บทวิจารณ์เกี่ยวกับการวิจัยเพื่อกำหนดเกณฑ์ระดับความสามารถของแบบทดสอบด้วย Modified Angoff Method ตามเกณฑ์ของ CEFR และข้อเสนอแนะ: กรณีศึกษาบทความวิจัย 2 เรื่อง

สุพัฒน์ สุกมลสันต์*
มหาวิทยาลัยแม่โจ้

บทคัดย่อ

กรอบแนวคิดของ CEFR เข้ามามีผลกระทบที่สำคัญต่อการเรียนการสอนในประเทศไทยตั้งแต่ปีพ.ศ. 2557 ปัจจุบันมีงานวิจัยบางเรื่องทำการศึกษาการกำหนดจุดตัดของคะแนนแบบทดสอบภาษาอังกฤษกับเกณฑ์ระดับของ CEFR โดยการใช้ Modified Angoff Methods จำนวน 2 วิธี ผลการวิจัยอาจก่อให้เกิดผลกระทบในวงกว้างต่อไปได้ เพราะว่าผู้วิจัยใช้แบบทดสอบที่มีผลกระทบในวงกว้าง ผู้เขียนพบว่ามีการวิจัยดังกล่าวมีจุดพกพร่องที่สำคัญหลายอย่าง เช่น เน้นที่ค่าความเที่ยงในการตัดสินใจมากกว่าค่าความตรง เป็นวิธีที่เป็นไปได้ยากมาก ใช้ค่าสถิติอย่างไม่เหมาะสมเพื่อให้ผู้เชี่ยวชาญประกอบการตัดสินใจ และใช้กระบวนการคิดไม่สมเหตุผล เป็นต้น จึงต้องการชี้ให้ผู้วิจัย และผู้ที่จะนำผลการวิจัยไปใช้ได้ตระหนักรู้ถึงข้อจำกัด รวมทั้งได้ให้ข้อเสนอแนะเพื่อให้ผู้วิจัยอื่นที่จะทำงานวิจัยในทำนองเดียวกันในอนาคตให้ได้ผลดียิ่งขึ้น

คำสำคัญ: วิธีดัดแปลงของ Angoff, การกำหนดเกณฑ์มาตรฐาน, กรอบอ้างอิงร่วมของยุโรป (CEFR)

*รศ.ดร. สุพัฒน์ สุกมลสันต์, ศูนย์ภาษา มหาวิทยาลัยแม่โจ้ เชียงใหม่

A Critical Review of Using Modified Angoff Methods for Cut-off Score Settings based on CEFR and Suggestions: A Case Study of 2 Research Articles

Abstract

CEFR has had important impacts on English language teaching and learning in Thailand since 2014. At present, there are some research studies on mapping cut-off scores of English tests with CEFR Levels by 2 Modified Angoff Methods. Their findings may have high impacts on the students and education administrators because they are high-stakes tests. The author found that there are many serious weaknesses in the studies, for example, they emphasize on reliability rather than validity, the methods are next to impossible to perform, they make use of statistics inappropriately for experts in the field to make their judgments, and their processes are illogical. Therefore, the author would like to point them out to the researchers and research consumers to make them realize their limitations. Some suggestions were provided for anyone who wants to conduct a similar study in the future to yield better results.

Keyword: Modified Angoff Method, Standard Setting, CEFR

ความนำ

เนื่องจากแนวคิดเกี่ยวกับกรอบอ้างอิงร่วมของยุโรป (Common European Framework of Reference : CEFR) เข้ามามีบทบาทในการเรียนการสอนของประเทศไทยมากตั้งแต่ปี พ.ศ. 2557 เพราะว่ากระทรวงศึกษาธิการได้กำหนดเป็นนโยบายให้การเรียนการสอน และการทดสอบประเมินผลภาษาอังกฤษในระดับประถมศึกษาและมัธยมศึกษาให้เป็นไปตามเกณฑ์ดังกล่าว (กระทรวงศึกษาธิการ, 2557) และในระดับอุดมศึกษาก็มีการเสนอแนะให้มหาวิทยาลัยต่างๆปรับปรุงคุณภาพของการเรียนการสอนภาษาอังกฤษให้สอดคล้องกับเกณฑ์ดังกล่าวด้วย (สำนักงานอุดมศึกษา, 2559) ดังนั้น จึงมีงานวิจัยที่เกี่ยวกับการกำหนดเกณฑ์มาตราฐานของคะแนนแบบทดสอบ (Standard Setting) ที่สร้างขึ้น หรือมีอยู่แล้วในประเทศเพื่อเทียบกับเกณฑ์ดังกล่าว ผู้เขียนได้อ่านพบบทความวิจัย 2 เรื่องที่เกี่ยวกับ Test Item Mapping และการเทียบคะแนนของแบบทดสอบสมิทธิภาพของมหาวิทยาลัยสองแห่งอย่างละเอียด เพราะว่าเป็นงานวิจัยที่เกี่ยวกับแบบทดสอบที่มีผลกระทบต่อผู้สอบและผู้ที่เกี่ยวข้องมาก (high-stakes test) ผู้เขียนมีวัตถุประสงค์ในการเขียนครั้งนี้ 2 ประการ คือ เพื่อชี้ให้ผู้อ่าน ผู้วิจัย และผู้ที่จะนำผลการวิจัยไปใช้ว่างานทั้ง 2 เรื่อง หรือเรื่องอื่นที่ดำเนินการวิจัยในทำนองเดียวกันมีข้อจำกัดที่สำคัญมากบางอย่าง และเพื่อเสนอแนะวิธีการต่างๆที่จะได้ผลการวิจัยดีกว่าเดิม สำหรับผู้ที่ต้องการทำการวิจัยเพื่อสร้างตารางเทียบคะแนนของแบบทดสอบของตนกับเกณฑ์ CEFR ในอนาคต และขอวิจารณ์บทความดังกล่าวในเชิงวิชาการ และผลกระทบที่อาจเกิดขึ้นในวงกว้างต่อวงการเรียนการสอน และการทดสอบภาษาอังกฤษในประเทศเท่านั้น โดยไม่ได้มีเจตนาที่จะก่อให้เกิดผลกระทบต่อผู้วิจัยที่ทำการวิจัยดังกล่าวแล้วแต่อย่างใด

วิธีการกำหนดระดับความสามารถ

วิธีกำหนดระดับความสามารถ หรือการกำหนดเกณฑ์มาตรฐาน (Standard Setting) หมายถึงกระบวนการของการบ่งชี้คะแนนต่ำสุดของแบบทดสอบชุดหนึ่งที่สามารถจำแนกผู้สอบออกที่มีความสามารถต่างกันได้ (Tannenbaum, 2011 อ้างถึงใน CaMLA, 2015) หรือหมายถึงวิธีที่ใช้เพื่อระบุระดับสัมฤทธิ์ผลหรือสมิทธิภาพของผู้สอบ และคะแนนจุดตัด (cutscores) คือคะแนนที่แบ่งระดับความสามารถดังกล่าวแล้ว (Bejar, 2008) การกำหนดคะแนนดังกล่าวมีมากกว่า 35 วิธี เมื่อนับถึงปี 1986 (Berk, 1986) และในปัจจุบันนี้คงจะมีจำนวนมากกว่านี้ เพราะว่าแต่ละวิธีมีข้อดีและข้อด้อยแตกต่างกัน และมักจะให้ผลของการกำหนดเกณฑ์แตกต่างกัน (Barman, 2008) วิธีต่างๆเหล่านี้สามารถจำแนกออกได้เป็น 2 ประเภท (Eckes, 2012) คือ เน้นที่การศึกษาตัวข้อทดสอบ (Item-centered studies) เช่น Angoff Method, Modified Angoff Method, Nedelsky Method และ Bookmark Method เป็นต้นและ เน้นที่ผู้สอบแบบทดสอบ (Person-centered studies) เช่น Borderline Groups method และ Contrast Groups Method เป็นต้น และนักวัดและประเมินผลบาง

คนจำแนกวิธีต่างๆเหล่านี้ตามแหล่งที่มาของข้อมูลเป็น 3 ประเภท (Berk, 1986) คือ วิธีที่อาศัยความคิดเห็นของผู้เชี่ยวชาญ (Judgmental Methods), วิธีที่อาศัยความคิดเห็นของผู้เชี่ยวชาญประกอบข้อมูลเชิงประจักษ์ของข้อทดสอบ (Judgmental-Empirical Methods) และ วิธีที่อาศัยข้อมูลเชิงประจักษ์ของข้อทดสอบประกอบความคิดเห็นของผู้เชี่ยวชาญ (Empirical-Judgmental Methods) สำหรับ Modified Angoff Method เป็นวิธีการกำหนดเกณฑ์มาตรฐานแบบหนึ่งในกลุ่มที่อาศัยความคิดเห็นของผู้เชี่ยวชาญประกอบข้อมูลเชิงประจักษ์ของข้อทดสอบ (Berk, 1986)

การกำหนดระดับความสามารถด้วย Angoff Method และ Modified Angoff Methods

1. Traditional Angoff Method (วิธีของ Angoff แบบประเพณีนิยม)

W.H. Angoff (1971) เป็นผู้นำเสนอวิธีกำหนดเกณฑ์ระดับความสามารถแบบนี้ขึ้น เพื่อกำหนดจุดตัดของคะแนน (Cut-off Score) ของแบบทดสอบชุดใดชุดหนึ่งที่สร้างขึ้น จุดตัดของคะแนนนี้แบ่งผู้สอบออกเป็น 2 กลุ่มเท่านั้น คือกลุ่มที่ประสบผลสำเร็จในการเรียน (Mastery Level) กับกลุ่มที่ยังไม่ประสบผลสำเร็จในการเรียน (Non-Mastery Level) และบริเวณจุดตัดของคะแนนเรียกว่า Minimum Competence Level (ระดับที่มีความสามารถต่ำที่สุด) หรือชื่ออย่างอื่นที่สื่อความหมายในทำนองเดียวกัน เช่น The borderline, Minimally Competent Candidate เป็นต้น วิธีนี้ดำเนินการโดยสังเขปดังนี้

1. สร้างแบบทดสอบมาตรฐานขึ้นมา 1 ชุด
2. แต่งตั้งผู้เชี่ยวชาญในเนื้อหาของแบบทดสอบขึ้น 1 ชุดหนึ่ง
3. ให้ผู้เชี่ยวชาญพิจารณาอย่างเป็นอิสระว่า ผู้สอบที่มีความสามารถต่ำที่สุดน่าจะทำข้อทดสอบแต่ละข้อได้ถูกต้องร้อยละเท่าใด เช่น 30%, 45% และ 55% เป็นต้น
4. นำค่าเฉลี่ยของร้อยละของการคาดการณ์ตอบถูกแต่ละข้อ และทุกๆข้อมาสรุปเป็นเกณฑ์จุดตัดมาตรฐานของแบบทดสอบที่นำมาพิจารณา

2. Modified Angoff Method (วิธีของ Angoff แบบดัดแปลง)

ต่อมามีนักทดสอบและประเมินผลได้ดัดแปลงวิธีของ Angoff แบบประเพณีนิยมเป็นแบบดัดแปลงหลายรูปแบบเช่น Modified Angoff Method 1 ดำเนินการเหมือนกับ Traditional Angoff Method แต่ให้คณะกรรมการประชุมร่วมกันเพื่อพิจาณาร้อยละของการตอบถูกจำนวน 2-3 ครั้ง Modified Angoff Method 2 ดำเนินการเหมือนกับ Traditional Angoff Method แต่ให้ข้อมูลรายข้อแก่คณะกรรมการก่อนการพิจารณาครั้งสุดท้าย Modified Angoff Method 3 ดำเนินการเหมือนกับ Traditional Angoff Method แต่ให้ผู้เชี่ยวชาญพิจารณาว่าอย่างอิสระว่า ผู้สอบที่มีความสามารถต่ำที่สุดน่าจะทำข้อทดสอบแต่ละข้อได้ถูกต้อง

หรือไม่ ถ้าคิดว่าได้ให้ระบุเป็น 1 แต่ถ้าคิดว่าไม่ได้ ให้ระบุเป็น 0 ในรอบที่ 1 แล้วให้ข้อมูลผลการสอบรายข้อ เช่น ค่าความยาก-ง่าย (difficulty index) และ/หรือ ค่าอำนาจจำแนก (discrimination index) แก่คณะกรรมการในรอบที่ 2 เพื่อประกอบการตัดสินใจว่าจะเปลี่ยนแปลงหรือไม่ อย่างไร แล้วประชุมคณะกรรมการเพื่ออภิปรายผลการตัดสินใจแต่ละข้อของคณะกรรมการเพื่อปรับแก้ไขเป็นครั้งสุดท้ายในรอบที่ 3 เพื่อหาข้อสรุปร่วมกันเป็นรายข้อ รวมทั้งเกณฑ์จุดตัดของคะแนนด้วย และเมื่อปี 1998 ได้มีผู้เสนอวิธี Modified Angoff Method 4 หรือ Extended Angoff Method ขึ้นมาใช้ (Plake et al., 1998) สำหรับการกำหนดจุดตัดของคะแนนแบบทดสอบที่วัดอเนกมิติ (Multi-dimensionality) และให้คะแนนหลายรูปแบบ (polytomously scored) แทนที่จะเป็นการให้คะแนนแบบถูก-ผิด (dichotomously scored) จากการทดสอบแบบเลือกตอบ (Multiple-choice) เท่านั้น เหมือน Traditional Angoff Method และ Modified Angoff Method อื่นๆ ซึ่งต่อมาก็มีผู้นำไปใช้เพื่อกำหนดจุดตัดมาตรฐานของคะแนนสอบ (standard setting) เพิ่มขึ้นเป็นหลายจุดหรือคะแนน

สำหรับงานวิจัยทั้ง 2 เรื่องดังกล่าวข้างต้น ใช้การกำหนดจุดตัดมาตรฐาน (Standard Setting) ของคะแนนวัดระดับความสามารถ ด้วย Modified Angoff Method 3 ที่มีการกำหนดคะแนนจุดตัดมากกว่า 1 จุด และมีรายละเอียดแตกต่างกันเล็กน้อย กล่าวคือ ในกรณีของงานวิจัยเรื่องที่ 1 ได้ใช้กำหนดจุดตัด 4 จุด คือ A2, B1, B2 และ C1 โดยกำหนดให้ผู้เชี่ยวชาญพิจารณาข้อทดสอบรายข้อ หากคิดว่าผู้สอบที่มีความสามารถต่ำสุดในแต่ละระดับความสามารถมีโอกาสตอบได้ถูก ให้ใส่รหัส = 1 แต่หากว่าคิดว่าคงตอบผิด ให้รหัส = 0 แล้วให้ข้อมูลด้านค่ายาก-ง่ายของข้อสอบรายข้อแก่ผู้เชี่ยวชาญในการพิจารณารอบที่ 2 ส่วนงานวิจัยเรื่องที่ 2 กำหนดจุดตัด 3 จุด คือ A2, B1 และ B2 เท่านั้น และกำหนดให้ผู้เชี่ยวชาญพิจารณาข้อทดสอบรายข้อว่า หากข้อใดคิดว่าผู้สอบที่มีความสามารถต่ำสุดในแต่ละระดับความสามารถมีโอกาสตอบได้ถูก ให้ระบุร้อยละของโอกาสที่จะตอบถูกตามที่ผู้วิจัยกำหนดไว้ให้ เช่น 5%, 10% หรือ 15% เป็นต้น แต่หากคิดว่าคงจะตอบผิด ให้ใส่ 0% แล้วให้ข้อมูลด้านค่ายาก-ง่าย และค่าอำนาจจำแนกของข้อสอบรายข้อแก่ผู้เชี่ยวชาญในการพิจารณารอบที่ 2

ข้อดีและข้อด้อยของการกำหนดจุดตัดมาตรฐานโดย Modified Angoff Method 3 ที่ใช้ในงานวิจัยทั้ง 2 เรื่อง

ข้อดี

1. เป็นวิธีที่ดำเนินการง่ายมาก สำหรับผู้ให้ข้อมูลเพราะว่าผู้เชี่ยวชาญในศาสตร์เพียงแต่ให้ความคิดเห็นว่าผู้สอบในระดับความสามารถต่างๆสามารถจะทำข้อสอบแต่ละข้อได้ถูกต้องหรือไม่ หรือถูกต้องสักร้อยละเท่าใด และอาจมีการปรับเปลี่ยนความคิดเห็นได้ในภายหลังเมื่อได้ข้อมูลของผลการสอบมาประกอบการตัดสินใจ ดังนั้นจึงไม่รู้สึกตรึงเครียดในการให้ข้อมูล (Berk, 1986; Stahl & VUE, 2019)
2. เป็นวิธีที่ดำเนินการง่ายมาก สำหรับผู้วิจัย เพราะว่าใช้เพียงสถิติเชิงพรรณา (Descriptive Statistics) เท่านั้น (Berk, 1986; Stahl & VUE, 2019)

   ดังนั้น Traditional Angoff Method และ Modified Angoff Method ต่างๆ ยกเว้น Extended Angoff Method จึงได้รับความนิยมอย่างแพร่หลาย โดยเฉพาะในหมู่ครูอาจารย์ และนักวัดและประเมินผลทางการศึกษาที่ต้องการกำหนดจุดตัดมาตรฐานของคะแนนวัดระดับความสามารถของแบบทดสอบที่มีผลกระทบน้อยต่อผู้สอบ (low-stakes test) เพียง 1 จุด คือคะแนนจุดตัดระหว่าง กลุ่มผู้ที่ประสบผลสำเร็จในการเรียน (Mastery Level) กับกลุ่มที่ยังไม่ประสบผลสำเร็จในการเรียน (Non-Mastery Level) เช่น ใช้เพื่อประเมินผลสัมฤทธิ์ในการเรียนในห้องเรียน หรือในกลุ่มโรงเรียน เป็นต้น
3. งานวิจัยทั้ง 2 เรื่อง ผู้วิจัยดำเนินการวิจัยตามขั้นตอนต่างๆได้อย่างละเอียด และครบถ้วนได้ค่อนข้างดี แต่ว่าวิธีการที่นำมาใช้ยังขาดสาระที่สำคัญมากบางอย่าง ซึ่งมีผลเสียต่อผลการวิจัย ดังที่จะได้กล่าวต่อไป

ข้อบกพร่อง

1. **งานวิจัยทั้ง 2 เรื่องไม่ได้เน้นที่ความตรง (Validity) ของการประเมิน แต่กลับเน้นที่ความเที่ยง (Reliability) ซึ่งนับว่าเป็นข้อบกพร่องที่ร้ายแรงที่สุดของวิธีการวิจัย**

   กล่าวคือ ความคิดเห็นที่สอดคล้องกันของผู้เชี่ยวชาญสำหรับงานวิจัยที่ใช้ Traditional Angoff Method หรือ Modified Angoff Method ไม่ได้หมายความว่าผู้เชี่ยวชาญตัดสินถูกต้องว่าผู้สอบแต่ละระดับความสามารถมีความรู้หรือความสามารถในระดับที่ตนประเมิน เนื่องจากการประเมินว่าผู้ตอบสามารถตอบถูกไม่ได้หมายความว่าผู้ตอบมีความสามารถอยู่ในระดับนั้นเสมอไป ดังนั้น การที่ผู้เชี่ยวชาญมีความคิดเห็นสอดคล้องกันมาก แสดงถึงความคงเส้นคงวาหรือมีความเที่ยงสูง (high reliability) แต่ไม่ได้แสดงว่ามีความตรงสูง (high validity) ด้วยโดยอัตโนมัติ เพราะว่าในการตัดสินดังกล่าวอาจ

ไม่ถูกต้องก็ได้ หากว่าผู้เชี่ยวชาญมีแนวความคิดเกี่ยวกับผู้สอบที่มีระดับความสามารถต่ำสุดของแต่ละระดับความสามารถไม่ถูกต้อง (Barman, 2008)

2. **การกำหนดจุดตัดของคะแนนโดย Modified Angoff Method ที่ผู้วิจัยใช้ในงานวิจัยทั้ง 2 เรื่อง ในรอบที่ 1 เป็นวิธีการที่เป็นไปได้ยากมากที่จะถูกต้อง** เพราะว่าอาศัยวิจารณาณของผู้เชี่ยวชาญในศาสตร์ (Judgmental Standard) เท่านั้นที่คาดการณ์ว่าผู้สอบที่มีความสามารถต่ำสุดในแต่ละระดับความสามารถจะสามารถทำข้อสอบแต่ละข้อได้ถูกต้องร้อยละเท่าใด หรือตอบได้ถูกหรือไม่ แนวคิดนี้อาศัยความเห็นของผู้เชี่ยวชาญเกี่ยวกับความยากของข้อทดสอบสำหรับผู้สอบที่มีความสามารถต่ำสุดในแต่ละระดับเป็นหลักสำคัญ ซึ่งการตัดสินใจดังกล่าวมักขึ้นอยู่กับปัจจัยหลายอย่าง เช่น ความคุ้นเคยของผู้เชี่ยวชาญกับกระบวนการนี้มาก่อน การรับรู้ว่าข้อสอบที่กำลังพิจารณาวัดความรู้ที่จำเป็นหรือไม่จำเป็นในขณะนั้น และวัตถุประสงค์ของแบบทดสอบ เป็นต้น (Burr et al., 2017) จึงเป็นแนวคิดที่ผู้เชี่ยวชาญในศาสตร์การวัดและประเมินผลทางการศึกษาที่มีชื่อเสียงมาก เช่น Berk (1996 อ้างถึงใน Ricker, 2006) วิจารณ์ว่าเป็น “ภาระงานทางด้านพุทธพิสัยที่เกือบเป็นไปไม่ได้” (nearly impossible cognitive task) และ Shepard (1995 อ้างถึงใน Ricker, 2006) ก็วิจารณ์ว่าเป็นแนวคิดที่ “เกินศักยภาพในการประมวลผลด้านพุทธิพิสัยของมนุษย์” (exceed human cognitive processing capacities) และเป็นภาระงานที่ยากมากสำหรับผู้เชี่ยวชาญ (Impara et al, 1997 & 1998 อ้างถึงใน Stahl, 2019) และแนวคิดดังกล่าวเป็นการกำหนดจุดตัดของคะแนนที่อาศัยความเป็นอัตนัย (subjectivity) ของผู้เชี่ยวชาญมาก จึงเป็นคะแนนจุดตัดที่ปราศจากเหตุผล หรือหลักการ เพราะกระทำตามอำเภอใจ (arbitrary) ของผู้เชี่ยวชาญ (Barman, 2008) ทั้งนี้ผู้เขียนเห็นด้วยเป็นอย่างยิ่งกับคำวิจารณ์ดังกล่าวเพราะว่าผู้เชี่ยวชาญไม่ได้เป็นผู้สอบแบบทดสอบเอง แต่เป็นเพียงการคาดการณ์หรือสมมติที่เป็นไปไม่ได้ว่า ผู้สอบที่มีความสามารถต่ำสุดในแต่ละระดับจะสามารถตอบข้อทดสอบได้ถูกต้องหรือไม่ หรือว่าจะถูกต้องสักร้อยละเท่าใด ประเด็นนี้เป็นจุดอ่อนมากที่สุด และได้รับการวิพากษ์วิจารณ์จากนักวัดและประเมินผลทั่วไปของวิธีการที่นำมาใช้เพื่อการวิจัย (Wilson & Santelices, 2017)

3. **งานวิจัยทั้ง 2 เรื่องใช้ค่าสถิติอย่างไม่เหมาะสมเพื่อประกอบการตัดสินใจ** กล่าวคือค่าสถิติของข้อทดสอบรายข้อที่นำมาให้ผู้เชี่ยวชาญใช้ประกอบการตัดสินใจในรอบที่ 2 คือ ค่าความยาก-ง่าย (difficulty index) และค่าอำนาจจำแนก (discrimination index) รวมทั้งการอภิปรายในรอบที่ 3 เป็นการให้ข้อมูลที่ไม่เหมาะสมที่จะทำให้ผลการกำหนดจุดตัดตรงตามความเป็นจริง เพราะว่า 1) เป็นค่าสถิติรวมของผู้สอบที่มีความสามารถหลายระดับ ไม่ใช่เฉพาะผู้สอบในแต่ละระดับความสามารถที่ผู้เชี่ยวชาญ

พิจารณา 2) ข้อมูลดังกล่าวมาจากคะแนนปรากฏ (observed scores) ไม่ใช่จากคะแนนจริง (true scores) ซึ่งเป็นสิ่งที่ควรใช้เพื่อให้ได้จุดตัดที่แท้จริง (true standard) เพราะว่าจะใช้สรุปอ้างอิง (generalize) ได้ดีกว่า (Berk, 1986 และ van der Linden, 1982 & 1984 อ้างถึงใน Ricker, 2006) 3) ความยาก-ง่ายของข้อทดสอบไม่ได้ขึ้นอยู่กับเนื้อหาที่ต้องการทดสอบเท่านั้น แต่ยังขึ้นอยู่กับภาษาที่ใช้ในข้อคำถาม (test stem) คุณภาพของตัวเลือกต่างๆของข้อทดสอบแบบเลือกตอบ ความสามารถของผู้สอบ และการเรียนการสอนด้วย รวมทั้งค่าอำนาจจำแนกไม่ได้บ่งบอกว่าเป็นข้อสอบไม่ดีในกรณีที่เป็นแบบทดสอบแบบอิงเกณฑ์ (Barman, 2008) อย่างที่นำมาเพื่อการวิจัยครั้งนี้ 4) การอภิปรายของผู้เชี่ยวชาญในขั้นสุดท้ายนั้น อาจเป็นไปได้ที่ผู้เชี่ยวชาญบางคนมีอิทธิพลเหนือความคิดของผู้เชี่ยวชาญอื่นแล้วชักจูงให้ตัดสินใจผิดพลาด (Hejri & Jalili, 2014) และจากการวิจัยเชิงอภิวิเคราะห์ (Meta-Analysis) พบว่า การให้ข้อมูลผลการสอบรายข้อแก่ผู้เชี่ยวชาญทำให้คะแนนจุดตัดต่ำลงกว่าการตัดสินใจในรอบที่ 1 และ 3 (Hurtz & Auerbach, 2003) และจากการวิจัยเชิงอภิวิเคราะห์ (Meta-Analysis) ได้ข้อสรุปว่า การให้ข้อมูลผลการสอบรายข้อแก่ผู้เชี่ยวชาญทำให้ผู้เชี่ยวชาญกำหนดจุดตัดลดลงกว่าเดิม (Hurtz & Auerbach, 2003) ดังนั้น ค่าสถิติต่างๆและข้อมูลที่นำมาใช้เหล่านั้นย่อมทำให้การตัดสินใจผิดพลาดไปด้วย

4. **Modified Angoff Method ใช้ตรรกะในการคิดไม่สมเหตุผล** กล่าวคือ Modified Angoff Method ที่ใช้ในการวิจัยทั้ง 2 เรื่องเป็นวิธีที่อาศัยความคิดเห็นของผู้เชี่ยวชาญก่อนใช้ข้อมูลเชิงประจักษ์ของข้อทดสอบประกอบการตัดสินใจภายหลัง (Judgmental-Empirical Methods) เป็นวิธีที่ใช้ตรรกะในการคิดที่ไม่สมเหตุสมผล (illogical thinking) เพราะว่าผู้เชี่ยวชาญได้ใช้ความรู้ความสามารถของตนพิจารณาตัดสินไปก่อนแล้ว ก่อนที่จะเห็นข้อมูลเชิงประจักษ์มาประกอบการตัดสินใจในภายหลัง จากการวิจัยที่ใช้ข้อมูลจำลอง (simulated data) ที่มีผู้เชี่ยวชาญจำนวน 4900 คน พิจารณาตัดสินโอกาสในการตอบข้อทดสอบ 2 รอบโดยไม่มีข้อมูลใดๆให้ประกอบการตัดสินใจ พบว่าความคิดเห็นแตกต่างกันอย่างมีนัยสำคัญทางสถิติ แต่สามารถเพิกเฉยได้ในทางปฏิบัติเพราะว่ามีขนาดน้อยมาก กล่าวคือ $p < 0.001$; Cohen's $d = -0.083$ เท่านั้น (Shulruf et al., 2015) แสดงให้เห็นอย่างชัดเจนว่า ผู้เชี่ยวชาญในศาสตร์มักไม่ค่อยเปลี่ยนการตัดสินใจของตนเองที่ได้กระทำแล้วตั้งแต่แรก ซึ่งอาจเป็นเพราะบุคคลเหล่านี้มีความมั่นใจในตัวเองสูงในการตัดสินใจ อนึ่ง จากบทความวิจัยทั้ง 2 เรื่องพบว่า ผู้วิจัยอนุญาตให้ผู้เชี่ยวชาญเปลี่ยนแปลงการตัดสินในในรอบแรกได้ภายหลังจากที่ได้เห็นข้อมูลผลการสอบรายข้อของแบบทดสอบแล้ว คือค่าความยาก-ง่าย ในกรณีของงานวิจัยเรื่องที่ 1 หรือค่าความยาก-ง่าย และค่าอำนาจจำแนก ในกรณีของ

งานวิจัยเรื่องที่ 2 แต่ไม่ได้มีการทดสอบว่ามีการเปลี่ยนแปลงอย่างมีนัยสำคัญหรือไม่ เพียงแต่บอกว่าความคิดเห็นของผู้เชี่ยวชาญในบางจุดตัดของคะแนนสอดคล้องกันมากขึ้น แต่ในบางจุดตัดกลับมีความคิดเห็นแตกต่างกันมากขึ้นในรายงานวิจัยเรื่องที่ 1 ส่วนงานวิจัยเรื่องที่ 2 ระบุแต่เพียงว่าความคิดเห็นของผู้เชี่ยวชาญมีความสอดคล้องกันมากขึ้นในรอบที่ 2 และ 3 วิธีกำหนดจุดตัดของคะแนนในระยะแรกๆอาศัยความคิดเห็นของผู้เชี่ยวชาญเป็นหลัก (Judgmental Methods) เช่น Angoff Method (1971) Ebel (1979) และ Nedelsy (1954) ต่อมามีวิธีต่างๆที่อาศัยความคิดเห็นของผู้เชี่ยวชาญก่อนใช้ข้อมูลเชิงประจักษ์ประกอบการตัดสินใจภายหลัง (Judgmental-Empirical Methods) เช่น Modified Angoff Method (1978), Beuk Method (1984) และ Hofstee Method (1983) และในปัจจุบันนี้นักวิจัยด้านการวัดและประเมินผลสนใจพัฒนาวิธีกำหนดจุดตัดโดยใช้ข้อมูลเชิงประจักษ์ก่อนเพื่อประกอบการตัดสินใจของผู้เชี่ยวชาญในภายหลัง (Empirical Judgmental Methods) เช่น Borderline Group Method (1982) Contrasting Groups Method (1982) และ Criterion Groups Method (1984) เป็นต้น (อ้างอิงใน Berk, 1986) รวมทั้ง Rasch IRT Model Method (Wang, 2003) ทั้งนี้อาจเป็นเพราะว่านักวิจัยเชื่อในตรรกะของความคิดว่าถูกต้องมากกว่า

5. **งานวิจัยทั้ง 2 เรื่อง ไม่มีวิธีตรวจสอบความสอดคล้องของความเห็นระหว่างผู้เชี่ยวชาญ (interjudge agreement)** กล่าวคือ เรื่องที่ 1 เมื่อใช้วิธีการแบบ 1 หรือ 0 ไม่มีวิธีการปรับแก้คะแนนจุดตัดต่างๆในรอบที่ 3 ที่เป็นที่ยอมรับทั่วไปในงานวิจัยเช่นนี้ และในกรณีที่งานวิจัยเรื่องที่ 2 การกำหนดให้ผู้เชี่ยวชาญเลือกร้อยละที่คาดว่าผู้สอบที่มีความสามารถต่ำที่สุดในแต่ละระดับตอบถูกต้องจากร้อยละที่ผู้วิจัยกำหนดไว้ล่วงหน้าอาจทำให้เกิดอคติ (bias) ในการให้ความคิดเห็นได้ (Berk, 1986) ดังนั้น ข้อบกพร่องดังกล่าวมีผลมีผลต่อความเที่ยงหรือความเชื่อมั่น (Reliability) ของการกำหนดจุดตัดของคะแนนในแต่ละระดับ และแม้ว่าในงานวิจัยเรื่องแรกจะมีการคำนวณหาค่าความคลาดเคลื่อนมาตรฐานของการตัดสิน (Standard Error of Judgment: SEJ) แต่ผู้วิจัยไม่ได้ใช้ค่านี้ในการปรับแก้จุดตัดในแต่ละระดับความสามารถเลย นอกจากระบุว่าจุดตัดของคะแนนในระดับใดมีค่าดังกล่าวมากหรือน้อยเท่านั้น นอกจากนี้ผู้วิจัยยังใช้ค่านี้ผิดเพราะว่าใช้ในรอบที่ 2 และ 3 ด้วย แทนที่จะใช้เฉพาะในรอบที่ 1 เท่านั้น (MacCann & Stanley, 2004)
6. **ความคิดเห็นเชิงประเมินของผู้เชี่ยวชาญในการพิจารณาข้อสอบจำนวนมากเป็นเวลานานๆ และหลายระดับความสามารถมีความเสถียรได้ยาก** กล่าวคือ การกำหนดจุดตัดของคะแนนด้วย Modified Angoff Method ที่นำมาใช้เพื่อการวิจัยครั้งนี้ เป็นการกำหนดจุดตัดหลายระดับ คือ 3-4 ระดับ และข้อสอบมีจำนวนมากคือ 100-

120 ข้อ ย่อมก่อให้เกิดความไม่คงที่ในความคิดเกี่ยวกับระดับความสามารถต่ำสุดของผู้สอบในแต่ละระดับเป็นเวลานานได้ง่าย (Impara, 1997; Impara & Plake, 1998 และ Boursicot & Roberts, 2006 อ้างถึงใน George et al., 2006; Burr et al., 2017)

7. **กรอบแนวคิดของ CEFR ที่นำมาใช้เพื่อการวิจัยไม่เหมาะสมกับประสบการณ์ และความรู้หรือความสามารถของผู้เชี่ยวชาญในงานวิจัยทั้ง 2 เรื่อง** เพราะว่าเป็นกรอบความรู้และความสามารถในการใช้ภาษาเพื่อการสื่อสารของผู้ใหญ่ในยุโรปในหลายบริบท (Read, 2019) ไม่ใช่การใช้ภาษาอังกฤษเพื่อการสื่อสารของนักเรียนหรือนักศึกษาในบริบทของไทย หรือบริบทของประเทศในเอเชียตะวันออก (East Asia counties) จึงเป็นเหตุให้นักการศึกษาในประเทศเหล่านี้เริ่มตระหนักว่าควรจะต้องปรับปรุงแก้ไขกรอบแนวคิดของ CEFR ให้เหมาะสมกับบริบทของตน เช่น จีน ไต้หวัน ญี่ปุ่น ออสเตรเลีย และนิวซีแลนด์ เป็นต้น หรือแม้แต่ในประเทศไทยก็มี FRELE-TH (Framework of Reference for English Language Education in Thailand) ซึ่งสร้างมาจากแนวคิดของ CEFR : Kulaporn Hiranburana et al, 2019) นอกจากนี้แนวคิดเกี่ยวกับคุณลักษณะของผู้สอบในแต่ละระดับความสามารถตามกรอบ CEFR ที่ใช้ในการวิจัยทั้ง 2 เรื่องเป็นกรอบแนวคิดกว้างๆด้านทักษะการสื่อสารด้านการฟัง และพูดเป็นหลัก ไม่ได้คอบคลุมทุกทักษะในแบบทดสอบที่นำมาวิจัย เช่น ทักษะการอ่านเข้าใจความ และคำศัพท์ เป็นต้น แต่ที่สำคัญมากก็คือ กรอบแนวคิดที่นำมาใช้ไม่ใช่ความรู้และความสามารถของผู้สอบที่ต่ำสุดในแต่ละระดับความสามารถตามแนวความคิดของ Traditional Angoff Method และ Modified Angoff Methods (Berk, 1986; Chinn, 2006; Burman, 2008) ปกติแล้วผู้ทำการวิจัยเรื่องจุดตัดมาตรฐานของคะแนนจะต้องให้คำนิยามลักษณะเฉพาะของผู้มีความสามารถต่ำสุดที่ชัดเจนในหลายมิติในเชิงปฏิบัติ แต่เป็นการยากมากที่จะให้คำนิยามดังกล่าวได้ถูกต้องเมื่อไม่มีผลการสอบเฉพาะผู้ที่มีความสามารถต่ำสุดในแต่ละระดับความสามารถ (Burr, 2017) ดังนั้น เมื่อประสบการณ์และความรู้และความสามารถของผู้เชี่ยวชาญไม่เหมาะสมกับกรอบแนวคิด กรอบแนวคิดไม่ถูกต้องชัดเจน และไม่มีข้อมูลผลการสอบของผู้มีความสามารถต่ำสุดในแต่ละระดับความ สามารถให้ผู้เชี่ยวชาญใช้ประกอบการพิจารณา ผลของความเห็นของผู้เชี่ยวชาญย่อมไม่ถูกต้องตามไปด้วย

8. **องค์ประกอบของผู้เชี่ยวชาญในงานวิจัยทั้ง 2 เรื่องไม่เหมาะสมสำหรับงานวิจัยที่ใช้ Modified Angoff Methods เพราะว่ามีลักษณะเป็นกลุ่มลักษณะเอกพันธุ์ (Homogeneous Group)** กล่าวคือ ในงานวิจัยเรื่องที่ 1 ใช้ผู้เชี่ยวชาญ 13 คนที่มีประสบการณ์ในการสอนภาษาอังกฤษระหว่าง 7-40 ปี สอนในระดับปริญญาตรีและบัณฑิตศึกษา วุฒิต่ำสุดคือปริญญาโท เป็นเพศชาย 1 คน และเพศหญิงอีก 12 คน ส่วน

เรื่องที่ 2 ใช้ผู้เชี่ยวชาญ 14 คน เป็นชาย 4 คน และหญิงอีก 10 คน มีประสบการณ์ในการสอนตั้งแต่ 5 ปีขึ้นไป และมีวุฒิทางการศึกษาระดับปริญญาโทและเอก ในสาขาการสอนภาษาอังกฤษและอื่นๆที่เกี่ยวข้อง ซึ่งจะเห็นได้ว่าผู้เชี่ยวชาญของงานแต่ละเรื่องเป็นกลุ่มลักษณะเอกพันธุ์ (Homogeneous Group) แต่จากการวรรณกรรมที่เกี่ยวข้องพบว่า กลุ่มผู้เชี่ยวชาญที่เหมาะสมในการกำหนดจุดตัดมาตรฐานของคะแนนควรเป็นกลุ่มอเนกพันธุ์ (Heterogeneous Group) เช่น มีวุฒิทางการศึกษาหลายระดับ มีประสบการณ์หลากหลาย มีตำแหน่งทางวิชาการหลายระดับ เป็นตัวแทนที่ดีของผู้ที่จะให้ข้อมูลหรือผู้มีส่วนได้ส่วนเสีย (stake-holders) กับผลของการกำหนดจุดตัด และควรต้องสามารถตอบข้อทดสอบได้ถูกต้องทุกข้อ ไม่ควรมีเฉพาะ “ผู้เชี่ยวชาญ” หรือ “ผู้ไม่เชี่ยวชาญ”หรือ “คนที่เข้มงวด (stringency)” หรือ “คนที่ใจดี (lenient)” หรือผู้มีวุฒิทางการศึกษาสูง เท่านั้น เพราะว่าปัจจัยเหล่านี้มีผลต่อการประเมินจุดตัดของคะแนน แต่ควรจะเป็นคณะผู้เชี่ยวชาญที่มีความหลากหลายในเรื่องดังกล่าวแล้ว (Morgan & Michaelides, 2005; Chinn, 2006; Verheggen et al., 2008; Hejri & Jaloli, 2014; Shulruf et al., 2015; Rezigalla, 2016)

9. **งานวิจัยทั้ง 2 เรื่องที่ไม่มีวิธีการตรวจสอบความถูกต้องหรือความตรง (Validity) ของการตัดสินใจของผู้เชี่ยวชาญโดยตรง** แต่อาศัยแนวความคิดทั่วไปที่เชื่อว่า หากผู้เชี่ยวชาญมีความคิดเห็นสอดคล้องกันมากเท่าใดการกำหนดคะแนนจุดตัดยิ่งมีความตรงมากเท่านั้น แต่ไม่มีการตรวจสอบโดยข้อมูลเชิงประจักษ์ (Ricker, 2006; Shulruf et al., 2016) หากว่าผู้วิจัยต้องการจะตรวจสอบความถูกต้องในการตัดสินใจจะต้องทำการวิจัยบางอย่างเพิ่มขึ้น เช่น ตรวจสอบความคลาดเคลื่อนมาตรฐานของการตัดสิน (Standard Error of Judging) ของผู้เชี่ยวชาญหลายชุดจากการทดสอบของแบบทดสอบชุดเดียวกันหลายครั้ง หรือจากแบบทดสอบคู่ขนาน (MacCann & Stanley, 2004) หรือด้วยวิธี Cluster Analysis (Tseng et al, 2015) มากกว่าเรื่องความคงเส้นคงวาของการตัดสินใจที่ได้จากเรื่องความสอดคล้องของความคิดเห็นของผู้เชี่ยวชาญ หรืออาจใช้คะแนนจริง (true score) จากการวิเคราะห์ข้อทดสอบรายข้อตามแนวทฤษฎีการตอบสนองของข้อทดสอบ (Item Response Theory) หรือวิเคราะห์ข้อมูลจากความเห็นของผู้เชี่ยวชาญตามแนวคิดของทฤษฎีอ้างอิงสรุป (Generalizability Theory) ก็จะมีประโยชน์มากขึ้นในการหาค่าความตรงของคะแนนกำหนดจุดมากกว่าการที่ไม่ได้ทำได้ (MacCann & Stanley, 2006; Schuwirth & van der Vleuten, 2006 อ้างถึงใน Shulruf et al., 2016)
10. **งานวิจัยทั้ง 2 เรื่องที่ใช้ Modified Angoff Method เป็นเพียงกรณีศึกษาเท่านั้น ไม่สามารถอ้างอิงสรุป (generalize) ผลการวิจัยไปยังแบบทดสอบชุดอื่นได้** เพราะว่า 1) ทำการวิเคราะห์แบบทดสอบดังกล่าวเพียงชุดเดียวเท่านั้น แม้ว่างานวิจัย

เรื่องที่ 2 จะมีข้อตกลงเบื้องต้นว่าแบบทดสอบสร้างขึ้นตาม Test Specification และมีทีมผู้สร้างตรวจสอบ รวมทั้งมีการวิเคราะห์หาคุณภาพรายข้อหลังการใช้จริงแล้ว แต่วิธีการดังกล่าวไม่ได้รับประกันว่าแบบทดสอบชุดอื่นจะเป็นแบบทดสอบคู่ขนานจริง (true parallel forms) กับชุดที่นำมาวิเคราะห์ในเชิงเนื้อหา ระดับความยาก-ง่าย และความแปรปรวนของคะแนน 2) ทำการศึกษาโดยใช้คะแนนปรากฏ (observed scores) ไม่ได้ใช้คะแนนจริง (true scores) เลย และ 3) ไม่มีการตรวจสอบด้วยการกระทำซ้ำ (iteration/replication) เช่น ใช้ผู้เชี่ยวชาญเพียงกลุ่มเดียว ทำการวิเคราะห์แบบทดสอบเพียงชุดเดียว และใช้วิธีการกำหนดจุดตัดของคะแนนเพียงวิธีเดียว เนื่องจากมีงานวิจัยพบว่า ผู้เชี่ยวชาญต่างกลุ่มกำหนดจุดตัดของแบบทดสอบชุดเดียวกันต่างกันแม้ว่าจะใช้วิธีเดียวกัน (Chang, 1999; Ricker, 2006)

11. **ผู้เชี่ยวชาญแต่ละคนไม่มีความเป็นอิสระในความคิดเห็นของตนเอง** เพราะว่า ในการประชุมรอบที่ 3 เพื่อพิจารณาผลการประเมินรายข้อ และคะแนนจุดตัดของแต่ละระดับความสามารถของงานวิจัยทั้ง 2 เรื่องอาจมีผู้เชี่ยวชาญบางคนมีอิทธิพลในการชักจูงผู้เชี่ยวชาญอื่นให้เห็นคล้อยตามได้ด้วย เช่น ผู้ที่มีวุฒิทางการศึกษาสูงมาก (ปริญญาเอก) มีอาวุโสมาก หรือมีประสบการณ์ในการสอนมาก จึงทำให้ความคิดเห็นของผู้เชี่ยวชาญแต่ละคนไม่เป็นอิสระ (Rezigalla, 2016)

สรุป

จากข้อบกพร่องต่างๆดังกล่าวแล้วข้างต้นของงานวิจัยทั้ง 2 เรื่อง ทำให้สามารถสรุปดังต่อไปนี้

1. คะแนนเกณฑ์จุดตัดของแบบทดสอบที่เทียบกับระดับต่างๆตามเกณฑ์ของ CEFR ยังไม่มีความตรง (Validity) ที่ถูกต้อง เนื่องจากยังขาดการตรวจสอบโดยวิธีใดวิธีหนึ่งที่น่าเชื่อถือได้
2. ผลการวิจัยทั้ง 2 เรื่องเป็นเพียงกรณีศึกษา (Case Study) เท่านั้น เนื่องจากอาศัยความคิดเห็นของผู้เชี่ยวชาญเพียงกลุ่มเดียว วิเคราะห์แบบทดสอบเพียงชุดเดียว วิเคราะห์โดยวิธีเดียวเท่านั้น และใช้เพียงสถิติเชิงพรรณา(Descriptive Statistics) เท่านั้น ดังนั้นผลการวิจัยยังไม่สามารถใช้เพื่อการอ้างอิงสรุป (generalize) ไปยังแบบทดสอบชุดอื่นได้

ข้อเสนอแนะสำหรับผู้วิจัยและผู้จะนำผลการวิจัยไปใช้

1. ในการเผยแพร่ผลการวิจัย ผู้วิจัยทั้ง 2 เรื่องควรระบุชื่อเรื่องของการวิจัยว่าเป็นกรณีศึกษา และบอกข้อจำกัดในการที่จะนำผลการวิจัยไปใช้ให้ชัดเจนในตอนต้นๆของรายงานวิจัย แทนที่จะอยู่ในตอนท้ายๆของรายงานอย่างที่นักวิจัยบางคนชอบกระทำ ซึ่งอาจถือได้ว่ามีความพยายามปกปิดความบกพร่องของงานวิจัยของตนเอง
2. ผู้วิจัยทั้ง 2 เรื่องควรต้องตรวจสอบความคงที่คงวา (Consistency) หรือความเที่ยง (Reliability) ของความคิดเห็นของผู้เชี่ยวชาญมากกว่าชุดเดียว กับแบบทดสอบมากกว่าหนึ่งชุด และใช้วิธีกำหนดจุดตัดของคะแนนมากกว่าหนึ่งวิธีเพื่อเปรียบเทียบผลการวิเคราะห์ และควรเข้าใจด้วยว่าค่าคะแนนต่างๆที่ผู้เชี่ยวชาญเห็นสอดคล้องกันนั้นไม่ใช่ค่าความตรง แต่เป็นค่าความเที่ยง ซึ่งแม้ว่าค่าความเที่ยงจะสูงมากเท่าใดก็ตามไม่ใช่สิ่งที่จะรับประกันว่าผลการวิจัยจะถูกต้อง
3. ผู้วิจัยทั้ง 2 เรื่องควรหาวิธีการตรวจสอบความตรง (validity) ของผลการวิจัยกับข้อมูลเชิงประจักษ์หลายๆครั้งด้วยวิธี Modified Angoff Method ที่เคยใช้ กับแบบทดสอบเดียวกันแต่ต่างชุด และผู้เชี่ยวชาญต่างกลุ่ม แล้วเปรียบผลการวิเคราะห์กับวิธีอื่นในการกำหนดจุดตัดของคะแนนที่ได้รับการพัฒนาขึ้นใหม่โดยใช้สถิติอ้างอิงชั้นสูง (Advanced Inferential Statistics) เช่น Cluster Analysis, Item Response Theory, Generalizability Theory และ Confirmatory Factor Analysis เป็นต้น เพราะมีงานวิจัยในทำทองเดียวกันนี้บางเรื่องใช้เวลาในการตรวจสอบผลการวิจัยถึง 5 ปี (MacCann & Stanley, 2004**)**
4. เนื่องจากเกณฑ์ CEFR ที่ใช้ในการวิจัยครั้งนี้เป็นเกณฑ์กลางของความสามารถทางภาษาเพื่อการสื่อสารของผู้ใหญ่ในบริบทของยุโรป (Read, 2019) ไม่ใช่เกณฑ์ความสามารถทางภาษาเพื่อการสื่อสารของนักเรียนหรือนักศึกษาในบริบทของประเทศไทย ดังนั้น การคัดเลือกผู้เชี่ยวชาญที่จะใช้เพื่อการกำหนดจุดตัดควรต้องคำนึงถึงประเด็นนี้เป็นสำคัญ เพราะนอกจากควรหาคณะผู้เชี่ยวชาญที่มีความหลากหลายแล้ว ผู้เชี่ยวชาญควรต้องมีประสบการณ์ตรงตามเกณฑ์ที่นำมาใช้ด้วย หรือหากว่าต้องปรับแก้เกณฑ์ดังกล่าวให้เหมาะสมกับนักเรียนหรือนักศึกษาในบริบทของไทยควรต้องระบุให้ชัดเจนในรายงานการวิจัยด้วย
5. เนื่องจากเกณฑ์ CEFR ที่ใช้ในการวิจัยครั้งนี้เป็นเกณฑ์กลางของความสามารถทางภาษาเพื่อการสื่อสารของผู้ใหญ่ในบริบทของยุโรป (Read, 2019) และข้อทดสอบของแบบทดสอบที่นำมาวิเคราะห์ครั้งนี้ไม่ได้สร้างขึ้นมาตามแนวคิดของ CEFR ตั้งแต่แรก ดังนั้นจึงจำเป็นอย่างยิ่งที่ผู้วิจัยควรต้องตรวจสอบความตรงเชิงเนื้อหา (Content Validity) ของข้อทดสอบรายข้อของแบบทดสอบที่จะนำมาวิเคราะห์เสียก่อนว่าวัดตรงตามเกณฑ์ของ CEFR หรือไม่ และวัดอยู่ในระดับใด โดยการหาค่าดัชนีความสอดคล้อง

ของข้อสอบรายข้อกับวัตถุประสงค์หลายอย่าง (Multidimensional Item-Objective Congruence Index) ตามวิธีของ Turner & Carlson (2003) ไม่ใช่โดยวิธีหาค่า IOC ที่นิยมใช้กันอย่างแพร่หลายในวงการศึกษาไทย ซึ่งมีที่มาของสูตรที่ไม่ถูกต้อง (จักรกฤษณ์ สำราญใจ, 2554)

6. การกำหนดคะแนนเกณฑ์มาตรฐานของแบบทดสอบแต่ละชุดควรคำนึงถึงนโยบายของการทดสอบด้วย ไม่ควรที่จะใช้คะแนนที่ได้จากการคำนวณเพียงอย่างเดียวเท่านั้น คะแนนเกณฑ์มาตรฐานไม่ควรจะต่ำจนเกินไป หรือสูงมากจนเกินไป แต่ควรจะเป็นจุดที่เป็นไปได้และพอดีระหว่างคะแนนที่กำหนด (cut scores) กับมาตรฐานของความสามารถจลย์ (performance standard) หรือความสามารถจริงในการใช้ภาษาของผู้สอบในแต่ละระดับความสามารถ (Kane, 2017)
7. นักวิจัยที่ประสงค์จะทำการวิจัยเกี่ยวกับการกำหนดระดับเกณฑ์มาตรฐานของแบบทดสอบชุดใดก็ตาม ควรจะต้องตระหนักรู้ว่าวิธีการต่างๆมีอยู่จำนวนมาก แต่ละวิธีให้ผลการวิจัยที่แตกต่างกัน หรือแม้แต่วิธีเดียวกัน เมื่อเปลี่ยนกลุ่มผู้เชี่ยวชาญในการให้ข้อมูล ก็มักจะได้ผลที่ไม่คงที่เหมือนเดิม (Kane, 2017) ทั้งนี้เพราะว่า วิธีต่างๆที่อาศัยความเห็นของผู้เชี่ยวชาญล้วนแต่เป็นการกำหนดเกณฑ์คะแนนแบบตามใจชอบ (arbitrary) จึงควรต้องมีการตรวจสอบความตรงของผลการวิจัยให้แน่ใจก่อนนำไปเผยแพร่ในวงกว้าง เพราะว่าทุกครั้งที่มีการใช้เกณฑ์จะเกิดผล 2 อย่าง คือ คนที่ผ่านเกณฑ์จะได้ประโยชน์ ส่วนคนที่ไม่ผ่านเกณฑ์จะเสียประโยชน์ การที่ผลการวิจัยทำให้คนไม่ผ่านเกณฑ์ทั้งๆที่เขาควรผ่านเกณฑ์เป็นการตัดสินผิดพลาดมาก (Kane, 2017)
8. สำหรับผู้ที่จะนำผลการวิจัยของงานวิจัยเรื่องใดเรื่องหนึ่งดังกล่าวแล้วไปใช้ ควรต้องพิจารณาเรื่องข้อบกพร่องต่างๆดังกล่าวแล้วข้างต้น แล้วพิจารณาว่าผลการวิจัยสอดคล้องกับบริบทของตนหรือไม่ และอาจเสี่ยงต่อการตัดสินใจมากน้อยเพียงใดต่อนักเรียน หรือผู้ที่นำผลการวิจัยมาใช้อ้างอิง เพราะว่าหากมีคดีฟ้องร้องในศาล ศาลจะไม่รับฟังความเห็นของผู้เชี่ยวชาญในการวิเคราะห์คะแนนจุดตัด แต่ศาลจะพิจารณาความตรงของข้อสอบกับหลักสูตรหรือสิ่งที่มุ่งทดสอบ (Mehrens & Popham, 1992; Sireci & Parker, 2006) ดังนั้นความตรงเชิงเนื้อหา (Content Validity) จึงมีความสำคัญมาก และสำคัญมากกว่าความเที่ยง (Reliability) หรือความสอดคล้อง (Agreement) ของความคิดเห็นของผู้เชี่ยวชาญในศาสตร์

บรรณานุกรม

กระทรวงศึกษาธิการ. (2557). **นโยบายการปฏิรูปการเรียนการสอนภาษาอังกฤษ.** Retrieved from http://english.obec.go.th/english/2013/index.php/th/2012-08-08-10-26-5/60-2014-04-05-08-29-13

จักรกฤษ์ สำราญใจ. (2554). *IOC = ความตรง?* **วารสารหลักสูตรและการสอน** มหาวิทยาลัยขอนแก่น ปีที่ 4 ฉบับที่ 1-2. Retrieved from https://www.scribd.com/doc/86608731/IOC

สำนักงานอุดมศึกษา. (2559). **นโยบายการยกระดับมาตรฐานภาษาอังกฤษในสถาบันอุดมศึกษา.** Retrieved from http://www.mua.go.th/users/bhes/front_home/Data%20Bhes_2559/04052559.pdf

Angoff, W. H. (1971). *Scales, norms, and equivalent scores*. In R. L. Thorndike (Ed.), **Educational Measurement** (pp. 508-600). Washington, DC: American Council on Education.

Assessment Strategies Inc. (2014). **The Angoff Method of Standard Setting**. Retrieved from https://bcrsp.ca/sites/default/files/documents/Angoff%20Method%20Article_1.pdf

Barman, A. (2008). *Standard setting in Student assessment: Is a Defensible Method Yet to Come?* **Annals of the Academy of Medicine Journal**. Retrieved from https://www.semanticscholar.org/paper/Standard-setting-in-student-assessment%3A-is-a-method-Barman/06dc3ba6d25568d904196d952233eeafe524a3ef

Bejar, I.I. (2008). *Standard Setting: What Is It? Why Is It Important?* **R & D Connections**, No. 7. Retrieved from https://www.ets.org/Media/Research/pdf/RD_Connections7.pdf

Berk, R. A. (1986). ***A Consumer's Guide to Setting Performance Standards on Criterion-Referenced Tests***. **Journal of Review of Educational Research**, Vol. 56, No. 1. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.858.9807&rep=rep1&type=pdf

Burrr, S.A.; Zahra, D.; Cookson, J. & Salih, V. M. (2017). **Angoff anchor statements: setting a flawed gold standard?** The Association for Medical Education in Europe (AMEE). Retrieved from https://www.mededpublish.org/manuscripts/1201

CaMLA. (2015). *Linking the Common European Framework of Reference and the CaMLA Speaking Test*. **Technical Report**. Cambridge Michigan Language Assessments. Retrieved from https://michiganassessment.org/wp-content/uploads/2014/12/EPT-Technical-Report-20140625.pdf

Chinn, R.N. (2006). **Considerations in Setting Cut Scores**. Council on Licensure, Enforcement and Regulation (CLEAR) Retrieved from https://www.clearhq.org/resources/Cut_Scores_RB_2006.pdf

Cizek, G. J. (1993). **Reconsidering Standards and Criteria**. Retrieved from https://www.jstor.org/stable/1435458?seq=1#page_scan_tab_contents

DeMauro, G.E. & Powers, D. E. (1993). **Logical Consistency of the Angoff Method of Standard Setting**. Retrieved from https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.1993.tb01537.x

Eckes, T. (2012). *Examinee-centered standard setting for large-scale assessments: The prototype group method.* **Journal of Psychological Test and Assessment Modeling**, Volume 54, (3), 257-283. Retrieved from https://www.testdaf.de/fileadmin/Redakteur/PDF/Forschung-Publikationen/Eckes_PTAM_2012-3.pdf

George, S., Haque, M.S. & Oyebode, F. (2006). *Standard Setting: Comparison of two methods*. **BMC Medical Education**. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1578558/

Hambleton, R.K., Powell, S. & Eignor, D.R. (1979). **Issues and Methods for Standard Setting**. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.853.1268&rep=rep1&type=pdf

Hejri, S.M. & Jalili, M. (2014). *Standard setting in medical education: fundamental concepts and emerging challenges*. **Medical Journal of the Islamic Republic of Iran**. Retrieved from

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4153516/

Hurtz, G.M. & Auerbach, M.A. (2003). *A Meta-Analysis of the Effects of Modifications to the Angoff Method on Cutoff Scores and Judgment Consensus*. **Journal of Educational and Psychological Measurement**, Vol. 63, Issue 4. Retrieved from https://journals.sagepub.com/doi/abs/10.1177/0013164403251284

Kane, M.T. (2017). *Using Empirical Results to Validate Performance Standards*. In Blomeke, S. and Gustafsson, J.E. (Eds). **Standard Setting in Education**. Springer, 2017.

Kulaporn Hiranburana et al. (2019). **Framework of Reference for English Language Education in Thailand (FRELE-TH) Based on the CEFR: Revisited in English Educational Reform**. Retrieved from https://www.researchgate.net/publication/331895094_FRELE-TH_Revisited/link/5c9219e6a6fdccd4602c2031/download

Lei Chang, L. (1999). **Judgmental Item Analysis of the Nedelsky and Angoff Standard-Setting Methods**. Retrieved from https://www.tandfonline.com/doi/abs/10.1207/s15324818ame1202_3

MacCann, R.G. & Stanley, G. (2004). *Estimating the Standard Error of the Judging in a modified-Angoff Standards Setting Procedure*. **Journal of Practical Assessment Research & Evaluation**, Vol. 9, No. 5. Retrieved from https://www.researchgate.net/publication/292661791_Estimating_the_standard_error_of_the_judging_in_a_modified-Angoff_standards_setting_procedure

MacCann, R.G. & Stanley, G. (2006**).** *The Use of Rasch Modeling to Improve Standard Setting***. Journal of Practical Assessment Research & Evaluation**, Vol. 11, No. 2. Retrieved from https://pareonline.net/pdf/v11n2.pdf

Mehrens, W.A. & Popham, W.J. (1999). *How to Evaluate the Legal Defensibility of High-Stakes Tests*. **Applied Measurement in Education** 5(3):265-283 July 1992. Retrieved from https://www.researchgate.net/publication/248940643_How_to_Evaluate_the_Legal_Defensibility_of_High-Stakes_Tests

Morgan, D.L. & Michaelides, M.P. (2005). **Setting Cut Scores for College Placement**. Retrieved from https://eric.ed.gov/?id=ED562865

Plake, B.S., Impara, J.C., Spines, R., Hertzog, M. & Giraud, G. (1998). **Setting Performance Standards on Polytomously Scored Assessments: An Adjustment to the Extended Angoff Method**. Retrieved from https://eric.ed.gov/?id=ED422355

Read, J. (2019). *The Influence of the Common European Framework of Reference (CEFR) in the Asia-Pacific Region*. **LEARN Journal**, Vol. 12, Issue 1. Retrieved from https://tci-thaijo.org/index.php/LEARN/article/download/168568/121292/

Rezigalla, A.A. (2016). **Angoff's Method: The Impact of Raters' Selection**. Retrieved from http://www.academia.edu/33258465/Angoff_s_method_The_impact_of_raters_selection.pdf

Ricker, K.L. (2006). *Setting Cut-Scores: A Critical Review of the Angoff and Modified Angoff Methods*. **The Alberta Journal of Educational Research**, Vol. 52, No. 1. Retrieved from https://journalhosting.ucalgary.ca/index.php/ajer/article/viewFile/55111/42163

Schuwirth, L. & van der Vleuten, C. (2006). **A plea for new psychometric models in educational assessment**. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/16573664

Shin, S.Y. & Lidster, R. (2017). *Evaluating different standard-setting methods in an ESL placement testing context*. **Journal of Language Testing**, Vol. 34, No. 3 Retrieved from https://journals.sagepub.com/doi/abs/10.1177/0265532216646605

Shulruf, B., Wilkinson, T. Weller, J. Jones, P. & Poole, P. (2016). **Insights into the Angoff method: Results from a simulation study**. Retrieved from https://www.researchgate.net/publication/301830910_Insights_into_the_Angoff_method_Results_from_a_simulation_study

Sireci, S. G. & Parker, P. (2006). *Validity on Trial: Psychometric and Legal Conceptualizations of Validity*. **Educational Measurement: Issues and**

Practice, *25*(3), 27-34. Retrieved from http://dx.doi.org/10.1111/j.1745-3992.2006.00065.x

Stahl, J.A. & VUE, Pearson. (2019). **Standard Setting Methodologies: Strengths and Weaknesses**. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.486.9283&rep=rep1&type=pdf

Thompson, N. (2018). **How do I conduct a modified Angoff study?** Assessment Systems. Retrieved from https://www.assess.com/conduct-a-modified-angoff-study/

Tseng, F.L., Chiou, J.M. & Sung, Y.T. (2515). **A validity study for Yes/No Angoff standard setting method using cluster analysis**. Retrieved from https://www.researchgate.net/publication/304293875_A_validity_study_for_YesNo_Angoff_standard_setting_method_using_cluster_analysis

Turner, R. & Carlson, L. A. (2003). **Indexes of Item-Objective Congruence for Multidimensional Items**. International Journal of Testing 3(2). Retrieved from https://www.researchgate.net/publication/247502723_Indexes_of_Item-Objective_Congruence_for_Multidimensional_Items

Verheggen, M.M., van Os, J, Muijtjens, A, & Schuwirth, L.W. (2008). **Is an Angoff standard an indication of minimal competence of examinees or of judges?** Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/17043915

Wang, N., Muijtjens, A.M., van Os, J. & Schuwirth, L.W. (2003). *Use of the Rasch IRT Model in Standard Setting: An Item-Mapping Method*. **Journal of Educational Measurement**, Vol. 40, No. 3. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-3984.2003.tb01106

Wayne, DB, Fudala, MJ, Butter, J, Siddall, VJ, Feinglass, J, Wade, LD & McGaghie, WC. (2005). **Comparison of two standard-setting methods for advanced cardiac life support training**. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/16199461

Wilson, M. & Santelices, M.V. (2017). *Weakness of the Traditional View of Standard Setting and a Suggested Alternative*. In Blomeke, S. & Gustafsson, J.E. (Eds). **Standard Setting in Education**. Springer, 2017.

Zieky, M. & Perie, M. (2004). **A Primer on Setting Cut Scores on Tests of Educational Achievement**. ETS. Retrieved from https://www.ets.org/Media/Research/pdf/Cut_Scores_Primer.pdf

Zumbo, B.D. (2016). *Standard-setting methodology: Establishing performance standards and setting cut scores to assist score interpretation*. **Journal of Applied Physiology, Nutrition and Metabolism**, Vol. 41, No. 6. Retrieved from https://www.researchgate.net/publication/295105068_Standard-setting_methodology_Establishing_performance_standards_and_setting_cut_scores_to_assist_score_interpretation