# การพัฒนาระบบตรวจให้คะแนนอัตโนมัติสำหรับแบบสอบเขียนตอบ วิชาภาษาไทย ระดับมัธยมศึกษาตอนต้น

# The Development of an Automated Scoring System for Assessing Writing Test in The Thai Language Subject at The Lower Secondary Level

**จิราพร ก้อนดินจี่[1*] และ ประกฤติยา ทักษิโณ[2]**

Chiraphon Kondinchi[1*] and Prakittiya Tuksino[2]

[1,2]สาขาวิชาการวัดและประเมินผลการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยขอนแก่น

(Educational Measurement and Evaluation program, Faculty of Education, Khon Kaen University)

---

## บทคัดย่อ

การวิจัยครั้งนี้มีความมุ่งหมาย คือ 1) เพื่อพัฒนาแบบสอบเขียนตอบ วิชาภาษาไทย ระดับมัธยมศึกษาตอนต้น 2) เพื่อพัฒนาระบบตรวจให้คะแนนอัตโนมัติสำหรับแบบสอบเขียนตอบ วิชาภาษาไทย ระดับมัธยมศึกษาตอนต้น และ 3) เพื่อเปรียบเทียบคุณภาพของการตรวจระหว่างผลที่ได้จากคนตรวจและผลที่ได้จากระบบตรวจให้คะแนน กลุ่มตัวอย่างในการวิจัย คือ นักเรียนที่กำลังศึกษาอยู่ในชั้นมัธยมศึกษาปีที่ 3 จำนวน 300 คน ครูผู้สอนในวิชาภาษาไทย จำนวน 5 คน โดยใช้วิธีการสุ่มแบบหลายขั้นตอน และผู้ตรวจที่มีวุฒิการศึกษาด้านการสอนวิชาภาษาไทย จำนวน 3 คน ใช้วิธีการเลือกแบบเจาะจง เครื่องมือที่ใช้ในการวิจัย คือ แบบสอบเขียนตอบ วิชาภาษาไทย ระดับมัธยมศึกษาตอนต้น เรื่อง การจับใจความ สรุปความ และย่อความ วิชาภาษาไทย และระบบตรวจให้คะแนนอัตโนมัติ ผลการวิจัยพบว่า 1) แบบสอบเขียนตอบ วิชาภาษาไทย ระดับมัธยมศึกษาตอนต้น มีจำนวน 6 ข้อ จาก 4 สถานการณ์ และมีคะแนนเต็ม 140 คะแนน ผลการตรวจสอบคุณภาพของแบบสอบพบว่าข้อสอบมีความยากอยู่ในเกณฑ์นำไปใช้ได้ (p = 0.43 – 0.70) อำนาจจำแนกเหมาะสม (r = 0.20 – 0.34) มีความตรงเชิงเนื้อหามีความเหมาะสมทุกข้อ (IOC = 1.00) และเกณฑ์การให้คะแนนของแบบสอบเป็นแบบพิจารณารายละเอียด (Analytic Rubrics) มีความตรงเชิงเนื้อหาทุกข้อ (IOC = 1.00) โดยทั้งข้อสอบและเกณฑ์มีความเที่ยงเหมาะสม ($\alpha$ = 0.731) 2) ระบบตรวจให้คะแนนอัตโนมัติเป็นระบบออนไลน์ชื่อ ASSWT (Automated Scoring System for Writing Test) ที่เป็นระบบการสอบพร้อมตรวจให้คะแนน ประกอบด้วย 3 ส่วน ได้แก่ การใช้งานสำหรับครู การใช้งานสำหรับผู้เข้าสอบ และการปฏิบัติงานของระบบ และ 3) ผลการเปรียบเทียบคุณภาพของการตรวจระหว่างคนและระบบตรวจให้คะแนนอัตโนมัติ พบว่า ระบบตรวจให้คะแนนอัตโนมัติสามารถให้คะแนนได้อย่างสม่ำเสมอกว่าคนตรวจ รวมถึงมีอำนาจจำแนกสูงกว่าในบางกรณี เมื่อวิเคราะห์ความเที่ยงระหว่างผู้ตรวจพบว่าค่าสัมประสิทธิ์อยู่ในระดับปานกลางถึงสูงมาก (r = 0.496 – 0.819) และมีนัยสำคัญทางสถิติทุกข้อ (p < 0.001) ค่าสัมประสิทธิ์สหสัมพันธ์ภายในชั้นอยู่ระดับดีถึงดีเยี่ยม (ICC = 0.815 – 0.945) เมื่อรวมกับระบบตรวจให้คะแนนอัตโนมัติแล้ว ค่าสัมประสิทธิ์สหสัมพันธ์ภายในชั้นยังคงอยู่ในระดับสูง และเมื่อวิเคราะห์ด้วยทฤษฎี G-Theory (P × i × r) พบว่า การใช้ระบบอัตโนมัติร่วมกับผู้ตรวจทำให้ค่าสัมประสิทธิ์การสรุปอ้างอิงเพิ่มขึ้นอย่างชัดเจน โดยค่าสัมประสิทธิ์การสรุปอ้างอิงสำหรับการตัดสินใจสัมพัทธ์ ($\boldsymbol{\rho_{\delta}^{2}}$) เพิ่มจาก 0.26 เป็น 0.62 และสำหรับการตัดสินใจสัมบูรณ์ ($\boldsymbol{\rho_{\Delta}^{2}}$) เพิ่มจาก 0.17 เป็น 0.51 แสดงให้เห็นว่าการนำระบบตรวจให้คะแนนอัตโนมัติมาตรวจจะช่วยลดแหล่งความแปรปรวนที่ไม่พึงประสงค์และเพิ่มความน่าเชื่อถือของการให้คะแนนอย่างมีนัยสำคัญ

**คำสำคัญ:** ระบบตรวจให้คะแนนอัตโนมัติ แบบสอบเขียนตอบ เกณฑ์การให้คะแนน ปัญญาประดิษฐ์ การตรวจให้คะแนนโดยคน



*ได้รับทุนอุดหนุนการวิจัยและนวัตกรรมจากสำนักงานการวิจัยแห่งชาติ (วช.)

*This project is funded by National Research Council of Thailand (NRCT)

## ABSTRACT

This study aimed to develop a Thai constructed-response test and an automated scoring system for assessing Thai writing at the lower secondary level, and to compare scoring outcomes between human raters and the automated system. The research was conducted in three phases. In Phase 1, a six-item, four-scenario constructed-response test totaling 140 points was developed along with analytic rubrics covering comprehension, summarization, and paraphrasing. The test showed appropriate item difficulty ($p = .43 – .70$), acceptable discrimination ($r = .20 –.34$), perfect content validity (IOC = 1.00), and satisfactory internal consistency (Cronbach's $\alpha = .731$). In Phase 2, the Automated Scoring System for Writing Test (ASSWT) was developed as a web-based application with three modules: a teacher interface, an examinee interface, and a backend scoring engine. The system was built using C#, JavaScript, HTML, CSS, and ASP.NET Core MVC (.NET 9) with SQL Server 2022 Express, and it utilized the o3-mini language model from OpenAI. In Phase 3, the system was evaluated against human scoring. The automated system yielded more consistent scores (lower standard deviations) and, in some cases, higher discrimination. Pearson correlations between system and human raters ranged from moderate to very high ($r = .496 – .819$, $p < .001$). Intraclass correlation coefficients were in the good to excellent range (.815 – .945). Generalizability theory analysis ($p \times i \times r$) indicated improved reliability when using the system, with generalizability coefficients increasing from $\boldsymbol{\rho_{\delta}^{2}}$ = .26 to .62 and from $\boldsymbol{\rho_{\Delta}^{2}}$ = .17 to .51. The results suggest that the automated system enhances scoring reliability and reduces unwanted variance.


**KEYWORDS:** Automated scoring system, Writing test, Analytic Rubrics, Artificial Intelligence, Human rater

---

**Corresponding author, E-mail: chiraphon.k@kkumail.com Tel. 081-3691627*

*Received: 28 September 2025 /Revised: 17 November 2025 /Accepted: 11 December 2025 /Published online: 26 December 2025*

## บทนำ

การทดสอบความสามารถทางการใช้ภาษาในระดับนานาชาติที่ได้รับการยอมรับจากทั่วโลก มีการใช้ข้อสอบแบบเขียนตอบเพื่อวัดความสามารถทางภาษาอังกฤษจากการเขียนตอบ เช่น Test of English as a Foreign Language (TOEFL), International English Language Testing System (IELTS), Common European Framework of Reference for Languages (CEFR) และในการทดสอบระดับชาติ เช่น การทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (Ordinary National Educational Test : O-NET) การทดสอบความสามารถพื้นฐานระดับชาติ (National Test : NT) ใช้แบบสอบเขียนตอบในการวัดความสามารถทางภาษาไทยเช่นเดียวกันกับระดับนานาชาติ เพื่อให้ผู้สอบสามารถแสดงทักษะในการใช้ภาษาได้อย่างเต็มศักยภาพ วัดความสามารถของผู้สอบได้เที่ยงตรงมากขึ้น และการจัดการเรียนรู้วิชาภาษาไทยมีการพัฒนาสาระการเขียนที่หลากหลายในทุกระดับชั้นอย่างต่อเนื่อง ซึ่งการเขียนเป็นเครื่องมือสื่อสารของมนุษย์ เพื่อถ่ายทอดความคิด ความเข้าใจ และประสบการณ์ของตนแก่ผู้อ่าน ใช้ในการเรียนรู้บันทึกและฝึกฝนในสิ่งต่าง ๆ เพื่อให้เกิดความเข้าใจในองค์ความรู้มากยิ่งขึ้น และในเรื่องการจับใจความ สรุปความ และย่อความ เป็นทักษะที่กำหนดเป็นตัวชี้วัดในหลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน พุทธศักราช 2551 ฉบับปรับปรุงแก้ไข พ.ศ. 2566 กลุ่มสาระการเรียนรู้ภาษาไทย ระดับมัธยมศึกษาปีที่ 1 – 3 มีตัวชี้วัดทั้งหมด 36 ตัวชี้วัด ซึ่งเป็นตัวชี้วัดที่เกี่ยวกับการจับใจความ สรุปความ และย่อความ

ทั้งหมด 16 ตัวชี้วัด (Pattani Provincial Education Office, 2023) ให้เห็นว่า หลักสูตรให้ความสำคัญกับการเรียนรู้เรื่อง การจับใจความ สรุปความ และย่อความ ในระดับชั้นมัธยมศึกษาตอนต้น

การทดสอบวิชาภาษาไทยเรื่อง การจับใจความ ย่อความ สรุปความ ต้องอาศัยการแสดงความรู้จากการเขียนตอบ (Essay test) ที่มีเกณฑ์ในการตรวจคำตอบที่ครอบคลุมคำตอบและมีหลายระดับคะแนน (Department of Local Administration, 2023) แบบสอบเขียนตอบหรือแบบเสนอคำตอบสามารถวัดความรู้ความสามารถของผู้เรียนได้จนถึงระดับการสร้างสรรค์ ข้อคำถามของแบบสอบที่เป็นการเขียนตอบมีการเขียนตอบแบบสั้นและแบบยาว ซึ่งขึ้นอยู่กับจุดมุ่งหมายของคำถาม และรู้การบูรณาการความคิดของนักเรียน ซึ่งเป็นการสะท้อนคำตอบของผู้เรียนด้วยตนเองได้ดี (Thongsilp et al., 2020) กระทรวงศึกษาธิการมีนโยบายยกระดับคุณภาพการศึกษา ซึ่งมุ่งไปที่การปฏิรูปการเรียนรู้ให้สัมพันธ์เชื่อมโยงระหว่างหลักสูตรการเรียนการสอนและการวัดประเมินผล และทักษะที่จำเป็นสำหรับศตวรรษที่ 21 ซึ่งเพิ่มการใช้ข้อสอบแบบเขียนตอบทั้งการเขียนตอบแบบสั้นและแบบยาวอย่างน้อยร้อยละ 30 ของการสอบแต่ละครั้ง ดังนั้น การวัดประเมินโดยใช้เครื่องมือมาตรฐานแบบเขียนตอบจึงเป็นเครื่องมือที่สำคัญที่จะส่งเสริมและนำไปสู่การพัฒนาทักษะด้านการเขียนเพื่อสะท้อนและถ่ายทอดองค์ความรู้ความเข้าใจในเรื่องราวต่าง ๆ พร้อมทั้งพัฒนาทักษะและความคิดสร้างสรรค์ที่เป็นทั้งศาสตร์และศิลป์ผ่านการเขียนตอบ เพื่อประเมินความรู้ความสามารถของผู้เรียน (Bureau of Educational Testing, 2022)

ปัญหาของการตรวจให้คะแนนแบบสอบเขียนตอบมีหลายประการ ได้แก่ เกณฑ์ที่กำหนดอาจจะไม่สอดคล้องกับสิ่งที่ต้องการวัดอย่างเพียงพอ ความลำบากในการประเมิน ความถูกต้องและความสมบูรณ์ของการตอบต้องอาศัยเวลาในการตรวจค่อนข้างมาก เป็นการเพิ่มภาระในการตรวจให้กับผู้สอน เนื่องจากต้องอ่านเพื่อวิเคราะห์และประเมินคำตอบเป็นค่าคะแนนตามเกณฑ์ที่กำหนด และอาจจะส่งผลให้ขาดความเที่ยงในการตรวจข้อสอบที่เหมาะสม อันเป็นความคลาดเคลื่อนจากผู้ตรวจ เช่น การอ่านลายมือผิดพลาด ให้คะแนนเพียงส่วนใดส่วนหนึ่ง ให้คะแนนตามความประทับใจ เป็นต้น หากใช้แบบสอบเขียนตอบในการทดสอบขนาดใหญ่ ต้องใช้งบประมาณและกำลังคนในการตรวจจำนวนมาก (Department of Local Administration, 2023) รวมถึงความคลาดเคลื่อนที่เกิดจากการใช้ความประทับใจส่วนตัว การไม่ชอบส่วนตัว การปล่อยคะแนน การกดคะแนน ผลตกค้างจากการตรวจ และความลำเอียงส่วนตัว (Thongsilp et al., 2020) และหากต้องการพิจารณาในกระบวนการแสดงวิธีทำของผู้เรียนเพื่อสะท้อนให้เห็นถึงกระบวนการแก้ปัญหาจำเป็นต้องอาศัยระยะเวลาในการให้ครูหรือผู้ประเมินตรวจให้คะแนน แล้วจึงนำผลการตรวจให้คะแนนที่ได้ไปสู่การรายงานผลการวินิจฉัยเป็นรายบุคคลและภาพรวมในลำดับต่อไป (Sukwichai et al., 2023) ดังนั้น การตรวจให้คะแนนแบบสอบเขียนตอบควรจะมีระบบตรวจให้คะแนนอัตโนมัติ เพื่อลดและแก้ปัญหาที่เกิดขึ้นจากการตรวจแบบสอบที่เป็นการเขียนตอบดังกล่าว

ความน่าเชื่อถือของผู้ตรวจพิจารณาจากการตรวจสอบคุณภาพของการตรวจ ซึ่งเป็นส่วนที่สำคัญเพื่อแสดงถึงความน่าเชื่อถือของผลการตรวจว่ามีคุณภาพและสามารถนำไปใช้วิเคราะห์ต่อไป โดยวิเคราะห์ความเที่ยงระหว่างผู้ประเมิน (Inter-rater reliability) และการวิเคราะห์สหสัมพันธ์ภายในชั้น (Intra-class correlation : ICC) (Janprasert et al., 2020 ; Thitikanpodchana & Tuksino, 2021 ; Broadfoot & Rockey, 2025) และใช้ค่าสัมประสิทธิ์การสรุปอ้างอิงจากทฤษฎี Generalizability Theory เพื่อมาเปรียบเทียบผลการตรวจระหว่างผู้ตรวจ (Thitikanpodchana & Tuksino, 2021)

ระบบการสอบแบบเขียนตอบที่มีการตรวจให้คะแนนอัตโนมัติสร้างขึ้นและพัฒนาครั้งแรกด้วยโครงการ Project Essay Grader™ (PEG) ของ Ellis Page เมื่อปี ค.ศ. 1966 ซึ่งมีการพัฒนาอย่างต่อเนื่องจนมีการเปลี่ยนแปลงในปี ค.ศ. 1990 ที่มีการพัฒนาข้อบกพร่องต่าง ๆ ทำให้มีความหลากหลายมากขึ้น เช่น Intelligent Essay Assessor™ (IEA) ซึ่งวิเคราะห์และให้คะแนนเรียงความโดยใช้เทคนิคการวิเคราะห์ข้อความ (Dikli, 2006) ระบบ e-rater®, ระบบBayesian essay test assessment system (BETSY) และระบบ CRASE การพัฒนาระบบการตรวจให้คะแนนแบบสอบเขียนตอบอัตโนมัติจะอาศัยเทคโนโลยีทางคอมพิวเตอร์ด้านการวิเคราะห์ข้อความ (Text Analysis) ที่เป็นลายลักษณ์อักษรด้วยวิธีการประมวลผลภาษาธรรมชาติ (Natural Language Processing : NLP) ซึ่งแนวทางที่นิยมนำมาใช้ในปัจจุบันมี 2 แนวทาง ได้แก่ (1) แนวทางอิงความรู้ (Knowledge based หรือ AI approach) และ (2) แนวทางอิงสถิติ (Statistical based หรือ Corpus

based approach) (Thongsilp et al., 2020) ระบบตรวจให้คะแนนอัตโนมัติในแบบสอบเขียนตอบ ถูกนำไปใช้ในการทดสอบระดับชาติ และข้อสอบมาตรฐานในระดับนานาชาติ เช่น TOEFL, GMAT, IELTS, SAT เป็นต้น ทั้งนี้ เพื่อเป็นการประหยัดทรัพยากรได้อย่างเหมาะสม (Dikli, 2006) ในปัจจุบัน เทคโนโลยีปัญญาประดิษฐ์ หรือ Artificial Intelligence (AI) เข้ามามีบทบาทในประเทศไทยมากขึ้น และมีโครงสร้างพื้นฐานของปัญญาประดิษฐ์มารองรับ เช่น AI for Thai : Thai AI Service Platform ที่มุ่งเน้นการตอบโจทย์บริบทประเทศไทยทั้งในภาคอุตสาหกรรมและการบริการต่าง ๆ และมี ThaiSC : NSTDA Supercomputer Center ที่ให้บริการทรัพยากรด้านการคำนวณด้วยระบบคอมพิวเตอร์สมรรถนะสูงที่ให้บริการหน่วยวิจัยทั้งองค์กรภาครัฐและเอกชน (Suchato et al., 2023) จะเห็นได้ว่าจากอดีตมาจนปัจจุบัน พัฒนาการของระบบตรวจให้คะแนนอัตโนมัติเป็นไปอย่างต่อเนื่องและหลากหลายมากขึ้น

งานวิจัยที่ใช้การตรวจให้คะแนนอัตโนมัติสำหรับการเขียนภาษาไทยที่เป็นแบบสอบเขียนตอบชนิดจำกัดคำตอบที่ใช้ทดสอบในชั้นเรียน เช่น แบบเติมคำ แบบเขียนคำตอบสั้นเป็นข้อ ๆ และแบบเขียนคำตอบสั้นแบบบรรยายไม่เกิน 3 หรือ 4 บรรทัด เป็นต้น กลุ่มตัวอย่างเป็นนักเรียนชั้นประถมศึกษาปีที่ 6 ชั้นมัธยมศึกษาปีที่ 4 และนักศึกษาในระดับปริญญาตรี ใช้เทคนิคการวิเคราะห์คำสำคัญด้วยวิธีการวิเคราะห์ทางสถิติ และการประมวลผลภาษาธรรมชาติ มาพัฒนาระบบการตรวจให้คะแนนด้วยโปรแกรมคอมพิวเตอร์ เช่น Power Builder, Weka, Eclipse, MATLAB, Java Platform, Java Matrix , PHP เป็นต้น (Thongsilp et al., 2020) ในประเทศไทยนั้น นักเรียนในระดับชั้นมัธยมศึกษาปีที่ 1 – 3 ทุกคนจะต้องเข้าร่วมการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (Ordinary National Educational Test : O-NET) เพื่อทดสอบความรู้ในทุกปี ซึ่งเป็นการสอบที่มีขนาดใหญ่ระดับประเทศ ประกอบด้วย 4 วิชาหลัก ได้แก่ ภาษาไทย ภาษาอังกฤษ คณิตศาสตร์ และวิทยาศาสตร์ แต่การตรวจโดยทั่วไปยังเป็นการตรวจด้วยมนุษย์ ซึ่งอาจทำให้เกิดค่าใช้จ่ายและกำลังคนในการตรวจปริมาณมากเช่นเดียวกับการสอบขนาดใหญ่อื่น ๆ และในต่างประเทศพบว่าระบบการตรวจให้คะแนนอัตนัยอัตโนมัติ มีผลดี คือ ประหยัดค่าใช้จ่ายและลดความคลาดเคลื่อนจากผู้ตรวจ การให้คะแนนเฉพาะตรงกลาง และใช้ความประทับใจ เป็นต้น (Thongsilp et al., 2020) โดยการตรวจให้คะแนนอัตโนมัติมีข้อจำกัด คือ สามารถตรวจอัตโนมัติได้เฉพาะในกรณีที่มุ่งเน้นไปที่คำตอบที่ถูกของผู้สอบ และนำผลการตอบที่ได้ไปเปรียบเทียบกับเกณฑ์ให้คะแนนที่สอดคล้องกับแผนที่เชิงโครงสร้างในแต่ละระดับแล้วสามารถให้คำแนะนำหรือแหล่งการเรียนรู้เพิ่มเติมเพื่อให้ผู้เรียนปรับปรุงได้ในทันที (Sukwichai et al., 2023) ซึ่งการสร้างและพัฒนาระบบตรวจให้คะแนนอัตโนมัติควรมีการสำรวจและสอบถามความต้องการของผู้ใช้ระบบ (User) เพื่อให้สอดคล้องเหมาะสมกับความต้องการใช้งานของผู้ใช้งานตามบริบทการใช้งานจริง การสัมภาษณ์เก็บข้อมูลจากครูผู้สอน ศึกษานิเทศก์ ผู้ทรงคุณวุฒิ ผู้ปกครอง และนักเรียน เกี่ยวกับปัญหา ความต้องการ ความเชื่อมั่น และแนวทางในการนำเทคโนโลยีมาใช้ในกิจกรรมการเรียนการสอน โดยสอบถามจากประสบการณ์ของกลุ่มเป้าหมายเพื่อเก็บข้อมูลแล้วนำไปใช้พัฒนาระบบให้เหมาะกับการเรียนการสอนในสถานการณ์จริง ตรงกับความต้องการใช้งานมากที่สุด (Suchato et al., 2023)

จากปัญหาและความสำคัญที่กล่าวมาในข้างต้น เพื่อแก้ปัญหาความคลาดเคลื่อนของการตรวจในคน ประหยัดเวลา งบประมาณ และมีคุณภาพการตรวจที่ใกล้เคียงหรือเท่ากันของผู้ตรวจ ผู้วิจัยจึงพัฒนาระบบตรวจให้คะแนนอัตโนมัติสำหรับแบบสอบเขียนตอบ วิชาภาษาไทย ระดับมัธยมศึกษาตอนต้น

**วัตถุประสงค์**

1. เพื่อพัฒนาแบบสอบเขียนตอบ เรื่อง การจับใจความ สรุปความ และย่อความ วิชาภาษาไทย ระดับมัธยมศึกษาตอนต้น
2. เพื่อพัฒนาระบบตรวจให้คะแนนอัตโนมัติสำหรับแบบสอบเขียนตอบ วิชาภาษาไทย ระดับมัธยมศึกษาตอนต้น
3. เพื่อเปรียบเทียบคุณภาพของการตรวจระหว่างผลที่ได้จากคนตรวจและผลที่ได้จากระบบตรวจให้คะแนนอัตโนมัติ

**นิยามศัพท์เฉพาะ**

1. การจับใจความ หมายถึง การเก็บประโยคหรือเนื้อหาของเรื่องที่เป็นสาระสำคัญ ความรู้ข้อมูล ที่น่าสนใจ และแนวความคิดหรือทัศนะของผู้เขียนของเรื่องที่อ่านในแต่ละย่อหน้า

2. การสรุปความ หมายถึง นำใจความสำคัญของเรื่องมาเรียบเรียงใหม่แบบสั้น ๆ กระชับ และเข้าใจง่าย เป็นสำนวนภาษาของตนเอง โดยครอบคลุมเนื้อหาทั้งหมด

3. การย่อความ หมายถึง การเขียนสาระสำคัญของเรื่อง โดยตัดข้อความที่เป็นพลความออก แล้วนำมาเขียนเป็นสำนวนภาษาของตนเอง และมีส่วนต้นที่บอกข้อมูลและที่มาของสารตามรูปแบบชนิดของสารที่ย่อความ

4. แบบสอบเขียนตอบวิชาภาษาไทย คือ แบบสอบที่ประกอบด้วยข้อสอบที่กำหนดสถานการณ์ให้ผู้สอบได้แสดงคำตอบออกมาด้วยตนเองด้วยการเขียนหรือพิมพ์ โดยไม่มีตัวเลือกของคำตอบ ในที่นี้จะเป็นการเขียนตอบในเรื่องการย่อความ จับใจความ และสรุปความ ที่มีความยาวของคำตอบไม่เกิน 1,000 อักขระ

5. ระบบตรวจให้คะแนนอัตโนมัติ หมายถึง การนำระบบคอมพิวเตอร์มาใช้ในการทดสอบและตรวจให้คะแนนจากผลการตอบข้อสอบเขียนตอบและแสดงผลให้ทราบโดยทันที เพื่อแสดงผลความสามารถของผู้สอบจากการสอบ โดยใช้เว็บไซต์ที่เป็นระบบออนไลน์ชื่อ Automated Scoring System for Writing Test (ASSWT) ที่มีการตรวจให้คะแนนคำตอบของข้อสอบเขียนตอบตามเกณฑ์การตรวจให้คะแนน เรื่อง การจับใจความ การย่อความ และการสรุปความ ด้วยเทคโนโลยีปัญญาประดิษฐ์ หรือ Artificial Intelligence (AI) โมเดล GPT-o3-mini จาก OpenAI

6. คุณภาพการตรวจให้คะแนนอัตโนมัติ หมายถึง การตรวจให้คะแนนที่มีความเที่ยงของผู้ตรวจเป็นไปตามที่ต้องการ โดยการวัดจากค่าสถิติ ได้แก่ ความเที่ยงระหว่างผู้ตรวจ (Inter-rater reliability) หมายถึง ความคงเส้นคงวาของผลการพิจารณาจากผู้ตรวจ 2 คนขึ้นไป โดยที่ผู้ตรวจมีความเป็นอิสระต่อกัน, สัมประสิทธิ์สหสัมพันธ์ภายในชั้น (Intraclass Correlation Coefficient: ICC) หมายถึง ความสอดคล้องของผลการตรวจอันแสดงถึงความน่าเชื่อถือของผู้ตรวจ และสัมประสิทธิ์การสรุปอ้างอิงความน่าเชื่อถือของผลการวัด (Generalizability Coefficient) หมายถึง ความน่าเชื่อถือของคุณภาพในผลการตรวจจากระบบตรวจให้คะแนนอัตโนมัติ เมื่อเทียบกับคุณภาพในผลการตรวจจากคน

**กรอบแนวคิดการวิจัย**

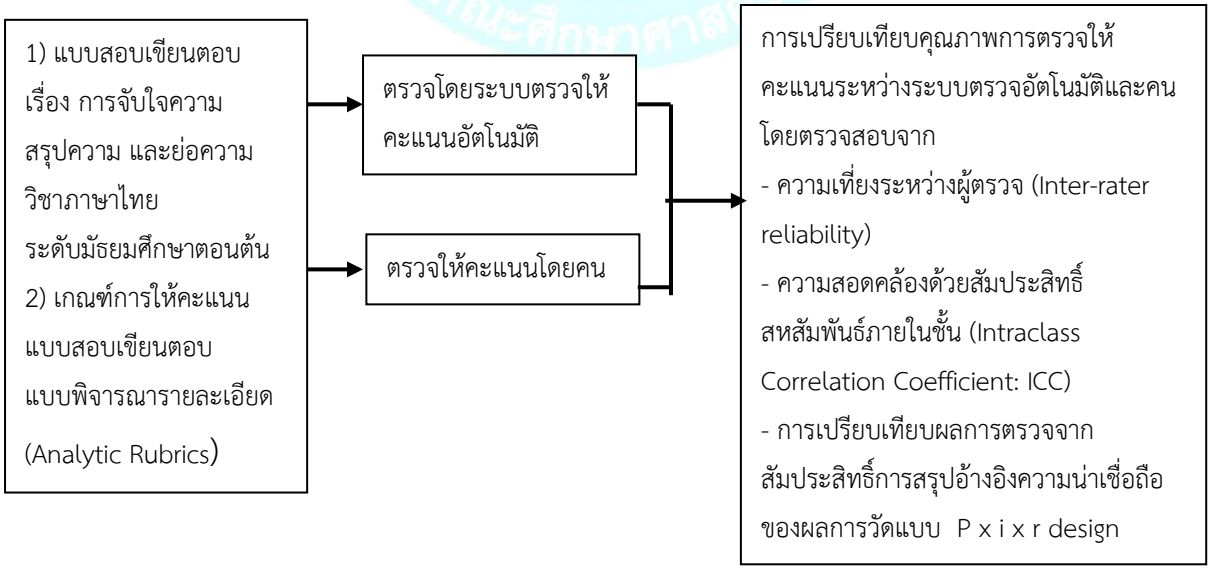


**Figure 1** กรอบแนวคิดการวิจัย

## วิธีดำเนินการวิจัย

การวิจัยนี้เป็นการวิจัยออกแบบ (Design Research) โดยผู้วิจัยได้แบ่งวิธีการดำเนินการวิจัยออกเป็น 3 ระยะ ได้แก่ ระยะที่ 1 การพัฒนาแบบสอบเขียนตอบ และเกณฑ์การให้คะแนนทั้งแบบปกติและแบบอัตโนมัติของแบบสอบเขียนตอบวิชาภาษาไทย ระดับมัธยมศึกษาตอนต้น ระยะที่ 2 การพัฒนาระบบตรวจให้คะแนนอัตโนมัติของแบบสอบเขียนตอบวิชาภาษาไทย ระดับมัธยมศึกษาตอนต้น และระยะที่ 3 การเปรียบเทียบคุณภาพของการตรวจระหว่างผลที่ได้จากผู้ตรวจและผลที่ได้จากระบบตรวจให้คะแนนอัตโนมัติ

### ประชากร และตัวอย่างวิจัย

ประชากร คือ นักเรียนชั้นมัธยมศึกษาตอนต้น (ม.1-3) ของโรงเรียนในทุกสังกัด จำนวน 1,661,035 คน (Office of the Basic Education Commission, 2023)

กลุ่มตัวอย่างในการวิจัย คือ นักเรียนที่กำลังศึกษาอยู่ในชั้นมัธยมศึกษาปีที่ 3 จำนวน 300 คน โดยใช้วิธีการสุ่มแบบหลายขั้นตอน โดยสุ่มนักเรียนชั้นมัธยมศึกษาปีที่ 3 จากสังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐานมา 1 ภาค (ภาคเหนือ ภาคตะวันออกเฉียงเหนือ ภาคกลาง และภาคใต้) แล้วสุ่มจังหวัดจากในภาคมา 5 จังหวัด เพื่อที่จะสุ่มจังหวัดละ 1 เขตพื้นที่การศึกษา แล้วสุ่ม 1 โรงเรียนในเขตพื้นที่การศึกษามัธยมศึกษานั้น และสุ่มนักเรียนชั้นมัธยมศึกษาปีที่ 3 โรงเรียนละ 60 คน รวมทั้งหมด 300 คน โดยคำนวณขนาดตัวอย่างตามกฎเกณฑ์ของ Smith (1978 as cited in Apaikawee et al., 2020) ที่เสนอกลุ่มตัวอย่างขั้นต่ำ $n_p$ = 25 คน (low), $n_p$ = 50 คน (middle) และ $n_p$ =100 คน (high) คือ $n_p$(กลุ่มตัวอย่าง) $n_i$(สถานการณ์) $n_r$(ผู้ตรวจ) อย่างน้อย 1,080 ค่า ซึ่งในการออกแบบการวัดใช้โดยโปรแกรม EduG เพื่อตรวจสอบแหล่งความคลาดเคลื่อน (source of variance) และประมาณค่าสัมประสิทธิ์การสรุปอ้างอิง (G-Coefficient) จะได้ ($n_p$ = 300) ($n_i$ = 6) ($n_r$ = 3) = 5,400 ค่า ซึ่งผ่านเกณฑ์ค่าขั้นต่ำของการประมาณค่าตามข้อเสนอดังกล่าว ครูผู้สอนวิชาภาษาไทย ระดับมัธยมศึกษาตอนต้น จำนวน 5 คน ใช้วิธีการเลือกแบบเจาะจง ซึ่งเป็นครูที่จะเข้าร่วมการใช้ระบบตรวจให้คะแนนอัตโนมัติจากโรงเรียนที่สุ่มกลุ่มตัวอย่างนักเรียน เนื่องจากกลุ่มตัวอย่างที่จะต้องเป็นผู้ที่สอน ผู้ใช้ระบบ และคุมสอบ ซึ่งมีทั้งหมด 5 โรงเรียน โรงเรียนละ 1 คน และผู้ตรวจที่มีวุฒิการศึกษาด้านการสอนวิชาภาษาไทย จำนวน 3 คน ใช้วิธีการเลือกแบบเจาะจง ซึ่งจะต้องมีวุฒิการศึกษาระดับปริญญาตรีด้านการสอนวิชาภาษาไทย และมีประสบการณ์สอนวิชาภาษาไทย ระดับมัธยมศึกษาตอนต้น ไม่ต่ำกว่า 5 ปี

### เครื่องมือวิจัย

1. แบบสอบเขียนตอบ เรื่อง การจับใจความ สรุปความ และย่อความ วิชาภาษาไทย ระดับมัธยมศึกษาตอนต้น โดยผู้วิจัยสร้างขึ้นตามตัวชี้วัดในหลักสูตรการเรียนรู้หลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน พ.ศ. 2551 ฉบับปรับปรุงแก้ไข พ.ศ. 2566 พร้อมทั้งศึกษาลักษณะข้อสอบอัตนัย O-NET ม.3 วิชาภาษาไทย ปีการศึกษา 2564 – 2566 และข้อสอบอัตนัย O-NET ป.6 วิชาภาษาไทย ปีการศึกษา 2559 – 2566 เพื่อนำมาปรับใช้ในการสร้างข้อสอบ โดยแบบสอบนี้มีข้อสอบจำนวน 6 ข้อ จาก 4 สถานการณ์ รวม 140 คะแนน แบ่งเป็นสถานการณ์ที่ 1 และ 2 ใช้วัดความสามารถ เรื่อง การจับใจความและการสรุปความ สถานการณ์ที่ 3 และ 4 ใช้วัดความสามารถ เรื่อง การย่อความ ในแต่ละสถานการณ์จะเป็นบทอ่านที่มีผู้เขียนไว้แล้ว ซึ่งผู้วิจัยได้คัดเลือกมาตามประเภทของบทอ่านในแผนผังการสร้างข้อสอบ และกำหนดเวลาในการสอบ 90 นาที โดยมีการตรวจและประเมินความเหมาะสมและความสอดคล้องจากผู้เชี่ยวชาญ และนำข้อสอบไปทดลองใช้กับนักเรียนชั้นมัธยมศึกษาปีที่ 1 – 3 จำนวน 30 คน

2. เกณฑ์การตรวจให้คะแนนแบบสอบเขียนตอบ เรื่อง การจับใจความ สรุปความ และย่อความ วิชาภาษาไทย ระดับมัธยมศึกษาตอนต้น ผู้วิจัยสร้างขึ้นโดยการพิจารณาจากองค์ประกอบของการจับใจความ สรุปความ และย่อความจากวรรณกรรมและงานวิจัยที่เกี่ยวข้องในบทที่ 2 ร่วมกับหลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน พ.ศ. 2551 ฉบับปรับปรุงแก้ไข พ.ศ. 2566 ความ และพัฒนาจากเกณฑ์การตรวจให้คะแนนข้อสอบอัตนัย O-NET ม.3 วิชาภาษาไทย ปีการศึกษา 2566

โดยสร้างเกณฑ์การตรวจให้คะแนนเป็นรูปแบบเกณฑ์การประเมินแบบพิจารณารายละเอียด (Analytic Scoring rubrics) จำแนกเป็น 3 เรื่อง ได้แก่ เกณฑ์การตรวจให้คะแนนการจับใจความ เกณฑ์การตรวจให้คะแนนการสรุปความ และเกณฑ์การตรวจให้คะแนนการย่อความ โดยมีการตรวจและประเมินความเหมาะสมและความสอดคล้องจากผู้เชี่ยวชาญ พร้อมทั้งนำไปทดลองใช้ตรวจให้คะแนนคำตอบร่วมกับแบบสอบในข้อ 1.

3. ระบบตรวจให้คะแนนอัตโนมัติของแบบสอบเขียนตอบ วิชาภาษาไทย ระดับมัธยมศึกษาตอนต้น (Automated Scoring System for Writing Test: ASSWT) ที่นำข้อสอบในข้อ 1. และเกณฑ์การตรวจให้คะแนนในข้อ 2. เข้ามาใช้งานในระบบ ซึ่งกำหนดเวลาสอบ 90 นาที จากนั้นจะนำแบบสอบเขียนตอบและเกณฑ์การตรวจให้คะแนนมาทดลองตรวจด้วยโมเดล AI ที่สามารถตรวจคำตอบตามจุดมุ่งหมายของการวัดของแบบสอบ ได้แก่ GPT-4o-mini, GPT-4o และ GPT-o3-mini มาทดลองตรวจคำตอบของกลุ่มตัวอย่างในขั้นการทดลองใช้เครื่องมือ แล้วนำผลการตรวจให้คะแนนของทุกโมเดลมาพิจารณาคัดเลือกในการสนทนากลุ่มรอบที่ 1 ร่วมกับผู้เชี่ยวชาญ เพื่อพิจารณาเลือกโมเดลการตรวจที่เหมาะสม สอดคล้องและตรวจคำตอบได้ดีที่สุด รวมถึงพิจารณาการทำงานของระบบในทุกด้าน ให้ข้อเสนอแนะในการปรับปรุงแก้ไข เพื่อให้ระบบสมบูรณ์มากที่สุดก่อนที่จะนำไปเก็บข้อมูลจริง

**การวิเคราะห์ข้อมูล**

การวิเคราะห์ข้อมูลได้จำแนกตามรูปแบบวิธีการดำเนินการวิจัยเป็น 3 ระยะ ดังนี้

1. ระยะที่ 1 ผู้วิจัยวิเคราะห์ข้อมูลของแบบสอบจากค่าสถิติ ได้แก่ ค่าเฉลี่ย คะแนนต่ำสุด คะแนนสูงสุด ส่วนเบี่ยงเบนมาตรฐาน (SD) และความเที่ยงของผู้ตรวจ (Inter-rater reliability) จากสถิติสหสัมพันธ์ภายในชั้น (ICC) ซึ่งตรวจสอบคุณภาพข้อสอบรายข้อจากค่าความยาก อำนาจจำแนก ความตรงเชิงเนื้อหา (Index of item objective congruence: IOC) พร้อมทั้งตรวจสอบคุณภาพของแบบสอบทั้งฉบับจากความตรงเชิงเนื้อหาและหาความเที่ยงของแบบสอบจากสัมประสิทธิ์แอลฟาของครอนบาค (Cronbach's Alpha Coefficient) ตามทฤษฎีการทดสอบแบบดั้งเดิม (Classical Test Theory: CTT) และตรวจสอบความสอดคล้องของเกณฑ์การตรวจให้คะแนนจากความตรงเชิงเนื้อหาของเกณฑ์ ซึ่งในการตรวจสอบความตรงเชิงเนื้อหาของข้อสอบรายข้อ แบบสอบทั้งฉบับ และความสอดคล้องของเกณฑ์ โดยจะนำมาจากผลการประเมินของผู้ทรงคุณวุฒิด้านการวัดประเมินผลและด้านภาษาไทย จำนวน 3 ท่าน

2. ระยะที่ 2 ผู้วิจัยได้สอบถามความต้องการใช้และปัญหาของการทดสอบออนไลน์ในระบบคอมพิวเตอร์ จากกลุ่มตัวอย่างนักเรียนที่กำลังศึกษาอยู่ในชั้นมัธยมศึกษาปีที่ 3 ที่จำนวน 5 คน และครูผู้สอนในวิชาภาษาไทย ระดับชั้นมัธยมศึกษาตอนต้น จำนวน 5 คน แล้วสรุปข้อมูลเพื่อนำไปออกแบบสร้างระบบตรวจให้คะแนนอัตโนมัติ เมื่อระบบสร้างเสร็จสิ้นก็จะนำไปพิจารณาในการสนทนากลุ่ม (Focus group) รอบที่ 1 โดยมีผู้เข้าร่วมคือ ครูที่ทดลองใช้ระบบ 5 คน ผู้เชี่ยวชาญในด้านระบบคอมพิวเตอร์ ด้านภาษาไทย และด้านการวัดและประเมินผล ด้านละ 1 ท่าน ร่วมกับผู้วิจัย เพื่อพิจารณาความเหมาะสมและให้ข้อเสนอแนะในการทำงานของระบบในทุกด้าน

3. ระยะที่ 3 ผู้วิจัยวิเคราะห์ค่าสถิติเชิงบรรยาย (descriptive statistics) จากการตรวจทั้งสองรูปแบบ ได้แก่ ค่าเฉลี่ย ค่าสูงสุด - ต่ำสุด ส่วนเบี่ยงเบนมาตรฐาน และค่าอำนาจจำแนกจากผลการตรวจของผู้ตรวจและผลการตรวจของระบบตรวจให้คะแนนอัตโนมัติ เพื่อเปรียบเทียบลักษณะการตรวจโดยระบบตรวจให้คะแนนอัตโนมัติและการตรวจโดยคน พร้อมทั้งตรวจสอบความเที่ยงระหว่างผู้ตรวจ (Inter-rater reliability) จากค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน (Pearson Correlation Coefficient) และสถิติสหสัมพันธ์ภายในชั้น (ICC) เพื่อตรวจสอบความสอดคล้องในการให้คะแนนว่ามีความน่าเชื่อถืออยู่ในระดับใด และเปรียบเทียบคุณภาพของผลการตรวจระหว่างผลที่ได้จากผู้ตรวจและผลที่ได้จากระบบตรวจให้คะแนนอัตโนมัติ โดยวิเคราะห์ตามทฤษฎี Generalizability Theory (G-Theory) จากสัมประสิทธิ์การสรุปอ้างอิงความน่าเชื่อถือของผลการวัด (Generalizability Coefficient) แบบ P x i x r design และวิเคราะห์ระบบ ASSWT จากการสนทนากลุ่มรอบที่ 2 โดยมีผู้เข้าร่วม คือ ผู้เชี่ยวชาญทางด้านภาษาไทย ด้านเทคโนโลยี และด้านการวัดประเมินผล ด้านละ

1 ท่าน ครูผู้ใช้ระบบ 5 ท่าน และผู้วิจัย ร่วมกันวิเคราะห์ให้ข้อเสนอแนะเกี่ยวกับผลการใช้ระบบการตรวจข้อสอบ และการรายงานผลของระบบ

## ผลการวิจัย

1. แบบสอบเขียนตอบ วิชาภาษาไทย ระดับมัธยมศึกษาตอนต้น มีจำนวน 6 ข้อ จาก 4 สถานการณ์ รวม 140 คะแนน แบ่งเป็นสถานการณ์ที่ 1 และ 2 ใช้วัดความสามารถ เรื่อง การจับใจความและการสรุปความ สถานการณ์ที่ 3 และ 4 ใช้วัดความสามารถ เรื่อง การย่อความ โดยข้อสอบมีค่าความเที่ยงของผู้ตรวจเป็น 0.312 อยู่ในระดับพอใช้ได้ มีค่าความยากอยู่ระหว่าง 0.43 – 0.70 มีค่าอำนาจจำแนกอยู่ระหว่าง 0.20 – 0.34 และความตรงเชิงเนื้อหา(IOC) เป็น 1.00 และมีค่าสัมประสิทธิ์แอลฟาของครอนบาค 0.731 ซึ่งเกณฑ์การให้คะแนนของแบบสอบเขียนตอบเป็นการให้คะแนนแบบพิจารณารายละเอียด (Analytic Rubrics) โดยจำแนกเป็น 3 ประเภท ได้แก่ (1) เกณฑ์การตรวจให้คะแนนการจับใจความ ประกอบด้วยเกณฑ์การให้คะแนนใน 3 ด้าน ประกอบด้วย ด้านเนื้อหา 2 คะแนน, ด้านการสะกดคำ 1 คะแนน และด้านความสมบูรณ์ของประโยคและภาษา 2 คะแนน ซึ่งผู้สอบจะตอบ 5 ประโยค ประโยคละ 5 คะแนน รวมทั้งหมด 25 คะแนน (2) เกณฑ์การตรวจให้คะแนนการสรุปความ ประกอบด้วยเกณฑ์การให้คะแนน 6 ด้าน ประกอบด้วย ด้านใจความสำคัญ 4 คะแนน, ด้านความยาวของคำตอบ 2 คะแนน, ด้านการเรียงลำดับและการเชื่อมโยงเนื้อหา 2 คะแนน, ด้านการคัดลอกข้อความ 3 คะแนน, ด้านหลักการสรุปความ 4 คะแนน และด้านการใช้ภาษา 8 คะแนน รวมทั้งหมด 23 คะแนน และ (3) เกณฑ์การตรวจให้คะแนนการย่อความ ประกอบด้วยเกณฑ์การให้คะแนน 4 ด้าน ประกอบด้วยด้านความยาวของคำตอบ 2 คะแนน, ด้านเนื้อหา 11 คะแนน, ด้านหลักการย่อความ 3 คะแนน และด้านการใช้ภาษา 6 คะแนน รวมทั้งหมด 22 คะแนน และผลการประเมินความสอดคล้องของเกณฑ์จากความตรงเชิงเนื้อหา(IOC) พบว่า เกณฑ์การจับใจความทุกข้อเป็น 1.00 เกณฑ์การสรุปความ ข้อ 1 และ 8 เป็น 0.67 ข้อ 2 – 7 เป็น 1.00 เกณฑ์การย่อความ ข้อ 1 และ 8 เป็น 0.67 และข้อ 2 – 7 มีค่าเป็น 1.00 และความเที่ยงของเกณฑ์การตรวจทั้งฉบับมีค่าสัมประสิทธิ์แอลฟาของครอนบาคเป็น 0.731

2. ระบบตรวจให้คะแนนอัตโนมัติ (ASSWT) ประกอบด้วย 3 ส่วน ได้แก่ (1) ระบบการใช้งานสำหรับครู ที่สามารถสร้างแบบทดสอบที่สามารถใส่บทอ่านได้ทั้งแบบรูปภาพและข้อความ การจัดการเกณฑ์การตรวจข้อสอบ การตรวจคำตอบที่สามารถเลือกตรวจและเลือกดูผลการตรวจได้ 2 รูปแบบ คือ การตรวจด้วยตนเองและการตรวจคำตอบโดยปัญญาประดิษฐ์ (AI) และการนำออกของข้อมูล (Export) ที่ครูสามารถจัดการและนำออกข้อมูลการตอบและผลการตรวจออกจากระบบได้หลากหลายรูปแบบทั้งการรายงานผลคะแนนรายกลุ่ม รายบุคคล โรงเรียน การนำออกคำตอบและข้อเสนอแนะของผู้ตรวจ (2) ระบบการใช้งานสำหรับผู้เข้าสอบที่ออกแบบมาเพื่อให้ผู้เข้าสอบสามารถทำแบบทดสอบออนไลน์ได้อย่างสะดวก โดยมีขั้นตอนและหน้าจอการใช้งานที่ชัดเจน และการรายงานผลพร้อมข้อเสนอแนะทันที และ (3) ระบบการตรวจให้คะแนนอัตโนมัติพัฒนาขึ้นด้วยระบบปฏิบัติการ Windows และตรวจให้คะแนนใช้โมเดลปัญญาประดิษฐ์ o3-mini จาก OpenAI

3. คุณภาพของการตรวจระหว่างผลที่ได้จากผู้ตรวจและผลที่ได้จากระบบตรวจให้คะแนนอัตโนมัติ

3.1 ผลการวิเคราะห์ค่าสถิติเชิงบรรยาย (descriptive statistics) จากการตรวจทั้งสองรูปแบบ พบว่าค่าเฉลี่ยของผลคะแนนของทุกข้อที่ตรวจด้วยผู้ตรวจต่ำกว่าการตรวจโดยระบบตรวจให้คะแนนอัตโนมัติ ส่วนเบี่ยงเบนมาตรฐาน (SD) ของการตรวจโดยผู้ตรวจอยู่ระหว่าง 6.49 – 7.52 และระบบตรวจให้คะแนนอัตโนมัติอยู่ระหว่าง 4.16 – 6.97 แสดงให้เห็นว่า การให้คะแนนของการตรวจให้คะแนนโดยผู้ตรวจมีการกระจายน้อยกว่า และอำนาจจำแนกของการตรวจด้วยผู้ตรวจอยู่ระหว่าง 0.28 – 0.44 ซึ่งอยู่ในระดับพอใช้ได้ถึงดี ส่วนค่าอำนาจจำแนกของการตรวจให้คะแนนจากระบบตรวจให้คะแนนอัตโนมัติอยู่ระหว่าง 0.18 – 0.32 ซึ่งส่วนใหญ่อยู่ในระดับพอใช้ได้ การตรวจทั้งสองรูปแบบมีแนวโน้มความสามารถในการจำแนกใกล้เคียงกัน แม้ในบางข้อผู้ตรวจจะสามารถให้คะแนนได้อย่างยืดหยุ่นและละเอียดกว่า แต่ระบบตรวจให้คะแนนอัตโนมัติก็ยังมีความสามารถของการตรวจส่วนใหญ่อยู่ในเกณฑ์ที่ยอมรับได้ ดัง **Figure 2 และ Figure 3**

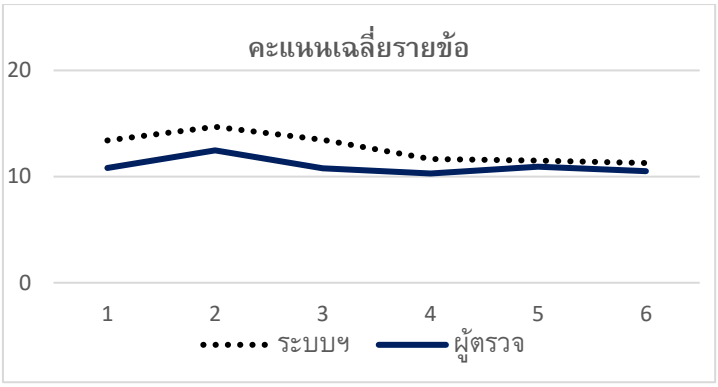


**Figure 2** คะแนนเฉลี่ยรายข้อของผลการตรวจทั้งระบบตรวจให้คะแนนอัตโนมัติและผู้ตรวจ

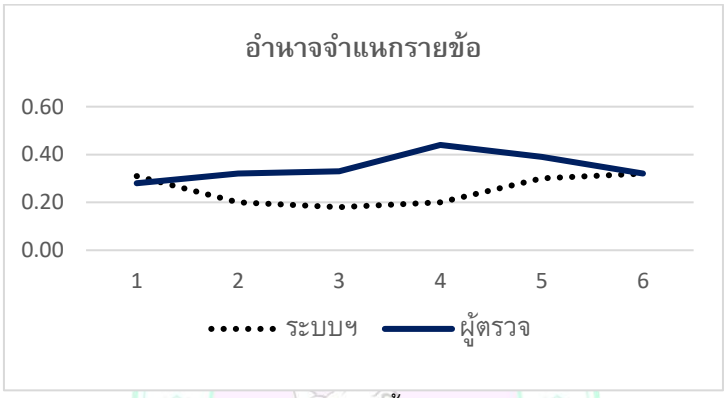


**Figure 3** อำนาจจำแนกรายข้อของผลการตรวจทั้งระบบตรวจให้คะแนนอัตโนมัติและผู้ตรวจ

3.2 การตรวจสอบความเที่ยงระหว่างผู้ตรวจ (Inter-rater reliability) จากค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน (Pearson Correlation Coefficient) ระหว่างคะแนนที่ได้จากการตรวจของผู้ตรวจที่เป็นคนและการตรวจด้วยระบบตรวจให้คะแนนอัตโนมัติ ซึ่งผู้วิจัยได้แปลผลขนาดของความสัมพันธ์ตามเกณฑ์ของโคเฮน (Cohen, 1988) คือ ค่าสัมประสิทธิ์ 0.10 – 0.30 อยู่ในระดับต่ำ 0.31 – 0.50 อยู่ในระดับปานกลาง และ 0.51 – 1.00 อยู่ในระดับสูง โดยข้อสอบมีความเที่ยงระหว่างผู้ตรวจอยู่ระหว่าง 0.496 - 0.819 พบว่าข้อสอบส่วนใหญ่อยู่ในระดับสูง เว้นแต่ข้อ 3 ที่มี ความเที่ยงระหว่างผู้ตรวจอยู่ในระดับปานกลาง ซึ่งแสดงให้เห็นถึงความสัมพันธ์ระดับปานกลางถึงระดับสูงมาก จะเห็นได้ว่าระบบตรวจให้คะแนนอัตโนมัติมีความสอดคล้องกับผู้ตรวจที่เป็นคนในระดับที่น่าเชื่อถือที่ระดับนัยสำคัญของทุกข้อ (Sig. 2-tailed < 0.001) อยู่ต่ำกว่าระดับนัยสำคัญทางสถิติที่กำหนดไว้ที่ 0.05 ดัง **Table 1**

**Table 1** ผลการตรวจสอบความเที่ยงระหว่างผู้ตรวจ (Inter-rater reliability) ของผู้ตรวจ 3 คนและระบบตรวจให้คะแนนอัตโนมัติ จากค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน (Pearson Correlation Coefficient)

| ข้อสอบ | Pearson Correlation Coefficient | P-value | ความเที่ยงระหว่างผู้ตรวจกับระบบ |
|---|---|---|---|
| 1 | 0.558 | <0.001 | สูง |
| 2 | 0.738 | <0.001 | สูง |
| 3 | 0.496 | <0.001 | ปานกลาง |
| 4 | 0.574 | <0.001 | สูง |
| 5 | 0.819 | <0.001 | สูง |
| 6 | 0.813 | <0.001 | สูง |

3.3 การตรวจสอบความเที่ยงระหว่างผู้ตรวจ (Inter-rater reliability) จากค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน (Pearson Correlation Coefficient) แบบรายบุคคล ระหว่างผู้ตรวจ 3 คน ระบบตรวจให้คะแนนอัตโนมัติ และคะแนนเกณฑ์การตรวจ ซึ่งแปลผลขนาดของความสัมพันธ์ตามเกณฑ์ของโคเฮน (Cohen, 1988) คือ 0.10 – 0.30 อยู่ในระดับต่ำ 0.31 – 0.50 อยู่ในระดับปานกลาง และ 0.51 – 1.00 อยู่ในระดับสูง โดยพบว่าความสัมพันธ์ระหว่างผู้ตรวจ (R1, R2, R3) มีความเที่ยงอยู่ในระดับปานกลางถึงสูง โดยเฉพาะความเที่ยงระหว่างผู้ตรวจที่ 1 และ 2 (R1 & R2) และความเที่ยงระหว่างผู้ตรวจที่ 2 และ 3 (R2 & R3) ที่มีความเที่ยงสูงอันแสดงถึงความสอดคล้องในการให้คะแนนกันอย่างชัดเจน ซึ่งบ่งชี้ถึงความน่าเชื่อถือในการประเมินร่วมกัน ส่วนความเที่ยงระหว่างผู้ตรวจทั้ง 3 คน กับระบบตรวจให้คะแนนอัตโนมัติอยู่ในระดับปานกลางถึงสูง โดยผู้ตรวจที่ 2 (R2) มีความใกล้เคียงกับระบบตรวจให้คะแนนอัตโนมัติมากที่สุด และผู้ตรวจคนที่ 3 (R3) มีความสอดคล้องต่ำสุด เมื่อพิจารณาจากคะแนนเกณฑ์ (Y) พบว่าความสัมพันธ์กับทั้งผู้ตรวจและความสัมพันธ์กับระบบตรวจให้คะแนนอัตโนมัติมีแนวโน้มลดลง โดยเฉพาะในข้อ 3 และ 4 ที่ค่าความเที่ยงต่ำกว่าคู่อื่น ๆ ดัง **Table 2**

**Table 2** ผลการตรวจสอบความเที่ยงระหว่างผู้ตรวจ (Inter-rater reliability) แบบรายบุคคล ระหว่างผู้ตรวจ 3 คน ระบบตรวจให้คะแนนอัตโนมัติ และคะแนนเกณฑ์การตรวจ

| ผู้ตรวจ | Pearson Correlation Coefficient | | | | | | P-value (Sig. 2-tailed) | ความเที่ยงระหว่างผู้ตรวจกับระบบ |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | | |
| R1 & R2 | 0.685 | 0.834 | 0.759 | 0.852 | 0.737 | 0.639 | <0.001 | สูง |
| R2 & R3 | 0.657 | 0.777 | 0.673 | 0.975 | 0.708 | 0.549 | <0.001 | สูง |
| R1 & R3 | 0.738 | 0.796 | 0.373 | 0.734 | 0.763 | 0.631 | <0.001 | ปานกลาง – สูง |
| R1 & S | 0.507 | 0.707 | 0.377 | 0.594 | 0.782 | 0.731 | <0.001 | ปานกลาง – สูง |
| R2 & S | 0.519 | 0.706 | 0.519 | 0.552 | 0.774 | 0.762 | <0.001 | สูง |
| R3 & S | 0.472 | 0.652 | 0.384 | 0.489 | 0.677 | 0.609 | <0.001 | ปานกลาง – สูง |
| Y & R1 | 0.753 | 0.831 | 0.257 | 0.319 | 0.697 | 0.705 | <0.001 | ต่ำ - สูง |
| Y & R2 | 0.723 | 0.807 | 0.264 | 0.280 | 0.713 | 0.690 | <0.001 | ต่ำ - สูง |
| Y & R3 | 0.687 | 0.781 | 0.175* | 0.242 | 0.685 | 0.598 | <0.001 | ต่ำ - สูง |
| Y & S | 0.545 | 0.704 | 0.183 | 0.230 | 0.663 | 0.748 | <0.001 | ต่ำ - สูง |

หมายเหตุ: * P-value (Sig. 2-tailed) = 0.002

โดยกำหนดให้ R1 – R3 แทนผู้ตรวจคนที่ 1 – 3

S แทนระบบตรวจให้คะแนนอัตโนมัติ

Y แทน คะแนนเกณฑ์ที่ตรวจโดยผู้เชียวชาญจากระยะที่ 1

3.4 การหาความสอดคล้องของการตรวจให้คะแนนเพื่อวิเคราะห์หาความเที่ยงระหว่างผู้ตรวจ (Inter-rater reliability) ทั้งสองรูปแบบ โดยใช้สถิติสหสัมพันธ์ภายในชั้น (Intraclass Correlation Coefficient : ICC) โมเดลอิทธิพลแบบผสมสองทาง ( Two-Way Mixed-Effects Model) ที่ช่วงความเชื่อมั่น 95% (Confident interval) โดยใช้ผลการตรวจจากผู้ตรวจที่เป็นผู้เชี่ยวชาญจำนวน 3 คน ผลการตรวจจากระบบตรวจให้คะแนนอัตโนมัติ และคะแนนที่เป็นเกณฑ์ในการตรวจ (Y) ที่เป็นผลการตรวจจากผู้เชี่ยวชาญที่เป็นผู้ตรวจในระยะที่ 1 ที่ตรวจด้วยรูปแบบการตรวจคำตอบทุกข้อของผู้สอบทุกคน และใช้เกณฑ์การแปลความหมายของ Koo and Li (2016) คือ น้อยกว่า 0.5 มีความสอดคล้องต่ำ,

0.5 - 0.74 มีความสอดคล้องปานกลาง, 0.75 - 0.89 มีความสอดคล้องดี และ 0.90 ขึ้นไป มีความสอดคล้องดีเยี่ยม จากการวิเคราะห์พบว่ามีค่า ICC เฉลี่ยสูงถึง 0.914 และ 0.930 ตามลำดับ ซึ่งจัดอยู่ในระดับดีเยี่ยม แสดงให้เห็นว่าการให้คะแนนของผู้ตรวจมีความเที่ยงสูง ดัง **Table 3**

**Table 3** ผลการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ภายในชั้น (Intraclass Correlation Coefficient: ICC) ของการตรวจให้คะแนน

| ผู้ตรวจ | ICC รายข้อ | | | | | | ความสอดคล้อง | ICC เฉลี่ย | ความสอดคล้อง |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | | | |
| R1 – R3 | 0.862 | 0.923 | 0.815 | 0.945 | 0.890 | 0.819 | ดี - ดีเยี่ยม | 0.914 | ดีเยี่ยม |
| R1 – R3 & S | 0.847 | 0.920 | 0.804 | 0.905 | 0.912 | 0.876 | ดี - ดีเยี่ยม | 0.930 | ดีเยี่ยม |
| R1 & R2 | 0.685 | 0.834 | 0.759 | 0.852 | 0.737 | 0.639 | ปานกลาง - ดี | 0.877 | ดี |
| R2 & R3 | 0.657 | 0.777 | 0.673 | 0.975 | 0.708 | 0.549 | ปานกลาง – ดีเยี่ยม | 0.871 | ดี |
| R1 & R3 | 0.738 | 0.796 | 0.373 | 0.287 | 0.438 | 0.493 | ต่ำ - ดี | 0.872 | ดี |
| R1 & S | 0.507 | 0.707 | 0.377 | 0.594 | 0.782 | 0.731 | ต่ำ - ดี | 0.868 | ดี |
| R2 & S | 0.519 | 0.706 | 0.519 | 0.552 | 0.774 | 0.762 | ปานกลาง - ดี | 0.860 | ดี |
| R3 & S | 0.472 | 0.652 | 0.384 | 0.265 | 0.489 | 0.609 | ต่ำ - ปานกลาง | 0.851 | ดี |
| Y & R1 | 0.753 | 0.831 | 0.257 | 0.319 | 0.697 | 0.705 | ต่ำ - ดี | 0.862 | ดี |
| Y & R2 | 0.723 | 0.807 | 0.264 | 0.280 | 0.713 | 0.690 | ต่ำ - ดี | 0.855 | ดี |
| Y & R3 | 0.687 | 0.781 | 0.175 | 0.242 | 0.685 | 0.598 | ต่ำ - ปานกลาง | 0.849 | ดี |
| Y & S | 0.545 | 0.704 | 0.183 | 0.230 | 0.663 | 0.748 | ต่ำ - ปานกลาง | 0.842 | ดี |

โดยกำหนดให้ R1 – R3 แทนผู้ตรวจคนที่ 1 – 3

S แทนระบบตรวจให้คะแนนอัตโนมัติ

Y แทน คะแนนเกณฑ์ที่ตรวจโดยผู้เชียวชาญที่ตรวจในระยะที่ 1

3.5 ผลการวิเคราะห์ความแปรปรวนของแบบสอบและเปรียบเทียบค่าความเที่ยงสัมประสิทธิ์ การสรุปอ้างอิงความน่าเชื่อถือของผลการวัด (Generalizability Coefficient) รูปแบบการตรวจ (P x i x r) ซึ่งได้วิเคราะห์ทั้งค่า G-study และ D-study โดยวิเคราะห์ใน 2 รูปแบบ ได้แก่ 1. แบบผู้ตรวจ 3 คน และ 2. แบบผู้ตรวจ 3 คนและระบบตรวจให้คะแนนอัตโนมัติ โดยผู้ตรวจตรวจให้คะแนนทุกข้อของผู้สอบทุกคน มีผลดังนี้

1) ผลการวิเคราะห์ G-Study (p x i x r)

เมื่อวิเคราะห์องค์ประกอบจากแหล่งความแปรปรวนของแบบสอบเขียนตอบ วิชาภาษาไทย ระดับมัธยมศึกษาตอนต้น ที่ใช้รูปแบบการตรวจให้คะแนนจากผู้ตรวจจำนวน 3 คน พบว่า มีความแปรปรวนรวมทั้งหมดเท่ากับ 2.426 โดยความแปรปรวนจากปฏิสัมพันธ์ระหว่างผู้สอบ ข้อสอบ และผู้ตรวจ มีมากที่สุด เท่ากับ 1.069 คิดเป็นร้อยละ 44.100 รองลงมา คือ ความแปรปรวนจากผู้สอบ เท่ากับ 0.744 คิดเป็นร้อยละ 30.700 และองค์ประกอบที่มีความแปรปรวนน้อยที่สุด คือ ความแปรปรวนที่เกิดจากผู้ตรวจ เท่ากับ 0.005 คิดเป็น ร้อยละ 0.200 โดยสามารถแสดงความสัมพันธ์ของค่าความแปรปรวน ดัง **Figure 4**

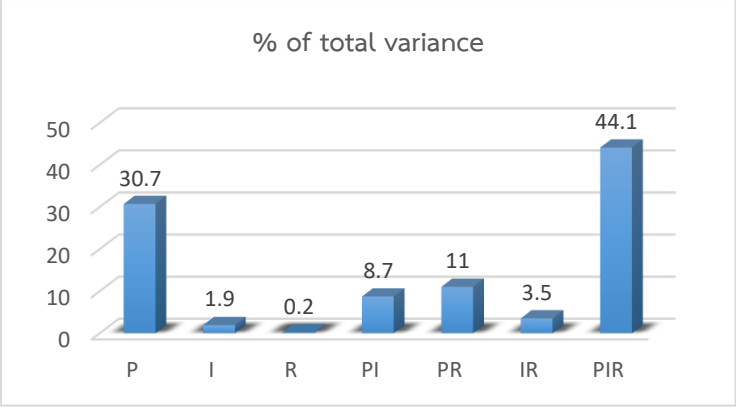


**Figure 4** ความแปรปรวนจาก G-Study (p x i x r) ของข้อสอบ 6 ข้อ และผู้ตรวจจำนวน 3 คน

เมื่อวิเคราะห์องค์ประกอบจากแหล่งความแปรปรวนในแบบสอบเขียนตอบ วิชาภาษาไทย ระดับมัธยมศึกษาตอนต้น จากผลตรวจที่ใช้รูปแบบการตรวจให้คะแนนจากผู้ตรวจจำนวน 4 คน (ผู้ตรวจ 3 คนและระบบตรวจให้คะแนนอัตโนมัติ) พบว่า มีความแปรปรวนรวมทั้งหมดเท่ากับ 2.211 โดยความแปรปรวนจากปฏิสัมพันธ์ระหว่างผู้สอบ ข้อสอบ และผู้ตรวจ มีมากที่สุด เท่ากับ 1.295 คิดเป็นร้อยละ 58.500 รองลงมาคือ ความแปรปรวนจากผู้สอบเท่ากับ 0.734 คิดเป็นร้อยละ 33.200 ผู้ตรวจมีความแปรปรวนเท่ากับ 0.023 คิดเป็นร้อยละ 1.100 และองค์ประกอบที่มีความแปรปรวนน้อยที่สุด คือ ความแปรปรวนที่เกิดจากปฏิสัมพันธ์ระหว่างผู้สอบและข้อสอบ เท่ากับ 0.009 คิดเป็นร้อยละ 0.400 โดยสามารถแสดงความสัมพันธ์ของค่าความแปรปรวนดัง **Figure 5**

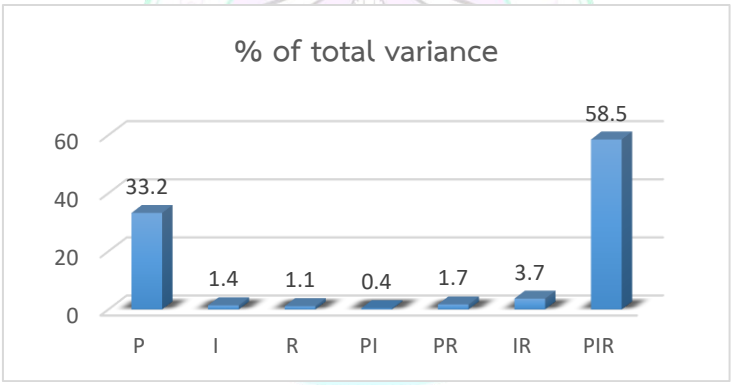


**Figure 5** ความแปรปรวนจาก G-Study (p x i x r) ของข้อสอบ 6 ข้อ และผู้ตรวจจำนวน 4 คน (ผู้ตรวจจำนวน 3 คนและระบบตรวจให้คะแนนอัตโนมัติ)

จาก **Figure 4 และ Figure 5** จะเห็นได้ว่าการวิเคราะห์องค์ประกอบของแหล่งความแปรปรวนของแบบสอบเขียนตอบ วิชาภาษาไทย ระดับมัธยมศึกษาตอนต้น จากการตรวจ 2 รูปแบบ ได้แก่ การตรวจให้คะแนนโดยผู้ตรวจจำนวน 3 คน และ การตรวจให้คะแนนโดยผู้ตรวจจำนวน 3 คนร่วมกับระบบตรวจให้คะแนนอัตโนมัติ พบว่า ทั้งสองรูปแบบมีแหล่งความแปรปรวนสูงสุดจากปฏิสัมพันธ์ระหว่างผู้สอบ ข้อสอบ และผู้ตรวจ (p x i x r) คือ 1.069 (ร้อยละ 44.1) และ 1.295 (ร้อยละ 58.5) ตามลำดับ แสดงว่าการเพิ่มระบบตรวจอัตโนมัติทำให้เกิดความแปรปรวนมากขึ้นเล็กน้อย แต่ในทางตรงกันข้าม ความแปรปรวนจากผู้สอบ (P) ลดลงเล็กน้อยเมื่อเพิ่มระบบอัตโนมัติเข้าไป โดยลดจาก 0.744 (ร้อยละ 30.7) เหลือ 0.734 (ร้อยละ 33.2) อย่างไรก็ตาม แบบสอบสามารถแยกแยะความสามารถผู้สอบได้ดีในการตรวจทั้งสองสถานการณ์ และความแปรปรวนจากผู้ตรวจ (R) อยู่ในระดับต่ำมาก (0.005 หรือร้อยละ 0.2 ในกรณีผู้ตรวจ 3 คน และ 0.009 หรือร้อยละ 0.4 เมื่อรวมคนกับระบบอัตโนมัติ) สะท้อนให้เห็นว่ามีความแปรปรวนน้อยมาก ดังนั้น ทั้งผู้ตรวจและระบบมีแนวโน้มการให้คะแนนอย่างสม่ำเสมอ

การเปรียบเทียบคุณภาพของการตรวจระหว่างผลที่ได้จากผู้ตรวจและผลที่ได้จากระบบตรวจให้คะแนนอัตโนมัติ โดยวิเคราะห์ตามทฤษฎี G-Theory จากสัมประสิทธิ์การสรุปอ้างอิงความน่าเชื่อถือของผลการวัด (Generalizability Coefficient) แบบ P x i x r design ในรูปแบบ G-Study พบว่า เมื่อกำหนดรูปแบบผู้ตรวจ 3 คน ค่าสัมประสิทธิ์สรุปอ้างอิงสำหรับการตัดสินใจสัมพัทธ์ (Generalizability Coefficient: $\boldsymbol{\rho_{\delta}^{2}}$) เท่ากับ 0.26 และค่าสัมประสิทธิ์สรุปอ้างอิงสำหรับการตัดสินใจสัมบูรณ์ (Dependability Coefficient: $\boldsymbol{\rho_{\Delta}^{2}}$) เท่ากับ 0.17 และรูปแบบผู้ตรวจ 3 คนร่วมกับระบบตรวจให้คะแนนอัตโนมัติมีค่าสัมประสิทธิ์สรุปอ้างอิงสำหรับการตัดสินใจสัมพัทธ์ ($\boldsymbol{\rho_{\delta}^{2}}$) เท่ากับ 0.62 และค่าสัมประสิทธิ์สรุปอ้างอิงสำหรับการตัดสินใจสัมบูรณ์ ($\boldsymbol{\rho_{\Delta}^{2}}$) เท่ากับ 0.51 แสดงว่า การนำระบบตรวจให้คะแนนอัตโนมัติมาร่วมตรวจกับการตรวจโดยผู้ตรวจทำให้มีค่าสัมประสิทธิ์การสรุปอ้างอิงความน่าเชื่อถือเพิ่มขึ้น ดัง **Figure 6**

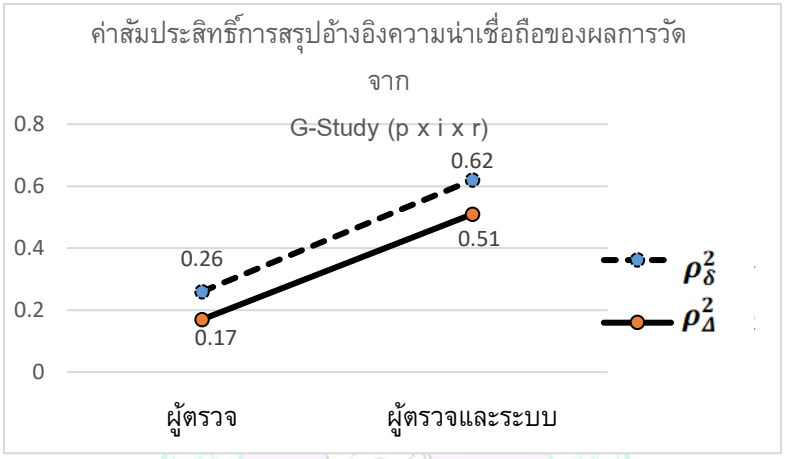


**Figure 6** ค่าสัมประสิทธิ์การสรุปอ้างอิงความน่าเชื่อถือของผลการวัด G-Study (p x i x r)

2) ผลการวิเคราะห์ D-Study (p x i x r)

ผลการวิเคราะห์สัมประสิทธิ์การสรุปอ้างอิงความน่าเชื่อถือของผลการวัด รูปแบบ D-Study พบว่า การเพิ่มจำนวนข้อสอบมีผลอย่างชัดเจนต่อการเพิ่มระดับความเที่ยงของเครื่องมือวัดจากสัมประสิทธิ์สรุปอ้างอิงสำหรับการตัดสินใจสัมพัทธ์ ( $\boldsymbol{\rho_{\delta}^{2}}$) เมื่อใช้เพียงผู้ตรวจคะแนนสัมประสิทธิ์จะอยู่ในระดับต่ำ และเพิ่มขึ้นอย่างต่อเนื่องเมื่อจำนวนข้อสอบมากขึ้นจาก 0.26 (ข้อสอบ 6 ข้อ) เพิ่มเป็น 0.64 (ข้อสอบ 40 ข้อ) ขณะที่การใช้ทั้งผู้ตรวจร่วมกับระบบตรวจให้คะแนนอัตโนมัติ ค่าดังกล่าวจะอยู่ในระดับสูงกว่าอย่างชัดเจน โดยเพิ่มจาก 0.62 (ข้อสอบ 6 ข้อ) เป็น 0.91 (ข้อสอบ 40 ข้อ) สะท้อนให้เห็นว่า การนำเทคโนโลยีระบบตรวจให้คะแนนอัตโนมัติมาใช้ร่วมกับผู้ตรวจจะเพิ่มความเที่ยงของการวัดและความน่าเชื่อถือของผลการตัดสินใจมากยิ่งขึ้น ดัง **Figure 7**

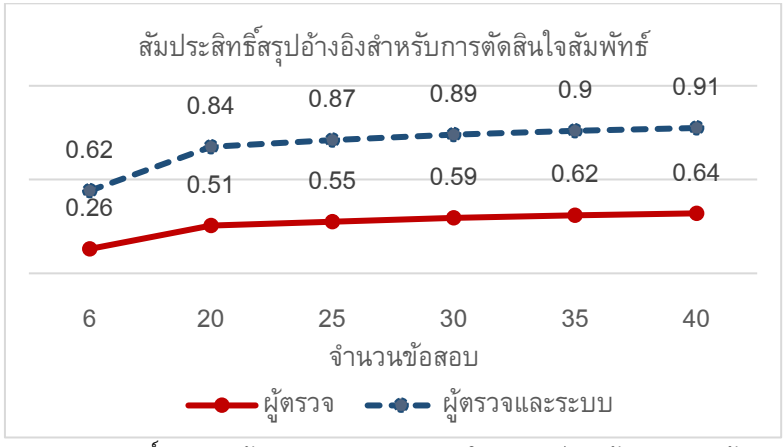


**Figure 7** ค่าสัมประสิทธิ์การสรุปอ้างอิงสำหรับการตัดสินใจสัมพัทธ์ของผู้ตรวจ และผู้ตรวจกับระบบ

ขณะเดียวกันค่าสัมประสิทธิ์สรุปอ้างอิงสำหรับการตัดสินใจสัมบูรณ์ ($\boldsymbol{\rho_{\Delta}^{2}}$) มีแนวโน้มในทิศทางเดียวกัน คือ การเพิ่มจำนวนข้อสอบและการใช้ระบบตรวจให้คะแนนอัตโนมัติร่วมกับผู้ตรวจส่งผลให้ค่าความเที่ยงเชิงสัมบูรณ์สูงขึ้นอย่างต่อเนื่อง โดยสำหรับการใช้ผู้ตรวจเพียงอย่างเดียวจะมีค่าจาก 0.17 (ข้อสอบ 6 ข้อ) เป็น 0.45 (ข้อสอบ 40 ข้อ) และหากใช้ทั้งสองวิธีร่วมกันจะเพิ่มขึ้นจาก 0.51 (ข้อสอบ 6 ข้อ) เป็น 0.81 (ข้อสอบ 40 ข้อ) การเพิ่มจำนวนข้อสอบและการนำระบบตรวจให้คะแนนอัตโนมัติมาใช้ควบคู่กับผู้ตรวจจะเพิ่มความเที่ยงและความน่าเชื่อถือของเครื่องมือวัดอย่างมีนัยสำคัญ แสดงว่าในการตรวจข้อสอบเขียนตอบควรพิจารณาเพิ่มจำนวนข้อสอบ พร้อมกับใช้เทคโนโลยีในการตรวจให้คะแนนร่วมกับผู้ตรวจที่เป็นคน เพื่อส่งเสริมคุณภาพและความเที่ยงตรงในการประเมินผลการเรียนรู้ให้สูงที่สุด ดัง **Figure 8**

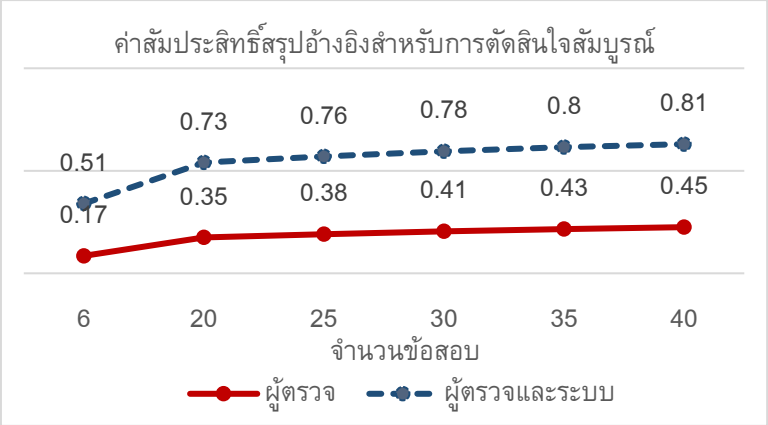


**Figure 8** ค่าสัมประสิทธิ์การสรุปอ้างอิงสำหรับการตัดสินใจสัมบูรณ์ของผู้ตรวจ และผู้ตรวจกับระบบ

## อภิปราย และข้อเสนอแนะ

แบบสอบเขียนตอบ วิชาภาษาไทย ระดับมัธยมศึกษาตอนต้น ที่พัฒนาขึ้น มีคุณภาพอยู่ในระดับที่เหมาะสมต่อ การใช้เป็นเครื่องมือวัดผลสัมฤทธิ์ทางการเรียนของผู้เรียน โดยข้อสอบทุกข้อมีความยาก (p) อยู่ระหว่าง 0.43 – 0.70 ซึ่งอยู่ในเกณฑ์ที่เหมาะสม แสดงให้เห็นว่าข้อสอบมีระดับความยากที่เหมาะสม สามารถทดสอบความสามารถของผู้เรียนได้อย่างมีประสิทธิภาพ และมีอำนาจจำแนก (r) อยู่ระหว่าง 0.20 – 0.34 ซึ่งอยู่ในระดับที่ยอมรับได้ตามเกณฑ์ที่เหมาะสม ถือว่าอยู่ในระดับที่สามารถแยกแยะผู้เรียนที่มีความสามารถสูงและต่ำได้อย่างมีประสิทธิภาพ สามารถสะท้อนผลสัมฤทธิ์ทางการเรียนของผู้เรียนอย่างแท้จริง สอดคล้องกับ Kanchanawasri (2013) ที่กล่าวว่า ข้อสอบที่มีคุณภาพควรมีค่าความยากอยู่ในช่วงระหว่าง 0.20 – 0.80 และมีอำนาจจำแนกตั้งแต่ 0.20 ขึ้นไป ผลการตรวจสอบความตรงเชิงเนื้อหาโดยผู้เชี่ยวชาญ พบว่า ทุกข้อมีค่า IOC เท่ากับ 1.00 ซึ่งถือว่าอยู่ในเกณฑ์ที่ดีมาก สอดคล้องกับ Rovinelli & Hambleton (1976) ที่ระบุว่าข้อสอบที่มีค่า IOC ตั้งแต่ 0.50 ขึ้นไป ถือว่าอยู่ในระดับที่ยอมรับได้ ซึ่งสะท้อนว่าแบบสอบมีความสอดคล้องกับจุดประสงค์ที่ตั้งไว้และครอบคลุมเนื้อหาที่ต้องการวัด ทั้งนี้ เนื่องจากผู้วิจัยสร้างแบบสอบที่ตรงตามแผนผังออกข้อสอบ และมีการวิเคราะห์ตัวชี้วัดจากหลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน พุทธศักราช 2551 ฉบับปรับปรุงแก้ไข พ.ศ. 2566 ในกลุ่มสาระการเรียนรู้ภาษาไทย ระดับมัธยมศึกษาปีที่ 1 – 3 และจากการวิเคราะห์ค่าความเที่ยงของแบบสอบด้วยค่าสัมประสิทธิ์แอลฟาของครอนบาค (Cronbach’s Alpha Coefficient) พบว่ามีค่าเท่ากับ 0.731 ซึ่งอยู่ในระดับที่ยอมรับได้ สอดคล้องกับเกณฑ์ของ Cronbach (1951) ที่กล่าวว่าแบบวัดที่มีคุณภาพควรมีความเที่ยง 0.70 ขึ้นไป ผลการตรวจสอบคุณภาพของแบบสอบเขียนตอบครั้งนี้แสดงถึงความสม่ำเสมอและน่าเชื่อถือสำหรับการนำไปใช้วัดและประเมินผลการเรียนรู้ของแบบสอบ ดังนั้นผลการวิจัยแสดงให้เห็นว่าแบบสอบเขียนตอบที่พัฒนาขึ้นมีคุณภาพในทุกด้าน สามารถนำไปใช้เป็นเครื่องมือวัดผลได้อย่างเหมาะสม

เกณฑ์การให้คะแนนแบบสอบเขียนตอบ วิชาภาษาไทย ระดับมัธยมศึกษาตอนต้น เป็นเกณฑ์การประเมินแบบพิจารณารายละเอียด (Analytic Scoring rubrics) ที่เน้นการพิจารณารายด้านอย่างเป็นระบบ ซึ่งสอดคล้องกับ Thongsilp et al. (2020) ที่กล่าวว่า เกณฑ์การประเมินแบบพิจารณารายละเอียด หมายถึง การพิจารณาที่คุณภาพของผลการปฏิบัติงาน ผลงานหรือชิ้นงาน ในลักษณะการพิจารณารายละเอียด และสอดคล้อง Chansima & Tuksino (2019) ที่กล่าวว่าเป็นการพิจารณาจากแต่ละส่วนของงาน ซึ่งกำหนดแนวทางการให้คะแนนจากคำนิยามหรือคำอธิบายลักษณะของส่วนงานนั้น ๆ ในแต่ละระดับคะแนนไว้อย่างชัดเจน และเกณฑ์การตรวจให้คะแนนในแต่ละวัตถุประสงค์มีการกำหนดองค์ประกอบและเกณฑ์การให้คะแนนอย่างชัดเจน ทั้งในด้านเนื้อหา ภาษา ความถูกต้องของรูปแบบ และหลักการทางภาษาไทย ในการออกแบบเกณฑ์ตรวจมีการกำหนดเงื่อนไขที่เน้นให้ความสำคัญกับเนื้อหาเป็นลำดับแรก เพื่อให้การตรวจแบบมีลำดับที่จะหยุดการให้คะแนนในเกณฑ์อื่นหากไม่ผ่านเงื่อนไขหลัก เพื่อความเป็นธรรมและลดความลำเอียง สอดคล้องกับ Wiboonsri (2013) ที่กล่าวว่า การใช้เกณฑ์เชิงเนื้อหาในการให้คะแนนแบบสอบความเรียงจะช่วยชี้ถึงความถูกต้องและความพอเพียงของเนื้อหาความรู้ที่ผู้ตอบควรจะมีในขอบข่ายของคำถามที่ต้องการ และเมื่อพิจารณาผลการตรวจสอบ ความตรงเชิงเนื้อหา (IOC) จากผู้เชี่ยวชาญ พบว่าเกณฑ์ส่วนใหญ่มีค่า IOC เท่ากับ 1.00 ซึ่งถือว่าอยู่ในระดับสูงมาก แสดงถึงความเห็นพ้องของผู้เชี่ยวชาญว่าข้อสอบวัดได้ตรงตามวัตถุประสงค์ และบางเกณฑ์ที่มีค่าเป็น 0.67 แต่ยังอยู่ในระดับเหมาะสม สอดคล้องกับ Rovinelli & Hambleton (1976) ที่กำหนดเกณฑ์ความตรงเชิงเนื้อหาที่ยอมรับทั่วไป คือ ไม่ต่ำกว่า 0.50 แสดงว่าแบบสอบมีความเหมาะสมและครอบคลุมเป้าหมายด้านเนื้อหาได้เป็นอย่างดี นอกจากนี้ ความเที่ยงของเกณฑ์การตรวจทั้งฉบับมีค่าสัมประสิทธิ์แอลฟาของครอนบาค (Cronbach's Alpha) เท่ากับ 0.731 ซึ่งสูงกว่าเกณฑ์ขั้นต่ำที่นิยมใช้อ้างอิงว่าอยู่ในระดับน่าเชื่อถือ (≥ 0.70) ตามข้อเสนอของ Cronbach (1951) แสดงว่าเกณฑ์การตรวจมีความสอดคล้องภายในที่ดี และสามารถนำไปใช้ประเมินผลการเรียนของนักเรียนได้อย่างมีความน่าเชื่อถือ แบบสอบเขียนตอบนี้มีทั้งความตรงเชิงเนื้อหาและความเที่ยงในระดับเหมาะสม ช่วยเสริมสร้างความน่าเชื่อถือของผลการประเมิน และสอดคล้องกับ Wiboonsri (2013) คือนอกจากการใช้เกณฑ์ประเมินในการตรวจแบบสอบแล้ว ผู้สร้างแบบสอบต้องพิจารณาถึงความเที่ยงระหว่างผู้ตรวจให้คะแนนเพื่อให้มีความเที่ยงที่มากขึ้นและลดความคลาดเคลื่อนของการตรวจให้น้อยลง ความคลาดเคลื่อนมีสาเหตุหลายประการ เช่น ภาวะความเหนื่อยอ่อน ความลำเอียง ความคาดหวังของผู้ตรวจ

ระบบตรวจให้คะแนนอัตโนมัติสำหรับแบบสอบเขียนตอบ วิชาภาษาไทย ระดับมัธยมศึกษาตอนต้น มีชื่อระบบว่า Automated Scoring System for Writing Test (ASSWT) ออกแบบและพัฒนาอย่างสอดคล้องกับความต้องการของผู้ใช้งานทั้งครูและนักเรียน ระบบแบ่งออกเป็น 3 ส่วนหลัก ได้แก่ ส่วนของครู มีการสร้างข้อสอบ กำหนดเกณฑ์ให้คะแนนตรวจคำตอบด้วยตนเองหรือผ่านระบบ และจัดการข้อมูลนักเรียนได้อย่างครบถ้วน ส่วนของผู้เข้าสอบที่มีขั้นตอนชัดเจนตั้งแต่เข้าสู่ระบบ เตรียมสอบ อ่านคำชี้แจง ทำแบบทดสอบ ตรวจคำตอบ และรับผลการสอบ โดยแสดงผลแบบทันที ส่วนสุดท้ายคือโครงสร้างทางเทคนิคของระบบ ส่วนนี้พัฒนาบนระบบปฏิบัติการ Windows และตรวจให้คะแนนด้วยโมเดล GPT-o3-mini จาก OpenAI ซึ่งมีความสามารถในการให้เหตุผลและวิเคราะห์คำตอบอย่างแม่นยำและรวดเร็ว จะเห็นได้ว่า ระบบ ASSWT เป็นระบบที่ตอบสนองต่อความต้องการของผู้ใช้ได้อย่างมีประสิทธิภาพ และสามารถต่อยอดโดยการนำไปใช้งานจริงในสภาพแวดล้อมการเรียนการสอน ขณะเดียวกัน แม้ระบบตรวจข้อสอบและรายงานผลที่พัฒนาขึ้นจะสามารถใช้งานได้ตามวัตถุประสงค์ แต่ยังมีข้อจำกัดในด้านการใช้งานของนักเรียนและข้อจำกัดด้านเทคนิคบางประการ ซึ่งในกรณีนี้ นักเรียนบางคนไม่สามารถใช้งานระบบได้อย่างราบรื่น เนื่องจากขาดความเข้าใจในการใช้งานและไม่ได้ใส่ใจคำชี้แจงที่ได้รับ ทำให้ต้องอาศัยการแนะนำเพิ่มเติมจากครูผู้คุมสอบ และการปรับปรุงระบบให้สามารถส่งข้อสอบอัตโนมัติเมื่อหมดเวลา และบันทึกคำตอบอัตโนมัติเพื่อป้องกันการสูญหายของข้อมูล แสดงถึงการพัฒนาที่เน้นการออกแบบระบบโดยยึดผู้ใช้เป็นศูนย์กลาง เพื่อให้สามารถตอบสนองความต้องการที่เกิดขึ้นจริงในการใช้งานด้านการนำเสนอผลคะแนน เป็นการใช้แถบสีช่วยให้ผู้ใช้งานเข้าใจข้อมูลได้ง่ายขึ้น ขณะเดียวกันการเสนอผลคะแนนในรูปแบบค่าร้อยละจะเป็นการช่วยให้ครูนำคะแนนไปใช้ประเมินผลหรือเทียบกับเกณฑ์ต่าง ๆ ได้สะดวกและเป็นมาตรฐานมากขึ้น อย่างไรก็ตาม ข้อจำกัดของระบบในการรองรับปริมาณผู้ใช้งาน

จำนวนมากพร้อมกัน ซึ่งส่งผลให้การประมวลผลล่าช้า แสดงให้เห็นถึงความจำเป็นในการใช้ระบบที่มีความยืดหยุ่นและสามารถเพิ่มจำนวนได้ สอดคล้องกับ Crossley et al. (2019) ที่ว่า ระบบ Language Tool สามารถตรวจจับข้อผิดพลาดในข้อความได้อย่างมีประสิทธิภาพ อย่างไรก็ตาม แม้ความแม่นยำจะอยู่ในระดับสูง แต่ระบบตรวจสอบข้อผิดพลาดกลับอยู่ในระดับต่ำ โดยเฉพาะในกลุ่มข้อผิดพลาดเกี่ยวกับเครื่องหมายวรรคตอน แม้ Language Tool จะไม่สามารถตรวจพบข้อผิดพลาดทั้งหมด แต่ข้อผิดพลาดที่ตรวจพบได้สามารถสะท้อนภาพรวมที่แม่นยำของข้อผิดพลาดด้านโครงสร้างและกลไกทางภาษาที่ผู้เขียนทำได้อย่างชัดเจน ดังนั้น การพัฒนาระบบเทคโนโลยีเพื่อการประเมินผลในสถานศึกษาควรคำนึงถึงทั้งประสบการณ์ผู้ใช้ การออกแบบที่รองรับพฤติกรรมจริง ความเสถียรของระบบ และรูปแบบการรายงานผลที่เข้าใจง่าย เพื่อให้สามารถตอบสนองความต้องการของผู้ใช้และนำไปใช้ได้จริงในบริบทการเรียนการสอน

การเปรียบเทียบคุณภาพของการตรวจจากการวิเคราะห์ค่าสถิติเชิงบรรยาย (descriptive statistics) จากการตรวจให้คะแนนทั้งสองรูปแบบพบว่า ส่วนเบี่ยงเบนมาตรฐาน (SD) มีระดับการกระจายคะแนนใกล้เคียงกัน โดยที่ผู้ตรวจอาจมีความแปรปรวนมากกว่าในบางข้อ ซึ่งอาจเกิดจากการพิจารณารูปแบบภาษาอย่างเป็นระบบมากขึ้น สอดคล้องกับ Sukwichai et al. (2023) ที่กล่าวว่าการตรวจให้คะแนนอัตโนมัติ มีข้อจำกัด คือ สามารถตรวจอัตโนมัติได้เฉพาะในกรณีที่มุ่งเน้นไปที่คำตอบที่ถูกของผู้สอบ และนำผลการตอบที่ได้ไปเปรียบเทียบกับเกณฑ์ให้คะแนนที่สอดคล้องกับแผนที่เชิงโครงสร้างในแต่ละระดับ อย่างไรก็ตาม การตรวจโดยระบบตรวจให้คะแนนอัตโนมัติมีช่วงค่าส่วนเบี่ยงเบนมาตรฐานที่แคบกว่าเล็กน้อย สะท้อนถึงแนวโน้มของการให้คะแนนที่สม่ำเสมอมากกว่า ผลการวิเคราะห์อำนาจจำแนกในการตรวจโดยภาพรวม พบว่า การตรวจโดยผู้ตรวจสามารถในการจำแนกผู้สอบได้ดีกว่า ซึ่งอาจเกิดจากความสามารถของมนุษย์ในการตีความเนื้อหา ความคิด และบริบทของผู้ตอบ ได้ลึกซึ้งกว่าระบบตรวจให้คะแนนอัตโนมัติ สอดคล้องกับ Crossley et al. (2019) ที่พบว่า แม้ Language Tool เป็นเครื่องมือที่มีความแม่นยำในการตรวจจับข้อผิดพลาดด้านโครงสร้างและกลไกในงานเขียนจะไม่ตรวจพบข้อผิดพลาดทั้งหมด แต่ข้อผิดพลาดที่พบสามารถสะท้อนภาพรวมของความสามารถในการเขียนของผู้เขียนได้อย่างแม่นยำ ดังนั้น การตรวจให้คะแนนข้อสอบเขียนตอบด้วยผู้ตรวจยังคงแสดงให้เห็นถึงความแม่นยำและสามารถจำแนกผู้เรียนที่ดีกว่าระบบตรวจให้คะแนนอัตโนมัติ อย่างไรก็ตาม ระบบตรวจให้คะแนนอัตโนมัติจะมีแนวโน้มที่ดีและให้ผลลัพธ์ที่ใกล้เคียงในหลายด้าน ซึ่งชี้ให้เห็นถึงศักยภาพของระบบตรวจให้คะแนนอัตโนมัติในการพัฒนาเพื่อนำไปใช้จริงในอนาคต โดยเฉพาะในบริบทที่ต้องการความรวดเร็วและประหยัดแรงงานในการตรวจข้อสอบ สอดคล้องกับ Thongsilp et al. (2020) กล่าวว่าในต่างประเทศมีการแก้ปัญหาการตรวจให้คะแนนแบบสอบเขียนตอบโดยใช้เทคโนโลยีคอมพิวเตอร์เป็นระบบการตรวจให้คะแนนอัตนัยอัตโนมัติ มีผลดี คือ ประหยัดค่าใช้จ่ายและลดความคลาดเคลื่อนจากผู้ตรวจ ให้คะแนนเฉพาะตรงกลาง และใช้ความประทับใจ เป็นต้น และสอดคล้องกับ Sukwichai et al. (2023) ที่กล่าวว่า การตรวจให้คะแนนอัตโนมัติสามารถให้ผลคะแนน คำแนะนำ หรือแหล่งในการเรียนรู้เพิ่มเติม เพื่อให้ผู้เรียนปรับปรุงได้ในทันที

การเปรียบเทียบคุณภาพของการตรวจจากความเที่ยงระหว่างผู้ตรวจ (Inter-rater reliability) ด้วยสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน (Pearson Correlation Coefficient) ระหว่างคะแนนจากการตรวจข้อสอบเขียนตอบโดยผู้ตรวจที่เป็นมนุษย์และระบบตรวจให้คะแนนอัตโนมัติ พบว่า การตรวจข้อสอบมีความสัมพันธ์ระดับปานกลางถึงสูงมาก โดยข้อสอบส่วนใหญ่แสดงให้เห็นความสัมพันธ์ในระดับสูง อันแสดงถึงความสอดคล้องกันระหว่างคะแนนที่ได้จากทั้งสองวิธีการตรวจอย่างมีนัยสำคัญ ยกเว้นข้อที่ 3 ที่มีค่าสัมประสิทธิ์อยู่ในระดับปานกลาง ซึ่งอาจสะท้อนถึงลักษณะคำตอบหรือเกณฑ์การให้คะแนนที่คลุมเครือ ทำให้ทั้งผู้ตรวจและระบบไม่สอดคล้องกันเท่าข้ออื่น ๆ นอกจากนี้ ค่าระดับนัยสำคัญทางสถิติ (Sig. 2-tailed) ของทุกข้อสอบมีค่าน้อยกว่า 0.001 จะเห็นว่าความสัมพันธ์ที่ตรวจพบในทุกข้อสอบมีนัยสำคัญทางสถิติอย่างชัดเจน แสดงว่าระบบตรวจให้คะแนนอัตโนมัติสามารถทำงานได้อย่างมีน่าเชื่อถือในระดับที่ใกล้เคียงกับผู้ตรวจที่เป็นมนุษย์ และเมื่อพิจารณาความเที่ยงระหว่างผู้ตรวจแต่ละราย (Inter-rater reliability) โดยวิเคราะห์แบบรายคู่ระหว่างผู้ตรวจ 3 คน (R1, R2, R3) และเปรียบเทียบกับระบบตรวจอัตโนมัติ พบว่า ความเที่ยงระหว่างผู้ตรวจส่วนใหญ่อยู่ในระดับปานกลางถึงสูง โดยเฉพาะคู่ของผู้ตรวจที่ 1 และ 2 (R1 & R2) และคู่ของผู้ตรวจ ที่ 2 และ 3 (R2 & R3) มีค่าสัมประสิทธิ์อยู่ในระดับสูง แสดงถึง

ความสอดคล้องกันของเกณฑ์และการตัดสินคะแนนอย่างเป็นระบบ ส่วนความเที่ยงระหว่างผู้ตรวจกับระบบตรวจให้คะแนนอัตโนมัติอยู่ในระดับปานกลางถึงสูงเช่นกัน โดยผู้ตรวจที่ 2 มีความสอดคล้องกับระบบอัตโนมัติมากที่สุด ขณะที่ผู้ตรวจที่ 3 มีความสอดคล้องต่ำที่สุด ซึ่งอาจสะท้อนความแตกต่างในการตีความเกณฑ์การให้คะแนนหรือความละเอียดในการประเมินของผู้ตรวจแต่ละคน และในส่วนของคะแนนเกณฑ์ พบว่าความสัมพันธ์ระหว่างคะแนนเกณฑ์กับทั้งผู้ตรวจและระบบตรวจให้คะแนนอัตโนมัติมีแนวโน้มลดลงในบางข้อ โดยเฉพาะข้อที่ 3 และข้อที่ 4 ซึ่งมีค่าความสัมพันธ์ต่ำกว่าคู่อื่น ๆ อันอาจเกิดจากลักษณะของการตีความคำตอบที่หลากหลาย ทำให้การให้คะแนนไม่สามารถสะท้อนความแม่นยำที่ตรงกับเกณฑ์มาตรฐานได้อย่างสมบูรณ์ ซึ่งสอดคล้องกับ Sukwichai et al. (2023) ที่กล่าวว่าการตรวจให้คะแนนอัตโนมัติมีข้อจำกัด คือ สามารถตรวจอัตโนมัติได้เฉพาะในกรณีที่มุ่งเน้นไปที่คำตอบที่ถูกของผู้สอบ และนำผลการตอบที่ได้ไปเปรียบเทียบกับเกณฑ์ให้คะแนนที่สอดคล้องกับแผนที่เชิงโครงสร้างในแต่ละระดับ ดังนั้น ระบบตรวจให้คะแนนอัตโนมัติมีแนวโน้มความสอดคล้องกับการให้คะแนนของผู้ตรวจในระดับที่น่าพึงพอใจ มีความน่าเชื่อถือในระดับหนึ่ง โดยเฉพาะในข้อสอบที่มีเกณฑ์ประเมินชัดเจน อย่างไรก็ตาม ความแตกต่างของความสัมพันธ์ในแต่ละข้อสะท้อนให้เห็นถึงความจำเป็นในการพัฒนาระบบให้สามารถประมวลผลลักษณะการตอบแบบเปิดได้อย่างลึกซึ้งมากขึ้น ตลอดจนต้องมีการพัฒนาและตรวจสอบความสอดคล้องกับเกณฑ์มาตรฐานอย่างต่อเนื่อง เพื่อให้สามารถนำไปใช้ในการประเมินจริงได้อย่างมีประสิทธิภาพ

การเปรียบเทียบคุณภาพของการตรวจจากความเที่ยงระหว่างผู้ตรวจ (Inter-rater reliability) ด้วยสัมประสิทธิ์สหสัมพันธ์ภายในชั้น (ICC) เพื่อประเมินผลความสอดคล้องอันแสดงถึงความน่าเชื่อถือของผู้ตรวจ พบว่า ระดับความน่าเชื่อถือที่อยู่ระดับดีถึงดีเยี่ยม โดยเฉพาะค่าความสอดคล้องระหว่างผู้ตรวจทั้งสามคน (R1–R3 = 0.914) และค่าความสอดคล้องระหว่างผู้ตรวจทั้งหมดกับระบบ (R1–R3 กับ S = 0.930) อยู่ในระดับดีเยี่ยม แสดงให้เห็นว่าผู้ตรวจมีแนวโน้มในการให้คะแนนที่สอดคล้องกันอย่างมาก และระบบตรวจอัตโนมัติสามารถจำลองเกณฑ์ของผู้ตรวจได้ใกล้เคียงในระดับที่น่าเชื่อถือ และเมื่อจับคู่เปรียบเทียบระหว่างผู้ตรวจแต่ละคนกับระบบ มีความสอดคล้องอยู่ระหว่าง 0.851 – 0.868 ซึ่งอยู่ในระดับดี แสดงให้เห็นว่าระบบตรวจสามารถสะท้อนรูปแบบการให้คะแนนของผู้ตรวจได้อย่างเหมาะสมในภาพรวม โดยเฉพาะในข้อที่มีเกณฑ์การประเมินชัดเจนและมีโครงสร้างคำตอบที่สม่ำเสมอ สอดคล้องกับ Thongsilp et al. (2020) ที่กล่าวว่า ความสอดคล้องในการให้คะแนนระหว่างผู้ตรวจ 1, ผู้ตรวจ 2 และระบบ มีความสอดคล้องกันในระดับดีมาก เนื่องจากเกณฑ์การตรวจมีการให้คะแนนชัดเจนและเป็นปรนัย ที่เข้าใจได้ตรงกันทั้งผู้ตรวจและระบบ โดยผู้ตรวจกับระบบนั้นใช้เกณฑ์เดียวกันและมีน้ำหนักคะแนนเท่ากัน ซึ่งเป็นเกณฑ์ที่ไม่ซับซ้อนเข้าใจง่ายทั้งระบบและมนุษย์ ทำให้ผู้ตรวจให้คะแนนได้สอดคล้องและมีความสัมพันธ์กันค่อนข้างสูง และสอดคล้องกับ Broadfoot and Rockey (2025) ที่กล่าวว่าการทดสอบความแม่นยำและความเที่ยงตรงเปรียบเทียบคะแนนจากโปรแกรม 4A กับการประเมินของผู้เชี่ยวชาญมนุษย์โดยใช้การทดสอบ McNemar โดยใช้ค่าสัมประสิทธิ์สหสัมพันธ์ภายในกลุ่ม (ICC) ระหว่างโปรแกรม 4A กับผู้เชี่ยวชาญมนุษย์ มีความสอดคล้องระดับสูง โดยมีความแม่นยำเท่ากับ 0.808 และความเที่ยงตรงเท่ากับ 0.913 เมื่อเทียบกับความแม่นยำ 0.923 และความเที่ยงตรง 1.000 ของผู้เชี่ยวชาญที่เป็นมนุษย์ อย่างไรก็ตาม ค่าความสอดคล้องในระดับปานกลางถึงต่ำปรากฏในบางคู่ของผู้ตรวจ โดยเฉพาะในข้อสอบข้อที่ 3 และ 4 ที่มีค่า ICC ต่ำกว่า 0.5 แสดงถึงความสอดคล้องที่อยู่ในระดับต่ำถึงปานกลาง ความคลาดเคลื่อนของค่าความสอดคล้องในข้อที่ 3 และ 4 แสดงว่าข้อสอบที่เปิดโอกาสให้ผู้เรียนแสดงคำตอบอย่างหลากหลายหรืออาจมีเกณฑ์การให้คะแนนที่ไม่ชัดเจนเพียงพอ และนำไปสู่การตีความที่แตกต่างกันระหว่างผู้ตรวจแต่ละราย และระหว่างผู้ตรวจกับระบบควรเพิ่มความเข้มข้นในการฝึกอบรมผู้ตรวจให้มีมาตรฐานเดียวกันในการให้คะแนน สอดคล้องกับ Sukwichai et al. (2023) ที่กล่าวว่าการตรวจให้คะแนนอัตโนมัติ มีข้อจำกัด คือ สามารถตรวจได้เฉพาะในกรณีที่มุ่งเน้นไปที่คำตอบที่ถูกของผู้สอบ และนำผลการตอบที่ได้ไปเปรียบเทียบกับเกณฑ์ให้คะแนนที่สอดคล้องกับแผนที่เชิงโครงสร้างในแต่ละระดับ และสอดคล้องกับ Crossley et al. (2019) ซึ่งผลการศึกษาพบว่า แม้ระบบ Language Tool จะไม่สามารถตรวจพบข้อผิดพลาดทั้งหมด แต่ข้อผิดพลาดที่ตรวจพบได้สามารถสะท้อนภาพรวมที่แม่นยำของข้อผิดพลาดด้านโครงสร้างและกลไกทางภาษาที่ผู้เขียนทำได้อย่างชัดเจน ดังนั้น การตรวจสอบด้วยค่าสถิติสหสัมพันธ์ภายในชั้น (ICC) ช่วยยืนยันว่าการตรวจข้อสอบแบบเขียนตอบ

โดยผู้ตรวจมีความน่าเชื่อถือในระดับสูง และระบบตรวจให้คะแนนอัตโนมัติมีศักยภาพในการตรวจที่ใกล้เคียงกับผู้ตรวจมนุษย์ในหลายด้าน

การเปรียบเทียบสัมประสิทธิ์การสรุปอ้างอิงความน่าเชื่อถือของผลการวัด (G - Theory) ของผลการตรวจระหว่างผลที่ได้จากผู้ตรวจและผลที่ได้จากระบบตรวจให้คะแนนอัตโนมัติ ด้วยการวิเคราะห์ G-Study โดยเปรียบเทียบผลการตรวจให้คะแนน 2 รูปแบบ ได้แก่ แบบผู้ตรวจ 3 คน และแบบผู้ตรวจ 3 คนกับระบบตรวจให้คะแนนอัตโนมัติ พบว่า แบบผู้ตรวจ 3 คน มีค่าสัมประสิทธิ์การสรุปอ้างอิงสำหรับการตัดสินใจสัมพัทธ์ ($\rho_{\delta}^{2}$) เท่ากับ 0.26 และสำหรับการตัดสินใจสัมบูรณ์ ($\rho_{\Delta}^{2}$) เท่ากับ 0.17 จากผลการวิเคราะห์สะท้อนให้เห็นว่าการให้คะแนนโดยผู้ตรวจเพียงอย่างเดียว ยังมีข้อจำกัดด้านความสม่ำเสมอและความเสถียรของการวัด โดยเฉพาะในการใช้งานจริงที่ต้องการความแม่นยำในการเปรียบเทียบผลระหว่างบุคคลหรือช่วงเวลา ซึ่งสอดคล้องกับ Department of Local Administration (2023) ที่กล่าวว่า ปัญหาของการตรวจให้คะแนนแบบสอบเขียนตอบมีหลายประการ ได้แก่ เกณฑ์ที่กำหนดอาจจะไม่สอดคล้องกับสิ่งที่ต้องการวัดอย่างเพียงพอ ความลำบากในการประเมิน ความถูกต้องและความสมบูรณ์ของการตอบต้องอาศัยเวลาในการตรวจค่อนข้างมาก ซึ่งเป็นการเพิ่มภาระการตรวจให้กับผู้สอน เนื่องจากต้องอ่านเพื่อวิเคราะห์และประเมินคำตอบเป็นค่าคะแนนตามเกณฑ์ที่กำหนด และอาจจะส่งผลให้ขาดความเที่ยงที่เหมาะสมในการตรวจข้อสอบ อันเป็นความคลาดเคลื่อนจากผู้ตรวจ เช่น การอ่านลายมือผิดพลาด ให้คะแนนเพียงส่วนใดส่วนหนึ่ง ให้คะแนนตามความประทับใจ เป็นต้น และสอดคล้องกับ Thongsilp et al. (2020) กล่าวว่า ความคลาดเคลื่อนที่เกิดจากการใช้ความประทับใจส่วนตัว การไม่ชอบส่วนตัว การปล่อยคะแนน การกดคะแนน ผลตกค้างจากการตรวจ ความลำเอียงส่วนตัว อย่างไรก็ตาม เมื่อมีการนำระบบตรวจให้คะแนนอัตโนมัติเข้ามาร่วมในการตรวจ พบว่าค่าสัมประสิทธิ์การสรุปอ้างอิงทั้งสองประเภทเพิ่มขึ้นอย่างมีนัยสำคัญ โดยค่า $\rho_{\delta}^{2}$เพิ่มขึ้นเป็น 0.62 และค่า $\rho_{\Delta}^{2}$ เพิ่มขึ้นเป็น 0.51 ซึ่งชี้ให้เห็นว่า การใช้ระบบตรวจอัตโนมัติร่วมกับการตรวจของมนุษย์ช่วยลดแหล่งความแปรปรวนที่ไม่พึงประสงค์และลดความแปรปรวนในการตัดสินใจของผู้ตรวจ ส่งผลต่อการตรวจให้คะแนนมีความน่าเชื่อถือ และสามารถนำไปใช้ในการตัดสินใจทั้งแบบสัมพัทธ์และสัมบูรณ์ได้อย่างมีประสิทธิภาพมากขึ้น โดยที่ใช้เวลาและงบประมาณในการตรวจน้อยกว่า เนื่องจากการตรวจโดยระบบให้คะแนนอัตโนมัติไม่ต้องมีค่าใช้จ่ายในการเดินทางและจ้างผู้ตรวจ สอดคล้องกับ Thongsilp et al. (2020) ที่กล่าวว่า การแก้ปัญหาในการตรวจให้คะแนนแบบสอบเขียนตอบ โดยใช้เทคโนโลยีคอมพิวเตอร์เป็นระบบการตรวจให้คะแนนอัตนัยอัตโนมัติ มีผลดี คือ ประหยัดค่าใช้จ่ายและลดความคลาดเคลื่อนจากผู้ตรวจ

ผลการเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิงเชิงตัดสินใจในวิธีการตรวจจากผู้ตรวจให้คะแนน และการตรวจจากระบบตรวจให้คะแนนอัตโนมัติ จากผลการวิเคราะห์ D-Study ประเมินค่าความเที่ยงของเครื่องมือวัด พบว่า ทั้ง$\rho_{\delta}^{2}$ และ $\rho_{\Delta}^{2}$ มีค่าเพิ่มขึ้นอย่างต่อเนื่องตามจำนวนข้อสอบที่มากขึ้น แสดงว่าการเพิ่มจำนวนข้อสอบจะช่วยลดความคลาดเคลื่อนจากความบังเอิญและเพิ่มความแม่นยำของการวัด ระบบตรวจให้คะแนนอัตโนมัติมีผลต่อการเพิ่มความเที่ยง ความน่าเชื่อถือและลดความแปรปรวนในการตรวจ เช่น ความลำเอียงหรือข้อจำกัดของผู้ตรวจแต่ละคน สอดคล้องกับ Thongsilp et al. (2020) ที่กล่าวว่า การแก้ปัญหาในการตรวจให้คะแนนแบบสอบเขียนตอบโดยใช้เทคโนโลยีคอมพิวเตอร์เป็นระบบการตรวจให้คะแนนอัตนัยอัตโนมัติมีผลดี คือ ประหยัดค่าใช้จ่ายและลดความคลาดเคลื่อนจากผู้ตรวจ และสอดคล้องกับ Department of Local Administration (2023) ที่กล่าวว่า ปัญหาของการตรวจให้คะแนนแบบสอบเขียนตอบมีหลายประการ ได้แก่ เกณฑ์ที่กำหนดอาจจะไม่สอดคล้องกับสิ่งที่ต้องการวัดอย่างเพียงพอ ความลำบากในการประเมิน ความถูกต้องและความสมบูรณ์ของการตอบต้องอาศัยเวลาในการตรวจค่อนข้างมาก เป็นการเพิ่มภาระการตรวจแก่ผู้สอนและอาจจะส่งผลให้ขาดความเที่ยงในการตรวจข้อสอบที่เหมาะสม ดังนั้น จากการอภิปรายมาข้างต้นชี้ให้เห็นว่า ระบบตรวจให้คะแนนอัตโนมัติมีแนวโน้มของคุณภาพของการตรวจให้คะแนนเท่ากับการตรวจให้คะแนนจากคน

## ข้อเสนอแนะในการนำผลวิจัยไปใช้

1. ครูผู้สอนสามารถนำแบบสอบเขียนตอบลักษณะนี้ไปประยุกต์ใช้ในการวัดผลและประเมินผลการเรียนรู้ในรายวิชาภาษาไทย และวิชาอื่น ๆ ที่มีการเขียนตอบในเชิงของการจับใจความ สรุปความ และย่อความ

2. ใช้เป็นเครื่องมือเสริมในการฝึกฝนทักษะของผู้เรียน โดยผู้เรียนสามารถฝึกฝนได้ด้วยตนเอง

3. ใช้สำหรับการประเมินเพื่อพัฒนาครูผู้สอน เพื่อใช้เปรียบเทียบและตรวจสอบความแม่นยำของเกณฑ์การให้คะแนนและเป็นแนวทางในการพัฒนาเทคนิคการประเมินผลที่มีมาตรฐานมากยิ่งขึ้น

4. การเลือกใช้ระบบตรวจให้คะแนนอัตโนมัติ เมื่อต้องการความรวดเร็วและความเที่ยงในการตรวจ และลดความเอนเอียงในการให้คะแนนของผู้ตรวจที่เป็นคน ทำให้ผลตรวจมีความน่าเชื่อถือสูงขึ้น

## ข้อเสนอแนะในการทำวิจัยครั้งต่อไป

1. การศึกษาผลกระทบระยะยาวของการใช้ระบบอัตโนมัติต่อทักษะการเขียนของผู้เรียน เช่น ลายมือ ความเป็นระเบียบของการเขียนบนกระดาษ การเขียนสะกดคำ การเลือกใช้คำ การเรียบเรียงคำในประโยค รูปแบบการใช้ประโยค เป็นต้น

2. การศึกษาความตรงเชิงโครงสร้าง (Construct validity) และใช้แนวทางวิเคราะห์เชิงปริมาณขั้นสูงในการวิเคราะห์คุณภาพของเครื่องมือ ได้แก่ ข้อสอบและเกณฑ์การตรวจให้คะแนน

3. การเปรียบเทียบระบบกับผู้ตรวจที่มีระดับประสบการณ์หลากหลาย ศึกษาความแม่นยำและสอดคล้องกับผู้ตรวจ เช่น เปรียบเทียบกับผู้ตรวจที่มีประสบการณ์สอนภาษาไทย ระดับมัธยมศึกษาตอนต้น 10 ปี, 15 ปี, 20 ปี และ 25 ปี หรือผู้ตรวจที่มีประสบการณ์ตรวจข้อสอบเขียนตอบในการสอบขนาดใหญ่ อาทิ NT O-NET เป็นต้น

4. การวิจัยเชิงคุณภาพเพิ่มเติม เพื่อศึกษาความคิดเห็นและประสบการณ์ของผู้ใช้จริง เช่น ความพึงพอใจในการแสดงผล ความเหมาะสมของการประกาศผล ข้อควรพัฒนาระบบการสอบจากมุมมองของผู้สอบหลังใช้งาน เป็นต้น

## References


Apaikawee, D., Tuksino, P., & Tangdhanakanond, K. (2020). A Study of the Results of Subjective Test Scoring by Applying the Many-Facet Rasch Model and Generalizability Theory. *Journal of Educational Measurement, Mahasarakham University, 26*(1), 110–124. [in Thai]

Broadfoot, P. & Rockey, J. (2025). Generative AI and the social functions of educational assessment. *Oxford Review of Education, 51*(2), 283–299.

Bureau of Educational Testing. (2022). *Manual for the use of standardized essay test instruments based on the Basic Education Core Curriculum B.E. 2551 (2008) (Revised B.E. 2560/2017) for primary education.* Bangkok: Aksornthai Press. [in Thai]

Chansima, N. & Tuksino, P. (2019). *Comparison of quality of scoring for essay test under different of scoring pattern and item characteristics and raters : application of generalizability theory* (Master's thesis). Faculty of Education, Khon Kaen University. [in Thai]

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed).* New Jersey: Lawrence Erlbaum Associates.

Cronbach, J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

Crossley, S.A., Bradfleld, F. & Bustamante, A. (2019). Using human judgments to examine the validity of automated grammar, syntax, and mechanical errors in writing. *Journal of Writing Research, 11*(2), 251- 270.

Department of Local Administration. (2023). *Notification of the allocation of the annual budget expenditure for the fiscal year 2023 on the Reading Test (RT) for Grade 1 students and the National Test (NT) for Grade 3 students in the academic year 2022*. Retrieved from https://www.dla.go.th/upload/document/type2/2023/1/28676_1_1673249130936.pdf?time=1673249864225 [in Thai]

Dikli, S. (2006). *An Overview of Automated Scoring of Essays*. *The Journal of Technology, Learning and Assessment, 5*(1), *1-36.*

Janprasert, B., Lawthong, N., & Ngadgratoke, S. (2020). Inter-rater Reliability of Alignment between Science Items and Indices. *Journal of Education Studies, 48*(3), 144–163. [in Thai]

Kanchanawasri, S. (2013). *Classical test theory* (7th ed.). Bangkok: Chulalongkorn University Press.

Koo, T.K. & Li, M.Y. (2016). A Guideline of Selecting and Reporting Intra-Class Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155 – 163.

Office of the Basic Education Commission. (2023). *Number of students and classrooms by gender and grade level, academic year 2023*. Bangkok: Office of the Basic Education Commission. [in Thai]

Pattani Provincial Education Office. (2023). *The Office of the Basic Education Commission (OBEC) revised the Basic Education Core Curriculum B.E. 2551 (2008)*, updated version B.E. 2566 (2023). Retrieved from https://www.ptnpeo.go.th/ednews/7191/ [in Thai]

Rovinelli, R.J. & Hambleton, R.K. (1976). *The use of content specialists in the assessment of criterion-referenced test item validity*. *Tijdschrift Voor Onderwijs Research, 2*, 49-60.

Suchato, A., Pratanwanich, N., Chomphooyod, P., & Wiriyachaiphon, P. (2023). *Complete research report on the study of applying artificial intelligence to develop reading skills for elementary school students.* Retrieved from https://www.onec.go.th/th.php/book/BookView/2008 [in Thai]

Sukwichai, S., Junpeng, P., Tawarungruang, C., & Intharah, T. (2023). Designing Automated Scoring System of Open-Ended Test by Providing Automatic Feedback to Diagnose Mathematical Proficiency Levels through Machine Learning. *Journal of Educational Measurement, Mahasarakham University, 29*(1), 210–230. [in Thai]

Thitikanpodchana, W., & Tuksino, P. (2021*). Comparisons of the generalizability coefficient scores of English writing ability test of mattayom 3 with different rater's linguistics background and scoring designs.* Retrieved from https://app.gs.kku.ac.th/gs/th/publicationfile/item/22nd-ngrc-2021/HMO18/HMO18.pdf [in Thai]

Thongsilp, A., Tangdhanakanond, K., & Chaimongkol, N. (2020). *Development of Automated Scoring System for Thai Writing Ability Test of Primary Education Level* (Doctoral dissertation). Faculty of Education, Chulalongkorn University. [in Thai]

Wiboonsri, Y. (2013). Measurement and Achievement Test Contstruction (11st ed.). Bangkok: Chulalongkorn University Press. [in Thai]

Zhang, M. & Williamson, D.M. (2023). Reliability improvement in writing assessment: The complementary role of AI-enhanced scoring systems. *Educational Measurement: Issues and Practice, 42*(1), 12-24.