DATA-DRIVEN ITEM ANALYSIS: การประเมินคุณภาพข้อสอบแบบเลือกตอบด้วยโปรแกรม CITAS*

DATA-DRIVEN ITEM ANALYSIS: EVALUATING THE QUALITY OF MULTIPLE CHOICE QUESTIONS USING THE CITAS PROGRAM

ปุริมปรัชญ์ คณิณพศุตย์
Purimpratch Khaninphasut
สำนักทะเบียนและวัดผล มหาวิทยาลัยสุโขทัยธรรมาธิราช
Office of Registration, Records and Evaluation, Sukhothai Thammathirat Open University
Corresponding Author E-mail: purim.orestou@gmail.com

บทคัดย่อ

บทความนี้มีวัตถุประสงค์เพื่ออธิบายวิธีการวิเคราะห์คุณภาพข้อสอบแบบเลือกตอบตามทฤษฎีการทดสอบแบบดั้งเดิม โดยใช้โปรแกรม CITAS เป็นเครื่องมือในการวิเคราะห์ข้อมูล พร้อมทั้งนำเสนอแนวทางการแปลความหมายของค่าสถิติที่ได้จากการวิเคราะห์ การวิเคราะห์ครอบคลุมทั้งระดับรายข้อ ได้แก่ ค่าความยาก ค่าอำนาจจำแนก และประสิทธิภาพของตัวลวง รวมถึงรายฉบับ ได้แก่ ค่าความเที่ยง และความคลาดเคลื่อนมาตรฐานของการวัด ซึ่งเป็นดัชนีสำคัญที่สะท้อนคุณภาพของแบบทดสอบโดยรวม โปรแกรม CITAS (Classical Item and Test Analysis Spreadsheet) เป็นเครื่องมือที่ใช้งานผ่านโปรแกรม Microsoft Excel ช่วยให้การคำนวณค่าสถิติเป็นไปอย่างรวดเร็ว ลดความคลาดเคลื่อนจากการคำนวณด้วยมือ และเหมาะสมกับการใช้งานในบริบทการประเมินระดับชั้นเรียนหรือการทดลองใช้เครื่องมือที่มีกลุ่มตัวอย่างขนาดเล็ก อย่างไรก็ตาม การใช้โปรแกรมวิเคราะห์เป็นเพียงเครื่องมือสนับสนุนการตัดสินใจ ผู้สอนและนักวิจัยจำเป็นต้องมีความเข้าใจหลักการของการวิเคราะห์ข้อสอบและการแปลความหมายของค่าสถิติอย่างถูกต้อง เพื่อให้สามารถพิจารณาปรับปรุง ตัดทิ้ง หรือคงข้อสอบแต่ละข้อได้อย่างเหมาะสม อันจะนำไปสู่การพัฒนาแบบทดสอบที่มีคุณภาพบนพื้นฐานของหลักฐานเชิงประจักษ์อย่างมีประสิทธิภาพ

คำสำคัญ: ข้อสอบแบบเลือกตอบ; การวิเคราะห์ข้อสอบ; คุณภาพข้อสอบ; โปรแกรม CITAS

*Received February 9, 2026; Revised February 24, 2026; Accepted February 26, 2026

Abstract

This article aims to explain how to analyze the quality of multiple-choice questions based on classical test theory, using the CITAS program, while providing guidelines for interpreting the statistical values obtained from the analysis. The analysis encompasses both item-level indices, including difficulty index, discrimination index, and distractor efficiency, as well as test-level indices, including reliability and standard error of measurement, which are key indicators reflecting the overall quality of the test. CITAS is a tool that operates through Microsoft Excel, facilitating rapid statistical computation, reducing errors from human calculation, and proving suitable for use in classroom assessment or tryouts with small samples. However, the use of analytical software serves merely as a decision-support tool. Teachers and researchers must possess a sound understanding of the principles of item analysis and the proper interpretation of statistical values in order to make appropriate decisions regarding the revision, elimination, or retention of test items. This will lead to the development of high-quality tests grounded in empirical evidence of effectiveness.

Keywords: Multiple Choice Questions; Item Analysis; Item Quality; CITAS Program

บทนำ

การประเมินผลการเรียนรู้ถือเป็นส่วนสำคัญของกระบวนการจัดการเรียนการสอน โดยข้อสอบแบบเลือกตอบ (Multiple Choice Questions: MCQs) เป็นเครื่องมือที่ได้รับความนิยมอย่างแพร่หลายในทุกระดับการศึกษา เนื่องจากสามารถวัดผลสัมฤทธิ์ทางการเรียนได้อย่างรวดเร็ว ครอบคลุมเนื้อหา และมีความเป็นปรนัยในการตรวจให้คะแนน (Haladyna, 2004; Coughlin & Featherstone, 2017) ข้อสอบที่มีคุณภาพจะต้องผ่านกระบวนการวิเคราะห์และปรับปรุงอย่างเป็นระบบ เพื่อให้มั่นใจว่าสามารถวัดความสามารถของผู้เรียนได้อย่างแท้จริง การวิเคราะห์คุณภาพข้อสอบจึงมีประโยชน์ต่อครูผู้สอนในการปรับปรุงการเรียนการสอนและข้อสอบให้มีประสิทธิภาพมากขึ้นต่อนักวิจัยในการพัฒนาเครื่องมือวัดที่มีคุณภาพ และต่อผู้สร้างแบบทดสอบในการผลิตข้อสอบที่สามารถจำแนกความสามารถของผู้สอบได้อย่างเหมาะสม (Crocker & Algina, 2008)

ด้วยทฤษฎีการทดสอบแบบดั้งเดิม (Classical Test Theory: CTT) ได้รับความนิยมและใช้กันอย่างแพร่หลาย โดยเฉพาะในบริบทของการประเมินในชั้นเรียน (Classroom Assessment) (Hambleton & Jones, 1993) แม้ว่า ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) จะมีข้อได้เปรียบในเชิงทฤษฎีและให้ข้อมูลที่ละเอียดมากกว่าทฤษฎีการทดสอบแบบดั้งเดิม แต่เมื่อนำมาใช้วิเคราะห์ในระดับชั้นเรียนมักประสบปัญหาในทางปฏิบัติ เนื่องจากต้องการตัวอย่างขนาดใหญ่ มีความซับซ้อนในการคำนวณและต้องมีความเชี่ยวชาญเฉพาะทางในการแปลผลการวิเคราะห์ข้อมูล (Fan, 1998; Thompson, 2016) ในขณะที่ทฤษฎีการทดสอบแบบดั้งเดิมสามารถนำไปใช้ได้กับกลุ่มตัวอย่างขนาดเล็ก คำนวณง่าย และให้ข้อมูลที่เพียงพอสำหรับการปรับปรุงข้อสอบทำให้เหมาะสมกับครูผู้สอนและนักวิจัยที่ทำงานในบริบทการศึกษาทั่วไป

การวิเคราะห์คุณภาพข้อสอบตามทฤษฎีการทดสอบแบบดั้งเดิม ประกอบด้วยดัชนีสำคัญหลายตัว ได้แก่ ค่าความยากง่ายของข้อสอบ (Item Difficulty) ค่าอำนาจจำแนก (Item Discrimination) ประสิทธิภาพของตัวเลือกลวง (Distractor Analysis) ค่าความเที่ยง (Reliability) (Nitko & Brookhart, 2011; Allen & Yen, 1979) การคำนวณดัชนีเหล่านี้สามารถทำได้ทั้งแบบคำนวณด้วยมือ และการใช้โปรแกรมคอมพิวเตอร์ ในยุคปัจจุบันที่เทคโนโลยีมีบทบาทสำคัญต่อการศึกษา การใช้โปรแกรมคอมพิวเตอร์ช่วยให้การวิเคราะห์มีความรวดเร็ว แม่นยำ และลดข้อผิดพลาดจากการคำนวณด้วยมือ โปรแกรม CITAS (Classical Item and Test Analysis Software) เป็นหนึ่งในเครื่องมือที่ออกแบบมาเพื่อการวิเคราะห์คุณภาพข้อสอบตามทฤษฎีการทดสอบแบบดั้งเดิมโดยเฉพาะ (Thompson, 2016) โปรแกรมนี้มีจุดเด่นในการใช้งานที่ไม่ซับซ้อน เหมาะสมกับครูผู้สอนที่ต้องการวิเคราะห์ข้อสอบในชั้นเรียน และนำเสนอค่าสถิติที่เกี่ยวข้องกับการวิเคราะห์ข้อสอบอย่างครอบคลุม รวมถึงการแสดงผลในรูปแบบกราฟที่ช่วยให้เข้าใจภาพรวมของข้อสอบได้ง่าย

บทความวิชาการนี้จึงมีจุดมุ่งหมายเพื่ออธิบายการวิเคราะห์คุณภาพข้อสอบแบบเลือกตอบโดยใช้โปรแกรม CITAS ควบคู่กับการนำเสนอหลักการและสูตรคำนวณทางสถิติที่อยู่เบื้องหลังค่าดัชนีต่าง ๆ เพื่อให้ผู้อ่านเข้าใจที่มาและความหมายของตัวเลขมากกว่าการรายงานค่าที่ได้จากโปรแกรมเพียงอย่างเดียว นอกจากนี้ ยังมุ่งเสนอแนวทางการแปลผลและการนำเสนอผลการวิเคราะห์ในรายงานผลการทดลองใช้เครื่องมือและการประเมินคุณภาพข้อสอบในการประเมินระดับชั้นเรียน เพื่อสนับสนุนการประกันคุณภาพการวัดและประเมินผลทางการศึกษาอย่างมีหลักฐานเชิงประจักษ์และนำไปใช้ได้จริงในทางปฏิบัติ

การวิเคราะห์คุณภาพข้อสอบด้วยโปรแกรม CITAS

โปรแกรม Classical Item and Test Analysis Spreadsheet (CITAS) เป็นโปรแกรมวิเคราะห์คุณภาพข้อสอบแบบเลือกตอบหรือข้อสอบที่ตรวจให้คะแนนแบบ 2 ค่า คือ 0 (ตอบผิด) กับ 1 (ตอบถูก) ที่วิเคราะห์ผ่าน Microsoft Excel® spreadsheet พัฒนาโดยบริษัท Assessment Systems เป็นโปรแกรมที่ไม่เสียค่าใช้จ่าย สามารถดาวน์โหลดได้ที่ลิงก์ https://assess.com/citas/ โปรแกรมนี้นำเสนอค่าดัชนีที่แสดงคุณภาพข้อสอบตามทฤษฎีการทดสอบแบบดั้งเดิม (Classical Test Theory)

ดัชนีคุณภาพข้อสอบที่โปรแกรมนำเสนอมีทั้งผลการวิเคราะห์คุณภาพข้อสอบรายข้อ (Item) และรายฉบับ (Test) ก่อนที่จะไปศึกษาตัวอย่างการใช้โปรแกรม CITAS และการแปลผลการวิเคราะห์คุณภาพข้อสอบ ผู้เขียนขอนำเสนอที่มาของสูตรสถิติที่โปรแกรมนี้นำมาเป็นพื้นฐานในการคำนวณเพื่อให้ผู้อ่านเข้าใจชัดเจนยิ่งขึ้น

1. การวิเคราะห์คุณภาพข้อสอบรายข้อ

การวิเคราะห์คุณภาพข้อสอบรายข้อ มีเป้าหมายหลักเพื่อนำค่าสถิติมาใช้พิจารณาข้อบกพร่องที่อาจเกิดขึ้นกับข้อสอบแต่ละข้อ เพื่อนำไปสู่การตัดสินใจว่า ควรจะปรับปรุง (Revise) เปลี่ยนใหม่ (Replace) หรือตัดทิ้ง (Retire) เช่น ตัวลวงไม่มีประสิทธิภาพเนื่องจากดึงดูดผู้สอบกลุ่มเก่งให้เลือกตอบ หรือข้อสอบมีความยากเกินระดับการเรียนรู้ของนักเรียน สำหรับดัชนีแสดงคุณภาพข้อสอบรายข้อ คือ ค่าความยาก (Item Difficulty) ค่าอำนาจจำแนก (Item Discrimination) และประสิทธิภาพตัวลวง (Distractor Efficiency)

1.1 ความยาก

ค่าความยากของข้อสอบ (Item Difficulty) หมายถึง สัดส่วนของนักเรียนที่ตอบข้อสอบข้อนั้นถูก เรียกว่า ค่า P (P-value) เช่น ถ้าค่า P = 0.95 หมายความว่า ถ้านักเรียน 95% ตอบถูก แสดงว่า ข้อสอบนั้นง่ายมาก เพราะมีนักเรียนถึง 95% ที่ทำข้อสอบข้อนั้นได้ถูกต้อง แต่ถ้าค่า P = 0.30 หมายความว่า นักเรียนเพียง 30% เท่านั้นที่ตอบถูก แสดงว่า ข้อสอบยากมาก โดยเฉพาะอย่างยิ่งเมื่อพิจารณาว่าข้อสอบแบบเลือกตอบ 4 ตัวเลือก ผู้สอบมีโอกาสเดาคำตอบที่ถูกต้องได้ถึง 25% ดังนั้น การที่นักเรียนตอบถูก 30% เมื่อหักโอกาสที่จะเดาถูก 25% ออก จะเหลือผู้สอบที่รู้จริงแค่ 5% เท่านั้น ข้อสอบที่มีค่า P น้อยกว่า 0.50 จึงถูกมองว่ายากมาก เพราะว่าเมื่อหักโอกาสที่จะเดาถูกออกจะเหลือผู้ที่รู้จริงไม่เกิน 25% (Thompson, 2016)

โดยทั่วไปมักพบค่า P อยู่ในช่วง 0.70 ถึง 0.80 ซึ่งสูงกว่าการเดามาก แสดงว่าข้อสอบมีความยากพอเหมาะ ไม่ง่ายจนเกินไป ทำให้นักเรียนที่มีความรู้จริงจะตอบถูก (Thompson, 2016) ทั้งนี้ แบบทดสอบบางชนิดอาจให้ค่า P อยู่ในช่วงที่แตกต่างออกไป

ซึ่งอาจจำเป็นต้องปรับเกณฑ์การพิจารณาให้เหมาะสมกับบริบทของการนำแบบทดสอบไปใช้ สำหรับประเทศไทยนิยมใช้เกณฑ์ค่า P อยู่ในช่วง 0.2 ถึง 0.8 (ศิริชัย กาญจนวาสี, 2556)

1.2 อำนาจจำแนก

ค่าอำนาจจำแนก (Item Discrimination) หมายถึง ความสามารถของข้อสอบในการแยกระหว่างผู้สอบที่มีระดับความรู้หรือความสามารถสูงออกจากผู้สอบที่มีระดับความรู้หรือความสามารถต่ำ โปรแกรมนี้จะหาค่าอำนาจจำแนกโดยใช้สูตรสหสัมพันธ์พอยต์-ไบซีเรียล (Point-biserial Correlation หรือ r_{pbis}) ซึ่งเป็นการหาความสัมพันธ์ระหว่างคะแนนรายข้อ(ที่ให้คะแนน 0 หรือ 1) กับคะแนนรวมของแบบทดสอบ เรียกว่า Item-total Correlation ค่าสถิตินี้ทำหน้าที่บ่งชี้ว่า นักเรียนที่ตอบข้อนั้นถูก มีแนวโน้มได้คะแนนรวมสูงหรือไม่ ซึ่งถือเป็นลักษณะสำคัญของข้อสอบที่มีคุณภาพดี โดยคำนวณจากสูตร (Crocker & Algina, 1986)

$$r_{pbis} = \frac{\bar{X}_i - \bar{X}_t}{S_t}\sqrt{\frac{p_i}{1-p_i}}$$

โดยที่ $\bar{X}_i$ แทน ค่าเฉลี่ยของคะแนนรวมของนักเรียนที่ตอบข้อสอบข้อที่ i ถูก

$\bar{X}_t$ แทน ค่าเฉลี่ยของคะแนนรวมของนักเรียนทั้งหมด

S_t แทน ส่วนเบี่ยงเบนมาตรฐานของคะแนนรวมของนักเรียนทั้งหมด

p_i แทน ค่าความยากของข้อสอบข้อที่ i

ค่าอำนาจจำแนกอยู่ในช่วง –1.0 ถึง 1.0 ข้อสอบที่ดีควรสามารถแยกผู้สอบที่มีความสามารถสูงออกจากผู้ที่มีความสามารถต่ำได้ และควรมีค่าพอยต์ไบซีเรียล (Point-Biserial) ค่อนข้างสูง อยู่ในช่วง 0.50 – 0.60 ค่าพอยต์ไบซีเรียลติดลบ หมายถึง ข้อสอบข้อนั้นมีค่าอำนาจจำแนกไม่ดีมาก ๆ เพราะผู้สอบที่มีความสามารถสูงกลับตอบผิด ในขณะที่ผู้สอบที่มีความสามารถต่ำกลับตอบถูก ซึ่งเป็นสถานการณ์ที่ผิดปกติอย่างยิ่ง ซึ่งอาจเกิดจากการเฉลยข้อสอบผิด คำถามมีความกำกวม สับสน หรือวัดคนละทักษะกับที่ต้องการวัด เกิดปัญหาในการให้คะแนนหรือการบันทึกข้อมูล ดังนั้น ข้อสอบข้อนั้นควรปรับปรุงแก้ไขหรือตัดทิ้งไป ส่วนค่าพอยต์ไบซีเรียลเท่ากับ 0.00 หมายถึง ข้อสอบไม่สามารถแยกระหว่างผู้สอบที่ได้คะแนนต่ำกับผู้สอบที่ได้คะแนนสูง ผู้สอบทั้งสองกลุ่มตอบถูกพอ ๆ กัน โดยทั่วไปค่าอำนาจจำแนก หรือ r_{pbis} ที่ใช้ได้ควรมีค่าเป็นบวกตั้งแต่ 0.2 ขึ้นไป (Thompson, 2016; ศิริชัย กาญจนวาสี, 2556)

1.3 ประสิทธิภาพตัวลวง

การประเมินว่า ตัวลวง (distractors) ทำหน้าที่ได้ดีหรือไม่นั้น หลักการ คือ ตัวลวงต้องลวงคนที่มีความสามารถต่ำมาตอบมากกว่าคนที่มีความสามารถสูง แต่ถ้ากลับทิศกัน แสดงว่า ข้อสอบนั้นมีปัญหาเกิดขึ้น ซึ่งอาจเกิดจากเฉลยผิด หรือตัวลวงนั้นอาจจะเป็นคำตอบที่ถูกอีกหนึ่งคำตอบ สำหรับการพิจารณาประสิทธิภาพของตัวลวงจะพิจารณาจากค่า r_{pbis} เช่นกัน โดยค่า r_{pbis} ของตัวลวงควรมีค่าเป็นลบ ขณะที่ค่า r_{pbis} ของตัวถูก ควรเป็นค่าเป็นบวก

หาก r_{pbis} ของตัวลวงเป็นบวก แสดงว่า ผู้สอบที่มีความสามารถสูงเลือกตัวลวงนั้น ทั้งที่โดยหลักการแล้ว เราต้องการให้ผู้สอบที่มีความสามารถต่ำเป็นผู้เลือกคำตอบที่ไม่ถูกต้อง นอกจากนี้ ในการพิจารณา ค่า P ของคำตอบถูก (key) ควรสูงกว่าค่า P ของตัวลวง (distractors) ทุกตัว กล่าวคือ เราไม่ต้องการให้มีผู้สอบเลือกตัวลวงใดมากกว่าคำตอบที่ถูกต้องแต่ในบางกรณี อาจเกิดจากตัวลวงนั้นเป็นตัวถูกอีกตัวหนึ่ง หรือคำตอบที่เฉลยไว้อาจไม่ถูกต้องจริง ๆ (Thompson, 2016)

2. การวิเคราะห์คุณภาพข้อสอบรายฉบับ

ดัชนีที่สำคัญในการบอกคุณภาพของแบบทดสอบรายฉบับ คือ ค่าความเที่ยง (Reliability) และความคลาดเคลื่อนมาตรฐานของการวัด (Standard Error of Measurement: SEM) โปรแกรม CITAS ได้นำเสนอสองค่านี้ โดยค่าความเที่ยงโปรแกรมใช้สูตร KR-20 เนื่องจากข้อสอบมีการให้คะแนนแบบ 2 ค่า คือ 0 กับ 1

ค่าความเที่ยง เป็นการแสดงถึงความคงเส้นคงวา (Consistency) หรือ ความสามารถในการวัดซ้ำ (Repeatability) ของแบบทดสอบ หากแบบทดสอบนั้นให้คะแนนคงที่หรือสม่ำเสมอ แสดงว่า แบบทดสอบนั้นมีความเที่ยง คำว่า “ความคงเส้นคงวา” หมายถึง หากนักเรียนคนหนึ่งมีคะแนนที่แท้จริง (True Score) อยู่ที่ 45 คะแนน จากคะแนนเต็ม 100 คะแนน ถ้านักเรียนคนนี้ทำแบบทดสอบฉบับเดิมซ้ำหลายครั้ง แล้วได้คะแนนเป็น 44 หรือ 43 ถือว่า แบบทดสอบมีความคงเส้นคงวา เพราะมีค่าใกล้เคียงกับคะแนนที่แท้จริง แต่ถ้าคะแนนที่ได้เป็น 34 แล้วครั้งต่อมาได้ 47 และครั้งถัดไปได้ 39 ถือว่า แบบทดสอบไม่มีความเที่ยง ค่าความเที่ยงจะอยู่ในช่วง 0 ถึง 1 เกณฑ์ในการระบุว่าแบบทดสอบฉบับนั้นมีความเที่ยงที่ยอมรับได้ คือ KR-20 ต้องมีค่า 0.7 ขึ้นไป (Salkind, 2010; Thompson, 2016)

ค่าความเที่ยงมีความสัมพันธ์กับความคลาดเคลื่อนมาตรฐานของการวัด (SEM) เพราะเป็นการนำแนวคิดเรื่องความคงเส้นคงวาของการวัดมาประยุกต์ใช้กับคะแนนของผู้สอบแต่ละคน SEM เป็นค่าที่แสดงว่า การวัดมีความแม่นยำหรือไม่ กล่าวคือ คะแนนที่ผู้สอบสอบได้ใกล้เคียงกับความรู้จริงหรือคะแนนจริง (True Score) มากน้อยแค่ไหน โดยการคำนวณค่า SEM เราจะนำค่าความเที่ยงแทนค่าไปในสูตร

$$SEM = S_X\sqrt{1 - r_{X\acute{X}}}$$

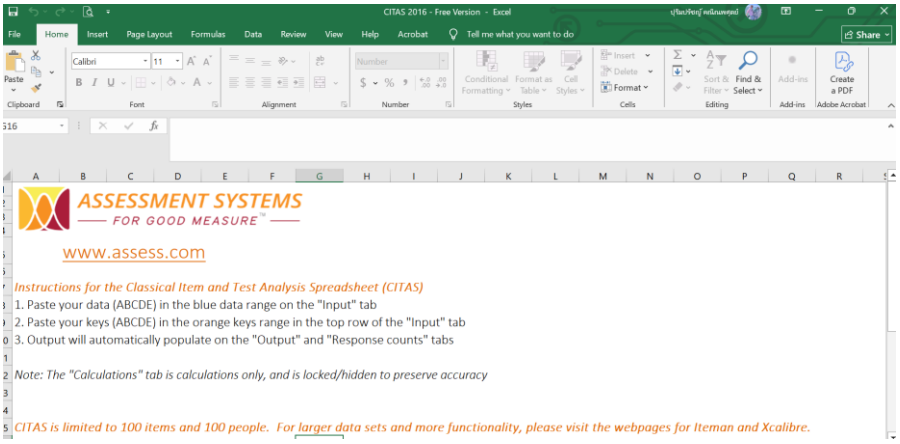
โดยที่ S_X แทน ส่วนเบี่ยงเบนมาตรฐานของคะแนนรวมทั้งฉบับ

$r_{X\acute{X}}$ แทน ค่าความเที่ยงของแบบทดสอบ ในที่นี้ คือ KR-20

การแปลความหมายของค่า SEM คือ ถ้าค่า SEM น้อย หมายถึง KR-20 หรือ $r_{X\acute{X}}$ ในสูตร มีค่าเข้าใกล้ 1 มาก จะทำให้ได้ค่าที่มาจากการถอดรากที่สองน้อยมาก แสดงว่าผลการวัดมีความแม่นยำสูง คะแนนที่สอบได้ใกล้เคียงกับความรู้จริง แต่ถ้าค่า SEM มาก แสดงว่า ผลการวัดขาดความแม่นยำ คะแนนที่สอบได้แตกต่างจากความรู้จริง

3. การใช้โปแกรม CITAS

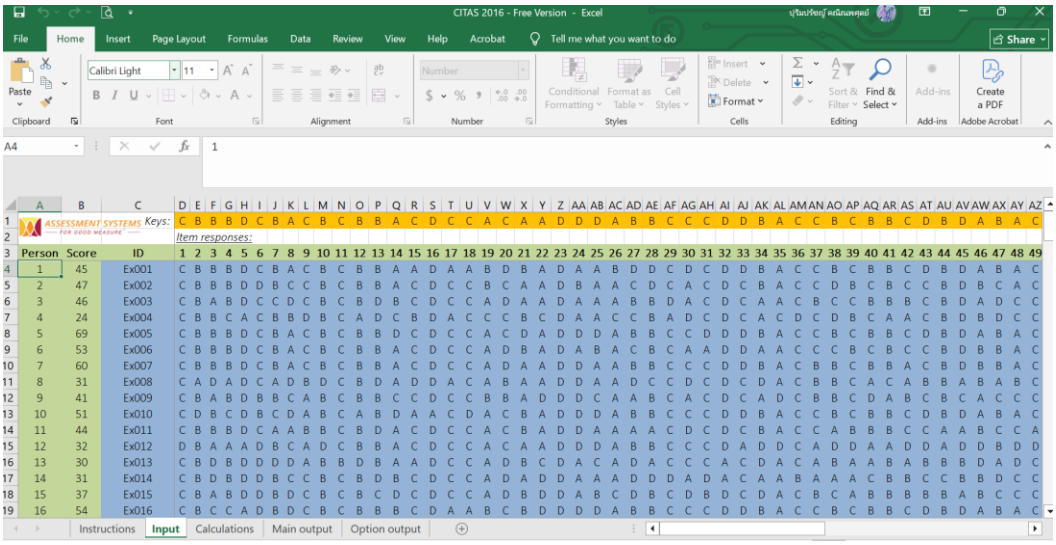
โปรแกรม CITAS เป็นโปรแกรมที่ใช้งานง่าย เพราะวิเคราะห์ผ่านโปรแกรม Microsoft Excel ที่ครูผู้สอนหรือนักวิจัยคุ้นเคย และไม่ต้องตรวจข้อสอบรายข้อให้มีค่าเป็น 0 กับ 1 ก่อนนำไปวิเคราะห์ ข้อจำกัดของโปรแกรมนี้ คือ สามารถวิเคราะห์ข้อสอบได้ 100 ข้อ และจำนวนผู้สอบไม่เกิน 100 คนเท่านั้น แต่สำหรับการสอนในระดับชั้นเรียนหรือการทดลองใช้เครื่องมือ (tryout) ที่ใช้กลุ่มตัวอย่างไม่มาก ครูผู้สอนหรือนักวิจัยสามารถใช้วิเคราะห์ได้สะดวกและประหยัดเวลา เพราะโปรแกรมนี้จะนำเสนอค่าสถิติให้กับครูผู้สอนหรือนักวิจัยนำไปเขียนรายงานได้โดยตรง สำหรับตัวอย่างข้อมูลและการรายงานผลวิเคราะห์คุณภาพข้อสอบ ผู้เขียนขอใช้ตัวอย่างข้อมูลที่อยู่ในโปรแกรม (Thomson, 2016) เพื่อให้สอดคล้องกับคู่มือภาษาอังกฤษ หากผู้อ่านต้องการศึกษาเพิ่มเติม



ภาพที่ 1 หน้าจอและคำชี้แจงของโปรแกรม CITAS
ที่มา: Thompson (2019), *CITAS* software

วารสารพุทธนวัตกรรมและการจัดการ
ปีที่ 9 ฉบับที่ 1 มกราคม – กุมภาพันธ์ 2569 น. 281-295.

การวิเคราะห์ข้อมูล เริ่มจากพิมพ์หรือ Copy ข้อมูลผลการตอบข้อสอบแต่ละข้อของนักเรียน โดยกรอกเป็นตัวเลือกที่นักเรียนเลือกตอบ และกรอกเฉลยคำตอบ (Keys) ลงในแท็บ “Input” ดังภาพที่ 2



ภาพที่ 2 ตัวอย่างการกรอกข้อมูลในโปรแกรม CITAS
ที่มา: Thompson, (2019), CITAS software

การรายงานค่าสถิติต่าง ๆ ในโปรแกรม CITAS จะรายงานทั้งค่าคุณภาพข้อสอบรายข้อและรายฉบับ กำหนด NC (Number-Correct) คือ จำนวนข้อที่ตอบถูก ดังตารางที่ 1

ตารางที่ 1 ความหมายของค่าสถิติต่าง ๆ ในโปรแกรม CITAS

ค่าสถิติรายฉบับ (Test-level statistics)	ค่าสถิติรายข้อ (Item Statistics)
Number of examinees (จำนวนผู้สอบ)	P (ค่าความยากของข้อสอบแต่ละข้อ)
Number of items (จำนวนข้อสอบ)	r_{pbis} (ค่าอำนาจจำแนกของข้อสอบแต่ละข้อ)
NC score mean (ค่าเฉลี่ยของคะแนนของจำนวนข้อที่ตอบถูก)	Number correct (จำนวนข้อที่ตอบถูก)
NC score standard deviation (ส่วนเบี่ยงเบนมาตรฐานของคะแนนของจำนวนข้อที่ตอบถูก)	Number incorrect (จำนวนข้อที่ตอบผิด)
NC score variance (ความแปรปรวนของคะแนนของจำนวนข้อที่ตอบถูก)	Mean score correct (ค่าเฉลี่ยของคะแนนคนที่ตอบถูก)
Minimum NC score (ค่าต่ำสุดของคะแนนของจำนวนข้อที่ตอบถูก)	Mean score incorrect (ค่าเฉลี่ยของคะแนนคนที่ตอบผิด)

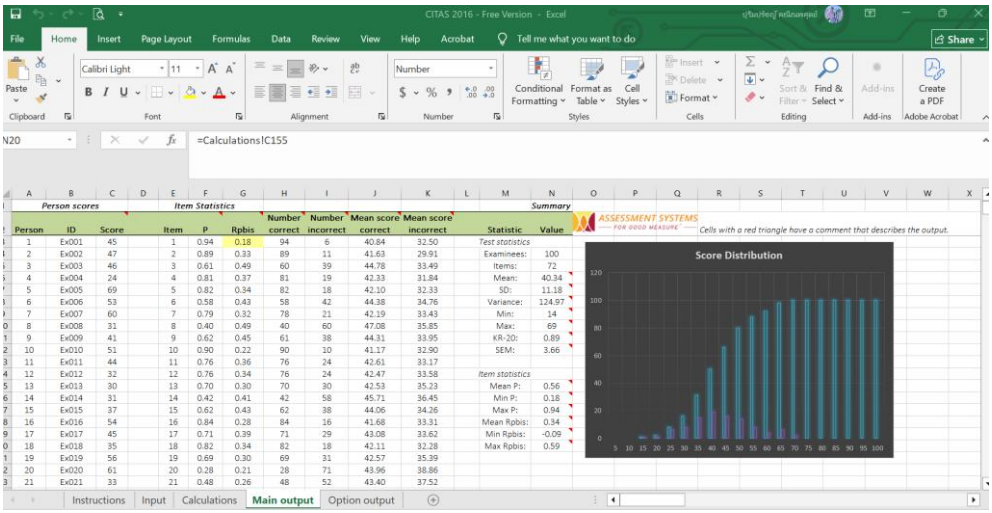
ตารางที่ 1 ความหมายของค่าสถิติต่าง ๆ ในโปรแกรม CITAS (ต่อ)

ค่าสถิติรายฉบับ (Test-level statistics)	ค่าสถิติรายข้อ (Item Statistics)
Maximum NC score (ค่าสูงสุดของคะแนนของจำนวนข้อที่ตอบถูก)	Mean P (ค่าเฉลี่ยของความยากทั้งฉบับ)
KR-20 (alpha) reliability (ความเที่ยง)	Minimum P (ค่าต่ำสุดของความยาก)
Standard error of measurement (SEM) (ความคลาดเคลื่อนมาตรฐานของการวัด)	Maximum P (ค่าสูงสุดของความยาก)
	Mean r_{pbis} (ค่าเฉลี่ยของค่าอำนาจจำแนกทั้งฉบับ)
	Minimum r_{pbis} (ค่าต่ำสุดของค่าอำนาจจำแนก)
	Maximum r_{pbis} (ค่าสูงสุดของค่าอำนาจจำแนก)

หมายเหตุ ในที่นี้ที่ค่าความเที่ยง KR-20 (alpha) หมายถึง กรณีที่ข้อสอบเป็นคะแนนแบบ 2 ค่า คือ (0, 1) ค่าความเที่ยงสูตร KR-20 จะมีค่าเท่ากับ Cronbach's alpha

4. การแปลผลข้อมูลการวิเคราะห์ด้วยโปรแกรม CITAS

ข้อมูลตัวอย่างที่ปรากฏในโปรแกรม ประกอบด้วย ข้อสอบ จำนวน 72 ข้อ ผู้สอบจำนวน 100 คน แบบทดสอบมีลักษณะแบบเลือกตอบ 5 ตัวเลือก โดยดูผลการวิเคราะห์ได้ที่แท็บ "Main Output" ดังภาพที่ 3



ภาพที่ 3 ผลการวิเคราะห์คุณภาพข้อสอบในแท็บ "Main Output"
ที่มา: Thompson, (2019), CITAS software

วารสารพุทธนวัตกรรมและการจัดการ
ปีที่ 9 ฉบับที่ 1 มกราคม – กุมภาพันธ์ 2569 น. 281-295.

Statistic	Value
Test statistics	
Examinees:	100
Items:	72
Mean:	40.34
SD:	11.18
Variance:	124.97
Min:	14
Max:	69
KR-20:	0.89
SEM:	3.66
Item statistics	
Mean P:	0.56
Min P:	0.18
Max P:	0.94
Mean Rpbis:	0.34
Min Rpbis:	-0.09
Max Rpbis:	0.59

การแปลความหมาย Test Statistics

แบบทดสอบฉบับนี้มีจำนวน 72 ข้อ มีนักเรียนเข้าสอบทั้งหมด 100 คน แบบทดสอบมีคะแนนเฉลี่ย เท่ากับ 40.32 คะแนน จากคะแนนเต็ม 72 คะแนน คะแนนมีการกระจายค่อนข้างมาก โดยมีส่วนเบี่ยงเบนมาตรฐาน (SD) เท่ากับ 11.01 และมีช่วงคะแนนตั้งแต่ 14 ถึง 69 คะแนน แบบทดสอบฉบับนี้ยังมีความเที่ยงที่เหมาะสม โดยมีค่า KR-20 เท่ากับ 0.89 และค่าความคลาดเคลื่อนมาตรฐานของการวัด (SEM) เท่ากับ 3.66

Statistic	Value
Test statistics	
Examinees:	100
Items:	72
Mean:	40.34
SD:	11.18
Variance:	124.97
Min:	14
Max:	69
KR-20:	0.89
SEM:	3.66
Item statistics	
Mean P:	0.56
Min P:	0.18
Max P:	0.94
Mean Rpbis:	0.34
Min Rpbis:	-0.09
Max Rpbis:	0.59

การแปลความหมาย Item Statistics

ค่าความยาก (P) : ค่าความยากเฉลี่ยของแบบทดสอบทั้งฉบับเท่ากับ 0.56 หมายถึง โดยเฉลี่ยมีคนตอบถูกประมาณ 56% ซึ่งถือว่าเป็นข้อสอบที่อยู่ในระดับปานกลางค่อนไปทางยาก ข้อสอบยากที่สุดมีคนตอบถูกเพียง 18% และข้อสอบง่ายที่สุด มีคนตอบถูกถึง 94%

ค่าอำนาจจำแนก (Rpbis) : ค่าอำนาจจำแนกเฉลี่ยเท่ากับ 0.34 ถือว่าอยู่นำไปใช้ได้ (ผ่านเกณฑ์มากกว่า 0.20) แสดงว่า ข้อสอบส่วนใหญ่ทำหน้าที่แยกคนเก่งกับคนอ่อนออกจากกันได้เหมาะสม ข้อสอบที่มีค่าอำนาจจำแนกต่ำสุด และติดลบ (-0.09) หมายความว่า คนเก่งตอบผิดแต่คนอ่อนกลับตอบถูก ข้อนี้ควรตัดทิ้งหรือปรับปรุง ส่วนข้อสอบข้อที่มีอำนาจจำแนกสูงสุด (0.59) ถือว่าเป็นข้อสอบที่ดีมากในการจำแนกคนเก่งออกจากคนอ่อน

จากการแปลผลการวิเคราะห์ดังกล่าว เป็นการแปลผลในภาพรวมของแบบทดสอบ เพื่อให้ผู้สอนหรือนักวิจัยเห็นภาพรวมทั้งหมด อย่างไรก็ตาม ในการพิจารณาคุณภาพของการวิเคราะห์ข้อสอบ เรามุ่งเน้นไปที่การพิจารณารายข้อเป็นหลัก เพื่อนำไปตัดสินใจว่าข้อสอบข้อใดที่ผ่านเกณฑ์นำไปใช้ได้ และข้อสอบข้อใดที่ไม่ผ่านเกณฑ์ ต้องมีการปรับปรุงข้อสอบก่อนนำไปใช้จริง ตัวอย่างต่อไปเป็นการแปลผลการวิเคราะห์คุณภาพข้อสอบรายข้อ ดังภาพที่ 4 และ 5

Person scores			*Item Statistics*						
Person	ID	Score	Item	P	Rpbis	Number correct	Number incorrect	Mean score correct	Mean score incorrect
1	Ex001	45	1	0.94	0.18	94	6	40.84	32.50
2	Ex002	47	2	0.89	0.33	89	11	41.63	29.91
3	Ex003	46	3	0.61	0.49	60	39	44.78	33.49
4	Ex004	24	4	0.81	0.37	81	19	42.33	31.84
5	Ex005	69	5	0.82	0.34	82	18	42.10	32.33
6	Ex006	53	6	0.58	0.43	58	42	44.38	34.76
7	Ex007	60	7	0.79	0.32	78	21	42.19	33.43
8	Ex008	31	8	0.40	0.49	40	60	47.08	35.85
9	Ex009	41	9	0.62	0.45	61	38	44.31	33.95
10	Ex010	51	10	0.90	0.22	90	10	41.17	32.90

ภาพที่ 4 ผลการวิเคราะห์คุณภาพข้อสอบรายข้อ ข้อที่ 1-10
ที่มา: Thompson (2019), CITAS software

จากภาพที่ 4 ข้อที่ 1 (item 1) เป็นข้อสอบที่ง่ายมาก เพราะมีค่า P เท่ากับ 0.94 และมีค่าอำนาจจำแนก (Rpbis) เท่ากับ 0.18 แต่อำนาจจำแนกที่เป็นบวกนี้ก็สะท้อนให้เห็นได้จากคะแนนเฉลี่ย โดยคะแนนเฉลี่ยของผู้สอบที่ตอบถูก (40.84) สูงกว่าคะแนนเฉลี่ยของผู้สอบที่ตอบผิด (32.50) แต่ก็ยังถือว่า เป็นข้อที่ควรปรับปรุง เพราะมีค่าความยากเกินเกณฑ์ที่กำหนด (ค่าความยากอยู่ระหว่าง 0.2 – 0.8) และมีค่าอำนาจจำแนกต่ำกว่าเกณฑ์ (ต่ำกว่า 0.2)

ข้อที่ 2 (item 2) เป็นข้อสอบที่ง่ายเช่นกัน แต่มีอำนาจจำแนกสูง (Rpbis) เท่ากับ 0.33 และมีความแตกต่างของคะแนนเฉลี่ยระหว่างผู้ที่ตอบถูก (41.63) กับตอบผิด (29.91) ค่อนข้างมาก แต่ข้อนี้ก็ยังคงต้องพิจารณาปรับปรุงข้อสอบ แม้ว่าค่าอำนาจจำแนกจะผ่านเกณฑ์ก็ตาม เพราะค่าความยากเกินเกณฑ์ที่กำหนด คือค่าความยากที่เหมาะสมควรอยู่ระหว่าง 0.2 – 0.8

ข้อที่ 3 (item 3) เป็นข้อสอบที่ใช้ได้ เพราะมีค่าความยากและค่าอำนาจจำแนกผ่านเกณฑ์ทั้งคู่ กล่าวคือ ค่าความยากค่อนไปทางง่าย (P = 0.61) และค่าอำนาจจำแนก (Rpbis) เท่ากับ 0.49 และมีความแตกต่างของคะแนนเฉลี่ยของผู้ที่ตอบถูก (44.78) กับตอบผิด (33.49) ค่อนข้างมาก ดังนั้น ข้อนี้สามารถนำไปเก็บข้อมูลจริงได้

Person scores			*Item Statistics*						
Person	ID	Score	Item	P	Rpbis	Number correct	Number incorrect	Mean score correct	Mean score incorrect
55	Ex055	41	55	0.54	0.41	54	46	44.57	35.37
56	Ex056	25	56	0.33	0.33	33	66	45.67	37.80
57	Ex057	56	57	0.69	0.47	69	31	43.86	32.52
58	Ex058	28	58	0.43	0.25	43	57	43.56	37.91
59	Ex059	31	59	0.20	0.13	20	80	43.15	39.64
60	Ex060	31	60	0.58	-0.09	57	42	39.33	41.33
61	Ex061	37	61	0.56	0.27	56	44	42.96	37.00
62	Ex062	61	62	0.48	0.23	48	52	43.04	37.85
63	Ex063	53	63	0.58	0.30	58	42	43.21	36.38

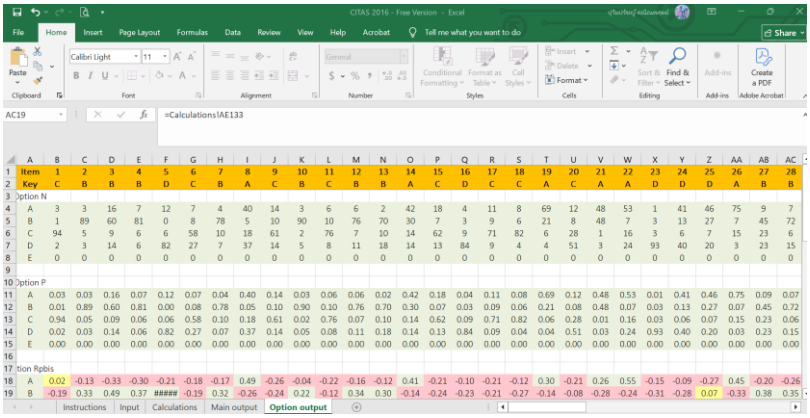
ภาพที่ 5 ผลการวิเคราะห์คุณภาพข้อสอบรายข้อ ข้อที่ 59-60
ที่มา: Thompson (2019), CITAS software

จากภาพที่ 5 ข้อที่ 59 (item 59) เป็นข้อสอบที่ยาก (P = 0.20) แต่ก็ยังผ่านเกณฑ์ขั้นต่ำของข้อสอบที่ใช้ได้ (ค่า P อยู่ระหว่าง 0.2-0.8) เมื่อพิจารณาค่าอำนาจจำแนก (Rpbis) เท่ากับ 0.13 ซึ่งยังคงมีค่าเป็นบวก แต่มีค่าต่ำกว่าเกณฑ์ 0.2 แสดงว่า ข้อนี้ยังจำแนกได้น้อย สังเกตได้จากคะแนนเฉลี่ยของผู้สอบที่ตอบถูก (43.15) สูงกว่าคะแนนเฉลี่ยของผู้สอบที่ตอบผิด (39.64) ดังนั้น ข้อนี้ควรปรับปรุง ก่อนนำไปใช้จริง

ข้อที่ 60 (item 60) เป็นข้อสอบที่มีความยากอยู่ในระดับปานกลางค่อนไปทางง่าย (P = 0.58) และผ่านเกณฑ์ของข้อสอบที่ใช้ได้ (ค่า P อยู่ระหว่าง 0.2-0.8) แต่เมื่อพิจารณาค่าอำนาจจำแนก (Rpbis) เท่ากับ -0.09 ซึ่งมีค่าเป็นลบ ข้อนี้ควรรีบตรวจสอบและปรับปรุงด่วน เพราะอาจมีตัวเลือกใดตัวเลือกหนึ่งที่เป็นตัวลวงที่ดึงดูดผู้สอบมากเกินไป ซึ่งอาจจะเป็นตัวถูกอีกตัวหนึ่ง หรือมีการเฉลยผิด สังเกตได้จากคะแนนเฉลี่ยของผู้สอบที่ตอบผิด (41.33) สูงกว่าคะแนนเฉลี่ยของผู้สอบที่ตอบผิด (39.33) ข้อนี้ควรปรับปรุง ก่อนนำไปใช้จริง

จากผลการวิเคราะห์ในโปรแกรม CITAS สังเกตว่า ผู้ออกแบบโปรแกรมใช้สีแดงและสีเหลือง แทน ค่าที่ยังไม่ผ่านเกณฑ์ โดย สีเหลือง แทน ค่าที่อำนาจจำแนกยังเป็นบวกแต่ไม่ผ่านเกณฑ์ (มากกว่า 0.2) และสีแดงแทนค่าที่อำนาจจำแนกที่ติดลบ บ่งบอกว่าต้องปรับปรุงด่วน

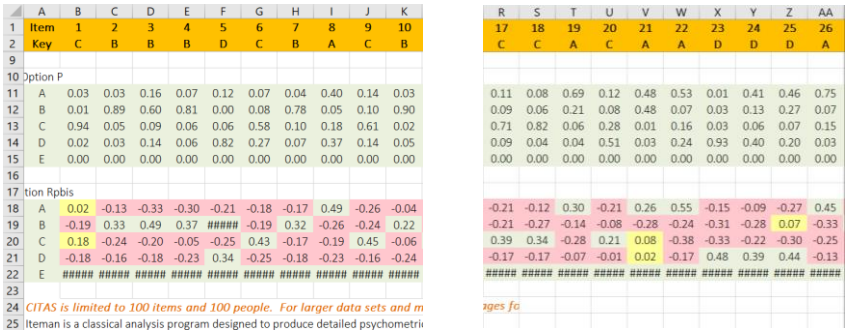
เมื่อพิจารณาคุณภาพข้อสอบรายข้อแล้ว ลำดับต่อมา คือ การพิจารณาการทำหน้าที่ของตัวลวง ว่าลวงคนอ่อนมาตอบได้หรือไม่ หรือมีผู้สอบเลือกตัวลวงนั้นหรือไม่ โดยดูผลการวิเคราะห์ได้ที่แท็บ “Option Output” ดังภาพที่ 6



ภาพที่ 6 ผลการวิเคราะห์คุณภาพตัวลวงในแท็บ “Option Output”
ที่มา: Thompson (2019), CITAS software

จากภาพที่ 6 เป็นหน้าต่างที่แสดงการวิเคราะห์ประสิทธิภาพตัวลวง โดยโปรแกรมนี้นำเสนอ 3 ค่า คือ จำนวนผู้สอบที่เลือกตัวเลือกแต่ละตัว ค่าสัดส่วนของจำนวนผู้สอบที่เลือกตัวเลือกแต่ละตัว (P) และค่าอำนาจจำแนก (Rpbis) ของตัวเลือกแต่ละตัว

การพิจารณาว่า ตัวลวงนั้นทำหน้าที่ได้ดีหรือไม่ ค่า Rpbis จะต้องติดลบ จึงจะถือว่าตัวลวงตัวนั้นใช้ได้ แต่ถ้ามีค่าเป็นบวก แสดงว่า ตัวลวงนี้ไปดึงดูดผู้สอบที่เก่งบางคนให้มาเลือกตอบ ดังนั้น ควรปรับปรุง ดังภาพที่ 7



ภาพที่ 7 ผลการวิเคราะห์คุณภาพตัวลวงในหัวข้อ Option Rpbis
ที่มา: Thompson (2019), CITAS software

จากภาพที่ 7 ข้อที่ 1 ตัวลวงที่ควรปรับปรุงคือ ตัวเลือก A เพราะมีค่า Rpbis เป็นบวก (0.02) แต่ถือว่ามีค่าน้อยมาก เนื่องจากมีผู้สอบเพียง 3 คนเท่านั้นที่เลือก A แม้ว่า

จะมีจำนวนผู้เลือก (N) น้อยก็ตาม ตัวลวงนี้อาจบังเอิญไปดึงดูดผู้สอบที่เก่งบางคนให้มาเลือกตอบ ซึ่งควรพิจารณาปรับปรุง แต่ในทางปฏิบัติหากพิจารณาแล้ว พบว่า ตัวลวงสอดคล้องกับข้อสอบและมาจากสิ่งที่นักเรียนเข้าใจผิด ก็อาจถือว่า ยังใช้ตัวลวงนี้ได้

ข้อ 21 ตัวลวงที่ควรปรับปรุง คือ ตัวเลือก C และ D เพราะมีค่า Rpbis เป็นบวก คือ 0.08 และ 0.02 ตามลำดับ

สรุป

การใช้โปรแกรมวิเคราะห์คุณภาพข้อสอบเป็นเพียงเครื่องมือที่ช่วยให้ครูผู้สอนหรือนักวิจัยทำงานสะดวกขึ้น และลดความคลาดเคลื่อนจากการคำนวณด้วยมือ อย่างไรก็ตาม ครูผู้สอนหรือนักวิจัยก็ควรมีความเข้าใจหลักการของการพิจารณาคุณภาพข้อสอบแบบเลือกตอบรวมทั้งการแปลความหมายของค่าสถิติต่าง ๆ ได้เป็นอย่างดี เพราะการที่ค่าบางค่าผิดแปลกไป จะทำให้เข้าใจธรรมชาติของตัวเลขมากขึ้น นำไปสู่การตัดสินใจที่จะปรับปรุง หรือคงข้อสอบข้อนั้นไว้ โดยทั่วไปเราจะเริ่มจากการพิจารณาค่าความยากและค่าอำนาจจำแนกของข้อสอบแต่ละข้อ โดยพิจารณาที่ตัวเฉลย ต่อมาเป็นการพิจารณาประสิทธิภาพตัวลวง โดยพิจารณาที่ค่าอำนาจจำแนกของตัวลวงเป็นหลัก และจำนวนคนตอบตัวลวงแต่ละตัวไม่ควรต่ำกว่า 5% จากนั้นก็พิจารณาคุณภาพของแบบทดสอบทั้งฉบับ คือ ค่าความเที่ยง ซึ่งค่าสถิติทุกค่าต้องเป็นไปตามเกณฑ์ที่กำหนด จึงจะถือว่า แบบทดสอบมีคุณภาพ สามารถนำไปใช้ได้

เอกสารอ้างอิง

ศิริชัย กาญจนวาสี. (2556). *ทฤษฎีการทดสอบแบบดั้งเดิม* (พิมพ์ครั้งที่ 7). กรุงเทพฯ: สำนักพิมพ์จุฬาลงกรณ์มหาวิทยาลัย.

Allen, M. J. & Yen, W. M. (1979). *Introduction to Measurement Theory*. Monterey, CA: Brooks/Cole Publishing Company.

Coughlin, P. A. & Featherstone, C. R. (2017). How to Write a High Quality Multiple Choice Question (MCQ): A Guide for Clinicians. *European journal of vascular and endovascular surgery: the official journal of the European Society for Vascular Surgery*, *54*(5), 654–658.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357–381.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Hambleton, R. K. & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38–47.

Kr-20. (2010). In N. J. Salkind (Ed.), *Encyclopedia of research design.* (pp. 668-668). Thousand Oaks, CA: SAGE Publications, Inc.

Nitko, A. J. & Brookhart, S. M. (2011). *Educational assessment of students* (6th ed.). Boston: Pearson Education.

Thompson, N. A. (2016). *Classical item and test analysis using CITAS*. St. Paul, MN: Assessment Systems Corporation.

_____. (2019). *CITAS* [Computer software]. St. Paul, MN: Assessment Systems Corporation.