การเปรียบเทียบผลการตรวจให้คะแนนของแบบสอบอัตนัยเมื่อรูปแบบการตรวจแตกต่างกัน : การประยุกต์ใช้ทฤษฎีการสรุปอ้างอิง

Comparisons of the Results of Essay Test by Different Scoring Designs: Application of Generalizability Theory

วีรภัทร ธิติกาญจน์พจนา (Weerapat Thitikanpodchana)* ดร.ประกฤติยา ทักษิโณ (Dr.Prakittiya Tuksino)**

บทคัดย่อ

การวิจัยนี้มีวัตถุประสงค์เพื่อ 1) เพื่อศึกษาความสอดคล้องของผลการตรวจให้คะแนนแบบสอบความสามารถในการเขียนภาษาอังกฤษของนักเรียนชั้นมัธยมศึกษาปีที่ 3 2) เพื่อเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิงของการให้คะแนนที่มีจำนวนผู้ตรวจ และรูปแบบการตรวจที่ต่างกัน กลุ่มตัวอย่างแบ่งเป็น 2 กลุ่ม คือ กลุ่มนักเรียนนักเรียนชั้นมัธยมศึกษาปีที่ 3 จำนวน 30 คน และกลุ่มผู้ตรวจให้คะแนนซึ่งเป็นครูผู้สอนวิชาภาษาอังกฤษ จำนวน 3 คน เครื่องมือที่ใช้คือ แบบสอบความสามารถในการเขียนภาษาอังกฤษ จำนวน 3 ข้อ วิเคราะห์ค่าสัมประสิทธิ์สรุปอ้างอิงโดยใช้โปรแกรม EduG ผลการวิจัยพบว่า 1) การศึกษาความสอดคล้องของผลการตรวจให้คะแนนระหว่างผู้ตรวจ พิจารณาโดยสถิติสหสัมพันธ์ภายในชั้น (Intra-Class Correlation: ICC) ที่มีความสอดคล้องกันโดยเฉลี่ย พบว่าความสอดคล้องของผลการตรวจให้คะแนนทั้ง 3 ข้ออยู่ในระดับพอใช้และดี 2) ค่าสัมประสิทธิ์การสรุปอ้างอิงของรูปแบบการตรวจข้อสอบเฉพาะข้อของผู้สอบทุกคน [p x (r: i)] มีค่าสูงกว่า รูปแบบการตรวจทุกข้อของผู้สอบทุกคน [p x i x r] ทั้งจำนวนผู้ตรวจ 2 คน และ 3 คน

ABSTRACT

This research aims 1) to study the agreement of the scoring results of English writing ability test of Mattayom 3 and 2) to compare the Generalizability Coefficient scores of scoring according to different raters and scoring designs. The sample populations were divided into 2 groups, including the group of 30 students in grade 9 and the group of 3 raters who were English language teaching teachers. The research instrument was English writing ability test with 3 items. Generalizability Coefficient scores were analyzed by EduG. The research findings were 1) the agreement of the scoring results for each items analyzed by Intra-Class Correlation (ICC) was found fair and good. 2) the Generalizability Coefficient scores of p x (r: i) design was higher than p x i x r design for both 2 raters and 3 raters scoring.

คำสำคัญ: ทฤษฎีการสรุปอ้างอิง รูปแบบการตรวจให้คะแนน แบบสอบอัตนัย

Keywords: Generalizability Theory, Scoring design, Essay Test

*นักศึกษาหลักสูตรศึกษาศาสตรมหาบัณฑิต สาขาวิชาการวัดและประเมินผลการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยขอนแก่น

**ผู้ช่วยศาสตราจารย์ สาขาวิชาการวัดและประเมินผลการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยขอนแก่น

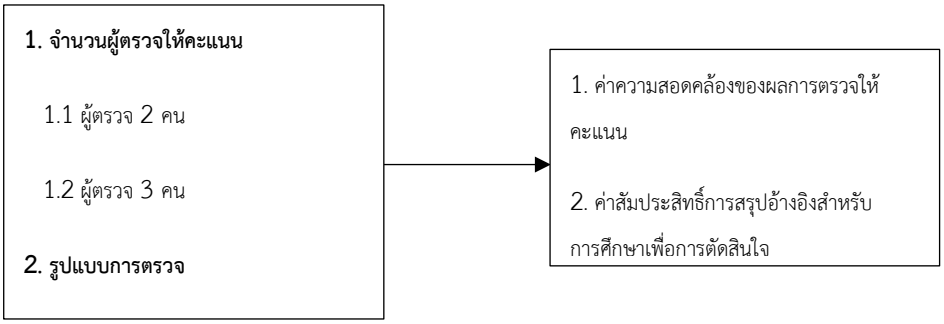
บทนำ

แบบทดสอบการเขียนที่มีประสิทธิภาพจะต้องมีความสัมพันธ์กับชีวิตจริงของผู้เรียนเพื่อกระตุ้นให้ผู้เรียนได้ดึงศักยภาพและถ่ายทอดความคิดของตนเองได้ ซึ่งมาจากเรื่องที่นักเรียนสนใจ และแบบทดสอบการเขียนที่ดีจะสามารถดึงความรู้ภูมิหลังและประสบการณ์ของนักเรียนได้ (Scott, 1996) ซึ่งตรงกับ Hughes (2003) ได้กล่าวว่า การประเมินการเขียนสามารถทำได้หลายรูปแบบ แต่อย่างไรก็ตามการประเมินการเขียนนั้นควรจะเป็นการประเมินตามสภาพจริงและงานเขียนนั้นควรจะมีความสัมพันธ์กับชีวิตจริงของผู้เรียน นอกจากนี้ผู้เรียนควรได้รับโอกาสในการประเมินหลายๆ ครั้ง โดยผู้สอนควรที่จะประเมินผู้เรียนอย่างต่อเนื่อง ไม่ใช่เพียงเมื่อมีการสอบกลางภาคและสอบปลายภาค (Coombe and Evans, 2001) แม้ว่าการประเมินตามสภาพจริงเป็นวิธีที่นำมาใช้ในการประเมินงานเขียนภาษาอังกฤษสำหรับการเรียนภาษาอังกฤษเป็นภาษาที่สอง (ESL) ความแปรปรวนจากการให้คะแนนของผู้ตรวจถือเป็นสิ่งสำคัญในการศึกษาความตรง (Kroll, 1998) ซึ่งความแปรปรวนเหล่านี้เกิดได้จากหลากหลายแหล่ง เช่น ประสบการณ์ของผู้ตรวจ (Cumming, 1990; Weigle, 1994) กระบวนการสอน (Santos, 1988) ภาษาแม่(Brown, 1995; Chalboub-Deville, 1995a, 1995b) และการตรวจเมื่อเวลาผ่านไป (White, 1984) และที่สำคัญคือการหารูปแบบการให้คะแนนที่เหมาะสม ซึ่งมี 2 วิธี ที่เป็นที่นิยมและใช้กันอย่างแพร่หลายนั่นก็คือ เกณฑ์การให้คะแนนแบบรูบริค (Rubric scoring) ได้แก่ แบบภาพรวม และ แบบแยกองค์ประกอบ โดยแต่ละเกณฑ์นั้นมีข้อดีและข้อเสียแตกต่างกันไป โดยทั่วไปเกณฑ์แบบแยกองค์ประกอบ (analytic scoring rubric) เป็นวิธีประเมินงานเขียนโดยแยกประเมินองค์ประกอบต่าง ๆ ของงานเขียน ด้วยการหักคะแนน และการให้คะแนนแยกตามองค์ประกอบต่าง ๆ (Harold S. Madsen, 1983 : 101 - 122) ทำให้ผู้เรียนได้ทราบถึงจุดบกพร่องของตนเอง ส่วนการให้คะแนนโดยใช้เกณฑ์แบบองค์รวม (holistic scoring rubric) มักใช้เวลาน้อยกว่าการใช้เกณฑ์แบบแยกองค์ประกอบ (กมลวรรณ ตังธนกานนท์, 2557) เกณฑ์แบบองค์รวมคำตอบจะไม่ถูกแบ่งเป็นส่วนๆ เป็นประเด็นเฉพาะแต่ผู้ตรวจจะอ่านคำตอบอย่างรวดเร็ว แล้วใช้ความประทับใจและใช้มาตรฐานบางอย่างกำหนดระดับคำตอบ (Mehrens & Lehmann, 1973) และเพื่อให้การประเมินงานเขียนมีประสิทธิภาพ ผู้ตรวจจะต้องเรียนรู้แหล่งความคลาดเคลื่อนของคะแนนและฝึกฝนความเชี่ยวชาญในการตรวจ เพื่อลดความคลาดเคลื่อนและความลำเอียงที่อาจเกิดขึ้น โดยที่แหล่งความคลาดเคลื่อนที่สำคัญของคะแนน อาทิ คุณภาพ ของเครื่องมือ ทักษะของผู้ตรวจ และการบริหารการตรวจให้คะแนน (ศิริชัย กาญจนวาสี, 2558) คะแนนของผู้สอบจะได้รับผลกระทบจากคุณลักษณะการให้คะแนนของผู้ตรวจ (Characteristics of rater) ยังเป็นปัจจัยหนึ่งพอๆ กับได้รับผลกระทบจากระดับความยากง่ายของข้อสอบและความสามารถของผู้สอบ นั่นอาจจะเป็นอีกปัจจัยที่มีอิทธิพลต่อความเที่ยงของผู้ตรวจ (rater reliability) ซึ่งต้องขึ้นอยู่กับมาตรฐานของการตัดสินใจของผู้ตรวจเองที่จะมีความเห็นสอดคล้องกันสูง (Hopkins & Antes,1990)

การประยุกต์ใช้ทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัด (Generalizability Theory) ซึ่งเป็นทฤษฎีทางสถิติสำหรับวิเคราะห์ความน่าเชื่อถือของผลการวัดในสถานการณ์การวัดในลักษณะต่าง ๆ ที่เป็นเป้าหมายของการนำเครื่องมือไปใช้ และสามารถนำผลที่ได้ไปใช้เป็นสารสนเทศสำหรับการตัดสินใจอย่างมีประสิทธิภาพ รวมถึงนำมาประยุกต์ใช้ในการศึกษาค่าความเที่ยง (ค่าสัมประสิทธิ์การสรุปอ้างอิง : G-Coefficient) ซึ่งวิธีการนี้จะมีความยืดหยุ่นที่จะรวมแหล่งความคลาดเคลื่อนของการวัดรวมทั้ง ผู้ประเมิน ผลงาน จำนวนครั้ง และรูปแบบของการประเมิน ค่าสัมประสิทธิ์การสรุปอ้างอิง (Generalizability coefficient) สามารถแปรความเป็นดัชนีความเห็นพ้องต้องกันของผู้ประเมินโดยจะต้องเป็นผู้ประเมินเดียวกัน ฝึกในทิศทางเดียวกัน และตรวจให้คะแนนภายใต้เงื่อนไขเดียวกัน (Cronbach et al. (1963) การออกแบบการตรวจให้คะแนน เป็นอีกวิธีหนึ่งที่นำมาช่วยแก้ไขแหล่งความคลาดเคลื่อนของการตรวจให้คะแนน รูปแบบการตรวจให้คะแนนซึ่งการออกแบบการตรวจให้คะแนนสามารถออกแบบได้หลายรูปแบบทั้งนี้ขึ้นอยู่สถานการณ์การสอบ เช่น ในสถานการณ์ที่บุคคลถูกตรวจด้วยผู้ตรวจทุกคนจากการทำข้อสอบเดียวกันทุกข้อเรียกว่า รูปแบบการตรวจ

ทุกข้อของผู้สอบทุกคน (p × r × i design) และในสถาณการณ์ที่ผู้ตรวจแบ่งกันตรวจข้อสอบแต่ละข้อที่มาจากบุคคลที่ทำข้อสอบเดียวกันทุกคน เรียกว่า รูปแบบการตรวจบางข้อของผู้สอบทุกคน [p x (i: r) design] ซึ่งการตรวจรูปแบบนี้จะช่วยประหยัดทรัพยากรทั้งจำนวนผู้ตรวจ ค่าใช้จ่ายและเวลาในการตรวจ รูปแบบการตรวจให้คะแนนที่มีความยุติธรรมมากที่สุดคือรูปแบบการตรวจทุกข้อของผู้สอบทุกคน แต่หากในความเป็นจริงแล้วการตรวจแบบนี้ทำให้เสียเวลาค่อนข้างมากและมีค่าใช้จ่ายสูง หากสามารถวางแผนการตรวจหรือหาวิธีการตรวจจะช่วยลดระยะเวลาในการตรวจ ทำให้เกิดความน่าเชื่อถือของคะแนนและให้ค่าสถิติที่ยอมรับได้ (Linacre and Wright, 2002 อ้างถึงใน น้ำผึ้ง อินทะเนตร, 2554) จากการศึกษาของชนิสรา สงวนไว้ (2558) และดรุณี อภัยกาวี (2562) พบว่า ค่าสัมประสิทธิ์การสรุปอ้างอิงของรูปแบบการตรวจ p x (i : r) สูงกว่ารูปแบบ p x i x r และการศึกษาของ จิรายุ เถาว์โท (2559) พบว่า ค่าสัมประสิทธิ์การสรุปอ้างอิงเมื่อรูปแบบการตรวจให้คะแนนเหมือนกันแต่จำนวนผู้ตรวจต่างกัน มีค่าแตกต่างกันอย่างมีนัยสำคัญ ยกเว้น รูปแบบการตรวจ (p: r) x i และ p x (i: r) ของผู้ตรวจ 2 และ 3 คน มีค่าไม่แตกต่างกัน แต่เมื่อใช้จำนวนผู้ตรวจเท่ากันและรูปแบบการให้คะแนนต่างกันพบว่าค่าสัมประสิทธิ์การสรุปอ้างอิงมีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติ เป็นต้น นอกจากนี้การศึกษาความตรงตามสภาพของผลการวัด ถ้าจำนวนผู้ตรวจมากกว่าจะมีค่าความตรงตามสภาพสูงกว่าจำนวนผู้ตรวจน้อยในทุกรูปแบบการตรวจ (อังคณา กุลนภาดล, 2555) หรือการศึกษาคะแนนในทุกเงื่อนไขที่ต่างกันจะมีความตรงตามสภาพสูง เช่นกัน (น้ำผึ้ง อินทะเนตร, 2554)

จากที่กล่าวมาข้างต้น ผู้วิจัยพบว่าแหล่งความคลาดเคลื่อนที่เกิดจากการตรวจแบบสอบการเขียนนั้นเกิดจากหลายแหล่ง โดยผู้วิจัยสนใจที่จะศึกษาขนาดของความแปรปรวนขององค์ประกอบเพื่อที่จะสามารถใช้ควบคุมแหล่งความคลาดเคลื่อนดังกล่าว โดยมีการกำหนดปัจจัยต่าง ๆ สำหรับเงื่อนไขและสถานการณ์การวัด ได้แก่ จำนวนผู้ตรวจ และรูปแบบการตรวจ เพื่อที่จะศึกษาการเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิงเมื่อมีการควบคุมปัจจัยดังกล่าว เพื่อเป็นประโยชน์สารสนเทศสำหรับการเลือกจำนวนผู้ตรวจและรูปแบบการตรวจที่เหมาะสมในการจัดการทดสอบ ให้เกิดความเที่ยงของคะแนนที่เชื่อถือได้ และสามารถนำไปใช้ได้จริงในทางปฏิบัติ



ภาพที่ 1 กรอบแนวคิดการวิจัย

วัตถุประสงค์การวิจัย

1. เพื่อศึกษาความสอดคล้องของผลการตรวจให้คะแนนแบบสอบความสามารถในการเขียนภาษาอังกฤษของนักเรียนชั้นมัธยมศึกษาปีที่ 3

2. เพื่อเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิงของการให้คะแนนที่มีจำนวนผู้ตรวจ และรูปแบบการตรวจที่ต่างกัน

วิธีการวิจัย

ประชากรและกลุ่มตัวอย่าง

ประชากรที่ใช้ในการศึกษาครั้งนี้คือ นักเรียนชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2563 ของโรงเรียนในสำนักงานเขตพื้นที่การศึกษามัธยมศึกษา เขต 25 จำนวน 10,324 คน

กลุ่มตัวอย่างแบ่งเป็น 2 กลุ่ม 1) กลุ่มนักเรียนนักเรียนชั้นมัธยมศึกษาปีที่ 3 ภาคเรียนที่ 2 ปีการศึกษา 2563 จำนวน 1 ห้องเรียน โรงเรียนสาธิตมหาวิทยาลัยขอนแก่น ฝ่ายมัธยมศึกษา (มอดินแดง) จำนวน 30 คน โดยใช้วิธีการเลือกกลุ่มตัวอย่างแบบเจาะจง (purposive sampling) และ 2) กลุ่มผู้ตรวจให้คะแนนซึ่งเป็นครูผู้สอนวิชาภาษาอังกฤษ โดยกำหนดคุณสมบัติของผู้ตรวจคือ ผู้มีวุฒิการศึกษาระดับปริญญาตรีขึ้นไปและทำหน้าที่สอนในรายวิชาภาษาอังกฤษและมีประสบการณ์สอนอย่างน้อย 5 ปี จำนวน 3 คน จาก โรงเรียนในสังกัดสำนักงานสำนักงานคณะกรรมการการศึกษาขึ้นพื้นฐาน ได้แก่ โรงเรียนบ้านนกเขาเปล้า โรงเรียนบ้านท่าลี่ โรงเรียนโรงเรียนอนุบาลลุมพุก (วันครู 2503) โดยใช้วิธีการเลือกกลุ่มตัวอย่างแบบเจาะจง (purposive sampling) เนื่องจากในงานวิจัยครั้งนี้ต้องใช้ครูผู้สอนวิชาภาษาอังกฤษชั้นมัธยมศึกษาปีที่ 3 เป็นผู้ตรวจแบบทดสอบ ซึ่งต้องใช้เวลา ความสนใจและความเต็มใจ ในการตรวจผลงานการเขียน

เครื่องมือ/ตัวแปรที่ใช้ในการวิจัย

1. เครื่องมือที่ใช้ในการวิจัย ได้แก่ แบบสอบความสามารถในการเขียนวิชาภาษาอังกฤษ ระดับชั้นมัธยมศึกษาปีที่ 3 จำนวน 3 ข้อ และเกณฑ์การให้คะแนนแบบแยกองค์ประกอบ (analytic scoring rubric) สร้างขึ้นโดยอ้างอิงตามหลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน 2551 กลุ่มสาระการเรียนรู้ภาษาต่างประเทศ โดยมุ่งเน้นในทักษะการเขียน ส่วนเกณฑ์การให้คะแนนนั้นพัฒนามาจากมาตรฐานระดับสมรรถภาพทางภาษาต่างประเทศในยุโรป (The Common EuropeanFramework of Reference for Languages: CEFR) เกณฑ์การให้คะแนนแบบแยกองค์ประกอบของวิเกิลและเกณฑ์การให้คะแนนแบบแยกองค์ประกอบ REEP ซึ่งถูกนำมาใช้ในการตรวจให้คะแนนงานเขียนของนักเรียนที่เรียนภาษาอังกฤษเป็นภาษาที่สอง โดยเกณฑ์การให้คะแนนดังกล่าวประกอบไปด้วย 4 ประเด็นย่อย ได้แก่ 1) เนื้อความและคำศัพท์ (Content and vocabulary) 2) การเรียบเรียงข้อความ (Organization) 3) หลักไวยากรณ์ (Structures) และ 4) รูปแบบการเขียน (Mechanics) นำแบบสอบดังกล่าวไปทดลองใช้ (Try out) กับกลุ่มทดลองซึ่งเป็นนักเรียนชั้นมัธยมศึกษาปีที่ 3 โรงเรียนสาธิตมหาวิทยาลัยขอนแก่น ฝ่ายมัธยมศึกษา (มอดินแดง) จำนวน 30 คน โดยใช้รูปแบบการตรวจทุกข้อของผู้สอบทุกคน ด้วยผู้ตรวจจำนวน 3 คน เพื่อตรวจสอบคุณภาพของแบบสอบ พบว่า ค่าความตรงเชิงเนื้อหา (IOC) มีค่าตั้งแต่ 0.80 ถึง 1.00 ค่าดัชนีความยากง่าย (p) มีค่าตั้งแต่ 0.27 ถึง 0.59 และมีค่าอำนาจจำแนก (r) ตั้งแต่ 0.25 ถึง 0.46

2. ตัวแปรที่ใช้ในการวิจัยหรือเงื่อนไขของการตรวจให้คะแนนที่ส่งผลต่อค่าสัมประสิทธิ์สรุปอ้างอิง

2.1 จำนวนผู้ตรวจ ได้แก่ จำนวน 2 คน และ 3 คน

2.2 รูปแบบการตรวจ ได้แก่ ผู้ตรวจตรวจให้คะแนนทุกข้อของผู้สอบทุกคน หรือ p x i x r และ ผู้ตรวจตรวจข้อสอบเฉพาะข้อของผู้สอบทุกคน หรือ p x (i: r)

การเก็บรวบรวมข้อมูล

ผู้วิจัยนำแบบสอบความสามารถในการเขียนภาษาอังกฤษที่ผ่านการปรับปรุงแก้ไขแล้วไปเก็บข้อมูลกับกลุ่มทดลอง จำนวน 30 ชุด จากนั้นนำกลับมาตรวจให้คะแนน โดยใช้เกณฑ์การให้คะแนนแบบแยกองค์ประกอบที่ประกอบไปด้วย 4 ประเด็นย่อย ใช้ผู้ตรวจคือครูผู้สอนวิชาภาษาอังกฤษจำนวน 3 คน ผู้วิจัยแจ้งวิธีการตรวจโดยใช้เกณฑ์การให้คะแนนแบบแยกองค์ประกอบให้ผู้ตรวจทราบและเกิดความเข้าใจตรงกัน จากนั้นนำผลการตรวจให้คะแนนมาวิเคราะห์ข้อมูล

การวิเคราะห์ข้อมูล

4.1 การศึกษาความสอดคล้องของผลการตรวจให้คะแนนแบบสอบความสามารถในการเขียนภาษาอังกฤษวิเคราะห์ด้วยสัมประสิทธิ์สหสัมพันธ์ภายในชั้น (Intraclass Correlation Coefficient) โดยใช้โปรแกรมสำเร็จรูป SPSS for Windows เวอร์ชัน 26

4.2 การวิเคราะห์ความแปรปรวนของแต่ละองค์ประกอบจากผลการตรวจให้คะแนน ในรูปแบบของการตรวจให้คะแนน 2 รูปแบบ ได้แก่ 1) ตรวจทุกข้อของผู้สอบทุกคน [p x i x r Design] และ 2) ตรวจเฉพาะข้อของผู้สอบบางคน [p x (i: r) Design] ด้วยการศึกษาเพื่อการสรุปอ้างอิง (Generalizability Study) จากนั้นวิเคราะห์ค่าสัมประสิทธิ์การสรุปอ้างอิง (G-Coefficient) โดยกำหนดเงื่อนไขจำนวนผู้ตรวจคือ จำนวน 2 คน และ จำนวน 3 คน โดยใช้การศึกษาการตัดสินใจ (Decision Study)

ผลการวิจัย

1. ผลการวิเคราะห์สัมประสิทธิ์สหสัมพันธ์ภายในชั้น (Intraclass Correlation Coefficient: ICC)

ผู้วิจัยได้ทำการวิเคราะห์ความเที่ยงระหว่างผู้ตรวจให้คะแนน เพื่อพิจารณาความสอดคล้องระหว่างผู้ตรวจให้คะแนน ด้วยโมเดลอิทธิพลผสมแบบสองทาง (Two-way mixed effects model) ที่ระดับความเชื่อมั่น 95% (Confidence interval) โดยใช้ผลการตรวจให้คะแนนของผู้ตรวจจำนวน 3 คน จากข้อสอบ 3 ข้อ ที่ตรวจด้วยรูปแบบการตรวจทุกข้อของผู้สอบทุกคนมาวิเคราะห์ด้วยโปรแกรมสำเร็จรูป SPSS for Window พิจารณาโดยสถิติสหสัมพันธ์ภายในชั้น (Intra-Class Correlation: ICC) และใช้เกณฑ์การแปลความหมาย ซึ่งค่า ICC จะอยู่ระหว่าง 0 ถึง 1 โดยค่า ICC อยู่ในช่วง 0.90 ถึง 1.00 แสดงว่ามีค่าความน่าเชื่อถืออยู่ในระดับดีมาก ค่า ICC อยู่ในช่วง 0.75 ถึง 0.89 แสดงว่ามีค่าความน่าเชื่อถืออยู่ในระดับดี ค่า ICC อยู่ในช่วง 0.50 ถึง 0.74 แสดงว่ามีค่าความน่าเชื่อถืออยู่ในระดับพอใช้ และค่า ICC อยู่ในช่วง 0.00 ถึง 0.49 แสดงว่ามีค่าความน่าเชื่อถืออยู่ในระดับต่ำ (Koo T.K. & Li M.Y., 2016) ซึ่งผลการวิเคราะห์ปรากฏดังตารางที่ 1

ตารางที่ 1 ค่าสัมประสิทธิ์สหสัมพันธ์ภายในชั้น ของการให้คะแนนข้อสอบของผู้ตรวจ

ข้อสอบ	รายบุคคล		โดยเฉลี่ย	
	ICC (95% CI)	ระดับเกณฑ์	ICC (95% CI)	ระดับเกณฑ์
ข้อ 1	0.560 (0.352-0.739)	พอใช้	0.793 (0.620- 0.894)	ดี
ข้อ 2	0.433 (0.210-0.646)	ต่ำ	0.696 (0.433-0.845)	พอใช้
ข้อ 3	0.730 (0.589-0.849)	พอใช้	0.890 (0.798-0.994)	ดี

จากตารางที่ 1 พบว่า ค่าความเที่ยงในการตรวจให้คะแนนระหว่างผู้ตรวจมีความสอดคล้องกันรายบุคคลอยู่ในระดับต่ำและพอใช้ โดยพบว่าข้อสอบ ข้อ 3 มีค่า $ICC_{(3,3)}$ มากที่สุด มีค่าตั้งแต่ 0.589-0.849 ข้อ 1 มีค่า $ICC_{(3,3)}$ รองลงมา โดยพบค่าตั้งแต่ 0.352-0.739 และข้อ 2 มีค่า $ICC_{(3,3)}$ ต่ำสุด โดยพบค่าตั้งแต่ 0.210-0.646 ส่วนค่าความเที่ยงในการตรวจให้คะแนนระหว่างผู้ตรวจมีความสอดคล้องกันโดยเฉลี่ยอยู่ในระดับพอใช้และดี โดยพบว่าข้อสอบ ข้อ 3 มีค่า $ICC_{(3,3)}$ มากที่สุด มีค่าตั้งแต่ 0.798-0.994 ข้อ 1 มีค่า $ICC_{(3,3)}$ รองลงมา มีค่าตั้งแต่ 0.620- 0.894 และข้อ 2 มีค่า $ICC_{(3,3)}$ ต่ำสุด มีค่าตั้งแต่ 0.433-0.845

2. ผลการวิเคราะห์ความแปรปรวนขององค์ประกอบจากแหล่งความแปรปรวนต่างๆ มีผลต่อสัมประสิทธิ์การสรุปอ้างอิง (G-Coefficient)

2.1 ผลการวิเคราะห์ขนาดความแปรปรวนของแต่ละองค์ประกอบที่มีผลต่อสัมประสิทธิ์การสรุปอ้างอิง (G-Coefficient) ของแบบสอบความสามารถในการเขียนภาษาอังกฤษ ที่ได้จากการออกแบบการวัดแบบ p x i x r หรือผู้ตรวจตรวจให้คะแนนทุกข้อของผู้สอบทุกคน (Two-Facet Full Crossed Design) และ p x (i: r) ผู้ตรวจตรวจข้อสอบเฉพาะข้อของผู้สอบทุกคน (Two-Facet Confounded Design) ที่ประกอบไปด้วยจำนวน ผู้สอบ 30 คน ข้อสอบ 3 ข้อ และผู้ตรวจ 3 คน ใช้การวิเคราะห์ความแปรปรวน 3 ทาง (3-WAY ANOVA) ปรากฏผลการวิเคราะห์แสดงดังตารางต่อไปนี้

ตารางที่ 2 ผลการวิเคราะห์ความแปรปรวนจาก G-Study (p x i x r) ของแบบสอบความสามารถในการเขียนภาษาอังกฤษที่มีจำนวนข้อสอบ 3 ข้อ และผู้ตรวจ 3 คน

Source of variance	df	SS	MS	Estimated variance components	% of total variance
P	29	1077.189	37.144	2.553	21.497
I	2	31.400	15.700	0.085	0.716
R	2	661.067	330.533	3.595	30.271
PI	58	481.711	8.305	2.307	19.426
PR	58	420.044	7.242	1.953	16.445
IR	4	4.466	1.116	-0.009	0.000
PIR	116	160.422	1.383	1.383	11.645
Total	269	2836.299	401.423	11.867	100.000

จากตารางที่ 2 พบว่า องค์ประกอบจากแหล่งความแปรปรวนของแบบสอบความสามารถในการเขียนภาษาอังกฤษที่ใช้รูปแบบการตรวจ ผู้ตรวจตรวจให้คะแนนทุกข้อของผู้สอบทุกคน มีความแปรปรวนรวมทั้งหมดเท่ากับ 11.867 องค์ประกอบที่มีความแปรปรวนมากที่สุดคือ ความแปรปรวนจากผู้ตรวจ (σ_r^2) มีค่าเท่ากับ 3.595 คิดเป็นร้อยละ 30.271 ของความแปรปรวนทั้งหมด รองลงมาคือความแปรปรวนจากผู้สอบ (σ_p^2) มีค่าเท่ากับ 2.553 คิดเป็นร้อยละ 21.497 ของความแปรปรวนทั้งหมด ส่วนองค์ประกอบที่มีความแปรปรวนน้อยที่สุดคือ ความแปรปรวนจากปฏิสัมพันธ์ระหว่างข้อสอบและผู้ตรวจ (σ_{ir}^2) มีค่าเท่ากับ -0.009 คิดเป็นร้อยละ 0 จากความแปรปรวนทั้งหมด

ตารางที่ 3 ผลการวิเคราะห์ความแปรปรวนจาก G-Study [p x (i:r)] ของแบบสอบความสามารถในการเขียนภาษาอังกฤษ ที่มีจำนวนข้อสอบ 3 ข้อ และผู้ตรวจ 3 คน

Source of variance	df	SS	MS	Estimated variance components	% of total variance
P	29	1077.189	37.144	3.756	30.452
R	2	31.400	15.700	-0.742	0.000
I:R	6	665.533	110.922	3.586	29.074
PR	58	481.711	8.305	1.656	13.426
PI:R	174	580.467	3.336	3.336	27.047
Total	269	2836.300		11.592	100.000

ในขณะที่รูปแบบการตรวจแบบผู้ตรวจตรวจข้อสอบเฉพาะข้อของผู้สอบทุกคน พบว่าองค์ประกอบที่มีความแปรปรวนมากที่สุดคือ ความแปรปรวนจากผู้สอบ (σ_p^2) มีค่าเท่ากับ 3.756 คิดเป็นร้อยละ 30.452 ของความแปรปรวนทั้งหมด รองลงมาคือความแปรปรวนจากปฏิสัมพันธ์ระหว่างผู้ตรวจและข้อสอบ(σ_{ir}^2) มีค่าเท่ากับ 3.586 คิดเป็นร้อยละ 29.074 ของความแปรปรวนทั้งหมด ส่วนองค์ประกอบที่มีความแปรปรวนน้อยที่สุดคือ ความแปรปรวนจากผู้ตรวจ(σ_r^2) มีค่าเท่ากับ -0.742 คิดเป็นร้อยละ 0 จากความแปรปรวนทั้งหมด แสดงผลดังตารางที่ 3

2.2 ผลการศึกษาเพื่อการตัดสินใจการสรุปอ้างอิง (D-Study) ของผลการตรวจให้คะแนนของแบบสอบที่มีจำนวนผู้ตรวจและรูปแบบการตรวจต่างกัน ของแบบสอบความสามารถในการเขียนภาษาอังกฤษที่มีจำนวนข้อ 3 ข้อ และมีจำนวนผู้ตรวจ 3 คน ซึ่งผู้วิจัยควบคุมแหล่งความคลาดเคลื่อนโดยกำหนดเงื่อนไขรูปแบบการตรวจ ได้แก่ ผู้ตรวจให้คะแนนทุกข้อของผู้สอบทุกคน (p x i x r Design) และผู้ตรวจตรวจข้อสอบเฉพาะข้อของผู้สอบทุกคน [p x (i: r) Design] นอกจากนี้ผู้วิจัยกำหนดเงื่อนไขของจำนวนผู้ตรวจคือ จำนวน 2 คน และ จำนวน 3 คน แสดงผลดังตารางต่อไปนี้

ตารางที่ 4 ผลการศึกษาเพื่อการตัดสินใจสรุปอ้างอิง (D-Study) ของผลการตรวจให้คะแนนของแบบสอบความสามารถในการเขียนภาษาอังกฤษ ที่มีจำนวนผู้ตรวจและรูปแบบการตรวจต่างกัน

รูปแบบการตรวจ (Design)		จำนวนผู้ตรวจ	
		2	3
p x i x r Design	ρ_{δ}^{2}	0.5637	0.6187
	ρ_{Δ}^{2}	0.4018	0.4769
p x (i: r) Design	ρ_{δ}^{2}	0.7948	0.9102
	ρ_{Δ}^{2}	0.7055	0.8301

จากตารางที่ 4 พบว่า รูปแบบการตรวจ ผู้ตรวจตรวจข้อสอบเฉพาะข้อของผู้สอบทุกคน [p x (i: r) Design] เมื่อกำหนดจำนวนผู้ตรวจ 2 คนและ 3 คน มีค่าสัมประสิทธิ์สรุปอ้างอิงสำหรับการตัดสินใจสัมพัทธ์ (ρ_{δ}^{2}) เท่ากับ 0.7948 และ 0.9102 ตามลำดับ และค่าสัมประสิทธิ์สรุปอ้างอิงสำหรับการตัดสินใจสัมบูรณ์ (ρ_{Δ}^{2}) มีค่าเท่ากับ 0.7055 และ 0.8301 ตามลำดับ ซึ่งสูงกว่ารูปแบบการตรวจผู้ตรวจให้คะแนนทุกข้อของผู้สอบทุกคน [p x i x r Design] ในทั้งเงื่อนไขการตรวจที่มีจำนวนผู้ตรวจ 2 คน และ 3 คน โดยมีค่าสัมประสิทธิ์สรุปอ้างอิงที่สูงสุดคือรูปแบบการตรวจ p x (i: r) Design เมื่อมีจำนวนผู้ตรวจ 3 คน

อภิปรายและสรุปผลการวิจัย

1. ผลการวิเคราะห์ความสอดคล้องของผลการตรวจให้คะแนนแบบสอบความสามารถในการเขียนภาษาอังกฤษของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ด้วยค่าสัมประสิทธิ์สหสัมพันธ์ภายในชั้น (Intraclass Correlation Coefficient :ICC) พบว่าความเที่ยงในการตรวจให้คะแนนระหว่างผู้ตรวจมีความสอดคล้องกันในระดับพอใช้ และดี เนื่องจากผู้ตรวจมีความเห็นไปในทิศทางที่ต่างกันในบางสถานการณ์คำถาม ถึงแม้ว่าจะมีเกณฑ์การให้คะแนนแบบแยกองค์ประกอบที่ให้ค่าความเที่ยงสูงกว่าเกณฑ์การให้คะแนนแบบภาพรวม (Weigle, 2002) เนื่องจากแบบสอบดังกล่าวให้นักเรียนสามารถเลือกหัวข้อในการเขียนตอบด้วยตนเอง จึงทำให้เกิดความคลาดเคลื่อนในการตรวจให้คะแนนที่มาจากหลายปัจจัย เช่น ความเข้มงวด/ใจดีของผู้ตรวจ ความลำเอียง คุณลักษณะการให้คะแนนของผู้ตรวจ (Characteristics of rater) หากผู้ตรวจมีปัจจัยดังกล่าวที่ใกล้เคียงกันจะทำให้ความเห็นสอดคล้องกันสูง (Hopkins & Antes,1990) สอดคล้องกับดรุณี อภัยกาวี(2562) ที่กล่าวว่าสาเหตุของแบบแผนการให้คะแนนที่แตกต่างกันนั้นอาจมาจากคุณลักษณะของผู้ตรวจ เช่น ประสบการณ์การทำงาน การจบการศึกษาตรงตามสาขาวิชา การได้รับการฝึกอบรม เป็นต้น หากมีการควบคุมปัจจัยดังกล่าวจะทำให้ผู้ตรวจมีแบบแผนในการตรวจไปในทิศทางเดียวกันมากขึ้น นอกจากนี้คุณภาพของเครื่องมือ ทักษะของผู้ตรวจ และการบริหารการตรวจให้คะแนน ยังมีผลต่อความคลาดเคลื่อนในการตรวจให้คะแนนเช่นกัน (ศิริชัย กาญจนวาสี, 2555)

2. ผลการวิเคราะห์ขนาดความแปรปรวนของแต่ละองค์ประกอบที่มีผลต่อสัมประสิทธิ์การสรุปอ้างอิง (G-Coefficient) จากรูปแบบการตรวจ p x i x r พบว่าแหล่งความแปรปรวนที่มากที่สุดคือ ความแปรปรวนจากผู้ตรวจ แสดงว่าผู้ตรวจมีการให้คะแนนผู้สอบแตกต่างกัน อาจเป็นผลมาจากจำนวนของข้อสอบซึ่งข้อสอบดังกล่าวเป็นข้อสอบการเขียนเรียงความภาษาอังกฤษซึ่งต้องใช้เวลาในการอ่านเนื้อเรื่องเพื่อตัดสินการให้คะแนน ประกอบกับเกณฑ์การให้คะแนนที่มีเกณฑ์ย่อยและระดับการให้คะแนนหลายระดับจึงอาจทำให้ผู้ตรวจเกิดความเมื่อยล้าในการตรวจ และข้อสอบดังกล่าวมีหัวข้อในการเขียนที่หลากหลายจึงอาจส่งผลต่อความคิดเห็นของผู้ตรวจด้วยเช่นกัน ซึ่งสอดคล้องกับผลการวิเคราะห์ความเที่ยงในการ

ตรวจให้คะแนนระหว่างผู้ตรวจที่มีความสอดคล้องแตกต่างกันปรากฏในหลายระดับ และรูปแบบการตรวจ p x (i: r) พบว่าแหล่งความแปรปรวนที่มีค่ามากที่สุดคือ ความแปรปรวนจากผู้สอบ แสดงว่า ผู้สอบมีความสามารถที่หลากหลายทำให้คะแนนที่ได้จากการทดสอบมีความแตกต่างกันจึงทำให้เกิดความแปรปรวน สอดคล้องกับ นิภาพร ฉันสิมา (2562) ที่พบว่าความแปรปรวนจากผู้สอบมีค่ามากที่สุดจากรูปแบบการตรวจแบบ p x (i: r) ซึ่งเกิดจากความสามารถในการทำข้อสอบที่แตกต่างกันของผู้สอบจึงส่งผลต่อค่าสัมประสิทธิ์การสรุปอ้างอิง

จากการศึกษาค่าสัมประสิทธิ์สรุปอ้างอิงที่มีรูปแบบการตรวจข้อสอบเฉพาะข้อของผู้สอบทุกคน หรือ p x *(i: r)* มีค่าสูงกว่ารูปแบบการตรวจให้คะแนนทุกข้อของผู้สอบทุกคน หรือ p x i x r ในทุกเงื่อนไขของการตรวจ เนื่องจากรูปการตรวจข้อสอบเฉพาะข้อของผู้สอบทุกคนนั้นจะลดโอกาสการเกิดความแปรปรวนจากการให้คะแนนของผู้ตรวจเพราะใช้คะแนนตรวจจากผู้ตรวจเพียงคนเดียวสำหรับข้อสอบข้อนั้น ๆ และผู้ตรวจจะมีความแม่นยำหรือความเที่ยงในการตรวจมากกว่าเพราะตรวจในเฉพาะข้อที่ตนเองได้รับมอบหมายเท่านั้น ซึ่งในขณะที่รูปแบบการตรวให้คะแนนทุกข้อของผู้สอบทุกคนนั้นใช้คะแนนจากผู้ตรวจทุกคน สอดคล้องกับการศึกษาของ ชนิสรา สงวนไว้ (2558) และดรุณี อภัยกาวี (2562) ซึ่งพบว่า รูปแบบการตรวจให้คะแนน ตรวจเฉพาะข้อของผู้สอบทุกคน มีค่าสัมประสิทธิ์การสรุปอ้างอิงสูงกว่ารูปแบบอื่น ๆ นอกจากนี้ยังพบว่าเมื่อจำนวนผู้ตรวจเพิ่มขึ้น ส่งผลต่อค่าสัมประสิทธิ์การสรุปอ้างอิงสำหรับการตัดสินใจเชิงสัมพัทธ์ (Relative coefficient) และค่าสัมประสิทธิ์การสรุปอ้างอิงสำหรับการตัดสินใจเชิงสัมบูรณ์ (Absolute coefficient) ที่สูงขึ้น จากการศึกษาจึงพบว่าเมื่อใช้รูปแบบการตรวจเฉพาะข้อของผู้สอบบางคน จะสามารถช่วยลดค่าใช้จ่าย ภาระงานและระยะเวลาในการตรวจ รวมถึงส่งผลต่อค่าสัมประสิทธิ์การสรุปอ้างอิงที่มากขึ้นทำให้เกิดความน่าเชื่อถือของคะแนนและให้ค่าสถิติที่ดีขึ้น

เอกสารอ้างอิง

จิรายุ เถาว์โท. (2559). การศึกษาค่าความเชื่อมั่นของคะแนนแบบทดสอบอัตนัยวิชาคณิตศาสตร์ของนักเรียน ชั้นมัธยมศึกษาปีที่ 2 ที่มีจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน โดยใช้ทฤษฎีการสรุปอ้างอิง. วารสารหาดใหญ่วิชาการ, 14(1), 1-14.

ชนิสรา สงวนไว้. (2558). การเปรียบเทียบความเที่ยงของแบบสอบวัดความสามารถในการแก้ปัญหาอย่าง สร้างสรรค์ทางคณิตศาสตร์ :การประยุกต์ใช้ทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัด. ปริญญานิพนธ์ ค.ม. (การวัดและประเมินผลการศึกษา). กรุงเทพฯ: บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย.

ดรุณี อภัยกาวี. (2562). ผลการตรวจให้คะแนนของแบบสอบอัตนัย เมื่อกลุ่มผู้ตรวจและรูปแบบการตรวจที่ต่างกัน. การวัดประเมินผล และวิจัยสัมพันธ์ แห่งประเทศไทยครั้งที่ 27, 108-124.

นิภาพร ฉันสิมา. (2562). การเปรียบเทียบการตรวจให้คะแนนของแบบทดสอบอัตนัยภายใต้ระดับความลึกความเข้าใจต่างกัน:การประยุกต์ใช้ทฤษฎีการสรุปอ้างอิง. การวัดผล ประเมินผล และวิจัยสัมพันธ์แห่งประเทศไทยครั้งที่ 27, 98-107.

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

ศิริชัย กาญจนวาสี. (2555). ทฤษฎีการทดสอบแนวใหม่ (พิมพ์ครั้งที่ 4). กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

ศิริชัย กาญจนวาศี และ คณะ. (2558). ยุทธศาสตร์การกระจายอำนาจการจัดการศึกษา. สำนักงานเลขาธิการสภาการศึกษา.

American Council for the Teaching of Foreign Langu. (2012). ACTFL Proficiency Guidelines 2012. Retrieved March 3, 2021, from https://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012/english/writing

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific performance test. Language Testing, 12, 1-15.

Coombe , C., & Evans, J. (2001). Writing assessment Scales: Making the right choice. TESOL Arabia News, 8(1), 7-9.

Cumming, A., Kantor, R., & Powers, D. E. (2001). Scoring TOEFL essays and TOEFL 2000 prototype written tasks: An investigation into raters' decision making and development of a preliminary analytic framework (TOEFL Monograph Series No. 22 ed.). Princeton: NJ: Educational Testing Service.

Harold , S. M. (1983). Techniques In Testing. New York and Oxford: Oxford University Press.

Hinkel, E. (1994). Native and non-native speakers' pragmatic interpretations of English texts. TESOL Quarterly 28(2), 353-76.

Hughes, A. (2003). Testing for Language Teachers (2nd ed.). Cambridge: Cambridge University Press.

IBM Corp. (2019). IBM SPSS Statistics for Windows, Version 26.0. Armonk, NY: IBM Corp.

Koo, T. K., & Li, M. Y. (2016). Guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2), 155-163. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4913118/

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

Scott, V. M. (1996). From Rethinking Foreign Language Writing. Boston: Heinle & Heinle.

Weigle, S. C. (2002). Assessing writing. Cambridge: Cambridge University Press.