

Academic Article

MANY-FACET RASCH MEASUREMENT PARTIAL CREDIT MODEL AND OPTIMIZING
RATING SCALE CATEGORY EFFECTIVENESS FOR RUBRICS:
CONCEPT AND ANALYSIS EXAMPLE

โมเดลการให้คะแนนบางส่วนแบบหลายองค์ประกอบของราสช์และการปรับประสิทธิภาพ
ระดับคะแนนสำหรับรูบริก: แนวคิดและตัวอย่างการวิเคราะห์

Received: October 9, 2024

Revised: December 1, 2024

Accepted: December 11, 2024

Purin Thepsathit^{1*} and Kamonwan Tangdhanakanond²

ภูรินทร์ เทพสถิตย์^{1*} และกมลวรรณ ตังธนากานนท์²

¹Independent Researcher

²Faculty of Education, Chulalongkorn University

¹นักวิชาการอิสระ

²คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

*Corresponding Author, E-mail: purin.tepsatit@gmail.com

Abstract

To develop rubrics through the analysis of The Many-Facet Rasch Measurement Partial Credit Model (MFRM-PCM) and optimizing rating scale category effectiveness benefits in helping determine an appropriate number of score levels in rubrics development. Rubrics development through this model allows for the investigation of validity and reliability by analyzing more than two facets related to assessment. This is suitable for performance assessments where raters are involved in assessment results. This academic article is written to provide knowledge and understanding to teachers and educators about the concept of MFRM-PCM and optimizing rating scale category effectiveness analysis, offering examples of analysis and interpretation. The aim is to help readers apply this in rubrics development in a school or institution context. Developing rubrics prepares teachers for competency-based learning, where learners demonstrate learning outcomes through performance or work that reflects their knowledge, attitudes, and skills. It also assists university

instructors in aligning their teaching with learning assessment standards in higher education, which emphasizes developing learners to achieve expected learning goals through assessment and feedback.

Keywords: Many-Facet Rasch Measurement Partial Credit Model, Performance Assessment, Educational Statistics, Rating Scale Category Effectiveness, Rubrics

บทคัดย่อ

การพัฒนา RUBRIC ผ่านการวิเคราะห์ด้วยโมเดลการให้คะแนนบางส่วนแบบหลายองค์ประกอบของ Rasch และการปรับปรุงประสิทธิภาพระดับคะแนน มีประโยชน์ในการช่วยตัดสินใจจำนวนระดับคะแนนของเกณฑ์การประเมินที่เหมาะสมในกระบวนการพัฒนา RUBRIC ได้ การพัฒนา RUBRIC ผ่านการวิเคราะห์ด้วยโมเดลนี้ สามารถตรวจสอบความตรงและความเที่ยงผ่านการวิเคราะห์ฟิตที่เกี่ยวกับการประเมินได้มากกว่า 2 ฟิต ซึ่งเหมาะสมกับการประเมินภาคปฏิบัติที่มีผู้ประเมินเข้ามาเกี่ยวข้องในการให้ผลการประเมิน บทความวิชาการนี้เขียนขึ้นเพื่อให้ความรู้และความเข้าใจกับครูหรือนักการศึกษาโดยทั่วไปเกี่ยวกับแนวคิดของการวิเคราะห์การปรับปรุงประสิทธิภาพระดับคะแนนและให้ตัวอย่างการวิเคราะห์และแปลผลโดยมีเป้าหมายในการให้ผู้อ่านสามารถนำไปประยุกต์ใช้ในการพัฒนา RUBRIC ในบริบทโรงเรียนหรือสถานศึกษาได้ การพัฒนา RUBRIC ช่วยให้ผู้เตรียมพร้อมสำหรับการจัดการเรียนรู้ฐานสมรรถนะที่จะประเมินผู้เรียนผ่านการปฏิบัติหรือผลงานที่สะท้อนความรู้ เจตคติ และทักษะของผู้เรียน และอาจารย์ระดับมหาวิทยาลัยที่ต้องจัดการเรียนรู้ให้สอดคล้องกับมาตรฐานการประเมินการเรียนรู้ในระดับอุดมศึกษาที่เน้นการพัฒนาผู้เรียนให้บรรลุเป้าหมายการเรียนรู้ที่คาดหวัง ผ่านการประเมินโดยใช้ RUBRIC และการให้ข้อมูลย้อนกลับ

คำสำคัญ: โมเดลการให้คะแนนบางส่วนแบบหลายองค์ประกอบของ Rasch การประเมินภาคปฏิบัติ สถิติทางการศึกษา การปรับปรุงประสิทธิภาพระดับคะแนน, RUBRIC

บทนำ (Introduction)

ในการประเมินภาคปฏิบัติ เครื่องมือทางการศึกษาที่มักถูกนำมาใช้ในการประเมินทักษะของผู้เรียนคือ RUBRIC (Rubrics) ซึ่งเป็นเครื่องมือที่เหมาะสมในการประเมินผลลัพธ์การปฏิบัติของผู้เรียนที่มีความซับซ้อนหรือมีลักษณะเป็นปลายเปิด (Open-ended) โดย RUBRIC มีประโยชน์ในการช่วยระบุคุณภาพของทักษะในแต่ละเกณฑ์การประเมิน (Criteria) ผ่านบรรยายลักษณะ (Description) ที่คาดหวังให้ผู้เรียนต้องแสดงออก ในระดับคะแนน (category/level) ที่แตกต่างกัน (Jönsson, 2022; Panadero & Jonsson, 2013; Popham, 1997; Stevens & Levi, 2012) RUBRIC ทำให้การประเมินมีความชัดเจน และช่วยให้ผู้เรียนรับรู้เกณฑ์การประเมินที่ใช้ในการประเมินทักษะของตนเอง การทำความเข้าใจข้อมูลย้อนกลับ (Feedback) ในข้อความบรรยายลักษณะที่คาดหวังของแต่ละระดับคะแนนใน RUBRIC ช่วยให้ผู้เรียนสามารถประเมินผลงานของตนเอง (Self-Assessment) และใช้ข้อมูลย้อนกลับดังกล่าวในการพัฒนาผลงานของตนเองให้บรรลุเกณฑ์การประเมินในระดับสูงขึ้นได้ การใช้ RUBRIC ในการประเมินตนเอง ช่วยให้ผู้เรียนสามารถเห็นช่องว่างระหว่างเป้าหมายการเรียนรู้ที่คาดหวังกับความสามารถตนเอง ช่วยให้ผู้เรียนมีปฏิสัมพันธ์กับการเรียนรู้และพัฒนากระบวนการเรียนรู้ควบคู่ไปกับการพัฒนาผลงานของตนเอง ซึ่งการประเมินตนเองโดยใช้ข้อมูลย้อนกลับถือเป็นองค์ประกอบสำคัญของการประเมินเพื่อความก้าวหน้าหรือพัฒนา (Formative Assessment) ซึ่งสามารถนำไปใช้ได้ทั้งในการศึกษาระดับการศึกษาขั้นพื้นฐานและอุดมศึกษา (Brookhart & Chen, 2015; Furtak, 2022; Harris & Brown, 2022; Jönsson, 2022; Pancorbo et al., 2021)

สำหรับประเภทของรูบริกจากการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง พบว่า รูบริกที่มักถูกนำมาใช้และมีการศึกษามากที่สุดคือ รูบริกแบบองค์รวม (holistic rubrics) และรูบริกแบบแยกองค์ประกอบ (Analytic Rubrics) (Brookhart, 2018; Jönsson, 2022) โดยรูบริกแบบองค์รวมเน้นการประเมินภาพรวมของผลงาน โดยในแต่ละระดับของคำอธิบายคุณภาพจะประกอบไปด้วยคุณภาพองค์ประกอบต่างๆ ที่ควรปรากฏในผลงาน ส่วนรูบริกแบบแยกองค์ประกอบจะแยกองค์ประกอบต่างๆ ออกมาเป็นข้อๆ และประเมินคุณภาพของแต่ละองค์ประกอบแยกออกจากกัน รูบริกแบบองค์รวมมีความเหมาะสมกับการประเมินแบบสรุปรวม (Summative) หรือการประเมินขนาดใหญ่ (Large-scale) เนื่องจากสามารถประเมินได้อย่างถูกต้องในเวลาที่รวดเร็ว ในขณะที่รูบริกแบบแยกองค์ประกอบเหมาะสำหรับการประเมินในห้องเรียนที่เป็นการประเมินเชิงพัฒนา เนื่องจากสามารถให้ข้อมูลย้อนกลับในการพัฒนาผู้เรียนได้ละเอียด ระบุจุดแข็งหรืออ่อนของทักษะผู้เรียนได้ แต่ต้องใช้เวลาในการประเมินมากกว่า (Jonsson & Svingby, 2007; Jönsson, 2022) โดยส่วนประกอบที่สำคัญของรูบริกมี 3 ส่วน ได้แก่ เกณฑ์การประเมินระดับคะแนน และคำบรรยายคุณภาพ โดยเกณฑ์การประเมินเป็นการระบุถึงประเด็นสำคัญหรือสิ่งที่ต้องสังเกตจากงานและ/หรือการปฏิบัติงาน โดยอาจมีได้ตั้งแต่ 1 ประเด็นขึ้นไป ระดับคะแนนหรือคุณภาพเป็นจำนวนระดับคุณภาพของงานหรือการปฏิบัติงานที่คาดหวังในแต่ละเกณฑ์การประเมินจากน้อยไปมาก ส่วนคำบรรยายคุณภาพเป็นคำอธิบายที่สะท้อนคุณภาพงานและ/หรือการปฏิบัติงานในแต่ละระดับคะแนนของเกณฑ์การประเมิน (Brookhart, 2018; Moskal, 2000; Popham, 1997)

อย่างไรก็ตาม ความท้าทายประการหนึ่งของการพัฒนารูบริกคือ การกำหนดจำนวนระดับคะแนนที่เหมาะสมในเกณฑ์การประเมินแต่ละข้อ โดยจากการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องกับการพัฒนารูบริก พบว่า ไม่มีกฎเกณฑ์ในการกำหนดจำนวนระดับคะแนนที่ชัดเจน (Moskal, 2000; Reddy, 2011) ผู้พัฒนารูบริกสามารถเริ่มจากการกำหนดจำนวนระดับคะแนนหรือระบุคำอธิบายคุณภาพงานหรือการปฏิบัติงานก่อนก็ได้ โดยผู้พัฒนารูบริกต้องวิเคราะห์เกณฑ์การประเมิน เพื่อพิจารณาว่าเกณฑ์การประเมินนั้นมีตัวชี้วัดคุณภาพอะไรบ้างที่ต้องสังเกตในการประเมิน เพื่อทำเป็นโครงสร้าง (Outline) ทักษะตามเกณฑ์การประเมิน หรือสามารถกำหนดความเข้มในการจำแนกความแตกต่างของทักษะผู้เรียนที่ต้องการประเมิน หรือใช้การกำหนดความคาดหวังต่อทักษะนั้นๆ เช่น กำหนดความคาดหวังให้ประกอบด้วย ดีมาก ดี พอใช้ ควรปรับปรุง และจากนั้นจึงระบุคำบรรยายคุณภาพงานหรือการปฏิบัติงานในแต่ละระดับ โดยมีการเสนอว่า จำนวนระดับคะแนนควรมีตั้งแต่ 2 ระดับขึ้นไป แต่ไม่เกิน 6 ระดับ ซึ่งโดยทั่วไปแล้วจะอยู่ในช่วง 3 ถึง 5 ระดับ (Brookhart, 2018; Clarkson & Luca, 2001; Griffith & Lim, 2012; Klimovski, 2013; Moskal, 2000; Popham, 1997)

จากความท้าทายในการกำหนดจำนวนระดับคะแนนข้างต้น จะเห็นได้ว่าผู้พัฒนารูบริกจะต้องวิเคราะห์และตัดสินใจว่าจำนวนระดับคะแนนของเกณฑ์การประเมินภายในรูบริกของตนเองจะมีจำนวนกี่ระดับ ซึ่งบางครั้งการกำหนดจำนวนระดับมากไปอาจส่งผลกระทบต่อผลการประเมิน เช่น ทำให้ผู้ประเมินสับสนและรบกวนการตัดสินใจ และยากต่อการให้ผลการประเมินอย่างคงเส้นคงวา (Thao, 2023) นอกจากนี้ยังมีประเด็นที่นักวิชาการนำเสนอว่า รูบริกสามารถมีระดับคะแนนที่แตกต่างกันตามเกณฑ์การประเมินแต่ละข้อ ซึ่งสามารถช่วยลดความลำเอียงหรืออคติ (Rater Effect) ในการประเมินของผู้ประเมินได้ และทักษะบางอย่างมีลักษณะเพียงแค่ 2 ระดับเท่านั้น เช่น ทำหรือไม่ทำ เตรียมพร้อมหรือไม่เตรียมพร้อม เป็นต้น (Johnson et al., 2009; Jonsson & Panadero, 2017; Linacre, 2002) ดังนั้น จากที่กล่าวมาข้างต้น จะเห็นได้ว่า ขั้นตอนการกำหนดระดับคะแนนในรูบริกมีความหลากหลาย ขึ้นอยู่กับผู้พัฒนาและลักษณะของทักษะที่ต้องการประเมิน และหากระดับคะแนนมีจำนวนที่มากไปก็อาจส่งผลกระทบต่อผลการประเมินหรือหากน้อยเกินไปก็อาจจำแนกทักษะตามที่ผู้ประเมินต้องการได้

อย่างไรก็ตาม ในการวัดและประเมินผลทางการศึกษา มีโมเดลการวัดที่ช่วยผู้พัฒนารูบริกในการตัดสินใจเกี่ยวกับจำนวนระดับคะแนนที่เหมาะสมกับการประเมินทักษะได้ คือ โมเดลการให้คะแนนบางส่วนแบบหลายองค์ประกอบของราสช์ (Many-facet Rasch measurement partial credit model: MFRM-PCM) โดยใช้การวิเคราะห์การปรับประสิทธิภาพของระดับคะแนน (Rating scale category effectiveness) ของ Linacre (2002) (Thepsathit & Tangdhanakanond, 2024) ซึ่งเป็น

การวิเคราะห์ระดับคะแนนหลังจากการนำเอารูบริกไปใช้ในการประเมินทักษะ และวิเคราะห์ประสิทธิผลของแต่ละระดับคะแนนแยกเป็นรายข้อผ่านดัชนีที่เกี่ยวข้องต่างๆ เพื่อตรวจสอบระดับคะแนนที่มีประสิทธิผลในการประเมินทักษะ ช่วยให้ผู้พัฒนามีข้อมูลเชิงปริมาณสำหรับช่วยในการตัดสินใจแก้ไขหรือปรับปรุงรูบริกสำหรับการนำไปใช้ (Linacre, 1994a; Linacre 2002; Pancorbo et al., 2021) โดยโมเดลการให้คะแนนบางส่วนแบบหลายองค์ประกอบของราสช์และการวิเคราะห์เพื่อปรับประสิทธิผลของระดับคะแนนมีรายละเอียดดังต่อไปนี้

โมเดลการให้คะแนนบางส่วนแบบหลายองค์ประกอบของราสช์ (Many-Facet Rasch Measurement Partial Credit Model)

โมเดลการให้คะแนนบางส่วนแบบหลายองค์ประกอบของราสช์ เป็นโมเดลที่พัฒนาต่อเนื่องมาจากโมเดลหลายองค์ประกอบของราสช์ (Many-facet Rasch measurement model) ที่เป็นโมเดลตระกูลราสช์ (Rasch model) ที่สามารถวิเคราะห์ฟาเซตได้ตั้งแต่ 3 ฟาเซตขึ้นไป โดยคำว่าฟาเซต หมายถึง องค์ประกอบที่เกี่ยวข้องกับการประเมิน เช่น ผู้เรียน เครื่องมือประเมิน หรือผู้ประเมิน โมเดลการให้คะแนนบางส่วนแบบหลายองค์ประกอบของราสช์มีสมการดังนี้ (Linacre, 1994a)

$$\ln\left(\frac{P_{nikj}}{1 - P_{nikj}}\right) = B_n - D_{ik} - C_j$$

เมื่อ $\ln(P_{nikj}/1 - P_{nikj})$ คือ ความน่าจะเป็นในการถูกประเมินโดยผู้ประเมิน j ด้วยระดับคะแนน k ในข้อคำถาม i ของผู้สอบ n ซึ่งสามารถอธิบายได้จาก $B_n - D_{ik} - C_j$ โดยที่ B_n คือ ความสามารถของผู้เรียนคนที่ n D_{ik} คือ ความยากของระดับคะแนนที่ k ในข้อคำถามที่ i และ C_j คือ การกดคะแนนของผู้ประเมินคนที่ j

จากสมการโมเดลดังกล่าว จะเห็นได้ว่าโมเดลนี้ จะคำนวณความน่าจะเป็นของการถูกประเมินโดยรวมเอาความสามารถของผู้เรียน ความยากของระดับคะแนนบนเครื่องมือ และการกดคะแนนของผู้ประเมินเข้ามาคำนวณร่วมกันโดยตรง ช่วยให้การวิเคราะห์เพื่อพัฒนาเครื่องมือการประเมินมีความตรงและความเที่ยงผ่านการวิเคราะห์ฟาเซตที่เกี่ยวข้องทั้งหมด รวมถึงสามารถวิเคราะห์การกดคะแนนของผู้ประเมินและค่าสถิติระดับคะแนนแยกรายข้อได้อีกด้วย (Eckes, 2009; Eckes, 2019; Myford & Wolfe, 2004; Thepsathit & Tangdhanakanond, 2024) โดยโมเดลการให้คะแนนบางส่วนแบบหลายองค์ประกอบนี้จะแตกต่างจากโมเดลหลายองค์ประกอบแบบปกติตรงสมการทางด้านขวา ซึ่งโมเดลหลายองค์ประกอบจะอธิบายความน่าจะเป็นของสมการทางด้านซ้ายด้วย $B_n - D_i - F_k - C_j$ ซึ่งเป็นการคำนวณความน่าจะเป็นโดยแยกความยากของข้อคำถามและความยากระดับคะแนนออกจากกัน (Linacre, 1994a) ดังนั้น ในโมเดลหลายองค์ประกอบแบบปกติ ข้อคำถามแต่ละข้อจะมีความยากของระดับคะแนนเท่ากัน คล้ายกับรูปแบบของเครื่องมือประเภทแบบวัดมาตราประมาณค่า (rating scale) แตกต่างจากโมเดลการให้คะแนนบางส่วนแบบหลายองค์ประกอบที่ข้อคำถามแต่ละข้อจะมีความยากของคะแนนแตกต่างกัน คล้ายกับลักษณะของเครื่องมือประเภทรูบริก

โมเดลการให้คะแนนบางส่วนแบบหลายองค์ประกอบของราสช์ จึงถูกนำมาวิเคราะห์เครื่องมือประเภทที่ให้ผลการตอบมากกว่าสองค่า (polytomous item) ที่ต้องอาศัยผู้ประเมินในการประเมิน ซึ่งในสถานการณ์ดังกล่าว มักจะเป็นการประเมินภาคปฏิบัติที่ใช้รูบริกและผู้ประเมินในการประเมินความสามารถในการปฏิบัติของผู้เรียน ซึ่งอาจเป็นได้ทั้งผลการปฏิบัติในการเรียนรู้ เช่น ดนตรี กีฬา การแสดง การเขียน หรือการนำเสนอผลงาน เป็นต้น และจากสมการของโมเดล ที่มีการกำหนดระดับของความยากเป็นรายข้อ จึงเหมาะสมกับรูบริกที่มีเกณฑ์การประเมินหลายข้อ ในลักษณะของรูบริกแบบแยกองค์ประกอบโดยรายละเอียดของแนวคิดและค่าสถิติอื่นๆ ที่ใช้ในการวิเคราะห์โมเดลหลายองค์ประกอบ สามารถศึกษาเพิ่มเติมได้จากงานของ Thepsathit and Tangdhanakanond (2024)

การวิเคราะห์เพื่อปรับประสิทธิภาพระดับคะแนน (An Analysis of Rating Scale Category Effectiveness)

Linacre (2002) ได้นำเสนอแนวทางในการวิเคราะห์เพื่อปรับประสิทธิภาพระดับคะแนนของมาตราประมาณค่า โดยใช้การวิเคราะห์จากสถิติระดับคะแนน (category statistics) จากการวิเคราะห์โมเดลการให้คะแนนบางส่วนแบบหลายองค์ประกอบของราสช์ด้วยโปรแกรม Facets (Linacre 2022) ซึ่งต่อมาเมื่อนักวิจัยได้นำเอาแนวทางดังกล่าวไปประยุกต์ใช้กับเครื่องมือการวัดและประเมินผลที่มีลักษณะเป็นเครื่องมือที่ให้คะแนนได้มากกว่าสองค่า ได้แก่ มาตราประมาณค่าหรือรูบริก ดังตัวอย่างงานวิจัยของ Chapman et al. (2024) Edwards et al. (2019) Hodges et al. (2019) Lallmamode et al. (2016) และ Yuen and Sawatdeenarunat (2020)

การวิเคราะห์เพื่อปรับประสิทธิภาพระดับคะแนน มีเป้าหมายในการศึกษาว่าระดับคะแนนแต่ละระดับของเกณฑ์การประเมินแต่ละข้อมีประสิทธิภาพในการประเมินผลงานหรือการปฏิบัติงานของผู้เรียนมากน้อยเพียงใด และควรพิจารณาปรับปรุงให้มีประสิทธิภาพขึ้นได้อย่างไร โดยพิจารณาจากค่าสถิติจำเป็นอย่างน้อย 4 ค่า ได้แก่ 1) Point Measure Correlation 2) Category Usage 3) Average Measure และ 4) Outfit Mean Square (MS) ซึ่งมีรายละเอียดดังนี้ (Linacre, 2002)

1) ค่า Point Measure Correlation คือ ค่าความสัมพันธ์ของเกณฑ์การประเมินภายในรูบริก ใช้ในการตรวจสอบว่าเกณฑ์การประเมินแต่ละข้อมุ่งวัดตัวแปรเดียวกันหรือไม่ โดยค่าสหสัมพันธ์ดังกล่าวควรมีค่าเป็นบวกในเกณฑ์การประเมินทุกข้อ การผ่านเกณฑ์ค่า Point Measure Correlation มีความจำเป็นในการยืนยันความคงที่ (Stability) และความถูกต้อง (Accuracy) ของผลการวัด รวมถึงเป็นหลักฐานที่จำเป็นในการอธิบายผลการวัดของกลุ่มตัวอย่างและความสามารถในการอ้างอิง (Inference) คุณสมบัติทางจิตมิติของเครื่องมือเมื่อไปใช้กับกลุ่มตัวอย่างอื่นๆ

2) ค่า Category Usage คือ ความถี่ของจำนวนการใช้ระดับคะแนนในการประเมิน โดยเป็นการนับจำนวนของระดับคะแนนแต่ละระดับภายในเกณฑ์การประเมินแต่ละข้อที่ถูกใช้ประเมินโดยผู้ประเมินทั้งหมด โดยแต่ละระดับคะแนนควรมีความถี่ของการถูกใช้ประเมินอย่างน้อย 10 ครั้ง การผ่านเกณฑ์ค่า Category Usage มีความจำเป็นในการยืนยันความคงที่ของผลการวัด

3) ค่า Average Measure คือ ค่าเฉลี่ยของค่าสถิติ Measure ของผู้เรียนในแต่ละระดับคะแนน ซึ่งค่า Measure ของพาเซตผู้เรียน หมายถึง ระดับความสามารถของผู้เรียนในการแสดงทักษะ โดยหากมีค่าน้อย หมายถึง ผู้เรียนมีความสามารถต่ำ ส่วนหากมีค่ามาก หมายถึง ผู้เรียนมีความสามารถสูง โดยค่าเฉลี่ยค่าสถิติ Measure ในแต่ละระดับคะแนนต้องมีค่าเรียงลำดับจากน้อยไปมาก สอดคล้องกับระดับคะแนนของเกณฑ์การประเมิน การเรียงลำดับจากน้อยไปมากของค่า Average Measure ในแต่ละระดับคะแนน มีความจำเป็นในการยืนยันความถูกต้องของผลการวัด รวมถึงเป็นหลักฐานที่จำเป็นในการอธิบายผลการวัดของกลุ่มตัวอย่างและความสามารถในการอ้างอิงคุณสมบัติทางจิตมิติของเครื่องมือเมื่อไปใช้กับกลุ่มตัวอย่างอื่นๆ

4) ค่า Outfit MS คือ ค่าสถิติที่แสดงผลการประเมินที่ผิดปกติไปจากที่ควรจะเป็น โดยค่าสถิติดังกล่าวจะมีค่าสูงหากพบผลการประเมินที่สูงหรือต่ำเกินจริง (Outliner) โดยค่า Outfit MS ควรมีค่าน้อยกว่า 2.0 การผ่านเกณฑ์ค่า Outfit MS มีความจำเป็นในการยืนยันความถูกต้องของผลการวัด รวมถึงเป็นหลักฐานที่จำเป็นของความสามารถในการอ้างอิงคุณสมบัติทางจิตมิติของเครื่องมือเมื่อไปใช้กับกลุ่มตัวอย่างอื่นๆ

โดยในบทความวิชาการนี้ จะนำเสนอตัวอย่างการวิเคราะห์เพื่อปรับประสิทธิภาพระดับคะแนนจากโปรแกรม Facets ประกอบด้วย คำสั่งในการวิเคราะห์ การแปลผลจาก print out ของโปรแกรม โดยกำหนดสถานการณ์คือ การพิจารณาปรับปรุงระดับคะแนนของเกณฑ์การประเมินแต่ละข้อของรูบริกสำหรับการประเมินทักษะ โดยมีพาเซตที่เกี่ยวข้องคือ ผู้เรียนจำนวน 50 คน รูบริกที่มีเกณฑ์การประเมิน 5 ข้อ และผู้ประเมิน จำนวน 3 คน

ตัวอย่างคำสั่งและการแปลผลการวิเคราะห์ (Example of Analysis' Command Code and Interpretation)

Figure 1 แสดงตัวอย่างคำสั่งในการวิเคราะห์โมเดลการให้คะแนนบางส่วนแบบหลายองค์ประกอบของรหัสด้วยโปรแกรม Facets โดยในคำสั่งแสดงฟาเซตที่จะวิเคราะห์จำนวน 3 ฟาเซต ประกอบด้วย ผู้เรียน 50 คน ผู้ประเมิน 3 คน และ เกณฑ์ประเมิน 5 ข้อ โดยในคำสั่งการวิเคราะห์โมเดล (Models =) จะมีการกำหนดให้วิเคราะห์ฟาเซตเกณฑ์การประเมินแยกเป็นรายข้อในรูปแบบของโมเดลการให้คะแนนบางส่วน โดยกำหนดคำสั่งเป็นสัญลักษณ์ # แทนสัญลักษณ์ ? ดังแสดงในบรรทัดที่ 8 ของคำสั่งการวิเคราะห์ใน Figure 1

Figure 1

Example of command code for MFRM-PCM in Facets Program

ตัวอย่างคำสั่งการวิเคราะห์โมเดลการให้คะแนนบางส่วนแบบหลายองค์ประกอบของรหัสด้วยโปรแกรม Facets

```
Title = FACETS essay
Facets = 3 ; student + rater + item
Inter-rater = 2
Positive = 1
Noncentered = 1
Pearsondf = Yes
Models=
?, ?, #, R5
*
Labels=
*
1, student
1-50
*
2, raters
1-3
*
3, items
1-5
```

Figure 2 แสดงผลการวิเคราะห์ Wright Map ซึ่งเป็นภาพรวมของผลการวัด ช่วยให้สามารถเปรียบเทียบผลการวัดข้ามฟาเซตได้ โดยคอลัมน์ที่ 1 แสดงค่า Measure ที่เปรียบเสมือนไม้บรรทัดของการวัด คอลัมน์ที่ 2 แสดงผลความสามารถของผู้เรียนจากเก่งไปอ่อนเรียงตามค่า Measure โดยจากภาพจะสังเกตได้ว่า ผู้เรียนส่วนใหญ่จะแสดงอยู่ในค่า Measure ตั้งแต่ 0 เป็นต้นไป แสดงให้เห็นว่ากลุ่มตัวอย่างที่เป็นผู้เรียนมีความสามารถค่อนข้างสูง คอลัมน์ที่ 3 แสดงการกดปล่อยคะแนนของผู้ประเมิน โดยหากมีค่ามาก แสดงว่าผู้ประเมินกดคะแนน หากมีค่าน้อยแสดงว่าผู้ประเมินปล่อยคะแนน จากภาพ สังเกตได้ว่า ผู้ประเมินคนที่ 1 มีการประเมินแบบกดคะแนนมากกว่าผู้ประเมินคนที่ 2 และ 3 ที่ค่อนข้างปล่อยคะแนนเท่าๆ กัน คอลัมน์ที่ 4 แสดงความยากของเกณฑ์การประเมินแต่ละข้อ โดยข้อที่มีความยากมากที่สุดคือเกณฑ์การประเมินข้อที่ 4 และ 5 ส่วนข้อที่มีความง่ายมากที่สุด คือ เกณฑ์การประเมินข้อที่ 1 คอลัมน์ที่ 5 ถึง 9 แสดงผลความยากของแต่ละระดับคะแนนแยกเป็นเกณฑ์การประเมินรายข้อ โดยมีระดับคะแนนตั้งแต่ 0 คะแนนไปถึง 4 คะแนน ตามลำดับ

Figure 2

Result of Wright Map analysis

ผลการวิเคราะห์ Wright Map

Measr	+student	-raters	-items	S.1	S.2	S.3	S.4	S.5
2	+	+	+	(4)	(4)	(4)	(4)	(4)
	*							
	*							
	**							
	*							
1	+	+	+	---	---	---	---	---
	*							
	*							
	**							
	**							

	***			3		3	3	3
	***				3			
	***			---	---	---	---	---
	***	1	4 5	---	---			
	***							2
*	0	*	*	3	2	2	2	*
	*****	2 3	2	---	---	---	---	---

	*		1					
	*			1	1	1	1	1
	*							
	*							
	*							
-1	+	+	+	(0)	(0)	(0)	(0)	(0)
Measr	* = 1	-raters	-items	S.1	S.2	S.3	S.4	S.5

เมื่อสำรวจข้อมูลภาพรวมของการประเมินแล้ว ในการวิเคราะห์เพื่อปรับประสิทธิภาพระดับคะแนนของรูบริก ให้ผู้วิเคราะห์พิจารณาผลการวิเคราะห์ค่า Separation ในฟาเซตนักเรียน ในตารางที่มีชื่อว่า Measurement Report โดยค่า Separation เป็นค่าแสดงสัดส่วนระหว่างส่วนเบี่ยงเบนมาตรฐานกับค่าเฉลี่ยความคลาดเคลื่อนของการประมาณ โดยค่า Separation ในฟาเซตนักเรียนจะเป็นค่าที่อธิบายว่า เกณฑ์การประเมินทั้งหมดในรูบริกสามารถจำแนกความสามารถผู้เรียน ได้กี่กลุ่ม ซึ่งค่า Separation ในฟาเซตผู้เรียนควรมีค่าน้อย 2.0 เพื่อยืนยันว่าเกณฑ์การประเมินสามารถจำแนกความสามารถผู้เรียนได้อย่างน้อยสองกลุ่ม (Bond et al., 2021; Boone et al., 2014) ซึ่งใน Figure 3 ผลการวิเคราะห์ค่า Separation ในฟาเซตนักเรียนมีค่าเท่ากับ 2.09 แสดงให้เห็นว่าเกณฑ์การประเมินทั้ง 5 ข้อสามารถจำแนกผู้เรียนได้อย่างน้อย 2 กลุ่ม

Figure 3

Result of students' facet analysis

ผลการวิเคราะห์ฟาเซตนักเรียน

Model, Populn: RMSE .23	Adj (True) S.D. .49	Separation 2.09	Strata 3.12	Reliability .81
Model, Sample: RMSE .23	Adj (True) S.D. .49	Separation 2.12	Strata 3.16	Reliability .82
Model, Fixed (all same) chi-squared: 197.9	d.f.: 49	significance (probability): .00		
Model, Random (normal) chi-squared: 36.5	d.f.: 48	significance (probability): .89		

ต่อมาให้พิจารณาฟาเซตผู้ประเมินและเกณฑ์การประเมิน ในตาราง Measurement Report โดยพิจารณาผลความสอดคล้องของการวัดกับโมเดล ด้วยค่าสถิติ Infit-Outfit Mean Square (MS) ซึ่งเป็นค่าที่ใช้ในการรายงานผลว่า ข้อมูลที่เก็บได้จากการประเมินสอดคล้องกับโมเดล (model fit) ในการวิเคราะห์หรือไม่ และเป็นหลักฐานยืนยันความตรงเชิงโครงสร้าง (construct validity) ของผลการวัด โดยค่า Infit-Outfit MS ของผู้ประเมินและเกณฑ์การประเมินที่สอดคล้องกับโมเดลควรมีค่าอยู่ในช่วง 0.5 ถึง 1.5 (Wright & Linacre, 1994)

จาก Figure 4 แสดงผลการวิเคราะห์ฟาเซตผู้ประเมินจำนวน 3 คน เมื่อพิจารณาค่า Infit-Outfit Mean Square (MnSq) พบว่า ผู้ประเมินจำนวน 3 คน มีค่า Infit MnSq อยู่ในช่วง 0.91 ถึง 1.06 และมีค่า Outfit MnSq อยู่ในช่วง 0.92 – 1.15 แสดงให้เห็นว่าผู้ประเมินทั้ง 3 คน ให้ผลการประเมินได้สอดคล้องกับโมเดล

Figure 4

Result of raters' facet analysis

ผลการวิเคราะห์ฟาเซตผู้ประเมิน

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	ZStd	Outfit MnSq	ZStd	Estim. Discrm	Correlation PtMea PtExp	Exact Obs %	Agree. Exp %	N raters
493	250	1.97	1.99	.25	.05	1.02	.3	.92	-.7	1.61	.70 .58	24.6	32.2	1 1
642	250	2.57	2.77	-.12	.05	.91	-1.1	.92	-.6	.97	.55 .54	29.4	34.3	2 2
648	250	2.59	2.80	-.13	.05	1.06	.7	1.15	1.2	.48	.35 .53	25.2	34.3	3 3
594.3	250.0	2.38	2.52	.00	.05	1.00	.0	1.00	-1.1		.53			Mean (Count: 3)
71.7	.0	.29	.37	.18	.00	.06	.8	.11	1.0		.14			S.D. (Population)
87.8	.0	.35	.46	.21	.00	.07	1.0	.13	1.2		.17			S.D. (Sample)

Model, Populn: RMSE .05 Adj (True) S.D. .17 Separation 3.36 Strata 4.81 Reliability (not inter-rater) .92
 Model, Sample: RMSE .05 Adj (True) S.D. .21 Separation 4.18 Strata 5.90 Reliability (not inter-rater) .95
 Model, Fixed (all same) chi-squared: 37.4 d.f.: 2 significance (probability): .00
 Model, Random (normal) chi-squared: 1.9 d.f.: 1 significance (probability): .17
 Inter-Rater agreement opportunities: 750 Exact agreements: 198 = 26.4% Expected: 251.9 = 33.6%

จาก Figure 5 แสดงผลการวิเคราะห์ฟาเซตเกณฑ์การประเมินจำนวน 5 ข้อ เมื่อพิจารณาค่า Infit-Outfit MnSq พบว่า ข้อรายการประเมินทั้ง 5 ข้อ มีค่า Infit MnSq อยู่ในช่วง 0.92 ถึง 1.10 และมีค่า Outfit MnSq อยู่ในช่วง 0.90 – 1.15 แสดงให้เห็นว่าเกณฑ์การประเมินทั้ง 5 มีผลการใช้ประเมินที่สอดคล้องกับโมเดล เมื่อพิจารณาความเที่ยงของเกณฑ์การประเมิน โดยพิจารณาจากค่า Separation และ ค่าความเที่ยงแบบราซส์ (Rasch Reliability) จากการประมาณประชากร (Population: Populn) พบว่า ค่า Separation มีค่าเท่ากับ 2.58 ซึ่งหมายถึง เกณฑ์การประเมิน 5 ข้อแบ่งออกได้เป็นข้อที่ง่ายและยากประมาณ 2 กลุ่ม และมีค่าความเที่ยงมีค่าเท่ากับ .87 โดยค่าความเที่ยงแบบโมเดลราซส์จะมีค่าต่ำกว่าค่าความเที่ยงแบบสัมประสิทธิ์แอลฟาของครอนบาคเล็กน้อย แต่สามารถใช้การแปลแบบเดียวกันได้ (Linacre, 2023) ดังนั้น ค่าความเที่ยงของเกณฑ์การประเมินทั้ง 5 ข้อ จึงมีความเที่ยงผ่านเกณฑ์ที่กำหนดไว้โดยมีค่า Separation มากกว่า 2.5 และค่าความเที่ยงมากกว่า .70 (Boone et al., 2014; George & Mallery, 2019; Hughes et al., 2022)

ต่อมา ให้พิจารณาค่า Point-Measure Correlation (PtMea Correlation) ซึ่งเป็นค่าที่แสดงสถิติสหสัมพันธ์ของเกณฑ์การประเมินแต่ละข้อ และเป็นค่าสถิติที่สำคัญอันดับแรกสำหรับการวิเคราะห์เพื่อปรับประสิทธิภาพระดับคะแนน โดยค่า PtMea Correlation ของเกณฑ์การประเมินทุกข้อควรมีค่าเป็นบวก เพื่อเป็นหลักฐานแสดงให้เห็นว่าเกณฑ์การประเมินทุกข้อมุ่งวัดทักษะเดียวกันในลักษณะของความเป็นเอกมิติ (Unidimensionality) (Linacre, 2002) โดยในตัวอย่างผลการวิเคราะห์ Figure 5 แสดงให้เห็นว่าเกณฑ์การประเมินทุกข้อมีค่า PtMea Correlation เป็นบวกทุกข้อ โดยมีค่าอยู่ในช่วง .52 ถึง .59

Figure 5

Result of criteria' facet analysis

ผลการวิเคราะห์ฟาส์เกตเกณฑ์การประเมิน

Total Score	Total Count	Obsvd Average	Fair(M) Average	- Measure	Model S.E.	Infit MnSq	ZStd	Outfit MnSq	ZStd	Estim. Discrm	Correlation PtMea	PtExp	N items
423	150	2.82	3.08	-.28	.07	.92	-.7	.90	-.5	1.06	.55	.51	1 1
385	150	2.57	2.82	-.07	.06	1.02	.1	1.07	.4	.98	.53	.54	2 2
365	150	2.43	2.60	-.04	.06	.99	.0	.95	-.3	1.01	.55	.55	3 3
316	150	2.11	2.18	.16	.06	.97	-.3	.93	-.5	1.04	.59	.57	4 4
294	150	1.96	1.97	.23	.07	1.10	1.0	1.15	1.1	.87	.52	.58	5 5
356.6	150.0	2.38	2.53	.00	.06	1.00	.0	1.00	.0		.55		Mean (Count: 5)
46.6	.0	.31	.41	.18	.00	.06	.6	.09	.7		.02		S.D. (Population)
52.1	.0	.35	.46	.20	.00	.07	.7	.10	.8		.03		S.D. (Sample)

Model, Populn: RMSE .06 Adj (True) S.D. .17 Separation 2.58 Strata 3.77 Reliability .87
 Model, Sample: RMSE .06 Adj (True) S.D. .19 Separation 2.93 Strata 4.24 Reliability .90
 Model, Fixed (all same) chi-squared: 37.2 d.f.: 4 significance (probability): .00
 Model, Random (normal) chi-squared: 3.6 d.f.: 3 significance (probability): .31

เมื่อวิเคราะห์ฟาส์เกตผู้ประเมินและเกณฑ์การประเมินแล้ว ต่อมาให้พิจารณาผลการวิเคราะห์สถิติระดับคะแนน (Category Statistics) รายข้อ ดัง Figure 6 โดยให้พิจารณาค่าสถิติ 3 ค่าสำหรับการวิเคราะห์เพื่อปรับประสิทธิภาพระดับคะแนน ค่าที่ 1 คือ ค่า Category Usage ที่แสดงในคอลัมน์ที่ 2 (Category Total) ค่าที่ 2 คือ ค่า Average Measure แสดงในคอลัมน์ที่ 6 (Avge Meas) และค่าที่ 3 คือ Outfit MS แสดงในคอลัมน์ที่ 8 (OUTFIT MnSq) โดยในตารางผลการวิเคราะห์แต่ละข้อแสดงผลเป็นระดับคะแนนไล่เรียงตั้งแต่ระดับ 0 คะแนนถึง 4 คะแนนแสดงตามแถวของแต่ละตาราง

จากผลการวิเคราะห์ใน Figure 6 แสดงให้เห็นว่า เกณฑ์การประเมินข้อที่ 1 และ 3 มีผลการวิเคราะห์เพื่อปรับประสิทธิภาพระดับคะแนนผ่านเกณฑ์ทุกข้อ พิจารณาได้จาก 1) ค่า Category Usage ที่มากกว่า 10 ในทุกระดับคะแนน หมายถึง ระดับคะแนนตั้งแต่ 0 ถึง 4 คะแนน ถูกใช้ประเมินอย่างน้อย 10 ครั้งทุกระดับ 2) ค่า Average Measure มีค่าเรียงลำดับจากน้อยไปมากทุกข้อ หมายถึง ระดับคะแนนจากน้อยไปมากสอดคล้องกับระดับความยากของทักษะจากง่ายไปยากตามลำดับ และ 3) ค่า Outfit MS น้อยกว่า 2.0 ในทุกระดับคะแนน หมายถึง ไม่มีผลการประเมินที่ผิดปกติเกิดขึ้นในทุกระดับคะแนน

สำหรับการวิเคราะห์เพื่อปรับประสิทธิภาพระดับคะแนนเกณฑ์การประเมินข้อที่มีปัญหา ในเกณฑ์การประเมินข้อที่ 2 พบว่า ระดับคะแนนที่ 2 คะแนน มีค่า Average Measure เท่ากันกับระดับคะแนนที่ 1 คะแนน (Avge Meas = 0.11) หมายถึง ระดับคะแนนที่ 2 คะแนน มีระดับความยากของทักษะเท่ากับระดับคะแนนที่ 1 หรือแปลผลได้ว่า ระดับคะแนนที่ 1 และ 2 มีความยากของทักษะไม่ต่างกัน จึงควรพิจารณาบูรณาการระดับคะแนนทั้งสองให้เป็นระดับคะแนนเดียวกัน โดยอาจพิจารณาจากระดับคะแนนที่ 2 หรือ 3 ไว้ตามความเหมาะสมของระดับความยากของทักษะ ส่งผลให้เกณฑ์การประเมินข้อที่ 2 ควรมีระดับคะแนนทั้งหมด 4 ระดับ (0 – 3 คะแนน)

การวิเคราะห์เพื่อปรับประสิทธิภาพระดับคะแนนเกณฑ์การประเมินข้อที่ 4 พบว่า ค่า Average Measure ของระดับคะแนนที่ 4 คะแนน (Avge Meas = 0.44) มีค่าน้อยกว่าค่า Average Measure ของระดับคะแนนที่ 3 คะแนน (Avge Meas = 0.45) หมายถึง ทักษะในระดับคะแนนที่ 3 คะแนน มีความยากมากกว่าในระดับคะแนนที่ 4 คะแนน แสดงให้เห็นถึงความไม่เรียงลำดับของความยากในเกณฑ์การประเมิน จึงควรพิจารณาบูรณาการระดับคะแนนทั้งสอง โดยอาจพิจารณาให้คงระดับคะแนนที่ 3 ไว้ เป็นระดับคะแนนที่สูงสุด ส่งผลให้เกณฑ์การประเมินข้อที่ 4 ควรมีระดับคะแนนทั้งหมด 4 ระดับ (0 – 3 คะแนน)

ส่วนการวิเคราะห์เพื่อปรับประสิทธิภาพระดับคะแนนเกณฑ์การประเมินข้อที่ 5 พบว่า ค่า Average Measure ของระดับคะแนนที่ 4 คะแนน (Avge Meas = 0.27) มีค่าน้อยกว่าค่า Average Measure ของระดับคะแนนที่ 3 คะแนน

(Avge Meas = 0.57) หมายถึง ทักษะในระดับคะแนนที่ 3 คะแนน มีความยากมากกว่าในระดับคะแนนที่ 4 คะแนน แสดงให้เห็นถึงความไม่เรียงลำดับของความยากในเกณฑ์การประเมิน นอกจากนี้ ยังพบว่า ค่า Outfit MS ในระดับคะแนนที่ 2 คะแนนมีค่าเท่ากับ 2.0 ซึ่งไม่ผ่านเกณฑ์ ที่กำหนดไว้ว่า ค่า Outfit MS ต้องมีค่าน้อยกว่า 2.0 จึงควรมีการพิจารณาขบรวมนระดับคะแนนที่ 2 กับระดับคะแนนที่ 3 ให้เป็นระดับคะแนนเดียวกัน ดังนั้น เกณฑ์การประเมินข้อที่ 5 จึงควรมีระดับคะแนนจำนวน 3 ระดับ (0 – 2 คะแนน) ตัวอย่างการปรับรูปกรก่อนและหลังการวิเคราะห์สถิติระดับคะแนนรายชื่อแสดงใน Figure 7

Figure 6

Result of category statistics analysis

ผลการวิเคราะห์สถิติระดับคะแนนรายชื่อ

เกณฑ์การประเมินข้อที่ 1	DATA				QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST PROBABLE		RASCH-THURSTONE		Cat
	Score	Category Total	Counts Used	Cum. %	Avge Meas	Exp. Meas	OUTFIT MnSq	Thresholds Measure	S.E.	Measure at Category -0.5		from		Thresholds	PEAK Prob	
เกณฑ์การประเมินข้อที่ 1	0	18	18	12%	12%	-.04	.04	.7		(-1.41)		low		low		100%
	1	18	18	12%	24%	.24	.18	1.4	.11	.27	-.46	-.92		-.50		27%
	2	14	14	9%	33%	.35	.34	.9	.51	.22	.03	-.19		-.04		13%
	3	23	23	15%	49%	.40	.53	.6	-.07	.20	.50	.25		.15		19%
	4	77	77	51%	100%	.84	.80	.8	-.55	.18	(1.33)	.90	.00	.44		100%
(Mean) (Modal) (Median)																
เกณฑ์การประเมินข้อที่ 2	Score	Category Total	Counts Used	Cum. %	Avge Meas	Exp. Meas	OUTFIT MnSq	Thresholds Measure	S.E.	Measure at Category -0.5		from		Thresholds	PEAK Prob	Cat
	0	30	30	20%	20%	-.12	-.12	1.3		(-1.21)		low		low		
	1	12	12	8%	28%	.11	.02	1.3	.86	.23	-.45	-.83		-.36		
	2	19	19	13%	41%	.11*	.18	.4	-.36	.21	-.01	-.21		-.14		
	3	21	21	14%	55%	.40	.38	1.2	.17	.19	.44	.19		.10		
เกณฑ์การประเมินข้อที่ 2	4	68	68	45%	100%	.63	.64	.9	-.67	.19	(1.25)	.83	.00	.37		100%
(Mean) (Modal) (Median)																
เกณฑ์การประเมินข้อที่ 3	Score	Category Total	Counts Used	Cum. %	Avge Meas	Exp. Meas	OUTFIT MnSq	Thresholds Measure	S.E.	Measure at Category -0.5		from		Thresholds	PEAK Prob	Cat
	0	26	26	17%	17%	-.20	-.14	.8		(-1.47)		low		low		
	1	19	19	13%	30%	.03	.01	1.2	.25	.24	-.51	-.99		-.55		
	2	28	28	19%	49%	.25	.18	1.0	-.30	.20	.03	-.21	-.02	-.15		
	3	18	18	12%	61%	.44	.39	.7	.72	.19	.54	.27		.24		
เกณฑ์การประเมินข้อที่ 3	4	59	59	39%	100%	.63	.65	1.0	-.67	.19	(1.42)	.97	.02	.50		100%
(Mean) (Modal) (Median)																
เกณฑ์การประเมินข้อที่ 4	Score	Category Total	Counts Used	Cum. %	Avge Meas	Exp. Meas	OUTFIT MnSq	Thresholds Measure	S.E.	Measure at Category -0.5		from		Thresholds	PEAK Prob	Cat
	0	36	36	24%	24%	-.42	-.28	.6		(-1.46)		low		low		
	1	24	24	16%	40%	-.04	-.13	.8	.20	.21	-.49	-.97		-.54		
	2	26	26	17%	57%	.21	.05	1.4	-.13	.19	.04	-.20		-.11		
	3	16	16	11%	68%	.45	.26	.8	.63	.20	.54	.27		.23		
เกณฑ์การประเมินข้อที่ 4	4	48	48	32%	100%	.44*	.53	1.1	-.71	.20	(1.38)	.95	.00	.48		100%
(Mean) (Modal) (Median)																
เกณฑ์การประเมินข้อที่ 5	Score	Category Total	Counts Used	Cum. %	Avge Meas	Exp. Meas	OUTFIT MnSq	Thresholds Measure	S.E.	Measure at Category -0.5		from		Thresholds	PEAK Prob	Cat
	0	36	36	24%	24%	-.47	-.34	.8		(-1.62)		low		low		
	1	31	31	21%	45%	-.09	-.18	1.2	-.11	.21	-.51	-1.05	-.11	-.66		
	2	27	27	18%	63%	.24	.01	2.0	.05	.19	.06	-.19		-.09		
	3	15	15	10%	73%	.57	.23	.4	.70	.20	.59	.31		.28		
เกณฑ์การประเมินข้อที่ 5	4	41	41	27%	100%	.27*	.51	1.4	-.64	.21	(1.45)	1.01	.04	.54		100%
(Mean) (Modal) (Median)																

Figure 7

The example of modifying rubrics before (upper) and after (down) the category statistics analysis

ตัวอย่างการปรับปรุงริกก่อน (บน) และหลัง (ล่าง) การวิเคราะห์สถิติระดับคะแนนรายข้อ

ข้อที่ 1	<u>คำอธิบายทักษะ</u> (4 คะแนน)	<u>คำอธิบายทักษะ</u> (3 คะแนน)	<u>คำอธิบายทักษะ</u> (2 คะแนน)	<u>คำอธิบายทักษะ</u> (1 คะแนน)	<u>คำอธิบายทักษะ</u> (0 คะแนน)
ข้อที่ 2	<u>คำอธิบายทักษะ</u> (4 คะแนน)	<u>คำอธิบายทักษะ</u> (3 คะแนน)	<u>คำอธิบายทักษะ</u> (2 คะแนน)	<u>คำอธิบายทักษะ</u> (1 คะแนน)	<u>คำอธิบายทักษะ</u> (0 คะแนน)
ข้อที่ 3	<u>คำอธิบายทักษะ</u> (4 คะแนน)	<u>คำอธิบายทักษะ</u> (3 คะแนน)	<u>คำอธิบายทักษะ</u> (2 คะแนน)	<u>คำอธิบายทักษะ</u> (1 คะแนน)	<u>คำอธิบายทักษะ</u> (0 คะแนน)
ข้อที่ 4	<u>คำอธิบายทักษะ</u> (4 คะแนน)	<u>คำอธิบายทักษะ</u> (3 คะแนน)	<u>คำอธิบายทักษะ</u> (2 คะแนน)	<u>คำอธิบายทักษะ</u> (1 คะแนน)	<u>คำอธิบายทักษะ</u> (0 คะแนน)
ข้อที่ 5	<u>คำอธิบายทักษะ</u> (4 คะแนน)	<u>คำอธิบายทักษะ</u> (3 คะแนน)	<u>คำอธิบายทักษะ</u> (2 คะแนน)	<u>คำอธิบายทักษะ</u> (1 คะแนน)	<u>คำอธิบายทักษะ</u> (0 คะแนน)



ข้อที่ 1	<u>คำอธิบายทักษะ</u> (4 คะแนน)	<u>คำอธิบายทักษะ</u> (3 คะแนน)	<u>คำอธิบายทักษะ</u> (2 คะแนน)	<u>คำอธิบายทักษะ</u> (1 คะแนน)	<u>คำอธิบายทักษะ</u> (0 คะแนน)
ข้อที่ 2	<u>คำอธิบายทักษะ</u> (3 คะแนน)	<u>คำอธิบายทักษะ</u> (2 คะแนน)	<u>คำอธิบายทักษะ</u> (1 คะแนน)		<u>คำอธิบายทักษะ</u> (0 คะแนน)
ข้อที่ 3	<u>คำอธิบายทักษะ</u> (4 คะแนน)	<u>คำอธิบายทักษะ</u> (3 คะแนน)	<u>คำอธิบายทักษะ</u> (2 คะแนน)	<u>คำอธิบายทักษะ</u> (1 คะแนน)	<u>คำอธิบายทักษะ</u> (0 คะแนน)
ข้อที่ 4	<u>คำอธิบายทักษะ</u> (3 คะแนน)		<u>คำอธิบายทักษะ</u> (2 คะแนน)	<u>คำอธิบายทักษะ</u> (1 คะแนน)	<u>คำอธิบายทักษะ</u> (0 คะแนน)
ข้อที่ 5	<u>คำอธิบายทักษะ</u> (2 คะแนน)			<u>คำอธิบายทักษะ</u> (1 คะแนน)	<u>คำอธิบายทักษะ</u> (0 คะแนน)

ข้อสังเกตและข้อเสนอแนะ (Remarks and Recommendations)

ตัวอย่างคำสั่งและการแปลผลการวิเคราะห์ข้างต้น เป็นการสรุปขั้นตอนที่สำคัญสำหรับการวิเคราะห์การปรับปรุงประสิทธิภาพระดับคะแนนจากผลการวิเคราะห์ด้วยโมเดลการให้คะแนนบางส่วนแบบหลายองค์ประกอบของราสซ์ ซึ่งมีข้อสังเกตบางประการคือ ในบางขั้นตอนยังมีรายละเอียดการวิเคราะห์ปลีกย่อยที่ช่วย (helpful) ในการวิเคราะห์ได้ เช่น การพิจารณาการแจกแจงความถี่ (distribution) ของผลการประเมินในแต่ละระดับคะแนน หรือตำแหน่งและระยะห่างระหว่างจุดตัด (thresholds) ที่เหมาะสม (Linacre, 2002) โดยการพิจารณาค่าเหล่านี้ช่วยในการอธิบายหรือตัดสินใจมากขึ้น รวมถึงการพิจารณาความสอดคล้องของข้อมูลกับโมเดล จากค่าสถิติ Chi-square โดยโมเดลราสซ์คาดหวังให้ข้อมูลกับโมเดลไม่สอดคล้องกัน เนื่องจากโมเดลราสซ์เป็นโมเดลสมบูรณ์ทางคณิตศาสตร์ (mathematical ideal) ดังนั้น จึงต้องการให้ข้อมูลที่สังเกตได้มีความแปรปรวนในระดับหนึ่ง และราสซ์โมเดลให้ความสำคัญกับความสอดคล้องของข้อมูลด้วยค่าสถิติ MS ที่นำไปใช้ได้จริง (Practical) มากกว่า Goodness-of-fit ที่เป็นการทดสอบทางสถิติเพียงอย่างเดียว (Statistical) (Linacre, 2003; Linacre, 2024; Raykov & DiStefano, 2021) และมีการศึกษาว่าการพิจารณาความสอดคล้องของข้อมูลกับโมเดลด้วยค่า Chi-square อาจไม่มีประสิทธิภาพ

เนื่องจากโมเดลราสซี้ใช้การประมาณผลการวิเคราะห์ด้วยฟังก์ชัน Maximum Likelihood (ML) (Maydeu-Olivares & Montaña, 2013)

ข้อสังเกตประการต่อมาคือ ค่า Separation ในฟาเซตเกณฑ์การประเมิน แม้ว่า Linacre (2024) จะกล่าวว่า ค่า Separation ในฟาเซตเกณฑ์การประเมินที่ต่ำกว่า 3 และมีความเที่ยงแบบราสซี้ต่ำกว่า .90 จะพิจารณาว่ามีค่าต่ำ ซึ่งเกิดจากมีตัวอย่างจำนวนน้อยเกินไปที่จะยืนยันว่าฟาเซตเกณฑ์การประเมินมีความยาก อย่างน้อย 3 ระดับ แต่เกณฑ์ดังกล่าวถูกอ้างถึงในกรณีที่เป็นการข้อสอบแบบถูกผิดแบบ 2 ค่า (dichotomous item) จึงอาจพิจารณาเป็นนัยได้ว่า สาเหตุที่ค่า Separation ควรมากกว่า 3 เป็นเพราะในแบบสอบควรต้องมีข้อสอบที่มีความยากและอำนาจในการจำแนกที่แตกต่างกัน ในระดับง่ายปานกลาง และยาก อย่างน้อย 3 กลุ่ม ซึ่งแตกต่างจากการประเมินทักษะที่ใช้รูบริกที่มีเกณฑ์การประเมินหลายข้อ ซึ่งไม่ได้ต้องการที่จะให้แต่ละข้อมีความยากหรือง่ายแตกต่างกัน แต่ต้องการความยากง่ายเกิดขึ้นแตกต่างกันในแต่ละระดับคะแนน

สำหรับข้อเสนอแนะในการวิเคราะห์ตามแนวทางนี้ คือ ไม่ควรปรับปรุงระดับคะแนนของรูบริกโดยพิจารณาผลการวิเคราะห์ด้วยค่าสถิติเพียงอย่างเดียว แต่ควรใช้การพิจารณาจากเนื้อหา (content) ของคำบรรยายคุณภาพงานหรือการปฏิบัติงาน ความเหมาะสม และความเป็นไปได้ในการปรับปรุงด้วย ตัวอย่างจากงานของ Janssen et al. (2015) ที่พิจารณาจำนวนระดับคะแนนที่เหมาะสมในการประเมินทักษะของผู้เรียน โดยผลการพิจารณาพบว่า แม้ว่า 4 ระดับคะแนนจะให้ผลการวิเคราะห์ที่มั่นคงกว่า แต่ทำให้การจำแนกทักษะของผู้เรียนไม่มีประสิทธิภาพเท่ากับ 6 ระดับคะแนน เป็นต้น โดยผู้พิจารณาประเด็นดังกล่าวอาจเป็นผู้พัฒนารูบริกเองหรือให้ผู้เชี่ยวชาญเนื้อหาวิชาเข้ามาพิจารณาร่วมด้วยก็ได้ เพื่อหาจุดสมดุลระหว่างเนื้อหาและค่าสถิติ (van der Linden, 2017; Wilson, 2013)

นอกจากนี้ จากการศึกษาของ Adams et al. (2012) ที่พบว่า โมเดลราสซี้สำหรับการวิเคราะห์เครื่องมือที่ให้คะแนนได้หลายค่า มีความเป็นไปได้ทางสมการที่จะเกิดการไม่เรียงลำดับของระดับคะแนนภายในเกณฑ์การประเมินรายข้อที่เกิดจากประมาณค่าพารามิเตอร์เทรชโฮลด์ (Threshold) ซึ่งแตกต่างจากโมเดล Graded Response Model ที่ในทางสมการจะบังคับให้ค่าพารามิเตอร์เทรชโฮลด์เรียงลำดับกัน อย่างไรก็ตาม การประมาณค่าพารามิเตอร์ไม่ได้เกี่ยวข้องโดยตรงกับความสามารถของผู้เรียน และมักมีปัญหาในกรณีที่ผู้ประเมินเข้ามาเกี่ยวข้อง ซึ่งการไม่เรียงลำดับกันของพารามิเตอร์เทรชโฮลด์สำหรับโมเดลราสซี้ อาจเกิดจากการที่ระดับคะแนนนั้นๆ มีการถูกใช้ประเมินน้อย และสามารถแปลความหมายได้วาระดับคะแนนนั้นๆ อาจไม่มีประสิทธิภาพในการวัดและอาจมีปัญหาเกี่ยวกับความสามารถในการจำแนก (discriminate) ทักษะ ซึ่งหากมีปัญหานี้เกิดขึ้น ควรมีการตรวจสอบและตัดสินใจปรับปรุงโดยผู้พัฒนาเครื่องมือ แต่หากสามารถยืนยันได้ว่า ระดับคะแนนเหล่านั้นเรียงลำดับกันอย่างถูกต้องตามทฤษฎีทางจิตวิทยาที่ใช้ในการพัฒนาเครื่องมือแล้ว ก็ถือว่าการไม่เรียงลำดับกันของพารามิเตอร์ไม่ได้เป็นปัญหาแต่อย่างใด นอกจากนี้ แม้จะมีการอ้างว่า ผู้ประเมินอาจทำให้เกิดการไม่เรียงลำดับกันของการประมาณพารามิเตอร์เนื่องจากผลการประเมินไม่ได้เกิดจากผู้เรียนโดยตรง แต่เป็นผลมาจากการสังเกตและตัดสินใจของผู้ประเมิน อย่างไรก็ตาม ผู้เขียนพิจารณาว่า ในกรณีทั่วไปของการประเมินประเมินภาคปฏิบัติส่วนใหญ่ ยากที่จะหลีกเลี่ยงการใช้ผู้ประเมิน ไม่ว่าผู้ประเมินคนนั้นจะเป็นผู้ประเมิน ผู้เชี่ยวชาญ ครูผู้สอน หรือผู้สังเกตภายนอก ซึ่งในงานของ Adams et al. (2012) ไม่ได้กล่าวถึงโมเดลหลายองค์ประกอบ แต่จากเหตุผลที่กล่าวไปข้างต้น เมื่อรู้ว่าส่วนหนึ่งของผลการประเมินได้รับอิทธิพลจากการประเมินของผู้ประเมิน การวิเคราะห์คุณสมบัติทางจิตมิติของเครื่องมือการประเมินภาคปฏิบัติด้วยโมเดลหลายองค์ประกอบที่ประมาณค่าพารามิเตอร์ผ่านทุกฟาเซตที่เกี่ยวข้องกับการประเมินภาคปฏิบัติ โดยเฉพาะฟาเซตผู้ประเมิน ก็เป็นเหตุเป็นผลมากกว่าที่จะใช้การวิเคราะห์ด้วยโมเดลการตอบสนองข้อสอบแบบหลายค่า (Polytomous item response theory) เช่น โมเดลการให้คะแนนบางส่วน โมเดลมาตรฐานค่า หรือ Graded Response Model และใช้การวิเคราะห์ Interrater Reliability หรือ Rater Agreement แยกออกไปจากผลการประเมิน (Bond et al., 2021) ดังที่นิยมกันทั่วไปในการพัฒนารูบริก (Brookhart & Chen, 2015) ดังนั้น

ผู้เขียนจึงเสนอแนะว่า หากจะวิเคราะห์คุณสมบัติทางจิตมิติเพื่อพัฒนาเครื่องมือใดก็ตามที่มีผู้ประเมินเข้ามาเกี่ยวข้อง การใช้การวิเคราะห์ด้วยโมเดลหลายองค์ประกอบมีความเป็นเหตุเป็นผลมากกว่า

สำหรับข้อเสนอแนะในการนำไปใช้ เนื่องจากโมเดลการให้คะแนนบางส่วนแบบหลายองค์ประกอบของราชภัฏต้องการตัวอย่างวิจัยจำนวนไม่มากในการวิเคราะห์เมื่อเปรียบเทียบกับโมเดลการวัดในทฤษฎีการตอบสนองข้อสอบแบบหนึ่ง สอง และ สามพารามิเตอร์ โดยโมเดลการให้คะแนนบางส่วนแบบหลายองค์ประกอบของราชภัฏกำหนดตัวอย่างวิจัยขั้นต่ำไว้ที่ 50 คน และเกณฑ์การประเมินแต่ละข้อต้องใช้ในการประเมินอย่างน้อย 100 ครั้ง (Linacre, 1994b) จากข้อกำหนดดังกล่าว สามารถใช้ตัวอย่างวิจัยขั้นต่ำที่ 50 คน และมีผู้ประเมินอย่างน้อย 2 คน ประเมินตัวอย่างวิจัยทุกคนก็จะผ่านจำนวนขั้นต่ำในวิเคราะห์ได้ ดังนั้น จึงมีความเป็นไปได้ในการนำไปใช้ในบริบทของโรงเรียน สำหรับการพัฒนาารูปกริเพื่อประเมินทักษะผู้เรียนเป็นการเตรียมพร้อมสำหรับการจัดการเรียนรู้ฐานสมรรถนะ (competency-based learning) ที่ใช้การประเมินผ่านการให้ภาระงานที่ให้ผู้เรียนปฏิบัติเพื่อแสดงผลการเรียนรู้ที่สะท้อนความรู้ เจตคติ ทักษะ หรือคุณลักษณะต่างๆ ของตนเอง รวมถึงใช้การประเมินตามสภาพจริง เช่น การประเมินภาคปฏิบัติ การใช้แฟ้มสะสมผลงาน หรือการประเมินตนเองหรือเพื่อน โดยมีข้อมูลย้อนกลับ (feedback) ให้กับผู้เรียนในการพัฒนาการเรียนรู้ของตนเอง (Gervis, 2016; Morke et al., 2012; Office of the Education Council, 2020)

นอกจากนี้ รูปกริยังมีความสำคัญในการศึกษาระดับอุดมศึกษา กล่าวคือ จากมาตรฐานการประกันคุณภาพการศึกษาของเครือข่ายมหาวิทยาลัยอาเซียน (Asean University Network Quality Assurance: AUN-QA) ที่มหาวิทยาลัยไทยใช้เป็นมาตรฐาน เมื่อพิจารณามาตรฐานดังกล่าวเกี่ยวกับการประเมินการจัดการเรียนรู้ พบว่า การประเมินต้องประเมินการเรียนรู้ของผู้เรียนกับเป้าหมายการเรียนรู้ที่คาดหวัง มีการประเมินที่ชัดเจน มีการสื่อสารกับผู้เรียน และประเมินผู้เรียนอย่างต่อเนื่อง ใช้รูปกริที่มีความตรง ความเที่ยง และความยุติธรรม (fairness) ในการประเมิน และมีการให้ข้อมูลย้อนกลับจากการประเมินอย่างเหมาะสม (ASEAN University Network, 2020) ดังนั้น รูปกริที่เป็นเครื่องมือการวัดและประเมินผลที่สามารถประเมินทักษะหรือผลงานที่มีความซับซ้อน สะท้อนผลลัพธ์การเรียนรู้ปัจจุบันกับเป้าหมาย และสามารถให้ข้อมูลย้อนกลับได้ (Harris & Brown, 2022; Jönsson, 2022; Stevens & Levi, 2012) จึงมีความเหมาะสมในการประเมินการเรียนรู้ตามลักษณะดังกล่าว

บทสรุป (Conclusion)

การพัฒนาารูปกริด้วยการวิเคราะห์เพื่อปรับประสิทธิภาพระดับคะแนนการวิเคราะห์โดยใช้โมเดลการให้คะแนนบางส่วนแบบหลายองค์ประกอบของราชภัฏ ช่วยให้ครูหรือนักการศึกษาที่ต้องการพัฒนาารูปกริสำหรับประเมินทักษะหรือผลงานของนักเรียนสามารถพิจารณาจำนวนของระดับคะแนนของเกณฑ์การประเมินแต่ละข้อที่เหมาะสมได้ การพัฒนาเครื่องมือและการวิเคราะห์คุณสมบัติทางจิตมิติเครื่องมือควรเป็นขั้นตอนที่ต่อเนื่องกัน ดังนั้นผู้พัฒนาารูปกริจึงควรมีความรู้ความเข้าใจเกี่ยวกับทฤษฎีและขั้นตอนการวัดประเมินผลด้วย จึงเป็นที่มาของการเขียนบทความวิชาการนี้ขึ้นเพื่อให้ความรู้และความเข้าใจกับครูหรือนักการศึกษาโดยทั่วไป การพัฒนาารูปกริด้วยแนวทางนี้ช่วยยืนยันคุณสมบัติทางจิตมิติของรูปกริ กล่าวคือ รูปกริสามารถประเมินทักษะหรือผลงานที่เป้าหมายได้อย่างมีความตรงและความเที่ยง รวมถึงสามารถคงคุณสมบัติทางจิตมิติดังกล่าวเมื่อนำไปใช้กับนักเรียนกลุ่มอื่นในอนาคต การพัฒนาารูปกริตามแนวทางนี้ ใช้จำนวนตัวอย่างไม่มาก จึงมีความเป็นไปได้ในการประยุกต์ใช้ในบริบทของสถานศึกษา การพัฒนาารูปกริเพื่อใช้ประเมินผลลัพธ์การเรียนรู้ของผู้เรียนสอดคล้องกับการจัดการเรียนรู้ฐานสมรรถนะที่ให้ผู้เรียนแสดงผลการเรียนรู้ผ่านการปฏิบัติหรือผลงานที่สะท้อนความรู้ เจตคติ และทักษะของผู้เรียน และสอดคล้องกับมาตรฐานการประเมินการเรียนรู้ในระดับอุดมศึกษาตามแนวทางของ AUN-QA ซึ่งกล่าวถึงการใช้อุปกริที่มีความตรง ความเที่ยง และความยุติธรรมสำหรับประเมินการเรียนรู้และให้ข้อมูลย้อนกลับแก่ผู้เรียน

References

- Adams, R. J., Wu, M. L., & Wilson, M. (2012). The Rasch rating model and the disordered threshold controversy. *Educational and Psychological Measurement*, 72(4), 547-573.
DOI: 10.1177/0013164411432166
- ASEAN University Network. (2020). *ASEAN University network quality assurance: Guide to AUN-QA assessment at programme level version 4.0*. ASEAN University Network.
- Bond, T. G., Yan, Z., & Heene, M. (2021). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Routledge.
- Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer.
- Brookhart, S. M., & Chen, F. (2015). The quality and effectiveness of descriptive rubrics. *Educational Review*, 67(3), 43–368. DOI:10.1080/00131911.2014.929565
- Brookhart, S. M. (2018). Appropriate criteria: Key to effective rubrics. *Frontiers in Education*, 3, Article 22.
DOI: 10.3389/educ.2018.00022
- Chapman, M., Chuang, P., Bitterman, T., & Elliott, H. (2024). *Development of a new WIDA writing scoring rubric for grades 1-12* (Report No. TR-2024-2). Wisconsin Center for Education Research.
- Clarkson, B., & Luca, J. (2001). *Improving assessment: Rubrics in a tertiary multimedia course*. Association for the Advancement of Computing in Education.
- Edwards, K. E., Edwards, A. S., & Wesolowski, B. C. (2019). The psychometric evaluation of a wind band performance rubric using the multifaceted Rasch partial credit measurement model. *Research Studies in Music Education*, 41(3), 343-367. DOI: 10.1177/1321103X18773103
- Eckes, T. (2009). *Introduction to many-facet Rasch measurement*. Peter Lang Edition.
- Eckes, T. (2019). Many-facet Rasch measurement: Implications for rater mediated language assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment volume I: Fundamental techniques* (pp. 153-176). Routledge.
- Furtak, E. M. (2022). *Formative assessment*. Routledge. DOI: 10.4324/9781138609877-REE62-1
- Gervis, J. (2016). The operational definition of competency-based education. *The Journal of Competency-Based Education*, 1(2), 98-106. DOI: 10.1002/cbe2.1011
- George, D., & Mallery, P. (2019). *IBM SPSS Statistics 26 Step by Step* (16th ed.). Routledge.
- Griffith, W. I., & Lim, H. (2012). Performance-based assessment: Rubrics, web 2.0 tools and language competencies. *MEXTESOL Journal*, 36(1), 1-12.
- Harris, L. R., & Brown, G. T. L. (2022). *Self-assessment*. Routledge. DOI: 10.4324/9781138609877-REE1-1
- Hodges, T. S., Wright, K. L., Wind, S. A., Matthews, S. D., Zimmer, W. K., & McTigue, E. (2019). Developing and examining validity evidence for the Writing Rubric to Inform Teacher Educators (WRITE). *Assessing Writing*, 40, 1-13. DOI: 10.1016/j.asw.2019.03.001

- Hughes, S., McMullen, C., Jackson, L., McNamara, G. Matthews, L., Camaradou, J., & Walker, A. (2022). Development and validation of the symptom burden questionnaire for long covid (SBQ-LC): Rasch analysis. *BMJ*, 377, Article e070230. DOI: 10.1136/bmj-2022-070230
- Janssen, G., Meier, V., & Trace, J. (2015). Building a better rubric: Mixed methods rubric revision. *Assessing Writing*, 26, 51-66. DOI: 10.1016/j.asw.2015.07.002
- Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring and validating performance tasks*. The Guilford Press.
- Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2), 130-144. DOI: 10.1016/j.edurev.2007.05.002
- Jonsson, A., & Panadero, E. (2017). The use and design of rubrics to support assessment for learning. In D. Carless, S. M. Bridges, C. K. Y. Chan & R. Glofcheski (Eds.), *Scaling up assessment for learning in higher education* (pp. 99–111). Springer.
- Jönsson, A. (2022). *Scoring rubrics*. Routledge DOI: 10.4324/9781138609877-REE2-1
- Klimovski, D. (2013). Does a well-defined scoring rubric lead to an improvement in student results? *Proceedings of the 2013 AAEE Conference, Australia*.
- Lallmamode, S. P., Daud, N. M., & Kassim, N. L. A. (2016). Development and initial argument-based validation of a scoring rubric used in the assessment of L2 writing electronic portfolios. *Assessing Writing*, 30, 44-62. DOI: 10.1016/j.asw.2016.06.001
- Linacre, J. M. (1994a). *Many-faceted Rasch measurement*. MESA Press.
- Linacre, J. M. (1994b). Sample size and item calibration stability. *Rasch Measurement Transactions*, 7(4), 328.
- Linacre, J. M. (2002). Understanding Rasch measurement: Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, 3(1), 85-106.
- Linacre, J. M. (2003). Rasch power analysis: Size vs. significance: Standardized Chi-Square fit statistic. *Rasch Measurement Transactions*, 17(1), 918–919.
- Linacre, J. M. (2023). *A User Guide to FACETS* (Version 3.87.0) [Computer software].
<https://www.winsteps.com/a/Facets-Manual.pdf>
- Linacre, J. M. (2024). *A User Guide to WINSTEPS* (Version 5.8.1) [Computer software].
<https://www.winsteps.com/a/Winsteps-Manual.pdf>
- Maydeu-Olivares, A., & Montaña, R. (2013). How should we assess the fit of Rasch-type models? Approximating the power of goodness-of-fit statistics in categorical data analysis. *Psychometrika*, 78(1), 116-133. DOI: 10.1007/S11336-012-9293-1
- Morke, A. M., Dornan, T., & Eika, B. (2012). Outcome (competency) based education: An exploration of its origins, theoretical basis, and empirical evidence. *Advances in Health Sciences Education*, 18, 851-863. DOI: 10.1007/s10459-012-9405-9
- Moskal, B. M. (2000). Scoring rubrics: What, when and how? *Practical Assessment, Research, and Evaluation*, 7(7), 1–5. DOI: 10.7275/a5vq-7q66

- Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using Many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, 5(2), 189-227.
- Office of the Education Council. (2020). *Active competency-based learning*. 21 Century Print.
- Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, 9, 129-144. DOI: 10.1016/j.edurev.2013.01.002
- Pancorbo, G., Primi, R., John, O. P., Santos, D., & De Fruyt, F. (2021). Formative assessment of social-emotional skills using rubrics: A review of knowns and unknowns. *Frontiers in Education*, 6, Article 687661. DOI: 10.3389/feduc.2021.687661
- Popham, W. J. (1997). What's wrong - and what's right - with rubrics. *Educational Leadership*, 55(2), 72-75.
- Raykov, T., & DiStefano, C. (2021). Evaluating restrictive models in educational and behavioral research: Local misfit overrides model tenability. *Educational and Psychological Measurement*, 81(5), 980-995. DOI: 10.1177/0013164420944566.
- Reddy, M. Y. (2011). Design and development of rubrics to improve assessment outcomes: A pilot study in a Master's level business program in India. *Quality Assurance in Education*, 19(1), 84-104. DOI: 10.1108/09684881111107771
- Stevens, S. D., & Levi, A. J. (2012). *Introduction to rubrics: An assessment tool to save grading time, convey effective feedback, and promote student learning* (2nd ed.). Routledge.
- Thao, N. H. (2023). End-of-term speaking assessment rubric for grade 6 English teachers teaching new English textbooks in Vietnam. *International Journal of Education and Pedagogy*, 5(1), 127-144.
- Thepsathit, P., & Tangdhanakanond, K. (2024). Many-facet Rasch measurement model: Concepts and necessary statistic for performance assessment. *Journal of Education and Innovation*, 26(1), 435-448.
- Van der Linden, W. J. (2017, 31 January). *IRT-van-der-Linden-2*. <https://www.rasch.org/audio/IRT-van-der-Linden-2.mp3>
- Wilson, M. (2013). Seeking a balance between the statistical and scientific elements in psychometrics. *Psychometrika*, 78(2), 211-236. DOI: 10.1007/S11336-013-9327-3
- Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8, 370-371.
- Yuen, B. P. L., & Sawatdeenarunat, S. (2020). Applying a rubric development cycle for assessment in higher education: An evidence-based case study of a science communication module. *Asian Journal of the Scholarship of Teaching and Learning*, 10(1), 53-68.