การศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน โดยใช้โมเดลการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราส์ช

น้ำผึ้ง อินทะเนตร[1] รศ.ดร.องอาจ นัยพัฒน์[2]
รศ.ดร.ผจงจิต อินทสุวรรณ[3] รศ.ดร.สุทธิวรรณ พีรศักดิ์โสภณ[4]

บทคัดย่อ

การวิจัยในครั้งนี้เป็นการศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ โดยใช้โมเดลการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราส์ช ภายใต้เงื่อนไขจำนวนผู้ตรวจต่างกันสามลักษณะ คือ 2 คน 3 คน และ 4 คน และรูปแบบการตรวจให้คะแนนต่างกันสามลักษณะ คือ รูปแบบที่ 1 ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน รูปแบบที่ 2 ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคนและรูปแบบที่ 3 ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน คุณลักษณะของคะแนนพิจารณาจากขนาดขององค์ประกอบความแปรปรวน ค่าความเชื่อมั่นและค่าความเที่ยงตรงตามสภาพ

เครื่องมือที่ใช้ในการวิจัยครั้งนี้คือแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ ระดับมัธยมศึกษาตอนต้นตามหลักสูตรการศึกษาขั้นพื้นฐาน พุทธศักราช 2551 จำนวน 12 ข้อ กลุ่มตัวอย่างที่ใช้ คือ 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. เมื่อวิเคราะห์ด้วยโมเดลหลายองค์ประกอบของราส์ช พบว่า ในทุกเงื่อนไขจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนน ความแปรปรวนของผู้สอบมีค่าสูงที่สุดรองลงมา คือ ความแปรปรวนของข้อสอบ และความแปรปรวนของผู้ตรวจมีค่าต่ำที่สุดและค่าความเชื่อมั่นแยกส่วนของผู้สอบของรูปแบบการตรวจที่ 1 มีค่าสูงสุดในทุกเงื่อนไขจำนวนผู้ตรวจ ค่าความเชื่อมั่นแยกส่วนของผู้สอบของรูปแบบการตรวจที่ 1 มีค่าสูงขึ้นเมื่อจำนวนผู้ตรวจเพิ่มขึ้น คะแนนความสามารถในทุกเงื่อนไขที่ต่างกันมีความเที่ยงตรงตามสภาพสูงและแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ

คำสำคัญ : แบบทดสอบปลายเปิด ทฤษฎีการสรุปอ้างอิง โมเดลหลายองค์ประกอบของราส์ช ความเชื่อมั่น ความเที่ยงตรงตามสภาพ

---

[1]ดุษฎีบัณฑิต สาขาวิชาการทดสอบและวัดผลการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
[2]รองศาสตราจารย์ ภาควิชาการวัดผลและวิจัยการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
[3]รองศาสตราจารย์ สถาบันวิจัยพฤติกรรมศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
[4]รองศาสตราจารย์ สำนักทดสอบทางการศึกษาและจิตวิทยา มหาวิทยาลัยศรีนครินทรวิโรฒ

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

ABSTRACT

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.

Keywords : Open-ended Test, Generalizability Theory, Many-Facet Rasch Model Reliability, Concurrent Validity

บทนำ

แบบทดสอบปลายเปิด (Open–Ended Test) เป็นเครื่องมือที่นอกจากจะวัดความเข้าใจ กระบวนการคิดในระดับสูงความสามารถในการเขียน ทักษะในการแก้ปัญหาได้เป็นอย่างดีแล้วยังเปิดโอกาสให้นักเรียนแสดงออกได้อย่างหลากหลายทั้งคำตอบและวิธีการในการแก้ปัญหา โดยการพัฒนาวิธีการหาคำตอบและการสื่อสารวิธีการแก้ปัญหาของตนเอง (NCTM. 1995 : online) อย่างไรก็ตามยังมีปัญหาในการใช้แบบทดสอบปลายเปิดอยู่หลายประการ เช่น มีความเชื่อมั่นต่ำ สิ้นเปลืองเวลา แรงงานและค่าใช้จ่ายในการตรวจมาก มีจุดอ่อนสำคัญอยู่ที่การให้คะแนน การตรวจยังมีความเป็นอัตนัย (Subjective) คะแนนแปรเปลี่ยนไปตามลักษณะของผู้ตรวจ (Linacre. 1993 : 3 ; Smith and Kulikowich. 2004 : 619) ส่งผลต่อความเที่ยงตรงและความยุติธรรมในการตัดสินตามมา (Linacre. 1993 : 3)

นักวัดผลทางการศึกษาได้ให้ข้อเสนอแนะวิธีการปรับปรุงเพื่อให้การวัดประเมินคำถามปลายเปิดมีความถูกต้องแม่นยำและน่าเชื่อถือมากขึ้น เช่น เมอเรนส์และลีแมน (Mehrens and Lehmann. 1973 : 228) กล่าวว่า เพื่อให้การตรวจคำถามปลายเปิดมีความเชื่อมั่นมากขึ้น ผู้ตรวจควรใช้วิธีการตรวจที่เหมาะสม ใช้เกณฑ์การตรวจกับนักเรียนทุกคน มีการออกแบบกฎเกณฑ์การให้คะแนน (Scoring rubric) มีการฝึกหรืออบรมผู้ตรวจ (Lane ; et. al. 1996 : 72) และควบคุมแหล่งความคลาดเคลื่อน เช่น ข้อสอบ ผู้ตรวจ ไว้ล่วงหน้าและวางแผนการตรวจเพื่อลดภาระ ลดเวลา โดยมีความเป็นไปได้ในทางปฏิบัติ (Practical) แต่ยังคงไว้ซึ่งความน่าเชื่อถือของคะแนนและให้ค่าทางสถิติที่ยอมรับได้ (Linacre and Wright. 2002 : 490-493) ทั้งนี้ คิม (Kim. 2000 : online) ไอรมณีรัทและคนอื่น ๆ (Iramaneerat ; et. al. 2007 : online) ได้ให้ข้อเสนอแนะไว้สอดคล้องกันว่าไม่จำเป็นที่ผู้ตรวจต้องตรวจให้คะแนนผู้สอบทุกคนทุกข้อ อาจให้ตรวจข้อสอบของผู้สอบบางคนหรือตรวจข้อสอบในบางคุณลักษณะหรือบางข้อและควรออกแบบการตรวจให้เหมาะสม ใช้ผู้ตรวจหรือเงื่อนไขการวัดให้น้อยที่สุด โดยที่ยังคงมีความน่าเชื่อถือในทางสถิติ (Engelhard. 1992: 188)

วิธีการทางสถิติที่สามารถค้นหา ตรวจสอบแหล่งความคลาดเคลื่อนได้หลายแหล่ง ได้แก่ โมเดลการสรุปอ้างอิงตามทฤษฎีการสรุปอ้างอิง (Generalizability Theory : GT) และโมเดลหลายองค์ประกอบของราส์ช (Many-Facet Rasch Model: MFRM) ตามทฤษฎีการตอบข้อสอบ (Item Response Theory) ทั้งนี้ GT ยอมรับว่าความแปรปรวนของผลการวัดที่เกิดขึ้นเกิดจากองค์ประกอบหลายแหล่งด้วยกัน (Shavelson and Webb. 1991 : 6) และสามารถประมาณค่าสัมประสิทธิ์การสรุปอ้างอิง (Generalizability coefficient) ที่แสดงถึงระดับความเชื่อถือได้ของคะแนน (Level of dependability) ได้ ในขณะที่ MFRM (Linacre. 1994) ปรับขยายมาจากราส์ชโมเดล (Rasch Model) จากเดิมที่ให้ค่าพารามิเตอร์ 2 ค่า คือ พารามิเตอร์ของผู้สอบ (Examinee parameter : EMBED Equation.3) และพารามิเตอร์ของข้อสอบ(Item parameter : EMBED Equation.3) โดยการเพิ่มพารามิเตอร์ของผู้ตรวจ (Rater parameter) ที่แสดงถึงความเข้มงวดของผู้ตรวจ (Rater severity) ได้จากที่กล่าวมาข้างต้นผู้วิจัยจึงสนใจศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน โดยใช้ GT และ MFRM เพื่อเป็นแนวทางในการเลือกจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนแบบทดสอบปลายเปิดที่มีความเชื่อมั่นความเที่ยงตรงของคะแนนและสอดคล้องกับสภาพจริงในทางปฏิบัติ

ความมุ่งหมายของการวิจัย

การวิจัยในครั้งนี้มีความมุ่งหมาย ดังนี้

1. เพื่อศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ ที่มีจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน โดยใช้โมเดลการสรุปอ้างอิงแยกได้เป็น

1.1 เพื่อศึกษาขนาดขององค์ประกอบความแปรปรวนของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน

1.2 เพื่อประมาณค่าและเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนน แบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน

1.3 เพื่อประมาณค่าและเปรียบเทียบความเที่ยงตรงตามสภาพของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน

2. เพื่อศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ ที่มีจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน โดยใช้โมเดลหลายองค์ประกอบของราส์ชแยกได้เป็น

2.1 เพื่อศึกษาขนาดขององค์ประกอบความแปรปรวนของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน

2.2 เพื่อประมาณค่าและเปรียบเทียบค่าความเชื่อมั่นแยกของแต่ละองค์ประกอบ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน

2.3 เพื่อประมาณค่าและเปรียบเทียบความเที่ยงตรงตามสภาพของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน

ขอบเขตของการวิจัย

ประชากรและกลุ่มตัวอย่าง

ประชากรที่ใช้ในการวิจัย คือ นักเรียนระดับชั้นมัธยมศึกษาปีที่ 4 สำนักงานเขตพื้นที่การศึกษาจังหวัดน่าน ปีการศึกษา 2552 จำนวน 3,476 คน กลุ่มตัวอย่างที่ใช้ในการวิจัย คือ นักเรียนระดับชั้นมัธยมศึกษาปีที่ 4 สำนักงานเขตพื้นที่การศึกษาจังหวัดน่าน ปีการศึกษา 2552 จำนวน 180 คน ซึ่งได้มาจากการสุ่มแบบสองขั้นตอน

ตัวแปรที่ศึกษา

ตัวแปรอิสระ 2 ตัว ได้แก่

1) จำนวนผู้ตรวจสามลักษณะ คือ 2 คน 3 คน และ 4 คน

2) รูปแบบการตรวจให้คะแนนสามลักษณะ คือ ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน

ตัวแปรตาม จำแนกเป็น 2 กลุ่ม คือ

1) เมื่อวิเคราะห์ด้วยโมเดลการสรุปอ้างอิง ได้แก่ ขนาดขององค์ประกอบความแปรปรวน ค่าสัมประสิทธิ์การสรุปอ้างอิงและความเที่ยงตรงตามสภาพ

2) เมื่อวิเคราะห์ด้วยโมเดลหลายองค์ประกอบของราส์ช ได้แก่ ขนาดขององค์ประกอบความแปรปรวน ค่าความเชื่อมั่นแยกส่วนและความเที่ยงตรงตามสภาพ

วิธีการดำเนินการวิจัย

เครื่องมือที่ใช้ในการวิจัยครั้งนี้ คือ แบบทดสอบปลายเปิดวิชาคณิตศาสตร์ที่ผู้วิจัยสร้างขึ้น จำนวน 12 ข้อ ข้อละ 4 คะแนน เป็นการวัดความสามารถทางคณิตศาสตร์ตามสาระการเรียนรู้คณิตศาสตร์ช่วงชั้นที่ 3 (ชั้นมัธยมศึกษาปีที่ 1-3) ตามหลักสูตรการศึกษาขั้นพื้นฐาน พุทธศักราช 2551 ได้แก่ สาระที่ 1 จำนวนและการดำเนินการ สาระที่ 2 การวัด สาระที่ 3 เรขาคณิต สาระที่ 4 พีชคณิตและสาระที่ 5 การวิเคราะห์ข้อมูลและความน่าจะเป็น ส่วนสาระที่ 6 ทักษะ/กระบวนการทางคณิตศาสตร์ใช้เป็นแนวทางในการสร้างกฎเกณฑ์การให้คะแนน ข้อสอบมีค่าความยากอยู่ระหว่าง 0.2499-0.7729 ค่าอำนาจจำแนกอยู่ระหว่าง 0.2015-0.6752 และค่าความเชื่อมั่นของแบบทดสอบทั้งฉบับเป็น 0.8771

ผู้วิจัยดำเนินการเก็บรวบรวมข้อมูลด้วยตนเอง ระหว่างวันที่ 24-31 สิงหาคม 2552 โดยได้อบรม ชี้แจงผู้ตรวจให้เข้าใจในกฎเกณฑ์การให้คะแนน จากนั้นผู้วิจัยนำผลการตรวจให้คะแนนข้อสอบจำนวน 12 ข้อ จากผู้ตรวจทั้ง 4 คนของนักเรียน 180 คน มาจัดกระทำข้อมูลตามเงื่อนไขและระดับของเงื่อนไขที่ต้องการศึกษา จึงนำผลการตรวจให้คะแนนมาวิเคราะห์ข้อมูล ดังนี้

1. วิเคราะห์ค่าสถิติพื้นฐานของคะแนนและค่าพารามิเตอร์ของแต่ละองค์ประกอบ

2. วิเคราะห์องค์ประกอบความแปรปรวน โดยใช้โมเดลการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราส์ช เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกันด้วยโปรแกรม GENOVA (Crick and Brennan. 1983) และโปรแกรม FACETS (Linacre. 2009) ตามลำดับ

3. ประมาณค่าสัมประสิทธิ์การสรุปอ้างอิง (EMBED Equation.3) ด้วยโปรแกรม GENOVA (Crick and Brennan. 1983) ประมาณค่าความเชื่อมั่นแยกส่วน (Separation reliability) ของแต่ละองค์ประกอบด้วยโปรแกรม FACETS (Linacre. 2009) เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกันและทดสอบความแตกต่างของค่าสัมประสิทธิ์การสรุปอ้างอิง ค่าความเชื่อมั่นแยกส่วนของแต่ละองค์ประกอบโดยใช้สูตร EMBED Equation. 3 ของวูดรัฟฟ์และเฟลดท์ (Woodruff and Feldt. 1986 : 393-413) ซึ่งเป็นการเปรียบเทียบภายในแต่ละโมเดล

4. ประมาณค่าความเที่ยงตรงตามสภาพ (Concurrent validity) ของคะแนน โดยพิจารณาจากค่าสัมประสิทธิ์สหสัมพันธ์ของคะแนนดิบ คะแนนความสามารถกับคะแนนที่ได้จากผลการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (Ordinary National Education Test : O-NET) ในช่วงชั้นที่ 3 (ระดับชั้นมัธยมศึกษาปีที่ 3) ปีการศึกษา 2551 และเปรียบเทียบค่าความเที่ยงตรงตามสภาพ ซึ่งเป็นการเปรียบเทียบภายในแต่ละ

โมเดล โดยการนำค่าสัมประสิทธิ์สหสัมพันธ์มาแปลงเป็นคะแนนมาตรฐาน (Z) ของฟิชเชอร์ แล้วทดสอบความแตกต่างโดยใช้สถิติไค-สแควร์ (Chi-Square ; EMBED Equation.3) (Meng ; Rosenthal and Robin. 1992 : 173) และทดสอบความแตกต่างรายคู่ โดยใช้สถิติ Z (Meng ; Rosenthal and Robin. 1992 : 173)

ผลการวิจัย

1. การวิเคราะห์ด้วยโมเดลการสรุปอ้างอิง ผลการวิจัยเป็นดังนี้

1.1 ขนาดองค์ประกอบความแปรปรวน พบว่า

รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน ขนาดขององค์ประกอบความแปรปรวนที่มีค่ามากที่สุดในทุกจำนวนผู้ตรวจเมื่อเทียบกับความแปรปรวนรวม คือ ปฏิสัมพันธ์ระหว่างผู้สอบกับข้อสอบ (43.0548-43.0817%) รองลงมา คือ ผู้สอบและข้อสอบ ส่วนความแปรปรวนของผู้ตรวจมีค่าค่อนข้างน้อย (0.0474-1.2261%) เมื่อใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน องค์ประกอบความแปรปรวนที่มีค่ามากที่สุดในทุกจำนวนผู้ตรวจ คือ ปฏิสัมพันธ์ระหว่างข้อสอบกับผู้สอบในแต่ละผู้ตรวจและความคลาดเคลื่อนเชิงสุ่มที่ไม่อาจระบุได้ (46.8346-48.3318%) รองลงมาคือผู้สอบในแต่ละผู้ตรวจและข้อสอบ ส่วนความแปรปรวนของผู้ตรวจมีค่าค่อนข้างน้อย (2.2087-4.0235%) และในรูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน องค์ประกอบความแปรปรวนที่มีค่ามากที่สุดในทุกจำนวนผู้ตรวจ คือ ปฏิสัมพันธ์ระหว่างข้อสอบกับผู้สอบในแต่ละผู้ตรวจและความคลาดเคลื่อนเชิงสุ่มที่ไม่อาจระบุได้ (48.4450-52.6648%) รองลงมาคือผู้สอบและข้อสอบในแต่ละผู้ตรวจส่วนความแปรปรวนของผู้ตรวจมีค่าค่อนข้างน้อย (0.0000-3.5661%)

1.2 การประมาณค่าและเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ พบว่า

1.2.1 เมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน พบว่า รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคนให้ค่าสัมประสิทธิ์การสรุปอ้างอิงสูงสุด (0.970-0.975) รองลงมาคือผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน (0.901-0.904) และผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (0.882-0.901) ตามลำดับในทุกจำนวนผู้ตรวจ

1.2.2 เมื่อรูปแบบการตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจต่างกัน พบว่า เมื่อใช้รูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน จำนวนผู้ตรวจที่มากกว่ามีค่าสัมประสิทธิ์การสรุปอ้างอิงสูงกว่าจำนวนผู้ตรวจที่น้อยกว่าโดยผู้ตรวจ 4 คน 3 คน และ 2 คน ให้ค่าสัมประสิทธิ์การสรุปอ้างอิงเป็น 0.904, 0.903 และ 0.901 ตามลำดับ ส่วนในรูปแบบการตรวจอื่นมีค่าสัมประสิทธิ์การสรุปอ้างอิงสูงสุดเมื่อใช้ผู้ตรวจ 3 คน รองลงมา คือ 2 คนและ 4 คน ตามลำดับ โดยผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคนมีค่าสัมประสิทธิ์การสรุปอ้างอิงเป็น 0.975, 0.975 และ 0.970 ตามลำดับและผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคนมีค่าสัมประสิทธิ์การสรุปอ้างอิงเป็น 0.901, 0.894 และ 0.882 ตามลำดับ

1.3 การประมาณค่าและเปรียบเทียบความเที่ยงตรงตามสภาพของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน พบว่า ค่าสัมประสิทธิ์สหสัมพันธ์ในทุกเงื่อนไขมีค่าสูงและใกล้เคียงกัน (r มีค่าตั้งแต่ 0.778 ถึง 0.796) และมีนัยสำคัญทางสถิติที่ระดับ .01 ในทุกเงื่อนไข เมื่อนำมาทดสอบความแตกต่าง พบว่า ค่าสัมประสิทธิ์สหสัมพันธ์ของคะแนนสอบกับคะแนน O-NET ในแต่ละเงื่อนไขจำนวนผู้ตรวจและรูปแบบการตรวจที่ต่างกัน มีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติที่ระดับ .01

2. การวิเคราะห์ด้วยโมเดลหลายองค์ประกอบของราส์ช ผลการวิจัยเป็นดังนี้

2.1 ขนาดขององค์ประกอบความแปรปรวน พบว่า ในทุกจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนความแปรปรวนของผู้สอบมีค่ามากที่สุด (63.9154-7.0979 %) เมื่อเทียบกับความแปรปรวนรวมรองลงมา คือ ความแปรปรวนของข้อสอบ (18.7373-26.4946 %) และความแปรปรวนของผู้ตรวจ (0.8444-13.9194%) ตามลำดับ

2.2 ค่าความเชื่อมั่นแยกส่วน (Separation reliability) ของแต่ละองค์ประกอบ

2.2.1 เมื่อจำนวนผู้ตรวจเท่ากันและรูปแบบการตรวจให้คะแนนต่างกัน พบว่า เมื่อใช้รูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคนและจำนวนผู้ตรวจ 4 คน 3 คน และ 2 คน ให้ค่าความเชื่อมั่นแยกส่วนของผู้สอบเป็น 0.97, 0.96 และ 0.94 ตามลำดับและให้ค่าความเชื่อมั่นแยกส่วนข้อสอบเป็น 0.99 ในทุกจำนวนผู้ตรวจซึ่งเป็นค่าที่สูงกว่ารูปแบบการตรวจให้คะแนนอื่นในทุกจำนวนผู้ตรวจ ส่วนความเชื่อมั่นแยกส่วนของผู้ตรวจมีค่าเป็น 0.96, 0.97 และ 0.98 ตามลำดับและมีแนวโน้มว่ามีค่าต่ำกว่ารูปแบบการตรวจให้คะแนนอื่นในทุกจำนวนผู้ตรวจ

2.2.2 เมื่อรูปแบบการตรวจให้คะแนนเหมือนกันและจำนวนผู้ตรวจต่างกัน พบว่า เมื่อใช้รูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน จำนวนผู้ตรวจที่มากกว่ามีค่าความเชื่อมั่นแยกส่วนของผู้สอบสูงกว่าจำนวนผู้ตรวจที่น้อย จำนวนผู้ตรวจที่มากกว่ามีค่าความเชื่อมั่นแยกส่วนของผู้ตรวจต่ำกว่าจำนวนผู้ตรวจที่น้อยกว่าและเมื่อผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคนมีแนวโน้มว่าจำนวนผู้ตรวจที่มากกว่ามีค่าความเชื่อมั่นแยกส่วนของข้อสอบสูงกว่าจำนวนผู้ตรวจที่น้อย

3. การประมาณค่าและเปรียบเทียบความเที่ยงตรงตามสภาพของคะแนนความสามารถจากแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน พบว่า ค่าสัมประสิทธิ์สหสัมพันธ์ในทุกเงื่อนไขมีค่าสูงและใกล้เคียงกัน (r มีค่าตั้งแต่ 0.769 ถึง 0.791) และมีนัยสำคัญทางสถิติที่ระดับ .01 ในทุกเงื่อนไข เมื่อนำมาทดสอบความแตกต่าง พบว่า ค่าสัมประสิทธิ์สหสัมพันธ์ของคะแนนความสามารถกับคะแนน O-NET ในแต่ละเงื่อนไขจำนวนผู้ตรวจและรูปแบบการตรวจที่ต่างกันมีค่าแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติที่ระดับ .01

อภิปรายผลการวิจัย

1. ความแปรปรวนของแต่ละองค์ประกอบ

1.1 เมื่อวิเคราะห์ด้วยโมเดลการสรุปอ้างอิง ขนาดความแปรปรวนขององค์ประกอบที่มีค่ามากที่สุด เมื่อใช้รูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน คือ ปฏิสัมพันธ์ระหว่างผู้สอบกับข้อสอบ และในรูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคนและผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน คือ ปฏิสัมพันธ์ระหว่างผู้สอบกับข้อสอบในแต่ละผู้ตรวจแต่ละคนและความคลาดเคลื่อนเชิงสุ่มที่ไม่อาจระบุได้ แสดงว่า อันดับที่ของผู้สอบมีการเปลี่ยนแปลงไปตามข้อสอบหรือผู้สอบไม่มีความคงเส้นคงว่าในการตอบข้อสอบแต่ละข้อ อาจเนื่องมาจากความยากของข้อสอบที่แตกต่างกัน และผู้สอบไม่ได้มีความรู้ความสามารถที่เท่าเทียมกันในการตอบข้อสอบแต่ละข้อ อีกทั้งมีเนื้อหาครอบคลุมทั้ง 6 สาระ จึงอาจทำให้มีเนื้อหาที่หลากหลาย ผู้สอบบางคนอาจมีความเชี่ยวชาญในบางเนื้อหาต่างกัน ประกอบกับการสุ่มกลุ่มผู้สอบให้กับผู้ตรวจแต่ละคน ซึ่งผู้วิจัยใช้การสุ่มอย่างง่ายหรือการสุ่มกลุ่มข้อสอบให้กับผู้ตรวจแต่ละคน ซึ่งผู้วิจัยใช้การสุ่มแบบแบ่งชั้น โดยมีสาระการเรียนรู้เป็นชั้นและข้อสอบเป็นหน่วยการสุ่ม จึงอาจทำให้ความสามารถของผู้สอบในแต่ละกลุ่มแตกต่างกันในแต่ละผู้ตรวจหรือความยากข้อสอบในแต่ละกลุ่มที่ผู้ตรวจแต่ละคนได้รับต่างกันและร่วมกับแหล่งความคลาดเคลื่อนที่ไม่อาจระบุได้ นั่นแสดงว่าหลังจากพยายามอธิบายความไม่สอดคล้องกันนี้ด้วยองค์ประกอบแหล่งต่าง ๆ แล้วยังมีความแปรปรวนอีกจำนวนหนึ่งที่ไม่สามารถอธิบายได้

นอกจากนี้ความแปรปรวนของผู้ตรวจในทุกเงื่อนไขมีค่าน้อยและน้อยกว่าองค์ประกอบอื่น แสดงว่าผู้ตรวจมีความแตกต่างกันในการให้คะแนนน้อยและในบางเงื่อนไขความแปรปรวนของผู้ตรวจมีค่าเป็นศูนย์ นั่นคือสามารถขจัดความแปรปรวนของผู้ตรวจได้อย่างสิ้นเชิง อาจเนื่องมาจากการมีแนวปฏิบัติการตรวจที่ชัดเจน ผู้ตรวจได้รับการฝึกฝนการตรวจตามเกณฑ์ที่สร้างขึ้น จึงมีความเป็น ปรนัยในการให้คะแนนและช่วยให้ผู้ตรวจมีความเห็นที่สอดคล้องกันมากขึ้น ประกอบกับผู้สอบได้รับการปิดชื่อ โรงเรียนจึงช่วยลดอคติของผู้ตรวจได้ สอดคล้องกับผลการวิจัยของเลนน์และคนอื่น ๆ (Lane ; et. al. 1996 : 87) สมิทและคูลลิโกวิท (Smith and Kulikowich. 2004 : 625) ที่พบว่าความแปรปรวนของผู้ตรวจมีค่าน้อยมาก เนื่องมาจากการฝึกอบรมผู้ตรวจและการใช้เกณฑ์เดียวกันกับผู้สอบทุกคน

1.2 เมื่อวิเคราะห์ด้วยโมเดลหลายองค์ประกอบของราส์ช พบว่า องค์ประกอบที่มีความแปรปรวนมากที่สุดในทุกเงื่อนไขจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนเมื่อเทียบกับความแปรปรวนรวม คือ ผู้สอบ รองลงมา คือ ข้อสอบและผู้ตรวจมีค่าความแปรปรวนน้อยที่สุด สอดคล้องกับผลการวิจัยของอัพเชอร์และเทอร์เนอร์ (Upshur and Turner. 1999 : 97) ที่พบว่าองค์ประกอบที่มีความแปรปรวนมากที่สุด คือ ผู้สอบ รองลงมา คือ ข้อสอบและผู้ตรวจ ตามลำดับและผลการวิจัยของบอน์คและออคคีย์ (Bonk and Ockey. 2003 : 96-97) แมคมานัส ทอมป์สันและมูลลอน (McManus ; Thompson and Mollon. 2006 : online) พบว่าความแปรปรวนของผู้สอบมีค่าสูงสุด รองลงมา คือ ผู้ตรวจและข้อสอบตามลำดับ ทั้งนี้เหตุผลสามารถอธิบายได้ตาม 1.1

2. การศึกษาค่าความเชื่อมั่น

2.1 ค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนในทุกจำนวนผู้ตรวจ รูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคนให้ค่าสัมประสิทธิ์การสรุปอ้างอิงสูงสุด รองลงมา คือ ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคนและผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคนตามลำดับ ทั้งนี้อาจเนื่องมาจากการกระจายของคะแนนที่แตกต่างกัน ดังที่กรอนลันด์ (Grounlund. 1976 : 117) ได้กล่าวไว้ว่า ถ้าความสามารถของกลุ่มผู้สอบแตกต่างกันมากจะทำให้มีค่าความเชื่อมั่นสูงกว่ากลุ่มผู้สอบที่มีความสามารถใกล้เคียงกัน ซึ่งรูปแบบการตรวจที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคนมีค่าพิสัยของคะแนนกว้างและส่วนเบี่ยงเบนมาตรฐานของคะแนนมีค่าสูงกว่ารูปแบบการตรวจอื่น ซึ่งอาจเกิดจากการสุ่มกลุ่มผู้สอบที่มีความสามารถต่างกันมากให้แต่ละผู้ตรวจทำให้คะแนนมีการกระจายมากตามมา

2.2 ค่าความเชื่อมั่นแยกส่วนในแต่ละองค์ประกอบในทุกเงื่อนไขจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนน องค์ประกอบผู้สอบมีค่าความเชื่อมั่นสูงเกิน 0.86 และมีนัยสำคัญทางสถิติแสดงว่าความสามารถของผู้สอบว่ามีการกระจายตลอดช่วงความสามารถ (Continuum) (Smith and Kulkowich. 2004 : 629) ซึ่งค่าดังกล่าวสามารถอธิบาย แปลความได้เช่นเดียวกับ KR 20 ครอนบาคและค่าสัมประสิทธิ์การสรุปอ้างอิง (Engelhard. 1994 : 97) นั่นคือ มีความเชื่อถือได้ของคะแนนความสามารถของผู้สอบ (Upshur and Turner. 1999 : 97) ทั้งนี้อาจเนื่องมาจากการสุ่มกลุ่มตัวอย่างที่ได้ผู้สอบที่มีความสามารถทางคณิตศาสตร์แตกต่างกัน ส่วนค่าความเชื่อมั่นแยกส่วนของข้อสอบ พบว่า มีค่าสูงเกิน 0.97 และมีนัยสำคัญทางสถิติเช่นกัน แสดงว่าค่าความยากของข้อสอบมีความแตกต่างกันและมีการกระจายอยู่ตลอดช่วงความสามารถทั้งนี้เมื่อความแปรปรวนของค่าความยากของข้อสอบมีค่าสูง ค่าความเชื่อมั่นแยกส่วนของข้อสอบมีค่าสูงด้วย (Smith and Kulikowich. 2004 : 619) ในขณะที่ค่าความเชื่อมั่นแยกส่วนของผู้ตรวจ ซึ่งแสดงถึงความแตกต่างกัน (Reliably different) ของความเข้มงวด (Sudweeks ; Reeve and Bradshaw. 2005 : 254) พบว่า มีค่าสูงเกิน 0.89 แสดงว่าผู้ตรวจมีความความเข้มงวดในการให้คะแนนต่างกัน แต่เมื่อพิจารณาค่า INFIT และ OUTFIT โดยภาพรวมพบว่า ผู้ตรวจมีความคงที่ในการให้คะแนน หรือกล่าวอีกนัยหนึ่งว่า แม้ว่าผู้ตรวจอาจให้คะแนนต่างกัน แต่ผู้ตรวจแต่ละคนต่างมีความสอดคล้อง (Consistentcy) ในการให้คะแนนข้อสอบทุกข้อ

3. การศึกษาค่าความเที่ยงตรงตามสภาพ

พิจารณาจากค่าสัมประสิทธิ์สหสัมพันธ์ของคะแนนสอบกับคะแนน O-NET พบว่า มีค่าสหสัมพันธ์สูง (r มีค่าตั้งแต่ 0.778 ถึง 0.796) และคะแนนความสามารถกับคะแนน O-NET มีค่าสหสัมพันธ์สูง (r มีค่าตั้งแต่ 0.769 ถึง 0.791) และแตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติที่ระดับ .01 ในทุกเงื่อนไข อาจเนื่องมาจากแบบทดสอบที่ใช้ในงานวิจัยมีคุณภาพกล่าวคือ แบบทดสอบผ่านการพิจารณาของผู้เชี่ยวชาญ ภาษาที่ใช้เหมาะสม ข้อคำถามชัดเจน ผู้ตอบจึงมีความเข้าใจ สามารถแสดงความรู้ ความคิดตามต้องการได้และข้อสอบมีความยากพอเหมาะมีอำนาจจำแนกที่ดี อีกทั้งคำชี้แจงในการตอบคำถามมีความชัดเจน ไม่เปิดโอกาสให้มีการเดาเนื่องด้วยต้องเขียนแสดงวิธีการได้มาซึ่งคำตอบ ประกอบกับผู้วิจัยดำเนินการสอบด้วยตนเองทำให้สามารถควบคุมดูแลการดำเนินการสอบได้ อีกทั้งผู้วิจัยได้อบรมผู้ตรวจให้เข้าใจเกณฑ์ในการให้คะแนนก่อนการตรวจ ดังคำกล่าว

ของเมอเรนส์และลีแมน (Mehrens and Lehmann. 1973 : 228) ที่ว่า เพื่อให้การตรวจคำถามปลายเปิดมีความเที่ยงตรงเพิ่มขึ้น ผู้ตรวจควรใช้วิธีการตรวจที่เหมาะสมใช้เกณฑ์การตรวจกับนักเรียนทุกคน มีการออกแบบกฎเกณฑ์การให้คะแนนและมีการฝึกหรืออบรมผู้ตรวจ (Lane ; et. al. 1996 : 72)

ข้อเสนอแนะ

ผลการวิจัยครั้งนี้ชี้ให้เห็นว่าความแปรปรวนของผู้ตรวจมีค่าค่อนข้างน้อย ค่าสัมประสิทธิ์การสรุปอ้างอิงและค่าความเชื่อมั่นแยกส่วนของผู้สอบมีค่าสูงและมีความเที่ยงตรงตามสภาพของคะแนนในทุกเงื่อนไข ดังนั้นจึงไม่จำเป็นที่ผู้ตรวจต้องตรวจข้อสอบทุกข้อของผู้สอบทุกคน สามารถเลือกรูปแบบการตรวจอื่นที่มีความสะดวกและประหยัดได้ จำนวนผู้ตรวจ 2-3 คน น่าจะเป็นจำนวนที่เหมาะสมและมีความเป็นไปได้ในปฏิบัติได้จริง นอกจากนี้ควรสร้างกฎเกณฑ์การให้คะแนนที่ชัดเจน มีการให้ข้อมูล ความรู้และอบรมผู้ตรวจให้เข้าใจและแม่นยำในการใช้กฎเกณฑ์การให้คะแนน

กิตติกรรมประกาศ

ขอขอบพระคุณมหาวิทยาลัยศรีนครินทรวิโรฒที่สนับสนุนทุนอุดหนุนการวิจัย ประจำปีงบประมาณ 2552 และมหาวิทยาลัยเชียงใหม่ที่สนับสนุนทุนตามโครงการพัฒนาอาจารย์ของมหาวิทยาลัยเชียงใหม่

เอกสารอ้างอิง

Bonk, W. J. and Ockey, G. 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

Crick, J. E. and Brennan, R. L. (1983). **Mannual for GENOVA : A Generalized Analysis of Variance System (ACT Technical Bulletin no. 43)**. Iowa. IA : American College Testing Program.

Engelhard, G. Jr. (1992). **The Measurement of Writing Ability With A Many-Facet Rasch Model**. Applied Measurement in Education. 5(3) : 171-191.

----------. (1994, Summer). **Examing Rater Errors in the Assessment of Written Composition With a Many-Faceted Rasch Model**. Journal of Education Measurement. 31(2) : 93-112.

Grounlund, N. E. (1976). **Measurement and Evaluation in Testing. 3rd ed**. New York. Macmillan.

Iramaneerat, C. ; Yudkowsky, R. ; Myford, C. M. ; 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, S. C. (2000). **Investigating the Generalizability of Scores from Different Rating Systems in Performance Assessment**. Retrieved July 15, 2008, from http://eric.ed. gov/ERICDocs/data/ericdocs2sql/ content_storage_01/0000019b/80/16/cc/1f.pdf

Lane, S. ; Lui, M. ; Ankenmann, R. D. ; et. al. (1996). **Generalizability and Validity of Mathematics Performance Assessment**. Journal of Education Measurement. 33(1) : 71-92.

Linacre, J. M. (1993). **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. Retrieved August 10, 2008, from http://www.eric.ed.gov/PDFS/ED364573.pdf

----------. (1994). **Many-Facet Rasch Measurement**. Chicago : MESA Press.

----------. (2009). **A User's Guide to FACETS [Computer Program Manual]**. Chicago : MESA Press.

Linacre, J. M. and Wright, B. D. (2002). **Construction of Measures from Many-facet Data**. Journal of Applied Measurement. 3(4) : 486-512.

McManus, I. C. ; Thompson, M. and Mollon, J. (2006). **Assessment of Examiner Leniency and Stringency in the MRCP(UK) Clinical Examination(PACES) Using Multi-Facet Rasch Modeling**. Retrieved March 20, 2009, from http://creativecommons.org/license/by/2.0

Mehrens, W. A. and Lehmann, I. J. (1973). **Measurement and Evaluation in Education and Psychology**. New York : Holt, Rinehart and Winston.

Meng, X. L. ; Rosenthal, R. and Rubin, D. B. (1992). **Comparing Correlated Correlation Coefficients**. Psychological Bulletin. 111(1) : 172-175.

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

Shavelson, R. J. and Webb, N. N. (1991). **Generalizability Theory : A Primer**. Sage Publications.

Smith, E.V. and Kulikowich, J.M. (2004). **An Application of Gerneralizability Theory and Many-Facet Rasch Measurement Using A Complex Problem-Solving Skills Assessment**. Educational and Psychological Measurement. 64(4) : 617-639.

Sudweeks,R. R. ; Reeve,S. and Bradshaw, W. S. (2005). **A Comparison of Generalizability Theory and Many-Facet Rash Measurement in an Analysis of College Sophomore Writing**. Assessment Writing. 9 (3) : 239-261.

Upsher,J. A. and Turner,C. 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 HYPERLINK "http://ltj.sagepub.com" http://ltj.sagepub.com/cgi/content/ abstract/16/1/82

Woodruff,D. J. and Feldt,L. S. (1986, September). **Test for Equality of Several Alpha Coefficients When Their Sample Estimates are Dependent**. Psychometrika. 51(2) : 393-413.