การศึกษาผลการตรวจให้คะแนนแบบสอบอัตนัย ประยุกต์ใช้โมเดลหลายองค์ประกอบของราส์ช และทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัด

A Study of the Results of Subjective Test Scoring by Applying the Many-Facet Rasch Model and Generalizability Theory

ดรุณี อภัยกาวี[1] ประกฤติยา ทักษิโณ[2] และกมลวรรณ ตังธนกานนท์[3]
Darunee Apaikawee[1], Prakrittiya Tuksino[2] and Kamonwan TangDhanakanond[3]
(Received: June 27, 2019; Revised: August 3, 2019; Accepted: August 9, 2019)

บทคัดย่อ

งานวิจัยเรื่องนี้มีความมุ่งหมายเพื่อ 1) ศึกษาความตรงตามสภาพ และ 2) ศึกษาค่าสัมประสิทธิ์การสรุปอ้างอิงของผลการตรวจให้คะแนนแบบสอบอัตนัย กลุ่มตัวอย่าง ได้แก่ ผู้ตอบข้อสอบ เป็นนักเรียนชั้นมัธยมศึกษาปีที่ 3 จำนวน 32 คน และ ผู้ตรวจให้คะแนนเป็นนักศึกษาสาขาวิชาคณิตศาสตร์ จำนวน 37 คน เครื่องมือที่ใช้ในการวิจัย คือ แบบสอบอัตนัยการคิดทางคณิตศาสตร์ จำนวน 3 ข้อ ประยุกต์ใช้โมเดลหลายองค์ประกอบของราส์ช (Many Facet Rasch Model) และ ทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัด (generalizability theory) ผลการวิจัยพบว่า

1) ค่าความตรงตามสภาพของผลการตรวจให้คะแนนแบบสอบอัตนัย 3 ข้อ ของผู้ตรวจให้คะแนนที่ต่างกัน เทียบกับคะแนนเกณฑ์กลางซึ่งมาจากคะแนนฉันทามติการตรวจของผู้เชี่ยวชาญ พบว่า ค่าสัมประสิทธิ์สหสัมพันธ์ของคะแนนที่ได้จากการตรวจของผู้ตรวจให้คะแนนเป็นกลางและเข้มงวดสูงกว่าผู้ตรวจให้คะแนนใจดี

2) ค่าสัมประสิทธิ์การสรุปอ้างอิง ในทุกคุณลักษณะของผู้ตรวจให้คะแนนที่ต่างกัน ภายใต้เงื่อนไขของรูปแบบการตรวจข้อสอบบางข้อของผู้สอบทุกคน [p x (i : r)] สูงกว่ารูปแบบการตรวจข้อสอบทุกข้อของผู้สอบทุกคน [p x i x r]

คำสำคัญ : โมเดลหลายองค์ประกอบของราส์ช ทฤษฎีการสรุปอ้างอิง คุณลักษณะของผู้ตรวจให้คะแนน

---

[1] นักศึกษาปริญญาเอก สาขาวิชาการวัดและประเมินผลการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยขอนแก่น
[2] ผู้ช่วยศาสตราจารย์ ภาควิชาการวัดและประเมินผลการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยขอนแก่น
[3] รองศาสตราจารย์ ภาควิชาวิจัยและจิตวิทยาการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
[1] Doctoral degree student in Educational Measurement and Evaluation Program, Faculty of Education, Khon Kaen University
[2] Assistant professor, Department of Educational Measurement and Evaluation, Faculty of Education, Khon Kaen University
[3] Associate professor, Department of Educational Research and Psychology, Faculty of Education, Chulalongkorn University
Corresponding Author E-mail: ajouiup@gmail.com

Abstract

The objectives of this research were: 1) to study the concurrent validity of the subjective test scores and 2) to study the generalizability coefficient of the subjective test scores. The sample consisted of 32 grade 9 students as the test takers and 37 students in mathematics as the raters. The research tool was a 3-item subjective test for measuring mathematical thinking with application of the many-facet Rasch model and generalizability theory. The research findings were as follows:

1. The concurrent validity of the 3-item subjective test scores rated by different raters was compared to the central score which came from the consensus of the experts and it was found that the correlation coefficients of the scores from the central rater and the severity rater were higher than those from the leniency rater.

2. The generalizability coefficient in every characteristic of the raters who assigned different scores, under the condition of the design that the raters scored some items of every test taker [p x (i : r)], was higher than the design in which the raters scored every item of every test taker [p x i x r].

Keywords : many facet Rasch model, generalizability theory, characteristics of rater

บทนำ

การตรวจให้คะแนนข้อสอบอัตนัย ข้อสอบที่สร้างคำตอบ (constructed – response question: CRQ) เป็นเครื่องมือที่เหมาะกับการวัดความสามารถทางสมองขั้นสูง สร้างง่ายแต่ยากต่อการตรวจให้คะแนนอย่างมีคุณภาพ เนื่องจากมีคำตอบถูกต้องนอกเหนือจากรูปแบบเดียว ดังนั้น การตรวจให้คะแนนที่มีประสิทธิภาพผู้ตรวจจะต้องเรียนรู้แหล่งความคลาดเคลื่อนของคะแนนและฝึกฝนความเชี่ยวชาญในการตรวจ เพื่อลดความคลาดเคลื่อนและความลำเอียงที่อาจเกิดขึ้น โดยที่แหล่งความคลาดเคลื่อนที่สำคัญของคะแนน อาทิ คุณภาพของเครื่องมือ ทักษะของผู้ตรวจ และการบริหารการตรวจให้คะแนน (ศิริชัย กาญจนวาสี, 2558) คุณภาพของการตรวจให้คะแนน หากลดความคลาดเคลื่อนได้จะทำให้การตรวจมีความยุติธรรม และผลการตรวจมีความน่าเชื่อถือได้ อาทิแหล่งความคลาดเคลื่อนภายนอกผู้ตรวจให้คะแนน ซึ่งเป็นสิ่งที่นอกเหนือการควบคุมของผู้ให้คะแนน เช่น เกณฑ์การตรวจให้คะแนน ความยากของข้อสอบ ความสามารถของผู้สอบหรืออาจเกิดจากตัวผู้เข้าสอบ ทั้งลายมือ ความถูกต้องของหลักการเขียน สิ่งเหล่านี้มีผลต่อการตรวจให้คะแนน แนวทางแก้ไขต้องชี้แจงให้ผู้เข้าสอบทราบอย่างชัดเจนว่าการตรวจให้คะแนนนั้นคำนึงถึงสิ่งใดบ้างเพื่อให้ผู้เข้าสอบเพิ่มความระมัดระวัง (พวงรัตน์ ทวีรัตน์, 2530) และแหล่งความคลาดเคลื่อนภายในผู้ตรวจ เป็นความคลาดเคลื่อนที่มาจากคุณลักษณะภายในตัวของผู้ตรวจเป็นหลัก เช่น อารมณ์ของผู้ตรวจขณะให้คะแนน ความประทับใจส่วนตัว (the halo error) ผู้ตรวจแต่ละคนก็มีมาตรฐานต่างกันทำให้คะแนนคลาดเคลื่อนไป หรือ การคล้อยตามกัน

จากพฤติกรรมส่วนตัวของผู้ตรวจ และความเข้มงวดหรือใจดีของผู้ตรวจ (Saal, Downey & Lahey, 1980; พวงรัตน์ ทวีรัตน์, 2530) ดังนั้น คะแนนของผู้สอบจะได้รับผลกระทบจากคุณลักษณะการให้คะแนนของผู้ตรวจ (characteristics of rater) พอ ๆ กับได้รับผลกระทบจากระดับความยากง่ายของข้อสอบและความสามารถของผู้สอบ นั่นอาจจะเป็นอีกปัจจัยที่มีอิทธิพลต่อความเที่ยงของผู้ตรวจ (rater reliability) ซึ่งต้องขึ้นอยู่กับมาตรฐานของการตัดสินใจของผู้ตรวจเองที่จะมีความเห็นสอดคล้องกันสูง (Hopkins & Antes, 1990) การออกแบบเกณฑ์การให้คะแนน (scoring rubric) การใช้เกณฑ์การให้คะแนนกับผู้สอบทุกคน (Mehrens & Lehmann, 1972)

จากการศึกษางานวิจัยที่ผ่านมา เกี่ยวกับผลการวิเคราะห์ความเที่ยงของผู้ตรวจ โดยใช้วิธีการตรวจเหมือนกันและวิธีการตรวจต่างกัน พบว่า แตกต่างกันอย่างไม่มีนัยสำคัญทางสถิติ (นิศารัตน์ คงสวัสดิ์, 2544) เนื่องจากผู้ตรวจใช้เกณฑ์การให้คะแนนเป็นแนวทางเดียวกัน จึงทำให้อคติของผู้ตรวจถูกขจัดทิ้งไปและมีความเป็นปรนัยของการตรวจให้คะแนนเพิ่มมากขึ้น ช่วยให้ผู้ตรวจมีความเห็นสอดคล้องกัน (น้ำผึ้ง อินทะเนตร, 2554) ถึงแม้ว่าจะกำหนดเกณฑ์การให้คะแนนที่ชัดเจน เป็นระบบแล้วก็ตาม ถ้าใช้ผู้ตรวจคนเดียวกันตรวจให้คะแนนแต่ใช้เวลาในการตรวจต่างกันก็ทำให้คะแนนไม่คงที่หรือเกิดความลำเอียงในการตรวจให้คะแนนกับผู้สอบได้ ซึ่งเกณฑ์การให้คะแนนแบบรูบริค (scoring rubric) เป็นแนวทางหนึ่งที่สามารถช่วยลดปัญหาเกี่ยวกับเวลาที่ใช้ในการตรวจลงได้และโดยทั่วไปการให้คะแนนโดยใช้เกณฑ์แบบองค์รวม (holistic scoring rubric) มักใช้เวลาน้อยกว่าการใช้เกณฑ์แบบแยกองค์ประกอบ (analytic scoring rubric) (กมลวรรณ ตังธนกานนท์, 2557) สอดคล้องกับ Cooney et al. (n.d. อ้างถึงใน น้ำผึ้ง อินทะเนตร, 2554) พบว่า วิธีการให้คะแนนแบบองค์รวมเหมาะสมกับคำถามเขียนตอบเพราะเป็นกฎเกณฑ์การให้คะแนนที่มีประสิทธิภาพ (effective) และ ปรีชา เนาว์เย็นผล (2544) กล่าวว่า ข้อดีของวิธีการให้คะแนนแบบองค์รวมเปิดโอกาสให้มีการพิจารณาคำตอบของผู้สอบได้อย่างรวดเร็ว ใช้เวลาน้อยในการตรวจ ง่ายต่อการนำไปใช้และวิธีการนี้ทำให้คะแนนมีความเที่ยงสูง หากแต่ไม่สามารถอธิบายเหตุผลของคะแนนที่ผู้สอบได้รับเพื่อใช้พัฒนาการเขียนตอบให้ดียิ่งขึ้น เมื่อเทียบกับวิธีการให้คะแนนแบบแยกองค์ประกอบ ที่ให้สารสนเทศที่จำเป็นย้อนกลับไปสู่ผู้สอบ เนื่องจากคำตอบที่ผู้สอบเขียนขึ้นมาอาจมีหลากหลายวิธีการขึ้นอยู่กับความสามารถของผู้สอบแต่ละคน ในปัจจุบันวิธีการให้คะแนนพบทั้งสองแบบ คือ แบบองค์รวมและแบบแยกองค์ประกอบ ซึ่งผู้ออกข้อสอบสามารถเลือกใช้วิธีการให้คะแนนแบบใดแบบหนึ่งหรือเลือกทั้งสองแบบก็ได้ ต่างมีข้อดี ข้อจำกัดแตกต่างกัน ควรเลือกใช้ให้สอดคล้องกับสิ่งที่ต้องการจะวัด และปรับเปลี่ยนไปได้ตลอดเวลาตามความเหมาะสม ซึ่งในงานวิจัยครั้งนี้ ผู้วิจัยใช้เกณฑ์การให้คะแนนแบบแยกองค์ประกอบ

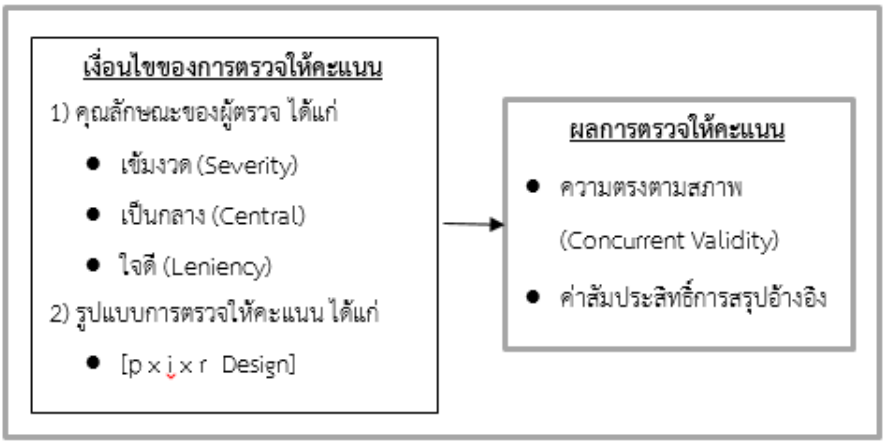
รูปแบบการตรวจให้คะแนนที่ยุติธรรมกับผู้สอบที่สุด คือ การให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน แล้วคำนวณคะแนนเฉลี่ย แต่การตรวจรูปแบบนี้ ต้องใช้เวลาในการตรวจและ มีค่าใช้จ่ายสูง กรณีที่มีผู้สอบจำนวนมาก หากแต่เป็นรูปแบบที่เหมาะกับสถานการณ์ภายในโรงเรียนที่คุณครูสามารถตรวจข้อสอบของนักเรียนทุกคน ดังนั้น หากมีการวางแผนการตรวจที่สามารถลดภาระ เวลาในการตรวจที่เป็นไปได้ในทางปฏิบัติ แต่มีความน่าเชื่อถือ จากข้อเสนอแนะของ น้ำผึ้ง อินทะเนตร (2554) กล่าวว่า รูปแบบการตรวจให้คะแนนข้อสอบทุกข้อของผู้สอบบางคนหรือการตรวจข้อสอบบางข้อของผู้สอบทุกคนไม่จำเป็นต้อง

ตรวจข้อสอบทุกข้อของผู้สอบทุกคน น่าจะเหมาะสมและมีประสิทธิภาพประสิทธิผลเช่นกัน ซึ่งรูปแบบการตรวจมี 2 แบบ คือ แบบไขว้สมบูรณ์ (full crossed) หรือ แบบบางส่วน (partially) ผู้ตรวจตรวจข้อสอบบางข้อ โดยมีการเชื่อมโยง หรือไขว้กันของผู้ตรวจและข้อสอบ ก็เพียงพอแล้วในการประมาณค่าโดยยังคงมีความน่าเชื่อถือในทางสถิติ (Brennan, 2001) จากการศึกษาของ Sudweeks, Reeve &Bradshaw (2005) พบว่าการใช้รูปแบบแฝงหรือสอดแทรก (nested) โดยให้ผู้ตรวจ แต่ละคนตรวจข้อสอบทุกข้อของผู้สอบบางคน จะได้คะแนนที่มีค่าความเที่ยงแตกต่างจากการออกแบบไขว้สมบูรณ์ (full crossed) เพียง 1.5 - 2.0% ถือว่ามีความแตกต่างกันน้อยมาก (ศิริชัย กาญจนวาสี, 2555) ดังนั้น ตัวแปรภายใต้เงื่อนไขของรูปแบบการตรวจให้คะแนน ผู้วิจัยจะทำการศึกษาใน 2 รูปแบบ เพื่อความเป็นไปได้ในการนำผลการศึกษาครั้งนี้ไปใช้ ที่เป็นประโยชน์และเหมาะสมต่อการจัดการทดสอบทั้งในระดับห้องเรียนจนถึงระดับชาติ คือ การตรวจข้อสอบทุกข้อของผู้สอบทุกคนและการตรวจข้อสอบบางข้อของผู้สอบทุกคน

วิธีการตรวจสอบแหล่งความแปรปรวนได้หลายแหล่งนอกเหนือจากทฤษฎีการสรุปอ้างอิง (generalizability theory) ที่ซึ่งมีความยืดหยุ่นที่จะรวมแหล่งความคลาดเคลื่อนได้หลายแหล่งและไม่ยึดข้อตกลงเบื้องต้นเกี่ยวกับคุณสมบัติเท่าเทียมกันของข้อสอบทำให้มีการนำโมเดลนี้มาใช้อย่างกว้างขวางในการวัดประเมินทางด้านการเขียน นั่นคือ โมเดลหลายองค์ประกอบของราสช์ (Many - Facet Rash Model) (Linacre, 1994) มาปรับขยายแนวคิดเดิมที่มุ่งเน้น 2 องค์ประกอบคือ ผู้สอบและข้อสอบ และให้ค่าพารามิเตอร์ 2 ค่า คือ พารามิเตอร์ของผู้สอบ (examinee parameter: θ) แสดงถึงค่าความสามารถ (ability or competence) ของผู้สอบ และพารามิเตอร์ของข้อสอบ (item parameter: b_i) ที่แสดงถึงระดับความยาก (difficulty) ของข้อสอบ โดยเพิ่มองค์ประกอบผู้ตรวจให้คะแนน (rater parameter) ที่แสดงถึงความเข้มงวดของผู้ตรวจ (rater severity) เข้าไปศึกษาในโมเดลด้วย โดย MFRM จะประมาณค่าความสามารถของผู้สอบ (θ) ที่เป็นอิสระจากกลุ่มผู้สอบ อิสระจากผู้ตรวจ และอิสระจากข้อสอบ (Linacre & Wright, 2002; Smith & Kulikowich, 2004) ดังนั้น ค่าความสามารถของผู้สอบ (θ) จึงมีค่าไม่แปรเปลี่ยนไปตามผู้สอบและข้อสอบ เนื่องจากค่าความสามารถของผู้สอบ(θ) ที่ได้จากการวิเคราะห์ MFRM ได้รับการปรับแก้จากความแตกต่างของความเข้มงวดของผู้ตรวจและความยากของข้อสอบ ซึ่งความสำคัญดังกล่าวมีประโยชน์สำหรับการนำมาใช้ในการสร้างและพัฒนาแบบสอบ (ผจงจิต อินทสุวรรณ, 2525) ทำให้สามารถนำคะแนนที่ได้ไปเปรียบเทียบกันได้โดยตรง จึงเป็นวิธีหาคะแนนที่ยุติธรรมสำหรับการตรวจข้อสอบ แม้ว่าจะได้รับการตรวจจากผู้ตรวจที่มีความเข้มงวดต่างกัน หรือ จากข้อสอบที่มีความยากต่างกันก็ตาม โดยผู้ตรวจทุกคนไม่จำเป็นต้องตรวจข้อสอบทุกข้อของผู้สอบทุกคน (Sudweeks, Reeve & Bradshaw, 2005) จึงทำให้สามารถนำแนวคิดทั้งสองมาใช้ศึกษาคุณลักษณะของผู้ตรวจและรูปแบบการตรวจที่ต่างกัน โดยทฤษฎีการสรุปอ้างอิงจะเป็นการศึกษาแหล่งความแปรปรวนในภาพรวม และโมเดลหลายองค์ประกอบของราสช์จะอธิบายแหล่งความแปรปรวนต่าง ๆ ในระดับแต่ละสมาชิกขององค์ประกอบ และนำเอาความแตกต่างของความเข้มงวดของผู้ตรวจและ ความยากของข้อสอบเข้าไปปรับแก้ในโมเดลด้วย แนวคิดทั้งสองจึงเป็นวิธีการทางสถิติที่สามารถ อธิบายคุณลักษณะของคะแนนได้ (Iramaneerat et al., 2007) ในการศึกษาครั้งนี้ กำหนดเกณฑ์ในการแบ่งคุณลักษณะของผู้ตรวจโดยพิจารณาค่าโลจิทของผู้ตรวจ (logit raters) ด้วยโปรแกรม FACET (Linacre, 2014) หากกลุ่มผู้ตรวจที่มีค่า

Measure มากกว่า +1.00 จะถือว่าเป็นผู้ตรวจที่มีคุณลักษณะเข้มงวด ผู้ตรวจที่มีค่า Measure อยู่ระหว่าง -1.00 ถึง +1.00 จะถือว่าเป็นผู้ตรวจที่มีคุณลักษณะเป็นกลาง และผู้ตรวจที่มีค่า Measure น้อยกว่า -1.00 จะถือว่าเป็นผู้ตรวจที่มีคุณลักษณะใจดี (Linacre, 1994) จากเกณฑ์ดังกล่าวสามารถแยกลักษณะของผู้ตรวจที่ต่างกัน เพื่อนำผลการตรวจ (คะแนน) ไปวิเคราะห์ความแปรปรวนจากแหล่งผู้ตรวจ และตอบคำถามการวิจัยตามทฤษฎีการสรุปอ้างอิง (G-Theory) ในลำดับถัดไป

การตรวจสอบความคลาดเคลื่อนที่เกิดจากผู้ตรวจ ในงานวิจัยที่ผ่านมา ภายใต้เงื่อนไข คุณลักษณะของผู้ตรวจและรูปแบบการตรวจให้คะแนน โดยประยุกต์ใช้โมเดลหลายองค์ประกอบของราส์ช อาทิ การศึกษาของ Rui (2010) พบว่า ระดับความเข้มงวดการให้คะแนนของผู้ตรวจมีความแตกต่างกัน โดยการให้คะแนนของผู้ตรวจจะมีความคงเส้นคงวา มีเพียงหนึ่งคนที่การให้คะแนนค่อนข้างสูงกว่าค่า INFIT(1.2) ระดับความเข้มงวดของผู้ตรวจส่วนใหญ่จะมีความผันแปรทั้งสองช่วงของการประเมินแต่เป็นที่ยอมรับได้ การศึกษาของบุษวรรษ์ แสนปลื้ม (2556) พบว่า คุณลักษณะของผู้ตรวจเข้มงวด/ใจดี ความลำเอียงด้านเพศและการทำหน้าที่ต่างกันของผู้ตรวจเมื่อเวลาผ่านไปมีผลต่อค่าความตรงตามสภาพของคะแนนแบบสอบอัตนัยและเสนอแนะว่าถึงแม้ผู้ตรวจจะมีลักษณะลำเอียง แต่เมื่อจำนวนผู้ตรวจเพิ่มขึ้น ค่าความตรงตามสภาพก็จะสูงขึ้น ควรเลือกผู้ตรวจเป็นกลาง 3 คน และการศึกษาของ ศุกลรัตน์ อิงชาติเจริญ (2557) พบว่า ผู้ประเมินกลุ่มเพื่อนมีแนวโน้มให้คะแนนชนิดปล่อยคะแนน/ใจดี ส่วนผู้ประเมินกลุ่มอาจารย์มีแนวโน้มให้คะแนนชนิดกดคะแนน/เข้มงวดเป็นอันดับสองรองจากผู้ประเมินกลุ่มตนเอง และทั้งหมดถือว่าเป็นการให้คะแนนที่มีความแม่นยำ (Infit MNSQ อยู่ระหว่าง 0.97 ถึง 1.05) การศึกษาความตรงตามสภาพของผลการวัด ถ้าจำนวนผู้ตรวจมากกว่าจะมีค่าความตรงตามสภาพสูงกว่าจำนวนผู้ตรวจน้อยในทุกรูปแบบการตรวจ (อังคณา กุลนภาดล, 2555) หรือการศึกษาคะแนนในทุกเงื่อนไขที่ต่างกันจะมีความตรงตามสภาพสูง เช่นกัน (น้ำผึ้ง อินทะเนตร, 2554) และการศึกษาที่ประยุกต์ใช้ทฤษฎีการสรุปอ้างอิง พบว่า ถ้าใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (p : r) x i จะให้ค่าสัมประสิทธิ์การสรุปอ้างอิงสูงขึ้น (น้ำผึ้ง อินทะเนตร, 2554 ; นภัสนันท์ ขวัญจ่า, 2556) การศึกษาของ ชนิสรา สงวนไว้ (2558) พบว่า ค่าสัมประสิทธิ์การสรุปอ้างอิงของรูปแบบการตรวจ p x (i : r) สูงกว่ารูปแบบ p x i x r และการศึกษาของ จิรายุ เถาว์โท (2559) พบว่า ค่าสัมประสิทธิ์การสรุปอ้างอิงเมื่อรูปแบบการตรวจให้คะแนนเหมือนกันแต่จำนวนผู้ตรวจต่างกัน มีค่าแตกต่างกันอย่างมีนัยสำคัญ ยกเว้นรูปแบบการตรวจ (p : r) x i และ p x (i : r) ของผู้ตรวจ 2 และ 3 คน มีค่าไม่แตกต่างกัน เมื่อใช้จำนวนผู้ตรวจเท่ากันแต่รูปแบบการให้คะแนนต่างกัน ค่าสัมประสิทธิ์การสรุปอ้างอิงมีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติ เป็นต้น จากประเด็นการศึกษาดังกล่าว ผู้วิจัยต้องการศึกษาผลการตรวจให้คะแนนแบบสอบอัตนัย ภายใต้เงื่อนไขคุณลักษณะของผู้ตรวจเข้มงวด/ใจดี และเป็นกลาง ที่ตรวจให้คะแนนด้วยรูปแบบการตรวจข้อสอบทุกข้อของผู้สอบทุกคน และตรวจให้คะแนนด้วยรูปแบบการตรวจข้อสอบบางข้อของผู้สอบทุกคน แสดงดังภาพที่ 1 ว่ามีผลต่อความตรงตามสภาพและมีความเที่ยงอย่างไร เพื่อเป็นสารสนเทศของการตรวจให้คะแนนแบบสอบอัตนัยให้มีความน่าเชื่อถือ ยุติธรรมและมีความเป็นปรนัย (Linacre, 1994) มากที่สุด



ภาพประกอบ 1 กรอบแนวคิดในการวิจัย

ความมุ่งหมายของการวิจัย

การศึกษาผลการตรวจให้คะแนนแบบสอบอัตนัยในครั้งนี้ ภายใต้เงื่อนไขของคุณลักษณะของผู้ตรวจให้คะแนน และรูปแบบการตรวจให้คะแนนที่ต่างกัน ผู้วิจัยศึกษา 2 ประเด็นคือ

1. เพื่อศึกษาความตรงตามสภาพของผลการตรวจให้คะแนนแบบสอบอัตนัย
2. เพื่อศึกษาค่าสัมประสิทธิ์การสรุปอ้างอิงของผลการตรวจให้คะแนนแบบสอบอัตนัย

วิธีดำเนินการวิจัย

1. ประชากรและกลุ่มตัวอย่าง

ประชากรและกลุ่มตัวอย่าง ได้แก่ กลุ่มนักเรียนระดับชั้นมัธยมศึกษาปีที่ 3 จำนวน 120 คน โดยการสุ่มแบบหลายขั้นตอน (multi – stage random sampling) และกลุ่มผู้ตรวจให้คะแนน คือนักศึกษาระดับปริญญาตรี คณะศึกษาศาสตร์ มหาวิทยาลัยขอนแก่น ที่ยินยอมเข้าร่วมกิจกรรมในโครงการวิจัยฯ และจะต้องมีคุณสมบัติผ่านการเรียนและมีผลการเรียนทางวิชาคณิตศาสตร์ และผลการเรียนทางวิชาการวัดและประเมินผลการศึกษา ตั้งแต่เกรด B ขึ้นไป จำนวน 37 คน (อ้างอิงหลักเกณฑ์คำประกาศเฮลซิงกิ ของศูนย์จริยธรรมการวิจัยในมนุษย์ มหาวิทยาลัยขอนแก่น : HE 613094)

ในการวิจัยครั้งนี้คำนวณขนาดตัวอย่างตามกฎเกณฑ์ของ Smith (1978) ที่เสนอขั้นต่ำ n_p= 25 คน (low), n_p= 50 คน (middle) และ n_p=100 คน (high) คือ $n_p \times n_i \times n_r$ อย่างน้อย 1,080 ค่า ซึ่งในการออกแบบการวัดใช้โปรแกรม EduG เพื่อตรวจสอบแหล่งความคลาดเคลื่อน (source of variance) และประมาณค่าสัมประสิทธิ์การสรุปอ้างอิง (G-Coefficient) จะได้ (n_p = 32) $\times$ (n_i = 3) $\times$ (n_r = 37) = 3,552 ผ่านเกณฑ์ค่าขั้นต่ำของการประมาณค่าตามข้อเสนอของสมิท

2. เครื่องมือ/ตัวแปรที่ใช้ในการวิจัย

2.1 เครื่องมือที่ใช้ในการวิจัย ได้แก่ แบบสอบอัตนัยการคิดทางคณิตศาสตร์ จำนวน 3 ข้อ วัดประเมินพฤติกรรมตั้งแต่ระดับการวิเคราะห์ขึ้นไปของ Bloom's Revised Taxonomy (Anderson et al., 2001) ที่สัมพันธ์กับเนื้อหาทางคณิตศาสตร์ สาระที่ 5 การวิเคราะห์ข้อมูลและความน่าจะเป็น และออกแบบเกณฑ์การให้คะแนนที่มีลักษณะเป็นเกณฑ์เฉพาะ (specific rubric) ของแต่ละข้อ โดยการเขียนบรรยายความสำเร็จในการตอบที่ถูกต้องเหมาะสม ตามกระบวนการทางคณิตศาสตร์ของ PISA (สสวท., 2560) 3 ด้าน คือ ด้านการคิดสถานการณ์ของปัญหาในเชิงคณิตศาสตร์ ด้านการใช้หลักการและกระบวนการทางคณิตศาสตร์ และ ด้านการตีความและประเมินผลลัพธ์ทางคณิตศาสตร์ โดยกำหนดมิติหรือระดับคุณภาพของลักษณะคำตอบเป็น 0-2 มิติ ให้สอดคล้องกับคำอธิบายของลักษณะคำตอบในแต่ละประเด็นการพิจารณา แล้วนำข้อสอบจำนวน 6 ข้อ ไปทดลองใช้ (try out) กับนักเรียนชั้นมัธยมศึกษาปีที่ 3 ที่ไม่ใช่กลุ่มตัวอย่าง ในโรงเรียนสาธิตมหาวิทยาลัยขอนแก่น ฝ่ายมัธยมศึกษา (ศึกษาศาสตร์) จำนวน 30 คน เพื่อตรวจสอบคุณภาพของข้อสอบที่ผ่านเกณฑ์ จำนวน 3 ข้อ (ใช้จริง) โดยมีค่าความตรงตามเนื้อหาของข้อคำถาม (content validity ratio: CVR_i) เท่ากับ 0.99 ค่าดัชนีความยากง่าย (p) มีค่าอยู่ระหว่าง 0.37 ถึง 0.67 ซึ่งทุกข้อมีค่าดัชนีอำนาจจำแนก (d) อยู่ระหว่าง 0.27 ถึง 0.56 และมีค่าความเที่ยงทั้งฉบับเท่ากับ 0.89

2.2 ตัวแปรที่ใช้ในการวิจัย หรือเงื่อนไขของการตรวจให้คะแนนที่ส่งผลต่อค่าความตรงตามสภาพและค่าสัมประสิทธิ์การสรุปอ้างอิง ดังนี้

2.2.1 คุณลักษณะของผู้ตรวจให้คะแนน ได้แก่ (1) ผู้ตรวจให้คะแนนเข้มงวด (2) ผู้ตรวจให้คะแนนเป็นกลาง และ (3) ผู้ตรวจให้คะแนนใจดี

2.2.2 รูปแบบการตรวจให้คะแนน ได้แก่ (1) ตรวจข้อสอบทุกข้อของผู้สอบทุกคน (p x i x r) (2) การตรวจข้อสอบบางข้อของผู้สอบทุกคน p x (i : r)

3. การเก็บรวบรวมข้อมูล

3.1 ผู้วิจัยนำแบบสอบอัตนัยการคิดทางคณิตศาสตร์ที่สร้างขึ้น หลังดำเนินการหาคุณภาพเครื่องมือและปรับแก้ไขเรียบร้อยแล้ว ขอความอนุเคราะห์โรงเรียนที่เป็นกลุ่มตัวอย่าง ดำเนินการทดสอบตามวัน เวลาที่นัดหมาย

3.2 ผู้วิจัยนำกระดาษคำตอบของนักเรียนมาคัดแยก และจัดเตรียมกระดาษคำตอบ และเกณฑ์การตรวจให้คะแนนแบบแยกองค์ประกอบเป็น 3 ประเด็นย่อย ตามกระบวนการทางคณิตศาสตร์ของ PISA (สสวท., 2560) แล้วชี้แจงรายละเอียดต่าง ๆ เกี่ยวกับวิธีการตรวจ เกณฑ์การตรวจให้คะแนน รวมถึงอธิบายตัวอย่างหรือแนวคำตอบในการตรวจให้คะแนน เพื่อลดความคลาดเคลื่อนอันเนื่องมาจากการให้คะแนน พร้อมทั้งให้คำแนะนำในการตรวจให้คะแนน และวิธีการบันทึกคะแนนลงในแบบบันทึกคะแนนที่ผู้วิจัยกำหนด

4. การวิเคราะห์ข้อมูล

4.1 การวิเคราะห์คุณลักษณะของผู้ตรวจให้คะแนน พิจารณาจากค่า Measure ในโปรแกรม FACET version 3.714.4 (Linacre, 2014) ที่แสดงถึงค่าคะแนนโลจิทของผู้ตรวจ ได้แก่ กลุ่มผู้ตรวจที่มีค่า

Measure มากกว่า +1.00 จะถือว่าเป็นผู้ตรวจที่มีคุณลักษณะเข้มงวด ผู้ตรวจที่มีค่า Measure อยู่ระหว่าง -1.00 ถึง +1.00 จะถือว่าเป็นผู้ตรวจที่มีคุณลักษณะเป็นกลาง และผู้ตรวจที่มีค่า Measure น้อยกว่า -1.00 จะถือว่าเป็นผู้ตรวจที่มีคุณลักษณะใจดี

4.2 การวิเคราะห์ผลการตรวจให้คะแนน พิจารณาค่าสัมประสิทธิ์การสรุปอ้างอิง ในโปรแกรม EduG version 6.1-e (Demo)

4.3 การวิเคราะห์ความตรงตามสภาพ (concurrent validity) ที่ได้จากการตรวจให้คะแนน จำนวน 3 ข้อ ของผู้ตรวจ (คะแนน X) เทียบกับคะแนนเกณฑ์กลาง (คะแนน Y) ซึ่งมาจากคะแนนฉันทามติการตรวจของผู้เชี่ยวชาญทางด้านคณิตศาสตร์ แล้วนำคะแนนทั้งสองส่วนมาคำนวณหาความสัมพันธ์ โดยหาสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน (Pearson product moment correlation coefficient) และแปลความหมาย (Hinkle D.E.,1998) ดังนี้ ค่า r มีค่า 0.90 – 1.00 เท่ากับ มีความสัมพันธ์กันระดับสูงมากมีค่า 0.70 – 0.89 เท่ากับ มีความสัมพันธ์กันระดับสูง มีค่า 0.50 – 0.69 เท่ากับ มีความสัมพันธ์กันระดับปานกลางมีค่า 0.30 – 0.49 เท่ากับ มีความสัมพันธ์กันระดับต่ำ และ มีค่า 0.00 – 0.29 เท่ากับ มีความสัมพันธ์กันระดับต่ำมาก

ผลการวิจัย

การศึกษาคุณลักษณะของผู้ตรวจให้คะแนน ผู้วิจัยพิจารณาจากค่า Measure in logit ของผู้ตรวจที่วิเคราะห์ด้วยโปรแกรม FACET (Linacre, 2014) เป็นค่าคะแนนโลจิท (logit) แสดงถึงระดับความเข้มงวดเป็นรายบุคคล จากผู้ตรวจ 37 คน พบว่า ผู้ตรวจให้คะแนนเข้มงวด มีค่าอยู่ในช่วง 1.500 ถึง 2.020 ผู้ตรวจให้คะแนนเป็นกลาง มีค่าอยู่ในช่วง 0.010 ถึง 0.330 และผู้ตรวจให้คะแนนใจดี มีค่าอยู่ในช่วง -2.00 ถึง -1.810 แสดงดังตารางที่ 1

ตาราง 1 ค่า Measure ที่แสดงถึงคุณลักษณะของผู้ตรวจให้คะแนนเป็นรายบุคคล

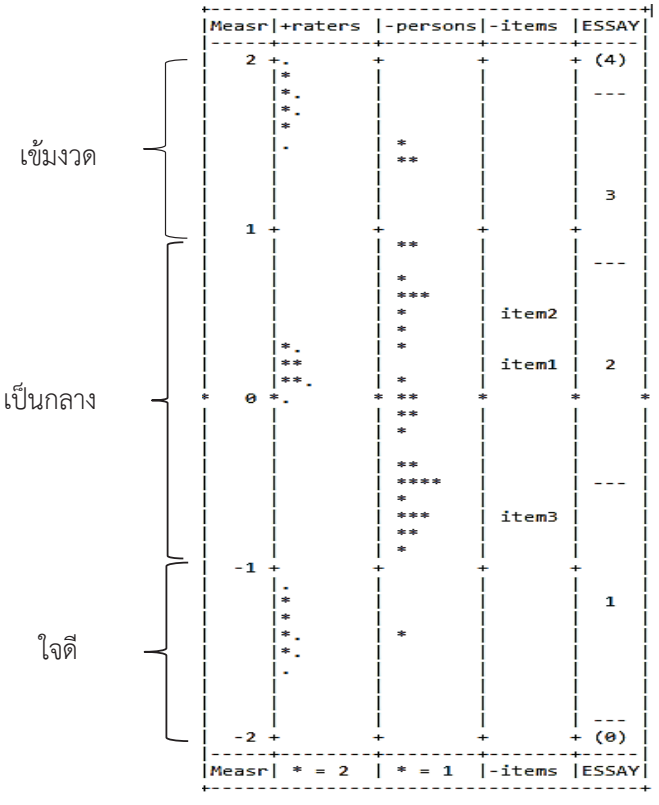
ผู้ตรวจ	Measure*	คุณลักษณะของผู้ตรวจ	ผู้ตรวจ	Measure*	คุณลักษณะของผู้ตรวจ
R01	1.700	เข้มงวด	R23	-1.900	ใจดี
R02	-1.860	ใจดี	R24	1.950	เข้มงวด
R03	0.210	เป็นกลาง	R26	-1.850	ใจดี
R04	0.180	เป็นกลาง	R27	0.100	เป็นกลาง
R06	-1.850	ใจดี	R28	1.760	เข้มงวด
R08	-1.920	ใจดี	R32	0.010	เป็นกลาง
R09	-1.810	ใจดี	R33	-1.860	ใจดี
R10	1.840	เข้มงวด	R34	0.240	เป็นกลาง
R11	0.110	เป็นกลาง	R35	1.800	เข้มงวด
R13	1.640	เข้มงวด	R37	0.200	เป็นกลาง
R14	0.270	เป็นกลาง	R38	0.310	เป็นกลาง

ตาราง 1 (ต่อ)

ผู้ตรวจ	Measure*	คุณลักษณะของผู้ตรวจ	ผู้ตรวจ	Measure*	คุณลักษณะของผู้ตรวจ
R15	-2.000	ใจดี	R39	0.200	เป็นกลาง
R16	-1.880	ใจดี	R41	0.260	เป็นกลาง
R17	0.330	เป็นกลาง	R42	-1.850	ใจดี
R18	1.670	เข้มงวด	R43	-1.850	ใจดี
R19	1.500	เข้มงวด	R44	0.180	เป็นกลาง
R20	2.020	เข้มงวด	R46	1.630	เข้มงวด
R21	1.840	เข้มงวด	R49	-1.840	ใจดี
R22	1.960	เข้มงวด	Expert	-0.050	เป็นกลาง

หมายเหตุ Measure* แสดงถึง คะแนนในหน่วยโลจิท (logit) ของแหล่งฟาเซทผู้ตรวจ (rater) ที่วิเคราะห์จากโปรแกรม FACET

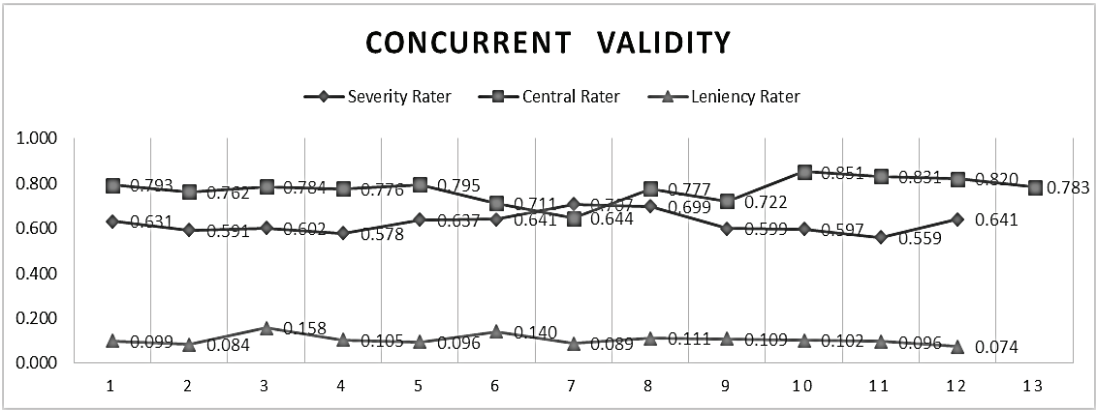
ทั้งนี้สามารถพิจารณาจากแผนที่ตัวแปร (variable map) การแจกแจงของคะแนนโลจิท ที่แสดงถึงระดับความเข้มงวดของผู้ตรวจให้คะแนน ประกอบให้เห็นความชัดเจน แสดงดังภาพประกอบ 2



ภาพประกอบ 2 แสดงระดับความเข้มงวดของผู้ตรวจให้คะแนนแบบสอบอัตนัย

ผู้วิจัยนำเสนอถึงผลการวิเคราะห์ตามจุดมุ่งหมาย 2 ข้อ รายละเอียดดังนี้

1) ความตรงตามสภาพของผลการตรวจให้คะแนนแบบสอบอัตนัย จากการตรวจให้คะแนนข้อสอบ 3 ข้อ ของผู้ตรวจเข้มงวด/เป็นกลาง และผู้ตรวจใจดี (คะแนน X) เทียบกับคะแนนเกณฑ์กลาง (คะแนน Y) ซึ่งมาจากคะแนนฉันทามติการตรวจของผู้เชี่ยวชาญ จำนวน 5 ท่าน (คุณสมบัติของผู้เชี่ยวชาญประกอบด้วย ผู้ที่มีประสบการณ์ด้านการสอนคณิตศาสตร์ระดับมัธยมศึกษา ไม่น้อยกว่า 5 ปีขึ้นไป จำนวน 3 คน และผู้ที่มีประสบการณ์ด้านคณิตศาสตร์และปฏิบัติงานกับสถาบันส่งเสริมการสอนวิทยาศาสตร์และเทคโนโลยี (สสวท.) ไม่น้อยกว่า 5 ปีขึ้นไป จำนวน 2 ท่าน ด้วยรูปแบบการตรวจข้อสอบทุกข้อของผู้สอบทุกคน ผลการวิเคราะห์ข้อสอบข้อที่ 1-3 พบว่า ค่าสัมประสิทธิ์สหสัมพันธ์ของคะแนนจากการตรวจของผู้ตรวจเป็นกลาง (ค่า r อยู่ช่วง 0.644 - 0.851) ถือว่าผลการตรวจสัมพันธ์กันในระดับสูง และผู้ตรวจเข้มงวด (ค่า r อยู่ช่วง 0.559 – 0.707) ถือว่าผลการตรวจสัมพันธ์กันในระดับปานกลาง แต่ยังสูงกว่าผู้ตรวจใจดี (ค่า r อยู่ช่วง 0.074 – 0.158) ถือว่าผลการตรวจสัมพันธ์กันในระดับต่ำ แสดงดังภาพประกอบ 3



ภาพประกอบ 3 แสดงค่าสัมประสิทธิ์สหสัมพันธ์ของผลการตรวจให้คะแนนของผู้ตรวจที่ต่างกัน

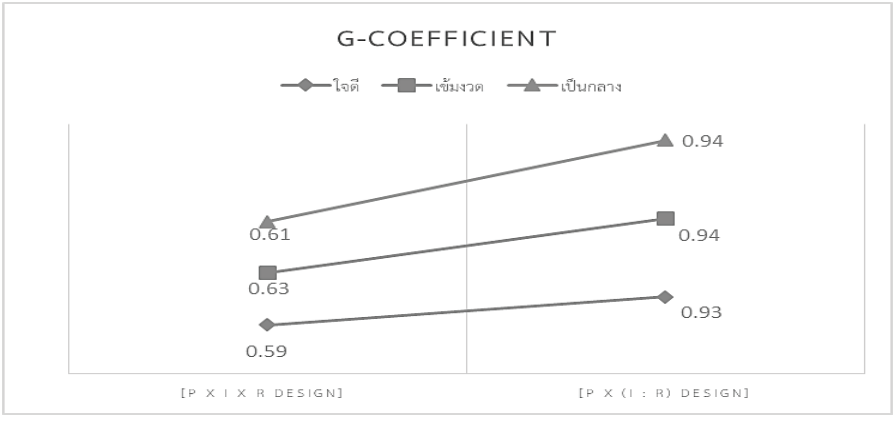
2) ค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนน ภายใต้เงื่อนไขของคุณลักษณะของผู้ตรวจให้คะแนนและรูปแบบการตรวจให้คะแนนที่ต่างกัน ผลการวิเคราะห์ ดังนี้

เมื่อผู้ตรวจให้คะแนนเข้มงวด (raters' severity) ภายใต้เงื่อนไขของรูปแบบการตรวจให้คะแนนที่ต่างกัน พบว่า รูปแบบการตรวจข้อสอบทุกข้อของผู้สอบทุกคน มีค่าสัมประสิทธิ์การสรุปอ้างอิง เท่ากับ 0.630 และรูปแบบการตรวจข้อสอบบางข้อของผู้สอบทุกคนมีค่าสัมประสิทธิ์การสรุปอ้างอิงเท่ากับ 0.940

เมื่อผู้ตรวจให้คะแนนเป็นกลาง (raters' central) ภายใต้เงื่อนไขของรูปแบบการตรวจให้คะแนนที่ต่างกัน พบว่า รูปแบบการตรวจข้อสอบทุกข้อของผู้สอบทุกคน มีค่าสัมประสิทธิ์การสรุปอ้างอิง เท่ากับ 0.610 และรูปแบบการตรวจข้อสอบบางข้อของผู้สอบทุกคนมีค่าสัมประสิทธิ์การสรุปอ้างอิงเท่ากับ 0.940

เมื่อผู้ตรวจให้คะแนนใจดี (raters' leniency) ภายใต้เงื่อนไขของรูปแบบการตรวจให้คะแนนที่ต่างกัน พบว่า รูปแบบการตรวจข้อสอบทุกข้อของผู้สอบทุกคน มีค่าสัมประสิทธิ์การสรุปอ้างอิง เท่ากับ 0.590

และรูปแบบการตรวจข้อสอบบางข้อของผู้สอบทุกคน มีค่าสัมประสิทธิ์การสรุปอ้างอิง เท่ากับ 0.930 แสดงดังภาพประกอบ 4



ภาพประกอบ 4 ค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนน ภายใต้เงื่อนไขของคุณลักษณะของผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน

อภิปรายผล

จากผลการวิเคราะห์ด้วยโปรแกรม FACET พบว่า คุณลักษณะการให้คะแนนของผู้ตรวจมีจำนวนใกล้เคียงกันโดยจำนวนผู้ตรวจเป็นกลาง 13 คน และจำนวนผู้ตรวจเข้มงวดเท่ากับผู้ตรวจใจดี คือ 12 คน นั่นคือคุณลักษณะของผู้ตรวจให้คะแนนที่เหมือนกัน มีความเข้าใจและตรวจได้สอดคล้องกับเกณฑ์การให้คะแนนที่ผู้วิจัยสร้างขึ้น เฉพาะลักษณะการให้คะแนนของแต่ละกลุ่มนั้น ๆ ซึ่งไม่สามารถอธิบายถึงความแม่นยำหรือความถูกต้องของผลการตรวจได้ ผู้วิจัยจึงมีการตรวจสอบผลการให้คะแนนของผู้ตรวจ โดยนำผลการตรวจให้คะแนนของผู้ตรวจแต่ละลักษณะมาเทียบกับคะแนนเกณฑ์ (คะแนนฉันทามติจากผู้เชี่ยวชาญ) เพื่อตรวจสอบความสอดคล้องหรือความตรงตามสภาพของผลการตรวจที่แสดงถึงความแม่นยำ และมีระบบการให้คะแนนที่เหมาะสม (Wolfe, 2004) ซึ่งในงานวิจัยนี้ ได้ตระหนักถึงคุณสมบัติของผู้ตรวจให้คะแนน จึงกำหนดคุณสมบัติเบื้องต้นคือ ต้องผ่านการเรียนในรายวิชาการวัดและประเมินผลการศึกษาและศึกษาในสาขาวิชาคณิตศาสตร์ โดยมีระดับผลการเรียน B+ ขึ้นไป จึงสามารถทำความเข้าใจและตรวจให้คะแนนเนื้อหาในข้อสอบได้ อีกทั้งก่อนดำเนินการให้คะแนน ผู้วิจัยได้ฝึกปฏิบัติจริง เพื่อสร้างความเข้าใจที่ตรงกันระหว่างผู้ตรวจในเรื่องต่าง ๆ เช่น ทบทวนเนื้อหา อธิบายและทำความเข้าใจถึงลักษณะของข้อสอบและเกณฑ์การตรวจ ตลอดจนเปิดโอกาสให้ผู้ตรวจได้ซักถามเพื่อให้ได้ผลการพิจารณาที่ถูกต้องและเข้าใจตรงกัน และผู้วิจัยนำเสนอผลการวิเคราะห์เพื่ออภิปรายผล 2 ประเด็นตามจุดมุ่งหมายของการวิจัย ดังนี้

1) ค่าความตรงตามสภาพของผลการตรวจให้คะแนนแบบสอบอัตนัย 3 ข้อ ระหว่างผู้ตรวจให้คะแนน (คะแนน X) เทียบกับคะแนนเกณฑ์กลาง (คะแนน Y) ซึ่งมาจากคะแนนฉันทามติการตรวจของผู้เชี่ยวชาญ ภายใต้เงื่อนไขของคุณลักษณะของผู้ตรวจให้คะแนน และรูปแบบการตรวจข้อสอบทุกข้อของผู้สอบ

ทุกคน พบว่า ผู้ตรวจให้คะแนนเข้มงวดและเป็นกลางให้ค่าความตรงตามสภาพค่อนข้างสูงกว่าผู้ตรวจใจดี ทั้งนี้เนื่องจากผู้ตรวจเข้มงวดและเป็นกลาง มีค่าคะแนนโลจิทใกล้เคียงกับค่าโลจิทของผู้เชี่ยวชาญ ทำให้ผลการตรวจมีค่าใกล้เคียงกัน จึงทำให้มีค่าความตรงตามสภาพสูงไปด้วย เช่นเดียวกับ บุษวรรษ์ แสนปลื้ม (2556) พบว่า การตรวจสอบความตรงตามสภาพของความสามารถในการเขียนของนักเรียนชั้นประถมศึกษาปีที่ 3 เมื่อผู้ตรวจมีคุณลักษณะเป็นกลาง ตรวจให้คะแนนด้วยวิธีแบบแยกองค์ประกอบในสัดส่วนของจำนวนผู้ตรวจให้คะแนน 2 คน มีความตรงตามสภาพสูงสุด และ อังคณา กุลนภาดล (2555) พบว่า ค่าความตรงตามสภาพของรูปแบบการตรวจข้อสอบทุกข้อของผู้สอบทุกคนสูงกว่ารูปแบบการตรวจอื่นในทุกจำนวนผู้ตรวจที่มากกว่าก็จะมีค่าความตรงตามสภาพสูงกว่าจำนวนผู้ตรวจน้อยในทุกรูปแบบการตรวจให้คะแนน

2) ค่าสัมประสิทธิ์การสรุปอ้างอิง ที่มีรูปแบบการตรวจข้อสอบบางข้อของผู้สอบทุกคนสูงกว่าการตรวจด้วยรูปแบบการตรวจข้อสอบทุกข้อของผู้สอบทุกคน ทุกลักษณะของผู้ตรวจให้คะแนนต่างกัน เช่นเดียวกับ น้ำผึ้ง อินทะเนตร (2554) ; นภัสนันท์ ขวัญจ่า (2556) ได้ประยุกต์ใช้ทฤษฎีการสรุปอ้างอิงในการศึกษา พบว่า ถ้าใช้รูปแบบการตรวจให้คะแนนที่ผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบบางคน (p : r) x i จะให้ค่าสัมประสิทธิ์การสรุปอ้างอิงสูงสูงกว่ารูปแบบ p x i x r และการศึกษาของ ชนิสรา สงวนไว้ (2558) พบว่า ค่าสัมประสิทธิ์การสรุปอ้างอิงของรูปแบบการตรวจ p x (i : r) สูงกว่ารูปแบบ p x i x r และการศึกษาของ จิรายุ เถาว์โท (2559) พบว่า ค่าสัมประสิทธิ์การสรุปอ้างอิงเมื่อรูปแบบการตรวจให้คะแนนเหมือนกันแต่จำนวนผู้ตรวจต่างกัน มีค่าแตกต่างกันอย่างมีนัยสำคัญ ยกเว้นรูปแบบการตรวจ (p : r) x i และ p x (i : r) ของผู้ตรวจ 2 และ 3 คน มีค่าไม่แตกต่างกัน เมื่อใช้จำนวนผู้ตรวจเท่ากัน แต่รูปแบบการให้คะแนนต่างกัน ค่าสัมประสิทธิ์การสรุปอ้างอิงมีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติ จากการศึกษาครั้งนี้ ทำให้ทราบว่าการตรวจให้คะแนนด้วยรูปแบบการตรวจข้อสอบบางข้อของผู้สอบทุกคน จะสามารถช่วยลดภาระงาน เวลาในการตรวจน้อยลงเมื่อเทียบกับรูปแบบการตรวจข้อสอบทุกข้อของผู้สอบทุกคนและให้ค่าสัมประสิทธิ์การสรุปอ้างอิงสูงขึ้น

ข้อเสนอแนะ

1. ข้อเสนอแนะในการนำผลวิจัยไปใช้

1.1 การตัดสินใจเลือกใช้รูปแบบการตรวจให้คะแนนให้มีประสิทธิภาพ ควรพิจารณาคุณภาพของการตรวจได้แก่ ความตรงตามสภาพ และค่าสัมประสิทธิ์การอ้างอิง ทั้งนี้ต้องขึ้นอยู่กับความพร้อมหรือสถานการณ์ในการนำไปใช้เดียวกัน กล่าวคือ

1.1.1 ถ้าเป็นสถานการณ์การตรวจให้คะแนนในระดับชั้นเรียน ซึ่งเป็นสถานการณ์ที่มีจำนวนครูผู้สอนหรือผู้ตรวจที่จำกัดและมีจำนวนนักเรียนไม่มากนัก จึงควรเลือกใช้รูปแบบการตรวจให้คะแนนข้อสอบทุกข้อของผู้สอบทุกคน (p x i x r) ถึงแม้จะเป็นรูปแบบที่ค่อนข้างใช้เวลาในการตรวจนานเมื่อเปรียบเทียบกับรูปแบบแบบแฝง (nested design) หรือ แบบผสม (confounded design) ซึ่งสามารถลดเวลาและภาระในการตรวจให้คะแนนได้มากกว่า แต่ก็เป็นรูปแบบที่สอดคล้องกับสถานการณ์จริงและไม่จำเป็นต้องใช้ผู้ตรวจมากนักเพื่อลดต้นทุนการพัฒนาผู้ตรวจ

1.1.2 ถ้าเป็นสถานการณ์การตรวจให้คะแนนในระดับท้องถิ่นหรือระดับชาติ (large scale) ซึ่งเป็นสถานการณ์การสอบที่มีผู้สอบจำนวนมาก สามารถเลือกรูปแบบการตรวจข้อสอบบางข้อของผู้สอบทุกคน p x (i : r) ซึ่งรูปแบบนี้เหมาะสำหรับสถานการณ์ที่มีผู้ตรวจให้คะแนนจำนวนมาก หรือ สามารถพัฒนาผู้ตรวจที่มีคุณภาพได้จำนวนมากเพียงพอ แต่ตรวจข้อสอบได้เฉพาะบางข้อเท่านั้น ดังนั้น รูปแบบนี้เหมาะกับสถานการณ์ที่มีจำนวนข้อสอบจำนวนมาก นั่นคือ จะทำให้ภาระงานในการตรวจให้คะแนนน้อยกว่าการที่ต้องตรวจข้อสอบทุกข้อของผู้สอบทุกคน และควรพิจารณาคุณภาพการตรวจและค่าคุ้มทุนร่วมด้วยให้สอดคล้องกับสภาพความพร้อมของการบริหารจัดการสอบที่สามารถปฏิบัติได้จริง

1.2 จากการศึกษาการตรวจให้คะแนนแบบสอบอัตนัยถึงแม้ว่าผู้ตรวจจะมีลักษณะการให้คะแนนที่ต่างกัน ก็สามารถทำให้ค่าสัมประสิทธิ์การสรุปอ้างอิงสูงขึ้นได้ เมื่อพิจารณาเลือกใช้รูปแบบการตรวจ [p x (i : r)] และเมื่อพิจารณาคุณลักษณะการให้คะแนนของผู้ตรวจที่ต่างกัน ควรมีการพัฒนาผู้ตรวจให้มีความแม่นยำในการตรวจให้คะแนน และ หากผู้ที่จะนำแบบสอบอัตนัยไปใช้ควรทำความเข้าใจเกี่ยวกับการบริหารการสอบ รูปแบบการตรวจให้คะแนนรวมถึงระยะเวลาในการทำข้อสอบให้มีความเหมาะสม

2. ข้อเสนอแนะในการวิจัยครั้งต่อไป

2.1 ควรศึกษาเพิ่มเติมถึงลักษณะการให้คะแนนที่ไม่สามารถระบุรูปแบบการตรวจ หรือ ความแม่นยำในการตรวจ เมื่อผู้ตรวจมีการจับคู่ลักษณะการให้คะแนนที่เหมือนกันหรือต่างกัน หรือลักษณะการทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป เพื่อเปรียบเทียบความตรงตามสภาพ หรือ ค่าสัมประสิทธิ์การสรุปอ้างอิง ว่าแบบแผนใดที่ให้ค่าความน่าเชื่อถือสูงสุด

2.2 จากผลการวิจัยที่พบว่า ผู้ตรวจให้คะแนนมีความแตกต่างกันในการให้คะแนนใกล้เคียงกัน แต่มีความเข้มงวด/ใจดีที่แตกต่างกัน เพื่อให้เกิดความเข้าใจในความคลาดเคลื่อนที่เกิดจากผู้ตรวจให้คะแนน ควรมีการศึกษาความคลาดเคลื่อนที่เกิดจากผู้ตรวจ อาทิ อิทธิพลการให้คะแนนแบบแม่นยำ (accuracy effect) การให้คะแนนตรงกลาง (central tendency) อิทธิพลจากฮาโล (Halo effect) ความลำเอียงในการให้คะแนน (bias rating) เป็นต้น

2.3 ควรมีการศึกษาหาแหล่งความแปรปรวนหรือแหล่งความคลาดเคลื่อนจากการตรวจให้คะแนนของข้อสอบที่มีระดับความลึกต่างกัน เช่น การศึกษารูปแบบเกณฑ์การให้คะแนนที่แตกต่างกัน ระดับความยากง่ายของแบบสอบอัตนัย และ ความเฉพาะเจาะจงของเนื้อหาสาระในแบบสอบอัตนัยที่ส่งผลต่อความน่าเชื่อถือของผลการตรวจให้คะแนน

เอกสารอ้างอิง

กมลวรรณ ตังธนกานนท์. (2557). *การวัดและประเมินทักษะการปฏิบัติ.* พิมพ์ครั้ง 1.กรุงเทพฯ : สำนักพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

จิรายุ เถาว์โท, อนุ เจริญวงศ์ระยับ และปัณณวิชญ์ ใบกุหลาบ. (2559). การศึกษาค่าความเชื่อมั่นของคะแนนแบบทดสอบอัตนัยวิชาคณิตศาสตร์ของนักเรียนชั้นมัธยมศึกษาปีที่ 2 ที่มีจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน โดยใช้ทฤษฎีการสรุปอ้างอิง.*วารสารหาดใหญ่วิชาการ*, 14(1),1-14.

ชนิสรา สงวนไว้. (2558). *การเปรียบเทียบความเที่ยงของแบบสอบวัดความสามารถในการแก้ปัญหาอย่างสร้างสรรค์ทางคณิตศาสตร์ : การประยุกต์ใช้ทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัด.* ปริญญานิพนธ์ ค.ม. (การวัดและประเมินผลการศึกษา). กรุงเทพฯ: บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย.

นภัสนันท์ ขวัญจ่า. (2556). *การเปรียบเทียบสัมประสิทธิ์การสรุปอ้างอิงของแบบวัดทักษะกระบวนการทางวิทยาศาสตร์ ชั้นมัธยมศึกษาปีที่ 4 ที่มีรูปแบบการตรวจให้คะแนนต่างกัน.* วิทยานิพนธ์ กศ.ม. (การวัดผลการศึกษา). มหาสารคาม: บัณฑิตวิทยาลัย มหาวิทยาลัยมหาสารคาม.

น้ำผึ้ง อินทะเนตร. (2554). *การศึกษาคุณลักษณะของคะแนนแบบทดสอบปลายเปิดวิชาคณิตศาสตร์ เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน โดยใช้โมเดลการสรุปอ้างอิงและโมเดลหลายองค์ประกอบของราส์ช.*วิทยานิพนธ์ กศ.ด.(การทดสอบและวัดผลการศึกษา). กรุงเทพฯ: บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ.

บุษวรรษ์ แสนปลื้ม. (2556). *การใช้วิธีการตรวจคุณลักษณะและสัดส่วนจำนวนผู้ตรวจให้คะแนนที่มีต่อความเที่ยงตรงของการวัดความสามารถในการเขียนของนักเรียนชั้นประถมศึกษาปีที่ 3.* วิทยานิพนธ์ กศ.ด. (การทดสอบและวัดผลการศึกษา). กรุงเทพฯ: บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ.

ปรีชา เนาว์เย็นผล. (2544). *กิจกรรมการเรียนการสอนคณิตศาสตร์โดยใช้การแก้ปัญหาปลายเปิดสำหรับนักเรียนชั้นมัธยมศึกษาปีที่ 1.* วิทยานิพนธ์ กศ.ด.(คณิตศาสตร์ศึกษา). กรุงเทพฯ : บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ.

ผจงจิต อินทสุวรรณ. (2525). Latent Trait Theory. *วารสารการวัดผลการศึกษา*, 3(3): 51-69.

พวงรัตน์ ทวีรัตน์. (2530). *การสร้างและพัฒนาแบบทดสอบวัดผลสัมฤทธิ์.* สำนักทดสอบทางการศึกษาและจิตวิทยา, มหาวิทยาลัยศรีนครินทรวิโรฒ.

พรรณี เจียมสุบุตร. (2543). *การเปรียบเทียบความเชื่อมั่นของแบบทดสอบวัดความสามารถในการแก้โจทย์ปัญหาทางคณิตศาสตร์ ที่มีจำนวนผู้ตรวจและวิธีการตรวจต่างกัน.* วิทยานิพนธ์ กศ.ด. (การทดสอบและวัดผลการศึกษา). กรุงเทพฯ: บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ. ถ่ายเอกสาร.

ศิริชัย กาญจนวาสี. (2555). *ทฤษฎีการทดสอบแนวใหม่.พิมพ์ครั้ง 4.* กรุงเทพฯ : สำนักพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

_____. (2556). *ทฤษฎีการทดสอบแบบดั้งเดิม.พิมพ์ครั้ง 7.*กรุงเทพฯ: สำนักพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

ศุกลรัตน์ อิงชาติเจริญ. (2557). *การพัฒนาโมเดลคุณภาพการให้คะแนนระหว่างกลุ่มผู้ประเมินในวิชาที่มีการเรียนรู้โดยใช้ปัญหาเป็นฐาน: การประยุกต์ใช้โมเดลหลายองค์ประกอบของราสช์.* กรุงเทพฯ: บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย.

อุษณี บัวศิริพันธ์. (2543). *การเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิงของแบบทดสอบวิชาคณิตศาสตร์ที่มีวิธีการตรวจ จำนวนผู้ตรวจและประสบการณ์ของผู้ประเมิน.* วิทยานิพนธ์ กศ.ด. (การทดสอบและวัดผลการศึกษา). กรุงเทพฯ: บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ. ถ่ายเอกสาร.

อังคณา กุลนภาดล. (2557). การเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนผังมโนทัศน์วิชาการวิจัยทางการศึกษา เมื่อรูปแบบการตรวจและจำนวนผู้ตรวจต่างกัน.*วารสารศึกษาศาสตร์* มหาวิทยาลัย*บูรพา*, 25(2).

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing A revision of Bloom's taxonomy of educational objectives.* New York: Addison Wesley Longman.

Brennan, R.L. (2001). *Generalizability theory*. New York: Springer.

Hopkins, C.D. & Antes, R.L. (1990). *Classroom Measurement and Evaluation*. 3rd Ed. Itasca, IL.

Iramaneerat, C., Yudkowsky, R., Myford, C. M., & Downing, S. M. (2008). *Quality control of an OSCE using generalizability theory and many-faceted Rasch measurement*. *Advances in Health Sciences Education*, *13*(4), 479.

Linacre, J. M. (1994). *Many-Facet Rasch Measurement*. Chicago: MESA Press.

_____. (2014). *FACETS (Version 3.71. 4) [Computer software].* Beaverton, Oregon: Winsteps.com

Wolfe, E. W, &. Myford, C. M. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of applied measurement*, 4(4), 386-422.

Mehrens, William A. & Lehmann, Irvin J.(1972). *Measurement and Evaluation in Education and Psychology.* New York : Holt, Rinehart and Winston.

Rui, Y. (2010). *A Many-facet Rasch Analysis of Rater Effects on an Oral English Proficiency Test.* Doctor of Philosophy. Purdue University West Lafayette, Indiana.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological bulletin*, 88(2), 413.

Smith, P. L. (1978). Sampling errors of variance components in small sample multifaceted generalizability studies. *Journal of Educational Statistics*, 3(4), 319-346.

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2004). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, 9(3), 239-261.