การตรวจสอบความแม่นยำของวิธีการประมาณค่าพารามิเตอร์พร้อมกันสำหรับการประเมินพัฒนาการความสามารถของผู้สอบในแบบทดสอบผสมที่มีความยาว ความยากและการให้คะแนน แตกต่างกัน

สุนีย์ เงินยวง[1]
ดร.องอาจ นัยพัฒน์[2]
ดร.ผจงจิต อินทสุวรรณ์[3]
ดร.สิริรัตน์ วิภาสศิลป์[4]

บทคัดย่อ

การศึกษาครั้งนี้มีจุดมุ่งหมายเพื่อตรวจสอบความแม่นยำของวิธีการประมาณค่าพารามิเตอร์พร้อมกันสำหรับการประเมินพัฒนาการความสามารถของผู้สอบ ($\Theta_2 - \Theta_1$) ด้วยแบบทดสอบผสมที่ประกอบด้วยข้อสอบแบบเลือกตอบที่ตรวจให้คะแนนสองค่าและข้อสอบแบบเขียนตอบที่ตรวจให้คะแนนหลายค่า

การศึกษาครั้งนี้ใช้โมเดลโลจิสติกแบบสามพารามิเตอร์ (3PL) จำลองผลการตอบข้อสอบแบบเลือกตอบ และใช้โมเดลเจเนอรอลไรส์พาร์เชียลเครดิต (GPCM) จำลองผลการตอบข้อสอบแบบเขียนตอบของผู้สอบ 1,000 คนที่ทดสอบซ้ำสองครั้ง ตามตัวแปรอิสระที่ศึกษา 4 ตัว ได้แก่ ระดับพัฒนาการความสามารถ 9 ระดับ ความยาวของแบบทดสอบผสม 3 ขนาด ซึ่งกำหนดจากจำนวนข้อสอบแบบเลือกตอบต่อจำนวนข้อสอบแบบเขียนตอบ (30 : 10, 24 : 8 และ 15 : 5) ความยากของแบบทดสอบผสมที่ใช้ในการทดสอบครั้งที่หนึ่ง 3 ระดับและการให้คะแนนของข้อสอบแบบเขียนตอบ 3 รูปแบบ รวมทั้งสิ้น 243 เงื่อนไข (9x3x3x3) การประเมินความแม่นยำของวิธีการประมาณค่าพารามิเตอร์พร้อมกันพิจารณาจากค่าความลำเอียง (BIAS) และ ค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (RMSE) ของค่าประมาณของพัฒนาการความสามารถ ($\Theta_2 - \Theta_1$)

ผลการศึกษา พบว่า

1. ค่าประมาณพารามิเตอร์ความสามารถที่ได้จากวิธีการประมาณค่าพารามิเตอร์พร้อมกันกับความสามารถจริงมีความสัมพันธ์ทางลบในระดับสูงอย่างมีนัยสำคัญทางสถิติที่ระดับ .01

2. ในทุกเงื่อนไขที่ศึกษา เมื่อส่วนเบี่ยงเบนมาตรฐานของระดับพัฒนาการความสามารถของผู้สอบเป็น 0.80 และ 1.00 ความแม่นยำของวิธีการประมาณค่าพารามิเตอร์พร้อมกัน มีค่าสูงกว่า เมื่อส่วนเบี่ยงเบนมาตรฐานเป็น 1.2 อย่างไม่มีนัยสำคัญที่ระดับ .05 พบว่า

---

[1] ดุษฎีบัณฑิต สาขาวิชาการทดสอบและวัดผลการศึกษา มหาวิทยาลัยศรีนครินทรวิโรฒ

[2] รองศาสตราจารย์ คณะศึกษาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ

[3] รองศาสตราจารย์ สถาบันวิจัยพฤติกรรมศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ

[4] รองศาสตราจารย์ ศูนย์วิชาการประเมินผล มหาวิทยาลัยสุโขทัยธรรมาธิราช

2.1 เมื่อความยากและการให้คะแนนคงที่ แบบทดสอบผสมขนาด 24:8 มีความแม่นยำสูงกว่าแบบทดสอบผสมขนาด 15 : 5 และขนาด 30 : 10 อย่างมีนัยสำคัญที่ระดับ .05

2.2 เมื่อความยาวและการให้คะแนนคงที่ แบบทดสอบผสมที่ใช้ในการสอบครั้งที่หนึ่งที่มีระดับความยากเท่ากับระดับความสามารถของผู้สอบ (0.00) มีความแม่นยำต่ำกว่า แบบทดสอบผสมที่ใช้ในการสอบครั้งที่หนึ่งที่มีระดับความยากไม่เท่ากับระดับความสามารถของผู้สอบ (-0.50 และ 0.50) อย่างมีนัยสำคัญที่ระดับ .05

2.3 เมื่อความยาวและความยากคงที่ แบบทดสอบผสมที่ให้คะแนนข้อสอบแบบเขียนตอบ 3 ระดับ (0/1/2) มีความแม่นยำสูงกว่าแบบทดสอบผสมที่ให้คะแนนข้อสอบแบบเขียนตอบ 4 ระดับ (0/1/2/3) และ 5 ระดับ (0/1/2/3/4) อย่างมีนัยสำคัญที่ระดับ .05

คำสำคัญ : พัฒนาการ แบบทดสอบผสม การประมาณค่าพารามิเตอร์พร้อมกัน

An Accuracy Investigation of Concurrent Calibration for the Assessment of Examinee's Growth Ability in Mixed-Format Test that has Different Test Lengths, Item Difficulties, and Scoring Categories.

Abstract

The main purpose of this study was to investigate the accuracy of concurrent calibration for the assessment of examinee's growth ability (Θ_2- Θ_1) in mixed–format test consisting in terms of multiple choice (MC) items and constructed–response (CR) items where the MC was dichotomous response model and the CR was polytomous response model.

In order to fulfill this purpose, the 3PL/GPCM model combination was then used to simulate the item responses data of 1,000 examinees, in which four factors–growth ability, test lengths (the number of MC:CR, i.e. 30 : 10, 24 : 8, and 15 : 5), item difficulties and scoring categories of mixed–format test – were manipulated. In total, there were 243 conditions (9x3x3x3) with respect to the four variables. The accuracy of concurrent calibration was determined from the degrees of bias (BIAS) and the root mean square errors (RMSE) of the estimated growth ability (Θ_2- Θ_1).

The results of the research were as follows :

1. Pearson's correlation coefficients between the estimated ability and the true ability were negatively high and statistically significant at the .01 level.

2. For all conditions, the accuracy of concurrent calibration when the standard deviations of growth ability were 0.80 and 1.00 was statistically insignificantly higher than when the standard deviation of growth ability was 1.2

2.1 When the item difficulties and scoring categories were fixed, the accuracy of concurrent calibration of the 24 : 8 mixed–format test was statistically significantly higher, at the .05 level, than those of the 15 : 5 and the 30 : 10 mixed–format test.

2.2 When the test lengths and scoring categories were fixed, the accuracy of concurrent calibration of the 1st mixed–format test with the 0.00 difficulty was statistically significantly higher, at the .05 level, than those of the 1st test mixed–format test with the -0.50 difficulty and the 0.50 difficulty.

2.3. When the test lengths and item difficulties were fixed, the accuracy of concurrent calibration of the mixed–format test with three–categories CR items (0/1/2) was statistically significantly higher, at the .05 level, than those of the mixed–format test with four–categories CR items (0/1/2/3) and with five–categories CR items (0/1/2/3/4).

Keywords : Growth, Mixed-Format Test, Concurrent Calibration

ภูมิหลัง

ปัจจุบันแนวโน้มการทดสอบในชั้นเรียนและการทดสอบระดับชาติมีการใช้แบบทดสอบที่ผสมข้อสอบแบบเลือกตอบ แบบเติมคำ แบบถูก-ผิด หรือแบบแสดงวิธีทำ ผสมกันอย่างน้อยสองลักษณะในฉบับเดียวมากขึ้น เนื่องจากแบบทดสอบผสมสามารถวัดความรู้ความสามารถได้ครอบคลุมกว่าการใช้แบบทดสอบที่มีข้อสอบลักษณะเดียว (Kim and Lee 2006: 53) ซึ่งการผสมข้อสอบในแบบทดสอบฉบับเดียวกันเป็นการเพิ่มจุดแข็งและชดเชยข้อด้อยของข้อสอบแต่ละประเภทได้ การใช้แบบทดสอบผสมสำหรับวัดพัฒนาการความสามารถของผู้เรียนจึงมีความเป็นไปได้และเป็นประโยชน์สำหรับผู้เกี่ยวข้องที่จะทำให้ได้สารสนเทศของผู้สอบได้ครอบคลุมมากยิ่งขึ้น แต่การใช้คะแนนดิบที่ได้จากผลการทดสอบผสมแต่ละครั้งมาคำนวณพัฒนาการความสามารถย่อมมีความซับซ้อนมากกว่าการคำนวณพัฒนาการความสามารถที่ได้จากการทดสอบด้วยแบบทดสอบที่มีข้อสอบประเภทเดียวกัน

เมื่อพิจารณาวิธีการคำนวณพัฒนาการโดยใช้ผลต่างของคะแนนดิบที่เกิดจากผลการทดสอบด้วยแบบทดสอบที่มีข้อสอบประเภทเดียวกันซ้ำสองครั้งด้วยแบบทดสอบฉบับเดิมหรือแบบทดสอบคู่ขนาน พบว่า การคำนวณผลต่างด้วยคะแนนดิบมีความเชื่อมั่นต่ำเนื่องจากผู้สอบที่มีความสามารถอยู่ในระดับต่ำจะมีโอกาสได้คะแนนพัฒนาการสูงกว่าผู้สอบที่มีความสามารถสูง นอกจากนี้คะแนนพัฒนาการยังขึ้นอยู่กับความยากง่ายของแบบทดสอบ จากข้อจำกัดของวิธีคำนวณการเปลี่ยนแปลงอย่างง่ายซึ่งเป็นการคำนวณตามทฤษฎีการทดสอบแบบมาตรฐานเดิมในเวลาต่อจึงมาได้มีการพัฒนาสูตรการคำนวณพัฒนาการความสามารถขึ้นหลายวิธี รวมถึงการพัฒนาวิธีการคำนวณการเปลี่ยนแปลงตามทฤษฎีการตอบสนอง (Item Response Theory : IRT) โดยเชื่อว่าโมเดล การตอบข้อสอบสามารถชดเชยข้อจำกัดของทฤษฎีการทดสอบแบบมาตรฐานเดิมได้หลายประการ (Fischer. 2003) ได้แก่ ความอิสระในการเลือกชุดของข้อสอบก่อนและหลังเรียน ค่าพารามิเตอร์ข้อสอบไม่มีการแปรเปลี่ยน (Invariance) เมื่อกลุ่มตัวอย่างเปลี่ยนไป ความแม่นยำของคะแนนการเปลี่ยนแปลงขึ้นอยู่กับระดับความยากของข้อสอบ ตลอดจนคะแนนที่คาดหวัง ความแปรปรวนของคะแนน ความสัมพันธ์ของคะแนน และความเชื่อมั่นของคะแนนซึ่งเป็นค่าพารามิเตอร์ของประชากรไม่ขึ้นอยู่กับการแจกแจงความสามารถของกลุ่มประชากรอ้างอิง ซึ่งการคำนวณคะแนนการเปลี่ยนแปลงความสามารถ (Θ_2- Θ_1) (May and Nicewander. 1998) เป็นวิธีหนึ่งที่ตั้งอยู่บนฐานทฤษฎีการตอบข้อสอบจึงลดปัญหาความคลาดเคลื่อนที่มักเกิดขึ้นจากคำนวณคะแนนผลต่างอย่างง่ายของคะแนนดิบและเป็นการคำนวณที่สามารถคำนวณได้ง่ายกว่าเมื่อเทียบกับโมเดลการวัดการเปลี่ยนแปลงตามทฤษฎีการตอบข้อสอบอื่น

เมื่อพิจารณาความเป็นไปได้ของการวัดพัฒนาการความสามารถโดยใช้แบบทดสอบผสมที่มีโครงสร้างการวัดเนื้อหาความสามารถเดียวกันแต่มีระดับความยากแตกต่างกัน พบว่า การศึกษาพัฒนาการความสามารถของกลุ่มประชากรหนึ่งกลุ่มโดยใช้แบบทดสอบผสมที่มีความยากต่างกันมีลักษณะสอดคล้องกับรูปแบบการเก็บรวบรวมข้อมูลสำหรับการปรับเทียบคะแนนสอบที่ได้จากการทดสอบผู้สอบสองกลุ่มประชากรที่มีระดับความสามารถไม่เท่าเทียมกันด้วยแบบทดสอบสองชุดที่มีความยากต่างกันแต่มีข้อสอบร่วมกัน (Common Item Non–Equivalent Group : CINEG) ดังนั้นผู้วิจัยจึงสนใจศึกษาพัฒนาการความสามารถโดยใช้แบบทดสอบผสมโดยนำวิธีการปรับเทียบมาใช้ร่วมในการประมาณค่าพารามิเตอร์ความสามารถ

การปรับเทียบคะแนนสอบตามทฤษฎีการตอบข้อสอบมี 3 ขั้นตอนหลัก (Kim. 2007 : 24) ได้แก่ ขั้นที่ 1 การเลือกและใช้โมเดลสำหรับประมาณค่าพารามิเตอร์ ขั้นที่ 2 การสร้างมาตรวัดร่วมและขั้นที่ 3 การแปลงคะแนนดิบให้อยู่บนมาตรวัดร่วมสำหรับการแปลความหมายและรายงานผลของคะแนน วิธีการสร้างมาตรวัดร่วมที่นิยมใช้มี 3 วิธี (Kolen and Brennan. 1995) ได้แก่ วิธีการประมาณค่าพารามิเตอร์แยกกัน (Separate calibration) วิธีการประมาณค่าพารามิเตอร์ที่กำหนดพารามิเตอร์ของข้อสอบร่วม (Fixed-item parameter calibration) และวิธีการประมาณค่าพารามิเตอร์พร้อมกัน (Concurrent calibration) ซึ่งวิธีการประมาณค่าพารามิเตอร์พร้อมกันเป็นวิธีที่นำข้อมูลผลการตอบของทุกกลุ่มผู้สอบที่ทดสอบด้วยแบบทดสอบต่างฉบับหรือต่างครั้งมารวมเป็นข้อมูลการตอบฉบับเดียวกันหรือครั้งเดียวกันแล้วจึงประมาณค่าความสามารถพร้อมกัน โดยโปรแกรมคอมพิวเตอร์ที่ยอมให้มีการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบหลายกลุ่มได้ เช่น โปรแกรม PARSCALE (Muraki and Bock. 1999) และโปรแกรม MULTILOG (Thissen. 1991) เป็นต้น

แม้ว่าวิธีการประมาณค่าพารามิเตอร์พร้อมกันเป็นวิธีการสร้างมาตรวัดร่วมที่สามารถขจัดปัญหาความไม่เท่าเทียมกัน (Paek and Young. 2005) และให้ค่าประมาณพารามิเตอร์ที่มีความคลาดเคลื่อนต่ำกว่าวิธีการประมาณค่าพารามิเตอร์แบบอื่นและสามารถนำไปใช้กับข้อมูลแบบทดสอบผสมเมื่อผู้สอบมีความสามารถไม่เท่าเทียมกันได้ แต่จากการศึกษางานวิจัยหลายงานวิจัย (Kim and Cohen. (1998) ; Li, Lissitz and Yang.(1999) ; Park. (2000) ; Hanson and Béguin. (2002) ; Kim. (2004); DeMars. (2004) ; Yao and Mao. (2004) ; Paek and Young. (2005) Kamata and Tate. (2005); Kim and Kolen. (2006) ; Yang. (2007) ; Hu, Rogers and Vukmirovic. (2008) ; Cao. (2008)) **พบว่า โมเดลการตอบข้อสอบสำหรับประมาณ**ค่าพารามิเตอร์ วิธีการสร้างมาตรวัดร่วม ขนาดความยาวของแบบทดสอบผสม ระดับคะแนนข้อสอบแบบให้คะแนนหลายค่าแต่ละข้อ ระดับความเท่าเทียมกันของความสามารถผู้สอบในการสอบแต่ละครั้งและจำนวนหน่วยตัวอย่าง เป็นปัจจัยสำคัญที่ส่งผลต่อความแม่นยำของวิธีการประมาณค่าพารามิเตอร์พร้อมกันที่อาจทำให้ผลต่างของค่าประมาณความสามารถนั้นมีค่าใกล้เคียงหรือเท่ากับผลต่างความสามารถที่แท้จริงได้ ดังนั้นเพื่อให้การศึกษาสามารถควบคุมปัจจัยที่เกี่ยวข้องให้ได้มากที่สุดผู้วิจัยจึงจำลองข้อมูลเพื่อตรวจสอบความแม่นยำของผลต่างค่าประมาณความสามารถที่ได้จากวิธีการสร้างมาตรวัดร่วมแบบประมาณค่าพารามิเตอร์พร้อมกัน ภายใต้เงื่อนไขพัฒนาการความสามารถของผู้สอบแตกต่างกัน และทดสอบซ้ำสองครั้งด้วยแบบทดสอบผสมมีความยาว ระดับความยากและการให้คะแนนแตกต่างกัน

วัตถุประสงค์ของการวิจัย

การศึกษาวิจัยครั้งนี้มีวัตถุประสงค์หลักเพื่อตรวจสอบความแม่นยำของวิธีการประมาณค่าพารามิเตอร์พร้อมกันเมื่อผู้สอบมีพัฒนาการความสามารถเพิ่มขึ้นแตกต่างกัน ด้วยแบบทดสอบผสมที่มีความยาว ความยาก และการให้คะแนนแตกต่างกัน โดยศึกษาภายใต้สถานการณ์จำลอง มีวัตถุประสงค์เฉพาะดังนี้

1. เพื่อศึกษาและเปรียบเทียบความสัมพันธ์ระหว่างค่าประมาณพารามิเตอร์ความสามารถที่ได้จากการใช้วิธีการประมาณค่าพารามิเตอร์พร้อมกัน กับความสามารถจริง เมื่อผู้สอบมีพัฒนาการความสามารถเพิ่มขึ้นแตกต่างกัน ภายใต้เงื่อนไข ความยาว ความยาก และการให้คะแนนของแบบทดสอบผสมแตกต่างกันเพื่อศึกษาและเปรียบเทียบความลำเอียงและค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนของค่าประมาณความสามารถ

กำลังสองได้จากวิธีการประมาณค่าพารามิเตอร์พร้อมกันเมื่อผู้สอบมีระดับพัฒนาการความสามารถเพิ่มขึ้นแตกต่างกัน ภายใต้เงื่อนไข ความยาว ความยากและการให้คะแนน ของแบบทดสอบผสมแตกต่างกัน เพื่อศึกษาและเปรียบเทียบความลำเอียงและค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนของพัฒนาการความสามารถกำลังสองที่ได้จากวิธีการประมาณค่าพารามิเตอร์พร้อมกัน เมื่อผู้สอบมีระดับพัฒนาการความสามารถเพิ่มขึ้นแตกต่างกัน ภายใต้เงื่อนไข ความยาว ความยาก และ การให้คะแนนของแบบทดสอบผสมแตกต่างกัน

สมมติฐานของการวิจัย

ผู้วิจัยได้ตั้งสมมติฐานการวิจัยไว้ดังนี้

1. ค่าประมาณความสามารถที่ได้จากวิธีการประมาณค่าพารามิเตอร์พร้อมกันมีความสัมพันธ์ระดับสูงกับค่าความสามารถจริงที่กำหนดในทุกระดับพัฒนาการความสามารถ

2. ค่าประมาณพารามิเตอร์ความสามารถของผู้สอบที่ได้จากวิธีการประมาณค่าพารามิเตอร์พร้อมกันมีค่าความลำเอียง (BIAS) และค่ารากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (RMSE) เข้าใกล้ศูนย์ทุกระดับพัฒนาการความสามารถของผู้สอบ

3. เมื่อระดับพัฒนาการความสามารถของผู้สอบมีส่วนเบี่ยงเบนมาตรฐานเท่ากับหรือต่ำกว่าส่วนเบี่ยงเบนมาตรฐานของความสามารถครั้งแรก (1 หรือ 0.8) ค่าประมาณความสามารถจากการใช้วิธีการประมาณค่าพารามิเตอร์พร้อมกันมีความแม่นยำมากกว่าเมื่อส่วนเบี่ยงเบนมาตรฐานของระดับพัฒนาการความสามารถของผู้สอบสูงกว่าส่วนเบี่ยงเบนมาตรฐานของความสามารถครั้งแรก (1.2)

3.1 แบบทดสอบผสมที่มีขนาดยาวกว่ามีความแม่นยำมากกว่าแบบทดสอบผสมที่มีขนาดสั้นกว่าในทุกระดับความยากและระดับการให้คะแนนของแบบทดสอบผสม

3.2 แบบทดสอบผสมที่มีระดับความยากสอดคล้องกับระดับความสามารถเริ่มต้นของผู้สอบ จะมีความแม่นยำมากกว่าแบบทดสอบผสมที่มีระดับความยากไม่สอดคล้องกับระดับความสามารถเริ่มต้นของผู้สอบในทุกขนาดความยาวและระดับการให้คะแนนของแบบทดสอบผสม

3.3 แบบทดสอบผสมที่มีระดับการให้คะแนนข้อสอบเขียนตอบสูงกว่า มีความแม่นยำมากกว่าแบบทดสอบผสมที่มีระดับการให้คะแนนข้อสอบเขียนตอบต่ำกว่า ในทุกขนาดความยาวและความยากของแบบทดสอบผสม

วิธีดำเนินการวิจัย

ขอบเขตของการวิจัย

การศึกษาครั้งนี้เป็นการศึกษาจากข้อมูลในสถานการณ์จำลองโดยมีขอบเขตการศึกษาดังนี้

ข้อมูลจำลองที่ใช้ศึกษา

ข้อมูลที่ศึกษาเป็นแบบแผนการตอบแบบทดสอบผสมสองครั้งของผู้สอบจำนวน 1,000 คน ที่ได้จากโปรแกรม WINGEN2 (Han. 2007) โดยใช้โมเดลโลจิสติกแบบสามพารามิเตอร์ในการจำลองผลการตอบข้อสอบ

แบบเลือกตอบ และใช้โมเดลเจเนอรอลไรส์พาร์เชียลเครดิตจำลองผลการตอบข้อสอบแบบเขียนตอบที่ให้คะแนนหลายค่า การกำหนดค่าพารามิเตอร์สำหรับจำลองแบบแผนการตอบมีดังนี้

1. การกำหนดพารามิเตอร์ของข้อสอบในแบบทดสอบผสมสำหรับการทดสอบสองครั้ง

การทดสอบครั้งที่ 1 แบบทดสอบผสมมีความยาว 3 ขนาด (จำนวนข้อสอบแบบเลือกต่อจำนวนข้อสอบแบบเขียนตอบ) ได้แก่ 30 : 10, 24 : 8 และ 15 : 5 แต่ละขนาดมีอัตราส่วนจำนวนข้อสอบเลือกตอบต่อจำนวนข้อสอบเขียนตอบคงที่เป็น 3 : 1 ความยากเฉลี่ยของข้อสอบในแบบทดสอบทั้งฉบับมี 3 ระดับ ได้แก่ ระดับความยากเฉลี่ย -0.5, 0.0, และ 0.5 และมีการให้คะแนนข้อสอบเขียนตอบ 3 รูปแบบ ได้แก่ 3 ระดับ (0/1/2 คะแนน) 4 ระดับ (0/1/2/3 คะแนน) และ 5 ระดับ (0/1/2/3/4 คะแนน) รวมแบบทดสอบผสมสำหรับใช้ในการทดสอบครั้งที่ 1 ทั้งสิ้น 27 ฉบับ (3x3x3)

การทดสอบครั้งที่ 2 แบบทดสอบผสมมีความยาว 3 ขนาดและมีระดับการให้คะแนนข้อสอบ รูปแบบเดียวกับการทดสอบครั้งที่ 1 แต่ข้อสอบเฉพาะในการทดสอบครั้งที่สองมีความยากเพิ่มขึ้นจากข้อสอบเฉพาะของการทดสอบครั้งที่ 1 ข้อละ 0.5 รวมแบบทดสอบผสมสำหรับใช้ในการทดสอบครั้งที่ 2 ทั้งสิ้น 27 ฉบับ (3x3x3)

2. การกำหนดค่าพารามิเตอร์ความสามารถของผู้สอบ

การทดสอบครั้งที่ 1 กำหนดค่าพารามิเตอร์ความสามารถผู้สอบมีการแจกแจงความสามารถปกติมีค่าเฉลี่ย 0.00 และส่วนเบี่ยงเบนมาตรฐานเป็น 1 และในการทดสอบครั้งที่ 2 กำหนดให้ผู้สอบมีความสามารถเพิ่มขึ้นโดยคำนวณจากสมการการแปลงเชิงเส้น $\grave{e}_2=(SD)\grave{e}_1+Mean$ โดยกำหนดให้ค่าเฉลี่ยพารามิเตอร์ความสามารถเพิ่มขึ้นจากการทดสอบครั้งที่หนึ่งเป็น 0.1, 0.5, และ 1 หน่วย และมีส่วนเบี่ยงเบนมาตรฐานเป็น 0.8, 1 และ 1.2 ตามลำดับรวมลักษณะความสามารถผู้สอบทั้งสิ้น 10 กลุ่ม [1+ (3x3) =10]

3. การจำลองผลแบบแผนการตอบของผู้สอบ 1,000 คน เป็นจำลองผลการทดสอบครั้งที่ 1และครั้งที่ 2 แยกกัน โดยกำหนดค่าพารามิเตอร์ข้อสอบและผู้สอบตามเงื่อนไขที่ศึกษา รวมแบบแผนการตอบข้อสอบครั้งที่หนึ่งและครั้งที่สองที่จำลองด้วยโปรแกรม WINGEN2 (Han. 2007) ทั้งสิ้น 243 ชุด (9x3x3x3) ทำซ้ำชุดละ 50 ครั้ง

4. แบบแผนการตอบแบบทดสอบผสมที่ได้จากการทดสอบครั้งที่ 1 และการทดสอบครั้งที่ 2 จะถูกนำรวมเข้าเป็นผลการตอบข้อสอบฉบับใหม่สำหรับใช้ประมาณค่าพารามิเตอร์ความสามารถผู้สอบโดยใช้วิธีการประมาณค่าพารามิเตอร์พร้อมกัน (Concurrent Calibration) โดยกำหนดให้ผลการตอบข้อสอบเฉพาะครั้งที่ 1 หรือครั้งที่ 2 ที่ผู้สอบไม่ได้ทำได้คะแนนเป็น 0 รวมข้อมูลที่ใช้วิธีการประมาณค่าพารามิเตอร์แบบพร้อมทั้งหมด 243 คู่

5. การตรวจสอบความแม่นยำของค่าประมาณความสามารถ พิจารณาจากค่าความลำเอียงและค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (RMSE) ของค่าประมาณความสามารถของผู้สอบแต่ละคนเมื่อเทียบกับค่าพารามิเตอร์ความสามารถที่กำหนดขึ้นในการจำลองข้อมูล ส่วนการตรวจสอบความแม่นยำของวิธีการประมาณค่าพารามิเตอร์พร้อมกันพิจารณาจากค่าความลำเอียงและค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (RMSE) ของค่าเฉลี่ยค่าประมาณพัฒนาการความสามารถที่ได้แต่ละครั้งที่ทำซ้ำ (50 ครั้ง) ในแต่ละเงื่อนไขกับค่าเฉลี่ยพัฒนาการความสามารถที่กำหนดขึ้นในการจำลองข้อมูล

ตัวแปรที่ศึกษา

1. ตัวแปรอิสระ มี 4 ตัวแปร ได้แก่

1.1 ระดับพัฒนาการความสามารถ 9 ระดับ (ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของความสามารถที่เพิ่มขึ้น) ได้แก่ (0.1, 1), (0.5, 1), (1, 1), (0.1, 0.8), (0.5, 0.8), (1, 0.8), (0.1, 1.2), (0.5, 1.2), และ (1, 1.2)

1.2 ความยาวของแบบทดสอบ 3 ขนาด (จำนวนข้อสอบแบบเลือกตอบต่อจำนวนข้อสอบแบบเขียนตอบ) ได้แก่ ขนาด 30:10 ขนาด 24:8 และขนาด 15:5

1.3 ระดับความยากของข้อสอบ 3 ระดับ (ความยากเฉลี่ยของแบบทดสอบผสมในการทดสอบครั้งที่ 1) ได้แก่ - 0.50, 0.00, และ 0.50

1.4 ระดับการให้คะแนนของข้อสอบแบบเขียนตอบ 3 รูปแบบ ได้แก่ 3 ระดับ (0, 1, และ 2 คะแนน) 4 ระดับ (0, 1, 2, และ 3 คะแนน) และ 5 ระดับ (0, 1, 2, 3, และ 4 คะแนน)

2. ตัวแปรตาม

2.1 ความแม่นยำของค่าประมาณความสามารถ

2.2 ความแม่นยำของวิธีประมาณค่าพารามิเตอร์พร้อมกัน

ผลการวิจัย

ผลการวิจัยพบว่า

1. ค่าประมาณพารามิเตอร์ความสามารถจากการใช้วิธีการประมาณค่าพารามิเตอร์พร้อมกันกับความสามารถจริง มีความสัมพันธ์ทางลบในระดับสูงอย่างมีนัยสำคัญทางสถิติที่ระดับ .01

2. ทุกเงื่อนไขของความยาว ความยากและรูปแบบการให้คะแนน และในทุกระดับพัฒนาการความสามารถทั้งหมดที่ศึกษา ความแม่นยำของวิธีการประมาณค่าพารามิเตอร์พร้อมกันเมื่อระดับพัฒนาการความสามารถที่มีส่วนเบี่ยงเบนเป็น 0.80 และ 1.00 สูงกว่าเมื่อระดับพัฒนาการที่มีส่วนเบี่ยงเบนเป็น 1.2 ในทุกระดับความยาว ความยากและรูปแบบการให้คะแนนของแบบทดสอบผสม แต่ไม่มีนัยสำคัญที่ระดับ .05

3. เมื่อแบบทดสอบผสมมีความยาก และรูปแบบการให้คะแนนคงที่ แบบทดสอบผสมขนาด 24:8 มีความแม่นยำของวิธีการประมาณค่าพารามิเตอร์พร้อมกันสูงกว่าแบบทดสอบผสมขนาด 15:5 และขนาด 30:10 อย่างมีนัยสำคัญที่ระดับ .05

4. เมื่อแบบทดสอบผสมมีความยาว และรูปแบบการให้คะแนนคงที่ แบบทดสอบผสมที่มีระดับความยากของใช้ในการสอบครั้งที่ 1 เท่ากับความสามารถในการสอบครั้งที่ 1 (0.00) มีความแม่นยำของวิธีการประมาณค่าพารามิเตอร์พร้อมกันต่ำกว่า แบบทดสอบที่มีความยากไม่เท่ากับความสามารถในการสอบครั้งที่ 1 (-0.50 และ 0.50) อย่างมีนัยสำคัญที่ระดับ .05

5. เมื่อแบบทดสอบผสมมีความยาว และความยากคงที่ แบบทดสอบผสมที่ให้คะแนน 3 ระดับมีความแม่นยำของของวิธีการประมาณค่าพารามิเตอร์พร้อมกัน สูงกว่ากรณีให้คะแนน 4 และ 5 ระดับ อย่างมีนัยสำคัญที่ระดับ .05

อภิปรายผลการวิจัย

ผู้วิจัยอภิปรายผลการวิจัยโดยมีรายละเอียดดังต่อไปนี้

1. ผลการวิจัย พบว่า “ค่าประมาณพารามิเตอร์ความสามารถจากการใช้วิธีการประมาณค่าพารามิเตอร์พร้อมกันกับความสามารถจริงมีความสัมพันธ์ทางลบในระดับสูงอย่างมีนัยสำคัญทางสถิติที่ระดับ.01” โดยที่ค่าประมาณความสามารถของผู้สอบที่ได้จากวิธีการประมาณค่าพารามิเตอร์พร้อมกันในการทดสอบครั้งที่ 1 มีความสัมพันธ์ทางลบในระดับสูงกับความสามารถจริงในการทดสอบครั้งที่ 1 และค่าประมาณความสามารถของผู้สอบที่ได้จากวิธีการประมาณค่าพารามิเตอร์พร้อมกันในการทดสอบครั้งที่ 2 มีความสัมพันธ์ทางลบในระดับสูงกับความสามารถจริงในการทดสอบครั้งที่ 2 นั่นแสดงว่าผลที่เกิดขึ้นจะไม่ทำให้ค่าพัฒนาการความสามารถซึ่งคำนวณจากผลต่างของค่าประมาณพารามิเตอร์ความสามารถในการทดสอบสองครั้งที่ได้จากวิธีการประมาณค่าพารามิเตอร์พร้อมกันจะมีลักษณะการแจกแจงที่แตกต่างจากพัฒนาการความสามารถจริง แต่อย่างไรก็ตาม การที่ค่าประมาณพารามิเตอร์ความสามารถในการทดสอบแต่ละครั้งที่ได้จากวิธีการประมาณค่าพารามิเตอร์พร้อมกันบิดเบือนไปจากความสามารถจริงนั้นน่าจะเป็นเพราะวิธีการประมาณค่าพารามิเตอร์พร้อมกันมีการปรับให้ค่าประมาณความสามารถให้อยู่บนมาตรวัดเดียวกัน และอาจเนื่องมาจากความแปรปรวนของกลุ่มผู้สอบในการทดสอบสองครั้งไม่เท่ากันซึ่งเป็นข้อจำกัดหนึ่งของโปรแกรม MULTILOG ที่กำหนดว่าความแปรปรวนในแต่ละกลุ่มต้องมีค่าเท่ากัน ดังนั้นการเลือกใช้โปรแกรม MULTILOG ในการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบสองกลุ่มที่มีความสามารถแตกต่างกันจึงอาจส่งผลต่อค่าประมาณความสามารถได้ ปัญหาที่เกิดขึ้นในการวิจัยครั้งนี้เป็นไปในลักษณะเช่นเดียวกับผลการวิจัยที่ผ่านมาที่พบว่าเมื่อกลุ่มผู้สอบสองกลุ่มมีความสามารถไม่เท่าเทียมกันค่าพารามิเตอร์ที่ได้จากวิธีการประมาณค่าพารามิเตอร์พร้อมกันจะมีความคลาดเคลื่อนมากกว่าวิธีอื่น (Béguin, Hanson and Glas. 2000 ; Kim and Kolen. 2006 ; Kim and Lee. 2006 ; Li, Lissitz and Yang. 1999 ; Tate. 2000) แต่วิธีการประมาณค่าพารามิเตอร์พร้อมกันเป็นวิธีการประมาณค่าพารามิเตอร์จากผลการทำแบบทดสอบผสมทั้งสองฉบับพร้อมกันโดยใช้การประมวลผลโดยคอมพิวเตอร์เพียง 1 ครั้ง จึงสามารถยืนยันได้ว่าค่าประมาณพารามิแตอร์อยู่บนมาตรวัดเดียวกัน

2. ผลการวิจัย พบว่า เมื่อส่วนเบี่ยงเบนมาตรฐานของพัฒนาการความสามารถเป็น 0.80 และ 1.00 วิธีการประมาณค่าพารามิเตอร์พร้อมกันมีความแม่นยำสูงกว่าเมื่อส่วนเบี่ยงเบนมาตรฐานของพัฒนาการความสามารถเป็น 1.2 ในทุกเงื่อนไขของแบบทดสอบผสมอย่างไม่มีนัยสำคัญที่ระดับ .05 ซึ่งสอดคล้องกับผลการศึกษาของเปค และ ยัง (Paek and Young. 2005) ที่พบว่าเมื่อค่าสัมบูรณ์ของพัฒนาการเฉลี่ยมีค่ามากขึ้นค่า MAE และ RMSE จะมีค่าสูงขึ้นและเมื่อค่าส่วนเบี่ยงเบนมาตรฐานของพัฒนาความสามารถเป็น 1.2 ค่า MAE และ RMSE สูงกว่าเมื่อค่าส่วนเบี่ยงเบนมาตรฐานของความสามารถที่เพิ่มขึ้นเป็น 0.8 และ 1

3. ผลการวิจัย พบว่า กรณีแบบทดสอบผสมขนาด 24 : 8 ความแม่นยำของวิธีการประมาณค่าพารามิเตอร์พร้อมกันต่ำกว่ากรณีแบบทดสอบผสมขนาด 30 : 10 และ 15 : 5 ซึ่งไม่สอดคล้องกับจากสมมติฐานที่ตั้งไว้ว่า “แบบทดสอบผสมขนาด 30 : 10 น่าจะมีความแม่นยำของวิธีการประมาณค่าพารามิเตอร์พร้อมกันค่าสูงกว่าแบบทดสอบผสมขนาด 24 : 8 และ 15 : 5” อาจจะเนื่องมาจากขนาดจำนวนข้อสอบร่วมที่ใช้ศึกษาสำหรับแบบทดสอบแต่ละขนาดแตกต่างกันดังเช่นผลการวิจัยของบาสตารี (Bastari. 2000) และอัญชลี ศรีกลชาญ (2552) ที่พบว่าเมื่อกลุ่มผู้สอบที่มีความสามารถไม่เท่าเทียมกัน วิธีการประมาณค่าพารามิเตอร์

พร้อมกันจะมีความคลาดเคลื่อนลดลงหากข้อสอบที่ตรวจให้คะแนนสองค่าที่ใช้เป็นข้อสอบร่วมมีจำนวนลดลง การวิจัยครั้งนี้ใช้ข้อสอบร่วมที่ผสมข้อสอบทั้ง 2 ประเภท โดยกำหนดจำนวนข้อสอบร่วมสำหรับแบบทดสอบผสมขนาด 30 : 10, 24 : 8 และ 15 : 5 ไว้ที่ 10 ข้อ : 4 ข้อ, 8 ข้อ : 3 ข้อ และ 5 ข้อ : 2 ข้อ ตามลำดับ ซึ่งพิจารณาแล้วจะพบว่าเมื่อข้อสอบแบบให้คะแนนหลายค่าในข้อสอบร่วมมีจำนวนคงที่ แบบทดสอบผสมขนาด 24 : 8 จะมีจำนวนข้อสอบแบบให้คะแนนสองค่ามากกว่าแบบทดสอบผสมขนาด 30 : 10 และ 15 : 5 ดังนั้นจึงอาจจะเป็นไปได้ว่าจำนวนข้อสอบแต่ลักษณะในข้อสอบร่วมนั่นเองที่ส่งผลต่อความแม่นยำของวิธีการประมาณค่าพารามิเตอร์พร้อมกัน แต่อย่างไรก็ตามไม่สามารถสรุปได้อย่างชัดเจนเนื่องจากจำนวนข้อสอบร่วมหรือจำนวนข้อสอบแต่ละลักษณะมีสัดส่วนที่ใกล้เคียงกันมาก

4. ผลการวิจัย พบว่า ความแม่นยำของวิธีการประมาณค่าพารามิเตอร์พร้อมกันกรณีระดับความยากของแบบทดสอบผสมที่ใช้ในการสอบครั้งที่ 1 เท่ากับความสามารถในการสอบครั้งที่ 1 (0.00) ต่ำกว่ากรณีความยากของแบบทดสอบผสมที่ใช้ในการสอบครั้งที่ 1 ไม่สอดคล้องกับความสามารถในการสอบครั้งที่ 1 (-0.50, 0.50) ซึ่งไม่สอดคล้องสมมติฐานที่ตั้งไว้ว่า “ความแม่นยำของวิธีการประมาณค่าพารามิเตอร์พร้อมกันเมื่อระดับความยากของแบบทดสอบผสมที่ใช้ในการสอบครั้งที่ 1 เท่ากับความสามารถในการสอบครั้งที่ 1 (0.00) น่าจะมีความแม่นยำมากกว่าเมื่อความยากของแบบทดสอบผสมที่ใช้ในการสอบครั้งที่ 1 ไม่สอดคล้องกับความสามารถในการสอบครั้งที่ 1 (-0.50,0.50)” ทั้งนี้อาจจะเนื่องมาจากการกำหนดให้แบบทดสอบสำหรับทดสอบครั้งที่สองมีค่าเฉลี่ยความยากเพิ่มขึ้น .50 จากการทดสอบครั้งแรก และการประมาณค่าพารามิเตอร์พร้อมกันเป็นการประมาณค่าพารามิเตอร์ความสามารถสองครั้งร่วมกัน แต่ข้อสมมติฐานที่ตั้งไว้นั้นไม่ได้ครอบคลุมถึงการแจกแจงความสามารถของผู้สอบในการทดสอบในครั้งที่สองด้วย ซึ่งการแจกแจงความสามารถของผู้สอบมีผลที่สำคัญมากต่อการปรับเทียบพารามิเตอร์พร้อมกัน โดยกรณีที่ผู้สอบสองกลุ่มผู้สอบมีความสามารถไม่เท่าเทียมกันแล้วความลำเอียง และ RMSE จะสูงกว่ากรณีที่ผู้สอบสองกลุ่มมีความสามารถเท่าเทียมกัน (Cao. 2008 : 91)

5. ผลการวิจัยพบว่าความแม่นยำของวิธีการประมาณค่าพารามิเตอร์พร้อมกันกรณีให้คะแนน 3 ระดับสูงกว่ากรณีให้คะแนน 4 และ 5 ระดับ ซึ่งไม่สอดคล้องกับสมมติฐานที่ตั้งไว้ว่า “ความแม่นยำของวิธีการประมาณค่าพารามิเตอร์พร้อมกัน เมื่อมีการให้คะแนนข้อสอบแบบ 5 ระดับ น่าจะสูงกว่าการให้คะแนน 3 ระดับ หรือ 4 ระดับ” อาจจะเนื่องมาจากสัดส่วนจำนวนข้อสอบแต่ละลักษณะในแบบทดสอบผสมแตกต่างกัน ดังผลของการศึกษาของลี ลิสิท และหยาง (Li, Lissitz and Yang. 1999) บาสตารี (Bastari. 2000) เทท (Tate. 2000) และอัญชลี ศรีกลชาญ (2552) พบว่า เมื่อกลุ่มผู้สอบมีความสามารถไม่เท่าเทียมกัน ผลการปรับเทียบจะมีความคลาดเคลื่อนลดลงเมื่อแบบทดสอบผสมที่มีข้อสอบทั้ง 2 ชนิดมีจำนวนข้อสอบที่ตรวจให้คะแนนสองค่ามากขึ้น ในการวิจัยครั้งนี้ผู้วิจัยได้กำหนดเงื่อนไขให้ข้อสอบแบบเขียนตอบมีรูปแบบการให้คะแนนเป็น 3 ระดับ 4 ระดับ และ 5 ระดับ โดยที่การให้ระดับคะแนนของข้อสอบแบบเขียนตอบมี 3 ระดับ (0/1/2) นั้นจะทำให้แบบทดสอบผสมขนาด 30:10 24:8 และ 15:5 มีสัดส่วนน้ำหนักคะแนนรวมของข้อสอบเลือกตอบสูงกว่าคะแนนรวมของแบบทดสอบให้คะแนนหลายค่า (30 คะแนน : 20 คะแนน, 24 คะแนน : 16 คะแนน, 15 คะแนน : 10 คะแนน) การให้คะแนนข้อสอบแบบเขียนตอบ 4 ระดับ (0/1/2/3) จะทำให้ข้อสอบแบบเลือกตอบมีคะแนนรวมเท่ากับข้อสอบแบบเขียนตอบ (30 คะแนน : 30 คะแนน, 24 คะแนน : 24 คะแนน, 15 คะแนน : 15 คะแนน) และการให้คะแนนข้อสอบแบบเขียนตอบ 5 ระดับ (0/1/2/3/4) จะทำให้ข้อสอบแบบเลือกตอบมี

คะแนนรวมต่ำกว่าข้อสอบแบบเขียนตอบ (30 คะแนน : 40 คะแนน, 24 คะแนน : 32 คะแนน, 15 คะแนน : 20 คะแนน) ซึ่งจะเห็นว่าเมื่อปรับลักษณะของแบบทดสอบทุกขนาดในรูปของคะแนนแล้วจะพบว่าการให้ระดับคะแนนข้อสอบแบบเขียนตอบมี 3 ระดับจะทำให้น้ำหนักคะแนนของข้อสอบแบบเลือกตอบสูงกว่าข้อสอบแบบเขียนตอบ อย่างไรก็ตามหากพิจารณาจำนวนระดับการให้คะแนนข้อสอบแบบเขียนตอบ 5 ระดับซึ่งใช้โมเดลเจอเนอรอลไรส์พาร์เชียลเครดิตประมาณค่า จะมีระดับขั้นการตอบ 4 ขั้น ซึ่งในการวิจัยครั้งนี้กำหนดให้ระดับความยากประจำขั้นเพิ่มขึ้นขั้นละ .50 ด้วยเหตุผลดังกล่าวจึงอาจจะเป็นปัจจัยหนึ่งที่ทำให้ความแม่นยำของวิธีการประมาณค่าพารามิเตอร์พร้อมกันเมื่อแบบทดสอบผสมให้คะแนนข้อสอบแบบเขียนตอบ 5 ระดับต่ำกว่ากรณีให้คะแนน 3 และ 4 ระดับ

ข้อเสนอแนะ

1. ข้อเสนอแนะในการนำผลวิจัยไปใช้

1.1 การศึกษาครั้งนี้ผู้วิจัยได้ศึกษาโดยการจำลองข้อมูล ดังนั้นในสถานการณ์จริง ข้อมูลสำหรับใช้ศึกษาวิธีการประมาณค่าพารามิเตอร์พร้อมกันด้วยโมเดล 3PL ร่วมกับ GPCM อาจจะมีผลที่แตกต่างออกไป เนื่องจากความเป็นไปได้ที่จะเกิดการละเมิดข้อตกลงเบื้องต้นมีสูง โดยเฉพาะผลการตอบข้อสอบแบบผสมที่มีข้อสอบหลายลักษณะในฉบับเดียวกันหรือการใช้โมเดลผสมสำหรับการประมาณค่าพารามิเตอร์ ตลอดจนปัจจัยอื่นที่เกี่ยวข้องกับความแม่นยำของวิธีการปรับเทียบได้เช่น จำนวนข้อสอบร่วม ลักษณะข้อสอบร่วม ลักษณะของแบบทดสอบผสม เป็นต้น

1.2 จากผลการวิจัยในภาพรวมพบว่าเมื่อผู้สอบมีพัฒนาการความสามารถเพิ่มขึ้น 0.1 วิธีการประมาณค่าพารามิเตอร์พร้อมกันมีความแม่นยำสูงกว่า กรณีที่ผู้สอบมีความสามารถเพิ่มขึ้น 0.5 หรือ 1.0 นั้น ดังนั้นการใช้วิธีการประมาณค่าพารามิเตอร์กันอาจจะต้องพิจารณาก่อนตัดสินใจนำไปใช้ อย่างไรก็ตามลักษณะความสามารถโดยทั่วไปมักจะมีค่าเพิ่มขึ้นไม่มากกว่า 0.5 ดังนั้นวิธีการประมาณค่าพารามิเตอร์พร้อมกัน จึงน่าจะมีความแม่นยำในการนำไปใช้ศึกษาพัฒนาการความสามารถของผู้สอบทางด้านการศึกษาได้

2. ข้อเสนอแนะในทำวิจัยครั้งต่อไป

2.1 ควรศึกษาผลของการเปลี่ยนแปลงค่าความยากของข้อสอบเฉพาะที่ไม่ใช่ข้อสอบร่วมว่าจะส่งผลอย่างไรต่อผลการประมาณค่าพารามิเตอร์พร้อมกัน เมื่อขนาด ทิศทางและการกระจายความยากของข้อสอบแตกต่างกัน

2.2 ควรศึกษาเปรียบเทียบผลการวัดพัฒนาการความสามารถที่ได้จากวิธีการประมาณค่าพารามิเตอร์พร้อมกันกับวิธีการวัดพัฒนาการหรือการวัดการเปลี่ยนแปลงวิธีอื่น

2.3 ศึกษาผลการใช้วิธีการประมาณค่าพารามิเตอร์พร้อมกัน เมื่อมีทดสอบมากกว่าสองครั้งหรือใช้แบบทดสอบผสมหลายฉบับ

2.4 ควรมีการศึกษาผลการใช้แบบทดสอบผสมสำหรับศึกษาพัฒนาการความสามารถของผู้สอบในสถานการณ์จริงเพื่อตรวจสอบองค์ความรู้ที่ได้จากการศึกษากับสถานการณ์การจำลอง

2.5 ควรมีการศึกษาผลการใช้โมเดลหลายมิติในการประมาณค่าผลการทดสอบด้วยแบบทดสอบผสมเพื่อดูความเหมาะสม ข้อจำกัดและความเป็นไปได้ในการนำไปใช้จริงเมื่อเทียบกับการใช้โมเดลแบบมิติเดียว

กิตติกรรมประกาศ งานวิจัยนี้ได้รับทุนอุดหนุนการวิจัยจากมหาวิทยาลัยศรีนครินทรวิโรฒประเภททุนงบประมาณแผ่นดิน ประจำปีงบประมาณ 2553 และขอบคุณมหาวิทยาลัยเชียงใหม่ในการสนับสนุน “ทุนพัฒนาจารย์”

เอกสารอ้างอิง

อัญชลี ศรีกลชาญ. (2552). **คุณภาพของการปรับเทียบคะแนนสำหรับแบบสอบรูปแบบผสม : การประยุกต์ใช้การปรับเทียบตามทฤษฎีการตอบสนองข้อสอบด้วยวิธีโค้งคุณลักษณะและการปรับค่าพารามิเตอร์พร้อมกัน.** วิทยานิพนธ์ครุศาสตรดุษฎีบัณฑิต (สาขาวิชาการวัดและประเมินผลการศึกษา). กรุงเทพฯ : จุฬาลงกรณ์มหาวิทยาลัย.

Bastari, B. (2000). **Linking multiple-choice and constructed-response items to a common proficiency scale**. Retrieved November 1, 2008, from, http://proquest.umi.com/pqdlink?did=731794221&Fmt=7&clientId=61839&RQT=309&VName=PQD

Béguin, A. A.; Hanson, B. A. and Glas, C. A. (2000). **Effect of multidimensionality on separate and concurrent estimation in IRT equating.** Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Cao, Y. (2008). **Mixed-Format test equating : Effects of Test dimensionality and common item sets.** Doctoral Dissertation, Department of Measurement Statistics and Evaluation, Faculty of the Graduate School, University of Maryland. Retrieved April 5, 2009 from http://drum.lib.umd.edu/bitstream/1903/8843/1/umi-umd-5871.pdf

DeMars, C. E. (2004). **A Comparison of the Recovery of Parameters Using the Nominal Response and Generalized Partial Credit Models.** Poster presented at the annual meeting of American Educational Research Association, San Diego, CA. Retrieved July 8, 2009 from http://www.jmu.edu/assessment/research/faculty/demars/DemarsAERA04_recovery.PDF

Han, K. T. (2007). **WinGen2 : Windows software that generates IRT parameters and item responses [computer program].** Amherst, MA: University of Massachusetts Amherst, Center for Educational Assessment. Retrieved May 13, 2009, from http://www.umass.edu/remp/software/simcata/wingen/WinGen_Manual_Han_Hambleton_2007.pdf

Hanson, B., and Béguin, A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design**.** **Applied Psychological Measurement.** 26(1) : 3-24.

Hu, H. ; Rogers, W. T. and Vukmirovic, Z. (2008). Investigation of IRT-based equating methods in the presence of outlier common items. **Applied Psychological Measurement.** 32(4) : 311-333.

Kamata, A. and Tate, R. (2005). The Performance of a Method for the Long-term Equating of Mixed-Format Assessment. **Journal of Educational Measurement**. 42, 2 : 193 - 213.

Kim, S. (2004). **Unidimensional IRT Scale Linking Procedures for Mixed-Format Tests and Their Robustness to Multidimensionality.** Unpublished doctoral dissertation, University of Iowa, Iowa City. Retrieved February 8, 2009, from http://proquest.umicom/pqdweb?index=0&did=765922701&SrchMode=1&sid=1&Fmt=14&VInst=PROD&VType=PQD&RQT=309&VName=PQD&TS=1251015767&clientId=29945

Kim, J. (2007). **A Comparison of Calibration Methods and Proficiency Estimators For Creating IRT Vertical Scales.** Theses and Dissertations. Retrieved February 8, 2009, from http://etd.lib.uiowa.edu/2007/jkim.pdf

Kim, S. H. and Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. **Applied Psychological Measurement**. 22, 131–143.

Kim, S. and Kolen, M. J. (2006). Robustness to Format Effects of IRT Linking Methods for Mixed-Format Tests. **Applied Measurement in Education.** 19(4) : 357–381.

Kim, S.H. and Lee, W.C. (2006). An Extension of Four IRT Linking Methods for Mixed-Format Tests. **Journal of Educational Measurement.** 43(1) : 53-76.

Kinsey, T. L. (2003). **A Comparison of IRT and Rasch Procedures in a Mixed–Item Format Test.** Doctor of Philosophy (Educational Research), University of North Texas, USA. Retrieved June 20, 2008, from http://digital.library.unt.edu/permalink/meta-dc-4316 : 1

Kolen, M. J. and Brennan, R.L. (1995). **Test equating: Methods and practices**. New York : Springer-Verlag.

Li, Y. H. ; Lissitz, R.W. and Yang, Y.N. (1999). **Estimating IRT Equating Coefficients for Tests with Polytomously and Dichotomously Scores Items.** Paper presented at the Annual Meeting of the National Council on Measurement in Education (Montreal, Quebec, Canada, April 19 – 23, 1999) Retrieved Jan 20, 2009, from http://eric.ed.gov/ERICWebPortal/custom/portlets/recordDetails/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=ED431800&ERICExtSearch_SearchType_0=no&accno=ED431800

May, K. and Nicewander, W.A. (1998). Measuring change conventionally and adaptively. **Educational and Psychological Measurement**. 58 : 529 – 554.

Muraki, E. and Bock, R. D. (1999). **PARSCALE**. Chicago, IL : Scientific Software.

Paek, I. and Young, M. J. (2005). Investigation of student growth recovery in a fixed-item linking procedure with a fixed-person prior distribution for mixed-format test data. **Applied Measurement in Education**. 18, 199–215.

Park, H. (2000). **Comparison of IRT Models For Ordered Polytomous Response Data. University of Minnesota.** Retrieved November 1, 2008, from http://proquest.umi.com/pqdlink?did=728340421&Fmt=2&clientId=61839&RQT=309&VName=PQD

Tate, R. (2000). Performance of a Proposed Method for the Linking of Mixed Format Tests with Constructed Response and Multiple Choice Items. **Journal of Educational Measurement**. 37(4) : 329 – 346.

Thissen, D. (1991). **Multilog user's guide: Multiple, categorical item analysis and test scoring using item response theory**. Chicago, IL : Scientific Software.

Yao, L. and Mao, X. (2004). **Unidimensional and Multidimensional Estimation of Vertical Scaled Tests with Complex Structure.** Retrieved November 2, 2008, from http://www.ctb.com/media/articles/pdfs/ResearchArticles/Unidimensional.pdf?FOLDER%3C%3Efolder_id=2534374302134982

Yang, S. (2007). **A Comparison of Unidimensional and Multidimensional Rasch Models Using Parameter Estimates and Fit Indices When Assumption of Unidimensionality is Violated.** Doctor of Philosophy, Ohio State University, Educational Policy and Leadership. **Retrieved November 1, 2008**, from, http://www.ohiolink.edu/etd/view.cgi?acc_num=osu1195695378