

การเปรียบเทียบคุณภาพของการตรวจให้คะแนนของแบบสอบอัตนัยวัดความสามารถในการ แก้ปัญหาทางคณิตศาสตร์ ที่ใช้การออกแบบการตรวจให้คะแนนที่ต่างกัน:

การประยุกต์ใช้ทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือผลของการวัด

Quality Comparison of Scoring of Essay Tests to Measure Mathematics

Problem Solving Ability Using Different Scoring Designs:

Application of Generalizability Theory

สุรัชย์ มาตราช^{1*} และ ประกฤติยา ทักซิโน²

Surachai Matharach^{1*} and Prakittiya Tuksino²

(Received: April 8, 2022; Revised: June 12, 2022; Accepted: June 21, 2022)

บทคัดย่อ

การวิจัยครั้งนี้มีความมุ่งหมาย 1. เพื่อศึกษาคุณภาพการตรวจให้คะแนนแบบสอบอัตนัยวัดความสามารถในการแก้ปัญหาทางคณิตศาสตร์ของผู้ตรวจให้คะแนนภายใต้รูปแบบการตรวจให้คะแนนต่างกัน 2. เพื่อศึกษาน้ำหนักของแหล่งความแปรปรวนขององค์ประกอบและค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนแบบสอบอัตนัยวัดความสามารถในการแก้ปัญหาทางคณิตศาสตร์ เมื่อจำนวนผู้ตรวจ และรูปแบบการตรวจให้คะแนนแตกต่างกัน 3. เพื่อเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิง (G-Coefficient) ของแบบสอบอัตนัยวัดความสามารถในการแก้ปัญหาทางคณิตศาสตร์ ภายใต้รูปแบบการตรวจให้คะแนนต่างกัน และจำนวนผู้ตรวจต่างกัน กลุ่มตัวอย่างในการวิจัยเป็นนักเรียนชั้นมัธยมศึกษาปีที่ 3 จำนวน 132 คน ได้จากการสุ่มแบบ 2 ขั้นตอน และกลุ่มผู้ตรวจให้คะแนนซึ่งเป็นครูผู้สอนวิชาคณิตศาสตร์ จำนวน 3 คน ได้จากการเลือกแบบเจาะจงเครื่องมือที่ใช้ในการวิจัยคือ แบบสอบอัตนัยวัดความสามารถในการแก้ปัญหาทางคณิตศาสตร์ ประกอบไปด้วยข้อสอบจำนวน 3 ข้อ มีค่าความยากตั้งแต่ 0.22 – 0.69 และค่าอำนาจจำแนกตั้งแต่ 0.56 – 0.80 ผลการวิจัยพบว่า 1) คุณภาพของการตรวจให้คะแนนแบบสอบอัตนัยวัดความสามารถในการแก้ปัญหาทางคณิตศาสตร์ ผู้ตรวจให้คะแนนคนที่ 1 มีความเห็นพ้องในการตรวจแบบสอบเมื่อเทียบกับเกณฑ์การตรวจให้คะแนนข้อที่ 3 มากที่สุด ผู้ตรวจให้คะแนนคนที่ 2 มีความเห็นพ้องในการตรวจแบบสอบเมื่อเทียบกับเกณฑ์การตรวจให้คะแนนข้อที่ 1 และ 2 มากที่สุด ส่วนผู้ตรวจให้คะแนนคนที่ 3 มีความเห็นพ้องในการตรวจแบบสอบเมื่อเทียบกับเกณฑ์การตรวจ

¹ นักศึกษาปริญญาโท สาขาวิชาการวัดและประเมินผลการศึกษา คณะศึกษาศาสตร์มหาวิทยาลัยขอนแก่น

² ผู้ช่วยศาสตราจารย์ สาขาวิชาการวัดและประเมินผลการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยขอนแก่น

¹ Master's degree student, Educational Measurement and Evaluation Program, Faculty of Education, Khon Kaen University

² Assistant Professor, Department of Measurement and Evaluation Program, Faculty of Education, Khon Kaen University

* Corresponding Author E-mail: Surachai_m@kkumail.com

ให้คะแนนทั้ง 3 ข้อน้อยที่สุด และความเที่ยงในการตรวจให้คะแนนระหว่างผู้ตรวจ (inter rater reliability) มีความสอดคล้องกันในระดับดีมาก 2) แหล่งความแปรปรวนในแต่ละรูปแบบการตรวจให้คะแนน รูปแบบการตรวจให้คะแนนแบบ $P \times I \times R$ แหล่งที่มีความแปรปรวนมากที่สุดคือ PI และแหล่งความแปรปรวนที่มีค่าน้อยที่สุดคือ PR และรูปแบบการตรวจให้คะแนนแบบ $P \times (I : R)$ แหล่งที่มีความแปรปรวนมากที่สุดคือ PI : R และแหล่งความแปรปรวนที่มีค่าน้อยที่สุดคือ PR 3) รูปแบบการตรวจให้คะแนนที่มีค่าสัมประสิทธิ์การสรุปอ้างอิงสูงที่สุดคือรูปแบบ $P \times I \times R$ และจำนวนผู้ตรวจ 3 คน ผลการตรวจมีค่าความเที่ยงมากกว่าผู้ตรวจ 2 คน

คำสำคัญ: ดัชนีความเห็นพ้องของผลการตรวจ สัมประสิทธิ์สหสัมพันธ์ภายในชั้น ทฤษฎีการสรุปอ้างอิง
ความน่าเชื่อถือของผลการวัด

Abstract

The objectives of this research were: 1) to study the quality of scoring the essay test measuring the mathematical problem-solving ability of the scorers under different scoring models; 2) to study the magnitude of the source of variance of the components and the reference summation coefficient of the essay test scores on mathematical problem solving when the number of scorers and scoring formats are different; 3) to compare the reference summation coefficients. (G-Coefficient) of the essay test that measures the ability to solve mathematical problems under different scoring formats and different number of scorers.

The sample consisted of 132 grade 9 students, obtained from two-stage cluster sampling, and a group of 3 scorers who were mathematics teachers, obtained through purposive sampling. The research instrument was an essay test measuring the ability to solve mathematical problems, with 3 items. The difficulty ranged from 0.22 to 0.69 and the discrimination ranged from 0.56 to 0.80. The results were as follows: 1) Regarding the quality of scoring of the essay test measuring the ability to solve mathematical problems, scorer 1 had the most consensus on the test, compared to the scoring criterion of item 3. Scorer 2 had the most consensus on the test, compared to the scoring criterion of item 1 and 2, while scorer 3 had the least consensus on the test, compared to the scoring criteria of all 3 items; and the reliability of the scoring between the scorers (Inter Rater Reliability) was consistent at a very good level.

2) Regarding the source of variance in each scoring format, in the $P \times I \times R$ scoring format, the source with the greatest variance was the PI, and the source with the lowest variance was the PR; and in the $P \times (I : R)$ scoring format, the source with the greatest variance was the PI: R, and the source with the lowest variance was then PR; 3) The scoring format with

the highest referencing coefficient of summation was the P x I x R format, and the results of 3 raters were more reliable than 2 raters.

Keywords: rater agreement index, intraclass correlation coefficient, Generalizability Theory

บทนำ

คณิตศาสตร์มีบทบาทสำคัญยิ่งต่อความสำเร็จในการเรียนรู้ในศตวรรษที่ 21 ซึ่งมีรายละเอียดตัวชี้วัดและสาระการเรียนรู้แกนกลาง กลุ่มสาระการเรียนรู้คณิตศาสตร์ (ฉบับปรับปรุง พ.ศ.2560) ระบุไว้หลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน พุทธศักราช 2551 โดยหลักสูตรได้คำนึงถึงการส่งเสริมให้ผู้เรียนมีทักษะที่จำเป็นสำหรับการเรียนรู้ในศตวรรษที่ 21 เป็นสำคัญ และ Programme for International Student Assessment (PISA) ได้นิยามความฉลาดรู้ด้านคณิตศาสตร์ของ PISA 2021 ไว้ว่า ความสามารถของแต่ละบุคคลในการให้เหตุผลทางคณิตศาสตร์ และสามารถแปลงปัญหา ใช้คณิตศาสตร์ และตีความผลลัพธ์ทางคณิตศาสตร์เพื่อใช้ในการแก้ปัญหาในบริบทของโลกชีวิตจริง รวมถึงการใช้แนวคิด กระบวนการ ข้อเท็จจริง และเครื่องมือทางคณิตศาสตร์เพื่อบรรยาย อธิบาย และคาดการณ์ปรากฏการณ์ต่าง ๆ การประเมินความฉลาดรู้ด้านคณิตศาสตร์ของ PISA 2021 มีลักษณะสำคัญคือเป็นการใช้คณิตศาสตร์ในการแก้ปัญหาจากสถานการณ์และบริบทต่าง ๆ ที่เกิดขึ้นจริงในโลก โดยผนวกการให้เหตุผลเชิงคณิตศาสตร์กับการแก้ปัญหาโดยใช้หลักการ กระบวนการ ข้อเท็จจริง และเครื่องมือทางคณิตศาสตร์ไว้ด้วยกัน เพื่อนำไปใช้ในการอธิบาย การให้เหตุผล และการคาดการณ์ปรากฏการณ์ต่าง ๆ นอกจากนี้การประเมิน PISA 2021 เป็นการประเมินที่เน้นความฉลาดรู้ด้านคณิตศาสตร์ (mathematical literacy) เพื่อให้นักเรียนอายุ 15 ปี ได้เห็นความสำคัญของคณิตศาสตร์และเป็นการสร้างประสบการณ์ให้นักเรียนได้พบเจอสถานการณ์และบริบทที่เกิดขึ้นในชีวิตจริง โดยวัดกระบวนการคิดทางคณิตศาสตร์ของนักเรียนผ่านสถานการณ์และบริบทในชีวิตจริง ซึ่งแนวคิดดังกล่าวยังเป็นแกนหลักในการวัดความฉลาดรู้ด้านคณิตศาสตร์

PISA ยังคำนึงถึงการเปลี่ยนแปลงอย่างรวดเร็วของเทคโนโลยีใหม่ ๆ ที่เข้ามามีบทบาทในชีวิตมากขึ้น และการพัฒนาทักษะที่จำเป็นสำหรับศตวรรษที่ 21 ด้วยเหตุนี้จึงทำให้มีการเน้นความสามารถในการให้เหตุผลทางคณิตศาสตร์และการคิดเชิงคำนวณ (computational thinking) ซึ่งเป็นกระบวนการในการแก้ปัญหา การคิดวิเคราะห์อย่างมีเหตุผลเป็นขั้นตอน รวมถึงบริบทที่สอดคล้องกับทักษะแห่งศตวรรษที่ 21 เข้ามาเป็นส่วนหนึ่งของกรอบการประเมินนี้ โดยกรอบโครงสร้างการประเมินผลการรู้เรื่องคณิตศาสตร์ PISA 2021 เป็นการสรุปแนวทางในการประเมินและแนวทางการสร้างข้อสอบ ซึ่งจะเชื่อมโยงกับ 3 องค์ประกอบที่เกี่ยวข้องกันได้แก่ 1) การให้เหตุผลทางคณิตศาสตร์ 2) เนื้อหาคณิตศาสตร์ และ 3) บริบทที่ใช้ในแบบทดสอบซึ่งสัมพันธ์กับทักษะแห่งศตวรรษที่ 21 (Institute for the Promotion of Teaching Science and Technology, 2020) ดังนั้นข้อสอบจึงสร้างขึ้นบนพื้นฐานของบริบทที่เกี่ยวข้องกับปัญหาที่ตั้งขึ้น และออกแบบให้ครอบคลุมสาระตามแนวคณิตศาสตร์เป็นเรื่อง ๆ นักเรียนนำคณิตศาสตร์เข้าไปมีส่วนร่วมในการแก้ปัญหาหรือสร้างความท้าทายที่เกิดขึ้นในโลกที่เป็นบริบทของชีวิตจริง ซึ่งในแต่ละกระบวนการแก้ปัญหาต้องอาศัยการให้เหตุผล

ทางคณิตศาสตร์ทั้งสิ้น ดังนั้นในการตรวจให้คะแนนข้อสอบนี้จึงต้องสร้างเกณฑ์ไว้ให้ดี เพื่อให้ผู้ตรวจจะได้มีแนวการตรวจที่ตรงกัน โดยอาศัยเกณฑ์การตรวจให้คะแนน (Rubric Scoring) ซึ่งเกณฑ์การตรวจให้คะแนน (rubric scoring) เป็นรูปแบบของการให้คะแนนที่เกิดจากการการรวมกันระหว่างเกณฑ์การให้คะแนนกับมาตราประมาณค่าหรือระดับคะแนน เพื่อระบุถึงความแตกต่างของคำตอบจากการทำแบบทดสอบ ซึ่งสามารถช่วยลดปัญหาเกี่ยวกับเวลาที่ใช้ในการตรวจลงได้ โดยเกณฑ์การให้คะแนนมี 2 รูปแบบที่นิยมและเป็นที่ยอมรับ คือ เกณฑ์การตรวจให้คะแนนแบบภาพรวมหรือแบบองค์รวม (holistic rubric) และเกณฑ์การประเมินแบบแยกส่วนหรือแบบแยกองค์ประกอบ (analytic rubric) ข้อดีสำหรับการตรวจให้คะแนนโดยใช้เกณฑ์แบบองค์รวม (holistic scoring rubric) คือใช้เวลาน้อยกว่าการใช้เกณฑ์การให้คะแนนแบบแยกองค์ประกอบ (analytic scoring rubric) (Tangdhanakanond, 2014) และเป็นการตรวจให้คะแนนโดยพิจารณาภาพรวมทั้งหมดของการตอบคำถามของผู้สอบ เปิดโอกาสให้มีการพิจารณาการตอบของผู้เข้าสอบได้อย่างรวดเร็ว ใช้เวลาน้อยในการตรวจ เหมาะกับการตรวจและประเมินคำตอบของผู้สอบที่ใช้ความคิดสร้างสรรค์มากกว่าการมีคำตอบที่ถูกต้อง ชัดเจน แน่นนอน แต่ข้อด้อยคือ ไม่สามารถวิเคราะห์ จุดแข็ง-จุดอ่อนของผู้เรียนได้ ซึ่งวิธีการที่จะลดความคลาดเคลื่อนของผลการตรวจให้คะแนนจากแบบสอบอัตนัย นอกจากวิธีการฝึกอบรมและชี้แจงเกณฑ์แก่ผู้ตรวจให้คะแนน และติดตามผลการอบรมและชี้แจงเกณฑ์ก่อนการตรวจให้คะแนน พบว่ามีการนำวิธีการทางสถิติมาช่วยในการตรวจสอบแหล่งความคลาดเคลื่อน หรือความแปรปรวนจากแหล่งต่างๆ ที่เกิดจากการวัดผลด้วย โดยเฉพาะความคลาดเคลื่อนจากข้อสอบและผู้ตรวจ ซึ่งการออกแบบการวัดในทฤษฎี G-Theory เป็นการออกแบบสถานการณ์ในการวัดให้มีเงื่อนไขหรือพาเซท ตามที่ผู้ประเมินสนใจ และคำนึงถึงแหล่งความคลาดเคลื่อนที่เป็นระบบอย่างน้อย 1 แหล่ง โดยระดับความน่าเชื่อถือของผลการวัดที่ได้มีชื่อเรียกเฉพาะว่าสัมประสิทธิ์การสรุปอ้างอิงความน่าเชื่อถือของผลการวัด (generalizability coefficient)

ความรู้ทฤษฎี G-Theory สามารถบอกได้ว่า ความแปรปรวนหรือความคลาดเคลื่อนของการวัดเกิดจากแหล่งใดมากน้อย โดยแหล่งของความคลาดเคลื่อนจะแตกต่างกันออกไปตามการออกแบบของผู้วิจัย ทั้งรูปแบบไขว้ (crossed) แบบแฝง (nested) หรือแบบแฝงบางส่วน (partially nested) ทำให้สามารถควบคุมได้ตรงจุด และประมาณค่าสัมประสิทธิ์การสรุปอ้างอิง (G-Coefficient) ที่เทียบเท่าสัมประสิทธิ์ความเชื่อมั่นหรือสัมประสิทธิ์ความเที่ยง (reliability) ในทฤษฎี CTT โดยมุ่งเน้นไปที่องค์ประกอบความแปรปรวนที่ส่งผลต่อความคลาดเคลื่อนในการวัด โดยแยกพิจารณาเพื่อตัดสินใจได้ใน 2 ลักษณะ คือการตัดสินใจเชิงสัมบูรณ์ (absolute decision) และการตัดสินใจเชิงสัมพัทธ์ (relative decision) ขึ้นอยู่กับการวางแผนการทดสอบว่าอิงกลุ่มหรืออิงเกณฑ์ แยกการวิเคราะห์เป็น 2 ขั้นตอนคือ ขั้นที่ 1 การศึกษาเพื่อการสรุปอ้างอิง (G-study) และขั้นที่ 2 การศึกษาเพื่อการตัดสินใจ (D-study) เป็นการใช้ข้อมูลจาก G-study ออกแบบการวัดที่ทำให้ความคลาดเคลื่อนน้อยที่สุดและให้ค่าสัมประสิทธิ์การสรุปอ้างอิง (G-coefficient) สูง เหมาะสม คำนวณค่าทั้งเวลาและค่าใช้จ่าย ตลอดจนเป็นไปได้กับสถานการณ์จริง (Kanjawasee, 2012) โดยการนำองค์ความรู้เกี่ยวกับองค์ประกอบความแปรปรวน รูปแบบความสัมพันธ์ และสัมประสิทธิ์การสรุปอ้างอิง ในทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัด มาออกแบบรูปแบบการตรวจให้คะแนนแบบสอบอัตนัยวิชาคณิตศาสตร์ เพื่อควบคุมแหล่งความคลาดเคลื่อน ค้นหาแบบการตรวจที่มีความเหมาะสม และมีความเป็นไปได้ในทางปฏิบัติ

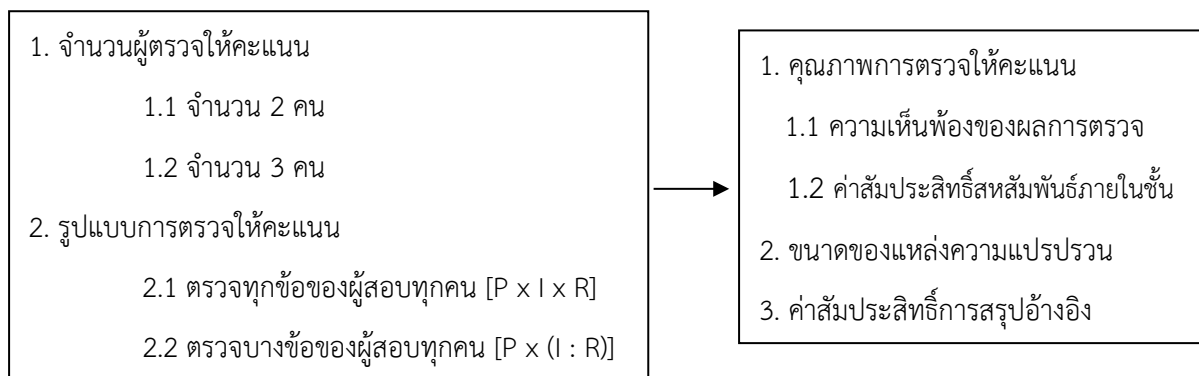
จากที่กล่าวมาข้างต้น ผู้วิจัยพบว่าแหล่งความคลาดเคลื่อนที่เกิดจากการตรวจแบบสอบการเขียนนั้นเกิดจากหลายแหล่ง โดยผู้วิจัยสนใจที่จะศึกษาขนาดของความแปรปรวนขององค์ประกอบเพื่อที่จะสามารถใช้ควบคุมแหล่งความคลาดเคลื่อนดังกล่าว โดยมีการกำหนดปัจจัยต่าง ๆ สำหรับเงื่อนไขและสถานการณ์การวัด ได้แก่ จำนวนผู้ตรวจ และรูปแบบการตรวจ เพื่อที่จะศึกษาคุณภาพการตรวจให้คะแนนและผลการเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิงเมื่อมีการควบคุมปัจจัยดังกล่าว ซึ่งก่อให้เกิดประโยชน์ต่อผู้บริหาร ครู ตลอดจนผู้ที่เกี่ยวข้องที่จะได้นำชุดเครื่องมือไปใช้วัดความสามารถในการแก้ปัญหาทางคณิตศาสตร์ของนักเรียน และนำไปพัฒนาชุดเครื่องมือวัดทักษะกระบวนการด้านอื่น ๆ ต่อไป

ความมุ่งหมายของการวิจัย

1. เพื่อศึกษาคุณภาพการตรวจให้คะแนนแบบสอบอัตนัยวัดความสามารถในการแก้ปัญหาทางคณิตศาสตร์ของผู้ตรวจให้คะแนนภายใต้รูปแบบการตรวจให้คะแนนต่างกัน
2. เพื่อศึกษาขนาดของแหล่งความแปรปรวนขององค์ประกอบและค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนแบบสอบอัตนัยวัดความสามารถในการแก้ปัญหาทางคณิตศาสตร์ เมื่อจำนวนผู้ตรวจ และรูปแบบการตรวจให้คะแนนแตกต่างกัน
3. เพื่อเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิง (G-Coefficient) ของแบบสอบอัตนัยวัดความสามารถในการแก้ปัญหาทางคณิตศาสตร์ ภายใต้รูปแบบการตรวจให้คะแนนต่างกัน และจำนวนผู้ตรวจต่างกัน

กรอบแนวคิดการวิจัย

จากการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง สามารถกำหนดกรอบแนวคิดการวิจัยได้ดังภาพ 1



ภาพ 1 กรอบแนวคิดการวิจัย

วิธีดำเนินการวิจัย

1. ประชากรและกลุ่มตัวอย่าง แบ่งเป็น 2 กลุ่ม ได้แก่

1.1 ประชากรและกลุ่มตัวอย่าง ที่เป็นนักเรียนที่กำลังศึกษาระดับชั้นมัธยมศึกษาปีที่ 3 สำนักงานเขตพื้นที่การศึกษามัธยมศึกษาขอนแก่น ได้มาจากการสุ่ม 2 ขั้นตอน คือ ขั้นตอนที่ 1 ทำการสุ่มโรงเรียน จากขนาดเล็ก กลาง และใหญ่ มาขนาดละ 1 โรงเรียน และขั้นตอนที่ 2 ทำการสุ่มห้องเรียน จึงได้กลุ่มตัวอย่างจากโรงเรียน 4 ขนาด ได้นักเรียนที่เป็นกลุ่มตัวอย่างโรงเรียนละ 33 คน รวมจำนวน 132 คน

1.2 ประชากรและกลุ่มตัวอย่าง ที่เป็นผู้ตรวจให้คะแนนแบบสอบอัตนัยวัดความสามารถในการแก้ปัญหาทางคณิตศาสตร์ซึ่งเป็นครูผู้สอนวิชาคณิตศาสตร์ ที่มีคุณวุฒิในการจบทางด้านการสอนคณิตศาสตร์โดยตรง ซึ่งมีประสบการณ์การทำงานไม่น้อยกว่า 3 ปี ในโรงเรียนในสังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน ซึ่งผู้วิจัยทำการสุ่มครูผู้ตรวจโดยใช้วิธีการเลือกกลุ่มตัวอย่างแบบเจาะจง (purposive sampling) จำนวน 3 คน

โดยขนาดของกลุ่มตัวอย่างที่สุ่มมาสำหรับใช้ในการศึกษามีขนาด $n_p \times n_i \times n_r$ เท่ากับ $132 \times 3 \times 3$ ได้ค่าสังเกต 1,188 ค่า ผ่านเกณฑ์การกำหนดขนาดของกลุ่มตัวอย่างอย่างน้อย 1,080 ค่าสังเกต ตามข้อเสนอของ Smith (1978, as cited in Buasiripan, 2000)

2. เครื่องมือที่ใช้ในการวิจัย

เครื่องมือที่ใช้ในการวิจัย เป็นแบบสอบอัตนัยวัดความสามารถในการแก้ปัญหาทางคณิตศาสตร์สำหรับนักเรียนชั้นมัธยมศึกษาปีที่ 3 ประกอบไปด้วยข้อสอบจำนวน 3 ข้อ การสร้างข้อคำถามจะสอดคล้องกับเนื้อหาวิชาคณิตศาสตร์ระดับชั้นมัธยมศึกษาปีที่ 3 ตามหลักสูตรแกนกลางการศึกษาขั้นพื้นฐานพุทธศักราช 2551 (ฉบับปรับปรุง พ.ศ. 2560) ที่มีบริบทหรือสถานการณ์ปัญหาที่สอดคล้องกับการเปลี่ยนแปลงในปัจจุบัน และสะท้อนความสามารถให้เหตุผลทางคณิตศาสตร์ของผู้สอบผ่านกระบวนการแก้ปัญหา แนวทางการสร้างข้อสอบตามกรอบการประเมินความฉลาดรู้ด้านคณิตศาสตร์ของ PISA จะเชื่อมโยงกับ 3 องค์ประกอบที่เกี่ยวข้องกัน ได้แก่ 1) การให้เหตุผลทางคณิตศาสตร์ 2) เนื้อหาคณิตศาสตร์ และ 3) บริบทที่ใช้ในแบบทดสอบ ซึ่งสัมพันธ์กับทักษะแห่งศตวรรษที่ 21 ดังนั้นข้อสอบจึงมีรูปแบบเป็นข้อคำถามที่มีคำตอบถูกเพียงคำตอบเดียว แต่อาจมีหลากหลายวิธีการได้มาซึ่งคำตอบ สร้างขึ้นบนพื้นฐานของบริบทที่เกี่ยวข้องกับปัญหาที่ตั้งขึ้น และออกแบบให้ครอบคลุมสาระตามแนวคณิตศาสตร์เป็นเรื่อง ๆ นักเรียนนำคณิตศาสตร์เข้าไปมีส่วนร่วมในการแก้ปัญหาหรือสร้างความท้าทายที่เกิดขึ้นในโลกที่เป็นบริบทของชีวิตจริง ซึ่งในแต่ละกระบวนการแก้ปัญหาต้องอาศัยการให้เหตุผลทางคณิตศาสตร์ กระบวนการหรือเทคนิคเฉพาะที่นักเรียนต้องอาศัยความรู้ มโนคติ การคิดวิเคราะห์ ประสบการณ์ และทักษะพื้นฐานต่าง ๆ ที่มีอยู่ไปประยุกต์ใช้ ตลอดจนการคิดหาแนวทางปฏิบัติเพื่อให้ปัญหาได้บรรลุตามจุดมุ่งหมายหรือคำตอบของปัญหานั้น ๆ โดยให้นักเรียนเขียนอธิบาย แสดงเหตุผลหรือแสดงวิธีการหาคำตอบที่มีวิธีการหรือขั้นตอนที่ชัดเจน ซึ่งมีค่าความตรงเชิงเนื้อหา (IOC) 0.80 – 1.00 มีค่าตั้งแต่ ค่าความยากตั้งแต่ 0.22 – 0.69 และค่าอำนาจจำแนกตั้งแต่ 0.56 – 0.80 และผลการตรวจสอบความสอดคล้องของความสอดคล้องของแบบสอบกับเกณฑ์การตรวจให้คะแนนแบบองค์รวม (holistic method) โดยผู้เชี่ยวชาญ พบว่า แนวคำตอบในแต่ละข้อคำถามของแต่ละข้อ และวิธีการให้คะแนนตามเกณฑ์การตรวจให้คะแนนแบบองค์รวม (holistic method) ที่ผู้วิจัยพัฒนาขึ้นทั้ง 3 ข้อ มีระดับความสอดคล้องกับข้อคำถามเมื่อเทียบกับเกณฑ์มีค่ามากกว่า 0.5 แสดงว่าเกณฑ์การตรวจให้คะแนนแบบองค์รวม (holistic method) สามารถนำมาใช้ตรวจให้คะแนนได้

ตาราง 1 แผนผังข้อสอบ (test blueprint) แบบสอบอัตนัยวัดความสามารถในการแก้ปัญหาทางคณิตศาสตร์
ระดับชั้นมัธยมศึกษาปีที่ 3

สาระ	มาตรฐาน	ตัวชี้วัด	สาระการเรียนรู้แกนกลาง	สถานการณ์/ คำถาม	ค่าความ ยาก	ค่าอำนาจ จำแนก
สาระที่ 1 จำนวน และ พีชคณิต	ค 1.2 เข้าใจและ วิเคราะห์แบบรูป ความสัมพันธ์ ฟังก์ชัน ลำดับและ อนุกรม และนำไปใช้	เข้าใจและใช้ ความรู้เกี่ยวกับ ฟังก์ชันกำลังสอง ในการแก้ปัญหา คณิตศาสตร์	ฟังก์ชันกำลังสอง ■ กราฟของฟังก์ชันกำลังสอง ■ การนำความรู้เกี่ยวกับ ฟังก์ชันกำลังสองไปใช้ในการแก้ ปัญหา	1) กล่องใส่พิชชา	0.69	0.80
สาระที่ 2 การวัด และ เรขาคณิต	มาตรฐาน ค 2.1 เข้าใจพื้นฐาน เกี่ยวกับการวัด วัด และคาดคะเนขนาด ของสิ่งที่ต้องการวัด และนำไปใช้	ประยุกต์ใช้ความรู้ เรื่องปริมาตรของ พีระมิด กรวย และ ทรงกลม ในการ แก้ปัญหา คณิตศาสตร์และ ปัญหาในชีวิตจริง	ปริมาตร ■ การหาปริมาตรของพีระมิด กรวย และทรงกลม ■ การนำความรู้เกี่ยวกับ ปริมาตรของพีระมิด กรวย และ ทรงกลม ไปใช้ในการแก้ปัญหา	2) รัศมีเป็น เท่าไร	0.27	0.56
สาระที่ 3 สถิติและ ความ น่าจะเป็น	มาตรฐาน ค 3.2 เข้าใจหลักการ นับเบื้องต้น ความ น่าจะเป็น และนำไป ใช้	เข้าใจเกี่ยวกับการ ทดลองสุ่มและนำ ผลที่ได้ไปหาความ น่าจะเป็นของ เหตุการณ์	ความน่าจะเป็น ■ เหตุการณ์จากการทดลองสุ่ม ■ ความน่าจะเป็น ■ การนำความรู้เกี่ยวกับ ความน่าจะเป็นไปใช้ในชีวิตจริง	3) ลง หรือ ไม่ลง	0.22	0.64

ตัวอย่าง สถานการณ์และคำถาม : รัศมีเป็นเท่าไร

แม่ค้าขายขนมบัวลอยนวดแป้งและส่วนผสมอื่น ๆ เข้าด้วยกัน ได้ขนาดเป็นทรงกลมตันรัศมี 6 เซนติเมตร หากแบ่งแป้งที่ผสมเรียบร้อยแล้วเป็นก้อนกลม 3 ลูกต่างขนาดกัน โดยลูกที่หนึ่งเป็นทรงกลมรัศมี 5 เซนติเมตร ลูกที่สองเป็นทรงกลมรัศมี 3 เซนติเมตร อยากทราบว่าลูกที่สามจะมีรัศมียาวกี่เซนติเมตร

ตาราง 2 ตัวอย่างเกณฑ์การตรวจให้คะแนนแบบองค์รวม

คะแนน	รูปแบบการตอบ
0	- ไม่เขียนข้อความหรือสัญลักษณ์ใด ๆ
1	- เขียนข้อความหรือสัญลักษณ์แต่ไม่เกี่ยวข้องกับประเด็นปัญหาหรือสถานการณ์ที่กำหนด
2	- มีการเขียนข้อความเพียงเล็กน้อย ไม่เพียงพอที่จะทำให้เข้าใจ และมีแนวโน้มไม่นำไปสู่คำตอบ
3	- มีการแสดงเหตุผลที่เลือกตัดสินใจ แต่ไม่มีการเขียนแสดงกระบวนการแก้ปัญหาและการให้เหตุผลทางคณิตศาสตร์ และไม่มีการสรุปคำตอบที่ชัดเจน
4	- มีการเขียนสรุปคำตอบที่ชัดเจน ถูกต้อง แต่ไม่มีสอดคล้องกับเหตุผลที่เลือกใช้ในกระบวนการแก้ปัญหาและการให้เหตุผลทางคณิตศาสตร์ที่ชัดเจน
5	- มีการเขียนสรุปคำตอบที่ชัดเจน ถูกต้อง แสดงเหตุผลที่เลือกตัดสินใจ แต่ไม่มีการเขียนแสดงกระบวนการแก้ปัญหาและการให้เหตุผลทางคณิตศาสตร์ที่ยังไม่มีความชัดเจน

ตาราง 2 (ต่อ)

คะแนน	รูปแบบการตอบ
6	- มีการเขียนคำตอบที่ชัดเจน ถูกต้อง พร้อมแสดงเหตุผลที่เลือกตัดสินใจ มีการเขียนแสดงกระบวนการแก้ปัญหา แต่การให้เหตุผลทางคณิตศาสตร์ยังไม่มีชัดเจน
7	- มีการเขียนคำตอบที่ชัดเจน ถูกต้อง พร้อมแสดงเหตุผลที่เลือกตัดสินใจ มีการเขียนแสดงกระบวนการแก้ปัญหา การให้เหตุผลทางคณิตศาสตร์มีความชัดเจน แนวทางการตอบแบบที่ 1 สูตรการหาพื้นที่รูปสี่เหลี่ยมจัตุรัส คือ ความยาวด้าน x ความยาวด้าน จะได้ ความยาวด้าน x ความยาวด้าน = 256 ความยาวด้าน x ความยาวด้าน = 16 x 16 ความยาวด้าน = 16 เซนติเมตร ดังนั้น กล่องใส่พิซซ่า ควรมีเส้นผ่านศูนย์กลางไม่เกิน 16 นิ้ว แนวทางการตอบแบบที่ 2 พื้นที่กล่องเป็นรูปสี่เหลี่ยมจัตุรัส = (x)(x) $256 = x^2$ $\sqrt{256} = x$ $x = 16$ ดังนั้น กล่องใส่พิซซ่า ควรมีเส้นผ่านศูนย์กลางไม่เกิน 16 นิ้ว

3. การเก็บรวบรวมข้อมูล

1. ผู้วิจัยนำเป็นแบบสอบถามวัดความสามารถในการแก้ปัญหาทางคณิตศาสตร์ที่สร้างขึ้น และขอความอนุเคราะห์โรงเรียนที่เป็นกลุ่มตัวอย่าง ดำเนินการสอบตามวัน และเวลาที่กำหนด

2. การตรวจให้คะแนน ผู้วิจัยได้สร้างเกณฑ์การตรวจให้คะแนนแบบวิธีแบบองค์รวม (holistic rubric) โดยผู้ตรวจให้คะแนนจะแบ่งคะแนนของแต่ละข้อออกเป็นส่วนๆ มีรูปแบบผู้ตรวจให้คะแนน จำนวน 2 คน และ 3 คน โดยก่อนที่ผู้ตรวจจะทำการตรวจแบบสอบ ผู้วิจัยได้อธิบายการตรวจและเกณฑ์การตรวจให้คะแนน รวมถึงอธิบายตัวอย่างหรือแนวคำตอบในการตรวจให้คะแนน เพื่อลดความคลาดเคลื่อนอันเนื่องมาจากการให้คะแนน พร้อมทั้งให้คำแนะนำในการตรวจให้คะแนนและเมื่อผู้ตรวจตรวจให้คะแนนจะต้องบันทึกคะแนนลงในแบบบันทึกคะแนนที่ผู้วิจัยจัดทำขึ้น

4. การวิเคราะห์ข้อมูล

1. เพื่อวิเคราะห์คุณภาพการตรวจให้คะแนนแบบสอบถามวัดความสามารถในการแก้ปัญหาทางคณิตศาสตร์ของผู้ตรวจให้คะแนนภายใต้รูปแบบการตรวจให้คะแนนต่างกัน ได้แก่ 1.1 การวิเคราะห์ร้อยละความเห็นพ้อง (rater agreement) ของผลการตรวจระหว่างผู้ตรวจเทียบกับคะแนนเกณฑ์มาตรฐาน หาได้จากการสร้างตารางแจกแจงความถี่แบบสองทางโดยโปรแกรม SPSS ผ่านเมนู analyze -> descriptive statistics -> crosstabs ทำการตรวจสอบความถี่ของผลการตรวจให้คะแนนที่ตรงกันและทำการคำนวณในรูปแบบของร้อยละ แปลความหมายของค่าสถิติร้อยละ และ 1.2 การวิเคราะห์สถิติสหสัมพันธ์ภายในชั้น (Intraclass Correlation Coefficient: ICC) โดยแปลผลตามเกณฑ์ของ Koo & Li (2016)

2. เพื่อวิเคราะห์ขนาดของแหล่งความแปรปรวนขององค์ประกอบและค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนแบบสอบอัตนัยวัดความสามารถในการแก้ปัญหาทางคณิตศาสตร์ เมื่อจำนวนผู้ตรวจ และรูปแบบการตรวจให้คะแนนแตกต่างกัน โดยการวิเคราะห์หาขนาดของแหล่งความแปรปรวน ได้แก่ ผู้สอบ ข้อสอบ และผู้ตรวจ โดยใช้โปรแกรม EduG

3. เพื่อเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิง (G-Coefficient) ของแบบสอบอัตนัยวัดความสามารถในการแก้ปัญหาทางคณิตศาสตร์ ภายใต้รูปแบบการตรวจให้คะแนนต่างกัน และจำนวนผู้ตรวจต่างกัน โดยทำการวิเคราะห์ค่าสัมประสิทธิ์การสรุปอ้างอิงสำหรับการตัดสินใจเชิงสัมบูรณ์และค่าสัมประสิทธิ์การสรุปอ้างอิงสำหรับการตัดสินใจเชิงสัมพัทธ์ ภายใต้เงื่อนไขในการตรวจให้คะแนนระหว่างผู้ตรวจ 2 คน และ 3 คน ที่มีรูปแบบการตรวจ 2 รูปแบบ ได้แก่ 1. ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน [$P \times I \times R$ Design] 2. ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน [$P \times (I : R)$ Design] โดยใช้โปรแกรม EduG

ผลการวิจัย

1. คุณภาพของการตรวจให้คะแนนแบบสอบอัตนัยวัดความสามารถในการแก้ปัญหาทางคณิตศาสตร์

1.1 ผลการวิเคราะห์ความเห็นพ้อง (rater agreement) ของผลการตรวจระหว่างผู้ตรวจ เทียบกับคะแนนเกณฑ์มาตรฐานที่ได้จากฉันทามติในการตรวจให้คะแนนจากผู้เชี่ยวชาญ แสดงได้ดังตาราง 3

ตาราง 3 ร้อยละความเห็นพ้องของผลการตรวจระหว่างผู้ตรวจเทียบกับคะแนนเกณฑ์มาตรฐาน

ข้อสอบ	ผู้ตรวจ	ร้อยละความเห็นพ้อง
ข้อ 1	คนที่ 1	54.55
	คนที่ 2	55.30
	คนที่ 3	26.52
ข้อ 2	คนที่ 1	44.70
	คนที่ 2	55.30
	คนที่ 3	32.58
ข้อ 3	คนที่ 1	46.21
	คนที่ 2	45.45
	คนที่ 3	34.09

จากตาราง 3 พบว่า ร้อยละความเห็นพ้องของผลการตรวจผู้ตรวจให้คะแนนคนที่ 1 มีความเห็นพ้องกับเกณฑ์การตรวจข้อที่ 1 ข้อที่ 2 และ ข้อที่ 3 เท่ากับ ร้อยละ 54.55, 44.70, 46.21 ตามลำดับ ร้อยละความเห็นพ้อง ของผลการตรวจผู้ตรวจให้คะแนนคนที่ 2 มีความเห็นพ้องกับเกณฑ์การตรวจข้อที่ 1 ข้อที่ 2 และ ข้อที่ 3 เท่ากับ ร้อยละ 55.30, 55.30, 45.45 ตามลำดับ และร้อยละความเห็นพ้องของผลการตรวจผู้ตรวจให้คะแนนคนที่ 3 มีความเห็นพ้องกับเกณฑ์การตรวจข้อที่ 1 ข้อที่ 2 และ ข้อที่ 3 เท่ากับ ร้อยละ 26.52, 32.58, 34.09 ตามลำดับ ซึ่งผู้ตรวจให้คะแนนคนที่ 1 มีความเห็นพ้องในการตรวจแบบสอบเมื่อเทียบกับเกณฑ์การตรวจให้คะแนนข้อที่ 3 มากที่สุด ผู้ตรวจให้คะแนนคนที่ 2 มีความเห็นพ้องในการตรวจแบบสอบเมื่อเทียบ

กับเกณฑ์การตรวจให้คะแนนข้อที่ 1 และ 2 มากที่สุด ส่วนผู้ตรวจให้คะแนนคนที่ 3 มีความเห็นพ้องในการตรวจแบบสอบเมื่อเทียบกับเกณฑ์การตรวจให้คะแนนทั้ง 3 ข้อ น้อยที่สุด

1.2 ผลการวิเคราะห์การวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ภายในชั้น (Intraclass Correlation Coefficient: ICC) แสดงได้ดังตาราง 4

ตาราง 4 ค่าสัมประสิทธิ์สหสัมพันธ์ภายในชั้น (Intraclass Correlation Coefficient: ICC)

ข้อสอบ	วิธีคำนวณ	Intraclass Correlation Coefficient: ICC (95% CI)	ระดับ
ข้อ 1	รายบุคคล	0.86 (0.82 - 0.89)	ดี
	โดยเฉลี่ย	0.96 (0.95 - 0.97)	ดีมาก
ข้อ 2	รายบุคคล	0.78 (0.73 - 0.83)	ดี
	โดยเฉลี่ย	0.94 (0.92 - 0.95)	ดีมาก
ข้อ 3	รายบุคคล	0.95 (0.93 - 0.96)	ดีมาก
	โดยเฉลี่ย	0.99 (0.98 - 0.99)	ดีมาก

หมายเหตุ : 95% CI = 95% Confidence Interval

จากตาราง 4 พบว่า ค่าความเที่ยงในการตรวจให้คะแนนค่าความสอดคล้องระหว่างผู้ตรวจให้คะแนนจากค่าสัมประสิทธิ์สหสัมพันธ์ภายในชั้น (Intraclass Correlation Coefficient: ICC) ที่มีผู้ตรวจให้คะแนน 3 คน ที่ระดับความเชื่อมั่น 95% จำแนกเป็นรายข้อซึ่งค่า ICC จะอยู่ระหว่าง 0 ถึง 1 โดยค่า ICC มากกว่าหรือเท่ากับ 0.90 แต่ไม่เกิน 1.00 ถือว่ามีความน่าเชื่อถืออยู่ในระดับดีมาก ค่า ICC มากกว่าหรือเท่ากับ 0.75 ถือว่ามีความน่าเชื่อถืออยู่ในระดับดี ค่า ICC มากกว่าหรือเท่ากับ 0.50 แต่น้อยกว่า 0.75 ถือว่ามีความน่าเชื่อถืออยู่ในระดับพอใช้ และค่า ICC น้อยกว่า 0.50 ถือว่ามีความน่าเชื่อถืออยู่ในระดับต่ำ ตามเกณฑ์ของ Koo & Li (2016) ผลการวิเคราะห์พบว่า ความสอดคล้องระหว่างผู้ตรวจให้คะแนนแบ่งตามข้อคำถาม โดยข้อที่ 1 โดยเฉลี่ยมีค่า ICC = 0.96 (0.95 - 0.97) ข้อที่ 2 โดยเฉลี่ยมีค่า ICC = 0.94 (0.92 - 0.95) และ ข้อที่ 3 โดยเฉลี่ยมีค่า ICC = 0.99 (0.98 - 0.99) ซึ่งทั้ง 3 ข้อคำถามถือว่ามีความน่าเชื่อถืออยู่ในระดับดีมาก

2. ผลการวิเคราะห์หาขนาดของแหล่งความแปรปรวน ได้แก่ ผู้สอบ ข้อสอบ และผู้ตรวจ ดังตาราง 5

ตาราง 5 ความแปรปรวนและร้อยละของแหล่งความแปรปรวนในแต่ละรูปแบบการตรวจให้คะแนน

	P x I x R			P x (I : R)		
	DF	Var.	%	DF	Var.	%
P	131	0.313	4.1	131	1.538	18.9
I	2	2.139	28.1			
R	2	0.699	9.2	2	-0.014	28.3
PI	262	3.674	48.2			
PR	262	-0.008	0.0	262	-1.233	0.0
IR	4	0.166	2.2			
PIR	524	0.627	8.2			
I : R	-			6	2.305	28.3
PI : R	-			786	4.301	52.8
ผลรวม	1187			1187		

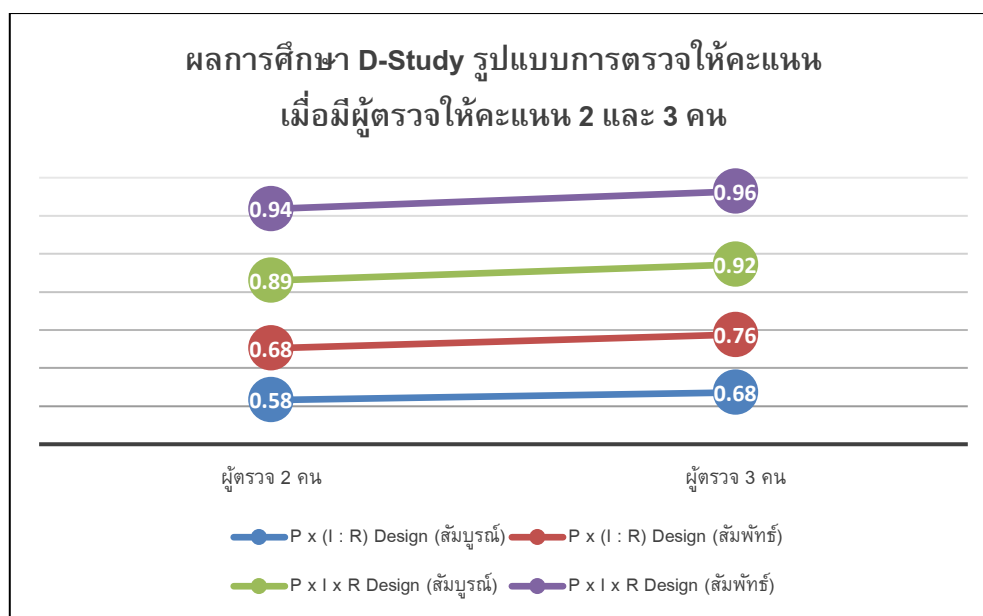
จากตาราง 5 การวิเคราะห์หาขนาดของแหล่งความแปรปรวนในแต่ละรูปแบบการตรวจให้คะแนน ผลการวิเคราะห์พบว่า รูปแบบการตรวจให้คะแนนแบบ $P \times I \times R$ แหล่งที่มีความแปรปรวนมากที่สุดคือ PI มีค่าเท่ากับ 3.674 และแหล่งความแปรปรวนที่มีค่าน้อยที่สุดคือ PR มีค่าเท่ากับ -0.008 และรูปแบบการตรวจให้คะแนนแบบ $P \times (I : R)$ แหล่งที่มีความแปรปรวนมากที่สุดคือ PI : R มีค่า 4.301 และแหล่งความแปรปรวนที่มีค่าน้อยที่สุดคือ PR มีค่าเท่ากับ -1.233

3. ผลการวิเคราะห์ค่าสัมประสิทธิ์การสรุปอ้างอิงสำหรับการตัดสินใจเชิงสัมบูรณ์และค่าสัมประสิทธิ์การสรุปอ้างอิงสำหรับการตัดสินใจเชิงสัมพัทธ์ ระหว่างผู้ตรวจ 2 คนและ 3 คน ที่มีรูปแบบการตรวจ 2 รูปแบบ ได้แก่ 1. ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน [$P \times I \times R$ Design] 2. ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน [$P \times (I : R)$ Design] มีผลดังตาราง 6

ตาราง 6 ผลการศึกษา D-Study รูปแบบการตรวจให้คะแนนเมื่อมีผู้ตรวจให้คะแนน 2 และ 3 คน

รูปแบบการตรวจ (Design)		จำนวนผู้ตรวจ (n_r)	
		2	3
$P \times I \times R$ Design	ρ_δ^2 (สัมพัทธ์)	0.94	0.96
	ρ_Δ^2 (สัมบูรณ์)	0.89	0.92
$P \times (I : R)$ Design	ρ_δ^2 (สัมพัทธ์)	0.68	0.76
	ρ_Δ^2 (สัมบูรณ์)	0.58	0.68

เพื่อแสดงให้เห็นการเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนสำหรับนำไปใช้ตัดสินใจเชิงสัมบูรณ์ (G – coefficient for absolute decisions) และค่าสัมประสิทธิ์การสรุปอ้างอิงสำหรับการตัดสินใจเชิงสัมพัทธ์ (relative coefficient) จึงได้ของคะแนนแบบสอบอัตนัยวัดความสามารถในการแก้ปัญหาทางคณิตศาสตร์ ตามรายละเอียดที่กล่าวมาแล้วข้างต้น ผู้วิจัยนำเสนอการเปรียบเทียบผลการศึกษา D-Study รูปแบบการตรวจให้คะแนนเป็นแผนภูมิแท่ง ดังแสดงในภาพ 2



ภาพ 2 ผลการศึกษา D-Study รูปแบบการตรวจให้คะแนน เมื่อมีผู้ตรวจให้คะแนน 2 และ 3 คน

จากตาราง 6 และภาพ 2 ผลการศึกษา D-study รูปแบบการตรวจให้คะแนนแบบ $P \times I \times R$ พบว่า เมื่อมีผู้ตรวจให้คะแนน 2 และ 3 คน มีสัมประสิทธิ์การสรุปอ้างอิงเชิงตัดสินใจสัมพัทธ์ (relative coefficient) คือ 0.94 , 0.96 ตามลำดับ และสัมประสิทธิ์การสรุปอ้างอิงเชิงตัดสินใจสัมบูรณ์ (absolute coefficient) คือ 0.89, 0.92 ตามลำดับ รูปแบบการตรวจให้คะแนนแบบ $P \times (I : R)$ พบว่า เมื่อมีผู้ตรวจให้คะแนน 2 และ 3 คน มีสัมประสิทธิ์การสรุปอ้างอิงเชิงตัดสินใจสัมพัทธ์ (relative coefficient) คือ 0.68, 0.76 ตามลำดับ และสัมประสิทธิ์การสรุปอ้างอิงเชิงตัดสินใจสัมบูรณ์ (absolute coefficient) คือ 0.58, 0.68 ตามลำดับ

อภิปรายผล

จากผลการวิจัย ผู้วิจัยนำเสนอประเด็นการอภิปราย ดังนี้

1. คุณภาพของการตรวจให้คะแนนแบบสอบอัตนัยวัดความสามารถในการแก้ปัญหาทางคณิตศาสตร์

1.1 ผู้ตรวจให้คะแนนคนที่ 1 มีความเห็นพ้องในการตรวจแบบสอบเมื่อเทียบกับคะแนนเกณฑ์มาตรฐานข้อที่ 3 มากที่สุด ผู้ตรวจให้คะแนนคนที่ 2 มีความเห็นพ้องในการตรวจแบบสอบเมื่อเทียบกับคะแนนเกณฑ์มาตรฐานข้อที่ 1 และ 2 มากที่สุด ส่วนผู้ตรวจให้คะแนนคนที่ 3 มีความเห็นพ้องในการตรวจแบบสอบเมื่อเทียบกับคะแนนเกณฑ์มาตรฐานทั้ง 3 ข้อ น้อยที่สุด เนื่องจากผู้ตรวจมีความเห็นไปในทิศทางที่ต่างกันในบางสถานการณ์คำถาม การชี้แจงเกณฑ์การตรวจให้คะแนนอาจยังไม่เพียงพอที่จะให้ผู้ตรวจให้คะแนนได้ใกล้เคียงหรือสอดคล้องกับคะแนนเกณฑ์มาตรฐาน สอดคล้องกับ Aphaikawee (2019) ที่กล่าวว่า สาเหตุของการแบบแผนการให้คะแนนที่แตกต่างกันนี้อาจมาจากคุณลักษณะของผู้ตรวจ เช่น ประสบการณ์การทำงาน การจบการศึกษาตรงตามสาขาวิชา การได้รับการฝึกอบรม เป็นต้น หากมีการควบคุมปัจจัยดังกล่าวข้างต้นจะทำให้ผู้ตรวจมีแบบแผนในการตรวจให้คะแนนและมีความสอดคล้องกับเกณฑ์มากขึ้น

1.2 ผลการวิเคราะห์ความสอดคล้องของผลการตรวจให้คะแนนแบบสอบอัตนัยวัดความสามารถในการแก้ปัญหาทางคณิตศาสตร์ของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ด้วยค่าสัมประสิทธิ์สหสัมพันธ์ภายในชั้น (Intraclass Correlation Coefficient: ICC) พบว่า ความเที่ยงในการตรวจให้คะแนนระหว่างผู้ตรวจ (inter rater reliability) มีความสอดคล้องกันในระดับดีมาก แสดงให้เห็นว่าผู้ตรวจให้คะแนนสอดคล้องไปในทิศทางเดียวกัน เนื่องมาจากการอธิบายการตรวจและเกณฑ์การตรวจให้คะแนน รวมถึงอธิบายตัวอย่างหรือแนวคำตอบในการตรวจให้คะแนน เพื่อลดความคลาดเคลื่อนอันเนื่องมาจากการให้คะแนน พร้อมทั้งให้คำแนะนำในการตรวจให้คะแนน ซึ่งสอดคล้องกับ Prasertsinponma & Tuksino (2021) ที่พบว่า ค่าความเที่ยงในการตรวจให้คะแนนค่าความสอดคล้องระหว่างผู้ตรวจให้คะแนนจากค่าสัมประสิทธิ์สหสัมพันธ์ภายในชั้น (Intraclass Correlation Coefficient: ICC) ที่มีผู้ตรวจให้คะแนน 3 คน ที่ระดับความเชื่อมั่น 95% จำแนกเป็นรายข้อ โดยข้อที่ 1 โดยเฉลี่ยมีค่า ICC = 0.96 (0.92-0.98) ข้อที่ 2 โดยเฉลี่ยมีค่า ICC = 0.97 (0.96-0.98) และ ข้อที่ 3 โดยเฉลี่ยมีค่า ICC = 0.97 (0.95-0.98) ซึ่งทั้ง 3 ข้อคำถามถือว่ามีความน่าเชื่อถืออยู่ในระดับดีมาก

2. ผลการวิเคราะห์แหล่งความแปรปรวนของแต่ละองค์ประกอบที่มีผลต่อการสัมประสิทธิ์การสรุปอ้างอิง (G-coefficient) ผลการประมาณค่าความแปรปรวนของทั้ง 3 องค์ประกอบ คือ ผู้สอบ (P) ข้อสอบ (I)

และผู้ตรวจ (R) เมื่อใช้รูปแบบการตรวจข้อสอบทุกข้อของผู้สอบทุกคน [$P \times I \times R$] พบว่า แหล่งความแปรปรวนที่มากที่สุด คือ แหล่งปฏิสัมพันธ์ระหว่างผู้สอบกับข้อสอบ (PI) แสดงว่าความสามารถของนักเรียนมีความแตกต่างกันและมีความถนัดในข้อสอบแต่ละข้อต่างกัน ซึ่งสอดคล้องกับ Bhu-iam (2007) ซึ่งกล่าวว่า ปฏิสัมพันธ์ระหว่างผู้สอบกับข้อสอบมีผลต่อนักเรียนเนื่องจากนักเรียนแต่ละคนมีความสามารถต่างกันทำให้เกิดความแปรปรวน และแหล่งที่ความแปรปรวนน้อยที่สุดคือ แหล่งปฏิสัมพันธ์ระหว่างผู้สอบและผู้ตรวจ (PR) รูปแบบการตรวจให้คะแนนแบบการตรวจข้อสอบบางข้อของผู้สอบทุกคน [$P \times (I : R)$] พบว่าแหล่งแปรปรวนมากที่สุดคือ PI : R แสดงว่า มีความแปรปรวนคงเหลือที่เกิดจากปฏิสัมพันธ์ระหว่างผู้สอบกับข้อสอบจากการตรวจของผู้สอบบางคนและยังไม่สามารถอธิบายได้ และแหล่งความแปรปรวนที่น้อยที่สุดคือปฏิสัมพันธ์ระหว่างผู้สอบกับผู้ตรวจ (PR) แสดงว่าผู้ตรวจสามารถตรวจให้คะแนนได้เหมาะสมกับกลุ่มตัวอย่างที่เป็นผู้สอบ และรูปแบบการตรวจให้คะแนนแบบการตรวจข้อสอบบางข้อของผู้สอบทุกคน [$P \times (I : R)$] มีความแปรปรวนที่มาจากผู้ตรวจมากกว่าความแปรปรวนจากผู้สอบ จะเห็นได้ว่า ซึ่งสอดคล้องกับ ผลการวิจัยของ Inthanate (2011); Chuithong (2011); Bhu-iam (2007) พบว่า ผู้ตรวจให้คะแนนเป็นแหล่งความคลาดเคลื่อนที่สำคัญที่ส่งผลต่อความเที่ยงของคะแนน และมีอิทธิพลต่อค่าสัมประสิทธิ์การสรุปอ้างอิง

3. ผลการประมาณค่าสัมประสิทธิ์การสรุปอ้างอิงของแบบสอบอัตนัยวัดความสามารถในการแก้ปัญหาทางคณิตศาสตร์ ที่มีการรูปแบบการตรวจให้คะแนน 2 ลักษณะ คือ ผู้ตรวจตรวจทุกข้อของผู้สอบทุกคน [$P \times I \times R$] และผู้ตรวจตรวจบางข้อของผู้สอบทุกคน [$P \times (I : R)$] สรุปผลได้ ดังนี้

3.1 รูปแบบการตรวจแบบผู้ตรวจตรวจทุกข้อของผู้สอบทุกคน [$P \times I \times R$] มีค่าสัมประสิทธิ์การสรุปอ้างอิงการตัดสินใจเชิงสัมพัทธ์ (relative coefficient) เท่ากับ 0.94 และ 0.96 เมื่อผู้ตรวจตรวจให้คะแนนแบบสอบอัตนัยวัดความสามารถในการแก้ปัญหาทางคณิตศาสตร์ 3 ข้อ จำนวน 2 และ 3 คน ตามลำดับ และมีค่าสัมประสิทธิ์การสรุปอ้างอิงการตัดสินใจเชิงสัมบูรณ์ (absolute coefficient) เท่ากับ 0.89 และ 0.92 เมื่อผู้ตรวจตรวจให้คะแนนแบบสอบอัตนัยวัดความสามารถในการแก้ปัญหาทางคณิตศาสตร์ 3 ข้อ จำนวน 2 และ 3 คน ตามลำดับ

3.2 รูปแบบการตรวจแบบผู้ตรวจตรวจบางข้อของผู้สอบทุกคน [$P \times (I : R)$] มีค่าสัมประสิทธิ์การสรุปอ้างอิงการตัดสินใจเชิงสัมพัทธ์ (relative coefficient) เท่ากับ 0.68 และ 0.76 เมื่อผู้ตรวจตรวจให้คะแนนแบบสอบอัตนัยวัดความสามารถในการแก้ปัญหาทางคณิตศาสตร์ 3 ข้อ จำนวน 2 และ 3 คน ตามลำดับ และมีค่าสัมประสิทธิ์การสรุปอ้างอิงการตัดสินใจเชิงสัมบูรณ์ (absolute coefficient) เท่ากับ 0.58 และ 0.68 เมื่อผู้ตรวจตรวจให้คะแนนแบบสอบอัตนัยวัดความสามารถในการแก้ปัญหาทางคณิตศาสตร์ 3 ข้อ จำนวน 2 และ 3 คน ตามลำดับ

จากการศึกษาเพื่อการตัดสินใจสรุปอ้างอิง (D-study) พบว่า 1) เจอนไขผู้ตรวจให้คะแนน 3 คน ส่งผลต่อสัมประสิทธิ์การสรุปอ้างอิงเชิงตัดสินใจสัมพัทธ์ (relative coefficient) และ สัมประสิทธิ์การสรุปอ้างอิงเชิงตัดสินใจสัมบูรณ์ (absolute coefficient) สูงกว่าผู้ตรวจให้คะแนน 2 คน เนื่องจากความเห็นพ้องกับคะแนนเกณฑ์มาตรฐานที่ต่างกัน และผลการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ภายในชั้น (ICC) พบว่า ความสอดคล้องของผู้ตรวจให้คะแนนรายบุคคล ผู้ตรวจบางคนอยู่ในระดับดี แต่ความสอดคล้องของผู้ตรวจ

ให้คะแนนโดยเฉลี่ยอยู่ในระดับดีมาก ผลการตรวจให้คะแนนจากผู้ตรวจ 2 คน จึงมีระดับความน่าเชื่อถือของผู้ตรวจให้คะแนนน้อยกว่าผลการตรวจให้คะแนนจากผู้ตรวจ 3 คน สอดคล้องกับ Taoto et al. (2016) กล่าวว่าในแต่ละรูปแบบการตรวจให้คะแนนที่เหมือนกัน เมื่อใช้จำนวนผู้ตรวจให้คะแนนมากขึ้น ค่าความเชื่อมั่นจะมีแนวโน้มเพิ่มมากขึ้น สอดคล้องกับผลการวิจัยของ Aphaikawi (2019) พบว่า ค่าสัมประสิทธิ์การสรุปอ้างอิงสำหรับตัดสินใจเชิงสมบูรณ์ของผลการให้คะแนน มีค่าเพิ่มขึ้นเมื่อผู้ตรวจตั้งแต่ 3 คนหรือจำนวนข้อสอบเพิ่มขึ้น 3 ข้อขึ้นไป จะทำให้ผลการตรวจยังมีความเที่ยงเพิ่มมากขึ้น และสอดคล้องกับ Thitikanpodchana & Tuksino (2021). ที่กล่าวว่าเมื่อจำนวนผู้ตรวจเพิ่มขึ้น ส่งผลต่อค่าสัมประสิทธิ์การสรุปอ้างอิงสำหรับการตัดสินใจเชิงสัมพัทธ์ (relative coefficient) และค่าสัมประสิทธิ์การสรุปอ้างอิงสำหรับการตัดสินใจเชิงสมบูรณ์ (absolute coefficient) และ 2) ผลการเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิงของแบบสอบอัตนัยวัดความสามารถในการแก้ปัญหาทางคณิตศาสตร์ ที่มีรูปแบบการตรวจที่ต่างกัน ภายใต้จำนวนข้อสอบและจำนวนผู้ตรวจต่างกันตามทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของการวัด จะได้ว่าเมื่อตรวจให้คะแนนด้วยรูปแบบการตรวจข้อสอบทุกข้อของผู้สอบทุกคน [$P \times I \times R$] มีค่าสัมประสิทธิ์การสรุปอ้างอิงการตัดสินใจเชิงสัมพัทธ์ (relative coefficient) และค่าสัมประสิทธิ์การสรุปอ้างอิงการตัดสินใจเชิงสมบูรณ์ (absolute coefficient) สูงกว่ารูปแบบการตรวจให้คะแนนแบบการตรวจข้อสอบบางข้อของผู้สอบทุกคน [$P \times (I : R)$] เนื่องจากรูปแบบการตรวจข้อสอบทุกข้อของผู้สอบทุกคน [$P \times I \times R$] รูปแบบนี้ผู้สอบทุกคนได้รับเงื่อนไขในการตรวจให้คะแนนจากผู้ตรวจคนเดียวกันเหมือนกันทุกคนอย่างเท่าเทียมกัน ส่วนรูปแบบการตรวจข้อสอบบางข้อของผู้สอบทุกคน [$P \times (I : R)$] รูปแบบนี้ผู้สอบแต่ละคนได้รับเงื่อนไขในการตรวจให้คะแนนจากผู้ตรวจต่างกัน โดยไม่ได้รับการตรวจจากผู้ตรวจทุกคน และอาจจะไม่ได้มีความถนัดในการตอบข้อสอบที่ได้รับการตรวจให้คะแนน รวมถึงขนาดของกลุ่มตัวอย่างที่เป็นผู้ตรวจให้คะแนนถือว่าค่อนข้างน้อย ทำให้มีความเห็นในการตรวจให้คะแนนต่างกัน สอดคล้องกับผลการวิเคราะห์ความเห็นพ้อง (rater agreement) ของผลการตรวจระหว่างผู้ตรวจ เทียบกับคะแนนเกณฑ์มาตรฐานที่ได้จากฉันทามติในการตรวจให้คะแนนจากผู้เชี่ยวชาญที่แสดงให้เห็นว่าผู้ตรวจให้คะแนนมีความเห็นพ้องที่ต่างกันไปใน การตรวจข้อสอบแต่ละข้อ และผลการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ภายในชั้น (ICC) พบว่า ความสอดคล้องระหว่างผู้ตรวจให้คะแนนแบ่งตามข้อคำถามทั้ง 3 ข้อคำถามถือว่ามีความน่าเชื่อถืออยู่ในระดับดีมาก ดังนั้น การตรวจให้คะแนนการทำแบบสอบจากกลุ่มตัวอย่างที่เป็นผู้สอบ ควรเลือกใช้รูปแบบการตรวจให้คะแนนแบบตรวจข้อสอบทุกข้อของผู้สอบทุกคน [$P \times I \times R$] จะให้ค่าว่าค่าสัมประสิทธิ์การสรุปอ้างอิงสูงกว่ารูปแบบการตรวจข้อสอบบางข้อของผู้สอบทุกคน [$P \times (I : R)$] เพราะเป็นสถานการณ์การตรวจให้คะแนนในระดับชั้นเรียน ซึ่งเป็นสถานการณ์ที่มีจำนวนครูผู้สอนหรือผู้ตรวจที่จำกัดและมีจำนวนนักเรียนไม่มากนัก จึงควรเลือกใช้รูปแบบการตรวจข้อสอบทุกข้อของผู้สอบทุกคน [$P \times I \times R$] ถึงแม้จะเป็นรูปแบบที่ค่อนข้างใช้เวลาในการตรวจนานเมื่อเปรียบเทียบกับรูปแบบการตรวจข้อสอบบางข้อของผู้สอบทุกคน [$P \times (I : R)$] ซึ่งสามารถลดเวลาและภาระในการตรวจให้คะแนนได้มากกว่า แต่ก็เป็นรูปแบบที่สอดคล้องกับสถานการณ์จริงและไม่จำเป็นต้องใช้ผู้ตรวจมากนักเพื่อลดต้นทุนการพัฒนาผู้ตรวจ

ข้อเสนอแนะ

1. ข้อเสนอแนะในการนำผลวิจัยไปใช้

1) ผลการวิเคราะห์ความเห็นพ้อง (rater agreement) ของผลการตรวจระหว่างผู้ตรวจ เทียบกับคะแนนเกณฑ์มาตรฐาน พบว่า ผู้ตรวจให้คะแนนคนที่ 3 มีผลการตรวจให้คะแนนที่มีความเห็นพ้องกับคะแนนเกณฑ์มาตรฐานน้อยกว่าผู้ตรวจคนอื่นในทุกข้อ ดังนั้น ในการคัดเลือกผู้ตรวจให้คะแนน ควรมีการเลือกผู้ตรวจที่มีคุณลักษณะเหมือนกับผู้ตรวจคนที่ 1 และ 2 ที่มีการยอมรับเกณฑ์การตรวจให้คะแนนมากกว่า ซึ่งจะช่วยให้ผลการวิเคราะห์ร้อยละความเห็นพ้องสูงกว่า ส่งผลให้ผลการตรวจให้คะแนนมีความสอดคล้องกับคะแนนเกณฑ์มาตรฐาน มีความน่าเชื่อถือมากขึ้น และลดความคลาดเคลื่อนในการตรวจให้คะแนนลงได้ ดังนั้นเพื่อลดความคลาดเคลื่อนอันเนื่องมาจากการให้คะแนน ก่อนการตรวจให้คะแนนควรมีการอบรมหรือชี้แจงเกณฑ์การตรวจให้คะแนนพร้อมทั้งให้คำแนะนำในการตรวจให้คะแนนแก่ผู้ตรวจให้คะแนน รวมถึงอธิบายตัวอย่างหรือแนวคำตอบในการตรวจให้คะแนน ให้ผู้ตรวจให้คะแนนยอมรับ และมีแนวทางการตรวจให้คะแนนไปในแนวเดียวกัน

2) ผลการวิเคราะห์แหล่งความแปรปรวนของแต่ละองค์ประกอบที่มีผลต่อการสัมประสิทธิ์การสรุปอ้างอิง (G-coefficient) ผลการประมาณค่าความแปรปรวนของทั้ง 3 องค์ประกอบ คือ ผู้สอบ (P) ข้อสอบ (I) และผู้ตรวจ (R) เมื่อใช้รูปแบบการตรวจข้อสอบทั้งสองรูปแบบ พบว่า มีความแปรปรวนที่มาจากผู้ตรวจมากกว่าความแปรปรวนจากผู้สอบและข้อสอบ ซึ่งอาจเกิดจากเกณฑ์การตรวจให้คะแนนที่เข้าใจยาก ดังนั้นควรใช้เวลาในการการอบรมหรือชี้แจงเกณฑ์การตรวจให้คะแนนมากขึ้น

2. ข้อเสนอแนะในการวิจัยครั้งต่อไป

1) ผลการวิจัยในครั้งนี้ ศึกษาผลการตรวจให้คะแนนโดยใช้เกณฑ์การตรวจให้คะแนนแบบองค์รวม (Holistic Rubric) พบว่า ค่าสัมประสิทธิ์การสรุปอ้างอิงในรูปแบบการตรวจด้วยรูปแบบการตรวจข้อสอบทุกข้อของผู้สอบทุกคน $[P \times I \times R]$ มีค่าสูงกว่า การตรวจข้อสอบบางข้อของผู้สอบทุกคน $[P \times (I : R)]$ จึงควรเปรียบเทียบกับผลการตรวจให้คะแนนแบบวิธีวิเคราะห์ย่อย (Analytic Rubric) ซึ่งเป็นเกณฑ์ที่มีความละเอียด มีรายละเอียดของแต่ละสเกลชัดเจน เพื่อใช้ในการตัดสินใจเลือกใช้เกณฑ์และรูปแบบการตรวจให้คะแนนให้เหมาะสมในการตรวจให้คะแนนครั้งต่อไป

2) การวิจัยครั้งนี้ผู้วิจัยได้เลือกกลุ่มตัวอย่างที่เป็นผู้ตรวจให้คะแนนแบบสอบอัตนัยวัดความสามารถในการแก้ปัญหาทางคณิตศาสตร์ซึ่งเป็นครูผู้สอนวิชาคณิตศาสตร์ ที่มีคุณวุฒิในการจบทางด้าน การสอนคณิตศาสตร์โดยตรง ในการออกแบบหรือศึกษาการตรวจให้คะแนนครั้งต่อไป ควรศึกษาผลการตรวจให้คะแนนโดยกลุ่มผู้ตรวจที่มีคุณวุฒิไม่ตรงเอก เพื่อเปรียบเทียบผลการตรวจให้คะแนนว่าแตกต่างกันหรือไม่

3) การคัดเลือกกลุ่มตัวอย่างที่เป็นผู้ตรวจควรตรวจสอบคุณลักษณะของผู้ตรวจว่ามีค่าความใจดี เป็นกลาง และเข้มงวดอย่างไร และคัดเลือกผู้ตรวจที่มีความเป็นกลางมาตรวจให้คะแนน จึงจะเห็นความสัมพันธ์ในการตรวจให้คะแนนมากยิ่งขึ้น

References

- Koo, T. K., & Li, M. Y. (2016). A Guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155-163.
- Miller, M. D., & Legg, S. M. (1993). Alternative assessment in a high-stakes environment. *Educational Measurement: Issues and Practice*, 12(2), 9-15.

Translate Thai References

- Aphaikawee, D. (2019). *Improving the effectiveness of the subjective test scoring* [Doctoral dissertation]. Khon Kaen University. (in Thai)
- Bhu-iam, W. (2007). *A comparison of generalizability coefficient and error variance between traditional and two-tier diagnostic tests In mathematics* [Master's thesis]. Chulalongkorn University. (in Thai)
- Buasiripan, U. (2000). *Comparison of reference sum coefficients of existing math tests. The methods of examination, the number of examiners, and the experience of the inspectors differ* [Master's thesis]. Srinakharinwirot University. (in Thai)
- Chuithong, S. (2011). *Model quality examination of the learning assessment toolkit based on actual conditions, mathematics learning subject group Mathayom 1 by applying the theory of reference summary* [Master's thesis]. Srinakharinwirot University. (in Thai)
- Institute for the Promotion of Teaching Science and Technology. (2020). *Mathematical Literacy*. <https://pisathailand.ipst.ac.th/about-pisa/mathematical-literacy/> (in Thai)
- Inthanate, N. (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* [Doctoral dissertation]. Srinakharinwirot University. (in Thai)
- Kanjanawasee, S. (2012). *New test theory* (4th Ed.). Chulalongkorn University Press. (in Thai)
- Prasertsinponma, D., & Tuksino, P. (2021). Comparison of results of subjective test for measuring mathematical competencies of primary 6: application of generalizability theory. *The 22nd National Graduate Research Conference (Online Conference)*, 22(1), 398-408. (in Thai)
- Saenpluem, B. (2013). *Using the Characteristics Testing Method and the Proportion of the Number of Reviewers on the Accuracy of the Writing Ability Measurement of Grade 3 Students* [Doctoral dissertation]. Srinakharinwirot University. (in Thai)

- Sa-nguanwai, C. (2015). *Comparison of the reliability of problem-solving ability tests mathematical creativity The application of summation theory references the reliability of measurement results* [Master's thesis]. Chulalongkorn University. (in Thai)
- Tangdhanakanond, K. (2014). *Measuring and evaluating practical skills*. Chulalongkorn University Press. (in Thai)
- Taoto, J., Jarernvongrayab, A., & Baikularb, P. (2016). A study of the reliability of mathematics essay test score for matayomsuksa 2 students: The different number of raters and scoring patterns using generalizability theory. *Hatyai Journal*, 14(1), 1-14. (in Thai)
- Thitikanpodchana, W., & Tuksino, P. (2021). Comparisons of the Results of Essay Test by Different Scoring Designs: Application of Generalizability Theory. *The 22nd National Graduate Research Conference (Online Conference)*, 22(1), 409-418. (in Thai)
- Tuksino, P. (2013). *Teaching documentation Educational research methodology*. Faculty of Education, Khon Kaen University. (in Thai)
- Umnacil, M. (2014). *Comparison of reliability of modified essay question test for measuring Scientific problem solving ability using different scoring methods Under different numbers of event: An application of generalizability theory* [Master's thesis]. Chulalongkorn University. (in Thai)