การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN

สุมาลี ถามั่งมี

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต
สาขาวิชาการวิจัยและสถิติทางวิทยาการปัญญา
วิทยาลัยวิทยาการวิจัยและวิทยาการปัญญา มหาวิทยาลัยบูรพา
กรกฎาคม 2561

คณะกรรมการควบคุมวิทยานิพนธ์และคณะกรรมการสอบวิทยานิพนธ์ ได้พิจารณาวิทยานิพนธ์ของ สุมาลี ถามั่งมี ฉบับนี้แล้ว เห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาการวิจัยและสถิติทางวิทยาการปัญญา 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 สาขาวิชาการวิจัยและสถิติทางวิทยาการปัญญา ของมหาวิทยาลัยบูรพา

..... คณบดีวิทยาลัยวิทยาการวิจัย
(ผู้ช่วยศาสตราจารย์ ดร. สุชาดา กรเพชรปาณี) และวิทยาการปัญญา
วันที่ 14 เดือน กรกฎาคม พ.ศ. 2561

ประกาศคุณูปการ

วิทยานิพนธ์ฉบับนี้สำเร็จลงได้ด้วยความกรุณาจาก ดร.ปิยะทิพย์ ประดุจพรม อาจารย์ที่ปรึกษาหลัก ที่กรุณาให้คำปรึกษา แนะนำแนวทางที่ถูกต้อง ตลอดจนแก้ไขข้อบกพร่องต่าง ๆ ด้วยความละเอียด ขอขอบพระคุณเป็นอย่างสูงไว้ ณ โอกาสนี้

ขอขอบพระคุณอาจารย์จากวิทยาลัยวิทยาการวิจัยและวิทยาการปัญญา ที่ประสิทธิ์ประสาทความรู้ วิชา ตลอดจนเพื่อน ๆ วิทยาลัยวิทยาการวิจัยและวิทยาการปัญญา วิทยาเขตสระแก้วทุกคน ที่ให้ความอนุเคราะห์ช่วยเหลือด้านต่าง ๆ ด้วยดีตลอดมาในช่วงการเรียน

ขอกราบขอบพระคุณ คุณพ่อเสริญ คุณแม่พิไล ทิพย์มาก และพี่น้อง ครอบครัว ทุกคน ที่ให้กำลังใจ สนับสนุนผู้วิจัยเสมอมา

คุณค่าและประโยชน์ของวิทยานิพนธ์ฉบับนี้ ผู้วิจัยขอมอบเป็นกตัญญูกตเวทิตาแด่บุพการี บูรพาจารย์ และผู้มีพระคุณทุกท่านทั้งในอดีตและปัจจุบัน ที่ทำให้ข้าพเจ้าเป็นผู้มีการศึกษาและประสบความสำเร็จมาจนตราบเท่าทุกวันนี้

สุมาลี ถามั่งมี

56910402: สาขาวิชา: การวิจัยและสถิติทางวิทยาการปัญญา;
วท.ม. (การวิจัยและสถิติทางวิทยาการปัญญา)
คำสำคัญ: การทำหน้าที่ต่างกันของข้อสอบ/ ทฤษฎีการตอบสนองข้อสอบ/ ข้อสอบ NT/
วิธี HGLM/ วิธี MIMIC/ วิธี BAYESIAN

สุมาลี ถามั่งมี: การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN (A COMPARISON OF THE DIFFERENTIAL ITEM FUNCTIONING FOR NATIONAL TEST ITEM AT THE GRADE THREE LEVEL USING HGLM, MIMIC, AND BAYESIAN METHODS) คณะกรรมการควบคุมวิทยานิพนธ์: ปิยะทิพย์ ประดุจพรม, Ph.D. 193 หน้า ปี พ.ศ. 2561.

การวิจัยนี้มีวัตถุประสงค์เพื่อ 1) วิเคราะห์คุณภาพของข้อสอบ NT ทั้ง 3 ด้าน 2) ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN และ 3) เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN ใช้ข้อมูลทุติยภูมิ ซึ่งเป็นผลการตอบข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2555 จำนวน 3 ด้าน ประกอบด้วย 1) ด้านภาษา 2) ด้านคำนวณ และ 3) ด้านเหตุผล กลุ่มตัวอย่างเป็นนักเรียนชั้นประถมศึกษาปีที่ 3 จังหวัดสระแก้ว จำนวน 2,000 คน ที่เข้าทดสอบ NT จากสำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน กระทรวงศึกษาธิการ วิเคราะห์คุณภาพของข้อสอบโดยใช้ทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ ด้วยโปรแกรม Xcalibre Version 4.2.2

ผลการวิจัยปรากฏว่า

1. ข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้าน มีค่าเฉลี่ยของค่าอำนาจจำแนกของข้อสอบ (a) ทั้งฉบับอยู่ในระดับค่อนข้างดี สามารถจำแนกผู้สอบได้ดี มีค่าความยากของข้อสอบ (b) ทั้งฉบับอยู่ในระดับยาก และค่าโอกาสการเดาของข้อสอบ (c) ทั้งฉบับเฉลี่ยไม่เกิน .30

2. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ระดับชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้าน โดยด้านภาษา วิธี HGLM ตรวจพบ DIF มากที่สุด คิดเป็นร้อยละ 30 รองลงมา วิธี BAYESIAN คิดเป็นร้อยละ 23.33 และวิธี MIMIC คิดเป็นร้อยละ 3.33 ส่วนด้านคำนวณ วิธี MIMIC ตรวจพบ DIF มากที่สุด คิดเป็นร้อยละ 26.67 รองลงมา วิธี BAYESIAN คิดเป็นร้อยละ 20 และวิธี HGLM คิดเป็นร้อยละ 16.67 และด้านเหตุผล วิธี HGLM ตรวจพบ DIF มากที่สุด คิดเป็นร้อยละ 56.67 รองลงมาคือ วิธี MIMIC และวิธี BAYESIAN ตรวจพบ DIF เท่ากัน คิดเป็นร้อยละ 36.67

3. ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้านปรากฏว่า ด้านคำนวณ วิธี HGLM ตรวจพบ DIF น้อยกว่าวิธี MIMIC และวิธี BAYESIAN คิดเป็นร้อยละ 10และ 3.33 ตามลำดับ อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05 ด้านภาษาและด้านเหตุผล วิธี HGLM ตรวจพบ DIF มากกว่าวิธี MIMIC และวิธี BAYESIAN คิดเป็นร้อยละ 26.67, 20, 6.67 และ 20 ตามลำดับ ส่วนวิธี MIMIC ตรวจพบ DIF น้อยกว่า วิธี BAYESIAN ในด้านภาษา คิดเป็นร้อยละ 20 วิธี MIMIC ตรวจพบ DIF มากกว่า วิธี BAYESIAN ในด้านคำนวณ คิดเป็นร้อยละ 6.67 และวิธี MIMIC ตรวจพบ DIF เท่ากับ วิธี BAYESIAN ในด้านเหตุผล คิดเป็นร้อยละ 36.67

56910402: MAJOR: RESEARCH AND STATISTICS IN COGNITIVE SCIENCE;
M.Sc (RESEARCH AND STATISTICS IN COGNITIVE SCIENCE)
KEYWORDS: DIFFERENTIAL ITEM FUNCTIONING / ITEM RESPONSE THEORY/ NT EXAM/ HGLM METHOD/ MIMIC METHOD/ BAYESIAN METHOD

SUMALEE THAMUNGMEE: A COMPARISON OF THE DIFFERENTIAL ITEM FUNCTIONING FOR NATIONAL TEST ITEM AT THE GRADE THREE LEVEL USING HGLM, MIMIC, AND BAYESIAN METHODS. ADVISORY COMMITTEE: PIYATHIP PRADUJPROM, Ph.D., 193 P. 2018.

The objectives of this research were 1) to analyze the item quality of three subjects of National Test (NT), 2) to examine the Differential Item Functioning (DIF) of NT tests using HGLM, MIMIC and BAYESIAN methods, and 3) to compare performance of differential Item functioning NT tests using HGLM, MIMIC, and BAYESIAN methods by using the secondary data of the results from NT items for academic year 2011 across according to three subjects from 2000 Grade three students. The item quality was analyzed according to three parameters using Xcalibre Version 4.2.2.

Results were as follows.

1. The NT for Grade 3 students on three subjects had a fairly good level of the discrimination parameter value (a), had a difficulty level of difficulty parameter value (b), and the guessing parameter value (c) did not exceed .30.

2. The results of the examination of DIF of the NT items from the grade 3 students across three subjects were shown as follows: For literacy, the HGLM was found to be the most DIF item and it could account for 30%, followed by the BAYESIAN method at 23.33% and MIMIC method at 3.33%, respectively. For numeracy, the MIMIC method was found to be the most DIF test and it could account for 26.67%, followed by the BAYESIAN method at 20%, and the HGLM method at 16.67%. For reasoning, the HGLM was the most DIF test and it could account for 56.67%, followed by the MIMIC and the BAYESIAN methods at 36.67%.

3. The results of the comparison of DIF of the NT items the grade 3 students across three subjects indicated that the HGLM method was better than the MIMIC method in terms of DIF on literacy and reasoning ability and the HGLM method could account for 26.27% and 20%, respectively. The HGLM method was better than the BAYESIAN method in terms of DIF on reasoning and literacy ability and the HGLM method could account for 20 % and 6.67%, respectively. The MIMIC method was better than the BAYESIAN method in terms of DIF on numeracy and it could account for 6.67% but The MIMIC method was worse than the BAYESIAN method in terms of DIF on literacy and it could account for 20%. For reasoning ability, both methods did not differ but the HGLM method was worse than the MIMIC and BAYESIAN methods in terms of DIF on numeracy ability that could account for 10% and 3.33%, respectively.

สารบัญ

สารบัญ (ต่อ)

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญตาราง (ต่อ)

ตารางที่ หน้า

สารบัญภาพ

ภาพที่ หน้า

สารบัญภาพ (ต่อ)

สารบัญภาพ (ต่อ)

บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

การวัดและประเมินผลการเรียนรู้ของผู้เรียนตามจุดมุ่งหมายของการวัดและประเมินผลการเรียนรู้ตามหลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน พุทธศักราช 2551 อยู่บนจุดมุ่งหมายพื้นฐานสองประการ คือ 1) การวัดและประเมินผลเพื่อพัฒนาผู้เรียน และ 2) การวัดและประเมินผลเพื่อตัดสินผลการเรียนรู้ (กระทรวงศึกษาธิการ, 2551, หน้า 28) โดยการวัดและประเมินผลทางการศึกษามีประโยชน์หลายประการโดยเฉพาะทางด้านการบริหาร ด้านการจัดการเรียนการสอนด้านการทดสอบ และการตัดเกรด ด้านการแนะแนวและการให้คำปรึกษา และหากพิจารณาถึงนโยบายระดับประเทศ ที่มีความเกี่ยวข้องกับการวัดและประเมินผลการเรียนรู้ของผู้เรียน ตั้งแต่ระดับชาติจนถึงระดับชั้นเรียนและทำการประเมินในองค์ประกอบที่แตกต่างกัน เป็นข้อกำหนดและข้อบังคับที่สถานศึกษาต้องปฏิบัติตามเพื่อเป็นการพัฒนาคุณภาพของผู้เรียนให้เต็มตามศักยภาพและสอดคล้องกับนโยบายและความมุ่งหมายของชาติ

ปัจจุบันโลกมีการเปลี่ยนแปลงไปอย่างรวดเร็วด้วยเทคโนโลยีที่ทันสมัย การแสวงหาความรู้อย่างต่อเนื่อง การรับรู้ข้อมูลอย่างคิดวิเคราะห์ สังเคราะห์ คิดคำนวณ และรู้จักใช้ความเป็นเหตุเป็นผลในการใช้ข้อมูลข่าวสารที่ได้รับอย่างมีประสิทธิภาพและคล่องแคล่ว ซึ่งเป็นสิ่งจำเป็นสำหรับคนในยุคปัจจุบัน การจัดการเรียนการสอนที่มุ่งเน้นเพียงผลสัมฤทธิ์เพียงอย่างเดียวไม่เพียงพออีกต่อไป ประเทศต่างๆ ในโลกล้วนมุ่งปลูกฝังและพัฒนาประชากรของตน โดยเฉพาะนักเรียนในระดับชั้นการศึกษาขั้นพื้นฐานให้มีทักษะด้านอื่นๆ อีกหลายด้าน เช่น ทักษะการสื่อสาร ทักษะการคิดวิเคราะห์ ทักษะการทำงานร่วมกัน เป็นต้น เพื่อการอยู่ในโลกแห่งการแข่งขันได้อย่างปลอดภัยและมีความสุข ในช่วงไม่กี่ปีที่ผ่านมา ได้มีการกล่าวถึงทักษะที่จำเป็น ในศตวรรษที่ 21 ที่เด็กและเยาวชนควรมี คือ ทักษะการเรียนรู้และนวัตกรรม หรือ 3R และ 4C ซึ่งมีองค์ประกอบ ดังนี้ 3R ได้แก่ การอ่าน (Reading) การเขียน (Writing) และคณิตศาสตร์ (Arithmetic) และ 4C ได้แก่ การคิดวิเคราะห์ (Critical Thinking) การสื่อสาร (Communication) ความคิดสร้างสรรค์ (Creativity) และความร่วมมือ (Collaboration) ในปีการศึกษา 2555 สำนักทดสอบทางการศึกษาได้รับมอบหมายให้มีการปรับเปลี่ยนแนวทางการประเมินคุณภาพการศึกษา เพื่อการประกันคุณภาพผู้เรียน ซึ่งรับผิดชอบการประเมินนักเรียนสังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน ในชั้นประถมศึกษาปีที่ 3 จากการสอบวัดผลสัมฤทธิ์ทางการเรียนมาเป็นการสอบวัดสมรรถนะสำคัญ 3 ด้านของนักเรียน คือ ความสามารถด้านภาษา (Literacy Ability) ความสามารถด้านคำนวณ (Numeracy Ability) และความสามารถด้านเหตุผล (Reasoning Ability) โดยดำเนินการระดมสมองของผู้ทรงคุณวุฒิ ผู้เชี่ยวชาญและผู้เกี่ยวข้องทุกฝ่าย เพื่อกำหนดนิยามและกรอบโครงสร้างของเครื่องมือวัดสมรรถนะของนักเรียนทั้ง 3 ด้าน เพื่อรองรับการประกันคุณภาพการศึกษาได้เป็นอย่างดี (สำนักทดสอบทางการศึกษา, 2558, หน้า 1)

การทดสอบเป็นการดำเนินการที่ตั้งอยู่บนพื้นฐานการวัดคุณลักษณะแฝงภายในตัวบุคคล (Traits) โดยใช้ข้อสอบเป็นสิ่งเร้าให้ผู้ทดสอบแสดงความสามารถออกมาตอบสนอง หากมีข้อมูลที่

สามารถยืนยันได้ว่า ข้อสอบที่สร้างขึ้นมีคุณสมบัติวัดได้ตรงตามสิ่งที่ต้องการวัด (Validity) และผลการวัดมีความคงเส้นคงวา (Reliability) ก็ย่อมมั่นใจได้ระดับหนึ่งว่า ข้อสอบที่สร้างขึ้นมีคุณภาพเพียงใดนั้น ผู้พัฒนาข้อสอบต้องมีความรู้ถึงแก่นแท้ของเนื้อหาวิชาที่จะวัดประกอบกับความสามารถทักษะการเขียนข้อสอบ และต้องวางแผนการสร้างข้อสอบอย่างรอบคอบ ครอบคลุมเนื้อหาที่ต้องการวัด รวมทั้งมีการตรวจสอบคุณภาพของข้อสอบ ต้องนำข้อสอบที่สร้างขึ้นมาไปทดลองสอบกับกลุ่มตัวอย่าง และนำผลการตอบของผู้สอบมาวิเคราะห์หาคุณภาพของข้อสอบเป็นรายข้อ ผลการวิเคราะห์คุณภาพข้อสอบเป็นรายข้อนี้จะทำให้ทราบว่าข้อสอบแต่ละข้อสามารถทำหน้าที่ได้ตรงตามที่ผู้พัฒนาข้อสอบต้องการหรือไม่ เพื่อเป็นข้อมูลพื้นฐานสำหรับการจัดทำเป็นแบบทดสอบที่เหมาะสมต่อไป

ปัจจุบันการใช้แบบทดสอบแบบเลือกตอบ (Multiple Choice Test) ยังคงใช้ประเมินความสามารถของผู้เรียนอย่างแพร่หลายทั้งผลการประเมินการเรียนรู้ในสถานศึกษาระดับชาติหรือแบบทดสอบคัดเลือกเข้าศึกษาต่อในระดับอุดมศึกษา เนื่องจากแบบทดสอบเลือกตอบหรือหลายตัวเลือกมีข้อดีหลายประการด้วยกัน คือ 1) เป็นแบบทดสอบที่เหมาะสำหรับการวัดความรู้ความ สามารถตั้งแต่ขั้นต่ำไปถึงขั้นสูง 2) ใช้เวลาในการตรวจสอบค่อนข้างน้อย เหมาะสำหรับผู้สอบจำนวนมาก 3) มีความตรงตามเนื้อหาและความเที่ยงค่อนข้างสูง 4) เหมาะสำหรับการพัฒนาเป็นแบบ ทดสอบมาตรฐาน และ 5) ให้สารสนเทศด้านการวินิจฉัยการเรียนรู้ของผู้เรียนได้ (สุพัฒนา หอมบุปผา, 2556, หน้า 1) นอกจากนี้ในการวัดความสามารถพื้นฐานของผู้เรียนระดับชาติ (National Test: NT) จัดสอบโดยสำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน (สพฐ.) กระทรวง ศึกษาธิการ เป็นการสอบเพื่อประเมินคุณภาพการศึกษาขั้นพื้นฐานของนักเรียนแต่ละโรงเรียน เพื่อนำข้อมูลมาเป็นแผนพัฒนานักเรียนให้สามารถอ่านออกเขียนได้ รู้จักคิดวิเคราะห์ โดยจะ ทดสอบตามมาตรฐานการเรียนรู้ของหลักสูตรการศึกษาขั้นพื้นฐาน ทั้ง 3 ด้าน คือ ด้านความสามารถทางภาษา (Literacy Ability) ความสามารถด้านคำนวณ (Numeracy Ability) และความสามารถด้านเหตุผล (Reasoning Ability) ให้สอดคล้องกับจุดเน้นการพัฒนาผู้เรียนคือนักเรียนชั้นประถม ศึกษาปีที่ 1-3 นักเรียนมีทักษะความสามารถในการอ่านออกเขียนได้ คิดเลขเป็น มีทักษะการคิดขั้นพื้นฐาน เป็นการสอบวัดความรู้นักเรียนชั้นประถมศึกษาปีที่ 3 และใช้ข้อสอบมาตรฐานเดียวกันทั้งประเทศ

สำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน จัดตั้งขึ้นตามกฎกระทรวงแบ่งส่วนราชการสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน กระทรวงศึกษาธิการ พ.ศ. 2546 ตามความในมาตรา 8 และมาตรา 11 แห่งพระราชบัญญัติระเบียบบริหารราชการกระทรวง พ.ศ. 2546 รัฐมนตรีว่าการกระทรวงศึกษาธิการออกกฎกระทรวงไว้ให้สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน มีภารกิจเกี่ยวกับการจัดและการส่งเสริมการศึกษาขั้นพื้นฐาน ให้แบ่งส่วนราชการสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน ออกเป็น 10 สำนัก ซึ่งหนึ่งในนั้นมีสำนักทดสอบทางการศึกษาอยู่ด้วย (กฎกระทรวงแบ่งส่วนราชการสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน พ.ศ. 2546, 2546, หน้า 17) สำนักทดสอบทางการศึกษา มีอำนาจหน้าที่ในการศึกษา วิจัย และพัฒนาการวัดและประเมินผลทางการศึกษา รวมทั้งจัดระบบ วิธีการสอบ และพัฒนาเครื่องมือวัดมาตรฐานสำหรับการประเมินผลการจัดการศึกษา และการทดสอบทางการศึกษาขั้นพื้นฐาน ดำเนินการสอบวัดความรู้ ความสามารถคุณลักษณะระดับต่าง ๆ ให้กับนักเรียนและประชาชนทั่วไป พัฒนาและส่งเสริมวิชาการด้านการทดสอบและประเมินผลทางการศึกษา รวมถึงการพัฒนาบุคลากรด้านการทดสอบและประเมินผล ดำเนินการ

เกี่ยวกับระบบข้อมูลและทะเบียนประวัติผู้สำเร็จการศึกษาและจัดทำระบบ การเทียบโอนผลการศึกษา รวมทั้งประสานความร่วมมือด้านการทดสอบทางการศึกษาทั้งในระดับชาติและระดับนานาชาติ และปฏิบัติงานร่วมกับหรือสนับสนุนการปฏิบัติงานของหน่วยงานอื่นที่เกี่ยวข้องหรือที่ได้รับมอบหมาย (กฎกระทรวงแบ่งส่วนราชการสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน พ.ศ. 2546, หน้า 17)

การประเมินคุณภาพการศึกษาขั้นพื้นฐาน ของสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน ดำเนินการประเมินคุณภาพการศึกษาขั้นพื้นฐานกับนักเรียนชั้นประถมศึกษาปีที่ 3 โดยวัดความสามารถพื้นฐานสำคัญ 3 ด้าน คือ ความสามารถด้านภาษา (Literacy Ability) ความสามารถด้านคำนวณ (Numeracy Ability) และความสามารถด้านเหตุผล (Reasoning Ability) ซึ่งเป็นความสามารถพื้นฐานเบื้องต้นสำคัญที่ใช้ในการเรียนรู้ในระดับที่สูงขึ้นและยังสะท้อนไปสู่การยกระดับผลการประเมินระดับชาติ (O-NET) และนานาชาติ (PISA) (คู่มือการจัดสอบ NT ชั้น ป.3, 2555, หน้า 4) ผลการประเมินที่ได้จะเป็นข้อมูลสำคัญที่สะท้อนคุณภาพการดำเนินงานการจัดการศึกษาของสถานศึกษา เขตพื้นที่การศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน จำเป็นต้องมีข้อมูลผลการเรียนรู้ไปเตรียมความพร้อมผู้เรียน และเป็นตัวบ่งชี้คุณภาพการศึกษาขั้นพื้นฐานในภาพรวม เพื่อใช้เป็นข้อมูลประกอบ การตัดสินใจในการกำหนดนโยบาย กำหนดยุทธศาสตร์ แผนการศึกษาของชาติ ระดับสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน ระดับเขตพื้นที่การศึกษา ระดับสถานศึกษา ให้มีคุณภาพมาตรฐานระดับสากลบนพื้นฐานของความเป็นไทย ให้นักเรียนได้รับการพัฒนาศักยภาพสูงสุด มีความรู้และทักษะที่แข็งแกร่งและเหมาะสมเป็นพื้นฐานสำคัญในการเรียนรู้ระดับสูงขึ้นไป และการดำรงชีวิตในอนาคต ซึ่งในการพัฒนาความแข็งแกร่งทางการศึกษาให้ผู้เรียนทุกระดับ ทุกประเภทรวมถึงเด็กพิการและด้อยโอกาสให้มีความรู้และทักษะแห่งโลกยุคใหม่ควบคู่กันไป โดยเฉพาะทักษะการอ่าน เขียน และการคิด เพื่อให้มีความพร้อมเข้าสู่การศึกษาระดับสูง และโลกของการทำงาน โดยได้กำหนดเป็นยุทธศาสตร์พัฒนาคุณภาพผู้เรียนทุกระดับทุกประเภทเพื่อให้นักเรียนระดับการศึกษาขั้นพื้นฐาน ทุกคนมีพัฒนาการเหมาะสมตามช่วงวัยและมีคุณภาพเพื่อพัฒนานักเรียนให้มีคุณภาพและมาตรฐานใกล้เคียงกัน อีกทั้งเป็นการส่งเสริมการประกันคุณภาพภายในของสถานศึกษาให้มีความเข้มแข็งเพื่อรองรับการประเมินภายนอกซึ่งได้กำหนดเป็นกลยุทธ์ในการดำเนินงาน เช่น ส่งเสริมสนับสนุนการนำผลการทดสอบ O-NET การประเมินผล PISA และเพื่อเป็นการประกันคุณภาพภายในสถานศึกษา และเพื่อเตรียมการให้ผู้เรียน มีความพร้อมสำหรับรองรับการประเมินภายนอกของสถานศึกษา ทั้งการทดสอบระดับชาติหรือระดับนานาชาติ โดยจะมุ่งประเมินให้ทัดเทียมกับการประเมินผลนักเรียนร่วมกับนานาชาติ (PISA) การศึกษาแนวโน้มการจัดการศึกษาคณิตศาสตร์และวิทยาศาสตร์ (TIMSS) เป็นต้น ที่มีรูปแบบการประเมินที่หลากหลายมุ่งเน้นคุณภาพของนักเรียนโดยพิจารณาจากความสามารถพื้นฐานที่สำคัญ (คู่มือการจัดสอบประเมินคุณภาพการศึกษาขั้นพื้นฐานเพื่อการประกันคุณภาพผู้เรียน ปีการศึกษา 2556 ระดับชั้น ป.3, 2556, หน้า 2-3)

ผลการประเมินนักเรียนจะเป็นข้อมูลสำคัญช่วยปรับปรุงและพัฒนาคุณภาพการศึกษา โดยพิจารณาจากผลการประเมินผลสัมฤทธิ์ทางการเรียน ที่มีความสำคัญทั้งระดับผู้เรียน ระดับสถานศึกษา ระดับเขตพื้นที่การศึกษา และระดับชาติ แบบทดสอบจึงเป็นเครื่องมือที่ใช้ในการวัดผลประเมินผลทางการศึกษา โดยตรวจสอบว่าผู้เข้าสอบนั้นจะมีคุณลักษณะแฝงหรือความสามารถอยู่ในระดับใด ดังนั้นการสร้างและการตรวจสอบคุณภาพของแบบทดสอบจะต้องคำนึงถึงความตรง (Validity) เป็นสำคัญ ทั้งนี้เพราะว่าความตรงเป็นคุณสมบัติที่แสดงถึงความสามารถในการวัดได้อย่าง

ถูกต้องแม่นยำถ้าผลการวัดได้ค่าที่ใกล้เคียงกับค่าคุณลักษณะที่แท้จริงเพียงใด ก็ถือว่าการวัดมีความตรงมากขึ้นเพียงนั้น

การทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) เป็นการตรวจสอบคุณภาพด้านความตรง โดยเป็นการตรวจสอบในประเด็นของความยุติธรรมของข้อสอบและแบบทดสอบ (Item and Test Unfairness) แต่เดิมใช้คำว่า ความลำเอียงของข้อสอบ (Item Bias) หรือความลำเอียงของแบบทดสอบ (Test Bias) ซึ่งต่อมามีการเปลี่ยนมาใช้คำที่เหมาะสมกว่าเป็นการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) จากผลการตรวจข้อสอบของผู้สอบกลุ่มต่าง ๆ ในประชากรมีมานานแล้ว แต่การศึกษาคุณภาพด้านความยุติธรรมของข้อสอบหรือแบบทดสอบระหว่างผู้สอบกลุ่มต่าง ๆ ปลายปี ค.ศ. 1960 มีการเสนอวิธีการต่างๆ เพื่อตรวจสอบความลำเอียงของข้อสอบ (Item Bias) ความลำเอียงของแบบทดสอบ (Test Bias) และความลำเอียงในการคัดเลือก (Selection Bias) โดยนิยามความลำเอียงว่าเป็นความคลาดเคลื่อนอย่างเป็นระบบ (Systematic Error) ซึ่งเป็นการขจัดข้อสอบที่ทำให้เกิดปัญหาความไม่ยุติธรรมระหว่างกลุ่มข้อสอบกลุ่มต่างๆที่มีลักษณะบางอย่างแตกต่างกัน (ศิริชัย กาญจนวาสี, 2555, หน้า 115)

การทดสอบแต่ละครั้ง ผู้สอบอาจจะมีลักษณะแตกต่างกัน เช่น เชื้อชาติ ศาสนา วัฒนธรรม ภูมิลำเนา สังคม เพศ ภาษา อายุ สภาพทางเศรษฐกิจ และประสบการณ์ เป็นต้น ทำให้ผู้สอบไม่ได้รับความยุติธรรมในการทำข้อสอบ โดยข้อสอบบางข้อ อาจมีความลำเอียงเข้าข้างผู้สอบกลุ่มย่อยบางกลุ่มของผู้เข้าสอบทั้งหมด ทำให้เกิดการได้เปรียบหรือเสียเปรียบระหว่างกลุ่มผู้สอบด้วยกัน ทั้ง ๆ ที่สอบด้วยข้อสอบข้อเดียวกันหรือแบบทดสอบฉบับเดียวกัน แสดงว่าแบบทดสอบหรือข้อสอบฉบับนั้นขาดความตรง คือ ไม่ได้วัดความสามารถหลักที่ต้องการวัด (Target Ability) เพียงอย่างเดียว แต่ยังวัดความสามารถแทรกซ้อนที่ไม่ต้องการวัด (Nuisance Ability) อีกด้วย เช่น แบบทดสอบวัดความสามารถด้านคำศัพท์ภาษาไทยฉบับหนึ่ง ข้อสอบบางข้ออาจถามความรู้สำหรับผู้ชายเป็นพิเศษ เช่น ความรู้เรื่องกีฬา ในบางข้ออาจถามความรู้สำหรับผู้หญิงเป็นพิเศษ เช่น ความรู้เรื่องการตัดเย็บ จากสถานการณ์นี้ แบบทดสอบวัดความสามารถคำศัพท์ในวิชาภาษาไทยเป็นความสามารถหลักที่ต้องการวัด ส่วนความสามารถด้านกีฬาและด้านการตัดเย็บเป็นความสามารถแทรกซ้อน ทำให้การตอบข้อสอบกลุ่มย่อยมีโอกาสการตอบถูกไม่เท่ากันขึ้นอยู่กับกลุ่มใดมีความสามารถแทรกซ้อนสูงกว่ากัน ทั้ง ๆ ที่ระดับความสามารถหลักที่ต้องการวัดเท่ากัน จึงทำให้ข้อสอบนั้นทำหน้าที่ต่างกัน

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เป็นการเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มผู้สอบอย่างน้อย 2 กลุ่มขึ้นไป กลุ่มแรก เรียกว่า กลุ่มเปรียบเทียบ (Focal Group: F) เป็นกลุ่มที่ผู้วิจัยสนใจศึกษา และคาดว่าเป็นกลุ่มที่เสียประโยชน์ในการตอบข้อสอบ และกลุ่มที่สอง เรียกว่า กลุ่มอ้างอิง (Reference Group: R) เป็นกลุ่มที่คาดว่าจะได้ประโยชน์จากการตอบข้อสอบได้ถูกต้อง วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่สำคัญๆ ได้แก่ การวิเคราะห์ความแปรปรวน (Analysis of Variance: ANOVA) วิธีการวิเคราะห์การถดถอยโลจิสติก (Logistic Regression: LR) วิธีแปลงค่าความยากของข้อสอบ (Transformed Item Difficulty: TID) วิธีวัดพื้นที่ความแตกต่างระหว่างโค้งการตอบสนองข้อสอบ (Item Response Theory – D2: IRT-D2) วิธีแมนเทล-แฮนส์เซล (Mantel-Haenszel: MH) วิธีไคสแควร์ของลอร์ด (Lord’s chi square) วิธีอัตราส่วนไลด์ลิฮูด ลอกลิเนียร์ (Loglinear Likelihood Ratio) และวิธี SIBTEST (ศิริชัย กาญจนวาสี, 2555, หน้า 124-125)

จากการศึกษาโมเดลเชิงเส้นตรงทั่วไประดับลดหลั่น (Hierarchical Generalized Linear Model: HGLM) ค่าอิทธิพลของตัวแปรภายนอกต่อโอกาสในการตอบข้อสอบในการวิเคราะห์ ระดับที่ 2 (ระดับผู้สอบ) สามารถดำเนินการวิเคราะห์ได้จากโปรแกรม HLM ด้วยโมเดลเชิงเส้นตรงทั่วไประดับลดหลั่น (HGLM) ได้ทำการวิเคราะห์การประมาณค่าพารามิเตอร์ความยากของข้อสอบ (b) ค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) จากโปรแกรม HLM ซึ่งมีลักษณะเป็นพารามิเตอร์แบบสุ่ม (Random Parameter) การดำเนินการวิเคราะห์ สามารถดำเนินการวิเคราะห์ในขั้นตอนเดียวตามโมเดล HGLM ด้วยโปรแกรมโมเดลเชิงเส้นตรงระดับลดหลั่น (HLM) ที่ผ่านมาส่วนใหญ่นักวิจัยได้ดำเนินการวิเคราะห์ในลักษณะแยกส่วน ซึ่งในการวิจัยครั้งนี้ผลการวิเคราะห์ข้อสอบ นอกจากจะให้ค่าพารามิเตอร์ข้อสอบ ค่าพารามิเตอร์ผู้สอบแล้ว ยังจะทราบว่าตัวแปรคุณลักษณะ ของผู้สอบตัวแปรใด สามารถอธิบายความแปรปรวนในค่าความสามารถของผู้สอบได้ และจะนำไปสู่การศึกษาในรายละเอียดเชิงลึกของการพัฒนาการทดสอบ โดยประโยชน์จากสารสนเทศที่ได้จากกระบวนการวิเคราะห์ที่น่าเชื่อถือเพื่อการวางแผนกำหนดนโยบาย ในการพัฒนาคุณภาพการศึกษาให้เกิดประสิทธิภาพต่อไป

Muthen and Muthen (2010) ได้พัฒนาโปรแกรม Mplus สำหรับวิเคราะห์ข้อมูลด้วยสถิติวิเคราะห์ขั้นสูง ที่ให้ผลการวิเคราะห์ข้อมูลที่มีความถูกต้องมากกว่าสถิติวิเคราะห์แบบเดิม โปรแกรม Mplus Version ใหม่ล่าสุด คือ โปรแกรม Mplus Version 7 ได้รับการพัฒนาให้เป็นโปรแกรมที่ใช้งานง่ายและสะดวก และได้รับการปรับปรุงให้ดีขึ้น สามารถวิเคราะห์ข้อมูลได้หลายประเภท

วิธี MIMIC เป็นวิธีหนึ่งที่ใช้ในโปรแกรม Mplus สำหรับการวิเคราะห์คุณภาพ ของข้อสอบตามทฤษฎีการตอบสนองข้อสอบ (IRT) ซึ่งวิธี MIMIC เป็นโมเดลลิสเรลที่มีตัวแปรแฝงเป็นตัวแปรเดียว โดยที่ตัวแปรแฝงนั้นได้รับอิทธิพลจากตัวแปรภายนอกสังเกตได้หลายตัวแปรและส่งอิทธิพลไปยังตัวแปรภายในสังเกตได้หลายตัวแปร กล่าวอีกอย่างหนึ่งคือเป็นโมเดลลิสเรลของคุณลักษณะแฝงที่มีหลายสาเหตุและวัดได้จากตัวบ่งชี้หลายตัวลักษณะโมเดลจะเห็นว่าการวัดตัวแปรภายนอกสังเกตได้ต้องมีข้อตกลงข้างต้นว่าไม่มีความคลาดเคลื่อนในการวัด วิธี MIMIC นี้เป็นประโยชน์มากในการตรวจสอบความเป็นเอกมิติ (Unidimensionality) การวัดผลการศึกษาสามารถวิเคราะห์ค่า พารามิเตอร์คุณลักษณะข้อสอบและค่าความสามารถของผู้สอบไม่สามารถสังเกตโดยตรงจึงต้องประมาณจากการตอบข้อสอบ การประมาณค่าพารามิเตอร์ในทฤษฎีการตอบสนองข้อสอบ

วิธี MIMIC มีข้อดีหลายประการในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) (Muthen et al., 1991) โดยใช้หลักทฤษฎีการตอบสนองข้อสอบ (IRT) ประมาณค่าการทำหน้าที่ต่างกันของข้อสอบ (DIF) จากการศึกษาของ Finch (2005) เปรียบเทียบประสิทธิภาพของโมเดล MIMIC กับการทดสอบโดยวิธีแมนเทล-แฮนส์เซล และวิธี SIBTEST (Shealy & Stout, 1993) และวิธีการทดสอบ IRT Likelihood Ratio (Thissenet et al., 1986) กับความคลาดเคลื่อนประเภทที่ 1 และอำนาจการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ซึ่งแสดงให้เห็นว่าวิธี MIMIC มีค่าสูงขึ้นและความคลาดเคลื่อนประเภทที่ 1 มีค่าลดลง เมื่อจำนวนข้อสอบมีจำนวน 50 ข้อ นอกจากนี้วิธี MIMIC ยังสามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบ Uniform DIF ได้เพียงอย่างเดียวในการนำทฤษฎีการตอบสนองข้อสอบมาใช้จึงเลือกวิธีที่เหมาะสม วิธีประมาณค่าพารามิเตอร์ของข้อสอบและความสามารถของผู้เข้าสอบก็เป็นอีกกระบวนการหนึ่งที่จำเป็นต้องเลือกใช้ให้เหมาะสมกับสภาพการวัดแต่ละครั้ง สำหรับทฤษฎีการตอบสนองข้อสอบนั้น วิธีการประมาณค่าพารามิเตอร์ของข้อสอบและ

ความสามารถของผู้เข้าสอบ มีหนึ่งวิธีที่น่าสนใจ คือ วิธีของ BAYESIAN ซึ่งมีข้อดีคือ สามารถใช้งานได้ง่าย มีความยืดหยุ่นสูง สามารถแก้ปัญหาทั้งง่ายและซับซ้อนได้ดี มีการกำหนดการแจกแจงเริ่มต้นของค่าพารามิเตอร์ (Prior Distribution) ที่ใช้ในการกำหนดช่วงของค่าพารามิเตอร์ที่ต้องการประมาณค่าเป็นประโยชน์อย่างยิ่งในการวิเคราะห์ข้อมูล

จะเห็นได้ว่าการวิเคราะห์ข้อสอบ ด้วยทฤษฎีการตอบสนองข้อสอบสามารถให้ทั้งสารสนเทศที่เป็นค่าพารามิเตอร์ของข้อสอบเป็นรายข้อ (Item Parameter) พารามิเตอร์ของผู้สอบเป็นรายบุคคล (Person Parameter) รวมทั้งความสามารถในการตรวจสอบการทำหน้าที่ต่างกัน ของข้อสอบ ในการวิจัยนี้จึงมุ่งศึกษาการวิเคราะห์คุณภาพของข้อสอบตามทฤษฎีการตอบสนองข้อสอบ (IRT) 1-Parameter Logistic Measurement Model (1-PL) ทั้งนี้ด้วยข้อจำกัดของวิธีการวิเคราะห์ด้วยวิธี HGLM สามารถวิเคราะห์ข้อสอบได้เพียง 1-PL ส่วนวิธี MIMIC สามารถวิเคราะห์ได้ 2-PL และวิธี BAYESIAN สามารถวิเคราะห์ได้ 3-PL เพื่อให้สามารถเปรียบเทียบผลการประมาณค่าพารามิเตอร์ความยากของข้อสอบได้ ในการวิจัยนี้ผู้วิจัยศึกษาเพียง 1-PL จากนั้นจึงตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยดำเนินการวิเคราะห์ด้วยวิธี HGLM ประยุกต์ใช้โปรแกรม HLM วิธี MIMIC ประยุกต์ใช้โปรแกรม Mplus และวิธี BAYESIAN ประยุกต์ใช้โปรแกรม WinBUGS ซึ่งโปรแกรมดังกล่าวสามารถวิเคราะห์สถิติขั้นสูงได้ดี และเป็นที่นิยมของนักสถิติและนักวัดผล ในขณะนี้ โดยศึกษาจากการสอบวัดผลสัมฤทธิ์ทางการเรียนเพื่อประเมินคุณภาพการศึกษาระดับชาติ (NT) ปีการศึกษา 2555 ชั้นประถมศึกษาปีที่ 3 ในด้านความสามารถทั้ง 3 ด้านได้แก่ ด้านภาษา (Literacy Ability) ด้านคำนวณ (Numeracy Ability) และด้านเหตุผล (Reasoning Ability) เพื่อเป็นแนวทางสำหรับผู้เกี่ยวข้องในการออกข้อสอบระดับชาติ ในการนำไปปรับปรุงและพัฒนาข้อสอบต่อไป

จากเหตุผลที่กล่าวมา ผู้วิจัยจึงสนใจศึกษาเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2555 ที่ทดสอบวัดความสามารถทั้ง 3 ด้าน ได้แก่ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ว่ามีข้อสอบข้อใดบ้างที่ทำหน้าที่ต่างกัน โดยประยุกต์ ใช้วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ 3 วิธี คือ วิธีการตรวจสอบการทำหน้าที่ต่างกันด้วยโมเดลสมการเชิงเส้นตรงระดับลดหลั่น (HGLM) วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยโมเดลสมการโครงสร้างมิมิค (MIMIC) และวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี BAYESIAN

วัตถุประสงค์ของการวิจัย

1. เพื่อวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้าน ได้แก่ ด้านภาษา ด้านคำนวณ และด้านเหตุผล โดยใช้หลักการของทฤษฎีการตอบสนองข้อสอบ (IRT) แบบ 3พารามิเตอร์

2. เพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้าน ได้แก่ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN

3. เพื่อเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้าน ได้แก่ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN

กรอบแนวคิดการวิจัย

จากการศึกษา แนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ จากการศึกษางานของ สุธาทิพย์ ตรีสิน และปิยะทิพย์ ประดุจพรม (2560) ได้เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ชั้นประถมศึกษาปีที่ 3 ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR ผลการศึกษาปรากฏว่า วิธี HGLM ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวนมากที่สุด คิดเป็นร้อยละ 69 ของข้อสอบทั้งฉบับ รองลงมาคือ วิธี IRT-LR ร้อยละ 54 และ วิธี MIMIC ร้อยละ 16 ตามลำดับ และงานวิจัยของ Acar and Kelecioglu (2010) ได้เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ในแบบทดสอบด้านสังคมศาสตร์และด้านวิทยาศาสตร์ ด้วยวิธี HGLM วิธี LR และวิธี IRT-LR ปรากฏว่า วิธี HGLM ตรวจพบ DIF ได้มากที่สุดในแบบทดสอบทั้งสองด้าน ส่วนวิธี LR และวิธี IRT-LR ตรวจพบ DIF ใกล้เคียงกัน สามารถเขียนเป็นกรอบแนวคิดในการวิจัย ดังภาพที่ 1-1

ภาพที่ 1-1 กรอบแนวคิดการวิจัย

สมมติฐานของการวิจัย

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ระดับชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้าน ได้แก่ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC วิธี BAYESIAN จากการศึกษางานวิจัย สุธาทิพย์ ตรีสิน และปิยะทิพย์ ประดุจพรม (2560) เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ 3 ด้าน ได้แก่ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2556 ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้ง 3 ด้าน ผลการศึกษาปรากฏว่า วิธี HGLM สามารถตรวจพบ DIF ได้มากที่สุด คิดเป็นร้อยละ 69 รองลงมาคือ วิธี IRT-LR คิดเป็นร้อยละ 54 และวิธีตรวจพบ DIF ได้น้อยที่สุด คิดเป็นร้อยละ 16 และงานวิจัยของ Ong, Lu, Lee, and Cohen (2015) ได้ศึกษาเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ระหว่างวิธี HGLM วิธี MIMIC และวิธี IRT ผลการศึกษา

ปรากฏว่า วิธี HGLM เป็นวิธีที่สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบและพบจำนวนข้อสอบที่ทำหน้าที่ต่างกันได้มากที่สุด ผู้วิจัยจึงตั้งสมมติฐานการวิจัย ดังนี้

1. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา วิธี HGLM ตรวจพบข้อสอบทำหน้าที่ต่างกัน (DIF) มากกว่าวิธี MIMIC

2. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา วิธี MIMIC ตรวจพบข้อสอบทำหน้าที่ต่างกัน (DIF) มากกว่าวิธี BAYESIAN

3. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา วิธี HGLM ตรวจพบข้อสอบทำหน้าที่ต่างกัน (DIF) มากกว่าวิธี BAYESIAN

4. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านคำนวณ วิธี HGLM ตรวจพบข้อสอบทำหน้าที่ต่างกัน (DIF) มากกว่าวิธี MIMIC

5. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านคำนวณ วิธี MIMIC ตรวจพบข้อสอบทำหน้าที่ต่างกัน (DIF) มากกว่าวิธี BAYESIAN

6. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านคำนวณ วิธี HGLM ตรวจพบข้อสอบทำหน้าที่ต่างกัน (DIF) มากกว่าวิธี BAYESIAN

7. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านเหตุผล วิธี HGLM ตรวจพบข้อสอบทำหน้าที่ต่างกัน (DIF) มากกว่าวิธี MIMIC

8. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านเหตุผล วิธี MIMIC ตรวจพบข้อสอบทำหน้าที่ต่างกัน (DIF) มากกว่าวิธี BAYESIAN

9. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านเหตุผล วิธี HGLM ตรวจพบข้อสอบทำหน้าที่ต่างกัน (DIF) มากกว่าวิธี BAYESIAN

ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

1. สำนักทดสอบทางการศึกษา (สพฐ.) สามารถนำผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เป็นข้อมูลสารสนเทศในการปรับปรุง พัฒนาคุณภาพของแบบทดสอบให้ดียิ่งขึ้น เหมาะกับผู้เข้าสอบ

2. วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ วิธี HGLM เป็นวิธีที่ตรวจพบ DIF ได้ดีกว่าวิธี MIMIC และวิธี BAYESIAN ในด้านภาษาและเหตุผล ส่วนด้านคำนวณวิธี MIMIC เป็นวิธีที่ตรวจพบ DIF ได้ดีกว่า วิธี HGLM และวิธี BAYESIAN

3. นักวัดผลการศึกษา สามารถนำผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ไปใช้ ในการพัฒนาคุณภาพของแบบทดสอบและเลือกข้อสอบที่มีคุณภาพมาใช้เพื่อให้เกิดประสิทธิภาพมากยิ่งขึ้น

ขอบเขตของการวิจัย

การวิจัยนี้ใช้ข้อมูลที่เป็นผลการตอบแบบทดสอบวัดความสามารถพื้นฐานของผู้เรียนระดับชาติ (National Test: NT) ของนักเรียนชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2555 ซึ่งประเมินความสามารถทั้ง 3 ด้าน ได้แก่ 1) ด้านภาษา 2) ด้านคำนวณ และ 3) ด้านเหตุผล ซึ่งเป็นข้อมูลทุติยภูมิ (Secondary

Data) จากสำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน (สพฐ.) กระทรวงศึกษาธิการ

ประชากรที่เป็นนักเรียนชั้นประถมศึกษาปีที่ 3 ที่เข้าทดสอบความสามารถพื้นฐานของผู้เรียนระดับชาติ (National Test: NT) ปีการศึกษา 2555 ซึ่งประเมินความสามารถ ทั้ง 3 ด้าน ได้แก่ 1) ด้านภาษา 2) ด้านคำนวณ และ 3) ด้านเหตุผล จำนวน 496,196 คน จากโรงเรียนทั้งหมด 28,204 โรง กลุ่มตัวอย่างเป็นนักเรียนชั้นประถมศึกษาปีที่ 3 จังหวัดสระแก้ว จำนวน 2,000 คน ที่เข้าทดสอบวัดความสามารถพื้นฐานของผู้เรียนระดับชาติ

1. ตัวแปรที่ศึกษา

1.1 ตัวแปรต้น เป็นวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจำนวน 3 วิธี ได้แก่

1.1.1 วิธี HGLM โดยใช้โปรแกรม HLM

1.1.2 วิธี MIMIC โดยใช้โปรแกรม Mplus

1.1.3 วิธี BAYESIAN โดยใช้โปรแกรม WinBUGS

1.2 ตัวแปรตาม เป็นผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยพิจารณาจากจำนวนข้อสอบที่ตรวจพบการทำหน้าที่ต่างกัน

นิยามศัพท์เฉพาะ

การสอบวัดความสามารถพื้นฐานของผู้เรียนระดับชาติ (National Test: NT) หมายถึง การทดสอบระดับชาติเป็นการประเมินคุณภาพผู้เรียนตามมาตรฐานและตัวชี้วัดของหลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน ที่ดำเนินการโดยหน่วยงานภายในประเทศ 2 หน่วยงาน คือ สถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน) และสำนักงานคณะกรรมการการศึกษา ขั้นพื้นฐาน ในการวิจัยครั้งนี้ คือ การทดสอบที่ดำเนินการโดย สำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน (สพฐ.) กระทรวงศึกษาธิการ

แบบทดสอบทางการศึกษาระดับชาติ (National Test: NT) หมายถึง ข้อสอบที่ใช้วัดผลการจัดการเรียนรู้ตามหลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน พ.ศ.2551 ในการวิจัยครั้งนี้ใช้แบบทดสอบของการทดสอบระดับชาติ (NT) ชั้นประถมศึกษาปีที่ 3 เป็นแบบทดสอบวัดด้านด้านภาษา (Literacy Ability) ด้านคำนวณ (Numeracy Ability) และด้านเหตุผล (Reasoning Ability)

การทำหน้าที่ต่างกันของข้อสอบ (Differential Item Function: DIF) หมายถึง ข้อสอบที่มีคุณสมบัติในการวัดความสามารถที่มุ่งวัดสิ่งเดียวกัน แต่ผู้สอบมีคุณลักษณะบางประการแตกต่างกัน จึงมีโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกัน จึงก่อให้เกิดความลำเอียงของข้อสอบนำไปสู่การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเพื่อพัฒนาให้ข้อสอบมีคุณภาพต่อไป

โมเดลเชิงเส้นตรงทั่วไประดับลดหลั่น (Hierarchical Generalized Linear Model: HGLM) หมายถึง รูปแบบหรือลักษณะการวิเคราะห์ข้อมูลเชิงเส้นทั่วไปที่มีการประยุกต์ปรับให้ดำเนินการวิเคราะห์ข้อมูลร่วมกับโมเดลการวิเคราะห์ข้อมูลแบบอื่นๆ และการวิเคราะห์พหุระดับที่มีข้อมูลสอดแทรกเป็นระดับลดหลั่นได้ โดยในระดับการวิเคราะห์ที่ 1 เป็นการวิเคราะห์ตามโมเดลเชิงเส้นทั่วไป (Generalized Linear Model: GLM) แล้วใช้ฟังก์ชันโยง (Link Function) ที่เป็นฟังก์ชั่นโยงแบบโลจิท (Logit Link Function) ในการปรับค่าเฉลี่ยของระดับการวิเคราะห์ที่ 1 นำมาสู่การวิเคราะห์ในระดับต่อไปได้โดยใช้

โมเดลการวิเคราะห์พหุระดับด้วยโมเดลเชิงเส้นตรงระดับลดหลั่น (HLM) โดยการวิเคราะห์ระดับที่ 1 ตัวแปรตามจึงเป็น Log-odds ของความน่าจะเป็นในการตอบข้อสอบได้ถูกต้อง

วิธี HGLM หมายถึง วิธีการทางสถิติที่นำมาใช้ในการวิเคราะห์ข้อมูลเชิงปริมาณในกรณีที่ข้อมูลมีลักษณะลดหลั่นตั้งแต่ 2 ระดับขึ้นไป ตัวแปรตามเป็นตัวแปรต่อเนื่องมีความสัมพันธ์เชิงเส้นตรงกับตัวแปรต้น มีการแจกแจงไม่เป็นโค้งปกติและไม่สามารถแปลงข้อมูลได้ เช่น ตัวแปรตามเป็นตัวแปรกลุ่ม ดังนั้น ข้อมูลที่นำมาวิเคราะห์จึงควรเป็นข้อมูลที่มีลักษณะเป็นตัวเลขที่ได้มาจากเครื่องมือที่มีคุณภาพ เช่น แบบสอบถาม แบบทดทดสอบ เป็นต้น เพื่อแก้ปัญหาความลำเอียงในการสรุปข้ามระดับ (Aggregation Bias) ปัญหาในการคำนวณค่าความคลาดเคลื่อนมาตรฐาน (Misestimated Standard Bias) และปัญหาความผันแปรของสัมประสิทธิ์การถดถอย (Heterogeneity of Regression)

วิธี MIMIC หมายถึง โมเดลสมการโครงสร้างที่มีตัวแปรแฝงเพียงตัวแปรเดียว โดยที่ตัวแปรแฝงนั้นได้รับอิทธิพลจากตัวแปรภายนอกสังเกตได้หลายตัวแปรและส่งอิทธิพลไปยังตัวแปรภายในสังเกตได้หลายตัวแปร ใช้โปรแกรม Mplus

วิธี BAYESIAN หมายถึง วิธีประมาณค่าพารามิเตอร์ของข้อสอบ และความสามารถของผู้เข้าสอบ ด้วยวิธี BAYESIAN ที่มีการกำหนด Prior Distribution ของค่าความยากของข้อสอบไว้ ใช้โปรแกรม WinBUGS

ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) หมายถึง วิธีการวัดที่อธิบายความสัมพันธ์ระหว่างคุณลักษณะภายในหรือความสามารถที่มีอยู่ภายในตัวบุคคลกับพฤติกรรมการตอบสนองข้อสอบของบุคคลนั้นว่ามีโอกาสตอบข้อสอบถูกมากน้อยเพียงไร โดยอธิบายความสัมพันธ์ดังกล่าวในรูปฟังก์ชันคณิตศาสตร์

ค่าอำนาจจำแนกของข้อสอบ (a) หมายถึง ค่าที่เป็นสัดส่วนโดยตรงกับความชันของโค้งคุณลักษณะของข้อสอบ ณ จุดเปลี่ยนโค้ง มีตั้งแต่ $-\alpha$ ถึง $+\alpha$ สำหรับค่าอำนาจจำแนกของข้อสอบที่เป็นลบ (-) แสดงว่า ข้อสอบไม่ดี จำแนกไม่ได้ ต้องตัดข้อสอบนั้นทิ้ง ค่าอำนาจจำแนกของข้อสอบที่เป็นศูนย์ (0) แสดงว่า ข้อสอบไม่มีค่าอำนาจจำแนก และค่าอำนาจจำแนกที่เป็นบวก (+) แสดงว่าข้อสอบดี จำแนกได้ สำหรับข้อสอบที่คัดเลือกไว้จะมีค่าอำนาจจำแนกของข้อสอบ (a) มีค่าตั้งแต่ 0.5 ถึง 2.5

ค่าความยากของข้อสอบ (b) หมายถึง ค่าที่แสดงถึงระดับความสามารถของผู้สอบที่จุดเปลี่ยนโค้งคุณลักษณะของข้อสอบที่มีความชันมากที่สุด มีค่าระหว่าง $-\alpha$ ถึง $+\alpha$ ซึ่งค่าความยากของข้อสอบเท่ากับ -2.5 แสดงว่าข้อสอบง่ายมาก และค่าความยากของข้อสอบเท่ากับ 2.5 แสดงว่าข้อสอบนั้นยากมาก สำหรับข้อสอบที่คัดเลือกไว้จะมีค่าความยากของข้อสอบตั้งแต่ -2.5 ถึง +2.5

ค่าการเดาของข้อสอบ (c) หมายถึง ความน่าจะเป็นของผู้เข้าสอบที่ไม่มีความสามารถในการตอบข้อสอบนั้นได้ถูกต้อง เป็นค่าที่แสดงถึงโอกาสการตอบข้อสอบถูก โดยไม่มีความรู้ในเรื่องนั้น ๆ มีค่าตั้งแต่ 0 ถึง 1 สำหรับข้อสอบที่คัดเลือกไว้จะมีค่าการเดาของข้อสอบไม่เกิน 0.3

เพศ (Gender) หมายถึง เพศของนักเรียนชั้นประถมศึกษาปีที่ 3 โดยจำแนกเป็นเพศหญิงและเพศชาย ที่เป็นกลุ่มตัวอย่างที่ตอบแบบทดสอบ NT ปีการศึกษา 2555

กลุ่มอ้างอิง (Reference Group: R) หมายถึง กลุ่มผู้สอบที่คาดว่าจะได้รับประโยชน์จาก

การตอบข้อสอบที่ทำหน้าที่ต่างกัน คือ เป็นกลุ่มที่มีความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องสูงกว่าผู้สอบอีกกลุ่มหนึ่งทั้งๆที่มีความสามารถเท่ากัน

กลุ่มเปรียบเทียบ (Focal Group: F) หมายถึง กลุ่มผู้สอบที่คาดว่าจะเสียประโยชน์จากการตอบข้อสอบที่ทำหน้าที่ต่างกัน คือเป็นกลุ่มที่มีความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องต่ำกว่าผู้สอบอีกกลุ่มหนึ่งทั้ง ๆ ที่มีความสามารถเท่ากัน

นักเรียน (Students) หมายถึง ผู้ที่กำลังศึกษาอยู่ในระดับชั้นประถมศึกษาปีที่ 3 ในปีการศึกษา 2555

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

การวิจัยนี้เพื่อเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถม ศึกษาปีที่ 3 ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN ซึ่งแบ่งการนำเสนอเป็น 6 ตอน ดังนี้

ตอนที่ 1 การสอบวัดความสามารถพื้นฐานของผู้เรียนระดับชาติ (National Test: NT) และงานวิจัยที่เกี่ยวข้อง

ตอนที่ 2 ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) และงานวิจัยที่เกี่ยวข้อง

ตอนที่ 3 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) และงานวิจัยที่เกี่ยวข้อง

ตอนที่ 4 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี HGLM และงานวิจัยที่เกี่ยวข้อง

ตอนที่ 5 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี MIMIC และงานวิจัยที่เกี่ยวข้อง

ตอนที่ 6 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี BAYESIAN และงานวิจัยที่เกี่ยวข้อง

ตอนที่ 1 การสอบวัดความสามารถพื้นฐานผู้เรียนระดับชาติ (National Test: NT) และงานวิจัยที่เกี่ยวข้อง

ความเป็นมาของการสอบวัดความสามารถพื้นฐานผู้เรียนระดับชาติ

การทดสอบระดับชาติเริ่มมาตั้งแต่ปี พ.ศ. 2478 ถึง พ.ศ. 2520 ถือเป็นการกระจายอำนาจให้โรงเรียนโดยมีระบบ Accountability กล่าวคือ ผู้จบชั้นประโยคต้องสอบได้คะแนนทดสอบระดับชาติไม่ต่ำกว่าร้อยละ 50 แต่เมื่อปีการศึกษา 2521 จนถึงปัจจุบัน ได้มอบให้โรงเรียนแต่ละโรงทำหน้าที่ในการทดสอบตัวประโยคแทนการสอบกลางระดับชาติ โดยให้คณะกรรมการการศึกษาจังหวัดพิจารณาอนุญาตให้โรงเรียนที่อยู่ขั้นมาตรฐานให้มีการเลื่อนขั้นอัตโนมัติ ซึ่งเป็นการกระจายอำนาจแบบไม่มี Accountability และไม่สามารถเทียบเคียงคุณภาพผู้เรียนของแต่ละสถานศึกษาได้ เนื่องจากแต่ละสถานศึกษาใช้การทดสอบต่างกัน (สัมพันธ์ พันธุ์พฤกษ์ และคณะ 2557, หน้า 14) ในปัจจุบันการทดสอบระดับชาติ เป็นการประเมินคุณภาพผู้เรียนตามมาตรฐานและตัวชี้วัดตามหลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน ที่ดำเนินการโดยหน่วยงานภายในประเทศ 2 หน่วยงาน คือ สถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน) และ สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน ได้แก่ สถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน) ดำเนินการทดสอบระดับชาติขั้นพื้นฐาน (Ordinary National Education Test: O-NET) วัดความรู้ ความคิดรวบยอด ตามมาตรฐานการเรียนรู้ในระดับประถมศึกษา (ป.4-6) ระดับมัธยมศึกษาตอนต้น (ม.1-3) และระดับมัธยมศึกษาตอนปลาย (ม.4-6) ของหลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน พ.ศ. 2551 โดยจัดสอบผู้เรียนทุกคนในชั้น ป.6 ม.3 และ ม.6 เพื่อนำผลการทดสอบไปใช้เป็นองค์ประกอบหนึ่งของการสำเร็จการศึกษาตามหลักสูตร ใช้เป็นองค์ประกอบ

ในการคัดเลือกเข้าศึกษาต่อ ในระดับมัธยมศึกษาตอนต้น มัธยมศึกษาตอนปลาย และสถาบันอุดมศึกษา ใช้ในการประกันคุณภาพการศึกษา ใช้ในการปรับปรุงการเรียนการสอนและการบริหารเพื่อยกระดับผลสัมฤทธิ์ทางการเรียนของนักเรียนตลอดจนใช้เป็นข้อมูลและสารสนเทศในการเทียบเคียงคุณภาพการศึกษาในระดับต่าง ๆ ตามพระราชบัญญัติการศึกษาแห่งชาติ พ.ศ. 2542 และที่แก้ไขเพิ่มเติม (ฉบับที่ 2) พ.ศ. 2545 ที่กำหนดไว้ในหมวด 4 มาตรา 26 เกี่ยวกับแนวทางการวัดและประเมินผลการเรียนรู้ซึ่งเป็นองค์ประกอบหนึ่งที่มีความสำคัญต่อกระบวนการพัฒนาผู้เรียน การปรับปรุงพัฒนาคุณภาพการศึกษาและการจัดการเรียนรู้ รวมทั้งใช้ผลการทดสอบในวัตถุประสงค์อื่นๆ และอีกหน่วยงาน คือ สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน (สพฐ.) กระทรวงศึกษาธิการ ดำเนินการประเมินคุณภาพผู้เรียนตามมาตรฐานและตัวชี้วัดตามหลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน เรียกว่า การทดสอบระดับชาติ (National Test: NT) โดยจัดสอบกับนักเรียนชั้นประถมศึกษาปีที่ 3 ในด้านความสามารถด้านภาษา (Literacy Ability) ความสามารถด้านคำนวณ (Numeracy Ability) และความสามารถด้านเหตุผล (Reasoning Ability) ส่วนการทดสอบระดับนานาชาติ เป็นการประเมินเพื่อเทียบเคียงคุณภาพของผู้เรียนระดับนานาชาติ ได้แก่การประเมินของ PISA (Programe for International Student Assessment) ประกอบด้วยการประเมิน 3 ด้าน ได้แก่ Reading Literacy, Mathematical Literacy และ Scientific Literacy โดยประเมินผู้เรียนที่มีอายุ 15 ปี และการประเมินของ TIMSS (Trends in International Mathematics and Science Study) ประกอบด้วยการประเมิน 2 วิชา คือ คณิตศาสตร์ (Mathematics Assessment) และวิทยาศาสตร์ (Science Assessment) โดยประเมินผู้เรียนในชั้นมัธยมศึกษาปีที่ 2

ความหมายของการสอบวัดความสามารถพื้นฐานผู้เรียนระดับชาติ

นักวิชาการได้ให้ความหมายของการการสอบวัดความสามารถพื้นฐานผู้เรียนระดับชาติไว้หลายท่าน ดังนี้

บุญชม ชลัษเฐียร (2550, หน้า 28-29) ได้ให้ความหมายว่า การวัดผลสัมฤทธิ์ระดับชาติ คือ กระบวนการดำเนินงานประเมินคุณภาพการศึกษาขั้นพื้นฐาน ซึ่งสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน กระทรวงศึกษาธิการกำหนดให้มีขึ้นเพื่อดูสัมฤทธิ์ผลของผู้เรียนอันเกิดจากการจัดการศึกษาตามหลักสูตรการศึกษา

จากความหมายต่าง ๆ สรุปได้ว่า การวัดผลสัมฤทธิ์ระดับชาติ คือ คะแนนที่ได้จากการสอบวัดผลสัมฤทธิ์ทางการเรียนชั้นประถมศึกษาปีที่ 3 ประถมศึกษาปีที่ 6 มัธยมศึกษาปีที่ 3 และมัธยมศึกษาปีที่ 6 โดยกรมวิชาการ กระทรวงศึกษาธิการ เพื่อนำไปประเมินคุณภาพการศึกษา ตามมาตรฐานการศึกษา ซึ่งในการวิจัยครั้งนี้ขอบเขตของการวิจัย คือผลสัมฤทธิ์ระดับชาติ ชั้นประถมศึกษาปีที่ 3 เท่านั้น

ความสำคัญของการทดสอบความสามารถพื้นฐานของผู้เรียนระดับชาติ

การดำเนินการประเมินคุณภาพการศึกษาขั้นพื้นฐาน ทำหน้าที่เป็นกลไกการควบคุมคุณภาพการศึกษาคุณภาพผู้เรียนเป็นเครื่องมือกระตุ้น สร้างแรงจูงใจในผลสัมฤทธิ์ผลักดันคุณภาพของงานให้เกิดขึ้นทั้งกับครูผู้สอนและตัวผู้เรียนอีกทางหนึ่งด้วย ดังนั้นผลที่เกิดขึ้นกับผู้เรียนจากการจัดการศึกษาที่มีคุณภาพควรสะท้อนให้เห็นได้จากคะแนนการทดสอบด้วยแบบทดสอบ หรือเครื่องมือวัดประเภทต่าง ๆ ที่สูงขึ้นหรือมีพัฒนาการที่ดีขึ้น ดังนั้น การประเมินคุณภาพการศึกษาระดับชาติ หรือการทดสอบวัด

ผลสัมฤทธิ์ทางการเรียนของผู้เรียน จึงเป็นเครื่องมือสำคัญยิ่งในการสร้างความมั่นใจเกี่ยวกับคุณภาพการศึกษา ทั้งนี้เพราะการประเมินระดับชาติจะทำหน้าที่เป็นมาตรฐานกลาง ที่ใช้เทียบเคียงผลที่เกิดขึ้น อีกทั้งยังทำให้ได้ข้อมูลซึ่งเป็นตัวบ่งชี้คุณภาพการศึกษาของชาติในภาพรวม ทำให้ได้ข้อมูลประกอบการตัดสินใจเชิงนโยบาย และการวางแผนการพัฒนาคุณภาพการศึกษาของกระทรวงศึกษาธิการโดยที่ข้อมูล ผลการประเมินถูกนำเสนอหลายระดับ ได้แก่ ระดับผู้เรียนรายบุคคล รายสถานศึกษา เขตพื้นที่การศึกษา ระดับประเทศ และจำแนกสังกัด กรณีมีสถานศึกษาสังกัดอื่น อาทิ โรงเรียนสาธิต เทศบาล เอกชน หรือ กรุงเทพมหานคร ฯ เข้าร่วมการประเมิน ดังนั้นผู้เกี่ยวข้องหรือผู้มีส่วนได้ส่วนเสีย สามารถใช้ประโยชน์จากข้อมูลที่รายงานแต่ละระดับ

สถานศึกษาสามารถนำข้อมูลไปใช้ในการแสดงระดับผลสัมฤทธิ์ทางการเรียนของผู้เรียนเป็นรายบุคคลอย่างต่อเนื่อง เป็นข้อมูลพื้นฐานเพื่อประเมินนักเรียนและโรงเรียนใช้ในการปรับปรุงพัฒนาการเรียนการสอน วางแผนพัฒนาคุณภาพการจัดการศึกษาของสถานศึกษา แสดงศักยภาพของสถานศึกษา รายงานต่อสาธารณะชน ชุมชน คณะกรรมการสถานศึกษา เพื่อแสดงความรับผิดชอบในการจัดการศึกษา เพื่อปรับปรุงพัฒนาหลักสูตร การเรียนการสอน รวมทั้งใช้เป็นข้อมูลสร้างความเข้าใจความตระหนักในการมีส่วนร่วมพัฒนาคุณภาพของผู้เกี่ยวข้องสำหรับข้อมูลสำนักงานเขตพื้นที่การศึกษา ระดับประเทศ สำนักทดสอบทางการศึกษา ได้ให้ความสำคัญกับการทดสอบวัดผลสัมฤทธิ์ระดับชาติ ดังนี้

1. ทำให้เปรียบเทียบผลการประเมินคุณภาพระหว่างชั้นเรียน ระดับสถานศึกษา ระดับเขตพื้นที่การศึกษา และระดับชาติ ตลอดจนการประเมินภายนอกอย่างสมเหตุสมผล

2. เพื่อกำกับติดตามและควบคุมคุณภาพการจัดการศึกษาขั้นพื้นฐานของประเทศในช่วงชั้นที่ 1 (ประถมศึกษาปีที่ 3) และช่วงชั้นที่ 3 (มัธยมศึกษาปีที่ 3) เพื่อให้เกิดการพัฒนาอย่างต่อเนื่อง ซึ่งเป็นส่วนหนึ่งของการประกันคุณภาพ

3. เพื่อให้ได้ข้อมูลย้อนกลับ สำหรับกระบวนการตัดสินใจ และกำหนดแผนพัฒนาคุณภาพการจัดการศึกษาขั้นพื้นฐานของประเทศ เขตพื้นที่การศึกษาและระดับสถานศึกษา

4. สามารถประเมินทั้งผลสัมฤทธิ์ทางวิชาการตามหลักสูตรและความถนัดทางการเรียน

5. ส่งเสริมและกระตุ้นให้สถานศึกษาให้ความสนใจอย่างจริงจังในการพัฒนาผลสัมฤทธิ์ ที่สำคัญของหลักสูตร

6. สามารถใช้ผลการประเมินให้เป็นประโยชน์ทั้งในระดับผู้เรียน ระดับชั้นเรียน ระดับสถานศึกษา ระดับเขตพื้นที่การศึกษา และระดับชาติ

7. สร้างแรงจูงใจกระตุ้นและทำนายให้ผู้เรียนทุกคนตั้งใจใฝ่หาสัมฤทธิผลทางการเรียน

8. เพื่อเป็นข้อมูลสร้างความมั่นใจเกี่ยวกับคุณภาพผู้เรียนผู้เกี่ยวข้องทั้งภายในและภายนอกสถานศึกษา

การสอบวัดความสามารถพื้นฐานผู้เรียนระดับชาติ (National Test: NT)

สำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน จัดขึ้นตามกฎกระทรวง แบ่งส่วนราชการสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน กระทรวงศึกษาธิการ พ.ศ. 2546 ตามความในมาตรา 8 และมาตรา 11 แห่งพระราชบัญญัติระเบียบบริหารราชการกระทรวง พ.ศ. 2546 รัฐมนตรีว่าการกระทรวงศึกษาธิการออกกฎกระทรวงไว้ให้สำนักงานคณะกรรมการการศึกษา

ขั้นพื้นฐาน มีภารกิจเกี่ยวกับการจัดและการส่งเสริมการศึกษาขั้นพื้นฐาน ให้แบ่งส่วนราชการสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน ออกเป็น 10 สำนัก ซึ่งหนึ่งในนั้น มีสำนักทดสอบทางการศึกษาอยู่ด้วย

1. อำนาจหน้าที่ของสำนักทดสอบทางการศึกษา

1.1 การศึกษา วิจัย และพัฒนาการวัดและประเมินผลทางการศึกษา รวมทั้งจัดระบบวิธีการสอบ และพัฒนาเครื่องมือวัดมาตรฐานสำหรับการประเมินผลการจัดการศึกษา และการทดสอบทางการศึกษาขั้นพื้นฐาน

1.2 ดำเนินการสอบวัดความรู้ ความสามารถ คุณลักษณะระดับต่าง ๆ ให้กับนักเรียนและประชาชนทั่วไป

1.3 พัฒนาและส่งเสริมวิชาการด้านการทดสอบและประเมินผลทางการศึกษา รวมถึงการพัฒนาบุคลากรด้านการทดสอบและประเมินผล

1.4 ดำเนินการเกี่ยวกับระบบข้อมูลและทะเบียนประวัติผู้สำเร็จการศึกษาและจัดทำระบบการเทียบโอนผลการศึกษา รวมทั้งประสานความร่วมมือด้านการทดสอบทางการศึกษาทั้งในระดับชาติและระดับนานาชาติ

1.5 ปฏิบัติงานร่วมกับหรือสนับสนุนการปฏิบัติงานของหน่วยงานอื่นที่เกี่ยวข้อง หรือที่ได้รับมอบหมาย

2. วิสัยทัศน์องค์กร

สำนักทดสอบทางการศึกษาเป็นองค์กรวิชาชีพชั้นสูงที่เป็นที่ยอมรับทั้งในระดับชาติและระดับนานาชาติในความเป็นผู้นำด้านการวิจัยพัฒนาและการให้บริการทางการทดสอบ และการประเมินผลทางการศึกษา ผลผลิตและบริการทางวิชาการของสำนักทดสอบทางการศึกษามีมาตรฐานระดับสากลในด้านความเที่ยงตรง และความน่าเชื่อถือ ความเป็นธรรม และความเป็นเลิศทางเทคโนโลยี

3. พันธกิจ

สำนักทดสอบทางการศึกษา มุ่งมั่นในการช่วยพัฒนาคุณภาพและมาตรฐานการศึกษาของชาติ ด้วยการศึกษาวิจัยพัฒนา และให้บริการด้านระบบ วิธีการ และเครื่องมือวัดและประเมินผลทางการศึกษาที่มีประสิทธิภาพ เที่ยงตรงน่าเชื่อถือ และเป็นธรรมเพื่อนำผล และข้อมูลการประเมินไปใช้ในการวางแผนปรับปรุงคุณภาพการศึกษาอย่างต่อเนื่องทั้งในระดับชาติ ระดับท้องถิ่น ระดับสถานศึกษา และระดับผู้เรียนเป็นรายบุคคล

4. แนวดำเนินการจัดสอบวัดความสามารถพื้นฐานของผู้เรียนระดับชาติ (National Test: NT)

สำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาแห่งชาติ มีภาระหน้าที่สำคัญอย่างหนึ่ง คือ จัดสอบวัดผลระดับชาติให้กับนักเรียนทุกคน ประเมินเพื่อศึกษาและพัฒนาผลสัมฤทธิ์ผู้เรียนให้เข้าสู่มาตรฐาน เพื่อเป็นหลักประกันการเรียนรู้ (Accountability) และเตรียมการให้ผู้เรียนมีความพร้อมสำหรับรองรับการประเมิน ทั้งการทดสอบระดับชาติ และนานาชาติโดยจะมุ่งประเมินให้ทัดเทียมกับนานาชาติ ที่มีกี่ประเมินที่หลากหลาย มุ่งเน้นคุณภาพของนักเรียนโดยพิจารณาจากความสามารถพื้นฐานหลักของนักเรียน และเพื่อโรงเรียนจะได้นำผลประเมินที่ได้ไปปรับปรุงการเรียนการสอน หรือนำไปใช้ในวัตถุประสงค์อื่น ๆ

การทดสอบความสามารถพื้นฐานของผู้เรียนระดับชาติ (National Test: NT) หมายถึง การประเมินคุณภาพผู้เรียนตามมาตรฐานและตัวชี้วัดของหลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน วัดด้านภาษา (Literacy Ability) ด้านคำนวณ (Numeracy Ability) และด้านเหตุผล (Reasoning Ability) ของนักเรียนชั้นประถมศึกษาปีที่ 3

ความสามารถด้านภาษา (Literacy Ability) เป็นความสามารถพื้นฐานที่สำคัญและจำเป็นต่อการเรียนรู้ที่สำนักงานคณะกรรมการการศึกษาขั้นฐานให้ความสำคัญเป็นอันดับแรกในการที่จะใช้ประเมินคุณภาพการศึกษาขั้นพื้นฐานโดยกำหนดนิยามไว้ดังนี้

1. ความสามารถด้านภาษา (Literacy Ability) หมายถึง ความสามารถในการอ่าน การฟัง การดู การพูด เพื่อรู้ เข้าใจ วิเคราะห์ สรุปสาระสำคัญ ประเมินสิ่งที่อ่าน ฟัง ดู จากสื่อประเภทต่าง ๆ และสื่อสารด้วยการพูด การเขียน ได้ถูกต้องตามหลักการใช้ภาษาอย่างสร้างสรรค์ เพื่อการนำไปใช้ในชีวิตประจำวัน การอยู่ร่วมกันในสังคมและการศึกษาตลอดชีวิต

2. รู้ หมายถึง สามารถบอกความหมาย เรื่องราว ข้อเท็จจริง และเหตุการณ์ต่าง ๆ

3. เข้าใจ หมายถึง สามารถแปลความ ตีความ ขยายความและอ้างอิง

4. วิเคราะห์ หมายถึง สามารถแยกแยะโครงสร้าง เรื่องราว ข้อเท็จจริง ข้อคิดเห็น เหตุผล และคุณค่า

5. สรุปสาระสำคัญ หมายถึง สามารถสรุปใจความสำคัญของเรื่องได้อย่างครอบคลุม

6. ประเมิน หมายถึง สามารถตัดสินความถูกต้อง ความชัดเจน ความเหมาะสม คุณค่าอย่างมีหลักเกณฑ์

7. สื่อประเภทต่าง ๆ หมายถึง สิ่งที่นำเสนอเรื่องราวและข้อมูลความรู้ต่าง ๆ ทั้งที่เป็นสื่อสิ่งพิมพ์ สื่ออิเล็กทรอนิกส์ และสื่อของจริง

8. สื่อสาร หมายถึง สามารถถ่ายทอดความรู้ ความเข้าใจ และความคิดจากการอ่าน ฟัง และดู โดยการพูดหรือเขียนอธิบาย วิเคราะห์ สรุปหรือประเมิน

9. สร้างสรรค์ หมายถึง สามารถสื่อสารความรู้ ความเข้าใจ เรื่องราว ทัศนะ และความคิดที่แปลกใหม่จากการอ่าน การฟัง และการดู เป็นคำพูด การเขียน หรือการกระทำได้อย่างหลากหลายและมีประโยชน์เพิ่มมากขึ้น

10. การนำไปใช้ในชีวิตประจำวัน การอยู่ร่วมกันในสังคมและการศึกษาตลอดชีวิต หมายถึง ความสามารถในการนำความรู้ ความเข้าใจ การวิเคราะห์ การสรุปสาระสำคัญนำไปใช้เป็นประโยชน์ในการแก้ไขปัญหา การตัดสินใจในการดำเนินชีวิต การอยู่ร่วมกับผู้อื่น และการพัฒนาตนเองอย่างต่อเนื่อง

ความสามารถด้านคำนวณ (Numeracy Ability) เป็นความสามารถที่เน้นการนำทักษะกระบวนการทางคณิตศาสตร์ไปใช้ในชีวิตประจำวันได้โดยมีการกำหนดนิยามไว้ดังนี้

1. ความสามารถด้านคำนวณ (Numeracy Ability) หมายถึง ความสามารถในการใช้ทักษะกระบวนการทางคณิตศาสตร์ ทักษะการคิดคำนวณ และความคิดรวบยอดทางคณิตศาสตร์ ในสถานการณ์ต่าง ๆ ที่เกี่ยวข้องกับชีวิตประจำวัน

2. ทักษะกระบวนการทางคณิตศาสตร์ หมายถึง ความสามารถในการแก้ปัญหาด้วยวิธีการ

ที่หลากหลาย การให้เหตุผล การสื่อสาร การสื่อความหมายทางคณิตศาสตร์ การนำเสนอ การเชื่อมโยงความรู้และการมีความคิดริเริ่มสร้างสรรค์

3. ทักษะการคิดคำนวณ หมายถึง ความสามารถในการบวก การลบ การคูณ และการหารได้อย่างถูกต้อง คล่องแคล่ว

4. ความคิดรวบยอดทางคณิตศาสตร์ หมายถึง ความรู้ ความเข้าใจเกี่ยวกับจำนวนนับ เศษส่วน ทศนิยม และร้อยละ ความยาว ระยะทาง น้ำหนัก พื้นที่ ปริมาตร ความจุ เวลา เงิน ทิศ แผนผัง ขนาดของมุม ชนิดและสมบัติของรูปเรขาคณิต แบบรูปและความสัมพันธ์ แผนภูมิและกราฟ การคาดคะเน การเกิดขึ้นของเหตุการณ์ต่าง ๆ

ความสามารถด้านเหตุผล (Reasoning Ability) เน้นความสามารถที่มีองค์ประกอบของทักษะเชิงเหตุผล ด้านสังคมศาสตร์ ด้านวิทยาศาสตร์ และทักษะชีวิตมารวมกันโดยกำหนดเป็นนิยามได้ดังนี้

1. ความสามารถด้านเหตุผล (Reasoning Ability) หมายถึง ความสามารถในการเชื่อมโยงความรู้ และประสบการณ์ด้านวิทยาศาสตร์และสิ่งแวดล้อม ด้านสังคมศาสตร์และเศรษฐศาสตร์ และด้านการดำเนินชีวิต โดยการวิเคราะห์ สังเคราะห์ ประเมินค่า แก้ปัญหา หรือตัดสินใจอย่างมีหลักการและเหตุผล บนพื้นฐานของข้อมูล สถานการณ์ หรือสารสนเทศที่เพียงพอ โดยยึดหลักคุณธรรมและจริยธรรม

2. ความรู้ หมายถึง ข้อเท็จจริง ทฤษฎี หลักการ กระบวนการที่ศึกษารวมทั้งคุณธรรมจริยธรรม

3. ประสบการณ์ หมายถึง ความรู้เดิมที่เกิดจากการเรียนรู้ ปฏิบัติหรือได้พบเห็นเรื่องต่าง ๆ ในระดับบุคคล สังคม และสังคมโลก

4. วิเคราะห์ หมายถึง ความสามารถในการเปรียบเทียบ บอกความต่าง ความเหมือน สรุปหลักการ บอกความสัมพันธ์เชื่อมโยงอย่างมีเหตุผล บนพื้นฐานของหลักการทางวิทยาศาสตร์ สังคมศาสตร์ และการดำเนินชีวิตอย่างมีคุณธรรมและจริยธรรม

5. สังเคราะห์ หมายถึง ความสามารถในการสร้างข้อสรุปใหม่ ออกแบบ คิดสร้างสรรค์ บนพื้นฐานของข้อมูลที่ผ่านการวิเคราะห์ ประเมินค่าแล้วอย่างสมเหตุสมผล

6. ประเมินค่า หมายถึง ความสามารถในการตัดสินใจเลือกทางเลือกอย่างสมเหตุสมผล มีประโยชน์ และสร้างสรรค์

7. เหตุผลทางวิทยาศาสตร์ หมายถึง การนำความรู้ ประสบการณ์ที่เกิดจากการเรียนรู้ มาประกอบการตัดสินใจในสถานการณ์ที่เกิดขึ้นในสังคมให้สมเหตุสมผลตามหลักเกณฑ์ทางวิทศาสตร์

8. เหตุผลทางสังคมศาสตร์ หมายถึง การนำความรู้ ประสบการณ์ จากกฎเกณฑ์ ความเชื่อ วัฒนธรรม ค่านิยมทางสังคมศาสตร์ มาประกอบการตัดสินใจในสถานการณ์ที่เกิดขึ้นในสังคมได้อย่างสมเหตุสมผล

9. เหตุผลทางการดำเนินชีวิต หมายถึง การนำความรู้ หลักการ กฎเกณฑ์ มาใช้ในการดำเนินชีวิตหรือประกอบการตัดสินใจในสถานการณ์ที่เกิดขึ้นในสังคมอย่างมีคุณธรรมและจริยธรรม

เครื่องมือที่ใช้ในการสอบวัดความสามารถพื้นฐานผู้เรียนระดับชาติ ชั้นประถมศึกษาปีที่ 3 แสดงรายละเอียด ดังนี้

ตารางที่ 2-1 รายละเอียดเครื่องมือที่ใช้ประเมินนักเรียน ชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2555

แบบวัดความสามารถ	จำนวน (ข้อ)	เวลา (นาที)
ด้านภาษา (Literacy Ability)	30	45
ด้านคำนวณ (Numeracy Ability)	30	45
ด้านเหตุผล (Reasoning Ability)	30	45
รวม	90	135

ตารางที่ 2-2 โครงสร้างแบบทดสอบด้านภาษา ชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2555

ตัวชี้วัด	ข้อที่	จำนวน (ข้อ)
1. อธิบายความหมายจากเรื่องที่อ่าน	1,5,7,9,12	5
2. คาดคะเนเหตุการณ์จากเรื่องที่อ่าน	8,11,13,16	4
3. สรุปเรื่องราวและข้อคิดจากเรื่องที่อ่าน	2,6,10,18,19,23,28	7
4. วิเคราะห์เรื่องที่อ่านได้อย่างถูกต้อง	3,15,17,20,21,29	6
5. นำข้อคิดที่ได้จากเรื่องที่อ่านไปใช้ในชีวิตประจำวัน	14,22,24,25,26,30	6
6. สื่อสารความคิดเห็นจากเรื่องที่อ่านอย่างมีเหตุผลและสร้างสรรค์	4,27	2
รวมทั้งหมด	30	30

ตารางที่ 2-3 โครงสร้างแบบทดสอบด้านคำนวณ ชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2555

สาระการเรียนรู้	ข้อที่	จำนวน (ข้อ)
1. จำนวนและการดำเนินการ	1 – 9, 17 – 19, 27	13
2. การวัด	10 – 16	7
3. เรขาคณิต	20 – 22	3
4. พีชคณิต	23 – 25	3
5. การวิเคราะห์ข้อมูลและความน่าจะเป็น	26, 28 – 30	4
รวมทั้งหมด	30	30

ตารางที่ 2-4 โครงสร้างแบบทดสอบด้านเหตุผล ชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2555

ตัวชี้วัด	ข้อที่	จำนวน (ข้อ)
1. วิเคราะห์ด้วยการเปรียบเทียบข้อมูล สถานการณ์ หรือสารสนเทศโดยใช้ความรู้ด้านวิทยาศาสตร์และสิ่งแวดล้อม ด้านสังคมและเศรษฐกิจ และด้านการดำเนินชีวิตอย่างมีเหตุผล	1-3, 12-14, 23-24	8
2. จัดกลุ่มข้อมูล สถานการณ์ หรือสารสนเทศโดยใช้ความรู้ด้านวิทยาศาสตร์และสิ่งแวดล้อม ด้านสังคมและเศรษฐกิจ และด้านการดำเนินชีวิตอย่างมีเหตุผล	4-6, 15,17, 25-26	7
3. นำข้อมูล สถานการณ์ หรือสารสนเทศมาวิเคราะห์และประยุกต์ใช้ในการวางแผน โดยใช้ความรู้ด้านวิทยาศาสตร์และสิ่งแวดล้อม ด้านสังคมและเศรษฐกิจ และด้านการดำเนินชีวิตอย่างมีเหตุผล	7-8, 16,18-19, 27	6
4. วิเคราะห์และกำหนดปัญหาและสามารถหาทางแก้ปัญหาได้อย่างมีประสิทธิภาพและประสิทธิผล เลือกใช้ข้อมูล สถานการณ์ หรือสารสนเทศในการตัดสินใจและ/หรือแก้ปัญหา โดยใช้ความรู้ด้านวิทยาศาสตร์และสิ่งแวดล้อม ด้านสังคมและเศรษฐกิจ และด้านการดำเนินชีวิตอย่างมีเหตุผล	9-11, 20-22, 28-30	9
รวมทั้งหมด	30	30

เครื่องมือที่ใช้ในการวิจัย

เครื่องมือที่ใช้ในการวิจัยครั้งนี้ เป็นผลการตอบข้อสอบ NT ของนักเรียนชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2555 ประกอบด้วย แบบทดสอบวัดความสามารถ 3 ด้าน ได้แก่ ด้านภาษา (Literacy Ability) ด้านคำนวณ (Numeracy Ability) และด้านเหตุผล (Reasoning Ability) สร้างโดยสำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน คณะกรรมการออกข้อสอบประกอบด้วย ครูผู้สอน ศึกษานิเทศก์ นักวิชาการ และผู้เชี่ยวชาญ ที่เกี่ยวข้อง การสร้างข้อสอบเป็นไปตามวัตถุประสงค์ที่กำหนดให้หลักสูตรชั้นประถมศึกษาปีที่ 3 ซึ่งประกอบด้วยข้อสอบดังต่อไปนี้

1. แบบทดสอบด้านภาษา จำนวน 30 ข้อ
2. แบบทดสอบด้านคำนวณ จำนวน 30 ข้อ
3. แบบทดสอบด้านเหตุผล จำนวน 30 ข้อ

แบบทดสอบด้านภาษา (Literacy Ability)

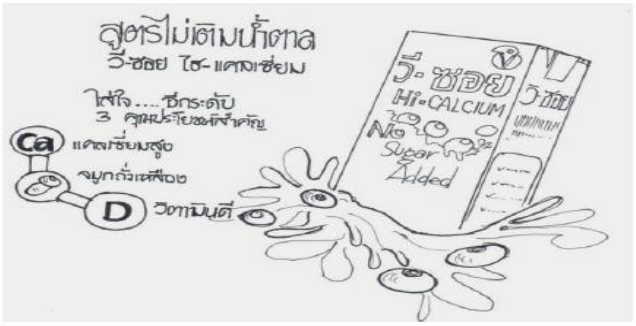
ตัวอย่างเครื่องมือ การวัดพฤติกรรมทางภาษา

อ่านข้อความต่อไปนี้แล้วจงตอบคำถาม

จอรูปสี่เหลี่ยม	เต็มเปี่ยมความรู้
เด็กเด็กชอบดู	เรียนรู้แสนสุข
ท่องโลกเว็บไซต์	ก้าวไกลล้ายุค
อย่าเพียงสนุก	เกิดทุกข์มีภัย

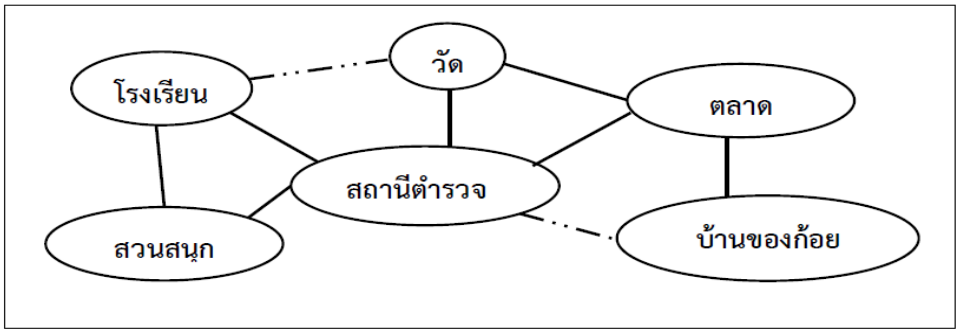
1. ข้อความใดในบทร้อยกรองนี้ ที่แสดงให้รู้ว่าเป็น "คอมพิวเตอร์"
   1) จอรูปสี่เหลี่ยม
   2) เต็มเปี่ยมความรู้
   3) ท่องโลกเว็บไซต์
   4) ก้าวไกลล้ำยุค

พิจารณาภาพแล้วตอบคำถาม



2. จากภาพ นักเรียนจะเลือกซื้อสินค้านี้หรือไม่ เพราะเหตุใด
   1) ซื้อ เพราะเป็นสูตรไม่เติมน้ำตาล
   2) ซื้อ เพราะเชื่อว่าดีต่อสุขภาพ
   3) ไม่ซื้อ เพราะคำโฆษณาเกินจริง
   4) ไม่ซื้อ เพราะราคาแพง

พิจารณาภาพแล้วตอบคำถาม

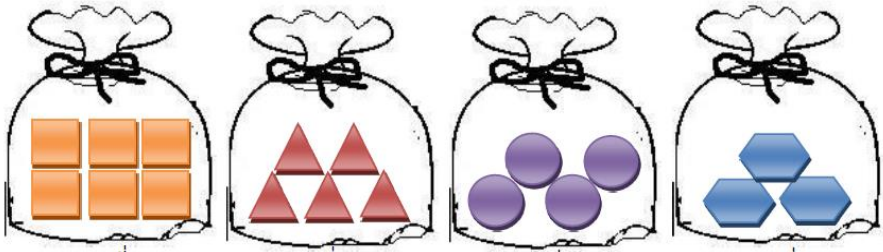


3. จากภาพเส้นทางใดที่ก้อยเดินทางไปโรงเรียนได้สะดวกที่สุด

1) ตลาด วัด โรงเรียน
2) สถานีตำรวจ วัด โรงเรียน
3) ตลาด สถานีตำรวจ โรงเรียน
4) สถานีตำรวจ สวนสนุก โรงเรียน

แบบทดสอบด้านคำนวณ (Numeracy Ability)
ตัวอย่างเครื่องมือ การวัดพฤติกรรมทางด้านการคิดคำนวณ

1. ภาพต่อไปนี้แสดงจำนวนขนมและราคาขนมที่อยู่ในถุง

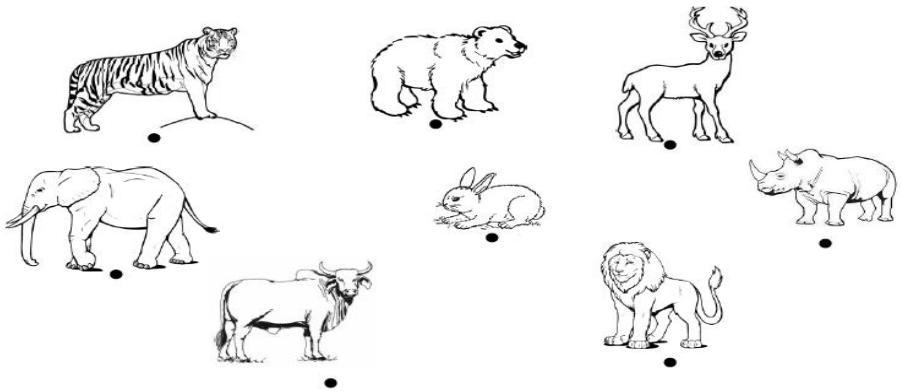


ถุงที่ 1 ราคาชิ้นละ 6 บาท | ถุงที่ 2 ราคาชิ้นละ 4 บาท | ถุงที่ 3 ราคาชิ้นละ 7 บาท | ถุงที่ 4 ราคาชิ้นละ 5 บาท

แดงมีเงิน 50 บาท ต้องการซื้อขนม 2 ถุง ให้ได้จำนวนชิ้นขนม**มากที่สุด**และยังมีเงินเหลือ เขาจะต้องเลือกซื้อขนมถุงใดบ้าง

1) ถุงที่ 1 และถุงที่ 3
2) ถุงที่ 2 และถุงที่ 3
3) ถุงที่ 1 และถุงที่ 4
4) ถุงที่ 2 และถุงที่ 4

2. แพนเค้กไปเที่ยวสวนสัตว์ เธอต้องการให้เส้นทางเดินดูสัตว์เป็นรูปสี่เหลี่ยม จากแผนภาพที่กำหนดให้ แพนเค้กต้องเดินดูสัตว์ตามข้อใด



1) เสือ ⟶ ช้าง ⟶ กระต่าย ⟶ หมี ⟶ เสือ
2) หมี ⟶ กวาง ⟶ กระต่าย⟶ แรด ⟶ หมี
3) แรด ⟶ สิงโต ⟶ กระต่าย ⟶ ช้าง ⟶ แรด
4) สิงโต⟶ กระต่าย ⟶ กวาง ⟶วัว ⟶ สิงโต

จงตอบคำถามจากแผนภูมิ

แผนภูมิแสดงน้ำหนักและราคาขายผลไม้ชนิดต่าง ๆ

ผลไม้	ราคาขาย/กก.
	40
	25
	10
	40
	35

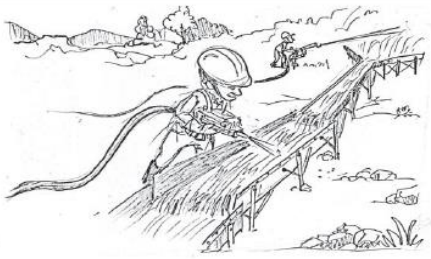
กำหนดให้ รูปผลไม้ 1 รูป แทนผลไม้หนัก 4 กิโลกรัม

3. แม่ค้าขายส้มและสับปะรด รวมกันได้มากหรือน้อยกว่าชมพู่กี่กิโลกรัม
   1) น้อยกว่า 4 กิโลกรัม
   2) มากกว่า 4 กิโลกรัม
   3) น้อยกว่า 8 กิโลกรัม
   4) มากกว่า 8 กิโลกรัม

แบบทดสอบด้านเหตุผล (Reasoning ability)
ตัวอย่างเครื่องมือ การวัดพฤติกรรมทางด้านเหตุผล

อ่านข้อความที่กำหนดแล้วตอบคำถาม

ชุมชนที่ 1 มีอาชีพระเบิดภูเขาหินปูนเพื่อนำมาทำถนน และสร้างอาคารบ้านเรือน

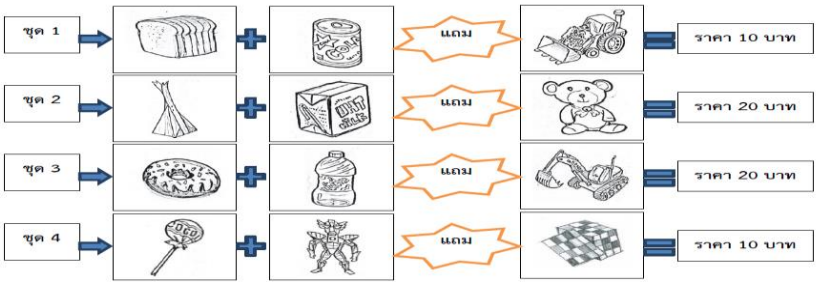


ชุมชนที่ 2 มีอาชีพทำเหมืองแร่ โดยการฉีดน้ำให้ก้อนแร่ไหลตามราง

1. อาชีพของทั้งสองชุมชนส่งผลกระทบต่อสิ่งใดมากที่สุด
   1) ป่าไม้ลดลง
   2) สัตว์ป่าลดลง
   3) น้ำป่าไหลหลาก
   4) ทรัพยากรลดลงและหมดไป

ดูรูปภาพที่กำหนดแล้วตอบคำถาม

แม่ค้าจัดของขายเป็นชุด ๆ ดังนี้



2. ถ้านักเรียนมีเงินอยู่ 20 บาท นักเรียนควรจะซื้อชุดใด
   1) ชุดที่ 1 เพราะได้ของราคาถูก
   2) ชุดที่ 2 เพราะได้ดื่มนมมีประโยชน์
   3) ชุดที่ 3 เพราะได้น้ำดื่มและของเล่นเหมือนกัน
   4) ชุดที่ 4 เพราะมีเงินเหลือพอที่จะซื้อ

อ่านข้อความที่กำหนดแล้วตอบคำถาม

กลุ่ม 1	 พนักงานขับรถ	 ตำรวจจราจร	 พนักงานโรงแรม	 บุรุษไปรษณีย์
กลุ่ม 2	 นักออกแบบ	 พ่อครัว	 ครู	 นักเขียน

3. แดงจะจัดอาชีพพยาบาล เป็นงานประเภทเดียวกันกับกลุ่ม 1 เพราะเหตุใด

1) เพราะเป็นอาชีพที่เสียสละ

2) เพราะเป็นอาชีพที่มีรายได้แน่นอน

3) เพราะเป็นอาชีพบริการประชาชน

4) เพราะเป็นอาชีพที่ไม่เสี่ยงอันตราย

5. หลักการของการทดสอบระดับชาติ

การทดสอบระดับชาติ (National Testing: NT) จะมีการจัดสอบปีการศึกษาละ 1 ครั้ง จัดสอบในช่วงเดือนกุมภาพันธ์-มีนาคม จัดสอบกับนักเรียนชั้นประถมศึกษาปีที่ 3 วัดด้านภาษา (Literacy Ability) ด้านคำนวณ (Numeracy Ability) และด้านเหตุผล (Reasoning Ability) ของนักเรียนชั้นประถมศึกษาปีที่ 3 (คู่มือการจัดสอบ NT ชั้นประถมศึกษาปีที่ 3, 2555, หน้า 5)

แนวปฏิบัติการจัดสอบ

ให้มีกรรมการกำกับห้องสอบ ห้องละ 2 คน ปฏิบัติตามระเบียบกระทรวงศึกษาธิการ ว่าด้วยการปฏิบัติของผู้กำกับการสอบ พ.ศ. 2548 อย่างเคร่งครัด การจัดห้องสอบให้จัดห้องสอบมีที่นั่งสอบไม่เกิน 35 คนต่อห้อง ในกรณีที่โรงเรียนมีห้องเรียนห้องเดียวและมีนักเรียนเกิน 35 คน แต่ไม่เกิน 40 คน อาจจัดห้องสอบเป็นห้องเดียวกันได้ แต่ควรจัดโต๊ะให้มีระยะห่างกันพอสมควรและไม่ควรจัดโต๊ะให้อยู่นอกห้องสอบ กรณีจำนวนผู้เข้าสอบเกิน 40 คน ให้จัดห้องสอบเพิ่ม ประกาศรายชื่อนักเรียนและแผนผังที่นั่งสอบติดหน้าห้องสอบทุกห้อง พร้อมข้อมูลรายละเอียดของนักเรียนแต่ละคน ที่ต้องระบายในกระดาษคำตอบ (คู่มือการจัดสอบ NT ชั้นประถมศึกษาปีที่ 3, 2555, หน้า 23)

การประกาศผลสอบ

การประกาศผลสอบสำนักทดสอบทางการศึกษาจะประกาศผลสอบ สำหรับเรียกดูรายงานผลของนักเรียนรายบุคคลทางเว็บไซต์ http://nt-result.bopp.go.th/student-user สำหรับเจ้าหน้าที่โรงเรียนทาง เว็บไซต์ http://nt-result.bopp.go.th และสำหรับเจ้าหน้าที่เขตพื้นที่การศึกษาทางเว็บไซต์ http://epcc2.bopp.go.th ช่วงเดือนมีนาคม ซึ่งสามารถตรวจสอบผลการทดสอบได้ทั้งระดับโรงเรียน (ใช้ username และ password) และระดับรายบุคคล (ใช้เลขบัตรประชาชน)

6. ประโยชน์ของการทดสอบทางการศึกษา

เพื่อให้ผลการประเมินสามารถนำไปพัฒนาคุณภาพการศึกษาได้อย่างมีประสิทธิภาพ ส่งผลถึงตัวผู้เรียน และปรับปรุงการจัดการเรียนการสอนของโรงเรียน และการกำกับดูแลแนวทางการจัดการศึกษาของเขตพื้นที่การศึกษา ให้เป็นไปตามจุดเน้นของสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐานที่สำคัญคือ ผู้เรียนในระดับชั้นประถมศึกษาปีที่ 3 ต้องอ่านออกเขียนได้ คิดเลขได้ และคิดวิเคราะห์อย่างมีเหตุผลอย่างเป็นรูปธรรม เขตพื้นที่การศึกษาจำเป็นต้องวิเคราะห์ผลจากการประเมินจัดทำเป็นแผนการยกระดับคุณภาพการศึกษาในทุกระดับ ตั้งแต่ตัวผู้เรียนรายบุคคล สภาพการจัดการเรียนการสอนของโรงเรียน สภาพปัญหาและความต้องการ บริบทในการบริหารจัดการ การกำกับดูแลการจัดการเรียนการสอนของเขตพื้นที่การศึกษา โดยมีจุดเน้นในแต่ละระดับ ดังนี้ (คู่มือการจัดสอบ NT ชั้นประถมศึกษาปีที่ 3, 2555, หน้า 28)

6.1 ผู้เรียนรายบุคคล ให้โรงเรียนวิเคราะห์ผลรายบุคคลของนักเรียน ที่สะท้อนถึงความสามารถผู้เรียนในความสามารถแต่ละด้าน และแจ้งให้นักเรียนทราบถึงจุดเด่น-จุดด้อย ที่ต้องเร่งพัฒนาและปรับปรุง เป็นรายบุคคลเพื่อให้ผู้เรียนสามารถปรับปรุงและพัฒนาตนเองได้อย่างถูกทิศทาง

6.2 โรงเรียนวิเคราะห์ผลภาพรวม เปรียบเทียบกับผลภาพรวมระดับเขตพื้นที่การศึกษา เพื่อทราบถึงจุดที่ต้องเร่งดำเนินการพัฒนา ปรับปรุง การจัดการเรียนการสอนของตน จัดทำเป็นแผนยกระดับ กำหนดเป้าหมายในการพัฒนา

6.3 เขตพื้นที่การศึกษาวิเคราะห์ผลภาพรวมของแต่ละโรงเรียน และภาพรวมของเขตพื้นที่การศึกษา ได้แนวทางในการส่งเสริมปรับปรุงการจัดการเรียนการสอนเพื่อยกระดับคุณภาพการศึกษา กำหนดเป็นแนวทางในการพัฒนาการจัดการศึกษาของเขตพื้นที่การศึกษา

งานวิจัยที่เกี่ยวข้องกับการสอบวัดความสามารถพื้นฐานของผู้เรียนระดับชาติ (NT) มีดังนี้

วรพรรณ ศรีกล่ำ (2559) ได้ศึกษาปัจจัยพหุระดับนักเรียนและระดับห้องเรียนที่ส่งผลต่อคะแนน การทดสอบระดับชาติ ด้านภาษาโรงเรียนที่มีผลคะแนนการทดสอบระดับชาติต่ำในจังหวัดพิษณุโลก กลุ่มตัวอย่าง ได้แก่ นักเรียนจำนวน 1,260 คน และครูจำนวน 68 ห้องเรียน ข้อมูลที่เก็บรวบรวม ใช้แบบสอบถาม จำนวน 2 ฉบับ แบบสอบถามระดับนักเรียนและระดับห้องเรียน 1) แบบสอบถามระดับนักเรียนประกอบด้วย 5 ตัวแปร แบบสอบถามวัดความรู้พื้นฐานเดิมแบบสอบถามวัดแรงจูงใจใฝ่สัมฤทธิ์ในการทำแบบทดสอบของการทดสอบระดับชาติ แบบสอบถามวัดเจตคติต่อการเรียนภาษาไทย แบบสอบถามวัดสภาพแวดล้อมทางบ้าน และแบบสอบถามวัดความเอาใจใส่ของผู้ปกครองในการส่งเสริมการเรียน ซึ่งมีค่าความเชื่อมั่นเท่ากับ 0.92 แบบสอบถามระดับห้องเรียนประกอบด้วย 2 ตัวแปร แบบสอบถามวัดคุณภาพการสอนครูภาษาไทยและแบบสอบถามวัดบรรยากาศในชั้นเรียน มีค่าความเชื่อมั่นเท่ากับ 0.87 ทำการวิเคราะห์ข้อมูลโดยการวิเคราะห์พหุระดับ (Multilevel Analysis) ผลการศึกษาปรากฏว่า โรงเรียนที่มีผลคะแนนการทดสอบระดับชาติต่ำ 1) ในระดับนักเรียน ตัวแปรความรู้พื้นฐานเดิมส่งผลต่อคะแนนการทดสอบระดับชาติ ด้านภาษาอย่างมีนัยสำคัญทางสถิติที่ .05 2) ในระดับห้องเรียนไม่มีตัวแปรอิสระใดที่ส่งผลต่อคะแนนการทดสอบระดับชาติ ด้านภาษา ตัวแปรความรู้พื้นฐานเดิมสามารถอธิบายความแปรปรวนของคะแนนการทดสอบระดับชาติ ด้านภาษาได้ร้อยละ 12.85

เอกลักษณ์ คล้ายสุบรรณ สังวรณ์ งัดกระโทก และนลินี ณ นคร (2559) ได้ศึกษาพัฒนาวิธีการวัดมูลค่าเพิ่มทางการศึกษาเพื่อใช้ประเมินคุณภาพสถานศึกษาด้วยการวัดมูลค่าเพิ่มจากผลสัมฤทธิ์ทางการเรียนและผลการประเมินและรับรองคุณภาพของโรงเรียน โดยงานวิจัยนี้ได้ใช้คะแนนจากการทดสอบระดับชาติ (NT) ศึกษาความสอดคล้องของผลการประเมินคุณภาพสถานศึกษาด้วยการวัดมูลค่าเพิ่มที่พัฒนาขึ้น ซึ่งมีการกำหนดน้ำหนักของการรวมคะแนนแตกต่างกันและเปรียบเทียบมูลค่าเพิ่มทางการศึกษาของสถานศึกษาที่มีบริบทต่างกัน 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 ผลการศึกษาปรากฏว่า โมเดลการวัดมูลค่าเพิ่มทางการศึกษาที่มีความกลมกลืนกับข้อมูลมากที่สุด คือ โมเดลที่ 1 (40:60) รองลงมาคือโมเดลที่ 2 (50:50) และโมเดลที่ 3 (60:40) (2) ความสอดคล้องของผลการประเมินคุณภาพสถานศึกษาด้วยการวัดมูลค่าเพิ่มระหว่างโมเดลที่ 1 (40:60) กับโมเดลที่ 2 (50:50) มีความสอดคล้องมากที่สุด มีความสอดคล้อง 91.67% รองลงมาระหว่างโมเดลที่ 2 (50:50) กับโมเดลที่ 3 (60:40) และมีความสอดคล้อง 88.54% และระหว่างโมเดลที่ 1 (40:60) กับโมเดลที่ 3 (60:40) มีความสอดคล้อง 80.21% ตามลำดับ และ (3) สถานศึกษาที่มีที่ตั้งและขนาดต่างกัน มีคะแนนมูลค่าเพิ่มทางการศึกษาไม่แตกต่างกัน

สุธาทิพย์ ตรีสิน และปิยะทิพย์ ประดุจพรม (2560) ได้ศึกษาการวิเคราะห์คุณภาพของแบบทดสอบระดับชาติ (NT) และตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านความสามารถด้านภาษา ความสามารถด้านคำนวณ และความสามารถด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR การดำเนินการวิจัยแบ่งเป็น 3 ระยะ ดังนี้ 1) วิเคราะห์คุณภาพของแบบทดสอบระดับชาติ ทั้ง 3 ด้าน 2) ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR และ 3) เปรียบเทียบผลการวิเคราะห์ การทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธีการตรวจสอบ 3 วิธี ข้อมูลที่นำมาใช้วิเคราะห์เป็นข้อมูลทุติยภูมิจากผลการตอบแบบทดสอบระดับชาติของนักเรียนชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2556 จำนวน 9,600 คน ผลการศึกษาปรากฏว่า แบบทดสอบระดับชาติ ชั้นประถม ศึกษาปีที่ 3 มีค่าความยากของข้อสอบ (b) อยู่ในระดับค่อนข้างยาก มีค่าอำนาจจำแนกของข้อสอบ (a) อยู่ในระดับที่สามารถจำแนกผู้สอบได้ดี และมีค่าโอกาสในการเดาของข้อสอบ (c) ไม่เกิน 0.3 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้งสามด้าน ชี้ให้เห็นว่า เพศส่งผลให้เกิดการทำหน้าที่ต่างกันของข้อสอบโดยเพศหญิงจะได้เปรียบในการตอบข้อสอบด้านภาษาและด้านเหตุผล ในขณะที่เพศชายจะได้เปรียบในการตอบข้อสอบด้านคำนวณ โดยวิธี HGLM ตรวจพบข้อสอบทำหน้าที่ต่างกันจำนวนมากที่สุด คิดเป็น

ร้อยละ 69 ของข้อสอบทั้งฉบับ รองลงมาคือ วิธี IRT-LR ร้อยละ 54 และวิธี MIMIC ร้อยละ 16 และด้านเหตุผลตามลำดับ การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ พบว่า วิธี HGLM ตรวจพบ DIF มากกว่าวิธี MIMIC ในด้านภาษา ด้านคำนวณ คิดเป็นร้อยละ 70,36, และ 53 ตามลำดับ และวิธี HGLM ตรวจพบ DIF มากกว่าวิธี IRT-LR ด้านภาษา และด้านคำนวณ คิดเป็นร้อยละ 37 และ 13 และวิธี IRT-LR ตรวจพบ DIF มากกว่าวิธี MIMIC ทั้ง 3 ด้านคิดเป็นร้อยละ 33,43 และ 40 ตามลำดับ ส่วนวิธี HGLM ตรวจพบ DIF น้อยกว่าวิธี IRT-LR ด้านคำนวณคิดเป็นร้อยละ 7 ($p < .05$)

Kjellstrom and Pettersson (2005) ได้ศึกษาระบบการทดสอบในประเทศสวีเดน การทดสอบระดับชาติของสวีเดนเป็นการมองที่เป้าหมายและองค์ความรู้ ซึ่งระบบที่สำคัญของการทดสอบคือความรู้ในหลักสูตร การศึกษาระบบการทดสอบทำให้ทราบถึงภาพรวมของการเรียนการสอนของวิชาคณิตศาสตร์และอิทธิพลต่างๆที่มีผลต่อการทดสอบระดับชาติ การประเมินที่แตกต่างไปจากเดิม โดยรวมมุ่งเน้นไปที่กระบวนการเรียนการสอนวิชาคณิตศาสตร์ ผลการศึกษาปรากฏว่าหลักสูตรพัฒนาหลักสูตรและการจัดกิจกรรมการเรียนการสอนให้มีประสิทธิภาพมากยิ่งขึ้น

Brown, De Four-Babb, Bristol, and Conrad (2014) ได้ศึกษาการทดสอบระดับชาติเกี่ยวกับคำพูดของครูในตรินิแดดและโตเบโก ข้อเสนอแนะของการทดสอบกล่าวถึง การรายงาน การตัดสินใจทางหลักสูตรที่ส่งผลกระทบต่อการเรียนรู้ของนักเรียน กลุ่มตัวอย่างประกอบด้วยครูประถมศึกษา จำนวน 133 คน แบ่งเป็นครูจากโรงเรียนที่มีประสิทธิภาพต่ำ จำนวน 79 คน และครูจากโรงเรียนที่มีประสิทธิภาพสูง จำนวน 54 คน และผู้บริหาร จำนวน 10 คน ผลการวิจัยเชิงปริมาณและเชิงคุณภาพ ปรากฏว่า ครูจำนวนมากรู้สึกไม่สบายใจกับการตีความข้อมูลที่นำเสนอในรายงานเกี่ยวกับการทดสอบระดับชาติ การศึกษาครั้งนี้มีการตัดสินใจในเรื่องของการจัดกิจกรรมการเรียนการสอนและการพัฒนาหลักสูตร มีความจำเป็นที่จะต้องฝึกอบรมครูในการใช้และการตีความข้อมูลจากการประเมินมีการสร้างสัญลักษณ์ให้กับโรงเรียนที่มีผลการประเมินที่มีประสิทธิภาพสูงและโรงเรียนที่มีประสิทธิภาพต่ำ

จากการศึกษางานวิจัยที่เกี่ยวข้อง สามารถสรุปได้ว่า การทดสอบระดับชาติ (NT) มีความสำคัญต่อการพัฒนาการศึกษาของประเทศเป็นอย่างมาก การทดสอบมุ่งเน้นการทดสอบความสามารถของนักเรียนทั้งสามด้าน คือ ความสามารถด้านภาษา ความสามารถด้านคำนวณ และความสามารถด้านเหตุผลจำนวนด้านละ 30 ข้อ เพื่อทำการประเมินระดับความสามารถของนักเรียนชั้นประถมศึกษาปีที่ 3 จากการศึกษาปรากฏว่า ผู้เรียน ครูผู้สอน กิจกรรมการเรียนการสอน รวมทั้งลักษณะของแบบทดสอบเป็นปัจจัยหนึ่งที่มีผลต่อการทดสอบระดับชาติ ที่ก่อให้เกิดความไม่เท่าเทียมกันในการทำข้อสอบ และส่งผลต่อคะแนนที่ได้จากการทำแบบทดสอบด้วย

ตอนที่ 2 ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) และงานวิจัยที่เกี่ยวข้อง

Hambleton, Swaminathan, and Rogers (1991, p. 174) กล่าวว่า ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) อยู่บนแนวคิด 2 ประการ ดังนี้ 1) พฤติกรรมในการตอบข้อสอบของผู้เข้าสอบนำไปใช้ทำนายความสามารถ (Ability) หรือลักษณะภายใน (Trait) ของผู้เข้าสอบโดยมีข้อตกลงเบื้องต้น ดังนี้

ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) เป็นทฤษฎีที่พัฒนาขึ้นเพื่อแก้ไขจุดด้อยของทฤษฎีการทดสอบแบบดังเดิม (Classical Test Theory: CTT) ซึ่งจุดด้อยที่สำคัญ คือ ค่าพารามิเตอร์ของข้อสอบจะแปรผันตามกลุ่มของผู้สอบ และคะแนนหรือการประมาณค่าความสามารถของผู้เข้าสอบไม่เป็นอิสระขึ้นอยู่กับข้อสอบหรือแบบทดสอบที่นำมาใช้ สำหรับทฤษฎีการตอบสนองข้อสอบ สามารถแบ่งออกเป็น 2 ประเภท คือ 1) ทฤษฎีการตอบสนองข้อสอบ แบบตรวจให้คะแนนแบบ 2 ค่า (Dichotomous Item Response Theory) และ 2) ทฤษฎีการตอบสนองข้อสอบแบบตรวจให้คะแนนมากกว่า 2 ค่า (Polytomous Item Response Theory) สำหรับงานวิจัยนี้กล่าวถึงเฉพาะในส่วนของทฤษฎีการตอบสนองข้อสอบ แบบตรวจให้คะแนน 2 ค่า (Dichotomous Item Response Theory) เท่านั้น โดยมีรายละเอียด ดังนี้

ทฤษฎีการตอบสนองข้อสอบ แบบตรวจให้คะแนนแบบ 2 ค่า (Dichotomous Item Response Theory) เป็นทฤษฎีที่อธิบายถึงความสัมพันธ์ระหว่างความสามารถของผู้สอบกับการตอบข้อสอบโดยใช้โค้งคุณลักษณะข้อสอบ (Item Characteristic Curve: ICC) ซึ่งมีการกำหนดคุณลักษณะข้อสอบด้วยค่าอำนาจจำแนก (a) ค่าความยาก (b) และค่าการเดา (c) มีหลักการตรวจให้คะแนนเพียง 2 ค่า เช่น ถูก-ผิด ใช่-ไม่ใช่ หรือ 0,1 เป็นต้น

1. ข้อตกลงเบื้องต้นของทฤษฎีการตอบสนองข้อสอบ

1.1 ความเป็นมิติเดียว (Unidimension) หมายถึง ผลการตอบข้อสอบของผู้เข้าสอบสามารถอธิบายความสามารถหรือคุณลักษณะภายในเพียงด้านใดด้านหนึ่งของผู้เข้าสอบและคุณลักษณะภายในด้านเดียวนี้มีความหมายเหมือนเป็นมิติเดียว ซึ่งข้อตกลงนี้ ชี้ให้เห็นว่าอาจมีคุณลักษณะของข้อสอบบางประการที่ส่งผลร่วมต่อการตอบข้อสอบเข้ามาเกี่ยวข้อง ดังนั้น จึงควรกำหนดความเป็นมิติเดียวให้เป็นลักษณะเด่น (Dominant) หรือลักษณะหลัก เพื่อที่จะนำไปอธิบายผลการตอบข้อสอบของผู้เข้าสอบได้

1.2 ความเป็นอิสระในการตอบข้อสอบ (Local Independence) หมายถึง เมื่อค่าความสามารถของผู้เข้าสอบเป็นค่าแน่นอน การตอบข้อสอบแต่ละข้อของผู้เข้าสอบคนหนึ่งจะมีความเป็นอิสระจากกัน กล่าวได้ว่า การตอบข้อสอบข้อใด ๆ ของผู้เข้าสอบจะไม่มีผลต่อข้อสอบข้ออื่น ๆ เลยแต่สิ่งที่ส่งผลต่อการตอบข้อสอบแต่ละข้อเป็นผลมาจากความสามารถของผู้เข้าสอบเท่านั้น ความเป็นอิสระในการตอบข้อสอบ ทำให้ค่าพารามิเตอร์ข้อสอบยังเป็นค่าคงที่ ไม่ว่าข้อสอบข้อนั้นอยู่ตำแหน่งใด ๆ ก็ตาม

1.3 โค้งคุณลักษณะของข้อสอบ (Item Characteristic Curve: ICC) หรือฟังก์ชันการตอบสนองข้อสอบ (Item Response Function: IRF) เป็นฟังก์ชันทางคณิตศาสตร์ที่แสดงความสัมพันธ์ระหว่างความน่าจะเป็นของการตอบข้อสอบถูกกับระดับความสามารถของผู้เข้าสอบ

2. โมเดลการตอบสนองข้อสอบ (Item Response Models)

โมเดลการตอบสนองข้อสอบ เป็นโมเดลแสดงความสัมพันธ์ระหว่างโอกาสตอบข้อสอบถูกกับความสามารถของผู้เข้าสอบในรูปแบบของโค้งคุณลักษณะเฉพาะของข้อสอบ ซึ่งมีลักษณะเป็นฟังก์ชั่นโลจิสติก (Logistic Function) หรือฟังก์ชันปกติสะสม (Normal Ogive Function) สามารถเรียกอีกอย่างหนึ่งว่า *“โมเดลโลจิสติกหรือโมเดลปกติสะสม”* โมเดลการตอบสนองข้อสอบ มี 3 รูปแบบ ดังต่อไปนี้

2.1 โมเดลการตอบสนองข้อสอบแบบ 1 พารามิเตอร์ (One - Parameter Model) เป็นโมเดลที่อธิบายผลการวิเคราะห์ข้อสอบด้วยค่าความยากของข้อสอบ (b) เรียกอีกอย่างว่า *"Rasch Model"* สามารถเขียนฟังก์ชันโลจิสติก ตามสมการที่ 1

$$\mathrm{Pi}(\theta) = \frac{1}{1+ e^{-(\theta - \mathrm{bi})}} \qquad \mathrm{i} = 1, 2, 3, \ldots \mathrm{n} \qquad (1)$$

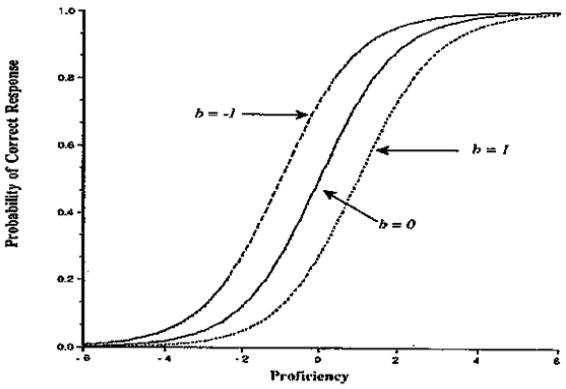
เมื่อ

$P_i(\theta)$ คือ ความน่าจะเป็นของผู้เข้าสอบที่มีความสามารถ (θ) จะตอบข้อสอบข้อที่ I ได้ถูกต้อง

b_i คือ ค่าความยากของข้อสอบข้อที่ i

θ คือ ความสามารถของผู้เข้าสอบ

e คือ 2.72



ภาพที่ 2-1 โค้งคุณลักษณะของข้อสอบแบบ 1 พารามิเตอร์ (Wainer, 2000, p. 68)

2.2 โมเดลการตอบสนองข้อสอบแบบ 2 พารามิเตอร์ (Two - Parameter Model)

เป็นโมเดลการวิเคราะห์ข้อสอบที่ใช้ค่าพารามิเตอร์แบบ 2 พารามิเตอร์ คือ ค่าความยากของข้อสอบ (b) และค่าอำนาจจำแนกของข้อสอบ (a) เขียนเป็นฟังก์ชันโลจิสติก ตามสมการที่ 2

$$\mathrm{Pi}(\theta) = \frac{1}{1+e^{-\mathrm{Da}_i(\theta - \mathrm{bi})}} \qquad \mathrm{i} = 1, 2, 3, \ldots \mathrm{n} \qquad (2)$$

เมื่อ

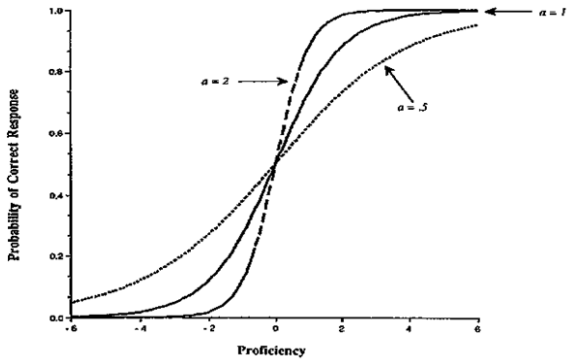
$P_i(\theta)$ คือ ความน่าจะเป็นของผู้เข้าสอบที่มีความสามารถ (θ) จะตอบข้อสอบข้อที่ I ได้ถูกต้อง

b_i คือ ค่าความยากของข้อสอบข้อที่ i

a_i คือ ค่าอำนาจจำแนกของข้อสอบข้อที่ i

θ คือ ความสามารถของผู้เข้าสอบ

D คือ 1.70



ภาพที่ 2-2 โค้งคุณลักษณะของข้อสอบ แบบ 2 พารามิเตอร์ (Wainer, 2000, p. 70)

2.3 โมเดลการตอบสนองข้อสอบ 3 พารามิเตอร์ (Three - Parameter Model)

เป็นโมเดลพัฒนามาจากโมเดลการวิเคราะห์ข้อสอบที่ใช้ค่าพารามิเตอร์ 3 พารามิเตอร์ คือ ค่าความยากของข้อสอบ (b) ค่าอำนาจจำแนกของข้อสอบ (a) และค่าการเดาของข้อสอบ (c) เขียนเป็นฟังก์ชันโลจิสติก ตามสมการที่ 3

$$\mathrm{Pi}(\theta) = \mathrm{C}_i + \frac{1-c_i}{1+e^{-\mathrm{Da}_i(\theta-\mathrm{bi})}} \qquad \mathrm{i} = 1, 2, 3,\ldots\mathrm{n} \tag{3}$$

เมื่อ

$P_i(\theta)$ คือ ความน่าจะเป็นของผู้เข้าสอบที่มีความสามารถ (θ) จะตอบข้อสอบข้อที่ I ได้ถูกต้อง

b_i คือ ค่าความยากของข้อสอบข้อที่ i

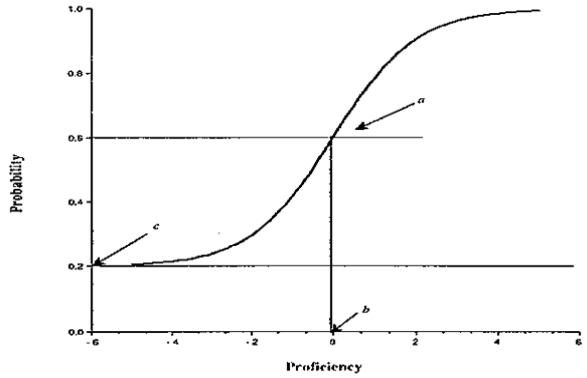
a_i คือ ค่าอำนาจจำแนกของข้อสอบข้อที่ i

c_i คือ ค่าการเดาของข้อสอบข้อที่ i

θ คือ ความสามารถของผู้เข้าสอบ

D คือ 1.70

e คือ 2.72



ภาพที่ 2-3 โค้งคุณลักษณะของข้อสอบแบบ 3 พารามิเตอร์ (Wainer, 2000, p. 71)

3. ความไม่เปลี่ยนแปลงของพารามิเตอร์

เมื่อโมเดลการตอบสนองข้อสอบมีความสอดคล้องกับข้อมูลที่มีอยู่จะทำให้ค่าพารามิเตอร์ของข้อสอบ (Item Parameter) และค่าพารามิเตอร์ความสามารถของผู้เข้าสอบ (Ability Parameter) ไม่เปลี่ยนแปลง ซึ่งเป็นคุณสมบัติสำคัญของทฤษฎีการตอบสนองข้อสอบ (IRT) ซึ่งโค้งคุณลักษณะของข้อสอบ (ICC) จะมีลักษณะเดียวกัน (a, b และ c) สำหรับทุกกลุ่มความสามารถของผู้เข้าสอบ นั่นคือ โค้งคุณลักษณะข้อสอบมีความคงที่ข้ามกลุ่มผู้เข้าสอบ

งานวิจัยที่เกี่ยวข้องกับทฤษฎีการตอบสนองข้อสอบ (IRT) มีดังนี้

สำราญ มีแจ้ง ประภัสสร วงษ์ดี และยุพิน โกณฑา (2552) ได้ศึกษาวิธีการปรับเทียบคะแนน O-NET ของนักเรียนระดับชั้นมัธยมศึกษาปีที่ 2 ระหว่างปี 2550 กับปี 2551 โดยวิธีการเทียบเป็นมาตราเดียวกัน โดยการใช้ทฤษฎีการตอบสนองข้อสอบ ผลการวิจัย พบว่า คะแนน O-NET ทั้ง 2 ปี ต้องมีการปรับคะแนนให้มีหน่วยเดียวกันก่อนจึงจะนำมาเทียบกันได้ ผลการปรับเทียบคะแนนโดยวิธีการปรับเทียบให้เป็นมาตราเดียวกัน ผลการศึกษาปรากฏว่า ค่าเฉลี่ยคะแนน O-NET เดิมกับค่าเฉลี่ยคะแนน O-NET ที่มีการปรับเทียบ ใน 4 กลุ่มสาระการเรียนรู้ ไม่แตกต่างกัน ส่วนผลการปรับเทียบคะแนน O-NET โดยใช้ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) แบบ 3 พารามิเตอร์ ปรากฏว่าค่าเฉลี่ยคะแนน O-NET เดิมกับค่าเฉลี่ยคะแนน O-NET ที่มีการปรับเทียบ แตกต่างกัน อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ทั้งสองวิธีให้คะแนนแปลที่มีค่าสัมประสิทธิ์สหสัมพันธ์สูง และมีความสัมพันธ์ทางบวกกับคะแนน O-NET เดิม

รุ่งนภา แสนอำนวย ประกฤติยา ทักษิโณ และชนะศึก นิชานนท์ (2555) ได้ศึกษาประสิทธิภาพของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนแบบผสม การประยุกต์ใช้ทฤษฎีการตอบสนองข้อสอบแบบตรวจให้คะแนนความรู้บางส่วน และทฤษฎีการตอบสนองข้อสอบแบบตรวจให้คะแนนบางส่วนแบบทั่วไป เพื่อศึกษาประสิทธิภาพของแบบทดสอบแบบผสม ปฏิสัมพันธ์ระหว่างโมเดลการตรวจให้คะแนน สัดส่วนของข้อสอบที่ตรวจให้คะแนนแบบสองค่าและมากกว่าสองค่า ความยาวของแบบทดสอบ และเพื่อเปรียบเทียบประสิทธิภาพของแบบทดสอบรูปแบบประสม เงื่อนไขที่ทำการศึกษามี 18 เงื่อนไข ประกอบ ด้วยโมเดลการตรวจให้คะแนน 2 โมเดล คือ โมเดลโลจิสติก 1 พารามิเตอร์กับโมเดลการตรวจให้คะแนนความรู้บางส่วน (PCM) และโมเดลโลจิสติก 3 พารามิเตอร์กับโมเดล

การตรวจให้คะแนนความรู้บางส่วนแบบทั่วไป (GPCM) สัดส่วนข้อสอบที่ตรวจให้คะแนนแบบสองค่าและมากกว่าสองค่า 3 สัดส่วน คือ 20:80 50:50 และ 80:20 และความยาวของแบบทดสอบ 3 เงื่อนไข คือ 10 30 และ 50 ข้อ การประเมินประสิทธิภาพของแบบทดสอบรูปแบบผสมพิจารณาจากดัชนีความคลาดเคลื่อน มาตรฐานในการประมาณค่า θ ดัชนีความลำเอียง (BIAS) พร้อมทั้งวิเคราะห์ความแปรปรวนแบบพหุจำแนก 3 ทางเพื่อเปรียบเทียบค่าเฉลี่ยของดัชนี θ และดัชนีความลำเอียง ผลการศึกษาพบว่า โมเดลโลจิสติก 1 พารามิเตอร์กับ PCM และโมเดลโลจิสติก 3 พารามิเตอร์กับ GPCM มีค่า θ และ BIAS ต่ำสุดที่สัดส่วนข้อสอบที่ตรวจให้คะแนนสองค่าและมากกว่าสองค่า คือ 20:80 และความยาวของแบบทดสอบ 50 ข้อ มีปฏิสัมพันธ์ระหว่างโมเดลการตรวจให้คะแนน สัดส่วนของข้อสอบที่ตรวจให้คะแนนสองค่าและมากกว่าสองค่า ความยาวของแบบทดสอบที่ส่งผลต่อค่า θ และ BIAS อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ส่วนปฏิสัมพันธ์รายคู่ ปรากฏว่า ปฏิสัมพันธ์ระหว่างโมเดลการตรวจให้คะแนนกับสัดส่วนของข้อสอบที่ตรวจให้คะแนนสองค่าและมากกว่าสองค่า ระหว่างโมเดลการตรวจให้คะแนนกับความยาวของแบบทดสอบ และระหว่างสัดส่วนของข้อสอบที่ตรวจให้คะแนนสองค่าและมากกว่าสองค่ากับความยาวของแบบทดสอบ ส่งผลต่อค่า θ และ BIAS อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นอกจากนี้พบว่าโมเดลการตรวจให้คะแนน สัดส่วนของข้อสอบที่ตรวจให้คะแนนสองค่าและมากกว่าสองค่า และความยาวของแบบทดสอบที่แตกต่างกันส่งผลต่อค่า θ และ BIAS อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

นุภาพรรณ ปลื้มใจ ปิยะทิพย์ ตินวร และโสพส สุขานนท์สวัสดิ์ (2558) ได้พัฒนาโปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ สำหรับการจัดสอบ O-NET ชั้นมัธยมศึกษาปีที่ 6 จำนวน 61 คน ด้วยรูปแบบของ Web Application โดยทำการทดสอบบนเว็บไซต์ www.onetcat.net/onetM6 ผลการศึกษาปรากฏว่า โปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์สำหรับการจัดสอบ O-NET ชั้นมัธยมศึกษาปีที่ 6 มีความถูกต้องและปลอดภัยในการใช้งาน เป็นที่ยอมรับของผู้เชี่ยวชาญและนักเรียนกลุ่มที่ทดลองใช้ประเมินโปรแกรมว่ามีความสะดวกในการใช้งาน

สุชาดา กรเพชรปาณี ปิยะทิพย์ ตินวร และโสพส สุขานนท์สวัสดิ์ (2559) ได้พัฒนาโปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ สำหรับการจัดสอบ O-NET ระดับชั้นประถม ศึกษาปีที่ 6 ชั้นมัธยมศึกษาปีที่ 3 และชั้นมัธยมศึกษาปีที่ 6 ทั้ง 8 กลุ่มสาระการเรียนรู้ โดยใช้ข้อมูลระหว่างปี 2551-2553 จากสถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน) เป็นข้อมูลแบบทุติยภูมิมีการวิเคราะห์คุณภาพของข้อสอบตามทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) แบบ 3 พารามิเตอร์ ทั้ง 3 ระดับชั้น โดยข้อสอบในคลังข้อสอบ O-NET อยู่ในระดับค่อนข้างยาก เป็นการพัฒนาโปรแกรมในรูปแบบของ Web Application ผู้ใช้สามารถเข้าใช้โปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ สำหรับการจัดสอบ O-NET ใน website://www. onetcat.net ผลการศึกษาปรากฏว่า การใช้งานของโปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ สำหรับการจัดสอบ O-NET อยู่ในเกณฑ์ดี เป็นที่พึงพอใจของนักเรียน

จารุจิตร สิทธิปรุ ปิยะทิพย์ ตินวร และโสพส สุขานนท์สวัสดิ์ (2559) ได้พัฒนาโปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ สำหรับการจัดสอบ O-NET ชั้นมัธยมศึกษาปีที่ 3 ในรูปแบบ ของ Web application วิเคราะห์คุณภาพของข้อสอบตามทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) แบบ 3 พารามิเตอร์ โดยใช้โปรแกรมสำเร็จรูป Xcalibre Version 4.1.7

จัดทำคลังข้อสอบ O-NET จำนวน 8 กลุ่มสาระการเรียนรู้ ระหว่างปี พ.ศ. 2551-2553 คุณภาพของข้อสอบอยู่ในระดับค่อนข้างยาก โปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ สำหรับการจัดสอบ O-NET ระดับชั้นมัธยมศึกษาปีที่ 3 มีประสิทธิภาพอยู่ในเกณฑ์ดี เป็นที่ยอมรับของผู้เชี่ยวชาญและนักเรียนกลุ่มตัวอย่างที่ทดลองใช้โปรแกรม ทั้งในด้านลักษณะทั่วไปและความสะดวกในการใช้โปรแกรม

สยามรัก สว่างศรี วราพร เอราวรรณ์ และทัศน์ศิรินทร์ สว่างบุญ (2560) ได้ศึกษาการสร้างแบบทดสอบวินิจฉัยทางการเรียนกลุ่มสาระการเรียนรู้ภาษาต่างประเทศ เรื่อง ความสามารถด้านไวยากรณ์วิชาภาษาอังกฤษ โดยใช้การสอบทางคอมพิวเตอร์ กลุ่มตัวอย่างเป็นนักเรียนชั้นมัธยมศึกษาปีที่ 2 ปีการศึกษา 2555 สังกัดสำนักงานเขตพื้นที่การศึกษามัธยมศึกษา เขต 27 จำนวน 1,271 คน เครื่องมือที่ใช้ในการวิจัย ประกอบด้วย 4 ด้าน ได้แก่ 1) แบบทดสอบเพื่อสำรวจข้อบกพร่องทางการเรียนรู้ เป็นแบบทดสอบชนิดเติมคำ จำนวน 80 ข้อ ใช้ทดสอบกับกลุ่มทดลอง จำนวน 155 คน 2) แบบทดสอบวินิจฉัยทางการเรียนกลุ่มสาระการเรียนรู้ภาษาต่างประเทศ เรื่อง ความสามารถด้านไวยากรณ์วิชาภาษาอังกฤษ เป็นชนิดเลือกตอบ 4 ตัวเลือก จำนวน 80 ข้อ ทดสอบครั้งที่ 1 กับกลุ่มทดลอง จำนวน 147 คน เพื่อหาคุณภาพข้อสอบรายข้อ ตามทฤษฎีการทดสอบแบบดั้งเดิม หาค่าความยาก ค่าอำนาจจำแนกรายข้อ และค่าความเชื่อมั่น 3) แบบทดสอบวินิจฉัยทางการเรียนกลุ่มสาระการเรียนรู้ภาษาต่างประเทศ เรื่องความสามารถด้านไวยากรณ์วิชาภาษาอังกฤษ จำนวน 60 ข้อ ทดสอบ ครั้งที่ 2 กับกลุ่มทดลอง จำนวน 599 คน เพื่อหาคุณภาพข้อสอบรายข้อตามทฤษฎีการตอบสนองข้อสอบ 4) แบบทดสอบวินิจฉัยทางการเรียน กลุ่มสาระการเรียนรู้ภาษาต่างประเทศ เรื่อง ความสามารถด้านไวยากรณ์วิชาภาษาอังกฤษ จำนวน 40 ข้อ เพื่อบรรจุไว้ในโปรแกรมคอมพิวเตอร์ใช้ในการวินิจฉัยข้อบกพร่องทางการเรียนของนักเรียนชั้นมัธยมศึกษาปีที่ 2 ผลการศึกษาปรากฏว่า คุณภาพของข้อสอบตามทฤษฎีการทดสอบแบบดั้งเดิม มีค่าอำนาจจำแนกของตัวถูก ตั้งแต่ 0.21 ถึง 0.66 ค่าความยากของตัวถูก ตั้งแต่ 0.28 - 0.71 ค่าอำนาจจำแนกของตัวลวง ตั้งแต่ 0.05 ถึง 0.50 และค่าความยากของตัวลวง ตั้งแต่ 0.05 - 0.49 ค่าความเชื่อมั่นทั้งฉบับ เท่ากับ 0.864 ส่วนการวิเคราะห์คุณภาพของข้อสอบตามทฤษฎีการตอบสนองข้อสอบ มีค่าอำนาจจำแนก (a) อยู่ระหว่าง 0.302 - 2.818 มีค่าความยาก (b) อยู่ระหว่าง -2.913 - 2.976 มีค่าโอกาสการเดาถูก (c) อยู่ระหว่าง 0.101 - 0.298 การทดสอบด้วยโปรแกรมการทดสอบวินิจฉัยทางการเรียนกลุ่มสาระการเรียนรู้ภาษาต่างประเทศ เรื่องความสามารถด้านไวยากรณ์วิชาภาษา อังกฤษ ของนักเรียนชั้นมัธยมศึกษาปีที่ 2 โดยใช้การทดสอบผ่านคอมพิวเตอร์นั้น ผู้เชี่ยวชาญ ครูผู้สอนภาษาอังกฤษและนักเรียน มีความคิดเห็นเกี่ยวกับการทดสอบผ่านคอมพิวเตอร์ว่ามีความถูกต้องเหมาะสมสามารถนำไปใช้งานได้อยู่ในระดับมาก

สุธาทิพย์ ตรีสิน และปิยะทิพย์ ประดุจพรม (2560) เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ระดับชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2556 ทั้ง 3 ด้าน ได้แก่ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR ปรากฏว่า แบบทดสอบระดับชาติ มีค่าอำนาจจำแนกของข้อสอบ (a) อยู่ในระดับที่สามารถจำแนกผู้เข้าสอบได้ดี มีค่าความยากของข้อสอบ (b) อยู่ในระดับค่อนข้างยาก และมีค่าโอกาสการเดาของข้อสอบ (c) ไม่เกิน .30

Muninsakorn Tinnaworn and Sukhanonsawat (2015) ได้พัฒนาโปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ สำหรับการจัดสอบ O-NET ชั้นประถมศึกษาปีที่ 6 จำนวน 8 กลุ่มสาระการเรียนรู้ กลุ่มผู้ทดลองใช้เป็นนักเรียนชั้นประถมศึกษาปีที่ 6 จังหวัดชลบุรี จำนวน 30 คน ผลการศึกษาปรากฏว่า ข้อสอบ O-NET ชั้นประถมศึกษาปีที่ 6 มีค่าความยากของข้อสอบ (b) อยู่ในระดับค่อนข้างยาก ซึ่งผ่านการวิเคราะห์คุณภาพของข้อสอบด้วยโมเดลของโลจิส แบบ 3 พารามิเตอร์ โปรแกรมมีความเหมาะสมอยู่ในระดับมากที่สุด ไม่มีปัญหาด้านการนำไปใช้และเป็นที่ยอมรับของผู้ทดลองใช้โปรแกรม

จากการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องกับทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) จะเห็นได้ว่าทฤษฎีการตอบสนองข้อสอบนั้น ได้เข้ามามีบทบาทสำคัญอย่างยิ่งในการสร้างและพัฒนาข้อสอบให้มีคุณภาพ น่าเชื่อถือ นักการศึกษาจึงได้มีการศึกษาวิจัยที่เกี่ยวข้องกับทฤษฎีการตอบสนองข้อสอบเพิ่มขึ้น เพื่อเพิ่มประสิทธิภาพในการพัฒนาแบบทดสอบต่าง ๆ และทำให้เกิดความก้าวหน้าทางการวัดและประเมินผลศึกษามากยิ่งขึ้น

ตอนที่ 3 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) และงานวิจัยที่เกี่ยวข้อง

ความหมายของการทำหน้าที่ต่างกันของข้อสอบ

ในการศึกษาเรื่องผลการสอบของกลุ่มผู้เข้าสอบย่อยจากกลุ่มผู้เข้าสอบทั้งหมดมีการศึกษามานานแล้ว แต่เพิ่งมีการศึกษาเรื่องของความยุติธรรมในการสอบระหว่างผู้เข้าสอบย่อยต่างกลุ่มกันอย่างจริงจัง ในช่วงปลายทศวรรษที่ 1960 โดยมีการนำเสนอวิธีการต่าง ๆ ในการตรวจสอบความลำเอียงของแบบทดสอบ (Test Bias) และความลำเอียงในการคัดเลือกผู้ที่จะเข้าสอบ (Selection Bias) เพิ่มขึ้นหลายวิธี ในช่วงเวลานั้น นักพัฒนาแบบทดสอบมีความสนใจวิธีการจำแนกข้อสอบที่ไม่เหมาะสมกับผู้เข้าสอบบางกลุ่มออกจากแบบทดสอบ ก่อนที่จะมีการพัฒนาให้เป็นแบบทดสอบฉบับสมบูรณ์ จึงมีการพัฒนาวิธีการตรวจสอบความลำเอียงของข้อสอบ (Item Bias) เพื่อใช้ในการจำแนกข้อสอบที่มีความลำเอียงกับกลุ่มผู้เข้าสอบบางกลุ่มที่มีลักษณะบางอย่างแตกต่างกัน เช่น เชื้อชาติ ศาสนา วัฒนธรรม ภูมิลำเนาสังคม เพศ ภาษา อายุและประสบการณ์ เป็นต้น เพื่อเป็นการพัฒนาแบบทดสอบให้มีคุณภาพที่เหมาะสมสามารถนำไปใช้ในการทดสอบต่อไปได้

การศึกษาเรื่องผลการสอบในช่วงแรก ๆ มีวัตถุประสงค์เพื่อคัดเลือกคนเข้าศึกษาต่อหรือเข้าทำงาน แต่มีหลักฐานปรากฏอย่างชัดเจนที่แสดงให้เห็นว่ามีความลำเอียงเกิดขึ้นกับกลุ่มคนต่างชาติ เพศ ทำให้ต้องมีการศึกษา ความลำเอียงในการคัดเลือกผู้เข้าสอบ เพื่อให้การศึกษานี้มีความถูกต้อง ชัดเจนเพิ่มมากขึ้น ในเวลาต่อมาจึงมีการศึกษาในระดับข้อสอบ (Item Level) ที่เรียกว่าความลำเอียงของข้อสอบ (Item Bias) แต่ในปัจจุบันนักวิจัยทางการวัดผลส่วนใหญ่ ใช้คำว่า ข้อสอบทำหน้าที่ต่างกันกับกลุ่มผู้เข้าสอบย่อยต่างกลุ่มกัน หรือเรียกสั้น ๆ ว่า ข้อสอบทำหน้าที่ต่างกัน (Differential Item Functioning: DIF) โดยเห็นว่าเป็นคำที่มีความหมายกลาง ๆ และมีความเหมาะสมในเชิงวิชาการมากกว่าคำว่าความลำเอียง (Bias) เป็นคำที่ใช้กันในทางสังคมและมีความหมายในเชิงลบ แต่อย่างไรก็ตามคำสองคำนี้มีจุดเน้นที่แตกต่างกัน โดยคำว่าความลำเอียงของข้อสอบ จะเน้นที่อิทธิพลที่สังเกตได้ของกลุ่มผู้เข้าสอบย่อยที่มุ่งศึกษา ส่วนคำว่าข้อสอบที่ทำหน้าที่ต่างกัน เน้นที่ลักษณะทางสถิติของข้อสอบที่ทำการตรวจสอบได้ด้วยวิธีการวิเคราะห์ทางสถิติ ซึ่งเป็นองค์ประกอบหนึ่งที่แสดงถึงความลำเอียง

ของข้อสอบ (Scheuneman & Bleistein, 1989; Angoff, 1993; Hambleton & Others, 1993; Zieky, 1993; Camilli & Shepard, 1994) จากจุดเน้นนี้แสดงให้เห็นว่าวิธีการทางสถิติที่นำมาใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นสิ่งที่มีความจำเป็นในการประเมินความลำเอียงของข้อสอบ แต่ถ้ามีการใช้เฉพาะวิธีการทางสถิติอย่างเดียวผลการตรวจสอบพบข้อสอบทำหน้าที่ต่างกันที่ไม่อาจจะสรุปได้ว่าข้อสอบข้อนั้นมีความลำเอียงหรือไม่ เนื่องจากการประเมินความลำเอียงของข้อสอบยังต้องรวมถึงการใช้วิธีการตัดสินข้อสอบ (Judgmental Method) โดยมีผู้เชี่ยวชาญพิจารณาเนื้อหาสาระของข้อสอบและจุดมุ่งหมายในการวัดของแบบทดสอบก่อนที่จะสรุปว่าข้อสอบข้อนั้นมีความลำเอียงหรือไม่

ปัจจุบันนี้นักวิจัยทางการวัดผลหลายท่านใช้คำว่าการทำหน้าที่ต่างกันของข้อสอบ แทนคำว่าความลำเอียงของข้อสอบ ซึ่งมีนักวิจัยทางการวัดผลได้ให้ความหมายของการทำหน้าที่ต่างกันของข้อสอบไว้ดังนี้

Holland and Wainer (1993, p. 453) กล่าวว่า การทำหน้าที่ต่างกันของข้อสอบ หมายถึง สารสนเทศทางสถิติของข้อสอบที่ได้จากกลุ่มผู้เข้าสอบต่างกลุ่มกันและมีความสามารถเท่ากัน แต่มีโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกัน

Camilli and Shepard (1994, p. 174) กล่าวว่า การทำหน้าที่ต่างกันของข้อสอบ หมายถึง การตรวจสอบความเป็นพหุมิติในการวัดของข้อสอบ ซึ่งแสดงได้จากการแจกแจงความสามารถหลัก (Primary Ability) ของกลุ่มผู้สอบตั้งแต่ 2 กลุ่มขึ้นไปมีความเท่ากันแต่มีการแจกแจงความสามารถรอง (Secondery Ability) แตกต่างกัน

Narayanan and Swaminathan (1996, pp. 257-274) กล่าวว่า การทำหน้าที่ต่างกันของข้อสอบ หมายถึง ผู้สอบมีความสามารถระดับเดียวกัน แต่มาจากกลุ่มย่อยแตกต่างกัน มีโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกัน

มีผู้ให้ความหมายของคำว่า การทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) ไว้หลายความหมายดังกล่าวไว้แล้วข้างต้น แต่ความหมายที่เป็นที่ยอมรับกันอย่างกว้างขวาง คือ ข้อสอบทำหน้าที่ต่างกันภายใต้เงื่อนไขผู้เข้าสอบที่มีความสามารถเท่ากัน แต่มาจากกลุ่มผู้สอบย่อยที่มีลักษณะต่างกัน มีความน่าจะเป็นในการตอบข้อสอบข้อนั้นไม่เท่ากัน

ดังนั้นจึงสรุปได้ว่า การทำหน้าที่ต่างกันของข้อสอบ (DIF) หมายถึง การที่ข้อสอบทำให้ผู้เข้าสอบจากกลุ่มต่างกันที่มีความสามารถหรือคุณลักษณะที่มุ่งวัดเท่ากัน มีโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกัน เป็นการเปรียบเทียบผลการตอบระหว่างผู้เข้าสอบ 2 กลุ่ม คือ กลุ่มอ้างอิง (Reference Group: R) และกลุ่มเปรียบเทียบ (Focal Group: F) กลุ่มอ้างอิง เป็นกลุ่มที่คาดว่าจะได้ประโยชน์จากการตอบข้อสอบ มีโอกาสในการตอบข้อสอบถูกได้มากกว่าผู้เข้าสอบกลุ่มเปรียบเทียบและกลุ่มเปรียบเทียบเป็นกลุ่มที่คาดว่าจะเสียประโยชน์จากการตอบข้อสอบ มีโอกาสตอบข้อสอบถูกได้น้อยกว่าผู้เข้าสอบกลุ่มอ้างอิง การทำหน้าที่ต่างกันของข้อสอบเกิดขึ้น เมื่อนำข้อสอบไปทดสอบกับผู้เข้าสอบกลุ่มย่อยต่างกันที่มีความสามารถหลัก (Primary Ability) เท่ากันหรือมีคุณลักษณะแฝง (Secondery Ability) แตกต่างกัน ทำให้ผู้เข้าสอบต่างกลุ่มที่นำมาจับคู่เปรียบเทียบมีโอกาสตอบข้อสอบถูกแตกต่างกัน

การทดสอบแต่ละครั้งผู้สอบระหว่างกลุ่มย่อยอาจมีลักษณะที่แตกต่างกัน เช่น เชื้อชาติ ศาสนา วัฒนธรรม ภูมิลำเนา สังคม เพศ ภาษา อายุ ประสบการณ์ เป็นต้น ผู้สอบกลุ่มย่อยอาจไม่ได้รับความยุติธรรมในการทำข้อสอบ โดยข้อสอบบางข้ออาจมีความลำเอียงเข้าข้างผู้สอบกลุ่มย่อยบางกลุ่มของผู้สอบทั้งหมด ซึ่งทำให้เกิดการได้เปรียบเสียเปรียบระหว่างผู้สอบกลุ่มย่อยด้วยกัน ทั้ง ๆ ที่สอบด้วยข้อสอบฉบับเดียวกัน สาเหตุดังกล่าวอาจเนื่องมาจากแบบสอบไม่ได้วัดความสามารถเป้าหมายที่ต้องการวัดเพียงอย่างเดียว แต่ยังวัดความสามารถแทรกซ้อนที่ไม่ต้องการวัดอีกด้วย ตัวอย่างเช่น แบบสอบวัดคำศัพท์ในวิชาภาษาอังกฤษฉบับหนึ่ง ข้อสอบบางข้ออาจถามความรู้ สำหรับผู้ชายเป็นพิเศษ เช่น ความรู้เรื่องกีฬา ในขณะที่ข้อสอบบางข้ออาจถามความรู้สำหรับผู้หญิงโดยเฉพาะ เช่น ความรู้เกี่ยวกับงานในบ้าน จากสถานการณ์ดังกล่าว ทักษะวัดคำศัพท์ในวิชาภาษาอังกฤษเป็นความสามารถเป้าหมาย (θ) ส่วนทักษะวัดความรู้ด้านกีฬา (η_1) และงานในบ้าน (η_2) เป็นความสามารถ แทรกซ้อน ข้อสอบทุกข้อในแบบสอบจะวัดความสามารถเป้าหมาย ส่วนข้อสอบบางข้อที่ทำหน้าที่ต่างกันจะวัดทั้งความสามารถเป้าหมายและความสามารถแทรกซ้อน นั่นคือ ถ้าผู้สอบกลุ่มย่อยกลุ่มใด มีความแทรกซ้อนสูงกว่าก็มีโอกาสในการตอบข้อสอบได้ถูกต้องมากกว่า ทั้ง ๆ ที่ระดับความสามารถเป้าหมายที่ต้องการวัดเท่ากัน จึงมีผลทำให้ข้อสอบทำหน้าที่ต่างกัน

การศึกษาถึงคุณภาพของข้อสอบจากผลการตรวจสอบข้อสอบของผู้สอบกลุ่มต่าง ๆ ในประชากรมีมานานแล้ว แต่การศึกษาคุณภาพด้านความยุติธรรมของข้อสอบหรือแบบสอบระหว่างผู้สอบกลุ่มต่าง ๆ เริ่มศึกษากันอย่างจริงจังในช่วงปลายทศวรรษของปี ค.ศ. 1960 มีเสนอวิธีการต่าง ๆ เพื่อตรวจสอบความลำเอียงของข้อสอบ (Item Bias) ความลำเอียงของแบบสอบ (Test Bias) และความลำเอียงในการคัดเลือก (Selection Bias) โดยนิยามความลำเอียงว่าเป็น ความคลาดเคลื่อนอย่างเป็นระบบ (Systematic Error) ที่เกิดขึ้นจากการวัด ความพยายามของการตรวจสอบความลำเอียงดังกล่าว ดำเนินไปเพื่อจำแนกข้อสอบ ที่ทำหน้าที่ไม่เหมาะสมหรือไม่ยุติธรรมสำหรับปรับปรุง หรือตัดข้อสอบนั้นออกจากแบบสอบ เป็นการขจัดข้อสอบที่ทำให้เกิดปัญหาความยุติธรรมระหว่างกลุ่มข้อสอบกลุ่มต่าง ๆ ที่มีลักษณะบางอย่างแตกต่างกัน เช่น เชื้อชาติ ศาสนา วัฒนธรรม ภูมิลำเนา สังคม เพศ ภาษา อายุ ประสบการณ์ เป็นต้น เพื่อพัฒนาแบบสอบให้มีคุณภาพเหมาะสมสำหรับนำไปใช้ทดสอบต่อไป (ศิริชัย กาญจนวาสี, 2555, หน้า 115)

ในเวลาต่อมา นักวัดผลการศึกษาได้ทำการศึกษาความลำเอียงของข้อสอบ (Item Bias) กันอย่างกว้างขวาง ทำให้เกิดความสับสนของการใช้คำและความหมาย มีประเด็นโต้แย้งกันว่าความลำเอียงของข้อสอบ เป็นผลการตัดสินว่าข้อสอบมีความยุติธรรมหรือไม่ อันส่งผลต่อการบรรลุจุดมุ่งหมายของการใช้แบบทดสอบหรือความลำเอียงของข้อสอบ เป็นสารสนเทศทางสถิติที่ได้จากข้อสอบเกี่ยวกับกลุ่มผู้สอบต่างกลุ่มกันตอบข้อสอบข้อเดียวกัน ความแตกต่างที่เกิดขึ้นอาจมาจากความไม่เหมาะสมของข้อคำถาม ซึ่งสามารถเกิดขึ้นได้หลายลักษณะ หรือประสบการณ์ของผู้สอบซึ่งอาจมีลักษณะพื้นฐานเดิมแตกต่างกันในหลายสถานการณ์จึงไม่เหมาะสมที่จะใช้คำว่า ข้อสอบลำเอียง (Biased Item) เนื่องจากเป็นภาษาที่มีความหมายในเชิงลบ ประกอบกับเกณฑ์ที่ใช้สำหรับตัดสินความลำเอียงยังมีความคลุมเครือและค่อนข้างสับสน ดังนั้น จึงควรเปลี่ยนมาใช้คำว่า การทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) ซึ่งเป็นคำที่มีความเป็นกลางและความเหมาะสมกว่า (Holland & Thayer, 1988; Holland & Wainer, 1993)

การทำหน้าที่ต่างกันของข้อสอบ (DIF) กับความลำเอียงของข้อสอบ (Item Bias) มีแนวคิดที่แตกต่างกัน สำหรับการทำหน้าที่ต่างกันของข้อสอบ เป็นกระบวนการที่เน้นการใช้วิธีทางการสถิติสำหรับการตรวจสอบ เพื่อให้ได้สารสนเทศเกี่ยวกับการทำหน้าที่ของข้อสอบสำหรับกลุ่มผู้สอบกลุ่มย่อยที่มีลักษณะเฉพาะบางอย่างแตกต่างกัน ส่วนความลำเอียงของข้อสอบ เป็นกระบวนการตัดสินความยุติธรรมของข้อสอบ โดยนำสารสนเทศการทำหน้าที่ต่างกันของข้อสอบมาวิเคราะห์เชิงตรรกะ (Logical Analysis) โดยผู้เชี่ยวชาญพิจารณาถึงการเขียนข้อสอบ เนื้อหาสาระของข้อสอบและจุดมุ่งหมายของการวัด เพื่อระบุว่าข้อสอบข้อนั้นลำเอียงเข้าข้างกลุ่มใดหรือไม่ เพราะเหตุใดจึงเป็นการตัดสินความลำเอียงของข้อสอบ (Camill & Shapard, 1994)

จากการศึกษาเอกสารและงานวิจัยผู้วิจัยสามารถสรุปได้ว่า การทำหน้าที่ต่างกันของข้อสอบแต่เดิมใช้คำว่า “ความลำเอียงของข้อสอบ” (Item Bias) ซึ่งเป็นภาษาที่ใช้กันในทางสังคมและมีความหมายในทางลบ แต่ระยะหลังนักวิจัยได้เปลี่ยนไปใช้คำใหม่ว่า “การทำหน้าที่ต่างกันของข้อสอบ” (Differential Item Functioning: DIF) แต่อย่างไรก็ตามคำสองคำนี้มีจุดเน้นที่แตกต่างกัน คำว่า “ความลำเอียงของข้อสอบ” เน้นที่อิทธิพลที่สังเกตได้ของกลุ่มผู้สอบย่อยที่มุ่งศึกษา ส่วนคำว่า “ข้อสอบที่ทำหน้าที่ต่างกัน” นั้นเน้นคุณลักษณะทางสถิติของข้อสอบที่ตรวจสอบได้ด้วยวิธีการวิเคราะห์ทางสถิติ สำหรับสิ่งที่ต้องให้ความสำคัญลำดับต่อมาได้แก่ ประเภทของการทำหน้าที่ต่างกันของข้อสอบ

ประเภทของการทำหน้าที่ต่างกันของข้อสอบ

การทำหน้าที่ต่างกันของข้อสอบ (DIF) เป็นการเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มผู้สอบอย่างน้อย 2 กลุ่มขึ้นไป ปกตินิยมทำการเปรียบเทียบ 2 กลุ่ม ประกอบด้วยกลุ่มแรกเรียกว่า กลุ่มเปรียบเทียบ (Focal Group หรือกลุ่ม F) เป็นกลุ่มที่สนใจศึกษาและคาดว่าจะเป็นกลุ่มที่เสียเปรียบในการตอบข้อสอบ และกลุ่มที่สองเรียกว่ากลุ่มอ้างอิง (Reference Group หรือกลุ่ม R) เป็นกลุ่มที่คาดว่าจะได้เปรียบในการตอบข้อสอบได้ถูกต้อง

ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ จะพบว่า ข้อสอบสามารถทำหน้าที่ต่างกันได้ 2 ประเภท (Mellenbergh, 1982) ได้แก่ การทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป (Uniform) และแบบอเนกรูป (Nonuniform) (ศิริชัย กาญจนวาสี, 2555, หน้า 118-119)

1. ข้อสอบทำหน้าที่ต่างกันแบบเอกรูป (Uniform DIF) หมายถึง ข้อสอบที่ทำให้ผู้สอบกลุ่มหนึ่งมีโอกาสในการตอบข้อสอบถูกมากกว่าผู้สอบอีกกลุ่มหนึ่งอย่างสม่ำเสมอกัน ในทุกระดับความสามารถ เมื่อพิจารณาโค้งคุณลักษณะข้อสอบของผู้สอบ 2 กลุ่ม จะไม่พบว่ามีปฏิสัมพันธ์ระหว่างระดับความสามารถของผู้สอบกับการเป็นสมาชิกของกลุ่ม (Group Membership)

2. ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป (Nonuniform DIF) หมายถึง ข้อสอบที่ทำให้โอกาสในการตอบข้อสอบถูกของผู้สอบระหว่างกลุ่มแตกต่างกันอย่างไม่สม่ำเสมอกันในทุกระดับความสามารถ เมื่อพิจารณาโค้งคุณลักษณะข้อสอบของผู้สอบสองกลุ่ม พบว่ามีปฏิสัมพันธ์ร่วมกันระหว่างความสามารถของผู้สอบ กับการเป็นสมาชิกของกลุ่ม เช่น ที่ระดับความสามารถหนึ่ง กลุ่มผู้สอบกลุ่ม R มีโอกาสในการตอบข้อสอบถูกมากกว่ากลุ่มผู้สอบกลุ่ม F แต่ที่ระดับความสามารถอีกระดับหนึ่งกลุ่มผู้สอบกลุ่ม F มีโอกาสในการตอบข้อสอบถูกมากกว่ากลุ่มผู้สอบกลุ่ม R

ตามทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) สามารถพิจารณา “ปฏิสัมพันธ์” ดังกล่าวได้จากความแตกต่างค่าพารามิเตอร์อำนาจจำแนกของข้อสอบ ระหว่างผู้สอบ

กลุ่มย่อยสองกลุ่ม กล่าวคือ ถ้าข้อสอบทำหน้าที่ต่างกันแบบเอกรูป แล้วโค้งลักษณะข้อสอบ (Item Characteristic Curves: ICCs) ระหว่างผู้สอบกลุ่มย่อยสองกลุ่มจะขนานกัน หรือมีฟังก์ชั่นการตอบสนองข้อสอบ (Item Response Functions: IRFs) เหมือนกัน แต่ถ้าข้อสอบมีหน้าที่ต่างกันแบบอเนกรูป แล้วโค้งลักษณะข้อสอบระหว่างผู้สอบกลุ่มย่อยสองกลุ่มจะไม่ขนานกัน หรือมีฟังก์ชันการตอบสนองข้อสอบต่างกัน ดังนั้นความแตกต่างระหว่างโค้งลักษณะข้อสอบทั้งสองแบบจะบ่งบอกถึงขนาดและทิศทางของข้อสอบที่ทำหน้าที่ต่างกัน ซึ่งสามารถคำนวณได้โดยใช้สูตรการคำนวณพื้นที่ของ Raju (1990)

ข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูป สามารถจำแนกได้เป็น 2 ลักษณะ (Swaminathan & Rogers, 1990) ดังนี้

1. ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปโดยมีปฏิสัมพันธ์ไม่เป็นลำดับ (Disordinal interaction) เป็นการทำหน้าที่ต่างกันสำหรับกลุ่มผู้สอบซึ่งเกิดขึ้น เมื่อโค้งลักษณะข้อสอบตัดกันระหว่างช่วงความสามารถของผู้สอบหรือเรียกว่าข้อสอบทำหน้าที่ต่างกันแบบไม่มีทิศทาง (Non-Unidirectional DIF)

2. ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปโดยมีปฏิสัมพันธ์เป็นลำดับ (Ordinal Interaction) เป็นการทำหน้าที่ต่างกันสำหรับกลุ่มผู้สอบซึ่งเกิดขึ้น เมื่อโค้งลักษณะข้อสอบต่างกันอย่างไม่สม่ำเสมอ แต่ไม่ตัดกัน หรืออาจตัดกันนอกช่วง ความสามารถของผู้สอบตรงปลายสุดของช่วงความสามารถต่ำหรือสูงอาจเรียกข้อสอบลักษณะนี้ว่า ข้อสอบทำหน้าที่ต่างกันแบบมีทิศทางเดียว (Unidirectional DIF)

โดยทั่วไปในแบบสอบมาตรฐานมักจะมีข้อสอบที่ต่างกันแบบเอกรูปมากกว่าข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูป แต่ในข้อมูลจริงจะมีข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปได้มากกว่า จะเห็นได้ว่าประเภทของการทำหน้าที่ต่างกันของข้อสอบแบ่งเป็น 2 ได้แก่ การทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป (Uniform DIF) เกิดขึ้นเมื่อผู้สอบกลุ่มหนึ่งมีโอกาสในการตอบข้อสอบถูกมากกว่าผู้สอบอีกกลุ่มหนึ่งในทุกระดับความสามารถ และการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป (Nonuniform DIF) เกิดขึ้นเมื่อโอกาสในการตอบข้อสอบถูกของผู้สอบระหว่างกลุ่มย่อย 2 กลุ่ม ไม่สม่ำเสมอ

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

แนวคิดเกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

การศึกษาเรื่องความยุติธรรมของข้อสอบ ในกรณีที่ข้อสอบทำให้ผู้สอบระหว่างกลุ่มย่อยเกิดการได้เปรียบเสียเปรียบกัน เดิมใช้คำว่า “ความลำเอียงของข้อสอบ” (Item Bias) ซึ่งเป็นภาษาที่ใช้กันในทางสังคมและมีความหมายในทางลบ แต่ระยะหลังนักวิจัยได้เปลี่ยนไปใช้คำใหม่ว่า “การทำหน้าที่ต่างกันของข้อสอบ” (Differential item Functioning: DIF) เนื่องจากเห็นว่าเป็นคำที่มีความหมายเป็นกลาง จึงมีความเหมาะสมเชิงวิชาการมากกว่า คำสองคำนี้มีจุดเนินที่แตกต่างกันที่คำว่า “การทำหน้าที่ต่างกันของข้อสอบ” เน้นที่คุณลักษณะทางสถิติของข้อสอบที่ตรวจสอบได้ด้วยวิธีการวิเคราะห์ทางสถิติ ซึ่งเป็นส่วนประกอบหนึ่งของสิ่งที่แสดงถึงความลำเอียงของข้อสอบ วิธีการทางสถิติที่ใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นเงื่อนไขที่จำเป็นในการตัดสินความลำเอียงของข้อสอบ เนื่องจากถ้าใช้วิธีการทางสถิติตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเพียงอย่างเดียวแล้วผลการตรวจสอบพบว่าข้อสอบทำหน้าที่ต่างกันนั้นยังสรุปไม่ได้ว่าข้อสอบมีความลำเอียงหรือไม่ ต้องให้ผู้เชี่ยวชาญพิจารณาเนื้อหาของข้อสอบและจุดมุ่งหมายในการวัดข้อสอบที่เรียกว่า “วิธีการตัดสินข้อสอบ”(Judgemental Method) (Camilli & Shepard, 1994, p. 135)

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF Detection) เป็นการเปรียบเทียบผลการตอบข้อสอบเป็นรายข้อระหว่างกลุ่มผู้สอบอย่างน้อย 2 กลุ่ม มีความสามารถหลัก (Primary Ability) ที่มุ่งวัดเท่ากัน แต่คาดว่าจะมีความได้เปรียบเสียเปรียบกัน โดยกลุ่มหนึ่งถือเป็น กลุ่มอ้างอิง (Reference Group) ซึ่งคาดว่าน่าจะได้เปรียบในการตอบข้อสอบข้อนั้น หรือมีโอกาสตอบข้อสอบได้ถูกต้องมากกว่า ส่วนอีกกลุ่มคือ กลุ่มเปรียบเทียบ (Focal Group) ซึ่งเป็นกลุ่มที่สนใจศึกษา และคาดว่าน่าจะเป็นกลุ่มที่เสียเปรียบ (ศิริชัย กาญจนวาสี, 2555, หน้า 120-126)

ในการเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบจำเป็นต้องจับคู่ (Matching) ผู้สอบตามความสามารถ ซึ่งเป็นเงื่อนไขสำคัญของการตรวจสอบ การทำหน้าที่ต่างกันของข้อสอบ เกณฑ์การจับคู่ (Matching Criteria) ที่นิยมใช้กันมี 2 วิธี ดังนี้

1. เกณฑ์ภายนอก (External Criterion)

การวิเคราะห์การทำหน้าที่ต่างกัน โดยใช้เกณฑ์ภายนอกนี้ สามารถนำไปใช้ได้ทั้งข้อสอบรายข้อและข้อสอบทั้งฉบับ โดยการใช้คะแนนจากแบบสอบอื่นเป็นเกณฑ์ภายนอกแล้วใช้เทคนิคการวิเคราะห์การถดถอย (Regression Analysis) เพื่อทำการเปรียบเทียบเส้นกราฟความสัมพันธ์ระหว่างตัวแปรเกณฑ์ กับตัวแปรทำนายระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ

หลักการนี้มีจุดมุ่งหมาย เพื่อสร้างสมการทำนายตัวแปรเกณฑ์ ซึ่งเป็นคะแนนของแบบสอบอื่นจากตัวแปรทำนายซึ่งเป็นคะแนนรายข้อ หรือคะแนนแบบสอบระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ จะใช้คะแนนรายข้อเป็นตัวแปรทำนาย แต่ถ้าเป็นการวิเคราะห์การทำหน้าที่ต่างกันของแบบสอบ จะใช้คะแนนรวมของแบบสอบทั้งฉบับเป็นตัวแปรทำนาย สำหรับตัวแปรเกณฑ์ที่ใช้เป็นเกณฑ์ภายนอก อาจใช้คะแนนรวมทั้งฉบับ หรือเกรดเฉลี่ยหรือผลสัมฤทธิ์ในงานที่เกี่ยวข้องของผู้สอบ (Cronbach, 1970)

การใช้เกณฑ์ภายนอกมีข้อดี คือเกณฑ์ที่ใช้มีความเป็นอิสระจากข้อสอบ และแบบสอบที่ต้องการตรวจสอบ แต่มีจุดอ่อนตรงที่ความเหมาะสมของเกณฑ์ที่จะนำมาใช้ ในทางปฏิบัติเป็นการยากที่จะหาตัวแปรเกณฑ์ภายนอกจากแบบสอบฉบับอื่นที่มีความตรงเชิงทำนาย และมีความยุติธรรมสำหรับกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ถ้าตัวแปรเกณฑ์ภายนอกขาดคุณสมบัติดังกล่าว จะทำให้ผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบหรือแบบสอบขาดความแม่นยำ และความสมบูรณ์

2. เกณฑ์ภายใน (Internal Criterion)

การวิเคราะห์การทำหน้าที่ต่างกัน โดยใช้เกณฑ์ภายในเป็นการนำวิธีการทางสถิติมาตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ หรือแบบสอบ โดยเน้นการพิจารณาจากโครงสร้างภายในของแบบสอบเป็นหลัก ด้วยการวิเคราะห์ผลจากการตอบข้อสอบและความสามารถหรือคะแนนจริงของผู้สอบที่ได้จากแบบทดสอบฉบับนั้น เพื่อนำมาเปรียบเทียบระหว่างผู้สอบจากกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ที่มีความสามารถหรือคะแนนจริงเท่ากัน ว่าจะมีผลการตอบหรือโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกันหรือไม่ เพื่อบ่งชี้การทำหน้าที่ต่างกันของข้อสอบ การวิเคราะห์ในลักษณะนี้นิยมใช้ค่าสถิติต่างๆ เป็นตัวบ่งชี้ถึงการทำหน้าที่ต่างกันของข้อสอบ ค่าสถิติทดสอบที่นิยมนำมาใช้ พอสรุปได้ดังนี้

2.1 การทดสอบปฏิสัมพันธ์ (Interaction)

ในระยะเริ่มแรกของการศึกษาความลำเอียงของข้อสอบ มีการใช้สถิติทดสอบเอฟ (F-test) จากการวิเคราะห์ความแปรปรวน (ANOVA) เพื่อทดสอบปฏิสัมพันธ์ระหว่างกลุ่มผู้สอบกับข้อสอบ

ถ้าการทดสอบมีนัยสำคัญเป็นสัญญาณการทำหน้าที่ต่างกันของข้อสอบ (Cleary & Hilton,1968; Jensen, 1974) จากนั้นจึงทำการวิเคราะห์ต่อด้วยวิธีการ Post Hoc เพื่อระบุข้อสอบที่มีผลต่อการเกิดปฏิสัมพันธ์ ซึ่งเป็นข้อที่ทำหน้าที่ต่างกัน วิธีการนี้มีข้อดีที่สามารถศึกษาผู้สอบหลาย ๆ กลุ่มได้สะดวก แต่มีจุดอ่อนในเรื่องการควบคุมกลุ่มต่าง ๆ ให้มีความสามารถที่ทัดเทียมกันขนาดกลุ่มตัวอย่างของกลุ่มต่าง ๆ และอัตราความคลาดเคลื่อนประเภทที่ 1 จะสูงขึ้นถ้าจำนวนข้อสอบเพิ่มมากขึ้น

2.2 การวัดความเบี่ยงเบนสัมพัทธ์ (Relative Deviation)

การคำนวณค่าความยากของข้อสอบ เช่น p, b เป็นต้น เมื่อคำนวณแยกระหว่างกลุ่มและแปลงให้เป็นค่าความยากมาตรฐาน สามารถนำมาพล็อตเปรียบเทียบเป็นรายข้อ ถ้าข้อใดเบี่ยงเบนไปจากแกนหลักที่คาดหมาย หรือเบี่ยงเบนเกินจากความคลาดเคลื่อนมาตรฐานของค่าความยากที่กำหนดย่อมแสดงถึงการทำหน้าที่ต่างกันของข้อสอบ (Cleary & Hilton, 1968; Angoff & Ford, 1973) รวมทั้งสามารถคำนวณค่าสหสัมพันธ์ระหว่างค่าความยากรายข้อระหว่างกลุ่ม เพื่อแสดงถึงการทำหน้าที่ต่างกันของแบบสอบ ค่าสหสัมพันธ์เข้าใกล้ 1.00 แสดงว่าค่าความยากสัมพัทธ์ของข้อสอบมีค่าใกล้เคียงกันระหว่างกลุ่ม ดังนั้นแบบสอบวัดคุณลักษณะคล้ายกันระหว่างกลุ่ม

วิธีการนี้มีข้อดีและข้อเสียคล้ายการทดสอบปฏิสัมพันธ์ นอกจากนี้ค่าความยากของข้อสอบ (p) มิใช่ตัวแทนของค่าความยากจริงของข้อสอบ และได้รับอิทธิพลจากค่าแทรกซ้อนอื่น ได้แก่ ค่าอำนาจจำแนก และความสามารถของผู้สอบ

2.3 การเปรียบเทียบน้ำหนักตัวประกอบ (Factor Loading)

การวิเคราะห์ตัวประกอบ (Factor Analysis) เป็นเทคนิคทางสถิติที่นิยมใช้ในการตรวจสอบความตรงเชิงทฤษฎีหรือโครงสร้าง (Construct Validity) เมื่อนำการวิเคราะห์ตัวประกอบมาใช้ในการวิเคราะห์โครงสร้างของแบบสอบแยกตามกลุ่มสอบ ความไม่สอดคล้องกันระหว่างน้ำหนักตัวประกอบบนคุณลักษณะสำคัญที่มุ่งวัด หรือ ความแตกต่างของค่าเฉลี่ยคะแนน ตัวประกอบ (Factor Scores) ระหว่างกลุ่มผู้สอบ ย่อมสะท้อนการทำหน้าที่ต่างกันของข้อสอบและแบบสอบ

การใช้เทคนิคการวิเคราะห์ตัวประกอบเชิงสำรวจ (Exporatory Factor Analysis: EFA) สำหรับศึกษาการทำหน้าที่ต่างกัน จะมีจุดอ่อนในเรื่องความไม่สอดคล้องระหว่างน้ำหนักตัวประกอบอาจเกิดความแตกต่างของความสามารถระหว่างกลุ่มก็ได้ แนวทางที่เหมาะสมจึงควรใช้เทคนิคการวิเคราะห์ตัวประกอบเชิงยืนยัน (Confirmatory Factor Analysis: CFA) นอกจากนี้ยังสามารถใช้ CFA สำหรับตรวจสอบความแตกต่างระหว่างกลุ่ม ในด้านคุณลักษณะหรือความสามารถหลักและความสามารถรองได้อีกด้วย (Camilli & Shepard, 1994, p. 135)

2.4 การเปรียบเทียบโอกาสการตอบข้อสอบถูก

การวิเคราะห์โอกาสตอบข้อสอบถูกของผู้สอบจากกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่มีความสามารถเท่ากัน เป็นแนวทางสำคัญที่นิยมใช้กันและเป็นที่ยอมรับในปัจจุบัน สำหรับบ่งชี้การทำหน้าที่ต่างกันของข้อสอบ มีการคำนวณค่าสถิติ 2 แนวทางดังนี้

2.4.1 เปรียบเทียบค่าสัดส่วนและความน่าจะเป็นในการตอบข้อสอบถูกของผู้สอบต่างกลุ่มที่มีความสามารถเท่ากัน เช่น วิธีแมนเทล-แฮนส์เซล (MH) เป็นต้น

2.4.2 เปรียบเทียบค่าฟังก์ชันการตอบสนองข้อสอบ หรือโค้งลักษณะข้อสอบระหว่างกลุ่มที่มีระดับความสามารถเท่ากัน เป็นวิธีที่อยู่บนพื้นฐานของทฤษฎี IRT เช่น วิธีวัดความแตกต่างของพื้นที่ วิธีวัดความแตกต่างค่าพารามิเตอร์ความยาก วิธีทดสอบไค-สแควร์ของลอร์ด เป็นต้น

วิธีการนี้มีข้อดีที่สำคัญได้แก่ การคำนวณค่าสถิติของข้อสอบมีความน่าเชื่อถือ มีกลไกควบคุมความสามารถของผู้สอบโดยการจับคู่กลุ่มความสามารถ เพื่อทำการเปรียบเทียบ ณ ตำแหน่งต่าง ๆ ที่มีความสามารถเท่ากัน จึงเป็นวิธีการที่ยอมรับกันทั่วไป แต่มีความจำกัดในด้านความสลับซับซ้อนของแนวคิดพื้นฐาน และการวิเคราะห์มีความจำเป็นต้องใช้โปรแกรมคอมพิวเตอร์โดยเฉพาะ

หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจะเป็นการเปรียบเทียบผลการตอบข้อสอบระหว่างผู้สอบสองกลุ่มที่มีระดับความสามารถเดียวกัน โดยกำหนดให้ผู้สอบกลุ่มหนึ่งเป็นกลุ่มอ้างอิงและผู้สอบอีกกลุ่มหนึ่งเป็นกลุ่มเปรียบเทียบ ถ้าข้อสอบทำหน้าที่ต่างกันแล้วโอกาสในการตอบข้อสอบถูกของผู้สอบแต่ละกลุ่มจะไม่เท่ากัน ต่อไปจะให้ความสำคัญในเรื่องขั้นตอนทางสถิติ นั่นคือ วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

การแบ่งกลุ่มวิธีการทางสถิติที่ใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ แบ่งได้หลายวิธี ซึ่ง Hambleton et al. (1991) จำแนกวิธีการตรวจสอบการทำหน้าที่ต่างกันออกเป็น 3 กลุ่มใหญ่ ๆ ดังนี้

1. กลุ่มวิธีที่ใช้ทฤษฎีการทดสอบแบบดั้งเดิม (Methods Using Classical Test Theory: CTT) วิธีในกลุ่มนี้พัฒนามาจากหลักการของทฤษฎีการทดสอบแบบดั้งเดิม ใช้คะแนนที่สังเกตได้ของผู้เข้าสอบแต่ละคนเป็นเกณฑ์การจับคู่กลุ่มผู้เข้าสอบย่อย และเปรียบเทียบค่าความยากของข้อสอบแต่ละข้อระหว่างกลุ่มผู้เข้าสอบย่อยที่สนใจศึกษา วิธีการในกลุ่มนี้ ได้แก่ การวิเคราะห์ความแปรปรวน (Analysis of Variance) วิธีสหสัมพันธ์ (Correlational Method) วิธีแปลงค่าความยากของข้อสอบ (Transformed Item Difficulty Method: TID) หรือวิธีการกำหนดจุดเดลต้า (Delta Plot Method) (Angoff, 1982) การวิเคราะห์ตัวลวง (Distractor Analysis) (Scheuneman, 1979) วิธีสหสัมพันธ์บางส่วน (Partial Correlation Methods) (Stricker,1982) และวิธีการทำให้เป็นมาตรฐาน (Standardization Method) (Dorans & Kulick, 1986)

ข้อดีของวิธีการกลุ่มนี้ คือ กระบวนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบไม่ยุ่งยาก เสียค่าใช้จ่ายไม่สูงนัก ใช้ตรวจสอบกันกลุ่มตัวอย่างขนาดเล็กได้ และสามารถอธิบายให้คนทั่วไปเข้าใจได้ง่าย ส่วนข้อเสียก็คือ ค่าสถิติของข้อสอบเปลี่ยนไปตามกลุ่มตัวอย่าง เมื่อกลุ่มตัวอย่างเปลี่ยนไปผลการตรวจพบข้อสอบทำหน้าที่ต่างกันก็เปลี่ยนไป ทำให้การสรุปอ้างอิงผลการศึกษาไปยังกลุ่มประชากรอาจมีความน่าเชื่อถือได้น้อยลง

2. กลุ่มวิธีที่ใช้ทฤษฎีการตอบสนองข้อสอบ (Methods Using Item Response Theory: IRT) วิธีการในกลุ่มนี้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ตามกรอบแนวคิดของทฤษฎีการตอบสนองข้อสอบ โดยปกติแล้วจะใช้ในการเปรียบเทียบเส้นโค้งลักษณะข้อสอบ (Item Characteristic Curves: ICCs) ของกลุ่มผู้เข้าสอบย่อยตามระดับความสามารถของผู้เข้าสอบ ถ้าเส้นโค้งลักษณะข้อสอบของกลุ่มผู้เข้าสอบย่อยสองกลุ่ม มีรูปร่างเหมือนกัน แสดงว่าข้อสอบข้อนั้นทำหน้าที่ไม่ต่างกัน แต่ถ้าเส้นโค้งลักษณะข้อสอบของกลุ่มผู้เข้าสอบย่อยสองกลุ่มมีรูปร่างแตกต่างกัน แสดงว่า ข้อสอบข้อนั้นทำหน้าที่ต่างกัน วิธีการในกลุ่มนี้ ได้แก่ วิธี Analysis of fit (Durovic, 1975 cited in Hambleton & Others, 1993) วิธี Difficulty shift (Wright,Mead & Draba, 1976 cited in Hambleton & Others,

1993) ซึ่งใช้โมเดล IRT แบบหนึ่งพารามิเตอร์ วิธี IRT Area (Ironson & Subkoviak, 1979: Raju, 1988, 1990) วิธี Two-Stage (Lord, 1980) ซึ่งใช้โมเดล IRT แบบสองหรือสามพารามิเตอร์ วิธี Plot (Hambleton & Rogers, 1991 cited in Hambleton & Others, 1993) และวิธี SIBTEST (Shealy & Stout, 1993)

ข้อดีของวิธีการในกลุ่มนี้ คือ การแก้ไขข้อบกพร่องของทฤษฎีการทดสอบแบบดั้งเดิมทำให้ค่าสถิติของข้อสอบไม่เปลี่ยนไปตามกลุ่มตัวอย่างที่สุ่มมาจากประชากรเดียวกัน การประมาณค่าความสามารถของผู้เข้าสอบเป็นอิสระจากค่าความยากของแบบทดสอบ โมเดลทางคณิตศาสตร์ง่ายต่อการจับคู่เส้นโค้งลักษณะข้อสอบตามระดับความสามารถของผู้เข้าสอบ ทำให้สามารถศึกษาความแตกต่างของผลการตอบข้อสอบตามระดับความสามารถของกลุ่มผู้เข้าสอบย่อยได้ไม่ต้องมีข้อจำกัดเบื้องต้นเรื่องแบบทดสอบคู่ขนานในการหาค่าสัมประสิทธิ์ความเที่ยงของแบบทดสอบและถ้าผลการตอบข้อสอบของกลุ่มผู้เข้าสอบสอดคล้องกับข้อตกลงเบื้องต้นของโมเดล IRT แล้ว วิธีการในกลุ่มนี้ก็น่าจะเป็นวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ได้ผลดี เนื่องจากเป็นวิธีที่มีทฤษฎีการตอบสนองข้อสอบสนับสนุนและใช้ค่าประมาณค่าความสามารถที่แท้จริงของผู้เข้าสอบแทนข้อสอบสลับซับซ้อน เสียค่าใช้จ่ายในการดำเนินการสูง และต้องการกลุ่มตัวอย่างขนาดใหญ่

3. กลุ่มวิธีที่ใช้วิธีไค-สแควร์ (Methods Using Chi-Square Methods) วิธีในกลุ่มนี้บางครั้งก็เรียกว่า กลุ่มวิธีไค-สแควร์ เนื่องจากใช้ค่าสถิติไค-สแควร์ แสดงการทำหน้าที่ต่างกันของข้อสอบ และใช้คะแนนของแบบทดสอบหรือคะแนนของแบบทดสอบที่ทำให้บริสุทธิ์เป็นเกณฑ์การจับคู่กลุ่มผู้เข้าสอบย่อยสองกลุ่มที่ทำการศึกษา ก่อนการเปรียบเทียบผลการตอบข้อสอบ วิธีการในกลุ่มนี้ ได้แก่ วิธีตารางการณ์จร (Contingengy Table Method) (Scheuneman, 1975; 1979) วิธีตารางการณ์ปรับใหม่ (Modified Contingengy Table Method) (Veale, 1977 cited in Hambleton & Others, 1993) วิธีลอก-ลิเนียร์ (Log-Linear Methods) (Mellenbergh, 1982) วิธีแมนเทล-แฮนส์เซล (Mantel-Haenszal Method: MH) (Holland & Thayer, 1986; 1988) และวิธีถดถอยโลจิสติก (Logistic Regression Methods: LR) (Swaminathan & Rogers, 1990) และวิธีการวิเคราะห์องค์ประกอบจำกัด (Restricted Factor Analysis Methods: RFA) (Oort, 1998)

ข้อดีของวิธีการในกลุ่มนี้ คือ กระบวนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ไม่ยุ่งยาก เสียค่าใช้จ่ายในการวิเคราะห์ข้อมูลไม่สูง ใช้ขนาดกลุ่มตัวอย่างไม่ใหญ่นัก และบางวิธีมีหลักการที่ดีในการจับคู่กลุ่มผู้เข้าสอบย่อยตามความสามารถของผู้เข้าสอบ และมีการทดสอบนัยสำคัญ ส่วนข้อเสียของวิธีการในกลุ่มนี้ก็คล้าย ๆ กับวิธีที่ใช้ทฤษฎีการทดสอบแบบดั้งเดิม

หลักการทำหน้าที่ต่างกันของข้อสอบ (DIF)

ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบดำเนินการโดยเปรียบเทียบผลการตอบของข้อสอบระหว่างผู้สอบ 2 กลุ่มที่มีความสามารถระดับเดียวกัน โดยกำหนดให้ผู้สอบกลุ่มหนึ่งเป็น “กลุ่มอ้างอิง” (Reference Group: R) ซึ่งเป็นกลุ่มที่คาดว่าจะได้ผลประโยชน์ในการตอบข้อสอบ คือมีโอกาสในการตอบข้อสอบถูกมากกว่าอีกกลุ่ม ส่วนอีกกลุ่มเป็น “กลุ่มเปรียบเทียบ หรือกลุ่มสนใจ” (Focal Group: F) ซึ่งเป็นกลุ่มที่คาดว่าจะเสียประโยชน์ในการตอบข้อสอบ คือ มีโอกาสตอบข้อสอบได้ถูกต้องน้อยกว่าผู้สอบอีกกลุ่มหนึ่ง สำหรับเกณฑ์ที่ใช้ในการจำแนกผู้สอบเป็นกลุ่มอ้างอิง มีหลายลักษณะ เช่น เพศ สีผิว เชื้อชาติ ภาษา วัฒนธรรม และภูมิลำเนา เป็นต้น

ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ จะเริ่มต้นด้วยการกำหนดประชากรให้ชัดเจน และแบ่งประชากรนั้นออกเป็น 2 กลุ่ม ดังที่กล่าวข้างต้นมาคือ กลุ่มอ้างอิง (R) ซึ่งคาดว่าจะได้ประโยชน์จากการที่ข้อสอบทำหน้าที่ต่างกัน หรืออาจกล่าวได้ว่าเป็นกลุ่มที่คาดว่าจะได้คะแนนมากกว่าอีกกลุ่มหนึ่ง ทั้งๆที่มีความสามารถที่แท้จริงเท่ากัน และกลุ่มเปรียบเทียบ (F) เป็นกลุ่มที่คาดว่าจะเสียประโยชน์จากการที่ข้อสอบทำหน้าที่ต่างกัน หรือเป็นกลุ่มที่คาดว่าจะได้คะแนนน้อยกว่ากลุ่มอ้างอิงนั่นเอง หลังจากที่มีการสอบแล้วนำคำตอบที่ได้ไปหาค่าพารามิเตอร์ของข้อสอบ ได้แก่ ค่าอำนาจจำแนกของข้อสอบ (a) ค่าความยากของข้อสอบ (b) และค่าโอกาสในการเดาของข้อสอบ (c) ดังได้กล่าวมาแล้วว่ามีการแบ่งกลุ่มผู้สอบเป็น 2 กลุ่ม ดังนั้นแต่ละกลุ่มก็มีชุดพารามิเตอร์ของข้อสอบเฉพาะกลุ่มของตน

การทำหน้าที่ต่างกันของข้อสอบดังกล่าวเป็นการพิจารณาจากค่าสถิติซึ่งสามารถคำนวณได้จากหลายวิธีโดยใช้คะแนนที่สังเกตได้ (Observed Score) และคะแนนที่สังเกตไม่ได้ (Latent Variable)

ลักษณะของข้อสอบโดยทั่วไปที่แสดงการทำหน้าที่ต่างกัน

1. มีเนื้อหาหรือภาษาที่ใช้ในข้อสอบยั่วยุให้ผู้สอบสนใจ โกรธเกิดการโต้แย้งหรือเกิดอารมณ์ไม่พอใจ

2. เนื้อหาหรือภาษาที่ใช้ในข้อสอบมีความหมายไปทางลบ ดูถูกเหยียดหยามหรือก้าวร้าวต่อผู้ตอบข้อสอบกลุ่มสนใจ

3. เนื้อหาหรือภาษาในข้อสอบแสดงว่าผู้ตอบข้อสอบกลุ่มสนใจมีปมด้อยเกี่ยวกับอำนาจหรือความเป็นผู้นำ

4. เนื้อหาหรือภาษาในข้อสอบหลายๆข้อให้ความสนใจเน้นความสำคัญและยกย่องผู้ตอบข้อสอบกลุ่มอ้างอิง

5. เนื้อหาหรือภาษาในข้อสอบมีสารสนเทศเป็นประโยชน์กับกลุ่มอ้างอิงมากกว่ากลุ่มสนใจ

ลักษณะข้อสอบที่แสดงการทำหน้าที่ต่างกันต่อเพศ

1. รูปแบบหรือโครงสร้างของข้อสอบเป็นปัญหาต่อผู้ตอบข้อสอบเพศใดเพศหนึ่งมากกว่าผู้ตอบข้อสอบอีกเพศหนึ่ง

2. เนื้อหาในข้อสอบมีสรรพนามเฉพาะเพศใดเพศหนึ่ง

3. เนื้อหาในข้อสอบกำหนดสถานการณ์ที่ผู้ตอบข้อสอบเพศใดเพศหนึ่งได้รับการฝึกฝนเฉพาะทางมีความสนใจและมีโอกาสพบเห็นในชีวิตประจำวันมากกว่า

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

วิธีการในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมีหลายวิธี สามารถจำแนกได้หลายลักษณะขึ้นอยู่กับเกณฑ์ที่ใช้จำแนก เช่น การใช้เกณฑ์การให้คะแนน แบ่งได้เป็น 2 กลุ่มวิธี คือ

1. กลุ่มวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนเป็นแบบ 2 ค่า (Dichotomous DIF Procedures) กลุ่มนี้ข้อสอบที่ตรวจสอบการทำหน้าที่ต่างกันมีการให้คะแนนเป็นแบบ 0-1 เช่น แบบทดสอบเลือกตอบที่ให้คะแนนตอบถูกเป็น 1 คะแนน และตอบผิดเป็น 0 คะแนน และกลุ่มวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนแบบหลายค่า (Polytomous DIF Procedures) เช่นข้อสอบวัดการปฏิบัติ (Performance Test) ข้อสอบที่ให้สร้างคำตอบเอง (Constructed - response Item) ไม่ว่าจะเป็นข้อสอบวัดการอ่าน (Reading Item) หรือการเขียน

(Writing Lethem) หรือแบบทดสอบเลือกตอบที่มีการให้คะแนนความรู้บางส่วน เช่น แบบทดสอบเลือกตอบแบบถูกผิด เป็นต้น การใช้เกณฑ์ที่ยึดทฤษฎีของการวิเคราะห์ข้อมูลแบ่งเป็น 2 กลุ่มวิธี คือ กลุ่มวิธีที่ยึดทฤษฎี IRT ที่วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยใช้คะแนนที่สังเกตไม่ได้หรือตัวแปรแฝงภายใต้ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory) และกลุ่มวิธีที่ไม่ใช้ IRT (Non IRT) กลุ่มนี้จะวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยใช้คะแนนสังเกตได้ภายใต้ทฤษฎีการทดสอบแบบดั้งเดิม (Classical Test Theory) การใช้เกณฑ์ข้อสอบเบื้องต้นของแบบจำลองแบ่งเป็น 2 กลุ่มวิธีคือ กลุ่มวิธีที่ยึดรูปแบบพาราเมตริก (Parametric Form) การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบมีข้อตกลงเบื้องต้นของแบบจำลองสำหรับอธิบายความสัมพันธ์ระหว่างคะแนนของข้อสอบและการจับคู่ตัวแปรและกลุ่มวิธีที่ยึดรูปแบบนันพาราเมตริก (Nonparametric Form) ซึ่งกลุ่มนี้จะไม่มีข้อตกลงเบื้องต้น

งานวิจัยที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) มีดังนี้

พีรญา สูงเนิน เสรี ชัดแช้ม และสมโภชน์ อเนกสุข (2552) ได้ศึกษาเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างข้อสอบรายข้อกับหมวดข้อสอบ โดยวิธี SIBTEST ภายใต้เงื่อนไขขนาดของกลุ่มตัวอย่างที่แตกต่างกัน กลุ่มตัวอย่างเป็นนักเรียนชั้นประถมศึกษาปีที่ 6 ปีการศึกษา 2546 สังกัดสำนักงานเขตพื้นที่การศึกษานครศรีธรรมราช นักเรียนที่เข้าสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ จำนวน 2,000 คน โดยใช้ข้อมูลทุติยภูมิ จากคะแนนแบบทดสอบวิชาภาษาไทย จำนวน 40 ข้อ ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ผลการศึกษาปรากฏว่า กลุ่มตัวอย่างขนาดเล็ก ข้อสอบทำหน้าที่ต่างกัน จำนวน 4 ข้อ คิดเป็นร้อยละ 10 กลุ่มตัวอย่างขนาดกลาง พบข้อสอบที่ทำหน้าที่แตกต่างกัน จำนวน 13 ข้อ คิดเป็นร้อยละ 32.50 และกลุ่มตัวอย่างขนาดใหญ่ พบข้อสอบที่ทำหน้าที่แตกต่างกัน จำนวน 15 ข้อ คิดเป็นร้อยละ 37.50 เมื่อนำไปตรวจสอบทีละหมวด ผลการศึกษาปรากฏว่า กลุ่มตัวอย่างขนาดเล็ก พบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 4 ข้อ คิดเป็นร้อยละ 10 กลุ่มตัวอย่างขนาดกลาง พบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 8 ข้อ คิดเป็นร้อยละ 20 และกลุ่มตัวอย่างขนาดใหญ่ พบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 16 ข้อ คิดเป็นร้อยละ 40 จากการศึกษาสรุปได้ว่า ขนาดกลุ่มตัวอย่างที่ใหญ่ทำให้สามารถพบข้อสอบที่ทำหน้าที่ต่างกันได้ดีกว่ากลุ่มตัวอย่างที่มีขนาดเล็ก

เรืองเดช ศิริกิจ (2554) ได้วิเคราะห์เปรียบเทียบโมเดลการประเมินคุณภาพการจัดการศึกษาวิชาคณิตศาสตร์: การประยุกต์ใช้โมเดลมูลค่าเพิ่มที่มีการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ และการวิเคราะห์การทำหน้าที่ต่างกันของตัวลวง จากการศึกษาปรากฏว่า ตัวแปรเพศมีคุณลักษณะที่พบการทำหน้าที่ต่างกันของข้อสอบ และการทำหน้าที่ต่างกันของตัวลวงในแบบทดสอบวิชาคณิตศาสตร์มากที่สุด เพศชายมีความสามารถในการแก้ปัญหาทางคณิตศาสตร์ได้ดีกว่าเพศหญิง สรุปได้ว่า เพศมีผลต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาคณิตศาสตร์ โดยเพศชายมีความสามารถในการแก้ปัญหาทางคณิตศาสตร์ดีกว่าเพศหญิง

ศิริรัตน์ สุคันธพฤกษ์ (2554) ได้ตรวจสอบการทำหน้าที่ต่างกันในแบบวัดความวิตกกังวลในการสอบคณิตศาสตร์โดยเปรียบเทียบระหว่าง Hierarchical Linear Model: HLM, Partial Credit Model: PCM และ Graded Reponse Model: GRM กลุ่มตัวอย่างที่ใช้ในการวิจัยเป็นนักเรียนสายวิทย์-คณิต ระดับชั้นมัธยมศึกษาปีที่ 6 ปีการศึกษา 2552 จำนวน 1,715 คน จากโรงเรียนทั้งสิ้น 29 โรงเรียน

ในสังกัดสำนักงานเขตพื้นที่การศึกษาพระนครศรีอยุธยา เขต 1 และเขต 2 สำนักงานเขตพื้นที่การศึกษาอ่างทองและสำนักงานเขตพื้นที่การศึกษานนทบุรี ซึ่งได้มาจากการสุ่มตัวอย่างแบบยกชั้น เครื่องมือที่ใช้ในการวิจัย คือ แบบวัดความวิตกกังวลในการสอบคณิตศาสตร์ โดยวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธี HLM วิธี PCM และวิธี GRM ด้วยโปรแกรม PRASCALE จากการเปรียบเทียบผลการวิเคราะห์ข้อมูลทั้ง 3 วิธี ผลการศึกษาปรากฏว่า ข้อคำถามที่ทำหน้าที่ต่างกันของข้อร่วมระหว่างวิธี HLM, วิธี PCM และ วิธี GRM มี 6 ข้อ จาก 39 ข้อ คิดเป็นร้อยละ 15.38 ข้อคำถามที่ทำหน้าที่ต่างกันของข้อร่วมระหว่าง วิธี HLM กับ วิธี PCM มี 7 ข้อ จาก 39 ข้อ คิดเป็นร้อยละ 17.94 และข้อคำถามที่ทำหน้าที่ต่างกันของข้อร่วมระหว่าง วิธี HLM กับ วิธี GRM มี 9 ข้อ จาก 39 ข้อคิดเป็นร้อยละ 23.07

สุพัฒนา หอมบุปผา ไพรัตน์ วงษ์นาม และสมพงษ์ ปั้นหุ่น (2556) ได้ศึกษาการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN ศึกษาลักษณะของข้อสอบที่เกิดจากการทำหน้าที่ต่างกันของข้อสอบ (DIF) ที่ได้จากการวิเคราะห์การทำหน้าที่ต่างกันด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN ข้อมูลที่ใช้ในการวิเคราะห์เป็นคะแนนการสอบวัดผลสัมฤทธิ์ทางการเรียนเพื่อประเมินคุณภาพการศึกษาระดับชาติ ของนักเรียนชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2553 ซึ่งมีการทดสอบใน 3 วิชา ได้แก่ วิชาภาษาไทย คณิตศาสตร์และวิทยาศาสตร์ ใช้กลุ่มตัวอย่าง จำนวน 1,000 คน จำแนกตามเพศ ซึ่งเป็นนักเรียนที่อยู่ในโรงเรียนเขตกรุงเทพ มหานครและปริมณฑล และนอกเขตกรุงเทพมหานคร ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ด้วยวิธี HGLM-2L วิธี MIMIC และวิธี BAYESIAN ผลการศึกษาปรากฏว่า ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทย คณิตศาสตร์และวิทยาศาสตร์ มีความ สัมพันธ์กันในระดับที่สูงมาก วิธีตรวจสอบที่พบการทำหน้าที่ต่างกันของข้อสอบมากที่สุด คือ วิธี HGLM -2L ส่วนวิธีที่ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบน้อยที่สุด คือ วิธี MIMIC สรุปได้ว่าการตรวจสอบการทำหน้าที่ต่างกันด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN ตรวจพบข้อสอบทำหน้าที่ต่างกันแตกต่างกัน

ชัยวัฒน์ หฤทัยพันธ์ (2558) ได้ศึกษาพัฒนาวิธีการสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยผู้เชี่ยวชาญ และเพื่อเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในด้านอัตราความถูกต้อง และอัตราความคลาดเคลื่อนของผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เมื่อใช้การวิเคราะห์ด้วยแบบวินิจฉัยโดยผู้เชี่ยวชาญ วิธีการประยุกต์ใช้เทคนิคการประชุมแบบเดลฟายจากกลุ่มผู้เชี่ยวชาญและวิธีการประยุกต์ใช้เทคนิคโปรโตคอลอะลาร์ด ตัวอย่างที่ใช้ในการวิจัย คือ ผู้เชี่ยวชาญจำนวน 21 คน และนักเรียนระดับชั้นมัธยมศึกษาปีที่ 6 จำนวน 139 คน ปีการศึกษา 2556 ซึ่งได้จากการเลือกตัวอย่างแบบเจาะจง เครื่องมือที่ใช้ในการวิจัยประกอบด้วย แบบวินิจฉัย การทำหน้าที่ต่างกันของข้อสอบจากผู้เชี่ยวชาญ แบบยืนยันการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยประยุกต์เทคนิคการประชุมแบบเดลฟาย แบบสอบถามสำหรับการตรวจสอบความลำเอียงของข้อสอบ สำหรับนักเรียน ชุดข้อสอบสาระการเรียนรู้สุขศึกษาและพลศึกษาสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสำหรับผู้เชี่ยวชาญ ข้อสอบที่คัดสรรมาได้นำมาผ่านการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยวิธีเมลเทล-แฮนส์เซล ด้วยโปรแกรม DDFS 1.0 และโปรแกรม DIFAS 5.0 ผลการศึกษาปรากฏว่า วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยการตัดสินของผู้เชี่ยวชาญ ที่สำคัญมี 3 วิธี ได้แก่ วิธีที่ 1 การวินิจฉัยการตรวจสอบการทำหน้าที่

ต่างกันของข้อสอบโดยผู้เชี่ยวชาญ วิธีที่ 2 การประยุกต์ใช้เทคนิคการประชุมแบบเดลฟายจากกลุ่มผู้เชี่ยวชาญ และวิธีที่ 3 การประยุกต์ใช้เทคนิคโปรโตคอลอะลาร์ดและข้อสอบที่ทำหน้าที่ต่างกันด้านเพศของแบบสอบสาระการเรียนรู้สุขศึกษาและพลศึกษา จากผลการวิเคราะห์ปรากฏว่า วิธีที่ 1 การตรวจสอบด้วยแบบวินิจฉัยโดยผู้เชี่ยวชาญ มีอัตราความถูกต้องโดยเฉลี่ยคิดเป็นร้อยละ 50 และมีอัตราความคลาดเคลื่อนของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยเฉลี่ยคิดเป็นร้อยละ 50 วิธีที่ 2 การประยุกต์ใช้เทคนิคการประชุมแบบเดลฟายจากกลุ่มผู้เชี่ยวชาญ มีอัตราความถูกต้องตามฉันทามติจากกลุ่มผู้เชี่ยวชาญ คิดเป็นร้อยละ 0 และมีอัตราความคลาดเคลื่อนของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบคิดเป็นร้อยละ 100 วิธีที่ 3 การประยุกต์ใช้เทคนิคโปรโตคอลอะลาร์ด มีอัตราความถูกต้องโดยเฉลี่ยคิดเป็นร้อยละ 25 และมีอัตราความคลาดเคลื่อนของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยเฉลี่ยคิดเป็นร้อยละ 75

พิชชา สุริอาจ และประกฤติยา ทักษิโณ (2559) ได้ศึกษาการพัฒนาแบบวัดความตระหนักต่อโลกในยุคศตวรรษที่ 21 ของนักเรียนมัธยมศึกษาตอนต้น โดยใช้แบบวัดเชิงสถานการณ์: การประยุกต์ใช้การทำหน้าที่ต่างกันของข้อสอบ กลุ่มตัวอย่างที่ใช้เป็นนักเรียนชั้นมัธยมศึกษาปีที่ 1-3 จังหวัดนครราชสีมา สังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน จำนวน 1,200 คน ได้มาจากการสุ่มแบบหลายขั้นตอน มีการวิเคราะห์ค่าอำนาจจำแนกตามทฤษฎีการวัดแบบดั้งเดิม วิเคราะห์ค่า ความเที่ยงโดยใช้โปรแกรม SPSS for windows วิเคราะห์ค่าอำนาจจำแนกตามทฤษฎีการตอบสนองข้อสอบโดยใช้โปรแกรม Multilog วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ (DIF) โดยวิธีโพลีโตมัสซิปเทสใช้โปรแกรมDIFPACK เวอร์ชั่น 1.7 วิเคราะห์องค์ประกอบเชิงยืนยันอันดับที่สอง ใช้โปรแกรม Mplus ผลการศึกษาปรากฏว่า ผลการพัฒนาแบบวัดความตระหนักต่อโลกในยุคศตวรรษที่ 21 ของนักเรียนมัธยมศึกษาตอนต้น ผลการศึกษาปรากฏว่า แบบวัดเชิงสถานการณ์ มี 6 องค์ประกอบ 14 ตัวบ่งชี้ ได้แก่ 1) ความตระหนักในมุมมองที่แตกต่าง 2) ความตระหนักในสภาพปัจจุบันของโลก 3) ความตระหนักในความแตกต่างของวัฒนธรรม 4) ความตระหนักในเรื่องพลวัตของโลก 5) ความตระหนักต่อทางเลือกของมนุษย์ 6) ความตระหนักต่อการเรียนรู้ในการทำงานกับบุคคลที่มีความแตกต่าง ข้อคำถามผ่านเกณฑ์ความตรงเชิงเนื้อหาและทดลองใช้ จำนวน 46 ข้อ ผลการตรวจสอบคุณภาพของแบบวัดความตระหนักต่อโลกในยุคศตวรรษที่ 21 ปรากฏว่า มีค่าอำนาจจำแนกตามทฤษฎีตอบสนองข้อสอบ (a) อยู่ระหว่าง 0.15 ถึง 3.22 ค่าความเที่ยงเท่ากับ 0.80 การทำหน้าที่ต่างกันของข้อคำถามตามตัวแปรเพศ พบว่า DIF จำนวน 3 ข้อ

สุธาทิพย์ ตรีสิน และปิยะทิพย์ ประดุจพรม (2560) ได้เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ 3 ด้าน ได้แก่ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ระดับชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2556 ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้ง 3 ด้าน ผลการศึกษาปรากฏว่า วิธี HGLM สามารถตรวจพบข้อสอบทำหน้าที่ต่างกันจำนวนมากที่สุด คิดเป็นร้อยละ 69 รองลงมาคือ วิธี IRT-LR คิดเป็นร้อยละ 54 และวิธี MIMIC พบข้อสอบทำหน้าที่ต่างกันน้อยที่สุด คิดเป็นร้อยละ 16 โดยเพศมีผลต่อการทำหน้าที่ต่างกันของข้อสอบ

Barnett and Ercikan (2006) ได้ศึกษาการทำหน้าที่ต่างกันของข้อสอบในการสอบวิชาคณิตศาสตร์ โดยใช้วิธี SIBTEST จำแนกตามเพศ ผลการศึกษาปรากฏว่า เพศชายมีความสามารถ

ในการแก้ปัญหาและวิธีการทางปัญญาที่สูงกว่าในการทำข้อสอบ ขณะที่เพศหญิงมีความสามารถในด้านการคำนวณสมการ ซึ่งการคำนวณไม่ได้ถูกจัดให้มีอยู่ในข้อสอบ จึงสรุปได้ว่า ข้อสอบวิชาคณิตศาสตร์นี้เกิดการทำหน้าที่ต่างกันของข้อสอบ โดยมีแนวโน้มว่าเพศชายมีความได้เปรียบมากกว่าเพศหญิง

Le (2009) ได้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ PISA วิชาวิทยาศาสตร์และแบบทดสอบด้านภาษาระหว่างประเทศ จำแนกตามเพศ เพื่อตรวจสอบความสัมพันธ์ระหว่างการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ มีการเก็บรวบรวมข้อมูลจาก 60 กลุ่มภาษาจาก 50 ประเทศ นักเรียนที่เข้าร่วมโครงการ จำนวน 83,000 คน ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้ทฤษฎีการตอบสนองข้อสอบ ลักษณะของข้อสอบมีทั้งข้อสอบแบบเลือกตอบและแบบตอบกลับปลายปิด มีแนวโน้มที่จะเข้าข้างเพศชาย จากการศึกษาแสดงให้เห็นถึงผลกระทบของประเทศและการทดสอบด้านภาษากับเพศ เพศชายจึงมีแนวโน้มที่จะมีความสามารถทางวิทยาศาสตร์ดีกว่าเพศหญิงและผลจากการศึกษาในครั้งนี้เป็นผลงานที่มีคุณค่าต่อการพัฒนาการทดสอบระหว่างประเทศสามารถนำไปใช้ในระดับสากลได้ สรุปได้ว่า เพศชายมีแนวโน้มที่จะมีความสามารถทางวิทยาศาสตร์ดีกว่าเพศหญิง

Park (2010) ได้ศึกษาการตรวจสอบการทำหน้าที่ต่างกันของแบบทดสอบคำศัพท์ EFL เพื่อตรวจสอบความเข้าใจในการอ่านและความรู้ด้านคำศัพท์ ในการศึกษาใช้วิธีการตรวจสอบ 3 วิธี ได้แก่ วิธี Likelihood ratio วิธี SIBTEST และวิธี Mantel-Haenszel ผลการศึกษาปรากฏว่า เพศหญิงมีความสามารถในการอ่านและจดจำคำศัพท์ได้ดีกว่าเพศชาย ส่วนเพศชายมีความสามารถในการใช้ศัพท์ภาษาศาสตร์ได้ดีกว่านักเรียนที่เรียนภาษาอังกฤษทั่วไปในระดับความรู้คำศัพท์เดียวกัน สรุปได้ว่าเพศหญิงมีความสามารถในการอ่านและจดจำคำศัพท์แบบคงทนของแบบทดสอบคำศัพท์ EFL ได้ดีกว่าเพศชาย

Williams and Lamprianou (2011) ได้ศึกษาการตรวจสอบความตรงของความแตกต่างระหว่างเพศในการประเมินผลการทำหน้าที่ต่างกันของข้อสอบ ของกลุ่มข้อสอบคณิตศาสตร์ เพื่อตรวจสอบแหล่งที่มาของการทำหน้าที่ต่างกันโดยเพศ ในการประเมินการทำแบบทดสอบวิชาคณิตศาสตร์และใช้ในการตรวจสอบการทำหน้าที่ต่างกันของกลุ่มข้อสอบว่ามีความเกี่ยวข้องกันหรือไม่ สอดคล้องกับ Rousso and Stout (1996) ได้ทำการศึกษาการทำหน้าที่ต่างกันของกลุ่มข้อสอบกับการปรับตัวด้วยวิธีการถดถอยโลจิสติก ผลการศึกษาปรากฏว่า ในการประเมินผลระดับชาติ โดยเฉพาะอย่างยิ่งโครงการสำหรับการประเมินนานาชาติ (PISA) ความสามารถจากการทดสอบคณิตศาสตร์เพศชายมักจะทำได้ดีกว่าเพศหญิง

Taylor and Lee (2012) ได้ศึกษาการทำหน้าที่ต่างกันของข้อสอบ (DIF) ด้วยรูปแบบผสมด้านการอ่านและคณิตศาสตร์ จำแนกตามเพศ ระดับชั้นประถมศึกษาปีที่ 4, 7 และ 10 การทดสอบตามเกณฑ์ของรัฐ ซึ่งประกอบด้วยข้อสอบที่มีตัวเลือกหลายตัวเลือกและสร้างการตอบสนองเพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี Poly-SIBTEST จากการศึกษาปรากฏว่าด้านการอ่าน การตีความ การวิเคราะห์ข้อความหรือความหมายโดยนัยมีแนวโน้มเข้าข้างเพศหญิง ส่วนการวิเคราะห์เนื้อหาของวิชาคณิตศาสตร์ เช่น เรขาคณิต การตีความทางสถิติความน่าจะเป็น พีชคณิต การแก้ปัญหาหลายขั้นตอนและการให้เหตุผลเชิงคณิตศาสตร์มีแนวโน้มเข้าข้างเพศชาย สรุปได้ว่า วิชาคณิตศาสตร์เพศชายมีแนวโน้มที่จะได้เปรียบมากกว่าเพศหญิง ส่วนในด้านการอ่าน การตีความ การวิเคราะห์ข้อความ

หรือความหมายโดยนัย เพศหญิงมีแนวโน้มที่จะได้เปรียบมากกว่าเพศชาย

Ong, Lu, Lee, and Cohen (2015) ได้ศึกษาเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ระหว่างวิธี HGLM วิธี MIMIC และวิธี IRT ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่าง ที่แตกต่างกัน เปรียบเทียบอัตราความคลาดเคลื่อนประเภทที่ 1 (Type 1 Error) ปรากฏว่า วิธี MIMIC มีอัตราความคลาดเคลื่อนประเภทที่ 1 น้อยกว่าวิธี HGLM และวิธี IRT เมื่อกลุ่มตัวอย่างมีขนาดเล็ก หากกลุ่มตัวอย่างมีจำนวนเพิ่มมากขึ้นอัตราความคลาดเคลื่อนก็จะเพิ่มขึ้น วิธี IRT เมื่อกลุ่มตัวอย่างมีขนาดเล็ก จะพบอัตราความคลาดเคลื่อนน้อย แต่ถ้ากลุ่มตัวอย่างมีจำนวนเพิ่มมากขึ้นอัตราความคลาดเคลื่อนก็จะเพิ่มขึ้น จากการศึกษาจึงสามารถสรุปได้ว่าวิธี MIMIC และวิธี IRT พบอัตรา ความคลาดเคลื่อนน้อยเมื่อกลุ่มตัวอย่างขนาดเล็ก หากกลุ่มตัวอย่างมีจำนวนเพิ่มมากขึ้น อัตราความคลาดเคลื่อนก็จะเพิ่มขึ้น ส่วนวิธี HGLM นั้นเป็นวิธีที่สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบและพบจำนวนข้อสอบที่ทำหน้าที่ต่างกันได้มากที่สุด

จากการศึกษางานวิจัยที่เกี่ยวข้อง ปรากฏว่า การทำหน้าที่ต่างกันของข้อสอบ (DIF) เป็นการขจัดข้อสอบที่จะก่อให้เกิดความลำเอียงต่อผู้เข้าสอบ เป็นกระบวนการที่เน้นการใช้วิธีการทางสถิติสำหรับการตรวจสอบว่าข้อสอบข้อใดมีความน่าจะเป็นที่จะก่อให้เกิดความลำเอียงได้บ้าง โดยคุณลักษณะของผู้เข้าสอบมีส่วนที่ทำให้เกิดการทำหน้าที่ต่างกัน เช่น เพศ ภาษา เชื้อชาติ สังคม ประสบการณ์ และภูมิลำเนา เป็นต้น วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมีหลายวิธี ผู้วิจัยได้เลือกใช้วิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่อยู่บนพื้นฐานของทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) ได้แก่ วิธี Hierarchical Generalized Linear Model (HGLM) วิธี Multiple Indicators and Multiple Causes (MIMIC) และวิธี BAYESIAN ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสำหรับงานวิจัยนี้

ตอนที่ 4 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี HGLM และงานวิจัยที่เกี่ยวข้อง

การวิเคราะห์ข้อสอบแบบพหุระดับเกิดขึ้นจากความพยายามที่นักวัดผล ต้องการศึกษาอิทธิพลของตัวแปรภายนอกที่เป็นตัวแปรทางจิตวิทยา ตัวแปรคุณลักษณะผู้สอบให้สามารถประมาณค่าร่วมในโมเดลการรวมกันเชิงเส้นไปพร้อมกับการประมาณค่าพารามิเตอร์ข้อสอบและพารามิเตอร์ผู้สอบแต่ที่ผ่านมาการประมาณค่าไปพร้อม ๆ กันทำให้เกิดผลการวิเคราะห์ที่คลาดเคลื่อน ความพยายามดังกล่าวทำให้เริ่มมีการวิเคราะห์แบบสองขั้นตอน คือ การวิเคราะห์ค่าความสามารถ ให้ผลการวิเคราะห์ตามหลักการของทฤษฎีการตอบสนองข้อสอบ นั่นคือค่าพารามิเตอร์ความสามารถผู้สอบ (θ_j) ต่อมานักวิจัยจึงนำค่า θ_j เหล่านี้ มาเป็นตัวแปรตามในการวิเคราะห์ถดถอย เพื่อมุ่งหาคำตอบใน 2 ประการหลักคือ ตัวแปร θ_j เหล่านี้ มีความผันแปรระหว่างผู้สอบหรือไม่ และหากมีความผันแปรเกิดขึ้น มีตัวแปรใดบ้างที่อธิบายความผันแปรที่เกิดขึ้นได้ โดยในกรณีนี้นักวิจัยจะนำตัวแปรทางจิตวิทยาหรือตัวแปรคุณลักษณะผู้สอบที่สนใจ เป็นตัวแปรทำนายในสมการถดถอยพหุ แต่นักวิจัยหลายคน เช่น Maier (2001); Hambleton & Swaminatan (1985); Adam, Wilson, & Wu (1997) ก็ให้ความรู้เกี่ยวกับความคลาดเคลื่อนที่อาจเกิดขึ้น จากการวิเคราะห์แบบสองขั้นตอนในสองประเด็นหลักคือ 1) ค่าความสามารถของผู้สอบที่ได้จากการประมาณค่าด้วยโมเดลการตอบสนองข้อสอบจะมีความแตกต่างกันของ

ค่าขนาดความคลาดเคลื่อนมาตรฐาน ณ ตำแหน่งค่าความสามารถของผู้สอบที่ต่างกัน การวิเคราะห์ที่ละเลยปัญหาความผันแปรของความคลาดเคลื่อนมาตรฐาน จะทำให้การวิเคราะห์ 2 ขั้นตอน มีการประมาณค่าที่ไม่คงที่ 2) การประมาณค่าความสามารถของผู้สอบ จะเกิดขึ้นภายหลังจากการประมาณค่าพารามิเตอร์ข้อสอบ ซึ่งจะรับผลจากการประมาณค่าครั้งแรกมาคำนวณต่อจะเกิดความลำเอียงและความไม่คงที่ของการประมาณค่า ซึ่งการวิเคราะห์ลักษณะนี้เป็นปัญหาของโมเดลการวิเคราะห์ถดถอย

นักวัดผลจึงนำหลักการของ Fisher (1983) ที่เสนอสมการรวมกันเชิงเส้น (Linear Combination) ที่สามารถดำเนินการได้ในลักษณะดังกล่าวแบบขั้นตอนเดียว ประกอบกับการพัฒนาสถิติหลายประการ ที่สามารถเอาชนะข้อจำกัด การประมาณค่าพารามิเตอร์ข้อสอบและพารามิเตอร์ผู้สอบไปพร้อมๆกัน เช่น Bock and Aikin (1981) ได้พัฒนาเทคนิควิเคราะห์แบบ MMLE ขึ้นสำหรับการวิเคราะห์ตามทฤษฎี IRT ซึ่งถือว่าเป็นวิธีการหลักของการประมาณค่าตามทฤษฎีการตอบสนองข้อสอบที่มีประสิทธิภาพมาก การประมาณค่าอีกวิธีหนึ่งเกิดจากการศึกษาของ Adam, Wilson, and Wu (1997) ที่ได้พัฒนาเทคนิคการวิเคราะห์ที่ชื่อ Random Coefficient Multinomail Logit Model (RCMLM) สามารถกำหนดให้ค่าพารามิเตอร์ผู้สอบเป็นตัวแปรสุ่มและสามารถรวมตัวแปรคุณลักษณะผู้สอบเป็นตัวแปรทำนายในสมการเดียวกันได้ต่อมา Adam, Wilson, and Wu (1997) ก็พัฒนาเทคนิคการวิเคราะห์กับโมเดลดั้งเดิมได้ด้วย เช่น โมเดลราสช์ทั้งแบบตัวแปรทวิภาคและพหุภาค

ในปี ค.ศ. 1998 Kamata นักศึกษาระดับปริญญาเอกของมหาวิทยาลัยมิชิแกน ประเทศสหรัฐอเมริกา ได้ทำดุษฎีนิพนธ์ภายใต้การดูแลของ Raudenbush ได้นำหลักการทางสถิติดังกล่าวมาเสนอรูปแบบการวิเคราะห์ข้อสอบ ภายใต้มโนทัศน์แบบพหุระดับเป็นคนแรกโดยงานดังกล่าว Kamata ได้เสนอเทคนิคทางสถิติที่สามารถวิเคราะห์ได้ด้วยโปรแกรม HLM ภายใต้โมเดลเชิงเส้นตรงทั่วไประดับลดหลั่น (Hierarchical Generalized Linear Model: HGLM) วิเคราะห์ข้อสอบแบบ 2 ระดับ และการตรวจสอบความคงที่ของพารามิเตอร์ (Parameter Recovery) ซึ่ง Kamata (2001) ได้เสนอความสมมูลของโมเดล HGLM กับโมเดลราสช์หรือโมเดล IRT แบบ 1 พารามิเตอร์ จากการศึกษาของ Kamata จะพิจารณาว่าการตอบข้อสอบของผู้สอบแต่ละคนเป็นโมเดลภายในผู้สอบ (Within-Student Model) และความผันแปรของประชากรผู้สอบเป็นโมเดลระหว่างผู้สอบ (Between-Student Model) การใช้แนวคิดพื้นฐานนี้เป็นการขยายแนวคิดของโมเดลทฤษฎีการตอบสนองข้อสอบ ว่าเป็นโมเดลพหุระดับที่มีตัวแปรแฝงเป็นตัวแปรตาม

การวิเคราะห์โมเดล HGLM ด้วย HLM

การวิเคราะห์ข้อมูลที่เป็นพหุระดับ (Multilevel Data) หากข้อมูลมีลักษณะโครงสร้างไม่เป็นเชิงเส้นตรง (Nonlinear Structural) และมีการกระจายของความคลาดเคลื่อนที่ไม่เป็นโค้งปกติ (Nonnomally Distributed Error) การวิเคราะห์ด้วยโมเดลเชิงเส้นตรงระดับลดหลั่น (HLM) อาจจะเกิดความไม่เหมาะสมในการวิเคราะห์ เพราะการแปลความหมายและการประมาณค่าอาจจะเกิดความผิดพลาด ดังนั้นโมเดลที่เหมาะสมกว่าและข้อมูลที่มีลักษณะเป็นแบบแบ่ง 2 ส่วน (Binary Response) ควรวิเคราะห์ด้วยโมเดลเชิงเส้นตรงทั่วไประดับลดหลั่น (Hierarchical Generalized Linear Model: HGLM) มากกว่าการวิเคราะห์ด้วย HLM (McCullagh & Nelder, 1989; Raudenbush & Bryk, 2002; Kamata, 2001)

การวิเคราะห์ข้อสอบแบบพหุระดับ

การวิเคราะห์พหุระดับเมื่อการตอบเป็นแบบทวิภาคได้นั้นคือ การใช้โมเดล HGLM ซึ่งโมเดลเชิงเส้นตรงทั่วไประดับลดหลั่น (HGLM) เป็นโมเดลที่มีลักษณะของการทำงานร่วมกันของสองโมเดลหลัก คือ โมเดลเชิงเส้นนัยทั่วไป (GLM: Generalized Linear Model) และโมเดลเชิงเส้นระดับลดหลั่น (HLM: Hierarchical Linear and Non- linear Model) โดยตัวแปรตามในระดับการวิเคราะห์ที่ 1 เป็นตัวแปรทวิภาค โมเดล HGLM จะนำหลักการกระจายแบบ Bemoulli เข้ามาใช้ในการสร้างสมการในระดับการวิเคราะห์ที่ 1 เพื่อให้เกิดการคำนวณทวนซ้ำ (Interactions) ตามโมเดล เชิงเส้นนัยทั่วไป (GLM) ก่อนแล้ว จึงใช้ฟังก์ชันการเชื่อมโยงหน้าที่แบบโลจิสเข้ามาทำหน้าที่จะสามารถทำให้เกิดฟังก์ชันเชื่อมโยง (Link Function) โดยการแปลงแบบโลจิสทำให้มีคุณสมบัติตรงตามการวิเคราะห์ถดถอยเชิงเส้นตรง ซึ่งจะมีความต่อเนื่องได้ตั้งแต่-α ถึง +α ขึ้นอยู่กับพิสัยของตัวแปรทำนาย ข้อมูลจากการวิเคราะห์ระดับที่ 1 จึงสามารถนำเข้าสู่การวิเคราะห์ระดับที่ 2 และระดับที่สูงขึ้นไป

Raudenbush and Bryk (2002, pp. 185-186) ได้กล่าวถึงลักษณะของโมเดลการวิเคราะห์เชิงเส้นตรงระดับลดหลั่นว่า มีองค์ประกอบหลัก คือ โมเดลการสุ่ม (Sampling Model) โมเดลการเชื่อมโยงหน้าที่ (Link Function Model) และโมเดลโครงสร้าง (Structural Model) ความสัมพันธ์ของโมเดล HGLM และ HLM ดังตารางที่ 2-5

ตารางที่ 2-5 ความสัมพันธ์ของหลักการวิเคราะห์ของสมการแบบ HLM และ HGLM
อิทธิฤทธิ์ พงษ์ปิยะรัตน์ (2551, หน้า 42)

สมการ	โมเดลการสุ่ม (Sampling Model)	โมเดลการเชื่อมโยงหน้าที่ (Link Function Model)	โมเดลโครงสร้าง (Structural Model)
HLM	ตัวแปรตามเป็นตัวแปรต่อเนื่อง การกระจายของตัวแปรตามเป็นการกระจายแบบโค้งปกติ มีค่าเฉลี่ยเท่ากับ μ_{ij} และการกระจายเท่ากับ σ^2 เขียนเป็นสมการได้ดังนี้ $Y_{ij} \mid \mu_{ij} \sim$ NID (μ_{ij} σ^2)	การวิเคราะห์ด้วย HLM ลักษณะทั่วไปไม่มีความจำเป็นต้องเปลี่ยนแปลงค่าดังกล่าวแต่ก็สามารถใช้ฟังก์ชันแบบ Lojit link ได้ ($\eta_{ij}=\mu_{ij}$=Identity Link Function)	การเปลี่ยนค่าของตัวทำนายเป็นจะมีความสัมพันธ์กับตัวแปรทำนายต่างๆในโมเดลสามารถแสดงในรูปสมการเชิงเส้นตรง ได้ดังนี้ $\eta_{ij}=\beta_{0i}+\beta_{1j}X_{1ij}+\beta_{2j}X_{2ij}+...+\beta_{pi}X_{pij}$
HGLM	ตัวแปรตามจะเป็นการตอบแบบทวิภาค (0,1) การกระจายจึงเป็นแบบไบโนเมียล ซึ่งเป็นกรณีหนึ่งของการกระจายแบบ Bernoulli (Y_{ij} $\varphi_{ij} \sim B(m_{ij}, \sigma^2)$)	การเชื่อมโยงหน้าที่ (Link Function) ในโมเดลนี้จะใช้ Logit Link เขียนเป็นสมการได้ดังนี้ $\eta_{ij} = \log\left[\frac{\varphi_{ij}}{1-\varphi_{ij}}\right]$ เมื่อ η_{ij} ค่าลอกของออดที่จะประสบความสำเร็จในการตอบข้อสอบข้อที่ i	การประมาณค่า จากสมการในโมเดลโครงสร้างของ HLM ก่อให้เกิดการทำนาย Log Odds ก็สามารถแปลงค่ากลับเป็นค่า Odds ได้ดังค่าเดิมโดยการคูณค่า exp(-η_{ij}) $\varphi_{ij} = \frac{1}{1+\exp(-\eta_{ij})}$

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

งานวิจัยที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี HGLM มีดังนี้

อิทธิฤทธิ์ พงษ์ปิยะรัตน์ (2551) ได้ศึกษาการวิเคราะห์ข้อสอบและการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ: การวิเคราะห์พหุระดับ พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ด้วยโมเดล HGLM จากโปรแกรม HLM สามารถตรวจสอบพบข้อสอบที่ทำหน้าที่ต่างกันตรงกับผลการวิเคราะห์จากโปรแกรม BILOG-MG ผลการวิเคราะห์ระดับนักเรียน ผลการศึกษาปรากฏว่าผลการเรียนวิชาคณิตศาสตร์ส่งผลต่อค่าเฉลี่ยของโอกาสในการตอบข้อสอบได้ถูกต้องในแต่ละโรงเรียน และผลการวิเคราะห์ระดับโรงเรียน ผลการศึกษาปรากฏว่า ขนาดของโรงเรียนและความเป็นผู้นำทางวิชาการของผู้บริหารส่งผลต่อค่าเฉลี่ยของโอกาสในการตอบข้อสอบได้ถูกต้องในโรงเรียน อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 และ .05 ตามลำดับ

สุพัฒนา หอมบุปผา ไพรัตน์ วงษ์นาม และสมพงษ์ ปั้นหุ่น (2556) ได้ศึกษาการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN ผลการศึกษาปรากฏว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทย คณิตศาสตร์ และวิทยาศาสตร์ มีความสัมพันธ์กันในระดับที่สูงมากอย่างมีนัยสำคัญทางสถิติที่ระดับ .01 วิธีการตรวจสอบที่พบการทำหน้าที่ต่างกันของข้อสอบมากที่ คือ วิธี HGLM-2L

สุธาทิพย์ ตรีสิน และปิยะทิพย์ ประดุจพรม (2560) ได้เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ 3 ด้าน ได้แก่ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ระดับชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2556 ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้ง 3 ด้าน ผลการศึกษาปรากฏว่า เพศมีผลต่อการทำหน้าที่ต่างกันของข้อสอบ โดยเพศหญิงจะได้เปรียบในการตอบข้อสอบด้านภาษาและด้านเหตุผล ส่วนเพศชายจะได้เปรียบในการตอบข้อสอบด้านคำนวณ โดยวิธี HGLM สามารถตรวจพบข้อสอบทำหน้าที่ต่างกันจำนวนมากที่สุด รองลงมาคือ วิธี IRT-LR และวิธี MIMIC พบข้อสอบทำหน้าที่ต่างกันน้อยที่สุด

Acar and Kelecioglu (2010) ได้ศึกษาเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ 3 วิธี คือวิธี HGLM วิธี LR และวิธี IRT-LR จำแนกตามเพศ กลุ่มตัวอย่างเป็นนักเรียนในประเทศตุรกี เครื่องมือที่ใช้ในการวิจัยเป็นแบบทดสอบของวิชาสังคมศาสตร์และวิทยาศาสตร์ ผลการศึกษาปรากฏว่า วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้ง 3 วิธี ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันในปริมาณที่ใกล้เคียงกัน แต่วิธี HGLM ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบวิชาสังคมศาสตร์และวิทยาศาสตร์มากที่สุด

Acar (2011) ได้ศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิชาวิทยาศาสตร์และสังคมศาสตร์ ปี 2006 จำนวน 25 ข้อ กลุ่มตัวอย่างที่ใช้จำนวน 10,727 คน ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี HGLM เมื่อมีการกำหนดขนาดกลุ่มตัวอย่างต่างกัน 8 ขนาด ผลการศึกษาปรากฏว่า ขนาดกลุ่มตัวอย่างที่แตกต่างกันมีผลต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

ขนาดกลุ่มตัวอย่างเพิ่มขึ้นมีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบดีขึ้น พบจำนวนข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบเพิ่มขึ้น สรุปได้ว่า เมื่อกลุ่มตัวอย่างมีขนาดใหญ่ขึ้น ตรวจพบจำนวนข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบเพิ่มขึ้นและมีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบดีขึ้น

Acar (2012) ได้ศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยใช้วิธี HGLM เปรียบเทียบกับวิธี IRT-LR เมื่อขนาดกลุ่มตัวอย่างต่างกัน จำแนกตามสถานะทางสังคมและเศรษฐกิจ พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี HGLM และ วิธี IRT– LR มีความสอดคล้องกันอย่างมีนัยสำคัญทางสถิติ จำนวนข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบแตกต่างกัน สรุปได้ว่าเมื่อขนาดของกลุ่มตัวอย่างและวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่แตกต่างกัน ทำให้ตรวจพบจำนวนข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบแตกต่างกัน

Acar (2013) ได้เปรียบเทียบค่าสัมประสิทธิ์ความคล้ายคลึงกันระหว่างกลุ่ม จากการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี HGLM และวิธี LR กลุ่มตัวอย่างจำนวน 10,727 คน ซึ่งเป็นนักเรียนในประเทศตุรกี ที่เข้าสอบวิชาวิทยาศาสตร์และสังคมศาสตร์ ปี 2006 เปรียบเทียบระหว่างวิธี HGLM-DIF กับวิธี LR-DIF ผลการศึกษาปรากฏว่า วิธี HGLM-DIF กับวิธี LR-DIF มีความคล้ายคลึงกัน รวมทั้งความสัมพันธ์ของกลุ่มและการตัดค่าสัมประสิทธิ์ความสัมพันธ์ของกลุ่มและการตัดค่าสัมประสิทธิ์จากทั้งสองวิธีค่อนข้างสมบูรณ์แบบ

Ong, Lu, Lee and Cohen (2015) ได้เปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ระหว่างวิธี HGLM วิธี MIMIC และวิธี IRT ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างที่แตกต่างกัน เปรียบเทียบอัตราความคลาดเคลื่อนประเภทที่ 1 (Type I Error) พบว่า วิธี MIMIC มีอัตราความคลาดเคลื่อนประเภทที่ 1 น้อยกว่าวิธี HGLM และวิธี IRT เมื่อกลุ่มตัวอย่างมีขนาดเล็ก หากกลุ่มตัวอย่างมีจำนวนเพิ่มมากขึ้นอัตราความคลาดเคลื่อนก็จะเพิ่มขึ้น วิธี IRT พบอัตราความคลาดเคลื่อนน้อย เมื่อกลุ่มตัวอย่างขนาดเล็ก แต่ถ้ากลุ่มตัวอย่างมีจำนวนเพิ่มมากขึ้นอัตราความคลาดเคลื่อนก็จะเพิ่มขึ้น จึงสามารถสรุปได้ว่า วิธี MIMIC และวิธี IRT พบอัตราความคลาดเคลื่อนน้อยเมื่อกลุ่มตัวอย่างขนาดเล็ก หากกลุ่มตัวอย่างมีจำนวนเพิ่มมากขึ้นอัตราความคลาดเคลื่อนก็จะเพิ่มขึ้น ส่วนวิธี HGLM นั้นเป็นวิธีที่สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบและพบจำนวนข้อสอบที่ทำหน้าที่ต่างกันได้มากที่สุด

จากการศึกษางานวิจัยที่เกี่ยวข้องพบว่า วิธี HGLM เป็นการวิเคราะห์ด้วยโมเดลสมการโครงสร้างเชิงเส้นตรงระดับลดหลั่น เมื่อทำการวิเคราะห์ในแบบทดสอบที่มีความยาวตั้งแต่ 20 ข้อ ขึ้นไป วิธี HGLM สามารถตรวจพบจำนวนข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบได้มากกว่า IRT วิธี BAYESIAN และวิธี MIMIC

ตอนที่ 5 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี MIMIC และงานวิจัยที่เกี่ยวข้อง

แบบจำลองกลุ่มนี้ประกอบด้วยแบบจำลองความสัมพันธ์ทั้งแบบที่มีและไม่มีความคลาดเคลื่อนในการวัดจะประกอบขึ้นด้วยตัวแปรสังเกตได้ทั้งหมด โดยไม่มีตัวแปรแฝง เขียนรูปสมการได้ดังนี้

$$Y = \beta Y + \Gamma X + \zeta \qquad (4)$$

หรือเขียนในรูปเมทริกซ์ได้ดังนี้

$$[Y] = [BE][Y] + [GA][X] + [Z] \quad (5)$$

เมทริกซ์พารามิเตอร์ LY, LX, TD และ TE จึงมีค่าเป็นศูนย์ทั้งหมดการกำหนดข้อมูลจำเพาะของโมเดลกำหนดรูปแบบและสถานะของเมทริกซ์ GA, BE, PH, c]t PH เท่านั้นโมเดลความสัมพันธ์โครงสร้างเชิงสาเหตุที่มีความคลาดเคลื่อนในการวัดมีตัวแปรครบทุกประเภทได้ตามโมเดลใหญ่ในโปรแกรมลิสเรลเมื่อเขียนในรูปสมการจะประกอบด้วยสมการการวัดสองสมการ และสมการโมเดลโครงสร้างหนึ่งสมการ ดังนี้

$$[X] = [LX][K] + [d] \quad (6)$$

$$[Y] = [L Y][E] + [e] \quad (7)$$

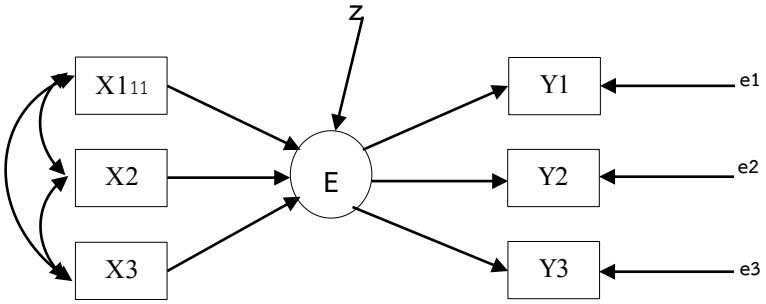
$$[E] = [BE][E] + [GA][K] + [z] \quad (8)$$

แบบจำลองกลุ่มนี้ยังแบ่งออกได้เป็น 3 แบบ ได้แก่

1. Regression Models and ANOVA Models
2. Path Analysis
3. Multiple Indicators and Multiple Causes Models หรือ MIMIC Models

โมเดลมิมิค (MIMIC Model)

MIMIC เป็นคำที่ย่อมาจาก Multiple Indicators and Multiple Causes ซึ่งหมายถึงโมเดลลิสเรลที่มีตัวแปรแฝงเพียงตัวแปรเดียว โดยที่ตัวแปรแฝงนั้นได้รับอิทธิพลจากตัวแปรภายนอกสังเกตได้หลายตัวแปร และส่งอิทธิพลไปยังตัวแปรภายในสังเกตได้หลายตัวแปร กล่าวอีกอย่างหนึ่ง คือเป็นโมเดลลิสเรลของคุณลักษณะแฝงที่มีหลายสาเหตุและวัดได้จากตัวบ่งชี้หลายตัว ดังแสดงตามภาพที่ 5 ในที่นี้มีตัวบ่งชี้ 3 ตัวแปร และมีตัวแปรสาเหตุ 3 ตัวแปรตามลักษณะโมเดลจะเห็นว่าการวัดตัวแปรภายนอกสังเกตได้ ต้องมีข้อตกลงข้างต้นว่า ไม่มีความคลาดเคลื่อนในการวัดและในการวิเคราะห์ข้อมูลจะกำหนดข้อมูลจำเพาะ เฉพาะรูปแบบและสถานะของเมทริกซ์ PH, BE, GA, PS, LY และ TE เท่านั้น ส่วนเมทริกซ์ TD และ LX มีค่าเป็นศูนย์ทั้งหมด โมเดลมิมิคนี้เป็นประโยชน์มากในการตรวจสอบความเป็นเอกมิติ (Unidimensionality) ในการวิจัยสาขาในการวัดผลการศึกษา แสดงดังภาพที่ 2-4



ภาพที่ 2-4 โมเดลย่อยของ MIMIC (Schumacker & Lomax, 2010, p. 294)

การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ (DIF) สามารถแบ่งออกเป็นการวัดองค์ประกอบและโครงสร้างองค์ประกอบ ในองค์ประกอบการวัด y^*_I ของข้อที่ *I* ลักษณะของตัวแปรแฝง y ที่ทดสอบเป็นการออกแบบการวัด และกลุ่มของตัวแปร z (ในที่นี้เป็นการศึกษาเพียง 1 กลุ่มตัวแปร) ที่เกี่ยวข้องกับการทำหน้าที่ต่างกันของข้อสอบ (DIF) ในการวิเคราะห์องค์ประกอบของโมเดล ดังนี้

สูตรสำหรับ MIMIC ของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) คือ

$$y_i^* = \lambda_i \theta + \beta_i z + \varepsilon_i, \tag{9}$$

เมื่อ y_i^* คือ ข้อที่ *i*

θ คือ องค์ประกอบ

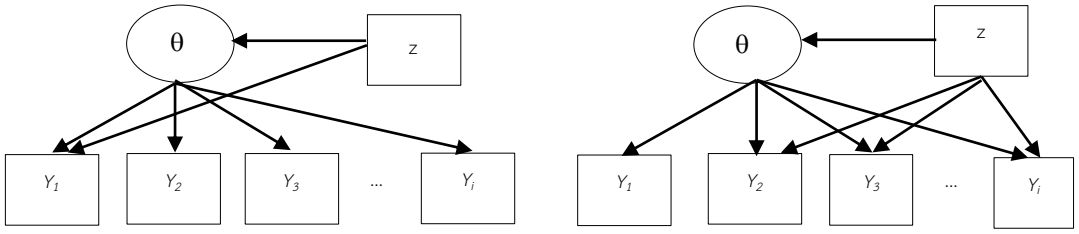
$\boldsymbol{\beta}'_I$ คือ สัมประสิทธิ์ของตัวแปรเพศและสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน

z คือ กลุ่มเพศและสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน

$\boldsymbol{\lambda}_i$ คือ น้ำหนักองค์ประกอบ

$\boldsymbol{\varepsilon}_i$ คือ ค่าความแปรปรวน

เมื่อ $\boldsymbol{\lambda}_i$ เป็นนำหนักองค์ประกอบและเกี่ยวข้องกับความชันของพารามิเตอร์ของข้อสอบข้อที่ I ในบริบทของทฤษฎีการตอบสนองข้อสอบ (IRT) แล้ว $\boldsymbol{\varepsilon}_i$ มีการแจกแจงแบบปกติสำหรับ Ordinal Probit และการแจกแจงแบบโลจิสติก สำหรับ Ordinal Logit และ $\boldsymbol{\beta}'_I$ คืออิทธิพลของกลุ่มตัวแปร z ต่อ y_i^* ถ้า $\boldsymbol{\beta}'_I = 0$ แล้วข้อสอบข้อที่ *i* มีค่าเท่ากันในทุกๆกลุ่ม ตรงกันข้าม ถ้า $\boldsymbol{\beta}'_I \neq 0$ จะเกิดการทำหน้าที่ต่างกันของข้อสอบ (DIF) แบบอเนกรูป เนื่องจากสมการไม่มีเทอมปฏิสัมพันธ์ เป็นตัวทำนาย ดังนั้น สมการ MIMIC จึงใช้แบบเอกรูปได้เพียงอย่างเดียว ดังภาพที่ 2-5



(a) (b)

ภาพที่ 2-5 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี MIMIC แบบเอกรูป
(Wang & Shin, 2010, p. 169)

ซึ่งวิธีหลายตัวชี้วัดหลายสาเหตุในรูปแบบองค์ประกอบเชิงยืนยัน (MIMIC) เป็นหลักการของ CFA กับตัวแปร แล้ววิธี MIMIC ยังสามารถนำไปใช้สำหรับการวิเคราะห์ DIF ได้ด้วย ซึ่งผลที่ได้ต้องมีค่าเป็น แบบ 2 ค่า (Dichotomous) ค่าพารามิเตอร์ของตัวชี้วัดไม่ต่อเนื่องเป็นสิ่งที่จำเป็น ในความเป็นจริงแล้ว มีหลายวิธีที่ตัวชี้วัดของค่าพารามิเตอร์ เป็นแบบ 2 ค่า (Dichotomous) โดยใช้ฟังก์ชันเชื่อมโยงที่เหมาะสม (เช่น การเชื่อมโยงแบบโลจิสหรือโพรบิต) ข้อตกลงเบื้องต้นคือตัวแปรแฝงเป็นตัวแปรต่อเนื่อง

และตัวแปรสังเกตได้เป็นการตอบแบบไบนารี (Binary) เมื่อ y_{ij}^{*} เป็นตัวแปรแฝงแบบ ต่อเนื่อง และตัวแปรสังเกตได้เป็นการตอบแบบ ไบนารี (Binary) ของข้อสอบ y_{ij} แล้วสามารถเขียนสมการได้ดังนี้

$$y_{ij} = \begin{cases} 1 & if \quad y_{ij}^{*} > 0 \\ 0 & if \quad y_{ij}^{*} \leq 0 \end{cases} \tag{10}$$

สูตรสำหรับวิธี MIMIC ของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) คือ

$$y_{ij}^{*} = \lambda_i \theta_j + \beta_i G_j + \varepsilon_{ij}, \tag{11}$$

เมื่อ λ_i เป็นน้ำหนักองค์ประกอบของข้อที่ *I* และ θ_j เป็นลักษณะของตัวแปร ส่วน $\boldsymbol{\beta}_i$ เป็นสัมประสิทธิ์ความชันสำหรับความแปรปรวนร่วม G_j ซึ่งเป็นกลุ่มตัวชี้วัดของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) และ $\boldsymbol{\varepsilon}_{ij}$ เป็นเศษเหลือ นอกจากนี้โมเดลการถดถอยเป็นสิ่งจำเป็นสำหรับการพยากรณ์ตัวแปรแฝง θโดยกลุ่มของตัวชี้วัด G_j เพื่อควบคุมความแตกต่างในลักษณะตัวแปรแฝงข้ามกลุ่มย่อย

$$\eta_j = yz_j + \zeta_j, \tag{12}$$

เมื่อ y เป็นความชันของกลุ่มตัวแปร G_j และ $\boldsymbol{\zeta}_j$ เป็นความคลาดเคลื่อนของสมการถดถอย $\boldsymbol{\beta}_i$ เป็นการทำหน้าที่ต่างกันของข้อสอบ (DIF) เป็นเอกรูป เมื่อ y เป็นผลต่างของค่าเฉลี่ยคุณลักษณะแฝงของกลุ่มเปรียบเทียบกับกลุ่มอ้างอิงและมีเกณฑ์การจับคู่ ตามตัวแปรคงที่ในสมการข้างต้น มีข้อตกลงกำหนดให้เป็น 0 ซึ่งจะไม่ปรากฏในสมการข้างต้น

$$a_i = \frac{\lambda_i \sqrt{\sigma_n^2}}{\sqrt{1-\lambda_1^2 \sigma_\zeta^2}}, \tag{13}$$

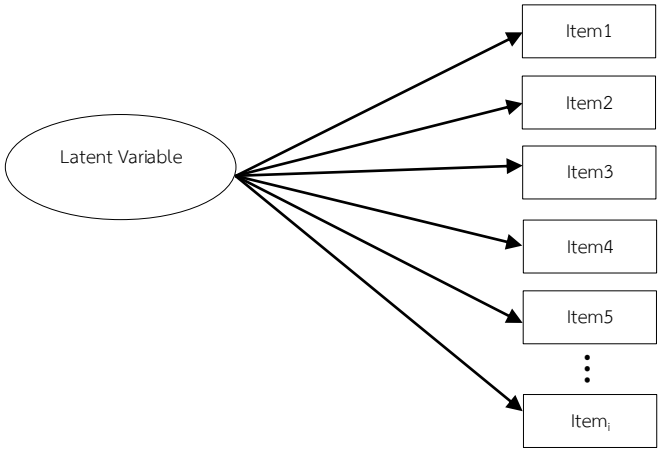
$$b_i = \frac{\left|(\tau_i - \beta_i z)\lambda_i^{-1} - \mu_\eta\right|}{(\sigma_\eta^2)^{1/2}}, \tag{14}$$

เมื่อ $\sigma^2{}_\eta$ เป็นตัวแปรสำหรับองค์ประกอบ θ_j และ $\sigma\zeta^2$ เป็นตัวแปรของความคลาดเคลื่อนของสมการถดถอยเชิงเส้นตรง ζ_j สำหรับการทำนายองค์ประกอบทั่วไป $\boldsymbol{\tau}_i$ เป็นความยากของข้อสอบข้อที่ *I* และ μ_η เป็นค่าเฉลี่ยขององค์ประกอบทั่วไป θ_j

ข้อดีหลายประการของการใช้โมเดล MIMIC ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ที่แสดงข้างต้น แสดงขนาดของการทำหน้าที่ต่างกันของข้อสอบ (DIF) โดยใช้หลักทฤษฎีการตอบสนองข้อสอบ (IRT) ประมาณค่าการทำหน้าที่ต่างกันของข้อสอบ (DIF) จากค่าพารามิเตอร์ตามทฤษฎีการตอบสนองข้อสอบ (IRT) ซึ่งเป็นประโยชน์ต่อผู้ปฏิบัติ

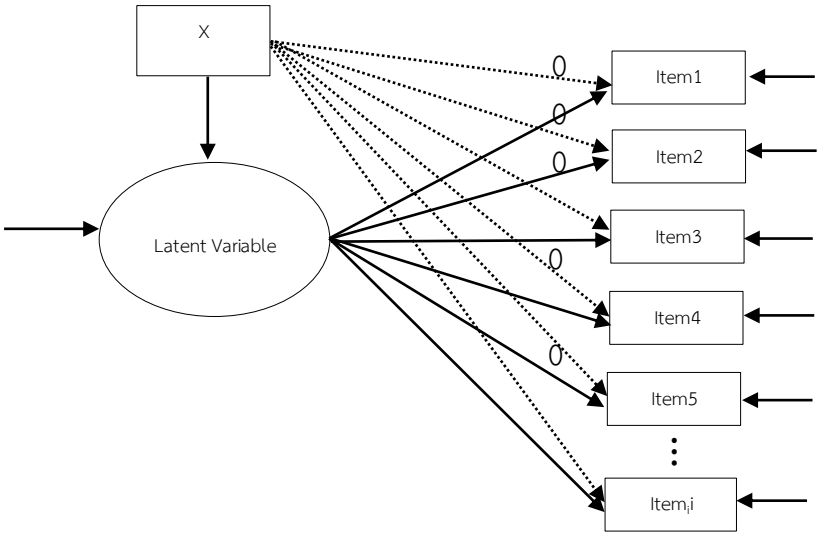
ในการวัดสิ่งต่างๆ (Measurement) สิ่งที่น่าสนใจคือความแตกต่างของกลุ่มในตัวแปรแฝง (Latent Variable) ในการศึกษาความแตกต่างของค่าเฉลี่ยของตัวแปรแฝงเป็นการศึกษาความไม่แปรเปลี่ยนของกลุ่ม (Invariant) ในขณะที่การศึกษาความแตกต่างของตัวแปรสังเกตได้ เช่น ค่าเฉลี่ยของข้อคำถามในแต่ละกลุ่ม ซึ่งการศึกษาการทำหน้าที่ต่างกันของข้อสอบเป็นการศึกษาความแตกต่างของตัวแปรสังเกตได้หรือตัวชี้วัด

โดยปกติรูปแบบการวิเคราะห์ข้อมูลตามทฤษฎีการตอบสนองข้อสอบ (IRT) เป็นการศึกษาคุณสมบัติบนพื้นฐานข้อตกลงความเป็นเอกมิติ (Unidimentsional) ของตัวแปรแฝงซึ่งสังเกตไม่ได้โดยตรง สำหรับตัวแปรแฝงในโมเดล IRT จะดูจากค่าเซต้า (θ) ซึ่งสามารถประมาณค่าได้โดยตรง ซึ่งมีอิทธิพลตรงต่อตัวชี้วัดหรือข้อคำถามที่สังเกตได้ ซึ่งเราสามารถอธิบายโมเดลการวิเคราะห์องค์ประกอบ ดังภาพที่ 2-6



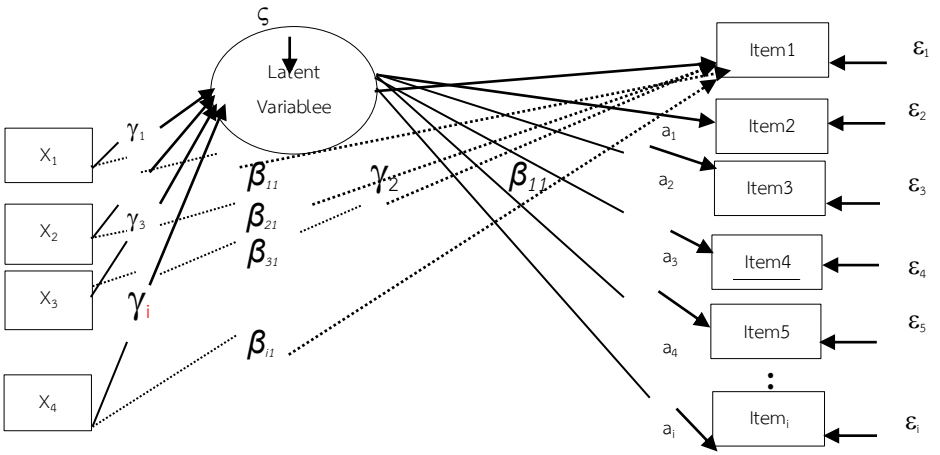
ภาพที่ 2-6 โมเดลการวิเคราะห์องค์ประกอบตามแนวคิด IRT (Riley & Dennis, 2015, p. 8)

จากภาพถ้าข้อคำถามหรือตัวชี้วัดเป็นตัวแปรจัดกลุ่ม (Dichotomous) และตัวแปรแฝงมีการแจกแจงแบบปกติ (Normal Distribution) ซึ่งมีลักษณะเช่นเดียวกับการแจกแจงโค้งความถี่สะสม ในโมเดล IRT น้ำหนักองค์ประกอบที่เกิดขึ้นบนตัวชี้วัดจะหมายถึง ค่าดัชนีประมาณค่าอำนาจจำแนกตามทฤษฎีการตอบสนองข้อสอบ ในขณะที่ค่าเฉลี่ย (Intercepts) ของแต่ละข้อคำถามคือค่าประมาณความยาก (Difficulty) ตามทฤษฎีการตอบสนองข้อสอบในกรณีที่มีตัวแปรแฝงมากกว่าหนึ่งตัวโปรแกรมที่พัฒนามาใช้ตามทฤษฎี IRT โดยปกติจะอนุมาน (Assumes) ว่ามีข้อมูลเป็นลักษณะมีความเป็นเอกมิติ (Fleishman, 2003, p. 6) ในการนำมาประยุกต์ใช้ในการวิจัยจึงทำได้กว้างยิ่งขึ้น จึงง่ายต่อการนำแนวคิดมาประยุกต์ใช้ในกรณีที่ต้องการนำแนวคิดของโมเดล MIMIC มาใช้ในกรณีที่ตัวแปรแฝงมีหลายมิติ (Multi – Dimensional) หลายองค์ประกอบ (Multi – Factor) การนำโปรแกรม MIMIC มาใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแสดงได้โดยใช้ตัวแปรสาเหตุ (Causes) เพียงตัวเดียว ดังภาพที่ 2-7



ภาพที่ 2-7 โมเดลการวิเคราะห์การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ MIMIC Model โดยใช้ตัวแปรสาเหตุ 1 ตัว (Brown, 2014, p. 308)

จากภาพที่ 2-7 เป็นการวิเคราะห์องค์ประกอบโดยใช้ตัวแปรทำนาย (X) จำนวนหนึ่งตัวแปรในการทำนายตัวแปรแฝงที่ประกอบด้วยตัวแปรสังเกตได้ที่เป็นข้อสอบหรือข้อคำถาม (Item) จำนวน I ตัว โดยการจำกัดความคลาดเคลื่อนจากการวัดความคลาดเคลื่อนจากการวัดของตัวแปรแฝง และให้อิสระกับความคลาดเคลื่อนของตัวแปรที่สอดคล้องกับโมเดลมากกว่าการประมาณค่าระหว่าง X กับตัวแปรแฝง (Latent Variable) อิทธิพลตรงของตัวแปร X ที่ทำนาย Item หลังจากที่มีอิทธิพลตรงไปยังตัวแปรแฝงแสดงทิศทางเดียว (Uniform) ในการทำหน้าที่ต่างกันของข้อสอบ (DIF) ซึ่งเป็นสิ่งที่แสดงความลำเอียง (Biased) ที่เกิดจากข้อสอบหรือข้อคำถามหรืออธิบายได้ว่าถ้าข้อสอบหรือข้อคำถาม (Item) ได้รับอิทธิพลอย่างมีนัยสำคัญจากตัวแปรสาเหตุ X แสดงให้เห็นว่าข้อสอบ หรือข้อคำถาม (Item) ขึ้นอยู่กับตัวแปรสาเหตุ X ไม่ได้อธิบายตัวแปรแฝงแสดงว่าข้อสอบหรือข้อคำถาม (Item) ข้อนั้นทำหน้าที่ต่างกันหรือมีความลำเอียง (Biased) ดังภาพที่ 2-8



ภาพที่ 2-8 โมเดลการวิเคราะห์การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ MIMIC Model

โดยใช้ตัวแปรสาเหตุมากกว่า 1 ตัว (Riley & Dennis, 2015, p. 10)

จากภาพเป็นการวิเคราะห์องค์ประกอบโดยใช้ตัวแปรทำนาย X_1 - X_i ในการทำนายตัวแปรแฝงที่ประกอบด้วยตัวแปรสังเกตได้ที่เป็นข้อสอบหรือข้อคำถาม (Item) จำนวน *I* ตัวในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) สามารถดำเนินการตามขั้นตอนดังนี้

1. วิเคราะห์องค์ประกอบในโมเดลวัดที่ประกอบด้วยตัวแปรแฝงและข้อคำถาม
2. เพิ่มความแปรปรวนร่วมในการทดสอบโมเดล
3. เพิ่มอิทธิพลตรงกับตัวแปรแฝง (γ) อิทธิพลตรง (a) และกำหนดให้มีค่าเท่ากับศูนย์
4. ตรวจสอบโมเดลดัชนีปรับแก้ (Modification Indices)
5. เพิ่มอิทธิพลตรงจากความแปรปรวนร่วมกับข้อคำถามที่มีค่าดัชนีปรับแก้สูงสุด
6. ดำเนินการในขั้นตอนที่ 4-5 จนกว่าจะไม่พบค่าดัชนีปรับแก้ (M.I) ที่ไม่มีนัยสำคัญประเมินความสอดคล้องของโมเดลและอิทธิพลตรง (β_i)

งานวิจัยที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี MIMIC มีดังนี้

โกศล จิตวิรัตน์ ทักษิณา เครือหงส์ และเนตร์พัณณา ยาวิราช (2554) ได้ศึกษาศักยภาพของโปรแกรม Mplus ในการวิเคราะห์ข้อมูลทางสถิติขั้นสูงในงานวิจัย เนื่องจากเป็นโปรแกรมที่มีศักยภาพสูงสุด สามารถวิเคราะห์โมเดลสมการโครงสร้างพหุระดับ ซึ่งเป็นวิธีการวิทยาการวิจัยที่พัฒนาการวิเคราะห์ข้อมูลให้สอดคล้องกับสภาพความสลับซับซ้อนตามความเป็นจริงของสภาพข้อมูลได้มากที่สุดในปัจจุบัน และยังเป็นโปรแกรมที่มีศักยภาพในการวิเคราะห์ข้อมูลทางสถิติขั้นสูงในงานวิจัยครอบคลุมทั้งโมเดลเดี่ยวและโมเดลพหุระดับ ในการจัดเตรียมข้อมูลผู้วิจัยต้องตระหนักถึงรายละเอียดตัวแปรที่จะต้องนำไประบุในชุดคำสั่ง สำหรับการระบุชื่อและชนิดตัวแปรที่ใช้ในการวิเคราะห์ โดยการจัดเตรียมข้อมูลนำเข้าสามารถจัดเตรียมด้วยโปรแกรม SPSS เสร็จแล้วแปลงไฟล์ข้อมูลเป็นภาษา ASCII ก่อนการวิเคราะห์ข้อมูลด้วยโปรแกรม Mplus

รติพร ถึงฝั่ง (2556) ได้ศึกษาการวิเคราะห์โมเดลมิมิค: การใช้ประโยชน์จากโปรแกรม LISREL รุ่นทดลองใช้เพื่องานวิจัย เพื่อเสนอวิธีการวิเคราะห์โมเดลมิมิคซึ่งสามารถนำมาใช้ประโยชน์ในการวิเคราะห์ข้อมูลงานวิจัย โมเดลมิมิคเป็นการวิเคราะห์ข้อมูลในลักษณะที่ตัวแปรสังเกตได้ (x-variables) หลายๆตัวแปรทำนายหรือส่งผลต่อตัวแปรแฝง (Eta) ซึ่งตัวแปรแฝงวัดได้จากตัวบ่งชี้ (y- variables) หลายตัวแปร ขั้นตอนการวิเคราะห์โมเดลมิมิคเริ่มต้นด้วยการนำเข้าข้อมูลจากโปรแกรม SPSS และตามด้วยการวิเคราะห์โมเดลมิมิคด้วยโปรแกรม LISREL ซึ่งมี 5 ขั้นตอนสำคัญ คือ 1) การเตรียมข้อมูลด้วยโปรแกรม PRELIS 2) การระบุโมเดลหรือวาดภาพโมเดลการวิจัย 3) การกำหนดการแสดงผลการวิเคราะห์ 4) การวิเคราะห์โมเดล และ 5) การปรับโมเดล ผลการวิเคราะห์จะประกอบไปด้วยค่าสถิติแสดงความสอดคล้องระหว่างโมเดลตามสมมติฐานกับข้อมูลเชิงประจักษ์และค่าสถิติแสดงผลของตัวแปรสังเกตที่มีต่อตัวแปรแฝง

Finch (2005) ได้เปรียบเทียบความสามารถของตัวแบบหลายตัวบ่งชี้หลายสาเหตุ (MIMIC) รูปแบบการวิเคราะห์ปัจจัยยืนยันเพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ได้อย่างถูกต้องด้วยวิธีการที่กำหนดขึ้น แม้ว่าแบบจำลอง MIMIC อาจมีแอพพลิเคชันในการระบุ DIF สำหรับตัวแปรหลายๆแบบ แต่ก็มีการตรวจสอบว่าเทคนิคการทำงานของ DIF มีความถูกต้องและไม่ถูกต้องอย่างไร

การใช้วิธีการมอนติคาร์โล การศึกษาครั้งนี้การจัดการจำนวนรายการ จำนวนผู้เข้าสอบ ความแตกต่างระหว่างความสามารถเฉลี่ยของกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ ระดับการตรวจสอบ DIF ของรายการหลักและจำนวนข้อสอบที่พบ DIF ในรายการหลัก ผลการศึกษาปรากฏว่า แบบจำลอง MIMIC มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นอย่างดี

Wang, Ching, and Chin (2009) ได้ศึกษาการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี MIMIC พร้อมประเมินประสิทธิภาพของโมเดล พบว่าวิธี MIMIC ตรวจสอบประสิทธิภาพได้สูงกว่ามาตรฐานของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ มีความสมดุลกันระหว่างกลุ่มหรือเปอร์เซ็นของการทำหน้าที่ต่างกันของข้อสอบเพียงเล็กน้อย

Mucherah, Finch, and Keaikitse (2012) ได้ศึกษาการทำหน้าที่ต่างกันของแบบวัดมโนภาพแห่งตนหรือแบบสอบถามที่อธิบายความเป็นตัวตนของชาวเคนย่า จำแนกตามเพศโดยใช้แบบจำลองสมการเชิงโครงสร้าง (MIMIC Model) การศึกษาเพื่อวิจัยว่า วัยรุ่นมีความเข้าใจความคิดของตัวเองมากน้อยเพียงไร จึงได้มีการตรวจสอบการทำหน้าที่ต่างกันของแบบทดสอบ โดยใช้วิธี MIMIC และวิธี SIBTEST กลุ่มตัวอย่างเป็นนักเรียนจากประเทศเคนย่า ซึ่งศึกษาอยู่ในระดับมัธยมศึกษา จำนวน 1,990 คน เป็นเพศชาย 983 คน เพศหญิง 1,007 คน ใช้แบบวัดมโนภาพแห่งตน (SDQ) จำนวนแบบทดสอบ 135 ข้อ ผลการศึกษาปรากฏว่า เพศหญิงจะตอบคำถามในเชิงบวก แต่ในวิชาคณิตศาสตร์เพศหญิงจะตอบคำถามได้น้อยกว่า และวิธีการตรวจสอบการทำหน้าที่ต่างกันด้วยวิธี MIMIC อยู่ในเกณฑ์ที่ใช้ได้ดี

จากการศึกษางานวิจัยที่เกี่ยวข้อง วิธี MIMIC มีข้อดีหลายประการของการใช้โมเดล MIMIC ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) วิธีนี้แสดงขนาดของการทำหน้าที่ต่างกันของข้อสอบ โดยใช้หลักทฤษฎีการตอบสนองข้อสอบ (IRT) ประมาณค่าการทำหน้าที่ต่างกันของข้อสอบ (DIF) จากค่าพารามิเตอร์ตามทฤษฎีการตอบสนองข้อสอบ (IRT) ซึ่งมีประโยชน์ต่อการวิเคราะห์ข้อมูล

ตอนที่ 6 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี BAYESIAN และงานวิจัยที่เกี่ยวข้อง

ในทฤษฎีความน่าจะเป็น สถิติการอนุมานและปัญญาประดิษฐ์บางครั้งจะพบคำว่า แบบเบส์ (BAYESIAN) มาขยายชื่อทฤษฎีหรือโมเดลต่าง ๆ โดยทุกครั้งที่พบคำขยายนี้ หมายความว่า ได้มีการนำปรัชญาหรือหลักการของทฤษฎีความน่าจะเป็นแบบ BAYESIAN (บางท่านเรียกการอนุมานแบบเบส์หรือสถิติแบบเบส์) มาใช้กับสาขาความรู้นั้นๆ ทฤษฎีความน่าจะเป็นแบบ BAYESIAN แปลความหมายของคำว่า ความน่าจะเป็น เป็นความเชื่อมั่นส่วนบุคคลในเหตุการณ์หนึ่งๆซึ่งต่างจากทฤษฎีความน่าจะเป็นของคอลโมโกรอฟ (ที่มักถูกเรียกว่าทฤษฎีความน่าจะเป็นเชิงความถี่) ซึ่งมักแปลความหมายของความน่าจะเป็น (โดยต้องแปลควบคู่ไปกับการทดลองเสมอ) ดังนั้นความน่าจะเป็นของเหตุการณ์ A คืออัตราส่วนของจำนวนครั้งของเหตุการณ์ A ที่ทดลองสำเร็จเทียบกับจำนวนครั้งที่ทดลองทั้งหมด จุดแตกต่างสำคัญระหว่างทฤษฎีทั้งสองประเภท คือ ความหมายความน่าจะเป็น พวก BAYESIAN มองความน่าจะเป็น เป็นความเชื่อส่วนบุคคล พวกเชิงความถี่มองความน่าจะเป็นเป็นคุณสมบัติหนึ่งที่ถูกฝังอยู่ในวัตถุ (ไม่ขึ้นอยู่กับตัวบุคคล)

การนำทฤษฎีไปใช้งาน ในการนำทฤษฎีความน่าจะเป็นเชิงความถี่ไปใช้จะต้องมีการทดลองเชิงแนวคิด (Conceptual Experiment) ควบคู่ไปด้วยเสมอ เหตุการณ์ใด ๆ ก็ตามที่ไม่มีการทดลองเชิงแนวคิดที่สมเหตุสมผลพอ จะไม่สามารถนำทฤษฎีความน่าจะเป็นเชิงความถี่ไปใช้งานได้ แต่สามารถนำทฤษฎีความน่าจะเป็นแบบ BAYESIAN มาอ้างความน่าจะเป็นประเภทนี้ได้ในมุมมองว่าความน่าจะเป็นแบบ BAYESIAN สามารถนำมาประยุกต์ใช้งานได้กว้างขวางมากกว่า กล่าวโดยสรุปทฤษฎีความน่าจะเป็นแบบ BAYESIAN มีปรัชญาที่ต่างจากทฤษฎีความน่าจะเป็นเชิงความถี่เกือบสิ้นเชิง ถึงแม้จะมีสัจพจน์พื้นฐานแบบเดียวกัน โดยในทฤษฎีความน่าจะเป็นแบบ BAYESIAN นั้นมองความน่าจะเป็นสถิติหรือการอนุมานเป็นเรื่องเดียวกัน

การประมาณค่าพารามิเตอร์ด้วยวิธีของ BAYESIAN มีแนวคิดว่า ค่าความสามารถ θ และค่าพารามิเตอร์ของข้อสอบ a, b และ c เป็นตัวแปรสุ่ม (Random Variable) จากการแจกแจงที่แสดงได้ด้วยฟังก์ชันความหนาแน่นร่วม (Joint Density Function) f(θ, b, a, c) และเรียกฟังก์ชัน (θ, b, a, c) นี้ว่า การแจกแจงเริ่มแรก (Prior Distribution) ของค่า θ, b, a, และ c ซึ่งทำให้การใช้ L(U/(θ, b, a, c) เพียงอย่างเดียวในการประมาณค่า θ, b, a และ c ถูกพิจารณาว่าเป็นการใช้ข้อมูลที่มีอยู่อย่างไม่ครบถ้วน เพราะยังมีการแจกแจงเริ่มต้นร่วมกับ f(θ, b, a, c) ที่ควรนำมาใช้ในการประมาณค่าพารา มิเตอร์ด้วย กล่าวคือถ้าพิจารณาความน่าจะเป็นร่วมของการตอบข้อสอบ P(U/θ, b, a, c) จะเห็นว่าการแจกแจงของตัวแปร U ขึ้นอยู่กับค่าความสามารถ θ และค่าพารามิเตอร์ของข้อสอบ a, b และ c ถ้าค่า θ, a, b และ c เปลี่ยนไปโอกาสที่ตัวแปร U มีค่าเท่ากับ u ก็จะเปลี่ยนไปด้วย ดังนั้นการทราบผลการตอบข้อสอบ u จึงน่าจะช่วยทำให้ทราบค่า θ, a, b และ c ได้ดียิ่งขึ้น ซึ่งอาจแสดงได้ด้วยการแจกแจงอย่างมีเงื่อนไขของค่า θ, a, b และ c เมื่อทราบผลการตอบข้อสอบ f(θ, b, a, c/U) และเรียกว่า การแจกแจงภายหลัง (Posterior Distribution)

การประมาณค่าด้วยวิธีของ BAYESIAN ดังกล่าว ทำให้สามารถจำแนกกระบวนการดำเนินการตามแนวคิดออกเป็น 2 กระบวนการ ดังนี้

1. กระบวนการกำหนดลักษณะของการแจกแจงเริ่มแรก มี 2 ขั้น (Swaminathan & Gifford, 1985, pp. 589-601) คือ

1.1 กำหนดให้การแจกแจงเริ่มแรกของค่าความสามารถ (θ) ค่าอำนาจจำแนก (a) ค่าความยาก (b) และค่าการเดา (c) เป็นอิสระต่อกัน

$$f(\theta, b, a, c) = f(\theta) \cdot f(b) \cdot f(a) \cdot f(c) \qquad (15)$$

กำหนดลักษณะของการแจกแจงของ f(θ) . f(b) . f(a) และ f(c) ดังนี้

1.1.1 การแจกแจงเริ่มแรกของค่าความสามารถ f(θ) มีข้อตกลงว่าสารสนเทศที่มีมาก่อนของค่าความสามารถของผู้เข้าสอบแต่ละคนไม่แตกต่างกัน สามารถใช้แทนกันได้ (Exchangeability) และค่าความสามารถเป็นตัวแปรสุ่มที่มีการแจกแจงเป็นปกติ (Normal Distribution)

1.1.2 การแจกแจงเริ่มแรกของค่าความยากของข้อสอบ f(b) อาจใช้กระบวนการเดียวกับการกำหนดการแจกแจงเริ่มแรกของค่าความสามารถ คือ มีข้อตกลงว่า f(b) มีการแจกแจงเป็นปกติหรืออาจจะไม่กำหนดการแจกแจงเริ่มแรกไว้ก็ได้ (Swaminathan & Gifford, 1985, p. 350)

1.1.3 การแจกแจงเริ่มแรกของค่าอำนาจจำแนกข้อสอบ f(a) เนื่องจากค่าอำนาจจำแนกของข้อสอบโดยทั่วไปจะต้องเป็นค่าบวก และเป็นความชั้นของเส้นโค้งลักษณะของข้อสอบ ณ จุดเปลี่ยน

โค้ง ดังนั้น การแจกแจงเริ่มแรกของค่าอำนาจจำแนก a_i จึงควรเป็นการแจกแจงแบบไคว์ (Chi Distribution)

1.1.4 การแจกแจงเริ่มแรกของค่าการเดาของข้อสอบ f(c) เนื่องจากค่าพารามิเตอร์ c_i จะมีขอบเขตอยู่ตั้งแต่ 0-1 ดังนั้น จึงมีข้อตกลงว่าควรมีการแจกแจงเบต้า (Beta Distribution)

1.2 กำหนดค่าที่เป็นตัวเลขของพารามิเตอร์ สำหรับการแจกแจงเริ่มแรก

1.2.1 พารามิเตอร์ของการแจกแจงเริ่มแรกของค่าความสามารถ θ ได้แก่ μ_θ และ σ_θ^2 อาจกำหนดให้ μ_θ = 0 และ σ_θ^2 = 1 ซึ่งจะทำให้การประมาณค่าความสามารถ θ มีความสะดวกและประมาณค่าได้รวดเร็วขึ้น (Swaminathan & Gifford, 1985, pp. 351-355)

1.2.2 พารามิเตอร์จากการแจกแจงเริ่มแรกของค่าความยาก b หากมีการกำหนดลักษณะการแจกแจงไว้ ก็ใช้ค่าเดียวกับพารามิเตอร์ของการแจกแจงเริ่มแรกของค่าความสามารถ θ

1.2.3 พารามิเตอร์ของการแจกแจงเริ่มแรกของค่าอำนาจจำแนก a ได้แก่ V_j W_j การกำหนดค่าของ V_j และ W_j ที่เหมาะสม อาจจะหาได้จากการกำหนดพิสัย (Range) ของค่า a คือ ถ้าให้ H เป็นขีดจำกัดบนของพิสัย และให้ L เป็นขีดจำกัดล่างของพิสัย จะหาค่า V_j และ W_j ได้จากสูตร

$$v_j = \frac{1}{2}(1+Z_{1/2}((H+L)/(H-L))^2) \quad (16)$$

$$w_j = \frac{1}{2}((H-L)/Z_{(1/2)\alpha})^2 \quad (17)$$

เมื่อ

$Z_{(1/2)\alpha}$ = ค่า Z ของการแจกแจงปกติมาตรฐาน (Standard Normal Distribution) ที่ระดับนัยสำคัญ α

V_j = Degree of Freedom

1.2.4 พารามิเตอร์ของการแจกแจงเริ่มแรกของค่าการเดา c ได้แก่ s_i และ t_i การกำหนดค่าของ s_i และ t_i ที่เหมาะสมอาจจะหาได้จากการสังเกตสัดส่วนการตอบถูกของกลุ่มผู้เข้าสอบที่มีความสามารถในระดับต่ำมาก กล่าวคือ ถ้าให้ m แทนจำนวนผู้เข้าสอบที่มีความสามารถระดับต่ำมาก และให้ M แทนสัดส่วนการตอบถูกของกลุ่ม m แล้วจะหาค่า s_i และ t_i ได้จากสูตร

$$S_i = mM \quad \text{และ} \quad t_i = m(1-M) - 2 \quad (18)$$

2. กระบวนการประมาณค่าพารามิเตอร์ของข้อสอบและความสามารถของผู้เข้าสอบ กระบวนการประมาณค่าพารามิเตอร์ด้วยวิธีของ BAYESIAN คือ การหาค่าประมาณ θ_l (l= 1,2,...,N) ปกติจะหาค่า θ_i , b_i , a_i, และ c_i (l = 1,2,...,n) ที่ทำให้ฟังก์ชันการแจกแจงภายหลัง $f(\theta, b, a, c/u)$ มีค่าสูงสุด ปกติจะหาค่า θ_i , b_i , a_i, และ c_i ที่ทำให้ $ln\, f(\theta, b, a, c/u)$ มีค่าสงสุด ถ้า $ln\, f(\theta, b, a, c/u)$ เป็นฟังก์ชันที่อนุพันธ์ได้ การหารค่า θ_i , b_i , a_i, และ c_i จะหาได้จากการอนุพันธ์ของ $ln\, f(\theta, b, a, c/u)$ มีค่าเท่ากับศูนย์

งานวิจัยที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี BAYESIAN มีดังนี้

ชนะศึก นิชานนท์ ศิริชัย กาญจนวาสี และ Wilson (2554) ได้ศึกษาประสิทธิภาพของการประมาณค่าพารามิเตอร์แบบเบส์โดยใช้การสรุปอ้างอิงความน่าเชื่อถือของโมเดลการตอบสนอง

ข้อสอบ เพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าพารามิเตอร์ของวิธีการสรุปอ้างอิงความน่าเชื่อถือของโมเดลการตอบสนองข้อสอบ (Generalizability in Item Response) 4 รูปแบบ ได้แก่ รูปแบบที่ 1 Original GIRM พัฒนาโดย Brigg and Wilson (2007) รูปแบบที่ 2 AGIRM A รูปแบบที่ 3 AGIRM B และรูปแบบที่ 4 Numerical Bayesien GIRM ผู้วิจัยเป็นผู้พัฒนาขึ้น ผลการวิจัยปรากฏว่าความลำเอียง ในการประมาณค่า รูปแบบที่ 1 กับ รูปแบบที่ 4 ให้ค่าประสิทธิภาพสูงที่สุด โดยรูปแบบที่ 4 สามารถประมาณค่าพารามิเตอร์ได้เฉพาะลักษณะการแจกแจงเริ่มแรกของผู้สอบและข้อสอบแบบปกติ สำหรับข้อสอบที่ไม่มีลักษณะการแจกแจงเริ่มแรกแบบปกติ พบว่า รูปแบบที่ 1 ให้ค่าประสิทธิภาพสูงที่สุด และเมื่อพิจารณาในด้านประสิทธิภาพขององค์ประกอบความแปรปรวนยุคลิด พบว่า รูปแบบที่ 2 ให้ค่าประสิทธิภาพสูงที่สุด ส่วนการแจกแจงเริ่มแรกของข้อสอบ พบว่า ส่งผลต่อการวัดประสิทธิภาพความลำเอียงในการประมาณค่าและความไม่แน่นอนในการประมาณค่าทุกรูปแบบ และส่งผลการวิเคราะห์ประสิทธิภาพองค์ประกอบความแปรปรวนยุคลิดเฉพาะในกรณีที่แจกแจงเริ่มแรกของผู้สอบเป็นแบบแกมม่าเท่านั้น

อัชฌา อระวีพร (2554) ได้ศึกษาการหาค่าตัวประมาณเบส์ด้วยโปรแกรมวินบั๊ก โปรแกรมวินบั๊กเป็นโปรแกรมทางสถิติสำหรับการประมาณค่าประมาณเบส์ โดยใช้วิธีของมาร์คอฟ เชน มอนติคาร์โล (MCMC) ในการประมาณค่าพารามิเตอร์ ตัวประมาณเบส์เป็นวิธีหนึ่งที่นิยมใช้เนื่องจากมีการใช้ฟังก์ชันการแจกแจง โดยหลักเกณฑ์มาช่วยในการประมาณค่าพารามิเตอร์ ซึ่งวิธีการนี้ค่อนข้างยุ่งยากในการพิสูจน์ในรูปแบบของการแจกแจง แต่โปรแกรมวินบั๊กสามารถช่วยคำนวณค่าของตัวประมาณจากการแจกแจงภายหลังจากตัวประมาณเบส์ โดยผู้ใช้โปรแกรมไม่จำเป็นต้องพิสูจน์ให้ได้รูปแบบของการแจกแจงก็สามารถประมาณค่าได้

อัชฌา อระวีพร (2555) ได้ศึกษาการวิเคราะห์เบส์จากโปรแกรมวินบั๊กสู่โปรแกรมอาร์ จากการศึกษาโปรแกรมวินบั๊กเป็นโปรแกรมทางสถิติสำหรับการประมาณค่าประมาณเบส์ โดยใช้วิธีของมาร์คอฟ เชน มอนติคาร์โล (MCMC) ซึ่งชุดคำสั่ง R2WinBUGS สร้างขึ้นเพื่ออำนวยความสะดวกให้ผู้ใช้สามารถเรียกโปรแกรมวินบั๊กได้จากโปรแกรมอาร์ ซึ่งสามารถเขียนคำสั่ง ข้อมูล แลประมวลผล โดยใช้โปรแกรมวินบั๊กพร้อมๆกับโปรแกรมอาร์ โดยผลลัพธ์ที่ได้สามารถเรียกดูได้จากโปรแกรมอาร์ ซึ่งผลลัพธ์ที่ได้จากตัวประมาณที่ได้ให้ค่าที่ใกล้เคียงกัน

สุพัฒนา หอมบุปผา ไพรัตน์ วงษ์นาม และสมพงษ์ ปั้นหุ่น (2556) ได้ศึกษาการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN ซึ่งในการนำทฤษฎีการตอบสนองข้อสอบมาใช้ ผู้วิจัยเลือกวิธีที่เหมาะสมแล้ว วิธีประมาณค่าพารามิเตอร์ของข้อสอบและความสามารถของผู้เข้าสอบก็เป็นอีกกระบวนการหนึ่งที่จำเป็นต้องเลือกใช้ให้เหมาะสมกับสภาพการวัดแต่ละครั้ง สำหรับทฤษฎีการตอบสนองข้อสอบ วิธีการประมาณค่าพารามิเตอร์ของข้อสอบและความสามารถของผู้เข้าสอบ มีหนึ่งวิธีที่น่าสนใจ คือ วิธีของเบส์ (BAYESIAN ESTIMATION)

Saengla Huffer and Kamata (2006) ได้ศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ สำหรับการปรับให้เกิดผลต่างระหว่างหน่วยด้วยวิธี BAYESIAN การศึกษาครั้งนี้เป็นการใช้ Logistic Regression Model ประเมินความแตกต่างของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบ 3 ระดับ มีการประมาณค่าพารามิเตอร์โดยใช้วิธี BAYESIAN ด้วยโปรแกรม WinBUGS 1.4

จากการศึกษางานวิจัยที่เกี่ยวข้อง ปรากฏว่า ปัจจุบันการใช้โปรแกรมสำเร็จรูปวิเคราะห์ข้อมูลมีปัญหาทางลิขสิทธิ์ทำให้การวิเคราะห์ข้อมูลและนำเสนอผลที่ได้จากการวิเคราะห์ทางสถิติมีปัญหา แต่โปรแกรม WinBUGS สามารถดาวโหลดโปรแกรมได้โดยไม่เสียค่าใช้จ่าย นอกจากนี้โปรแกรม WinBUGS ยังสามารถช่วยแก้ปัญหาสำหรับผู้ที่ไม่เข้าใจการประมาณค่าในตัวแบบ BAYESIAN ซึ่งต้องใช้สถิติเชิง อนุมานและการพิสูจน์รูปแบบการแจกแจงทางสถิติ ก็สามารถประมาณค่าพารามิเตอร์จากฟังก์ชันการแจกแจงภายหลังซึ่งผลลัพธ์ที่ได้จะให้ค่าที่ใกล้เคียงกัน

บทที่ 3
วิธีดำเนินการวิจัย

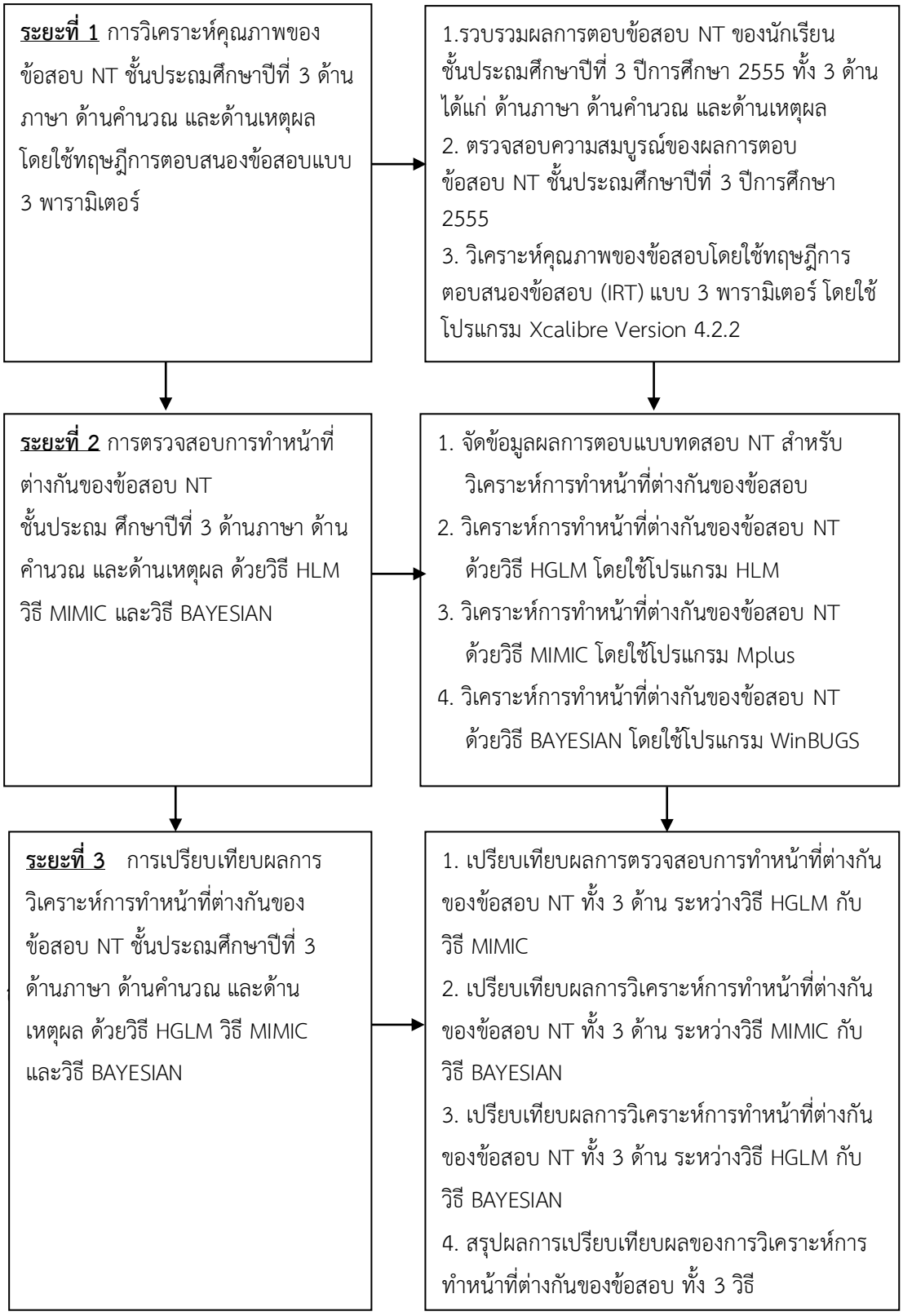
การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ความสามารถทั้ง 3 ด้าน ได้แก่ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ตามหลักทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN และเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN ซึ่งมีการดำเนินการวิจัยเป็น 3 ระยะ ดังนี้

ระยะที่ 1 การวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล โดยใช้หลักการของทฤษฎีการตอบสนองข้อสอบ (IRT) แบบ 3 พารามิเตอร์

ระยะที่ 2 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN

ระยะที่ 3 การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN

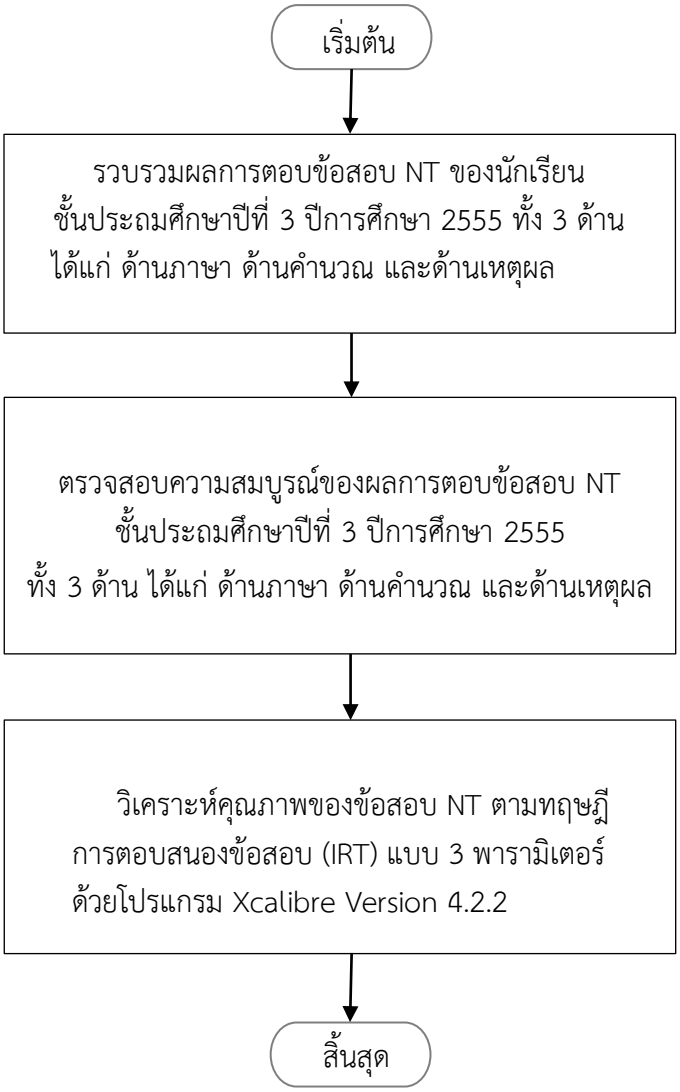
โดยมีขั้นตอนการดำเนินการวิจัย ดังภาพที่ 3-1



ภาพที่ 3-1 ขั้นตอนการดำเนินการวิจัย

ระยะที่ 1 การวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้าน ได้แก่ ด้านภาษา ด้านคำนวณ และด้านเหตุผล โดยใช้หลักการของทฤษฎีการตอบสนองข้อสอบ (IRT) แบบ 3 พารามิเตอร์

การวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้าน ได้แก่ 1) ด้านภาษา 2) ด้านคำนวณ และ 3) ด้านเหตุผล ดังภาพที่ 3-2



ภาพที่ 3-2 ขั้นตอนการวิเคราะห์คุณภาพของข้อสอบ NT ระดับชั้นประถมศึกษาปีที่ 3

จากภาพที่ 3-2 แสดงขั้นตอนการวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้าน ได้แก่ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ดังนี้

1. ผู้วิจัยขอหนังสือขอความอนุเคราะห์ขอข้อมูลเพื่อการวิจัย จากวิทยาลัยวิทยาการวิจัยและวิทยาการปัญญา มหาวิทยาลัยบูรพา เพื่อขอผลการตอบข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ปีการศึกษา

2555 จำนวน 3 ด้าน ประกอบด้วย 1) ด้านภาษา 2) ด้านคำนวณ และ 3) ด้านเหตุผล จากสำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน (สพฐ.) กระทรวงศึกษาธิการ

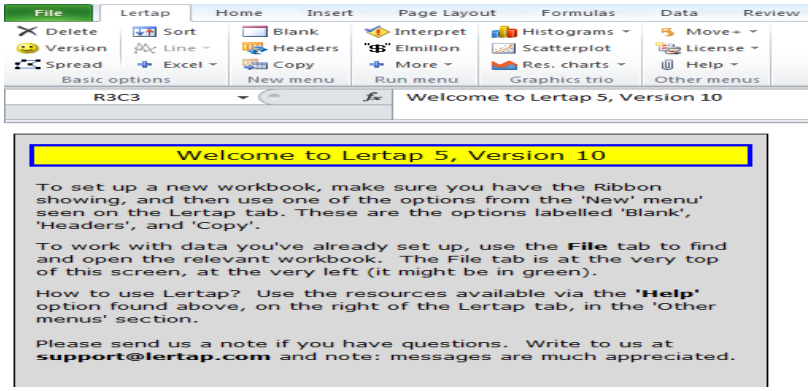
2. ตรวจสอบความสมบูรณ์ของผลการตอบข้อสอบ NT ปีการศึกษา 2555 ทั้งข้อคำถามตัวเลือก และเฉลยคำตอบที่ถูกต้อง รวมทั้งตรวจสอบความสมบูรณ์ ของคำตอบที่ผู้สอบทำการตอบ

3. วิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้าน ได้แก่ 1) ด้านภาษา 2) ด้านคำนวณ และ 3) ด้านเหตุผล ตามหลักการทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ โดยใช้โปรแกรม Xcalibre Version 4.2.2

ขั้นตอนการวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้วยโปรแกรม Xcalibre Version 4.2.2

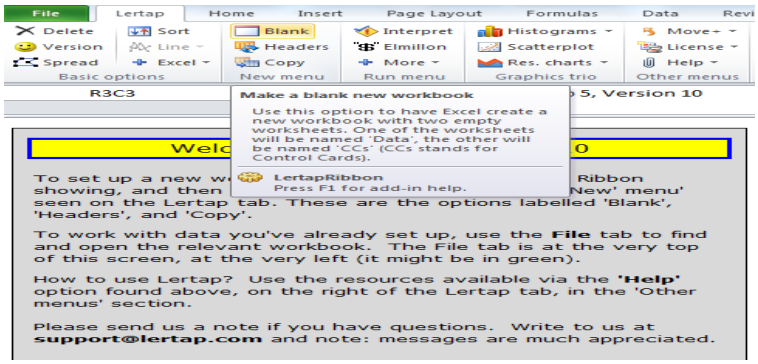
1. เริ่มจากการเตรียมไฟล์ข้อมูลผลการตอบข้อสอบ ด้วยโปรแกรม Lertap 5 ตามลำดับขั้นตอน ดังนี้

1.1 เปิดโปรแกรม Lertap เพื่อทำการวิเคราะห์คุณภาพของข้อสอบด้วยโปรแกรม Lertap ดังภาพที่ 3-3

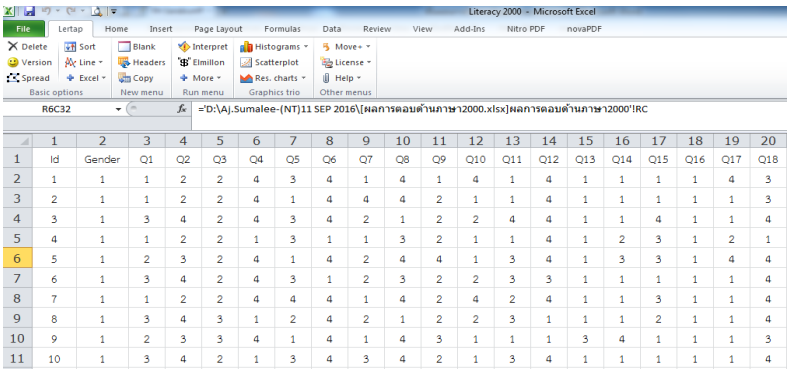


ภาพที่ 3-3 หน้าต่างโปรแกรม Lertap 5

1.2 เลือก Blank เปิดไฟล์ที่บันทึกไว้ Copy มาวางใน Excel ดังภาพที่ 3-4



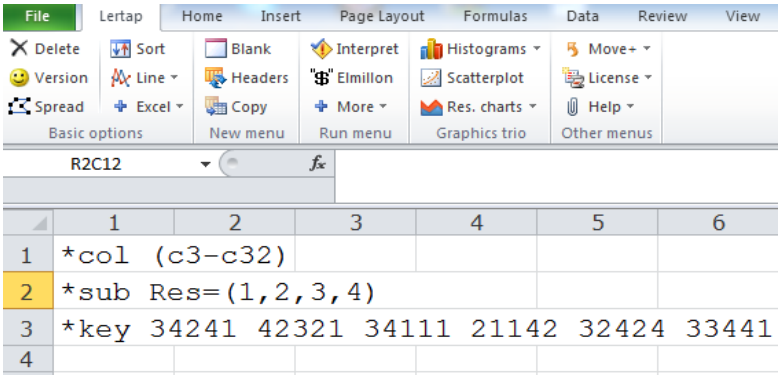
ภาพที่ 3-4 เลือกคำสั่ง Blank เพื่อ Copy ไฟล์ที่บันทึกมาวางใน Excel



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	Id	Gender	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18
2	1	1	1	2	2	4	3	4	1	4	1	4	1	4	1	1	1	1	4	3
3	2	1	1	2	2	4	1	4	4	4	2	1	1	4	1	1	1	1	1	3
4	3	1	3	4	2	4	3	4	2	1	2	2	4	4	1	1	4	1	1	4
5	4	1	1	2	2	1	3	1	1	3	2	1	1	4	1	2	3	1	2	1
6	5	1	2	3	2	4	1	4	2	4	4	1	3	4	1	3	3	1	4	4
7	6	1	3	4	2	4	3	1	2	3	2	2	3	3	1	1	1	1	1	4
8	7	1	1	2	2	4	4	4	1	4	2	4	2	4	1	1	3	1	1	4
9	8	1	3	4	3	1	2	4	2	1	2	2	3	1	1	1	2	1	1	4
10	9	1	2	3	3	4	1	4	1	4	3	1	1	1	3	4	1	1	1	3
11	10	1	3	4	2	1	3	4	3	4	2	1	3	4	1	1	1	1	1	4

ภาพที่ 3-5 ไฟล์ข้อมูลสำหรับการวิเคราะห์

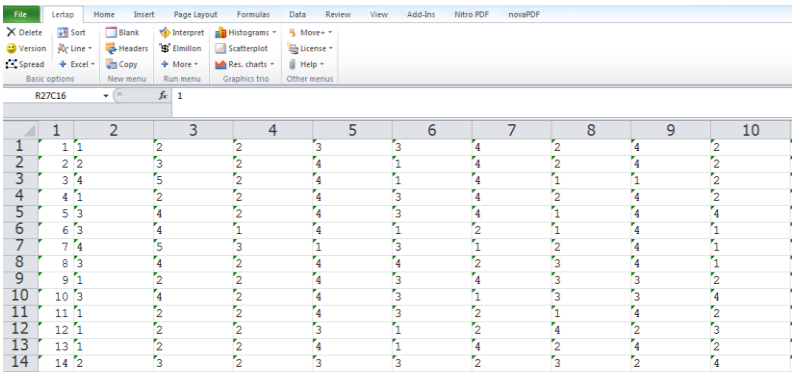
1.3 เลือก Blank เปิดไฟล์เฉลยคำตอบ Copy มาวางที่หน้า CCs ดังภาพที่ 3-6



	1	2	3	4	5	6
1	*col (c3-c32)					
2	*sub Res=(1,2,3,4)					
3	*key 34241 42321 34111 21142 32424 33441					
4						

ภาพที่ 3-6 ไฟล์ข้อมูลเฉลยคำตอบ

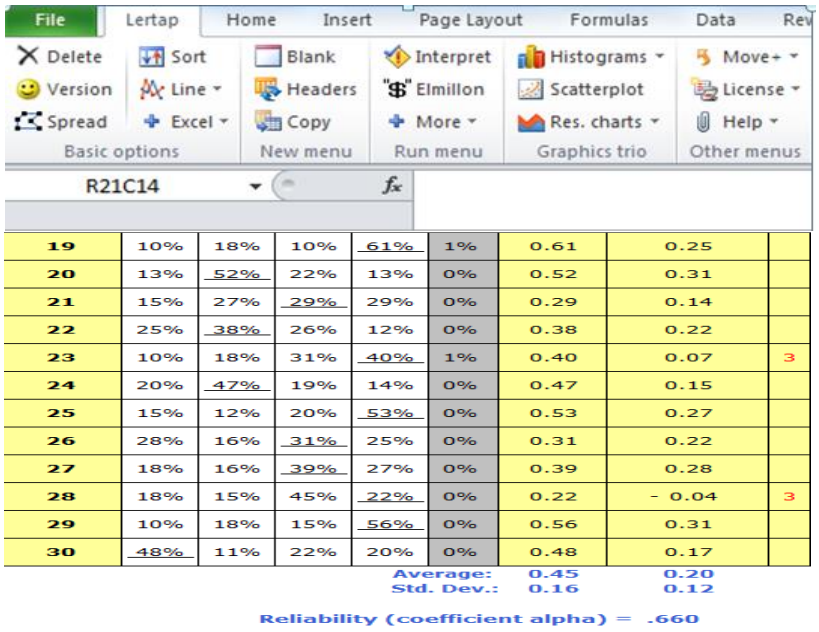
1.4 เลือก Interpert รอจนกว่าจะ Run เสร็จ



	1	2	3	4	5	6	7	8	9	10
1	1	1	2	2	3	3	4	2	4	2
2	2	2	3	2	4	1	4	2	4	2
3	3	4	5	2	4	1	4	1	1	2
4	4	1	2	2	4	3	4	2	4	2
5	5	3	4	2	4	3	4	1	4	4
6	6	3	4	1	4	1	2	1	4	1
7	7	4	5	3	1	3	1	2	4	1
8	8	3	4	2	4	4	2	3	4	1
9	9	1	2	2	4	3	4	3	3	2
10	10	3	4	2	4	3	1	3	3	4
11	11	1	2	2	4	3	2	1	4	2
12	12	1	2	2	3	1	2	4	2	3
13	13	1	2	2	4	1	4	2	4	2
14	14	2	3	2	3	3	2	3	2	4

ภาพที่ 3-7 หน้าต่างแสดงข้อมูล เมื่อใช้คำสั่ง Interpert

1.5 เลือก Eimilion รอจนกว่าจะ Run เสร็จ ตรวจเช็คค่า Reliability



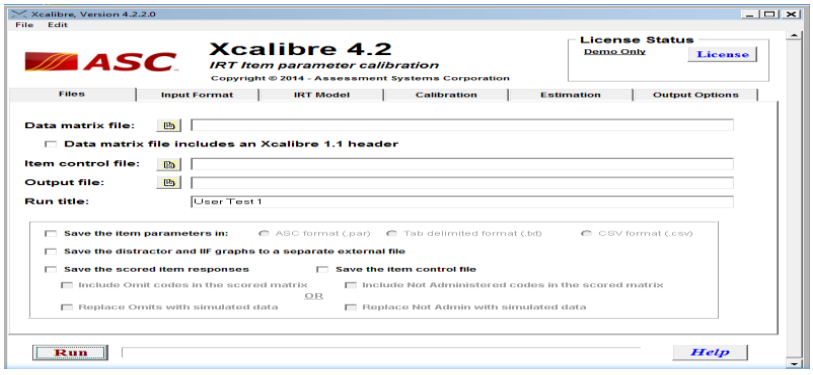
ภาพที่ 3-8 หน้าต่างแสดงข้อมูล เมื่อใช้คำสั่ง Eimilion

1.6 เลือก More เลือก Item Scores and Correlation รอจนกว่าโปรแกรมจะ Run เสร็จ

1.7 ตั้งโฟลเดอร์ใหม่ Save ไฟล์ที่ได้จากการ Run ด้วยโปรแกรม Lertap เตรียมไปทำการวิเคราะห์ด้วยโปรแกรม Xcalibre version 4.2.2 ต่อไป

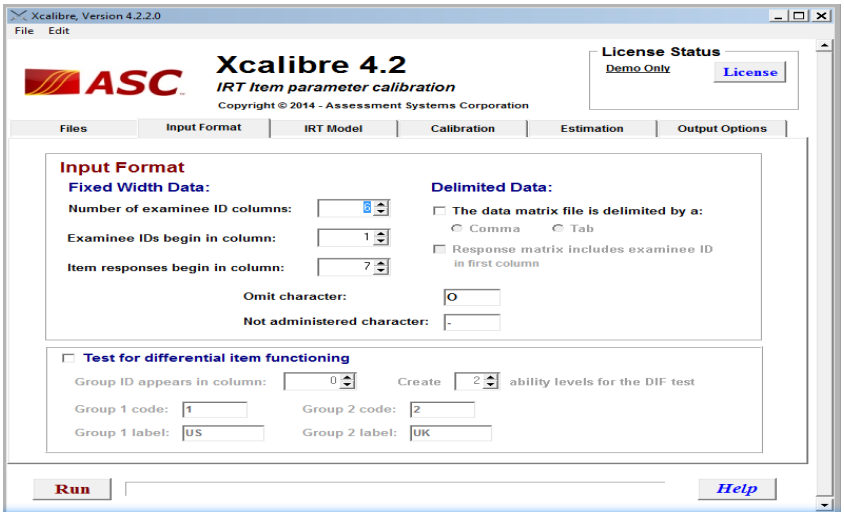
2. ขั้นตอนการวิเคราะห์คุณภาพของข้อสอบด้วยโปรแกรม Xcalibre version 4.2.2

2.1 เปิดโปรแกรม Xcalibre version 4.2.2 ดังภาพที่ 3-9



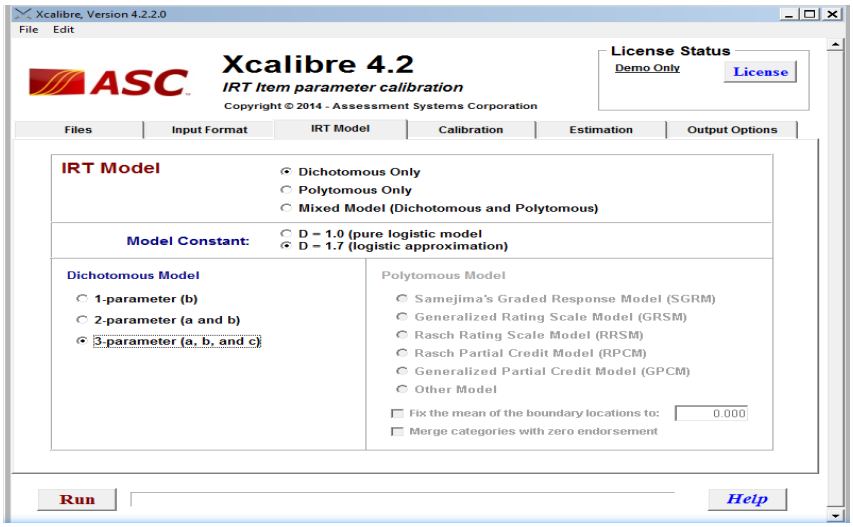
ภาพที่ 3-9 หน้าต่างโปรแกรม Xcalibre version 4.2.2

2.2 หน้าต่างสำหรับป้อนข้อมูลเพื่อระบุคอลัมน์ของเมทริกซ์ ดังภาพที่ 3-10



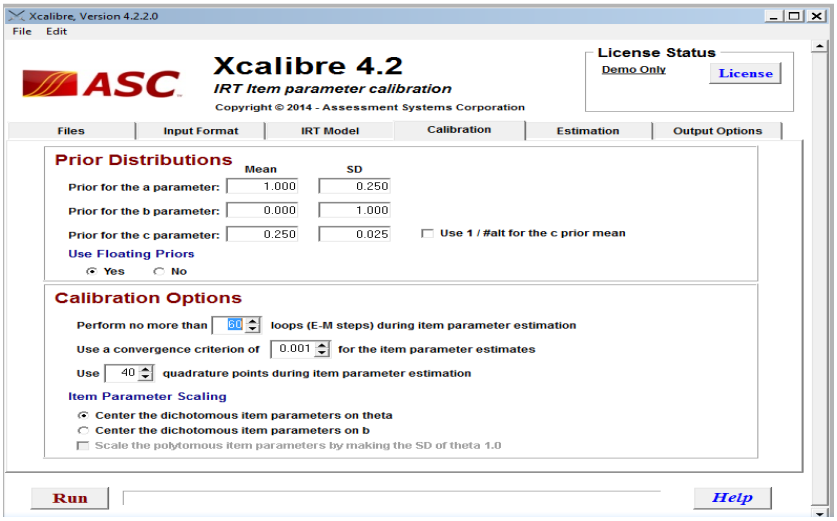
ภาพที่ 3-10 หน้าต่างสำหรับป้อนข้อมูล เพื่อระบุคอลัมน์ของเมทริกซ์

2.3 หน้าต่างโมเดลทฤษฎีการตอบสนองข้อสอบ ใช้สำหรับการประมาณค่าพารามิเตอร์ของข้อสอบ ดังภาพที่ 3-11



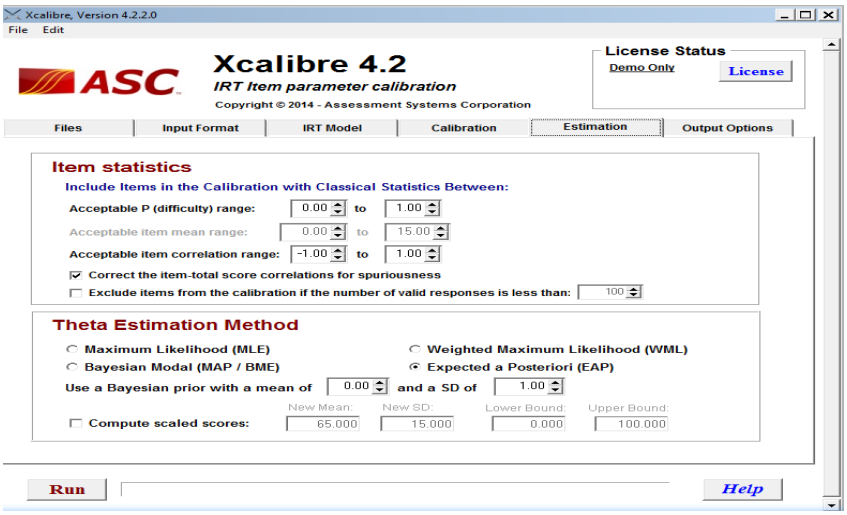
ภาพที่ 3-11 หน้าต่าง IRT Model

2.4 หน้าต่างการเปรียบเทียบที่สามารถระบุวิธีการก่อนและหลังกระบวนการประมาณค่าพารามิเตอร์ นอกจากนี้ยังสามารถระบุตัวเลือกที่จะใช้ในการเปรียบเทียบข้อสอบในทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ ดังภาพที่ 3-12



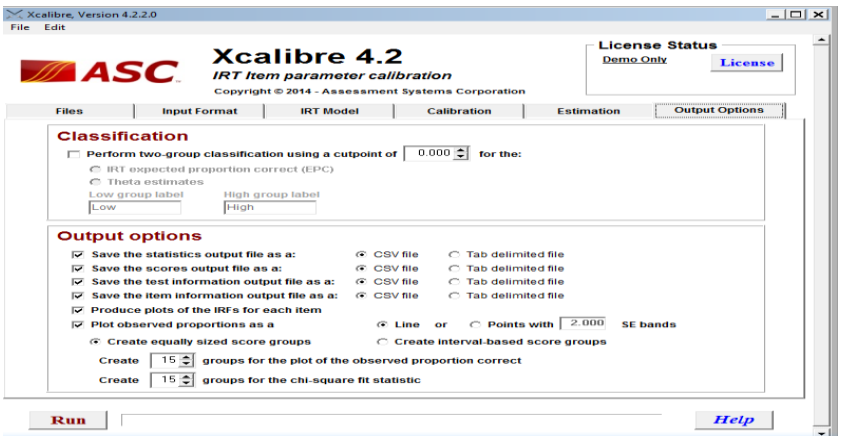
ภาพที่ 3-12 หน้าต่างการประมาณค่าพารามิเตอร์

2.5 หน้าต่างการประเมินความยากของข้อสอบที่ได้รับการยอมรับและมีความสัมพันธ์ในช่วงของการสอบเทียบ นอกจากนี้สามารถระบุวิธีการประมาณค่า θ คือการถูกนำมาใช้สำหรับการให้คะแนน IRT ดังภาพที่ 3-13



ภาพที่ 3-13 หน้าต่างการประมาณความยากของข้อสอบ

2.6 หน้าต่างตัวเลือกระบุผลลัพธ์ ว่าจะได้รับการจัดหมวดหมู่เช่นเดียวกับตัวเลือกการส่งออกที่เพิ่มขึ้น ดังภาพที่ 3-14



ภาพที่ 3-14 หน้าต่างตัวเลือกระบุผลลัพธ์ของข้อมูลการวิเคราะห์ด้วยโปรแกรม Xcalibre Version 4.2.2

2.7 แสดงผลการวิเคราะห์คุณภาพของข้อสอบด้วยโปรแกรม Xcalibre Version 4.2.2 ดังภาพที่ 3-15

Table 9: Item Parameters for All Calibrated Items

Seq.	Item ID	P	R	a	b	c	Flag(s)
1	1	0.488	0.336	0.610	0.284	0.075	
2	2	0.488	0.336	0.604	0.269	0.075	
3	2	0.735	0.237	0.473	-0.899	0.240	
4	4	0.624	0.225	0.431	0.046	0.252	
5	3	0.353	0.153	0.509	2.476	0.248	
6	4	0.681	0.239	0.442	-0.412	0.258	
7	1	0.199	0.073	0.890	3.050	0.184	K, Hb
8	4	0.172	0.047	1.086	3.246	0.173	Hb
9	1	0.480	0.274	0.633	0.949	0.240	
10	4	0.507	0.148	0.400	1.239	0.267	
11	1	0.359	0.138	0.531	2.406	0.254	
12	4	0.626	0.266	0.524	0.096	0.267	
13	1	0.572	0.279	0.599	0.366	0.245	
14	1	0.429	0.321	0.961	1.076	0.237	
15	1	0.500	0.359	0.888	0.641	0.229	
16	1	0.221	-0.078	1.142	4.000	0.224	K, Hb
17	4	0.502	0.312	0.713	0.746	0.240	
18	3	0.176	-0.022	1.174	3.511	0.182	K, Hb
19	4	0.578	0.283	0.579	0.352	0.251	
20	2	0.505	0.309	0.698	0.727	0.239	

ภาพที่ 3-15 ผลการวิเคราะห์คุณภาพของข้อสอบด้วยโปรแกรม Xcalibre Version 4.2.2

ระยะที่ 2 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ระดับชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้าน ได้แก่ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ดังภาพที่ 3-16

ภาพที่ 3-16 ขั้นตอนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้าน ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN

จากภาพที่ 3-16 แสดงขั้นตอนการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถม ศึกษาปีที่ 3 ทั้ง 3 ด้าน ด้วยวิธี HGLMวิธี MIMIC และวิธี BAYESIAN ดังนี้

1. การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ NT ด้วยวิธี HGLM

เป็นการวิเคราะห์โดยโมเดลเชิงเส้นตรงทั่วไประดับลดหลั่น ด้วยการประยุกต์ใช้โปรแกรม HLM โดยการจัดเตรียมข้อมูลสำหรับการวิเคราะห์ ซึ่งผู้วิจัยกำหนดให้ตัวแปรมีค่าดังต่อไปนี้

ความสามารถด้านภาษา ผู้วิจัยกำหนดให้ ตัวแปรเพศชายมีค่าเท่ากับ 0 และตัวแปรเพศหญิงมีค่าเท่ากับ 1

ความสามารถด้านคำนวณ ผู้วิจัยกำหนดให้ ตัวแปรเพศชายมีค่าเท่ากับ 1 และตัวแปรเพศหญิงมีค่าเท่ากับ 0

ความสามารถด้านเหตุผล ผู้วิจัยกำหนดให้ ตัวแปรเพศชายมีค่าเท่ากับ 0 และตัวแปรเพศหญิงมีค่าเท่ากับ 1

การวิเคราะห์แบ่งออกเป็น 2 ระดับ ด้วยโปรแกรมคอมพิวเตอร์สำเร็จรูป แล้วดำเนินการตามขั้นตอนต่อไปนี้

ขั้นตอนที่ 1 เตรียมไฟล์สำหรับวิเคราะห์ข้อมูล

ระดับที่ 1: ระดับข้อสอบ

ประกอบด้วยลำดับของผู้สอบ (ID) ลำดับของแบบทดสอบ (ITEM) ผลการตอบแบบทดสอบของผู้สอบ (Response) และตัวแปรดัมมี่ของแบบทดสอบ (ITEM....., n) ดังภาพที่ 3-17

	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	1	0	0	1	0	0
4	1	4	0	0	0	0	1	0
5	1	5	0	0	0	0	0	1
6	1	6	0	0	0	0	0	0
7	1	7	0	0	0	0	0	0
8	1	8	0	0	0	0	0	0
9	1	9	0	0	0	0	0	0
10	1	10	0	0	0	0	0	0

ภาพที่ 3-17 ตัวอย่างการจัดไฟล์ข้อมูลสำหรับการวิเคราะห์ด้วยวิธี HGLM ระดับที่ 1: ระดับข้อสอบ

ระดับที่ 2: ระดับผู้สอบ ประกอบด้วยลำดับของผู้สอบ (ID) ตัวแปรเพศ (Gender) ดังภาพที่ 3-18

	ID	gender
1	1	0
2	2	0
3	3	0
4	4	0
5	5	0
6	6	0
7	7	0
8	8	0
9	9	0
10	10	0

ภาพที่ 3-18 ตัวอย่างการจัดไฟล์ข้อมูลสำหรับการวิเคราะห์ด้วยวิธี HGLM ระดับที่ 2: ระดับผู้สอบ

ขั้นตอนที่ 2 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT

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

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยการศึกษาอิทธิพลของตัวแปรระดับข้อสอบและระดับผู้สอบ ที่มีโอกาสในการตอบข้อสอบถูก ในการวิเคราะห์ด้วยโปรแกรม HLM โดยสามารถแบ่งการวิเคราะห์เป็น 2 ระดับ ดังนี้

โมเดล HGLM วิเคราะห์ระดับที่ 1: ระดับข้อสอบ

การวิเคราะห์ระดับข้อสอบ ใช้หลักการวิเคราะห์ที่ข้อสอบสอดแทรกอยู่ในตัวบุคคล ผลการวิเคราะห์ระดับนี้จะแสดงค่าความยากของข้อสอบ ซึ่งสามารถเขียนสมการวิเคราะห์ ดังนี้

สมการโมเดลการวิเคราะห์ระดับที่ 1 ระดับข้อสอบ

$$\eta_{ij} = \beta_{0j} + \beta_{1j}X_{1j} + \beta_{2j}X_{2j} + ... + \beta_{(k-1)jx(k-1)j} \qquad (19)$$

โมเดล HGLM วิเคราะห์ระดับที่ 2: ระดับผู้สอบ

การวิเคราะห์ระดับผู้สอบ ผลการวิเคราะห์ได้ค่าพารามิเตอร์ข้อสอบ และค่าความสามารถของผู้สอบในสมการระดับผู้สอบ โดยสามารถเขียนตัวแปรคุณลักษณะของผู้สอบเข้าสู่สมการ เพื่ออธิบายความผันแปรของโอกาสในการตอบแบบทดสอบได้ถูกต้องของผู้สอบ สามารถเขียนสมการได้ดังนี้

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Gender + \gamma_{0jm} \quad (20)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}Gender \quad (21)$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}Gender \quad (22)$$

.

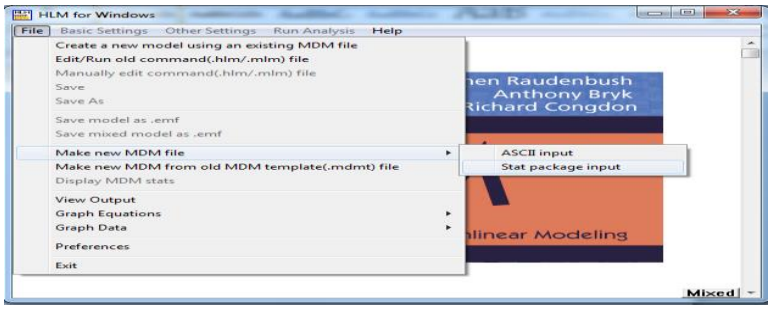
.

.

$$\beta_{(k-1)j} = \gamma_{(k-1)0} + \gamma_{(k-1)1}\ Gender \quad (23)$$

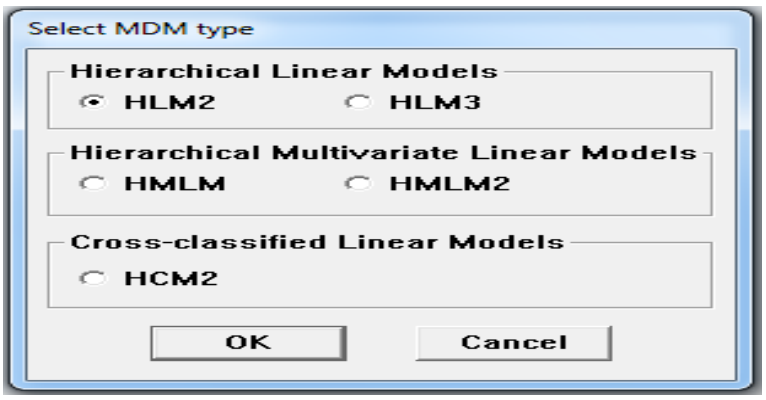
การวิเคราะห์ข้อมูลโดยใช้โปรแกรม HLM

1. เปิดโปรแกรม HLM เลือกเมนูไฟล์ Make new MDM file เลือก Select MDM type



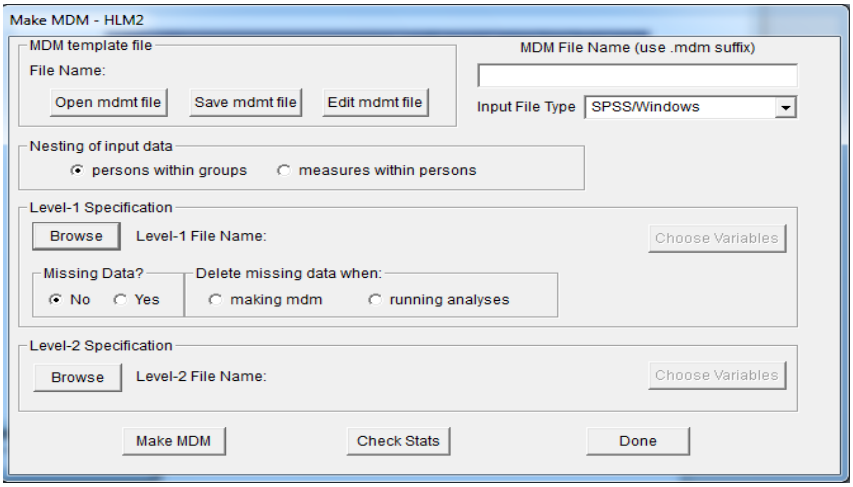
ภาพที่ 3-19 การสร้างแฟ้มข้อมูล MDM จากโปรแกรม SPSS

2. หน้าต่างแสดงชนิดของข้อมูล เลือก HLM 2 เลือก OK



ภาพที่ 3-20 หน้าต่างสำหรับเลือกประเภทของโมเดล

3. เลือกโมเดล HLM 2 จะได้หน้าต่าง Make MDM – HLM 2



ภาพที่ 3-21 หน้าต่างสำหรับเลือกประเภทของโมเดล

4. ผู้วิจัยต้องระบุรายละเอียดของข้อมูลให้โปรแกรมทราบ เพื่อสร้างแฟ้มข้อมูล MDM โดยดำเนินการ ดังนี้

4.1 ภายใต้หัวข้อ File Name มี 3 ทางเลือก คือ

4.1.1 Open Mdmt File ใช้ในกรณีที่ผู้วิจัยมีแฟ้มข้อมูล MDM ที่สร้างไว้แล้ว และจะนำมาสร้างโมเดลใหม่

4.1.2 Save Mdmt file ใช้ในกรณีที่ผู้วิจัยสร้างแฟ้มข้อมูล MDM ใหม่ ซึ่งกรณีนี้ผู้วิจัยต้องพิมพ์ชื่อแฟ้มข้อมูล MDM ที่จะสร้างลงในช่องว่างภายใต้ MDM File Name โดยต้องตามหลังด้วย Mdm

4.1.3 Edit Mdmt file ใช้ในกรณีที่ผู้วิจัยมีแฟ้มข้อมูล MDM ที่สร้างไว้แล้วและต้องการนำมาแก้ไขใหม่

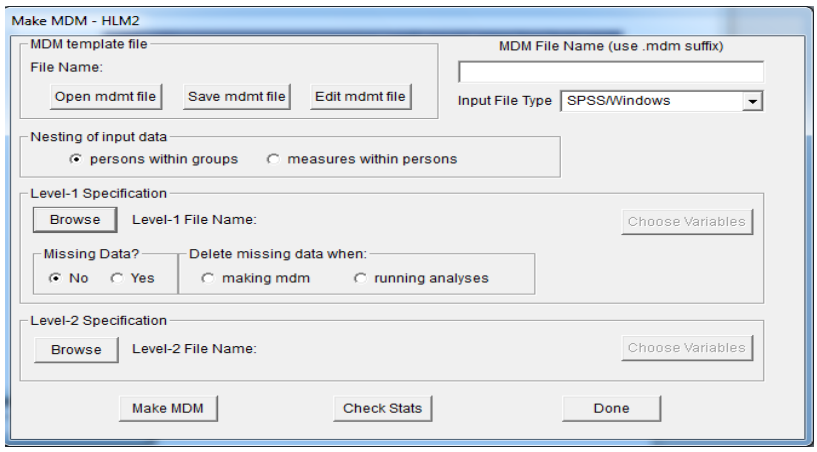
4.2 ภายใต้หัวข้อ Nesting of Input Data มี 2 ทางเลือก คือ

4.2.1 Persons Within Groups ใช้ในกรณีที่ข้อมูลระดับบุคคลรวมอยู่ภายใต้กลุ่ม เช่น ข้อมูลนักเรียนหลายๆคนที่อยู่ภายใต้โรงเรียนแต่ละโรงเรียน ไม่มีการวัดซ้ำ

4.2.2 Measures Within Persons ใช้ในกรณีที่ค่าที่วัดได้หลายๆค่าอยู่ภายใต้บุคคลแต่ละคน คือ มีการวัดซ้ำนั่นเอง

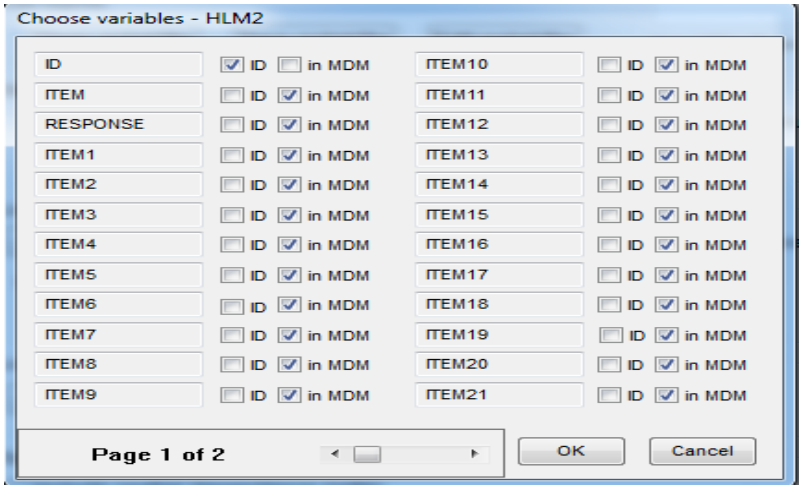
4.3 ภายใต้หัวข้อ Level – 1 Specfication ผู้วิจัยระบุรายละเอียดของข้อมูลระดับที่ 1 พร้อมทั้งระบุชื่อแฟ้มข้อมูล SPSS ระดับที่ 1

4.3.1 คลิกที่ Browse เพื่อระบุแหล่งของข้อมูลระดับที่ 1 พร้อมทั้งระบุชื่อแฟ้มข้อมูล SPSS ระดับที่ 1 แล้วคลิก Open จะกลับมาที่หน้าต่าง Make MDM – HLM 2 ภายใต้หัวข้อ Level 1 Specification จะปรากฏชื่อของแฟ้มข้อมูลระดับที่ 1 และที่ปุ่ม Choose Variable จะเป็นตัวหนา ดังภาพที่ 3-22



ภาพที่ 3-22 การระบุชื่อและแหล่งข้อมูลของแฟ้มข้อมูลระดับที่ 1

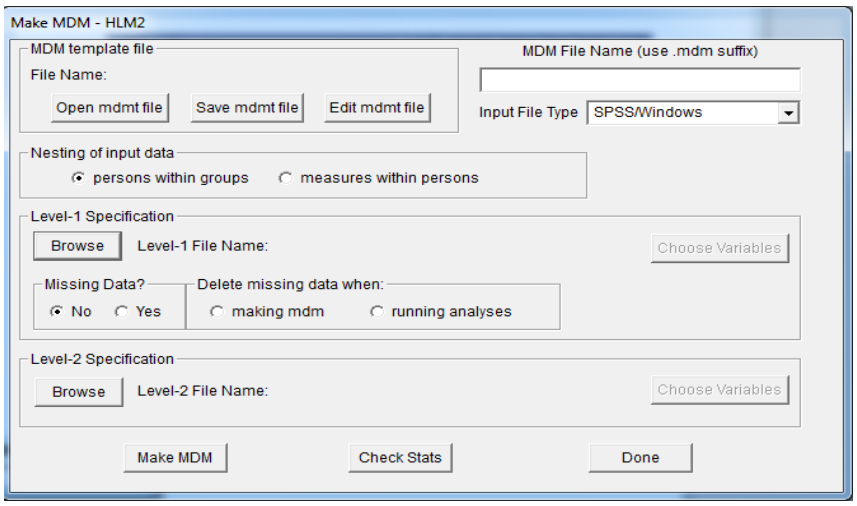
4.3.2 คลิกปุ่ม Choose Variables จะได้หน้าต่าง Choose variables HLM 2 ซึ่งมีรายชื่อของตัวแปรในระดับที่ 1 เลือกตัวแปรที่จะนำมาวิเคราะห์โดยคลิกเครื่องหมายถูกหลังชื่อตัวแปรที่ต้องการ ดังภาพที่ 3-23 แล้วคลิก OK จะกลับมาที่หน้าต่าง Make MDM – HLM 2



ภาพที่ 3-23 เลือกตัวแปร ระดับที่ 1

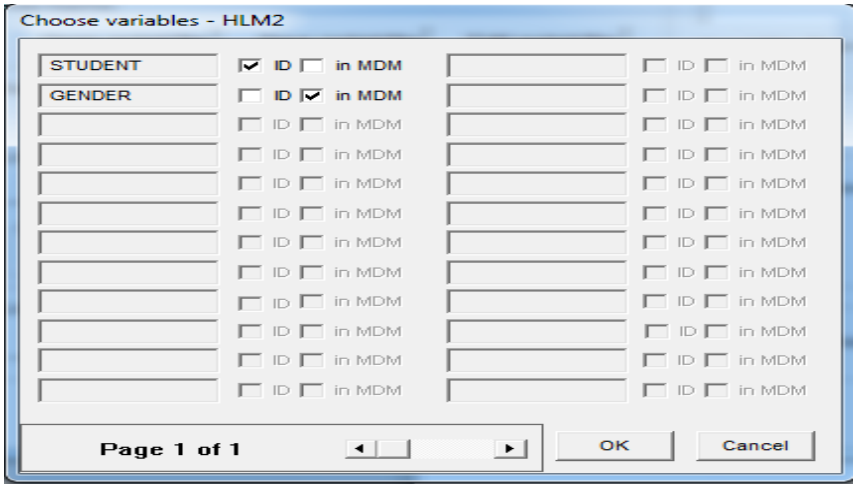
4.4 ภายใต้หัวข้อ Level 2 Specification ผู้วิจัยระบุรายละเอียดของข้อมูลระดับที่ 2 โดยดำเนินการ ดังนี้

4.4.1 คลิกที่ Browse เพื่อระบุแหล่งของข้อมูลระดับที่ 2 พร้อมทั้งระบุชื่อแฟ้มข้อมูล SPSS ระดับที่ 2 แล้วคลิก Open จะกลับมาที่หน้าต่าง Make MDM – HLM 2 ภายใต้หัวข้อ Level 2 Specification จะปรากฏชื่อของแฟ้มข้อมูลระดับที่ 2 และที่ปุ่ม Choose Variables จะเป็นตัวหนา ดังภาพที่ 3-24



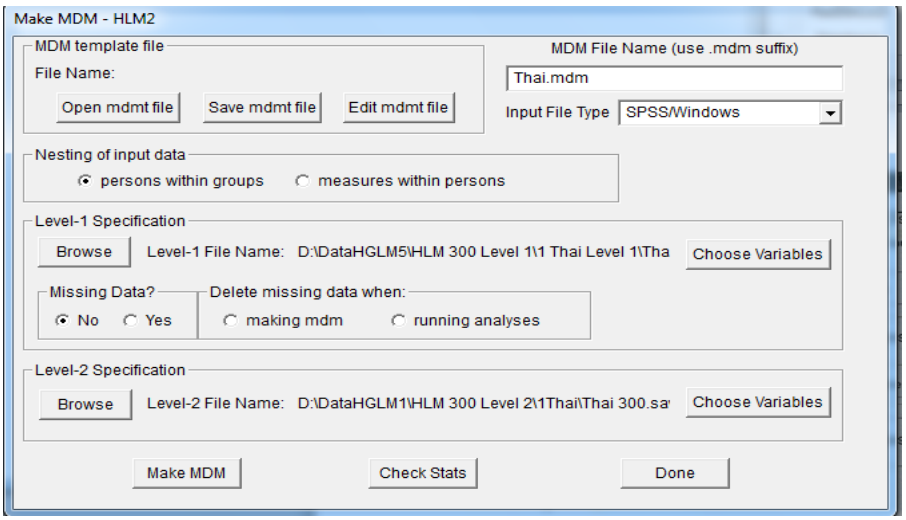
ภาพที่ 3-24 การระบุชื่อและแหล่งของแฟ้มข้อมูลระดับที่ 2

4.4.2 คลิกปุ่ม Choose Variables จะได้หน้าต่าง Choose Variables HLM 2 ซึ่งมีรายชื่อของตัวแปรในระดับที่ 2 เลือกตัวแปรที่จะนำมาวิเคราะห์ โดยคลิกเครื่องหมายถูกหลังชื่อตัวแปรที่ต้องการ ดังภาพที่ 3-25 แล้วคลิก OK จะกลับมาที่หน้าต่าง Make MDM – HLM 2



ภาพที่ 3–25 การเลือกตัวแปรระดับที่ 2

4.5 ในกล่องด้านบนขวามือภายใต้ MDM File Name ให้พิมพ์ชื่อแฟ้มข้อมูล MDM ที่ต้องการเก็บไว้ประมวลผล แล้วคลิกปุ่ม Save Mdmt File จะได้หน้าต่าง Save Template File ให้พิมพ์ชื่อแฟ้มข้อมูล MDM Template File ลงในช่องหลัง File name แล้วคลิก Save จะกลับมาที่หน้าต่าง Make MDM – HLM 2

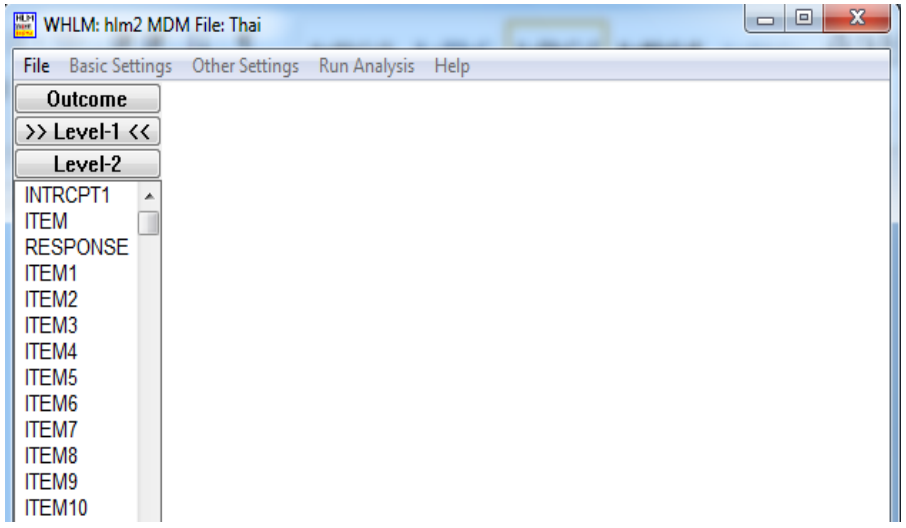


ภาพที่ 3–26 การระบุชื่อแฟ้มข้อมูล MDM ที่ต้องการเก็บไว้

4.6 คลิก Make MDM หน้าจอจะขึ้นค่าสถิติของข้อมูลประมาณ 3 วินาทีก็จะหายไป

4.7 คลิก Check Stats เพื่อตรวจสอบค่าสถิติพื้นฐาน เสร็จแล้วคลิกเครื่องหมาย X เพื่อปิดหน้าต่าง

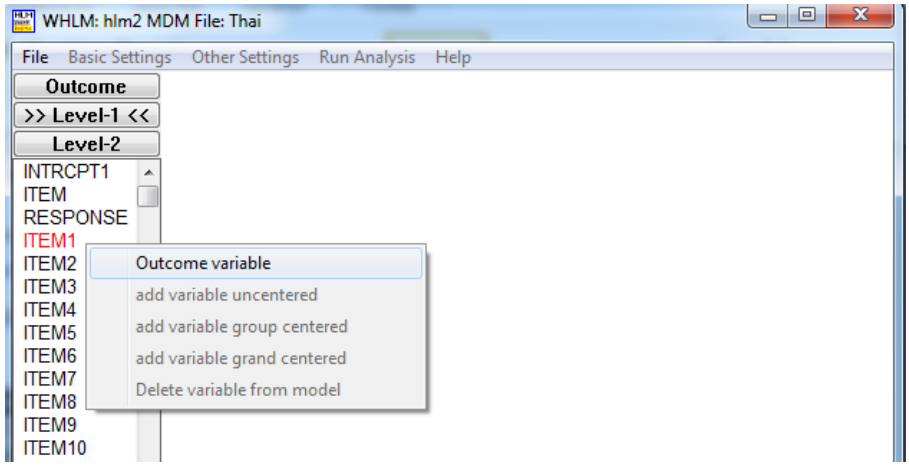
4.8 คลิก Done แสดงว่าจบสิ้นกระบวนการสร้างแฟ้มข้อมูล MDM และจะได้หน้าต่างก่อนการกำหนดลักษณะเฉพาะของโมเดล ดังภาพที่ 3-27



ภาพที่ 3-27 หน้าต่างก่อนกำหนดลักษณะเฉพาะของโมเดล

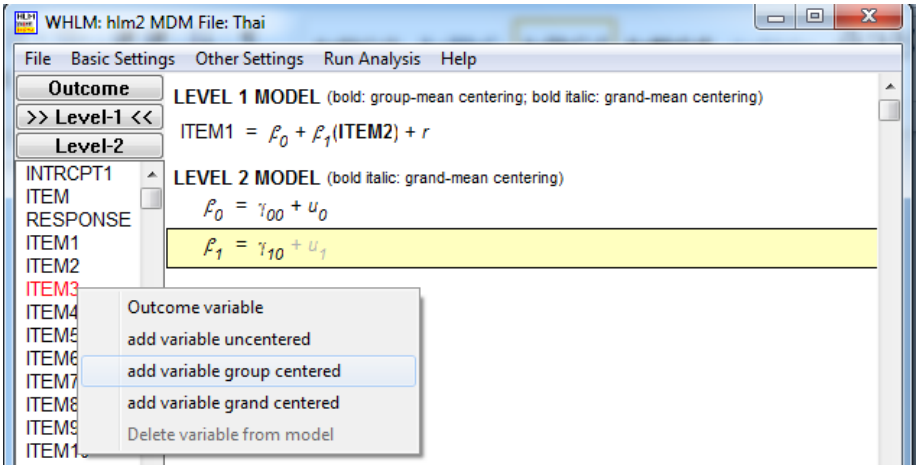
5. การกำหนดลักษณะเฉพาะของโมเดลระดับที่ 1 ซึ่งค่าสัมประสิทธิ์ในโมเดลระดับที่ 1 จะไปเป็นตัวแปรตามในโมเดลระดับที่ 2 การกำหนดลักษณะเฉพาะของโมเดลระดับที่ 1 ดำเนินการดังนี้

5.1 คลิกปุ่ม Level 1 เลือกชื่อตัวแปรที่จะให้เป็นตัวแปรตาม Outcome Variable ดังภาพที่ 3-28



ภาพที่ 3-28 การกำหนดตัวแปรตามในโมเดลระดับที่ 1

5.2 เลือกตัวแปรที่ต้องการให้เป็นตัวพยากรณ์ทีละตัวแปร โดยกำหนดวิธีการแปลงค่าของตัวพยากรณ์แต่ละตัว แล้วคลิก Add Variable Group Centered ดังภาพที่ 3-29 ซึ่งตัวพยากรณ์จะไปปรากฏเป็นตัวพยากรณ์ในสมการระดับที่ 1

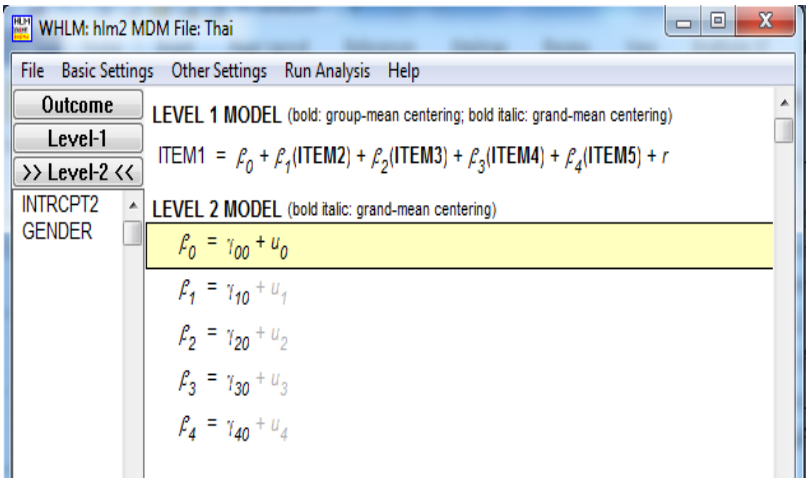


ภาพที่ 3-29 เลือกตัวแปรที่ใช้ในการวิเคราะห์ เป็นตัวพยากรณ์ระดับที่ 1

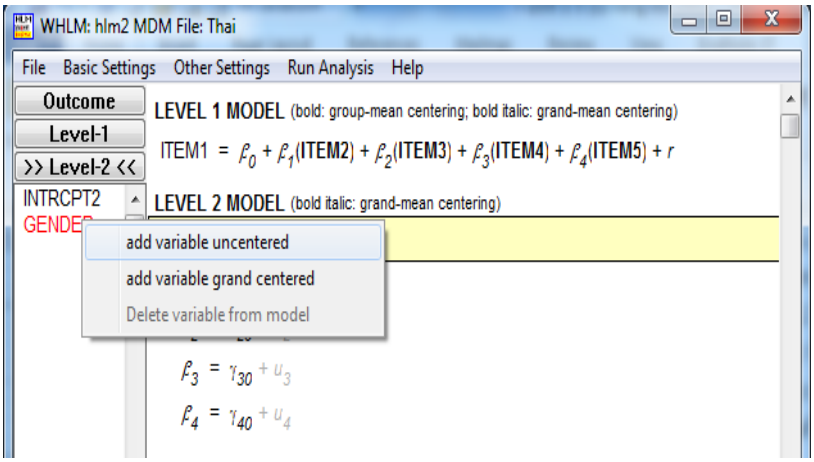
6. การกำหนดลักษณะเฉพาะของโมเดลระดับที่ 2 โดยมีค่าสัมประสิทธิ์ในโมเดลระดับที่ 1 มาเป็นตัวแปรตามในโมเดลระดับที่ 2 โดยดำเนินการดังนี้

6.1 คลิกปุ่ม Level 2 กำหนดค่าสัมประสิทธิ์ที่ได้จาโมเดลระดับที่ 1 มาเป็นตัวแปรตามในโมเดลระดับที่ 2 ว่าเป็นตัวแปรแบบสุ่มหรือเป็นค่าคงที่ ซึ่งทำได้โดยคลิกที่ตัว u_0 และ u_1 ถ้าต้องการให้

เป็นตัวแปรแบบสุ่มให้คลิกให้เป็นตัวเข้ม ถ้าต้องการให้เป็นค่าคงที่คลิกให้เป็นตัวจาง ดังภาพที่ 3-30 และภาพที่ 3-31

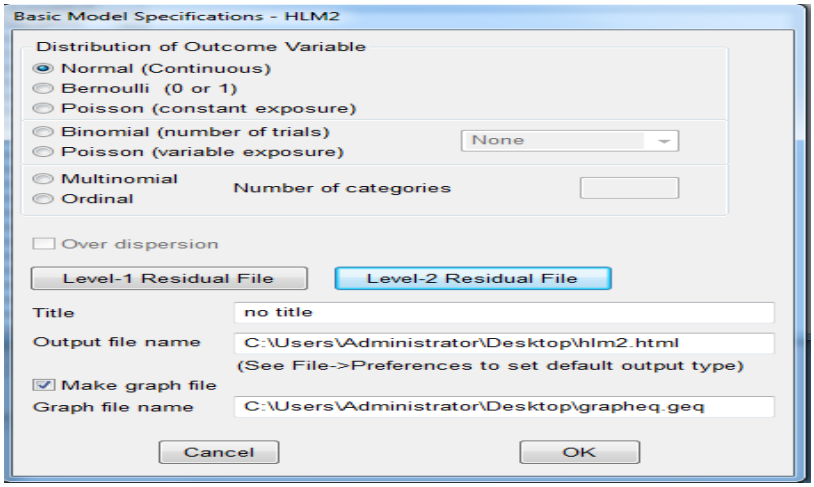


ภาพที่ 3-30 เลือกตัวแปรที่ใช้ในการวิเคราะห์ เป็นตัวพยากรณ์ระดับที่ 2



ภาพที่ 3-31 หน้าต่างหลังกำหนดลักษณะเฉพาะของโมเดล

6.2 กำหนดลักษณะของผลที่ได้ โดยใช้เมนู Basic Settings ในหน้าต่างหลังกำหนดลักษณะเฉพาะของโมเดล ให้คลิกที่เมนู Basic Settings จะได้หน้าต่าง Basic Model Specifications HLM 2 ดังภาพที่ 3-32



ภาพที่ 3-32 หน้าต่าง Basic Model Specifications HLM 2

6.3 ประมวลผลโดยคลิกเมนู Run Analysis ซึ่งผลที่ได้เก็บในรูป Text File ในโฟลเดอร์เดียวกับแฟ้มข้อมูล MDM การเรียกแฟ้มผลลัพธ์มาดูหลังจากเสร็จสิ้นการประมวลผล ผู้วิจัยทำได้โดยคลิกคำสั่ง View Output ที่อยู่ภายใต้เมนู File

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For INTRCPT1, B0					
INTRCPT2, G00	-0.311212	0.173470	-1.794	297	0.073
GENDER, G01	0.829153	0.247276	3.353	297	0.001
For ITEM1 slope, B1					
INTRCPT2, G10	-0.494859	0.242543	-2.040	5940	0.041
GENDER, G11	-0.541024	0.340260	-1.590	5940	0.112
For ITEM2 slope, B2					
INTRCPT2, G20	0.705502	0.235307	2.998	5940	0.003
GENDER, G21	-0.873242	0.333644	-2.617	5940	0.009
For ITEM3 slope, B3					
INTRCPT2, G30	-0.196033	0.236780	-0.828	5940	0.408
GENDER, G31	-0.295240	0.333473	-0.885	5940	0.376
For ITEM4 slope, B4					
INTRCPT2, G40	1.180960	0.244133	4.837	5940	0.000
GENDER, G41	-0.851612	0.346020	-2.461	5940	0.014
For ITEM5 slope, B5					
INTRCPT2, G50	0.378351	0.233162	1.623	5940	0.104
GENDER, G51	-0.869624	0.330914	-2.628	5940	0.009
For ITEM6 slope, B6					
INTRCPT2, G60	0.761626	0.235987	3.227	5940	0.002
GENDER, G61	-0.525801	0.338722	-1.552	5940	0.120
For ITEM7 slope, B7					
INTRCPT2, G70	-0.139203	0.235995	-0.590	5940	0.555
GENDER, G71	-0.378739	0.332910	-1.138	5940	0.256

ภาพที่ 3-33 ผลการวิเคราะห์ด้วยโปรแกรม HLM

หลังจากทำการวิเคราะห์ข้อมูลเรียบร้อยแล้ว จึงพิจารณาผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี HGLM โดยพิจารณาจากค่า *p*-value ของแต่ละข้อ ถ้าพบว่าข้อใดมีค่า *p*-value ที่นัยสำคัญทางสถิติที่ระดับ .05 แสดงว่า ข้อสอบข้อนั้นทำหน้าที่ต่างกัน (DIF)

การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ NT ด้วยวิธี MIMIC

การวิเคราะห์ด้วยโมเดลลิสเรล มีขั้นตอนการดำเนินการดังนี้

ขั้นตอนที่ 1 เตรียมไฟล์ข้อมูลสำหรับการวิเคราะห์

การวิเคราะห์ด้วยโมเดลลิสเรล ผู้วิจัยได้เตรียมข้อมูลในรูปแบบไฟล์ .dat เพื่อวิเคราะห์

การทำหน้าที่ต่างกันของข้อสอบ ด้วยโปรแกรม Mplus ประกอบด้วย ลำดับผู้สอบ (ID) เพศ (Gender) และผลการตอบแบบทดสอบของผู้สอบ (Response) ดังภาพที่ 3-34

	ID	SEX	Item1	Item2	Item3	Item4	Item5
1	1	0	1	0	1	0	0
2	2	0	0	0	0	0	0
3	3	0	1	0	1	0	0
4	4	0	0	0	1	1	0
5	5	0	0	0	0	0	0
6	6	0	1	0	0	0	0
7	7	0	0	0	0	0	0
8	8	0	1	0	1	0	0
9	9	0	0	0	1	0	0
10	10	0	0	0	0	0	0

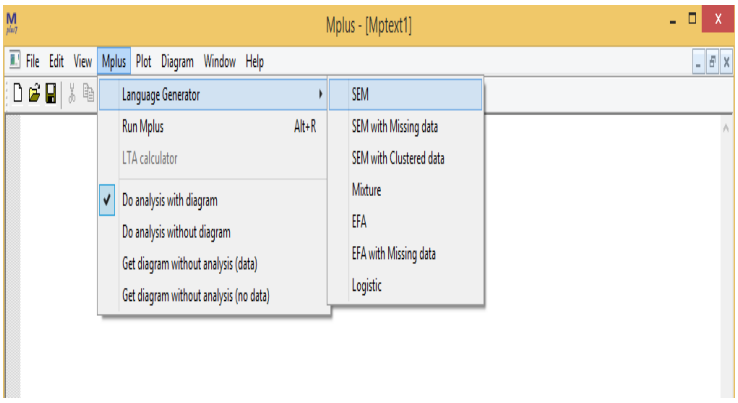
ภาพที่ 3-34 การจัดไฟล์ข้อมูลสำหรับวิเคราะห์ด้วยวิธี MIMIC ในรูปแบบไฟล์ .dat

ขั้นตอนที่ 2 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี MIMIC สามารถดำเนินการตามขั้นตอน ดังนี้

1. เปิดโปรแกรม Mplus แล้วสร้างโปรเจคการวิเคราะห์ขึ้นมา จากนั้นคลิก Mplus ที่แถบเมนูด้านบน

2. คลิกแถบเมนูย่อย Language Generator คลิกแถบเมนูย่อย SEM จะทำให้ได้หน้าต่าง Structural Equation Modeling

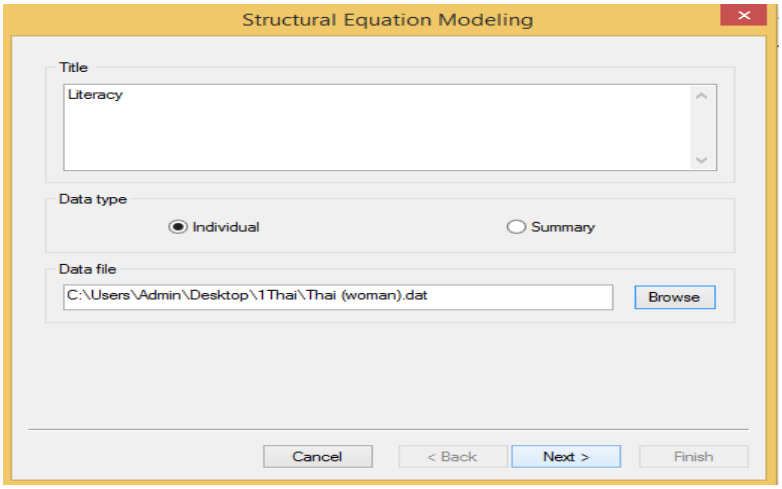
3. การระบุรายละเอียดของการวิเคราะห์ข้อมูล ดังภาพที่ 3-35



ภาพที่ 3-35 ตัวอย่างการจัดไฟล์ข้อมูลสำหรับวิเคราะห์ด้วยวิธี MIMIC ในรูปแบบไฟล์ .dat

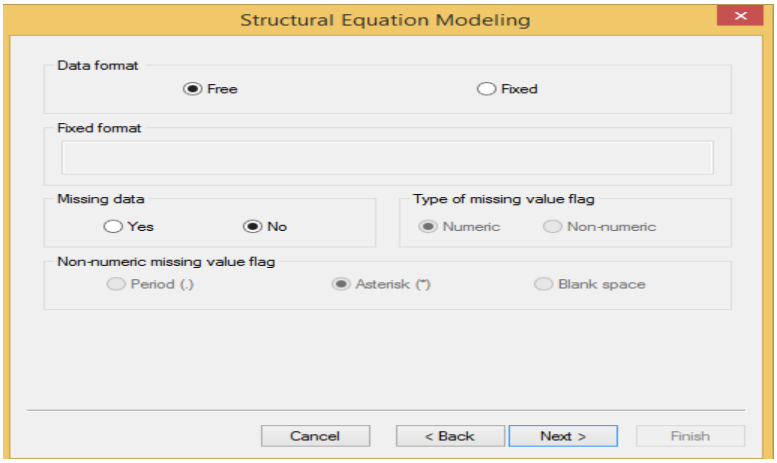
4. พิมพ์ชื่อการวิเคราะห์ข้อมูลลงใน Title

5. คลิก Browse เพื่อดึงไฟล์ข้อมูลที่เตรียมไว้ (นามสกุล.dat)



ภาพที่ 3-36 หน้าจอ Data File ของข้อมูลสำหรับวิเคราะห์ด้วยวิธี MIMIC ในรูปแบบไฟล์ .dat

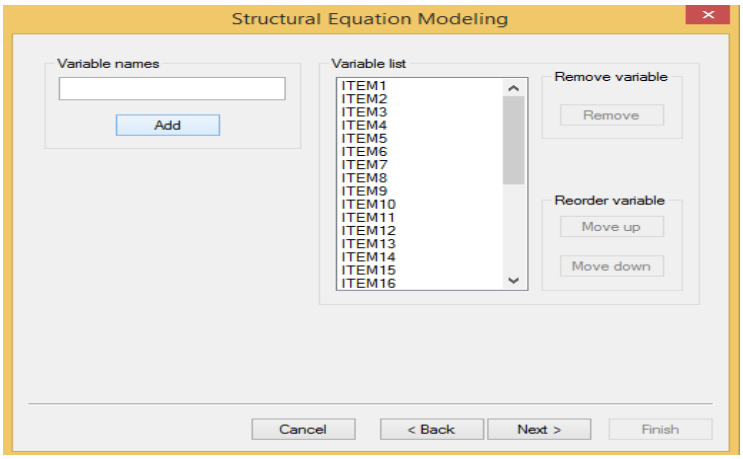
6. เลือกไฟล์ข้อมูลที่เตรียมไว้
7. คลิก Open และจะกลับไปที่หน้าต่าง Structural Equation Modeling
8. สังเกตว่าช่อง Data File จะปรากฏข้อความระบุตำแหน่งไฟล์ จากนั้นคลิก Next
9. เลือก Free ในช่อง Data Format
10. เลือก No (ในกรณีที่ไม่มีข้อมูลขาดหายในไฟล์ข้อมูล)



ภาพที่ 3-37 การระบุรายละเอียดของการวิเคราะห์ข้อมูล

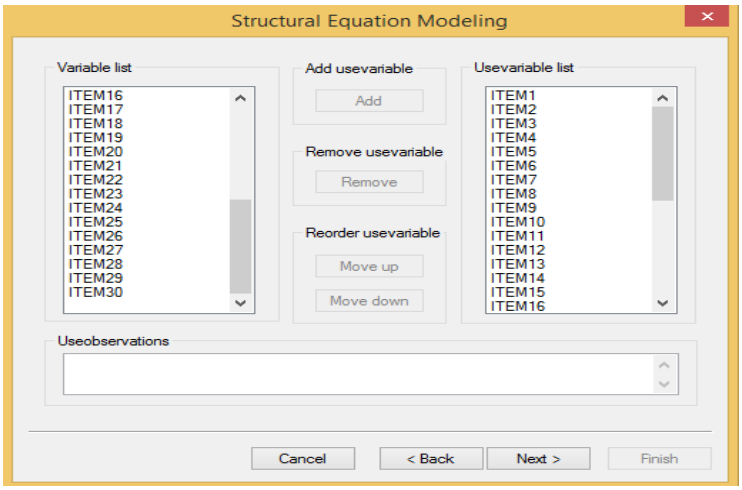
11. คลิก Next จะเกิดหน้าต่างใหม่ชื่อ Structural Equation Modeling ที่ใช้ระบุตัวแปร
12. สร้างตัวแปรโดยระบุชื่อตัวแปรที่มีในข้อมูล ในช่อง Variable Names เริ่มจากตัวแปรที่ปรากฏในไฟล์ข้อมูลตามลำดับทีละตัวหรือทีละชุด
13. คลิก Add เพื่อกำหนดชื่อตัวแปรจากไฟล์ข้อมูล ทันทีที่คลิกชื่อตัวแปรจะไปปรากฏอยู่ในช่อง Variable List ทำซ้ำจนกว่าตัวแปรในไฟล์ข้อมูลครบถ้วน

14. เมื่อตัวแปรที่สร้างมีการเรียงลำดับและจำนวนครบถ้วนตามไฟล์ข้อมูลที่เตรียมไว้ ให้คลิก Next



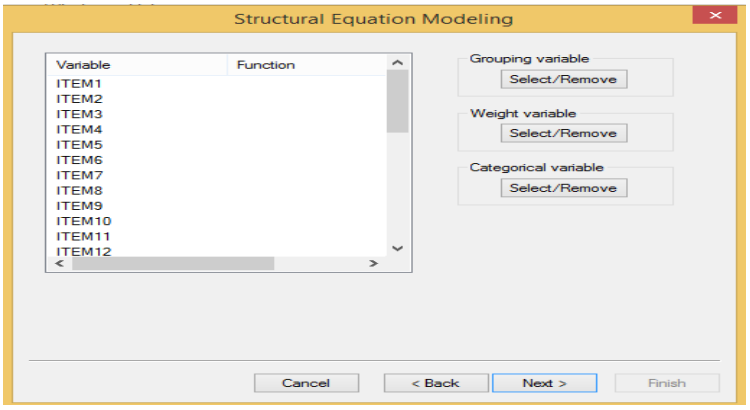
ภาพที่ 3-38 การระบุรายละเอียดของตัวแปรสำหรับการวิเคราะห์ข้อมูล

15. เลือกตัวแปรที่สร้างไว้ในช่อง Variable List คลิก Add เพื่อเลือกตัวแปรที่สร้างไว้เข้าสู่การวิเคราะห์ข้อมูล ตัวแปรจะเลื่อนไปอยู่ในช่อง Usevariable List ทางขวามือ เมื่อเลือกตัวแปรครบถ้วนตามจำนวนตัวแปรในโมเดลแล้ว คลิก Next



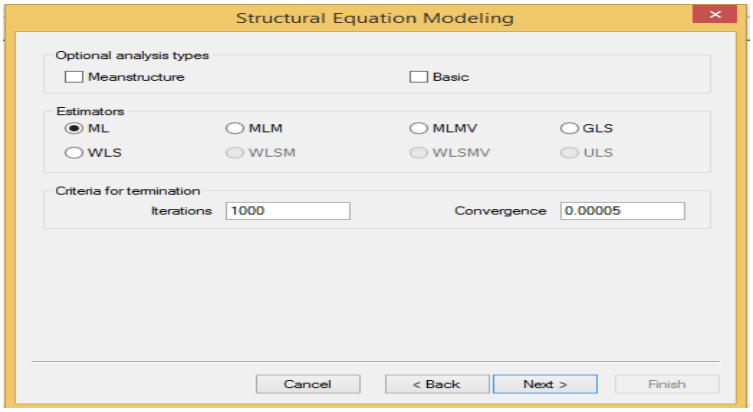
ภาพที่ 3-39 การนำตัวแปรที่สร้างไว้เข้าสู่การวิเคราะห์ข้อมูล

16. กำหนดคุณลักษณะของตัวแปร ในครั้งนี้เป็นการวิเคราะห์องค์ประกอบเชิงยืนยันของตัวแปรที่มีคุณสมบัติเชิงเมทริกซ์ ไม่มีการกำหนดน้ำหนักของตัวแปร จึงไม่ต้องดำเนินการกับคำสั่งด้านขวามือ จากนั้นคลิก Next



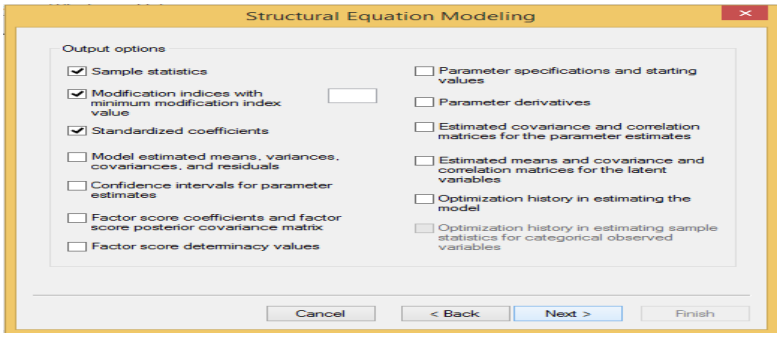
ภาพที่ 3-40 กำหนดคุณลักษณะของตัวแปร

17. ตรวจสอบว่าในกรอบ Optional Analysis Types ไม่มีการทำเครื่องหมายใดๆและกรอบ Estimator เลือกที่ ML จากนั้นคลิก Next



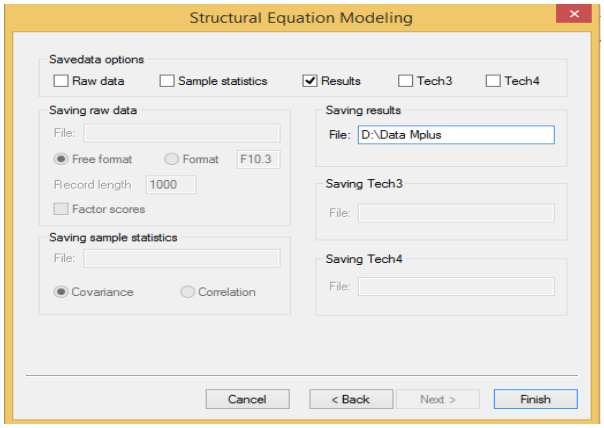
ภาพที่ 3-41 ตรวจสอบ Optional Analysis Types

18. คลิก Output Options ดังภาพที่ 3-42



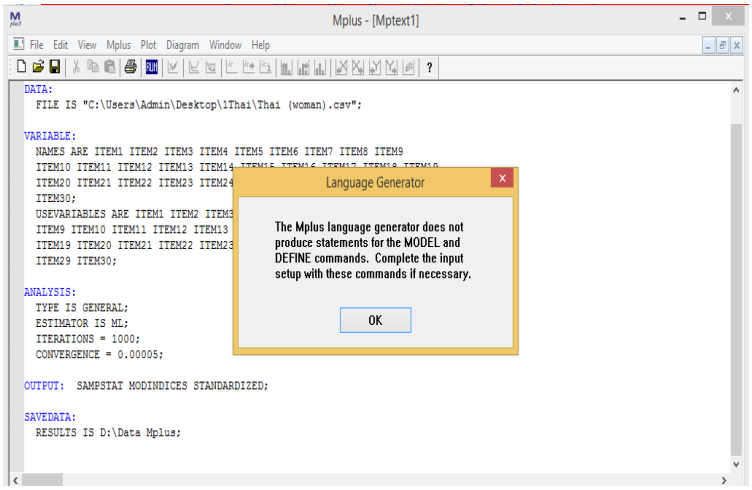
ภาพที่ 3-42 การเลือกให้แสดงผลลัพธ์ (Output)

19. คลิก เลือก Result ในกรอบ Savedata Options ระบุแหล่งที่ต้องการบันทึกผลการวิเคราะห์ข้อมูลในช่อง Saving Result โดยใช้การคัดลอก Address ของโฟลเดอร์ที่ต้องการบันทึกผลการวิเคราะห์โดยการ คลิกขวา แล้วเลือก Copy กลับมาที่โปรแกรม Mplus แล้วคลิกขวา ในช่อง File แล้วเลือก Paste จะพบว่ามีการระบุตำแหน่งการบันทึกผลการวิเคราะห์ข้อมูลปรากฏอยู่ จากนั้นคลิก Finish



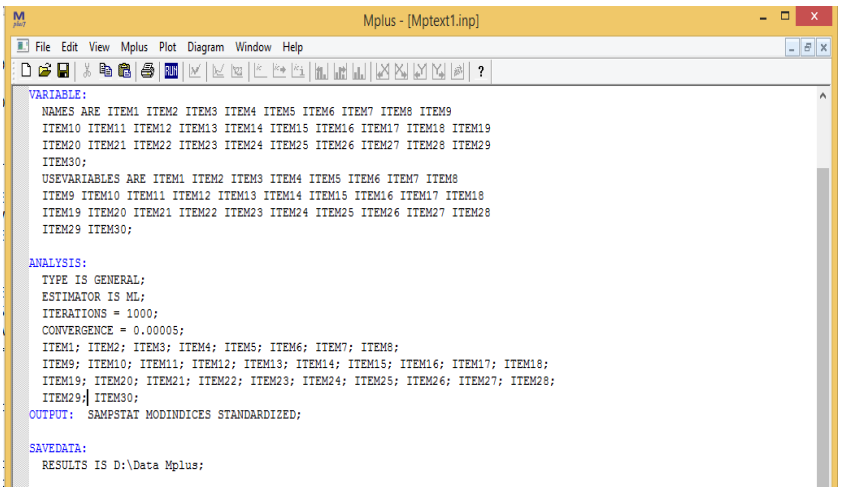
ภาพที่ 3-43 ระบุแหล่งที่ต้องการบันทึกผลการวิเคราะห์ข้อมูล

20. เมื่อดำเนินการสร้างคำสั่งการวิเคราะห์ข้อมูลด้วย Language Generator แล้วจะพบว่าโปรแกรม Mplus จะแจ้งให้ผู้ใช้โปรแกรมทราบว่าชุดคำสั่งที่สร้างขึ้นนั้นยังเป็นคำสั่งที่ไม่สมบูรณ์จำเป็นต้องดำเนินการเขียนคำสั่งในส่วนของ MODEL ด้วยตนเอง โดยให้พิมพ์คำสั่งแทรก เข้าไปหลังคำสั่ง Analysis



ภาพที่ 3-44 การสร้างคำสั่งการวิเคราะห์ข้อมูลด้วย Language Generator

21. พิมพ์คำสั่งตามลักษณะโมเดล



ภาพที่ 3-45 แสดงผลคำสั่งตามลักษณะโมเดล

22. หลังจากที่เขียนคำสั่งเพื่ออธิบายลักษณะโมเดลแล้วทำการวิเคราะห์ข้อมูล จะพบว่าโปรแกรม Mplus จะทำการวิเคราะห์ข้อมูลด้วย Dos ซึ่งแสดงให้เห็นเป็นหน้าต่างสีดำในระยะเวลาไม่นานนัก จากนั้น จะปรากฏหน้าต่างไฟล์ผลการิเคราะห์ซึ่งเป็นนามสกุล .out

MODEL RESULTS

FACTOR	ON				
SEX		-0.086	0.113	-0.761	0.447
ITEM1	ON				
SEX		0.014	0.065	0.216	0.829
ITEM2	ON				
SEX		-0.489	0.059	-8.229	0.000
ITEM3	ON				
SEX		-0.010	0.083	-0.126	0.900
ITEM4	ON				
SEX		-0.023	0.092	-0.256	0.798
ITEM5	ON				
SEX		-0.022	0.061	-0.371	0.711
ITEM6	ON				
SEX		-0.016	0.072	-0.225	0.822
ITEM7	ON				
SEX		0.058	0.061	0.954	0.340

ภาพที่ 3-46 แสดงผลลัพธ์ของผลการวิเคราะห์

3. การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ NT ด้วยวิธี BAYESIAN

การวิเคราะห์ข้อมูลในครั้งนี้ใช้ข้อมูลจริง (Real Data) ทำการประมวลผลภายใต้การเขียนคำสั่งการประมวลผลด้วยโปรแกรม WinBUGS ดังนี้

ศึกษาเอกสารและงานวิจัยต่าง ๆ ทั้งในประเทศและต่างประเทศที่เกี่ยวกับแนวคิดและวิธีการหลักการในการประมาณค่าพารามิเตอร์ความยากของข้อสอบ (b) พารามิเตอร์ความสามารถของผู้สอบ (θ) และการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ตามทฤษฎีการตอบสนองข้อสอบ (IRT)

ด้วยโปรแกรม WinBUGS เนื่องจากการวิจัยครั้งนี้ทั้ง 3 วิธี คือ วิธี HGLM วิธี MIMIC และวิธี BAYESIAN อยู่บนพื้นฐานของทฤษฎีการตอบสนองข้อสอบ

การศึกษาแนวคิดและหลักการ วิธีการใช้งาน จุดเด่น ข้อจำกัด และวิธีการเขียนคำสั่งตามสมการ รวมถึงวิธีการเขียนข้อมูลจริง (Real Data) ในการประมาณค่าพารามิเตอร์ความยากของข้อสอบ (b) พารามิเตอร์ความสามารถของผู้สอบ (θ) และตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ด้วยโปรแกรม WinBUGS มีการดำเนินการตามขั้นตอนดังนี้

ขั้นตอนที่ 1 เตรียมไฟล์ข้อมูลสำหรับการวิเคราะห์

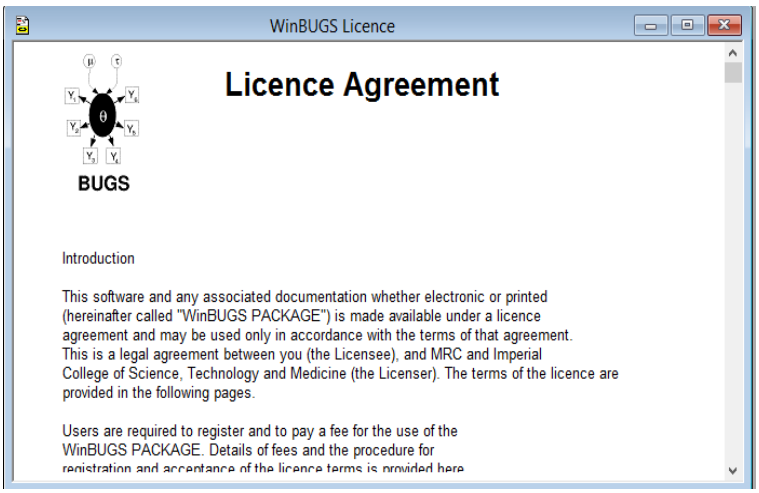
การวิเคราะห์ด้วยโมเดลลิสเรล ผู้วิจัยได้เตรียมข้อมูลในรูปแบบไฟล์ .dat เพื่อวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยโปรแกรม WinBUGS ประกอบด้วย ลำดับผู้สอบ (ID) เพศ (Gender) และผลการตอบแบบทดสอบของผู้สอบ (Response) ดังภาพที่ 3-47

	ID	SEX	Item1	Item2	Item3	Item4	Item5
1	1	0	1	0	1	0	0
2	2	0	0	0	0	0	0
3	3	0	1	0	1	0	0
4	4	0	0	0	1	1	0
5	5	0	0	0	0	0	0
6	6	0	1	0	0	0	0
7	7	0	0	0	0	0	0
8	8	0	1	0	1	0	0
9	9	0	0	0	1	0	0
10	10	0	0	0	0	0	0

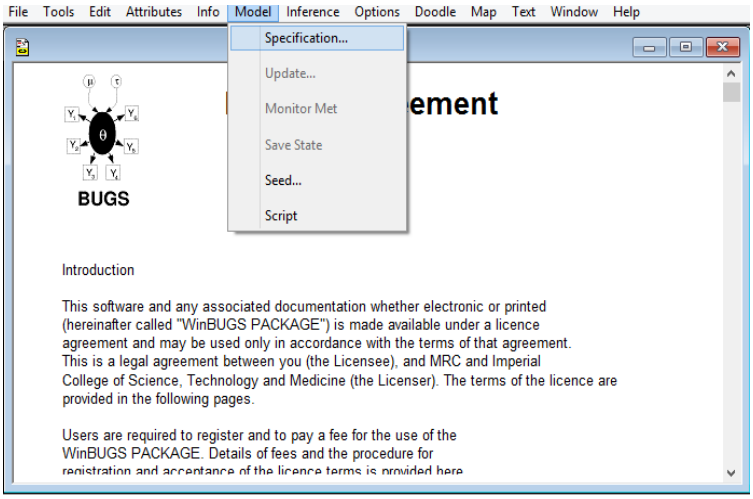
ภาพที่ 3-47 การจัดไฟล์ข้อมูลสำหรับวิเคราะห์ด้วยวิธี BAYESIAN ในรูปแบบไฟล์ .dat

ขั้นตอนที่ 2 ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ด้วยวิธี BAYESIAN ดังนี้

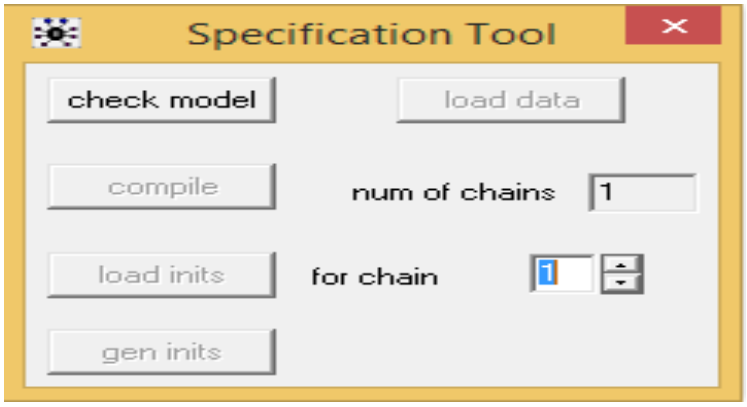
2.1 เปิดโปรแกรม WinBUGS ไปที่แถบเมนูเลือกคำสั่ง Model ตามด้วย Specification หลังจากนั้นจะได้หน้าจอแสดงหน้าต่าง Specification Tool และเลือกปุ่ม Check Model เพื่อตรวจสอบรูปแบบที่กำหนดว่าถูกต้องหรือไม่ ถ้าถูกต้องไม่มีข้อผิดพลาด ด้านล่างของหน้าจอจะเขียนคำว่า “Model is syntactically correct” ดังภาพที่ 3-48 ถึงภาพที่ 3-49



ภาพที่ 3-48 หน้าต่างโปรแกรม WinBUGS

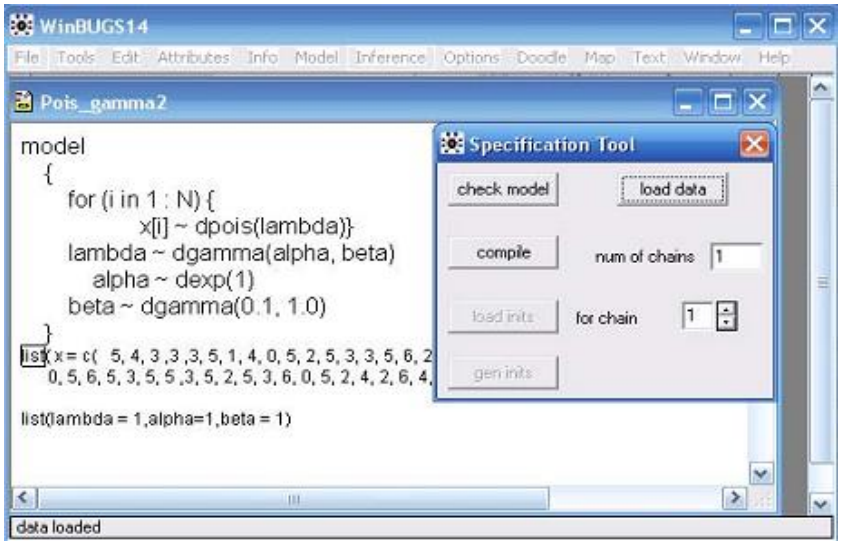


ภาพที่ 3-49 คำสั่ง Model ตามด้วย Specification

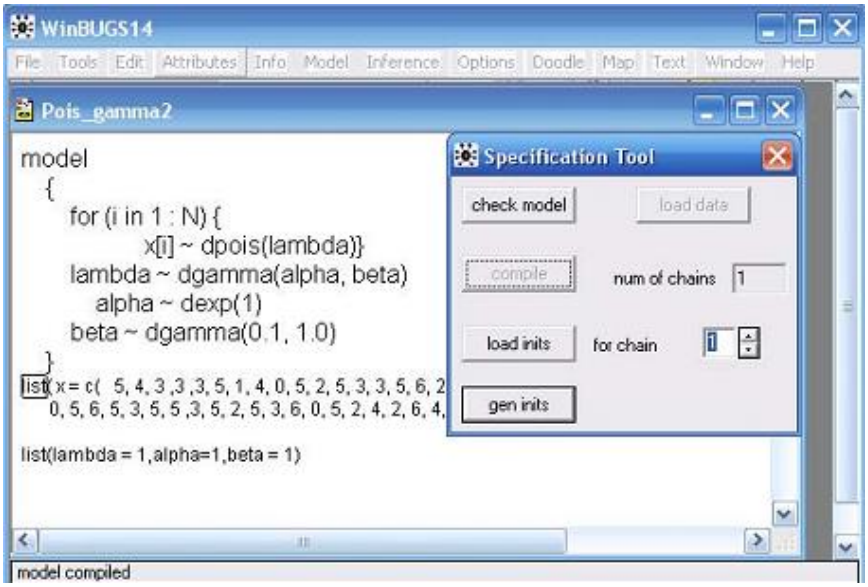


ภาพที่ 3-50 หน้าต่าง Specification Tool

2.2 เมื่อข้อมูลถูกต้องให้เรียกข้อมูลที่มีอยู่โดยเลือกเน้นคำว่า List ก่อนที่จะเลือกปุ่ม Load Data หลังจากที่โปรแกรมเรียกข้อมูลครบจะปรากฏคำว่า “Data Loaded” ด้านล่างของหน้าจอประมวลผล (Compile) เลือกปุ่ม Compile เพื่อยืนยันการประมวลผลของโปรแกรม จะปรากฏคำว่า “Model Compile” ดังภาพที่ 3-51 ถึงภาพที่ 3-52

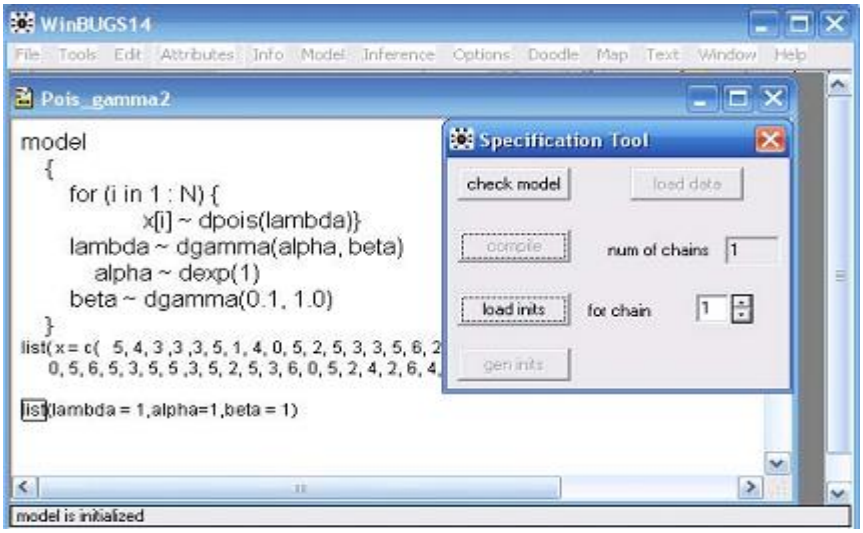


ภาพที่ 3-51 หน้าจอการเรียกข้อมูลเข้าโปรแกรม

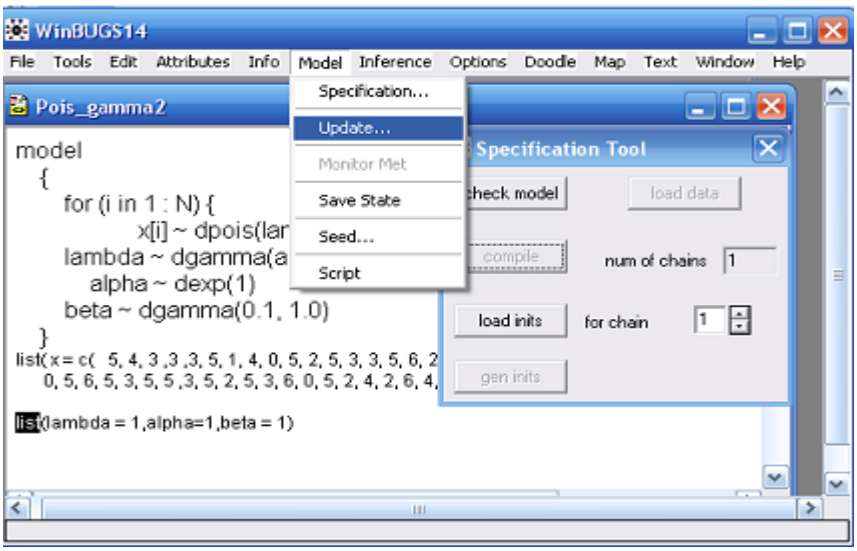


ภาพที่ 3-52 หน้าจอเพื่อประมวลผลรูปแบบและข้อมูลที่เรียกเข้า

2.3 เรียกค่าเริ่มต้น (Load Initial Values) ซึ่งเป็นค่าเริ่มต้นของค่าพารามิเตอร์ไว้หลังคำว่า list หลังจากนั้นเน้นคำว่า List เลือกปุ่ม Load Inits ใต้หน้าต่างจะขึ้นคำว่า “Model is Initialized” แต่ถ้าไม่มีการกำหนดค่าเริ่มต้นให้เลือกปุ่ม Gen Inits เพื่อให้โปรแกรมสร้างค่าเริ่มต้นให้เลือกแถบเมนู Model และเลือกที่ปุ่ม Update เพื่อสุ่มข้อมูลเริ่มต้น ดังภาพที่ 3-53 ถึงภาพที่ 3-54

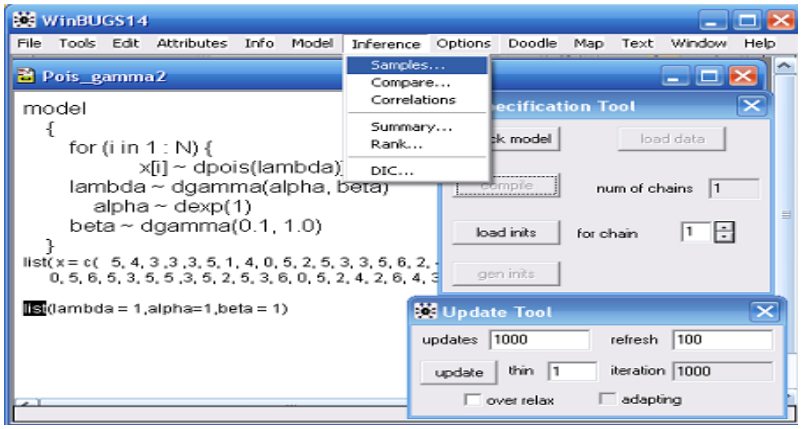


ภาพที่ 3-53 หน้าจอการเรียกค่าเริ่มต้น



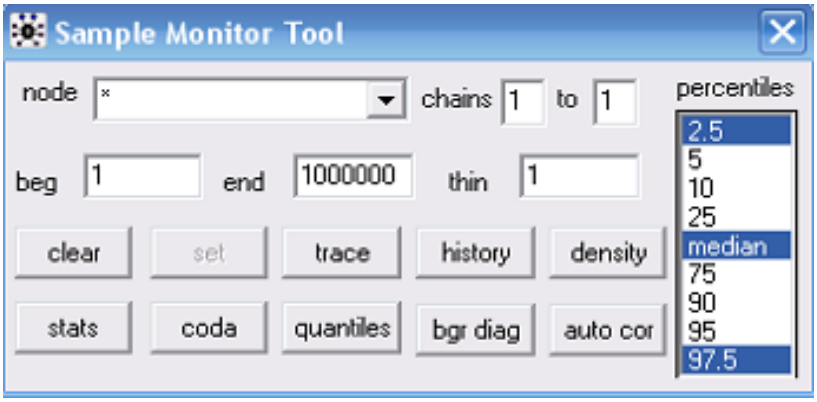
ภาพที่ 3-54 หน้าจอการสร้างค่าทดลอง

2.4 กำหนดค่าพารามิเตอร์ (Monitor Parameters) ไปที่แถบเมนูที่คำว่า Inference และเลือกคำสั่ง Samples จะได้หน้าต่างใหม่ เพื่อใส่ค่าพารามิเตอร์ที่ต้องการประมาณและเลือกปุ่ม Set เพื่อกำหนดค่าพารามิเตอร์ที่ต้องการ จากนั้นสร้างค่าการแจกแจงโดยหลักเกณฑ์ (Generate Posterior Values) โดยการเลือกปุ่ม Update ที่หน้าต่าง Update Tool ตามจำนวนซ้ำที่ต้องการ ดังภาพที่ 3-55

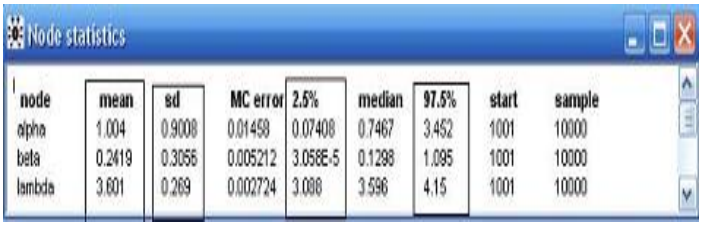


ภาพที่ 3-55 หน้าจอการเลือกค่าพารามิเตอร์

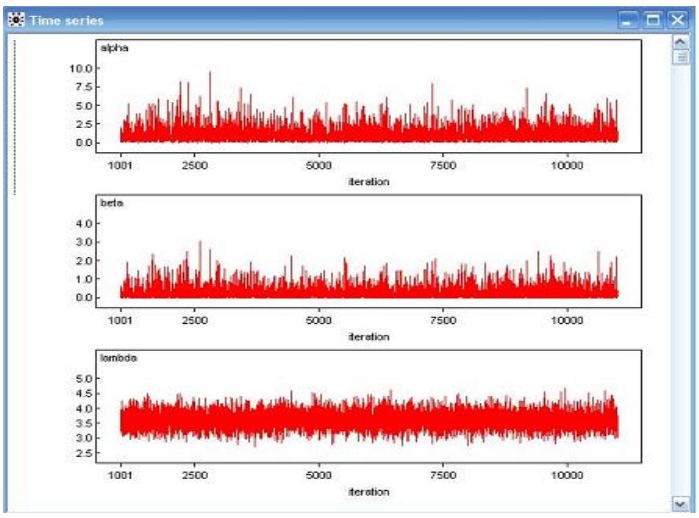
2.5 ถ้าต้องการค่าประมาณจากการแจกแจงโดยหลักเกณฑ์ทุกค่าที่กำหนดขึ้น ให้ใส่เครื่องหมาย * ในช่อง Node และเลือกปุ่ม Stats จะได้ค่าพารามิเตอร์ที่ต้องการ โดยพิจารณาจากค่าเฉลี่ย (Mean) ค่าส่วนเบี่ยงเบนมาตรฐาน (SD) ค่าความคลาดเคลื่อนจากการสุ่มแบบ MC (Markov Chain) และช่วงความเชื่อมั่น 95% (2.5% และ 97.5%) ค่าเริ่มต้นและค่าทั้งหมดจากการสุ่ม เลือกปุ่ม History จะแสดงกราฟการสุ่มตัวอย่างของค่าพารามิเตอร์ และปุ่ม Density จะแสดงลักษณะการแจกแจงของค่าพารามิเตอร์ ดังภาพที่ 3-56 ถึงภาพที่ 3-59



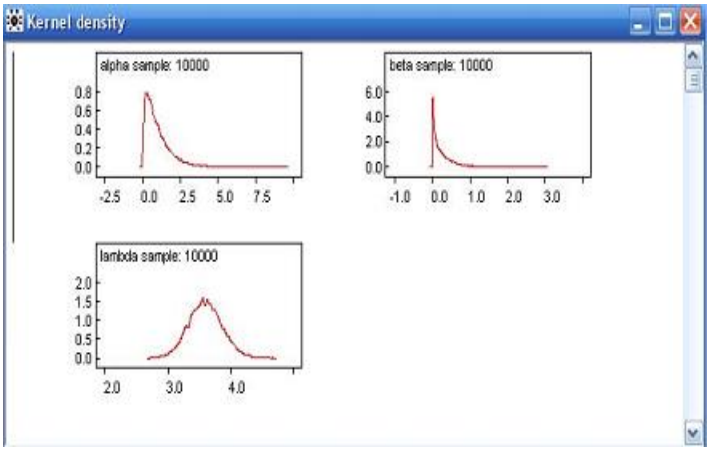
ภาพที่ 3-56 การเลือกค่าพารามิเตอร์ทั้งหมดที่ต้องการประมาณ



ภาพที่ 3-57 แสดงค่าพารามิเตอร์ที่ต้องการประมาณค่า



ภาพที่ 3-58 แสดงค่า alpha, beta, lambda



ภาพที่ 3-59 แสดงการแจกแจงของค่าพารามิเตอร์

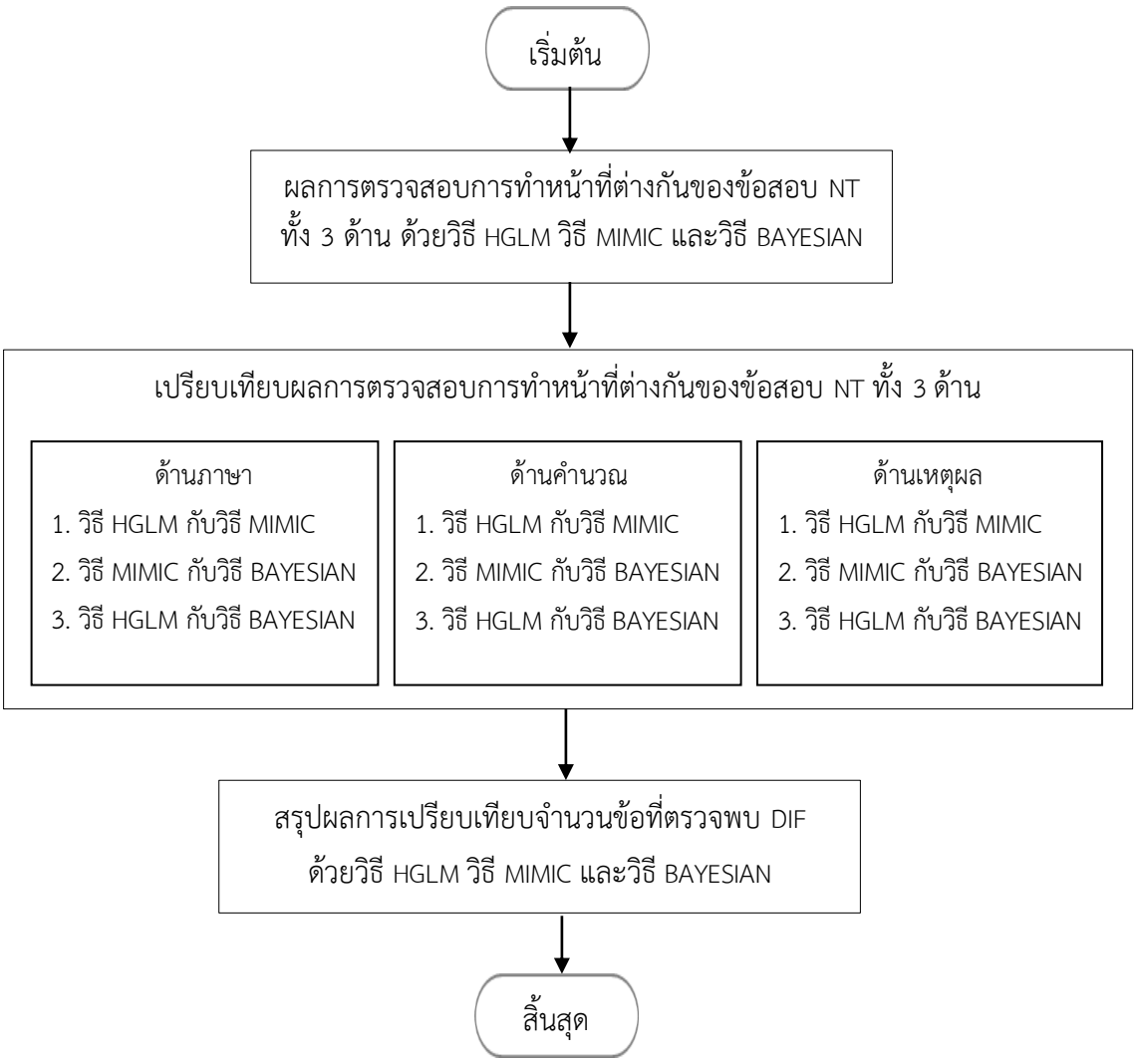
2.6 ผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ ด้วยโปรแกรม WinBUGS

mean	sd	MC_error		val2.5pc	median	val97.5pc		start	sample
b[1]	2.272	0.1658	0.01154	1.951	2.269	2.622	5001	5000	
b[2]	1.017	0.1688	0.0109	0.6856	1.023	1.348	5001	5000	
b[3]	5.012	0.2428	0.0186	4.467	5.024	5.478	5001	5000	
b[4]	6.172	0.2902	0.02393	5.675	6.165	6.749	5001	5000	
b[5]	0.8524	0.1549	0.01105	0.5492	0.8451	1.159	5001	5000	
b[6]	4.283	0.2369	0.02053	3.766	4.29	4.755	5001	5000	
b[7]	1.912	0.1674	0.01219	1.589	1.915	2.256	5001	5000	
b[8]	3.504	0.2093	0.01733	3.126	3.493	3.89	5001	5000	
b[9]	3.473	0.2088	0.01827	3.092	3.461	3.919	5001	5000	
b[10]	-0.5041	0.1556	0.00911	-0.805	-0.5009	-0.2087	5001	5000	
b[11]	3.938	0.2237	0.01783	3.568	3.932	4.419	5001	5000	
b[12]	6.006	0.3	0.02473	5.444	5.993	6.574	5001	5000	
b[13]	-1.854	0.1703	0.01208	-2.173	-1.862	-1.512	5001	5000	
b[14]	-0.06278	0.1651	0.01447	-0.357	-0.05154	0.2568	5001	5000	
b[15]	-0.3314	0.1557	0.008322		-0.6135	-0.3354	-0.01225	5001	5000
b[16]	-0.9961	0.1547	0.009665		-1.302	-0.9971	-0.6851	5001	5000
b[17]	-0.197	0.1516	0.009064		-0.4976	-0.1963	0.1042	5001	5000
b[18]	3.489	0.209	0.01797	3.089	3.474	3.94	5001	5000	
b[19]	5.394	0.2562	0.02006	4.884	5.413	5.866	5001	5000	
b[20]	4.482	0.2415	0.02119	4.054	4.465	5.022	5001	5000	
b[21]	3.898	0.2189	0.01796	3.497	3.89	4.337	5001	5000	
b[22]	2.672	0.2002	0.01611	2.275	2.669	3.088	5001	5000	
b[23]	5.341	0.2719	0.02288	4.782	5.338	5.88	5001	5000	
b[24]	3.306	0.2081	0.01757	2.914	3.295	3.753	5001	5000	
b[25]	4.071	0.2343	0.02076	3.601	4.073	4.483	5001	5000	
b[26]	2.694	0.1815	0.01524	2.352	2.692	3.053	5001	5000	
b[27]	3.725	0.2098	0.01914	3.341	3.717	4.226	5001	5000	
b[28]	3.811	0.1901	0.01444	3.462	3.81	4.22	5001	5000	
b[29]	4.961	0.2339	0.01854	4.534	4.969	5.444	5001	5000	
b[30]	-0.7157	0.1454	0.009216		-0.9927	-0.7187	-0.4199	5001	5000

ภาพที่ 3-60 แสดงผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ ด้วยโปรแกรม WinBUGS

ระยะที่ 3 การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้าน ได้แก่ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN

การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้าน ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN มีขั้นตอนการดำเนินการ ดังนี้



ภาพที่ 3-61 ขั้นตอนการเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT

ผู้วิจัยเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ตามสมมติฐานการวิจัย ดังนี้

1. ผลการเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา วิธี HGLM ตรวจพบข้อสอบทำหน้าที่ต่างกัน (DIF) มากกว่าวิธี MIMIC

2. ผลการเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา วิธี MIMIC ตรวจพบข้อสอบทำหน้าที่ต่างกัน (DIF) มากกว่าวิธี BAYESIAN

3. ผลการเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา วิธี HGLM ตรวจพบข้อสอบทำหน้าที่ต่างกัน (DIF) มากกว่าวิธี BAYESIAN

4. ผลการเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านคำนวณ วิธี HGLM ตรวจพบข้อสอบทำหน้าที่ต่างกัน (DIF) มากกว่าวิธี MIMIC

5. ผลการเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านคำนวณ วิธี MIMIC ตรวจพบข้อสอบทำหน้าที่ต่างกัน (DIF) มากกว่าวิธี BAYESIAN

6. ผลการเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านคำนวณวิธี HGLM ตรวจพบข้อสอบทำหน้าที่ต่างกัน (DIF) มากกว่าวิธี BAYESIAN

7. ผลการเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านเหตุผล วิธี HGLM ตรวจพบข้อสอบทำหน้าที่ต่างกัน (DIF) มากกว่าวิธี MIMIC

8. ผลการเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านเหตุผลวิธี MIMIC ตรวจพบข้อสอบทำหน้าที่ต่างกัน (DIF) มากกว่าวิธี BAYESIAN

9. ผลการเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านเหตุผลวิธี HGLM ตรวจพบข้อสอบทำหน้าที่ต่างกัน (DIF) มากกว่าวิธี BAYESIAN

บทที่ 4
ผลการวิจัย

การวิจัยนี้มีวัตถุประสงค์เพื่อวิเคราะห์คุณภาพของข้อสอบ NT ตามหลักทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT และเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้าน ได้แก่ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN ผู้วิจัยนำเสนอผลการวิจัย เป็น 3 ตอน ดังนี้

ตอนที่ 1 ผลการวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้าน ได้แก่ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ตามหลักการของทฤษฎีการตอบสนองข้อสอบ (IRT) แบบ 3 พารามิเตอร์

ตอนที่ 2 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้าน ได้แก่ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN

ตอนที่ 3 ผลการเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษา ปีที่ 3 ทั้ง 3 ด้าน ได้แก่ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN

ผู้วิจัยได้กำหนดสัญลักษณ์ที่ใช้ในการวิจัย ดังนี้

M หมายถึง ค่าเฉลี่ย

SD หมายถึง ส่วนเบี่ยงเบนมาตรฐาน

n หมายถึง จำนวนตัวอย่าง

a หมายถึง ค่าอำนาจจำแนกของข้อสอบ

b หมายถึง ค่าความยากของข้อสอบ

c หมายถึง ค่าโอกาสการเดาของข้อสอบ

ตอนที่ 1 การวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้าน ได้แก่ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ตามหลักการของทฤษฎีการตอบสนองข้อสอบ (IRT) แบบ 3 พารามิเตอร์

การวิเคราะห์คุณภาพของข้อสอบ NT เป็นการวิเคราะห์ค่าพารามิเตอร์ของข้อสอบตามหลักทฤษฎีการตอบสนองข้อสอบ (IRT) ประกอบด้วย ค่าอำนาจจำแนกของข้อสอบ (a) ค่าความยากของข้อสอบ (b) และค่าโอกาสการเดาของข้อสอบ (c) โดยใช้โปรแกรม Xcalibre Version 4.2.2 การประมาณค่าพารามิเตอร์ของข้อสอบ ซึ่งมีเกณฑ์การคัดเลือกข้อสอบ ดังนี้

เกณฑ์การคัดเลือกข้อสอบ (Urry, 1997)

1) ค่าอำนาจจำแนกข้อสอบ (a) มีค่าตั้งแต่ 0.50 ถึง 2.50

2) ค่าความยากของข้อสอบ (b) มีค่าตั้งแต่ -2.50 ถึง 2.50

3) ค่าโอกาสการเดาของข้อสอบ (c) มีค่าไม่เกิน 0.30

โดยเกณฑ์การประเมินค่าความยากของข้อสอบ (b) ชั้นประถมศึกษาปีที่ 3 มีดังนี้

ค่าความยากเฉลี่ยมากกว่า	2.0000			หมายถึง	ข้อสอบยากมาก
ค่าความยากเฉลี่ยตั้งแต่	1.0001	ถึง	2.0000	หมายถึง	ข้อสอบยาก
ค่าความยากเฉลี่ยตั้งแต่	0.5001	ถึง	1.0000	หมายถึง	ข้อสอบค่อนข้องยาก
ค่าความยากเฉลี่ยตั้งแต่	-0.4999	ถึง	0.5000	หมายถึง	ข้อสอบปานกลาง
ค่าความยากเฉลี่ยตั้งแต่	-0.9999	ถึง	-0.5000	หมายถึง	ข้อสอบค่อนข้างง่าย
ค่าความยากเฉลี่ยตั้งแต่	-1.9999	ถึง	-1.0000	หมายถึง	ข้อสอบง่าย
ค่าความยากเฉลี่ยน้อยกว่า	-2.0000			หมายถึง	ข้อสอบง่ายมาก

สรุปผลค่าพารามิเตอร์ของข้อสอบรายข้อ ชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้าน ได้แก่ 1) ด้านภาษา 2) ด้านคำนวณ และ 3) ด้านเหตุผล ดังตารางที่ 4-1

ตารางที่ 4-1 ผลการวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา จำนวน 30 ข้อ ตามหลักการของทฤษฎีการตอบสนองข้อสอบ (IRT) แบบ 3 พารามิเตอร์

ข้อสอบ NT	ข้อที่	ค่าพารามิเตอร์		
		a	b	c
ด้านภาษา	1	0.546	-0.266	0.066
	2	0.543	-0.281	0.067
	3	0.566	-1.094	0.225
	4	0.495	0.121	0.407
	5	0.408	2.297	0.256
	6	0.370	-0.971	0.255
	7	0.800	2.714	0.201
	8	1.013	4.000	0.182
	9	0.549	0.549	0.241
	10	0.313	0.985	0.264
	11	0.456	2.265	0.271
	12	0.612	-0.430	0.259
	13	0.592	-0.278	0.247
	14	0.865	0.606	0.236
	15	0.754	0.401	0.228
	16	1.241	3.849	0.216
	17	0.742	0.561	0.229
	18	1.170	2.834	0.182
	19	0.546	-0.328	0.247
	20	0.604	0.419	0.247
	21	1.156	2.179	0.219
	22	0.672	1.377	0.235
	23	1.071	3.939	0.375
	24	0.587	0.875	0.261
	25	0.550	0.454	0.253
	26	0.937	1.174	0.234
	27	0.851	1.040	0.234
	28	1.162	4.000	0.217
	29	0.656	-0.016	0.257
	30	0.744	0.387	0.252

จากตารางที่ 4-1 ข้อสอบ NT ด้านภาษา จำนวน 30 ข้อ เป็นข้อสอบมีค่าอำนาจจำแนกของข้อสอบ (a) ตั้งแต่ 0.313 ถึง 1.241 ค่าความยากของข้อสอบ (b) ตั้งแต่ -1.094 ถึง 4.000 และค่าโอกาสการเดาของข้อสอบ (c) ตั้งแต่ 0.066 ถึง 0.407

ตารางที่ 4-2 ผลการวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านคำนวณ จำนวน 30 ข้อ ตามหลักการของทฤษฎีการตอบสนองข้อสอบ (IRT) แบบ 3 พารามิเตอร์

ข้อสอบ NT	ข้อที่	ค่าพารามิเตอร์		
		a	b	c
ด้านคำนวณ	1	0.272	0.743	0.155
	2	0.389	0.823	0.142
	3	0.539	1.840	0.257
	4	0.540	3.002	0.232
	5	0.666	1.666	0.246
	6	0.544	2.273	0.248
	7	0.901	1.717	0.223
	8	0.742	1.843	0.305
	9	0.833	3.409	0.207
	10	1.040	1.592	0.202
	11	0.595	2.707	0.219
	12	1.099	1.702	0.228
	13	1.056	1.582	0.218
	14	0.929	2.015	0.214
	15	0.851	3.491	0.248
	16	1.382	1.586	0.209
	17	0.503	1.070	0.253
	18	0.741	0.748	0.222
	19	0.969	1.264	0.230
	20	1.037	2.045	0.184
	21	0.610	0.773	0.239
	22	0.811	2.796	0.371
	23	0.657	-0.066	0.244
	24	1.197	1.847	0.215
	25	0.829	1.946	0.210
	26	1.140	4.000	0.201
	27	1.144	2.796	0.195
	28	0.686	0.958	0.257
	29	1.026	1.473	0.222
	30	1.381	1.855	0.239

จากตารางที่ 4-2 ข้อสอบ NT ด้านคำนวณ จำนวน 30 ข้อ เป็นข้อสอบมีค่าอำนาจจำแนกของข้อสอบ (a) ตั้งแต่ 0.272 ถึง 1.382 ค่าความยากของข้อสอบ (b) ตั้งแต่ -0.066 ถึง 4.000 และค่าโอกาสการเดาของข้อสอบ (c) ตั้งแต่ 0.142 ถึง 0.371

ตารางที่ 4-3 ผลการวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านเหตุผล จำนวน 30 ข้อ ตามหลักการของทฤษฎีการตอบสนองข้อสอบ (IRT) แบบ 3 พารามิเตอร์

ข้อสอบ NT	ข้อที่	ค่าพารามิเตอร์		
		a	b	c
ด้านเหตุผล	1	0.670	1.616	0.233
	2	0.999	-0.537	0.247
	3	0.902	0.528	0.348
	4	0.890	0.155	0.241
	5	0.364	2.750	0.284
	6	0.732	1.568	0.231
	7	0.944	2.896	0.277
	8	0.673	1.877	0.272
	9	0.486	1.369	0.249
	10	0.821	1.686	0.252
	11	0.953	0.333	0.230
	12	0.952	0.098	0.243
	13	0.903	2.441	0.280
	14	1.550	1.827	0.238
	15	0.429	1.666	0.247
	16	0.520	2.775	0.284
	17	1.164	3.993	0.258
	18	0.770	0.784	0.232
	19	0.846	0.374	0.218
	20	1.111	3.174	0.239
	21	0.887	2.112	0.269
	22	0.740	1.665	0.217
	23	0.966	-0.073	0.258
	24	0.855	0.420	0.259
	25	0.640	0.768	0.251
	26	0.793	0.824	0.252
	27	1.119	3.674	0.269
	28	0.524	0.890	0.256
	29	0.757	-0.123	0.228
	30	0.745	0.241	0.245

จากตารางที่ 4-3 ข้อสอบ NT ด้านเหตุผล จำนวน 30 ข้อ ซึ่งเป็นข้อสอบที่มีค่าอำนาจจำแนกของข้อสอบ (a) ตั้งแต่ 0.364 ถึง 1.550 ค่าความยากของข้อสอบ (b) ตั้งแต่ -0.537 ถึง 3.993 และค่าโอกาสการเดาของข้อสอบ (c) ตั้งแต่ 0.217 ถึง 0.348

จากการวิเคราะห์คุณภาพของข้อสอบ โดยการวิเคราะห์ค่าพารามิเตอร์ตามทฤษฎีการตอบสนองข้อสอบ (IRT) แบบ 3 พารามิเตอร์ ปรากฏว่า

ข้อสอบ NT ด้านภาษา มีค่าอำนาจจำแนกของข้อสอบ (a) มีค่าเฉลี่ยเท่ากับ 0.719 ค่าความยากของข้อสอบ (b) มีค่าเฉลี่ยเท่ากับ 1.112 และค่าโอกาสการเดาของข้อสอบ (c) มีค่าเฉลี่ยเท่ากับ 0.235 ข้อสอบ NT ด้านภาษา มีค่าอำนาจจำแนกข้อสอบอยู่ในระดับค่อนข้างดี มีค่าความยากของข้อสอบอยู่ในระดับยาก และมีค่าโอกาสของการเดาของข้อสอบ เฉลี่ยไม่เกิน 0.30

ข้อสอบ NT ด้านคำนวณ มีค่าอำนาจจำแนกของข้อสอบ (a) มีค่าเฉลี่ยเท่ากับ 0.837 ค่าความยากของข้อสอบ (b) มีค่าเฉลี่ยเท่ากับ 1.850 และค่าโอกาสการเดาของข้อสอบ (c) มีค่าเฉลี่ยเท่ากับ 0.228 ข้อสอบ NT ด้านคำนวณ มีค่าอำนาจจำแนกข้อสอบอยู่ในระดับค่อนข้างดี มีค่าความยากของข้อสอบ อยู่ในระดับยาก และมีค่าโอกาสของการเดาของข้อสอบ เฉลี่ยไม่เกิน 0.30

ข้อสอบ NT ด้านเหตุผล มีค่าอำนาจจำแนกของข้อสอบ (a) มีค่าเฉลี่ยเท่ากับ 0.823 ค่าความยากของข้อสอบ (b) มีค่าเฉลี่ยเท่ากับ 1.392 และค่าโอกาสการเดาของข้อสอบ (c) มีค่าเฉลี่ยเท่ากับ 0.254 ข้อสอบ NT ด้านเหตุผล มีค่าอำนาจจำแนกข้อสอบอยู่ในระดับค่อนข้างดี มีค่าความยากของข้อสอบ อยู่ในระดับยาก และมีค่าโอกาสของการเดาของข้อสอบ เฉลี่ยไม่เกิน 0.30

สามารถสรุปได้ว่า ข้อสอบ NT ทั้ง 3 ด้าน มีค่าอำนาจจำแนกของข้อสอบ (a) อยู่ในระดับที่สามารถจำแนกผู้สอบได้ค่อนข้างดี มีค่าความยากของข้อสอบ (b) อยู่ในระดับยาก และมีค่าโอกาสของการเดาของข้อสอบ (c) เฉลี่ยไม่เกิน 0.30

ตารางที่ 4-4 สรุปค่าเฉลี่ยพารามิเตอร์ของข้อสอบ NT ตามหลักการทฤษฎีการตอบสนองข้อสอบแบบ 3 พารามิเตอร์

ข้อสอบ NT	ค่าเฉลี่ยของค่าอำนาจจำแนก (a)	ค่าเฉลี่ยของค่าความยากของข้อสอบ (b)	ค่าเฉลี่ยของค่าโอกาสการเดาของข้อสอบ (c)	การแปลความหมาย
ด้านภาษา	0.720	1.112	0.235	ยาก
ด้านคำนวณ	0.837	1.849	0.228	ยาก
ด้านเหตุผล	0.824	1.392	0.254	ยาก
รวม	0.794	1.451	0.239	ยาก

จากตารางที่ 4-4 แสดงค่าเฉลี่ยพารามิเตอร์ของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ค่าอำนาจจำแนกของข้อสอบ (a) เฉลี่ยรวมทั้ง 3 ด้าน เท่ากับ 0.794 ค่าความยากของข้อสอบ (b) เฉลี่ยรวมทั้ง 3 ด้าน เท่ากับ 1.451 และค่าโอกาสการเดาของข้อสอบ (c) เฉลี่ยรวมทั้ง 3 ด้าน เท่ากับ 0.239 โดยมีค่าความยากของข้อสอบ (b) อยู่ในระดับยาก

ตอนที่ 2 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล จำนวนด้านละ 30 ข้อ ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN

ตารางที่ 4-5 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM

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

ตารางที่ 4-5 (ต่อ)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี HGLM		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
28	**DIF**	NO-DIF	**DIF**
29	NO-DIF	NO-DIF	**DIF**
30	NO-DIF	NO-DIF	DIF
จำนวนข้อที่พบ DIF	(9 ข้อ) 30.00%	(5 ข้อ) 16.67%	(17 ข้อ) 56.67%

หมายเหตุ DIF หมายถึง ข้อสอบที่พบว่าทำหน้าที่ต่างกัน
NO-DIF หมายถึง ข้อสอบที่ไม่พบว่าทำหน้าที่ต่างกัน

จากตารางที่ 4-5 แสดงผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ NT ทั้ง 3 ด้าน ด้วยวิธี HGLM ปรากฏว่า

ด้านภาษา ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 9 ข้อ ได้แก่ ข้อที่ 2, 4, 10, 12, 13, 19, 26, 27 และ 28 คิดเป็นร้อยละ 30 ของข้อสอบทั้งหมด

ด้านคำนวณ ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 5 ข้อ ได้แก่ ข้อที่ 12, 23, 24, 25 และ 26 คิดเป็นร้อยละ 16.67 ของข้อสอบทั้งหมด

ด้านเหตุผล ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 17 ข้อ ได้แก่ ข้อที่ 3, 4, 8, 9, 10, 11, 12, 16, 17, 19, 22, 23, 24, 25, 28, 29 และ 30 คิดเป็นร้อยละ 56.67 ของข้อสอบทั้งหมด

ตารางที่ 4-6 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี MIMIC

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี MIMIC		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
1	NO DIF	NO DIF	NO DIF
2	**DIF**	NO DIF	NO DIF
3	NO DIF	NO DIF	NO DIF
4	NO DIF	NO DIF	NO DIF
5	NO DIF	**DIF**	NO DIF
6	NO DIF	NO DIF	**DIF**
7	NO DIF	NO DIF	NO DIF
8	NO DIF	NO DIF	NO DIF
9	NO DIF	NO DIF	**DIF**
10	NO DIF	NO DIF	**DIF**
11	NO DIF	**DIF**	NO DIF
12	NO DIF	**DIF**	NO DIF
13	NO DIF	NO DIF	**DIF**

ตารางที่ 4-6 (ต่อ)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี MIMIC		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
14	NO DIF	NO DIF	DIF
15	NO DIF	NO DIF	DIF
16	NO DIF	NO DIF	DIF
17	NO DIF	NO DIF	DIF
18	NO DIF	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	DIF
23	NO DIF	DIF	NO DIF
24	NO DIF	DIF	NO DIF
25	NO DIF	DIF	NO DIF
26	NO DIF	DIF	NO DIF
27	NO DIF	NO DIF	NO DIF
28	NO DIF	NO DIF	DIF
29	NO DIF	NO DIF	NO DIF
30	NO DIF	NO DIF	DIF
จำนวนข้อที่พบ DIF	(1 ข้อ) 3.33%	(8 ข้อ) 26.67%	(11 ข้อ) 36.67%

หมายเหตุ DIF หมายถึง ข้อสอบที่พบว่าทำหน้าที่ต่างกัน

NO-DIF หมายถึง ข้อสอบที่ไม่พบว่าทำหน้าที่ต่างกัน

จากตารางที่ 4-6 แสดงผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ NT ทั้ง 3 ด้าน ด้วยวิธี MIMIC ปรากฏว่า

ด้านภาษา ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 1 ข้อ ได้แก่ ข้อที่ 2 คิดเป็นร้อยละ 3.33 ของข้อสอบทั้งหมด

ด้านคำนวณ ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 8 ข้อ ได้แก่ ข้อที่ 5, 11, 12, 18, 23, 24, 25 และ 26 คิดเป็นร้อยละ 26.67 ของข้อสอบทั้งหมด

ด้านเหตุผล ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 11 ข้อ ได้แก่ ข้อที่ 6, 9, 10, 13, 14, 15, 16, 17, 22, 28 และ 30 คิดเป็นร้อยละ 36.67 ของข้อสอบทั้งหมด

ตารางที่ 4-7 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี BAYESIAN

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี BAYESIAN		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
1	NO DIF	NO DIF	NO DIF
2	DIF	DIF	NO DIF
3	NO DIF	NO DIF	NO DIF
4	NO DIF	NO DIF	DIF
5	NO DIF	NO DIF	NO DIF
6	NO DIF	NO DIF	NO DIF
7	NO DIF	NO DIF	NO DIF
8	NO DIF	NO DIF	DIF
9	NO DIF	NO DIF	NO DIF
10	DIF	NO DIF	NO DIF
11	NO DIF	NO DIF	NO DIF
12	NO DIF	NO DIF	DIF
13	NO DIF	NO DIF	DIF
14	DIF	NO DIF	NO DIF
15	DIF	NO DIF	DIF
16	DIF	NO DIF	NO DIF
17	DIF	NO DIF	DIF
18	NO DIF	NO DIF	NO DIF
19	NO DIF	NO DIF	DIF
20	NO DIF	DIF	NO DIF
21	NO DIF	DIF	DIF
22	NO DIF	NO DIF	NO DIF
23	NO DIF	DIF	NO DIF
24	NO DIF	NO DIF	NO DIF
25	NO DIF	DIF	DIF
26	NO DIF	DIF	DIF
27	NO DIF	NO DIF	NO DIF
28	NO DIF	NO DIF	NO DIF
29	NO DIF	NO DIF	NO DIF
30	DIF	NO DIF	DIF
จำนวนข้อที่พบ DIF	7 ข้อ (23.33%)	6 ข้อ (20.00%)	11 ข้อ (36.67%)

หมายเหตุ DIF หมายถึง ข้อสอบที่พบว่าทำหน้าที่ต่างกัน
NO-DIF หมายถึง ข้อสอบที่ไม่พบว่าทำหน้าที่ต่างกัน

จากตารางที่ 4-7 แสดงผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ NT ทั้ง 3 ด้าน ด้วยวิธี BAYESIAN ปรากฏว่า

ด้านภาษา ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 7 ข้อ ได้แก่ ข้อที่ 2,10,14,15,16,17 และ 30 คิดเป็นร้อยละ 23.33 ของข้อสอบทั้งหมด

ด้านคำนวณ ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 6 ข้อ ได้แก่ ข้อที่ 2,20,21,23,25 และ 26 คิดเป็นร้อยละ 20.00 ของข้อสอบทั้งหมด

ด้านเหตุผล ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน จำนวน 11 ข้อ ได้แก่ ข้อที่ 4,8,12,13,15, 17,19,21,25,26 และ 30 คิดเป็นร้อยละ 36.67 ของข้อสอบทั้งหมด

ตารางที่ 4-8 สรุปผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN

แบบทดสอบ	จำนวน (ข้อ)	วิธี HGLM		วิธี MIMIC		วิธี BAYESIAN	
		จำนวนข้อที่พบ DIF	ร้อยละ	จำนวนข้อที่พบ DIF	ร้อยละ	จำนวนข้อที่พบ DIF	ร้อยละ
ด้านภาษา	30	9	30.00	1	3.33	7	23.33
ด้านคำนวณ	30	5	16.67	8	26.67	6	20.00
ด้านเหตุผล	30	17	56.67	11	36.67	11	36.67
รวมทั้ง 3 ด้าน	90	31	34.44	20	22.22	24	26.67

จากตารางที่ 4-8 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้าน ปรากฏว่า

วิธี HGLM ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ NT ด้านภาษา ตรวจพบ DIF จำนวน 9 ข้อ คิดเป็นร้อยละ 30.00 ของข้อสอบทั้งหมด ด้านคำนวณ ตรวจพบ DIF จำนวน 5 ข้อ คิดเป็นร้อยละ 16.67 ของข้อสอบทั้งหมด และด้านเหตุผล ตรวจพบ DIF จำนวน 17 ข้อ คิดเป็นร้อยละ 56.67 ของข้อสอบทั้งหมด

วิธี MIMIC ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบความสามารถพื้นฐานของผู้เรียนระดับชาติ ด้านภาษา จำนวน 1 ข้อ คิดเป็นร้อยละ 3.33% ของข้อสอบทั้งหมด ด้านคำนวณพบ DIF จำนวน 8 ข้อ คิดเป็นร้อยละ 26.67 ของข้อสอบทั้งหมด และด้านเหตุผลตรวจพบ DIF จำนวน 11 ข้อ คิดเป็นร้อยละ 36.67 ของข้อสอบทั้งหมด

วิธี BAYESIAN ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ NT ด้านภาษา ตรวจพบ DIF จำนวน 7 ข้อ เป็นร้อยละ 23.33 ของข้อสอบทั้งหมด สำหรับด้านคำนวณ ตรวจพบ DIF จำนวน 6 ข้อ คิดเป็นร้อยละ 20.00 ของข้อสอบทั้งหมด และด้านเหตุผล ตรวจพบ DIF จำนวน 11 ข้อ คิดเป็นร้อยละ 36.67 ของข้อสอบทั้งหมด

ตอนที่ 3 การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้าน ได้แก่ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN

การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ทั้ง 3 วิธี ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN

ตารางที่ 4-9 ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ NT ด้านภาษา ทั้ง 3 วิธี

ข้อที่	เปรียบเทียบวิธีการตรวจสอบ DIF					
	ด้านภาษา		ด้านภาษา		ด้านภาษา	
	วิธี HGLM	วิธี MIMIC	วิธี MIMIC	วิธี BAYESIAN	วิธี HGLM	วิธี BAYESIAN
1	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
2	DIF	DIF	DIF	DIF	DIF	DIF
3	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
4	DIF	NO DIF	NO DIF	NO DIF	DIF	NO DIF
5	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
6	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
7	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
8	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
9	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
10	DIF	NO DIF	NO DIF	DIF	DIF	DIF
11	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
12	DIF	NO DIF	NO DIF	NO DIF	DIF	NO DIF
13	DIF	NO DIF	NO DIF	NO DIF	DIF	NO DIF
14	NO DIF	NO DIF	NO DIF	DIF	NO DIF	DIF
15	NO DIF	NO DIF	NO DIF	DIF	NO DIF	DIF
16	NO DIF	NO DIF	NO DIF	DIF	NO DIF	DIF
17	NO DIF	NO DIF	NO DIF	DIF	NO DIF	DIF
18	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
19	DIF	NO DIF	NO DIF	NO DIF	DIF	NO DIF
20	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
21	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
22	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
23	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
24	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
25	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
26	DIF	NO DIF	NO DIF	NO DIF	DIF	NO DIF
27	DIF	NO DIF	NO DIF	NO DIF	DIF	NO DIF
28	DIF	NO DIF	NO DIF	NO DIF	DIF	NO DIF
29	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
30	NO DIF	NO DIF	NO DIF	DIF	NO DIF	DIF
จำนวนข้อที่พบ DIF	9 ข้อ	1 ข้อ	1 ข้อ	7 ข้อ	9 ข้อ	7 ข้อ
ร้อยละ	30	3.33	3.33	23.33	30	23.33

ตารางที่ 4-10 ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ NT ด้านคำนวณ ทั้ง 3 วิธี

ข้อที่	วิธีการตรวจสอบ DIF					
	ด้านคำนวณ		ด้านคำนวณ		ด้านคำนวณ	
	วิธี HGLM	วิธี MIMIC	วิธี MIMIC	วิธี BAYESIAN	วิธี HGLM	วิธี BAYESIAN
1	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
2	NO DIF	NO DIF	NO DIF	DIF	NO DIF	DIF
3	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
4	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
5	NO DIF	DIF	DIF	NO DIF	NO DIF	NO DIF
6	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
7	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
8	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
9	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
10	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
11	NO DIF	DIF	DIF	NO DIF	NO DIF	NO DIF
12	DIF	DIF	DIF	NO DIF	DIF	NO DIF
13	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
14	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
15	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
16	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
17	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
18	NO DIF	DIF	DIF	NO DIF	NO DIF	NO DIF
19	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
20	NO DIF	NO DIF	NO DIF	DIF	NO DIF	DIF
21	NO DIF	NO DIF	NO DIF	DIF	NO DIF	DIF
22	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
23	DIF	DIF	DIF	DIF	DIF	DIF
24	DIF	DIF	DIF	NO DIF	DIF	NO DIF
25	DIF	DIF	DIF	DIF	DIF	DIF
26	DIF	DIF	DIF	DIF	DIF	DIF
27	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
28	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
29	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
30	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
จำนวนข้อที่พบ DIF	5 ข้อ	8 ข้อ	8 ข้อ	6 ข้อ	5 ข้อ	6 ข้อ
ร้อยละ	16.67	26.67	26.67	20	16.67	20

ตารางที่ 4-11 ผลการเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ด้านเหตุผล ระหว่างวิธี HGLM กับวิธี MIMIC วิธี MIMIC กับวิธี BAYESIAN และ วิธี HGLM กับวิธี BAYESIAN

ข้อที่	วิธีการตรวจสอบ DIF					
	ด้านเหตุผล		ด้านเหตุผล		ด้านเหตุผล	
	วิธี HGLM	วิธี MIMIC	วิธี MIMIC	วิธี BAYESIAN	วิธี HGLM	วิธี BAYESIAN
1	NO DIF	NO DIF	NO DIF	NO DIF	วิธี HGLM	NO DIF
2	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
3	DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
4	NO DIF	NO DIF	NO DIF	DIF	DIF	DIF
5	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
6	NO DIF	DIF	DIF	NO DIF	NO DIF	NO DIF
7	DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
8	DIF	NO DIF	NO DIF	DIF	DIF	DIF
9	DIF	DIF	DIF	NO DIF	DIF	NO DIF
10	DIF	DIF	DIF	NO DIF	DIF	NO DIF
11	DIF	NO DIF	NO DIF	NO DIF	DIF	NO DIF
12	DIF	NO DIF	NO DIF	DIF	DIF	DIF
13	NO DIF	DIF	DIF	DIF	DIF	DIF
14	NO DIF	DIF	DIF	NO DIF	NO DIF	NO DIF
15	NO DIF	DIF	DIF	DIF	NO DIF	DIF
16	DIF	DIF	DIF	NO DIF	NO DIF	NO DIF
17	DIF	DIF	DIF	DIF	DIF	DIF
18	NO DIF	NO DIF	NO DIF	NO DIF	DIF	NO DIF
19	DIF	NO DIF	NO DIF	DIF	NO DIF	DIF
20	NO DIF	NO DIF	NO DIF	NO DIF	DIF	NO DIF
21	NO DIF	NO DIF	NO DIF	DIF	NO DIF	DIF
22	DIF	DIF	DIF	NO DIF	NO DIF	NO DIF
23	DIF	NO DIF	NO DIF	NO DIF	DIF	NO DIF
24	DIF	NO DIF	NO DIF	NO DIF	DIF	NO DIF
25	DIF	NO DIF	NO DIF	DIF	DIF	DIF
26	NO DIF	NO DIF	NO DIF	DIF	DIF	DIF
27	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF	NO DIF
28	DIF	DIF	DIF	NO DIF	DIF	NO DIF
29	DIF	NO DIF	NO DIF	NO DIF	DIF	NO DIF
30	DIF	DIF	DIF	DIF	DIF	DIF
จำนวนข้อที่พบ DIF	17 ข้อ	11 ข้อ	11 ข้อ	11 ข้อ	17 ข้อ	11 ข้อ
ร้อยละ	56.67	36.67	36.67	36.67	56.67	36.67

ตารางที่ 4-12 ผลการเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบNT ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN ทั้ง 3 ด้าน

แบบทดสอบ NT	เปรียบเทียบร้อยละของการตรวจพบ DIF		
	วิธี HGLM กับ วิธี MIMIC	วิธี MIMIC กับ วิธี BAYESIAN	วิธี HGLM กับ วิธี BAYESIAN
ด้านภาษา	วิธี HGLM > วิธี MIMIC (8 ข้อ, 26.67%)	วิธี MIMIC < วิธี BAYESIAN (6 ข้อ, 20%)	วิธี HGLM > วิธี BAYESIAN (2 ข้อ, 6.67%)
ด้านคำนวณ	วิธี HGLM < วิธี MIMIC (3 ข้อ, 10%)*	วิธี MIMIC > วิธีBAYESIAN (2 ข้อ, 6.67%)	วิธี HGLM < วิธี BAYESIAN (1 ข้อ, 3.33%)*
ด้านเหตุผล	วิธี HGLM > วิธี MIMIC (6 ข้อ, 20%)	วิธี MIMIC = วิธี BAYESIAN (0 ข้อ, 0 %)	วิธี HGLM > วิธี BAYESIAN (6 ข้อ, 20%)

หมายเหตุ *$p<.05$

วิธี HGLM > วิธี MIMIC หมายถึง วิธี HGLM ตรวจพบ DIF ได้มากกว่า วิธี MIMIC

วิธี HGLM > วิธี BAYESIAN หมายถึง วิธี HGLM ตรวจพบ DIF ได้มากกว่า วิธี BAYESIAN

วิธี HGLM < วิธี MIMIC หมายถึง วิธี HGLM ตรวจพบ DIF ได้น้อยกว่า วิธี MIMIC

วิธี HGLM < วิธี BAYESIAN หมายถึง วิธี HGLM ตรวจพบ DIF ได้น้อยกว่า วิธี BAYESIAN

วิธี MIMIC > วิธี BAYESIAN หมายถึง วิธี MIMIC ตรวจพบ DIF ได้มากกว่า วิธี BAYESIAN

วิธี MIMIC < วิธี BAYESIAN หมายถึง วิธี MIMIC ตรวจพบ DIF ได้น้อยกว่า วิธี BAYESIAN

วิธี MIMIC = วิธี BAYESIAN หมายถึง วิธี MIMIC ตรวจพบ DIF ได้เท่ากับ วิธี BAYESIAN

จากตารางที่ 4-12 ผลการเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้าน ได้แก่ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ปรากฏว่า

ด้านภาษา วิธี HGLM ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน มากกว่าวิธี MIMIC จำนวน 8 ข้อ คิดเป็นร้อยละ 26.67 และวิธี HGLM ตรวจพบ DIF มากกว่าวิธี BAYESIAN จำนวน 2 ข้อ คิดเป็นร้อยละ 6.67 และวิธี MIMIC ตรวจพบ DIF น้อยกว่าวิธี BAYESIAN จำนวน 6 ข้อ คิดเป็นร้อยละ 20 ของข้อสอบทั้งหมด

ด้านคำนวณ วิธี HGLM ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน น้อยกว่าวิธี MIMIC จำนวน 3 ข้อ คิดเป็นร้อยละ 10 และวิธี HGLM ตรวจพบ DIF น้อยกว่าวิธี BAYESIAN จำนวน 1 ข้อ คิดเป็นร้อยละ 3.33 และวิธี MIMIC ตรวจพบ DIF มากกว่าวิธี BAYESIAN จำนวน 2 ข้อ คิดเป็นร้อยละ 6.67 ของข้อสอบทั้งหมด

ด้านเหตุผล วิธี HGLM ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน มากกว่า วิธี MIMIC จำนวน 6 ข้อ คิดเป็นร้อยละ 20 และวิธี HGLM ตรวจพบ DIF มากกว่าวิธี BAYESIAN จำนวน 6 ข้อ คิดเป็นร้อยละ 20 และวิธี MIMIC ตรวจพบ DIF เท่ากับวิธี BAYESIAN จำนวน 0 ข้อ คิดเป็นร้อยละ 0 ของข้อสอบทั้งหมด

สรุปผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้าน พบว่า ด้านคำนวณ วิธี HGLM ตรวจพบ DIF น้อยกว่าวิธี MIMIC และน้อยกว่าวิธี BAYESIAN คิดเป็นร้อยละ 10 และร้อยละ 3.33 แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ .05 ด้านภาษา วิธี HGLM ตรวจพบ DIF มากกว่าวิธี MIMIC และมากกว่าวิธี BAYESIAN คิดเป็นร้อยละ

26.67 และร้อยละ 6.67 วิธี MIMIC ตรวจพบ DIF น้อยกว่าวิธี BAYESIAN คิดเป็นร้อยละ 20 ด้านคำนวณ วิธี MIMIC ตรวจพบ DIF มากกว่าวิธี BAYESIAN ส่วนด้านเหตุผล วิธี HGLM ตรวจพบ DIF มากกว่าวิธี MIMIC และวิธี BAYESIAN คิดเป็นร้อยละ 20 เท่ากัน วิธี MIMIC ตรวจพบ DIF เท่ากับวิธี BAYESIAN ผลการตรวจสอบไม่แตกต่างกัน

บทที่ 5
สรุปและอภิปรายผล

การวิจัยนี้มีวัตถุประสงค์เพื่อ 1) วิเคราะห์คุณภาพของข้อสอบ NT ทั้ง 3 ด้าน 2) ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN และ 3) เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN ใช้ข้อมูลทุติยภูมิ ซึ่งเป็นผลการตอบข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2555 จำนวน 3 ด้าน ประกอบด้วย 1) ด้านภาษา 2) ด้านคำนวณ และ 3) ด้านเหตุผล จากสำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน กระทรวงศึกษาธิการ วิเคราะห์คุณภาพของข้อสอบโดยใช้ทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ ด้วยโปรแกรม Xcalibre Version 4.2.2 ซึ่งมีวิธีดำเนินการวิจัยเป็น 3 ระยะ ดังนี้ ระยะที่ 1 การวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ตามหลักการของทฤษฎีการตอบ สนองข้อสอบ (IRT) แบบ 3 พารามิเตอร์ ระยะที่ 2 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN ระยะที่ 3 การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN

สรุปผลการวิจัย

1. ผลการวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2555 ทั้ง 3 ด้าน ได้แก่ ด้านภาษา ด้านคำนวณ และด้านเหตุผล โดยการวิเคราะห์ตามทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ ปรากฏว่า ข้อสอบ NT ด้านภาษา มีค่าอำนาจจำแนกของข้อสอบ (a) มีค่าเฉลี่ยเท่ากับ 0.719 มีค่าอำนาจจำแนกข้อสอบอยู่ในระดับค่อนข้างดี ค่าความยากของข้อสอบ (b) มีค่าเฉลี่ยเท่ากับ 1.112 มีค่าความยากของข้อสอบอยู่ในระดับยาก และค่าโอกาสการเดาของข้อสอบ (c) มีค่าเฉลี่ยเท่ากับ 0.235 สำหรับ ด้านคำนวณ มีค่าอำนาจจำแนกของข้อสอบ (a) มีค่าเฉลี่ยเท่ากับ 0.837 มีค่าอำนาจจำแนกข้อสอบอยู่ในระดับค่อนข้างดี ค่าความยากของข้อสอบ (b) มีค่าเฉลี่ยเท่ากับ 1.850 มีค่าความยากของข้อสอบ อยู่ในระดับยาก และค่าโอกาสการเดาของข้อสอบ (c) มีค่าเฉลี่ยเท่ากับ 0.228 และด้านเหตุผล มีค่าอำนาจจำแนกของข้อสอบ (a) มีค่าเฉลี่ยเท่ากับ 0.823 มีค่าอำนาจจำแนกข้อสอบอยู่ในระดับค่อนข้างดี ค่าความยากของข้อสอบ (b) มีค่าเฉลี่ยเท่ากับ 1.392 มีค่าความยากของข้อสอบ อยู่ในระดับยากและค่าโอกาสการเดาของข้อสอบ (c) มีค่าเฉลี่ยเท่ากับ 0.254 สามารถสรุปได้ว่า ข้อสอบ NT ทั้ง 3 ด้าน มีค่าอำนาจจำแนกของข้อสอบ (a) อยู่ในระดับที่สามารถจำแนกผู้สอบได้ค่อนข้างดี มีค่าความยากของข้อสอบ (b) อยู่ในระดับยาก และมีค่าโอกาสของการเดาของข้อสอบ (c) เฉลี่ยไม่เกิน 0.30

2. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ระดับชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2555 ทั้ง 3 ด้าน ได้แก่ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN ปรากฏว่า วิธี HGLM ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ

NT ด้านภาษา ตรวจพบ DIFจำนวน 9 ข้อ คิดเป็นร้อยละ 30.00 ของข้อสอบทั้งหมด ด้านคำนวณ ตรวจพบ DIF จำนวน 5 ข้อ คิดเป็นร้อยละ 16.67 ของข้อสอบทั้งหมด และด้านเหตุผล ตรวจพบ DIF จำนวน 17 ข้อ คิดเป็นร้อยละ 56.67 ของข้อสอบทั้งหมด สำหรับ วิธี MIMIC ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบความสามารถพื้นฐานของผู้เรียนระดับชาติ ด้านภาษา จำนวน 1 ข้อ คิดเป็นร้อยละ 3.33% ของข้อสอบทั้งหมด ด้านคำนวณพบ DIF จำนวน 8 ข้อ คิดเป็นร้อยละ 26.67 ของข้อสอบทั้งหมด และด้านเหตุผล พบ DIF จำนวน 11 ข้อ คิดเป็นร้อยละ 36.67 ของข้อสอบทั้งหมด และ วิธี BAYESIAN ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ NT ด้านภาษา ตรวจพบ DIF จำนวน 7 ข้อ เป็นร้อยละ 23.33 ของข้อสอบทั้งหมด สำหรับด้านคำนวณ ตรวจพบ DIF จำนวน 6 ข้อ คิดเป็นร้อยละ 20.00 ของข้อสอบทั้งหมด และด้านเหตุผล ตรวจพบ DIF จำนวน 11 ข้อ คิดเป็นร้อยละ 36.67 ของข้อสอบทั้งหมด

3. ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ NT ระดับชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2555 ทั้ง 3 ด้าน ได้แก่ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN ปรากฏว่า ด้านภาษา วิธี HGLM ตรวจพบ DIF มากกว่าวิธี MIMIC คิดเป็นร้อยละ 26.67 และวิธี HGLM ตรวจพบ DIF มากกว่าวิธี BAYESIAN คิดเป็นร้อยละ 6.67 และวิธี MIMIC ตรวจพบ DIF น้อยกว่าวิธี BAYESIAN คิดเป็นร้อยละ 20 สำหรับด้านคำนวณ วิธี HGLM ตรวจพบ DIF น้อยกว่าวิธี MIMIC คิดเป็นร้อยละ 10 อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05 และวิธี HGLM ตรวจพบ DIF น้อยกว่าวิธี BAYESIAN คิดเป็นร้อยละ 3.33 อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05 และวิธี MIMIC ตรวจพบ DIF มากกว่าวิธี BAYESIAN คิดเป็นร้อยละ 6.67 และด้านเหตุผล วิธี HGLM ตรวจพบ DIF มากกว่าวิธี MIMIC คิดเป็นร้อยละ 20 และวิธี HGLM ตรวจพบ DIF มากกว่าวิธี BAYESIAN คิดเป็นร้อยละ 20 และวิธี MIMIC ตรวจพบ DIF เท่ากับวิธี BAYESIAN คิดเป็นร้อยละ 36.67

อภิปรายผล

ผลการวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2555 ตามหลักทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ ทั้ง 3 ด้าน ได้แก่ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ด้วยวิธี HGLM วิธี MIMIC และ วิธี BAYESIAN และเปรียบเทียบผลของการทำหน้าที่ต่างกันของข้อสอบ NT ทั้ง 3 วิธี มีประเด็นที่ควรอภิปราย ดังนี้

1. ผลการวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้าน ได้แก่ ด้านภาษา ด้านคำนวณ และด้านเหตุผล โดยใช้หลักการทฤษฎีการตอบสนองข้อสอบ (IRT) แบบ 3 พารามิเตอร์ ประกอบด้วย ค่าอำนาจจำแนกของข้อสอบ (a) ค่าความยากของข้อสอบ (b) และค่าโอกาสการเดาของข้อสอบ (c) มีค่าเฉลี่ยของค่าอำนาจจำแนกของข้อสอบ (a) รวมทั้ง 3 ด้าน เท่ากับ 0.794 สามารถจำแนกผู้เข้าสอบได้ดี โดยด้านคำนวณ มีค่าเฉลี่ยของค่าอำนาจจำแนกของข้อสอบ (a) สูงที่สุด มีค่าเท่ากับ 0.837 และด้านภาษา มีค่าเฉลี่ยของค่าอำนาจจำแนกของข้อสอบ (a) น้อยที่สุด มีค่าเท่ากับ 0.720 ส่วนค่าเฉลี่ยของค่าความยากของข้อสอบ (b) รวมทั้ง 3 ด้าน เท่ากับ 1.451 อยู่ในระดับยาก โดยด้านคำนวณ มีค่าความยากของข้อสอบ (b) สูงที่สุด เท่ากับ 1.849 และ

ด้านภาษา มีค่าความยากของข้อสอบ (b) น้อยที่สุด เท่ากับ 1.112 และมีค่าเฉลี่ยของค่าโอกาสการเดาของข้อสอบ (c) รวมทั้ง 3 ด้าน เท่ากับ 0.239

2. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธีการตรวจสอบ 3 วิธี ปรากฏว่า ด้านภาษา วิธี HGLM ตรวจพบ DIF ได้มากที่สุด รองลงมาคือ วิธี BAYESIAN และวิธี MIMIC ตามลำดับ ส่วนด้านคำนวณ วิธี MIMIC ตรวจพบ DIF ได้มากที่สุด รองลงมาคือ วิธี BAYESIAN และวิธี HGLM ตามลำดับ และด้านเหตุผล วิธี HGLM ตรวจพบ DIF ได้มากที่สุด รองลงมาคือ วิธี BAYESIAN และวิธี MIMIC ซึ่งตรวจพบ DIF ได้จำนวนเท่ากัน สอดคล้องกับงานวิจัยของ Ong, Lu, Lee, and Cohen (2015) ได้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR ผลการศึกษา ปรากฏว่า วิธี HGLM ตรวจสอบการทำหน้าที่ต่างกันได้มากที่สุด รองลงมา คือ วิธี IRT-LR และวิธี MIMIC สอดคล้องกับงานวิจัยของ สุธาทิพย์ ตรีสิน และปิยะทิพย์ ประดุจพรม (2560) ศึกษาเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR ผลการศึกษาปรากฏว่า วิธี HGLM ตรวจพบการทำหน้าที่ต่างกันของข้อสอบจำนวนมากที่สุด รองลงมาคือวิธี IRT-LR และวิธี MIMIC

3. ผลการเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ทั้ง 3 วิธี ปรากฏว่า ด้านภาษา วิธี HGLM ตรวจพบ DIF ได้มากที่สุด รองลงมาคือ วิธี BAYESIAN และวิธี MIMIC ตามลำดับ ส่วนด้านคำนวณ วิธี MIMIC ตรวจพบ DIF ได้มากที่สุด รองมาวิธี BAYESIAN และวิธี HGLM ตามลำดับ อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05 และด้านเหตุผล วิธี HGLM ตรวจพบ DIF ได้มากที่สุด รองลงมาคือ วิธี MIMIC และวิธี BAYESIAN ตรวจพบ DIF ได้เท่ากัน สอดคล้องกับงานวิจัยของ Ong, Lu, Lee, and Cohen (2015) ได้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR ผลการศึกษา ปรากฏว่า วิธี HGLM ตรวจสอบการทำหน้าที่ต่างกันได้มากที่สุด รองลงมา คือ วิธี IRT-LR และวิธี MIMIC สอดคล้องกับงานวิจัยของ สุธาทิพย์ ตรีสิน และปิยะทิพย์ ประดุจพรม (2560) ศึกษาการเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR ผลการศึกษาปรากฏว่า วิธี HGLM ตรวจพบการทำหน้าที่ต่างกันของข้อสอบ จำนวนมากที่สุด รองลงมาคือวิธี IRT-LR และวิธี MIMIC และสอดคล้องกับงานวิจัยของสุพัฒนา หอมบุปผา และคณะ ศึกษาการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ วิธีที่พบการทำหน้าที่ต่างกันของข้อสอบมากที่สุด คือ วิธี HGLM ลำดับที่สอง คือวิธี BAYESIAN และลำดับสุดท้ายคือวิธี MIMIC ส่วนความสามารถด้านคำนวณพบการทำหน้าที่ต่างกันของข้อสอบมากที่สุด คือ วิธี MIMIC วิธี BAYESIAN และวิธี HGLM ตามลำดับ สรุปได้ว่า วิธี HGLM มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมากกว่าวิธี MIMIC และวิธี BAYESIAN อาจเนื่องมาจากการวิเคราะห์ข้อสอบตามโมเดลการตอบสนองข้อสอบด้วยวิธีพหุระดับ (HGLM) มีความเหมาะสมทั้งในมิติที่ลักษณะการวิเคราะห์สอดคล้องกับธรรมชาติของข้อมูลของการทดสอบ และการกำหนดกรอบการวิเคราะห์อีกหลายประการ คือ สามารถดำเนินการศึกษาการทดสอบที่มีลักษณะเป็นการประเมินหลายมิติและการศึกษาที่มีหลายคุณลักษณะแฝงได้สะดวกมากขึ้น มีความสะดวกและเหมาะสมกับธรรมชาติ บริบทของข้อมูลในการศึกษาความแปรปรวนระหว่างสถานที่คงที่ (Fixed) ทางสังคม เช่น สถานศึกษา ที่ตั้ง สามารถวิเคราะห์ตัวแปร

ต่าง ๆ ได้อย่างเหมาะสมกับระดับของตัวแปรและลักษณะการวิเคราะห์ของโมเดลการตอบสนองข้อสอบ และสามารถวิเคราะห์ตัวแปรแฝงที่กำหนดในโมเดลการตอบสนองข้อสอบได้อย่างเดียวกับลักษณะของตัวแปรบรรยาย (explanatory variable) ส่วนผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบบางส่วนที่ไม่สอดคล้องตามสมมติฐานนั้น อาจเป็นเพราะในงานวิจัยนี้ใช้กลุ่มตัวอย่างเฉพาะจังหวัดสระแก้ว ที่ตั้งทางภูมิศาสตร์อยู่นอกเมือง ขาดประสบการณ์สภาพแวดล้อมและการฝึกปฏิบัติที่แตกต่างกันระหว่างนักเรียนในเมืองและนอกเมือง จึงทำให้ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบไม่สอดคล้องตามสมมติฐาน

ข้อเสนอแนะสำหรับการนำผลการวิจัยไปใช้

1. สำนักทดสอบทางการศึกษา สามารถนำผลการวิเคราะห์คุณภาพของข้อสอบระดับชาติที่ผ่านเกณฑ์การวิเคราะห์ตามทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) แบบ 3 พารามิเตอร์ ไปใช้สอบในครั้งต่อไป เพื่อวัดความสามารถของนักเรียนชั้นประถมศึกษาปีที่ 3 ของสำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน (สพฐ.)

2. นักวิจัยและนักวัดผลทางการศึกษาที่มีความสนใจเกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการตรวจสอบที่อยู่บนพื้นฐานทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) ทั้ง 3 วิธี คือ วิธี HGLM วิธี MIMIC และวิธี BAYESIAN สามารถเลือกใช้วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้

3. นักวิจัยและนักวัดผลทางการศึกษาที่มีความสนใจเกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสามารถเลือกใช้ขนาดกลุ่มตัวอย่างที่มีความเหมาะสมที่จะทำให้มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบดีที่สุด กลุ่มตัวอย่างขนาดใหญ่ (2,000 คน) จะทำให้สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันได้ดีกว่ากลุ่มตัวอย่างขนาดเล็ก

ข้อเสนอแนะสำหรับการวิจัยต่อไป

1. วิธี HGLM มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมากกว่าวิธี MIMIC และวิธี BAYESIAN ในแบบทดสอบที่มีการให้คะแนนแบบ 2 ค่า จึงควรมีการนำไปเปรียบเทียบเพิ่มเติมกับวิธีการตรวจสอบอื่น ๆ และตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบที่มีการให้คะแนนมากกว่า 2 ค่า

2. ควรมีการเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างข้อสอบรายข้อกับหมวดข้อสอบ ด้วยวิธี HGLM กับวิธีอื่นๆ เช่น SIBTEST เป็นต้น และใช้ผลการตอบข้อสอบอื่น ๆ เช่น การใช้ข้อสอบกลาง, O-NET รวมทั้งศึกษาตัวแปรต่าง ๆ ที่ส่งผลต่อประสิทธิภาพของการตรวจสอบการทำหน้าที่ต่างกัน เช่น อัตราส่วนระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ เป็นต้น

บรรณานุกรม

กระทรวงศึกษาธิการ. (2546). กฎกระทรวงแบ่งส่วนราชการสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน พ.ศ. 2546. *ราชกิจจานุเบกษา, 120*(63), 17.

กระทรวงศึกษาธิการ. (2551). *หลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน พุทธศักราช ๒๕๕๑*. พิมพ์ครั้งที่ 3. กรุงเทพฯ: โรงพิมพ์ชุมนุมสหกรณ์การเกษตรแห่งประเทศไทย จำกัด.

โกศล จิตวิรัตน์ ทักษิณา เครือหงส์ และเนตร์พัณณา ยาวิราช. (2554). ศักยภาพของโปรแกรม Mplus กับการวิเคราะห์สถิติขั้นสูงในงานวิจัย. *วารสารสมาคมนักวิจัย, 16*(3), 52-65.

จารุจิตร สิทธิปรุ ปิยะทิพย์ ตินวร และโสพส สุขานนท์สวัสดิ์. (2559). การพัฒนาโปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ สำหรับการจัดสอบ O-NET ระดับชั้นมัธยมศึกษาปีที่ 3. *วารสารการวัดผลการศึกษา มหาวิทยาลัยมหาสารคาม, 22*(2), 47-62.

ชนะศึก นิชานนท์ ศิริชัย กาญจนวาสี และMark Wilson. (2554). ประสิทธิภาพของการประมาณค่าพารามิเตอร์แบบเบส์โดยใช้การสรุปอ้างอิงความน่าเชื่อถือของโมเดลการตอบสนองข้อสอบ. *SDU Res. J, 7*(2), 59-75.

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

ณรงค์ จันทรมหา. (2554). การเปรียบเทียบค่าความเที่ยงของแบบทดสอบผลสัมฤทธิ์ทางการเรียนที่มีจำนวนข้อสอบทำหน้าที่ต่างกันแตกต่างกัน. *วิทยาการวิจัยและวิทยาการปัญญา, 8*(1), 58-71.

ธเกียรติกมล ทองงอก โชติกา ภาษีผล และศิริชัย กาญจนวาสี. (2556). ประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบภายใต้วิธีถดถอยโลจิสติกสำหรับข้อสอบที่ตรวจให้คะแนนแบบทวิวิภาค: เปรียบเทียบขนาดอิทธิพลสองเกณฑ์. *วารสารวิจัย มสด. สาขมนุษยศาสตร์สังคมศาสตร์, 9*(2), 31-49.

นงลักษณ์ วิรัชชัย. (2552). *โมเดลลิสเรล: สถิติวิเคราะห์สำหรับการวิจัย.* กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

นุภาพรรณ ปลื้มใจ ปิยะทิพย์ ตินวร และโสพส สุขานนท์สวัสดิ์. (2558). การพัฒนาโปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ สำหรับการจัดสอบ O-NET ระดับชั้นมัธยมศึกษาปีที่ 6. *วิทยาการวิจัยและวิทยาการปัญญา, 13*(2), 109-125.

ปริญญา เรืองทิพย์ และเดชา วรรณภากุล. (2554). ผลวิเคราะห์ความลำเอียงของข้อสอบด้วยวิธีแปลงค่าความยาก. *วารสารการวัดผลการศึกษา มหาวิทยาลัยมหาสารคาม, 17*(2), 57-66.

ปิยะทิพย์ ตินวร ม.ร.ว. สมพร สุทัศนีย์ และเสรี ชัดแช้ม. (2550). การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ: การเปรียบเทียบประสิทธิภาพระหว่างวิธีการวิเคราะห์องค์ประกอบจำกัดกับวิธีถดถอยโลจิสติกส์. *วิทยาการวิจัยและวิทยาการปัญญา, 5*(1), 63-80.

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

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

เรืองเดช ศิริกิจ. (2554). *การวิเคราะห์เปรียบเทียบโมเดลการประเมินคุณภาพการจัดการศึกษาวิชาคณิตศาสตร์: การประยุกต์ใช้โมเดลมูลค่าเพิ่มที่มีการวิเคราะห์การทำหน้าที่ต่างกันของตัวลวง*. กรุงเทพฯ: จุฬาลงกรณ์มหาวิทยาลัย.

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

วริษษา ชะม้อย. (2551). การเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ตรวจให้คะแนนแบบพหุวิภาคระหว่างวิธีโครงสร้างความแปรปรวนร่วมและค่าเฉลี่ยกับวิธีการวิเคราะห์ฟังก์ชันเชิงจำแนกแบบโลจิสติก. *วารสารอิเล็กทรอนิกส์ทางการศึกษา. 3*(1), 364-376.

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

ศิริชัย กาญจนวาสี. (2556). *ทฤษฎีการทดสอบแบบดั้งเดิม* (พิมพ์ครั้งที่ 7). กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

ศิริชัย กาญจนวาสี ทวีวัฒน์ ปิตยานนท์ และดิเรก ศรีสุโข. (2555). *การเลือกใช้สถิติที่เหมาะสมสำหรับการวิจัย* (พิมพ์ครั้งที่ 6). กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

ศิริรัตน์ สุคันธพฤกษ์. (2554). การวิเคราะห์ข้อคำถามในแบบวัดความวิตกกังวลในการสอบคณิตศาสตร์: การเปรียบเทียบระหว่างไฮราชิคอลลิเนียร์โมเดล พาเชียลเครดิตโมเดล และเกรดเรสพอนส์โมเดล. *วิทยาการวิจัยและวิทยาการปัญญา, 24*(2), 214-271.

สยามรัก สว่างศรี วราพร เอราวรรณ์ และทัศน์ศิรินทร์ สว่างบุญ (2560). การสร้างแบบทดสอบวินิจฉัยทางการเรียนกลุ่มสาระการเรียนรู้ภาษาต่างประเทศ เรื่อง ความสามารถด้านไวยากรณ์วิชาภาษาอังกฤษของนักเรียนชั้นมัธยมศึกษาปีที่ 2. *วารสารการวัดผลการศึกษา มหาวิทยาลัยมหาสารคาม, 23*(พิเศษ), 207-217.

สัมพันธ์ พันธุ์พฤกษ์. (2557). เอกสารประกอบการบรรยายการพัฒนาสมรรถนะด้านการวัดและประเมินผลการเรียนรู้. กรุงเทพฯ: สถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน).

สุชาดา กรเพชรปาณี ปิยะทิพย์ ตินวร และโสพส สุขานนท์สวัสดิ์ (2559). การพัฒนาโปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ สำหรับการจัดสอบ O-NET. *วิทยาการวิจัยและวิทยาการปัญญา, 14*(1), 14-31.

สุธาทิพย์ ตรีสิน และปิยะทิพย์ ประดุจพรม. (2560). การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ชั้นประถมศึกษาปีที่ 3 ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR. *วิทยาการวิจัยและวิทยาการปัญญา, 15*(2), 109-119.

สุพัฒนา หอมบุปผา ไพรัตน์ วงษ์นาม และสมพงษ์ ปั้นหุ่น. (2556). การเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN. *วารสารวิจัยราชภัฏพระนคร, 8*(2), 10-24.

สุภมาส อังศุโชติ สมถวิล วิจิตรวรรณา และรัชนีกูล ภิญโญภานุวัฒน์. (2554). *สถิติวิเคราะห์สำหรับการวิจัยทางสังคมศาสตร์และพฤติกรรมศาสตร์: เทคนิคการใช้โปรแกรม LISREL*. (พิมพ์ครั้งที่ 3). กรุงเทพฯ: เจริญดีมั่นคงการพิมพ์.

สุวิมล ติรกานันท์. (2553). *การวิเคราะห์ตัวแปรพหุในงานวิจัยทางสังคมศาสตร์*. กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

สำนักทดสอบทางการศึกษา. (2555). *ผลการประเมินคุณภาพผู้เรียนระดับชาติ ปีการศึกษา 2555 บทสรุปและข้อเสนอแนะเชิงนโยบาย.* กรุงเทพฯ: โรงพิมพ์ชุมนุมสหกรณ์การเกษตรแห่งประเทศไทย จำกัด.

สำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน. (2555). *คู่มือการจัดสอบประเมินคุณภาพการศึกษาขั้นพื้นฐานเพื่อการประกันคุณภาพผู้เรียน ปีการศึกษา 2555.* กรุงเทพฯ: กระทรวงศึกษาธิการ

สำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน. (2556). *คู่มือการจัดสอบประเมินคุณภาพการศึกษาขั้นพื้นฐานเพื่อการประกันคุณภาพผู้เรียน ปีการศึกษา 2556.* กรุงเทพฯ: กระทรวงศึกษาธิการ

สำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน. (2558). *คู่มือการจัดสอบประเมินคุณภาพการศึกษาขั้นพื้นฐานเพื่อการประกันคุณภาพผู้เรียน ปีการศึกษา 2558.* กรุงเทพฯ: กระทรวงศึกษาธิการ

สำราญ มีแจ้ง. (2552). การปรับเทียบคะแนน O-NET ระหว่างปีโดยวิธีการเทียบเป็นมาตราเดียวกันโดยใช้ทฤษฎีการตอบสนองข้อสอบ. *วิทยาการวิจัยและวิทยาการปัญญา, 7*(2), 83-92.

สำราญ มีแจ้ง ประภัสสร วงษ์ดี และยุพิน โกณฑา. (2552). การปรับเทียบคะแนน O-NET ระหว่างปี โดยวิธีการเทียบเป็นมาตราเดียวกันกับโดยใช้ทฤษฎีการตอบสนองข้อสอบ. *วิทยาการวิจัยและวิทยาการปัญญา, 7*(2), 81-92.

อัชฌา อระวีพร. (2554). การหาค่าตัวประมาณเบส์ด้วยโปรแกรมวินบั๊ก. *วารสารวิทยาศาสตร์ลาดกระบัง, 20*(2), 45-60.

อัชฌา อระวีพร. (2555). การวิเคราะห์เบส์จากโปรแกรมวินบั๊กสู่โปรแกรมอาร์. *NU Science Journal, 9*(1), 30-44.

อิทธิฤทธิ์ พงษ์ปิยะรัตน์. (2551). *การวิเคราะห์ข้อสอบและการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ: การวิเคราะห์พหุระดับ.* กรุงเทพฯ: จุฬาลงกรณ์มหาวิทยาลัย, ม.ป.ท.

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., & Hulya, K. (2010). Comparison of Differential Item Functioning Determination Techniques HGLM, LR, and IRT-LR. *Education Sciences*: *Theory & Practice, 11*(1), 639-649.

Acar, T. (2011). Sample Size in Differential Item Functioning: An Application of Hierarchical Linear Modeling. *Educational Sciences Theory & Practice, 11*(1), 284-288.

Acar, T. (2012). *Determination of a Differential Item Functioning Procedure Using the Hierarchical Generalized Linear Model: A Comparison Study With Logistic Regression and Likelihood Ratio Procedure*., 1-8.

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

Alan J. K., & Yoonsun, L. (2008). Simulated Test of Differential Item Functioning Using SIBTEST With and Without Impact. *Journal of Education Measurement Fall, 45*(3), 271-285.

Angoff, W. H. (1982). Summary and derivation of equating methods used at ETS. In P. Holland & D. Rubin (Eds.), Test equating. New York: Academic Press.

Brown, T. A. (2014). *Cofirmatory factor analysis for applied research*. NY.: Guilford Press.

Camilli, G., & Shepard, L. A. (1994). *Methods for indentifying biased test items.* L.: Sage Publication.

Cleary, T.A., & Hillton, T. L. (1968). Comparison of logistic regression and analysis of Variance differential item functioning detection methods. *Educational and Psychological Measurement, 2*(2), 29-35.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization Approach to assessing unexpected differential item performance on the Scholastic aptitude test, *Journal of Educational Measurement, 23*(4), 355-368.

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

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item Response theory.* Newbury Park, CA.: Sage Publications.

Holland, P. W., & Thayer, D.T. (1998). Differential item performance and the Mantel-Haenszel procedure. In Wainer, H., & Braun, H. L. (eds.), *Test Validity,* pp. 129-145, NJ.: Lawrence Erlbaum Associates.

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

Jensen, A. R. (1974). How biased are culture-loaded tests? *Genetic Psychology Monographs, 40*, 185-244.

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.

Mellenbergh, G. J. (1982) Contingency table models for assessing item bias. *Journal of Education Statistics, 7*(2), 105-118.

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.

Muninsakorn, Y., Tinnaworn, P., & Sukhanonsawat, S. (2015). Development of the Computerized Adaptive Testing Program for O-NET at the Grade 6 Level. *In Burapha University International Conference 2015: Moving Forword to a Prosperous and Sustainable Community, July 10-12, 2015 Bangsaen Heritafe Hotel Chonburi.* Thailand: Burapha University.

Muthen, B. O., Kao, C. F., & Burstein, L. (1991). Instructionally sensitive psychometrics: Achievement test items. *Journal of Educational Measurement, 28*(1), 1-22.

Muthen, L. K., & Muthen, B. O. (2007). *Mplus User's Guide* (5th ed.). Los Anggeles: Auther.

Muthen, L. K., & Muthen, B. O. (2010). Mplus User's Guide (6th ed.). Los Anggeles: Auther.

Narayanan, P., & Swaminathan, H. (1996). Identification of item that show nonuniform DIF. *Applied Psychological Measurement, 20*(3), 257-274.

Ong, M. L.,Lu, Lee, S., & Cohen, A. (2015). *A comparison of the hierarchical generalized linear modal, 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, 348-357.

Ong, Y., Williams, J., & Lamprianou, I. (2011). Exploration of the Validity of Gender Differences in Mathematics Assessment Using Differential Bundle Functioning. *International Journal of Testing, 11,* 271-293.

Oort, F.J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structral Equation Modeling, 5*(2), 107-124.

Park, C. (2010). Differential Item Functioning Analysis of an EFA Vocabulary Test. *English Teaching, 65*(3), 23-42.

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

Saengla Chaimongkol, Fred W. Huffer, and Akihito Kamata. (2006). A Bayesian Approach for Fitting a Random Effect Differential Item Functioning Across Group Units. *Thaland Statistician, 4,* 27-41.

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

Shealy, R., & Stout, W. F. (1993a). An item response theory model for test bias. In P.W. Holland & Wainer (Eds), *Differential Item Functioning : Theory and practic (197-239).* Hillsadale, NJ. : Lawerence Erlbaum.

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

Swaminathan, H., & Roger, 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.

Wainer, H. (2000). *Computerized Adaptive Testing: A Primer.* NJ.: Lawrence Erlbaum Associates, Inc., Publishers.

Yoke, M. O., Julian, S. W., & Iasonas, L. (2011). Exploration of the Validity of Gender Differencs in Mathematices Assessment Using Differential Bundle Functioning. *Internatinoal Journal of Testing*, *11*, 271-293.

ภาคผนวก

ภาคผนวก ก
หนังสือขอความอนุเคราะห์ขอข้อมูลเพื่อการวิจัยและแบบรายงานผล การพิจารณาจริยธรรมการวิจัยในคน

ที่ ศธ ๖๖๒๘/๐๑๐๒

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

๒๙ เมษายน ๒๕๕๙

เรื่อง ขอความอนุเคราะห์ขอข้อมูลเพื่อการวิจัย

เรียน ผู้อำนวยการสำนักทดสอบทางการศึกษา

สิ่งที่ส่งมาด้วย ๑. เค้าโครงวิทยานิพนธ์ฉบับย่อ จำนวน ๑ ชุด
๒. แบบทดสอบความสามารถด้านภาษา ด้านการคิดคำนวณ และด้านเหตุผล ชั้นประถมศึกษาปีที่ ๓ ปีการศึกษา ๒๕๕๕ จำนวน ๑ ชุด
๓. แบบรายงานผลการพิจารณาจริยธรรมการวิจัยในคน จำนวน ๑ ชุด

ด้วย นางสุมาลี ถามั่งมี รหัสประจำตัว ๕๖๙๑๐๔๐๒ นิสิตหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาการวิจัยและสถิติทางวิทยาการปัญญา ได้รับอนุมัติให้ทำวิทยานิพนธ์เรื่อง "การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ระดับชั้นประถมศึกษาปีที่ ๓: ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN" ซึ่งอยู่ในความควบคุมดูแลของ ดร.ปิยะทิพย์ ประดุจพรม อาจารย์ที่ปรึกษาหลัก ในการนี้ ผู้วิจัยมีความประสงค์ขอความอนุเคราะห์ขอข้อสอบ NT (National Test) ระดับชั้นประถมศึกษาปีที่ ๓ ปีการศึกษา ๒๕๕๕ จำนวน ๓ ด้าน คือ ด้านภาษา ด้านการคิดคำนวณ และด้านเหตุผล พร้อมเฉลยและขอผลการตอบข้อสอบ NT (National Test) ระดับชั้นประถมศึกษาปีที่ ๓ ปีการศึกษา ๒๕๕๕ จำแนกเพศชาย และเพศหญิง จำนวน ๓ ด้าน คือ ด้านภาษา ด้านการคิดคำนวณ และด้านเหตุผล ของจังหวัดสระแก้ว

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

แบบรายงานผลการพิจารณาจริยธรรมการวิจัยในคน
วิทยาลัยวิทยาการวิจัยและวิทยาการปัญญา
มหาวิทยาลัยบูรพา

๑. ชื่อเรื่องวิทยานิพนธ์

ชื่อเรื่องวิทยานิพนธ์ (ภาษาไทย) การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ระดับชั้นประถมศึกษาปีที่ ๖: ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN

ชื่อเรื่องวิทยานิพนธ์ (ภาษาอังกฤษ) A COMPARISON OF THE DIFFERENTIAL ITEM FUNCTIONING FOR NATIONAL TEST ITEM AT THE GRADE 3 LEVEL USING HGLM, MIMIC AND BAYESIAN METHODS

๒. ชื่อนิสิต (นาย, นาง, นางสาว): สุมาลี ถามั่งมี

หลักสูตรวิทยาศาสตรมหาบัณฑิต (M.Sc.) สาขาวิชาการวิจัยและสถิติทางวิทยาการปัญญา

☑ ภาคปกติ ☐ ภาคพิเศษ

รหัสประจำตัว ๕๖๙๑๐๔๐๒ คณะ/วิทยาลัยวิทยาการวิจัยและวิทยาการปัญญา

๓. หน่วยงานที่สังกัด: วิทยาลัยวิทยาการวิจัยและวิทยาการปัญญา

๔. ผลการพิจารณาของคณะกรรมการพิจารณาจริยธรรมการวิจัยในคน:

คณะกรรมการพิจารณาจริยธรรมการวิจัยในคน ได้พิจารณารายละเอียดวิทยานิพนธ์เรื่องดังกล่าวข้างต้นแล้ว ในประเด็นที่เกี่ยวกับ

๑) การเคารพในศักดิ์ศรี และสิทธิของคนที่ใช้เป็นตัวอย่างการวิจัย

๒) วิธีการอย่างเหมาะสมในการได้รับความยินยอมจากกลุ่มตัวอย่างก่อนเข้าร่วมโครงการวิจัย (Informed consent) รวมทั้งการป้องกันสิทธิประโยชน์ และรักษาความลับกลุ่มตัวอย่างในการวิจัย

๓) การดำเนินการวิจัยอย่างเหมาะสม เพื่อไม่ก่อความเสียหายต่อสิ่งที่ศึกษาวิจัย ไม่ว่าจะเป็นสิ่งที่มีชีวิตหรือไม่มีชีวิต

คณะกรรมการพิจารณาจริยธรรมการวิจัยในคน มีมติเห็นชอบ ดังนี้

(√) รับรองโครงการวิจัย

( ) ไม่รับรอง

๕. วันที่ที่ให้การรับรอง: ๒๗ เดือน เมษายน พ.ศ. ๒๕๕๙

ลงนาม.....

(ผู้ช่วยศาสตราจารย์ ดร.สุชาดา กรเพชรปาณี)

ประธานกรรมการพิจารณาจริยธรรมการวิจัยในคน

คณบดีวิทยาลัยวิทยาการวิจัยและวิทยาการปัญญา

วันที่ ๒๗ เมษายน พ.ศ. ๒๕๕๙

ภาคผนวก ข
ตัวอย่าง Print Out ผลการวิเคราะห์คุณภาพของข้อสอบ NT

IRT Item Parameter
Calibration Report

User Test 1

Report created on 11/9/2016

Xcalibre 4.2.2.0: IRT Item Parameter Estimation Software

ASC

Introduction

This report provides the results of the IRT item parameter calibration by the computer program Xcalibre Version 4.2.2.0 (Assessment Systems Corporation, 2014) 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\Aj.Sumalee-(NT)11 SEP 2016\Reasoning\LrtpXcal_Data.txt
C:\Users\Admin\Desktop\Aj.Sumalee-(NT)11 SEP 2016\Reasoning\LrtpXcal_ICF.txt
The Windows paths for the output files produced by this analysis were:
C:\Users\Admin\Desktop\Aj.Sumalee-(NT)11 SEP 2016\Reasoning\Reason-11 Sep 2016.rtf
C:\Users\Admin\Desktop\Aj.Sumalee-(NT)11 SEP 2016\Reasoning\Reason-11 Sep 2016.csv
C:\Users\Admin\Desktop\Aj.Sumalee-(NT)11 SEP 2016\Reasoning\Reason-11 Sep 2016 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	2000	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	-
Save item parameters	No	Item parameter format	N/A
Save data matrix	No	Omit codes are	N/A
Not Admin codes are	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	Merge empty poly categories	N/A

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)	1.000 (0.250)
b parameter prior mean (sd)	0.000 (1.000)	c parameter prior mean (sd)	0.250 (0.025)
Theta estimation method	EAP	Bayesian prior mean (sd)	0.000 (1.000)
Maximum E-M loops	60	Convergence criterion	0.001
Quadrature points	40	Center dich item parameters on	theta
Acceptable P range	0.00 to 1.00	Acceptable item-corr range	-1.00 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 statistics output	Yes	Delimiter	Comma

Specification	Value	Specification	Value
Save scores output	Yes	Delimiter	Comma
Save test information output	Yes	Delimiter	Comma
Save item information 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.

Item 7 failed to converge on this loop
Item 17 failed to converge on this loop
Item 27 failed to converge on this loop
Maximum change after Loop 1 was 4.0000 for Item 17 for the b parameter
Maximum change after Loop 2 was -2.1698 for Item 17 for the a parameter
Item 17 failed to converge on this loop
Maximum change after Loop 3 was 0.8302 for Item 17 for the a parameter
Maximum change after Loop 4 was 0.7978 for Item 17 for the b parameter
Maximum change after Loop 5 was -0.4198 for Item 17 for the b parameter
Maximum change after Loop 6 was 0.1432 for Item 17 for the b parameter
Maximum change after Loop 7 was 0.0467 for Item 27 for the a parameter

Maximum change after Loop 8 was -0.0082 for Item 5 for the b parameter
Maximum change after Loop 9 was -0.0073 for Item 5 for the b parameter
Maximum change after Loop 10 was -0.0063 for Item 5 for the b parameter
Maximum change after Loop 11 was -0.0054 for Item 5 for the b parameter
Maximum change after Loop 12 was -0.0046 for Item 5 for the b parameter
Maximum change after Loop 13 was -0.0038 for Item 5 for the b parameter
Maximum change after Loop 14 was -0.0032 for Item 5 for the b parameter
Maximum change after Loop 15 was -0.0026 for Item 5 for the b parameter
Maximum change after Loop 16 was -0.0022 for Item 1 for the a parameter
Maximum change after Loop 17 was -0.0021 for Item 1 for the a parameter
Maximum change after Loop 18 was -0.0020 for Item 1 for the a parameter
Maximum change after Loop 19 was -0.0019 for Item 1 for the a parameter
Maximum change after Loop 20 was -0.0018 for Item 1 for the a parameter
Maximum change after Loop 21 was -0.0017 for Item 1 for the a parameter
Maximum change after Loop 22 was -0.0013 for Item 1 for the a parameter
Maximum change after Loop 23 was -0.0009 for Item 1 for the a 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	0.823	0.243	0.364	1.550
b	30	1.392	1.195	- 0.537	3.993
c	30	0.254	0.026	0.217	0.348

Table 5: Summary Statistics for the Total Scores

Test	Items	Alpha	Mean	SD	Skew	Min	Q1	Median	Q3	Max	IQR
Full Test	30	0.714	13.659	4.739	0.152	3	10.00	14.0	17.00	27	7.00

Table 6: Summary Statistics for the Theta Estimates

Test	Examinees	Mean	SD	Skew	Min	Q1	Median	Q3	Max	IQR
Full Test	2000	0.034	0.914	0.238	-1.662	-0.767	0.003	0.759	2.772	1.525

Table 7: Overall Model Fit

Test	Items	Chi-square	df	p	-2LL
Full Test	30	1108.315	360	0.000	70886

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	4	2	4	1	Y	M	Included
2	1	1	4	1	Y	M	Included
3	1	4	4	1	Y	M	Included
4	1	1	4	1	Y	M	Included
5	1	3	4	1	Y	M	Included
6	4	3	4	1	Y	M	Included
7	3	2	4	1	Y	M	Included
8	2	2	4	1	Y	M	Included
9	4	4	4	1	Y	M	Included
10	4	4	4	1	Y	M	Included
11	4	4	4	1	Y	M	Included
12	4	4	4	1	Y	M	Included
13	1	4	4	1	Y	M	Included
14	4	2	4	1	Y	M	Included
15	1	1	4	1	Y	M	Included
16	1	1	4	1	Y	M	Included
17	3	2	4	1	Y	M	Included
18	2	2	4	1	Y	M	Included
19	1	1	4	1	Y	M	Included
20	3	4	4	1	Y	M	Included
21	4	4	4	1	Y	M	Included
22	4	4	4	1	Y	M	Included
23	2	2	4	1	Y	M	Included
24	2	2	4	1	Y	M	Included
25	2	2	4	1	Y	M	Included
26	3	3	4	1	Y	M	Included
27	4	3	4	1	Y	M	Included
28	2	2	4	1	Y	M	Included

Seq.	Item ID	Key	Options	Domain	Inclusion	Item Type	Status
29	2	2	4	1	Y	M	Included
30	3	3	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	4	0.385	0.242	0.670	1.616	0.233	
2	1	0.731	0.394	0.999	- 0.537	0.247	
3	1	0.593	0.336	0.902	0.528	0.348	
4	1	0.596	0.381	0.890	0.155	0.241	
5	1	0.414	0.079	0.364	2.750	0.284	
6	4	0.376	0.246	0.732	1.568	0.231	
7	3	0.304	0.071	0.944	2.896	0.277	
8	2	0.392	0.185	0.673	1.877	0.272	
9	4	0.459	0.205	0.486	1.369	0.249	
10	4	0.371	0.217	0.821	1.686	0.252	
11	4	0.553	0.380	0.953	0.333	0.230	
12	4	0.607	0.402	0.952	0.098	0.243	
13	1	0.331	0.123	0.903	2.441	0.280	
14	4	0.299	0.216	1.550	1.827	0.238	
15	1	0.440	0.178	0.429	1.666	0.247	
16	1	0.367	0.087	0.520	2.775	0.284	
17	3	0.261	- 0.094	1.164	3.993	0.258	K, Hb
18	2	0.482	0.295	0.770	0.784	0.232	
19	1	0.538	0.381	0.846	0.374	0.218	
20	3	0.248	0.040	1.111	3.174	0.239	K, Hb
21	4	0.342	0.155	0.887	2.112	0.269	

Seq.	Item ID	P	R	a	b	c	Flag(s)
22	4	0.349	0.259	0.740	1.665	0.217	
23	2	0.648	0.391	0.966	- 0.073	0.258	
24	2	0.560	0.337	0.855	0.420	0.259	
25	2	0.512	0.264	0.640	0.768	0.251	
26	3	0.490	0.287	0.793	0.824	0.252	
27	4	0.276	- 0.010	1.119	3.674	0.269	K, Hb
28	2	0.513	0.230	0.524	0.890	0.256	
29	2	0.636	0.365	0.757	- 0.123	0.228	
30	3	0.585	0.330	0.745	0.241	0.245	

ภาคผนวก ค
ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT
ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN

ตารางที่ ค-1 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ด้วยวิธี HGLM

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี HGLM		
	ความสามารถด้านภาษา (*p*-value)	ความสามารถด้านคำนวณ (*p*-value)	ความสามารถด้านเหตุผล (*p*-value)
1	0.168 (NO DIF)	0.273 (NO DIF)	0.150 (NO DIF)
2	**0.000 (DIF)**	0.340 (NO DIF)	0.080 (NO DIF)
3	0.363 (NO DIF)	0.688 (NO DIF)	**0.024 (DIF)**
4	**0.001 (DIF)**	0.881 (NO DIF)	0.129 (NO DIF)
5	0.942 (NO DIF)	0.238 (NO DIF)	0.150 (NO DIF)
6	0.386 (NO DIF)	0.590 (NO DIF)	0.214 (NO DIF)
7	0.065 (NO DIF)	0.320 (NO DIF)	**0.010 (DIF)**
8	0.117 (NO DIF)	0.971 (NO DIF)	**0.047 (DIF)**
9	0.539 (NO DIF)	0.624 (NO DIF)	**0.021 (DIF)**
10	**0.007 (DIF)**	0.544 (NO DIF)	**0.020 (DIF)**
11	0.611 (NO DIF)	0.305 (NO DIF)	**0.012 (DIF)**
12	**0.009 (DIF)**	**0.037 (DIF)**	**0.001 (DIF)**
13	**0.026 (DIF)**	0.798 (NO DIF)	0.938 (NO DIF)
14	0.871 (NO DIF)	0.917 (NO DIF)	0.332 (NO DIF)
15	0.090 (NO DIF)	0.901 (NO DIF)	0.052 (NO DIF)
16	0.102 (NO DIF)	0.935 (NO DIF)	**0.043 (DIF)**
17	0.150 (NO DIF)	0.082 (NO DIF)	**0.013 (DIF)**
18	0.278 (NO DIF)	0.299 (NO DIF)	0.229 (NO DIF)
19	**0.019 (DIF)**	0.850 (NO DIF)	**0.001 (DIF)**
20	0.177 (NO DIF)	0.295 (NO DIF)	0.613 (NO DIF)
21	0.062 (NO DIF)	0.362 (NO DIF)	0.229 (NO DIF)
22	0.247 (NO DIF)	0.768 (NO DIF)	**0.029 (DIF)**
23	0.845 (NO DIF)	**0.004 (DIF)**	**0.000 (DIF)**
24	0.895 (NO DIF)	**0.047 (DIF)**	**0.000 (DIF)**
25	0.826 (NO DIF)	**0.010 (DIF)**	**0.004 (DIF)**
26	**0.023 (DIF)**	**0.003 (DIF)**	0.313 (NO DIF)
27	**0.028 (DIF)**	0.867 (NO DIF)	0.073 (NO DIF)
28	**0.024 (DIF)**	0.660 (NO DIF)	**0.029 (DIF)**
29	0.854 (NO DIF)	0.672 (NO DIF)	**0.004 (DIF)**
30	0.396 NO DIF)	0.999 (NO DIF)	**0.028 (DIF)**
จำนวนข้อที่พบ DIF	9 ข้อ (30.00%)	5 ข้อ (16.67%)	17 ข้อ (56.67%)

ตารางที่ ค-2 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ด้วยวิธี MIMIC

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี MIMIC		
	ความสามารถด้านภาษา (*p*-value)	ความสามารถด้านคำนวณ (*p*-value)	ความสามารถด้านเหตุผล (*p*-value)
1	0.000 (NO DIF)	0.000 (NO DIF)	0.000 (NO DIF)
2	**0.026 (DIF)**	0.000 (NO DIF)	0.000 (NO DIF)
3	0.000 (NO DIF)	0.000 (NO DIF)	0.000 (NO DIF)
4	0.000 (NO DIF)	0.000 (NO DIF)	0.000 (NO DIF)
5	0.000 (NO DIF)	**0.007 (DIF)**	0.000 (NO DIF)
6	0.000 (NO DIF)	0.000 (NO DIF)	**0.029 (DIF)**
7	0.000 (NO DIF)	0.000 (NO DIF)	0.000 (NO DIF)
8	0.000 (NO DIF)	0.000 (NO DIF)	0.000 (NO DIF)
9	0.000 (NO DIF)	0.000 (NO DIF)	**0.002 (DIF)**
10	0.000 (NO DIF)	0.000 (NO DIF)	**0.027 (DIF)**
11	0.000 (NO DIF)	**0.028 (DIF)**	0.000 (NO DIF)
12	0.000 (NO DIF)	**0.031(DIF)**	0.000 (NO DIF)
13	0.000 (NO DIF)	0.000 (NO DIF)	**0.031 (DIF)**
14	0.000 (NO DIF)	0.000 (NO DIF)	**0.006 (DIF)**
15	0.000 (NO DIF)	0.000 (NO DIF)	**0.001 (DIF)**
16	0.000 (NO DIF)	0.000 (NO DIF)	**0.028 (DIF)**
17	0.000 (NO DIF)	0.000 (NO DIF)	**0.010 (DIF)**
18	0.000 (NO DIF)	**0.006 (DIF)**	0.000 (NO DIF)
19	0.000 (NO DIF)	0.000 (NO DIF)	0.000 (NO DIF)
20	0.000 (NO DIF)	0.000 (NO DIF)	0.000 (NO DIF)
21	0.000 (NO DIF)	0.000 (NO DIF)	0.000 (NO DIF)
22	0.000 (NO DIF)	0.000 (NO DIF)	**0.014 (DIF)**
23	0.000 (NO DIF)	**0.008 (DIF)**	0.000 (NO DIF)
24	0.000 (NO DIF)	**0.025 (DIF)**	0.000 (NO DIF)
25	0.000 (NO DIF)	**0.028 (DIF)**	0.000 (NO DIF)
26	0.000 (NO DIF)	**0.031 (DIF)**	0.000 (NO DIF)
27	0.000 (NO DIF)	0.000 (NO DIF)	0.000 (NO DIF)
28	0.000 (NO DIF)	0.000 (NO DIF)	**0.003 (DIF)**
29	0.000 (NO DIF)	0.000 (NO DIF)	0.000 (NO DIF)
30	0.000 (NO DIF)	0.000 (NO DIF)	**0.006 (DIF)**
จำนวนข้อที่พบ DIF	1 ข้อ (30.00%)	8 ข้อ (16.67%)	11 ข้อ (56.67%)

ตารางที่ ค-3 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ด้วยวิธี WinBUGS

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี BAYESIAN					
	ความสามารถด้านภาษา		ความสามารถด้านคำนวณ		ความสามารถด้านเหตุผล	
	Val 2.5pc	Val 97.5pc	Val 2.5pc	Val 97.5pc	Val 2.5pc	Val 97.5pc
1	-0.575 (NO DIF)	0.305 (NO DIF)	-0.71 (NO DIF)	0.155 (NO DIF)	-0.523 (NO DIF)	0.166 (NO DIF)
2	**0.988 (DIF)**	**2.058 (DIF)**	**0.147 (DIF)**	**1.040 (DIF)**	-0.829 (NO DIF)	0.011 (NO DIF)
3	-0.319 (NO DIF)	0.682 (NO DIF)	-0.319 (NO DIF)	0.599 (NO DIF)	-0.034 (NO DIF)	0.852 (NO DIF)
4	-0.299 (NO DIF)	0.929 (NO DIF)	-2.259 (NO DIF)	0.710 (NO DIF)	0.718 **(DIF)**	2.069 **(DIF)**
5	-0.523 (NO DIF)	0.166 (NO DIF)	-0.641 (NO DIF)	0.335 (NO DIF)	-0.339 (NO DIF)	0.609 (NO DIF)
6	-0.549 (NO DIF)	0.584 (NO DIF)	-0.417 (NO DIF)	0.377 (NO DIF)	-0.857 (NO DIF)	0.038 (NO DIF)
7	-0.829 (NO DIF)	0.011 (NO DIF)	-0.545 (NO DIF)	0.312 (NO DIF)	-0.271 (NO DIF)	0.606 (NO DIF)
8	-0.376 (NO DIF)	0.432 (NO DIF)	-0.271 (NO DIF)	0.606 (NO DIF)	**0.118 (DIF)**	**1.530 (DIF)**
9	-0.285 (NO DIF)	0.681 (NO DIF)	-0.400 (NO DIF)	0.346 (NO DIF)	-1.050 (NO DIF)	0.217 (NO DIF)
10	**-0.777 (DIF)**	**-0.016 (DIF)**	-0.463 (NO DIF)	0.425 (NO DIF)	-0.688 (NO DIF)	0.146 (NO DIF)
11	-0.438 (NO DIF)	0.529 (NO DIF)	-0.594 (NO DIF)	0.254 (NO DIF)	-0.382 (NO DIF)	0.738 (NO DIF)
12	-0.229 (NO DIF)	0.953 (NO DIF)	-0.849 (NO DIF)	0.112 (NO DIF)	**0.741 (DIF)**	**2.088 (DIF)**
13	-0.688 (NO DIF)	0.146 (NO DIF)	-0.205 (NO DIF)	0.634 (NO DIF)	**-2.239 (DIF)**	**-1.294 (DIF)**
14	**-1.030 (DIF)**	**-0.143 (DIF)**	-0.300 (NO DIF)	0.587 (NO DIF)	-0.180 (NO DIF)	0.829 (NO DIF)
15	**-1.190 (DIF)**	**-0.350 (DIF)**	-0.236 (NO DIF)	0.625 (NO DIF)	**-1.741 (DIF)**	**-0.712 (DIF)**
16	**-0.942 (DIF)**	**-0.158 (DIF)**	-0.326 (NO DIF)	0.711 (NO DIF)	-0.184 (NO DIF)	1.071 (NO DIF)
17	**-0.889 (DIF)**	**-0.116 (DIF)**	-0.821 (NO DIF)	0.001 (NO DIF)	**-1.085 (DIF)**	**-0.005 (DIF)**
18	-0.763 (NO DIF)	0.214 (NO DIF)	-0.599 (NO DIF)	0.394 (NO DIF)	-0.917 (NO DIF)	0.244 (NO DIF)
19	-0.184 (NO DIF)	1.071 (NO DIF)	-0.330 (NO DIF)	0.722 (NO DIF)	**0.769 (DIF)**	**1.986 (DIF)**
20	-0.180 (NO DIF)	0.829 (NO DIF)	**0.045 (DIF)**	**1.158 (DIF)**	-0.812 (NO DIF)	0.228 (NO DIF)
21	-0.376 (NO DIF)	0.649 (NO DIF)	0.024 **(DIF)**	**1.052 (DIF)**	**-1.313 (DIF)**	**-0.143 (DIF)**
22	-0.444 (NO DIF)	0.493 (NO DIF)	-0.167 (NO DIF)	0.729 (NO DIF)	-0.092 (NO DIF)	0.971 (NO DIF)
23	-0.148 (NO DIF)	1.098 (NO DIF)	**-1.052 (DIF)**	-0.302 **(DIF)**	-0.289 (NO DIF)	0.592 (NO DIF)
24	-0.289 (NO DIF)	0.592 (NO DIF)	-0.641 (NO DIF)	0.323 (NO DIF)	-0.096 (NO DIF)	1.058 (NO DIF)
25	-0.337 (NO DIF)	0.866 (NO DIF)	**-1.011 (DIF)**	**-0.142 (DIF)**	**0.355 (DIF)**	**1.737 (DIF)**
26	-0.674 (NO DIF)	0.128 (NO DIF)	**-1.201 (DIF)**	**-0.346 (DIF)**	**-1.742 (DIF)**	**-0.810 (DIF)**
27	-0.465 (NO DIF)	0.452 (NO DIF)	-0.188 (NO DIF)	0.688 (NO DIF)	-0.116 (NO DIF)	0.775 (NO DIF)
28	-0.806 (NO DIF)	0.161 (NO DIF)	-0.136 (NO DIF)	0.836 (NO DIF)	-0.746 (NO DIF)	0.418 (NO DIF)
29	-0.178 (NO DIF)	0.924 (NO DIF)	-0.448 (NO DIF)	0.334 (NO DIF)	-0.167 (NO DIF)	0.729 (NO DIF)
30	**-0.778 (DIF)**	**-0.027 (DIF)**	-0.393 (NO DIF)	0.428 (NO DIF)	**-1.567 (DIF)**	**-0.585 (DIF)**
จำนวนข้อที่พบ DIF	7 ข้อ (23.33%)		6 ข้อ (20.00%)		11 ข้อ (36.67%)	

ภาคผนวก ง
ผลการตอบข้อสอบ NT ของนักเรียนชั้นประถมศึกษาปีที่ 3
ปีการศึกษา 2555 ทั้ง 3 ด้าน

ตารางที่ ง-1 แสดงข้อมูลดิบของผลการตอบข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา จำนวน 30 ข้อ จำนวน 2,000 คน

ID	SEX	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	รวม
1	1	1	0	0	0	1	0	0	0	0	1	1	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	9
2	1	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	1	7
3	1	1	0	0	1	0	1	1	0	0	1	1	0	1	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	1	11
4	1	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	5
5	1	0	0	0	0	0	1	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	1	8
6	1	1	0	0	0	0	0	1	0	0	0	0	0	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	7
7	1	0	0	0	1	0	0	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	1	9
8	1	0	0	0	0	1	0	1	0	0	1	1	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	9
9	1	0	0	0	1	0	0	0	0	0	1	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	8
10	1	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	1	1	1	0	0	0	0	0	1	0	0	0	0	1	0	8
11	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	1	5
12	1	0	0	0	0	0	0	1	0	0	1	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	9
13	1	0	0	0	0	1	0	0	1	0	1	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	7
14	1	0	0	0	0	1	0	1	0	0	1	0	0	1	1	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	9
15	1	1	0	0	0	1	0	1	0	0	1	0	0	1	1	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	1	11
16	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	1	0	0	1	0	7
17	1	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	5
18	1	1	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0	1	0	0	1	0	1	0	0	0	1	0	0	0	1	8
19	1	0	0	1	1	0	0	1	0	1	1	0	0	0	1	0	0	1	0	0	0	0	1	1	0	1	0	0	1	0	1	12

ตารางที่ ง-1 (ต่อ)

ID	SEX	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	รวม
20	1	0	0	0	0	0	1	1	0	1	0	0	0	0	1	0	1	0	1	0	1	0	0	0	0	1	0	0	0	0	1	9
21	1	0	0	0	0	1	0	1	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	1	0	1	1	0	1	0	1	10
22	1	1	0	0	0	1	0	0	1	1	1	0	0	0	0	0	1	0	1	0	0	1	0	0	1	1	0	1	1	0	1	13
23	1	1	0	0	0	0	0	0	1	0	1	0	1	0	0	1	1	1	0	0	0	0	1	0	0	1	0	0	0	1	0	10
24	1	1	0	0	0	0	1	0	0	0	1	0	0	1	0	0	1	1	0	0	0	0	1	0	0	1	1	0	0	0	1	10
25	1	1	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	1	1	0	0	1	1	0	0	1	11
26	1	0	0	0	0	1	0	1	1	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1	8
27	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0	1	1	0	0	6
28	1	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	6
29	1	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	1	1	0	0	1	0	1	0	0	0	0	0	0	0	1	8
30	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	1	0	5
31	1	1	0	0	0	0	0	1	0	0	1	0	0	1	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	1	8
32	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	1	0	0	0	1	0	1	0	1	0	1	0	1	9
33	1	0	0	0	0	1	0	1	1	0	0	0	0	1	1	1	1	1	0	0	0	1	0	0	1	0	1	0	0	0	1	12
34	1	1	0	0	0	0	0	1	1	0	1	1	0	1	0	1	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	10
35	1	0	0	0	0	1	0	0	0	0	0	1	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	6
36	1	0	0	0	0	1	0	1	1	0	0	0	0	1	1	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	9
37	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	6
38	1	0	0	0	0	1	0	1	1	0	1	0	0	1	1	1	1	1	0	0	0	0	0	1	0	0	0	0	0	1	11

ตารางที่ ง-1 (ต่อ)

ID	SEX	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	รวม
39	1	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	6
40	1	1	0	0	0	0	1	1	0	0	0	0	0	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	9
41	1	1	0	0	0	0	0	0	1	0	0	0	0	1	0	1	1	0	0	0	0	0	1	1	0	0	0	1	0	0	1	9
42	1	0	0	0	0	1	0	1	1	0	0	0	0	1	1	1	0	1	0	0	0	0	1	0	0	0	1	0	0	0	1	10
43	1	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	5
44	1	1	0	0	0	1	0	0	0	0	1	0	1	0	1	1	1	0	0	1	0	0	1	0	1	0	0	0	1	0	1	12
45	1	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	6
46	1	0	0	0	0	1	0	1	0	0	0	0	0	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	8
47	1	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0	0	0	1	0	0	1	0	0	0	1	0	0	0	7
48	1	0	0	0	0	1	0	1	0	0	1	0	0	1	1	1	1	1	0	0	0	0	0	0	1	0	1	1	1	0	1	13
49	1	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	1	7
50	1	1	0	0	0	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	8
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
2000	1	1	0	0	0	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	5	
รวม		120	203	121	804	274	631	362	356	1095	311	129	1337	1025	1054	1204	1062	381	166	265	309	672	171	377	289	691	336	361	197	1135		

ตารางที่ ง-2 แสดงข้อมูลดิบของผลการตอบข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านคำนวณ จำนวน 30 ข้อ จำนวน 2,000 คน

ID	SEX	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	รวม
1	1	1	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	7
2	1	0	0	0	0	0	0	1	1	0	1	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	1	1	0	1	1	10
3	1	1	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	7
4	1	0	0	1	1	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	1	0	0	1	0	8
5	1	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	6
6	1	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	1	1	0	1	0	9
7	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	1	0	6
8	1	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	1	0	7
9	1	0	0	1	0	0	0	1	0	0	1	0	0	0	1	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	1	10
10	1	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	1	0	7
11	1	1	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	1	0	8
12	1	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	1	1	0	8
13	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	1	0	1	1	0	0	1	0	0	1	1	9
14	1	0	0	1	0	0	1	0	0	0	1	0	0	0	1	1	0	0	1	0	0	0	0	1	0	0	1	0	0	0	1	9
15	1	0	0	1	0	0	0	1	0	0	1	0	0	0	1	1	0	1	1	0	0	0	0	1	0	0	1	1	0	1	1	12
16	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	3
17	1	0	0	0	1	0	0	0	1	0	0	0	0	0	1	1	1	0	0	0	0	1	1	0	1	0	0	0	0	1	1	10
18	1	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	5
19	1	0	0	0	0	0	0	0	1	1	1	0	0	0	1	1	0	1	0	0	1	0	0	0	0	1	0	0	0	0	1	9

ID	SEX	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	รวม
20	1	0	0	1	1	0	0	0	1	0	0	0	0	0	1	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	1	9
21	1	0	0	0	1	0	0	1	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	6
22	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	1	8
23	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	3
24	1	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	4
25	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
26	1	1	0	1	0	0	0	0	1	0	0	1	0	0	1	0	0	1	0	1	0	0	1	1	0	1	0	0	0	1	1	12
27	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	4
28	1	0	0	0	1	0	0	1	0	0	0	0	0	1	1	0	0	1	0	0	0	0	1	0	0	0	1	0	1	0	0	8
29	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	4
30	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	3
31	1	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	1	0	0	0	1	1	1	0	0	1	0	0	1	1	10
32	1	0	0	1	0	0	0	1	1	1	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	1	1	0	0	0	0	9
33	1	0	0	1	0	0	0	1	1	1	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	1	1	0	0	0	0	9
34	1	1	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	1	8
35	1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	1	0	1	0	0	0	0	1	0	1	1	0	0	1	0	9
36	1	0	0	1	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	6
37	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	4
38	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	5

ID	SEX	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	รวม
39	1	1	0	1	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	6
40	1	0	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	1	7
41	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	3
42	1	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	1	1	0	0	1	1	9
43	1	0	0	0	0	0	0	1	0	0	1	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	6
44	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	4
45	1	1	0	1	0	0	0	0	0	1	0	0	0	1	1	0	1	1	0	0	0	0	1	0	0	0	1	0	1	0	0	10
46	1	1	0	1	0	0	0	0	0	0	1	1	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	8
47	1	1	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	6
48	1	0	1	0	0	0	1	0	0	1	0	0	0	0	1	0	1	1	0	0	1	0	0	1	0	0	0	1	0	1	0	10
49	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	4
50	1	0	0	0	1	0	0	0	1	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	7
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
2000	1	1	0	0	0	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	7
รวม		152	659	297	285	411	679	542	503	645	424	311	537	596	414	427	989	326	328	204	321	401	1267	330	481	965	402	309	718	631		

ตารางที่ ง-3 แสดงข้อมูลดิบของผลการตอบข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านเหตุผล จำนวน 30 ข้อ จำนวน 2,000 คน

ID	SEX	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	รวม
1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	1	5
2	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1	1	1	0	0	1	0	0	0	0	0	1	1	0	0	8
3	1	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	1	7
4	1	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	1	0	0	1	1	0	1	0	0	0	0	0	0	0	1	10
5	1	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	4
6	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	1	6
7	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1	1	1	0	0	1	0	0	0	0	0	1	1	0	0	8
8	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	4
9	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	1	5
10	1	1	0	0	0	1	1	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	6
11	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	3
12	1	0	0	0	0	1	1	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	1	8
13	1	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	1	7
14	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	1	0	1	0	1	7
15	1	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	1	7
16	1	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	6
17	1	0	0	0	0	1	0	1	0	1	0	0	1	0	0	0	0	0	1	0	0	1	0	0	1	1	0	1	0	0	0	9
18	1	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	5
19	1	0	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0	1	9
20	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
21	1	1	0	0	0	0	1	0	0	0	1	0	1	1	1	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	8
22	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	5
23	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	4

ID	SEX	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	รวม
24	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	5
25	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	6
26	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	1	6
27	1	0	1	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0	1	0	1	0	0	0	0	1	0	1	0	0	0	8
28	1	1	0	0	0	0	1	0	1	0	1	0	1	1	1	0	0	0	1	0	1	0	1	0	0	0	1	0	0	0	0	11
29	1	1	0	0	0	0	1	0	0	0	0	1	0	1	0	1	0	0	1	0	0	1	0	1	0	1	0	0	1	0	0	10
30	1	0	0	0	0	0	1	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	0	1	0	0	0	1	1	0	1	9
31	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	1	1	7
32	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	1	6
33	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	1	6
34	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	1	0	1	0	1	0	0	0	1	0	0	7
35	1	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	5
36	1	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	5
37	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	1	6
38	1	0	0	1	0	1	0	1	0	0	0	1	1	0	1	0	0	0	0	0	1	0	1	0	1	0	0	1	0	0	0	10
39	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	1	6
40	1	1	0	0	0	1	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	8
41	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	1	5
42	1	1	0	0	1	1	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	1	0	0	9
43	1	1	0	0	0	1	0	0	0	0	0	0	1	1	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	1	10
44	1	1	0	0	0	1	0	0	0	0	0	0	0	1	1	0	1	1	0	1	0	0	0	0	0	0	1	0	0	0	1	9
45	1	0	0	0	1	0	0	0	1	0	1	0	0	1	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	1	0	7
46	1	0	0	1	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	1	1	0	1	0	1	0	0	1	0	1	1	11
47	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	1	0	0	0	1	7
48	1	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	1	0	1	0	1	0	0	0	1	0	0	0	1	0	0	9
48	1	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	1	0	1	0	1	0	0	0	1	0	0	0	1	0	0	9

ID	SEX	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	รวม
49	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0	7
50	1	1	0	0	0	1	0	0	0	1	0	1	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	7
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
2000	1	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	1	7
รวม		492	144	374	270	828	752	476	204	234	392	295	285	699	637	280	332	313	449	282	614	637	532	245	313	279	981	551	320	239	1169	

ภาคผนวก จ

ตัวอย่าง Print Out ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้วยวิธี HGLM โดยใช้โปรแกรม 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: 2 February 2018, Friday
Time: 8:40:20

SPECIFICATIONS FOR THIS NONLINEAR HLM2 RUN
Problem Title: no title
The data source for this run = literacy
The command file for this run = D:\P_kung\28-1-61\Literacy_HLM\liter.hlm
Output file name = D:\P_kung\28-1-61\Literacy_HLM\Outliteracy.txt
The maximum number of level-1 units = 60000
The maximum number of level-2 units = 2000
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
	GENDER, G01
# ITEM1 slope, B1	INTRCPT2, G10
	GENDER, G11
# ITEM2 slope, B2	INTRCPT2, G20
	GENDER, G21
# ITEM3 slope, B3	INTRCPT2, G30
	GENDER, G31
# ITEM4 slope, B4	INTRCPT2, G40
	GENDER, G41
# ITEM5 slope, B5	INTRCPT2, G50
	GENDER, G51
# ITEM6 slope, B6	INTRCPT2, G60
	GENDER, G61
# ITEM7 slope, B7	INTRCPT2, G70
	GENDER, G71
# ITEM8 slope, B8	INTRCPT2, G80
	GENDER, G81
# ITEM9 slope, B9	INTRCPT2, G90
	GENDER, G91
# ITEM10 slope, B10	INTRCPT2, G100
	GENDER, G101
# ITEM11 slope, B11	INTRCPT2, G110
	GENDER, G111
# ITEM12 slope, B12	INTRCPT2, G120
	GENDER, G121
# ITEM13 slope, B13	INTRCPT2, G130
	GENDER, G131
# ITEM14 slope, B14	INTRCPT2, G140

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

'#' - The residual parameter variance for this level-1 coefficient has been set
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*(GENDER) + U0
B1 = G10 + G11*(GENDER)
B2 = G20 + G21*(GENDER)
B3 = G30 + G31*(GENDER)
B4 = G40 + G41*(GENDER)
B5 = G50 + G51*(GENDER)
B6 = G60 + G61*(GENDER)
B7 = G70 + G71*(GENDER)
B8 = G80 + G81*(GENDER)
B9 = G90 + G91*(GENDER)
B10 = G100 + G101*(GENDER)
B11 = G110 + G111*(GENDER)
B12 = G120 + G121*(GENDER)
B13 = G130 + G131*(GENDER)
B14 = G140 + G141*(GENDER)
B15 = G150 + G151*(GENDER)
B16 = G160 + G161*(GENDER)
B17 = G170 + G171*(GENDER)

B18 = G180 + G181*(GENDER)
B19 = G190 + G191*(GENDER)
B20 = G200 + G201*(GENDER)
B21 = G210 + G211*(GENDER)
B22 = G220 + G221*(GENDER)
B23 = G230 + G231*(GENDER)
B24 = G240 + G241*(GENDER)
B25 = G250 + G251*(GENDER)
B26 = G260 + G261*(GENDER)
B27 = G270 + G271*(GENDER)
B28 = G280 + G281*(GENDER)
B29 = G290 + G291*(GENDER)

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

The value of the likelihood function at iteration 2 = -3.405288E+004

RESULTS FOR NON-LINEAR MODEL WITH THE LOGIT LINK FUNCTION: Unit-Specific Model
(macro iteration 248)

Tau
INTRCPT1,B0 0.01470

Tau (as correlations)
INTRCPT1,B0 1.000

Random level-1 coefficient	Reliability estimate
INTRCPT1, B0	0.074

The value of the likelihood function at iteration 2 = -8.513362E+004

The outcome variable is RESPONSE

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

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For INTRCPT1, B0					
INTRCPT2, G00	-0.172471	0.063605	-2.712	1998	0.007
GENDER, G01	0.076372	0.089837	0.850	1998	0.396
For ITEM1 slope, B1					
INTRCPT2, G10	-0.416090	0.091588	-4.543	59940	0.000
GENDER, G11	0.176998	0.128503	1.377	59940	0.168
For ITEM2 slope, B2					
INTRCPT2, G20	-0.028252	0.089846	-0.314	59940	0.753
GENDER, G21	-0.825317	0.130595	-6.320	59940	0.000
For ITEM3 slope, B3					
INTRCPT2, G30	-0.910468	0.096557	-9.429	59940	0.000
GENDER, G31	-0.124585	0.136952	-0.910	59940	0.363
For ITEM4 slope, B4					
INTRCPT2, G40	-1.427752	0.105765	-13.499	59940	0.000
GENDER, G41	-0.557142	0.159159	-3.501	59940	0.001
For ITEM5 slope, B5					
INTRCPT2, G50	-0.873754	0.096070	-9.095	59940	0.000
GENDER, G51	-0.009850	0.135201	-0.073	59940	0.942
For ITEM6 slope, B6					
INTRCPT2, G60	-0.802192	0.095179	-8.428	59940	0.000
GENDER, G61	-0.116977	0.134879	-0.867	59940	0.386
For ITEM7 slope, B7					
INTRCPT2, G70	-0.703784	0.094073	-7.481	59940	0.000
GENDER, G71	0.241691	0.131075	1.844	59940	0.065
For ITEM8 slope, B8					
INTRCPT2, G80	-1.226693	0.101633	-12.070	59940	0.000
GENDER, G81	-0.228161	0.145833	-1.565	59940	0.117
For ITEM9 slope, B9					
INTRCPT2, G90	-1.304078	0.103135	-12.644	59940	0.000

GENDER, G91	-0.089647	0.145974	-0.614	59940	0.539
For ITEM10 slope, B10					
INTRCPT2, G100	-0.827478	0.095485	-8.666	59940	0.000
GENDER, G101	0.361061	0.132112	2.733	59940	0.007
For ITEM11 slope, B11					
INTRCPT2, G110	-1.449373	0.106255	-13.640	59940	0.000
GENDER, G111	-0.076380	0.150199	-0.509	59940	0.611
For ITEM12 slope, B12					
INTRCPT2, G120	-1.501080	0.107465	-13.968	59940	0.000
GENDER, G121	-0.414747	0.158646	- 2.614	59940	0.009
For ITEM13 slope, B13					
INTRCPT2, G130	-0.817328	0.095361	-8.571	59940	0.000
GENDER, G131	0.294225	0.132290	2.224	59940	0.026
For ITEM14 slope, B14					
INTRCPT2, G140	-0.665532	0.093680	-7.104	59940	0.000
For ITEM15 slope, B15					
INTRCPT2, G150	-1.337457	0.103816	-12.883	59940	0.000
GENDER, G151	0.241816	0.142776	1.694	59940	0.090
For ITEM16 slope, B16					
INTRCPT2, G160	-1.271385	0.102487	-12.405	59940	0.000
GENDER, G161	0.230901	0.141249	1.635	59940	0.102
For ITEM17 slope, B17					
INTRCPT2, G170	-0.490994	0.092132	-5.329	59940	0.000
GENDER, G171	0.185685	0.129085	1.438	59940	0.150
For ITEM18 slope, B18					
INTRCPT2, G180	-1.392392	0.104983	-13.263	59940	0.000
GENDER, G181	-0.162687	0.149787	-1.086	59940	0.278
For ITEM19 slope, B19					
INTRCPT2, G190	-1.463963	0.106591	-13.734	59940	0.000
GENDER, G191	-0.368095	0.156181	-2.357	59940	0.019
For ITEM20 slope, B20					
INTRCPT2, G200	-1.290921	0.102872	-12.549	59940	0.000
GENDER, G201	-0.198969	0.147246	-1.351	59940	0.177
For ITEM21 slope, B21					
INTRCPT2, G210	-1.081875	0.099105	-10.916	59940	0.000
GENDER, G211	-0.265889	0.142495	-1.866	59940	0.062

For ITEM22 slope, B22					
INTRCPT2, G220	-0.991485	0.097703	-10.148	59940	0.000
GENDER, G221	-0.160988	0.139028	-1.158	59940	0.247
For ITEM23 slope, B23					
INTRCPT2, G230	-0.355595	0.091199	-3.899	59940	0.000
GENDER, G231	-0.025267	0.128687	-0.196	59940	0.845
For ITEM24 slope, B24					
INTRCPT2, G240	-1.442132	0.106090	-13.593	59940	0.000
GENDER, G241	-0.019665	0.149078	-0.132	59940	0.895
For ITEM25 slope, B25					
INTRCPT2, G250	-1.297486	0.103003	-12.597	59940	0.000
GENDER, G251	0.031782	0.144150	0.220	59940	0.826
For ITEM26 slope, B26					
INTRCPT2, G260	-0.101297	0.090041	-1.125	59940	0.261
For ITEM27 slope, B27					
INTRCPT2, G270	-0.894653	0.096345	-9.286	59940	0.000
GENDER, G271	-0.304605	0.138632	-2.197	59940	0.028
For ITEM28 slope, B28					
INTRCPT2, G280	-1.471312	0.106762	-13.781	59940	0.000
GENDER, G281	-0.351755	0.156105	-2.253	59940	0.024
For ITEM29 slope, B29					
INTRCPT2, G290	-0.873754	0.096070	-9.095	59940	0.000
GENDER, G291	-0.025021	0.135331	-0.185	59940	0.854

Fixed Effect	Coefficient	Odds Ratio	Confidence Interval
For INTRCPT1, B0			
INTRCPT2, G00	-0.172471	0.841582	(0.743,0.953)
GENDER, G01	0.076372	1.079364	(0.905,1.287)
For ITEM1 slope, B1			
INTRCPT2, G10	-0.416090	0.659621	(0.551,0.789)
GENDER, G11	0.176998	1.193629	(0.928,1.536)
For ITEM2 slope, B2			

INTRCPT2, G20	-0.028252	0.972143	(0.815,1.159)
GENDER, G21	-0.825317	0.438096	(0.339,0.566)
For ITEM3 slope, B3			
INTRCPT2, G30	-0.910468	0.402336	(0.333,0.486)
GENDER, G31	-0.124585	0.882863	(0.675,1.155)
For ITEM4 slope, B4			
INTRCPT2, G40	-1.427752	0.239847	(0.195,0.295)
GENDER, G41	-0.557142	0.572844	(0.419,0.783)
For ITEM5 slope, B5			
INTRCPT2, G50	-0.873754	0.417382	(0.346,0.504)
GENDER, G51	-0.009850	0.990199	(0.760,1.291)
For ITEM6 slope, B6			
INTRCPT2, G60	-0.802192	0.448345	(0.372,0.540)
For ITEM7 slope, B7			
INTRCPT2, G70	-0.703784	0.494710	(0.411,0.595)
GENDER, G71	0.241691	1.273400	(0.985,1.646)
For ITEM8 slope, B8			
INTRCPT2, G80	-1.226693	0.293261	(0.240,0.358)
GENDER, G81	-0.228161	0.795996	(0.598,1.059)
For ITEM9 slope, B9			
INTRCPT2, G90	-1.304078	0.271423	(0.222,0.332)
GENDER, G91	-0.089647	0.914254	(0.687,1.217)
For ITEM10 slope, B10			
INTRCPT2, G100	-0.827478	0.437151	(0.363,0.527)
GENDER, G101	0.361061	1.434851	(1.108,1.859)
For ITEM11 slope, B11			
INTRCPT2, G110	-1.449373	0.234717	(0.191,0.289)
GENDER, G111	-0.076380	0.926464	(0.690,1.244)
For ITEM12 slope, B12			
INTRCPT2, G120	-1.501080	0.222889	(0.181,0.275)
GENDER, G121	-0.414747	0.660507	(0.484,0.901)
For ITEM13 slope, B13			
INTRCPT2, G130	-0.817328	0.441610	(0.366,0.532)
GENDER, G131	0.294225	1.342086	(1.036,1.739)
For ITEM14 slope, B14			
INTRCPT2, G140	-0.665532	0.514000	(0.428,0.618)

GENDER, G141	0.021419	1.021650	(0.789,1.323)
For ITEM15 slope, B15			
INTRCPT2, G150	-1.337457	0.262512	(0.214,0.322)
GENDER, G151	0.241816	1.273560	(0.963,1.685)
For ITEM16 slope, B16			
INTRCPT2, G160	-1.271385	0.280443	(0.229,0.343)
GENDER, G161	0.230901	1.259735	(0.955,1.662)
For ITEM17 slope, B17			
INTRCPT2, G170	-0.490994	0.612018	(0.511,0.733)
GENDER, G171	0.185685	1.204043	(0.935,1.551)
For ITEM18 slope, B18			
INTRCPT2, G180	-1.392392	0.248480	(0.202,0.305)
For ITEM19 slope, B19			
INTRCPT2, G190	-1.463963	0.231318	(0.188,0.285)
GENDER, G191	-0.368095	0.692052	(0.510,0.940)
For ITEM20 slope, B20			
INTRCPT2, G200	-1.290921	0.275017	(0.225,0.336)
GENDER, G201	-0.198969	0.819576	(0.614,1.094)
For ITEM21 slope, B21			
INTRCPT2, G210	-1.081875	0.338959	(0.279,0.412)
GENDER, G211	-0.265889	0.766524	(0.580,1.013)
For ITEM22 slope, B22			
INTRCPT2, G220	-0.991485	0.371025	(0.306,0.449)
GENDER, G221	-0.160988	0.851303	(0.648,1.118)
For ITEM23 slope, B23			
INTRCPT2, G230	-0.355595	0.700756	(0.586,0.838)
GENDER, G231	-0.025267	0.975050	(0.758,1.255)
For ITEM24 slope, B24			
INTRCPT2, G240	-1.442132	0.236423	(0.192,0.291)
GENDER, G241	-0.019665	0.980527	(0.732,1.313)
For ITEM25 slope, B25			
INTRCPT2, G250	-1.297486	0.273218	(0.223,0.334)
GENDER, G251	0.031782	1.032293	(0.778,1.369)
For ITEM26 slope, B26			
INTRCPT2, G260	-0.101297	0.903665	(0.757,1.078)
GENDER, G261	0.289487	1.335742	(1.041,1.713)

For ITEM27 slope, B27			
INTRCPT2, G270	-0.894653	0.408749	(0.338,0.494)
GENDER, G271	-0.304605	0.737414	(0.562,0.968)
For ITEM28 slope, B28			
INTRCPT2, G280	-1.471312	0.229624	(0.186,0.283)
GENDER, G281	-0.351755	0.703452	(0.518,0.955)
For ITEM29 slope, B29			
INTRCPT2, G290	-0.873754	0.417382	(0.346,0.504)
GENDER, G291	-0.025021	0.975289	(0.748,1.272)

ภาคผนวก ฉ
ตัวอย่าง Print Out ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้วยวิธี MIMIC โดยใช้โปรแกรม Mplus

Mplus VERSION 7
MUTHEN & MUTHEN
01/05/2018 8:57 PM
INPUT INSTRUCTIONS
TITLE: Run factor NUMERACY
DATA: FILE = Numeracy 2000.csv;
VARIABLE:
NAMES = ID SEX Item1 - Item30;
USEVARIABLES SEX Item1 - Item30;
CATEGORICAL ARE Item1 Item2- Item30;

Observed dependent variables
Binary and ordered categorical (ordinal)

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

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
FACTOR BY				
ITEM1	-0.223	0.038	-5.932	0.000
ITEM2	0.250	0.051	4.948	0.000
ITEM3	0.011	0.038	0.298	0.766
ITEM4	0.077	0.045	1.697	0.090
ITEM5	0.225	0.040	5.652	0.000
ITEM6	0.112	0.039	2.849	0.004
ITEM7	-0.443	0.038	-11.522	0.000
ITEM8	0.264	0.037	7.205	0.000
ITEM9	-0.113	0.042	-2.697	0.007
ITEM10	-0.488	0.039	-12.623	0.000
ITEM11	0.096	0.040	2.399	0.016
ITEM12	0.128	0.042	3.066	0.002
ITEM13	0.211	0.038	5.553	0.000
ITEM14	-0.413	0.039	-10.484	0.000

ITEM15	0.119	0.050	2.384	0.017
ITEM16	0.151	0.037	4.036	0.000
ITEM17	-0.261	0.037	-7.032	0.000
ITEM18	0.257	0.042	6.178	0.000
ITEM19	0.213	0.039	5.475	0.000
ITEM20	0.112	0.049	2.278	0.023
ITEM21	0.135	0.040	3.382	0.001
ITEM22	-0.135	0.045	-3.013	0.003
ITEM23	-0.445	0.036	-12.400	0.000
ITEM24	0.168	0.041	4.079	0.000
ITEM25	0.125	0.039	3.187	0.001
ITEM26	-0.353	0.037	-9.588	0.000
ITEM27	-0.033	0.043	-0.766	0.444
ITEM28	0.255	0.043	5.950	0.000
ITEM29	-0.510	0.036	-14.233	0.000
ITEM30	-0.427	0.041	-10.482	0.000
FACTOR ON				
SEX	-0.033	0.137	-0.243	0.808
ITEM1 ON				
SEX	0.085	0.063	1.356	0.175
ITEM2 ON				
SEX	-0.090	0.090	-1.005	0.315
ITEM3 ON				
SEX	-0.030	0.058	-0.517	0.605
ITEM4 ON				
SEX	-0.010	0.069	-0.150	0.881
ITEM5 ON				
SEX	0.118	0.076	1.561	0.119
ITEM6 ON				
SEX	0.064	0.065	0.976	0.329
ITEM7 ON				
SEX	0.054	0.079	0.678	0.498
ITEM8 ON				
SEX	0.003	0.068	0.041	0.967
ITEM9 ON				
SEX	0.031	0.062	0.492	0.622

ITEM10	ON				
	SEX	0.031	0.083	0.376	0.707
ITEM11	ON				
	SEX	0.093	0.065	1.440	0.150
ITEM12	ON				
	SEX	0.176	0.070	2.516	0.012
ITEM13	ON				
	SEX	-0.026	0.065	-0.404	0.686
ITEM14	ON				
	SEX	-0.008	0.077	-0.104	0.917
ITEM15	ON				
	SEX	-0.010	0.067	-0.152	0.879
ITEM16	ON				
	SEX	-0.005	0.066	-0.079	0.937
ITEM17	ON				
	SEX	0.129	0.065	1.990	0.047
ITEM18	ON				
	SEX	0.098	0.077	1.276	0.202
ITEM19	ON				
	SEX	0.031	0.072	0.438	0.662
ITEM20	ON				
	SEX	-0.097	0.077	-1.255	0.210
ITEM21	ON				
	SEX	-0.073	0.070	-1.055	0.292
ITEM22	ON				
	SEX	-0.044	0.065	-0.671	0.503
ITEM23	ON				
	SEX	0.217	0.082	2.637	0.008
ITEM24	ON				
	SEX	0.167	0.070	2.400	0.016
ITEM25	ON				
	SEX	0.220	0.063	3.472	0.001
ITEM26	ON				
	SEX	0.242	0.072	3.384	0.001
ITEM27	ON				
	SEX	-0.015	0.064	-0.240	0.810

ITEM28 ON				
SEX	-0.038	0.074	-0.509	0.611
ITEM29 ON				
SEX	0.026	0.083	0.311	0.756
Thresholds				
ITEM1$1	0.020	0.040	0.506	0.613
ITEM2$1	1.385	0.057	24.270	0.000
ITEM3$1	0.426	0.041	10.401	0.000
ITEM4$1	1.036	0.048	21.401	0.000
ITEM5$1	1.126	0.050	22.405	0.000
ITEM6$1	0.852	0.045	18.801	0.000
ITEM7$1	0.448	0.041	10.900	0.000
ITEM8$1	0.607	0.042	14.310	0.000
ITEM9$1	0.687	0.043	15.894	0.000
ITEM10$1	0.485	0.041	11.711	0.000
ITEM11$1	0.845	0.045	18.685	0.000
ITEM12$1	1.103	0.050	22.160	0.000
ITEM13$1	0.601	0.042	14.186	0.000
ITEM14$1	0.533	0.042	12.768	0.000
ITEM15$1	0.810	0.045	18.103	0.000
ITEM16$1	0.789	0.044	17.750	0.000
ITEM17$1	0.083	0.040	2.087	0.037
ITEM18$1	1.028	0.048	21.296	0.000
ITEM19$1	0.990	0.048	20.817	0.000
ITEM20$1	1.221	0.052	23268	0.000
ITEM21$1	0.954	0.047	20.327	0.000
ITEM22$1	0.820	0.045	18.277	0.000
ITEM23$1	-0.228	0.040	-5.688	0.000
ITEM24$1	1.058	0.049	21.658	0.000
ITEM25$1	0.817	0.045	18.218	0.000
ITEM26$1	0.171	0.040	4.299	0.000
ITEM27$1	0.831	0.045	18.453	0.000
ITEM28$1	0.994	0.048	20.871	0.000
ITEM29$1	0.383	0.041	9.399	0.000
ITEM30$1	0.487	0.041	11.774	0.000

Residual Variances

FACTOR	1.000	0.000	999.000	999.000

STANDARDIZED MODEL RESULTS

	StdYX Estimate	Std Estimate
FACTOR BY		
ITEM1	-0.222	-0.223
ITEM2	0.250	0.250
ITEM3	0.011	0.011
ITEM4	0.077	0.077
ITEM5	0.225	0.225
ITEM6	0.112	0.112
ITEM7	-0.443	-0.443
ITEM8	0.264	0.264
ITEM9	-0.113	-0.113
ITEM10	-0.487	-0.488
ITEM11	0.096	0.096
ITEM12	0.127	0.128
ITEM13	0.211	0.211
ITEM14	-0.413	-0.413
ITEM15	0.119	0.119
ITEM16	0.151	0.151
ITEM17	-0.260	-0.261
ITEM18	0.257	0.257
ITEM19	0.213	0.213
ITEM20	0.112	0.112
ITEM21	0.135	0.135
ITEM22	-0.135	-0.135
ITEM23	-0.442	-0.445
ITEM24	0.167	0.168
ITEM25	0.125	0.125
ITEM26	-0.350	-0.353
ITEM27	-0.033	-0.033
ITEM28	0.254	0.255
ITEM29	-0.509	-0.510

ITEM30	-0.428	-0.428
FACTOR ON		
SEX	-0.017	-0.033
ITEM1 ON		
SEX	0.043	0.085
ITEM2 ON		
SEX	-0.045	-0.090
ITEM3 ON		
SEX	-0.015	-0.030
ITEM4 ON		
SEX	-0.005	-0.010
ITEM5 ON		
SEX	0.059	0.118
ITEM6 ON		
SEX	0.032	0.064
ITEM7 ON		
SEX	0.027	0.054
ITEM8 ON		
SEX	0.001	0.003
ITEM9 ON		
SEX	0.015	0.031
ITEM10 ON		
SEX	0.016	0.031
ITEM11 ON		
SEX	0.046	0.093
ITEM12 ON		
SEX	0.088	0.176
ITEM13 ON		
SEX	-0.013	-0.026
ITEM14 ON		
SEX	-0.004	-0.008
ITEM15 ON		
SEX	-0.005	-0.010
ITEM16 ON		
SEX	-0.003	-0.005

ITEM17	ON		
	SEX	0.065	0.129
ITEM18	ON		
	SEX	0.049	0.098
ITEM19	ON		
	SEX	0.016	0.031
ITEM20	ON		
	SEX	-0.049	-0.097
ITEM21	ON		
	SEX	-0.037	-0.073
ITEM22	ON		
	SEX	-0.022	-0.044
ITEM23	ON		
	SEX	0.108	0.217
ITEM24	ON		
	SEX	0.083	0.167
ITEM25	ON		
	SEX	0.109	0.220
ITEM26	ON		
	SEX	0.120	0.242
ITEM27	ON		
	SEX	-0.008	-0.015
ITEM28	ON		
	SEX	-0.019	-0.038
ITEM29	ON		
	SEX	0.013	0.026
ITEM11	WITH		
	ITEM10	-0.312	-0.312
ITEM14	WITH		
	ITEM13	-0.182	-0.182
ITEM24	WITH		
	ITEM23	-0.193	-0.193
	ITEM1	0.106	0.106
ITEM23	WITH		
	ITEM22	-0.189	-0.189
	ITEM13	0.127	0.127

ITEM17	0.137	0.137
ITEM21	-0.109	-0.109
ITEM20	-0.119	-0.119
ITEM2 WITH		
ITEM1	-0.205	-0.205
ITEM4 WITH		
ITEM3	-0.174	-0.174
ITEM19 WITH		
ITEM18	-0.200	-0.200
ITEM6 WITH		
ITEM5	-0.180	-0.180
ITEM30 WITH		
ITEM15	0.178	0.178
ITEM2	0.098	0.098
ITEM23	-0.134	-0.134
ITEM18	0.166	0.166
ITEM7 WITH		
ITEM6	-0.149	-0.149
ITEM16 WITH		
ITEM9	0.120	0.120
ITEM3	0.089	0.089
ITEM25 WITH		
ITEM24	-0.129	-0.129
ITEM16	0.100	0.100
ITEM12 WITH		
ITEM8	0.122	0.122
ITEM15 WITH		
ITEM14	-0.099	-0.099
ITEM7	0.131	0.131
ITEM10	0.119	0.119
ITEM9	0.099	0.099
ITEM8 WITH		
ITEM7	-0.094	-0.094
ITEM28 WITH		
ITEM14	0.164	0.164
ITEM7	0.165	0.165

ITEM5	-0.145	-0.145
ITEM5 WITH		
ITEM4	-0.144	-0.144
ITEM21 WITH		
ITEM5	0.141	0.141
ITEM3	0.098	0.098
ITEM17	-0.115	-0.115
ITEM27 WITH		
ITEM26	-0.123	-0.123
ITEM18	0.122	0.120
ITEM26 WITH		
ITEM23	0.142	0.142
ITEM17	0.091	0.091
ITEM3 WITH		
ITEM2	-0.125	-0.125

Residual Variances

FACTOR	1.000	1.000

ภาคผนวก ช

ตัวอย่าง Print Out ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้วยวิธี BAYESIAN โดยใช้โปรแกรม WinBUGS

ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT
ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้วยวิธี BAYESIAN

mean	sd	MC_error	val2.5pc	median	val97.5	pc	start	sample
b[1]	2.272	0.1658	-0.575	1.951	0.305	2.622	5001	5000
b[2]	1.017	0.1688	0.988	0.6856	2.058	1.348	5001	5000
b[3]	5.012	0.2428	-0.319	4.467	0.682	5.478	5001	5000
b[4]	6.172	0.2902	-0.299	5.675	0.929	6.749	5001	5000
b[5]	0.8524	0.1549	-0.523	0.5492	0.166	1.159	5001	5000
b[6]	4.283	0.2369	-0.549	3.766	0.584	4.755	5001	5000
b[7]	1.912	0.1674	-0.829	1.589	0.011	2.256	5001	5000
b[8]	3.504	0.2093	-0.376	3.126	0.432	3.890	5001	5000
b[9]	3.473	0.2088	-0.285	3.092	0.681	3.919	5001	5000
b[10]	-0.5041	0.1556	-0.777	-0.805	-0.016	-0.2087	5001	5000
b[11]	3.938	0.2237	-0.438	3.568	0.529	4.419	5001	5000
b[12]	6.006	0.300	-0.229	5.444	0.953	6.574	5001	5000
b[13]	-1.854	0.1703	-0.688	-2.173	0.146	-1.512	5001	5000
b[14]	-0.062	0.1651	-1.030	-0.357	-0.143	0.256	5001	5000
b[15]	-0.331	0.1557	-1.190	-0.6135	-0.350	-0.012	5001	5000
b[16]	-0.996	0.1547	-0.942	-1.302	-0.158	-0.685	5001	5000
b[17]	-0.197	0.1516	-0.889	-0.4976	-0.116	0.104	5001	5000
b[18]	3.489	0.209	-0.763	3.089	0.214	3.940	5001	5000
b[19]	5.394	0.2562	-0.184	4.884	1.071	5.866	5001	5000
b[20]	4.482	0.2415	-0.180	4.054	0.829	5.022	5001	5000
b[21]	3.898	0.2189	-0.376	3.497	0.649	4.337	5001	5000
b[22]	2.672	0.2002	-0.444	2.275	0.493	3.088	5001	5000
b[23]	5.341	0.2719	-0.148	4.782	1.098	5.88	5001	5000
b[24]	3.306	0.2081	-0.289	2.914	0.592	3.753	5001	5000
b[25]	4.071	0.2343	-0.337	3.601	0.866	4.483	5001	5000
b[26]	2.694	0.1815	-0.674	2.352	0.128	3.053	5001	5000
b[27]	3.725	0.2098	-0.465	3.341	0.452	4.226	5001	5000
b[28]	3.811	0.1901	-0.806	3.462	0.161	4.220	5001	5000
b[29]	4.961	0.2339	-0.178	4.534	0.924	5.444	5001	5000
b[30]	-0.715	0.1454	-0.778	-0.993	-0.027	-0.419	5001	5000

ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT
ชั้นประถมศึกษาปีที่ 3 ด้านคำนวณ ด้วยวิธี BAYESIAN

mean	sd	MC_error	val2.5pc	median	val97.5	pc	start	sample
b[1]	-0.0427	0.1560	-0.71	-0.3081	0.155	0.305	5001	5000
b[2]	5.922	0.2558	0.147	5.424	1.040	6.418	5001	5000
b[3]	1.715	0.1692	-0.319	1.376	0.599	2.052	5001	5000
b[4]	4.315	0.2210	-2.259	3.868	0.710	4.760	5001	5000
b[5]	4.593	0.2180	-0.641	4.159	0.335	5.037	5001	5000
b[6]	3.435	0.1906	-0.417	3.083	0.377	3.806	5001	5000
b[7]	1.726	0.1808	-0.545	1.359	0.312	2.070	5001	5000
b[8]	2.456	0.1688	-0.271	2.124	0.606	2.797	5001	5000
b[9]	2.778	0.1718	-0.400	2.443	0.346	3.115	5001	5000
b[10]	1.868	0.1804	-0.463	1.517	0.425	2.205	5001	5000
b[11]	3.383	0.2058	-0.594	2.988	0.254	3.792	5001	5000
b[12]	4.425	0.2219	-0.849	3.994	0.112	4.878	5001	5000
b[13]	2.422	0.1871	-0.205	2.066	0.634	2.804	5001	5000
b[14]	2.111	0.1679	-0.300	1.811	0.587	2.466	5001	5000
b[15]	3.343	0.1779	-0.236	3.008	0.625	3.687	5001	5000
b[16]	3.232	0.2063	-0.326	2.841	0.711	3.653	5001	5000
b[17]	0.1984	0.1510	-0.821	-0.097	0.001	0.478	5001	5000
b[18]	4.219	0.2088	-0.599	3.816	0.394	4.655	5001	5000
b[19]	4.039	0.2409	-0.330	3.589	0.722	4.519	5001	5000
b[20]	5.179	0.2627	0.045	4.718	1.158	5.690	5001	5000
b[21]	3.965	0.2298	0.024	3.549	1.052	4.436	5001	5000
b[22]	3.370	0.214	-0.167	2.996	0.729	3.819	5001	5000
b[23]	-1.166	0.1625	-1.052	-1.489	-0.302	-0.862	5001	5000
b[24]	4.195	0.2285	-0.641	3.800	0.323	4.678	5001	5000
b[25]	3.193	0.1765	-1.011	2.879	-0.142	3.590	5001	5000
b[26]	0.4959	0.1606	-0.201	0.173	-0.346	0.798	5001	5000
b[27]	3.387	0.1956	-0.188	3.008	0.688	3.776	5001	5000
b[28]	4.166	0.2324	-0.136	3.694	0.836	4.642	5001	5000
b[29]	1.481	0.1647	-0.448	1.184	0.334	1.787	5001	5000
b[30]	1.919	0.1686	-0.393	1.609	0.428	2.240	5001	5000

ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT
ชั้นประถมศึกษาปีที่ 3 ด้านเหตุผล ด้วยวิธี BAYESIAN

	mean	sd	MC_error	val2.5pc	median	val97.5pc		start	sample
b[1]	2.803	0.1885	-0.523	2.425	0.166	3.187	5001	5000	
b[2]	5.88	0.3008	-0.829	5.324	0.011	6.509	5001	5000	
b[3]	3.442	0.2025	-0.034	3.056	0.852	3.836	5001	5000	
b[4]	4.345	0.2426	0.718	3.880	2.069	4.842	5001	5000	
b[5]	0.5835	0.1688	-0.339	0.2279	0.609	0.9183	5001	5000	
b[6]	1.543	0.1798	-0.857	1.197	0.038	1.909	5001	5000	
b[7]	2.768	0.1930	-0.271	2.400	0.606	3.152	5001	5000	
b[8]	4.766	0.2624	0.118	4.280	1.530	5.293	5001	5000	
b[9]	4.981	0.2479	-1.050	4.508	0.217	5.52	5001	5000	
b[10]	3.342	0.2069	-0.688	2.965	0.146	3.775	5001	5000	
b[11]	4.202	0.2278	-0.382	3.778	0.738	4.628	5001	5000	
b[12]	3.86	0.2307	0.741	3.428	2.088	4.35	5001	5000	
b[13]	1.903	0.1901	-2.239	1.545	-1294	2.296	5001	5000	
b[14]	1.849	0.1887	-0.180	1.477	0.829	2.205	5001	5000	
b[15]	4.053	0.222	-1.741	3.609	-0.712	4.499	5001	5000	
b[16]	3.731	0.2041	-0.184	3.36	1.071	4.156	5001	5000	
b[17]	3.77	0.2038	-1.085	3.39	-0.005	4.185	5001	5000	
b[18]	3.279	0.2295	-0.917	2.856	0.244	3.736	5001	5000	
b[19]	3.984	0.2286	0.769	3.559	1.986	4.487	5001	5000	
b[20]	2.416	0.1918	-0.812	2.056	0.228	2.797	5001	5000	
b[21]	2.242	0.1997	-1.313	1.827	-0.143	2.654	5001	5000	
b[22]	2.4	0.1897	-0.092	2.054	0.971	2.802	5001	5000	
b[23]	4.254	0.2418	-0.289	3.78	0.592	4.742	5001	5000	
b[24]	3.678	0.2102	-0.096	3.306	1.058	4.09	5001	5000	
b[25]	4.025	0.2616	0.355	3.571	1.737	4.584	5001	5000	
b[26]	0.6462	0.1718	-0.742	0.3157	-0.810	0.9922	5001	5000	
b[27]	2.372	0.1888	-0.116	2.014	0.775	2.757	5001	5000	
b[28]	3.95	0.2264	-0.746	3.515	0.418	4.414	5001	5000	
b[29]	4.603	0.2528	-0.167	4.124	0.729	5.087	5001	5000	
b[30]	-0.5414	0.1719	-1.567	-0.8781	-0.585	-0.2139	5001	50(	

ภาคผนวก ซ

ผลการทดสอบทางสถิติ Chi – Square ของผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ระหว่างวิธี HGLM กับวิธี MIMIC ทั้ง 3 ด้าน

ตารางที่ ซ-1 ผลการทดสอบทางสถิติ Chi – square ผลการตรวจสอบการทำหน้าที่ต่างกัน ของข้อสอบ ระหว่างวิธี HGLM กับวิธี MIMIC ด้านภาษา

HGLM * MIMIC Crosstabulation

			MIMIC		Total
			DIF	NO DIF	
HGLM	DIF	Count	1_a	8_a	9
		% within HGLM	11.1%	88.9%	100.0%
		% within MIMIC	100.0%	27.6%	30.0%
		% of Total	3.3%	26.7%	30.0%
	NO DIF	Count	0_a	1_a	1
		% within HGLM	0.0%	100.0%	100.0%
		% within MIMIC	0.0%	3.4%	3.3%
		% of Total	0.0%	3.3%	3.3%
	NO-DIF	Count	0_a	20_a	20
		% within HGLM	0.0%	100.0%	100.0%
		% within MIMIC	0.0%	69.0%	66.7%
		% of Total	0.0%	66.7%	66.7%
Total		Count	1	29	30
		% within HGLM	3.3%	96.7%	100.0%
		% within MIMIC	100.0%	100.0%	100.0%
		% of Total	3.3%	96.7%	100.0%

Each subscript letter denotes a subset of MIMIC 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	2.414[a]	2	.299
Likelihood Ratio	2.490	2	.288
N of Valid Cases	30		

a. 4 cells (66.7%) have expected count less than 5. The minimum expected count is .03.

ตารางที่ ซ-2 ผลการทดสอบทางสถิติ Chi – square ผลการตรวจสอบการทำหน้าที่ต่างกัน ของข้อสอบ ระหว่างวิธี HGLM กับวิธี MIMIC ด้านคำนวณ

HGLM * MIMIC Crosstabulation

			MIMIC			Total
			DIF	NO DIF	NO-DIF	
HGLM	DIF	Count	5	0	0	5
		% within HGLM	100.0%	0.0%	0.0%	100.0%
		% within MIMIC	62.5%	0.0%	0.0%	16.7%
		% of Total	16.7%	0.0%	0.0%	16.7%
	NO-DIF	Count	3	21	1	25
		% within HGLM	12.0%	84.0%	4.0%	100.0%
		% within MIMIC	37.5%	100.0%	100.0%	83.3%
		% of Total	10.0%	70.0%	3.3%	83.3%
Total		Count	8	21	1	30
		% within HGLM	26.7%	70.0%	3.3%	100.0%
		% within MIMIC	100.0%	100.0%	100.0%	100.0%
		% of Total	26.7%	70.0%	3.3%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	16.500[a]	2	.000
Likelihood Ratio	16.449	2	.000
N of Valid Cases	30		

a. 4 cells (66.7%) have expected count less than 5. The minimum expected count is .17.

ตารางที่ ซ-3 ผลการทดสอบทางสถิติ Chi – square ผลการตรวจสอบการทำหน้าที่ต่างกัน ของข้อสอบ ระหว่างวิธี HGLM กับวิธี MIMIC ด้านเหตุผล

HGLM * MIMIC Crosstabulation

			MIMIC		Total
			DIF	NO DIF	
HGLM	DIF	Count	7	10	17
		% within HGLM	41.2%	58.8%	100.0%
		% within MIMIC	63.6%	52.6%	56.7%
		% of Total	23.3%	33.3%	56.7%
	NO DIF	Count	4	9	13
		% within HGLM	30.8%	69.2%	100.0%
		% within MIMIC	36.4%	47.4%	43.3%
		% of Total	13.3%	30.0%	43.3%
Total		Count	11	19	30
		% within HGLM	36.7%	63.3%	100.0%
		% within MIMIC	100.0%	100.0%	100.0%
		% of Total	36.7%	63.3%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.344[a]	1	.558		
Continuity Correction[b]	.042	1	.838		
Likelihood Ratio	.346	1	.556		
Fisher's Exact Test				.708	.421
N of Valid Cases	30				

a. 1 cells (25.0%) have expected count less than 5. The minimum expected count is 4.77.

b. Computed only for a 2x2 table

ภาคผนวก ฌ

ผลการทดสอบทางสถิติ Chi – Square ของผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ระหว่างวิธี MIMIC กับวิธี BAYESIAN ทั้ง 3 ด้าน

ตารางที่ ฌ-1 ผลการทดสอบทางสถิติ Chi – square ผลการตรวจสอบการทำหน้าที่ต่างกัน ของข้อสอบ ระหว่างวิธี MIMIC กับวิธี BAYESIAN ด้านภาษา

MIMIC * BAYESIAN Crosstabulation

			BAYESIAN		Total
			DIF	NO DIF	
MIMIC	DIF	Count	1	0	1
		% within MIMIC	100.0%	0.0%	100.0%
		% within BAYESIAN	14.3%	0.0%	3.3%
		% of Total	3.3%	0.0%	3.3%
	NO DIF	Count	6	23	29
		% within MIMIC	20.7%	79.3%	100.0%
		% within BAYESIAN	85.7%	100.0%	96.7%
		% of Total	20.0%	76.7%	96.7%
Total		Count	7	23	30
		% within MIMIC	23.3%	76.7%	100.0%
		% within BAYESIAN	100.0%	100.0%	100.0%
		% of Total	23.3%	76.7%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	3.399[a]	1	.065		
Continuity Correction[b]	.411	1	.521		
Likelihood Ratio	3.027	1	.082		
Fisher's Exact Test				.233	.233
N of Valid Cases	30				

a. 2 cells (50.0%) have expected count less than 5. The minimum expected count is .23.

b. Computed only for a 2x2 table

ตารางที่ ฌ-2 ผลการทดสอบทางสถิติ Chi – square ผลการตรวจสอบการทำหน้าที่ต่างกัน ของข้อสอบ ระหว่างวิธี MIMIC กับวิธี BAYESIAN ด้านคำนวณ

MIMIC * BAYESIAN Crosstabulation

			BAYESIAN		Total
			DIF	NO DIF	
MIMIC	DIF	Count	3	5	8
		% within MIMIC	37.5%	62.5%	100.0%
		% within BAYESIAN	50.0%	20.8%	26.7%
		% of Total	10.0%	16.7%	26.7%
	NO DIF	Count	3	19	22
		% within MIMIC	13.6%	86.4%	100.0%
		% within BAYESIAN	50.0%	79.2%	73.3%
		% of Total	10.0%	63.3%	73.3%
Total		Count	6	24	30
		% within MIMIC	20.0%	80.0%	100.0%
		% within BAYESIAN	100.0%	100.0%	100.0%
		% of Total	20.0%	80.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	2.088[a]	1	.148		
Continuity Correction[b]	.863	1	.353		
Likelihood Ratio	1.914	1	.167		
Fisher's Exact Test				.300	.175
N of Valid Cases	30				

a. 2 cells (50.0%) have expected count less than 5. The minimum expected count is 1.60.

b. Computed only for a 2x2 table

ตารางที่ ฌ-3 ผลการทดสอบทางสถิติ Chi – square ผลการตรวจสอบการทำหน้าที่ต่างกัน ของข้อสอบ ระหว่างวิธี MIMIC กับวิธี BAYESIAN ด้านเหตุผล

MIMIC * BAYESIAN Crosstabulation

			BAYESIAN		Total
			DIF	NO DIF	
MIMIC	DIF	Count	4	7	11
		% within MIMIC	36.4%	63.6%	100.0%
		% within BAYESIAN	36.4%	36.8%	36.7%
		% of Total	13.3%	23.3%	36.7%
	NO DIF	Count	7	12	19
		% within MIMIC	36.8%	63.2%	100.0%
		% within BAYESIAN	63.6%	63.2%	63.3%
		% of Total	23.3%	40.0%	63.3%
Total		Count	11	19	30
		% within MIMIC	36.7%	63.3%	100.0%
		% within BAYESIAN	100.0%	100.0%	100.0%
		% of Total	36.7%	63.3%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.001[a]	1	.979		
Continuity Correction[b]	.000	1	1.000		
Likelihood Ratio	.001	1	.979		
Fisher's Exact Test				1.000	.646
N of Valid Cases	30				

a. 1 cells (25.0%) have expected count less than 5. The minimum expected count is 4.03.

b. Computed only for a 2x2 table

ภาคผนวก ญ
ผลการทดสอบทางสถิติ Chi – Square ของผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ระหว่างวิธี HGLM กับวิธี BAYESIAN ทั้ง 3 ด้าน

ตารางที่ ญ-1 ผลการทดสอบทางสถิติ Chi – square ผลการตรวจสอบการทำหน้าที่ต่างกัน ของข้อสอบ ระหว่างวิธี HGLM กับวิธี BAYESIAN ด้านภาษา

HGLM * BAYESIAN Crosstabulation

			BAYESIAN		Total
			DIF	NO DIF	
HGLM	DIF	Count	2	7	9
		% within HGLM	22.2%	77.8%	100.0%
		% within BAYESIAN	28.6%	30.4%	30.0%
		% of Total	6.7%	23.3%	30.0%
	NO-DIF	Count	5	16	21
		% within HGLM	23.8%	76.2%	100.0%
		% within BAYESIAN	71.4%	69.6%	70.0%
		% of Total	16.7%	53.3%	70.0%
Total		Count	7	23	30
		% within HGLM	23.3%	76.7%	100.0%
		% within BAYESIAN	100.0%	100.0%	100.0%
		% of Total	23.3%	76.7%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.009[a]	1	.925		
Continuity Correction[b]	.000	1	1.000		
Likelihood Ratio	.009	1	.925		
Fisher's Exact Test				1.000	.657
N of Valid Cases	30				

a. 2 cells (50.0%) have expected count less than 5. The minimum expected count is 2.10.

b. Computed only for a 2x2 table

ตารางที่ ญ-2 ผลการทดสอบทางสถิติ Chi – square ผลการตรวจสอบการทำหน้าที่ต่างกัน ของข้อสอบ ระหว่างวิธี HGLM กับวิธี BAYESIAN ด้านคำนวณ

HGLM * BAYESIAN Crosstabulation

			BAYESIAN		Total
			DIF	NO DIF	
HGLM	DIF	Count	3	2	5
		% within HGLM	60.0%	40.0%	100.0%
		% within BAYESIAN	50.0%	8.3%	16.7%
		% of Total	10.0%	6.7%	16.7%
	NO-DIF	Count	3	22	25
		% within HGLM	12.0%	88.0%	100.0%
		% within BAYESIAN	50.0%	91.7%	83.3%
		% of Total	10.0%	73.3%	83.3%
Total		Count	6	24	30
		% within HGLM	20.0%	80.0%	100.0%
		% within BAYESIAN	100.0%	100.0%	100.0%
		% of Total	20.0%	80.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	6.000[a]	1	.014		
Continuity Correction[b]	3.375	1	.066		
Likelihood Ratio	4.948	1	.026		
Fisher's Exact Test				.041	.041
N of Valid Cases	30				

a. 2 cells (50.0%) have expected count less than 5. The minimum expected count is 1.00.

b. Computed only for a 2x2 table

ตารางที่ ญ-3 ผลการทดสอบทางสถิติ Chi – square ผลการตรวจสอบการทำหน้าที่ต่างกัน ของข้อสอบ ระหว่างวิธี HGLM กับวิธี BAYESIAN ด้านเหตุผล

HGLM * BAYESIAN Crosstabulation

			BAYESIAN		Total
			DIF	NO DIF	
HGLM	DIF	Count	6	11	17
		% within HGLM	35.3%	64.7%	100.0%
		% within BAYESIAN	54.5%	57.9%	56.7%
		% of Total	20.0%	36.7%	56.7%
	NO DIF	Count	5	8	13
		% within HGLM	38.5%	61.5%	100.0%
		% within BAYESIAN	45.5%	42.1%	43.3%
		% of Total	16.7%	26.7%	43.3%
Total		Count	11	19	30
		% within HGLM	36.7%	63.3%	100.0%
		% within BAYESIAN	100.0%	100.0%	100.0%
		% of Total	36.7%	63.3%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.032[a]	1	.858		
Continuity Correction[b]	.000	1	1.000		
Likelihood Ratio	.032	1	.859		
Fisher's Exact Test				1.000	.579
N of Valid Cases	30				

a. 1 cells (25.0%) have expected count less than 5. The minimum expected count is 4.77.

b. Computed only for a 2x2 table