Journal of Education Studies
วารสารครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ปีที่ 50 ฉบับที่ 3 EDUCU5003003 (14 pages) doi:

การตรวจสอบคุณสมบัติทางจิตมิติของรูบริกสำหรับการประเมินทักษะดนตรีไทย: การประยุกต์ใช้โมเดลการให้คะแนนบางส่วนแบบหลายองค์ประกอบของราสช์

Psychometric Properties Evaluation of Rubric for Assessing Thai Music Performance: An Application of Many-Facet Rasch Measurement Partial Credit Model

ภูรินท์ เทพสถิตย์[1*] และ กมลวรรณ ตังธนกานนท์[2]
Purin Thepsathit[1*] and Kamonwan Tangdhanakanond[2]

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อตรวจสอบคุณสมบัติทางจิตมิติของรูบริกสำหรับการประเมินทักษะดนตรีไทย โดยใช้โมเดลการให้คะแนนบางส่วนแบบหลายองค์ประกอบของราสช์ ตัวอย่างวิจัย คือ 1) นักเรียนระดับชั้นมัธยมศึกษาที่บรรเลงเดี่ยวเครื่องดนตรีไทยประเภทเครื่องตี 4 ชนิด ประกอบด้วย ระนาดเอก ระนาดทุ้ม ฆ้องวงใหญ่ และฆ้องวงเล็ก จำนวน 84 คน และ 2) ผู้ประเมินทักษะดนตรีไทยที่มีความเชี่ยวชาญและผ่านการอบรม จำนวน 6 ท่าน เครื่องมือวิจัย คือ รูบริกสำหรับการประเมินทักษะดนตรีไทย ประกอบด้วยเกณฑ์ 8 ด้าน 12 ข้อรายการประเมิน โดยแต่ละข้อรายการประเมินจะประกอบไปด้วยระดับคุณภาพ 5 ระดับ วิเคราะห์คุณสมบัติทางจิตมิติโดยใช้โมเดลการให้คะแนนบางส่วนแบบหลายองค์ประกอบของราสช์ผ่านฟาเซตที่เกี่ยวข้อง 4 ฟาเซต ประกอบด้วย ฟาเซตนักเรียน ฟาเซตผู้ประเมิน ฟาเซตเครื่องดนตรี และฟาเซตข้อรายการประเมิน ผลการวิจัยพบว่า 1) ผลความสอดคล้องกลมกลืนกับโมเดลของฟาเซตทั้ง 4 ฟาเซต ดัชนีคุณภาพฟาเซตผู้ประเมินและข้อรายการประเมิน ค่า point-measure correlation ของฟาเซตข้อรายการประเมิน และดัชนีประสิทธิผลระดับคุณภาพ แสดงถึงความตรงเชิงโครงสร้างของรูบริก และ 2) ผล Chi-square ฟาเซตนักเรียนและผู้ประเมิน และดัชนีคุณภาพฟาเซตนักเรียน แสดงถึงความเที่ยงของรูบริก

คำสำคัญ *:* การตรวจสอบคุณสมบัติทางจิตมิติ, รูบริก, การประเมินทักษะดนตรีไทย, โมเดลการให้คะแนนบางส่วนแบบหลายองค์ประกอบของราสช์

Article Info: Received 1 June, 2022; **Received in revised form** 22 June, 2022; **Accepted** 28 June, 2022

[1] นิสิตมหาบัณฑิต สาขาวิชาวิธีวิทยาการพัฒนานวัตกรรมทางการศึกษา ภาควิชาวิจัยและจิตวิทยาการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
อีเมล : purin.tepsatit@gmail.com
Graduate Student, Division of Innovation Development in Education Department of Educational Research and Psychology, Faculty of Education, Chulalongkorn University Email: purin.tepsatit@gmail.com

[2.] อาจารย์ประจำภาควิชาวิจัยและจิตวิทยาการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อีเมล : kamonwan.t@chula.ac.th
Lecturer of Department of Educational Research and Psychology, Faculty of Education, Chulalongkorn University
Email: kamonwan.t@chula.ac.th

* Corresponding Author

Abstract

The purpose of this study was to evaluate the psychometric properties of a rubric for assessing Thai music performance using the Many-facet Rasch measurement partial credit model. The samples of this study were 1) 84 secondary students who played 4 Thai classical Instruments including Ranad Eak, Ranad Thum, Khong Wong Yai and Khong Wong Lek; and 2) 6 trained raters who had experiences in Thai music performance. The 8 criterion and 12 item rubric for assessing Thai music performance included 5 categories for each item that was used in this study. The data were analyzed using the Many-facet Rasch measurement partial credit model across 4 facets including student, rater, instrument, and item. The results revealed that 1) The results of fit statistics across 4 facets, facet quality index of rater and item facet, point-measure correlation of item facet and category effectiveness index indicated the construct validity of the rubric and 2) Chi-square of student and rater facet and facet quality index of student facet indicated reliability of the rubric.

Keywords: psychometric properties evaluation, rubric, Thai music performance assessment, Many-facet Rasch measurement partial credit model

บทนำ

ผลการศึกษางานวิจัยที่เกี่ยวข้องกับการพัฒนาเครื่องมือวัดและประเมินผลทางดนตรีไทย พบว่า การตรวจสอบคุณสมบัติทางจิตมิติไม่เหมาะสมตามหลักการวัดและประเมินผลและตรวจสอบคุณสมบัติทางจิตมิติด้วยแนวคิดทฤษฎีการทดสอบแบบดั้งเดิมเท่านั้น (Classical Test Theory: CTT) ซึ่งทฤษฎี CTT มีข้อจำกัด คือ คุณสมบัติทางจิตมิติของเครื่องมือ เช่น ความยาก จะเปลี่ยนไปตามกลุ่มตัวอย่างที่ใช้ (ศิริชัย กาญจนวาสี, 2563) โดยคุณสมบัติทางจิตมิตินั้น หมายถึง ค่าสถิติต่าง ๆ สำหรับใช้อธิบายคุณสมบัติของเครื่องมือ เช่น ความตรง ความเที่ยง หรือความยาก และหากการประเมินมีองค์ประกอบอื่นที่เกี่ยวข้อง ควรแสดงค่าสถิติที่เกี่ยวข้องกับองค์ประกอบนั้นด้วย เช่น ความสอดคล้องของการประเมินระหว่างผู้ประเมิน (American Educational Research Association et al., 2014)

เครื่องมือการวัดและประเมินชนิดหนึ่งที่มีความเหมาะสมกับการประเมินทักษะดนตรี คือ รูบริก ซึ่งมีลักษณะคล้ายกับมาตรประมาณค่า (rating scale) ที่มีคำอธิบายเกี่ยวกับทักษะที่คาดหวังในแต่ละระดับ ช่วยให้เกิดความเป็นปรนัยของผลการประเมิน สามารถนำผลการประเมินที่ได้รับจากรูบริกไปใช้ในการพัฒนาผู้เรียน และกระตุ้นให้ผู้เรียนรับรู้ถึงความสามารถของตนและกระตุ้นตนเองในการซ้อมได้ (DeLuca & Bolden, 2014) องค์ประกอบที่สำคัญของรูบริกประกอบไปด้วย 4 องค์ประกอบ ได้แก่ 1) เกณฑ์ หมายถึง ทักษะเป้าหมายที่ต้องการประเมิน 2) ระดับคุณภาพ หมายถึง ระดับคุณภาพของทักษะที่ต้องการประเมิน มีลักษณะเรียงลำดับจากทักษะที่ปฏิบัติได้ง่ายไปสู่ปฏิบัติได้ยาก 3) คะแนน หมายถึง ตัวเลขที่ใช้สำหรับการให้คะแนนระดับคุณภาพแต่ละระดับ และ 4) คำอธิบายระดับคุณภาพ หมายถึง คำอธิบายระดับคุณภาพแต่ละระดับ โดยอธิบายเพื่อสะท้อนให้เห็นถึงเกณฑ์การผ่านของทักษะที่ต้องการในแต่ละระดับคุณภาพ

โมเดลการให้คะแนนบางส่วนแบบหลายองค์ประกอบของราสช์ (Many-facet Rasch Measurement Partial Credit Model: MFRM-PCM) เป็นโมเดลที่พัฒนาขึ้นบนแนวคิดของทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) ซึ่งเครื่องมือที่ผ่านการตรวจสอบด้วยทฤษฎี IRT นั้นจะมีคุณสมบัติทางจิตมิติที่คงที่ไม่แปรเปลี่ยน (invariance) ไปตามกลุ่มตัวอย่างที่ใช้ (ศิริชัย กาญจนวาสี, 2563) โดยโมเดล MFRM-PCM ใช้สำหรับการตรวจสอบเครื่องมือสำหรับการประเมินที่เกี่ยวข้องมากกว่า 2 ฟาเซตและมีผลการตอบมากกว่าสองค่า (polytomous) เช่น 1 - 5 คะแนน คำว่า ฟาเซตนั้น หมายถึง

องค์ประกอบที่เกี่ยวข้องและมีอิทธิพลต่อการประเมิน เช่น นักเรียน ผู้ประเมิน และข้อรายการประเมิน โดยแต่ละฟาเซตจะมีอิทธิพลต่อกัน ซึ่งโมเดล MFRM-PCM จะวิเคราะห์แต่ละฟาเซตอย่างเป็นอิสระและนำผลการประเมินที่ได้จากแต่ละฟาเซตมาใช้วิเคราะห์สิ่งที่ต้องการประเมินร่วมกัน (Linacre, 2022a) โมเดล MFRM-PCM พัฒนาขึ้นจากแนวคิดของ 2 โมเดล ได้แก่ โมเดลการให้คะแนนบางส่วน (Partial Credit Model) และโมเดลหลายองค์ประกอบของราสช์ (Many-Facet Rasch Measurement Model) โดยโมเดลจะประมาณค่าคุณสมบัติทางจิตมิติจากทุกฟาเซตที่เกี่ยวข้องและอนุญาตให้แต่ละระดับคุณภาพของข้อรายการประเมินแต่ละข้อมีความยากแตกต่างกันได้ ซึ่งหมายถึง ข้อรายการประเมินทุกข้ออาจไม่จำเป็นต้องมีจำนวนระดับคุณภาพที่เท่ากันก็ได้ ดังนั้นการวิเคราะห์ด้วยโมเดล MFRM-PCM ช่วยให้ผู้วิเคราะห์สามารถตรวจสอบคุณสมบัติทางจิตมิติของข้อรายการประเมินรายข้อและสามารถนำผลการวิเคราะห์ไปใช้เพื่อตัดสินใจสำหรับการปรับปรุงประสิทธิผลของระดับคุณภาพในแต่ละข้อรายการประเมินได้ (Eckes, 2009; Linacre, 2002; Linacre, 2022a) สำหรับการตรวจสอบคุณสมบัติทางจิตมิติครั้งนี้ เกี่ยวข้องกับ 4 ฟาเซต ประกอบด้วย ฟาเซตนักเรียน ฟาเซตผู้ประเมิน ฟาเซตเครื่องดนตรี และฟาเซตข้อรายการประเมิน โดยผลการวิเคราะห์แต่ละฟาเซตช่วยยืนยันหลักฐานคุณสมบัติทางจิตมิติของรูบริก โดยฟาเซตนักเรียนวิเคราะห์เพื่อยืนยันว่าข้อรายการประเมินมีความสามารถในการจำแนกทักษะของนักเรียน และยืนยันถึงความเที่ยงของรูบริก ฟาเซตผู้ประเมินวิเคราะห์การกดการปล่อยคะแนนที่ต่างกันของการประเมินทักษะนักเรียน ฟาเซตข้อรายการประเมินวิเคราะห์ความยากที่แตกต่างกันของข้อรายการประเมินและวิเคราะห์เพื่อยืนยันว่านักเรียนมีความสามารถแตกต่างกันมากพอ ซึ่งผลการวิเคราะห์ช่วยยืนยันถึงความตรงเชิงโครงสร้างของรูบริก (construct validity) (Linacre, 2022a) ส่วนฟาเซตเครื่องดนตรีช่วยยืนยันว่าเครื่องดนตรีแต่ละชนิดได้รับอิทธิพลการประเมินจากผู้ประเมินไม่แตกต่างกัน

ผู้วิจัยต้องการตรวจสอบคุณสมบัติทางจิตมิติของรูบริกสำหรับการประเมินทักษะดนตรีไทยที่พัฒนาขึ้นด้วยโมเดล MFRM-PCM ซึ่งรูบริกนี้มีเป้าหมายในการพัฒนาเพื่อประเมินเชิงพัฒนาการ (formative assessment) สำหรับการประเมินทักษะดนตรีไทย 4 ชนิด ประกอบด้วย ระนาดเอก ระนาดทุ้ม ฆ้องวงใหญ่ และฆ้องวงเล็ก โดยผู้วิจัยนำเสนอวิธีการตรวจสอบคุณสมบัติทางจิตมิติและอธิบายค่าดัชนีบ่งชี้ถึงคุณสมบัติทางจิตมิติต่าง ๆ รวมถึงเกณฑ์การตรวจสอบค่าดัชนีบ่งชี้ถึงคุณสมบัติทางจิตมิติที่เกี่ยวข้อง โดยมีฟาเซตในการตรวจสอบทั้งสิ้น 4 ฟาเซต

วัตถุประสงค์

เพื่อตรวจสอบคุณสมบัติทางจิตมิติของรูบริกสำหรับการประเมินทักษะดนตรีไทยโดยใช้โมเดลการให้คะแนนบางส่วนแบบหลายองค์ประกอบของราสช์

วิธีการวิจัย

ประชากรและกลุ่มตัวอย่าง คือ 1) นักเรียนระดับชั้นมัธยมศึกษา ที่บรรเลงเพลงเดี่ยวเครื่องดนตรีไทยประเภทเครื่องตี 4 ชนิด ได้แก่ ระนาดเอก ระนาดทุ้ม ฆ้องวงใหญ่ และฆ้องวงเล็ก และ 2) ผู้ประเมินทักษะดนตรีไทย ที่เป็นผู้มีประสบการณ์ในการสอนทักษะดนตรีไทย คัดเลือกกลุ่มตัวอย่างด้วยวิธีการเลือกแบบเฉพาะเจาะจง สำหรับกลุ่มตัวอย่างที่เป็นนักเรียน จะเป็นคลิปวิดีโอการบรรเลงเดี่ยวเครื่องดนตรีไทยทั้ง 4 ชนิดของนักเรียนที่ส่งเข้าร่วมการประกวดดนตรีไทย โดยขอความร่วมมือในการเก็บข้อมูลจาก 2 หน่วยงาน ได้แก่ 1) รายการประลองเพลงประเลงมโหรีของธนาคารกรุงเทพ และ 2) รายการการประกวดเดี่ยวเครื่องดนตรีไทยและขับร้องระดับชาติของมหาวิทยาลัยราชภัฏบ้านสมเด็จเจ้าพระยา โดยมีเกณฑ์ในการคัดเลือกกลุ่มตัวอย่างที่เป็นนักเรียน คือ 1) คัดเลือกเฉพาะการบรรเลงเดี่ยวเครื่องดนตรีไทยประเภทเครื่องตี 4 ชนิด ย้อนหลังไม่เกิน 10 ปี 2) คัดเลือกกลุ่มตัวอย่างจากบทเพลงที่บรรเลง โดยจะต้องเป็นเพลงเดี่ยวมาตรฐาน อัตราจังหวะ 3 ชั้นเดี่ยวทั้งเถา หรือเดี่ยวอัตราจังหวะเฉพาะ และ 3) คัดเลือกผลงานที่มีคุณภาพในด้านเสียงและด้านภาพ โดยทุกผลงานจะต้องมีคุณภาพของเสียงที่ดังชัดเจน และคุณภาพของภาพที่สามารถเห็นท่านั่ง บุคลิกภาพ หรือการจับไม้ได้อย่างชัดเจน คัดเลือก

จำนวนชนิดละ 21 คน รวมทั้งสิ้น 84 คน จำนวนกลุ่มตัวอย่างกำหนดจากเกณฑ์ขั้นต่ำ 2 ข้อ สำหรับการวิเคราะห์ด้วยโมเดล MFRM-PCM ได้แก่ 1) ต้องมีจำนวนไม่น้อยกว่า 50 คน (Linacre, 1994) และ 2) แต่ละข้อรายการประเมินต้องถูกประเมินอย่างน้อย 100 ครั้ง (Linacre, 1994) โดยงานวิจัยนี้นักเรียน 1 คน จะถูกประเมิน 2 ครั้งจากผู้ประเมินที่แตกต่างกัน ดังนั้น ข้อรายการประเมิน 1 ข้อจะถูกประเมินทั้งหมด 168 ครั้ง ซึ่งมากกว่าเกณฑ์ที่กำหนดไว้

สำหรับกลุ่มตัวอย่างที่เป็นผู้ประเมินทักษะของนักเรียน มีเกณฑ์การคัดเลือก คือ 1) เป็นผู้ที่มีประสบการณ์ในสอนวิชาดนตรีไทยหรือทักษะดนตรีไทยอย่างน้อย 2 ปี 2) เป็นผู้ที่มีประสบการณ์ในการพัฒนาและส่งผู้เรียนเข้าประกวดดนตรีไทย หรือเป็นผู้ที่เคยเป็นกรรมการตัดสินการแข่งขันดนตรีไทยอย่างน้อย 2 ปี และ 3) ต้องไม่เคยเป็นกรรมการการตัดสินในรายการประกวดดนตรีไทยที่เป็นที่มาของกลุ่มตัวอย่างที่เป็นนักเรียน โดยมีกลุ่มตัวอย่างที่เป็นผู้ประเมินที่ผ่านเกณฑ์การคัดเลือกทั้งสิ้น 6 คน

เครื่องมือที่ใช้ในการวิจัย คือ รูบริกสำหรับประเมินทักษะดนตรีไทยแบบแยกองค์ประกอบ (analytic rubric) ที่มีเกณฑ์ประเมิน 8 ด้าน 12 ข้อรายการประเมิน โดยแต่ละข้อรายการประเมินจะมีคำอธิบายระดับคุณภาพ 5 ระดับ มีเป้าหมายของคำอธิบายเพื่อประเมินและให้ข้อมูลย้อนกลับเชิงพัฒนาการหรือความก้าวหน้า โดยรูบริกสำหรับการประเมินทักษะดนตรีไทยนี้ พัฒนาขึ้นมาจากเกณฑ์มาตรฐานดนตรีไทยและเกณฑ์การประเมินของสำนักทบวงมหาวิทยาลัย (2544) เป็นหลัก ผ่านการตรวจสอบความตรงเชิงเนื้อหาด้วยการวิเคราะห์ค่าความสอดคล้องระหว่างทักษะและข้อรายการประเมินในรูบริก (IOC) โดยผู้ทรงคุณวุฒิ 5 ท่าน ประกอบไปด้วยผู้ทรงคุณวุฒิทางด้านดนตรีไทยระดับอุดมศึกษา 1 ท่าน ระดับมัธยมศึกษา 3 ท่าน และด้านการวัดและการประเมินผลจำนวน 1 ท่าน ผลการวิเคราะห์ พบว่า ทุกข้อรายการประเมินมีดัชนี IOC อยู่ในช่วง 0.6 ถึง 1.0 ซึ่งมากกว่า 0.5 ขึ้นไป ผ่านเกณฑ์ความตรงเชิงเนื้อหา โดยผู้วิจัยได้ปรับปรุงคำอธิบายระดับคุณภาพเพิ่มเติมตามคำแนะนำของผู้ทรงคุณวุฒิ และรูบริกนี้ผ่านการทดลองใช้ด้วยกลุ่มตัวอย่างที่เป็นนักเรียนจำนวน 40 คน และผู้ประเมินจำนวน 5 ท่าน ผลการทดลองใช้ พบว่า รูปแบบการประเมินที่ใช้สามารถใช้โมเดล MFRM-PCM ในการวิเคราะห์ได้ และข้อรายการประเมินทุกข้อผ่านเกณฑ์ความตรงเชิงโครงสร้างและความเที่ยง สามารถนำไปวิเคราะห์เพื่อตัดสินใจปรับปรุงประสิทธิผลระดับคุณภาพของข้อรายการประเมินได้

ตาราง 1

เกณฑ์และข้อรายการประเมินของรูบริกสำหรับการประเมินทักษะดนตรีไทย

เกณฑ์การประเมิน	ข้อรายการประเมิน
ด้านที่ 1 การเตรียมพร้อมเครื่องดนตรี	1. เครื่องดนตรี 2. ไม้ตี และ 3. เสียงเครื่องดนตรี
ด้านที่ 2 ท่านั่งและบุคลิกภาพ	4. ท่านั่งและบุคลิกภาพ
ด้านที่ 3 การจับไม้	5. การจับไม้
ด้านที่ 4 พื้นฐานและเทคนิคการบรรเลง	6. พื้นฐานการบรรเลง 7. การบรรเลงเทคนิคพื้นฐาน และ 8. การบรรเลงเทคนิคพิเศษ
ด้านที่ 5 ความถูกต้องของทำนองหลักและความเหมาะสมกับผู้บรรเลงแต่ละบุคคล	9. ความถูกต้องของทำนองหลักและความเหมาะสมกับผู้บรรเลงแต่ละบุคคล
ด้านที่ 6 ความถูกต้องของการบรรเลงเดี่ยวและความสอดคล้องกับแนวทำนองและจังหวะ	10. ความถูกต้องของการบรรเลงเดี่ยวและความสอดคล้องกับแนวทำนองและจังหวะ
ด้านที่ 7 คุณภาพเสียงตลอดการบรรเลง	11. คุณภาพเสียงตลอดการบรรเลง
ด้านที่ 8 ความสอดคล้องสัมพันธ์ของทักษะ	12. ความสอดคล้องสัมพันธ์ของทักษะ

การเก็บรวมรวบข้อมูล ผู้วิจัยดำเนินการอบรมผู้ประเมินด้วยเอกสารการอบรบผู้ประเมินที่จัดทำขึ้น มีเนื้อหาเกี่ยวกับ 1) วิธีการใช้ วัตถุประสงค์ และขอบเขตของรูบริกสำหรับการประเมินทักษะดนตรีไทย 2) ทักษะดนตรีไทยที่เป้าหมายของการประเมิน 3) ข้อมูลทั่วไปและการได้มาของตัวอย่างวิจัย และ 4) ความคลาดเคลื่อนที่อาจเกิดขึ้นจากผู้ประเมิน (rater bias/rater effect) และดำเนินการให้ผู้ประเมินกรอกแบบฟอร์มยินยอมเป็นผู้ประเมินและยินยอมที่จะไม่เผยแพร่คลิปวิดีโอของตัวอย่างวิจัย จากนั้นผู้วิจัยนำคลิปการบรรเลงเดี่ยวเครื่องดนตรีไทยของกลุ่มตัวอย่าง จำนวน 84 คน สุ่มให้กับผู้ประเมินแต่ละท่าน ใช้รูปแบบการประเมินแบบเชื่อมต่อประสาน ประเมินในรูปแบบออนไลน์ผ่าน Google Forms โดยผู้ประเมิน 1 ท่าน ประเมินนักเรียน 28 คน โดยใช้ข้อรายการประเมินทุกข้อ ซึ่งนักเรียน 1 คน ถูกประเมินโดยผู้ประเมินทั้งสิ้น 2 คน และผู้ประเมินทุกคนจะประเมินนักเรียนทับซ้อนกันกับผู้ประเมินคนก่อนหน้าและถัดไปจำนวน 14 คน (overlap) ผู้วิจัยสุ่มกลุ่มตัวอย่างให้ผู้ประเมินแต่ละท่านโดยมีเครื่องดนตรีชนิดละเท่า ๆ กัน ใช้รูปแบบการประเมินแบบเชื่อมต่อประสาน (linking design) (Engelhard, 1997) เป็นรูปแบบที่แนะนำสำหรับการวิเคราะห์ด้วยโมเดล MFRM-PCM เพื่อให้สอดคล้องกับภาระงานของผู้ประเมินที่ลดลงเมื่อเปรียบเทียบกับการให้ผู้ประเมินประเมินนักเรียนทุกคนและทุกข้อรายการประเมิน (complete design) หลีกเลี่ยงความเมื่อยล้า (fatigue) (กมลวรรณ ตังธนกานนท์, 2563) ที่ส่งผลต่อความเที่ยงของผู้ประเมิน โดยไม่มีผลกระทบต่อการประมาณค่าของโมเดล และคงคุณสมบัติทางจิตมิติภายในฟาเซตและระหว่างฟาเซตได้ (Linacre, 1997; Uto, 2021) จากนั้นดำเนินการรวบรวมผลการประเมินของผู้ประเมินทั้ง 6 ท่าน เพื่อนำไปใช้ในการวิเคราะห์เพื่อตรวจสอบคุณสมบัติทางจิตมิติด้านความตรงเชิงโครงสร้างและความเที่ยงของรูบริก จากนั้นวิเคราะห์ปรับปรุงประสิทธิผลของระดับคุณภาพแต่ละข้อรายการประเมินด้วยโมเดล MFRM-PCM

การวิเคราะห์ข้อมูล ผู้วิจัยวิเคราะห์ข้อมูลด้วยโมเดล MFRM-PCM โดยใช้โปรแกรม Minifac มีคุณสมบัติทางจิตมิติ 2 ด้านที่ต้องการตรวจสอบ ประกอบด้วย 1) ความตรงเชิงโครงสร้าง หมายถึง รูบริกสามารถประเมินทักษะดนตรีไทยที่เป็นเป้าหมายได้อย่างสอดคล้องกลมกลืนไปในทิศทางเดียวกันจากหลักฐานทางเชิงปริมาณที่ได้จากการวิเคราะห์ด้วยโมเดลผ่านฟาเซตที่เกี่ยวข้องกับการประเมิน โดยแต่ละฟาเซตจะมีหน้าที่ในการยืนยันและสนับสนุนความตรงเชิงของรูบริก พิจารณาจาก Fit Mean-Square ดัชนีคุณภาพฟาเซต และดัชนีประสิทธิผลระดับคุณภาพ และ 2) ความเที่ยง หมายถึง ความสามารถในการประเมินที่คงไว้ซึ่งคุณสมบัติทางจิตมิติด้านความตรงและความยุติธรรมของรูบริก เมื่อนำไปใช้กับกลุ่มตัวอย่างอื่น พิจารณาจาก Fixed/Random (χ^2) และดัชนีคุณภาพฟาเซต มีรายละเอียดของดัชนีบ่งชี้ถึงคุณสมบัติทางจิตมิติ ดังนี้

1) Chi-squared (χ^2) แบ่งการทดสอบเป็น 1.1) สำหรับโมเดลสรุปรวม ต้องการตรวจสอบความแตกต่างระหว่างโมเดลเชิงประจักษ์ (empirical model) และโมเดลราสช์ โดยต้องมีนัยสำคัญทางสถิติที่ระดับ .05 เพื่อแสดงให้เห็นว่าโมเดลราสช์ไม่สามารถทำนายโมเดลเชิงประจักษ์ได้อย่างสมบูรณ์ (Eckes, 2009; Linacre, 2022a) เนื่องจากโมเดลราสช์ต้องการให้ข้อมูลเชิงประจักษ์มีความแตกต่างกันเพื่อใช้สำหรับการวิเคราะห์ 1.2) สำหรับฟาเซตนักเรียน ฟาเซตผู้ประเมิน ฟาเซตเครื่องดนตรี และฟาเซตข้อรายการประเมินแบ่งเป็น (fixed) χ^2 มีสมมติฐาน คือ ข้อมูลภายในฟาเซตไม่แตกต่างกัน และ (random) χ^2 มีสมมติฐาน คือ กลุ่มตัวอย่างมาจากการสุ่มจากประชากรที่มีการแจงแจกแบบปกติ (normal distribution) สำหรับ (fixed) χ^2 ใช้สำหรับตรวจสอบความแตกต่างของความสามารถนักเรียน การกดปล่อยคะแนนของผู้ประเมิน และความยากง่ายของข้อรายการประเมินว่ามีความแตกต่างกันหรือไม่ โดยต้องมีนัยสำคัญทางสถิติที่ระดับ .05 เพื่อแสดงให้เห็นว่ามีความแตกต่างและช่วยยืนยันความตรงเชิงโครงสร้างของรูบริก ส่วนฟาเซตเครื่องดนตรี ต้องไม่มีนัยสำคัญทางสถิติ เพื่อแสดงให้เห็นว่า อิทธิพลการประเมินที่มีต่อแต่ละเครื่องดนตรีนั้นไม่ต่างกันสำหรับฟาเซตนักเรียนและฟาเซตผู้ประเมินนั้น แสดงผลการทดสอบ (random) χ^2 เพิ่มเติม เพื่อยืนยันว่า กลุ่มตัวอย่างที่ใช้วิเคราะห์ครั้งนี้เป็นนักเรียนที่มีความสามารถและเป็นผู้ประเมินที่มีการกดปล่อยคะแนนเท่ากับการสุ่มมาจากประชากรจากการแจกแจงแบบปกติ (Linacre, 2022a)

2) เมเชอร์ (measure) หน่วยเป็น Logit แสดงค่าความสามารถของนักเรียน การกดปล่อยคะแนนของผู้ประเมิน อิทธิพลการถูกประเมินของเครื่องดนตรี และความยากของข้อรายการประเมินภายในรูบริก โดยมีค่าอยู่ในช่วง $-\infty$ ถึง ∞ หากค่าเป็นบวก หมายถึง มีความสามารถสูง มีการกดคะแนน มีอิทธิพลการถูกประเมินแบบกดคะแนน และข้อรายการประเมินมีความยาก ส่วนค่าเป็นลบ หมายถึง มีความสามารถต่ำ มีการปล่อยคะแนน มีอิทธิพลการถูกประเมินแบบปล่อยคะแนน และข้อรายการประเมินมีความง่าย สำหรับฟาเซตข้อรายการประเมิน ควรมีค่าเมเชอร์อยู่ในช่วง -0.99 ถึง 0.99 เพื่อแสดงถึงความไม่ยากหรือไม่ง่ายจนเกินไปของข้อรายการประเมิน (Krishnan & Idris, 2018)

3) Fit Mean Square (Fit MS) คำนวณจากค่า χ^2 และ df เป็นค่าที่แสดงให้เห็นถึงความสอดคล้องกลมกลืนของข้อมูลกับโมเดล มีค่าที่คาดหวังเท่ากับ 1 หมายถึงผลการประเมินสอดคล้องกับโมเดลอย่างสมบูรณ์ ค่า Fit MS พิจารณาจากการวิเคราะห์ 2 ค่า ประกอบกัน ได้แก่ Infit MS ที่ละเอียดอ่อนต่อผลการประเมินที่เป็นแบบแผน (inlying) หมายถึง ค่า Infit MS จะมีค่าที่ใกล้เคียง 1 ต่อเมื่อปฏิบัติทักษะในข้อรายการประเมินที่ง่ายได้มากกว่าข้อรายการประเมินที่ยากอย่างเป็นระบบ และ Outfit MS ที่ละเอียดอ่อนต่อผลการประเมินแบบสุดโต่ง (outlying) หมายถึง ค่า Outfit MS จะมีค่าใกล้เคียง 1 ต่อเมื่อไม่มีผลการประเมินที่ผิดไปจากที่ควรจะเป็น ตัวอย่างผลการประเมินแบบสุดโต่ง เช่น มีความสามารถระดับต่ำแต่ถูกประเมินได้คะแนนในข้อที่ยากที่สุด จึงพิจารณาว่า การที่ปฏิบัติข้อที่ยากที่สุดได้เป็นผลสุดโต่งหรือคาดเดาไม่ได้ ค่าทั้ง 2 ควรมีค่าอยู่ที่ระหว่าง 0.5 ถึง 1.5 เพื่อยืนยันความสอดคล้องกลมกลืนกับโมเดลและเป็นหลักฐานยืนยันความตรงเชิงโครงสร้าง ค่า Fit MS แสดงผลพร้อมค่า Z-Standardized (ZSTD) ที่เป็นค่าที่ถูกปรับให้เป็นมาตรฐานจากผลการทดสอบ *t*-test ของค่า MS โดยมีสมมติฐานว่า ข้อมูลสอดคล้องกลมกลืนกับโมเดลอย่างสมบูรณ์ ค่า ZSTD ตั้งแต่ $\pm$2.0 พิจารณาว่าข้อมูลไม่สอดคล้องกลมกลืนกับโมเดลอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 Linacre, 2022a; Linacre, 2022b)

4) ดัชนีคุณภาพฟาเซต พิจารณาจาก 4.1) ดัชนีจำแนก (separation index) แสดงให้เห็นถึงสัดส่วนของความแตกต่างที่แท้จริง (True Standard Deviation: True SD) กับความคลาดเคลื่อนที่เกิดจากการวัด (Root Mean Square Error: RMSE) มีค่าตั้งแต่ 0 ขึ้นไป เป็นดัชนีที่แสดงให้เห็นถึงจำนวนระดับความแตกต่างของข้อมูลภายในฟาเซต โดยอนุมานว่าข้อมูลเป็นการแจกแจงแบบปกติ ดังนั้น ผลการประเมินที่มีค่าที่สูงมากหรือต่ำมาก เป็นผลมาจากความสามารถที่สุดโต่งที่อธิบายไม่ได้หรือเป็นความบังเอิญ มีสูตรคำนวณ คือ $True\ SD/RMSE$ 4.2) ดัชนีความเที่ยงจำแนก (reliability of separation) แสดงให้เห็นถึงสัดส่วนของความแปรปรวนที่แท้จริง (true variance) กับความแปรปรวนที่สังเกตได้ (observed variance) มีค่าอยู่ระหว่าง 0 ถึง 1 ยิ่งมีค่าเข้าใกล้ 1 หมายถึงมีความแตกต่างกันของข้อมูลภายในฟาเซตมาก มีสูตรคำนวณ คือ $True\ SD^2/(RMSE^2 + True\ SD^2)$ และ 4.3) ดัชนีระดับชั้น (strata index) เป็นดัชนีที่แสดงจำนวนระดับคล้ายกับดัชนีจำแนก แต่อนุมานว่าข้อมูลที่มีผลการประเมินความสามารถที่สูงมากหรือต่ำมาก เป็นผลจากความสามารถที่เกิดขึ้นจริง มีสูตรคำนวณ คือ $4 \times (True\ SD/RMSE) + 1)/3$ สำหรับฟาเซตนักเรียนต้องการให้มีค่าดัชนีคุณภาพที่สูง เพื่อยืนยันว่ารูบริกมีความสามารถในการจำแนกความสามารถของผู้เรียนได้ เป็นดัชนีที่บ่งชี้ถึงความเที่ยง ฟาเซตข้อรายการประเมินต้องการให้มีค่าที่สูง เพื่อยืนยันว่านักเรียนมีจำนวนกลุ่มตัวอย่างเพียงพอที่จะยืนยันความตรงเชิงโครงสร้างของรูบริก ฟาเซตผู้ประเมินต้องการให้มีค่าที่สูง เพื่อให้ผู้ประเมินประเมินอย่างอิสระแตกต่างกันได้ตามความเชี่ยวชาญของตนและเป็นการยืนยันความตรงเชิงโครงสร้าง ส่วนฟาเซตเครื่องดนตรีต้องการให้มีค่าที่ต่ำ เนื่องจากไม่ต้องการให้มีอิทธิพลของการประเมินที่เกิดจากจากผู้ประเมินหรือรูบริกของเครื่องดนตรีแต่ละชนิดแตกต่างกัน

5) ดัชนีประสิทธิผลระดับคุณภาพ เป็นดัชนีที่บ่งชี้ถึงความตรงเชิงโครงสร้าง ประกอบไปด้วยดัชนีบ่งชี้ที่จำเป็น ดังนี้ (Linacre, 2002) 5.1) ค่า point-measure correlation (PT_{cor}) ของฟาเซตข้อรายการประเมิน ค่า PT_{cor} เป็นการวิเคราะห์ค่า Pearson point-biserial correlation ในรูปแบบของโมเดล MFRM-PCM วิเคราะห์เพื่อพิจารณาความสอดคล้องไปในทิศทางเดียวกันของข้อรายการประเมินภายในรูบริกกับทักษะที่ต้องการประเมิน ช่วยยืนยันความตรงเชิงโครงสร้างของรูบริก โดยค่า PT_{cor} ของแต่ละข้อรายการประเมินควรมีค่าเป็นบวก (Linacre, 2022a; Mohaffyza et al., 2015) เป็นค่าที่

ช่วยยืนยันความคงที่และความถูกต้องของการประมาณค่า ช่วยในการอธิบายคุณสมบัติทางจิตมิติของรูบริกจากกลุ่มตัวอย่างที่ใช้และยืนยันผลการประมาณค่าเมื่อนำไปใช้กับกลุ่มตัวอย่างอื่น 5.2) จำนวนการใช้ระดับคุณภาพ (category usage) ช่วยในการยืนยันความคงที่ของการประมาณค่า การใช้ระดับคุณภาพจำนวนอย่างน้อย 10 ครั้ง ช่วยยืนยันว่าระดับคุณภาพนั้น ๆ สามารถใช้ประเมินทักษะที่ต้องการในระดับนั้นได้จริง ระดับคุณภาพที่มีการใช้ไม่ถึง 10 ครั้ง ควรพิจารณานำระดับคุณภาพนั้นออกหรือยุบรวมระดับคุณภาพนั้นกับระดับคุณภาพต่อไปที่มีจำนวนการใช้ผ่านเกณฑ์ 5.3) ค่าเฉลี่ยเมเชอร์ (average measure: AM) เป็นค่าความสามารถเฉลี่ยในแต่ละระดับคุณภาพจากการประมาณค่าจากฟาเซตที่เกี่ยวข้อง นักเรียนที่มีความสามารถที่สูงควรต้องถูกประเมินในระดับคุณภาพที่สูง ดังนั้น ค่าเมเชอร์เฉลี่ยควรมีค่าจากต่ำไปสูงเรียงขึ้นไปตามลำดับ (monotonicity) สอดคล้องกับความสามารถ ค่าเฉลี่ยเมเชอร์ช่วยยืนยันความถูกต้องของการประมาณค่า และช่วยในการอธิบายคุณสมบัติทางจิตมิติของรูบริกจากกลุ่มตัวอย่างที่ใช้ รวมถึงยืนยันผลการประมาณค่าเมื่อนำไปใช้กับกลุ่มตัวอย่างอื่น และ 5.4) ค่า Outfit MS เป็นที่ละเอียดอ่อนต่อผลการประเมินแบบสุดโต่ง หากมีค่ามากกว่า 1.0 จะแสดงให้เห็นถึงผลการประเมินแบบสุดโต่งหรือผลการประเมินที่ผิดไปจากที่ควรเป็น ค่า Outfit MS ควรมีค่าไม่เกิน 2.0 ซึ่งหมายถึง ค่าสุดโต่งที่เกิดขึ้นสามารถประมาณได้ด้วยโมเดล หากค่า AM ของระดับคุณภาพใดมีค่าน้อยกว่าระดับคุณภาพก่อนหน้า หรือค่า Outfit MS ของระดับคุณภาพใดมีค่ามากกว่า 2.0 ควรพิจารณานำระดับคุณภาพนั้นออกหรือยุบรวมระดับคุณภาพนั้นกับระดับคุณภาพก่อนหน้า

ผลการวิจัย

ตอนที่ 1 ผลการวิเคราะห์โมเดลสรุปและฟาเซต

ผลการวิเคราะห์โมเดลสรุป (summary model) พบว่า โมเดลเชิงประจักษ์มีความแตกต่างจากโมเดลราสช์อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 (χ^2 = 41926.10, *df* = 839, $p < .01$)

ผลการวิเคราะห์ฟาเซตนักเรียนหลังการนำนักเรียน 5 คน ออกจากการวิเคราะห์เนื่องจากมีผลการประเมินที่ไม่สมบูรณ์ของผู้ประเมิน พบว่า มีนักเรียนจำนวน 18 คน มีค่า Infit - Outfit MS ไม่อยู่ในช่วง 0.5 ถึง 1.5 แสดงความไม่สอดคล้องกลมกลืนกับโมเดล มีระดับความแตกต่างของความสามารถนักเรียนประมาณ 1 ถึง 2 ระดับ (เก่ง/อ่อน) (ดัชนีจำแนก = 1.47, ดัชนีระดับชั้น = 2.29 และดัชนีความเที่ยงเท่ากับ 0.68) (Linacre, 2022b) ยืนยันถึงความเที่ยงของรูบริก การทดสอบ (fixed) χ^2 พบว่า ฟาเซตนักเรียนมีความสามารถแตกต่างกันอย่างน้อย 2 คน อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 (χ^2 = 279.2, *df* = 78, $p < .01$) และการทดสอบ (random) χ^2 พบว่า กลุ่มตัวอย่างที่ใช้ในการวิเคราะห์เป็นนักเรียนที่มีความสามารถเท่ากับการสุ่มมาจากประชากรที่มีการแจกแจงแบบปกติ (χ^2 = 56, *df* = 77, $p > .01$)

ผลการวิเคราะห์ฟาเซตผู้ประเมิน พบว่า ผู้ประเมินทั้ง 6 คน มีความสอดคล้องกลมกลืนกับโมเดล มีระดับความแตกต่างของการกดปล่อยคะแนนประมาณ 7 ถึง 10 ระดับ (ดัชนีจำแนก = 7.52, ดัชนีระดับชั้น = 10.35 และดัชนีความเที่ยงเท่ากับ 0.98) การทดสอบ (fixed) χ^2 พบว่า ผู้ประเมินประเมินโดยมีการกดปล่อยแตกต่างกันอย่างน้อย 2 คน อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 (χ^2 = 307, *df* = 5, $p < .01$) และการทดสอบ (random) χ^2 พบว่า กลุ่มตัวอย่างที่ใช้ในการวิเคราะห์ครั้งนี้เป็นผู้ประเมินที่มีการกดปล่อยคะแนนเท่ากับการสุ่มมาจากประชากรที่มีการแจกแจงแบบปกติ (χ^2 = 4.9, *df* = 4, $p > .01$)

ผลการวิเคราะห์ฟาเซตเครื่องดนตรี พบว่า เครื่องดนตรีทั้ง 4 ชนิด มีความสอดคล้องกลมกลืนกับโมเดล ไม่พบระดับความแตกต่างของระดับอิทธิพลของการถูกประเมินที่แตกต่างกันหรือพบเพียงเล็กน้อย (ดัชนีจำแนก = 0.00, ดัชนีระดับชั้น = 0.33 และ ดัชนีความเที่ยง = 0.00) การทดสอบ (fixed) χ^2 พบว่า เครื่องดนตรีแต่ละชนิดได้รับอิทธิพลของการถูกประเมินไม่แตกต่างกัน (χ^2 = 1.6, *df* = 3, $p > .01$)

ผลการวิเคราะห์ฟาเซตข้อรายการประเมิน พบว่า ข้อรายการประเมินมีระดับความแตกต่างของความยากประมาณ 3 ถึง 4 ระดับ จากผลการวิเคราะห์ดัชนีคุณภาพฟาเซต (ดัชนีจำแนก = 2.71, ดัชนีระดับชั้น = 3.95 และดัชนีความเที่ยง = 0.88) ผลการวิเคราะห์ช่วยยืนยันว่ากลุ่มตัวอย่างที่เป็นนักเรียนมีคุณภาพเพียงพอที่จะยืนยันความสามารถในการจำแนกระดับทักษะของนักเรียนด้วยข้อรายการประเมินภายในรูบริก และยืนยันว่ามีการจำแนกที่มากพอสำหรับยืนยันความตรงเชิงโครงสร้างของรูบริกด้วย (Linacre, 2022a) การทดสอบ (fixed) χ^2 พบว่า รูบริกมีความยากของข้อรายการประเมินแตกต่างกันอย่างน้อย 2 ข้อ อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 (χ^2 = 81.8, *df* = 11, $p < .01$)

ตาราง 2

ผลการวิเคราะห์สรุปรายฟาเซต

	ฟาเซต			
	นักเรียน (*N* = 79)	ผู้ประเมิน (*N* = 6)	เครื่องดนตรี (*N* = 4)	ข้อรายการฯ (*N* = 12)
ค่าเฉลี่ยเมเชอร์	2.00 (*SD* = 0.74)	0.00 (*SD* = 0.86)	0.00 (*SD* = 0.05)	0.00 (*SD* = 0.42)
เมเชอร์สูงสุด	4.64	0.96	0.04	0.62
เมเชอร์ต่ำสุด	0.26	-1.40	-0.08	-0.86
Infit MS (*M*)	0.98 (*SD* = 0.30)	0.99 (*SD* = 0.17)	1.00 (*SD* = 0.06)	0.99 (*SD* = 0.14)
Infit MS สูงสุด	1.73	1.22	1.08	1.17
Infit MS ต่ำสุด	0.46	0.78	0.97	0.63
Outfit MS (*M*)	1.02 (*SD* = 0.41)	1.02 (*SD* = 0.11)	1.02 (*SD* = 0.09)	1.02 (*SD* = 0.14)
Outfit MS สูงสุด	2.22	1.09	1.10	1.56
Outfit MS ต่ำสุด	0.23	0.80	0.98	0.58
ดัชนีจำแนก	1.47	7.52	0.00	2.71
ดัชนีระดับชั้น	2.29	10.35	0.33	3.95
ดัชนีความเที่ยงฯ	0.68	0.98	0.00	0.88
(Fixed) χ^2	279.2 (df = 78, $p < .01$)	307 (df = 5, $p < .01$)	1.6 (df = 3, $p > .01$)	81.8 (df = 11, $p < .01$)
(Random) χ^2	56 (df = 77, $p > .01$)	4.9 (df = 4, $p > .01$)	1.0 (df = 2, $p > .01$)	9.6 (df = 10, $p > .01$)

ผลการวิเคราะห์ฟาเซตข้อรายการประเมิน พบว่า ข้อรายการประเมินทั้ง 12 ข้อมีค่าเมเชอร์อยู่ในช่วง -0.99 ถึง 0.99 แสดงถึงความยากง่ายที่เหมาะสม ข้อที่ง่ายที่สุด คือ ข้อรายการประเมินที่ 1 (เมเชอร์ = -0.82) และข้อที่ยากที่สุด คือ ข้อรายการประเมินที่ 8 (เมเชอร์ = 0.62) การตรวจสอบความสอดคล้องกลมกลืน พบว่า ข้อรายการประเมินที่ 10 (ความถูกต้องของการบรรเลงเดี่ยวและความสอดคล้องกับแนวทำนองและจังหวะ) ไม่สอดคล้องกลมกลืนกับโมเดล มีค่า Outfit MS เท่ากับ 1.56 ซึ่งมากกว่า 1.5 แสดงถึงการใช้ระดับคุณภาพด้านบนสุดต่ำสุดมากเกินไปหรือเกิดค่าผิดปกติในการประเมิน ผลการวิเคราะห์พบว่าไม่สอดคล้องกลมกลืนกับโมเดลอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 (Outfit ZSTD > 1.96) (Linacre, 2022b)

ตาราง 3

ผลการวิเคราะห์ฟาเซตข้อรายการประเมิน

เกณฑ์การประเมิน	ข้อรายการประเมิน	เมเซอร์	*SE*	Infit		Outfit	
				MS	ZSTD	MS	ZSTD
ด้านที่ 1	1	-0.86 (ง่าย)	0.21	1.08	0.4	0.93	-0.1
	2	-0.58 (ง่าย)	0.15	0.96	-0.1	0.90	-0.1
	3	-0.39 (ง่าย)	0.18	0.99	0.0	0.97	0.0
ด้านที่ 2	4	0.06 (ยาก)	0.11	1.12	1.0	1.07	0.4
ด้านที่ 3	5	0.10 (ยาก)	0.12	1.17	1.0	1.43	2.1
ด้านที่ 4	6	0.24 (ยาก)	0.12	1.09	0.6	0.98	0.0
	7	0.07 (ยาก)	0.14	1.04	0.3	1.01	0.1
	8	0.62 (ยาก)	0.12	0.89	-0.8	0.84	-1.2
ด้านที่ 5	9	0.20 (ยาก)	0.15	0.90	-0.8	1.03	0.2
ด้านที่ 6	10*	-0.15 (ง่าย)	0.13	1.12	0.7	1.56	2.6
ด้านที่ 7	11	0.56 (ยาก)	0.15	0.93	-0.6	0.90	-0.6
ด้านที่ 8	12	0.13 (ยาก)	0.12	0.63	-3.1	0.58	-3.8

ตอนที่ 2 ผลการวิเคราะห์ปรับปรุงประสิทธิผลของระดับคุณภาพ

1) ผลการวิเคราะห์ค่า PT_{cor} ของข้อรายการประเมินทุกข้อ (ตาราง 3) มีค่าเป็นบวก แสดงถึงความสอดคล้องไปในทิศทางเดียวกันของข้อรายการประเมินกับทักษะที่ต้องการประเมิน

2) ผลการวิเคราะห์รายข้อ พบว่า 2.1) ข้อรายการประเมินที่ 1 มีจำนวนการใช้ระดับคุณภาพที่ 1 2 และ 3 น้อยกว่า 10 ครั้ง และมีค่าเฉลี่ยเมเซอร์ในระดับคุณภาพที่ 4 น้อยกว่าระดับคุณภาพที่ 3 ควรยุบรวมระดับคุณภาพที่ 1 2 3 และ 4 เข้าด้วยกัน 2.2) ข้อรายการประเมินที่ 2 6 8 9 11 และ 12 มีจำนวนการใช้ระดับคุณภาพที่ 1 และ 2 น้อยกว่า 10 ครั้ง ควรยุบรวมระดับคุณภาพที่ 1 และ 2 เข้ากันกับระดับคุณภาพที่ 3 2.3) ข้อรายการประเมินที่ 3 และ 7 มีจำนวนการใช้ระดับคุณภาพที่ 1 2 และ 3 น้อยกว่า 10 ครั้ง ควรยุบรวมระดับคุณภาพที่ 1 2 และ 3 เข้ากันกับระดับที่ 4 2.4) ข้อรายการประเมินที่ 4 มีจำนวนการใช้ระดับคุณภาพที่ 1 และ 2 น้อยกว่า 10 ครั้ง และมีค่าเฉลี่ยเมเซอร์ในระดับคุณภาพที่ 3 น้อยกว่าระดับคุณภาพที่ 2 ควรยุบรวมระดับคุณภาพที่ 1 2 และ 3 เข้าด้วยกัน 2.5) ข้อรายการประเมินที่ 5 มีจำนวนการใช้ระดับคุณภาพที่ 1 และ 2 น้อยกว่า 10 ครั้ง และมีค่าเฉลี่ยเมเซอร์ในระดับคุณภาพที่ 2 น้อยกว่าระดับคุณภาพที่ 1 ควรยุบรวมระดับคุณภาพที่ 1 และ 2 เข้ากันกับระดับคุณภาพที่ 3 และ 2.6) ข้อรายการประเมินที่ 10 มีจำนวนการใช้ระดับคุณภาพที่ 1 และ 2 น้อยกว่า 10 ครั้ง และระดับคุณภาพที่ 4 มีค่า Outfit MS มากกว่า 2.0 ควรยุบรวมระดับคุณภาพที่ 1 2 และ 4 เข้ากันกับระดับคุณภาพที่ 3

ตาราง 4

ผลการวิเคราะห์การปรับปรุงประสิทธิผลระดับคุณภาพของรายข้อรายการประเมิน

ข้อ	จำนวนการใช้ระดับคุณภาพ					ค่าเฉลี่ยเมเซอร์					Outfit MS				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1	0*	0*	2*	24	132	-	-	2.11	1.98*	2.99	-	-	1.6	0.8	1.1
2	0*	1*	10	17	130	-	0.78	1.32	1.62	2.77	-	0.6	0.8	1.0	0.9
3	0*	2*	1*	24	131	-	0.96	0.97	1.49	2.55	-	1.2	0.6	0.9	1.0
4	1*	5*	30	37	85	-0.33	0.93	0.91*	1.8	2.39	0.6	1.4	0.6	1.4	1.2
5	2*	2*	15	46	93	0.39	-0.12*	1.13	1.56	2.21	1.1	0.3	1.8	1.7	1.1
6	0*	5*	15	37	101	-	0.27	0.66	1.31	2.11	-	0.8	0.8	1.1	1.1
7	0*	3*	9*	52	94	-	0.32	0.74	1.46	2.29	-	0.8	0.9	1.1	1.0
8	0*	6*	16	71	65	-	-0.09	0.05	1.11	1.06	-	1.0	0.4	0.9	1.0
9	0*	0*	10	38	110	-	-	0.33	1.14	2.11	-	-	0.8	1.2	0.9
10	1*	1*	12	45	99	-0.27	0.71	0.98	1.98	2.36	0.3	0.9	0.8	2.5*	1.2
11	0*	0*	12	60	86	-	-	-0.15	1.02	1.90	-	-	0.6	1.0	1.0
12	1*	3*	20	69	65	-0.55	0.36	0.60	1.38	2.82	0.3	0.7	0.6	0.4	0.7

หมายเหตุ : สัญลักษณ์ * หมายถึง ระดับคุณภาพควรได้รับการปรับปรุงประสิทธิผล

อภิปรายผล

ผู้วิจัยขอแบ่งการอภิปรายผลเป็น 2 ช่วง คือ การอภิปรายสรุปดัชนีบ่งชี้ที่เชื่อมโยงให้เห็นถึงหลักฐานของคุณสมบัติทางจิตมิติของรูบริกสำหรับการประเมินทักษะดนตรีไทยที่พัฒนาขึ้น และการวิเคราะห์ค่าสถิติที่เป็นดัชนีบ่งชี้รายฟาเซตทั้ง 4 ฟาเซต รายละเอียดดังต่อไปนี้

การอภิปรายสรุปดัชนีบ่งชี้ พบว่า ด้านความตรงเชิงโครงสร้าง จากผลการวิเคราะห์ความสอดคล้องกลมกลืนกับโมเดลด้วยสถิติ Fit MS ในฟาเซตข้อรายการประเมิน พบว่า ข้อรายการประเมินส่วนใหญ่มีความตรงเชิงโครงสร้าง โดยข้อรายการประเมินที่ 10 ที่แสดงความไม่สอดคล้องกลมกลืนกับโมเดลนั้น เมื่อทำการปรับปรุงด้วยการวิเคราะห์ประสิทธิผลระดับคุณภาพแล้ว ก็ช่วยทำให้ข้อรายการประเมินนี้มีความตรงเชิงโครงสร้างมากขึ้น และค่า PT_{cor} ที่มีผลเป็นบวกทุกข้อ ช่วยยืนยันความตรงเชิงโครงสร้างของรูบริก แสดงให้เห็นว่า ข้อรายการประเมินทุกข้อมุ่งประเมินในทักษะเดียวกัน นอกจากนี้ ผลการวิเคราะห์ค่า Chi-square ของฟาเซตนักเรียนและผู้ประเมินและดัชนีคุณภาพฟาเซตของฟาเซตผู้ประเมินและฟาเซตข้อรายการประเมิน ยังช่วยยืนยันถึงความตรงเชิงโครงสร้างของรูบริกด้วย เนื่องจาก ค่าสถิติที่กล่าวมานี้ช่วยยืนยันว่า นักเรียนมีความสามารถและผู้ประเมินมีการกดปล่อยคะแนนที่หลากหลายมากพอที่จะช่วยยืนยันการวิเคราะห์คุณสมบัติทางจิตมิติของรูบริกที่พัฒนาขึ้น นอกจากนี้ การวิเคราะห์ปรับปรุงประสิทธิผลระดับคุณภาพด้วยดัชนีประสิทธิผลระดับคุณภาพ ช่วยในการยืนยันความคงที่ของคุณสมบัติทางจิตมิติเมื่อนำไปใช้กับกลุ่มตัวอย่างอื่น เนื่องจากเป็นการปรับจำนวนระดับคุณภาพจากฟาเซตที่เกี่ยวข้อง ยืนยันด้วยค่าสถิติแล้วว่ามีความแตกต่างหลากหลายมากพอ ดังนั้น เมื่อปรับจำนวนระดับคุณภาพตามผลการวิเคราะห์แล้วนั้น ก็ทำให้รูบริกมีโครงสร้างที่สอดคล้องกับคุณภาพทักษะในภาพรวมมากขึ้น

ในส่วนความเที่ยง มีดัชนีบ่งชี้จากผลการวิเคราะห์ดัชนีคุณภาพฟาเซตของฟาเซตนักเรียนนั้น มีค่าสูงกว่าเกณฑ์ที่ยอมรับได้ แสดงถึงความสามารถของรูบริกในการจำแนกทักษะของผู้เรียนได้ ค่าสถิติที่อยู่ในดัชนีคุณภาพฟาเซตนี้ช่วยในเรื่องของการยืนยันว่า รูบริกสำหรับการประเมินทักษะดนตรีไทยนี้ เมื่อนำไปใช้ประเมินครั้งถัดไป ก็จะได้ผลของคุณสมบัติทางจิตมิติที่ไม่เปลี่ยนไป นอกจากนี้ จากการวิเคราะห์ค่า inter-rater agreement ของผู้ประเมินที่ผ่านเกณฑ์นั้น

ก็ช่วยยืนยันความเที่ยงของการนำรูบริกไปใช้ เนื่องจากผลการวิเคราะห์ที่ได้ แสดงถึงความประเมินอย่างเป็นอิสระซึ่งกันและกันของผู้ประเมิน แต่ก็มีความสอดคล้องเห็นไปในทิศทางเดียวกับกับทักษะที่กำลังประเมินจากการวิเคราะห์ค่า Fit MS จะช่วยยืนยันว่า หากนำรูบริกไปใช้กับผู้ประเมินทั่วไป ก็จะมีคุณสมบัติทางจิตมิติที่ไม่เปลี่ยนไปมากนัก หากผู้วิจัยดำเนินการออกแบบการวิจัยนี้ให้ผู้ประเมินต้องประเมินให้เห็นพ้องกันอย่างสมบูรณ์นั้น ค่าสถิติที่จากการวิเคราะห์คุณสมบัติทางจิตมิติในครั้งนี้ก็จะไม่สามารถยืนยันได้ว่าจะคงที่หากนำไปใช้กับผู้ประเมินโดยทั่วไป นอกจากนี้ ผลการวิเคราะห์ Chi-square และดัชนีคุณภาพฟาเซตของฟาเซตผู้ประเมิน ฟาเซตเครื่องดนตรี และฟาเซตข้อรายการประเมิน ยังช่วยยืนยันถึงความเที่ยงของรูบริกด้วย นอกจากนี้การวิเคราะห์ฟาเซตเครื่องดนตรีนั้น ค่าเมเซอร์ที่วิเคราะห์ได้ยังช่วยยืนยันว่าไม่มีความแตกต่างกันของอิทธิพลการถูกประเมินเมื่อใช้ประเมินในเครื่องดนตรีที่ต่างกันอีกด้วย

ผลการวิเคราะห์ฟาเซตนักเรียน พบว่า มีนักเรียนจำนวน 18 คน จาก 79 คนที่มีค่า Infit-Outfit MS ไม่สอดคล้องกลมกลืนกับโมเดล การพิจารณาค่า Fit MS สำหรับฟาเซตนักเรียนนั้น มีจุดประสงค์เพื่อใช้ศึกษาลักษณะนิสัยหรือรูปแบบการตอบที่มีลักษณะผิดปกติ เช่น ไม่ใส่ใจตอบ ผู้วิจัยสามารถนำข้อมูลของนักเรียนที่ไม่สอดคล้องกลมกลืนออกจากการวิเคราะห์ จนกว่าจะได้ผลค่าดัชนีที่ต้องการ (Linacre, 2010) สำหรับงานวิจัยในครั้งนี้ ผู้วิจัยพิจารณาว่านักเรียนที่ไม่สอดคล้องกลมกลืนกับโมเดลนั้นเป็นผลการประเมินจากผู้ประเมินที่ใช้ข้อรายการประเมิน ดังนั้น การนำผลการประเมินของนักเรียนที่ไม่สอดคล้องกลมกลืนกับโมเดลออกจึงเป็นเรื่องไม่สมเหตุสมผล ควรวินิจฉัยผลที่เกิดขึ้นจากตรวจสอบฟาเซตผู้ประเมินและฟาเซตข้อรายการประเมินกับโมเดลมากกว่า ซึ่งจะเห็นได้ว่า ผู้ประเมินทุกคนมีความสอดคล้องกลมกลืนกับโมเดล แต่มีข้อรายการประเมินบางข้อที่ไม่สอดคล้องกลมกลืนกับโมเดล คือ ข้อรายการประเมินที่ 10 ซึ่งจากการวินิจฉัยพบว่า ข้อรายการประเมินที่ 10 มีค่าผิดปกติเกิดขึ้น ดังนั้น ความไม่สอดคล้องกลมกลืนของนักเรียนบางส่วนที่เกิดขึ้นอาจต้องพิจารณาปรับปรุงในส่วนข้อรายการประเมินนี้นอกจากนี้จากการทดลองของผู้วิจัยเอง ช่วยยืนยันว่า หากนำเอาผลการวิเคราะห์ครั้งนี้ไปทำการลบนักเรียนที่ไม่สอดคล้องกลมกลืนกับโมเดลออก เมื่อวิเคราะห์ซ้ำ พบว่า จะมีนักเรียนที่ไม่สอดคล้องกลมกลืนกับโมเดลอีก กระทั่งการลบครั้งที่ 4 จึงจะให้ผลทุกคนสอดคล้องกลมกลืนกับโมเดล อย่างไรก็ตามสิ่งที่ผู้วิจัยพบ คือ หากลบข้อมูลของนักเรียนไม่สอดคล้องกลมกลืนกับโมเดลออกจนหมด ดัชนีบ่งชี้คุณสมบัติทางจิตมิติส่วนใหญ่มีคุณภาพลดลงและไม่ผ่านเกณฑ์ที่ยอมรับได้ ดังนั้น ผู้วิจัยจึงพิจารณาใช้แนวทางที่ศึกษาที่กล่าวไว้ข้างต้น คือ การคงนักเรียนที่ไม่สอดคล้องกลมกลืนกับโมเดลไว้ เนื่องจากพิจารณาว่าเป็นผลจากข้อรายการประเมินและผู้ประเมิน และรักษาสารสนเทศที่จะได้รับจากการวิเคราะห์ รวมถึงคงไว้ซึ่งหลักฐานคุณสมบัติทางจิตมิติที่มีคุณภาพของรูบริกด้วย

ส่วนผลการวิเคราะห์ดัชนีคุณภาพของฟาเซตนักเรียน ผู้วิจัยเลือกพิจารณาดัชนีระดับชั้น เนื่องจากพิจารณาว่าการให้คะแนนที่สูงเป็นคะแนนที่เกิดขึ้นจริงจากความสามารถที่สูงของนักเรียนมากกว่าเป็นเหตุบังเอิญ (Wright & Masters, 2002) ดังนั้น จึงพิจารณาได้ว่า กลุ่มตัวอย่างมีความสามารถประมาณ 2 ระดับ (เก่งและอ่อน) (ดัชนีระดับชั้น = 2.29) ส่วนดัชนีความเที่ยงจำแนก สามารถเปรียบเทียบได้กับความเที่ยงแบบสัมประสิทธิ์แอลฟาของครอนบาคในทฤษฎีการวัดแบบดั้งเดิม โดยความเที่ยงแบบสัมประสิทธิ์แอลฟานั้นจะมีค่าที่สูงกว่าดัชนีความเที่ยงจำแนก (Linacre, 2022a) โดย Fisher (2007, as cited in Mohaffyza et al., 2015) ได้นำเสนอค่าที่ยอมรับได้ของความเที่ยงในฟาเซตนักเรียนและข้อรายการประเมินไว้ คือ ต้องมีค่าตั้งแต่ 0.67 ขึ้นไป นอกจากนี้ดัชนีความเที่ยงจำแนกยังคำนวณจากดัชนีจำแนกที่วิเคราะห์ความแตกต่างของความสามารถในการแจกแจงแบบปกติซึ่งจะมีค่าน้อยกว่าดัชนีระดับชั้น ที่คำนวณจากความสามารถของกลุ่มตัวอย่างทั้งหมดที่ใช้ในงานวิจัย (Wright and Masters, 1982 as cited in Fisher, 1992)

ผลการวิเคราะห์ฟาเซตผู้ประเมิน แสดงให้เห็นว่า ผู้ประเมินมีการกดปล่อยคะแนนที่ต่างกัน โดยมีดัชนีความเที่ยงจำแนกเข้าใกล้ 1 ช่วยยืนยันความตรงเชิงโครงสร้างของรูบริก (Edwards et al., 2019) โมเดลราสช์นั้นอนุญาตให้ผู้ประเมินประเมินอย่างเป็นอิสระซึ่งกันและกัน แตกต่างจากแนวคิดของทฤษฎีการวัดแบบดั้งเดิมและทฤษฎีการสรุปผลอ้างอิงของ

การวัด (generalizability theory) ดังนั้นจึงควรพิจารณาค่า Fit MS เป็นหลัก โดยหากต้องการให้ประเมินโดยมีคะแนนเท่ากันควรใช้ทฤษฎีอื่นในการวิเคราะห์ (Linacre, 2022a)

ผลการวิเคราะห์ฟาเซตเครื่องดนตรี ค่าเมเชอร์แสดงให้เห็นว่าเครื่องดนตรีแต่ละชนิดได้รับอิทธิพลของการประเมินไม่แตกต่างกัน สอดคล้องกับการที่เครื่องดนตรีทั้ง 4 ชนิดเป็นประเภทเครื่องตีเหมือนกัน และมีแบบแผนการบรรเลงคล้ายกัน (สงบศึก ธรรมวิหาร, 2540) และสอดคล้องกับการศึกษาเกณฑ์การประเมิน สำนักทบวงมหาวิทยาลัย (2544) ที่ระบุเกณฑ์ของเครื่องดนตรีแต่ละชนิดเท่ากัน โดยมีรายละเอียดของการเตรียมพร้อมเครื่องดนตรี การจับไม้ และการบรรเลงเทคนิคพิเศษบางส่วนที่แตกต่างกันเท่านั้น ดังนั้น การพัฒนารูบริกโดยยึดเกณฑ์นี้ จึงน่าจะทำให้เครื่องดนตรีแต่ละชนิดได้รับอิทธิพลของการประเมินไม่แตกต่างกัน

ผลการวิเคราะห์ฟาเซตข้อรายการประเมิน แสดงให้เห็นว่า ข้อรายการประเมินที่ยากที่สุด คือ ข้อที่ 8 การบรรเลงเทคนิคพิเศษ สอดคล้องกับการบรรเลงเทคนิคพิเศษเป็นลักษณะเฉพาะของการบรรเลงเพลงเดี่ยวและเป็นส่วนที่ต้องใช้เวลาในการฝึกฝนมากที่สุด เป็นส่วนใช้ในการแสดงฝีมือ ความแม่นยำ และความเข้าใจในบทเพลง (รณฤทธิ์ ไหมทอง, 2021) ส่วนข้อรายการที่ง่ายที่สุด คือ ที่ 1 เครื่องดนตรี ซึ่งอยู่ในเกณฑ์ด้านการเตรียมพร้อมเครื่องดนตรี ซึ่งหากพิจารณาเพิ่มเติมจะพบว่า ข้อรายการประเมินทั้ง 3 ข้อที่อยู่ในเกณฑ์ด้านนี้มีความยากน้อยที่สุด 3 อันดับแรก ผู้วิจัยพิจารณาว่า เกณฑ์การประเมินเรื่องนี้ เป็นพื้นฐานของผู้บรรเลงดนตรีไทยทุกคน ที่จะต้องเตรียมพร้อมเครื่องดนตรีก่อนการเล่น และมักถูกระบุอยู่ในส่วนแรกของเกณฑ์หรือแบบฝึกหัดดนตรีต่าง ๆ (ชยพร ไชยสิทธิ์ และคณะ, 2017; สำนักทบวงมหาวิทยาลัย, 2544) ดังนั้นจึงเป็นข้อที่ง่ายที่สุด และเมื่อพิจารณาทั้ง 12 ข้อรายการจากดัชนีคุณภาพของฟาเซตข้อรายการประเมิน พบว่ามีความยากง่ายเหมาะสม โดยมีความยากง่ายอยู่ประมาณ 3 ระดับ (ยาก ปานกลาง และง่าย) และเมื่อผลการวิเคราะห์ค่า PT_{cor} ของฟาเซตข้อรายการประเมินไปในทิศทางเดียวกัน ผู้วิจัยพิจารณาว่า เนื่องจากข้อรายการประเมินแต่ละข้อวัดในทักษะเดียวกัน ดังนั้น ความยากง่ายของข้อรายการประเมินจึงไม่แตกต่างกันมากนัก ส่วนดัชนีความเที่ยงที่สูง แสดงให้เห็นว่า ตัวอย่างที่ใช้ในการวิเคราะห์ครั้งนี้มีมากเพียงพอที่จะยืนยันความตรงเชิงโครงสร้างของรูบริก (Edwards et al., 2019; Linacre, 2022a) และผลการปรับปรุงประสิทธิผลระดับคุณภาพ แสดงให้เห็นว่า ข้อรายการประเมินส่วนใหญ่มีความถี่สูงในการถูกประเมินช่วงระดับคุณภาพที่ 3 ขึ้นไป แสดงให้เห็นว่า ควรมีระดับคุณภาพไม่เกิน 3 ระดับ สอดคล้องกับระดับความยากจากการวิเคราะห์ดัชนีฟาเซตข้อรายการประเมิน

ข้อเสนอแนะ

ข้อเสนอแนะในการนำผลวิจัยไปใช้ คือ สำหรับผู้ที่สนใจหรือต้องการพัฒนาเครื่องมือประเภทรูบริก สามารถศึกษาแนวทางการวิเคราะห์ด้วยโมเดล MRFM-PCM ไปใช้สำหรับการพิจารณาปรับปรุงคุณภาพของรูบริก ทั้งเรื่องของจำนวนข้อรายการประเมินที่ควรมี จำนวนระดับคุณภาพรายข้อ และการแสดงหลักฐานคุณสมบัติทางจิตมิติที่เกี่ยวข้องได้ ซึ่งข้อดีของโมเดล MFRM-PCM คือ สามารถพิจารณาผลการวิเคราะห์ผ่านฟาเซตที่เกี่ยวข้องมากกว่า 2 ฟาเซตได้ ซึ่งในการประเมินทักษะปฏิบัติโดยทั่วไปแล้วมักมีอย่างน้อย 3 ฟาเซตที่เกี่ยวข้อง (นักเรียน ผู้ประเมิน และข้อรายการประเมิน) รวมถึงสามารถพิจารณาปรับปรุงประสิทธิผลระดับคุณภาพของข้อรายการประเมินให้สอดคล้องกับความสามารถที่แท้จริงและสถานการณ์การประเมินจริง

ข้อเสนอแนะในการวิจัยครั้งต่อไป คือ ควรนำผลการวิเคราะห์คุณสมบัติด้วยโมเดล MFRM-PCM ของรูบริกไปใช้ปรับปรุงระดับคุณภาพแต่ละข้อรายการประเมิน และวิพากษ์กับผู้ทรงคุณวุฒิหรือผู้ประเมินเพื่อพัฒนาเป็นรูบริกฉบับสมบูรณ์ และพัฒนาเป็นคู่มือการใช้ที่อาจประกอบไปด้วยภาระงาน (task) และรูบริก

รายการอ้างอิง

ภาษาไทย

กมลวรรณ ตังธนกานนท์. (2563). *การวัดและประเมินทักษะการปฏิบัติ* (พิมพ์ครั้งที่ 3). สำนักพิมพ์จุฬาลงกรณ์มหาวิทยาลัย.

ชยพร ไชยสิทธิ์, มนัส วัฒนไชยยศ, และ บรรจง ชลวิโรจน์. (2560). การพัฒนาชุดฝึกระนาดทุ้มเพลงสาธุการของนักเรียนชั้นมัธยมศึกษาปีที่ 4 สาขาวิชาชีพปี่พาทย์ วิทยาลัยนาฏศิลปนครศรีธรรมราช สถาบันบัณฑิตพัฒนศิลป์. *วารสารวิทยาลัยนครราชสีมา, 11*(3), 202-212.

รณฤทธิ์ ไหมทอง. (2564). การประพันธ์เดี่ยวระนาดเอกเพลงเทวาประสิทธิ์ สองชั้น เพื่อการประกวดเดี่ยวเครื่องดนตรีไทยระดับชาติ สำหรับนักเรียนระดับชั้นประถมศึกษา. *วารสารวิชาการ คณะมนุษยศาสตร์และสังคมศาสตร์มหาวิทยาลัยราชภัฏพระนคร, 5*(2), 109 – 130.

ศิริชัย กาญจนวาสี. (2563). *ทฤษฎีการทดสอบแนวใหม่* (พิมพ์ครั้งที่ 5). สำนักพิมพ์จุฬาลงกรณ์มหาวิทยาลัย.

สงบศึก ธรรมวิหาร. (2540). *ดุริยางค์ไทย.* สำนักพิมพ์จุฬาลงกรณ์มหาวิทยาลัย.

สำนักทบวงมหาวิทยาลัย. (2544). *เกณฑ์มาตรฐานดนตรีไทยและเกณฑ์การประเมิน.* ภาพพิมพ์.

ภาษาอังกฤษ

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Chen, W. H., Lenderking, W., Jin, Y., Wyrwich, K. W., Gelhorn, H., & Revicki, D. A. (2014). Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. *Quality of Life Research, 23*(2), 485-493. DOI: 10.1007/s11136-013-0487-5

DeLuca, C., & Bolden, B. (2014). Music performance assessment: Exploring three approaches for quality rubric construction. *Music Educators Journal, 101*(1), 70-76.

Eckes, T. (2009). *Introduction to many-facet Rasch measurement*. Peter Lang Edition.

Edwards, A. S., Edwards, K. E., & 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*, 1-25.

Engelhard, G. (1997). Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement*, *1*(1), 19-33.

Fisher, W. (1992). Reliability, separation, strata statistics. *Rasch Measurement Transections, 6*(3), 238.

Krishnan, S., & Idris, N. (2018). Using partial credit model to improve the quality of an instrument. *International Journal of Evaluation and Research in Education, 7(*4), 313-316.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(4), 328.

Linacre, J. M. (1997, August). *Judging plans and facets*. https://www.rasch.org/rn3.htm

Linacre, J. M. (2002). Understanding Rasch measurement: Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85-106.

Linacre, J. M. (2010). When to stop removing Items and persons in Rasch misfit analysis. *Rasch measurement transactions, 23*(4).

Linacre, J. M. (2022a). *A user guide to FACETS* (Version 3.84.0) [Computer software]. https://www.winsteps.com/a/Facets-Manual.pdf

Linacre, J. M. (2022b). *A user guide to WINSTEPS* (Version 5.2.3) [Computer software]. https://www.winsteps.com/a/Winsteps-Manual.pdf

Mohaffyza, M., Sulaiman, N., Lai, C. S., & Salleh, K. M. (2015). Measuring the validity and reliability of research instruments. *Procedia – Social and Behavioral Sciences, 204*, 164-171. https://doi.org/10.1016/j.sbspro.2015.08.129

Uto, M. (2021). Accuracy of performance-test linking based on a many-facet Rasch model. *Behavior Research Methods, 53*(4). 1440-1454.

Wright, B. D., & Masters, G. N. (2002). Number of person or item strata. *Rasch Measurement Transaction, 16*(3), 888.