การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ชั้นประถมศึกษาปีที่ 3 ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR

A Comparison of Differential Item Functioning Detection in National Tests of Literacy, Numeracy, and Reasoning Abilities at the Grade Three Level Using HGLM, MIMIC and IRT-LR Methods

สุธาทิพย์ ตรีสิน [1]*, ปิยะทิพย์ ประดุจพรม [1]

Suthathip Treesin [1]*, Piyathip Pradujprom [1]

[1] College of Research Methodology and Cognitive Science, Burapha University, Thailand

บทคัดย่อ

การวิจัยนี้มีวัตถุประสงค์เพื่อวิเคราะห์คุณภาพของแบบทดสอบระดับชาติ (NT) และตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผลด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR การดำเนินการวิจัยแบ่งเป็น 3 ระยะ ดังนี้ 1) วิเคราะห์คุณภาพของแบบทดสอบระดับชาติ ทั้ง 3 ด้าน 2) ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR และ 3) เปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธีการตรวจสอบ 3 วิธีข้อมูลที่นำมาใช้วิเคราะห์เป็นข้อมูลทุติยภูมิ จากผลการตอบแบบทดสอบระดับชาติของนักเรียนชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2556 จำนวน 9,600 คน

ผลการวิจัยปรากฏว่า

1) แบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 มีค่าความยากของข้อสอบ (b) อยู่ในระดับค่อนข้างยากมีค่าอำนาจจำแนกของข้อสอบ (a) อยู่ในระดับที่สามารถจำแนกผู้สอบได้ดี และมีค่าโอกาสในการเดาของข้อสอบ (c) ไม่เกิน 0.3

2) การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้ง 3 ด้าน ชี้ให้เห็นว่า เพศส่งผลให้เกิดการทำหน้าที่ต่างกันของข้อสอบ โดยเพศหญิงจะได้เปรียบในการตอบข้อสอบด้านภาษา และด้านเหตุผล ในขณะที่เพศชายจะได้เปรียบในการตอบข้อสอบด้านคำนวณ โดยวิธี HGLM ตรวจพบข้อสอบทำหน้าที่ต่างกัน จำนวนมากที่สุด คิดเป็นร้อยละ 69 ของข้อสอบทั้งฉบับ รองลงมาคือ วิธี IRT-LR ร้อยละ 54 และวิธี MIMIC ร้อยละ 16 ตามลำดับ

3) การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ พบว่า วิธี HGLM ตรวจพบ DIF มากกว่าวิธี MIMIC ในด้านภาษา ด้านคำนวณ และด้านเหตุผล คิดเป็นร้อยละ 70, 36 และ 53 ตามลำดับ และวิธี HGLM ตรวจพบ DIF มากกว่าวิธี IRT-LR ด้านภาษา และด้านคำนวณ คิดเป็นร้อยละ 37 และ 13 และวิธี IRT-LR ตรวจพบ DIF มากกว่าวิธี MIMIC ทั้ง 3 ด้าน คิดเป็นร้อยละ 33, 43 และ 40 ตามลำดับ ส่วนวิธี HGLM ตรวจพบ DIF น้อยกว่าวิธี IRT-LR ด้านคำนวณ คิดเป็นร้อยละ 7 ($p<.05$)

คำสำคัญ: การทำหน้าที่ต่างกันของข้อสอบ, วิธี HGLM, วิธี MIMIC, วิธี IRT-LR, แบบทดสอบระดับชาติ

**Corresponding author. E-mail: suthathip.cm@gmail.com*

ABSTRACT

The objectives of this research were to analyze the quality of National Tests (NT) and to investigate the possibility of Differential Item Functioning (DIF) in three subjects: Literacy, Numeracy, and Reasoning by using HGLM, MIMIC, and IRT-LR methods. The research methods were divided into three phases: 1) Analyzing the quality of NT exam item for three subjects; 2) Testing DIF detection of the items in NT using HGLM, MIMIC, and IRT-LR methods; and 3) Comparing the results of DIF three methods using secondary data from NT examination of 9,600 Grade three students academic year 2013.

Results were as follows:

1. The national tests had IRT difficulty parameter values at relatively difficult levels, discrimination parameter values capable of differentiating examiners at a good level, and guessing parameters not exceeding 0.30.

2. The examination of possible DIF in the three subjects revealed that gender affected the test scores; female students had an advantage when answering the Literacy, and Reasoning subjects, while male students had an advantage in the Numeracy subject. In addition, the HGLM method indicated that the three most common DIF tests could account for 69% of the test, followed by the IRT-LR at 54% and MIMIC at 16%, respectively.

3. Comparison of the DIF test results revealed that the HGLM method outperformed the MIMIC method in terms of DIF detection, namely 70% for Literacy, 36% for Numeracy, and 53% for Reasoning subjects. The HGLM method also outperformed the IRT-LR method in terms of DIF detection, namely 37% for Literacy and 13% for Numeracy subjects. The IRT-LR method outperformed the MIMIC method in terms of DIF detection, namely 33% for Literacy, 43% for Numeracy, and 40% for Reasoning subjects. Also, the HGLM method outperformed the IRT-LR method in terms of DIF detection for only Numeracy subjects (7%) ($p<.05$).

Keywords: differential item functioning, hierarchical generalized linear modeling, multiple-indicator multiple-causes, item response theory–likelihood ratio, national tests

ความนำ

กระทรวงศึกษาธิการมีแผนการพัฒนาการศึกษาของประเทศไทยให้มีประสิทธิภาพมากยิ่งขึ้น เพื่อให้สอดคล้องกับรัฐธรรมนูญแห่งราชอาณาจักรไทย พุทธศักราช 2550 และความต้องการของท้องถิ่น ซึ่งกำหนดเป้าหมายหลักเน้นไปที่ผู้เรียนให้ได้รับการศึกษาที่มีคุณภาพและสถาบันการศึกษาทุกระดับทุกประเภท จะต้องผ่านการรับรองมาตรฐานทางการศึกษา (สำนักงานปลัดกระทรวงศึกษาธิการ, 2554) เนื่องจากมาตรฐานการเรียนรู้เป็นกลไกสำคัญในการพัฒนาการศึกษา และยังเป็นเครื่องมือในการตรวจสอบเพื่อประกันคุณภาพการศึกษาโดยใช้การประเมินคุณภาพภายในและภายนอก รวมถึงการทดสอบระดับพื้นที่การศึกษาและการทดสอบระดับชาติ ซึ่งระบบการตรวจสอบเพื่อประกันคุณภาพดังกล่าวเป็น

สิ่งสำคัญที่ช่วยสะท้อนภาพการจัดการศึกษาว่า สามารถพัฒนาผู้เรียนให้มีคุณภาพตามที่มาตรฐานการเรียนรู้กำหนดเพียงใด (กระทรวงศึกษาธิการ, 2551)

สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน (สพฐ.) กำหนดให้มีการทดสอบระดับชาติ (National Testing: NT) เป็นการทดสอบเพื่อวัดความรู้ของนักเรียนในแต่ระดับชั้นว่า มีความรู้ความสามารถในระดับใด เพื่อนำไปวิเคราะห์สภาพปัญหาการจัดการเรียนการสอน โดยยึดตามหลักการของการศึกษาในศตวรรษที่ 21 เน้นการประเมินความสามารถของผู้สอบ 3 ด้าน คือ ด้านภาษา (Literacy) ด้านคำนวณ (Numeracy) และด้านเหตุผล (Reasoning Abilities) มีนโยบายพัฒนาระบบการทดสอบการวัดและประเมินผลเพื่อให้เป็นเครื่องมือที่ช่วยส่งเสริมการปฏิรูปการเรียนรู้และการพัฒนาคุณภาพของผู้เรียนให้มีมาตรฐานเทียบเท่านานาชาติ ปัจจุบันมีการตรวจสอบคุณภาพของเครื่องมือหลากหลายวิธี ซึ่งวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) เป็นวิธีหนึ่งที่นักวัดผลการศึกษาส่วนใหญ่ใช้ในการตรวจสอบคุณภาพของข้อสอบ

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) เป็นการเปรียบเทียบผลการตอบข้อสอบระหว่างผู้สอบ 2 กลุ่ม ที่มีความสามารถในระดับเดียวกัน ประกอบด้วย 1) กลุ่มเปรียบเทียบ (Focal Group: F) เป็นกลุ่มผู้สอบที่สนใจศึกษาและคาดว่าจะเสียเปรียบในการตอบข้อสอบ คือ เป็นกลุ่มผู้สอบที่มีโอกาสในการตอบข้อสอบได้ถูกน้อยกว่ากลุ่มอ้างอิง และ 2) กลุ่มอ้างอิง (Reference Group: R) เป็นกลุ่มผู้สอบที่คาดว่าจะได้เปรียบในการตอบข้อสอบ คือ เป็นกลุ่มผู้สอบที่มีโอกาสในการตอบข้อสอบได้ถูกมากกว่ากลุ่มเปรียบเทียบ ซึ่งการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นการขจัดข้อสอบที่จะก่อให้เกิดความไม่เท่าเทียมกันในแบบทดสอบ เมื่อผู้สอบมีคุณลักษณะแตกต่างกัน เช่น เพศ ภาษา หรือภูมิลำเนา (De Ayala, 2009) โดย Le (2009) ได้ศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างเพศ ในโครงการประเมินผลนักเรียนนานาชาติ (Programme for International Assessment: PISA) พบว่า เพศมีผลต่อผลการทดสอบในโครงการประเมินผลนักเรียนนานาชาติ และ Taylor and Lee (2012) ได้ศึกษาการทำหน้าที่ต่างกันของข้อสอบในข้อสอบการอ่านและคณิตศาสตร์ เมื่อจำแนกตามเพศ พบว่า ข้อสอบที่คาดว่าเพศชายจะได้เปรียบ คือ เรขาคณิต ความน่าจะเป็น และพีชคณิต ส่วนข้อสอบคณิตศาสตร์ที่เพศหญิงจะได้เปรียบ คือ การตีความทางสถิติ การแก้ปัญหาหลายขั้นตอน และการให้เหตุผลเชิงคณิตศาสตร์

จากการศึกษางานวิจัย พบว่า Acar and Kelecioglu (2010) ได้เปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ในแบบทดสอบด้านสังคมศาสตร์และด้านวิทยาศาสตร์ ระหว่างวิธี Hierarchical Generalized Linear Modeling (HGLM) วิธี Logistics Regression (LR) และ วิธี Item Response Theory – Likelihood Ratio (IRT-LR) พบว่า ทั้ง 3 วิธี ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันในจำนวนที่ใกล้เคียงกัน แต่วิธี HGLM เป็นวิธีที่สามารถตรวจสอบพบข้อสอบที่คาดว่าจะเกิดการทำหน้าที่ต่างกันมากที่สุดในแบบทดสอบทั้งสองด้าน และ Finch (2005) ได้เปรียบเทียบความสามารถของตัวแบบหลายตัวบ่งชี้หลายสาเหตุ (Multiple-Indicators Multiple-Causes: MIMIC) รูปแบบการวิเคราะห์ปัจจัยยืนยันเพื่อระบุการทำหน้าที่ต่างกันของข้อสอบพบว่า วิธี MIMIC มีประสิทธิภาพในการตรวจพบข้อสอบที่ทำหน้าที่ต่างกันในแบบทดสอบที่มีความยาวของข้อสอบ จำนวน 50 ข้อขึ้นไป ในขณะที่ Li, Hunter, and Oshima (2013) ได้ศึกษาการตรวจสอบการทำหน้าที่ต่างกันในการทดสอบการอ่านและเหตุผลที่เป็นไปได้ระหว่างเพศ ด้วยวิธี IRT-LR และ วิธี Mantel-Haenszel (MH) พบว่า เพศมีผลต่อการทำแบบทดสอบในการอ่านและเหตุผลที่เป็นไปได้ วิธี IRT-LR สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ดีในแบบทดสอบที่มีความยาวตั้งแต่ 20 ข้อขึ้นไป

ผู้วิจัยได้เลือกใช้วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้งหมด 3 วิธี คือ วิธี HGLM วิธี MIMIC และวิธี IRT-LR ซึ่งทั้ง 3 วิธี อยู่ภายใต้พื้นฐานของทฤษฎีการตอบสนองข้อสอบ (IRT) และสามารถตรวจสอบ DIF

ได้ดีในแบบทดสอบที่มีการตรวจให้คะแนนแบบ 2 ค่า (Dichotomous) สำหรับตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ (NT) ด้านภาษา ด้านคำนวณ และด้านเหตุผล ชั้นประถมศึกษาปีที่ 3 เพื่อศึกษาผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบของทั้ง 3 วิธี ในแบบทดสอบทั้ง 3 ด้าน เมื่อจำแนกตามเพศว่า ข้อสอบเกิดการทำหน้าที่ต่างกันหรือไม่ และทั้ง 3 วิธี วิธีใดสามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันได้มากกว่ากัน

วัตถุประสงค์ของการวิจัย

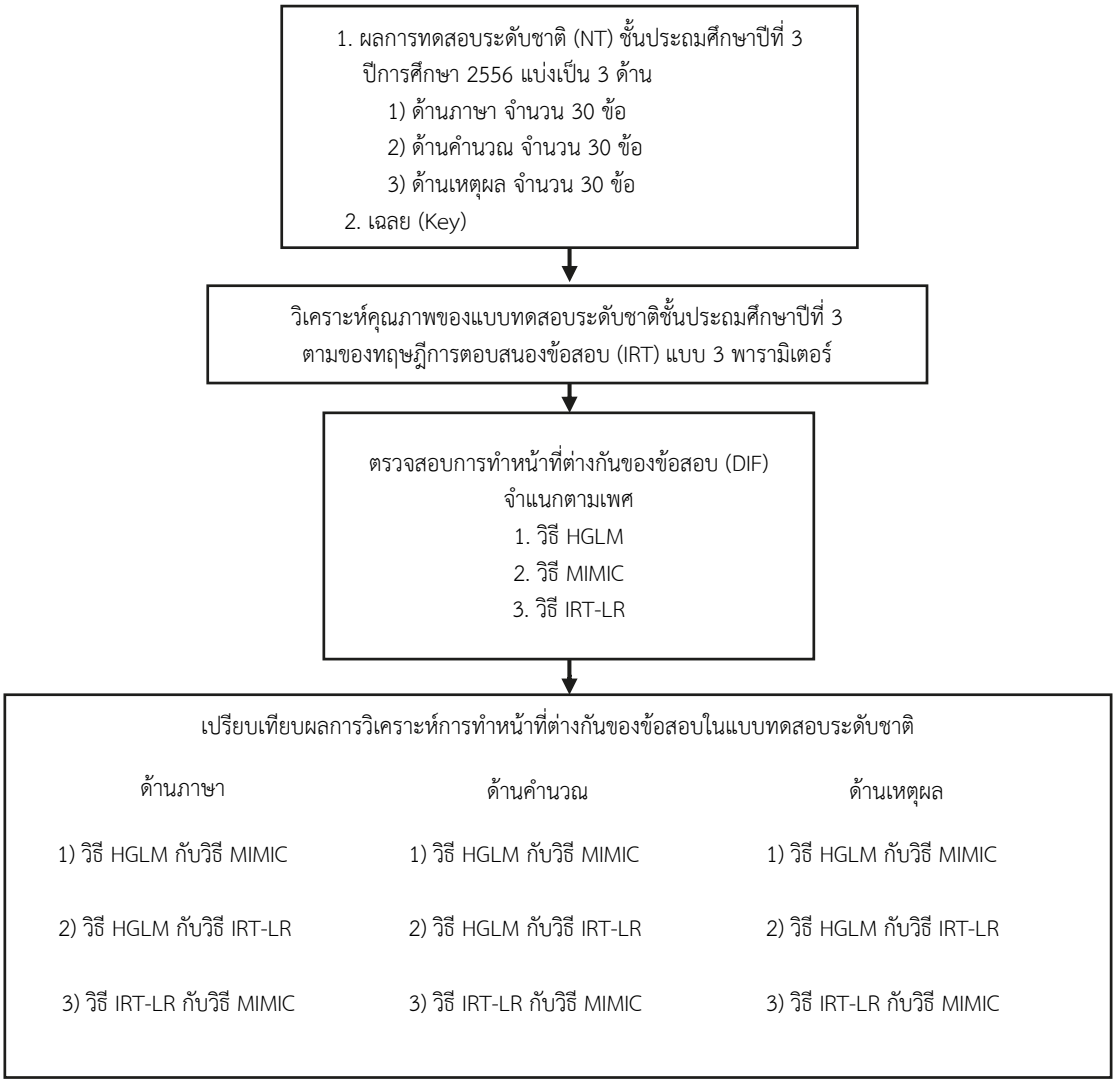
1. เพื่อวิเคราะห์คุณภาพของแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล โดยใช้ทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์

2. เพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และ วิธี IRT-LR จำแนกตามเพศ

3. เพื่อเปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และ วิธี IRT-LR จำแนกตามเพศ

กรอบแนวคิดการวิจัย

จากการศึกษางานวิจัยของ Acar and Kelecioglu (2010) ได้เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ในแบบทดสอบด้านสังคมศาสตร์และด้านวิทยาศาสตร์ ด้วยวิธี HGLM วิธี LR และวิธี IRT-LR พบว่า วิธี HGLM ตรวจพบ DIF ได้มากที่สุดในแบบทดสอบทั้ง 2 ด้าน ส่วนวิธี LR และวิธี IRT-LR ตรวจพบ DIF ได้ใกล้เคียงกัน และงานวิจัยของ Kabasakal, Arsan, Gok, and Kelecioglu (2014) ศึกษาประสิทธิภาพในการตรวจสอบ DIF ด้วยวิธี MH วิธี SIBTEST และวิธี IRT-LR พบว่า วิธี IRT-LR มีประสิทธิภาพในการตรวจสอบ DIF ได้ดีในแบบทดสอบที่มีความยาวของข้อสอบไม่เกิน 20 ข้อ และสามารถตรวจสอบ DIF ได้ดีในแบบทดสอบที่มีการตรวจให้คะแนนแบบ 2 ค่า นอกจากนี้งานวิจัยของ Li, Hunter, and Oshima (2013) ได้ศึกษาการตรวจสอบการทำหน้าที่ต่างกันในแบบทดสอบด้านการอ่าน และด้านเหตุผล ด้วยวิธี IRT-LR และ วิธี MH พบว่า วิธี IRT-LR สามารถตรวจสอบ DIF ได้ดีในแบบทดสอบที่มีความยาวตั้งแต่ 20 ข้อขึ้นไป และเพศมีผลต่อการทำแบบทดสอบด้านการอ่านและด้านเหตุผล จึงสามารถเขียนเป็นกรอบแนวคิดการวิจัย ตามภาพที่ 1



ภาพที่ 1 กรอบแนวคิดการวิจัย เรื่อง การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ

สมมติฐานการวิจัย

1. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านภาษา วิธี HGLM ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันมากกว่าวิธี MIMIC

2. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านภาษา วิธี HGLM ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันมากกว่าวิธี IRT-LR

3. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านภาษา วิธี IRT-LR ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันมากกว่าวิธี MIMIC

4. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านคำนวณ วิธี HGLM ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันมากกว่าวิธี MIMIC

5. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านคำนวณ วิธี HGLM ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันมากกว่าวิธี IRT-LR

6. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านคำนวณ วิธี IRT-LR ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันมากกว่าวิธี MIMIC

7. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านเหตุผล วิธี HGLM ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันมากกว่าวิธี MIMIC

8. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านเหตุผล วิธี HGLM ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันมากกว่าวิธี IRT-LR

9. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านเหตุผล วิธี IRT-LR ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันมากกว่าวิธี MIMIC

วิธีดำเนินการวิจัย

การวิจัยนี้ใช้ข้อมูลทุติยภูมิ (Secondary data) จากสำนักทดสอบทางการศึกษาของสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน (สพฐ.) เป็นผลการตอบแบบทดสอบระดับชาติ (NT) ของนักเรียนชั้นประถมศึกษาปีที่ 3 และทำการสุ่มกลุ่มตัวอย่างแบบแบ่งชั้นภูมิ (Stratified random sampling) โดยการจำแนกประชากรออกตามคุณลักษณะเฉพาะ สุ่มแบบแบ่งกลุ่ม (Cluster random sampling) เพื่อจำแนกภูมิภาค จังหวัดและเพศ จากการศึกษางานวิจัยของพีรญา สูงเนิน, เสรี ชัดแช้ม และสมโภชน์ อเนกสุข (2552) พบว่า เมื่อใช้กลุ่มตัวอย่างขนาดใหญ่ จำนวน 1,000 คนขึ้นไป จะมีประสิทธิภาพในการตรวจพบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบได้มากขึ้น จึงสุ่มอย่างง่าย (Simple random sampling) เป็นเพศชาย 4,800 คน และเพศหญิง 4,800 คน รวมทั้งหมดเป็น 9,600 คน

วิธีดำเนินการวิจัย แบ่งเป็น 3 ระยะ ดังนี้

ระยะที่ 1 การวิเคราะห์คุณภาพของแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล โดยใช้ทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ ดังนี้

1. รวบรวมผลการตอบแบบทดสอบระดับชาติ ปีการศึกษา 2556 จำนวน 3 ด้าน ประกอบด้วย 1) ด้านภาษา 2) ด้านคำนวณ และ 3) ด้านเหตุผล จากสำนักทดสอบทางการศึกษา ของสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน (สพฐ.)

2. ตรวจสอบความสมบูรณ์ของผลการตอบแบบทดสอบระดับชาติ ปีการศึกษา 2556 ทั้งข้อคำถามตัวเลือก และเฉลยคำตอบที่ถูกต้อง รวมทั้งตรวจสอบความสมบูรณ์ของคำตอบที่ผู้สอบตอบ

3. วิเคราะห์คุณภาพของแบบทดสอบระดับชาติตามทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ โดยใช้โปรแกรมคอมพิวเตอร์สำเร็จรูป Xcalibre Version 4.1

ระยะที่ 2 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 จำนวน 3 ด้าน คือ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และ วิธี IRT-LR จำแนกตามเพศ ดังนี้

1. วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธี HGLM จำแนกตามตัวแปรเพศ โดยใช้โปรแกรม HLM แบ่งเป็น 2 ระดับ ดังนี้

1.1 การวิเคราะห์ระดับที่ 1 คือ ระดับข้อสอบ

การวิเคราะห์ระดับข้อสอบ ใช้การกำหนดให้ข้อสอบสอดแทรกอยู่ในตัวบุคคล ผลการวิเคราะห์ระดับนี้ แสดงค่าความยากของข้อสอบ (b) ซึ่งสามารถเขียนสมการได้ ดังนี้ (Kamata, 2001)

สมการโมเดลการวิเคราะห์ระดับที่ 1 คือ ระดับข้อสอบ

$$\eta_{ij} = \beta_{0j} + \beta_{1j}x_{1j} + \beta_{2j}x_{2j} + ... + \beta_{29j}x_{29j} \quad (1)$$

เมื่อ η_{ij} แทน ค่าลอคของออดส์ (Odds) ที่จะตอบข้อสอบข้อที่ i ได้ถูกต้องของผู้สอบคนที่ j จึงสามารถเขียนเป็นสมการโครงสร้างของระดับการวิเคราะห์ระดับที่ 1

ได้ โดย β_{0j} แทน ค่าความสามารถของผู้สอบคนที่ j และ x_{29j} แทน ตัวแปรดัมมี่ที่ 29 สำหรับบุคคลที่ j ซึ่งสามารถพิจารณาได้ว่า x_{29j} เป็นตัวแปรอิทธิพลของรายข้อ

1.2 การวิเคราะห์ระดับที่ 2 คือ ระดับผู้สอบ

การวิเคราะห์ระดับผู้สอบ ใช้การกำหนดให้ผู้สอบแต่ละคนสอดแทรกในแต่ละเพศ ผลการวิเคราะห์ระดับนี้ แสดงค่าความยากของข้อสอบ (b) และค่าความสามารถของผู้สอบ (θ) ในสมการระดับผู้สอบซึ่งสามารถเขียนสมการได้ ดังนี้

สมการโมเดลการวิเคราะห์ระดับที่ 2 คือ ระดับผู้สอบ จำแนกตามเพศ

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Gender + u_{ojm} \quad (2)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}Gender \quad (3)$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}Gender \quad (4)$$

$$\vdots$$

$$\beta_{29j} = \gamma_{290} + \gamma_{291}Gender \quad (5)$$

เมื่อ γ_{00} เป็นค่า Intercept ของ β_{0j} คือ เป็นค่าเฉลี่ยอิทธิพลของข้อสอบข้ออ้างอิงต่อโอกาสในการตอบข้อสอบถูกในโรงเรียนที่ m โดย u_{ojm} เป็นค่าส่วนที่เหลือของ β_{0j} คือ เป็นค่าส่วนเบี่ยงเบนของโอกาสในการตอบข้อสอบถูกต้องคนที่ j จากค่าเฉลี่ยโอกาสในการตอบข้อสอบถูกในโรงเรียนที่ m

2. วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธี MIMIC จำแนกตามตัวแปรเพศ โดยใช้โปรแกรม Mplus ดังนี้

2.1 การวิเคราะห์องค์ประกอบเชิงยืนยัน (Confirmatory Factor Analysis: CFA) โดยปราศจากตัวแปรทำนาย (X)

2.2 การวิเคราะห์โมเดล CFA และเพิ่มตัวแปรทำนาย (X) โดยไม่มีอิทธิพลทางตรง (Direct effect) ต่อข้อสอบ

2.3 การเพิ่มอิทธิพลทางตรงต่อข้อสอบและบังคับให้มีค่าเป็น 0 เพื่อกำหนดให้ตัวแปรต้นไม่มีอิทธิพลทางตรงต่อข้อสอบ (Y1 on X@0)

2.4 ตรวจสอบค่าดัชนีปรับแก้ (Modification indices) ว่า ข้อใดมีค่าดัชนีปรับแก้สูงที่สุด

2.5 เพิ่มอิทธิพลทางตรงจากตัวแปรทำนายไปที่ข้อสอบ ในข้อสอบที่มีดัชนีปรับแก้สูงที่สุดแล้ววิเคราะห์โมเดลที่ปรับแก้อีกครั้ง

2.6 ดำเนินการวิเคราะห์ซ้ำอีกครั้ง ในข้อที่ 2.4 และ 2.5 จนไม่พบดัชนีปรับแก้ที่มีนัยสำคัญทางสถิติ

2.7 ตรวจสอบดัชนีตรวจสอบความกลมกลืนของโมเดล และตรวจสอบอิทธิพลทางตรงที่มีนัยสำคัญทางสถิติถ้าพบข้อสอบที่มีอิทธิพลทางตรง และมีนัยสำคัญทางสถิติที่ระดับ .05 แสดงว่า ข้อสอบข้อนั้นเป็นข้อสอบที่เกิดการทำหน้าที่ต่างกัน

3. วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธี IRT-LR จำแนกตามตัวแปรเพศ โดยใช้โปรแกรม IRTPRO ดังนี้

3.1 เลือกวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบผลการสอบตามทฤษฎีการตอบสนองข้อสอบแบบเอกมิติ (Unidimensional IRT)

3.2 กำหนดตัวแปรเพศ ความสามารถด้านภาษาและด้านเหตุผล ให้เพศหญิงเป็นกลุ่มอ้างอิง (Reference group: R) และเพศชายเป็นกลุ่มเปรียบเทียบ (Focal group: F) ส่วนผลการสอบความสามารถด้านคำนวณ ให้เพศชายเป็นกลุ่มอ้างอิง (R) และเพศหญิงเป็นกลุ่มเปรียบเทียบ (F)

3.3 เลือกโมเดลสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เพื่อวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ โดยกำหนดให้วิเคราะห์ข้อสอบทุกข้อ จากนั้นพิจารณาว่าข้อสอบข้อใดมี DIF โดยดูจาก ค่า *p*-value ที่มีนัยสำคัญทางสถิติที่ระดับ .05

ระยะที่ 3 การเปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 จำนวน 3 ด้าน คือ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR

ผลการวิจัย

1. ผลการวิเคราะห์คุณภาพของแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้าน คือ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ตามหลักการทฤษฎีตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ แสดงดังตารางที่ 1

ตารางที่ 1 ผลการวิเคราะห์คุณภาพของแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล

แบบทดสอบ	ค่าอำนาจจำแนกของข้อสอบ (a)	ค่าความยากของข้อสอบ (b)	ค่าโอกาสการเดาของข้อสอบ (c)	ค่าความเที่ยงทั้งฉบับ
ด้านภาษา	0.228 – 1.330	-1.404 – 2.306	0.158 – 0.269	0.746
ด้านคำนวณ	0.408 – 1.089	-0.147 – 2.737	0.095 – 0.282	0.764
ด้านเหตุผล	0.394 – 1.360	-0.444 – 3.292	0.154 – 0.260	0.774

จากตารางที่ 1 ปรากฏว่า แบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 มีค่าความยากของข้อสอบ (b) อยู่ในระดับค่อนข้างยาก มีค่าอำนาจจำแนกของข้อสอบ (a) อยู่ในระดับที่สามารถจำแนกผู้สอบได้ดี มีค่าโอกาสการเดาของข้อสอบ (c) ไม่เกิน 0.3 และค่าความเที่ยงของแบบทดสอบทั้งฉบับ (Reliability) โดยใช้วิธีการหาค่าสัมประสิทธิ์แอลฟาของครอนบราค (Cronbach's alpha coefficient) อยู่ในระดับดี

2. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และ วิธี IRT-LR จำแนกตามเพศ แสดงดังตารางที่ 2

ตารางที่ 2 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และ วิธี IRT-LR จำแนกตามเพศ

แบบทดสอบ	จำนวนข้อสอบ (ข้อ)	วิธี HGLM		วิธี MIMIC		วิธี IRT-LR	
		จำนวนข้อที่พบ DIF	ร้อยละ	จำนวนข้อที่พบ DIF	ร้อยละ	จำนวนข้อที่พบ DIF	ร้อยละ
ด้านภาษา	30	27	90	6	20	16	53
ด้านคำนวณ	30	13	43	2	7	15	50
ด้านเหตุผล	30	22	73	6	20	18	60
รวมทั้ง 3 ด้าน	90	62	69	14	16	49	54

จากตารางที่ 2 แสดงผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ปรากฏว่า วิธี HGLM ตรวจพบ DIF จำนวน 62 ข้อ คิดเป็นร้อยละ 69 ของข้อสอบทั้งฉบับ รองลงมาคือ วิธี IRT-LR ตรวจพบ DIF จำนวน 49 ข้อ คิดเป็นร้อยละ 54 และวิธี MIMIC ตรวจพบ DIF จำนวน 14 ข้อ คิดเป็นร้อยละ 16 ตามลำดับ

2.1 วิธี HGLM เมื่อพิจารณาด้านภาษา ตรวจพบ DIF จำนวน 27 ข้อ คิดเป็นร้อยละ 90 ของข้อสอบทั้งหมด ด้านคำนวณ ตรวจพบ DIF จำนวน 13 ข้อ คิดเป็นร้อยละ 43 และด้านเหตุผล ตรวจพบ DIF จำนวน 22 ข้อ คิดเป็นร้อยละ 73

2.2 วิธี MIMIC เมื่อพิจารณาด้านภาษา ตรวจพบ DIF จำนวน 6 ข้อ คิดเป็นร้อยละ 20 ของข้อสอบทั้งหมด ด้านคำนวณ ตรวจพบ DIF จำนวน 2 ข้อ คิดเป็นร้อยละ 7 และด้านเหตุผล ตรวจพบ DIF จำนวน 6 ข้อ คิดเป็นร้อยละ 20

2.3 วิธี IRT-LR เมื่อพิจารณาด้านภาษา ตรวจพบ DIF จำนวน 16 ข้อ คิดเป็นร้อยละ 53 ของข้อสอบทั้งหมด ด้านคำนวณ ตรวจพบ DIF จำนวน 15 ข้อ คิดเป็นร้อยละ 50 และด้านเหตุผล ตรวจพบ DIF จำนวน 18 ข้อ คิดเป็นร้อยละ 60

3. ผลการเปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้าน คือ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และ วิธี IRT-LR แสดงดังตารางที่ 3

ตารางที่ 3 ผลการเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธีวิเคราะห์ 3 วิธี

แบบทดสอบ	เปรียบเทียบร้อยละของการตรวจพบ DIF		
	วิธี HGLM กับ วิธี MIMIC	วิธี HGLM กับ วิธี IRT-LR	วิธี IRT-LR กับ วิธี MIMIC
ด้านภาษา	วิธี HGLM > วิธี MIMIC (70%)*	วิธี HGLM > วิธี IRT-LR (37%)*	วิธี IRT-LR > วิธี MIMIC (33%)*
ด้านคำนวณ	วิธี HGLM > วิธี MIMIC (36%)*	วิธี HGLM < วิธี IRT-LR (7%)*	วิธี IRT-LR > วิธี MIMIC (43%)*
ด้านเหตุผล	วิธี HGLM > วิธี MIMIC (53%)*	วิธี HGLM > วิธี IRT-LR (13%)*	วิธี IRT-LR > วิธี MIMIC (40%)*

หมายเหตุ *p < .05

วิธี HGLM > *วิธี* MIMIC *หมายถึง วิธี* HGLM *ตรวจพบ* DIF *มากกว่าวิธี* MIMIC วิธี HGLM > วิธี IRT-LR *หมายถึง วิธี* HGLM *ตรวจพบ* DIF *มากกว่าวิธี* IRT-LR
วิธี HGLM < วิธี IRT-LR *หมายถึง วิธี* HGLM *ตรวจพบ* DIF *น้อยกว่าวิธี* IRT-LR วิธี IRT-LR > วิธี MIMIC *หมายถึง วิธี* IRT-LR *ตรวจพบ* DIF *มากกว่าวิธี* MIMIC

จากตารางที่ 3 แสดงให้เห็นว่า ด้านภาษา วิธี HGLM ตรวจพบ DIF มากกว่าวิธี MIMIC และวิธี IRT-LR คิดเป็นร้อยละ 70 และ 37 ตามลำดับ ส่วนวิธี IRT-LR ตรวจพบ DIF มากกว่าวิธี MIMIC คิดเป็นร้อยละ 33 ของข้อสอบทั้งหมด ส่วนด้านคำนวณ วิธี HGLM ตรวจพบ DIF มากกว่าวิธี MIMIC คิดเป็นร้อยละ 36 ของข้อสอบทั้งหมด ส่วนวิธี IRT-LR ตรวจพบ DIF มากกว่าวิธี HGLM และวิธี MIMIC คิดเป็นร้อยละ 7 และ 43 ตามลำดับ และด้านเหตุผล วิธี HGLM ตรวจพบ DIF มากกว่าวิธี MIMIC และวิธี IRT-LR คิดเป็นร้อยละ 53 และ 13 ตามลำดับ ส่วนวิธี IRT-LR ตรวจพบ DIF มากกว่าวิธี MIMIC คิดเป็นร้อยละ 40 ของข้อสอบทั้งหมด อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

การอภิปรายผล

1. การวิเคราะห์คุณภาพของแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล

การวิเคราะห์คุณภาพของแบบทดสอบโดยใช้หลักการทฤษฎีการตอบสนองข้อสอบ (IRT) แบบ 3 พารามิเตอร์ ประกอบด้วย ค่าอำนาจจำแนกของข้อสอบ (a) ค่าความยากของข้อสอบ (b) และค่าโอกาสการเดาของข้อสอบ (c) แบบทดสอบระดับชาติด้านภาษา มีค่าอำนาจจำแนกของข้อสอบ (a) ทั้งฉบับอยู่ในระดับค่อนข้างดี มีค่าความยากของข้อสอบ (b) ทั้งฉบับอยู่ในระดับค่อนข้างยาก และมีค่าโอกาสการเดาของข้อสอบ (c) ทั้งฉบับไม่เกิน 0.3 สำหรับด้านคำนวณ มีค่าอำนาจจำแนกของข้อสอบ (a) ทั้งฉบับอยู่ในระดับค่อนข้างดี มีค่าความยากของข้อสอบ (b) ทั้งฉบับอยู่ในระดับยาก และมีค่าโอกาสการเดาของข้อสอบ (c) ทั้งฉบับไม่เกิน 0.3 ส่วนด้านเหตุผล มีค่าอำนาจจำแนกของข้อสอบ (a) ทั้งฉบับอยู่ในระดับค่อนข้างดี มีค่าความยากของข้อสอบ (b) ทั้งฉบับอยู่ในระดับค่อนข้างยาก และมีค่าโอกาสการเดาของข้อสอบ (c) ทั้งฉบับไม่เกิน 0.3

2. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการตรวจสอบ 3 วิธี จำแนกตามเพศ

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ทั้ง 3 วิธี คือ วิธี HGLM วิธี MIMIC และวิธี IRT-LR จำแนกตามเพศ พบว่า วิธี HGLM สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน (DIF) ได้มากที่สุด รองลงมา คือ วิธี IRT-LR และวิธี MIMIC ตามลำดับ เพราะวิธี HGLM ตรวจสอบ DIF ได้ดีในแบบทดสอบที่มีการตรวจให้คะแนนแบบ 2 ค่า ซึ่งจากการศึกษางานวิจัยของ Acar (2013) ที่เปรียบเทียบผลการตรวจสอบ DIF ระหว่างวิธี HGLM และวิธี LR ผลการศึกษาพบว่า วิธี HGLM จะตรวจสอบ DIF ได้ดีกว่าวิธี LR ในแบบทดสอบที่มีการตรวจให้คะแนนแบบ 2 ค่า สอดคล้องกับงานวิจัยของ Ong, Lu, Lee, and Cohen (2015) ที่ได้ตรวจสอบ DIF ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR ผลการศึกษาพบว่า วิธี HGLM ตรวจพบ

DIF ได้มากที่สุด รองลงมา คือ วิธี IRT-LR และวิธี MIMIC และสอดคล้องกับงานวิจัยของ Acar and Kelecioglu (2010) ที่ตรวจสอบ DIF ด้วยวิธี HGLM วิธี LR และวิธี IRT-LR ผลการศึกษาพบว่า วิธี HGLM ตรวจพบ DIF ได้มากที่สุด ส่วนวิธี LR และวิธี IRT-LR ตรวจพบ DIF ได้ใกล้เคียงกันจากการศึกษาของ Kabasakal et al. (2014) ศึกษาประสิทธิภาพในการตรวจสอบ DIF ด้วยวิธี MH วิธี SIBTEST และวิธี IRT-LR ผลการศึกษาพบว่า วิธี IRT-LR จะมีประสิทธิภาพในการตรวจสอบ DIF ได้ดีในแบบทดสอบที่มีความยาวของข้อสอบไม่เกิน 20 ข้อ

3. การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ทั้ง 3 วิธี จำแนกตามเพศ

การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ พบว่า วิธี HGLM และวิธี IRT-LR ตรวจพบ DIF มากกว่าวิธี MIMIC ทั้ง 3 ด้าน คือ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ส่วนวิธี HGLM ตรวจพบ DIF มากกว่าวิธี IRT-LR ในด้านภาษา และด้านเหตุผล สอดคล้องกับงานวิจัยของ Ong et al. (2015) ที่ได้เปรียบเทียบผลการตรวจสอบ DIF ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR ผลการศึกษาพบว่า วิธี HGLM ตรวจพบ DIF ได้มากกว่าวิธี IRT-LR และวิธี MIMIC เมื่อใช้กลุ่มตัวอย่างขนาดใหญ่ (2,000 คน) และสอดคล้องกับงานวิจัยของ Acar and Kelecioglu (2010) ที่เปรียบเทียบผลการตรวจสอบ DIF ด้วยวิธี HGLM วิธี LR และวิธี IRT-LR ในแบบทดสอบด้านสังคมศาสตร์และด้านวิทยาศาสตร์ ผลการศึกษาพบว่า วิธี HGLM ตรวจพบ DIF ได้มากกว่าวิธี LRและวิธี IRT-LR ในแบบทดสอบทั้ง 2 ด้าน และยังสอดคล้องกับผลการศึกษาของ Acar (2013) ที่พบว่า วิธี HGLM มีประสิทธิภาพในการตรวจสอบ DIF ได้ดีกว่าในแบบทดสอบที่มีการตรวจให้คะแนนแบบ 2 ค่า ส่วนวิธี IRT-LR ตรวจพบ DIF มากกว่าวิธี HGLM เพราะวิธี IRT-LR มีประสิทธิภาพในการตรวจสอบ DIF ได้ดีกว่าในแบบทดสอบที่มีเนื้อหาด้านคำนวณ สอดคล้องกับงานวิจัยของ Yildirim and Berberoglu (2009) ที่พบว่า วิธี IRT-LR มีประสิทธิภาพในการตรวจสอบ DIF ได้ดีในด้านความสามารถทางคณิตศาสตร์ของโครงการประเมินผลนักเรียนนานาชาติ (PISA, 2003)

ข้อเสนอแนะในการนำผลการวิจัยไปใช้

1. สำนักทดสอบทางการศึกษาสามารถนำผลการวิเคราะห์คุณภาพของแบบทดสอบระดับชาติที่ผ่านเกณฑ์การวิเคราะห์คุณภาพโดยใช้หลักการของทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ ไปใช้สอบในครั้งต่อไปเพื่อใช้สำหรับวัดความสามารถของนักเรียนชั้นประถมศึกษาปีที่ 3 ของสำนักทดสอบทางการศึกษาสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน (สพฐ.)

2. นักวิจัยและนักวัดผลการศึกษาที่สนใจเกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการตรวจสอบที่อยู่บนพื้นฐานของทฤษฎีการตอบสนองข้อสอบ (IRT) ทั้ง 3 วิธี คือ วิธี HGLM วิธี MIMIC และวิธี IRT-LR เมื่อกลุ่มตัวอย่างมีขนาดใหญ่ (2,000 คน) ควรเลือกใช้ วิธี HGLM ในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน และเมื่อกลุ่มตัวอย่างมีขนาดเล็ก (300 คน) ควรเลือกใช้ วิธี MIMIC และวิธี IRT-LR

ข้อเสนอแนะในการวิจัยต่อไป

1. วิธี HGLM มีประสิทธิภาพในการตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ มากกว่า วิธี MIMIC และวิธี IRT-LR เมื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับข้อสอบที่มีการให้คะแนนแบบ 2 ค่าจึงควรมีการเปรียบเทียบเพิ่มเติมกับวิธีการตรวจสอบอื่น ๆ และศึกษาตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบที่มีการตรวจให้คะแนนแบบมากกว่า 2 ค่า

2. ควรมีการศึกษาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการตรวจให้คะแนนแบบมากกว่า 2 ค่า ด้วยวิธี Standard MIMIC (M-ST) วิธี MIMIC with Scale Purification (M-SP) และวิธี MIMIC with Pure Anchor (M-PA) ว่า วิธีใดมีประสิทธิภาพในการตรวจสอบ DIF มากกว่ากัน

เอกสารอ้างอิง

กระทรวงศึกษาธิการ. (2551). *หลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน พุทธศักราช 2551*. กระทรวงศึกษาธิการ.

พีรญา สูงเนิน, เสรี ชัดแช้ม และสมโภชน์ อเนกสุข (2552). การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ: การเปรียบเทียบระหว่างรายข้อกับรายหมวดข้อสอบ โดยใช้วิธีซิปเทสท์. *วิทยาการวิจัยและวิทยาการปัญญา*, *6*(2), 49-62.

สำนักงานปลัดกระทรวงศึกษาธิการ. (2554). *แผนพัฒนาการศึกษาของกระทรวงศึกษาธิการฉบับที่สิบเอ็ด พ.ศ 2555-2559*. กรุงเทพฯ: กระทรวงศึกษาธิการ.

Acar, T., (2013). Comparison of the group and intercept coefficient from HGLM and LR-DIF method. *British Journal of Science*, *10*(1), 12-20.

Acar, T., & Kelecioglu, H. (2010). Comparison of differential item functioning determination techniques: HGLM, LR and IRT-LR. *Educational Sciences: Theory and Practice*, *10*(2), 639-649.

De Ayala, R. J. (2009). The theory and practice of item response theory. New York: Guilford Press.

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*(4), 278-295.

Kabasakal, K. A., Arsan, N., Gök, B., & Kelecioglu, H. (2014). Comparing performances (Type I Error and Power) of IRT likelihood ratio SIBTEST and Mantel-Haenszel methods in the determination of differential item functioning. *Educational Sciences: Theory and Practice*, *14*(6), 2186-2193.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*(1), 79-93.

Le, L. T. (2009). Investigating gender differential item functioning across countries and test languages for PISA science items. *International Journal of Testing*, *9*(2), 122-133.

Li, H., Hunter, C. V., & Oshima, T. C. (2013). Gender DIF in reading tests: A synthesis of research. *In New Developments in Quantitative Psychology* (pp. 489-506). New York, Springer.

Ong, M. L., Lu, L., Lee, S., & Cohen, A. (2015). A comparison of the hierarchical generalized linear model, multiple-indicators multiple-causes, and the item response theory-likelihood ratio test for detecting differential item functioning. In Mellsap, R. E., Bolt, D. M., Van der Ark, L. A. & Wang, W. C. (Eds.), *Quantitative Psychology Research* (pp. 343-357). DOI:10.1007-978-3-319-07503-7_22.

Taylor, C. S., & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed item formats. *Applied Measurement in Education*, *25*(3), 246-280.

Yildirim, H. H., & Berberoglu, G. (2009). Judgmental and statistical DIF analyses of the PISA-2003 mathematics literacy items. *International Journal of Testing*, *9*(2), 108-121.