิทธิ์การสรุปอ้างอิงมีค่าสูงขึ้น ดังคำกล่าวของคลอสเซอร์และลินาเคอร์ (Clausen; &

Linacre. 1999: online) ที่ว่าหากมีคะแนนสุดโต่ง (Extreme) เกิดขึ้น ในการประมาณค่าสัมประสิทธิ์การสรุปอ้างอิงจะนำค่าสุดโต่งนี้ไปคำนวณด้วย และเนื่องด้วยคะแนนสุดโต่งนี้ไม่มีความแปรปรวนของความคลาดเคลื่อนของคะแนน (Score error variance) ทำให้ค่าเฉลี่ยของความคลาดเคลื่อน (Average error) มีค่าลดลงส่งผลให้ค่าสัมประสิทธิ์การสรุปอ้างอิงมีค่าสูงขึ้น

เมื่อพิจารณาตามจำนวนผู้ตรวจที่ต่างกัน โดยใช้รูปแบบการตรวจให้คะแนนเหมือนกัน พบว่า จำนวนผู้ตรวจที่มากกว่าจะมีค่าสัมประสิทธิ์การสรุปอ้างอิงสูงกว่าจำนวนผู้ตรวจที่น้อยกว่าเมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน ส่วนในรูปแบบการตรวจอื่น ค่าสัมประสิทธิ์การสรุปอ้างอิงมีค่าสูงสุดเมื่อใช้ผู้ตรวจจำนวน 3 คน รองลงมาคือ 2 คนและ 4 คน ตามลำดับ และเมื่อเปรียบเทียบตามสมมติฐานข้อ 1.2 ที่ว่า “ค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนที่มีจำนวนผู้ตรวจที่มากกว่า จะมีค่าสูงกว่า จำนวนผู้ตรวจที่น้อยกว่า ในทุกรูปแบบการตรวจให้คะแนน” พบว่า ค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนที่ใช้รูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคนและมีผู้ตรวจจำนวน 4 คน มีค่าสูงกว่าผู้ตรวจจำนวน 2 คน และเมื่อใช้ผู้ตรวจ 3 คน มีค่าสูงกว่าผู้ตรวจ 2 คน อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ส่วนกรณีผู้ตรวจ 4 คนนั้นมีค่าสูงกว่า 3 คน อย่างไม่มีนัยสำคัญทางสถิติที่ระดับ .01 และค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนที่ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน เมื่อใช้ผู้ตรวจจำนวน 3 คน มีค่าสูงกว่าผู้ตรวจ 2 คน อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 แสดงว่าเมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน จำนวนผู้ตรวจที่มากกว่ามีแนวโน้มว่าจะให้ค่าสัมประสิทธิ์การสรุปอ้างอิงสูงกว่าจำนวนผู้ตรวจที่น้อยกว่า สอดคล้องกับคำกล่าวของ เมอเรนส์และลีแมน (Mehrens; & Lehmann. 1973: 234-236) ศิริชัย กาญจนวาสี (2550: 37) ที่ว่า “ไม่ว่าจะใช้รูปแบบการตรวจใด ถ้าเป็นไปได้ควรใช้ผู้ตรวจหลาย ๆ คนในการให้คะแนนผู้สอบแต่ละคนแล้วหาคะแนนเฉลี่ยหรือคะแนนรวมแทนความสามารถของผู้สอบจะช่วยลดความแปรปรวนอันเนื่องมาจากผู้ตรวจและความคลาดเคลื่อนที่เหลือ ซึ่งจะทำให้ค่าความเชื่อมั่นมีค่าเพิ่มสูงขึ้น และสอดคล้องกับงานวิจัยของโอเบรียนและโจนส์ (O'Brien; & Jones. 1986) เบรนนอน เกาและคอลตัน (Brennan; Gao; & Colton. 1995) สวาร์ทซ์และคณะ (Swartz; et al. 1999) สดวีคส์รีฟและบรัดชอว์ (Sudweeks; Reeve; & Bradshaw. 2005) ชัยฤทธิ์ ศิลาดเดช (2540) และอุษณีย์ บัวศิริพันธ์ (2543) ที่ใช้รูปแบบการตรวจให้คะแนนแบบไขว้ (Fully crossed) แล้วพบว่าจำนวนผู้ตรวจมีอิทธิพลต่อค่าความเชื่อมั่น ทั้งนี้ความเชื่อมั่นมีค่าสูงขึ้นเมื่อจำนวนผู้ตรวจเพิ่มขึ้น

แต่อย่างไรก็ตามเมื่อพิจารณาค่าสัมประสิทธิ์การสรุปอ้างอิงในทุกเงื่อนไขจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนถือว่ามีค่าค่อนข้างสูง (มีค่าตั้งแต่ 0.882 ถึง 0.975) การที่มีความเชื่อมั่นสูง อาจเกิดจากการควบคุมแหล่งความคลาดเคลื่อนไว้ล่วงหน้าของผู้วิจัย ซึ่งองค์ประกอบสำคัญที่ต้อง

ควบคุมคือข้อสอบ และผู้ตรวจ ซึ่งผู้วิจัยควบคุมองค์ประกอบข้อสอบโดยให้ผู้สอบทุกคนทำข้อสอบชุดเดียวกัน และผู้ตรวจทุกคน ควบคุมโดยการฝึกอบรม ซึ่งแรงกดดันการให้คะแนน (Ebel. 1951) และเมื่อพิจารณาความแตกต่างของค่าสัมประสิทธิ์การสรุปอ้างอิงที่ได้ถือว่าไม่ต่างกันมากนัก สอดคล้องกับผลการวิจัยของ สัตวิคส์ รีฟและแบรดชอว์ (Sudweeks; Reeve; & Bradshaw. 2005: 259) ที่พบว่าการใช้แบบแผนแบบแฝง (Nested) ที่ให้ผู้ตรวจแต่ละคนตรวจข้อสอบทุกข้อของนักเรียนที่ต่างกัน จะได้คะแนนที่มีค่าความเชื่อมั่นแตกต่างจากแบบแผนแบบไขว้สมบูรณ์ (Fully crossed) ไม่มากนัก แสดงว่า แม้ว่าค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนในแต่ละเงื่อนไขจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนน จะมีค่าแตกต่างกัน แต่ยังมีค่าสูงเกิน 0.88 ในทุกเงื่อนไข ดังนั้น จึงสามารถเลือกใช้รูปแบบการตรวจให้คะแนนและจำนวนผู้ตรวจที่เหมาะสมกับการปฏิบัติได้ แม้ว่ารูปแบบการตรวจให้คะแนนที่ยุติธรรมที่สุดคือการให้ผู้ตรวจทุกคนตรวจข้อสอบทุกข้อของผู้สอบทุกคน แต่รูปแบบการตรวจลักษณะนี้ทำได้ยาก ต้องใช้เวลาในการตรวจค่อนข้างมาก (Time-consuming) และมีค่าใช้จ่ายสูง (Expensive) (Linacre; & Wright. 2002: 492; Sudweeks; Reeve; & Bradshaw. 2005) โดยเมื่อพิจารณาในด้านเวลาที่ใช้ในการตรวจ พบว่า โดยเฉลี่ยผู้ตรวจใช้เวลาตรวจข้อละประมาณ 3 นาที หากตรวจข้อสอบทุกข้อของผู้สอบทุกคน ผู้ตรวจแต่ละคนจะใช้เวลาตรวจประมาณ $(3 \times 12 \times 180) = 6480$ นาที หรือ 108 ชั่วโมง แต่หากใช้รูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคนหรือผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน กรณีนี้ให้ผู้ตรวจ 4 คน ผู้ตรวจแต่ละคนจะใช้เวลาประมาณ 27 ชั่วโมง เท่านั้น ซึ่งประหยัดเวลาในการตรวจได้ค่อนข้างมาก

นอกจากนี้เมื่อพิจารณาจำนวนครั้งที่ผู้ตรวจแต่ละคนจะต้องตรวจ เช่น ในกรณีผู้ตรวจจำนวน 4 คน เมื่อใช้รูปแบบที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน พบว่า จำนวนครั้งที่ผู้ตรวจแต่ละคนตรวจเป็น 2160 ครั้ง, 540 ครั้ง และ 540 ครั้ง ตามลำดับ จะเห็นว่า รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคนและตรวจข้อสอบบางข้อของผู้สอบทุกคน ผู้ตรวจแต่ละคนจะตรวจเพียง 25 % ของรูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน และหากใช้ผู้ตรวจจำนวน 3 คน และ 2 คน ผู้ตรวจแต่ละคนจะตรวจ 50 % และ 75 % ของรูปแบบที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน ตามลำดับ ดังนั้น แม้ว่าจำนวนผู้ตรวจที่มากกว่าจะมีแนวโน้มให้ค่าความเชื่อมั่นของคะแนนสูงกว่าจำนวนผู้ตรวจที่น้อยกว่า แต่แตกต่างกันไม่มากนัก การวางแผนการตรวจที่เหมาะสม จึงสามารถลดภาระ ลดเวลาในการตรวจ แต่ยังคงไว้ซึ่งความน่าเชื่อถือของคะแนน และมีค่าทางสถิติที่ยอมรับได้ (Linacre. 2002: 490-493; Nakamura. 2002: online) ทำให้เกิดประสิทธิภาพของการใช้เวลา (Time efficient) และค่าใช้จ่าย (Cost efficient) (Bonk; & Ockey. 2003: 90) อีกทั้งยังสอดคล้องกับความเป็นไปได้หรือความสะดวกในเชิงปฏิบัติ (ศิริชัย กาญจนวาสี.

2550: 36) โดยเฉพาะการตรวจให้คะแนนแบบทดสอบในระดับโรงเรียน ที่มีครูผู้ตรวจที่มีความเชี่ยวชาญจำนวนจำกัดและมีภาระงานค่อนข้างมาก

2.2 การประมาณค่าความเชื่อมั่นแยกส่วนในแต่ละองค์ประกอบ เมื่อเปรียบเทียบตามสมมติฐานที่ตั้งไว้ในข้อ 2.1 ที่ว่า “ค่าความเชื่อมั่นแยกส่วนของผู้สอบและข้อสอบ ที่มีรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน จะมีค่าสูงกว่า รูปแบบการตรวจให้คะแนนรูปแบบอื่น ส่วนค่าความเชื่อมั่นแยกส่วนของผู้ตรวจ จะมีค่าต่ำกว่า รูปแบบการตรวจให้คะแนนรูปแบบอื่น ในทุกจำนวนผู้ตรวจให้คะแนน” สรุปได้ว่า ค่าความเชื่อมั่นแยกส่วนของผู้สอบและข้อสอบที่มีรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคนมีค่าสูงกว่ารูปแบบการตรวจให้คะแนนรูปแบบอื่นในทุกเงื่อนไขจำนวนผู้ตรวจ ทั้งนี้อาจเกิดจากรูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคนและผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคนนั้นมีลักษณะที่ผู้สอบแนบในผู้ตรวจและข้อสอบแนบในผู้ตรวจ ทำให้ผู้ตรวจได้ตรวจข้อสอบของผู้สอบที่แตกต่างกันหรือได้ตรวจข้อสอบต่างข้อกัน เป็นลักษณะของข้อมูลขาดหาย (Missing data) ซึ่งส่งผลให้ค่าความเชื่อมั่นแยกส่วนของผู้สอบมีค่าต่ำลง ดังที่แอนดริช (Andrich, 1982: online) ได้กล่าวไว้ หากมีข้อมูลขาดหายหรือไม่สมบูรณ์ (Incomplete data) จะทำให้การประมาณค่าความคลาดเคลื่อนมาตรฐานมีค่าสูงขึ้น ซึ่งส่งผลให้ค่าความเชื่อมั่นแยกส่วนของผู้สอบมีค่าลดลง ประกอบกับมีคะแนนสุดโต่ง (Extreme scores) เกิดขึ้น ซึ่งในการประมาณค่าความเชื่อมั่นแยกส่วนของผู้สอบใน MFRM จะไม่นำคะแนนสุดโต่งนี้มาคำนวณเมื่อนำออกไปจึงทำให้ค่าความเชื่อมั่นแยกส่วนของผู้สอบลดลง สอดคล้องกับคำกล่าวของฟิชเชอร์ (Fisher, 1936: online) ที่ว่าหากมีคะแนนสุดโต่งหรือมีข้อมูลขาดหายจะส่งผลทำให้ค่าความเชื่อมั่นแยกส่วนของผู้สอบมีค่าลดลง ในขณะที่ค่าความเชื่อมั่นแยกส่วนของข้อสอบที่มีค่าลดลงนั้นอาจเกิดจากผู้ตรวจได้ตรวจข้อสอบต่างข้อกัน เช่น ผู้ตรวจบางคนได้ชุดข้อสอบที่ง่าย บางคนได้ชุดข้อสอบที่ยาก จึงทำให้ความแปรปรวนของค่าความยากของข้อสอบที่ผู้ตรวจแต่ละคนได้รับมีค่าน้อยและทำให้ค่าความเชื่อมั่นแยกส่วนของข้อสอบมีค่าลดลงตามมาเช่นกัน ซึ่งหากความแปรปรวนของค่าความยากของข้อสอบว่ามีมากจะทำให้ค่าความเชื่อมั่นแยกส่วนของข้อสอบมีค่าสูงขึ้น (Fisher, 1936: online; Smith; & Kulikowich, 2004: 619)

เมื่อเปรียบเทียบค่าความเชื่อมั่นแยกส่วนของแต่ละองค์ประกอบตามสมมติฐานที่ตั้งไว้ในข้อ 2.2 ที่ว่า “ค่าความเชื่อมั่นแยกส่วนของผู้สอบและข้อสอบ ที่มีจำนวนผู้ตรวจที่มากกว่า จะมีค่าสูงกว่าจำนวนผู้ตรวจที่น้อยกว่า ส่วนค่าความเชื่อมั่นแยกส่วนของผู้ตรวจ จะมีค่าต่ำกว่า จำนวนผู้ตรวจที่น้อยกว่า ในทุกรูปแบบการตรวจให้คะแนน” พบว่า ความเชื่อมั่นแยกส่วนของผู้สอบในกรณีที่ใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคนที่จำนวนผู้ตรวจที่มากกว่ามีค่าสูงกว่าจำนวนผู้ตรวจที่น้อยกว่า และความเชื่อมั่นแยกส่วนของข้อสอบมีแนวโน้มว่ามีค่าไม่แตกต่าง

กันในทุกรูปแบบการตรวจให้คะแนนและจำนวนผู้ตรวจ และความเชื่อมั่นแยกส่วนของผู้ตรวจที่จำนวนผู้ตรวจที่มากกว่ามีแนวโน้มจะมีค่าต่ำกว่าจำนวนผู้ตรวจที่น้อยกว่าในทุกรูปแบบการตรวจให้คะแนน แสดงว่าจำนวนผู้ตรวจส่งผลต่อค่าความเชื่อมั่นแยกส่วนของแต่ละองค์ประกอบ แต่เมื่อพิจารณาว่าความเชื่อมั่นแยกส่วนของแต่ละองค์ประกอบจะเห็นว่ามีค่าแตกต่างกันเพียงเล็กน้อย

เมื่อพิจารณาโดยภาพรวมจะเห็นว่า ค่าความเชื่อมั่นแยกส่วนของแต่ละองค์ประกอบในทุกเงื่อนไขจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนน ในองค์ประกอบผู้สอบให้ค่าความเชื่อมั่นสูง คือ มีค่าตั้งแต่ 0.87 ถึง 0.99 ซึ่งเป็นตามที่คาดหวังคือต้องการให้มีค่าสูง เพื่อแสดงถึงความสามารถของผู้สอบว่ามีการกระจายตลอดช่วงความสามารถ (Continuum) (Smith; & Kulikowich. 2004: 629) ซึ่งค่าดังกล่าวสามารถอธิบาย แปลความได้เช่นเดียวกับ KR 20 ครอนบาคและค่าสัมประสิทธิ์การสรุปอ้างอิง (Wright; & Master. 1982; Engelhard. 1994: 97; Nakamura. 2000: 4) และค่าที่ได้ทุกค่าในทุกเงื่อนไขมีนัยสำคัญทางสถิติ แสดงว่า ความสามารถของผู้สอบมีการกระจายอยู่ตลอดช่วงความสามารถอย่างมีนัยสำคัญทางสถิติ นั่นคือ มีความเชื่อถือได้ของคะแนนความสามารถของผู้สอบที่ได้จากข้อสอบและผู้ตรวจ (Upshur; & Turner. 1999: 97) ทั้งนี้อาจเนื่องมาจากการสุ่มกลุ่มตัวอย่างที่มีความสามารถทางคณิตศาสตร์แตกต่างกันคือมีทั้งผู้สอบที่มีความสามารถต่ำจนถึงมีความสามารถสูง ประกอบกับข้อสอบที่ใช้มีค่าความยากกระจายตลอดช่วงความสามารถเช่นกัน ทำให้มีประสิทธิภาพในการวัดประเมินผู้สอบ ในส่วนของค่าความเชื่อมั่นแยกส่วนของข้อสอบ พบว่ามีค่าความเชื่อมั่นค่อนข้างสูงเช่นกัน คือ มีค่าตั้งแต่ 0.89 ถึง 0.99 เป็นไปตามที่คาดหวังไว้ แสดงว่าความยากของข้อสอบมีความแตกต่างกันมีข้อที่ยากจนถึงข้อที่ง่าย และความเชื่อมั่นแยกส่วนของข้อสอบมีนัยสำคัญทางสถิติ นั่นคือ ความยากของข้อสอบมีการกระจายอยู่ตลอดช่วงความสามารถ (Continuum) อย่างมีนัยสำคัญทางสถิติ สอดคล้องกับผลการวิจัยของเอนเกลฮาร์ด (Engelhard. 1992) มายฟอร์ด และวูล์ฟ (Myford; & Wolfe. 2000: online) นากามูระ (Nakamura. 2000: online) บัสเตอร์ก (Basturk. 2008) บาซแมน ไลนซ์ และมาสัน (Bachman; Lynch; & Mason. 1993: online) สมิทและคูลิกอวิช (Smith; & Kulikowich. 2004) กูเลอร์และเกลบัล (Guler; & Gelbal. 2010) ที่พบว่าค่าความเชื่อมั่นแยกส่วนของผู้สอบมีค่าสูงเกิน 0.92 และผลการศึกษาของสมิทและคูลิกอวิช (Smith; & Kulikowich. 2004) บัสเตอร์ก (Basturk. 2008) และไหหยาง (Haiyang. 2010) ที่พบว่าค่าความเชื่อมั่นแยกส่วนของข้อสอบมีค่าสูงเกิน 0.87 และเมื่อพิจารณาร่วมกับความแปรปรวนของข้อสอบในโมเดลหลายองค์ประกอบของราสส์ พบว่า มีความสอดคล้องกัน นั่นคือ เมื่อความแปรปรวนของข้อสอบมีมาก ทำให้ความเชื่อมั่นแยกส่วนของข้อสอบมีค่าสูงด้วย (Smith; & Kulikowich. 2004: 619) และเมื่อพิจารณาความสอดคล้อง (Fit) ของข้อสอบกับโมเดล ข้อสอบส่วน

ใหญ่อยู่ในช่วงที่ยอมรับได้ แสดงว่าข้อสอบมีประสิทธิภาพในการวัด ทำให้ข้อมูลที่ได้มีประสิทธิภาพในการวัดผล (Linacre. 2007)

เมื่อพิจารณาค่าความเชื่อมั่นแยกส่วนของผู้ตรวจ พบว่า มีค่าความเชื่อมั่นสูง คือ มีค่าตั้งแต่ 0.89 ถึง 0.99 แสดงว่า ผู้ตรวจมีความแตกต่างกันในประเด็นของความเข้มงวดและความใจดีของการให้คะแนน เนื่องจากค่าความเชื่อมั่นแยกส่วนของผู้ตรวจแสดงถึงความแตกต่างกัน (Reliably different) ของระดับความเข้มงวด (Sudweeks; Reeve; & Bradshaw. 2005: 254) อาจเนื่องด้วยผู้ตรวจแต่ละคนได้ตรวจกลุ่มผู้สอบที่มีความสามารถแตกต่างกัน จึงทำให้คะแนนในแต่ละกลุ่มแตกต่างกัน ส่งผลให้มีความแตกต่างกันค่อนข้างมากระหว่างผู้ตรวจแต่ละคน (Myford; & Wolfe. 2004: 197-198) สอดคล้องกับผลการวิจัยของ เอนเกลฮาร์ด (Engelhard. 1992) นากามูระ (Nakamura. 2000: online) เทอร์เนอร์ (Turner. 2003: online) ปาร์ค (Park. 2004: online) กูเลอร์ และเกลบัล (Guler; & Gelbal. 2010) ที่พบความเชื่อมั่นแยกส่วนของผู้ตรวจค่อนข้างสูง ซึ่งแสดงถึงความเข้มงวด/ใจดีของผู้ตรวจที่มีความแตกต่างกัน แต่อย่างไรก็ตามเมื่อพิจารณาค่า INFIT และ OUTFIT จะพบว่า ผู้ตรวจมีความคงที่ในการให้คะแนนโดยภาพรวม หรือกล่าวอีกนัยหนึ่งว่า การให้คะแนนของผู้ตรวจทั้ง 4 คนนั้นอาจแตกต่างกัน แต่ผู้ตรวจแต่ละคนต่างมีความสอดคล้อง (Consistency) ของการให้คะแนนในข้อสอบทุกข้อ หรือมีความคงที่ในการให้คะแนนในข้อสอบทุกข้อ (Nakamura. 2000) แต่อย่างไรก็ตามหากผลต่างของค่าความเข้มงวดไม่เกิน 0.5 แสดงว่าอยู่ในช่วงที่ยอมรับได้ คือ ความเข้มงวด/ใจดีของผู้ตรวจไม่ต่างกันมากนัก (Guler; & Gelbal. 2010) นอกจากนี้เมื่อพิจารณาร่วมกับความแปรปรวนของผู้ตรวจ จะพบว่าความแปรปรวนของผู้ตรวจมีค่อนข้างต่ำเมื่อเทียบกับความแปรปรวนรวมทั้งหมด แสดงว่าความแปรปรวนของผู้ตรวจมีค่าน้อยเมื่อเทียบกับความแปรปรวนแหล่งอื่นๆ นั่นคือ ผู้ตรวจมีความแตกต่างกันไม่มากในการให้คะแนน โดยภาพรวมถึงแม้ผู้ตรวจจะมีความเข้มงวดหรือใจดีแตกต่างกันแต่การให้คะแนนไม่ได้แตกต่างกันมากนัก ความแตกต่างของคะแนนที่ได้จึงขึ้นกับจำนวนผู้ตรวจน้อยมาก สอดคล้องกับงานวิจัยของนากามูระ (Nakamura. 2002: online) ที่พบว่า คะแนนที่ได้ไม่ขึ้นอยู่กับการให้คะแนนของผู้ตรวจ การใช้ระบบการตรวจเป็นคู่สามารถลดภาระงานของผู้ตรวจได้ โดยที่ยังมีความน่าเชื่อถือทางสถิติ ทั้งนี้ผู้ตรวจทั้ง 4 คนไม่จำเป็นต้องตรวจข้อสอบนักเรียนทุกคน ขัดแย้งกับผลการวิจัยของ สมิทและคูลิโกวิช (Smith; & Kulikowich. 2004) ที่พบว่า ความเชื่อมั่นแยกส่วนของผู้ตรวจมีค่าเป็น 0.00 อย่างไรก็ตามแม้ว่าผู้ตรวจจะมีความสอดคล้องภายในตนเองหรือมีความคงที่ในการให้คะแนน แต่ควรได้รับการฝึกอบรมเพิ่มเติมเพื่อให้การให้คะแนนมีความสอดคล้องกับผู้ตรวจคนอื่น (Haiyang. 2010: 99) และควรให้ข้อมูลย้อนกลับ (Feedback) เพื่อให้เกิดความตระหนักในการให้คะแนน การสร้างข้อสอบ อันจะทำให้เกิดการปรับปรุงระบบการวัดผล (Examining) ต่อไป (Basturk. 2008: 442)

3. การศึกษาค่าความเที่ยงตรงตามสภาพ

ผู้วิจัยแบ่งการอภิปรายผลการศึกษาค่าความเที่ยงตรงตามสภาพ ออกเป็น 2 ประเด็น คือ 1) การศึกษาค่าความเที่ยงตรงตามสภาพของคะแนนสอบ และ 2) การศึกษาค่าความเที่ยงตรงตามสภาพของคะแนนความสามารถ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน ดังนี้

3.1 การศึกษาค่าความเที่ยงตรงตามสภาพ พิจารณาจากค่าสัมประสิทธิ์สหสัมพันธ์ของคะแนนสอบกับคะแนนที่ได้จากผลการทดสอบทางการศึกษาระดับชาติด้านพื้นฐาน (Ordinary National Education Test: O-NET) พบว่า มีค่าสหสัมพันธ์สูง (r มีค่าตั้งแต่ 0.778 ถึง 0.796) และแตกต่างกันอย่างไม่มีนัยสำคัญที่ระดับ .01 ในทุกเงื่อนไขจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนน ซึ่งไม่เป็นไปตามสมมติฐานที่ตั้งไว้ในข้อ 1.3 ที่ว่า “ค่าความเที่ยงตรงตามสภาพของคะแนนที่มีรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน จะมีค่าสูงกว่า รูปแบบการตรวจให้คะแนนรูปแบบอื่น ในทุกจำนวนผู้ตรวจให้คะแนน” และข้อ 1.4 ที่ว่า “ค่าความเที่ยงตรงตามสภาพของคะแนนที่มีจำนวนผู้ตรวจที่มากกว่า จะมีค่าสูงกว่า จำนวนผู้ตรวจที่น้อยกว่า ในทุกรูปแบบการตรวจให้คะแนน” แสดงว่า ไม่ว่าจะใช้รูปแบบการตรวจให้คะแนนต่างกันแต่มีจำนวนผู้ตรวจเท่ากัน หรือใช้จำนวนผู้ตรวจที่ต่างกันแต่มีรูปแบบการตรวจให้คะแนนเหมือนกัน ไม่พบความแตกต่างของค่าความเที่ยงตรงตามสภาพ นั่นคือ จำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนไม่มีผลต่อความเที่ยงตรงตามสภาพ ทั้งนี้อาจเนื่องมาจาก ข้อสอบมีความเที่ยงตรงเชิงเนื้อหา ผู้ตรวจตรวจข้อสอบอย่างสอดคล้องกับผลการตอบของผู้สอบ และผู้สอบทำข้อสอบเต็มที่ตามศักยภาพของตนเอง ประกอบกับเครื่องมือที่ใช้วัดมีคุณภาพ กล่าวคือ แบบทดสอบปลายเปิดวิชาคณิตศาสตร์ที่ผู้วิจัยสร้างนี้ผ่านการพิจารณาจากผู้เชี่ยวชาญ ทำให้มีความเหมาะสมทั้งภาษา ข้อคำถามชัดเจน มีความเป็นปรนัยในข้อคำถาม ผู้สอบจึงมีความเข้าใจจึงสามารถแสดงความรู้ ความคิดตามต้องการได้ ข้อสอบมีความยากง่ายพอเหมาะ ทำให้มีอำนาจจำแนกที่ดี ส่งผลให้เครื่องมือวัดมีความเที่ยงตรงสูงขึ้นด้วย อีกทั้งคำชี้แจงในการตอบคำถามมีความชัดเจน ไม่เปิดโอกาสให้มีการเดาเนื่องด้วยต้องเขียนแสดงวิธีการได้มาซึ่งคำตอบ ตลอดจนการอธิบายแนวคิด แสดงเหตุผลหรือยกตัวอย่างประกอบ เป็นต้น

นอกจากนี้เมื่อพิจารณาถึงด้านกลุ่มตัวอย่างที่ใช้ในการวิจัย จะพบว่า กลุ่มตัวอย่างมีความแตกต่างกัน (Heterogeneous) กล่าวคือ กลุ่มตัวอย่างเป็นผู้สอบที่คละความสามารถ ซึ่งได้มาจากการสุ่ม ทำให้ได้ผู้สอบมาจากหลายลักษณะ มีความเป็นตัวแทนที่ดี ประกอบกับผู้สอบมีความตั้งใจในการตอบข้อสอบ ผู้วิจัยได้เสริมแรงจูงใจโดยการให้รางวัลตอบแทน เช่น ปากกา ยางลบ ดินสอ เพื่อเป็นกำลังใจให้แก่ผู้สอบด้วย อีกทั้งผู้วิจัยเป็นผู้ดำเนินการสอบด้วยตนเองในทุกโรงเรียนที่เป็นกลุ่มตัวอย่าง จึงทำให้สามารถควบคุมและดูแลการดำเนินการสอบได้เป็นอย่างดี อีกทั้งในการให้คะแนนของผู้ตรวจนั้น ผู้วิจัยได้พัฒนาเกณฑ์การให้คะแนน โดยได้ทดลองใช้และผ่านการแก้ไข

ปรับปรุงจากผู้ตรวจที่ได้ทดลองใช้ จนมีความชัดเจน และก่อนจะใช้จริง ผู้วิจัยได้อบรมชี้แจง อธิบายให้ผู้ตรวจทุกคนได้เข้าใจเกณฑ์ในการให้คะแนน เปิดโอกาสให้ซักถาม เมื่อข้อสอบมีความเป็นปรนัยในการให้คะแนน จึงทำให้มีค่าความเที่ยงตรงของการวัดมากขึ้น สามารถวัดได้ตรงกับความเป็นจริงตามความสามารถของผู้สอบนั้น ๆ สอดคล้องกับคำกล่าวของเมอเรนส์และลีแมน (Mehrens; & Lehmann. 1972: 228) ที่ว่าเพื่อให้การตรวจคำถามปลายเปิดมีความเที่ยงตรงมากขึ้น ผู้ตรวจควรใช้วิธีการตรวจที่เหมาะสม ใช้เกณฑ์การตรวจกับนักเรียนทุกคน มีการออกแบบกฎเกณฑ์การให้คะแนน (Scoring rubric) ตลอดจนการฝึกหรืออบรมผู้ตรวจ (Lane; et al. 1996: 72)

3.2 การศึกษาความเที่ยงตรงตามสภาพ พิจารณาจากค่าสัมประสิทธิ์สหสัมพันธ์ของคะแนนความสามารถกับคะแนนที่ได้จากผลการทดสอบทางการศึกษาระดับชาตินัยพื้นฐาน (Ordinary National Education Test: O-NET) พบว่า มีค่าสหสัมพันธ์สูง (r มีค่าตั้งแต่ 0.769 ถึง 0.791) และมีค่าแตกต่างกันอย่างไม่มีนัยสำคัญที่ระดับ .01 ในทุกเงื่อนไขจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนน แสดงว่า จำนวนผู้ตรวจและรูปแบบการตรวจไม่มีผลต่อความเที่ยงตรงตามสภาพ ซึ่งไม่เป็นไปตามสมมติฐานที่ตั้งไว้ในข้อ 2.3 ที่ว่า “ค่าความเที่ยงตรงตามสภาพของคะแนนที่มีรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน จะมีค่าสูงกว่า รูปแบบการตรวจให้คะแนนรูปแบบอื่น ในทุกจำนวนผู้ตรวจให้คะแนน” และข้อ 2.4 ที่ว่า “ค่าความเที่ยงตรงตามสภาพของคะแนนที่มีจำนวนผู้ตรวจที่มากกว่า จะมีค่าสูงกว่า จำนวนผู้ตรวจที่น้อยกว่า ในทุกรูปแบบการตรวจให้คะแนน” ผลการวิจัยนี้สามารถอธิบายได้ด้วยเหตุผลเช่นเดียวกับข้อ 3.1

4. ภาพรวมทั้งสองโมเดล

ในทัศนะของผู้วิจัย แม้ว่าการวิจัยในครั้งนี้ผู้วิจัยจะไม่ได้มุ่งเปรียบเทียบคุณลักษณะของคะแนนจากการวิเคราะห์ด้วยโมเดลการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราล์ซ แต่การใช้โมเดลทั้งสองวิเคราะห์คุณลักษณะคะแนนนั้นทำให้สามารถค้นหาแหล่งความคลาดเคลื่อนของการวัดได้หลายแหล่งได้พร้อมกัน ๆ กัน อีกทั้งสามารถประมาณค่าน้ำหนักหรือขนาดของแต่ละแหล่งได้ แต่ข้อมูลที่ได้มีความละเอียดที่แตกต่างกัน กล่าวคือ ในโมเดลการสรุปอ้างอิงจะให้สารสนเทศเป็นภาพรวมหรือรายกลุ่ม (Group level) เช่น ผู้สอบ ข้อสอบ ผู้ตรวจในภาพรวม ในส่วนของโมเดลหลายองค์ประกอบของราล์ซ จะให้รายละเอียดเป็นรายบุคคลหรือรายสมาชิก (Individual level) เช่น ผู้สอบแต่ละคน ข้อสอบแต่ละข้อ ผู้ตรวจแต่ละคน ซึ่งการวิจัยในครั้งนี้ผลการวิเคราะห์คุณลักษณะของคะแนน พบว่า ทั้งสองโมเดลให้ผลการวิเคราะห์ที่มีทิศทางหรือแนวโน้มที่สอดคล้องกัน ทั้งในประเด็นของความแปรปรวนของแต่ละองค์ประกอบที่พบว่า ความแปรปรวนของผู้สอบ ข้อสอบ มีค่าสูงกว่าความแปรปรวนของผู้ตรวจ และความแปรปรวนของผู้ตรวจมีค่าค่อนข้างน้อยในทุกเงื่อนไขจำนวน

ผู้ตรวจและรูปแบบการตรวจ นั่นคือ ผู้ตรวจมีความแตกต่างกันในการให้คะแนนน้อย และเมื่อพิจารณาจากค่าดัชนีความสอดคล้องของผู้ตรวจให้คะแนน (Rater Agreement Index: RAI) (Burry Stock, 1996) ของผู้ตรวจเมื่อตรวจข้อสอบทุกข้อของผู้สอบทุกคนและจำนวนผู้ตรวจ 2 คน 3 คน และ 4 คน มีค่าเป็น 0.8782, 0.8785 และ 0.8826 ตามลำดับ ซึ่งมีค่าสูง แสดงว่าผู้ตรวจมีความสอดคล้องกันในการให้คะแนนสูง แม้ว่าในโมเดลหลายองค์ประกอบของราล์ซ จะพบว่า ความเชื่อมั่นแยกส่วนของผู้ตรวจมีค่าสูง ซึ่งแสดงถึงความแตกต่างของความเข้มงวด/ใจดี ก็ตามแต่ยังอยู่ในช่วงที่ยอมรับได้ และเมื่อพิจารณาในประเด็นของค่าความเชื่อมั่นของคะแนนที่ได้ คือ ค่าสัมประสิทธิ์การสรุปอ้างอิงและค่าความค่าเชื่อมั่นแยกของผู้สอบ ในแต่ละเงื่อนไขจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนน แม้จะมีค่าแตกต่างกัน แต่ทั้งสองโมเดลต่างให้ค่าความเชื่อมั่นสูงคือเกินกว่า 0.86 ซึ่งอยู่ในเกณฑ์ที่ยอมรับได้ดังที่นักวิจัยได้เสนอว่าแบบทดสอบควรมีค่าความเชื่อมั่นไม่ต่ำกว่า 0.70 (สุวิมล ติรภานันท์, 2550: 182; อ้างอิงจาก Jacobs, 1991; Frankel; & Wallen, 1999) และในประเด็นของความเที่ยงตรงตามสภาพของคะแนนซึ่งพิจารณาจากค่าสัมประสิทธิ์สหสัมพันธ์ของคะแนนดิบ คะแนนความสามารถกับคะแนนสอบ O-NET จะเห็นว่ามีค่าสหสัมพันธ์สูง (r มีค่าตั้งแต่ 0.769 ถึง 0.796) และมีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติที่ระดับ .01 ซึ่งแสดงถึง คะแนนที่ได้จากทุกเงื่อนไขจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนน มีความเที่ยงตรงตามสภาพไม่แตกต่างกัน นอกจากนี้เมื่อหาค่าสหสัมพันธ์แบบเพียร์สันของคะแนนในทุกเงื่อนไข พบว่า คะแนนสอบกับคะแนนสอบมีค่า 0.977-0.999 คะแนนสอบกับคะแนนความสามารถมีค่า 0.947-0.987 คะแนนความสามารถกับคะแนนความสามารถมีค่า 0.958-0.998 จะเห็นว่าคะแนนสอบและคะแนนความสามารถที่ได้ทั้งจากทุกเงื่อนไขมีค่าสหสัมพันธ์กันสูง

จะเห็นว่า ผลการวิจัยที่ได้จากทั้งสองโมเดลให้สารสนเทศที่เติมเต็ม (Complementary) ซึ่งกันและกัน ทำให้เกิดความเพียงพอในการตัดสินใจเลือกรูปแบบการตรวจให้คะแนนและจำนวนผู้ตรวจที่เหมาะสม เพื่อทำให้การทดสอบมีประสิทธิภาพสูงสุด (Cost-effective) (Iramaneerat; et al. 2007) อย่างไรก็ตามการพิจารณาจำนวนผู้ตรวจที่เหมาะสม จะต้องพิจารณาว่าเป็นไปได้หรือความสามารถที่จะปฏิบัติได้จริง มีความสะดวกและประหยัด นอกจากนี้การใช้โมเดลหลายองค์ประกอบของราล์ซ ทำให้สามารถคำนวณค่าความเชื่อมั่นแยกส่วนในแต่ละแหล่งความแปรปรวนได้ในขั้นตอนเดียว ทำให้สามารถให้สารสนเทศที่เป็นประโยชน์ต่อไปได้ เช่น เมื่อพบความไม่สอดคล้องกับโมเดล (Misfit) ในแต่ละแหล่งของความแปรปรวน สามารถให้ผลย้อนกลับ (Feedback) แก่แหล่งนั้นได้ เช่น หากพบว่าผู้ตรวจคนใดไม่สอดคล้องกับโมเดลหรือผู้ตรวจมีความแตกต่างกันของความเข้มงวด/ใจดี สามารถให้ข้อมูลย้อนกลับแก่ผู้ตรวจคนดังกล่าว เพื่อการปรับปรุง แก้ไขต่อไป ดังนั้นการวิเคราะห์ข้อมูลชุดเดียวกัน โดยใช้สองโมเดล จึงทำให้ข้อมูลหรือผลการวิจัยมีความน่าเชื่อถือมากขึ้น (Magno, 2009:

online; Guler; & Gelbal. 2010) รวมถึงพิจารณาข้อสอบที่ไม่สอดคล้องกับโมเดล เช่น ให้คะแนนสูงเกินค่าคาดหวัง เพื่อการปรับปรุงข้อสอบต่อไป

ในการวางแผนการตรวจจึงไม่จำเป็นที่ผู้ตรวจต้องตรวจให้คะแนนผู้สอบทุกคนทุกข้อ อาจให้ผู้ตรวจตรวจข้อสอบของผู้สอบบางคน หรือ ตรวจข้อสอบในบางคุณลักษณะหรือข้อสอบบางข้อ การวางแผนการตรวจอาจทำได้ทั้งแบบบางส่วน (Partially) หรือแบบไขว้สมบูรณ์ (Full crossed) (Kim. 2000: online; Myford; & Wolfe. 2003: 392; Iramaneerat; et al. 2007: online) โดยที่ยังคงมีความน่าเชื่อถือในทางสถิติ (Engelhard. 1992: 188; Linacre.1994; Linacre; et al. 1994; Brennan. 2001) นั่นคือ การออกแบบรูปแบบการตรวจให้คะแนนที่เหมาะสมหรือหาวิธีการตรวจเพื่อลดภาระลดเวลาในการตรวจ เสียค่าใช้จ่ายน้อยที่สุด พยายามให้ผู้ตรวจหรือเงื่อนไขการวัดให้น้อยที่สุด โดยที่มีความเป็นไปได้ในทางปฏิบัติ (Practical) แต่ยังคงไว้ซึ่งความน่าเชื่อถือ (Reliable) และยังมีอำนาจ (Power) ของคะแนน ตลอดจนให้ค่าทางสถิติที่ยอมรับได้ (Statistically well-founded) (Kim. 2000: online)

สรุปโดยภาพรวม ได้ว่า เมื่อใช้รูปแบบการตรวจให้คะแนนทั้งสามรูปแบบ คือ ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน ในจำนวนผู้ตรวจสามลักษณะ คือ 4 คน 3 คน และ 2 คน พบว่า ผู้ตรวจมีความแตกต่างกันในการตรวจให้คะแนนแต่อยู่ในช่วงที่ยอมรับได้ ค่าความเชื่อมั่นซึ่งพิจารณาจากค่าสัมประสิทธิ์การสรุปอ้างอิงและค่าความเชื่อมั่นแยกส่วน มีค่าแตกต่างกันในบางเงื่อนไข แต่ค่าความเชื่อมั่นที่ได้ในทุกเงื่อนไข มีค่าสูงเกิน 0.86 และค่าความเที่ยงตรงตามสภาพของคะแนนสอบ คะแนนความสามารถมีค่าสูงและไม่แตกต่างกันในทุกเงื่อนไข จึงทำให้สามารถเลือกใช้รูปแบบการตรวจให้คะแนนหรือจำนวนผู้ตรวจในการตรวจแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ที่ประหยัด เหมาะสมกับบริบทจริง สะดวกในการปฏิบัติโดยที่คะแนนที่ได้มีความเที่ยงตรงตามสภาพและมีค่าความเชื่อมั่นสูง

ข้อเสนอแนะ

จากผลการศึกษาคุณลักษณะคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนแตกต่างกัน โดยใช้โมเดลการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราล์ซ ผู้วิจัยมีข้อเสนอแนะในการนำผลการวิจัยไปใช้และการทำวิจัยครั้งต่อไป ดังนี้

1. ข้อเสนอแนะในการนำผลการวิจัยไปใช้

ผลการวิจัยครั้งนี้ชี้ให้เห็นว่า ทั้งสองโมเดลต่างให้ผลการศึกษาไปในทิศทางเดียวกัน กล่าวคือ มีความแปรปรวนของผู้ตรวจค่อนข้างน้อยในทุกเงื่อนไขจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนน แสดงว่า ผู้ตรวจมีความแตกต่างกันน้อยในการให้คะแนน แม้ว่าจะมีความเข้มงวดหรือใจดีแตกต่างกันก็ตาม ค่าสัมประสิทธิ์การสรุปอ้างอิงและความเชื่อมั่นแยกส่วนของผู้สอบมีค่าสูงและมีค่าใกล้เคียงกัน แม้ว่าจะแตกต่างกันไปตามจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนที่ต่างกัน เมื่อพิจารณาถึงความเที่ยงตรงตามสภาพของคะแนน จะเห็นว่าในทุกเงื่อนไขมีค่าสูงและแตกต่างกันอย่างไม่มีนัยสำคัญ ดังนั้นผู้วิจัยขอเสนอแนะหากจะนำผลการวิจัยไปใช้ ดังนี้

1. ผู้ใช้สามารถเลือกรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน หรือผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน ไม่จำเป็นต้องให้ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน และเลือกใช้จำนวนผู้ตรวจที่เหมาะสมซึ่งผู้ตรวจ 2-3 คน น่าจะเหมาะสม โดยพิจารณาถึงความเป็นไปได้หรือความสามารถที่จะปฏิบัติได้จริง มีความสะดวกและประหยัด

2. ผู้ใช้ควรสร้างกฎเกณฑ์การให้คะแนนให้ชัดเจน และควรเป็นเกณฑ์เฉพาะ (Specific rubric) ของแต่ละชิ้นงานหรือแต่ละข้อ ควรให้ข้อมูล ความรู้และอบรมผู้ตรวจให้เข้าใจเกณฑ์การตรวจให้คะแนนและแม่นยำในการใช้เกณฑ์ โดยอาจมีการยกตัวอย่างการตอบของผู้สอบในแต่ละระดับคะแนน และมีการทดลองใช้ ตลอดจนมีการตรวจสอบผลการตรวจย้อนหลัง (Recheck) เพื่อดูความสอดคล้องกันของผู้ตรวจ อย่างไรก็ตามหากนำไปใช้แล้วยังพบว่าผู้ตรวจบางคนมีความเข้มงวดหรือใจดีมากเกินไปหรือให้คะแนนไม่สอดคล้องกับผู้อื่น สามารถให้ข้อมูลย้อนกลับแก่ผู้ตรวจต่อไป และในการคัดเลือกผู้ตรวจนั้นควรเลือกผู้ที่มีความชำนาญหรือเชี่ยวชาญในสิ่งที่ประเมินหรือตรวจให้คะแนน

3. ควรมีการฝึกฝนให้นักเรียนมีประสบการณ์ในการเขียนตอบคำถามปลายเปิด หรืออธิบายความคิด ความรู้ของตนเองในห้องเรียนอย่างเพียงพอ ทั้งนี้หากนำแบบทดสอบปลายเปิดมาใช้ทดสอบวัดความรู้ แต่ขาดการฝึกปฏิบัติในห้องเรียน อาจทำให้การวัดประเมินขาดประสิทธิภาพได้

2. ข้อเสนอแนะสำหรับการทำวิจัยครั้งต่อไป

ผลการวิจัยครั้งนี้ชี้ให้เห็นคุณลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ ในสามประเด็นใหญ่ ๆ คือ ความแปรปรวนขององค์ประกอบ ความเชื่อมั่น และความเที่ยงตรงตามสภาพ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน ผู้วิจัยจึงขอเสนอแนะสำหรับผู้สนใจที่จะศึกษาเกี่ยวกับคุณลักษณะคะแนนจากแบบทดสอบปลายเปิด ดังต่อไปนี้

1. จากผลการศึกษาที่พบว่า ความแปรปรวนของความคลาดเคลื่อนเชิงสุ่มที่ไม่อาจระบุได้ ในรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน และผู้ตรวจ

ตรวจข้อสอบบางข้อของผู้สอบทุกคน มีค่าค่อนข้างสูง จึงควรมีการศึกษาคุณลักษณะของคะแนนจากแบบทดสอบปลายเปิด โดยการเพิ่มเงื่อนไขในการศึกษาเพื่อควบคุมแหล่งที่คิดว่าเป็นความคลาดเคลื่อนไว้ล่วงหน้า เช่น จำนวนครั้งในการตรวจให้คะแนน จำนวนสเกล(ระดับคะแนน)ในกฎเกณฑ์การให้คะแนน เป็นต้น

2. จากผลการศึกษาที่พบว่า ผู้ตรวจมีความแตกต่างกันในการให้คะแนนไม่มาก แต่มีความเข้มงวดและใจดีที่แตกต่างกัน เพื่อให้เกิดความเข้าใจในความคลาดเคลื่อนที่เกิดจากผู้ตรวจมากขึ้น จึงควรมีการศึกษาคความคลาดเคลื่อนที่เกิดจากผู้ตรวจในประเด็นอื่น ๆ นอกเหนือจากความเข้มงวด/ใจดี เช่น การจำกัดช่วง (Restriction range) การให้คะแนนตรงกลาง (Central tendency) ผลจากฮาโล (Halo effect) ความลำเอียงในการตรวจ (Bias rating) หรือวิธีการตรวจให้คะแนน

3. จากผลการศึกษาที่พบว่า การวิเคราะห์คะแนนโดยใช้ทั้งสองโมเดลคือ โมเดลการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของวาร์ธ ทำให้ได้สารสนเทศที่เติมเต็มกัน ทำให้ได้ข้อมูลที่สมบูรณ์และมีความน่าเชื่อถือมากขึ้น ดังนั้น จึงควรมีการประยุกต์ใช้ทั้งสองโมเดล ในการวิเคราะห์คะแนนที่ได้จากผู้ประเมินในการประเมินรูปแบบอื่น ๆ





บรรณานุกรม

- จิตติมา ชอบเอียด. (2551). การใช้ปัญหาปลายเปิดเพื่อส่งเสริมทักษะการให้เหตุผลและทักษะการสื่อสารทางคณิตศาสตร์ของนักเรียนชั้นมัธยมศึกษาปีที่ 1. วิทยานิพนธ์ กศ.ม. (การมัธยมศึกษา). กรุงเทพฯ: บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ. ถ่ายเอกสาร.
- แดง กลางท่าได้. (2531). การประยุกต์ทฤษฎีการสรุปอ้างอิงในการหาความเชื่อมั่นของการประเมินความตรงเชิงเนื้อหา. วิทยานิพนธ์ ค.ม. (การวัดและประเมินผลการศึกษา). กรุงเทพฯ: บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย. ถ่ายเอกสาร.
- ชาญวิทย์ จรัสสุทธิศิธร. (2550). การแสดงหลักฐานความเที่ยงตรงเชิงโครงสร้างของคะแนนการประเมินทักษะกระบวนการทางวิทยาศาสตร์โดยใช้ทฤษฎีมาตรฐานเดิมและทฤษฎีการตอบข้อคำถามตามโมเดลของราล์ฟแบบหลายองค์ประกอบ. วิทยานิพนธ์ กศ.ด. (การทดสอบและวัดผลการศึกษา). กรุงเทพฯ: บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ. ถ่ายเอกสาร.
- ชัยฤทธิ์ ศิลาเดช. (2540). การพัฒนาแฟ้มสะสมงานในการประเมินผลการเรียนวิชาภาษาอังกฤษ ระดับมัธยมศึกษาปีที่ 3. วิทยานิพนธ์ กศ.ด. (การทดสอบและวัดผลการศึกษา). กรุงเทพฯ: บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ. ถ่ายเอกสาร.
- บุญเชิด ภิญโญนนตพงษ์. (2547). การวัดประเมินการเรียนรู้(การวัดประเมินแนวใหม่). เอกสารประกอบคำสอน. คณะศึกษาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ. อัดสำเนา.
- ปวีณา ปิอาทิตย์. (2545). การศึกษาจำนวนผู้ประเมินและจำนวนงานเขียนที่เหมาะสม เมื่อใช้เกณฑ์การให้คะแนนที่ต่างกัน. วิทยานิพนธ์ ค.ม. (การวัดและประเมินผลการศึกษา). กรุงเทพฯ: บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย. ถ่ายเอกสาร.
- ปรีชา เนาว์เย็นผล. (2544). กิจกรรมการเรียนการสอนคณิตศาสตร์โดยใช้การแก้ปัญหาปลายเปิดสำหรับนักเรียนชั้นมัธยมศึกษาปีที่ 1. วิทยานิพนธ์ กศ.ด. (คณิตศาสตร์ศึกษา). กรุงเทพฯ: บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ. ถ่ายเอกสาร.
- ผจงจิต อินทสุวรรณ. (2525). Latent Trait Theory. วารสารวัดผลการศึกษา. 3(3): 51-69.
- (2545). ทฤษฎีการตอบสนองข้อคำถาม. กรุงเทพฯ: สถาบันวิจัยพฤติกรรมศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ.

พรรณี เจียมสุบุตร. (2543). การเปรียบเทียบความเชื่อมั่นของแบบทดสอบวัดความสามารถในการแก้โจทย์ปัญหาทางคณิตศาสตร์ ที่มีจำนวนผู้ตรวจและวิธีการตรวจต่างกัน.

วิทยานิพนธ์ กศ.ม. (การวัดผลการศึกษา). กรุงเทพฯ: บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ. ถ่ายเอกสาร.

ไพรัตน์ วงษ์นาม. (2533). สัมประสิทธิ์การสรุปอ้างอิงของแบบทดสอบแบบความเรียง.

วิทยานิพนธ์ ค.ด. (การวิจัยทางการศึกษา). กรุงเทพฯ: บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย. ถ่ายเอกสาร.

ไมตรี อินทร์ประสิทธิ์. (2547, มกราคม-มิถุนายน). การสอนโดยใช้วิธีการแบบเปิดในชั้นเรียนญี่ปุ่น.

KKU Journal of Mathematics Education 1. (1): 1-9.

เยาวดี ราชชัยกุล วิบูลย์ศรี. (2553). การสร้างวัดผลและการสร้างแบบสอบผลสัมฤทธิ์.

พิมพ์ครั้งที่ 9. กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย

ล้วน สายยศ และอังคณา สายยศ. (2539). เทคนิคการวัดผลการเรียนรู้. กรุงเทพฯ: สุวีริยาสาส์น

ศิริชัย กาญจนวาสี. (2550). ทฤษฎีการสอบแนวใหม่. กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

สถาบันทดสอบทางการศึกษาแห่งชาติ. (2552). ค่าสถิติพื้นฐานผลการทดสอบทางการศึกษา

ระดับชาติขั้นพื้นฐาน (O-NET) ปีการศึกษา 2551. สืบค้นเมื่อ 18 กันยายน 2552, จาก www.niets.or.th

----- (2553). คู่มือนักเรียนในการสอบ O-NET ป.6, ม.3. สืบค้นเมื่อ 20 พฤษภาคม 2554, จาก <http://www.niets.or.th/upload-files/uploadfile/7/5573f00263d32e2e7a2839e5f9fdd56.pdf>

----- (2554). ความเป็นมา. สืบค้นเมื่อ 20 พฤษภาคม 2554, จาก www.niets.or.th

----- (n.d). ระบบการพัฒนาข้อสอบของ สทศ. สืบค้นเมื่อ 20 พฤษภาคม 2554, จาก <http://www.niets.or.th/upload-files/uploadfile/5/b6e3c25bf733dcc3e0180aff4c8922cd.pdf>

สุภมาส อังคุโชติ. (2543). การปรับคะแนนเฉลี่ยสะสมระดับมัธยมศึกษาตอนปลาย: การ

เปรียบเทียบความตรงเชิงทำนาย. ปรินญานินพนธ์ ค.ด. (การวัดและประเมินผลการศึกษา). กรุงเทพฯ: บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย. ถ่ายเอกสาร.

สุวิมล ติรกานนท์. (2546). การใช้สถิติในการวิจัยทางสังคมศาสตร์: แนวทางสู่การปฏิบัติ.

กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

สุวิมล ติรกานนท์. (2550). การสร้างเครื่องมือวัดตัวแปรในการวิจัยทางสังคมศาสตร์: แนวทางสู่การปฏิบัติ. กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

สำนักงานรับรองมาตรฐานและการประเมินคุณภาพการศึกษา (องค์การมหาชน). (2547).

พระราชบัญญัติการศึกษาแห่งชาติ พ.ศ. 2542 แก้ไขเพิ่มเติม (ฉบับที่ 2) พ.ศ. 2545.

กรุงเทพฯ: พรินทวนกราฟฟิค.

สัมพันธ์ พันธุ์พฤกษ์. (2553). Powerpoint การวัดผลสัมฤทธิ์ทางการศึกษาของนักเรียน: การทดสอบทางการศึกษาระดับชาติด้านพื้นฐาน(O-NET) ชั้น ป.6 และ ม. 3 ปีการศึกษา 2553. สืบค้นเมื่อ 20 พฤษภาคม 2554, จาก <http://www.niets.or.th/upload-files/uploadfile/7/bd3b809af9889c82cc8fe5814e44aa05.pdf>

อนันต์ ศรีโสภา. (2525). **การวัดผลการศึกษา**. พิมพ์ครั้งที่ 3. กรุงเทพฯ: ไทยวัฒนาพานิช

อุษณีย์ บัวศิริพันธ์. (2543). **การเปรียบเทียบค่าสัมประสิทธิ์สรุปอ้างอิงของแบบทดสอบวิชาคณิตศาสตร์ที่มีวิธีการตรวจ จำนวนผู้ตรวจและประสบการณ์ของผู้ตรวจ แตกต่างกัน**. วิทยานิพนธ์ กศ.ม. (การวัดผลการศึกษา). กรุงเทพฯ: บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ. ถ่ายเอกสาร.

Andrich, David. (1982). An Index of Person Separation in Latent Trait Theory, the Traditional KR-20 Index, and the Guttman Scale Response Pattern. **Education Research and Perspectives**. 9(1): 95-104. Retrieved March 18,2009, from <http://www.rasch.org/erpERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED368154>

Arter, Judy. (1999). Teaching About Performance Assessment. **Educational Measurement; Issues and Practice**. 18(2): 30-44.

Bachman, Lyle F.; Lynch, Brian K.; & Mason, Maureen. (1993). **Investigating Variability in Tasks and Rater Judgments in a Performance Test of Foreign Language Speaking**. Paper presented at the Annual Language Testing Research Colloquium. Cambridge, England. Retrieved January 18, 2008, from <http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED368154>

Basturk, R. (2008, August). Applying the Many-Facet Rasch Model to Evaluate PowerPoint Presentation Performance in Higher Education. **Assessment & Evaluation in Higher Education**. 33(4): 431-444.

Baxter; et al. (1992, Spring). Evaluation of Procedure-Based Scoring for Hand-On Science Assessment. **Journal of Educational Measurement**. 29(1): 1-17.

Becker, Jerry P.; & Shimada, Shigeru. (1997). **The Open-Ended Approach: A new Proposal For Teaching Mathematics**. Reston, Virginia: National Council of Teachers of Mathematics.

- Bloom, Benjamin S. (1956). **Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain**. New York: David McKay.
- Bonk, William J.; & Ockey, Gary J. (2003). A Many-Facet Rasch Analysis of the Second Language Group Oral Discussion Task. **Language Testing**. 20(1): 89-110.
Retrieved August 5, 2007, from <http://ltj.sagepub.com/cgi/content/abstract/20/1/89>
- Brennan, Robert L. (1983). **Element of Generalizability Theory**. Iowa: The American College Testing Program.
- (1997. Winter). A Perspective on the History of Generalizability Theory. **Educational Measurement: Issues and Practice**. 16(4): 14-20.
- (2001). **Generalizability Theory**. Iowa: Springer-Verlag New York.
- Brennan, Robert L.; Gao, Xiaohong; & Colton, Dean A. (1995, April). Generalizability Analysis Work Key Listening and Writing Testing. **Educational and Psychological Measurement**. 55(2): 157-176.
- Buray-Stock, Judith A.; et al. (1996). Rater Agreement Indexes for Performance Assessment. **Educational and Psychological Measurement**. 52(6): 251-262.
- Cai, J.; Lane, S.; & Jackabcsin, M.S. (1996). The Role of Open-Ended Tasks and Holistic Scoring Rubrics : Assessing Student's Mathematical Reasoning and Communication. In *Communication in Mathematics, K-12, 1996 Yearbook* (pp. 137-145), Reston, Virginia: National Council of Teachers of Mathematics
- Chicago Board of Education [CBE]. (2000). **Introduction to Score Rubrics** . Retrieved June 3, 2006, from http://intranet.cps.k12.il.us/Assessment/Ideas_and_Rubrics/Intro_Scoring/Intro_scoring.html
- Chi, Eunlim. (2001). Comparison Holistic and Analytic Scoring for Performance Assessment with Many-Facet Rasch Model. **Journal of Applied Measurement**. 2(4): 379-388.
- Chui, Christopher W.T.; & Wolfe, Edward W. (2002, September). A Method for Analyzing Sparse Data Matrices in the Generalizability Theory Frame work. **Applied Psychological Measurement**. 26(3): 321-338.
- Clauser, Brian; & Linacre, John M. (1999). Relating Cronbach and Rasch Reliabilities. **Rasch Measurement Transactions**. 13(2): 696. Retrieved August 20, 2007 from <http://www.rasch.org/rmt/rmt132i.htm>

- Coffman, William E. (1971). Essay Examinations. In **Education Measurement**. Thorndike, R. L. editor. pp. 271-280. 2nd ed. Washington. D.C.: American Council on Education.
- Conway, Kathleen D. (1999). **Assessing Open-Ended Problems**. Retrieved March 15, 2008, from <http://wilsontxt.hwwilson.com/pdfhtml/05891/2QVB6/USJ.html>.
- Cooney, Thomas J.; et al. (n.d.). **Opend-Ended Assessment in Math**. Retrieved June 6, 2006, from <http://www.heinemann.com/math/reasons.cfm>
- Crick, J. E.; & Brennan, Robert L. (1983). **Mannual for GENOVA: A Generalized Analysis of Variance System** (ACT Technical Bulletin no. 43). Iowa. IA: American College Testing Program.
- Cronbach, Lee J.; et al. (1972). **The Dependability of Behavioral Measurement: Theory of Generalizability of Score and Profiles**. New York : John wiley & Son.
- Cunningham, George K. (1986). **Educational and Psychological Measurement** . New York. Macmillan.
- Dunbar, S.B.; Koretz, D.M.; & Hoover, H.D. (1991). Quality Control in the Development and Use of Performance Assessment. **Applied Measurement in Education**. 4:289-304.
- Engelhard, George, Jr. (1992). The Measurement of Writing Ability With a Many-Facet Rasch Model. **Applied Measurement in Education**. 5(3): 171-191.
- (1994, Summer). Examining Rater Errors in the Assessment of Written Composition With a Many-Faceted Rasch Model. **Journal of Education Measurement**. 31(2): 93 -112.
- (1996, Spring). Evaluating Rater Accuracy in Performance Assessment. **Journal of Education Measurement**. 33(1): 56-70.
- Ebel, Robert L. (1951). Estimation of Reliability of Rating. **Psychometrika**. 16: 407-424.
- Glatzer, David J.; & Glatzer, J. (2005). **PREPARING FOR THE Pennsylvania PSSA Mathematics GRADE 11**. Amsco School. United States of America.
- Guler, N. & Gelbal S. (2010). A Study Based on Classic Test Theory and Many Facet Rasch Model. **Egitim Arastirmalari-Eurasian Journal of Educational Research**. 38: 108-125.

- Fisher, R. A. (1936). **The Use of Multiple Measurement in Taxonomic Problems**. *Annals of Eugenics*. 7: 179-188. Retrieved August 25, 2008, from <http://www.library.adelaide.edu.au/digitised/Fisher/138.pdf>
- Fisher, W. Jr. (1992). Reliability Statistic. **Rasch Measurement Transaction**. 6(3): 238. Retrieved August 20, 2007, from <http://www.rasch.org/rmt63i.htm>
- Foong, Pui Y. (n.d). **Using Short Open-ended Mathematics Questions to Promote Thinking and Understanding**. National Institute of Education, Singapore. Retrieved March 15, 2008, from <http://math.unipa.it/~grim/SiFoong.PDF>
- Gronlund, Norman E. (1976). **Measurement and Evaluation in Testing**. 3rd ed. New York. Macmillan.
- Gulliksen, Horold. (1950). **Theory of Mental Tests**. New York: John Wiley&Son
- Haiyang, S. (2010, April). An Application of Classical Test Theory and Many-facet Rasch Measurement in Analyzing the Reliability of an English Test for Non-English Major Graduates. **Chinese Journal of Applied Linguistics(Bimonthly)**. 33(2): 87-102.
- Hambleton, Ronald K.; Swaminathan, H.; & Roger, H. Jane. (1991). **Fundamentals of Item Response Theory**. Newbury Park: Sage Publication.
- Hancock, Lynn C. (1995). Enhancing Mathematics Learning with Open-Ended Questions. **The Mathematics Teacher**. 88(6): 496-499.
- Hopkins, Charles D.; & Antes, Richard L. (1990). **Classroom Measurement and Evaluation**. 3rd ed. Itasca, IL.
- Hurman, J. (1997). **Assessing New Assessment; How Do They Measure Up?**. Los Angeles: Graduate School of Education and Information Studies, University of California.
- Ibrahim, A. M. (1984). The Dependability of Need Assessment Data: An Application of Generalizability Theory to the Rating of Educational Goals. **Dissertation Abstracts International**. 45: 499A.
- Iramaneerat, Cherdasak; et al. (2007). **Quality Control of an OSCE using Generalizability Theory and Many-Faceted Rasch Measurement**. *Adv Health Sci Educ Theory Pract*. Retrieved August 20, 2007, from <http://www.springerlink.com/content/b5627765j1441q3v/fulltext.pdf>

- Kim, Sungsook C. (2000). **Investigating the Generalizability of Scores from Different Rating Systems in Performance Assessment**. Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000). Retrieved July 15, 2008, from http://eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/16/cc/1f.pdf
- Klein, Stephen P.; et al. (1998). Analytic Versus Holistic Scoring of Science Performance Tasks. **Applied Measurement in Education**. 11(2): 121-137.
- Kristie, W.; Andrea, K.; & Gina, K. (1998). **Alternative Approaches to Scoring the Effects of Using Different Scoring Methods on the Validity of Score from a Performance Assessment**. Retrieved June 15, 2007, from <http://eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED427080>
- Kubiszyn, Tom; & Borich, Gary. (1996). **Educational Testing and Measurement: Classroom Application and Practice**. 5th ed. Haper Collins College Publishers.
- Lane, Suzanne; et al. (1996). Generalizability and Validity of Mathematics Performance Assessment. **Journal of Education Measurement**. 33(1): 71-92.
- Lane, Suzanne; & Stone, Clement A. (2006). Performance Assessment. In **Educational Measurement**. Brennan, Robert L., editor. pp. 387-431. 4th ed. New York: American Council on Education and Praeger Publishers.
- Lee, Yong W.; & Kantor, Robert. (2005). **Dependability of Scores for a New ESL Speaking Test: Evaluating Prototype Tasks**. Retrieved August 2, 2008, from <http://www.ets.org/Media/Research/pdf/RM-04-07.pdf>
- Linacre, John M. (1993, April). **Generalizability Theory and Many-Facet Rasch Measurement**. Paper presented at the 1993 Annual Meeting of the American Education Research Association Atlanta, Georgia, April 13, 1993.
- (1994). **Many-Facet Rasch Measurement**. Chicago: MESA Press.
- (2009). **A User's Guide to FACETS [Computer Program Manual]**. Chicago: MESA Press.
- Linacre, John M.; et al. (1994). Measurement with Judges: Many-Faceted Conjoint Measurement. **International Journal of Educational Research**. 21(6). 569-577.

- Linacre, John M.; & Wright, Benjamin D. (2002). Construction of Measures from Many-facet Data. **Journal of Applied Measurement**. 3(4): 486-512.
- Loe, Q. J.; & Chen, C. X. (2004). **The Open-ended Approach in Reforming Traditional Teaching**. Paper Presented at the 10 th International Congress on Mathematics Education. Retrieved March 15, 2008, from www.icmeorganisers.dk/tsg14/#paper.
- Lord, Frederic M., & Novick, Melvin R. (1968). **Statistical Theories of Mental Test Scores**. Reading, Massachusetts: Addison-Wesley.
- Lunz, Mary E.; Wright, Benjamin D.; & Linacre, John M. (1990). Measuring the Impact of Judge Severity on Examination Score. **Applied Measurement in Education**. 3(4): 331-345. Retrieved May 14, 2008, from www.Rasch.org/memo47.htm
- Lunz, Mary E.; Wright, Benjamin D. (n.d). **Latent Trait Model for Performance Examinations**. Retrieved November 10, 2008, from <http://www.ipn.uni-kiel.de/aktuell/buecher/rostbuch/c06.pdf>
- Lynch, Brian K.; & McNamara, Tim F. (1998). **Using G-theory and Many-facet Rasch measurement in the development of performance assessment of the ESL speaking skills of immigrant**. Retrieved August 20, 2007, from <http://ltj.sagepub.com/cgi/content/abstract/15/2/158>
- Magno, Carlo. (2009, April). Demonstrating the Difference between Classical Test Theory and Item Response Theory Using Derived Test Data. **The International Journal of Educational and Psychological Assessment**. 1(1): 1-11. Retrieved september 20, 2010, from <http://tijepa.books.officelive.com/Documents/article1v1.pdf>
- McManus, IC.; Thompson, M.; & Mollon, J. (2006) **Assessment of Examiner Leniency and Stringency in the MRCP(UK) Clinical Examination(PACES) Using Multi-Facet Rasch Modelling**. Retrieved March 20, 2009, from <http://creativecommons.org/licenses/by/2.0>
- Magone, Maria E.; et al. (1994). Validating The Cognitive Complexity and Content Quality of Mathematics Performance Assessment. **International Journal of Education Research**. 21(3): 317-340.
- Mehrens, William A.; & Lehmann, Irvin J. (1973). **Measurement and Evaluation in Education and Psychology**. New York: Holt, Rinehart and Winston

- Meng, Xiao Li; Rosenthal, Robert; & Rubin, Donald B. (1992). Comparing Correlated Correlation Coefficients. **Psychological Bulletin**. 111(1): 172-175.
- Mertler, Craig A. (2001). Designing scoring rubrics for your classroom. **Practical Assessment, Research & Evaluation**, 7(25). Retrieved July 29, 2008 from <http://PAREonline.net/getvn.asp?v=7&n=25>
- Messick, Samuel. (1989). Validity. In **Educational Measurement**. Linn, Robert L., editor. 3rd ed. New York: American Council on Education and Praeger Publishers.
- Moskal, Barbara M. (2000). Scoring rubrics: what, when and how?. **Practical Assessment, Research & Evaluation**. Retrieved July 29, 2008, from <http://PAREonline.net/getvn.asp?v=7&n=3>
- Myford, Carol M.; & Wolfe, Edward W. (2000). **Monitoring Sources of Variability Within the Test of Spoken English Assessment System**. Education Testing Service, Princeton, New Jersey. Retrieved June 21, 2008, from www.ets.org/Media/Research/pdf/RR-00-06-Myford.pdf
- Myford, Carol M.; & Wolfe, Edward W. (2003). Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part I. **Journal of Applied Measurement**. 4(4): 386-422.
- (2004). Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part II. **Journal of Applied Measurement**. 5(2): 189-277.
- Nakamura, Yuji. (2002). An Application of Many – Faceted Rasch Model to Writing Test Analysis. **Curriculum Innovation Testing and Evaluation : Proceedings of the 1st Annual JALT Pan-SIG Conference**. May 11-12, 2002. Kyoto University: Kyoto Institute of Technology. Retrieved June 15, 2007, from <http://jalt.org/pansig/2002/HTML/Nakamura.htm>
- (2004). A Comparison of Holistic and Analytic Scoring Methods in the Assessment of Writing. **The Interface Between Interlanguage, Pragmatics and Assessment : Proceedings of the 3rd Annual JALT Pan-SIG Conference**. May 22-23, 2004. Tokyo University: Tokyo Keizai University. Retrieved January 20, 2008, from <http://jalt.org/pansig/2004/HTML/Nakamura.htm>

- National Council of Teacher of Mathematic [NCTM]. (1995). **Assessment Standards for School Mathematic**. Retrieved March 21, 2003, from [http://standards.nctm.org/Previous/ AssStads/Intro.htm](http://standards.nctm.org/Previous/AssStads/Intro.htm)
- Nitko, Anthony J. (1996). **Educational Assessment of Student**. 2nd ed. Englewood Cliffs, N.J.: Merrill.
- O'Brien, R. M.; & Jones, B. (1986). The Reliability of School – level Aggregate Variable: An Application of Generalizability Theory. **Journal of Research and Development in Education**. 20: 21-27.
- Park, Taejoon. (2004). An Investigation of an ESL Placement Test of Writing Using Many-Facet Rasch measurement. **TESOL & Applied Linguistics**. 4(1). Retrieved July 22, 2008, from <http://journals.tc-library.org/index.php/tesol/article/viewFile/41/48>
- Partnership for Reform Initiatives in Science and Mathematics [PRISM]. (2001). **Opened Response Questioning Strategies**. Retrieved June 26, 2006, from <http://www.murraystate.edu/prism/kyprism/openrp1.htm>
- Paulukonis, Susan T.; Myford, Carol M.; & Heller, Joan I. (n.d). **Formative Evaluation of a Performance Assessment Scoring System**. Retrieved October 21, 2008 from http://books.google.co.th/books?id=w9VGRj8jth4C&pg=PA15&lpg=PA15&dq=formative+evaluation+of+a+performance+assessment+scoring+system&source=bl&ots=MtwcNcfUDL&sig=oZKSBfBspXa-P0GsftoC26CbLfl&hl=th&sa=X&oi=book_result&resnum=1&ct=result
- Rudner, Lawrence M. (1992). Reducing Errors Due to the Use of Judges. **Practical Assessment, Research & Evaluation**. 3(3). Retrieved September 16, 2008 from <http://PAREonline.net/getvn.asp?v=3&n=3>
- Schumacker, Randall E.; & Smith, Everett V., Jr. (2007, June). Reliability: A Rasch Perspective. **Educational and Psychological Measurement**. 67(3): 394-409.
- Shavelson, Rechar J.; & Webb, Noreen N. (1991). **Generalizability Theory: A Primer**. Sage Publications.
- Smith, Evertt V., Jr. (2001). Evidence for the Reliability of Measure and Validity of Measure Interpretation: A Rasch Measurement Perspective. **Journal of Applied Measurement**. 2(3): 281-311.

- Smith, Everett V., Jr.; & Kulikowich, Jonna M. (2004, August). An Application of Generalizability Theory and Many-Facet Rasch Measurement Using A Complex Problem-Solving Skills Assessment. *Educational and Psychological Measurement*. 64(4): 617-639.
- Stenmark, Jean K. (1991). **Mathematics Assessment : Myths, Models, Good Questions, and Practical Suggestions**. 6th ed. Reston, Virginia: National Council of Teachers of Mathematics.
- Stemler, Steven E. (2004). **A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability**. Retrieved June 26, 2006, from www.pareonline.net/getvn.asp?v=9&n=4
- Stiggins, Richard J. (1994). **Student-Centered Classroom Assessment**. New York: Maxwell Macmillan College.
- Sudweeks, Richard R.; Reeve, Suzanne; & Bradshaw, William S. (2005). A Comparison of Generalizability Theory and Many-Facet Rasch Measurement in an Analysis of College Sophomore Writing. *Assessment Writing*. 9(3): 239-261.
- Swartz, Carl M.; et al. (1999, June). Using Generalizability Theory to Estimate the Reliability of Writing Scores Derived from Holistic and Analytical Scoring Methods. *Educational and Psychological Measurement*. 59(3): 492-506.
- Takahashi, A. (2000). **Open-Ended Solving on Web**. Retrieved June 6, 2006, from http://www.mste.uiue /users/aki/open_ended/
- The Online Teacher Resource. (2009). **Why Rubrics?**. Retrieved September 6, 2009, from <http://www.teach-nology.com/tutorials/teaching/rubrics/>
- Thompson, Denisse R.; & Senk, Sharon L. (1998). Using Rubrics in High School Mathematics Courses. *The Mathematics Teacher*. 91(9): 786-793.
- Tindal, Gerald; Yovanoff, Paul; & Geller, Josh P. (2010, May). Generalizability Theory Applied to Reading Assessment for Students with Significant Cognitive Disabilities. *The Journal of Special Education*. 44(1): 3-17.
- Turner, Jere. (2003). **Examining an Art Portfolio Assessment Using a Many-Facet Rasch Measurement Model**. Retrieved January 12, 2007, from <http://proquest.umi.com/pqdlink?did=764823501&Fmt=7&clientId=79356&RQS=309&VName=PQD>

- Uebersax, John. (2000). **Raw Agreement Indices**. Retrieved August 13, 2006, from <http://ourworld.compuserve.com/homepages/jsuebersax/raw.htm>
- (2003). **Statistical Methods for Rater Agreement**. Retrieved August 13, 2006, from <http://ourworld.compuserve.com/homepage/jsuebersax/agree/htm>
- Upsher, John A.; Turner, Carolyn E. (1999). Systematic Effect in the Rating of Second-Language Speaking Ability: Test Method and Learner Discourse. **Language Testing**. 16(82): 82-111. Retrieved November 11, 2008, from <http://ltj.sagepub.com/cgi/content/abstract/16/1/82>
- Sanchez, Wendy; & Ice, Nicole. (2004, July/August). **Open-Ended Items Better Reveal Students' Mathematical Thinking**. Retrieved June 10, 2009, from <http://www.nctm.org/news/release.aspx?id=754>
- Woodruff, David J.; & Feldt, Leonard S. (1986, September). Test for Equality of Several Alpha Coefficients When Their Sample Estimates are Dependent. **Psychometrika**. 51(2): 393-413.
- Wright, Benjamin D.; & Master, Geofferey N. (1982). **Rating Scale Analysis: Rasch Measurement**. MESA Press.
- Zimmaro, Dawn M. (2004). **Developing Grading Rubrics**. Retrieved September 10, 2008, from www.utexas.edu/academic/mec



ภาคผนวก ก
การสร้างและพัฒนาเครื่องมือ

- รายชื่อผู้เชี่ยวชาญตรวจสอบคุณภาพเครื่องมือ
- การวิเคราะห์หลักสูตรกลุ่มสาระการเรียนรู้คณิตศาสตร์
- ค่าดัชนีความสอดคล้อง (IOC)
- การวิเคราะห์องค์ประกอบเชิงยืนยัน
- ค่าความยาก (p) และค่าอำนาจจำแนก (r) ของแบบทดสอบ

รายชื่อผู้เชี่ยวชาญตรวจคุณภาพเครื่องมือ

1. รศ.ดร. เกียรติสุดา ศรีสุข
สายวิชาประเมินผลและวิจัยการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยเชียงใหม่
2. อาจารย์สุรชัย อินทสังข์
ภาควิชาหลักสูตรและการสอน คณะศึกษาศาสตร์ มหาวิทยาลัยรามคำแหง
3. อาจารย์ศักดิ์ดา สวาทะนันท์
กลุ่มสาระการเรียนรู้คณิตศาสตร์ โรงเรียนสาธิตมหาวิทยาลัยเชียงใหม่
4. อาจารย์สุชาติ สิริมีนันท์
กลุ่มสาระการเรียนรู้คณิตศาสตร์ โรงเรียนบางกะปิ
สำนักงานเขตพื้นที่การศึกษากรุงเทพมหานคร เขต 1
5. อาจารย์นิพนธ์ สารถ้อย
ศึกษานิเทศก์ สำนักงานเขตพื้นที่การศึกษา น่าน เขต 2

รายชื่อผู้เชี่ยวชาญตรวจให้คะแนนแบบทดสอบปลายเปิด

1. อาจารย์สุรพล กาละดี
กลุ่มสาระการเรียนรู้คณิตศาสตร์ โรงเรียนสา สำนักงานเขตพื้นที่การศึกษา น่าน เขต 1
2. อาจารย์ณพวิศก พลทิพย์
กลุ่มสาระการเรียนรู้คณิตศาสตร์ โรงเรียนยาบห้วนา สำนักงานเขตพื้นที่การศึกษา น่าน เขต 1
3. อาจารย์สมคิด นันคำ
กลุ่มสาระการเรียนรู้คณิตศาสตร์ โรงเรียนบ้านห้วยแอ้ว สำนักงานเขตพื้นที่การศึกษา น่าน เขต 1
4. อาจารย์ณัฐพร การศิริ
กลุ่มสาระการเรียนรู้คณิตศาสตร์ โรงเรียนบ้านหลวง สำนักงานเขตพื้นที่การศึกษา น่าน เขต 1
5. อาจารย์สุภาพรรณ สิทธิวงศ์
กลุ่มสาระการเรียนรู้คณิตศาสตร์ โรงเรียนสา สำนักงานเขตพื้นที่การศึกษา น่าน เขต 1
6. อาจารย์เวธิต ยะติ
กลุ่มสาระการเรียนรู้คณิตศาสตร์ โรงเรียนน่านน้อย สำนักงานเขตพื้นที่การศึกษา น่าน เขต 1
7. อาจารย์นิวัฒน์ ใจบาล
กลุ่มสาระการเรียนรู้คณิตศาสตร์ โรงเรียนสันติสุขพิทยาคม สำนักงานเขตพื้นที่การศึกษา น่าน เขต 1

ตาราง 7 การวิเคราะห์หลักสูตรกลุ่มสาระการเรียนรู้คณิตศาสตร์ ระดับชั้นมัธยมศึกษาตอนต้น

สาระที่	ชั้น	มาตรฐานการเรียนรู้/ ตัวชี้วัด	ระดับพฤติกรรม						รวม
			ความรู้-ความจำ	ความเข้าใจ	การนำไปใช้	การวิเคราะห์	การสังเคราะห์	การประเมินค่า	
1 จำนวน และการ ดำเนินการ	มาตรฐาน ค 1.1 เข้าใจถึงความหลากหลายของการแสดงจำนวนและการใช้จำนวนในชีวิตจริง								
	ม.1	1. ระบุหรือยกตัวอย่าง และเปรียบเทียบจำนวนเต็มบวก จำนวนเต็มลบ ศูนย์ เศษส่วนและทศนิยม	-	-	-	-	-	-	-
		2. เข้าใจเกี่ยวกับเลขยกกำลังที่มีเลขชี้กำลัง เป็นจำนวนเต็ม และเขียนแสดงจำนวน ให้อยู่ในรูปสัญกรณ์วิทยาศาสตร์ (scientific notation)	-	-	-	-	-	-	-
	ม.2	1. เขียนเศษส่วนในรูปทศนิยมและเขียนทศนิยมซ้ำในรูปเศษส่วน	-	-	-	-	-	-	-
		2. จำแนกจำนวนจริงที่กำหนดให้ และ ยกตัวอย่างจำนวนตรรกยะและจำนวน อตรรกยะ	-	-	-	-	-	-	-
		3. อธิบายและระบุรากที่สองและรากที่สามของจำนวนจริง	-	-	-	-	-	-	-
	4. ใช้ความรู้เกี่ยวกับอัตราส่วน สัดส่วน และร้อยละในการแก้โจทย์ปัญหา	-	-	1	-	-	-	1	
ม.3	-	-	-	-	-	-	-	-	
มาตรฐาน ค 1.2 เข้าใจถึงผลที่เกิดขึ้นจากการดำเนินการของจำนวนและความสัมพันธ์ระหว่างการดำเนินการต่าง ๆ และใช้การดำเนินการในการแก้ปัญหา									
ม.1	1. บวก ลบ คูณ หารจำนวนเต็ม และ นำไปใช้แก้ปัญหา ตระหนักถึงความสมเหตุสมผลของคำตอบ อธิบายผลที่เกิดขึ้นจากการบวก การลบ การคูณ การหาร และบอกความสัมพันธ์ของ การบวกกับการลบ การคูณกับการหารของจำนวนเต็ม								
	2. บวก ลบ คูณ หารเศษส่วนและทศนิยม และ นำไปใช้แก้ปัญหา ตระหนักถึงความสมเหตุสมผลของคำตอบ อธิบายผลที่เกิดขึ้นจากการบวก	-	-	-	-	-	-	-	-

ตาราง 7 (ต่อ)

สาระที่	ชั้น	มาตรฐานการเรียนรู้/ ตัวชี้วัด	ระดับพฤติกรรม						รวม
			ความรู้-ความจำ	ความเข้าใจ	การนำไปใช้	การวิเคราะห์	การสังเคราะห์	การประเมินค่า	
1 จำนวน และการ ดำเนิน การ		การ ลบ การคูณ การหาร และบอกความสัมพันธ์ของการบวกกับการลบ การคูณกับการหารของเศษส่วนและทศนิยม	-	-	1	-	-	-	1
		3. อธิบายผลที่เกิดขึ้นจากการยกกำลังของ จำนวนเต็ม เศษส่วนและทศนิยม	-	-	-	-	-	-	-
		4. คูณและหารเลขยกกำลังที่มีฐานเดียวกัน และเลขชี้กำลังเป็นจำนวนเต็ม	-	-	-	-	-	-	-
	ม.2	1. หารากที่สองและรากที่สามของจำนวนเต็มโดยการแยกตัวประกอบและนำไปใช้ในการแก้ปัญหา พร้อมทั้งตระหนักถึงความสมเหตุสมผลของคำตอบ	-	-	-	-	-	-	-
		2. อธิบายผลที่เกิดขึ้นจากการหารากที่สองและรากที่สามของจำนวนเต็ม เศษส่วน และทศนิยม บอกความสัมพันธ์ของการยกกำลังกับการหารากของจำนวนจริง	-	-	-	-	-	-	-
	ม.3	-	-	-	-	-	-	-	-
	มาตรฐาน ค 1.3 ใช้การประมาณค่าในการคำนวณและแก้ปัญหา								
	ม.1	1. ใช้การประมาณค่าในสถานการณ์ต่าง ๆ ได้อย่างเหมาะสม รวมถึงใช้ในการพิจารณาความสมเหตุสมผลของคำตอบที่ได้จากการคำนวณ	-	-	1	-	-	-	1
	ม.2	2. หาค่าประมาณของรากที่สอง และรากที่สามของจำนวนจริง และนำไปใช้ในการแก้ปัญหา พร้อมทั้งตระหนักถึงความสมเหตุสมผลของคำตอบ	-	-	-	-	-	-	-
	ม.3	-	-	-	-	-	-	-	-

ตาราง 7 (ต่อ)

สาระที่	ชั้น	มาตรฐานการเรียนรู้/ ตัวชี้วัด	ระดับพฤติกรรม						รวม
			ความรู้-ความจำ	ความเข้าใจ	การนำไปใช้	การวิเคราะห์	การสังเคราะห์	การประเมินค่า	
2 การวัด		มาตรฐาน ค 1.4 : เข้าใจระบบจำนวนและนำเสนอสมบัติเกี่ยวกับจำนวนไปใช้							
	ม.1	1. นำความรู้และสมบัติเกี่ยวกับจำนวนเต็มไปใช้ในการแก้ปัญหา	—	—	1	—	—	—	1
	ม.2	1. บอกความเกี่ยวข้องของจำนวนจริง จำนวนตรรกยะ และจำนวนอตรรกยะ	—	—	—	—	—	—	—
	ม.3	—	—	—	—	—	—	—	—
		มาตรฐาน ค 2.1 : เข้าใจพื้นฐานเกี่ยวกับการวัด วัดและคาดคะเนขนาดของสิ่งที่ต้องการวัด							
	ม.1	—	—	—	—	—	—	—	—
	ม.2	1. เปรียบเทียบหน่วยความยาว หน่วยพื้นที่ในระบบเดียวกัน และต่างระบบ และเลือกใช้หน่วยการวัดได้อย่างเหมาะสม 2. คาดคะเนเวลา ระยะทาง พื้นที่ ปริมาตรและน้ำหนักได้อย่างใกล้เคียง และอธิบายวิธีการที่ใช้ในการคาดคะเน 3. ใช้การคาดคะเนเกี่ยวกับการวัดในสถานการณ์ต่าง ๆ ได้อย่างเหมาะสม	—	—	—	—	—	—	—
	ม.3	1. หาพื้นที่ผิวของปริซึมและทรงกระบอก 2. หาปริมาตรของปริซึม ทรงกระบอก พีระมิดกรวย และทรงกลม 3. เปรียบเทียบหน่วยความจุ หรือหน่วยปริมาตรในระบบเดียวกันหรือต่างระบบ และเลือกใช้หน่วยการวัดได้อย่างเหมาะสม 4. ใช้การคาดคะเนเกี่ยวกับการวัดในสถานการณ์ต่าง ๆ ได้อย่างเหมาะสม	—	—	—	—	—	—	—
			—	1	—	—	—	—	1
			—	—	—	—	—	—	—
			—	—	—	—	—	—	—

ตาราง 7 (ต่อ)

สาระที่	ชั้น	มาตรฐานการเรียนรู้/ ตัวชี้วัด	ระดับพฤติกรรม						รวม
			ความรู้-ความจำ	ความเข้าใจ	การนำไปใช้	การวิเคราะห์	การสังเคราะห์	การประเมินค่า	
2 การวัด	มาตรฐาน ค 2.2: แก้ปัญหาเกี่ยวกับการวัด								
	ม.1	-	-	-	-	-	-	-	-
	ม.2	1. ใช้ความรู้เกี่ยวกับความยาวและพื้นที่ แก้ปัญหา ในสถานการณ์ต่าง ๆ	-	-	-	-	-	-	-
	ม.3	1. ใช้ความรู้เกี่ยวกับพื้นที่ พื้นที่ผิว และ ปริมาตรในการแก้ปัญหาในสถานการณ์ต่าง ๆ	-	-	1	-	-	-	1
3 เรขาคณิต	มาตรฐาน ค 3.1 : อธิบายและวิเคราะห์รูปเรขาคณิตสองมิติและสามมิติ								
	ม.1	1. สร้างและบอกขั้นตอนการสร้างพื้นฐานทางเรขาคณิต	-	-	-	-	-	-	-
		2. สร้างรูปเรขาคณิตสองมิติโดยใช้การสร้างพื้นฐานทางเรขาคณิต และบอกขั้นตอนในการสร้างโดยไม่เน้นการพิสูจน์	-	-	-	-	-	-	-
		3. สืบเสาะ สังเกต และคาดการณ์เกี่ยวกับสมบัติทางเรขาคณิต	-	-	-	-	-	-	-
		4. อธิบายลักษณะของรูปเรขาคณิตสามมิติจากภาพที่กำหนดให้	-	-	-	-	-	-	-
		5. ระบุนภาพสองมิติที่ได้จากการมองด้านหน้า (front view) ด้านข้าง (side view) หรือ ด้านบน (top view) ของรูปเรขาคณิตสามมิติที่กำหนดให้	-	-	-	-	-	-	-
		6. วาดหรือประดิษฐ์รูปเรขาคณิตสามมิติที่ประกอบขึ้นจากลูกบาศก์ เมื่อกำหนดภาพสองมิติที่ได้จากการมองด้านหน้า ด้านข้างและด้านบนให้	-	1	-	-	-	-	1
ม.2	-	-	-	-	-	-	-	-	
ม.3	1. อธิบายลักษณะและสมบัติของปริซึม พีระมิด ทรงกระบอก กรวย และทรงกลม	-	-	-	-	-	-	-	

ตาราง 7 (ต่อ)

สาระที่	ชั้น	มาตรฐานการเรียนรู้/ ตัวชี้วัด	ระดับพฤติกรรม						รวม
			ความรู้-ความจำ	ความเข้าใจ	การนำไปใช้	การวิเคราะห์	การสังเคราะห์	การประเมินค่า	
3 เรขาคณิต		มาตรฐาน ค 3.2 : ใช้การนิกภาพ (visualization) ให้เหตุผลเกี่ยวกับปริภูมิ (spatial reasoning) และใช้แบบจำลองทางเรขาคณิต (geometric model) ในการแก้ปัญหา							
	ม.1	–	–	–	–	–	–	–	–
	ม.2	1. ใช้สมบัติเกี่ยวกับความเท่ากันทุกประการของรูปสามเหลี่ยมและสมบัติของเส้นขนานในการให้เหตุผลและแก้ปัญหา	–	–	–	–	–	–	–
		2. ใช้ทฤษฎีบทพีทาโกรัสและบทกลับในการให้เหตุผลและแก้ปัญหา	–	–	–	1	–	–	1
		3. เข้าใจเกี่ยวกับการแปลงทางเรขาคณิตในเรื่องการเลื่อนขนาน การสะท้อน และการหมุน และนำไปใช้	–	1	–	–	–	–	1
4 พีชคณิต		4. บอกภาพที่เกิดขึ้นจากการเลื่อนขนาน การสะท้อน และการหมุนรูปต้นแบบ และอธิบายวิธีการที่จะได้ภาพที่ปรากฏเมื่อกำหนดรูปต้นแบบและภาพนั้นให้	–	–	–	–	–	–	–
	ม.3	1. ใช้สมบัติของรูปสามเหลี่ยมคล้ายในการให้เหตุผลและการแก้ปัญหา	–	–	–	–	–	–	–
		มาตรฐาน ค 4.1 : เข้าใจและวิเคราะห์แบบรูป (pattern) ความสัมพันธ์ และฟังก์ชัน							
	ม.1	1. วิเคราะห์และอธิบายความสัมพันธ์ของแบบรูปที่กำหนดให้	–	–	–	1	–	–	1
	ม.2	–	–	–	–	–	–	–	–
	ม.3	–	–	–	–	–	–	–	–

ตาราง 7 (ต่อ)

สาระที่	ชั้น	มาตรฐานการเรียนรู้/ ตัวชี้วัด	ระดับพฤติกรรม						รวม
			ความรู้-ความจำ	ความเข้าใจ	การนำไปใช้	การวิเคราะห์	การสังเคราะห์	การประเมินค่า	
4 พีช คณิต		มาตรฐาน ค 4.2 : ใช้นิพจน์ สมการ อสมการ กราฟ และตัวแบบเชิงคณิตศาสตร์ (mathematical model) อื่น ๆ แทนสถานการณ์ต่าง ๆ ตลอดจนแปลความหมายและนำไปใช้แก้ปัญหา							
	ม.1	1. แก้สมการเชิงเส้นตัวแปรเดียวอย่างง่าย	-	-	-	-	-	-	-
		2. เขียนสมการเชิงเส้นตัวแปรเดียวจากสถานการณ์หรือปัญหาอย่างง่าย	-	-	-	-	-	-	-
		3. แก้โจทย์ปัญหาเกี่ยวกับสมการเชิงเส้นตัวแปรเดียวอย่างง่าย พร้อมทั้งตระหนักถึงความสมเหตุสมผลของคำตอบ	-	-	-	-	-	-	-
		4. เขียนกราฟบนระนาบในระบบพิกัดฉากแสดงความเกี่ยวข้องของปริมาณสองชุดที่กำหนดให้	-	-	-	-	-	-	-
		5. อ่านและแปลความหมายของกราฟบนระนาบในระบบพิกัดฉากที่กำหนดให้	-	-	-	-	-	-	-
	ม.2	1. แก้โจทย์ปัญหาเกี่ยวกับสมการเชิงเส้นตัวแปรเดียว พร้อมทั้งตระหนักถึงความสมเหตุสมผลของคำตอบ	-	1	-	-	-	-	1
		2. หาพิกัดของจุด และอธิบายลักษณะของรูปเรขาคณิตที่เกิดขึ้นจากการเลื่อนขนาน การสะท้อน และการหมุนบนระนาบในระบบพิกัดฉาก	-	-	-	-	-	-	-
	ม.3	1. ใช้ความรู้เกี่ยวกับสมการเชิงเส้นตัวแปรเดียวในการแก้ปัญหา พร้อมทั้งตระหนักถึงความสมเหตุสมผลของคำตอบ	-	-	-	-	-	-	-
		2. เขียนกราฟแสดงความเกี่ยวข้องระหว่างปริมาณสองชุดที่มีความสัมพันธ์เชิงเส้น	-	-	-	-	-	-	-
		3. เขียนกราฟของสมการเชิงเส้นสองตัวแปร	-	-	-	-	-	-	-

ตาราง 7 (ต่อ)

สาระที่	ชั้น	มาตรฐานการเรียนรู้/ ตัวชี้วัด	ระดับพฤติกรรม						รวม
			ความรู้-ความจำ	ความเข้าใจ	การนำไปใช้	การวิเคราะห์	การสังเคราะห์	การประเมินค่า	
4 พีช คณิต		4. อ่านและแปลความหมาย กราฟของระบบ สมการเชิงเส้นสองตัวแปร และกราฟอื่น ๆ	-	-	-	-	-	-	-
		5. แก่ระบบสมการเชิงเส้นสองตัวแปร และ นำไปใช้แก้ปัญหา พร้อมทั้งตระหนักถึงความ สมเหตุสมผลของคำตอบ	-	-	1	-	-	-	1
5 การ วิเคราะห์ ข้อมูล และ ความ น่าจะเป็น		มาตรฐาน ค 5.1 เข้าใจและใช้วิธีการทางสถิติในการ วิเคราะห์ข้อมูล							
	ม.1	-	-	-	-	-	-	-	-
	ม.2	1. อ่านและนำเสนอข้อมูลโดยใช้แผนภูมิรูปวงกลม	-	-	-	-	-	-	-
	ม.3	1. กำหนดประเด็น และเขียนข้อคำถามเกี่ยวกับ ปัญหาหรือสถานการณ์ต่าง ๆ รวมทั้งกำหนดวิธี การศึกษาและการเก็บรวบรวมข้อมูลที่เหมาะสม	-	-	-	-	-	-	-
		2. หาค่าเฉลี่ยเลขคณิต มัธยฐาน และฐานนิยม ของข้อมูลที่ไม่ได้แจกแจงความถี่ และเลือกใช้ได้ อย่างเหมาะสม	-	-	-	1	-	-	1
		3. นำเสนอข้อมูลในรูปแบบที่เหมาะสม	-	-	-	-	-	-	-
		4. อ่าน แปลความหมาย และวิเคราะห์ข้อมูลที่ได้ จากการนำเสนอ	-	-	-	-	-	-	-
		มาตรฐาน ค 5.2 : ใช้วิธีการทางสถิติและความรู้เกี่ยวกับ ความน่าจะเป็นในการคาดการณ์ได้อย่างสมเหตุสมผล							
	ม.1	1. อธิบายได้ว่าเหตุการณ์ที่กำหนดให้ เหตุการณ์ ใดจะมีโอกาสเกิดขึ้นได้มากกว่ากัน	-	-	-	-	-	-	-
	ม.2	1. อธิบายได้ว่าเหตุการณ์ที่กำหนดให้ เหตุการณ์ ใดเกิดขึ้นแน่นอน เหตุการณ์ใดไม่เกิดขึ้นแน่นอน และเหตุการณ์ใดมีโอกาสเกิดขึ้นได้มากกว่ากัน	-	-	-	-	-	-	-

ตาราง 7 (ต่อ)

สาระที่	ชั้น	มาตรฐานการเรียนรู้/ ตัวชี้วัด	ระดับพฤติกรรม						รวม
			ความรู้-ความจำ	ความเข้าใจ	การนำไปใช้	การวิเคราะห์	การสังเคราะห์	การประเมินค่า	
5 การ วิเคราะห์ ข้อมูล และ ความ น่าจะเป็น	ม.3	1.หาความน่าจะเป็นของเหตุการณ์จากการ ทดลองสุ่มที่ผลแต่ละตัวมีโอกาสเกิดขึ้น เท่า ๆ กัน และใช้ความรู้เกี่ยวกับความน่าจะเป็นในการ คาดการณ์ได้อย่างสมเหตุสมผล	-	1	-	-	-	-	1
		มาตรฐาน ค 5.3 : ใช้ความรู้เกี่ยวกับสถิติและความน่าจะเป็น ช่วยในการตัดสินใจและแก้ปัญหา							
	ม.1	-	-	-	-	-	-	-	-
	ม.2	-	-	-	-	-	-	-	-
	ม.3	1. ใช้ความรู้เกี่ยวกับสถิติและความน่าจะเป็น ประกอบการตัดสินใจในสถานการณ์ต่างๆ	-	-	-	1	-	-	1
		2. อภิปรายถึงความคลาดเคลื่อนที่อาจเกิดขึ้นได้ จากการนำเสนอข้อมูลทางสถิติ	-	-	-	-	-	-	-
รวม			-	5	6	4	-	-	15

ดัชนีความสอดคล้อง (IOC) ของผู้เชี่ยวชาญในการตรวจสอบคุณภาพของเครื่องมือ

ข้อที่	รายการ	ผู้เชี่ยวชาญคนที่					รวม	IOC	แปลผล
		1	2	3	4	5			
1	ข้อคำถาม	1	1	1	1	1	5	1.00	ใช้ได้
	เกณฑ์การให้คะแนน	1	1	1	1	1	5	1.00	ใช้ได้
2	ข้อคำถาม	1	1	1	1	1	5	1.00	ใช้ได้
	เกณฑ์การให้คะแนน	1	1	0	1	1	4	0.80	ใช้ได้
3	ข้อคำถาม	1	1	1	1	1	5	1.00	ใช้ได้
	เกณฑ์การให้คะแนน	1	1	1	1	1	5	1.00	ใช้ได้
4	ข้อคำถาม	1	1	1	1	1	5	1.00	ใช้ได้
	เกณฑ์การให้คะแนน	1	0	1	1	0	3	0.60	ใช้ได้
5	ข้อคำถาม	1	1	1	1	1	5	1.00	ใช้ได้
	เกณฑ์การให้คะแนน	1	1	1	1	1	5	1.00	ใช้ได้
6	ข้อคำถาม	1	1	1	1	1	5	1.00	ใช้ได้
	เกณฑ์การให้คะแนน	1	1	1	1	1	5	1.00	ใช้ได้
7	ข้อคำถาม	1	1	1	0	1	4	0.80	ใช้ได้
	เกณฑ์การให้คะแนน	1	1	1	0	1	4	0.80	ใช้ได้
8	ข้อคำถาม	1	1	1	1	1	5	1.00	ใช้ได้
	เกณฑ์การให้คะแนน	1	1	1	1	1	5	1.00	ใช้ได้
9	ข้อคำถาม	1	1	1	1	1	5	1.00	ใช้ได้
	เกณฑ์การให้คะแนน	1	1	1	1	1	5	1.00	ใช้ได้
10	ข้อคำถาม	1	1	1	1	1	5	1.00	ใช้ได้
	เกณฑ์การให้คะแนน	1	1	1	1	1	5	1.00	ใช้ได้
11	ข้อคำถาม	1	1	0	1	1	4	0.80	ใช้ได้
	เกณฑ์การให้คะแนน	1	1	0	1	1	4	0.80	ใช้ได้
12	ข้อคำถาม	1	1	1	1	1	5	1.00	ใช้ได้
	เกณฑ์การให้คะแนน	1	1	1	1	1	5	1.00	ใช้ได้
13	ข้อคำถาม	1	1	1	1	1	5	1.00	ใช้ได้
	เกณฑ์การให้คะแนน	1	1	1	1	1	5	1.00	ใช้ได้
14	ข้อคำถาม	1	1	1	1	1	5	1.00	ใช้ได้
	เกณฑ์การให้คะแนน	1	1	1	1	1	5	1.00	ใช้ได้
15	ข้อคำถาม	1	0	0	1	0	2	0.40	ปรับปรุง
	เกณฑ์การให้คะแนน	1	0	0	1	0	2	0.40	ปรับปรุง

ผลการวิเคราะห์องค์ประกอบเชิงยืนยัน

Goodness of Fit Statistics

Degrees of Freedom = 52
 Minimum Fit Function Chi-Square = 66.12 (P=0.090)
 Normal Theory Weighted Least Squares Chi-Square = 60.17 (P=0.20)
 Estimated Non-centrality Parameter (NCP) = 8.17
 90 Percent Confidence Interval for NCP = (0.0 ; 31.61)

 Minimum Fit Function Value = 0.37
 Population Discrepancy Function Value (F0) = 0.046
 90 Percent Confidence Interval for F0 = (0.0 ; 0.18)
 Root Mean Square Error of Approximation (RMSEA) = 0.030
 90 Percent Confidence Interval for RMSEA = (0.0 ; 0.058)
 P-Value for Test of Close Fit (RMSEA < 0.05) = 0.86

 Expected Cross-Validation Index (ECVI) = 0.63
 90 Percent Confidence Interval for ECVI = (0.58 ; 0.76)
 ECVI for Saturated Model = 0.87
 ECVI for Independence Model = 14.05

 Chi-Square for Independence Model with 66 Degrees of Freedom = 2490.16
 Independence AIC = 2514.16
 Model AIC = 112.17
 Saturated AIC = 156.00
 Independence CAIC = 2564.47
 Model CAIC = 221.19
 Saturated CAIC = 483.05

 Normed Fit Index (NFI) = 0.99
 Non-Normed Fit Index (NNFI) = 0.99
 Parsimony Normed Fit Index (PNFI) = 0.77
 Comparative Fit Index (CFI) = 0.99
 Incremental Fit Index (IFI) = 0.99
 Relative Fit Index (RFI) = 0.97

 Critical N (CN) = 213.84

 Root Mean Square Residual (RMR) = 0.059
 Standardized RMR = 0.036
 Goodness of Fit Index (GFI) = 0.95
 Adjusted Goodness of Fit Index (AGFI) = 0.92
 Parsimony Goodness of Fit Index (PGFI) = 0.63

ค่าความยาก (P) และค่าอำนาจจำแนก (r) ของแบบทดสอบปลายเปิดวิชาคณิตศาสตร์

ข้อ	P	r
1	.6750	.6170
2	.4959	.5582
3	.7729	.3817
4	.3793	.3750
5	.5794	.2015
6	.4126	.6752
7	.2499	.4998
8	.3459	.5918
9	.6832	.5000
10	.4248	.4835
11	.5919	.5498
12	.3541	.5247

ภาคผนวก ข
การวิเคราะห์ข้อมูลการวิจัย

- ตัวอย่างคำสั่งและผลการวิเคราะห์ด้วยโมเดลการสรุปอ้างอิง
- ตัวอย่างคำสั่งและผลการวิเคราะห์ด้วยโมเดลหลายองค์ประกอบของราล์ช

ผลการวิเคราะห์คุณลักษณะคะแนนด้วยโมเดลการสุบอ้างอิง
เมื่อใช้ผู้ตรวจ 4 คน ตรวจข้อสอบทุกข้อของผู้สอบทุกคน (P x I x R DESIGN)

```

GGGGGGGGGG EEEEEEEEEEE NN NN 00000000 VV VV AAAAAAAAAA
GGGGGGGGGG EEEEEEEEEEE NNN NN 0000000000 VV VV AAAAAAAAAA
GG G EE EEEEEEE NNNN NN 00 00 VV VV AA AA
GG EE NN NN NN 00 00 VV VV AA AA
GG EE NN NN NN 00 00 VV VV AA AA
GG EE NN NN NN 00 00 VV VV AA AA
GG EE NN NN NN 00 00 VV VV AA AA
GG GGGG EEEEEEE NN NNNN 00 00 VV VV AAAAAAAAAA
GG GGGG EE NN NNN 00 00 VV VV AAAAAAAAAA
GG GG EE NN NN 00 00 VV VV AA AA
GG GG EE NN NN 00 00 VV VV AA AA
GG GG EE NN NN 00 00 V V AA AA
GG GG EE NN NN 00 00 VVVV AA AA
GGGGGGGGGG EEEEEEEEEEE NN NN 0000000000 VVVV AA AA
GGGGGGGGGG EEEEEEEEEEE NN NN 00000000 VV AA AA

```

A GENERAL PURPOSE ANALYSIS OF VARIANCE SYSTEM

GENOVA IS A FORTRAN 77 PROGRAM FOR ANALYSIS OF VARIANCE
AND GENERALIZABILITY ANALYSES WITH BALANCED DESIGNS

AUTHORS

Joe E. Crick, Ed.D.
Chief Technology & Information Officer
Vice President Applications and Database Services
National Board of Medical Examiners
Philadelphia, PA 19104

Robert L. Brennan, Ed.D.
Director, Iowa Testing Program
University of Iowa
Iowa City, Iowa 52242

VERSION 3.1
January, 2001

GENOVA has been checked for accuracy of output, however the authors
can make no assurances that the program is totally without error.

GENOVA was developed in part under contract No. N00123-78-C-1206 with the Navy Personnel Research and Development Center (NPRDC); Robert L. Brennan Principal Investigator. GENOVA does not necessarily reflect NPRDC positions or policy, and no official endorsement should be inferred

GENOVA VERSION 3.1
PAGE 1

CONTROL CARD INPUT LISTING

```

COLUMN 111111111222222223333333334444444445555555556666666667777777778
1234567890123456789012345678901234567890123456789012345678901234567890
GSTUDY P X I X R DESIGN -- RANDOM MODEL
COMMENT
COMMENT # RECORDS = 180
COMMENT # VALUES PER RECORD = 48
COMMENT
OPTIONS RECORDS ALL CORRELATION
EFFECT * P 180 0
EFFECT + I 12 0
EFFECT + R 4 0
FORMAT (48F1.0)
PROCESS
GENOVA VERSION 3.1
PAGE 2
G STUDY

```

P X I X R DESIGN -- RANDOM MODEL

EXPANDED MAIN AND INTERACTION EFFECT TABLE

(** = INFINITE)	P 180	I 12	R 4	PRIMARY INDICES	TOTAL NUMBER INDICES	DEGREES OF FREEDOM
SAMPLE SIZE UNIVERSE SIZE	****	****	****			
* P	1	0	0	1	1	179
* I	0	1	0	1	1	11
* R	0	0	1	1	1	3
* PI	1	1	0	2	2	1969
* PR	1	0	1	2	2	537
* IR	0	1	1	2	2	33
* PIR	1	1	1	3	3	5907

GENOVA VERSION 3.1
PAGE 33
G STUDY

P X I X R DESIGN -- RANDOM MODEL

CELL MEAN SCORES

*** GRAND MEAN = 2.3001157 ***

MEAN SCORES FOR EFFECT: I

SUBSCRIPT NOTATION: (I)

(1) =	2.670833	(2) =	2.529167	(3) =	3.258333	(4) =	1.922222
(5) =	2.606944	(6) =	2.050000	(7) =	1.483333	(8) =	1.648611
(9) =	2.930556	(10) =	2.131944	(11) =	2.297222	(12) =	2.072222

MEAN SCORES FOR EFFECT: R

SUBSCRIPT NOTATION: (R)

(1) =	2.149074	(2) =	2.386574	(3) =	2.341204	(4) =	2.323611
-------	----------	-------	----------	-------	----------	-------	----------

MEAN SCORES FOR EFFECT: IR

SUBSCRIPT NOTATION: (I,R)

(1,1) =	2.505556	(1,2) =	2.722222	(1,3) =	2.744444	(1,4) =	2.711111
(2,1) =	2.466667	(2,2) =	2.527778	(2,3) =	2.661111	(2,4) =	2.461111
(3,1) =	3.033333	(3,2) =	3.366667	(3,3) =	3.377778	(3,4) =	3.255556
(4,1) =	1.711111	(4,2) =	1.977778	(4,3) =	2.022222	(4,4) =	1.977778
(5,1) =	2.650000	(5,2) =	2.550000	(5,3) =	2.583333	(5,4) =	2.644444
(6,1) =	1.922222	(6,2) =	2.033333	(6,3) =	2.122222	(6,4) =	2.122222
(7,1) =	1.277778	(7,2) =	1.661111	(7,3) =	1.400000	(7,4) =	1.594444
(8,1) =	1.366667	(8,2) =	1.794444	(8,3) =	1.633333	(8,4) =	1.800000
(9,1) =	2.816667	(9,2) =	3.144444	(9,3) =	3.038889	(9,4) =	2.722222
(10,1) =	2.022222	(10,2) =	2.255556	(10,3) =	2.072222	(10,4) =	2.177778
(11,1) =	2.133333	(11,2) =	2.322222	(11,3) =	2.355556	(11,4) =	2.377778
(12,1) =	1.883333	(12,2) =	2.283333	(12,3) =	2.083333	(12,4) =	2.038889

GENOVA
PAGE 34
G STUDY

VERSION

3.1

P X I X R DESIGN -- RANDOM MODEL

ANOVA TABLE

(** = INFINITE) P I R
SAMPLE SIZE 180 12 4
UNIVERSE SIZE ****

EFFECT	DEGREES OF FREEDOM	SUMS OF SQUARES FOR MEAN SCORES	SUMS OF SQUARES FOR SCORE EFFECTS	MEAN SQUARES	F STATISTIC	(QF = QUASI F RATIO) F-TEST DEGREES OF FREEDOM	NUMERATOR	DENOMINATOR
P	179	53033.68750	7323.48738	40.91334	10.46048 QF	179 QF		2066 QF
I	11	47853.53472	2143.33461	194.84860	35.97062 QF	11 QF		283 QF
R	3	45780.46250	70.26238	23.42079	12.24129 QF	3 QF		38 QF
PI	1969	62621.25000	7444.22789	3.78072	25.79258	1969		5907
PR	537	53252.75000	148.60012	0.27710	1.89038	537		5907
IR	33	47982.62778	58.83067	1.78275	12.16216	33		5907
PIR	5907	63765.00000	865.85683	0.14658				
MEAN		45710.20012						
TOTAL	8639		18054.79988					

NOTE: FOR GENERALIZABILITY ANALYSES, F-STATISTICS SHOULD BE IGNORED

GENOVA
PAGE 35
G STUDY

VERSION

3.1

P X I X R DESIGN -- RANDOM MODEL

G STUDY RESULTS

(** = INFINITE) P I R
SAMPLE SIZE 180 12 4
UNIVERSE SIZE ****

FORM

QFM = QUADRATIC

EFFECT	DEGREES OF FREEDOM	MODEL USING ALGORITHM	VARIANCE USING EMS EQUATIONS	COMPONENTS STANDARD ERROR
P	179	0.7708773	0.7708773	0.0896341
I	11	0.2630996	0.2630996	0.1061489
R	3	0.0099572	0.0099572	0.0068605
PI	1969	0.9085334	0.9085334	0.0301158
PR	537	0.0108761	0.0108761	0.0014244
IR	33	0.0090898	0.0090898	0.0023676
PIR	5907	0.1465815	0.1465815	0.0026967

NOTE: THE "ALGORITHM" AND "EMS" ESTIMATED VARIANCE COMPONENTS WILL BE IDENTICAL IF THERE ARE NO NEGATIVE ESTIMATES

GENOVA VERSION 3.1
PAGE 36
G STUDY

P X I X R DESIGN -- RANDOM MODEL

EXPECTED MEAN SQUARE EQUATIONS

(** = INFINITE) P I R
SAMPLE SIZE 180 12 4
UNIVERSE SIZE ****

EMS(P) = 1.00*VC(PIR) + 12.00*VC(PR) + 4.00*VC(PI) + 48.00*VC(P)
EMS(I) = 1.00*VC(PIR) + 180.00*VC(IR) + 4.00*VC(PI) + 720.00*VC(I)
EMS(R) = 1.00*VC(PIR) + 180.00*VC(IR) + 12.00*VC(PR) + 2160.00*VC(R)
EMS(PI) = 1.00*VC(PIR) + 4.00*VC(PI)
EMS(PR) = 1.00*VC(PIR) + 12.00*VC(PR)
EMS(IR) = 1.00*VC(PIR) + 180.00*VC(IR)
EMS(PIR) = 1.00*VC(PIR)

GENOVA VERSION 3.1
PAGE 37
G STUDY

P X I X R DESIGN -- RANDOM MODEL

VARIANCE - COVARIANCE MATRIX FOR ESTIMATED VARIANCE COMPONENTS (V)

	P	I	R	PI	PR	IR	PIR
P	0.0080343						
I	0.0000004	0.0112676					
R	0.0000000	0.0000001	0.0000471				
PI	-0.0000756	-0.0000050	0.0000000	0.0009070			
PR	-0.0000005	0.0000000	0.0000000	0.0000002	0.0000020		
IR	0.0000000	-0.0000014	-0.0000005	0.0000000	0.0000000	0.0000056	
PIR	0.0000002	0.0000000	0.0000000	-0.0000018	-0.0000006	0.0000000	0.0000073

GENOVA VERSION 3.1
PAGE 38
G STUDY

P X I X R DESIGN -- RANDOM MODEL

CORRELATION MATRIX FOR ESTIMATED VARIANCE COMPONENTS

	P	I	R	PI	PR	IR	PIR
P	1.0000000						
I	0.0000441	1.0000000					
R	0.0000046	0.0001604	1.0000000				
PI	-0.0279989	-0.0015762	-0.0000041	1.0000000			
PR	-0.0039729	-0.0000056	-0.0011535	0.0035318	1.0000000		
IR	-0.0000040	-0.0055761	-0.0287586	0.0001417	0.0009983	1.0000000	
PIR	0.0006268	0.0000353	0.0001820	-0.0223863	-0.1577663	-0.0063279	1.0000000

GENOVA VERSION 3.1
PAGE 39

CONTROL CARD INPUT LISTING

COLUMN 11111111122222222223333333334444444445555555556666666667777777778
1234567890123456789012345678901234567890123456789012345678901234567890
COMMENT FIRST SET OF D STUDY CONTROL CARDS
DSTUDY #1 -- P X I X R DESIGN
DEFFECT \$ P
DEFFECT I
DEFFECT R
ENDDSTUDY

GENOVA VERSION 3.1
PAGE 40
D STUDY

#1 -- P X I X R DESIGN

D STUDY DESIGN NUMBER 001-001

OBJECT OF MEASUREMENT : P FACETS : I R
G STUDY POPULATION SIZE : INFINITE G STUDY UNIVERSE SIZES : INFINITE INFINITE
D STUDY POPULATION SIZE : INFINITE D STUDY UNIVERSE SIZES : INFINITE INFINITE
D STUDY SAMPLE SIZE : 180 D STUDY SAMPLE SIZES : 12 4

EFFECT	VARIANCE COMPONENTS IN TERMS OF G STUDY UNIVERSE (OF ADMISSIBLE OBSERVATIONS) SIZES					VARIANCE COMPONENTS IN TERMS OF D STUDY UNIVERSE (OF GENERALIZATION) SIZES				
	VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE COR- RECTIONS	D STUDY SAMPLING FRE- QUENCIES	VARIANCE COMPONENTS FOR MEAN SCORES		VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS	FINITE UNIVERSE COR- RECTIONS	D STUDY SAMPLING FRE- QUENCIES	VARIANCE COMPONENTS FOR MEAN SCORES	
				ESTIMATES	STANDARD ERRORS				ESTIMATES	STANDARD ERRORS
P	0.77088	1.0000	1	0.77088	0.08963	0.77088	1.0000	1	0.77088	0.08963
I	0.26310	1.0000	12	0.02192	0.00885	0.26310	1.0000	12	0.02192	0.00885
R	0.00996	1.0000	4	0.00249	0.00172	0.00996	1.0000	4	0.00249	0.00172
PI	0.90853	1.0000	12	0.07571	0.00251	0.90853	1.0000	12	0.07571	0.00251
PR	0.01088	1.0000	4	0.00272	0.00036	0.01088	1.0000	4	0.00272	0.00036
IR	0.00909	1.0000	48	0.00019	0.00005	0.00909	1.0000	48	0.00019	0.00005
PIR	0.14658	1.0000	48	0.00305	0.00006	0.14658	1.0000	48	0.00305	0.00006

QFM = QUADRATIC FORM

	VARIANCE	STANDARD DEVIATION	STANDARD ERROR OF VARIANCE	
UNIVERSE SCORE	0.77088	0.87800	0.08963	
EXPECTED OBSERVED SCORE	0.85236	0.92323	0.08960	
LOWER CASE DELTA	0.08148	0.28545	0.00253	
UPPER CASE DELTA	0.10609	0.32571	0.00936	
MEAN	0.02934	0.17129		
				GENERALIZABILITY COEFFICIENT = 0.90440 (9.46048)
				PHI = 0.87903 (7.26642)

NOTE: SIGNAL/NOISE RATIOS ARE IN PARENTHESES

GENOVA VERSION 3.1
PAGE 41
D STUDY

#1 -- P X I X R DESIGN

D STUDY DESIGN NUMBER 001-001

VARIANCE - COVARIANCE MATRIX FOR ESTIMATED VARIANCE COMPONENTS (FOR MEAN SCORES) IN UNIVERSE
OF GENERALIZATION (w)

	P	I	R	PI	PR	IR	PIR
P	0.0080343						
I	0.0000000	0.0000782					
R	0.0000000	0.0000000	0.0000029				
PI	-0.0000063	0.0000000	0.0000000	0.0000063			
PR	-0.0000001	0.0000000	0.0000000	0.0000000	0.0000001		
IR	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	
PIR	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000

GENOVA VERSION 3.1
PAGE 42

CONTROL CARD INPUT LISTING

COLUMN 11111111122222222233333333344444444555555556666666667777777778
1234567890123456789012345678901234567890123456789012345678901234567890

COLUMNS 1-12 BLANK WHEN NOT EXPECTED: CARD IMAGE =

ผลการวิเคราะห์คุณลักษณะคะแนนด้วยโมเดลหลายองค์ประกอบของวราสิทธิ์

เมื่อใช้ผู้ตรวจ 4 คน ตรวจข้อสอบทุกข้อของผู้สอบทุกคน

Facets (Many-Facet Rasch Measurement) Version No. 3.66.1 Copyright ?(c) 1987-2009, John M. Linacre. All rights reserved.
6/1/2554 17:05:36

Scores of the open-ended mathematics test 6/1/2554 17:05:36
Table 1. Specifications from file
"D:\thesis20_12_53\thesis17_3_52\Facet\Facet_score1_1\open_end math test score1_1.txt".

Title = Scores of the open-ended mathematics test 6/1/2554 17:05:36
Data file = (D:\thesis20_12_53\thesis17_3_52\Facet\Facet_score1_1\open_end math test score1_1.txt)
Output file = D:\thesis20_12_53\thesis17_3_52\Facet\Facet_score1_1\open_end math test score1_1.out.txt

; Data specification
Facets = 3
Non-centered = 1
Positive = 1
Labels =
1,examinees ; (elements = 180)
2,raters ; (elements = 4)
3,items ; (elements = 12)
Model = ?B,?B,?,OPEN-ENDED MATHEMATICS,1
Rating (or other) scale = OPEN-ENDED MATHEMATICS,R4,General,Ordinal
; Output description
Arrange tables in order = mN,1N,0fN
Bias/Interaction direction = ability ; leniency, easiness: higher score = positive logit
Fair score = Mean
Pt-biserial = Measure
Heading lines in output data files = Y
Inter-rater coefficients reported for facet = 2
Omit unobserved elements = yes
Scorefile = open-ended mathematicsds.txt
Barchart = Yes
Total score for elements = Yes
T3onscreen show only one line on screen iteration report = Y
T4MAX maximum number of unexpected observations reported in Table 4 = 100
T8NBC show table 8 numbers-barcharts-curves = NBC
Unexpected observations reported if standardized residual >= 2
Usort unexpected observations sort order = (1,2,3),(3,1,2),(Z,3)
Vertical ruler definitions = 1N,3A,1*,2L,2A
WHexact - Wilson-Hilferty standardization = Y
Zscore (Bias size/significance) not less than = 1, 2 ; bias size, t-score

; Convergence control
Convergence = .5, .01
Iterations (maximum) = 0 ; unlimited
Xtreme scores adjusted by = .3, .5 ;(estimation, bias)

Scores of the open-ended mathematics test 6/1/2554 17:05:36
Table 2. Data Summary Report.

Assigning models to "D:\thesis20_12_53\thesis17_3_52\Facet\Facet_score1_1\open_end math test score1_1.txt"
Total lines in data file = 722
Total data lines = 722
Responses matched to model: ?B,?B,?,OPEN-ENDED MATHEMATICS,1 = 8640
Total non-blank responses found = 8640
Number of blank lines (Edit Data=) = 2
Valid responses used for estimation = 8640

Scores of the open-ended mathematics test 6/1/2554 17:05:36
Table 3. Iteration Report.

Iteration	Max. Score Elements	Residual % Categories	Max. Logit Elements	Change Steps
PROX 1			3.8501	
PROX 2			-.2518	
JMLE 3	-198.8799	-9.3	1322.6792	-.6452 -1.4704
JMLE 4	205.9536	4.3	253.6611	-.0884 -.0663
JMLE 5	107.2549	2.1	120.3757	-.0753 -.0247
JMLE 6	55.7002	1.2	64.4302	-.0510 -.0205
JMLE 7	28.2949	.8	34.1714	-.0370 -.0168
JMLE 8	16.6477	.6	18.3838	-.0280 -.0140
JMLE 9	12.1522	.4	11.0931	-.0220 -.0117
JMLE 10	-10.8948	-.4	10.7514	-.0177 -.0098
JMLE 11	-9.5874	-.3	9.7712	-.0144 -.0083
JMLE 12	-8.3054	-.3	8.5856	-.0119 -.0069
JMLE 13	-7.1101	-.2	7.3998	-.0099 -.0058
JMLE 14	-6.0459	-.2	6.3042	-.0082 -.0048
JMLE 15	-5.1108	-.2	5.3325	-.0068 -.0040
JMLE 16	-4.3054	-.1	4.4897	-.0057 -.0034
JMLE 17	-3.6150	-.1	3.7688	-.0047 -.0028
JMLE 18	-3.0325	-.1	3.1570	-.0040 -.0024
JMLE 19	-2.5364	-.1	2.6408	-.0033 -.0020
JMLE 20	-2.1211	-.1	2.2068	-.0028 -.0016
JMLE 21	-1.7737	-.1	1.8429	-.0023 -.0014
JMLE 22	-1.4792	-.1	1.5381	-.0019 -.0011
JMLE 23	-1.2341	.0	1.2831	-.0016 -.0010
JMLE 24	-1.0308	.0	1.0702	-.0013 -.0008
JMLE 25	-.8623	.0	.8925	-.0011 -.0007
JMLE 26	-.7129	.0	.7442	-.0009 -.0006
JMLE 27	-.5977	.0	.6199	-.0008 -.0005
JMLE 28	-.4976	.0	.5168	-.0006 -.0004

Subset connection O.K.

Scores of the open-ended mathematics test 6/1/2554 17:05:36
Table 4. Unexpected Responses - appears after Table 8.

Scores of the open-ended mathematics test 6/1/2554 17:05:36
Table 5. Measurable Data Summary.

Cat	Score	Exp.	Resd	StRes
2.30	2.30	2.30	.00	-.02
1.45	1.45	1.07	.96	1.02
1.45	1.45	1.07	.96	1.03

Data log-likelihood chi-square = 20122.7832

Approximate degrees of freedom = 8443

Chi-square significance prob. = .0000

Responses used for estimation	=	Count	Mean	S.D.	Params
Count of measurable responses	=	8640.00	2.30	1.45	197
Raw-score variance of observations	=	2.09	100.00%		
Variance explained by Rasch measures	=	1.17	56.07%		
Variance of residuals	=	0.92	43.93%		

Scores of the open-ended mathematics test 6/1/2554 17:05:36
Table 6.0 All Facet Vertical "Rulers".

Vertical = (1N,3A,1*,2A,2A,S) Yardstick (columns lines low high extreme)= 0,8,-2,4,End

Measr	+examinees	-items	+examinees	-raters	-raters	OPEN-
4 +						(4)
3 +	173					
	163					
	121 146 178					
	142 143 176					
	53					
2 +						
	55 113 166					
	37 52 61 98 120					
	4 56 67 75 95 107 112 148					
	51 68 81 82 129 169 171					
	6 12 42 59 69 130 134 140 160					
	2 22 70 114 118 119 132 158 168 170					
1 +	1 3 17 24 54 58 71 104					
	16 23 29 30 94 99 111 138 164					
	36 41 73 86 103 139 157 159					
	11 28 80 88 93 150 152					
	5 15 64 74 105 145 154					
	7 155					
	27 65 78 92 126 174					
	25 84 90 135 137 153 161					
0 *	40 83 87 102 125 133 147 151 156 165 180					
	46 57 62 96 101 123 131 162					
	8 9 26 35 66 79 122 124 144 167					
	10 33 43 44 48 77 97 117 128 136 149 175					
	19 32 34 45 72 100 109 172 177					
	13 31 50 60 85 106 127 179					
	14 21 49 63 110 115					
	18 39 89 91 108					
-1 +	47 141					
	38 76 116					
	20					
-2 +						(0)
Measr	+examinees	-items	+examinees	-raters	-raters	OPEN-

S.1: Model = ?B,?B,?,OPEN-ENDED MATHEMATICS

Scores of the open-ended mathematics test 6/1/2554 17:05:36
Table 6.1 examinees Facet Summary.

Logit:

312364699779785245684777785744 1123 1 1
+---Q---+---S---+---M---+---S---+---Q---+
-2 -1 0 1 2 3 4

Infit MnSq:

12221111
246544375448432212
+---Q---S---M---S---Q---+
0 1 2 3 4 5 6

Outfit MnSq:

11222111
130753113687417112 12 1 21 1
+---S---+M---S---Q---+
0 1 2 3 4 5 6

Infit ZStd:

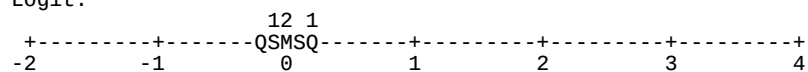
1 1 1
1 1 2 1123213262569807980894173864641 12 311 11
+---+---+Q---+---S---+---M---+---S---+Q---+
-7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9

Outfit ZStd:

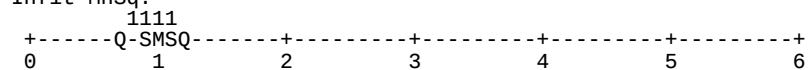
11 1
1 1 2 1242417448249589866907752354 11 121 2 1 1
+---+---+Q---+---S---+---M---+---S---+Q---+
-7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9

Scores of the open-ended mathematics test 6/1/2554 17:05:36
Table 6.2 raters Facet Summary.

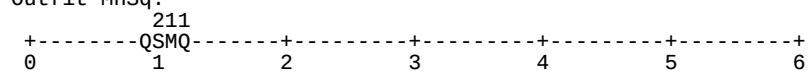
Logit:



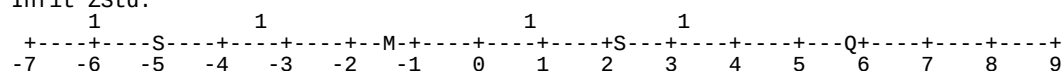
Infit MnSq:



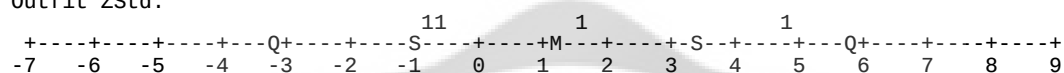
Outfit MnSq:



Infit ZStd:

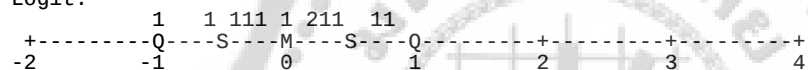


Outfit ZStd:

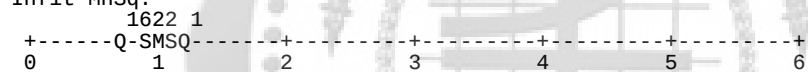


Scores of the open-ended mathematics test 6/1/2554 17:05:36
Table 6.3 items Facet Summary.

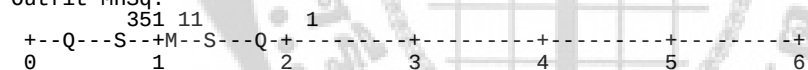
Logit:



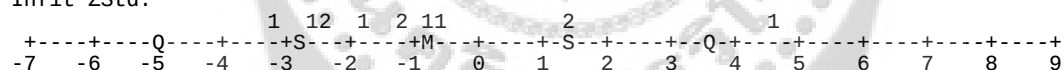
Infit MnSq:



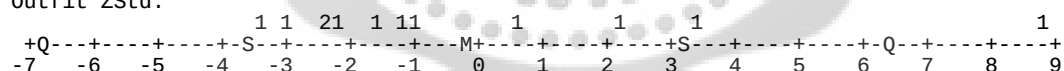
Outfit MnSq:



Infit ZStd:



Outfit ZStd:



Scores of the open-ended mathematics test 6/1/2554 17:05:36
Table 7.1.1 examinees Measurement Report (arranged by mN).

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	ZStd	Outfit MnSq	ZStd	Estim. Discrm	Correlation PtMea PtExp	Num examinees
188	48	3.9	3.93	3.00	.47	.79	-.1	1.20	.5	.97	.05 .14	173 173
186	48	3.9	3.90	2.66	.37	.88	.0	2.68	2.0	.80	-.51 .18	163 163
185	48	3.9	3.88	2.53	.34	1.00	.1	1.41	.8	.93	-.02 .19	121 121
185	48	3.9	3.88	2.53	.34	1.27	.6	1.65	1.1	.93	-.04 .19	146 146
185	48	3.9	3.88	2.53	.34	1.27	.6	2.04	1.5	.91	-.08 .19	178 178
184	48	3.8	3.86	2.42	.32	1.20	.5	1.69	1.1	.90	-.09 .21	143 143
184	48	3.8	3.86	2.42	.32	.90	.0	1.00	.1	.96	.12 .21	176 176
183	48	3.8	3.85	2.33	.30	.84	-.1	1.70	1.2	.90	-.05 .22	142 142
182	48	3.8	3.83	2.24	.28	.74	-.4	.90	.0	.90	.01 .23	53 53
173	48	3.6	3.69	1.75	.20	.89	-.2	1.92	1.9	.73	-.17 .32	113 113
173	48	3.6	3.69	1.75	.20	1.34	1.0	1.24	.6	.92	.13 .32	166 166
172	48	3.6	3.67	1.71	.20	1.03	.2	1.11	.3	1.00	.30 .33	55 55
171	48	3.6	3.65	1.67	.19	1.17	.6	1.06	.2	.94	.22 .33	52 52
170	48	3.5	3.64	1.63	.19	1.06	.2	.68	-.8	1.15	.49 .34	98 98
170	48	3.5	3.64	1.63	.19	.81	-.5	1.97	2.1	.72	-.08 .34	120 120
169	48	3.5	3.62	1.60	.18	.98	.0	.64	-.9	1.31	.62 .35	37 37
169	48	3.5	3.62	1.60	.18	.64	-1.3	.59	-1.1	1.01	.46 .35	61 61
167	48	3.5	3.58	1.53	.18	.75	-.9	.74	-.6	1.05	.44 .36	67 67
166	48	3.5	3.57	1.50	.17	.74	-1.0	.93	.0	1.07	.43 .37	95 95
166	48	3.5	3.57	1.50	.17	1.36	1.3	1.43	1.2	.76	.01 .37	112 112
165	48	3.4	3.55	1.47	.17	1.10	.4	1.67	1.7	.72	.00 .37	4 4
165	48	3.4	3.55	1.47	.17	1.03	.2	1.37	1.0	.97	.38 .37	56 56
165	48	3.4	3.55	1.47	.17	1.57	1.9	1.76	1.9	.84	.08 .37	75 75
165	48	3.4	3.55	1.47	.17	.90	-.3	1.48	1.3	.62	-.12 .37	107 107
164	48	3.4	3.53	1.44	.17	2.00	3.2	5.52	6.9	.49	-.32 .38	148 148
163	48	3.4	3.51	1.41	.17	1.20	.8	1.24	.7	.90	.28 .39	129 129
163	48	3.4	3.51	1.41	.17	.79	-.8	.61	-1.2	1.17	.56 .39	169 169
162	48	3.4	3.49	1.39	.16	1.08	.3	1.17	.5	.58	-.11 .39	81 81
162	48	3.4	3.49	1.39	.16	1.29	1.1	3.14	4.4	.45	-.31 .39	171 171
160	48	3.3	3.46	1.33	.16	.64	-1.7	.49	-1.9	1.34	.71 .40	51 51
160	48	3.3	3.46	1.33	.16	.96	.0	.84	-.4	1.15	.49 .40	68 68
160	48	3.3	3.46	1.33	.16	.84	-.6	.94	.0	.98	.38 .40	82 82
159	48	3.3	3.44	1.31	.16	.69	-1.4	.60	-1.4	1.10	.54 .41	12 12
159	48	3.3	3.44	1.31	.16	.69	-1.4	1.07	.3	.91	.35 .41	134 134
158	48	3.3	3.42	1.28	.16	.76	-1.1	.69	-1.0	1.10	.54 .41	6 6
158	48	3.3	3.42	1.28	.16	1.05	.2	.83	-.5	1.10	.46 .41	130 130
157	48	3.3	3.40	1.26	.15	.97	.0	1.18	.6	1.09	.46 .42	140 140
156	48	3.3	3.38	1.23	.15	.90	-.4	.68	-1.1	1.32	.68 .42	69 69
155	48	3.2	3.36	1.21	.15	.79	-1.0	.70	-1.0	1.23	.62 .42	59 59
154	48	3.2	3.34	1.19	.15	.94	-.2	.93	-.1	1.31	.56 .43	42 42
154	48	3.2	3.34	1.19	.15	.70	-1.5	.63	-1.4	1.26	.62 .43	160 160
153	48	3.2	3.32	1.17	.15	1.83	3.3	2.72	4.3	.39	.04 .43	70 70
153	48	3.2	3.32	1.17	.15	1.42	1.8	1.35	1.2	.97	.32 .43	118 118
153	48	3.2	3.32	1.17	.15	1.05	.3	.87	-.3	1.11	.47 .43	158 158
152	48	3.2	3.30	1.14	.15	1.69	2.8	1.66	2.1	.93	.30 .44	114 114
152	48	3.2	3.30	1.14	.15	1.23	1.1	1.29	1.0	1.03	.39 .44	170 170
151	48	3.1	3.28	1.12	.15	.57	-2.4	.55	-1.9	1.35	.71 .44	119 119
150	48	3.1	3.26	1.10	.15	.87	-.6	.72	-1.0	1.36	.69 .45	132 132
150	48	3.1	3.26	1.10	.15	.64	-2.0	.69	-1.2	.90	.43 .45	168 168
149	48	3.1	3.24	1.08	.14	1.10	.5	1.23	.8	.45	.00 .45	2 2
149	48	3.1	3.24	1.08	.14	.91	-.4	.80	-.7	1.24	.55 .45	22 22
146	48	3.0	3.17	1.02	.14	1.10	.5	1.16	.7	.95	.34 .46	1 1
146	48	3.0	3.17	1.02	.14	.82	-.9	.95	-.1	.95	.45 .46	58 58
145	48	3.0	3.15	1.00	.14	1.13	.7	1.63	2.2	.56	.10 .46	3 3
144	48	3.0	3.13	.98	.14	1.41	2.0	1.48	1.7	.18	-.18 .47	54 54
144	48	3.0	3.13	.98	.14	1.43	2.0	1.25	1.0	.75	.30 .47	71 71
143	48	3.0	3.11	.96	.14	.70	-1.7	.61	-1.8	1.34	.72 .47	17 17
143	48	3.0	3.11	.96	.14	.81	-1.0	.76	-1.0	.91	.44 .47	104 104
142	48	3.0	3.08	.94	.14	1.28	1.4	1.12	.5	.78	.25 .47	24 24
141	48	2.9	3.06	.92	.14	1.09	.5	.92	-.2	1.23	.57 .47	29 29
141	48	2.9	3.06	.92	.14	.84	-.8	1.02	.1	.83	.44 .47	138 138
141	48	2.9	3.06	.92	.14	.96	-.1	1.01	.1	.93	.42 .47	164 164
140	48	2.9	3.04	.90	.14	1.94	4.1	2.47	4.6	.23	-.03 .48	94 94
139	48	2.9	3.02	.89	.14	1.52	2.5	2.82	5.5	.40	.09 .48	30 30
139	48	2.9	3.02	.89	.14	.97	-.1	.92	-.2	1.20	.54 .48	111 111
137	48	2.9	2.97	.85	.13	.90	-.5	.89	-.4	.84	.46 .49	23 23
136	48	2.8	2.95	.83	.13	.84	-.8	.77	-1.0	.88	.44 .49	16 16
135	48	2.8	2.92	.81	.13	1.22	1.2	1.28	1.2	.59	.18 .49	99 99
134	48	2.8	2.90	.80	.13	1.28	1.5	1.09	.4	1.32	.67 .49	103 103
131	48	2.7	2.83	.74	.13	.57	-2.9	.68	-1.6	1.27	.70 .50	36 36
131	48	2.7	2.83	.74	.13	1.30	1.6	1.19	.9	.97	.45 .50	41 41
131	48	2.7	2.83	.74	.13	.90	-.5	.81	-.9	1.53	.69 .50	157 157
131	48	2.7	2.83	.74	.13	.55	-3.1	.52	-2.7	1.64	.78 .50	159 159
130	48	2.7	2.81	.73	.13	1.24	1.3	1.16	.8	.58	.42 .50	86 86
129	48	2.7	2.78	.71	.13	1.23	1.3	1.07	.4	1.15	.53 .50	139 139
128	48	2.7	2.76	.69	.13	1.32	1.7	1.50	2.1	1.29	.49 .51	73 73
127	48	2.6	2.73	.67	.13	.87	-.7	.82	-.9	1.40	.59 .51	150 150
125	48	2.6	2.68	.64	.13	1.70	3.5	2.30	4.9	.17	.12 .51	93 93
124	48	2.6	2.66	.62	.13	.84	-.9	.78	-1.1	1.31	.66 .51	28 28
123	48	2.6	2.64	.61	.13	.79	-1.2	.84	-.8	.77	.43 .51	11 11
123	48	2.6	2.64	.61	.13	1.63	3.2	1.75	3.2	.58	.29 .51	80 80
123	48	2.6	2.64	.61	.13	1.26	1.4	1.16	.8	.80	.57 .51	88 88
121	48	2.5	2.59	.57	.13	1.42	2.2	1.44	2.0	.89	.49 .52	152 152
119	48	2.5	2.54	.54	.13	1.11	.7	1.23	1.1	.80	.39 .52	64 64
119	48	2.5	2.54	.54	.13	1.03	.2	1.14	.7	1.05	.40 .52	145 145
116	48	2.4	2.46	.49	.13	.62	-2.6	.59	-2.4	1.22	.73 .52	105 105
116	48	2.4	2.46	.49	.13	1.13	.8	1.31	1.5	.86	.38 .52	154 154
114	48	2.4	2.41	.46	.13	.38	-4.9	.37	-4.4	1.68	.81 .52	5 5
113	48	2.4	2.38	.44	.13	.74	-1.7	.71	-1.6	1.21	.59 .52	15 15
113	48	2.4	2.38	.44	.13	1.27	1.5	1.43	2.0	.50	.36 .52	74 74
111	48	2.3	2.33	.41	.13	1.18	1.0	1.11	.6	1.06	.44 .53	155 155
110	48	2.3	2.31	.39	.13	.49	-3.7	.46	-3.6	1.77	.77 .53	7 7
105	48	2.2	2.18	.31	.13	.74	-1.6	.75	-1.4	1.27	.52 .53	78 78
102	48	2.1	2.11	.26	.13	.65	-2.2	.66	-2.0	1.56	.85 .53	126 126
101	48	2.1	2.08	.25	.13	.68	-2.0	.64	-2.1	1.39	.67 .53	27 27
100	48	2.1	2.06	.23	.13	1.97	4.4	2.21	4.9	-.04	.17 .53	174 174
99	48	2.1	2.03	.21	.13	.76	-1.4	.70	-1.6	1.48	.75 .53	92 92
98	48	2.0	2.00	.20	.13	.61	-2.5	.58	-2.5	1.22	.69 .53	65 65
96	48	2.0	1.96	.16	.13	.62	-2.5	.59	-2.4	1.52	.69 .53	161 161
94	48	2.0	1.91	.13	.13	.79	-1.2	.83	-.8	.78	.65 .53	84 84
94	48	2.0	1.91	.13	.13	.97	-.1	.88	-.6	1.04	.64 .53	90 90
94	48	2.0	1.91	.13	.13	1.03	.2	.96	-.1	1.39	.65 .53	135 135
94	48	2.0	1.91	.13	.13	.80	-1.1	.76	-1.3	1.47	.79 .53	153 153

93	48	1.9	1.88	.11	.13	1.57	2.8	1.66	2.9	.91	.45	.53	137	137
92	48	1.9	1.86	.10	.13	.85	- .8	.82	- .9	1.30	.55	.53	25	25
90	48	1.9	1.81	.06	.13	.65	-2.1	.64	-2.0	1.16	.76	.53	87	87
90	48	1.9	1.81	.06	.13	1.35	1.8	1.21	1.0	.74	.18	.53	180	180
88	48	1.8	1.76	.03	.13	.78	-1.2	.69	-1.7	1.38	.72	.53	83	83
88	48	1.8	1.76	.03	.13	1.11	.6	1.04	.2	.92	.57	.53	102	102
88	48	1.8	1.76	.03	.13	.51	-3.2	.54	-2.7	1.64	.75	.53	156	156
87	48	1.8	1.74	.01	.13	.65	-2.1	.59	-2.3	1.39	.70	.53	40	40
87	48	1.8	1.74	.01	.13	1.36	1.8	1.27	1.3	.64	.56	.53	133	133
85	48	1.8	1.69	-.03	.13	.25	-5.9	.28	-5.1	1.21	.65	.52	151	151
84	48	1.8	1.67	-.04	.13	.75	-1.4	.70	-1.5	1.36	.70	.52	165	165
83	48	1.7	1.64	-.06	.13	1.10	.5	1.07	.3	.64	.56	.52	125	125
83	48	1.7	1.64	-.06	.13	.72	-1.5	.83	-.8	.91	.38	.52	147	147
82	48	1.7	1.62	-.08	.13	.81	-1.0	.85	-.6	1.12	.59	.52	131	131
82	48	1.7	1.62	-.08	.13	.97	.0	.98	.0	.78	.23	.52	162	162
81	48	1.7	1.60	-.10	.13	1.27	1.4	1.30	1.4	.30	.35	.52	123	123
80	48	1.7	1.57	-.11	.13	.49	-3.3	.48	-3.0	1.34	.67	.52	46	46
80	48	1.7	1.57	-.11	.13	.93	-.3	.84	-.7	1.11	.70	.52	57	57
80	48	1.7	1.57	-.11	.13	.87	-.6	.89	-.4	.83	.65	.52	101	101
79	48	1.6	1.55	-.13	.14	.52	-3.0	.51	-2.7	1.21	.70	.52	96	96
77	48	1.6	1.50	-.17	.14	.77	-1.2	.76	-1.1	1.24	.44	.51	62	62
76	48	1.6	1.48	-.19	.14	.60	-2.3	.66	-1.7	1.51	.62	.51	26	26
76	48	1.6	1.48	-.19	.14	.84	-.8	.81	-.8	1.15	.55	.51	35	35
75	48	1.6	1.46	-.21	.14	.94	-.2	.89	-.4	.81	.50	.51	66	66
75	48	1.6	1.46	-.21	.14	1.07	.3	1.01	.0	.73	.40	.51	122	122
75	48	1.6	1.46	-.21	.14	.67	-1.8	.76	-1.1	1.38	.55	.51	144	144
73	48	1.5	1.42	-.25	.14	.62	-2.1	.61	-2.0	1.14	.69	.51	8	8
72	48	1.5	1.39	-.27	.14	.37	-4.1	.44	-3.1	1.22	.63	.50	79	79
72	48	1.5	1.39	-.27	.14	.45	-3.4	.48	-2.9	1.19	.61	.50	167	167
71	48	1.5	1.37	-.29	.14	1.24	1.1	1.07	.3	.86	.54	.50	9	9
70	48	1.5	1.35	-.31	.14	.73	-1.4	.80	-.8	1.09	.56	.50	124	124
69	48	1.4	1.33	-.33	.14	.68	-1.7	.64	-1.7	1.24	.56	.50	149	149
69	48	1.4	1.33	-.33	.14	.84	-.7	.79	-.9	1.09	.37	.50	175	175
68	48	1.4	1.31	-.35	.14	1.47	2.0	1.35	1.4	.81	.44	.50	33	33
66	48	1.4	1.27	-.39	.14	.42	-3.5	.47	-2.8	1.30	.68	.49	43	43
65	48	1.4	1.25	-.41	.15	.61	-2.1	.60	-1.9	1.39	.59	.49	10	10
65	48	1.4	1.25	-.41	.15	1.82	3.2	2.31	4.2	.25	.07	.49	48	48
65	48	1.4	1.25	-.41	.15	1.11	.5	.95	-.1	1.19	.43	.49	77	77
65	48	1.4	1.25	-.41	.15	1.23	1.1	1.36	1.4	.80	.37	.49	128	128
64	48	1.3	1.23	-.43	.15	1.05	.2	.97	.0	.71	.41	.49	44	44
64	48	1.3	1.23	-.43	.15	.86	-.6	.91	-.3	.67	.40	.49	97	97
64	48	1.3	1.23	-.43	.15	1.11	.5	1.10	.5	.71	.47	.49	117	117
64	48	1.3	1.23	-.43	.15	.69	-1.6	.73	-1.1	1.18	.62	.49	136	136
63	48	1.3	1.21	-.45	.15	1.23	1.0	1.17	.7	.93	.55	.48	19	19
63	48	1.3	1.21	-.45	.15	.59	-2.2	.60	-1.9	1.09	.57	.48	109	109
63	48	1.3	1.21	-.45	.15	.35	-4.1	.36	-3.6	1.45	.73	.48	177	177
62	48	1.3	1.19	-.47	.15	.45	-3.2	.51	-2.5	1.26	.65	.48	34	34
61	48	1.3	1.17	-.50	.15	1.17	.8	1.13	.5	.93	.41	.48	45	45
60	48	1.3	1.15	-.52	.15	.97	.0	.96	.0	.95	.45	.47	32	32
60	48	1.3	1.15	-.52	.15	.71	-1.4	.67	-1.4	1.19	.64	.47	100	100
59	48	1.2	1.13	-.54	.15	.90	-.3	.90	-.3	1.05	.49	.47	72	72
59	48	1.2	1.13	-.54	.15	.86	-.5	.76	-1.0	1.14	.66	.47	172	172
58	48	1.2	1.11	-.56	.15	.76	-1.1	.67	-1.4	1.24	.59	.47	60	60
58	48	1.2	1.11	-.56	.15	.75	-1.1	.73	-1.1	1.34	.52	.47	179	179
57	48	1.2	1.09	-.59	.15	.90	-.3	.91	-.2	.91	.06	.47	50	50
57	48	1.2	1.09	-.59	.15	.98	.0	.90	-.3	.88	.67	.47	85	85
56	48	1.2	1.07	-.61	.16	.78	-.9	.73	-1.1	1.19	.62	.46	13	13
55	48	1.1	1.05	-.64	.16	.90	-.4	.85	-.5	1.08	.61	.46	127	127
54	48	1.1	1.03	-.66	.16	.49	-2.7	.49	-2.4	1.24	.61	.46	106	106
53	48	1.1	1.01	-.69	.16	.86	-.5	.79	-.8	1.14	.61	.45	31	31
52	48	1.1	.99	-.71	.16	.64	-1.6	.66	-1.4	1.08	.46	.45	63	63
52	48	1.1	.99	-.71	.16	1.02	.1	.94	-.1	.96	.61	.45	110	110
50	48	1.0	.95	-.76	.16	1.16	.7	.93	-.1	1.07	.41	.44	14	14
49	48	1.0	.93	-.79	.17	1.46	1.7	1.48	1.7	.65	.16	.44	21	21
49	48	1.0	.93	-.79	.17	1.28	1.1	1.23	.9	.85	.42	.44	49	49
49	48	1.0	.93	-.79	.17	.93	-.1	.74	-1.0	1.26	.57	.44	115	115
48	48	1.0	.91	-.82	.17	.67	-1.4	.69	-1.2	1.06	.36	.43	18	18
48	48	1.0	.91	-.82	.17	1.16	.7	1.18	.7	.86	.49	.43	91	91
46	48	1.0	.88	-.87	.17	.75	-1.0	.73	-1.0	1.06	.52	.42	39	39
46	48	1.0	.88	-.87	.17	.65	-1.5	.68	-1.3	1.11	.32	.42	89	89
46	48	1.0	.88	-.87	.17	1.29	1.1	1.18	.7	.89	.44	.42	108	108
43	48	.9	.82	-.97	.18	.68	-1.3	.72	-1.1	1.17	.36	.41	47	47
43	48	.9	.82	-.97	.18	.74	-1.0	.75	-.9	1.10	.39	.41	141	141
39	48	.8	.75	-1.10	.19	.95	-.1	.97	.0	.95	.42	.39	38	38
37	48	.8	.71	-1.17	.19	1.52	1.7	1.28	1.0	.78	.31	.38	76	76
37	48	.8	.71	-1.17	.19	1.44	1.5	1.13	.5	.92	.49	.38	116	116
36	48	.8	.69	-1.20	.19	1.36	1.2	1.03	.2	.93	.66	.38	20	20
Total Score	Total Count	Obsvd Average	Fair-M Average	Measure	Model S.E.	Infit MnSq	ZStd	Outfit MnSq	ZStd	Estim. Discrm	Correlation PtMea	PtExp	Num examinees	
110.5	48.0	2.3	2.31	.44	.16	.97	-.3	1.05	-.1		.43		Mean (Count: 180)	
44.1	.0	.9	1.01	.91	.05	.33	1.7	.58	1.8		.25		S.D. (Population)	
44.2	.0	.9	1.01	.91	.05	.33	1.7	.59	1.8		.25		S.D. (Sample)	

Model, Populn: RMSE .16 Adj (True) S.D. .89 Separation 5.43 Strata 7.57 Reliability .97
Model, Sample: RMSE .16 Adj (True) S.D. .89 Separation 5.45 Strata 7.60 Reliability .97
Model, Fixed (all same) chi-square: 4534.3 d.f.: 179 significance (probability): .00
Model, Random (normal) chi-square: 168.7 d.f.: 178 significance (probability): .68

Scores of the open-ended mathematics test 6/1/2554 17:05:36
Table 7.2.1 raters Measurement Report (arranged by mN).

Total Score	Total Count	Obsvd Average	Fair-M Average	Measure	Model S.E.	Infit MnSq	ZStd	Outfit MnSq	ZStd	Estim. Discrm	Correlation PtMea	PtExp	Exact Obs %	Agree. Exp %	N raters
4651	2160	2.2	2.14	.16	.02	1.10	3.2	1.20	4.7	.91	.72	.72	61.4	38.4	1 1
5023	2160	2.3	2.42	-.02	.02	.82	-6.0	.97	-6	.97	.72	.72	61.9	39.0	4 4
5063	2160	2.3	2.45	-.04	.02	1.02	.7	1.06	1.5	1.01	.72	.71	67.8	39.1	3 3
5158	2160	2.4	2.52	-.09	.02	.90	-3.4	.97	-7	1.09	.72	.71	68.6	39.1	2 2
4973.8	2160.0	2.3	2.39	.00	.02	.96	-1.4	1.05	1.2		.72		Mean (Count: 4)		
192.7	.0	.1	.14	.09	.00	.11	3.6	.09	2.3		.08		S.D. (Population)		
222.5	.0	.1	.17	.11	.00	.12	4.2	.11	2.6		.08		S.D. (Sample)		
Model, Populn: RMSE .02 Adj (True) S.D. .09 Separation 4.12 Strata 5.82 Reliability (not inter-rater) .94															
Model, Sample: RMSE .02 Adj (True) S.D. .11 Separation 4.79 Strata 6.72 Reliability (not inter-rater) .96															
Model, Fixed (all same) chi-square: 71.9 d.f.: 3 significance (probability): .00															
Model, Random (normal) chi-square: 2.9 d.f.: 2 significance (probability): .23															
Inter-Rater agreement opportunities: 12960 Exact agreements: 8410 = 64.9% Expected: 5041.3 = 38.9%															

Scores of the open-ended mathematics test 6/1/2554 17:05:36
Table 7.3.1 items Measurement Report (arranged by mN).

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	ZStd	Outfit MnSq	ZStd	Estim. Discrm	Correlation PtMea	PtExp	Nu items
1068	720	1.5	1.25	.85	.04	1.07	1.4	.94	-.9	1.30	.64	.67	7 7
1194	720	1.7	1.45	.66	.04	.96	-.7	.92	-1.2	1.08	.66	.69	8 8
1384	720	1.9	1.80	.39	.04	.87	-2.6	.87	-2.2	1.00	.69	.70	4 4
1477	720	2.1	1.99	.26	.04	.88	-2.3	.86	-2.3	1.11	.74	.70	6 6
1494	720	2.1	2.02	.24	.04	.91	-1.7	.80	-3.3	1.38	.77	.70	12 12
1547	720	2.1	2.14	.16	.04	.97	-.5	.90	-1.5	1.32	.79	.70	10 10
1650	720	2.3	2.36	.02	.04	.93	-1.2	1.15	2.2	.49	.60	.69	11 11
1824	720	2.5	2.73	-.23	.04	.94	-1.2	1.04	.6	1.03	.72	.68	2 2
1878	720	2.6	2.83	-.30	.04	.84	-3.2	.80	-2.9	1.00	.73	.67	5 5
1923	720	2.7	2.92	-.37	.04	.88	-2.3	.83	-2.3	1.37	.74	.66	1 1
2109	720	2.9	3.24	-.64	.04	1.26	4.6	1.31	3.4	.74	.55	.62	9 9
2347	720	3.3	3.55	-1.03	.04	1.09	1.4	2.18	8.7	.34	.28	.55	3 3
1657.9	720.0	2.3	2.36	.00	.04	.97	-.7	1.05	-.2		.66		Mean (Count: 12)
357.6	.0	.5	.68	.52	.00	.12	2.1	.37	3.3		.13		S.D. (Population)
373.5	.0	.5	.71	.54	.00	.12	2.2	.39	3.5		.14		S.D. (Sample)

Model, Populn: RMSE .04 Adj (True) S.D. .52 Separation 13.47 Strata 18.30 Reliability .99
Model, Sample: RMSE .04 Adj (True) S.D. .54 Separation 14.08 Strata 19.10 Reliability .99
Model, Fixed (all same) chi-square: 2031.1 d.f.: 11 significance (probability): .00
Model, Random (normal) chi-square: 10.9 d.f.: 10 significance (probability): .36

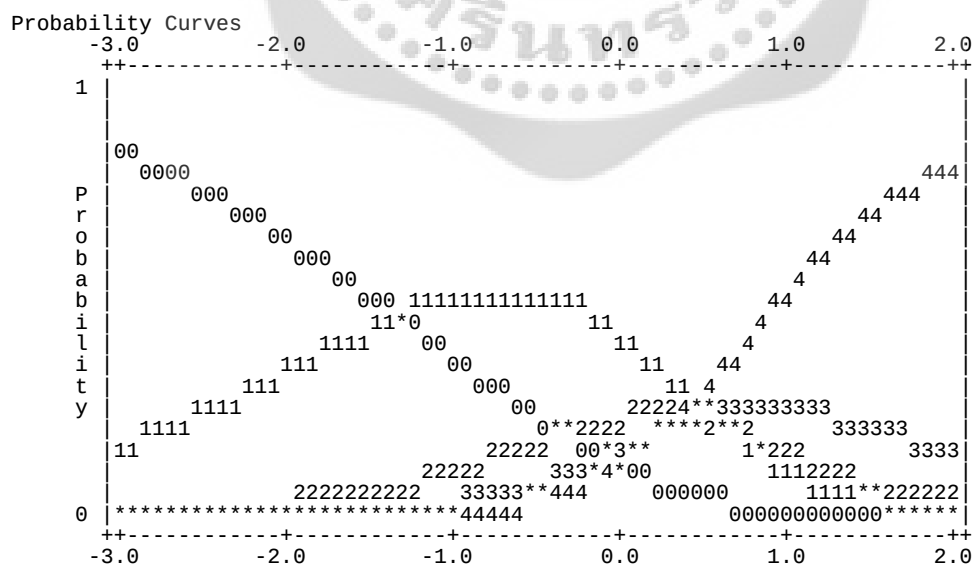
Scores of the open-ended mathematics test 6/1/2554 17:05:36
Table 8.1 Category Statistics.

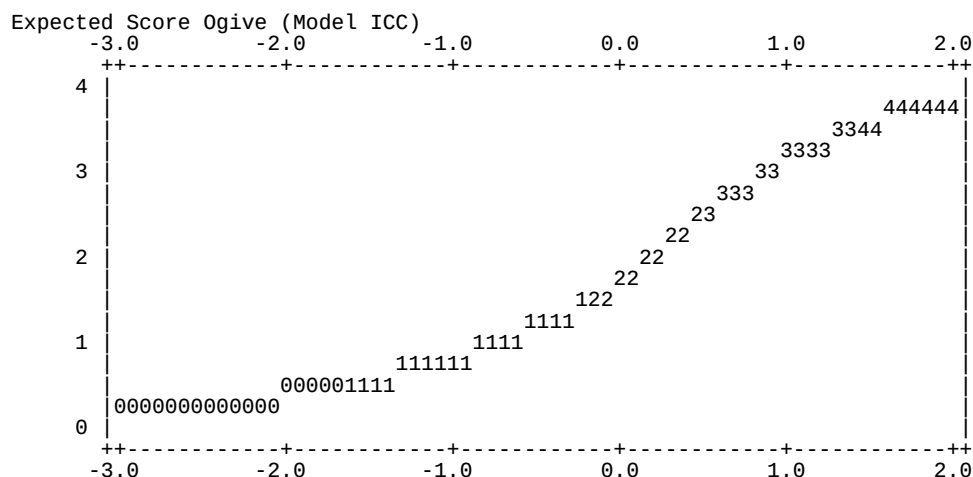
Model = ?B,?B,?,OPEN-ENDED MATHEMATICS
Rating (or partial credit) scale = OPEN-ENDED MATHEMATICS,R4,G,0

DATA Category	Counts Used	%	Cum. %	QUALITY CONTROL Avg Meas	Exp. Meas	OUTFIT MnSq	RASCH-ANDRICH Thresholds Measure	S.E.	EXPECTATION Measure at Category -0.5	MOST PROBABLE from	RASCH-THURSTONE Thresholds	Cat PEAK Prob
0	984	11%	11%	-.60	-.67	1.0	(-2.50)			low	low	100%
1	2340	27%	38%	-.32	-.25	1.0	-1.33	.04	-.70	-1.64	-1.33	52%
2	1223	14%	53%	.27	.23	1.2	.63	.03	.19	-.18	.05	23%
3	1263	15%	67%	.89	.76	1.2	.46	.03	.87	.52	.49	24%
4	2830	33%	100%	1.32	1.36	.9	.24	.03	(1.94)	1.40	.44	100%
(Mean)---(Modal)---(Median)-----												

Scale structure for "OPEN-ENDED MATHEMATICS"

Measr:-3.0 -2.0 -1.0 0.0 1.0 2.0
+ + + + +
Mode:<0----(^)-----01-----^-----14-----(^)-4>
Median:<0----(^)-----01-----^-----12-^--23--^--34-----(^)-4>
Mean:<0----(^)-----01-----^-----12-^--23--^--34-----(^)>
+ + + + +
Measr:-3.0 -2.0 -1.0 0.0 1.0 2.0





Scores of the open-ended mathematics test 6/1/2554 17:05:36
Table 4.1 Unexpected Responses (100 residuals sorted by 1,2,3).

Scores of the open-ended mathematics test 6/1/2554 17:05:36
Table 9.1 Bias Iteration Report.

Bias/Interaction: 1. examinees, 2. raters

There are empirically 720 Bias terms

Iteration	Max. Score Residual Elements % Categories	Max. Logit Change Elements Steps
BIAS 1	5.2629 -84.9	.8785
BIAS 2	1.3180 -17.0	.3040
BIAS 3	.1026 -1.1	.0804
BIAS 4	.0051 .0	.0044

Scores of the open-ended mathematics test 6/1/2554 17:05:36
Table 11.1 Bias/Interaction Measurement Summary.

Bias/Interaction: 1. examinees, 2. raters

Cat	Score	Exp.	Resd	StRes	
2.30	2.30	2.30	.00	-.02	Mean (Count: 8640)
1.45	1.45	1.07	.95	1.02	S.D. (Population)
1.45	1.45	1.07	.95	1.02	S.D. (Sample)

Count of measurable responses = 8640.00
Raw-score variance of observations = 2.09 100.00%
Variance explained by Rasch measures = 1.17 56.07%
Variance of residuals = 0.92 43.93%
Variance explained by bias/interactions = 0.01 0.59%
Variance remaining in residuals = 0.91 43.34%

Scores of the open-ended mathematics test 6/1/2554 17:05:36
Table 12.1 Bias/Interaction Summary Report.

Bias/Interaction: 1. examinees, 2. raters (higher score = higher bias measure)

Bias/Interaction Size:

1	211	122	4526776072549296843521519670683532212111	1	11
-1	-Q	-S	-M	-S	-Q
			0		
			0		
			0		

Bias/Interaction Significance:

1	11333223479578607126746931450586961952915787523233611	1	2	12
-2	-Q	-S	-M	-S
			0	
			0	
			0	

Scores of the open-ended mathematics test 6/1/2554 17:05:36
Table 13.1.1 Bias/Interaction Report (arranged by mN).

Bias/Interaction: 1. examinees, 2. raters (higher score = higher bias measure)

Obsvd Score	Exp. Score	Obsvd Count	Obs-Exp Average	Bias Size	Model S.E.	t	d.f.	Prob.	Infit MnSq	Outfit MnSq	Sq	examinees Num exa	raters measr N r
27.6	27.6	12.0	.00	.00	.32	.00			1.0	1.0	Mean (Count: 720)		
11.2	11.1	.0	.13	.18	.12	.49			.4	.7	S.D. (Population)		
11.2	11.1	.0	.13	.18	.12	.49			.4	.7	S.D. (Sample)		

Fixed (all = 0) chi-square: 172.2 d.f.: 720 significance (probability): 1.00

Scores of the open-ended mathematics test 6/1/2554 17:05:36

Table 13.1.2 Bias/Interaction Report (arranged by 0fN).

Bias/Interaction: 1. examinees, 2. raters (higher score = higher bias measure)

Obsvd Score	Exp. Score	Obsvd Count	Obs-Exp Average	Bias Size	Model S.E.	t	d.f.	Prob.	Infit MnSq	Outfit MnSq	Sq	examinees Num exa	raters measr N r
27.6	27.6	12.0	.00	.00	.32	.00			1.0	1.0	Mean (Count: 720)		
11.2	11.1	.0	.13	.18	.12	.49			.4	.7	S.D. (Population)		
11.2	11.1	.0	.13	.18	.12	.49			.4	.7	S.D. (Sample)		

Fixed (all = 0) chi-square: 172.2 d.f.: 720 significance (probability): 1.00





ภาคผนวก ค

คู่มือการตรวจให้คะแนนแบบทดสอบ
ปลายเปิดวิชาคณิตศาสตร์

คู่มือการตรวจให้คะแนน
แบบทดสอบปลายเปิดวิชาคณิตศาสตร์
ชั้นมัธยมศึกษาตอนต้น (ชั้นมัธยมศึกษาปีที่ 1-3)



คู่มือการตรวจให้คะแนน
แบบทดสอบปลายเปิดวิชาคณิตศาสตร์
ชั้นมัธยมศึกษาตอนต้น (ชั้นมัธยมศึกษาปีที่ 1-3)

คำชี้แจง

1. คู่มือการตรวจให้คะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ ชั้นมัธยมศึกษาตอนต้น (ชั้นมัธยมศึกษาปีที่ 1-3) ประกอบด้วย 3 องค์ประกอบ คือ
 - 1.1 แบบทดสอบปลายเปิดวิชาคณิตศาสตร์ ชั้นมัธยมศึกษาตอนต้น จำนวน 12 ข้อ
 - 1.2 แนวทางการตอบ
 - 1.3 กฎเกณฑ์การให้คะแนน (Scoring rubric)
2. การตอบของผู้สอบอาจไม่ตรงกับแนวทางการตอบที่เสนอไว้ ผู้ตรวจต้องใช้ประสบการณ์การสอนคณิตศาสตร์ร่วมกับกฎเกณฑ์การให้คะแนนที่ให้ไว้ ประกอบการพิจารณาให้คะแนนได้ตามความเหมาะสม
3. การตรวจให้คะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ ชั้นมัธยมศึกษาตอนต้น (ชั้นมัธยมศึกษาปีที่ 1-3) กำหนดให้ตรวจความถูกต้องเหมาะสมของคำตอบในแต่ละข้อโดยใช้แนวทางการตอบประกอบการพิจารณาให้คะแนน โดยใช้กฎเกณฑ์การให้คะแนน (Scoring rubric) ซึ่งเป็นกฎเกณฑ์การให้คะแนนแบบรวม (Holistic scoring method) ที่มีลักษณะเป็นเกณฑ์เฉพาะ (Specific rubric) ของข้อสอบแต่ละข้อ โดยกำหนดเป็น 5 ระดับคะแนน คือ 0, 1, 2, 3 และ 4 คะแนน ตามระดับของความถูกต้องเหมาะสมจากน้อยไปหามาก

1. ราคาปกติของตู้เย็นที่ขาย คือ 8,000 บาท ตะวันซื้อเงินสดทางร้านลดราคาให้ 25 % จากราคาปกติ จากนั้นตะวันนำไปขายให้เพื่อน โดยเพิ่มราคาขึ้นอีก 25 % จากราคาที่ซื้อมา อยากทราบว่า ราคาตู้เย็นที่ตะวันขายให้เพื่อนจะน้อยกว่า มากกว่า หรือ เท่ากับ 8,000 บาท จงอธิบาย

แนวการตอบ

ราคาตู้เย็นที่ตะวันขายให้เพื่อนน้อยกว่า 8,000 บาท คำนวณได้ดังนี้

ราคาปกติของตู้เย็น คือ 8,000 บาท ตะวันซื้อเงินสดทางร้านลดราคาให้ 25% จากราคาปกติ ลดราคาให้ 25 % จากราคาปกติ แสดงว่า

ตั้งราคาขายไว้ 100 บาท ขายจริง 75 บาท

$$\text{ตั้งราคาขายไว้ } 8,000 \text{ บาท ขายจริง } \frac{8,000 \times 75}{100} = 6,000 \text{ บาท}$$

(นักเรียนอาจหาราคาที่ลดก่อน แล้วนำมาลบออกจากราคาปกติที่ขาย จะเป็นราคาขายจริง)

แสดงว่าตะวันซื้อตู้เย็นในราคา 6,000 บาท

จากนั้นตะวันนำไปขายให้เพื่อน โดยปรับราคาขึ้นอีก 25 % จากราคาที่ซื้อมา

ปรับราคาเพิ่มขึ้น 25 % จากราคาที่ซื้อมา แสดงว่า

ถ้าซื้อมาราคา 100 บาท ราคาขาย เป็น 125 บาท

$$\text{ถ้าซื้อมาราคา } 6,000 \text{ บาท ราคาขาย เป็น } \frac{6,000 \times 125}{100} = 7,500 \text{ บาท}$$

(นักเรียนอาจหาราคาที่เพิ่มขึ้นก่อน แล้วนำมาบวกกับราคาที่ซื้อมา จะเป็นราคาขาย)

ดังนั้น ตะวันขายตู้เย็นในราคา 7,500 บาท ซึ่งมีค่าน้อยกว่า 8,000 บาท อยู่ 500 บาท

เกณฑ์การให้คะแนน

ระดับคะแนน	รายละเอียดของเกณฑ์
0	ไม่ตอบ หรือคำตอบไม่ถูกต้อง โดยไม่เขียนคำอธิบายหรือปรากฏร่องรอยการคิดใด ๆ ในการแก้ปัญหา
1	-คำตอบไม่ถูกต้อง แต่ปรากฏร่องรอยการคิดหรือเขียนคำอธิบาย โดยใช้หลักทางคณิตศาสตร์ที่ไม่ถูกต้อง เช่น เข้าใจว่าการลดราคา 25 % จากราคาปกติ มีค่าเท่ากับการเพิ่มราคาขึ้นอีก 25 % จากราคาที่ซื้อมา -แสดงวิธีคิดที่แสดงถึงความเข้าใจปัญหาได้เล็กน้อย แต่ไม่สามารถดำเนินการแก้ปัญหาได้ ไม่ได้คำตอบ
2	-คำตอบไม่ถูกต้อง แต่ได้แสดงวิธีแก้ปัญหาโดยใช้หลักการทางคณิตศาสตร์ได้ถูกต้องบางส่วน เช่น แสดงวิธีหารราคาซื้อมาได้ แต่หารราคาขายใหม่ไม่ถูกต้อง -แสดงวิธีแก้ปัญหาโดยใช้หลักการทางคณิตศาสตร์ได้ถูกต้อง แต่คำนวณผิดพลาด -คำตอบถูกต้อง แต่ไม่แสดงแนวคิดวิธีการแก้ปัญหา
3	แสดงวิธีการหารราคาซื้อและราคาใหม่ที่ขายโดยใช้หลักการทางคณิตศาสตร์ได้อย่างถูกต้อง คำนวณผิดพลาดเล็กน้อย
4	แสดงวิธีการหารราคาซื้อและราคาใหม่ที่ขายโดยใช้หลักการทางคณิตศาสตร์ได้อย่างถูกต้อง สมบูรณ์ คำนวณถูกต้องและคำตอบถูกต้อง

2. เตี้ยมีลวดเส้นหนึ่งยาว $7\frac{4}{5}$ เมตร และต้องการตัดแบ่งเป็นลวดสามเส้น ซึ่งยาว $1\frac{1}{2}$ เมตร, $2\frac{3}{4}$

เมตร และ $3\frac{3}{5}$ เมตร เตี้ยจะสามารถตัดให้ได้ขนาดตามต้องการหรือไม่ จงอธิบาย

แนวการตอบ

วิธีที่ 1 ใช้การหาผลรวมของลวดที่ต้องการทั้ง 3 เส้น แล้วเปรียบเทียบกับลวดที่มีอยู่ (ใช้เศษส่วน)

เตี้ยต้องการใช้ลวด 3 เส้นที่มีความยาวเป็น $1\frac{1}{2}$ เมตร $2\frac{3}{4}$ เมตร และ $3\frac{3}{5}$ เมตร

สามารถหาผลบวกของลวดทั้ง 3 เส้น ได้เป็น

$$\begin{aligned} 1\frac{1}{2} + 2\frac{3}{4} + 3\frac{3}{5} &= \frac{3}{2} + \frac{11}{4} + \frac{18}{5} \\ &= \frac{30 + 55 + 72}{20} \end{aligned}$$

$$= \frac{157}{20} = 7\frac{17}{20}$$

ความยาวของลวดทั้งหมดที่เตี้ยต้องการคือ $7\frac{17}{20}$

เตี้ยมีลวดยาว $7\frac{4}{5}$ เมตร ซึ่ง $7\frac{4}{5} = \frac{39}{5} = \frac{156}{20} = 7\frac{16}{20}$ เมตร

เมื่อทำการเปรียบเทียบค่าของ $7\frac{4}{5}$ กับ $7\frac{17}{20}$ จะเห็นว่า $7\frac{16}{20}$ น้อยกว่า $7\frac{17}{20}$

(การเปรียบเทียบนักเรียนอาจใช้หลักการคูณไขว้ได้)

แสดงว่า เตี้ยไม่สามารถนำลวดที่มีอยู่มาตัดให้ได้ขนาดตามต้องการทั้งสามเส้นได้ เพราะลวดที่มีอยู่มีความยาวน้อยกว่าผลรวมความยาวลวดที่ต้องการ

วิธีที่ 2 ใช้การหาผลรวมของลวดที่ต้องการทั้ง 3 เส้น แล้วเปรียบเทียบกับลวดที่มีอยู่ (ใช้ทศนิยม)

เต็ยมีลวดยาว $7\frac{4}{5}$ เมตร หรือ 7.80 เมตร

เต็ยต้องการใช้ลวดสามเส้นที่มีความยาวเป็น $1\frac{1}{2}$ เมตร หรือ 1.50 เมตร

$2\frac{3}{4}$ เมตร หรือ 2.75 เมตร และ

$3\frac{3}{5}$ เมตร หรือ 3.60 เมตร ตามลำดับ

ความยาวของลวดทั้งหมดที่เต็ยต้องการคือ $1.50 + 2.75 + 3.60 = 7.85$ เมตร

แต่เต็ยมีลวดอยู่ 7.80 เมตร ซึ่งน้อยกว่าความยาวลวดที่เต็ยต้องการ

ดังนั้นเต็ยไม่สามารถนำลวดที่มีอยู่มาตัดให้ได้ขนาดตามต้องการทั้งสามเส้นได้

วิธีที่ 3 ใช้การตัดแบ่งลวดทีละเส้นออกจากความยาวลวดที่มีอยู่

เต็ยมีลวดยาว $7\frac{4}{5}$ เมตร หรือ $\frac{39}{5}$ เมตร

ตัดแบ่งเป็นลวดเส้นแรกยาว $1\frac{1}{2}$ เมตร เหลือลวดยาว $\frac{39}{5} - \frac{3}{2} = \frac{78 - 15}{10} = \frac{63}{10}$ เมตร

ตัดแบ่งเป็นลวดเส้นที่สองยาว $2\frac{3}{4}$ เมตร

เหลือลวดยาว $\frac{63}{10} - \frac{11}{4} = \frac{126 - 55}{20} = \frac{71}{20}$ เมตร

ต้องการลวดเส้นที่สามยาว $3\frac{3}{5}$ หรือ $\frac{18}{5}$ เมตร

เปรียบเทียบความยาวลวดที่เหลือคือ $\frac{71}{20}$ เมตร กับความยาวลวดที่ต้องการคือ $\frac{18}{5}$ เมตร ได้เป็น

$$\frac{18}{5} = \frac{18 \times 4}{5 \times 4} = \frac{72}{20} \text{ ซึ่งมากกว่า } \frac{71}{20}$$

จะเห็นว่าลวดที่เหลือมีความยาวน้อยกว่าลวดเส้นที่สามที่ต้องการ

แสดงว่า เต็ยไม่สามารถนำลวดที่มีอยู่มาตัดให้ได้ขนาดตามต้องการทั้งสามเส้นได้ เพราะลวดที่มีอยู่มีความยาวน้อยกว่าผลรวมความยาวลวดที่ต้องการ

เกณฑ์การให้คะแนน

ระดับคะแนน	รายละเอียดของเกณฑ์
0	ไม่ตอบ หรือตอบไม่ถูกต้อง โดยไม่แสดงวิธีการแก้ปัญหาใด ๆ
1	- แสดงวิธีการแก้ปัญหาไม่ถูกต้อง ไม่ได้คำตอบหรือคำตอบไม่ถูกต้อง - แสดงวิธีการแก้ปัญหาได้เล็กน้อย ไม่สมบูรณ์ แต่ไม่ได้คำตอบหรือคำตอบไม่ถูกต้อง
2	- คำตอบถูกต้อง แต่ไม่แสดงวิธีการแก้ปัญหาหรือแสดงวิธีแก้ปัญหาไม่ชัดเจน - แสดงวิธีการแก้ปัญหาได้ถูกต้องบางส่วน คำนวณผิดพลาด (หาผลบวกหรือผลลบของเศษส่วนไม่ถูกต้อง หรือ เปรียบเทียบไม่ถูกต้อง) คำตอบไม่ถูกต้อง
3	แสดงวิธีแก้ปัญหาได้ถูกต้อง คำนวณผิดพลาดเล็กน้อย หรือไม่แสดงการเปรียบเทียบ คำตอบถูกต้อง
4	แสดงวิธีแก้ปัญหาได้ถูกต้อง สมบูรณ์ทุกขั้นตอน คำตอบถูกต้อง

3. จงพิจารณาสถานการณ์แต่ละข้อต่อไปนี้ว่า ใช้การประมาณค่าได้เหมาะสม หรือไม่ เพราะเหตุใด

1. โดยปกติการอบขนมปังชนิดหนึ่ง จะใช้อุณหภูมิ 330 องศาฟาเรนไฮต์ เป็นเวลา 22 นาที เพื่อเป็นการประหยัดพลังงานและเวลา หมักแป้งจึงประมาณค่าอุณหภูมิเป็น 300 องศาฟาเรนไฮต์ และประมาณเวลาเป็น 20 นาที นักเรียนเห็นด้วยการประมาณค่าของหมักแป้งหรือไม่ เพราะเหตุใด

2. การขับรถจากกรุงเทพมหานคร ถึงจังหวัดน่านของรถโดยสาร ใช้เวลา 9 ชั่วโมง 45 นาที แต่พนักงานบนรถประกาศให้ผู้โดยสารทราบว่า จะใช้เวลาเดินทางประมาณ 10 ชั่วโมง นักเรียนเห็นด้วยกับการประมาณค่าของพนักงานบนรถหรือไม่ เพราะเหตุใด

3. เอะเจต้องใช้น็อตขนาดเส้นผ่านศูนย์กลาง 0.8 ซม. ในการซ่อมจักรยาน แต่เมื่อไปซื้อที่ร้านปรากฏว่าน็อตขนาด 0.8 ซม. หหมด เอะเจจึงซื้อน็อต ขนาด 1 ซม. มาแทน เพราะคิดสามารถใช้แทนได้ นักเรียนเห็นด้วยกับการประมาณค่าของเอะเจหรือไม่ เพราะเหตุใด

4. ในการผลิตยาชนิดหนึ่ง 1 เม็ด ต้องผสมสาร A 1.25 กรัม และสาร B 0.75 กรัม เพื่อความสะดวก บริษัทผลิตยา จึงผสมสาร A โดยประมาณเป็น 1 กรัม และสาร B ประมาณเป็น 1 กรัม นักเรียนเห็นด้วยกับการประมาณค่าของบริษัทผลิตยาหรือไม่ เพราะเหตุใด

แนวการตอบ

- ข้อ 1 ไม่เห็นด้วย เนื่องจาก หากใช้อุณหภูมิและเวลาที่ไม่พอดี จะทำให้ขนมปังที่ได้ไม่สุก ไม่ฟู
- ข้อ 2 เห็นด้วย เนื่องจาก มีความเป็นไปได้ว่าในการخبร้อาจมากกว่า 9 ชั่วโมง 45 นาที เพราะอาจมีการแวะพักระหว่างทาง การจราจรติดขัด เป็นต้น จึงเป็นการประมาณค่าที่เหมาะสม
- ข้อ 3 ไม่เห็นด้วย เนื่องจาก เขาต้องใช้เนื้อที่มีเส้นผ่านศูนย์กลางขนาด 0.8 ซม. ในการซ่อมจักรยาน หากใช้เนื้อ 1 ซม. จะไม่สามารถใช้ซ่อมจักรยานได้ เพราะมีขนาดที่ไม่พอดีคือ หลวมไป
- ข้อ 4 ไม่เห็นด้วย เนื่องจาก ในการผสมยา หากใช้ส่วนผสมที่ไม่ตรงตามสัดส่วนที่กำหนดไว้ จะทำให้ยาชนิดนี้มีคุณภาพผิดพลาดไป ไม่ได้มาตรฐาน อาจเกิดผลเสียหรือเป็นอันตรายกับผู้ใช้ได้
- หมายเหตุ** สามารถอธิบายเหตุผลได้หลายแบบ

เกณฑ์การให้คะแนน

ระดับคะแนน	รายละเอียดของเกณฑ์
0	ไม่ตอบ หรือ ตอบไม่ถูกต้องแม้แต่ข้อเดียว
1	คำตอบถูกต้องและแสดงเหตุผลประกอบถูกต้อง 1 ข้อ
2	คำตอบถูกต้องและแสดงเหตุผลประกอบถูกต้อง 2 ข้อ
3	คำตอบถูกต้องและแสดงเหตุผลประกอบถูกต้อง 3 ข้อ
4	คำตอบถูกต้องและแสดงเหตุผลประกอบถูกต้องทั้ง 4 ข้อ

4. เชือกสามเส้น ยาว 45, 60 และ 90 เมตร ต้องการตัดเชือกทั้งสามเส้นนี้ให้เป็นเส้นสั้น ๆ ที่มีความยาวเป็นจำนวนเต็มเมตรและยาวเท่า ๆ กันทุกเส้น โดยที่ตัดแล้วจะต้องไม่มีเศษเชือกเหลือ จงอธิบายว่าเชือกเส้นสั้น ๆ แต่ละเส้นที่ตัดแล้วจะมีความยาวเป็นเท่าใดได้บ้าง บอกมาให้ครบถ้วน

แนวการตอบ

ในการแบ่งเชือกที่มีความยาว 45, 60 และ 90 เมตร ให้เป็นเส้นสั้น ๆ ที่ยาวเท่า ๆ กัน โดยไม่ให้มีเศษเหลือและมีความยาวเป็นจำนวนเต็มเมตร ต้องหาจำนวนเต็มที่หารทั้ง 45, 60 และ 90 ลงตัว (หรือหาตัวประกอบร่วม 45, 60, และ 90) ซึ่งได้แก่ 1, 3, 5, และ 15

แสดงว่า ความยาวเชือกเส้นสั้น ๆ จะยาว 1, 3, 5, และ 15 เมตร และสามารถตัดได้เป็น

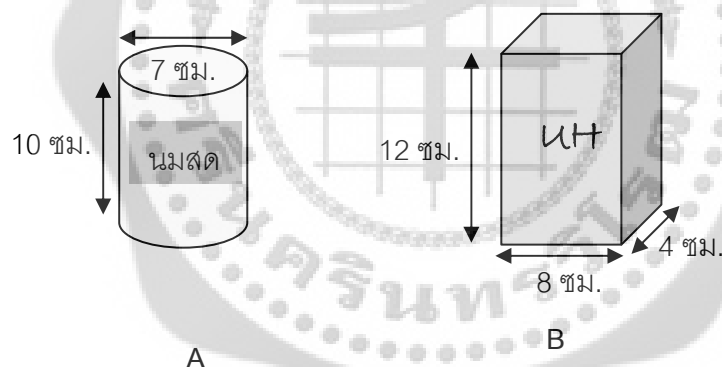
- แบ่งเชือกให้ยาวเส้นละ 1 เมตร ได้เชือก $45+60+90 = 195$ เส้น
- แบ่งเชือกให้ยาวเส้นละ 3 เมตร ได้เชือก $15+20+30 = 65$ เส้น
- แบ่งเชือกให้ยาวเส้นละ 5 เมตร ได้เชือก $9+12+18 = 39$ เส้น
- แบ่งเชือกให้ยาวเส้นละ 15 เมตร ได้เชือก $3+4+6 = 13$ เส้น

หมายเหตุ สามารถหา ห.ร.ม. ของ 45, 60 และ 90 ซึ่งคือ 15 แล้วจึงหาตัวประกอบของ 15 ได้เป็น 1 เมตร, 3 เมตร, 5 เมตร และ 15 เมตร

เกณฑ์การให้คะแนน

ระดับคะแนน	รายละเอียดของเกณฑ์
0	ไม่ตอบหรือตอบไม่ถูกต้อง โดยไม่แสดงวิธีแก้ปัญหาใด ๆ หรือแสดงแนวคิดที่ไม่ถูกต้อง
1	แสดงวิธีการแบ่งเชือกได้ถูกต้อง 1 วิธี
2	แสดงวิธีการแบ่งเชือกได้ถูกต้อง 2 วิธี
3	แสดงวิธีการแบ่งเชือกได้ถูกต้อง 3 วิธี
4	แสดงวิธีการแบ่งเชือกได้ถูกต้องทั้ง 4 วิธี

5. นมสดบรรจุอยู่ในกระป๋องสูง 10 ซม. เส้นผ่านศูนย์กลางยาว 7 ซม. (ดังรูป A) และนม UHT บรรจุอยู่ในกล่องทรงสี่เหลี่ยมมุมฉาก กว้าง 4 ซม. ยาว 8 ซม. และ สูง 12 ซม. (ดังรูป B)



นมชนิดใดมีปริมาตรมากกว่ากัน จงอธิบาย

แนวการตอบ

1. หาปริมาตรของนมสด

นมสดที่บรรจุในกระป๋องสูง 10 ซม. เส้นผ่านศูนย์กลางยาว 7 ซม.

จากปริมาตรทรงกระบอก = $\pi r^2 h$ (เมื่อ r คือรัศมี และ h คือ ความสูง)

นมสดที่บรรจุในกระป๋อง มีปริมาตร = $\pi(3.5)^2(10) = 122.5\pi$ ลบ.ซม.

หรือหากแทนค่า π ด้วย 3.14 จะได้เป็น $122.5 \times 3.14 = 384.65$ ลบ.ซม.

หากแทนค่า π ด้วย $\frac{22}{7}$ จะได้เป็น $122.5 \times \frac{22}{7} = 385$ ลบ.ซม.

2. หาปริมาตรของนม UHT

นม UHT บรรจุอยู่ในกล่องทรงสี่เหลี่ยมมุมฉาก กว้าง 4 ซม. ยาว 8 ซม. และ สูง 12 ซม.

จากปริมาตรของทรงสี่เหลี่ยมมุมฉาก = ความกว้าง×ความยาว × ความสูง (หรือ พื้นฐาน×สูง)

$$= 4 \times 8 \times 12$$

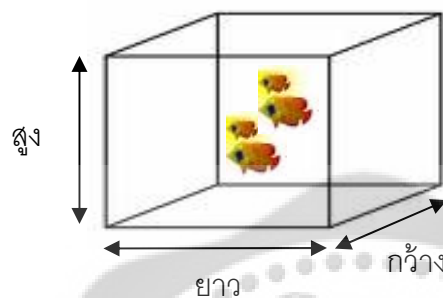
$$= 384 \text{ ลบ.ซม.}$$

ดังนั้น นมสด มีปริมาตรมากกว่า นม UHT

เกณฑ์การให้คะแนน

ระดับคะแนน	รายละเอียดของเกณฑ์
0	ไม่ตอบ หรือตอบไม่ถูกต้องโดยไม่เขียนแนวคิด วิธีการใด ๆ ในการแก้ปัญหา
1	- มีร่องรอยของการแก้ปัญหาเล็กน้อย เช่น เขียนสูตรการหาปริมาตรของรูปทรงกระบอก รูปทรงสี่เหลี่ยมมุมฉาก แต่แสดงวิธีแก้ปัญหา ไม่ได้คำตอบ - ใช้วิธีการคำนวณปริมาตรทรงกระบอกและทรงสี่เหลี่ยมมุมฉากไม่ถูกต้อง - คำตอบถูกต้อง แต่ไม่แสดงวิธีการแก้ปัญหาหรือวิธีการแก้ปัญหาไม่ชัดเจน
2	- แสดงวิธีการหาปริมาตรของทรงกระบอกและทรงสี่เหลี่ยมมุมฉากได้ถูกต้อง แต่คำนวณหาปริมาตรของรูปทรงทั้งสองไม่ถูกต้อง - แสดงวิธีการหาปริมาตรของทรงกระบอกหรือทรงสี่เหลี่ยมมุมฉากได้ถูกต้องหนึ่งรูปทรง และคำนวณหาปริมาตรของรูปทรงถูกต้องหนึ่งรูปทรง
3	แสดงวิธีการหาปริมาตรของทรงกระบอกและทรงสี่เหลี่ยมมุมฉากได้ถูกต้อง แต่คำนวณหาปริมาตรได้ถูกต้องเพียงหนึ่งรูปทรง
4	แสดงวิธีการหาปริมาตรของทรงกระบอกและทรงสี่เหลี่ยมมุมฉากได้ถูกต้อง สมบูรณ์ คำนวณถูกต้องทั้งสองรูปทรง สรุปคำตอบถูกต้อง

6. ฝ้าย ต้องการขนย้ายตู้เลี้ยงปลารูปทรงสี่เหลี่ยมมุมฉากที่มีปริมาตร 5,819 ลบ.นิ้ว กว้าง 11 นิ้ว และสูง 23 นิ้ว โดยได้เตรียมกล่องกว้าง 14 นิ้ว ยาว 26 นิ้ว และสูง 26 นิ้ว สำหรับใส่ตู้เลี้ยงปลา ทั้งนี้เพื่อป้องกันความเสียหายจากการขนย้าย ฝ้ายจึงใส่โฟมกันกระแทกหนา 1.5 นิ้ว ทั้งด้านข้าง ด้านบนและด้านล่างของกล่อง อยากทราบว่า ฝ้ายจะสามารถใส่ตู้เลี้ยงปลาลงในกล่องนี้ได้พอดีหรือไม่ จงอธิบาย



แนวการตอบ

ตู้ปลาที่มีปริมาตร 5,819 ลบ.นิ้ว กว้าง 11 นิ้ว และ สูง 23 นิ้ว

ให้ ตู้เลี้ยงปลา ยาว a นิ้ว

จากปริมาตรของทรงสี่เหลี่ยมมุมฉาก = ความกว้าง $\times$ ความยาว $\times$ ความสูง

$$\begin{array}{rcl} \text{จะได้} & 5,819 & = 11 \times a \times 23 \\ & a & = 23 \end{array}$$

ดังนั้น ตู้เลี้ยงปลาที่มีความยาว 23 นิ้ว

นั่นคือ ตู้เลี้ยงปลาที่มีขนาด กว้าง 11 นิ้ว ยาว 23 นิ้ว และสูง 23 นิ้ว

กล่อง กว้าง 14 นิ้ว ยาว 26 นิ้ว และสูง 26 นิ้ว ต้องการใส่โฟมกัน กระแทกหนาแต่ละด้านละ 1.5 ซม.

ดังนั้น กล่องที่เหลือจะมีขนาด กว้าง $14 - 3 = 11$ นิ้ว ยาว $26 - 3 = 23$ นิ้ว และสูง $26 - 3 = 23$ นิ้ว

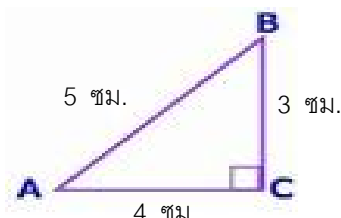
ดังนั้น ฝ้าย สามารถใส่ตู้เลี้ยงปลาลงในกล่องได้พอดี

หมายเหตุ สามารถหาผลบวกของขนาดตู้ปลากับโฟมกันกระแทก จะได้เป็นขนาดกล่อง โดยจะกว้าง $11 + 3 = 14$ นิ้ว ยาว $23 + 3 = 26$ นิ้ว และ สูง $23 + 3 = 26$ นิ้ว ซึ่งเท่ากับขนาดของกล่องที่มีอยู่

เกณฑ์การให้คะแนน

ระดับคะแนน	รายละเอียดของเกณฑ์
0	ไม่มีคำตอบ คำตอบไม่ถูกต้อง โดยไม่ได้แสดงแนวคิดในการหาคำตอบ
1	-แสดงร่องรอยในการแก้ปัญหาได้เล็กน้อย เช่น วาดรูป เขียนสูตรการหาปริมาตร แต่ไม่ได้คำตอบหรือคำตอบไม่ถูกต้อง -แสดงวิธีการแก้ปัญหาไม่ถูกต้อง เช่น ใช้การเปรียบเทียบค่าปริมาตร -คำตอบถูกต้อง แต่ไม่แสดงวิธีการแก้ปัญหา
2	-แสดงวิธีการแก้ปัญหาได้ถูกต้องบางส่วน เช่น แสดงวิธีหาความยาวตู้ปลาได้ แต่แสดงวิธีการแก้ปัญหากล่องใส่โฟมกันกระแทกไม่ถูกต้อง -แสดงวิธีการแก้ปัญหาถูกต้อง แต่ผิดพลาดในการคำนวณค่อนข้างมาก เช่น คำนวณหาความยาวตู้ปลาและขนาดของกล่องเมื่อใส่โฟมกันกระแทกไม่ถูกต้อง -คำตอบถูกต้อง แต่แสดงวิธีการแก้ปัญหาไม่สมบูรณ์ ชัดเจน
3	แสดงวิธีการแก้ปัญหาถูกต้อง แต่คำนวณผิดพลาด เช่น -คำนวณความยาวของตู้ปลาถูกต้อง แต่คำนวณขนาดของกล่องเมื่อใส่โฟมกันกระแทกผิดพลาด เช่น หักความหนาของโฟมออกเพียงด้านเดียว -คำนวณความยาวของตู้ปลาไม่ถูกต้อง แต่หาขนาดของกล่องเมื่อใส่โฟมถูกต้อง
4	คำตอบถูกต้อง แสดงวิธีการในการแก้ปัญหาได้ถูกต้อง ชัดเจน สมบูรณ์

7. ครูให้นักเรียนวาดรูปสามเหลี่ยมมุมฉากที่มีด้านตรงข้ามมุมฉากยาว 5 ซม. ด.ช. ต้น วาดรูป โดยให้ด้านตรงข้ามมุมฉากยาว 5 ซม. และด้านประกอบมุมฉากยาว 3 ซม. และ 4 ซม. ดังรูป



แต่ ด.ช. ใจ บอกครูว่า เขาสามารถวาดรูปสามเหลี่ยมมุมฉากรูปอื่น ที่มีด้านตรงข้ามมุมฉากยาว 5 ซม. แต่ด้านประกอบมุมฉากไม่ได้ยาว 3 และ 4 ซม. ได้ นักเรียนคิดว่า ด.ช. ใจ กล่าวถูกต้องหรือไม่ จงอธิบายและยกตัวอย่างประกอบ

แนวการตอบ

ด.ช. ใจ กล่าวถูกต้อง เพราะรูปสามเหลี่ยมมุมฉากที่มีด้านตรงข้ามมุมฉากยาว 5 ซม. ไม่จำเป็นต้องด้านประกอบมุมฉากต้องยาว 3 และ 4 ซม. เสมอไป

ทั้งนี้สามารถกำหนดความยาวของด้านประกอบมุมฉาก 1 ค่า แล้วใช้ทฤษฎีพีทาโกรัส คำนวณหาความยาวของด้านประกอบมุมฉากอีกด้านได้ ตัวอย่างเช่น

ด้านตรงข้ามมุมฉากยาว 5 ซม. ให้ด้านประกอบมุมฉากด้านหนึ่งยาว 1 ซม. อีกด้านยาว a ซม.

$$\text{จะได้} \quad 5^2 = 1^2 + a^2$$

$$25 = 1 + a^2$$

$$a = \sqrt{24}$$

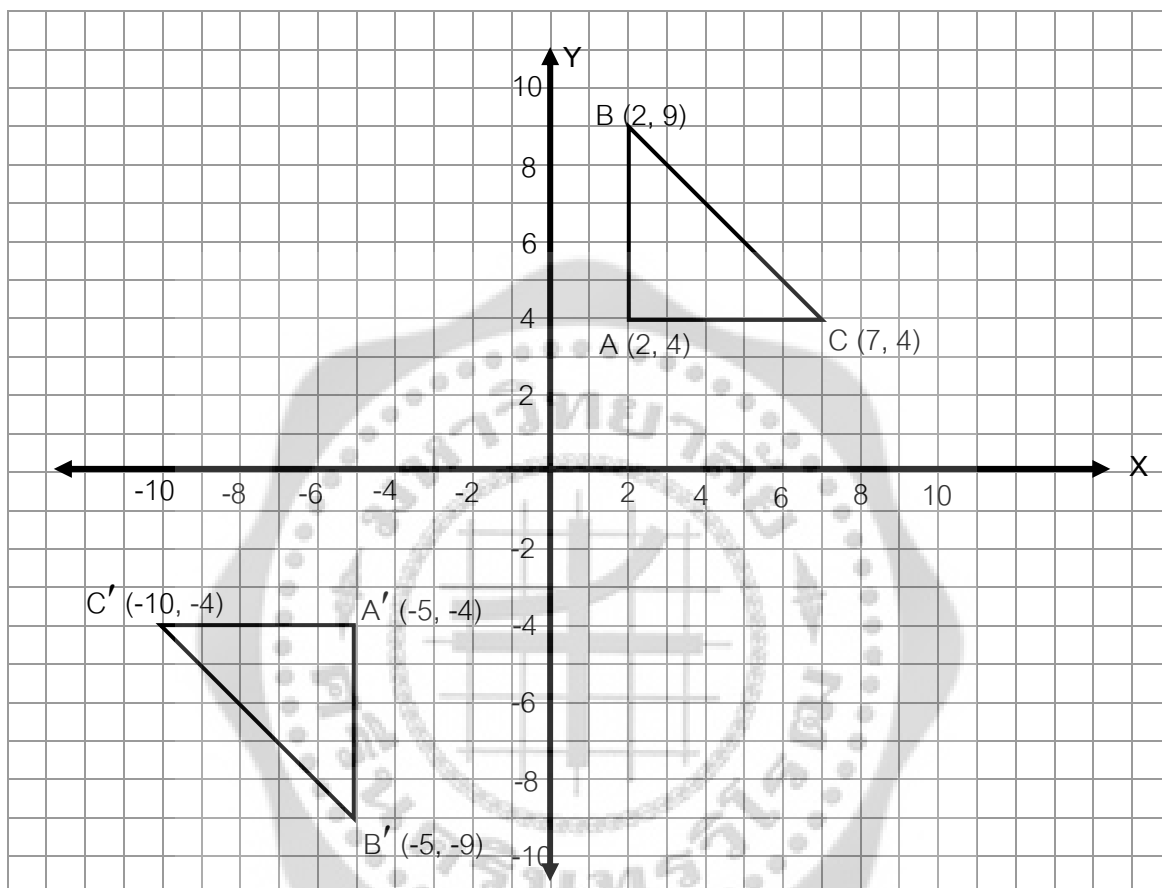
ดังนั้น รูปสามเหลี่ยมมุมฉากที่มีด้านตรงข้ามมุมฉากยาว 5 ซม. ด้านประกอบมุมฉากยาว 1 ซม. และ $\sqrt{24}$ หรือ $2\sqrt{6}$ ซม.

หมายเหตุ ด้านประกอบมุมฉาก อาจมีความยาวได้หลากหลาย เช่น $(\sqrt{2}, \sqrt{23})$, $(\sqrt{3}, \sqrt{22})$, $(2, \sqrt{21})$, $(\sqrt{5}, \sqrt{20})$, $(\sqrt{6}, \sqrt{19})$, $(\sqrt{7}, \sqrt{18})$, $(\sqrt{8}, \sqrt{17})$, $(\sqrt{10}, \sqrt{15})$, $(\sqrt{11}, \sqrt{14})$, $(\sqrt{12}, \sqrt{13})$, หรือ $\frac{5}{\sqrt{2}}, \frac{5}{\sqrt{2}}$ ซม. เป็นต้น

เกณฑ์การให้คะแนน

ระดับคะแนน	รายละเอียดของเกณฑ์
0	ไม่มีคำตอบ คำตอบไม่ถูกต้อง โดยไม่ได้แสดงเหตุผลหรือยกตัวอย่างประกอบ
1	คำตอบไม่ถูกต้อง โดยอธิบายแสดงเหตุผล หรือยกตัวอย่างประกอบ
2	คำตอบถูกต้อง แต่ - ไม่อธิบายวิธีคิดหรือแสดงเหตุผลว่าคำตอบได้มาอย่างไรและไม่ยกตัวอย่างประกอบ - อธิบายวิธีคิดหรือแสดงเหตุผลไม่ถูกต้อง หรือยกตัวอย่างประกอบไม่ถูกต้อง
3	คำตอบถูกต้อง แสดงเหตุผลได้ถูกต้อง แต่ไม่ได้ยกตัวอย่างประกอบ หรือยกตัวอย่างประกอบ แต่การให้เหตุผลไม่สมบูรณ์หรือชัดเจน
4	คำตอบถูกต้อง ใช้ความรู้มาให้เหตุผลได้อย่างชัดเจนสมบูรณ์ และยกตัวอย่างได้ถูกต้อง

8. กำหนดรูป $\triangle ABC$ มีจุด $A(2,4)$, $B(2,9)$ และ $C(7,4)$ เป็นรูปต้นแบบ และรูป $\triangle A'B'C'$ มีจุด $A'(-5,-4)$, $B'(-5,-9)$ และ $C'(-10,-4)$ จงอธิบายวิธีการแปลงทางเรขาคณิตจากรูป $\triangle ABC$ ไปเป็นรูป $\triangle A'B'C'$



แนวการตอบ

ตัวอย่างวิธีการแปลงทางเรขาคณิต จากรูป $\triangle A'B'C'$ เป็นรูป $\triangle ABC$

วิธีที่ 1 รูป $\triangle A'B'C'$ เกิดจากการสะท้อนรูป $\triangle ABC$ โดยแกน X เป็นเส้นสะท้อน จากนั้นสะท้อนรูปที่ได้ โดยมีแกน Y เป็นเส้นสะท้อน แล้วเลื่อนขนานรูปที่ได้ตามแนวแกน X ไปทางซ้ายเป็นระยะทาง 3 หน่วย

วิธีที่ 2 รูป $\triangle A'B'C'$ เกิดจากการสะท้อนรูป $\triangle ABC$ โดยมีแกน Y เป็นเส้นสะท้อน จากนั้นสะท้อนรูปที่ได้ โดยมีแกน X เป็นเส้นสะท้อน แล้วเลื่อนขนานรูปที่ได้ตามแนวแกน X ไปทางซ้ายเป็นระยะทาง 3 หน่วย

วิธีที่ 3 รูป $\triangle A'B'C'$ เกิดจากการหมุนรูป $\triangle ABC$ ตามเข็มนาฬิกาด้วยมุมขนาด 180° องศา รอบจุดกำเนิด แล้วเลื่อนขนานรูปที่ได้ตามแนวแกน X ไปทางซ้ายเป็นระยะทาง 3 หน่วย

วิธีที่ 4 รูป $\triangle A'B'C'$ เกิดจากการสะท้อนรูป $\triangle ABC$ โดยมีเส้นตรง $x = y$ เป็นเส้นสะท้อน จากนั้นเลื่อนขนานรูปที่ได้ตามแกน X ไปทางซ้ายเป็นระยะทาง 1 หน่วย แล้วเลื่อนขนานตามแกน Y ลงมาเป็นระยะทาง 2 หน่วย

หมายเหตุ สามารถอธิบายการแปลงรูป $\triangle A'B'C'$ เป็นรูป $\triangle ABC$ ได้หลายวิธี)

เกณฑ์การให้คะแนน

ระดับคะแนน	รายละเอียดของเกณฑ์
0	ไม่ตอบ หรือ ไม่เขียนอธิบายแสดงแนวคิด วิธีการใด ๆ ในการแปลงรูปจาก $\triangle ABC$ ไปยัง $\triangle A'B'C'$
1	อธิบายการแปลงรูปจาก $\triangle ABC$ ไปยัง $\triangle A'B'C'$ ได้ถูกต้องเพียงเล็กน้อย
2	อธิบายการแปลงรูปจาก $\triangle ABC$ ไปยัง $\triangle A'B'C'$ ได้ถูกต้องบางส่วน ไม่สมบูรณ์ หรือไม่ชัดเจน
3	อธิบายการแปลงรูปจาก $\triangle ABC$ ไปยัง $\triangle A'B'C'$ ได้ถูกต้อง แต่ขาดรายละเอียดเล็กน้อย เช่น ไม่ระบุเส้นสะท้อน ไม่ระบุจุดหมุนหรือทิศทางการหมุน ไม่ระบุระยะทางในการเลื่อนขนานหรือทิศทางการเลื่อนขนาน
4	อธิบายการแปลงรูปจาก $\triangle ABC$ ไปยัง $\triangle A'B'C'$ ได้ถูกต้อง ครบถ้วน สมบูรณ์

9. กำหนดความสัมพันธ์ระหว่างจำนวนเป็น

2, 6, 12, 20, 30

จงพิจารณาว่า 100 อยู่ในลำดับที่กำหนดให้หรือไม่ เพราะเหตุใด จงอธิบาย

แนวการตอบ

วิธีที่ 1 ใช้ความสัมพันธ์ระหว่างจำนวนที่กำหนดให้เป็น

ลำดับที่ 1 คือ 2

ลำดับที่ 2 คือ 6 เกิดจากจำนวนในลำดับที่ 1 บวกด้วย 4 ($6 = 2 + 4$)

ลำดับที่ 3 คือ 12 เกิดจากจำนวนในลำดับที่ 2 บวกด้วย 6 ($12 = 6 + 6$)

ลำดับที่ 4 คือ 20 เกิดจากจำนวนในลำดับที่ 3 บวกด้วย 8 ($20 = 12 + 8$)

ลำดับที่ 5 คือ 30 เกิดจากจำนวนในลำดับที่ 4 บวกด้วย 10 ($30 = 20 + 10$)

เมื่อสังเกตความสัมพันธ์ของจำนวนที่กำหนดให้ จะเห็นว่า จำนวนที่บวกเข้าไปเพิ่มทีละ

2 เริ่มจาก บวกเพิ่ม 4 ต่อไปบวกเพิ่ม 6 บวกเพิ่ม 8 บวกเพิ่ม 10 ตามลำดับ

ดังนั้นจึงคาดการณ์ว่า ลำดับต่อไป จะบวกเพิ่ม 12, 14, 16, 18, 20, ... ไปเรื่อยๆ ทำให้ได้จำนวนในลำดับถัดไป เป็น

2, 6, 12, 20, 30, 42, 56, 72, 90, 110, ... จะเห็นได้ว่า 100 ไม่ได้อยู่ในลำดับนี้

วิธีที่ 2 ใช้ความสัมพันธ์ระหว่าง ลำดับที่ของจำนวนกับ จำนวนที่กำหนดให้ ได้เป็น

ลำดับที่ 1 จำนวนที่กำหนดให้คือ 2 เกิดจาก 1×2

ลำดับที่ 2 จำนวนที่กำหนดให้คือ 6 เกิดจาก 2×3

ลำดับที่ 3 จำนวนที่กำหนดให้คือ 12 เกิดจาก 3×4

ลำดับที่ 4 จำนวนที่กำหนดให้คือ 20 เกิดจาก 4×5

ลำดับที่ 5 จำนวนที่กำหนดให้คือ 30 เกิดจาก 5×6

ดังนั้น ลำดับที่ 6 มีค่า $6 \times 7 = 42$

ลำดับที่ 7 มีค่า $7 \times 8 = 56$

ลำดับที่ 8 มีค่า $8 \times 9 = 72$

ลำดับที่ 9 มีค่า $9 \times 10 = 90$

ลำดับที่ 10 มีค่า $10 \times 11 = 110$

จะเห็นได้ว่า 100 ไม่ได้อยู่ในลำดับนี้

วิธีที่ 3 ใช้วิธีการหาความสัมพันธ์ในชุดตัวแปรก่อน

ลำดับที่ 1 จำนวนที่กำหนดให้คือ 2 เกิดจาก $1^2 + 1$

ลำดับที่ 2 จำนวนที่กำหนดให้คือ 6 เกิดจาก $2^2 + 2$

ลำดับที่ 3 จำนวนที่กำหนดให้คือ 12 เกิดจาก $3^2 + 3$

ลำดับที่ 4 จำนวนที่กำหนดให้คือ 20 เกิดจาก $4^2 + 4$

ลำดับที่ 5 จำนวนที่กำหนดให้คือ 30 เกิดจาก $5^2 + 5$

เมื่อสังเกตความสัมพันธ์ระหว่างลำดับที่ของจำนวน กับจำนวนที่กำหนดให้ จะพบว่าค่าของจำนวนในลำดับใด เกิดจากลำดับที่ของจำนวนนั้นยกกำลังสองบวกกับลำดับที่ของจำนวนนั้น

ดังนั้น เมื่อให้ n เป็นจำนวนนับ สามารถเขียนความสัมพันธ์ของแบบรูปของจำนวนชุดนี้ในรูปของตัวแปรได้เป็น ลำดับที่ n จำนวนที่กำหนดจะได้เป็น $n^2 + n$

พิจารณาว่า 100 อยู่ในลำดับที่กำหนดให้หรือไม่ จาก $100 = n^2 + n$

ไม่สามารถหาค่า n ที่เป็นจำนวนเต็มบวกที่สอดคล้องกับสมการได้

ดังนั้น 100 ไม่อยู่ในลำดับที่กำหนดให้

เกณฑ์การให้คะแนน

ระดับคะแนน	รายละเอียดของเกณฑ์
0	ไม่ตอบ หรือ ตอบไม่ถูกต้องโดยไม่ได้เขียนคำอธิบายใดๆ
1	- แสดงร่องรอยการแก้ปัญหาเพียงเล็กน้อย เช่น - แสดงวิธีแก้ปัญหาไม่ถูกต้อง - คำตอบถูกต้อง แต่ไม่แสดงวิธีการแก้ปัญหาหรือแสดงวิธีแก้ปัญหาไม่ถูกต้อง
2	แสดงวิธีแก้ปัญหาได้ถูกต้องบางส่วน เช่น หาความสัมพันธ์ระหว่างจำนวนไม่ถูกต้อง ทำให้คำนวณจำนวนในลำดับถัดไปไม่ถูกต้อง
3	- คำตอบถูกต้อง โดยอธิบายการแก้ปัญหาได้ถูกต้อง แต่ไม่เป็นขั้นตอนหรือเป็นระบบที่ชัดเจนหรือสมบูรณ์ เช่น เขียนการเชื่อมโยงระหว่างจำนวนแต่ไม่อธิบาย - แสดงวิธีแก้ปัญหาได้ถูกต้อง แต่คำนวณผิดพลาด เช่น หาความสัมพันธ์ระหว่างจำนวนได้ถูกต้อง แต่คำนวณจำนวนในลำดับถัดไปไม่ถูกต้อง
4	แสดงวิธีแก้ปัญหาย่างถูกต้อง เป็นขั้นตอนหรือเป็นระบบที่ชัดเจน สมบูรณ์ คำตอบถูกต้อง

10. ถ้าพ่อแบ่งเงินให้แบม 4 บาท แล้วจะทำให้แบมมีเงินเป็นสองเท่าของปิม แต่ถ้าพ่อแบ่งเงินให้ปิม 8 บาท แล้วจะทำให้ปิมมีเงินเป็นสามเท่าของแบม ก่อนที่พ่อจะให้เงิน แบมและปิมมีเงินคนละกี่บาท จงอธิบาย (เมื่อให้ a แทนจำนวนเงินของแบม และ b แทนจำนวนเงินของปิม)

แนวการตอบ

ให้ แบมมีเงิน a บาท และปิม มีเงิน b บาท

พ่อแบ่งเงินให้แบม 4 บาท จะทำให้แบมมีเงินเป็นสองเท่าของปิม จะได้ $a+4 = 2b$ (1)

แต่ถ้าพ่อแบ่งเงินให้ปิม 8 บาท จะทำให้ปิมมีเงินเป็นสามเท่าของแบม จะได้ $b+8 = 3a$...(2)

$$\text{จาก (1)} \quad 2b - a = 4 \quad \dots\dots\dots(3)$$

$$\text{จาก (2)} \quad b - 3a = -8 \quad \dots\dots\dots(4)$$

$$2 \times (4) \quad 2b - 6a = -16 \quad \dots\dots\dots(5)$$

$$(3) - (5) \quad 5a = 20$$

$$a = 4$$

แทนค่า a ลงใน (1) ได้เป็น $2b - 4 = 6$ จะได้ $b = 4$

นั่นคือ ก่อนที่พ่อจะให้เงินแบบ มีเงิน 4 บาท และปืมมีเงิน 4 บาท

เกณฑ์การให้คะแนน

ระดับคะแนน	รายละเอียดของเกณฑ์
0	ไม่ตอบ หรือคำตอบไม่ถูกต้องโดยไม่ได้แสดงแนวคิด วิธีการใด ๆ ในการแก้ปัญหา
1	-แสดงวิธีคิดหรือแนวคิดเล็กน้อยในการแก้ปัญหา เช่น กำหนดค่าของตัวแปร ไม่ได้คำตอบ - กำหนดสมการไม่ถูกต้อง ไม่สามารถแก้สมการได้ ไม่ได้คำตอบ
2	- กำหนดสมการไม่ถูกต้อง แสดงขั้นตอนวิธีการแก้สมการได้ถูกต้อง คำตอบไม่ถูกต้อง - กำหนดสมการถูกต้อง แสดงขั้นตอนวิธีการแก้สมการได้บางส่วน ไม่สมบูรณ์ ไม่ได้คำตอบ - ใช้การคาดเดาหรือกำหนดค่า และตรวจสอบกับเงื่อนไขโจทย์ได้ถูกต้อง
3	- กำหนดสมการถูกต้อง แสดงขั้นตอนวิธีการแก้สมการได้ถูกต้อง แต่คำนวณผิดพลาด คำตอบไม่ถูกต้อง - มีคำตอบถูกต้องเพียง 1 คำตอบ หรือ ตอบคำถามไม่ครบถ้วน
4	กำหนดสมการได้ถูกต้อง แสดงขั้นตอนวิธีการแก้สมการได้ถูกต้อง สมบูรณ์และเป็นระบบ คำตอบถูกต้อง ครบถ้วน

11. จากการสำรวจรายได้ต่อเดือนของพนักงานของบริษัทแห่งหนึ่ง จำนวน 7 คน พบว่า มีรายได้เป็น 7200, 4300, 8400, 3100, 8000, 44500 และ 4300 บาท ตามลำดับ

11.1 จงหาค่าเฉลี่ยเลขคณิต มัธยฐาน และฐานนิยม

11.2 ถ้าต้องการเลือกค่ากลางของรายได้ต่อเดือนของพนักงานหนึ่งคน เพื่อเป็นตัวแทนในการเจรจาปรับขึ้นเงินเดือน นักเรียนคิดว่า ควรเลือกค่ากลางใด เพราะเหตุใด

แนวการตอบ

11.1 คำนวณหาค่าเฉลี่ยเลขคณิต ค่ามัธยฐาน และค่าฐานนิยม

1. ค่าเฉลี่ยเลขคณิต มีค่าเป็น

$$\frac{3100 + 4300 + 4300 + 7200 + 8000 + 8400 + 44500}{7} = 11400$$

ดังนั้น ค่าเฉลี่ยเลขคณิตของรายได้ มีค่าเป็น 11400 บาท

2. เรียงข้อมูลจากมากไปน้อยหรือน้อยไปมาก จะได้

3100, 4300, 4300, 7200, 8000, 8400, 44500

ดังนั้น ค่ามัธยฐาน ซึ่งมีค่าเป็น 7200 บาท

3. ค่าฐานนิยม มีค่าเป็น 4300 บาท

11.2 ควรเลือกใช้ค่ามัธยฐาน เนื่องจากถ้าหากมีข้อมูลบางตัวมีความผิดปกติ เช่น มีข้อมูลบางตัวมีค่าสูงเกินไปหรือต่ำมากเกินไป ค่ากลางที่เหมาะสม คือ ค่ามัธยฐาน และจากโจทย์ พบว่า 44500 เป็นรายได้ที่สูงผิดปกติ คือ มีค่าสูงกว่าค่าอื่น ๆ มาก ดังนั้นค่ากลางที่เหมาะสม คือ ค่ามัธยฐาน ซึ่งมีค่าเป็น 7200 บาท

ทั้งนี้หากใช้ค่าฐานนิยม จะทำให้เป็นค่ากลางที่มีค่าต่ำเกินไป (มีพนักงานถึง 4 คนที่มีเงินเดือนสูงกว่า 4,300 บาท) และหากใช้ค่าเฉลี่ยเลขคณิต จะทำให้เป็นค่ากลางที่มีค่าสูงเกินไป (มีพนักงานถึง 6 คนที่มีเงินเดือนต่ำกว่า 11400 บาท)

เกณฑ์การให้คะแนน

ระดับคะแนน	รายละเอียดของเกณฑ์
0	ไม่ตอบ หรือ ตอบไม่ถูกต้อง(คำนวณค่ากลางผิดทั้ง 3 ค่า, เลือกใช้ไม่ถูกต้อง) โดยไม่ได้แสดงแนวคิดหรือ เหตุผลใด ๆ
1	แสดงวิธีการหาค่ากลางถูกต้องไม่เกิน 2 คน เลือกใช้ค่ากลางไม่ถูกต้อง ไม่ให้เหตุผล หรือแสดงเหตุผลไม่ถูกต้อง หรือตอบคำถามไม่ครบถ้วน
2	-แสดงวิธีการหาค่ากลางถูกต้องไม่เกิน 2 คน เลือกใช้ค่ากลางถูกต้อง โดยไม่ได้แสดง หรืออธิบายเหตุผล หรือแสดงเหตุผลไม่ถูกต้อง -แสดงวิธีการหาค่ากลางถูกต้องทั้ง 3 คน เลือกใช้ค่ากลางไม่ถูกต้อง โดยไม่ได้แสดง หรืออธิบายเหตุผล หรือแสดงเหตุผลไม่ถูกต้อง หรือตอบคำถามไม่ครบถ้วน
3	-แสดงวิธีการหาค่ากลางถูกต้องไม่เกิน 2 คน เลือกใช้ค่ากลางถูกต้อง โดยได้แสดง หรืออธิบายเหตุผล ถูกต้อง -แสดงวิธีการหาค่ากลางถูกต้องทั้ง 3 คน เลือกใช้ค่ากลางถูกต้อง ไม่ได้แสดงเหตุผล หรือแสดงเหตุผลได้แต่ไม่ชัดเจน หรือ สมบูรณ์ หรือตอบคำถามไม่ครบถ้วน
4	แสดงวิธีการหาค่ากลางถูกต้องทั้ง 3 คน เลือกใช้ค่ากลางถูกต้อง และ สามารถ อธิบายเหตุผลโดยใช้ความรู้มาให้เหตุผลได้อย่างถูกต้อง ชัดเจนและสมบูรณ์

12. ถุงใบที่หนึ่งมีลูกบอลสีแดง 4 ลูกและสีดำ 3 ลูก ถุงใบที่สองมีลูกบอลสีแดง 5 ลูกและสีดำ 4 ลูก ถ้าสุ่มหยิบได้ลูกบอลสีแดงแล้วจะได้รับรางวัล หากให้สุ่มหยิบลูกบอล 1 ลูก นักเรียนจะสุ่มหยิบลูกบอลจากถุงใดจึงจะมีโอกาสมากกว่าที่จะได้รับรางวัล จงอธิบายเหตุผล

แนวการตอบ

ถุงใบที่หนึ่งมีลูกบอลสีแดง 4 ลูกและสีดำ 3 ลูก

มีโอกาสที่จะได้รับรางวัลเป็น $\frac{4}{7}$ หรือประมาณ 0.57 หรือ 57 %

ส่วนถุงใบที่สองมีลูกบอลสีแดง 5 ลูกและสีดำ 4 ลูก

มีโอกาสที่จะได้รับรางวัลเป็น $\frac{5}{9}$ หรือประมาณ 0.56 หรือ 56 %

ดังนั้น ควรสุ่มหยิบลูกบอลจากถุงใบที่หนึ่งจึงจะมีโอกาสมากกว่าที่จะได้รับรางวัล

หมายเหตุ นักเรียนอาจพิจารณาจาก

1. การเปรียบเทียบ $\frac{4}{7}$ กับ $\frac{5}{9}$ จะได้ว่า $\frac{4}{7}$ มีค่ามากกว่า $\frac{5}{9}$
2. การเปรียบเทียบอัตราส่วนของลูกบอลสีแดงกับลูกบอลสีดำของถุงใบที่หนึ่งและถุงใบที่สองได้เป็น $\frac{4}{3}$ กับ $\frac{5}{4}$ จะได้ว่า $\frac{4}{3}$ มีค่ามากกว่า $\frac{5}{4}$

เกณฑ์การให้คะแนน

ระดับคะแนน	รายละเอียดของเกณฑ์
0	ไม่มีคำตอบ คำตอบไม่ถูกต้องโดยไม่ได้แสดงแนวคิดใด ๆ ในการแก้ปัญหา
1	- แสดงแนวคิดในการแก้ปัญหาเพียงเล็กน้อย - แสดงวิธีแก้ปัญหาไม่ถูกต้อง เช่น พิจารณาจากจำนวนลูกบอลสีแดง พิจารณาจากผลต่างของลูกบอลทั้งสองสี คำตอบไม่ถูกต้อง
2	- แสดงวิธีแก้ปัญหาได้บางส่วน สรุปคำตอบไม่ถูกต้อง หรือไม่ได้คำตอบ - สรุปคำตอบถูกต้อง แต่ไม่อธิบายการแก้ปัญหาหรือแสดงการแก้ปัญหาไม่ชัดเจน
3	- แสดงวิธีแก้ปัญหาได้ถูกต้อง แต่ไม่สมบูรณ์ เช่น ไม่แสดงการเปรียบเทียบ คำตอบถูกต้อง - แสดงวิธีแก้ปัญหาได้ถูกต้อง แต่คำนวณผิดพลาด คำตอบไม่ถูกต้อง
4	แสดงวิธีแก้ปัญหาได้ถูกต้อง สมบูรณ์ คำตอบถูกต้อง

ประวัติย่อผู้วิจัย



ประวัติย่อผู้วิจัย

ชื่อ ชื่อสกุล	น้ำผึ้ง อินทะเนตร
สถานที่เกิด	จังหวัดน่าน
สถานที่อยู่ปัจจุบัน	239/35 ถนนห้วยแก้ว ต. สุเทพ อ.เมืองเชียงใหม่ จ.เชียงใหม่
ตำแหน่งหน้าที่การงาน	พนักงานมหาวิทยาลัย ตำแหน่ง อาจารย์
สถานที่ทำงาน	สายวิชาประเมินผลและวิจัยการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยเชียงใหม่
ประวัติการศึกษา	
พ.ศ. 2541	ศึกษาศาสตรบัณฑิต (คณิตศาสตร์) เกียรตินิยมอันดับ 1 เหรียญทอง จาก มหาวิทยาลัยเชียงใหม่
พ.ศ. 2546	ศึกษาศาสตรมหาบัณฑิต (คณิตศาสตร์ศึกษา) จาก มหาวิทยาลัยเชียงใหม่
พ.ศ. 2554	การศึกษาดุษฎีบัณฑิต (การทดสอบและวัดผลการศึกษา) จาก มหาวิทยาลัยศรีนครินทรวิโรฒ