การเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบรูปแบบผสมโดยวิธี LRT วิธี LOR Z และ วิธี HGLM

A Comparison of Differential Item Functioning for Mixed-Format Tests Using LRT, LOR Z and HGLM Methods

เสาวลักษณ์ บุญจันทร์[1] ไพรัตน์ วงษ์นาม[2] อาวีพร ปานทอง[3]

Saowaluk Boonjun[1], Pairatana Wongnam[2] and Aweeporn Panthong[3]

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์ เพื่อเปรียบเทียบผลการประมาณค่าพารามิเตอร์ข้อสอบ ในแบบทดสอบรูปแบบผสม ระหว่างวิธี LRT วิธี LOR Z และ วิธี HGLM และเพื่อเปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ (DIF) ในแบบทดสอบรูปแบบผสม จำแนกตามเพศ ระหว่างวิธี LRT วิธี LOR Z และ วิธี HGLM ประชากรในการวิจัยครั้งนี้ เป็นนักเรียนของประเทศไทยที่มีช่วงอายุ 15 ปี 3 เดือน จนถึง 16 ปี 2 เดือน จำนวน 8,249 คน ที่เข้าร่วมสอบข้อสอบ PISA กลุ่มตัวอย่างได้มาจากการสุ่มอย่างง่าย (Simple Random Sampling) โดยใช้ชุดของแบบทดสอบเป็นหน่วยการสุ่ม โดยทำการสุ่มนักเรียนที่ทำแบบทดสอบวิชาละ 1 ฉบับ ได้กลุ่มตัวอย่างทั้งหมด 644 คน โดยใช้คะแนนที่ได้จากแบบทดสอบประเมินผลนักเรียนระดับนานาชาติ PISA 2015 วิเคราะห์ข้อมูลด้วยโปรแกรม IRT PRO โปรแกรม DIFAS และโปรแกรม HLM

ผลการวิจัยพบว่า

1. ผลการประมาณค่าพารามิเตอร์ความยากของข้อสอบทั้ง 3 รายวิชา ได้แก่ การรู้เรื่องวิทยาศาสตร์ (Scientific Literacy) การรู้เรื่องคณิตศาสตร์ (Mathematical Literacy) และการรู้เรื่องการอ่าน (Reading Literacy) ด้วยวิธี LRT และวิธี HGLM มีความสัมพันธ์กันทางบวก ส่วนวิธี LOR Z สัมพันธ์ทางลบกับวิธี LRT และวิธี HGLM อย่างมีนัยสำคัญทางสถิติที่ระดับ .01

2. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ทั้ง 3 รายวิชา ได้แก่ การรู้เรื่องวิทยาศาสตร์ (Scientific Literacy) การรู้เรื่องคณิตศาสตร์ (Mathematical Literacy) และการรู้เรื่องการอ่าน (Reading Literacy) ระหว่างวิธี LRT วิธี LOR Z และวิธี HGLM จำแนกตามเพศ พบว่า วิธี LOR Z ตรวจพบการทำหน้าที่ต่างกันของข้อสอบมากที่สุด จำนวน 15 ข้อ รองลงมาวิธี HGLM ตรวจพบการทำหน้าที่ต่างกันของข้อสอบ จำนวน 5 ข้อ และวิธี LRT ไม่พบการทำหน้าที่ต่างกันของข้อสอบ

คำสำคัญ : การทำหน้าที่ต่างกันของข้อสอบ แบบทดสอบรูปแบบผสม

[1]นักศึกษาระดับปริญญาเอก สาขาวิชาวิจัย วัดผล และสถิติการศึกษา มหาวิทยาลัยบูรพา, Doctor of Philosophy Program in Educational Research, Measurement and Statistics, Burapha University

[2]รองศาสตราจารย์ ดร. สาขาวิชาวิจัย วัดผล และสถิติการศึกษา มหาวิทยาลัยบูรพา, Assoc. Prof. Dr. of Program Philosophy Program in Educational Research, Measurement and Statistics, Burapha University

[3]อาจารย์ ดร. สาขาวิชาคณิตศาสตร์และสถิติ มหาวิทยาลัยราชภัฏนครสวรรค์, Doctor of Program in Mathematics and Statistics, Nakhon Sawan Rajabhat University

*ผู้ติดต่อ, อีเมล์: เสาวลักษณ์ บุญจันทร์, bjsaowaluk@gmail.com

รับเมื่อ 22 มีนาคม 2562 แก้ไข 28 เมษายน 2562 ตอบรับเมื่อ 29 เมษายน 2562

ABSTRACT

The purposes of this research were to: 1) compare the results of parameter estimation of ability parameters for mixed–format tests using LRT, LOR Z and HGLM methods, and 2) compare the result analysis of differential item functioning for mixed–format tests using LRT, LOR Z and HGLM methods, classified by gender. The samples were 644 Thai students drawn from the population of 8,249 students, aged 15 years 3 months to 16 years 2 months, who completed the 2015 Program for International Student Assessment. The samples were obtained through simple random sampling, using a test as a random unit, by selected the students who taking the tests for one subject. The data analysis was done through IRT PRO, DIFAS, and HLM programs.

The findings were as follows:

1. The analysis results of compared estimation of the difficulty parameters of all three subjects concerning Scientific Literacy, Mathematical Literacy and Reading Literacy through the use of LRT and HGLM methods showed positive relationships, where as the LOR Z method had a negative relationship with LRT and HGLM methods at a statiscal significant level of 0.01.

2. The results of the examination of the differential item functioning (DIF) in all three subjects: Scientific Literacy, Mathematical Literacy and Reading Literacy using the LRT, LOR Z and HGLM methods, classified by gender, revealed the LOR Z method found DIF in 15 items, followed by the HGLM method, with five items, whereas the LRT method did not find DIF.

Keywords : Differential Item Functioning, Mixed–Format Tests

ภูมิหลัง

มาตรฐานการศึกษาของชาติ ตามพระราชบัญญัติการศึกษาแห่งชาติ พุทธศักราช 2542 แก้ไขเพิ่มเติม (ฉบับที่ 2) พุทธศักราช 2545 มีอุดมการณ์และหลักการจัดการศึกษาเพื่อพัฒนาสังคมไทยให้เป็นสังคมแห่งการเรียนรู้ให้คนไทยได้รับโอกาสเท่าเทียมกันทางการศึกษา และพัฒนาคนได้อย่างต่อเนื่องตลอดชีวิต หน่วยงานที่มีหน้าที่โดยตรงคือสถานศึกษาที่จะต้องมีการจัดกระบวนการเรียนรู้ เพิ่มพูนความรู้ทักษะให้นักเรียนเป็นบุคคลที่สามารถเรียนรู้ และสามารถพัฒนาตนเองได้ ในหมวด 6 ว่าด้วยมาตรฐานและการประกันคุณภาพการศึกษา กำหนดให้หน่วยงานต้นสังกัดและสถานศึกษาจัดทำรายงานเกี่ยวกับคุณภาพของสถานศึกษา เพื่อนำไปสู่การพัฒนาคุณภาพ (มาตรา 48) ซึ่งมุมมองของคุณภาพการศึกษามีแนวคิดที่กว้างไกลขึ้นในทุกกลุ่มที่มีความเกี่ยวข้องกับการศึกษา ทำให้เพิ่มความกดดันในการปฏิบัติที่จะทำให้เกิดประสิทธิผลของระบบการศึกษามากขึ้น และมีความต้องการประเมินคุณภาพการจัดการศึกษาให้มีความถูกต้องในการจัดการศึกษาทุกระดับ (กระทรวงศึกษาธิการ, 2545)

จากการทดสอบเพื่อประเมินผลสัมฤทธิ์ทางการเรียนที่มีความสำคัญทั้งระดับผู้เรียน ระดับสถานศึกษา ระดับเขตพื้นที่การศึกษา และระดับชาติ การพัฒนาแบบทดสอบจึงต้องคำนึงถึงคุณภาพรายข้อและแบบทดสอบทั้งฉบับอย่างรอบด้าน โดยเฉพาะประเด็นด้านความตรง (Validity) ซึ่งถือเป็นหัวใจสำคัญ เพราะว่าความตรงเป็นคุณสมบัติของแบบสอบที่แสดงถึงความสามารถในการวัดได้ถูกต้อง แม่นยำ ถ้าผลการวัดได้ค่าที่ใกล้เคียงกับคุณลักษณะที่แท้จริงเพียงใด ก็ถือว่าการวัดมีความตรงมากขึ้นเพียงนั้น ส่วนการทำหน้าที่ต่างกันของข้อสอบและแบบสอบก็เป็นลักษณะหนึ่ง ของการตรวจสอบคุณภาพด้านความตรง โดยเป็นการตรวจสอบในประเด็นของความยุติธรรมของข้อสอบและแบบทดสอบ (Item and test unfairness) ซึ่งจะเกิดขึ้นในกรณีที่ผู้สอบกลุ่มย่อยต่างกันและมีลักษณะเฉพาะบางอย่างแตกต่างกัน มีความได้เปรียบหรือเสียเปรียบทั้งที่มีความสามารถจริงเท่ากัน แต่เดิมใช้คำว่า ความลำเอียง

ของข้อสอบ (Item bias) หรือความลำเอียงของแบบสอบ (Test bias) ซึ่งต่อมามีการเปลี่ยนมาใช้คำที่เหมาะสมและเป็นวิชาการมากกว่า เป็นคำว่าการทำหน้าที่ต่างกันของข้อสอบ (Differential item functioning : DIF) (ศิริชัย กาญจนวาสี, 2550, หน้า 115) การศึกษาเกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ มีนักวัดผลหลายท่านได้ ศึกษาทั้งในสภาพของข้อมูลจริง หรือการจำลองข้อมูลขึ้นมา ศึกษากับแบบทดสอบที่สร้างขึ้นเอง หรือแบบทดสอบมาตรฐานที่สร้างโดยหน่วยงานอื่น ๆ ของประเทศ โดยมีการศึกษาตัวแปรต้นและตัวแปรตาม ได้แก่ วิธีการตรวจสอบ กลุ่มเปรียบเทียบที่ศึกษา เช่น เพศ ภาษา ศาสนา เชื้อชาติ สังคม เป็นต้น ขนาดของกลุ่มตัวอย่าง ความยาวของแบบทดสอบ ลักษณะของแบบทดสอบ เป็นต้น ในส่วนตัวแปรตามที่ศึกษา ได้แก่ ผลการตรวจสอบการทำหน้าที่ต่างกันหรือความลำเอียงของข้อสอบ ความสอดคล้องของผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธีการตรวจสอบ เป็นต้น Thissen, Steinberg, and wainer (1988 cited in de Ayala, 2009) ได้นำเสนอวิธีสำหรับใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยใช้อัตราส่วนความควรจะเป็น (LRT) โดยการทดสอบความแตกต่างระหว่างเพศ ภูมิลำเนาเชื้อชาติ ศาสนา ซึ่งวิธีดังกล่าวใช้ทดสอบอัตราส่วนความควรจะเป็นเพื่อทดสอบนัยสำคัญของการทำหน้าที่ต่างกันของข้อสอบ โดยใช้โปรแกรม IRTLRDIF (Thissen, 2011) และ Bolt (2002) ได้ทำการเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี LRT วิธี Poly-SIBTEST และวิธี DFIT พบว่า LRT มีการควบคุมความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบดีกว่าวิธี Poly-SIBTEST และวิธี DFIT สุพัฒนา หอมบุปผา (2556) ได้ทำการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN ผลการศึกษาพบว่า วิธี HGLM มีความไวในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ดีกว่าวิธี MIMIC และวิธี BAYESIAN (Jose Luis Padilla, 2012) ได้นำเสนอการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี LOR Z โดยใช้โปรแกรม DIFAS and Randall (D. Penfield, 2013) ได้ทำการศึกษาตรวจสอบการทำหน้าที่ต่างกันของข้อสอบตามตัวแปรที่ศึกษา จำนวน 3 ตัวแปร คือ เพศ ประเภทสถานศึกษา และที่ตั้งภูมิศาสตร์ของสถานศึกษา โดยใช้โปรแกรม DIFAS 5.0 (Penfield, 2005) สถิติที่ใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (Differential item functioning: DIF) ด้วยวิธีการวิเคราะห์แมนเทล-แฮนส์เซลร่วมกับอัตราส่วนแต้มต่อมาตรฐาน (Standardized mantel-haenszel log-odds ratio: LOR Z) พบว่าเป็นวิธีที่มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

การทำหน้าที่ต่างกันของข้อสอบที่ตรวจให้คะแนนทั้งแบบทวิภาคและพหุวิภาค หรือที่เรียกว่า แบบทดสอบรูปแบบผสม (Mixed format tests) ปัจจุบันมีการศึกษาน้อยมาก พบว่า มีการศึกษาการให้คะแนนแบบ 2 ค่า (Dichotomous) และการให้คะแนนหลายค่า (Polytomous) แต่เนื่องจากข้อสอบปัจจุบันมีการผสมผสานระหว่างข้อสอบการให้คะแนนแบบ 2 ค่า และการให้คะแนนหลายค่า (Polytomous) อยู่ในฉบับเดียวกัน ซึ่งงานวิจัยในประเทศยังไม่ได้ศึกษาเกี่ยวกับข้อสอบลักษณะนี้ อีกการศึกษาเกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ มีนักวัดผลหลายท่านได้ศึกษาตัวแปรเพศ พบว่า เพศส่งผลให้เกิดการทำหน้าที่ต่างกันของข้อสอบ ผู้วิจัยจึงมีความสนใจที่จะศึกษา การเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบรูปแบบผสม โดยใช้คะแนนโครงการประเมินผลนักเรียนนานาชาติ หรือ PISA เนื่องจากเป็นข้อสอบประเมินผลนักเรียนระดับนานาชาติ ซึ่งเป็นข้อสอบรูปแบบผสมที่มีการตรวจให้คะแนนทั้งแบบทวิวิภาคและพหุวิภาค โดยใช้การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยวิธีอัตราส่วนไลค์ลิฮูด (Likehood Raito Test: LRT) วิธีการวิเคราะห์แมนเทล-แฮนส์เซลร่วมกับอัตราส่วนแต้มต่อมาตรฐาน (Standardized Mantel-Haenszel Log-Odds Ratio: LOR Z) และการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยการประยุกต์ใช้โมเดลเชิงเส้นตรงทั่วไปแบบลดหลั่น (HGLM)

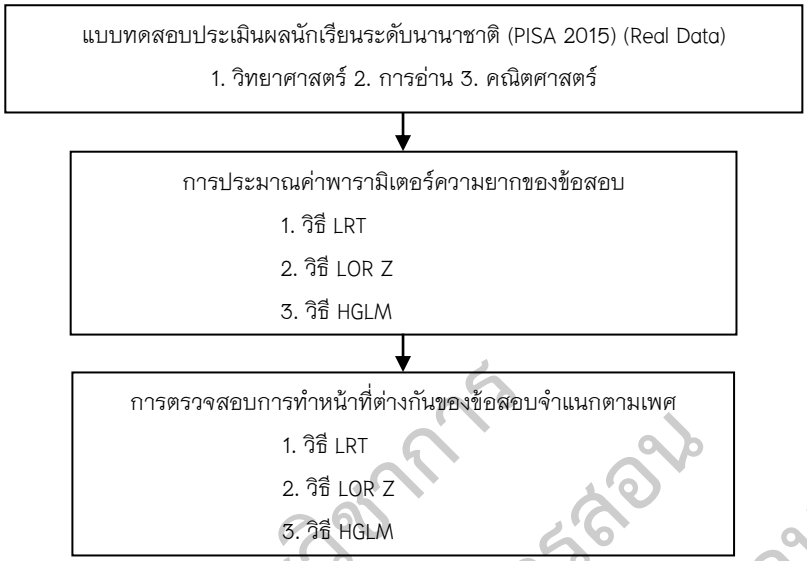
วัตถุประสงค์ของการวิจัย

1. เพื่อเปรียบเทียบผลการประมาณค่าพารามิเตอร์ข้อสอบในแบบทดสอบรูปแบบผสม ระหว่างวิธี LRT วิธี LOR Z และวิธี HGLM

2. เพื่อเปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ (DIF) ในแบบทดสอบรูปแบบผสม จำแนกตามเพศ ระหว่างวิธี LRT วิธี LOR Z และวิธี HGLM

กรอบแนวคิดของการวิจัย

จากการศึกษาแนวคิดทฤษฎี และงานวิจัยต่าง ๆ ที่เกี่ยวข้อง ผู้วิจัยได้วิเคราะห์ สังเคราะห์เพื่อกำหนดกรอบแนวคิดของการวิจัย ซึ่งกรอบแนวคิดของการวิจัยแสดงให้เห็นว่าการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี LRT วิธี LOR Z และวิธี HGLM ในแบบทดสอบรูปแบบผสม โดยทั้งสองวิธีมีความสามารถในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ดังต่อไปนี้



ภาพประกอบ กรอบแนวคิดของการวิจัย

วิธีดำเนินการวิจัย

ประชากรที่ใช้ในการวิจัย

การวิจัยครั้งนี้เป็นการศึกษาจากข้อมูลทุติยภูมิของโครงการประเมินผลนักเรียนนานาชาติ ปี 2558 หรือ PISA 2015 จากประชากรนักเรียนของประเทศไทยที่มีช่วงอายุ 15 ปี 3 เดือน จนถึง 16 ปี 2 เดือน ครอบคลุมทุกพื้นที่ภูมิศาสตร์ของประเทศ 9 พื้นที่ และสังกัดโรงเรียน 9 กลุ่ม จำนวน 8,249 คน ที่เข้าร่วมสอบข้อสอบ PISA จำนวนทั้งหมด 66 ฉบับ โดยพิจารณาให้มีปริมาณของพื้นที่และจำนวนโรงเรียนใกล้เคียงกันมากที่สุด รายละเอียดพื้นฐานตามกรอบการสุ่มจำแนกตามภาคพื้นที่ภูมิศาสตร์และตามสังกัดโรงเรียน

กลุ่มตัวอย่างที่ใช้ในการวิจัย

การวิจัยครั้งนี้เป็นการศึกษาจากข้อมูลทุติยภูมิของโครงการประเมินผลนักเรียนนานาชาติ ปี 2558 หรือ PISA 2015 จากประชากรนักเรียนของประเทศไทยที่มีช่วงอายุ 15 ปี 3 เดือน จนถึง 16 ปี 2 เดือน โดยกลุ่มตัวอย่างได้มาจากการสุ่มอย่างง่าย (Simple Random Sampling) โดยใช้ชุดของแบบทดสอบเป็นหน่วยสุ่ม โดยทำการสุ่มนักเรียนที่ทำแบบทดสอบวิชาละ 1 ฉบับ ได้กลุ่มตัวอย่างทั้งหมด จำนวน 644 คน

คุณภาพเครื่องมือที่ใช้ในการวิจัย

จากรายงานของ OECD (2015) ความเที่ยงของเครื่องมือมีค่าอยู่ระหว่าง .81–.88 แบบทดสอบการรู้เรื่องวิทยาศาสตร์ มีค่าความเที่ยง เท่ากับ .88 แบบทดสอบการรู้เรื่องคณิตศาสตร์ มีค่าความเที่ยง เท่ากับ .81 แบบทดสอบการรู้เรื่องการอ่าน มีค่าความเที่ยง เท่ากับ .86

การวิเคราะห์ข้อมูล

การวิเคราะห์ข้อมูลแบ่งออกเป็น 3 ตอน คือ

ตอนที่ 1 การวิเคราะห์ค่าสถิติพื้นฐาน

ตอนที่ 2 การวิเคราะห์ผลการประมาณค่าพารามิเตอร์ความยากข้อสอบ ในแบบทดสอบรูปแบบผสม ระหว่างวิธี LRT วิธี LOR Z และวิธี HGLM

ตอนที่ 3 การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ (DIF) ในแบบทดสอบรูปแบบผสม จำแนกตามเพศ ระหว่างวิธี LRT วิธี LOR Z และวิธี HGLM

สถิติที่ใช้ในการวิเคราะห์ข้อมูล

1. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยวิธีอัตราส่วนไลค์ลิฮูด (Likehood Raito Test: LRT)

2. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยวิธีการวิเคราะห์แมนเทล–แฮนส์เซลร่วมกับอัตราส่วนแต้มต่อมาตรฐาน (Standardized Mantel–Haenszel Log–Odds Ratio: LOR Z)

3. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยประยุกต์ใช้โมเดลเชิงเส้นตรงทั่วไปแบบลดหลั่น (HGLM)

ผลการวิเคราะห์ข้อมูลและสรุปผลการวิจัย

จากการวิเคราะห์ค่าพารามิเตอร์ความยากของข้อสอบและการทำหน้าที่ต่างกันของข้อสอบครั้งนี้ ประกอบด้วยวิชาการรู้เรื่องวิทยาศาสตร์ (Scientific Literacy) วิชาการรู้เรื่องคณิตศาสตร์ (Mathematical Literacy) และวิชาการรู้เรื่องการอ่าน (Reading Literacy) ซึ่งผู้วิจัยได้ทำการวิเคราะห์ด้วยวิธี LRT วิธี LOR Z และวิธี HGLM โดยการประยุกต์ใช้โปรแกรมโมเดลเชิงเส้นตรงระดับลดหลั่น (HLM) โปรแกรม IRT PRO และโปรแกรม DIFAS ได้ผลการวิจัยดังนี้

1. ผลการเปรียบเทียบผลการประมาณค่าพารามิเตอร์ความยากของข้อสอบ ในแบบทดสอบรูปแบบผสม ระหว่างวิธี LRT วิธี LOR Z และวิธี HGLM ค่าพารามิเตอร์ความยากของวิชาการรู้เรื่องวิทยาศาสตร์ (Scientific Literacy) พบว่า ผลการประมาณค่าพารามิเตอร์ความยากของข้อสอบด้วยวิธี LRT และวิธี HGLM มีความสัมพันธ์กันทางบวก ส่วนวิธี LOR Z สัมพันธ์ทางลบกับวิธี LRT และวิธี HGLM อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ส่วนผลการประมาณค่าพารามิเตอร์ความยากของข้อสอบวิชาการรู้เรื่องคณิตศาสตร์ (Mathematical Literacy) พบว่า ผลการประมาณค่าพารามิเตอร์ความยากของข้อสอบด้วยวิธี LRT และวิธี HGLM มีความสัมพันธ์กันทางบวก ส่วนวิธี LOR Z สัมพันธ์ทางลบกับวิธี LRT และวิธี HGLM อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และผลการประมาณค่าพารามิเตอร์ความยากของข้อสอบวิชาการรู้เรื่องการอ่าน (Reading Literacy) พบว่า วิธี LRT และวิธี HGLM มีความสัมพันธ์กันทางบวก ส่วนวิธี LOR Z สัมพันธ์ทางลบกับวิธี LRT และวิธี HGLM อย่างมีนัยสำคัญทางสถิติที่ระดับ .01

2. ผลการเปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ (DIF) ในแบบทดสอบรูปแบบผสม ระหว่างวิธี LRT วิธี LOR Z และวิธี HGLM จำแนกตามเพศ ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ทั้ง 3 รายวิชา ได้แก่ การรู้เรื่องวิทยาศาสตร์ (Scientific Literacy) จำนวน 29 ข้อ การรู้เรื่องคณิตศาสตร์ (Mathematical Literacy) จำนวน 24 ข้อ และการรู้เรื่องการอ่าน (Reading Literacy)ระหว่างวิธี LRT วิธี LOR Z และวิธี HGLM พบว่า วิธี LOR Z ตรวจพบการทำหน้าที่ต่างกันของข้อสอบมากที่สุด จำนวน 15 ข้อ รองลงมาวิธี HGLM ตรวจพบการทำหน้าที่ต่างกันของข้อสอบ จำนวน 5 ข้อ และวิธี LRT ไม่พบการทำหน้าที่ต่างกันของข้อสอบ

อภิปรายผล

1. ผลการประมาณค่าพารามิเตอร์ความยากของข้อสอบทั้ง 3 รายวิชา ได้แก่ การรู้เรื่องวิทยาศาสตร์ (Scientific Literacy) การรู้เรื่องคณิตศาสตร์ (Mathematical Literacy) และการรู้เรื่องการอ่าน (Reading Literacy) ด้วยวิธี LRT และวิธี HGLM มีความสัมพันธ์กันทางบวก ส่วนวิธี LOR Z สัมพันธ์ทางลบกับวิธี LRT และวิธี HGLM อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 อาจเนื่องมาจากการประมาณค่าพารามิเตอร์ความยากของข้อสอบ (δ) คือ การสร้างโมเดลการวิเคราะห์ที่มีตัวแปรตามเป็นผลการตอบที่ให้คะแนนแบบทวิภาค (0, 1) ซึ่งมีการกระจายแบบทวินาม (Binomial) และเป็นกรณีหนึ่งของการกระจายแบบเบอร์นูลลี (Bernouli) ซึ่ง Raudenbush & Bryk (2002) ได้เสนอให้ใช้ฟังก์ชันการเชื่อมโยง (Link Function) แบบฟังก์ชันโลจิท (Logit Link Function) เพื่อทำให้การประมาณค่าระดับที่ 1 เชื่อมโยงไปสู่การวิเคราะห์ในระดับที่ 2 และระดับที่สูงขึ้นไปได้ โดยระดับการวิเคราะห์ที่ 1 จะเป็นระดับข้อสอบ ใช้โมเดลเชิงเส้นทั่วไป (GLM) ในการวิเคราะห์แบบทวนซ้ำ (iterations) ก่อนเข้าสู่สมการวิเคราะห์ที่ 2 ดังนั้น ในระดับการวิเคราะห์ที่ 1 จึงไม่สามารถใส่ตัวแปรทำนายเข้าสู่สมการได้ ส่วนการวิเคราะห์ระดับที่ 2 ก็จะเป็นการวิเคราะห์ด้วยโมเดลเชิงเส้นทั่วไประดับลดหลั่น (HGLM) อิทธิฤทธิ์ พงษ์ปิยะรัตน์ (2551) สอดคล้องกับงานวิจัยของ Kim (2003) พบว่า ค่าพารามิเตอร์ของข้อสอบจากการประมาณค่าด้วยโมเดล HGLM–2 มีความสัมพันธ์กับค่าพารามิเตอร์ของข้อสอบ ที่ประมาณค่าด้วยโมเดลรายข้ออย่างสมบูรณ์ (r = 1.00) และสอดคล้องกับการศึกษา ของนพดล มีชั้นช่วง (2544) ที่ศึกษาค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าความยากที่ประมาณค่า ด้วยวิธีแมกซิมัมไลค์ลิฮูด วิธีฮิวริสติก และวิธีของเบย์ พบว่า ค่าสัมประสิทธิ์สหสัมพันธ์มีค่าตั้งแต่ 263 ถึง 446 วิธีแมกซิมัมไลค์ลิฮูด และวิธีของเบย์มีนัยสำคัญทางสถิติที่ระดับ .05 และสอดคล้องกับการศึกษาของ อิทธิฤทธิ์ พงษ์ปิยะรัตน์ (2551) ที่ได้ศึกษาการประมาณ

ค่าพารามิเตอร์ความยากของข้อสอบที่วิเคราะห์ค่าพารามิเตอร์ความยากของข้อสอบด้วยโมเดล HGLM–2L ซึ่งเมื่อเชื่อมโยงโมเดลการวิเคราะห์ค่าพารามิเตอร์ความยากของข้อสอบกับหลักการวิเคราะห์เศษเหลือที่ Goldstein (1997) ได้กล่าวไว้ว่า ค่าส่วนที่เหลือที่เกิดจากการวิเคราะห์ถดถอย เมื่อตัวแปรตามเป็นผลสัมฤทธิ์ทางการเรียน ส่วนที่เหลือที่ได้จากการวิเคราะห์เป็นผลสัมฤทธิ์ที่ได้รับการขจัดอิทธิพลแทรกซ้อนต่าง ๆ ออกไป จึงสามารถเป็นตัวบ่งชี้ความสามารถเฉพาะของผลสัมฤทธิ์ทางการเรียนที่เกิดจากการจัดการศึกษาได้ ซึ่งในการวิเคราะห์ข้อสอบพหุระดับนี้ ค่าส่วนที่เหลือก็คือค่าอิทธิพลการสุ่ม (Random Effect) ซึ่งเป็นค่าส่วนเบี่ยงเบนของโอกาสในการตอบข้อสอบถูกของผู้สอบ จากค่าเฉลี่ยของโอกาสในการตอบข้อสอบถูกของนักเรียนในโรงเรียน จึงเป็นค่าที่แสดงถึงความสามารถเฉพาะของผู้สอบแต่ละคนในโรงเรียนที่ m ซึ่งค่าดังกล่าว มีการแจงแจงเป็นแบบปกติค่าเฉลี่ยเท่ากับศูนย์และค่าความแปรปรวนเป็น $\tau(U_{ojm} \sim N[0,\tau])$ และจากกรณีที่ค่าสัมประสิทธิ์สหสัมพันธ์ของค่าพารามิเตอร์ วิธี LOR Z สัมพันธ์ทางลบกับวิธี LRT และวิธี HGLM เนื่องมาจากการประมาณค่าพารามิเตอร์ความยากของด้วยวิธี LOR Z เป็น Classical test theory แต่การประมาณค่าพารามิเตอร์ความยากของด้วยวิธี LRT และวิธี HGLM เป็น Item response theory ซึ่งสอดคล้องกับงานวิจัยของ Nese Guler (2013) ที่ศึกษาเปรียบเทียบการประมาณค่าพารามิเตอร์ระหว่าง Classical test theory และ Item response theory พบว่าการประมาณค่าพารามิเตอร์ความยากของข้อสอบระหว่าง Classical test theory และ Item response theory มีค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ –.9986 นั่นคือความสัมพันธ์การทางลบ

2. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ทั้ง 3 รายวิชา ได้แก่ การรู้เรื่องวิทยาศาสตร์ (Scientific Literacy) การรู้เรื่องคณิตศาสตร์ (Mathematical Literacy) และการรู้เรื่องการอ่าน (Reading Literacy) ระหว่างวิธี LRT วิธี LOR Z และวิธี HGLM จำแนกตามเพศ พบว่า วิธี LOR Z ตรวจพบการทำหน้าที่ต่างกันของข้อสอบมากที่สุด จำนวน 15 ข้อ รองลงมาวิธี HGLM ตรวจพบการทำหน้าที่ต่างกันของข้อสอบ จำนวน 5 ข้อ และวิธี LRT ไม่พบการทำหน้าที่ต่างกันของข้อสอบ ซึ่งสอดคล้องกับการศึกษาของ Fukahara & Kamata (2007) ได้ศึกษาการประเมินประสิทธิภาพการทำงานของวิธี MIMIC แบบละเมิดข้อตกลงเบื้องต้น โดยการทำชุดข้อสอบ พบว่า การทำหน้าที่ต่างกันของข้อสอบ (DIF) มีแนวโน้มที่จะการประมาณค่าได้ภายใต้ข้อสอบที่ไม่เป็นอิสระกัน ซึ่งสอดคล้องกับการศึกษาของ เกษร หว่างจิตร (2539) พบว่า ระบุลักษณะของข้อสอบที่ทำหน้าที่ต่างกัน ของข้อสอบระหว่างกลุ่มผู้สอบ เมื่อจำแนกตามตัวแปรเพศ มีความแตกต่างกัน โดยเมื่อจำแนกผู้สอบตามเพศ พบว่า มีข้อสอบ ที่มีการระบุว่าเกิด DIF มากที่สุด และสอดคล้องกับผลการศึกษาของรักชนก ยี่สุ่นศรี (2544) ที่ศึกษาการทำหน้าที่ต่างกันของข้อสอบระหว่างกลุ่มผู้สอบเมื่อจำแนกตามเพศ พบว่าเมื่อจำแนกกลุ่มผู้สอบตามเพศจะมีข้อสอบที่ถูกระบุว่าเกิด DIF ทั้งในแบบสอบวิชาภาษาอังกฤษและวิชาคณิตศาสตร์ Maier & Casselman (1970) กล่าวว่า เพศชายจะมีความสามารถในการแก้ปัญหาคณิตศาสตร์ เพศหญิงจะมีความสามารถทางภาษา และเป็นการถกเถียงกัน ในการเรียนที่ต่างกันระหว่างเพศหญิงและเพศชาย เมื่อคนส่วนใหญ่เชื่อว่า เพศชายเรียนวิชาคณิตศาสตร์และวิชาวิทยาศาสตร์เก่งกว่าเพศหญิง ขณะที่เพศหญิงเรียนภาษาศาสตร์และสังคมศาสตร์เก่งกว่าเพศชาย ซึ่งสอดคล้องกับนักวิจัยชาวอเมริกัน และจากการศึกษาของ Professor Nicole Else–Quest ยังพบว่า เพศหญิงสามารถเรียนรู้และนำวิชาคณิตศาสตร์และวิชาวิทยาศาสตร์ไปใช้ได้ไม่แพ้เพศชาย ยิ่งไปกว่านั้น เพศหญิงส่วนใหญ่ทำได้ดีกว่าเพศชาย และสอดคล้องกับงานวิจัยของ ประกฤติญา ทักษิโณ (2552) ได้ศึกษาผลของข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบประเมินการรู้เรื่องด้านวิทยาศาสตร์ต่อการประเมินคุณภาพการจัดการศึกษาโดยใช้การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยประยุกต์ใช้โมเดลเชิงเส้นตรงทั่วไปแบบลดหลั่น (HGLM) ผลการวิจัยพบว่า ข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบวิชาวิทยาศาสตร์ไม่มีผลต่อการประเมินคุณภาพการจัดการศึกษา และเมื่อมีการควบคุมอิทธิพลของคุณลักษณะของนักเรียนและสถานศึกษา พบว่าข้อสอบที่ทำหน้าที่ต่างกันทำให้การจัดระดับคุณภาพการจัดการศึกษาของสถานศึกษาแตกต่างกันอย่างมีนัยสำคัญที่ระดับ .05 สอดคล้องกับงานวิจัยของ พนัส จันทรเปล่ง (2554) ได้ศึกษาผลการทำหน้าที่ต่างกันของข้อสอบ (DIF) ที่มีการตรวจให้คะแนนทวิภาคและพหุวิภาค และการทำหน้าที่ต่างกันของแบบทดสอบ (DTF) โดยใช้ตัวประมาณค่าการทำหน้าที่ต่างกันของแบบทดสอบทั่วไป พบว่า การทำหน้าที่

ต่างกันของข้อสอบ มีข้อสอบที่เอนเอียงเข้าข้างนักเรียนหญิง 3 ข้อ นักเรียนที่เรียนพิเศษวิทยาศาสตร์นอกสถานศึกษา 1 ข้อ นักเรียนที่มาจากครอบครัวที่มีเศรษฐกิจฐานะทางบ้านต่ำ 2 ข้อ นักเรียนที่มาจากครอบครัวฐานะทางบ้านสูง 5 ข้อ และนักเรียนที่ครอบครัวมีความมั่งคั่งสูง 2 ข้อ และสอดคล้องกับงานวิจัยของ สุพัฒนา หอมบุปผา (2556) ได้ตรวจสอบการข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบด้วยวิธี HGLM วิธี MIMIC วิธี BAYSIEN โดยใช้คะแนนการสอบ ONET วิชาภาษาไทย คณิตศาสตร์ และวิทยาศาสตร์ ผลการวิจัยพบว่า วิธีการตรวจสอบที่ทำหน้าที่ต่างกันของข้อสอบมากที่สุด คือ วิธี HGLM–2L ส่วนวิธีที่ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบน้อยที่สุด คือ วิธี MIMIC และสอดคล้องกับ Finch (2005) เปรียบเทียบประสิทธิภาพของโมเดล MIMIC กับการทดสอบโดยวิธีแมนเทลเฮนเซล (Mantel and Haenzel, 1959) วิธี SIBTEST (Shealy and Stout, 1993) และวิธีการทดสอบ IRT likelihood ratio (Thissen et al., 1988) เกณฑ์การเปรียบเทียบพิจารณาจากความคลาดเคลื่อนประเภทที่ 1 และอำนาจการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ได้แสดงให้เห็นว่าวิธี MIMIC มีอำนาจการตรวจสอบสูงขึ้น และความคลาดเคลื่อนประเภทที่ 1 มีค่าลดลงเมื่อข้อสอบ 50 ข้อ สอดคล้องกับ Lei, Chen and Yu (2006) ได้ศึกษาเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบแบบปรับเหมาะโดยใช้คอมพิวเตอร์จำลองข้อมูลการทำหน้าที่ต่างกันของข้อสอบทั้งแบบมีทิศทางและไม่มีทิศทาง ภายใต้เงื่อนไขที่ศึกษา คือ ขนาดกลุ่มตัวอย่างของกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบการแจกแจงความสามารถที่แตกต่างกัน ข้อสอบแบ่งเป็น 3 แบบ คือ ข้อสอบไม่เกิดการทำหน้าที่ต่างกัน ข้อสอบทำหน้าที่ต่างกันแบบมีทิศทาง ข้อสอบทำหน้าที่ต่างกันแบบไม่มีทิศทาง ใช้วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ 3 วิธี ได้แก่ วิธีถดถอยโลจิสติก วิธีการทดสอบอัตราส่วนไลด์ลิฮู้ดแบบ IRT (IRT likelihood ratio test) และวิธีแคทซิบ (CATSIB) ผลการศึกษาพบว่า วิธีถดถอยโลจิสติก (Logistic regression) และวิธีการทดสอบอัตราส่วนไลด์ลิฮู้ดแบบ IRT (IRT likelihood ratio test) ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้งแบบมีทิศทางและแบบลบไม่มีทิศทางได้ดีเท่าเทียมกัน และดีกว่าวิธีแคทซิบ (CATSIB) ในขณะที่วิธีแคทซิบ (CATSIB) สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบมีทิศทางได้ดีกว่าแบบไม่มีทิศทาง และ Stark, Chernyshenko and Drasgow (2006) ได้ศึกษาพัฒนาและทดสอบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยการทดสอบอัตราส่วนไลด์ลิฮู้ดแบบ IRT (IRT likelihood ratio test) ซึ่งสามารถใช้ได้ทั้งการวิเคราะห์องค์ประกอบเชิงยืนยันและทฤษฎีการตอบสนองข้อสอบ โดยใช้ข้อมูลจำลองในการตรวจสอบความตรงของทั้งสองวิธี เครื่องมือที่ใช้เป็นแบบวัดมิติเดียว จำนวน 15 ข้อ ศึกษาตัวแปร 8 ตัว วิเคราะห์วิธีวิเคราะห์องค์ประกอบเชิงยืนยัน โดยใช้โปรแกรมลิสเรล 8 และการทดสอบด้วยทฤษฎีตอบสนองข้อสอบด้วยวิธี likelihood ratio test โดยใช้โปรแกรมคอมพิวเตอร์ MULTILOG ผลการวิจัยพบว่า ทฤษฎีการตอบสนองข้อสอบ วิธีทดสอบ likelihood ratio test ให้ผลดีกว่าวิธีวิเคราะห์องค์ประกอบเชิงยืนยัน กรณีกลุ่มตัวอย่างมีขนาดใหญ่และข้อมูลเป็นแบบ Dichotomous มิติเดียว การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้วิธีทฤษฎีการตอบสนองข้อสอบให้ผลดีกว่า และวิธีทฤษฎีการตอบสนองข้อสอบหลาย ๆ วิธี จะมีความแกร่งในการทดสอบ (Robust) หากมีการฝ่าฝืนข้อตกลงเบื้องต้นในเรื่องความเป็นเอกมิติ

ข้อเสนอแนะ

ข้อเสนอแนะในการนำไปใช้

1. การประมาณค่าพารามิเตอร์ความยากของข้อสอบ ควรเลือกวิเคราะห์ด้วยวิธี LRT จากโปรแกรม IRT PRO เพราะมีการวิเคราะห์ที่ง่าย สะดวก และไม่ซับซ้อน วิเคราะห์ในขั้นตอนเดียว ซึ่งผลการวิเคราะห์ได้สอดคล้องกันกับวิธี LOR Z และวิธี HGLM

2. การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ ควรเลือกการวิเคราะห์ด้วยวิธี LOR Z โดยใช้โปรแกรม DIFAS เนื่องจากวิเคราะห์ง่าย ไม่ซับซ้อน ใช้งานสะดวก และที่สำคัญผลการวิเคราะห์สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้มากกว่าวิธี LRT และวิธี HGLM

3. การศึกษาครั้งนี้ได้ข้อค้นพบว่า ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบรูปแบบผสม ควรใช้วิธี LOR Z และวิธี HGLM มากกว่าวิธี LRT เนื่องจากการวิเคราะห์ไม่พบการทำหน้าที่ต่างกันของข้อสอบ

4. ควรมีการปรับปรุงลักษณะข้อสอบ ทั้งข้อคำถามและตัวเลือกคำตอบให้มีความเป็นปรนัยมากยิ่งขึ้น ในข้อที่ตรวจพบการทำหน้าที่ต่างกันของข้อสอบ เมื่อจำแนกตามเพศเพื่อหลีกเลี่ยงการทำหน้าที่ต่างกันของข้อสอบ

ข้อเสนอแนะสำหรับการทำวิจัยครั้งต่อไป

1. ควรศึกษาการประมาณค่าพารามิเตอร์ของข้อสอบด้วยวิธีอื่น ๆ และโปรแกรมอื่น ๆ

2. ควรศึกษาการเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ในแบบทดสอบรูปแบบผสมด้วยวิธีอื่น ๆ และโปรแกรมอื่น ๆ

3. ควรศึกษาการเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ในแบบทดสอบรูปแบบผสมโดยใช้วิธีการจำลองข้อมูล (Simulation Data)

4. ควรศึกษาการเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ในแบบทดสอบรูปแบบผสมโดยวิเคราะห์ตัวแปรอื่น ๆ เช่น สังกัดโรงเรียน สถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน เป็นต้น โดยใช้คะแนนจากแบบทดสอบอื่น ๆ ที่สามารถระบุตัวแปรอื่น ๆ ได้

เอกสารอ้างอิง

กระทรวงศึกษาธิการ. (2545). *พระราชบัญญัติการศึกษาแห่งชาติ พ.ศ. 2542 และที่แก้ไขเพิ่มเติม (ฉบับที่ 2) พ.ศ. 2545.* กรุงเทพฯ: โรงพิมพ์คุรุสภาลาดพร้าว.

เกษร หว่างจิตร. (2539). *การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบสำหรับแบบสอบคัดเลือก ระดับบัณฑิตศึกษา วิชาภาษาไทยและภาษาอังกฤษ.* วิทยานิพนธ์ ค.ม. กรุงเทพฯ: จุฬาลงกรณ์มหาวิทยาลัย.

นพดล มีชั้นช่วง. (2544). *การเปรียบเทียบผลของการประมาณค่าพารามิเตอร์ตามทฤษฎีการตอบสนอง ข้อสอบระหว่างวิธีแมกซิมัมไลค์ลิฮูด วิธีฮิวริสติก และวิธีของเบย์ของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนคณิตศาสตร์ชั้นมัธยมศึกษาปีที่ 1.* วิทยานิพนธ์ กศ.ม. มหาสารคาม: มหาวิทยาลัยมหาสารคาม.

ประกฤติญา ทักษิโณ. (2552). *การประเมินคุณภาพการจัดการศึกษาวิชาวิทยาศาสตร์ของสถานศึกษาขั้นพื้นฐาน: การประยุกต์ใช้การทำหน้าที่ต่างกันของข้อสอบและโมเดลมูลค่าเพิ่ม.* วิทยานิพนธ์ ค.ม. กรุงเทพฯ: จุฬาลงกรณ์มหาวิทยาลัย.

พนัส จันทรเปล่ง. (2554). *การวิเคราะห์เปรียบเทียบโมเดลประเมินคุณภาพการจัดการศึกษาวิทยาศาสตร์: การประยุกต์ใช้โมเดลเพิ่มพหุระดับที่มีการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบและแบบสอบ.* วิทยานิพนธ์ ค.ม. กรุงเทพฯ: จุฬาลงกรณ์มหาวิทยาลัย.

รักชนก ยี่สุ่นศรี. (2544). *การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบและแบบสอบด้วยกระบวนการ ดี เอฟ ไอ ที สำหรับแบบสอบคัดเลือกบุคคลเข้าศึกษาในสถาบันอุดมศึกษา วิชาภาษาอังกฤษและวิชาคณิตศาสตร์.* วิทยานิพนธ์ ค.ด. กรุงเทพฯ: จุฬาลงกรณ์มหาวิทยาลัย.

ศิริชัย กาญจนวาสี. (2550). *ทฤษฎีการทดสอบแบบดั้งเดิม* พิมพ์ครั้งที่ 6. กรุงเทพฯ: จุฬาลงกรณ์มหาวิทยาลัย.

สุพัฒนา หอมบุปผา. (2556). *การเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี HGLM วิธี Mimic และวิธี bayesian.* ดุษฎีนิพนธ์ ปร.ด. ชลบุรี: มหาวิทยาลัยบูรพา.

อาวีพร ปานทอง. (2558). *การเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบให้คะแนนหลายค่าโดยวิธีทดสอบอัตราส่วนความควรจะเป็นวิธีเบส์เซียนและวิธีโพลี–ซิปเทสท์.* ดุษฎีนิพนธ์ ปร.ด. ชลบุรี: มหาวิทยาลัยบูรพา.

อิทธิฤทธิ์ พงษ์ปิยะรัตน์. (2551). *การวิเคราะห์ข้อสอบและการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ: การวิเคราะห์พหุระดับ.* วิทยานิพนธ์ ค.ด. กรุงเทพฯ: จุฬาลงกรณ์มหาวิทยาลัย.

Bolt–Lee, C, (2002). New competencies for accounting students. *The CPA journal*, 72(1), 68–71.

Carvajal, J., &Skorupski, W.P. (2010). The Effects of Small Sample Size on Identifying Ploytomous DIF Using the Liu–Agresti Estimator of the Cumulative Common Odds Ratio. *Educational and Psychological Measurement*, 70(6), 914–925.

Finch, H. (2005). *The MIMIC model as a method for detecting DIF*: Comparison with Mantel–Haenszel, SIBTEST and the IRT likelihood ratio. *Applied Psychological Measurement*, 29(5), 278–295.

Fukuhara, H., & Kamata, A. (2007). *DIF detection in a presence of locally dependent items*. Florida: Tampa.

Goldstein, H. (1997). Methods in school effectiveness research. *School Effectiveness and School Improvement*, 8(4), 369–395.

Jose Luis Padilla. (2012). What Cognitive Interviews Tell Us about Bias in Cross–cultural Research: An Illustration Using Quality–of–life Items. *Measurement*, 58, 468–475.

Lei, Chen and Yu (2006) Lei Zou, Lei Chen, J. Xu Yu, Y. Lu. (2006). *A Novel Spectral Coding in a Large Graph Database*. in Proc: of EDBT.

Maier, N. R. F., & Casseiman, G. G. (1970). Locating the difficulty in insight problems: Individual and sex differences. *Psychological Rep*, 26, 103–107.

Nese Guler. (2013). A study on multiline linear regression analysis. *Social and behavioral sciences*, 106(201), 234–236.

Kim, S., and Walker, M. E. & Mchale, F. (2009). *Evaluating subpopulation invariance of linking functioms to determine the anchor composition fot a mixed format test (Research Report 09–36)*. Princetom NJ: Educational Testing Service.

OECD. (2015). *PISA 2015 Technical Report*. Paris: OECD.

Penfield, R. D. (2005). DIFAS: Differential Item Functioning Analysis System. *Applied Psychological Measurement*, 29(2), 150–151.

Penfield, R. D. (2013). *Item analysis. In K. Geisinger (Ed.), APA handbook of testing and assessment in psychology*. Washington, DC: American Psychological Association.

Raudenbush, S. W., & Bryk, A. S. (2002). A hierarchical model for studying school effects. *Sociology of Education*, 59, 1–17.

R.J. DE AYALA. (2009). *The Theory and Practice of Item Response Theory*. New York: The Guilford Press.

Shealy, R., & Stout, W. (1993, June). A model–based standardization approach that separates true bias/DIF from group ability differences and detects testbias/DTF as well as item bias/DIF. *Psychometrika*, 58(2), 159–194.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, 91, 1202–1306.

Thissen, D., Steinberg, L., &Wainer, H. (1988). *Use of item response theory in the study of group differences in trace lines*. In H. Wainer, &H. I. Braun (Eds.), Test validity (pp. 147–169). Hillsdale, NJ: Erlbaum.