

การเปรียบเทียบความเที่ยงและความตรงตามเกณฑ์ในการตรวจให้คะแนนของผู้ตรวจ
ที่มีคุณลักษณะต่างกันของแบบสอบอัตนัยสำหรับวัดสมรรถนะ
ทางวิทยาศาสตร์ของนักเรียนชั้นมัธยมศึกษาปีที่ 3

The Comparison of Reliability and Validity of the Raters' Scoring Criteria
with Different Characteristics of Raters of Essay Tests for Measuring Scientific
Competence of Grade 9 Students

พนิดา จังหะ^{1*} และ ประกฤติยา ทักซิโน²

Phanida Changwa^{1*} and Prakittiya Tuksino²

(Received: April 8, 2022; Revised: June 14, 2022; Accepted: June 21, 2022)

บทคัดย่อ

การวิจัยในครั้งนี้มีความมุ่งหมายของการวิจัย คือ 1) เพื่อศึกษาค่าความสอดคล้องของผลการตรวจให้คะแนนแบบสอบอัตนัยวัดสมรรถนะทางวิทยาศาสตร์ ภายใต้คุณลักษณะของผู้ตรวจที่แตกต่างกัน โดยใช้สถิติสหสัมพันธ์ภายในชั้น (Intra-Class Correlation: ICC) 2) เพื่อเปรียบเทียบความตรงตามเกณฑ์การให้คะแนนแบบองค์รวม ของแบบสอบอัตนัยสำหรับวัดสมรรถนะทางวิทยาศาสตร์ โดยพิจารณาจากความเห็นพ้อง (rater agreement) ภายใต้คุณลักษณะของผู้ตรวจที่แตกต่างกัน 3. เพื่อเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิง (G-coefficient) ภายใต้คุณลักษณะของผู้ตรวจที่แตกต่างกัน ที่มีรูปแบบการตรวจ 2 รูปแบบ ได้แก่ รูปแบบการตรวจข้อสอบบางข้อของผู้สอบทุกคน [$(p \times i : r)$] และ รูปแบบการตรวจทุกข้อของผู้สอบทุกคน [$(p \times i \times r)$] กลุ่มตัวอย่างแบ่งออกเป็น 2 กลุ่ม คือ กลุ่มนักเรียนชั้นมัธยมศึกษาปีที่ 3 จำนวน 100 คน และกลุ่มผู้ตรวจให้คะแนน ซึ่งแบ่งออกเป็นกลุ่มผู้ตรวจให้คะแนนที่เป็นผู้ตรวจที่มีคุณวุฒิตรงเอก จำนวน 3 คน และกลุ่มผู้ตรวจให้คะแนนที่เป็นผู้ตรวจที่มีคุณวุฒิไม่ตรงเอก จำนวน 3 คน เครื่องมือที่ใช้ในการวิจัย คือ 1) แบบสอบอัตนัยวัดสมรรถนะทางวิทยาศาสตร์ของนักเรียนชั้นมัธยมศึกษาปีที่ 3 จำนวน 3 สถานการณ์ รวม 9 ข้อคำถาม 2) เกณฑ์การให้คะแนนแบบองค์รวม (holistic scoring rubric) วิเคราะห์ค่าสัมประสิทธิ์การสรุปอ้างอิงโดยใช้โปรแกรม EduG ผลการวิจัยพบว่า 1) การศึกษาความสอดคล้องของผลการตรวจให้คะแนนระหว่างผู้ตรวจ พิจารณาโดยใช้สถิติสหสัมพันธ์ภายในชั้น (Intra-Class Correlation: ICC) ที่มีความสอดคล้องกัน

¹ นักศึกษาปริญญาโท สาขาวิชาการวัดและประเมินผลการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยขอนแก่น

² ผู้ช่วยศาสตราจารย์ สาขาวิชาการวัดและประเมินผลการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยขอนแก่น

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

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

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

โดยเฉลี่ย พบว่าความสอดคล้องของผลการตรวจให้คะแนน ทั้ง 3 สถานการณ์ 9 ข้อ อยู่ในระดับสอดคล้องต่ำไปจนถึงระดับดีมาก ทั้งกลุ่มผู้ตรวจให้คะแนนที่เป็นกลุ่มผู้ตรวจที่มีคุณวุฒิตรงเอก และกลุ่มผู้ตรวจที่มีคุณวุฒิไม่ตรงเอก 2) ค่าความตรงตามเกณฑ์การให้ตรวจให้คะแนนโดยพิจารณาจากความเห็นพ้อง (rater agreement) ของผลการตรวจให้คะแนนของผู้ตรวจ (x) กับเกณฑ์มาตรฐาน (y) พบว่าในกลุ่มผู้ตรวจที่มีคุณวุฒิตรงเอก มีค่าความเห็นพ้อง ตั้งแต่ร้อยละ 14 ถึง ร้อยละ 84 และในกลุ่มผู้ตรวจที่มีคุณวุฒิไม่ตรงเอกมีค่าความเห็นพ้อง ตั้งแต่ ร้อยละ 27 ถึง ร้อยละ 89 3) ค่าสัมประสิทธิ์การสรุปอ้างอิงของรูปแบบการตรวจข้อสอบบางข้อของผู้สอบทุกคน [$p \times (i : r)$] มีค่าสูงกว่า รูปแบบการตรวจทุกข้อของผู้สอบทุกคน [$p \times i \times r$] ทั้งกลุ่มผู้ตรวจที่มีคุณวุฒิตรงเอก และกลุ่มผู้ตรวจที่มีคุณวุฒิไม่ตรงเอก

คำสำคัญ: เกณฑ์การตรวจให้คะแนนแบบองค์รวม ดัชนีความเห็นพ้องของการตรวจ ทฤษฎีการสรุปอ้างอิง
ค่าความเชื่อมั่นของผลการตรวจ

Abstract

This research aimed 1) to study the inter-rater reliability of the raters' scoring criteria of the essay test for measuring scientific competence by Intra-Class Correlation: ICC under different raters' characteristics. 2) to compare the validity of the raters' scoring criteria with the holistic scoring rubric of the essay test by Rater Agreement under different raters' characteristics. And 3) to compare the G-coefficient, under different raters' characteristics, for Cross design [$p \times i \times r$] and Nested design [$p \times (i : r)$]. The sample was divided into 2 groups: the group of 100 Grade 9 students and the group of raters. The group of raters comprised 3 raters who were science majors and another 3 raters who were non-science majors. The research instruments were 1) an essay test to measure the scientific competency of grade 9 students in 3 situations, containing 9 questions; and 2) a holistic scoring rubric. Generalizability Coefficient scores were analyzed by EduG. The research findings were 1) the reliability of the scoring results for each item, analyzed by the Intra-Class Correlation (ICC) statistics, were found to be from low to very good for all of the raters in the group, both science majors and non-science majors. 2) the validity of the raters' scoring criteria for each item analyzed by Rater Agreement between the raters' scores (x) and standard scores (y) revealed that the 3 raters who were science majors had the agreement index from 14 percent to 84 percent, and the 3 raters who were non-science majors had the agreement index from 27 percent to 89 percent. 3) The Generalizability Coefficient scores of the $p \times (i : r)$ design was higher than the $p \times i \times r$ design for all raters in the group.

Keywords: holistic scoring rubric, rater agreement, G Theory, inter-rater reliability

บทนำ

การนำประเทศเข้าสู่สังคมโลกในศตวรรษที่ 21 นั้น จำเป็นอย่างยิ่งที่จะต้องเตรียมความพร้อมในเรื่องของทรัพยากรมนุษย์ ให้มีทั้งความรู้ ความสามารถ และสมรรถนะที่จำเป็นให้สามารถปรับตัว และรู้เท่าทันต่อกระแสความเปลี่ยนแปลงของโลก (Ministry of Education, 2017) การจัดการศึกษาจะต้องพัฒนาและสามารถสร้างเสริมทักษะสำคัญในศตวรรษที่ 21 ซึ่งหนึ่งในสิ่งที่ผู้เรียนจำเป็นต้องมี คือ การรู้เรื่องทางวิทยาศาสตร์ (scientific literacy) The Institute for the Promotion of Teaching Science and Technology, (2020) ได้กล่าวถึง บุคคลที่ได้ชื่อว่าฉลาดรู้ด้านวิทยาศาสตร์ (scientifically literate person) จะต้องเป็นผู้ที่สามารถใช้สมรรถนะทางวิทยาศาสตร์ ได้แก่ 1) การอธิบายปรากฏการณ์ในเชิงวิทยาศาสตร์ (explain phenomena scientifically) 2) การประเมินและออกแบบกระบวนการสืบเสาะหาความรู้ทางวิทยาศาสตร์ (evaluate and design scientific enquiry) และ 3) การแปลความหมายข้อมูลและการใช้ประจักษ์พยานในเชิงวิทยาศาสตร์ (interpret data and evidence scientifically) ดังนั้น ในการจัดการเรียนรู้ จึงจำเป็นที่จะต้องมีการวัดและประเมินผล เพื่อให้ทราบว่าผู้เรียนเกิดการเรียนรู้ตามวัตถุประสงค์ของการจัดกิจกรรมการเรียนรู้หรือไม่ (Tuksino, 2013)

แบบสอบที่จะวัดสมรรถนะทางวิทยาศาสตร์ของผู้เรียนได้ดี ควรจะเลือกใช้แบบสอบที่กระตุ้นให้ผู้เรียนได้แสดงความคิดความเข้าใจ และความรู้ความสามารถของผู้เรียนผ่านการเขียนแสดงวิธีการคิดและหาคำตอบ จึงจำเป็นต้องมีการสร้างแบบสอบอัตนัย (essay test) เพื่อเปิดโอกาสให้ผู้เรียนได้แสดงออกอย่างหลากหลายทั้งในด้านของวิธีการแก้ปัญหาการสื่อสารวิธีการแก้ปัญหาของผู้เรียน การให้เหตุผล รวมทั้งการแสดงความคิดสร้างสรรค์ แต่แบบสอบอัตนัย (essay test) ยังคงมีปัญหาในเรื่องของความเที่ยงต่ำ สิ้นเปลืองเวลา แรงงาน และค่าใช้จ่ายในการตรวจมาก (Coffman, 1971; Chiu & Wolfe, 2002) และจุดอ่อนที่สำคัญที่สุดคือเรื่องของการตรวจให้คะแนน (Pinyoanuntapong, 2004) การตรวจให้คะแนนจึงไม่ควรใช้ผู้ตรวจคนเดียว เพราะจะทำให้ขาดความเที่ยงในการให้คะแนน ดังที่ผลการวิจัย พบว่า เมื่อเพิ่มจำนวนผู้ตรวจ จะทำให้ความเที่ยงสูงขึ้น (Brennan & Johnson, 1995; Swartz et al., 1999) และควรควบคุมแหล่งความคลาดเคลื่อนอื่น เช่น ข้อสอบ ผู้ตรวจ (Phadungphon, 2017) การวัดและประเมินผู้เรียนด้วยแบบสอบอัตนัย (essay test) ยังคงมีปัญหาในเรื่องของความเที่ยงของคะแนนที่ได้ ซึ่งจะขึ้นอยู่กับผู้ตรวจเป็นหลัก และปัจจุบันปัญหาของระบบการศึกษาไทย ยังคงมีปัญหาในเรื่องของอัตราค่าจ้างการบรรจุครู ทำให้โรงเรียนขนาดเล็กจำนวนมาก มีปัญหาในเรื่องของการขาดแคลนครู ส่งผลให้ภายในโรงเรียนนั้น มีครูที่สอนไม่ครบชั้น และสอนไม่ตรงตามคุณวุฒิ หรือวิชาเอกที่จบมา ทำให้ในการวัดประเมินสมรรถนะทางวิทยาศาสตร์ ยังคงมีคุณครูที่สอนไม่ตรงตามคุณวุฒิ หรือวิชาเอกที่จบมา มาเป็นผู้ตรวจแบบสอบ และทำให้ค่าความเที่ยงของการให้คะแนนต่ำ (Phusing, 2020) ดังนั้น เพื่อให้การตรวจมีความเที่ยงเพิ่มมากขึ้น จะต้องมีการออกแบบวิธีการตรวจให้คะแนน และเกณฑ์การให้คะแนนที่เหมาะสม เพื่อให้เกิดความเป็นปรนัยในการตรวจ และมีความเที่ยงเพิ่มมากขึ้น (Umnacil, 2014) และนำเอาทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัดมาช่วยในการประเมินค่าสัมประสิทธิ์การสรุปอ้างอิงคะแนนของแบบสอบ (Kanjawasee, 2007) โดยวิธีการออกแบบฟาเซตซึ่งฟาเซตที่นิยมนำมาออกแบบ คือ 2 ฟาเซต ได้แก่ จำนวนข้อสอบ (i) และจำนวนผู้ตรวจ (r) และ

มีวิธีการออกแบบการตรวจให้คะแนน 2 ลักษณะ คือ ผู้ตรวจตรวจให้คะแนนทุกข้อของผู้สอบทุกคน (two-facet crossed design) หรือ $p \times i \times r$ และ ผู้ตรวจตรวจข้อสอบบางข้อของผู้สอบทุกคน (two-facet nested design) หรือ $p \times (i : r)$ ซึ่งผลการวิจัยพบว่า เมื่อเพิ่มจำนวนผู้ตรวจจะทำให้ความเที่ยงสูงขึ้น (Brennan & Johnson, 1995; Swartz et al., 1999; Intanate, 2011; Sanguanwai, 2015)

ดังนั้น ผู้วิจัยจึงได้สนใจที่จะพัฒนาเกณฑ์การให้คะแนนแบบองค์รวม (holistic scoring rubric) เพื่อใช้ในการตรวจให้คะแนนแบบสอบอัตนัยสำหรับการสมรรถนะทางวิทยาศาสตร์ของนักเรียนชั้นมัธยมศึกษาปีที่ 3 และศึกษาเกี่ยวกับการเปรียบเทียบความเที่ยงและความตรงตามเกณฑ์การให้คะแนนแบบองค์รวม (holistic scoring rubric) ของแบบสอบอัตนัยสำหรับการสมรรถนะทางวิทยาศาสตร์ ภายใต้คุณลักษณะของผู้ตรวจที่แตกต่างกัน คือ กลุ่มผู้ตรวจที่มีคุณวุฒิตรงเอก เป็นครูผู้สอนในรายวิชาวิทยาศาสตร์ และกลุ่มผู้ตรวจที่มีคุณวุฒิไม่ตรงเอก ที่เป็นครูผู้สอนในรายวิชาอื่น ๆ ได้แก่ ครูผู้สอนรายวิชาภาษาไทย ครูผู้สอนรายวิชาคณิตศาสตร์ และครูผู้สอนรายวิชาสังคมศึกษา โดยประยุกต์ใช้ทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัด ซึ่งเป็นทฤษฎีทางสถิติสำหรับวิเคราะห์ความน่าเชื่อถือของผลการวัดในสถานการณ์การวัดในลักษณะต่าง ๆ ที่เป็นเป้าหมายของการนำเครื่องมือไปใช้ เพื่อนำผลที่ได้ไปใช้ตัดสินใจการสรุปอ้างอิงการตรวจแบบสอบอัตนัยสำหรับวัดสมรรถนะทางวิทยาศาสตร์ ว่าควรผู้ตรวจที่มีคุณลักษณะที่แตกต่างกัน เมื่อใช้เกณฑ์การตรวจให้คะแนนแบบองค์รวม ในตรวจให้คะแนนแบบสอบอัตนัยวัดสมรรถนะทางวิทยาศาสตร์ ควรใช้รูปแบบการตรวจให้คะแนนแบบใด จึงจะมีค่าความเที่ยงของการให้คะแนนที่เหมาะสม และเพื่อให้ได้แบบสอบและเกณฑ์การให้คะแนนของแบบสอบอัตนัยสำหรับการวัดสมรรถนะทางวิทยาศาสตร์ของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ที่มีมาตรฐาน สามารถนำไปใช้วัดสมรรถนะทางวิทยาศาสตร์ของผู้เรียนได้จริง

ความมุ่งหมายของการวิจัย

1. เพื่อศึกษาค่าความสอดคล้องของผลการตรวจให้คะแนนแบบสอบอัตนัยวัดสมรรถนะทางวิทยาศาสตร์ ภายใต้คุณลักษณะของผู้ตรวจที่แตกต่างกัน
2. เพื่อเปรียบเทียบความตรงตามเกณฑ์การให้คะแนนแบบองค์รวม ของแบบสอบอัตนัยสำหรับวัดสมรรถนะทางวิทยาศาสตร์ ภายใต้คุณลักษณะของผู้ตรวจที่แตกต่างกัน
3. เพื่อเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิง ภายใต้คุณลักษณะของผู้ตรวจที่แตกต่างกัน ที่มีรูปแบบการตรวจ 2 รูปแบบ ได้แก่ รูปแบบการตรวจข้อสอบบางข้อของผู้สอบทุกคน $[(p \times (i : r))]$ และรูปแบบการตรวจทุกข้อของผู้สอบทุกคน $[p \times i \times r]$

วิธีดำเนินการวิจัย

1. ประชากรและกลุ่มตัวอย่าง

ประชากรที่ใช้ในการศึกษาแบ่งออกเป็น 2 กลุ่ม ได้แก่ กลุ่มผู้สอบและกลุ่มผู้ตรวจให้คะแนน

- 1) กลุ่มผู้สอบ คือ นักเรียนชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2564 สังกัดสำนักงานเขตพื้นที่การศึกษา มัธยมศึกษาขอนแก่น จำนวน 84 โรงเรียน มีจำนวนนักเรียนทั้งสิ้น 10,575 คน
- 2) กลุ่มผู้ตรวจให้คะแนน

เป็นคุณครูผู้ที่มีใบประกอบวิชาชีพครูและมีวุฒิการศึกษาระดับปริญญาตรี มีจำนวนทั้งสิ้น 9 คน

กลุ่มตัวอย่างที่ใช้ในการวิจัยแบ่งออกเป็น 2 กลุ่ม ได้แก่ กลุ่มผู้สอบและกลุ่มผู้ตรวจให้คะแนน แบบสอบอัตนัย 1) กลุ่มผู้สอบ คือ นักเรียนชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2564 สังกัดสำนักงานเขตพื้นที่ การศึกษามัธยมศึกษาขอนแก่น ได้มาจากการใช้การสุ่มหลายขั้นตอน (multistage random sampling) กล่าวคือ ขั้นที่ 1. แบ่งโรงเรียนโดยใช้ขนาดของโรงเรียนเป็นเกณฑ์ ได้แก่ โรงเรียนขนาดเล็ก กลาง ใหญ่ และ ขนาดใหญ่พิเศษ ขั้นที่ 2 สุ่มโรงเรียน โดยการสุ่มอย่างง่าย (simple random sampling) เลือกตัวแทน โรงเรียนแต่ละขนาด จะได้ตัวแทนจำนวน 4 โรงเรียน และขั้นที่ 3 ใช้การสุ่มผู้เรียนโดยการสุ่มอย่างง่าย (simple random sampling) โรงเรียนละ 25 คน รวมเป็นนักเรียนจำนวน 100 คน 2) ผู้ตรวจให้คะแนน แบบสอบอัตนัย จะแบ่งออกเป็น 2 กลุ่ม คือ 2.1) กลุ่มผู้ตรวจที่มีคุณวุฒิตรงเอก คือ คุณครูที่สอนในรายวิชา วิทยาศาสตร์ จำนวน 3 คน โดยกำหนดคุณสมบัติ คือ เป็นผู้ที่มีใบประกอบวิชาชีพครูและมีวุฒิการศึกษาระดับ ปริญญาตรีในสาขาของวิทยาศาสตร์ ได้แก่ คณะศึกษาศาสตร์ คณะครุศาสตร์ และคณะวิทยาศาสตร์ สาขาวิชา ฟิสิกส์ เคมี ชีววิทยา และวิทยาศาสตร์ทั่วไป 2.2) กลุ่มผู้ตรวจที่มีคุณวุฒิไม่ตรงเอก คือ คุณครูที่สอนในรายวิชา อื่น ๆ จำนวน 3 คน โดยกำหนดคุณสมบัติ คือ เป็นผู้ที่มีใบประกอบวิชาชีพครูและมีวุฒิการศึกษาระดับ ปริญญาตรีในสาขาอื่น ๆ ที่ไม่ใช่สาขาของวิทยาศาสตร์ ซึ่งในงานวิจัยนี้ มีคุณครูที่สอนในรายวิชาภาษาไทย คณิตศาสตร์ และสังคมศึกษา โดยใช้วิธีการเลือกกลุ่มตัวอย่างแบบเจาะจง (purposive sampling) เนื่องจาก งานวิจัยในครั้งนี้ต้องใช้ครูผู้สอนทั้งที่มีคุณวุฒิตรงเอกและคุณวุฒิไม่ตรงเอกในการตรวจให้คะแนนแบบสอบ อัตนัยวัดสมรรถนะทางวิทยาศาสตร์ โดยใช้เกณฑ์การให้คะแนนแบบองค์รวม (holistic scoring rubric) ซึ่งต้องใช้เวลา และความเต็มใจในการตรวจให้คะแนนของนักเรียน

2. เครื่องมือที่ใช้ในการวิจัย

เครื่องมือที่ใช้ในการวิจัย ได้แก่ แบบสอบอัตนัยวัดสมรรถนะทางวิทยาศาสตร์สำหรับนักเรียน ชั้นมัธยมศึกษาปีที่ 3 มีจำนวน 3 สถานการณ์ 9 ข้อ โดยประกอบด้วย 1) สถานการณ์ผลกระทบของ แผ่นดินไหว ที่ส่งผลกระทบต่อระบบนิเวศ 2) สถานการณ์ของการดำรงชีวิตของพืชและสิ่งแวดล้อม 3) สถานการณ์ ของประโยชน์และโทษของพันธุวิศวกรรม โดยทั้ง 3 สถานการณ์จะมีข้อย่อย จำนวน 3 ข้อย่อย ที่กระตุ้นให้ ผู้เรียนแสดงออกถึงความรู้ด้านเนื้อหาวิทยาศาสตร์ตามกรอบการประเมินสมรรถนะ 3 ด้าน ของ PISA ได้แก่ การอธิบายปรากฏการณ์ในเชิงวิทยาศาสตร์, การประเมินและออกแบบกระบวนการสืบเสาะหาความรู้ ทางวิทยาศาสตร์ และการแปลความหมายข้อมูลและการใช้ประจักษ์พยานในเชิงวิทยาศาสตร์ จากนั้นนำไป ตรวจสอบความตรงเชิงเนื้อหาโดยผู้เชี่ยวชาญ และนำแบบสอบดังกล่าวไปทดลองใช้ (try out) กับกลุ่มทดลอง ซึ่งเป็นนักเรียนชั้นมัธยมศึกษาปีที่ 3 จำนวน 50 คน และตรวจให้คะแนนโดยผู้ตรวจให้คะแนน จำนวน 3 คน เพื่อตรวจสอบคุณภาพของแบบสอบ พบว่า ค่าความตรงเชิงเนื้อหา (IOC) มีค่าตั้งแต่ 0.80 ถึง 1.00 ดัชนีค่า ความยากง่าย (p) มีค่าตั้งแต่ 0.42 ถึง 0.74 และมีค่าอำนาจจำแนก (r) มีค่าตั้งแต่ 0.50 ถึง 0.86 และ เกณฑ์ การให้คะแนนแบบองค์รวม (holistic scoring rubric) โดยมีการพัฒนาและกำหนดประเด็นการประเมิน 3 ประเด็น ตามกรอบการประเมินสมรรถนะของ PISA ได้แก่ 1. การอธิบายปรากฏการณ์ในเชิงวิทยาศาสตร์ 2. การประเมินและออกแบบกระบวนการสืบเสาะหาความรู้ทางวิทยาศาสตร์ และ 3. การแปลความหมาย

ข้อมูลและการใช้ประจักษ์พยานในเชิงวิทยาศาสตร์ และมีคะแนนประเด็นละ 3 คะแนน รวมข้อสอบทั้งฉบับมีคะแนนเต็ม 27 คะแนน และจากนั้นนำไปตรวจสอบความตรงเชิงเนื้อหา (IOC) พบว่า มีค่าตั้งแต่ 0.80 ถึง 1.00

3. การเก็บรวบรวมข้อมูล

ผู้วิจัยนำแบบสอบอัตนัยวัดสมรรถนะทางวิทยาศาสตร์ของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ที่ผ่านการปรับปรุงแก้ไขแล้ว ไปเก็บข้อมูลกับกลุ่มตัวอย่าง จำนวน 100 คน แล้วนำผลการตอบมาตรวจให้คะแนนโดยใช้เกณฑ์การให้คะแนนแบบองค์รวม (holistic scoring rubric) และใช้กลุ่มผู้ตรวจให้คะแนนจำนวน 2 กลุ่ม ได้แก่ กลุ่มผู้ตรวจที่มีคุณวุฒิตรงเอก จำนวน 3 คน และกลุ่มผู้ตรวจที่มีคุณวุฒิไม่ตรงเอก จำนวน 3 คน จากนั้นนำผลการตรวจให้คะแนนมาวิเคราะห์ข้อมูล ได้แก่ 1. วิเคราะห์ค่าความสอดคล้องของผลการตรวจให้คะแนน โดยใช้สถิติสหสัมพันธ์ภายในชั้น (Intra-Class Correlation: ICC) 2. วิเคราะห์ค่าสัมประสิทธิ์การสรุปอ้างอิง และ 3. วิเคราะห์ค่าความตรงตามเกณฑ์การให้ตรวจให้คะแนนโดยพิจารณาจากความเห็นพ้อง (rater agreement) ของผลการตรวจให้คะแนนของผู้ตรวจ (x) กับเกณฑ์มาตรฐาน (y) โดยใช้โปรแกรมสำเร็จรูป SPSS โดยนำผลการตรวจให้คะแนนโดยใช้เกณฑ์การให้คะแนนแบบองค์รวม (holistic scoring rubric) มาวิเคราะห์หาค่าความตรงของเกณฑ์การให้ตรวจให้คะแนนโดยพิจารณาจากความเห็นพ้อง (rater agreement) ของผลการตรวจให้คะแนนของผู้ตรวจ (x) กับเกณฑ์มาตรฐาน (y) โดยคะแนนเกณฑ์มาตรฐาน (y) ได้มาจากคะแนนชั้นตามติ จากผู้เชี่ยวชาญที่มีคุณสมบัติ คือ ครูที่มีวิทยฐานะชำนาญการพิเศษ และจบการศึกษาในระดับปริญญาโท สาขาทางวิทยาศาสตร์ และสาขาการวัดและประเมินผล จำนวน 5 คน และตรวจให้คะแนนโดยใช้เกณฑ์การให้คะแนนแบบองค์รวม (holistic scoring rubric) และ 3. วิเคราะห์ค่าสัมประสิทธิ์การสรุปอ้างอิงเชิงตัดสินใจสัมพัทธ์ โดยใช้โปรแกรม EduG ในการตรวจให้คะแนนของผู้ตรวจที่มีคุณวุฒิตรงเอก และผู้ตรวจที่มีคุณวุฒิไม่ตรงเอก ที่มีรูปแบบการตรวจ 2 รูปแบบ ได้แก่ รูปแบบการตรวจข้อสอบบางข้อของผู้สอบทุกคน $[p \times (i : r)]$ และ รูปแบบการตรวจทุกข้อของผู้สอบทุกคน $[p \times i \times r]$

4. การวิเคราะห์ข้อมูล

1) การวิเคราะห์ค่าความสอดคล้องของผลการตรวจให้คะแนนแบบสอบอัตนัยวัดสมรรถนะทางวิทยาศาสตร์ โดยใช้สถิติสหสัมพันธ์ภายในชั้น (Intra-Class Correlation: ICC) โดยใช้โปรแกรมสำเร็จรูป SPSS

2) การวิเคราะห์ค่าความตรงตามเกณฑ์การให้ตรวจให้คะแนนโดยพิจารณาจากความเห็นพ้อง (rater agreement) ของผลการตรวจให้คะแนนของผู้ตรวจ (x) กับเกณฑ์มาตรฐาน (y) โดยใช้โปรแกรมสำเร็จรูป SPSS

3) การวิเคราะห์ค่าสัมประสิทธิ์การสรุปอ้างอิงเชิงตัดสินใจสัมพัทธ์ โดยใช้โปรแกรม EduG ในการตรวจให้คะแนนของผู้ตรวจที่มีคุณวุฒิตรงเอก และผู้ตรวจที่มีคุณวุฒิไม่ตรงเอก ที่มีรูปแบบการตรวจ 2 รูปแบบ ได้แก่ รูปแบบการตรวจข้อสอบบางข้อของผู้สอบทุกคน $[p \times (i : r)]$ และ รูปแบบการตรวจทุกข้อของผู้สอบทุกคน $[p \times i \times r]$

ผลการวิจัย

1. ผลการวิเคราะห์ค่าความสอดคล้องของการตรวจให้คะแนนแบบสอบอัตนัยวัดสมรรถนะทางวิทยาศาสตร์ โดยใช้สถิติสหสัมพันธ์ภายในชั้น (Intra-Class Correlation: ICC)

ผู้วิจัยได้ทำการวิเคราะห์ค่าความสอดคล้องของการตรวจให้คะแนน เพื่อวิเคราะห์หาความเที่ยงของการตรวจให้คะแนนระหว่างผู้ตรวจ (inter rater reliability) ตั้งแต่ 2 คน ขึ้นไป โดยใช้สถิติสหสัมพันธ์ภายในชั้น (Intra-Class Correlation: ICC) โมเดลอิทธิพลผสมแบบสองทาง (two-way mixed effects model) ที่ช่วงความเชื่อมั่น 95% (confidence interval) โดยใช้ผลการตรวจให้คะแนนจากกลุ่มผู้ตรวจที่มีคุณสมบัติตรงเอก และกลุ่มผู้ตรวจที่มีคุณสมบัติไม่ตรงเอก ที่ตรวจด้วยรูปแบบการตรวจทุกข้อของผู้สอบทุกคน มาวิเคราะห์ด้วยโปรแกรมสำเร็จรูป SPSS เพื่อพิจารณาโดยใช้สถิติสหสัมพันธ์ภายในชั้น (Intra-Class Correlation: ICC) และใช้เกณฑ์การแปลความหมายของ Koo & Li 2016 ซึ่งผลการวิเคราะห์ปรากฏดังตาราง 1 และตาราง 2

ตาราง 1 ค่าสัมประสิทธิ์สหสัมพันธ์ภายในชั้น (Intra-Class Correlation: ICC) ของการตรวจให้คะแนนแบบสอบโดยกลุ่มผู้ตรวจที่มีคุณสมบัติตรงเอก

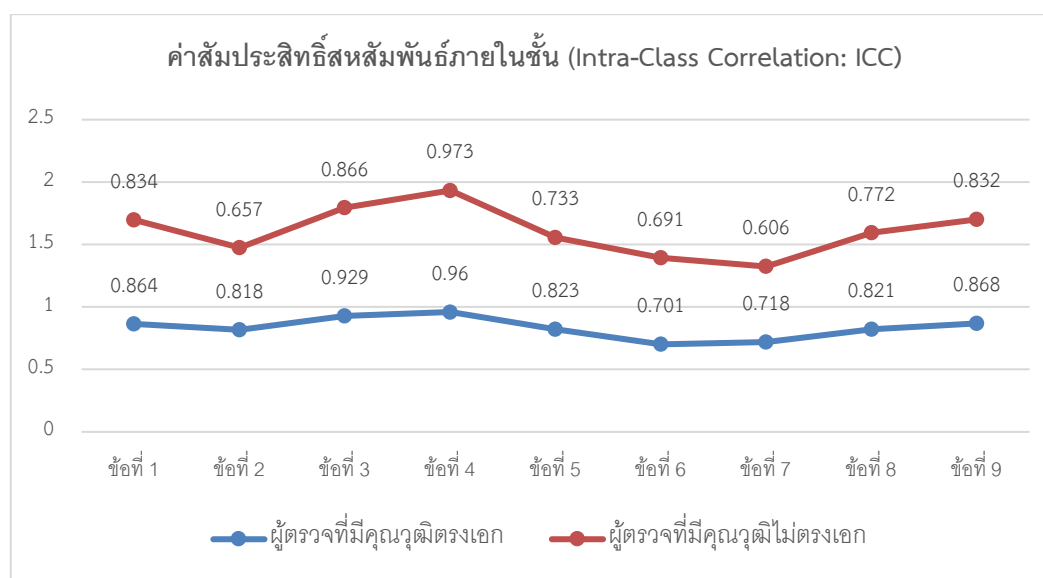
สถานการณ์	ข้อ	รายบุคคล		โดยเฉลี่ย	
		ICC (95% CI)	ระดับเกณฑ์	ICC (95% CI)	ระดับเกณฑ์
1	1	0.613 (0.523-0.699)	พอใช้	0.864 (0.814-0.903)	ดี
	2	0.529 (0.431-0.626)	พอใช้	0.818 (0.752-0.870)	ดี
	3	0.767 (0.701-0.824)	ดี	0.929 (0.904-0.949)	ดีมาก
2	4	0.857 (0.812-0.894)	ดี	0.960 (0.945-0.971)	ดีมาก
	5	0.538 (0.441-0.634)	พอใช้	0.823 (0.759-0.874)	ดี
	6	0.369 (0.266-0.479)	ต่ำ	0.701 (0.592-0.786)	พอใช้
3	7	0.389 (0.285-0.497)	ต่ำ	0.718 (0.615-0.798)	พอใช้
	8	0.534 (0.436-0.630)	พอใช้	0.821 (0.756-0.872)	ดี
	9	0.621 (0.532-0.706)	พอใช้	0.868 (0.820-0.906)	ดี

จากตาราง 1 พบว่า ค่าความสอดคล้องในการตรวจให้คะแนนระหว่างผู้ตรวจ กลุ่มผู้ตรวจที่มีคุณสมบัติตรงเอก มีความสอดคล้องกันรายบุคคล อยู่ในระดับต่ำ พอใช้ และดี โดยพบว่า สถานการณ์ที่ 2 ข้อย่อยที่ 1 มีค่า ICC_(2,1) มากที่สุด คือ มีค่าตั้งแต่ 0.812 ถึง 0.894 และ สถานการณ์ที่ 2 ข้อย่อยที่ 3 มีค่า ICC_(2,3) ต่ำที่สุด คือ มีค่าตั้งแต่ 0.266 ถึง 0.479 ส่วนค่าความสอดคล้องในการตรวจให้คะแนนระหว่างผู้ตรวจ กลุ่มผู้ตรวจที่มีคุณสมบัติตรงเอก มีความสอดคล้องกันโดยเฉลี่ย อยู่ในระดับพอใช้ ดี และดีมาก โดยพบว่า สถานการณ์ที่ 2 ข้อย่อยที่ 1 มีค่า ICC_(2,1) มากที่สุด คือ มีค่าตั้งแต่ 0.945 ถึง 0.971 และ สถานการณ์ที่ 2 ข้อย่อยที่ 3 มีค่า ICC_(2,3) ต่ำที่สุด คือ มีค่าตั้งแต่ 0.592 ถึง 0.786

ตาราง 2 ค่าสัมประสิทธิ์สหสัมพันธ์ภายในชั้น (Intra-Class Correlation: ICC) ของการตรวจให้คะแนนแบบสอบโดยกลุ่มผู้ตรวจที่มีคุณวุฒิไม่ตรงเอก

สถานการณ์	ข้อ	รายบุคคล		โดยเฉลี่ย	
		ICC (95% CI)	ระดับเกณฑ์	ICC (95% CI)	ระดับเกณฑ์
1	1	0.557 (0.461-0.651)	พอใช้	0.834 (0.774-0.882)	ดี
	2	0.324 (0.222-0.436)	ต่ำ	0.657 (0.533-0.755)	พอใช้
	3	0.618 (0.529-0.703)	พอใช้	0.866 (0.818-0.905)	ดี
2	4	0.900 (0.867-0.927)	ดีมาก	0.973 (0.963-0.981)	ดีมาก
	5	0.407 (0.304-0.514)	ต่ำ	0.733 (0.636-0.809)	พอใช้
	6	0.359 (0.256-0.469)	ต่ำ	0.691 (0.579-0.779)	พอใช้
3	7	0.278 (0.177-0.390)	ต่ำ	0.606 (0.463-0.719)	พอใช้
	8	0.458 (0.356-0.562)	ต่ำ	0.772 (0.689-0.837)	ดี
	9	0.554 (0.457-0.647)	พอใช้	0.832 (0.771-0.880)	ดี

จากตาราง 2 พบว่า ค่าความสอดคล้องในการตรวจให้คะแนนระหว่างผู้ตรวจ กลุ่มผู้ตรวจที่มีคุณวุฒิไม่ตรงเอก มีความสอดคล้องกันรายบุคคล อยู่ในระดับต่ำ พอใช้ และดีมาก โดยพบว่า สถานการณ์ที่ 2 ข้อย่อยที่ 1 มีค่า ICC_(2,1) มากที่สุด คือ มีค่าตั้งแต่ 0.867 ถึง 0.927 และ สถานการณ์ที่ 3 ข้อย่อยที่ 1 มีค่า ICC_(3,1) ต่ำที่สุด คือ มีค่าตั้งแต่ 0.177 ถึง 0.390 ส่วนค่าความสอดคล้องในการตรวจให้คะแนนระหว่างผู้ตรวจกลุ่มผู้ตรวจที่มีคุณวุฒิไม่ตรงเอก มีความสอดคล้องกันโดยเฉลี่ย อยู่ในระดับพอใช้ ดี และดีมาก โดยพบว่า สถานการณ์ที่ 2 ข้อย่อยที่ 1 มีค่า ICC_(2,1) มากที่สุด คือ มีค่าตั้งแต่ 0.963 ถึง 0.981 และ สถานการณ์ที่ 3 ข้อย่อยที่ 1 มีค่า ICC_(3,1) ต่ำที่สุด คือ มีค่าตั้งแต่ 0.463 ถึง 0.719



ภาพ 1 เปรียบเทียบค่าความสอดคล้องของผลการตรวจให้คะแนน โดยใช้สถิติสหสัมพันธ์ภายในชั้น (Intra-Class Correlation: ICC) ของกลุ่มผู้ตรวจที่มีคุณลักษณะที่แตกต่างกัน

จากภาพ 1 เมื่อพิจารณาค่าความสอดคล้องของผลการตรวจให้คะแนน โดยใช้สถิติสหสัมพันธ์ภายในชั้น (Intra-Class Correlation: ICC) ของผู้ตรวจที่มีคุณวุฒิตรงเอก พบว่ามีค่าระหว่าง 0.701 – 0.960 และค่า ICC ของผู้ตรวจที่มีคุณวุฒิไม่ตรงเอก พบว่ามีค่าอยู่ระหว่าง 0.606 – 0.973

2. ผลการวิเคราะห์ค่าความตรงตามเกณฑ์การให้ตรวจให้คะแนนโดยพิจารณาจากความเห็นพ้อง (rater agreement) ของผลการตรวจให้คะแนนของผู้ตรวจ (x) กับเกณฑ์มาตรฐาน (y)

ผู้วิจัยได้ใช้โปรแกรม SPSS วิเคราะห์หาค่าความตรงตามเกณฑ์การให้ตรวจให้คะแนนโดยพิจารณาจากความเห็นพ้อง (rater agreement) ของผลการตรวจให้คะแนนของผู้ตรวจที่มีคุณลักษณะที่แตกต่างกัน (x) กับเกณฑ์มาตรฐาน (y) โดยคะแนนเกณฑ์มาตรฐาน (y) ได้มาจากคะแนนฉันทามติ จากผู้เชี่ยวชาญที่มีคุณสมบัติ คือ ครูที่มีวิทยฐานะชำนาญการพิเศษ และจบการศึกษาในระดับปริญญาโท สาขาทางวิทยาศาสตร์ และสาขาการวัดและประเมินผล จำนวน 5 คน และตรวจให้คะแนนโดยใช้เกณฑ์การให้คะแนนแบบองค์รวม (holistic scoring rubric) จากนั้นนำคะแนนที่ได้จากผู้เชี่ยวชาญมาคำนวณหาค่าความเห็นพ้อง (rater agreement) กับคะแนนจากกลุ่มผู้ตรวจให้คะแนน จากสูตร ดังนี้ (Gronlund, 2003 cited in Tuksino, 2013)

$$\text{ร้อยละความเห็นพ้อง} = \frac{\text{จำนวนผู้สอบที่ได้คะแนนเท่ากันจากผู้ตรวจ 2 คน}}{\text{จำนวนผู้สอบทั้งหมด}} \times 100 =$$

ซึ่งผลการวิเคราะห์ปรากฏดังตาราง 3

ตาราง 3 ค่าความตรงตามเกณฑ์การให้ตรวจให้คะแนนโดยพิจารณาจากความเห็นพ้อง (rater agreement) ของผลการตรวจให้คะแนนของผู้ตรวจที่มีคุณลักษณะที่แตกต่างกัน (x) กับเกณฑ์มาตรฐาน (y) (ร้อยละ)

สถานการณ์	ข้อ	ผู้ตรวจที่มีคุณวุฒิตรงเอก			ผู้ตรวจที่มีคุณวุฒิไม่ตรงเอก		
		คนที่ 1	คนที่ 2	คนที่ 3	คนที่ 1	คนที่ 2	คนที่ 3
1	1	50	73	60	35	52	27
	2	44	63	49	38	42	38
	3	73	81	53	64	75	59
2	4	81	84	70	84	88	89
	5	33	65	55	63	60	71
	6	18	71	39	63	52	74
3	7	66	63	50	75	75	69
	8	29	62	40	70	42	63
	9	46	55	40	38	41	46

จากตาราง 3 ค่าความเห็นพ้อง (rater agreement) ของผู้ตรวจ (x) กับเกณฑ์มาตรฐาน (y) พบว่า คุณลักษณะของผู้ตรวจที่แตกต่างกัน ส่งผลให้ค่าความเห็นพ้องในแต่ละข้อคำถามแตกต่างกัน โดยผู้ตรวจที่มี คุณวุฒิตรงเอก คนที่ 2 มีค่าความเห็นพ้องกับเกณฑ์มาตรฐาน (y) สูงกว่าผู้ตรวจคนอื่น ๆ ในข้อคำถามย่อยที่ 1, 2, 3, 5, และ 9 รองลงมาคือผู้ตรวจที่มีคุณวุฒิที่ไม่ตรงเอก คนที่ 3 มีค่าความเห็นพ้องกับเกณฑ์มาตรฐาน (y) สูงกว่าผู้ตรวจคนอื่น ๆ ในข้อคำถามย่อยที่ 4 และ 6 ส่วนของผู้ตรวจที่มีค่าความเห็นพ้องกับเกณฑ์ มาตรฐาน (y) ต่ำ ได้แก่ ผู้ตรวจที่มีคุณวุฒิตรงเอก คนที่ 1 โดยมีค่าความเห็นพ้องต่ำในข้อคำถามย่อยที่ 5, 6 และ 8 ผู้ตรวจที่มีคุณวุฒิตรงเอก คนที่ 3 โดยมีค่าความเห็นพ้องต่ำในข้อคำถามย่อยที่ 3, 4 และ 7 ผู้ตรวจที่มี คุณวุฒิไม่ตรงเอก คนที่ 1 โดยมีค่าความเห็นพ้องต่ำในข้อคำถามย่อยที่ 2 และ 9 และผู้ตรวจที่มีคุณวุฒิไม่ตรงเอก คนที่ 3 โดยมีค่าความเห็นพ้องต่ำในข้อคำถามย่อยที่ 1 และ 2

3. ผลการวิเคราะห์ค่าสัมประสิทธิ์การสรุปอ้างอิงโดยใช้ทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือ ของผลการวัด (G-theory) โดยใช้โปรแกรม EduG

ผู้วิจัยได้นำผลการตรวจให้คะแนนแบบสอบอัตนัยวัดสมรรถนะทางวิทยาศาสตร์ของนักเรียน ชั้นมัธยมศึกษาปีที่ 3 โดยใช้เกณฑ์การให้คะแนนแบบองค์รวม (holistic scoring rubric) โดยตรวจให้คะแนน จากกลุ่มผู้ตรวจที่มีคุณวุฒิตรงเอก และกลุ่มผู้ตรวจที่มีคุณวุฒิไม่ตรงเอก มาเปรียบเทียบค่าความเที่ยง (สัมประสิทธิ์การสรุปอ้างอิง) ของการตรวจให้คะแนนแบบสอบ ผ่านรูปแบบการตรวจ 2 รูปแบบ ได้แก่ รูปแบบการตรวจข้อสอบบางข้อของผู้สอบทุกคน [$p \times (i:r)$] และ รูปแบบการตรวจทุกข้อของผู้สอบทุกคน [$p \times i \times r$] มีผลการวิเคราะห์ปรากฏดังตาราง 4

ตาราง 4 ผลการวิเคราะห์ค่าสัมประสิทธิ์การสรุปอ้างอิงโดยใช้ทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของ ผลการวัด (G-theory)

รูปแบบการตรวจ (design)	คุณลักษณะของผู้ตรวจ	
	ผู้ตรวจที่มีคุณวุฒิตรงเอก	ผู้ตรวจที่มีคุณวุฒิไม่ตรงเอก
$p \times i \times r$ Design	ρ_{δ}^2	0.86
	ρ_{Δ}^2	0.80
$p \times (i:r)$ Design	ρ_{δ}^2	0.87
	ρ_{Δ}^2	0.40

จากตาราง 4 การเปรียบเทียบค่าความเที่ยง (สัมประสิทธิ์การสรุปอ้างอิง) ของการตรวจให้คะแนน แบบสอบอัตนัยวัดสมรรถนะทางวิทยาศาสตร์ของนักเรียนชั้นมัธยมศึกษาปีที่ 3 โดยใช้เกณฑ์การให้คะแนน แบบองค์รวม (holistic scoring rubric) พบว่า รูปแบบการตรวจให้คะแนนแบบผู้ตรวจตรวจข้อสอบทุกข้อ ของผู้สอบทุกคน [$p \times i \times r$] เมื่อกำหนดคุณลักษณะของผู้ตรวจที่แตกต่างกัน กลุ่มผู้ตรวจที่มีคุณวุฒิตรงเอก จะมีค่าความเที่ยงของการตรวจให้คะแนนสูงกว่ากลุ่มผู้ตรวจที่มีคุณวุฒิไม่ตรงเอก โดยมีค่าสัมประสิทธิ์การสรุป อ้างอิงสำหรับตัดสินใจสัมพัทธ์ (ρ_{δ}^2) เท่ากับ 0.86 และมีค่าสัมประสิทธิ์การสรุปอ้างอิงสำหรับตัดสินใจสมบูรณ์

(ρ_{Δ}^2) เท่ากับ 0.80 และหากตรวจให้คะแนนโดยรูปแบบการตรวจข้อสอบบางข้อของผู้สอบทุกคน [$(p \times i : r)$] กลุ่มผู้ตรวจที่มีคุณวุฒิไม่ตรงเอจะมีค่าความเที่ยงของการตรวจให้คะแนนใกล้เคียงกับกลุ่มผู้ตรวจที่มีคุณวุฒิตรงเอ โดยมิต่ำสัมประสิทธิ์การสรุปอ้างอิงสำหรับตัดสินใจสัมพัทธ์ (ρ_{δ}^2) เท่ากับ 0.87 และมีค่าสัมประสิทธิ์การสรุปอ้างอิงสำหรับตัดสินใจสมบูรณ์ (ρ_{Δ}^2) เท่ากับ 0.40

อภิปรายผล

1. ผลการวิเคราะห์ค่าความสอดคล้องของผลการตรวจให้คะแนนแบบสอบอัตนัยวัดสมรรถนะทางวิทยาศาสตร์ โดยใช้เกณฑ์การให้คะแนนแบบองค์รวม (holistic scoring rubric) โดยใช้สถิติสหสัมพันธ์ภายในชั้น (Intra-Class Correlation: ICC) พบว่า ความสอดคล้องของผลการตรวจให้คะแนนแบบสอบอัตนัยทั้ง 3 สถานการณ์ 9 ข้อ นั้น อยู่ในระดับสอดคล้องต่ำไปจนถึงระดับดีมาก แสดงให้เห็นว่า ผู้ตรวจให้คะแนนยังมีความเห็นไปในทิศทางที่ต่างกันบางสถานการณ์และข้อคำถาม โดยกลุ่มผู้ตรวจที่มีคุณวุฒิไม่ตรงเอ มีค่าความสอดคล้องของผลการตรวจให้คะแนนโดยเฉลี่ย สูงกว่ากลุ่มผู้ตรวจที่มีคุณวุฒิตรงเอ ซึ่งคุณลักษณะที่แตกต่างกันในกลุ่มของผู้ตรวจให้คะแนน อาจเป็นสาเหตุที่ทำให้เกิดความคลาดเคลื่อนของการตรวจให้คะแนนสอดคล้องกับ Aphaikawi (2019) ที่กล่าวว่า สาเหตุของการให้คะแนนที่แตกต่างกัน อาจจะมาจากรูปลักษณ์ของผู้ตรวจ เช่น ประสบการณ์ในการทำงาน การศึกษาตรงตามสาขาวิชา เป็นต้น ดังนั้นเพื่อลดความคลาดเคลื่อนที่อาจจะเกิดขึ้น จึงควรควบคุมแหล่งความคลาดเคลื่อนต่าง ๆ เช่น เพิ่มการฝึกอบรม เพื่อเพิ่มทักษะให้กับผู้ตรวจ คัดเลือกผู้ตรวจที่มีลักษณะความเข้มงวด ใจดี ใกล้เคียงกัน เพื่อให้ผลการตรวจมีค่าความสอดคล้องที่สูงขึ้น (Hopkins & Antes, 1990)

2. ผลการวิเคราะห์และเปรียบเทียบค่าความตรงของเกณฑ์การให้ตรวจให้คะแนน โดยพิจารณาจากความเห็นพ้อง (rater agreement) ของผลการตรวจให้คะแนนของผู้ตรวจ (x) กับเกณฑ์มาตรฐาน (y) ของกลุ่มผู้ตรวจที่มีคุณลักษณะที่แตกต่างกัน พบว่า ผู้ตรวจที่มีคุณวุฒิตรงเอ มีค่าความเห็นพ้องกับเกณฑ์มาตรฐาน (y) สูงกว่าผู้ตรวจคนอื่น ๆ ซึ่งแสดงให้เห็นว่า ผู้ตรวจที่มีคุณวุฒิตรงเอ มีความเชี่ยวชาญและมีประสบการณ์เดิมในด้านของเนื้อหาและการประเมินสมรรถนะในรายวิชาวิทยาศาสตร์ มากกว่าผู้ตรวจที่มีคุณวุฒิไม่ตรงเอ จึงทำให้ผลของการตรวจให้คะแนนมีค่าความเห็นพ้องใกล้เคียงกับเกณฑ์มาตรฐาน (y) ดังนั้นหากต้องการให้ค่าความเห็นพ้องของกลุ่มผู้ตรวจที่มีคุณวุฒิไม่ตรงเอมีค่าสูงมากขึ้น จึงควรมีเกณฑ์การให้คะแนนที่เหมาะสมกับผู้ตรวจที่มีคุณวุฒิไม่ตรงเอ หรือต้องมีการจัดอบรมการใช้เกณฑ์การให้คะแนนเพื่อพัฒนาทักษะให้กับผู้ตรวจ (Aphaikawi, 2019)

3. ผลการวิเคราะห์ค่าสัมประสิทธิ์การสรุปอ้างอิงโดยใช้ทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของการวัด (G-theory) พบว่า กลุ่มผู้ตรวจที่มีคุณวุฒิไม่ตรงเอ เมื่อใช้รูปแบบการตรวจให้คะแนนแบบผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน [$p \times i \times r$] มีค่าสัมประสิทธิ์การสรุปอ้างอิงสำหรับตัดสินใจสัมพัทธ์ (ρ_{δ}^2) และค่าสัมประสิทธิ์การสรุปอ้างอิงสำหรับตัดสินใจสมบูรณ์ (ρ_{Δ}^2) เท่ากับ 0.00 ซึ่งถือว่าไม่มีค่าความเที่ยงเลย ทั้งนี้ เนื่องจากกลุ่มผู้ตรวจที่มีคุณวุฒิไม่ตรงเอ มาจากครูผู้สอนรายวิชาภาษาไทย คณิตศาสตร์ และสังคมศึกษา ซึ่งมีธรรมชาติของรายวิชาที่แตกต่างกัน เมื่อตรวจให้คะแนนแบบสอบอัตนัยในรายวิชาวิทยาศาสตร์

ซึ่งผู้ตรวจไม่มีประสบการณ์ในด้านของการสอนรายวิชาวิทยาศาสตร์เลย จึงทำให้ความรู้พื้นฐานด้านเนื้อหาและความเห็นในเรื่องของการลงคะแนน มีความแตกต่างกัน หรือคุณลักษณะที่แฝงอยู่กับผู้ตรวจ เช่น การเข้มงวด/ใจดี ของผู้ตรวจแตกต่างกัน จึงส่งผลต่อการให้คะแนนมีความแตกต่างกัน ซึ่งสอดคล้องกับงานวิจัยของ Taoto (2016) ที่พบว่า ความแปรปรวนด้านผู้ตรวจให้คะแนนแบบสอบจะมีค่าน้อยสุด เนื่องมาจากมีการคัดเลือกผู้ตรวจที่มีคุณสมบัติเหมาะสม มีความรู้ความสามารถให้ด้านการสอนและตัวชี้วัดทางวิทยาศาสตร์ นอกจากนี้ยังมีการอบรมและฝึกปฏิบัติจริง ตลอดจนมีคู่มือการดำเนินเพื่อเพิ่มความชัดเจนในแนวทางการดำเนินงาน ดังนั้น หากต้องการให้กลุ่มผู้ตรวจที่มีคุณวุฒิไม่ตรงเอก มีผลการตรวจให้คะแนนที่มีค่าสัมประสิทธิ์การสรูปอ้างอิงที่สูงขึ้น จึงจำเป็นต้องมีการ คัดเลือกผู้ตรวจที่มีลักษณะความเข้มงวด ใจดี ใกล้เคียงกัน เพื่อให้ผลการตรวจมีค่าความเที่ยงของการตรวจให้คะแนนที่สูงขึ้น (Hopkins & Antes, 1990) และต้องมีการฝึกอบรม และฝึกปฏิบัติจริง เพื่อพัฒนาทักษะให้กับผู้ตรวจ (Aphaikawi, 2019) รวมทั้ง ต้องมีคู่มือเกณฑ์การให้คะแนนที่มีการกำหนดแนวทาง และวิธีการให้คะแนนอย่างชัดเจน (Welk et al., 2004)

4. ผลการวิเคราะห์ค่าสัมประสิทธิ์การสรูปอ้างอิงโดยใช้ทฤษฎีการสรูปอ้างอิงความน่าเชื่อถือของผลการวัด (G-theory) พบว่า รูปแบบการตรวจข้อสอบบางข้อของผู้สอบทุกคน $[(p \times (i : r))]$ มีค่าสัมประสิทธิ์การสรูปอ้างอิงสูงกว่า รูปแบบการตรวจให้คะแนนแบบผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน $[p \times i \times r]$ ทั้งในกลุ่มผู้ตรวจที่มีคุณวุฒิตรงเอก และกลุ่มผู้ตรวจที่มีคุณวุฒิไม่ตรงเอก เนื่องจากการตรวจข้อสอบทุกข้อของผู้สอบทุกคน อาจจะทำให้ผู้ตรวจเกิดอาการเมื่อยล้า หรืออาจจะมีข้อคำถามบางข้อ หรือบางสถานการณ์ที่ไม่ถนัด แต่หากเป็นการตรวจในรูปแบบการตรวจข้อสอบบางข้อของผู้สอบทุกคน ผู้ตรวจจะได้รับผิดชอบเฉพาะบางข้อเท่านั้นซึ่งจะทำให้ผู้ตรวจมีความเชี่ยวชาญในเนื้อหา และเกณฑ์การให้คะแนนมากยิ่งขึ้น สอดคล้องกับผลการวิจัยของ Kwanja, (2013); Sanguanwai, (2015); Saosin, (2018) ที่พบว่า แบบสอบที่มีการออกแบบการตรวจให้คะแนนแบบที่ผู้ตรวจ ตรวจให้คะแนนเฉพาะเหตุการณ์ของผู้สอบทุกคน มีค่าความเที่ยงสูงกว่าการที่ผู้ตรวจให้คะแนนเหตุการณ์ทุกเหตุการณ์ของผู้สอบทุกคน

ข้อเสนอแนะ

1. ข้อเสนอแนะในการนำผลวิจัยไปใช้

1. แบบสอบอัตนัยวัดสมรรถนะทางวิทยาศาสตร์ของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ที่ผู้วิจัยพัฒนาขึ้น สามารถนำเอาไปใช้สำหรับวัดสมรรถนะทางวิทยาศาสตร์ของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ได้ โดยในการตรวจให้คะแนนแบบสอบอัตนัยโดยใช้เกณฑ์การตรวจให้คะแนนแบบองค์รวม (holistic scoring rubric) ที่ผู้วิจัยพัฒนาขึ้นนั้น สำหรับกลุ่มผู้ตรวจที่มีคุณวุฒิตรงเอก สามารถตรวจให้คะแนนได้ทั้งรูปแบบการตรวจข้อสอบบางข้อของผู้สอบทุกคน $[(p \times (r : i))]$ และ รูปแบบการตรวจให้คะแนนแบบผู้ตรวจตรวจข้อสอบทุกข้อของผู้สอบทุกคน $[p \times i \times r]$ เนื่องจากให้ค่าความเที่ยงของการให้คะแนนอยู่ในระดับที่เหมาะสม คือ มากกว่า 0.7 ทั้งสองรูปแบบการตรวจ แต่เพื่อให้ประหยัดเวลาในการตรวจให้คะแนน และลดความเหนื่อยล้าของผู้ตรวจ ควรตรวจให้คะแนนรูปแบบการตรวจข้อสอบบางข้อของผู้สอบทุกคน $[(p \times (r : i))]$ แต่ในกรณีที่ผู้ตรวจให้คะแนนเป็นกลุ่มผู้ตรวจที่มีคุณวุฒิไม่ตรงเอก ให้ใช้รูปแบบการตรวจให้คะแนนที่เป็น

รูปแบบการตรวจข้อสอบบางข้อของผู้สอบทุกคน $[(p \times (i : r))]$ เท่านั้น เนื่องจากจะให้ค่าความเที่ยงของการตรวจให้คะแนนอยู่ในระดับสูง

2. ข้อเสนอแนะในการวิจัยครั้งต่อไป

1. จากการศึกษางานวิจัยในครั้งนี้อย่างถี่ถ้วน พบว่ากลุ่มผู้ตรวจที่มีคุณวุฒิตรงเอก มีค่าความสอดคล้องของการตรวจให้คะแนนสูง ซึ่งกลุ่มผู้ตรวจที่มีคุณวุฒิตรงเอกเป็นคุณครูที่จบจากสาขา/วิชาเอกวิทยาศาสตร์เหมือนกัน แต่กลุ่มผู้ตรวจที่มีคุณวุฒิไม่ตรงเอก มีคุณครูที่จบจากสาขา/วิชาเอกคณิตศาสตร์ ภาษาไทย และสังคมศึกษา ซึ่งมีความแตกต่างกัน ดังนั้นในการศึกษาครั้งต่อไป ควรศึกษากลุ่มผู้ตรวจที่มีคุณวุฒิไม่ตรงเอกที่เป็นคุณครูที่จบจากสาขา/วิชาเอกเดียวกัน แล้วนำผลการตรวจให้คะแนนมาเปรียบเทียบกัน ว่าส่งผลต่อคุณภาพของการตรวจให้คะแนนที่แตกต่างหรือไม่

2. ในการศึกษาความสอดคล้องของผลการตรวจให้คะแนนในงานวิจัยนี้จะเห็นว่าค่าความสอดคล้องของผลการตรวจให้คะแนนในบางข้อยังอยู่ในระดับต่ำ ดังนั้นในการวิจัยครั้งต่อไป จึงควรมีการคัดเลือกคุณลักษณะของกลุ่มผู้ตรวจ เช่น ความใจดี ความเข้มงวด หรือ ควรมีการอบรมการใช้เกณฑ์การตรวจให้คะแนน และนำผลการตรวจให้คะแนนก่อนและหลังการอบรม มาเปรียบเทียบกัน เพื่อศึกษาหาความแปรปรวนของการตรวจให้คะแนนในครั้งต่อไป และควบคุมแหล่งความแปรปรวนที่อาจจะส่งผลต่อการตรวจให้คะแนนเหล่านั้น

References

- Brennan, R. L., & Johnson, E. G. (1995). Generalizability of Performance Assessments. *Journal of Educational Measurement*, 14(4), 9-12.
- Chiu, C., & Wolfe, E. (2002). A Method for Analyzing Sparse Data Matrices in the Generalizability Theory Framework. *SAGE Journal*, 26(3), 321-338.
- Coffman, W. E. (1971). On the Reliability of Ratings of Essay Examinations in English. *JSTOR Journal*, 5(1), 24-36.
- Hopkins, C. D., & Antes, R. L. (1990). *Classroom Measurement and Evaluation*. Peacock Press.
- 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.
- Swartz, C. W., Hooper, S. R., Montgomery, J. W., Wakely, M. B., Kruif, R. E. L., Reed, M., Brown, T. T., Levine, M. D., & White, K. P. (1999). Using Generalizability Theory to Estimate the Reliability of Writing Scores Derived from Holistic and Analytical Scoring Methods. *Sage Journal*, 59(3), 492-506.
- Welk et al., (2004). Reliability of accelerometry-based activity monitors: a generalizability study. *Ovid Journal*, 36(9), 1637-1645.

Translate Thai References

- Aphaikawi, D. (2019). Scoring results of subjective exams when different groups of inspectors and examination patterns. *The 27th Thailand Measurement Evaluation and Research*, 108-124. (in Thai)
- Intanate, N. (2011). *Characteristic 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. (2007). *Modern test theories*. Chulalongkorn University Press. (in Thai)
- Kwanja, N. (2013). *Comparison of summaries reference coefficients of the process skills scale Grade 4 science with different scoring patterns* [Master's thesis]. Mahasarakham University. (in Thai)
- Ministry of Education. (2017). *Thailand Education Plan B.E. 2560 - 2579 (A.D. 2017 – 2036)*. Office of the Education Council Press. (in Thai)
- Phadungphon, S. (2017). Comparison of reliability of modified essay question test for measuring the abilities in using scientific method in physic under different numbers of event and rater: an application of generalizability theory. *Educational Electronic Journal*, 12(4), 381-393. (in Thai)
- Phusing N. (2020). Scienceteacher development model throughstem education for the schools with non-science majoringteachers (nsmt). *Journal of MCU Ubon Review*, 5(3), 439-454. (in Thai)
- Pinyoanuntapong, B. (2004). *Measurement and evaluation*. Srinakharinwirot University Press. (in Thai)
- Sanguanwai, C. (2015). *Comparison of test reliability for Measuring Mathematical Creative problem-solving ability: Application of Generalizability theory* [Master's thesis]. Chulalongkorn University. (in Thai)
- Saosin, K. (2019). Comparison of Reliability of Math Problem Solving Proficiency Test with Sub-analytical Scoring At the lower secondary level: application of summary theory referring to the reliability of measurement results. *Educational Electronic Journal*, 13(3), 423-438. (in Thai)
- Taoto, J. (2016). A study of the confidence values of students' math subjective test scores. Secondary school with different number of examiners and scoring patterns using the theory of summaries. Reference. *Hat Yai Academic Journal*, 14(1), 1-14. (in Thai)

- The Institute for the Promotion of Teaching Science and Technology, (2020). *Scientific Literacy*. <https://pisathailand.ipst.ac.th/about-pisa/scientific-literacy/> (in Thai)
- Tuksino, P. (2013). *Teaching documentation educational research methodology*. Khon Kaen University Press. (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 number of events: an application of generalizability theory* [Master's thesis]. Chulalongkorn University. (in Thai)