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

สุธาทิพย์ ตรีสิน

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต
สาขาวิชาการวัดและเทคโนโลยีทางวิทยาการปัญญา
วิทยาลัยวิทยาการวิจัยและวิทยาการปัญญา มหาวิทยาลัยบูรพา
สิงหาคม 2560
ลิขสิทธิ์ของมหาวิทยาลัยบูรพา

คณะกรรมการควบคุมวิทยานิพนธ์และคณะกรรมการสอบวิทยานิพนธ์ ได้พิจารณาวิทยานิพนธ์ของ สุธาทิพย์ ตรีสิน ฉบับนี้แล้ว เห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาการวัดและเทคโนโลยีทางวิทยาการปัญญา ของมหาวิทยาลัยบูรพาได้

คณะกรรมการควบคุมวิทยานิพนธ์

..... อาจารย์ที่ปรึกษาหลัก
(ดร.ปิยะทิพย์ ประดุจพรม)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธาน
(รองศาสตราจารย์ ดร.เสรี ชัดแช้ม)

..... กรรมการ
(ดร.ปิยะทิพย์ ประดุจพรม)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ภัทราวดี มากมี)

..... กรรมการ
(ดร.กนก พานทอง)

วิทยาลัยวิทยาการวิจัยและวิทยาการปัญญาอนุมัติให้รับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาการวัดและเทคโนโลยีทางวิทยาการปัญญาของมหาวิทยาลัยบูรพา

..... คณบดีวิทยาลัยวิทยาการวิจัย
(ผู้ช่วยศาสตราจารย์ ดร.สุชาดา กรเพชรปาณี) และวิทยาการปัญญา
วันที่ 8 เดือน สิงหาคม พ.ศ. 2560

ประกาศคุณูปการ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยความกรุณาจาก ดร.ปิยะทิพย์ ประดุจพรม อาจารย์ที่ปรึกษาหลัก ที่ให้คำปรึกษา แนะนำ ช่วยเหลือ และตรวจสอบแก้ไขข้อบกพร่องต่าง ๆ ของวิทยานิพนธ์ด้วยความเอาใจใส่เป็นอย่างดีตลอดมา ผู้วิจัยรู้สึกซาบซึ้งเป็นอย่างมาก และขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ โอกาสนี้

ขอขอบคุณ รองศาสตราจารย์ ดร.เสรี ชัดแช้ม ผู้ช่วยศาสตราจารย์ ดร.ภัทราวดี มากมี ดร.กนก พานทอง ดร.วรวุฒิ เพ็งพันธ์ และอาจารย์ประจำวิทยาลัยวิทยาการวิจัยและวิทยาการปัญญา และผู้เชี่ยวชาญทุกท่าน ที่ให้คำแนะนำ คำปรึกษาในการดำเนินงานวิจัย วิเคราะห์ข้อมูล และช่วยตรวจสอบความถูกต้องในการวิเคราะห์ข้อมูล ที่ขาดไม่ได้ ขอขอบพระคุณบิดามารดา ที่สนับสนุนทุนการศึกษา อบรมสั่งสอน รวมถึงการให้กำลังใจ ทำให้ผู้วิจัยสามารถทำสิ่งที่ฝันสำเร็จลุล่วงโดยสมบูรณ์

นอกจากนี้ ผู้วิจัยขอขอบคุณ ครอบครัว พี่น้อง และพ้องเพื่อนทุกท่าน ที่มิได้กล่าวนามมา ณ ที่นี้ ซึ่งมีส่วนช่วยให้การวิจัยครั้งนี้สำเร็จได้โดยสมบูรณ์ คุณค่าและประโยชน์ของวิทยานิพนธ์ฉบับนี้ ผู้วิจัยขอมอบเป็นกตัญญูกตเวทิตาแด่ บุพการี บูรพาจารย์ และผู้มีพระคุณทุกท่านทั้งในอดีตและปัจจุบัน ที่ทำให้ผู้วิจัยเป็นผู้มีการศึกษา และประสบความสำเร็จมาจนตราบเท่าทุกวันนี้

สุธาทิพย์ ตรีสิน

55910392: สาขาวิชา: การวัดและเทคโนโลยีทางวิทยาการปัญญา;
วท.ม. (การวัดและเทคโนโลยีทางวิทยาการปัญญา)
คำสำคัญ: การทำหน้าที่ต่างกันของข้อสอบ/ วิธี HGLM/ วิธี MIMIC/ วิธี IRT-LR/
แบบทดสอบระดับชาติ

สุธาทิพย์ ตรีสิน: การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ระดับชั้นประถมศึกษาปีที่ 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) คณะกรรมการควบคุมวิทยานิพนธ์: ปิยะทิพย์ ประดุจพรม, Ph.D., 148 หน้า. ปี พ.ศ. 2560.

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 MIMICและวิธี IRT-LR และ 3) เปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธีการตรวจสอบ 3 วิธีข้อมูลที่นำมาใช้วิเคราะห์เป็นข้อมูลทุติยภูมิ จากผลการตอบแบบทดสอบระดับชาติของนักเรียนชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2556 จำนวน 9,600 คน

ผลการวิจัยปรากฏว่า

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

2) การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้ง 3 ด้าน ชี้ให้เห็นว่า เพศส่งผลให้เกิดการทำหน้าที่ต่างกันของข้อสอบ โดยเพศหญิงจะได้เปรียบในการตอบข้อสอบด้านภาษาและด้านเหตุผล ในขณะที่เพศชายจะได้เปรียบในการตอบข้อสอบด้านคำนวณ โดยวิธี HGLM ตรวจพบข้อสอบทำหน้าที่ต่างกัน จำนวนมากที่สุดคิดเป็นร้อยละ 69 ของข้อสอบทั้งฉบับ รองลงมาคือวิธี IRT-LR ร้อยละ 54 และวิธี MIMIC ร้อยละ 16 ตามลำดับ

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

55910392: MAJOR: MEASUREMENT AND TECHNOLOGY IN COGNITIVE SCIENCE
M.Sc. (MEASUREMENT AND TECHNOLOGY IN COGNITIVE SCIENCE)
KETWORDS: DIFFERENTIAL ITEM FUNCTIONING: DIF/ HIERARCHICAL GERNERALIZRD LINEAR MODELING: HGLM/ MULTIPLE-INDICATORS MULTIPLE-CAUSES: MIMIC/ ITEM RESPONSE THEORY – LIKELIHOOD RATIO: IRT-LR/ NATIONAL TESTS: NT

SUTHATHIP TREESIN: 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. ADVISORY COMMITTEE: PIYATHIP PRADUJPROM, Ph.D. 148 P. 2017.

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; 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%) ($p < .05$).

สารบัญ

สารบัญ (ต่อ)

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญตาราง (ต่อ)

สารบัญภาพ

บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

กระทรวงศึกษาธิการมีนโยบายพัฒนาการศึกษา ว่าด้วยการที่ประเทศไทยกำลังเข้าสู่สังคมขนาดใหญ่ ทั้งระดับภูมิภาคและสังคมโลก และเมื่อประเทศไทยมีการแข่งขันทางเศรษฐกิจเพื่อที่จะให้สามารถอยู่รอดในสถานการณ์ทั้งในปัจจุบันและอนาคตต่อไป รวมไปถึงการที่จะรวมตัวกันเป็นประชาคมอาเซียน คือ การเพิ่มความสามารถในการแข่งขันของคนและประเทศ จึงมีความสำคัญอย่างมากที่การศึกษาจึงต้องเดินหน้าสู่การพัฒนาและเตรียมความพร้อมให้สอดคล้องกับสังคมโลก ด้วยเหตุนี้รัฐบาลจึงมีนโยบายทางด้านการศึกษาให้ทุกหน่วยงานที่เกี่ยวข้อง รวมถึงสังคมต้องให้ความตระหนักทางด้านการศึกษา เพื่อช่วยกันยกระดับคุณภาพการศึกษา โดยมีเป้าหมายที่มุ่งพัฒนาผู้เรียนให้มีความสามารถในการคิด วิเคราะห์ เรียนรู้ด้วยตนเอง มีคุณลักษณะที่พึงประสงค์ พร้อมทั้งมีทักษะที่จำเป็นสำหรับศตวรรษที่ 21 (กระทรวงศึกษาธิการ, 2556, หน้า 1) ในการประเมินทักษะแห่งศตวรรษที่ 21 มีโครงการประเมินผลนักเรียนนานาชาติ (Programme for International Assessment: PISA) ซึ่งมีเกณฑ์การประเมินด้านทักษะการอ่าน คณิตศาสตร์ และวิทยาศาสตร์ โดยมุ่งหวังที่จะวัดความสามารถและทักษะของเด็กที่มีอายุ 15 ปี ว่ามีความพร้อมในการนำความรู้และทักษะไปใช้ได้ดีเพียงไร PISA จึงเรียกความพร้อมนี้ว่า “ความรู้พื้นฐาน” (Literacy) (วรพจน์ วงศ์กิจรุ่งเรือง และอธิป จิตตฤกษ์, 2556, หน้า 7) ซึ่งในหลายปีที่ผ่านมาคุณภาพการศึกษาของไทยยังไม่ดีเท่าที่ควร ผลสัมฤทธิ์ทางการเรียนในวิชาหลักของการศึกษาขั้นพื้นฐาน ยังมีค่าเฉลี่ยต่ำกว่าร้อยละ 50 โดยผลการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (Ordinary National Educational Testing: O-NET) ปี พ.ศ. 2553 พบว่าคะแนนเฉลี่ยของวิชาภาษาอังกฤษ และคณิตศาสตร์ลดลง พร้อมทั้งมาตราฐานความสามารถยังได้คะแนนต่ำในเรื่องของการคิดวิเคราะห์ สังเคราะห์ และความคิดสร้างสรรค์ (กระทรวงศึกษาธิการ, 2554, หน้า 6) โดยผลการประเมินการจัดอันดับความสามารถจาก International Institute for Management Development: IMD พบว่า ในปี พ.ศ. 2556 การศึกษาไทยอยู่อันดับที่ 51 จาก 60 ประเทศที่เข้าร่วมประเมินเพื่อจัดอันดับนี้

กระทรวงศึกษาธิการจึงมีแผนการพัฒนาการศึกษาของประเทศไทยให้มีประสิทธิภาพมากยิ่งขึ้น เพื่อให้สอดคล้องกับรัฐธรรมนูญแห่งราชอาณาจักรไทย พุธศักราช 2550 และความต้องการของท้องถิ่น ซึ่งกำหนดเป้าหมายหลักเน้นไปที่ผู้เรียนให้ได้รับการศึกษามีคุณภาพและสถาบันการศึกษาทุกระดับทุกประเภทจะต้องผ่านการรับรองมาตราฐานทางการศึกษา (แผนการพัฒนาการศึกษา กระทรวงศึกษาธิการ, 2554, หน้า 14) หลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน พุทธศักราช 2551 จึงมีจุดหมายเพื่อมุ่งพัฒนาผู้เรียนให้เป็นคนดี มีปัญญา มีศักยภาพในการศึกษาต่อ ประกอบอาชีพ และมีคุณลักษณะอันพึงประสงค์ เพื่อให้สามารถอยู่ร่วมกับผู้อื่นในสังคมได้อย่างมีความสุข จึงกำหนดมาตรฐานการเรียนรู้เพื่อพัฒนาผู้เรียนให้เกิดความสมดุล โดยคำนึงถึงหลักพัฒนาการทางสมองและพหุปัญญา จึงให้ผู้เรียนเรียนรู้ 8 กลุ่มสาระการเรียนรู้ ประกอบไปด้วย ภาษาไทย คณิตศาสตร์ วิทยาศาสตร์ สังคมศึกษา ศาสนา และวัฒนธรรม สุขศึกษาและพลศึกษา ศิลปะ การงานอาชีพและ

เทคโนโลยี และภาษาต่างประเทศ ซึ่งแต่ละกลุ่มสาระการเรียนรู้ ได้กำหนดมาตรฐานการเรียนรู้เป็นเป้าหมายสำคัญของการพัฒนาคุณภาพของผู้เรียน เนื่องจากมาตรฐานการเรียนรู้เป็นกลไกสำคัญในการพัฒนาการศึกษา และยังเป็นเครื่องมือในการตรวจสอบเพื่อประกันคุณภาพการศึกษา โดยใช้ประเมินคุณภาพภายในและภายนอก รวมถึงการทดสอบระดับพื้นที่การศึกษา และการทดสอบระดับชาติ ซึ่งระบบการตรวจสอบเพื่อประกันคุณภาพดังกล่าวเป็นสิ่งสำคัญที่ช่วยสะท้อนภาพการจัดการศึกษาว่าสามารถพัฒนาผู้เรียนให้มีคุณภาพตามที่มาตรฐานการเรียนรู้กำหนดเพียงใด (กระทรวงศึกษาธิการ, 2551)

สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน (สพฐ.) จึงกำหนดให้มีการทดสอบระดับชาติ (National Testing: NT) เป็นการทดสอบเพื่อวัดความรู้ของนักเรียนในแต่ระดับชั้น ว่ามีความรู้ความสามารถในระดับใด เพื่อนำไปวิเคราะห์สภาพปัญหาการจัดการเรียนการสอน ซึ่งในระดับชั้นประถมศึกษาปีที่ 3 จะเน้นการประเมิน 3 ด้าน คือ ด้านภาษา (Literacy) ด้านคำนวณ (Numeracy) และด้านเหตุผล (Reasoning Abilities) โดยการบูรณาการเนื้อหาทั้ง 8 กลุ่มสาระไว้ด้วยกัน การศึกษาในศตวรรษที่ 21 นั้นได้ผสมผสานองค์ความรู้ ทักษะเฉพาะด้าน ความชำนาญการและความรู้เท่าทันด้านต่าง ๆ เข้าด้วยกัน เพื่อให้ประสบความสำเร็จทั้งในด้านการงานและการดำเนินชีวิต และผลของการประเมินคุณภาพผู้เรียนระดับชาติปีการศึกษา 2555 พบว่า ผู้เรียนระดับชั้นประถมศึกษาปีที่ 3 มีคะแนนเฉลี่ยสูงทางความสามารถด้านเหตุผล และหากจำแนกตามภูมิภาค สถานที่ตั้งของโรงเรียน พบว่า ภูมิภาคที่มีคะแนนเฉลี่ยสูงหรือเท่ากับค่าเฉลี่ยของประเทศ คือ ภาคตะวันออกเฉียงเหนือ

การศึกษาศตวรรษที่ 21 มีนโยบายในการพัฒนาระบบการทดสอบ วัดและประเมินผล เพื่อให้เป็นเครื่องมือที่ช่วยส่งเสริมการปฏิรูปการเรียนรู้และการพัฒนาคุณภาพของผู้เรียนให้มีมาตราฐานเทียบเท่านานาชาติ ปัจจุบันมีวิธีตรวจสอบคุณภาพของเครื่องมือหลากหลายวิธีซึ่งวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) จึงเป็นหนึ่งวิธีที่นักวัดผลการศึกษาใช้ในการตรวจสอบคุณภาพของข้อสอบ การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เป็นการเปรียบเทียบผลการตอบข้อสอบระหว่างผู้สอบ 2 กลุ่ม ที่มีความสามารถในระดับเดียวกัน ประกอบด้วย 1) กลุ่มเปรียบเทียบ (Focal Group: F) เป็นกลุ่มผู้สอบที่สนใจศึกษาและคาดว่าจะเสียเปรียบในการตอบข้อสอบ คือ เป็นกลุ่มผู้สอบที่มีโอกาสในการตอบข้อสอบได้ถูกน้อยกว่ากลุ่มอ้างอิง 2) กลุ่มอ้างอิง (Referent Group: R) เป็นกลุ่มผู้สอบที่คาดว่าจะได้เปรียบในการตอบข้อสอบ คือ เป็นกลุ่มผู้สอบที่มีโอกาสในการตอบข้อสอบได้ถูกมากกว่ากลุ่มเปรียบเทียบ ซึ่งการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นการขจัดข้อสอบที่จะก่อให้เกิดความไม่เท่าเทียมกันในแบบทดสอบ เมื่อผู้สอบมีคุณลักษณะแตกต่างกัน จะก่อให้เกิดความไม่เท่าเทียมในการตอบแบบทดสอบฉบับนั้น เช่น เพศ ภาษา หรือภูมิลำเนา (De Ayala, 2009, p. 325) ผู้สอบดังกล่าวอาจไม่ได้รับความยุติธรรมในการทำข้อสอบ โดยข้อสอบบางข้อมีความลำเอียงเข้าข้างกลุ่มผู้สอบย่อยบางกลุ่ม ซึ่งอาจเกิดการได้เปรียบเสียเปรียบ ทั้ง ๆ ที่ทำข้อสอบเดียวกัน (Holland & Wainer, 1993, pp. 103-105) ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ พบว่า ข้อสอบสามารถทำหน้าที่แตกต่างกันได้ 2 ประเภท (Mellenbergh, 1982) ได้แก่ การทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป (Uniform) และแบบอเนกรูป (Non-Uniform) จากการศึกษางานวิจัยพบว่า แบบทดสอบหรือข้อคำถามที่มุ่งหวังคำตอบเป็นการตีความ หรือวิเคราะห์ มีแนวโน้มที่ลำเอียงเข้าทางเพศหญิง เช่นเดียวกับ

ข้อสอบที่เกี่ยวกับการแก้ปัญหา หรือเรขาคณิต (Taylor & Lee, 2012) สอดคล้องกับ Mendes-Barnett and Ercikan (2006) พบว่า เพศชายจะมีความถนัดด้านการแก้ปัญหาที่ซับซ้อนมากกว่าเพศหญิงในขณะที่เพศหญิงจะมีความถนัดด้านการคำนวณสมการมากกว่าเพศชาย

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมีหลากหลายวิธี เช่น การวิเคราะห์ความแปรปรวน (Analysis of Variance: ANOVA) วิธีการวิเคราะห์การถดถอยโลจิสติก (Logistic regression: LR) วิธีแปลงค่าความยากของข้อสอบ (Transformed Item Difficulty: TID) วิธีแมนเทล-แฮนส์เซล (Mantel-Haenszel: MH) วิธีวัดพื้นที่ความแตกต่างระหว่างโค้งการตอบสนองข้อสอบ (Item Response Theory: IRT) วิธีไคสแควร์ของลอร์ด (Lord's Chi-square (χ^2) วิธีอัตราส่วนไลด์ลิฮูด ลอกลิเนียร์ (Loglinear Likelihood Ratio) และวิธีซิปเทสท์ (SIBTEST) (ศิริชัย กาญจนวาสี, 2555, หน้า 125) จากการศึกษางานวิจัยที่เกี่ยวข้อข้อง พบว่า Acar and Kelecioglu (2010) ศึกษาเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ในแบบทดสอบสังคมศาสตร์และวิทยาศาสตร์ ระหว่างวิธี Hierarchical Generalized Linear Modeling (HGLM) วิธี Logistics Regression (LR) และ วิธี Item Response Theory – Likelihood Ratio (IRT-LR) พบว่า ทั้ง 3 วิธีตรวจพบข้อสอบที่ทำหน้าที่ต่างกันในจำนวนที่ใกล้เคียงกัน แต่วิธี HGLM จะเป็นวิธีที่สามารถตรวจสอบพบข้อสอบที่คาดว่าจะเกิดการทำหน้าที่ต่างกันมากที่สุดในแบบทดสอบทั้งสองด้าน และ Finch (2005) ศึกษาเปรียบเทียบความสามารถของตัวแบบหลายตัวบ่งชี้หลายสาเหตุ Multiple-Indicators Multiple-Causes (MIMIC) รูปแบบการวิเคราะห์ปัจจัยยืนยันเพื่อระบุการทำหน้าที่ต่างกันของข้อสอบ พบว่า วิธี MIMIC มีประสิทธิภาพในการตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบที่มีความยาวของข้อสอบ จำนวน 50 ข้อขึ้นไป ในขณะที่ Li, Hunter, and Oshima (2013) ศึกษาการตรวจสอบการทำหน้าที่ต่างกันในการทดสอบ การอ่าน และเหตุผล ที่เป็นไปได้ระหว่างเพศด้วยวิธี IRT-LR และ วิธี Mantel-Haenszel (MH) พบว่า วิธี IRT-LR สามารถตรวจสอบการทำหน้าที่ต่างกันได้ดีในแบบทดสอบที่มีความยาวตั้งแต่ 20 ข้อขึ้นไป และเพศมีผลต่อการทำหน้าที่ต่างกันของข้อสอบในด้านการอ่านและเหตุผล

จากการศึกษางานวิจัยที่เกี่ยวข้อง ผู้วิจัยสนใจศึกษาเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ปีการศึกษา 2556 ที่ทดสอบความสามารถทั้ง 3 ด้าน ประกอบด้วย ด้านภาษา ด้านคำนวณ และด้านเหตุผล ว่ามีข้อสอบข้อใดบ้างที่ทำหน้าที่ต่างกัน โดยการประยุกต์ใช้วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ 3 วิธี คือ วิธีการตรวจสอบการตรวจสอบการทำหน้าที่ต่างกันด้วยโมเดลสมการเชิงเส้นตรงระดับลดหลั่น (HGLM) วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยโมเดลสมการโครงสร้างมิมิค (MIMIC) และวิธีการตรวจสอบการหน้าที่ต่างกันของข้อสอบด้วยอัตราส่วนไลด์ลิฮูด (IRT-LR)

วัตถุประสงค์ของการวิจัย

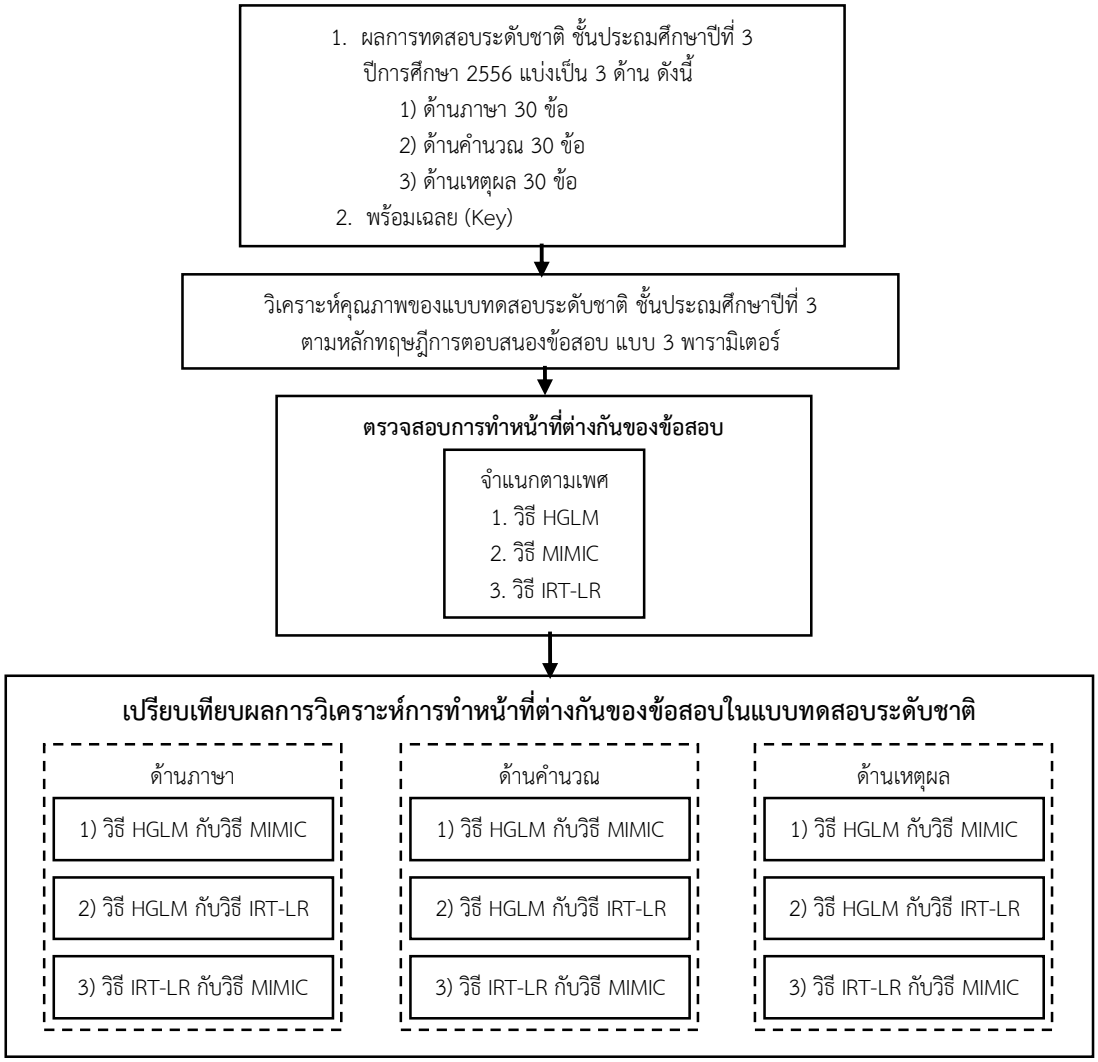
1. เพื่อวิเคราะห์คุณภาพของแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล โดยใช้ทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์

2. เพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR จำแนกตามเพศ

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

กรอบแนวคิดการวิจัย

ผู้วิจัยได้วิเคราะห์และสังเคราะห์งานวิจัยที่เกี่ยวข้องต่าง ๆ เพื่อกำหนดเป็นกรอบแนวคิดการวิจัยครั้งนี้ได้ ดังภาพที่ 1-1



ภาพที่ 1-1 กรอบแนวคิดการวิจัย

สมมติฐานของการวิจัย

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR เป็นวิธีวิเคราะห์ข้อสอบที่อยู่บนพื้นฐานทฤษฎีการตอบสนองข้อสอบ ทั้ง 3 วิธี (Ong, Lu, Lee & Cohen, 2015) และได้เปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ระหว่างวิธี HGLM วิธี MIMIC และวิธี IRT-LR โดยใช้การจำลองข้อมูลเชิงลำดับชั้น พบว่า วิธี MIMIC มีการควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 (Type I Error Rates) ได้ดีกว่าวิธี HGLM และวิธี IRT-LR เมื่อกลุ่มตัวอย่างมีขนาดเล็ก และถ้ากลุ่มตัวอย่างมีขนาดเพิ่มขึ้นจะทำให้อัตราความคลาดเคลื่อนประเภทที่ 1 (Type I Error Rates) เพิ่มขึ้นด้วย และวิธี IRT-LR ก็สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 (Type I Error Rates) ได้ดีเมื่อกลุ่มตัวอย่างมีขนาดเล็ก และถ้ากลุ่มตัวอย่างมีขนาดใหญ่จะทำให้อัตราความคลาดเคลื่อนประเภทที่ 1 (Type I Error Rates) เพิ่มขึ้น โดย Finch (2005) ได้เปรียบเทียบวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ระหว่าง วิธี MIMIC กับวิธี MH วิธี SIBTEST และวิธี IRT-LR แบบ 2 พารามิเตอร์ และ 3 พารามิเตอร์ พบว่า เมื่อตรวจสอบจากแบบทดสอบที่มีขนาดความยาวสั้น ทั้งแบบ 2 พารามิเตอร์ และ 3 พารามิเตอร์ วิธี MIMIC มีอัตราความคลาดเคลื่อนที่สูง และ Woods (2009) ได้เปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ระหว่างวิธี MIMIC กับวิธี IRT-LR เมื่อกลุ่มตัวอย่างมีขนาดเล็กกับแบบทดสอบที่มีรูปแบบการตอบแบบหลายตัวเลือก (Multiple Choice) พบว่า วิธี MIMIC มีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ดีกว่า วิธี IRT-LR ขณะที่ สุพัฒนา หอมบุปผา (2556) พบว่า วิธี HGLM สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันได้มากกว่า วิธี MIMIC และวิธี BAYSIAN ส่วนวิธีที่ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันน้อยที่สุดคือ วิธี MIMIC จากเหตุผลดังกล่าว ผู้วิจัยจึงตั้งสมมติฐาน ดังนี้

1. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านภาษา วิธี HGLM ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันมากกว่า วิธี MIMIC
2. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านภาษา วิธี HGLM ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันมากกว่า วิธี IRT-LR
3. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านภาษา วิธี IRT-LR ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันมากกว่า วิธี MIMIC
4. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านคำนวณ วิธี HGLM ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันมากกว่า วิธี MIMIC
5. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านคำนวณ วิธี HGLM ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันมากกว่า วิธี IRT-LR
6. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านคำนวณ วิธี IRT-LR ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันมากกว่า วิธี MIMIC
7. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านเหตุผล วิธี HGLM ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันมากกว่า วิธี MIMIC
8. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านเหตุผล วิธี HGLM ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันมากกว่า วิธี IRT-LR

9. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านเหตุผล วิธี IRT-LR ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันมากกว่า วิธี MIMIC

ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

1. สำนักทดสอบทางการศึกษาสามารถนำผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ นำข้อสอบที่ไม่ก่อให้เกิดความลำเอียงระหว่างเพศมาใช้ในการทดสอบครั้งต่อไป เพื่อใช้สำหรับสอบวัดความรู้ความสามารถของนักเรียนได้ตรงตามความสามารถที่แท้จริงของผู้สอบ

2. การทดสอบระดับชาติครั้งต่อไป เกิดประสิทธิภาพสำหรับการทดสอบมากยิ่งขึ้น เมื่อพิจารณาจากเนื้อหาของข้อคำถาม เพื่อหลีกเลี่ยงข้อคำถามที่ก่อให้เกิดการลำเอียงต่อผู้สอบ

3. นักวัดผลและประเมินผล นำผลการเปรียบเทียบผลการตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ ว่าวิธีใดสามารถตรวจพบได้มากกว่านั้น นำไปสู่การเลือกใช้วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

ขอบเขตของการวิจัย

การวิจัยครั้งนี้ใช้ข้อมูลทุติยภูมิ (Secondary Data) ของผลการทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2556 ของสำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน (สพฐ.) กระทรวงศึกษาธิการ

ประชากร เป็นนักเรียนชั้นประถมศึกษาปีที่ 3 ที่ทำแบบทดสอบระดับชาติ ปีการศึกษา 2556 ของสำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน (สพฐ.) กระทรวงศึกษาธิการทั้งประเทศ จำนวน 706,372 คน มาจากโรงเรียน จำนวน 30,283 โรง

2. ตัวแปรที่ศึกษา

2.1 ตัวแปรต้น มี 2 ตัว ได้แก่

2.1.1 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ จำนวน 3 วิธี ได้แก่

2.1.1.1 วิธี HGLM

2.1.1.2 วิธี MIMIC

2.1.1.3 วิธี IRT-LR

2.1.2 เพศ

2.1.2.1 เพศชาย

2.1.2.2 เพศหญิง

2.2 ตัวแปรตาม คือ ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยพิจารณาจากจำนวนข้อที่ตรวจพบการทำหน้าที่ต่างกันของข้อสอบ (DIF)

นิยามศัพท์เฉพาะ

การทดสอบระดับชาติ (National Testsing: NT) หมายถึง การทดสอบวัดความรู้และความคิดของนักเรียนชั้นประถมศึกษาปีที่ 3 ที่มุ่งวัดความสามารถของผู้เรียนทั้งหมด 3 ด้าน คือ ด้านภาษา ด้านคำนวณ และเหตุผล โดยสำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน กระทรวงศึกษาธิการ

การทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) หมายถึง ข้อสอบที่มีคุณสมบัติในการวัดความสามารถที่มุ่งวัดเดียวกัน แต่ผู้สอบที่มีคุณลักษณะบางประการแตกต่างกัน จึงมีโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกัน จึงก่อให้เกิดความลำเอียงของข้อสอบ นำไปสู่การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เพื่อพัฒนาให้ข้อสอบมีคุณภาพตรงตามมาตราฐานต่อไป

การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกัน หมายถึง การนำผลการตรวจพบข้อสอบที่ก่อให้เกิดการทำหน้าที่ต่างกันของข้อสอบ ระหว่างวิธีการตรวจสอบสองวิธี โดยทำการเปรียบเทียบ จำนวนข้อ (ร้อยละ) ที่พบการทำหน้าที่ต่างกันของข้อสอบ

ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) หมายถึง ทฤษฎีการวัดที่อธิบายความสัมพันธ์ระหว่างความสามารถที่อยู่ภายในตัวของบุคคล กับผลการตอบข้อคำถามหรือแบบทดสอบโดยใช้โค้งคุณลักษณะข้อสอบ แบบ 3 พารามิเตอร์ คือ ค่าอำนาจจำแนกของข้อสอบ (a) ค่าความยากของข้อสอบ (b) และค่าโอกาสในการเดาของข้อสอบ (c)

โมเดลเชิงเส้นตรงทั่วไประดับลดหลั่น (Hierarchical Generalized Linear Model: HGLM) หมายถึง รูปแบบหรือลักษณะการวิเคราะห์ข้อมูลเชิงเส้นทั่วไป ซึ่งถูกปรับให้ทำการวิเคราะห์ข้อมูลกับโมเดลการวิเคราะห์ข้อมูลแบบอื่น ๆ และการวิเคราะห์พหุระดับที่มีข้อมูลสอดแทรกเป็นระดับลดหลั่นได้

โมเดลมิมิค (Multiple-Indicators Multiple-Causes Model: MIMIC Model) หมายถึง โมเดลสมาการโครงสร้างที่มีตัวแปรแฝงเพียงตัวแปรเดียว โดยที่ตัวแปรแฝงนั้นได้รับอิทธิพลจากตัวแปรภายนอกสังเกตได้หลายตัวแปรและสามารถส่งอิทธิพลไปยังตัวแปรภายในสังเกตได้หลายตัวแปร

อัตราส่วนไลค์ลิฮูด (Likelihood Ratio: LR) หมายถึง การใช้หลักของอัลกอริทึมในการประมาณค่าความเป็นไปได้สูงสุดในการประมาณค่าพารามิเตอร์ สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ทั้งแบบเอกรูป และอเนกรูป

วิธี Hierarchical Generalized Linear Model: HGLM หมายถึง การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยวิธีการวิเคราะห์ข้อมูลแบบพหุระดับ มีการวิเคราะห์ 2 ขั้น ด้วยโปรแกรม HLM ตามหลักการวิเคราะห์โมเดลสมการเชิงเส้นตรงระดับลดหลั่น

วิธี Multiple-Indicators Multiple-Causes: MIMIC หมายถึง การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยการวิเคราะห์ข้อมูลด้วยโมเดลของคุณลักษณะแฝงที่มีหลายสาเหตุและสามารถวัดได้จากตัวบ่งชี้หลายตัว ด้วยโปรแกรม Mplus

วิธี Item Response Theory – Likelihood Ratio: IRT-LR หมายถึง การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยการวิเคราะห์ข้อมูลเพื่อหาโค้งคุณลักษณะข้อสอบ และประมาณค่าพารามิเตอร์ว่ามีการทำหน้าที่ต่างกันหรือไม่ ด้วยโปรแกรมสำเร็จรูป IRTPRO ตามหลักทฤษฎีการตอบสนองข้อสอบ (IRT)

เพศ (Gender) หมายถึง เพศของผู้เรียนที่เป็นกลุ่มตัวอย่างในการศึกษาวิจัยในครั้งนี้ แบ่งเป็นเพศชาย และเพศหญิง

กลุ่มเปรียบเทียบ (Focal Group: F) หมายถึง กลุ่มที่สนใจศึกษาคาดว่าจะเป็นกลุ่มที่เสียเปรียบในการตอบข้อสอบโดยมีข้อสอบชุดเดียวกับกลุ่มอ้างอิง

กลุ่มอ้างอิง (Reference Group: R) หมายถึง กลุ่มที่คาดว่าจะได้เปรียบในการตอบข้อสอบได้ถูกต้องในการทำข้อสอบชุดเดียวกันกับกลุ่มเปรียบเทียบ

ค่าอำนาจจำแนกของข้อสอบ (a-parameter) หมายถึง ค่าที่แสดงถึงประสิทธิภาพข้อสอบที่มีความสามารถในการจำแนกคุณลักษณะบางประการของผู้สอบได้ เมื่ออำนาจจำแนกมีค่าเข้าใกล้ 1 หรือมากกว่า 1 แสดงว่าข้อสอบนั้นมีอำนาจจำแนกได้ดี และถ้าค่าอำนาจจำแนกติดลบ แสดงว่าข้อสอบข้อนั้นไม่สามารถจำแนกกลุ่มผู้สอบได้ โดยค่าอำนาจจำแนกของข้อสอบในการวิจัยนี้อยู่ระหว่าง 0.50 ถึง 2.50

ค่าความยากของข้อสอบ (b-parameter) หมายถึง ค่าที่แสดงถึงสัดส่วนของผู้สอบที่ตอบข้อสอบได้ถูกต้องจากผู้สอบทั้งหมด ซึ่งเมื่อค่าความยากของข้อสอบเข้าใกล้ 2 แสดงว่าข้อสอบข้อนั้นมีความยากที่เหมาะสมต่อการนำไปใช้ แต่หากข้อสอบข้อใดมีค่าความยากของข้อสอบติดลบ แสดงว่าข้อสอบเป็นข้อสอบที่ง่ายเกินไป โดยการวิจัยนี้กำหนดค่าความยากของข้อสอบอยู่ที่ -2.50 ถึง 2.50

ค่าโอกาสการเดาของข้อสอบ (c-parameter) หมายถึง ค่าที่แสดงถึงค่าการเดาข้อสอบที่ผู้สอบมีโอกาสเดาข้อสอบได้ถูกต้อง ซึ่งความสามารถของผู้สอบในเรื่องนั้น ๆ เป็น 0 เมื่อข้อสอบมีค่าโอกาสการเดามากกว่า 0.3 แสดงว่าข้อสอบข้อนั้นไม่มีประสิทธิภาพ แต่หากข้อสอบข้อใดมีค่าโอกาสการเดาของข้อสอบเป็น 0 แสดงว่าข้อสอบนั้นเป็นข้อสอบที่ดี เหมาะสำหรับการนำไปใช้ โดยการวิจัยนี้กำหนดค่าโอกาสการเดาข้อสอบได้ถูกไว้ที่ 0 ถึง 1

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

การวิจัยครั้งนี้ผู้วิจัยได้ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR โดยนำเสนอเป็น 5 ตอน ดังนี้

ตอนที่ 1 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) และงานวิจัยที่เกี่ยวข้อง

ตอนที่ 2 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ด้วยวิธี HGLM และงานวิจัยที่เกี่ยวข้อง

ตอนที่ 3 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ด้วยวิธี MIMIC และงานวิจัยที่เกี่ยวข้อง

ตอนที่ 4 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ด้วยวิธี IRT-LR และงานวิจัยที่เกี่ยวข้อง

ตอนที่ 5 การทดสอบระดับชาติ (NT) และงานวิจัยที่เกี่ยวข้อง

ตอนที่ 1 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) และงานวิจัยที่เกี่ยวข้อง

การทำหน้าที่ต่างกันของข้อสอบ (DIF) มีการเริ่มศึกษาในช่วงปลายของปี ค.ศ. 1960 แรกเริ่มนั้นใช้คำว่า ความลำเอียงของข้อสอบ (Item Bias) หรือความลำเอียงของแบบสอบ (Test Bias) การทำหน้าที่ต่างกันของข้อสอบนั้น มีนักวัดผลได้ศึกษาไว้อย่างมากมาย และได้มีความเห็นขัดแย้งกันว่าการลำเอียงของข้อสอบ เป็นการดูว่าข้อสอบฉบับนั้นมีความยุติธรรมในการวัดหรือไม่ เป็นการประเมินความสัมพันธ์ระหว่างการวัดคุณลักษณะที่มุ่งหวังกับความสามารถที่แท้จริงของผู้สอบ เมื่อผู้สอบต่างกลุ่มกันตอบข้อสอบข้อเดียวกัน ความแตกต่างกันของประสบการณ์หรือคุณลักษณะอื่นๆ ทำให้ไม่เหมาะสมที่จะใช้คำว่า ข้อสอบลำเอียง (Item Bias) เพราะมีความหมายไปทางลบ จึงเปลี่ยนมาใช้คำว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (Diferential Item Functioning: DIF) ที่มีความเหมาะสมมากกว่า (Holland & Wainer, 1993, pp. 103-105) ต่อมาได้มีการนิยามการทำหน้าที่ของข้อสอบนี้ว่าเป็นการคลาดเคลื่อนอย่างเป็นระบบ (Systematic Error) ของการวัดเป็นการตรวจสอบความยุติธรรมของข้อสอบ ว่าข้อสอบสามมารถตรวจสอบได้ตรงกับคุณลักษณะที่แท้จริงของผู้สอบมากเพียงใด โดยในการทำหน้าที่ต่างกันของข้อสอบนั้น ก็จะตรวจสอบคุณภาพข้อสอบที่คาดว่าจะมีการเกิดการทำหน้าที่ต่างกันหรือเกิดการลำเอียงในการวัดคุณลักษณะ ซึ่งการตรวจสอบนี้เป็นการกำจัดข้อสอบที่ก่อให้เกิดปัญหาด้านความยุติธรรมระหว่างกลุ่มของผู้สอบที่มีลักษณะต่างกัน เช่น เพศ ศาสนา วัฒนธรรม สังคม ภูมิลำเนา ประสบการณ์ หรืออายุ เป็นต้น เพื่อทำให้ข้อสอบมีคุณภาพที่จะนำไปใช้ทดสอบ (ศิริชัย กาญจนวาสี, 2555, หน้า 115)

รูปแบบของข้อสอบเป็นอีกปัจจัยหนึ่งที่ก่อให้เกิดการทำหน้าที่ต่างกันของข้อสอบ เมื่อผู้สอบมีคุณลักษณะบางอย่างแตกต่างกัน กล่าวได้คือ เพศชายจะมีความถนัดในการทำข้อสอบแบบปรนัย (Multiple-Choice Item) ในขณะที่เพศหญิงมีความถนัดในการทำข้อสอบแบบอัตนัยหรือแบบเขียนตอบ (Constructed-Response Item) จึงทำให้การตอบสนองข้อสอบของกลุ่มผู้สอบมีค่าแตกต่างกัน

(Garner & Englehard, 1999; Zenisky & colleagues, 2004) รวมถึงความสามารถในการเรียนรู้ที่อาจก่อให้เกิดการต่างกันของความสามารถในการตอบข้อสอบ จากการศึกษางานวิจัยพบว่า เพศชายมีความถนัดในการใช้องค์ความรู้ที่สูง เช่น วิชาคิดคำนวณ การแก้ปัญหา และการประยุกต์การใช้แนวคิดในการแก้ปัญหา (Fennema & Carpenter, 1981; Marshall, 1984) ซึ่งในขณะที่เพศหญิงถนัดในการใช้องค์ความรู้ที่ต่ำกว่า (Carlton & Harris, 1992)

นักวิจัยทางการวัดหลายท่านได้ให้ความหมายของการทำหน้าที่ต่างกันของข้อสอบดังนี้

ความลำเอียงของข้อสอบ (Item Bias) หมายถึง ผู้สอบที่มีความสามารถหรือคุณลักษณะที่ต้องการวัดเท่ากัน แต่มาจากกลุ่มประชากรที่แตกต่างกัน เมื่อได้รับข้อสอบสำหรับวัดความสามารถจึงมีโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกัน (Hulin, Drasgow, & Parson, 1983)

ความลำเอียงของข้อสอบ (Item Bias) หมายถึง อัตราการตอบข้อสอบได้ถูกต้องไม่เท่ากันระหว่างกลุ่มผู้สอบที่นำมาทำการศึกษา (Scheuneman, 1979)

การทำหน้าที่ต่างกันของข้อสอบ (DIF) หมายถึง สารสนเทศทางสถิติของข้อสอบที่ได้จากผลการตอบของผู้สอบต่างกลุ่มกัน และมีความสามารถเท่ากัน แต่มีโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกัน (Holland & Wainer, 1993, pp. 103-104)

การทำหน้าที่ต่างกันของข้อสอบ หมายถึง การวัดความเป็นพหุมิติในข้อสอบ ที่แสดงผลระหว่างกลุ่มผู้สอบ 2 กลุ่มขึ้นไปที่มีความสามารถหลัก (Primary Ability) เหมือนกัน แต่มีความสามารถรอง (Secondary Ability) ที่แตกต่างกัน (Camilli & Shrpard, 1994, pp. 15-17)

การทำหน้าที่ต่างกันของข้อสอบ หมายถึง ข้อสอบที่ทำให้ผู้สอบสองกลุ่มที่มีคุณลักษณะที่มุ่งวัด ได้รับโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกัน โดยมีความสามารถหลัก (Primary Ability) และคุณลักษณะแฝง (Latent Trait) ในระดับที่เท่ากัน แต่มีความสามารถรอง (Secondary Ability) ที่แตกต่างกัน จึงทำให้ผู้สอบมีโอกาสในการตอบข้อสอบถูกต้องแตกต่างกัน (ศิริชัย กาญจนวาสี, 2555, หน้า 117)

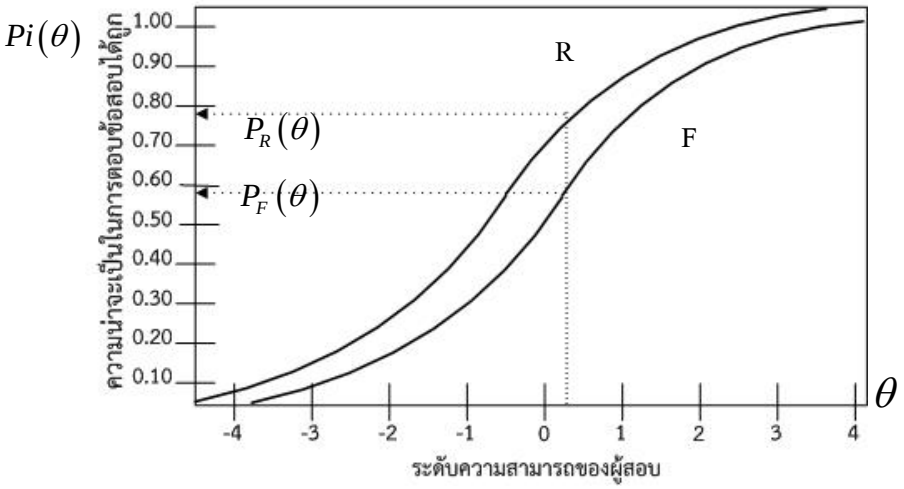
ดังนั้นจึงสรุปได้ว่า การทำหน้าที่ต่างกันของข้อสอบ หมายถึง ข้อสอบที่มีคุณสมบัติในการวัดความสามารถที่มุ่งวัดเดียวกัน แต่ผู้สอบที่มีคุณลักกษณะบางประการแตกต่างกัน จึงมีโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกัน จึงก่อให้เกิดการลำเอียงของข้อสอบ นำไปสู่การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เพื่อพัฒนาให้ข้อสอบมีคุณภาพตรงตามมาตราฐานต่อไป

ประเภทของการทำหน้าที่ต่างกันของข้อสอบ

การทำหน้าที่ต่างกันของข้อสอบ (DIF) โดยปกตินิยมเปรียบเทียบผลการตอบข้อสอบของผู้สอบตั้งแต่ 2 กลุ่มขึ้นไป กลุ่มแรกคือ กลุ่มผู้สอบที่คาดว่าจะเสียเปรียบในการทำข้อสอบ เรียกว่า กลุ่มเปรียบเทียบ (Focal Group หรือ F) กลุ่มสองคือ กลุ่มผู้สอบที่คาดว่าจะได้เปรียบในการทำข้อสอบ เรียกว่า กลุ่มอ้างอิง (Referent Group หรือ R) ซึ่งในการทำหน้าที่ต่างกันของข้อสอบ จะสามารถแยกได้ 2 ประเภท คือ 1) ข้อสอบทำหน้าที่ต่างกันแบบเอกรูป (Uniform DIF) และ 2) ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป (Nonuniform DIF) ดังนี้

1. ข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูป (Uniform DIF) หมายถึง ข้อสอบที่ทำหน้าที่ต่างกันโดยผู้สอบกลุ่มหนึ่งมีโอกาสตอบข้อสอบถูกมากกว่าผู้สอบอีกหนึ่งกลุ่มอย่างสม่ำเสมอ ในทุกระดับ

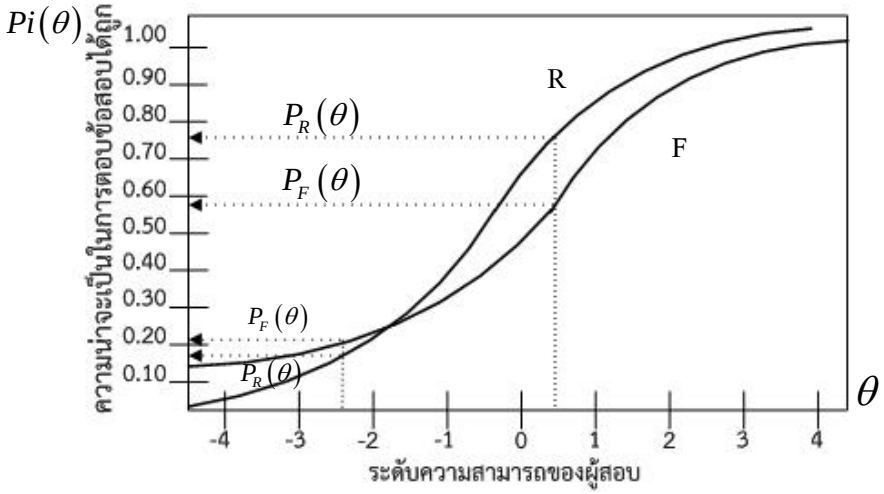
ความสามารถ และเมื่อพิจารณาโค้งคุณลักษณะข้อสอบของผู้สอบ 2 กลุ่ม พบว่า ไม่มีปฏิสัมพันธ์ภายในระดับความสามารถของผู้สอบ ดังภาพที่ 2-1



ภาพที่ 2-1 การทำหน้าที่ต่างกันแบบเอกรูป (Uniform DIF) (Osterlind & Everson, 2009, p. 12)

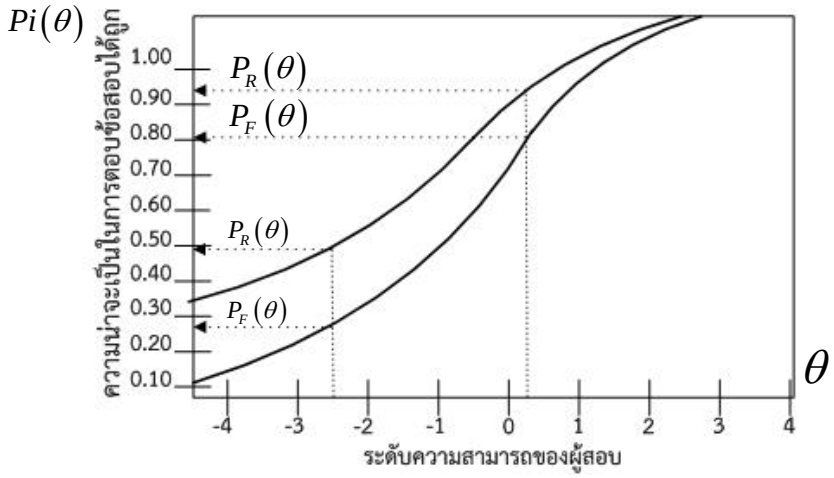
2. ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป (Nonuniform DIF) หมายถึง ข้อสอบที่ทำหน้าที่ต่างกันโดยผู้สอบกลุ่มหนึ่งมีโอกาสตอบข้อสอบถูกมากกว่าผู้สอบอีกหนึ่งกลุ่มอย่างไม่สม่ำเสมอ ในทุกระดับความสามารถ และเมื่อพิจารณาโค้งคุณลักษณะข้อสอบของผู้สอบ 2 กลุ่ม พบว่า มีปฏิสัมพันธ์ภายในระดับความสามารถของผู้สอบ โดยการที่ผู้สอบกลุ่มหนึ่งได้เปรียบที่ช่วงความสามารถหนึ่งและเสียเปรียบที่อีกช่วงความสามารถหนึ่ง ความแตกต่างของโค้งข้อสอบ

2.1 ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปโดยมีปฏิสัมพันธ์ไม่เป็นลำดับ (Disordinal interaction) เป็นการทำหน้าที่ต่างกันสำหรับกลุ่มผู้สอบซึ่งเกิดขึ้น เมื่อโค้งคุณลักษณะข้อสอบตัดกันระหว่างช่วงความสามารถของผู้สอบหรือเรียกว่าข้อสอบทำหน้าที่ต่างกันแบบไม่มีทิศทาง (Non-Unidirectional DIF)



ภาพที่ 2-2 การทำหน้าที่ต่างกันแบบไม่มีทิศทาง (Non-Unidirectional DIF) (Ayala, 2009, p. 326)

2.2 ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป โดยมีการปฏิสัมพันธ์เป็นลำดับ (Ordinal Interaction) เป็นการทำหน้าที่ต่างกัน สำหรับกลุ่มผู้สอบซึ่งเกิดขึ้น เมื่อโค้งคุณลักษณะข้อสอบต่างกันอย่างไม่สม่ำเสมอ แต่ไม่ตัดกัน หรืออาจตัดกันนอกช่วง ความสามารถของผู้สอบตารางปลายสุดของช่วงความสามารถต่ำหรือสูง อาจเรียกข้อสอบลักษณะนี้ว่าข้อสอบทำหน้าที่ต่างกันแบบมีทิศทางเดียว (Unidirectional DIF)



ภาพที่ 2-3 การทำหน้าที่ต่างกันแบบมีทิศทาง (Unidirectional DIF) (ศิริชัย กาญจนวาสี, 2555, หน้า 120)

หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

การตรวจการทำหน้าที่ต่างกันของข้อสอบ (DIF Detection) เป็นการเปรียบเทียบผลการตอบสนองของข้อสอบเป็นรายข้อระหว่างกลุ่มผู้สอบอย่างน้อย 2 กลุ่ม ที่มีความสามารถหลัก (Primary Ability) ที่มุ่งวัดเท่ากัน แต่คาดว่าจะไม่มีความได้เปรียบหรือเสียเปรียบกัน โดยกลุ่มหนึ่งถือเป็น กลุ่มอ้างอิง (Reference Group) ซึ่งคาดว่าน่าจะได้เปรียบในการตอบข้อสอบข้อนั้น หรือมีโอกาสตอบข้อสอบได้ถูกต้องมากกว่า ส่วนอีกกลุ่มคือ กลุ่มเปรียบเทียบ (Foacal Group) ซึ่งเป็นกลุ่มสนใจศึกษาและค่าว่าน่าจะได้เป็นกลุ่มที่เสียเปรียบในการเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบจำเป็นต้องจับคู่ (Matching) ผู้สอบตามความสามารถซึ่งเป็นเงื่อนไขสำคัญของการตรวจการทำหน้าที่ต่างกันของข้อสอบ

เกณฑ์การจับคู่ (Matching Criteria) ที่นิยมใช้กันมี 2 วิธีดังนี้

1. เกณฑ์ภายนอก (External Criterion)

การวิเคราะห์การทำหน้าที่ต่างกัน โดยใช้เกณฑ์ภายนอกนี้สามารถนำไปใช้ได้ทั้งข้อสอบรายข้อและแบบสอบทั้งฉบับ โดยการใช้คะแนนจากแบบสอบอื่นเป็นเกณฑ์ภายนอกแล้วใช้เทคนิคการการถดถอย (Regression Analysis) เพื่อเปรียบเทียบเส้นกราฟความสัมพันธ์ระหว่างตัวแปรเกณฑ์กับตัวแปรทำนายระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ

หลักการและจุดมุ่งหมายนี้เพื่อสร้างสมการทำนายตัวแปรเกณฑ์ ซึ่งเป็นคะแนนของแบบสอบอื่นจากตัวแปรทำนายซึ่งเป็นคะแนนรายข้อ หรือคะแนนแบบสอบระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ จะใช้คะแนนรายข้อเป็นตัวแปรทำนาย

แต่ถ้าเป็นการวิเคราะห์การทำหน้าที่ต่างกันของแบบทดสอบจะใช้คะแนนรวมของแบบสอบทั้งสองฉบับเป็นตัวแปรทำนาย สำหรับตัวแปรเกณฑ์ที่ใช้เป็นเกณฑ์ภายนอก อาจใช้คะแนนรวมทั้งฉบับหรือเกรดเฉลี่ย หรือผลสัมฤทธิ์ในงานที่เกี่ยวข้องของผู้สอบ (Cronbach & Furby, 1970) สมการทำนายสำหรับกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ดังนี้

กลุ่มอ้างอิง (R) $Y_i = A_R + B_R X_i$ (1)

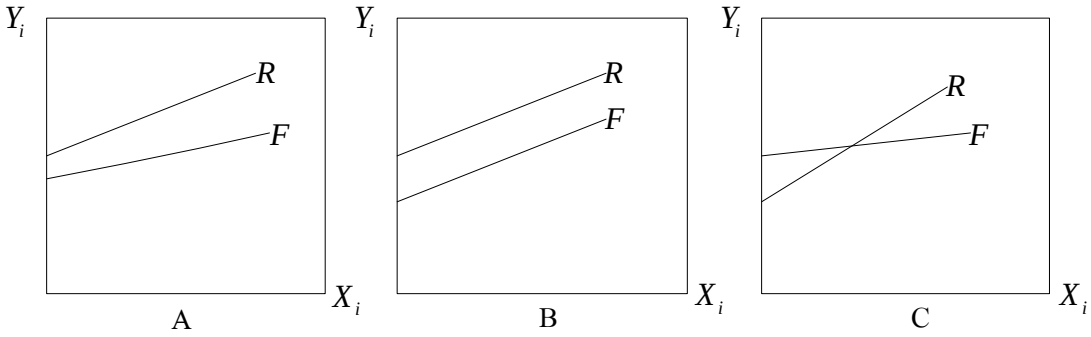
กลุ่มเปรียบเทียบ (F) $Y_i = A_F + B_F X_i$ (2)

เมื่อ Y_i = คะแนนของตัวแปรเกณฑ์ภายนอก

X_i = คะแนนของตัวแปรทำนาย

A = ค่าคงที่หรือค่าตัดแกน (Intercept)

B = ค่าความชัน (Slope)



ภาพที่ 2-4 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบและแบบทดสอบ โดยการใช้วิธีวิเคราะห์สมการถดถอย (ศิริชัย กาญจนวาสี, 2555, หน้า 121)

การใช้เกณฑ์ภายนอกมีข้อดี คือ เกณฑ์ที่ใช้มีความเป็นอิสระจากข้อสอบ และแบบสอบที่ดีต้องมีการตรวจสอบ แต่มีจุดอ่อนตรงที่ความเหมาะสมของเกณฑ์ที่จะนำมาใช้ในทางปฏิบัติหาตัวแปรเกณฑ์ภายนอกจากแบบทดสอบฉบับอื่นที่มีความตรงเชิงทำนายและมีความยุติธรรมสำหรับกลุ่มอ้างอิงและกลุ่มเปรียบเทียบได้ยาก หากตัวแปรเกณฑ์ภายนอกขาดคุณสมบัติดังกล่าวจะทำให้ผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบหรือแบบทดสอบขาดความแม่นยำและความสมบูรณ์

2. เกณฑ์ภายใน (Internal Criterion)

การวิเคราะห์การทำหน้าที่ต่างกัน โดยใช้เกณฑ์ภายในเป็นการนำวิธีทางสถิติมาตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ หรือแบบทดสอบ หรือแบบสอบ โดยเน้นการพิจารณาจากโครงสร้างภายในของแบบทดสอบหลักด้วยการวิเคราะห์ผลการตอบข้อสอบและความสามารถหรือคะแนนจริงของผู้สอบจากแบบทดสอบฉบับนั้น เพื่อนำมาเปรียบเทียบระหว่างผู้สอบจากกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ที่มีความสามารถหรือคะแนนจริงเท่ากัน ว่าจะมีผลการตอบหรือโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกันหรือไม่ เพื่อบ่งชี้ถึงการทำหน้าที่ต่างกันของข้อสอบ

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF Detection) สามารถจำแนกตามลักษณะการตรวจให้คะแนนได้เป็น 2 ประเภทคือ ข้อสอบที่มีการให้คะแนนแบบทวิภาค หรือสองค่า (Dichotomous Scoring) และข้อสอบที่มีการให้คะแนนแบบพหุวิภาค หรือหลายค่า (Polytomous Scoring) โดยวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแต่ละประเภท ยังสามารถจำแนกได้อีก 2 แบบได้แก่ แบบลักษณะของตัวแปรเกณฑ์ ซึ่งแบ่งเป็นกลุ่มวิธีที่ใช้คะแนนสังเกตได้ (Observed Score) และกลุ่มวิธีที่ใช้คะแนนสังเกตไม่ได้หรือคะแนนนของตัวแปรแฝง (Latent Variable) และลักษณะของสถิติในการวิเคราะห์สามารถแบ่งออกเป็นกลุ่มที่ใช้สถิติพาราเมตริก (Parametric Approach) และกลุ่มที่ใช้สถิติแบบนันพาราเมตริก (Nonparametric Approach) ซึ่งวิธีการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ ที่นิยมใช้กันโดยทั่วไป บางวิธีได้แก่ วิธีแมนเทล – แฮนส์เซล (Mantel-Haenzel: MH) วิธีถดถอยโลจิสติก (Logistic Regression: LR) วิธีซิปเทสท์ (SIBTEST) และวิธีตรวจสอบด้วยทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) จะประกอบด้วยวิธีการวัดความแตกต่างของพื้นที่ วิธีวัดความแตกต่างของค่าพารามิเตอร์ b และวิธีทดสอบไค – สแควร์ (Lord's χ^2 - Test) (Feinstein, 1995; Potenza and Dorans, 1995; ศิริชัย กาญจนวาสี, 2555, หน้า 124-151) และ Wiberg (2007, p. 24) ได้มีการเปรียบเทียบคุณสมบัติของวิธีวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยกำหนดการแสดงให้แสดงค่า ของวิธีการวิเคราะห์ ดังนี้

ตารางที่ 2-1 การเปรียบเทียบคุณสมบัติของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

วิธีการตรวจสอบ	Par/Non-p	Item Scores	U/N
Mantel-Haenszel	Non-p	D/P	U
Standardization	Non-p	D	U
Chi-square Methods	Non-p	D	U
SIBTEST	Non-p	D/P	U/N
Logistic Regression	Par	D/P	U/N
Likelihood Ratio Test	Par	D/P	U/N
b Parameter Indices	Par	D	U/N
General IRT-LR	Par	D/P	U/N
IRT LRT	Par	D/P	U/N
IRT Methods	Par	D/P	U/N
Lord’s Chi-squared Test	Par	D	U/N
Log Linear Models	Par	D/P	U/N
Mixed Effect Models	Par	D/P	U/N

หมายเหตุ Par หมายถึง Parametric Non-p หมายถึง Non-parametric
D หมายถึง Dichotomusly P หมายถึง Polytomously
U หมายถึง Uniform N หมายถึง Non-uniform

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนแบบทวิภาค

กลุ่มที่ใช้วิธีคะแนนสังเกตได้

วิธีในกลุ่มนี้มักวิเคราะห์ตามทฤษฎีการทดสอบแบบดั้งเดิม (CTT) หรือกลุ่มที่ไม่ใช้ทฤษฎีการตอบสนองข้อสอบ (Non-IRT Approach) โดยใช้คะแนนรวมของผู้สอบเป็นเกณฑ์การจับคู่ของกลุ่มผู้สอบ วิธีการตรวจสอบที่สำคัญในกลุ่มนี้ คือ

การวิเคราะห์ความแปรปรวน (ANOVA)

วิธีการวิเคราะห์การถดถอยโลจิสติก (Logistic Regression: LR)

วิธีการแปลงค่าความยากของข้อสอบ (Transformed Item Difficulty: TID)

วิธีแมนเทล-แฮนส์เซล (Mantel - Haenszel: MH)

วิธีดัชนีมาตรฐาน (Standardization: STND) การปรับให้เป็มาตราฐานด้วยน้ำหนักตัวประกอบ

กลุ่มวิธีใช้คุณลักษณะแฝง

วิธีในกลุ่มนี้ใช้คุณลักษณะหรือตัวแปรแฝง ซึ่งวิเคราะห์บนพื้นฐานของทฤษฎีการตอยสนองข้อสอบ (IRT) สำหรับใช้เป็นเกณฑ์การจัดกลุ่มผู้สอบ วิธีการตรวจสอบสำคัญมีดังนี้

วิธีวัดพื้นที่ความแตกต่างระหว่างโค้งการตอบสนองข้อสอบ (IRT-D^2)

วิธีไค-สแควร์ของลอร์ด (Lord's χ^2)

วิธีอัตราส่วนไลค์ลิฮูดทั่วไป (General IRT Linklihood Ratio)

วิธีอัตราส่วนไลค์ลิฮูด ลอกลิเนียร์ (Loglinear IRT Linklihood Ratio)

วิธีซิปเทสท์ (SIBTEST)

เกณฑ์สำหรับเปรียบเทียบคุณภาพของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (Criteria For Comparison between DIF Detecting Methods)

คุณภาพของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ พิจารณาจากดัชณีบ่งชี้สำคัญ 2 ตัว ประกอบด้วย อำนาจการทดสอบ หรืออัตราความถูกต้องของการตรวจพบข้อที่ทำหน้าที่ต่างกัน (Power Rate) และความคลาดเคลื่อนของการตรวจสอบ หรืออัตราตามความคลาดเคลื่อนประเภทที่ 1 (Type I Error Rate) ซึ่งเป็นช่องว่างที่ก่อให้เกิดความคลาดเคลื่อนในลักษณะที่ตรวจพบว่า ข้อสอบทำหน้าที่ต่างกัน โดยที่แท้จริงแล้วข้อสอบนั้นไม่ได้ทำหน้าที่ต่างกัน โดยมีการคำนวณค่าสถิติ ที่มุ่งหมายเพื่อทดสอบนัยสำคัญของผลการตรวจสอบ โดยมีสมมติฐานศูนย์ของการทดสอบ คือ ข้อสอบไม่ได้ทำหน้าที่ต่างกัน หรือ Ho : No DIF ผลการทดสอบสมมติฐานของวิธีการตรวจสอบ DIF ด้วยวิธีต่าง ๆ นำไปสู่การตัดสินใจว่า ยอมรับสมมติฐานศูนย์ (Accept Ho) หรือปฏิเสธสมมติฐานศูนย์ (Reject Ho) โดยผลการตัดสินใจมีโอกาสเกิดขึ้นได้ 4 ลักษณะ ดังนี้

ตารางที่ 2-2 คุณภาพของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF)

การตัดสินใจตามผลการตรวจสอบ	Ho: No DIF	
	ความเป็นจริง	
	Ho ถูก	Ho ผิด
Accept Ho	ตัดสินถูก (True Negative) ระดับความเชื่อมั่น (1-α)	ตัดสินผิด (Type II Error, β) False Negative
Reject Ho	ตัดสินผิด (Type I Error, α) False Positive	ตัดสินถูก (True Positive) ระดับความเชื่อมั่น (1-β)

จากค่าสถิติที่คำนวณตามวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ จะนำไปสู่การตัดสินใจ และสามารถสรุปผลการตรวจสอบได้ ดังนี้

1. ตัดสินใจถูก สามารถเกิดขึ้นได้ 2 ลักษณะ คือ
   ข้อสอบ No DIF ตามความเป็นจริง (True Negative)
   ข้อสอบ DIF ตามความเป็นจริง (True Positive)
2. ตัดสินใจผิด สามารถเกิดขึ้นได้ 2 ลักษณะ คือ
   ข้อสอบ DIF ทั้งที่ความเป็นจริงแล้วข้อสอบ No DIF (False Positive)
   ข้อสอบ No DIF ทั้งที่ความเป็นจริงแล้วข้อสอบ DIF (False Negative)

เนื่องจากอำนาจการทดสอบ (β) กับ β เป็นค่าดัชนีที่มีสเกลผกผันกันและมีอัตราความคลาดเคลื่อนประเภทที่ 1 (α) กับ 1-α ต่างก็เป็นดัชนีที่มีสเกลที่มีค่าผกผันเช่นเดียวกัน ดังนั้น การพิจารณาดัชนีบ่งชี้คุณภาพ 2 ตัว คือ อัตราความถูกต้องของการตรวจสอบ DIF (Power Rate) และอัตราความคลาดเคลื่อนประเภทที่ 1 (Type I Error Rate) ก็เพียงพอที่จะให้สารสนเทศครบทั้ง 4 เหตุการณ์

สรุป คือ การทำหน้าที่ต่างกันของข้อสอบ (DIF) นั้นเป็นการวิเคราะห์ข้อสอบที่คาดว่ามีการทำหน้าที่ต่างกันระหว่างผู้สอบ 2 กลุ่ม ที่มีคุณลักษณะบางอย่างที่แตกต่างกัน ซึ่งสามารถข้อสอบสามารถทำหน้าที่ต่างกันได้ 2 ลักษณะ คือ ทำหน้าที่ต่างกันอย่างเอกรูป (Uniform DIF) และทำหน้าที่ต่างกันแบบอเนกรูป (Non-uniform DIF) โดยการตรวจสอบสามารถทำได้หลากหลายวิธี จำแนกได้ 3 กลุ่ม คือ 1) กลุ่มที่ใช้คะแนนสังเกตุได้ หรือคะแนนจริง 2) กลุ่มที่ใช้สถิติพาราเมตริก หรือนันพาราเมตริก และ 3) กลุ่มที่ใช้วิธีโมเดล IRT ดังที่กล่าวมา เป็นการตรวจสอบว่าข้อสอบมีการทำหน้าที่ต่างกันหรือไม่ และเป็นแนวทางของการพัฒนาข้อสอบให้มีคุณภาพมากยิ่งขึ้นจากเดิม

งานวิจัยที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกัน (DIF) มีดังนี้

พีรญา สูงเนิน (2552) ศึกษาเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบแบบพหุมิติ ระหว่างข้อสอบรายข้อกับหมวดข้อสอบ โดยวิธีซิปเทสท์ ภายใต้เงื่อนไขขนาดของกลุ่มตัวอย่างที่ต่างกัน คือ ขนาดเล็ก 300 คน ขนาดกลาง 1,000 คน และขนาดใหญ่ 2,000 คน ซึ่งกลุ่มตัวอย่างคือนักเรียนชั้นประถมศึกษาปีที่ 6 สังกัดพื้นที่การศึกษานครศรีธรรมราช ปีการศึกษา 2546

ของนักเรียนที่เข้าสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติจำนวน 2,000 คน โดยใช้ข้อมูล ทุติยภูมิจากคะแนนแบบทดสอบวิชาภาษาไทยชั้นประถมศึกษาปีที่ 6 จำนวน 40 ข้อ เมื่อนำไปตรวจสอบการทำหน้าที่ต่างกัน พบว่า กลุ่มตัวอย่างขนาดเล็ก พบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 4 ข้อ คิดเป็นร้อยละ 10 กลุ่มตัวอย่างขนาดกลาง พบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 13 ข้อ คิดเป็นร้อยละ 32.50 และกลุ่มตัวอย่างขนาดใหญ่ พบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 15 ข้อ คิดเป็นร้อยละ 37.50 เมื่อนำไปตรวจสอบทีละหมวด พบว่า กลุ่มตัวอย่างขนาดเล็ก พบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 4 ข้อ คิดเป็นร้อยละ 10 กลุ่มตัวอย่างขนาดกลาง พบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 8 ข้อ คิดเป็นร้อยละ 20 และกลุ่มตัวอย่างขนาดใหญ่ พบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 16 ข้อ คิดเป็นร้อยละ 40 สรุปได้ว่า ขนาดกลุ่มตัวอย่างที่ใหญ่ทำให้สามารถพบข้อสอบที่ทำหน้าที่ต่างกันได้ดีดว่ากลุ่มตัวอย่างที่มีขนาดเล็ก

ศิริรัตน์ สุคันธพฤกษ์ (2554) ศึกษาตรวจสอบการทำหน้าที่ต่างกันในแบบวัดความวิตกกังวล ในการสอบคณิตศาสตร์ โดยเปรียบเทียบระหว่าง Hierarchical Linear Model: HLM, Partial Credit Model: PCM และ Graded Response Model: GRM โดยกลุ่มตัวอย่างที่ใช้ในการวิจัยเป็นนักเรียนมัธยมศึกษาปีที่ 6 สายวิทย์-คณิต ปีการศึกษา 2552 จำนวน 1,715 คน จาก 29 โรงเรียน ในสังกัดสำนักงานเขตพื้นที่การศึกษาพระนครศรีอยุธยาเขต 1 และเขต 2 สำนักงานเขตพื้นที่การศึกษาอ่างทองและสำนักงานเขตพื้นที่การศึกษานนทบุรี ซึ่งได้มาจากการสุ่มตัวอย่างแบบยกชั้น เครื่องมือที่ใช้ในการวิจัย คือ แบบวัดความวิตกกังวลในการสอบคณิตศาสตร์ โดยวิเคราะการทำหน้าที่ต่างกันของข้อสอบด้วย วิธี HLM โดยใช้โปรแกรม HLM และวิธี PCM กับวิธี GRM ด้วยโปรแกรม PRASCALE โดยเปรียบเทียบผลการวิเคราะห์ข้อมูลทั้ง 3 วิธี พบว่า ข้อคำถามที่ทำหน้าที่ต่างกันของข้อร่วมระหว่าง HLM, PCM และ GRM มี 6 ข้อ จาก 39 ข้อ คิดเป็นร้อยละ 15.38 ข้อคำถามที่ทำหน้าที่ต่างกันของข้อร่วมระหว่าง HLM กับ PCM มี 7 ข้อ จาก 39 ข้อ คิดเป็นร้อยละ 17.94 และข้อคำถามที่ทำหน้าที่ต่างกันของข้อร่วมระหว่าง HLM กับ GRM มี 9 ข้อ จาก 39 ข้อ คิดเป็นร้อยละ 23.07

ชัยวัฒน์ หฤทัยพันธ์ (2558) ศึกษาพัฒนาวิธีการสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยผู้เชี่ยวชาญ และประการที่สองเพื่อเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในด้านอัตราความถูกต้อง และอัตรา ความคลาดเคลื่อนของผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับข้อสอบที่คัดสรรมาสำหรับ การทดลองเมื่อใช้วิธีการวิเคราะห์ด้วยแบบวินิจฉัยโดยผู้เชี่ยวชาญ วิธีการประยุกต์ใช้เทคนิคการประชุม แบบเดลฟายจากกลุ่มผู้เชี่ยวชาญ และวิธีการประยุกต์ใช้เทคนิคโปรโตคอลอะลาว์ด ตัวอย่างที่ใช้ในการ วิจัย คือ ผู้เชี่ยวชาญจำนวน 21 คน และนักเรียนระดับมัธยมศึกษาปีที่ 6 ปีการศึกษา 2556 จำนวน 139 คน ซึ่งได้จากการเลือกตัวอย่างแบบเจาะจง เครื่องมือที่ใช้ในการวิจัยประกอบด้วยแบบวินิจฉัย การทำหน้าที่ต่างกันของข้อสอบจากผู้เชี่ยวชาญ แบบยืนยันการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยประยุกต์เทคนิคการประชุมแบบเดลฟาย แบบสอบถามสำหรับการตรวจสอบความลำเอียงของข้อสอบ สำหรับนักเรียน ชุดข้อสอบสาระการเรียนรู้สุขศึกษาและพลศึกษาสำหรับการตรวจสอบการทำหน้าที่ ต่างกันของข้อสอบสำหรับผู้เชี่ยวชาญ ข้อสอบที่คัดสรรมาได้นำมาผ่านการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยวิธีแมนเทล-แฮนส์เซล ด้วยโปรแกรม DDFS 1.0 และโปรแกรม DIFAS 5.0 พบว่า วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยการตัดสินของ ผู้เชี่ยวชาญที่สำคัญมี 3 วิธี ได้แก่ วิธีที่ 1

การวินิจฉัยการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดย ผู้เชี่ยวชาญ วิธีที่ 2 การประยุกต์ใช้เทคนิคการประชุมแบบเดลฟายจากกลุ่มผู้เชี่ยวชาญ และวิธีที่ 3 การประยุกต์ใช้เทคนิคโปรโตคอลอะลาว์ดและข้อสอบที่ทำหน้าที่ต่างกันด้านเพศของแบบสอบสาระ การเรียนรู้สุขศึกษาและพลศึกษาจากผลการวิเคราะห์เปรียบเทียบอัตราความถูกต้องระหว่างวิธี การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีทางสถิติกับการตรวจสอบโดยผู้เชี่ยวชาญพบว่า วิธีที่ 1 การตรวจสอบด้วยแบบวินิจฉัยโดยผู้เชี่ยวชาญ มีอัตราความถูกต้องโดยเฉลี่ยคิดเป็นร้อยละ 50 และมีอัตราความคลาดเคลื่อนของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยเฉลี่ยคิดเป็นร้อยละ50 วิธีที่ 2 การประยุกต์ใช้เทคนิคการประชุมแบบเดลฟายจากกลุ่มผู้เชี่ยวชาญ มีอัตราความถูกต้อง ตามฉันทามติจากกลุ่มผู้เชี่ยวชาญคิดเป็นร้อยละ 0 และมีอัตราความคลาดเคลื่อนของการตรวจสอบ การทำหน้าที่ต่างกันของข้อสอบคิดเป็นร้อยละ 100 วิธีที่ 3 การประยุกต์ใช้เทคนิคโปรโตคอลอะลาว์ด มีอัตราความถูกต้องโดยเฉลี่ยคิดเป็นร้อยละ 25 และมีอัตราความคลาดเคลื่อนของการตรวจสอบ การทำหน้าที่ต่างกันของข้อสอบโดยเฉลี่ยคิดเป็นร้อยละ 75

พิชชา สุริอาจ และประกฤติยา ทักษิโณ (2559) ศึกษาการพัฒนาแบบวัดความตระหนักต่อโลกในยุคศตวรรษที่ 21 ของนักเรียนมัธยมศึกษาตอนต้น โดยในแบบวัดเชิงสถานการ์ โดยการประยุกต์ใช้การทำหน้าที่ต่างกันของข้อสอบ กลุ่มตัวอย่าง คือ นักเรียนชั้นมัธยมศึกษปีที่ 1-3 สังกัดสำนักงานคณะกรรมการการศึกษาพื้นฐาน จังหวัดนครราชสีมา จำนวน 1,200 คน ได้มาจากการสุ่มแบบหลายขั้นตอน เครื่องมือที่ใช้ในการวิจัย คือ แบบวัดความตระหนักต่อโลกในยุคศตวรรษที่ 21 ของนักเรียนมัธยมศึกษาตอนต้นโดยใช้แบบวัดเชิงสถานการณ์ วิเคราะห์ค่าอำนาจจำแนกของข้อสอบตามทฤษฎีการวัดแบบดั้งเดิม วิเคราะห์ค่าความเที่ยงใช้โปรแกรม SPSS for windows วิเคราะห์ค่าอำนาจจำแนกของข้อสอบตามทฤษฎีตอบสนองข้อสอบใช้โปรแกรม Multilog วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) โดยวิธี Polytomous SIBTEST ใช้โปรแกรม DIFPACK Version 1.7 วิเคราะห์องค์ประกอบเชิงยืนยันอันดับที่สอง ใช้โปรแกรม Mplus พบว่า 1) ผลการพัฒนาแบบวัดความตระหนักต่อโลกในยุคศตวรรษที่ 21 ของนักเรียนมัธยมศึกษาตอนต้น พบว่า แบบวัดเชิงสถานการณ์ มี 6 องค์ประกอบ 14 ตัวบงชี้ คือ 1. ความตระหนักในมุมมองที่แตกต่าง (3 ตัวบ่งชี้) 2. ความตระหนักในสภาพปัจจุบันของโลก (2 ตัวบ่งชี้) 3. ความตระหนักในความแตกต่างของวัฒนธรรม (3 ตัวบ่งชี้) 4. ความตระหนักในเรื่องพลวัตของโลก (2 ตัวบ่งชี้) 5. ความตระหนักต่อทางเลือกของมนุษย์ (2 ตัวบ่งชี้) และ 6. ความตระหนักต่อการเรียนรู้ในการทำงานกับบุคคลที่มีความแตกต่าง (2 ตัวบ่งชี้) ข้อคำถามผ่านเกณฑ์ความตรงเชิงเนื้อหาและทดลองใช้ จำนวน 46 ข้อ 2) ผลการตรวจสอบคุณภาพของแบบวัดความตระหนักต่อโลกในยุคศตวรรษที่ 21 ของนักเรียนมัธยมศึกษาตอนต้นโดยใช้แบบวัดเชิงสถานการณ์ พบว่า 2.1) มีค่าอำนาจจำแนกของข้อสอบตามทฤษฎีแบบดั้งเดิม อยู่ระหว่าง 0.20 ถึง 0.81 และค่าอำนาจจำแนกของข้อสอบตามทฤษฎีตอบสนองข้อสอบ อยู่ระหว่าง 0.15 ถึง 3.22 ค่าความเที่ยง เท่ากับ 0.80 2.2) การทำหนาที่ต่างกันของข้อคำถาม ตามตัวแปรเพศ พบว่า DIF จำนวน 3 ข้อ และ 2.3) ผลการวิเคราะห์องค์ประกอบเชิงยืนยันอันดับที่สองหลังการปรับโครงสร้างแบบวัดหลังการตัด DIF มีความสอดคลองกลมกลืนกับข้อมูลเชิงประจักษ์ มีค่าเท่ากับ 30.629, ค่า df เท่ากับ 27, p-value เท่ากับ 0.2866, /df เท่ากับ 1.13, TLI เท่ากับ 0.998, CFI เท่ากับ 0.999 RMSEA เท่ากับ 0.011 และ SRMR เท่ากับ 0.014

Mendes-Barnett and Ercikan (2006) ได้ศึกษาการทำหน้าที่ต่างกันของข้อสอบในการสอบวิชาคณิตศาสตร์ โดยใช้วิธีซิปเทสท์ โดยใช้ตัวแปรเพศ ผลการวิจัยพบว่าเพศชายมีความสามารถในการแก้ปัญหา และวิธีการทางปัญญาที่สูงเป็นอย่างมากที่จัดไว้ในข้อสอบ ขณะที่เพศหญิงมีความสามารถในด้านการคำนวณสมการ ซึ่งการคำนวณไม่ได้ถูกจัดให้อยู่ในข้อสอบ จึงสรุปได้ว่าข้อสอบวิชาคณิตศาสตร์นี้เกิดการทำหน้าที่ต่างกัน โดยลำเอียงเข้าทางเพศชายมากกว่าเพศหญิง

Breland and Lee (2007) ได้ศึกษาการทำหน้าที่ต่างกันของข้อสอบของการทดสอบปรับเหมาะด้วยคอมพิวเตอร์ที่มีภาษาอังกฤษและภาษาต่างประเทศ ในรายการข้อสอบของ TOEFL-CBT จากกรณีตัวอย่างจำนวน 5,660 ซึ่งใช้วิธีวิเคราะห์สมการถดถอยโลจิสติก สำหรับข้อสอบที่มีการให้คะแนนแบบหลายค่า โดยวิเคราะห์จากตัวแปรที่แยกตามประเภทของข้อสอบคือ ข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูป (Uniform DIF) กับข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูป (Non-Uniform DIF) และตัวแปรเพศ จากการวิจัยพบว่าข้อสอบทำหน้าที่ต่างกันอย่างอเนกรูป มีการทำหน้าที่ต่างกันของข้อสอบการเรียงความ (Writing) ที่เพศหญิงจะได้เปรียบ จึงสามารถสรุปได้ว่า ข้อสอบ TOEFL-CBT มีการทำหน้าที่ต่างกันของข้อสอบอยู่เพียงเล็กน้อยทางด้านตัวแปรเพศ

Barnes and Wells (2009) ได้ศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบของโครงการ National Doctoral Program Survey: (NDPS) ระหว่างตัวแปรเพศและเชื้อชาติ ทางด้านความคิด โดยผลการวิจัยพบว่า ข้อคำถามจำนวน 29 จาก 48 ข้อ พบ DIF โดยที่เพศหญิงที่มีผิวสีมีโอกาสมากขึ้นหรือน้อยลงที่จะมีความเห็นตรงกับเพื่อนชาวผิวขาวเพศชายของเขา และควรมีความระมัดระวังในการใช้ข้อคำถามสำหรับกลุ่มผู้เรียนที่มีความหลากหลาย เช่น เพศ และเชื้อชาติ เป็นต้น จึงสามารถสรุปได้ว่า เพศ และเชื้อชาติมีผลต่อความคิด และการตอบคำถามของผู้เรียนที่ศึกษาในสถานศึกษาที่มีลักษณะบางอย่างแตกต่างกัน

Le (2009) ศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างเพศ การอยู่นอกพื้นที่และการทดสอบภาษาในการทดสอบด้านวิทยาศาสตร์ ใน PISA โดยใช้ข้อมูลการทดลองภาคสนามของ PISA รอบ 3 เพื่อตรวจสอบความสัมพันธ์ระหว่าง การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างเพศ การอยู่นอกพื้นที่และการทดสอบภาษาในการทดสอบด้านวิทยาศาสตร์และรูปแบบอื่น ๆ ที่กำหนดไว้ใน PISA โดยมุ่งเน้นที่บริบทความสามารถและความรู้ทางวิทยาศาสตร์ ข้อมูลที่ใช้ได้รับการรวบรวมจาก 60 กลุ่มภาษาทดสอบโดย 50 ประเทศที่เข้าร่วมด้วยรวมประมาณ 83,000 คน ที่เป็นนักเรียนอายุ 15 ปี งานวิจัยนี้ใช้วิธี IRT เพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างเพศสำหรับแต่ละกลุ่มภาษาและทั่วโลก ข้อสอบแบบเลือกตอบและแบบตอบกลับแบบปิดมีแนวโน้มที่จะเข้าข้างเพศชาย การศึกษายังแสดงให้เห็นถึงผลกระทบของประเทศและการทดสอบภาษากับเพศ การทำหน้าที่ต่างกันของข้อสอบในข้อมูลระหว่างประเทศ ผลการวิจัยพบว่า มีคุณค่าการมีส่วนร่วมในการพัฒนาการทดสอบเพื่อการใช้งานระหว่างประเทศ

Taylor and Lee (2012) ศึกษาการทำหน้าที่ต่างกันของข้อสอบ (DIF) สำหรับการอ่านระดับชั้นประถมศึกษาปีที่ 4 มัธยมศึกษาปีที่ 1 และปีที่ 4 ของข้อสอบคณิตศาสตร์จากการทดสอบตามเกณฑ์ของรัฐ การทดสอบประกอบด้วย ข้อสอบที่มีตัวเลือกหลายตัวเลือกและสร้างการตอบสนองเพื่อตรวจสอบ DIF เมื่อจำแนกตามเพศ โดยใช้ Poly-SIBTEST และขั้นตอน Rasch โดยขั้นตอน Rasch ถูกตั้งค่าสถานะไว้เพิ่มเติมสำหรับข้อสอบที่ DIF มากกว่าที่ทำในข้อสอบพร้อมกันโดยเฉพาะ

ข้อสอบที่มีหลายตัวเลือก สำหรับการอ่านและการทดสอบทางคณิตศาสตร์ทั้งสองแบบ ข้อสอบที่เพศชาย ได้เปรียบในขณะที่รายการตอบสนองการสร้างที่เข้าข้างเพศหญิง การวิเคราะห์เนื้อหาแสดงให้เห็นว่ามีการอ่านค่า อ่านหนังสือที่ถูกตั้งค่าสถานะไว้ การตีความข้อความหรือความหมายโดยนัย เพศชายมีแนวโน้มที่จะได้เปรียบจากสิ่งต่าง ๆ จึงมีการระบุการตีความและการวิเคราะห์ข้อความที่ให้ข้อมูลที่เหมาะสม รายการส่วนใหญ่ที่เป็นที่เข้าข้างเพศหญิงขอให้นักเรียนทำการตีความของตัวเองและวิเคราะห์ทั้งข้อความวรรณกรรมและข้อมูลโดยได้รับการสนับสนุนจากหลักฐานจากข้อความ การวิเคราะห์เนื้อหาของรายการคณิตศาสตร์แสดงให้เห็นว่าข้อสอบที่เข้าข้างเพศชาย คือ เรขาคณิต, ความน่าจะเป็นและพีชคณิต ข้อสอบคณิตศาสตร์ที่เข้าข้างเพศหญิง การตีความทางสถิติ, การแก้ปัญหาหลายขั้นตอนและการให้เหตุผลเชิงคณิตศาสตร์

จากการศึกษางานวิจัยที่เกี่ยวข้อง พบว่า การทำหน้าที่ต่างกันของข้อสอบ (DIF) เป็นการขจัดข้อสอบที่จะก่อให้เกิดการลำเอียงต่อผู้ทำข้อสอบได้ โดยการพิจารณาจากค่าทางสถิติ ที่แสดงให้เห็นว่าข้อสอบข้อใดมีความน่าจะเป็นที่จะก่อให้เกิดความลำเอียงขึ้นบ้าง โดยคุณลักษณะของผู้สอบเป็นส่วนที่ก่อให้เกิดการทำหน้าที่ต่างกัน เช่น เพศ ภาษา และภูมิลำเนา เป็นต้น โดยการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมีด้วยกันหลายวิธี โดยผู้วิจัยได้เลือกใช้วิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่อยู่บนพื้นฐานของทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) คือ วิธี Hierarchical Generalized Linear Model (HGLM) วิธี Multiple Indicators and Multiple Causes Model (MIMIC) และวิธี Item Response Theory – Likelihood Ratio (IRT-LR) ในการนำมาตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสำหรับงานวิจัยครั้งนี้

ตอนที่ 2 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ด้วยวิธี HGLM และงานวิจัยที่เกี่ยวข้อง

การวิเคราะห์ข้อสอบแบบพหุระดับ เกิดขึ้นจากความพยายามที่นักวัดผล ต้องการศึกษาอิทธิพลของตัวแปรภายนอกที่เป็นตัวแปรทางจิตวิทยา ตัวแปรคุณลักษณะผู้สอบให้สามารถประมาณค่าร่วมในโมเดลการรวมกันเชิงเส้นไปพร้อมกับการประมาณค่าพารามิเตอร์ข้อสอบและพารามิเตอร์ผู้สอบ การจากการศึกษาการประมาณค่าพร้อมกันทำให้เกิดผลการวิเคราะห์ที่คลาดเคลื่อน ความพยายามดังกล่าว จึงเริ่มที่การวิเคราะห์แบบสองขั้นตอน คือ การวิเคราะห์ค่าความสามารถของผู้สอบ ให้ผลการวิเคราะห์ตามหลักการของทฤษฎีการตอบสนองข้อสอบ นั่นคือ ค่าพารามิเตอร์ความสามารถของผู้สอบ (θ_j) ต่อมานักวิจัยจึงนำค่า θ_1 เหล่านี้ มาเป็นตัวแปรตามในการวิเคราะห์ถดถอย เพื่อมุ่งหาคำตอบใน 2 ประการหลัก คือ ตัวแปร θ_1 เหล่านี้ มีความแปรผันระหว่างผู้สอบหรือไม่ และหากมีความแปรผันเกิดขึ้น มีตัวแปรใดบ้างที่สามารถอธิบายความผันแปรที่เกิดขึ้นได้ โดยในกรณีนักวิจัยจะนำตัวแปรทางจิตวิทยาหรือตัวแปรคุณลักษณะผู้สอบที่สนใจ เป็นตัวแปรทำนายในสมการถดถอยพหุ แต่นักวัดผลหลายคน เช่น Adam, Wilson, and Wu (1997) ให้ความรู้เกี่ยวกับความคลาดเคลื่อนที่อาจเกิดขึ้น จากการวิเคราะห์แบบ 2 ขั้นตอนในประเด็นหลักคือ 1) ค่าความสามารถของผู้สอบที่ได้จากการประมาณค่าด้วยโมเดลการตอบสนองข้อสอบจะมีความแตกต่างกันของขนาดค่าความคลาดเคลื่อนมาตรฐาน ที่ตำแหน่งความสามารถของผู้สอบที่ต่างกัน การวิเคราะห์ที่ละเลยปัญหาความผันแปรของ

ความคลาดเคลื่อนมาตรฐาน (Heteroschastic Measurement Errors) จะทำให้การวิเคราะห์ 2 ขั้นตอน มีการประมาณค่าที่ไม่คงที่ 2) การประมาณค่าความสามารถของผู้สอบ จะเกิดขึ้นภายหลังจากการประมาณค่าพารามิเตอร์ข้อสอบ ซึ่งจะรับผลจากการประมาณค่าครั้งแรกมาคำนวณต่อจะเกิดความลำเอียงและความไม่คงที่ของการประมาณค่า ซึ่งการวิเคราะห์ลักษณะนี้เป็นปัญหาของโมเดลการวิเคราะห์ถดถอย

ผู้วิจัยจึงได้นำหลักการของ Fisher (1983) ที่เสนอการรวมกันเชิงเส้น (Linear Combination) ที่สามารถดำเนินการได้ ในลักษณะดังกล่าวแบบขั้นตอนเดียว ประกอบกับการพัฒนาสถิติหลายประการ ที่สามารถเอาชนะข้อจำกัด การประมาณค่าพารามิเตอร์ข้อสอบและพารามิเตอร์ผู้สอบไปพร้อมกัน เช่น Bock and Aikin (1981) ได้พัฒนาเทคนิควิเคราะห์แบบ MMLE ขึ้นสำหรับการวิเคราะห์ตามทฤษฎี IRT ซึ่งถือว่าเป็นวิธีการหลักของการประมาณค่าตามทฤษฎีการตอบสนองข้อสอบ ที่มีประสิทธิภาพมาก การประมาณค่าอีกวิธีหนึ่งเกิดจากการศึกษาของ Adam et al. (1997) ที่ได้พัฒนาเทคนิคการวิเคราะห์ที่ชื่อ Random Coefficient Multinomial Logit Model (RCMLM) สามารถกำหนดให้ค่าพารามิเตอร์ผู้สอบเป็นตัวแปรสุ่มและสามารถรวมตัวแปรคุณลักษณะผู้สอบเป็นตัวแปรทำนายในสมการเดียวกันได้ ต่อมาได้พัฒนาเทคนิคการวิเคราะห์กับโมเดลดั้งเดิมได้ด้วย เช่น โมเดลราสช์ทั้งแบบตัวแปรทวิภาคและพหุภาค

Kamata (1998) ได้นำหลักการทางสถิติดังกล่าวมาเสนอรูปแบบการวิเคราะห์ข้อสอบภายใต้มโนทัศน์แบบพหุระดับเป็นคนแรก โดยงานดังกล่าว ได้เสนอเทคนิคทางสถิติที่สามารถวิเคราะห์ได้ด้วยโปรแกรม HLM ภายใต้โมเดลเชิงเส้นตรงทั่วไประดับลดหลั่น (Hierarchical Generalized Linear Model: HGLM) วิเคราะห์ข้อสอบแบบ 2 ระดับ และตรวจสอบความคงที่ของพารามิเตอร์ (Parameter Recovery) ซึ่ง Kamata (2001) ได้เสนอความสมมูลของโมเดล HGLM กับโมเดลราสช์หรือโมเดล IRT แบบ 1 พารามิเตอร์ พิจารณาว่าการตอบข้อสอบของผู้สอบแต่ละคนเป็นโมเดลระหว่างผู้สอบ (Between-student Model) การใช้แนวคิดพื้นฐานนี้เป็นการขยายแนวคิดของโมเดลทฤษฎีการตอบสนองข้อสอบ ว่าเป็นโมเดลพหุระดับที่มีตัวแปรแฝงเป็นตัวแปรตาม

การวิเคราะห์ข้อสอบแบบพหุระดับ

การวิเคราะห์ระดับเมื่อการตอบเป็นแบบทวิภาค คือ การใช้โมเดล HGLM โมเดลเชิงเส้นตรงทั่วไประดับลดหลั่น (HGLM) เป็นโมเดลที่มีลักษณะของการทำงานร่วมกันของ 2 โมเดลหลัก คือ โมเดลเชิงเส้นนัยทั่วไป (Generalized Linear Model: GLM) และโมเดลเชิงเส้นระดับลดหลั่น (Hierarchical Linear and Non–linear Model: HLM) โดยตัวโปรตามในระดับการวิเคราะห์ที่ 1 เป็นตัวแปรทวิภาค โมเดล HGLM จะนำหลักการกระจายแบบ Bernoulli เข้ามาใช้ในการสร้างสมการในระดับการวิเคราะห์ที่ 1 เพื่อให้เกิดการคำนวณทวนซ้ำ (Interactions) ตามโมเดลเชิงเส้นนัยทั่วไป (GLM) ก่อนแล้ว จึงใช้ฟังก์ชั่นการชื่อมโยงหน้าที่แบบโลจิท เข้ามาทำหน้าที่ จะสามารถทำให้เกิดฟังก์ชั่นเชื่อมโยง (Link Function) โดยการแปลงแบบโลจิท ทำให้มีคุณสมบัติตรงตามการวิเคราะห์ถดถอยเชิงเส้นตรง ซึ่งจะมีความต่อเนื่องได้ตั้งแต่ - ถึง + ขึ้นอยู่กับพิสัยของการทำนาย ข้อมูลจากการวิเคราะห์ระดับที่ 1 จึงสามารถนำเข้าสู่การวิเคราะห์ระดับที่ 2 และระดับที่สูงขึ้นไป ได้รายละเอียดของการวิเคราะห์แต่ละขั้นตอนแสดงได้ดังนี้

โมเดลการวิเคราะห์ HGLM ที่สมมูลกับโมเดลราสช์นั้น เป็นการขยายแนวคิดของการตอบแบบทวิภาค หากตัวแปรตาม Y_{ij} เป็นการตอบข้อสอบข้อที่ i (ระดับที่ 1) ของผู้สอบคนที่ j (ระดับที่ 2) สามารถกล่าวได้ว่าความแปรปรวนของตัวแปรตาม Y_{ij} เป็นการกระจายแบบไบไมเมียล (Binomail Distribution) ค่าคาดหมายของความน่าจะเป็นที่จะตอบข้อสอบได้ถูกต้องข้อที่ i ของผู้สอบคนที่ j เขียนได้ ดังนี้

$$E\left(Y_{ij} / P_{ij}\right) = P_{ij} \tag{3}$$

โดยมีความแปรปรวนเท่ากับ $Var\left(Y_{ij} / P_{ij}\right) = P_{ij}\left(1 - P_{ij}\right)$ (4)

เมื่อ P_{ij} แทน ความน่าจะเป็นบุคคลที่ j (j=1 ถึงคนที่ n) สามารถทำข้อสอบที่ i ได้ถูกต้อง

เมื่อมีการกระจายเป็นแบบไบโนเมียล สามารถเลือกใช้ฟังก์ชั่นการเชื่อมโยงหน้าที่ได้หลายประเภท (Raudenbush & Bryk, 2002, pp. 29-37) ได้เสนอให้ใช้ฟังก์ชั่นการเชื่อมโยงหน้าที่แบบโลจิท ดังนั้นสามารถเขียนเป็นสมการระการวิเคราะห์ระดับที่ 1 คือ

$$\eta_{il} = \log\left(\frac{P_{il}}{1 - P_{il}}\right) \tag{5}$$

เมื่อแทน η_{il} แทน ค่าลอคของออดส์ (Odds) ที่จะตอบข้อสอบข้อที่ i ได้ถูกต้องของผู้สอบคนที่ j จึงสามารถเขียนเป็นสมการโครงสร้างของระดับการวิเคราะห์ระดับที่ 1 ได้

$$\eta_{il} = \beta_{0i} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \ldots + \beta_{kj}X_{kij} \tag{6}$$

$$= \beta_{0j} + \sum_{q=1}^{k} \beta_{qj}X_{qij} \tag{7}$$

เมื่อ X_{qij} เป็นตัวแปรดัมมี่ที่ q สำหรับบุคคลที่ j ซึ่งสามารถพิจารณาได้ว่า X_{qij} เป็นตัวแปรอิทธิพลของรายข้อ เพื่อให้เป็นไปตามข้อตกลงเบื้องต้นของสมการพหุระดับที่โมเดลจะต้องเป็นเมตริกซ์เอกลักษณ์ (Identity Matrix) จึงมีความจำเป็นต้องกำหนดให้ตัวแปรดัมมี่ ที่ X_{qij} ตัวใดตัวหนึ่งเป็นศูนย์ ทำให้เกิดเมตริกซ์แบบเต็มอันดับ (Full Rank) โดยนิยมตัด (Drop) ข้อสอบข้อสุดท้ายเพราะมีความสะดวกในการวิเคราะห์

เมื่อ $\eta_{ij} = \log\left(\frac{P_{il}}{1 - P_{il}}\right) = \beta_{0j} + \beta_{qj}$ (8)

ดังนั้นความน่าจะเป็นของบุคคลที่ j จะตอบข้อสอบที่ i ได้ถูกต้องจึงเท่ากับ

$$P_{ij} = \frac{1}{1 + \exp\left(-\eta_{ij}\right)} \text{ หรือ } P_{ij} = \frac{1}{1 + e^{-\eta_{ij}}} \tag{9}$$

การวิเคราะห์ในระดับที่ 2 (Level 2) ค่าความยากของข้อสอบจึงเป็นอิทธิพลคงที่กับกลุ่มผู้สอบทั้งหมด แต่จะเป็ยอิทธิพลแบบสุ่มไปตามข้อสอบแต่ละข้อ สามารถแสดงสมการ การวิเคราะห์ในระดับที่ 2 ได้ดังนี้

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{10}$$

$$\beta_{1j} = \gamma_{10} \tag{11}$$

.

.

.

$$\beta_{(k-1)i} = \gamma_{k-1)}0 \tag{12}$$

เมื่อ u_{0j} เป็นองค์ประกอบแบบสุ่มของ Intercept β_{0j} ซึ่งเป็นค่าที่แสดงความสามารถของผู้สอบคนที่ j หากรวมสมการระดับที่ 1 และระดับที่ 2 เข้าด้วยกัน สามารถแสดงสมการความน่าจะเป็นที่บุคคลที่ j จะตอบข้อสอบข้อที่ i ได้ดังนี้

$$\eta_{il} = \gamma_{00} + \gamma_{qo} + u_{oj} \tag{13}$$

$$P_{ij} = \frac{1}{1+\exp\left\{\left[u_{oj} - \left(-\gamma_{qo} - \gamma_{oo}\right)\right]\right\}} \tag{14}$$

จากสมการข้างต้น Kamata (2001) แสดงให้เห็นว่าเป็นสมการคู่ขนาน (Equivalent) กับสมการความน่าจะเป็นที่บุคคลที่ j จะตอบข้อสอบข้อที่ i ได้ถูกต้องของโมเดลราสช์

โมเดลราสช์

$$P_{ij} = \frac{\exp[\theta_j - \delta_i]}{1+\exp[\theta_j - \delta_i]} = \frac{1}{1+\exp[-\theta_j - \delta_i]} \tag{15}$$

โมเดลเชิงเส้นตรงทั่วไประดับลดหลั่น 2 ระดับ (HGLM)

$$P_{ij} = \frac{1}{1+\exp\{-[u_{0j} - (-\gamma_0 - \gamma_{00})]\}} \tag{16}$$

การประยุกต์วิเคราะห์ข้อสอบแบบพหุระดับด้วยโมเดลเชิงเส้นตรงทั่วไประดับลดหลั่น 2 ระดับ (HGLM-2L)

จากแนวคิดการวิเคราะห์ข้อสอบแบบพหุระดับ (HGLM-2L)

ระดับการวิเคราะห์ที่ 1 ระดับข้อสอบ (Between Items Within Person) ผลจากการวิเคราะห์จะทำให้ได้สารสนเทศพารามิเตอร์ของข้อสอบ $\left(\delta_i\right)$ และพารามิเตอร์ผู้สอบ $\left(\theta_I\right)$

ระดับการวิเคราะห์ที่ 1 ระดับข้อสอบ

เป็นการวิเคราะห์อิทธิพลระหว่างการตอบสนองข้อสอบทั้งฉบับที่สอดแทรกอยู่ในผู้สอบแต่ละคน ในแต่ละโรงเรียน โดยอิทธิพลของข้อสอบจะมีความคงที่กับผู้สอบแต่ละคน แต่จะมีความแปรผันแบบสุ่มไปตามข้อสอบแต่ละข้อ แสดงรายละเอียดที่ผู้สอบคนที่ j ในโรงเรียนที่ m จะสามารถตอบข้อสอบข้อที่ i ได้ถูกต้อง ดังสมการ

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\eta_{ijm} \tag{17}$$

$$=\beta_{ojm}+\beta_{1jm}x_{1jm}+\beta_{2jm}+\ldots+\beta_{(k-1)jm}x_{(k-1)jm} \tag{18}$$

$$=\beta_{oj}+\sum_{q-1}^{k-1}\beta_{qjm}x_{ojm} \tag{19}$$

$$=\beta_{oj}+\beta_{ojm} \tag{20}$$

เมื่อ x_{ojm} แทน ตัวแปรดัมมี่ที่ q สำหรับข้อสอบข้อที่ i ผู้สอบที่ j ในโรงเรียน m ซึ่งมีค่าเป็น 1 เมื่อ q = i และมีค่าเป็น 0 เมื่อ q≠ i

β_{ojm} แทน ค่าจุดตัดแกนตั้ง (Intercept) หรืออิทธิพลของข้อสอบข้อที่ Drop ซึ่งเรียกว่า ข้อสอบอ้างอิง (Refference)

β_{qjm} แทน ค่าสัมประสิทธิ์ตัวแปรดัมมี่ของข้อสอบข้อที่ q เมื่อเปรียบเทียบกับข้อสอบข้ออ้างอิงของผู้สอบคนที่ j ในโรงเรียน m

ระดับการวิเคราะห์ที่ 2 ระดับผู้สอบ

เป็นการวิเคราะห์ค่าความสามารถเฉพาะในการตอบข้อสอบระหว่างผู้สอบ และค่าอิทธิพลของข้อสอบรายข้อสำหรับผู้สอบภายในโรงเรียน ค่าอิทธิพลของข้อสอบ (Item Effect) จึงมีค่าคงที่ระหว่างโรงเรียน แต่ผันแปรแบบสุ่มไปตามรายข้อสอบแต่ละข้อ ได้ดังนี้

$$\beta_{ojm}=\gamma_{00m}+u_{ojm} \tag{21}$$

$$\beta_{1jm}=\gamma_{10m} \tag{22}$$

$$\beta_{2jm}=\gamma_{20m} \tag{23}$$

.

.

.

$$\beta_{(k-1)jm}=\gamma_{(k-1)om} \tag{24}$$

เมื่อ γ_{00m} เป็นค่า Intercept ของ β_{ojm} คือ เป็นค่าเฉลี่ยอิทธิพลของข้อสอบข้ออ้างอิงต่อโอกาสในการตอบข้อสอบถูกในโรงเรียนที่ m

u_{ojm} เป็นค่าส่วนที่เหลือของ β_{ojm} คือ เป็นค่าส่วนเบี่ยงเบนของโอกาสในการตอบข้อสอบถูกต้องคนที่ j จากค่าเฉลี่ยโอกาสในการตอบข้อสอบถูกในโรงเรียนที่ m ซึ่งถือเป็นค่าความสามารถเฉพาะของผู้สอบคนที่ j ในโรงเรียนที่ m มีการแจกแจงเป็นโค้งปกติ ค่าเฉลี่ยเท่ากับศูนย์ และความแปรปรวนเท่ากับ

$$\tau[r_{ojm} \sim N(0, \tau)]$$

ในการวิเคราะห์ด้วยโปรแกรม HLM จะรายงานผลในไฟล์ส่วนที่เหลือ (Residual File) โดยรายงานค่าความสามารถของบุคคลที่สอบ ด้วยการวิเคราะห์จากสถิติ EB และ OLS ซึ่งวิเคราะห์ข้อมูลระดับที่ 1 เป็นการวิเคราะห์เพื่อให้ได้สาระสนเทศค่าพารามิเตอร์ความยากของแบบสอบ ในขั้นตอนนี้จะเป็นการวิเคราะห์ด้วยโมเดลเชิงเส้นทั่วไป (HGLM) ไม่มีการเพิ่มตัวแปรทำนายในระดับการวิเคราะห์นี้ การวิเคราะห์ข้อมูลระดับที่ 2 เป็นการวิเคราะห์เพื่อให้ได้สารสนเทศค่าพารามิเตอร์ความสามารถของผู้สอบ ในขั้นตอนนี้เป็นการวิเคราะห์ด้วยโมเดล HGLM สามารถพิจารณาเพิ่มตัวแปรทำนายระดับผู้สอบในสมการการวิเคราะห์ได้ โดยเพิ่มในสมการแรกของการวิเคราะห์ซึ่งก็คือสมการของแบบสอบข้ออ้างอิง (Reference Item) ที่ได้ตัดออกเพื่อทำให้เป็นเมตริกซ์เอกลักษณ์ตามข้อตกลงเบื้องต้นของการวิเคราะห์ข้อมูลด้วยโปรแกรมโมเดลเชิงเส้นตรงระดับลดหลั่น

การวิเคราะห์โมเดล HGLM ด้วยโปรแกรมสำเร็จรูป HLM

การวิเคราะห์ข้อมูลที่เป็นพหุระดับ (Multilevel Data) หากข้อมูลมีลักษณะโครงสร้างไม่เป็นเชิงเส้นตรง (Nonlinear Structural) และมีการกระจายของความคลาดเคลื่อนที่ไม่เป็นโค้งปกติ (Nonnormally Distributed Error) การวิเคราะห์ด้วยโมเดลเชิงเส้นระดับลดหลั่น (HLM) อาจจะเกิดความไม่เหมาะสมในการวิเคราะห์ เพราะการแปลความหมายและการประมาณค่าอาจจะเกิดความผิดพลาด ดังนั้นโมเดลที่เหมาะสมกว่าและข้อมูลที่มีลักษณะเป็นแบบแบ่ง 2 ส่วน (Binary Response) ควรทำการวิเคราะห์ด้วยโมเดลเชิงเส้นตรงทั่วไประดับลดหลั่น (Hierarchical Generalized Linear Model: HGLM) มากกว่าการวิเคราะห์ด้วย HLM (Raudenbush & Bryk, 2002; Kamata, 2001)

Raudenbush and Bryk (2002, pp. 185-186) ได้กล่าวถึงลักษณะของโมเดลการวิเคราะห์เชิงเส้นตรงระดับลดหลั่นว่ามีองค์ประกอบหลัก คือ โมเดลการสุ่ม (Sampling Model) โมเดลการเชื่อมโยงหน้าที่ (Link Function Model) และโมเดลโครงสร้าง (Structural Model) โดยแสดงความสัมพันธ์ของโมเดล HGLM และ HLM ดังนี้

ตารางที่ 2-3 ความสัมพันธ์ของหลักการวิเคราะห์ของสมการแบบ HLM และ HGLM

สมการ	โมเดลการสุ่ม (Sampling Model)	โมเดลการเชื่อมโยงหน้าที่ (Link Function Model)	โมเดลโครงสร้าง (Structural Model)
HLM	ตัวแปรตามเป็นตัวแปรต่อเนื่อง การกระจายของตัวแปรเป็นการกระจายแบบโค้งปกติ มีค่าเฉลี่ยเท่ากับ μ_{ij} และกระจายเท่ากับ σ^2 เขียนเป็นสมการ ดังนี้ $Y_{ij}\left\|\mu_{ij}\right. \sim NID\left(\mu_{ij}\sigma^2\right)$	การวิเคราะห์ด้วย HLM ลักษณะทั่วไปไม่มีความจำเป็นต้องเปลี่ยนแปลงค่าดังกล่าว แต่ก็สามารถใช้ฟังก์ชันแบบ Logit link ได้ ($\eta_{ij} = \mu_{ij}$ = Identity Link Function)	การเปลี่ยนค่าของตัวทำนายเป็น η_{ij} จะมีความสัมพันธ์กับตัวแปรทำนายต่าง ๆ ในโมเดลสามารถแสดงในรูปแบบสมการเชิงเส้นตรง ดังนี้ $\eta_{ij} = \beta_{0i} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \beta_{pj}X_{pij}$
HGLM	ตัวแปรตามจะเป็นการตอบแบบทวิภาค (0, 1) การกระจายจึงเป็นแบบไบโนเมียลซึ่งกรณีหนึ่งของการกระจายแบบ Bernoulli $\left(Y_{ij}\left\|\varphi_{ij}\right. \sim B\left(m_{ij},\sigma^2\right)\right)$	การเชื่อมโยงหน้าที่ (Link Function) ในโมเดลนี้จะใช้ Logit Function หรือ Logit Link เขียนเป็นสมการ ดังนี้ $\eta_{ij} = \log\left(\frac{\varphi_{ij}}{1-\varphi_{ij}}\right)$ เมื่อ η_{ij} คือค่า log Odds ที่จะประสบความสำเร็จในการตอบข้อสอบข้อที่ i	การประมาณค่า βs จากสมการในโมเดล โครงสร้างของ HLM ก่อให้เกิดการทำนาย Log Odds ก็สามารถแปลงค่ากลับเป็นค่า Odds ได้ดังค่าเดิมโดยคูณค่า $\exp\left(\eta_{ij}\right)$ $\varphi_{ij} = \frac{1}{1+\exp\left(-\eta_{ij}\right)}$

จากความสัมพันธ์เชิงโครงสร้างสมการทั้งสองโมเดลของ Sampling Model, Link Function และ Structural Model จะเห็นได้ว่าสมการวิเคราะห์ด้วยโมเดลการวิเคราะห์ HLM จัดเป็นกรณีเฉพาะ (Special Case) ของการวิเคราะห์แบบ HGLM โดยแตกต่างกันที่ประเภทของตัวแปรตามเป็นปัจจัยสำคัญ

งานวิจัยที่เกี่ยวข้องกับวิธี HGLM มีดังนี้

Seo (2009) ศึกษาผลกระทบจากสภาพภูมิอากาศในวิทยาลัยที่มีต่อการดื่มสุราของนักเรียน โดยใช้โมเดลการวิเคราะห์เชิงเส้นลำดับชั้นทั่วไป (HGLM) ซึ่งทำการศึกษาจากผลการประเมินสุขภาพจากกลุ่มตัวอย่าง จำนวน 76,542 คน จาก 113 วิทยาลัยในสหรัฐอเมริกา วิเคราะห์ผลโดยวิธีของฮอกซ์ ที่มี 5 ขั้นตอน จากการศึกษาพบว่า สภาพภูมิอากาศมีผลทำให้นักเรียนในวิทยาลัยดื่มสุรากันมากขึ้น โดยส่วนใหญ่เปอร์เซ็นของนักเรียนที่ดื่มสุราจะเป็นเพศชาย ซึ่งการตรวจสอบในลำดับขั้นสามารถตรวจพบตัวแปรแฝงในการดื่มสุราของนักเรียนได้อีกหลายอย่าง นอกเหนือจากสภาพภูมิอากาศ คืออายุและเชื้อชาติ ที่เป็นเหตุผลของการชักชวนกันดื่มสุรากันมากยิ่งขึ้น

Acar and Kelecioglu (2010) ศึกษาเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ระหว่างวิธี HGLM วิธี LR และวิธี IRT-LR โดยมีวัตถุประสงค์การวิจัยครั้งนี้เป็นการตรวจสอบความสอดคล้องระหว่างวิธีที่ใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ 3 วิธี ซึ่งมีเงื่อนไขเป็นเพศในการตรวจสอบ มีกลุ่มตัวอย่างเป็นนักเรียนในประเทศตุรกี และเครื่องมือที่ใช้ในการวิจัยคือแบบทดสอบของวิชาสังคมศาสตร์และวิทยาศาสตร์ จากการศึกษาพบว่า วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้ง 3 วิธี คือ วิธี HGLM วิธี LR และวิธี IRT-LR มีการตรวจพบข้อสอบที่ทำหน้าที่ต่างกันในปริมาณที่ใกล้เคียงกัน แต่ไม่ตรงกัน โดยวิธี LR และวิธี IRT-LR จะตรวจพบข้อสอบที่ทำหน้าที่ต่างกันน้อยทั้งแบบทดสอบสังคมศาสตร์และวิทยาศาสตร์ ขณะที่วิธี HGLM จะตรวจสอบพบข้อสอบที่ทำหน้าที่ต่างกันในแบบทดสอบทั้ง 2 แบบทดสอบในปริมาณที่มากที่สุด

Acar (2013) ศึกษาเปรียบเทียบค่าสัมประสิทธิ์ความคล้ายคลึงกันระหว่างกลุ่มจากการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี HGLM และวิธี LR กลุ่มตัวอย่างได้รับการศึกษาที่ดำเนินการผ่านวิธีการสุ่มตัวอย่างแบบแบ่งชั้นสุ่มและข้อมูลที่ได้รวบรวมมาจากนักเรียน 10,727 ข้อมูลที่ใช้ในการศึกษาครั้งนี้ประกอบด้วย 25 รายการในรายการย่อยประเทศตุรกี SSIE ที่ถูกนำมาใช้ในประเทศตุรกีในปี 2006 การสืบสวนเชิงประจักษ์เปรียบเทียบวิธี HGLM-DIF กับถดถอยโลจิสติกวิธี DIF ได้จัดให้มีหลักฐานยืนยันว่าขั้นตอน HGLM-DIF เทียบเท่ากับวิธี LR-DIF ความคล้ายคลึงกันระหว่างทั้งสองวิธีมีการระบุในแง่มุมต่าง ๆ ของผลรวมทั้งความสัมพันธ์ของกลุ่มและการตัดค่าสัมประสิทธิ์ ความสัมพันธ์ของกลุ่มและการตัดค่าสัมประสิทธิ์จากทั้งสองวิธีที่ค่อนข้างสมบูรณ์แบบ HGLM-DIF อาจจะขอแนะนำสำหรับการวิเคราะห์ DIF ถ้าตัวแปรผลเป็นสองค่าเพราะขั้นตอน HGLM-DIF ไม่ต้องใช้ความพยายามมากขึ้นและใช้เวลามากกว่าขั้นตอน LR-DIF

Ong, Lu, Lee, and Cohen (2015) ได้ศึกษาเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ระหว่างวิธี HGLM วิธี MIMIC และวิธี IRT ในการทดสอบแบบจำลองข้อมูลเชิงลำดับชั้น โดยมีเงื่อนไขเป็นขนาดของกลุ่มตัวอย่าง โดยเปรียบเทียบที่อัตราความคลาดเคลื่อนประเภทที่ 1 (Type I Error) พบว่า วิธี MIMIC มีการควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ดีกว่าวิธี HGLM และวิธี IRT หากมีกลุ่มตัวอย่างที่เล็ก เพราะหากเมื่อขนาดของกลุ่มตัวอย่างเพิ่มขึ้น อัตราความคลาดเคลื่อนก็เพิ่มขึ้นด้วย และวิธี IRT ก็สามารถควบคุมความคลาดเคลื่อนได้ดีหากมีกลุ่มตัวอย่างที่มีขนาดเล็ก แต่เมื่อกลุ่มตัวอย่างมีขนาดที่ใหญ่ขึ้นก็เกิดความผิดพลาดของความคลาดเคลื่อนเช่นเดียวกัน จากผลวิจัยสรุปได้ว่า วิธี MIMIC เมื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแล้วสามารถควบคุมอัตราความคลาดเคลื่อนได้ดีเมื่อกลุ่มตัวอย่างมีขนาดไม่ใหญ่มาก และวิธี HGLM ก็เป็นวิธีที่สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแล้วพบข้อสอบที่ทำหน้าที่ต่างกันมากที่สุด

จากการศึกษางานวิจัยที่เกี่ยวข้องพบว่า วิธี HGLM เป็นการวิเคราะห์ด้วยโมเดลสมการโครงสร้างเชิงเส้นตรงระดับลดหลั่น เมื่อทำการวิเคราะห์ในแบบทดสอบที่มีความยาวตั้งแต่ 20 ข้อขึ้นไป วิธี HGLM สามารถตรวจพบจำนวนข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบได้มากกว่าวิธี IRT วิธี SIBTEST และวิธี MIMIC และสามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดี เมื่อกลุ่มตัวอย่างมีขนาดใหญ่ คือ 1,000 คน ขึ้นไป

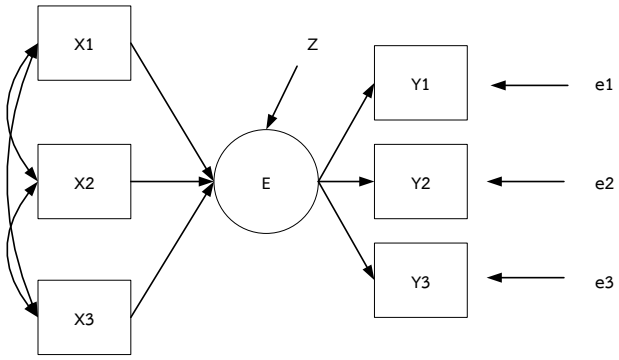
ตอนที่ 3 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ด้วยวิธี MIMIC และงานวิจัยที่เกี่ยวข้อง

ปัจจุบันการศึกษาความสัมพันธ์ระหว่างตัวแปร ในรูปแบบโครงสร้างสมการเส้นตรง ถูกนำมาใช้กันอย่างแพร่หลาย ไม่ว่าจะเป็นการศึกษาความสัมพันธ์ระหว่างตัวแปรสังเกตได้กับตัวแปรสังเกตได้ ตัวแปรแฝง (Latent Variables) กับตัวแปรสังเกตได้ หรือตัวแปรแฝงกับตัวแปรแฝง โดยเฉพาะการตรวจสอบและปรับทฤษฎี เช่น การวิเคราะห์ถดถอยพหุ การวิเคราะห์ความสัมพันธ์เชิงสาเหตุ (Path Analysis) การวิเคราะห์องค์ประกอบเชิงยืนยัน (Comfirmatory Factor Analysis) การววิเคราะห์ความสัมพันธ์เชิงโครงสร้างระหว่างตัวแปรแฝงตามระบบสมการเชิงเส้น (Structural Equation Modeling) การวิเคราะห์โมเดลสมการโครงสร้าง (Structural Equation Model: SEM) ส่วนใหญ่นิยมใช้การวิเคราะห์องค์ประกอบเชิงยืนยัน และวิเคราะห์ความสัมพันธ์เชิงสาเหตุ

แต่ยังมีวิธีวิเคราะห์ข้อมูลอีกหลากหลายวิธี หนึ่งในนั้นคือ การวิเคราะห์โมเดลมิมิค (Multiple Indicators and Multiple Causes Model: MIMIC Model) เป็นหนึ่งในทางเลือกที่จะใช้ในการวิเคราะห์ข้อมูลที่สามารถใช้ประโยชน์จากโปรแกรมวิเคราะห์ได้อย่างเต็มประสิทธิภาพ (รติพร ถึงฝั่ง, 2556) ในการวิเคราะห์ข้อมูลด้วยแบบจำลองสมการโครงสร้างนั้นสามารถวิเคราะห์ได้ด้วยหลากหลายโปรแกรม เช่น Amos, EQS, LISREL และ Mplus (Bowen, 2011, p. 5) โปรแกรมลิสเรล (LISREL) เป็นโปรแกรมที่ให้ผู้ทำวิจัยได้ศึกษาความสัมพันธ์ระหว่างตัวแปรแฝงด้วยกันเพื่อที่จะได้ทดสอบเนื้อหาของทฤษฎี และวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรแฝงกับตัวชี้วัดเพื่อทดสอบคุณภาพของการวัดอีกด้วย สุชาต ประสิทธิ์รัฐสินธุ์, กรรณิการ์ สุขเกษม, โศภิต ผ่องศรี และถนอมรัตน์ ประสิทธิเมตต์ (2551, หน้า 3) และโปรแกรม Mplus เป็นโปรแกรมถูกพัฒนาขึ้นเพื่อให้นักวิจัยใช้สำหรับการวิเคราะห์ข้อมูลด้วยสถิติวิเคราะห์ขั้นสูง และต่อมาได้ถูกพัฒนาให้โปรแกรมมีการใช้งานได้ง่ายและสะดวก

โมเดลมิมิค (MIMIC Model)

MIMIC เป็นคำที่ย่อมาจาก Multiple Indicators and Multiple Causes ซึ่งหมายถึง โมเดลมีตัวแปรแฝงเพียงตัวแปรเดียว โดยที่ตัวแปรแฝงนั้น ได้รับอิทธิพลจากตัวแปรภายนอกสังเกตได้หลายตัวแปร และส่งอิทธิพลไปยังตัวแปรภายในสังเกตได้หลายตัวแปร กล่าวอีกอย่างหนึ่งคือ เป็นโมเดลของคุณลักษณะแฝงที่มีหลายสาเหตุและวัดได้จากตัวบ่งชี้หลายตัว ดังแสดงดังภาพที่ 2-5 ในที่นี้มีตัวบ่งชี้ 3 ตัวแปร และมีตัวแปรสาเหตุ 3 ตัวแปรตามลักษณะ โมเดลจะเห็นว่าการวัดตัวแปรภายนอกสังเกตได้ ต้องมีข้อตกลงข้างต้นว่า ไม่มีความคลาดเคลื่อนในการวัด และในการวิเคราะห์ข้อมูลจะกำหนดข้อมูลจำเพาะ เฉพาะรูปแบบและสถานะของเมทริกซ์ PH, BE, GA, PS, LY และ TE เท่านั้น ส่วนเมทริกซ์ TD และ LX มีค่าเป็นศูนย์ทั้งหมด โมเดลมิมิคนี้เป็นประโยชน์มากในการตรวจสอบความเป็นเอกมิติ (Unidimensionality) ในการวิจัยสาขาการวัดผลการศึกษา แสดงดังภาพที่ 2-5



ภาพที่ 2-5 โมเดลย่อยของ MIMIC (Schumacker & Lomax, 2010, p. 294)

การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ (DIF) ด้วยวิธี MIMIC

วิธี MIMIC ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ (DIF) สามารถแบ่งออกเป็นการวัดองค์ประกอบและโครงสร้างองค์ประกอบ ในองค์ประกอบการวัด ของข้อสอบที่ i ลักษณะของตัวแปรแฝง y ที่ทดสอบเป็นการออกแบบการวัด และกลุ่มของตัวแปร z (ในที่นี้เป็นการศึกษาเพียง 1 กลุ่มตัวแปร) ที่เกี่ยวข้องกับการทำหน้าที่ต่างกันของข้อสอบ (DIF) ในการวิเคราะห์องค์ประกอบของโมเดล ดังนี้

สูตรสำหรับวิธี MIMIC ของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) (Muthen et al., 1991) คือ

$$y_i^* = \lambda_i \theta + \beta_i^{'} z + \varepsilon_{i'} \tag{25}$$

เมื่อ y_i^* คือ ข้อที่ i

θ คือ องค์ประกอบ

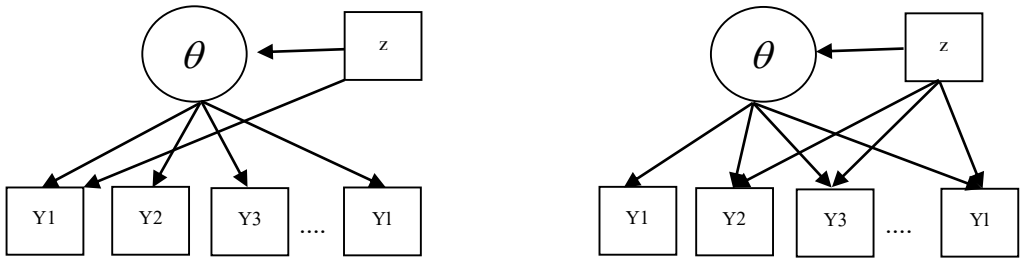
$\beta_i^{'}$ คือ สัมประสิทธิ์ของตัวแปรเพศ

Z คือ กลุ่มเพศ

λ_i คือ น้ำหนักองค์ประกอบ

ε_i คือ ค่าความแปรปรวน

เมื่อ λ_i เป็นน้ำหนักองค์ประกอบและเกี่ยวข้องกับความชันของพารามิเตอร์ของข้อสอบข้อที่ iในบริบทของทฤษฎีการตอบสนองข้อสอบ (IRT) แล้ว ε_i มีการแจกแจงแบบปกติสำหรับ Ordinal Probit และการแจกแจงแบบโลจิสติก สำหรับ Ordinal Logit และ $\beta_i^{'}$ คือ อิทธิพลของกลุ่มตัวแปร z ต่อ y_i^* ถ้า $\beta_i^{'}$ = 0 แล้วข้อสอบข้อที่ i มีค่าเท่ากันในทุก ๆ กลุ่ม ตรงกันข้าม ถ้า $\beta_i^{'}$ ≠ 0 จะเกิดการทำหน้าที่ต่างกันของข้อสอบ (DIF) แบบอเนกรูป เนื่องจากสมการไม่มีเทอมปฏิสัมพันธ์เป็นตัวทำนาย (θZ) ดังสมการ MIMIC จึงใช้แบบเอกรูปได้เพียงอย่างเดียว ดังแสดงในภาพที่ 2-6



(a) (b)

ภาพที่ 2-6 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี MIMIC แบบเอกรูป (Wang & Shih, 2010, p. 169)

ซึ่งวิธีหลายตัวชี้วัดหลายสาเหตุในรูปแบบองค์ประกอบเชิงยืนยัน (MIMIC) เป็นหลักการของ CFA กับตัวแปร แล้ววิธีMIMIC ยังสามารถนำไปใช้สำหรับการวิเคราะห์ DIF ได้ด้วย Muthen et al. (1991) ซึ่งผลที่ได้ต้องมีค่าเป็นแบบ 2 ค่า (Dichotomous) ค่าพารามิเตอร์ของตัวชี้วัดไม่ต่อเนื่องเป็นสิ่งที่จำเป็น ในความเป็นจริงแล้ว มีหลายวิธีที่ตัวชี้วัดของค่าพารามิเตอร์เป็นแบบ 2 ค่า (Dichotomous) โดยใช้ฟังก์ชั่นเชื่อมโยงที่เหมาะสม (เช่น การเชื่อมโยงแบบโลจิทหรือโพรบิต) ข้อตกลงเบื้องต้นคือ ตัวแปรแฝงเป็นตัวแปรต่อเนื่องและตัวแปรสังเกตได้เป็นการตอบแบบไบนารี (Binary) เมื่อ y^*_{ij} เป็นตัวแปรแฝงแบบต่อเนื่องและตัวแปรสังเกตได้เป็นการตอบแบบไบนารี (Binary) ของข้อสอบแล้วสามารถเขียนสมการได้ดังนี้

$$y_{ij} = \begin{cases} 1 \; if \; y^*_{ij} > \\ 0 \; if \; y^*_{ij} \leq 0 \end{cases} \tag{26}$$

สูตรสำหรับวิธี MIMIC ของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) (Muthen et al., 1991) คือ

$$y^*_{ij} = \lambda_i \theta_j + \beta_i G_j + \varepsilon_{ij}, \tag{27}$$

เมื่อ λ_i เป็นน้ำหนักองค์ประกอบของข้อที่ iและ θ_j เป็นลักษณะของตัวแปร ส่วน β_i เป็นสัมประสิทธิ์ความชันสำหรับความแปรปรวนร่วม G_i ซึ่งเป็นกลุ่มตัวชี้วัดของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) และ ε_{ij} เป็นเศษเหลือ นอกจากนี้โมเดลการถดถอยเป็นสิ่งจำเป็นสำหรับการพยากรณ์ตัวแปรแฝง θ โดยกลุ่มของตัวชี้วัด G_j เพื่อควบคุมความแตกต่างในลักษณะตัวแปรแฝงข้ามกลุ่มย่อย

$$\eta_j = yz_j + \zeta_{j'} \tag{28}$$

เมื่อ γ เป็นความชันของกลุ่มตัวแปร G_j และ ζ_j เป็นความคลาดเคลื่อนของสมการถดถอย β_i เป็นการทำหน้าที่ต่างกันของข้อสอบ (DIF) แบบเอกรูป เมื่อ γ เป็นผลต่างของค่าเฉลี่ยคุณลักษณะแฝงกลุ่มเปรียบเทียบกับกลุ่มอ้างอิงและมีเกณฑ์การจับคู่ ตามตัวแปรคงที่ในสมการข้างต้นมีข้อตกลงกำหนดให้เป็น 0 ซึ่งจะไม่ปรากฏในสมการข้างต้น

$$a_i = \frac{\lambda_i \sqrt{\sigma_\eta^2}}{\sqrt{1-\lambda_i^2}\sigma_\zeta^{2'}} \tag{29}$$

$$b = \frac{[(\tau_i - \beta_i z)\lambda_i^{-1} - \mu_n]}{(\sigma_n^2)^{1/2}} \tag{30}$$

เมื่อ σ_η^2 เป็นตัวแปรสำหรับองค์ประกอบ θ_j และ σ_ζ^2 เป็นตัวแปรของความคลาดเคลื่อนของสมการถดถอยเชิงเส้นตรง ζ_i สำหรับการทำนายองค์ประกอบทั่วไป τ_i เป็นความยากของข้อสอบข้อที่ i และ μ_η เป็นค่าเฉลี่ยขององค์ประกอบทั่วไป θ_j

มีข้อดีหลายประการของการใช้โมเดล MIMIC ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ที่แสดงข้างต้น (Muthen et al. 1991) แสดงขนาดของการทำหน้าที่ต่างกันของข้อสอบ โดยใช้หลักทฤษฎีการตอบสนองข้อสอบ (IRT) ประมาณค่าการทำหน้าที่ต่างกันของข้อสอบจากค่าพารามิเตอร์ตามทฤษฎีการตอบสนองข้อสอบ (IRT) ซึ่งมีประโยชน์กับผู้ปฏิบัติ

Finch (2005) เปรียบเทียบประสิทธิภาพของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธี MIMIC กับวิธี MH (Mantel and Haenszel, 1959) และวิธี SIBTEST (Shealy and Stout, 1993) และวิธี IRT-LR (Thissen, Steinberg, and Gerrard, 1986) กับความคลาดเคลื่อนประเภทที่ 1และอำนาจการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ได้แสดงให้เห็นว่าวิธี MIMIC มีค่าสูงขึ้นและความคลาดเคลื่อนประเภทที่ 1 มีค่าลดลงกับจำนวนข้อสอบ 50 ข้อ นอกจากนี้วิธี MIMIC ยังสามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบ Uniform DIF ได้ด้วย และได้มีการประเมินประสิทธิภาพการทำงานของวิธี MIMIC แบบละเมิดข้อตกลงเบื้องต้น โดยการทำชุดข้อสอบ พบว่าการทำหน้าที่ต่างกันของข้อสอบ (DIF) มีแนวโน้มที่จะ Underestimated เมื่อข้อสอบไม่เป็นอิสระ

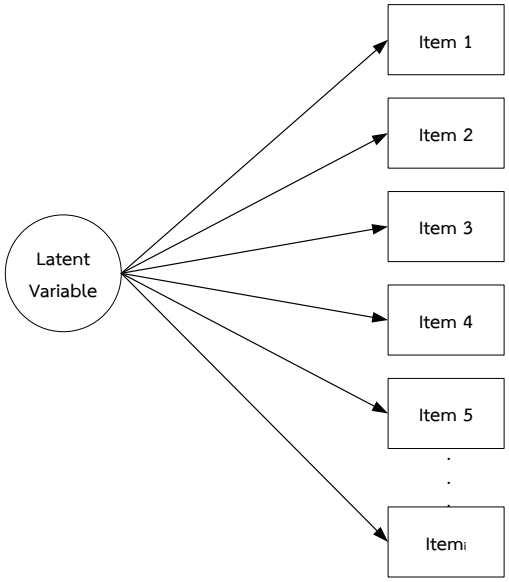
ซึ่งจากงานวิจัยครั้งนี้ ผู้วิจัยได้แทนตัวแปร คือ β_i แทนค่าพารามิเตอร์ความยากของข้อสอบ และค่าพารามิเตอร์ความสามารถผู้วิจัยใช้ตัวแปร θ_j ดังนี้

$$\theta = \gamma' z + \zeta, \tag{31}$$

เมื่อ γ คือ เวกเตอร์ของสัมประสิทธิ์การถดถอยของกลุ่มตัวแปร z ที่จะอธิบายความต่างระหว่างกลุ่มใน θ ซึ่งมักจะอ้างถึงในการวิเคราะห์ DIF และ ζ คือ Normally Distributed กับค่าเฉลี่ย 0 และตัวแปรอิสระของ z ซึ่งการรวมกันของสมการ 1, 2 และ 3 Reveals ว่า β_i คือ เงื่อนไขการทดสอบของ θ ซึ่งเป็นไปตามเงื่อนไขของการทำหน้าที่ต่างกันของข้อสอบ (DIF) เช่นเดียวกันกับที่ได้กล่าวมาแล้ว

นอกจากนี้ ในการวัดสิ่งต่าง (Measurement) สิ่งน่าสนใจคือความแตกต่างของกลุ่มในตัวแปรแฝง (Latent Variable) ในการศึกษาความแตกต่างของค่าเฉลี่ยของตัวแปรแฝง เป็นการศึกษาความไม่แปรเปลี่ยนของกลุ่ม (Invariant) ในขณะที่การศึกษาความแตกต่างของตัวแปรสังเกตได้ เช่น ค่าเฉลี่ยของข้อคำถามในแต่ละกลุ่ม ซึ่งการศึกษาการทำหน้าที่ต่างกันของข้อคำถามเป็นการศึกษาความแตกต่างของตัวแปรสังเกตได้หรือตัวชี้วัด

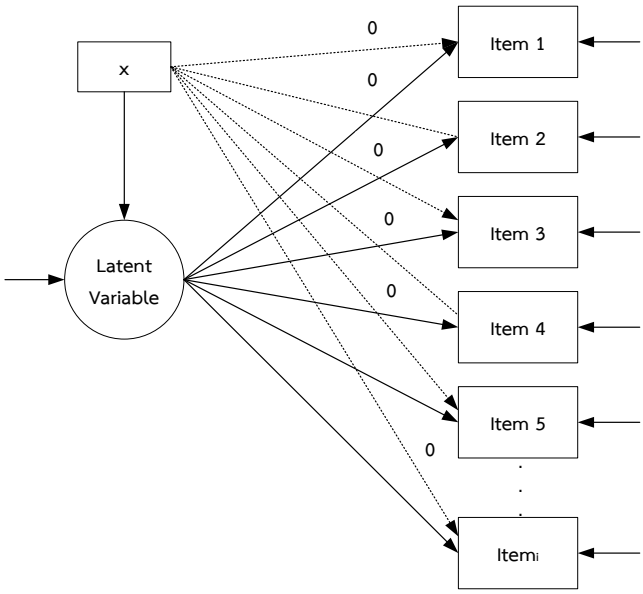
โดยปกติรูปแบบการวิเคราะห์ข้อมูลตามทฤษฎีการตอบสนองข้อสอบ (IRT) เป็นการศึกษาคุณสมบัติอยู่บนพื้นฐานข้อตกลงความเป็นเอกมิติ (Unidimensional) ของตัวแปรแฝงซึ่งสังเกตไม่ได้โดยตรง สำหรับตัวแปรแฝงในโมเดล IRT จะดูจากค่าเซต้า (θ) ซึ่งสามารถประมาณค่าได้โดยตรงซึ่งมีอิทธิพลตรงต่อตัวชี้วัดหรือข้อคำถามที่สังเกตได้ ซึ่งสามารถอธิบายในโมเดลการวิเคราะห์องค์ประกอบ แสดงดังภาพที่ 2-7



ภาพที่ 2-7 โมเดลการวิเคราะห์องค์ประกอบตามแนวคิด IRT (Riley & Dennis, 2015, p. 8)

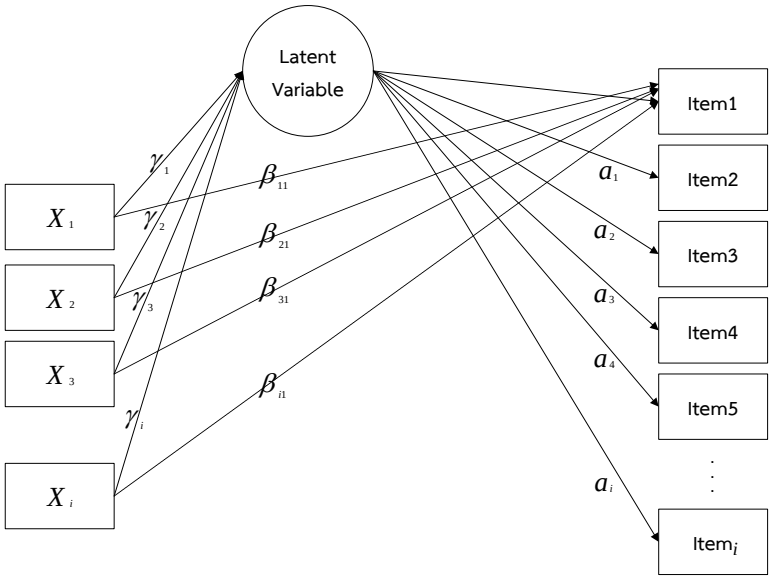
จากภาพที่ 2-7 ถ้าข้อคำถามหรือตัวชี้วัดเป็นตัวแปรจัดกลุ่ม (Dichotomous) และตัวแปรแฝงมีการแจกแจงแบบปกติ (Normal Distribution) ซึ่งจะมีลักษณะเช่นเดียวกับการแจกแจงโค้ง

ความถี่สะสมในโมเดล IRT น้ำหนักองค์ประกอบที่เกิดขึ้นบนตัวชี้วัดจะหมายถึง ค่าดัชนีประมาณค่าอำนาจจำแนกของข้อสอบตามทฤษฎีการตอบสนองข้อสอบ ในขณะที่ค่าเฉลี่ย (Intercepts) ของแต่ละข้อคำถามคือค่าความยากของข้อสอบ (Difficulty) ตามทฤษฎีการตอบสนองข้อสอบในกรณีที่มีตัวแปรแฝงมากกว่าหนึ่งตัวโปรแกรมที่พัฒนามาใช้ตามทฤษฎี IRT โดยปกติจะอนุมาน (Assumes) ว่ามีข้อมูลเป็นลักษณะมีความเป็นเอกมิติ ในการนำมาประยุกต์ใช้ในการวิจัยจึงทำได้กว้างยิ่งขึ้น ดังนั้น จึงง่ายต่อการนำแนวคิดมาประยุกต์ใช้ในกรณีที่ต้องการนำแนวคิดของโมเดล MIMIC มาใช้ในกรณีที่ตัวแปรแฝงมีหลายมิติ (Multi-dimensional) หลายองค์ประกอบ (Multi-Factor) การนำโมเดล MIMIC มาใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถามแสดงได้ตามโดยใช้ตัวแปรสาเหตุ (Causes) เพียงตัวเดียว แสดงได้ดังภาพที่ 2-8



ภาพที่ 2-8 โมเดลดารวิเคราะห์การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี MIMIC โดยใช้ตัวแปรสาเหตุ 1 ตัว (Brown, 2014, p. 308)

จากภาพที่ 2-8 เป็นการวิเคราะห์องค์ประกอบโดยใช้ตัวแปรทำนาย (x) จำนวนหนึ่งตัวแปร ในการทำนายตัวแปรแฝงที่ประกอบด้วยตัวแปรสังเกตได้ที่เป็นข้อสอบหรือข้อคำถาม (Item) จำนวน i ตัว โดยการจำกัดความคาดเคลื่อนจากการวัดของตัวแปรแฝง และให้อิสระกับความคลาดเคลื่อนของตัวแปรวัดที่สอดคล้องกับโมเดลมากกว่าในการประมาณค่าระหว่าง x กับ ตัวแปรแฝง (Latent Variable) อิทธิพลตรงของตัวแปร x ที่ทำนาย Item หลังจากที่มีอิทธิพลตรงไปยังตัวแปรแฝงแสดงทิศทางเดียว (Uniform) ในการทำหน้าที่ต่างกันของข้อคำถาม (DIF) ซึ่งเป็นสิ่งที่แสดงความลำเอียง (Biased) ที่เกิดจากข้อสอบหรือข้อคำถามหรืออธิบายได้ว่า ถ้าข้อสอบหรือข้อคำถาม (Item) ได้รับอิทธิพลอย่างมีนัยสำคัญจากตัวแปรสาเหตุ x ไม่ได้อธิบายตัวแปรแฝงแสดงว่าข้อสอบหรือข้อคำถาม (Item) ข้อนั้นทำหน้าที่ต่างกันหรือมีความลำเอียง (Biased) แสดงได้ดังภาพที่ 2-9



ภาพที่ 2-9 โมเดลการวิเคราะห์การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี MIMIC โดยใช้ตัวแปรสาเหตุมากกว่า 1 ตัว (Riley & Dennis, 2015, p. 10)

จากภาพที่ 2-9 เป็นการวิเคราะห์องค์ประกอบโดยใช้ตัวแปรทำนาย $X_1 - X_i$ ในการทำนายตัวแปรแฝงที่ประกอบด้วยตัวแปรสังเกตได้ที่เป็นข้อสอบหรือข้อคำถาม (Item) จำนวน i ตัว ในการตรวจสอบการทำหน้าที่ต่างกันของข้อคำถาม (DIF) สามารถดำเนินการตามขั้นตอนดังนี้

1. การวิเคราะห์องค์ประกอบในโมเดลวัดที่ประกอบด้วยตัวแปรแฝงและข้อคำถาม
2. การเพิ่มความแปรปรวนร่วมในการทดสอบโมเดล
3. การเพิ่มอิทธิพลตรงกับตัวแปรแฝง (γ) อิทธิพลตรง (a) และกำหนดให้มีค่าเท่ากับ 0
4. การตรวจสอบโมเดลดัชนีปรับแก้ (Modification Indices)
5. การเพิ่มอิทธิพลตรงจากความแปรปรวนร่วมกับข้อคำถามที่มีค่าดัชนีปรับแก้สูงสุด
6. การดำเนินในขั้นตอนที่ 4-5 จนกว่าจะไม่พบค่าดัชนีปรับแก้ (M.I.) ที่ไม่มีนัยสำคัญ

ประเมินความสอดคล้องของโมเดลและอิทธิตรง (β_i)

งานวิจัยที่เกี่ยวข้องกับวิธี MIMIC มีดังนี้

สุพัฒนา หอมบุปผา (2556) ได้ศึกษาเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN โดยจำแนกกลุ่มตัวอย่างตามเพศ และสถานที่ตั้งของโรงเรียน จากการศึกษาพบว่า ผลการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ของค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) ของวิชาคณิตศาสตร์ของท้ะง 3 วิธี มีความสัมพันธ์ในระดับที่สูงมาก เมื่อพิจารณาผลการประมาณค่าด้วยวิธี MIMIC เทียบกับผลการประมาณค่าด้วยวิธี BAYESIAN พบว่า MIMIC มีความสัมพันธ์กับ Baysian ในระดับสูงมากที่สุด เท่ากับ .999 และมีนัยสำคัญทางสถิติที่ระดับ .01 รองลงมาคือผลการประมาณค่าด้วยวิธี HGLM-2L เทียบกับผลการประมาณค่าด้วยวิธี BAYESIAN พบว่า HGLM-2L สัมพันธ์กับ Baysian ในระดับสูงมาก เท่ากับ .996 และมีนัยสำคัญทางสถิติที่ระดับ .01

และผลการประมาณค่าด้วยวิธี HGLM-2L เทียบกับผลการประมาณค่าด้วยวิธี MIMIC พบว่า เซต้า HGLM-2L สัมพันธ์กับ MIMIC ในระดับสูงมาก เท่ากับ .994 และมีนัยสำคัญทางสถิติที่ระดับ .01

Finch (2005) ศึกษาเปรียบเทียบความสามารถของตัวแบบหลายตัวบ่งชี้หลายสาเหตุ (MIMIC) รูปแบบการวิเคราะห์ปัจจัยยืนยันเพื่อระบุกรณีของการทำงานของรายการที่แตกต่างกัน (DIF) ได้อย่างถูกต้องด้วยวิธีการที่กำหนดขึ้น แม้ว่าแบบจำลอง MIMIC อาจมีแอพพลิเคชันในการระบุ DIF สำหรับตัวแปรกลุ่มหลาย ๆ แบบ แต่ก็มีการตรวจสอบว่าเทคนิคการทำงานของ DIF มีความถูกต้องและไม่ถูกต้องอย่างไร การใช้วิธีการมอนติคาร์โลในการศึกษาครั้งนี้สำหรับการจัดการจำนวนรายการ จำนวนผู้สอบ ความแตกต่างระหว่างความสามารถเฉลี่ยของกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ ระดับการตรวจพบ DIF ของรายการหลักและจำนวนข้อสอบที่ DIF ในรายการหลัก ผลการทดลองแสดงให้เห็นว่าแบบจำลอง MIMIC มีประสิทธิภาพสำหรับการระบุ DIF สำหรับข้อสอบ 50 ข้อ หรือเมื่อโมเดลโลจิสติกสองพารามิเตอร์มีข้อมูลอยู่การ DIF แต่มีอัตราการ DIF ที่ไม่ถูกต้องสำหรับ 20 ข้อที่มีข้อมูลโลจิสติกสามพารามิเตอร์

Wang, Ching, and Chih (2009) ได้ศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี MIMIC พร้อมประเมินประสิทธิภาพของโมเดล พบว่า วิธี MIMIC กับการตรวจสอบที่มีประสิทธิภาพสูงกว่ามาตราฐานของวิธี MIMIC ในการควบคุมอัตราการเป็นเท็จและอัตราผลการตอบที่เป็นจริงสูงขึ้น โดยรูปแบบของการทำหน้าที่ต่างกันของข้อสอบมีความสมดุลกันระหว่างกลุ่มหรือเปอร์เซ็นของการทำหน้าที่ต่างกันของข้อสอบเพียงเล็กน้อย ระหว่างวิธี MIMIC ที่มีประสิธิภาพสูงกว่ามาตราฐาน หรือวิธี MIMIC ที่เป็นแบบมาตราฐาน พบว่า วิธีที่สูงกว่ามาตราฐานนั้นดีกว่าวิธีที่เป็นเพียงมาตราฐาน เพราะรูปแบบของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในการทดสอบจริงอาจไม่สมดุลกัน และมีเปอร์เซ็นที่จะตรวจพบข้อสอบที่ทำหน้าที่ต่างกันมากขึ้น

Mucherah, Finch, and Keaikitse (2012) ศึกษาการทำหน้าที่ต่างกันของแบบวัดมโนภาพแห่งตน (Self-Description Questionnaire) หรือเป็นแบบสอบที่ให้อธิบายความตัวเอง (Self-Concept Scale) สำหรับชาวเคนย่า จำแนกเพศหญิงและเพศชาย โดยใช้แบบจำลองสมการเชิงโครงสร้าง (MIMIC Model) ซึ่งเป็นงานวิจัยที่จะแสดงให้เห็นว่า วัยรุ่นมีความเข้าใจความคิดตัวเองมากน้อยเพียงไร จึงได้มีการตรวจสอบการทำหน้าที่ต่างกันของแบบสอบเหล่านี้ โดยใช้วิธี MIMIC และวิธี SIBTEST มีกลุ่มตัวอย่างเป็นนักเรียนมัธยมจากทั่วประเทศเคนย่า จำนวน 1990 คน เป็นเพศชาย 983 คน เพศหญิง 1,007 คน โดยมีแบบวัดมโนภาพแห่งตน (SDQ) จำนวน 135 ข้อ จากการศึกษาพบว่า เพศหญิงจะมีการตอบคำถามเข้าในเชิงบวก และตัวแปรเพศยังมีความแตกต่างกันเรื่องของคณิตศาสตร์พร้อมทั้ง โดยที่เพศเพศหญิงจะมีความคิดในเชิงลบทางด้านนี้ และวิธีการตรวจสอบการทำหน้าที่ต่างกันด้วยวิธี MIMIC และวิธี SIBTEST อยู่ในเกณฑ์ที่ใช้ได้

จากการศึกษางานวิจัยที่เกี่ยวข้อง วิธี MIMIC สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ดีกว่าวิธี HGLM ในแบบทดสอบที่มีความยาวไม่เกิน 20 ข้อ และสามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดี เมื่อแบบทดสอบมีความยาวไม่เกิน 20 ข้อ แต่เมื่อข้อสอบมีความยาวมากกว่า 20 ข้อ จะทำให้ประสิทธิภาพในการควบคุมความคลาดเคลื่อนประเภทที่ 1 ลดลงด้วย

ตอนที่ 4 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ด้วยวิธี IRT-LR และงานวิจัยที่เกี่ยวข้อง

ทฤษฎีการตอบสนองข้อสอบด้วย IRT นั้น เป็นทฤษฎีการวัดที่อธิบายความสัมพันธ์ของคุณลักษณะภายในและความสามารถของตัวบุคคลที่จะสามารถตอบข้อสอบได้ถูก โดยสามารถแบ่งออกเป็น 2 ประเภทด้วยกัน คือ การตอบสนองข้อสอบแบบตรวจให้คะแนนแบบ 2 ค่า (Dichotomous IRT) เป็นโมเดลที่ใช้การตรวจให้คะแนนแบบ 2 ค่า เช่น ข้อสอบที่มีการตรวจให้คะแนน แบบถูก/ผิด แบบ 0,1 (คือการตอบผิดให้ 0, ตอบถูกให้ 1) หรือ ใช่/ไม่ใช่ เป็นต้น และอีกหนึ่งประเภท คือ การตอบสนองข้อสอบแบบการตรวจให้คะแนนแบบหลายค่า (Polytomous IRT) นั่นคือข้อสอบที่มีการตรวจการให้คะแนนแบบมากกว่า 2 ค่าขึ้นไป เช่นข้อสอบที่ตรวจให้คะแนนแบบมาตรประมาณค่า (Rating Scale) หรือแบบการตรวจให้คะแนนแบบให้คะแนนความรู้บางส่วน (Partial Credit) เป็นต้น (ศิริชัย กาญจนวาสี, 2555, หน้า 51-52)

โมเดลการตอบสนองข้อสอบ (IRT Models) เดิมมีแนวคิดพื้นฐานมาจากทฤษฎีการวัดแบบดั้งเดิม (Classical Test Theory) โดยโมเดลการตอบสนองข้อสอบจะวัดความสัมพันธ์ระหว่างตัวแปรอิสระ ที่ประกอบด้วยตัวแปรแฝง คือ ความสามารถของผู้สอบ (θ) หรือค่าพารามิเตอร์ของผู้สอบ (a,b,c) และตัวแปรอิสระที่เป็นตัวแปรที่สังเกตุได้ คือ โอกาสในการตอบข้อสอบได้ถูกต้องของผู้สอบ ซึ่งทฤษฎีการตอบสนองข้อสอบ (IRT) เป็นการอธิบายความสัมพันธ์ระหว่างความสามารถภายในของตัวบุคคล (Latent Trait or Ability) กับผลของการตอบข้อสอบถูกหรือโค้งคุณลักษณะข้อสอบ (Item Characteristic Curve: ICC) ที่มีการกำหนดลักษณะข้อสอบด้วยพารามิเตอร์ของข้อสอบ นั่นคือ ค่าอำนาจจำแนกของข้อสอบ (a) ค่าความยากของข้อสอบ (b) และค่าโอกาสในการเดาของข้อสอบได้ถูก (c) เพราะเหตุนี้ IRT จึงอยู่บนพื้นฐานของความคิดที่จำเป็น 2 ประการ คือ 1) ผลการตอบข้อสอบได้ถูกต้องของผู้สอบ ที่อธิบายความสามารถที่มีอยู่ภานในตัวผู้สอบได้ และ 2) ความสัมพันธ์ระหว่างผลการตอบข้อสอบกับความสามารถที่อยู่ภายในตัวของผู้สอบ ซึ่งอธิบายได้ด้วยโค้งคุณลักษณะข้อสอบ (ICC) และ ฟังก์ชั่นของลักษณะข้อสอบ (Embertson & Resie, 2000, p. 8)

แนวคิดทฤษฎีการตอบสนองข้อสอบ

พัฒนาการของทฤษฎีการทดสอบ

ทฤษฎีการทดสอบเริ่มแรกมีการพัฒนามาจากผลงาน และความพยายามของนักจิตวิทยาทั้งในยุโรปและอเมริกา นักจิตวิทยาได้หาวิธีการแก้ปัญหาการวัดต่าง ๆ เพื่อพัฒนาศาสตร์แห่งการวัดและตรวจสอบจนมีความมั่นคง ในศตวรรษที่ 12 เมื่อมีเริ่มผลิตกระดาษขึ้นใช้แทนการสอบปากเปล่า และปี พ.ศ. 2293 มหาวิทยาลัยเคมบริด์จ ได้เริ่มใช้การสอบด้วยกระดาษอย่างเป็นทางการ เพื่อวัดผลการเรียนของนักศึกษา มหาวิทยาลัยออกซ์ฟอร์ด ได้ใช้ข้อสอบข้อเขียนสาหรับวัดจำแนกความสามารถของนักศึกษา เพื่อวัดผลระดับผ่านหรือระดับเกียรตินิยม

แนวคิดพื้นฐานของทฤษฎีการตอบสนองข้อสอบ

ทฤษฎีนี้เกิดขึ้นท่ามกลางข้อจากัดของทฤษฎีการทดสอบแบบดั้งเดิมหลายประการ คือ (Hambleton, Swaminathan, & Rogers, 1991, pp. 7-12)

1. ค่าสถิติของข้อสอบ เช่น ความยากจะขึ้นอยู่กับลักษณะของกลุ่มผู้สอบ กล่าวคือ ถ้าผู้สอบมีความสามารถสูง ข้อสอบจะกลายเป็นข้อสอบที่ง่าย แต่ถ้าผู้สอบมีความสามารถต่ำ

ข้อสอบดังกล่าวจะกลายเป็นข้อสอบที่ยาก ส่วนค่าอำนาจจำแนกของข้อสอบขึ้นอยู่กับความเป็นเอกพันธ์ของความสามารถของผู้สอบ ถ้าผู้สอบมีความสามารถแตกต่างกันมากข้อสอบก็จะมีค่าอำนาจจำแนกของข้อสอบสูงซึ่งมีผลทำให้ความเที่ยงของแบบทดสอบมีค่าสูงตามไปด้วย เนื่องจากความเที่ยงของแบบทดสอบมีความสัมพันธ์ทางบวกกับค่าความแปรปรวนของคะแนนจากแบบทดสอบ

2. การเปรียบเทียบความสามารถของผู้สอบนั้น จะต้องใช้แบบทดสอบฉบับเดียวกันหรือแบบทดสอบคู่ขนาน ปัญหาที่เกิดขึ้น คือ แบบทดสอบวัดผลสัมฤทธิ์ และแบบทดสอบวัดความถนัดนั้น ส่วนใหญ่แล้วจะเหมาะสมกับผู้ที่มีความสามารถปานกลาง ดังนั้น ความถูกต้องแม่นยำของการวัด ผู้สอบที่มีความสามารถสูงและผู้สอบที่มีความสามารถต่ำจึงลดลง

3. ค่าความเที่ยงของแบบทดสอบถูกนิยามในรูปของผลที่ได้จากการใช้แบบทดสอบคู่ขนาน ซึ่งในทางปฏิบัติจริงนั้นนับว่าเป็นเรื่องยากที่จะให้การสอบ 2 ครั้งมีสภาพที่เหมือนกัน ถึงแม้ว่าแบบทดสอบคู่ขนานจะขนานกันจริง แต่ผู้สอบอาจจะมีลักษณะที่แตกต่างกันไปจากการสอบครั้งแรกเกี่ยวกับแรงจูงใจ ความกังวล การลืม หรือการพัฒนาตนเองในบางทักษะเป็นต้น

4. ทฤษฎีการทดสอบแบบดั้งเดิมไม่สามารถบอกได้ว่าผู้สอบจะตอบข้อสอบอย่างไร ยกเว้นแต่ว่าจะได้ใช้ข้อสอบข้อนั้นกับผู้สอบที่มีลักษณะคล้ายคลึงกันมาแล้ว

5. ทฤษฎีการทดสอบแบบดั้งเดิมใช้ค่าความแปรปรวนของความคลาดเคลื่อนในการวัด (Variance of Error of Measurement) เหมือนกันกับผู้สอบทุกคน ซึ่งตามความเป็นจริงแล้วผู้สอบที่มีความสามารถสูงและต่ำ จะมีค่าความแปรปรวนของความคลาดเคลื่อนในการวัดต่างจากผู้สอบที่มีความสามารถปานกลาง

หลักการของทฤษฎีการตอบสนองข้อสอบ

ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory) มีความเชื่อเกี่ยวกับค่าพารามิเตอร์ของข้อสอบ (Item Parameter) คือ ค่าความยากของข้อสอบ (b) ค่าอำนาจจำแนกของข้อสอบ (a) ค่าโอกาสในการเดาข้อสอบ (c) ของข้อสอบแต่ละข้อว่าเป็นคุณลักษณะที่คงที่ในตัวข้อสอบนั้น เพราะฉะนั้นค่าพารามิเตอร์เหล่านี้จึงไม่ควรแปรเปลี่ยนไปตามกลุ่มผู้สอบ และในทำนองเดียวกัน ค่าพารามิเตอร์ของผู้สอบ (Person Parameter) หรือความสามารถที่แท้จริงของผู้สอบก็เป็นคุณลักษณะที่มีอยู่ภายในตัวผู้สอบจึงไม่ควรแปรเปลี่ยนไปตามชุดข้อสอบที่เลือกใช้ แต่เนื่องจากความสามารถของผู้สอบเป็นคุณลักษณะแฝงไม่สามารถสังเกต หรือวัดได้โดยตรง (Unobservable) จึงจาเป็นต้องใช้การทำนาย (Predict) หรืออธิบาย (Explain) คุณลักษณะดังกล่าว โดยอาศัยผลที่ได้จากการตอบแบบทดสอบ (Test Performance) หรือคะแนน (Score) ซึ่งเป็นสิ่งที่สามารถสังเกตและวัดได้ (Observable) นักวัดผลจึงได้พยายามหาความสัมพันธ์ระหว่างผลที่ได้จากการตอบแบบทดสอบหรือคะแนน (Test Performance or Score) กับระดับความสามารถ (Ability) ของผู้ตอบแต่ละคน เพื่อเขียนเป็นโมเดลทางคณิตศาสตร์ (Mathematical Model) ความสัมพันธ์ระหว่างผลที่ได้จากการตอบแบบทดสอบกับระดับความสามารถของผู้สอบ สามารถเขียนในรูปของความสัมพันธ์ ได้ดังนี้

$$P = f\left(U_i / \theta_1, \theta_2, \theta_3, \ldots . \theta_k ; \beta_k\right) \qquad (32)$$

เมื่อ P แทน ผลการตอบแบบทดสอบ (Test Performance)

f แทน ฟังก์ชัน (Function)

U_i แทน ผลการตอบแบบทดสอบข้อที่ I (ตอบถูก = 1, ตอบผิด = 0)

$\theta_1,\theta_2,\theta_3,.....\theta_k$ แทน ระดับความสามารถ (Ability) ที่ 1, 2, 3, ...,k

β_k แทน ค่าพารามิเตอร์ของข้อสอบข้อที่ j

เนื่องจากความสัมพันธ์ดังกล่าวเป็นเพียงฟังก์ชันความสัมพันธ์ในลักษณะทั่ว ๆ ไป นักวัดผลการศึกษาจึงต้องหาโมเดลทางคณิตศาสตร์ที่เหมาะสม เพื่อใช้แทนฟังก์ชันความสัมพันธ์ดังกล่าว โดยอาศัยข้อตกลงเบื้องต้นของทฤษฎีการตอบสนองข้อสอบ

ข้อตกลงเบื้องต้นของทฤษฎีการตอบสนองข้อสอบ

ข้อตกลงเบื้องต้นของทฤษฎีการตอบสนองข้อสอบที่สาคัญ ดังนี้ (Hambleton et al., 1991, pp. 9-12, ศิริชัย กาญจนวาสี, 2555, หน้า 75-76)

1. ความเป็นเอกมิติ (Unidimentional) การวัดตามแนวทฤษฎีการตอบสนองข้อสอบ ได้ระบุความเป็นเอกมิติของคุณลักษณะในข้อตกลงเบื้องต้นนี้ว่า ข้อคาถาม หรือข้อสอบทุกข้อในแบบทดสอบนั้นจะต้องมุ่งวัดความสามารถ หรือคุณลักษณะเดียวเท่านั้น โดยทั่วไปแล้วข้อตกลงข้อนี้เป็นไปได้ค่อนข้างยากเนื่องจากมีปัจจัยที่มีผลต่อคะแนนสอบ เช่น ปัจจัยด้านความรู้ความเข้าใจ (Cognitive) บุคลิกภาพและปัจจัยเกี่ยวกับการจัดการสอบ ปัจจัยเหล่านี้อาจรวมถึงแรงจูงใจ ความวิตกกังวลในการสอบความสามารถในการทำงานได้รวดเร็ว ความรู้เกี่ยวกับการใช้กระดาษคำตอบ เมื่อเป็นเช่นนี้สิ่งที่ทำให้ข้อตกลงนี้เป็นไปได้ คือ การพิจารณาว่าแบบทดสอบฉบับนั้นมีองค์ประกอบใดหรือปัจจัยใดที่เด่นที่สุด ก็ถือว่าแบบทดสอบได้วัดในสิ่งนั้นการตรวจสอบความเป็นเอกมิติของคุณลักษณะที่ใช้ในการทดสอบมีวิธีการตรวจสอบได้หลายวิธี สรุปที่สำคัญได้ ดังนี้ (ชนะศึก นิชานนท์, 2553)

1.1 การหาค่าความสัมพันธ์ระหว่างค่าน้ำหนักองค์ประกอบรายข้อ (Factor Loading) ขององค์ประกอบที่หนึ่งกับค่าสหสัมพันธ์แบบไบซีเรียล (Biserial Correlation Coefficient) ของข้อสอบรายข้อกับคะแนนรวม ถ้ามีค่าสัมประสิทธิ์สหสัมพันธ์มากกว่า .80 ทำให้สามารถสรุปได้ว่า ข้อสอบหรือแบบสอบฉบับนั้นมีความเป็นเอกมิติของคุณลักษณะที่ใช้ในการทดสอบ

1.2 การวิเคราะห์องค์ประกอบ (Factor Analysis) ของข้อสอบทั้งฉบับ พิจารณาได้จาก ค่าไอเกน (Eigen Value) โดยผลการวิเคราะห์องค์ประกอบใดมีค่าไอแกนในองค์ประกอบใดองค์ประกอบหนึ่งสูงกว่าค่าอื่นอย่างชัดเจน สามารถสรุปได้ว่าข้อสอบหรือแบบสอบฉบับนั้นมีความเป็นเอกมิติของ คุณลักษณะในการทดสอบ

1.3 การใช้โปรแกรม TESTFACT ในการพิจารณาความเป็นมิติของแบบสอบจากการวิเคราะห์องค์ประกอบเชิงยืนยัน (Confirmatory Factor Analysis) ซึ่งพัฒนาขึ้นโดย (Wilson & Hoskens, 1991 อ้างถึงใน ชนะศึก นิชานนท์, 2553) โดยวิเคราะห์ข้อสอบและทดสอบความตรงของโครงสร้าง ด้วย x^2 สาหรับ Likelihood Ratio G^2 ในการตรวจสอบมิติของแบบสอบดัชนีที่ใช้นี้ทดสอบด้วยการกำหนดจำนวน องค์ประกอบของชุดข้อมูลไว้ล่วงหน้าแล้วทดสอบด้วย x^2 ที่ประมาณค่าด้วยวิธี G^2 เพื่อทดสอบความเหมาะสมของโมเดล เมื่อค่า G^2 ไม่มีนัยสำคัญแสดงว่าข้อมูลมีจำนวนองค์ประกอบเท่าที่กำหนดในการทดสอบ

1.4 การวิเคราะห์ การจัดกลุ่มระดับชั้น (Hierarchical Cluster Analysis) เป็นเทคนิคสำหรับการทดสอบความเป็นพหุมิติของชุดแบบสอบ โดยการพิจารณาการแบ่งกลุ่มของตัวแปร โดยกระบวนการแบ่งกลุ่มนี้เป็นการแบ่งกลุ่มจำนวนข้อสอบที่มีลักษณะคล้ายคลึงกันให้อยู่ในกลุ่มเดียวกัน นอกจากนี้ยังใช้วิธีในการหมุนซ้ำ (Iteration) จนสำเร็จหรืออยู่ในระดับที่น่าพอใจ ซึ่งทำให้ได้ความเป็นไปได้ของผลลัพธ์ (Outcome) โดยสามารถช่วยอธิบายให้ข้อมูลมีความถูกต้องมากยิ่งขึ้น ซึ่งการวิเคราะห์ด้วยวิธีการนี้สามารถใช้โปรแกรมคอมพิวเตอร์สาเร็จรูป CCPROX และ HCA ในการวิเคราะห์ได้

1.5 การใช้โปรแกรม DETECT เป็นการตรวจสอบมิติแฝงเชิงยืนยันแบบ Nonparametric ซึ่งจะใช้ในการประมาณค่าจำนวนของมิติแฝงที่มีคุณลักษณะเด่นในชุดของข้อมูล และสามารถตรวจสอบความเป็นเอกมิติของแบบสอบ โดยระบุคุณลักษณะเด่นของมิติแฝงในแต่ละข้อ ซึ่งผู้ใช้โปรแกรมสามารถระบุจำนวนมิติแฝงสูงสุดที่ต้องการศึกษาได้ เนื่องจากการจัดกลุ่มชุดของข้อสอบ แต่กระบวนการดังกล่าวยังมีลักษณะแบบไม่เป็นทางการเท่าใดนัก เนื่องจากการระบุการจัดกลุ่มเพื่อจาแนกความแตกต่างของมิติจะอาศัยกระบวนการในการระบุความเป็นหนึ่งเดียว

1.6 การใช้โปรแกรม DIMTEST เป็นกระบวนการ ตรวจสอบสมมติฐานของแบบสอบด้วย Nonparametric Statistical โดยมีลักษณะคล้ายคลึงกับการตรวจสอบด้วยโปรแกรม DETECT โดยการตรวจสอบจะตรวจสอบความสัมพันธ์ระหว่างชุดข้อสอบย่อยภายใต้เงื่อนไขความแปรปรวนร่วมของข้อสอบ ซึ่งแตกต่างจากการใช้โปรแกรม DETECT ที่มีลักษณะคล้ายการวิเคราะห์องค์ประกอบเชิงยืนยัน

2. ความเป็นอิสระในการตอบข้อสอบ (Local Independent)

ความเป็นอิสระในการตอบข้อสอบ หมายถึง ความน่าจะเป็นในการตอบข้อสอบแต่ละข้อได้ถูกต้องเป็นอิสระจากกัน นั่นคือ การตอบข้อสอบข้อใดข้อหนึ่งถูกหรือผิดจะไม่มีผลกระทบต่อการตอบข้ออื่น ๆ ด้วย หรืออาจจะกล่าวในเชิงคณิตศาสตร์ได้ว่า ความเป็นอิสระในการตอบข้อสอบ หมายถึง ความน่าจะเป็นในการตอบข้อสอบถูกทั้งหมดมีค่าเท่ากับ ผลคูณของความน่าจะเป็นในการตอบข้อสอบถูกเป็นรายข้อ นั่นคือ ผู้สอบที่มีความสามารถ (θ) จะมีความน่าจะเป็นที่จะตอบข้อสอบทั้งข้อ 1 และข้อ 2 ถูกเท่ากัน ซึ่งได้มาจาก ความน่าจะเป็นในการตอบข้อสอบข้อที่ 1 ถูก และความน่าจะเป็นในการตอบข้อสอบข้อที่ 2 ถูก คือ ถ้าผู้สอบมีความสามารถ (θ) เท่ากับ 1.5 มีความน่าจะเป็นในการตอบข้อสอบข้อที่ 1 ถูกเท่ากับ 0.5 และมีความน่าจะเป็นในการตอบข้อสอบข้อที่ 2 ถูก เท่ากับ 0.6 ดังนั้นผู้สอบที่มีความสามารถ เท่ากับ 1.5 มีความน่าจะเป็นในการตอบข้อสอบทั้งสองข้อถูกภายใต้เงื่อนไขความเป็นอิสระ มีค่าเท่ากับ 0.3 อย่างไรก็ตาม Hambleton and Swaminathan (1985) กล่าวว่า ถ้าแบบทดสอบมีความเป็นเอกมิติอยู่แล้ว ความเป็นอิสระในการตอบข้อสอบก็จะเกิดขึ้นตามไปด้วย

3. โค้งคุณลักษณะข้อสอบ (Item Characteristic Curve)

โค้งคุณลักษณะข้อสอบเป็นฟังก์ชันทางคณิตศาสตร์ สามารถใช้อธิบายความสัมพันธ์ระหว่างความน่าจะเป็น หรือโอกาสที่ผู้สอบจะตอบข้อสอบถูกกับระดับความสามารถที่วัดได้โดยใช้ชุดของข้อสอบ หรือแบบทดสอบฉบับนั้น ทั้งนี้ ความน่าจะเป็น หรือโอกาสในการตอบข้อสอบถูกจะขึ้นอยู่กับโค้งลักษณะข้อสอบในแต่ละโมเดลที่เลือกใช้ โดยที่รูปร่าง (Shape) ของ

โค้งคุณลักษณะข้อสอบในแต่ละข้อมีคุณสมบัติไม่แปรเปลี่ยน (Invariant) ไปตามกลุ่มตัวอย่างที่ใช้ ดังนั้น จึงทำให้ความน่าจะเป็น หรือโอกาสในการตอบข้อสอบถูกในแต่ละข้อไม่แปรเปลี่ยนด้วย คุณสมบัตินี้ถือเป็นลักษณะเด่นของโมเดลต่าง ๆ ในทฤษฎีการตอบสนองข้อสอบ โค้งคุณลักษณะข้อสอบมีหลายรูปแบบขึ้นอยู่กับว่าเลือกใช้พารามิเตอร์ของข้อสอบกี่พารามิเตอร์

4. ข้อสอบที่ใช้ต้องไม่เป็นข้อสอบประเภทความเร็ว (Speediness)

ผู้สอบทุกคนควรมีโอกาสในการทำข้อสอบทุกข้อ เพื่อให้คะแนนรวมจากการสอบเป็นค่าความสามารถที่แท้จริงของผู้สอบไม่มีข้อจำกัดเกี่ยวกับเวลาในการสอบ สรุปได้ว่า ข้อตกลงเบื้องต้นของทฤษฎีการตอบสนองข้อสอบ คือ ข้อสอบต้องมีความเป็นเอกมิติ กล่าวคือ วัดความสามารถหรือคุณลักษณะเดียว ความน่าจะเป็นในการตอบข้อสอบแต่ละข้อได้ถูกต้อง จะต้องเป็นอิสระต่อกัน นอกจากนี้ โค้งคุณลักษณะข้อสอบเป็นฟังก์ชันทางคณิตศาสตร์ที่อธิบายความสัมพันธ์ระหว่างความน่าจะเป็นในการตอบข้อสอบถูกกับระดับความสามารถของผู้สอบ และข้อสอบที่ใช้ต้องไม่เป็นข้อสอบประเภทความเร็ว

พารามิเตอร์ในทฤษฎีการตอบสนองข้อสอบ

พารามิเตอร์ในทฤษฎีการตอบสนองข้อสอบ แบ่งออกเป็น 2 ชนิด คือ พารามิเตอร์ข้อสอบ (Item Parameter) ได้แก่ ค่าความยากของข้อสอบ (b) ค่าอำนาจจำแนกของข้อสอบ (a) ค่าโอกาสการเดาของข้อสอบ (c) และความรอบคอบ (γ) ส่วนพารามิเตอร์ของผู้สอบ (Person Parameter) ได้แก่ ระดับความสามารถ หรือ คุณลักษณะของผู้สอบ (θ) ซึ่งค่าพารามิเตอร์ต่าง ๆ มีลักษณะและการแปลความหมาย ดังนี้ (ศิริชัยกาญจนวาสี, 2550, หน้า 53-55)

1. พารามิเตอร์ข้อสอบ (Item Parameter)

พารามิเตอร์ข้อสอบประกอบด้วย ค่าความยากของข้อสอบ (b) ค่าอำนาจจำแนกของข้อสอบ (a) ค่าโอกาสการเดาของข้อสอบ (c) และความรอบคอบ ซึ่งพารามิเตอร์แต่ละชนิดมีรายละเอียด ดังนี้

1.1 ค่าความยากของข้อสอบ (b)

ค่าความยากของข้อสอบได้มาจากค่าความสามารถที่ตรงจุดเปลี่ยนโค้ง (Inflexion Point) ซึ่งเป็นจุดที่โค้งมีความชันมากที่สุด หรือมีความหมายอีกนัยหนึ่งก็คือ ผู้สอบที่มีความสามารถถึงระดับ ณ จุดเปลี่ยนโค้งข้อสอบนั้น จะมีโอกาสตอบข้อสอบนี้ถูกอยู่ 0.5 หรือในทางปฏิบัติ กล่าวได้ว่า จากจุดบนแกน y ที่แสดงถึงตาแหน่งโอกาสในการตอบข้อสอบข้อนี้ถูกมีอยู่ 0.5 ถ้าลากเส้นตั้งขนานกับแกน x จนพบกับเส้นโค้ง ซึ่งจะเป็นจุดเปลี่ยนโค้งด้วยนั้น ในทางตรงกันข้าม เมื่อลากเส้นตั้งฉากจากจุดดังกล่าวให้มาจรดแกน x ค่าที่วัดได้ในแกน x คือ ค่าความยากของข้อสอบ ข้อนั้น ๆ ซึ่งข้อสอบทั้งหมดในแบบทดสอบฉบับหนึ่ง ๆ ที่นามาวิเคราะห์จะมีค่าความยากของข้อสอบกระจายอยู่ในแกน x จากค่า -∞ ถึง ∞ แต่ในทางปฏิบัติ นิยมใช้ช่วง -3 ถึง +3 และแบบทดสอบทั่วไปมักจะมีค่า b อยู่ระหว่าง -2.5 ถึง +2.5 ถ้าค่า bi เข้าใกล้ -2.5 แสดงว่าข้อสอบง่าย ตรงกันข้าม ถ้าค่า b อยู่ใกล้ +2.5 แสดงว่าข้อสอบยาก ในกรณีที่เป็นรูปแบบที่มีค่าพารามิเตอร์ 3 ตัว หรือคานึงถึงโอกาสในการเดาคำตอบ (c) ค่าความยากซึ่งเริ่มต้นจากจุดบนแกน y นั้น จะใช้จุดตั้งต้นตรงที่ค่าโอกาสในการตอบข้อสอบถูก กล่าวคือ ค่าความยากมีค่าเริ่มต้นที่ $\frac{1-c}{2}$

1.2 ค่าอำนาจจำแนกของข้อสอบ (a)

ค่าอำนาจจำแนกของข้อสอบ (a) เป็นสัดส่วนกับค่าความชัน (Slope) ของ pi (θ) ที่จุดเปลี่ยนโค้ง หรือที่จุด θ = b โดยทฤษฎีแล้ว ค่าอำนาจจำแนกของข้อสอบ จะมีค่าอยู่ในช่วง -∞ ถึง ∞ แต่ในทางการนำมาใช้ประโยชน์นั้น ข้อสอบข้อใดที่มีค่า a ติดลบ ย่อมแสดงว่า ข้อสอบข้อนั้นไม่ดี และควรจะต้องถูกตัดออกไป ส่วนข้อสอบที่มีค่า a สูงขึ้น ย่อมแสดงว่า ความน่าจะเป็นของการตอบข้อสอบข้อนั้น ๆ เพิ่มขึ้นเมื่อระดับความสามารถของผู้สอบสูงขึ้น ตามปกติ ค่า a มีค่าไม่เกิน +2.5 ในทางปฏิบัตินิยมใช้ข้อสอบที่มีค่า a อยู่ระหว่าง +0.5 ถึง +2.5

1.3 ค่าโอกาสการเดาของข้อสอบ (c)

ค่าโอกาสการเดาของข้อสอบ (c) เป็นค่าที่อยู่ปลายโค้งด้านต่ำ (Lower Asymptote) ของข้อสอบ ค่านี้เป็นค่าแทนความน่าจะเป็น หรือโอกาสที่คนซึ่งมีความสามารถต่ำ แต่สามารถตอบข้อสอบข้อนั้นได้ถูกต้องโดยการเดา ในทางทฤษฎีพารามิเตอร์การเดามีค่าระหว่าง 0.00 ถึง 1.00 โดยทั่วไปนิยมใช้ข้อสอบที่มีค่าโอกาสการเดาของข้อสอบไม่เกิน 0.30

1.4 ความรอบคอบ (γ)

McDonald (1967 อ้างถึงใน Hambleton & Swaminatan, 1985) ได้เสนอพารามิเตอร์ที่แสดงถึงความรอบคอบของผู้สอบ เป็นค่าพารามิเตอร์ที่บ่งชี้ว่าผู้สอบ ที่มีความสามารถสูงอาจจะตอบข้อสอบได้ไม่ถูกต้องเสมอไป ซึ่งอาจเกิดความไม่รอบคอบในการพิจารณาคำตอบ หรือผู้สอบอาจจะมีสารสนเทศอื่น ๆ เกี่ยวกับผู้ออกข้อสอบทำให้เลือกตอบในตัวเลือกที่ไม่ใช่คำตอบที่ถูกต้อง โดย Barton and Lord (1981 อ้างถึงใน Hambleton & Swaminatan, 1985) กล่าวว่า พารามิเตอร์ตัวนี้จะเหมาะสมในการศึกษาทางทฤษฎีเท่านั้น ซึ่งในทางปฏิบัติแล้วไม่สามารถพบพารามิเตอร์นี้ได้

2. พารามิเตอร์ผู้สอบ

พารามิเตอร์ผู้สอบ (θ) เป็นระดับความสามารถของผู้สอบ (θ) ที่ประมาณได้จากโมเดลตามทฤษฎีการตอบสนองข้อสอบ นิยมปรับให้เป็นคะแนนมาตรฐานที่มีค่าเฉลี่ยเป็น 0 และส่วนเบี่ยงเบนมาตรฐานเป็น 1 ซึ่งพารามิเตอร์ผู้สอบมีค่าระหว่าง -∞ ถึง ∞ แต่ส่วนใหญ่จะมีค่าอยู่ในช่วง -3.0 ถึง +3.0 ค่าที่เป็นลบ แสดงว่าผู้สอบมีความสามารถต่ำ และค่าที่เป็นบวก แสดงว่าผู้สอบมีความสามารถสูง สรุปได้ว่า พารามิเตอร์ตามทฤษฎีการตอบสนองข้อสอบจำแนกเป็น 2 ชนิด คือ พารามิเตอร์ข้อสอบ และพารามิเตอร์ผู้สอบ ซึ่งพารามิเตอร์ข้อสอบประกอบด้วยพารามิเตอร์ความยากของข้อสอบ พารามิเตอร์อำนาจจำแนกของข้อสอบ พารามิเตอร์โอกาสการเดาของข้อสอบ และความรอบคอบ ส่วนพารามิเตอร์ผู้สอบเป็นพารามิเตอร์ที่แสดงระดับความสามารถของผู้สอบ ซึ่งข้อตกลงเบื้องต้น และพารามิเตอร์ ที่กล่าวมานี้ มีความหมายเด่นชัดในกรณีที่ข้อสอบนั้นให้คะแนนแบบสองค่า ในการประยุกต์ทฤษฎี เพื่อใช้กับข้อสอบที่ให้คะแนนแบบมากกว่าสองค่า ข้อตกลงเบื้องต้นทั้งหมดก็เทียบเคียงในทำนองเดียวกัน แตกต่างกันเพียงรายละเอียดปลีกย่อยเกี่ยวกับเงื่อนไขเฉพาะของแต่ละโมเดลเท่านั้นโมเดลการตอบสนองข้อสอบแบบตรวจให้คะแนนสองค่า (Dichotomous IRT Models)เนื่องจากข้อมูลที่ได้จากการสอบมีหลายลักษณะ ได้แก่ ข้อมูลแบบมีสองค่า (Dichotomous) และข้อมูลแบบมีมากกว่าสองค่า (Polytomous) ดังนั้น จึงมีการพัฒนารูปแบบเพื่อให้สอดคล้องกับลักษณะของข้อมูลดังกล่าวขึ้นมากมาย แต่สาหรับข้อมูลที่เป็นแบบมี 2 ค่า

รูปแบบที่นิยมใช้เป็นรูปแบบโลจิสติก (Logistic Model) ซึ่งแตกต่างไปตามจานวนพารามิเตอร์ที่ใช้ในแต่ละรูปแบบ มีรายละเอียด ดังนี้

โมเดลการตอบสนองข้อสอบแบบ 1 พารามิเตอร์ (One – Parameter Model)

รูปแบบนี้บางครั้งเรียกว่า รูปแบบราสช์ (Rasch Model) คือ โมเดลที่มีการแปรเปลี่ยนค่าพารามิเตอร์เพียงพารามิเตอร์ค่าความยากของข้อสอบ (b) เพียงอย่างเดียว โค้งคุณลักษณะข้อสอบสามารถเขียนสมการ ดังนี้ (Hambleton et al., 1991, pp. 12-14)

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \qquad (i = 1, 2, 3, ..., n) \tag{33}$$

เมื่อ $P_i(\theta)$ แทน ความน่าจะเป็นที่ผู้ตอบซึ่งมีความสามารถจะตอบข้อสอบข้อที่ i ได้ถูกต้อง

b_i แทน ค่าพารามิเตอร์ความยากของข้อสอบข้อที่ i

e แทน ค่าคงที่มีค่าเท่ากับ 2.718

n แทน ลำดับข้อสอบข้อที่ i

ถึงแม้ว่ารูปแบบนี้จะเป็นกรณีเฉพาะของรูปแบบ 2 พารามิเตอร์ และ 3 พารามิเตอร์ แต่ก็ยังมีคุณสมบัติพิเศษที่ทำให้นิยมใช้กัน คือ ประการแรก เนื่องจากรูปแบบนี้มี จำนวนพารามิเตอร์ไม่มากจึงสะดวกต่อการใช้งาน ประการที่สอง ปัญหาที่เกิดจากการประมาณค่าพารามิเตอร์มีน้อยกว่าการประมาณค่าพารามิเตอร์สาหรับรูปแบบที่มีพารามิเตอร์หลาย ๆ ตัว

โมเดลการตอบสนองข้อสอบแบบ 2 พารามิเตอร์ (Two – Parameter Model)

โมเดลแบบ 2 พารามิเตอร์ เป็นโค้งคุณลักษณะข้อสอบและเป็นฟังก์ชันของการแจกแจงที่มี 2 พารามิเตอร์ คือ ค่าความยากของข้อสอบ (b) ค่าอำนาจจำแนกของข้อสอบ (a) และโมเดล 2PL มีความเหมาะสมสาหรับการวัดคุณลักษณะแฝงที่แต่ละคนมีไม่เท่ากัน สามารภเขียนเป็นสมการได้ ดังนี้ (Hambleton et al., 1991, pp. 14-17)

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad (i = 1, 2, 3, ..., n) \tag{34}$$

เมื่อ $P_i(\theta)$ แทน เป็นค่าความน่าจะเป็นของผู้สอบที่มีความสามารถ θ สามารถตอบข้อสอบข้อที่ i ได้ถูกต้อง

b_i แทน ค่าพารามิเตอร์ความยากของข้อสอบข้อที่ i

a_i แทน ค่าพารามิเตอร์อำนาจจำแนกของข้อสอบข้อที่ i

D แทน ค่าสเกลองค์ประกอบ มีค่าเท่ากับ 1.7

$P_i(\theta)$ จากโค้งความถี่สะสมกับโค้งโลจีสติก จะมีค่าที่ต่างกันน้อยกว่า 0.01

สาหรับทุกค่าของ θ จากรูปแบบนี้อยู่บนข้อตกลงที่ว่าการเดาคาตอบจะไม่เกิดขึ้น ซึ่งถ้าจะเป็นเช่นนี้ได้ก็ต่อเมื่อค่าพารามิเตอร์ $a_i > 0$ (ข้อสอบที่มีความสัมพันธ์ด้านบวกระหว่างคะแนนจากการสอบกับความสามารถของผู้สอบที่วัดโดยแบบทดสอบนั้น) และค่าความน่าจะเป็นในการตอบข้อสอบได้ถูกจะลดลงถึงศูนย์เมื่อความสามารถลดลง

โมเดลการตอบสนองข้อสอบแบบ 3 พารามิเตอร์ (Three – Parameter Model)

โมเดลแบบ 3 พารามิเตอร์ เป็นการปรับปรุงมาจาก 2 พารามิเตอร์โดยการเพิ่มพารามิเตอร์ที่ 3 คือ พารามิเตอร์โอกาสการเดาของข้อสอบ หรือพารามิเตอร์ (c) เข้าไปในรูปแบบนี้ ดังในข้อสอบแบบหลายตัวเลือก ความน่าจะเป็นของการตอบถูกมากกว่า 0 แม้ว่าผู้สอบจะมีความสามารถต่ำ ซึ่งสามารถเขียนในรูปแบบสมการได้ ดังนี้ (Hambleton et al., 1991, pp. 17-18)

$$P_i(\theta) = c_i + (1-c_i)\frac{e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}} \qquad (i = 1, 2, 3, ..., k) \tag{35}$$

เมื่อ $P_i(\theta)$ แทน ความน่าจะเป็นของผู้สอบที่มีความสามารถตอบข้อสอบข้อที่ i ได้ถูกต้อง

b_i แทน พารามิเตอร์ความยากของข้อสอบข้อที่ i

a_i แทน พารามิเตอร์อำนาจจำแนกของข้อสอบข้อที่ i

c_i แทน พารามิเตอร์โอกาสการเดาข้อสอบได้ถูกของข้อสอบข้อที่ i

D แทน ค่าสเกลองค์ประกอบ มีค่าเท่ากับ 1.7

พารามิเตอร์ c_i เป็นจุดต่ำสุดที่โค้งคุณลักษณะข้อสอบ ซึ่งพารามิเตอร์นี้จะใช้เมื่อคิดว่าการเดาเป็นองค์ประกอบในการตอบข้อสอบ บางครั้งเรียกพารามิเตอร์นี้ว่าโอกาสที่จะตอบข้อสอบได้ถูกต้อง สำหรับคนที่มีความสามารถต่ำในการปรับรูปแบบสามพารามิเตอร์ให้เป็นรูปแบบ 2 พารามิเตอร์ ต้องอยู่บนข้อตกลงที่ว่า $c_i = 0$

ฟังก์ชันสารสนเทศ (Information Function)

ฟังก์ชันสารสนเทศจะเกี่ยวข้องกับการคัดเลือกข้อสอบ การพัฒนาแบบทดสอบ และการประเมินความแม่นยำของการประมาณค่าความสามารถของผู้สอบ ซึ่งในที่นี้จะกล่าวถึงฟังก์ชันสารสนเทศของข้อสอบ และฟังก์ชันสารสนเทศของแบบทดสอบ

1. ฟังก์ชันสารสนเทศของข้อสอบ (Item Information Function)

การอธิบายข้อสอบและแบบทดสอบ การเปรียบเทียบประสิทธิภาพของแบบทดสอบของทฤษฎีการตอบสนองข้อสอบ และการเปรียบเทียบประสิทธิภาพของโมเดล ต้องอาศัยคุณสมบัติของฟังก์ชันสารสนเทศของข้อสอบ โดยฟังก์ชันสารสนเทศของข้อสอบเป็นความสัมพันธ์ของอัตราส่วนระหว่าง กำลังสองค่าอนุพันธ์ของโอกาสในการตอบข้อสอบถูกของผู้สอบที่ระดับความสามารถนั้น ๆ กับผลคูณของโอกาสในการตอบข้อสอบถูก และผิดของผู้สอบในระดับความ สามารถนั้น ๆ เขียนเป็นสมการ ได้ดังนี้

$$I_i(\theta) = \frac{\left[P_i'(\theta)\right]^2}{P_i(\theta)Q_i(\theta)} \qquad (i = 1, 2, 3, ..., k) \tag{36}$$

เมื่อ $I_i(\theta)$ แทน ค่าฟังก์ชันสารสนเทศ หรือค่าสารสนเทศที่ได้รับจากข้อสอบข้อที่ i สาหรับผู้ตอบที่มีความสามารถ θ

$P_i'(\theta) = P_i'$ แทน ความชันของฟังก์ชันการตอบข้อสอบข้อที่ i ณ ตำแหน่งความสามารถ θ

$P_i(\theta) = P_i$ แทน ความน่าจะเป็นที่ผู้ตอบซึ่งมีความสามารถ θ จะตอบข้อสอบข้อที่ i ได้ถูกต้อง

$$Q_i(\theta) = Q_i = 1 - P_i(\theta) \tag{37}$$

ฟังก์ชันสารสนเทศของข้อสอบแต่ละข้อรวมกันเป็นฟังก์ชันสารสนเทศของแบบทดสอบ ซึ่งค่าฟังก์ชันสารสนเทศของข้อสอบขึ้นอยู่กับค่าความชันของฟังก์ชันการตอบข้อสอบ และค่าความแปรปรวนที่เงื่อนไขแต่ละระดับของความสามารถหรือคุณลักษณะ ค่าความชันสูง และค่าความแปรปรวนต่ำทำให้ค่าสารสนเทศของข้อสอบมีค่าสูง และทำให้ค่าของความคลาดเคลื่อนมาตรฐานในการวัดมีค่าต่ำ การแจกแจงของฟังก์ชันสารสนเทศของข้อสอบมีลักษณะเป็นรูประฆังคว่ำ ค่าสารสนเทศที่สูงที่สุดจะอยู่ที่จุด b บนสเกลความสามารถสำหรับโมเดลการตอบแบบโลจิสติกแบบ 1 และ 2 พารามิเตอร์ ส่วนโมเดล 3 พารามิเตอร์นั้น ค่าสารสนเทศของข้อสอบข้อที่ i จะสูงสุดที่จุด $\theta_{\max}$ เมื่อ

$$Q_{\max} = b_i + \frac{1}{Da_i}\left[In\left(\frac{1+\sqrt{(1+8c_i)}}{2}\right)\right] \tag{38}$$

เมื่อ $\theta_{\max}$ แทน ค่าความสามารถของผู้สอบสูงสุด

D แทน ค่าคงที่ซึ่งมีค่าเท่ากับ 1.70

สำหรับโมเดล 1 พารามิเตอร์นั้น ค่าสูงสุดของสารสนเทศของข้อสอบจะคงที่ และขณะเดียวกันโมเดล 2 พารามิเตอร์ ค่าสูงสุดของสารสนเทศของข้อสอบจะเป็นสัดส่วนโดยตรงกับกำลังสองของค่าอำนาจจำแนกของข้อสอบ ถ้าค่าอำนาจจำแนกของข้อสอบสูงก็จะทำให้ค่าสารสนเทศของข้อสอบมีค่ามาก ส่วนโมเดล 3 พารามิเตอร์นั้น ค่าสารสนเทศของข้อสอบสูงสุดจะมีค่า ดังนี้

$$I_i(\theta_{\max}) = D^2a^2\frac{\left[1 - 20c_i - 8c_i^2 + (1+8c_i)^{3/2}\right]}{8\left(1-c_i^2\right)} \tag{39}$$

เมื่อ $I_i(\theta_{\max})$ แทน ค่าสารสนเทศของข้อสอบสูงสุด

D แทน ค่าคงที่ซึ่งมีค่าเท่ากับ 1.70

ถ้าค่า c_i ลดลง ค่าสารสนเทศของข้อสอบก็จะเพิ่มขึ้น

ลักษณะสารสนเทศของข้อสอบ

1) ผลรวมค่าสารสนเทศของข้อสอบทุกข้อคือสารสนเทศของแบบทดสอบ

2) ค่าฟังก์ชันสารสนเทศขึ้นอยู่กับค่าความชันของฟังก์ชันการตอบข้อสอบ และค่าความแปรปรวนที่มีเงื่อนไขที่แต่ละระดับของความสามารถ ถ้าค่าความชันมาก และค่าความแปรปรวนต่ำทำให้ค่าสารสนเทศของข้อสอบมีค่าสูง ซึ่งทำให้ค่าความคลาดเคลื่อนมาตรฐานในการวัดมีค่าต่ำ

2. ฟังก์ชันสารสนเทศของแบบทดสอบ (Test Information)

การวิเคราะห์ตามทฤษฎี IRT จะใช้แบบแผนการตอบสนองแบบทดสอบเป็นรายข้อในการประมาณค่าความสามารถของผู้สอบ ดังนั้น การประเมินคุณภาพของแบบทดสอบจึงสามารถพิจาณาความถูกต้องแม่นยำในการประมาณค่าความสามารถของผู้ตอบ โดยใช้ดัชนีตัวหนึ่งเรียกว่าสารสนเทศของแบบทดสอบ (Test Information; $I(\theta)$) ซึ่งเป็นค่าฟังก์ชันสารสนเทศของแบบทดสอบอันเกิดจากผลรวมเชิงพีชคณิตของฟังก์ชันสารสนเทศของข้อสอบแต่ละข้อรวมเข้าด้วยกันทั้งฉบับ ณ ตำแหน่ง θ เดียวกัน (ศิริชัย กาญจนวาสี, 2550, หน้า 65) ดังสูตร

$$I(\theta)=\sum_{i=1}^{k} I_i(\theta) \tag{40}$$

เมื่อ $I(\theta)$ แทน ค่าฟังก์ชันสารสนเทศที่ได้รับจากแบบสอบสาหรับผู้ที่มีระดับความสามารถเท่ากับ θ

$I_i(\theta)$ แทน ค่าฟังก์ชันสารสนเทศที่ได้รับจากข้อสอบข้อที่ i สาหรับผู้สอบที่มีระดับความสามารถเท่ากับ θ

ลักษณะของฟังก์ชันสารสนเทศของแบบทดสอบ

1) ฟังก์ชันสารสนเทศของแบบทดสอบเป็นสิ่งที่ถูกกำหนดขึ้นสำหรับชุดของข้อสอบที่แต่ละจุดของสเกลความสามารถ

2) ค่าสารสนเทศของแบบทดสอบเป็นผลมาจากคุณภาพและจำนวนของข้อสอบ

3) ณ ตำแหน่งความสามารถเดียวกัน เส้นถดถอยที่มีความชันมากกว่าจะให้ค่าสารสนเทศของแบบทดสอบสูงกว่าเส้นถดถอยที่มีความชันน้อยกว่า

4) ข้อสอบที่มีค่าความแปรปรวนต่ำจะส่งผลให้ค่าสารสนเทศของแบบทดสอบสูง

5) ค่าสารสนเทศของแบบทดสอบจะไม่ขึ้นอยู่กับการจัดหมวดหมู่เฉพาะของข้อสอบและข้อสอบแต่ละข้อเป็นอิสระจากกัน

6) ค่าสารสนเทศของแบบทดสอบมีความสัมพันธ์แบบผกผันกับค่าความคลาดเคลื่อน

มาตรฐานในการประมาณค่าความสามารถที่ระดับเดียวกัน ดังสมการ

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}} \tag{41}$$

เมื่อ $SE(\theta)$ แทน ค่าความคลาดเคลื่อนมาตรฐานในการประมาณค่าที่ระดับความสามารถ θ

$I(\theta)$ แทน ค่าฟังก์ชันสารสนเทศที่ได้รับจากแบบสอบสาหรับผู้ที่มีระดับความสามารถเท่ากับ

สรุปว่าฟังก์ชันสารสนเทศตามทฤษฎีการตอบสนองข้อสอบนี้ประกอบด้วย ฟังก์ชันสารสนเทศของข้อสอบ และฟังก์ชันสารสนเทศของแบบทดสอบ ซึ่งฟังก์ชันสารสนเทศของข้อสอบขึ้นอยู่กับค่าความชันและค่าความแปรปรวน กล่าวคือ ถ้าค่าความชันสูงแต่ค่าความแปรปรวนต่ำจะทำให้ค่าสารสนเทศของข้อสอบมีค่าสูง และผลรวมของค่าสารสนเทศของข้อสอบทุกข้อ คือ สารสนเทศของแบบทดสอบ ซึ่งสารสนเทศของแบบทดสอบจะแสดงถึงความถูกต้องแม่นยำในการประมาณค่าความสามารถของผู้สอบ ที่เกิดจากผลรวมทางพีชคณิตของฟังก์ชันสารสนเทศของข้อสอบแต่ละข้อทั้งฉบับ ณ ตำแหน่ง θ เดียวกัน นอกจากนี้ค่าสารสนเทศของแบบทดสอบจะแปรผกผันกับค่าความคลาดเคลื่อนมาตรฐานในการประมาณค่าความสามารถอีกด้วย

วิธีอัตราส่วนไลค์ลิฮูด (Likelihood Ratio: LR)

วิธีอัตราส่วนไลค์ลิฮูด หรือวิธี IRT-LR โดยทั่วไป ใช้หลักของอัลกอริทึมในการประมาณค่าความเป็นไปได้สูงสุดในการประมาณค่าพารามิเตอร์ ในข้อมูลที่จำกัดวิธี IRT-LR ใช้การประมาณการกำลังสองน้อยมากสำหรับแบบจำลองการตอบสนองข้อสอบของข้อสอบแบบปกติ และยังใช้ข้อมูลที่ต่ำกว่าเกณฑ์ของการตอบสนองรูปแบบการจำแนกผู้ตอบข้อสอบ (Marie Wiberg, 2007) โดยที่วิธี IRT-LR จะประเมินความสำคัญระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบระหว่างความแตกต่างของค่าพารามิเตอร์ (Kim & Cohen, 1998) และ Thissen, David, Lynne Steinberg, and Howard Wainer (1988) เสนอไว้ว่าวิธีการทดสอบ IRT-LR เป็นที่นิยมเพราะมีการเปรียบเทียบของพารามิเตอร์และวัดพื้นที่ที่ต้องมีการประมาณการที่ถูกต้อง ในวิธี IRT-LR ข้อสอบบางข้ออาจจะเข้าข้างผู้สอบทั้งกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ซึ่งจะเรียกข้อสอบเหล่านั้นว่า DIF-free โดยมีข้อจำกัดอยู่ระหว่างสองกลุ่ม

วิธี IRT-LR สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ทั้งแบบเอกรูปและอเนกรูป ซึ่งมีการเปรียบเทียบกันกับแบบจำลองที่มีข้อจำกัดเท่ากันทั้งสองกลุ่ม สถิติที่ใช้ในการทดสอบ คือ ความแตกต่างระหว่างค่าความเป็นไปได้ของ $-2\log$ สำหรับกลุ่ม L_C และกลุ่ม L_A ตามค่าที่ถูกกำหนด Acar (2010) โดยสามารถเขียนสมการการวิเคราะห์ได้ ดังนี้

$$G^2(d.f.) = -2\log L_C - (-2\log L_A) \tag{42}$$

เมื่อ G^2 แสดงการกระจายของค่าไค-สแควร์ χ^2 และค่าพารามิเตอร์รายข้อ ทั้งยังสามารถพิจารณาจากค่าของ *p*-value ที่มีนัยสำคัญทางสถิติที่ .05 เมื่อข้อสอบมีการทำหน้าที่ต่างกัน

งานวิจัยที่เกี่ยวข้องกับวิธี IRT-LR มีดังนี้

รุ่งนภา แสนอานวยผล (2555) ศึกษาประสิทธิภาพของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนรูปแบบผสมโดยการประยุกต์ใช้ทฤษฎีการตอบสนองข้อสอบแบบตรวจให้คะแนนความรู้บางส่วน และทฤษฎีการตอบสนองข้อสอบแบบตรวจให้คะแนนความรู้บางส่วนแบบทั่วไป เพื่อศึกษาประสิทธิภาพของแบบทดสอบรูปแบบผสม ปฏิสัมพันธ์ระหว่างโมเดลการตรวจสอบการให้คะแนน ทั้งแบบข้อสอบที่ตรวจให้คะแนนแบบสองค่าและมากกว่าสองค่า รวมทั้งความยาวของแบบทดสอบ เพื่อเปรียบเทียบประสิทธิภาพของแบบทดสอบ เมื่อโมเดลการตรวจให้คะแนนสัดส่วนของข้อสอบ และความยาวของข้อสอบแตกต่างกัน และศึกษาขนาดอิทธิพลของโมเดลการตรวจให้คะแนน สัดส่วนของข้อสอบและความยาวของแบบสอบ ซึ่งมีเงื่อนไขทั้งหมด 18 เงื่อนไข ซึ่งโมเดลการตรวจให้คะแนน 2 โมเดล คือ โมเดลโลจิสติก 1 พารามิเตอร์และโมเดลการตรวจให้คะแนนความรู้บางส่วน (PCM) และใช้โมเดลโลจิสติกแบบ 3 พารามิเตอร์ สำหรับการตรวจให้คะแนนความรู้บางส่วนแบบทั่วไป (GPCM) และข้อสอบที่ตรวจให้คะแนนแบบสองค่าและมากกว่าสองค่า แบ่งออกเป็น 3 สัดส่วน คือ 20:80 50:50 80:20 รวมถึงความยาวของแบบทดสอบ มี 3 เงื่อนไข คือ 10 30 และ 50 ข้อ โดยการประเมินประสิทธิภาพแบบทดสอบพิจารณาจากดัชนีความคลาดเคลื่อนมาตราฐานของการประมาณค่า $(SE(\theta))$ และ BIAS ในการหาค่าปฏิสัมพันธ์ จากการศึกษาผลวิจัยพบว่าโมเดลโลจิสติก 1 พารามิเตอร์ กับการตรวจให้คะแนนความรู้บางส่วน (PCM) และโมเดลโลจิสติกแบบ 3 พารามิเตอร์กับการตรวจให้คะแนนความรู้บางส่วนแบบทั่วไป (GPCM) มีค่า $(SE(\theta))$ และ BIAS ต่ำสุดในสัดส่วนของการตรวจให้คะแนนแบบสองค่าและมากกว่าสองค่า คือ 20:80 และมีความยาวของแบบสอบ 50 ข้อ และมีปฏิสัมพันธ์ระหว่างโมเดลการตรวจให้คะแนน กับสัดส่วนการตรวจให้คะแนนแบบสองค่าและมากกว่าสองค่า รวมทั้งความยาวของแบบทดสอบ ที่ส่งผลต่อค่า $(SE(\theta))$ และ BIAS อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 และการศึกษาขนาดอิทธิพล พบว่า ความยาวของแบบสอบมีอิทธิพลต่อค่า $(SE(\theta))$ ทุกระดับความสามารถในระดับที่สูงมาก และมีอิทธิพลต่อค่า BIAS ในระดับปานกลาง สรุปได้ว่า ความยาวของแบบสอบส่งผลทั้งค่า $(SE(\theta))$ และค่า BIAS รวมถึงปฏิสัมพันธ์ระหว่างโมเดลการตรวจให้คะแนนและสัดส่วนการตรวจให้คะแนน อยู่ในระดับปานกลางค่อนไปในทางสูงมาก

Teresi (2007) ศึกษาการประเมินผลการวัดความเท่าเทียมโดยการใช้ทฤษฎีการตอบสนองข้อสอบแบบอัตราส่วนไลค์ลิฮูด (IRT-LR) สำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ในแบบวัดความสามารถและความทุกข์ในการทำงาน โดยมีเงื่อนไขในการตรวจสอบเพศ อายุ และการแข่งขัน มีกลุ่มตัวอย่างเป็นผู้ป่วยมะเร็ง จำนวน 1,714 คน เครื่องมือที่ใช้ในการวิจัยเป็นแบบวัดความสามารถทางกายภาพ จำนวน 23 ข้อ และแบบวัดทางอารมณ์ จำนวน 15 ข้อ จากการวิจัยพบว่า อายุ และความสามารถในการแข่งขัน มีผลต่อการที่จะได้รับการปฏิบัติที่แตกต่างกันออกไป ซึ่งเป็นปัจจัยที่อาจจะส่งผลต่อผู้ป่วยที่จะเลือกเข้ารับการรักษาของสถานรักษาพยาบาล

Acar and Kelecioglu (2010) ศึกษาเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ระหว่างวิธี HGLM วิธี LR และวิธี IRT-LR โดยมีวัตถุประสงค์การวิจัยครั้งนี้เป็นการตรวจสอบความสอดคล้องระหว่างวิธีที่ใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ 3 วิธี ซึ่งมีเงื่อนไขเป็นเพศ

ในการตรวจสอบ มีกลุ่มตัวอย่างเป็นนักเรียนในประเทศตุรกี และเครื่องมือที่ใช้ในการวิจัยคือแบบทดสอบของวิชาสังคมศาสตร์และวิทยาศาสตร์ จากการศึกษาพบว่า วิธีการตรวจสอบการทำหน้าที่ต่างกันขอข้อสอบทั้ง 3 วิธี คือ วิธี HGLM วิธี LR และวิธี IRT-LR มีการตรวจพบข้อสอบที่ทำหน้าที่ต่างกันในปริมาณที่ใกล้เคียงกัน แต่ไม่ตรงกัน โดยวิธี LR และวิธี IRT-LR จะตรวจพบข้อสอบที่ทำหน้าที่ต่างกันน้อยทั้งแบบทดสอบสังคมศาสตร์และวิทยาศาสตร์ ขณะที่วิธี HGLM จะตรวจสอบพบข้อสอบที่ทำหน้าที่ต่างกันในแบบทดสอบทั้ง 2 แบบทดสอบในปริมาณที่มากที่สุด

Pae (2012) ศึกษาการทำงานของข้อสอบที่แตกต่างระหว่างเพศ ในการทดสอบย่อยภาษาอังกฤษของการทดสอบความถนัดทางวิชาการของนักวิชาการวิทยาลัยเกาหลี (Korean College Scholastic Aptitude Test: KCSAT) ในช่วงเวลาเก้าปีในชุดข้อมูล 3 จุด โดยใช้ทั้งอัตราส่วน Mantel-Haenszel (MH) และทฤษฎีการตอบสนองข้อสอบแบบอัตราส่วยไลค์ลิฮูด (IRT-LR) ขั้นตอนต่าง ๆ นอกจากนี้การศึกษายังระบุถึงปัจจัย 2 ประการ ได้แก่ กลยุทธ์การอ่านและการรับรู้ความสนใจ ซึ่งอธิบายถึงความแปรปรวนของขนาดการทำหน้าที่ต่างกันที่มีต่อเพศ โดยอาศัยการวิเคราะห์การถดถอยเชิงเส้นหลายแบบ จากงานวิจัยพบว่า ปฏิสัมพันธ์การทำหน้าที่ต่างกันของข้อสอบระหว่างข้อสอบรายข้อกับเพศ และความสัมพันธ์ที่สำคัญระหว่างความแตกต่างของเพศ ความสนใจในผลการทดสอบและขนาดในการทำหน้าที่ต่างกันของเพศ โดยการรับรู้และการอ่านมีผลต่อการทำหน้าที่ต่างกันของข้อสอบของเพศ

Li, Hunter, and Oshima (2013) ศึกษาการตรวจสอบการทำหน้าที่ต่างกันในการทดสอบการอ่านและเหตุผลที่เป็นไปได้ระหว่างเพศ จากการศึกษาค้นคว้าอย่างละเอียดแล้ว 1,210 ข้อ จาก 18 บทความถูกรวมไว้ในการวิเคราะห์ขั้นสุดท้าย พบว่า 23.3% ของข้อสอบแสดงให้เห็นถึงการทำหน้าที่ต่างกันของข้อสอบในตัวแปรเพศ อย่างไรก็ตามมีการเปลี่ยนแปลงเปอร์เซ็นต์ของข้อสอบที่กำหนดว่าแสดงการทำหน้าที่ต่างกัน ระหว่างการศึกษาตั้งแต่ 0 ถึง 77% ของข้อสอบที่ทำหน้าที่ต่างกันที่ครึ่งหนึ่งลำเอียงเข้าทางเพศชายและอีกครึ่งหนึ่งที่ลำเอียงเข้าข้างเพศหญิง โดยรูปแบบนี้เป็นจริงสำหรับการศึกษาโดยใช้การทดสอบการตอบสนองต่อข้อสอบ (IRT-LR) และสำหรับผู้ที่ใช้ Mantel-Haenszel (MH) นอกจากนี้รายการจากการทดสอบที่สั้นกว่ามีแนวโน้มที่จะได้รับการพิจารณาว่าทำหน้าที่ต่างกัน มากกว่าที่ข้อสอบจากการทดสอบที่ยาวกว่า รูปแบบการตรวจสอบการทำหน้าที่ต่างกันอื่น ๆ จะขึ้นอยู่กับวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ และการทดสอบ

Kabasakal, Arsan, Gok, and Kelecioglu (2014) ศึกษาแบบจำลองเปรียบเทียบผลอัตราความคลาดเคลื่อนประเภทที่ 1 และประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของวิธี Mantel-Haenszel (MH) SIBTEST และวิธีการตอบสนองต่อข้อสอบ (IRT-LR) ภายใต้เงื่อนไขบางประการ ปัจจัยการจัดการคือขนาดตัวอย่างความแตกต่างระหว่างกลุ่มความยาวการทดสอบเปอร์เซ็นต์ของการทำงานของรายการที่แตกต่างกัน (DIF) และรูปแบบพื้นฐานที่ใช้ในการสร้างข้อมูล ผลการวิจัยแสดงให้เห็นว่า SIBTEST มีข้อผิดพลาดประเภทที่สูงที่สุดในการตรวจจับ DIF สม่ำเสมอแต่ MH มีอำนาจสูงสุดในทุกสภาวะ นอกจากนี้เปอร์เซ็นต์ของ DIF และโมเดลพื้นฐานยังมีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 1 ของ IRT-LR ความแตกต่างของความแตกต่างระหว่างกลุ่มความยาวทดสอบเปอร์เซ็นต์ของ DIF รูปแบบและความสัมพันธ์ระหว่างความแตกต่างของความสามารถร้อยละของ DIF ความแตกต่างของความสามารถความยาวในการทดสอบความยาวในการทดสอบร้อยละของ

DIF ความยาวของแบบทดสอบมีผลต่อวิธีการ SIBTEST อัตราความคลาดเคลื่อนประเภทที่ 1 ในขั้นตอน MH ปัจจัยที่มีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 1 ได้แก่ ขนาดตัวอย่างระยะเวลาในการทดสอบเปอร์เซ็นต์ DIF ความแตกต่างของความสามารถเปอร์เซ็นต์ของ DIF ความแตกต่างของความสามารถรูปแบบความสามารถและความแตกต่างของ DIF ไม่มีผลต่อประสิทธิภาพของ SIBTEST และ MH แต่รูปแบบพื้นฐานมีผลต่ออัตราการใช้ IRT-LR

จากการศึกษางานวิจัยที่เกี่ยวข้อง วิธี IRT-LR สามารถตรวจพบข้อสอบการทำหน้าที่ต่างกันระหว่างเพศได้มากกว่าวิธี SIBTEST และยังสามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ดีในแบบทดสอบที่มีความยาวไม่เกิน 30 ข้อ และตรวจพบข้อสอบการทำหน้าที่ต่างกันได้น้อยกว่าวิธี HGLM และควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้น้อย เมื่อแบบทดสอบมีความยาวกว่า 30 ข้อ และขนาดของกลุ่มตัวอย่างมีขนาดใหญ่

ตอนที่ 5 การทดสอบระดับชาติ (NT) และงานวิจัยที่เกี่ยวข้อง

สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน หรือ สพฐ. ได้มีการวางแผนการศึกษาเพื่อพัฒนาคุณภาพผู้เรียนทุกระดับและทุกประเภท ให้นักเรียนที่อยู่ในระดับการศึกษาขั้นพื้นฐานทุกคนมีพัฒนาการรวมทั้งให้นักเรียนระดับประถมศึกษามีพัฒนาการที่เหมาะสมตามช่วงวัยและมีคุณภาพ โดยการนำระบบการทดสอบกลางของสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน มาเป็นตัวช่วยในการจัดการเรียนการสอน และยังได้สนับสนุนให้มีการทดสอบ O-NET และการประเมินผลการศึกษาระดับชาติอย่าง PISA เพื่อพัฒนานักเรียนให้มีคุณภาพการศึกษาที่มีมาตราฐานเดียวกันทั้งประเทศ โดยในปีการศึกษา 2557 สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐานได้มีนโยบายทางด้านการศึกษาเพื่อพัฒนาความสามารถในการแข่งขันของประเทศร่วมประกอบกับการเตรียมความพร้อมที่จะรวมตัวกับประชาคมอาเซียน จึงได้ร่วมมือกับสำนักทดสอบทางการศึกษาเพื่อจัดการทดสอบระดับชาติขึ้นมา เพื่อเป็นส่วนหนึ่งในการประกันคุณภาพการศึกษาภายในสถานศึกษา และเพื่อที่จะเตรียมความพร้อมให้นักเรียนเพื่อที่จะรับการประเมินจากภายนอกสถานศึกษา ทั้งแบบทดสอบระดับชาติหรือระดับนานาชาติ (สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน, 2557)

การทดสอบระดับชาติ มีนโยบายที่จะใช้ประเมินกับนักเรียนชั้นประถมศึกษาปีที่ 3 ที่เป็นนักเรียนที่ศึกษาอยู่ภายในโรงเรียนในสังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน โรงเรียนในสังกัดสำนักงานคณะกรรมการส่งเสริมการศึกษาเอกชน และตำรวจตระเวนชายแดน ซึ่งข้อสอบประเมินตามความสามารถพื้นฐานที่สำคัญ 3 ด้าน คือ ด้านภาษา (Literacy) ด้านคำนวน (Numeracy) และด้านเหตุผล (Reasoning Abilities) โดยมีข้อสอบจำนวน 30 ข้อในการประเมินในแต่ละด้าน รวมทั้งหมด 3 ด้าน มีข้อสอบจำนวน 90 ข้อ สำนักทดสอบทางการศึกษา ได้ให้นิยามในแต่ละด้านความสามารถ 3 ด้าน ดังนี้

1. ด้านภาษา (Literacy) หมายถึง ความสามารถในการอ่าน ฟัง ดู และพูด เพื่อความรู้เข้าใจ สามารถวิเคราะห์ สรุปสาระสำคัญ ประเมินสิ่งที่อ่าน ฟัง ดู จากสื่อประเภทต่าง ๆ และสื่อสารด้วยการพูด การเขียน ได้ถูกต้องตามหลักการใช้ภาษาได้อย่างสร้างสรรค์ เพื่อนำไปใช้ในชีวิตประจำวัน การอยู่ร่วมกันในสังคม และการศึกษาตลอดชีวิต โดยทางสำนักทดสอบทางการศึกษาได้ให้ความหมายคำสำคัญไว้ดังนี้

1.1 รู้ หมายถึง สามารถบอกความหมาย เรื่องราว ข้อเท็จจริง และเหตุการณ์ต่าง ๆ

1.2 เข้าใจ หมายถึง สามารถแปลความ ตีความ ขยายความ และอ้างอิง

1.3 วิเคราะห์ หมายถึง สามารถแยกแยะโครงสร้าง เรื่องราว ข้อเท็จจริง ข้อคิดเห็น คุณค่า และเหตุผล

1.4 สรุปสาระสำคัญ หมายถึง สามารถสรุปใจความสำคัญของเรื่องได้อย่างครอบคลุม

1.5 ประเมิน หมายถึง สามารถตัดสินความถูกต้อง ความชัดเจน ความเหมาะสม คุณค่า อย่างมีหลักเกณฑ์

1.6 สื่อประเภทต่าง ๆ หมายถึง สิ่งที่นำเสนอเรื่องราวและข้อมูลความรู้ต่าง ๆ ทั้งที่เป็นสื่อสิ่งพิมพ์ สื่ออิเล็กทรอนิกส์ และสื่อของจริง

1.7 สื่อสาร หมายถึง สามารถถ่ายถอดความรู้ ความเข้าใจ และความคิด จากการอ่าน ฟัง และดู โดยการพูดและอธิบาย วิเคราะห์ สรุป หรือประเมิน

1.8 สร้างสรรค์ หมายึง สามารถสื่อสารความรู้ความเข้าใจ เรื่องราว ทัศนะและความคิดที่แปลกใหม่จากการอ่าน การฟังและดู แสดงออกมาเป็นคำพูด การเขียน หรือการกระทำได้อย่างหลากหลายและมีประโยชน์เพิ่มมากขึ้น

1.9 การนำไปใช้ในชีวิตประจำวัน การอยู่ร่วมกันในสังคมและการศึกษาตลอดชีวิต หมายถึง ความสามารถในการนำความรู้ ความเข้าใจ การวิเคราะห์ การสรุปสาระสำคัญไปใช้ประโยชน์ในการแก้ไขปัญหา การตัดสินใจในการดำเนินชีวิต การอยู่ร่วมกับผู้อื่น และการพัฒนาตนเองอย่างต่อเนื่อง

2. ด้านคำนวณ (Numeracy) หมายถึง ความสามารถของการใช้ทักษะในกระบวนการทางคณิตศาสตร์ การคิดคำนวณ และความคิดรวบยอดทางคณิตศาสตร์ในสถานการณ์ต่าง ๆ ที่เกี่ยวข้องกับชีวิตประจำวัน ซึ่งคำสำคัญมีความหมายดังนี้

2.1 ทักษะกระบวนการทางคณิตศาสตร์ หมายถึง ความสามารถในการแก้ปัญหาด้วยวิธีการที่หลากหลาย การให้เหตุผล การสื่อสาร การสื่อความหมายทางคณิตศาสตร์ การนำเสนอ การเชื่อมโยงความรู้ และการมีความคิดริเริ่มสร้างสรรค์

2.2 ทักษะการคิดคำนวณ หมายถึง ความสามารถในการบวก ลบ คูณ และหาร ได้อย่างถูกต้อง และคล่องแคล่ว

2.3 ความคิดรวบยอดทางคณิตศาสตร์ หมายถึง ความรู้ความเข้าใจเกี่ยวกับ จำนวนนับ เศษส่วน ทศนิยม ร้อยละ ความยาว น้ำหนัก ระยะทาง พื้นที่ ปริมาตร เวลา เงิน ความจุ แผนผัง ทิศ และขนาดของมุม ชนิดและสมบัติของรูปเรขาคณิต แบบรูปและความสัมพันธ์ แผนภูมิและกราฟ การคาดคะเนการเกิดขึ้นของเหตุการณ์ต่าง ๆ

3. ด้านเหตุผล (Reasoning Abilities) หมายถึง ความสามารถในการเชื่อมโยงความรู้และประสบการณ์ด้านวิทยาศาสตร์ และสิ่งแวดล้อม ด้านสังคมศาสตร์ และเศรษฐศาสตร์ ด้านการดำเนินชีวิต โดยการวิเคราะห์ สังเคราะห์ ประเมินค่า และตัดสินใจอย่างมีหลักการ และเหตุผล บนพื้นฐานข้อมูลสถานการณ์ หรือสารสนเทศที่เพียงพอ โดยยึดหลักคุณธรรมและจริยธรรม โดยมีความหมายของคำสำคัญดังนี้

3.1 ความรู้ หมายถึง ข้อเท็จจริง ทฤษฎี หลักการ กระบวนการที่ศึกษารวมทั้งคุณธรรมจริยธรรม

3.2 ประสบการณ์ หมายถึง ความรู้เดิมที่เกิดจากการเรียนรู้ ปฏิบัติ หรือได้พบเห็นเรื่องต่าง ๆ ในระดับบุคคล สังคม และสังคมโลก

3.3 วิเคราะห์ หมายถึง ความสามารถในการเปรียบเทียบ บอกความแตกต่าง ความเหมือน สรุปหลักการ บอกความสัมพันธ์เชื่อมโยงอย่างมีเหตุผลบนพื้นฐานของหลักการทางวิทยาศาสตร์ สังคมศาสตร์ และการดำเนินชีวิต อย่างมีคุณธรรมและจริยธรรม

3.4 สังเคราะห์ หมายถึง ความสามารถในการสร้างข้อสรุปใหม่ ออกแบบ คิดสร้างสรรค์ บนพื้นฐานของข้อมูลที่ผ่านการวิเคราะห์ ประเมินแล้วอย่างสมเหตุสมผล

3.5 ประเมินค่า หมายถึง ความสามารถในการตัดสินใจเลือกทางเลือกอย่างสมเหตุสมผล มีประโยชน์และสร้างสรรค์

3.6 เหตุผลทางวิทยาศาสตร์ หมายถึง การนำความรู้ ประสบการณ์ที่เกิดจากการเรียนรู้มาประกอบการตัดสินใจในสถานการณ์ที่เกิดขึ้นในสังคม ให้สมเหตุสมผลตามหลักเกณฑ์ทางวิทยาศาสตร์

3.7 เหตุผลทางสังคมศาสตร์ หมายถึง การนำความรู้ ประสบการณ์จากกฎเกณฑ์ ความเชื่อ วัฒนธรรม ค่านิยมทางสังคมศาสตร์มาประกอบการตัดสินใจในสถานการณ์ที่เกิดขึ้นในสังคมได้อย่างสมเหตุสมผล

3.8 เหตุผลทางการดำเนินชีวิต หมายถึง การนำความรู้ หลักการ กฎเกณฑ์ มาใช้ในการดำรงชีวิตหรือประกอบการตัดสินใจในสถานการณ์ที่เกิดขึ้นในสังคมอย่างมีคุณธรรมจริยธรรม

เครื่องมือที่ใช้ในการประเมินนักเรียนชั้นประถมศึกษาปีที่ 3 มีรายละเอียด ดังนี้

ตารางที่ 2-4 รายละเอียดเครื่องมือที่ใช้ประเมินนักเรียน ชั้นประถมศึกษาปีที่ 3

แบบทดสอบความสามรถ	จำนวน (ข้อ)	เวลา (นาที)
ด้านภาษา (Literacy)	30	60
ด้านคำนวณ (Numeracy)	30	60
ด้านเหตุผล (Reasoning Abilities)	30	60
รวม	90	180

โครงสร้างของแบบทดสอบระดับชาติ ปีการศึกษา 2556

1. ด้านภาษา มีโครงสร้างในการประเมิน ดังนี้

ตารางที่ 2-5 โครงสร้างข้อสอบด้านภาษาของแบบทดสอบระดับชาติ ปีการศึกษา 2556

ตัวชี้วัด	จำนวน (ข้อ)
1. บอกความหมายของคำและประโยคจากเรื่องที่ฟัง ดู อ่าน	5
2. บอกความหมายของเครื่องหมาย/สัญลักษณ์	3
3. ตอบคำถามจากเรื่องที่ฟัง ดู อ่าน	5
4. บอกเล่าเรื่องราวที่ได้จากการฟัง ดู อ่าน อย่างง่าย ๆ	4
5. คาดคะเนเหตุการณ์ที่จะเกิดขึ้นจากเรื่องที่ฟัง ดู อ่าน	6
6. สื่อสารความรู้ ความเข้าใจ ข้อคิดเห็นจากเรื่องที่ ฟัง ดู อ่าน	7
รวม	30

2. ด้านคำนวณ มีโครงสร้างในการประเมิน ดังนี้

ตารางที่ 2-6 โครงสร้างข้อสอบด้านคำนวณของแบบทดสอบระดับชาติ ปีการศึกษา 2556

ตัวชี้วัด	จำนวน (ข้อ)
1. ใช้ทักษะกระบวนการทางคณิตศาสตร์ หรือทักษะการคิดคำนวณเพื่อตัดสินใจเลือกแนวทางปฏิบัติหรือหาคำตอบ จากสถานการณ์ต่าง ๆ ในชีวิตประจำวันที่กำหนดเกี่ยวกับความคิดรวบยอดทางคณิตศาสตร์ เรื่องจำนวนและการดำเนินการตามขอบข่ายเนื้อหา	10
2. ใช้ทักษะกระบวนการทางคณิตศาสตร์ หรือทักษะการคิดคำนวณเพื่อตัดสินใจเลือกแนวทางปฏิบัติหรือหาคำตอบจากสถานการณ์ต่าง ๆ ในชีวิตประจำวันที่กำหนดเกี่ยวกับความคิดรวบยอดทางคณิตศาสตร์ เรื่องการวัดตามขอบข่ายเนื้อหา	8
3. ใช้ทักษะกระบวนการทางคณิตศาสตร์ หรือทักษะการคิดคำนวณเพื่อตัดสินใจเลือกแนวทางปฏิบัติหรือหาคำตอบ จากสถานการณ์ต่าง ๆ ในชีวิตประจำวันที่กำหนดเกี่ยวกับความคิดรวบยอดทางคณิตศาสตร์ เรื่องเรขาคณิตตามขอบข่ายเนื้อหา	4
4. ใช้ทักษะกระบวนการทางคณิตศาสตร์ หรือทักษะการคิดคำนวณเพื่อตัดสินใจเลือกแนวทางปฏิบัติหรือหาคำตอบ จากสถานการณ์ต่าง ๆ ในชีวิตประจำวันที่กำหนดเกี่ยวกับความคิดรวบยอดทางคณิตศาสตร์ เรื่องพีชคณิตตามขอบข่ายเนื้อหา	4

ตารางที่ 2-6 (ต่อ)

ตัวชี้วัด	จำนวน (ข้อ)
5. ใช้ทักษะกระบวนการทางคณิตศาสตร์ หรือทักษะการคิดคำนวณเพื่อตัดสินใจเลือกแนวทางปฏิบัติหรือหาคำตอบ จากสถานการณ์ต่าง ๆ ในชีวิตประจำวันที่กำหนดเกี่ยวกับความคิดรวบยอดทางคณิตศาสตร์ เรื่องการวิเคราะห์ข้อมูลและความน่าจะเป็นตามขอบข่ายเนื้อหา	4
รวม	30

3. ด้านเหตุผล มีโครงสร้างข้อสอบ ดังนี้

ตารางที่ 2-7 โครงสร้างข้อสอบด้านเหตุผลของแบบทดสอบระดับชาติ ปีการศึกษา 2556

ตัวชี้วัด	จำนวน (ข้อ)
1. มีความรู้ความเข้าใจในข้อมูล สถานการณ์ หรือสารสนเทศทางด้านวิทยาศาสตร์และสิ่งแวดล้อม ด้านสังคมและเศรษฐกิจ และด้านการดำเนินชีวิต	4
2. วิเคราะห์ข้อมูล สถานการณ์ หรือสารสนเทศโดยใช้ความรู้ด้านวิทยาศาสตร์และสิ่งแวดล้อม ด้านสังคมและเศรษฐกิจ และด้านการดำเนินชีวิตอย่างมีเหตุผล	8
3. สามารถสร้างข้อสรุปใหม่ ออกแบบ วางแผน บนพื้นฐานของข้อมูลสถานการณ์หรือสารสนเทศที่ผ่านการวิเคราะห์โดยใช้องค์ความรู้ด้านวิทยาศาสตร์และสิ่งแวดล้อม ด้านสังคมและเศรษฐกิจ และด้านการดำเนินชีวิตอย่างมีเหตุผล	10
4. ให้ข้อสนับสนุน/โต้แย้งเพื่อการตัดสินใจและแก้ปัญหาอย่างมีหลักการและเหตุผลคำนึงถึงหลักคุณธรรมและจริยธรรม ค่านิยม ความเชื่อ ในกรณีที่สถานการณ์ที่ต้องการตัดสินใจ/หรือมีปัญหา	8

ลักษณะเครื่องมือเป็นแบบทดสอบแบบเลือกตอบ (Multiple Choices) 4 ตัวเลือก มีเกณฑ์การประเมิน ดังนี้

1. เกณฑ์การประเมินของแบบทดสอบด้านภาษา มีช่วงคะแนน ดังนี้

ปรับปรุง 0-10 คะแนน
พอใช้ 11-15 คะแนน
ดี 16-20 คะแนน
ดีมาก 21-30 คะแนน

2. เกณฑ์การประเมินของแบบทดสอบด้านคำนวน มีช่วงคะแนน ดังนี้

ปรับปรุง	0-6	คะแนน
พอใช้	7-11	คะแนน
ดี	12-16	คะแนน
ดีมาก	17-30	คะแนน

3. เกณฑ์การประเมินของแบบทดสอบด้านเหตุผล มีช่วงคะแนน ดังนี้

ปรับปรุง	0-8	คะแนน
พอใช้	9-13	คะแนน
ดี	14-19	คะแนน
ดีมาก	20-30	คะแนน

ตัวอย่างแบบทดสอบระดับชาติ (NT)

ด้านภาษา

อ่านข้อความที่กำหนด แล้วตอบคำถามข้อ 1

ก๋วยเตี๋ยวเรือรสเด็ด ! แม่แดงชวนชิม ชามละ 10 บาท

1. คำว่า รสเด็ด มีความหมายตรงกับข้อใด
   1) รสเผ็ด
   2) รสอร่อย
   3) รสเปรี้ยว
   4) รสกลมกล่อม

อ่านข้อความที่กำหนด แล้วตอบคำถามข้อ 2-3

พ่อแม่สมัยใหม่ มักปล่อยให้ลูกชมโทรทัศน์รายการการ์ตูนต่างๆ ซึ่งเป็นการทำลายสุขภาพจิตของลูกเพราะการที่เลี้ยงลูกด้วยรายการโปรดของน้อง ๆ หนู ๆ อาจทำให้ลูกมีความสุขได้ แต่ระยะนาน ๆ ไปมันจะทำลายสุขภาพทั้งทางร่างกาย และจิตใจ จากรายงานการวิจัย พบว่า เด็กที่ป่วยเป็นโรคสมาธิสั้นเกือบ 40% มาจากการชมโทรทัศน์ แม้จะไม่ส่งผลโดยตรง แต่ผลทางอ้อมก็คือเด็กจะมีการสื่อสารทางเดียว ปลีกตัวจากสังคม โอกาสที่จะได้เล่นกับเพื่อน ๆ น้อยลง

2. จากเรื่องที่อ่าน การสื่อสารทางเดียว หมายถึงข้อใด
   1) เป็นการดูอย่างเดียว
   2) เป็นการฟังอย่างเดียว
   3) เป็นการอ่านอย่างเดียว
   4) เป็นการรับมูลอย่างเดียว

3. ข้อความนี้ กล่าวถึงสิ่งใดเป็นสำคัญ
   1) วิธีการดูแลลูก
   2) ความรักของพ่อแม่
   3) ผลเสียจากการดูโทรทัศน์
   4) การดูโทรทัศน์อย่างถูกวิธี

4. “เพลงนี้ โดนใจ จริง ๆ ค่ะ” จากข้อความ คำที่ขีดเส้นใต้มีความหมายตรงกับข้อใด
   1) ถึงใจ
   2) ถูกใจ
   3) พอใจ
   4) ชอบใจ

อ่านข้อความที่กำหนด แล้วตอบคำถามข้อ 5

ปรีชาและคุณพ่อไปเยี่ยมคุณย่าที่โรงพยาบาล ขณะเดินผ่านตึกผู้ป่วยใน คุณพ่อ บอกว่า “ตรงนี้เขาห้ามใช้เสียงนะ” ปรีชาสงสัยว่าคุณพ่อรู้ได้อย่างไร พ่อจึงชี้ที่ป้ายสัญลักษณ์ที่ติดไว้ข้างฝา

5. นักเรียนคิดว่า คุณพ่อชี้ที่ป้ายสัญลักษณ์ใด
   1) 
   2) 
   3) 
   4) 

ด้านคำนวณ

1.

การเก็บมะนาวของชูศักดิ์ วันแรกเก็บได้ 20 ผล และวันถัดมาจะเก็บได้เพิ่มขึ้นจากวันเดิม วันละ 5 ผล ทุกวัน

ใน 1 สัปดาห์ ชูศักดิ์ เก็บมะนาวได้ทั้งหมดกี่ผล

1) 25
2) 100
3) 175
4) 245

2. การสะสมเงินของนักเรียน 3 คน

ชื่อนักเรียน	จำนวนเงินสะสมวันละ (บาท)
หน่อย	50
น้อย	70
นก	40

ในเวลา 3 วัน ทั้งสามคน มีเงินรวมกันกี่บาท

1) 480
2) 380
3) 320
4) 160

3.

สุดาขายไข่ไก่ ถุงละสิบฟอง ในราคา 35 บาท วันแรก ขายได้ 12 ถุง วันที่สอง ขายได้ 7 ถุง วันที่สาม ขายได้ 11 ถุง

จากข้อมูลข้างบน ให้นักเรียนพิจารณาข้อความต่อไปนี้

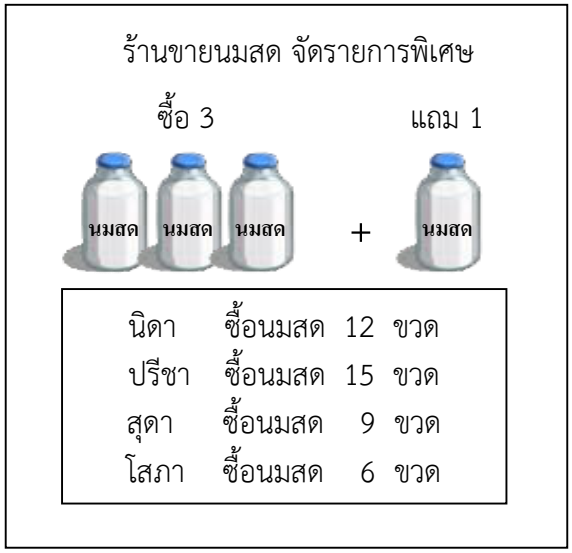
ก. วันแรกขายไข่ไก่ได้เงินมากกว่าวันที่สอง 105 บาท

ข. วันที่สามขายไข่ไก่ได้เงินมากกว่าวันที่สอง 140 บาท

ข้อใดถูกต้อง

1) ข้อ ก ถูก และข้อ ข ผิด
2) ข้อ ก ผิด และข้อ ข ถูก
3) ถูกทั้งข้อ ก และ ข้อ ข
4) ผิดทั้งข้อ ก และ ข้อ ข

ใช้ข้อมูลต่อไปนี้ ตอบคำถามข้อ 4–5



4. ถ้านมสดราคาขวดละ 12 บาท ทางร้านจะได้เงินจากนิดาและโสภารวมเป็นเงินกี่บาท

1) 120
2) 122
3) 216
4) 218

5. ร้านค้าแถมนมสดให้ปรีชาและสุดา เป็นจำนวนกี่ขวด

1) 4
2) 6
3) 8
4) 10

ด้านเหตุผล

พิจารณาภาพ แล้วตอบคำถามข้อ 1

1. จากภาพ ประชาชนในชุมชนนี้มีปัญหาเกี่ยวกับระบบทางเดินหายใจเป็นส่วนใหญ่ ข้อใดอธิบายสาเหตุของปัญหาของคนในชุมชนนี้ได้เหมาะสมที่สุด
   1) การเผาขยะทำให้สารพิษตกค้าง
   2) การเผาขยะเป็นการกระจายมลพิษ
   3) การเผาขยะในปริมาณที่มากเกินไป
   4) การเผาขยะเพิ่มความรุนแรงของกลิ่น

อ่านข้อความแล้วตอบคำถามข้อ 2

ชิดปลูกพริกใส่กระถางไว้ใต้ต้นไม้ใหญ่ พริกมีลำต้นเรียวสูง ใบมีสีเขียวอ่อน ออกดอกและให้ผลน้อย ทั้ง ๆ ที่ดูแลบำรุงรักษาอย่างดี เมื่อเขานำกระถางต้นพริกไปวางไว้กลางแดดและดูแลรักษาเหมือนเดิม จากนั้นพบว่าใบของพริกค่อย ๆ เปลี่ยนเป็นสีเขียวเข้ม มีดอกและให้ผลมากขึ้น

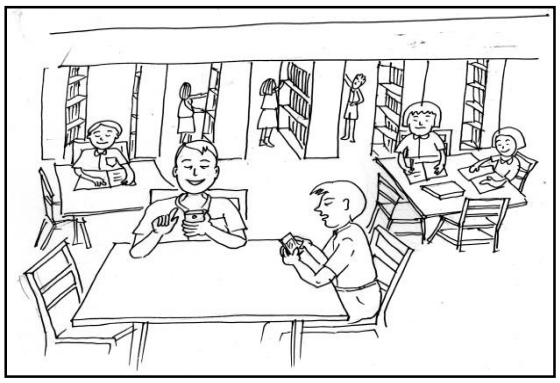
2. การที่ต้นพริกเจริญงอกงามขึ้น มีดอกและให้ผลเพิ่มขึ้นเกิดจากปัจจัยใดมากที่สุด
   1) มีการดูแลแต่งกิ่ง
   2) มีการบำรุงรักษาที่ดี
   3) ได้รับแสงที่พอเหมาะ
   4) ได้รับสารอาหารจากดินมากขึ้น

3. บุคคลใดมีความเสี่ยงต่อการเป็นโรคอ้วนน้อยที่สุด เพราะเหตุใด
   1) อ้อยชอบกินแครอท เพราะช่วยในการขับถ่าย
   2) อ้นชอบกินทุเรียน เพราะเป็นผลไม้ที่ให้พลังงานสูง
   3) อั้มชอบกินมะม่วงสุก เพราะความหวานของมะม่วงทำให้สดชื่น
   4) อ้อมชอบกินมันฝรั่งทอด เพราะมีแป้งมากพอกับที่ร่างกายต้องการ

อ่านข้อความแล้วตอบคำถามข้อ 4

ร้าน ก ขายอาหารจานด่วนประเภทไข่เจียว ปลาทอด และผัดผัก โดยใช้ผักที่มีใบสีเขียว ไม่เป็นรู ไม่มีหนอนกัดกิน ต่อมาพบว่ามีลูกค้าที่รับประทานอาหารร้านนี้เป็นประจำหลายคนป่วยเป็นโรคมะเร็ง

4. ข้อสรุปใดเป็นเหตุที่อาจจะทำให้ลูกค้าร้าน ก เป็นโรคมะเร็ง
   1) ใช้ปลาที่ไม่สดทำอาหาร
   2) รับประทานอาหารเดิมซ้ำ ๆ
   3) สารเคมีที่หลงเหลืออยู่ในผัก
   4) วิธีการปรุงอาหารไม่เป็นไปตามหลักการ



พิจารณาภาพแล้วตอบคำถามข้อ 5

5. ถ้านักเรียนเป็นเพื่อนกับแดงและดำ คำพูดใดควรใช้มากที่สุด
   1) น่าจะเอามาแบ่งให้เพื่อน ๆ เล่นบ้าง
   2) มีความสุขจริงนะ ขอเล่นบ้างได้ไหม
   3) เดี๋ยวฉันจะเก็บเงินเพื่อมาซื้อเล่นบ้าง
   4) ที่บ้านของฉันก็มี แต่ฉันจะเล่นในเวลาว่าง

การประเมินคุณภาพผู้เรียนระดับชาติ ปีการศึกษา 2556 สามารถแบ่งกลุ่มในคุณลักษณะของผู้เรียน คือ ภูมิภาค โดยแบ่งตาม กรมการปกครอง กระทรวงมหาดไทย ซึ่งแบ่งเป็น 4 ภาค ดังนี้

1. ภาคเหนือ ประกอบด้วย 17 จังหวัด กำแพงเพชร เชียงราย เชียงใหม่ ตาก นครสวรรค์ น่าน พะเยา พิจิตร พิษณุโลก เพชรบูรณ์ แพร่ แม่ฮ่องสอน ลำปาง ลำพูน สุโขทัย อุตรดิตถ์ อุทัยธานี

2. ภาคตะวันออกเฉียงเหนือ (อีสาน) ประกอบด้วย 19 จังหวัด กาฬสินธุ์ ขอนแก่น ชัยภูมิ นครพนม นครราชสีมา บุรีรัมย์ มหาสารคาม มุกดาหาร ยโสธร ร้อยเอ็ด เลย ศรีสะเกษ สกลนคร สุรินทร์ หนองคาย หนองบัวลาภู อานาจเจริญ อุดรธานี อุบลราชธานี

3. ภาคกลาง ประกอบด้วย 26 จังหวัด กรุงเทพฯ กาญจนบุรี ชัยนาท นครนายก นครปฐม นนทบุรี ปทุมธานี ประจวบคีรีขันธ์ พระนครศรีอยุธยา เพชรบุรี ราชบุรี ลพบุรี สมุทรปราการ สมุทรสงคราม สมุทรสาคร สระบุรี สิงห์บุรี สุพรรณบุรี อ่างทอง จันทบุรี ฉะเชิงเทรา ชลบุรี ตราด ปราจีนบุรี ระยอง สระแก้ว

4. ภาคใต้ ประกอบด้วย 14 จังหวัด กระบี่ ชุมพร ตรัง นครศรีธรรมราช นราธิวาส ปัตตานี พังงา พัทลุง ภูเก็ต ยะลา ระนอง สงขลา สตูล สุราษฎร์ธานี

สรุปได้ว่า การทดสอบระดับชาติ (NT) มุ่งเน้นประเมินที่ความสามารถของนักเรียน ว่ามีความรู้ เข้าใจ และสามารถนำไปใช้ในการดำเนินชีวิตได้มากน้อยเพียงใด ทั้งยังเป็นการเตรียมความพร้อมของตัวนักเรียนที่จะต้องรับการประเมินจากภายในและภายนอกโรงเรียน ที่เป็นส่วนหนึ่งที่จะพัฒนาให้การเรียนการสอนของประเทศมีคุณภาพและได้มาตราฐานตามหลักสากลในภายภาคหน้า

งานวิจัยที่เกี่ยวข้องกับการทดสอบระดับชาติ (NT) มีดังนี้

วรพรรณ ศรีกล่ำ (2559) ศึกษาปัจจัยพหุระดับนักเรียนและระดับห้องเรียนที่ส่งผลต่อคะแนนการทดสอบระดับชาติ ด้านภาษา โรงเรียนที่มีผลคะแนนการทดสอบระดับชาติ ต่ำในจังหวัดพิษณุโลก กลุ่มตัวอย่าง ได้แก่ นักเรียนจำนวน 1,260 คน และครูจำนวน 68 ห้องเรียน ข้อมูลที่เก็บรวบรวมข้อมูลโดยใช้แบบสอบถาม 2 ฉบับ แบบสอบถามระดับนักเรียนและระดับห้องเรียน 1) แบบสอบถามระดับนักเรียนประกอบด้วย 5 ตัวแปร แบบสอบถามวัดความรู้พื้นฐานเดิม แบบสอบถามวัดแรงจูงใจ ใฝ่สัมฤทธิ์ในการทำแบบทดสอบในการทดสอบระดับชาติ แบบสอบถามวัดเจตคติต่อการเรียนภาษาไทย แบบสอบถามวัดสภาพแวดล้อมทางบ้าน และแบบสอบถามวัดความเอาใจใส่ผู้ปกครอง ในการส่งเสริมการเรียน ซึ่งมีค่าความเชื่อมั่นเท่ากับ 0.92 แบบสอบถามระดับห้องเรียนประกอบด้วย 2 ตัวแปร แบบสอบถามวัดคุณภาพการสอนครูภาษาไทยและแบบสอบถามวัดบรรยากาศในชั้นเรียน มีค่าความเชื่อมั่นเท่ากับ 0.87 ทำการวิเคราะห์ข้อมูลโดยการวิเคราะห์พหุระดับ (Multilevel Analysis) ผลการวิเคราะห์ พบว่า โรงเรียนที่มีผลคะแนนการทดสอบระดับชาติต่ำ 1) ในระดับนักเรียน ตัวแปรความรู้พื้นฐานเดิมส่งผลต่อคะแนนการทดสอบระดับชาติ ด้านภาษาอย่างมีนัยสำคัญทางสถิติที่ .05 2) ในระดับห้องเรียนไม่มีตัวแปรอิสระใดที่ส่งผลต่อคะแนนการทดสอบระดับชาติ ด้านภาษา ตัวแปรความรู้พื้นฐานเดิมสามารถอธิบายความแปรปรวนของคะแนนการทดสอบระดับชาติ ด้านภาษา ได้ร้อยละ 12.85

เอกลักษณ์ คล้ายสุบรรณ สังวรณ์ งัดกระโทก และนลินี ณ นคร (2559) ศึกษาพัฒนาวิธีการวัดมูลค่าเพิ่มทางการศึกษาเพื่อใช้ประเมินคุณภาพ สถานศึกษาด้วยการวัดมูลค่าเพิ่มจากผลสัมฤทธิ์ทางการเรียนและผลการประเมินและรับรองคุณภาพของโรงเรียน โดยงานวิจัยนี้ได้ใช้คะแนนจาก

การทดสอบระดับชาติ (NT) ศึกษาความสอดคล้องของผลการประเมินคุณภาพสถานศึกษาด้วยการวัดมูลค่าเพิ่มที่พัฒนาขึ้นซึ่งมีการกำหนดน้ำหนักของการรวมคะแนนแตกต่างกัน และเปรียบเทียบมูลค่าเพิ่มทางการศึกษาของสถานศึกษาที่มีบริบทต่างกัน การวิจัยนี้ใช้ข้อมูลทุติยภูมิ กลุ่มตัวอย่างที่ใช้ในการวิจัย คือ โรงเรียนประถมศึกษาในสังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน จำนวน 96 โรงเรียน และมีนักเรียนชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2557 จำนวน 7,988 คน ซึ่งได้จากการสุ่มแบบหลายขั้นตอน ตัวแปรที่ศึกษาประกอบด้วย ตัวแปรระดับนักเรียน ได้แก่ เพศ การศึกษาของผู้ปกครอง ความสัมพันธ์ในครอบครัว ผลการประเมินคุณภาพการศึกษาระดับชาติขั้นพื้นฐาน ตัวแปรระดับสถานศึกษา ได้แก่ ผลการประเมินและรับรองคุณภาพของโรงเรียน ที่ตั้ง และขนาดของสถานศึกษา เครื่องมือที่ใช้ คือ แบบบันทึกข้อมูลพื้นฐานของโรงเรียน ผลการประเมินคุณภาพการศึกษาระดับชาติขั้นพื้นฐาน และ ผลการประเมินคุณภาพสถานศึกษา รอบ 3 ผู้วิจัยวิเคราะห์ข้อมูลคะแนนคุณภาพการศึกษาที่เป็นผลรวมของคะแนนผลสัมฤทธิ์ทางการเรียนกับผลการประเมินและรับรองคุณภาพของโรงเรียนโดยมีการถ่วงน้ำหนักต่างกันสามโมเดล คือ 40:60, 50:50, และ 60:40 สถิติที่ใช้ในการวิเคราะห์ข้อมูล คือ การวิเคราะห์พหุระดับ ผลการวิจัยพบว่า โมเดลการวัดมูลค่าเพิ่มทางการศึกษาที่มีความกลมกลืนกับข้อมูลมากที่สุด คือ โมเดลที่ 1 (40:60) รองลงมาคือโมเดลที่ 2 (50:50) และ โมเดลที่ 3 (60:40) (2) ความสอดคล้องของผลการประเมินคุณภาพสถานศึกษาด้วยการวัดมูลค่าเพิ่มระหว่างโมเดลที่ 1 (40:60) กับโมเดลที่ 2 (50:50) มีความสอดคล้องมากที่สุด มีความสอดคล้อง 91.67% รองลงมา ระหว่างโมเดลที่ 2 (50:50) กับโมเดลที่ 3 (60:40) มีความสอดคล้อง 88.54% และระหว่างโมเดลที่ 1 (40:60) กับโมเดลที่ 3 (60:40) มีความสอดคล้อ 80.21% ตามลำดับ และ(3) สถานศึกษาที่มีที่ตั้ง และขนาดต่างกันมีคะแนนมูลค่าเพิ่มทางการศึกษาไม่แตกต่างกัน

Kjellstrom and Pettersson (2005) ศึกษาระบบการทดสอบในประเทศสวีเดนโดยการทดสอบระดับชาติ ของสวีเดนเป็นการมองที่เป้าหมายและองค์ความรู้ ซึ่งระบบที่สำคัญของระบบการทดสอบนี้คือ ความรู้ภายในหลักสูตร การศึกษาครั้งนี้ยังบอกเกี่ยวกับภาพรวมของการเรียนการสอนในวิชาคณิตศาสตร์และอธิบายถึงอิทธิพลต่าง ๆ ที่มีผลต่อการทดสอบระดับชาติ การเปลี่ยนแปลงของการประเมินที่แตกต่างออกไปจากเดิม เพื่อไปสู่เป้าหมาย เกณฑ์ในการอ้างอิงระบบการให้คะแนนและการแก้ปัญหาของนักเรียน โดยรวมมุ่งเน้นไปที่กระบวนการเรียนการสอบของวิชาคณิตศาสตร์ พบว่า หลักสูตรและระบบการเรียนการสอนมีผลต่อคะแนนของการทดสอบระดับชาติ อย่างไรก็ตามทางสถานศึกษาจำเป็นต้องมีการพัฒนาหลักสูตรและการเรียนการสอนให้มีประสิทธิภาพมากยิ่งขึ้น

Brown, De Four-Babb, Bristol, and Conrad (2014) ศึกษาการทดสอบระดับชาติและข้อติชมการวินิจฉัย เกี่ยวกับคำพูดของครูในตรินิแดดและโตเบโก โดยมีการกล่าวถึงข้อเสนอแนะของการทดสอบ ในตรินิแดดและโตเบโก รวมถึงขอบเขตที่พวกเขาใช้ในรายงานการตัดสินใจทางหลักสูตรที่ส่งผลกระทบต่อการเรียนรู้ของนักเรียน โดยกลุ่มตัวอย่างประกอบด้วยครูประถมศึกษา จำนวน 133 คน แบ่งเป็น 79 คน จากโรงเรียนประสิทธิภาพต่ำและ 54 คน จากโรงเรียนประสิทธิภาพสูง และผู้บริหาร 10 คน ผลการวิจัยเชิงปริมาณและข้อมูลเชิงคุณภาพพบว่า มีครูจำนวนมากรู้สึกไม่สบายใจกับการตีความข้อมูลที่นำเสนอในรายงานเกี่ยวกับการทดสอบระดับชาติ โดยครูในโรงเรียนที่มีประสิทธิภาพสูงผ่านการทำงานร่วมกันของแผนกเพื่อใช้การศึกษาครั้งนี้ตัดสินใจเรื่องการสอนและหลักสูตร ความจำเป็นในการฝึกอบรมครูในการใช้และการตีความข้อมูลการประเมิน ปัญหาอื่น ๆ

ที่เกิดขึ้นจากข้อมูลและหัวข้อที่เป็นไปได้สำหรับการวิจัยเพิ่มเติมรวม การสร้างตราสินค้าของโรงเรียนเป็นโรงเรียนที่ดีและโรงเรียนไม่ดีตามเกี่ยวกับประสิทธิภาพของโรงเรียนในการทดสอบ

จากการศึกษางานวิจัยที่เกี่ยวข้องสามารถสรุปได้ว่าการทดสอบระดับชาติ มีความสำคัญต่อระบบการพัฒนาการศึกษาของประเทศเป็นอย่างมาก โดยมุ่งเน้นการทดสอบความสามารถของนักเรียนทั้งหมด 3 ด้าน คือ ด้านภาษา ด้านคำนวณ และด้านเหตุผล จำนวนด้านละ 30 ข้อ เพื่อทำการประเมินระดับความสามารถของนักเรียนระดับชั้นประถมศึกษาปีที่ 3 และยังพบว่าผู้เรียน ครูผู้สอน ระบบการเรียนการสอน รวมทั้งลักษณะของแบบทดสอบ ก็เป็นปัจจัยหนึ่งที่ส่งผลต่อผลการทดสอบระดับชาติ ที่ก่อให้เกิดความไม่เท่าเทียมกันในการทำข้อสอบ และส่งผลต่อคะแนนที่ได้จากการทำแบบทดสอบอีกด้วย

บทที่ 3
วิธีดำเนินการวิจัย

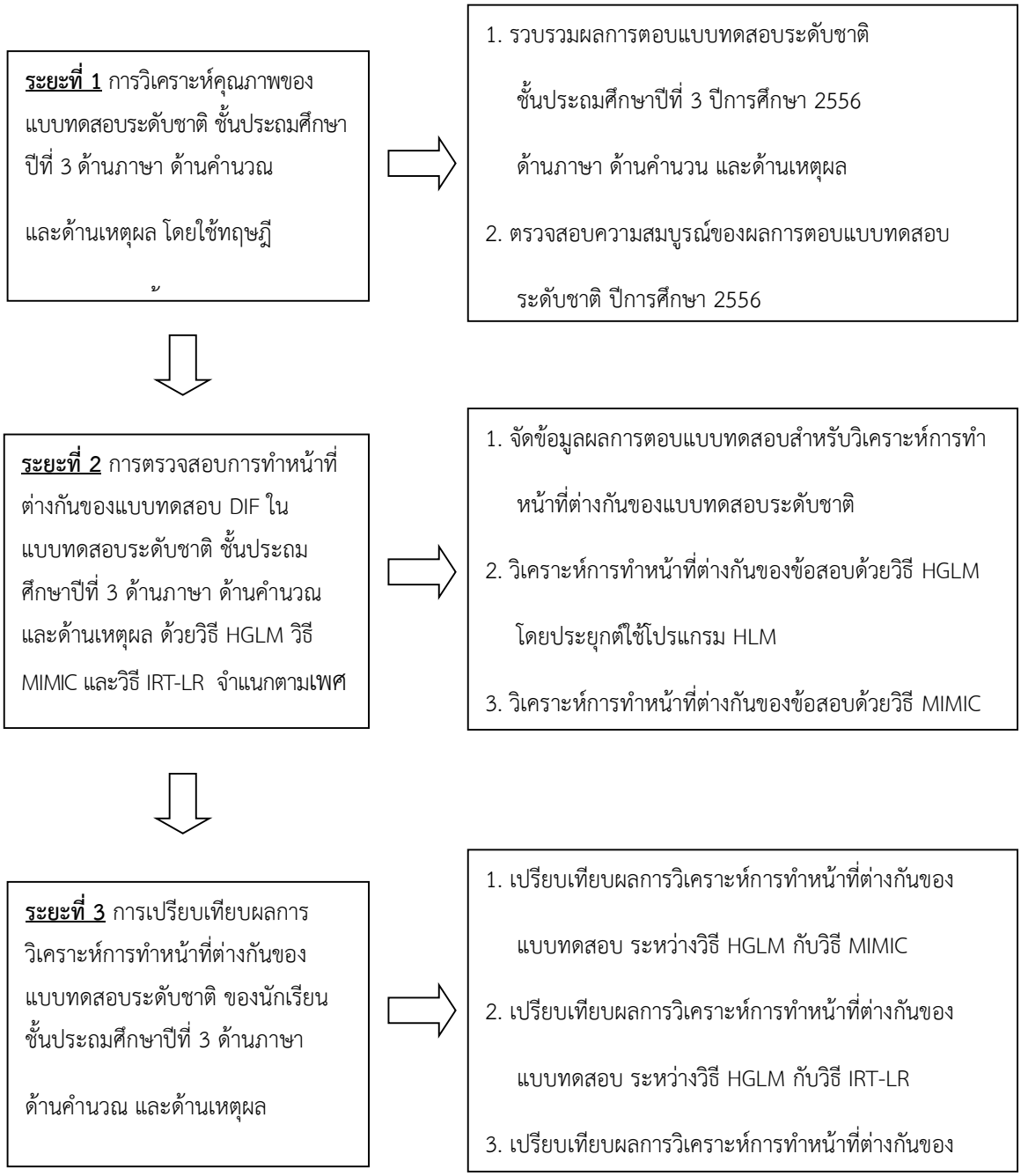
การวิจัยนี้มีวัตถุประสงค์เพื่อวิเคราะห์คุณภาพของแบบทดสอบระดับชาติ ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ และเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT โดยผู้วิจัยเสนอวิธีดำเนินการวิจัยเป็น 3 ระยะ ดังนี้

ระยะที่ 1 การวิเคราะห์คุณภาพของแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล โดยใช้ทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์

ระยะที่ 2 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR จำแนกตามเพศ

ระยะที่ 3 การเปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR จำแนกตามเพศ

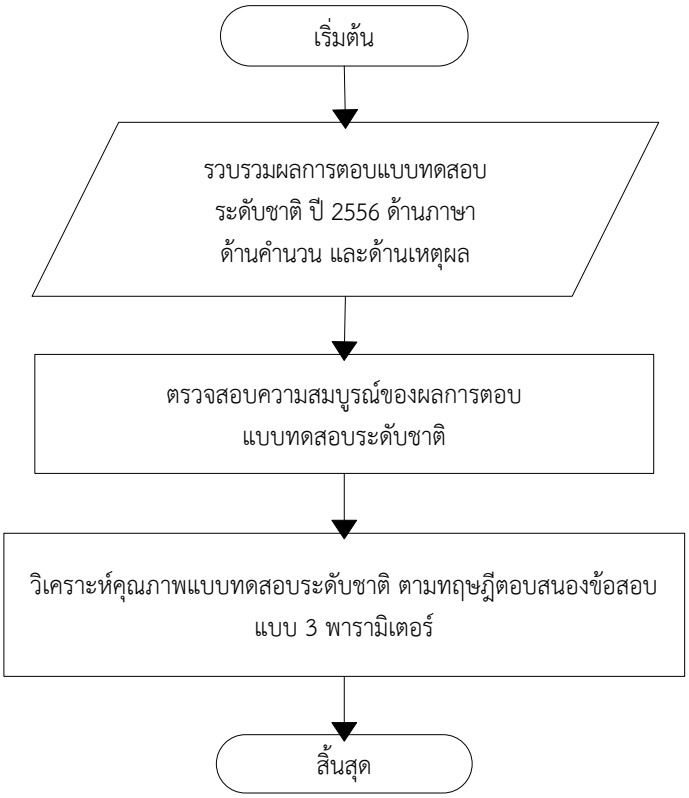
โดยมีวิธีดำเนินการวิจัยเป็น 3 ระยะ แสดงดังภาพที่ 3-1



ภาพที่ 3-1 ขั้นตอนการดำเนินงานวิจัย

ระยะที่ 1 การวิเคราะห์คุณภาพของแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล โดยใช้ทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์

การวิเคราะห์คุณภาพแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ตามความสามารถ จำนวน 3 ด้าน คือ 1) ด้านภาษา 2) ด้านคำนวณ และ 3) ด้านเหตุผล ดังภาพที่ 3-2



ภาพที่ 3-2 ขั้นตอนวิเคราะห์คุณภาพของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 โดยใช้ทฤษฎีการตอบสนองข้อสอบ ด้านภาษา ด้านคำนวณ และด้านเหตุผล แบบ 3 พารามิเตอร์

จากภาพที่ 3-2 แสดงขั้นตอนการวิเคราะห์คุณภาพของแบบทดสอบ ชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2556 ด้านภาษา ด้านคำนวน และด้านเหตุผล ดังนี้

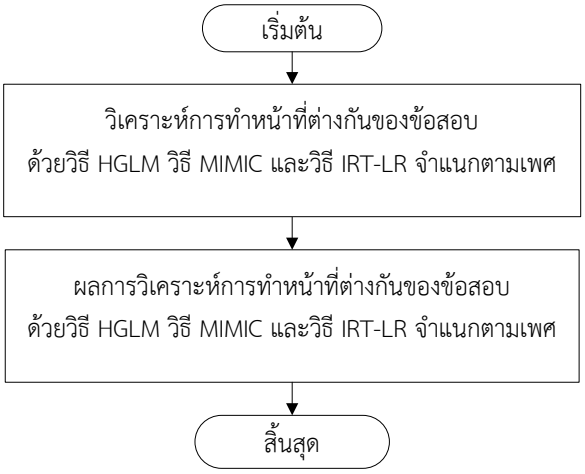
1. ผู้วิจัยทำการดำเนินการขอหนังสือขอความอนุเคราะห์ขอข้อมูลเพื่อการวิจัย เพื่อขอผลการตอบแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2556 จำนวน 3 ด้านประกอบด้วย 1) ด้านภาษา 2) ด้านคำนวณ และ 3) ด้านเหตุผล จากสำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน (สพฐ.)

2. ตรวจสอบความสมบูรณ์ของผลการตอบแบบทดสอบระดับชาติ ปีการศึกษา 2556 ทั้งข้อคำถาม ตัวเลือก และเฉลยคำตอบที่ถูกต้อง รวมทั้งตรวจสอบความสมบูรณ์ของคำตอบที่ผู้สอบทำการตอบ

3. วิเคราะห์คุณภาพของแบบทดสอบระดับชาติ ตามทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ โดยใช้โปรแกรมคอมพิวเตอร์สำเร็จรูป Xcalibre Version 4.1 ทั้งหมด 3 ด้าน คือ 1) ด้านภาษา 2) ด้านคำนวณ และ 3) ด้านเหตุผล

ระยะที่ 2 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR จำแนกตามเพศ

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ดังภาพที่ 3-3



ภาพที่ 3-3 ขั้นการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR จำแนกตามเพศ

จากภาพที่ 3-3 แสดงขั้นตอนการวิเคราะห์การทำหน้าที่ต่างกันของแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2556 ด้านภาษา ด้านคำนวน และด้านเหตุผล ดังนี้

1. การวิเคราะห์การทำหน้าที่ต่างกันของแบบทดสอบระดับชาติ ด้วยวิธี HGLM

เป็นการวิเคราะห์โดยโมเดลเชิงเส้นตรงทั่วไประดับลดหลั่น ด้วยการประยุกต์ใช้โปรแกรม HLM โดยมีตัวแปรเพศ และที่ตั้งโรงเรียนเป็นตัวแปรแฝง จากนั้นผู้วิจัยได้จัดเตรียมข้อมูลสำหรับการวิเคราะห์ ซึ่งผู้วิจัยกำหนดให้ตัวแปรมีค่าดังต่อไปนี้

ด้านภาษา ผู้วิจัยกำหนดให้ ตัวแปรเพศชายมีค่าเท่ากับ 0 และตัวแปรเพศหญิงมีค่าเท่ากับ 1 ส่วนตัวแปรที่ตั้งโรงเรียน ผู้วิจัยกำหนดให้โรงเรียนที่อยู่ในเมืองมีค่าเท่ากับ 1 และโรงเรียนที่อยู่นอกเมืองมีค่าเท่า 0

ด้านคำนวณ ผู้วิจัยกำหนดให้ ตัวแปรเพศชายมีค่าเท่ากับ 1 และตัวแปรเพศหญิงมีค่าเท่ากับ 0 ส่วนตัวแปรที่ตั้งโรงเรียน ผู้วิจัยกำหนดให้โรงเรียนที่อยู่ในเมืองมีค่าเท่ากับ 1 และโรงเรียนที่อยู่นอกเมืองมีค่าเท่า 0

ด้านเหตุผล ผู้วิจัยกำหนดให้ ตัวแปรเพศชายมีค่าเท่ากับ 0 และตัวแปรเพศหญิงมีค่าเท่ากับ 1 ส่วนตัวแปรที่ตั้งโรงเรียน ผู้วิจัยกำหนดให้โรงเรียนที่อยู่ในเมืองมีค่าเท่ากับ 1 และโรงเรียนที่อยู่นอกเมืองมีค่าเท่า 0

การวิเคราะห์แบ่งออกเป็น 2 ระดับ ด้วยโปรแกรมคอมพิวเตอร์สำเร็จรูป แล้วดำเนินตามขั้นตอนต่อไปนี้

ขั้นตอนที่ 1 เตรียมไฟล์ข้อมูลสำหรับการวิเคราะห์

ระดับที่ 1: ระดับข้อสอบ

ประกอบด้วยลำดับของผู้สอบ (ID) ลำดับของแบบทดสอบ (ITEM) ผลการตอบแบบทดสอบของผู้สอบ (Response) และตัวแปรดัมมี่ของแบบทดสอบ (ITEM,....,n) ดังภาพที่ 3-4

	ID	ITEM	Response	ITEM1	ITEM2	ITEM3	ITEM4	ITEM5
1	1	1	1	1	0	0	0	0
2	1	2	1	0	1	0	0	0
3	1	3	0	0	0	1	0	0
4	1	4	1	0	0	0	1	0
5	1	5	1	0	0	0	0	1
6	1	6	1	0	0	0	0	0
7	1	7	1	0	0	0	0	0
8	1	8	1	0	0	0	0	0
9	1	9	1	0	0	0	0	0
10	1	10	0	0	0	0	0	0

ภาพที่ 3-4 ตัวอย่างการจัดไฟล์ข้อมูลสำหรับวิเคราะห์ด้วยวิธี HGLM ระดับที่ 1: ระดับข้อสอบ

ระดับที่ 2: ระดับผู้สอบ

ประกอบด้วยลำดับของผู้สอบ (ID) ตัวแปรเพศ (Gender) และตัวแปรที่ตั้งโรงเรียน (Area) โดยผู้วิจัยได้กำหนดดังภาพที่ 3-5

	ID	Gender	Area
1	1	1	1
2	2	1	1
3	3	1	1
4	4	1	1
5	5	1	1
6	6	1	1
7	7	1	1
8	9	1	1
9	10	1	1

ภาพที่ 3-5 ตัวอย่างการจัดไฟล์ข้อมูลสำหรับวิเคราะห์ด้วยวิธี HGLM ระดับที่ 2: ระดับผู้สอบ

ขั้นตอนที่ 2 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF)

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ด้วยการประยุกต์ใช้โมเดลเชิงเส้นตรงทั่วไปแบบลดหลั่น (HGLM) ที่มีผลการตอบแบบ 2 ค่า (Dichotomous) ผู้วิจัยได้วิเคราะห์ตามขั้นตอนดังนี้

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยการศึกษาอิทธิพลของตัวแปรระดับข้อสอบและระดับผู้สอบ ที่มีโอกาสในการตอบข้อสอบถูก ในการวิเคราะห์ด้วยโปรแกรม HLM โดยสามารถแบ่งการวิเคราะห์เป็น 2 ระดับ ดังนี้

โมเดล HGLM วิเคราะห์ระดับที่ 1: ระดับข้อสอบ

การวิเคราะห์ระดับข้อสอบ ใช้หลักการวิเคราะห์ที่ข้อสอบสอดแทรกอยู่ในตัวบุคคล ผลการวิเคราะห์ระดับนี้จะแสดงค่าความยากของข้อสอบ ซึ่งสามารถเขียนสมการการวิเคราะห์ ดังนี้

สมการโมเดลการวิเคราะห์ระดับที่ 1 ระดับข้อสอบ

$$\eta_{ij} = \beta_{0j} + \beta_{1j}x_{1j} + \beta_{2j}x_{2j} + ... + \beta_{29j}x_{29j} \quad (43)$$

โมเดล HGLM วิเคราะห์ระดับที่ 2: ระดับผู้สอบ

การวิเคราะห์ระดับผู้สอบ ใช้หลักการวิเคราะห์ที่ผู้สอบแต่ละคนสอดแทรกในแต่ละที่ตั้งโรงเรียน ผลการวิเคราะห์ได้ค่าพารามิเตอร์ข้อสอบ และค่าความสามารถของผู้สอบในสมการระดับผู้สอบ โดยสามารถเขียนตัวแปรคุณลักษณะของผู้สอบเข้าสู่สมการ เพื่ออธิบายความผันแปรของโอกาสในการตอบแบบทดสอบได้ถูกต้องของผู้สอบในแต่ละที่ตั้งโรงเรียน สามารถเขียนสมการได้ ดังนี้

สมการโมเดลการวิเคราะห์ระดับที่ 2 ระดับผู้สอบ จำแนกตามเพศ ดังนี้

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Gender + u_{0j} \quad (44)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}Gender \quad (45)$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}Gender \quad (46)$$

.

.

.

$$\beta_{29j} = \gamma_{290} + \gamma_{291}Gender \quad (47)$$

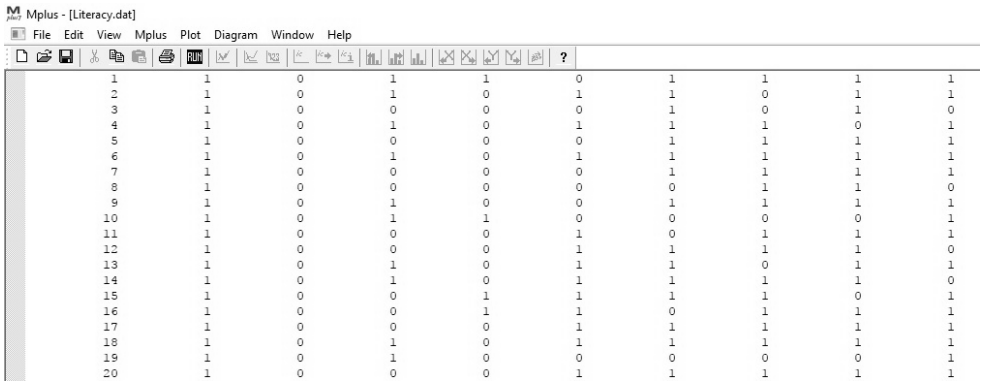
หลังจากทำการวิเคราะห์ข้อมูล จึงพิจารณาผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี HGLM โดยดูจากค่า *p*-value ของแต่ละข้อ ว่าข้อใดมีค่า *p*-value ที่นัยสำคัญทางสถิติที่ระดับ .05 แสดงว่า ข้อสอบข้อนั้นเกิดการทำหน้าที่ต่างของข้อสอบ

2. การวิเคราะห์การทำหน้าที่ต่างกันของแบบทดสอบระดับชาติ ด้วยวิธี MIMIC

การวิเคราะห์ด้วยโมเดลลิสเรลของคุณลักษณะแฝงที่มีหลายสาเหตุและวัดได้จากตัวบ่งชี้หลายตัว ผู้วิจัยได้จำแนกตามเพศ และที่ตั้งโรงเรียนตามภูมิศาสตร์ โดยมีขั้นตอนการดำเนินการ ดังนี้

ขั้นตอนที่ 1 เตรียมไฟล์ข้อมูลสำหรับการวิเคราะห์

การวิเคราะห์โมเดลลิสเรล ผู้วิจัยได้เตรียมข้อมูลให้อยู่ในรูปแบบไฟล์ .dat เพื่อวิเคราะห์การทำหน้าที่ต่างกันของแบบทดสอบโปรแกรม Mplus ประกอบด้วยลำดับผู้สอบ (ID) ที่ตั้งโรงเรียน (Area) เพศ (Gender) และผลการตอบแบบทดสอบของผู้สอบ (Response) ดังภาพที่ 3-6


Mplus - [Literacy.dat]

File Edit View Mplus Plot Diagram Window Help

1	1	0	1	1	0	1	1	1	1
2	1	0	1	0	1	1	0	1	1
3	1	0	0	0	0	1	0	1	0
4	1	0	1	0	1	1	1	0	1
5	1	0	0	0	0	1	1	1	1
6	1	0	1	0	1	1	1	1	1
7	1	0	0	0	0	1	1	1	1
8	1	0	0	0	0	0	1	1	0
9	1	0	1	0	0	1	1	1	1
10	1	0	1	1	0	0	0	0	1
11	1	0	0	0	1	0	1	1	1
12	1	0	0	0	1	1	1	1	0
13	1	0	1	0	1	1	0	1	1
14	1	0	1	0	1	1	1	1	0
15	1	0	0	1	1	1	1	0	1
16	1	0	0	1	1	0	1	1	1
17	1	0	0	0	1	1	1	1	1
18	1	0	1	0	1	1	1	1	1
19	1	0	1	0	0	0	0	0	1
20	1	0	0	0	1	1	1	1	1

ภาพที่ 3-6 ตัวอย่างการจัดไฟล์ข้อมูลสำหรับวิเคราะห์ด้วยวิธี MIMIC ในรูปแบบไฟล์ .dat

ขั้นตอนที่ 2 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี MIMIC สามารถดำเนินการตามขั้นตอน ดังนี้

1. การวิเคราะห์โมเดล CFA โดยปราศจากตัวแปรทำนาย (X)

2. การวิเคราะห์โมเดล CFA และเพิ่มตัวแปรทำนาย (X) โดยไม่มีอิทธิพลทางตรง (Direct Effect) ต่อข้อสอบ

3.การเพิ่มอิทธิพลทางตรงต่อข้อสอบและบังคับให้มีค่าเป็น 0 เพื่อกำหนดให้ ตัวแปรต้นไม่มีผลทางตรงต่อข้อสอบ (Y1 on X@0)

4. ตรวจดูค่าดัชนีปรับแก้ (Modification Indices) ว่าข้อใดมีค่าดัชนีปรับแก้สูงที่สุด

5. เพิ่มอิทธิพลทางตรงจากตัวแปรทำนายไปที่ข้อสอบ ในข้อสอบที่มีดัชนีปรับแก้สูงที่สุด แล้วทำการวิเคราะห์โมเดลที่ปรับแก้อีกครั้ง

6. ดำเนินการซ้ำ ในข้อที่ 4-5 จนไม่พบดัชนีปรับแก้ที่มีนัยสำคัญทางสถิติ

7. ประเมินความกลมกลืนของโมเดล และตรวจผลทางตรงที่มีนัยสำคัญทางสถิติ ถ้าหากพบข้อสอบที่มีผลทางตรง และมีนัยสำคัญทางสถิติที่ระดับ .05 แสดงว่าข้อสอบข้อนั้น คือข้อที่เกิดการทำหน้าที่ต่างกันของข้อสอบ

3. การวิเคราะห์การทำหน้าที่ต่างกันของแบบทดสอบระดับชาติ ด้วยวิธี IRT-LR

การวิเคราะห์ด้วยสถิติทางคณิตศาสตร์ โดยการเปรียบเทียบความแตกต่างของค่าพารามิเตอร์ใน กลุ่มอ้างอิงและกลุ่มเปรียบเทียบ มีวิธีดำเนินการ ดังนี้

ขั้นตอนที่ 1 เตรียมไฟล์ข้อมูลสำหรับวิเคราะห์

การวิเคราะห์ข้อมูลด้วยวิธี IRT-LR ผู้วิจัยได้เตรียมข้อมูลในรูปแบบของไฟล์ .sav เพื่อวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยโปรแกรม IRTPRO ประกอบด้วยลำดับผู้สอบ (ID) เพศ (Gender) ที่ตั้งโรงเรียน (Area) และผลการตอบแบบทดสอบของผู้สอบ (Response) ดังภาพที่ 3-7

	ID	Gender	Area	Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8	Item9	Item10	Item11	Item12	Item13	Item14	Item15
1	1	1	1	1	0	0	0	0	0	0	0	1	0	1	0	1	0	0
2	2	1	1	0	0	0	0	1	1	0	0	1	0	1	0	1	0	1
3	3	1	1	0	0	1	1	0	1	1	1	1	0	0	1	1	0	1
4	4	1	1	0	0	1	1	0	1	0	0	0	0	1	0	0	0	0
5	5	1	1	0	0	0	0	0	0	0	1	1	0	1	1	0	1	0
6	6	1	1	1	0	1	1	0	1	1	1	1	0	0	1	1	0	0
7	7	1	1	1	0	1	1	1	0	1	1	1	1	1	1	0	1	0
8	8	1	1	1	0	0	0	0	1	0	0	1	0	1	0	0	0	0
9	9	1	1	1	0	0	1	1	1	1	1	1	1	1	0	1	0	1
10	10	1	1	0	0	0	1	1	1	1	1	1	0	0	0	1	0	0
11	11	1	1	0	0	1	1	1	0	0	0	1	1	1	0	0	1	0
12	12	1	1	0	1	0	0	1	1	1	1	0	0	0	0	0	0	1
13	13	1	1	0	0	1	1	1	0	0	0	1	0	0	1	1	0	1
14	14	1	1	0	0	1	0	1	1	1	0	0	1	1	1	0	0	0
15	15	1	1	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0

ภาพที่ 3-7 ตัวอย่างการจัดไฟล์ข้อมูลสำหรับวิเคราะห์ด้วยวิธี IRT-LR ในรูปแบบไฟล์ .sav

ขั้นตอนที่ 2 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ด้วยวิธี IRT-LR จะประเมินความสำคัญระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ในความแตกต่างของแต่ละพารามิเตอร์ สามารถเขียนสมการการวิเคราะห์ ดังนี้

สมการโมเดลเดลการวิเคราะห์ ดังนี้

$$G^2(d.f.) = -2\log L_C - (-2\log L_A) \tag{52}$$

1. เลือกวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยข้อสอบตามทฤษฎีการตอบสนองข้อสอบแบบเอกมิติ (Unidimensional IRT)

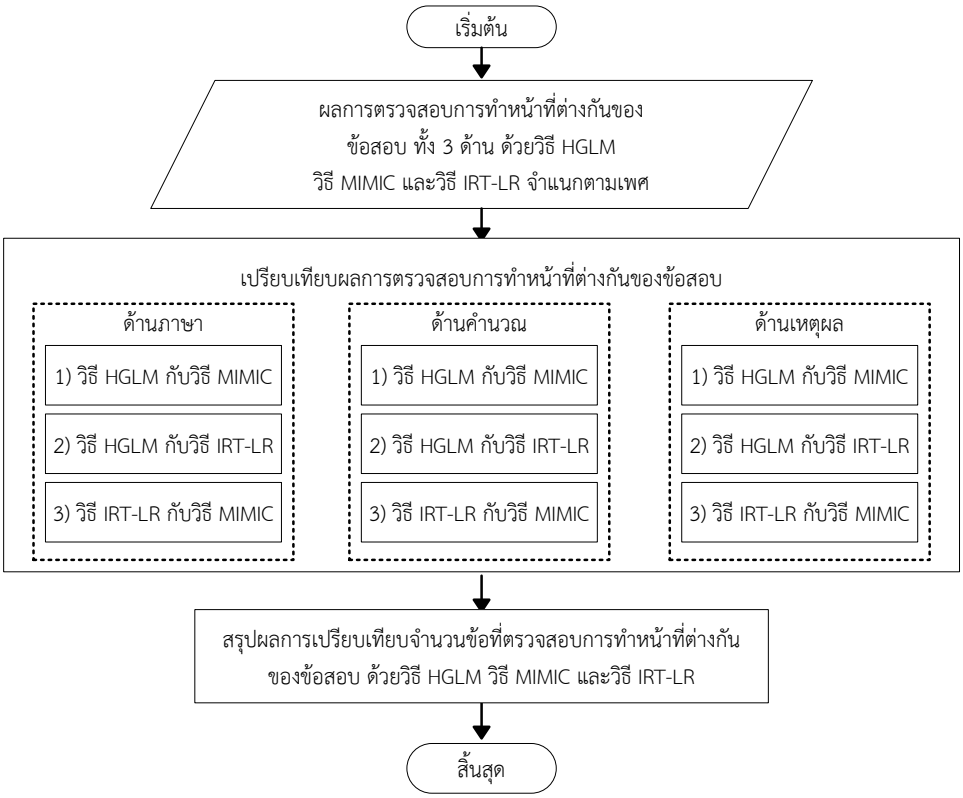
2. กำหนดตัวแปรเพศ ในความสามารถด้านภาษา และด้านเหตุผล ให้เพศหญิงเป็นกลุ่มอ้างอิง (Referent Group: R) และเพศชายเป็นกลุ่มเปรียบเทียบ (Focal Group: F) ส่วนความสามารถด้านคำนวณ ให้เพศชายเป็นกลุ่มอ้างอิง (R) และเพศหญิงเป็นกลุ่มเปรียบเทียบ (F)

3. เลือกโมเดลสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เพื่อวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ โดยกำหนดให้ทำการวิเคราะห์ข้อสอบทุกข้อเป็นข้อของแบบทดสอบ จากนั้นพิจารณาว่าข้อสอบข้อใดมี DIF โดยดูจากค่า p-value ที่มีนัยสำคัญทางสถิติที่ระดับ .05

จากการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ ผู้วิจัยได้ดำเนินการรวบรวมผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ของวิธี HGLM วิธี MIMIC และวิธี IRT-LR และนำไปสู่ขั้นตอนการเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ต่อไป

ระยะที่ 3 การเปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ของนักเรียนระดับชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR จำแนกตามเพศ

การเปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ มีขั้นตอนดำเนินการ ดังนี้



ภาพที่ 3-8 ขั้นตอนการเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

ผู้วิจัยทำการเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ จำแนกตามเพศ ตามสมมติฐานการวิจัย ดังนี้

1. เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ระหว่างวิธี HGLM กับ วิธี MIMIC
2. เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ระหว่างวิธี HGLM กับ วิธี IRT-LR
3. เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ระหว่างวิธี IRT-LR กับ วิธี MIMIC
4. เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านคำนวณ ระหว่างวิธี HGLM กับ วิธี MIMIC

5. เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านคำนวณ ระหว่างวิธี HGLM กับ วิธี IRT-LR

6. เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านคำนวณ ระหว่างวิธี IRT-LR กับ วิธี MIMIC

7. เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านเหตุผล ระหว่างวิธี HGLM กับ วิธี MIMIC

8. เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านเหตุผล ระหว่างวิธี HGLM กับ วิธี IRT-LR

9. เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านเหตุผล ระหว่างวิธี IRT-LR กับ วิธี MIMIC

บทที่ 4
ผลการวิจัย

การวิจัยนี้มีวัตถุประสงค์เพื่อวิเคราะห์คุณภาพของแบบทดสอบระดับชาติ ตรวจสอบการทำหน้าที่ต่างกันของแบบทดสอบระดับชาติ และเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR โดยผู้วิจัยนำเสนอผลการวิจัยเป็น 3 ตอน ดังนี้

ตอนที่ 1 ผลการวิเคราะห์คุณภาพของแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล โดยใช้ทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์

ตอนที่ 2 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR จำแนกตามเพศ

ตอนที่ 3 ผลการเปรียบเทียบการวิเคราะห์การทำหน้าที่ต่างกันของแบบทดสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR จำแนกตามเพศ

ผู้วิจัยได้กำหนดสัญลักษณ์ที่ใช้ในการวิจัย ดังนี้

M	หมายถึง	ค่าเฉลี่ย
SD	หมายถึง	ส่วนเบี่ยงเบนมาตรฐาน
n	หมายถึง	จำนวนกลุ่มตัวอย่าง
a	หมายถึง	ค่าอำนาจจำแนกของข้อสอบ
b	หมายถึง	ค่าความยากของข้อสอบ
c	หมายถึง	ค่าโอกาสการเดาข้อสอบ

ตอนที่ 1 ผลการวิเคราะห์คุณภาพของแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 โดยใช้ทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์

การวิเคราะห์คุณภาพของแบบทดสอบระดับชาติ เป็นการวิเคราะห์ค่าพารามิเตอร์ของข้อสอบในทฤษฎีการตอบสนองข้อสอบ ประกอบด้วย ค่าอำนาจจำแนกของข้อสอบ (a) ค่าความยากของข้อสอบ (b) และค่าโอกาสในการเดาของข้อสอบ (c) โดยใช้โปรมแกรมคอมพิวเตอร์สำเร็จรูป Xcaliber Version 4.1 ในการประมาณค่าพารามิเตอร์ของข้อสอบ (Urry, 1977) ดังนี้

1. ค่าอำนาจจำแนกของข้อสอบ (a) มีค่าระหว่าง 0.50 ถึง 2.50
2. ค่าความยากของข้อสอบ (b) มีค่าระหว่าง -2.50 ถึง 2.50
3. ค่าโอกาสในการเดาของข้อสอบ (c) ไม่เกิน 0.30

ตารางที่ 4-1 ผลการวิเคราะห์คุณภาพของแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา จำนวน 30 ข้อ โดยใช้ทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์

แบบทดสอบ NT	ข้อที่	ค่าพารามิเตอร์		
		a	b	c
ด้านภาษา	1	0.427	0.749	0.197
	2	0.742	2.306	0.215
	3	0.887	-0.363	0.158
	4	0.685	-0.957	0.183
	5	0.579	0.582	0.223
	6	0.585	-0.844	0.219
	7	1.144	-0.298	0.213
	8	0.635	0.856	0.206
	9	0.587	-1.404	0.210
	10	0.548	1.357	0.259
	11	0.696	-0.341	0.260
	12	1.146	0.056	0.209
	13	0.384	1.441	0.190
	14	1.330	0.497	0.238
	15	0.450	1.589	0.193
	16	0.690	0.646	0.171
	17	1.240	-0.257	0.195
	18	0.986	0.998	0.235

ตารางที่ 4-1 (ต่อ)

แบบทดสอบ NT	ข้อที่	ค่าพารามิเตอร์		
		a	b	c
ด้านภาษา	19	1.241	-0.665	0.197
	20	0.901	-0.208	0.189
	21	0.549	0.801	0.215
	22	1.071	1.720	0.174
	23	0.552	1.886	0.201
	24	0.580	0.576	0.195
	25	0.558	1.804	0.239
	26	0.228	1.886	0.207
	27	1.160	1.148	0.269
	28	0.847	0.321	0.250
	29	0.523	1.006	0.177
	30	0.645	0.215	0.209
ค่าความเที่ยงทั้งฉบับ (Reliability)			0.746	

จากตารางที่ 4-1 แบบทดสอบด้านภาษา จำนวน 30 ข้อ มีค่าความเที่ยงทั้งฉบับ (Reliability) เท่ากับ 0.746 ข้อสอบมีค่าอำนาจจำแนกของข้อสอบ (a) ตั้งแต่ 0.228 ถึง 1.330 ค่าความยากของข้อสอบ (b) ตั้งแต่ -1.404 ถึง 2.306 และค่าโอกาสการเดาของข้อสอบ (c) ตั้งแต่ 0.158 ถึง 0.269

ตารางที่ 4-2 ผลการวิเคราะห์คุณภาพของแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านคำนวณ จำนวน 30 ข้อ โดยใช้ทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์

แบบทดสอบ NT	ข้อที่	ค่าพารามิเตอร์		
		a	b	c
ด้านคำนวณ	1	1.122	2.268	0.095
	2	1.138	0.963	0.119
	3	1.055	1.654	0.282
	4	1.156	0.572	0.216
	5	1.024	0.816	0.174
	6	1.170	1.672	0.210
	7	2.089	1.630	0.194
	8	1.117	-0.147	0.199

ตารางที่ 4-2 (ต่อ)

แบบทดสอบ NT	ข้อที่	ค่าพารามิเตอร์		
		a	b	c
ด้านคำนวณ	9	1.096	-0.028	0.182
	10	1.255	0.943	0.271
	11	0.974	0.871	0.253
	12	0.601	2.418	0.260
	13	0.852	0.282	0.181
	14	0.711	1.459	0.207
	15	1.042	1.180	0.265
	16	0.845	0.620	0.176
	17	0.991	-0.043	0.231
	18	0.883	0.864	0.217
	19	0.890	0.738	0.162
	20	0.408	0.921	0.230
	21	0.647	2.737	0.276
	22	0.934	2.236	0.174
	23	0.965	0.954	0.232
	24	0.832	0.736	0.197
	25	1.056	0.993	0.213
	26	0.991	0.362	0.210
	27	1.268	1.443	0.184
	28	1.216	1.417	0.241
	29	1.146	1.033	0.165
	30	1.122	2.268	0.095
ค่าความเที่ยงทั้งฉบับ (Reliability)			0.764	

จากตารางที่ 4-2 แบบทดสอบด้านคำนวณ จำนวน 30 ข้อ มีค่าความเที่ยงทั้งฉบับ (Reliability) เท่ากับ 0.764 ข้อสอบมีค่าอำนาจจำแนกของข้อสอบ (a) ตั้งแต่ 0.408 ถึง 2.089 ค่าความยากของข้อสอบ (b) ตั้งแต่ -0.147 ถึง 2.737 และค่าโอกาสการเดาของข้อสอบ (c) ตั้งแต่ 0.095 ถึง 0.282

ตารางที่ 4-3 ผลการวิเคราะห์คุณภาพของแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านเหตุผล จำนวน 30 ข้อ โดยใช้ทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์

แบบทดสอบ NT	ข้อที่	ค่าพารามิเตอร์		
		a	b	c
ด้านเหตุผล	1	0.543	0.321	0.165
	2	0.833	0.180	0.197
	3	0.704	0.126	0.157
	4	0.884	-0.406	0.158
	5	0.657	-0.444	0.166
	6	0.477	-0.253	0.176
	7	0.944	0.034	0.186
	8	0.628	0.516	0.213
	9	0.502	0.480	0.189
	10	0.753	0.782	0.157
	11	0.807	0.080	0.154
	12	0.685	0.312	0.159
	13	0.578	0.688	0.177
	14	0.935	0.514	0.190
	15	0.627	0.948	0.172
	16	1.219	1.377	0.260
	17	0.773	2.333	0.203
	18	0.922	0.366	0.154
	19	0.752	0.454	0.160
	20	0.661	3.292	0.241
	21	0.852	0.651	0.212
	22	1.097	0.324	0.239
	23	1.004	0.348	0.185
	24	1.360	0.238	0.225
	25	0.394	2.359	0.213
	26	0.882	0.406	0.161
	27	0.745	1.165	0.160

ตารางที่ 4-3 (ต่อ)

แบบทดสอบ NT	ข้อที่	ค่าพารามิเตอร์		
		a	b	c
	28	0.781	1.472	0.184
	29	0.834	2.643	0.220
	30	0.764	1.335	0.176
ค่าความเที่ยงทั้งฉบับ (Reliability)			0.774	

จากตารางที่ 4-3 แบบทดสอบด้านคำนวณ จำนวน 30 ข้อ มีค่าความเที่ยงทั้งฉบับ (Reliability) เท่ากับ 0.774 ข้อสอบมีค่าอำนาจจำแนกของข้อสอบ (a) ตั้งแต่ 0.394 ถึง 1.360 ค่าความยากของข้อสอบ (b) ตั้งแต่ -0.444 ถึง 3.292 และค่าโอกาสการเดาของข้อสอบ (c) ตั้งแต่ 0.154 ถึง 0.260

จากการวิเคราะห์คุณภาพของแบบทดสอบระดับชาติ โดยการวิเคราะห์ค่าพารามิเตอร์ตามทฤษฎีการตอบสนองข้อสอบ (IRT) พบว่า

แบบทดสอบระดับชาติ ด้านภาษา มีค่าอำนาจจำแนกของข้อสอบ (a) มีค่าเฉลี่ยเท่ากับ 0.753 ค่าความยากของข้อสอบ (b) มีค่าเฉลี่ยเท่ากับ 0.570 และค่าโอกาสการเดาของข้อสอบ (c) มีค่าเฉลี่ยเท่ากับ 0.210 สามารถสรุปได้ว่า แบบทดสอบด้านภาษา มีอำนาจจำแนกข้อสอบอยู่ในระดับค่อนข้างดี มีค่าความยากของข้อสอบอยู่ในระดับค่อนข้างยาก และมีค่าโอกาสการเดาของข้อสอบไม่เกิน 0.3

แบบทดสอบระดับชาติ ด้านคำนวณ มีค่าอำนาจจำแนกของข้อสอบ (a) มีค่าเฉลี่ยเท่ากับ 1.020 ค่าความยากของข้อสอบ (b) มีค่าเฉลี่ยเท่ากับ 1.128 และมีค่าโอกาสการเดาของข้อสอบ (C) มีค่าเฉลี่ยเท่ากับ 0.204 สามารถสรุปได้ว่า แบบทดสอบด้านคำนวณ มีอำนาจจำแนกของข้อสอบอยู่ในระดับค่อนข้างดี มีค่าความยากของข้อสอบอยู่ในระดับยาก และมีค่าโอกาสการเดาของข้อสอบไม่เกิน 0.3

แบบทดสอบระดับชาติ ด้านเหตุผล มีค่าอำนาจจำแนกของข้อสอบ (a) มีค่าเฉลี่ยเท่ากับ 0.787 ค่าความยากของข้อสอบ (b) มีค่าเฉลี่ยเท่ากับ 0.755 และมีค่าโอกาสการเดาของข้อสอบ (c) มีค่าเฉลี่ยเท่ากับ 0.187 สามารถสรุปได้ว่า แบบทดสอบด้านเหตุผล มีอำนาจจำแนกของข้อสอบอยู่ในระดับค่อนข้างดี มีค่าความยากของข้อสอบอยู่ในระดับค่อนข้างยาก และมีค่าโอกาสการเดาของข้อสอบไม่เกิน 0.3

สามารถสรุปได้ว่า แบบทดสอบระดับชาติ ทั้ง 3 ด้าน มีค่าอำนาจจำแนกของข้อสอบ (a) อยู่ในระดับที่สามารถจำแนกผู้สอบได้ดี มีค่าความยากของข้อสอบ (b) อยู่ในระดับค่อนข้างยาก และมีค่าโอกาสการเดาของข้อสอบ (c) ไม่เกิน 0.3

ตอนที่ 2 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR จำแนกตามเพศ

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้าน คือ ด้านภาษา ด้านคำนวณ และด้านเหตุผล มีจำนวนข้อสอบด้านละ 30 ข้อ รวมทั้งหมด 90 ข้อ ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR

ตารางที่ 4-4 ผลการตรวจสอบการทำหน้าที่ต่างกันของแบบทดสอบระดับชาติ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM จำแนกตามเพศ

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี HGLM		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
1	NO-DIF	NO-DIF	NO-DIF
2	**DIF**	NO-DIF	**DIF**
3	**DIF**	**DIF**	**DIF**
4	**DIF**	NO-DIF	**DIF**
5	NO-DIF	NO-DIF	**DIF**
6	**DIF**	NO-DIF	**DIF**
7	**DIF**	**DIF**	**DIF**
8	**DIF**	**DIF**	**DIF**
9	**DIF**	**DIF**	**DIF**
10	**DIF**	NO-DIF	**DIF**
11	**DIF**	NO-DIF	**DIF**
12	**DIF**	**DIF**	**DIF**
13	**DIF**	**DIF**	**DIF**
14	**DIF**	NO-DIF	**DIF**
15	**DIF**	NO-DIF	NO-DIF
16	**DIF**	**DIF**	NO-DIF
17	**DIF**	**DIF**	NO-DIF
18	**DIF**	NO-DIF	**DIF**
19	**DIF**	**DIF**	**DIF**
20	**DIF**	**DIF**	**DIF**
21	**DIF**	NO-DIF	NO-DIF
22	**DIF**	NO-DIF	NO-DIF
23	**DIF**	NO-DIF	**DIF**

ตารางที่ 4-4 (ต่อ)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี HGLM		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
24	**DIF**	NO-DIF	**DIF**
25	**DIF**	**DIF**	NO-DIF
26	**DIF**	NO-DIF	**DIF**
27	**DIF**	**DIF**	NO-DIF
28	**DIF**	NO-DIF	**DIF**
29	**DIF**	NO-DIF	NO-DIF
30	NO-DIF	NO-DIF	**DIF**
จำนวนข้อที่พบ DIF	(27 ข้อ) 90%	(13 ข้อ) 43%	(22 ข้อ) 73%

หมายเหตุ DIF หมายถึง ข้อสอบที่พบว่าทำหน้าที่ต่างกันของข้อสอบ
NO-DIF หมายถึง ข้อสอบที่ไม่พบว่าทำหน้าที่ต่างกันของข้อสอบ

จากตารางที่ 4-4 แสดงผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM จำแนกตามเพศ พบว่า

ด้านภาษา ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 27 ข้อ ได้แก่ ข้อที่ 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28 และ 29 คิดเป็นร้อยละ 90 ของข้อสอบทั้งหมด

ด้านคำนวณ ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 13 ข้อ ได้แก่ ข้อที่ 3, 7, 8, 9, 12, 13, 16, 17, 19, 20, 23, 25 และ 27 คิดเป็นร้อยละ 43 ของข้อสอบทั้งหมด

ด้านเหตุผล ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 22 ข้อ ได้แก่ ข้อที่ 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 18, 19, 20, 22, 23, 24, 26, 28 และ 30 คิดเป็นร้อยละ 73 ของข้อสอบทั้งหมด

ตารางที่ 4-5 ผลการตรวจสอบการทำหน้าที่ต่างกันของแบบทดสอบระดับชาติ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี MIMIC จำแนกตามเพศ

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี MIMIC		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
1	**DIF**	**DIF**	**DIF**
2	NO-DIF	NO-DIF	NO-DIF
3	NO-DIF	NO-DIF	NO-DIF
4	NO-DIF	NO-DIF	NO-DIF
5	**DIF**	NO-DIF	NO-DIF
6	NO-DIF	NO-DIF	NO-DIF

ตารางที่ 4-5 (ต่อ)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี MIMIC		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
7	NO-DIF	NO-DIF	NO-DIF
8	NO-DIF	NO-DIF	NO-DIF
9	NO-DIF	NO-DIF	NO-DIF
10	NO-DIF	NO-DIF	NO-DIF
11	NO-DIF	NO-DIF	NO-DIF
12	**DIF**	NO-DIF	**DIF**
13	NO-DIF	NO-DIF	NO-DIF
14	NO-DIF	NO-DIF	NO-DIF
15	NO-DIF	NO-DIF	**DIF**
16	NO-DIF	NO-DIF	NO-DIF
17	NO-DIF	NO-DIF	**DIF**
18	NO-DIF	NO-DIF	NO-DIF
19	NO-DIF	NO-DIF	NO-DIF
20	NO-DIF	NO-DIF	NO-DIF
21	NO-DIF	NO-DIF	NO-DIF
22	NO-DIF	NO-DIF	NO-DIF
23	NO-DIF	NO-DIF	NO-DIF
24	NO-DIF	NO-DIF	NO-DIF
25	NO-DIF	NO-DIF	NO-DIF
26	**DIF**	NO-DIF	NO-DIF
27	NO-DIF	NO-DIF	NO-DIF
28	NO-DIF	NO-DIF	NO-DIF
29	**DIF**	NO-DIF	**DIF**
30	**DIF**	**DIF**	**DIF**
จำนวนข้อที่พบ DIF	(6 ข้อ) 20%	(2 ข้อ) 7%	(6 ข้อ) 20%

หมายเหตุ DIF หมายถึง ข้อสอบที่พบว่าทำหน้าที่ต่างกันของข้อสอบ
NO-DIF หมายถึง ข้อสอบที่ไม่พบว่าทำหน้าที่ต่างกันของข้อสอบ

จากตารางที่ 4-5 แสดงผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยด้วยวิธี MIMIC จำแนกตามเพศ พบว่า

ด้านภาษา ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 6 ข้อ ได้แก่ ข้อที่ 1, 5, 12, 26, 29 และ 30 คิดเป็นร้อยละ 20 ของข้อสอบทั้งหมด

ด้านคำนวณ ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 2 ข้อ ได้แก่ ข้อที่ 1 และ 30 คิดเป็นร้อยละ 7 ของข้อสอบทั้งหมด

ด้านเหตุผล ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 6 ข้อ ได้แก่ ข้อที่ 1, 12, 15, 17, 29 และ 30 คิดเป็นร้อยละ 20 ของข้อสอบทั้งหมด

ตารางที่ 4-6 ผลการตรวจสอบการทำหน้าที่ต่างกันของแบบทดสอบระดับชาติ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี IRT-LR จำแนกตามเพศ ดังนี้

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี IRT-LR		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
1	DIF	NO-DIF	DIF
2	NO-DIF	DIF	DIF
3	DIF	DIF	NO-DIF
4	DIF	DIF	NO-DIF
5	DIF	DIF	DIF
6	NO-DIF	NO-DIF	DIF
7	DIF	DIF	DIF
8	NO-DIF	DIF	NO-DIF
9	NO-DIF	DIF	DIF
10	NO-DIF	NO-DIF	NO-DIF
11	DIF	DIF	DIF
12	DIF	DIF	DIF
13	NO-DIF	DIF	NO-DIF
14	NO-DIF	NO-DIF	NO-DIF
15	NO-DIF	NO-DIF	DIF
16	NO-DIF	DIF	DIF
17	DIF	DIF	DIF
18	NO-DIF	DIF	NO-DIF
19	DIF	NO-DIF	DIF
20	NO-DIF	DIF	DIF
21	NO-DIF	NO-DIF	DIF
22	NO-DIF	NO-DIF	DIF

ตารางที่ 4-6 (ต่อ)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี IRT-LR		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
23	**DIF**	NO-DIF	NO-DIF
24	**DIF**	NO-DIF	**DIF**
25	**DIF**	NO-DIF	NO-DIF
26	DIF	NO-DIF	**DIF**
27	**DIF**	NO-DIF	NO-DIF
28	NO-DIF	NO-DIF	NO-DIF
29	**DIF**	NO-DIF	NO-DIF
30	**DIF**	**DIF**	**DIF**
จำนวนข้อที่พบ DIF	(16 ข้อ) 53%	(15 ข้อ) 50%	(18 ข้อ) 60%

หมายเหตุ DIF หมายถึง ข้อสอบที่พบว่าทำหน้าที่ต่างกันของข้อสอบ
NO-DIF หมายถึง ข้อสอบที่ไม่พบว่าทำหน้าที่ต่างกันของข้อสอบ

จากตารางที่ 4-6 แสดงผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยด้วยวิธี IRT-LR จำแนกตามเพศ พบว่า

ด้านภาษา ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 16 ข้อ ได้แก่ข้อ ที่ 1, 3, 4, 5, 7, 11, 12, 17, 19, 23, 24, 25, 26, 27, 29 และ 30 คิดเป็นร้อยละ 53 ของข้อสอบทั้งหมด

ด้านคำนวณ ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 15 ข้อ ได้แก่ข้อที่ 2, 3, 4, 5, 7, 8, 9, 11, 12, 13, 16, 17, 18, 20 และ 30 คิดเป็นร้อยละ 50 ของข้อสอบทั้งหมด

ด้านเหตุผล ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 18 ข้อ ได้แก่ข้อที่ 1, 2, 3, 5, 6, 7, 9, 11, 12, 15, 16, 17, 19, 20 21, 22, 24, 26 และ 30 คิดเป็นร้อยละ 60

จากผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ (DIF) ในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล พบว่า

วิธี HGLM ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ด้านภาษา จำนวน 27 ข้อ คิดเป็นร้อยละ 90 ของข้อสอบด้านภาษา สำหรับด้านคำนวณพบ DIF จำนวน 13 ข้อ คิดเป็นร้อยละ 43 ของข้อสอบด้านคำนวณ และด้านเหตุผลพบ DIF จำนวน 22 ข้อ คิดเป็นร้อยละ 73 ของข้อสอบด้านเหตุผล

วิธี MIMIC ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ด้านภาษา จำนวน 6 ข้อ คิดเป็นร้อยละ 20 ของข้อสอบด้านภาษา สำหรับด้านคำนวณพบ DIF จำนวน 2 ข้อ คิดเป็นร้อยละ 7 ของข้อสอบด้านคำนวณ และด้านเหตุผลพบ DIF จำนวน 6 ข้อ คิดเป็นร้อยละ 20 ของข้อสอบ

วิธี IRT-LR ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ด้านภาษา จำนวน 16 ข้อ คิดเป็นร้อยละ 53 ของข้อสอบด้านภาษา สำหรับด้านคำนวณพบ DIF จำนวน

15 ข้อ คิดเป็นร้อยละ 50 ของข้อสอบด้านคำนวณ และด้านเหตุผลพบ DIF จำนวน 18 ข้อ คิดเป็นร้อยละ 60 ของข้อสอบด้านเหตุผล

ตอนที่ 3 ผลการเปรียบเทียบการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR จำแนกตามเพศ

การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้าน คือ ด้านภาษา ด้านคำนวณ และด้านเหตุผล จำนวนด้านละ 30 ข้อ โดยเปรียบเทียบระหว่าง 2 วิธี ว่าวิธีใดตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบจำนวนมากกว่ากัน

ตารางที่ 4-7 การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของแบบทดสอบระดับชาติ ด้านภาษา ระหว่างวิธี HGLM กับวิธี MIMIC จำแนกตามเพศ ดังนี้

ข้อที่	วิธีการตรวจสอบ DIF	
	วิธี HGLM	วิธี MIMIC
1	NO-DIF	**DIF**
2	**DIF**	NO-DIF
3	**DIF**	NO-DIF
4	**DIF**	NO-DIF
5	NO-DIF	**DIF**
6	**DIF**	NO-DIF
7	**DIF**	NO-DIF
8	**DIF**	NO-DIF
9	**DIF**	NO-DIF
10	**DIF**	NO-DIF
11	**DIF**	NO-DIF
12	**DIF**	**DIF**
13	**DIF**	NO-DIF
14	**DIF**	NO-DIF
15	**DIF**	NO-DIF
16	**DIF**	NO-DIF
17	**DIF**	NO-DIF
18	**DIF**	NO-DIF
19	**DIF**	NO-DIF
20	**DIF**	NO-DIF
21	**DIF**	NO-DIF
22	**DIF**	NO-DIF
23	**DIF**	NO-DIF

ตารางที่ 4-7 (ต่อ)

ข้อที่	วิธีการตรวจสอบ DIF	
	วิธี HGLM	วิธี MIMIC
24	**DIF**	NO-DIF
25	**DIF**	NO-DIF
26	**DIF**	**DIF**
27	**DIF**	NO-DIF
28	**DIF**	NO-DIF
29	**DIF**	**DIF**
30	NO-DIF	DIF
จำนวนข้อที่พบ DIF	(27 ข้อ) 90%	(6 ข้อ) 20%

จากตารางที่ 4-7 พบว่า วิธี HGLM ตรวจพบข้อสอบทำหน้าที่ต่างกันมากกว่าวิธี MIMIC ในด้านภาษา จำนวน 21 ข้อ คิดเป็นร้อยละ 70 ของข้อสอบทั้งหมด

ตารางที่ 4-8 การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของแบบทดสอบระดับชาติ ด้านภาษา ระหว่างวิธี HGLM กับวิธี IRT-LR จำแนกตามเพศ ดังนี้

ข้อที่	วิธีการตรวจสอบ DIF	
	วิธี HGLM	วิธี IRT-LR
1	NO-DIF	**DIF**
2	**DIF**	NO-DIF
3	**DIF**	**DIF**
4	**DIF**	**DIF**
5	NO-DIF	DIF
6	**DIF**	NO-DIF
7	**DIF**	**DIF**
8	**DIF**	NO-DIF
9	**DIF**	NO-DIF
10	**DIF**	NO-DIF
11	**DIF**	**DIF**
12	**DIF**	**DIF**
13	**DIF**	NO-DIF
14	**DIF**	NO-DIF
15	**DIF**	NO-DIF
16	**DIF**	NO-DIF
17	**DIF**	**DIF**

ตารางที่ 4-8 (ต่อ)

ข้อที่	วิธีการตรวจสอบ DIF	
	วิธี HGLM	วิธี IRT-LR
18	**DIF**	NO-DIF
19	**DIF**	**DIF**
20	**DIF**	NO-DIF
21	**DIF**	NO-DIF
22	**DIF**	NO-DIF
23	**DIF**	**DIF**
24	**DIF**	**DIF**
25	**DIF**	**DIF**
26	**DIF**	**DIF**
27	**DIF**	**DIF**
28	**DIF**	NO-DIF
29	**DIF**	**DIF**
30	NO-DIF	**DIF**
จำนวนข้อที่พบ DIF	(27 ข้อ) 90%	(16 ข้อ) 53%

จากตารางที่ 4-8 พบว่า วิธี HGLM ตรวจพบข้อสอบทำหน้าที่ต่างกันมากกว่าวิธี IRT-LR ในด้านภาษา จำนวน 11 ข้อ คิดเป็นร้อยละ 37 ของข้อสอบทั้งหมด

ตารางที่ 4-9 การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของแบบทดสอบระดับชาติ ด้านภาษา ระหว่างวิธี IRT-LR กับวิธี MIMIC จำแนกตามเพศ ดังนี้

ข้อที่	วิธีการตรวจสอบ DIF	
	วิธี IRT-LR	วิธี MIMIC
1	NO-DIF	**DIF**
2	**DIF**	NO-DIF
3	NO-DIF	NO-DIF
4	NO-DIF	NO-DIF
5	NO-DIF	**DIF**
6	**DIF**	NO-DIF
7	**DIF**	NO-DIF
8	**DIF**	NO-DIF
9	**DIF**	NO-DIF
10	**DIF**	NO-DIF
11	NO-DIF	NO-DIF

ตารางที่ 4-9 (ต่อ)

ข้อที่	วิธีการตรวจสอบ DIF	
	วิธี IRT-LR	วิธี MIMIC
12	DIF	DIF
13	DIF	NO-DIF
14	NO-DIF	NO-DIF
15	DIF	NO-DIF
16	NO-DIF	NO-DIF
17	NO-DIF	NO-DIF
18	NO-DIF	NO-DIF
19	DIF	NO-DIF
20	DIF	NO-DIF
21	DIF	NO-DIF
22	DIF	NO-DIF
23	DIF	NO-DIF
24	NO-DIF	NO-DIF
25	DIF	NO-DIF
26	DIF	DIF
27	NO-DIF	NO-DIF
28	DIF	NO-DIF
29	NO-DIF	DIF
30	NO-DIF	DIF
จำนวนข้อที่พบ DIF	(16 ข้อ) 53%	(6 ข้อ) 20%

จากตารางที่ 4-9 พบว่า วิธี IRT-LR ตรวจพบข้อสอบทำหน้าที่ต่างกันมากกว่าวิธี MIMIC ในด้านภาษา จำนวน 10 ข้อ คิดเป็นร้อยละ 33 ของข้อสอบทั้งหมด

ตารางที่ 4-10 การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของแบบทดสอบระดับชาติ ด้านคำนวณ ระหว่างวิธี HGLM กับวิธี MIMIC จำแนกตามเพศ ดังนี้

ข้อที่	วิธีการตรวจสอบ DIF	
	วิธี HGLM	วิธี MIMIC
1	NO-DIF	DIF
2	NO-DIF	NO-DIF
3	DIF	NO-DIF
4	NO-DIF	NO-DIF
5	NO-DIF	NO-DIF

ตารางที่ 4-10 (ต่อ)

ข้อที่	วิธีการตรวจสอบ DIF	
	วิธี HGLM	วิธี MIMIC
6	NO-DIF	NO-DIF
7	**DIF**	NO-DIF
8	**DIF**	NO-DIF
9	**DIF**	NO-DIF
10	NO-DIF	NO-DIF
11	NO-DIF	NO-DIF
12	**DIF**	NO-DIF
13	**DIF**	NO-DIF
14	NO-DIF	NO-DIF
15	NO-DIF	NO-DIF
16	**DIF**	NO-DIF
17	**DIF**	NO-DIF
18	NO-DIF	NO-DIF
19	**DIF**	NO-DIF
20	**DIF**	NO-DIF
21	NO-DIF	NO-DIF
22	NO-DIF	NO-DIF
23	**DIF**	NO-DIF
24	NO-DIF	NO-DIF
25	**DIF**	NO-DIF
26	NO-DIF	NO-DIF
27	**DIF**	NO-DIF
28	NO-DIF	NO-DIF
29	NO-DIF	NO-DIF
30	NO-DIF	**DIF**
จำนวนข้อที่พบ DIF	(13 ข้อ) 43%	(2 ข้อ) 7%

จากตารางที่ 4-10 พบว่า วิธี HGLM ตรวจพบข้อสอบทำหน้าที่ต่างกันมากกว่าวิธี MIMIC ในด้านคำนวณ จำนวน 11 ข้อ คิดเป็นร้อยละ 36 ของข้อสอบทั้งหมด

ตารางที่ 4-11 การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของแบบทดสอบระดับชาติ ด้านคำนวณ ระหว่างวิธี HGLM กับวิธี IRT-LR จำแนกตามเพศ ดังนี้

ข้อที่	วิธีการตรวจสอบ DIF	
	วิธี HGLM	วิธี IRT-LR
1	NO-DIF	NO-DIF
2	NO-DIF	**DIF**
3	**DIF**	**DIF**
4	NO-DIF	**DIF**
5	NO-DIF	**DIF**
6	NO-DIF	NO-DIF
7	**DIF**	**DIF**
8	**DIF**	**DIF**
9	**DIF**	**DIF**
10	NO-DIF	NO-DIF
11	NO-DIF	**DIF**
12	**DIF**	**DIF**
13	**DIF**	**DIF**
14	NO-DIF	NO-DIF
15	NO-DIF	NO-DIF
16	**DIF**	**DIF**
17	**DIF**	**DIF**
18	NO-DIF	**DIF**
19	**DIF**	NO-DIF
20	**DIF**	**DIF**
21	NO-DIF	NO-DIF
22	NO-DIF	NO-DIF
23	**DIF**	NO-DIF
24	NO-DIF	NO-DIF
25	**DIF**	NO-DIF
26	NO-DIF	NO-DIF
27	**DIF**	NO-DIF
28	NO-DIF	NO-DIF
29	NO-DIF	NO-DIF
30	NO-DIF	**DIF**
จำนวนข้อที่พบ DIF	(13 ข้อ) 43%	(15 ข้อ) 50%

จากตารางที่ 4-11 พบว่า วิธี HGLM ตรวจพบข้อสอบทำหน้าที่ต่างกันน้อยกว่าวิธี IRT-LR ในด้านคำนวณ จำนวน 2 ข้อ คิดเป็นร้อยละ 7 ของข้อสอบทั้งหมด

ตารางที่ 4-12 การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของแบบทดสอบระดับชาติ ด้านคำนวณ ระหว่างวิธี IRT-LR กับวิธี MIMIC จำแนกตามเพศ ดังนี้

ข้อที่	วิธีการตรวจสอบ DIF	
	วิธี IRT-LR	วิธี MIMIC
1	NO-DIF	**DIF**
2	**DIF**	NO-DIF
3	**DIF**	NO-DIF
4	**DIF**	NO-DIF
5	**DIF**	NO-DIF
6	NO-DIF	NO-DIF
7	**DIF**	NO-DIF
8	**DIF**	NO-DIF
9	**DIF**	NO-DIF
10	NO-DIF	NO-DIF
11	**DIF**	NO-DIF
12	**DIF**	NO-DIF
13	**DIF**	NO-DIF
14	NO-DIF	NO-DIF
15	NO-DIF	NO-DIF
16	**DIF**	NO-DIF
17	**DIF**	NO-DIF
18	**DIF**	NO-DIF
19	NO-DIF	NO-DIF
20	**DIF**	NO-DIF
21	NO-DIF	NO-DIF
22	NO-DIF	NO-DIF
23	NO-DIF	NO-DIF
24	NO-DIF	NO-DIF
25	NO-DIF	NO-DIF
26	NO-DIF	NO-DIF
27	NO-DIF	NO-DIF
28	NO-DIF	NO-DIF
29	NO-DIF	NO-DIF
30	**DIF**	**DIF**
จำนวนข้อที่พบ DIF	(15 ข้อ) 50%	(2 ข้อ) 7%

จากตารางที่ 4-12 พบว่า วิธี IRT-LR ตรวจพบข้อสอบทำหน้าที่ต่างกันมากกว่าวิธี MIMIC ในด้านคำนวณ จำนวน 13 ข้อ คิดเป็นร้อยละ 43 ของข้อสอบทั้งหมด

ตารางที่ 4-13 การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของแบบทดสอบระดับชาติ ด้านเหตุผล ระหว่างวิธี HGLM กับวิธี MIMIC จำแนกตามเพศ ดังนี้

ข้อที่	วิธีการตรวจสอบ DIF	
	วิธี IRT-LR	วิธี MIMIC
1	NO-DIF	**DIF**
2	**DIF**	NO-DIF
3	**DIF**	NO-DIF
4	**DIF**	NO-DIF
5	**DIF**	NO-DIF
6	**DIF**	NO-DIF
7	**DIF**	NO-DIF
8	**DIF**	NO-DIF
9	**DIF**	NO-DIF
10	**DIF**	NO-DIF
11	**DIF**	NO-DIF
12	**DIF**	**DIF**
13	**DIF**	NO-DIF
14	**DIF**	NO-DIF
15	NO-DIF	**DIF**
16	NO-DIF	NO-DIF
17	NO-DIF	**DIF**
18	**DIF**	NO-DIF
19	**DIF**	NO-DIF
20	**DIF**	NO-DIF
21	NO-DIF	NO-DIF
22	**DIF**	NO-DIF
23	**DIF**	NO-DIF
24	**DIF**	NO-DIF
25	NO-DIF	NO-DIF
26	**DIF**	NO-DIF
27	NO-DIF	NO-DIF
28	**DIF**	NO-DIF
29	NO-DIF	**DIF**
30	**DIF**	**DIF**
จำนวนข้อที่พบ DIF	(22 ข้อ) 73%	(6 ข้อ) 20%

จากตารางที่ 4-13 พบว่า วิธี HGLM ตรวจพบข้อสอบทำหน้าที่ต่างกันมากกว่าวิธี MIMIC ในด้านเหตุผล จำนวน 16 ข้อ คิดเป็นร้อยละ 53 ของข้อสอบทั้งหมด

ตารางที่ 4-14 การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของแบบทดสอบระดับชาติ ด้านเหตุผล ระหว่างวิธี HGLM กับวิธี IRT-LR จำแนกตามเพศ ดังนี้

ข้อที่	วิธีการตรวจสอบ DIF	
	วิธี HGLM	วิธี IRT-LR
1	NO-DIF	**DIF**
2	**DIF**	**DIF**
3	**DIF**	NO-DIF
4	**DIF**	NO-DIF
5	**DIF**	**DIF**
6	**DIF**	**DIF**
7	**DIF**	**DIF**
8	**DIF**	NO-DIF
9	**DIF**	**DIF**
10	**DIF**	NO-DIF
11	**DIF**	**DIF**
12	**DIF**	**DIF**
13	**DIF**	NO-DIF
14	**DIF**	NO-DIF
15	NO-DIF	**DIF**
16	NO-DIF	**DIF**
17	NO-DIF	**DIF**
18	DIF	NO-DIF
19	**DIF**	**DIF**
20	**DIF**	**DIF**
21	NO-DIF	**DIF**
22	**DIF**	**DIF**
23	**DIF**	NO-DIF
24	**DIF**	**DIF**
25	NO-DIF	NO-DIF
26	**DIF**	**DIF**
27	NO-DIF	NO-DIF
28	**DIF**	NO-DIF
29	NO-DIF	NO-DIF
30	**DIF**	**DIF**
จำนวนข้อที่พบ DIF	(22 ข้อ) 73%	(18 ข้อ) 60%

จากตารางที่ 4-14 พบว่า วิธี HGLM ตรวจพบข้อสอบทำหน้าที่ต่างกันมากกว่าวิธี IRT-LR ในด้านเหตุผล จำนวน 4 ข้อ คิดเป็นร้อยละ 13 ของข้อสอบทั้งหมด

ตารางที่ 4-15 การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของแบบทดสอบระดับชาติด้านเหตุผล ระหว่างวิธี IRT-LR กับวิธี MIMIC จำแนกตามเพศ ดังนี้

ข้อที่	วิธีการตรวจสอบ DIF	
	วิธี IRT-LR	วิธี MIMIC
1	DIF	DIF
2	DIF	NO-DIF
3	NO-DIF	NO-DIF
4	NO-DIF	NO-DIF
5	DIF	NO-DIF
6	DIF	NO-DIF
7	DIF	NO-DIF
8	NO-DIF	NO-DIF
9	DIF	NO-DIF
10	NO-DIF	NO-DIF
11	DIF	NO-DIF
12	DIF	DIF
13	NO-DIF	NO-DIF
14	NO-DIF	NO-DIF
15	DIF	DIF
16	DIF	NO-DIF
17	DIF	DIF
18	NO-DIF	NO-DIF
19	DIF	NO-DIF
20	DIF	NO-DIF
21	DIF	NO-DIF
22	DIF	NO-DIF
23	NO-DIF	NO-DIF
24	DIF	NO-DIF
25	NO-DIF	NO-DIF
26	DIF	NO-DIF
27	NO-DIF	NO-DIF
28	NO-DIF	NO-DIF
29	NO-DIF	DIF
30	DIF	DIF
จำนวนข้อที่พบ DIF	(18 ข้อ) 60%	(6 ข้อ) 20%

จากตารางที่ 4-15 พบว่า วิธี IRT-LR ตรวจพบข้อสอบทำหน้าที่ต่างกันมากกว่าวิธี MIMIC ในด้านเหตุผล จำนวน 12 ข้อ คิดเป็นร้อยละ 40 ของข้อสอบทั้งหมด

ตารางที่ 4-16 การเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR ทั้ง 3 ด้าน

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 4-16 ผลการเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR พบว่า

ด้านภาษา วิธี HGLM พบข้อสอบที่ทำหน้าที่ต่างกันมากกว่าวิธี MIMIC และวิธี IRT-LR คิดเป็นร้อยละ 70 และ 37 ตามลำดับ และวิธี IRT-LR พบข้อสอบที่ทำหน้าที่ต่างกันมากกว่าวิธี MIMIC คิดเป็นร้อยละ 33 ของข้อสอบทั้งหมด

ด้านคำนวณ วิธี HGLM พบข้อสอบที่ทำหน้าที่ต่างกันมากกว่าวิธี MIMIC คิดเป็นร้อยละ 36 และวิธี IRT-LR พบข้อสอบที่ทำหน้าที่ต่างกันมากกว่าวิธี MIMIC คิดเป็นร้อยละ 43 ส่วนวิธี HGLM พบข้อสอบที่ทำหน้าที่ต่างกันน้อยกว่าวิธี IRT-LR คิดเป็นร้อยละ 7 ของข้อสอบทั้งหมด

ด้านเหตุผล วิธี HGLM พบข้อสอบที่ทำหน้าที่ต่างกันมากกว่าวิธี MIMIC และวิธี IRT-LR คิดเป็นร้อยละ 13 และ 53 ตามลำดับ และวิธี IRT-LR พบข้อสอบที่ทำหน้าที่ต่างกันมากกว่าวิธี MIMIC คิดเป็นร้อยละ 40 ของข้อสอบทั้งหมด

บทที่ 5
สรุปและอภิปรายผล

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อวิเคราะห์คุณภาพของแบบทดสอบระดับชาติ ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ และเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ของนักเรียนชั้นประถมศึกษาปีที่ 3 จำนวน 3 ด้าน คือ ด้านภาษา ด้านคำนวณ และด้านเหตุผล โดยมีวิธีการดำเนินการวิจัยเป็น 3 ระยะ ดังนี้ ระยะที่ 1 การวิเคราะห์คุณภาพของแบบทดสอบระดับชาติ โดยใช้ทฤษฎีการตอบสนองข้อสอบ (IRT) แบบ 3 พารามิเตอร์ ระยะที่ 2 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ จำแนกตามเพศ ด้วยวิธีการตรวจสอบ 3 วิธี คือ วิธี HGLM วิเคราะห์ด้วยโปรแกรมสำเร็จรูป HLM วิธี MIMIC วิเคราะห์ด้วยโปรแกรมสำเร็จรูป Mplus และวิธี IRT-LR วิเคราะห์ด้วยโปรแกรมสำเร็จรูป IRTPRO ระยะที่ 3 การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธีการตรวจสอบ ทั้ง 3 วิธี จำแนกตามเพศ

สรุปผลการวิจัย

1. ผลการวิเคราะห์คุณภาพของแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล โดยการวิเคราะห์ตามทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ พบว่า แบบทดสอบระดับชาติด้านภาษา มีค่าอำนาจจำแนกของข้อสอบ (a) เฉลี่ยเท่ากับ 0.753 ค่าความยากของข้อสอบ (b) เฉลี่ยเท่ากับ 0.570 และค่าโอกาสการเดาของข้อสอบ (c) เฉลี่ยเท่ากับ 0.210 แสดงให้เห็นว่า แบบทดสอบระดับชาติด้านภาษา มีค่าอำนาจจำแนกของข้อสอบ (a) ทั้งฉบับอยู่ในระดับค่อนข้างดี มีค่าความยากของข้อสอบ (b) ทั้งฉบับอยู่ในระดับค่อนข้างยาก และมีค่าโอกาสการเดาของข้อสอบ (c) ทั้งฉบับไม่เกิน 0.3 สำหรับด้านคำนวณ มีค่าอำนาจจำแนกของข้อสอบ (a) เฉลี่ยเท่ากับ 1.020 มีค่าความยากของข้อสอบ (b) เฉลี่ยเท่ากับ 1.128 และค่าโอกาสในการเดาของข้อสอบ (c) เฉลี่ยเท่ากับ 0.204 แสดงให้เห็นว่า แบบทดสอบระดับชาติด้านคำนวณ มีค่าอำนาจจำแนกของข้อสอบ (a) ทั้งฉบับอยู่ในระดับค่อนข้างดี มีค่าความยากของข้อสอบ (b) ทั้งฉบับอยู่ในระดับยาก และมีค่าโอกาสการเดาของข้อสอบ (c) ทั้งฉบับไม่เกิน 0.3 ส่วนด้านเหตุผล มีค่าอำนาจจำแนกของข้อสอบ (a) เฉลี่ยเท่ากับ 0.787 มีค่าความยากของข้อสอบ (b) เฉลี่ยเท่ากับ 0.755 และค่าโอกาสการเดาของข้อสอบ (c) เฉลี่ยเท่ากับ 0.187 แสดงให้เห็นว่า แบบทดสอบระดับชาติด้านเหตุผล มีค่าอำนาจจำแนกของข้อสอบ (a) ทั้งฉบับอยู่ในระดับค่อนข้างดี มีค่าความยากของข้อสอบ (b) ทั้งฉบับอยู่ในระดับค่อนข้างยาก และมีค่าโอกาสการเดาของข้อสอบ (c) ทั้งฉบับไม่เกิน 0.3

2. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล จำแนกตามเพศ พบว่า วิธี HGLM ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน (DIF) ด้านภาษา จำนวน 27 ข้อ คิดเป็นร้อยละ 90 ของข้อสอบด้านภาษา สำหรับด้านคำนวณ พบ DIF จำนวน 13 ข้อ คิดเป็นร้อยละ 43 ของข้อสอบด้านคำนวณ และด้านเหตุผล พบ DIF จำนวน 22 ข้อ คิดเป็นร้อยละ 73 ของข้อสอบด้านเหตุผล ส่วนวิธี MIMIC ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน (DIF) ด้านภาษา จำนวน 6 ข้อ คิดเป็นร้อยละ 20 ของข้อสอบ

ด้านภาษา สำหรับด้านคำนวณ พบ DIF จำนวน 2 ข้อ คิดเป็นร้อยละ 7 ของข้อสอบด้านคำนวณ และด้านเหตุผล พบ DIF จำนวน 6 ข้อ คิดเป็นร้อยละ 20 ของข้อสอบ และวิธี IRT-LR ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน (DIF) ด้านภาษา จำนวน 16 ข้อ คิดเป็นร้อยละ 53 ของข้อสอบด้านภาษา สำหรับด้านคำนวณ พบ DIF จำนวน 15 ข้อ คิดเป็นร้อยละ 50 ของข้อสอบด้านคำนวณ และด้านเหตุผล พบ DIF จำนวน 18 ข้อ คิดเป็นร้อยละ 60 ของข้อสอบด้านเหตุผล อย่างมีนัยสำคัญทางสถิติที่ .05

3. ผลการเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล จำแนกตามเพศ ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR พบว่า ด้านภาษา วิธี HGLM พบ DIF มากกว่าวิธี MIMIC และวิธี IRT-LR คิดเป็นร้อยละ 70 และ 37 ของข้อสอบทั้งหมด และวิธี IRT-LR พบ DIF มากกว่าวิธี MIMIC คิดเป็นร้อยละ 33 ของข้อสอบทั้งหมด ส่วนด้านคำนวณ วิธี HGLM พบ DIFมากกว่าวิธี MIMIC คิดเป็นร้อยละ 36 ของข้อสอบทั้งหมด สำหรับวิธี HGLM พบ DIF น้อยกว่าวิธี IRT-LR คิดเป็นร้อยละ 7 ของข้อสอบทั้งหมด และวิธี IRT-LR พบ DIF มากกว่าวิธี MIMIC คิดเป็นร้อยละ 43 ของข้อสอบทั้งหมด และด้านเหตุผล วิธี HGLM พบ DIF มากกว่าวิธี MIMIC และวิธี IRT-LR คิดเป็นร้อยละ 53 และ 13 ของข้อสอบทั้งหมด ส่วนวิธี IRT-LR พบ DIF มากกว่าวิธี MIMIC คิดเป็นร้อยละ 40 ของข้อสอบทั้งหมด อย่างมีนัยสำคัญทางสถิติที่ .05

อภิปรายผลการวิจัย

ผลการวิเคราะห์คุณภาพของแบบทดสอบระดับชาติ โดยใช้ทฤษฎีการตอบสนองข้อสอบแบบ 3 พารามิเตอร์ การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ จำแนกตามเพศ และการเปรียบเทียบผลการตรวจพบข้อสอบที่ทำหน้าต่างกันของข้อสอบในแบบทดสอบระดับชาติ ทั้ง 3 วิธี มีประเด็นที่ควรอภิปราย ดังนี้

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) ทั้งฉบับอยู่ในระดับค่อนข้างยากและมีค่าโอกาสการเดาของข้อสอบ (c) ทั้งฉบับไม่เกิน 0.3

2. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธีการตรวจสอบ 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 ผลการศึกษาพบว่าวิธี HGLM จะตรวจสอบ DIF ได้ดีกว่าวิธี LR ในแบบทดสอบที่มีการตรวจให้คะแนนแบบ 2 ค่าสอดคล้องกับงานวิจัยของ 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-LRผลการศึกษาพบว่า วิธี HGLM ตรวจพบ DIF ได้มากที่สุด ส่วนวิธี LR และวิธี IRT-LR ตรวจพบ DIF ได้ใกล้เคียงกันจากการศึกษาของ Kabasakal, Arsan, Gok, and Kelecioglu (2014) ศึกษาประสิทธิภาพในการตรวจสอบ DIFด้วยวิธี 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-LRตรวจพบ DIF มากกว่าวิธี MIMIC ทั้ง 3 ด้าน คือ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ส่วนวิธี HGLM ตรวจพบ DIFมากกว่าวิธี IRT-LR ในด้านภาษา และด้านเหตุผล สอดคล้องกับงานวิจัยของ Ong, Lu, Lee, and Cohen (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)

ข้อเสนอแนะสำหรับการนำผลการวิจัยไปใช้

จากผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ และการเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ มีข้อเสนอแนะในการนำผลการวิจัยไปใช้ดังนี้

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

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

ข้อเสนอแนะสำหรับการวิจัยต่อไป

1. วิธี HGLM มีประสิทธิภาพในการตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบมากกว่า วิธี MIMIC และ วิธี 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

บรรณานุกรม

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

กระทรวงศึกษาธิการ. (2556, 11 กรกฎาคม). *นโยบายด้านการศึกษารัฐมนตรีว่าการกระทรวงศึกษาธิการ*. กรุงเทพฯ: กระทรวงศึกษาธิการ.

ชนะศึก นิชานนท์. (2553). ประสิทธิภาพของการประมาณค่าพารามิเตอร์แบบเบส์โดยใช้การสรุปอ้างอิงความน่าเชื่อถือของโมเดลการตอบสนองข้อสอบ. *วารสารวิจัย มสด*, *11*(2), 61-75.

ชัยวัฒน์ หฤทัยพันธ์. (2558). การพัฒนาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยการตัดสินของผู้เชี่ยวชาญ. *วารสารครุศาสตร์*, *43*(1), 1-18.

พิชชา สุริอาจ และประกฤติยา ทักษิโณ. (2559). การพัฒนาแบบวัดความตระหนักต่อโลกในยุคศตวรรษที่ 21 ของนักเรียนมัธยมศึกษาตอนต้นโดยใช้แบบวัดเชิงสถานการณ์: การประยุกต์ใช้การทำหน้าที่ต่างกันของข้อสอบ. *วารสารศึกษาศาสตร์ ฉบับวิจัยบัณฑิตศึกษา*, *10*(พิเศษ), 94-100

พีรญา สูงเนิน. (2552). การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ: การเปรียบเทียบระหว่างข้อสอบรายข้อกับหมดข้อสอบ โดยใช้วิธีซิปเทสท์. *วิทยาการวิจัยและวิทยาการปัญญา*, *6*(2), 49-62.

รติพร ถึงฝั่ง. (2556). การวิเคราะห์โมเดลมิมิค: การใช้ประโยชน์จากโปรแกรม LISREL รุ่งทดลองใช้เพื่องานวิจัย. *วารสารสมาคมวิจัย*, *18*(2), 128-140.

รุ่งนภา แสนอานวยผล. (2555). ประสิทธิภาพของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนรูปแบบผสม: การประยุกต์ ใช้ทฤษฎีการตอบสนองข้อสอบแบบตรวจให้คะแนนความรู้บางส่วน และทฤษฎีการตอบสนองข้อสอบแบบตรวจให้คะแนนความรู้บางส่วนแบบทั่วไป. *วารสารศึกษาศาสตร์ มหาวิทยาลัยขอนแก่น*, *35*(1), 58-66.

วรพจน์ วงศ์กิจรุ่งเรือง และอธิป จิตตฤกษ์. (2556). *ทักษแห่งอนาคตใหม่: การศึกษาเพื่อศตวรรษที่ 21* (พิมพ์ครั้งที่ 2). กรุงเทพฯ: สำนักพิมพ์ openworlds.

วรพรรณ ศรีกล่ำ. (2559). ปัจจัยพหุระดับที่ส่งผลต่อคะแนนการสอบประเมินคุณภาพการศึกษาระดับชาติ ด้านความสามารถทางภาษา: การศึกษาของโรงเรียน ที่มีผล NT ต่ำ ในจังหวัดพิษณุโลก. *วารสารราชภัฏสุราษฎร์ธานี*, *3*(2), 81-98.

ศิริชัย กาญจนวาสี. (2555). *ทฤษฎีการทดสอบแนวใหม่* (พิมพ์ครั้งที่ 4). กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

ศิริรัตน์ สุคันธพฤกษ์. (2554). การวิเคราะห์ข้อคำถามในแบบวัดความวิตกกังวลในการสอบคณิตศาสตร์: การเปรียบเทียบระหว่างไฮราชิคอลลิเนียร์โมเดล พาเชียลเครดิตโมเดล และเกรดเรสพอนส์โมเดล. *วารสารวิธีวิทยาการวิจัย*, *24*(2), 241-271.

สุชาต ประสิทธิ์รัฐสินธุ์, กรรณิการ์ สุขเกษม, โศภิต ผ่องศรี และถนอมรัตน์ ประสิทธิเมตต์. (2551). *แบบจำลองสมการโครงสร้าง: การใช้โปรแกรม LISREL, PRELIS และ SIMPLIS*. กรุงเทพฯ: ห้างหุ้นส่วนจำกัดสามลดา.

สุพัฒนา หอมบุปผา. (2556). การเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี HGLM วิธีMIMIC และวิธี BAYESIAN. *วารสารวิจัยราชภัฏพระนคร*, *8*(2), 10-24.

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

สำนักทดสอบทางการศึกษาและสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน. (2557). *ผลการประเมินคุณภาพผู้เรียนระดับชาติ ปีการศึกษา 2555 บทสรุปและข้อเสนอแนะเชิงนโยบาย*. กรุงเทพฯ: โรงพิมพ์ชุมนุมสหกรณ์การเกษตรแห่งประเทศไทย.

สำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน. (2557). *คู่มือการจัดสอบประเมินคุณภาพการศึกษาขั้นพื้นฐาน เพื่อประกันคุณภาพผู้เรียน ปีการศึกษา 2557*. กรุงเทพฯ: กระทรวงศึกษาธิการ.

เอกลักษณ์ คล้ายสุบรรณ, สังวรณ์ งัดกระโทก และนลินี ณ นคร. (2559). โมเดลการวัดมูลค่าเพิ่มทางการศึกษาสาหรับวัดคุณภาพสถานศึกษาด้วยการใช้ผลรวมของผลสัมฤทธิ์ทางการเรียนและผลการประเมินและรับรองคุณภาพของโรงเรียน. *Veridian E-Journal*, *9*(1), 1041-1052.

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.

Acar, T., (2013). Comparison of the Group and Intercept Coefficient from HGLM and LR-DIF Method. *British Journal of Science*, *10*(1), 12-20.

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel Item Response Model: An Approach to Error In Variables Regression. *Journal of Educational and Behavioral Statistics*, *22*(1), 47-76.

Avila, M. L., Stinson, J., Kiss, A., Brandão, L. R., Uleryk, E., & Feldman, B. M. (2015). A critical review of scoring options for clinical measurement tools. *BMC Research Notes*, *8*(1), 612.

Barnes, B. J., & Wells, C. S. (2009). Differential Item Functional Analysis by Gender and Race of the National Doctoral Program Survey. *International Journal of Doctoral Studies*, *4*, 77-96.

Bock, R. D., & Aitikin, M. (1981). Marginal Maximum Likelihood Estimation of Item Parameters: *An Application of an EM Algorithm. Psychometrika*, *46*, 443-459.

Bowen, N. K., & Guo, S. (2011). *Structural equation modeling.* Oxford, UK: Oxford University Press.

Breland, H., Lee, Y. (2007). Investigating uniform and non-uniform gender DIF in computer-based ESL writing assessment. *Applied Measurement in Education*, *20*(4), 377-403.

Brown, L. I., Bristol, L., De Four-Babb, J., & Conrad, D. A. (2014). National Tests and Diagnostic Feedback: What Say Teachers in Trinidad and Tobago?. *The Journal of Educational Research*, *107*(3), 241-251.

Brown, T. A. (2014). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items* (Vol. 4). Sage.

Carlton, S. T., & Harris, A. M. (1992). Characteristics associated with differential item functioning on the Scholastic Aptitude Test: Gender and majority/minority group comparisons. *ETS Research Report Series*, *1992*(2).

Cronbach, L. J., & Furby, L. (1970). How should we measure change - Or should we?. *Psychological Bulletin*, *74*, 68-80.

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

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Feinstein, Z. S. (1995). Effectt of Differing Item Parameters on Closed - Interval DIF Statistics. *Applied Psychological Measurement*, *19*(2), 131-142.

Fennema, E., & Carpenter, T. E (1981). Sex-related differences in mathematics: Results from the national assessment. *Mathematics Teacher*, *74*, 554–559.

Finch, W. 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.

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika*, *48*(1), 3-26.

Garner, M., & Engelhard, G. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education*, *12*, 29–51.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Boston: Kluwer Nijhoff Publishing.

Hambleton, R. K., Swaminatan, H., & Roger, H. J. 1991. *Fundamentals of Item Response Theory*. Newbury Park, California: SAGE Publications.

Holland, P. W., Wainer, H. (1993). *Differential Item Functioning*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Hulin, C. L., Darsgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Homewood, Illnois: Dow Jones - Irwin.

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. (1998). *One-parameter hierarchical generalized linear logistic model: an application of HGLM to IRT*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, California: American Educational Research Association.

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

Kjellstrom, K., & Pettersson, A. (2005). The curriculum's view of knowledge transferred to national tests in mathematics in Sweden. *ZDM*, *37*(4), 308-316.

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). Springer, New York: Springer Science Business Media.

Marshall, S. E. (1984). Sex differences in children's mathematics achievement: Solving computations and story problems. *Journal of Educational Psychology*, *76*, 194–204.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Education Statistics*, *7*, 105-118.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*(2), 127-143.

Mendes-Barnett, S., & Ercikan, K. (2006). Examining sources of gender DIF in mathematics assessments using a confirmatory multidimensional model approach. *Applied Measurement in Education*, *19*(4), 289-304.

Mucherah, W., Finch, W. H., & Keaikitse, S. (2012). Differential Bundle Functioning Analysis of the Self-Description Questionnaire Self-Concept Scale for Kenyan Female and Male Students Using the MIMIC Model. *International Journal of Testing*, *12*(1), 78-99.

Muthén, B. O., Kao, C. F., & Burstein, L. (1991). Instructionally sensitive psychometrics: Application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement*, *28*(1), 1-22.

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.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (Vol. 161). Sage Publications.

Pae, T. I. (2012). Causes of gender DIF on an EFL language test: A multiple-data analysis over nine years. *Language Testing*, *29*(4), 533-554.

Potenza, M. T., & Dorans, N. J. (1995). DIF Assessment For Polytomous Scored Items: A Framework For Classification And Evaluation. *Applied Paychological Measurement*, *19*(1), 23-37.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Riley, B., B., & Dennis, M., L., (2015) *Distinguishing between Treatment Effects and DIF in a Substance Abuse Outcome Measures Using Multiple Indicator Multiple Causes (MIMIC) Models*. Retrieved from http://slideplayer.com/slide/2753983/

Scheuneman, J. D. (1979). A method for assessing bias in test items. *Journal of Educational Measurement*, *16*, 143-152.

Seo, D. C., & Li, K. (2009). Effects of college climate on students' binge drinking: Hierarchical generalized linear model. *Annals of Behavioral Medicine*, *38*(3), 262-268.

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

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361-370.

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.

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Cook, K. F., Crane, P. K., Gibbons, L. E., & Cella, D. (2007). Evaluating measurement equivalence using the item response theory log-likelihood ratio (IRTLR) method to assess differential item functioning (DIF): applications (with illustrations) to measures of physical functioning ability and general distress. *Quality of Life Research*, *16*(1), 43-68.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin*, *99*(1), 118.

Urry, V. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Education Measuarement*, *14*, 181–196.

Wang, W., Shih, C.-L., & Yang, C. C. (2009). The MIMIC method with scale purification for detecting differential item functioning. *Educational and Psychological Measurement*, *69*, 713-731.

Wang, W., & Shih, C. –L. (2010). MIMIC Methods for Assessing Differential Item Functioning in Polytomous Items. *Applied Psychological Measurement*, *34*(3), 166-180.

Wiberg, M. (2007). *Measuring and Detecting Differential Item Functioning in Criterion-Referenced Licensing Test* (EM No. 60). Umea, Sweden: Umea University, Department of Educational Measurement.

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research*, *44*(1), 1-27.

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

Zenisky, A. L., Hambleton, R. K., & Robin, F. (2004). DIF detection and interpretation in large-scale science assessments: Informing item writing practices. *Educational Assessment*, *9*, 61–78.

ภาคผนวก

ภาคผนวก ก
หนังสือขอความอนุเคราะห์ขอข้อมูลเพื่อการวิจัย

ที่ ศธ ๖๖๒๘/๐๐๔๖

วิทยาลัยวิทยาการวิจัยและวิทยาการปัญญา
มหาวิทยาลัยบูรพา
ต.แสนสุข อ.เมือง จ.ชลบุรี ๒๐๑๓๑

๒๗ กุมภาพันธ์ ๒๕๕๙

เรื่อง ขอความอนุเคราะห์ขอข้อมูลเพื่อการวิจัย

เรียน ผู้อำนวยการสำนักทดสอบทางการศึกษา

สิ่งที่ส่งมาด้วย ๑. เค้าโครงวิทยานิพนธ์ฉบับย่อ จำนวน ๑ ชุด
๒. ผลการวิเคราะห์คุณภาพข้อสอบวัดความสามารถด้านภาษา ด้านคำนวณ และด้านเหตุผล ชั้นประถมศึกษาปีที่ ๓ ปีการศึกษา ๒๕๕๖ จำนวน ๑ ชุด
๓. แบบรายงานผลการพิจารณาจริยธรรมการวิจัยในคน จำนวน ๑ ชุด

ด้วย นางสาวสุธาทิพย์ ตรีสิน รหัสประจำตัว ๕๕๙๑๐๓๙๒ นิสิตหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาการวัดและเทคโนโลยีทางวิทยาการปัญญา ได้รับอนุมัติให้ทำวิทยานิพนธ์เรื่อง “การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ด้านภาษา ด้านการคิดคำนวณ และด้านเหตุผล ระดับชั้นประถมศึกษาปีที่ ๓ ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR” ซึ่งอยู่ในความควบคุมดูแลของ ดร.ปิยะทิพย์ ตินวร อาจารย์ที่ปรึกษาหลัก ในการนี้ ผู้วิจัยมีความประสงค์ขอความอนุเคราะห์ขอข้อมูลการทดสอบของนักเรียน รหัสโรงเรียน เพศ และผลการตอบแบบทดสอบของแบบทดสอบระดับชาติ (National Testing) ในปีการศึกษา ๒๕๕๖ จำนวน ๓ ด้าน คือ ด้านภาษา ด้านการคิดคำนวณ และด้านเหตุผล

จึงเรียนมาเพื่อโปรดพิจารณา วิทยาลัยวิทยาการวิจัยและวิทยาการปัญญา หวังเป็นอย่างยิ่งว่าคงจะได้รับความอนุเคราะห์จากท่านด้วยดี และขอขอบคุณมา ณ โอกาสนี้

ขอแสดงความนับถือ

(ผู้ช่วยศาสตราจารย์ ดร.สุชาดา กรเพชรปาณี)
คณบดีวิทยาลัยวิทยาการวิจัยและวิทยาการปัญญา

วิทยาลัยวิทยาการวิจัยและวิทยาการปัญญา
โทร. ๐ ๓๘๑๐ ๒๐๗๗-๘
โทร/ โทรสาร ๐ ๓๘๓๙ ๓๔๘๔
http ://www.rmcs.buu.ac.th

ภาคผนวก ข

ตัวอย่าง Printout ผลการวิเคราะห์คุณภาพของแบบทดสอบระดับชาติ

IRT Item Parameter Calibration Report

User Test 1

Report created on 24/5/2016

Xcalibre 4.1: IRT Item Parameter Estimation Software

Copyright © 2012 - Assessment Systems Corporation

Introduction

This report provides the results of the IRT item parameter calibration by the computer program Xcalibre Version 4.1.7.0 (Assessment Systems Corporation, 2012) for User Test 1. The output is divided into four sections:

1. Specifications
2. E-M Algorithm
3. Summary statistics
4. Item-by-item results.

The statistical output is also recorded in a comma-separated value (CSV) file of the same name.

Specifications

This section records the input/output specifications and settings for historical purposes.

The Windows paths for the input files used in this analysis were:

C:\Users\Admin\Desktop\NT-3 part(24-5-59)\Literacy-NT\LrtpXcal_Data.txt
C:\Users\Admin\Desktop\NT-3 part(24-5-59)\Literacy-NT\LrtpXcal_ICF.txt

The Windows paths for the output files produced by this analysis were:

Output_Literacy
Output_Literacy.csv
Output_Literacy Scores.csv

Table 1 presents the file specifications. Table 2 presents the IRT specifications used to perform the IRT item parameter calibration. Table 3 presents the flag specifications.

Table 1: File Specifications

Specification	Value	Specification	Value
Number of examinees	32000	Total Items	30
Calibrated Items	30	Pretest Items	0
Excluded Items	0	Number of domains	1
Classic Data Header	No	Delimited input	Yes
Delimiter for input	Tab	Number of ID columns	N/A
ID begins in column	N/A	Responses begin in column	N/A
Omit character	O	Not Admin character	N
Save item parameters	No	Item parameter format	N/A
Save data matrix	No	Include omit codes in matrix	N/A
Include Not Admin codes in matrix	N/A	Score Not Admin as omits	No
Plot the IRFs	Yes	Save the IRFs and IIFs	No
Produce the fit line	Yes	# Groups for Plot	15
Type of score groups	Equally sized	# Groups for Chi-square	15
Perform classification	No	Classify using	N/A
Two-group cutpoint	N/A	Low group label	N/A
High group label	N/A	Group IDs included in matrix	No

Table 2: IRT Calibration Specifications

Specification	Value	Specification	Value
IRT Specification	Dichotomous	Model constant	1.7
Polytomous IRT Model	N/A	Dichotomous IRT Model	3-parameter
Center the boundary locations	No	Centered value	N/A
Floating priors	Yes	a parameter prior mean (sd)	0.800 (0.300)
b parameter prior mean (sd)	0.000 (1.000)	c parameter prior mean (sd)	0.250 (0.025)
Theta estimation method	MLE	Bayesian prior mean (sd)	N/A
Maximum E-M loops	60	Convergence criterion	0.010
Quadrature points	25	Center dich item parameters on	theta
Acceptable P range	0.00 to 1.00	Acceptable item-corr range	0.00 to 1.00
Acceptable item mean range	0.00 to 15.00	Correct for spuriousness	Yes
Fit statistic critical alpha	0.050	Minimum a	0.05
Maximum a	6.00	Minimum b	-4.00
Maximum b	4.00	Minimum c	0.00
Maximum c	0.70	Minimum theta	-7.00
Maximum theta	7.00	Treat scored items as poly	No
Center poly parameters on theta	No	Test for DIF	No
Group status column	N/A	Ability levels for DIF Test	N/A
Group 1 code	N/A	Group 2 code	N/A
Group 1 label	N/A	Group 2 label	N/A
Exclude items with low N	No	Minimum valid N	N/A
Compute scaled scores	No	Mean (SD) of scaled scores	N/A
Minimum scaled score	N/A	Maximum scaled score	N/A
Save delimited output	Yes	Delimiter	Comma
Save scores output	Yes	Delimiter	Comma
Save delimited output	Yes	Delimiter	Comma

Table 3: Flag Specifications

Specification	Value	Specification	Value
Low a Flag Bound	0.30	High a Flag Bound	4.00
Low b Flag Bound	-3.00	High b Flag Bound	3.00
Low c Flag Bound	0.00	High c Flag Bound	0.40
Key Flag	K	Fit Flag	F
Low a Flag	La	High a Flag	Ha
Low b Flag	Lb	High b Flag	Hb
Low c Flag	Lc	High c Flag	Hc

E-M Algorithm

Xcalibre uses the expectation-maximization approach to calibrate item parameters. The estimation process is iterative, and repeated in loops until the convergence criterion is satisfied. The following list presents the item with the largest parameter change after each loop, and the value of the change.

The number of loops needed is evidence regarding the fit of the data; if many loops are required, or convergence is never reached, it means that the data does not fit well with the selected IRT model.

Maximum change after Loop 1 was 2.5560 for Item 26 for the b parameter
Maximum change after Loop 2 was -0.3983 for Item 2 for the a parameter
Maximum change after Loop 3 was 0.0604 for Item 22 for the a parameter
Maximum change after Loop 4 was -0.0349 for Item 29 for the b parameter
Maximum change after Loop 5 was -0.0278 for Item 29 for the b parameter
Maximum change after Loop 6 was -0.0229 for Item 29 for the b parameter
Maximum change after Loop 7 was -0.0193 for Item 29 for the b parameter
Maximum change after Loop 8 was -0.0166 for Item 29 for the b parameter
Maximum change after Loop 9 was -0.0144 for Item 29 for the b parameter
Maximum change after Loop 10 was -0.0127 for Item 29 for the b parameter
Maximum change after Loop 11 was -0.0124 for Item 26 for the b parameter
Maximum change after Loop 12 was -0.0123 for Item 26 for the b parameter
Maximum change after Loop 13 was -0.0120 for Item 26 for the b parameter
Maximum change after Loop 14 was -0.0117 for Item 26 for the b parameter
Maximum change after Loop 15 was -0.0114 for Item 26 for the b parameter
Maximum change after Loop 16 was -0.0110 for Item 26 for the b parameter
Maximum change after Loop 17 was -0.0106 for Item 26 for the b parameter
Maximum change after Loop 18 was -0.0102 for Item 26 for the b parameter
Maximum change after Loop 19 was -0.0098 for Item 26 for the b parameter

Summary statistics

Table 4 presents the summary statistics for the item parameters for all calibrated items. Table 5 summarizes the total scores for the full test for just the calibrated items. Table 6 summarizes the theta estimates for the full test. Table 7 provides the overall model fit chi-square(s) for the full test. Definitions of these statistics are found in the Xcalibre manual.

Table 4: Summary Statistics for All Calibrated Items

Parameter	Items	Mean	SD	Min	Max
a	30	0.787	0.311	0.197	1.388
b	30	0.603	0.972	-1.319	2.568
c	30	0.205	0.045	0.115	0.298

Table 5: Summary Statistics for the Total Scores

Test	Items	Alpha	Mean	SD	Skew	Min	Q1	Median	Q3	Max	IQR
Full Test	30	0.752	15.314	5.016	0.022	0	12.00	15.0	19.00	30	7.00

Table 6: Summary Statistics for the Theta Estimates

Test	Examinees	Mean	SD	Skew	Min	Q1	Median	Q3	Max	IQR
Full Test	31996	-0.033	1.058	-0.671	-7.000	-0.694	0.002	0.657	7.000	1.351

Table 7: Overall Model Fit

Test	Items	Chi-square	df	p	-2LL
Full Test	30	4441.039	360	0.000	1099782

Table 8 presents the item control information and item status for each item

Table 8: Item Control and Item Status for All Items

Seq.	Item ID	Key	Options	Domain	Inclusion	Item Type	Status
1	1	2	4	1	Y	M	Included
2	2	4	4	1	Y	M	Included
3	3	3	4	1	Y	M	Included
4	4	2	4	1	Y	M	Included
5	5	1	4	1	Y	M	Included
6	6	2	4	1	Y	M	Included
7	7	2	4	1	Y	M	Included
8	8	4	4	1	Y	M	Included
9	9	3	4	1	Y	M	Included
10	10	4	4	1	Y	M	Included
11	11	1	4	1	Y	M	Included
12	12	1	4	1	Y	M	Included
13	13	4	4	1	Y	M	Included
14	14	4	4	1	Y	M	Included
15	15	2	4	1	Y	M	Included
16	16	4	4	1	Y	M	Included
17	17	2	4	1	Y	M	Included
18	18	3	4	1	Y	M	Included
19	19	1	4	1	Y	M	Included
20	20	3	4	1	Y	M	Included
21	21	3	4	1	Y	M	Included
22	22	4	4	1	Y	M	Included
23	23	1	4	1	Y	M	Included
24	24	2	4	1	Y	M	Included
25	25	3	4	1	Y	M	Included
26	26	1	4	1	Y	M	Included
27	27	3	4	1	Y	M	Included
28	28	3	4	1	Y	M	Included
29	29	1	4	1	Y	M	Included
30	30	2	4	1	Y	M	Included

Table 9 presents the classical statistics, the item parameters, and any flags for each calibrated item.
The K flag indicates that the keyed alternative did not have the highest correlation with total score. The F flag indicates that the item fit statistic (z Resid for dichotomous / chi-square for polytomous) was significant, and the item did not fit the IRT model.The La, Lb, and Lc flags indicate that the a/b/c parameters were lower than the minimum acceptable value.The Ha, Hb, and Hc flags indicate that the a/b/c parameters were higher than the maximum acceptable value

Table 9: Item Parameters for All Calibrated Items

Seq.	Item ID	P	R	a	b	c	Flag(s)
1	1	0.479	0.193	0.424	0.871	0.195	
2	2	0.269	0.091	0.696	2.568	0.218	K
3	3	0.648	0.391	0.918	-0.409	0.115	F
4	4	0.750	0.311	0.706	-0.957	0.141	F
5	5	0.520	0.243	0.546	0.439	0.196	
6	6	0.732	0.269	0.608	-0.790	0.216	
7	7	0.666	0.412	1.107	-0.336	0.183	
8	8	0.433	0.255	0.662	0.874	0.195	
9	9	0.797	0.257	0.594	-1.319	0.184	
10	10	0.432	0.181	0.725	1.330	0.288	
11	11	0.667	0.282	0.748	-0.132	0.298	
12	12	0.575	0.406	1.268	0.099	0.221	
13	13	0.410	0.169	0.409	1.400	0.178	
14	14	0.483	0.370	1.300	0.498	0.233	
15	15	0.376	0.165	0.482	1.676	0.203	
16	16	0.442	0.294	0.657	0.602	0.136	
17	17	0.650	0.429	1.276	-0.217	0.205	
18	18	0.407	0.241	1.050	1.022	0.251	
19	19	0.746	0.427	1.261	-0.645	0.175	
20	20	0.621	0.355	0.838	-0.238	0.153	
21	21	0.460	0.219	0.604	0.945	0.234	
22	22	0.236	0.169	1.169	1.694	0.169	
23	23	0.328	0.147	0.604	1.918	0.212	
24	24	0.487	0.249	0.557	0.581	0.181	
25	25	0.371	0.151	0.619	1.787	0.252	
26	26	0.461	0.093	0.197	2.078	0.193	K, F, La
27	27	0.398	0.226	1.388	1.127	0.287	
28	28	0.537	0.311	0.958	0.419	0.266	
29	29	0.423	0.228	0.547	0.956	0.168	
30	30	0.553	0.290	0.693	0.252	0.215	

ภาคผนวก ค

ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ
ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR

ตารางที่ ค-1 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี HGLM

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี HGLM		
	ด้านภาษา	**ด้านคำนวณ**	**ด้านเหตุผล**
	เพศ (*p*-value)	เพศ (*p*-value)	เพศ (*p*-value)
1	.602 (NO-DIF)	.555 (NO-DIF)	.970 (NO-DIF)
2	.003 (DIF)	.462 (NO-DIF)	.004 (DIF)
3	.000 (DIF)	.043 (DIF)	.000 (DIF)
4	.000 (DIF)	.880 (NO-DIF)	.000 (DIF)
5	.304 (NO-DIF)	.891 (NO-DIF)	.000 (DIF)
6	.000 (DIF)	.833 (NO-DIF)	.000 (DIF)
7	.000 (DIF)	.000 (DIF)	.001 (DIF)
8	.000 (DIF)	.000 (DIF)	.009 (DIF)
9	.000 (DIF)	.000 (DIF)	.000 (DIF)
10	.004 (DIF)	.844 (NO-DIF)	.000 (DIF)
11	.000 (DIF)	.839 (NO-DIF)	.000 (DIF)
12	.000 (DIF)	.001 (DIF)	.000 (DIF)
13	.001 (DIF)	.000 (DIF)	.000 (DIF)
14	.000 (DIF)	.764 (NO-DIF)	.000 (DIF)
15	.003 (DIF)	.543 (NO-DIF)	.279 (NO-DIF)
16	.000 (DIF)	.000 (DIF)	.643 (NO-DIF)
17	.000 (DIF)	.000 (DIF)	.351 (NO-DIF)
18	.000 (DIF)	.084 (NO-DIF)	.000 (DIF)
19	.000 (DIF)	.022 (DIF)	.000 (DIF)
20	.000 (DIF)	.008 (DIF)	.001 (DIF)
21	.000 (DIF)	.056 (NO-DIF)	.654 (NO-DIF)
22	.002 (DIF)	.572 (NO-DIF)	.000 (DIF)
23	.004 (DIF)	.006 (DIF)	.000 (DIF)
24	.000 (DIF)	.976 (NO-DIF)	.000 (DIF)
25	.000 (DIF)	.037 (DIF)	.850 (NO-DIF)
26	.000 (DIF)	.808 (NO-DIF)	.000 (DIF)
27	.000 (DIF)	.001 (DIF)	.066 (NO-DIF)
28	.000 (DIF)	.053 (NO-DIF)	.001 (DIF)
29	.000 (DIF)	.795 (NO-DIF)	.164 (NO-DIF)
30	.453 (NO-DIF)	.098 (NO-DIF)	.001 (DIF)
รวม	27 ข้อ	13 ข้อ	22 ข้อ

ตารางที่ ค-2 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี MIMIC

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี MIMIC		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
	เพศ (*p*-value)	เพศ (*p*-value)	เพศ (*p*-value)
1	.026 (DIF)	.003 (DIF)	.028 (DIF)
2	.000 (NO-DIF)	.000 (NO-DIF)	.000 (NO-DIF)
3	.000 (NO-DIF)	.000 (NO-DIF)	.000 (NO-DIF)
4	.000 (NO-DIF)	.000 (NO-DIF)	.000 (NO-DIF)
5	.026 (DIF)	.000 (NO-DIF)	.000 (NO-DIF)
6	.000 (NO-DIF)	.000 (NO-DIF)	.000 (NO-DIF)
7	.000 (NO-DIF)	.000 (NO-DIF)	.000 (NO-DIF)
8	.000 (NO-DIF)	.000 (NO-DIF)	.000 (NO-DIF)
9	.000 (NO-DIF)	.000 (NO-DIF)	.000 (NO-DIF)
10	.000 (NO-DIF)	.000 (NO-DIF)	.000 (NO-DIF)
11	.000 (NO-DIF)	.000 (NO-DIF)	.000 (NO-DIF)
12	.025 (DIF)	.000 (NO-DIF)	.026 (DIF)
13	.000 (NO-DIF)	.000 (NO-DIF)	.000 (NO-DIF)
14	.000 (NO-DIF)	.000 (NO-DIF)	.000 (NO-DIF)
15	.000 (NO-DIF)	.000 (NO-DIF)	.027 (DIF)
16	.000 (NO-DIF)	.000 (NO-DIF)	.000 (NO-DIF)
17	.000 (NO-DIF)	.000 (NO-DIF)	.029 (DIF)
18	.000 (NO-DIF)	.000 (NO-DIF)	.000 (NO-DIF)
19	.000 (NO-DIF)	.000 (NO-DIF)	.000 (NO-DIF)
20	.000 (NO-DIF)	.000 (NO-DIF)	.000 (NO-DIF)
21	.000 (NO-DIF)	.000 (NO-DIF)	.000 (NO-DIF)
22	.000 (NO-DIF)	.000 (NO-DIF)	.000 (NO-DIF)
23	.000 (NO-DIF)	.000 (NO-DIF)	.000 (NO-DIF)
24	.000 (NO-DIF)	.000 (NO-DIF)	.000 (NO-DIF)
25	.000 (NO-DIF)	.000 (NO-DIF)	.000 (NO-DIF)
26	.027 (DIF)	.000 (NO-DIF)	.000 (NO-DIF)
27	.000 (NO-DIF)	.000 (NO-DIF)	.000 (NO-DIF)
28	.000 (NO-DIF)	.000 (NO-DIF)	.000 (NO-DIF)
29	.026 (DIF)	.000 (NO-DIF)	.029(DIF)
30	.026 (DIF)	.031 (DIF)	.030 (DIF)
รวม	6 ข้อ	2 ข้อ	6 ข้อ

ตารางที่ ค-3 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี IRT-LR

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี IRT-LR		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
	เพศ (*p*-value)	เพศ (*p*-value)	เพศ (*p*-value)
1	.000 (DIF)	.404 (NO-DIF)	.000 (DIF)
2	.802 (NO-DIF)	.000 (DIF)	.002 (DIF)
3	.017 (DIF)	.000 (DIF)	.088 (NO-DIF)
4	.002 (DIF)	.000 (DIF)	.221 (NO-DIF)
5	.000 (DIF)	.025 (DIF)	.000 (DIF)
6	.540 (NO-DIF)	.610 (NO-DIF)	.000 (DIF)
7	.044 (DIF)	.000 (DIF)	.000 (DIF)
8	.058 (NO-DIF)	.000 (DIF)	.623 (NO-DIF)
9	.080 (NO-DIF)	.049 (DIF)	.000 (DIF)
10	.499 (NO-DIF)	.184 (NO-DIF)	.822 (NO-DIF)
11	.010 (DIF)	.003 (DIF)	.000 (DIF)
12	.000 (DIF)	.000 (DIF)	.000 (DIF)
13	.461 (NO-DIF)	.000 (DIF)	.119 (NO-DIF)
14	.322 (NO-DIF)	.224 (NO-DIF)	.085 (NO-DIF)
15	.434 (NO-DIF)	.218 (NO-DIF)	.007 (DIF)
16	.726 (NO-DIF)	.001 (DIF)	.000 (DIF)
17	.030 (DIF)	.000 (DIF)	.025 (DIF)
18	.151 (NO-DIF)	.000 (DIF)	.122 (NO-DIF)
19	.000 (DIF)	.224 (NO-DIF)	.016 (DIF)
20	.465 (NO-DIF)	.003 (DIF)	.023 (DIF)
21	.723 (NO-DIF)	.118 (NO-DIF)	.000 (DIF)
22	.146 (NO-DIF)	.283 (NO-DIF)	.010 (DIF)
23	.012 (DIF)	.590 (NO-DIF)	.073 (NO-DIF)
24	.006 (DIF)	.140 (NO-DIF)	.000 (DIF)
25	.046 (DIF)	.777 (NO-DIF)	.521 (NO-DIF)
26	.000 (DIF)	.133 (NO-DIF)	.001 (DIF)
27	.000 (DIF)	.322 (NO-DIF)	.063 (NO-DIF)
28	.223 (NO-DIF)	.336 (NO-DIF)	.302 (NO-DIF)
29	.001 (DIF)	.052 (NO-DIF)	.275 (NO-DIF)
30	.000 (DIF)	.014 (DIF)	.000 (DIF)
รวม	16 ข้อ	15 ข้อ	18 ข้อ

ภาคผนวก ง

ตัวอย่าง Print Out การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ
ด้วยวิธี HGLM โดยโปรแกรม HLM

Program: HLM 6 Hierarchical Linear and Nonlinear Modeling
Authors: Stephen Raudenbush, Tony Bryk, & Richard Congdon
Publisher: Scientific Software International, Inc. (c) 2000
techsupport@ssicentral.com
www.ssicentral.com

Module: HLM2.EXE (6.04.27107.1)
Date: 15 December 2016, Thursday
Time: 16:40:42

SPECIFICATIONS FOR THIS NONLINEAR HLM2 RUN

Problem Title: no title

The data source for this run = Literacy.mdm
The command file for this run = D:\HGLM-dif\Literacy\Model2.hlm
Output file name = D:\HGLM-dif\Literacy\DIF2.txt
The maximum number of level-1 units = 288000
The maximum number of level-2 units = 9600
The maximum number of micro iterations = 14
Method of estimation: restricted PQL
Maximum number of macro iterations = 100

Distribution at Level-1: Bernoulli

Weighting Specification

	Weighting?	Weight Variable Name	Normalized?
Level 1	no		
Level 2	no		
Precision	no		

The outcome variable is RESPONSE

The model specified for the fixed effects was:

Level-1 Coefficients	Level-2 Predictors
INTRCPT1, B0	INTRCPT2, G00
	FEMALE, G01
# ITEM1 slope, B1	INTRCPT2, G10
	FEMALE, G11
# ITEM2 slope, B2	INTRCPT2, G20
	FEMALE, G21
# ITEM3 slope, B3	INTRCPT2, G30
	FEMALE, G31
# ITEM4 slope, B4	INTRCPT2, G40
	FEMALE, G41
# ITEM5 slope, B5	INTRCPT2, G50
	FEMALE, G51
# ITEM6 slope, B6	INTRCPT2, G60
	FEMALE, G61
# ITEM7 slope, B7	INTRCPT2, G70
	FEMALE, G71
# ITEM8 slope, B8	INTRCPT2, G80
	FEMALE, G81
# ITEM9 slope, B9	INTRCPT2, G90
	FEMALE, G91
# ITEM10 slope, B10	INTRCPT2, G100
	FEMALE, G101
# ITEM11 slope, B11	INTRCPT2, G110
	FEMALE, G111
# ITEM12 slope, B12	INTRCPT2, G120
	FEMALE, G121
# ITEM13 slope, B13	INTRCPT2, G130

FEMALE, G131
ITEM14 slope, B14 INTRCPT2, G140
FEMALE, G141
ITEM15 slope, B15 INTRCPT2, G150
FEMALE, G151
ITEM16 slope, B16 INTRCPT2, G160
FEMALE, G161
ITEM17 slope, B17 INTRCPT2, G170
FEMALE, G171
ITEM18 slope, B18 INTRCPT2, G180
FEMALE, G181
ITEM19 slope, B19 INTRCPT2, G190
FEMALE, G191
ITEM20 slope, B20 INTRCPT2, G200
FEMALE, G201
ITEM21 slope, B21 INTRCPT2, G210
FEMALE, G211
ITEM22 slope, B22 INTRCPT2, G220
FEMALE, G221
ITEM23 slope, B23 INTRCPT2, G230
FEMALE, G231
ITEM24 slope, B24 INTRCPT2, G240
FEMALE, G241
ITEM25 slope, B25 INTRCPT2, G250
FEMALE, G251
ITEM26 slope, B26 INTRCPT2, G260
FEMALE, G261
ITEM27 slope, B27 INTRCPT2, G270
FEMALE, G271
ITEM28 slope, B28 INTRCPT2, G280
FEMALE, G281
ITEM29 slope, B29 INTRCPT2, G290
FEMALE, G291

'#' - The residual parameter variance for this level-1 coefficient has been set to zero.

The model specified for the covariance components was:

Tau dimensions

INTRCPT1

Summary of the model specified (in equation format)

Level-1 Model

Prob(Y=1|B) = P

log[P/(1-P)] = B0 + B1*(ITEM1) + B2*(ITEM2) + B3*(ITEM3) + B4*(ITEM4) + B5*(ITEM5) + B6*(ITEM6) + B7*(ITEM7) + B8*(ITEM8) + B9*(ITEM9) + B10*(ITEM10) + B11*(ITEM11) + B12*(ITEM12) + B13*(ITEM13) + B14*(ITEM14) + B15*(ITEM15) + B16*(ITEM16) + B17*(ITEM17) + B18*(ITEM18) + B19*(ITEM19) + B20*(ITEM20) + B21*(ITEM21) + B22*(ITEM22) + B23*(ITEM23) + B24*(ITEM24) + B25*(ITEM25) + B26*(ITEM26) + B27*(ITEM27) + B28*(ITEM28) + B29*(ITEM29)

Level-2 Model

B0 = G00 + G01*(FEMALE) + U0

B1 = G10 + G11*(FEMALE)

B2 = G20 + G21*(FEMALE)

B3 = G30 + G31*(FEMALE)

B4 = G40 + G41*(FEMALE)

B5 = G50 + G51*(FEMALE)

B6 = G60 + G61*(FEMALE)

B7 = G70 + G71*(FEMALE)

B8 = G80 + G81*(FEMALE)

B9 = G90 + G91*(FEMALE)

B10 = G100 + G101*(FEMALE)

B11 = G110 + G111*(FEMALE)

B12 = G120 + G121*(FEMALE)

B13 = G130 + G131*(FEMALE)

B14 = G140 + G141*(FEMALE)
B15 = G150 + G151*(FEMALE)
B16 = G160 + G161*(FEMALE)
B17 = G170 + G171*(FEMALE)
B18 = G180 + G181*(FEMALE)
B19 = G190 + G191*(FEMALE)
B20 = G200 + G201*(FEMALE)
B21 = G210 + G211*(FEMALE)
B22 = G220 + G221*(FEMALE)
B23 = G230 + G231*(FEMALE)
B24 = G240 + G241*(FEMALE)
B25 = G250 + G251*(FEMALE)
B26 = G260 + G261*(FEMALE)
B27 = G270 + G271*(FEMALE)
B28 = G280 + G281*(FEMALE)
B29 = G290 + G291*(FEMALE)

Level-1 variance = 1/[P(1-P)]

The value of the likelihood function at iteration 2 = -1.914100E+005

RESULTS FOR NON-LINEAR MODEL WITH THE LOGIT LINK FUNCTION: Unit-Specific Model
(macro iteration 6)

Tau
INTRCPT1,B0 0.41200

Tau (as correlations)
INTRCPT1,B0 1.000

Random level-1 coefficient	Reliability estimate
INTRCPT1, B0	0.724

The value of the likelihood function at iteration 2 = -4.091102E+005

The outcome variable is RESPONSE

Final estimation of fixed effects: (Unit-specific model)

Fixed Effect	Coefficient	Standard Error	Approx. T-ratio	d.f.	P-value
For INTRCPT1, B0					
INTRCPT2, G00	0.257182	0.031461	8.175	9598	0.000
FEMALE, G01	-0.033415	0.044479	-0.751	9598	0.453
For ITEM1 slope, B1					
INTRCPT2, G10	-0.254477	0.042381	-6.004	287940	0.000
FEMALE, G11	-0.031176	0.059935	-0.520	287940	0.602
For ITEM2 slope, B2					
INTRCPT2, G20	-1.377658	0.045323	-30.396	287940	0.000
FEMALE, G21	0.195291	0.063491	3.076	287940	0.003

For ITEM3 slope, B3					
INTRCPT2, G30	0.115562	0.042670	2.708	287940	0.007
FEMALE, G31	0.625713	0.061661	10.148	287940	0.000
For ITEM4 slope, B4					
INTRCPT2, G40	0.696604	0.044418	15.683	287940	0.000
FEMALE, G41	0.567090	0.065230	8.694	287940	0.000
For ITEM5 slope, B5					
INTRCPT2, G50	-0.231277	0.042382	-5.457	287940	0.000
FEMALE, G51	0.061724	0.059935	1.030	287940	0.304
For ITEM6 slope, B6					
INTRCPT2, G60	0.694467	0.044409	15.638	287940	0.000
FEMALE, G61	0.328912	0.063976	5.141	287940	0.000
For ITEM7 slope, B7					
INTRCPT2, G70	0.198641	0.042820	4.639	287940	0.000
FEMALE, G71	0.649419	0.062145	10.450	287940	0.000
For ITEM8 slope, B8					
INTRCPT2, G80	-0.679181	0.042788	-15.873	287940	0.000
FEMALE, G81	0.416752	0.060220	6.920	287940	0.000
For ITEM9 slope, B9					
INTRCPT2, G90	1.068081	0.046511	22.964	287940	0.000
FEMALE, G91	0.509145	0.068777	7.403	287940	0.000
For ITEM10 slope, B10					
INTRCPT2, G100	-0.600025	0.042651	-14.068	287940	0.000
FEMALE, G101	0.175474	0.060181	2.916	287940	0.004
For ITEM11 slope, B11					
INTRCPT2, G110	0.433111	0.043416	9.976	287940	0.000
FEMALE, G111	0.294187	0.062133	4.735	287940	0.000
For ITEM12 slope, B12					
INTRCPT2, G120	-0.091807	0.042434	-2.164	287940	0.030
FEMALE, G121	0.376862	0.060377	6.242	287940	0.000
For ITEM13 slope, B13					
INTRCPT2, G130	-0.686596	0.042802	-16.041	287940	0.000
FEMALE, G131	0.205208	0.060328	3.402	287940	0.001
For ITEM14 slope, B14					
INTRCPT2, G140	-0.511790	0.042533	-12.033	287940	0.000
FEMALE, G141	0.399475	0.060057	6.652	287940	0.000

For ITEM15 slope, B15					
INTRCPT2, G150	-0.825750	0.043116	-19.152	287940	0.000
FEMALE, G151	0.180758	0.060722	2.977	287940	0.003
For ITEM16 slope, B16					
INTRCPT2, G160	-0.639479	0.042716	-14.971	287940	0.000
FEMALE, G161	0.368119	0.060170	6.118	287940	0.000
For ITEM17 slope, B17					
INTRCPT2, G170	0.132089	0.042698	3.094	287940	0.002
FEMALE, G171	0.611344	0.061687	9.910	287940	0.000
For ITEM18 slope, B18					
INTRCPT2, G180	-0.799111	0.043049	-18.563	287940	0.000
FEMALE, G181	0.268739	0.060546	4.439	287940	0.000
For ITEM19 slope, B19					
INTRCPT2, G190	0.579423	0.043927	13.191	287940	0.000
FEMALE, G191	0.780507	0.065491	11.918	287940	0.000
For ITEM20 slope, B20					
INTRCPT2, G200	0.103652	0.042652	2.430	287940	0.015
FEMALE, G201	0.450002	0.061093	7.366	287940	0.000
For ITEM21 slope, B21					
INTRCPT2, G210	-0.505458	0.042525	-11.886	287940	0.000
FEMALE, G211	0.320718	0.060035	5.342	287940	0.000
For ITEM22 slope, B22					
INTRCPT2, G220	-1.550727	0.046374	-33.440	287940	0.000
FEMALE, G221	0.208980	0.064804	3.225	287940	0.002
For ITEM23 slope, B23					
INTRCPT2, G230	-1.076223	0.043918	-24.505	287940	0.000
FEMALE, G231	0.178778	0.061728	2.896	287940	0.004
For ITEM24 slope, B24					
INTRCPT2, G240	-0.506363	0.042526	-11.907	287940	0.000
FEMALE, G241	0.462226	0.060084	7.693	287940	0.000
For ITEM25 slope, B25					
INTRCPT2, G250	-0.934059	0.043424	-21.510	287940	0.000
FEMALE, G251	0.260316	0.060980	4.269	287940	0.000
For ITEM26 slope, B26					
INTRCPT2, G260	-0.603685	0.042657	-14.152	287940	0.000
FEMALE, G261	0.403762	0.060126	6.715	287940	0.000

For ITEM27 slope, B27					
INTRCPT2, G270	-0.899890	0.043321	-20.773	287940	0.000
FEMALE, G271	0.375884	0.060733	6.189	287940	0.000
For ITEM28 slope, B28					
INTRCPT2, G280	-0.233954	0.042382	-5.520	287940	0.000
FEMALE, G281	0.422874	0.060199	7.025	287940	0.000
For ITEM29 slope, B29					
INTRCPT2, G290	-0.801008	0.043054	-18.605	287940	0.000
FEMALE, G291	0.458121	0.060428	7.581	287940	0.000

Fixed Effect	Coefficient	Odds Ratio	Confidence Interval
For INTRCPT1, B0			
INTRCPT2, G00	0.257182	1.293281	(1.216,1.376)
FEMALE, G01	-0.033415	0.967137	(0.886,1.055)
For ITEM1 slope, B1			
INTRCPT2, G10	-0.254477	0.775322	(0.714,0.842)
FEMALE, G11	-0.031176	0.969305	(0.862,1.090)
For ITEM2 slope, B2			
INTRCPT2, G20	-1.377658	0.252168	(0.231,0.276)
FEMALE, G21	0.195291	1.215665	(1.073,1.377)
For ITEM3 slope, B3			
INTRCPT2, G30	0.115562	1.122504	(1.032,1.220)
FEMALE, G31	0.625713	1.869579	(1.657,2.110)
For ITEM4 slope, B4			
INTRCPT2, G40	0.696604	2.006925	(1.840,2.189)
FEMALE, G41	0.567090	1.763130	(1.552,2.004)
For ITEM5 slope, B5			
INTRCPT2, G50	-0.231277	0.793520	(0.730,0.862)
FEMALE, G51	0.061724	1.063668	(0.946,1.196)
For ITEM6 slope, B6			
INTRCPT2, G60	0.694467	2.002642	(1.836,2.185)
FEMALE, G61	0.328912	1.389455	(1.226,1.575)

For ITEM7 slope, B7			
INTRCPT2, G70	0.198641	1.219744	(1.122,1.327)
FEMALE, G71	0.649419	1.914428	(1.695,2.162)
For ITEM8 slope, B8			
INTRCPT2, G80	-0.679181	0.507032	(0.466,0.551)
FEMALE, G81	0.416752	1.517026	(1.348,1.707)
For ITEM9 slope, B9			
INTRCPT2, G90	1.068081	2.909790	(2.656,3.188)
FEMALE, G91	0.509145	1.663869	(1.454,1.904)
For ITEM10 slope, B10			
INTRCPT2, G100	-0.600025	0.548798	(0.505,0.597)
FEMALE, G101	0.175474	1.191811	(1.059,1.341)
For ITEM11 slope, B11			
INTRCPT2, G110	0.433111	1.542047	(1.416,1.679)
FEMALE, G111	0.294187	1.342035	(1.188,1.516)
For ITEM12 slope, B12			
INTRCPT2, G120	-0.091807	0.912281	(0.839,0.991)
FEMALE, G121	0.376862	1.457704	(1.295,1.641)
For ITEM13 slope, B13			
INTRCPT2, G130	-0.686596	0.503286	(0.463,0.547)
FEMALE, G131	0.205208	1.227780	(1.091,1.382)
For ITEM14 slope, B14			
INTRCPT2, G140	-0.511790	0.599421	(0.551,0.652)
FEMALE, G141	0.399475	1.491042	(1.325,1.677)
For ITEM15 slope, B15			
INTRCPT2, G150	-0.825750	0.437906	(0.402,0.477)
FEMALE, G151	0.180758	1.198125	(1.064,1.350)
For ITEM16 slope, B16			
INTRCPT2, G160	-0.639479	0.527567	(0.485,0.574)
FEMALE, G161	0.368119	1.445013	(1.284,1.626)
For ITEM17 slope, B17			
INTRCPT2, G170	0.132089	1.141209	(1.050,1.241)
FEMALE, G171	0.611344	1.842907	(1.633,2.080)
For ITEM18 slope, B18			
INTRCPT2, G180	-0.799111	0.449729	(0.413,0.489)
FEMALE, G181	0.268739	1.308313	(1.162,1.473)

For ITEM19 slope, B19			
INTRCPT2, G190	0.579423	1.785009	(1.638,1.946)
FEMALE, G191	0.780507	2.182579	(1.920,2.482)
For ITEM20 slope, B20			
INTRCPT2, G200	0.103652	1.109214	(1.020,1.206)
FEMALE, G201	0.450002	1.568316	(1.391,1.768)
For ITEM21 slope, B21			
INTRCPT2, G210	-0.505458	0.603229	(0.555,0.656)
FEMALE, G211	0.320718	1.378117	(1.225,1.550)
For ITEM22 slope, B22			
INTRCPT2, G220	-1.550727	0.212094	(0.194,0.232)
FEMALE, G221	0.208980	1.232420	(1.085,1.399)
For ITEM23 slope, B23			
INTRCPT2, G230	-1.076223	0.340880	(0.313,0.372)
FEMALE, G231	0.178778	1.195755	(1.059,1.350)
For ITEM24 slope, B24			
INTRCPT2, G240	-0.506363	0.602684	(0.554,0.655)
FEMALE, G241	0.462226	1.587604	(1.411,1.786)
For ITEM25 slope, B25			
INTRCPT2, G250	-0.934059	0.392956	(0.361,0.428)
FEMALE, G251	0.260316	1.297340	(1.151,1.462)
For ITEM26 slope, B26			
INTRCPT2, G260	-0.603685	0.546793	(0.503,0.594)
FEMALE, G261	0.403762	1.497447	(1.331,1.685)
For ITEM27 slope, B27			
INTRCPT2, G270	-0.899890	0.406614	(0.374,0.443)
FEMALE, G271	0.375884	1.456278	(1.293,1.640)
For ITEM28 slope, B28			
INTRCPT2, G280	-0.233954	0.791398	(0.728,0.860)
FEMALE, G281	0.422874	1.526342	(1.356,1.717)
For ITEM29 slope, B29			
INTRCPT2, G290	-0.801008	0.448876	(0.413,0.488)
FEMALE, G291	0.458121	1.581100	(1.405,1.780)

ภาคผนวก จ

ตัวอย่าง Print Out การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ
ด้วยวิธี MIMIC โดยโปรแกรม Mplus

Mplus VERSION 7
MUTHEN & MUTHEN
06/29/2017 3:09 PM

INPUT INSTRUCTIONS

TITLE: Literacy Gender

DATA:
FILE IS "F:\DATA for MPlus\Literacy Gender.dat";

VARIABLE:
NAMES ARE GENDER ITEM1 ITEM2 ITEM3 ITEM4 ITEM5 ITEM6 ITEM7 ITEM8 ITEM9 ITEM10 ITEM11 ITEM12 ITEM13 ITEM14 ITEM15 ITEM16 ITEM17 ITEM18 ITEM19 ITEM20 ITEM21 ITEM22 ITEM23 ITEM24 ITEM25 ITEM26 ITEM27 ITEM28 ITEM29 ITEM30;
USEVARIABLES ARE ITEM1 ITEM2 ITEM3 ITEM4 ITEM5 ITEM6 ITEM7 ITEM8 ITEM9 ITEM10 ITEM11 ITEM12 ITEM13 ITEM14 ITEM15 ITEM16 ITEM17 ITEM18 ITEM19 ITEM20 ITEM21 ITEM22 ITEM23 ITEM24 ITEM25 ITEM26 ITEM27 ITEM28 ITEM29 ITEM30 GENDER;
CATEGORICAL ARE ITEM1-ITEM30;

MODEL RESULTS

				Two-Tailed
Estimate		S.E.	Est./S.E.	P-Value
LITERACY BY				
ITEM1	1.000	0.000	999.000	999.000
ITEM2	0.506	0.056	9.114	0.000
ITEM3	1.861	0.095	19.687	0.000
ITEM4	1.541	0.084	18.458	0.000
ITEM5	1.160	0.069	16.734	0.000
ITEM6	1.319	0.074	17.780	0.000
ITEM7	1.984	0.097	20.377	0.000
ITEM8	1.132	0.068	16.720	0.000
ITEM9	1.321	0.077	17.252	0.000
ITEM10	0.772	0.059	13.185	0.000

ITEM11	1.374	0.076	18.023	0.000
ITEM12	1.932	0.097	19.930	0.000
ITEM13	0.768	0.058	13.158	0.000
ITEM14	1.683	0.088	19.186	0.000
ITEM15	0.812	0.060	13.636	0.000
ITEM16	1.366	0.075	18.183	0.000
ITEM17	2.110	0.103	20.571	0.000
ITEM18	1.212	0.071	17.082	0.000
ITEM19	2.143	0.105	20.440	0.000
ITEM20	1.789	0.090	19.774	0.000
ITEM21	1.039	0.066	15.807	0.000
ITEM22	0.843	0.066	12.801	0.000
ITEM23	0.754	0.060	12.513	0.000
ITEM24	1.170	0.070	16.759	0.000
ITEM25	0.676	0.057	11.934	0.000
ITEM26	0.414	0.051	8.047	0.000
ITEM27	1.000	0.065	15.353	0.000
ITEM28	1.409	0.078	18.143	0.000
ITEM29	1.027	0.066	15.470	0.000
ITEM30	1.417	0.078	18.225	0.000
LITERACY ON				
GENDER	0.161	0.010	15.461	0.000
ITEM30 ON				
GENDER	-0.246	0.026	-9.596	0.000
ITEM1 ON				
GENDER	-0.197	0.026	-7.527	0.000
ITEM5 ON				
GENDER	-0.168	0.026	-6.494	0.000
ITEM26 ON				
GENDER	0.151	0.027	5.650	0.000
ITEM12 ON				
GENDER	-0.109	0.025	-4.434	0.000
ITEM29 ON				
GENDER	0.083	0.026	3.149	0.002
ITEM2 ON				
GENDER	0.000	0.000	999.000	999.000

ITEM3 ON GENDER	0.000	0.000	999.000	999.000
ITEM4 ON GENDER	0.000	0.000	999.000	999.000
ITEM6 ON GENDER	0.000	0.000	999.000	999.000
ITEM7 ON GENDER	0.000	0.000	999.000	999.000
ITEM8 ON GENDER	0.000	0.000	999.000	999.000
ITEM9 ON GENDER	0.000	0.000	999.000	999.000
ITEM10 ON GENDER	0.000	0.000	999.000	999.000
ITEM11 ON GENDER	0.000	0.000	999.000	999.000
ITEM13 ON GENDER	0.000	0.000	999.000	999.000
ITEM14 ON GENDER	0.000	0.000	999.000	999.000
ITEM15 ON GENDER	0.000	0.000	999.000	999.000
ITEM16 ON GENDER	0.000	0.000	999.000	999.000
ITEM17 ON GENDER	0.000	0.000	999.000	999.000
ITEM18 ON GENDER	0.000	0.000	999.000	999.000
ITEM19 ON GENDER	0.000	0.000	999.000	999.000
ITEM20 ON GENDER	0.000	0.000	999.000	999.000
ITEM21 ON GENDER	0.000	0.000	999.000	999.000
ITEM22 ON GENDER	0.000	0.000	999.000	999.000

ITEM23 ON				
GENDER	0.000	0.000	999.000	999.000
ITEM24 ON				
GENDER	0.000	0.000	999.000	999.000
ITEM25 ON				
GENDER	0.000	0.000	999.000	999.000
ITEM27 ON				
GENDER	0.000	0.000	999.000	999.000
ITEM28 ON				
GENDER	0.000	0.000	999.000	999.000
Thresholds				
ITEM1$1	-0.001	0.018	-0.029	0.977
ITEM2$1	0.647	0.020	33.114	0.000
ITEM3$1	-0.217	0.018	-11.879	0.000
ITEM4$1	-0.552	0.019	-28.852	0.000
ITEM5$1	-0.014	0.018	-0.779	0.436
ITEM6$1	-0.551	0.019	-28.802	0.000
ITEM7$1	-0.265	0.018	-14.468	0.000
ITEM8$1	0.247	0.018	13.519	0.000
ITEM9$1	-0.760	0.020	-37.790	0.000
ITEM10$1	0.201	0.018	11.045	0.000
ITEM11$1	-0.401	0.019	-21.525	0.000
ITEM12$1	-0.096	0.018	-5.280	0.000
ITEM13$1	0.252	0.018	13.750	0.000
ITEM14$1	0.150	0.018	8.248	0.000
ITEM15$1	0.332	0.018	18.001	0.000
ITEM16$1	0.224	0.018	12.283	0.000
ITEM17$1	-0.226	0.018	-12.397	0.000
ITEM18$1	0.317	0.018	7.198	0.000
ITEM19$1	-0.485	0.019	-25.683	0.000
ITEM20$1	-0.210	0.018	-11.509	0.000
ITEM21$1	0.146	0.018	8.048	0.000
ITEM22$1	0.743	0.020	37.110	0.000
ITEM23$1	0.476	0.019	25.257	0.000

ITEM24$1	0.147	0.018	8.077	0.000
ITEM25$1	0.395	0.019	21.208	0.000
ITEM26$1	0.203	0.018	11.156	0.000
ITEM27$1	0.375	0.019	20.206	0.000
ITEM28$1	-0.013	0.018	-0.693	0.488
ITEM29$1	0.318	0.018	17.254	0.000
ITEM30$1	-0.149	0.018	-8.221	0.000
Residual Variances				
LITERACY	0.090	0.008	10.863	0.000

STANDARDIZED MODEL RESULTS

StdYX Estimate	Std Estimate
LITERACY BY	
ITEM1	0.311	0.311
ITEM2	0.157	0.158
ITEM3	0.573	0.579
ITEM4	0.476	0.480
ITEM5	0.361	0.361
ITEM6	0.408	0.411
ITEM7	0.610	0.617
ITEM8	0.351	0.352
ITEM9	0.409	0.411
ITEM10	0.240	0.240
ITEM11	0.425	0.428
ITEM12	0.598	0.601
ITEM13	0.239	0.239
ITEM14	0.519	0.524
ITEM15	0.252	0.253
ITEM16	0.422	0.425
ITEM17	0.647	0.657
ITEM18	0.375	0.377
ITEM19	0.657	0.667
ITEM20	0.551	0.557
ITEM21	0.322	0.324

ITEM22	0.262	0.262
ITEM23	0.234	0.235
ITEM24	0.362	0.364
ITEM25	0.210	0.210
ITEM26	0.128	0.129
ITEM27	0.310	0.311
ITEM28	0.436	0.438
ITEM29	0.317	0.320
ITEM30	0.441	0.441
LITERACY ON		
GENDER	0.259	0.518
ITEM30 ON		
GENDER	-0.123	-0.246
ITEM1 ON		
GENDER	-0.099	-0.197
ITEM5 ON		
GENDER	-0.084	-0.168
ITEM26 ON		
GENDER	0.075	0.151
ITEM12 ON		
GENDER	-0.054	-0.109
ITEM29 ON		
GENDER	0.041	0.083
ITEM2 ON		
GENDER	0.000	0.000
ITEM3 ON		
GENDER	0.000	0.000
ITEM4 ON		
GENDER	0.000	0.000
ITEM6 ON		
GENDER	0.000	0.000
ITEM7 ON		
GENDER	0.000	0.000
ITEM8 ON		
GENDER	0.000	0.000

ITEM9 ON GENDER	0.000	0.000
ITEM10 ON GENDER	0.000	0.000
ITEM11 ON GENDER	0.000	0.000
ITEM13 ON GENDER	0.000	0.000
ITEM14 ON GENDER	0.000	0.000
ITEM15 ON GENDER	0.000	0.000
ITEM16 ON GENDER	0.000	0.000
ITEM17 ON GENDER	0.000	0.000
ITEM18 ON GENDER	0.000	0.000
ITEM19 ON GENDER	0.000	0.000
ITEM20 ON GENDER	0.000	0.000
ITEM21 ON GENDER	0.000	0.000
ITEM22 ON GENDER	0.000	0.000
ITEM23 ON GENDER	0.000	0.000
ITEM24 ON GENDER	0.000	0.000
ITEM25 ON GENDER	0.000	0.000
ITEM27 ON GENDER	0.000	0.000
ITEM28 ON GENDER	0.000	0.000

Thresholds		
ITEM1$1	-0.001	-0.001
ITEM2$1	0.646	0.647
ITEM3$1	-0.214	-0.217
ITEM4$1	-0.547	-0.552
ITEM5$1	-0.014	-0.014
ITEM6$1	-0.547	-0.551
ITEM7$1	-0.262	-0.265
ITEM8$1	0.246	0.247
ITEM9$1	-0.756	-0.760
ITEM10$1	0.201	0.201
ITEM11$1	-0.398	-0.401
ITEM12$1	-0.095	-0.096
ITEM13$1	0.251	0.252
ITEM14$1	0.148	0.150
ITEM15$1	0.331	0.332
ITEM16$1	0.223	0.224
ITEM17$1	-0.223	-0.226
ITEM18$1	0.315	0.317
ITEM19$1	-0.478	-0.485
ITEM20$1	-0.208	-0.210
ITEM21$1	0.146	0.146
ITEM22$1	0.741	0.743
ITEM23$1	0.475	0.476
ITEM24$1	0.146	0.147
ITEM25$1	0.394	0.395
ITEM26$1	0.202	0.203
ITEM27$1	0.374	0.375
ITEM28$1	-0.012	-0.013
ITEM29$1	0.315	0.318
ITEM30$1	-0.149	-0.149
Residual Variances		
LITERACY	0.933	0.933

ภาคผนวก ฉ

ตัวอย่าง Print Out การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ
ด้วยวิธี IRT-LR โดยโปรแกรม IRTPRO

IRTPRO Version 2.1
Output generated by IRTPRO estimation engine Version 4.54 (32-bit)

Project: DIF Literacy
Description:
Date: 06 December 2016
Time: 10:32 AM

Table of Contents

2PL Model Item Parameter Estimates for Group 1, logit: aθ + c or a(θ – b)

Item	Label	a		s.e.	c		s.e.	b	s.e.
1	ITEM1	2	0.57	0.04	1	0.28	0.04	-0.49	0.05
2	ITEM2	4	0.27	0.04	3	-0.94	0.04	3.46	0.55
3	ITEM3	6	1.17	0.05	5	1.02	0.05	-0.87	0.03
4	ITEM4	8	1.00	0.05	7	1.56	0.06	-1.56	0.05
5	ITEM5	10	0.65	0.04	9	0.34	0.04	-0.53	0.05
6	ITEM6	12	0.75	0.05	11	1.36	0.05	-1.82	0.08
7	ITEM7	14	1.24	0.06	13	1.17	0.05	-0.94	0.03
8	ITEM8	16	0.60	0.04	15	-0.14	0.04	0.23	0.07
9	ITEM9	18	0.83	0.05	17	1.82	0.06	-2.19	0.09
10	ITEM10	20	0.40	0.04	19	-0.14	0.03	0.35	0.10
11	ITEM11	22	0.72	0.04	21	1.07	0.04	-1.49	0.07
12	ITEM12	24	1.30	0.06	23	0.85	0.05	-0.66	0.03
13	ITEM13	26	0.41	0.04	25	-0.22	0.03	0.53	0.11
14	ITEM14	28	0.95	0.05	27	0.19	0.04	-0.19	0.04
15	ITEM15	30	0.39	0.04	29	-0.36	0.03	0.93	0.15
16	ITEM16	32	0.75	0.04	31	-0.03	0.04	0.05	0.05
17	ITEM17	34	1.31	0.06	33	1.14	0.05	-0.87	0.03
18	ITEM18	36	0.63	0.04	35	-0.25	0.04	0.39	0.07
19	ITEM19	38	1.44	0.06	37	1.79	0.07	-1.24	0.03
20	ITEM20	40	1.12	0.05	39	0.97	0.05	-0.87	0.03
21	ITEM21	42	0.52	0.04	41	0.01	0.03	-0.01	0.07

22	ITEM22	44	0.41	0.04	43	-1.06	0.04	2.58	0.30
23	ITEM23	46	0.31	0.04	45	-0.63	0.03	2.04	0.30
24	ITEM24	48	0.58	0.04	47	0.03	0.04	-0.05	0.06
25	ITEM25	50	0.28	0.04	49	-0.51	0.03	1.80	0.30
26	ITEM26	52	0.29	0.03	51	-0.19	0.03	0.66	0.17
27	ITEM27	54	0.41	0.04	53	-0.42	0.03	1.02	0.15
28	ITEM28	56	0.73	0.04	55	0.38	0.04	-0.52	0.04
29	ITEM29	58	0.51	0.04	57	-0.29	0.03	0.57	0.09
30	ITEM30	60	0.79	0.04	59	0.66	0.04	-0.83	0.04

2PL Model Item Parameter Estimates for Group 2, logit: aθ + c or a(θ – b)

Item	Label	a		s.e.	c		s.e.	b	s.e.
1	ITEM1	62	0.44	0.04	61	-0.06	0.03	0.13	0.07
2	ITEM2	64	0.25	0.04	63	-0.91	0.03	3.58	0.53
3	ITEM3	66	1.09	0.05	65	1.12	0.04	-1.03	0.05
4	ITEM4	68	0.82	0.05	67	1.59	0.04	-1.94	0.11
5	ITEM5	70	0.57	0.04	69	0.06	0.03	-0.10	0.05
6	ITEM6	72	0.77	0.05	71	1.32	0.04	-1.71	0.10
7	ITEM7	74	1.33	0.06	73	1.34	0.05	-1.01	0.04
8	ITEM8	76	0.57	0.04	75	-0.04	0.03	0.06	0.05
9	ITEM9	78	0.75	0.05	77	1.88	0.05	-2.50	0.16
10	ITEM10	80	0.39	0.04	79	-0.19	0.03	0.50	0.09
11	ITEM11	82	0.86	0.05	81	1.03	0.04	-1.20	0.06
12	ITEM12	84	1.11	0.05	83	0.60	0.04	-0.54	0.04
13	ITEM13	86	0.35	0.04	85	-0.25	0.03	0.70	0.11
14	ITEM14	88	1.00	0.05	87	0.13	0.03	-0.13	0.03
15	ITEM15	90	0.42	0.04	89	-0.41	0.03	0.97	0.11
16	ITEM16	92	0.71	0.04	91	-0.05	0.03	0.07	0.04
17	ITEM17	94	1.54	0.07	93	1.30	0.05	-0.85	0.03
18	ITEM18	96	0.69	0.04	95	-0.31	0.03	0.46	0.05
19	ITEM19	98	1.65	0.08	97	2.16	0.07	-1.31	0.04
20	ITEM20	100	1.06	0.05	99	0.90	0.04	-0.85	0.04
21	ITEM21	102	0.54	0.04	101	0.04	0.03	-0.08	0.06
22	ITEM22	104	0.51	0.04	103	-1.11	0.04	2.18	0.17
23 I	TEM23	106	0.46	0.04	105	-0.66	0.03	1.44	0.13

24	ITEM24	108	0.62	0.04	107	0.19	0.03	-0.30	0.05
25	ITEM25	110	0.39	0.04	109	-0.43	0.03	1.11	0.12
26	ITEM26	112	0.09	0.03	111	0.02	0.03	-0.26	0.33
27	ITEM27	114	0.64	0.04	113	-0.30	0.03	0.48	0.06
28	ITEM28	116	0.82	0.04	115	0.45	0.03	-0.54	0.05
29	ITEM29	118	0.52	0.04	117	-0.12	0.03	0.22	0.06
30	ITEM30	120	0.75	0.04	119	0.24	0.03	-0.32	0.04

DIF Statistics for Graded Items

Group 1	Group 2	Total	χ^2	d.f.	*p*	χ^2 a	d.f.	*p*	χ^2 c|a	d.f.
p										
1	1	52.4	2	0.0001	6.3	1	0.0119	46.1	1	0.0001
2	2	0.4	2	0.8024	0.1	1	0.7582	0.3	1	0.5568
3	3	8.1	2	0.0174	1.2	1	0.2808	6.9	1	0.0084
4	4	12.5	2	0.0019	6.4	1	0.0113	6.1	1	0.0133
5	5	35.2	2	0.0001	2.1	1	0.1502	33.2	1	0.0001
6	6	1.2	2	0.5399	0.1	1	0.7452	1.1	1	0.2887
7	7	6.2	2	0.0440	1.1	1	0.2900	5.1	1	0.0236
8	8	5.7	2	0.0575	0.2	1	0.6331	5.5	1	0.0192
9	9	5.0	2	0.0801	1.0	1	0.3114	4.0	1	0.0450
10	10	1.4	2	0.4993	0.1	1	0.7870	1.3	1	0.2516
11	11	9.1	2	0.0105	5.0	1	0.0253	4.1	1	0.0424
12	12	16.4	2	0.0003	6.0	1	0.0140	10.4	1	0.0013
13	13	1.6	2	0.4614	1.4	1	0.2333	0.1	1	0.7240
14	14	2.3	2	0.3220	0.6	1	0.4525	1.7	1	0.1923
15	15	1.7	2	0.4342	0.3	1	0.5665	1.3	1	0.2474
16	16	0.6	2	0.7258	0.6	1	0.4243	0.0	1	0.9654
17	17	7.0	2	0.0301	6.4	1	0.0116	0.6	1	0.4266
18	18	3.8	2	0.1512	1.1	1	0.2914	2.7	1	0.1027
19	19	16.2	2	0.0003	4.1	1	0.0420	12.1	1	0.0005
20	20	1.5	2	0.4648	0.7	1	0.3939	0.8	1	0.3698
21	21	0.6	2	0.7234	0.2	1	0.6924	0.5	1	0.4837
22	22	3.9	2	0.1461	2.8	1	0.0926	1.0	1	0.3137
23	23	8.8	2	0.0124	7.9	1	0.0050	0.9	1	0.3407
24	24	10.8	2	0.0045	0.5	1	0.4733	10.3	1	0.0013
25	25	6.2	2	0.0459	4.6	1	0.0323	1.6	1	0.2096

26	26	54.8	2	0.0001	17.1	1	0.0001	37.7	1	0.0001
27	27	20.8	2	0.0001	17.2	1	0.0001	3.7	1	0.0557
28	28	3.0	2	0.2229	2.4	1	.1241	0.6	1	0.4260
29	29	14.8	2	0.0006	0.1	1	0.8199	14.8	1	0.0001
30	30	72.9	2	0.0001	0.6	1	0.4468	72.3	1	0.0001

ภาคผนวก ช

ตัวอย่าง Print Out การทดสอบทางสถิติ Chi square Test ของผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR ของข้อสอบด้านภาษา ด้านคำนวณ และด้านเหตุผล

ตารางที่ ช-1 ผลการทดสอบทางสถิติ Chi square ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR

Methods * Response Crosstabulation

			Response		Total
			NO-DIF	DIF	
Methods	HGLM	Count	3_a	27_b	30
		% within Methods	10.0%	90.0%	100.0%
		% within Response	7.3%	55.1%	33.3%
		% of Total	3.3%	30.0%	33.3%
		Residual	-10.7	10.7	
	MIMIC	Count	24_a	6_b	30
		% within Methods	80.0%	20.0%	100.0%
		% within Response	58.5%	12.2%	33.3%
		% of Total	26.7%	6.7%	33.3%
		Residual	10.3	-10.3	
	IRT-LR	Count	14_a	16_a	30
		% within Methods	46.7%	53.3%	100.0%
		% within Response	34.1%	32.7%	33.3%
		% of Total	15.6%	17.8%	33.3%
		Residual	.3	-.3	
Total		Count	41	49	90
		% within Methods	45.6%	54.4%	100.0%
		% within Response	100.0%	100.0%	100.0%
		% of Total	45.6%	54.4%	100.0%

Each subscript letter denotes a subset of Response categories whose column proportions do not differ significantly from each other at the .05 level.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	29.657[a]	2	.000
Likelihood Ratio	33.070	2	.000
Linear-by-Linear Association	8.041	1	.005
N of Valid Cases	90		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 13.67.

ผลงานวิจัย

สุธาทิพย์ ตรีสิน และปิยะทิพย์ ประดุจพรม. (2560). การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ชั้นประถมศึกษาปีที่ 3 ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR. *วิทยาการวิจัยและวิทยาการปัญญา*, *15*(2), (In press)