Journal of Education Studies

Volume 48
Issue 3 *July - September 2020*

Article 9

1-1-2020

Inter-rater Reliability of Alignment between Science Items and Indices(ความเที่ยงระหว่างผู้ประเมินของความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัดวิทยาศาสตร์)

Budsayarat Janprasert

Nattaporn Lawthong

Sungworn Ngudgratoke

Follow this and additional works at: https://digital.car.chula.ac.th/educujournal

 Part of the Education Commons

Recommended Citation

Janprasert, Budsayarat; Lawthong, Nattaporn; and Ngudgratoke, Sungworn (2020) "Inter-rater Reliability of Alignment between Science Items and Indices(ความเที่ยงระหว่างผู้ประเมินของความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัดวิทยาศาสตร์)," *Journal of Education Studies*: Vol. 48: Iss. 3, Article 9.
Available at: https://digital.car.chula.ac.th/educujournal/vol48/iss3/9

This Article is brought to you for free and open access by Chula Digital Collections. It has been accepted for inclusion in Journal of Education Studies by an authorized editor of Chula Digital Collections. For more information, please contact ChulaDC@car.chula.ac.th.

Journal of Education Studies
วารสารครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ปีที่ 48 ฉบับที่ 3 (กรกฎาคม - กันยายน 2563) หน้า 144-163

ความเที่ยงระหว่างผู้ประเมินของความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัดวิทยาศาสตร์

Inter-rater Reliability of Alignment between Science Items and Indices

บุษยารัตน์ จันทร์ประเสริฐ[1] ณัฏฐภรณ์ หลาวทอง[2] และ สังวรณ์ งัดกระโทก[3]
Budsayarat Janprasert[1] Nattaporn Lawthong[2] and Sungworn Ngudgratoke[3]

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อ 1) ตรวจสอบความเที่ยงระหว่างผู้ประเมินของความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัดวิทยาศาสตร์ ระดับชั้นมัธยมศึกษาตอนต้น และ 2) ประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัดวิทยาศาสตร์ ระดับชั้นมัธยมศึกษาตอนต้น ตัวอย่างในการวิจัย ได้แก่ ข้อสอบรู้วิทยาศาสตร์ชั้นมัธยมศึกษาตอนต้นของโรงเรียนมัธยมศึกษา จำนวน 1,089 ข้อ ที่ได้จากการสุ่มแบบหลายขั้นตอน (multi-stage random sampling) และผู้เชี่ยวชาญในการประเมินความสอดคล้องฯ จำนวน 20 คน เครื่องมือการวิจัย คือ แบบประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด วิเคราะห์ความเที่ยงระหว่างผู้ประเมิน (inter-rater reliability) ด้วยสถิติแคปปาของฟลีส (Fleiss' kappa statistic) และสหสัมพันธ์ภายในชั้น (intra-class correlation: ICC) และวิเคราะห์ค่าเฉลี่ยคะแนนความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด

ผลการวิจัย พบว่า 1) ในส่วนของการประเมินระดับความซับซ้อนทางปัญญา มีค่าความเที่ยงระหว่างผู้ประเมินที่วิเคราะห์ด้วยสถิติแคปปาของฟลีส (Fleiss' kappa statistic: K_f) อยู่ในระดับดี (K_f = 0.510) 2) ในส่วนของการประเมินระดับความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด มีค่าความเที่ยงระหว่างผู้ประเมินที่วิเคราะห์ด้วยสถิติสหสัมพันธ์ภายในชั้น (intra-class correlation: ICC) อยู่ในระดับดีมาก (ICC = 0.954, Sig. = .000) และ 3) ข้อสอบร้อยละ 92.93 มีความสอดคล้องในแนวเดียวกันกับตัวชี้วัดที่ระบุ โดยมีค่าเฉลี่ยคะแนนความสอดคล้องฯ อยู่ระหว่าง 3.20–4.00

คำสำคัญ: ความสอดคล้องในแนวเดียวกัน, ความเที่ยงระหว่างผู้ประเมิน, สถิติแคปปาของฟลีส, สถิติสหสัมพันธ์ภายในชั้น

Article Info: Received 13 June, 2018; **Received in revised form** 2 August, 2020; **Accepted** 6 August, 2020

[1] นิสิตดุษฎีบัณฑิตสาขาวิชาการวัดและประเมินผลการศึกษา ภาควิชาวิจัยและจิตวิทยาการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อีเมล: bjanprasert@gmail.com
Ph.D. Candidate in Educational Measurement and Evaluation Department of Educational Research and Psychology, Faculty of Education, Chulalongkorn University Email: bjanprasert@gmail.com
[2] อาจารย์ประจำสาขาวิชาการวัดและประเมินผลการศึกษา ภาควิชาวิจัยและจิตวิทยาการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อีเมล: nuttaporn.l@chula.ac.th
Lecturer in Educational Measurement and Evaluation Department of Educational Research and Psychology, Faculty of Education, Chulalongkorn University Email: nuttaporn.l@chula.ac.th
[3] อาจารย์ประจำสาขาวิชาศึกษาศาสตร์ มหาวิทยาลัยสุโขทัยธรรมาธิราช อีเมล: sungworn@hotmail.com
Lecturer in School of Education, Sukhothai Thammathirat Open University Email: sungworn@hotmail.com

หมายเหตุ: งานวิจัยนี้ได้รับการสนับสนุนทุน 90 ปี จุฬาลงกรณ์มหาวิทยาลัย กองทุนรัชดาภิเษกสมโภช

Abstract

This research aimed to: 1) examine the inter-rater reliability of alignment between science items and indicators at the lower secondary level; and 2) evaluate the alignment between science items and indicators at the lower secondary level. Research subjects were 1,089 science test items used at the lower secondary level, chosen by using a multi-stage random sampling procedure. Analysis relied on 20 expert panelists to evaluate the alignment. The data were analyzed for inter-rater reliability by Fleiss' kappa statistic and the intra-class correlation (ICC), and mean scores of alignment between science test items and indices were calculated.

The findings revealed that 1) in the cognitive complexity evaluation part, there was good inter-rater reliability, as demonstrated by the Fleiss' kappa statistic (K_f = 0.510), 2) in the evaluation of alignment between science test items, there was excellent inter-rater reliability, demonstrated by the intra-class correlation (ICC = 0.954, Sig. = .000), and 3) 92.93 percent of the items aligned with the specified indices by mean scores of 3.20-4.00.

Keywords: alignment, inter-rater reliability, Fleiss' kappa statistic, intra-class

บทนำ

ความสอดคล้องในแนวเดียวกัน (alignment) ของมาตรฐานการเรียนรู้และการประเมินผลการเรียนรู้ของผู้เรียนเป็นสิ่งที่จำเป็นและส่งผลต่อระบบการจัดการศึกษาแบบอิงมาตรฐานที่มีประสิทธิภาพ (Webb, 2002) ทั้งนี้ ความสอดคล้องในแนวเดียวกัน (alignment) หมายถึง ระดับที่ผลการเรียนรู้ที่คาดหวัง (เช่น มาตรฐาน ตัวชี้วัด วัตถุประสงค์) การประเมินผล (เช่น ข้อสอบ) ตลอดจนองค์ประกอบต่าง ๆ ในระบบการศึกษา มีความเชื่อมโยงสัมพันธ์และประสานกัน เพื่อนำไปสู่การเรียนรู้อย่างมีประสิทธิภาพของผู้เรียน (Webb, 1997a)

แท้จริงแล้วความสอดคล้องในแนวเดียวกันไม่ใช่เรื่องใหม่ในแวดวงการวัดและประเมินทางการศึกษา (Ananda, 2003; Case & Zucker, 2005; Impara, 2001; Resnick et al., 2004; Webb, 1999) เพราะการศึกษาความสอดคล้องในบริบทของการวัดและประเมินผลอิงมาตรฐานเป็นประเด็นเดียวกับการศึกษาเรื่อง "ความตรง" (validity) ซึ่งเมื่อใดก็ตามที่มีการประเมินอิงมาตรฐาน ควรมีการประเมินความสอดคล้องด้วย ผลการวิเคราะห์ความสอดคล้องจะเป็น "หลักฐานของการประเมินความตรงของการแปลและใช้ผลการประเมิน" (evidence of assessment's validity) (Webb, 1977 อ้างถึงใน สังวรณ์

งัดกระโทก, 2555) สำหรับประเทศไทยที่เริ่มใช้หลักสูตรอิงมาตรฐานตั้งแต่ที่กระทรวงศึกษาได้ประกาศใช้หลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน พุทธศักราช 2544 และมีการปรับปรุงพุทธศักราช 2551 แต่ยังไม่ชัดเจนว่ามีการศึกษาการวัดความสอดคล้องในแนวเดียวระหว่างองค์ประกอบในการจัดการศึกษา อาจเป็นเพราะประเด็นเกี่ยวกับการวัดความสอดคล้องต่าง ๆ ยังไม่เป็นที่รู้จักแพร่หลายในหมู่นักวิชาการมากนัก ถึงแม้ว่าประเด็นดังกล่าวจะเป็นสิ่งจำเป็นมากในการวัดและประเมินผลอิงมาตรฐาน

วิธีการวัดความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับมาตรฐานและตัวชี้วัดทำโดยให้ผู้เชี่ยวชาญพิจารณาตัดสินความสอดคล้องระหว่างข้อสอบกับมาตรฐานและตัวชี้วัดที่ต้องการวัด ข้อสอบแต่ละข้อจะถูกประเมินด้วยผู้เชี่ยวชาญ จำนวนตั้งแต่ 2 คนขึ้นไป (Porter & Smithson, 2001; Rothman et al., 2002; Webb, 1997a, 1997b) โดยผู้เชี่ยวชาญจะพิจารณาความสอดคล้องทั้งในมิติด้านเนื้อหา (content dimension) และมิติด้านความซับซ้อนทางปัญญา (cognitive complexity dimension) ซึ่งการจำแนกความซับซ้อนทางปัญญาสามารถจำแนกได้หลายรูปแบบ ขึ้นอยู่กับแนวคิดที่ใช้ในการจำแนก เช่น แนวคิดการจำแนกลำดับขั้นของกระบวนการทางปัญญาของ Bloom ที่ได้รับการปรับปรุงใหม่ (revised Bloom’s taxonomy) (Anderson et al., 2001) จำแนกความซับซ้อนทางปัญญาได้เป็น 6 ระดับ (ได้แก่ จำ เข้าใจ ประยุกต์ใช้ วิเคราะห์ ประเมินค่า และสร้างสรรค์) แนวคิดของ Webb (1999) จำแนกได้เป็น 4 ระดับ (ได้แก่ ระลึกได้ ทักษะ/ ความคิดรวบยอด การคิดเชิงกลยุทธ์ และการขยายความคิด) แนวคิดของ Porter and Smithson (2001) จำแนกได้เป็น 5 ระดับ (จำ ปฏิบัติตามขั้นตอน สื่อสารความเข้าใจ แก้ปัญหาในสถานการณ์ใหม่ และคาดคะเน/ สรุปอ้างอิง/ พิสูจน์) แนวคิดของ Marzano (Marzano & Kendall, 2001) จำแนกได้เป็น 4 ระดับ (ได้แก่ ดึงความรู้เดิมออกมาใช้ สร้างความคิดรวบยอด วิเคราะห์ความรู้ ใช้ประโยชน์จากความรู้ อภิปัญญา และจัดระบบแห่งตน) อย่างไรก็ตาม Näsström and Henriksson (2008) ได้ทำวิเคราะห์กรอบแนวคิดหรือลำดับขั้นในการจำแนกความซับซ้อนทางปัญญาจากแนวคิด 9 แนวคิด และสรุปว่า revised Bloom’s taxonomy มีความครอบคลุมและจำเพาะ ตลอดจนมีค่าความเที่ยงระหว่างผู้ตรวจที่สูงกว่าเกณฑ์การจำแนกตามแนวคิดอื่น ๆ งานวิจัยนี้จึงเลือกใช้แนวคิดของ revised Bloom’s taxonomy มาใช้ในการศึกษาระดับความซับซ้อนทางปัญญาของข้อสอบ

อย่างไรก็ตาม จากการศึกษางานวิจัยความสอดคล้องในแนวเดียวกันที่ผ่านมา ยังขาดหลักฐานที่แสดงถึงความน่าเชื่อถือของผลการประเมินโดยผู้เชี่ยวชาญ ในการยืนยันคุณภาพในด้านความสอดคล้องภายในของผู้ประเมิน ซึ่งเป็นสิ่งที่สำคัญและส่งผลต่อความน่าเชื่อถือของการประเมิน ดังนั้น ในงานวิจัยนี้จึงได้มีการดำเนินการวิเคราะห์เพื่อตรวจสอบความเห็นพ้องกันระหว่างผู้ประเมิน (inter-rater agreement: IRA) หรือเรียกอีกอย่างหนึ่งว่าความเที่ยงระหว่างผู้ประเมิน (inter-rater reliability: IRR) ซึ่งเป็นความคงเส้นคงวาของผลการพิจารณาของผู้ประเมิน (Rosnow & Rosenthal, 1991) และเป็นคุณสมบัติพื้นฐานในการออกแบบและการประเมินคุณภาพของเครื่องมือที่ใช้ในการวิจัย (Gisev et al., 2013) ซึ่งสถิติที่ใช้ในการวิเคราะห์ความเที่ยงระหว่างผู้ประเมินมีหลากหลาย (Gisev et al., 2013; Hallgren, 2012; McHugh, 2012) เช่น การหาร้อยละของความเห็นพ้องระหว่างผู้ประเมิน (percent agreement) สถิติแคปปาของโคเฮน (Cohen, 1960) ใช้สำหรับกรณีที่มีผู้ประเมินจำนวน 2 คน และข้อมูลอยู่ในระดับนามบัญญัติ และสถิติแคปปาแบบถ่วงน้ำหนัก (Cohen, 1968) ใช้สำหรับข้อมูลระดับเรียงอันดับ สถิติแคปปาของฟลีส (Fleiss' kappa statistic: K_f) (Fleiss, 1971) สำหรับกรณีที่มีผู้ประเมินตั้งแต่ 3 คนขึ้นไป สัมประสิทธิ์สหสัมพันธ์เพียร์สัน (Pearson's correlation coefficient) สหสัมพันธ์โร (Spearman Rho) สหสัมพันธ์ภายในชั้น (intra-class correlation: ICC) และสัมประสิทธิ์แอลฟาของคริปเพนดอร์ฟ (Krippendorff's alpha coefficient) สามารถใช้สำหรับกรณีที่มีผู้ประเมินหลายคน และข้อมูลอยู่ในระดับเรียงอันดับ อันตรภาค และอัตราส่วน อย่างไรก็ตาม การเลือกใช้สถิติที่เหมาะสมกับบริบทของการศึกษาเป็นสิ่งที่สำคัญ เนื่องจากการวิเคราะห์ด้วยสถิติที่แตกต่างกัน นำไปสู่ผลลัพธ์และการแปลผลที่แตกต่างกัน ในงานวิจัยนี้ได้เลือกใช้สถิติแคปปาของฟลีส ในการวิเคราะห์ความเที่ยงระหว่างผู้ประเมินในการประเมินระดับความซับซ้อนทางปัญญา เนื่องจากข้อมูลในส่วนนี้เป็นข้อมูลที่อยู่ในระดับนามบัญญัติ และมีจำนวนผู้ประเมินตั้งแต่ 3 คนขึ้นไป และเลือกใช้สถิติสหสัมพันธ์ภายในชั้น เนื่องจากข้อมูลในส่วนนี้เป็นข้อมูลระดับเรียงอันดับ

ข้อมูลจากที่กล่าวมาข้างต้น ผู้วิจัยสนใจศึกษาการตรวจสอบความเที่ยงระหว่างผู้ประเมินของเครื่องมือในการประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัดวิทยาศาสตร์ ระดับชั้นมัธยมศึกษาตอนต้น ทั้งในมิติด้านเนื้อหาและมิติด้านความซับซ้อนทางปัญญา เพื่อเป็นการตรวจสอบและยืนยันคุณภาพในด้านความสอดคล้องภายในของ

ผู้ประเมิน และประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัดฯ ในการตรวจสอบหลักฐานของการประเมินความตรงของการแปลและใช้ผลการประเมินแบบอิงมาตรฐาน และให้สารสนเทศเกี่ยวกับความสอดคล้องของมาตรฐานและการประเมินวิทยาศาสตร์ในโรงเรียนมัธยมศึกษา อันจะนำมาใช้เป็นแนวทางในการพัฒนาการประเมินผล รวมถึงการเรียนการสอนให้สอดคล้องกับมาตรฐานที่กำหนด ให้มุ่งไปสู่เป้าหมายเดียวกัน

วัตถุประสงค์

1. เพื่อตรวจสอบความเที่ยงระหว่างผู้ประเมินของความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัดวิทยาศาสตร์ ระดับชั้นมัธยมศึกษาตอนต้น

2. เพื่อประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัดวิทยาศาสตร์ ระดับชั้นมัธยมศึกษาตอนต้น

วิธีการวิจัย

1. ประชากรและตัวอย่างในการวิจัย

ประชากรในการวิจัยครั้งนี้ ประกอบด้วย

1) ข้อสอบที่ครูสร้างขึ้น (teacher-made test) ในการประเมินระดับชั้นเรียน กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับชั้นมัธยมศึกษาตอนต้น (มัธยมศึกษาปีที่ 1–3) ของโรงเรียนมัธยมศึกษา สังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน ในกรุงเทพมหานคร เพื่อประเมินผลการเรียนรู้ของผู้เรียน ทั้งข้อสอบระหว่างภาคและข้อสอบประจำภาค ในปีการศึกษา 2559

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

ตัวอย่างในการวิจัย ประกอบด้วย

1) ข้อสอบที่ครูสร้างขึ้นในการประเมินระดับชั้นเรียน กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ชั้นมัธยมศึกษาตอนต้น (มัธยมศึกษาปีที่ 1–3) ของโรงเรียนมัธยมศึกษาสังกัด

สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน ในกรุงเทพมหานคร ในปีการศึกษา 2559 จำนวน 1,089 ข้อ โดยเป็นข้อสอบที่วัดตัวชี้วัดที่สำนักทดสอบทางการศึกษาแห่งชาติ (สทศ.) ใช้ในการออกข้อสอบ O-NET วิทยาศาสตร์ ชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2559 ซึ่งมีจำนวน 40 ตัวชี้วัด การได้มาซึ่งตัวอย่างข้อสอบใช้วิธีการสุ่มแบบหลายขั้นตอน (multi-stage sampling) ได้แบบสอบทั้งหมด 48 ชุด จากโรงเรียน 4 โรง จากนั้นจึงทำการคัดเลือกเฉพาะข้อสอบที่ครูระบุว่าใช้ในการวัดตัวชี้วัดทั้ง 40 ตัวชี้วัด

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

2. เครื่องมือในการวิจัย

เครื่องมือในการวิจัยครั้งนี้ ได้แก่ แบบประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด กลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับชั้นมัธยมศึกษาตอนต้น ซึ่งมีข้อรายการที่ประกอบด้วย 2 ส่วนหลัก ดังนี้

ส่วนที่ 1 เป็นรายละเอียดของข้อสอบที่ถูกประเมิน ประกอบด้วย ข้อคำถามและตัวเลือก (กรณีเป็นข้อสอบหลายตัวเลือก) คะแนนรายข้อ ระดับชั้น และมาตรฐานและตัวชี้วัดที่ข้อสอบข้อนั้น ๆ ต้องการวัด

ส่วนที่ 2 เป็นส่วนการประเมินความสอดคล้องของข้อสอบเป็นรายข้อใน 3 ประเด็นหลัก ได้แก่ 1) ระดับความซับซ้อนทางปัญญา (cognitive complexity) ซึ่งเป็นลำดับขั้นของกระบวนการทางปัญญา (cognitive process dimension) ตามแนวคิด revised Bloom's taxonomy (Anderson et al., 2001) ซึ่งแบ่งเป็น 6 ขั้น ได้แก่ จำ เข้าใจ ประยุกต์ใช้ วิเคราะห์ ประเมินค่า และสร้างสรรค์

2) ระดับความสอดคล้องในแนวเดียวกันระหว่างเนื้อหาของข้อสอบกับเนื้อหาตัวชี้วัดที่กำหนด มีลักษณะเป็นมาตรประมาณค่า 5 ระดับ (0-4) โดย 0 คือ ไม่สอดคล้อง 1 คือ ค่อนข้างสอดคล้อง 2 คือ ไม่แน่ใจ 3 คือ ค่อนข้างสอดคล้อง และ 4 คือ สอดคล้องโดยตรง กำหนดเกณฑ์ค่าเฉลี่ยคะแนนความสอดคล้อง ต้องมีค่าตั้งแต่ 3.00 ขึ้นไป จึงจะถือว่าข้อสอบจะมีความสอดคล้องในแนวเดียวกันกับตัวชี้วัด

3) ข้อเสนอแนะเพิ่มเติม เป็นส่วนที่ให้ผู้เชี่ยวชาญเขียนแสดงข้อสังเกตหรือข้อเสนอแนะอื่น ๆ เพิ่มเติมเกี่ยวกับข้อสอบแต่ละข้อ ตัวอย่างข้อรายการประเมินในแบบประเมินความสอดคล้องฯ แสดงดังภาพ 1

การตรวจสอบคุณภาพของแบบประเมินฯ ก่อนนำไปใช้จริง โดยให้ผู้เชี่ยวชาญจำนวน 7 ท่าน ตรวจสอบความถูกต้องและความเหมาะสมของรูปแบบและภาษา และผู้วิจัยได้ทำการสุ่มข้อสอบจำนวน 40 ข้อ (ตัวชี้วัดละ 1 ข้อ) เพื่อจัดทำเป็นแบบประเมินความสอดคล้องฯ ฉบับทดลองใช้ และให้ผู้เชี่ยวชาญทั้ง 7 ท่าน ทดลองประเมินความสอดคล้องฯ ของข้อสอบ พบว่า ในส่วนของการประเมินระดับความซับซ้อนทางปัญญา ได้ค่าสถิติแคปปาของฟลีส เท่ากับ 0.596 แสดงว่ามีความสอดคล้องระหว่างผู้ประเมินอยู่ในระดับดีมาก (Fleiss et al., 2003) และใช้สถิติสหสัมพันธ์ภายในชั้น (intra-class correlation: ICC) ในส่วนของการประเมินระดับความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด ได้ค่าสหสัมพันธ์ภายในชั้นที่วิเคราะห์จากโมเดล two-way mixed-effects model โดยมีผู้ประเมินจำนวน 7 คน (ใช้สัญลักษณ์คือ ICC (3,7)) ที่ระดับความเชื่อมั่น 95% มีค่าเท่ากับ 0.943 แสดงว่า มีความสอดคล้องระหว่างผู้ประเมินอยู่ในระดับดีมาก (Portney & Watkins, 2015)

ภาพ 1

ตัวอย่างข้อรายการประเมินในแบบประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัดฯ

ข้อที่ N24 2-3123.21-221	ในการแยกสาร A, B และ C โดยใช้วิธีโครมาโทกราฟี ผลปรากฏว่า สารทั้ง 3 ชนิด เคลื่อนที่ไปบนตัวดูดซับได้ 5, 20 และ 12 cm ตามลำดับ ข้อใดกล่าวถูกต้อง 1. สาร A ถูกดูดซับได้ดีที่สุด 2. สาร A ละลายได้ดีกว่าสาร B 3. สาร C ดูดซับได้มากกว่าสาร A 4. สาร C ละลายได้น้อยกว่าสาร A
คะแนนรายข้อ	0.50 คะแนน
ระดับชั้น	มัธยมศึกษาปีที่ 2
มาตรฐาน	ว 3.1 เข้าใจสมบัติของสาร ความสัมพันธ์ระหว่างสมบัติของสารกับโครงสร้างและแรงยึดเหนี่ยวระหว่างอนุภาค มีกระบวนการสืบเสาะหาความรู้และจิตวิทยาศาสตร์ สื่อสารสิ่งที่เรียนรู้ นำความรู้ไปใช้ประโยชน์
ตัวชี้วัด	ม. 2/3 ทดลองและอธิบายหลักการแยกสารด้วยวิธีการกรอง การตกผลึก การสกัด การกลั่น และโครมาโท กราฟี และนำความรู้ไปใช้ประโยชน์

1. ระดับความซับซ้อนทางปัญญา *(เลือกได้เพียงระดับเดียว)*

☐ จำ ☐ เข้าใจ ☐ ประยุกต์ใช้ ☐ วิเคราะห์ ☐ ประเมินค่า ☐ สร้างสรรค์

2. ระดับความสอดคล้องระหว่างข้อสอบกับตัวชี้วัด *(เลือกได้เพียงระดับเดียว)*

☐ ไม่สอดคล้อง (0) ☐ ค่อนข้างไม่สอดคล้อง (1) ☐ ไม่แน่ใจ (2) ☐ ค่อนข้างสอดคล้อง (3) ☐ สอดคล้องโดยตรง (4)

3. ข้อเสนอแนะเพิ่มเติม (ถ้ามี) .....

3. การเก็บและรวบรวมข้อมูล

เก็บรวบรวมข้อมูลโดยให้ผู้เชี่ยวชาญ จำนวน 20 คน ประเมินความสอดคล้องระหว่างกรอบมาตรฐานและตัวชี้วัดกับข้อสอบในการประเมินผลการเรียนรู้ของโรงเรียน จำนวน 1,089 ข้อ มีขั้นตอนการดำเนินงานดังนี้

1) จัดประชุมและฝึกปฏิบัติผู้เชี่ยวชาญ โดยมีจุดประสงค์เพื่อสร้างความเข้าใจที่ตรงกันระหว่างผู้เชี่ยวชาญในเรื่องต่าง ๆ เช่น การทบทวนนิยามของระดับความซับซ้อนทางปัญญา ตามแนวคิด revised Bloom's taxonomy ตลอดจนแนวทางในการดำเนินงานของผู้เชี่ยวชาญ โดยในการประชุมได้มีการทดลองให้ผู้เชี่ยวชาญทดลองฝึกปฏิบัติกระบวนการฉันทามติ และฝึกประเมินความสอดคล้องระหว่างข้อสอบกับตัวชี้วัดเป็นรายบุคคล

2) หลังจากการประชุมผู้เชี่ยวชาญแล้ว ผู้เชี่ยวชาญแต่ละคนจะทำหน้าที่ประเมินความสอดคล้องระหว่างตัวชี้วัดกับข้อสอบได้อย่างอิสระ โดยใช้แบบประเมินความสอดคล้องระหว่างตัวชี้วัดกับข้อสอบฯ และเอกสารคู่มือในการดำเนินงานสำหรับผู้เชี่ยวชาญที่ผู้วิจัยจัดทำขึ้น แต่เนื่องจากข้อสอบในการศึกษาครั้งนี้มีจำนวนมาก (1,089 ข้อ) ผู้วิจัยจึงออกแบบให้ข้อสอบแต่ละข้อถูกประเมินโดยผู้เชี่ยวชาญ จำนวน 5 คน แต่จะมีข้อสอบจำนวน 40 ข้อ (สุ่มมาจากตัวชี้วัดละ 1 ข้อ) ที่ถูกประเมินด้วยผู้เชี่ยวชาญทุกคน (20 คน) ข้อมูลในส่วนนี้จะถูกไปใช้ในการตรวจสอบความเที่ยงระหว่างผู้ประเมิน ทั้งนี้ ผู้เชี่ยวชาญแต่ละคนจะได้ประเมินข้อสอบอย่างสุ่ม จำนวนคนละ 287-312 ข้อ ที่ถูกแบ่งเป็นแบบประเมินชุดย่อย ได้คนละ 7 ชุด (จากทั้งหมด 25 ชุด (ชุด A-Y)) แบบประเมินแต่ละชุดย่อยประกอบด้วยข้อสอบจำนวน 41–46 ข้อ รายละเอียดตาราง 1

ตาราง 1

แผนผังการออกแบบการประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด

ลำดับที่ของแบบประเมินย่อย	ผู้เชี่ยวชาญคนที่																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
2	B	B	B	B	B	C	D	D	I	I	I	I	M	L	O	O	O	O	O	K
3	C	C	C	C	F	F	E	E	J	J	J	J	U	U	P	P	P	P	P	U
4	D	D	D	F	G	G	F	F	K	K	K	K	V	V	Q	Q	Q	Q	Q	V

ตาราง 1 (ต่อ)

แผนผังการออกแบบการประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด

ลำดับที่ของแบบประเมินย่อย	ผู้เชี่ยวชาญคนที่																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
5	E	E	E	H	H	H	G	G	L	L	L	L	W	W	R	R	R	R	R	W
6	G	U	V	W	X	N	H	H	M	M	M	M	X	X	S	S	S	S	S	X
7	U	V	W	X	Y	Y	I	J	N	N	N	N	Y	Y	T	T	T	T	T	Y

หมายเหตุ: แบบประเมิน 25 ชุดย่อย (A–Y) แต่ละชุดมีข้อสอบ 41–46 ข้อ, ผู้เชี่ยวชาญแต่ละคนประเมินข้อสอบจำนวน 287–312 ข้อ (7 ชุดย่อย)

4. การวิเคราะห์ข้อมูล

4.1 การวิเคราะห์ความเที่ยงระหว่างผู้ประเมิน (inter-rater reliability: IRR) เพื่อตรวจสอบความสอดคล้องระหว่างผู้ประเมินโดยใช้ข้อมูลจากแบบประเมินฯ ที่มีข้อสอบจำนวน 40 ข้อ (ชุดย่อย A) โดยผู้ประเมิน 20 คน มีดังนี้

1) วิเคราะห์ความเที่ยงระหว่างผู้ประเมินของแบบประเมินฯ ในส่วนของการประเมินระดับความซับซ้อนทางปัญญาปัญญา โดยเลือกใช้สถิติแคปปาของฟลีส (Fleiss' kappa statistic: K_f) (Fleiss, 1971) เนื่องจากข้อมูลในส่วนนี้เป็นข้อมูลที่อยู่ในระดับนามบัญญัติ และมีจำนวนผู้ประเมินตั้งแต่ 3 คนขึ้นไป สูตรในการคำนวณสถิติแคปปาของฟลีส ดังนี้

$$K_f = \frac{\bar{P}-\bar{P}e}{1-\bar{P}e} \qquad \text{เมื่อ} \qquad \bar{P} = \frac{1}{N}\sum_{i=1}^{N} p_i$$

โดย $p_i = \frac{1}{n(n-1)}\sum_{j=1}^{k} n_{ij^2} - \mathrm{n}$ ดังนั้น $\bar{P} = \frac{1}{Nn(n-1)}\sum_{i=1}^{N}\sum_{j=1}^{k} n_{ij^2} - n$ และ $\bar{P}_e = \sum_{j=1}^{k} p_{j^2}$

โดยที่ N แทน จำนวนข้อสอบที่ถูกประเมิน,

n แทน จำนวนผู้ประเมินข้อสอบแต่ละข้อ

k แทน จำนวนประเภทของการประเมิน (number of categories of scale)

i แทน ข้อสอบที่ถูกประเมิน,

j แทน ประเภทที่ถูกประเมิน (categories of scale)

กำหนดเกณฑ์การพิจารณาระดับความสอดคล้องตามแนวทางของ Fleiss et al. (2003) ดังนี้

ค่า K_f อยู่ระหว่าง 0.00 – 0.39 หมายถึง ผู้ประเมินมีความสอดคล้องในการประเมินระดับต่ำ

ค่า K_f อยู่ระหว่าง 0.40 – 0.74 หมายถึง ผู้ประเมินมีความสอดคล้องในการประเมินระดับดี

ค่า K_f อยู่ระหว่าง 0.75 – 1.00 หมายถึง ผู้ประเมินมีความสอดคล้องในการประเมินระดับดีมาก

คำนวณค่าความคลาดเคลื่อนมาตรฐานของสถิติแคปปาของฟลีส ($SE(K_f)$) ซึ่ง Fleiss (1971) คำนวณได้จากรากที่สองของค่าความแปรปรวนของสถิติแคปปา ($Var(K_f)$) สูตรในการคำนวณ $Var(K_f)$ ดังนี้

$$Var(K_f) = \frac{2}{Nn(n-1)} x \frac{\sum_j p_j^2 - (2n-3)(\sum_j p_j^2)^2 + 2(n-2)\sum_j p_j^3}{(1-\sum_j p_j^2)^2}$$

2) วิเคราะห์ความเที่ยงระหว่างผู้ประเมินของแบบประเมินฯ ในส่วนของการประเมินระดับความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด (มาตรประมาณค่า 0–4) โดยใช้สถิติสหสัมพันธ์ภายในชั้น (intra-class correlation: ICC) ด้วยโปรแกรม SPSS เนื่องจากข้อมูลในส่วนนี้เป็นข้อมูลระดับเรียงอันดับ (Gisev et al., 2013) สหสัมพันธ์ภายในชั้นที่วิเคราะห์จากโมเดล two-way mixed-effects model โดยมีผู้ประเมินจำนวน 20 คน (ใช้สัญลักษณ์ คือ ICC (3,20)) ที่ระดับความเชื่อมั่น 95% โดยกำหนดเกณฑ์การพิจารณาระดับความสอดคล้องตามแนวทางของ Portney and Watkins (2015) ดังนี้

ICC > 0.50	หมายถึง ผู้ประเมินมีความสอดคล้องในการประเมินระดับต่ำ
0.50 ≤ ICC < 0.75	หมายถึง ผู้ประเมินมีความสอดคล้องในการประเมินระดับพอใช้
0.75 ≤ ICC ≤ 0.90	หมายถึง ผู้ประเมินมีความสอดคล้องในการประเมินระดับดี

ICC > 0.90 หมายถึง ผู้ประเมินมีความสอดคล้องในการประเมินระดับดีมาก

$$SEM = S \times \sqrt{1 - ICC}$$

คำนวณค่าความคลาดเคลื่อนมาตรฐานในการวัด (standard error of measurement: SEM) จากสูตร

4.2 วิเคราะห์ความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัดที่กำหนด โดยการคำนวณค่าเฉลี่ยคะแนนความสอดคล้องในแนวเดียวกันฯ ของคะแนนการประเมินจากผู้เชี่ยวชาญแต่ละคนเป็นรายข้อ ทั้งนี้ กำหนดให้ข้อสอบแต่ละต้องมีค่าเฉลี่ยคะแนนความสอดคล้องฯ ตั้งแต่ 3.00 คะแนนขึ้นไป จึงจะถือว่าข้อสอบข้อนั้น ๆ มีความสอดคล้องในแนวเดียวกันกับตัวชี้วัดที่กำหนด (Anderson et al., 2015)

ผลการวิจัย

1. ผลการวิเคราะห์ความเที่ยงระหว่างผู้ประเมิน ในส่วนของการประเมินระดับความซับซ้อนทางปัญญา โดยใช้สถิติแคปปาของฟลีส (Fleiss' kappa statistic: K_f) พบว่ามีค่า K_f เท่ากับ 0.510 มีค่าความคลาดเคลื่อนมาตรฐานของ K_f เท่ากับ 0.017 แสดงว่ามีค่าความสอดคล้องกันระหว่างผู้ประเมินอยู่ในระดับดี รายละเอียดดังตาราง 2

ตาราง 2

ค่าสถิติแคปปาของฟลีสของการประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด

จำนวนข้อสอบ (ข้อ)	จำนวนผู้ประเมิน (คน)	Fleiss' kappa	Var (K_f)	SE (K_f)
40	20	0.510	0.0003	0.017

2. ผลการวิเคราะห์ความเที่ยงระหว่างผู้ประเมิน ในส่วนของการประเมินระดับความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด ที่มีลักษณะเป็นมาตรประมาณค่า 5 ระดับ (0–4) โดยใช้สถิติสหสัมพันธ์ภายในชั้น (intra-class correlation: ICC) ด้วยโปรแกรม SPSS ที่วิเคราะห์จากโมเดล two-way mixed-effects model ที่ระดับความเชื่อมั่น 95% พบว่า มีค่าสหสัมพันธ์ภายในชั้น เท่ากับ 0.954 อย่างมีนัยสำคัญทางสถิติ (Sig. = .000) แสดงว่า มีความสอดคล้องระหว่างผู้ประเมินในระดับดีมาก มีค่าความคลาดเคลื่อนมาตรฐาน

ในการวัด (standard error of measurement: SEM) เท่ากับ 0.02 รายละเอียดแสดงดังตาราง 3

ตาราง 3

สถิติสหสัมพันธ์ภายในชั้นของการประเมินระดับความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด

ผู้ประเมินคนที่	MEAN ± *SD*	ผู้ประเมินคนที่	MEAN ± *SD*
1	3.68 ± 0.57	11	3.83 ± 0.38
2	3.40 ± 0.67	12	3.88 ± 0.33
3	3.70 ± 0.72	13	3.88 ± 0.33
4	2.90 ± 0.81	14	3.85 ± 0.43
5	3.43 ± 0.98	15	3.83 ± 0.55
6	3.30 ± 0.72	16	3.63 ± 0.90
7	3.83 ± 0.38	17	3.75 ± 0.59
8	3.75 ± 0.59	18	3.33 ± 0.83
9	3.80 ± 0.41	19	3.43 ± 0.96
10	3.65 ± 0.48	20	3.48 ± 0.75
ICC (ช่วงความเชื่อมั่น 95%) = 0.954 (0.929 - 0.973) Sig. = .000 SEM = 0.02			

3. ผลการวิเคราะห์ความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัดที่กำหนด จากการคำนวณค่าเฉลี่ยคะแนนความสอดคล้องในแนวเดียวกันฯ ของคะแนนการประเมินจากผู้เชี่ยวชาญแต่ละคนเป็นรายข้อ ตามแผนผังการออกแบบการประเมินความสอดคล้องฯ ในตาราง 1 พบว่า จากข้อสอบ 1,089 ข้อ ข้อสอบส่วนใหญ่มีความสอดคล้องในแนวเดียวกันกับตัวชี้วัดที่ระบุ (มีค่าเฉลี่ยตั้งแต่ 3.00 ขึ้นไป) จำนวน 1,012 ข้อ คิดเป็นร้อยละ 92.93 โดยมีค่าเฉลี่ยคะแนนความสอดคล้องฯ อยู่ระหว่าง 3.20–4.00 และมีข้อสอบ จำนวน 77 ข้อ คิดเป็นร้อยละ 7.03 ที่ไม่สอดคล้องในแนวเดียวกันกับตัวชี้วัดที่ระบุ (มีค่าเฉลี่ยน้อยกว่า 3.00) โดยมีค่าเฉลี่ยคะแนนความสอดคล้องฯ อยู่ระหว่าง 0.60–2.90 ตัวอย่างของผลการวิเคราะห์ข้อสอบที่ได้รับการสุ่มเพื่อนำเสนอจำนวน 10 ข้อ แสดงดังตาราง 4

ตาราง 4

ค่าสถิติพรรณาของคะแนนความสอดคล้องในแนวเดียวกันของข้อสอบ (สุ่มตัวอย่างข้อสอบ 10 ข้อ)

ข้อ	มาตรฐานและตัวชี้วัด	จำนวนผู้ประเมิน (คน)	Min	Max	MEAN (SE)	*SD*
1	1123	5	2.00	4.00	3.20 (0.37)	0.84
2	1233	5	0.00	2.00	1.20 (0.37)	0.84
3	2131	20	3.00	4.00	3.95 (0.05)	0.22
4	3111	5	3.00	4.00	3.80 (0.20)	0.45
5	3111	5	4.00	4.00	4.00 (0.00)	0.00
6	4122	20	1.00	3.00	2.30 (0.18)	0.80
7	4131	5	0.00	2.00	0.60 (0.40)	0.89
8	5131	5	2.00	4.00	3.40 (0.40)	0.89
9	6116	5	3.00	4.00	3.60 (0.24)	0.55
10	7131	5	3.00	4.00	3.80 (0.20)	0.45

อภิปรายผล

1. จากผลการวิจัยที่พบว่า ในส่วนของการประเมินระดับความซับซ้อนทางปัญญา มีค่าความเที่ยงระหว่างผู้ประเมิน ที่วิเคราะห์ด้วยสถิติแคปปาของฟลีส (Fleiss' kappa statistic: K_f) อยู่ในระดับดี (K_f = 0.510) (Fleiss et al., 2003) และพบว่า ในส่วนของการประเมินระดับความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด มีค่าความเที่ยงระหว่างผู้ประเมินที่วิเคราะห์ด้วยสถิติสหสัมพันธ์ภายในชั้น (ICC) ที่ระดับความเชื่อมั่น 95% อยู่ในระดับดีมาก อย่างมีนัยสำคัญทางสถิติ (ICC = 0.95, ฆig. = .000) (Portney & Watkins, 2015) แสดงให้เห็นว่ามีความสอดคล้องภายในของผลการประเมิน ผู้ประเมินมีความคงเส้นคงวา และมีความน่าเชื่อถือในการประเมินความสอดคล้องระหว่างข้อสอบกับตัวชี้วัด ที่เป็นเช่นนี้อาจมีสาเหตุมาจากหลายปัจจัย ซึ่งสรุปใน 3 ประเด็นหลัก ดังนี้

1.1 การประเมินความซับซ้อนทางปัญญาโดยเลือกใช้แนวคิด revised Bloom's taxonomy (Anderson et al., 2001) ซึ่งผลการวิจัยพบว่า แบบประเมินฯ มีค่าความสอดคล้องกันระหว่างผู้ประเมินในระดับดี (K_f = 0.510) สอดคล้องกับงานวิจัยของ Näsström and Henriksson (2008) ที่ทำการวิเคราะห์กรอบแนวคิดหรือลำดับขั้นใน

การจำแนกความซับซ้อนทางปัญญาจากแนวคิด 9 แนวคิด โดยใช้ผู้ประเมินจำนวน 2 คน พบว่า การจำแนกความซับซ้อนทางปัญญาโดยใช้แนวคิด revised Bloom's taxonomy มีความครอบคลุมและจำเพาะ ตลอดจนมีค่าความเที่ยงระหว่างผู้ตรวจที่สูงกว่าเกณฑ์การจำแนกตามแนวคิดอื่น ๆ (K_f เท่ากับ 0.36–0.46) อย่างไรก็ตาม ผลการวิจัยนี้มีค่า K_f สูงกว่าของ Näsström and Henriksson (2008) เล็กน้อย ที่เป็นเช่นนี้อาจเป็นเพราะงานวิจัยครั้งนี้มีจำนวน ผู้ประเมินมากกว่า ทั้งนี้ Webb (2007) กล่าวว่า จำนวนผู้เชี่ยวชาญยิ่งมากจะทำให้ผลการศึกษายิ่งมีความเที่ยงเพิ่มมากขึ้น โดยทั่วไปควรมีผู้เชี่ยวชาญในการพิจารณาตัดสินความสอดคล้องในแนวเดียวกัน จำนวน 5–8 คน นอกจากนี้ การจำแนกความซับซ้อนทางปัญญาโดยใช้แนวคิด revised Bloom's taxonomy อาจง่ายต่อการสื่อความให้เป็นที่เข้าใจตรงกันในหมู่ครูและนักวิชาการของประเทศไทย เนื่องจากมีความคุ้นเคยมากกว่าการจำแนกด้วยแนวคิดอื่น ๆ จึงส่งผลให้มีค่าความเที่ยงระหว่างผู้ประเมินอยู่ในระดับดี ตามเกณฑ์ของ Fleiss et al. (2003)

1.2 ผู้เชี่ยวชาญที่ทำหน้าที่เป็นผู้ประเมินมีคุณสมบัติที่เหมาะสม จึงสามารถประเมินความสอดคล้อง ในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัดได้อย่างคงเส้นคงวา ซึ่งสอดคล้องกับ Davis-Becker and Buckendahl (2013) และ La Marca et al. (2000) ที่กล่าวว่าผู้เชี่ยวชาญจะต้องมีความรู้และคุ้นเคยในเนื้อหาของวิชาที่ต้องการประเมิน ลักษณะของผู้สอบ หลักสูตร ตลอดจนมาตรฐานการเรียนรู้ที่ต้องการประเมิน ซึ่งเป็นคุณสมบัติของผู้เชี่ยวชาญในงานวิจัยนี้ โดยผู้วิจัยได้กำหนดเกณฑ์การคัดเลือกผู้เชี่ยวชาญว่าทุกคนต้องเป็นผู้ที่สำเร็จการศึกษาอย่างน้อยในระดับปริญญาตรีทางด้านการสอนวิทยาศาสตร์ และสำเร็จการศึกษาอย่างน้อยในระดับปริญญาโทสาขาวิชาการวัดและประเมินทางการศึกษาหรือสาขาวิชาที่เกี่ยวข้อง และ/หรือมีประสบการณ์สอนกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับชั้นมัธยมศึกษาตอนต้น ไม่ต่ำกว่า 3 ปี

1.3 มีการประชุมและฝึกปฏิบัติผู้เชี่ยวชาญก่อนทำการประเมินจริง โดยมีกิจกรรมเพื่อสร้างความเข้าใจที่ตรงกันระหว่างผู้เชี่ยวชาญในเรื่องต่าง ๆ เช่น การทบทวนนิยามของระดับความซับซ้อนทางปัญญา ตามแนวคิด revised Bloom's taxonomy การชี้แจงแนวทางในการดำเนินงานของผู้เชี่ยวชาญ การทดลองให้ผู้เชี่ยวชาญทดลองฝึกปฏิบัติกระบวนการฉันทามติ และฝึกประเมินความสอดคล้องระหว่างข้อสอบกับตัวชี้วัดเป็นรายบุคคล

และเปิดโอกาสให้ผู้เชี่ยวชาญได้อภิปรายในภายในกลุ่มเพื่อให้ได้ผลการพิจารณาที่ถูกต้องและเข้าใจตรงกัน ซึ่งสอดคล้องกับ Webb (1999) ที่กล่าวว่าการฝึกปฏิบัติก่อนดำเนินการประเมินความสอดคล้องในแนวเดียวกันฯ มีความสำคัญต่อการศึกษาความสอดคล้องในแนวเดียวกัน เนื่องจากการฝึกปฏิบัติจะช่วยให้ผู้เชี่ยวชาญทุกคนมีความเข้าใจถึงกระบวนการประเมินความสอดคล้องในแนวเดียวกัน เข้าใจถึงหลักการจำแนกระดับความซับซ้อนทางปัญญา ตลอดจนวิธีการในการพิจารณาตัดสินความสอดคล้องของผู้เชี่ยวชาญในการประเมินจริง เพื่อให้สามารถทำหน้าที่ประเมินอย่างอิสระเป็นรายบุคคล นอกจากนี้ หลังจากการประชุมผู้เชี่ยวชาญแล้ว ผู้เชี่ยวชาญ แต่ละคนจะทำหน้าที่ประเมินความสอดคล้องระหว่างตัวชี้วัดกับข้อสอบได้อย่างอิสระ โดยในการวิจัยครั้งนี้ ผู้เชี่ยวชาญจะได้รับเอกสารคู่มือประกอบการดำเนินงานประเมินความสอดคล้องในแนวเดียวกันฯ สำหรับผู้เชี่ยวชาญ ที่ผู้วิจัยจัดทำขึ้น ประเด็นหลักของคู่มือฯ ประกอบด้วย 1) คำชี้แจงในการดำเนินงานของผู้เชี่ยวชาญ 2) นิยามศัพท์คำอธิบาย และตัวอย่างของการจำแนกพฤติกรรมการเรียนรู้ด้านพุทธิพิสัยตามแนวคิด revised Bloom's taxonomy และ 3) รายละเอียดของตัวชี้วัดและสาระการเรียนรู้แกนกลางกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับชั้นมัธยมศึกษาตอนต้น ตามหลักสูตรแกนกลางขั้นพื้นฐาน พุทธศักราช 2551

2. ผลการวิจัยที่พบว่า ข้อสอบส่วนใหญ่ (ร้อยละ 92.93) มีความสอดคล้องในแนวเดียวกันกับตัวชี้วัดที่ระบุ ซึ่งสอดคล้องกับผลการตรวจสอบคุณภาพของข้อสอบที่ใช้ในการวัดและประเมินผลในชั้นเรียนของสถานศึกษาของสำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน (2559) ที่พบว่า มีข้อสอบที่ใช้ในการวัดและประเมินผลในชั้นเรียนของสถานศึกษาที่ตรงตามมาตรฐานและตัวชี้วัดกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ระดับชั้นมัธยมศึกษาปีที่ 3 ร้อยละ 98.54 ที่เป็นเช่นนี้อาจเป็นเพราะโรงเรียนได้รับการสนับสนุนด้านการวัดและประเมินผลการเรียนรู้จากสำนักงานเขตพื้นที่การศึกษาหรือหน่วยงานต้นสังกัด เช่น การสร้างความรู้ความเข้าใจให้กับบุคลากรในสถานศึกษาที่เกี่ยวกับการวัดและประเมินผลการเรียนรู้ตามหลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน การส่งเสริมและสนับสนุนการเรียนรู้ในกลุ่มสาระการเรียนรู้ต่าง ๆ การส่งเสริมให้ครูบุคลากรในสถานศึกษามีความรู้ความเข้าใจในแนวปฏิบัติการวัดและประเมินผลรูปแบบต่าง ๆ โดยเน้นการประเมินตามสภาพจริง การส่งเสริมสนับสนุนให้สถานศึกษาพัฒนาเครื่องมือ และการบริหารจัดการด้าน

การวัดและประเมินผลการเรียนรู้ รวมถึงการให้คำปรึกษา แนะนำเกี่ยวกับการวัดและประเมินผลการเรียนรู้ (กระทรวงศึกษาธิการ, 2551)

ข้อเสนอแนะ

ข้อเสนอแนะในการนำไปใช้

1. โรงเรียนที่ออกข้อสอบควรคัดเลือกข้อสอบที่มีความสอดคล้องในแนวเดียวกันกับมาตรฐานตัวชี้วัด จัดเก็บไว้ใช้ในการวัดและประเมินผลระดับชั้นเรียน ตลอดจนสามารถนำไปพัฒนาเป็นฐานข้อมูลหรือคลังข้อสอบต่อไป

2. ในการประเมินความสอดคล้องในแนวเดียวกันระหว่างองค์ประกอบทางการศึกษาจำเป็นต้องคัดเลือกผู้เชี่ยวชาญที่ทำหน้าที่ประเมินข้อสอบที่มีความรู้ความสามารถในเนื้อหาของตัวชี้วัดและระดับที่ศึกษา และการวัดและประเมินผล มีการอบรมและฝึกปฏิบัติผู้เชี่ยวชาญที่ทำหน้าที่ประเมินให้ความเข้าใจที่ตรงกันถึงแนวทางในการดำเนินงาน เพื่อช่วยลดปัญหาอันอาจเกิดจากอิทธิพลของผู้ประเมิน และช่วยเพิ่มความเที่ยงของผลการประเมิน

ข้อเสนอแนะในการวิจัยครั้งต่อไป

1. ควรมีการวิเคราะห์ทางสถิติ เช่น การประยุกต์ใช้ทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัด (G-theory) เพื่อศึกษาจำนวนและรูปแบบการประเมินที่เหมาะสมในการประเมินความสอดคล้องในแนวเดียวกันฯ ที่จะทำให้ผลการประเมินมีความน่าเชื่อถือโดยอยู่บนพื้นฐานของการใช้ทรัพยากร ทั้งด้านเวลา แรงงาน และค่าใช้จ่ายที่เหมาะสมและมีความเป็นไปได้ในทางปฏิบัติ

2. ควรมีการศึกษาแหล่งความแปรปรวนอื่น ๆ ที่อาจส่งผลต่อความเที่ยงในการประเมินความสอดคล้องในแนวเดียวกันระหว่างข้อสอบกับตัวชี้วัด เช่น จำนวนข้อในแบบประเมิน อิทธิพลของผู้ประเมิน หรือผลสัมฤทธิ์ทางวิทยาศาสตร์ของโรงเรียนที่ใช้ในการศึกษา เป็นต้น

รายการอ้างอิง

ภาษาไทย

กระทรวงศึกษาธิการ. (2551). *แนวปฏิบัติการวัดและประเมินผลการเรียนรู้ ตามหลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน พุทธศักราช 2551*. ชุมนุมสหกรณ์การเกษตรแห่งประเทศไทย.

กระทรวงศึกษาธิการ. (2552). *หลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน พุทธศักราช 2551*. ชุมนุมสหกรณ์การเกษตรแห่งประเทศไทย.

สังวรณ์ งัดกระโทก. (2555). การวัดความสอดคล้องของมาตรฐานการเรียนรู้กับการจัดการเรียนการสอนและการประเมิน. ใน *เอกสารการสอนชุดวิชา การวัดและประเมินอิงมาตรฐานการเรียนรู้ (หน่วยที่ 6)*. มหาวิทยาลัยสุโขทัยธรรมาธิราช.

สำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน. (2559). *การติดตามและตรวจสอบคุณภาพของข้อสอบที่ใช้ในการวัดและประเมินผลในชั้นเรียนของสถานศึกษา สังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน*. ชุมนุมสหกรณ์การเกษตรแห่งประเทศไทย.

ภาษาอังกฤษ

Ananda, S. (2003). *Rethinking issues of alignment under "no child left behind"*. WestEd. https://files.eric.ed.gov/fulltext/ED476416.pdf

Anderson, D., Irvin, S., Alonzo, J., & Tindal, G. A. (2015). Gauging item alignment through online systems while controlling for rater effects. *Educational Measurement: Issues and Practice*, *34*(1), 22-33.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M. C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives* (Complete ed.). Longman.

Case, B., & Zucker, S. (2005, July). *Methodologies for alignment of standards and assessments* [Paper presentation]. China-US Conference on Alignment of Assessments and Instruction, Beijing, China.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213-233.

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. https://doi.org/10.1037/h0031619

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). John Wiley & Sons.

Gisev, N., Bell, J. S., & Chen, T. F. (2013). Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, *9*(3), 330-338.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23-34.

Impara, J. C. (2001). *Alignment: One element of an assessment's instructional unity* [Paper presentation]. 2001 annual meeting of the National Council on Measurement in Education, Seattle, USA. http://www.unl.edu/BIACO/NCME/Alignment%20revised.pdf

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

Marzano, R. J., & Kendall, J. S. (2001). *The new taxonomy of educational objectives*. Corwin.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276-282.

Näsström, G., & Henriksson, W. (2008). Alignment of standards and assessment: A theoretical and empirical study of methods for alignment. *Electronic Journal of Research in Educational Psychology*, *6*(3), 667-690.

Porter, A. C., & Smithson, J. L. (2001). *Defining, developing, and using curriculum indicators. CPRE research report series RR-048*. University of Pennsylvania Graduate School of Education.

Portney, L. G., & Watkins, M. P. (2015). *Foundations of clinical research: Applications to practice* (3rd ed.). Davis Company.

Resnick, L. B., Rothman R., Slattery, J. B., & Vranek, J. L. (2004). Benchmarking and alignment of standards and testing. *Educational Assessment, Evaluation and Accountability*, *9*(1-2), 1–27. https://doi.org/10.1080/10627197.2004.9652957

Rosnow, R. L., & Rosenthal, R. (1991). If you're looking at the cell means, you're not looking at only the interaction (unless all main effects are zero). *Psychological Bulletin*, *110*(3), 574-576.

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). Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California. https://eric.ed.gov/?id=ED466642

Webb, N. L. (1997a). *Criteria for alignment of expectations and assessments in mathematics and science education* (Research Monograph No. 8). Council of Chief State School Officers.

Webb, N. L. (1997b). Determining alignment of expectations and assessments in mathematics and science education. *NISE Brief*, *1*(2), 1-10.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states* (Research Monograph No. 18). Council of Chief State School Officers.

Webb, N. L. (2002). *An analysis of the alignment between mathematics standards and assessments for three states* [Paper presentation]. American Educational Research Association, New Orleans, USA.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education*, *20*(1), 7-25.