



การเปรียบเทียบสถิติบูตสแตรปของการออกแบบการปรับเทียบสำหรับแบบสอบรูปแบบผสมระหว่าง วิธีที่ต่างกัน

A COMPARISON OF BOOTSTRAPPED STATISTICS IN EQUATING DESIGNS FOR MIXED FORMAT TEST AMONG DIFFERENT METHODS

นางสาวพนิดา พานิชวัฒน์ *

Panida Panidvadtana

ผศ.ดร.ณัฐภรณ์ หลาวทอง **

Asst. Prof. Nuttaporn Lawthong, Ph.D.

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อ 1) เพื่อเปรียบเทียบสถิติบูตสแตรป (bootstrapped statistics) ของการออกแบบการปรับเทียบคะแนนแบบสอบรูปแบบผสมโดยใช้ข้อสอบร่วมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT) ระหว่างวิธีการตรวจที่แตกต่างกัน 2) เพื่อเปรียบเทียบค่ารากที่สองของความแตกต่างยกกำลังสอง (RMSD) ของการออกแบบการปรับเทียบคะแนนแบบสอบรูปแบบผสมโดยใช้ข้อสอบร่วมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT) ระหว่างวิธีของเลวินและวิธี CLE กลุ่มตัวอย่างเป็นนักเรียนระดับชั้นมัธยมศึกษาปีที่ 4 สำนักงานเขตพื้นที่การศึกษามัธยมศึกษาเขต 1 ปีการศึกษา 2556 จังหวัดกรุงเทพมหานครจำนวน 1,609 คน เครื่องมือที่ใช้ในการวิจัยครั้งนี้ ได้แก่ 1) แบบสอบคู่ขนานวิชาฟิสิกส์ ได้แก่ แบบสอบฉบับใหม่และฉบับอ้างอิง ประกอบด้วยข้อสอบแบบหลายตัวเลือกจำนวน 16 ข้อ และข้อสอบแบบสร้างคำตอบจำนวน 8 ข้อ แบบสอบทั้งสองฉบับมีข้อสอบร่วมจำนวน 12 ข้อ 2) เกณฑ์การให้คะแนนที่ใช้ในการตรวจข้อสอบแบบสร้างคำตอบในแบบสอบฉบับใหม่และฉบับอ้างอิง 3) ผู้ตรวจให้คะแนน 2 คน โดยวิธีการตรวจทั้ง 3 วิธี คือ 1) วิธีตรวจข้อสอบร่วมแบบสร้างคำตอบ โดยไม่มีการตรวจให้คะแนนข้ามกลุ่ม 2) วิธีตรวจข้อสอบร่วมแบบสร้างคำตอบ โดยมีการตรวจให้คะแนนข้ามกลุ่มทั้งหมดของข้อสอบร่วม 3) วิธีตรวจข้อสอบร่วมแบบสร้างคำตอบ โดยมีการตรวจให้คะแนนข้ามกลุ่มครึ่งหนึ่งของข้อสอบร่วม

ผลการวิจัยพบว่า 1) วิธีตรวจข้อสอบร่วมแบบสร้างคำตอบ โดยมีการตรวจให้คะแนนข้ามกลุ่มครึ่งหนึ่งของข้อสอบร่วม เมื่อใช้วิธี CLE มีค่า RMSE ค่า Equating error และค่า Bias ต่ำที่สุด คือ .802, .251 และ .840 ตามลำดับ 2) วิธี CLE มีค่า RMSD จากวิธีการตรวจวิธีที่ 1, 2 และ 3 คือ 1.421, 1.481 และ 1.481 ตามลำดับ ซึ่งเป็นค่าที่ต่ำกว่าวิธีของเลวินจากวิธีการตรวจวิธีที่ 1, 2 และ 3 คือ 1.717, 1.804 และ 1.793 ตามลำดับ

* นิสิตมหาบัณฑิตสาขาวิชาการวัดและประเมินผลการศึกษา ภาควิชาวิจัยและจิตวิทยาการศึกษา

คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

E-mail Address: panidapanidvadtana@hotmail.co.th

** อาจารย์ประจำสาขาวิชาการวัดและประเมินผลการศึกษา ภาควิชาวิจัยและจิตวิทยาการศึกษา

คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

E-mail Address: Nuttaporn.l@chula.ac.th

ISSN 1905-4491

Abstract

The purposes of this research were 1) to compare bootstrapped statistics of equating mixed-format test of nonequivalent groups with anchor test equating design among scoring methods 2) to compare root mean squared difference (RMSD) of equating mixed-format test of nonequivalent groups with anchor test equating design between Levine method and CLE method. The sample were 1,607 students from Matayomsuksa 4. They were used for data collecting with 1) two parallel mixed - format tests of the Physics. There were new form and reference form which consisted 16 multiple-choice items and 8 construct-response items. Those form had 12 items in anchor test 2) rubric scoring used for constructed-response items in Physics test, new and reference forms. 3) two raters. Three scoring methods were 1) constructed-response anchor items scoring methods with no trend scoring method 2) constructed-response anchor items scoring methods with full anchor test trend scoring method 3) constructed-response anchor items scoring methods with half anchor test trend scoring method.

The results of the study revealed that 1) constructed-response anchor items scoring methods with half anchor test trend scoring method with CLE method resulted in smallest RMSE , Equating error and Bias were .802, .251, .840. 2) CLE method had RMSD from scoring methods of 1, 2, 3 method were 1.421, 1.481, 1.481, which lower than those of equating with Levine method were 1.717, 1.804, 1.793.

คำสำคัญ : การออกแบบการปรับเทียบ / แบบสอบรูปแบบผสม / สถิติบูทสเตรป

KEYWORDS : EQUATING DESIGNS / MIXED FORMAT TEST / BOOTSTRAPPED STATISTICS

บทนำ

การทดสอบขนาดใหญ่มีการใช้ข้อสอบแบบสร้างคำตอบ (construct-response: CR) ร่วมกับข้อสอบแบบหลายตัวเลือก (multiple-choice: MC) มากขึ้น โดยข้อสอบแบบหลายตัวเลือก (MC) มีจุดเด่น คือ การให้คะแนนมีความเป็นปรนัยและมีความเที่ยงสูง สามารถวัดได้ครอบคลุมเนื้อหา อีกทั้งยังสามารถวัดระดับผลการเรียนรู้ขั้นสูงได้ เช่น ความรู้ ความเข้าใจ การนำไปใช้ และการวิเคราะห์ ขณะที่ข้อสอบแบบสร้างคำตอบ (CR) ขาดความเป็นปรนัยและความเที่ยงในการให้คะแนน สร้างให้ครอบคลุมเนื้อหาทั้งหมดทำได้ยาก (Messick, 1994 cited in Kim *et al.*, 2010) แต่อย่างไรก็ตาม ข้อสอบแบบสร้างคำตอบสามารถวัดระดับผลการเรียนรู้ขั้นสูงได้ดีกว่าข้อสอบแบบอื่น สร้างได้ง่าย อีกทั้งยังส่งเสริมผู้สอบได้แสดงความคิดเห็น ความคิดสร้างสรรค์ และทัศนคติ (ศิริชัย กาญจนวาสี, 2555) เนื่องจากข้อสอบแบบหลายตัวเลือกและแบบสร้างคำตอบต่างก็มีจุดแข็งและจุดอ่อน ดังนั้นในการทดสอบจึงมีการสร้างแบบสอบรูปแบบผสม ที่ประกอบด้วยข้อสอบแบบสร้างคำตอบและข้อสอบแบบหลายตัวเลือก การทดสอบที่ใช้แบบสอบรูปแบบผสม เช่น แบบสอบทางการศึกษาแห่งชาติขั้นพื้นฐาน (O-NET) การทดสอบภาษาอังกฤษ TOEFL เป็นต้น ในสถานการณ์ทดสอบปัจจุบันมีความจำเป็นต้องใช้แบบสอบหลายฉบับในการวัดคุณลักษณะเดียว ซึ่งการพัฒนาแบบสอบแต่ละฉบับต้องมั่นใจว่าแบบสอบนั้น ๆ มีเนื้อหาและคุณลักษณะทางสถิติ (statistical characteristics) คล้ายคลึงกัน และสำหรับแบบสอบที่บรรจุข้อสอบแบบสร้างคำตอบ ต้องมีเกณฑ์การให้คะแนน (scoring rubrics) เฉพาะข้อสอบแต่ละข้อ (Kim *et al.*, 2010b) ด้วยเหตุนี้จึงมีความจำเป็นต้องใช้กระบวนการปรับเทียบคะแนน (equating process) เพื่อปรับเทียบผลที่ได้จากการทำแบบสอบต่างฉบับให้อยู่ในมาตรวัดเดียวกัน เพื่อให้คะแนนของแบบสอบต่างฉบับสามารถนำมาเปรียบเทียบกันได้

การปรับเทียบคะแนนระหว่างแบบสอบต้องพิจารณาถึงการออกแบบการปรับเทียบคะแนน (equating designs) ซึ่งในการปรับเทียบแบบสอบรูปแบบผสมนั้น ข้อสอบร่วมที่ใช้ควรเป็นข้อสอบแบบหลาย

ตัวเลือกผสมกับข้อสอบแบบสร้างคำตอบ แต่การใช้ข้อสอบแบบสร้างคำตอบอาจจะมีปัญหาเกี่ยวกับมาตรฐานการให้คะแนน จึงมีการประยุกต์การตรวจให้คะแนนข้ามกลุ่ม (trend scoring method) โดยผู้ประเมินกลุ่มหนึ่งทำการตรวจข้อสอบรวมแบบสร้างคำตอบของผู้สอบทั้ง 2 กลุ่ม ซึ่ง Kim *et al.* (2010) พบว่า การประยุกต์การตรวจให้คะแนนข้ามกลุ่ม (trend scoring method) ให้ค่าความลำเอียง (bias) และค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสอง (RMSE) ในการปรับเทียบน้อยกว่ารูปแบบที่ใช้ข้อสอบรวมแบบหลายตัวเลือกและรูปแบบที่ใช้ข้อสอบรวมแบบผสมที่ไม่มีการตรวจให้คะแนนข้ามกลุ่มในข้อสอบรวมแบบสร้างคำตอบ

การออกแบบวิธีการเก็บรวบรวมข้อมูล (designs for data collection) เพื่อการปรับเทียบคะแนนที่ได้รับความนิยมอีกรูปแบบหนึ่ง คือ รูปแบบผู้สอบกลุ่มไม่เท่าเทียมกัน โดยใช้ข้อสอบรวม ซึ่งผู้สอบแต่ละกลุ่มทำแบบสอบฉบับเดียว โดยแบบสอบแต่ละฉบับมีข้อสอบรวมกันจำนวนหนึ่ง มักใช้กับสถานการณ์ที่กลุ่มผู้สอบต่างประชากรกัน เช่น กลุ่มผู้สอบต่างโปรแกรม เวลา หรือ ระดับ เป็นต้น (ศิริชัย กาญจนวาสี, 2555)

การปรับเทียบคะแนนต้องใช้วิธีการปรับเทียบคะแนน (equating method) ในการเชื่อมโยงคะแนนซึ่งมีอยู่หลายวิธี ไม่ว่าจะเป็นวิธีการปรับเทียบค่าเฉลี่ย วิธีปรับเทียบเชิงเส้นตรง วิธีการปรับเทียบอิกวิเปอร์เซ็นต์ไทล์ และวิธีปรับเทียบโดยใช้ทฤษฎีการตอบสนองข้อสอบ ซึ่งงานวิจัยที่ผ่านมาพบว่าวิธีการปรับเทียบโดยใช้ทฤษฎีการตอบสนองข้อสอบจะให้ผลการปรับเทียบที่ดี แต่อย่างไรก็ตามวิธีการดังกล่าวจำเป็นต้องใช้กลุ่มข้อมูลขนาดใหญ่ ในกรณีที่กลุ่มตัวอย่างมีขนาดเล็กนิยมใช้วิธีการปรับเทียบแบบดั้งเดิม เช่น วิธี CLE (Chained Linear Equating), วิธีของ Tucker (Tucker method), วิธีของเลวิน (Levine method), วิธีการประมาณค่าความถี่ (frequency estimate) หรือวิธีการปรับเทียบอิกวิเปอร์เซ็นต์ไทล์แบบลูกโซ่ (chained equipercentile equating) โดยวิธีของเลวิน (Levine method) มีความเหมาะสมมากกว่าเมื่อกลุ่มตัวอย่างแต่ละกลุ่มมีความสามารถแตกต่างกัน และวิธีการดังกล่าวจะมีความถูกต้องมากขึ้นเมื่อกลุ่มตัวอย่างมีขนาดใหญ่ (Mroch *et al.*, 2009) เช่นเดียวกับวิธี CLE (Chained Linear Equating) ที่สามารถนำมาใช้กับกลุ่มตัวอย่างใหม่และกลุ่มเก่าที่มีความสามารถแตกต่างกัน หรือเมื่อข้อสอบรวมกับข้อสอบทั้งหมดมีความสัมพันธ์กันสูงได้ (Puhan, 2010) ดังนั้นวิธีการปรับเทียบที่เหมาะสมกับกลุ่มตัวอย่างขนาดเล็กที่มีความสามารถแตกต่างกัน และมีการเก็บข้อมูลจากการทดสอบจริง คือ วิธีของเลวิน (Levine method) และวิธี CLE (Chained Linear Equating)

ผู้วิจัยจึงสนใจศึกษาการออกแบบการปรับเทียบคะแนน ภายใต้รูปแบบผู้สอบกลุ่มไม่เท่าเทียมกัน โดยใช้ข้อสอบรวม (NEAT) ซึ่งข้อสอบรวมที่ใช้เป็นข้อสอบรูปแบบผสมที่มีการตรวจให้คะแนนแบบข้ามกลุ่มในข้อสอบรวมแบบสร้างคำตอบ และไม่มีการตรวจให้คะแนนแบบข้ามกลุ่ม แต่ในการตรวจให้คะแนนแบบข้ามกลุ่มในงานวิจัยที่ผ่านมา ผู้ตรวจกลุ่มหนึ่ง (ผู้ตรวจกลุ่ม B) ตรวจข้อสอบรวมแบบสร้างคำตอบทั้งสองฉบับ ซึ่งผู้ตรวจกลุ่มดังกล่าวต้องรับหน้าที่ในการตรวจข้อสอบจำนวนมาก ด้วยเหตุนี้ผู้วิจัยจึงต้องการลดภาระของผู้ตรวจกลุ่มหนึ่ง/คนหนึ่งลง แต่ยังคงใช้การตรวจให้คะแนนข้ามกลุ่ม (trend scoring method) ด้วยการออกแบบให้ผู้ตรวจแต่ละกลุ่มตรวจข้อสอบรวมแบบสร้างคำตอบของผู้สอบทั้ง 2 กลุ่มอย่างละเท่า ๆ กัน นอกจากนี้ในงานวิจัยที่ผ่านมาใช้ผู้ตรวจ 2 กลุ่ม คือ ผู้ตรวจกลุ่ม A และผู้ตรวจกลุ่ม B ผู้วิจัยจึงสนใจศึกษากรณีที่ผู้ตรวจ 2 คน คือ ผู้ตรวจ 1 และ ผู้ตรวจ 2 เพื่อเพิ่มทางเลือกในการนำไปใช้งานจริง

วัตถุประสงค์

1. เพื่อเปรียบเทียบสถิติบูทสเตรป (bootstrapped statistics) ของการออกแบบการปรับเทียบคะแนนแบบสอบรูปแบบผสมโดยใช้ข้อสอบร่วมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT) ระหว่างวิธีการตรวจที่แตกต่างกัน

2. เพื่อเปรียบเทียบค่ารากที่สองของค่าเฉลี่ยความแตกต่างยกกำลังสอง (RMSD) ของการออกแบบการปรับเทียบคะแนนแบบสอบรูปแบบผสมโดยใช้ข้อสอบร่วมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT) ระหว่างวิธีของเลวิน (Levine method) และวิธี CLE (Chained Linear Equating)

สมมติฐานการวิจัย

1. วิธีตรวจข้อสอบร่วมแบบสร้างคำตอบ โดยมีการตรวจให้คะแนนข้ามกลุ่มทั้งสองวิธีมีค่าสถิติบูทสเตรป (bootstrapped statistics) ต่ำกว่าวิธีตรวจข้อสอบร่วมแบบสร้างคำตอบ โดยไม่มีการตรวจให้คะแนนข้ามกลุ่ม สำหรับการออกแบบการปรับเทียบคะแนนแบบสอบรูปแบบผสมโดยใช้ข้อสอบร่วมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT)

2. ค่ารากที่สองของค่าเฉลี่ยความแตกต่างยกกำลังสอง (RMSD) ของวิธีของเลวิน (Levine method) มีค่าต่ำกว่าวิธี CLE (Chained Linear Equating) สำหรับการออกแบบการปรับเทียบคะแนนแบบสอบรูปแบบผสมโดยใช้ข้อสอบร่วมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT)

ขอบเขตของการวิจัย

1. ตัวแปรอิสระ ได้แก่

1.1 วิธีการตรวจ 3 วิธี ได้แก่ 1) วิธีตรวจข้อสอบร่วมแบบสร้างคำตอบ โดยไม่มีการตรวจให้คะแนนข้ามกลุ่ม 2) วิธีตรวจข้อสอบร่วมแบบสร้างคำตอบ โดยมีการตรวจให้คะแนนข้ามกลุ่มทั้งหมดของข้อสอบร่วม 3) วิธีตรวจข้อสอบร่วมแบบสร้างคำตอบ โดยมีการตรวจให้คะแนนข้ามกลุ่มครึ่งหนึ่งของข้อสอบร่วม

1.2 วิธีการปรับเทียบคะแนน ได้แก่ วิธีของเลวิน (Levine method) และวิธี CLE (Chained Linear Equating)

2. ตัวแปรตาม ได้แก่

2.1 สถิติบูทสเตรป (bootstrapped statistics) ของการปรับเทียบคะแนน ได้แก่ ค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสอง (RMSE) ค่าความลำเอียง (bias) และค่าความคลาดเคลื่อนในการปรับเทียบ (equating error) ซึ่งใช้ในการเปรียบเทียบวิธีการตรวจทั้ง 3 วิธี

2.2 ค่ารากที่สองของค่าเฉลี่ยความแตกต่างยกกำลังสอง (RMSD) ซึ่งใช้ในการเปรียบเทียบวิธีการปรับเทียบเชิงเส้นตรงทั้ง 2 วิธี

วิธีดำเนินการวิจัย

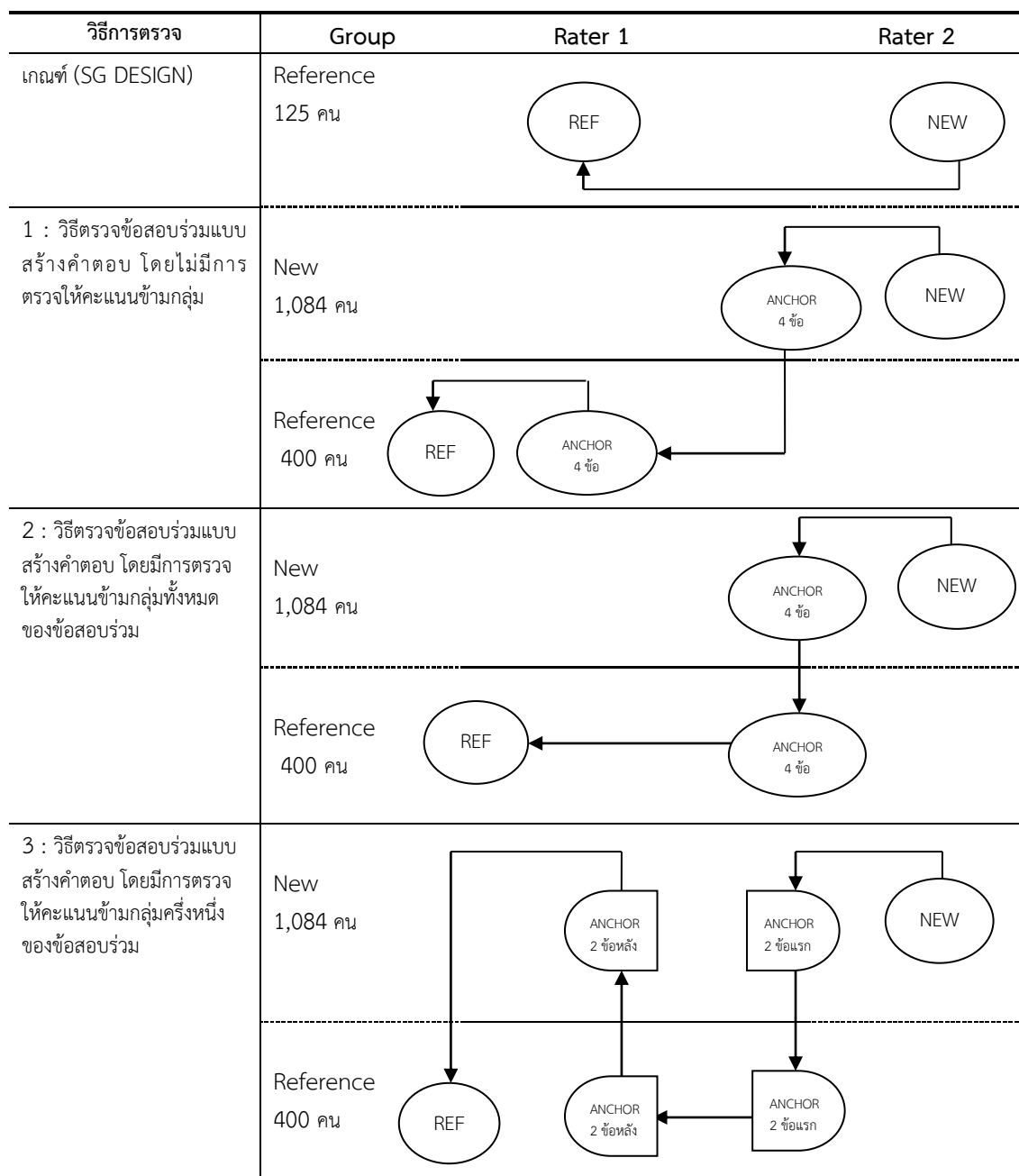
การวิจัยนี้เป็นงานวิจัยเชิงทดลองที่มีวิธีการตรวจ 3 วิธี สำหรับการออกแบบการปรับเทียบคะแนนแบบสอบรูปแบบผสมโดยใช้ข้อสอบร่วมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT) ได้แก่ 1) วิธีตรวจข้อสอบร่วมแบบสร้างคำตอบ โดยไม่มีการตรวจให้คะแนนข้ามกลุ่ม โดยผู้ตรวจ 1 ตรวจแบบสอบฉบับใหม่ (new) และผู้ตรวจ 2 ตรวจแบบสอบฉบับอ้างอิง (reference) 2) วิธีตรวจข้อสอบร่วมแบบสร้างคำตอบ โดยมีการตรวจให้คะแนนข้ามกลุ่มทั้งหมดของข้อสอบร่วม (ผู้ตรวจคนหนึ่ง ตรวจข้อสอบร่วมแบบสร้างคำตอบของผู้สอบทั้ง 2 กลุ่ม) โดยผู้ตรวจ 2 ตรวจแบบสอบฉบับใหม่ (new) และตรวจข้อสอบร่วมแบบสร้างคำตอบของผู้สอบกลุ่มใหม่และกลุ่มอ้างอิง ส่วนผู้ตรวจ 1 ตรวจแบบสอบฉบับอ้างอิง (reference) ส่วนที่เหลืออยู่ 3) วิธีตรวจข้อสอบร่วม

แบบสร้างคำตอบ โดยมีการตรวจให้คะแนนข้ามกลุ่มครั้งหนึ่งของข้อสอบรวม (ผู้ตรวจทั้ง 2 คน ตรวจข้อสอบรวมแบบสร้างคำตอบของผู้สอบทั้ง 2 กลุ่ม โดยแบ่งกันตรวจคนละครั้งหนึ่งของข้อสอบรวม) โดยผู้ตรวจ 2 ตรวจแบบสอบฉบับใหม่ (new) และตรวจข้อสอบรวมแบบสร้างคำตอบ (2 ข้อแรก) ของผู้สอบกลุ่มใหม่และกลุ่มอ้างอิง ส่วนผู้ตรวจ 1 ตรวจแบบสอบฉบับอ้างอิง (reference) และตรวจข้อสอบรวมแบบสร้างคำตอบ (2 ข้อหลัง) ของผู้สอบกลุ่มใหม่และกลุ่มอ้างอิง โดยการออกแบบการเปรียบเทียบที่เป็นเกณฑ์ ใช้รูปแบบผู้สอบกลุ่มเดียว โดยผู้สอบในรูปแบบดังกล่าวทำแบบสอบฉบับใหม่ (new) หลังจากนั้นผ่านไปประมาณ 3 สัปดาห์ ผู้สอบกลุ่มเดียวกันนี้ทำแบบสอบฉบับอ้างอิง (reference) ซึ่งแบบสอบทั้ง 2 ฉบับนี้เป็นแบบสอบคู่ขนานที่มีข้อสอบรวมรูปแบบผสม โดยผู้ตรวจ 1 ตรวจแบบสอบฉบับใหม่ (new) และผู้ตรวจ 2 ตรวจแบบสอบฉบับอ้างอิง (reference) ในการเชื่อมต่อคะแนนจากแบบสอบฉบับใหม่/ผู้ตรวจ 2 ไปยังคะแนนจากแบบสอบฉบับอ้างอิง/ผู้ตรวจ 1 ใช้วิธีการเปรียบเทียบอีควิเปอร์เซ็นไทล์แบบลูกโซ่ ที่ปรับการแจกแจงคะแนนด้วยล็อกเชิงเส้น (chained equipercentile equating with loglinear presmoothed) 1 ดังภาพ 1

การกำหนดประชากรและกลุ่มตัวอย่าง

ประชากรในการวิจัยครั้งนี้ เป็นนักเรียนช่วงชั้นที่ 3 ระดับชั้นมัธยมศึกษาปีที่ 4 ปีการศึกษา 2556 ของโรงเรียนในเขตพื้นที่การศึกษาที่ 1 จังหวัดกรุงเทพมหานคร จำนวน 67 โรงเรียน จำนวน 121,710 คน เป็นนักเรียนระดับมัธยมศึกษาปีที่ 4 จำนวน 20,585 คน

กลุ่มตัวอย่างที่ใช้ในการวิจัยครั้งนี้ คำนวณขนาดกลุ่มตัวอย่างจากนักเรียนระดับชั้นมัธยมศึกษาปีที่ 4 จำนวน 20,585 คน โดยใช้จากตารางแสดงขนาดของกลุ่มตัวอย่างที่เหมาะสมสำหรับการวิจัย (ศิริชัย กาญจนวาสี และคณะ, 2537) ในระดับความเชื่อมั่น $95\% \pm 5\%$ พบว่าควรใช้ขนาดกลุ่มตัวอย่าง 1,481 คน ซึ่งในการวิจัยครั้งนี้ผู้วิจัยใช้กลุ่มตัวอย่าง 1,609 คน แบ่งเป็นกลุ่มใหม่ จำนวน 1,084 คน กลุ่มอ้างอิง จำนวน 400 คน และกลุ่มอ้างอิงที่ใช้เป็นเกณฑ์ จำนวน 124 คน รวมเป็น 1,609 คน โดยผู้สอบกลุ่มอ้างอิงที่ใช้เป็นเกณฑ์ทำแบบสอบฉบับใหม่ (new) หลังจากนั้นผ่านไปประมาณ 3 สัปดาห์ ผู้สอบกลุ่มนี้ทำแบบสอบฉบับอ้างอิง (reference) ซึ่งแบบสอบทั้ง 2 ฉบับนี้เป็นแบบสอบคู่ขนานที่มีข้อสอบรวมรูปแบบผสม โดยผู้ตรวจ 1 ตรวจแบบสอบฉบับใหม่ (new) และผู้ตรวจ 2 ตรวจแบบสอบฉบับอ้างอิง (reference) ในการเชื่อมต่อคะแนนจากแบบสอบฉบับใหม่/ผู้ตรวจ 2 ไปยังคะแนนจากแบบสอบฉบับอ้างอิง/ผู้ตรวจ 1 ใช้วิธีการเปรียบเทียบอีควิเปอร์เซ็นไทล์แบบลูกโซ่ ที่ปรับการแจกแจงคะแนนด้วยล็อกเชิงเส้น (chained equipercentile equating with loglinear presmoothed) วิธีการสุ่มตัวอย่างที่ใช้เป็นแบบหลายขั้นตอน (multi - stage random sampling) โดยสุ่มโรงเรียนแบบแบ่งชั้น (stratified random sampling) โดยมีขนาดโรงเรียนเป็นตัวแปรสำหรับแบ่งชั้น คือ ขนาดใหญ่พิเศษ ขนาดใหญ่ ขนาดกลาง และขนาดเล็ก โดยใช้จำนวนนักเรียนเป็นเกณฑ์ในการแบ่งกลุ่มขนาดโรงเรียน ซึ่งจะได้จำนวนประชากรโรงเรียน และนักเรียนที่แบ่งตามขนาดของโรงเรียน จากนั้นผู้วิจัยสุ่มโรงเรียนอย่างง่าย (simple random sampling) ตามขนาดของโรงเรียน



ภาพ 1 การออกแบบการปรับเทียบคะแนนแบบสอบรูปแบบผสมโดยใช้ข้อสอบรวม
สำหรับกลุ่มไม่เท่าเทียมกัน (NEAT) ระหว่างวิธีการตรวจที่แตกต่างกัน

เครื่องมือที่ใช้ในการวิจัย

1. แบบสอบวิชาฟิสิกส์รูปแบบผสม ระดับมัธยมศึกษาชั้นปีที่ 4 ภาคการศึกษาต้น ปีการศึกษา 2556 เรื่อง "งานและพลังงาน" ที่ผู้วิจัยสร้างขึ้น 2 ฉบับ ซึ่งเป็นแบบสอบคู่ขนาน ได้แก่ แบบสอบฉบับใหม่ จำนวน 24 ข้อ ประกอบด้วยข้อสอบแบบหลายตัวเลือก 16 ข้อ คะแนนเต็มข้อละ 1 คะแนน ซึ่งมีค่าความเที่ยงเป็น .634 และข้อสอบแบบสร้างคำตอบ 8 ข้อ คะแนนเต็มข้อละ 4 คะแนน ซึ่งมีค่าความเที่ยงเป็น .824 ดังนั้นคะแนนเต็มของแบบสอบฉบับนี้จึงเท่ากับ 48 คะแนน และแบบสอบฉบับอ้างอิง จำนวน 24 ข้อ ประกอบด้วยข้อสอบแบบหลายตัวเลือก 16 ข้อ คะแนนเต็มข้อละ 1 คะแนน ซึ่งมีค่าความเที่ยงเป็น .602 และข้อสอบแบบสร้างคำตอบ 8 ข้อ คะแนนเต็มข้อละ 4 คะแนน ซึ่งมีค่าความเที่ยงเป็น .872 ดังนั้นคะแนนเต็มของแบบสอบฉบับนี้จึงเท่ากับ 48 คะแนน โดยแบบสอบทั้งสองฉบับมีความเป็นคู่ขนานกันทั้งด้านเนื้อหา และความยากของ

แบบสอบ ซึ่งมีข้อสอบร่วมกัน จำนวน 12 ข้อ ประกอบด้วยข้อสอบแบบหลายตัวเลือกจำนวน 8 ข้อ คะแนนเต็มข้อละ 1 คะแนน และข้อสอบแบบสร้างคำตอบจำนวน 4 ข้อคะแนนเต็มข้อละ 4 คะแนน คะแนนเต็มของข้อสอบรวมจึงเท่ากับ 24 คะแนน

2. เกณฑ์การให้คะแนน (rubric score) ที่ใช้ในตรวจข้อสอบแบบสร้างคำตอบในแบบสอบวิชาฟิสิกส์ระดับมัธยมศึกษาปีที่ 4 เรื่อง "งานและพลังงาน" โดยเกณฑ์การให้คะแนนดังกล่าวมีความเที่ยงสูง เนื่องจากคะแนนจากการตรวจให้คะแนนข้อสอบแบบสร้างคำตอบในแบบสอบเรื่องงานและพลังงานของผู้ตรวจคน 1 และผู้ตรวจ 2 มีความสัมพันธ์กันทางบวกในระดับสูง ($r = .96$) อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

3. ผู้ตรวจให้คะแนน 2 คน ได้แก่ ผู้ตรวจ 1 และผู้ตรวจ 2 โดยผู้ตรวจ 2 คือ ผู้วิจัย ส่วนผู้ตรวจ 1 คือ ครูผู้สอนวิชาฟิสิกส์ ที่มีประสบการณ์ในโรงเรียนมาแล้วอย่างน้อย 3 ปี โดยที่ผู้ตรวจคะแนนทั้งสองคนมีความสอดคล้องในการให้คะแนนระหว่างผู้ประเมินเป็นรายข้อสำหรับแบบสอบฉบับใหม่และแบบสอบฉบับอ้างอิงอยู่ในระดับสูงอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 โดยมีค่าสัมประสิทธิ์สหพันธ์แบบเพียร์สัน (r) ในแต่ละข้ออยู่ระหว่าง .86 - .99 และ .84 -1.00 ตามลำดับ

การดำเนินการทดลองและเก็บรวบรวมข้อมูล

วิธีการจัดการสอบ และการตรวจให้คะแนนตามการออกแบบการปรับเทียบคะแนน (equating designs) มีขั้นตอนการดำเนินการดังนี้

1. การทดสอบช่วงที่ 1 มีขั้นตอนในการจัดการสอบและการตรวจให้คะแนน คือ ผู้วิจัยดำเนินการจัดการสอบกับ 1) กลุ่มอ้างอิงที่ใช้เป็นเกณฑ์ โดยให้ผู้สอบกลุ่มดังกล่าวทำแบบสอบฉบับใหม่ 2) กลุ่มใหม่ โดยให้ผู้สอบกลุ่มดังกล่าวทำแบบสอบฉบับใหม่ หลังจากนั้นผู้ตรวจ 1 และ 2 (ผู้วิจัย) ทำการตรวจแบบสอบของกลุ่มอ้างอิงที่ใช้เป็นเกณฑ์ และกลุ่มใหม่ ตามวิธีการตรวจที่ได้กำหนดไว้ ดังภาพ 1

2. การทดสอบช่วงที่ 2 ซึ่งมีขั้นตอนในการจัดการสอบและการตรวจให้คะแนน คือ ผู้วิจัยดำเนินการจัดการสอบกับ 1) กลุ่มอ้างอิงที่ใช้เป็นเกณฑ์ โดยให้ผู้สอบกลุ่มดังกล่าวทำแบบสอบฉบับอ้างอิง ซึ่งทำการทดสอบหลังจากการทดสอบครั้งแรกประมาณ 3 สัปดาห์ 2) กลุ่มอ้างอิง โดยให้ผู้สอบกลุ่มดังกล่าวทำแบบสอบฉบับอ้างอิง หลังจากนั้นผู้ตรวจ 1 และ 2 (ผู้วิจัย) ทำการตรวจแบบสอบของกลุ่มอ้างอิงที่ใช้เป็นเกณฑ์ และกลุ่มอ้างอิง ตามวิธีการตรวจที่ได้กำหนดไว้ ดังภาพ 1

การวิเคราะห์ข้อมูล

ผู้วิจัยนำผลจากการสอบของผู้สอบกลุ่มใหม่ และกลุ่มอ้างอิงทั้ง 2 กลุ่มมาวิเคราะห์ข้อมูล ดังนี้

1. ค่าสถิติที่เกี่ยวข้องกับแบบสอบและผู้สอบทั้งสองกลุ่ม ภายใต้รูปแบบผู้สอบกลุ่มไม่เท่าเทียมกัน โดยใช้ข้อสอบรวม โดยสรุป ใช้โปรแกรมคอมพิวเตอร์

2. ค่าสถิติที่เกี่ยวข้องกับแบบสอบและผู้สอบทั้งสองกลุ่ม ภายใต้สำหรับการออกแบบการปรับเทียบที่ใช้เป็นเกณฑ์ โดยสรุป ใช้โปรแกรมคอมพิวเตอร์

3. การวิเคราะห์ความสอดคล้องในการให้คะแนนข้อสอบรวมแบบสร้างคำตอบระหว่างผู้ตรวจ A และ B ใช้โปรแกรมคอมพิวเตอร์

4. การปรับเทียบคะแนนแบบสอบรูปแบบผสมโดยใช้ข้อสอบรวมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT) โดยใช้วิธีการปรับเทียบเชิงเส้นตรง 2 วิธี ได้แก่ วิธีของเลวีน (Levine method) และวิธี CLE (Chained linear equating) ใช้โปรแกรม R (Albano, 2014)

5. การปรับเทียบคะแนนสำหรับการออกแบบการปรับเทียบที่ใช้เป็นเกณฑ์ โดยใช้วิธีการปรับเทียบอีควิเปอร์เซ็นไทล์แบบลูกโซ่ ที่ปรับการแจกแจงคะแนนด้วยล็อกเชิงเส้น (chained equipercentile equating with loglinear presmoothed) ใช้โปรแกรม R (Albano, 2014)

6. การเปรียบเทียบค่ารากที่สองของค่าเฉลี่ยความแตกต่างยกกำลังสอง (RMSD) ของวิธีการเปรียบเทียบเชิงเส้นตรง 2 วิธี ระหว่างวิธีของเลวิน (Levine method) และวิธี CLE (Chained linear equating) ใช้โปรแกรม Microsoft excel ตามสมการ $RMSD = \sqrt{\sum_{i=0}^{48} w_i [\hat{e}_i(x_i) - e_i(x_i)]^2}$

7. การเปรียบเทียบสถิติบูตสเตรป (bootstrapped statistics) ของวิธีการตรวจ 3 วิธี สำหรับการออกแบบการปรับเทียบคะแนนแบบสอบรูปแบบผสมโดยใช้ข้อสอบร่วมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT) ซึ่งพิจารณาจาก

7.1 ค่าเฉลี่ยน้ำหนักของรากที่สองของค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสอง (weighted average RMSE) หรือเรียกสั้น ๆ ว่าค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสอง (RMSE) โดยผู้วิจัยต้องคำนวณค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสองในแต่ละจุดคะแนน ($RMSE_i$) ก่อนด้วยโปรแกรม R (Albano, 2014) หลังจากนั้นนำค่า $RMSE_i$ ในแต่ละจุดคะแนนแทนในสมการ $\sqrt{\sum_i w_i RMSE_i^2}$ เพื่อหาค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสอง (RMSE) โดยใช้โปรแกรม Microsoft excel

7.2 ค่าเฉลี่ยน้ำหนักของค่าความลำเอียง (weighted average root mean square bias) หรือเรียกสั้น ๆ ว่าค่าความลำเอียง (bias) โดยผู้วิจัยต้องคำนวณค่าความลำเอียง (bias) ในแต่ละจุดคะแนน ($Bias_i$) ก่อนด้วยโปรแกรม R (Albano, 2014) หลังจากนั้นนำค่า $Bias_i$ ในแต่ละจุดคะแนนแทนในสมการ $\sqrt{\sum_i w_i Bias_i^2}$ เพื่อหาค่าความลำเอียง (bias) โดยใช้โปรแกรม Microsoft excel

7.3 ค่าเฉลี่ยน้ำหนักของความคลาดเคลื่อนในการปรับเทียบ (weighted average equating error) หรือเรียกสั้น ๆ ว่าค่าความคลาดเคลื่อนในการปรับเทียบ (equating error) โดยผู้วิจัยต้องคำนวณค่าความคลาดเคลื่อนในการปรับเทียบในแต่ละจุดคะแนน (SEE_i) ก่อนด้วยโปรแกรม R (Albano, 2014) หลังจากนั้นนำค่า SEE_i ในแต่ละจุดคะแนนแทนในสมการ $\sqrt{\sum_i w_i SEE_i^2}$ เพื่อหาค่าความคลาดเคลื่อนในการปรับเทียบ (equating error) โดยใช้โปรแกรม Microsoft excel

ผลการวิจัย

การเปรียบเทียบสถิติบูตสเตรปของการออกแบบการปรับเทียบสำหรับแบบสอบรูปแบบผสมระหว่างวิธีที่ต่างกัน สรุปผลการวิจัย ดังนี้

ผลการเปรียบเทียบค่ารากที่สองของค่าเฉลี่ยความแตกต่างยกกำลังสอง (RMSD) ของวิธีการเปรียบเทียบเชิงเส้นตรง 2 วิธี ระหว่างวิธีของเลวิน (Levine method) และวิธี CLE (Chained linear equating) ใช้โปรแกรม Microsoft excel

ค่ารากที่สองของค่าเฉลี่ยความแตกต่างยกกำลังสอง (RMSD) ของวิธี CLE (Chain Linear Equating) มีค่าต่ำกว่าวิธีของเลวิน (Levine method) ในทุกวิธีการตรวจของการออกแบบการปรับเทียบคะแนนแบบสอบรูปแบบผสมโดยใช้ข้อสอบร่วมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT) โดยค่ารากที่สองของค่าเฉลี่ยความแตกต่างยกกำลังสอง (RMSD) ของวิธี CLE (Chain Linear Equating) มีค่ามากที่สุดอยู่ที่วิธีการตรวจวิธีที่ 2: Full Trend และ 3: Half Trend คือ 1.481 และมีค่าต่ำที่สุดอยู่ที่วิธีการตรวจวิธีที่ 1: Non Trend คือ 1.421 ขณะที่ค่ารากที่สองของค่าเฉลี่ยความแตกต่างยกกำลังสอง (RMSD) ของวิธีของเลวิน (Levine method) มีค่ามากที่สุดอยู่ที่วิธีการตรวจวิธีที่ 2: Full Trend รองลงมา คือ วิธีการตรวจวิธีที่ 3: Half Trend และ 1: Non Trend ตามลำดับ โดยมีค่ารากที่สองของค่าเฉลี่ยความแตกต่างยกกำลังสอง

(RMSD) เท่ากับ 1.804 1.793 และ 1.717 ตามลำดับ ดังนั้นวิธีการปรับเทียบคะแนนที่เหมาะสมสำหรับการออกแบบการปรับเทียบคะแนนแบบสอบรูปร่างผสมโดยใช้ข้อสอบร่วมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT) คือ วิธี CLE (Chain Linear Equating) รายละเอียดดังตาราง 1

ตาราง 1 ผลการวิเคราะห์ค่ารากที่สองของค่าเฉลี่ยความแตกต่างยกกำลังสอง (RMSD) ระหว่างวิธีการปรับเทียบเชิงเส้นตรง 2 วิธี สำหรับการออกแบบการปรับเทียบคะแนนแบบสอบรูปร่างผสมโดยใช้ข้อสอบร่วมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT)

วิธีการตรวจ	วิธีการปรับเทียบคะแนน	
	วิธี CLE (Chained Linear Equating)	วิธีของเลวิน (Levine method)
วิธีที่ 1: Non Trend	1.421	1.717
วิธีที่ 2: Full Trend	1.481	1.804
วิธีที่ 3: Half Trend	1.481	1.793

ผลการเปรียบเทียบสถิติบูตสแตรป (bootstrapped statistics) ของวิธีการตรวจ 3 วิธีสำหรับการออกแบบการปรับเทียบคะแนนแบบสอบรูปร่างผสมโดยใช้ข้อสอบร่วมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT)

การปรับเทียบคะแนนด้วยวิธี CLE (Chained Linear Equating) สำหรับการออกแบบการปรับเทียบคะแนนแบบสอบรูปร่างผสมโดยใช้ข้อสอบร่วมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT) พบว่าวิธีการตรวจที่มีค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสอง (RMSE) ต่ำที่สุด คือ วิธีการตรวจวิธีที่ 3: Half Trend, 2: Full Trend และ 1: Non Trend ตามลำดับ ส่วนวิธีการตรวจที่มีค่าความลำเอียง (bias) ต่ำที่สุด คือ วิธีการตรวจวิธีที่ 3: Half Trend, 2: Full Trend และ 1: Non Trend ตามลำดับ ส่วนวิธีการตรวจที่มีค่าความคลาดเคลื่อนในการปรับเทียบ (equating error) ต่ำที่สุด ได้แก่ วิธีการตรวจวิธีที่ 3: Half Trend, 1: Non Trend และ 2: Full Trend ตามลำดับ

การปรับเทียบคะแนนด้วยวิธีของเลวิน (Levine method) สำหรับการออกแบบการปรับเทียบคะแนนแบบสอบรูปร่างผสมโดยใช้ข้อสอบร่วมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT) พบว่า วิธีการตรวจที่มีค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสอง (RMSE) ต่ำที่สุด คือ วิธีการตรวจวิธีที่ 1: Non Trend, 2: Full Trend และ 3: Half Trend ตามลำดับ ส่วนวิธีการตรวจวิธีที่มีค่าความลำเอียง (bias) ต่ำที่สุด คือ วิธีการตรวจวิธีที่ 2: Full Trend, 1: Non Trend และ 3: Half Trend ตามลำดับ ส่วนวิธีการตรวจที่มีค่าความคลาดเคลื่อนในการปรับเทียบ (equating error) ต่ำที่สุด ได้แก่ วิธีการตรวจวิธีที่ 3: Half Trend, 1: Non Trend และ 2: Full Trend ตามลำดับ แต่ถ้าเปรียบเทียบค่าสถิติบูตสแตรปทั้ง 3 ค่า ที่ได้จากการปรับเทียบเชิงเส้นตรงทั้ง 2 วิธี พบว่า ค่าสถิติบูตสแตรปทั้ง 3 ค่า ที่ได้จากการวิธีของเลวิน (Levine method) ในแต่ละวิธีการตรวจมีค่ามากกว่าวิธี CLE (Chained Linear Equating) รายละเอียดดังตาราง 2

ตาราง 2 ผลการวิเคราะห์สถิติบูทสแตรป (bootstrapped statistics) สำหรับการออกแบบการปรับเทียบคะแนนแบบสอบรูปแบบผสมโดยใช้ข้อสอบร่วมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT) ด้วยวิธีการปรับเทียบเชิงเส้นตรง 2 วิธี

วิธีการตรวจ	วิธี CLE (Chained Linear Equating)			วิธีของเลวิน (Levine method)		
	BIAS	EQUATING ERROR	RMSE	BIAS	EQUATING ERROR	RMSE
Non Trend	0.818	0.260	0.858	0.976	0.274	1.014
Full Trend	0.803	0.261	0.844	0.975	0.282	1.015
Half Trend	0.802	0.251	0.840	0.988	0.271	1.024

อภิปรายผลและข้อเสนอแนะ

การอภิปรายตามสมมติฐานการวิจัยข้อที่ 1 การออกแบบการปรับเทียบคะแนน (equating designs) ที่มีการตรวจให้คะแนนแตกต่างกัน

จากการวิจัยในครั้งนี้ พบว่า วิธีปรับเทียบคะแนนด้วยวิธี CLE (Chained Linear Equating) ให้ค่ารากที่สองของค่าเฉลี่ยความแตกต่างยกกำลังสอง (RMSE) ต่ำกว่าวิธีของเลวิน (Levine method) ด้วยเหตุนี้ผู้วิจัยอภิปรายการออกแบบการปรับเทียบด้วยวิธี CLE (Chained Linear Equating) เป็นหลักจากการเปรียบเทียบสถิติบูทสแตรปของการออกแบบการปรับเทียบคะแนนแบบสอบรูปแบบผสมโดยใช้ข้อสอบร่วมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT) พบว่า สอดคล้องกับสมมติฐานข้อที่ 1 นั่นคือ วิธีตรวจข้อสอบร่วมแบบสร้างคำตอบ โดยมีการตรวจให้คะแนนข้ามกลุ่มทั้ง 2 วิธีมีค่าสถิติบูทสแตรปส่วนใหญ่ต่ำกว่าวิธีตรวจข้อสอบร่วมแบบสร้างคำตอบ โดยไม่มีการตรวจให้คะแนนข้ามกลุ่ม ถ้าพิจารณาค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสอง (RMSE) ซึ่งเป็นค่าความคลาดเคลื่อนที่มาจากค่าความลำเอียง (bias) และค่าความคลาดเคลื่อนในการปรับเทียบ (equating error) พบว่า วิธีการตรวจวิธีที่ 3: Half Trend มีค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสอง (RMSE) ต่ำที่สุด ซึ่งเป็นผลมาจากค่าความลำเอียง (bias) และค่าความคลาดเคลื่อนในการปรับเทียบ (equating error) มีค่าต่ำ ขณะที่วิธีการตรวจวิธีที่ 2: Full Trend มีค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสอง (RMSE) ต่ำเป็นลำดับที่ 2 ซึ่งเป็นผลมาจากค่าความลำเอียง (bias) มีค่าต่ำเป็นลำดับที่ 2 ส่วนวิธีการตรวจวิธีที่ 1: Non Trend มีค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสอง (RMSE) มากที่สุด เนื่องจากมีค่าความลำเอียง (bias) มีค่ามากที่สุด แต่อย่างไรก็ตาม ค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสอง (RMSE) และค่าความลำเอียง (bias) ที่ได้จากการตรวจวิธีที่ 2: Full Trend และ 3: Half Trend มีค่าใกล้เคียงกัน จึงได้กล่าวว่า ในการออกแบบการปรับเทียบคะแนนแบบสอบรูปแบบผสมโดยใช้ข้อสอบร่วมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT) ด้วยวิธี CLE (Chained Linear Equating) ที่ใช้วิธีตรวจข้อสอบร่วมแบบสร้างคำตอบ โดยมีการตรวจให้คะแนนข้ามกลุ่ม ทำให้ประสิทธิภาพในการปรับเทียบคะแนนมากกว่าวิธีตรวจข้อสอบร่วมแบบสร้างคำตอบ โดยไม่มีการตรวจให้คะแนนข้ามกลุ่ม ซึ่งสอดคล้องกับงานวิจัยของ Kim *et al.* (2010a) ที่ได้ศึกษาเปรียบเทียบสถิติบูทสแตรปของการออกแบบการปรับเทียบคะแนนแบบสอบรูปแบบผสมโดยใช้ข้อสอบร่วมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT) ด้วยวิธี CLE (Chained Linear Equating) ซึ่งพบว่า การปรับเทียบคะแนนที่ใช้วิธีตรวจข้อสอบร่วมแบบสร้างคำตอบ โดยมีการตรวจให้คะแนนข้ามกลุ่ม มีค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสอง (RMSE) ต่ำที่สุด ขณะที่การปรับเทียบคะแนนที่ใช้วิธีตรวจข้อสอบร่วมแบบสร้างคำตอบ โดยไม่มีการตรวจให้คะแนนข้ามกลุ่ม ให้ค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสอง (RMSE) และค่าความลำเอียง (bias) มากที่สุด อีกทั้งยังสอดคล้องกับงานวิจัยของ Kim *et al.* (2010b) ได้ศึกษาสถิติบูทสแตรปของการออกแบบการปรับเทียบคะแนนแบบสอบแบบสร้างคำตอบโดยใช้ข้อสอบร่วมสำหรับกลุ่มไม่เท่าเทียมกัน

(NEAT) พบว่า การปรับเทียบคะแนนที่ใช้วิธีตรวจสอบข้อสอบรวมแบบสร้างคำตอบ โดยมีการตรวจให้คะแนนข้ามกลุ่ม มีระดับความคลาดเคลื่อนของการปรับเทียบคะแนนระหว่างแบบสอบต่ำ แสดงให้เห็นว่าความคลาดเคลื่อนในการปรับเทียบคะแนนและความลำเอียงในการปรับเทียบคะแนนสามารถควบคุมได้ด้วยการตรวจให้คะแนนแบบข้ามกลุ่มในข้อสอบรวมแบบสร้างคำตอบ (trend construct-response anchor) ซึ่งทำให้ค่าความลำเอียง (bias) และค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสอง (RMSE) ต่ำกว่าข้อสอบรวมแบบสร้างคำตอบที่ไม่มีการตรวจให้คะแนนแบบข้ามกลุ่ม (no-trend construct-response anchor) และเมื่อเปรียบเทียบระหว่างวิธีตรวจสอบข้อสอบรวมแบบสร้างคำตอบ โดยมีการตรวจให้คะแนนข้ามกลุ่ม ทั้ง 2 วิธี พบว่า วิธีการตรวจวิธีที่ 3: Half Trend มีค่าความลำเอียง (bias) และค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสอง (RMSE) ต่ำกว่าวิธีการตรวจวิธีที่ 2: Full Trend อาจเป็นเพราะการที่ผู้ตรวจให้คะแนนทั้ง 2 คนร่วมกันตรวจข้อสอบรวมแบบสร้างคำตอบของผู้สอบทั้ง 2 กลุ่ม โดยแบ่งกันตรวจคนละครึ่งหนึ่งของข้อสอบรวมเป็นการลดความลำเอียง และความเมื่อยล้าในการตรวจให้คะแนนเมื่อเทียบกับอีกวิธีหนึ่งที่ผู้ตรวจให้คะแนนเพียงคนเดียวต้องทำหน้าที่ในการตรวจให้คะแนนข้อสอบรวมสร้างคำตอบทั้ง 4 ข้อของผู้สอบทั้ง 2 กลุ่มจากแบบสอบจำนวน 1,609 ฉบับ ซึ่งสอดคล้องกับ Johnson และคณะ (2009) ที่กล่าวว่า ในการประเมินข้อสอบแบบอัตนัย (essay test) และทักษะปฏิบัติต่างๆ ควรมีผู้ประเมิน (rater) หลายคน เพื่อให้คะแนนที่ได้จากการประเมินมีความเที่ยงมากขึ้น ซึ่งหากมีผู้ประเมินคนเดียวและผู้ประเมินเป็นผู้ออกข้อสอบหรือกำหนดงานที่ใช้ในการประเมิน ผู้ประเมินอาจนึกถึงคำตอบที่ตัวเองเฉลยไว้แล้วละเลยคำตอบอื่นๆที่เป็นไปได้ ดังนั้นควรมีผู้ประเมินอีกหนึ่งคนเพื่อให้ได้มุมมองคำตอบอื่น ๆ ที่เป็นไปได้ และสามารถใช้อ้างอิงในการตรวจสอบความเที่ยงระหว่างผู้ประเมินได้ด้วย (Lenel, 1990 cited in Johnson *et al.*, 2009)

แต่อย่างไรก็ตาม เมื่อพิจารณาสถิติบุทสเตรปของการปรับเทียบคะแนนที่ได้จากวิธีของเลวิน (Levine method) มีความสอดคล้องกับวิธี CLE (Chained Linear Equating) กล่าวคือ วิธีการตรวจวิธีที่ 3: Half Trend มีค่าความคลาดเคลื่อนในการปรับเทียบ (equating error) ต่ำที่สุดในขณะเดียวกันก็มีความขัดแย้งกันระหว่างผลได้จากทั้ง 2 วิธี กล่าวคือ ในวิธีของเลวิน (Levine method) วิธีตรวจที่มีค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสอง (RMSE) ต่ำที่สุด คือ วิธีการตรวจวิธีที่ 1: Non Trend ส่วนรูปแบบที่มีค่าความลำเอียง (bias) ต่ำที่สุด คือ วิธีการตรวจวิธีที่ 2: Trend ขณะที่วิธี CLE (Chained Linear Equating) วิธีการตรวจที่มีค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสอง (RMSE) และค่าความลำเอียง (bias) ต่ำที่สุด คือ วิธีการตรวจวิธีที่ 3: Half Trend แสดงให้เห็นว่า นอกจากการข้อสอบรวมแบบสร้างคำตอบที่มีผลต่อค่าสถิติบุทสเตรปแล้ว วิธีการปรับเทียบที่ใช้อยู่มีผลต่อค่าสถิติดังกล่าวเช่นกัน ดังนั้นในการปรับเทียบคะแนนจึงต้องใช้วิธีการตรวจและวิธีการปรับเทียบคะแนนให้มีความเหมาะสม ซึ่งสำหรับบริบทของการวิจัยในครั้งนี้ วิธี CLE (Chained Linear Equating) มีความเหมาะสมในการปรับเทียบคะแนนมากกว่าวิธีของเลวิน (Levine method)

ด้วยเหตุนี้ ในการปรับเทียบคะแนนสำหรับการออกแบบการปรับเทียบคะแนนแบบสอบรูปแบบผสมโดยใช้ข้อสอบรวมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT) ด้วยวิธี CLE (Chained Linear Equating) ควรใช้วิธีตรวจสอบข้อสอบรวมแบบสร้างคำตอบ โดยมีการตรวจให้คะแนนข้ามกลุ่มครึ่งหนึ่งของข้อสอบรวม (ผู้ตรวจทั้ง 2 คน ตรวจข้อสอบรวมแบบสร้างคำตอบของผู้สอบทั้ง 2 กลุ่ม โดยแบ่งกันตรวจคนละครึ่งหนึ่งของข้อสอบรวม) หรือวิธีตรวจสอบข้อสอบรวมแบบสร้างคำตอบ โดยมีการตรวจให้คะแนนข้ามกลุ่มทั้งหมดของข้อสอบรวม (ผู้ตรวจคนหนึ่ง ตรวจข้อสอบรวมแบบสร้างคำตอบของผู้สอบทั้ง 2 กลุ่ม) เนื่องจากการตรวจให้คะแนนข้ามกลุ่ม (trend scoring method) ในข้อสอบรวมแบบสร้างคำตอบช่วยลดความลำเอียงในการปรับเทียบคะแนนได้ (Kim *et al.*, 2010a) และทำให้การปรับเทียบคะแนนในบริบทดังกล่าวมีประสิทธิภาพ

การอภิปรายตามสมมติฐานการวิจัยข้อที่ 2 วิธีการปรับเทียบเชิงเส้นตรง (equating methods)

จากการวิเคราะห์ค่ารากที่สองของค่าเฉลี่ยความแตกต่างยกกำลังสอง (RMSD) พบว่า ค่ารากที่สองของค่าเฉลี่ยความแตกต่างยกกำลังสอง (RMSD) ของวิธี CLE (Chained Linear Equating) มีค่าต่ำกว่าวิธีของเลวิน (Levine method) ในทุกวิธีการตรวจของการออกแบบการปรับเทียบคะแนนแบบสอบรูปแบบผสมโดยใช้ข้อสอบร่วมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT) ดังนั้นวิธีการปรับเทียบคะแนนที่เหมาะสมในการนำมาใช้ในการปรับเทียบแบบสอบรูปแบบผสม โดยใช้ข้อสอบร่วมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT) ที่มีการทดสอบจริงในชั้นเรียน คือ วิธี CLE (Chained Linear Equating) นอกจากนี้เมื่อพิจารณาค่าสถิติบูทสเตรปของวิธีการตรวจทั้ง 3 วิธี ของการออกแบบการปรับเทียบคะแนนแบบสอบรูปแบบผสมโดยใช้ข้อสอบร่วมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT) พบว่า ค่าความลำเอียง (bias) ค่าความคลาดเคลื่อนในการปรับเทียบ (equating error) และ ค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสอง (RMSE) ที่ได้จากวิธี CLE (Chained Linear Equating) มีค่าต่ำกว่าสถิติบูทสเตรปดังกล่าวที่ได้จากวิธีของเลวิน (Levine method) ซึ่งพบว่า ขัดแย้งกับสมมติฐานข้อที่ 2 ที่ผู้วิจัยตั้งไว้ว่า ค่ารากที่สองของค่าเฉลี่ยความแตกต่างยกกำลังสอง (RMSD) ของวิธีการปรับเทียบเชิงเส้นตรงด้วยวิธีของเลวิน (Levine method) มีค่าต่ำกว่าวิธี CLE (Chained Linear Equating) สำหรับการออกแบบการปรับเทียบคะแนนแบบสอบรูปแบบผสมโดยใช้ข้อสอบร่วมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT) ที่ใช้ข้อมูลจากการทดสอบจริง ซึ่งเมื่อพิจารณางานวิจัยของ Suh *et al.* (2009) ที่ได้เปรียบเทียบวิธีการปรับเทียบเชิงเส้นตรงภายใต้รูปแบบ NEAT ซึ่งพบว่า วิธีของเลวิน (Levine method) มีค่ารากที่สองของค่าเฉลี่ยความแตกต่างยกกำลังสอง (RMSD) ต่ำกว่าวิธี CLE (Chained Linear Equating) แต่อย่างไรก็ตามงานวิจัยดังกล่าวใช้กลุ่มตัวอย่างขนาดใหญ่ คือ 40,000 คน จากการทดสอบ MBE ที่ใช้แบบสอบแบบหลายตัวเลือก ในขณะที่งานวิจัยในครั้งนี้ใช้กลุ่มตัวอย่างขนาดเล็กเพียง 1,609 คน และแบบสอบที่ใช้เป็นแบบสอบรูปแบบผสม ซึ่งบริบทของการทดสอบดังกล่าวไม่ว่าจะเป็นจำนวนกลุ่มตัวอย่าง และรูปแบบของแบบสอบนั้นต่างจากงานวิจัยในครั้งนี้ จึงเป็นไปได้ว่าผลการปรับเทียบที่ได้อาจแตกต่างกัน นอกจากนี้งานวิจัยดังกล่าวยังได้อภิปรายเพิ่มเติมว่า ถึงแม้ว่าค่ารากที่สองของค่าเฉลี่ยความแตกต่างยกกำลังสอง (RMSD) ของวิธีของเลวิน (Levine method) มีค่าต่ำกว่าวิธีอื่น ๆ แต่วิธี CLE (Chained Linear Equating) ก็สามารถนำมาใช้ได้เมื่อมีกลุ่มตัวอย่างมีความแตกต่างกัน หรือเมื่อคะแนนของข้อสอบร่วมทั้งหมดกับข้อสอบทั้งหมดมีความสัมพันธ์กันสูง และเมื่อพิจารณาค่าสัมประสิทธิ์สหสัมพันธ์ของคะแนนสอบทั้งฉบับและคะแนนของข้อสอบร่วมสำหรับวิธีการตรวจทั้ง 3 วิธี พบว่า คะแนนสอบทั้งฉบับและคะแนนของข้อสอบร่วมมีความสัมพันธ์กันสูง โดยค่า r มีค่าอยู่ระหว่าง .91 - .94 จึงเป็นไปได้ว่าวิธี CLE (Chained Linear Equating) มีความเหมาะสมในการนำมาใช้กับบริบทของงานวิจัยนี้ที่ใช้กับแบบสอบรูปแบบผสม และกลุ่มตัวอย่างมีขนาดเล็ก อีกทั้งในงานวิจัยของ Suh *et al.* (2009) ไม่ได้ใช้การตรวจข้อสอบแบบข้ามกลุ่ม เช่นเดียวกับงานวิจัยของ Mroch *et al.* (2009) ที่ได้ทำการประเมินวิธีการปรับเทียบเชิงเส้นตรงด้วยการวิเคราะห์จากเอกสารและงานวิจัยที่เกี่ยวข้อง ซึ่งไม่ได้มีการเก็บข้อมูลจริง งานวิจัยดังกล่าวพบว่า วิธีของเลวิน (Levine method) มีค่ารากที่สองของค่าเฉลี่ยความแตกต่างยกกำลังสอง (RMSD) ต่ำกว่าวิธี CLE (Chained Linear Equating) และวิธีของ Tucker (Tucker method) แต่อย่างไรก็ตามเมื่อพิจารณางานวิจัยของ Puhan (2010) ที่ได้ศึกษาผลลัพธ์ของวิธีการปรับเทียบเชิงเส้นตรง 3 วิธีกับกลุ่มตัวอย่างขนาดใหญ่ โดยใช้แบบสอบแบบหลายตัวเลือกภายใต้กลุ่มตัวอย่างที่มีความสามารถแตกต่างกัน โดยใช้เกณฑ์ในการเปรียบเทียบ คือ ค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสอง (RMSE) ค่าความลำเอียง (bias) และค่าความคลาดเคลื่อนในการปรับเทียบ (equating error) ซึ่งผลการวิจัยพบว่า ในกรณีที่กลุ่มตัวอย่างมีความสามารถแตกต่างกันอย่างมาก วิธีของเลวิน (Levine method) ให้การปรับเทียบที่ดีกว่าเนื่องจากมีค่าความลำเอียง (bias) และค่ารากที่

สองของค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสอง (RMSE) ต่ำ ในกรณีที่กลุ่มตัวอย่างมีความสามารถแตกต่างกัน ระดับปานกลางไปจนถึงแตกต่างกันมาก วิธีของเลวิน (Levine method) ให้ผลการเปรียบเทียบที่ดี แต่ถ้า ในกรณีที่กลุ่มตัวอย่างมีความสามารถแตกต่างกันไม่มากนัก วิธี CLE (Chained Linear Equating) มีค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสอง (RMSE) ค่าความลำเอียง (bias) และค่าความคลาดเคลื่อนในการเปรียบเทียบ (equating error) ต่ำกว่าทั้งสองวิธีดังกล่าว นอกจากนี้ Puhan ยังได้อธิบายเพิ่มเติมว่า ถ้าผู้สอบสองกลุ่มมีความสามารถแตกต่างกัน โดยที่คะแนนจากแบบสอบทั้งฉบับและคะแนนจากข้อสอบรวมทั้งหมดมีความสัมพันธ์กันสูง วิธี CLE (Chained Linear Equating) และ วิธีของเลวิน (Levine method) ต่างเป็นวิธีการที่เหมาะสมที่ในการนำมาใช้ในการเปรียบเทียบ และเมื่อพิจารณาวิจัยของ Puhan (2012) ที่ได้ประเมินวิธีการเปรียบเทียบคะแนน 2 วิธี ระหว่างวิธีของ Tucker (Tucker method) และ วิธี CLE (Chained Linear Equating) ภายใต้สถานการณ์ที่สมมติขึ้น คือ เมื่อการทดสอบมีผู้ตรวจหลายคน และคะแนนของแบบสอบทั้งฉบับกับคะแนนจากข้อสอบรวมทั้งหมดมีความสัมพันธ์กันสูงมาก ผลการวิจัยพบว่า วิธีของ Tucker (Tucker method) มีความแม่นยำในการเปรียบเทียบคะแนนต่ำกว่าวิธี CLE (Chained Linear Equating)

จากงานวิจัยที่ผ่านมาที่ได้ทำการศึกษาถึงประสิทธิภาพของวิธีการเปรียบเทียบคะแนนเชิงเส้นตรง พบว่า เมื่อกลุ่มตัวอย่างมีความสามารถแตกต่างกัน และคะแนนของแบบสอบทั้งฉบับกับคะแนนของข้อสอบรวมทั้งหมดมีความสัมพันธ์กันสูง วิธี CLE (Chained Linear Equating) และวิธีของเลวิน (Levine method) ต่างมีความเหมาะสมในการนำมาใช้ในการ เปรียบเทียบคะแนน แต่อย่างไรก็ตามงานวิจัยที่ผ่านมาส่วนใหญ่ทำการศึกษาเกี่ยวกับกลุ่มตัวอย่างขนาดใหญ่ อีกทั้งยังใช้แบบสอบแบบหลายตัวเลือก และในงานวิจัยก็ทำการศึกษาจากการจำลองข้อมูล สถานการณ์สมมติ และจากวิเคราะห์เอกสารและงานวิจัยที่เกี่ยวข้อง แต่ในการวิจัยครั้งนี้เป็นการศึกษากับกลุ่มตัวอย่างขนาดเล็ก ในสภาพการทดสอบจริงที่ใช้แบบสอบรูปแบบผสมที่มีข้อสอบรวมรูปแบบผสมรวมกับการตรวจให้คะแนนแบบข้ามกลุ่มในข้อสอบรวมแบบสร้างคำตอบ ซึ่งยังไม่มีงานวิจัยใดที่ทำการเปรียบเทียบวิธีการเปรียบเทียบเชิงเส้นตรงทั้ง 2 วิธี ในบริบทดังกล่าว และเมื่อพิจารณาผลของการวิจัยในครั้งนี้ที่พบว่าวิธี CLE (Chain Linear Equating) ให้ค่ารากที่สองของค่าเฉลี่ยความแตกต่างยกกำลังสอง (RMSD) ต่ำกว่าวิธีของเลวิน (Levine method) ซึ่งมีความเป็นไปได้ที่อาจเกิดขึ้นเนื่องจากงานวิจัยที่ผ่านมาพบว่าวิธี CLE (Chain linear equating) มีความเหมาะสมในการนำมาใช้กับกลุ่มตัวอย่างที่มีความสามารถแตกต่างกัน โดยที่คะแนนของแบบสอบทั้งฉบับและคะแนนของข้อสอบรวมทั้งหมดมีความสัมพันธ์กันสูง ซึ่งเมื่อพิจารณาคะแนนของกลุ่มใหม่และกลุ่มอ้างอิงที่ได้จากงานวิจัยนี้ในแต่ละวิธีการตรวจ พบว่า คะแนนเฉลี่ยของผู้สอบทั้งสองกลุ่มมีความแตกต่างกัน และค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างคะแนนของแบบสอบทั้งฉบับและคะแนนของข้อสอบรวมทั้งหมดก็มีค่าสูง ดังนั้นจึงสามารถสรุปได้ว่าวิธี CLE (Chained Linear Equating) มีความเหมาะสมมากกว่าวิธีของเลวิน (Levine method) ในการนำมาใช้เปรียบเทียบคะแนนภายใต้การออกแบบการเปรียบเทียบคะแนนแบบสอบรูปแบบผสมโดยใช้ข้อสอบรวมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT) โดยกลุ่มตัวอย่างมีขนาดเล็กและมีความสามารถแตกต่างกัน เนื่องจากวิธี CLE (Chained Linear Equating) มีค่ารากที่สองของค่าเฉลี่ยความแตกต่างยกกำลังสอง (RMSD) ต่ำกว่าวิธีของเลวิน (Levine method) อีกทั้งค่าสถิติบูทสเตรปของวิธีการตรวจทั้ง 3 วิธี ได้แก่ ค่าความลำเอียง (bias) ค่าความคลาดเคลื่อนในการเปรียบเทียบ (equating error) และค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสอง (RMSE) ที่ได้จากวิธี CLE (Chained Linear Equating) มีค่าต่ำกว่าค่าสถิติบูทสเตรปดังกล่าวที่ได้จากวิธีของเลวิน (Levine method) อีกด้วย

ข้อเสนอแนะสำหรับการนำไปปฏิบัติ

จากผลการวิจัยดังกล่าว ผู้วิจัยมีข้อเสนอแนะดังนี้

1. ในการแปลความหมายคะแนนที่ปรับเทียบแล้ว ควรระวังในการสื่อความหมายหรือการนำไปใช้อ้างอิง เนื่องจากความหมายของคะแนนเหล่านี้ไม่ได้แปลความหมายเหมือนจากการสอบ เพราะคะแนนที่ได้เป็นค่าความสัมพันธ์เมื่อเปรียบเทียบกับผู้สอบอีกกลุ่มหนึ่ง ที่มีการสอบคนละช่วงเวลากัน แต่มีการใช้ข้อสอบร่วมกัน

2. ผลการวิจัยในครั้งนี้ พบว่า ค่าสถิติบูทสเตรป (bootstrapped statistics) ของวิธีตรวจสอบข้อสอบร่วมแบบสร้างคำตอบ โดยมีการตรวจให้คะแนนข้ามกลุ่มครั้งหนึ่งของข้อสอบร่วมมีค่าต่ำที่สุดในการแปลงคะแนนจากผู้สอบกลุ่มใหม่ไปยังกลุ่มอ้างอิง สำหรับการออกแบบการปรับเทียบคะแนนแบบสอบรูปแบบผสมโดยใช้ข้อสอบร่วมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT) ดังนั้นในการออกแบบการปรับเทียบคะแนนแบบสอบรูปแบบผสมโดยใช้ข้อสอบร่วมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT) ควรให้ผู้ตรวจทั้งสองคนทำหน้าที่ตรวจข้อสอบร่วมแบบสร้างคำตอบของผู้สอบทั้งสองกลุ่ม โดยแบ่งกันตรวจคนละครั้งหนึ่ง และถ้าต้องการให้การปรับเทียบคะแนนในรูปแบบดังกล่าวมีค่ารากที่สองของค่าเฉลี่ยความแตกต่างยกกำลังสอง (RMSD) ต่ำ ๆ ควรเลือกใช้วิธี CLE (Chained Linear Equating) ในการแปลงคะแนนของผู้สอบให้อยู่ในสเกลเดียวกันสำหรับบริบทที่กลุ่มตัวอย่างมีขนาดเล็ก ภายใต้การออกแบบการปรับเทียบคะแนนแบบสอบรูปแบบผสมโดยใช้ข้อสอบร่วมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT) ที่ทำการทดสอบจริงในชั้นเรียน

3. ผลการวิจัยครั้งนี้ พบว่า คะแนนเฉลี่ยของผู้สอบทั้ง 2 กลุ่มจากแบบสอบเรื่องงานและพลังงาน มีคะแนนเฉลี่ยต่ำกว่าครั้งหนึ่ง และผู้สอบจำนวนไม่น้อยที่ไม่ทำข้อสอบแบบสร้างคำตอบ โดยเลือกทำเฉพาะข้อสอบแบบปรนัยเท่านั้น ดังนั้นครูควรมีการจัดการเรียนการสอน หรือการทดสอบในห้องเรียนที่ส่งเสริมให้นักเรียนได้แสดงวิธีทำ หรือเขียนคำตอบของตนเองมากยิ่งขึ้น เนื่องจากสารสนเทศที่ได้จากการเขียนตอบของนักเรียนจะทำให้เห็นถึงมโนทัศน์ของนักเรียนที่ไม่ถูกต้องเกี่ยวกับเรื่องนั้นๆ ซึ่งครูสามารถนำสารสนเทศดังกล่าวไปใช้ในการปรับปรุงการจัดการเรียนการสอนให้ประสิทธิภาพมากยิ่งขึ้น

ข้อเสนอแนะสำหรับการวิจัยในครั้งต่อไป

1. ควรมีการเปรียบเทียบสถิติบูทสเตรปในการออกแบบการปรับเทียบทั้ง 3 รูปแบบในรายวิชาอื่นๆ อย่างเช่น วิชาภาษาไทย วิชาชีววิทยา และวิชาภาษาอังกฤษ เป็นต้น เพื่อที่จะได้ทราบว่าการออกแบบการปรับเทียบแบบใดที่เหมาะสมที่จะนำไปใช้ในรายวิชาต่างๆ อย่างเช่น วิชาภาษาไทย วิชาชีววิทยา และวิชาภาษาอังกฤษ เป็นต้น เพื่อที่จะได้ทราบว่าวิธีการตรวจวิธีใดที่เหมาะสมในการนำไปใช้ในรายวิชาต่าง ๆ

2. ควรมีการออกแบบวิธีการตรวจ และวิธีการปรับเทียบคะแนน สำหรับการออกแบบการปรับเทียบคะแนนแบบสอบรูปแบบผสมโดยใช้ข้อสอบร่วมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT) ในแบบสอบที่มีลักษณะหลายมิติ (multidimensional) เนื่องจากในปัจจุบันมีการใช้ข้อสอบที่มีลักษณะเนื้อหาหลายมิติมากขึ้น หากมีความต้องการปรับเทียบคะแนนในแบบสอบประเภทดังกล่าว จะได้มีวิธีการตรวจและวิธีการปรับเทียบคะแนนที่มีความเหมาะสมและมีประสิทธิภาพ

3. ควรมีการศึกษาสัดส่วน/จำนวนของข้อสอบร่วมรูปแบบผสมที่ใช้ในการออกแบบการปรับเทียบคะแนนแบบสอบรูปแบบผสมโดยใช้ข้อสอบร่วมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT) เมื่อใช้วิธีตรวจสอบข้อสอบร่วมแบบสร้างคำตอบ โดยมีการตรวจให้คะแนนข้ามกลุ่ม ซึ่งจากการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องลักษณะและสัดส่วนข้อสอบร่วม พบว่า ข้อสอบร่วมควรสร้างให้มีรายละเอียดคล้ายคลึงกับแบบสอบทั้งฉบับ โดยข้อสอบร่วมควรมีความยาวพอที่จะเป็นตัวแทนเนื้อหาในแบบสอบ ซึ่ง Hambleton *et al.* (1991) แนะนำว่า ข้อสอบร่วมควรมีจำนวนไม่น้อยกว่า 20 % - 25% ของข้อสอบทั้งฉบับ ขณะที่ Angoff *et*

al. (1971) แนะนำว่า ถ้าแบบสอบมีจำนวนข้อสอบทั้งหมด 40 ข้อ ควรใช้ข้อสอบรวมอย่างต่ำ 20% และในกรณีที่แบบสอบมีข้อสอบจำนวนมาก จำนวนข้อสอบรวมที่เพียงพอ คือ 30% ซึ่งการวิจัยในครั้งนี้ได้สร้างข้อสอบรวมให้กระจายตามจุดประสงค์เชิงพฤติกรรม แต่อย่างไรก็ตามแบบสอบทั้งสองฉบับมีจำนวนข้อสอบเฉพาะและจำนวนข้อสอบรวมเท่ากันตามงานวิจัยของ Kim et al. (2010) ซึ่งอาจทำให้ผู้สอบใช้เวลาในการทำข้อสอบรวมมากเกินไป ดังนั้นในงานวิจัยครั้งต่อไปอาจศึกษาสัดส่วนของข้อสอบรวมสำหรับบริบทดังกล่าวในสัดส่วนต่างๆ อย่างเช่น 20% 30% 40% และ 50% สำหรับการออกแบบการปรับเทียบคะแนนแบบสอบรูปแบบผสมโดยใช้ข้อสอบรวมสำหรับกลุ่มไม่เท่าเทียมกัน (NEAT) เมื่อใช้วิธีตรวจข้อสอบรวมแบบสร้างคำตอบโดยมีการตรวจให้คะแนนข้ามกลุ่ม

รายการอ้างอิง

ภาษาไทย

ศิริชัย กาญจนวาสี. (2555). *ทฤษฎีการทดสอบแนวใหม่*. (พิมพ์ครั้งที่ 2). กรุงเทพมหานคร: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

ภาษาอังกฤษ

Albabo, A. D. (2014). *An R Package for observed-score linking and equating*. Retrieved March 15, 2014. from <http://cran.r-project.org/web/packages/equate/index.html>.

Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance : Designing, scoring, and validating performance tasks*. New York: The Guilford Press.

Kim, S., Walker, M. E., & McHale, F. (2010a). Comparisons among designs for equating for equating mixed-format tests in large-scale assessment. *Journal of Educational Measurement*, 47(1), 36-53.

Kim, S., E.Walker, M. E., & McHale, F. (2010b). Investigating the effectiveness of equating designs for constructed-response tests in large-scale assessments. *Journal of Educational Measurement*, 47(2), 186-201.

Mroch, A. A., Suh, Y., Kane, M. T., & Ripkey, D. R. (2009). An evaluation of five linear equating methods for the NEAT design. *Measurement: Interdisciplinary Research & Perspective*, 7(3-4), 174-193.

Puhan, G. (2010). A comparison of chained linear and poststratification linear equating under different testing conditions. *Journal of Educational Measurement*, 47(1), 54-75.

Suh, Y., Mroch, A. A., Kane, M. T., & Ripkey, D. R. (2009). An Empirical Comparison of Five Linear Equating Methods for the NEAT Design. *Measurement: Interdisciplinary Research & Perspective*, 7(3-4), 147-173. doi: 10.1080/15366360903418048.