

คุณภาพของการเชื่อมโยงคะแนนแนวตั้งแบบพหุมิติสำหรับแบบทดสอบ
รูปแบบผสมด้วยวิธีโค้งลักษณะแบบทดสอบ
The Quality of Multidimensional Vertical Scaling for Test
with Mixed Item Types on Test Characteristic Function Method

ศิวะทัศน์ สุขสุวรรณ (Siwatussana Suksuwan)*

ดร. พัทธี จันทรเพ็ง (Dr. Putcharee Junpeng)**

ดร. สังวรณัฏฐ์ งัดกระโทก (Dr. Sungworn Ngudgratoke)***

บทคัดย่อ

การประเมินผลการเรียนรู้ในปัจจุบันนิยมใช้แบบทดสอบรูปแบบผสมซึ่งวัดคุณลักษณะได้หลากหลาย โดยดำเนินการภายใต้ทฤษฎีการตอบสนองข้อสอบแบบพหุมิติรวมทั้งการเชื่อมโยงคะแนนด้วย ซึ่งปัจจัยสำคัญที่ส่งผลต่อคุณภาพของการเชื่อมโยงคะแนนคือข้อสอบรวมและโครงสร้างมิติความสามารถ ดังนั้นการศึกษาครั้งนี้จึงมีวัตถุประสงค์เพื่อศึกษาคุณภาพของการเชื่อมโยงคะแนนแนวตั้งแบบพหุมิติสำหรับแบบทดสอบรูปแบบผสมด้วยวิธีโค้งคุณลักษณะแบบทดสอบภายใต้โครงสร้างมิติความสามารถและสัดส่วนน้ำหนักคะแนนของข้อสอบรวมที่ต่างกัน โดยศึกษาทั้ง 6 เงื่อนไข (2X3) ประกอบด้วย สัดส่วนน้ำหนักคะแนนข้อสอบรวม 3 สัดส่วนคือ 20% 25% และ 30% และโครงสร้างมิติความสามารถ 2 ลักษณะ คือ แบบไม่ซับซ้อน และแบบซับซ้อน ใช้การวิจัยเชิงทดลองโดยอาศัยหลักการมอนติคาร์โลเพื่อออกแบบการเก็บข้อมูลสำหรับผู้สอบกลุ่มไม่เท่าเทียมกันโดยใช้ข้อสอบรวมกับกลุ่มตัวอย่าง 3,000 คน มี 3 ขั้นตอน คือ 1) การสร้างค่าพารามิเตอร์ ที่แท้จริงของข้อสอบและแบบแผนการตอบของผู้สอบแต่ละระดับ 2) การประมาณค่าพารามิเตอร์ข้อสอบ และ 3) การเชื่อมโยงคะแนน เกณฑ์ที่ใช้เปรียบเทียบคุณภาพการเชื่อมโยงคะแนน คือ ดัชนีรากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย ดัชนีความเอนเอียง และการวิเคราะห์ความแปรปรวนแบบการวัดซ้ำ ผลการวิจัยพบว่า โครงสร้างมิติความสามารถและสัดส่วนน้ำหนักคะแนนข้อสอบรวมไม่ส่งผลต่อคุณภาพของการเชื่อมโยงคะแนน แต่โครงสร้างมิติความสามารถแบบไม่ซับซ้อนและสัดส่วนน้ำหนักคะแนนข้อสอบรวม 25% ให้คุณภาพการเชื่อมโยงคะแนนดีกว่าเงื่อนไขอื่นๆ ซึ่งเป็นประโยชน์ในการเลือกใช้สัดส่วนข้อสอบรวมและโครงสร้างมิติความสามารถให้เหมาะสมสำหรับการเชื่อมโยงคะแนนแนวตั้งแบบพหุมิติในสถานการณ์จริง

Abstract

Numerous assessments contain a mixture of multiple choice and constructed response item types, which are found to measure more than one trait. Thus, there is a need for multidimensional dichotomous and polytomous item response theory modelling solutions, including multidimensional linking

คำสำคัญ : ทฤษฎีการตอบสนองข้อสอบแบบพหุมิติ แบบทดสอบรูปแบบผสม การปรับเทียบคะแนนแบบทดสอบ

Key Words: Multidimensional Item Response Theory, Mixed Format Test, Test Equating

* นักศึกษาระดับปริญญาเอก สาขาวิชาการวัดและประเมินผลการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยขอนแก่น

** อาจารย์ประจำสาขาวิชาการวัดและประเมินผลการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยขอนแก่น

*** รองผู้อำนวยการสถาบันวิจัยและพัฒนา มหาวิทยาลัยสุโขทัยธรรมาธิราช



scores. Common items are most important for linking scores. Practitioners need empirical data to inform their selection of the number common items' for linking scores. Previous research conducted in tests composed exclusively of multiple-choice items demonstrated that common-item sets should be representative of the overall test. The extent to which the number of common items in a mixed-format test differed measured different aspects of the construct. The purpose of this experimental study was study the quality of linking method according to MIRT model by test characteristic function method--under testing condition of differences of dimensional structures and proportion of common item sets. There were 6 conditions (2X3) consisted of dimensional structure -- approximate simple structure and mixed structure, and 3 condition of proportion common items with mixed format test -- 20%, 25% and, 30% of total items, respectively. This research was based on data simulation using Monte Carlo Method and included 3,000 examinees. Data simulation consisted of three steps - generating true item parameters and response patterns for each grade level, calibrating multidimensional item response theory parameter model, and equating the procedure of linking. Root-mean squared error (RMSE) and BIAS were used as criteria to compare the quality of linking scores in each condition. Results showed that there was not an interaction between dimensional structure and the proportion of common item set which affected on quality of linking scores. The proportion of common items set and dimensional structure affected the difference between the stability and the accuracy of linking method. In other words, the difference in quality linking method depended on proportion of common items and dimensional structure, especially the condition of 25% and approximate simple structure which approximated to multidimensional linking scores.

ความเป็นมาและความสำคัญของปัญหา

การวัดทางด้านการศึกษาและจิตวิทยาในปัจจุบันแบบทดสอบที่สร้างขึ้นมักจะใช้แบบทดสอบรูปแบบผสม (Cao, 2008) และสามารถวัดคุณลักษณะได้มากกว่าหนึ่งคุณลักษณะ (Min, 2007) และหากเรายังใช้กระบวนการวัดและการวิเคราะห์ข้อสอบตามทฤษฎีการตอบสนองข้อสอบ (Unidimensional Item Response Theory: UIRT) โดยมีความเชื่อว่าข้อสอบที่สร้างขึ้นนั้นวัดมิติเชิงเดียว (Single Dimension) จะทำให้ฝ่าฝืนข้อตกลงเบื้องต้นได้ง่าย จึงได้มีนักวิชาการด้านการวัดผลการศึกษาได้พยายามคิดทฤษฎีการทดสอบเพื่อแก้ปัญหาที่เกิดขึ้นเรียกว่า “ทฤษฎีการตอบสนองข้อสอบแบบพหุมิติ” (Multidimensional Item Response Theory: MIRT) ขึ้นโดยโมเดล MIRT ถือว่าคุณลักษณะแฝงของบุคคลมากกว่าหนึ่งองค์ประกอบส่งผลต่อการตอบข้อสอบ ดังนั้นพารามิเตอร์ความสามารถของผู้สอบจึงมีตั้งแต่สองพารามิเตอร์ขึ้นไป การทดสอบมีแนวโน้มเพิ่มขึ้นทั้งจำนวนผู้สอบและความถี่ของการทดสอบ จึงทำให้เกิดข้อคำถามว่า

ทำอย่างไรถึงจะได้สารสนเทศเกี่ยวกับผู้สอบที่ถูกต้องมากที่สุดเพื่อนำไปสู่การตัดสินใจที่ดีที่สุด และที่สำคัญทำอย่างไรจึงจะรักษาความลับเกี่ยวกับแบบทดสอบไว้ได้ (Kolen & Brennan, 2004) ดังนั้นจึงมีความจำเป็นต้องใช้แบบทดสอบหลายฉบับในการวัดครั้งเดียวกันหรือใช้ในเวลาที่แตกต่างกัน ซึ่งแบบทดสอบจะต้องมีคุณสมบัติที่คู่ขนานกัน แต่ในความเป็นจริงกระทำได้ยากมากและต้องลงทุนใช้ทรัพยากรมากอาจไม่คุ้มทุนหรือทันกาล ในทางปฏิบัตินิยมแก้ปัญหาด้วยการใช้แบบทดสอบต่างฉบับที่มีความแตกต่างกัน จึงมีความจำเป็นที่จะต้องใช้เทคนิคการเชื่อมโยงคะแนนระหว่างแบบทดสอบ (Test Linking Scores) เพื่อทำแบบทดสอบต่างฉบับให้อยู่ในมาตรฐานเดียวกันและสามารถเปรียบเทียบคะแนนกันได้ และการเชื่อมโยงคะแนนตามทฤษฎีการตอบสนองข้อสอบแบบพหุมิตินี้อยู่หลายวิธีด้วยกันแต่วิธีที่มีการนำมาประยุกต์ใช้กับแบบทดสอบรูปแบบผสมคือวิธีโค้งลักษณะแบบทดสอบ (Test Characteristic Function: TCF)

สำหรับปัจจัยที่ส่งผลต่อคุณภาพของการเชื่อมโยงคะแนนนั้นมียุหลายปัจจัยด้วยกัน เช่น โครงสร้างมิติความสามารถ (Min, 2007) และข้อสอบรวม (Cao, 2008) โดยเฉพาะสัดส่วนข้อสอบรวมเท่าใดจึงจะเหมาะสมต่อการเชื่อมโยงคะแนนตามโมเดลการตอบสนองข้อสอบแบบพหุมิติ สำหรับการวิจัยที่ผ่านพบว่า ข้อสอบรวมจำนวน 20% เหมาะสำหรับการเชื่อมโยงคะแนนตามโมเดลการตอบสนองข้อสอบแบบเอกมิติ (Kolen and Brennan, 2004) และ 20%-25% เหมาะสำหรับการเชื่อมโยงคะแนนตามโมเดลการตอบสนองข้อสอบแบบพหุมิติ (Reckase, 2009) แต่การวิจัยดังกล่าวศึกษาเกี่ยวกับข้อสอบที่มีการตรวจให้คะแนน 2 ค่า เท่านั้น สำหรับข้อสอบรวมรูปแบบผสมนั้น Yao and Boughton (2009) ได้ทำการศึกษาคุณภาพการเชื่อมโยงคะแนนในแนวนอนแต่ยังไม่ได้ศึกษาในประเด็นของข้อสอบรวมและโครงสร้างมิติความสามารถ ดังนั้นคณะผู้วิจัยจึงสนใจที่จะศึกษาคุณภาพของการเชื่อมโยงคะแนนแนวตั้งแบบพหุมิติสำหรับแบบทดสอบรูปแบบผสมอันจะเป็นประโยชน์สำหรับการตัดสินใจเลือกใช้สัดส่วนน้ำหนักคะแนนข้อสอบรวมและโครงสร้างมิติความสามารถสำหรับการเชื่อมโยงคะแนนตามทฤษฎีการตอบสนองข้อสอบแบบพหุมิติในสถานการณ์จริงต่อไป

วัตถุประสงค์การวิจัย

เพื่อศึกษาคุณภาพของการเชื่อมโยงคะแนนแนวตั้งแบบพหุมิติสำหรับแบบทดสอบรูปแบบผสมภายใต้โครงสร้างมิติความสามารถและสัดส่วนน้ำหนักคะแนนของข้อสอบรวมที่ต่างกัน

วิธีดำเนินการวิจัย

การศึกษาค้นคว้าครั้งนี้ใช้รูปแบบการวิจัยเชิงทดลอง โดยศึกษากับ 6 เงื่อนไข (2X3) ประกอบด้วย สัดส่วนน้ำหนักคะแนนข้อสอบรวม 3 สัดส่วนคือ 20% 25% และ 30% และ โครงสร้างมิติความสามารถ 2 ลักษณะ คือ แบบไม่ซับซ้อน และแบบซับซ้อน โดยอาศัยหลักการมอนติคาร์โลเพื่อออกแบบการเก็บข้อมูลสำหรับผู้สอบกลุ่มไม่เท่าเทียมกัน โดยใช้ข้อสอบรวมกับกลุ่มตัวอย่าง 3,000 คน ซึ่งมีรายละเอียดดังนี้

1. ตัวแปรที่ใช้ในการวิจัย

1) ตัวแปรอิสระ มี 2 ตัวแปร ประกอบด้วย

(1) สัดส่วนน้ำหนักคะแนนของข้อสอบรวมต่อข้อสอบทั้งฉบับ จำนวน 3 สัดส่วน คือ 20% 25% และ 30%

(2) มิติโครงสร้างความสามารถมี 2 ลักษณะ คือ มิติโครงสร้างมีไม่ซับซ้อน (APSS) และซับซ้อน (MS)

2) ตัวแปรตาม มี 1 ตัวแปร คือ คุณภาพของการเชื่อมโยงคะแนนโดยพิจารณา 2 ส่วน คือ ความถูกต้องของการเชื่อมโยงคะแนนพิจารณาจากดัชนีค่าความเอนเอียง (Bias) และความคงที่ของการเชื่อมโยงคะแนนพิจารณาจากดัชนีค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE)

2. เงื่อนไขที่ใช้ในการจำลองข้อมูล

เงื่อนไขที่ใช้ในการจำลองข้อมูลของการศึกษาในครั้งนี้ประกอบด้วย การเชื่อมโยงคะแนนวิธีโค้งคุณลักษณะแบบทดสอบ (Test characteristic Function: TCF) มีลักษณะคล้ายกับวิธีการเชื่อมโยงคะแนนตามโมเดล UIRT การเชื่อมโยงคะแนนของข้อสอบที่ตรวจให้คะแนนแบบ 2 ค่าตามโมเดล M-3PL สามารถเขียนเป็นสมการได้ดังนี้

$$\vec{a}_i^* = A^{-1} \vec{a}_i \quad (1)$$

$$\vec{d}_i^* = \vec{d}_i + \vec{a}_i^T A^{-1} \vec{B} \quad (2)$$

$$\vec{c}_i^* = \vec{c}_i \quad (3)$$

$$\text{หรือ } d_{\delta ki}^* = d_{\delta ki} + \vec{a}_i^T A^{-1} \vec{B} \quad \text{เมื่อ } k = 1, \dots, K_i \text{ สำหรับโมเดล M-2PPC} \quad (4)$$

$$\text{ดังนั้น } \vec{\theta}_j^* = A^T \vec{\theta}_j + \vec{B} \quad (5)$$

$$\text{โดยที่ } P_{ijk}(\vec{\theta}_j^*, \vec{\beta}_i^*) = P_{ijk}(\vec{\theta}_j, \vec{\beta}_i) \quad (6)$$



เมื่อ A คือ เมตริกซ์การหมุนแกน คือ พารามิเตอร์ที่สัมพันธ์กับค่าความยาก $\hat{d}_i$ คือ เวกเตอร์ของพารามิเตอร์อำนาจจำแนก C_i คือ พารามิเตอร์ค่าการเดา $\bar{B}$ คือ เวกเตอร์การแปลงคะแนน คือ เวกเตอร์ความสามารถของผู้สอบคนที่ i P_{ijk} คือ ความน่าจะเป็นของพารามิเตอร์การแปลงคะแนนที่ได้รับในข้อสอบข้อที่ i

การศึกษาตามโมเดล MIRT ประเด็นหนึ่งที่น่าสนใจคือโครงสร้างมิติความสามารถซึ่งโดยทั่วไปจะประกอบไปด้วย 3 ลักษณะ คือ โครงสร้างอย่างง่าย (Simple Structure: SS) โครงสร้างที่ไม่ซับซ้อน (Approximate Simple Structure: APSS) และโครงสร้างที่ซับซ้อน (Mixed Structure: MS) แต่การศึกษาครั้งนี้เลือกศึกษาเฉพาะโครงสร้างที่ไม่ซับซ้อนและโครงสร้างที่ซับซ้อนกับกลุ่มตัวอย่างจำนวน 3,000 คน เพราะมีจำนวนที่มีความเพียงพอและให้ผลการประมาณค่าพารามิเตอร์ที่ถูกต้องและคงที่สำหรับโมเดล MIRT (Yao & Boughton, 2009) โดยออกแบบการเก็บรวบรวมข้อมูลรูปแบบผู้สอบกลุ่มไม่เท่าเทียมกันโดยใช้ข้อสอบร่วม แต่จากการศึกษาที่ผ่านมายังไม่มีผลการศึกษาที่ชัดเจนเกี่ยวกับจำนวนข้อสอบร่วม (Reckase, 2009) ดังนั้นคณะผู้วิจัยจึงเลือกศึกษาจำนวนข้อสอบร่วมในมิติของสัดส่วนน้ำหนักคะแนนของข้อสอบร่วมต่อข้อสอบทั้งฉบับจำนวน 3 สัดส่วน คือ 20% 25% และ 30% นอกจากนี้คณะผู้วิจัยได้ใช้สัดส่วนของคะแนนระหว่างข้อสอบของแบบทดสอบรูปแบบผสม (Mixed-Format Test) 70%:30% ทั้งนี้เพื่อให้สอดคล้องกับการศึกษาของ Yao and Boughton (2005)

3. การศึกษาการจำลองข้อมูล

1) การสร้างค่าพารามิเตอร์ที่แท้จริงและแบบแผนการตอบข้อสอบของผู้สอบ

ในการสร้างค่าพารามิเตอร์ที่แท้จริงค่าอำนาจจำแนก (MDISC) และค่าความยาก (MDIFF) นั้นคณะผู้วิจัยเลือกใช้ค่าที่นำเสนอโดย Yon (2006) เนื่องจากเป็นค่าที่สอดคล้องกับการศึกษาในครั้งนี้ โดยมีการแบ่งเป็นค่าที่ได้จากแบบทดสอบในระดับที่ต่ำกว่ากับค่าที่ได้จากแบบทดสอบในระดับที่สูงกว่า ซึ่งมีค่าเฉลี่ยเท่ากับ 1.0 (0.6 0.8 1.0 1.2 และ 1.3) ซึ่งเป็นค่าที่คำนวณได้จากสมการ

$$MDISC_i = \sqrt{\sum_{k=1}^m a_{ik}^2} \quad (7)$$

เมื่อ $MDISC_i$ แทน ค่าอำนาจจำแนกรวมของข้อสอบข้อที่ i ใน m มิติ

m แทน จำนวนของมิติใน

θ -Space

a_{ik} แทน ส่วนประกอบของ $\hat{d}_i$

เวกเตอร์ของมิติที่ k

จากหลักการดังกล่าวผู้วิจัยจึงกำหนดให้ขนาดของมุมในแต่ละระดับมีค่าเท่ากันจากการสุ่มแบบการกระจายเอกรูป (Uniform Distribution) โดยใช้โปรแกรม MATLAB เนื่องจากการกำหนดขนาดของมุมมีผลต่อค่าอำนาจจำแนกของข้อสอบโดยตรง ซึ่งการศึกษาครั้งนี้ถ้าอยู่ในจำนวนข้อเดียวกันและมีโครงสร้างเดียวกัน จะมีขนาดของมุมเท่ากันทุกระดับตั้งแต่ระดับที่ 1-2

สำหรับชุดของพารามิเตอร์ค่าที่สัมพันธ์กับความยาก (d_i) สำหรับข้อสอบที่ตรวจให้คะแนน 2 ค่าของการศึกษาในครั้งนี้ คณะผู้วิจัยเลือกใช้ค่าที่นำเสนอโดย Yon (2006) เนื่องจากเป็นค่าที่สอดคล้องกับการศึกษาในครั้งนี้ โดยมีการแบ่งเป็นค่าที่ได้จากแบบทดสอบในระดับที่ต่ำกว่ากับค่าที่ได้จากแบบทดสอบในระดับที่สูงกว่า สำหรับค่าความยากสำหรับผู้สอบในระดับที่สูงกว่ามีค่าเฉลี่ย 0.9 (-0.5 2.0 1.0 0.0 และ 2.2) ส่วนค่าความยากสำหรับผู้สอบในระดับที่ต่ำกว่ามีค่าเฉลี่ย 1.5 (0.0 2.5 1.5 0.5 และ 3.0) ซึ่งเป็นค่าที่คำนวณได้จากสมการ

$$MDIFF_i = \frac{-d_i}{MDISC_i} \quad (8)$$

โดย $MDIFF_i$ แทน ค่าความยากรวมของข้อสอบข้อที่ i

$MDISC_i$ แทน ค่าอำนาจจำแนกรวมของข้อสอบข้อที่ i

d_i แทน ระยะทางจากจุดตำแหน่งมิติของ θ ซึ่งเป็นจุดที่มีความชันสูงสุด

จากข้อมูลข้างต้นจะพบว่าค่าความยาก (MDIFF) ของผู้สอบในระดับที่ 1 ทำแบบทดสอบในระดับที่ 2

จะมากกว่าผู้สอบระดับที่ 2 ทำแบบทดสอบในระดับที่ 2 เพื่อใช้สำหรับการแปลงคะแนนจากระดับที่ 1 (L1) ไปยังระดับที่ 2 (L2) สำหรับค่าพารามิเตอร์ความยากแต่ละระดับขั้นการตอบ ($d_{\delta_k i}$) ของข้อสอบที่ตรวจให้คะแนนมากกว่า 2 ค่าแต่ละข้อในแต่ละระดับ คณะผู้วิจัยได้สร้างจากการแจกแจงปกติแบบ 2 ตัวแปร โดยมีค่าเฉลี่ยเท่ากับ 0 ส่วนเบี่ยงเบนมาตรฐานเท่ากับ 1

$$P_{ij1} = P(x_{ij} = 1/a_i, c_i, d_i, \theta_j) = c_i + (1 - c_i) \frac{\exp(a'_i \theta_j + d_i)}{1 + \exp(a'_i \theta_j + d_i)} \quad (9)$$

เมื่อ $P(x_{ij} = 1/a_i, c_i, d_i, \theta_j)$ คือความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องของผู้สอบคนที่ j สำหรับข้อสอบข้อที่ i , X_{ij} คือผลการตอบข้อสอบของผู้สอบคนที่ j (ตอบถูกได้ 1 คะแนน และตอบผิดได้ 0 คะแนน) คือ เวกเตอร์ค่าพารามิเตอร์อำนาจจำแนกของข้อสอบข้อที่ i , c_i คือค่าพารามิเตอร์การเดาของข้อสอบข้อที่ i , d_i คือค่าพารามิเตอร์ที่สัมพันธ์กับค่าความยากของข้อสอบข้อที่ i และ θ_j คือ เวกเตอร์ความสามารถของผู้สอบคนที่ j โดยผลที่ได้จากการคำนวณ P_{ij} จะนำไปเปรียบเทียบกับกลุ่มแบบเอกรูป (Uniform Random) ซึ่งเรียกว่า

ในการสร้างสายอักขระของผู้สอบแต่ละคนสำหรับข้อสอบที่มีการตรวจให้คะแนน 2 ค่า (0, 1) สามารถคำนวณ P_{ij} ได้จากโมเดลที่การประมาณค่าความสามารถ (θ) ในมิติต่างกันสามารถชดเชยกันได้ (Compensatory Model) ซึ่งเป็นโมเดลโลจิสติกชนิด 3 พารามิเตอร์ (Reckase, 2009) ดังแสดงในสมการ

P^* โดยกำหนดให้

$$x_{ij} = \begin{cases} 0, & P_{ij} < P^* \\ 1, & P_{ij} \geq P^* \end{cases} \quad (10)$$

สำหรับข้อสอบที่มีรูปแบบการตรวจให้คะแนนมากกว่า 2 ค่า (0, 1 และ 2) ใช้โมเดลการตอบสนองข้อสอบแบบพหุมิติแบบ Partial Credit ชนิด 2 พารามิเตอร์ ดังนี้

$$P_{ijk} = P(x_{ij} = k - 1/a_i, d_{\delta_{ki}}, \theta_j) = \frac{\exp(k - 1) a'_i \theta_j - \sum_{t=1}^k d_{\delta_{ti}}}{\sum_{m=1}^{k_j} \exp(m - 1) a'_i \theta_j - \sum_{t=1}^m d_{\delta_{ti}}} \quad (11)$$

เมื่อ $P(x_{ij} = k - 1/a_i, d_{\delta_{ki}}, \theta_j)$ คือความน่าจะเป็นในการตอบข้อสอบของผู้สอบคนที่ j สำหรับข้อสอบข้อที่ i , X_{ij} คือ ผลการตอบข้อสอบของผู้สอบคนที่ j (ตั้งแต่ 0 คะแนนถึง K-1 คะแนน) $d_{\delta_{ki}}$ คือ ค่าพารามิเตอร์ความยากแต่ละระดับขั้นการตอบ โดยที่ $d_{\delta_{1i}} = 0$, K_i คือ จำนวนของระดับขั้นการตอบสำหรับข้อสอบข้อที่ i ซึ่ง $k = 1, 2, \dots, K_i$ โดยผลที่ได้จากการคำนวณจะนำไปเปรียบเทียบกับกลุ่มแบบเอกรูป ซึ่งเรียกว่าโดยกำหนดให้

$$x_{ij} = \begin{cases} 2, & 0 < P^* \leq P_{ij1} \\ 1, & P_{ij1} < P^* \leq P_{ij2} \\ 0, & P_{ij2} < P^* < 1 \end{cases} \quad (12)$$

ในการสร้างลักษณะการแจกแจงความสามารถ (Generate Ability Distribution) ในครั้งนี้จะจำลองข้อมูลให้มีลักษณะการแจกแจงปกติแบบ 2 ตัวแปร (Bivariate Normal Distribution) โดยใช้โปรแกรม MATLAB ในการจำลองข้อมูล ซึ่งการสร้างสเกลการเปรียบเทียบแนวตั้งที่แท้จริงนั้นคุณสมบัติและชุดการกระจายความสามารถของผู้สอบจะถูกสร้างขึ้นให้อยู่ในแกนที่สามารถเปรียบเทียบกันได้ ซึ่งการศึกษาครั้งนี้จะศึกษากับกลุ่มผู้สอบที่มีความสามารถแตกต่างกัน 2 ระดับขั้น คือ ระดับที่ 1 และ 2 ตามลำดับ โดยมีระดับความสัมพันธ์ของมิติความสามารถในแต่ละระดับขั้นเท่ากัน คือ 0.5



2) การประมาณค่าพารามิเตอร์

สำหรับการประมาณค่าพารามิเตอร์ของข้อสอบคณะผู้วิจัยใช้โปรแกรม BMIRTII ซึ่งพัฒนาโดย Yao (2003 อ้างถึงใน Reckase, 2009) ซึ่งออกแบบเพื่อให้สามารถทำการประมาณค่าพารามิเตอร์ของข้อสอบ และค่าพารามิเตอร์ของผู้สอบสำหรับโมเดลการตอบสนองข้อสอบแบบพหุมิติแบบโลจิสติก ชนิด 3 พารามิเตอร์ สำหรับข้อสอบที่มีการตรวจให้คะแนน 2 ค่า และโมเดลการตอบสนองข้อสอบแบบพหุมิติแบบพาเซียลเครดิต ชนิด 2 พารามิเตอร์ สำหรับข้อสอบที่มีการตรวจให้คะแนนมากกว่า 2 ค่า นอกจากนี้ยังสามารถตรวจสอบความเป็นพหุมิติของข้อมูลผลการตอบของผู้สอบโดยใช้ดัชนี Akaike's Information Criterion (AIC) ได้อีกด้วย

3) การเชื่อมโยงคะแนน สำหรับการศึกษาในครั้งนี้คณะผู้วิจัยใช้วิธีการคุณลักษณะแบบทดสอบ (TCF) โดยใช้โปรแกรม LinkMIRT

4. การวิเคราะห์ข้อมูล

การวิเคราะห์ข้อมูลสำหรับการศึกษานี้ครั้งนี้จะทำการวิเคราะห์ดัชนีรากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE) ดังสมการที่ 13 ดัชนีความเอนเอียง (BIAS) ดังสมการที่ 14

$$RMSE = \left(\frac{\sum_{i=1}^I (a_{1i}^* - a_{1i})^2}{I} \right)^{1/2} \quad (13)$$

$$BIAS = \sum_{i=1}^I \frac{(a_{1i}^* - a_{1i})}{I} \quad (14)$$

เมื่อ a_{1i}^* คือ พารามิเตอร์ค่าความอำนาจจำแนกได้แปลงแล้ว

a_{1i} คือ ค่าพารามิเตอร์อำนาจจำแนกในมิติที่ 1 ของข้อสอบข้อที่ i

I คือ จำนวนของข้อสอบรวม

ทั้งนี้ในการคำนวณค่า RMSE และ ค่า BIAS สำหรับค่าพารามิเตอร์อื่นๆ ของข้อสอบสามารถใช้สูตรดังสมการข้างต้นในการคำนวณได้

นอกจากนี้ยังดำเนินการวิเคราะห์ความแปรปรวนแบบวัดซ้ำ (Repeated Measure ANOVA) เพื่อศึกษาปฏิสัมพันธ์ของตัวแปรอิสระต่อคุณภาพของการเชื่อมโยงคะแนน

ผลการวิจัย

จากการวิเคราะห์ข้อมูลโดยใช้ Repeated Measure ANOVA พบว่า โครงสร้างมิติความสามารถและสัดส่วนน้ำหนักคะแนนของข้อสอบรวม ไม่รวมส่งผลกระทบต่อคุณภาพของการเชื่อมโยงคะแนน โดยเฉพาะคุณภาพด้านความถูกต้องของการประมาณค่าพารามิเตอร์ของข้อสอบ แต่อย่างไรก็ตามพบว่า มีเพียงการแปลงคะแนนของค่าพารามิเตอร์ความยากของระดับขั้นการตอบที่ 2 (d_{δ_2}) ในส่วนของข้อสอบที่ตรวจให้คะแนนมากกว่า 2 ค่า ที่ส่งผลกระทบต่อคุณภาพด้านความคงที่ของการแปลงค่าพารามิเตอร์สำหรับการเชื่อมโยงคะแนนอย่างมีนัยสำคัญทางสถิติที่ระดับ .05

เมื่อพิจารณาอิทธิพลหลัก พบว่า ทั้งโครงสร้างมิติความสามารถและสัดส่วนน้ำหนักคะแนนของข้อสอบรวมต่างส่งผลกระทบต่อคุณภาพของการเชื่อมโยงคะแนนทั้งความคงที่และความถูกต้องของการแปลงค่าพารามิเตอร์สำหรับการเชื่อมโยงคะแนนอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 โดยเฉพาะอย่างยิ่งการแปลงค่าพารามิเตอร์ของข้อสอบแบบตรวจให้คะแนนมากกว่า 2 ค่า ดังรายละเอียดในตารางที่ 1



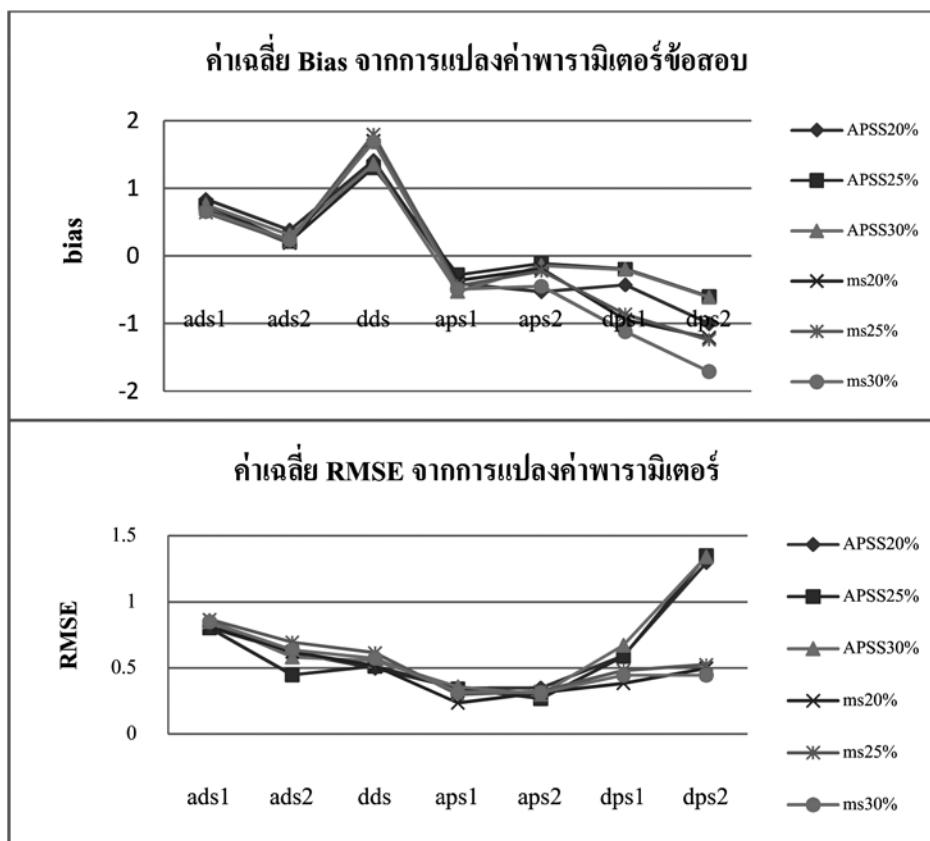
ตารางที่ 1 สถิติทดสอบ (F) องศาความเป็นอิสระ (df) และขนาดอิทธิพล (η^2) ของ Bias, Ln RMSE

แหล่ง	Ln RMSE							BIAS						
	ข้อสอบปรนัย			ข้อสอบอัตนัย				ข้อสอบปรนัย			ข้อสอบอัตนัย			
	a_1	a_2	d_i	a_1	a_2	d_{δ_2}	d_{δ_3}	a_1	a_2	d_i	a_1	a_2	d_{δ_2}	d_{δ_3}
ระหว่างกลุ่ม														
โครงสร้าง														
F	10.744 [*]	2.150	4.469 [*]	0.128	1.908	49.853 [*]	587.114 [*]	2.647	3.430	523.59 [*]	4.827 [*]	13.726 [*]	665.179 [*]	1655.408 [*]
df	1	1	1	1	1	1	1	1	1	1	1	1	1	1
η^2	0.22	0.05	0.10	0.00	0.05	0.57	0.94	0.07	0.08	0.93	0.11	0.27	0.95	0.98
ภายในกลุ่ม														
สัดส่วนข้อสอบรวม														
F	0.132	1.905	0.075	3.274 [*]	1.068	3.805 [*]	1.158	0.597	1.121	10.944	4.689 [*]	20.365 [*]	23.352 [*]	119.711 [*]
df	1.548	1.719	2	2	1.712	1.656	2	2	2	2	2	2	2	2
η^2	0.00	0.04	0.00	0.08	0.03	0.09	0.03	0.02	0.03	0.02	0.11	0.35	0.38	0.76
สัดส่วนข้อสอบรวม X โครงสร้าง														
F	0.083	2.619	0.747	1.00	2.535	4.328 [*]	0.151	0.481	0.439	1.158	1.531	3.619	0.002	0.966
df	1.548	1.719	2	2	1.712	1.656	2	2	2	2	2	2	2	2
η^2	0.00	0.06	0.02	0.03	0.06	0.10	0.00	0.01	0.01	0.03	0.04	0.09	0.00	0.03

*P<.05

สำหรับคุณภาพของการเชื่อมโยงคะแนนเมื่อพิจารณาตามอิทธิพลหลัก ในส่วนของโครงสร้างมีความสามารถพบว่า โครงสร้างแบบไม่ซับซ้อน (APSS) จะส่งผลให้คุณภาพของการเชื่อมโยงคะแนนดีกว่าโครงสร้างแบบซับซ้อน (MS) โดยเฉพาะด้านความถูกต้องของการแปลงค่าพารามิเตอร์ของข้อสอบ

สำหรับอิทธิพลหลักในส่วนของสัดส่วนน้ำหนักคะแนนของข้อสอบรวม พบว่า สัดส่วนน้ำหนักคะแนนข้อสอบรวม 25% จะส่งผลให้คุณภาพของการเชื่อมโยงคะแนนดีกว่าสัดส่วนอื่นๆ โดยเฉพาะอย่างยิ่งด้านความถูกต้องของการแปลงค่าพารามิเตอร์ของข้อสอบ ดังรายละเอียดในภาพที่ 1



ภาพที่ 1 ค่า Bias เฉลี่ย และค่า RMSE เฉลี่ยแต่ละพารามิเตอร์สำหรับข้อสอบร่วม

เมื่อ ads1 ads2 แทน อำนาจจำแนกในมิติที่ 1, 2 ของข้อสอบที่ตรวจให้คะแนน 2 ค่า

dds แทน ความยากของข้อสอบที่ตรวจให้คะแนน 2 ค่า

aps1 aps2 แทน อำนาจจำแนกในมิติที่ 1, 2 ของข้อสอบที่ตรวจให้คะแนนมากกว่า 2 ค่า

dps1 dps2 แทน ความยากแต่ละระดับของการตอบของข้อสอบที่ตรวจให้คะแนนมากกว่า 2 ค่า

อภิปรายผลการวิจัย

จากผลการวิจัยสรุปได้ว่า โครงสร้างมิติความสามารถและสัดส่วนน้ำหนักคะแนนของข้อสอบร่วมไม่รวมส่งผลต่อคุณภาพของการเชื่อมโยงคะแนน แต่โครงสร้างมิติความสามารถและสัดส่วนน้ำหนักคะแนนของข้อสอบร่วมต่างส่งผลต่อคุณภาพของการเชื่อมโยงคะแนน โดยโครงสร้างมิติความสามารถแบบไม่ซับซ้อน

จะให้คุณภาพการเชื่อมโยงคะแนนที่ดีกว่าโครงสร้างมิติความสามารถแบบซับซ้อน และสัดส่วนน้ำหนักคะแนนของข้อสอบร่วม 25% จะให้คุณภาพการเชื่อมโยงคะแนนดีกว่าสัดส่วนอื่น ๆ ซึ่งขัดแย้งกับ Kolen and Brennan (2004) ที่กล่าวว่าสัดส่วนข้อสอบร่วมที่ใช้สำหรับการเชื่อมโยงคะแนนตามทฤษฎีการตอบสนองข้อสอบ (IRT) ที่เหมาะสมคือ 20% แต่สอดคล้องกับ Reckase (2009) ที่กล่าวว่าสัดส่วนข้อสอบร่วมที่เหมาะสมสำหรับการเชื่อมโยงคะแนนตามโมเดลการตอบสนองข้อสอบแบบพหุมิติคือ 20%-25% ทั้งนี้อาจเนื่องมาจากมิติความสามารถที่จะทำการวัดในครั้งนี้ใช้เพียง 2 มิติเท่านั้น จึงทำให้ข้อสอบที่มีโครงสร้างมิติความสามารถแบบไม่ซับซ้อนซึ่งแบ่งมิติความสามารถออกเป็น 2 มิติอย่างชัดเจน ให้คุณภาพของการเชื่อมโยงคะแนนที่ดีกว่า นอกจากนี้สัดส่วนน้ำหนักคะแนนข้อสอบร่วมมีอิทธิพลทำให้การเชื่อมโยงคะแนนมีคุณภาพมากขึ้น (Cao, 2008)

หากข้อสอบรวมประกอบด้วยข้อสอบที่ตรวจให้คะแนนแบบ 2 ค่า และมากกว่า 2 ค่า (Kim & Lee, 2006; Kirkpatrick, 2005; Sykes et al., 2002; Tate, 2000) และหากเพิ่มจำนวนข้อสอบที่มีการตรวจให้คะแนนแบบ 2 ค่ามากขึ้นให้อยู่ในระดับที่เหมาะสม (Yao & Boughton, 2009) ซึ่งงานวิจัยในครั้งนี้ทำให้เกิดองค์ความรู้ใหม่เกี่ยวกับสัดส่วนน้ำหนักคะแนนของข้อสอบรวมต่อข้อสอบทั้งฉบับสำหรับใช้ในการเชื่อมโยงคะแนนตามทฤษฎีการตอบสนองข้อสอบแบบพหุมิติสำหรับแบบทดสอบรูปแบบผสม

ข้อเสนอแนะ

แต่อย่างไรก็ตามเนื่องจากการศึกษาในครั้งนี้ใช้วิธีการเชื่อมโยงคะแนนโดยวิธี TCF ซึ่งเป็นวิธีการที่ยังไม่มีความชัดเจนเกี่ยวกับการแปลงคะแนนค่าพารามิเตอร์ที่สัมพันธ์กับค่าความยาก ดังนั้นในการวิจัยครั้งต่อไปควรได้มีการศึกษาเปรียบเทียบกับวิธีการเชื่อมโยงคะแนนวิธีอื่นๆ เช่น วิธีหมุนแกนแบบไม่ตั้งฉาก และควรได้มีการศึกษาเปรียบเทียบกับโมเดลการตอบสนองข้อสอบแบบพหุมิติที่มีรูปแบบการตรวจให้คะแนนมากกว่า 2 ค่า โมเดลอื่นๆ เช่น Grade Response Model, Generalize Partial Credit Model อีกด้วย

เอกสารอ้างอิง

- Ackerman, Terry A. (1994). Using multidimensional item response theory to understand what items and test Are measuring. **Apply Measurement in Education**, 7(4), 255-278.
- Cao, Y. (2008). **Mixed-format test equating: effects of test dimensionality and common item sets**. Ph. D. Dissertation, Department of measurement, statistics and evaluation, University of Maryland, USA.
- Kim, S. , & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. **Applied Measurement in Education**, 19, 357-381.
- Kim, S. & Lee, W. C. (2006). An extension of four IRT linking methods for mixed format tests. **Journal of Educational Measurement**, 43(1), 53-76.
- Kirkpatrick, R. K. (2005). **The effects of item format in common item equating**. Ph. D. Dissertation, Measurement and Evaluation, University of Iowa.
- Kolen, M. J., & Brennan, R. L. (2004). **Test equating: methods and practices**. 2nd ed. New York, NY: Springer-Verlag.
- Min, K. S. (2007). Evaluation of linking methods for multidimensional IRT calibrations. **Asia Pacific Education Review**, 8(1), 41-55.
- Oshima, T. C., Davy, T. C. & Lee, K. (2000). Multidimensional linking: Four practical approaches. **Journal of Educational Measurement**, 37(4), 357-373.
- Patz, R. J., & Yao, L. (2007). Methods and models for vertical scaling. In N. J. Dorans, M. Pommerich & P. W. Holland (Eds.), **Linking and aligning scores and scales**. (pp. 253-272). New York: Springer.
- Reckase, M. D., & Martineau, J. A. (2004). **The vertical scaling of science achievement tests**. Paper commissioned by The Committee on test design for K-12 science achievement center for education national research council. [n. p.].
- Reckase, Mark D. (2009). **Multidimensional item response theory**. New York: Springer.



- Sykes, R. C., Hou, L.-L., Hanson, B., & Wang, Z. (2002). **Multidimensionality and the equating of a mixed-format math examination.** Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.
- Tate, R. L. (2000). Performance of a proposed method for the linking of mixed format tests with construct response and multiple choice items. **Journal of Educational Measurement**, 37, 329-346.
- Yao, L. & Schwarz, R. (2005). A multidimensional partial credit model with associated item and test statistics. **Paper presented at the Annual Meeting at the American Educational Research Association, Montreal, Quebec, Canada.**
- Yao, L. & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. **Applied Psychological Measurement**, 31(2), 83-105.
- Yao, L., & Boughton, K. A. (2005). **Multidimensional parameter recovery: Markov chain Monte Carlo versus NOHARM.** Unpublished manuscript.
- Yao, L. & Boughton, K. A. (2009). Multidimensional linking for tests with mixed format item types. **Journal of Educational Measurement**, 46(2), 177-197.
- Yon, H. (2006). **Multidimensional item response theory (MIRT) approaches to vertical scaling.** Ph. D. Dissertation, Measurement and quantitative method, Michigan State University.