

การสรุปอ้างอิงความน่าเชื่อถือของการประเมินความสอดคล้องในแนวเดียวกัน ระหว่างข้อสอบกับตัวชี้วัดวิทยาศาสตร์: การประยุกต์ใช้ทฤษฎีการสรุปอ้างอิง ความน่าเชื่อถือของผลการวัด*

A Generalizability Study of Science Items-Indicators Alignment Evaluation: An Application of Generalizability Theory*

บุษยรัตน์ จันทร์ประเสริฐ¹ ณัฐภรณ์ หลาวทอง² สังวรณ์ จัตุระโทก³

Budsarat Janprasert¹ Nattaporn Lawthong² Sungworn Ngudgratoke³

¹นักศึกษาปริญญาตรีบัณฑิต สาขาวิชาการวัดและประเมินผลการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

²ผู้ช่วยศาสตราจารย์ ดร.ประจำสาขาวิชาการวัดและประเมินผลการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

³ผู้ช่วยศาสตราจารย์ ดร.ประจำสาขาวิชาศึกษาศาสตร์ มหาวิทยาลัยสุโขทัยธรรมาธิราช

*งานวิจัยเรื่องนี้ได้รับการสนับสนุนทุนวิจัยจากทุน 90 ปี จุฬาลงกรณ์มหาวิทยาลัย กองทุนรัชดาภิเษกสมโภช

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อ 1) ศึกษาขนาดขององค์ประกอบความแปรปรวนของคะแนนการประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัดกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับชั้นมัธยมศึกษาตอนต้น 2) เปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนการประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัดฯ เมื่อจำนวนผู้ประเมินและรูปแบบการออกแบบการประเมินต่างกัน และ 3) วิเคราะห์จำนวนผู้ประเมินที่ทำให้เกิดความน่าเชื่อถือของการประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัดฯ ในระดับที่ยอมรับได้ ตัวอย่างในการวิจัย ได้แก่ ข้อสอบในการประเมินระดับชั้นเรียน กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ชั้นมัธยมศึกษาปีตอนต้น ของโรงเรียนมัธยมศึกษา สังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน ในกรุงเทพมหานคร จำนวน 1,089 ข้อ และผู้เชี่ยวชาญในการประเมินความสอดคล้องฯ จำนวน 20 คน เครื่องมือที่ใช้ในการวิจัย คือ แบบประเมินความสอดคล้องใน แนวเดียวกัน จากการตรวจสอบความเที่ยงระหว่างผู้ประเมินด้วยสถิติเคลปาของฟลีส (Fleiss' kappa statistic: Kf) มีค่าเท่ากับ 0.510 และสถิติสหสัมพันธ์ภายในชั้น (intra-class correlation: ICC) ที่ระดับความเชื่อมั่น 95% มีค่าเท่ากับ 0.954 วิเคราะห์และเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิง (G-coefficient) ตามทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัด เมื่อจำนวนผู้ประเมินและรูปแบบการออกแบบการประเมินต่างกัน ผลการวิจัยพบว่า 1) ค่าความแปรปรวนของคะแนนความสอดคล้องในแนวเดียวกันของข้อสอบมีค่าสูงสุด และความแปรปรวนอันเกิดจากผู้ประเมินมีค่าต่ำสุด ตามลำดับ 2) ค่าสัมประสิทธิ์การสรุปอ้างอิงมีค่าเพิ่มขึ้นเมื่อจำนวนผู้ประเมินเพิ่มขึ้น ในทุกรูปแบบการออกแบบการประเมินทั้งในส่วนของการประเมินระดับความซับซ้อนทางปัญญา และในส่วนของการประเมินระดับความสอดคล้อง

ในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด ด้วยมาตรประมาณค่า 5 ระดับ และ 3) การประเมินระดับความซับซ้อนทางปัญญาต้องการจำนวนผู้ประเมินมากกว่าการประเมินระดับความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด ที่มีลักษณะเป็นมาตรประมาณค่า 5 ระดับ

คำสำคัญ: ความสอดคล้องในแนวเดียวกัน ทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัด
สัมประสิทธิ์การสรุปอ้างอิง

Abstract

This research aimed to (1) study the relative magnitude of variance components of the evaluating alignment scores between science items and indicators at the lower secondary level; (2) compare the generalizability coefficients of the evaluation scores of the science items-indicators alignment when the numbers of raters and evaluation designs are different; and (3) analyze the number of raters that contribute to reliability of the analysis of the science items-indicators alignment. Research subjects were 1,089 science classroom test items used at the lower secondary level in secondary schools under the Office of the Basic Education Commission in Bangkok Metropolis, and 20 expert panelists who evaluated the alignment. The employed research instrument was an alignment assessment rating scale. The inter-rater reliability examination using the Fleiss' kappa statistic was 0.510, and 95% confidence interval for the intra-class correlation was 0.954. The G-coefficients were analyzed with the use of generalizability theory and compared. The findings revealed that (1) the maximum value of the variance was in alignment between items and indices scores, and the minimum value of the variance was in raters; (2) the generalizability coefficient increased when the number of raters increased in both cognitive demand evaluation and alignment between items and indices level evaluation; and (3) the cognitive demand evaluation needed greater number of raters than did the alignment between items and indicators level evaluation.

Keywords: Alignment, Generalizability Theory, G-coefficient

บทนำ

ความสอดคล้องในแนวเดียวกัน หรือ “alignment” ในทางการศึกษา หมายถึง ระดับความเชื่อมโยงสอดคล้องสัมพันธ์กันขององค์ประกอบต่างๆ ในระบบการศึกษา เช่น ผลการเรียนรู้ที่คาดหวัง (เช่น มาตรฐานการเรียนรู้ ตัวชี้วัด วัตถุประสงค์) การจัดการเรียนการสอน และการประเมินผล (เช่น ข้อสอบ) (Ananda, 2003; Bhola, Impara, & Buckendahl, 2003; Biggs, 2003; Case, Jorgensen, & Zucker, 2004; La Marca, 2001; Webb, 1997) ความสอดคล้องในแนวเดียวกันระหว่างมาตรฐานกับการประเมินผลเป็นหลักฐานที่สำคัญของการประเมินความถูกต้องของ

การตีความคะแนนที่ได้จากการวัด (La Marca, 2001; Rothman, 2003) เนื่องจากช่วยแสดงให้เห็นถึงความสอดคล้องระหว่างมาตรฐานกับการประเมินผล ช่วยตรวจสอบว่าการจัดการเรียนการสอนมีความสอดคล้องครบถ้วนหรือไม่ นอกจากนี้ยังช่วยในการตรวจสอบได้ว่าผู้เรียนสามารถบรรลุผลการเรียนรู้ที่คาดหวังของความรู้และทักษะต่างๆ ที่กำหนดในมาตรฐานหรือไม่ ซึ่งเป็นการส่งเสริมโอกาสที่ผู้เรียนจะบรรลุตามมาตรฐานและการเรียนรู้ที่คาดหวัง (Linn, 1994)

วิธีการศึกษาความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับมาตรฐานและตัวชี้วัด โดยทั่วไปต้องอาศัยผู้เชี่ยวชาญพิจารณาตัดสินความสอดคล้องระหว่างเนื้อหาของข้อสอบกับมีเนื้อหาของมาตรฐานและตัวชี้วัดที่ต้องการวัด เช่น วิธีการของ Webb (1997, 1999) วิธีการของ Porter and Smithson (2001, 2002) และวิธีการของบริษัท Achieve (Rothman, Slattery, Vranek, & Resnick, 2002) มีข้อกำหนดให้ผู้เชี่ยวชาญจะต้องมีความรู้และคุ้นเคยในเนื้อหาของวิชาที่ต้องการประเมิน ลักษณะของผู้ถูกประเมิน หลักสูตร ตลอดจนมาตรฐานการเรียนรู้ที่ต้องการประเมิน (Davis-Becker & Buckendahl, 2013; La Marca, Redfield, & Winter, 2000) ทั้งนี้ Webb (2007) กล่าวว่า การเพิ่มจำนวนผู้เชี่ยวชาญจะส่งผลให้ค่าความเที่ยงเพิ่มสูงขึ้น และเสนอแนะว่าในการศึกษาความสอดคล้องในแนวเดียวกัน ควรมีผู้เชี่ยวชาญจำนวน 5 – 7 คน Herman, Webb, and Zuniga (2007) และ Lombardi, Sebuen, Conley and Snow (2010) กล่าวว่า ผู้เชี่ยวชาญจำนวน 6 คน จะทำให้ความเที่ยงมีค่าอยู่ในเกณฑ์ที่ยอมรับได้ หรือทำให้ผลการประเมินมีความน่าเชื่อถือ อย่างไรก็ตามจำนวนผู้เชี่ยวชาญที่เหมาะสมนั้นยังไม่มีข้อสรุป ทั้งนี้หากเป็นงานวิจัยที่ต้องการมีการประเมินความสอดคล้องในแนวเดียวกันของข้อสอบจำนวนมาก จำนวนผู้เชี่ยวชาญก็จะมากขึ้นด้วย ส่งผลต่อการบริหารจัดการ งบประมาณ เวลาในการศึกษา ดังนั้นเพื่อช่วยลดปัญหาดังกล่าว จึงจำเป็นต้องมีการศึกษาการออกแบบการศึกษาที่ผู้เชี่ยวชาญแต่ละคนอาจไม่จำเป็นต้องประเมินข้อสอบทุกข้อ แต่จะมีการวิเคราะห์ความน่าเชื่อถือของผลการประเมินสำหรับสถานการณ์ของการประเมินลักษณะต่างๆ ให้สามารถสรุปความเที่ยงในสถานการณ์หรือเงื่อนไขการวัดในลักษณะอื่นๆ ได้ และให้ข้อสรุปถึงสถานการณ์ที่ให้ความเที่ยงที่เหมาะสมในระดับที่สามารถยอมรับได้ จึงต้องอาศัยหลักการแนวคิดของทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัด (generalizability theory: G-theory) เพื่อให้ได้คำตอบดังกล่าว

ทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัด เป็นทฤษฎีที่ได้ขยายแนวคิดของความเที่ยงในทฤษฎีการทดสอบแบบดั้งเดิม (classical test theory) ให้สามารถสรุปอ้างอิงค่าความเที่ยงไปยังสถานการณ์หรือเงื่อนไขการวัดในลักษณะต่างๆ ได้ (ศิริชัย กาญจนวาสี, 2555) โดยใช้การประมาณค่าความแปรปรวนจากแหล่งต่างๆ ด้วยการวิเคราะห์ความแปรปรวน (analysis of variance: ANOVA) รวมทั้งสามารถจำแนกแหล่งความแปรปรวนที่เป็นไปได้ทั้งหมด เพื่อให้ได้ค่าความแปรปรวนของความคลาดเคลื่อนที่ถูกต้อง ซึ่งจะช่วยให้ประมาณค่าความเที่ยงได้ถูกต้องแม่นยำ สามารถใช้ออกแบบรูปแบบการวัด เพื่อให้ได้ค่าความเที่ยงที่สูงขึ้นได้ในหลายสถานการณ์วัด โดยการเพิ่มหรือลดจำนวนองค์ประกอบความแปรปรวนหรือพาเซตต่างๆ เช่น จำนวนข้อสอบ จำนวนครั้งในการวัด จำนวนผู้ตรวจหรือจำนวนผู้ประเมิน เป็นต้น ซึ่งแนวคิดของทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัดสามารถนำมาประยุกต์ใช้ในการศึกษาความน่าเชื่อถือของการประเมินความสอดคล้องในแนวเดียวกันได้ (Lombardi, Sebuen, Conley, & Snow, 2010)

งานวิจัยนี้สนใจการศึกษาขนาดขององค์ประกอบความแปรปรวนของคะแนนการประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัดกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับชั้นมัธยมศึกษาตอนต้น และเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนการประเมินความสอดคล้องในแนวเดียวกัน เมื่อจำนวนผู้ประเมินและรูปแบบการออกแบบการประเมินต่างกัน ตลอดจนวิเคราะห์จำนวนผู้ประเมินที่ทำให้เกิดความน่าเชื่อถือของการประเมินความสอดคล้องในแนวเดียวกัน ในระดับที่ยอมรับได้ ผลการวิจัยจะให้สารสนเทศอันเป็นแนวทางสำหรับออกแบบสถานการณ์หรือเงื่อนไขของการวัดความสอดคล้องในแนวเดียวกัน เพื่อนำไปสู่การตัดสินใจเลือกสถานการณ์การประเมินที่มีความน่าเชื่อถือ ภายใต้ทรัพยากรในการบริหารจัดการที่เหมาะสมและมีความเป็นไปได้ในทางปฏิบัติ

วัตถุประสงค์

1. เพื่อศึกษาขนาดขององค์ประกอบความแปรปรวนของคะแนนการประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัดกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับชั้นมัธยมศึกษาตอนต้น
2. เพื่อเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนการประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัดกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับชั้นมัธยมศึกษาตอนต้น เมื่อจำนวนผู้ประเมินและรูปแบบการออกแบบการประเมินต่างกัน
3. เพื่อวิเคราะห์จำนวนผู้ประเมินที่ทำให้เกิดความน่าเชื่อถือของการประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัดกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับชั้นมัธยมศึกษาตอนต้น ในระดับที่ยอมรับได้

วิธีดำเนินการวิจัย

1. ประชากรและตัวอย่างในการวิจัย

ประชากรในการวิจัยครั้งนี้ ได้แก่

1) ข้อสอบในการประเมินผลการเรียนรู้ระดับชั้นเรียน กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับชั้นมัธยมศึกษาตอนต้น (มัธยมศึกษาปีที่ 1 – 3) ของโรงเรียนมัธยมศึกษา สังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน ในกรุงเทพมหานคร ซึ่งประกอบด้วยข้อสอบที่ครูสร้างขึ้น (teacher-made test) เพื่อประเมินผลการเรียนรู้ของผู้เรียนทั้งข้อสอบระหว่างภาคและข้อสอบประจำภาค ในปีการศึกษา 2559

2) ผู้เชี่ยวชาญในการตรวจสอบความสอดคล้องในแนวเดียวกัน ได้แก่ ครูสอนวิทยาศาสตร์ ชั้นมัธยมศึกษาตอนต้น และ/หรือนักวิชาการ และ/หรืออาจารย์ระดับอุดมศึกษาที่สอนในรายวิชาวิทยาศาสตร์ และ/หรือสาขาที่เกี่ยวข้องกับการวัดและประเมินผลการศึกษา

ตัวอย่างในการวิจัย ได้แก่

1) ข้อสอบในการประเมินผลการเรียนรู้ระดับชั้นเรียน กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ชั้นมัธยมศึกษาตอนต้นของโรงเรียนมัธยมศึกษา สังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน ในกรุงเทพมหานคร จำนวนทั้งหมด 1,089 ข้อ ซึ่งประกอบด้วยข้อสอบที่ครูสร้างขึ้นเพื่อประเมินผลการเรียนรู้ของผู้เรียน ทั้งข้อสอบระหว่าง

ภาคและข้อสอบประจำภาค ในปีการศึกษา 2559 โดยเป็นข้อสอบที่โรงเรียนใช้วัดตัวชี้วัดที่สำนักทดสอบทางการศึกษาแห่งชาติ (สทศ.) ใช้ในการออกข้อสอบ O-NET วิทยาศาสตร์ ระดับชั้นมัธยมศึกษาปีที่ 3 ประจำปีการศึกษา 2559 ซึ่งมีจำนวน 40 ตัวชี้วัด การได้มาซึ่งตัวอย่างข้อสอบใช้วิธีการสุ่มแบบหลายขั้นตอน (multi-stage sampling) โดยทำการสุ่มอย่างง่ายโรงเรียนมัธยมศึกษา สังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน ในกรุงเทพมหานคร จำนวน 4 โรงเรียนเพื่อขอข้อสอบมาใช้ในการวิจัย ได้แบบสอบทั้งหมด 48 ชุด จากนั้นทำการคัดเลือกเฉพาะข้อสอบที่ครูระบุว่าใช้ในการวัดตัวชี้วัดที่ สทศ. ใช้ในการออกข้อสอบ O-NET ทั้ง 40 ตัวชี้วัด ได้ข้อสอบที่เป็นตัวอย่างในการวิจัยจำนวนทั้งหมด 1,089 ข้อ

2) ผู้เชี่ยวชาญในการตรวจสอบความสอดคล้องในแนวเดียวกัน จำนวน 20 คน ประกอบด้วย ครูสอนวิทยาศาสตร์ ชั้นมัธยมศึกษาตอนต้น และ/หรือนักวิชาการ และ/หรืออาจารย์ระดับอุดมศึกษาที่สอนในรายวิชาวิทยาศาสตร์ และ/หรือสาขาวิชาที่เกี่ยวข้องกับการวัดและประเมินผลการศึกษา ซึ่งได้มาจากการเลือกแบบเจาะจง (purposive sampling)

2. ตัวแปรที่ใช้ในการวิจัย

ตัวแปรตาม มี 2 ตัว ได้แก่

1) ขนาดขององค์ประกอบความแปรปรวนของคะแนนการประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบตัวชี้วัดวิทยาศาสตร์

2) ค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนการประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัดวิทยาศาสตร์

ตัวแปรอิสระ มี 2 ตัว ได้แก่

1) จำนวนผู้ประเมิน มีตั้งแต่ 1 – 20 คน

2) รูปแบบการออกแบบการประเมิน มี 2 ลักษณะ คือ ผู้ประเมินแต่ละคนประเมินข้อสอบทุกข้อ และผู้ประเมินแต่ละคนประเมินข้อสอบบางข้อ

3) รูปแบบข้อรายการในการประเมิน มี 2 รูปแบบ คือ ข้อรายการในการประเมินระดับความซับซ้อนทางปัญญา และข้อรายการในการประเมินระดับความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด ที่มีลักษณะเป็นมาตรฐานค่า 5 ระดับ

3. เครื่องมือที่ใช้ในการวิจัย

เครื่องมือในการวิจัยครั้งนี้ ได้แก่ แบบประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัดกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับชั้นมัธยมศึกษาตอนต้น ซึ่งแต่ละข้อรายการประกอบด้วย 2 ส่วนหลัก คือ

ส่วนที่ 1 รายละเอียดของข้อสอบที่ถูกประเมิน ประกอบด้วย ข้อคำถามและตัวเลือก (กรณีเป็นข้อสอบหลายตัวเลือก) คะแนนรายข้อ ระดับชั้น และมาตรฐานและตัวชี้วัดที่ข้อสอบข้อนั้นๆ ต้องการวัด

ส่วนที่ 2 การประเมินความสอดคล้องของข้อสอบโดยผู้เชี่ยวชาญใน 3 ประเด็นหลัก ได้แก่

1) ระดับความซับซ้อนทางปัญญา (cognitive complexity) ซึ่งเป็นลำดับขั้นของกระบวนการทางปัญญา (cognitive process dimension) ตามแนวคิดการจำแนกพฤติกรรมการเรียนรู้ด้านพุทธิพิสัยของ Bloom ที่ได้รับการ-

ปรับปรุงใหม่ (revised Bloom's taxonomy) (Anderson et al, 2001) ซึ่งแบ่งเป็น 6 ชั้น ได้แก่ จำ เข้าใจ ประยุกต์ใช้ วิเคราะห์ ประเมินค่า และสร้างสรรค์

2) ระดับความสอดคล้องระหว่างเนื้อหาข้อสอบกับเนื้อหาตัวชี้วัดที่กำหนด มีลักษณะเป็นมาตรฐานประมาณค่า 5 ระดับ (0 = ไม่สอดคล้อง, 1 = ค่อนข้างสอดคล้อง, 2 = ไม่แน่ใจ, 3 = ค่อนข้างสอดคล้อง, 4 = สอดคล้องโดยตรง) กำหนดเกณฑ์ค่าเฉลี่ยคะแนนความสอดคล้อง ต้องมีค่าตั้งแต่ 3.00 ขึ้นไป จึงจะถือว่าข้อสอบจะมีความสอดคล้องในแนวเดียวกันกับตัวชี้วัด

3) ข้อเสนอแนะเพิ่มเติม เป็นส่วนที่ให้ผู้เชี่ยวชาญเขียนแสดงความคิดเห็น หรือข้อสังเกต หรือข้อเสนอแนะอื่นๆ เพิ่มเติมเกี่ยวกับข้อสอบแต่ละข้อ ตัวอย่างข้อรายการประเมินในแบบประเมินความสอดคล้องฯ แสดงดังภาพที่ 1

การตรวจสอบคุณภาพของเครื่องมือก่อนนำไปใช้จริง ทำโดยให้ผู้เชี่ยวชาญด้านมาตรฐานและตัวชี้วัดวิทยาศาสตร์ ชั้นมัธยมศึกษาตอนต้น และด้านการวัดและประเมินผลทางการศึกษา จำนวน 7 ท่าน ตรวจสอบความถูกต้องและความเหมาะสมของรูปแบบและภาษา ตลอดจนได้มีการทดลองใช้แบบประเมินความสอดคล้องฯ ดังกล่าว โดยผู้วิจัยทำการสุ่มข้อสอบจำนวน 40 ข้อ จากทั้งหมด 1,089 ข้อ ซึ่งเป็นการสุ่มข้อสอบ 1 ข้อ จากแต่ละตัวชี้วัด เพื่อจัดทำเป็นแบบประเมินความสอดคล้องฯ ฉบับทดลองใช้ และให้ผู้เชี่ยวชาญทั้ง 7 ท่าน ทดลองประเมินความสอดคล้องฯ ของข้อสอบ จากนั้นทำการวิเคราะห์ความเที่ยงระหว่างผู้ประเมิน (inter-rater reliability) เพื่อตรวจสอบความสอดคล้องกันระหว่างผู้ประเมิน โดยวิเคราะห์สถิติแคปปาของฟลีส (Fleiss's kappa statistic: K_p) (Fleiss, 1971) ในการประเมินระดับความซับซ้อนทางปัญญา เนื่องจากเป็นข้อมูลระดับนามบัญญัติ (Gisev, Bell, & Chen, 2013) ได้ค่า K_p เท่ากับ 0.596 แสดงว่ามีความสอดคล้องระหว่างผู้ประเมินอยู่ในระดับดี (Fleiss, Levin, & Paik, 2003) และใช้สถิติสหสัมพันธ์ภายในชั้น (intra-class correlation: ICC) ในการประเมินระดับความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด เนื่องจากเป็นข้อมูลต่อเนื่อง (continuous data) (Gisev, Bell, & Chen, 2013) ซึ่งจากการทดลองใช้แบบประเมินดังกล่าว พบว่า ค่าสัมประสิทธิ์สหสัมพันธ์ภายในชั้นที่วิเคราะห์จากโมเดลอิทธิพลผสมแบบสองทาง (two-way mixed-effects model) โดยมีผู้ประเมินจำนวน 7 คน ที่ระดับความเชื่อมั่น 95% มีค่าเท่ากับ 0.943 แสดงว่ามีความสอดคล้องระหว่างผู้ประเมินอยู่ในระดับดีมาก (Portney & Watkins, 2015)

ข้อที่ U3 2-4122.1-312	วัตถุ A น้ำหนัก 90 N ถูกแรงขนาด 50 N กดติดกับฝาผนังที่มีแรงเสียดทานดังรูป จงหาขนาดและทิศทางของแรงปฏิกิริยาในแนวตั้งฉากที่กระทำต่อวัตถุ A 1. 40 N ทิศขึ้นในแนวตั้ง 2. 50 N ทิศไปทางขวา 3. 90 N ทิศขึ้นในแนวตั้ง 4. 140 N ทิศไปทางขวา
คะแนนรายข้อ	0.50 คะแนน
ระดับชั้น	มัธยมศึกษาปีที่ 2
มาตรฐาน	ว 4.1 เข้าใจธรรมชาติของแรงแม่เหล็กไฟฟ้า แรงโน้มถ่วง และแรงนิวเคลียร์ มีกระบวนการสืบเสาะหาความรู้ สื่อสารสิ่งที่เรียนรู้และนำความรู้ไปใช้ประโยชน์ อย่างถูกต้องและมีคุณธรรม
ตัวชี้วัด	ม. 2/2 อธิบายแรงลัพธ์ที่กระทำต่อวัตถุที่หยุดนิ่งหรือวัตถุเคลื่อนที่ด้วยความเร็วคงตัว

1. ระดับความซับซ้อนทางปัญญา (เลือกได้เพียงระดับเดียว)

☐ จำ

☐ เข้าใจ

☐ ประยุกต์ใช้

☐ วิเคราะห์

☐ ประเมินค่า

☐ สร้างสรรค์

2. ระดับความสอดคล้องระหว่างข้อสอบกับตัวชี้วัด (เลือกได้เพียงระดับเดียว)

☐ ไม่สอดคล้อง

☐ ก่อนข้างไม่สอดคล้อง

☐ ไม่แน่ใจ

☐ ก่อนข้างสอดคล้อง

☐ สอดคล้องโดยตรง

(0) (1) (2) (3) (4)

3. ข้อเสนอแนะเพิ่มเติม (ถ้ามี)

.....

.....

ภาพที่ 1 ตัวอย่างข้อรายการประเมินในแบบประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัดฯ

4. การเก็บและรวบรวมข้อมูล

การประเมินความสอดคล้องระหว่างกรอบมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินผลการเรียนรู้ของโรงเรียน ดำเนินการโดยให้ผู้เชี่ยวชาญพิจารณาเพื่อประเมินความสอดคล้องระหว่างเนื้อหาของข้อสอบกับเนื้อหาของตัวชี้วัดที่โรงเรียนกำหนด มีดังนี้

1) จัดประชุมและฝึกปฏิบัติผู้เชี่ยวชาญ โดยมีจุดประสงค์เพื่อสร้างความเข้าใจที่ตรงกันระหว่างผู้เชี่ยวชาญในเรื่องต่างๆ เช่น การทบทวนนิยามของระดับความซับซ้อนทางปัญญา ตามแนวคิด revised Bloom’s taxonomy (Anderson et al, 2001) ตลอดจนชี้แจงแนวทางในการดำเนินงานของผู้เชี่ยวชาญ โดยในการประชุมได้มีการให้ผู้เชี่ยวชาญทดลอง ฝึกปฏิบัติกระบวนการจั่นทามติ และฝึกประเมินความสอดคล้องระหว่างข้อสอบกับตัวชี้วัดเป็นรายบุคคล ตลอดจนเปิดโอกาสให้ผู้เชี่ยวชาญได้อภิปรายในภายในกลุ่มเพื่อให้ได้ผลการพิจารณาที่ถูกต้องและเป็นที่เข้าใจตรงกัน

2) หลังจากการประชุมผู้เชี่ยวชาญแล้ว ผู้เชี่ยวชาญแต่ละคนจะทำหน้าที่ประเมินความสอดคล้องระหว่างตัวชี้วัดกับข้อสอบได้อย่างอิสระ โดยใช้แบบประเมินความสอดคล้องระหว่างตัวชี้วัดกับข้อสอบฯ และเอกสารคู่มือ

ในการดำเนินงานสำหรับผู้เชี่ยวชาญที่ผู้วิจัยจัดทำขึ้น เนื่องจากข้อสอบที่ใช้ในการศึกษาครั้งนี้มีจำนวนมาก คือ 1,089 ข้อ ดังนั้น ผู้วิจัยได้ออกแบบการวิจัยที่ผู้ประเมินแต่ละคนไม่ต้องประเมินข้อสอบทุกข้อ และข้อสอบแต่ละข้ออาจถูกประเมินด้วยผู้เชี่ยวชาญที่ไม่เหมือนกัน โดยผู้เชี่ยวชาญแต่ละคนจะประเมินข้อสอบคนละประมาณ 287 - 312 ข้อ ซึ่งผู้เชี่ยวชาญสามารถเลือกดำเนินการประเมินได้ 2 รูปแบบ คือ รูปแบบเอกสาร หรือรูปแบบออนไลน์ ซึ่งขึ้นอยู่กับความสะดวกของผู้เชี่ยวชาญแต่ละท่าน ทั้งนี้ข้อสอบที่ใช้ในการศึกษาทั้งหมดจะถูกสุ่มเข้าสู่ชุดข้อสอบย่อย (subset) จำนวน 25 ชุดๆ ละประมาณ 41 – 46 ข้อ ผู้เชี่ยวชาญแต่ละคนจะได้ประเมินข้อสอบอย่างสุ่ม จำนวน 287 - 312 ข้อ (7 ชุดย่อย) ทั้งนี้ ข้อสอบแต่ละข้อจะถูกประเมินโดยผู้เชี่ยวชาญ จำนวน 5 คน แต่จะมีข้อสอบ 1 ชุดย่อย (จำนวน 40 ข้อ) ที่สุ่มมาจากตัวชี้วัดละ 1 ข้อ จะถูกประเมินโดยผู้เชี่ยวชาญทุกคน (20 คน) เพื่อนำผลมาใช้ในการศึกษาสัมประสิทธิ์การสรุปอ้างอิงของคะแนนการประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบตัวชี้วัดวิทยาศาสตร์ เมื่อจำนวนผู้ประเมินและรูปแบบการออกแบบการประเมินต่างกัน โดยใช้ทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัด (G-theory) ตลอดจนใช้ในการวิเคราะห์ความเที่ยงระหว่างผู้ประเมิน (inter-rater reliability) โดยพบว่า การประเมินระดับความซับซ้อนทางปัญญาปัญญา มีค่า K_r เท่ากับ 0.510 แสดงว่ามีความสอดคล้องระหว่างผู้ประเมินอยู่ในระดับดีมาก (Fleiss, Levin, & Paik, 2003) และการประเมินระดับความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด มี ค่า ICC ที่วิเคราะห์จากโมเดลอิทธิพลผสมแบบสองทาง (two-way mixed-effects model) โดยมีผู้ประเมินจำนวน 20 คน ที่ระดับความเชื่อมั่น 95% เท่ากับ 0.954 แสดงว่ามีความสอดคล้องระหว่างผู้ประเมินอยู่ในระดับดีมาก (Portney & Watkins, 2015)

5. การวิเคราะห์ข้อมูล

การประมาณค่าและเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิงของผลการประเมินความสอดคล้องในแนวเดียวกัน โดยใช้ข้อมูลจากแบบประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด เมื่อจำนวนผู้ตรวจและรูปแบบการประเมินต่างกัน โดยใช้การวิเคราะห์ความแปรปรวนแบบสองทาง (two-way ANOVA) มีขั้นตอนการดำเนินการดังต่อไปนี้

1. นำผลการประเมินความสอดคล้องในแนวเดียวกันฯ ของข้อสอบจำนวน 40 ข้อ มาจัดกระทำตามเงื่อนไขจำนวนผู้ประเมิน ตั้งแต่ 1 – 20 คน ตามรูปแบบการประเมิน 2 รูปแบบ ดังนี้

- 1) ผู้ประเมินแต่ละคนประเมินข้อสอบทุกข้อ (i x r design)
- 2) ผู้ประเมินแต่ละคนประเมินข้อสอบบางข้อ หรือที่เรียกว่า ผู้ประเมินสอดแทรก (nested) ภายในข้อสอบ (r : i design)

2. ทำการวิเคราะห์หองค์ประกอบความแปรปรวน โดยใช้โมเดลการสรุปอ้างอิง เมื่อจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกัน ซึ่งกำหนดให้ข้อสอบ (item: i) เป็นสิ่งที่ถูกประเมิน และผู้ประเมิน (rater: r) เป็นแหล่งความแปรปรวน ใช้การวิเคราะห์ความแปรปรวนแบบสองทาง (two-way ANOVA) ของแต่ละรูปแบบการประเมิน โดยจำแนก ผลการวิเคราะห์ตามลักษณะของแบบประเมินความสอดคล้องฯ แบ่งเป็น 2 ส่วน ได้แก่ การประเมินระดับความซับซ้อนทางปัญญา และการประเมินระดับความสอดคล้องในแนวเดียวกันฯ ด้วยมาตรประมาณค่า 5 ระดับ

3. ประเมินค่าและเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิงสำหรับการตัดสินใจเชิงสัมบูรณ์ (ρ^2_{Abs}) โดยจำแนกใน 2 ส่วน คือ ส่วนของการประเมินความซับซ้อนทางปัญญา และส่วนของการประเมินระดับความสอดคล้องในแนวเดียวกันฯ ด้วยมาตรประมาณค่า 5 ระดับ ตามเงื่อนไขดังนี้

3.1 เมื่อจำนวนผู้ประเมินเท่ากัน แต่มีรูปแบบการออกแบบการประเมินที่แตกต่างกัน

3.2 เมื่อรูปแบบการออกแบบการประเมินเหมือนกัน แต่จำนวนผู้เชี่ยวชาญต่างกัน

สูตรในการคำนวณสัมประสิทธิ์การสรุปอ้างอิงสำหรับการตัดสินใจเชิงสัมบูรณ์ (ρ^2_{Abs}) (Shavelson & Webb, 1991) มีดังนี้

$$\rho^2_{Abs} = \frac{\sigma_i^2}{\sigma_i^2 + \sigma_{Abs}^2}$$

σ_i^2 คือ ความแปรปรวนของคะแนนความสอดคล้องในแนวเดียวกันของข้อสอบ

σ_{Abs}^2 คือ ความแปรปรวนของความคลาดเคลื่อนแบบสัมบูรณ์ (absolute error variance)

ทั้งนี้ กำหนดเกณฑ์ความน่าเชื่อถือของการประเมินในระดับที่ยอมรับได้ คือ มีค่าสัมประสิทธิ์การสรุปอ้างอิงสำหรับตัดสินใจเชิงสัมบูรณ์ (ρ^2_{Abs}) ตั้งแต่ 0.70 ขึ้นไป (Nunnally, 1987)

ผลการวิจัย

1. ผลการวิเคราะห์ขนาดขององค์ประกอบความแปรปรวนของคะแนนการประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบตัวชี้วัดกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับชั้นมัธยมศึกษาตอนต้น พบว่า ทั้งการประเมินระดับความซับซ้อนทางปัญญา และการประเมินระดับความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด ด้วยมาตรประมาณค่า 5 ระดับ มีผลเป็นไปในทิศทางเดียวกัน กล่าวคือ ความแปรปรวนจากแหล่งต่างๆ เมื่อเทียบกับความแปรปรวนรวม มีค่าเรียงลำดับจากมากไปหาน้อย ดังนี้ ความแปรปรวนของคะแนนความสอดคล้องในแนวเดียวกันของข้อสอบ (σ_i^2) (ร้อยละ 54.72 และ 51.12 ตามลำดับ) รองลงมาคือ ความแปรปรวนของความคลาดเคลื่อนหรือส่วนเหลือ ($\sigma_{ir,c}^2$) (ร้อยละ 42.44 และ 36.36 ตามลำดับ) ซึ่งเป็นความแปรปรวนอันเนื่องมาจากความแตกต่างของความสอดคล้องในแนวเดียวกันของข้อสอบแต่ละข้อที่มีได้รับการประเมินด้วยผู้ประเมินที่แตกต่างกัน และ/หรือเป็นแหล่งความแปรปรวนที่ไม่สามารถวัดได้ และความแปรปรวนอันเกิดจากผู้ประเมิน (σ_r^2) (ร้อยละ 2.84 และ 12.52 ตามลำดับ) ตามลำดับ รายละเอียดดังตารางที่ 1

ตารางที่ 1 G-Study สำหรับประมาณค่าองค์ประกอบความแปรปรวนของการวัดในการประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด

แหล่งความแปรปรวน	การประเมินระดับความซับซ้อนทางปัญญา					การประเมินระดับความสอดคล้อง ¹ ด้วยมาตรประมาณค่า				
	SS	df	MS	σ^2	% of Total σ^2	SS	df	MS	σ^2	% of Total σ^2
Items (i)	390.639	39	10.016	0.482	54.72	200.899	39	5.151	0.249	51.12
Rater (r)	26.364	19	1.388	0.025	2.84	49.524	19	2.607	0.061	12.52
Residual (ir,e)	276.986	741	0.374	0.374	42.44	131.226	741	0.177	0.177	36.36
Total	693.989	799		0.881	100.00	381.649	799		0.487	100.00

2. ผลการเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนการประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัดกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับชั้นมัธยมศึกษาตอนต้น เมื่อจำนวนผู้ประเมินและรูปแบบการออกแบบการประเมินต่างกัน พบว่า ค่าสัมประสิทธิ์การสรุปอ้างอิงสำหรับตัดสินใจเชิงสัมบูรณ์ (ρ^2_{Abs}) ของคะแนน การประเมินความสอดคล้องในแนวเดียวกัน⁴ มีค่าเพิ่มขึ้นเมื่อจำนวนผู้ประเมินเพิ่มขึ้น ในทุกรูปแบบการออกแบบการประเมิน ทั้งการประเมินระดับความซับซ้อนทางปัญญา และการประเมินระดับความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด ด้วยมาตรประมาณค่า 5 ระดับ รายละเอียดดังตารางที่ 2

3. ผลการพิจารณาจำนวนผู้ประเมินที่ทำให้เกิดความน่าเชื่อถือของการประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด⁴ ในระดับที่ยอมรับได้ โดยการพิจารณาค่าสัมประสิทธิ์การสรุปอ้างอิงสำหรับตัดสินใจเชิงสัมบูรณ์ (ρ^2_{Abs}) ตั้งแต่ 0.70 ขึ้นไป (Nunnally, 1987) ในการประเมินระดับความซับซ้อนทางปัญญา พบว่า ผู้เชี่ยวชาญ (n_r) จำนวน 2 คน สามารถให้ผลการตัดสินที่มีความน่าเชื่อถือในระดับที่ยอมรับได้ ในทุกรูปแบบการออกแบบการประเมิน โดยให้ค่า ρ^2_{Abs} ตั้งแต่ 0.71 ขึ้นไป และพบว่า การประเมินระดับความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด ด้วยมาตรประมาณค่า 5 ระดับ ต้องการผู้ประเมินจำนวน ตั้งแต่ 3 คนขึ้นไป จึงจะให้ผลการตัดสินที่มีความน่าเชื่อถือในระดับที่ยอมรับได้ ในทุกรูปแบบการออกแบบการประเมิน โดยให้ค่า ρ^2_{Abs} ตั้งแต่ 0.76 ขึ้นไป นอกจากนี้ ยังพบว่า ทั้ง 2 รูปแบบการออกแบบการประเมิน คือ รูปแบบที่ 1 ผู้ประเมินแต่ละคนประเมินข้อสอบทุกข้อ และรูปแบบที่ 2 ผู้ประเมินแต่ละคนประเมินข้อสอบบางข้อ ให้ค่าสัมประสิทธิ์การสรุปอ้างอิงที่ใกล้เคียงกัน เมื่อจำนวนผู้ประเมินเท่ากัน รายละเอียดดังตารางที่ 2

ตารางที่ 2 D-Study ของการประเมินในส่วนของการประเมินระดับความสอดคล้องในแนวเดียวกันฯ

Design	n _r =	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
การประเมินระดับความซับซ้อนทางปัญญา																					
i x r	ρ^2_{Abs}	0.547	0.707	0.784	0.830	0.858	0.880	0.894	0.906	0.915	0.923	0.931	0.936	0.940	0.943	0.947	0.951	0.954	0.956	0.958	0.960
r : i	ρ^2_{Abs}	0.547	0.707	0.784	0.829	0.858	0.879	0.894	0.906	0.916	0.924	0.930	0.935	0.940	0.944	0.948	0.951	0.954	0.956	0.958	0.960
การประเมินระดับความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด ด้วยมาตรประมาณค่า 5 ระดับ																					
i x r	ρ^2_{Abs}	0.511	0.675	0.759	0.808	0.841	0.862	0.880	0.892	0.902	0.912	0.919	0.926	0.929	0.936	0.940	0.943	0.947	0.950	0.954	0.954
r : i	ρ^2_{Abs}	0.511	0.677	0.758	0.807	0.839	0.863	0.880	0.893	0.904	0.913	0.920	0.926	0.931	0.936	0.940	0.944	0.947	0.950	0.952	0.954

อภิปรายผล

การอภิปรายผลการวิจัย ประกอบด้วย 3 ประการ มีรายละเอียดดังต่อไปนี้

1. จากผลการวิจัยที่พบว่า ความแปรปรวนของคะแนนความสอดคล้องในแนวเดียวกันของข้อสอบมีค่ามากที่สุด เมื่อเทียบกับความแปรปรวนรวม ทั้งการประเมินระดับความซับซ้อนทางปัญญา และการประเมินระดับความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด ด้วยมาตรประมาณค่า 5 ระดับ แสดงว่าความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัดของข้อสอบแต่ละข้อมีความแตกต่างกัน อาจเนื่องมาจากข้อสอบแต่ละข้อได้รับการสุ่มจากโรงเรียนที่แตกต่างกัน ซึ่งมีการบริหารจัดการด้านการวัดและประเมินผลแตกต่างกัน และ/หรือ มีคะแนนผลสัมฤทธิ์ทางการเรียนด้านวิทยาศาสตร์ที่พิจารณาจากคะแนนการทดสอบทางการศึกษาระดับชาตินั้นพื้นฐาน (O-NET) แตกต่างกัน ส่วนความแปรปรวนอันเกิดจากผู้ประเมินมีค่าน้อยที่สุด เมื่อเทียบกับความแปรปรวนรวม แสดงให้เห็นว่าผู้ประเมินแต่ละคนมีการประเมินที่สอดคล้องกัน หรือมีความแตกต่างกันในการประเมินค่อนข้างน้อย ซึ่งสอดคล้องกับงานวิจัยของกนกวรรณ เอี่ยมชัย (2539) ที่พบว่า ผู้ตรวจเป็นแหล่งความแปรปรวนในการตรวจให้คะแนนแบบทดสอบความเรียงประยุกต์น้อยที่สุด สอดคล้องกับงานวิจัยของสมคิด เทียรพิสุทธ์ (2550) ที่พบว่า ความแปรปรวนของผู้ตรวจมีค่าน้อยมาก และสอดคล้องกับงานวิจัยของจิรายุ เถาว์โท อนุ เจริญวงศ์ระยัย และปณณวิชญ์ ใบกุหลาบ (2559) ซึ่งพบว่า ความแปรปรวนของผู้ตรวจให้คะแนนแบบทดสอบอัตนัยมีค่าน้อยที่สุด ทั้งนี้เนื่องจากงานวิจัยนี้ได้มีการคัดเลือกผู้เชี่ยวชาญที่ทำหน้าที่ประเมินที่มีคุณสมบัติเหมาะสม มีความรู้ความสามารถในด้านตัวชี้วัดและการสอนวิทยาศาสตร์ระดับชั้นมัธยมศึกษาตอนต้น อีกทั้งเป็นผู้ที่สำเร็จการศึกษาน้อยในระดับปริญญาตรีวิชาเอกด้านการสอนวิทยาศาสตร์ และสำเร็จการศึกษาน้อยในระดับปริญญาโททางการวัดและประเมินทางการศึกษา และผู้เชี่ยวชาญส่วนใหญ่มีประสบการณ์ด้านการสอนวิทยาศาสตร์ไม่ต่ำกว่า 6 ปี นอกจากนี้ยังมีการจัดประชุมและฝึกปฏิบัติก่อนทำหน้าที่ประเมินจริง เพื่อชี้แจงแนวทางการดำเนินงาน และทำความเข้าใจร่วมกับเกี่ยวกับรายละเอียดของเครื่องมือในการประเมินความสอดคล้องฯ ตลอดจนมีคู่มือสำหรับการดำเนินงานสำหรับผู้ประเมินที่ช่วยเพิ่มความชัดเจนของแนวทางในการดำเนินงาน ซึ่งสอดคล้องกับ Webb (1999) ที่ให้ความสำคัญกับการฝึกปฏิบัติของผู้เชี่ยวชาญก่อนดำเนินการประเมินความสอดคล้องในแนวเดียวกันฯ เพื่อให้ผู้เชี่ยวชาญทุกคน

มีความเข้าใจถึงกระบวนการประเมินความสอดคล้องในแนวเดียวกัน หลักการจำแนกระดับความซับซ้อนทางปัญญา ตลอดจนการใช้ดุลพินิจในการตัดสินความสอดคล้องๆ ได้อย่างถูกต้องและปฏิบัติไปในทิศทางเดียวกัน และสอดคล้องกับ Patz, Junker, Johnson, and Mariano (2002) และ Welk, Schaben, and Morrow (2004) ที่กล่าวว่า ความคลาดเคลื่อนของการประเมินที่มาจากผู้ประเมินจะมีค่าน้อยมาก เมื่อมีเครื่องมือการประเมินที่กำหนดและอธิบาย รายละเอียดของรายการประเมิน และมีเกณฑ์การประเมินที่ชัดเจน

2. จากผลการเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนการประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัดกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับชั้นมัธยมศึกษาตอนต้น เมื่อจำนวนผู้ประเมินและรูปแบบการออกแบบการประเมินต่างกัน พบว่า ค่าสัมประสิทธิ์การสรุปอ้างอิงสำหรับตัดสินใจเชิงสัมบูรณ์ (ρ^2_{Abs}) ของคะแนนการประเมินความสอดคล้องในแนวเดียวกันๆ มีค่าเพิ่มขึ้นเมื่อจำนวนผู้ประเมินเพิ่มขึ้น ในทุกรูปแบบการออกแบบการประเมิน ทั้งการประเมินระดับความซับซ้อนทางปัญญา และการประเมินระดับ ความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด ด้วยมาตรประมาณค่า 5 ระดับ โดยค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนการประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด เมื่อมีจำนวนผู้ประเมินที่มากกว่า จะมีค่าความเที่ยงสูงกว่าจำนวนผู้เชี่ยวชาญที่น้อยกว่า ในทุกรูปแบบการออกแบบการประเมิน ซึ่งเป็นไปตามแนวคิดของ Herman, Webb, & Zuniga (2007) Lombardi, Sebuén, Conley, and Snow (2010) Porter and Smithson (2001) และ Webb (2007) ที่กล่าวว่า จำนวนผู้เชี่ยวชาญในการประเมินความสอดคล้องในแนวเดียวกัน ยิ่งมาก จะทำให้ผลการประเมินยิ่งมีความเที่ยงเพิ่มมากขึ้น สอดคล้องกับงานวิจัยที่เป็นการศึกษาความเที่ยงในการตรวจประเมินที่พบว่า การเพิ่มจำนวนผู้ตรวจประเมินจะทำค่าสัมประสิทธิ์การสรุปอ้างอิงความน่าเชื่อถือของผลการวัดเพิ่มขึ้น เช่น งานวิจัยของวัชรียา ศรีวลีรัตน์ (2558) ที่พบว่า หากเพิ่มจำนวนผู้ตรวจ ค่าสัมประสิทธิ์การสรุปอ้างอิงความน่าเชื่อถือของ ผลการวัดจะเพิ่มขึ้น และค่าความแปรปรวนของความคลาดเคลื่อนลดลงอย่างมีนัยสำคัญทางสถิติ เช่นเดียวกับงานวิจัยของดวงใจ สีเขียว (2549) ที่พบว่า คะแนนการประเมินผลสรุปรวม/สอบสอนของนิสิต/นักศึกษาฝึกประสบการณ์วิชาชีพครูที่มีจำนวนผู้ประมาณเกินมากกว่าจะมีค่าความเที่ยงสูงกว่ากรณีที่มีผู้ประเมินน้อยกว่า ในทุกสถานการณ์การประเมิน และงานวิจัยของอังคณา กุลนาคล (2557) ที่พบว่า ค่าสัมประสิทธิ์การสรุปอ้างอิงความน่าเชื่อถือของ ผลการวัดของคะแนนผังมโนทัศน์วิชาการวิจัยทางการศึกษาที่มีจำนวนผู้ตรวจมากกว่า จะมีค่าสูงกว่าจำนวนผู้ตรวจน้อยกว่า ในทุกรูปแบบการตรวจ

นอกจากนี้ ผลการวิจัยพบว่า รูปแบบการออกแบบการประเมิน 2 รูปแบบ คือ รูปแบบที่ 1 ผู้ประเมินทำการประเมินข้อสอบทุกข้อ และรูปแบบที่ 2 ผู้ประเมินทำการประเมินข้อสอบบางข้อ ให้ค่าสัมประสิทธิ์การสรุปอ้างอิงที่ใกล้เคียงกัน เมื่อจำนวนผู้ประเมินเท่ากัน แสดงให้เห็นว่าในการประเมินความสอดคล้องในแนวเดียวกัน ผู้ประเมินแต่ละคนไม่จำเป็นต้องประเมินข้อสอบทุกข้อก็สามารถได้ผลการประเมินที่มีความน่าเชื่อถือใกล้เคียงกับการศึกษา โดยให้ผู้ประเมินแต่ละคนประเมินข้อสอบทุกข้อ ซึ่งเหมาะกับบริบทการศึกษาในกรณีที่มีข้อสอบที่ต้องประเมินจำนวนมาก ที่อาจต้องใช้ผู้เชี่ยวชาญในการประเมินหลายคน การออกแบบการประเมินโดยผู้ประเมินแต่ละคนประเมินข้อสอบบางข้อจะช่วยประหยัดเวลา แรงงาน และงบประมาณในการศึกษา ตลอดจนมีความเป็นไปได้ในการปฏิบัติ พร้อมทั้งยังให้ผลการศึกษาที่มีความน่าเชื่อถือ

3. ผลการพิจารณาจำนวนผู้ประเมินที่ทำให้เกิดความน่าเชื่อถือของการประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัดฯ ในระดับที่ยอมรับได้ โดยการพิจารณาค่าสัมประสิทธิ์การสรุปอ้างอิงสำหรับตัดสินใจเชิงสัมบูรณ์ (ρ^2_{Abs}) ตั้งแต่ 0.70 ขึ้นไป (Nunnally, 1987) พบว่า ในส่วนของการประเมินระดับความซับซ้อนทางปัญญา จำนวนผู้ประเมิน (n_r) 2 คน สามารถให้ผลการตัดสินใจที่มีความน่าเชื่อถืออยู่ในระดับที่ยอมรับได้ในทุกรูปแบบการออกแบบการประเมิน ขณะที่ในส่วนของการประเมินระดับความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด ที่มีลักษณะเป็นมาตรฐานค่า 5 ระดับ ต้องการผู้ประเมินจำนวน ตั้งแต่ 3 คนขึ้นไป จึงจะให้ผลการตัดสินใจที่มีความน่าเชื่อถืออยู่ในระดับที่ยอมรับได้ ในทุกรูปแบบการออกแบบการประเมิน การที่การประเมินระดับความซับซ้อนทางปัญญาดำเนินการจำนวนผู้ประเมินมากกว่าการประเมินระดับความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด ที่มีลักษณะเป็นมาตรฐานค่า 5 ระดับ เป็นไปตามแนวคิดทฤษฎี ทั้งนี้ความแตกต่างนี้อาจเกิดจากคำอธิบายที่ใช้ในการประเมิน โดยในการประเมินระดับความซับซ้อนทางปัญญาตามแนวคิดของ revised Bloom's taxonomy ที่ประกอบด้วย 6 ระดับ ผู้ประเมินสามารถเข้าใจคำอธิบายและตัวอย่างของพฤติกรรมการเรียนรู้ในแต่ละระดับ เพื่อพิจารณาคัดสินระดับความซับซ้อนทางปัญญาของข้อสอบแต่ละข้อได้ดีกว่าการประเมินในส่วนของความสอดคล้องระหว่างข้อสอบกับตัวชี้วัดที่เป็นมาตรฐานค่า 5 ระดับ ที่ผู้ประเมินมีโอกาสที่ต้องใช้การตีความ และมีความเป็นปรนัยน้อยกว่า การประเมินตัดสินแต่ละระดับของความซับซ้อนทางปัญญา ซึ่งสอดคล้องกับงานวิจัยของ Lombardi, Sebuen, Conley, & Snow (2010) ที่พบว่า ในการศึกษาความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับมาตรฐานวิชาคณิตศาสตร์และภาษาอังกฤษ ในส่วนการประเมินระดับความซับซ้อนทางปัญญามีค่าความเที่ยงมากกว่าการประเมินด้วยมาตรฐานค่า 3 ระดับ เพื่อประเมินระดับความรู้ที่ผู้สอบต้องใช้ในการทำข้อสอบแต่ละข้อ ที่เรียกว่า rigor scale (โดยที่ 1 = ต่ำกว่าระดับความรู้ที่ผู้สอบควรมีในการสอบเข้ามหาวิทยาลัย, 2 = เท่ากับระดับความรู้ที่ผู้สอบควรมีในการสอบเข้ามหาวิทยาลัย 3 = สูงกว่าระดับความรู้ที่ผู้สอบควรมีในการสอบเข้ามหาวิทยาลัย) ทั้งนี้ ในภาพรวมพบว่า มาตรฐานประเมินความซับซ้อนทางปัญญา มีคำอธิบายแต่ละระดับที่ชัดเจนมากกว่า จึงทำให้ง่ายต่อการประเมินและมีความเป็นปรนัยในการประเมิน

ข้อเสนอแนะ

ข้อเสนอแนะในการนำไปใช้

1. ออกแบบการประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด โดยใช้รูปแบบที่ผู้ประเมิน แต่ละคนประเมินข้อสอบบางข้อ โดยเฉพาะในกรณีที่ในการศึกษามีข้อสอบจำนวนมาก ซึ่งสามารถช่วยประหยัดเวลาและงบประมาณ ตลอดจนมีความเป็นไปได้ในการบริหารจัดการและช่วยลดความซับซ้อนในการดำเนินงาน แต่ยังสามารถให้ค่าความเที่ยงที่ใกล้เคียงกับการประเมินรูปแบบที่ผู้ประเมินแต่ละคนประเมินข้อสอบทุกข้อ

2. ออกแบบการประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัดในระดับที่น่าเชื่อถือ โดยออกแบบการศึกษาที่มีผู้ประเมินอย่างน้อย 2 คน ในกรณีที่เป็นการประเมินระดับความซับซ้อนทางปัญญา และมีผู้ประเมินอย่างน้อย 3 คน ในกรณีที่ประเมินระดับความสอดคล้องฯ โดยใช้มาตรฐานค่า

ข้อเสนอแนะในการวิจัยครั้งต่อไป

1. เนื่องจากงานวิจัยนี้ เป็นการประเมินความสอดคล้องในแนวเดียวกันๆ ของข้อสอบที่ใช้ประเมินระดับชั้นเรียน โดยใช้เกณฑ์ความน่าเชื่อถือของการวัด ที่มีค่าความเที่ยงตั้งแต่ 0.70 ขึ้นไป ของ Nunnally (1987) ทั้งนี้ หากเป็นการศึกษาความสอดคล้องในแนวเดียวกันของข้อสอบในระดับที่สูงกว่าข้อสอบที่ประเมินระดับชั้นเรียน เช่น ข้อสอบประเมินระดับชาติ อาจใช้เกณฑ์ในการตัดสินความน่าเชื่อถือของผลการวัดที่มีความเข้มงวดมากขึ้น เช่น เกณฑ์ค่าความเที่ยง 0.80 (Sattler, 1988) หรือเกณฑ์ค่าความเที่ยง 0.90 (Gay, 1992)

2. ควรมีการศึกษาแหล่งความแปรปรวนอื่นๆ นอกเหนือจากจำนวนผู้ประเมินและรูปแบบการออกแบบประเมิน เช่น มาตรฐานการเรียนรู้ อิทธิพลของผู้ประเมิน (rater effects) และระดับผลสัมฤทธิ์ทางการเรียนของโรงเรียน เป็นต้น

บรรณานุกรม

- กนกวรรณ เอี่ยมชัย. (2539). *การศึกษาค่าความเชื่อมั่นของแบบทดสอบ เอ็ม อี คิว วิชาแนวคิดพื้นฐานและหลักการพยาบาลที่ตรวจให้คะแนนต่างกัน* (วิทยานิพนธ์ปริญญาการศึกษามหาบัณฑิต ไม่ได้ตีพิมพ์). สาขาวิชาการวัดผลการศึกษา มหาวิทยาลัยศรีนครินทรวิโรฒ, กรุงเทพมหานคร.
- จิรายุ เถาว์โท อนุ เจริญวงศ์ระยับ และปณณวิชญ์ ใบกุหลาบ. (2559). การศึกษาค่าความเชื่อมั่นของคะแนนแบบทดสอบอัตนัยวิชาคณิตศาสตร์ของ นักเรียนชั้นมัธยมศึกษาปีที่ 2 ที่มีจำนวนผู้ตรวจและรูปแบบการตรวจให้คะแนนต่างกันโดยใช้ทฤษฎีการสรุปอ้างอิง, *วารสารหาดใหญ่วิชาการ*, 14(1), 1-14.
- ดวงใจ สีเขียว. (2549). *การพัฒนาระบบการประเมินนิสิตนักศึกษาฝึกประสบการณ์วิชาชีพครู ตามแนวทางการประเมินแบบ 360 องศา โดยการประยุกต์ใช้ทฤษฎีการสรุปอ้างอิง* (วิทยานิพนธ์ปริญญาครุศาสตรดุษฎีบัณฑิต ไม่ได้ตีพิมพ์). สาขาการวัดและประเมินผลการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพมหานคร.
- วัชรียา ศรีวลีรัตน์. (2558). การเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิงความน่าเชื่อถือของผลการวัดของแบบทดสอบเรียงความวิชาภาษาไทย ชั้นมัธยมศึกษาปีที่ 6 ที่มีจำนวนข้อสอบและจำนวนผู้ตรวจต่างกัน. *วารสารวิจัยและพัฒนาหลักสูตร*, 5(2), 14-19.
- ศิริชัย กาญจนวาสี. (2555). *ทฤษฎีการทดสอบแนวใหม่*. กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.
- สมคิด เทียรพิสุทธิ์. (2550). *การเปรียบเทียบค่าความเชื่อมั่นของแบบทดสอบความเรียงประยุกต์วิชาคณิตศาสตร์ที่มีวิธีการตรวจและจำนวนผู้ตรวจต่างกัน*. (วิทยานิพนธ์ปริญญาการศึกษามหาบัณฑิต ไม่ได้ตีพิมพ์). สาขาวิชาการวัดผลการศึกษา มหาวิทยาลัยทักษิณ, สงขลา.
- อังคณา กุลนาคดล. (2557). การเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิงของคะแนนแผนผังโนทัศน์ วิชาการวิจัยทางการศึกษา เมื่อรูปแบบการตรวจและจำนวนผู้ตรวจต่างกัน. *วารสารศึกษาศาสตร์*, 25(3), 39-50.
- Ananda, S. (2003). *Rethinking issues of alignment under "No Child Left Behind."* Knowledge Brief. San Francisco: WestEd.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., . . .

Wittrock, M. C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives* (Complete ed.). New York: Longman.

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice*, 22(3), 21-29.

Biggs, J. (2003). *Teaching for quality learning at university*. Glasgow: The Society for Research into Higher Education & Open University Press.

Case, B., Jorgensen, M. A., & Zucker, S. (2004). *Alignment in educational assessment*. San Antonio: Harcourt Assessment.

Davis-Becker, S. L., & Buckendahl, C. W. (2013). A proposed framework for evaluating alignment studies. *Educational Measurement: Issues and Practice*, 32(1), 23-33.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378-382. doi:10.1037/h0031619

Fleiss, J. L., Levin, B. A., & Paik, M. C. (2003). *Statistical methods for rates and proportions*. Hoboken, NJ: John Wiley & Sons, Inc.

Gay, L. R. (1992). *Educational research competencies for analysis and application*. New York: Macmillan Publishing.

Gisev, N., Bell, J. S. and Chen, T. F. (2013). Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, 9, 330–338.

Herman, J. L., Webb, N. M., & Zuniga, S. A. (2007). Measurement issues in the alignment of standards and assessments. *Applied Measurement in Education*, 20(1), 101-126. doi:10.1080/08957340709336732

La Marca, P. M. (2001). Alignment of standards and assessments as an accountability criterion. *Practical Assessment, Research & Evaluation*, 7(21).

La Marca, P. M., Redfield, D., Winter, P. C., Bailey, A., & Hansche, D. (2000). *State standards and state assessment systems: A guide to alignment*. Washington: Council of Chief State School Officers.

Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher*, 23(9), 4-14.

Lombardi, A., Seburn, M., Conley, D., & Snow, E. (2010). *A generalizability investigation of cognitive demand and rigor ratings of items and standards in an alignment study*. Paper presented at the Annual Meeting of the American Educational Research Association, Denver, CO.

Nunnally, J. C. (1987). *Psychometric theory*. 2nded. New York, NY: McGraw-Hill.

- Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, 27(4), 341-384.
- Porter, A. C., & Smithson, J. L. (2001). *Defining, developing, and using curriculum indicators*. Consortium for Policy Research in Education (CPRE) Research Report Series RR-048. Philadelphia: CPRE, University of Pennsylvania Graduate School of Education.
- Porter, A. C., & Smithson, J. L. (2002). *Alignment of assessments, standards and instruction using curriculum indicator data*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.
- Portney, L. G., & Watkins, M. P. (2015). *Foundations of clinical research: Applications to practice*. 3rd ed. Philadelphia, PA: F. A. Davis Company.
- Rothman, R. (2003). *Imperfect matches: The alignment of standards and tests*. Paper commissioned by the Committee on Test Design for K-12 Science Achievement.
- Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). *Benchmarking and alignment of standards and testing (CSE Technical Report No. CSE-TR-566)*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.
- Sattler, J. M. (1988). *Assessment of children*. 3rd ed. San Diego, CA: J.M. Sattler.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.
- Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (Research Monograph No. 6). Washington, DC: Council of Chief State School Officers.
- Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states* (Research Monograph No. 18). Washington, DC: Council of Chief State School Officers.
- Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education*, 20, 7-25.
- Welk, G. J., Schaben, J. A., & Morrow, J. J. (2004). Reliability of accelerometry-based activity monitors: a generalizability study. *Medicine and science in sports and exercise*, 36(9), 1637-1645.