

การจัดสร้างแบบทดสอบทดสอบภาษาอังกฤษ Engenius English Language Tests (EELT) สำหรับนักเรียนระดับมัธยมศึกษาและการเทียบคะแนนผลการทดสอบกับ แบบทดสอบ English Language Test for International Students (ELTiS)

ศศ วัฒนโกคากุล¹

สุพัฒน์ สุกมลสันต์²

¹ ผู้นิพนธ์ประสานงาน โทรศัพท์ 02-4414401 ต่อ 1544 อีเมล : sasa.wat@mahidol.edu

รับเมื่อ 30 มีนาคม 2565 วันที่แก้ไขบทความ 20 มิถุนายน 2565 ตอบรับเมื่อ 21 มิถุนายน 2565

บทคัดย่อ

การเทียบคะแนนของแบบทดสอบที่สร้างขึ้นใหม่กับแบบทดสอบมาตรฐานที่มีจำนวนจำกัดเป็นสิ่งสำคัญมากเพื่อความปลอดภัยของแบบทดสอบ และความน่าเชื่อถือของผลการสอบ การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อจัดสร้างแบบทดสอบ Engenius English Language Test (EELT) สำหรับนักเรียนระดับมัธยมศึกษาชั้นปีที่ 1-6 ให้เป็นแบบทดสอบมาตรฐานในระดับสากลจำนวน 6 ชุดที่คู่ขนานจริง กับแบบทดสอบชุดเดิมคือ แบบทดสอบ English Language Test for International Students (ELTiS) และเพื่อสร้างตารางเทียบคะแนนของแบบทดสอบทั้งสองด้วย 2 วิธี คือ เทียบตำแหน่งเปอร์เซ็นต์ไทล์ และวิธีการถดถอยเชิงเส้นตรง กลุ่มตัวอย่างคือ ผู้สอบแบบทดสอบแต่ละชุดมี 300 คน มาจากนักเรียนแต่ละระดับชั้น 50 คน รวมเป็นทั้งสิ้น 1,800 คน ซึ่งได้มาจากการสุ่มอย่างง่ายจากผู้ที่มาสมัครสอบโดยไม่เสียค่าใช้จ่าย แบบทดสอบ EELT ชุดใหม่ทั้ง 6 ชุดสร้างจากตารางกำหนดลักษณะเฉพาะของแบบทดสอบ ELTiS และตรวจสอบความตรงเชิงเนื้อหาแล้วว่ามีคุณภาพอยู่ในระดับสูงมากโดยผู้เชี่ยวชาญด้านการเรียนการสอน และการวัดและประเมินผลทางภาษาอังกฤษ 5 ท่าน และเก็บข้อมูลจากการจัดสอบแบบใช้ผู้สอบกลุ่มเดียวและสลับแบบทดสอบ แล้ววิเคราะห์ข้อมูลด้วย Independent Samples t-test วิธีเทียบคะแนนเชิงเส้นตรงของเลอวิน สหสัมพันธ์แบบเพียร์สัน วิธีเทียบคะแนนโดยใช้ตำแหน่งของคะแนน และวิธีเทียบคะแนนเชิงเส้นตรงการถดถอย ผลการวิจัยพบว่า แบบทดสอบ EELT ชุดที่ 1, 2 และ 4 เป็นแบบทดสอบคู่ขนานจริงกับแบบทดสอบ ELTiS แต่แบบทดสอบ EELT ชุดที่ 3, 5 และ 6 เป็นแบบทดสอบคู่ขนานเทียม และได้ตารางเทียบคะแนนระหว่างแบบทดสอบ EELT กับ ELTiS จำนวน 2 ตารางจากผลการเทียบคะแนน 2 วิธี ผลของการวิจัยได้แบบทดสอบใหม่ 3 ชุดที่สามารถใช้แทนแบบทดสอบมาตรฐานได้ ส่วนอีก 3 ชุดควรต้องได้รับการปรับปรุงแก้ไขบางอย่างก่อนที่จะนำไปใช้จริงต่อไป

คำสำคัญ : การเทียบคะแนนแบบทดสอบ วิธีเทียบคะแนนโดยใช้ตำแหน่งของคะแนน วิธีเทียบคะแนนเชิงเส้นตรงการถดถอย, แบบทดสอบ Engenius English Language Test (EELT) แบบทดสอบ English Language Test for International Students (ELTiS) การสอบแบบใช้ผู้สอบกลุ่มเดียวและสลับแบบทดสอบ

¹ ผู้ช่วยศาสตราจารย์ประจำคณะศิลปศาสตร์ มหาวิทยาลัยมหิดล

² รองศาสตราจารย์ประจำวิทยาลัยนานาชาติ มหาวิทยาลัยแม่โจ้

A Development of the Engenius English Language Tests (EELT) for High School Students and their Equivalence with the English Language Test for International Students (ELTiS)

Sasa Watanapokakul ¹

Suphat Sukamolson ²

¹ Corresponding author, Tel. 02-4414401 ext. 1544, E-mail : sasa.wat@mahidol.edu

Received 30 March 2022; Revised 20 June 2022; Accepted 21 June 2022

Abstract

Equating the scores of a newly created test to a limited number of standardized tests is very important for test security and the reliability of the test results. The purposes of this research were to construct six true parallel forms of Engenius English Language Test (EELT) for Grades 1-6 students as an international standardized test with the original English Language Test for International Students (ELTiS) and to construct scoreboards for the two tests by two equating methods namely Equipercentile Method and Linear Regression Method. The subjects taking each test form consisted of 300 students, 50 students from each grade level, for a total of 1,800 students. They were randomly selected from candidates for taking English tests for free. All six new forms of EELT tests were constructed from a test item specification table of the ELTiS test and were verified for content validity at a very high level by five experts in the fields of English teaching and learning, language measurement, and evaluation. Data were collected from the Single Group Design with Counterbalancing and were analyzed by means of Independent Samples t-tests, Levine's tests, Pearson Correlation, Equipercentile Method and Linear Regression Method. It was found that EELT Forms 1, 2 and 4 were true parallel forms with ELTiS, but EELT Forms 3, 5, and 6 were pseudo-parallel forms. Two tables to equate the scores between the EELT and ELTiS tests were constructed based on the two equating methods. The results showed that three new tests could be used in place of the standardized test, while the other three should undergo some revisions before they could be used in the future.

Keywords : Test score equating, Equipercentile method, Linear regression method, Engenius English Language Test (EELT), English Language Test for International Students (ELTiS), Single Group Design with Counterbalancing

¹ Assistant Professor, Ph.D., Faculty of Liberal Arts, Mahidol University

² Associate Professor, Ph.D., International College, Maejo University

1. หลักการและเหตุผล

เนื่องจากสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน มีนโยบายยกระดับมาตรฐานภาษาอังกฤษของนักเรียนในโรงเรียนมัธยมศึกษาทั่วประเทศให้เป็นผู้ที่มีความพร้อมทางวิชาการและทักษะการสื่อสารภาษาอังกฤษในระดับที่ใช้งานได้เป็นอย่างดี ทั้งนี้เพื่อเสริมสร้างสมรรถนะและความสามารถในการสื่อสารภาษาอังกฤษของนักเรียนไทย โดยกำหนดให้นักเรียนระดับชั้นมัธยมศึกษาปีที่ 1-6 ก่อนสำเร็จการศึกษาให้มีระดับความสามารถด้านการใช้ภาษาอังกฤษตามเกณฑ์ Common European Framework of Reference for Languages หรือที่เรียกว่า CEFR (Council of Europe, 2020) ที่ B1 และให้ครูสอนภาษาอังกฤษในระดับประถมศึกษาและมัธยมศึกษาที่มีความสามารถในระดับ B1 และ B2 (คุรุสภา, 2563) และแผนการศึกษาแห่งชาติ พ.ศ. 2560-2579 ได้กำหนดตัวชี้วัดตามเป้าหมายของแผนการศึกษาแห่งชาติในด้านภาษาอังกฤษไว้ โดยมีเป้าหมายที่จะพัฒนานักเรียนให้นักเรียนระดับมัธยมศึกษาตอนต้น มัธยมศึกษาตอนปลาย และปริญญาตรีก่อนสำเร็จการศึกษาให้มีระดับความสามารถด้านการใช้ภาษาอังกฤษตามเกณฑ์ CEFR ในระยะสุดท้ายของแผนคือ ปี พ.ศ. 2575-2579 ให้อยู่ในระดับ B1, B2 และ C1+ ตามลำดับ (สำนักงานเลขาธิการสภาการศึกษา, 2560) นอกจากนี้โรงเรียนต่าง ๆ ยังมีนโยบายในการพัฒนาทักษะการใช้ภาษาอังกฤษของนักเรียนโดยกำหนดให้มีการจัดทดสอบความรู้ภาษาอังกฤษของนักเรียนทุกปีการศึกษา ทั้งนี้เพื่อให้นักเรียนได้ทราบระดับทักษะภาษาอังกฤษของตนเองและสามารถพัฒนาทักษะภาษาอังกฤษของตนเองต่อไปได้

ด้วยเหตุนี้โครงการแลกเปลี่ยนภาษาและวัฒนธรรมนานาชาติเอ็นจีเนียส อินเตอร์เนชันแนล (Engenius International Student Exchange Program) จึงได้มีการจัดโครงการสอบประเมินความสามารถทางภาษาอังกฤษตามมาตรฐานของ CEFR ที่ใช้อธิบายระดับความสามารถทางภาษาของนักเรียนระดับชั้นมัธยมศึกษา ซึ่งในปัจจุบันการทดสอบใช้ข้อสอบภาษาอังกฤษชื่อ ELTIS (English Language Test for International Students) เป็นแบบทดสอบมาตรฐานเพื่อวัดทักษะวิชาภาษาอังกฤษระดับมัธยมศึกษา เฉพาะสำหรับนักเรียนต่างชาติ (ที่ไม่ได้ใช้ภาษาอังกฤษเป็นหลัก) โดยทำการประเมินความสามารถด้านทักษะการฟังและการอ่านภาษาอังกฤษ เพื่อให้นักเรียนได้รับการทดสอบทักษะภาษาอังกฤษอย่างได้มาตรฐานจากข้อสอบที่ได้รับการรับรองมาตรฐานสากล นักเรียนจะได้ทราบระดับทักษะภาษาอังกฤษของตนเอง และสามารถวางแผนในการพัฒนาทักษะภาษาอังกฤษของตนเองต่อไปในอนาคตได้ รวมทั้งครูผู้สอนยังจะได้ทราบระดับทักษะภาษาอังกฤษโดยรวมของนักเรียนในโรงเรียนด้วย ทั้งนี้เพื่อให้สามารถวางแผนหรือส่งเสริมนักเรียนในทักษะภาษาอังกฤษด้านที่ยังขาดอยู่ของนักเรียนต่อไปได้ในอนาคต

ทั้งนี้ทางโครงการฯ มีความตั้งใจที่จะจัดทดสอบวัดความสามารถทางภาษาอังกฤษให้กับนักเรียนในระดับมัธยมศึกษาตามมาตรฐานความสามารถทางภาษาอังกฤษตามเกณฑ์ของ CEFR ให้กับนักเรียนเป็นประจำทุกปี แต่หากใช้แบบทดสอบ ELTIS ซึ่งมีจำนวนเพียง 1 ชุดซ้ำอาจทำให้ผลการสอบของนักเรียนที่ทำการทดสอบครั้งที่ 2 เป็นต้นไปมีผลกระทบนำพา (Carry-over Effects) จากการสอบครั้งแรก ซึ่งอาจเกิดจากสาเหตุหลายอย่าง เช่น การเรียนรู้ การจดจำข้อสอบ การฝึกซ้อม การเดา และสภาพแวดล้อมที่ต่างกัน เป็นต้น และอาจมีผลเสียต่อความตรง (Validity) และความเที่ยง (Reliability) ของแบบทดสอบ และทำให้ผลการทดสอบไม่ตรงกับความเป็นจริงได้ (Kontos, 2007; Shah & Brown, 2020) ดังนั้นทางโครงการฯ จึงมีความต้องการที่จะจัดสร้างแบบทดสอบชื่อ Engenius English Language Test (EELT) ขึ้นหลายชุด และทำการเทียบคะแนนกับคะแนนของแบบทดสอบ ELTIS ที่มีอยู่แล้ว ทั้งนี้เพื่อให้แบบทดสอบชุดใหม่มีความเป็นมาตรฐานและแก้ปัญหาจากผลกระทบนำพา (Carry-over Effects) จากการสอบครั้งแรก ดังกล่าวแล้ว

2. การทบทวนวรรณกรรม

การเทียบคะแนนของแบบทดสอบมีหลากหลายวิธี แต่สามารถจำแนกออกเป็น 2 จำพวกใหญ่ (Kolen & Brennan, 2004; Brennan et al., 2009) คือ

1. วิธีเทียบคะแนนแบบมาตรฐาน (Standard Methods) โดยใช้คะแนนที่ปรากฏ (Observed scores) ซึ่งมีวิธีย่อยอีก 13 วิธี คือ

- 1.1 วิธีเทียบคะแนนโดยใช้คะแนนเฉลี่ย (Mean Equating Method)
- 1.2 วิธีเทียบคะแนนเชิงเส้นตรง (Linear Equating Method)
- 1.3 วิธีเทียบคะแนนเชิงเส้นตรงการถดถอย (Linear Regression Method)
- 1.4 วิธีเทียบคะแนนเชิงเส้นตรงของทักเกอร์ (Tucker Linear Equating Method)

1.5 วิธีเทียบคะแนนเชิงเส้นตรงของเลวีนโดยใช้คะแนนที่ปรากฏ (Levine Observed-Score Linear Equating Method)

1.6 วิธีเทียบคะแนนเชิงเส้นตรงของเลวีนโดยใช้คะแนนจริง (Levine True-Score Linear Equating Method)

1.7 วิธีเทียบคะแนนโดยใช้ตำแหน่งของคะแนน (Equipercentile Equating Method)

1.8 วิธีเทียบคะแนนโดยการประมาณค่าความถี่ (Frequency Estimation Equating Method)

1.9 วิธีเทียบคะแนนเชิงเส้นตรงของบราวน์และฮอลแลนด์ (Braun-Holland Linear Equating Method)

1.10 วิธีเทียบคะแนนโดยใช้ตำแหน่งของคะแนนต่อเนื่อง (Chained Equipercentile Equating Method)

1.11 วิธีเทียบคะแนนปรากฏในพื้นที่ (Local Observed-Score Equating Method)

1.12 วิธีเทียบคะแนนโดยการประมาณค่าความถี่ที่ปรับแก้ (Modified Frequency Estimation Equating Method)

1.13 วิธีเทียบคะแนนโดยใช้หลักการกระจายปกติ (Gaussian Kernel Equating Method)

2. วิธีเทียบคะแนนแบบมาตรฐาน (Standard Methods) โดยใช้คะแนนที่ปรากฏ (Observed scores) หรือคะแนนจริง (True scores) ซึ่งมีวิธีย่อยอีก 7 วิธี คือ

2.1 วิธีเปลี่ยนค่าเฉลี่ย 2 ค่า (Mean/Mean Transformation Method)

2.2 วิธีเปลี่ยนค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐาน (Mean/Sigma Transformation Method)

2.3 วิธีฝ่าฝืนการเปลี่ยนค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐาน (Robust Mean/Sigma Transformation Method)

2.4 วิธีใช้ค่าลักษณะความโค้งของแฮบารา (Haebara Characteristic Curve Method)

2.5 วิธีใช้ค่าลักษณะความโค้งของสต็อกกิงและลอร์ด (Stocking-Lord Characteristic Curve Method)

2.6 วิธีใช้ลักษณะความโค้งเฉพาะ (Characteristic Curve Transformation Method)

2.7 วิธีใช้ค่าลักษณะความโค้งแบบราซ (Rasch Characteristic Curve Method)

แต่ละวิธีเหมาะกับรูปแบบ (Design) ของการวิจัยเพื่อการเทียบที่ต่างกัน และมีข้อดกกลางเบื้องต้นแตกต่างกัน รวมทั้งมีข้อดีและข้อเสียแตกต่างกัน แต่ปัจจุบันนี้ยังไม่เป็นที่ยอมรับกันทั่วไปว่าวิธีใดดีที่สุด

สำหรับการวิจัยครั้งนี้ ผู้วิจัยสนใจที่จะเทียบคะแนนของแบบทดสอบโดย Equipercentile Method และ Linear Regression Method เพราะว่าเป็นวิธีที่ได้รับความนิยมกันทั่วไปอย่างแพร่หลาย เนื่องจากเป็นวิธีที่ไม่ซับซ้อน แต่ได้ผลลัพธ์ที่ดีเทียบเท่ากับหลายวิธีที่ซับซ้อนมากกว่า (Phillips, 1983; Rapp, 1999; Heh, 2007; สรัญญา จันทร์ชูสกุล และคณะ, 2560) โดยมีงานวิจัยที่สนับสนุนจากผลการศึกษาระียบเทียบกับวิธีต่าง ๆ อย่างต่อเนื่องมานานแล้ว เช่น จากการศึกษาเปรียบเทียบการเทียบคะแนนโดย Equipercentile Observed-score Method กับ Item Response Theory (IRT) True-score Method ในกรณีที่ใช้แบบทดสอบรวม (Anchor Test) และการกระจายของคะแนนดิบได้รับการปรับเรียบด้วยวิธีที่เหมาะสมมาก่อนให้เป็นโค้งปกติ (Normal Curve) แล้ว การเทียบคะแนนของทั้ง 2 วิธีได้ผลดีเช่นเดียวกัน (Lord & Wingersky, 1984) ผลการศึกษานี้สอดคล้องกับผลการศึกษาของ Kolen (1981 อ้างถึงใน Lord & Wingersky, 1984) ว่า Equipercentile Observed-score Method เป็นวิธีเทียบคะแนนที่ดีที่สุดเมื่อเปรียบเทียบกับวิธีอื่นอีก 9 วิธี ทั้งนี้หากว่าการกระจายของคะแนนของแบบทดสอบที่นำมาเทียบทั้งคู่เป็นโค้งปกติ

ต่อมาในปี 1987 ได้มีการศึกษาเปรียบเทียบผลการเทียบคะแนนโดย Equipercentile Method จำนวน 14 รูปแบบที่ต่างกัน (Fairbank Jr., 1987) และมี 7 รูปแบบที่ทำการปรับลักษณะโค้งของการกระจายของคะแนนก่อน (Presmoothing) เพื่อให้เป็นโค้งปกติ จากผลการศึกษาพบว่า การเทียบคะแนนด้วยวิธี Negative Hypergeometric Model มีความถูกต้องมากที่สุด เมื่อมีการปรับลักษณะโค้งของการกระจายของคะแนนมาก่อน (Presmoothing) และวิธี Orthogonal Regression Model และ Cubic Splines Model เหมาะสมในกรณีที่ทำการปรับลักษณะโค้งของการกระจายของคะแนนภายหลัง (Postsmoothing) และยังพบว่าวิธี Linear Regression Method และ Orthogonal Regression Method เป็นวิธีเทียบคะแนนที่ดีที่สุด เมื่อต้องการเทียบคะแนนของแบบทดสอบ X ไปยังแบบทดสอบ Y ผลการศึกษานี้สอดคล้องกับผลการศึกษาที่มีมาก่อนแล้ว ของ Lord และ Novick (Lord & Novick, 1968) ที่พบว่า คะแนนของแบบทดสอบ 2 ชุดสามารถเทียบกันได้ด้วย Linear Regression Method เมื่อการกระจายของแบบทดสอบทั้งคู่มีลักษณะเหมือนกัน แม้ว่าค่าเฉลี่ยของคะแนน และค่าเบี่ยงเบนมาตรฐานจะแตกต่างกัน หรือความยาก-ง่ายของคะแนนต่างกัน

อนึ่ง จากการวิจัยของนักวิจัยหลายคนในระยะต่อมาพบว่า ในกรณีของคะแนนดิบของแบบทดสอบที่ต้องการนำคะแนนมาเทียบกัน มีรูปแบบการกระจายที่เหมือนกัน ก็ไม่จำเป็นที่จะต้องปรับการกระจายของคะแนนให้เป็นโค้งปกติก่อน หรือหลังการเทียบคะแนนแล้วก็ได้ผลลัพธ์ที่คล้ายคลึงกัน แต่ในกรณีที่การกระจายของคะแนนของแบบทดสอบทั้งคู่แตกต่างกัน การปรับคะแนน (smoothing) ยังมีความจำเป็นที่จะต้องทำ (Livingston, 1992; Hanson et al., 1994; Kolen & Brennan, 2004; Cui & Kolen, 2009) นอกจากนี้ Hanson (1991) ยังให้คำแนะนำว่า ถ้าขนาดของกลุ่มตัวอย่างมีขนาดเล็ก ควรจะทำการปรับลักษณะโค้งของการกระจายของคะแนนก่อนการเทียบคะแนน โดย Equipercentile Method แต่ถ้ากลุ่มตัวอย่างมีขนาดใหญ่ ไม่ควรทำการปรับลักษณะโค้งของการกระจายของคะแนนก่อนการเทียบคะแนน เพราะจะทำให้เกิดความคลาดเคลื่อนมากและแนะนำให้ควรทำการทดสอบการกระจายของคะแนนของแบบทดสอบทั้ง 2 ชุดเสียก่อนเพื่อความแน่ใจด้วย Fit statistics เช่น Goodness of fit เป็นต้น

ต่อมาในปี 2011 Moses and Liu (2011) ได้มีการศึกษาเปรียบเทียบวิธีการเทียบคะแนนหลายรูปแบบและหลายเงื่อนไข และพบว่า การเทียบคะแนนโดย Equipercentile Method ที่ใช้ Loglinear Presmoothing Model, Kernel Equating Model และ Cubic Spline Posts smoothing Model ให้ผลการเทียบคะแนนที่ดีและคล้ายคลึงกัน และสรุปว่า ถ้าการกระจายของคะแนนของแบบทดสอบที่ต้องการเทียบคะแนน มีลักษณะคล้ายกัน การปรับคะแนนด้วยวิธีต่าง ๆ จะได้ผลการเทียบคะแนนที่คล้ายคลึงกัน แต่หากว่าการกระจายของคะแนนของแบบทดสอบที่ต้องการเทียบคะแนนมีลักษณะแตกต่างกัน ก็จำเป็นต้องมีการปรับการกระจายของคะแนน และใช้รูปแบบหรือวิธีการเทียบคะแนนที่เหมาะสมกว่า

นอกจากนี้ ในปี 2017 สรัญญา จันทรชูสกุล และคณะ (2560) ได้ศึกษาเปรียบเทียบคุณภาพของการเทียบคะแนนสอบภาษาอังกฤษระหว่าง Equipercentile Method, Linear Method และ Linear Regression Method เพื่อเทียบคะแนนสอบภาษาอังกฤษระหว่างคะแนน SEPT (Silpakorn English Proficiency Test) และคะแนน TOEIC และเพื่อเปรียบเทียบคุณภาพของวิธีปรับเทียบระหว่าง 3 วิธีดังกล่าว ผลการวิจัยพบว่า การเทียบคะแนนทั้ง 3 วิธีได้ผลดีไม่แตกต่างกัน มีค่าความคลาดเคลื่อนไม่แตกต่างกัน และมีประสิทธิภาพในการเทียบไม่แตกต่างกัน แต่ Linear Method และ Linear Regression Method มีความสะดวกในการใช้งานมากกว่า Equipercentile Method เพราะวิธีหลังนี้ให้ค่าคะแนนบางค่าที่เทียบแล้วซ้ำค่ากันได้

เมื่อไม่นานมานี้ มีงานวิจัยเกี่ยวกับการใช้แบบทดสอบร่วม (Anchor Test) เพื่อการศึกษาเชิงเปรียบเทียบผลของการเทียบคะแนน 4 วิธี กับกลุ่มผู้สอบจำนวนน้อย คือ ประมาณ 200 คน ในระดับชั้น Grade 8 (Gubes & Uyar, 2020) คือ Tucker Linear Equating Model, Presmoothed Equipercentile Model, Levine Linear Equating Model และ Simplified Circle Arc Equating Model ผลการศึกษาพบว่า ไม่ว่าข้อมูลดิบจะมีความแตกต่างของการกระจายของคะแนน (Differential Item Function : DIF) หรือไม่ก็ตาม ผลลัพธ์ของการเทียบคะแนนแต่ละรูปแบบจะคล้ายคลึงกัน และมีค่าความคลาดเคลื่อนมาตรฐานในการเทียบคะแนน (Standard Error of Equating : SEE) น้อย แต่เมื่อพิจารณาอคติ (Bias) และค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Square Error : RMSE) ประกอบ พบว่ารูปแบบการเทียบคะแนนที่ถูกต้องมากที่สุด คือ Presmoothed Equipercentile Model

ต่อมา มีผลงานวิจัยล่าสุดที่เกิดขึ้นในปี 2021 เกี่ยวกับการเทียบคะแนนที่ปรากฏ (Observed scores) โดย Traditional Equipercentile Method ที่ใช้คะแนนดิบจากการทดสอบโดยตรง กับแบบใหม่ที่เปลี่ยนคะแนนดิบที่แยกประเภท (discrete) ให้เป็นคะแนนที่มีความต่อเนื่อง (continuous) แล้วทำให้คะแนนที่ปรากฏ (Observed scores) ของแบบทดสอบเกณฑ์ และคะแนนของแบบทดสอบพยากรณ์มีการกระจายเป็นโค้งปกติ ก่อนนำคะแนนที่อยู่ในตำแหน่ง Percentile ที่เท่ากันมาเทียบกัน หรือที่เรียกว่า Kernel Method การวิจัยครั้งนี้พบว่า การเทียบคะแนนโดย Kernel Method มีความถูกต้องมากกว่าวิธีดั้งเดิม เพราะว่ามีค่าความคลาดเคลื่อนมาตรฐานในการเทียบคะแนน (SEE) น้อยกว่า (Wiberg, 2021) ซึ่งขัดแย้งกับผลการวิจัยของ von Davier และคณะที่ทำได้ในปี 2006 (von Davier et al., 2006) ที่พบว่า การเทียบคะแนนโดย Traditional Equipercentile Method และแบบ Kernel Method ได้ผลดีเช่นเดียวกัน เมื่อมีจำนวนผู้สอบแบบทดสอบทั้ง 2 ชุดเท่ากัน หรือแตกต่างกันก็ตาม และมีแบบทดสอบร่วมภายนอก (External Anchor Test) หรือภายใน (Internal Anchor Test) รวมทั้งยังขัดแย้งกับผลการวิจัยของนักวิจัยหลายท่านที่ทำการศึกษามาก่อนแล้วที่ต่างก็พบว่า การเทียบคะแนนโดยรูปแบบทั้งสองดังกล่าวแล้วได้ผลลัพธ์ไม่แตกต่างกัน (Liu & Low, 2008; Grant et al., 2009; Arıkan & Gelbal, 2018)

จากการทบทวนวรรณกรรมที่เกี่ยวข้อง สามารถสรุปได้ว่า

1. การเทียบคะแนนที่ปรากฏ (Observed scores) ของแบบทดสอบต่างชุดที่ทดสอบความตรงเชิงเนื้อหา (Content Validity) ที่เหมือนกันและที่มีการกระจายของคะแนนในลักษณะคล้ายคลึงกันโดยวิธี Traditional/Conventional Equipercentile Method ได้ผลดีไม่ต่างจากวิธีอื่นที่ซับซ้อนมากกว่า เช่น IRT, Presmoothed Equipercentile หรือ Post-smoothed Equipercentile Methods, Kernel Method เป็นต้น
2. การเทียบคะแนนที่ปรากฏ (Observed scores) ของแบบทดสอบต่างชุดที่ทดสอบความตรงเชิงเนื้อหา (Content Validity) ที่เหมือนกัน และมีการกระจายของคะแนนในลักษณะคล้ายคลึงกันโดย Linear Regression Method ได้ผลดีไม่ต่างจากวิธีอื่น เช่น Traditional/Conventional Equipercentile Method และ Linear Method เป็นต้น
3. การเทียบคะแนนโดยใช้คะแนนที่ปรากฏ (Observed scores) ของแบบทดสอบที่ต้องการนำคะแนนมาเทียบกัน มีรูปแบบการกระจายที่เหมือนกัน ก็ไม่จำเป็นที่จะต้องปรับการกระจายของคะแนนให้เป็นโค้งปกติก่อน หรือหลังการเทียบคะแนนแล้วก็ได้ผลลัพธ์ที่คล้ายคลึงกัน แต่ในกรณีที่การกระจายของคะแนนของแบบทดสอบทั้งคู่แตกต่างกัน การปรับคะแนน (smoothing) ยังมีความจำเป็นที่จะต้องทำ และถ้ากลุ่มตัวอย่างมีขนาดเล็ก ควรจะทำการปรับลักษณะโค้งของการกระจายของคะแนนก่อนการเทียบคะแนน (Presmoothing) โดย Equipercentile Method แต่ถ้ากลุ่มตัวอย่างมีขนาดใหญ่ไม่ควรทำการ Presmoothing เพราะจะทำให้เกิดความคลาดเคลื่อนมาก และควรจะทำทดสอบการกระจายของคะแนนสอบเสียก่อน เพื่อความแน่ใจว่าการกระจายของคะแนนของแบบทดสอบที่จะนำคะแนนมาเทียบกันมีการกระจายในลักษณะที่ไม่แตกต่างกัน
4. การเทียบคะแนนโดย Equipercentile Method เป็นวิธีที่ได้รับความนิยมมาก เพราะเป็นวิธีที่ไม่ซับซ้อน และได้ผลลัพธ์คล้ายกับหลายวิธีที่มีความยุ่งยากและซับซ้อนกว่า เช่น IRT, Linear, และ Linear Regression Methods แต่ก็มีปัญหาในการนำไปใช้เพราะอาจได้ค่าคะแนนบางค่าที่ซ้ำกัน

3. วัตถุประสงค์ของการวิจัย

3.1 เพื่อจัดสร้างแบบทดสอบวัดระดับทักษะภาษาอังกฤษด้านการฟังและการอ่านชื่อ EELT สำหรับนักเรียนระดับมัธยมศึกษาชั้นปีที่ 1-6 ให้เป็นแบบทดสอบมาตรฐานในระดับสากลจำนวน 6 ชุดที่คู่ขนานจริง (True Parallel Forms) กับแบบทดสอบชุดเดิมคือ ELTiS

3.2 เพื่อเทียบคะแนนของแบบทดสอบ EELT ที่สร้างขึ้นใหม่กับคะแนนของแบบทดสอบ ELTiS ด้วย 2 วิธี คือ Equipercentile Method และ Linear Regression Method

4. สมมติฐานการวิจัย

4.1 คะแนนสอบของ EELT ชุดใหม่แต่ละชุดกับคะแนนสอบของ ELTiS ชุดเดิมมีความสัมพันธ์กันอย่างมีนัยสำคัญที่ $p = .05$

4.2 ค่าเฉลี่ยของคะแนนสอบของ EELT ชุดใหม่แต่ละชุดกับคะแนนสอบของ ELTiS ชุดเดิมต่างกันอย่างไม่มีนัยสำคัญที่ $p = .05$

4.3 ค่าความแปรปรวนของคะแนนสอบของ EELT ชุดใหม่แต่ละชุดกับของคะแนนสอบของ ELTiS ชุดเดิมต่างกันอย่างไม่มีนัยสำคัญที่ $p = .05$

5. คำถามการวิจัย

5.1 แบบทดสอบ EELT ใหม่ทั้ง 6 ชุด เป็นแบบทดสอบคู่ขนานจริง (True Parallel Forms) กับแบบทดสอบชุดเดิม คือ ELTiS หรือไม่

5.2 แบบทดสอบ EELT ใหม่ทั้ง 6 ชุด เป็นมีคุณภาพทางด้านความตรงเชิงเนื้อหา (Content Validity) และมีความเที่ยง (Reliability) สูงในระดับที่น่าพอใจหรือไม่

5.3 ตารางเทียบคะแนนระหว่างแบบทดสอบ EELT ชุดใหม่กับ ELTiS จำเป็นต้องมีกี่ตาราง

6. ขอบเขตการวิจัย

6.1 สร้างแบบทดสอบ EELT ใหม่จำนวน 6 ชุดให้มีลักษณะคู่ขนานจริง (True parallel forms) กับแบบทดสอบ ELTIS เดิม ด้านเนื้อหา ความยาก-ง่ายโดยเฉลี่ยของแบบทดสอบ และค่าความแปรปรวนของคะแนน (Variance)

6.2 พหุวิจัย (Subjects) ได้แก่ นักเรียนมัธยมศึกษาชั้นปีที่ 1 ถึง 6 สอบแบบทดสอบ ELTIS และ EELT ชุดละ 300 คน มาจากระดับชั้นละ 50 คนเท่ากัน และมีความสามารถทางภาษาอังกฤษหลากหลาย (Heterogeneous groups)

6.3 ทำการเทียบคะแนนของแบบทดสอบโดยใช้คะแนนที่ปรากฏ (Observed scores) 2 วิธีเท่านั้น คือ Equipercentile Method และ Linear Regression Method

6.4 ตารางเทียบคะแนนระหว่างแบบทดสอบ ELTIS ชุดเดิมกับแบบทดสอบ EELT ชุดใหม่อาจมีเพียงชุดเดียว หรือหลายชุด ทั้งนี้ขึ้นอยู่กับผลของการเทียบคะแนนแต่ละชุด และแต่ละวิธีว่ามีลักษณะแตกต่างกัน หรือคล้ายคลึงกันมากน้อยเพียงใด รวมทั้งความสะดวกในการนำไปใช้

7. ข้อตกลงเบื้องต้น

7.1 แบบทดสอบ ELTIS ชุดเดิมเป็นแบบทดสอบมาตรฐานที่ความตรง (Validity) และความเที่ยง (Reliability) สูงในระดับที่น่าพอใจ

7.2 นักเรียนทุกคนที่เข้ารับการสอบแบบทดสอบ EELT ชุดใหม่แต่ละชุดและ ELTIS มีความตั้งใจจริงในการสอบ เนื่องจากคะแนนผลการสอบใช้เพื่อวัดสมิทธิภาพทางภาษาอังกฤษ (English Proficiency) ผู้สอบแต่ละคนที่สมัครใจมาสอบเอง

7.3 สิ่งแวดล้อมและบรรยากาศในการสอบแบบทดสอบ ELTIS ชุดเดิมกับแบบทดสอบ EELT ชุดใหม่ ของผู้สอบในสถานที่แตกต่างกันมีผลต่อคะแนนการสอบโดยเฉลี่ยแตกต่างกันอย่างไม่มีนัยสำคัญ เนื่องจากแบบทดสอบแต่ละชุดมีเพียง 50 ข้อและใช้เวลาสอบไม่นาน

8. ข้อจำกัดของการวิจัย

8.1 ตัวแปรทางด้านกายภาพและจิตวิทยาบางอย่างของนักเรียนที่เข้ารับการทดสอบแบบทดสอบ ELTIS ชุดเดิมกับแบบทดสอบ EELT ชุดใหม่ที่อาจมีผลต่อคะแนนการสอบไม่อาจควบคุมได้ เช่น เพศสภาพ ความเมื่อยล้า ความน่าเบื่อหน่าย สิ่งแวดล้อมและบรรยากาศในการสอบ เป็นต้น

8.2 สิ่งแวดล้อมและบรรยากาศในการสอบแบบทดสอบของบางสถานที่สอบในบางขณะเวลาอาจอยู่นอกเหนือการควบคุมได้ของผู้ดำเนินการสอบ เช่น มีเสียงข้างนอกรบกวนสมาธิผู้สอบ หรือกระแสไฟฟ้าตกหรือดับ เป็นต้น

8.3 เนื่องจากการวิจัยนี้ดำเนินการเก็บข้อมูลในปี พ.ศ. 2565 ซึ่งอยู่ในระหว่างการระบาดของโรคโควิด-19 ไม่สามารถจัดสอบให้กับนักเรียนเป็นกลุ่มใหญ่ได้ จึงต้องสอบเป็นกลุ่มย่อยจำนวนหลายครั้ง

9. นิยามศัพท์เชิงปฏิบัติการ

9.1 แบบทดสอบวัดสมิทธิภาพทั่วไปทางภาษาอังกฤษ (General English proficiency test) หมายถึง แบบทดสอบที่มุ่งวัดความสามารถทางภาษาอังกฤษด้านทักษะ Listening Comprehension จำนวน 5 ด้าน และด้านทักษะ Reading Comprehension อีกจำนวน 3 ด้าน อย่างเช่น แบบทดสอบ ELTIS ชุดเดิม และ EELT ที่ผู้วิจัยสร้างขึ้นใช้ เนื้อหาต่าง ๆ ของแบบทดสอบดังกล่าวนี้ไม่ได้เป็นส่วนหนึ่งของหลักสูตรในการเรียนการสอนอย่างเป็นทางการที่ผู้สอบเคยเรียนมาแล้วโดยเฉพาะจากหลักสูตรใดหลักสูตรหนึ่ง หากแต่เป็นเนื้อหาที่ผู้สร้างแบบทดสอบรวบรวมขึ้นเพราะมีความเห็นว่าตรงตามทฤษฎีและแนวปฏิบัติตามเกณฑ์ของ CEFR

9.2 เกณฑ์วัดระดับความสามารถทางภาษาของ CEFR หมายถึง เกณฑ์ที่สภาแห่งยุโรป (Council of Europe) กำหนดขึ้นในปี พ.ศ. 2544 โดยแบ่งระดับความสามารถทางภาษาออกเป็น 6 ระดับ คือ A1, A2, B1, B2, C1 และ C2 ตามรายละเอียดในหนังสือชื่อ Common European Framework of Reference for Languages : Learning, Teaching, Assessment (Council of Europe, 2020)

9.3 การเทียบคะแนน (Score equating) คือ กระบวนการทางสถิติที่ใช้การปรับคะแนนของแบบทดสอบเพื่อให้คะแนนของแบบทดสอบชุดหนึ่งสามารถใช้แทนคะแนนของแบบทดสอบอีกชุดหนึ่งได้ การปรับคะแนนเป็นการปรับระดับความยากง่ายของแบบทดสอบ ไม่ใช้การปรับเนื้อหาของแบบทดสอบ (Kolen & Brennan, 2004)

9.4 ตารางเทียบคะแนน (Equating score table) หมายถึง ตารางที่ใช้เทียบคะแนนของแบบทดสอบ EELT ชุดใหม่กับแบบทดสอบ ELTiS ชุดเดิมโดยประมาณ เพราะว่าการเทียบคะแนนโดยวิธีใด ๆ ก็ตามมักจะมีค่าความคลาดเคลื่อนมาตรฐานในการประมาณค่า (Standard Error of Estimate) จากการคำนวณด้วยเสมอ ซึ่งขนาดของความคลาดเคลื่อนนี้อาจเกิดจากปัจจัยหลายอย่างที่อยู่นอกเหนือการควบคุมของผู้วิจัยได้ เช่น ความไม่เท่าเทียมกันอย่างแท้จริงของแบบทดสอบชุดใหม่กับแบบทดสอบชุดเดิม หรือกลุ่มตัวอย่างในการทดสอบมีระดับความสามารถแตกต่างกัน เป็นต้น

9.5 แบบทดสอบ ELTiS ชุดเดิม หมายถึง แบบทดสอบมาตรฐานตามเกณฑ์ของ CEFR ที่ผู้วิจัยมีอยู่แล้ว และเป็นแบบทดสอบที่มีคุณภาพสูงทางด้านค่าความตรงและค่าความเที่ยง เป็นต้น

9.6 แบบทดสอบ EELT ชุดใหม่ หมายถึง แบบทดสอบ EELT จำนวน 6 ชุดที่สร้างขึ้นตามโครงสร้าง และตารางลักษณะเฉพาะของข้อทดสอบ (Table of test item specifications) ของแบบทดสอบ ELTiS ชุดเดิม โดยคณะกรรมการผู้เชี่ยวชาญที่ผู้วิจัยเชิญมาจากสถาบันการศึกษาต่าง ๆ จำนวน 5 ท่าน

9.7 แบบทดสอบคู่ขนานจริง (True parallel forms) หมายถึง แบบทดสอบ 2 ชุดที่มีความตรงเชิงเนื้อหาเหมือนกัน มีคะแนนค่าเฉลี่ยหรือค่าความยาก-ง่าย และค่าความแปรปรวนของคะแนนต่างกันอย่างไม่มีนัยสำคัญ

9.8 แบบทดสอบคู่ขนานเทียม (Pseudo parallel forms) หมายถึง แบบทดสอบ 2 ชุดที่มีความตรงเชิงเนื้อหาเหมือนกัน แต่มีคะแนนค่าเฉลี่ยหรือค่าความยาก-ง่าย หรือค่าความแปรปรวนของคะแนนต่างกันอย่างไม่มีนัยสำคัญ

10. ประโยชน์ที่คาดว่าจะได้รับ

10.1 โครงการฯ ได้แบบทดสอบมาตรฐานที่สามารถนำไปใช้ทดสอบวัดความรู้และความสามารถทางภาษาอังกฤษของนักเรียนในระดับชั้นมัธยมศึกษาเพื่อนำผลคะแนนจากการสอบมาเป็นเกณฑ์ประกอบการพิจารณาให้ทุนการศึกษาของโครงการแลกเปลี่ยนภาษาและวัฒนธรรมนานาชาติได้อย่างเหมาะสม

10.2 โครงการฯ ได้แบบทดสอบมาตรฐานสำหรับโครงการแลกเปลี่ยนภาษาและวัฒนธรรมนานาชาติเอ็นจีเนียส อินเตอร์เนชันแนลเพิ่มขึ้น และแก้ปัญหาเรื่องผลกระทบนำพา (Carry-over Effects) ได้

10.3 ทำให้การทดสอบเกิดความเป็นธรรม (Fairness) สำหรับผู้สอบที่ต้องสอบแบบทดสอบมาตรฐานหลายครั้ง เพราะว่าแบบทดสอบหลายชุดสามารถเทียบคะแนนผลการสอบกันได้ เช่น เนื้อหาที่สอบ ความยาก และความเที่ยงของแบบทดสอบ เป็นต้น (Baghaei, 2010)

10.4 ทำให้การทดสอบของโครงการฯ มีความน่าเชื่อถือ (Credibility) มากขึ้น เพราะมีแบบทดสอบที่จำเป็นต้องใช้หลายชุดสำหรับนำผลการสอบมาใช้ประกอบการตัดสินใจบางประการ เช่น การให้ทุนการศึกษาของโครงการแลกเปลี่ยนภาษาและวัฒนธรรมนานาชาติ หรือเพื่อการศึกษาต่อ เป็นต้น (สุพัฒน์ สุขมลสันต์, 2560; Albano, 2010)

10.5 ทำให้แบบทดสอบมีความปลอดภัย (Security) เพราะว่าแบบทดสอบมาตรฐานที่จำเป็นต้องใช้หลาย ๆ ครั้ง จำเป็นจะต้องมีหลาย ๆ ชุดที่มีลักษณะคู่ขนานกันทางเนื้อหา และความยากง่าย (Dorans 2004; Albano, 2010)

10.6 ช่วยหลีกเลี่ยงการใช้แบบทดสอบชุดเดิมซ้ำมากเกินไป (Overexposure) ซึ่งจะเป็นผลเสียต่อผลการสอบเพราะขาดความน่าเชื่อถือในผลการสอบอันเกิดจากผลของการฝึกฝนการสอบ (Practice Effects) หรือผลกระทบนำพา (Carry-over Effects) เพราะว่าใช้แบบทดสอบเดิมหลายครั้งกับผู้สอบเดิม (Albano, 2010)

10.7 โครงการฯ สามารถใช้คะแนนที่เทียบกันได้ (Comparability) จากการสอบต่างเวลา หรือต่างกลุ่มผู้สอบสำหรับใช้ประกอบการตัดสินใจเกี่ยวกับอนาคตของผู้สอบซึ่งจะได้รับผลกระทบจากการตัดสินใจมาก (Shin, 2015)

10.8 โครงการฯ ได้รับประสบการณ์ตรงในการสร้างแบบทดสอบมาตรฐานและการเทียบคะแนนของแบบทดสอบชุดใหม่กับแบบทดสอบมาตรฐาน

11. วิธีการวิจัย

11.1 ประชากรและกลุ่มตัวอย่าง

11.1.1 ประชากร (Population) ได้แก่ นักเรียนชั้นมัธยมศึกษาปีที่ 1 ถึง 6 ทั่วประเทศในปีการศึกษา พ.ศ. 2564 จำนวน 625,487 คน 619,794 คน 602,565 คน 391,167 คน 362,530 คน และ 333,632 คน ตามลำดับ (กองทุนเพื่อความเสมอภาคทางการศึกษา, 2565)

11.1.2 พลวิทย์หรือกลุ่มตัวอย่าง (Samples) ได้แก่ นักเรียนชั้นมัธยมปีที่ 1-6 ชั้นละ 50 คน รวมเป็น 300 คน เพื่อสอบแบบทดสอบ EELT ชุดใหม่แต่ละชุด (Form) และแบบทดสอบ ELTiS ชุดเดิม ดังนั้นกลุ่มตัวอย่างทั้งหมดจึงมีจำนวน 1800 คน ตัวอย่างเหล่านี้ได้มาจากการสุ่มเลือกนักเรียนของแต่ละระดับชั้นที่สอบแบบทดสอบทั้ง 2 ชุดครบถ้วนด้วยวิธีสุ่มอย่างง่าย (Simple Random Sampling Technique)

11.2 เครื่องมือวิจัย การสร้าง และการหาคุณภาพ

11.2.1 เครื่องมือวิจัยได้แก่แบบทดสอบ ELTiS ชุดเดิมที่มีอยู่แล้ว ซึ่งเป็นแบบทดสอบมาตรฐาน และแบบทดสอบ EELT ชุดใหม่ 6 ชุดที่สร้างขึ้น ก่อนที่จะสร้างแบบทดสอบ EELT โครงการฯ ศึกษาลักษณะเฉพาะของแบบทดสอบ ELTiS ซึ่งมีดังนี้

1) ทักษะที่มุ่งทดสอบ

แบบทดสอบ ELTiS เน้นวัดทักษะด้าน การฟัง การอ่าน โครงสร้างประโยค และ คำศัพท์ และเหมาะสมสำหรับนักเรียนอายุ 13-18 ปี (ระดับชั้นมัธยมศึกษาปีที่ 1-6) พัฒนาขึ้นโดย บริษัท Ballard & Tighe ผู้จัดทำหนังสือแบบเรียนภาษาอังกฤษมากกว่า 40 ปี ก่อตั้งโดย Wanda Ballard และ Phyllis Tighe ผู้มีความมุ่งมั่นที่จะพัฒนาแบบเรียนภาษาอังกฤษให้มีประสิทธิภาพสูงสุดในการพัฒนาศักยภาพทางภาษาอังกฤษของผู้เรียน จึงร่วมมือกับ CSJET (The Council on Standards for International Educational Travel) ซึ่งเป็นหน่วยงานที่ดูแลและควบคุมมาตรฐานของโครงการนักเรียนแลกเปลี่ยนเพื่อพัฒนาแบบทดสอบนี้ขึ้น

2) รูปแบบของแบบทดสอบ ELTiS

- แบบทดสอบ ELTiS เป็นข้อสอบปรนัย จำนวน 50 ข้อ แบ่งเป็น 2 ส่วน คือ Listening Comprehension จำนวน 24 ข้อ และ Reading Comprehension อีก 26 ข้อ

- ใช้เวลาในการทดสอบ 1 ชั่วโมง 10 นาที โดยแบ่งเป็น Listening Comprehension 25 นาที และ Reading Comprehension อีก 45 นาที

- โครงสร้างของแบบทดสอบ ELTiS ประกอบด้วย 2 ส่วนหลัก ดังนี้

ส่วนที่ 1 Listening Comprehension แบ่งเป็น 5 ด้าน ได้แก่

1. คำสั่งในชั้นเรียนของครู
2. ภาษาทางคณิตศาสตร์
3. บทสนทนาขนาดสั้น
4. บทสนทนาขนาดยาวในห้องเรียน
5. เนื้อหาในบทเรียนวิชาพื้นต่าง ๆ

ส่วนที่ 2 Reading Comprehension แบ่งเป็น 3 ด้าน ได้แก่

1. คำศัพท์
2. หลักไวยากรณ์
3. บทความจากเนื้อหาวิชาต่าง ๆ เช่น วิชาภาษาอังกฤษ วิชาวิทยาศาสตร์ และสังคมศาสตร์

11.2.2 โครงการฯ เชิญและแต่งตั้งคณะกรรมการผู้เชี่ยวชาญด้านการเรียนการสอน และการวัดและประเมินผล ภาษาอังกฤษจำนวน 5 ท่านจากสถาบันการศึกษาและมหาวิทยาลัยบางแห่ง เพื่อสร้างแบบทดสอบ EELT ชุดใหม่ 6 ชุด บุคคลเหล่านี้มีวุฒิการศึกษาที่เกี่ยวข้องในระดับปริญญาโท 3 ท่านและปริญญาเอก 3 ท่าน และมีประสบการณ์ในการสอนโดยเฉลี่ยคนละประมาณ 10 ปี และมีอาจารย์ชาวไทย (วุฒิการศึกษาระดับปริญญาเอก) เป็นผู้ตรวจสอบข้อสอบและพิสูจน์อักษร 1 คน และอาจารย์ชาวอเมริกัน (วุฒิการศึกษาระดับปริญญาเอก) อีก 1 คนเป็นผู้ตรวจภาษา ทั้งสองท่านมีประสบการณ์ด้านการสอนภาษาอังกฤษและการออกข้อสอบมากกว่า 10 ปี

11.2.3 คณะกรรมการฯ จัดทำตารางลักษณะเฉพาะของข้อสอบ (Test item specification table) ตามลักษณะของแบบทดสอบ ELTiS และสร้างแบบทดสอบ EELT ชุดใหม่ จำนวน 6 ชุด ให้มีลักษณะคู่ขนานกับแบบทดสอบ ELTiS ชุดเดิมในด้านเนื้อหาที่มุ่งทดสอบ

11.2.4 หลังจากคณะกรรมการฯ สร้างแบบทดสอบ EELT จำนวน 6 ชุดเสร็จแล้ว โครงการฯ เชิญผู้ทรงคุณวุฒิ จำนวน 3 ท่าน ที่มีความเชี่ยวชาญด้านการสอนภาษาอังกฤษและการประเมินผลดำเนินการหาความสอดคล้องของเนื้อหา กับ วัตถุประสงค์ของการทดสอบ (Item-Objective Congruence : IOC) ซึ่งเป็นปัจจัยหนึ่งที่สำคัญมากของค่าความตรงเชิงเนื้อหา (Content Validity) โดยใช้สูตรของ Rovinelli and Hambleton (1977) ผลปรากฏว่า แบบทดสอบ EELT ชุดใหม่ทั้ง 6 ชุด มีค่า IOC อยู่ในช่วง 0.50-0.79 ดังนั้นคณะกรรมการผู้สร้างแบบทดสอบจึงได้ปรับแก้ไขข้อสอบตามคำแนะนำของผู้ทรงคุณวุฒิ หลังจากนั้นจึงให้ผู้ทรงคุณวุฒิทั้ง 3 ท่านตรวจสอบความตรงเชิงเนื้อหาของแบบทดสอบทั้ง 6 ชุดอีกครั้ง และค่า IOC ของแบบทดสอบ ทั้ง 6 ชุด อยู่ในช่วง 0.90–0.95 ซึ่งเป็นค่าที่สูงมาก (Turner & Carlson, 2003) แสดงว่า แบบทดสอบทั้ง 6 ชุดสามารถวัดเนื้อหา เช่นเดียวกับแบบทดสอบ ELTiS ชุดเดิมได้

12. การเก็บรวบรวมข้อมูล

12.1 โครงการฯ ประกาศเชิญชวนให้นักเรียนชั้นมัธยมศึกษาชั้นปีที่ 1–6 ของโรงเรียนต่าง ๆ ในกรุงเทพมหานคร ที่สนใจ จะทดสอบวัดระดับสมรรถภาพทางภาษาอังกฤษของตนสมัครเข้ารับการทดสอบได้โดยไม่เสียค่าใช้จ่ายผ่านระบบ online

12.2 โครงการฯ ดำเนินการจัดสอบแก่ผู้สมัครสอบที่โรงเรียนของนักเรียนเอง ด้วยรูปแบบให้ผู้สอบกลุ่มเดียวและสลับ แบบทดสอบ (Single Group Design with Counterbalancing) (Kolen & Brennan, 2004; Yu & Osborn Popp, 2005) กล่าวคือ แต่ละกลุ่มมีการสอบใน 2 ลักษณะ คือ 1) จัดให้มีการสอบแบบทดสอบ ELTiS ชุดเดิมแล้วจึงสอบแบบทดสอบ EELT ชุดใหม่ 1 ชุด และมีการสอบแบบทดสอบ EELT ชุดใหม่ 1 ชุด แล้วจึงสอบแบบทดสอบ ELTiS ชุดเดิม ทั้งนี้เพื่อแก้ปัญหาที่อาจ เกิดจากผลกระทบของลำดับที่ของการสอบ (Order Effects)

12.3 ดำเนินการสุ่มเลือกแบบง่ายในการเลือกผู้สอบแต่ละชั้นปีที่สอบแบบทดสอบทั้งสองครบถ้วนสำหรับแบบทดสอบ EELT ชุดใหม่แต่ละชุดมาชั้นละ 50 คน รวมเป็น 300 คนต่อชุด หรือทั้งหมด 1800 คน

12.4 รวบรวมคะแนนผลการสอบของผู้สอบแต่ละคน ในการสอบแบบทดสอบแต่ละชุดเพื่อการวิเคราะห์ต่อไป

13. การวิเคราะห์ข้อมูล

13.1 วิเคราะห์หาค่าความตรงเชิงเนื้อหา (Content Validity) ของแบบทดสอบ EELT ชุดใหม่แต่ละชุดกับของ ELTiS โดยใช้สูตรของ Rovinelli and Hambleton (1977)

13.2 ทดสอบความสัมพันธ์ของคะแนนสอบของ EELT ชุดใหม่แต่ละชุดกับคะแนนสอบของ ELTiS ชุดเดิมว่ามีมากน้อย เพียงใดด้วย Pearson Correlation

13.3 ทดสอบความแตกต่างของค่าเฉลี่ยของคะแนนสอบของ EELT ชุดใหม่แต่ละชุดกับคะแนนสอบของ ELTiS ชุดเดิม ว่ามีนัยสำคัญหรือไม่ด้วย Independent Samples t-test

13.4 ทดสอบความแตกต่างของค่าความแปรปรวนของคะแนนสอบของ EELT ชุดใหม่แต่ละชุดกับของคะแนนสอบ ELTiS ชุดเดิมว่ามีนัยสำคัญหรือไม่ด้วย Levine's test

13.5 คำนวณหาค่าความเที่ยง (Reliability) ของ EELT ชุดใหม่แต่ละชุด โดยวิธี KR₂₁

13.6 เปรียบคะแนนของแบบสอบ EELT ชุดใหม่แต่ละชุดกับคะแนนสอบของ ELTiS ชุดเดิมด้วย Equipercentile Method และ Linear Regression Method แล้วจัดทำตารางเทียบรวมของแต่ละวิธี

14. ผลการวิจัย

14.1 ความสัมพันธ์ของคะแนนสอบของ EELT แต่ละชุดกับคะแนนสอบของ ELTiS

ตารางที่ 1 ความสัมพันธ์ของคะแนนสอบของ EELT แต่ละชุดกับคะแนนสอบ ELTiS

Test Forms	ELTiS	R ² x 100
EELT_01	0.971*	94.28%
EELT_02	0.990*	98.01%
EELT_03	0.982*	96.43%
EELT_04	0.943*	88.92%
EELT_05	0.956*	91.39%
EELT_06	0.946*	89.46%

* p < 0.05

ตารางที่ 1 แสดงให้เห็นว่าคะแนนสอบของ EELT แต่ละชุดกับคะแนนสอบ ELTiS มีค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficients : r_{xy} หรือ R) ระหว่าง 0.946–0.990 ซึ่งสูงมากจนเกือบอยู่ในระดับสมบูรณ์ (Hopkins, 2002) ซึ่งแสดงว่าผลการสอบของแบบทดสอบทั้ง 2 ชุด มีความสอดคล้องกันสูงมาก และจากค่า R² x 100 จะเห็นได้ว่าแบบทดสอบแต่ละคู่สามารถวัดเนื้อหาที่สอดคล้องกันได้ประมาณร้อยละ 89–98 ซึ่งนับว่าสูงมาก หรือกล่าวอีกนัยหนึ่งก็คือ แบบทดสอบ EELT แต่ละชุดมีความตรง (Validity) กับแบบทดสอบ ELTiS สูงมาก

14.2 การทดสอบความแตกต่างของค่าเฉลี่ยของคะแนนสอบของ EELT แต่ละชุดกับคะแนนสอบของ ELTiS

ตารางที่ 2 การทดสอบความแตกต่างของค่า mean และ variance ของคะแนน EELT แต่ละชุดกับของคะแนน ELTiS และค่าความเที่ยงของแบบทดสอบ EELT แต่ละชุด

Test Forms	df	Mean	SD	t-test		Levine's test		KR ₂₁
				t-value	Sig	F-value	Sig	
EELT_01	598	33.61	11.147	0.467	0.641	0.005	0.942	0.930
ELTiS		33.19	11.228					
EELT_02	598	31.50	11.113	1.532	0.126	1.830	0.177	0.924
ELTiS		32.84	10.358					
EELT_03	598	34.09	10.771	1.814	0.070	6.248	0.013*	0.925
ELTiS		35.60	9.637					
EELT_04	598	37.85	9.285	1.185	0.236	0.156	0.693	0.912
ELTiS		36.97	8.964					
EELT_05	598	37.97	9.509	2.420	0.016*	0.056	0.813	0.917
ELTiS		36.08	9.686					
EELT_06	598	37.29	9.855	2.427	0.015*	0.001	0.999	0.921
ELTiS		35.34	9.824					

* p < 0.05

ตารางที่ 2 แสดงให้เห็นว่าค่าเฉลี่ยและค่าความแปรปรวน (variance) ของคะแนนสอบ EELT ชุดที่ 1, 2 และ 4 แตกต่างจากค่าดังกล่าวของแบบทดสอบ ELTiS อย่างมีนัยสำคัญ (p = .05) แต่ค่าเฉลี่ยของคะแนนสอบ EELT ชุดที่ 5 และ 6 รวมทั้งค่าความแปรปรวนของ EELT ชุดที่ 3 แตกต่างจากค่าดังกล่าวของแบบทดสอบ ELTiS อย่างมีนัยสำคัญ (p = .05)

ดังนั้น จากผลการวิจัยในตารางที่ 1 และ 2 จึงสามารถสรุปได้ว่าแบบทดสอบ EELT ชุดที่ 1, 2 และ 4 เป็นแบบทดสอบคู่ขนานจริง (True Parallel Forms) กับแบบทดสอบ ELTiS ส่วนแบบทดสอบ EELT ชุดที่ 3, 5 และ 6 เป็นแบบทดสอบคู่ขนานเทียม (Pseudo Parallel Forms) กับแบบทดสอบ ELTiS

14.3 การเทียบคะแนนของแบบสอบ EELT ชุดใหม่โดยรวมกับคะแนนสอบของ ELTiS ชุดเดิมจาก Equipercentile Method

ตารางที่ 3 ตารางเทียบคะแนนของแบบสอบ EELT ชุดใหม่โดยรวมกับคะแนนสอบของ ELTiS ชุดเดิมจาก Equipercentile Method

EELT	ELTiS	EELT	ELTiS	EELT	ELTiS	EELT	ELTiS
11	11	29	29	38	37	44	44
12	12	29	29	38	38	44	44
14	14	30	29	38	38	45	44
15	15	30	29	39	38	45	44
15	16	30	30	39	39	45	45
16	17	31	30	39	39	45	45
17	18	31	30	40	39	45	45
18	19	32	31	40	39	46	45
19	19	32	32	40	39	46	45
20	20	33	32	40	39	46	45
20	21	33	33	40	40	46	45
21	21	33	33	41	40	46	46
22	22	34	33	41	40	46	46
22	23	34	33	41	40	47	46
23	23	34	34	42	41	47	47
24	24	35	34	42	41	48	47
21	25	35	34	42	41	48	47
25	25	35	35	42	42	48	47
26	26	36	35	42	42	48	48
26	26	36	36	43	42	48	48
26	26	36	36	43	42	49	48
27	27	37	36	43	43	49	49
28	27	37	36	44	43	49	49
28	28	37	36	44	43	50	49
29	29	38	37	44	43	---	---

ตารางที่ 3 เกิดจากค่าเฉลี่ยของคะแนนเทียบกันระหว่างแบบสอบ EELT ชุดใหม่แต่ละชุดกับคะแนนสอบของ ELTiS โดย Equipercentile Method เนื่องจากคะแนนแต่ละค่าที่อยู่ใน Percentile Rank ที่เท่ากันมีค่าใกล้เคียงกันมาก ซึ่งเกิดจากความสัมพันธ์ของคะแนนของแบบสอบแต่ละคู่สูงมากดังแสดงให้เห็นแล้วในตารางที่ 1 เป็นที่น่าสังเกตว่า จำนวนคะแนนน้อยที่สุดในตารางไม่ใช่ 1 คะแนน หากแต่เป็น 11 คะแนนซึ่งเป็นคะแนนต่ำที่สุดของผู้สอบ และอยู่ที่ Percentile Rank ที่ 1 และคะแนนสูงที่สุดคือ 50 คะแนนจึงอยู่ที่ Percentile Rank ที่ 99 นอกจากนี้จะพบว่ามีความคล้ายคลึงกันของค่าที่เท่ากัน หรือซ้ำกัน

14.4 การเทียบคะแนนของแบบสอบ EELT ชุดใหม่โดยรวมกับคะแนนสอบของ ELTiS ชุดเดิมจาก Linear Regression Method

ตารางที่ 4 ตารางเทียบคะแนนของแบบสอบ EELT ชุดใหม่โดยรวม (EELT) กับคะแนนสอบของ ELTiS ชุดเดิมจาก Linear Regression Method

EELT	ELTiS	EELT	ELTiS
1	3	26	26
2	4	27	27
3	5	28	28
4	6	29	29
5	7	30	30
6	8	31	31
7	8	32	32
8	9	33	33
9	10	34	34
10	11	35	35
11	12	36	36
12	13	37	37
13	14	38	37
15	15	39	38
15	16	40	39
16	17	41	40
17	18	42	41
18	19	43	42
19	20	44	43
20	21	45	44
21	22	46	45
22	22	47	46
23	23	48	47
24	24	49	48
25	25	50	49

ตารางที่ 4 เกิดจากค่าเฉลี่ยของคะแนนเทียบกันระหว่างแบบสอบ EELT ชุดใหม่แต่ละชุดกับคะแนนสอบของ ELTiS โดย Linear Regression Method จากสมการพยากรณ์ซึ่งได้จาก Linear Regression Analysis ดังต่อไปนี้ คือ

$$y'_1 = 0.323 + 0.978X, y'_2 = 3.781 + 0.923X, y'_3 = 5.652 + 0.879X$$

$$y'_4 = 2.508 + 0.910X, y'_5 = -0.910 + 0.974X, y'_6 = 0.155 + 0.944X$$

เมื่อ $y'_1 \dots y'_6$ = คะแนนพยากรณ์ของ ELTiS เมื่อเทียบกับคะแนน EELT ชุดที่ 1-6 และ X = คะแนนของ EELT โดยสมการข้างต้นมาจาก $y' = a + bX$ เมื่อ a = ค่าคงที่ b = ค่าสัมประสิทธิ์การถดถอย และ X = ค่าตัวแปรต้น ซึ่งในที่นี้คือ คะแนน EELT แต่ละชุด

เนื่องจากคะแนนพยากรณ์ของ ELTiS เมื่อเทียบกับคะแนน EELT แต่ละชุดมีค่าใกล้เคียงกันมาก ซึ่งเกิดจากความสัมพันธ์ของคะแนนของแบบสอบแต่ละคู่สูงมากดังแสดงให้เห็นแล้วในตารางที่ 1 ดังนั้น เพื่อความสะดวกในการนำตารางเทียบคะแนนไปใช้ต่อไป ผู้วิจัยจึงนำค่าดังกล่าวมาหาค่าเฉลี่ยจึงได้ค่าตามที่ปรากฏอยู่ในตารางที่ 4

15. สรุปผลการวิจัย

15.1 แบบทดสอบ EELT ที่สร้างขึ้นใหม่ 6 ชุด เป็นแบบทดสอบคู่ขนานจริง (True Parallel Forms) กับแบบทดสอบ ELTIS จำนวน 3 ชุด คือ ชุดที่ 1, 2 และ 4 และเป็นแบบทดสอบคู่ขนานเทียม (Pseudo Parallel Forms) กับแบบทดสอบ ELTIS จำนวน 3 ชุด คือ ชุดที่ 3, 5 และ 6 แบบทดสอบใหม่ทั้ง 6 ชุดมีคุณภาพสูงในด้านความตรงเชิงเนื้อหา และค่าความเที่ยงในการวัดและประเมินผล

15.2 ตารางเทียบคะแนนระหว่าง EELT ที่สร้างขึ้นใหม่ 6 ชุด โดยรวมกับคะแนนแบบทดสอบ ELTIS มี 2 ตารางตามรายละเอียดในตารางที่ 3 และ 4

16. อภิปรายผล

ผลของการวิจัยครั้งนี้มีประเด็นที่น่าสนใจคือ ผลลัพธ์ที่แตกต่างกันในบางด้านระหว่างคะแนนในตารางเทียบจากแบบทดสอบชุดเดียวกัน ดังมีรายละเอียดตามตารางที่ 3 และ 4 ที่เป็นเช่นนี้อาจเป็นเพราะว่า

16.1 กลุ่มตัวอย่างในการวิจัยครั้งนี้สำหรับการสอบแบบทดสอบแต่ละคู่มีจำนวนน้อย คือ เพียง 300 คน และผู้สอบเหล่านี้ไม่มีพื้นความรู้ทางภาษาอังกฤษในทุกระดับความสามารถ โดยเฉพาะอย่างยิ่งในระดับต่ำมาก จึงทำให้คะแนนเทียบในตารางที่ 3 เริ่มที่ Percentile Rank ที่ 1 ที่คะแนน 11 แทนที่จะเป็นคะแนน 1 หรือ 0

16.2 การที่คะแนนในตารางที่ 3 ซึ่งเกิดจากการเทียบคะแนนโดย Equipercentile Method มีซ้ำกันจำนวนค่อนข้างมาก อาจเกิดจากพิสัย (range) ของคะแนนดิบของแบบทดสอบแต่ละคู่เทียบแคบ คือ 0-50 แต่ Percentile Rank มีพิสัยที่มากกว่าเป็น 2 เท่า คือ 1-99 นอกจากนี้ การวิเคราะห์มีการปิดคะแนนทศนิยมให้เป็นจำนวนเต็มด้วย จึงทำให้เกิดมีค่าคะแนนซ้ำกันหลายจำนวน ผลการวิจัยครั้งสอดคล้องกับงานวิจัยของ สัจญญา จันทรชูสกุล และคณะ (2560)

16.3 การเทียบคะแนนระหว่างแบบทดสอบมีหลากหลายวิธี แต่ยังมีวิธีใดที่ดีที่สุดที่จะใช้ได้ในทุกกรณี แต่ละวิธีมีค่าความคลาดเคลื่อนมาตรฐานในการเทียบคะแนน (Standard Error of Equating: SEE) ที่มากหรือน้อยแตกต่างกัน ยังไม่มีวิธีใดที่ให้ค่าที่แน่นอน (absolute value) ได้ และนักวิจัยบางคนยังไม่จัดให้ Linear Regression Method เป็นวิธีหนึ่งในการเทียบคะแนนตามคำจำกัดความของการเทียบคะแนนอย่างเคร่งครัด กล่าวคือ การเทียบคะแนนของแบบทดสอบ X และ Y คะแนนของแบบทดสอบทั้งสองสามารถแทนคะแนนของกันและกันได้ (Kolen & Brennan, 2004) แต่ว่า Linear Regression Method ทำให้คะแนนของ X แทนคะแนนของ Y ได้เพียงอย่างเดียวเท่านั้น

17. ข้อเสนอแนะ

17.1 ข้อเสนอแนะเพื่อการนำผลการวิจัยไปใช้

17.1.1 ควรจะใช้ค่าคะแนนในตารางที่ 4 เพื่อการเทียบคะแนน EELT ที่สร้างขึ้นใหม่ทั้ง 6 ชุดกับคะแนน ELTIS ไม่มีความจำเป็นต้องใช้ตารางเทียบคะแนนเฉพาะระหว่าง EELT แต่ละชุดกับ ELTIS ก็ได้ เนื่องจากค่าจากตารางดังกล่าวแตกต่างจากค่าในตารางรวมไม่มาก ทั้งนี้เพราะว่าค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficients) ระหว่างคะแนนของแบบทดสอบแต่ละคู่สูงมาก (ดูตารางที่ 1)

17.1.2 เนื่องจากการเทียบคะแนนระหว่างแบบทดสอบมีหลากหลายวิธี แต่ยังมีวิธีใดที่ดีที่สุดที่จะใช้ได้ในทุกกรณี แต่ละวิธีมีค่าความคลาดเคลื่อนมาตรฐานในการเทียบคะแนน (Standard Error of Equating : SEE) ที่มากหรือน้อยแตกต่างกัน ไม่ได้ให้ค่าที่แน่นอน (absolute value) ดังนั้น เมื่อใช้ค่าเทียบคะแนนตามตารางที่ 4 แล้วก็อาจลองตรวจสอบค่าคะแนนที่เท่ากันจากตารางที่ 3 ด้วยก็ได้เพื่อความแน่ใจ

17.2 ข้อเสนอแนะเพื่อการวิจัยต่อไป

17.2.1 เนื่องจากแบบทดสอบ EELT ที่สร้างขึ้นใหม่ทั้ง 6 ชุดยังไม่ได้ผ่านการทดลองใช้ (Pilot Study) มาก่อน จึงควรนำผลการทดสอบในครั้งนี้ของแบบทดสอบทุกชุดมาทำการวิเคราะห์ข้อทดสอบรายข้อ (Test Item Analysis) ทั้งแบบประเพณีนิยม (Traditional Models) และแบบทฤษฎีการตอบสนองของข้อทดสอบ (Item Response Theory Models : IRT) โดยเฉพาะอย่างยิ่งแบบทดสอบชุดที่ 3, 5 และ 6 เพื่อปรับปรุงแก้ไขคุณภาพของแบบทดสอบให้ดียิ่งขึ้นในด้าน คุณภาพของข้อทดสอบ และคุณภาพของตัวเลือก (choices) เช่น ค่าความยาก-ง่าย ค่าอำนาจจำแนก ค่าความตรงรายข้อ และค่าการเดาถูก

เป็นต้น รวมทั้งวิเคราะห์ความเป็นคู่ขนานที่แท้จริงของเนื้อหาของแบบทดสอบด้วย เช่น areas of text, text types, text lengths และ readability เป็นต้น

17.2.2 หลังจากปรับปรุงแก้ไขแบบทดสอบ EELT ที่สร้างขึ้นใหม่ทั้ง 6 ชุดแล้ว ควรนำไปใช้จริงอีกสักครั้งหนึ่ง แล้วนำผลการสอบมาตรวจสอบคุณภาพทางด้านการวัดและประเมินผลอีกครั้งหนึ่ง โดยเฉพาะอย่างยิ่ง ควรศึกษาว่าแบบทดสอบแต่ละชุดเป็นแบบทดสอบคู่ขนานจริงกับแบบทดสอบ ELTIS หรือไม่

18. เอกสารอ้างอิง

ภาษาไทย

กองทุนเพื่อความเสมอภาคทางการศึกษา. (2565). จำนวนนักเรียน จำแนกตามเพศและช่วงชั้น ปีการศึกษา 2564.

สืบค้นจาก <https://isee.eef.or.th/screen/studentdata/student.html>

ครูสภา. (2563). หลักเกณฑ์การเทียบเคียงผลการทดสอบและประเมินสมรรถนะทางวิชาชีพครูด้านความรู้และ
ประสบการณ์วิชาชีพ ตามมาตรฐานวิชาชีพครู ประจำปี พ.ศ. 2563-2565. สืบค้นจาก <https://www.ksp.or.th/ksp2018/wp-content/uploads/2020/11/Announce-of-teachers-licence-test.pdf>

สร้อยญา จันทรัฐสกุล, ไชยยศ ไพวิทยศิริธรรม และยุวรี ผลพันธุ์. (2560). การเปรียบเทียบคุณภาพของการปรับเทียบ
คะแนนสอบภาษาอังกฤษระหว่างวิธีอีคิวเปอร์เซ็นต์ไคล์ วิธีเชิงเส้นตรง และวิธีสมการถดถอย. *วารสาร Veridian E-Journal*, Silpakorn University, 10(2), 2444-2455. สืบค้นจาก <https://he02.tci-thaijo.org/index.php/Veridian-E-Journal/article/view/103591/83359>

สุพัฒน์ สุกมลสันต์. (2560). การเทียบคะแนนแถบระดับความสามารถของแบบทดสอบวัดสมรรถภาพทางภาษาอังกฤษชุดใหม่
ของมหาวิทยาลัยแม่โจ้ (New MJU-TEP) กับระดับของ CEFR. (รายงานวิจัย). เชียงใหม่ : ศูนย์ภาษา มหาวิทยาลัยแม่โจ้.

สำนักงานเลขาธิการสภาการศึกษา. (2560). แผนการศึกษาแห่งชาติ พ.ศ. 2560-2579. กรุงเทพมหานคร : สำนักพิมพ์
พริกหวานกราฟฟิค จำกัด. สืบค้นจาก <http://www.lampang.go.th/public60/EducationPlan2.pdf>

ภาษาอังกฤษ

Albano, A. (2010). *Statistical equating methods*. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.168.826&rep=rep1&type=pdf>

Arikan, C. A. & Gelbal, S. (2018). A comparison of traditional and Kernel equating methods. *International Journal of Assessment Tools in Education*, 5(3), 417–427. Retrieved from <https://dergipark.org.tr/tr/pub/ijate/issue/37036/409826>

Baghaei, P. (2010). Test score equating and fairness in language assessment. *Journal of English Language Studies*, 1(3), 113-128.

Brennan, R. L., Wang, T., Kim, S. & Seol, J. (2009). *Equating recipes (Version 1) [Computer software and manual]* (CASMA Monograph No. 1). Iowa City, IA : Center for Advanced Studies in Measurement and Assessment, The University of Iowa. Retrieved from <http://www.education.uiowa.edu/casma>

Council of Europe. (2020). *Common European framework of reference for languages : Learning, teaching, Assessment—Companion volume*. Strasbourg : Council of Europe Publishing. Retrieved from www.coe.int/lang-cefr

Cui, Z. & Kolen, M. J. (2009). Evaluation of two new smoothing methods in equating : The cubic B-spline presmoothing method and the direct presmoothing method. *Journal of Educational Measurement*, 46(2), 135–158.

Dorans, N. J. (2004). Equating, concordance, and expectation. *Applied Psychological Measurement*, 28(4), 227-246.

Fairbank, B. A. Jr. (1987). The use of presmoothing and posts smoothing to increase the precision of equipercentile equating. *Journal of Applied Psychological Measurement*, 11(3), 245-262. Retrieved

- from <https://conservancy.umn.edu/bitstream/handle/11299/104061/v11n3p245.pdf?sequence=1>
- Grant, M. C., Zhang, L. & Damiano, I. (2009). *An evaluation of Kernel equating : Parallel equating with classical methods in the SAT subject tests program* (Research Report ETS RR-09-06). Princeton, New Jersey : Educational Testing Service.
- Gubes, N. & Uyar, S. (2020). Comparing performance of different equating methods in presence and absence of DIF items in Anchor Test. *International Journal of Progressive Education*, 16(3), 111-122.
- Hanson, B. A. (1991). A comparison of bivariate smoothing methods in common-Item equipercentile equating. *Journal of Applied Psychological Measurement*, 15(4), 391-408. Retrieved from <https://conservancy.umn.edu/bitstream/handle/11299/114469/1/v15n4p391.pdf>
- Hanson, B. A., Zeng, L. & Colton, D. (1994). *A comparison of presmoothing and postsmoothing methods in equipercentile equating* (ACT Research Report Series 94-4). Iowa : The American College Testing Program. Retrieved from https://www.act.org/content/dam/act/unsecured/documents/ACT_RR94-04.pdf
- Heh, V. K. (2007). *Equating accuracy using small samples in the random group design* (Doctoral Dissertation). Ohio, USA : The College of Education of Ohio University. Retrieved from <https://etd.ohiolink.edu/!etd.sendfile?accession=ohiou1178299995&disposition=inline>
- Hopkins, W. G. (2002). *A new view of statistics : Correlation coefficients*. Retrieved from <https://www.sportsci.org/resource/stats/effectmag.html>
- Kolen, M. J. & Brennan, R. J. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York : Springer-Verlag.
- Kontos, D. L. (2007). *Investigation of validity, reliability, and practice effects of the immediate postconcussion assessment and cognitive test (impact) and traditional paper-pencil neuropsychological tests* (Master's thesis). North Carolina, USA : University of North Carolina at Chapel Hill. Retrieved from <https://doi.org/10.17615/ht0k-ge23>
- Liu, J. & Low, A. C. (2008). A comparison of the Kernel equating method with traditional equating methods using SAT data. *Journal of Educational Measurement*, 45(4), 309-323. Retrieved from <https://doi.org/10.1111/j.1745-3984.2008.00067.x>
- Livingston, S. (1992). *Small-sample equating with log-linear smoothing* (Research Report No. RR-92-4). Princeton, New Jersey : Educational Testing Service. Retrieved from <https://doi.org/10.1002/j.2333-8504.1992.tb01434.x>
- Lord, F. M. & Wingersky, M. S. (1984). Comparison of IRT true-Score and equipercentile observed-score "equatings". *Journal of Psychological Measurement*, 8(4), 453-461.
- Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA : Addison-Wesley.
- Moses, T. & Liu, J. (2011). *Smoothing and equating methods applied to different types of test score distributions and evaluated with respect to multiple equating criteria* (Research Report ETS RR-11-20). Princeton, New Jersey : Educational Testing Service. Retrieved from <https://files.eric.ed.gov/fulltext/ED523686.pdf>
- Phillips, S. E. (1983). Comparison of equi-percentile and response theory equating when the scaling test method is applied to multilevel achievement batter. *Applied Psychological Measurement*, 7(3), 267-281.
- Rapp, J. (1999). Multilevel linear and equi-percentile methods for equating PET. *National Institute for Testing and Evaluation*. Retrieved from <https://www.nite.org.il/wp-content/uploads/2018/06/e266.pdf>

- Rovinelli, R. J. & Hambleton, R. K. (1977). On the use of content specialists in the assessment of criterion-referenced test item validity. *Tijdschrift voor Onderwijsresearch*, 2(2), 49–60. Retrieved from <https://files.eric.ed.gov/fulltext/ED121845.pdf>
- Shah, C. H. & Brown, J. D. (2020). Reliability and validity of the short-form 12 item version 2 (SF-12v2) health-related quality of life survey and disutilities associated with relevant conditions in the U.S. older adult population. *Journal of Clinical Medicine*, 9(3), 661. Retrieved from <https://www.mdpi.com/2077-0383/9/3/661>
- Shin, M. (2015). *An investigation of subtest score equating methods under classical test theory and item response theory frameworks* (Doctoral dissertation). Massachusetts, USA : University of Massachusetts Amherst. Retrieved from https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1397&context=dissertations_2
- Turner, R. C. & Carlson, L. (2003). Indexes of item-objective congruence for multidimensional items. *International Journal of Testing*, 3(2), 163-171. Retrieved from https://www.tandfonline.com/doi/abs/10.1207/S15327574IJT0302_5
- von Davier, A. A. et al. (2006). *An evaluation of the kernel equating method: a special study with pseudotests constructed from real test data* (Research Report RR-06-02). Princeton, New Jersey : Educational Testing Service. Retrieved from <https://files.eric.ed.gov/fulltext/EJ1111405.pdf>
- Wiberg, M. (2021). On the use of different linkage plans with different observed-score equipercentile equating methods. *Practical Assessment, Research & Evaluation*, 26(23), 1-16. Retrieved from <https://scholarworks.umass.edu/pare/vol26/iss1/23>
- Yu, C. H. & Osborn Popp, S. E. (2005). Test equating by common items and common subjects : concepts and applications. *Practical Assessment Research and Evaluation*, 10(4), 1–19.