



Academic Article

MANY-FACET RASCH MEASUREMENT MODEL: CONCEPTS AND NECESSARY
STATISTIC FOR PERFORMANCE ASSESSMENT

โมเดลหลายองค์ประกอบของราสช์: มโนทัศน์และค่าสถิติที่สำคัญสำหรับ
การประเมินภาคปฏิบัติ

Received: February 17, 2023

Revised: March 27, 2023

Accepted: April 5, 2023

Purin Thepsathit^{1*} and Kamonwan Tangdhanakanond²

ภูรินท์ เทพสถิตย์^{1*} และกมลวรรณ ตังธณานนท์²

^{1,2}Faculty of Education, Chulalongkorn University, Bangkok 10330, Thailand

^{1,2}คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

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

Abstract

Many-Facet Rasch Measurement Model is a measurement model that is widely used for the analysis of educational instruments, which can be found especially in a performance assessment situation related to more than two facets, including examinee, item, and rater. The psychometric properties of the instruments are analyzed across each facet of the model and are invariant. However, this model is quite complex because the interpretation of the model's analysis is different depending on each facet, and the interpretation criterion is different depending on the analyst's objectives. Therefore, this article will present the Many-Facet Rasch Measurement Model concepts, writing the command for the model's analysis, and the necessary statistic for the model's analysis. Readers could use these guidelines for analyzing and interpreting their performance assessment using Many-Facet Rasch Measurement Model's analysis properly.

Keywords: Many-Facet Rasch Measurement Model, Performance Assessment, Educational Data Analysis, Educational Statistics

บทคัดย่อ

โมเดลหลายองค์ประกอบของราสช์ เป็นโมเดลที่นิยมใช้ในการวิเคราะห์เครื่องมือการวัดและประเมินผลทางการศึกษา ซึ่งพบได้ทั่วไปโดยเฉพาะการประเมินภาคปฏิบัติที่เกี่ยวข้องกับฟาเซตของการประเมินมากกว่า 2 ฟาเซตขึ้นไป ประกอบด้วย ผู้สอบ ข้อรายการประเมิน และผู้ประเมิน คุณสมบัติทางจิตมิติของเครื่องมือวัดและประเมินจะถูกวิเคราะห์ในโมเดลผ่านทุกฟาเซตที่เกี่ยวข้อง โดยคุณสมบัติทางจิตมิติของเครื่องมือวัดและประเมินที่วิเคราะห์ด้วยโมเดลจะมีค่าที่คงที่ ไม่แปรเปลี่ยนไปตามกลุ่ม

อย่างไรก็ตาม เนื่องจากโมเดลหลายองค์ประกอบเป็นโมเดลที่มีความซับซ้อน เพราะเป็นโมเดลที่มีการแปลผลแตกต่างกันไปตามแต่ละฟาเซตที่ศึกษา รวมถึงมีเกณฑ์การวิเคราะห์ที่แตกต่างกันตามจุดประสงค์ของผู้วิเคราะห์ ดังนั้นบทความนี้ จึงจะนำเสนอโมทัศน์ของโมเดลหลายองค์ประกอบของราสส์ การเขียนคำสั่งเพื่อวิเคราะห์โมเดลหลายองค์ประกอบของราสส์ และค่าสถิติสำคัญในการวิเคราะห์โมเดลหลายองค์ประกอบ เพื่อนำไปใช้เป็นแนวทางในการวิเคราะห์และแปลผลการประเมินภาคปฏิบัติทางการศึกษาได้อย่างถูกต้องเหมาะสม

คำสำคัญ: โมเดลหลายองค์ประกอบของราสส์ การประเมินภาคปฏิบัติ การวิเคราะห์ข้อมูลทางการศึกษา สถิติทางการศึกษา

บทนำ (Introduction)

โมเดลหลายองค์ประกอบของราสส์ (Many-Facet Rasch Measurement Model: MFRM) เป็นหนึ่งในโมเดลตระกูลราสส์ที่พัฒนาโดย Linacre (1994) และเป็นโมเดลในกลุ่มทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) โดยโมเดลหลายองค์ประกอบของราสส์ แตกต่างจากโมเดลแบบดั้งเดิมหรือโมเดลในกลุ่มทฤษฎีการตอบสนองข้อสอบอื่นๆ กล่าวคือ สามารถใช้วิเคราะห์รูปแบบการวัดทางจิตวิทยาและการวัดทางการศึกษาที่มีฟาเซต (องค์ประกอบ) มากกว่า 2 ฟาเซตขึ้นไปได้ เช่น การประเมินภาคปฏิบัติที่มีฟาเซตเกี่ยวข้องมากกว่า 2 ฟาเซตขึ้นไป ประกอบด้วย ผู้สอบ ข้อสอบ และผู้ประเมิน (Bond et al., 2021) ซึ่งเป็นรูปแบบที่จะสามารถพบได้ในการประเมินภาคปฏิบัติโดยทั่วไป (Johnson et al., 2009) โดยอาจเพิ่มองค์ประกอบที่เกี่ยวข้องกับการวัด เช่น ชุดข้อสอบ ภาระงาน วันที่สอบ หรือลักษณะนิสัยของผู้สอบ เพื่อประมาณค่าพารามิเตอร์ร่วมกันได้ (Bond et al., 2021; Linacre, 1994; Linacre, 2022a)

โมเดลหลายองค์ประกอบของราสส์ พัฒนาต่อยอดมาจากโมเดลราสส์แบบดั้งเดิม ที่ศึกษาความสัมพันธ์ของตัวแปรแฝง (ตัวแปรอิสระ) กับตัวแปรตาม โดยตัวแปรแฝงนั้น ได้แก่ ความสามารถของผู้สอบ หรือความยากของข้อสอบ ส่วนตัวแปรตามคือโอกาสในการตอบข้อสอบถูกของผู้สอบ โดยมีแนวคิดคือ ผลคะแนนที่สังเกตได้ของผู้สอบ (ตัวแปรตาม) สามารถใช้อธิบายความสามารถของผู้สอบและความยากของข้อสอบ (ตัวแปรแฝง) ผ่านความน่าจะเป็นได้ ดังนั้นแนวคิดนี้ จึงเป็นที่มาของโมเดลราสส์สำหรับการตรวจให้คะแนนแบบ 2 ค่า (Dichotomous Rasch model) (Bond et al., 2021) ซึ่งอธิบายได้ดังสมการนี้ (Bond et al., 2021; Linacre, 2022a; Linacre, 2022b)

$$\ln\left(\frac{P_{ni}}{1 - P_{ni}}\right) = B_n - D_i$$

เมื่อ $\ln(P_{ni}/1 - P_{ni})$ คือ ความน่าจะเป็นการตอบข้อสอบข้อที่ i ถูกของผู้สอบคนที่ n ซึ่งสามารถอธิบายได้จาก $B_n - D_i$ โดยที่ B_n คือ ความสามารถของผู้สอบคนที่ n ส่วน D_i คือความยากของข้อสอบข้อที่ i

จากสมการข้างต้น จะสังเกตได้ว่า 1) โมเดลราสส์อธิบายความสามารถของผู้สอบและความยากของข้อสอบที่เป็นตัวแปรแฝงผ่านความน่าจะเป็นการตอบข้อสอบถูกของผู้สอบ และ 2) โมเดลราสส์ใช้เพียงความยากของข้อสอบในการอธิบายความน่าจะเป็นเท่านั้น โดยไม่ได้กล่าวถึงอำนาจจำแนกเหมือนโมเดลในทฤษฎีการตอบสนองข้อสอบอื่นๆ ดังนั้น โมเดลราสส์จึงมีความเฉพาะและถูกพัฒนามาจากแนวคิดที่แตกต่างจากโมเดลในทฤษฎีการตอบสนองแบบ 1, 2 และ 3 พารามิเตอร์

ต่อมา Andrich (1978) และ Master (1982) ได้พัฒนาโมเดลราสส์แบบมาตราประมาณค่า (Rasch Rating Scale Model) และโมเดลราสส์แบบการให้คะแนนบางส่วน (Rasch Partial Credit Model) ตามลำดับ เพื่อใช้สำหรับเครื่องมือที่ตรวจให้คะแนนที่มากกว่า 2 ค่า (Polytomous) โดยโมเดลราสส์แบบมาตราประมาณค่า จะประมาณค่าพารามิเตอร์ความยากของแต่ละระดับขั้นการตอบหรือระดับคะแนน (Thresholds) เท่ากันทั้งฉบับ ในขณะที่โมเดลราสส์แบบการให้คะแนนบางส่วนจะประมาณค่าพารามิเตอร์ โดยอนุญาตให้ความยากของแต่ละระดับขั้นการตอบ (Thresholds) ของแต่ละข้อรายการประเมินมีความแตกต่าง

กันได้ (Linacre, 2000) โดยโมเดลราสส์แบบมาตรประมาณค่าสามารถอธิบายได้ดังสมการนี้ (Bond et al., 2021; Linacre, 2022a)

$$\ln\left(\frac{P_{nik}}{1 - P_{nik}}\right) = B_n - D_i - F_k$$

เมื่อ $\ln(P_{nik}/1 - P_{nik})$ คือ ความน่าจะเป็นในการตอบบนระดับคะแนนที่ k ในข้อคำถามที่ i ของผู้ตอบคนที่ n ซึ่งสามารถอธิบายได้จาก $B_n - D_i - F_k$ โดยที่ B_n คือ ความสามารถของผู้ตอบคนที่ n D_i คือความยากของข้อคำถามข้อที่ i และ F_k คือความยากของระดับคะแนนที่ k

จากสมการดังกล่าว จะสังเกตได้ว่า ความยากของข้อคำถามและความยากของระดับคะแนนแต่ละระดับของมาตรประมาณค่านี้ ถูกประมาณค่าอย่างเป็นอิสระกัน โดยข้อคำถามแต่ละข้อจะมีความยากแตกต่างกันได้ แต่จะมีความยากของแต่ละระดับคะแนนที่เท่ากัน ซึ่งแตกต่างจากโมเดลราสส์แบบการให้คะแนนบางส่วน ซึ่งอธิบายได้ดังสมการต่อไปนี้ (Bond et al., 2021; Linacre, 2022a)

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

เมื่อ $\ln(P_{nik}/1 - P_{nik})$ คือ ความน่าจะเป็นในการตอบบนระดับคะแนนที่ k ในข้อคำถามที่ i ของผู้ตอบคนที่ n ซึ่งสามารถอธิบายได้จาก $B_n - D_{ik}$ โดยที่ B_n คือ ความสามารถของผู้ตอบคนที่ n และ D_{ik} คือความยากของระดับคะแนนที่ k ของข้อคำถามข้อที่ i

จากสมการดังกล่าว จะสังเกตได้ว่า ความยากของข้อคำถามและความยากของระดับคะแนนแต่ละระดับถูกประมาณค่าร่วมกัน โดยระดับคะแนนแปรผันไปตามแต่ละข้อคำถาม ดังนั้น แต่ละข้อคำถามในแบบวัดชุดนี้ ไม่จำเป็นต้องมีความยากของระดับคะแนนที่เหมือนกัน ดังนั้น จากคุณสมบัติของโมเดลที่ต่างกันนี้ โมเดลราสส์แบบมาตรประมาณค่า จึงนิยมนำไปใช้ในการวิเคราะห์เครื่องมือทางจิตวิทยาหรือมาตรประมาณค่าที่ต้องการให้เครื่องมือมีจำนวนของระดับคะแนนที่เท่ากันทั้งฉบับ ส่วนโมเดลราสส์แบบการให้คะแนนบางส่วน นิยมนำไปใช้กับการวิเคราะห์เครื่องมือที่เป็นข้อสอบ ทั้งข้อสอบที่สามารถให้คะแนนได้บางส่วน หรือรูปrikที่สามารถให้คะแนนได้หลายค่า (Andrich, 1978; Kanjanawasee, 2022; Master, 1982)

ดังนั้น โมเดลหลายองค์ประกอบของราสส์จึงถูกพัฒนาต่อยอดมาจากโมเดลราสส์แบบการตรวจให้คะแนนมากกว่า 2 ค่า ซึ่งเพิ่มพาดของการศึกษาผู้ประเมินเข้าไปในสมการ โดย Linacre (1994) ที่เป็นผู้พัฒนาโมเดล ได้เขียนสมการของโมเดลหลายองค์ประกอบไว้ดังนี้

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

เมื่อ $\ln(P_{nikj}/1 - P_{nikj})$ คือ ความน่าจะเป็นในการถูกประเมินโดยผู้ประเมินคนที่ j ด้วยระดับคะแนนที่ k ในข้อคำถามที่ i ของผู้สอบคนที่ n ซึ่งสามารถอธิบายได้จาก $B_n - D_i - F_k - C_j$ โดยที่ B_n คือ ความสามารถของผู้ตอบคนที่ n D_i คือความยากของข้อคำถามข้อที่ i F_k คือความยากของระดับคะแนนที่ k และ C_j คือ การกตคะแนนของผู้ประเมินคนที่ j

คุณสมบัติของโมเดลหลายองค์ประกอบของราสส์นี้ ช่วยให้สามารถวิเคราะห์เครื่องมือการวัดและประเมินผลที่มีพาดเกี่ยวข้องมากกว่า 2 พาดขึ้นไปได้ ซึ่งพบได้โดยทั่วไปในการประเมินภาคปฏิบัติ ที่มีผู้ประเมินเข้ามาเกี่ยวข้อง จากเดิมที่ผู้วิเคราะห์ จะวิเคราะห์ผู้ประเมินผ่านการวิเคราะห์ความเที่ยงด้วยวิธี inter-rater reliability หรือ intra-rater reliability ซึ่งอาจพิจารณาได้ว่าเป็นเพียงการวิเคราะห์ความเที่ยงแบบทฤษฎีดั้งเดิม (Classical Test Theory) ที่เป็นเพียงการบ่งชี้ถึง

ความสัมพันธ์ของการให้คะแนนของผู้ประเมินต่างคนกันหรือการให้คะแนนของผู้ประเมินคนเดียวกันต่างช่วงเวลากัน แต่ไม่สามารถระบุอิทธิพลการกดหรือปล่อยคะแนนของผู้ประเมินแต่ละคนได้ (Bond et al., 2021) รวมถึงไม่สามารถวิเคราะห์ความตรงเชิงโครงสร้างของรูปแบบการประเมินที่มีผลมาจากผู้ประเมินได้ แต่โมเดลหลายองค์ประกอบสามารถประมาณค่าโดยนำฟาเซตผู้ประเมินเข้ามาเป็นส่วนหนึ่งในการวิเคราะห์ได้ทั้งความตรงเชิงโครงสร้างและความเที่ยง ช่วยให้สามารถพัฒนารูปแบบการประเมินให้มีคุณภาพมากขึ้น และช่วยให้การวิเคราะห์เพื่อพัฒนาเครื่องมือทางการศึกษาสำหรับการประเมินทักษะปฏิบัติมีความตรงและความเที่ยงผ่านการวิเคราะห์ฟาเซตที่เกี่ยวข้องกับรูปแบบการประเมินได้ทั้งหมด รวมถึงสามารถวิเคราะห์ลักษณะการกดหรือปล่อยคะแนนของผู้ประเมินได้อีกด้วย (Eckes, 2009; Eckes, 2019; Myford & Wolfe, 2004)

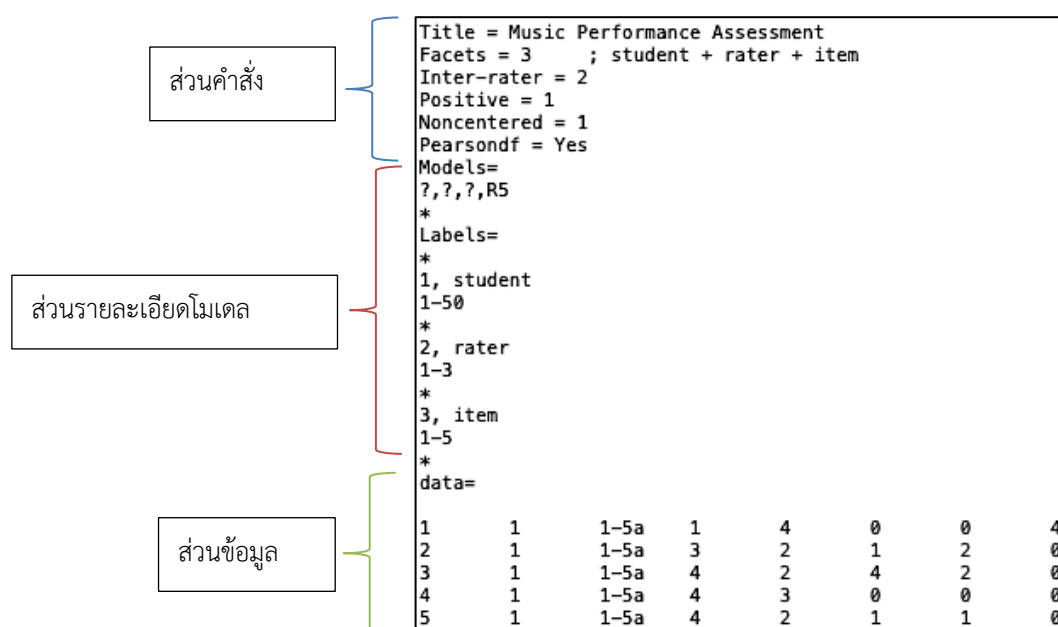
การเขียนคำสั่งโมเดลหลายองค์ประกอบของราสช์

ในบทความวิชาการฉบับนี้ ผู้เขียนขอเสนอตัวอย่างคำสั่งการวิเคราะห์และผลการวิเคราะห์จากโปรแกรมที่ใช้สำหรับการวิเคราะห์โมเดลหลายองค์ประกอบโดยเฉพาะ ได้แก่โปรแกรม Facets (Linacre, 2022a) โดยขอเสนอตัวอย่างคำสั่งในการวิเคราะห์โมเดลหลายองค์ประกอบดัง Figure 1 ซึ่งเป็นสถานการณ์การประเมินทักษะดนตรีไทยของนักเรียน 50 คน ที่ประเมินโดยผู้ประเมิน 3 คน ใช้เครื่องมือการประเมินแบบมาตรประมาณค่าซึ่งมีข้อรายการประเมินจำนวน 5 ข้อ โดยแต่ละข้อจะมีระดับคะแนนตั้งแต่ 0 ถึง 4 คะแนน

Figure 1

Rasch's multi-component model command for Facets program

คำสั่งโมเดลหลายองค์ประกอบของราสช์สำหรับโปรแกรม Facets



สำหรับคำสั่งในการวิเคราะห์โมเดลหลายองค์ประกอบของราสช์ สามารถแบ่งออกเป็น 3 ส่วน ประกอบด้วย 1) ส่วนคำสั่ง 2) ส่วนรายละเอียดโมเดล และ 3) ส่วนข้อมูล โดยในส่วนแรกบรรทัดที่ 1 (Title) ผู้วิเคราะห์สามารถระบุชื่อของรูปแบบการวัดได้ บรรทัดที่ 2 (Facets) ผู้วิเคราะห์ต้องระบุว่าการวิเคราะห์ครั้งนี้ มีฟาเซตที่เกี่ยวข้องจำนวนกี่ฟาเซต

จากตัวอย่างคำสั่งในภาพ 1 เป็นการวิเคราะห์ 3 ฟาเซต ได้แก่ ฟาเซตผู้สอบ ฟาเซตผู้ประเมิน และฟาเซตข้อรายการประเมิน บรรทัดที่ 3 (Inter-rater) หากต้องการให้โปรแกรมวิเคราะห์ค่าสถิติที่เกี่ยวข้องกับความสอดคล้องของผู้ประเมิน ให้ระบุว่าฟาเซตใดเป็นฟาเซตผู้ประเมิน บรรทัดที่ 4 (Positive) ผู้วิเคราะห์ต้องกำหนดให้เป็นฟาเซตผู้สอบ เพื่อให้การประมาณความสามารถของผู้สอบสูงต่ำตามความเป็นจริง บรรทัดที่ 5 (Noncentered) ผู้วิเคราะห์ต้องกำหนดให้เป็นฟาเซตผู้สอบ เนื่องจากต้องการให้ฟาเซตผู้สอบไม่ได้มีค่าเฉลี่ยของผลการวัดเท่ากับ 0 เหมือนฟาเซตอื่นๆ บรรทัดที่ 6 (Persondf) ควรระบุให้เป็น Yes เนื่องจากให้โมเดลวิเคราะห์ค่า Chi-square เพื่อตรวจสอบความสอดคล้องกลมกลืนของข้อมูลเชิงประจักษ์กับโมเดล บรรทัดที่ 7 (Models) เป็นการระบุว่าต้องการให้โปรแกรมวิเคราะห์ข้อมูลอย่างไร โดยเครื่องหมาย ? จะหมายถึงฟาเซต ตัวอักษร B หมายถึงให้โปรแกรมวิเคราะห์ความลำเอียง (Bias) ในที่นี้มักให้วิเคราะห์กับฟาเซตผู้ประเมิน ส่วน R5 เป็นการระบุว่า เครื่องมือที่ใช้วิเคราะห์มีระดับคะแนนทั้งสิ้น 5 ระดับ เมื่อเขียนคำสั่งในส่วนนี้แล้วให้ใส่เครื่องหมาย * ในบรรทัดถัดไปเพื่อไปยังส่วนรายละเอียดโมเดลต่อไป

ในส่วนรายละเอียดโมเดล เป็นการเพิ่มเติมข้อมูลให้โปรแกรมทราบว่า ข้อมูลของผู้วิเคราะห์ในแต่ละฟาเซตมีจำนวนกี่ตัวอย่าง โดยใช้คำสั่ง Labels จากนั้นใส่เครื่องหมาย * เพื่อระบุรายละเอียดที่ละฟาเซต จากตัวอย่างคำสั่งในภาพ 1 ระบุรายละเอียดโมเดลในส่วนฟาเซตที่ 1 คือ ฟาเซตผู้สอบ โดย 1-50 หมายถึง มีจำนวนของนักเรียนในการประเมินครั้งนี้ 50 คน ฟาเซตที่ 2 คือ ฟาเซตผู้ประเมิน โดย 1-3 หมายถึง มีจำนวนของผู้ประเมิน 3 คน และในส่วนฟาเซตที่ 3 คือ ฟาเซตข้อรายการประเมิน โดย 1-5 หมายถึง มีจำนวนข้อรายการประเมินทั้งหมด 5 ข้อ เมื่อเขียนคำสั่งในส่วนนี้แล้วให้ใส่เครื่องหมาย * ในบรรทัดถัดไปเพื่อไปยังส่วนข้อมูลต่อไป

ในส่วนข้อมูล ผู้วิเคราะห์ต้องระบุข้อมูลให้ตรงกับส่วนคำสั่งและส่วนรายละเอียดโมเดลที่ได้ระบุไว้ก่อนหน้านี้ โดยใส่ตำแหน่งให้ถูกต้องตามฟาเซต จากภาพ 1 คอลัมน์แรก คือ ฟาเซตผู้สอบ ซึ่งจะระบุว่าเป็นผู้สอบคนที่เท่าไร จากนั้นคอลัมน์ 2 จะเป็นการระบุว่ามีนักเรียนคนนั้นถูกประเมินโดยผู้ประเมินคนที่เท่าไร คอลัมน์ที่ 3 เป็นการระบุว่ามีผู้ประเมินคนไหนใช้ข้อรายการประเมินทั้งหมดกี่ข้อในการประเมินนักเรียนคนนี้ จากตัวอย่างภาพ 1 ระบุ 1-5a หมายถึง นักเรียนคนนี้ถูกประเมินโดยข้อรายการประเมินทั้ง 5 ข้อ และจากนั้นในคอลัมน์ที่เหลือให้ผู้วิเคราะห์ใส่ผลการประเมินแต่ละข้อรายการประเมินไล่เรียงกันไปในแต่ละคอลัมน์ จากตัวอย่างภาพ 1 ผู้วิเคราะห์แทนผลประเมินระดับต่ำที่สุดด้วย 0 คะแนน ไล่เรียงกันไปจนถึงระดับคะแนนที่สูงที่สุดซึ่งเป็น 4 คะแนน เนื่องจากเครื่องมือที่ใช้วิเคราะห์มีระดับคะแนนทั้งสิ้น 5 ระดับ

ค่าสถิติที่สำคัญในการวิเคราะห์โมเดลหลายองค์ประกอบของราสช์

โมเดลหลายองค์ประกอบของราสช์ เป็นโมเดลที่มีความซับซ้อน มีค่าสถิติที่ใช้ในการวิเคราะห์หลายค่า ซึ่งต้องอาศัยการวิเคราะห์ที่แตกต่างกันไปในแต่ละฟาเซต รวมถึงการแปลผลที่อาศัยการศึกษาทฤษฎีและเอกสารที่เกี่ยวข้องของผู้วิเคราะห์ เกณฑ์การแปลผลค่าสถิติบางค่ามีเกณฑ์ที่ไม่ตายตัว ขึ้นอยู่กับฟาเซตที่ต้องการศึกษาและการนำไปใช้ของผู้วิเคราะห์ บทความนี้นำเสนอค่าสถิติที่สำคัญสำหรับการวิเคราะห์ ประกอบด้วย 1) Measure 2) Fit Mean square 3) Fixed Chi-square 4) Random Chi-square 5) Separation Index 6) Strata Index 7) Reliability of Separation Index 8) Point-Measure Correlation และ 9) Rasch Rater Agreement

1. Measure คือค่าสถิติที่บอกถึงความเข้มของผลการวัด โดยจะแปลผลแตกต่างกันไปตามฟาเซตที่ศึกษา มีค่าอยู่ในช่วง $-\infty$ ถึง ∞ โดย Measure ที่มีค่าสูง หากเป็นฟาเซตผู้สอบ หมายถึงผู้สอบเป็นคนที่มีความสามารถสูง หากเป็นฟาเซตข้อรายการประเมิน (Item) หมายถึง ข้อรายการประเมินยาก และหากเป็นฟาเซตผู้ประเมิน หมายถึง ผู้ประเมินมีลักษณะการประเมินแบบกดคะแนน Measure คำนวณจากการแปลงคะแนนเชิงประจักษ์ที่ได้จากการวัด มาประมาณค่าพารามิเตอร์ต่างๆ ในโมเดล ผ่านความน่าจะเป็นในการตอบหรือถูกประเมินและความสอดคล้องกลมกลืนของโมเดลกับข้อมูลเชิงประจักษ์ โดยค่า Measure

จะแสดงในหน่วยลอจิท (Logits) ดังนั้น เมื่อโมเดลถูกตรวจสอบแล้วว่าเชื่อมต่อกันอย่างสมบูรณ์ จะทำให้สามารถเปรียบเทียบค่า Measure ระหว่างฟาเซตและสามารถคำนวณความน่าจะเป็นของการตอบหรือการถูกประเมินโดยใช้ค่า Measure จากแต่ละฟาเซตได้ (Linacre, 2022a)

Measure จะถูกประมาณด้วยวิธีการทำซ้ำ (iteration) ผ่านฟังก์ชันการประมาณค่าด้วยวิธี Joint maximum likelihood estimates ซึ่งเป็นวิธีการประมาณค่าพารามิเตอร์ให้มีความใกล้เคียงกับประชากรมากที่สุด (Myung, 2003) ดังนั้นค่า Measure จึงถูกนำเสนอคู่กับความคลาดเคลื่อนมาตรฐาน (Standard Error: SE) ซึ่งสะท้อนความแม่นยำของผลการวัดเมื่อนำไปใช้กับกลุ่มตัวอย่างอื่น (Linacre, 2022a)

สำหรับเกณฑ์ของการแปลค่า Measure มีผู้ที่ระบุเกณฑ์เอาไว้ใน 2 ฟาเซต ได้แก่ ฟาเซตข้อรายการประเมิน และฟาเซตผู้ประเมิน โดยฟาเซตข้อรายการประเมินนั้น หากผู้วิจัยต้องการให้ข้อรายการประเมินแต่ละข้อมีความยากง่ายในระดับปานกลาง ค่า Measure ของข้อรายการประเมินแต่ละข้อ ควรอยู่ในช่วง -0.99 ถึง 0.99 (Krishnan & Idris, 2018) ส่วนฟาเซตผู้ประเมิน หากต้องการให้ผู้ประเมินประเมินโดยมีการกดหรือปล่อยคะแนนไม่ต่างกัน ผู้ประเมินแต่ละคนควรมีค่า Measure อยู่ในช่วง -1.0 ถึง 1.0 (Myford & Wolfe, 2004) ส่วนฟาเซตผู้สอบ ค่า Measure สามารถมีค่าได้ตั้งแต่ศูนย์ถึงอนันต์ และไม่ได้มีการระบุถึงเกณฑ์ที่เหมาะสมของค่า Measure ในฟาเซตผู้สอบ อาจเป็นเพราะฟาเซตผู้สอบ มักตรวจสอบในส่วนของความสอดคล้องกลมกลืนกับโมเดลมากกว่า หากผู้สอบมีผลการวัดที่สูงมากหรือต่ำมากกว่าปกติ แต่มีค่าสถิติ Fit Mean Square ที่อยู่ในเกณฑ์เหมาะสม ก็ถือว่าคะแนนของผู้สอบที่มีความสามารถตามที่กล่าวมานั้น ยังมีความสอดคล้องกลมกลืนกับโมเดลอยู่ (Bond et al., 2021)

2. Fit Mean square (Fit MS) คือ ค่าสถิติที่แสดงถึงความสอดคล้องกลมกลืนระหว่างข้อมูลกับโมเดลราสซ์ และแสดงถึงความตรงเชิงโครงสร้าง (Construct Validity) แต่เนื่องจากโมเดลราสซ์เป็นโมเดลที่มีลักษณะเป็นโมเดลเชิงยืนยัน (Confirmatory Model) ข้อมูลจึงต้องสอดคล้องกลมกลืนกับโมเดล ดังนั้น ค่าสถิติ Fit MS จึงสามารถพิจารณาได้ว่าเป็นการควบคุมคุณภาพของผลการวัด (Bond et al., 2021; Linacre, 2022a) อย่างไรก็ตาม หากมีข้อมูลที่ไม่สอดคล้องกลมกลืนกับโมเดล ผู้วิเคราะห์ไม่ควรตัดทิ้งในทันที แต่ควรที่จะตรวจสอบถึงสาเหตุของความไม่สอดคล้องที่เกิดขึ้น เพื่อที่จะทำการตัดสินใจยอมรับ ปฏิเสธ หรือปรับปรุงข้อมูล (Smith, 1996) หากมีความไม่สอดคล้องเกิดขึ้น ควรตรวจสอบเบื้องต้นว่ามีความคลาดเคลื่อนเกิดขึ้นในขั้นตอนการกรอกข้อมูลสำหรับการวิเคราะห์หรือมีความผิดพลาดที่เกิดขึ้นจากผู้สอบหรือไม่ หากพิจารณาแล้วไม่พบความผิดปกติ อาจต้องพิจารณาว่าความไม่สอดคล้องกลมกลืนเกิดขึ้นจากความกำกวมของภาษาในข้อรายการประเมินหรืออาจเกิดจากความไม่เป็นเอกมิตีของข้อรายการประเมินหรือไม่ ผู้วิเคราะห์ควรพิจารณาว่าหากปฏิเสธความไม่สอดคล้องของข้อรายการประเมินข้อใดข้อหนึ่งโดยการนำข้อนั้นออกจากมาตรประมาณค่าแล้ว จะส่งผลกระทบต่อความเป็นตัวแทน (Representation) ของทักษะที่ต้องการประเมินหรือไม่ หากไม่มีข้อรายการประเมินที่สามารถทดแทนความเป็นตัวแทนนั้นได้ ผู้วิเคราะห์ไม่ควรนำข้อรายการประเมินนั้นออกจากมาตรประมาณค่าดังกล่าว (Bohlig et al., 1998; Smith, 1996)

ค่าสถิติ Fit MS ถูกคำนวณจากการหารค่า Chi-square และ Degree of Freedom ที่ได้จากการตรวจสอบความสอดคล้องกลมกลืน โดยค่า Fit MS จะมีค่าอยู่ในช่วง 0 ถึง ∞ โดยสำหรับค่าที่โมเดลราสซ์คาดหวัง จะมีค่าเท่ากับ 1.0 หากค่า Fit MS ที่มีค่าต่ำกว่า 1.0 สามารถแปลผลได้ว่า ข้อมูลเชิงประจักษ์ถูกทำนายได้ง่ายเกินไป (Too Predictable) หรือสอดคล้องกลมกลืนมากเกินไป (Overfit) ส่วนค่า Fit MS ที่มีค่ามากกว่า 1.0 สามารถแปลผลได้ว่า มีข้อมูลที่ไม่สามารถทำนายได้ (Unpredictable) โดยหากดูจากผลการวิเคราะห์ที่มาจากโปรแกรม Facets ค่านี้จะแสดงผลว่าเป็นค่า Outfit Mean Square ส่วนค่า Infit MS คือค่าความสอดคล้องแบบถ่วงน้ำหนัก (Information-weighted Fit) โดยค่า Outfit MS จะไวต่อผลการตอบหรือการประเมินที่เป็นค่าสุดโต่ง (Outlier-sensitive) ส่วน Infit MS จะไวต่อผลการตอบหรือการประเมินที่เป็นแบบแผน (Inlier-sensitive) (Linacre, 2002b; Wright & Linacre, 1994) โดยค่า Fit MS จะถูกนำเสนอไปกับค่า Z-Standardized (Zstd)

ที่เป็นผลการทดสอบ t-test ที่ถูกปรับให้เป็นค่ามาตรฐานสำหรับการพิจารณานัยสำคัญทางสถิติของค่า Fit MS โดยถ้า Zstd มีค่าตั้งแต่ 1.96 ขึ้นไปหรือตั้งแต่ -1.96 ลงมา จะแสดงถึงการมีนัยสำคัญทางสถิติที่ระดับ .05 หากค่า Outfit หรือ Infit MS ไม่สอดคล้องกลมกลืนกับโมเดล และมีค่า Z-Standardized ที่แสดงถึงนัยสำคัญทางสถิติ จะแปลผลได้ว่า ผลของความไม่สอดคล้องกลมกลืนนี้ เกิดขึ้นอย่างเป็นระบบ หรือไม่ได้เป็นเรื่องบังเอิญ (Linacre, 2022a; Linacre, 2022b)

สำหรับการตรวจสอบความสอดคล้องกลมกลืนของข้อมูลกับโมเดลนั้น อาจพิจารณาได้ว่า ค่า Fit MS เป็นค่าที่ใช้ในการตรวจสอบว่าข้อมูลสอดคล้องกับโมเดลซึ่งสามารถนำไปใช้ประโยชน์ได้หรือไม่ (Usefully) ส่วนค่า Z-Standardized เป็นค่าที่ตรวจสอบว่าข้อมูลสอดคล้องกับโมเดลในเชิงสถิติหรือไม่ (Statistically) (Linacre, 2003) ค่า Fit MS ที่มากกว่า 1.5 จะถูกพิจารณาว่าทำให้ผลการวัดบิดเบือนไป ส่วนค่า Zstd ที่มากกว่า 2.0 จะถูกพิจารณาว่ามีความผิดปกติที่ใหญ่มากพอที่จะต้องตรวจสอบ (Linacre, 2022a) ซึ่งแนวทางของการตรวจสอบความสอดคล้องกลมกลืนนั้น Linacre (2022b) ให้เริ่มจาก 1) ตรวจสอบค่า Outfit MS ก่อน Infit MS 2) ตรวจสอบ Fit MS ก่อนค่า Z-Standardized และ 3) ตรวจสอบค่า Z-Standardized ที่มีค่าสูงก่อนค่าที่ต่ำหรือค่าที่ติดลบ

อย่างไรก็ตาม เนื่องจากโมเดลหลายองค์ประกอบของราสซ์เป็นโมเดลที่มีสมมติฐานคือโมเดลสามารถทำนายข้อมูลได้อย่างสมบูรณ์ แต่ในความเป็นจริง ผู้วิเคราะห์ต้องใช้ข้อมูลเชิงประจักษ์ในการประมาณค่าพารามิเตอร์ จึงไม่สามารถทำให้โมเดลเชิงประจักษ์สอดคล้องกับโมเดลของราสซ์ได้อย่างสมบูรณ์ ค่า Fit MS ที่แสดงถึงความสอดคล้องนั้น เป็นไปได้ที่น้อยมากที่จะมีค่าเป็น 1.0 เท่ากันทุกฟาเซต ดังนั้น จึงต้องมีการกำหนดช่วงที่ยอมรับได้ของค่า Fit MS โดยช่วงที่ยอมรับได้ทั่วไปที่ส่งผลต่อการวัดจะอยู่ในช่วง 0.5 ถึง 1.5 (Linacre, 2022a; Linacre, 2022b) อย่างไรก็ตาม ไม่ได้มีการกำหนดเกณฑ์ไว้อย่างตายตัว โดย Wright and Linacre (1994) ได้เสนอแนวทางการปรับช่วงของค่า Fit MS ตามการนำไปใช้ โดยกำหนดให้ช่วงมีความกว้างหรือแคบตามวัตถุประสงค์ของการวัด เช่น กำหนดในช่วง 0.8 ถึง 1.2 สำหรับการนำไปใช้สำหรับข้อสอบแบบเลือกตอบ ซึ่งจะเห็นได้ว่ามีการกำหนดให้ช่วงแคบลงเพื่อป้องกันการทำนายไม่ได้หรือทำนายได้ง่ายเกินไป หรือกำหนดในช่วง 0.4 ถึง 1.2 สำหรับการประเมินตัดสิน ซึ่งจะเห็นได้ว่ามีการขยายช่วงจาก 0.5 ในแนวทางปกติ ไปสู่ 0.4 เพื่อยอมรับผลการประเมินที่มีความสอดคล้องกันมากขึ้นของผู้ประเมินได้

3. Fixed Chi-square คือ ค่าสถิติการทดสอบ Chi-square สำหรับการตรวจสอบความแตกต่างของข้อมูลภายในฟาเซต Fixed Chi-square มีสมมติฐานหลัก คือ ข้อมูลภายในฟาเซตไม่แตกต่างกัน ผลการวิเคราะห์จะแสดงแตกต่างกันไปในแต่ละฟาเซต ดังนั้น หากค่า Fixed Chi-square มีนัยสำคัญทางสถิติ จะแปลผลได้ว่า ข้อมูลภายในฟาเซตนั้นมีความแตกต่างกัน ซึ่งเกณฑ์ของการวิเคราะห์จะขึ้นอยู่กับจุดประสงค์ของผู้วิจัย หากต้องการให้เครื่องมือมีคุณสมบัติทางจิตมิติที่คงที่เมื่อนำไปใช้กับตัวอย่างวิจัยอื่นๆ ค่า Fixed Chi-square ควรมีนัยสำคัญทางสถิติ แต่หากต้องการให้ผู้ประเมินประเมินโดยมีลักษณะการกดหรือปล่อยไม่ต่างกัน ไม่ควรให้มีผลการทดสอบที่มีนัยสำคัญทางสถิติ (Linacre, 2022a) อย่างไรก็ตาม Myford and Wolfe (2004) ได้ทำการทดลอง พบว่า ค่า Fixed Chi-square ในฟาเซตผู้ประเมินมีความอ่อนไหวมาก ดังนั้น หากผู้ประเมินประเมินแบบกดหรือปล่อยคะแนนไม่ต่างกันมาก ก็อาจมีนัยสำคัญทางสถิติเกิดขึ้นได้

4. Random Chi-square คือ ค่าสถิติการทดสอบ Chi-square สำหรับการตรวจสอบผลการวัดของตัวอย่างกับประชากร Random Chi-square มีสมมติฐานหลัก คือ กลุ่มตัวอย่างมีความสามารถใกล้เคียงกับการสุ่มจากประชากรที่มีการแจกแจงแบบปกติ (Normal Distribution) ดังนั้น หากค่า Random Chi-square มีนัยสำคัญทางสถิติในฟาเซตผู้สอบ จะหมายถึงผู้สอบที่เป็นตัวอย่างวิจัยในครั้งนี้ มีความสามารถแตกต่างจากกลุ่มประชากรที่มีการแจกแจงแบบปกติ เกณฑ์การวิเคราะห์ขึ้นอยู่กับผู้วิเคราะห์ อย่างไรก็ตาม หากต้องการให้ผลการวัดนี้เกิดขึ้นอย่างคงที่เมื่อนำไปใช้กับกลุ่มตัวอย่างอื่น ควรให้ผลการทดสอบไม่มีนัยสำคัญทางสถิติในฟาเซตผู้สอบและผู้ประเมิน (Linacre, 2022a)

5. Separation Index คือ ดัชนีที่แสดงถึงจำนวนระดับของความแตกต่างของข้อมูลภายในฟาเซต โดยประมาณค่าตามสมมติฐานว่าข้อมูลนั้นเป็นการแจกแจงแบบปกติ ดังนั้น ข้อมูลภายในฟาเซตที่มีผลการวัดสูงมากหรือต่ำมากจะไม่มีผลต่อการวิเคราะห์ดัชนี Separation เนื่องจากจะถูกพิจารณาว่าเกิดจากผลการวัดที่สุตโต่งหรือบังเอิญ (Wright & Masters, 2002) ดัชนี Separation คำนวณจากสัดส่วนของส่วนเบี่ยงเบนมาตรฐานที่แท้จริง ($True SD$) กับความคลาดเคลื่อนที่เกิดจากการวัด (Root Mean Square Error: RMSE) โดยมีค่าได้ตั้งแต่ 0 ขึ้นไป โดยเกณฑ์การแปลผลดัชนี Separation จะมีค่าแตกต่างกันตามแต่ละฟาเซต สำหรับฟาเซตผู้สอบ ดัชนี Separation ที่มีค่าต่ำกว่า 2.0 จะถูกพิจารณาว่ามีความแตกต่างต่ำ ฟาเซตข้อรายการประเมิน ดัชนี Separation ที่มีค่าต่ำกว่า 3.0 จะถูกพิจารณาว่ามีความแตกต่างต่ำ ส่วนฟาเซตผู้ประเมิน ขึ้นอยู่กับวัตถุประสงค์ของผู้วิเคราะห์ หากต้องการให้ผู้ประเมินประเมินโดยมีลักษณะการกดหรือปล่อยคะแนนไม่ต่างกัน ก็ควรให้ดัชนี Separation มีค่าต่ำ (Linacre, 2022b)

6. Strata Index คือ ดัชนีที่แสดงถึงจำนวนระดับของความแตกต่างของข้อมูลภายในฟาเซต แต่แตกต่างจากดัชนี Separation ในเรื่องของการที่ไม่ได้อนุมานว่าข้อมูลมีการแจกแจงแบบปกติ ดังนั้นดัชนี Strata จะคำนวณโดยข้อมูลจากตัวอย่างวิจัยทุกข้อมูล โดยอนุมานว่าข้อมูลที่มีความสามารถสูงมากหรือต่ำมากนั้นเป็นผลจากความสามารถของผู้สอบที่เกิดขึ้นจริง ดังนั้นหากผู้วิเคราะห์พิจารณาแล้วว่า ความสามารถนี้เป็นความสามารถที่เกิดขึ้นจริง ก็สามารถใช้ดัชนีนี้ในการอธิบายความแตกต่างของข้อมูลภายในฟาเซตแทนดัชนี Separation ได้ โดยดัชนี Strata จะมีค่าที่สูงกว่าดัชนี Separation เสมอ (Linacre, 2022a; Linacre, 2022b) อย่างไรก็ตาม หากตัวอย่างวิจัยมีขนาดใหญ่และมีการแจกแจงแบบปกติ การใช้ดัชนี Separation จะเหมาะสมกว่าดัชนี Strata (Bond et al., 2021)

7. Reliability of Separation Index คือ ดัชนีที่แสดงถึงความเที่ยงในแนวคิดแบบทฤษฎีการตอบสนองข้อสอบ ที่แสดงถึงความสามารถในการนำไปใช้ได้คงที่ ของคุณสมบัติทางจิตมิติ (Reproducibility/Replicability) (American Educational Research Association et al., 2014; Bond et al., 2021; Linacre, 2022b) ดัชนีความเที่ยงจำแนกที่สูงเป็นการยืนยันโครงสร้างผลการวัดของโมเดล

ดัชนี Reliability of Separation คำนวณจากสัดส่วนของความแปรปรวนที่แท้จริงกับความแปรปรวนที่สังเกตได้ โดยดัชนีนี้จะมีค่าอยู่ในช่วง 0 ถึง 1 คล้ายกับความเที่ยงในแบบทฤษฎีดั้งเดิม แต่จะมีค่าต่ำกว่าเล็กน้อยเนื่องจากคำนวณมาจากดัชนี Separation ที่ไม่ได้นำเอาข้อมูลแบบสุตโต่งมาคำนวณ (Linacre, 2022a) เกณฑ์การแปลผลดัชนี Reliability of Separation สำหรับฟาเซตผู้สอบ ค่าที่ต่ำกว่า 0.8 ถือว่ามีความเที่ยงจำแนกต่ำ ฟาเซตข้อรายการประเมิน ค่าที่ต่ำกว่า 0.9 ถือว่ามีความเที่ยงจำแนกต่ำ ส่วนฟาเซตผู้ประเมิน หากต้องการให้ผู้ประเมินประเมินโดยมีลักษณะการกดหรือปล่อยไม่ต่างกัน ควรมีค่าดัชนี Reliability of Separation ที่เข้าใกล้ 0 (Linacre, 2022b)

8. Point-Measure Correlation (PT_{cor}) คือ ค่าสถิติสหสัมพันธ์ (Correlation) ในรูปแบบของโมเดลราสช์แบบ Pearson Point Biserial Correlation ของข้อมูลภายในฟาเซต แสดงถึงความสอดคล้องไปในทิศทางเดียวกันของข้อมูลกับความเป็นเอกมิติ (Unidimensionality) ของโมเดลราสช์ ผู้วิเคราะห์สามารถใช้ค่านี้ในการวิเคราะห์ว่าข้อรายการประเมินมุ่งวัดในตัวแปรเดียวกันหรือไม่ และสามารถใช้ในการตรวจสอบความสอดคล้องของผลการประเมินกับตัวแปรที่ต้องการวัดของผู้ประเมินได้ ในกรณีที่เครื่องมือที่ให้คะแนนมากกว่า 2 ค่า PT_{cor} ช่วยในการวินิจฉัยลักษณะรูปแบบการตอบหรือการประเมินในแต่ละข้อรายการประเมินได้ ค่า PT_{cor} โดยทั่วไปนั้น ควรมีค่าเป็นบวกขึ้นไป เพื่อยืนยันว่าข้อรายการประเมินมุ่งวัดในตัวแปรเดียวกัน และช่วยยืนยันความตรงเชิงเนื้อหาและโครงสร้างของเครื่องมือ (Krishnan & Idris, 2018; Linacre, 2022a)

9. Rasch Rater Agreement คือ ค่าสถิติที่แสดงถึงความสอดคล้องของผลการประเมินระหว่างผู้ประเมินบนพื้นฐานแนวคิดแบบโมเดลราสช์ โดย Rasch Rater Agreement จะนำเสนอค่าสถิติที่เกี่ยวข้องทั้งหมด 2 ค่า ประกอบด้วยค่าความสอดคล้องที่สังเกตได้ (Exact Agreement Observed) และค่าความสอดคล้องที่คาดหวัง (Exact Agreement

Expected) โดยจะแสดงผลเป็นร้อยละ ซึ่งค่าความสอดคล้องที่คาดหวังจะถูกคำนวณจากความน่าจะเป็นของผลการประเมินจากฟาเซตที่เกี่ยวข้องทุกฟาเซต การแปลผลการวิเคราะห์ค่า Rasch Rater Agreement นั้นขึ้นอยู่กับผู้วิเคราะห์ว่าต้องการให้ผลการประเมินของผู้ประเมินเป็นอย่างไร หากต้องการผู้ประเมินประเมินสอดคล้องกันทั้งหมด ควรให้ค่าความสอดคล้องที่สังเกตได้มากกว่าค่าความสอดคล้องที่คาดหวังอย่างมากหรือมากกว่าร้อยละ 90 ขึ้นไป แต่หากให้ผู้ประเมินประเมินได้อย่างเป็นอิสระซึ่งกันและกันแต่ยังมุ่งประเมินในสิ่งเดียวกัน ควรให้ค่าความสอดคล้องที่สังเกตได้ใกล้เคียงหรือมากกว่าค่าความสอดคล้องที่คาดหวังเล็กน้อย หากค่าความสอดคล้องที่สังเกตน้อยกว่าค่าความสอดคล้องที่คาดหวัง แสดงถึงความไม่สอดคล้องของผู้ประเมินซึ่งเกิดขึ้นได้ทั่วไปในกรณีที่ผู้ประเมินไม่ได้ผ่านการอบรม (Linacre, 2022a)

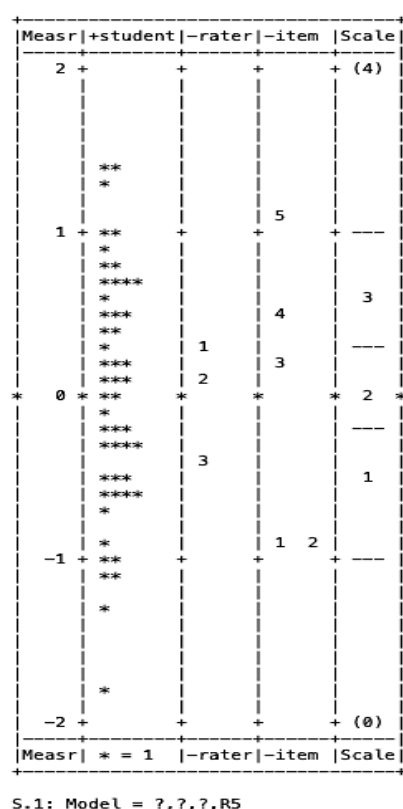
ตัวอย่างการแปลผลการวิเคราะห์

ตัวอย่างการแปลผลในสถานการณ์ตัวอย่างการประเมินทักษะดนตรีไทย ผู้เขียนจะเน้นการแปลผลในแต่ละฟาเซต ซึ่งครอบคลุมการวิเคราะห์คุณสมบัติทางจิตมิติด้วยโมเดลหลายองค์ประกอบ ตัวอย่างผลการวิเคราะห์ แสดงดังต่อไปนี้

Figure 2

Example of Wright Map analysis results

ตัวอย่างผลการวิเคราะห์ Wright Map



จาก Figure 2 แสดงผลการวัดของทั้ง 3 ฟาเซตในภาพรวมที่ปรับให้ในมาตรฐานเดียวกัน โดยคอลัมน์ 1 แสดงค่า Measure เปรียบเทียบกับไม้บรรทัดเพื่อใช้เทียบแต่ละฟาเซต คอลัมน์ 2 แสดงค่า Measure ของผู้สอบ โดยสัญลักษณ์ * เท่ากับผู้สอบ 1 คน โดยผู้สอบที่มีความสามารถสูงสุด 2 คน มีค่า Measure ประมาณ 1.4 ส่วนผู้สอบที่มีความสามารถต่ำสุด มีค่า Measure

ประมาณ -1.8 คอลัมน์ 3 แสดงค่า Measure ของผู้ประเมิน โดยผู้ประเมินที่มีการกดคะแนนมากที่สุดคือผู้ประเมินคนที่ 1 ส่วนผู้ประเมินที่ปล่อยคะแนนมากที่สุดคือผู้ประเมินคนที่ 3 คอลัมน์ที่ 4 แสดงค่า Measure ของข้อรายการประเมิน โดยข้อรายการประเมินที่ยากที่สุดคือข้อที่ 5 ส่วนข้อรายการประเมินที่ง่ายที่สุดคือข้อที่ 1 และ 2 ส่วนคอลัมน์ 5 แสดงค่า Measure ของมาตรในเครื่องมือ โดยแสดงค่า Measure ของแต่ละระดับคะแนนในเครื่องมือ

Figure 3

Sample test taker FASET analysis results

ตัวอย่างผลการวิเคราะห์ฟาสเซตผู้สอบ

Music Performance Assessment 2/5/2023 8:13:20 PM
Table 7.1.2 student Measurement Report (arranged by FN).

Total Score	Total Count	Obsvd Average	Fair(M) Average	+ Measure	Model S.E.	Infit MnSq	ZStd	Outfit MnSq	ZStd	Estim. Discrm	Correlation PtMea	PtExp	Nu student
19	15	1.27	.87	-.64	.25	1.91	2.0	6.71	4.9	.18	.35	.70	14 14
20	15	1.33	.95	-.57	.25	2.10	2.3	6.47	4.9	.02	.29	.71	1 1
21	15	1.40	1.03	-.51	.24	1.14	.4	3.28	2.9	.59	.56	.72	12 12
25	15	1.67	1.41	-.28	.24	2.41	2.7	2.70	2.5	.71	.49	.75	33 33
41	15	2.73	3.12	.66	.25	1.87	1.9	2.54	2.0	.03	.30	.71	38 38
12	15	.80	.47	-1.10	.27	2.18	2.2	2.51	1.8	.47	.30	.60	42 42
40	15	2.67	3.04	.60	.24	1.62	1.5	2.42	1.9	.40	.45	.72	30 30
36	15	2.40	2.66	.36	.24	1.60	1.4	2.29	1.9	.65	.54	.75	23 23
44	15	2.93	3.34	.84	.25	1.75	1.7	2.10	1.5	.91	.50	.67	6 6
12	15	.80	.47	-1.10	.27	2.09	2.1	1.22	.5	1.27	.53	.60	43 43
37	14	2.64	2.91	.51	.26	1.07	.3	1.83	1.3	.82	.63	.74	9 9
51	15	3.40	3.69	1.36	.30	1.18	.5	1.78	1.0	.90	.42	.55	3 3
46	15	3.07	3.46	.97	.26	1.40	1.0	1.77	1.1	.60	.46	.64	7 7
25	15	1.67	1.41	-.28	.24	1.48	1.1	1.52	1.0	1.15	.68	.75	13 13
36	15	2.40	2.66	.36	.24	1.45	1.1	1.50	.9	.70	.60	.75	21 21
41	15	2.73	3.12	.66	.25	1.26	.7	1.45	.8	.87	.62	.71	8 8
24	15	1.60	1.30	-.34	.24	.94	.0	.97	.1	1.06	.75	.75	18 18
26	15	1.73	1.51	-.22	.24	1.13	.4	.88	.0	.98	.74	.76	25 25
43	15	2.87	3.27	.78	.25	.78	-.5	1.11	.3	.42	.57	.68	19 19
19	15	1.27	.87	-.64	.25	.76	-.5	.99	.1	.74	.70	.70	27 27
9	15	.60	.34	-1.35	.30	.80	-.2	.75	-.1	.93	.54	.53	44 44
42	15	2.80	3.20	.72	.25	.75	-.5	.89	.0	1.12	.72	.70	28 28
26	15	1.73	1.51	-.22	.24	.75	-.5	.83	-.1	.47	.69	.76	32 32
32	15	2.13	2.22	.13	.24	.80	-.4	.75	-.3	1.12	.79	.76	36 36
41	15	2.73	3.12	.66	.25	.73	-.6	1.13	.4	1.19	.75	.71	24 24
45	15	3.00	3.40	.91	.26	1.03	.2	.71	-.2	1.21	.69	.66	34 34
13	15	.87	.51	-1.03	.27	1.16	.5	.71	-.2	1.15	.61	.61	10 10
47	15	3.13	3.51	1.04	.27	1.02	.1	.69	-.2	1.16	.64	.63	46 46
33	15	2.20	2.33	.19	.24	.66	-.8	1.02	.1	1.00	.77	.76	17 17
50	15	3.33	3.65	1.27	.29	1.05	.2	.66	-.2	1.22	.60	.57	49 49
34	15	2.27	2.45	.25	.24	.72	-.6	.62	-.6	1.43	.85	.76	50 50
22	15	1.47	1.11	-.45	.24	.80	-.4	.62	-.6	1.13	.78	.73	20 20
38	15	2.53	2.86	.48	.24	.84	-.3	.61	-.6	1.42	.81	.73	48 48
29	15	1.93	1.86	-.04	.24	.63	-.9	.58	-.8	.85	.82	.76	41 41
18	15	1.20	.80	-.70	.25	.68	-.8	.55	-.7	1.14	.75	.69	37 37
15	15	1.00	.62	-.89	.26	.58	-1.1	.54	-.7	.88	.76	.65	15 15
19	15	1.27	.87	-.64	.25	.67	-.8	.54	-.8	1.20	.77	.70	26 26
26	15	1.73	1.51	-.22	.24	.48	-1.5	.53	-.9	1.12	.84	.76	2 2
24	15	1.60	1.30	-.34	.24	.53	-1.3	.47	-1.1	1.30	.87	.75	40 40
21	15	1.40	1.03	-.51	.24	.60	-1.0	.45	-1.1	1.52	.85	.72	16 16
28	15	1.87	1.74	-.10	.24	.37	-1.9	.37	-1.5	1.75	.94	.76	4 4
35	15	2.33	2.56	.31	.24	.41	-1.9	.36	-1.4	1.19	.89	.75	11 11
31	15	2.07	2.10	.07	.24	.36	-2.0	.39	-1.4	1.37	.90	.76	47 47
5	15	.33	.20	-1.81	.40	.41	-.8	.33	-.7	1.04	.63	.41	29 29
30	15	2.00	1.98	.02	.24	.31	-2.2	.33	-1.6	1.27	.91	.76	22 22
13	15	.87	.51	-1.03	.27	.38	-1.9	.30	-1.3	1.41	.82	.61	31 31
39	15	2.60	2.96	.54	.24	.29	-2.6	.29	-1.5	1.40	.90	.73	35 35
34	15	2.27	2.45	.25	.24	.28	-2.5	.32	-1.6	1.44	.93	.76	5 5
32	15	2.13	2.22	.13	.24	.30	-2.3	.27	-1.9	1.39	.92	.76	39 39
51	15	3.40	3.69	1.36	.30	.39	-1.5	.22	-1.1	1.19	.74	.55	45 45
30.0	15.0	2.00	2.00	.01	.25	.98	-.2	1.26	.2		.68		Mean (Count: 50)
11.7	.1	.79	1.07	.73	.03	.56	1.4	1.32	1.5		.18		S.D. (Population)
11.9	.1	.79	1.08	.74	.03	.56	1.4	1.33	1.5		.18		S.D. (Sample)

Model, Populn: RMSE .26 Adj (True) S.D. .68 Separation 2.68 Strata 3.90 Reliability .88
Model, Sample: RMSE .26 Adj (True) S.D. .69 Separation 2.71 Strata 3.94 Reliability .88
Model, Fixed (all same) chi-squared: 351.2 d.f.: 49 significance (probability): .00
Model, Random (normal) chi-squared: 42.9 d.f.: 48 significance (probability): .68

จาก Figure 3 แสดงผลการวิเคราะห์ฟาเซตผู้สอบ ซึ่งมีจำนวน 50 คน โดยผู้สอบที่มีค่า Infit-Outfit MS ไม่อยู่ในช่วงเกณฑ์ที่กำหนดไว้ คือ 0.5-1.5 หมายถึงข้อมูลไม่สอดคล้องกลมกลืนกับโมเดล เช่น ผู้สอบคนที่ 3 (student 3) มีค่า Infit MS อยู่ในเกณฑ์คือ 1.18 แต่มีค่า Outfit MS มากกว่าเกณฑ์ คือ 1.78 หมายถึง ผู้สอบคนนี้อาจถูกประเมินด้วยผลการประเมินแบบสุดโต่งในบางข้อรายการประเมิน หากพบกรณีผู้สอบที่มีคะแนนไม่สอดคล้องกับโมเดล ผู้วิเคราะห์ควรตรวจสอบความถูกต้องของการกรอกข้อมูลในเบื้องต้น หากพบว่าถูกต้องแล้ว ผู้วิเคราะห์อาจพิจารณาทดลองวิเคราะห์อีกครั้งโดยตัดผู้สอบที่ไม่สอดคล้องกับโมเดลออก จากนั้นตรวจสอบว่าค่าสถิติที่ต้องการจะศึกษาในฟาเซตอื่น ๆ ยังอยู่ในเกณฑ์ที่ยอมรับได้ของผู้วิเคราะห์หรือไม่ โดยการวิเคราะห์สามารถทำได้ทั้ง 2 แนวทาง คือ การวิเคราะห์โดยใช้ผู้สอบทั้งหมด หรือตัดผู้สอบที่ไม่สอดคล้องกับโมเดลออก (Linacre, 2010) ฟาเซตผู้สอบมีดัชนี Separation และดัชนี Reliability of Separation ผ่านเกณฑ์ที่กำหนดไว้ ($Sep. Index > 2.0$, $Rel. Index > 0.8$) สามารถแปลผลได้ว่าผู้สอบมีระดับความสามารถประมาณ 3 กลุ่ม ($Sep. Index = 2.68$) และมีความเที่ยงที่แสดงว่าผู้สอบมีความสามารถแตกต่างกันมากพอในการยืนยันความตรงเชิงโครงสร้าง ($Rel. Index = 0.88$) Fixed Chi-square มีนัยสำคัญทางสถิติ แสดงว่าผู้สอบมีความสามารถแตกต่างกัน ส่วน Random Chi-square ไม่มีนัยสำคัญ แสดงว่าตัวอย่างผู้สอบมีความสามารถใกล้เคียงกับการสุ่มมาจากประชากร

Figure 4

Example of evaluator FASET analysis results

ตัวอย่างผลการวิเคราะห์ฟาเซตผู้ประเมิน

Music Performance Assessment 2/5/2023 8:13:20 PM
Table 7.2.2 rater Measurement Report (arranged by fN).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model S.E.	Infit MnSq	Outfit MnSq	Estim. Discrm	Correlation PtMea	Exact Agree. Obs %	Exp %	N rater
419	250	1.68	1.37	.31	1.31	2.9	.78	.67	36.1	41.7	1 1
598	249	2.40	2.70	-.37	.99	.0	.88	.74	35.7	41.4	3 3
483	250	1.93	1.83	.07	.64	-4.2	1.32	.84	39.7	43.0	2 2
500.0	249.7	2.00	1.96	.00	.98	-.5		.75			Mean (Count: 3)
74.1	.5	.30	.55	.28	.27	3.0		.07			S.D. (Population)
90.7	.6	.37	.68	.35	.34	3.6		.08			S.D. (Sample)

Model, Populn: RMSE .06 Adj (True) S.D. .28 Separation 4.46 Strata 6.28 Reliability (not inter-rater) .95
 Model, Sample: RMSE .06 Adj (True) S.D. .34 Separation 5.51 Strata 7.68 Reliability (not inter-rater) .97
 Model, Fixed (all same) chi-squared: 62.0 d.f.: 2 significance (probability): .00
 Model, Random (normal) chi-squared: 1.9 d.f.: 1 significance (probability): .16
 Inter-Rater agreement opportunities: 748 Exact agreements: 278 = 37.2% Expected: 314.5 = 42.0%

จาก Figure 4 แสดงผลการวิเคราะห์ฟาเซตผู้ประเมิน ซึ่งมีจำนวน 3 คน โดยผู้ประเมินคนที่ 1 มีผลการวิเคราะห์ความสอดคล้องกลมกลืนกับโมเดลที่ไม่ผ่านเกณฑ์ โดยมีค่า Infit MS เท่ากับ 1.31 และมีค่า Outfit MS เท่ากับ 2.07 ซึ่งมากกว่าเกณฑ์ที่กำหนดไว้ในช่วง 0.5-1.5 หากพบความสอดคล้องกับโมเดลของผู้ประเมิน ผู้วิเคราะห์ควรตรวจสอบความถูกต้องของการกรอกข้อมูลในเบื้องต้น ทั้งนี้การนำเอาผลการประเมินของผู้ประเมินคนที่ไม่สอดคล้องออกจากการวิเคราะห์ อาจทำให้ข้อมูลไม่เชื่อมต่อกัน ซึ่งจะส่งผลกระทบต่อผลการแปลผลการวิเคราะห์ได้ (Linacre, 2022b) ดัชนี Separation และดัชนี Reliability of Separation แสดงถึงความแตกต่างของการกดหรือปล่อยคะแนนของผู้ประเมินทั้ง 3 คน Rasch Rater Agreement แสดงความไม่สอดคล้องของผลการประเมินของผู้ประเมินแต่ละคน โดยมีค่า Rasch Rater Agreement ที่สังเกตได้น้อยกว่าค่าที่คาดหวัง ค่า Point-Measure Correlation (PtMea) มีค่าเป็นบวกทั้งหมด แสดงว่าผู้ประเมินประเมินทักษะไปในทิศทางเดียวกัน Fixed Chi-square มีนัยสำคัญทางสถิติ แสดงว่าผู้ประเมินประเมินโดยกดหรือปล่อยคะแนนต่างกัน ส่วน Random Chi-square ไม่มีนัยสำคัญแสดงว่าตัวอย่างผู้ประเมินประเมินโดยกดหรือปล่อยคะแนนใกล้เคียงกับการสุ่มมาจากประชากร

Figure 5

Example of FASET analysis results of assessment items

ตัวอย่างผลการวิเคราะห์ฟาเซตข้อรายการประเมิน

Music Performance Assessment 2/5/2023 8:13:20 PM
Table 7.3.2 item Measurement Report (arranged by fN).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	S.E.	Infit MnSq	ZStd	Outfit MnSq	ZStd	Estim. Discrm	Correlation PtMea	PtExp	N item
107	149	.72	.46	1.13	.09	1.08	.5	2.06	3.9	1.04	.52	.53	5 5
457	150	3.05	3.37	-.87	.08	1.14	1.0	1.05	.3	.99	.59	.62	2 2
267	150	1.78	1.62	.17	.07	.99	.0	1.08	.5	.96	.66	.67	3 3
203	150	1.35	1.05	.51	.07	.87	-1.1	1.00	.0	1.03	.66	.64	4 4
466	150	3.11	3.43	-.93	.08	.82	-1.3	1.10	.5	.95	.61	.61	1 1
300.0	149.8	2.00	1.99	.00	.08	.98	-.2	1.26	1.1		.61		Mean (Count: 5)
141.4	.4	.94	1.21	.80	.01	.12	.9	.40	1.4		.05		S.D. (Population)
158.1	.4	1.05	1.36	.89	.01	.13	1.1	.45	1.6		.06		S.D. (Sample)

Model, Populn: RMSE .08 Adj (True) S.D. .79 Separation 9.84 Strata 13.45 Reliability .99
Model, Sample: RMSE .08 Adj (True) S.D. .89 Separation 11.01 Strata 15.02 Reliability .99
Model, Fixed (all same) chi-squared: 444.6 d.f.: 4 significance (probability): .00
Model, Random (normal) chi-squared: 4.0 d.f.: 3 significance (probability): .26

จาก Figure 5 แสดงผลการวิเคราะห์ฟาเซตข้อรายการประเมิน ซึ่งมีจำนวน 5 ข้อ โดยข้อรายการประเมินที่ 5 เป็นข้อที่ยากที่สุด (ผลการวัด = 1.13) และเป็นข้อที่มีผลการวิเคราะห์ความสอดคล้องกลมกลืนไม่ผ่านเกณฑ์ (Outfit MS = 2.06) ซึ่งหมายถึง ผลการประเมินด้วยข้อรายการประเมินนี้อาจมีคะแนนแบบสุดโต่ง หากพบความไม่สอดคล้องของข้อรายการประเมิน ผู้วิเคราะห์สามารถนำข้อดังกล่าวออกจากการวิเคราะห์ได้ อย่างไรก็ตาม ควรพิจารณาว่าหากนำข้อดังกล่าวออกจากเครื่องมือแล้ว ข้อรายการประเมินที่เหลืออยู่ควรเป็นตัวแทนที่ดีและครอบคลุมทักษะดนตรีที่ต้องการประเมิน (Bohlig et al., 1998) ฟาเซตข้อรายการประเมินมีดัชนี Separation และดัชนี Reliability of Separation ผ่านเกณฑ์ที่กำหนดไว้ (*Sep. Index* > 3.0, *Rel. Index* > 0.9) แสดงถึงข้อรายการประเมินมีความยากแตกต่างกันมากพอที่จะจำแนกความสามารถของผู้สอบได้ ค่า Point-Measure Correlation (PtMea) มีค่าเป็นบวกทั้งหมด แสดงว่าข้อรายการประเมินทุกข้อมุ่งประเมินในทักษะเดียวกัน Fixed Chi-square มีนัยสำคัญทางสถิติ แสดงว่าข้อรายการประเมินมีความยากแตกต่างกัน

บทสรุป (Conclusion)

โมเดลหลายองค์ประกอบของราสส์ เป็นโมเดลที่มีประโยชน์สำหรับการวัดประเมินผลทางการศึกษา โดยเฉพาะการวัดประเมินผลที่มีฟาเซตที่เกี่ยวข้องมากกว่า 2 ฟาเซตขึ้นไป ซึ่งพบทั่วไปในการวัดและประเมินผลภาคปฏิบัติ การประเมินทักษะ หรือ การประเมินสมรรถนะ ซึ่งโมเดลสามารถให้สารสนเทศของคุณสมบัติทางจิตมิติได้อย่างครบถ้วน และให้สารสนเทศของแต่ละฟาเซตในโมเดลสำหรับการตรวจสอบและพัฒนากการประเมิน เช่น สามารถตรวจสอบความผิดปกติของผลการประเมินที่เกิดขึ้นกับผู้สอบที่มีคะแนนไม่สอดคล้องกับโมเดล โดยอาจเกิดจากการกรอกข้อมูล หรืออาจเกิดจากผลการประเมินของผู้ประเมินที่ประเมิน โดยมีการกดหรือปล่อยคะแนนต่างกัน สามารถออกแบบหรือพัฒนารูปแบบการอบรมผู้ประเมินบางท่านเพิ่มเติมจากผลการวิเคราะห์ที่พบว่าผู้ประเมินยังประเมินได้ไม่สอดคล้องกับโมเดล เนื่องจากมีผลการประเมินที่สุดโต่ง หรือสามารถปรับปรุงข้อรายการประเมินที่ผลการวิเคราะห์พบว่าไม่สอดคล้องกับโมเดล โดยพิจารณาปรับปรุงจากผลการวิเคราะห์ค่า Infit-Outfit MS ที่สามารถบอกลักษณะของการใช้ข้อรายการประเมินนี้ได้ ซึ่งบทความนี้ได้นำเสนอแนวคิดของโมเดลหลายองค์ประกอบของราสส์ การเขียนคำสั่งสำหรับการวิเคราะห์โมเดลหลายองค์ประกอบด้วยโปรแกรม Facets ค่าสถิติที่สำคัญสำหรับการวิเคราะห์โมเดล

หลายองค์ประกอบของราชส์ และตัวอย่างการแปลผลการวิเคราะห์ เพื่อให้ผู้อ่านนำไปปรับใช้ในการวิเคราะห์ด้วยโมเดลได้อย่างเหมาะสม

อย่างไรก็ตาม นอกจากค่าสถิติที่นำเสนอในบทความนี้ ยังมีค่าสถิติอื่นๆ ที่น่าสนใจ เช่น ค่าสถิติสำหรับแนวทางการปรับปรุงประสิทธิภาพของระดับคะแนน (Linacre, 2002a) ซึ่งใช้สำหรับพัฒนาระดับคะแนนของเครื่องมือแบบตรวจให้คะแนน 2 ค่าให้มีประสิทธิภาพมากขึ้น รวมถึงค่าสถิติที่ใช้สำหรับการคำนวณค่าสถิติสำคัญนำเสนอในบทความนี้ เช่น True SD, RMSE หรือค่า Fair Average ที่ใช้ในการศึกษาผลการวัดของผู้สอบหลังการจัดความลำเอียงที่เกิดขึ้น ซึ่งโมเดลหลายองค์ประกอบได้ให้สารสนเทศเอาไว้สำหรับศึกษาต่อไป

References

- American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association
- Andrich, D. (1978). A rating scale formulation for ordered response categories. *Psychometrika*, 43(4), 561 – 573. <https://doi.org/10.1007/BF02293814>
- Bohlig, M., Fisher, W. P., Masters, G. N., & Bond, T. (1998). Content validity and misfitting Items. *Rasch Measurement Transactions*, 12(1), 607.
- Bond, T. G., Yan, Z., & Heene, M. (2021). *Applying the Rasch model: Fundamental measurement in human sciences* (4th ed.). London: Routledge.
- Eckes, T. (2009). *Introduction to many-facet Rasch measurement*. Peter Lang Edition.
- Eckes, T. (2019). Many-facet Rasch measurement: Implications for rather mediated language assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment volume I: Fundamental techniques*. (pp. 153-176). Routledge.
- 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. <https://doi.org/10.1177/1321103X18773103>
- Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing Performance: Designing, Scoring and Validating Performance Tasks*. The Guilford Press.
- Kanjanawasee, S. (2020). *Modern test theories* (5th ed.). Bangkok: CU Press.
- Kim, P. (2020). Many-Facet Partial Credit Model analysis of Dörnyei's L2 motivational self-system. *Journal of Second and Multiple Language Acquisition*, 8(4), 111-128.
- Krishnan, S., & Idris, N. (2018). Using partial credit model to improve the quality of an instrument. *International Journal of Evaluation and Research in Education (IJERE)*, 7(4), 313-316. [10.11591/ijere.v7.i4.pp313-316](https://doi.org/10.11591/ijere.v7.i4.pp313-316)
- Linacre, J. M. (1994). *Many-faceted Rasch measurement*. MESA Press.

- Linacre, J. M. (2000). Comparing “Partial Credit Models” (PCM) and “Rating Scale Models” (RSM). *Rasch Measurement Transactions*, 14, 768.
- Linacre, J. M. (2010). When to stop removing items and persons in Rasch misfit analysis? *Rasch Measurement Transactions*, 23(4), 1241.
- Linacre, J. M. (2002a). Understanding Rasch measurement: Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, 3(1), 85-106.
- Linacre, J. M. (2002b). What do Infit and Outfit, Mean-square and Standardized mean? *Archives of Rasch Measurement*, 16, 871-882.
- 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. (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>
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174.
- 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.
- Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47(1), 90-100. [https://doi.org/10.1016/S0022-2496\(02\)00028-7](https://doi.org/10.1016/S0022-2496(02)00028-7)
- Smith, R. M. (1996). Polytomous mean-square fit statistics. *Rasch Measurement Transactions*, 10(3), 516-517.
- Wesolowski, B. C., Amend, R. M., Barnstead, T. S., Edwards, A. S., Everhart, M., Goins, Q., Grogan III, R. J., Herceg, A. M., Jenkins, S. I., Johns, P. M., McCarver, C. J., Schaps, R. E., Sorrell, G. W., & Williams, J. D. (2017). The development of a secondary-level solo wind instrument performance rubric using the Multifaceted Rasch Partial Credit Measurement Model. *Journal of Research in Music Education*, 65, 95-119.
- Wesolowski, B. C. (2019). An examination of differential item functioning in a rubric to assess solo music performance. *Musicae Scientiae*, 1-15. <https://doi.org/10.1177/1029864919859928>
- Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8, 370-371.
- Wright, B. D., & Masters, G. N. (2002). Number of person or item strata. *Rasch Measurement Transaction*, 16(3), 888.