หอสมุดกลาง สถาบันวิทยบริการ จุฬาลงกรณ์มหาวิทยาลัย

บทที่ 2

วรรณคดีที่เกี่ยวข้อง

คำว่า Test Equating หรือการปรับเทียบคะแนนระหว่างแบบสอบเป็นศัพท์เฉพาะที่นักจิตมิติ(Psychometricians) ใช้ในกระบวนการวัดและประเมินผลหรือการทดสอบ ในช่วงแรกยังไม่เป็นที่นิยมในหมู่นักวัดผลทั่ว ๆ ไปเท่าใดนัก จะมีการศึกษากันก็เฉพาะในกลุ่มของนักจิตมิติเท่านั้น(Brennan, 1987) ในช่วงนั้นตำราหรือเอกสารที่เกี่ยวข้องก็ยังมีเพียงน้อย ตำราหลักที่สำคัญได้แก่ Scales, Norms, and Equivalent Scores ของ Angoff ที่พิมพ์ออกมาในปี 1971 แต่ต่อมาในราวปี 1980 เป็นต้นมาเอกสารที่เกี่ยวข้องกับ เรื่องนี้ได้เพิ่มขึ้นอย่างผิดสังเกต และนักวัดผลตามที่ต่างๆ ได้เล็งเห็นถึงความสำคัญของการ ปรับเทียบคะแนนเพราะการสอบได้เกิดขึ้นอย่างกว้างขวางและมากมายกับกลุ่มผู้สอบจำนวน มากจึงต้องใช้ แบบสอบหลาย ๆ ฉบับ และนักวัดผลยอมรับว่าคะแนนจากการสอบแต่ละฉบับไม่ สามารถที่จะนำมาเปรียบเทียบกันได้โดยตรง จำเป็นต้องใช้กระบวนการปรับเทียบคะแนน การปรับเทียบคะแนนระหว่างแบบสอบได้มีการศึกษาและพัฒนาขึ้นมาตามลำดับ แรกที่เดียวจะมีการปรับเทียบคะแนนเฉพาะตามแนวทฤษฎีการวัดแบบดั้งเดิม ซึ่งได้แก่ การปรับเทียบแบบอีควิเปอร์เซ็นไตล์และแบบเส้นตรงเท่านั้น แต่ต่อมาเมื่อได้มีทฤษฎีการตอบสนองรายข้อ หรือ IRT เกิดขึ้นมา จึงได้มีการปรับเทียบคะแนนตามแนวทฤษฎีนี้ขึ้นมาอีก ในการวิจัยครั้งนี้ผู้วิจัยสนใจศึกษาการปรับเทียบคะแนนระหว่างแบบสอบตามแนว IRT ดังนั้นการเสนอวรรณคดีที่เกี่ยวข้องจึงเน้นหนักไปทางทฤษฎีการตอบสนองรายข้อ ซึ่งจะเสนอดังนี้

1) แนวคิดเชิงทฤษฎีของการปรับเทียบคะแนนระหว่างแบบสอบ
2) แบบแผนการในการปรับเทียบคะแนน
3) การปรับเทียบคะแนนตามแนวทฤษฎีการวัดแบบดั้งเดิม
4) การปรับเทียบคะแนนตามแนวทฤษฎีการตอบสนองรายข้อ
5) ความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนน
6) งานวิจัยที่เกี่ยวข้องกับการปรับเทียบคะแนน

<u>แนวคิดเชิงทฤษฎีของการปรับเทียบคะแนนระหว่างแบบสอบ</u>

แองกอฟ(Angoff, 1984) ได้กล่าวถึงการปรับเทียบคะแนนระหว่างแบบสอบว่า เนื่องจากแบบสอบสองฉบับใดๆยากที่สร้างให้มีความเท่าทียมกันในระดับความยาก จึงจำเป็นต้องมีการปรับเทียบคะแนนระหว่างแบบสอบเกิดขึ้น โดยเป็นระบบการเปลี่ยนสเกลของแบบสอบฉบับหนึ่งให้ไปอยู่ในสเกลของอีกฉบับหนึ่ง หลังจากปรับเปลี่ยนสเกลแล้วแบบสอบทั้งสองจึงมีความเท่าเทียมกัน และการที่จะปรับเปลี่ยนสเกลไปหากันได้ก็ไม่ผิดอะไรกันกับการเปลี่ยนสเกลของอุณหภูมิจาก องศาเซลเซียส ไปเป็นองศาฟาเรนไฮท์ หรือเปลี่ยนหน่วยของ ความยาวจากจากนิ้วไปเป็น เซ็นติเมตร จากแนวคิดลักษณะนี้บอกถึงเงื่อนไขที่สำคัญของการปรับเทียบคะแนน 2 ประการ คือ ประการแรกเครื่องมือทั้งสองจะต้องวัดคุณลักษณะอันเดียวกัน เช่น ความยาวหรืออุณหภูมิ และในแง่ของการวัดทางจิตวิทยาก็ต้องมีความชัดเจน เช่น ไม่สามารถที่จะเปลี่ยนสเกลของแบบสอบทางภาษาไปสู่สเกลของแบบสอบเกี่ยวกับการคำนวณได้ เช่นเดียวกันกับไม่สามารถเปลี่ยนหน่วยของอุณหภูมิไปเป็นหน่วยของความยาวได้ ประการที่สองคือการปรับเทียบคะแนนจะต้องมีความเป็นหนึ่งเดียว(unique) ยกเว้นเฉพาะความคลาดเคลื่อนเชิงสุ่มที่เกี่ยวข้องกับความไม่เที่ยงตรงของข้อมูลเท่านั้น กล่าวคือการปรับเปลี่ยนหน่วยต้องมีความเป็นอิสระจากผู้สอบซึ่งเป็นบุคคลที่ให้ข้อมูลของการปรับเทียบ การทำให้เกิดความเป็นหนึ่งสามารถทำได้โดยใช้แบบสอบคู่ขนาน

ฟลานากาน และลอร์ด(Flanagan, 1951 และ Lord, 1950 อ้างถึงใน Angoff, 1984) ได้ให้ความหมายของคะแนนที่เท่าเทียมกันหรือคะแนนสมมูล(equivalent score) ว่า คะแนนจากแบบสอบสองฉบับคือ ฉบับ X และฉบับ Y (โดยที่แบบสอบทั้งสองวัดในสิ่งเดียวกันและมีความเที่ยงเท่ากัน) จะเป็นคะแนนที่สมมูลกันเมื่อตำแหน่งเปอร์เซ็นไตล์ของกลุ่มผู้สอบทั้งสองอยู่ที่ตำแหน่งเดียวกัน จากความหมายนี้เป็นการนำไปสู่การปรับเทียบคะแนนแบบอีควิเปอร์เซ็นไตล์นั่นเอง ส่วนความหมายของการปรับเทียบคะแนนแบบเส้นตรง อาจกล่าวได้ว่าคะแนนจากแบบสอบ 2 ฉบับจะสมมูลกันก็ต่อเมื่อมีคะแนนมาตรฐานเท่ากัน คือ

$Z_x = Z_y$ หรือ

$$\frac{Y - M_y}{S_y} = \frac{X - M_x}{S_x} \qquad (1)$$

เมื่อจัดรูปสมการนี้ใหม่จะได้ $Y = Ax + B$ เมื่อ

$$A = \frac{S_y}{S_x} \qquad (2)$$

$$B = M_Y - AM_x \qquad (3)$$

นั่นคือ ค่า A เป็นความชันของกราฟที่ปรับคะแนน และ B ก็คือ จุดตัดบนแกน Y การปรับเทียบคะแนนแบบอีควิเปอร์เซ็นไทล์ จะมีค่าเท่ากับแบบเส้นตรงเมื่อการแจงแจงของคะแนนดิบมีความคล้ายคลึงกัน

นอกจากวิธีการปรับเทียบแบบอีควิเปอร์เซ็นไทล์ และแบบเส้นตรงแล้ว ในเรื่องนี้ฟลานากานยังได้เสนอแนะวิธีการปรับเทียบคะแนนเมื่อแบบสอบทั้งสองเป็นแบบสอบคู่ขนานว่าอาจจะใช้วิธี เทียบค่าเฉลี่ยโดยคำนวณค่าเฉลี่ยของคะแนนทั้งสองชุดถ้าความแตกต่างของค่าเฉลี่ยอยู่ภายในขอบเขตความแปรผันเชิงสุ่มแล้ว ถือว่าคะแนนทั้งสองชุดนั้นสามารถเปรียบเทียบกันได้ แต่ถ้าความแตกต่างมีนัยสำคัญให้ใช้วิธีบวกเข้า หรือลบออกเท่ากับจำนวนที่แตกต่างจากคะแนนอีกชุดหนึ่ง และอีกวิธีหนึ่งคือใช้เทคนิคสมการถดถอย ประมาณค่าดีที่สุดของคะแนน จากแบบสอบชุดหนึ่งซึ่งรู้ค่าของอีกชุดหนึ่ง

มโนทัศน์ที่กล่าวมานั้นเป็นเรื่องการปรับเทียบคะแนนระหว่างแบบสอบตามแนวทฤษฎีแบบดั้งเดิม ที่มุ่งเน้นความสำคัญไปที่การสร้างแบบสอบคู่ขนาน โดยมีเงื่อนไขว่าเมื่อแบบสอบเป็นคู่ขนานแล้ว คะแนนที่ได้จากแบบสอบทั้งสองชุดย่อมเป็นคะแนนสมมูลกัน ในเรื่องแบบสอบคู่ขนานกัลลิคเซ็น(Guliliksen, 1950) ได้เสนอแนะการสร้างไว้ดังนี้

1) วิเคราะห์หาความยากและอำนาจจำแนกของข้อสอบเป็นรายข้อ แล้วพล็อตจุดลงกราฟ

2) พิจารณาข้อสอบที่เกาะอยู่ในกลุ่มเดียวกันว่า แต่ละกลุ่มวัดความสามารถด้านใด

3) แยกข้อสอบที่อยู่ในกลุ่มเดียวกันนั้นให้อยู่คนละชุดโดยการสุ่ม

การพัฒนาแบบสอบคู่ขนานโดยวิธีนี้ทำให้สามารถควบคุมความหลากหลายของคำถามของแบบสอบแต่ละชุดได้ แต่วิธีการเช่นนี้ยังมีจุดอ่อนอยู่ 5 ประการ คือ

1) สำหรับข้อสอบข้อเดียวกันความยากของข้อสอบที่วิเคราะห์จากฉบับเริ่มแรกกับฉบับอื่นที่มีข้อสอบนั้นอยู่ จะมีค่าไม่เท่ากัน ทั้งนี้อาจเนื่องจากตำแหน่งของการเรียงข้อสอบและบริบทต่างๆ ในข้อสอบมีผลกระทบต่อค่าความยาก

2) การแบ่งข้อสอบจากฉบับเริ่มแรกไปสู่ฉบับอื่นที่สมมูลกันโดยไม่ทำให้สูญเสียลักษณะการสุ่มเนื้อหาในแต่ละชุดนั้นเป็นเรื่องทำไม่ได้

3) การปรับปรุงแบบสอบชุดหลังทำได้ยาก เนื่องจากจะต้องคงสภาพความยากและจำนวนข้อในแต่ละตอนให้เหมือนฉบับเดิม

4) การดำเนินการสอบที่จัดขึ้นต่างเวลากัน มีความหมายถึงการเปลี่ยนแปลงสถานการณ์การสอบ ซึ่งมีผลกระทบต่อการแปรเปลี่ยนของการแจกแจงคะแนนดิบ จะนำมาเปรียบเทียบกันโดยตรงไม่ได้

5) ความสัมพันธ์ของข้อสอบข้อเดียวกันในแบบสอบต่างชุดมีค่าแตกต่างกัน ซึ่งทำให้ค่าความสอดคล้องภายในเปลี่ยนแปลงไป ผลที่ตามก็คือคะแนนจากการสอบจะมีการแจกแจงไม่เหมือนกัน

ลอร์ด(Lord, 1980) ได้กล่าวถึงการปรับเทียบคะแนนระหว่างแบบสอบว่า วัตถุประสงค์ของการปรับเทียบคะแนนคือเพื่อกำหนดความเท่าเทียมกันระหว่างคะแนนดิบจากแบบสอบสองฉบับ และเนื่องจากวิธีการปรับเทียบคะแนนระหว่างแบบสอบเป็นวิธีการหาหลักฐานเชิงประจักษ์ จึงจำเป็นต้องกำหนดแบบแผนในการเก็บรวบรวมข้อมูลและกฎเกณฑ์ของการแปลงคะแนนจากแบบสอบหนึ่งไปสู่แบบสอบอีกฉบับหนึ่ง ลอร์ดได้กำหนดเงื่อนไขของการปรับเทียบคะแนนซึ่งนักวัดผลหลายคนได้เห็นด้วย ที่ว่าแบบสอบฉบับ X และฉบับ Y จะสามารถนำมาปรับเทียบกันได้ก็ต่อเมื่อแบบสอบทั้งสองมีคุณสมบัติ 4 ประการคือ

1) แบบสอบทั้งสองฉบับจะต้องวัดความสามารถเดียวกัน(same ability) คือแบบสอบทั้งสองวัดในคุณลักษณะเดียวกัน คุณลักษณะนี้อาจเป็นคุณลักษณะแฝง หรือความสามารถหรือทักษะ อย่างใดอย่างหนึ่งก็ได้

2) มีความเสมอภาค(equity) คือเมื่อทุกกลุ่มมีความสามารถเดียวกัน การแจกแจงคะแนนของแบบสอบ Y หลังจากที่มีการแปลงคะแนนแล้วจะมีการแจกแจกเหมือนกับการแจกแจงของคะแนนจากแบบสอบ X

3) ประชากรไม่แปรเปลี่ยน(population invariance) คือการแปลงคะแนนต้องเป็นไปในลักษณะเดียวกันไม่ว่าคะแนนจะมาจากกลุ่มตัวอย่างใดก็ตาม

4) มีความสมมาตร(symmetry) หมายถึงการแปลงคะแนนสามารถแปรเปลี่ยนกลับได้ เช่นการแปลงคะแนนจากฉบับ X ไปยังฉบับ Y มีผลเช่นเดียวกันกับฉบับ Y ไปสู่ฉบับ X

โดยสรุป การปรับเทียบคะแนนระหว่างแบบสอบเป็นกระบวนการทางสถิติที่นำมาใช้ในการปรับคะแนนของแบบสอบสองฉบับที่วัดคุณลักษณะเดียวกันให้อยู่ในสเกลเดียวกัน คะแนนจากแบบสอบทั้งสองฉบับนั้นจึงเปรียบเทียบกันได้อย่างมีความหมาย วิธีการปรับเทียบโดยเน้นที่การสร้างแบบสอบคู่ขนานยังมีจุดอ่อนอยู่มากไม่สามารถเป็นไปตามเงื่อนไขที่กำหนดได้ จึงจำเป็นต้องหาเทคนิควิธีอื่นๆมาใช้ เพื่อให้การปรับเทียบคะแนนบรรลุตามเงื่อนไขเหล่านี้ได้วิธี ดังกล่าวก็คือ การปรับเทียบตามทฤษฎีการตอบสนองรายข้อ

<u>แบบแผนการเก็บรวบรวมข้อมูลในการปรับเทียบคะแนน</u>(Design of Equating)

ในการปรับเทียบคะแนนระหว่างแบบสอบแต่ละครั้งจำเป็นต้องมีแบบแผนในการเก็บรวบรวมข้อมูล ต่อไปนี้เป็นแบบแผนที่นำมาใช้ได้ในการปรับเทียบคะแนน

1.แบบแผนกลุ่มเดียว(Single-group Design) ตามแบบแผนนี้ต้องนำแบบสอบสองฉบับไปสอบกับกลุ่มผู้สอบเพียงกลุ่มเดียว ดังนั้นกลุ่มผู้สอบแต่ละกลุ่มจะได้รับการสอบทั้งสองฉบับ ระดับความยากของแบบสอบไม่มีส่วนเกี่ยวข้องกับระดับความสามารถของผู้สอบ อย่างไรก็ตามตามแบบแผนนี้ความเมื่อยล้าของผู้สอบอาจมีผลต่อการปรับเทียบคะแนน

2. แบบแผนกลุ่มสมมูล(Equivalent-group Design)แบบสอบสองฉบับใด ๆ ที่จะนำคะแนนมาปรับเทียบกันจะนำไปสอบกับกลุ่มผู้สอบต่างกลุ่ม กลุ่มที่เลือกมาทำแบบสอบจะเลือกมาโดยการสุ่ม เมื่อใช้การปรับคะแนนตามแบบแผนนี้สามารถแก้ปัญหาความเมื่อยล้าจากการทำแบบสอบได้ อย่างไรก็ตามเนื่องจากกลุ่มผู้สอบไม่เป็นกลุ่มเดียวกัน อาจจะมีความสามารถแตกต่างกันเล็กน้อย จึงอาจเกิดความลำเอียงเกิดขึ้นกับการปรับเทียบคะแนนได้

3.แบบแผนข้อสอบร่วม(Anchor-test Design) ตามแบบแผนนี้แบบสอบสองฉบับที่นำมาปรับเทียบคะแนนกันจะนำไปสอบกับผู้สอบต่างกลุ่มกันแต่แบบสอบ แต่ละฉบับจะบรรจุชุดของข้อสอบร่วมหรืออาจจะเป็นชุดของข้อสอบร่วมที่แยกออกเป็นฉบับต่างหากก็ได้ แบบสอบร่วมนี้จะนำไปสอบทั้งสองกลุ่มพร้อมกับแบบสอบที่ต้องการนำมาปรับเทียบ ผู้สอบสองกลุ่มไม่จำเป็นต้องมีความเท่าเทียมกัน ตามแบบแผนนี้จะช่วยแก้ปัญหาต่าง ๆ ที่เกิดกับแบบแผนที่ 1 และแบบแผนที่ 2 ได้

จากแบบแผนที่กล่าวมาสามารถนำมาปรับปรุงเพื่อประโยชน์ในการปรับเทียบได้ เช่น แทนที่จะใช้แบบแผนข้อสอบร่วมกับกลุ่มผู้สอบทั้งสองกลุ่ม อาจจะใช้เป็นกลุ่มผู้สอบร่วมแทนได้โดย

ให้กลุ่มผู้สอบกลุ่มหนึ่งทำแบบสอบทั้งสองฉบับ

<u>การปรับเทียบคะแนนตามแนวทฤษฎีการวัดแบบดั้งเดิม</u>

การปรับเทียบคะแนนตามแนวทฤษฎีการวัดแบบดั้งเดิมจากการอธิบายของ แองกอฟ (Angoff,1984) สามารถจัดแบ่งออกได้เป็น 3 ชนิดด้วยกัน คือ 1)การปรับเทียบคะแนนแบบอีควิเปอร์เซ็นไตล์(equipercentile equating) 2)การปรับเทียบคะแนนแบบเส้นตรง (linear equating) และ 3)การปรับเทียบคะแนนแบบถดถอย(regression method)

การปรับเทียบคะแนนแบบอีควิเปอร์เซ็นไตล์ยึดถือนิยามของการปรับเทียบที่ว่า คะแนนจากแบบสอบ X และแบบสอบ Y จะมีความเท่าเทียมกันก็ต่อเมื่อคะแนนทั้งสองอยู่ที่ตำแหน่งเปอร์เซ็นไตล์เดียวกัน และลักษณะการแจกแจงคะแนนของประชากรผู้สอบจะต้องมีลักษณะเหมือนกัน

ขณะที่การปรับเทียบคะแนนแบบอีควิเปอร์เซ็นไตล์ยึดหลักการว่าการแจกแจงของคะแนนที่แปลงแล้วจะต้องมีลักษณะเหมือนกัน สำหรับคะแนนดิบแล้วจะเป็นไปตามเงื่อนไขนี้ได้ยาก จึงจำเป็นต้องใช้วิธีการปรับเทียบคะแนนแบบเส้นตรงแทน ซึ่งเป็นผลให้ความสัมพันธ์ระหว่างคะแนนดิบไม่เป็นเส้นตรง ลักษณะเช่นนี้เกิดได้จากการที่แบบสอบทั้งสองมีความเที่ยงไม่เท่ากันและผู้สอบมีความสามารถแตกต่างกัน ดังนั้นเงื่อนไขของความเสมอภาคจึงไม่บรรลุ ปัญหาต่อมาของการปรับเทียบคะแนนดิบแบบอีควิเปอร์เซ็นไตล์คือการปรับเทียบคะแนนของกลุ่มที่ขึ้นต่อกัน

เมื่อแบบสอบสองฉบับที่นำมาปรับเทียบคะแนนมีค่าความยากใกล้เคียงกัน เช่นการปรับเทียบในแนวระดับ การแจกแจงของคะแนนดิบจะแตกต่างกันเพียงแต่ในสองโมเมนต์แรกเมื่อแบบสอบทั้งสองนำไปสอบกับผู้สอบกลุ่มเดียวกัน ในกรณีเช่นนี้การแปลงคะแนนดิบในเชิงเส้นตรงจะเป็นการยืนยันว่าลักษณะการแจกแจงมีลักษณะเหมือนกัน ดังนั้นคะแนน x และ y จากแบบสอบ X และแบบสอบ Y จึงสามารถนำมาปรับเทียบคะแนนตามแบบเส้นตรงได้ คือ

$$y = ax + b \qquad (4)$$

โดยสัมประสิทธิ์ a และ b กำหนดได้จากความสัมพันธ์ต่อไปนี้

$$\mu_y = a\mu_x + b \qquad (5)$$

$$\sigma_y = a\sigma_x \qquad (6)$$

หอสมุดกลาง สถาบันวิทยบริการ จุฬาลงกรณ์มหาวิทยาลัย

เมื่อ μ_y และ μ_x เป็นค่าเฉลี่ยของคะแนน y และ x ตามลำดับ ขณะที่ σ_y และ σ_x เป็นส่วนเบี่ยงเบนมาตรฐานตามลำดับ ดังนั้นจึงได้

$$Y = \frac{\sigma_y}{\sigma_x} X + \left[\mu_y - \frac{\sigma_y}{\sigma_x}\mu_x\right] \qquad (7)$$

สมการเหล่านี้เป็นกรณีเฉพาะของการปรับเทียบคะแนนแบบอีควิเปอร์เซ็นไทล์ เมื่อข้อตกลงเกี่ยวกับโมเมนต์ได้เป็นไปตามเงื่อนไข

จากปัญหาที่พบกับการใช้วิธีอีควิเปอร์เซ็นไทล์และวิธีแบบเส้นตรง อาจจะมีความจำเป็นต้องใช้วิธีการปรับเทียบแบบถดถอยในการปรับเทียบคะแนนแทน วิธีการนี้อาจยึดหลักที่เป็นไปได้ 2 ประการ คือ

1. เป็นการทำนายคะแนนของแบบสอบหนึ่งจากแบบสอบหนึ่ง

2. เป็นการกำหนดความสัมพันธ์ระหว่างคะแนน 2 ชุด โดยใช้เกณฑ์ภายนอก

ตามหลักการแรกอาจมีปัญหาเกิดขึ้นเนื่องจากในสถานการณ์การถดถอยที่มีตัวแปรอิสระและตัวแปรตามซึ่งไม่มีความสัมพันธ์กันในเชิงสมมาตร ส่วนหลักการที่สองอธิบายได้ว่า เมื่อ $R_x(w|x)$ แทนค่าเกณฑ์ภายนอก w ที่ทำนายจาก x ผ่านสมการถดถอย ในทำนองเดียวกัน ให้ $R_y(w|y)$ แทนค่าของ w ที่ทำนายจาก y ดังนั้นความสัมพันธ์ระหว่าง x และy จึงกำหนดได้ว่า

$$R_x(w|x) = R_y(w|y) \qquad (8)$$

จากนั้นนำไปพล็อตกราฟความสัมพันธ์และสามารถปรับคะแนนจากแบบสอบหนึ่งไปยังอีกแบบสอบหนึ่งได้โดยใช้กราฟนี้

ลอร์ด(Lord, 1980) ได้ชี้ให้เห็นปัญหาของวิธีการนี้ว่า จะไม่เป็นไปตามเงื่อนไขเมื่อคะแนนดิบที่นำมาปรับเทียบได้มาจากแบบสอบที่ไม่คู่ขนานกันและความยากไม่เท่ากัน ในการทำนายคะแนนจากคะแนน x (หรือ y) ตามปกติแล้วจะอยู่บนข้อสมมุติที่ว่าคะแนน x (หรือy) จะต้องวัดมาโดยไม่มีค่าความคลาดเคลื่อน ปัญหาประการที่สองก็คือความสัมพันธ์ระหว่างคะแนน x และ y ที่พบในวิธีการนี้จะมีการเปลี่ยนแปลงจากกลุ่มหนึ่งไปสู่กลุ่มหนึ่งเว้นเสียแต่แบบสอบทั้งสองนั้นมีความสัมพันธ์กับเกณฑ์อย่างเท่ากัน เนื่องจากการที่ได้อภิปรายมาแล้วยึดอยู่บนพื้นฐานที่ว่าแบบสอบทั้งสองจะต้องมีความเที่ยงเท่ากัน ในทางปฏิบัติทำได้ยากมาก ดังนั้นวิธีการปรับเทียบคะแนนแบบถดถอยจึงเป็นไปได้ยาก

การปรับเทียบคะแนนตามแนวทฤษฎีการตอบสนองรายข้อ

ข้อตกลงเบื้องต้น

แฮมเบิลตันและสวามินาธาน(Hambleton และ Swaminathan, 1985)ได้เสนอข้อตกลงเบื้องต้นของ IRT ไว้ 4 ประการ คือ

1) ความเป็นเอกมิติ (unidimentional)

2) ความเป็นอิสระ (local independence)

3) โค้งคุณลักษณะข้อสอบ (item charactoristic curve)

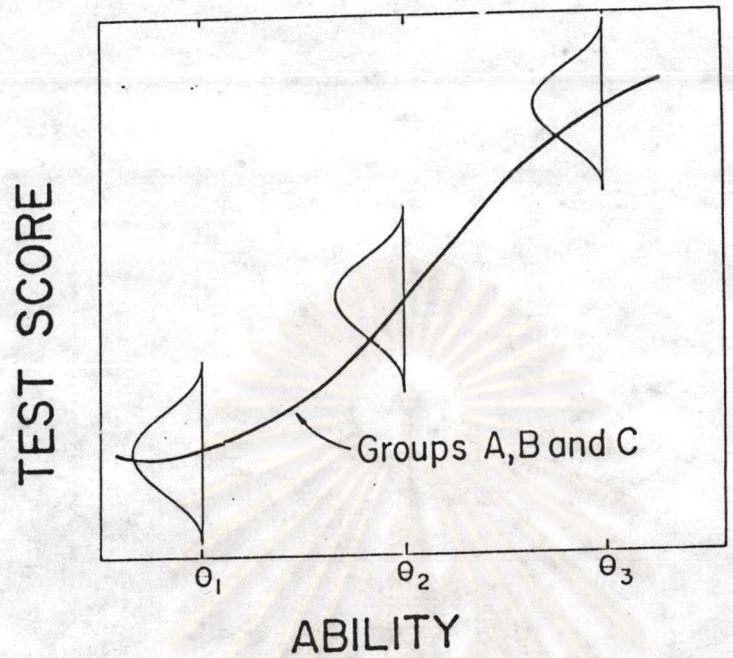
4) ความเร็วในการทำแบบสอบ (speedness)

ความเป็นเอกมิติ

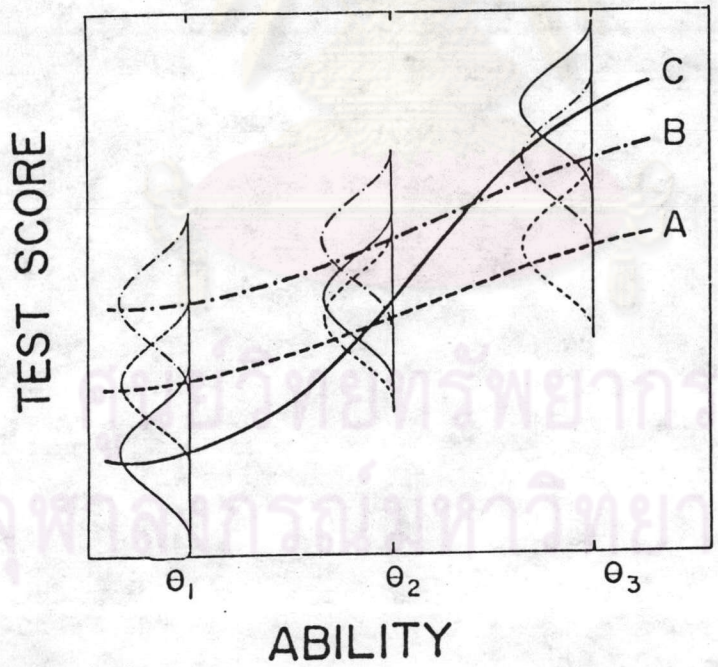
โดยทั่วไปมีข้อสมมุติว่ามีเพียงคุณลักษณะเดียวหรือความสามารถเดียวเท่านั้นที่จำเป็นต่อการอธิบายคะแนนที่ได้จากการตอบของผู้สอบ ตามโมเดลของการตอบสนองรายข้อกำหนดว่าการที่มีความสามารถเดียวหรือคุณลักษณะเดียวคือความเป็นเอกมิติ หรือ unidimentional โดยทั่วไปแล้วข้อตกลงข้อนี้เป็นไปได้ค่อนข้างยาก เนื่องจากมีปัจจัยหลายปัจจัยที่มีผลต่อคะแนนสอบ เช่น ปัจจัยทางด้านความรู้ความเข้าใจ(cognitive) บุคลิกภาพ และปัจจัยเกี่ยวกับการจัดการสอบ ปัจจัยเหล่านี้อาจรวมถึงแรงจูงใจ ความวิตกกังวลในการสอบ ความสามารถในการทำงานได้รวดเร็ว ความรู้เกี่ยวกับการใช้กระดาษคำตอบ เมื่อเป็นเช่นนี้สิ่งที่จำเป็นที่ทำให้ข้อตกลงนี้เป็นไปได้ คือการพิจารณาว่าแบบสอบฉบับนั้นมีองค์ประกอบใดหรือปัจจัยใดที่เด่นที่สุด ก็ถือว่าแบบสอบได้วัดในสิ่งนั้น

ข้อตกลงด้านความเป็นเอกมิตินี้เป็นเรื่องปกติสำหรับผู้สร้างแบบสอบที่ต้องคำนึงอยู่แล้ว เพราะไม่เช่นนั้นจะมีปัญหาในเรื่องการแปลผลคะแนน สมมุติว่าแบบสอบฉบับหนึ่งประกอบด้วยข้อสอบ n ข้อ และนำไปสอบกับกลุ่มประชากรย่อยกลุ่มต่าง ๆ จำนวน r กลุ่ม แล้วพิจารณาการแจกแจงของคะแนนจากแบบสอบว่าแต่ละระดับความสามารถของกลุ่มผู้สอบเหล่านั้น มีการแจกแจงเหมือนกันหรือไม่ถ้ามีการแจกแจงเหมือนกันแสดงว่าแบบสอบนั้นมีความเป็นเอกมิติ (ดังแสดงในรูปที่ 1) ถ้าการแจกแจงมีการแปรเปลี่ยนในแต่ละระดับความสามารถแสดงว่าแบบสอบ

นั้นวัดความสามารถมากกว่า 1 ความสามารถ(ดังรูปที่ 2)



รูปที่ 1 แสดงการแจกแจงของคะแนนของแบบสอบที่เป็นเอกมิติ



รูปที่ 2 แสดงการแจกแจงของคะแนนของแบบสอบที่ไม่เป็นเอกมิติ

ความเป็นอิสระ (local independence)

ข้อตกลงเบื้องต้นเกี่ยวกับความเป็นอิสระ กล่าวว่าคำตอบของผู้สอบที่ตอบข้อสอบแต่ละข้อมีความเป็นอิสระจากกันอย่างมีนัยสำคัญ ข้อตกลงเบื้องต้นข้อนี้จะเป็นจริงก็ต่อเมื่อคะแนนของการตอบในข้อสอบข้อหนึ่งข้อใดจะไม่มีผลต่อคะแนนในการตอบข้ออื่น ๆ ตัวอย่างเช่น เนื้อหาในข้อสอบต้องไม่เป็นการแนะคำตอบสำหรับข้อสอบข้ออื่น เมื่อเกิดความเป็นอิสระนี้จริงความน่าจะเป็นของแบบแผนการตอบข้อสอบแต่ละข้อจึงเป็นไปได้หลาย ๆ แบบ เช่น ความน่าจะเป็นที่เกิดขึ้นในแบบแผนการตอบของข้อสอบ 5 ข้อ $u=(1\ 0\ 1\ 1\ 0)$ เมื่อ 1 แทนคำตอบที่ถูกต้อง และ 0 แทนคำตอบที่ผิด มีค่าเท่ากับ $P_1(1 - P_2)P_3P_4(1 - P_5)$ เมื่อ P_i เป็นความน่าจะเป็นที่ผู้สอบจะตอบข้อสอบข้อที่ i ได้ถูกต้อง และ $1-P_i$ เป็นความน่าจะเป็นที่ผู้สอบจะตอบข้อสอบผิด เมื่อเป็นเช่นนี้การจัดเรียงข้อสอบในลักษณะต่าง ๆ จะไม่กระทบกระเทือนต่อผลการสอบ

ถ้าให้ U_i เมื่อ $i = 1,2...,n$ แทนคำตอบแบบทวิ(ตอบถูกได้ 1 ตอบผิดได้ 0)ของผู้สอบที่ทำข้อสอบ n ข้อ P_i เป็นความน่าจะเป็นในการตอบถูกของข้อสอบ i และ $Q = 1 - P$ จากข้อตกลงเกี่ยวกับความเป็นอิสระจะได้

$$\text{Prob}[U_1 = u_1, U_2 = u_2, \ldots, U_n = u_n \mid \theta] = \text{Prob}[U_1 = u_1 \mid \theta]$$
$$\text{Prob}[U_2 = u_2 \mid \theta] \ldots \text{Prob}[U_n = u_n \mid \theta]$$

ถ้ากำหนดว่า $P_i(\theta) = \text{Prob}[U_i = 1 \mid \theta]$ และ $Q_i(\theta) = \text{Prob}[U_i = 0 \mid \theta]$ จากนั้นจะได้

$$\begin{aligned}\text{Prob}[U_1 = u_1, U_2 = u_2, \ldots, U_n = u_n \mid \theta] &= P_1(\theta)^{u_1}Q_1(\theta)^{1-u_1}P_2(\theta)^{u_2}Q_2(\theta)^{1-u_2}\ldots P_n(\theta)^{u_n}Q_n(\theta)^{1-u_n} \\ &= \prod P_i(\theta)^{u_i}Q_i(\theta)^{1-u_i}\end{aligned}$$

จากสมการนี้หมายความว่าความน่าจะเป็นของแบบแผนการตอบของผู้สอบแต่ละคนมีค่าเท่ากับผลคูณของความน่าจะเป็นที่เกี่ยวข้องกับคำตอบของผู้สอบในแต่ละข้อ

ผลของข้อตกลงเบื้องต้นในเรื่องความเป็นอิสระนี้คือ ค่าความถี่ของคะแนนผู้สอบที่ระดับความสามารถใด ๆ เป็นดังนี้

$$f(x|\theta) = \sum_{i=1}^{n} \prod P_i\theta^{u_i}Q_i\theta^{1-u_i} \qquad (9)$$

เมื่อ X เป็นคะแนนจากการสอบซึ่งมีค่าระหว่าง 0 ถึง n

สิ่งหนึ่งที่อธิบายได้เกี่ยวกับข้อตกลงเบื้องต้นด้านความเป็นอิสระ ในกรณีที่เมื่อ θ มีความเป็นเอกมิติ หรือข้อตกลงเบื้องต้นเกี่ยวกับคุณลักษณะแฝงเป็นจริงคือ ประการแรกสมมุติว่าแบบสอบวัดความสามารถร่วม สำหรับผู้สอบที่มีความสามารถ θ คำตอบที่ได้มามีความเป็นอิสระอย่างมีนัยสำคัญ แต่ถ้าไม่มีความเป็นอิสระ จะเกิดผลว่าผู้สอบบางคนจะมีคะแนนที่คาดหวังสูงกว่าผู้สอบคนอื่นที่มีความสามารถระดับเดียวกัน สิ่งที่ตามมาคือ จำเป็นต้องใช้ความสามารถหลาย ๆ อย่างมาใช้อธิบายคะแนนจากแบบสอบ ลักษณะนี้เป็นการละเมิดข้อตกลงด้านความเป็นเอกมิติประการที่สอง ข้อตกลงด้านความเป็นอิสระแสดงให้เห็นเป็นนัยว่าเมื่อเกิดความเป็นอิสระที่ระดับความสามารถใด ๆ จะมีเพียงหนึ่งความสามารถเท่านั้นที่อธิบายความสัมพันธ์ระหว่างชุดของข้อสอบในแบบสอบได้

เนื่องจากความเท่าเทียมกันระหว่างข้อตกลงเบื้องต้นเกี่ยวกับความเป็นอิสระและความเป็นเอกมิติของคุณลักษณะแฝง การตรวจสอบข้อตกลงเบื้องต้นเกี่ยวกับความเป็นอิสระสามารถทำได้โดยใช้ การวิเคราะห์ปัจจัย(factor analysis) และการตรวจสอบความมีนัยสำคัญทางสถิติสามารถทำได้โดยใช้ χ^2

<u>โค้งคุณลักษณะข้อสอบ</u>

การแจกแจงของคะแนนจากการสอบแบบทวิสำหรับระดับความสามารถหนึ่ง สามารถเขียนได้ดังนี้

$$f_i(u_i \mid \theta) = P_i(\theta)^{u_i} Q_i(\theta)^{1-u_i}$$

เนื่องจาก $f_i(u_i \mid \theta) = P_i(\theta)$ ถ้า $u_i = 1$

และ $f_i(u_i \mid \theta) = Q_i(\theta)$ ถ้า $u_i = 0$ (10)

เส้นโค้งที่เชื่อมโยงค่าเฉลี่ยของการแจกแจง ตามสมการ (10) เป็นการถดถอยของคะแนนจากการสอบ เรียกโค้งนี้ว่าโค้งคุณลักษณะข้อสอบ(item characteristic curve หรือ item characteristic function, ICC) ซึ่งเป็นฟังค์ชันทางคณิตศาสตร์ที่แสดงความสัมพันธ์ของความน่าจะเป็นในการทำข้อสอบได้ถูกกับระดับความสามารถของผู้สอบ และเป็นฟังค์ชันการถดถอยในลักษณะที่ไม่เป็นเส้นตรง

โค้งคุณลักษณะของข้อสอบไม่มีการเปลี่ยนแปลงในกลุ่มประชากร กล่าวคือสำหรับคุณลักษณะแฝงของกลุ่มผู้สอบที่ระดับความสามารถใด ๆ จะไม่แปรเปลี่ยนถ้าการแจกแจงมีลักษณะเหมือนกัน จากนั้นโค้งที่เชื่อมโยงค่าเฉลี่ยของการแจกแจงเหล่านี้ต้องมีลักษณะเหมือนกันด้วย นั่นคือโค้งคุณลักษณะข้อสอบจึงไม่แปรเปลี่ยนทั้งกลุ่มประชากร

ความเร็วในการทำแบบสอบ

ข้อตกลงเบื้องต้นในข้อนี้แสดงเป็นนัยว่า โมเดลการตอบสนองรายข้อจะมีความเหมาะสมก็ต่อเมื่อการสอบไม่อยู่ภายใต้เงื่อนไขของความเร็วในการทำแบบสอบ กล่าวคือผู้สอบที่ทำข้อสอบผิดจะต้องมีสาเหตุมาจากความสามารถของเขาเท่านั้น ไม่ได้เป็นสามเหตุมาจากการทำข้อสอบไม่ทัน ข้อตกลงในข้อนี้จำเป็นต้องมีเพราะเป็นส่วนที่แฝงอยู่ในข้อตกลงเบื้องต้นเกี่ยวกับความเป็นเอกมิติ เนื่องจากเมื่อความเร็วในการสอบเข้ามามีส่วนเกี่ยวข้องในการสอบหมายถึงมีคุณลักษณะหรือความสามารถอย่างน้อย 2 ประการที่เข้ามาเกี่ยวข้องกับการสอบ ได้แก่ ความเร็ว กับคุณลักษณะที่วัดโดยเนื้อหาของข้อสอบ การที่จะพิจารณาว่าแบบสอบเป็นแบบสอบที่เกี่ยวข้องกับความเร็วในการสอบหรือไม่อาจทำได้โดยการนับจำนวนผู้ที่ทำข้อสอบไม่ครบทุกข้อ ก็จะสามารถทราบได้

โมเดลการตอบสนองรายข้อ

เนื่องจากข้อมูลที่ได้มาจากการสอบมีหลายลักษณะ ได้แก่ ข้อมูลแบบทวิ(dichotomus) ข้อมูลแบบพหุ(multi-chotomous) และข้อมูลแบบต่อเนื่อง(continuous)ดังนั้นจึงมีผู้พัฒนาโมเดลเพื่อให้สอดคล้องกับลักษณะข้อมูลดังกล่าวขึ้นมามากมาย แต่สำหรับข้อมูลที่เป็นแบบทวิโมเดลที่นิยมใช้ได้แก่ โมเดลลอจิสติค(logistic model) ชนิด หนึ่ง สอง และสามพารามิเตอร์ซึ่งแต่ละโมเดลมีรายละเอียดดังนี้

โมเดลลอจิสติค 2 พารามิเตอร์(Two-Parameter Logistic Model)

โมเดลนี้ได้เสนอโดยเบิร์นบวม(Birnbaum)เมื่อปี 1957 เป็นโมเดลโค้งคุณลักษณะข้อสอบและเป็นฟังค์ชันของการแจกแจงที่มี 2 พารามิเตอร์ ซึ่งสามารถเขียนได้ดังนี้

$$P_i(\theta) = \frac{e^{Da(\theta - b_i)}}{1 + e^{Da(\theta - b_i)}}$$

$$(i = 1,2,\ldots,n) \qquad (11)$$

เมื่อ $P_i(\theta)$ เป็นความน่าจะเป็นของผู้สอบที่มีความสามารถ θ สามารถตอบข้อสอบ i ได้ถูกต้อง b_i และ a_i เป็นพารามิเตอร์ของข้อสอบ i D เป็นค่าองค์ประกอบมาตร (scaling factor) ซึ่งถ้ากำหนดให้มีค่าเท่ากับ 1.7 แล้ว ค่า $P_i(\theta)$ จากโค้งความถี่สะสมกับโค้งลอจิสติคจะมีค่าต่างกันน้อยกว่า .01 สำหรับทุกค่าของ θ

จากโมเดลนี้อยู่บนข้อตกลงที่ว่าการเดาตอบจะไม่เกิดขึ้น ซึ่งจะเป็นเช่นนี้ได้ก็ต่อเมื่อค่าพารามิเตอร์ $a_i > 0$ (ข้อสอบที่มีค่าความสัมพันธ์ด้านบวกระหว่างคะแนนจากการสอบกับความสามารถของผู้สอบที่วัดโดยแบบสอบนั้น) และค่าความน่าจะเป็นในการตอบข้อสอบได้ถูกจะลดลงถึงศูนย์เมื่อค่าความสามารถลดลง

โมเดลลอจิสติค 3 พารามิเตอร์ (Three-Parameter Logistic Model)

โมเดลลอจิสติค 3 พารามิเตอร์ เป็นการปรับปรุงมาจากโมเดล 2 พารามิเตอร์ เพียงแต่เพิ่มพารามิเตอร์ตัวที่ 3 คือพารามิเตอร์การเดาคำตอบหรือพารามิเตอร์ c_i เข้าไปเท่านั้นเอง โมเดลนี้เขียนในรูปสมการเชิงคณิตศาสตร์ได้เป็น

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$

$$(i = 1, 2, \ldots, n) \qquad (12)$$

เมื่อ $P_i(\theta)$ เป็นความน่าจะเป็นที่ผู้สอบที่มีความสามารถ (θ) จะตอบข้อสอบข้อที่ i ได้ถูกต้อง

b_i เป็นพารามิเตอร์ความยาก

a_i เป็นพารามิเตอร์อำนาจจำแนก

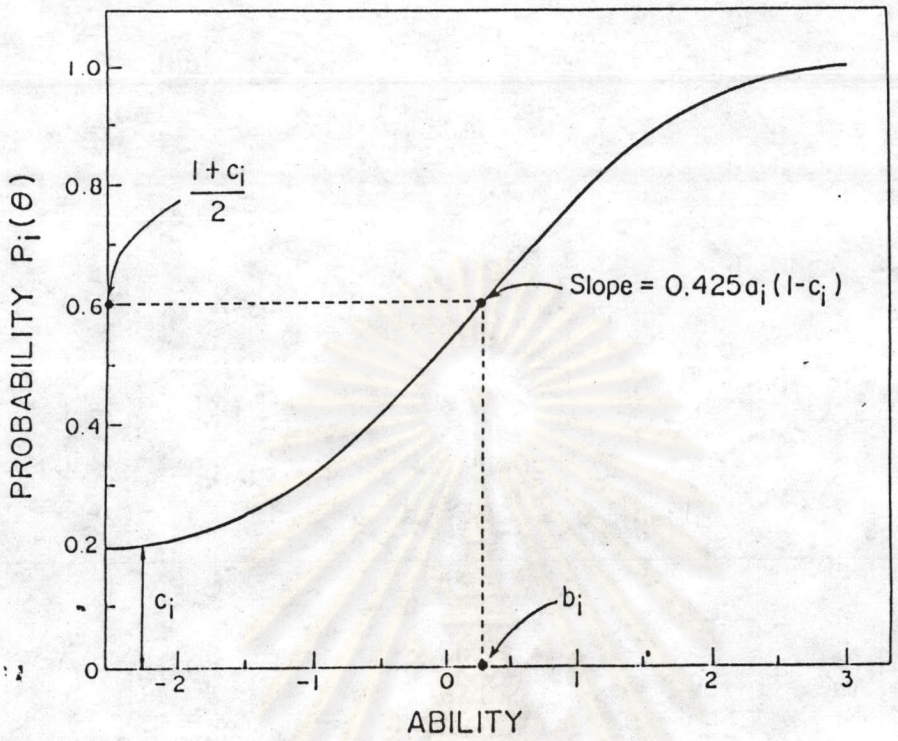
D เป็นค่า scaling factor

พารามิเตอร์ c_i เป็นจุดต่ำสุดที่โค้งคุณลักษณะข้อสอบ ซึ่งเป็นพารามิเตอร์ที่แทนความน่าจะเป็นของผู้สอบที่มีความสามารถต่ำจะตอบข้อสอบได้ถูกต้อง โมเดลที่มีค่าพารามิเตอร์นี้จะใช้ก็ต่อเมื่อคิดว่าการเดาเป็นองค์ประกอบในการตอบข้อสอบ บางครั้งเรียกพารามิเตอร์นี้ว่าเป็นโอกาสที่จะตอบคำตอบได้ถูกต้องสำหรับคนที่มีความสามารถต่ำ

รูปที่ 3 เป็นการอธิบายโมเดลโค้งคุณลักษณะข้อสอบ 3 พารามิเตอร์ โดยค่า b จะเป็นค่าของความสามารถ ณ จุดที่เส้นโค้งมีความชันมากที่สุด ความชันของเส้นกราฟที่จุด b นี้มีค่าเท่ากับ 0.425a(1-c) เมื่อ a คืออำนาจจำแนก ค่านี้จะมีค่าสูงเมื่อโค้งคุณลักษณะข้อสอบ (item characteristic curve,ICC)มีค่าสูง จุดต่ำสุดของเส้นกราฟคือ พารามิเตอร์ c อธิบายได้ว่าเมื่อพารามิเตอร์ c มีค่าไม่เป็นศูนย์ ค่าความน่าจะเป็นที่สอดคล้องกับตำแหน่งพารามิเตอร์ b คือ (1+c)/2 แต่ถ้า c เท่ากับศูนย์ความน่าจะเป็นที่ตำแหน่งพารามิเตอร์ b คือ 50 เปอร์เซ็นต์ และถ้า c ไม่เท่ากับศูนย์ความน่าจะเป็นจะมากกว่า 50 เปอร์เซ็นต์

ในการปรับโมเดล 3 พารามิเตอร์ ให้เป็นโมเดล 2 พารามิเตอร์ ต้องอยู่บนข้อตกลงว่าพารามิเตอร์การเดามีค่าเท่ากับศูนย์ ข้อตกลงนี้ดูเหมือนว่าจะเป็นจริงได้โดยเฉพาะอย่างยิ่งเมื่อข้อสอบไม่ยากจนเกินไป

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย



รูปที่ 3 โมเดลโค้งคุณลักษณะข้อสอบ 3 พารามิเตอร์

<u>โมเดลลอจิสติคหนึ่งพารามิเตอร์</u>(One-Parameter Logistic Model)

โมเดลนี้บางครั้งเรียกว่า ราสช์โมเดล(Rasch Model)เนื่องจากโมเดลนี้ได้พัฒนาโดยนักคณิตศาสตร์ชาวเดนมาร์ค ชื่อ ยอร์ช ราสช์(Georg Rasch)ในปี 1966 โค้งคุณลักษณะข้อสอบตามโมเดลนี้คือ

$$P_i(\theta) = \frac{e^{D\bar{a}_i(\theta - b_i)}}{1 + e^{D\bar{a}_i(\theta - b_i)}}$$

$$(i = 1,2,\ldots,n) \qquad (13)$$

ถึงแม้โมเดลนี้จะเป็นกรณีเฉพาะของโมเดล 2 และ 3 พารามิเตอร์ แต่ก็ยังมีคุณสมบัติพิเศษที่ทำให้นิยมใช้กันคือ ประการแรก เนื่องจากโมเดลนี้มีจำนวนพารามิเตอร์ไม่มาก จึงสะดวกต่อการใช้งาน ประการที่สอง ปัญหาที่เกิดจากการประมาณค่าพารามิเตอร์มีน้อยกว่าการประมาณค่าพารามิเตอร์สำหรับโมเดลที่มีพารามิเตอร์หลาย ๆ ตัว

การนำ IRT มาใช้ในการปรับเทียบคะแนน

การปรับเทียบคะแนนตามแนวทฤษฎีการวัดแบบดั้งเดิมที่ได้อภิปรายมาแล้วไม่สา-มารถเป็นไปตามเงื่อนไขของ ความเสมอภาค ความสมมาตร และความไม่แปรเปลี่ยนได้ การปรับเทียบคะแนนตามแนวทฤษฎีการตอบสนองรายข้อสามารถแก้ปัญหาเหล่านี้ได้ ถ้าโมเดลการตอบสนองรายข้อมีความสอดคล้องกับข้อมูล (Kolen,1981)

จากทฤษฎี IRT ความสามารถของผู้สอบ(θ) มีความเป็นอิสระจากชุดของข้อสอบที่ตอบ และเนื่องจากการประมาณค่า θ มีความคงเส้นคงวาเมื่อทราบค่าพารามิเตอร์ข้อสอบ ดังนั้นการประมาณค่า $\hat{\theta}$ จาก θ ไม่มีอิทธิพลมาจากชุดของคำตอบจึงไม่มีความจำเป็นที่ต้องคำนึงถึงว่าข้อสอบที่ทำนั้นยากหรือง่าย การประมาณค่าความสามารถเป็นการเปรียบเทียบกลุ่มตัวอย่างกับความคลาดเคลื่อนของการเลือกกลุ่มตัวอย่าง ดังนั้นตามขอบเขตของ IRT จึงไม่จำเป็นต้องมีการปรับเทียบคะแนน ซึ่งหลักการนี้เป็นจริงทั้งสถานการณ์การปรับเทียบคะแนนตามแนวดิ่งและตามแนวระดับ

นักจิตมิติหลายคนได้อธิบายว่าการรายงานคะแนนโดยใช้พารามิเตอร์ความสามารถเป็นการยากแก่การเข้าใจของคนทั่วไป จึงจำเป็นต้องใช้คะแนนมาตร(scale score)แทน นอกจากนี้ยังสามารถเปลี่ยนคะแนนความสามารถ (ability score) ให้เป็นคะแนนมาตรได้ กล่าวคือเมื่อรู้ค่า θ ก็สามารถหาค่าคะแนนจริง ξ หรือสัดส่วนของคะแนนที่ถูก π ได้จากสมการ

$$\xi = \Sigma\, p_i(\theta) \qquad (14)$$

$$\pi = \quad /n \qquad (15)$$

เมื่อ n คือจำนวนข้อสอบในแบบสอบ และเนื่องจากไม่ทราบค่า θ ดังนั้นจึงใช้ค่าประมาณของความสามารถคือ $\hat{\theta}$ แทน จะได้ $\hat{\xi}$ และ $\hat{\pi}$ โดยที่ค่า $\hat{\pi}$ มีค่ามากกว่า 0 และมีค่าน้อยกว่า 1 ขณะที่$\hat{\xi}$มีค่ามากกว่า 0 และน้อยกว่า n เมื่อใช้โมเดล 1 หรือ 2 พารามิเตอร์

แต่สำหรับโมเดล 3 พารามิเตอร์ค่าขีดจำกัดล่างจะเปลี่ยนไป เนื่องจาก $P_i(\theta)$มากกว่าหรือเท่ากับ c_i ซึ่งเป็นค่าพารามิเตอร์ของการเดาคำตอบ ดังนั้น

$$\sum_{i=1}^{n} c_i < \hat{\xi} < n$$

ขณะที่ $(\Sigma c_i)/n < \hat{\pi} < 1$

โดยสรุปสำหรับ IRT เมื่อรู้ค่าพารามิเตอร์ข้อสอบแล้ว การปรับเทียบคะแนนไม่มีความจำเป็น และการรายงานผลอาจอยู่ในรูปของคะแนนที่แปลงหรือคะแนนมาตรก็ได้

สถานการณ์ที่กล่าวมาจะใช้เมื่อไม่ทราบค่าพารามิเตอร์ข้อสอบ แต่เมื่อทราบค่าพารามิเตอร์ สเกลของ θ จะคงที่ถึงแม้ว่าจะไม่นำไปใช้ในการแปลงคะแนนก็ตาม อย่างไรก็ตาม เมื่อไม่รู้ค่าพารามิเตอร์ความสามารถและพารามิเตอร์ข้อสอบ ฟังค์ชันการตอบสนองรายข้อไม่มีการแปรเปลี่ยน การแปลงความสามารถและพารามิเตอร์ข้อสอบจึงไม่เป็นไปในเชิงเส้นตรง จึงจำเป็นต้องเลือกมาตรความสามารถ(θ) และพารามิเตอร์ข้อสอบที่กำหนดขึ้นมาเองตามความคิดของตนเอง(arbitary metric) สำหรับโมเดล 1 พารามิเตอร์ การกำหนดมาตรให้คงที่มีความเหมาะสมเมื่อค่าเฉลี่ยของ θ และความยาก bให้มีค่าเป็นศูนย์ ตามปกติแล้วสำหรับโมเดล 2 และ 3 พารามิเตอร์จะกำหนดให้ค่าเฉลี่ยของ θ (หรือ b)ให้เป็น ศูนย์และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 1

เกี่ยวกับความไม่แปรเปลี่ยนของโมเดลการตอบสนองรายข้อ ความสามารถของผู้สอบไม่มีอิทธิพลจากการจัดการสอบ และพารามิเตอร์ข้อสอบยังคงไม่แปรเปลี่ยนไม่ว่ากลุ่มผู้สอบจะเปลี่ยนไปก็ตาม อย่างไรก็ตามค่าพารามิเตอร์ข้อสอบจากกลุ่มผู้สอบสองกลุ่มที่วิเคราะห์แยกจากกันอาจเกิดการแปรเปลี่ยนจึงเกิดความแตกต่างกันขึ้นเนื่องจากการกำหนดความคงที่ของมาตรขึ้นมานั่นเอง และยังมีความสัมพันธ์เชิงเส้นตรงเกิดขึ้นระหว่างพารามิเตอร์ข้อสอบและพารามิเตอร์ความสามารถจากกลุ่มตัวอย่างทั้งสอง

สมมุติว่ากลุ่มตัวอย่างกลุ่มเดียวทำการสอบแบบสอบ 2 ฉบับ X และ Y โดยที่ทั้งสองฉบับนี้วัดในคุณลักษณะเดียวกัน สำหรับโมเดล 1 พารามิเตอร์ เมื่อกำหนด b ให้คงที่โดยมีค่าเฉลี่ยเท่ากับ 0 จะได้ความสัมพันธ์ดังนี้

$$\theta_x - \mu_{\theta x} = \theta_y - \mu_{\theta y} \qquad (16)$$

$$\theta_y = \theta_x + (\mu_{\theta y} - \mu_{\theta x}) \qquad (17)$$

สำหรับโมเดล 2 และ 3 พารามิเตอร์ เนื่องจากค่าเฉลี่ยเท่ากับ 0 และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 1 จะได้

$$\frac{\theta_x - \mu_{\theta x}}{\sigma_{\theta x}} = \frac{\theta_y - \mu_{\theta y}}{\sigma_{\theta y}} \qquad (18)$$

หรือ

$$\theta_y = \frac{\sigma_{\theta y}}{\sigma_{\theta x}} \theta_x + \left[\mu_{\theta y} - \frac{\sigma_{\theta y}}{\sigma_{\theta x}} \mu_{\theta x}\right] \qquad (19)$$

เมื่อ $\mu_{\theta x}$ และ $\sigma_{\theta x}$ แทนค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของ θ สำหรับแบบสอบฉบับ X และสำหรับแบบสอบฉบับ Y ใช้ในทำนองเดียวกัน

สมการ (16)- (19) เป็นการกำหนดความสัมพันธ์ระหว่าง θ จากแบบสอบทั้งสองฉบับ ความสัมพันธ์นี้เป็นความสัมพันธ์เชิงเส้นตรง ซึ่งจัดรูปใหม่ได้เป็น

$$\theta_y = \alpha\theta_x + \beta \qquad (20)$$

สำหรับโมเดล 1 พารามิเตอร์ ค่า $\alpha = 1$ จากสมการ (20) เมื่อรู้ค่าคงที่ α และ β ก็สามารถปรับเทียบความสามารถจากแบบสอบทั้งสองได้

ตามแบบแผนการเก็บรวบรวมข้อมูลแบบที่ 1 คือแบบแผนกลุ่มเดียว แบบสอบทั้งสองได้นำไปใช้กับกลุ่มตัวอย่างกลุ่มเดียวกัน วิธีการที่ง่ายที่สุดคือการทำให้แบบสอบทั้งสองเสมือนว่าสอบในคราวเดียวกัน มีการเชื่อมโยงคำตอบเข้าด้วยกันและประมาณค่าพารามิเตอร์ความสามารถและพารามิเตอร์ข้อสอบพร้อมกัน การทำเช่นนี้ทำให้พารามิเตอร์ความสามารถและพารามิเตอร์ข้อสอบอยู่ในสเกลเดียวกัน จึงไม่จำเป็นต้องปรับเทียบคะแนน

ในกรณีที่ไม่สามารถวิเคราะห์พร้อมกันได้จำเป็นต้องมีการปรับเทียบ เนื่องจากผู้สอบทุกคนมีคู่ลำดับของความสามารถ(θ_x, θ_y)ความสัมพันธ์ระหว่างคู่ลำดับเหล่านี้จะเป็นสิ่งกำหนดค่า α และ β ในสมการ (20) ได้

ถ้ามีการกำหนดมาตรของ θ ให้มีค่าคงที่สำหรับสถานการณ์การปรับเทียบ 2 สถานการณ์ จึงได้ $\mu_{\theta x} = \mu_{\theta y} = 0$ และ $\sigma_{\theta x} = \sigma_{\theta y} = 1$ ดังนั้นเมื่อแทนค่าในสมการ (19) จะได้ $\theta_x = \theta_y$ จึงไม่จำเป็นต้องปรับเทียบคะแนนเช่นกัน ส่วนในมาตรของพารามิเตอร์ b ถ้ามีการกำหนดให้คงที่ สมการ (19) จะให้ค่าคงที่ของการปรับเทียบ

ในแบบแผนการเก็บรวบรวมข้อมูลแบบที่ 3 คือแบบแผนข้อสอบร่วม ค่าพารามิเตอร์ความยากและพารามิเตอร์อำนาจจำแนกของข้อสอบร่วมมีความสัมพันธ์เชิงเส้นตรง และเนื่องจากค่าพารามิเตอร์มีเป็นคู่ (b_x, b_y) และ (a_x, a_y) ความสัมพันธ์ระหว่างค่าพารามิเตอร์เหล่านี้หาได้ด้วยการกำหนด θ ในแต่ละกลุ่มให้คงที่ และหาความสัมพันธ์จาก
สมการต่อไปนี้

$$b_y = \alpha b_x + \beta \qquad (21)$$

และ $$a_y = a_x / \alpha \qquad (22)$$

เมื่อ $$\alpha = \sigma_{b_y} / \sigma_{b_x} \qquad (23)$$

และ $$\beta = \mu_{b_y} - \alpha_{b_x} \qquad (24)$$

สถานการณ์เช่นนี้คล้ายกับแบบแผนที่ 1 ยกเว้นเกิดสารสนเทศเพิ่มขึ้นมาเท่านั้น ซึ่งได้แก่ ความชันของกราฟความยากจะเป็นสัดส่วนผกผันกับความชันของเส้นกราฟอำนาจจำแนกของข้อสอบ

แบบแผนการเก็บรวบรวมข้อมูลอีกประเภทหนึ่งที่คล้ายกันกับที่กล่าวมาในแบบที่ 1 คือ แบบแผนกลุ่มผู้สอบร่วม หมายถึงมีกลุ่มผู้สอบจำนวนหนึ่งที่ทำแบบสอบทั้งสองฉบับที่แตกต่างกัน แบบแผนนี้ก็เป็นส่วนย่อยของแบบแผนกลุ่มเดี่ยวนั่นเอง ตามแบบแผนนี้ข้อสอบแต่ละข้ออยู่บนสเกลเดียวกันและสามารถปรับเทียบความสามารถได้ในกรณีเช่นนี้มาตรของพารามิเตอร์ความสามารถและพารามิเตอร์ข้อสอบได้ถูกกำหนดให้คงที่

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

<u>การกำหนดค่าคงที่ของการปรับเทียบคะแนน</u>

เมื่อมีคู่ลำดับเช่น $(\theta_x, \theta_y), (b_x, b_y), (a_x, a_y)$ แล้วนำค่าเหล่านี้มาพล็อตกราฟจะได้ความสัมพันธ์ในรูปเส้นตรงซึ่งจะได้ค่าความชัน และจุดตัด แต่เป็นที่น่าเสียดายเพราะเนื่องจากพารามิเตอร์เหล่านี้ได้มาจากการประมาณค่า คู่ลำดับทุกคู่จึงไม่อยู่บนเส้นตรงเดียวกันแต่จะเป็นจุดที่อยู่ในลักษณะกระจายออกจากเส้นตรง ดังนั้นการกำหนดค่าคงที่ตามสมการเส้นตรงจึงสามารถทำได้โดยใช้วิธีการต่อไปนี้ คือ

ก) วิธีการถดถอย(regression method)

ข) วิธีการ Mean and sigma(Mean and sigma procedure)

ค) วิธีการ Robust mean and sigma

ง) วิธีโค้งคุณลักษณะข้อสอบ

แต่ละวิธีมีรายละเอียดดังนี้

วิธีการถดถอย

ความสัมพันธ์เชิงเส้นตรงระหว่างตัวแปร 2 ตัว ส่วนมากสามารถพบเห็นได้จากวิธีการถดถอย ในกรณีนี้ได้

$$Y = \alpha X + \beta + e \qquad (25)$$

เมื่อ $y = \theta_x$ และ $x = \theta_y$ ในกรณีที่การปรับเทียบคะแนนอยู่ในรูปของความสามารถ แต่เมื่อใช้กับการปรับค่าความยากจะเปลี่ยนเป็น $y = b_y$, $x = b_x$ ค่าความคลาดเคลื่อน e เป็นค่าที่อิสระและมีการแจกแจงเหมือนกับตัวแปรเชิงสุ่ม การประมาณค่าของสัมประสิทธิ์ α และ β กำหนดได้จากสมการต่อไปนี้

$$\hat{\alpha} = r_{xy}s_y/s_x \qquad (26)$$

และ

$$\hat{\beta} = y - \alpha x \qquad (27)$$

เมื่อ r_{xy} เป็นค่าสหสัมพันธ์ระหว่าง x กับ y และ $\bar{y}$ กับ $\bar{x}$ เป็นค่าเฉลี่ยของ y และ x และ s_y กับ s_x เป็นค่าส่วนเบี่ยงเบนมาตรฐานตามลำดับ

วิธีการนี้มีจุดอ่อนที่ความสัมพันธ์ไม่มีความสมมาตร สัมประสิทธิ์การถดถอยจะได้รับผลจากแบบสอบที่เลือกให้เป็นฐานเท่านั้น นอกจากนี้ยังมีข้อสมมุติว่าการวัดค่า x ไม่มีความผิดพลาด เนื่องจากเป็นการไม่สมเหตุสมผลที่แบบสอบฉบับหนึ่งได้รับเลือกให้เป็นฐาน วิธีการนี้จึงให้ผลการปรับเทียบไม่มีความสมมาตร ยิ่งกว่านี้ค่าความคลาดเคลื่อนไม่จำเป็นต้องมีการแจกแจงเหมือนกัน เนื่องจากพารามิเตอร์ความสามารถและพารามิเตอร์ข้อสอบทุกตัวที่ได้มาจากการประมาณค่ามีความคลาดเคลื่อนมาตรฐานในการประมาณค่าแตกต่างกัน

ข้อยกเว้นในกรณีที่ขาดการสมมาตรคือเมื่อมีความเหมาะสมกับราสช์โมเดล ซึ่งในกรณีนี้ค่า $\alpha = 1$ ดังนั้น $\hat{\beta} = \bar{y} - \bar{x}$ และ $y = x + (\bar{y} - \bar{x})$ และ $x = y + (\bar{x} - \bar{y})$

ความสัมพันธ์เช่นนี้คือการสมมาตร

วิธี Mean และ sigma

วิธีการนี้ใช้หลักการที่ว่าจากสมการ $y = \alpha x + \beta$ จะได้

$$\bar{y} = \alpha\bar{x} + \beta \quad (28)$$

และ $$s_y = \alpha s_x \quad (29)$$

โดย $$\alpha = s_y/s_x \quad (30)$$

และ $$\beta = \bar{y} - \alpha\bar{x} \quad (31)$$

ผลที่ได้จากสมการเหล่านี้จะเกิดความสมมาตร และสำหรับราสช์โมเดลวิธีการนี้และวิธีการถดถอยให้ผลอย่างเดียวกัน

วิธี Robust Mean และ Sigma

ขณะที่วิธี Mean and Sigma มีความสมมาตร โดย $x = (y - \beta)/\alpha$ แต่ก็ยังไม่มีการอธิบายในเรื่องความจริงที่ว่า ทุก ๆ พารามิเตอร์ข้อสอบและพารามิเตอร์ความสามารถถูกประมาณค่ามาโดยมีความเที่ยงตรงที่แปรเปลี่ยน ยิ่งกว่านั้นสิ่งอื่น ๆ ยังมีผลต่อการคำนวณค่าสัมประสิทธิ์ด้วย

วิธี robust mean and sigma เสนอโดย ลินน์, เลไวน์, ฮัสติงส์ และ วอดรอป (Linn, Levine, Hastings และ Wardrop, 1981) เนื่องจากคู่ลำดับ (x,y) เป็นคู่ของการประมาณค่าของความสามารถจากแบบสอบสองฉบับหรือความยากของข้อสอบในสองกลุ่มผู้สอบ ทุก ๆ ค่าของ x และ y มีความคลาดเคลื่อนมาตรฐานของการประมาณค่าของตัวมันเอง น้ำหนักสำหรับแต่ละคู่จึงเป็นส่วนกลับของความแปรปรวนที่ประมาณค่าจากกลุ่มที่ใหญ่กว่าสองกลุ่ม

ถ้าใช้ค่าความสามารถในการหาค่าคงที่ในการปรับเทียบคะแนน การประมาณค่าความแปรปรวนจะเป็นส่วนกลับของฟังค์ชันสารสนเทศที่ประมาณค่า ณ ตำแหน่งความสามารถนั้นเนื่องจากความแปรปรวนที่มากกว่าจะให้ค่าฟังค์ชันสารสนเทศที่น้อยกว่า ดังนั้นความสามารถที่มีค่าความแปรปรวนมากกว่าจะได้รับน้ำหนักน้อย เมื่อความยากที่ประมาณค่าจากสองกลุ่มได้นำมาใช้

ในการหาค่าคงที่ของการปรับเทียบ แมตริกซ์สารสนเทศสำหรับข้อสอบแต่ละข้อก็สามารถหาได้ และจะได้ค่าองค์ประกอบในแนวทแยงที่เหมาะสมสำหรับการประมาณค่าความแปรปรวน มิติของแมตริกซ์สารสนเทศจะมากหรือน้อยขึ้นอยู่กับโมเดลที่ใช้ เช่น สำหรับโมเดล 3 พารามิเตอร์ จะมีมิติแบบ 3 x 3 วิธีการนี้สามารถสรุปได้ดังนี้

1. กำหนด w_j สำหรับทุกคู่ลำดับ (x_j, y_j) คือ

$$w_j = \max\{v(x_j), v(y_j)\}, \quad j = 1, \ldots, k$$

เมื่อ $v(\cdot)$ แทนความแปรปรวนของการประมาณค่าความแปรปรวนลำดับที่ j

2. ค่าน้ำหนักของมาตร

$$w'_j = w_j / \left(\sum_{j=1}^{n} w\right)_j$$

3. คำนวณ

$$x'_j = w'_j x_j \qquad (j = 1, \ldots, k)$$

และ $$y'_j = w'_j y_j \qquad (j = 1, \ldots, k)$$

4. กำหนด $\bar{x}$, $\bar{y}$, $\bar{s}_x$, $\bar{s}_y$ เมื่อ x, y, s_x, และ s_y เป็นค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของน้ำหนักคะแนน

5. กำหนดค่า α และ β โดยใช้สมการ (26) และใช้การให้น้ำหนักค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐาน

สต็อกกิงและลอร์ด(Stocking และ Lord,1983) ได้ชี้ให้เห็นว่า แม้ว่าวิธีการนี้จะพัฒนามาจากวิธี mean and sigma แต่ก็ไม่สามารถอธิบายส่วนประกอบอื่น ๆ ที่เหลือได้ เขาจึงแนะนำให้ใช้การให้น้ำหนักโดยยึดระยะทางที่ตั้งฉากกับเส้นการปรับเทียบ วิธีการนี้ใช้การคำนวณในขั้นที่ 1-5 ก่อน แล้วเพิ่มเติมด้วยการคำนวณต่อไปนี้

6. เมื่อได้ค่า α และ β และได้เส้นตรงแล้ว หาค่าระยะทางที่ตั้งฉากกับเส้นตรงสำหรับทุกจุด(x_j, y_j) โดย

$$d_j = (y_j - \alpha x_j - \beta_j)/[\alpha^2 + \beta^2)]^{1/2}$$

แล้วหาค่ามัธยฐาน (M)

7. คำนวณหาค่าน้ำหนักตุ๊กกี(Tukey weights) ซึ่งกำหนดไว้ว่า

$$T_i = \begin{cases} [1 - (d_i/6M)^2]^2 & \text{เมื่อ } d_i < 6M \\ 0 \end{cases}$$

8. ให้น้ำหนักใหม่กับทุกจุดของ (x_i', y_i') โดยใช้น้ำหนัก

$$u_i = w_i' T \{ \sum_{i=1}^{n} w_i' T_i \}^{-1}$$

9. ทำซ้ำในขั้นที่ 3 โดยใช้ u_i แทน w_i และคำนวณค่า α และ β ในขั้นที่ 5

10. ทำซ้ำในขั้นที่ 6, 7, 8 และ 9 จนกระทั่งได้ค่า α และ β น้อยกว่าค่าที่กำหนด

<u>วิธีโค้งคุณลักษณะข้อสอบ</u>

ขณะที่วิธี robust mean and sigma ได้รับความนิยม แต่วิธีนี้ก็ยังมีข้อเสียคือ เมื่อใช้ค่าประมาณของพารามิเตอร์ข้อสอบในการหาเส้นของการปรับเทียบ จะมีการใช้เพียงความสัมพันธ์ของความยากข้อสอบเท่านั้น คือ

$$b_y = \alpha b_x + \beta$$

ความสัมพันธ์ระหว่างอำนาจจำแนก เช่น $a_y = a_x/\alpha$ ไม่ได้ถูกนำมาใช้ในการกำหนดค่า α ซึ่งค่านี้เป็นสารสนเทศที่สำคัญซึ่งสามารถนำมาใช้ในการให้น้ำหนักคล้ายกับวิธีการที่กล่าวมาแล้ว และค่าเฉลี่ยของ α สามารถกำหนดได้ วิธีการโค้งคุณลักษณะข้อสอบเป็นวิธีการหนึ่งที่เสนอโดย แฮบารา(Haebara,1980)และสต็อกกิงและลอร์ด(Stocking และ Lord, 1983) สามารถนำมาแก้ปัญหาได้ กล่าวคือ

คะแนนจริง ξ ของผู้สอบที่มีความสามารถ θ_a สำหรับแบบสอบ X คือ

$$\xi_{xa} = \sum_{i=1}^{n} P(\theta_a, a_{xi}, b_{xi}, c_{xi}) \qquad (32)$$

คะแนนจริงของผู้สอบ a ที่มีความสามารถ a สำหรับแบบสอบ Y คือ

$$\xi_{ya} = \Sigma P(\theta_a, a_{yi}, b_{yi}, c_{yi}) \qquad (33)$$

เมื่อ $$b_{yi} = \alpha b_{xi} + \beta \qquad (34)$$

$$a_{yi} = a_{xi}/\alpha \qquad (35)$$

I16701719

และ $c_{xi} = c_{xi}$ (36)

ค่าคงที่ α และ β จะต้องเลือกค่าที่ทำให้ความแตกต่างระหว่าง และ มีค่าน้อยที่สุด ลอร์ดและสต็อกกิงได้เสนอเกณฑ์ที่เหมาะสมในการเลือก คือ

$$F = (1/N)[\Sigma_{a=1}^{n}(\xi_{xa} - \xi_{xa})^2 \quad (37)$$

เมื่อ N เป็นจำนวนผู้สอบ และ F เป็นฟังค์ชันของ α และ β ซึ่งถูกทำให้มีค่าต่ำสุดเมื่อ

$$\partial F/\partial \alpha = \partial F/\partial \beta \quad (38)$$

สมการเหล่านี้ไม่ใช่สมการเชิงเส้นตรงและต้องมีการหาคำตอบเป็นขั้นตอนเรียงตามลำดับโดยใช้ วิธีนิวตัน-ราฟสัน(Newton-Raphson procedure) ลอร์ดและสต็อกกิงได้เปรียบเทียบกับวิธี mean and sigma และพบว่าวิธีโค้งคุณลักษณะข้อสอบให้ผลการแปลงอำนาจจำแนกข้อสอบได้ดีกว่า และวิธีโค้งคุณลักษณะข้อสอบนี้เป็นวิธีการที่มีเหตุผลเนื่องจากได้ใช้สารสนเทศที่มีอยู่อย่างครบถ้วน

<u>การปรับเทียบคะแนนจริง</u>

ในบางครั้งการรายงานผลการปรับเทียบโดยใช้ค่าความสามารถ(θ-scale)ไม่เป็นที่ยอมรับของคนบางกลุ่ม ค่า θ จึงสามารถแปลงเป็นคะแนนจริงได้โดยใช้สมการ (15) จึงเป็นไปได้ที่จะปรับเทียบคะแนนของแบบสอบชุดต่างๆ

ให้ θ_x แทนระดับความสามารถของผู้สอบแบบสอบฉบับ X และ ξ_x เป็นคะแนนจริงของผู้สอบ

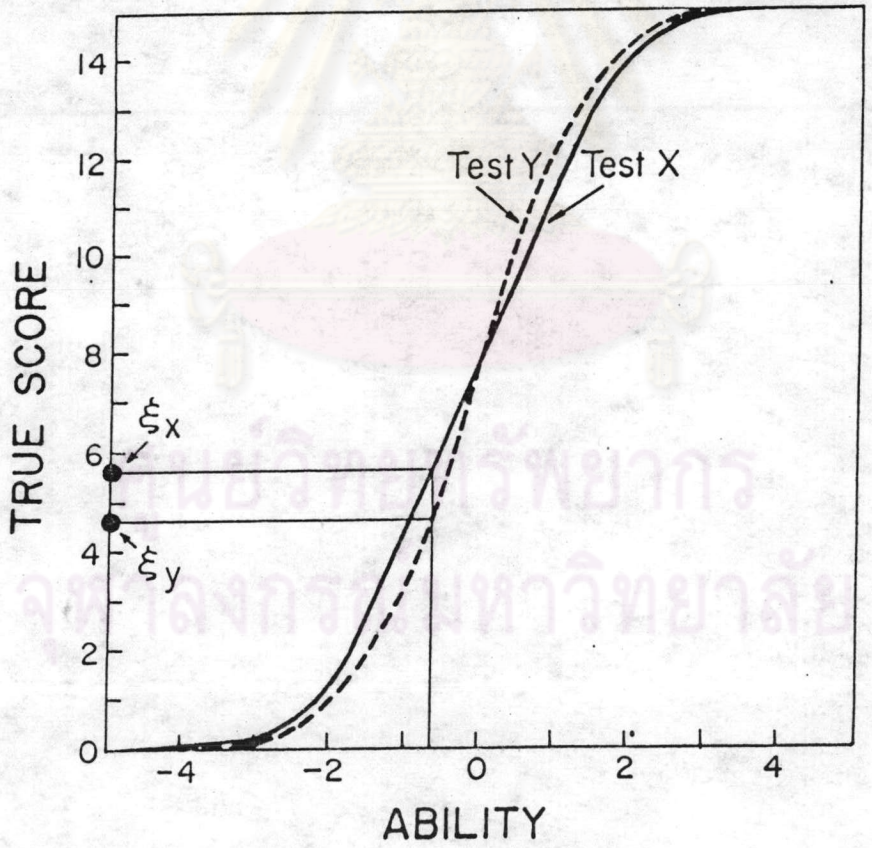
โดย $\xi_x = \sum_{i=1}^{n} P_i(\theta_x)$ (39)

ในทำนองเดียวกันถ้า θ_y เป็นความสามารถของผู้สอบแบบสอบฉบับ Y และมีคะแนนจริง ξ_y จึงได้

$$\xi_y = \sum_{j=1}^{m} P_j(\theta_y) = \sum_{j=1}^{m} P_j(\alpha\theta_x + \beta) \qquad (40)$$

เมื่อ $\theta_y = \alpha\theta_x + \beta$ เป็นเส้นตรงที่แสดงถึงความสัมพันธ์ระหว่าง θ_y และ θ_x สำหรับค่า θ_x แต่ละค่าสามารถหาคู่ลำดับ(ξ_x , ξ_y) ได้ ดังนั้นจึงสามารถปรับเทียบคะแนนจริงจากแบบสอบทั้งสองได้ ดังรูปที่ 4

เส้นตรงความสัมพันธ์ระหว่าง θ_x และ θ_y สามารถหาได้จากวิธีการใดวิธีการหนึ่งที่ได้อธิบายมาแล้ว อย่างไรก็ตามตามหลักการของวิธีโค้งคุณลักษณะข้อสอบแล้วจะสอดคล้องกับหลักการของคะแนนจริง จึงเป็นวิธีการที่เหมาะสมมากที่สุดสำหรับการปรับเทียบคะแนนจริง



รูปที่ 4 แสดงการปรับเทียบคะแนนจริง

กราฟระหว่างคู่ของ ξ_x กับ ξ_y ไม่เป็นเส้นตรง สิ่งนี้ไม่เป็นปัญหาเมื่อรู้ค่า θ_x และ θ_y เนื่องจากในกรณีนี้ความสัมพันธ์ระหว่าง ξ_x และ ξ_y สามารถหาได้ อย่างไรก็ตาม โดยทั่วไปเมื่อ θ_x และ θ_y ได้มาจากการประมาณค่า ดังนั้นความสัมพันธ์ระหว่าง $\hat{\theta}_x$ และ $\hat{\theta}_y$ จะได้รับอิทธิพลจากความคลาดเคลื่อน ความคลาดเคลื่อนเชิงสุ่มจะมีค่ามากสำหรับค่าพารามิเตอร์ข้อสอบและค่าพารามิเตอร์ความสามารถที่มีลักษณะสุดโต่ง จึงทำให้ค่าความ สัมพันธ์ระหว่าง ξ_x และ ξ_y จะมีความคลาดเคลื่อนมาก และในขอบเขตเช่นนี้จำเป็นต้องใช้วิธีการเทียบส่วน วิธีการเทียบที่แองกอฟ(Angoff,1984) ได้อธิบายไว้ในเรื่องการปรับเทียบแบบอีควิเปอร์เซ็นไตล์สามารถนำมาประยุกต์ใช้ได้ในสถานการณ์นี้ จุดนี้เองที่ทำให้การปรับเทียบคะแนนจริงมีจุดด้อยที่ชัดเจน คือ ข้อได้เปรียบของการปรับเทียบ θ_x และ θ_y ที่มีข้อดีคือมีความสัมพันธ์จะสูญเสียไป

ปัญหานี้สามารถหลีกเลี่ยงได้ถ้าความสัมพันธ์ที่ไม่เป็นเชิงเส้นตรงระหว่าง ξ_x และ ξ_yไม่ได้รับการกำหนด และหาค่า $\hat{\theta}_x$ และ $\hat{\theta}_y$ โดยใช้ความสัมพันธ์ $\theta_y = \alpha\theta_x + \beta$ แทนจากนี้สามารถคำนวณ $\hat{\xi}_x$ และ $\hat{\xi}_y$ โดยใช้สมการ (34) และ (35) แล้วทำตารางขึ้นมาการแปลงคะแนนจากแบบสอบหนึ่งไปยังแบบสอบหนึ่งโดยการใช้ตารางนั้นอีกครั้งหนึ่ง

<u>การปรับเทียบคะแนนดิบโดยใช้ IRT</u>

การปรับเทียบคะแนนตามแนว IRT สามารถทำได้โดยการทราบค่าพารามิเตอร์ความสามารถ (θ) หรือคะแนนจริง คำถามที่ตามมาคือ สามารถปรับเทียบคะแนนที่ได้จากการสอบได้หรือไม่

คะแนนจริง จะอยู่บนสเกลเดียวกันกับคะแนนดิบ (r) ก็ต่อเมื่อ

$$r = \sum_{i=1}^{n} U_i$$

ยิ่งกว่านั้น ถ้า IRT มีความสมเหตุสมผลจะได้ $E(r) = \xi$

ดังนั้นสิ่งที่ทำได้ในตอนนี้คือ

1. หาความสัมพันธ์ระหว่างคะแนนจริง ξ_x และ ξ_y จากแบบสอบทั้งสองดังที่ได้กล่าวมาแล้ว

2. ให้การกระทำกับความสัมพันธ์คล้ายกับเป็นความสัมพันธ์ระหว่างคะแนนดิบ r_x และ r_y แล้วปรับเทียบคะแนนดิบ

ลอร์ด(Lord, 1980) ได้อธิบายว่าความสัมพันธ์ที่มีอยู่ระหว่าง ξ_x และ ξ_y ไม่จำเป็นต้องเหมือนกับที่มีอยู่ระหว่างคะแนนดิบ r_x และ r_y เรื่องนี้สามารถอธิบายได้จากโมเดล 3 พารามิเตอร์ คือ $\xi_x > \sum_{i=1}^{n} c_i$ และ $\xi_y > \sum_{i=1}^{m} c_i$ ขณะที่คะแนนที่ได้จากการสอบ(คะแนนดิบ) r_x และ r_y อาจมีค่าเป็นศูนย์ ดังนั้นการปรับเทียบคะแนนโดยใช้คะแนนจริงจะไม่ให้สารสนเทศการปรับเทียบของผู้สอบที่มีคะแนนดิบต่ำกว่าระดับการเดาเพื่อหลีกเลี่ยงปัญหานี้อาจจะใช้สูตรการแปลงคะแนนเข้ามาช่วยได้ อย่างไรก็ตามถ้าจะกล่าวกันตามหลักเกณฑ์โดยตรงแล้วความสัมพันธ์ระหว่างคะแนนจริง ξ_x และ ξ_y อาจจะไม่นำมาใช้ในการปรับเทียบคะแนนดิบ r_x และ r_y

ทฤษฎีการตอบสนองรายข้อ(IRT) ได้ให้วิธีการสำหรับทำนายการแจกแจงคะแนนดิบของแบบสอบที่กำหนดได้ เมื่อได้ลักษณะการแจกแจงคะแนนดิบของแบบสอบฉบับ X และ ฉบับ Y อาจจะใช้วิธีการปรับเทียบแบบอีควิเปอร์เซ็นไทล์ แต่ข้อได้เปรียบอย่างมากของ IRT คือ แบบสอบที่นำมาใช้นั้นไม่จำเป็นต้องมีระดับความยากใกล้เคียงกัน

ตามทฤษฎีแล้วการแจกแจงของคะแนนดิบ $f(r|\theta)$ ของแบบสอบสามารถกำหนดจากความเป็นเอกลักษณ์(identity)(Lord, 1980)

$$\sum_{r=1}^{n} f(r|\theta)t^r = \prod_{i=1}^{n} [Q_i(\theta) + tP_i(\theta)] \quad (41)$$

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

<u>ความคลาดเคลื่อนของการปรับเทียบคะแนน</u>

วิธีการปรับเทียบคะแนนทุกวิธีไม่ว่าจะเป็นการปรับเทียบด้วยวิธีใด เมื่อกลุ่มตัวอย่างผู้สอบได้รับการสุ่มมาจากกลุ่มประชากร ย่อมมีความแปรผันเชิงสุ่มเกิดขึ้น จึงนิยมใช้เทคนิคการประมาณค่าความคลาดเคลื่อนเชิงสุ่ม(estimated sampling errors) ความคลาดเคลื่อน

เชิงสุ่มนี้มีข้อตกลงว่ากลุ่มตัวอย่างได้มาจากการสุ่ม และใช้ความคลาดเคลื่อนมาตรฐานของการเทียบมาตรา(standard error of equating, SEE) เป็นปริมาณที่ใช้วัดความแปรผันที่เกิดขึ้นนี้

เนื่องจากการปรับเทียบแต่ละวิธี มีวิธีการและข้อตกลงเบื้องต้นที่แตกต่ากัน ดังนั้นค่า SEE ของการปรับเทียบคะแนนในแต่ละวิธีจึงมีสูตรการหาแตกต่างกัน การศึกษาครั้งนี้มีขอบเขตอยู่ที่การปรับเทียบคะแนนตามแนว IRT จึงขอเสนอเฉพาะความคลาดเคลื่อนมาตรฐานและวิธีการประเมินความคลาดเคลื่อนของการปรับเทียบคะแนนตามแนวนี้เท่านั้น

ความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนตามแนว IRT

ลอร์ด(Lord,1982) ได้พัฒนาสูตรขึ้นมาใช้สำหรับหาความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนน ดังนี้

ใน IRT เป็นคะแนนที่คาดหวังในการตอบแบบสอบ X ได้ถูกที่ระดับความสามารถ θ จะได้

$$\xi = \sum_{g=1}^{n_x} P_g(\theta) \qquad (42)$$

เมื่อ $P_i(\theta)$ เป็นฟังค์ชันการตอบสนองรายข้อ คือเป็นความน่าจะเป็นในการตอบข้อสอบ i ได้ถูกที่ระดับความสามารถ θ ถ้ามีการสอบครั้งที่ 2 โดยใช้แบบสอบฉบับ Y ที่วัดความสามารถเดียวกันกับ X คะแนนที่คาดหวังในการตอบได้ถูกคือ η ซึ่งอาจเขียนได้ดังนี้

$$\eta = \sum_{h=1}^{n_y} P_h(\theta) \qquad (43)$$

สมการ (42) และ (43) เป็นสมการพาราเมตริกซึ่งเป็นฟังค์ชันความสัมพันธ์ระหว่าง ξ และ η ซึ่งเป็นสมการทางคณิตศาสตร์ ที่ระดับ θ ใด ๆ สมการทั้งสองจะกำหนดคู่ของ ξ และ η โดยแต่ละคู่ของ (ξ, η) สามารถเทียบกันได้ ถ้าให้ความสนใจถึงความคลาดเคลื่อนเชิงสุ่มในการประมาณค่า η และ ξ ตามสมการ (42), (43) จะต้องมีการใช้พารามิเตอร์ข้อสอบ

เพราะค่าเหล่านี้เป็นแหล่งความคลาดเคลื่อนในการปรับเทียบตามแนว IRT

ในการปรับเทียบตามแนว IRT ส่วนมากมักจะใช้ข้อสอบร่วมที่ให้ผู้สอบทุกคนได้ตอบ ทั้งนี้ก็เพื่อให้พารามิเตอร์ข้อสอบของแบบสอบ Y อยู่บนสเกลเดียวกันกับพารามิเตอร์ข้อสอบของแบบสอบ X แบบสอบร่วมที่ใช้คือฉบับ W อาจจะเป็นแบบสอบร่วมภายนอกหรือภายในก็ได้ผู้สอบกลุ่มที่ 1 ทำแบบสอบฉบับ X และ W ส่วนกลุ่มที่ 2 ทำแบบสอบฉบับ Y และ W ในทางปฏิบัติเมื่อใช้แบบสอบหลาย ๆ ฉบับ เช่น ฉบับคือ A,B,...,X,Y,Z,... มักจะทำการวิเคราะห์ทันทีหลังจากสอบแต่ละฉบับเสร็จเพื่อที่จะนำไปปรับเทียบกับฉบับก่อน เมื่อสอบกลุ่มต่อไปเสร็จจะไม่ต้องเสียเวลาที่จะต้องนำข้อมูลก่อนหน้านี้มาวิเคราะห์รวมกันใหม่ กรณีเช่นนี้ก็คือการแยกวิเคราะห์ระหว่างแต่ละกลุ่มซึ่งเป็นกรณีที่นำมาพิจารณาในครั้งนี้ นั่นคือข้อตกลงเบื้องต้นของการประมาณค่าความแปรปรวนเชิงสุ่มที่จะกล่าวต่อไปนี้ ไม่รวมถึงกรณีที่นำคะแนนผู้สอบกลุ่มที่ 1 และกลุ่มที่ 2 มาวิเคราะห์รวมกันและพร้อม ๆ กัน

เมื่อมีการวิเคราะห์หาค่าพารามิเตอร์แยกกันระหว่างกลุ่ม 1 และกลุ่ม 2 พารามิเตอร์ข้อสอบและ θ ในสมการ (43) จะมีจุดเริ่มต้นของสเกลแตกต่างไปจากค่าพารามิเตอร์ในสมการ (42) จึงเป็นไปไม่ได้ที่จำกำจัดค่า θ จากสมการ (42) (43) เพื่อหาค่า η และในลักษณะเช่นนี้ต้องใช้แบบสอบร่วมในการแปลงสเกลของกลุ่มที่ 1 ให้อยู่บนกลุ่มที่ 2 วิธีการเช่นนี้เป็นการเพิ่มความแปรปรวนเชิงสุ่มของการแปลงค่าพารามิเตอร์และเป็นการยุ่งยากในการกำหนดความแปรปรวนเชิงสุ่มของการปรับเทียบที่จะเกิดตามมา วิธีการและสูตรต่าง ๆ ต่อไปนี้จะเป็นการหลีกเลี่ยงปัญหาเนื่องจากไม่มีการแปลงพารามิเตอร์ข้อสอบ

สมการ (42) และ (43) ยังคงไม่เปลี่ยนแปลงยกเว้นตัวห้อย(subscripts) ที่ใช้เท่านั้น โดยเฉพาะอย่างยิ่ง θ_1 และ θ_2 ต้องแตกต่างกันเนื่องจากกลุ่ม 1 และ 2 มีสเกลความสามารถแตกต่างกัน จึงได้

$$\zeta = \sum_g P_{g1}(\theta_1) \qquad (44)$$

$$\eta = \sum_g P_{g4}(\theta_2) \qquad (45)$$

ตอนนี้ฟังค์ชันการตอบสนองรายข้อเขียนเป็น P_{gp} เมื่อ $p = 1,2,3,4$ ซึ่งหมายถึงแบบสอบ X ใช้กับกลุ่มที่ แบบสอบ W ใช้กับกลุ่มที่ 1 แบบสอบ W ใช้กับกลุ่มที่ 2 และแบบสอบ Y ใช้กับกลุ่มที่ 2 ตามลำดับ และ $g = 1,2,\ldots,n_p$ เมื่อ n_p เป็นจำนวนข้อของแบบสอบแต่ละฉบับ

ถ้าให้ ω เป็นคะแนนที่คาดว่าจะตอบได้ถูกสำหรับแบบสอบร่วม W จะได้

$$\omega = \sum_g P_{g2}(\theta_1) \qquad (46)$$

$$\omega = \sum_g P_{g3}(\theta_2) \qquad (47)$$

การที่จะได้ค่าความสัมพันธ์ระหว่าง η และ ξ จะต้องกำจัดค่า θ_1, θ_2 และ ω จากสมการ(44) (45) (46) (47) วิธีการก็คือขั้นที่ 1 เป็นการปรับเทียบระหว่าง ω ไปยังโดยใช้สมการ (44) กับ (46) จากนั้นปรับเทียบ η ไปยัง ω โดยใช้สมการ (47) และ (45) วิธีการนี้เป็นการปรับเทียบ η ไปยัง ξ สำหรับสถานการณ์ที่กลุ่มผู้สอบกลุ่ม 1 และกลุ่ม 2 มีค่าพารามิเตอร์ไม่อยู่บนสเกลเดียวกัน

การประมาณค่าการปรับเทียบจากสมการ (44) ถึง (47) หลังจากที่ได้แทนค่าพารามิเตอร์ข้อสอบที่ได้จากการประมาณโดยวิธี maximum likelihood มีการใช้เครื่องหมาย $\hat{}$ ในสมการ จึงได้

$$\xi = \sum_g \hat{P}_{g1}(\theta_1) \qquad (48)$$

$$\hat{\omega} = \sum_g \hat{P}_{g2}(\theta_1) \qquad (49)$$

$$\hat{\omega} = \sum_g \hat{P}_{g3}(\theta_2) \qquad (50)$$

$$\hat{\eta} = \sum_g \hat{P}_{g4}(\theta_2) \qquad (51)$$

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

สมการเหล่านี้แสดงว่า $\hat{\eta}$ เป็นฟังค์ชันของการประมาณค่าพารามิเตอร์ข้อสอบทั้งหมดที่สมนัยกับค่า ξ

สำหรับข้อสอบข้อ g แทนที่จะใช้ค่า a_g, b_g และ c_g แทน 3 พารามิเตอร์ ที่ใช้กันโดยทั่วไป แต่ใช้ t_{1gp}, t_{2gp} และ t_{3gp} แทนตามลำดับ เราจำเป็นต้องทำการอนุพันธ์ สำหรับ r = 1,2,3, ที่ได้รับจาก (48) - (51)

$$\frac{\partial \eta}{\partial t_{rg4}} = P_{g4}^{(r)}(\theta_2)$$

$$\frac{\partial \omega}{\partial t_{rg3}} = P_{g3}^{(r)}(\theta_2)$$

$$\frac{\partial \omega}{\partial t_{rg2}} = P_{g2}^{(r)}(\theta_1) \qquad (52)$$

เมื่อ P' แทนอนุพันธ์ของ P_{gp} ที่เกี่ยวข้องกับ t_{rgp}

ในทำนองเดียวกันได้

$$\frac{\partial \eta}{\partial \theta_2} = \sum_g P'_{g4}(\theta_2)$$

$$\frac{\partial \omega}{\partial \theta_1} = \sum_g P'_{g2}(\theta_1)$$

เมื่อ P' แทนอนุพันธ์ที่เกี่ยวข้องกับ θ เมื่อใช้สูตรสำหรับการอนุพันธ์ฟังค์ชัน จาก (48-51) สำหรับ r = 1,2,3 ได้

$$\frac{\partial \theta_2}{\partial t_{rg3}} = -\frac{P_{g3}^{(r)}(\theta_2)}{\sum_g P'_{g3}(\theta_2)}$$

$$\frac{\partial \theta_1}{\partial t_{rg1}} = -\frac{P_{g1}^{(r)}(\theta_1)}{\sum_g P'_{g1}(\theta_1)}$$

$$\frac{\partial \theta_2}{\partial \omega} = \frac{1}{\sum_g P'_{g3}(\theta_2)}$$

ใช้กฎลูกโซ่ในการอนุพันธ์(chain rule) จะได้

$$\frac{\partial \eta}{\partial t_{rg3}} = \frac{\partial \eta}{\partial \theta_2}\frac{\partial \theta_2}{\partial t_{rg3}} = -P^{(r)}_{g3}(\theta_2)\frac{\sum_g P'_{g4}(\theta_2)}{\sum_g P'_{g3}(\theta_2)} \qquad (53)$$

$$\frac{\partial \eta}{\partial t_{rg2}} = \frac{\partial \eta}{\partial \theta_2}\frac{\partial \theta_2}{\partial \omega}\frac{\partial \omega}{\partial t_{rg2}} = P^{(r)}_{g2}(\theta_1)\frac{\sum_g P'_{g4}(\theta_2)}{\sum_g P'_{g3}(\theta_2)} \qquad (54)$$

$$\frac{\partial \eta}{\partial t_{rg1}} = \frac{\partial \eta}{\partial \theta_2}\frac{\partial \theta_2}{\partial \omega}\frac{\partial \omega}{\partial \theta_1}\frac{\partial \theta_1}{\partial t_{rg1}} = -P^{(r)}_{g1}(\theta_1)\frac{\sum_g P'_{g2}(\theta_1)}{\sum_g P'_{g1}(\theta_1)}\frac{\sum_g P'_{g4}(\theta_2)}{\sum_g P'_{g3}(\theta_2)} \qquad (55)$$

เมื่อให้ค่า $\mathfrak{B}$ ก็สามารถแสดง $\hat{\eta}$ ที่เป็นชุดของ $\hat{t}_{rgp} - t_{rgp}$ ($r = 1,2,3$ $g = 1,2,\ldots,n_p$; $p = 1,2,3,4$) และเขียน η'_{rgp} แทน $\partial\eta/\partial t_{rgp}$ และ η''_{rgpshq} แทน $\partial^2\eta/\partial t_{rgp}\partial t_{shq}$ ได้

$$\hat{\eta} = \eta + \sum_p\sum_g\sum_r (\hat{t}_{rgp} - t_{rgp})\eta'_{rgp}$$

$$+ \frac{1}{2}\sum_p\sum_q\sum_g\sum_h\sum_r\sum_s (\hat{t}_{rgp} - t_{rgp})(\hat{t}_{shq} - t_{shq})\eta''_{rgpshq} + \ldots \qquad (56)$$

และสุดท้ายจะได้ว่า

$$\mathrm{Var}\,\hat{\eta} \doteq \sum_{p=1}^{4}\sum_{g=1}^{n_p}\sum_{r=1}^{3}\sum_{s=1}^{3}\eta'_{rgp}\eta'_{sgp}\,\mathrm{Cov}(\hat{t}_{rgp},\hat{t}_{sgp}) \qquad (57)$$

<u>การประเมินความเพียงพอของการปรับเทียบคะแนน</u>

การปรับเทียบคะแนนรูปแบบใดก็ตามจะให้ผลดีที่สุดเมื่อ คะแนนที่ได้จากแบบสอบที่นำมาปรับเทียบคะแนน มีคุณสมบัติเป็นไปตามเงื่อนไขต่าง ๆ ที่กำหนดไว้ในรูปแบบการปรับเทียบแต่ละรูปแบบ แต่ในสภาพการณ์จริงอาจมีข้อจำกัดบางประการซึ่งไม่สามารถทำให้ได้ข้อมูลตรงตามเงื่อนไขได้ จึงมีความจำเป็นที่ต้องมีการตรวจสอบความเพียงพอของการเทียบมาตรา ซึ่งเป็นการประเมินประสิทธิภาพของวิธีการปรับเทียบคะแนน ซึ่งมีผู้เสนอแนวความคิดไว้หลายวิธี

ภาวิณี ศรสุขวัฒนานันท์(2528) ได้ประยุกต์ดัชนีในการประเมินความเพียงพอ โดยใช้ดัชนีเปรียบเทียบเปอร์เซ็นไตล์ (the percentile comparison index)ของโคเลนและวิทนีย์ (Kolen and Whitney, 1982) ดัชนีความแตกต่าง(Discrepancy Index) ของปีเตอเซ็นและคณะ(petersen and others, 1982) ตามคำแนะนำของสเตราด์(Stroud, 1982) มาประยุกต์ใช้

แนวคิดของโคเลนและวิทนีย์ที่ใช้ดัชนีเปรียบเทียบเปอร์เซ็นไตล์นั้น เป็นการใช้ข้อมูลคะแนนของผู้สอบเป็นเกณฑ์ในการหาความแตกต่าง ข้อมูลเหล่านี้ได้มาจากกลุ่มสอบทานผล ซึ่งเป็นกลุ่มตัวอย่างที่สุ่มมาจากประชากรเดียวกันกับกลุ่มตัวอย่างปรับเทียบคะแนน และไม่มีหน่วยตัวอย่างซ้ำกันเลย และให้ได้รับการทดสอบด้วยแบบสอบปรับเทียบคะแนนทั้งสองชุด คือฉบับ X และฉบับ Y ให้คะแนนแบบสอบชุด X เป็นคะแนนเกณฑ์แล้วนำคะแนนของแบบสอบฉบับ Y ไปแปลงคะแนนให้อยู่ในมาตราคะแนนของ X คือ X^* ด้วยวิธีการปรับเทียบคะแนนที่ระบุไว้ ณ ตำแหน่งเปอร์เซ็นไตล์เดียวกัน มีสูตรคำนวนค่าดัชนีดังนี้

$$C = \frac{\Sigma(X - X^*)^2}{nk} \qquad (58)$$

เมื่อ n เป็นจำนวนคะแนนดิบของกลุ่มสอบทานผล

k เป็นจำนวนข้อสอบในแบบสอบร่วมที่ใช้

ถ้าค่า C ที่มีค่าน้อย หมายความว่า รูปแบบการปรับเทียบคะแนนที่นำมาสร้างคะแนนสมมูล นั้นมีความเหมาะสมและเพียงพอที่ให้ผลการแปลงคะแนนคงเส้นคงวา

สำหรับแนวคิดของปีเตอร์เซ็นและคณะที่ใช้ดัชนีความแตกต่างนั้น คะแนนเกณฑ์ที่ใช้คือผลการแปลงคะแนนด้วยรูปแบบอิงทฤษฎีการตอบข้อสอบ มีสูตรการคำนวณดัชนี ดังนี้

$$\text{total error} = \frac{\Sigma fd^2}{nS_t^2} \qquad (59)$$

เมื่อ $d = t - t'$

n = จำนวนคะแนนที่ใช้

S_t^2 = ความแปรปรวนของคะแนน t

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

ค่าดัชนีที่ได้มีลักษณะเป็นค่ามาตรฐาน ค่าเหล่านี้สามารถนำมาเปรียบเทียบกันได้โดยตรง ถึงแม้ในสถานการณ์ที่ได้ข้อมูลมาต่างกันก็ตาม

จากแนวความคิดของโคเลนและวิทนีย์ ที่ใช้คะแนนของผู้สอบเองเป็นเกณฑ์ทำให้มีความเป็นอิสระ ไม่ขึ้นกับกระบวนการแปลงคะแนนในรูปแบบอื่น เช่นวิธีของปีเตอร์เซ็นและคณะเสนอ ผู้วิจัยจึงเลือกวิธีการประเมินความเพียงพอจากการวิเคราะห์กลุ่มสอบทานผล โดยใช้ค่าดัชนีความแตกต่างตามแนวคิดของ ภาวิณี ศรีสุขวัฒนานันท์ ซึ่งแปลงมาจากสูตรของ โคเลนไปใช้ตามแนวคิดของปีเตอร์เซ็นและคณะ คือใช้ความแปรปรวนเป็นตัวถ่วงน้ำหนัก เพื่อให้ได้ค่าเป็นมาตรฐาน สูตรที่ใช้คือ

$$C = \frac{\Sigma(X - X^{*})^{2}}{nS_{x}^{2}} \qquad (60)$$

เมื่อ X เป็นคะแนนจากแบบสอบฉบับ X

X^{*} เป็นคะแนนจากแบบสอบฉบับ X ที่ได้จากการนำคะแนนจากแบบสอบฉบับ Y ไปแปลงจากตารางการปรับเทียบคะแนน

n เป็นจำนวนคนในกลุ่มสอบทานผล

S_{x}^{2} เป็นค่าความแปรปรวนของแบบสอบฉบับ X

<u>เกณฑ์ที่นำมาใช้ตรวจสอบ</u>

ในกรณีที่แบบสอบปรับเทียบคะแนนกับตัวมันเอง เกณฑ์ที่ใช้ในการปรับเทียบคือคะแนนที่ได้จากการสอบแบบสอบครั้งแรกและครั้งหลังแบบสอบทั้งสองทำหน้าที่คล้ายกับเป็นแบบสอบสองฉบับที่แตกต่างกัน แต่ที่จริงแล้วเป็นแบบสอบเดียวกัน การปรับเทียบคะแนนในอุดมคติอาจเกิดขึ้นกับคะแนนของแบบสอบฉบับแรก กล่าวคือ การแปลงคะแนนดิบไปเป็คะแนนมาตรอาจเหมือนกันทั้งแบบสอบฉบับก่อนและฉบับหลัง การหาเกณฑ์ที่นำมาใช้ในการตรวจสอบนี้ค่อนข้างเป็นเรื่องลำบาก ซึ่งอาจต้องคำนวณค่าการปรับที่แท้จริงโดยการปรับเทียบแบบสอบเหล่านี้กับลักษณะที่เป็นอุดมคติให้มากที่สุดเท่าที่จะทำได้

คะแนนเกณฑ์ที่ใช้คือ คะแนนแปลงจากคะแนนดิบชุดเดียวกัน ค่านี้ทำให้เป็นมาตรฐานด้วยการทำให้เป็นสัดส่วนของความเบี่ยงเบนมาตรฐานของคะแนนเกณฑ์ แล้วนำไปใช้เปรียบเทียบผลที่ได้จากการวิเคราะห์ในสถานการณ์ที่ต่างกัน ค่าดัชนีหรือค่าความคลาดเคลื่อนรวม (TE) เมื่อส่วนเบี่ยงเบนมาตรฐานของคะแนนเกณฑ์ที่ใช้มีค่าเท่ากับ 100 มีความหมายเชิงคุณภาพของการปรับเทียบคะแนนระหว่างแบบสอบดังนี้

TE < 25	คือ น่าพอใจอย่างมาก
25 ≤ TE < 100	คือ น่าพอใจ
100 ≤ TE < 225	คือ ปานกลาง
225 ≤ TE < 400	คือ ไม่น่าพอใจ
400 ≤ TE	คือ ไม่น่าพอใจอย่างมาก

งานวิจัยที่เกี่ยวข้องกับการปรับเทียบคะแนน

งานวิจัยที่เกี่ยวข้องกับการปรับเทียบมาตราในประเทศมีค่อนข้างน้อย ส่วนใหญ่เป็นงานวิจัยของต่างประเทศซึ่งศึกษาในลักษณะต่าง ๆ กัน ดังนี้

ภาวิณี(2528) ได้ทำการวิจัยเปรียบเทียบผลการปรับเทียบคะแนนระหว่างวิธีอีควิเปอร์เซ็นไทล์ วิธีเชิงเส้นตรง และวิธี IRT 3-พารามิเตอร์ โดยใช้ข้อสอบร่วมที่มีขนาดต่างกัน 3 ขนาดคือร้อยละ 60 ร้อยละ 40 และร้อยละ 20 ในการสอบผลสัมฤทธิ์ทางการเรียนและการสอบคัดเลือกโดยใช้กลุ่มตัวอย่างกลุ่มละ 1500 คน การประเมินผลการปรับเทียบคะแนนใช้การวิเคราะห์หาความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนน และการวิเคราะห์กลุ่มสอบทานผล ผลการวิจัยพบว่า 1) วิธีการปรับเทียบทุกวิธีให้ผลในระดับที่ยอมรับได้ 2) ความยาวของแบบสอบร่วมเป็นปัจจัยที่มีผลกระทบต่อความแม่นยำและความเพียงพอของวิธีการปรับเทียบ 3) สถานการณ์การสอบผลสัมฤทธิ์ทางการเรียนและการสอบคัดเลือกภายใต้วิธีการปรับเทียบเดียวกันให้ผลที่แตกต่างกัน

สุนิสา(2534) ได้วิจัยตรวจสอบคุณภาพของการปรับเทียบคะแนนเชิงเส้นตรง จากการใช้แบบสอบร่วมภายในขนาดความยาวต่างกัน 4 ขนาด คือ 10 15 20 และ 25 ข้อ โดยการเปรียบเทียบความคลาดเคลื่อนมาตรฐานของการเปรับเทียบคะแนน(SEE) และดัชนีความแตกต่าง

(C) จากการวิเคราะห์กลุ่มสอบทานผล เครื่องมือที่ใช้ในการวิจัยเป็นแบบสอบวัดผลสัมฤทธิ์ปลายภาคเรียน ชั้นมัธยมศึกษาปีที่ 2 ภาคต้น จำนวน 2 ฉบับ ฉบับละ 60 ข้อ และแบบสอบร่วมที่ใช้ในการวิจัยจำนวน 25 ข้อ กลุ่มตัวอย่างประกอบด้วย 2 กลุ่ม คือกลุ่มเป็นกลุ่มที่ใช้ในการปรับเทียบคะแนน จำนวน 810 คน และกลุ่มสอบทานผลจำนวน 117 คน ผลการวิจัยพบว่า (1)คะแนนสมมูลของแบบสอบฉบับ Y น้อยกว่าคะแนนจากแบบสอบฉบับ X จากการปรับเทียบคะแนนโดยใช้แบบสอบร่วม 4 ขนาด(2)ค่าความคลาดเคลื่อนมาตรฐานของการปรับเทียบคะแนนเชิงเส้นตรงในแบบสอบร่วมที่ยาวกว่าให้ความคลาดเคลื่อนน้อยกว่า โดยพิจารณาจากค่าความคลาดเคลื่อนมาตรฐานของการใช้แบบสอบร่วม 4 ขนาด ณ ระดับคะแนนมาตรฐานเดียวกัน 7 ระดับ (3) ค่าประสิทธิภาพสัมพัทธ์ที่ได้จากอัตราส่วนร้อยละของความคลาดเคลื่อนมาตรฐาน เมื่อใช้แบบสอบร่วมขนาด 10 15 และ 20 ข้อ เทียบกับแบบสอบร่วม 25 ข้อ ให้ค่า 86.26 89.08 และ 93.30 ตามลำดับ (4) ค่าดัชนีความแตกต่างของการปรับเทียบคะแนน จากการใช้แบบสอบร่วม 4 ขนาด คือ 0.4319 0.3886 0.3630 และ 0.3354 ตามลำดับ ซึ่งแตกต่างกันอย่างมีนัยสำคัญที่ระดับ .05

เรคเคส(Reckase,1979)ได้ทำการศึกษาความสัมพันธ์ระหว่างผลของการประมาณค่าพารามิเตอร์กับขนาดของกลุ่มตัวอย่างและความยาวของแบบสอบร่วมโดยใช้ข้อมูลจากแบบวัด ITED(Iowa Tests of Education Development) ผลการวิจัยพบว่าการเชื่อมโยงจะได้ผลดีกว่าเมื่อขนาดกลุ่มตัวอย่างมีค่าเพิ่มขึ้น การเชื่อมโยงจะให้ผลที่คงที่เมื่อขนาดกลุ่มตัวอย่างเท่ากับ 300 สำหรับราช์โมเดล แต่สำหรับโมเดล 3 พารามิเตอร์ ต้องใช้กลุ่มตัวอย่างน้อยที่สุด 1,000 คน ถ้าขนาดกลุ่มตัวอย่างมีจำนวนน้อยโปรแกรม LOGIST มีแนวโน้มที่จะให้ค่าความยากในลักษณะสุดโต่ง และค่าอำนาจจำแนกจะเข้าใกล้ 0 นอกจากนี้เขายังศึกษาให้เห็นว่าข้อสอบร่วมมีจำนวน 5 ข้อถึง 10 ข้อ ก็มีความเพียงพอเมื่อกลุ่มตัวอย่างมีขนาดใหญ่ คืออย่างน้อยที่สุด 300 คน

โคเวลล์(Cowell,1981) ได้ศึกษาโดยใช้ขนาดกลุ่มตัวอย่างเป็นตัวแปรอิสระข้อมูลที่ใช้เป็นคะแนนจากการสอบ TOEFL เขาได้เปรียบเทียบการปรับเทียบคะแนนตามแนว IRT หลาย ๆ แบบโดยใช้ขนาดกลุ่มตัวอย่างขนาดใหญ่ คือ 2,000-3,000 คน และขนาดเล็กประมาณ 300 คน เกณฑ์ที่ใช้คือคะแนนสมมูลที่แปลงมาจากการปรับเทียบเชิงเส้นตรง ผลพบว่าข้อแตกต่างของการปรับเทียบระหว่างวิธีการปรับเทียบ และขนาดกลุ่มตัวอย่างมีค่าน้อยมาก

แบบสอบที่ใช้มีค่าความยากใกล้เคียงกันและกลุ่มตัวอย่างมีความสามารถเกือบเท่ากัน ผลยังพบว่าการปรับเทียบตามโมเดล IRT 3-พารามิเตอร์ให้ผลอย่างคงที่ถึงแม้กลุ่มตัวอย่างจะมีขนาดเล็ก ผลที่แตกต่างกันระหว่างการใช้กลุ่มตัวอย่างขนาดเล็ก กับกลุ่มตัวอย่างขนาดใหญ่มีค่าน้อยกว่าความแตกต่างจากการใช้โมเดล 1 และ 3 พารามิเตอร์

ฮิวลิน ลิสสัค และ ดราสโกว์(Hulin, lissak, และ Drasgow,1982)ได้ใช้การจำลองข้อมูลเพื่อศึกษา ขนาดของกลุ่มตัวอย่าง ความยาวของแบบสอบ และการประมาณค่าพารามิเตอร์ด้วยโปรแกรม LOGIST โดยมีการเปลี่ยนความยาวแบบสอบและเปลี่ยนขนาดของกลุ่มตัวอย่าง กับ IRT โมเดล 3 พารามิเตอร์ และพารามิเตอร์ความสามารถมีการแจกแจงแบบโค้งปกติ ผลการวิจัยพบว่าความยาวของแบบสอบและขนาดของกลุ่มตัวอย่างมีผลต่อการประมาณค่าเมื่อประมาณค่าพารามิเตอร์ 3 พารามิเตอร์โดยใช้ข้อสอบ 60 ข้อกับกลุ่มผู้สอบจำนวน 1,000 คน พบว่าให้ผลออกมาในลักษณะที่คงที่ ถ้าลดข้อสอบลงครึ่งหนึ่งคือเหลือ 30 ข้อ และเพิ่มขนาดของกลุ่มตัวอย่างเป็นสองเท่าจะให้ผลใกล้เคียงกัน และพบว่าโปรแกรม LOGIST ใช้กับกลุ่มตัวอย่างขนาด 15 คนไม่ได้

คาลด์เวลล์(Caldwell, 1984) ได้ทำการศึกษาเพื่อวัดประสิทธิผลสัมพัทธ์ในโมเดลการปรับเทียบความยากของแบบสอบ โมเดลที่ใช้คือ โมเดลเชิงเส้นตรง แบบแผนที่ 4 และราสช์โมเดล ในการปรับเทียบได้พิจารณาถึง การเพิ่มค่าเฉลี่ยและการลดส่วนเบี่ยงเบนมาตรฐานสำหรับการสอบครั้งที่ 2 ซึ่งคาดว่าโมเดลเชิงเส้นตรงจะมีประสิทธิผลน้อยกว่าราสช์โมเดล เนื่องจากโมเดลเชิงเส้นตรงมีค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานบรรจุอยู่ในสูตร ส่วนราสช์โมเดลไม่มี และคาดว่าประสิทธิผลสัมพัทธ์ของราสช์โมเดลจะเพิ่มขึ้นขณะที่ความแตกต่างของค่าเฉลี่ยเพิ่มขึ้น ได้มีการประเมินแบบสอบร่วมทั้งสองแบบ แบบสอบร่วมชุดหนึ่งประกอบด้วยข้อสอบที่มีความยากในระดับปานกลาง ส่วนอีกชุดหนึ่งมีความยากสูงสุด แบบสอบที่มีค่าความยากสูงนี้ได้รับการพิจารณาว่าเป็นกรณีที่มีประสิทธิผลต่ำสุด และแบบสอบร่วมที่มีความยากปานกลางได้รับการพิจารณาว่าให้ประสิทธิผลที่ดีกว่า ผลการวิจัยพบว่าเมื่อพิจารณาคะแนนทั้งหมด ราสช์โมเดลเหนือกว่าโมเดลเชิงเส้นตรงทั้งในแบบสอบร่วมที่มีความยากระดับปานกลางและระดับสูง แต่เมื่อพิจารณาคะแนนจุดตัดที่แสดงถึงความสามารถต่ำสุด ราสช์โมเดลเหนือกว่าในลักษณะที่ให้ความลำเอียงเพียงเล็กน้อยแต่โมเดลเชิงเส้นตรง ให้ค่าความคลาดเคลื่อนในระดับต่ำ ดัชนีความแตกต่างของราสช์โมเดลมีค่าเฉลี่ยเกือบเป็นศูนย์และมีการแปรเปลี่ยนในช่วงกว้าง

โมเดลเชิงเส้นตรงให้ค่าความลำเอียงด้านลบแต่มีการแปรเปลี่ยนที่น้อยกว่า

ลุย(Liou,1984)ได้ศึกษาเพื่อเสนอโมเดลการแยกองค์ประกอบของความคลาดเคลื่อนของการปรับเทียบคะแนนตามแนว IRT โดยใช้สถานการณ์จำลอง องค์ประกอบของความคลาดเคลื่อนที่นำมาประเมินได้แก่ ความคลาดเคลื่อนเนื่องจากการละเมิดข้อตกลงของโมเดล IRT ความคลาดเคลื่อนเนื่องจากการประมาณค่าพารามิเตอร์ความสามารถ ความคลาดเคลื่อนเนื่องจากการแปลงสเกลความสามารถเชิงเส้นตรงและความคลาดเคลื่อนส่วนที่เหลือ ผลการวิจัยพบว่าโมเดลที่เสนอนี้มีประสิทธิภาพดีพอที่จะเป็นเครื่องมือในการทั้งขนาดและแหล่งของความคลาดเคลื่อนในการปรับเทียบคะแนนตามแนว IRT

สแก็กส์(Skaggs, 1984) ได้ทำการตรวจสอบวิธีการปรับเทียบคะแนน 4 วิธีภายใต้เงื่อนไขที่พารามิเตอร์ข้อสอบและพารามิเตอร์ความสามารถแปรเปลี่ยนอย่างเป็นระบบ วิธีการปรับเทียบที่ใช้ได้แก่ IRT 2 แบบ คือ ราสช์โมเดล กับ โมเดล 3-พารามิเตอร์ และแบบดั้งเดิมอีก 2 แบบ คือการปรับเทียบแบบเส้นตรงและแบบอีควิเปอร์เซ็นไทล์ ข้อมูลที่ใช้ในการศึกษาใช้การจำลองขึ้นมา ตัวแปรที่ใช้ในการศึกษาได้แก่ ค่าความยาก อำนาจจำแนก และระดับโอกาสการเดา และค่าเฉลี่ยของความสามารถเกณฑ์ที่ใช้ในการปรับเทียบคือคะแนนจริงจากโมเดล 3-พารามิเตอร์ ข้อแตกต่างระหว่างวิธีการปรับเทียบทั้ง 4 แบบกับเกณฑ์นำมาวิเคราะห์ หาค่าสถิติเชิงบรรยายและแสดงด้วยกราฟและจุดประสงค์หลักของการศึกษาครั้งนี้เพื่อทดสอบความแกร่งของราสช์โมเดลเมื่อมีการละเมิดข้อตกลง ผลการวิจัยพบว่าราสช์โมเดลยังไม่มีความแกร่งเมื่อมีการละเมิดในข้อตกลง กรณีนี้เป็นจริงเมื่อค่าเฉลี่ยของอำนาจจำแนกมีค่าไม่เท่ากัน และโมเดล 3-พารามิเตอร์ ก็ประสบปัญหาเช่นกันเมื่ออำนาจจำแนกไม่เท่ากัน โดยภาพรวมการปรับเทียบคะแนนแบบอีควิเปอร์เซ็นไทล์ให้ผลดีเกือบทั้งหมด การวิจัยครั้งนี้ได้ให้ข้อเสนอแนะว่าการใช้ราสช์โมเดล และ โมเดล 3-พารามิเตอร์ ต้องใช้ด้วยความระมัดระวัง

ฮิลล์ สับฮิยาห์ และ เฮิร์ช(Hills, Subhiyah และ Hirsch, 1988)ได้ศึกษาเปรียบเทียบวิธีการปรับเทียบคะแนน 5 วิธี คือวิธีการเชิงเส้นตรง และวิธี IRT 4 วิธีการ ได้แก่ วิธี IRT 3 พารามิเตอร์ คือ Concurent method(IRTCON), Fixed-paramater method(IRTFIX),Formular method(IRTFOR) และ ราสช์โมเดล และศึกษาความยาวแบบสอบร่วม 6 ขนาด คือ 30 25 20 15 10 และ 5 ข้อ ซึ่งสุ่มมาจากแบบสอบร่วม

30 ข้อ ว่าแบบสอบร่วมทั้ง 5 ขนาด มีประสิทธิภาพเทียบเท่ากับแบบสอบร่วม 30 ข้อ โดยการใช้การปรับเทียบคะแนนแบบ Concurrent IRT

ข้อมูลได้จากการสอบนักเรียนในรัฐฟลอริดาปี 1986 1984 ตามโครงการสอบของ Florida's statewide Student Assesment Test Part II แบบสอบที่ใช้เป็นแบบวัดความสามารถขั้นต่ำ (Minimum-Competency Test) ใช้กับโรงเรียนมัธยมศึกษาตอนปลาย แบบสอบประกอบด้วยแบบสอบฉบับย่อยเป็นแบบทดสอบสายภาษา และสายคณิตศาสตร์ การศึกษาใช้แบบสอบปี 1984 เป็นเกณฑ์ และใช้แบบสอบปี 1986 เทียบไปสู่แบบสอบปี 1984 กลุ่มตัวอย่างเป็นนักเรียนระดับ 9-11 ของโรงเรียนในรัฐฟลอริดาจำนวน 6,000 คน สอบแบบสอบในแต่ละปีปีละ 3,000 คน ซึ่งทำทั้งแบบสอบย่อยสายภาษาและคณิตศาสตร์

ผลการประเมินวิธีการปรับเทียบคะแนนทั้ง 5 วิธี ไม่มีการสรุปอย่างเป็นแบบแผน เพราะได้นำผลการปรับเทียบคะแนนเชิงเส้นตรงมาเป็นพื้นฐานในการเปรียบเทียบ นั่นคือเมื่อปรับเทียบคะแนนในแต่ละวิธีแล้ว นำมาหาความแตกต่างของคะแนนที่เทียบตามวิธีการต่าง ๆ กับคะแนนที่ได้จากการปรับเทียบแบบเส้นตรง แต่ผู้วิจัยได้อภิปรายผลว่า ในโปรแกรมการสอบทั่ว ๆ ไปวิธีการ IRT ไม่ใช่วิธีการเดียวที่จะใช้ในการพัฒนาข้อสอบและปรเมินข้อสอบ แต่วิธีการเชิงเส้นตรงก็เป็นวิธีการที่ดีที่ควรนำมาใช้ เป็นวิธีการที่รู้จักกันอย่างกว้างขวาง สำหรับสถานการณ์ของการทดสอบความสามารถขั้นต่ำนี้ วิธีการแบบ IRT ก็สามารถใช้ได้กับการกระจายที่มีความเบ้มาก และวิธีการ IRTCON ยังให้ผลที่ใกล้เคียงกัน ซึ่งวิธีการปรับเทียบแบบราสช์จะง่ายกว่า

ในเรื่องจำนวนข้อกระทงของแบบสอบร่วมที่ใช้ในการปรับเทียบ ข้อสอบร่วมตั้งแต่ 10 ข้อขึ้นไป ที่สุ่มจาก 30 ข้อ โดยการปรับเทียบคะแนนแบบ IRTCON มีประสิทธิภาพเพียงพอในการปรับเทียบคะแนนเท่ากับแบบสอบร่วมขนาด 30 ข้อ ยกเว้นแบบสอบร่วม 5 ข้อที่ไม่มีประสิทธิภาพเท่า ผู้วิจัยได้อภิปรายผลว่า ผลที่ได้นี้ต่างจากงานวิจัยของวิงเกอร์สกี และ ลอร์ด (Wingersky และ Lord, 1984) ราจู เอ็ดเวิร์ด และ ออสเบิร์ก (Raju, Edwards และ Osberg,1983) และ ราจู โบด ลาเซน และสเตนเฮิร์ท(Raju, Bode, Larsen และ Steinhaus,1986)ซึ่งชี้ให้เห็นว่าแบบสอบร่วมขนาด 5-6 ข้อ ก็เพียงพอในการปรับเทียบคะแนนด้วยวิธีการ IRT 3 พารามิเตอร์ และผู้วิจัยได้ชี้ให้เห็นว่าเมื่อใช้รูปแบบ IRTCON เป็นวิธีการปรับเทียบคะแนน ที่ผู้สร้างแบบสอบสามารถลดจำนวนข้อสอบร่วมได้มาก การใช้ข้อสอบร่วมจำนวนน้อยเป็นการรักษาการสอบที่เป็นความลับได้ดี

เมียโอ(Miao,1989) ได้ทำการวิจัยเกี่ยวกับโมเดลของการปรับเทียบคะแนนที่มีแบบแผนต่าง ๆ กัน สำหรับราสช์โมเดลการประเมินประสิทธิผลของการปรับเทียบใช้ค่า item drift ซึ่งเป็นการหาความแตกต่างระหว่างคะแนนเริ่มแรกและคะแนนตอนสุดท้าย การวิจัยครั้งนี้เป็นการหาทางเลือกใหม่โดยใช้การประมาณค่าแบบ conditional maximun likelihood กับการปรับเทียบคะแนนตามแนวราสช์โมเดลเพื่อกำจัด item drift ข้อมูลที่ใช้ในการวิจัยเป็นคะแนนจากแบบสอบการจัดตำแหน่งในวิชาการอ่านและการฟังภาษาสเปน แล้วปรับเทียบคะแนนตามราสช์โมเดล จากนั้นใช้วิธีการใหม่ที่มีแบบสอบร่วมและนำผลจากการประมาณค่าพารามิเตอร์ไปใช้ประโยชน์ การปรับเทียบตามแนวราสช์โมเดลและวิธีการใหม่นำมาเปรียบเทียบกันโดยพิจารณาจากค่าดัชนีความแตกต่าง ผลการวิจัยพบว่า 1)ถ้ามีการละเมิดเงื่อนไขด้านความยากและความสามารถที่เท่ากัน วิธีการใหม่ให้ค่าความลำเอียง(square bias) ใกล้เคียงกับราสช์โมเดล แต่วิธีการใหม่มีความแปรปรวนของความแตกต่างน้อยกว่า

ทริสคาริ(Triscari, 1990)ได้ทำการวิจัยเปรียบเทียบวิธีการปรับเทียบคะแนน 5 วิธี โดยใช้แบบแผนการเก็บรวบรวมข้อมูลแบบความไม่เท่าเทียมกันของกลุ่มประชากร (nonequivalent population) และใช้ข้อสอบร่วมแบบภายในโดยมีการเชื่อมโยงข้อสอบแบบเลือกตอบและแบบเรียงความ วิธีการปรับเทียบที่ใช้คือตามแนวทฤษฎีการวัดแบบดั้งเดิมคือ แบบเส้นตรงและแบบอีควิเปอร์เซ็นไทล์ และแบบ IRT ชนิด 1, 2 และ 3 พารามิเตอร์ ซึ่งใช้การให้คะแนนตามแนว partial credit model การเปรียบเทียบใช้วิธีการจำลองสถานการณ์เมื่อทราบค่าคะแนนจริงที่เท่าเทียมกันของแบบสอบที่เป็นฐาน การประเมินประสิทธิผลใช้ค่าความแตกต่างระหว่างคะแนนจริงของแบบสอบฐาน นอกจากนี้ยังได้เก็บข้อมูลจริงเพื่อหาความสอดคล้องกับข้อมูลที่จำลองขึ้นมาด้วย ผลการวิจัยจากการจำลองข้อมูลพบว่า การปรับเทียบตามแนว IRT ที่ใช้ partial credit model และมีการวิเคราะห์รวมให้ความคลาดเคลื่อนน้อยที่สุด และการปรับเทียบแบบเส้นตรงให้ความคลาดเคลื่อนมากที่สุดจากการเก็บข้อมูลจริงพบว่ามีความสอดคล้องกับการจำลองข้อมูลร้อยละ 70

กลอวากิ (Glowacki, 1991)ได้ตรวจสอบโมเดลของการปรับเทียบคะแนนที่มีความเหมาะสมกับการสอบของบัณฑิตวิทยาลัยแห่งมหาวิทยาลัยอลาบามา ปัญหาในการวิจัยคือโมเดลของการปรับเทียบที่ตรวจสอบ มีการแจกแจงคะแนนดิบ หรือคะแนนที่ผ่านจากการสอบแบบสอบการอ่าน และคณิตศาสตร์ แตกต่างกันหรือไม่ โมเดลที่ใช้ในการตรวจสอบคือ โมเดล

เชิงเส้นตรง อีควิเปอร์เซ็นไทล์ และ IRT ชนิด 1,2 และ 3 พารามิเตอร์ ผลการวิจัยพบว่า วิธีการปรับเทียบคะแนนทั้ง 5 โมเดล ในการสอบการอ่านและคณิตศาสตร์ให้ผลที่คล้ายคลึงกัน แสดงว่าโมเดลทั้งหมดสามารถนำมาใช้กับการปรับเทียบคะแนนได้ โดยไม่มีโมเดลใดที่ดีที่สุด

อเยอร์เว(Ayerve, 1992)ได้ทำการเปรียบเทียบประสิทธิผลการปรับเทียบคะแนนตามแนวดิ่งโดยใช้วิธีอีควิเปอร์เซ็นไทล์ และโมเดล IRT ชนิด 3 พารามิเตอร์ โดยเปลี่ยนขนาดของกลุ่มตัวอย่าง ความยาวแบบสอบ และความยาวของแบบสอบร่วม ประเด็นที่ทำการเปรียบเทียบได้แก่ ก)ประสิทธิผลในเงื่อนไขที่กำหนดทั้งหมด ข)ประสิทธิผลในแต่ละวิธีของทุก ๆ ตัวแปรอิสระ และ ค)ตรวจสอบผลของทุกตัวแปรอิสระในแต่ละวิธีภายใต้เงื่อนไขที่เปลี่ยนแปลง กลุ่มตัวอย่างที่ใช้ 3 ขนาด คือ 200 500 และ 1000 ความยาวแบบสอบใช้ 2 ขนาด คือ 30 และ 60 ข้อ และความยาวของแบบสอบร่วม 2 ขนาดคือ 5 และ 10 ข้อ การวิเคราะห์ข้อมูลใช้ Wieghted Mean Square Error และ Unwieghted Mean Square Error การวิจัยครั้งนี้ใช้วิธีการจำลองสถานการณ์ ผลการวิจัยพบว่าโดยส่วนรวมแล้ววิธีการปรับเทียบแบบอีควิเปอร์เซ็นไทล์และวิธี IRT ไม่แตกต่างกันอย่างมีนัยสำคัญ สำหรับวิธีอีควิเปอร์เซ็นไทล์ ความยาวของแบบสอบและความยาวของแบบสอบร่วมเป็นปัจจัยที่สำคัญ ขณะที่ตามวิธี IRT ขนาดของกลุ่มตัวอย่างเป็นปัจจัยสำคัญ คือ กลุ่มตัวอย่าง 200 คน มีแนวโน้มที่จะให้ผลไม่มีความเที่ยงตรง ขณะที่กลุ่มตัวอย่างขนาดใหญ่คือ 500 และ 1000 คน ให้ผลที่เที่ยงตรงกว่า

จากงานวิจัยที่รวบรวมมานี้ แสดงให้เห็นว่าในการศึกษาเกี่ยวกับการปรับเทียบคะแนนระหว่างแบบสอบเป็นไปได้หลายแนว เช่นศึกษาเปรียบเทียบรูปแบบการปรับเทียบคะแนนศึกษาความยาวของแบบสอบร่วม หารูปแบบที่เหมาะสม ส่วนงานวิจัยที่ใช้ขนาดของกลุ่มตัวอย่างเป็นตัวแปรมีเพียง 2 เรื่อง(Cowell, 1981 และ Ayerve, 1982) ซึ่งได้ผลไปในทำนองเดียวกันว่ากลุ่มตัวอย่างขนาดใหญ่จะให้ผลการปรับเทียบที่ดีกว่า และในรูปแบบ IRT การปรับเทียบตามโมเดล 3-พารามิเตอร์ ต้องการกลุ่มตัวอย่างที่มากกว่าตามโมเดล 1 และ 2 พารามิเตอร์ ส่วนในเรื่องความยาวของแบบสอบร่วมได้ผลไปในลักษณะเดียวกันว่า จำนวนข้อสอบร่วมที่มากกว่าจะให้ผลการปรับเทียบที่เที่ยงตรงกว่า และในเรื่องวิธีการปรับเทียบคะแนนนั้นยังได้ข้อสรุปที่ขัดแย้งกันอยู่ ทั้งนี้ขึ้นอยู่กับสถานการณ์ที่ทำการศึกษา