

การประยุกต์ใช้ทฤษฎีการตอบสนองแบบพหุวิภาค ในการพัฒนาคคลังข้อสอบการรู้เรื่องการอ่านภาษาอังกฤษ Applying Polytomous Item Response Theory to the Development of an English Reading Literacy Item Bank

พิสิฐ พินิจสกุล^{1*} เทิดศักดิ์ สุพันธ์² เบญจมาภรณ์ เสนารัตน์³ และ สมประสงค์ เสนารัตน์³
Pisit Pinitakul^{1*}, Terdsak Suphandee², Benjamaporn Senarat³, and Somprasong Senarat³
(Received: April 24, 2025; Revised: June 3, 2025; Accepted: June 9, 2025)

บทคัดย่อ

การวิจัยครั้งนี้มีความมุ่งหมายเพื่อพัฒนาคคลังข้อสอบการรู้เรื่องการอ่านภาษาอังกฤษ โดยประยุกต์ใช้ทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค (Polytomous Item Response Theory: IRT) ด้วยโมเดล Graded Response Model (GRM) กลุ่มตัวอย่างคือนักเรียนมัธยมศึกษาตอนปลายในภาคตะวันออกเฉียงเหนือจำนวน 4,000 คน ได้มาโดยการสุ่มแบบหลายขั้นตอน เครื่องมือที่ใช้เป็นแบบทดสอบเชิงสถานการณ์การอ่านภาษาอังกฤษจำนวน 438 ข้อ ผลการวิเคราะห์ความตรงเชิงเนื้อหาโดยผู้เชี่ยวชาญ พบว่า ข้อสอบทุกข้อผ่านเกณฑ์ โดยมีค่าดัชนีความสอดคล้องอยู่ระหว่าง 4.00-5.00 จากมาตราส่วน 5 ระดับ แบ่งข้อสอบออกเป็น 18 ฉบับ ฉบับละ 24 ข้อ มีข้อสอบรวม จำนวน 6 ข้อมีการให้คะแนน 8 ระดับ นำไปทดสอบกับกลุ่มตัวอย่าง และนำมาวิเคราะห์ข้อมูลด้วยโปรแกรม R แพ็กเกจ mirt ได้ข้อสอบที่มีคุณภาพตามเกณฑ์จำนวน 339 ข้อ โดยมีผลสรุปค่าพารามิเตอร์ของข้อสอบโดยรวมพบว่าอำนาจจำแนก (α) ทั้ง 3 ระดับมีค่าเฉลี่ยอยู่ที่ 1.295 ส่วนเบี่ยงเบนมาตรฐานอยู่ที่ 0.466 ในส่วนของความยากทุกระดับ (β) มีค่าเฉลี่ยอยู่ที่ -0.643 ส่วนเบี่ยงเบนมาตรฐานอยู่ที่ 1.181 ลักษณะของข้อสอบในคลังมีความครอบคลุมตั้งแต่ระดับความสามารถต่ำถึงสูง โดยข้อสอบสามารถประเมินผู้เรียนในช่วงความสามารถตั้งแต่ระดับ A1 ถึง B1 ตาม CEFR และมีข้อสอบบางส่วนที่สามารถวัดได้ถึงระดับความสามารถสูงกว่าค่าเฉลี่ย ($\theta > 2$) ซึ่งสะท้อนศักยภาพของคลังที่มีประสิทธิภาพ คลังข้อสอบถูกจัดเก็บในฐานข้อมูลออนไลน์สร้างขึ้นโดยใช้โปรแกรม PhpMyAdmin และ

¹ นักศึกษาปริญญาเอก สาขาวิชาวิจัยและประเมินผลการศึกษา คณะครุศาสตร์และการพัฒนามนุษย์
มหาวิทยาลัยราชภัฏร้อยเอ็ด

² ผู้ช่วยศาสตราจารย์ คณะครุศาสตร์และการพัฒนามนุษย์ มหาวิทยาลัยราชภัฏชัยภูมิ

³ ผู้ช่วยศาสตราจารย์ คณะครุศาสตร์และการพัฒนามนุษย์ มหาวิทยาลัยราชภัฏร้อยเอ็ด

¹ Doctoral degree student, Educational Research and Evaluation Program, Faculty of Education and Human Development, Roi Et Rajabhat University

² Assistant Professor, Faculty of Education and Human Development, Chaiyaphum Rajabhat University

³ Assistant Professor, Faculty of Education and Human Development, Roi Et Rajabhat University

*Corresponding author: E-mail: pisit_pinitakul@hotmail.com

ฐานข้อมูลในคลังข้อสอบนี้จะนำไปพัฒนาโปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (Computerized Adaptive Testing: CAT) ต่อไป

คำสำคัญ การรู้เรื่องการอ่านภาษาอังกฤษ ทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค โมเดล GRM
คลังข้อสอบ

Abstract

This study aimed to develop an English reading literacy item bank by applying Polytomous Item Response Theory (IRT) using the Graded Response Model (GRM). The sample consisted of 4,000 upper secondary school students in northeastern Thailand, selected through multi-stage random sampling. The instrument used was a situational English reading test comprising 438 items. All items demonstrated acceptable content validity based on expert ratings (4.00–5.00 on a five-point scale). The items were divided into 18 test forms, each containing 24 items, with 6 anchor items shared across forms. The scoring was based on an 8-level rubric. The tests were administered to the sample, and data were analyzed using the mirt package in R. A total of 339 items met the quality criteria. The overall item parameter estimates showed that the average discrimination (α) across the three levels was 1.295 with a standard deviation of 0.466. The average difficulty (β) across all thresholds was -0.643 with a standard deviation of 1.181. The item bank covers a wide range of ability levels, from low to high. The test items can assess learners from CEFR levels A1 to B1, with some items capable of measuring abilities higher than average ($\theta > 2$), reflecting the effectiveness of the item bank. The item bank was stored in an online database developed using PhpMyAdmin. This item bank will be used for further development into a Computerized Adaptive Testing (CAT) system.

Keywords: English reading literacy, polytomous item response theory, graded response model, item bank

บทนำ

การรู้เรื่องการอ่านภาษาอังกฤษ (English reading literacy) เป็นทักษะสำคัญสำหรับผู้เรียนในโลกยุคดิจิทัลที่เต็มไปด้วยการสื่อสารข้ามภาษาและข้ามวัฒนธรรม ภาษาอังกฤษในฐานะภาษากลางช่วยให้ผู้เรียนเข้าถึงองค์ความรู้ ข่าวสาร และข้อมูลจากแหล่งต่างประเทศได้อย่างมีประสิทธิภาพ อีกทั้งยังเป็นรากฐานสำคัญของการพัฒนาทักษะภาษาอื่น ๆ เช่น การเขียนและการพูด (Brown, 2004) Richards (2015) เน้นว่าทักษะการอ่านไม่ได้จำกัดเพียงการถอดรหัสข้อความเท่านั้น แต่ยังรวมถึงความสามารถในการตีความ วิเคราะห์ และเชื่อมโยงข้อมูลในบริบทจริงด้วย ในขณะที่ Brown (2004) ชี้ให้เห็นว่า การอ่านเป็นทักษะรับที่ไม่สามารถ

สังเกตได้โดยตรง การประเมินจึงต้องอาศัยพฤติกรรมตอบสนอง เช่น การเลือกคำตอบ สรุปลักษณะ หรือเขียนอธิบาย เพื่ออนุมานระดับความเข้าใจของผู้เรียนอย่างมีนัยสำคัญ ดังนั้น การออกแบบเครื่องมือวัดผลที่มีคุณภาพจึงมีบทบาทสำคัญในการสะท้อนศักยภาพที่แท้จริงของผู้เรียน

แม้ว่าการรู้เรื่องการอ่าน (reading literacy) จะเป็นทักษะพื้นฐานในด้านภาษา แต่ในทางทฤษฎีการวัดผลทางภาษา ทั้งนี้การอ่านหรือ reading comprehension ความเข้าใจในการอ่านเกี่ยวข้องกับกระบวนการทางความคิดหลายขั้น โดยพฤติกรรมการอ่านไม่สามารถสังเกตได้โดยตรง ผู้ประเมินต้องอาศัยการ “อนุมาน” ผ่านพฤติกรรมของผู้สอบ เช่น การตอบคำถาม หรือการสรุปลักษณะ ซึ่งอาจคลาดเคลื่อนจากความเข้าใจจริงได้ (Alderson, 2010) ยิ่งไปกว่านั้น บริบทของการอ่าน (reading context) มีผลอย่างมากต่อการประเมิน เช่น บทความยาว บทสนทนา หรือข้อความเชิงวิชาการ ล้วนใช้ภาษาและกลยุทธ์การอ่านที่แตกต่างกัน ซึ่งส่งผลต่อ “validity” ของการวัด หากแบบทดสอบไม่ครอบคลุมประเภทข้อความและประเภทคำถามอย่างเหมาะสม การวัดอาจไม่สะท้อนทักษะการอ่านที่แท้จริงของผู้เรียนอย่างรอบด้าน (Bachman & Palmer, 2010) ด้วยข้อจำกัดในเชิงทฤษฎีเหล่านี้ นักวิจัยด้านการวัดผลภาษาเสนอว่า ควรใช้เทคนิคหรือโมเดลที่สามารถประเมินระดับความเข้าใจในหลายมิติได้อย่างแม่นยำและลึกซึ้งมากขึ้น (Dai et al., 2021)

โดยทั่วไปการวัดทักษะการอ่านมักใช้ข้อสอบแบบเลือกตอบที่มีการให้คะแนนแบบจำแนกสองค่า (dichotomous scoring) คือให้คะแนน 1 เมื่อคำตอบถูกต้อง และ 0 เมื่อผิด แม้จะใช้งานสะดวก แต่ข้อสอบลักษณะนี้มีข้อจำกัดสำคัญคือไม่สามารถสะท้อนระดับความเข้าใจที่หลากหลายของผู้เรียนได้อย่างครบถ้วน อีกทั้งยังต้องใช้จำนวนข้อสอบจำนวนมากเพื่อให้ครอบคลุมทุกระดับของผู้สอบ ส่งผลให้ใช้เวลาในการสร้างและการทำแบบทดสอบมาก และอาจเกิดความคลาดเคลื่อนในการวัดที่มาจากความเหนื่อยล้าหรือการที่มีโจทย์คำถามให้อ่านมาก และอาจทำให้เกิดการเดาในการตอบข้อคำถามได้ (Hambleton & Swaminathan, 1985; Wheadon, 2011) ด้วยเหตุนี้ โมเดลตอบสนองข้อสอบแบบพหุวิภาค (polytomous IRT) หรือโมเดลแบบตรวจให้คะแนนมากกว่า 2 ค่า เป็นโมเดลความสัมพันธ์ที่ไม่ใช่เชิงเส้นตรง (non-linear relationship model) ที่อธิบายความสัมพันธ์ระหว่างความสามารถของผู้สอบกับโอกาสของการเลือกตอบแต่ละรายการคำตอบที่มีให้ (Kanjawasee, 2020) และมีโมเดลตอบสนองข้อสอบแบบพหุวิภาคที่เป็นที่นิยมหลายวิธีการ เช่น Graded Response Model (GRM) ซึ่งสามารถให้คะแนนข้อสอบได้หลายระดับจากการตอบเพียงหนึ่งข้อ ช่วยลดจำนวนข้อสอบ และระยะเวลาในการทดสอบลงได้อย่างมาก การที่ประยุกต์ทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาคจะสามารถแก้ปัญหาในประเด็นนี้ได้และได้สารสนเทศที่หลากหลายและมีประสิทธิภาพสามารถให้ค่าประมาณความสามารถของผู้เรียนที่มีความละเอียดและมีประสิทธิภาพสูงขึ้น (Ostini & Nering, 2006) ในการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบในข้อสอบทางภาษานั้นยังมีการประยุกต์ทฤษฎีตอบสนองข้อสอบเข้ามายังไม่มาก ซึ่งเห็นได้ในงานวิจัยของต่างประเทศที่ทำวิจัยเรื่องเกี่ยวกับการทดสอบทางภาษาที่เป็นข้อสอบมาตรฐาน (Wainer et al., 2007; Wheadon, 2011)

จุดแข็งสำคัญของการประยุกต์ใช้โมเดลพหุวิภาค เช่น Graded Response Model (GRM) คือความสามารถในการประเมินพฤติกรรมการตอบของผู้สอบในหลายระดับคะแนนจากการตอบเพียงข้อเดียว (Chattergoon, 2020; Dai et al., 2021; Yen, 1996) จึงเหมาะสมกับข้อสอบที่ใช้การให้คะแนนแบบ rubric

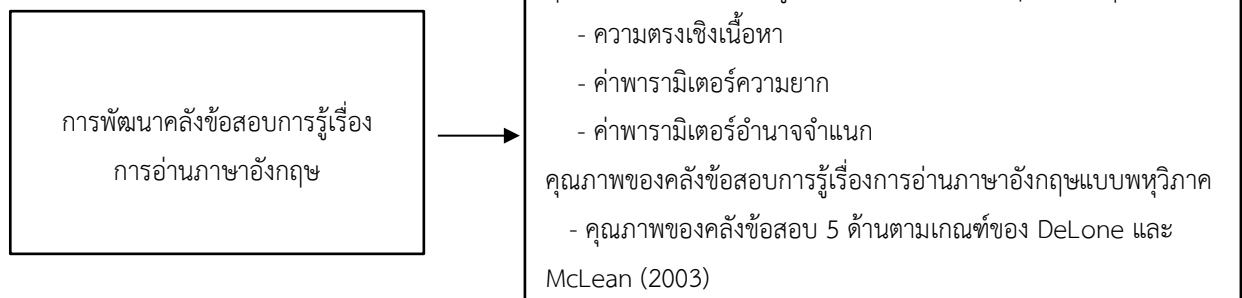
หรือกลุ่มของข้อย่อยหลายข้อ ที่ใช้ stimulus ร่วมกัน ข้อสอบแบบ testlet หรือข้อสอบเชิงสถานการณ์ที่วัดหลายทักษะในหนึ่งข้อ เช่น แบบทักษะการรู้เรื่องการอ่านภาษาอังกฤษ (Yen, 1996) ในกรณีของข้อสอบแบบ testlet ที่ใช้คำถามย่อยหลายข้อกับสิ่งเร้าเดียวกัน (stimulus) การให้คะแนนแบบรวม (rubric) ที่เป็นข้อมูลเชิงลำดับเหมาะสำหรับการวิเคราะห์ด้วย GRM ซึ่งสามารถจัดการกับข้อจำกัดด้าน local independence ได้อย่างมีประสิทธิภาพ (Pinitsakul et al., 2020) นอกจากนี้ ข้อสอบที่ผ่านการวิเคราะห์ด้วย GRM จะมีพารามิเตอร์ความยาก threshold (β) จำนวนหลายระดับภายในข้อเดียวทั้งนี้ขึ้นอยู่กับจำนวนตัวเลือก ความหลากหลายของค่า β ช่วยให้สามารถวัดระดับความสามารถของผู้เรียนได้ครอบคลุมและต่อเนื่อง ทั้งในกลุ่มที่มีความสามารถต่ำ ปานกลาง และสูง ในขณะเดียวกันการกำหนดพารามิเตอร์อำนาจจำแนก (α) ให้เหมาะสมในแต่ละข้อจะช่วยให้ข้อสอบสามารถแยกแยะผู้เข้าสอบที่มีระดับความสามารถต่างกันได้อย่างแม่นยำ ซึ่งเป็นปัจจัยสำคัญต่อประสิทธิภาพของโมเดล GRM โดยเฉพาะอย่างยิ่งเมื่อถูกนำไปประยุกต์ใช้ในระบบการสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (Computerized Adaptive Testing: CAT) ที่ต้องอาศัยข้อมูลพารามิเตอร์ในการคัดเลือกข้อสอบให้สอดคล้องกับระดับความสามารถของผู้สอบในแต่ละช่วงเวลา โดยระบบสามารถคำนวณหาข้อสอบที่ให้ค่าสารสนเทศสูงสุด (maximum information) ตรงกับระดับความสามารถของผู้สอบในแต่ละขั้นตอนของการทดสอบ ส่งผลให้การวัดมีความแม่นยำมากขึ้น แม้จะใช้จำนวนข้อสอบน้อยลงก็ตาม (Lord, 1980; Weiss, 1982; van der Linden & Glas, 2000) การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (Computerized Adaptive Testing: CAT) จะสามารถดำเนินการได้อย่างมีประสิทธิภาพ ก็ต่อเมื่อมีคลังข้อสอบที่ครอบคลุมช่วงความสามารถของผู้สอบ และมีพารามิเตอร์ที่แม่นยำเพียงพอในการประมาณค่าความสามารถอย่างต่อเนื่อง (Chattergoon, 2020; Embretson & Reise, 2000; van der Linden & Glas, 2000) ที่ผ่านมามีงานวิจัย เช่น Wainer et al. (2007) และ Ravand (2015) ที่เสนอแนวทางการวิเคราะห์ข้อสอบแบบ testlet ด้วยโมเดลเฉพาะเพื่อรองรับโครงสร้างข้อมูลที่มีความพัวพันภายในกลุ่มข้อสอบ อย่างไรก็ตาม การประยุกต์ใช้โมเดลพหุวิภาค เช่น GRM กับข้อสอบที่มีการให้คะแนนหลายระดับยังคงได้รับการศึกษาน้อย โดยเฉพาะในบริบทของข้อสอบเชิงสถานการณ์ด้านการอ่านภาษาอังกฤษ ซึ่งจำเป็นต้องมีการพัฒนาอย่างเป็นระบบเพื่อรองรับการใช้งานในระบบการสอบแบบปรับเหมาะ (CAT)

ด้วยเหตุนี้ งานวิจัยนี้จึงมุ่งพัฒนากล้องข้อสอบการรู้เรื่องการอ่านภาษาอังกฤษที่ออกแบบขึ้นโดยเฉพาะสำหรับการนำไปใช้ในระบบ CAT ในอนาคต ข้อสอบแต่ละชุดถูกออกแบบให้อยู่ในรูปแบบข้อสอบเชิงสถานการณ์แบบ testlet ซึ่งสามารถวัดหลายทักษะในข้อเดียว ใช้การให้คะแนนแบบพหุระดับตามเกณฑ์ rubric และผ่านการวิเคราะห์ค่าพารามิเตอร์ด้วย GRM อย่างเป็นระบบ คลังข้อสอบลักษณะนี้มีศักยภาพในการรองรับการประเมินเฉพาะบุคคล ช่วยลดเวลาในการสอบ เพิ่มประสิทธิภาพในการวัดผลที่เที่ยงตรงต่อระดับความสามารถของผู้เรียนแต่ละรายได้อย่างแท้จริง (Dai et al., 2021; Senarat et al., 2013; Thissen et al., 2000)

ความมุ่งหมายของการวิจัย

เพื่อพัฒนากล้องข้อสอบการรู้เรื่องการอ่านภาษาอังกฤษโดยประยุกต์ใช้ทฤษฎีตอบสนองข้อสอบแบบพหุวิภาค

กรอบแนวคิดในการวิจัย



ภาพ 1 กรอบแนวคิดในการวิจัย การพัฒนาคำข้อสอบการรู้เรื่อง การอ่านภาษาอังกฤษ

วิธีดำเนินการวิจัย

1. ประชากรและตัวอย่าง

ประชากรคือ นักเรียนระดับชั้นมัธยมศึกษาตอนปลายที่กำลังศึกษาอยู่ในปีการศึกษา 2567 ในภาคตะวันออกเฉียงเหนือ จำนวน 20 จังหวัด

ตัวอย่าง คือ นักเรียนระดับชั้นมัธยมศึกษาตอนปลายที่กำลังศึกษาอยู่ในปีการศึกษา 2567 จำนวน 4,000 คน ได้มาโดยการสุ่มแบบหลายขั้นตอน (multi-stage random sampling) ตามขั้นตอนดังนี้

- สุ่มจังหวัดในภาคตะวันออกเฉียงเหนือ จำนวน 20 จังหวัด มาร้อยละ 70 (ใช้กลุ่มตัวอย่างในแต่ละจังหวัดประมาณร้อยละ 10 ของประชากร) ด้วยวิธีการสุ่มอย่างง่ายโดยใช้จังหวัดเป็นหน่วยการสุ่ม ได้กลุ่มตัวอย่าง จำนวน 14 จังหวัด จากนั้น แบ่งกลุ่มโรงเรียนจาก 14 จังหวัด ออกเป็น 3 กลุ่มตามขนาดโรงเรียน คือ ขนาดใหญ่ ขนาดกลาง และขนาดเล็ก และหลังจากนั้นสุ่มโรงเรียนในแต่ละกลุ่มด้วยวิธีการสุ่มอย่างง่าย โดยใช้โรงเรียนเป็นหน่วยการสุ่ม และสุ่มได้โรงเรียน รวม 35 แห่ง มีนักเรียนเข้าร่วมทดสอบรวม ประมาณ 4,000 คน โดยเฉลี่ยโรงเรียนละ 100–300 คน แล้วแต่ขนาดโรงเรียน

2. เครื่องมือที่ใช้ในการวิจัย ขั้นตอนการสร้างแบบทดสอบการรู้เรื่อง การอ่านภาษาอังกฤษมีดังนี้

2.1 ศึกษาเอกสาร ทฤษฎี และงานวิจัยที่เกี่ยวข้องกับการสร้างแบบทดสอบ วิเคราะห์ สังเคราะห์ แนวคิด ทฤษฎี งานวิจัยทั้งในประเทศและต่างประเทศ รวมทั้งฐานข้อมูลต่าง ๆ ในอินเทอร์เน็ตที่เกี่ยวข้องกับเนื้อหาการรู้เรื่อง การอ่านภาษาอังกฤษที่สอดคล้องกับระดับความสามารถทางภาษาตามกรอบ CEFR จากระดับ A1-B1 ข้อสอบเชิงสถานการณ์แบบเทสต์เลต โมเดลการตอบสนองข้อสอบแบบพหุวิภาค

2.2 สร้างแบบทดสอบ โดยการวิเคราะห์เนื้อหา เพื่อกำหนดนิยามเนื้อหาที่ต้องการวัดซึ่งผู้วิจัยได้กำหนดเนื้อหา ตามความสำคัญของเนื้อหาการรู้เรื่อง การอ่านภาษาอังกฤษ และเขียนข้อสอบตามนิยามเนื้อหาเป็นแบบเลือกตอบ (multiple choice Item) จำนวน 438 ข้อ ข้อละ 3 คำถาม 8 ตัวเลือก ซึ่งวัดทักษะการอ่านหาใจความสำคัญ การอ่านหาข้อมูลเฉพาะเจาะจง และการอ่านหารายละเอียด โดยมีลักษณะเป็นข้อสอบเชิงสถานการณ์ (testlet test) ในแต่ละสถานการณ์ประกอบด้วย บทความมีความยาวไม่เกิน 200 คำ มีเนื้อหาทั่วไปอิงระดับภาษาตาม CEFR อยู่ระหว่าง A1 A2 และ B1 โดยสรุปใน 1 ข้อมี 1 บทความ 3 คำถามและ มี 8 ตัวเลือกที่ใช้ตอบทั้ง 3 คำถามในตัวเลือกเดียวกัน

2.3 สร้างเกณฑ์การให้คะแนน จากการให้คะแนนเต็ม 7 คะแนนแบ่งเป็น 8 ระดับคือ 7 6 5 4 3 2 1 และ 0 แบบพหุวิภาค (polytomous) ในรูปแบบ rubric score ในการให้คะแนนของตัวเลือก โดยมีตัวอย่างดังตาราง 1

ตาราง 1 เกณฑ์การให้คะแนนการตอบแต่ละตัวเลือก

1. การอ่านหาใจความสำคัญ (4 คะแนน)	2. การอ่านหาข้อมูลเฉพาะเจาะจง (2 คะแนน)	3. การอ่านหารายละเอียด (1 คะแนน)	คะแนนที่ได้
✓	✓	✓	7
✓	✓	✗	6
✓	✗	✓	5
✓	✗	✗	4
✗	✓	✓	3
✗	✓	✗	2
✗	✗	✓	1
✗	✗	✗	0

จากตาราง 1 มีการให้คะแนนดังนี้ โดยมีคะแนนเต็ม 7 คะแนน ซึ่งถ้าตอบถูกในคำถามวัดการอ่านหาใจความสำคัญจะได้ 4 คะแนน ถ้าตอบถูกในคำถามวัดการอ่านหาข้อมูลเฉพาะเจาะจงจะได้ 2 คะแนน และถ้าตอบถูกในคำถามวัดการอ่านหารายละเอียดจะได้ 1 คะแนน จากตารางซึ่งเห็นได้ว่าถ้าตอบทุกทั้ง 3 ประเด็น การวัดในข้อคำถามนั้นจะได้คะแนนรวมเต็ม 7 คะแนน และในแต่ละด้านการวัดถ้าตอบผิดด้านใดด้านหนึ่งจะได้คะแนนแตกต่างกันออกไปตามน้ำหนักคะแนนของแต่ละด้าน และรวมเป็นคะแนนรวมในแต่ละสถานการณ์บทความ และถ้าผู้สอบตอบผิดทั้งสามด้านจะได้ 0 คะแนน ทั้งนี้ในแต่ละข้อนั้นจะคิดเป็นคะแนนรวมจากทั้ง 3 ด้านที่ผู้สอบเลือกตอบจากตัวเลือกทั้ง 8 ตัวเลือก โดยข้อสอบ 1 ข้อมีลักษณะดังตัวอย่างในภาพ 2

2.4 การหาคุณภาพของข้อสอบ พิจารณาความตรงเชิงเนื้อหาทักษะการอ่านหาใจความสำคัญ การอ่านหาข้อมูลเฉพาะเจาะจง และการอ่านหารายละเอียด โดยผู้เชี่ยวชาญ 15 ท่าน ด้วยวิธีการตรวจสอบความตรงเชิงเนื้อหาแบบ CVI (Content Validity Index) และผู้เชี่ยวชาญ 4 ท่านตรวจสอบความเหมาะสมของเกณฑ์การให้คะแนน และคัดเลือกข้อสอบที่มีค่า CVI ตั้งแต่ 4.00 ขึ้นไปจากการประเมิน rating scales 1-5 (Waltz et al., 2010) ซึ่งมีข้อผ่านเกณฑ์ร้อยละ 98.63 ส่วนข้อที่เหลือนำไปแก้ไขตามคำแนะนำความเที่ยง พบว่ามีค่า Cronbach Alpha อยู่ที่ 0.84 และนำไปพิมพ์เป็นแบบทดสอบฉบับสมบูรณ์

บทความ

The Netherlands, is a small country in Western Europe, also commonly known as Holland. Netherlands literally means "the low land". In the past, almost the whole country was covered by sea until the Dutch, citizens of the Netherlands, built dikes to keep out seawater. Finally, they got their man-made lands. The country is bordered by Germany and Belgium. Dutch people, therefore, can speak many languages such as Dutch, German, French and English. Its capital city is Amsterdam.

คำถาม 1 ข้อที่ประกอบด้วย 3 ประเด็นการวัด ต่อ 1 บทความ

What is the main idea of this passage? (การอ่านหาใจความสำคัญ) **Why is Netherlands called "the low land"?** (การอ่านหาข้อมูลเฉพาะเจาะจง) **Who are the citizens of the Netherlands?** (การอ่านหารายละเอียด)

- a. The Netherlands is a small country in Western Europe/ The whole country was covered by sea/ Dutch
- b. The Netherlands is a small country in Western Europe/ The whole country was covered by sea/ German
- c. The Netherlands is a small country in Western Europe/ It is a small country/ Dutch
- d. The Netherlands is a small country in Western Europe/ It is a small country/ German
- e. Amsterdam is capital city of the Netherlands/ The whole country was covered by sea/ Dutch
- f. Amsterdam is capital city of the Netherlands/ The whole country was covered by sea/ German
- g. They can speak many languages/ It is a small country/ Dutch
- h. They can speak many languages/ It is a small country/ Belgian

ภาพ 2 ตัวอย่างข้อสอบ 1 ข้อ 3 คำถาม 8 ตัวเลือก

3. การเก็บรวบรวมข้อมูล

3.1. นำแบบทดสอบไปทดสอบกับตัวอย่าง จำนวน 4,000 คน ดำเนินการจัดการทดสอบออนไลน์ด้วยระบบทดสอบที่สร้างขึ้นมาเอง โดยแบ่งแบบทดสอบออกเป็น 18 ชุดและให้กลุ่มตัวอย่างทำข้อสอบกลุ่มละ 6 ชุด และกำหนดเวลาในการทำข้อสอบจำนวน 1 ชั่วโมง เมื่อสิ้นสุดการทดสอบหรือครบกำหนดเวลา 1 ชั่วโมงแล้วระบบจะปิดอัตโนมัติ จากนั้นนำผลการทดสอบของนักศึกษามาตรวจสอบความสมบูรณ์ของข้อมูล เพื่อนำไปใช้ในการวิเคราะห์ข้อมูลต่อไป

3.2 การจัดทำคลังข้อสอบออนไลน์ พัฒนาระบบคลังข้อสอบออนไลน์โดยใช้โปรแกรม MySQL ร่วมกับ PhpMyAdmin เพื่อออกแบบฐานข้อมูลสำหรับจัดเก็บข้อสอบและพารามิเตอร์ของข้อสอบอย่างเป็นระบบ ฐานข้อมูลถูกออกแบบให้สามารถรองรับโครงสร้างของข้อสอบแต่ละชุด ลักษณะการใช้งาน รวมถึงข้อมูลพารามิเตอร์ต่าง ๆ ที่ได้จากการวิเคราะห์ เช่น พารามิเตอร์ความยาก (difficulty) และพารามิเตอร์อำนาจจำแนก (discrimination) ระบบรองรับการเพิ่มและแก้ไขข้อสอบผ่านส่วนจัดการ (backend) สำหรับผู้ดูแลระบบ โดยในส่วนของผู้เข้าสอบ ระบบจะแสดงเฉพาะคำถามและตัวเลือกคำตอบโดยไม่มีการแสดงเฉลย

เพื่อป้องกันการเปิดเผยข้อมูลในคลัง ขณะที่ผู้ดูแลระบบสามารถเข้าถึงข้อมูลผลการตอบข้อสอบ และค่าพารามิเตอร์ที่เกี่ยวข้องได้จากฐานข้อมูลเพื่อการวิเคราะห์ในภายหลัง

3.3 ตรวจสอบข้อตกลงเบื้องต้น พบว่าค่าเฉลี่ยของสถิติ Yen's Q3 ในทุกชุดแบบทดสอบ (ชุดที่ 1-18) มีค่าอยู่ระหว่าง -0.031 ถึง -0.025 ซึ่งอยู่ในระดับต่ำกว่า 0.20 ตามเกณฑ์ของ Yen (1993) แสดงว่าไม่เป็นการละเมิดข้อตกลงเบื้องต้นของ Local Item Dependence (LID) ในแต่ละชุด ถือว่าละเอียดพิถีพิถันของ testlet และสามารถใช้โมเดล GRM ในการวิเคราะห์ได้ และการตรวจสอบ normality test พบว่าค่าสถิติ Kolmogorov-Smirnov ไม่มีนัยสำคัญทางสถิติ ($p > .05$) แสดงว่าข้อมูลมีการแจกแจงปกติ

3.4 คัดเลือกข้อสอบเข้าคลังโดยพิจารณาจากค่าพารามิเตอร์จากการวิเคราะห์ข้อมูลด้วยโมเดล GRM โดยพิจารณาจากค่าพารามิเตอร์ของข้อสอบกับเกณฑ์ที่กำหนดไว้ คือ ค่าพารามิเตอร์อำนาจจำแนก มีค่าตั้งแต่ 0.70 ขึ้นไปและไม่เกิน 2.50 ค่าพารามิเตอร์ความยากไม่เกิน 3.000 สอดคล้องงานวิจัยของ Loken & Rulison (2010)

3.5 การนำข้อสอบเข้าคลังและประเมินคุณภาพคลัง เมื่อคัดเลือกข้อสอบที่ผ่านเกณฑ์ตามข้อ 3.2 แล้ว ระบบจะดำเนินการบันทึกข้อมูลข้อสอบและพารามิเตอร์ที่เกี่ยวข้องเข้าสู่ฐานข้อมูลคลังข้อสอบออนไลน์โดยอัตโนมัติ ข้อมูลที่จัดเก็บจะรวมถึงข้อความคำถาม ตัวเลือกคำตอบ รหัสข้อสอบ พารามิเตอร์ของแต่ละตัวเลือก (เช่น ค่า threshold ในโมเดล GRM) และข้อมูลการอ้างอิง และในการประเมินคุณภาพของคลังข้อสอบ ได้ดำเนินการประเมินด้วยตนเอง (self-evaluation) โดยผู้วิจัย โดยประยุกต์อิงจากกรอบแนวคิดของ DeLone & McLean (2003) ซึ่งประกอบด้วยองค์ประกอบหลัก 5 ด้าน ได้แก่ 1) คุณภาพของระบบ (system quality) 2) คุณภาพของข้อมูล (information quality) 3) คุณภาพของการบริการ (service quality) 4) การใช้งาน (use) 5) ผลกระทบต่อประสิทธิภาพ (net benefit) การประเมินแต่ละด้านดำเนินการโดยใช้รายการตรวจสอบ (checklist) ที่จัดทำขึ้นจากเกณฑ์ของแต่ละองค์ประกอบ

4. การวิเคราะห์ข้อมูล

วิเคราะห์สถิติเชิงบรรยายและวิเคราะห์ค่าพารามิเตอร์ของข้อสอบ ด้วยทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค (polytomous item response theory) ด้วยโปรแกรม R แพ็กเกจ MIRT และแพ็กเกจ grm_model (Chalmers, 2012) วิเคราะห์ค่าพารามิเตอร์ของข้อสอบ โดยพิจารณาค่าอำนาจจำแนก (a) และ ค่าความยาก (b) ซึ่งโมเดล graded response มีลักษณะดังนี้

สมการโมเดล graded response

$$P(Y \geq k|\theta) = \frac{1}{1 + \exp[-a(\theta - b_k)]}$$

โดยที่

$P(Y \geq k|\theta)$ คือ ความน่าจะเป็นที่ผู้ตอบจะได้คะแนน k หรือสูงกว่า

a คือ พารามิเตอร์อำนาจจำแนก

θ คือ ความสามารถของผู้ตอบ (Ability)

b_k คือ พารามิเตอร์ความยากสำหรับคะแนน k

$\exp$ คือ ฟังก์ชันการยกกำลังฐาน e

ผลการวิจัย

การนำเสนอผลการวิจัยจะนำเสนอแบ่งเป็น 2 ตอนตามลำดับดังต่อไปนี้

1. ผลการวิเคราะห์ข้อสอบ

1.1 ผลการวิเคราะห์ความสอดคล้องระหว่างข้อสอบกับนิยามเนื้อหาการอ่านภาษาอังกฤษ ผลการพิจารณา พบว่าค่าความสอดคล้องระหว่างข้อสอบกับนิยามเนื้อหา ผ่านเกณฑ์จำนวน 438 ข้อจาก 438 ข้อ มีค่า CVI สูงกว่า 4.00 และดำเนินการปรับแก้ไขตามข้อเสนอแนะของผู้เชี่ยวชาญทุกท่าน

1.2 ผลการประมาณค่าพารามิเตอร์ข้อสอบ โดยทำการวิเคราะห์ข้อมูลด้วยโมเดลการตอบสนองข้อสอบแบบพหุวิภาค โมเดล Graded response model ด้วยแพ็คเกจ mirt ในโปรแกรม R ประกอบด้วย อำนาจจำแนก (α) และความยาก (β_1 ถึง β_7) จากข้อสอบทั้งหมดจำนวน 438 ข้อที่ยังได้คัดเลือกเข้าคลังข้อสอบ อิงตามทุกระดับ CEFR A1, A2 และ B1 โดยสรุปจะได้ดังตาราง 2

ตาราง 2 ผลการประมาณค่าพารามิเตอร์ข้อสอบก่อนคัดเลือกเข้าคลัง

ค่าพารามิเตอร์	Mean	S.D.	Min	Max
รวมอำนาจจำแนกทุกระดับ (α)	1.128	0.571	-0.205	3.691
รวมความยากทุกระดับ (β)	-0.776	3.885	-125.958	37.957

α แทนค่าอำนาจจำแนก β แทนค่าความยาก

* ข้อสอบที่ไม่เป็นไปตามเกณฑ์ และถูกคัดออก

1.3 ผลการตรวจสอบ item fit จากการวิเคราะห์คุณภาพข้อสอบรายชุดด้วยดัชนีความสอดคล้องระหว่างแบบจำลอง ($S-X^2$) ด้วยโปรแกรม R พบว่ามีข้อสอบบางรายการที่ไม่สอดคล้องกับแบบจำลอง GRM อย่างมีนัยสำคัญ ($p < .05$) จึงดำเนินการคัดออกตามหลักการของทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค ทั้งนี้จากข้อสอบทั้งหมด 438 ข้อ ซึ่งแบ่งออกเป็น 18 ชุด ๆ ละ 30 ข้อ โดยแต่ละชุดมีข้อสอบเฉพาะตัว 24 ข้อ และข้อสอบรวมอีก 6 ข้อ พบว่ามีข้อสอบที่ไม่ผ่านเกณฑ์ item fit รวมทั้งสิ้น 99 ข้อ คิดเป็น 22.60% ของข้อสอบทั้งหมด ซึ่งเมื่อคัดออกแล้ว เหลือข้อสอบที่ผ่านเกณฑ์ item fit จำนวน 339 ข้อ (77.40%) ที่มีความสอดคล้องกับโมเดล GRM อย่างมีนัยสำคัญ และสามารถนำไปวิเคราะห์ค่าพารามิเตอร์ขั้นต่อไป โดยใช้เกณฑ์การคัดเลือกคือ อำนาจจำแนก (α) อยู่ในช่วง 0.70–2.50 และค่าความยาก (b_1 – b_7) ต้องไม่เกิน 3.00 ในทุกระดับ ซึ่งเป็นเกณฑ์มาตรฐานตามข้อเสนอของ Loken & Rulison (2010)

1.4 ผลการตรวจสอบความสอดคล้องของโมเดลด้วยสถิติ M2 จากทั้ง 18 ชุดแบบทดสอบด้วยโปรแกรม R พบว่าแบบจำลอง GRM มีค่าดัชนีความสอดคล้องของโมเดลในระดับที่น่าพึงพอใจ โดยค่าดัชนี RMSEA อยู่ในช่วงระหว่าง 0.000 ถึง 0.080, ค่า AIC อยู่ในช่วง 8875.28 ถึง 23862.71, ค่า BIC อยู่ในช่วง 9422.7 ถึง 24656.70, และค่า Likelihood-ratio χ^2 (G^2) อยู่ในช่วง -3744.88 ถึง -11691.36 ซึ่งอยู่ในเกณฑ์ที่ยอมรับได้ตามเกณฑ์มาตรฐานในการวิเคราะห์ของทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค (polytomous IRT) (Embretson & Reise, 2000)

2. ผลการจัดทำคลังข้อสอบ

2.1 ผลการคัดเลือกข้อสอบ หลังจากคัดเลือกข้อสอบแล้วเหลือข้อสอบระดับ A1 จำนวน 117 ข้อ ระดับ A2 จำนวน 105 ข้อและระดับ B1 จำนวน 111 ข้อ และข้อสอบรวม 6 ข้อ รวมทั้งสิ้น 339 ข้อ จากทั้งหมด 438 ข้อคิดเป็นร้อยละ 77.40 โดยมีผลสรุปค่าพารามิเตอร์ของข้อสอบโดยรวมพบว่าอำนาจจำแนก (α) ทั้ง 3 ระดับมีค่าเฉลี่ยอยู่ที่ 1.278 ส่วนเบี่ยงเบนมาตรฐานอยู่ที่ 0.426 มีค่าต่ำสุดคือ 0.70 และค่าสูงสุดคือ 2.472 ในส่วนของความยากทุกระดับ (β) มีค่าเฉลี่ยอยู่ที่ -0.643 ส่วนเบี่ยงเบนมาตรฐานอยู่ที่ 1.181 มีค่าต่ำสุดคือ -6.227 และค่าสูงสุดคือ 2.890 ซึ่งหากแจกแจงรายละเอียดของ Threshold ทุกระดับจาก β_1 ถึง β_7 หลังจากคัดเลือกข้อที่เป็นไปตามเกณฑ์แล้วแยกตามระดับ CEFR A1, A2 และ B1 ตาราง 3

ตาราง 3 ตารางพิจารณาคุณภาพของข้อสอบการรู้เรื่องการอ่านภาษาอังกฤษ

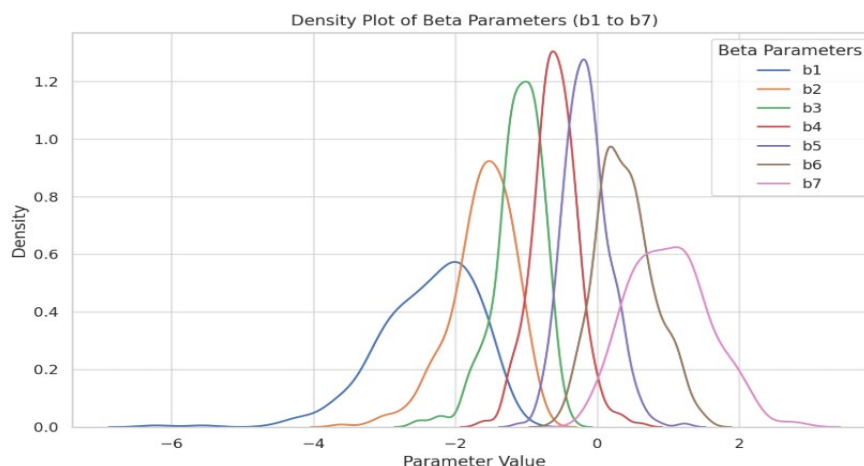
ระดับ A1				
ค่าพารามิเตอร์	Mean	S.D.	Min	Max
อำนาจจำแนก (α)	1.295	0.446	0.720	2.362
ความยาก (β_1)	-2.377	0.673	-4.225	-1.084
ความยาก (β_2)	-1.604	0.404	-2.694	-0.719
ความยาก (β_3)	-1.077	0.296	-2.136	-0.395
ความยาก (β_4)	-0.526	0.300	-1.530	0.392
ความยาก (β_5)	-0.113	0.336	-1.093	1.226
ความยาก (β_6)	0.392	0.413	-0.583	1.513
ความยาก (β_7)	0.954	0.569	-0.107	2.181
ระดับ A2				
ค่าพารามิเตอร์	Mean	S.D.	Min	Max
อำนาจจำแนก (α)	1.307	0.460	0.700	2.472
ความยาก (β_1)	-2.409	1.194	-6.227	-1.366
ความยาก (β_2)	-1.635	0.514	-3.615	-0.910
ความยาก (β_3)	-1.117	0.443	-2.544	-0.445
ความยาก (β_4)	-0.603	0.304	-1.632	0.613
ความยาก (β_5)	-0.179	0.132	-0.808	0.861
ความยาก (β_6)	0.387	0.155	-0.410	1.440
ความยาก (β_7)	0.964	0.257	-0.035	2.251
ระดับ B1				
ค่าพารามิเตอร์	Mean	S.D.	Min	Max
อำนาจจำแนก (α)	1.234	0.418	0.700	2.393

ตาราง 3 (ต่อ)

ระดับ A1				
ค่าพารามิเตอร์	Mean	S.D.	Min	Max
ความยาก (β_1)	-2.432	0.675	-4.485	-1.302
ความยาก (β_2)	-1.660	0.485	-3.030	-0.815
ความยาก (β_3)	-1.156	0.343	-2.238	-0.520
ความยาก (β_4)	-0.667	0.290	-1.299	0.158
ความยาก (β_5)	-0.157	0.271	-0.870	0.725
ความยาก (β_6)	0.428	0.367	-0.296	1.445
ความยาก (β_7)	1.179	0.551	0.237	2.890

*จากข้อสอบทุกระดับ ระดับละ 146 ข้อ ระดับ A1 ผ่านเกณฑ์ 80.13% ระดับ A ผ่านเกณฑ์ 71.91% ระดับ B1 ผ่านเกณฑ์ 76.02%

2.2 เมื่อนำค่าพารามิเตอร์ความยากของข้อสอบที่เป็นไปตามเกณฑ์ มาเขียนกราฟในภาพรวม เพื่อแสดงถึงลักษณะการกระจายของค่าความยากของข้อสอบ พบว่า โค้งการแจกแจงของกราฟมีลักษณะค่อนข้างสมมาตร แสดงว่าข้อสอบในคลังข้อสอบมีความยากง่ายเหมาะสม สามารถนำไปวัดทักษะการรู้เรื่อง การอ่านภาษาอังกฤษได้ ดังภาพ 3

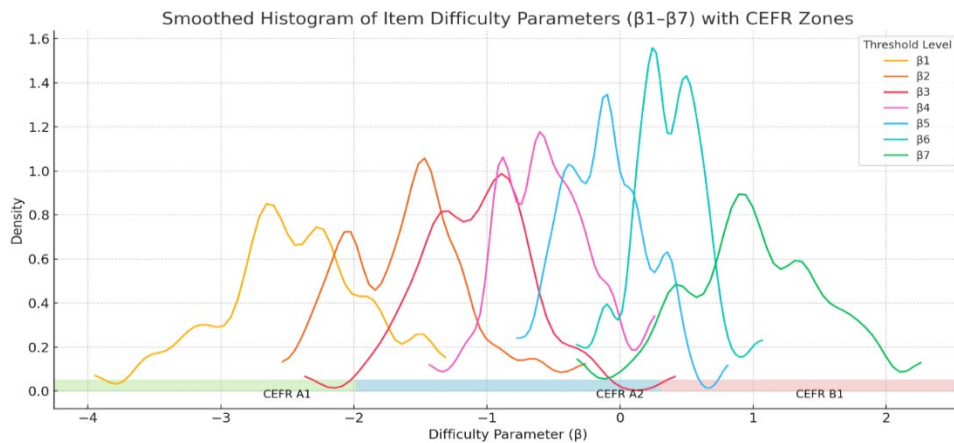


ภาพ 3 ค่าความยากพารามิเตอร์ของข้อสอบที่เป็นไปตามเกณฑ์

จากภาพ 3 แสดงการกระจายของพารามิเตอร์ Threshold β_1 ถึง β_7 ที่ได้จากการวิเคราะห์ข้อสอบในโมเดล Graded Response Model (GRM) สำหรับข้อสอบแบบมี 8 ตัวเลือก (polytomous) แต่ละค่า β ทำหน้าที่เป็นจุดตัด (threshold) ในการแยกผู้ที่เลือกตัวเลือกในระดับที่สูงขึ้นจากระดับที่ต่ำกว่า ในแต่ละข้อคำถาม โดยพารามิเตอร์ β_1 มีค่ากระจายอยู่ในช่วงประมาณ -3.1 ถึง -1.5, β_2 อยู่ในช่วง -2.7 ถึง -1.3, β_3 อยู่ระหว่าง -1.9 ถึง -1.0, β_4 อยู่ในช่วงประมาณ -1.5 ถึง -0.7, β_5 กระจายระหว่าง -1.1 ถึง -0.4,

β_6 อยู่ในช่วง -0.6 ถึง 0.0 และ β_7 มีค่ากระจายอยู่ในช่วงประมาณ -0.5 ถึง 0.5 ซึ่งแสดงให้เห็นลำดับของค่าพารามิเตอร์ที่เพิ่มสูงขึ้นตามระดับของตัวเลือกในแต่ละข้อคำถาม ข้อมูลนี้สะท้อนรูปแบบการจัดเรียงพารามิเตอร์ threshold อย่างเป็นระบบตามลักษณะของโมเดล GRM

2.3 ลักษณะความหนาแน่นของจำนวนข้อสอบ



ภาพ 4 ลักษณะความหนาแน่นของจำนวนข้อสอบอิงตามระดับ CEFR

จากการวิเคราะห์การกระจายของข้อสอบในแต่ละระดับ threshold (β_1 – β_7) ตามภาพ 4 พบว่าข้อสอบในคลังมีการกระจายตัวในแต่ละระดับความยากไม่สม่ำเสมอ โดยข้อสอบส่วนใหญ่กระจุกตัวอยู่ในระดับกลาง ได้แก่ β_4 และ β_3 ซึ่งมีสัดส่วนร้อยละ 20 และ 18 ตามลำดับ รองลงมาคือระดับ β_5 ซึ่งมีร้อยละ 17 ข้อสอบในกลุ่มนี้มีค่าพารามิเตอร์ความยากที่สอดคล้องกับระดับความสามารถของผู้เรียนในช่วง A2 สำหรับระดับ β_2 และ β_6 ซึ่งมีข้อสอบอยู่ร้อยละ 10 และ 12 ตามลำดับนั้น ครอบคลุมกลุ่มผู้เรียนที่อยู่ในช่วงระดับ A1 และ B1 สะท้อนให้เห็นถึงความพยายามในการกระจายข้อสอบไปยังระดับต้นและระดับสูง อย่างไรก็ตาม สัดส่วนยังถือว่าน้อยเมื่อเทียบกับกลุ่มระดับกลาง ส่วนข้อสอบในระดับ β_1 และ β_7 ซึ่งอยู่ที่ร้อยละ 6 รองรับกลุ่มผู้เรียนที่มีสมรรถนะต่ำมาก A1 และ β_7 ซึ่งอยู่ที่ร้อยละ 7 เป็นข้อสอบที่มีความยากอยู่ในช่วงปลายของช่วงค่าความสามารถระดับสูงขึ้นมา B1

2.4 ผลการจัดทำคลังข้อสอบ พบว่า คลังข้อสอบออนไลน์ที่สร้างขึ้นสามารถจัดเก็บข้อสอบที่ประกอบด้วยข้อมูล ดังนี้ หมายเลขข้อสอบ ตัวเลข ตัวอักษร บทความ ตัวเลือก คำอ่านจำแนก ค่าความยาก โดยข้อมูลถูกจัดเก็บไว้ในรูปของตาราง การเข้าสู่ระบบการใช้งานโปรแกรมบน Web Browser ได้หลากหลายช่องทาง ได้แก่ <https://thaistringtest.reru.ac.th/pisit/survey/register.php> ผ่านช่องทางผู้ดูแลระบบ ดังตาราง 4 ภาพ 5 และภาพ 6

ตาราง 4 โครงสร้างฐานข้อมูล

ชื่อตาราง (Table name)	ชื่อฟิลด์ (Field Name)	ชนิดของข้อมูล (Data type)	คำอธิบาย (Description)
Reading Test	ID	item	ข้อที่
	Question	Text	ข้อความคำถาม
	Answer	Text	คำตอบ
	Choice 1	Text	ตัวเลือกที่ 1
	Choice 2	Text	ตัวเลือกที่ 2
	Choice 3	Text	ตัวเลือกที่ 3
	Choice 4	Text	ตัวเลือกที่ 4
	Choice 5	Text	ตัวเลือกที่ 5
	Choice 6	Text	ตัวเลือกที่ 6
	Choice 7	Text	ตัวเลือกที่ 7
	Choice 8	Text	ตัวเลือกที่ 8
	(a)	Numeric	ค่าอำนาจจำแนก
	(b)	Numeric	ค่าจุดตัดความยาก

Item Bank

Series	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
--------	---	---	---	---	---	---	---	---	---	---	----	----	----	----	----	----	----	----	----

Serie 1

มีข้อมูลการตอบแบบทดสอบ จากกระบวนการเก็บข้อมูล 3,870 รายการ

Paragraph 1

Items	a	b1	b2	b3	b4	b5	b6	b7	ref
1	1.1150	-3.0930	-2.6940	-1.9000	-1.5300	-1.0930	-0.4830	0.4000	b1

ข้อมูลการทำแบบสำรวจ 127 รายการ

My favorite animal is the dog. Dogs are friendly and loyal. They make great pets. I have a dog named Max. He is small and has brown fur. Max likes to run and chase birds. He also likes to cuddle and sleep next to me. I take care of Max by feeding him and giving him water every day. Max is a wonderful companion and I am happy to have Max.

What is the main idea of the paragraph?

What is the name of the dog?

What activity does Max like to do every day?

a Favorite animal, the dog, and how they care for it | Max | Go for walks. (7)

b Favorite animal, the dog, and how they care for it | Max | Play with toys. (6)

c Favorite animal, the dog, and how they care for it | Rex | Go for walks. (5)

d Favorite animal, the dog, and how they care for it | Rex | Play with toys. (4)

e Different breeds of dogs. | Max | Go for walks. (3)

f Different breeds of dogs. | Max | Chase birds. (2)

g Different breeds of dogs. | Rex | Go for walks. (1)

h Different breeds of dogs. | Buddy | Sleep next to the writer. (0)

ภาพ 5 ตัวอย่างหน้าจอคลังข้อสอบการรู้เรื่องการอ่านภาษาอังกฤษ

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30

ชุดที่ 7
Next 8

Paragraph

Last summer I went to the aquarium with my parents. We saw many kinds of fish and sea animals. My favorite was the dolphin show. The dolphins were very smart and did many tricks. We also saw sharks, turtles, and colorful fish. There was a touch pool where I could touch starfish and sea urchins. The aquarium was very big and interesting. I had a great time learning about sea life.

What is the main idea of the paragraph?

What was the writers favorite part of the aquarium visit?

What could the writer touch in the touch pool?

☒ The writer enjoyed visiting the aquarium | dolphin show | jellyfish and sea urchins
☐ The writer enjoyed learning about sea life | shark tank | jellyfish and sea urchins
☐ The writer enjoyed learning about sea life | dolphin show | starfish and sea urchins
☐ The writer enjoyed learning about sea life | dolphin show | jellyfish and sea urchins
☐ The writer enjoyed visiting the aquarium | shark tank | jellyfish and sea urchins
☐ The writer enjoyed visiting the aquarium | dolphin show | starfish and sea urchins
☐ The writer enjoyed learning about sea life | shark tank | starfish and sea urchins

Next

ภาพ 6 ผลการทดสอบที่ให้รายละเอียดกับผู้ดูแลระบบ

2.5 ผลการประเมินคลังข้อสอบ

การประเมินคุณภาพของคลังข้อสอบในงานวิจัยนี้อิงตามกรอบแนวคิดของ DeLone & McLean (2003) ซึ่งประกอบด้วยองค์ประกอบหลัก 5 ประการ ได้แก่ 1) คุณภาพของระบบ (system quality) 2) คุณภาพของข้อมูล (information quality) 3) คุณภาพของบริการ (service quality) 4) การใช้งาน (use) และ 5) ผลกระทบสุทธิ (net benefit) โดยสามารถเชื่อมโยงกับหลักการในกรอบ CEFR (Common European Framework of Reference for Languages) ได้ดังนี้

1) System Quality: ระบบมีเสถียรภาพและรองรับการเลือกข้อสอบแบบปรับระดับได้อย่างแม่นยำตามโมเดล IRT ซึ่งสะท้อนหลักการของ CEFR ที่เน้นการประเมินทักษะอย่างต่อเนื่องและสอดคล้องกับระดับความสามารถผู้เรียน

2) Information Quality: ข้อมูลในคลังครอบคลุมทั้งค่าพารามิเตอร์ เนื้อหา และระดับความยากตามหลัก IRT โดยสอดคล้องกับระดับความสามารถด้านการอ่านใน CEFR เช่น ทักษะการจับใจความและการตีความ

3) Service Quality: แม้ระบบเป็นแบบอัตโนมัติ แต่มีการออกแบบเพื่ออำนวยความสะดวกแก่ผู้ใช้ทั้งด้านการเข้าถึง การแสดงผลลัพธ์ และความเข้าใจง่าย ซึ่งสอดคล้องกับหลักการ CEFR ในด้านความเป็นธรรมและความโปร่งใสในการประเมิน

4) Use: ระบบมีความยืดหยุ่น รองรับผู้ใช้หลากหลายกลุ่ม ข้อสอบครอบคลุมระดับ CEFR ตั้งแต่ A1 ถึง B1 ทำให้สามารถใช้งานได้ในบริบทการเรียนรู้ที่แตกต่าง

5) Net Benefit: ระบบสามารถวัดความสามารถได้แม่นยำในช่วงกว้าง รวมถึงระดับสูง ($\theta > 2$) ลดข้อจำกัดของการสอบแบบเดิม และส่งเสริมการใช้ผลประเมินเพื่อการพัฒนาหลักสูตรและการเรียนรู้ตามแนวทาง CEFR

อภิปรายผล

1. ผลด้านการวิเคราะห์ข้อสอบ จากการวิเคราะห์พารามิเตอร์ของข้อสอบจำนวน 438 ข้อ ก่อนการคัดเลือกเข้าสู่คลัง ด้วยโมเดล Graded Response Model (GRM) พบว่า ข้อสอบมีค่าอำนาจจำแนก (α) มีค่าเฉลี่ยที่ 1.278 ซึ่งอยู่ในระดับที่ถือว่าดีตามเกณฑ์ของ Baker (2001) อยู่ในช่วง 0.7–2.5 อย่างไรก็ตาม ยังพบข้อสอบบางส่วนที่มีค่า α ต่ำหรือสูงกว่าช่วงดังกล่าว ซึ่งอาจมีผลมาจากผู้สอบที่คละความสามารถจากกลุ่มตัวอย่างที่กระจายเก็บข้อมูล สำหรับค่าพารามิเตอร์ความยาก (threshold parameters: β_1 – β_7) มีลักษณะเรียงลำดับจากง่ายไปยากตามแนวคิดของ Samejima (1969) โดยค่าเฉลี่ยของ threshold β_1 แรกเริ่มจากประมาณ -2.4 และเพิ่มขึ้นต่อเนื่องไปจนถึง β_7 ที่เฉลี่ยประมาณ 0.89 ซึ่งแสดงให้เห็นถึงโครงสร้างภายในของข้อสอบที่มีแนวโน้มสอดคล้องกับการออกแบบข้อสอบแบบพหุวิภาค และมีศักยภาพในการวัดผู้เรียนตั้งแต่ระดับความสามารถต่ำถึงสูง ทั้งนี้ได้มีการคัดข้อสอบที่ไม่ผ่านเกณฑ์อำนาจจำแนก ต่ำกว่า 0.70 และไม่เกิน 2.5 และความยากไม่เกิน 3.00 สอดคล้องงานวิจัยของ Loken & Rulison (2010) เพื่อนำเข้าคลังต่อไป

2. ผลด้านคุณภาพของคลังข้อสอบ ข้อสอบที่ผ่านการคัดเลือกเข้าสู่คลังจำนวน 339 ข้อ มีลักษณะทางพารามิเตอร์ที่เหมาะสมมากขึ้นจากการกรองด้วยเกณฑ์เชิงเทคนิค โดยเฉพาะในด้านค่าความยากซึ่งมีค่าเฉลี่ยรวมอยู่ที่ -0.643 และช่วงค่าระหว่าง -6.227 ถึง 2.890 แสดงถึงการกระจายของข้อสอบที่ครอบคลุมช่วงความสามารถของผู้เรียนได้อย่างกว้างพอสมควร แม้ว่าค่าความยากโดยรวมจะมีแนวโน้มต่ำหรือปานกลาง อันสะท้อนบริบทของกลุ่มผู้สอบที่เป็นนักเรียนมัธยมศึกษาตอนปลายทั่วไป อย่างไรก็ตาม ลักษณะเช่นนี้กลับสอดคล้องกับผลการศึกษาหลายฉบับในบริบทไทย เช่น งานของ Raksombat et al. (2020) และ Romkaew et al. (2021) ที่พบว่าคลังข้อสอบที่มีข้อสอบระดับง่ายถึงกลาง มีความเหมาะสมต่อการวัดผลกับผู้เรียนกลุ่มใหญ่ และเอื้อต่อการเริ่มต้นการประมาณค่าความสามารถได้อย่างแม่นยำในระบบ CAT ในขณะเดียวกัน การจัดเรียงข้อสอบตามระดับ threshold และการกระจายค่าพารามิเตอร์อย่างเป็นระบบ อย่างไรก็ตามลักษณะดังกล่าวกลับกลายเป็นจุดแข็งสำคัญของคลังข้อสอบ เพราะสามารถตอบสนองการวัดผลในกลุ่มผู้เรียนระดับพื้นฐานได้ดี โดยข้อสอบส่วนใหญ่มี threshold ครอบคลุมช่วงความสามารถระหว่างระดับ A1-B1 ตามกรอบ CEFR ได้อย่างต่อเนื่อง โดยเฉพาะใน threshold กลางถึงต่ำ (β_1 - β_4) ที่มีความหนาแน่นของข้อสอบสูง ซึ่งสอดคล้องกับแนวทางการพัฒนาคลังข้อสอบเพื่อรองรับ CAT ซึ่งช่วยให้ระบบสามารถเลือกข้อสอบได้เหมาะสมยิ่งขึ้นตามระดับความสามารถจริง

เมื่อพิจารณาคุณภาพของคลังข้อสอบตามกรอบแนวคิดของ DeLone & McLean (2003) ซึ่งประกอบด้วย 5 ด้านหลัก พบว่าคลังข้อสอบที่พัฒนาขึ้นมีความเหมาะสมทั้งในด้านโครงสร้าง 1) ด้านคุณภาพของระบบ (system quality) ระบบฐานข้อมูลออนไลน์มีเสถียรภาพสูง สามารถประมวลผลและคัดเลือกข้อสอบตามค่าความสามารถของผู้สอบได้แบบอัตโนมัติ ซึ่งสอดคล้องกับแนวคิดของ van der Linden & Glas (2000) ที่ชี้ว่าระบบ CAT ต้องอาศัยกลไกที่สามารถปรับเปลี่ยนข้อสอบอย่างต่อเนื่องตามค่าพารามิเตอร์ของผู้สอบ 2) ด้านคุณภาพของข้อมูล (information quality) ข้อสอบแต่ละข้อมีค่าพารามิเตอร์จากโมเดล GRM อย่างครบถ้วน ทั้งค่าความยากแบบ threshold และอำนาจจำแนก ซึ่งสะท้อนความแม่นยำในการวัดตามแนวคิดของ Samejima (1969) และ Thissen et al. (2000) 3) ด้านคุณภาพของบริการ (service quality) ระบบสามารถให้ข้อมูลย้อนกลับที่สื่อสารได้กับทั้งผู้เรียนและครูผู้สอนอย่างชัดเจน ซึ่งสอดคล้องกับความคาดหวังของระบบประเมินภาษาในบริบท CEFR ที่มุ่งให้ผู้เรียนเข้าใจระดับความสามารถของตน 4) ด้านการใช้งาน (Use) ระบบมีความยืดหยุ่นสูง และสามารถเข้าถึงได้ง่ายผ่านเว็บเบราว์เซอร์ ไม่จำกัดอุปกรณ์หรือเวลา สอดคล้องกับข้อเสนอของ DeLone & McLean (2003) ที่เน้นความสามารถในการใช้งานจริงและตอบสนองผู้ใช้หลากหลายกลุ่ม 5) ด้านผลกระทบสุทธิ (Net Benefit) คลังข้อสอบนี้สามารถวัดความสามารถของผู้เรียนได้ในเวลาที่หลากหลาย รวมถึงระดับที่สูงกว่า $\theta > 2$ ซึ่งโดยทั่วไปมักเป็นข้อจำกัดของการสอบแบบดั้งเดิม (Dai et al., 2021) ข้อมูลจากระบบยังสามารถนำไปใช้ในการวิเคราะห์ผลเพื่อวางแผนการเรียนการสอนได้อย่างเป็นระบบ และการประยุกต์ใช้งานในระบบทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (CAT) (DeLone & McLean, 2003)

ข้อเสนอแนะ

1. ข้อเสนอแนะในการวิจัยครั้งต่อไป

การพัฒนาและตรวจสอบคุณภาพของข้อสอบเพิ่มเติมควรดำเนินการสร้างข้อสอบเพิ่มเติมเพื่อเพิ่มความหลากหลายของระดับความยาก โดยเฉพาะในช่วงค่าความสามารถระดับสูง ($\theta > 2$) ซึ่งยังมีจำนวนข้อสอบน้อยในคลัง จากผลที่พบว่าค่าความยากโดยเฉลี่ยอยู่ที่ -0.643 จึงควรเน้นการสร้างข้อสอบที่ยากขึ้นเพื่อให้ครอบคลุมกลุ่มผู้สอบที่มีความสามารถสูง และรองรับการเลือกข้อสอบแบบปรับเหมาะในอนาคต

2. ข้อเสนอแนะสำหรับการนำไปใช้ประโยชน์

2.1 ควรส่งเสริมให้ครูผู้สอน โดยเฉพาะกลุ่มสาระภาษาต่างประเทศหรือผู้รับผิดชอบด้านการวัดผลในโรงเรียนมีความเข้าใจในการใช้คลังข้อสอบที่พัฒนาขึ้นด้วยทฤษฎีการตอบสนองข้อสอบแบบพหุวิภาค (polytomous IRT) และสร้างข้อสอบในรูปแบบนี้เพื่อเลือกใช้ข้อสอบให้เหมาะสมกับระดับผู้เรียนในแต่ละระดับชั้น หรือใช้ข้อสอบที่ให้สารสนเทศสูงอิงตามความสามารถของผู้เรียนในโรงเรียนของตน ซึ่งจะช่วยยกระดับคุณภาพของการวัดผลให้เที่ยงตรงมากขึ้น และลดการใช้ข้อสอบที่ไม่มีคุณภาพ

2.2 การพัฒนาระบบคลังข้อสอบให้รองรับการสอบแบบ CAT เต็มรูปแบบ ควรพัฒนาระบบฐานข้อมูลให้สามารถเชื่อมต่อกับโปรแกรมสอบแบบ Adaptive ได้ เช่น การเชื่อมต่อกับระบบประมาณค่า θ ปรับตามแบบเรียลไทม์ การเลือกข้อสอบอัตโนมัติตามพารามิเตอร์

References

- Alderson, J. C. (2010). *Assessing Reading*. Cambridge University Press.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford University Press.
- Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation.
- Brown, D. H. (2004). *Language assessment: Principles and classroom practices*. Pearson Education.
- Chalmers, R. P. (2012). MIRT: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1–29.
- Chattergoon, R. (2020). *Using polytomous item response theory models to validate learning progressions* [Doctoral dissertation]. University of Colorado.
- Dai, S., Vo, T. T., Kehinde, O. J., He, H., Xue, Y., Demir, C., & Wang, X. (2021). Performance of polytomous IRT models with rating scale data: An investigation over sample size, instrument length, and missing data. *Frontiers in Education*, 6, Article 721963.

- DeLone, W. H., & McLean, E. R. (2003). The DeLone and McLean model of information systems success: A ten-year update. *Journal of Management Information Systems*, 19(4), 9–30.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Psychology Press.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer Academic.
- Loken, E., & Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *The British Journal of Mathematical and Statistical Psychology*, 63(3), 509–525.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.
- Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. SAGE.
- Pinitasakul, P., Senarat, S., & Senarat, B. (2020). Item response theory and testlet test. *Ubon Ratchathani Journal of Research and Evaluation*, 9(2), 74–84.
- Ravand, H. (2015). Assessing testlet effect, impact, differential testlet, and item functioning using cross-classified multilevel measurement modeling. *SAGE Open*, 5(2), 1–9.
- Richards, J. C. (2015). *Key issues in language teaching*. Cambridge University Press.
- Samejima, F. (1969). *Estimation of Latent Ability Using a Response Pattern of Graded Scores*. Psychometrika Monograph.
- Senarat, S., Tayraukham, S., Piyapimonsit, C., & Tongkhambanjong, S. (2013). Development of a computerized adaptive testing for diagnosing the cognitive process of grade 7 students in learning algebra, using multidimensional item response theory. *Educational Research and Reviews*, 8(13), 1009–1021.
- Thissen, D., Mislevy, R. J., & Wainer, H. (2000). *Testing algorithms*. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed.). Lawrence Erlbaum Associates.
- van der Linden, W. J., & Glas, C. A. W. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Kluwer Academic.
- Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge University Press.
- Waltz, C. F., Strickland, O. L., & Lenz, E. R. (2010). *Measurement in nursing and health research* (4th ed.). Springer.

- Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, 6(4), 473–492.
- Wheadon, B. C. (2011). *An item response theory approach to the maintenance of standards in public examinations in England* [Doctoral dissertation]. Durham University.
<http://etheses.dur.ac.uk/615/>
- Yen, W. M. (1996). *Polytomous item response theory models and their applications in large scale testing programs: Review of literature* (ETS Research Memorandum No. 96-08). Educational Testing Service.

Translate Thai References

- Kanjanawasee, S. (2020). *Modern test theory* (5th ed.). Chulalongkorn University. (in Thai).
- Raksombat, S., Pradujphom, P., & Panthong, K. (2020). Item bank classified by content and difficulty levels using item response theory. *Journal of Educational Measurement Mahasarakham University*, 26(2), 322–333. (in Thai).
- Romkaew, N., Senarat, S., & Saengprom, N. (2021). Application of item response theory in development of the item bank of basic knowledge in physical education. *Journal of Educational Measurement Mahasarakham University*, 27(2), 82–98. (in Thai).