# การประยุกต์ใช้โปรแกรม R สำหรับการปรับเทียบคะแนนตามทฤษฎีการตอบข้อสอบ
# Application of R Program for Test Score Equating based on Item Response Theory

อนุสรณ์ เกิดศรี[1*], สังวรณ์ งัดกระโทก[2] และ ณภัทร ชัยมงคล[3]

[1]ค.ด. (การวัดและประเมินผลการศึกษา)/ ดร. สำนักทะเบียนและวัดผล มหาวิทยาลัยสุโขทัยธรรมาธิราช
*Corresponding Author E-mail: k.anusorn@hotmail.com
[2]Ph.D. (Measurement and Quantitative Methods)/รองศาสตราจารย์ ดร. สาขาศึกษาศาสตร์ มหาวิทยาลัยสุโขทัยธรรมาธิราช
sungworn@hotmail.com
[3]ค.ด. (การวัดและประเมินผลการศึกษา)/ ดร. คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย nhabhat.c@chula.ac.th

*Received: December 6, 2022 / Revised: December 13, 2022 / Accepted: December 30, 2022*

## บทคัดย่อ

การปรับเทียบคะแนนเป็นเทคนิคทางสถิติที่ใช้บ่อยในโปรแกรมการทดสอบเมื่อการทดสอบมีแบบทดสอบหลายฉบับ (ฟอร์ม) ที่นำไปใช้กับกลุ่มผู้สอบต่างกันเนื่องจากการทดสอบหลายครั้งกับผู้เข้าสอบหลายกลุ่มอาจส่งผลให้ข้อสอบถูกเปิดเผยมากเกินไป จึงอาจส่งผลกระทบต่อความปลอดภัยของการทดสอบ การใช้แบบทดสอบหลายฉบับจะทำให้การเปิดเผยข้อสอบลดลง อย่างไรก็ตามการใช้แบบทดสอบหลายฉบับจะทำให้คะแนนที่ได้มีความแตกต่างกัน เนื่องจากเป็นประเมินด้วยเนื้อหาคล้าย ๆ กัน และมีความยากแตกต่างกัน การเทียบเคียงคะแนนจากแบบทดสอบฉบับหนึ่งไปยังอีกฉบับหนึ่งจึงถูกนำมาใช้เพื่อให้เกิดความยุติธรรมในการทดสอบ ทฤษฎีการตอบข้อสอบถูกนำมาใช้เพื่อปรับเทียบคะแนนของแบบทดสอบ ในโปรแกรม R สามารถปรับเทียบคะแนนด้วยแพ็คเกจ equateIRT บทความนี้มีวัตถุประสงค์เพื่อนำเสนอ 1) แนวคิดพื้นฐานของการปรับเทียบคะแนน 2) การออกแบบวิธีการเก็บข้อมูลสำหรับการปรับเทียบคะแนน และ 3) การประยุกต์ใช้โปรแกรม R ในการปรับเทียบคะแนนแบบทดสอบ

**คำสำคัญ :** การปรับเทียบคะแนนแบบทดสอบ, ทฤษฎีการตอบข้อสอบ, โปรแกรมอาร์

## Abstract

Equating is a statistical technique frequently utilized in testing programs because administrations throughout multiple occasions and examinee groups can result in item overexposure and threaten the test's security. In reality, item exposure can be reduced by using several test forms; however, utilizing different tests would result in employing different score scales that assess the relevant construct at various degrees of difficulty. Equating scores from one test to another test is used to ensure fairness of tests. IRT models are implemented for equating various forms. Program R can equate scores from multiple test forms using the package called equateIRT. This article aims to present 1) fundamental concepts of linking and equating scores, 2) data collection designing for equating and 3) the application of R programming in equating test scores.

**Keywords:** Equating Test Scores, Item Response Theory, R program

## บทนำ

การบริหารการสอบในโปรแกรมการทดสอบที่มีการจัดสอบหลายครั้งกับกลุ่มผู้เข้าสอบมากกว่าหนึ่งกลุ่ม มักหลีกเลี่ยงการใช้ฟอร์มของแบบทดสอบที่ซ้ำกับแบบทดสอบที่ผ่านการจัดสอบมาแล้ว เนื่องจากอาจนำไปสู่การเปิดเผยข้อสอบบ่อยครั้งเกินไป ทำให้กลุ่มผู้สอบจำข้อสอบเหล่านั้นได้ ซึ่งอาจส่งผลกระทบต่อความปลอดภัยของการทดสอบ ในทางปฏิบัติการเปิดเผยข้อสอบสามารถจำกัดได้โดยใช้แบบทดสอบฉบับหรือฟอร์มอื่น อย่างไรก็ตาม การใช้แบบทดสอบหลายฟอร์มหรือหลายฉบับ (Test forms) นำไปสู่คะแนนที่ได้มาจากหลากหลายมาตรวัดที่มีระดับความยากง่ายต่างกัน หน่วยงานที่บริหารโปรแกรมการทดสอบจึงใช้วิธีการทางสถิติที่หลากหลายเพื่อแปลงคะแนนจากแบบทดสอบฉบับหนึ่งไปยังแบบทดสอบฉบับอื่น โดยมีเป้าหมายเพื่อทำให้เกิดความเท่าเทียมกันของคะแนนในเรื่องการแปลและการใช้ผลคะแนน วิธีการเหล่านี้เรียกว่าการเทียบเคียง (Linking) คะแนน เป็นการปรับระดับความยากของแบบทดสอบในแบบฟอร์มหนึ่งไปยังอีกแบบฟอร์มหนึ่ง ซึ่งนำมาใช้เป็นฐานในการปรับเทียบ วิธีการเทียบเคียงและการปรับเทียบนั้นแตกต่างกันตามทฤษฎีการทดสอบที่นำมาใช้ ไม่ว่าจะเป็นการใช้ทฤษฎีการทดสอบแบบดั้งเดิม หรือการใช้ทฤษฎีการตอบข้อสอบ วิธีการเหล่านี้มีการดำเนินการค่อนข้างซับซ้อนและต้องใช้ซอฟต์แวร์เฉพาะในการประมวลผลซึ่งมีทั้งซอฟต์แวร์ที่ต้องเสียค่าใช้จ่าย เช่น Bilog-MG flexMIRT และ IRTPRO หรือซอฟต์แวร์ที่เปิดให้ใช้งานโดยไม่เสียค่าใช้จ่ายโดยไม่เสียค่าใช้จ่าย เช่น โปรแกรม R

R เป็นภาษาและสภาพแวดล้อมสำหรับการคำนวณเชิงสถิติและกราฟิกที่เปิดให้ใช้งานครั้งแรกเมื่อปี ค.ศ. 1993 ผู้ใช้งานสามารถดาวน์โหลดได้โดยไม่เสียค่าใช้จ่าย จาก http://www.R-project.org โปรแกรม R สามารถจัดเตรียมชุดคำสั่งสำหรับการวิเคราะห์ข้อมูลที่หลากหลาย เช่น การสร้างแบบจำลองเชิงเส้น และไม่เป็นเชิงเส้น การทดสอบทางสถิติแบบดั้งเดิม การวิเคราะห์อนุกรมเวลา การจัดหมวดหมู่ การจัดกลุ่ม และฟังก์ชันการใช้งานสำหรับนำเสนอข้อมูลเป็นภาพกราฟิกที่สวยงาม และจุดเด่นของโปรแกรม R อีกประการหนึ่งคือ การมีลิขสิทธิ์แบบโอเพ่นซอร์ส (open source) เป็นการเปิดเผยโค้ดเพื่อให้กลุ่มความร่วมมือของชุมชนทางวิชาการนำเทคนิควิธีการวิเคราะห์ที่พัฒนาขึ้นใหม่ มาสร้างเป็นแพ็คเกจหรือชุดคำสั่งให้ผู้ใช้งานทั่วไปสามารถนำไปใช้หรือนำไปแก้ไขได้อย่างอิสระโดยไม่มีค่าใช้จ่าย

บทความนี้จึงได้นำเสนอทางเลือกในการปรับเทียบคะแนนโดยการประยุกต์ใช้โปรแกรม R ในการวิเคราะห์และประมวลผลคะแนนของการทดสอบ ด้วยแพ็คเกจ equateIRT ที่พัฒนาโดย Battauz (2015) การนำเสนอเนื้อหาในบทความจะครอบคลุม 4 ประเด็น ดังนี้ 1) ความแตกต่างของการเทียบเคียงและการปรับเทียบคะแนน 2) การออกแบบการเก็บรวบรวมข้อมูล 3) การปรับเทียบคะแนนโดยการใช้ทฤษฎีการตอบข้อสอบ และ 4) การประยุกต์ใช้โปรแกรม R สำหรับการปรับเทียบคะแนนตามทฤษฎีการตอบข้อสอบ

### 1. ความแตกต่างของการเทียบเคียง (Linking) และการปรับเทียบ (Equating)

ในมุมของผู้ใช้คะแนนมักเกิดข้อสงสัยว่าผลการทดสอบต่าง ๆ มีความเกี่ยวข้องกันอย่างไร ผู้ใช้ผลคะแนนบางคนระมัดระวังและลังเลที่จะทำการเปรียบเทียบคะแนนจากการทดสอบต่าง ๆ บางคนสันนิษฐานว่าการทดสอบทั้งหมดสามารถเชื่อมโยงในลักษณะที่นำไปสู่การเปรียบเทียบอย่างง่ายและการอนุมานไปยังการแปลผลคะแนนได้อย่างถูกต้องสำหรับ “การทดสอบที่แตกต่างกัน” ซึ่งอาจหมายถึงการทดสอบเดียวกันที่มีแบบทดสอบสองฉบับที่สร้างขึ้นตามผังการออกข้อสอบที่ระบุอย่างชัดเจน “แบบทดสอบที่แตกต่างกัน” หมายถึง การวัดของโครงสร้างเดียวกัน (เช่น คณิตศาสตร์) ที่สร้างขึ้นตามข้อกำหนดที่แตกต่างกัน เช่น การทดสอบความสามารถทางคณิตศาสตร์ของบริษัท ACT และการทดสอบของ College Board จะเห็นได้ว่าการทดสอบโดยสององค์กร วัดความสามารถทางคณิตศาสตร์เหมือนกัน แต่มีความแตกต่างกันในเรื่องของโครงสร้างที่วัด เมื่อต้องการเทียบคะแนนสองการทดสอบจากแบบทดสอบ 2 ฉบับนี้ จะเรียกว่าการเทียบเคียง หรือ linking ซึ่งเป็นคำที่มีความหมายกว้างและมีความเข้มงวดน้อยกว่าการปรับเทียบ หรือ equating (Dorans & Puhan, 2017)

Dorans and Puhan (2017) กล่าวว่า การเทียบเคียงคะแนนของการทดสอบเป็นคำที่มีนัยกว้าง จึงได้จัดกลุ่มเพื่ออธิบายแนวคิดการเทียบเคียงคะแนนการทดสอบ เป็น 3 ประเภท ดังนี้

1) การปรับเทียบคะแนน จัดเป็นซับเซตหรือส่วนหนึ่งในวิธีการเทียบเคียงคะแนนที่ผ่านการปรับเทียบจะสามารถใช้แทนกันได้ (Exchangeable) แสดงว่าคะแนนจะต้องมาจากแบบทดสอบที่มีเนื้อหาเดียวกันหรือมีโครงสร้างแบบเดียวกัน

2) การปรับมาตรคะแนน Scaling เป็นการเทียบเคียงการกระจายของคะแนนจากฉบับต่าง ๆ ให้เทียบเคียงกันได้ เนื่องจากแบบทดสอบอาจจะวัดโครงสร้างที่เหมือนกันและอาจจะมีบางส่วนต่างกัน เช่น แบบทดสอบที่วัดความสามารถทางคณิตศาสตร์ของ ACT และของ SAT ซึ่งมีเนื้อหาบางส่วนต่างกัน จึงทำให้แบบทดสอบทั้งสอบฉบับมีโครงสร้างไม่เหมือนกัน

3) การทำนายคะแนน (Predicting) เป็นการนำคะแนนจากการทดสอบหนึ่งไปทำนายอีกการทดสอบหนึ่ง โดยคาดหวังให้มีความคลาดเคลื่อนในการทำนายต่ำ ๆ เช่น คะแนนสอบ SAT กับ GPA ของการเรียนระดับอุดมศึกษาในชั้นปีแรก

ดังนั้น เมื่อกล่าวถึงการปรับเทียบคะแนน จึงมีนัยถึงกระบวนการทางสถิติที่ใช้เพื่อปรับคะแนนจากแบบทดสอบต่างฉบับกัน ซึ่งอาจมีความยากแตกต่างกันให้สามารถเทียบเคียงกันได้ และสามารถใช้คะแนนของแบบทดสอบฉบับใดฉบับหนึ่งแทนกันได้ ประหนึ่งว่ามาจากแบบทดสอบฉบับเดียวกัน ในกรณีที่แบบทดสอบทั้งสองฉบับไม่เป็นแบบทดสอบที่มีความคู่ขนาน (Parallel form) หรือมีเนื้อหาบางส่วนแตกต่างกันบางประการ เช่นนี้จะถูกเรียกว่า การเทียบเคียง (Linking) ซึ่งมีข้อตกลงเบื้องต้นที่ผ่อนปรนมากกว่าการปรับเทียบ โดยที่ Dorans and Holland (2000) กล่าวถึงข้อตกลงเบื้องต้นที่จำเป็นที่จะทำให้การเทียบเคียงคะแนน มีฐานะเท่ากับการปรับเทียบคะแนน ซึ่งมีข้อตกลงเบื้องต้น 5 ประการ ดังนี้

1) ข้อตกลงด้านความเป็นโครงสร้างเดียวกัน (Same construct) วัดในเนื้อหาเดียวกัน แบบทดสอบที่วัดโครงสร้างต่างกันไม่ควรนำมาปรับเทียบคะแนนกัน

2) ข้อตกลงด้านความเท่าเทียมกันของความเที่ยง (Equal reliability) การทดสอบที่วัดโครงสร้างเดียวกัน เนื้อหาเดียวกันแต่มีความเที่ยงต่างกันไม่ควรนำมาปรับเทียบกัน

3) ข้อตกลงด้านความสมมาตร (Symmetry) ฟังก์ชันการปรับเทียบสำหรับการเทียบคะแนนของแบบทดสอบ Y กับ คะแนนของแบบทดสอบ X ควรเป็นฟังก์ชันที่ผกผันกัน สำหรับการเทียบคะแนนของแบบทดสอบ X ไปยังแบบทดสอบ Y

4) ข้อตกลงด้านความเท่าเทียมกัน (Equity) ความสามารถของผู้สอบหรือคะแนนที่ได้จากการทำแบบทดสอบต่างฉบับที่ถูกปรับเทียบแล้วควรมีคะแนนเท่ากัน และการแจกแจงของคะแนนเฉลี่ยหรือคะแนนคาดหวังที่เป็นความสามารถของผู้สอบ (Distribution of performance) จากแบบทดสอบต่างฉบับกันที่ได้รับการปรับเทียบควรจะมีการแจกแจงเหมือนกัน

5) ข้อตกลงด้านความไม่แปรเปลี่ยนของประชากร (Population invariance) ตามกลุ่ม เนื่องจากฟังก์ชันการปรับเทียบที่ใช้เชื่อมโยงคะแนนจากแบบทดสอบ X และ Y ควรมีค่าไม่แปรเปลี่ยนไปตามประชากรกลุ่มย่อย ถ้าประชากรที่ใช้ในการปรับเทียบมีกลุ่มย่อย เช่น เพศชาย และหญิง สมการที่ใช้ปรับเทียบของกลุ่มเพศชาย และกลุ่มเพศหญิงต้องเป็นสมการเดียวกัน ถ้าไม่ใช่สมการเดียวกันอาจเกิดเหตุการณ์ได้เปรียบหรือเสียเปรียบกันระหว่างกลุ่ม โดยเป็นการแสดงถึงคุณภาพของการปรับเทียบในกลุ่มย่อย

**2. การออกแบบการเก็บรวบรวมข้อมูลสำหรับการปรับเทียบ**

การออกแบบการเก็บรวบรวมข้อมูลสำหรับการปรับเทียบคะแนน มีหลายรูปแบบซึ่งสามารถศึกษารายละเอียดเพิ่มเติมได้จาก (Kolen & Brennan, 2014) ในบทความนี้จะนำเสนอรูปแบบการออกแบบการเก็บรวบรวมข้อมูลสำหรับการปรับเทียบคะแนน 3 รูปแบบ ซึ่งเป็นรูปแบบพื้นฐาน ได้แก่ 1) รูปแบบผู้สอบกลุ่มเดียว (Single-group design) 2) ผู้สอบกลุ่มเดียวทำแบบทดสอบทั้งสองชุดที่ได้รับการจัดให้สมดุล (Single-group design with counterbalancing) และ 3) รูปแบบผู้สอบกลุ่มไม่เท่าเทียมกันด้วยการใช้แบบทดสอบร่วม (Nonequivalent-groups with anchor test) ซึ่งเป็นรูปแบบที่นิยมใช้กันมาก รายละเอียดแต่ละรูปแบบมีดังนี้ (Livingston & Kim, 2010)

**2.1 รูปแบบผู้สอบกลุ่มเดียว (single-group design)**

การออกแบบแบบกลุ่มเดียวจะสุ่มตัวอย่างจากประชากรเป้าหมาย P เพื่อทำแบบทดสอบที่แตกต่างกันสองฉบับ คือ X และ Y โดยผู้สอบแต่ละกลุ่มต้องทำแบบทดสอบทั้งสองฉบับ ทำฉบับใดก่อนก็ได้ แล้วตามด้วยอีกฉบับ (Kolen & Brennan, 2014)

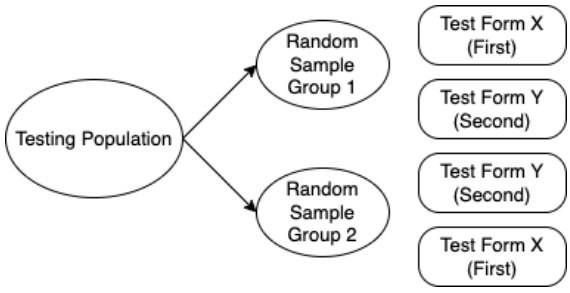

| ประชากร | ตัวอย่าง | แบบทดสอบ X | แบบทดสอบ Y |
|---|---|---|---|
| P | 1 | ✓ | ✓ |

**ภาพที่ 1** การออกแบบการเก็บรวบรวมข้อมูล รูปแบบผู้สอบกลุ่มเดียว

จุดเด่นของการออกแบบกลุ่มเดี่ยว คือ ตอบข้อโต้แย้งในเรื่องความเท่าเทียมของระดับความสามารถของกลุ่มตัวอย่างนั้นใกล้เคียงกันหรือไม่ สำหรับข้อจำกัดของการออกแบบแบบกลุ่มเดียว คือ การจัดแบบทดสอบสองชุดแยกกันให้กับกลุ่มเดียวกัน การออกแบบลักษณะดังกล่าวอาจไม่สามารถนำไปใช้งานได้จริง ในการจัดโปรแกรมการทดสอบไม่สามารถจัดให้ผู้สอบทำแบบทดสอบที่สมบูรณ์สองฉบับ เนื่องจากความเหนื่อยล้าของผู้สอบ ทำให้ศักยภาพในการทำข้อสอบลดลง และการออกแบบลักษณะนี้ยังเปิดเผยข้อสอบทั้งหมดทุกแบบฟอร์มให้แก่ผู้สอบ ซึ่งอาจไม่เหมาะสมเนื่องจากความเสี่ยงด้านความปลอดภัยของการทดสอบ

**2.2 รูปแบบผู้สอบกลุ่มเดียวที่จัดให้สมดุล (Single group design with counterbalancing)**

การออกแบบลักษณะนี้มีการเก็บข้อมูลจากตัวอย่าง 2 กลุ่ม ที่มาจากประชากรกลุ่มเดียวกัน โดยตัวอย่างกลุ่มที่ 1 ทำแบบทดสอบ X ก่อน แล้วจึงทำแบบทดสอบ Y และตัวอย่างกลุ่มที่ 2 ทำแบบทดสอบ Y ก่อน แล้วจึงทำแบบทดสอบ X เพื่อแก้ปัญหาเรื่องความเหนื่อยล้า (Kolen & Brennan, 2014)



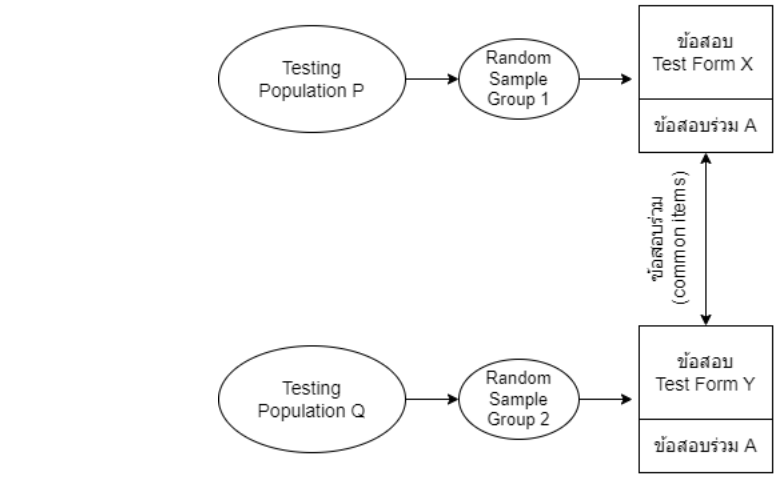

| ประชากร | ตัวอย่าง | แบบทดสอบ X (ลำดับที่ 1) | แบบทดสอบ Y (ลำดับที่ 1) | แบบทดสอบ X (ลำดับที่ 2) | แบบทดสอบ Y (ลำดับที่ 2) |
|---|---|---|---|---|---|
| P | 1 | ✓ | | | ✓ |
| P | 2 | | ✓ | ✓ | |

**ภาพที่ 2** การออกแบบการเก็บรวบรวมข้อมูล รูปแบบผู้สอบกลุ่มเดียวที่จัดให้สมดุล

จุดเด่นของการออกแบบแบบกลุ่มเดียวสมดุล คือ การควบคุมอิทธิพลของลำดับจากการทำแบบทดสอบ ซึ่งประสบการณ์ในการทำแบบทดสอบ A อาจทำให้ผลการเรียนของนักเรียนเปลี่ยนไปในแบบทดสอบ B อย่างไรก็ตามสำหรับข้อจำกัด คือ การจัดการทดสอบสองครั้งแบบแยกกันให้กับนักเรียนกลุ่มย่อยกลุ่มเดียวกันนั้นอาจไม่เหมาะสมในการนำไปปฏิบัติจริง

**2.3 รูปแบบผู้สอบกลุ่มไม่เท่าเทียมกันด้วยการใช้แบบทดสอบร่วม (Nonequivalent-groups with anchor test: NEAT)**

การออกแบบการเก็บข้อมูล จะเก็บข้อมูลจากประชากร 2 กลุ่ม คือ ตัวอย่างกลุ่ม 1 สุ่มจากประชากร P และตัวอย่างกลุ่ม 2 สุ่มจากประชากร Q กลุ่มผู้สอบทั้ง 2 กลุ่ม มีลักษณะไม่เท่าเทียมหรือแตกต่างกัน อันเนื่องมาจากเป็นกลุ่มผู้สอบต่างโปรแกรมการศึกษาหรือระดับชั้นกัน การขจัดความแตกต่างดังกล่าวสามารถใช้แบบทดสอบร่วม A ซึ่งเป็นการวัดความสามารถทั่วไปสำหรับทั้งสองกลุ่ม ความสามารถที่ไม่เท่าเทียมกันจะถูกควบคุมหรือขจัดออกด้วยแบบทดสอบร่วม (Kolen & Brennan, 2014)




| ประชากร | ตัวอย่าง | แบบทดสอบ X | แบบทดสอบ Y | แบบทดสอบร่วม A |
|---|---|---|---|---|
| P | 1 | ✓ | | ✓ |
| Q | 2 | | ✓ | ✓ |

**ภาพที่ 3** การออกแบบการเก็บรวบรวมข้อมูล รูปแบบผู้สอบกลุ่มไม่เท่าเทียมกันด้วยการใช้แบบทดสอบร่วม

การออกแบบการเก็บข้อมูลสำหรับปรับเทียบคะแนนแบบ NEAT นั้น ขั้นตอนการเลือกแบบทดสอบที่ใช้เป็นแบบทดสอบร่วมมีความสำคัญอย่างยิ่ง โดยต้องแสดงความเป็นตัวแทนของเนื้อหาตามสัดส่วนที่ใกล้เคียงกันของแบบทดสอบทั้งฉบับ อาจกล่าวได้ว่าแบบทดสอบร่วม (Anchor test) เป็น “ฉบับย่อ” ของแบบทดสอบฉบับเต็ม (Kolen & Brennan, 2014)

**3. การปรับเทียบคะแนนโดยการใช้ทฤษฎีการตอบข้อสอบ**

การปรับเทียบคะแนนโดยใช้ทฤษฎีการตอบข้อสอบ (Item Response Theory : IRT) สามารถแก้ไขข้อจำกัดของการปรับเทียบที่ใช้ทฤษฎีการทดสอบแบบดั้งเดิม ในด้านความไม่เสมอภาค ความไม่สมมาตร และความไม่ผันแปรตามกลุ่มผู้สอบซึ่งต้องเลือกใช้โมเดล IRT ที่มีความสอดคล้องกับข้อมูล (Kolen & Brennan, 2014) โดยในการปรับเทียบคะแนน

ตามวิธีการนี้อาศัยรูปแบบความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องของผู้สอบข้อสอบใด ๆ โดยผลการตอบข้อสอบขึ้นอยู่กับระดับความสามารถของผู้สอบ ซึ่งประมาณค่าได้จาก โค้งคุณลักษณะข้อสอบ (Item Characteristic Curve : ICC) ของแต่ละรูปแบบโดยอาจอยู่ในรูปแบบ 1, 2 หรือ 3 พารามิเตอร์สำหรับข้อสอบที่มีการให้คะแนนแบบสองค่า (Dichotomous Scoring) โดยไม่ได้ขึ้นอยู่กับการ แจกแจงคะแนนของกลุ่มตัวอย่างซึ่งมีข้อกำหนดสำคัญตามที่กล่าวมาแล้ว การดำเนินการทั้งหมดควรอยู่บนข้อตกลงเบื้องต้นเกี่ยวกับทฤษฎี IRT ที่ Hambleton and Swaminathan (1985) เสนอไว้ 4 ประการ ดังนี้ 1) ความเป็นเอกมิติ (Unidimensional) คือ คุณลักษณะหรือความสามารถที่เป็นตัวกำหนดพฤติกรรมการตอบข้อสอบแต่ละข้อ (ตอบผิดหรือถูก)มีลักษณะเด่นหรือความสำคัญเพียงลักษณะเดียว 2) ความเป็นอิสระ (Local independence) การตอบข้อสอบแต่ละข้อของผู้สอบมีความเป็นอิสระจากกัน หรือไม่มีอิทธิพลต่อกัน ทั้งระหว่างข้อสอบและระหว่างผู้สอบ 3) โค้งคุณลักษณะข้อสอบสามารถอธิบายพฤติกรรมการตอบข้อสอบ (Item Characteristic Curves/Item Response Model) ความสัมพันธ์ระหว่างระดับความสามารถของผู้สอบกับโอกาสในการตอบข้อสอบถูกสามารถแสดงได้ด้วยโค้งคุณลักษณะข้อสอบ 4) ข้อสอบที่ใช้ต้องไม่เป็นข้อสอบประเภทความเร็ว (Non-speededness of the test) ผู้สอบทุกคนควรมีโอกาสทำข้อสอบทุกข้อ เพื่อให้คะแนนรวมจากการสอบเป็นค่าความสามารถที่แท้จริง ไม่มีข้อจำกัดเกี่ยวกับเวลาที่ใช้ในการสอบ

การปรับเทียบคะแนนภายใต้ทฤษฎีตอบข้อสอบจำเป็นต้องมีฟังก์ชันการแปลงที่เชื่อมโยงแบบทดสอบ X กับแบบทดสอบ Y บนมาตรของทฤษฎีการตอบข้อสอบ การแปลงมาตรของ IRT จะแตกต่างกันไปขึ้นอยู่กับการออกแบบการเก็บข้อมูลสำหรับการปรับเทียบคะแนนที่นำมาใช้ ตัวอย่างเช่น ทั้งในการออกแบบการเก็บข้อมูลสำหรับการปรับเทียบคะแนนแบบกลุ่มเดียวและกลุ่มเท่าเทียม ความสามารถของผู้สอบที่สอบแบบทดสอบฉบับ X และแบบทดสอบฉบับ Y เป็นความสามารถของคนกลุ่มเดียวกันมีความเท่าเทียมกัน จึงไม่น่ากังวลในเรื่องความต่างของความสามารถของผู้สอบ ดังนั้นการปรับเทียบคะแนนสำหรับผู้สอบกลุ่มเดียวจึงมีความจำเป็นน้อย เนื่องจากอยู่บนข้อตกลงเบื้องต้นเดียวกัน คือ ความสามารถมีการแจกแจงลักษณะเดียวกัน มีค่าเฉลี่ยและความแปรปรวนเท่า ๆ กัน ขณะที่ใช้การออกแบบการเก็บข้อมูลแบบ NEAT กลุ่มผู้สอบจะมีความสามารถไม่เท่าเทียมกัน ดังนั้นจำเป็นต้องมีการแปลงมาตรของค่าพารามิเตอร์ภายใต้ทฤษฎีตอบข้อสอบ เนื่องจากค่าพารามิเตอร์ของ แบบทดสอบและผู้สอบเป็นคนละกลุ่มและมาจากแบบทดสอบต่างฉบับกัน ความแตกต่างดังกล่าวจะต้องถูกปรับให้อยู่ในมาตรเดียวกัน หรือเรียกว่า การเทียบเคียงพารามิเตอร์ตามทฤษฎีการตอบข้อสอบ (IRT parameter linking) ซึ่งเป็นการจัดวางค่าประมาณของพารามิเตอร์ตามทฤษฎีตอบข้อสอบจากแบบทดสอบที่แยกฉบับกัน ให้อยู่บนมาตรเดียวกันโดยใช้ข้อสอบร่วม (Common items หรือ Common scale) โดยทั่วไปการปรับเทียบความสามารถจากแบบทดสอบ Y ไปยังแบบทดสอบ X จะกำหนดข้อตกลงเบื้องต้นว่าเป็นการแปลงคะแนนโดยใช้สมการเชิงเส้น (González & Wiberg, 2017) มีสูตรการคำนวณดังนี้

$$\theta_{x_j} = A\theta_{y_j} + B$$

เมื่อ $\theta_{x_j}$ แทน ความสามารถของผู้สอบคนที่ j ที่ทำข้อสอบแบบทดสอบ X

$\theta_{y_j}$ แทน ความสามารถของผู้สอบคนที่ j ที่ทำข้อสอบแบบทดสอบ Y

ความสัมพันธ์ระหว่างพารามิเตอร์ข้อสอบในแบบทดสอบทั้งสองฉบับมีดังนี้

$$b_{x_i} = Ab_{y_i} + B$$

$$a_{x_i} = \frac{a_{y_i}}{A}$$

$$c_{x_i} = c_{y_i}$$

เมื่อ $A$ แทน ค่าสัมประสิทธิ์การปรับเทียบพารามิเตอร์อำนาจ

$B$ แทน ค่าสัมประสิทธิ์การปรับเทียบพารามิเตอร์ความยาก

$b_{x_i}$ แทน ค่าพารามิเตอร์ความยากของข้อสอบข้อที่ i แบบทดสอบ X

$b_{y_i}$ แทน ค่าพารามิเตอร์ความยากของข้อสอบข้อที่ i แบบทดสอบ Y

$a_{x_i}$ แทน ค่าพารามิเตอร์อำนาจจำแนกของข้อสอบข้อที่ i แบบทดสอบ X

$a_{y_i}$ แทน ค่าพารามิเตอร์อำนาจจำแนกของข้อสอบข้อที่ i แบบทดสอบ Y

$c_{x_i}$ แทน ค่าพารามิเตอร์โอกาสการเดาข้อสอบข้อที่ i แบบทดสอบ X

$c_{y_i}$ แทน ค่าพารามิเตอร์โอกาสการเดาข้อสอบข้อที่ i แบบทดสอบ Y

กล่าวโดยสรุป การปรับเทียบคะแนนโดยการใช้ทฤษฎีการตอบข้อสอบเกี่ยวข้องกับการดำเนินการ 3 ขั้นตอน คือ ก) การเลือกรูปแบบการเก็บรวบรวมข้อมูล ข) การประมาณค่าพารามิเตอร์ข้อสอบบนแบบทดสอบร่วม (สเกลร่วม) และ ค) การปรับเทียบคะแนนสอบ ในหัวข้อต่อไปนี้จะกล่าวถึงการประมาณค่าพารามิเตอร์ข้อสอบบนแบบทดสอบร่วม ซึ่งเป็นวิธีการคำนวณค่าสัมประสิทธิ์สำหรับการปรับเทียบพารามิเตอร์ข้อสอบที่เป็นแนวคิดพื้นฐาน 3 วิธี คือ 1) วิธีค่าเฉลี่ย - ค่าเฉลี่ย 2) วิธีค่าเฉลี่ยและซิกมา และ 3) วิธีโค้งลักษณะข้อสอบ ซึ่งมีวิธีการคำนวณที่ง่ายไม่ซับซ้อน และเป็นพื้นฐานให้ผู้อ่านเข้าใจแนวคิดของวิธีการคำนวณค่าสัมประสิทธิ์สำหรับการปรับเทียบที่ซับซ้อน ซึ่งนำเสนอรายละเอียดในหัวข้อถัดไป สำหรับผู้ที่สนใจวิธีการอื่น ๆ สามารถศึกษาเพิ่มเติมได้จากหนังสือหลักการและการประยุกต์ใช้ทฤษฎีตอบข้อสอบของ Hambleton and Swaminathan (1985) และหนังสือเรื่อง Test equating, scaling, and linking ของ Kolen and Brennan (2014)

**3.1 วิธีค่าเฉลี่ย - ค่าเฉลี่ย (Mean-mean method)**

การปรับค่าพารามิเตอร์อำนาจจำแนก (a) และพารามิเตอร์ความยาก (b) โดยใช้ค่าเฉลี่ยของ a หาค่าคงที่ A และค่าเฉลี่ยของ b หาค่าคงที่ B โดยใช้สูตรการคำนวณ ดังนี้

$$A = \frac{\mu(a_y)}{\mu(a_x)},$$

$$B = \mu(b_x) - A\mu(b_y)$$

เมื่อ $\mu(a_y)$ แทน ค่าเฉลี่ยพารามิเตอร์อำนาจจำแนกของข้อสอบร่วมจากแบบทดสอบ Y

$\mu(a_x)$ แทน ค่าเฉลี่ยพารามิเตอร์อำนาจจำแนกของข้อสอบร่วมจากแบบทดสอบ X

$\mu(b_y)$ แทน ค่าเฉลี่ยพารามิเตอร์ความยากของข้อสอบร่วมจากแบบทดสอบ Y

$\mu(b_x)$ แทน ค่าเฉลี่ยพารามิเตอร์ความยากของข้อสอบร่วมจากแบบทดสอบ X

**3.2 วิธีค่าเฉลี่ยและซิกมา (Mean and sigma method)**

$$A = \frac{\sigma(b_x)}{\sigma(b_y)}$$

เมื่อ $\sigma(b_x)$ แทน ส่วนเบี่ยงเบนพารามิเตอร์ความยากของข้อสอบร่วมจากแบบทดสอบ X

$\sigma(b_y)$ แทน ส่วนเบี่ยงเบนพารามิเตอร์ความยากของข้อสอบร่วมจากแบบทดสอบ Y

**3.3 วิธีโค้งคุณลักษณะข้อสอบ (Characteristic curve method)**

วิธีโค้งลักษณะข้อสอบ มีด้วยกัน 2 วิธี คือ วิธีของ Haebara ซึ่งนำเสนอในปี 1980 แสดงความแตกต่างระหว่างโค้งคุณลักษณะข้อสอบ เป็นผลรวมของผลต่างกำลังสองระหว่างโค้งคุณลักษณะข้อสอบ ของข้อสอบแต่ละข้อสำหรับแต่ละความสามารถของผู้สอบ และวิธีของ Stocking และ Lord ซึ่งนำเสนอในปี 1983 แสดงความแตกต่างระหว่างโค้งคุณลักษณะข้อสอบ เป็นผลต่างกำลังสองระหว่างโค้งคุณลักษณะข้อสอบสำหรับแต่ละความสามารถของผู้สอบ (Kolen & Brennan, 2014) วิธีการทั้ง 2 วิธี ให้ผลลัพธ์ที่ใกล้เคียงกันในการคำนวณค่าสัมประสิทธิ์การปรับเทียบพารามิเตอร์ข้อสอบ

วิธีโค้งลักษณะข้อสอบสามารถคำนวณได้จากการนำสารสนเทศที่ได้จากค่าพารามิเตอร์ของข้อสอบทั้งค่าพารามิเตอร์ความยาก และอำนาจจำแนก มาใช้ในการคำนวณค่าคงที่ สำหรับการปรับเทียบคะแนนระหว่างแบบทดสอบ ดังนี้

ให้ $T_{xa}$ แทนคะแนนจริงของข้อสอบที่มีความสามารถ ( $\theta_a$ ) จากการทำแบบทดสอบ X ซึ่งมีข้อสอบร่วม k ข้อ

$$T_{xa} = \sum_{i=1}^{K} P(\theta_a, b_{x_{ci}}, a_{x_{ci}}, c_{x_{ci}})$$

ในทำนองเดียวกันให้ $T_{ya}$ แทนคะแนนจริงของผู้สอบที่มีความสามารถเท่ากัน ( $\theta_a$ ) จากการทำแบบทดสอบ Y ซึ่งมีข้อสอบร่วม k ข้อ

$$T_{ya} = \sum_{i=1}^{K} P(\theta_a, b_{y_{ci}}, a_{y_{ci}}, c_{y_{ci}})$$

สำหรับชุดของข้อสอบร่วม k ข้อ

$$b_{y_{ci}} = \alpha b_{x_{ci}} + \beta$$

$$a_{y_{ci}} = \frac{a_{x_{ci}}}{\alpha}$$

$$c_{y_{ci}} = c_{x_{ci}}$$

การคำนวณค่าคงที่ $\alpha$ และ $\beta$ ได้มาจากการทำให้ฟังก์ชัน F มีค่าต่ำสุด

$$F = \frac{1}{N} \sum_{a=1}^{N} (T_{xa} - T_{ya})^2$$

เมื่อ N เป็นจำนวนข้อสอบ และ F เป็นฟังก์ชันแสดงความแตกต่างระหว่างคะแนนจริงที่ได้จากแบบทดสอบทั้งสองฉบับของผู้สอบที่มีความสามารถเดียวกัน การคำนวณค่าคงที่ $\alpha$ และ $\beta$ ใช้กระบวนการกระทำซ้ำหลายรอบจนได้ค่าที่ดีที่สุด

**4. การประยุกต์ใช้โปรแกรม R สำหรับการปรับเทียบคะแนนตามทฤษฎีการตอบข้อสอบ**

โปรแกรม R เป็นทั้งภาษาและสภาพแวดล้อมสำหรับการคำนวณเชิงสถิติ ดังนั้นการเริ่มทำงานบนสภาพแวดล้อมของ R จึงมีลักษณะเหมือนกับการสร้างพื้นที่ว่าง ๆ ที่มีอุปกรณ์อำนวยความสะดวกครบครันพร้อมสำหรับทำงาน ที่เรียกว่า working space การเก็บข้อมูลต่าง ๆ ภายใน working space จะถูกเรียกว่า วัตถุ (Object) ซึ่งมีนัยกว้างกว่าคำว่าตัวแปร (Variable) ที่ใช้ในแวดวงการวิจัย คำว่า วัตถุ ในสภาพแวดล้อมของ R จึงเกี่ยวข้องกับคำ 2 คำ ซึ่งระบุคุณสมบัติของวัตถุ คือ 1) ชนิดของข้อมูล ได้แก่ อักขระ ตัวเลข จำนวนเต็ม ตรรกะ และจำนวนเชิงซ้อน และ 2) โครงสร้างข้อมูล ได้แก่ เวกเตอร์ ลิสต์ เมตริกซ์ อะเรย์ แฟคเตอร์ และเดต้าเฟรม ดังนั้นการนำข้อมูลจากโปรแกรมสำเร็จรูป เช่น SPSS หรือ Microsoft Excel ข้อมูลเหล่านั้นจะถูกเก็บเป็นวัตถุโดยมีโครงสร้างและชนิดของข้อมูลตามไฟล์ต้นฉบับ

การวิเคราะห์ข้อมูลสำหรับการปรับเทียบคะแนนจำเป็นต้องใช้โปรแกรมคอมพิวเตอร์สำหรับการประมาณค่าพารามิเตอร์ข้อสอบและพารามิเตอร์ผู้สอบตามแบบจำลองของทฤษฎีการตอบข้อสอบที่เลือกใช้เพื่อคำนวณเมทริกซ์ความแปรปรวนร่วมของการประมาณค่าพารามิเตอร์ ตัวอย่างของโปรแกรมคอมพิวเตอร์สำหรับการวิเคราะห์ผลการตอบข้อสอบตามทฤษฎีการตอบข้อสอบที่จำหน่ายเชิงพาณิชย์โดย Vector Psychometric Group ได้แก่ โปรแกรม flexMIRT และ โปรแกรม IRTPRO (Han & Paek, 2014) หรือโปรแกรม R ซึ่งเป็นซอฟต์แวร์ที่เปิดให้ใช้โดยไม่เสียค่าใช้จ่าย และจำเป็นต้องติดตั้งแพ็คเกจ ltm (Rizopoulos, 2006) หรือ แพ็คเกจ mirt (Chalmers, 2012) เพิ่มเติมเพื่อใช้ในการวิเคราะห์ เนื้อหาในส่วนนี้ประกอบด้วย 1) การเตรียมข้อมูลสำหรับการวิเคราะห์เพื่อปรับเทียบคะแนนด้วยโปรแกรม R 2) การเขียนคำสั่งภาษา R สำหรับวิเคราะห์ปรับเทียบคะแนนตามทฤษฎีการตอบข้อสอบ และ 3) การแปลผลการปรับเทียบคะแนนตามทฤษฎีการตอบข้อสอบ

**4.1 การเตรียมข้อมูลสำหรับการวิเคราะห์เพื่อปรับเทียบคะแนนด้วยโปรแกรม R**

ขั้นการเตรียมข้อมูลผู้วิเคราะห์ต้องเตรียมข้อมูลผลการตอบข้อสอบของแบบทดสอบแต่ละฉบับด้วยโปรแกรมตารางการคำนวณหรือโปรแกรมไมโครซอฟต์เอ็กเซล ซึ่งใช้การตั้งชื่อตัวแปรในแถวแรกของแต่ละคอลัมน์ ดังภาพที่ 4 และ 5

| ID | Item1 | Item2 | Item3 | Item4 | Item5 | Item6 | Anchor1 | Anchor2 |
|---|---|---|---|---|---|---|---|---|
| keys | คีย์ข้อ 1 | คีย์ข้อ 2 | คีย์ข้อ 3 | คีย์ข้อ 4 | คีย์ข้อ 5 | คีย์ข้อ 6 | คีย์ข้อร่วม 1 | คีย์ข้อร่วม 2 |
| P001 | | | | | | | | |
| P002 | | | | | | | | |
| P003 | | | | | | | | |

ผลการตอบข้อสอบ

ไฟล์ข้อมูล ผลการตอบแบบทดสอบ X จากกลุ่มผู้สอบ P (ชื่อไฟล์ DataX.csv)

**ภาพที่ 4** แบบแผนการเตรียมข้อมูลของแบบทดสอบฉบับ X สำหรับการปรับเทียบคะแนนแบบ NEAT

| ID | Item7 | Item8 | Item9 | Item10 | Item11 | Item12 | Anchor1 | Anchor2 |
|---|---|---|---|---|---|---|---|---|
| keys | คีย์ข้อ 7 | คีย์ข้อ 8 | คีย์ข้อ 9 | คีย์ข้อ 10 | คีย์ข้อ 11 | คีย์ข้อ 12 | คีย์ข้อร่วม 1 | คีย์ข้อร่วม 2 |
| Q001 | | | | | | | | |
| Q002 | | | | | | | | |
| Q003 | | | | | | | | |

ผลการตอบข้อสอบ

ไฟล์ข้อมูล ผลการตอบแบบทดสอบ Y จากกลุ่มผู้สอบ Q (ชื่อไฟล์ DataY.csv)

**ภาพที่** 5 แบบแผนการเตรียมข้อมูลของแบบทดสอบฉบับ Y สำหรับการปรับเทียบคะแนนแบบ NEAT

จากแบบแผนการคีย์ข้อมูลจะสังเกตว่า การตั้งชื่อตัวแปรสำหรับเก็บผลการตอบรายข้อของแบบทดสอบแต่ละฉบับจะตั้งชื่อไม่เหมือนกัน ส่วนข้อสอบร่วมจะตั้งชื่อเหมือนกัน เนื่องจากในการวิเคราะห์ข้อมูลจะต้องเชื่อมโยงข้อมูลของแบบทดสอบทั้งสองฉบับเข้าด้วยกันเพื่อนำมาวิเคราะห์พร้อมกัน การตั้งชื่อตัวแปรไม่ซ้ำกันจะเป็นการบอกคอมพิวเตอร์ว่าเป็นข้อสอบจากแบบทดสอบคนละฉบับกัน โดยผู้เขียนการเก็บไฟล์ข้อมูลไว้ที่ ไดร์ฟ C ซับไดเร็คทอรี equateIRTdata ซึ่งเขียนเป็น path ที่อยู่ของไฟล์ได้ดังนี้ “C:/equateIRTdata”

เมื่อเตรียมไฟล์ข้อมูลดังภาพ 4 และบันทึกไว้ที่ “C:/equateIRTdata” เรียบร้อย จึงกำหนดไดเร็คทอรีการทำงาน (working directory) ให้ตรงกับไดเร็คทอรีหลักที่ใช้เก็บข้อมูลซึ่งมีเส้นทางอยู่ที่ “C:/equateIRT” โดยใช้คำสั่งกำหนดไดเร็คทอรีการทำงานและนำไฟล์ข้อมูลเข้าสู่โปรแกรม R ด้วยคำสั่ง ต่อไปนี้

```
setwd(“C:/equateIRTdata”)          #กำหนดไดเร็คทอรีการทำงาน
#อ่านข้อมูลจากไฟล์ DataX.csv และ DataY.csv ไปบันทึกที่วัตถุ DataX และ DataY ตามลำดับ โดยระบุว่ามีการเก็บชื่อตัวแปรในบรรทัดแรกของไฟล์ (header = TRUE)
DataX <- read.csv("DataX.csv", header = TRUE)
DataY <- read.csv("DataY.csv", header = TRUE)
```

เมื่อนำข้อมูลเข้าสู่ R เรียบร้อยแล้ว ขั้นตอนต่อไปที่ต้องดำเนินการ คือ การตรวจผลการตอบผลการตอบข้อสอบ ซึ่งดำเนินการโดยใช้ แพ็คเกจ CTT หลังจากนั้น เมื่อได้ผลการตอบข้อสอบแล้วจึงใช้ แพ็คเกจ ltm ในการประมาณค่าพารามิเตอร์ข้อสอบเพื่อนำไปใช้ในการปรับเทียบคะแนนต่อไป การตรวจผลการตอบข้อสอบโดยใช้ แพ็คเกจ CTT มีคำสั่งดังนี้

**4.2 การเขียนคำสั่งภาษา R สำหรับวิเคราะห์ปรับเทียบคะแนนตามทฤษฎีการตอบข้อสอบ**

| ลำดับ | คำสั่ง | คำอธิบาย |
|---|---|---|
| 1 | install.packages("CTT")<br>library(CTT) | ติดตั้งแพ็คเกจ CTT<br>เรียกแพ็คเกจ CTT เพื่อใช้งาน |
| 2 | respX <- DataX[ -1 , ]<br>keyX <- DataX[ 1, ]<br>scoreX <- score(respX, keyX, output.scored=TRUE) | นำข้อมูลที่ไม่ใช่แถวแรกเก็บใน respX<br>นำข้อมูลที่เป็นแถวแรกเก็บใน keyX<br>ตรวจให้ข้อสอบฉบับXและบันทึกผลเป็นคะแนน [ 0/1] ไว้ที่วัตถุชื่อว่า scoreX |
| 3 | respY <- DataY[ -1 , ]<br>keyY<- DataY[ 1, ]<br>scoreY <- score(respY, keyY, output.scored=TRUE) | นำข้อมูลที่ไม่ใช่แถวแรกเก็บใน respX<br>นำข้อมูลที่เป็นแถวแรกเก็บใน keyX<br>ตรวจให้ข้อสอบฉบับ Y และบันทึกผลเป็นคะแนน [ 0/1] ไว้ที่วัตถุชื่อว่า scoreY |
| 4 | install.packages("ltm")<br>library(ltm) | ติดตั้งแพ็คเกจ ltm<br>เรียกแพ็คเกจ ltm เพื่อใช้งาน |
| 5.1 | one.pl.X<-rasch(scoreX$scored)<br>one.pl.Y<-rasch(scoreY$scored) | ประมาณค่าพารามิเตอร์ข้อสอบของแบบทดสอบ X และ Y โดยโมเดล IRT 1 พารามิเตอร์ (1PL) และบันทึกไว้ที่วัตถุ one.pl.X และ one.pl.Y |
| 5.2 | two.pl.X<-ltm(scoreX$scored~ z1, IRT.param=TRUE)<br>two.pl.Y<-ltm(scoreY$scored~ z1, IRT.param=TRUE) | ประมาณค่าพารามิเตอร์ข้อสอบของแบบทดสอบ X และ Y โดยโมเดล IRT 2 พารามิเตอร์ (2PL) และบันทึกไว้ที่วัตถุ two.pl.X และ two.pl.Y |
| 5.3 | three.pl.X<-tpm(scoreX$scored, IRT.param=TRUE)<br>three.pl.Y<-tpm(scoreY$scored, IRT.param=TRUE) | ประมาณค่าพารามิเตอร์ข้อสอบของแบบทดสอบ X และ Y โดยโมเดล IRT 3 พารามิเตอร์ (3PL) และบันทึกไว้ที่วัตถุ three.pl.X และ three.pl.Y |

แพ็กเกจ equateIRT มีฟังก์ชันในการนำเข้าข้อมูลจากโปรแกรมสำหรับการประมาณค่าพารามิเตอร์ข้อสอบ ในบทความนี้ขอนำเสนอตัวอย่างการนำเข้าพารามิเตอร์ข้อสอบและเมทริกซ์ความแปรปรวนร่วมของข้อสอบที่เป็นผลจากการวิเคราะห์โดยใช้แพ็กเกจ ltm ตัวอย่างที่นำเสนอในลำดับต่อไปเป็นการนำผลการวิเคราะห์ ข้อที่ 5.2 มาใช้ซึ่งเป็นการประมาณค่าพารามิเตอร์ข้อสอบของแบบทดสอบ X และ Y ด้วยโมเดล IRT แบบ 2 พารามิเตอร์

| ลำดับ | คำสั่ง | คำอธิบาย |
|---|---|---|
| 6 | install.packages("equateIRT")<br>library("equateIRT") | ติดตั้งแพ็คเกจ equateIRT<br>เรียกแพ็คเกจ equateIRT เพื่อใช้งาน |
| 7 | est.mod2pl.X <- import.ltm(two.pl.X, display = TRUE) | นำเข้าพารามิเตอร์ข้อสอบและเมทริกซ์ความแปรปรวนร่วมของแบบทดสอบ X จากผลการประมาณค่าพารามิเตอร์ข้อสอบด้วยแพ็คเกจ ltm และบันทึกผล ไว้ที่วัตถุชื่อว่า est.mod2pl.X |
| 8 | est.mod2pl.Y <- import.ltm(two.pl.Y, display = TRUE) | นำเข้าพารามิเตอร์ข้อสอบและเมทริกซ์ความแปรปรวนร่วมของแบบทดสอบ Y จากผลการประมาณค่าพารามิเตอร์ข้อสอบด้วยแพ็คเกจ ltm และบันทึกผล ไว้ที่วัตถุชื่อว่า est.mod2pl.Y |
| 9 | ocefXY <-list(est.mod2pl.X$coef , est.mod2pl.Y$coef)<br>varXY <- list(est.mod2pl.X$var, est.mod2pl.Y$var) | จัดเตรียมข้อมูลพารามิเตอร์ข้อสอบ ($coef) และเมทริกซ์ความแปรปรวนร่วม ($var) ให้อยู่ในวัตถุประเภท list |
| 10 | test <- paste("test", 1:2, sep = "") | สร้าง vector เพื่อระบุว่าแบบทดสอบที่ใช้ปรับเทียบคะแนนในตัวอย่างมี ฉบับคือ 2 test และ 1test 1 ใช้) 2แทน X และ แทน 2 Y เพื่อความสะดวกในการรันลำดับอัตโนมัติในกรณีที่มีแบบทดสอบมากกว่า (ฉบับ 2 |
| 11 | linkp(ocefXY)<br>## output ###<br>[,1] [,2]<br>[1,] 20 10<br>[2,] 10 20 | ตรวจสอบข้อสอบร่วมระหว่างแบบทดสอบจาก output เป็นเมทริกซ์ 2x2 การอ่านเมทริกซ์ เส้นแนวทแยงมุมของเมทริกซ์จะระบุจำนวนข้อสอบของแต่ละฟอร์ม ในตัวอย่างคือ แบบทดสอบ X และ Y มีความยาวฉบับละ ข้อ และองค์ประกอบของเมทริกซ์จะระบุจำนวน 20 ข้อสอบร่วมระหว่างแบบฟอร์ม จากตัวอย่างมีข้อสอบร่วมจำนวน ข้อ 10 |
| 12 | mod2plXY <- modIRT(coef = ocefXY,<br>var = varXY,<br>names = test, digits = 4) | เรียกใช้คำสั่ง modIRT เพื่อจัดกระทำข้อมูลค่าพารามิเตอร์ข้อสอบ (coef) และ เมทริกซ์ความแปรปรวนร่วม (var) ให้อยู่ในรูปแบบที่สามารถนำมาวิเคราะห์ได้ โดยนำข้อมูลที่เก็บไว้ในวัตถุชื่อ ocefXY และ varXY จากข้อ 9 มาใช้ |
| 13 | eqT.1xy <- direc(mods = mod2plXY, which=c(1,(2,<br>method = "mean-sigma")<br>eqT2.xy <- direc(mods = mod2plXY, which=c(1,(2,<br>method = "Stocking-Lord") | ปรับเทียบคะแนนของแบบทดสอบ X (new Form) ไปยังแบบทดสอบ Y (old Form) ด้วยวิธี mean-sigma และบันทึก ผลไว้ที่ตัวแปร eqT.1xy และ eqT.2xy ปรับเทียบคะแนนด้วยวิธีการของ Stocking-Lord |
| 14 | summary(eqT.1xy)<br>## output ###<br>Link: test1.test2<br>Method: mean-sigma<br>Equating coefficients:<br>Estimate StdErr<br>A 1.08509 0.049425<br>B -0.17894 0.035983<br><br>summary(eqT.2xy)<br>## output ###<br>Link: test.1test 2<br>Method: Stocking-Lord<br>Equating coefficients:<br>Estimate StdErr<br>A 024722.0 01681.1<br>B -025384.0 13719.0 | แสดงผลการวิเคราะห์การปรับเทียบคะแนน |
| 15 | score(eqT1.xy, method = "TSE", se = TRUE) | แสดงคะแนนหลังการปรับเทียบด้วยวิธี true-score equating (TSE) ต้องต้องการให้เทียบเป็นคะแนนสังเกตได้สามารถปรับ method ให้เห็น “OBE” ซึ่งหมายถึง observed-score equating |

จากคำสั่งด้านบนผู้เขียนได้ยกตัวอย่าง การปรับเทียบคะแบบตามทฤษฎีตอบข้อสอบซึ่งคำนวณหาค่าสัมประสิทธิ์การปรับเทียบพารามิเตอร์เพียง 2 วิธี คือ วิธี mean-sigma และวิธี Stocking-Lord เพื่อนำเสนอให้ผู้อ่านเห็นตัวอย่างการใช้งานและวิธีการพิจารณาคุณภาพในการปรับเทียบคะแนนซึ่งจะกล่าวในส่วนต่อไป ผู้อ่านสามารถเลือกใช้วิธีอื่นได้โดยเปลี่ยนค่าอาร์กิวเมนต์ method ของฟังก์ชัน direc ซึ่งแสดงเป็นตัวหนา

direc(mods = mod2plXY, which=c(1,2), **method** = "**mean-sigma**")

เป็นวิธีการอื่น ดังต่อไปนี้ "mean-mean", "mean-gmean", "mean-sigma", "Haebara" หรือ "Stocking-Lord" รายละเอียดของวิธีการแต่ละแบบศึกษาได้จาก Battauz (2015)

**4.3 การแปลผลการปรับเทียบคะแนนตามทฤษฎีการตอบข้อสอบ**

ผลการวิเคราะห์การปรับเทียบคะแนนด้วยวิธี mean-sigma พบว่า ค่าคงที่ในการปรับเทียบพารามิเตอร์อำนาจจำแนก A เท่ากับ 1.085 และมีความคลาดเคลื่อนมาตรฐาน 0.049 และค่าคงที่ในการปรับเทียบพารามิเตอร์ความยาก B เท่ากับ -0.179 และมีความคลาดเคลื่อนมาตรฐาน 0.036

ผลการวิเคราะห์การปรับเทียบคะแนนด้วยวิธี Stocking-Lord พบว่า ค่าคงที่ในการปรับเทียบพารามิเตอร์อำนาจจำแนก A เท่ากับ 1.017 และมีความคลาดเคลื่อนมาตรฐาน 0.025 และค่าคงที่ในการปรับเทียบพารามิเตอร์ความยาก B เท่ากับ -0.137 และมีความคลาดเคลื่อนมาตรฐาน 0.025

ขั้นตอนต่อไปเป็นการนำค่าคงที่ในการปรับเทียบไปแทนค่าในสูตรดังที่กล่าวมาแล้ว เพื่อคำนวณพารามิเตอร์ข้อสอบของแบบทดสอบ X โดยเทียบกับแบบทดสอบ Y และแปลงคะแนนจากแบบทดสอบ X เมื่อเทียบเคียงไปยังแบบทดสอบ Y ซึ่งไม่ต้องดำเนินการคำนวณเองด้วยมือ สามารถสั่งให้โปรแกรม R คำนวณโดยใช้คำสั่งในลำดับที่ 15 ผลลัพธ์ที่ได้ดัง ตารางที่ 1

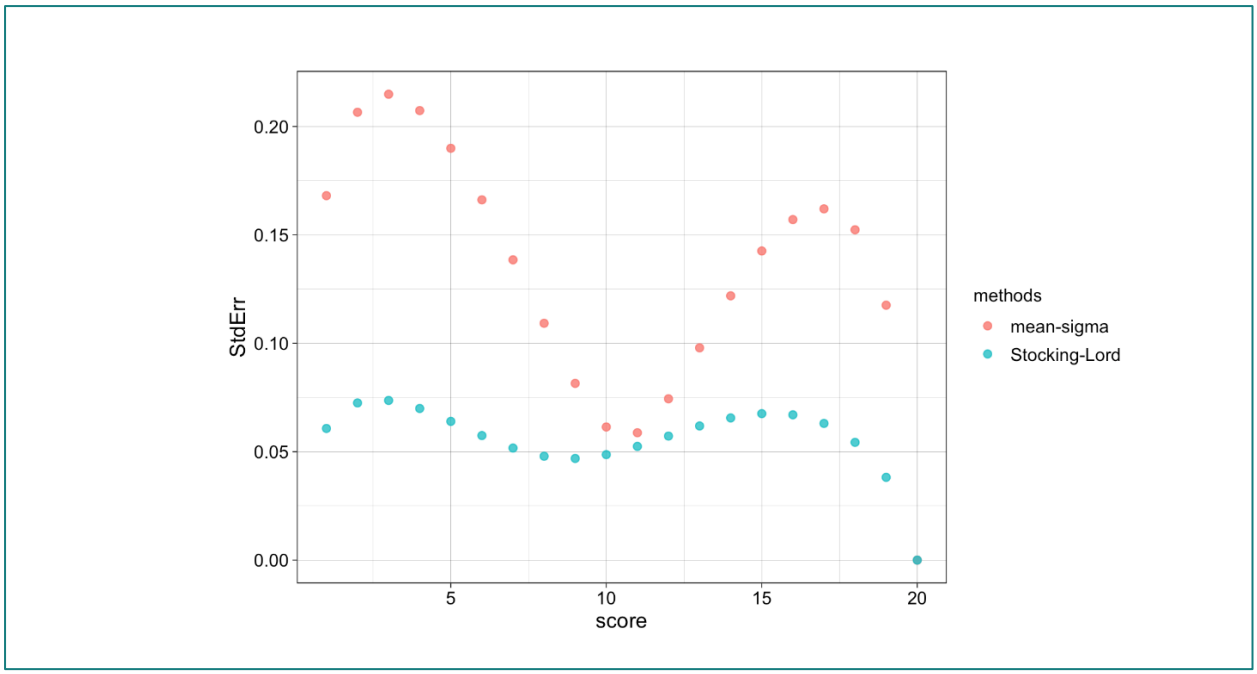
การแปลผลตารางที่ 1 การปรับเทียบคะแนน สำหรับวิธี true-score equating จะมีการรายงานค่าความสามารถ (theta) ในคอลัมน์ที่ 1 สำหรับคอลัมน์ที่ 2 คือคะแนนในฟอร์มที่เลือกใช้เป็นฐานที่ต้องการนำคะแนนไปปรับเทียบ (test2 หรือ แบบทดสอบ Y) คอลัมน์ที่ 3 คือ คะแนนของแบบทดสอบ X ที่ปรับเทียบไปยังแบบทดสอบ Y (test1.as.test2) และคอลัมน์ที่ 4 คือ ความคลาดเคลื่อนมาตรฐานจากการปรับเทียบคะแนน (StdErr) ความคลาดเคลื่อนมาตรฐานที่มีค่าต่ำจะสะท้อนความถูกต้องของการปรับเทียบคะแนน ตัวอย่างการแปลผลข้อมูลในแถวที่หนึ่ง ผู้สอบที่มีความสามารถ -2.79 ทำข้อสอบในแบบทดสอบ X ได้ 1 คะแนน เมื่อนำผลคะแนนจากแบบทดสอบ X ที่ได้ 1 คะแนนไปปรับเทียบ จะเทียบได้กับคะแนนจากแบบทดสอบ Y 1.64 คะแนน โดยมีความคลาดเคลื่อนมาตรฐานในการปรับเทียบ เท่ากับ 0.168

สำหรับวิธี observed-score equating จะรายงานเฉพาะคะแนนสังเกตได้เท่านั้น ตัวอย่างการแปลผลข้อมูลในแถวที่หนึ่ง ผู้สอบทำข้อสอบในแบบทดสอบ X ได้ 0 คะแนน จะเทียบได้กับคะแนนจากแบบทดสอบ Y 0.582 คะแนน โดยมีความคลาดเคลื่อนมาตรฐานในการปรับเทียบ เท่ากับ 0.079

**ตารางที่ 1** ผลการการปรับเทียบคะแนนโดยใช้วิธี true-score equating (ซ้าย) และผลการปรับเทียบคะแนนโดยใช้วิธี observed-score equating (ขวา)

| true-score equating | | | | observed-score equating | | |
|---|---|---|---|---|---|---|
| theta | test2 | test1.as.test2 | StdErr | test2 | test1.as.test2 | StdErr |
| -2.792 | 1 | 1.644 | 0.168 | 0 | 0.582 | 0.079 |
| -2.021 | 2 | 2.859 | 0.207 | 1 | 1.731 | 0.145 |
| -1.543 | 3 | 3.948 | 0.215 | 2 | 2.837 | 0.177 |
| -1.182 | 4 | 4.966 | 0.207 | 3 | 3.891 | 0.186 |
| -0.883 | 5 | 5.939 | 0.190 | 4 | 4.904 | 0.182 |
| -0.621 | 6 | 6.881 | 0.166 | 5 | 5.885 | 0.170 |
| -0.383 | 7 | 7.800 | 0.138 | 6 | 6.842 | 0.152 |
| -0.160 | 8 | 8.704 | 0.109 | 7 | 7.780 | 0.131 |
| 0.054 | 9 | 9.597 | 0.082 | 8 | 8.703 | 0.108 |
| 0.264 | 10 | 10.483 | 0.061 | 9 | 9.615 | 0.086 |
| 0.474 | 11 | 11.365 | 0.059 | 10 | 10.517 | 0.068 |
| 0.687 | 12 | 12.246 | 0.074 | 11 | 11.419 | 0.055 |
| 0.910 | 13 | 13.129 | 0.098 | 12 | 12.322 | 0.059 |
| 1.148 | 14 | 14.019 | 0.122 | 13 | 13.227 | 0.073 |
| 1.409 | 15 | 14.918 | 0.143 | 14 | 14.136 | 0.091 |
| 1.707 | 16 | 15.832 | 0.157 | 15 | 15.051 | 0.108 |
| 2.068 | 17 | 16.769 | 0.162 | 16 | 15.975 | 0.123 |
| 2.546 | 18 | 17.742 | 0.152 | 17 | 16.912 | 0.133 |
| 3.319 | 19 | 18.775 | 0.118 | 18 | 17.867 | 0.137 |
| 42.162 | 20 | 20.000 | 0.000 | 19 | 18.846 | 0.131 |
| | | | | 20 | 19.861 | 0.108 |

การตรวจสอบความถูกต้องของการปรับเทียบคะแนน สามารถพิจารณาได้จากความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนในแต่ละวิธีการที่ใช้ ซึ่งได้ยกตัวอย่างความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนที่ดำเนินการด้วยวิธี mean-sigma และวิธี Stocking-Lord ดังภาพที่ 5




**ภาพที่ 6** การเปรียบเทียบความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนน

จากภาพที่ 6 พบว่า การปรับเทียบคะแนนด้วย Stocking-Lord มีความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนน้อยกว่าวิธี mean-sigma ในทุกช่วงคะแนน ซึ่งเป็นหลักฐานที่ใช้สนับสนุนการตัดสินใจเลือกวิธีการปรับเทียบคะแนนที่มีความคลาดเคลื่อนต่ำ

## บทสรุป

โปรแกรม R เป็นสภาพแวดล้อมที่เอื้อต่อการคำนวณและการวิเคราะห์ทางสถิติตั้งแต่ขั้นพื้นฐานจนถึงขั้นสูงดังตัวอย่างที่ได้นำเสนอ การใช้งานโปรแกรม R ในครั้งแรกอาจรู้สึกยากเนื่องจากผู้ใช้ส่วนมากคุ้นเคยกับกับใช้งานโปรแกรมประเภทคลิกและรัน แต่โปรแกรม R เป็นโปรแกรมที่ต้องเขียนคำสั่งเพื่อเรียกใช้แพ็กเกจและฟังก์ชันต่าง ๆ สำหรับประมวลผล การใช้งานโปรแกรม R จึงมีความยืดหยุ่นสูงมาก เนื่องจากเปิดโอกาสให้ผู้ใช้สามารถปรับค่าอาร์กิวเมนต์ต่าง ๆ ในฟังก์ชันได้หลากหลาย รวมทั้งสามารถนำโค้ดของฟังก์ชันมาใช้ ซึ่งมีนักวิชาการหลากหลายศาสตร์ร่วมกันพัฒนา ปรับปรุงแก้ไข เพื่อให้วิเคราะห์งานเฉพาะได้ตรงความต้องการของผู้ใช้มากยิ่งขึ้น นอกจากนี้ยังมีชุมชนทางวิชาการ อาทิ Journal of Statistical Software เป็นวารสารที่นำเสนอบทความเกี่ยวซอฟต์แวร์ทางสถิติที่พัฒนาขึ้นใหม่ ๆ มีเว็บไซต์ https://stackoverflow.com ที่เป็นแหล่งถามตอบปัญหาให้ผู้ใช้มือใหม่เข้าไปเรียนรู้เพิ่มเติม และเนื่องจาก R เป็นโปรแกรมสาธารณะสามารถใช้ได้โดยไม่เสียค่าใช้จ่าย จึงได้รับความนิยมอย่างแพร่หลายโดยเฉพาะในต่างประเทศที่มีกฎหมายด้านลิขสิทธิ์ที่เข้มงวด และมีบทลงโทษรุนแรง การประยุกต์ใช้งานโปรแกรม R สำหรับการวิจัยและการทดสอบทางการศึกษาในประเทศไทยจึงเป็นเรื่องน่าสนใจและควรสนับสนุนให้มีการใช้เป็นวงกว้าง บทความนี้ถือเป็นส่วนหนึ่งในการสนับสนุนและแบ่งบันความรู้ในสังคมวิชาการของประเทศไทยในการประยุกต์ใช้โปรแกรม R ทางด้านการวัดและการทดสอบทางการศึกษา

## เอกสารอ้างอิง

Battauz, M. (2015). equateIRT: An R package for IRT test equating. *Journal of statistical software, 68*(7), 1-22. https://doi.org/10.18637/jss.v068.i07

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of statistical software, 48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06

Dorans, N. J. & Holland, P.W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, 37: 281-306. https://doi.org/10.1111/j.1745-3984.2000.tb01088.x

Dorans, N. J., & Puhan, G. (2017). Contributions to score linking theory and practice. In R. E. Bennett & M. von Davier (Eds.), *Advancing Human Assessment: The Methodological, Psychological and Policy Contributions of ETS* (pp. 79-132). Springer International Publishing. https://doi.org/10.1007/978-3-319-58689-2_4

González, J., & Wiberg, M. (2017). Item response theory equating. In *Applying Test Equating Methods: Using R* (pp. 111-136). Springer International Publishing. https://doi.org/10.1007/978-3-319-51824-4_5

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: principles and applications. *Springer Netherlands*. https://books.google.co.th/books?id=jKFMUFI-e1UC

Han, K.T., & Paek, I. (2014). A Review of commercial software packages for multidimensional IRT modeling. *Applied Psychological Measurement, 38*(6), 486-498. https://doi.org/10.1177/0146621614536770

Kolen, M., & Brennan, R. (2014). *Test equating, scaling, and linking. Methods and practices*. (3rd revised ed.) https://doi.org/10.1007/978-1-4939-0317-7

Livingston, S. A., & Kim, S. (2010). Random-groups equating with samples of 50 to 400 test takers. *Journal of Educational Measurement, 47*(2), 175-185. http://www.jstor.org/stable/20778946

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and Item response analysis. *Journal of statistical software, 17*(5), 1-25. https://doi.org/10.18637/jss.v017.i05