การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้วยวิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล

อรุณี แปลงกาย

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต
สาขาวิชาการวิจัยและสถิติทางวิทยาการปัญญา
วิทยาลัยวิทยาการวิจัยและวิทยาการปัญญา มหาวิทยาลัยบูรพา
กรกฎาคม 2561

คณะกรรมการควบคุมวิทยานิพนธ์และคณะกรรมการสอบวิทยานิพนธ์ ได้พิจารณาวิทยานิพนธ์ของ อรุณี แปลงกาย 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)

วิทยาลัยวิทยาการวิจัยและวิทยาการปัญญา อนุมัติให้รับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาการวิจัยและสถิติทางวิทยาการปัญญา ของมหาวิทยาลัยบูรพา

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

ประกาศคุณูปการ

วิทยานิพนธ์ เรื่อง การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ระหว่างวิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซลฉบับนี้ สำเร็จลุล่วงได้ด้วยความกรุณาจาก ดร.ปิยะทิพย์ ประดุจพรม อาจารย์ที่ปรึกษาหลัก ซึ่งเฝ้าอบรมบ่มเพาะ ถ่ายทอดความรู้ ประสิทธิ์ประสาทวิชา และให้คำแนะนำเป็นอย่างดี ทั้งในส่วนของเนื้อหาวิทยานิพนธ์และวิธีการปรับตัว ตลอดระยะเวลาที่ดำเนินการวิจัย ทำให้ผู้วิจัยสามารถฟันฝ่าอุปสรรคต่าง ๆ มาได้ด้วยดี และความเอาใจใส่ ช่วยเหลือในทุกขั้นตอนของการทำวิทยานิพนธ์ ผู้วิจัยรู้สึกซาบซึ้งและระลึกถึงพระคุณอันหาที่เปรียบมิได้ในครั้งนี้เสมอ จึงขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ โอกาสนี้

ขอกราบขอบพระคุณผู้ช่วยศาสตราจารย์ ดร.สุชาดา กรเพชรปาณี คณบดีวิทยาลัยวิทยาการวิจัยและวิทยาการปัญญา สำหรับการสนับสนุน ส่งเสริมกิจกรรมเชิงวิชาการ และการให้แนวคิดในการทำงานอย่างเป็นระบบ ทำให้ผู้วิจัยมีความมุ่งมั่นในการทำงานจนประสบความสำเร็จ รวมถึงเจ้าหน้าที่ของวิทยาลัยวิทยาการวิจัยและวิทยาการปัญญาทุกคนที่ให้บริการ และการประสานงานที่ดีเยี่ยมอยู่เสมอ

ขอกราบขอบพระคุณบุคลากรทุกท่านและหน่วยงานที่เกี่ยวข้องทุกแห่งสำหรับการอำนวยความสะดวกตลอดการดำเนินการวิจัย สำหรับการเสียสละเวลาอันมีค่าเพื่อให้ข้อมูลสำคัญสำหรับการศึกษาในครั้งนี้ นอกจากนี้ ขอขอบคุณมิตรภาพของเพื่อนทุกกลุ่ม สำหรับการสนับสนุน และกระตุ้นให้ผู้วิจัยมีแรงกาย แรงใจมากเพียงพอที่จะดำเนินการวิจัยจนผ่านพ้นไปได้ด้วยดี ผู้วิจัยขอขอบคุณในมิตรภาพที่มอบให้เสมอมา

สุดท้ายนี้ ผู้วิจัยกราบขอบพระคุณคณาจารย์ทุกท่าน ที่ได้ประสิทธิ์ประสาทวิชาความรู้ และขอบคุณครอบครัวที่คอยให้ความช่วยเหลือเป็นกำลังใจตลอดมา ประโยชน์ของวิทยานิพนธ์ฉบับนี้ ผู้วิจัยขอมอบเป็นกตัญญูกตเวทิตาแด่บุพการี บูรพาจารย์ และผู้มีพระคุณทุกท่านทั้งในอดีต และปัจจุบันที่ทำให้ข้าพเจ้าเป็นผู้มีการศึกษา และประสบความสำเร็จมาจนตราบเท่าทุกวันนี้

อรุณี แปลงกาย

56910403: สาขาวิชา: การวิจัยและสถิติทางวิทยาการปัญญา;
วท.ม. (การวิจัยและสถิติทางวิทยาการปัญญา)
คำสำคัญ: การทำหน้าที่ต่างกันของข้อสอบ/ ข้อสอบ NT/ วิธีการทดสอบอัตราส่วนไลค์ลิฮูด/ วิธีซิปเทสท์/ วิธีแมนเทล-แฮนส์เซล

อรุณี แปลงกาย: การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้วยวิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล (A COMPARISON OF THE IRT LIKELIHOOD RATIO TEST, SIBTEST, AND THE MANTEL-HAENSZEL METHOD FOR DETECTING DIFFERENTIAL ITEM FUNCTIONING USING RESULTS FROM THE GRADE 3 NATIONAL TEST) คณะกรรมการควบคุมวิทยานิพนธ์: ปิยะทิพย์ ประดุจพรม, Ph.D., 256 หน้า. ปี พ.ศ. 2561.

การวิจัยนี้มีวัตถุประสงค์เพื่อวิเคราะห์คุณภาพของข้อสอบของแบบทดสอบระดับชาติ (NT) และตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ภายใต้เงื่อนไขกลุ่มตัวอย่างขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) ด้วยวิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล การดำเนินการวิจัยแบ่งเป็น 3 ระยะ ดังนี้ 1) วิเคราะห์คุณภาพของแบบทดสอบระดับชาติ ทั้ง 3 ด้าน 2) ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้ง 3 วิธี และ 3) เปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการตรวจสอบ 3 วิธี ข้อมูลที่นำมาใช้วิเคราะห์เป็นข้อมูลทุติยภูมิ ซึ่งเป็นผลการตอบแบบทดสอบระดับชาติของนักเรียน ชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2556 จำนวนทั้งหมด 706,372 คน

ผลการวิจัยปรากฏว่า

1. แบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 มีค่าความยากของข้อสอบ (b) อยู่ในระดับค่อนข้างยาก มีค่าอำนาจจำแนกของข้อสอบ (a) อยู่ในระดับที่สามารถจำแนกผู้สอบได้ดีมาก และมีค่าโอกาสในการเดาของข้อสอบ (c) ไม่เกิน .30

2. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้ง 3 ด้าน พบว่า กลุ่มตัวอย่างขนาดใหญ่ ส่งผลให้เกิดการทำหน้าที่ต่างกันของข้อสอบ จะได้เปรียบในการตอบข้อสอบ ด้านภาษา ด้านคำนวณ และด้านเหตุผล มากกว่าขนาดเล็ก และขนาดกลาง โดยวิธีแมนเทล-แฮนส์เซล ตรวจพบข้อสอบทำหน้าที่ต่างกัน จำนวนมากที่สุด คิดเป็นร้อยละ 34 ของข้อสอบทั้งหมด รองลงมาคือ วิธีซิปเทสท์ และวิธีการทดสอบอัตราส่วนไลค์ลิฮูด คิดเป็นร้อยละ 14 ของข้อสอบทั้งหมด

3. การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ พบว่า วิธีแมนเทล-แฮนส์เซล ตรวจพบ DIF มากกว่าวิธีการทดสอบอัตราส่วนไลค์ลิฮูดในด้านภาษา ด้านคำนวณ และด้านเหตุผล คิดเป็นร้อยละ 80, 13 และ 80 และวิธีแมนเทล-แฮนส์เซล ตรวจพบ DIF มากกว่าวิธีซิปเทสท์ ทั้ง 3 ด้าน คิดเป็นร้อยละ 87, 13 และ 73 วิธีการทดสอบอัตราส่วนไลค์ลิฮูด ตรวจพบ DIF มากกว่าวิธีซิปเทสท์ในด้านภาษา คิดเป็นร้อยละ 7 และวิธีการทดสอบอัตราส่วนไลค์ลิฮูด ตรวจพบ DIF เท่ากันกับวิธีซิปเทสท์ ในด้านคำนวณ ส่วนวิธีซิปเทสท์ ตรวจพบ DIF มากกว่า วิธีการทดสอบอัตราส่วนไลค์ลิฮูด ด้านเหตุผล คิดเป็นร้อยละ 7 ($p < .05$)

56910403: MAJOR: RESEARCH AND STATISTICS IN COGNITIVE SCIENCE;
M.Sc. (RESEARCH AND STATISTICS IN COGNITIVE SCIENCE)
KEYWORDS: DIFFERENTIAL ITEM FUNCTIONING / NATIONAL TESTS: NT/ IRT-LR METHOD / SIBTEST METHOD / MANTEL-HAENSZEL METHOD

ARUNEE PLANGKAY: A COMPARISON OF THE IRT LIKELIHOOD RATIO TEST, SIBTEST, AND THE MANTEL-HAENSZEL METHOD FOR DETECTING DIFFERENTIAL ITEM FUNCTIONING USING RESULTS FROM THE GRADE 3 NATIONAL TEST. ADVISORY COMMITTEE: PIYATHIP PRADUJPROM, Ph.D. 256 P. 2018.

The objectives of this research were to analyze the quality of test items from National Tests (NT), and to determine if differential item functioning (DIF) was present. Three subjects were involved: Literacy, Numeracy, and Reasoning. Different sample sizes were used, from small (n=300), medium (n=1,000), to large (n=2,000). Three DIF methods were compared: IRT-LR, SIBTEST, and Mantel-Haenszel. The research procedures were divided into three phases: 1) Analyzing the item quality of NT for the three subjects; 2) Testing DIF detection of the items in NT using IRT-LR, SIBTEST, and Mantel-Haenszel methods; 3) Comparing the results of the three methods of DIF using secondary data from 706,372 NT Grade three students in the academic year 2013.

Results were as follows:

1. Items on the NT had relatively high difficult levels, very good discrimination levels, and guessing parameters which did not exceed .30.

2. The examination of results in the three subjects revealed that the larger samples had higher scores on Literacy, Numeracy, and Reasoning. The Mantel-Haenszel method detected the greatest number of DIF items across all three subjects, finding DIF on 34% of the test items; SIBTEST and IRT-LR both found DIF in 14% of the items.

3. Comparison of the DIF test results revealed that the Mantel-Haenszel method outperformed the IRT-LR method in terms of DIF detection, namely 80% for Literacy, 13% for Numeracy, and 80% for Reasoning. The Mantel-Haenszel method also outperformed the SIBTEST method in terms of DIF detection, namely 87% for Literacy, 13% for Numeracy, and 73% for Reasoning subjects. The IRT-LR method outperformed the SIBTEST method in terms of DIF detection, namely 7% for Literacy, and same DIF as for Numeracy subjects. Also, the SIBTEST method outperformed the IRT-LR method in terms of DIF detection on the Reasoning subject (7%) ($p < .05$).

สารบัญ

สารบัญ (ต่อ)

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญตาราง (ต่อ)

สารบัญตาราง (ต่อ)

สารบัญตาราง (ต่อ)

สารบัญตาราง (ต่อ)

สารบัญภาพ

สารบัญภาพ (ต่อ)

สารบัญภาพ (ต่อ)

บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

การวัดประเมินผลการศึกษามีความสำคัญกับการศึกษา ซึ่งเป็นตัวบ่งชี้ที่สำคัญของความสำเร็จของการจัดการศึกษาของประเทศที่พัฒนาแล้วทำให้องค์การระหว่างประเทศของโลกได้ให้ความสำคัญกับการประเมินผลการศึกษา โดยกำหนดการทดสอบ PISA (Programme for International Student Assessment: PISA) ขึ้นมาเพื่อเป็นการประเมินผลการศึกษาของแต่ละประเทศว่าประเทศใดสามารถที่จะพัฒนาผู้เรียนให้รู้จักคิดวิเคราะห์ มีเหตุผล และรู้จักแก้ปัญหาอยู่ในระดับใด (ปนัดดา หัสปราบ, 2557, หน้า 1) สำหรับประเทศไทยจึงได้กำหนดหลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน พุทธศักราช 2551 ขึ้นมาเพื่อการวัดผลและประเมินผลการเรียนรู้ โดยแบ่งเป็น 4 ระดับ คือ 1) ระดับชั้นเรียน 2) ระดับสถานศึกษา 3) ระดับเขตพื้นที่การศึกษา และ 4) ระดับชาติ ซึ่งการประเมินระดับชาติจะประเมินในระดับชั้นประถมศึกษาปีที่ 3 ชั้นประถมศึกษาปีที่ 6 ชั้นมัธยมศึกษาปีที่ 3 และชั้นมัธยมศึกษาปีที่ 6 (สำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน, 2557, หน้า 2-4)

การวัดผลประเมินผลทางการศึกษาเพื่อการตรวจสอบความรู้ความสามารถของนักเรียนหรือเป็นการวัดคุณลักษณะที่ต้องการวัดอยู่ในระดับใด (พีรญา สูงเนิน เสรี ชัดแช้ม และสมโภชน์ อเนกสุข 2552, หน้า 51) และการทดสอบระดับชาติ (National Testing: NT) เป็นการประเมินคุณภาพทางการศึกษาที่มีจุดมุ่งหมาย เพื่อประเมินคุณภาพการศึกษาขั้นพื้นฐาน โดยจะนำข้อมูลเพื่อจัดทำแผนการพัฒนาให้นักเรียนได้มีความสามารถอ่านออกเขียนได้ รู้จักคิดวิเคราะห์ สังเคราะห์ และเป็นการวัดผลสัมฤทธิ์ทางการเรียนของผู้เรียนแต่ละช่วงชั้น จะมีการทดสอบที่ชั้นเรียนสูงสุดของช่วงชั้น คือ ระดับชั้นประถมศึกษาปีที่ 3 ซึ่งการทดสอบจะส่งผลสะท้อนให้เห็นถึงการประสบความสำเร็จของการพัฒนาคุณภาพการศึกษาที่สอดคล้องกับพระราชบัญญัติการศึกษาแห่งชาติ โดยมีสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐานเป็นหน่วยงานที่รับผิดชอบในการจัดการศึกษา ซึ่งจะจัดสอบโดยสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน ในปีการศึกษา 2555 โดยสำนักทดสอบทางการศึกษาได้กำหนดนโยบายจัดสอบระดับชาติ (NT) ในระดับชั้นประถมศึกษาปีที่ 3 มีทั้งหมด 3 ด้าน คือ ความสามารถด้านภาษา (Literacy Ability) ความสามารถด้านคำนวณ (Numeracy Ability) และความสามารถด้านเหตุผล (Reason Ability) การวัดทั้ง 3 ด้านนี้เพื่อให้สอดคล้องกับจุดเน้นการพัฒนาคุณภาพผู้เรียน คือ ชั้นประถมศึกษาปีที่ 1-3 นักเรียนมีทักษะความสามารถในการอ่านออกเขียนได้ คิดเลขเป็น และมีทักษะการคิดขั้นพื้นฐาน ซึ่งลักษณะของแบบทดสอบจะเป็นแบบปรนัยเลือกตอบ แบบ 4 ตัวเลือก ข้อสอบด้านภาษา มีจำนวนข้อสอบ 30 ข้อ ใช้เวลาในการทดสอบ 50 นาที ข้อสอบด้านคำนวณ มีจำนวนข้อสอบ 30 ข้อ ใช้เวลาในการทดสอบ 50 นาที ข้อสอบด้านเหตุผล มีจำนวนข้อสอบ 30 ข้อ ใช้เวลาในการทดสอบ 50 นาที (สำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน, 2557, หน้า 1-4)

แบบทดสอบจึงเป็นเครื่องมือที่ใช้การวัดผลประเมินผลทางการศึกษา โดยจะเป็นการตรวจสอบว่าผู้สอบนั้นจะมีคุณลักษณะแฝงหรือความสามารถอยู่ในระดับดี การสร้างเครื่องมือในการทดสอบและการตรวจสอบคุณภาพของแบบทดสอบจะต้องคำนึงถึงความตรงตามเนื้อหา (Content Validity) ความตรงตามเกณฑ์ (Criterion Validity) และความตรงเชิงโครงสร้าง (Construct Validity) เป็นสำคัญ (พีรญา สูงเนิน เสรี ชัดแช้ม และสมโภชน์ อเนกสุข, 2552) ความตรงเป็นคุณสมบัติของข้อสอบที่แสดงว่า คะแนนจากข้อสอบสามารถสรุปอ้างอิงไปยังสิ่งที่วัดได้อย่างเหมาะสม (ปิยะทิพย์ ตินวร ม.ร.ว.สมพร สุทัศนีย์ และเสรี ชัดแช้ม, 2550) ซึ่งแบบทดสอบในบางฉบับยังขาดความตรงหรือแสดงถึงการทำหน้าที่ต่างกันของข้อสอบ และเมื่อนำไปทดสอบกับกลุ่มย่อยตั้งแต่ 2 กลุ่มขึ้นไป พบว่าผลการวิเคราะห์การตอบของผู้สอบที่มีความสามารถที่เท่ากันแต่โอกาสในการตอบได้ถูกต้องไม่เท่ากัน คือ แบบทดสอบเกิดปัจจัยที่เอื้อต่อผู้สอบบางกลุ่มเท่านั้น และนักการศึกษาพยายามที่จะตัดข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบออกเพื่อให้เกิดความยุติธรรมกับผู้สอบก็ยังเกิดปัญหาอีกเช่นเดิม และในต่างประเทศมีการศึกษาการทำหน้าที่ต่างกันของข้อสอบโดยใช้ข้อสอบ PISA 2003 และใช้วิธีการวิเคราะห์องค์ประกอบจำกัด (Restricted Factor Analysis: RFA) วิธีแมนเทล-แฮนส์เซล และทฤษฎีการตอบสนองข้อสอบในการวิเคราะห์อัตราส่วนไลค์ลิฮูด พบว่า มีข้อสอบที่ทำหน้าที่ต่างกัน (ชัยวัฒน์ หฤทัยพันธ์, 2558)

จากงานวิจัยของ Li, Hunter, and Oshima (2013) ได้ศึกษาการตรวจสอบการทำหน้าที่ต่างกันในการทดสอบ การอ่านและเหตุผล ที่เป็นไปได้ระหว่างเพศด้วยวิธีอัตราส่วนไลค์ลิฮูด และวิธีแมนเทล-แฮนส์เซล พบว่า เพศมีผลต่อการทำแบบทดสอบในการอ่านและเหตุผลที่เป็นไปได้ วิธีอัตราส่วนไลค์ลิฮูด สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ดีในแบบทดสอบที่มีความยาวตั้งแต่ 20 ข้อขึ้นไป และ Kabasakal, Arsan, Gok, and Kelecioglu (2014) ได้ศึกษาแบบจำลองการเปรียบเทียบอัตราความคลาดเคลื่อนประเภทที่ 1 และอัตราความถูกต้องโดยใช้วิธีอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล ภายใต้เงื่อนไข ขนาดกลุ่มตัวอย่าง ความยาวแบบทดสอบที่ทำหน้าที่แตกต่างกันของข้อสอบ พบว่า วิธีซิปเทสท์ มีอัตราความคลาดเคลื่อนประเภทที่ 1 สูงสุดแบบทดสอบที่ทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปแต่วิธีแมนเทล–แฮนส์เซล มีอัตราความถูกต้องสูงภายใต้เงื่อนไขของขนาดกลุ่มตัวอย่างความยาวแบบทดสอบ นอกจากนี้ค่าร้อยละของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปภายใต้เงื่อนไขขนาดกลุ่มตัวอย่าง ความยาวแบบทดสอบจะมีอิทธิพลต่ออัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีอัตราส่วนไลค์ลิฮูด และค่าร้อยละของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ มีความแตกต่างความสามารถในระยะเวลาในการทดสอบ ระยะเวลาในการทดสอบการทำหน้าที่ต่างกันของข้อสอบความยาวแบบทดสอบส่งผลให้ วิธีซิปเทสท์มีความคลาดเคลื่อนประเภทที่ 1 ส่วนวิธีแมนเทล–แฮนส์เซล ปัจจัยที่มีผลต่อความคลาดเคลื่อนประเภทที่ 1 คือ ขนาดกลุ่มตัวอย่าง ระยะเวลาในการทดสอบของการทำหน้าที่ต่างกันของข้อสอบที่มีความแตกต่างความสามารถ และไม่มีปัจจัยใดที่ส่งผลให้เกิดอัตราความถูกต้อง วิธีซิปเทสท์ และวิธีแมนเทล–แฮนส์เซล แต่จะมีผลกระทบต่อการประเมินอัตราความถูกต้องต่อวิธีอัตราส่วนไลค์ลิฮูด และการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างวิธีการวิเคราะห์องค์ประกอบจำกัด (RFA) กับวิธีถดถอยโลจิสติก (LR)

ภายใต้เงื่อนไข 18 เงื่อนไข โดยกำหนดขนาดกลุ่มตัวอย่างขนาดใหญ่ จำนวน 2,000 คน กลุ่มตัวอย่างขนาดกลาง จำนวน 1,000 คน และกลุ่มตัวอย่างขนาดเล็ก จำนวน 300 คน พบว่า วิธีถดถอยโลจิสติก มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบดีกว่าเกือบทุกเงื่อนไขโดยพิจารณาจากอัตราความถูกต้องประเภทที่ 1 และอัตราความคลาดเคลื่อนประเภทที่ 1 (ปิยะทิพย์ ตินวร ม.ร.ว.สมพร สุทัศนีย์ และเสรี ชัดแช้ม, 2550) การศึกษาขนาดกลุ่มตัวอย่างของ Stout, Li, Nandakumar, and Bolt (1997) ได้เสนอการวิเคราะห์โดยใช้โปรแกรม SIBTEST กลุ่มตัวอย่างขนาดเล็กที่ควรใช้ คือ ขนาดกลุ่ม จำนวน 100 คน จากงานวิจัยของ Narayanan and Swaminathan (1994) ได้ศึกษาการใช้กลุ่มตัวอย่างขนาดเล็ก จำนวน 300 คน ก็พอที่จะตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้อย่างมีประสิทธิภาพ และ (พีรญา สูงเนิน เสรี ชัดแช้ม และสมโภชน์ อเนกสุข, 2552) ได้ศึกษาเพื่อพัฒนาเกณฑ์ตัดสินข้อสอบลำเอียงทางเพศ พบว่า วิธีซิปเทสท์และวิธีแมนเทล–แฮนส์เซล ควรใช้กลุ่มตัวอย่างจำนวน 600 คนขึ้นไป ได้เปรียบเทียบประสิทธิภาพวิธีแมนเทล-แฮนส์เซล และวิธีซิปเทสท์ พบว่า เมื่อกลุ่มตัวอย่างมีขนาดจำนวน 200 คน และจำนวน 600 คน วิธีทั้งสองสามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ถูกต้องร้อยละ 50 แต่ถ้ากลุ่มตัวอย่างขนาดจำนวน 1,000 คน สามารถตรวจสอบได้ถูกต้อง ร้อยละ 100

การทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) มีขนาดทิศทางการทำหน้าที่ต่างกันเปลี่ยนไปตามความสามารถของผู้เข้าสอบที่แตกต่างกัน แบ่งออกเป็น 2 ประเภท คือ 1) ข้อสอบทำหน้าที่ต่างกันแบบเอกรูป (Uniform DIF) โอกาสของผู้เข้าสอบที่คะแนนสูงกว่าจะไม่ปฏิสัมพันธ์กันตลอดช่วงความสามารถของผู้เข้าสอบที่มีคะแนนต่ำกว่า 2) ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป (Nonuniform DIF) โอกาสของผู้เข้าสอบที่คะแนนสูงกว่าจะปฏิสัมพันธ์กันไม่ตลอดช่วงความสามารถของผู้เข้าสอบที่มีคะแนนต่ำกว่าตามทฤษฎีการตอบสนองข้อสอบ (IRT) โดยสามารถพิจารณาปฏิสัมพันธ์จากความแตกต่างของพารามิเตอร์อำนาจจำแนกของข้อสอบระหว่างผู้เข้าสอบกลุ่มย่อย 2 กลุ่ม และถ้าหากข้อสอบระหว่างผู้เข้าสอบ 2 กลุ่มย่อยมีค่าพารามิเตอร์อำนาจจำแนกของข้อสอบระหว่างผู้เข้าสอบกลุ่มย่อย 2 กลุ่ม ซึ่งมีค่าพารามิเตอร์อำนาจจำแนกเท่ากันแล้วนั้น โค้งคุณลักษณะของข้อสอบ (Item Characteristic Curves: ICC) ของผู้เข้าสอบ 2 กลุ่มจะขนานกันนั้นก็แสดงว่าข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอถ้าข้อสอบระหว่างกลุ่มผู้เข้าสอบ 2 กลุ่มย่อยมีค่าพารามิเตอร์อำนาจจำแนกไม่เท่ากันแล้ว โค้งคุณลักษณะข้อสอบ (Item Characteristic Curves: ICC) ของผู้เข้าสอบ 2 กลุ่มจะไม่ขนานกัน นั่นก็แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ การทำหน้าที่ต่างกันของข้อสอบมาวิเคราะห์เชิงตรรกะ (Logical Analysis) นอกจากนี้การทำหน้าที่ต่างกันของข้อสอบเป็นการเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มผู้สอบอย่างน้อย 2 กลุ่มขึ้นไป ประกอบด้วยกลุ่มแรกที่เรียกว่ากลุ่มเปรียบเทียบ (Focal Group: F) เป็นกลุ่มที่สนใจศึกษาและคาดว่าจะเป็นกลุ่มที่เสียเปรียบในการตอบข้อสอบ และกลุ่มที่สองเรียกว่า กลุ่มอ้างอิง (Reference Group: R) เป็นกลุ่มที่คาดว่าจะได้เปรียบในการตอบข้อสอบได้ถูกต้อง ซึ่งในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ (ศิริชัย กาญจนวาสี, 2555, หน้า 115-118)

การเปรียบเทียบผลการตอบข้อสอบในระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบต้องจับคู่ (Matching) ผู้สอบตามความสามารถ โดยเป็นเงื่อนไขของการตรวจสอบการทำหน้าที่ต่างกัน

ของข้อสอบเกณฑ์การจับคู่ (Matching Criteria) ที่นิยมใช้กันมี 2 วิธี ดังนี้ (ศิริชัย กาญจนวาสี, 2555, หน้า 120-122) เกณฑ์ภายนอก (External Criterion) การวิเคราะห์การทำหน้าที่ต่างกัน จะใช้เกณฑ์ภายนอกนี้สามารถใช้ได้ทั้งข้อสอบรายข้อและข้อสอบทั้งฉบับ โดยการใช้คะแนนจากข้อสอบอื่นเป็นเกณฑ์ภายนอกแล้วใช้เทคนิคการวิเคราะห์การถดถอย (Regression Analysis) เพื่อเป็นการเปรียบเทียบเส้นกราฟความสัมพันธ์ระหว่างตัวแปรเกณฑ์ กับตัวแปรทำนายระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบและเกณฑ์ภายใน (Internal Criterion) การวิเคราะห์การทำหน้าที่ต่างกัน โดยใช้เกณฑ์ภายในเป็นการนำวิธีการทางสถิติมาตรวจสอบการทำหน้าที่ต่างกันของข้อสอบหรือแบบสอบจะเป็นการเน้นการพิจารณาจากโครงสร้างภายในของแบบสอบเป็นหลัก และการวิเคราะห์ผลจากการตอบข้อสอบและความสามารถหรือคะแนนจริงของผู้เข้าสอบได้จากแบบทดสอบฉบับนั้น เพื่อนำมาเปรียบเทียบระหว่างผู้เข้าสอบจากกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่มีความสามารถหรือคะแนนจริงเท่ากัน ว่าจะมีผลการตอบหรือโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกันหรือไม่เพื่อบ่งชี้การทำนายของข้อสอบ

ผู้วิจัยจึงมีความสนใจที่จะศึกษาการเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ระหว่างวิธีอัตราส่วนไลค์ลิฮูด วิธีชิปเทสท์ และวิธีแมนเทล-แฮนส์เซล ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก จำนวน 300 คน ขนาดกลาง จำนวน 1,000 คน และขนาดใหญ่ จำนวน 2,000 คน และข้อค้นพบที่ได้จะเป็นแนวทางในการดำเนินการจัดทำแบบทดสอบระดับชาติในชั้นประถมศึกษาปีที่ 3 ในครั้งต่อไปว่าการที่จะนำแบบทดสอบที่จะนำมาทดสอบจะต้องมีความเป็นกลางไม่เกิดการทำหน้าที่ต่างกันของข้อสอบต่อไป

วัตถุประสงค์ของการวิจัย

1. เพื่อวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 โดยใช้หลักการของทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)

2. เพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้วยวิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีชิปเทสท์ และวิธีแมนเทล-แฮนส์เซล ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)

3. เพื่อเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้วยวิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีชิปเทสท์ และวิธีแมนเทล-แฮนส์เซล ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)

กรอบแนวคิดการวิจัย

จากการศึกษาแนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้อง ผู้วิจัยได้วิเคราะห์และสังเคราะห์การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล เป็นวิธีวิเคราะห์ข้อสอบที่อยู่บนพื้นฐานทฤษฎีตอบสนองข้อสอบทั้ง 3 วิธี (Li, Hunter, & Oshima, 2013) และได้เปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล โดย Kabasakal, Arsan, Gok, and Kelecioglu (2014) ได้ศึกษาประสิทธิภาพในการตรวจสอบ DIF ด้วยวิธีแมนเทล-แฮนส์เซล วิธีซิปเทสท์ และวิธีการทดสอบอัตราส่วนไลค์ลิฮูด พบว่า วิธีการทดสอบอัตราส่วนไลค์ลิฮูด มีประสิทธิภาพในการตรวจสอบ DIF ได้ดี และงานวิจัยของ Li, Hunter, and Oshima (2013) ได้ศึกษาการตรวจสอบการทำหน้าที่ต่างกันในแบบทดสอบด้านการอ่าน และด้านเหตุผล ด้วยวิธีการทดสอบอัตราส่วนไลค์ลิฮูด และวิธีแมนเทล-แฮนส์เซล พบว่า วิธีการทดสอบอัตราส่วนไลค์ลิฮูด มีประสิทธิภาพในการตรวจสอบ DIF ได้ดีในแบบทดสอบที่มีความยาวของข้อสอบตั้งแต่ 20 ข้อขึ้นไป มากำหนดเป็นกรอบแนวคิดการวิจัย ดังภาพที่ 1-1

เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ชั้นประถมศึกษาปีที่ 3								
ด้านภาษา			ด้านคำนวณ			ด้านเหตุผล		
ขนาดกลุ่มตัวอย่าง			ขนาดกลุ่มตัวอย่าง			ขนาดกลุ่มตัวอย่าง		
300 คน	1,000 คน	2,000 คน	300 คน	1,000 คน	2,000 คน	300 คน	1,000 คน	2,000 คน
วิธี IRT-LR กับ วิธี SIBTEST	วิธี IRT-LR กับ วิธี SIBTEST	วิธี IRT-LR กับ วิธี SIBTEST	วิธี IRT-LR กับ วิธี SIBTEST	วิธี IRT-LR กับ วิธี SIBTEST	วิธี IRT-LR กับ วิธี SIBTEST	วิธี IRT-LR กับ วิธี SIBTEST	วิธี IRT-LR กับ วิธี SIBTEST	วิธี IRT-LR กับ วิธี SIBTEST
วิธี IRT-LR กับวิธี MH	วิธี IRT-LR กับวิธี MH	วิธี IRT-LR กับวิธี MH	วิธี IRT-LR กับวิธี MH	วิธี IRT-LR กับวิธี MH	วิธี IRT-LR กับวิธี MH	วิธี IRT-LR กับวิธี MH	วิธี IRT-LR กับวิธี MH	วิธี IRT-LR กับวิธี MH
วิธี SIBTEST กับวิธี MH	วิธี SIBTEST กับวิธี MH	วิธี SIBTEST กับวิธี MH	วิธี SIBTEST กับวิธี MH	วิธี SIBTEST กับวิธี MH	วิธี SIBTEST กับวิธี MH	วิธี SIBTEST กับวิธี MH	วิธี SIBTEST กับวิธี MH	วิธี SIBTEST กับวิธี MH

ภาพที่ 1-1 กรอบแนวคิดการวิจัย

สมมติฐานของการวิจัย

1. เมื่อกลุ่มตัวอย่างขนาดเล็ก (300 คน) ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา โดยใช้วิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล แตกต่างกัน

2. เมื่อกลุ่มตัวอย่างขนาดเล็ก (300 คน) ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านคำนวณ โดยใช้วิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล แตกต่างกัน

3. เมื่อกลุ่มตัวอย่างขนาดเล็ก (300 คน) ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านเหตุผล โดยใช้วิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล แตกต่างกัน

4. เมื่อกลุ่มตัวอย่างขนาดกลาง (1,000 คน) ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา โดยใช้วิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล แตกต่างกัน

5. เมื่อกลุ่มตัวอย่างขนาดกลาง (1,000 คน) ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านคำนวณ โดยใช้วิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล แตกต่างกัน

6. เมื่อกลุ่มตัวอย่างขนาดกลาง (1,000 คน) ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านเหตุผล โดยใช้วิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล แตกต่างกัน

7. เมื่อกลุ่มตัวอย่างขนาดใหญ่ (2,000 คน) ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา โดยใช้วิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล แตกต่างกัน

8. เมื่อกลุ่มตัวอย่างขนาดใหญ่ (2,000 คน) ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านคำนวณ โดยใช้วิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล แตกต่างกัน

9. เมื่อกลุ่มตัวอย่างขนาดใหญ่ (2,000 คน) ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านเหตุผล โดยใช้วิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล แตกต่างกัน

ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

1. ได้วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบแบบหลายตัวเลือก วิธีใดเหมาะสมกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเมื่อกลุ่มตัวอย่างที่ต่างกัน

2. ได้วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ทั้ง 3 ด้าน คือ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ตรวจพบ DIF ได้ดีในกลุ่มตัวอย่างขนาดใหญ่ (2,000 คน)

3. ได้วิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเมื่อกลุ่มตัวอย่างขนาดต่างกัน 3 ขนาด คือ ขนาดใหญ่ (2,000 คน) ขนาดกลาง (1,000 คน) และขนาดเล็ก (300 คน) ว่าวิธีแมนเทล-แฮนส์เซล ตรวจพบ DIF ได้มากที่สุด รองลงมาคือ วิธีซิปเทสท์ และวิธีการทดสอบอัตราส่วนไลค์ลิฮูด

ขอบเขตของการวิจัย

การวิจัยครั้งนี้ใช้ข้อมูลทุติยภูมิ ที่เป็นผลการตอบแบบทดสอบระดับชาติ (NT) ชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้าน คือ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ของนักเรียนสังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐานโดยสำนักทดสอบทางการศึกษา จำนวนทั้งหมด 706,372 คน

ตัวแปรที่ศึกษา

1. ตัวแปรต้น มี 2 ตัว ได้แก่

1.1 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ จำนวน 3 วิธี ได้แก่

1.1.1 วิธีการทดสอบอัตราส่วนไลค์ลิฮูด (วิธี IRT-LR)

1.1.2 วิธีซิปเทสท์ (วิธี SIBTEST)

1.1.3 วิธีแมนเทล-แฮนส์เซล (วิธี MH)

1.2 ขนาดกลุ่มตัวอย่าง จำนวน 3 ขนาด ได้แก่

1.2.1 ขนาดเล็ก (300 คน)

1.2.2 ขนาดกลาง (1,000 คน)

1.2.3 ขนาดใหญ่ (2,000 คน)

2. ตัวแปรตาม คือ ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) โดยพิจารณาจากความถี่ (จำนวนข้อ) และร้อยละที่ตรวจพบการทำหน้าที่ต่างกันของข้อสอบ

นิยามศัพท์เฉพาะ

การทดสอบระดับชาติ (National Testing: NT) หมายถึง การประเมินคุณภาพผู้เรียนระดับชาติของผู้เรียน ชั้นประถมศึกษาปีที่ 3 ตามมาตรฐานและตัวชี้วัดตามหลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน พุทธศักราช 2551 ซึ่งดำเนินการจัดสอบโดยสำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน ความสามารถด้านภาษา (Literacy Ability) ความสามารถด้านคำนวณ (Numeracy Ability) และความสามารถด้านเหตุผล (Reason Ability)

แบบทดสอบระดับชาติ (National Test: NT) หมายถึง แบบทดสอบประเมินคุณภาพผู้เรียนระดับชาติของผู้เรียน ชั้นประถมศึกษาปีที่ 3 ตามมาตรฐานและตัวชี้วัดตามหลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน พุทธศักราช 2551 ซึ่งดำเนินการจัดสอบโดยสำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน ความสามารถด้านภาษา (Literacy Ability) ความสามารถด้านคำนวณ (Numeracy Ability) และความสามารถด้านเหตุผล (Reason Ability)

การทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) หมายถึง ข้อสอบที่ผู้ตอบข้อสอบ ซึ่งมีความสามารถหรือคุณลักษณะที่ต้องการวัดเท่ากัน มีโอกาสตอบข้อสอบ

ข้อนั้นได้ถูกต้องไม่เท่ากัน เนื่องจากอยู่ในกลุ่มผู้เข้าสอบย่อยที่มีลักษณะต่างกันในที่นี้ คือ กลุ่มผู้ตอบข้อสอบนักเรียนชาย กับกลุ่มผู้ตอบนักเรียนหญิง

การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกัน หมายถึง การนำผลการตรวจพบข้อสอบที่ก่อให้เกิดการทำหน้าที่ต่างกันของข้อสอบ ระหว่างวิธีการตรวจสอบสองวิธี โดยทำการเปรียบเทียบ จำนวนข้อ (ร้อยละ) ที่พบการทำหน้าที่ต่างกันของข้อสอบ

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

วิธีการทดสอบอัตราส่วนไลค์ลิฮูด (Item Response Theory-Likelihood Ratio: IRT-LR) หมายถึง การใช้หลักของอัลกอริทึมในการประมาณค่าความเป็นไปได้ของค่าพารามิเตอร์ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบเอกรูป และอเนกรูป โดยใช้โปรแกรมสำเร็จรูป IRTPRO ตามทฤษฎีการตอบสนองข้อสอบ IRT

วิธีซิปเทสท์ (SIBTEST) หมายถึง วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่อยู่ภายในทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) ที่พัฒนาโดย Shealy and Stout (1993) ซึ่งในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบจะคำนวณดัชนีจากสัดส่วนของการตอบข้อสอบถูกของผู้เข้าสอบกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบในข้อสอบ NT ซึ่งจะใช้คะแนนที่ได้มาจับคู่เปรียบเทียบระหว่างกลุ่มผู้เข้าสอบที่ข้อสอบมีความตรง

วิธีแมนเทล-แฮนส์เซล (Mantel-Haenszel: MH) หมายถึง วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยทำการทดสอบอัตราส่วนเปรียบเทียบด้วยไค-สแควร์ (Chi-Square Statistic) จะนำผลการตอบข้อสอบของกลุ่มผู้เข้าสอบระหว่าง 2 กลุ่ม กลุ่มหนึ่งเรียกว่า “กลุ่มเปรียบเทียบ” (Focal Group) และอีกกลุ่มหนึ่งเรียกว่า “กลุ่มอ้างอิง” (Reference Group) เอามาพิจารณาเปรียบเทียบสัดส่วนการตอบข้อสอบว่า 2 กลุ่ม นี้ที่มีความสามารถในระดับเดียวกัน

ขนาดกลุ่มตัวอย่าง (Sample Size) หมายถึง จำนวนของผู้เข้าสอบตอบข้อคำถามทั้งกลุ่มอ้างอิงและกลุ่มเปรียบเทียบซึ่งผู้วิจัยกำหนดขึ้นเพื่อการวิจัยครั้งนี้ โดยกำหนดไว้ 3 ขนาด คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)

กลุ่มเปรียบเทียบ (Focal Group: F) หมายถึง กลุ่มผู้เข้าสอบที่คาดว่าจะเสียเปรียบจากการตอบข้อสอบทำหน้าที่ต่างกันของข้อสอบ เป็นกลุ่มที่แนวโน้มที่ตอบข้อสอบได้ถูกต้องต่ำกว่าผู้เข้าสอบอีกกลุ่มหนึ่งที่มีความสามารถในระดับเดียวกัน

กลุ่มอ้างอิง (Reference Group: R) หมายถึง กลุ่มผู้เข้าสอบที่คาดว่าจะได้เปรียบจากการตอบข้อสอบทำหน้าที่ต่างกันของข้อสอบ เป็นกลุ่มที่มีแนวโน้มที่ตอบข้อสอบได้ถูกต้องสูงกว่าผู้เข้าสอบอีกกลุ่มหนึ่งที่มีความสามารถในระดับเดียวกัน

ค่าอำนาจจำแนกของข้อสอบ (a-parameter) หมายถึง ค่าแสดงถึงข้อสอบที่มีความสามารถในการจำแนกคุณลักษณะของข้อสอบ ค่าอำนาจจำแนกข้อสอบมีค่าเข้าใกล้ 1 แสดงถึงข้อสอบมีอำนาจจำแนกได้ดี ค่าอำนาจจำแนกเป็นลบ แสดงว่าข้อสอบใช้ไม่ได้ ค่าอำนาจจำแนกเป็นศูนย์แสดงว่าข้อสอบจำแนกไม่ได้ และค่าอำนาจจำแนกของข้อสอบอยู่ระหว่าง

0.50 ถึง 2.50

ค่าความยากของข้อสอบ (b-parameter) หมายถึง ค่าความยากข้อสอบที่เข้าใกล้ 2 แสดงว่าข้อสอบมีความยากค่อนข้างดี ถ้าข้อสอบมีค่าความยากของข้อสอบที่ติดลบ แสดงว่าข้อสอบค่อนข้างง่าย และค่าความยากของข้อสอบอยู่ระหว่าง -2.50 ถึง 2.50

ค่าโอกาสการเดาของข้อสอบ (c-parameter) หมายถึง ค่าแสดงการเดาของข้อสอบมีโอกาสการเดาข้อสอบได้ถูกต้อง เป็น 0 และค่าการเดาข้อสอบมากกว่า 0.3 แสดงว่าค่อนข้างใช้ไม่ได้ ถ้าค่าโอกาสการเดาข้อสอบเป็น 0 แสดงว่าข้อสอบใช้ได้ ค่าโอกาสการเดาข้อสอบอยู่ระหว่าง 0 ถึง 1

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

ผู้วิจัยได้ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ด้วยวิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีชิปเทสท์ และวิธีแมนเทล-แฮนส์เซล โดยนำเสนอเป็น 6 ตอน ดังนี้

ตอนที่ 1 การทดสอบความสามารถพื้นฐานของผู้เรียนระดับชาติ (National Testing: NT) และงานวิจัยที่เกี่ยวข้อง

ตอนที่ 2 ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) และงานวิจัยที่เกี่ยวข้อง

ตอนที่ 3 การทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) และงานวิจัยที่เกี่ยวข้อง

ตอนที่ 4 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการทดสอบอัตราส่วนไลค์ลิฮูด และงานวิจัยที่เกี่ยวข้อง

ตอนที่ 5 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีชิปเทสท์ และงานวิจัยที่เกี่ยวข้อง

ตอนที่ 6 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมนเทล-แฮนส์เซล และงานวิจัยที่เกี่ยวข้อง

ตอนที่ 1 การทดสอบความสามารถพื้นฐานของผู้เรียนระดับชาติ (National Testing: NT) และงานวิจัยที่เกี่ยวข้อง

การทดสอบความสามารถพื้นฐานของผู้เรียนระดับชาติ (National Testing: NT) คือ การสอบประเมินคุณภาพการศึกษาระดับชาติขั้นพื้นฐาน เพื่อการประกันคุณภาพผู้เรียน ตรวจสอบกำกับดูแล และพัฒนาคุณภาพการศึกษาของโรงเรียน ของสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน (สพฐ.) จัดสอบโดย สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน ให้กับนักเรียนในช่วงชั้นที่ 1 หรือ ชั้นประถมศึกษาปีที่ 3 โดยพระราชบัญญัติการศึกษาแห่งชาติ พ.ศ. 2542 มาตรา 47 ได้กำหนดให้มีระบบการประกันคุณภาพการศึกษา เพื่อให้มีการพัฒนาคุณภาพและมาตรฐานการศึกษาในทุกระดับ และมาตรา 48 โดยให้หน่วยงานต้นสังกัดและสถานศึกษา จัดให้มีระบบการประกันคุณภาพการศึกษาภายในสถานศึกษา และการประกันคุณภาพภายใน ยังเป็นส่วนหนึ่งของกระบวนการบริหารการศึกษาที่ต้องดำเนินการอย่างต่อเนื่อง มีการจัดทำรายงานประจำปีเสนอต่อหน่วยงานต้นสังกัด หน่วยงานที่เกี่ยวข้องเปิดเผยต่อสาธารณชน และนำไปสู่การพัฒนาคุณภาพมาตรฐานการศึกษา เพื่อรองรับการประกันคุณภาพภายนอก การประเมินคุณภาพการศึกษาขั้นพื้นฐานจึงเป็นกระบวนการ เพื่อให้ข้อมูลที่จะเป็นตัวบ่งชี้ถึงผลสำเร็จในการจัดการศึกษา ซึ่งเป็นส่วนประกอบสำคัญส่วนหนึ่งในการประกันคุณภาพภายใน แนวทางการวัดและประเมินผลการเรียนรู้ เพื่อให้ได้ข้อมูลสารสนเทศที่แสดงพัฒนาการความก้าวหน้า และความสำเร็จทางการเรียน

ของผู้เรียน โดยสถานศึกษาต้องจัดให้มีการประเมินผลการเรียนให้เป็นมาตรฐานเดียวกัน ทั้งในระดับชั้นเรียน ระดับสถานศึกษา ระดับเขตพื้นที่การศึกษา และระดับชาติ ข้อมูลที่ได้จากการประเมินจะนำไปใช้ในการพัฒนาคุณภาพของผู้เรียน และคุณภาพการจัดการศึกษาของสถานศึกษา และเพื่อเป็นสารสนเทศรองรับบริบทของการประเมินภายนอก

ดังนั้นการประเมินคุณภาพทางการศึกษา เป็นกลไกสำคัญประการหนึ่ง ที่สะท้อนให้เห็นถึงความสำเร็จของการพัฒนาคุณภาพการศึกษาตามเจตนารมณ์ของพระราชบัญญัติการศึกษาแห่งชาติรวมทั้งเป็นข้อมูลพื้นฐานประกอบการวางแผนและพัฒนาคุณภาพของหน่วยงานต่าง ๆ ที่เกี่ยวข้องกับการจัดการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐานเป็นหน่วยงานที่รับผิดชอบในการจัดการศึกษาตั้งแต่ระดับปฐมวัยจนถึงระดับมัธยมศึกษา จำเป็นต้องใช้ประโยชน์จากสารสนเทศผลการประเมินคุณภาพการศึกษาของสถานศึกษาในสังกัด ไปใช้ในการวางแผนเพื่อพัฒนาและปรับปรุงคุณภาพการศึกษาอย่างเป็นระบบและต่อเนื่อง เพื่อให้เกิดการพัฒนาคุณภาพผู้เรียน อีกทั้งการพัฒนาหลักสูตร กระบวนการจัดการศึกษาที่ส่งเสริมให้ผู้เรียนสามารถพัฒนาตามธรรมชาติ และเต็มตามศักยภาพของตนเองต่อไป

ในปีการศึกษา 2555 สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน โดยสำนักทดสอบทางการศึกษาได้กำหนดนโยบายจัดการทดสอบความสามารถพื้นฐานของผู้เรียนระดับชาติ (National Testing: NT) ในระดับชั้นประถมศึกษาปีที่ 3 ทั้ง 3 ด้าน คือ ด้านภาษา (Literacy Ability) ด้านคำนวณ (Numeracy Ability) และด้านเหตุผล (Reason Ability) ให้สอดคล้องกับจุดเน้นการพัฒนาคุณภาพผู้เรียน คือชั้นประถมศึกษาปีที่ 1-3 นักเรียนมีทักษะความสามารถในการอ่านออกเขียนได้ คิดเลขเป็น มีทักษะการคิดขั้นพื้นฐาน (สำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน กระทรวงศึกษาธิการ, 2555, หน้า 1) โดยมีกรอบการประเมินผลสัมฤทธิ์ทางการเรียนตามหลักสูตร ปีการศึกษา 2555 ซึ่งมีตัวชี้วัดด้านภาษา 6 ตัวชี้วัด ตัวชี้วัดด้านคำนวณ 5 ตัวชี้วัด และตัวชี้วัดด้านเหตุผล 4 ตัวชี้วัด รวม 15 ตัวชี้วัด และใช้เครื่องมือที่ใช้ในการประเมินผลสัมฤทธิ์ทางการเรียนของผู้เรียน คือ แบบทดสอบ โดยลักษณะของเครื่องมือเป็นปรนัยแบบเลือกตอบ มีจำนวนแบบทดสอบที่ใช้ในการสอบ แบ่งเป็น 3 ด้าน คือ ด้านภาษา มีแบบทดสอบ 30 ข้อ เวลา 50 นาที ด้านคำนวณ มีแบบทดสอบ 30 ข้อ เวลา 50 นาที และด้านเหตุผล มีแบบทดสอบ 30 ข้อ เวลา 50 นาที ผลการประเมินผลสัมฤทธิ์ทางการเรียนของผู้เรียนระดับชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2555 พบว่า ด้านเหตุผลมีคะแนนเฉลี่ยสูงสุด เท่ากับ 45.92 ด้านภาษามีคะแนนเฉลี่ยเท่ากับ 42.94 และด้านคำนวณมีคะแนนเฉลี่ย เท่ากับ 37.45 ซึ่งการประเมินคุณภาพผู้เรียนในระดับชาติ ระดับชั้นประถมศึกษาปีที่ 3 จำแนกออกเป็น 4 ภูมิภาค ได้แก่ ภาคเหนือ ภาคกลาง ภาคตะวันออกเฉียงเหนือ และภาคใต้ พบว่า ภาคตะวันออกเฉียงเหนือ เป็นภูมิภาคที่มีคะแนนค่าเฉลี่ยสูงกว่าคะแนนเฉลี่ยระดับประเทศ โดยมีคะแนนสูงสุดเท่ากับ 43.21 ภาคเหนือมีคะแนนค่าเฉลี่ย เท่ากับ 41.57 ภาคกลางมีคะแนนค่าเฉลี่ย เท่ากับ 41.17 และภาคใต้ มีคะแนนค่าเฉลี่ยเท่ากับ 40.89 โดยคะแนนเฉลี่ยระดับประเทศ เท่ากับ 42.10 (สำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน กระทรวงศึกษาธิการ, 2557, หน้า 4-6)

ผลการประเมินคุณภาพการศึกษาขั้นพื้นฐาน เพื่อการประกันคุณภาพผู้เรียน (Nation Test: NT) ชั้นประถมศึกษาปีที่ 3 พบว่า ด้านภาษา ด้านคำนวณ และด้านเหตุผล คะแนนเฉลี่ยต่ำ

กว่าร้อยละ 50 ทั้ง 3 ด้าน เนื่องจากปีการศึกษา 2555 สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐานได้มีการปรับกรอบการประเมินไปเป็น 3 ด้าน คือ ด้านภาษา (Literacy Ability) ด้านคำนวณ (Numeracy Ability) และด้านเหตุผล (Reason Ability) ซึ่งเดิมจะเป็นการประเมิน 3 กลุ่มสาระการเรียนรู้ คือ กลุ่มสาระการเรียนรู้ภาษาไทย กลุ่มสาระการเรียนรู้คณิตศาสตร์ และกลุ่มสาระการเรียนรู้วิทยาศาสตร์ ดังนั้นจึงทำให้นักเรียนครูผู้สอนมีการนำข้อสอบแบบเดิมมาให้นักเรียนทดลองทำเพราะยังไม่มีตัวอย่างข้อสอบแบบปรับใหม่ออกมาให้ทดลองทำ ส่งผลให้นักเรียนไม่คุ้นเคยกับการประเมินผลสัมฤทธิ์ในครั้งนี้ การประเมินผลสัมฤทธิ์ทางการเรียนของผู้เรียนจำแนกตามกลุ่มสถานศึกษาที่มีบริบทแตกต่างกัน

1. กลุ่มสถานศึกษาที่เป็นต้นแบบส่งเสริมสนับสนุนในการพัฒนาคุณภาพในภาพรวม ได้แก่ โรงเรียนดีศรีตำบล โรงเรียนในฝัน รุ่น 1-3 โรงเรียนมาตรฐานสากล โรงเรียนจุฬาภรณ์ โรงเรียนอัตราการแข่งขันสูง โรงเรียนเฉลิมพระเกียรติ และโรงเรียนไทยรัฐวิทยา สรุปได้ดังนี้

1.1 โรงเรียนดีศรีตำบล ผลการประเมินผลสัมฤทธิ์ทางการเรียนชั้นประถมศึกษาปีที่ 3 พบว่าคะแนนเฉลี่ยรวมต่ำกว่าระดับประเทศ เมื่อพิจารณาความสามารถ พบว่าทุกความสามารถมีคะแนนเฉลี่ยต่ำกว่าคะแนนเฉลี่ยระดับประเทศทุกด้าน เนื่องจากกระบวนการคัดเลือกโรงเรียนดีศรีตำบลในบางเขตพื้นที่ ยังมีข้อบกพร่อง เช่น ไม่ได้คัดเลือกจากโรงเรียนที่มีความพร้อมในแต่ละตำบลจริงๆ ระยะเวลาในจัดสรรงบประมาณโดยมุ่งเน้นทางด้านครุภัณฑ์ในการก่อสร้างอาคารสถานที่ จัดซื้อคอมพิวเตอร์ และเน้นด้านคุณธรรมจริยธรรมมากกว่า แต่ไม่เน้นไปที่การขาดแคลนครูตามสาขาวิชา และไม่เน้นทางวิชาการจึงทำให้คะแนนเฉลี่ยต่ำกว่าคะแนนเฉลี่ยระดับประเทศทุกกลุ่มสาระการเรียนรู้

1.2 โรงเรียนในฝัน รุ่นที่ 1-3 ผลการประเมินผลสัมฤทธิ์ทางการเรียนระดับชั้นประถมศึกษาปีที่ 3 ของกลุ่มโรงเรียนในฝัน พบว่า โรงเรียนในฝันรุ่นที่ 2 คะแนนเฉลี่ยรวมสูงกว่าคะแนนระดับประเทศ และโรงเรียนในรุ่นที่ 1,3 คะแนนเฉลี่ยรวมต่ำกว่าคะแนนระดับประเทศ เนื่องจากโรงเรียนในรุ่นที่ 3 ได้ผ่านการประเมินรับรองเป็นต้นแบบโรงเรียนในฝันแล้ว จึงทำให้ไม่ต่อเนื่องในการส่งเสริมการจัดการศึกษารวมไปถึงการนิเทศที่รับผิดชอบโรงเรียนในฝันยังทำหน้าที่รับผิดชอบโรงเรียนดีศรีตำบลอีกด้วยจึงทำให้การบริหารจัดการไม่ดีเท่าที่ควรทำให้โรงเรียนในฝันรุ่นที่ 3 มีผลคะแนนเฉลี่ยรวมต่ำกว่าระดับประเทศ

1.3 โรงเรียนมาตรฐานสากล ผลการประเมินผลสัมฤทธิ์ทางการเรียนระดับชั้นประถมศึกษาปีที่ 3 มีคะแนนเฉลี่ยรวมสูงกว่าระดับประเทศ เท่ากับ 47.18 พบว่า มีคะแนนเฉลี่ยสูงกว่าคะแนนเฉลี่ยระดับประเทศทุกด้าน เนื่องจากการคัดเลือกโรงเรียนเพื่อพัฒนาและยกระดับเป็นโรงเรียนมาตรฐานสากล โรงเรียนจะมีการคัดเลือกนักเรียนที่สติปัญญาความพร้อมเข้าสู่โรงเรียน อีกทั้งผู้บริหาร ครูผู้สอนได้รับการพัฒนาเพื่อเพิ่มสมรรถนะ และมีภาคีเครือข่ายรวมไปถึงโรงเรียนร่วมพัฒนาทั้งในประเทศและต่างประเทศ จึงส่งผลให้มีคะแนนเฉลี่ยสูงกว่าระดับประเทศทุกกลุ่มสาระการเรียนรู้

1.4 โรงเรียนอัตราการแข่งขันสูง ผลการประเมินผลสัมฤทธิ์ทางการเรียนระดับชั้นประถมศึกษาปีที่ 3 พบว่า ความสามารถทั้ง 3 ด้าน มีคะแนนเฉลี่ยรวมสูงกว่าระดับประเทศ เนื่องจากโรงเรียนมีกระบวนการการคัดเลือกนักเรียนที่เข้าศึกษามีการแข่งขันสูงและเป็นนักเรียน

ที่มีผลการเรียนที่ดีถึงดีมาก และโรงเรียนก็เน้นการพัฒนาทางด้านวิชาการ จึงส่งให้คะแนนเฉลี่ยสูงกว่าระดับประเทศ

1.5 โรงเรียนเฉลิมพระเกียรติ ผลการประเมินผลสัมฤทธิ์ทางการเรียนในระดับชั้นประถมศึกษาปีที่ 3 มีคะแนนเฉลี่ยรวมสูงกว่าระดับประเทศ เท่ากับ 53.13 พบว่า ด้านภาษาและด้านเหตุผลมีค่าเฉลี่ย เท่ากับ 46.98 และ 46.07 สูงกว่าคะแนนเฉลี่ยระดับประเทศ และด้านคำนวณมีคะแนนเฉลี่ย เท่ากับ 37.07 ใกล้เคียงกับคะแนนเฉลี่ยระดับประเทศ

1.6 โรงเรียนไทยรัฐวิทยา ผลการประเมินผลสัมฤทธิ์ทางการเรียนในระดับชั้นประถมศึกษาปีที่ 3 มีคะแนนค่าเฉลี่ยรวมต่ำกว่าคะแนนเฉลี่ยระดับประเทศ เท่ากับ 40.96 พบว่ามีคะแนนเฉลี่ยต่ำกว่าคะแนนระดับประเทศในทุกด้าน เนื่องจากผลสัมฤทธิ์ในภาพรวมในรายกลุ่มสาระการเรียนรู้ต่ำกว่าระดับประเทศในทุกระดับชั้น ซึ่งโรงเรียนไทยรัฐวิทยามีจุดเน้นด้านคุณลักษณะของผู้เรียน ความซื่อสัตย์ เพื่อสร้างพลเมืองโลกที่ดี อาจทำให้โครงการยกระดับผลสัมฤทธิ์ของโรงเรียนไม่ดีเท่าที่ควร และบางส่วนเป็นโรงเรียนขนาดเล็ก ส่งผลให้ขาดความพร้อมหลายด้าน ครูผู้สอนไม่ตรงตามเอกวิชาที่สอน

2. กลุ่มสถานศึกษาต้นแบบส่งเสริมสนับสนุนในการพัฒนาคุณภาพการศึกษาเฉพาะด้านระดับชั้นประถมศึกษาปีที่ 3 คือ โรงเรียนส่งเสริมนิสัยรักการอ่าน (โรงเรียนต้นแบบ, โรงเรียนแกนนำ และโรงเรียนเครือข่าย) โรงเรียนยกระดับผลสัมฤทธิ์ทางการเรียนในวิชาวิทยาศาสตร์และคณิตศาสตร์ และโครงการยกระดับคุณภาพ โรงเรียนขนาดเล็กผลการทดสอบ ดังนี้

2.1 โรงเรียนส่งเสริมนิสัยรักการอ่าน (โรงเรียนต้นแบบ) ระดับชั้นประถมศึกษาปีที่ 3 พบว่า มีคะแนนเฉลี่ยสูงกว่าคะแนนเฉลี่ยระดับประเทศ ซึ่งมีค่าเฉลี่ยใกล้เคียงกับค่าเฉลี่ยระดับประเทศ เท่ากับ 42.10 และพบว่ามี 2 ด้าน ที่มีคะแนนค่าเฉลี่ยสูงกว่าระดับประเทศ คือ ด้านภาษา เท่ากับ 42.10 และด้านเหตุผล เท่ากับ 42.10

2.2 โรงเรียนส่งเสริมนิสัยรักการอ่าน (โรงเรียนแกนนำ) ระดับชั้นประถมศึกษาปีที่ 3 พบว่า มีคะแนนเฉลี่ยสูงกว่าคะแนนเฉลี่ยระดับประเทศ เท่ากับ 42.57 ซึ่งมีค่าใกล้เคียงกับค่าเฉลี่ยระดับประเทศ เท่ากับ 42.10 และมี 2 ด้านที่มีค่าคะแนนเฉลี่ยสูงกว่าระดับประเทศ คือ ด้านภาษา เท่ากับ 43.61 และด้านคำนวณ เท่ากับ 37.62 โรงเรียนที่มีวัตถุประสงค์และการบริหารจัดการเฉพาะด้านซึ่งเน้นด้านการอ่านส่งผลให้ผลสัมฤทธิ์ทางการเรียนสูงกว่าระดับชาติ

2.3 โรงเรียนส่งเสริมนิสัยรักการอ่าน (โรงเรียนเครือข่าย) ระดับชั้นประถมศึกษาปีที่ 3 พบว่า มีค่าเฉลี่ยต่ำกว่าคะแนนเฉลี่ยระดับประเทศ เท่ากับ 42.03 และมี 1 ด้านที่มีค่าเฉลี่ยสูงกว่าระดับประเทศ คือ ด้านเหตุผล เท่ากับ 46.15 ซึ่งอาจเป็นเพราะโรงเรียนที่เริ่มเข้าร่วมโครงการการบริหารจัดการ การสนับสนุนด้านสื่อ วัสดุ อุปกรณ์ บุคลากร และงบประมาณ และการสนับสนุนอย่างเต็มที่

2.4 โรงเรียนห้องสมุดมีชีวิต (โรงเรียนต้นแบบ) พบว่ามีคะแนนเฉลี่ยต่ำกว่าคะแนนเฉลี่ยระดับประเทศ เท่ากับ 41.87 เนื่องจากการนิเทศกำกับ ติดตามอย่างเป็นระบบ จะมี 1 ด้านเท่านั้นที่มีค่าเฉลี่ยสูงกว่าระดับประเทศ คือ ความสามารถด้านภาษา เท่ากับ 43.07 ซึ่งโรงเรียนอาจมีวัตถุประสงค์เฉพาะด้าน และเน้นกิจกรรม ส่งเสริมการอ่าน การใช้ภาษา อาจส่งผลให้มีคะแนนค่าเฉลี่ยเฉพาะด้านสูงกว่าระดับประเทศ

2.5 โรงเรียนห้องสมุดมีชีวิต (โรงเรียนแกนนำ) พบว่า มีคะแนนเฉลี่ยสูงกว่าคะแนนเฉลี่ยระดับประเทศ เท่ากับ 42.91 ในทุกด้าน อาจเป็นเพราะเป็นโรงเรียนที่มีวัตถุประสงค์เฉพาะด้านเน้นกิจกรรมต่างๆ ส่งเสริมการอ่าน การใช้ภาษา และระยะเวลาการประเมินความก้าวหน้าอย่างต่อเนื่อง เพื่อเลื่อนระดับเป็นโรงเรียนต้นแบบ จึงทำให้ส่งผลให้ค่าเฉลี่ยทุกด้านสูงกว่าระดับประเทศ

2.6 โรงเรียนห้องสมุดมีชีวิต (โรงเรียนเครือข่าย) พบว่า มีคะแนนเฉลี่ยต่ำกว่าคะแนนเฉลี่ยระดับประเทศ เท่ากับ 41.94 จะมีเพียง 1 ด้านที่มีค่าเฉลี่ยสูงกว่าระดับประเทศ คือ ความสามารถด้านเหตุผล เท่ากับ 46.02 อาจเป็นเพราะเป็นที่มีวัตถุประสงค์เฉพาะด้าน เน้นกิจกรรมส่งเสริมการอ่าน การใช้ภาษา และช่วงระยะเวลาของการประเมินความก้าวหน้าอย่างต่อเนื่อง เพื่อเลื่อนระดับเป็นโรงเรียนแกนนำ ทำให้ส่งผลให้ค่าเฉลี่ยด้านดังกล่าวสูงกว่าระดับประเทศ

2.7 โครงการยกระดับคุณภาพโรงเรียนขนาดเล็ก เป็นสถานศึกษาต้นแบบส่งเสริมสนับสนุนในการพัฒนาคุณภาพการศึกษาเฉพาะด้าน พบว่า มีคะแนนเฉลี่ยสูงกว่าคะแนนเฉลี่ยระดับประเทศ เท่ากับ 43.78 ค่าเฉลี่ยใกล้เคียงค่าเฉลี่ยระดับประเทศ เท่ากับ 42.10 เนื่องจากการสนับสนุน สื่อ วัสดุ อุปกรณ์ อาคารสถานที่ งบประมาณ การสร้างเครือข่าย และโรงเรียนที่ผ่านกระบวนการคัดเลือกและฝึกอบรมเพื่อพัฒนาบุคลากร ทำให้ส่งผลคะแนนเฉลี่ยสูงกว่าระดับประเทศ

2.8 โรงเรียนส่งเสริมวิทยาศาสตร์และคณิตศาสตร์ พบว่า มีค่าคะแนนเฉลี่ยรวมสูงสุด เท่ากับ 43.78 มีคะแนนใกล้เคียงกับค่าเฉลี่ยระดับประเทศ เท่ากับ 42.10 มีคะแนนเฉลี่ยความสามารถด้านภาษา ความสามารถด้านคำนวณ และความสามารถด้านเหตุผลสูงกว่าระดับประเทศ เนื่องจากการสนับสนุนในหลายด้าน โดยเฉพาะครูผู้สอนที่จบการศึกษาตรงตามสาขาวิชาเอก ทำให้มีความรู้ความสามารถเฉพาะด้าน ทั้งสนับสนุนด้านวิชาการ วัสดุ อุปกรณ์ที่ทันสมัย งบประมาณเพียงพอ

3. โรงเรียนที่มีขนาดของสถานศึกษาแตกต่างกัน ได้แก่ โรงเรียนที่มีขนาดเล็ก (มีจำนวนนักเรียน 0-120 คน) โรงเรียนที่มีขนาดกลาง (มีจำนวนนักเรียนอยู่ระหว่าง 121-300 คน) โรงเรียนที่มีขนาดใหญ่ (มีจำนวนนักเรียนอยู่ระหว่าง 301-500 คน) และโรงเรียนที่มีขนาดใหญ่พิเศษ (มีจำนวนนักเรียนตั้งแต่ 501) ผลสัมฤทธิ์ทางการเรียนระดับประถมศึกษาปีที่ 3 ของโรงเรียนที่มีขนาดต่างกัน โรงเรียนที่มีขนาดใหญ่พิเศษมีคะแนนสูงสุดในทุกด้าน รองลงมาเป็นโรงเรียนขนาดเล็ก โรงเรียนขนาดกลาง เนื่องจากโรงเรียนขนาดใหญ่พิเศษ โดยส่วนใหญ่เป็นโรงเรียนประจำจังหวัด จึงทำให้มีความพร้อมในทุกด้าน ส่วนโรงเรียนขนาดเล็ก มีนักเรียนน้อยทำให้ครูผู้สอนดูแลได้ทั่วถึง มีการสนับสนุนเป็นรูปธรรม และโรงเรียนขนาดกลาง มักประสบปัญหาบุคลากรมีจำนวนไม่เพียงพอนักเรียนมีจำนวนมาก จึงทำให้ส่งเสริมการจัดการเรียนรู้ค่อนข้างยาก (สำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน กระทรวงศึกษาธิการ, 2557, หน้า 53-81)

สรุป คือ จากผลการประเมินคุณภาพการศึกษาขั้นฐานพื้นฐาน เพื่อการประกันคุณภาพผู้เรียน ระดับชั้นประถมศึกษาปีที่ 3 พบว่า ความสามารถด้านภาษา ความสามารถด้านคำนวณ และความสามารถด้านเหตุผล มีคะแนนเฉลี่ยต่ำกว่าร้อยละ 50 ทั้ง 3 ด้าน เนื่องจากในปีการศึกษา 2555 สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐานได้ปรับเปลี่ยนกรอบการประเมินผล จากเดิม

ที่ประเมินผลสัมฤทธิ์ทางการเรียน 3 กลุ่มสาระการเรียนรู้ คือ ภาษาไทย คณิตศาสตร์ และวิทยาศาสตร์ เปลี่ยนเป็นการประเมินความสามารถด้านภาษา (Literacy Ability) ความสามารถด้านคำนวณ (Numeracy Ability) และความสามารถด้านเหตุผล (Reason Ability) จึงทำให้ไม่มีตัวอย่างข้อสอบเพื่อให้ครูผู้สอนนำไปเทียบเคียง ในการสร้างเครื่องมือวัดและประเมินผลผู้เรียน

งานวิจัยที่เกี่ยวข้องกับการทดสอบความสามารถพื้นฐานของผู้เรียนระดับชาติ (NT) มีดังนี้

วรพรรณ ศรีกล่ำ อนุ เจริญวงศ์ระยับ และนิคม นาคอ้าย (2559) ศึกษาปัจจัยพหุระดับที่ส่งผลต่อคะแนนการสอบประเมินคุณภาพการศึกษาระดับชาติ ด้านความสามารถทางภาษาการศึกษาของโรงเรียนที่มีผล NT ต่ำในจังหวัดพิษณุโลก มีวัตถุประสงค์เพื่อศึกษาปัจจัยพหุระดับได้แก่ ปัจจัยพหุระดับนักเรียนและระดับห้องเรียนที่ส่งผลคะแนนการสอบ NT ด้านความสามารถทางภาษาโรงเรียนที่มีคะแนน NT ต่ำในจังหวัดพิษณุโลก กลุ่มตัวอย่าง ได้แก่ นักเรียนจำนวน 1,260 คน และครูจำนวน 68 ห้องเรียน ข้อมูลที่เก็บรวบรวมข้อมูลโดยใช้แบบทดสอบ 2 ฉบับ (แบบสอบถามระดับนักเรียนและระดับห้องเรียน) แบบสอบถามระดับนักเรียน ประกอบด้วย 5 ตัวแปร แบบสอบถามวัดความรู้พื้นฐานเดิม แบบสอบถามวัดแรงจูงใจใฝ่สัมฤทธิ์ในการทำข้อสอบ NT แบบสอบถามวัดเจตคติต่อการเรียนภาษาไทยแบบสอบถามวัดสภาพทางบ้าน และแบบสอบถามวัดความเอาใจใส่ผู้ปกครองในการส่งเสริมการเรียน ซึ่งมีค่าความเชื่อมั่นเท่ากับ 0.92 แบบสอบถามระดับห้องเรียน ประกอบด้วย 2 ตัวแปร แบบสอบวัดคุณภาพการสอนครูภาษาไทยและแบบสอบถามวัดบรรยากาศในชั้นเรียน มีค่าความเชื่อมั่นเท่ากับ 0.87 ทำการวิเคราะห์ข้อมูล โดยการวิเคราะห์พหุระดับ (Multilevel Analysis) และผลการวิเคราะห์ที่สำคัญ พบว่า โรงเรียน ที่มีคะแนน NT ต่ำ 1) ในระดับนักเรียน ตัวแปรความรู้พื้นฐานเดิมส่งผลต่อคะแนนการสอบ NT ด้านความสามารถทางภาษาอย่างมีนัยสำคัญทางสถิติ 2) ในระดับห้องเรียนไม่มีตัวแปรอิสระใดที่ส่งผลต่อคะแนน NT ด้านความสามารถทางภาษา ตัวแปรความรู้พื้นฐานเดิมสามารถอธิบายความแปรปรวนของคะแนนการสอบ NT ด้านความสามารถทางภาษา ได้ร้อยละ 12.85

เอกลักษณ์ คล้ายสุบรรณ สังวรณ์ งัดกระโทก และนลินี ณ นคร (2559) ศึกษาโมเดลการวัดมูลค่าเพิ่มทางการศึกษาสำหรับวัดคุณภาพสถานศึกษาด้วยการใช้ผลรวมของผลสัมฤทธิ์ทางการเรียนและผลการประเมินและรับรองคุณภาพของโรงเรียน โดยมีวัตถุประสงค์เพื่อ 1) พัฒนาวิธีการวัดมูลค่าเพิ่มทางการศึกษาเพื่อใช้ประเมินคุณภาพสถานศึกษาด้วยการวัดมูลค่าเพิ่มจากผลสัมฤทธิ์ทางการเรียนและผลการประเมินและรับรองคุณภาพของโรงเรียน 2) ศึกษาความสอดคล้องของผลการประเมินคุณภาพสถานศึกษาด้วยการวัดมูลค่าเพิ่มที่พัฒนาขึ้น ซึ่งมีการกำหนดน้ำหนักของการรวมคะแนนแตกต่างกัน และ 3) เปรียบเทียบมูลค่าเพิ่มทางการศึกษาของสถานศึกษาที่มีบริบทต่างกัน การวิจัยนี้ใช้ข้อมูลทุติยภูมิ กลุ่มตัวอย่างที่ใช้ในการวิจัย คือ โรงเรียนประถมศึกษาในสังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน จำนวน 96 โรงเรียน และมีนักเรียนชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2557 จำนวน 7,988 คน ซึ่งได้จากการสุ่มแบบหลายขั้นตอน ตัวแปรที่ศึกษาประกอบด้วย ตัวแปรระดับนักเรียน ได้แก่ เพศ การศึกษาของผู้ปกครองความสัมพันธ์ในครอบครัว ผลการประเมินคุณภาพการศึกษาระดับชาติขั้นพื้นฐาน

ตัวแปรระดับสถานศึกษา ได้แก่ ผลการประเมินและรับรองคุณภาพของโรงเรียน ที่ตั้ง และขนาดของสถานศึกษา เครื่องมือที่ใช้ คือ แบบบันทึกข้อมูลพื้นฐานของโรงเรียน ผลการประเมินคุณภาพการศึกษาระดับชาติขั้นพื้นฐานและ ผลการประเมินคุณภาพสถานศึกษา รอบ 3 ผู้วิจัยวิเคราะห์ข้อมูลคะแนนคุณภาพการศึกษาที่เป็นผลรวมของคะแนนผลสัมฤทธิ์ทางการเรียนกับผลการประเมินและรับรองคุณภาพของโรงเรียนโดยมีการถ่วงน้ำหนักต่างกันสามโมเดล คือ 40:60, 50:50, และ 60:40 สถิติที่ใช้ในการวิเคราะห์ข้อมูล คือ การวิเคราะห์พหุระดับผลการวิจัยสรุปได้ว่า (1) โมเดลการวัดมูลค่าเพิ่มทางการศึกษาที่มีความกลมกลืนกับข้อมูลมากที่สุด คือ โมเดลที่ 1 (40:60) รองลงมาคือโมเดลที่ 2 (50:50) และ โมเดลที่ 3 (60:40) (2) ความสอดคล้องของผลการประเมินคุณภาพสถานศึกษาด้วยการวัดมูลค่าเพิ่มระหว่างโมเดลที่ 1 (40:60) กับโมเดลที่ 2 (50:50) มีความสอดคล้องกันมากที่สุด มีความสอดคล้อง 91.67% และรองลงมา ระหว่างโมเดลที่ 2 (50:50) กับโมเดลที่ 3 (60:40) มีความสอดคล้อง 88.54% และระหว่างโมเดลที่ 1 (40:60) กับโมเดลที่ 3 (60:40) มีความสอดคล้อง 80.21% ตามลำดับ และ(3) สถานศึกษาที่มีที่ตั้ง และขนาดต่างกันมีคะแนนมูลค่าเพิ่มทางการศึกษาไม่แตกต่างกัน

สุธาทิพย์ ตรีสิน และปิยะทิพย์ ประดุจพรม (2560) ได้ศึกษาการเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ชั้นประถมศึกษาปีที่ 3 ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR โดยมีวัตถุประสงค์เพื่อวิเคราะห์คุณภาพของแบบทดสอบระดับชาติ (NT) และตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR 1) วิเคราะห์คุณภาพของแบบทดสอบระดับชาติ ทั้ง 3 ด้าน 2) ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR และ 3) เปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธีการตรวจสอบ 3 วิธี ข้อมูลวิเคราะห์เป็นข้อมูลทุติยภูมิ จากผลการตอบแบบทดสอบระดับชาติของนักเรียนชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2556 จำนวน 9,600 คน ผลการวิจัยพบว่า แบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 มีค่าความยากของข้อสอบ (b) อยู่ในระดับค่อนข้างยากมีค่าอำนาจจำแนกของข้อสอบ (a) อยู่ในระดับที่สามารถจำแนกผู้สอบได้ดี และมีค่าโอกาสในการเดาของข้อสอบ (c) ไม่เกิน 0.3 และการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้ง 3 ด้าน จะเห็นได้ว่าเพศจะส่งผลให้เกิดการทำหน้าที่ต่างกันของข้อสอบ โดยเพศหญิงจะได้เปรียบในการตอบข้อสอบ ด้านภาษาและด้านเหตุผล

Kjellstrom and Pettersson (2005) ได้ศึกษาความรู้ความเข้าใจของหลักสูตรการศึกษานำไปทดสอบระดับชาติในวิชาคณิตศาสตร์ในประเทศสวีเดน จะเกี่ยวกับระบบการทดสอบของประเทศสวีเดน โดยมีการทดสอบระดับชาติ ที่เป็นการมุ่งเป้าหมายองค์ความรู้ ระบบการทดสอบที่สำคัญในหลักสูตรของการเรียนการสอนรายวิชาคณิตศาสตร์ ที่ได้อธิบายอิทธิพลต่อการทดสอบระดับชาติ ซึ่งได้เปลี่ยนแปลงการประเมินที่แตกต่างไปจากเดิม เพื่อไปสู่เป้าหมาย เกณฑ์ในการอ้างอิงการให้คะแนนและการแก้ปัญหาของนักเรียน ของกระบวนการเรียนการสอนรายวิชาคณิตศาสตร์ พบว่า หลักสูตรการศึกษาในการเรียนการสอนในรายวิชาคณิตศาสตร์มีผลต่อคะแนนระดับชาติ ดังนั้นสถานศึกษาจึงควรปรับปรุงหลักสูตรการศึกษาให้มีประสิทธิภาพยิ่งขึ้น

Brown, De Four-Babb, Bristol and Conrad (2014) ได้ศึกษาการทดสอบระดับชาติเกี่ยวกับคำพูดของครูในตรินิแดดและโตเบโก ซึ่งได้มีการกล่าวถึงข้อเสนอแนะของการทดสอบในตรินิแดดและโตเบโก รวมไปถึงขอบเขตที่ใช้ในการตัดสินใจของหลักสูตรที่ส่งผลกระทบต่อการเรียนรู้ของนักเรียน โดยกลุ่มตัวอย่างโดยจะประกอบไปด้วยครูประถมศึกษา จำนวน 133 คน แบ่งเป็น 79 คน จากโรงเรียนที่มีประสิทธิภาพต่ำ ส่วน 54 คน จากโรงเรียนประสิทธิภาพสูง และผู้บริหาร 10 คน ซึ่งผลการวิจัยเชิงปริมาณและข้อมูลเชิงคุณภาพ พบว่า มีครูจำนวนมากที่รู้สึกไม่สบายใจกับการตีความข้อมูลที่นำเสนอในรายงานที่เกี่ยวข้องกับการทดสอบระดับชาติ ซึ่งครูในโรงเรียนที่มีประสิทธิภาพสูง ที่ผ่านการทำงานร่วมกันเพื่อเป็นข้อมูลในการตัดสินใจเรื่องการเรียนการสอนในหลักสูตร ดังนั้นจึงมีความจำเป็นที่จะให้มีการฝึกอบรมครูในการใช้ข้อมูลการประเมินของปัญหา อื่นๆ ที่จะเกิดขึ้นจากข้อมูลและหัวข้อที่จะเป็นไปได้สำหรับการวิจัยเพิ่มเติม เพื่อการแสดงถึงสัญลักษณ์ของโรงเรียนที่มีประสิทธิภาพที่ดี และโรงเรียนที่ไม่มีประสิทธิภาพในการทดสอบระดับชาติ

สรุปจากการศึกษางานวิจัยที่เกี่ยวข้องการทดสอบระดับชาตินั้นมีความสำคัญกับการศึกษาเป็นอย่างมาก เนื่องจากการทดสอบระดับชาติมีเป้าหมายในการพัฒนาผู้เรียนเป็นสำคัญ ซึ่งจะมีการนำผลของการทดสอบระดับชาติไปวิเคราะห์ผลเพื่อให้มีการปรับปรุงหลักสูตรการเรียนการสอนให้สอดคล้องกับผู้เรียนรวมไปถึงการฝึกอบรมครูผู้สอนให้มีกระบวนการจัดการเรียนการสอนที่มีประสิทธิภาพมากขึ้น เพื่อให้ส่งผลให้กับผลสัมฤทธิ์ทางเรียนดีขึ้นอย่างมีประสิทธิภาพ

ตอนที่ 2 ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) และงานวิจัยที่เกี่ยวข้อง

ทฤษฎีการตอบสนองข้อสอบ หรือ IRT สามารถจำแนกได้เป็น 2 ประเภท ได้แก่ ทฤษฎีการตอบสนองข้อสอบแบบตรวจให้คะแนน 2 ค่า (Binary or Dichotomous IRT) ซึ่งเป็นโมเดลการตอบสนองข้อสอบที่ใช้กับการตรวจคะแนนรายข้อแบบ 2 ค่า เช่น ข้อสอบหรือข้อคำถามที่ตรวจให้คะแนนแบบ 0,1 (ตอบผิดได้ 0, ตอบถูกได้ 1) แบบถูก/ ผิด ใช่/ ไม่ใช่ และทฤษฎีการตอบสนองข้อสอบแบบตรวจให้คะแนนมากกว่า 2 ค่า (Polytomous IRT) ซึ่งเป็นโมเดลการตอบสนองข้อสอบที่ใช้กับการตรวจคะแนนรายข้อแบบมากกว่า 2 ค่า เช่น ข้อสอบหรือข้อคำถามมาตรประมาณค่า (Rating Scale) การตรวจข้อสอบแบบให้คะแนนความรู้บางส่วน (Partial Credit)

แนวคิดพื้นฐานของ IRT

โมเดลการวัด (Measurement Model) โมเดลการวัดในที่นี้เป็นโมเดลเชิงคณิตศาสตร์ (Mathematical Model) ซึ่งเป็นระบบความสัมพันธ์ระหว่างตัวแปรอิสระที่รวมกันอย่างเหมาะสมในการทำนายตัวแปรตาม

1) โมเดลการวัดแบบดั้งเดิม (CTT Model)

$$X_p = T_p + E_p \qquad (1)$$

เมื่อ X_p เป็นคะแนนที่สังเกตได้ของผู้สอบ (คะแนนดิบรวม) เกิดจากการรวมเชิงเส้นตรงของผลบวกของตัวแปรแฝง 2 ตัว ได้แก่ คะแนนจริงของผู้สอบ (T_p) คะแนนความคลาดเคลื่อน (E_p) โดยมีข้อตกลงเบื้องต้นของคะแนนความคลาดเคลื่อนว่า (1) ค่าเฉลี่ยของ E ของกลุ่มผู้สอบมีค่าเป็นศูนย์ (2) ค่า E ไม่มีความสัมพันธ์กับ T และ E อื่นๆ จากโมเดลดังกล่าวมีข้อจำกัดหลายด้านที่สำคัญได้แก่ (1) ค่า X เป็นค่าเฉพาะที่ได้จากการวัดแต่ละครั้ง ซึ่งได้จากข้อสอบที่มีค่าความยากและอำนาจจำแนกเฉพาะของฉบับที่ใช้วัดนั้น การเปรียบเทียบคะแนน (X หรือ T) ระหว่างแบบสอบที่ใช้วัดคุณลักษณะเดียวกัน จึงต้องอยู่บนพื้นฐานข้อตกลงเบื้องต้นของแบบสอบที่มีข้อสอบคู่ขนานกัน (Parallelism) (2) คะแนนรายข้อไม่ได้ เชื่อมโยงกับพฤติกรรมการตอบข้อสอบและคะแนนจริงของผู้สอบ ปัจจัยสำคัญเกี่ยวกับคุณลักษณะข้อสอบจึงถูกละเลยไปจากโมเดลการวัดค่าพารามิเตอร์ของข้อสอบ เช่น ค่าความยาก ค่าอำนาจจำแนก เป็นต้น จึงแปรผันไปตามกลุ่มผู้สอบ

2) โมเดลการตอบสนองข้อสอบ (IRT Models)

โมเดลการวัดเป็นระบบความสัมพันธ์ระหว่างตัวแปรอิสระที่รวมกันสำหรับทำนายตัวแปรตามสำหรับโมเดลการตอบสนองของข้อสอบ ตัวแปรอิสระประกอบด้วย ตัวแปรแฝง คือความสามารถที่แท้จริงของผู้สอบ (θ) และคุณลักษณะของข้อสอบ (B) หรือค่าพารามิเตอร์ของข้อสอบ (a,b,c) ส่วนตัวแปรอิสระเป็นตัวแปรที่สังเกตได้ คือ โอกาสการตอบข้อสอบได้ถูกต้อง

ทฤษฎีการตอบสนองข้อสอบเป็นทฤษฎีการวัดที่อธิบายความสัมพันธ์ระหว่างคุณลักษณะภายในหรือความสามารถที่มีอยู่ภายในตัวบุคคลกับพฤติกรรมการตอบสนองข้อสอบของบุคคลนั้น โดยจะใช้โค้งลักษณะข้อสอบ (Item Characteristic Curve: ICC) มีการกำหนดความยาก (b) อำนาจจำแนก (a) และโอกาสการเดาข้อสอบถูก (c) IRT จึงอยู่บนฐานความคิดที่สำคัญ 2 ประการ คือ 1) ผลการตอบข้อสอบหรือข้อคำถามของผู้ตอบ สามารถอธิบายได้ด้วยความสามารถที่มีอยู่ภายในของผู้ตอบ และ 2) ความสัมพันธ์ระหว่างผลการตอบข้อสอบกับความสามารถที่มีอยู่ภายในสามารถอธิบายได้ด้วยฟังก์ชันลักษณะข้อสอบ หรือโค้งลักษณะข้อสอบ (ICC) อันมีลักษณะเป็นฟังก์ชันทางคณิตศาสตร์ เรียกว่าฟังก์ชันโลจิส (Logistic Function) หรือ ใกล้เคียงกับฟังก์ชันปกติสะสม (Normal Ogive Function)

ทฤษฎีการตอบสนองข้อสอบ พยายามอธิบายความสัมพันธ์ระหว่างคุณลักษณะระหว่างคุณลักษณะภายในหรือความสามารถที่มีอยู่ภายในตัวบุคคล กับพฤติกรรมการตอบสนองข้อสอบของบุคคลนั้นว่ามีโอกาสตอบข้อสอบถูกมากน้อยเพียงไร ทฤษฎีนี้มีพื้นฐานความเชื่อว่า พฤติกรรมการตอบสนองต่อข้อสอบของผู้สอบ ซึ่งเป็นสิ่งที่สังเกตได้โดยตรงว่าถูกหรือผิดจะถูกกำหนดโดยคุณลักษณะภายในหรือความสามารถที่อยู่ภายในตัวบุคคล ซึ่งเป็นสิ่งที่ไม่สามารถสังเกตได้โดยตรง ทฤษฎีนี้ได้โดยอธิบายความสัมพันธ์ดังกล่าวในรูปของฟังก์ชันคณิตศาสตร์ หรือโมเดลที่แสดงความสัมพันธ์ระหว่างระดับความสามารถ คุณลักษณะ ของข้อสอบและโอกาสของการตอบข้อสอบได้ถูก ที่เรียกว่า ฟังก์ชันการตอบสนองข้อสอบ ซึ่งมีลักษณะความสัมพันธ์เป็นแบบฟังก์ชันโลจิสหรือฟังก์ชันปกติสะสม

ฟังก์ชันการตอบสนองข้อสอบสามารถนำมาใช้ศึกษาความสัมพันธ์ระหว่างความน่าจะเป็นในการตอบข้อสอบแต่ละข้อได้ถูกต้อง [Pi (θ)] กับระดับความสามารถของผู้สอบที่วัดได้โดยแบบสอบฉบับนั้น (θ) เมื่อนำมาเขียนเป็นกราฟได้โค้งลักษณะข้อสอบ (Item Characteristic Curve: ICC)

โค้งลักษณะข้อสอบมีได้หลายลักษณะขึ้นอยู่กับโมเดล (Model) หรือแบบจำลองที่ใช้อธิบายความสัมพันธ์ดังกล่าว โมเดลที่นิยมใช้กันคือ โมเดลแบบหนึ่งพารามิเตอร์ (One-Parameter Model) โมเดลแบบสองพารามิเตอร์ (Two- Parameter Model) และโมเดลแบบสามพารามิเตอร์ (Three-Parameter Model)

โมเดลการตอบสนองข้อสอบ (Item Response Model)

คุณลักษณะของโมเดลการตอบสนองข้อสอบเป็นระบบความสัมพันธ์ระหว่างโอกาสตอบข้อสอบถูก (P_i) กับความสามารถที่มีอยู่ภายในผู้ตอบ (θ) ในรูปของโค้งลักษณะข้อสอบ (ICC) ซึ่งมีลักษณะเป็นฟังก์ชันโลจิส (Logistic Function) หรือฟังก์ชันปกติสะสม (Normal Ogive Function) บางครั้ง อาจเรียกว่า โมเดลโลจิส หรือ โมเดลปกติสะสม

โมเดลปกติสะสม ใช้ฟังก์ชันปกติสะสมแสดงความสัมพันธ์ระหว่างผลการตอบข้อสอบกับความสามารถของผู้สอบ ส่วนโมเดลโลจิสฟังก์ชันโลจิสแสดงความสัมพันธ์ระหว่างผลการตอบกับความสามารถดังกล่าว ซึ่งฟังก์ชันทั้งสองให้ผลลัพธ์ของการประมาณค่าใกล้เคียงกันมาก แต่ฟังก์ชันโลจิสมีลักษณะของสูตรทางคณิตศาสตร์ และวิธีคำนวณง่ายและสะดวกกว่า โมเดลโลจิสยังมีความสามารถทนทานต่อความคาดเคลื่อนที่เกิดขึ้นกับผู้สอบที่มีความสามารถสูงจะตอบข้อสอบได้ดีกว่าจึงทำให้โมเดลโลจิสเป็นที่นิยมกันมากในการนำไปใช้จริง

พารามิเตอร์ในทฤษฎีการตอบสนองข้อสอบ

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

1. พารามิเตอร์ข้อสอบ (Item Parameter)

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

1.1 ค่าความยากของข้อสอบ (b)

ค่าความยากของข้อสอบได้มาจากค่าความสามารถที่ตรงจุดเปลี่ยนโค้ง (Inflexion Point) ซึ่งเป็นจุดที่โค้งมีความชันมากที่สุด หรือมีความหมายอีกนัยหนึ่งก็คือ ผู้สอบที่มีความสามารถถึงระดับ ณ จุดเปลี่ยนโค้งข้อสอบนั้น จะมีโอกาสตอบข้อสอบนี้ถูกอยู่ 0.5 หรือในทางปฏิบัติจากจุดบนแกน y ที่แสดงถึงตำแหน่งโอกาสในการตอบข้อสอบข้อนี้ถูกมีอยู่ 0.5 ถ้าลากเส้นตั้งขนานกับแกน x จนพบกับเส้นโค้ง ซึ่งจะเป็นจุดเปลี่ยนโค้งด้วยนั้น ในทางตรงกันข้าม เมื่อลากเส้นตั้งฉากจากจุดดังกล่าวให้มาจรดแกน x ค่าที่วัดได้ในแกน x คือ ค่าความยากของข้อสอบข้อนั้น ๆ ซึ่งข้อสอบทั้งหมดในแบบทดสอบฉบับหนึ่ง ๆ ที่นำมาวิเคราะห์จะมีค่าความยากของข้อสอบกระจายอยู่ในแกน x จากค่า $-\infty$ ถึง ∞ แต่ในทางปฏิบัติ นิยมใช้ช่วง -3 ถึง +3 และแบบทดสอบทั่วไป มักจะมีค่า b อยู่ระหว่าง -2.5 ถึง +2.5 ถ้าค่า bi เข้าใกล้ -2.5 แสดงว่าข้อสอบง่าย ตรงกันข้ามถ้าค่า b อยู่ใกล้ +2.5 แสดงว่าข้อสอบยาก ในกรณีที่เป็นรูปแบบที่มีค่าพารามิเตอร์ 3 ตัว หรือคำนึงถึงโอกาส

ในการเดาคำตอบ (c) ค่าความยากซึ่งเริ่มต้นจากจุดบนแกน y นั้น จะใช้จุดตั้งต้นตรงที่ค่าโอกาสในการตอบข้อสอบถูก กล่าวคือ ค่าความยากมีค่าเริ่มต้นที่ $\frac{1-c}{2}$

1.2 ค่าอำนาจจำแนกของข้อสอบ (a)

ค่าอำนาจจำแนกของข้อสอบ (a) เป็นสัดส่วนกับค่าความชัน (Slope) ของ pi (θ) ที่จุดเปลี่ยนโค้ง หรือที่จุด θ = b ซึ่งทฤษฎีค่าอำนาจจำแนกของข้อสอบ จะมีค่าอยู่ในช่วง $-\infty$ถึง ∞ แต่ในทางการนำมาใช้ประโยชน์นั้น ข้อสอบข้อใดที่มีค่า a ติดลบ แสดงว่าข้อสอบข้อนั้นไม่ดี และควรจะต้องถูกตัดออกไป ส่วนข้อสอบที่มีค่า a สูงขึ้น แสดงว่า ความน่าจะเป็นของการตอบข้อสอบข้อนั้นเพิ่มขึ้นเมื่อระดับความสามารถของผู้สอบสูงขึ้น ตามปกติ ค่า a มีค่าไม่เกิน +2.5 ในทางปฏิบัตินิยมใช้ข้อสอบที่มีค่า a อยู่ระหว่าง +0.5 ถึง +2.5

1.3 ค่าโอกาสการเดาของข้อสอบ (c)

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

1.4 ความรอบคอบ (γ)

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

2. พารามิเตอร์ผู้สอบ

พารามิเตอร์ผู้สอบ (θ) เป็นระดับความสามารถของผู้สอบ (θ) ที่ประมาณได้จากโมเดลตามทฤษฎีการตอบสนองข้อสอบ นิยมปรับให้เป็นคะแนนมาตรฐานที่มีค่าเฉลี่ยเป็น 0 และส่วนเบี่ยงเบนมาตรฐานเป็น 1 ซึ่งพารามิเตอร์ผู้สอบมีค่าระหว่าง $-\infty$ ถึง ∞ แต่ส่วนใหญ่จะมีค่าอยู่ในช่วง-3.0 ถึง +3.0 ค่าที่เป็นลบ แสดงว่าผู้สอบมีความสามารถต่ำ และค่าที่เป็นบวก แสดงว่าผู้สอบมีความสามารถสูง สรุปได้ว่า พารามิเตอร์ตามทฤษฎีการตอบสนองข้อสอบจำแนกเป็น 2 ชนิด คือ พารามิเตอร์ข้อสอบ และพารามิเตอร์ผู้สอบ ซึ่งพารามิเตอร์ข้อสอบประกอบด้วยพารามิเตอร์ความยากของข้อสอบ พารามิเตอร์อำนาจจำแนกของข้อสอบ พารามิเตอร์โอกาสการเดาของข้อสอบ และความรอบคอบ ส่วนพารามิเตอร์ผู้สอบเป็นพารามิเตอร์ที่แสดงระดับความสามารถของผู้สอบ ซึ่งข้อตกลงเบื้องต้น และพารามิเตอร์ ที่กล่าวมานี้ มีความหมายเด่นชัดในกรณีที่ข้อสอบนั้นให้คะแนนแบบสองค่าในการประยุกต์ทฤษฎีเพื่อใช้กับข้อสอบที่ให้คะแนนแบบมากกว่าสองค่า ข้อตกลงเบื้องต้นทั้งหมดก็เทียบเคียงในทำนองเดียวกันแตกต่างกันเพียงรายละเอียดปลีกย่อยเกี่ยวกับเงื่อนไข

เฉพาะของแต่ละโมเดลเท่านั้นโมเดลการตอบสนองข้อสอบแบบตรวจให้คะแนนสองค่า Dichotomous IRT Models) เนื่องจากข้อมูลที่ได้จากการสอบมีหลายลักษณะ ได้แก่ ข้อมูลแบบมีสองค่า (Dichotomous) และข้อมูลแบบมีมากกว่าสองค่า (Polytomous) ดังนั้น จึงมีการพัฒนารูปแบบเพื่อให้สอดคล้องกับลักษณะของข้อมูลดังกล่าวขึ้นมากมาย แต่สำหรับข้อมูลที่เป็นแบบมี 2 ค่ารูปแบบที่นิยมใช้เป็นรูปแบบโลจิสติก (Logistic Model) ซึ่งแตกต่างไปตามจำนวนพารามิเตอร์ที่ใช้ในแต่ละรูปแบบ มีรายละเอียด ดังนี้

โมเดลการตอบสนองข้อสอบแบบ 1 พารามิเตอร์ (One-Parameter Model) รูปแบบนี้บางครั้งเรียกว่า รูปแบบราสช์ (Rasch Model) คือ โมเดลที่มีการแปรเปลี่ยนค่าพารามิเตอร์เพียงพารามิเตอร์ค่าความยากของข้อสอบ (b) เพียงอย่างเดียวโค้งคุณลักษณะข้อสอบสามารถเขียนสมการดังนี้ (Hambleton et al., 1991, pp. 12-14)

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

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

b_i แทน ค่าพารามิเตอร์ความยากของข้อสอบข้อที่ i

e แทน ค่าคงที่มีค่าเท่ากับ 2.718

n แทน ลำดับข้อสอบข้อที่ i

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

โมเดลการตอบสนองข้อสอบแบบ 2 พารามิเตอร์ (Two-Parameter Model)

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

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

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

b_i แทน ค่าพารามิเตอร์ความยากของข้อสอบข้อที่ i

a_i แทน ค่าพารามิเตอร์อำนาจจำแนกของข้อสอบข้อที่ i

D แทน ค่าสเกลองค์ประกอบ มีค่าเท่ากับ 1.7

$P_i(\theta)$ จากโค้งความถี่สะสมกับโค้งโลจิสติก จะมีค่าที่ต่างกันน้อยกว่า 0.01

สำหรับทุกค่าของ θ จากรูปแบบนี้อยู่บนข้อตกลงที่ว่าการเดาคำตอบจะไม่เกิดขึ้น ซึ่งถ้าจะเป็นเช่นนี้ได้ก็ต่อเมื่อค่าพารามิเตอร์ $a_i > 0$ (ข้อสอบที่มีความสัมพันธ์ด้านบวกระหว่างคะแนนจากการสอบกับความสามารถของผู้สอบที่วัดโดยแบบทดสอบนั้น) และค่าความน่าจะเป็นในการตอบข้อสอบได้ถูกจะลดลงถึงศูนย์เมื่อความสามารถลดลง

โมเดลการตอบสนองข้อสอบแบบ 3 พารามิเตอร์ (Three-Parameter Model)

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

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad (i = 1, 2, 3, \ldots, k) \qquad (4)$$

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

b_i แทน พารามิเตอร์ความยากของข้อสอบข้อที่ i

a_i แทน พารามิเตอร์อำนาจจำแนกของข้อสอบข้อที่ i

c_i แทน พารามิเตอร์โอกาสการเดาข้อสอบได้ถูกของข้อสอบข้อที่ i

D แทน ค่าสเกลองค์ประกอบ มีค่าเท่ากับ 1.7

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

1. ข้อตกลงเบื้องต้น (Assumptions) และการตรวจสอบ

ฟังก์ชันทางคณิตศาสตร์ที่ใช้ใน IRT ได้กำหนดความน่าจะเป็นในการตอบข้อสอบที่ตอบถูกนั้นจะขึ้นอยู่กับความสามารถของผู้เข้าสอบและคุณลักษณะของข้อสอบ ดังนั้นโมเดลการตอบสนองข้อสอบจึงอยู่บนพื้นฐานของความเชื่อหรือข้อตกลงเบื้องต้นหลายประการด้วยกัน ซึ่งลักษณะข้อมูลที่จะทำให้โมเดลสามารถนำไปใช้ได้อย่างเหมาะสม ข้อตกลงเบื้องต้นบางประการก็ไม่สามารถที่จะตรวจสอบไม่ได้โดยตรง แต่สามารถเก็บรวบรวมข้อมูลได้โดยทางอ้อมเพื่อนำมาสนับสนุนในข้อตกลงเบื้องต้นที่สำคัญของ IRT คือดังนี้

2. ความเป็นเอกมิติ (Undimensionality)

ข้อตกลงเบื้องต้นที่ใช้กันทั่วไปสำหรับ IRT คือ ข้อคำถาม/ ข้อสอบทุกข้อในเครื่องมือ/ แบบสอบมุ่งวัดเพียงคุณลักษณะเดียว หรือ ความสามารถเดียว (One Ability) ซึ่งเรียกว่า ความเป็นเอกมิติ (Unidimensionality)

การตรวจสอบความเป็นเอกมิติของเครื่องมือ หรือแบบสอบถาม สามารถทำได้โดยใช้เทคนิคทางสถิติ ได้แก่ การวิเคราะห์ตัวประกอบ (Factor Analysis) เพื่อคำนวณค่า ไอเกน (Eigen value) สำหรับศึกษาอัตราส่วนระหว่างค่าไอเกนของตัวประกอบแรกกับตัวประกอบถัดไป ถ้ามีอัตราส่วนที่สูงแสดงถึงเครื่องมือ หรือข้อสอบวัดคุณลักษณะเด่นเดียว (Single Dominant Factor) หรือทำการวิเคราะห์ให้เกิดความมั่นใจยิ่งขึ้น ด้วยการวิเคราะห์ตัวประกอบเชิงยืนยัน (Confirmatory Factor Analysis) เพื่อตรวจสอบยืนยันว่า เครื่องมือหรือแบบสอบมุ่งวัดเพียงคุณลักษณะเดียวหรือความสามารถเดียว

3. ความเป็นอิสระ (Local Independence)

แนวคิดเกี่ยวกับ "ความเป็นอิสระระหว่างข้อสอบและผู้สอบ" มีความเกี่ยวข้อง และเชื่อมโยงมาจาก "ความเป็นเอกมิติของแบบสอบ" ความเป็นอิสระระหว่างข้อสอบและผู้เข้าสอบ หมายถึง เมื่อมีการควบคุมความสามารถ (θ) ที่ส่งผลต่อการตอบข้อสอบ หรือให้ θ คงที่แล้ว ผลการตอบข้อสอบแต่ละข้อจะต้องเป็นอิสระจากกัน หรือเมื่อควบคุมอิทธิพลของ θ แล้ว ผลการตอบสนองข้อสอบรายข้อ ไม่มีความสัมพันธ์กัน นั่นคือ โมเดลการตอบสนองข้อสอบมีเพียง θ ปัจจัยเดียวเท่านั้นที่มีอิทธิพลต่อการตอบรายข้อ ความเป็นอิสระสามารถจำแนกพิจารณาเป็นความอิสระระหว่างข้อสอบและความเป็นอิสระระหว่างผู้สอบดังนี้

1) ความเป็นอิสระระหว่างข้อสอบ

เมื่อสุ่มผู้เข้าสอบ ซึ่งมีความสามารถ θ ขึ้นมา 1 คน ในการข้อสอบ k ข้อ ให้ U_j เป็นผลการตอบหรือคะแนนข้อสอบข้อที่ j หลังจากควบคุม θ ของผู้สอบแล้ว คะแนนผลการตอบของผู้สอบนั้นในแต่ละข้อไม่สัมพันธ์กัน

ถ้าผลการตอบข้อสอบรายข้อของผู้สอบคนเดียวกันเป็นอิสระจากกันความน่าจะเป็นของแผนการตอบข้อสอบ k ข้อของผู้สอบที่มีความสามารถ θ จะเท่ากับผลคูณระหว่างความน่าจะเป็นของผลการตอบข้อสอบแต่ละข้อ

2) ความเป็นอิสระระหว่างผู้สอบ

เมื่อสุ่มข้อสอบขึ้นมา 1 ข้อ ในการตอบข้อสอบของผู้สอบ n คน ให้ U_i เป็นผลการตอบหรือคะแนนข้อสอบของผู้สอบคนที่ i หลังจากควบคุม θ ของผู้สอบแต่ละคนแล้ว คะแนนผลการตอบข้อนั้นของผู้สอบแต่ละคนไม่สัมพันธ์กัน

ถ้าผลการตอบข้อสอบข้อเดียวกันของผู้สอบแต่ละคนเป็นอิสระจากกันความน่าจะเป็นของแบบแผนการตอบข้อสอบของผู้สอบ n คน จะเท่ากับ ผลคูณระหว่างความน่าจะเป็นของผลการตอบข้อนั้นของผู้สอบแต่ละคน

การตรวจสอบความเป็นอิสระระหว่างข้อสอบและผู้สอบสามารถกระทำได้ โดยการพิจารณาเมตริกซ์ ความแปรปรวนและความแปรปรวนร่วม (Variance - Covariance Matrix) หรือเมตริกซ์สหสัมพันธ์ (Correlation Matrix) ของคะแนนคำตอบรายข้อ สำหรับกลุ่มผู้สอบ

ที่มีช่วงคะแนนความสามารถเท่ากัน โดยค่านอกแนวทแยงมุมควรมีค่าต่ำหรือเข้าใกล้ 0

4. โมเดลการตอบสนองข้อสอบ (Item Response Models)

IRT อยู่บนพื้นฐานความเชื่อว่า ฟังก์ชันลักษณะข้อสอบ หรือโค้งลักษณะข้อสอบ (ICC) สามารถสะท้อนความสัมพันธ์จริงระหว่างความสามารถของผู้สอบกับลักษณะของข้อสอบและผลการตอบข้อสอบ โมเดลการตอบสนองข้อสอบเสนอ ICC ซึ่งเป็นฟังก์ชันโลจิสติก ด้วยรูปลักษณ์ที่แตกต่างกัน ตามจำนวนพารามิเตอร์ที่ใช้บรรยายลักษณะของข้อสอบ โมเดลการตอบสนองข้อสอบแบบตรวจให้คะแนน 2 ค่า (Dichotomous Item Models) ที่ใช้กันแพร่หลายได้แก่ โมเดลการตอบสนองข้อสอบแบบ 1 พารามิเตอร์ โมเดลการตอบสนองข้อสอบแบบ 2 พารามิเตอร์ โมเดลการตอบสนองข้อสอบแบบ 3 พารามิเตอร์ การเลือกใช้จึงขึ้นกับจุดมุ่งหมายของงานและธรรมชาติของข้อมูล

โมเดลการตอบสนองข้อสอบแบบ 1 พารามิเตอร์ มีข้อตกลงเบื้องต้นว่า ข้อสอบแต่ละข้อมีพารามิเตอร์ c=0 และพารามิเตอร์ a เท่ากัน แต่มีความแตกต่างกันเฉพาะพารามิเตอร์ b เท่านั้น โมเดลนี้จึงเหมาะสำหรับใช้กับข้อสอบอิงเกณฑ์ที่ไม่สลับซับซ้อน ข้อสอบที่ค่อนข้างเรียบง่ายสำหรับพัฒนาคลังข้อสอบที่มีความเป็นเอกพันธ์

โมเดลการตอบสนองข้อสอบแบบ 2 พารามิเตอร์ มีข้อตกลงเบื้องต้นว่าข้อสอบแต่ละข้อมีพารามิเตอร์ c=0 มีความแตกต่างกันของพารามิเตอร์ a และ b โมเดลนี้จึงเหมาะสำหรับใช้กับข้อสอบที่ต้องเติมคำตอบ หรือข้อสอบแบบเลือกตอบที่ไม่ยากมากนักและกลุ่มผู้สอบมีความพร้อมในการตอบ

โมเดลการตอบสนองข้อสอบแบบ 3 พารามิเตอร์ มีข้อตกลงเบื้องต้นว่า ข้อสอบแต่ละข้อมีความแตกต่างกันได้ทั้ง พารามิเตอร์ a, b, และ c โมเดลนี้จึงเหมาะสำหรับใช้กับข้อสอบเลือกตอบแบบทั่วไปข้อสอบแบบหลายตัวเลือก เนื่องจากผู้สอบสามารถเดาคำตอบได้

การตรวจสอบความเหมาะสมของโมเดลการตอบสนองข้อสอบกับข้อมูล (Model Data Fit) ไม่ว่าจะเลือกใช้โมเดลใดก็ตาม โมเดลจะต้องสอดคล้องกับข้อมูล เพื่อให้ผลการวิเคราะห์มีความถูกต้องและน่าเชื่อถือ การตรวจสอบความสอดคล้องควรพิจารณาทั้งความไม่แปรเปลี่ยนของค่าประมาณความสามารถ และความไม่แปรเปลี่ยนของค่าประมาณพารามิเตอร์ของข้อสอบ

ความไม่แปรเปลี่ยนของค่าประมาณความสามารถ ตรวจสอบได้โดยการเปรียบเทียบค่าประมาณความสามารถของผู้สอบที่ได้จากกลุ่มข้อสอบที่แตกต่างกัน เช่น กลุ่มข้อสอบง่าย กลุ่มข้อสอบยาก หรือกลุ่มข้อสอบจากคลังข้อสอบเดียวกัน แต่มีความครอบคลุมเนื้อหาแตกต่างกัน เป็นต้น ค่าประมาณความสามารถจะถือว่าไม่แปรเปลี่ยน เมื่อความแตกต่างเกิดขึ้นไม่เกินความคลาดเคลื่อนมาตรฐานของการประมาณค่า (ศิริชัย กาญจนวาสี, 2555, หน้า 77)

ความไม่แปรเปลี่ยนของค่าประมาณพารามิเตอร์ของข้อสอบ ตรวจสอบได้โดยเปรียบเทียบค่าประมาณพารามิเตอร์แต่ละตัวของข้อสอบที่ได้จากกลุ่มตัวอย่างประชากรผู้สอบหลายกลุ่ม เช่น กลุ่มผู้สอบชาย/ หญิง กลุ่มผู้สอบจำแนกตามภูมิภาค เป็นต้น ค่าประมาณพารามิเตอร์ของข้อสอบจะถือว่าไม่แปรเปลี่ยน เมื่อผลการพล็อตกราฟออกมาเป็นเส้นตรง โดยมีการกระจาย ไม่แตกต่างจากผลที่ได้จากกลุ่มตัวอย่าง 2 กลุ่ม ซึ่งเป็นกลุ่มสุ่มที่ทัดเทียมกัน

5. การสอบที่ไม่แข่งขันด้านเวลา (Nonspeeded Test Administration)

IRT ถือว่าความสามารถ (θ) เป็นปัจจัยสำคัญต่อผลการสอบ ความเร็วในการตอบจะต้องไม่มีอิทธิพลต่อผลการตอบ การจัดการสอบจึงไม่อยู่ในสถานการณ์ที่สอบแข่งขันด้วยเวลา การสอบจะต้องอยู่ในลักษณะที่ผู้สอบซึ่งมีความสามารถมีเวลาเพียงพอในการทำข้อสอบ (Power Test Administration)

การตรวจสอบถึงความเหมาะสม ของมิติด้านเวลา สำหรับการดำเนินการสอบสามารถพิจารณาได้จาก สัดส่วนหรือร้อยละของจำนวนผู้สอบที่ทำข้อสอบได้ครบทุกข้อ โดยผู้สอบส่วนใหญ่ (เช่น ร้อยละ 80 เป็นต้น) สามารถตอบข้อสอบได้ครบหรือเกือบครบทุกข้อ นอกจากนี้ควรพิจารณาเปรียบเทียบระหว่างความแปรปรวนของจำนวนข้อที่เว้น กับความแปรปรวนของจำนวนข้อที่ตอบผิด ถ้าอัตราส่วนของความแปรปรวนเข้าใกล้ 0 แสดงว่าการจัดการสอบเป็นไปตามข้อตกลงเบื้องต้นข้อนี้ (ศิริชัย กาญจนวาสี, 2555, หน้า 75-78)

คุณสมบัติของความไม่แปรเปลี่ยนของค่าพารามิเตอร์ (Invariance)

เมื่อ IRT Model สอดคล้อง (Fit) กับข้อมูลจะทำให้เกิดความไม่แปรเปลี่ยนของพารามิเตอร์ของข้อสอบ (Item Parameter) และพารามิเตอร์ความสามารถของผู้สอบ (Ability Parameter) ซึ่งเป็นคุณสมบัติของ IRT

1) ความไม่แปรเปลี่ยนของพารามิเตอร์ข้อสอบ (Item Parameter) ประกอบด้วย ค่าพารามิเตอร์ของข้อสอบไม่แปรเปลี่ยนไปตามกลุ่มผู้สอบ และ ICC ลักษณะเดียวกัน (a, b, c) สำหรับทุกกลุ่มความสามารถของผู้สอบ แสดงว่า ICC มีความคงที่ข้ามกลุ่มผู้สอบ

2) ความไม่แปรเปลี่ยนของพารามิเตอร์ความสามารถของผู้สอบ (Ability Invariance) ประกอบด้วย ค่าพารามิเตอร์ของผู้สอบไม่แปรเปลี่ยนไปตามชุดของข้อสอบ และเมื่อนำข้อสอบต่างชุด (ทุกข้อมุ่งวัดสิ่งเดียวกัน) เช่น ข้อสอบค่อนข้างง่าย กับข้อสอบชุดค่อนข้างยากมาสอบวัดผู้สอบกลุ่มเดียวกัน ค่า θ ที่ประมาณได้จากข้อสอบทั้ง 2 ชุด มีความแตกต่างกันไม่เกิน SEE แสดงว่าการประมาณค่าความสามารถมีความคงที่ข้ามชุดของข้อสอบ

สรุป คือ ทฤษฎีการตอบสนองข้อสอบเป็นทฤษฎีการวัดที่อธิบายความสัมพันธ์ระหว่างคุณลักษณะภายในหรือความสามารถที่มีอยู่ภายในตัวบุคคลกับพฤติกรรมการตอบสนองข้อสอบของบุคคลนั้น โดยจะใช้โค้งลักษณะข้อสอบ (Item Characteristic Curve: ICC) มีการกำหนดความยาก (b) อำนาจจำแนก (a) และโอกาสการเดาข้อสอบถูก (c) IRT จึงอยู่บนฐานความคิดที่สำคัญ 2 ประการ คือ 1) ผลการตอบข้อสอบหรือข้อคำถามของผู้ตอบ สามารถอธิบายได้ด้วยความสามารถที่มีอยู่ภายในของผู้ตอบ และ 2) ความสัมพันธ์ระหว่างผลการตอบข้อสอบกับความสามารถที่มีอยู่ภายใน สามารถอธิบายได้ด้วยฟังก์ชันลักษณะข้อสอบ หรือโค้งลักษณะข้อสอบ (ICC) อันมีลักษณะเป็นฟังก์ชันทางคณิตศาสตร์ เรียกว่าฟังก์ชันโลจิส (Logistic Function) หรือใกล้เคียงกับฟังก์ชันปกติสะสม (Normal Ogive Function)

งานวิจัยที่เกี่ยวข้องกับทฤษฎีการตอบสนองข้อสอบ (IRT) มีดังนี้

ชนะศึก นิชานนท์ (2554) ได้ศึกษาประสิทธิภาพของการประมาณค่าพารามิเตอร์แบบเบส์โดยใช้การสรุปอ้างอิงความน่าเชื่อถือของโมเดลการตอบสนองข้อสอบมีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าพารามิเตอร์ของวิธีการสรุปอ้างอิงความน่าเชื่อถือของโมเดลการ

ตอบสนองข้อสอบ (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 Bayesian GIRM ผู้วิจัยเป็นผู้พัฒนาขึ้นนอกจากนี้ยังศึกษาถึงอิทธิพลของขนาดกลุ่มตัวอย่างและจำนวนข้อสอบ รวมทั้งยังศึกษาความไว (Sensitivity) ของรูปแบบต่าง ๆ ต่อการกำหนดลักษณะการแจกแจงเริ่มแรกของค่าพารามิเตอร์ของข้อสอบและผู้สอบที่ส่งผลต่อประสิทธิภาพของวิธีการประมาณ ซึ่งวัดได้จากดัชนี 3 ประเภท ได้แก่ ความลำเอียง ในการประมาณค่า (Biased estimator) คำนวณจากการวิเคราะห์ค่าความคลาดเคลื่อนเฉลี่ย (Mean Average Deviation) ความไม่แน่นอนในการประมาณค่า (Uncertainty estimator) คำนวณจากการวิเคราะห์ค่าส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation) และประสิทธิภาพในการประมาณค่าองค์ประกอบความแปรปรวนซึ่งคำนวณจากการวิเคราะห์ระยะทางยุคลิด (Euclidean Distance) ข้อมูลที่ใช้ในการศึกษาครั้งนี้ใช้ข้อมูลจำลอง (Simulation) จากโปรแกรม R และทำการประมวลผลภายใต้จากการเขียนคำสั่งการประมวลด้วยโปรแกรม WinBUGS ด้วย Package R2 WinBUGS พบว่า เมื่อเปรียบเทียบประสิทธิภาพในการประมาณค่าของรูปแบบของวิธีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัดด้วยโมเดลการตอบสนองข้อสอบ (GIRM) พบว่า ความลำเอียงในการประมาณค่า รูปแบบที่ 1 กับรูปแบบที่ 4 ให้ค่าประสิทธิภาพสูงที่สุด โดยรูปแบบที่ 4 สามารถประมาณค่าพารามิเตอร์ได้เฉพาะลักษณะการแจกแจงเริ่มแรกของผู้สอบและข้อสอบแบบปกติ สำหรับความไม่แน่นอนในการประมาณค่า พบว่า รูปแบบที่ 4 ให้ค่าประสิทธิภาพสูงที่สุด สำหรับลักษณะการแจกแจงเริ่มแรกของผู้สอบและข้อสอบแบบปกติส่วนลักษณะการแจกแจงเริ่มแรกของค่าพารามิเตอร์ตัวใดตัวหนึ่งของผู้สอบหรือข้อสอบที่ไม่มีลักษณะการแจกแจงเริ่มแรกแบบปกติ พบว่า รูปแบบที่ 1 ให้ค่าประสิทธิภาพสูงที่สุด และเมื่อพิจารณาในด้านประสิทธิภาพขององค์ประกอบความแปรปรวนยุคลิด พบว่า รูปแบบที่ 2 ให้ค่าประสิทธิภาพสูงที่สุด การศึกษาอิทธิพลของขนาดกลุ่มตัวอย่างและจำนวนข้อสอบ พบว่า ขนาดกลุ่มตัวอย่าง ส่งผลต่อประสิทธิภาพในด้านความลำเอียงและความไม่แน่นอน ในการประมาณค่าในทุกรูปแบบและด้านประสิทธิภาพขององค์ประกอบความแปรปรวนยุคลิด พบว่า ขนาดกลุ่มตัวอย่างไม่ส่งผลต่อประสิทธิภาพในการประมาณค่า สำหรับความยาวแบบสอบ พบว่า ส่งผลต่อการวัดประสิทธิภาพในความลำเอียงในการประมาณค่า และการวิเคราะห์ประสิทธิภาพองค์ประกอบความแปรปรวนยุคลิด ในทุกรูปแบบ ส่วนความไม่แน่นอนในการประมาณค่า พบว่า การแจกแจงเริ่มแรกของผู้เข้าสอบไม่ส่งผลต่อประสิทธิภาพด้านความลำเอียง ส่วนความไม่แน่นอน ในการประมาณค่าและประสิทธิภาพขององค์ประกอบความแปรปรวนยุคลิด พบว่า การแจกแจงเริ่มแรกของข้อสอบ พบว่า ส่งผลต่อการวัดประสิทธิภาพความลำเอียงในการประมาณค่าความไม่แน่นอนในการประมาณค่าทุกรูปแบบ และส่งผลการวิเคราะห์ประสิทธิภาพองค์ประกอบความแปรปรวนยุคลิดเฉพาะในกรณีที่การแจกแจงเริ่มแรกของผู้สอบเป็นแบบแกมม่าเท่านั้น

ศาตรา แสนปัญญา (2555) ได้ศึกษาพัฒนาการความสามารถทางวิทยาศาสตร์ของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ไปยังชั้นมัธยมศึกษาปีที่ 4 โดยการศึกษาภาคตัดขวาง (Cross-Sectional Study) เมื่อปรับเทียบคะแนนแนวตั้ง (Vertical Equating) ตามทฤษฎีการตอบสนองข้อสอบ (IRT) แบบโลจิสติก 3 พารามิเตอร์ ร่วมกับวิธีค่าเฉลี่ยและซิกมา (Mean and Sigma Method) และสมการถดถอยโดยวิธีกำลังสองน้อยที่สุด (OLS) ผลที่ได้นำไปสร้างสมการโดยวิธีการถดถอย (Regression

Method) เพื่อทำนายความสามารถทางวิทยาศาสตร์ของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ในอนาคต หลังจากมีพัฒนาการแล้ว เครื่องมือที่ใช้ในการวิจัยครั้งนี้เป็นแบบวัดความสามารถทางวิทยาศาสตร์ของมัธยมศึกษาปีที่ 3 และมัธยมศึกษาปีที่ 4 ระดับชั้นละ 1 ฉบับ ฉบับละ 30 ข้อ มีความเชื่อมั่นของแบบทดสอบเท่ากับ 0.70 และ 0.71 ตามลำดับ และมีข้อสอบร่วมภายใน 6 ข้อ ผลการวิจัย พบว่า พัฒนาการทางวิทยาศาสตร์ (Growth Rate) ของนักเรียนชั้นมัธยมศึกษาปีที่ 3 เมื่อขึ้นชั้นมัธยมศึกษาปีที่ 4 กลุ่ม สูง ปานกลาง และต่ำ มีพัฒนาการเรียงลำดับจากมากไปน้อยตามลำดับ และสมการมีสัมประสิทธิ์การทำนาย (r^2) ของคะแนนสอบ (x) ความสามารถแฝง (θ) และคะแนนจริง (T) เท่ากับ 0.997 1.000 และ 0.979 ตามลำดับ โดยค่าสัมประสิทธิ์การทำนายของสมการทั้งสามสมการมีค่านัยสำคัญทางสถิติที่ระดับ 0.05

อนุชิต กลิ่นกำเนิด อรจิรา สิทธิ์ศักดิ์ และทัศนวรรณ ศูนย์กลาง (2555) ได้ศึกษาผลการประเมินระบบบริหารจัดการการเรียนรู้แบบปรับเหมาะ กรณีศึกษาเรื่อง องค์ประกอบของระบบสารสนเทศ มีวัตถุประสงค์เพื่อพัฒนาและหาประสิทธิภาพของระบบบริหารจัดการเรียนรู้แบบปรับเหมาะพัฒนาในรูปของเว็บแอพพลิเคชั่น (Web Application) นักเรียนเข้าเรียนเนื้อหาโดยแบ่งนักเรียนออกเป็น 3 ระดับ คือ กลุ่มเก่ง กลุ่มปานกลาง และกลุ่มอ่อน ทำการวัดค่าความสามารถของนักเรียนก่อนเรียนและหลังเรียนผ่านระบบบริหารจัดการเรียนรู้แบบปรับเหมาะในการวัดค่าความสามารถของนักเรียนใช้ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) และการประมาณค่าความสามารถ (θ) โดยใช้กลวิธีของเบย์ (Bayesian Updating) สำหรับครูผู้สอนระบบได้จัดเตรียมเครื่องมือสำหรับช่วยให้ผู้สอนเตรียมเนื้อหาบทเรียนให้เหมาะสำหรับนักเรียนแต่ละระดับโดยใช้เทคนิคสื่อหลายมิติแบบปรับตัว และจัดเตรียมแบบทดสอบโดยระบุค่าพารามิเตอร์ของข้อสอบเพื่อนำไปใช้ในการทำแบบทดสอบแบบปรับเหมาะได้ ผลการวิจัยระบบบริหารจัดการเรียนรู้แบบปรับเหมาะพบว่าค่าความสามารถทางการเรียนก่อนเรียนและหลังเรียนแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 โดยค่าความสามารถทางการเรียนหลังเรียนสูงกว่าก่อนเรียน และผลการประเมินความพึงพอใจของครูผู้สอนที่มีต่อระบบพบว่าครูผู้สอน มีความพึงพอใจต่อระบบในภาพรวมอยู่ในระดับความพึงพอใจมากที่สุด (X = 4.68, $S.D.$ = 0.47)

นุภาพรรณ ปลื้มใจ ปิยะทิพย์ ตินวร และโสพส สุขานนท์สวัสดิ์ (2558) ได้พัฒนาโปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์สำหรับการจัดสอบ O-NET ระดับชั้นมัธยมศึกษาปีที่ 6 โดยมีวัตถุประสงค์เพื่อวิเคราะห์คุณภาพข้อสอบ O-NET จัดทำคลังข้อสอบ O-NET และพัฒนาโปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ สำหรับการจัดสอบ O-NET ระดับชั้นมัธยมศึกษาปีที่ 6 จำนวน 8 กลุ่มสาระการเรียนรู้ ได้พัฒนาโปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ในรูปแบบของ Web Application การดำเนินการวิจัยมี 4 ขั้นตอน ดังนี้ 1) วิเคราะห์คุณภาพของข้อสอบ O-NET ตามทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) แบบ 3 พารามิเตอร์จำนวน 8 กลุ่มสาระการเรียนรู้ 2) จัดทำคลังข้อสอบ O-NET จำนวน 8 กลุ่มสาระการเรียนรู้ 3) พัฒนาโปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร และ 4) ประเมินความคิดเห็นของผู้ทดลองใช้โปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์โดยทำการทดสอบบนเว็บไซต์(www.onetcat.net/onetM6) กับนักเรียนระดับชั้นมัธยมศึกษาปีที่ 6 ที่กำลังศึกษาอยู่ในภาคเรียนที่ 2 ปีการศึกษา 2557 จำนวน 61 คน ผลการวิจัยปรากฏว่า 1. ข้อสอบ

O-NET ระดับชั้นมัธยมศึกษาปีที่ 6 จำนวน 8 กลุ่มสาระการเรียนรู้ที่ผ่านเกณฑ์การคัดเลือก จำนวน 1,197 ข้อ มีค่าอำนาจจำแนกของข้อสอบเฉลี่ย เท่ากับ 1.3693 ค่าความยากของข้อสอบเฉลี่ย เท่ากับ 0.8624 และค่าการเดาของข้อสอบเฉลี่ย เท่ากับ0.2024 แสดงให้เห็นว่า ข้อสอบที่อยู่ในคลังข้อสอบ O-NET ค่อนข้างยาก 2. คลังข้อสอบ O-NET ระดับชั้นมัธยมศึกษาปีที่ 6 สามารถบรรจุข้อสอบแบบหลายตัวเลือก ชนิด 4 ตัวเลือก ได้ไม่จำกัดขึ้นอยู่กับขนาดของServer 3. โปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ สำหรับการจัดสอบ O-NET ระดับชั้นมัธยมศึกษาปีที่ 6 มีความถูกต้องและปลอดภัยในการใช้งานเป็นที่ยอมรับของผู้เชี่ยวชาญ และนักเรียนที่ทดลองใช้โปรแกรมประเมินว่ามีความสะดวกในการนำไปใช้งาน

จารุจิตร สิทธิปรุ ปิยะทิพย์ ตินวร และโสพส สุขานนท์ (2559) ได้พัฒนาโปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์สำหรับการจัดสอบ O-NET ระดับชั้นมัธยมศึกษาปีที่ 3 โดยมีวัตถุประสงค์ เพื่อวิเคราะห์คุณภาพข้อสอบ O-NET จัดทำคลังข้อสอบ O-NET และพัฒนาโปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ สำหรับการจัดสอบ O-NET ระดับชั้นมัธยมศึกษาปีที่ 3 จำนวน 8 กลุ่มสาระการเรียนรู้ ซึ่งผู้วิจัยพัฒนาโปรแกรมการทดสอบ แบบปรับเหมาะด้วยคอมพิวเตอร์ในรูปแบบของ Web Application วิเคราะห์คุณภาพข้อสอบโดยใช้ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) แบบ 3 พารามิเตอร์ โดยใช้โปรแกรมสำเร็จรูป Xcalibre Version 4.1.7 ซึ่งแบ่งวิธีดำเนินการวิจัยออกเป็น 4 ขั้นตอน ดังนี้ 1) การวิเคราะห์คุณภาพข้อสอบ O-NET จำนวน 8 กลุ่มสาระการเรียนรู้ 2) การจัดทำคลังข้อสอบ O-NET จำนวน 8 กลุ่มสาระการเรียนรู้ ระหว่างปี พ.ศ. 2551–2553 3) การพัฒนาโปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ และ 4) การประเมินประสิทธิภาพของโปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ โดยสรุปผลระดับความคิดเห็นในการทดลองใช้งานโปรแกรม ซึ่งกลุ่มตัวอย่างเป็นนักเรียนระดับชั้นมัธยมศึกษาปีที่ 3 โรงเรียนอ่างศิลาพิทยาคม จังหวัดชลบุรี จำนวน 30 คนเครื่องมือที่ใช้ในการวิจัย ได้แก่ 1) ข้อสอบ O-NET จำนวน 8 กลุ่มสาระการเรียนรู้ และ 2) โปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ โดยทำการทดสอบใน website: www.onetcat.net/onetm3 และวิเคราะห์ระดับความคิดเห็นของนักเรียนที่ทดลองใช้โปรแกรม ด้วยสถิติพื้นฐาน ได้แก่ ค่าเฉลี่ย และส่วนเบี่ยงเบนมาตรฐาน ผลการวิจัยปรากฏว่า 1. การวิเคราะห์คุณภาพข้อสอบ O-NET 8 กลุ่มสาระการเรียนรู้ แสดงให้เห็นว่า ข้อสอบ O-NET ระดับชั้นมัธยมศึกษาปีที่ 3 มีค่าความยากของข้อสอบ (b) เฉลี่ยค้อนข้างยาก 2. คลังข้อสอบ O-NET สามารถบรรจุข้อสอบแบบหลายตัวเลือก (Multiple Choice) ชนิด 4 ตัวเลือกได้ไม้จำกัด ทั้งนี้ขึ้นอยู่กับขนาดของ Server โดยได้บรรจุข้อสอบ O-NET ที่วิเคราะห์ตามทฤษฎีการตอบสนองข้อสอบแบบ 3 พารามิเตอร์ และผ่านเกณฑ์การคัดเลือกข้อสอบ ทั้ง 8 กลุ่มสาระการเรียนรู้ ในระดับชั้นมัธยมศึกษาปีที่ 3 จำนวนทั้งหมด 469 ข้อ

สุชาดา กรเพชรปาณี ปิยะทิพย์ ตินวร และโสพส สุขานนท์สวัสดิ์ (2559) ได้พัฒนาโปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์สำหรับการจัดสอบ O-NET โดยมีวัตถุประสงค์เพื่อจัดทำคลังข้อสอบ O-NET และ พัฒนาโปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์สำหรับการจัดสอบ O-NET ระดับชั้นประถมศึกษาปีที่ 6 ชั้นมัธยมศึกษาปีที่ 3 และ ชั้นมัธยมศึกษา

ปีที่ 6 ระดับชั้นละ 8 กลุ่มสาระการเรียนรู้ การจัดทำคลังข้อสอบใช้ฐานข้อมูล MySQL และคัดเลือกข้อสอบ O-NET ของสถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน) ระหว่างปี พ.ศ. 2551-2553 ที่ผ่านการวิเคราะห์ข้อสอบตามทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ ในระดับชั้นประถมศึกษาปีที่ 6 จำนวน 258 ข้อ ชั้นมัธยมศึกษาปีที่ 3 จำนวน 469 ข้อ และชั้นมัธยมศึกษาปีที่ 6 จำนวน 1,197 ข้อโดยข้อสอบในคลังข้อสอบ O-NET อยู่ในระดับค่อนข้างยากพัฒนาโปรแกรมในรูปแบบของ Web Application โดยใช้ภาษา PHP พัฒนาโปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ตามแนวคิดของ Thompson and Weiss (2011) ผู้ใช้สามารถเข้าใช้โปรแกรมคอมพิวเตอร์สำหรับการจัดสอบ O-NET http://www.onetcat.net การประเมินความเหมาะสมในการนำไปใช้ของโปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ ในเดือนมิถุนายน พ.ศ. 2557 โดยนักเรียนระดับชั้นประถมศึกษาปีที่ 6 จำนวน 224 คน ชั้นมัธยมศึกษาปีที่ 3 จำนวน 432 คน และชั้นมัธยมศึกษาปีที่ 6 จำนวน 435 คน ปรากฏว่าการใช้งานของโปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์สำหรับการจัดสอบ O-NET อยู่ในเกณฑ์ดี เป็นที่พึงพอใจของนักเรียน

สุชาดา สกลกิจรุ่งโรจน์ เสรี ชัดแช้ม และสมพร สุทัศนีย์ (2559) ได้ศึกษาโปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ สำหรับมาตรวัดความสุขของคนไทย เป็นโปรแกรมที่พัฒนาขึ้นตามแนวคิดทฤษฎีการตอบสนองของข้อสอบร่วมกับการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ สำหรับวัดความอยู่ดีมีสุขเชิงอัตวิสัย การวิจัยนี้มีวัตถุประสงค์เพื่อตรวจสอบประสิทธิภาพของโปรแกรมในด้านความประหยัดและความรวดเร็วในการทดสอบ และศึกษาความสัมพันธ์ของผลการประมาณค่าความสุขที่ได้จากการทดสอบด้วยโปรแกรมการทดสอบ แบบปรับเหมาะด้วยคอมพิวเตอร์ สำหรับมาตรวัดความสุขของคนไทย กับการทดสอบด้วยข้อคำถามทั้งหมดในคลังข้อคำถาม เครื่องมือที่ใช้ในการวิจัย คือ โปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ สำหรับมาตรวัดความสุขของคนไทยที่พัฒนาขึ้นในงานวิจัยก่อนหน้านี้ ในรูปแบบเว็ปแอปพลิเคชั่น ซึ่งสามารถเข้าใช้งานผ่านทางเว็ปไซต์ http://www.thscat.com/test/ โดยสามารถเลือกทดสอบได้ 2 ลักษณะ ได้แก่ 1) การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ (THS-CAT) และ 2) การทดสอบด้วยข้อคำถามทั้งหมดในคลังข้อคำถาม (THS-ทุกข้อ) กลุ่มตัวอย่างที่ใช้ในการวิจัยมีจำนวนทั้งสิ้น 30 คน กลุ่มตัวอย่างจะได้รับการทดสอบทั้งสองรูปแบบ คือ การทดสอบด้วย THS-CAT และ THS-ทุกข้อ วิเคราะห์ข้อมูลด้วยสถิติบรรยาย และวิเคราะห์ความสัมพันธ์ของผลการประมาณค่าความสุขที่ได้จากการทดสอบทั้งสองรูปแบบ ด้วยการหาค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน ผลการวิจัย ปรากฏว่า โปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ สำหรับมาตรวัดความสุขของคนไทยสามารถดำเนินการสอบโดยใช้จำนวนข้อคำถามและระยะเวลาในการทดสอบน้อยกว่าการทดสอบด้วยข้อคำถามทั้งหมด ในคลังข้อคำถาม อีกทั้งมีความแม่นยำในการประมาณค่าสูงผลการประมาณค่าที่ได้จากการทดสอบทั้งสองรูปแบบมีความสัมพันธ์กันทางบวก

สุทธิวรรณ พีรศักดิ์โสภณ ปิยพงษ์ คล้ายคลึง และสมกิจ กิจพูนวงศ์ (2560) ได้ศึกษาคุณภาพแบบทดสอบความถนัดทางการเรียน SWUSAT ปีการศึกษา 2553-2555 โดยมีวัตถุประสงค์เพื่อ 1) วิเคราะห์คุณภาพของแบบทดสอบความถนัดทางการเรียน SWUSAT ตามทฤษฎีการทดสอบแบบมาตรฐานเดิมและวิธีทฤษฎีการตอบข้อสอบ 2) หาความสัมพันธ์ระหว่างค่าพารามิเตอร์ข้อสอบที่

วิเคราะห์โดยทฤษฎีการทดสอบแบบมาตรฐานเดิมและวิธีทฤษฎีการตอบข้อสอบ และ3) เปรียบเทียบจำนวนข้อสอบที่ผ่านเกณฑ์ใช้ได้ตามวิธีทฤษฎีแบบมาตรฐานเดิมและวิธีทฤษฎีการตอบข้อสอบข้อมูลที่นำมาวิเคราะห์คุณภาพได้จากผลการตอบแบบวัดความถนัดทางการเรียน SWUSAT ของนิสิตระดับปริญญาตรี ชั้นปีที่ 1 ระหว่างปีการศึกษา 2553-2555 จำนวน 27 ฉบับ การวิเคราะห์ข้อมูลใช้การวิเคราะห์ข้อสอบเป็นรายข้อตามแนวทฤษฎีการทดสอบแบบมาตรฐานเดิมและทฤษฎีการตอบข้อสอบและค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างพารามิเตอร์ของข้อสอบ ผลการวิจัยสรุปได้ดังนี้ ผลการวิเคราะห์ข้อสอบความถนัดทางการเรียน SWUSAT53-55 ด้วยวิธีการทดสอบแบบมาตรฐานเดิม พบว่า แบบทดสอบฉบับที่มีค่าความยากอยู่ในเกณฑ์ใช้ได้ทุกข้อ คือ N53-2, R54-1, S54-2, R55-2, S55-2 และ S55-3 แบบทดสอบฉบับที่มีค่าอำนาจจำแนกอยู่ในเกณฑ์ใช้ได้ทุกข้อ คือ S53-1, S53-2, S54-3, S54-4, S55-2 และ S55-3 นอกนั้นไม่พบแบบทดสอบฉบับใดที่มีค่าความยากหรือค่าอำนาจจำแนกอยู่ในเกณฑ์ใช้ได้ทุกข้อผลการวิเคราะห์ข้อสอบความถนัดทางการเรียน SWUSAT53-55 ด้วยวิธีทฤษฎีการตอบข้อสอบ พบว่า แบบทดสอบที่มีค่าความยากอยู่ในเกณฑ์ใช้ได้ทุกข้อมีเพียงฉบับเดียว คือ S55–3 แบบทดสอบที่มีค่าความยากเฉลี่ยต่ำสุดคือ S54-3 และความยากเฉลี่ยสูงสุดคือ V55-2 แบบทดสอบที่มีค่าอำนาจจำแนกอยู่ในเกณฑ์ใช้ได้ทุกข้อ คือ V53-1, V53-2, N53-1, R53-1, S53-1, V54-1, N54-1, R54-2, S54-2, V55-1, V55-2, N55-2, R55-2, S55-1, S55-2 และ S55-3 แบบทดสอบที่มีค่าอำนาจจำแนกเฉลี่ยต่ำสุดคือ V53-2 และที่มีค่าอำนาจจำแนกเฉลี่ยสูงสุดคือ N55-1 ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าความยากที่วิเคราะห์โดยวิธีทฤษฎีการทดสอบแบบมาตรฐานเดิมและวิธีทฤษฎีการตอบข้อสอบ พบว่าทุกฉบับมีความสัมพันธ์กันทางลบอย่างมีนัยสำคัญทางสถิติ แต่ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าอำนาจจำแนกที่วิเคราะห์โดยวิธีทฤษฎีการทดสอบแบบมาตรฐานเดิมและวิธีทฤษฎีการตอบข้อสอบ พบว่า มีความสัมพันธ์กันบางฉบับ คือ V53-2, R53-1, S53-1, S53-2, V54-2, S54-2, N54-2, R54-2, S54-1, V55-1, N55-2, R55-1 และ R55-2 นอกนั้นไม่พบว่ามีความสัมพันธ์กัน 3) การเปรียบเทียบจำนวนข้อที่ใช้ได้ระหว่างการวิเคราะห์ด้วยวิธีการทดสอบ แบบมาตรฐานเดิมและวิธีทฤษฎีการตอบข้อสอบ พบว่า ฉบับที่มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ คือ V53-2, S54-3 และ V54-1 นอกนั้นไม่พบความแตกต่างกัน

สุธาทิพย์ ตรีสิน และปิยะทิพย์ ประดุจพรม (2560) ได้เปรียบเทียบผลการตรวจสอบการทำาหน้าที่ต่างกันของข้อสอบในแบบทดสอบ ระดับชาติ ด้านภาษา ด้านคำานวณ และด้านเหตุผล ชั้นประถมศึกษาปีที่ 3 ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR โดยมีวัตถุประสงค์เพื่อวิเคราะห์คุณภาพของแบบทดสอบระดับชาติ (NT) และตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR ดังนี้ 1) วิเคราะห์คุณภาพของแบบทดสอบระดับชาติ ทั้ง 3 ด้าน 2) ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR และ3) เปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการตรวจสอบ 3 วิธีข้อมูลที่นำมาใช้ วิเคราะห์เป็นข้อมูลทุติยภูมิ จากผลการตอบแบบทดสอบระดับชาติของนักเรียนชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2556 จำนวน 9,600 คน ผลการวิจัยปรากฏว่า 1) แบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 มีค่าความยากของข้อสอบ (b) อยู่ในระดับค่อนข้างยากมีค่าอำนาจจำแนกของข้อสอบ (a) อยู่ในระดับที่สามารถจำแนกผู้สอบได้ดี และมี

ค่าโอกาสในการเดาของข้อสอบ (c) ไม่เกิน 0.3 2) การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้ง 3 ด้าน ชี้ให้เห็นว่า เพศส่งผลให้เกิดการทำหน้าที่ต่าง กันของข้อสอบ โดยเพศหญิงจะได้เปรียบในการตอบข้อสอบด้านภาษา และด้านเหตุผล ในขณะที่เพศชาย จะได้เปรียบในการตอบข้อสอบด้านคำนวณ โดยวิธี HGLM ตรวจพบข้อสอบทำหน้าที่ต่างกัน จำนวนมากที่สุด คิดเป็นร้อยละ 69 ของข้อสอบทั้งฉบับ รองลงมาคือ วิธี IRT-LR ร้อยละ 54 และวิธี MIMIC ร้อยละ 16 ตามลำดับ 3) การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ พบว่า วิธี HGLM ตรวจพบ DIF มากกว่า วิธี MIMIC ในด้านภาษา ด้านคำนวณ และด้านเหตุผล คิดเป็นร้อยละ 70, 36 และ 53 ตามลำดับ และวิธี HGLM ตรวจ พบ DIF มากกว่าวิธี IRT-LR ด้านภาษา และด้านคำนวณ คิดเป็นร้อยละ 37 และ 13 และวิธี IRT-LR ตรวจพบ DIF มากกว่าวิธี MIMIC ทั้ง 3 ด้าน คิดเป็นร้อยละ 33, 43 และ 40 ตามลำดับ ส่วนวิธี HGLM ตรวจพบ DIF น้อยกว่า วิธี IRT-LR ด้านคำนวณคิดเป็นร้อยละ 7 ($p<.05$)

Yildirim and Berberoglu (2009) ได้วิเคราะห์ข้อมูลการตัดสินใจและทางสถิติของรายการวิชาคณิตศาสตร์ PISA-2003 การเปรียบเทียบลักษณะของมนุษย์ในกลุ่มภาษาและวัฒนธรรมที่แตกต่างกันกลายเป็นสิ่งที่สำคัญมากขึ้นในแนวทางการประเมินผลการศึกษา ความสนใจที่เพิ่มขึ้นในการศึกษาเปรียบเทียบระหว่างประเทศผลลัพธ์ในด้านภาษาและกลุ่มทางวัฒนธรรมที่แตกต่างกันทำให้ความสนใจของนักวิจัย ในการศึกษาครั้งนี้มีสมมติฐานที่สร้างขึ้นเกี่ยวกับการทำงานของข้อสอบที่แตกต่างกัน (DIF) จะได้รับการศึกษาโดยใช้ชุดข้อมูลสำหรับการประเมินนักศึกษาต่างชาติ PISA 2003 ข้อสอบที่พบ DIF ผ่านการวิเคราะห์องค์ประกอบ Mantel-Haenzsel (MH) และการวิเคราะห์อัตราส่วนไลค์ลิฮูดความเป็นไปได้ในการตอบสนองข้อสอบ (IRT-LR) ได้รับการพิจารณาพร้อม ๆ กัน สำหรับการประเมินเนื้อหาที่ชาญฉลาด กลุ่มผู้ทดสอบ ด้านทักษะทางปัญญาที่วัดได้จากข้อบกพร่องด้านข้อผิดพลาดในการแปลและด้านการใช้คำเชิงปริมาณเป็นข้อผิดพลาดสำคัญสามข้อที่อาจก่อให้เกิด DIF ในการศึกษาของ PISA 2003 ผลการวิจัยพบว่าศึกษาการวิเคราะห์ข้อมูลการตัดสินใจและทางสถิติของวิชาคณิตศาสตร์ PISA-2003 พบว่า วิธี IRT-LR มีประสิทธิภาพในการตรวจสอบ DIF ได้ดีในด้านความสามารถทางคณิตศาสตร์ของโครงการประเมินผลนักเรียนนานาชาติ (PISA, 2003)

Muninsakorn, Tinnaworn, and Sukhanonsawat (2016) ได้พัฒนาโปรแกรมทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์สำหรับการจัดสอบ O-NET ระดับชั้นประถมศึกษาปีที่ 6 โดยมีวัตถุประสงค์เพื่อวิเคราะห์คุณภาพของข้อสอบ O-NET จัดทำคลังข้อสอบ O-NET และพัฒนาโปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ สำหรับการจัดสอบ O-NET ระดับชั้นประถมศึกษาปีที่ 6 จำนวน 8 กลุ่มสาระการเรียนรู้วิธีดำเนินการวิจัยแบ่งเป็น 4 ขั้นตอน ดังนี้ 1) วิเคราะห์คุณภาพของข้อสอบ O-NET จำนวน 8 กลุ่มสาระการเรียนรู้ 2) จัดทำคลังข้อสอบ O-NET จำนวน 8 กลุ่มสาระการเรียนรู้ 3) พัฒนาโปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ และ 4) ประเมินความคิดเห็นของผู้ทดลองใช้โปแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ ซึ่งเป็นนักเรียนระดับชั้นประถมศึกษาปีที่ 6 จังหวัดชลบุรี จำนวน 30 คน ผลการวิจัยปรากฏว่า 1. ข้อสอบ O-NET ระดับชั้นประถมศึกษาปีที่ 6 มีค่าความยากของข้อสอบ (b) ในระดับค่อนข้างยาก 2. คลังข้อสอบ O-NET ระดับชั้นประถมศึกษาปีที่ 6 บรรจุข้อสอบแบบหลายตัวเลือก (Multiple Choice) ชนิด 4 ตัวเลือก ได้โดยไม่จำกัด ขึ้นอยู่กับขนาดของ Server ซึ่งมีข้อสอบจำนวน 258 ข้อ

ที่ผ่านเกณฑ์การวิเคราะห์คุณภาพของข้อสอบตามโมเดลโลจิส แบบ 3 พารามิเตอร์ใน 8 กลุ่มสาระการเรียนรู้ 3. โปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ สำหรับการจัดสอบ O-NET ระดับชั้นประถมศึกษาปีที่ 6 มีความเหมาะสมของโปรแกรมในระดับมากที่สุด ไม่มีปัญหาด้านการนำไปใช้และเป็นที่ยอมรับของที่ผู้ทดลองใช้โปรแกรม

Moghadamzadeh, Salehi, and Khodaie (2011) ได้เปรียบเทียบฟังก์ชันสารสนเทศของข้อสอบ และฟังก์ชันสารสนเทศของแบบสอบในโมเดลการตอบสนองข้อสอบแบบ 1 พารามิเตอร์ 2 พารามิเตอร์และ 3 พารามิเตอร์ โดยเริ่มต้นจากการสร้างฟังก์ชันสารสนเทศภายใต้ทฤษฎีการตอบสนองข้อสอบ จากนั้นจึงใช้ข้อมูลการตอบข้อสอบที่ได้จากคลังข้อสอบซึ่งเป็นข้อสอบแบบหลายตัวเลือก ซึ่งได้มาจากข้อมูลการสอบของนักศึกษามหาวิทยาลัยอิหร่านที่ทำการทดสอบระดับชาติที่มีชื่อว่า “National Organization for Educational Testing (NOET)” ซึ่งเป็นการทดสอบคัดเลือกเข้าศึกษาต่อในระดับมหาวิทยาลัย ในสาขาด้านคณิตศาสตร์และฟิสิกส์ ใน ปี ค.ศ. 2009 ผู้วิจัยเลือกข้อมูลการทดสอบ 2,000 ชุด ด้วยวิธีการเลือกอย่างเป็นระบบ จากนั้นจึงวิเคราะห์ข้อมูลโดยใช้โปรแกรม SPSS และ BILOG ผลการศึกษาปรากฏว่าค่าสารสนเทศของข้อสอบ และค่าสารสนเทศของแบบสอบที่ได้จากการประมาณค่าด้วยโมเดลการตอบสนองข้อสอบแบบ 2 พารามิเตอร์ มีค่าสูงกว่าโมเดลการตอบสนองข้อสอบแบบ 1 พารามิเตอร์ และแบบ 3 พารามิเตอร์ นอกจากนี้ ผลการวิจัยยังชี้ให้เห็นว่าค่าฟังก์ชันสารสนเทศของข้อสอบ และค่าฟังก์ชันสารสนเทศของแบบสอบที่ได้จากโมเดลการตอบสนองข้อสอบแบบ 1 พารามิเตอร์ ไม่แตกต่างจากค่าที่ได้จากโมเดลการตอบสนองข้อสอบแบบ 3 พารามิเตอร์ ซึ่งหมายความว่าโมเดลการตอบสนองข้อสอบแบบ 2 พารามิเตอร์ ทำให้การทดสอบมีความแม่นยำมากกว่าโมเดลการตอบสนองข้อสอบแบบหนึ่งพารามิเตอร์และแบบสามพารามิเตอร์ โดยเหตุผลที่โมเดลการตอบสนองข้อสอบแบบสามพารามิเตอร์มีค่าสารสนเทศต่ำว่าอาจเนื่องมาจากพารามิเตอร์การเดาที่เพิ่มเข้ามาในโมเดลซึ่งทำให้เกิดความคลาดเคลื่อนในการประมาณค่าความสามารถได้มากกว่า เมื่อมีความคลาดเคลื่อนในการประมาณค่ามากกว่าจึงทำให้สารสนเทศของแบบสอบมีค่าต่ำกว่าด้วย

Kose and Demirtasli (2012) ได้เปรียบเทียบโมเดลการตอบสนองข้อสอบแบบเอกมิติและโมเดลการตอบสนองข้อสอบแบบพหุมิติภายใต้ความยาวของแบบสอบ และขนาดกลุ่มตัวอย่างที่แตกต่างกัน โดยใช้ข้อมูลการทดสอบภาษาตุรกี โดยขนาดของกลุ่มตัวอย่าง และความยาวของแบบสอบเป็นตัวแปรต้นที่ถูกจัดกระทำในการศึกษาครั้งนี้ ข้อมูลการทดสอบได้จากนักเรียนเกรด 8 จำนวน 1,516 คน ผลการศึกษาปรากฏว่าการประมาณค่าพารามิเตอร์ข้อสอบ และพารามิเตอร์ความสามารถของบุคคลโดยใช้โมเดลการตอบสนองข้อสอบแบบพหุมิติมีค่าความคลาดเคลื่อนในการประมาณค่าต่ำกว่า ขณะที่มีความแม่นยำในการวัดสูงกว่า นอกจากนี้พบว่าข้อมูลการทดสอบมีความสอดคล้องกับโมเดลการตอบสนองข้อสอบแบบพหุมิติเป็นอย่างดี และยังพบด้วยว่าความยาวของแบบสอบ และขนาดของกลุ่มตัวอย่างไม่มีใด ๆ ต่อความสอดคล้องของข้อมูลการทดสอบกับโมเดลการตอบสนองข้อสอบแบบเอกมิติ แต่ขนาดกลุ่มตัวอย่างที่ใหญ่ขึ้น และแบบสอบที่มีความยาวมากขึ้นสามารถช่วยลดความคลาดเคลื่อนในการทดสอบได้และมีผลต่อความสอดคล้องของข้อสอบ

กับโมเดลการตอบสนองข้อสอบแบบพหุมิติ

สรุปจากการศึกษางานวิจัยที่เกี่ยวข้อง ทฤษฎีการตอบสนองข้อสอบ จะมีความสัมพันธ์ระหว่างคุณลักษณะระหว่างคุณลักษณะภายในหรือความสามารถที่มีอยู่ภายในตัวบุคคลกับพฤติกรรมการตอบสนองข้อสอบของบุคคลนั้นว่ามีโอกาสตอบข้อสอบถูกมากน้อยเพียงไร ทฤษฎีนี้มีพื้นฐานความเชื่อว่าพฤติกรรมการตอบสนองต่อ ข้อสอบของผู้สอบ ซึ่งเป็นสิ่งที่สังเกตได้โดยตรงว่าถูกหรือผิดจะถูกกำหนดโดยคุณลักษณะภายในหรือความสามารถที่อยู่ภายในตัวบุคคลซึ่งเป็นสิ่งที่ไม่สามารถสังเกตได้โดยตรง ทฤษฎีนี้ได้โดยอธิบายความสัมพันธ์ดังกล่าวในรูปของฟังก์ชันคณิตศาสตร์หรือโมเดลที่แสดงความสัมพันธ์ระหว่างระดับความสามารถคุณลักษณะของข้อสอบและโอกาสของการตอบข้อสอบได้ถูก ที่เรียกว่า ฟังก์ชันการตอบสนองข้อสอบ ซึ่งมีลักษณะความสัมพันธ์เป็นแบบฟังก์ชันโลจิสหรือฟังก์ชันปกติสะสม

ตอนที่ 3 การทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) และงานวิจัยที่เกี่ยวข้อง

การตรวจสอบคุณภาพของข้อที่เป็นผลมาจากการตัดสินว่าข้อสอบนั้นมีความยุติธรรมหรือไม่อย่างไร ซึ่งจะเป็นข้อมูลสารสนเทศทางสถิติและจะมีความสัมพันธ์ที่ส่งผลไปยังคุณลักษณะที่ข้อสอบมุ่งที่จะวัด กับมวลประสบการณ์ของผู้สอบในกลุ่มต่าง ๆ ที่มีบางอย่างที่แตกต่างกัน เช่น เชื้อชาติ ศาสนา วัฒนธรรม ภูมิลำเนา สังคม เพศ ภาษา อายุ และประสบการณ์เป็นต้น โดยข้อสอบที่จะมุ่งวัดนั้นอาจทำให้เกิดการได้เปรียบเสียเปรียบทั้งที่ความสามารถของผู้สอบมีความเท่าเทียมกันทำให้เกิดความแตกต่างกับคำถามของข้อสอบจึงทำใช้ คำที่ว่า ข้อสอบลำเอียง (Biased item) ซึ่งคำที่ว่านี้จะมีความหมายในเชิงลบ แล้วเกณฑ์ที่ใช้สำหรับในการตัดสินว่าข้อสอบนั้นมีความลำเอียงยังไม่ชัดเจนเท่าที่ควรนัก จึงได้เปลี่ยนมาใช้คำว่า การทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) จะเห็นได้ว่าความหมายจะเป็นกลางสำหรับผู้สอบ (Holland & Thayer, 1988: Honlland & Wainer, 1993; ศิริชัย กาญจนวาสี, 2555, หน้า 115-116) และมีนักวัดผลหลายท่านที่ได้ให้ความหมายของการทำหน้าที่ต่างกันของข้อสอบไว้ดังนี้

ความลำเอียงของข้อสอบ หมายถึง ผู้ที่เข้าสอบจะมีสัดส่วนในการตอบคำถามข้อสอบได้ถูกต้องไม่เท่ากันของแต่ละกลุ่มประชากรที่จะศึกษา และเมื่อกลุ่มผู้เข้าสอบได้คะแนนที่เท่ากันนั้นข้อสอบก็เป็นเอกพันธ์ (Scheuneman, 1979; ศิริชัย กาญจนวาสี, 2555, หน้า 116)

ความลำเอียงของข้อสอบ หมายถึง ข้อสอบที่มีค่าความยากจะมีความสัมพัทธ์ต่อเนื่องกับกลุ่มผู้เข้าสอบกลุ่มหนึ่งมากกว่ากลุ่มผู้เข้าสอบอีกกลุ่ม (Rudner, Getson, & Knight, 1980)

ความลำเอียงของข้อสอบ หมายถึง แนวโน้มความลำเอียงที่ใช้คะแนนจากผลการสอบนำมาตัดสินผลเป็นจะทำให้เกิดความไม่ยุติธรรม (Popham, 1981) ความลำเอียงของข้อสอบหมายถึง ข้อสอบที่วัดความสามารถในการที่เลือกคำตอบมีโอกาสในการตอบที่ถูกได้แตกต่างกันและโอกาสการตอบในเชิงบวกจะแตกต่างกันในการวัดเจตคติ เนื่องจากผู้เข้าสอบมีคุณลักษณะของการในปริมาณที่เท่ากัน แต่จะมาจากกลุ่มประชากรย่อยที่แตกต่างกัน (Hulin, Drasgow, & Parson, 1983)

ความลำเอียงของข้อสอบ หมายถึง ในการเลือกตอบข้อสอบนั้นจะมีโอกาสในการตอบถูกต้องของผู้เข้าสอบกลุ่มหนึ่งมีค่าคะแนนต่ำกว่าหรือคะแนนอาจสูงกว่าอีกกลุ่มหนึ่งของผู้เข้าสอบโดยที่มีความสามารถในระดับเดียวกัน (Lederman, 1990)

การทำหน้าที่ต่างกันของข้อสอบ หมายถึง จะเป็นแบบพหุมิติของข้อสอบที่จะวัดจะดูได้จากข้อมูลการแจกแจงความสามารถหลัก (Primary Abillity) ของผู้เข้าสอบตั้งแต่ 2 กลุ่ม ขึ้นไปที่มีความเท่ากันโดยมีการแจกแจงความสามารถรองลงมา (Secondary) จะแตกต่างจากกัน (Camilli & Shepard, 1994)

การทำหน้าที่แตกต่างกันของข้อสอบ หมายถึง จะเป็นการตอบสนองที่มีฟังก์ชันคำนวณจากกลุ่มผู้เข้าสอบที่เป็นกลุ่มย่อยที่ต่างกัน และก็มีค่าที่ไม่เท่ากัน (Narayanan & Swaminathan, 1996)

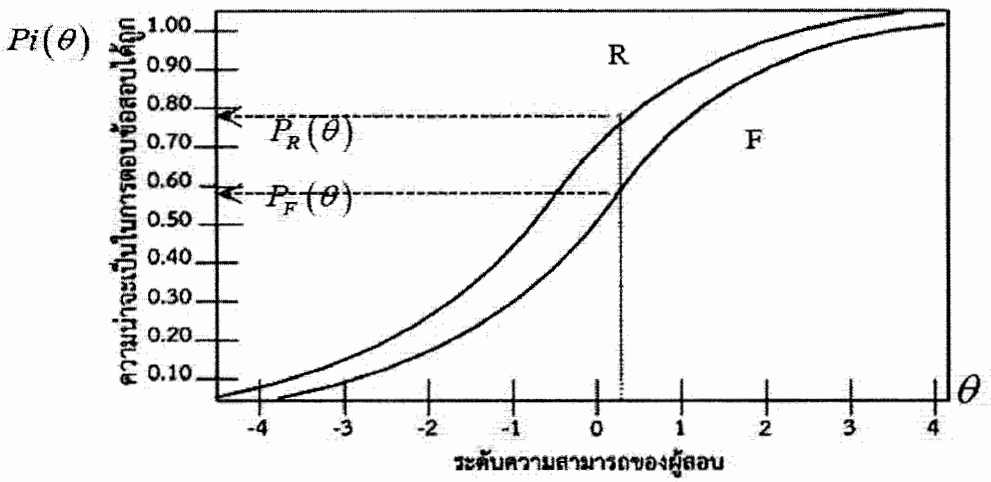
การทำหน้าที่ต่างกันของข้อสอบ ตามที่ได้กล่าวมานี้สามารถที่สรุปได้ว่า การทำหน้าที่ต่างกันของข้อสอบ หมายถึง ผู้เข้าสอบข้อสอบที่ต่างกลุ่มกันที่มีความสารถหรือคุณลักษณที่มุ่งวัดเท่ากัน ซึ่งทำให้มีโอกาสในการตอบข้อสอบที่ถูกได้แตกต่างกันออกไป หรือมีฟังก์ชันการตอบสนองข้อสอบที่แตกต่างกัน และการทำหน้าที่ต่างกันของข้อสอบจะเกิดขึ้นเมื่อนำข้อสอบไปทดสอบกับผู้เข้าสอบกลุ่มย่อยที่ต่างกัน โดยจะมีความสามารถหลัก (Primary Ability) ในระดับเดียวกันหรือมีคุณลักษณะแฝง (Latent Trait) ที่จะวัดเท่ากัน และจะมีความสามารถรองลงมา (Secondary Ability) ที่แตกต่างซึ่งจะทำให้ผู้เข้าสอบที่ต่างกลุ่มกันเมื่อนำมาเปรียบเทียบกันจะทำให้โอกาสในการตอบข้อสอบนั้นแตกต่างกัน (ศิริชัย กาญจนวาสี, 2555, หน้า 116-117)

ประเภทของการทำหน้าที่ต่างกันของข้อสอบ

การเปรียบเทียบผลการตอบของข้อสอบ ที่ทำหน้าที่ต่างกันของข้อสอบจะเป็นการเปรียบเทียบกลุ่มผู้เข้าสอบ 2 กลุ่มขึ้นไป โดยจะประกอบด้วยกลุ่มแรกคือ กลุ่มเปรียบเทียบ (Focal Group หรือกลุ่ม F) เป็นกลุ่มที่น่าสนใจที่จะศึกษาและคาดว่าจะเป็นกลุ่มที่จะเสียเปรียบในตอบข้อสอบ ส่วนกลุ่มที่ 2 เป็นกลุ่มอ้างอิง (Reference Group หรือ กลุ่ม R) เป็นกลุ่มที่คาดว่าจะได้เปรียบในการตอบข้อสอบได้ถูกต้องมากกว่ากลุ่มแรก

การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ ซึ่งข้อสอบจะทำหน้าที่ต่างกันของข้อสอบแตกต่างกันได้ 2 ประเภท (Mellenbergh, 1982) คือ การทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป (Unifrom DIF) และแบบอเนกรูป (Nonunifrom DIF)

1. ข้อสอบทำหน้าที่ต่างกันแบบเอกรูป (Unifrom DIF) หมายถึง ข้อสอบที่กลุ่มผู้เข้าสอบกลุ่มหนึ่งมีโอกาสในการตอบข้อสอบได้มากกว่าอีกกลุ่มผู้เข้าสอบอย่างสม่ำเสมอกัน ในทุกระดับความสามารถ และพิจารณาโค้งลักษณะข้อสอบของผู้สอบทั้ง 2 กลุ่ม จะเห็นได้ว่าไม่มีปฏิสัมพันธ์กันระหว่างระดับความสามารถของผู้สอบกับการเป็นสมาชิกของกลุ่ม (Group Membership) ดังภาพที่ 2-1



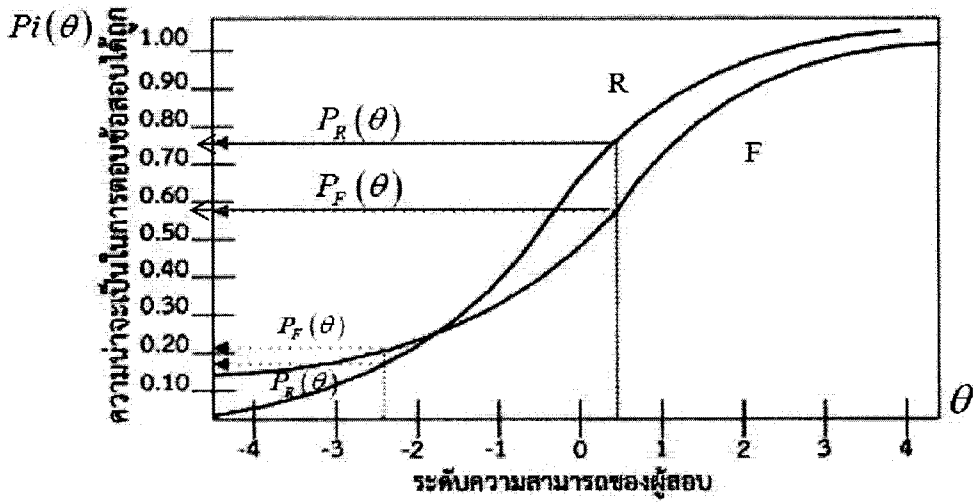
ภาพที่ 2-1 ข้อสอบทำหน้าที่ต่างกันแบบเอกรูป (Uniform DIF) (ศิริชัย กาญจนวาสี, 2555, หน้า 118)

2. ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป (Nonunifrom DIF) หมายถึง ข้อสอบที่ผู้เข้าสอบมีโอกาสในการตอบถูกในระหว่างผู้เข้าสอบทั้ง 2 กลุ่มมีความแตกต่างกันไม่สม่ำเสมอกันในทุกระดับความสามารถ และเมื่อพิจารณาโค้งคุณลักษณะของข้อสอบทั้งกลุ่ม 2 กลุ่ม จะเห็นได้ว่าจะมีปฏิสัมพันธ์กันในระหว่างระดับความสามารถของผู้เข้าสอบ กับการเป็นสมาชิกกลุ่ม เช่น ระดับความสามารถหนึ่งกลุ่มผู้เข้าสอบกลุ่มอ้างอิง (R) จะมีโอกาสในการตอบข้อสอบถูกได้มากกว่ากลุ่มผู้เข้าสอบเปรียบเทียบ (F) และที่ระดับความสามารถอีกระดับหนึ่งกลุ่มผู้เข้าสอบกลุ่มเปรียบเทียบ (F) มีโอกาสตอบข้อสอบได้มากกว่ากลุ่มอ้างอิง (R)

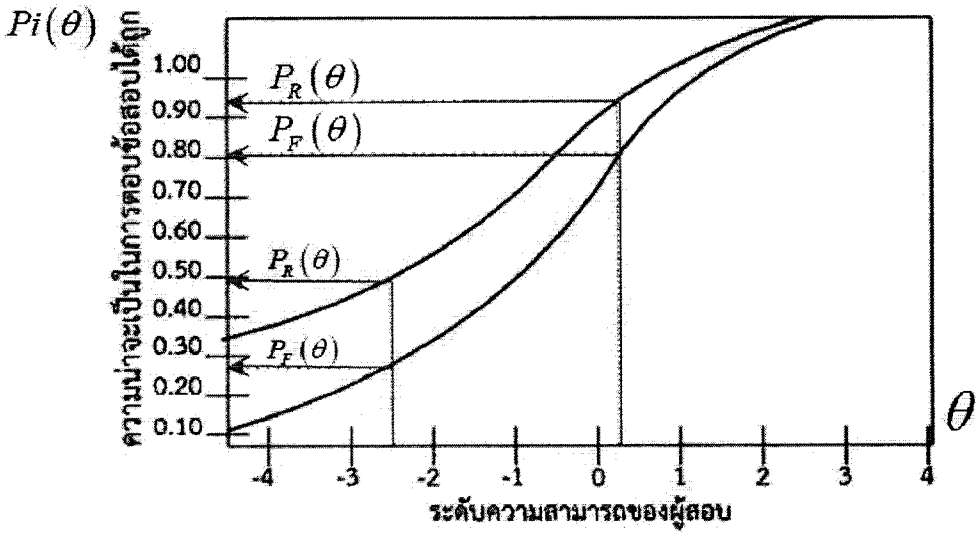
ตามทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) สามารถพิจารณา “ปฏิสัมพันธ์”ที่ได้จากความแตกต่างของค่าพารามิเตอร์อำนาจจำแนกของข้อสอบในระหว่างผู้เข้าสอบกลุ่มย่อยสองกลุ่ม ถ้าข้อสอบทำหน้าที่ต่างกันแบบเอกรูป โค้งลักษณะของข้อสอบ (Iterm Charactreistic Curves: ICCs) ในกลุ่มผู้เข้าสอบทั้ง 2 กลุ่มจะเป็นเส้นที่ขนานกัน หรือฟังก์ชันการตอบสนองข้อสอบ (Item Response Functions: IRFs) เหมือนกัน ส่วนข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป โค้งลักษณะข้อสอบระหว่างผู้เข้าสอบกลุ่มย่อยทั้ง 2 กลุ่มไม่ขนาน หรือ มีฟังก์ชันการตอบสนองข้อสอบต่างกัน และความแตกต่างระหว่างโค้งลักษณะข้อสอบทั้งสองแบบ จะแสดงถึงขนาดและทิศทางของข้อสอบที่ทำหน้าที่ต่างกัน

ข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูป จำแนกได้เป็น 2 ลักษณะ (Swaminathan & Roger, 1990)

1. ข้อสอบทำหน้าที่ต่างแบบอเนกรูปที่มีปฏิสัมพันธ์ไม่เป็นลำดับ (Disordinal interaction) เป็นการทำหน้าที่ต่างกันสำหรับกลุ่มผู้เข้าสอบที่เกิดขึ้น เมื่อโค้งลักษณะข้อสอบจะตัดกันระหว่างช่วงความสามารถของผู้เข้าสอบเรียกว่า ข้อสอบทำหน้าที่ต่างกันแบบไม่มีทิศทาง (Non-Unidirectional DIF) ดังภาพที่ 2-2



ภาพที่ 2-2 ข้อสอบทำหน้าที่ต่างกันแบบไม่มีทิศทาง (Non-unidirectional DIF)
(ศิริชัย กาญจนวาสี, 2555, หน้า 119)



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

หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF Detection) เป็นการเปรียบเทียบผลการตอบข้อสอบเป็นรายข้อระหว่างกลุ่มผู้สอบอย่างน้อย 2 กลุ่ม ที่มีความสามารถหลัก (Primary Ability) ที่มุ่งวัดเท่ากัน แต่คาดว่าจะมีความได้เปรียบในการตอบข้อสอบข้อนั้น หรือมีโอกาสตอบ

ข้อสอบได้ถูกต้องมากกว่า ส่วนอีกกลุ่มคือ กลุ่มเปรียบเทียบ (Focal Group) ซึ่งเป็นกลุ่มที่สนใจศึกษาและคาดว่าน่าจะเป็นกลุ่มที่เสียเปรียบ

ในการเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบจำเป็นต้องจับคู่ (Matching) ผู้สอบตามความสามารถ ซึ่งเป็นเงื่อนไขสำคัญของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เกณฑ์การจับคู่ (Matching Criteria) ที่นิยมใช้กันมี 2 วิธี ดังนี้

1. เกณฑ์ภายนอก (External Criterion)

การวิเคราะห์การทำน้าที่ต่างกัน โดยใช้คะแนนจากแบบสอบอื่นเป็นเกณฑ์ภายนอก แล้วใช้เทคนิคการวิเคราะห์การถดถอย (Regression Analysis) เพื่อทำการเปรียบเทียบเส้นกราฟความสัมพันธ์ระหว่างตัวแปรเกณฑ์ กับตัวแปรทำนายระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ

หลักการนี้มีจุดมุ่งหมาย เพื่อสร้างสมการทำนายตัวแปรเกณฑ์ ซึ่งเป็นคะแนนของ แบบสอบอื่นจากตัวแปรทำนายซึ่งเป็นคะแนนรายข้อหรือคะแนนแบบสอบระหว่างกลุ่มอ้างอิง และกลุ่มเปรียบเทียบในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ จะใช้คะแนนรวมข้อเป็นตัวแปรทำนาย แต่ถ้าเป็นการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ จะใช้คะแนนรวมของแบบสอบ ทั้งฉบับเป็นตัวแปรทำนายสำหรับตัวแปรเกณฑ์ที่ใช้เป็นเกณฑ์ภายนอก อาจใช้คะแนนรวมทั้งฉบับ หรือเกรดเฉลี่ย หรือผลสัมฤทธิ์ในงานที่เกี่ยวข้องของผู้สอบ (Cronbach, 1970) สมการทำนายสำหรับกลุ่มอ้างอิงและกลุ่มเปรียบเทียบแสดงได้ดังนี้

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

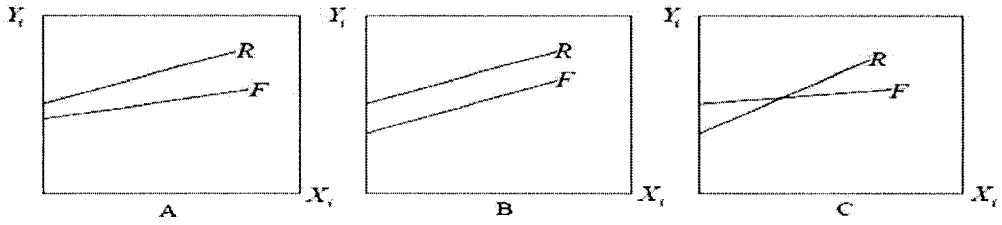
กลุ่มเปรียบเทียบ(F) $Y_i = A_F + B_F X_i$

เมื่อ Y_i = คะแนนของตัวแปรเกณฑ์ภายนอก

X_i = คะแนนของตัวแปรทำนาย

A = ค่าคงที่หรือค่าตัดแกน (Intercept)

B = ค่าความชัน (Slope)



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

จากฟังก์ชันการทำนายทั้ง 2 สมการ สามารถเปรียบเทียบค่าตัดแกน (A) และค่าความชัน (B) ของเส้นกราฟระหว่างกลุ่มอ้างอิง (R) และกลุ่มเปรียบเทียบ (F) ได้ ถ้าเส้นกราฟดังกล่าวมีค่าความชันหรือค่าตัดแกน แตกต่างกันสำหรับข้อสอบใดหรือแบบสอบใดแสดงว่าข้อสอบ

หรือแบบสอบนั้นมีการทำหน้าที่ต่างกัน โดยลำเอียงเข้าข้างกลุ่มผู้สอบที่มีค่าตัดแกนสูงกว่าหรือค่าความชันที่สูงกว่า

การใช้เกณฑ์ภายนอกมีข้อดี คือเกณฑ์ที่ใช้มีความเป็นอิสระจากข้อสอบ และแบบสอบที่ต้องการตรวจสอบ แต่มีจุดอ่อนตรงที่ความเหมาะสมของเกณฑ์ที่จะนำมาใช้ ในทางปฏิบัติเป็นการยากที่จะหาตัวแปรเกณฑ์ภายนอกจากแบบสอบฉบับอื่นที่มีความตรงเชิงทำนาย และมีความยุติธรรมสำหรับกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ถ้าตัวแปรเกณฑ์ภายนอกขาดคุณสมบัติดังกล่าวจะทำให้ผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบหรือแบบสอบขาดความแม่นยำและความสมบูรณ์

2. เกณฑ์ภายใน (Internal Criterion)

การวิเคราะห์การทำหน้าที่ต่างกัน โดยใช้เกณฑ์ภายในเป็นการนำวิธีการทางสถิติมาตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ หรือแบบสอบ โดยเน้นการพิจารณาโครงสร้างภายในของแบบสอบเป็นหลัก ด้วยการวิเคราะห์ผลการจากตอบข้อสอบและความสามารถหรือคะแนนจริงของผู้สอบที่ได้จากแบบสอบฉบับนั้น เพื่อนำมาเปรียบเทียบระหว่างผู้สอบจากกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ที่มีความสามารถหรือคะแนนจริงเท่ากัน ว่าจะมีผลการตอบหรือโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกันหรือไม่ เพื่อบ่งชี้ถึงการทำหน้าที่ต่างกันของข้อสอบ การวิเคราะห์ในลักษณะนี้นิยมใช้ค่าสถิติต่าง ๆ เป็นตัวชี้ถึงการทำหน้าที่ต่างกันของข้อสอบ ค่าสถิติทดสอบที่นิยมนำมาใช้พอสรุปได้ดังนี้

2.1 การทดสอบปฏิสัมพันธ์ (Interction)

ในระยะเริ่มแรกของการศึกษาความลำเอียงของข้อสอบ มีการใช้สถิติทดสอบ เอฟ (F-test) จากการวิเคราะห์ความแปรปรวน (ANOVA) เพื่อทดสอบปฏิสัมพันธ์ระหว่างกลุ่มผู้สอบกับข้อสอบ ถ้าการทดสอบมีนัยสำคัญเป็นสัญญาณของการทำหน้าที่ต่างกันของข้อสอบ (Cleary & Hilton 1968; Jensen, 1974) จากนั้นจึงทำการวิเคราะห์ต่อด้วยวิธีการ Post Hoc เพื่อระบุข้อสอบที่มีผลต่อการเกิดปฏิสัมพันธ์ ซึ่งเป็นข้อที่ทำหน้าที่ต่างกัน

วิธีการนี้มีข้อดีที่สามารถศึกษาผู้สอบหลายๆกลุ่มได้สะดวก แต่มีจุดอ่อนในเรื่องการควบคุมกลุ่มต่างๆให้มีความสามารถทัดเทียมกัน ขนาดกลุ่มตัวอย่างของกลุ่มต่าง ๆ และอัตราความคลาดเคลื่อนประเภทที่ 1 จะสูงขึ้น ถ้าจำนวนข้อสอบเพิ่มขึ้น

2.2 การวัดความเบี่ยงเบนสัมพันธ์ (Reletive Deviation)

การคำนวณค่าความยากของข้อสอบ เช่น p, b เป็นต้น เมื่อคำนวณแยกระหว่างกลุ่มและแปลงให้เป็นค่ายากมาตรฐาน สามารถนำมาเปรียบเทียบเป็นรายข้อ ถ้าข้อใดเบี่ยงเบนไปแกนหลักที่คาดหมาย หรือเบี่ยงเบนเกินจากความคลาดเคลื่อนมาตรฐานของค่าความยากที่กำหนดย่อมแสดงถึงหน้าที่ต่างกันของข้อสอบ (Cleary & Kilton, 1968; Angoff & Ford, 1973) รวมทั้งสามารถคำนวณค่าสหพันธ์เข้าใกล้ 1.00 แสดงว่าความยากสัมพันธ์ของข้อสอบมีค่าใกล้เคียงกันระหว่างกลุ่ม ดังนั้นแบบสอบวัดคุณลักษณะคล้ายกันระหว่างกลุ่ม

วิธีการนี้มีข้อดีและข้อเสียคล้ายการทดสอบปฏิสัมพันธ์ นอกจากนี้ค่าความยากของข้อสอบ (P) มิใช่ตัวแทนของค่าความยากจริงของข้อสอบ และได้รับอิทธิพลจากค่าแทรกซ้อนอื่นได้แก่ ค่าอำนาจจำแนก และความสามารถของผู้สอบ

การเปรียบเทียบน้ำหนักตัวประกอบ (Factor Loading)

การวิเคราะห์ตัวประกอบ (Factor Analysis) เป็นเทคนิคทางสถิติที่นิยมใช้ในการตรวจสอบความตรงเชิงทฤษฎีหรือโครงสร้าง (Construct Validity) เมื่อนำการวิเคราะห์ตัวประกอบมาใช้ในการวิเคราะห์โครงสร้างของแบบสอบแยกตามกลุ่มผู้สอบ ความไม่สอดคล้องกันระหว่างน้ำหนักตัวประกอบคุณลักษณะสำคัญที่มุ่งวัด หรือ ความแตกต่างของค่าเฉลี่ยคะแนนตัวประกอบ (Factor Scores) ระหว่างกลุ่มผู้สอบ ย่อมสะท้อนการทำหน้าที่ต่างกันของข้อสอบและแบบสอบ

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

การเปรียบเทียบโอกาสตอบข้อสอบถูก

การวิเคราะห์โอกาสตอบข้อสอบถูกของผู้สอบจากกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่มีความสามารถเท่ากัน เป็นแนวทางสำคัญที่นิยมใช้กันและเป็นที่ยอมรับในปัจจุบัน สำหรับบ่งชี้การทำหน้าที่ต่างกันของข้อสอบ มีการคำนวณค่าสถิติ 2 แนวทาง ดังนี้

1) เปรียบเทียบค่าสัดส่วนหรือความน่าจะเป็นการตอบข้อสอบถูกของผู้สอบต่างกลุ่มที่มีความสามารถเท่ากัน เช่น วิธีแมนเทล-แฮนส์เซล เป็นต้น

2) เปรียบเทียบค่าฟังก์ชันการตอบสนองข้อสอบ หรือโค้งลักษณะข้อสอบระหว่างกลุ่มที่มีระดับความสามารถเท่ากัน เป็นวิธีที่อยู่บนพื้นฐานของทฤษฎี IRT เช่น วิธีวัดความแตกต่างของพื้นที่ วิธีวัดความแตกต่างของค่าพารามิเตอร์ความยาก วิธีการทดสอบไค-สแควร์ของ ลอร์ด (Lord's χ^2-test) เป็นต้น

วิธีการนี้มีข้อดีที่สำคัญ ได้แก่ การคำนวณค่าสถิติของข้อสอบมีความน่าเชื่อถือมีกลไกลควบคุมความสามารถของผู้สอบโดยการจับคู่กลุ่มความสามารถ เพื่อทำการเปรียบเทียบ ณ ตำแหน่งต่าง ๆ ที่มีความสามารถเท่ากัน จึงเป็นวิธีการที่ยอมกันทั่วไป แต่มีข้อจำกัดในด้านความสลับซับซ้อนของแนวคิดพื้นฐาน และการวิเคราะห์มีความจำเป็นต้องใช้โปรแกรมคอมพิวเตอร์โดยเฉพาะ

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF Detection) จำแนกตามลักษณะการตรวจให้คะแนนได้เป็น 2 ประเภท คือ ข้อสอบที่มีการให้คะแนนแบบทวิวิภาค หรือสองค่า (Dichotomous Scoring) และข้อสอบที่มีการให้คะแนนแบบพหุวิภาค หรือหลายค่า (Polytomous Scoring) วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแต่ละประเภท ยังสามารถจำแนกได้อีก 2 มิติ ได้แก่ มิติลักษณะของตัวแปรเกณฑ์ ซึ่งแบ่งเป็นกลุ่มวิธีที่ใช้คะแนนสังเกตได้ (Observed Score) และกลุ่มวิธีที่ใช้คะแนนสังเกตไม่ได้หรือคะแนนของตัวแปรแฝง (Latent variable) และมิติลักษณะของสถิติวิเคราะห์ ซึ่งแบ่งเป็นกลุ่มวิธีที่ใช้สถิติราพาเมตริก (Parametric Approach) รายชื่อวิธีการตรวจการทำหน้าที่ต่างกันของข้อสอบที่สำคัญ ๆ ดังตารางที่ 2-1

ตารางที่ 2-1 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการตรวจให้คะแนนแบบทวิวิภาค (Dichotomous DIF) และพหุวิภาค (Polytomous DIF) (ศิริชัย กาญจนวาสี, 2555, หน้า 124)

ประเภทและตัวแปรเกณฑ์	พาราเมตริก	นันพาราเมตริก
1. DIF แบบทวิวิภาค		
1.1 คะแนนที่สังเกตได้ (observed score)	ANOVA Logistic Regression	TID MH STND
1.2 คุณลักษณะ/ตัวแปรแฝง (latent variable)	IRT-D^2 Lord's χ^2 General IRTLR Loglinear IRTLR	SIBTEST
2. DIF แบบพหุวิภาค		
2.1 คะแนนที่สังเกตได้ (observed score)	ANOVA Polytomous Logistic Regression	Polytomous STND
2.2 คุณลักษณะ/ตัวแปรแฝง (latent variable)	General IRTLR PCM	GMH Polytomous SIBTEST GPCM

1). วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนแบบทวิวิภาค

1.1 กลุ่มวิธีที่ใช้คะแนนที่สังเกตได้

วิธีการตรวจสอบคะแนนที่สังเกตได้แบบทวิวิภาคจะวิเคราะห์ตามทฤษฎีการทดสอบแบบดั้งเดิม (CTT) และกลุ่มที่ไม่ใช้ทฤษฎีการตอบสนองข้อสอบ (NON-IRT approach) โดยใช้คะแนนรวมของผู้เข้าสอบเป็นเกณฑ์การจับคู่ของกลุ่มผู้สอบ วิธีการตรวจสอบที่สำคัญในกลุ่มนี้ คือ

การวิเคราะห์ความแปรปรวน (ANOVA) (Cleary & Hilton, 1968)

วิธีการวิเคราะห์การถดถอยโลจิสติก (Logistic Regression: LR) (Swaminathan & Rogers, 1990)

วิธีแปลงค่าความยากของข้อสอบ (Transformed Item Difficulty: TID) (Cleary & Hilton, 1968; Angoff & Ford, 1973)

วิธีแมนเทล-แฮนส์เซล (Mantel-Haenszel: MH) (Holland & Thayer, 1988: 1989)

วิธีดัชนีมาตรฐาน (Standardization: STND) การปรับให้เป็นมาตรฐานด้วยน้ำหนักตัวประกอบ (Dorans & Kulick, 1986)

1.2 กลุ่มวิธีที่ใช้คุณลักษณะแฝง

วิธีการตรวจสอบคุณลักษณะ/ ตัวแปรแฝงแบบทวิวิภาค ซึ่งเป็นการวิเคราะห์อยู่บนพื้นฐานของทฤษฎีการตอบสนองข้อสอบ (IRT) สำหรับใช้เป็นเกณฑ์การจับคู่กลุ่มสอบ วิธีการตรวจสอบที่สำคัญในกลุ่มนี้ คือ

วิธีวัดพื้นที่ความแตกต่างระหว่างโค้งการตอบสนองข้อสอบ (IRT-D^2) (Linn et al., 1981; Shepard et al., 1984; Raju, 1990; Kim & Cohen, 1991)

วิธีไค-สแควร์ของลอร์ด (Lord's χ^2) (Lord, 1980)

วิธีอัตราส่วนไลค์ลิฮูดทั่วไป (General IRT Likelihood Ratio) (Thissen, Steinberg & Wainer, 1993)

วิธีอัตราส่วนไลค์ลิฮูด ลอกลิเนียร์ (Loglinear Likelihood Ratio) (Thissen, Steinberg & Wainer, 1993)

วิธีชิปเทสท์ (SIBTEST) (Shealy & Stout, 1993)

2.) วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนแบบพหุวิภาค

2.1 กลุ่มวิธีที่ใช้คะแนนที่สังเกตได้

วิธีการวิเคราะห์ความแปรปรวน (ANOVA) (Cleary & Hilton, 1968)

วิธีการวิเคราะห์การถดถอยโลจิสติกพหุวิภาค (Polytomous Logistic Regression) (Swaminathan & Rogers, 1990)

วิธีดัชนีมาตรฐานพหุวิภาค (Polytomous Standardization) (Dorans & Thayer, 1988, 1989)

วิธีแมนเทล–แฮนส์เซลทั่วไป (General Mantel-Haenzel; GMH) (Holland & Thayer, 1988, 1989)

2.2 กลุ่มวิธีที่ใช้คุณลักษณะแฝง

วิธีอัตราส่วนไลค์ลิฮูดในรูปทั่วไป (General Mantel Likelihood Ratio) (Thissen, Steinberg & Wainer, 1993)

วิธีการให้คะแนนบางส่วน (Partial Credit Model: PCM) (Master, 1982)

วิธีชิปเทสท์พหุวิภาค (Polytomous SIBTEST) (Shealy & Stout, 1993)

วิธีการให้คะแนนบางส่วนทั่วไป (Generalized Partial Credit Model: GPCM) (Muraki, 1992, 1993) (ศิริชัย กาญจนวาสี, 2555, หน้า 116-126)

สรุป คือ การทำหน้าที่ต่างกันของข้อสอบ ตามที่ได้กล่าวมานี้สามารถที่สรุปได้ว่า การทำหน้าที่ต่างกันของข้อสอบ หมายถึง ผู้เข้าสอบข้อสอบที่ต่างกลุ่มกันที่มีความสามารถหรือคุณลักษณะที่มุ่งวัดเท่ากัน ซึ่งทำให้มีโอกาสในการตอบข้อสอบที่ถูกได้แตกต่างกันออกไป หรือมีฟังก์ชันการตอบสนองข้อสอบที่แตกต่างกัน และการทำหน้าที่ต่างกันของข้อสอบจะเกิดขึ้นเมื่อนำข้อสอบไปทดสอบกับผู้เข้าสอบกลุ่มย่อยที่ต่างกัน โดยจะมีความสามารถหลัก (Primary Ability) ในระดับเดียวกันหรือมีคุณลักษณะแฝง (Latent Trait) ที่จะวัดเท่ากัน และจะมี

ความสามารถรองลงมา (Secondary Ability) ที่แตกต่างซึ่งจะทำให้ผู้เข้าสอบที่ต่างกลุ่มกันเมื่อนำมาเปรียบเทียบกันจะทำให้โอกาสในการตอบข้อสอบนั้นแตกต่างกัน โดยการเปรียบเทียบผลการตอบของข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบจะเป็นการเปรียบเทียบกลุ่มผู้เข้าสอบ 2 กลุ่มขึ้นไปโดยประกอบด้วยกลุ่มแรกคือ กลุ่มเปรียบเทียบ (Focal Group หรือกลุ่ม F) เป็นกลุ่มที่นาสนใจที่จะศึกษาและคาดว่าจะเป็นกลุ่มที่จะเสียเปรียบในตอบข้อสอบ ส่วนกลุ่มที่ 2 เป็นกลุ่มอ้างอิง (Reference Group หรือ กลุ่ม R) เป็นกลุ่มที่คาดว่าจะได้เปรียบในการตอบข้อสอบได้ถูกต้องมากกว่ากลุ่มแรก

งานวิจัยที่เกี่ยวข้องกับการทำหน้าที่ต่างกันของข้อสอบ (DIF) มีดังนี้

พีรญา สูงเนิน เสรี ชัดแช้ม และสมโภชน์ อเนกสุข (2552) ได้ปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบแบบพหุมิติ ระหว่างข้อสอบรายข้อกับหมวดข้อสอบโดยวิธีซิปเทสท์ ภายใต้เงื่อนไขขนาดของกลุ่มตัวอย่างที่ต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) กลุ่มตัวอย่างเป็นนักเรียนชั้นประถมศึกษาปีที่ 6 สังกัดเขตพื้นที่การศึกษานครศรีธรรมราช ปีการศึกษา 2546 ที่เข้าสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ จำนวน 2,000 คน ข้อมูลทุติยภูมิที่ใช้เป็นคะแนนจากแบบทดสอบวิชาภาษาไทย ชั้นประถมศึกษาปีที่ 6 จำนวน 40 ข้อ จำแนกเป็น 2 หมวดข้อสอบ คือ หมวดที่ 1 วัดด้านโครงสร้างความรู้ จำนวน 15 ข้อ และหมวดที่ 2 วัดด้านกระบวนการ จำนวน 25 ข้อ วิเคราะห์ค่าสถิติพื้นฐานโดยใช้โปรแกรม SPSS ตรวจสอบความตรงเชิงโครงสร้างด้วยการวิเคราะห์องค์ประกอบเชิงยืนยันอันดับสอง โดยใช้โปรแกรม LISREL 8.50 และตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยใช้โปรแกรม SIBTEST ผลการวิจัยปรากฏว่า ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก ขนาดกลาง และขนาดใหญ่ ระหว่างการตรวจสอบการทำหน้าที่ต่างกันเป็น รายข้อกับรายหมวดข้อสอบ พบข้อสอบทำหน้าที่ต่างกันแตกต่างกันการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นรายข้อ เมื่อกลุ่มตัวอย่างขนาดเล็ก พบข้อสอบทำหน้าที่ต่างกันจำนวน 4 ข้อ คิดเป็นร้อยละ 10 ขนาดกลางพบจำนวน 13 ข้อ คิดเป็นร้อยละ 32.5 และขนาดใหญ่พบจำนวน 15 ข้อ คิดเป็นร้อยละ 37.5 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นรายหมวดข้อสอบ เมื่อกลุ่มตัวอย่างขนาดเล็ก พบข้อสอบทำหน้าที่ต่างกันจำนวน 4 ข้อ คิดเป็นร้อยละ 10 ขนาดกลางพบจำนวน 8 ข้อ คิดเป็นร้อยละ 20 และขนาดใหญ่ พบจำนวน 16 ข้อ คิดเป็นร้อยละ 40 การตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบ ชี้ให้เห็นว่า หมวดที่ 2 ภายใต้เงื่อนไขกลุ่มตัวอย่างขนาดกลาง มีนัยสำคัญทางสถิติที่ระดับ .05

ณรงค์ จันทรมหา (2554) ได้เปรียบเทียบค่าความเที่ยงของแบบทดสอบผลสัมฤทธิ์ทางการเรียนที่มีจำนวนข้อสอบทำหน้าที่ต่างกันแตกต่างกัน 7 เงื่อนไข คือ 0% 5% 10% 15% 20% 25% และ 30% กลุ่มตัวอย่างเป็นนักเรียนชั้นประถมศึกษาปีที่ 6 สังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน ปีการศึกษา 2546 จำนวน 2,000 คน ที่ทำแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ วิชาภาษาไทย จำนวน 40 ข้อ วิเคราะห์ค่าสถิติพื้นฐานโดยใช้โปรแกรม SPSS ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้โปรแกรมซิปเทสท์ และทดสอบความแตกต่างของค่าความเที่ยงของแบบทดสอบ โดยใช้การทดสอบสถิติซี ผลการวิจัย พบว่า ค่าความเที่ยงของแบบทดสอบที่มีจำนวนข้อสอบที่ต่างกัน 0% 5% 10% 15% 20% 25% และ30% ไม่แตกต่างกัน

ศิริรัตน์ สุคันธพฤกษ์ โชติกา ภาษี และศิริชัย กาญจนวาสี (2554) ได้วิเคราะห์ข้อคำถามในแบบวัดความวิตกกังวลในการสอบคณิตศาสตร์การเปรียบเทียบระหว่างไฮราชิคอลลิเนียร์โมเดล พาเชียลเครดิตโมเดล และเกรดเรสพอนส์โมเดล ซึ่งมีการตรวจสอบการทำหน้าที่ต่างกันในแบบวัดความวิตกกังวล ในการสอบคณิตศาสตร์ โดยเปรียบเทียบระหว่างไฮราชิคอลลิเนียร์โมเดล พาเชียลเครดิตโมเดลและเกรดเรสพอนส์โมเดลโดยกลุ่มตัวอย่างที่ใช้ในการวิจัยเป็นนักเรียนมัธยมศึกษาปีที่ 6 สายวิทย์-คณิต ปีการศึกษา 2552 จำนวน 1,715 คน จาก 29 โรงเรียนในสังกัดสำนักงานเขตพื้นที่การศึกษาพระนครศรีอยุธยาเขต 1 และเขต 2 สำนักงานเขตพื้นที่การศึกษาอ่างทองและสำนักงานเขตพื้นที่การศึกษานนทบุรี ซึ่งได้มาจากการสุ่มตัวอย่างแบบยกชั้น เครื่องมือที่ใช้ในการวิจัย คือ แบบวัดความวิตกกังวลในการสอบคณิตศาสตร์ การวิเคราะห์ข้อมูล 3 ขั้นตอน คือ วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วย Hierarchical Linear Model (HLM) โดยใช้โปรแกรม HLM การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วย Partial Credit Model (PCM) และ Graded Response Model: GRM ด้วยโปแกรม PRASCALE และเปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วย HLM โดยใช้โปรแกรม HLM กับ PCM กับ GRM ด้วยโปรแกรม PRASCALE ผลการวิจัย พบว่า ข้อคำถามที่ทำหน้าที่ต่างกันของข้อร่วมระหว่าง HLM, PCM และ GRM มี 6 ข้อ จาก 39 ข้อ คิดเป็นร้อยละ 15.38 ข้อคำถามที่ทำหน้าที่ต่างกันของข้อร่วมระหว่าง HLM กับ PCM มี 7 ข้อ จาก 39 ข้อ คิดเป็นร้อยละ 17.94 และข้อคำถามที่ทำหน้าที่ต่างกันของร่วมระหว่าง HLM กับ GRM มี 9 ข้อ จาก 39 ข้อ คิดเป็นร้อยละ 23.07

ชัยวัฒน์ หฤทัยพันธ์ (2558) ได้พัฒนาวิธีการสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยผู้เชี่ยวชาญ และประการที่สองเพื่อเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในด้านอัตราความถูกต้อง และอัตราความคลาดเคลื่อนของผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกับข้อสอบที่คัดสรรมาสำหรับ การทดลองเมื่อใช้วิธีการวิเคราะห์ด้วยแบบวินิจฉัยโดยผู้เชี่ยวชาญ วิธีการประยุกต์ใช้เทคนิคการประชุม แบบเดลฟายจากกลุ่มผู้เชี่ยวชาญและวิธีการประยุกต์ใช้เทคนิคโปรโตคอลอะลาว์ด ตัวอย่างที่ใช้ในการวิจัย คือ ผู้เชี่ยวชาญจำนวน 21 คนและนักเรียนระดับมัธยมศึกษาปีที่ 6 ปีการศึกษา 2556 จำนวน 139 คน ซึ่งได้จากการเลือกตัวอย่างแบบเจาะจง เครื่องมือที่ใช้ในการวิจัยประกอบด้วยแบบวินิจฉัยการทำหน้าที่ต่างกันของข้อสอบจากผู้เชี่ยวชาญ แบบยืนยันการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยประยุกต์เทคนิคการประชุมแบบเดลฟาย แบบสอบถามสำหรับการตรวจสอบความลำเอียงของข้อสอบสำหรับนักเรียน ชุดข้อสอบสาระการเรียนรู้สุขศึกษาและพลศึกษาสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบสำหรับผู้เชี่ยวชาญ ข้อสอบที่คัดสรรมาได้นำมาผ่านการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยวิธีแมนเทล-แฮนส์เซล ด้วยโปรแกรม DDFS 1.0 และโปรแกรม DIFAS 5.0 พบว่า วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยการตัดสินของผู้เชี่ยวชาญที่สำคัญมี 3 วิธี ได้แก่ วิธีที่ 1 การวินิจฉัยการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยผู้เชี่ยวชาญ วิธีที่ 2 การประยุกต์ใช้เทคนิคการประชุมแบบเดลฟายจากกลุ่มผู้เชี่ยวชาญ และวิธีที่ 3 การประยุกต์ใช้เทคนิคโปรโตคอลอะลาว์ด 2) ข้อสอบที่ทำหน้าที่ต่างกันด้านเพศของแบบสอบสาระการเรียนรู้สุขศึกษาและพลศึกษาจากผลการวิเคราะห์เปรียบเทียบอัตราความถูกต้องระหว่างวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีทางสถิติกับการตรวจสอบโดยผู้เชี่ยวชาญ พบว่า วิธีที่ 1 การตรวจสอบด้วยแบบ

วินิจฉัยโดยผู้เชี่ยวชาญมีอัตราความถูกต้องโดยเฉลี่ยคิดเป็น ร้อยละ 50 และมีอัตราความคลาดเคลื่อนของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยเฉลี่ยคิดเป็นร้อยละ 50 วิธีที่ 2 การประยุกต์ใช้เทคนิคการประชุมแบบเดลฟายจากกลุ่มผู้เชี่ยวชาญ มีอัตราความถูกต้องตามฉันทามติกลุ่มเชี่ยวชาญคิดเป็นร้อยละ 0 และมีอัตราความคลาดเคลื่อนของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบคิดเป็นร้อยละ 100 วิธีที่ 3 การประยุกต์ใช้เทคนิคโปรโตคอลอะลาว์ดมีอัตราความถูกต้องโดยเฉลี่ยคิดเป็นร้อยละ 25 และมีอัตราความคลาดเคลื่อนของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยเฉลี่ยคิดเป็นร้อยละ 75

สุธาทิพย์ ตรีสิน และปิยะทิพย์ ประดุจพรม (2560) ได้เปรียบเทียบผลการตรวจสอบการทำาหน้าที่ต่างกันของข้อสอบในแบบทดสอบ ระดับชาติ ด้านภาษา ด้านคำานวณ และด้านเหตุผล ชั้นประถมศึกษาปีที่ 3 ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR โดยมีวัตถุประสงค์เพื่อวิเคราะห์คุณภาพของแบบทดสอบระดับชาติ (NT) และตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR ดังนี้ 1) วิเคราะห์คุณภาพของแบบทดสอบระดับชาติทั้ง 3 ด้าน 2) ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR และ 3) เปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการตรวจสอบ 3 วิธีข้อมูลที่นำมาใช้ วิเคราะห์เป็นข้อมูลทุติยภูมิ จากผลการตอบแบบทดสอบระดับชาติของนักเรียนชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2556 จำนวน 9,600 คน ผลการวิจัยปรากฏว่า 1) แบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 มีค่าความยากของข้อสอบ (b) อยู่ในระดับค่อนข้างยากมีค่าอำนาจจำแนกของข้อสอบ (a) อยู่ในระดับที่สามารถจำแนกผู้สอบได้ดี และมีค่าโอกาสในการเดาของข้อสอบ (c) ไม่เกิน 0.3 2) การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้ง 3 ด้าน ชี้ให้เห็นว่า เพศส่งผลให้เกิดการทำหน้าที่ต่าง กันของข้อสอบ โดยเพศหญิงจะได้เปรียบในการตอบข้อสอบด้านภาษา และด้านเหตุผล ในขณะที่เพศชาย จะได้เปรียบในการตอบข้อสอบด้านคำนวณ โดยวิธี HGLM ตรวจพบข้อสอบทำหน้าที่ต่างกัน จำนวนมากที่สุด คิดเป็นร้อยละ 69 ของข้อสอบทั้งฉบับ รองลงมาคือ วิธี IRT-LR ร้อยละ 54 และวิธี MIMIC ร้อยละ 16 ตามลำดับ 3) การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ พบว่า วิธี HGLM ตรวจพบ DIF มากกว่า วิธี MIMIC ในด้านภาษา ด้านคำนวณ และด้านเหตุผล คิดเป็นร้อยละ 70, 36 และ 53 ตามลำดับ และวิธี HGLM ตรวจ พบ DIF มากกว่าวิธี IRT-LR ด้านภาษา และด้านคำนวณ คิดเป็นร้อยละ 37 และ 13 และวิธี IRT-LR ตรวจพบ DIF มากกว่าวิธี MIMIC ทั้ง 3 ด้าน คิดเป็นร้อยละ 33, 43 และ 40 ตามลำดับ ส่วนวิธี HGLM ตรวจพบ DIF น้อยกว่า วิธี IRT-LR ด้านคำนวณ คิดเป็นร้อยละ 7 ($p<.05$)

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

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

Barnes and Wells (2009) ได้ศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบของโครงการ National Doctoral Program Survey (NDPS) ระหว่างตัวแปรเพศกับเชื้อชาติทางด้านความคิด โดยผลการวิจัยพบว่า ข้อคำถามจำนวน 29 จาก 48 ข้อ พบ DIF โดยที่เพศหญิงที่มีผิวสี มีโอกาสมากหรือน้อยที่จะมีความคิดตรงกับเพื่อนเพศชายชาวผิวขาวของเขา และควรให้มีการระมัดระวังในการใช้ข้อคำถามสำหรับกลุ่มผู้เรียนที่มีความหลากหลาย เช่น เพศ และเชื้อชาติ จึงสามารถสรุปได้ว่า เพศ และเชื้อชาติมีผลต่อความคิด และการตอบคำถามของผู้เรียนที่ศึกษาในสถานศึกษาที่มีลักษณะที่แตกต่างกัน

Le (2009) ได้ศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างเพศการอยู่นอกพื้นที่และการทดสอบภาษาในการทดสอบด้านวิทยาศาสตร์ ใน PISA โดยใช้ข้อมูลการทดลองภาคสนามของ PISA รอบ 3 เพื่อตรวจสอบความสัมพันธ์ระหว่าง การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างเพศ การอยู่นอกพื้นที่และการทดสอบภาษาในการทดสอบด้านวิทยาศาสตร์และรูปแบบอื่น ๆ ที่กำหนดไว้ใน PISA โดยมุ่งเน้นที่บริบทความสามารถและความรู้ทางวิทยาศาสตร์ข้อมูลที่ใช้ได้รับการรวบรวมจาก 60 กลุ่มภาษาทดสอบโดย 50 ประเทศที่เข้าร่วมด้วยรวมประมาณ 83,000 คน ที่เป็นนักเรียนอายุ 15 ปี งานวิจัยนี้ใช้วิธี IRT เพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างเพศสำหรับแต่ละกลุ่มภาษาและทั่วโลก ข้อสอบแบบเลือกตอบและแบบตอบกลับ แบบปิดมีแนวโน้มที่จะเข้าข้างเพศชาย การศึกษายังแสดงให้เห็นถึงผลกระทบของประเทศและการทดสอบภาษากับเพศ การทำหน้าที่ต่างกันของข้อสอบในข้อมูลระหว่างประเทศ ผลการวิจัย ปรากฏว่ามีคุณค่าการมีส่วนร่วมในการพัฒนาการทดสอบเพื่อการใช้งานระหว่างประเทศ

Taylor and Lee (2012) ได้ศึกษาการทำหน้าที่ต่างกันของข้อสอบ (DIF) สำหรับการอ่านระดับชั้นประถมศึกษาปีที่ 4 มัธยมศึกษาปีที่ 1 ของข้อสอบคณิตศาสตร์จากการทดสอบตามเกณฑ์ของรัฐ การทดสอบประกอบด้วย ข้อสอบที่มีตัวเลือกหลายตัวเลือกและสร้างการตอบสนองเพื่อตรวจสอบ DIF เมื่อจำแนกตามเพศ โดยใช้ Poly-SIBTEST และขั้นตอน Rasch โดยขั้นตอน Rasch ถูกตั้งค่าสถานะไว้เพิ่มเติมสำหรับข้อสอบที่ DIF มากกว่าที่ทำในข้อสอบพร้อมกันโดยเฉพาะ 20 ข้อสอบที่มีหลายตัวเลือก สำหรับการอ่านและการทดสอบทางคณิตศาสตร์ทั้งสองแบบ ข้อสอบที่เพศชาย ได้เปรียบในขณะที่รายการตอบสนองการสร้างที่เข้าข้างเพศหญิง การวิเคราะห์เนื้อหาแสดงให้เห็นว่ามีการอ่านค่า อ่านหนังสือที่ถูกตั้งค่าสถานะไว้ การตีความข้อความหรือความหมายโดยนัย เพศชายมีแนวโน้มที่จะได้เปรียบจากสิ่งต่าง ๆ จึงมีการระบุการตีความและการวิเคราะห์

ข้อความที่ให้ข้อมูลที่เหมาะสม ข้อสอบที่เข้าข้างเพศหญิงขอให้นักเรียนทำการตีความของตัวเองและวิเคราะห์ทั้งข้อความวรรณกรรมและข้อมูลโดยได้รับการสนับสนุนจากหลักฐานจากข้อความ การวิเคราะห์เนื้อหาของรายการคณิตศาสตร์แสดงให้เห็นว่าข้อสอบที่เข้าข้างเพศชาย คือ เรขาคณิต ความน่าจะเป็นและพีชคณิต ข้อสอบคณิตศาสตร์ที่เข้าข้างเพศหญิง การตีความทางสถิติ การแก้ปัญหาหลายขั้นตอนและการให้เหตุผลเชิงคณิตศาสตร์

สรุปจากการศึกษางานวิจัยที่เกี่ยวข้องการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นผลมาจากการตัดสินว่าข้อสอบนั้นมีความยุติธรรมหรือไม่อย่างไร ซึ่งเป็นข้อมูลสารสนเทศทางสถิติมีความสัมพันธ์ที่ส่งผลไปยังคุณลักษณะที่ข้อสอบมุ่งที่จะวัด กับมวลประสบการณ์ของผู้สอบในกลุ่มต่างๆ ที่มีบางอย่างที่แตกต่างกัน เช่น เชื้อชาติ ศาสนา วัฒนธรรม ภูมิลำเนา สังคม เพศ ภาษา อายุ และประสบการณ์เป็นต้น การทำหน้าที่ต่างกันของข้อสอบของผู้ที่เข้าสอบจะมีสัดส่วนในการตอบคำถามข้อสอบได้ถูกต้องไม่เท่ากันของแต่ละกลุ่มประชากรที่จะศึกษา และเมื่อกลุ่มผู้เข้าสอบได้คะแนนที่เท่ากันนั้นข้อสอบก็เป็นเอกพันธ์ ซึ่งการทำหน้าที่ต่างกันของข้อสอบ จะเป็นแบบพหุมิติของข้อสอบที่จะวัด

ตอนที่ 4 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการทดสอบอัตราส่วนไลค์ลิฮูด และงานวิจัยที่เกี่ยวข้อง

ทฤษฎีการตอบสนองข้อสอบ IRT นี้ สามารถจำแนกได้แป็น 2 ประเภท ได้แก่ ทฤษฎีการตอบสนองข้อสอบแบบตรวจให้คะแนน 2 ค่า (Binary or Dichotomous IRT) ซึ่งเป็นโมเดลการตอบสนองข้อสอบที่ใช้กับการตรวจสอบคะแนนรายข้อแบบ 2 ค่า เช่น ข้อสอบหรือข้อคำถามที่ตรวจให้คะแนนแบบ 0,1 (ตอบผิดได้ 0, ตอบถูกได้ 1) แบบถูก/ ผิด ใช่/ ไม่ใช่ เป็นต้น และทฤษฎี การตอบสนองข้อสอบแบบตรวจให้คะแนนมากกว่า 2 ค่า (Polytomous IRT) ซึ่งเป็นโมเดลการตอบสอบ ข้อสอบที่ใช้กับการตรวจคะแนนรายข้อแบบมากกว่า 2 ค่า เช่น ข้อสอบหรือข้อคำถามมาตรประมาณค่า (Rating Scale) การตรวจข้อสอบแบบให้คะแนนความรู้บางส่วน (Partial Credit) เป็นต้น (ศิริชัย กาญจนวาสี, 2555, หน้า 51)

โมเดลการตอบสนองข้อสอบ (IRT Models) เดิมมีแนวคิดพื้นฐานมาจากทฤษฎีการวัดแบบดั้งเดิม (Classical Test Theory) โดยโมเดลการตอบสนองข้อสอบจะวัดความสัมพันธ์ระหว่างตัวแปรอิสระ ที่ประกอบด้วยตัวแปรแฝง ดังนี้ ความสามารถของผู้สอบ (θ) หรือค่าพารามิเตอร์ของผู้สอบ (a,b,c) และตัวแปรอิสระที่เป็นตัวแปรที่สังเกตุได้นั่น คือ โอกาสในการตอบข้อสอบได้ถูกต้องของผู้สอบซึ่งทฤษฎีการตอบสนองข้อสอบ (IRT) เป็นการอธิบายความสัมพันธ์ระหว่างความสามารถภายในของตัวบุคคล (Latent Trait or Ability) กับผลของการตอบข้อสอบถูกหรือโค้งคุณลักษณะข้อสอบ (ItemCharacteristic Curve: ICC) ที่มีการกำหนดลักษณะข้อสอบด้วยพารามิเตอร์ของข้อสอบ คือค่าอำนาจจำแนกของข้อสอบ (a) ค่าความยากของข้อสอบ (b) และค่าโอกาสในการเดาของข้อสอบได้ถูก (c) เพราะเหตุนี้ IRT จึงอยู่บนพื้นฐานของความคิด 2 ประการ คือ 1) ผลการตอบข้อสอบได้ถูกต้องของผู้สอบ ที่อธิบายความสามารถที่มีอยู่ภายในตัวผู้สอบได้ และ 2) ความสัมพันธ์ระหว่างผลการตอบข้อสอบกับความสามารถที่อยู่ภายในตัวของผู้สอบ ซึ่งอธิบาย

ได้ด้วยโค้งคุณลักษณะข้อสอบ (ICC) และ ฟังก์ชันของลักษณะข้อสอบ (Embertson & Resie, 2000, p. 8)

แนวคิดทฤษฎีการตอบสนองข้อสอบ

ทฤษฎีการทดสอบเริ่มแรกมีการพัฒนามาจากผลงาน และความพยายามของนักจิตวิทยาทั้งในยุโรปและอเมริกา นักจิตวิทยาได้หาวิธีการแก้ปัญหาการวัดต่าง ๆ เพื่อพัฒนาศาสตร์แห่งการวัดและตรวจสอบจนมีความมั่นคง ในศตวรรษที่ 12 เมื่อมีเริ่มผลิตกระดาษขึ้นใช้แทนการสอบปากเปล่า และปี พ.ศ. 2293 มหาวิทยาลัยเคมบริด์จ ได้เริ่มใช้การสอบด้วยกระดาษอย่างเป็นทางการเพื่อวัดผลการเรียนของนักศึกษา มหาวิทยาลัยออกซ์ฟอร์ดได้ใช้ข้อสอบข้อเขียนสำหรับวัดจำแนกความสามารถของนักศึกษาเพื่อวัดผลระดับผ่านหรือระดับเกียรตินิยม

แนวคิดพื้นฐานของทฤษฎีการตอบสนองข้อสอบ

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

1. ค่าสถิติของข้อสอบ เช่น ความยากจะขึ้นอยู่กับลักษณะของกลุ่มผู้สอบ กล่าวคือ ถ้าผู้สอบมีความสามารถสูง ข้อสอบจะกลายเป็นข้อสอบที่ง่าย แต่ถ้าผู้สอบมีความสามารถต่ำข้อสอบดังกล่าวจะกลายเป็นข้อสอบที่ยาก ส่วนค่าอำนาจจำแนกของข้อสอบขึ้นอยู่กับความเป็นเอกพันธ์ของความสามารถของผู้สอบ ถ้าผู้สอบมีความสามารถแตกต่างกันมากข้อสอบก็จะมีค่าอำนาจจำแนกของข้อสอบสูงซึ่งมีผลทำให้ความเที่ยงของแบบทดสอบมีค่าสูงตามไปด้วย เนื่องจากความเที่ยงของแบบทดสอบมีความสัมพันธ์ทางบวกกับค่าความแปรปรวนของคะแนนจากแบบทดสอบ

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

3. ค่าความเที่ยงของแบบทดสอบถูกนิยามในรูปของผลที่ได้จากการใช้แบบทดสอบคู่ขนาน ซึ่งในทางปฏิบัติจริงนั้นนับว่าเป็นเรื่องยากที่จะให้การสอบ 2 ครั้ง มีสภาพที่เหมือนกัน ถึงแม้ว่าแบบทดสอบคู่ขนานจะขนานกันจริง แต่ผู้สอบอาจจะมีลักษณะที่แตกต่างกันไปจากการสอบครั้งแรกเกี่ยวกับแรงจูงใจ ความกังวล การลืม หรือการพัฒนาตนเองในบางทักษะ เป็นต้น

4. ทฤษฎีการทดสอบแบบดั้งเดิมไม่สามารถบอกได้ว่าผู้สอบจะตอบข้อสอบอย่างไร ยกเว้นแต่ว่าจะได้ใช้ข้อสอบข้อนั้นกับผู้สอบที่มีลักษณะคล้ายคลึงกันมาแล้ว

5. ทฤษฎีการทดสอบแบบดั้งเดิมใช้ค่าความแปรปรวนของความคลาดเคลื่อนในการวัด (Variance of Error of Measurement) เหมือนกันกับผู้สอบทุกคน ซึ่งตามความเป็นจริงแล้ว ผู้สอบที่มีความสามารถสูงและต่ำ จะมีค่าความแปรปรวนของความคลาดเคลื่อนในการวัดต่างจากผู้สอบที่มีความสามารถปานกลาง

หลักการของทฤษฎีการตอบสนองข้อสอบ

ทฤษฎีการตอบสนองข้อสอบ (Item Response Theory) มีความเชื่อเกี่ยวกับค่าพารามิเตอร์ของข้อสอบ (Item Parameter) คือ ค่าความยากของข้อสอบ (b) ค่าอำนาจจำแนก

ของข้อสอบ (a) ค่าโอกาสในการเดาข้อสอบ (c) ของข้อสอบแต่ละข้อว่าเป็นคุณลักษณะที่คงที่ในตัวข้อสอบนั้น เพราะฉะนั้นค่าพารามิเตอร์เหล่านี้จึงไม่ควรแปรเปลี่ยนไปตามกลุ่มผู้สอบ และในทำนองเดียวกัน ค่าพารามิเตอร์ของผู้สอบ (Person Parameter) หรือความสามารถที่แท้จริงของผู้สอบก็เป็นคุณลักษณะที่มีอยู่ภายในตัวผู้สอบจึงไม่ควรแปรเปลี่ยนไปตามชุดข้อสอบที่เลือกใช้ แต่เนื่องจากความสามารถของผู้สอบเป็นคุณลักษณะแฝงไม่สามารถสังเกต หรือวัดได้โดยตรง (Unobservable) จึงจำเป็นต้องใช้การทำนาย (Predict) หรืออธิบาย (Explain) คุณลักษณะดังกล่าวโดยอาศัยผลที่ได้ จากการตอบแบบทดสอบ (Test Performance) หรือคะแนน (Score) ซึ่งเป็นสิ่งที่สามารถสังเกตและวัดได้ (Observable) นักวัดผลจึงได้พยายามหาความสัมพันธ์ระหว่างผลที่ได้จากการตอบแบบทดสอบหรือคะแนน (Test Performance or Score) กับระดับความสามารถ (Ability) ของผู้ตอบแต่ละคน เพื่อเขียนเป็นโมเดลทางคณิตศาสตร์ (Mathematical Model) ความสัมพันธ์ระหว่างผลที่ได้จากการตอบแบบทดสอบกับระดับความสามารถของผู้สอบ สามารถเขียนในรูปของความสัมพันธ์ ได้ดังนี้

$$P=f(U_i/\theta_1,\theta_2,\theta_3,\ldots\ldots\theta_k;\beta_k) \qquad (5)$$

เมื่อ P แทน ผลการตอบแบบทดสอบ (Test Performance)
f แทน ฟังก์ชัน (Function)
U_i แทน ผลการตอบแบบทดสอบข้อที่ I (ตอบถูก = 1, ตอบผิด = 0)
$\theta_1,\theta_2,\theta_3,\ldots\ldots\theta_k$ แทน ระดับความสามารถ (Ability) ที่ 1, 2, 3, ...,k
β_k แทน ค่าพารามิเตอร์ของข้อสอบข้อที่ j

เนื่องจากความสัมพันธ์ดังกล่าวเป็นเพียงฟังก์ชันความสัมพันธ์ในลักษณะทั่วไป นักวัดผลการศึกษาจึงต้องหาโมเดลทางคณิตศาสตร์ที่เหมาะสม เพื่อใช้แทนฟังก์ชันความสัมพันธ์ดังกล่าว โดยอาศัยข้อตกลงเบื้องต้นของทฤษฎีการตอบสนองข้อสอบ

ข้อตกลงเบื้องต้นของทฤษฎีการตอบสนองข้อสอบ

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

1. ความเป็นเอกมิติ (Unidimentional) การวัดตามแนวทฤษฎีการตอบสนองข้อสอบได้ระบุความเป็นเอกมิติของคุณลักษณะในข้อตกลงเบื้องต้นว่า ข้อคำถาม หรือข้อสอบทุกข้อในแบบทดสอบนั้นจะต้องมุ่งวัดความสามารถ หรือคุณลักษณะเดียวเท่านั้น โดยทั่วไปแล้วข้อตกลงข้อนี้เป็นไปได้ค่อนข้างยากเนื่องจากมีปัจจัยที่มีผลต่อคะแนนสอบ เช่น ปัจจัยด้านความรู้ความเข้าใจ (Cognitive) บุคลิกภาพและปัจจัยเกี่ยวกับการจัดการสอบ ปัจจัยเหล่านี้อาจรวมถึงแรงจูงใจ ความวิตกกังวลในการสอบความสามารถในการทำงานได้รวดเร็ว ความรู้เกี่ยวกับการใช้กระดาษคำตอบเมื่อเป็นเช่นนี้สิ่งที่ทำให้ข้อตกลงนี้เป็นไปได้ คือ การพิจารณาว่าแบบทดสอบฉบับนั้นมีองค์ประกอบใดหรือปัจจัยใดที่เด่นที่สุด ก็ถือว่าแบบทดสอบได้วัดในสิ่งนั้นการตรวจสอบความเป็นเอกมิติของคุณลักษณะที่ใช้ในการทดสอบมีวิธีการตรวจสอบได้หลายวิธี ดังนี้

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

1.2 การวิเคราะห์องค์ประกอบ (Factor Analysis) ของข้อสอบทั้งฉบับ พิจารณาได้จากค่าไอเกน (Eigen Value) โดยผลการวิเคราะห์องค์ประกอบใดมีค่าไอแกนในองค์ประกอบใดองค์ประกอบหนึ่งสูงกว่าค่าอื่นอย่างชัดเจน สามารถสรุปได้ว่าข้อสอบหรือแบบสอบฉบับนั้นมีความเป็นเอกมิติของคุณลักษณะในการทดสอบ

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

1.4 การวิเคราะห์ การจัดกลุ่มระดับชั้น (Hierarchical Cluster Analysis) เป็นเทคนิคสำหรับการทดสอบความเป็นพหุมิติของชุดแบบสอบ โดยการพิจารณาการแบ่งกลุ่มของตัวแปรโดยกระบวนการแบ่งกลุ่มนี้เป็นการแบ่งกลุ่มจำนวนข้อสอบที่มีลักษณะคล้ายคลึงกันให้อยู่ในกลุ่มเดียวกัน นอกจากนี้ยังใช้วิธีในการหมุนซ้ำ (Iteration) จนสำเร็จหรืออยู่ในระดับที่น่าพอใจซึ่งทำให้ได้ความเป็นไปได้ของผลลัพธ์ (Outcome) โดยสามารถช่วยอธิบายให้ข้อมูลมีความถูกต้องมากยิ่งขึ้น ซึ่งการวิเคราะห์ด้วยวิธีการนี้สามารถใช้โปรแกรมคอมพิวเตอร์สำเร็จรูป CCPROX และ HCA ในการวิเคราะห์ได้

1.5 การใช้โปรแกรม DETECT เป็นการตรวจสอบมิติแฝงเชิงยืนยัน แบบ Nonparametric ซึ่งจะใช้ในการประมาณค่าจำนวนของมิติแฝงที่มีคุณลักษณะเด่นในชุดของข้อมูลและสามารถตรวจสอบความเป็นเอกมิติของแบบสอบ โดยระบุคุณลักษณะเด่นของมิติแฝงในแต่ละข้อซึ่งผู้ใช้โปรแกรมสามารถระบุจำนวนมิติแฝงสูงสุดที่ต้องการศึกษาได้ เนื่องจากการจัดกลุ่มชุดของข้อสอบแต่กระบวนการดังกล่าวยังมีลักษณะแบบไม่เป็นทางการเท่าใดนัก เนื่องจากการระบุการจัดกลุ่มเพื่อจำแนกความแตกต่างของมิติจะอาศัยกระบวนการในการระบุความเป็นหนึ่งเดียว

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

2. ความเป็นอิสระในการตอบข้อสอบ (Local Independent) ความเป็นอิสระในการตอบข้อสอบ หมายถึง ความน่าจะเป็นในการตอบข้อสอบแต่ละข้อได้ถูกต้องเป็นอิสระจากกัน คือการตอบข้อสอบข้อใดข้อหนึ่งถูกหรือผิดจะไม่มีผลกระทบต่อการตอบข้ออื่นด้วย หรืออาจจะกล่าวใน

เชิงคณิตศาสตร์ได้ว่า ความเป็นอิสระในการตอบข้อสอบ หมายถึง ความน่าจะเป็นในการตอบข้อสอบถูกทั้งหมดมีค่าเท่ากับ ผลคูณของความน่าจะเป็นในการตอบข้อสอบถูกเป็นรายข้อ คือ ผู้สอบที่มีความสามารถ (θ) จะมีความน่าจะเป็นที่จะตอบข้อสอบทั้งข้อ 1 และข้อ 2 ถูกเท่ากัน ซึ่งได้มาจากความน่าจะเป็นในการตอบข้อสอบข้อที่ 1 ถูก และความน่าจะเป็นในการตอบข้อสอบข้อที่ 2 ถูก คือ ถ้าผู้สอบมีความสามารถ (θ) เท่ากับ 1.5 มีความน่าจะเป็นในการตอบข้อสอบข้อที่ 1 ถูกเท่ากับ 0.5 และมีความน่าจะเป็นในการตอบข้อสอบข้อที่ 2 ถูก เท่ากับ 0.6 ดังนั้นผู้สอบที่มีความสามารถเท่ากับ 1.5 มีความน่าจะเป็นในการตอบข้อสอบทั้งสองข้อถูกภายใต้เงื่อนไขความเป็นอิสระ มีค่าเท่ากับ 0.3 และ Hambleton and Swaminathan (1985) กล่าวว่า ถ้าแบบทดสอบมีความเป็นเอกมิติอยู่แล้ว ความเป็นอิสระในการตอบข้อสอบก็จะเกิดขึ้นตามไปด้วย

3. โค้งคุณลักษณะข้อสอบ (Item Characteristic Curve) โค้งคุณลักษณะข้อสอบเป็นฟังก์ชันทางคณิตศาสตร์ สามารถใช้อธิบายความสัมพันธ์ระหว่างความน่าจะเป็น หรือโอกาสที่ผู้สอบจะตอบข้อสอบถูกกับระดับความสามารถที่วัดได้โดยใช้ชุดของข้อสอบ หรือแบบทดสอบฉบับนั้น ทั้งนี้ ความน่าจะเป็น หรือโอกาสในการตอบข้อสอบถูกจะขึ้นอยู่กับโค้งลักษณะข้อสอบในแต่ละโมเดลที่เลือกใช้ โดยที่รูปร่าง (Shape) ของโค้งคุณลักษณะข้อสอบในแต่ละข้อมีคุณสมบัติไม่แปรเปลี่ยน (Invariant) ไปตามกลุ่มตัวอย่างที่ใช้ ดังนั้น จึงทำให้ความน่าจะเป็น หรือโอกาสในการตอบข้อสอบถูกในแต่ละข้อไม่แปรเปลี่ยนด้วยคุณสมบัตินี้ถือเป็นลักษณะเด่นของโมเดลต่าง ๆ ในทฤษฎีการตอบสนองข้อสอบโค้งคุณลักษณะข้อสอบมีหลายรูปแบบขึ้นอยู่กับว่าเลือกใช้พารามิเตอร์ของข้อสอบกี่พารามิเตอร์

4. ข้อสอบที่ใช้ต้องไม่เป็นข้อสอบประเภทความเร็ว (Speediness) ผู้สอบทุกคนควรมีโอกาสในการทำข้อสอบทุกข้อ เพื่อให้คะแนนรวมจากการสอบเป็นค่าความสามารถที่แท้จริงของผู้สอบไม่มีข้อจำกัดเกี่ยวกับเวลาในการสอบสรุปได้ว่าข้อตกลงเบื้องต้นของทฤษฎีการตอบสนองข้อสอบ คือ ข้อสอบต้องมีความเป็นเอกมิติ กล่าวคือ วัดความสามารถหรือคุณลักษณะเดียว ความน่าจะเป็นในการตอบข้อสอบแต่ละข้อได้ถูกต้อง จะต้องเป็นอิสระต่อกัน นอกจากนี้ โค้งคุณลักษณะข้อสอบเป็นฟังก์ชันทางคณิตศาสตร์ที่อธิบายความสัมพันธ์ระหว่าง ความน่าจะเป็นในการตอบข้อสอบถูกกับระดับความสามารถของผู้สอบ และข้อสอบที่ใช้ต้องไม่เป็นข้อสอบประเภทความเร็ว

วิธีการทดสอบอัตราส่วนไลค์ลิฮูด

วิธีการทดสอบอัตราส่วนไลค์ลิฮูด (IRT-LR) ใช้หลักของอัลกอริทึมในการประมาณค่าความเป็นไปได้สูงสุดในการประมาณค่าพารามิเตอร์ ในข้อมูลที่จำกัดวิธีการทดสอบอัตราส่วนไลค์ลิฮูด (IRT-LR) ใช้การประมาณการกำลังสองน้อยมากสำหรับแบบจำลองการตอบสนองข้อสอบของข้อสอบแบบปกติ และยังใช้ข้อมูลที่ต่ำกว่าเกณฑ์ของการตอบสนองรูปแบบการจำแนกผู้ตอบข้อสอบ (Marie Wiberg, 2007) โดยที่วิธี IRT-LR จะประเมินความสำคัญระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบระหว่างความแตกต่างของค่าพารามิเตอร์ (Kim & Cohen, 1998) และ Thissen, David, Lynne Steinberg, and Howard Wainer (1988) เสนอไว้ว่าวิธีการทดสอบ IRT-LR เป็นที่นิยม เพราะมีการเปรียบเทียบของพารามิเตอร์และวัดพื้นที่ที่ต้องมีการประมาณการที่ถูกต้อง ในวิธี IRT-LR ข้อสอบบางข้ออาจจะเข้าข้างผู้สอบทั้งกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ซึ่งจะเรียกข้อสอบเหล่านั้นว่า DIF-

free โดยมีข้อจำกัดอยู่ระหว่างสองกลุ่มวิธี IRT-LR สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ทั้งแบบเอกรูปและอเนกซึ่งมีการเปรียบเทียบกันกับแบบจำลองที่มีข้อจำกัดเท่ากันทั้งสองกลุ่มสถิติที่ใช้ในการทดสอบ คือ ความแตกต่างระหว่างค่าความเป็นไปได้ของ -2log สำหรับกลุ่ม L_C และกลุ่ม L_A ตามค่าที่ถูกกำหนด Acar (2010) โดยสามารถเขียนสมการการวิเคราะห์ได้ ดังนี้

$$G^2(d.f.)=-2logL_c-(-2logL_A) \quad (6)$$

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

David, Lynne, and Meg (1986) ได้เสนอวิธีการทดสอบอัตราส่วนความเป็นไปได้ที่ใช้ในการตรวจสอบการทำหน้าที่เบี่ยงเบนของข้อสอบ โดยเป็นการทดสอบผลการตอบข้อสอบระหว่างผู้สอบกลุ่มย่อย 2 กลุ่ม ซึ่งวิธีทดสอบอัตราส่วนความเป็นไปได้ แบ่งเป็น 3 วิธี คือ 1) วิธีการทดสอบอัตราส่วนความเป็นไปได้ในทฤษฎีการตอบข้อสอบ ในรูปทั่วไป (General LR) ซึ่งเป็นวิธีการตรวจสอบการทำหน้าที่เบี่ยงเบนของข้อสอบซึ่งใช้การประมาณค่าพารามิเตอร์ในโมเดลการตอบข้อสอบวิธีความเป็นไปได้สูงสุดแบบมาร์จินอล (Marginal Maximum likelihood: MML) 2) วิธีการทดสอบอัตราส่วนความเป็นไปได้ในทฤษฎีการตอบข้อสอบแบบลอกลิเนียร์ (Loglinear: LR) ซึ่งเป็นการประมาณค่าพารามิเตอร์แบบวิธีความเป็นไปได้สูงสุด (Maximum likelihood: ML) และ 3) วิธีการทดสอบอัตราส่วนความเป็นไปได้ของทฤษฎีการตอบข้อสอบสารสนเทศที่มีขอบเขตแบบจำกัด (Limited Information: LR) วิธีนี้เป็นการประมาณค่าพารามิเตอร์ในโมเดลการตอบข้อสอบแบบ Normal Ogive เป็นวิธีกำลังสองน้อยที่สุดในรูปทั่วไป (Generalized least Squares: GLS) ซึ่ง 3 วิธีดังกล่าวจะใช้ในการทดสอบอัตราส่วนความเป็นไปได้เพื่อทดสอบหาค่านัยสำคัญของการทำหน้าที่เบี่ยงเบนของข้อสอบ ซึ่งความเหมาะสมของโมเดลของฟังก์ชันในทฤษฎีการตอบข้อสอบ (IRT) ซึ่งค่าความเหมาะสมของฟังก์ชันเป็นดัชนีโมเดลความเหมาะสมกับข้อมูลกับการประมาณค่าความเป็นไปได้สูงสุดของการประมาณค่าพารามิเตอร์ข้อสอบ

Embertson and Resie (2000) โมเดลการตอบสนองข้อสอบ (IRT Models) เดิมมีแนวคิดพื้นฐานมาจากทฤษฎีการวัดแบบดั้งเดิม (Classical Test Theory) โดยโมเดลการตอบสนองข้อสอบเป็นโมเดลการวัดความสัมพันธ์ระหว่างตัวแปรอิสระ ที่จะประกอบไปด้วยตัวแปรแฝง คือ ความสามารถของผู้สอบ (θ) หรือค่าพารามิเตอร์ของผู้สอบ (a,b,c) และตัวแปรอิสระที่สามารถสังเกตุได้ คือ โอกาสในการตอบข้อสอบได้ถูกต้องของผู้สอบ ทฤษฎีการตอบสนองข้อสอบ (IRT) จะเป็นการอธิบายความสัมพันธ์ความสามารถภายในของตัวบุคคล (Latent Trait or Ability) กับผลของการตอบข้อสอบถูกหรือโค้งคุณลักษณะข้อสอบ (ItemCharacteristic Curve: ICC) โดยมีการกำหนดลักษณะข้อสอบด้วยพารามิเตอร์ของข้อสอบ คือค่าอำนาจจำแนกของข้อสอบ (a) ค่าความยากของข้อสอบ (b) และค่าโอกาสในการเดาของข้อสอบได้ถูก (c) ดังนั้น ทฤษฎีการตอบสนองข้อสอบ (IRT) จึงอยู่บนพื้นฐานของความคิด 2 ประการ คือ 1) ผลการตอบข้อสอบได้ถูกต้องของผู้สอบ ที่จะอธิบายความสามารถที่มีอยู่ภายในตัวผู้สอบได้ และ 2) ความสัมพันธ์ระหว่างผล

การตอบข้อสอบกับความสามารถที่อยู่ภายในตัวของผู้สอบ โดยจะอธิบายได้ด้วยโค้งคุณลักษณะข้อสอบ (ICC) และ ฟังก์ชันของลักษณะข้อสอบ

สรุป คือ ทฤษฎีการตอบสนองข้อสอบ หรือ IRT นี้ สามารถจำแนกได้แป็น 2 ประเภท ได้แก่ ทฤษฎีการตอบสนองข้อสอบแบบตรวจให้คะแนน 2 ค่า ซึ่งเป็นโมเดลการตอบสนองข้อสอบที่ใช้กับการตรวจสอบคะแนนรายข้อแบบ 2 ค่า เช่น ข้อสอบหรือข้อคำถามที่ตรวจให้คะแนนแบบ 0, 1 (ตอบผิดได้ 0, ตอบถูกได้ 1) แบบถูก/ ผิด ใช่/ ไม่ใช่ ซึ่งโมเดลการตอบสนองข้อสอบ (IRT Models) เดิมมีแนวคิดพื้นฐานมาจากทฤษฎีการวัดแบบดั้งเดิม โดยโมเดลการตอบสนองข้อสอบเป็นโมเดลการวัดความสัมพันธ์ระหว่างตัวแปรอิสระ ที่จะประกอบไปด้วยตัวแปรแฝง คือ ความสามารถของผู้สอบ (θ) หรือค่าพารามิเตอร์ของผู้สอบ (a, b และ c) และตัวแปรอิสระที่สามารถสังเกตุได้ คือ โอกาสในการตอบข้อสอบได้ถูกต้องของผู้สอบ ทฤษฎีการตอบสนองข้อสอบ (IRT) จะเป็นการอธิบายความสัมพันธ์ความสามารถภายในของตัวบุคคล กับผลของการตอบข้อสอบถูกหรือโค้งคุณลักษณะข้อสอบ

งานวิจัยที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการทดสอบอัตราส่วนไลค์ลิฮูด มีดังนี้

รุ่งนภา แสนอำนวยผล ประกฤติยา ทักษิโณ และชนะศึก นิชานนท์ (2555) ได้ศึกษาประสิทธิภาพของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนรูปแบบผสมการประยุกต์ ใช้ทฤษฎีการตอบสนองข้อสอบแบบตรวจให้คะแนนความรู้บางส่วน และทฤษฎีการตอบสนองข้อสอบแบบตรวจให้คะแนนความรู้บางส่วนแบบทั่วไป มีวัตถุประสงค์ เพื่อศึกษาประสิทธิภาพของแบบทดสอบรูปแบบผสม และเพื่อศึกษาปฏิสัมพันธ์ระหว่างโมเดลการตรวจให้คะแนนสัดส่วนของข้อสอบที่ตรวจให้คะแนนแบบสองค่าและมากกว่าสองค่า และความยาวของแบบทดสอบ และเปรียบเทียบประสิทธิภาพของแบบทดสอบรูปแบบผสม เงื่อนไขที่ทำการศึกษามี 18 เงื่อนไข ประกอบด้วยโมเดลการตรวจให้คะแนน 2 โมเดล คือโมเดลโลจีสติก 1 พารามิเตอร์กับโมเดลการตรวจให้ คะแนนความรู้บางส่วน (PCM) และโมเดลโลจีสติก 3 พารามิเตอร์กับโมเดลการตรวจ ให้คะแนนความรู้บางส่วนแบบทั่วไป (GPCM) สัดส่วนข้อสอบที่ตรวจให้คะแนนแบบสองค่าและมากกว่าสองค่า 3 สัดส่วน คือ 20:80 50:50 และ 80:20 และความยาวของแบบทดสอบ 3 เงื่อนไข คือ 10 30 และ 50 ข้อ การประเมินประสิทธิภาพของแบบทดสอบรูปแบบ ผสมพิจารณาจากดัชนีความคลาดเคลื่อนมาตรฐานในการประมาณค่า SE(θ) ค่าความลำเอียง (Bias) พร้อมทั้งวิเคราะห์ความแปรปรวนแบบพหุจำแนก 3 ทาง (Three-way MANOVA) เพื่อเปรียบเทียบค่าเฉลี่ยของดัชนี SE(θ) และค่าความลำเอียง (Bias) พบว่า โมเดลโลจีสติก1 พารามิเตอร์กับ PCM และโมเดลโลจีสติก 3 พารามิเตอร์กับ GPCM มีค่า SE (θ) และค่าความลำเอียง (Bias) ต่ำสุดที่สัดส่วนข้อสอบที่ตรวจให้คะแนนสองค่าและมากกว่าสองค่า คือ 20:80 และความยาว ของแบบทดสอบ 50 มีปฏิสัมพันธ์ระหว่างโมเดลการตรวจให้คะแนน สัดส่วนของข้อสอบที่ตรวจให้คะแนนสองค่า และมากกว่าสองค่าและความยาวของแบบทดสอบที่ส่งผลต่อค่า SE(θ) และค่าความลำเอียง (Bias) อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ส่วนปฏิสัมพันธ์รายคู่ พบว่า ปฏิสัมพันธ์ระหว่างโมเดลการตรวจให้คะแนนกับสัดส่วนของข้อสอบที่ตรวจให้คะแนน สองค่าและมากกว่าสองค่า ระหว่างโมเดลการตรวจให้คะแนน กับความยาวของแบบทดสอบ และระหว่างสัดส่วน ของข้อสอบที่ตรวจให้คะแนนสองค่าและมากกว่าสองค่ากับ

ความยาวของแบบทดสอบ ส่งผลต่อค่า SE(θ) และค่าความลำเอียง (Bias) อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นอกจากนี้พบว่าโมเดลการตรวจให้คะแนน สัดส่วนของข้อสอบที่ตรวจให้คะแนนสองค่าและมากกว่าสองค่า และความยาวของแบบทดสอบที่แตกต่างกันส่งผลต่อค่า SE(θ) และค่าความลำเอียง (Bias) ที่ต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05

สุธาทิพย์ ตรีสิน และปิยะทิพย์ ประดุจพรม (2560) ได้ศึกษาการเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ชั้นประถมศึกษาปีที่ 3 ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR มีวัตถุประสงค์เพื่อวิเคราะห์คุณภาพของแบบทดสอบระดับชาติ (NT) และตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ด้วยวิธี HGLM วิธี MIMIC และวิธี IRT-LR ปรากฏว่า แบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 มีค่าความยากของข้อสอบ (b) อยู่ในระดับค่อนข้างยาก มีค่าอานาจจาแนกของข้อสอบ (a) อยู่ในระดับที่สามารถจำแนกผู้สอบได้ดี และมีค่าโอกาสในการเดาของข้อสอบ (c) ไม่เกิน 0.3 ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้ง 3 ด้าน ชี้ให้เห็นว่า เพศส่งผลให้เกิดการทำหน้าที่ต่างกันของข้อสอบ โดยเพศหญิงจะได้เปรียบในการตอบข้อสอบด้านภาษา และด้านเหตุผล ในขณะที่เพศชาย จะได้เปรียบในการตอบข้อสอบด้านคำนวณ โดยวิธี HGLM ตรวจพบข้อสอบทำหน้าที่ต่างกันจำนวนมากที่สุด คิดเป็นร้อยละ 69 ของข้อสอบทั้งฉบับ รองลงมาคือ วิธี IRT-LR ร้อยละ 54 และวิธี MIMIC ร้อยละ 16 ตามลำดับ และการเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ พบว่า วิธี HGLM ตรวจพบ DIF มากกว่าวิธี MIMIC ในด้านภาษา ด้านคำนวณ และด้านเหตุผลคิดเป็นร้อยละ 70, 36 และ 53 ตามลำดับ และวิธี HGLMตรวจ พบ DIF มากกว่า วิธี IRT-LR ด้านภาษา และด้านคานวณ คิดเป็นร้อยละ 37 และ 13 และวิธี IRT-LR ตรวจพบ DIF มากกว่าวิธี MIMIC ทั้ง 3 ด้าน คิดเป็นร้อยละ 33, 43 และ 40 ตามลำดับ ส่วนวิธี HGLM ตรวจพบ DIF น้อยกว่า วิธี IRT-LR ด้านคำนวณ คิดเป็นร้อยละ 7 ($p < .05$)

ธีระวัฒน์ สุขีสาร ดุษฎี โยเหลา เสกสรรค์ ทองคำบรรจง และนิยะดา จิตต์จรัส (2555) ได้ศึกษาความเที่ยงตรงของการประมาณค่าในการวิเคราะห์โมเดลสมการโครงสร้างพหุระดับภายใต้เงื่อนไขวิธีการประมาณค่าและขนาดตัวอย่างที่ต่างกัน มีวัตถุประสงค์เพื่อศึกษาความเที่ยงตรงของวิธีการประมาณค่าพารามิเตอร์สำหรับงานวิจัยโมเดลสมการโครงสร้างพหุระดับภายใต้เงื่อนไขขนาดตัวอย่างของแต่ละระดับการวิเคราะห์ที่แตกต่างกัน ซึ่งข้อมูลที่ใช้ในการวิเคราะห์ เป็นข้อมูลทุติยภูมิ โดยคัดเลือกข้อมูลจากโครงการการศึกษาโอกาสในการเรียนรู้ทางด้านคณิตศาสตร์และวิทยาศาสตร์ของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ของสถาบันส่งเสริมการสอนวิทยาศาสตร์และเทคโนโลยี (สสวท.) โดยมีการสุ่มข้อมูลตามเงื่อนไขขนาดตัวอย่างระดับที่สอง 50 โรงเรียน 30 โรงเรียนและ 15 โรงเรียน ขนาดตัวอย่างระดับที่หนึ่ง โรงเรียนละ 10 คน โรงเรียนละ 5 คนและขนาดเท่ากับแต่ละโรงเรียนที่เก็บจริง (10-40) วิเคราะห์ข้อมูลแต่ละเงื่อนไขด้วยวิธีการประมาณค่า 2 วิธี คือ Full Information Maximum Likelihood (FIML) และ Robust Maximum Likelihood (RML) และก็ตรวจสอบความเที่ยงตรงโดยการพิจารณาจากค่าการประมาณค่าน้ำหนักองค์ประกอบค่าสัมประสิทธิ์เส้นทางและค่าสัมประสิทธิ์อิทธิพลข้ามระดับโดยการเปรียบเทียบค่าสถิติกับค่าพารามิเตอร์ของประชากร ด้วยการวิเคราะห์สถิติทดสอบที่ พบว่า การทดลองความเที่ยงตรง

ของการประมาณค่าพารามิเตอร์ระหว่าง วิธีการประมาณค่าพารามิเตอร์แบบ FIML และ RML พบว่า เมื่อกลุ่มตัวอย่างระดับที่สอง 50 โรงเรียน และ 30 โรงเรียน ส่วนกลุ่มตัวอย่างระดับที่หนึ่ง โรงเรียนละ 5 คน โรงเรียนละ 10 คน และจำนวนนักเรียนที่เก็บจริงแต่ละโรงเรียน (10-40) ซึ่งทั้งสองวิธีประมาณค่าน้ำหนักองค์ประกอบ ค่าสัมประสิทธิ์เส้นทางและค่าสัมประสิทธิ์อิทธิพลข้ามระดับไม่ต่างกัน คือการประมาณค่าทั้งสองวิธีให้ผลการประมาณค่าที่มีความเที่ยงตรงไม่ต่างกันส่วนกรณีขนาดตัวอย่างระดับที่สอง 15 โรงเรียน ตัวอย่างระดับที่หนึ่งเก็บจริงแต่ละโรงเรียน (10-40 คน) วิธีการประมาณค่าแบบ RML ประมาณค่าต่าง ๆ ครบ ส่วนวิธีการประมาณค่าพารามีเตอร์แบบ FIML ไม่สามารถประมาณค่าต่าง ๆ ได้ครบ

Teresi (2007) ได้ศึกษาการประเมินผลการวัดความเท่าเทียมโดยการใช้ทฤษฎีการตอบสนองข้อสอบแบบอัตราส่วนไลค์ลิฮูด (IRT-LR) สำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ในแบบวัดความสามารถและความทุกข์ในการทำงาน ซึ่งมีเงื่อนไขในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ คือ เพศ อายุ โดยมีกลุ่มตัวอย่างเป็นผู้ป่วยมะเร็ง จำนวน 1,714 คน จะใช้เครื่องมือในการวิจัยเป็นแบบวัดความสามารถทางกายภาพ จานวน 23 ข้อ และแบบวัดทางอารมณ์ จานวน 15 ข้อ จากการศึกษาพบว่า อายุ และความสามารถ มีผลต่อการที่จะได้รับการปฏิบัติที่แตกต่างกันออกไป เป็นปัจจัยที่อาจจะส่งผลต่อผู้ป่วยที่จะเลือกเข้ารับการรักษาของสถานรักษาพยาบาล

Acar and Kelecioglu (2010) ได้ศึกษาเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ระหว่างวิธี HGLM วิธี LR และวิธี IRT-LR มีวัตถุประสงค์ของการวิจัยครั้งนี้คือ การตรวจสอบความสอดคล้องระหว่างวิธีที่ใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ 3 วิธี มีเงื่อนไขใช้ในการตรวจสอบ คือ เพศ โดยมีกลุ่มตัวอย่างเป็นนักเรียนในประเทศตุรกี ซึ่งจะใช้เครื่องมือในการวิจัยคือ แบบทดสอบของวิชาสังคมศาสตร์และวิทยาศาสตร์ จากการศึกษาพบว่า วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้ง 3 วิธี คือ วิธี HGLM วิธี LR และวิธี IRT-LR มีการตรวจพบข้อสอบที่ทำหน้าที่ต่างกันในปริมาณที่ใกล้เคียงกัน แต่ไม่ตรงกัน โดยวิธี LR และวิธี IRT-LR ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันได้น้อยทั้งแบบทดสอบสังคมศาสตร์และวิทยาศาสตร์ ส่วนวิธี HGLM ตรวจสอบพบข้อสอบที่ทำหน้าที่ต่างกันในแบบทดสอบทั้ง 2 แบบทดสอบในปริมาณที่มากที่สุด

Pae (2012) ได้ศึกษาการทำงานของข้อสอบที่แตกต่างระหว่างเพศ ในการทดสอบย่อยภาษาอังกฤษของการทดสอบความถนัดทางวิชาการของนักวิชาการวิทยาลัยเกาหลี (Korean College Scholastic Aptitude Test: KCSAT) ในช่วงเวลา 9 ปี ชุดข้อมูล 3 จุด โดยใช้ทั้งอัตราส่วน Mantel-Haenszel (MH) และทฤษฎีการตอบสนองข้อสอบแบบอัตราส่วนไลค์ลิฮูด (IRT-LR) ขั้นตอนต่าง ๆ การศึกษายังระบุถึงปัจจัย 2 ประการ ได้แก่ กลยุทธ์การอ่านและการรับรู้ความสนใจซึ่งอธิบายถึงความแปรปรวนของขนาดการทำหน้าที่ต่างกันที่มีต่อเพศ โดยใช้การวิเคราะห์การถดถอยเชิงเส้นหลายแบบ จากการศึกษา พบว่า ปฏิสัมพันธ์การทำหน้าที่ต่างกันของข้อสอบระหว่างข้อสอบรายข้อกับเพศ และความสัมพันธ์ที่สำคัญระหว่างความแตกต่างของเพศ ความสนใจผลของการทดสอบและขนาดทดสอบในการทำหน้าที่ต่างกันของข้อสอบต่อเพศ ซึ่งการรับรู้และการอ่านมีผลต่อการทำหน้าที่ต่างกันของข้อสอบต่อเพศ

Li, Hunter, and Oshima (2013) ได้ศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในการทดสอบการอ่านและเหตุผลที่เป็นไปได้ระหว่างเพศ จากการศึกษาอย่างละเอียดพบว่าข้อสอบ 1,210 ข้อ จาก 18 บทความได้รวบรวมไว้ในการวิเคราะห์ขั้นสุดท้าย พบว่า 23.3% ของข้อสอบ แสดงให้เห็นถึงการทำหน้าที่ต่างกันของข้อสอบในตัวแปรเพศ และมีการเปลี่ยนแปลงค่าเปอร์เซ็นต์ของข้อสอบที่ได้กำหนดไว้ว่าแสดงการทำหน้าที่ต่างกันของข้อสอบระหว่างการศึกษาตั้งแต่ 0 ถึง 77% ของข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบที่ครึ่งหนึ่งลำเอียงเข้าข้างเพศชายและอีกครึ่งหนึ่งที่ลำเอียงเข้าข้างเพศหญิง โดยรูปแบบที่เป็นจริง สำหรับการศึกษาโดยใช้การทดสอบการตอบสนองต่อข้อสอบ (IRT-LR) และสำหรับผู้ที่ใช้ วิธี Mantel-Haenszel (MH) นอกจากนี้แบบทดสอบที่สั้นกว่ามีแนวโน้มที่จะได้รับการพิจารณาทำหน้าที่ต่างกันของข้อสอบมากกว่าที่แบบทดสอบที่ยาวกว่า รูปแบบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบอื่น ๆ จะขึ้นอยู่กับวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ และแบบทดสอบ

Kabasakal, Arsan, Gok, and Kelecioglu (2014) ได้ศึกษาแบบจำลองเปรียบเทียบผลตราความคลาดเคลื่อนประเภทที่ 1 และประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของวิธี Mantel-Haenszel (MH) วิธี SIBTEST และวิธีการทดสอบอัตราส่วนไลค์ลิฮูด (IRT-LR) โดยมีภายใต้เงื่อนไขบางประการ ปัจจัยการจัดการคือขนาดตัวอย่างความแตกต่างระหว่างกลุ่มที่มีความสามารถความยาวแบบทดสอบร้อยละของการทำหน้าที่ต่างกันของข้อสอบ (DIF) และรูปแบบพื้นฐานที่ใช้ในการสร้างข้อมูลผลการวิจัยแสดงให้เห็นว่า SIBTEST มีข้อผิดพลาดประเภทที่สูงที่สุดในการตรวจสอบแบบข้อสอบทำหน้าที่ต่างกันแบบเอกรูป ส่วนวิธี MH มีอัตราความถูกต้องสูงภายใต้เงื่อนไขทั้งหมด นอกจากนี้ร้อยละของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) และโมเดลพื้นฐานยังมีผลที่อัตราความคลาดเคลื่อนประเภทที่ 1 ของ IRT-LR ความแตกต่างของความสามารถระหว่างกลุ่มความยาวแบบทดสอบร้อยละของ DIF รูปแบบ และความสัมพันธ์ระหว่างความแตกต่างของความสามารถร้อยละของ DIF ความแตกต่างของความสามารถความยาวแบบทดสอบความยาวในการทดสอบร้อยละของ DIF ความยาวของแบบทดสอบมีผลต่อ วิธี SIBTEST อัตราความคลาดเคลื่อนประเภทที่ 1 ในขั้นตอน MH ปัจจัยที่มีผลต่ออัตราความคลาดเคลื่อนประเภทที่ 1 ได้แก่ ขนาดตัวอย่างระยะเวลาในการทดสอบร้อยละ DIFความแตกต่างของความสามารถร้อยละของ DIF ความแตกต่างของความสามารถรูปแบบความสามารถและความแตกต่างของ DIF ไม่มีผลต่อประสิทธิภาพของ วิธี SIBTEST และ วิธี MH แต่รูปแบบพื้นฐานมีผลต่ออัตราการใช้ วิธี IRT-LR

Lopez (2012) ได้ศึกษาเรื่องตรวจสอบและจัดหมวดหมู่ของประเภท DIF ใช้ Parametric และ Nonparametric โดยการเปรียบเทียบวิธีการทดสอบอัตราส่วนไลค์ลิฮูด (IRT-LR) วิธี SIBTEST และวิธีการถดถอยโลจิสติก มีวัตถุประสงค์ของการตรวจสอบนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของสามวิธีการที่สำหรับการตรวจสอบรายการที่แตกต่างกันทำงาน (DIF) ประสิทธิภาพการทำงานของข้ามทดสอบอคติรายการพร้อมกัน ความน่าจะเป็นทฤษฎีการตอบสนองข้อสอบวิธีการทดสอบอัตราส่วนไลค์ลิฮูด (IRT-LR) และการถดถอยโลจิสติกถูกตรวจสอบในช่วงของเงื่อนไขการทดลองรวมทั้งความยาวการทดสอบที่แตกต่างกันขนาดตัวอย่าง DIF และการทดสอบการทำงานที่แตกต่างกัน (DTF) ขนาดและหมายถึงความแตกต่างในการกระจายลักษณะพื้นฐานของกลุ่มเปรียบเทียบในที่นี้จะเรียกว่าเป็นข้อมูลอ้างอิงและกลุ่มโฟกัส นอกจากนี้ในแต่ละขั้นตอนได้ดำเนินการ

โดยใช้เทียบทั้งหมดอื่น ๆ วิธีการซึ่งในรูปแบบพื้นฐานวิธีการทดสอบอัตราส่วนไลค์ลิฮูด (IRT-LR) วิธี CSIBEST การจับคู่การทดสอบย่อยและลักษณะ LOGREG ประมาณการอยู่บนพื้นฐานของการทดสอบทุกรายการยกเว้นหนึ่งภายใต้การศึกษาและวิธีการที่ผู้ประกาศข่าวอย่างต่อเนื่องซึ่งในรูปแบบพื้นฐานที่ตรงกับการทดสอบย่อย และประมาณการลักษณะอยู่บนพื้นฐานของเซตที่กำหนดไว้ล่วงหน้าระยะเวลอิสระของรายการ DIF ข้อมูลการตอบสนองสำหรับการอ้างอิงและกลุ่มโฟกัสที่ถูกสร้างขึ้นโดยใช้พารามิเตอร์ที่รู้จักกันในรายการที่อยู่บนพื้นฐานของสามพารามิเตอร์โลจิสติกรูปแบบทฤษฎีการตอบสนองข้อสอบ (3 PLM) ประเภทต่าง ๆ ของ DIF ถูกจำลองโดยขยับพารามิเตอร์การสร้างรายการของรายการที่เลือกเพื่อให้บรรลุ DIF ต้องการและขนาด DTF ขึ้นอยู่กับพื้นที่ระหว่างรายการที่กลุ่มฟังก์ชั่นการตอบสนอง อัตราความถูกต้อง ความผิดพลาดประเภทและประเภทที่สามอัตราความผิดพลาดถูกคำนวณสำหรับแต่ละสภาพขึ้นอยู่กับการทดลองซ้ำ 100 และผลการวิเคราะห์ผ่านการวิเคราะห์ความแปรปรวน ผลการศึกษาพบว่าวิธีการที่แตกต่างกันในการรับรู้ความสามารถมี LOGREG เมื่อดำเนินการโดยใช้วิธีการอื่น ๆ ทั้งหมดให้สมดุลที่ดีที่สุดของอำนาจจำแนกและอัตราความผิดพลาด การจำลองครั้งนี้ แต่ยังไม่มีวิธีการที่มีประสิทธิภาพในการระบุชนิดของ DIF

สรุปจากการศึกษางานวิจัยที่เกี่ยวข้อง วิธี IRT-LR สามารถตรวจสอบข้อสอบการทำหน้าที่ต่างกันของเพศได้มากกว่าวิธี SIBTEST และสามารถควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้ดีในแบบทดสอบที่มีความยาวไม่เกิน 30 ข้อ และตรวจพบข้อสอบการทำหน้าที่ต่างกันได้น้อยกว่าวิธี HGLM และควบคุมอัตราความคลาดเคลื่อนประเภทที่ 1 ได้น้อย เมื่อแบบทดสอบมีความยาวกว่า 30 ข้อและขนาดของกลุ่มตัวอย่างมีขนาดใหญ่

ตอนที่ 5 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีซิปเทสท์ และงานวิจัยที่เกี่ยวข้อง

วิธีซิปเทสท์ (Simulataneous Item Bias Test: SIBTEST) ได้พัฒนาโดยเชียเลียและสเตาท์ (Shealy & Stout, 1993) เพื่อนำมาใช้การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) การทำหน้าที่ต่างกันของแบบทดสอบ (Differential Test Functioning: DTF) และการทำหน้าที่ต่างกันของกลุ่มข้อสอบ (Differential Bundle Functioning: DBF) สามารถใช้ได้กับแบบทดสอบเอกมิติ (Unidimensional Test) และแบบทดสอบพหุมิติ (Multidimensional Test) (Stout, Li & Nandakumar, 1997) วิธีซิปเทสท์เป็นสถิติแบบนันพาราเมตริก (Nonparametric) ซึ่งไม่ต้องใช้ฟังก์ชันการตอบสนองข้อสอบหรือการประมาณค่าความสามารถแฝง วิธีซิปเทสท์ใช้กับข้อสอบแบบเอกรูป (Uniform DIF) จึงทำให้ไม่มีความไวในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป (Nonuniform DIF) (Li & Stout, 1996) ข้อดีของวิธีซิปเทสท์ คือ คำนวณได้ง่าย ไม่ซับซ้อน ประหยัดค่าใช้จ่าย ซึ่งไม่จำเป็นที่จะต้องใช้กับกลุ่มตัวอย่างที่มีขนาดใหญ่ และเป็นสถิติทดสอบนัยสำคัญ (Narayanan & Swaminathan, 1996) และนำไปประยุกต์ใช้กับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนแบบพหุวิภาค (Polytomous DIF)

วิธีซิปเทสท์ใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบเอกมิติจะต้องมีข้อตกลงในแบบทดสอบ คือ ข้อสอบในแบบทดสอบจะต้องมุ่งวัดคุณลักษณะ หรือความสามารถแฝง

ลักษณะเดียวเท่านั้น ซึ่งความสามารถแฝงประเภทที่หนึ่ง เรียกว่า ความสามารถเป้าหมายที่ต้องการวัด (Target Ability: θ) แต่จะมีความสามารถแฝงอีกประเภทหนึ่งที่มีอิทธิพลต่อการตอบข้อสอบ เรียกว่า ความสามารถแทรกซ้อนที่ไม่ต้องการวัด (Nuisance Ability: η) เช่น แบบทดสอบคำศัพท์ในรายวิชาภาษาต่างประเทศ ซึ่งในข้อสอบบางข้ออาจมีคำถามความรู้สำหรับผู้ชายเกี่ยวกับอิเล็กทรอนิกส์ และบางข้ออาจถามความรู้สำหรับผู้หญิงเกี่ยวกับการเย็บปักถักร้อย เป็นต้น จากแบบทดสอบดังกล่าวเป็นทักษะความรู้เกี่ยวกับคำศัพท์วิชาภาษาอังกฤษ ซึ่งเป็นเป้าหมายที่ต้องการวัด (θ) ส่วนความรู้เกี่ยวกับอิเล็กทรอนิกส์และการเย็บปักถักร้อย จัดเป็นความสามารถแทรกซ้อนที่ไม่ต้องที่จะวัด (η_1 และ η_2) ตามลำดับ ข้อสอบทุกข้อในแบบทดสอบจะวัดความสามารถเป้าหมาย แต่ข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบจะวัดทั้งความสามารถเป้าหมาย และความสามารถแทรกซ้อน

ถ้าให้ฟังก์ชันการตอบสนองข้อสอบ (IRF) ข้อที่ i ซึ่งขึ้นอยู่กับความสามารถ θ เพียงอย่างเดียวแทนด้วย $P_i(\theta)$ ส่วน IRF ข้อที่ I ที่ขึ้นอยู่กับความสามารถทั้ง θ และ η แทนด้วย $P_i(\theta, \eta)$ ฟังก์ชันการตอบสนองข้อสอบของข้อสอบดังกล่าวแบบ 3 พารามิเตอร์เป็นดังนี้ (Shealy & Stout,1993)

$$P_1(\theta) = c_i + \frac{(1-c)}{1+exp\,[-1.7a_{10}(\theta-b_{10})]}, i = 1, \ldots, N \quad (7)$$

$$P_1(\theta,\eta) = c_i + \frac{(1-c)}{1+exp\{-1.7[a_{10}(\theta-b_{10})+a_{in}(\eta-b_{in})]\}}, i = 1, \ldots, N \quad (8)$$

ดังนั้นฟังก์ชันความน่าจะเป็นอย่างมีเงื่อนไขของแบบแผนการตอบข้อสอบทั้งฉบับเป็นดังนี้

$$P[U/(\Theta=\theta,\eta=)] = \prod_{i=1}^{N} p_i(\theta,\eta)^{U_i}[1 - P_i(\theta,\eta)]^{1-Ui} \quad (9)$$

ซีเลียและสเตาท์ (Shealy & Stout, 1993) ได้ใช้ marginal IRFs อธิบายการทำหน้าที่ต่างกันของข้อสอบ

$$M_{ig}(\theta) = \int_n P_i(\theta,\eta) f_g(\eta/\theta) d\eta \quad (10)$$

เมื่อ $M_{ig}(\theta)$ = marginal IRF สำหรับความสามารถเป้าหมายที่ต้องการวัด θ ของผู้เข้าสอบกลุ่มอ้างอิงหรือกลุ่มเปรียบเทียบ

$P_i(\theta,\eta)$ = IRF ของข้อสอบข้อที่ i

$f_g(\eta \mid \theta)$ = การแจกแจงแบบมีเงื่อนไขของกลุ่มผู้เข้าสอบ

การเปรียบเทียบ marginal IRF ระหว่างกลุ่มอ้างอิง (R) กับกลุ่มเปรียบเทียบ (F) จะทำให้ทราบถึงทิศทางของการได้เปรียบหรือเสียเปรียบ กล่าวคือ ถ้า $M_{iF}(\theta) < M_{iR}(\theta)$ ทุกค่าของ θ แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบมีทิศทางเดียว โดยข้อสอบจะเข้าข้างผู้สอบกลุ่มอ้างอิง และ ถ้า $M_{iF}(\theta) >$

$M_{iR}(\theta)$ ทุกค่าของ θ แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบมีทิศทางเดียว โดยข้อสอบจะเข้าข้างผู้สอบกลุ่มเปรียบเทียบการทำทำหน้าที่ต่างกันของข้อสอบแบบมีทิศทางเดียวอาจเรียก อีกอย่างหนึ่งว่า “การทำหน้าที่ต่างกันแบบไม่ตัดกัน” (Noncrossing DIF)

วิธีชิปเทสท์มีหลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป โดยการเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ จะแบ่งแบบทดสอบออกเป็น 2 ชุด ย่อย (Subtest) คือ (1) ชุดแบบทดสอบที่มีความตรง (Valid Subtest) คือ แบบทดสอบที่ใช้เป็นเกณฑ์ในการเปรียบเทียบ (Matching Subtest) โดยแบบทดสอบประกอบด้วยข้อสอบที่ทำหน้าที่ไม่ต่างกัน (2) ชุดทดสอบที่ต้องการศึกษา (Studied Subtest) ประกอบด้วยข้อสอบที่สงสัยว่าทำหน้าที่ต่างกัน และข้อสอบชุดแรกมีจำนวน n ข้อ (ข้อที่ 1 ถึง n) แล้วแบบทดสอบชุดที่สองจะมีจำนวน N-n ข้อ (ข้อที่ n+1 ถึง N) เมื่อ N เป็นจำนวนข้อสอบทั้งหมด

ฟังก์ชันการตอบสนองข้อสอบของแบบทดสอบที่ต้องการจะศึกษาจากผู้เข้าสอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ กำหนดในรูปฟังก์ชัน marginal ดังนี้

$$M_{SR}(\theta) = \sum_{i=n+1}^{N} M_{iR}(\theta) \tag{11}$$

$$M_{SF}(\theta) = \sum_{i=n+1}^{N} M_{iF}(\theta) \tag{12}$$

เมื่อ $M_{SR}(\theta)$ แทน ผลรวมของ Marginal IRFs ของข้อสอบที่ต้องการศึกษาจากผู้เข้าสอบกลุ่มอ้างอิง ณ ระดับความสามารถ θ

เมื่อ $M_{SF}(\theta)$ แทน ผลรวม Marginal IRFs ของข้อสอบที่ต้องการศึกษาจาก ผู้เข้าสอบกลุ่มเปรียบเทียบ ณ ระดับความสามารถ θ

ขนาดของความสามารถแตกต่างระหว่าง $M_{SR}(\theta)$ กับ $M_{SF}(\theta)$ แสดงถึงปริมาณของการทดสอบของการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป หรือการทำหน้าที่ต่างกันแบบไม่ตัดกันจากชุดแบบทดสอบที่ต้องการศึกษา ณ ระดับความสามารถ θ สามารถคำนวณในรูปการอินทิเกรทดังนี้

$$\beta_{uni} = \int_{\theta}[M_{SR}(\theta) - M_{SF}(\theta)]f_p(\theta)d\theta \tag{13}$$

เมื่อ β_{uni} แทน ดัชนีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป

$f_p(\theta)$ แทน ฟังก์ชันความหนาแน่นของโอกาสการแจกแจงความสามารถของผู้สอบ θ ทั้ง 2 กลุ่ม

ดัชนี β_{uni} ที่คำนวณได้จากสูตรดังกล่าวข้างต้น นำมาทดสอบสมมติฐานของการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป ดังนี้

$$H_0 : \beta_{uni} = 0 \quad (14)$$

$$H_1 : \beta_{uni} > 0 \quad (15)$$

สมมติฐานอื่น (H_1) ใช้ทดสอบการทำหน้าที่ต่างกันของข้อสอบที่มีลักษณะทิศทางเดียวที่เข้าข้างผู้เข้าสอบกลุ่มเปรียบเทียบ และค่าประมาณของ β_{uni} คำนวณได้จากคะแนนรวมของชุดแบบทดสอบที่มีความตรงและชุดแบบทดสอบที่ต้องการศึกษา ซึ่งกำหนดด้วยสัญลักษณ์ดังนี้

$$X = \sum_{i=1}^{n} U_i \quad (16)$$

$$Y = \sum_{i=n+1}^{n} U_i \quad (17)$$

เมื่อ X แทน คะแนนรวมของชุดแบบทดสอบที่มีความตรง(ใช้เป็นเกณฑ์ในการเปรียบเทียบ

Y แทน คะแนนรวมของชุดแบบทดสอบที่ต้องการศึกษา

U_i แทน ผลการตอบข้อสอบข้อที่ I (ตอบถูกต้องได้ 1 คะแนน และตอบผิดได้ 0 คะแนน)

นำผลคะแนนเฉลี่ยของการตอบข้อสอบที่ต้องการศึกษาระหว่างผู้เข้าสอบกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบที่มีความสามารถระดับเดียวกันมาจับคู่เปรียบเทียบกัน จะพิจารณาจากคะแนนรวมที่เท่ากันของชุดแบบทดสอบที่มีความตรง (X=k) ด้วยสัญลักษณ์ดังนี้

$$\bar{Y}_{RK} - \bar{Y}_{FK} \qquad ; K = 0,1,2\ldots,n \quad (18)$$

เมื่อ $\bar{Y}_{RK}$ = ค่าเฉลี่ยของคะแนนรายข้อ จากชุดแบบทดสอบที่ต้องการของผู้เข้าสอบกลุ่มอ้างอิง ซึ่งได้คะแนน X=k

$\bar{Y}_{FK}$ = ค่าเฉลี่ยของคะแนนรายข้อ จากชุดแบบทดสอบที่ต้องการศึกษาของผู้เข้าสอบกลุ่มเปรียบเทียบ ซึ่งได้คะแนน X=k

K = คะแนนรวมจากชุดแบบทดสอบที่มีความตรง

ค่า $\bar{Y}_{RK} - \bar{Y}_{FK}$ เป็นผลแตกต่างของผลการตอบข้อสอบในชุดแบบทดสอบที่ต้องการศึกษาระหว่างผู้เข้าสอบกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบที่มีความสามารถระดับเดียวกัน

ถ้า $\bar{Y}_{RK} - \bar{Y}_{FK} = 0$ ทุกคะแนน k หมายถึงว่าข้อสอบที่ต้องการศึกษาทำหน้าที่ไม่ต่างกัน

และ ถ้า $\bar{Y}_{RK} - \bar{Y}_{FK} > 0$ ทุกคะแนน K หมายถึงข้อสอบทำหน้าที่ต่างกันแบบเอกรูป โดยจะลำเอียงเข้าข้างผู้เข้าสอบกลุ่มอ้างอิง ค่าความแตกต่างของผลการตอบข้อสอบสามารถประมาณค่าในรูป β_{uni} ดังนี้

$$\hat{\beta}_{uni} = \sum_{K=0}^{n} \hat{P}_K \left(\bar{Y}_{RK} - \bar{Y}_{FK}\right) \quad (19)$$

$$\hat{P}_k = \frac{\left(J_{Rk} + J_{Fk}\right)}{\sum_{k=0}^{n}\left(J_{Rk} + J_{Fk}\right)} \quad (20)$$

เมื่อ P_k = สัดส่วนของจำนวนผู้เข้าสอบทั้งหมด(กลุ่มอ้างอิงและกลุ่มเปรียบเทียบ) ซึ่งตอบชุดแบบทดสอบที่มีความตรงแล้วได้คะแนนรวม X = k จากจำนวนผู้สอบทั้งหมด

J_{FK} = จำนวนผู้เข้าสอบกลุ่มเปรียบเทียบซึ่งตอบชุดแบบทดสอบที่มีความตรงแล้วได้คะแนนรวม X=k

J_{RK} = จำนวนผู้เข้าสอบกลุ่มอ้างอิงซึ่งตอบชุดแบบทดสอบที่มีความตรงแล้วได้คะแนนรวม X=k (ศิริชัย กาญจนวาสี, 2555, หน้า 139-142)

สรุป คือ วิธีซิปเทสท์ใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบเอกมิติจะต้องมีข้อตกลงในแบบทดสอบ คือ ข้อสอบในแบบทดสอบจะต้องมุ่งวัดคุณลักษณะ หรือความสามารถแฝงลักษณะเดียวเท่านั้น ซึ่งความสามารถแฝงประเภทที่หนึ่ง เรียกว่า ความสามารถเป้าหมายที่ต้องการวัด (Target Ability: θ) แต่จะมีความสามารถแฝงอีกประเภทหนึ่งที่มีอิทธิพลต่อการตอบข้อสอบเรียกว่า ความสามารถแทรกซ้อนที่ไม่ต้องการวัด และวิธีซิปเทสท์มีหลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปโดยการเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ จะแบ่งแบบทดสอบออกเป็น 2 ชุด ย่อย (Subtest) คือ (1) ชุดแบบทดสอบที่มีความตรง (Valid Subtest) คือแบบทดสอบที่ใช้เป็นเกณฑ์ในการเปรียบเทียบ (Matching Subtest) โดยแบบทดสอบประกอบด้วยข้อสอบที่ทำหน้าที่ไม่ต่างกัน (2) ชุดทดสอบที่ต้องการศึกษา (Studied Subtest) ประกอบด้วยข้อสอบที่สงสัยว่าทำหน้าที่ต่างกัน และข้อสอบชุดแรกมีจำนวน n ข้อ (ข้อที่ 1 ถึง n) แล้วแบบทดสอบชุดที่สองจะมีจำนวน N-n ข้อ (ข้อที่ n+1 ถึง N) เมื่อ N เป็นจำนวนข้อสอบทั้งหมด

งานวิจัยที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีซิปเทสท์มีดังนี้

พีรญา สูงเนิน เสรี ชัดแช้ม และสมโภชน์ อเนกสุข (2552) ได้ศึกษาการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติเป็นการเปรียบเทียบระหว่างข้อสอบรายข้อกับหมวดข้อสอบโดยใช้วิธีซิบเทสท์ ซึ่งมีวัตถุประสงค์ของศึกษาวิจัยเพื่อเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบสอบพหุมิติ ระหว่างการตรวจสอบเป็นรายข้อกับรายหมวดข้อสอบภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) กลุ่มประชากรที่ศึกษาเป็นนักเรียนชั้นประถมศึกษาปีที่ 6 สังกัดเขตพื้นที่การศึกษานครศรีธรรมราช ปีการศึกษา 2546 ที่เข้าสอบวัดผลสัมฤทธิ์ระดับชาติรายวิชาภาษาไทย จำนวนที่เข้าสอบ 2,000 คน และจำนวนข้อสอบ 40 ข้อ จำแนกเป็น 2 หมวดข้อสอบ คือ หมวดที่ 1 วัดด้านโครงสร้างความรู้ จำนวน 15 ข้อ และหมวดที่ 2 วัดด้านกระบวน จำนวน 25 ข้อ จะใช้วิธีการ

สุ่มแบบแบ่งชั้น (Stratified Random Sampling) แบบจัดสรรเท่าเทียมกัน (Equal Allocation) โดยแบ่งชั้นตามระดับความสามารถเป็น 3 ระดับ คือ ดี พอใช้ และปรับปรุง ใช้นักเรียนเป็นหน่วยการสุ่ม สุ่มมาจำนวน 2,000 คน แบ่งเป็นเพศชาย จำนวน 1,000 คน และเพศหญิง จำนวน 1,000 คน การวิเคราะห์ข้อมูลค่าสถิติพื้นฐานจะใช้โปรแกรม SPSS ตรวจสอบความตรงเชิงโครงสร้างด้วยการวิเคราะห์องค์ประกอบเชิงยืนยันอันดับสอง โดยใช้โปรแกรม LISREL 8.50 และขั้นตอนต่อไปก็จะเป็นการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบและหมวดข้อสอบโดยใช้โปรแกรม SIBTEST ผลการวิจัยพบว่า (1) ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก ขนาดกลาง และขนาดใหญ่ ระหว่างการตรวจสอบการทำหน้าที่ต่างกันเป็นรายข้อกับรายหมวดข้อสอบ พบว่าข้อสอบทำหน้าที่ต่างกันแตกต่าง (2) การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบรายข้อ เมื่อกลุ่มตัวตัวอย่างขนาดเล็ก พบข้อสอบทำหน้าที่ต่างกันของข้อสอบ จำนวน 4 ข้อ คิดเป็นร้อยละ 10 ขนาดกลางพบจำนวน 13 ข้อ คิดเป็นร้อยละ 32.5 และขนาดใหญ่ พบจำนวน 15 ข้อคิดเป็นร้อยละ 37.5 (3) การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นรายหมวดข้อสอบ เมื่อกลุ่มตัวอย่างขนาดเล็ก พบว่าข้อสอบทำหน้าที่ต่างกันจำนวน 4 ข้อ คิดเป็นร้อยละ 10 ขนาดกลางพบว่าจำนวน 8 ข้อ คิดเป็นร้อยละ 20 ขนาดใหญ่ พบจำนวน 16 ข้อ คิดเป็นร้อยละ 40 และ (4) การตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบ ชี้ให้เห็นว่า หมวดที่ 2 ภายใต้เงื่อนไขกลุ่มตัวอย่างขนาดกลางมีนัยสำคัญทางสถิติ ที่ระดับ .50

ณรงค์ จันทรมหา (2554) ได้ศึกษาเปรียบเทียบค่าความเที่ยงของแบบทดสอบผลสัมฤทธิ์ทางการเรียนที่มีจำนวนข้อสอบทำหน้าที่ต่างกันแตกต่างกันโดยมีวัตถุประสงค์เพื่อเปรียบเทียบค่าความเที่ยงของแบบทดสอบผลสัมฤทธิ์ทางการเรียนที่มีจำนวนข้อสอบทำหน้าที่ต่างกันแตกต่างกัน 7 เงื่อน คือ 0% 5% 10% 15% 20% 25% และ 30% กลุ่มตัวอย่างเป็นนักเรียนชั้นประถมศึกษาปีที่ 6 สังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน ปีการศึกษา 2546 จำนวน 2,000 คน ที่ทำแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ วิชาภาษาไทย จำนวน 40 ข้อ วิเคราะห์ค่าสถิติพื้นฐานโดยใช้โปรแกรม SPSS ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้โปรแกรมซิปเทสท์ และทดสอบความแตกต่างของค่าความเที่ยงของแบบทดสอบ โดยใช้การทดสอบสถิติซี ผลการวิจัยปรากฏว่า ค่าความเที่ยงของแบบทดสอบที่มีจำนวนข้อสอบทำหน้าที่ต่างกัน 0% 5% 10% 15% 20% 25% และ 30% ไม่แตกต่างกัน

Apinyapibal, Lawthong, and Kanjanawasee (2015) ได้ศึกษาเรื่องการวิเคราะห์เปรียบเทียบประสิทธิภาพของรายการที่แตกต่างกัน การตรวจสอบการทำงานสำหรับ Dichotomously คะแนนรายการหมู่ วิธีโลจิสติกการถดถอย วิธี SIBTEST และวิธีการ Raschtree ที่แตกต่างกันส่วนใหญ่รายการที่ทำงาน (DIF) วิธีการตรวจสอบที่ใช้ในรูปแบบ Rasch จะขึ้นอยู่กับการเปรียบเทียบของพารามิเตอร์ของการทดสอบการประเมินระหว่างสองกลุ่มหรือมากกว่ากลุ่มที่มีการกำหนดไว้ล่วงหน้าเช่นกลุ่มของผู้ชาย พบว่า ความหลากหลายของวิธีการทางสถิติที่สามารถใช้ได้สำหรับการตรวจสอบรายการที่แตกต่างกันการทำงาน (DIF) ในรูปแบบ Rasch ส่วนใหญ่เหล่านี้วิธีการประกอบด้วยสองวิธี โดยวิธีแรกจะขึ้นอยู่กับการเปรียบเทียบของพารามิเตอร์รายการประมาณการกลุ่มที่กำหนดไว้ล่วงหน้าของอาสาสมัคร วิธีที่สองจะขึ้นอยู่กับการเปรียบเทียบระหว่างกลุ่มที่เป็นไปได้ทั้งหมดของอาสาสมัครโดยไม่คำนึงถึงตัวแปรคน วัตถุประสงค์ของงานวิจัยนี้

คือการเปรียบเทียบประสิทธิภาพของรายการที่แตกต่างกันในการทำงาน (DIF) การตรวจสอบระหว่าง 3 วิธี คือ (1) วิธีการถดถอยโลจิสติกตามทฤษฎีการทดสอบคลาสสิก (2) วิธี SIBTEST บนพื้นฐานของทฤษฎีการตอบสนองข้อสอบและ (3) วิธี Raschtree ขึ้นอยู่กับแบบที่ใช้การแบ่งส่วน การจำลองการตรวจหา DIF ในช่วงแรกของการศึกษาพบข้อดีและข้อเสียของทั้งสามวิธีการจากการทบทวนวรรณกรรม ขั้นที่สองของการวิจัยเปรียบเทียบประสิทธิภาพของการตรวจสอบ DIF โดยใช้การจำลองข้อมูลประเภทของรายการจำนวน DIF ความยาวทดสอบและขนาดของกลุ่มตัวอย่าง ซึ่งการวิจัยครั้งนี้ การวิจัยครั้งนี้เมื่อเทียบกับประสิทธิภาพของสามวิธีของการตรวจสอบ DIF ของรายการคะแนน Dichotomously วิธีการโลจิสติกถดถอย วิธี SIBTEST และ วิธี Raschtree โดยตรวจสอบเอกสารงานวิจัยที่เกี่ยวข้องผลการวิจัยพบว่า (1) โลจิสติกการถดถอยที่มีประสิทธิภาพในการตรวจสอบทั้งในข้อสอบทำหน้าที่ต่างกันแบบเอกรูป (Uniform) และข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูป (Non-Uniform DIF) และสามารถนำมาใช้ในการตรวจสอบคะแนน Polytomously DIF รูปแบบการวิเคราะห์มีความยืดหยุ่นและง่ายต่อการใช้งาน ก็ยังสามารถนำมาใช้ในการตรวจสอบ DIF กับหลายกลุ่มสอบและสามารถใช้สถิติในการทดสอบความสำคัญของสมมติฐาน แต่มีข้อจำกัดที่มีความมั่นคงสถิติที่แตกต่างกันตามกลุ่มตัวอย่างและคะแนนรวมเป็นเกณฑ์ในการจับคู่ (2) วิธี SIBTEST สามารถนำมาใช้ในการตรวจสอบทั้งในเครื่องแบบและนอกเครื่องแบบ DIF และการตรวจสอบการทดสอบที่แตกต่างกันทั้งทำงาน (DTF) และกำหน้าที่ต่างกัน (DBF) นอกจากนี้ยังสามารถนำมาใช้ในการตรวจสอบ Polytomously ฝ่าย DIF และการวิเคราะห์การทดสอบทั้งในมิติเดียวและหลายมิติมันสามารถคำนวณได้ง่ายไม่ซับซ้อนประหยัดและค่าใช้จ่ายและไม่ต้องใช้กลุ่มตัวอย่างที่มีขนาดใหญ่นอกจากนี้จะใช้การทดสอบนัยสำคัญทางสถิติสมมติฐานแต่มีข้อจำกัด เกี่ยวกับความไวในการตรวจสอบที่ไม่ได้ทิศทางเดียว DIF (3) จุดที่โดดเด่นของวิธี Raschtree ที่สามารถตรวจจับทั้งสองกลุ่มที่กำหนดไว้ล่วงหน้าและแฝง ผลการวิเคราะห์สามารถแสดงในรูปแบบของแผนภาพที่ชัดเจนอธิบายถึงกลุ่มรายการ และสามารถเข้าใจได้ง่ายโดยสอบ แต่มีข้อจำกัด คล้ายกับวิธีการถดถอยโลจิสติกการศึกษาต่อไปจะแสดงการเปรียบเทียบประสิทธิภาพของการตรวจสอบ DIF จากข้อมูลจำลองโดยใช้การจำลองข้อมูลภายใต้หนึ่งพารามิเตอร์รูปแบบทฤษฎีการตอบสนองรายการข้อมูลจะถูกจัดให้สอดคล้องกับปัจจัยสี่ตัวแปร: สามประเภทของรายการตัวเลขสองของ DIF สองระดับความยาวของการทดสอบและการสามระดับของขนาดตัวอย่าง จะมีทั้งหมด 36 เงื่อนไข ต้องมีการศึกษา (3 ประเภท x 2 ขนาด x 2 ขนาด x 3 ขนาด) เพื่อทำให้การเปรียบเทียบที่ชัดเจนของการรับรู้ความสามารถของสามวิธีการ

สรุปจากการศึกษางานวิจัยที่เกี่ยวข้อง วิธีซิปเทสท์สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในขนาดกลุ่มตัวอย่างที่มีขนาดเล็ก ขนาดกลาง และขนาดใหญ่ นั้น สามารถตรวจสอบพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบในกลุ่มตัวอย่างที่มีขนาดใหญ่ก็จะตรวจพบข้อสอบได้มากกว่ากลุ่มขนาดตัวอย่างขนาดเล็ก

ตอนที่ 6 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมนเทล-แฮนส์เซล และงานวิจัยที่เกี่ยวข้อง

วิธีแมนเทล-แฮนส์เซล พัฒนาโดย (Camilli & Shepard, 1994; cited Mantalel-Haenszel, 1959) เดิมเป็นสถิติสำหรับเปรียบเทียบอัตราส่วนแต้มต่อร่วม และการทดสอบอัตราส่วนเปรียบเทียบด้วยไคสแควร์ และฮอนแลน์ (Holland, 1985; cited Holland & Thayer, 1988) ได้นำวิธีแมนเทล-แฮนส์เซล ไปประยุกต์ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ และหน่วยงานการบริการทดสอบทางการศึกษาของประเทศสหรัฐอเมริกา วิธีแมนเทล-แฮนส์เซล เป็นที่ยอมรับจากนักวิจัย เป็นวิธีที่ใช้ง่าย สะดวก และประหยัด โดยใช้หลักการของตารางการณ์จรแบบทฤษฎีการตอบข้อสอบ และไม่มีการคำนวณซ้ำ สามารถนำไปใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ง่ายมีขั้นตอนการคำนวณที่ไม่สลับซับซ้อน เป็นการทดสอบทางสถิติแบบนอนพาราเมตริก ไม่จำเป็นต้องใช้โมเดลประมาณค่า

หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยวิธีแมนเทล-แฮนส์เซล คือ จะนำเอาผลการตอบของข้อสอบระหว่างผู้สอบสองกลุ่มมาเปรียบเทียบกัน นั้น คือ กลุ่มเปรียบเทียบกับกลุ่มอ้างอิง ซึ่งจะเปรียบเทียบทุกระดับความสามารถของผู้เข้าสอบกลุ่มย่อยสองกลุ่มที่มีระดับความสามารถเท่าเทียมกัน ในทางการปฏิบัติจะใช้คะแนนรวมของแบบทดสอบเป็นเกณฑ์การจับคู่กลุ่มผู้สอบ

การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบเป็นรายข้อ เมื่อจับคู่กลุ่มผู้สอบแล้ว จะนำข้อมูลผลการตอบข้อสอบนั้น ทั้งสองกลุ่มย่อยมาจัดลงตารางการณ์จรแบบ 2x2 (กลุ่มผู้สอบ 2 กลุ่ม x ผลการตอบ 2 แบบ) ตารางการณ์จร 1 ตารางแทนคะแนนรวม 1 ระดับ และถ้ามีคะแนนรวมของกลุ่มผู้เข้าสอบทั้งสิ้น k ระดับ จะต้องมีตารางการณ์ 2x2 ทั้งหมด k ตารางสำหรับตารางการณ์จรแบบ 2x2 ของข้อสอบแต่ละข้อที่มีคะแนนรวมระดับ j ดังตารางที่ 2-2

ตารางที่ 2-2 ผลการตอบข้อสอบข้อหนึ่งระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่มีคะแนนรวม j

	คะแนนผลของการตอบข้อสอบที่ศึกษา		
กลุ่ม	ตอบถูก (1)	ตอบผิด (0)	รวม
R	A_j	B_j	nR_j
F	C_j	D_j	nF_j
รวม	m_{1j}	m_{0j}	T_j

เมื่อ A_j = จำนวนผู้เข้าสอบกลุ่มอ้างอิงที่มีคะแนนรวม j ซึ่งตอบข้อสอบถูก

B_j = จำนวนผู้เข้าสอบกลุ่มอ้างอิงที่มีคะแนนรวม j ซึ่งตอบข้อสอบผิด

C_j = จำนวนผู้เข้าสอบกลุ่มเปรียบเทียบที่มีคะแนนรวม j ซึ่งตอบข้อสอบถูก

D_j = จำนวนผู้เข้าสอบกลุ่มเปรียบเทียบที่มีคะแนนรวม j ซึ่งตอบข้อสอบผิด

m_{1j} = จำนวนผู้เข้าสอบทั้งหมดที่มีคะแนนรวม j ซึ่งตอบข้อสอบถูก

m_{0j} = จำนวนผู้เข้าสอบทั้งหมดที่มีคะแนนรวม j ซึ่งตอบข้อสอบผิด
n_{Rj} = จำนวนผู้เข้าสอบกลุ่มอ้างอิงที่มีคะแนนรวม j
n_{Fj} = จำนวนผู้เข้าสอบกลุ่มเปรียบเทียบที่มีคะแนนรวม j
T_j = จำนวนผู้เข้าสอบทั้งหมดที่มีคะแนนรวม j

จากตารางที่ 2-2 นำมาคำนวณสัดส่วนของผลการตอบข้อสอบถูกและผิด ดังตารางที่ 2-3

ตารางที่ 2-3 สัดส่วนของผลการตอบข้อสอบข้อหนึ่งระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่มีคะแนน j

	คะแนนผลของการตอบข้อสอบที่ศึกษา		
กลุ่ม	ตอบถูก (1)	ตอบผิด (0)	รวม
R	pR_j	qR_j	1.00
F	pF_j	qF_j	1.00

เมื่อ pRj = สัดส่วนของผู้สอบกลุ่มอ้างอิงที่มีคะแนนรวม j ซึ่งตอบข้อสอบถูก
qR_j = สัดส่วนของผู้สอบกลุ่มอ้างอิงที่มีคะแนนรวม j ซึ่งตอบข้อสอบผิด
pF_j = สัดส่วนของผู้สอบกลุ่มเปรียบเทียบที่มีคะแนนรวม j ซึ่งตอบข้อสอบถูก
qF_j = สัดส่วนของผู้สอบกลุ่มเปรียบเทียบที่มีคะแนนรวม j ซึ่งตอบข้อสอบผิด

ในการทดสอบสมมติฐานของการทำหน้าที่ต่างกันของข้อสอบจะตั้งสมมติฐานศูนย์ (H_0) และสมมติฐานอื่น (H_1) ดังนี้

$$H_0: \frac{pR_j}{qR_j} = \frac{pF_j}{qF_j} \qquad ; j = 1,2,3,\ldots,k \tag{21}$$

$$H_1: \frac{pR_j}{qR_j} = \alpha\frac{pF_j}{qF_j} \qquad ; j = 1,2,3,\ldots,k \text{ เมื่อ } \alpha \neq 1 \tag{22}$$

สมมติฐานศูนย์ตั้งขึ้นเป็นอิสระอย่างมีเงื่อนไขระหว่างกลุ่มผู้สอบและคะแนนผลการตอบข้อสอบที่ศึกษา ดังนั้นสัดส่วนคะแนนที่ได้จากตารางที่ 3.2 ภายใต้สมมติฐานศูนย์ที่ตั้งไว้สามารถคำนวณเป็นค่าคาดหมาย (Expected values) ในแต่ละเซลได้ดังนี้

$$E(A_j) = \frac{nR_j m_{1j}}{T_j} \tag{23}$$

$$E(B_j) = \frac{nR_j m_{0j}}{T_j} \tag{24}$$

$$E(C_j) = \frac{nF_j m_{1j}}{T_j} \quad (25)$$

$$E(D_j) = \frac{nF_j m_{0j}}{T_j} \quad (26)$$

สำหรับพารามิเตอร์ α ภายใต้สมมติฐานอื่นเรียกว่า อัตราส่วนแต้มต่อร่วม (Common oddsratio) ซึ่งคำนวณได้จาก

$$\alpha = \frac{\frac{pR_j}{qR_j}}{\frac{pF_j}{qF_j}} = \frac{pR_j qF_j}{qR_j pF_j} \quad (27)$$

ถ้า α = 1 แสดงโอกาสของการตอบข้อสอบระหว่างผู้สอบทั้งสองกลุ่ม มีค่าเท่ากัน

ถ้า α > 1 แสดงว่าผู้สอบกลุ่มอ้างอิงมีโอกาสตอบข้อสอบถูกมากกว่าผู้สอบกลุ่มเปรียบเทียบ

ถ้า α < 1 แสดงว่าผู้สอบกลุ่มเปรียบเทียบมีโอกาสตอบข้อสอบถูกมากกว่าผู้สอบกลุ่มอ้างอิง

แมนเทล-แฮนเซล ได้เสนอวิธีประมาณค่า α จากตารางแบบ 2 x 2 จำนวน k ระดับ ดังนี้

$$\hat{\alpha}_{MH} = \frac{\sum_{j=1}^{K} A_j D_j / T_j}{\sum_{j=1}^{K} B_j C_j / T_j} \quad (28)$$

$\hat{\alpha}_{MH}$ เป็นค่าประมาณขนาดอิทธิพลของการทำหน้าที่ต่างกันของข้อสอบ (DIF effect size) ซึ่งมีค่าอยู่ระหว่าง 0 ถึง ∞ ฮอลแลนด์และเธเยอร์ (1985 cited in Holland ; & Thayer, 1988) ได้เสนอให้แปลง $\hat{\alpha}_{MH}$ ค่าเป็นสเกลมาตรฐานเดลต้า (Dalta scale ; MH_{D-DIF} หรือ MH_{DIF}) ดังนี้

$$MH_{DIF} = -2.35 \ln (\alpha_{MH}) \quad (29)$$

ค่า MH_{DIF} ดังกล่าวสามารถนำไปพิจารณาค่าความยากของข้อสอบ กล่าวคือ ถ้า MH_{DIF} มีค่าเป็นศูนย์ แสดงว่าข้อสอบนั้นมีความยากสำหรับแต่ละกลุ่มเท่ากัน ถ้า MH_{DIF} มีค่าเป็นลบ แสดงว่าข้อสอบยากสำหรับผู้สอบกลุ่มเปรียบเทียบมากกว่ากลุ่มอ้างอิง และถ้า MH_{DIF} มีค่าเป็นบวก แสดงว่าแสดงว่าข้อสอบยากสำหรับผู้สอบกลุ่มอ้างอิงมากกว่ากลุ่มเปรียบเทียบ ส่วนการประมาณค่าความคลาดเคลื่อนมาตรฐานของ MH_{DIF} สามารถคำนวณได้จากสูตร ดังนี้

$$SE(MH_{DIF}) = 2.35\sqrt{\text{var}[\text{In}(\hat{\alpha}_{MH})]} \quad (30)$$

$$\text{โดยที่ } \text{Var}[\text{In}(\hat{\alpha}_{MH})] = \frac{\sum_{j=1}^{k} U_j V_j / T_j^2}{2\left[\sum_{j=1}^{k} A_j D_j / T_j\right]^2} \quad (31)$$

ขณะที่ $U_i = A_jD_j+\alpha_{MH}(B_jC_j)$

และ $V_j = (A_j+D_j) +\alpha_{MH}(B_j+C_j)$

ในการทดสอบนัยสำคัญของสมมติฐาน จะต้องนำค่า α_{MH} หรือค่า MHDIF ไปทดสอบกับสถิติแมนเทล-แฮนส์เซลไค-สแควร์ (χ^2_{MH}) ที่ระดับชั้นความเป็นอิสระเท่ากับ 1 (df = 1) สถิติ χ^2_{MH} มีสูตรในการคำนวณดังนี้

$$\chi^2_{HM} = \frac{\left[|\sum_{j=1}^{k} A_j - E(A_j)| - 0.5\right]^2}{\sum_{j=1}^{k} Var(A_j)} \quad (32)$$

$$\text{โดยที่ } E(A_j) = \frac{nR_j m_{1j}}{T_j} \quad (33)$$

$$Var(A_j) = \frac{nR_j\, nF_j m_{1j} m_{0j}}{T_j^2(T_j-1)} \quad (34)$$

เมื่อ E(Aj) = ค่าคาดหมายของจำนวนผู้สอบกลุ่มอ้างอิงที่ระดับคะแนน j ซึ่งตอบข้อสอบถูก

Var(Aj) = ค่าความแปรปรวนของจำนวนผู้สอบกลุ่มอ้างอิงที่ระดับคะแนน j ซึ่งตอบข้อสอบถูก

สำหรับเกณฑ์ในการตัดสินข้อสอบที่ทำหน้าที่ต่างกัน คือ ข้อสอบที่มีค่าแตกต่างจาก 1 อย่างมีนัยสำคัญทางสถิติหรือค่าแตกต่างจาก 0 อย่างมีนัยสำคัญทางสถิติ โดยมีเกณฑ์ในการพิจารณา ดังนี้

สำหรับค่าแปลผลการทดสอบได้ดังนี้

1) ค่า = 1 หรือไม่แตกต่างจาก 1 อย่างมีนัยสำคัญ แสดงว่าข้อสอบนั้นทำหน้าที่ไม่แตกต่างกันระหว่างกลุ่ม (No DIF)

2) ค่า > 1 แสดงว่าข้อสอบนั้นทำหน้าที่แตกต่างกันระหว่างกลุ่มโดยจะเข้าข้างกลุ่มอ้างอิง

3) ค่า< 1 แสดงว่าข้อสอบนั้นทำหน้าที่แตกต่างกันระหว่างกลุ่มโดยจะเข้าข้างกลุ่มเปรียบเทียบ

ส่วนค่า MHDIF แปลผลการทดสอบได้ดังนี้

1) ค่า MHDIF เท่ากับ 0 หรือไม่แตกต่างจาก 0 อย่างมีนัยสำคัญทางสถิติ แสดงว่าข้อสอบนั้นทำหน้าที่ไม่แตกต่างกันระหว่างกลุ่ม (No DIF)

2) ค่า MHDIF แตกต่างจาก 0 อย่างมีนัยสำคัญทางสถิติ และมีค่าเป็นบวก (positive) แสดงว่าข้อสอบนั้นทำหน้าที่แตกต่างกันระหว่างกลุ่ม โดยจะเข้าข้างกลุ่มเปรียบเทียบ (F)

3) ค่า MHDIF แตกต่างจาก 0 อย่างมีนัยสำคัญทางสถิติ และมีค่าเป็นลบ (negative) แสดงว่าข้อสอบนั้นทำหน้ามีแตกต่างกันระหว่างกลุ่ม โดยจะเข้าข้างกลุ่มอ้างอิง (R) นอกจากนี้ขนาดของ |MHDIF| สามารถนำไปใช้แปลผลถึงระดับของการทำหน้าที่แตกต่างกันของข้อสอบได้
ถ้า 0 < |MHDIF| < 1.00 แสดงว่า ข้อสอบทำหน้าที่แตกต่างกันเล็กน้อย
ถ้า 1.00 ≤ |MHDIFI ≤ 1.50 แสดงว่า ข้อสอบทำหน้าที่แตกต่างกันปานกลาง แต่ถ้า |MH DIF| > 1.50 แสดงว่าข้อสอบทำหน้าที่แตกต่างกันมาก (Zieky, 1993)

การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยวิธี MH นี้ ใช้คะแนนรวมของแบบทดสอบเป็นเกณฑ์การจับคู่ ซึ่งมีจุดอ่อนในด้านความไม่เป็นอิสระของคะแนนรวมกับคะแนนรายข้อ ที่ทำการศึกษาฮอลแลนด์และเทเยอร์ (Holland & Thayer, 1988) ได้เสนอวิธีแก้จุดอ่อนดังกล่าว เพื่อทำให้เกณฑ์การจับคู่ผู้สอบระหว่างกลุ่มมีความบริสุทธิ์ยิ่งขึ้น โดยใช้วิธีการ 2 ขั้นตอน ดังนี้

ขั้นตอนแรก นำคะแนนรวมของแบบสอบทั้งฉบับเป็นเกณฑ์การจับคู่ผู้สอบระหว่างกลุ่มย่อย 2 กลุ่ม แล้ววิเคราะห์การทำหน้าที่ต่างกันของข้อสอบเป็นรายข้อ เมื่อพบว่าข้อสอบข้อใดทำหน้าที่ต่างกันให้นำคะแนนของข้อสอบข้อนั้นออกจากคะแนนรวมของผู้สอบแต่ละคน

ขั้นตอนที่สอง ใช้คะแนนรวมของแบบสอบที่นำเอาคะแนนข้อสอบที่ทำหน้าที่ต่างกันซึ่งตรวจพบในขั้นตอนแรกออกไป เพื่อใช้เป็นเกณฑ์การจับคู่แล้ววิเคราะห์การทำหน้าที่ต่างกันของข้อสอบซ้ำอีกครั้งหนึ่ง สำหรับนำไปใช้สรุปผลการตรวจสอบ

ที่ผ่านมามีผู้นำวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ วิธี MH มาศึกษา เช่น คลอสเซอร์และคณะ (Clauser et al., 1991) ใช้วิธี MH วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบอย่างสม่ำเสมอ พบว่าวิธี MH จะใช้ได้ผลดีในกรณีที่ข้อสอบมีค่าอำนาจจำแนกสูง แต่จะไม่สามารถวิเคราะห์ข้อสอบที่มีค่าความยากมากได้ ซึ่งสอดคล้องกับข้อค้นพบของเมเซอร์และคณะ (Mazor et al., 1991) ซุควีคส์ และทอลแมน (Sudweeks & Tolman, 1990) ที่พบว่าข้อสอบที่ทำหน้าที่ต่างกันมักเป็นข้อสอบที่ยาก บากิและเฟอราราา (Baghi & Ferrara, 1990) พบว่า เมื่อใช้กลุ่มตัวอย่างขนาด 750 คนขึ้นไป วิธี MH ใช้แทนวิธีทฤษฎีการตอบข้อสอบ 3 พารามิเตอร์ สวามินารานและโรเจอร์ (Swaminathan & Rogers, 1990) ได้ศึกษาเอาข้อมูลจำลองพบว่าวิธี MH วิเคราะห์ได้ดีกว่าการถดถอยแบบโลจีสเล็กน้อยโดยตรวจค้นได้ถูกต้องร้อยละ 75 กรณีใช้กลุ่มตัวอย่างจำนวน 250 คน และตรวจค้นได้ถูกต้องร้อยละ 100 กรณีกลุ่มตัวอย่างจำนวน 500 คน กรณีที่การทำหน้าที่ต่างกันของข้อสอบอย่างสม่ำเสมอ และกรณีที่ไม่สม่ำเสมอที่ติดกันปลายข้างใดข้างหนึ่ง แต่วิธี MH มีค่าใช้จ่ายง่ายน้อยกว่าวิธีโลจีสติกประมาณ 3-4 เท่า แฮมเบิลตันและคณะ (Hambleton et al., 1986) พบว่า MH ให้ค่าใกล้เคียงกับทฤษฎีการตอบข้อสอบ ทั้งที่ใช้ค่าความแตกต่างของค่าเฉลี่ยกำลัง

สองและการตรวจสอบความแตกต่างของพื้นที่รวมใต้โค้ง แต่วิธี MH มีค่าใช้จ่ายต่ำกว่าและใช้เวลาน้อยกว่าลินเคอร์ (Linacre, 1988) พบว่า ใช้ได้ดีในสถานการณ์จำลองทุกสถานการณ์พอ ๆ กับโปรแกรม PROX ของราสช์ (Rasch) แต่ควรใช้เกณฑ์ในการคำนวณการทำหน้าที่ต่างกันของข้อสอบและความคลาดเคลื่อนหลาย ๆ เกณฑ์ และใช้ตัวประมาณค่าความคลาดเคลื่อนมาตรฐานมากกว่า 1 ตัว เฟิร์ลแรนและคณะ (Perlman et al., 1988) พบว่า วิธี MH มีปัญหาด้านความเชื่อมั่น เมื่อจำนวนกลุ่มตัวอย่างน้อยกว่า 660 คน ริสเซน สไตน์เบอร์ก และวายเนอร์ (Thissen, Steinberg & Wainer, 1988) พบว่าวิธี MH ให้ผลการวิเคราะห์คล้ายกับวิธีทฤษฎีการตอบข้อสอบแบบการถดถอยเชิงเส้น และอาจใช้วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบก่อนใช้วิธีทฤษฎีการตอบข้อสอบแบบการถดถอยเชิงเส้น (ศิริชัย กาญจนวาสี, 2555, หน้า 126-131)

สรุป คือ การทำหน้าที่ต่างกันของข้อสอบโดยวิธีแมนเทล-แฮนส์เซล คือจะนำเอาผลการตอบของข้อสอบระหว่างผู้สอบสองกลุ่มมาเปรียบเทียบกัน นั้น คือ กลุ่มเปรียบเทียบกับกลุ่มอ้างอิง ซึ่งจะเปรียบเทียบทุกระดับความสามารถของผู้เข้าสอบกลุ่มย่อยสองกลุ่มที่มีระดับความสามารถเท่าเทียมกัน ซึ่งจะใช้คะแนนรวมของแบบทดสอบเป็นเกณฑ์การจับคู่กลุ่มผู้สอบ

งานวิจัยที่เกี่ยวข้องการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมนเทล-แฮนส์เซล มีดังนี้

ธเกียรติกมล ทองงอก โชติกา ภาษี และศิริชัย กาญจนวาสี (2556) ได้ศึกษาประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบภายใต้วิธีถดถอยโลจิสติกสำหรับข้อสอบที่ตรวจให้คะแนนแบบทวิวิภาค เปรียบเทียบขนาดอิทธิพลสองเกณฑ์ ที่จะเป็นการสอบเพื่อสรุปบุคคลจากคะแนนที่ได้จากเครื่องมือต้องดำเนินด้วยความยุติธรรม ซึ่งมีวัตถุประสงค์ที่จะศึกษา คือ 1) เพื่อเปรียบเทียบอัตราความถูกต้องและอัตราความคลาดเคลื่อนประเภทที่ 1 ในวิธีถดถอยโลจิสติกระหว่างการวัดขนาดอิทธิพลตามเกณฑ์ Jodoin and Gierl และการวัดขนาดอิทธิพลตามเกณฑ์ Zumbo and Thomas ภายใต้เงื่อนไขเดียวกันของปัจจัยที่แปรเปลี่ยน 4 ปัจจัยและปฏิสัมพันธ์สองทางระหว่างปัจจัยที่แปรเปลี่ยน 4 ปัจจัย 2) เพื่อเปรียบเทียบอัตราความถูกต้องและอัตราความคลาดเคลื่อนประเภทที่ 1 ในวิธีถดถอยโลจิสติก ด้วยการวัดขนาดอิทธิพลตามเกณฑ์ Jodoin and Gierl และการวัดขนาดอิทธิพลตามเกณฑ์ Zumbo and Thomas ภายใต้เงื่อนไขต่างกันของปัจจัยที่แปรเปลี่ยน 4 ปัจจัยและปฏิสัมพันธ์สองทางระหว่างปัจจัยที่แปรเปลี่ยน 4 ปัจจัย และจะมีวิธีการวิจัย ดังนี้ คือ ตัวแปรที่ศึกษาตัวแปรอิสระ (1) เกณฑ์ขนาดอิทธิพลเกณฑ์ Jodoin, Gierl, Zumbo, and Thomas (2) รูปแบบของข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูปและแบบอเนกรูป (3) ขนาดของการทำหน้าที่ต่างกันขนาด 0.1 0.2 และ 0.4 (4) จำนวนข้อสอบที่ทำหน้าที่ต่างกันทั้งฉบับคิดเป็นร้อยละ 10 และ 20 และ (5) ความยาวของแบบสอบทั้งฉบับ 40 และ 50 ข้อ ส่วนตัวแปรตาม (1) อัตราความถูกต้องคำนวณจากจำนวนของข้อสอบที่ตรวจสอบได้ถูกต้องว่าทำหน้าที่ต่างกันต่อจำนวนข้อสอบที่ทำหน้าที่ต่างกันทั้งหมดในแบบสอบคำนวณเป็นค่าร้อยละ (2) อัตราความคลาดเคลื่อนประเภทที่ 1 เป็นการระบุผิดพลาดว่าข้อสอบทำหน้าที่ต่างกัน ซึ่งในความจริงข้อสอบทำหน้าที่ไม่ต่างกันคำนวณจากสัดส่วนของจำนวนข้อสอบที่ตรวจสอบผิดพลาดว่าทำหน้าที่ต่างกันทั้งที่ในความเป็นจริงข้อสอบไม่ได้ทำหน้าที่ต่างกันคำนวณเป็นค่าร้อยละและใช้การวิเคราะห์ข้อมูลจะเป็นการวิเคราะห์ค่าพารามิเตอร์ใช้โมเดลที่ใช้ทฤษฎีการตอบสนองของข้อสอบด้วยโปรแกรม

MULTILOG-MG พัฒนาขึ้นโดย DAVID Thissen ที่มีคุณสมบัติในการประมาณค่าพารามิเตอร์ของข้อสอบและผู้สอบ ความเหมาะสมสำหรับการตอบสนองที่มีการให้คะแนนแบบทวิวิภาค วิเคราะห์คุณภาพข้อสอบด้วยโปรแกรม SPSS และวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธีถดถอยโลจิสติก โดยมีการเขียนคำสั่งประมวลด้วยโปรแกรม R จากนั้นก็นำมาเปรียบเทียบอัตราความถูกต้องและอัตราความคลาดเคลื่อนประเภทที่ 1 ที่ภายใต้ตัวแปรอิสระ โดยสถิติการวิเคราะห์ความแปรปรวนหลายตัวแปร (MANOVA) ที่ระดับนัยสำคัญ .01 จะกำหนดการวิเคราะห์ให้มีปฏิสัมพันธ์ระหว่างตัวแปรอิสระไม่เกินอันดับสอง

ผลการวิจัยปรากฏว่า 1) ประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบภายใต้วิธีถดถอยโลจิสติก ระหว่างการวัดขนาดอิทธิพลตามเกณฑ์ Jodoin, Gierl, Zumbo และ Thomas ภายใต้การจำลองตามปัจจัยที่แปรเปลี่ยนและศึกษาปฏิสัมพันธ์สองทางระหว่างปัจจัยที่แปรเปลี่ยน 4 ปัจจัย ในการวัดขนาดอิทธิพลตามเกณฑ์ Jodoin และ Gierl มีความถูกต้องสูงกว่าเกณฑ์ Zumbo และ Thomas เกือบทุกเงื่อนไข คือ (1.1) รูปแบบของข้อสอบที่ทำหน้าที่ต่างกัน และขนาดของการทำหน้าที่ต่างกัน มีผลต่ออัตราความถูกต้องของการวัดขนาดอิทธิพลตามเกณฑ์ทั้งสองเกณฑ์ (1.2) ความยาวของแบบสอบทั้งฉบับ ไม่มีผลต่ออัตราความถูกต้องและ TE เนื่องจากกำหนดความยาวของแบบสอบใกล้เคียงกันมากระหว่าง 40 กับ 50 ข้อทำให้ไม่มีความแตกต่างกัน และ (1.3) ปฏิสัมพันธ์สองทางของปัจจัยที่แปรเปลี่ยนระหว่างรูปแบบของการทำหน้าที่ต่างกันกับขนาดของการทำหน้าที่ต่างกันปฏิสัมพันธ์สองทาง ระหว่างรูปแบบของข้อสอบที่ทำหน้าที่ต่างกันกับจำนวนข้อสอบที่ทำหน้าที่ต่างกัน และปฏิสัมพันธ์สองทางระหว่างจำนวนข้อสอบที่ทำหน้าที่ต่างกันกับขนาดของการทำหน้าที่ต่างกัน มีผลต่ออัตราความถูกต้องของการวัดขนาดอิทธิพลตามเกณฑ์ทั้งสองเกณฑ์ 2) ประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบของวิธีโลจิสติก โดยการวัดขนาดอิทธิพลของการทำหน้าที่ต่างกัน ปัจจัยที่ศึกษา พบว่า (2.1) รูปแบบของข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปมีอัตราความถูกต้องสูงกว่าแบบเอกรูป รูปแบบอเนกรูปตามเกณฑ์ Jodoin and Gierl ให้ค่า TE ต่ำกว่าเกณฑ์ Zumbo and Thomas (2.2) ขนาดของการทำหน้าที่ต่างกัน มีผลต่อประสิทธิภาพการตรวจสอบระหว่างการวัดขนาดอิทธิพลตามเกณฑ์ Jodoin and Gierl และเกณฑ์ Zumbo and Thomas และ (2.3) จำนวนข้อสอบที่ทำหน้าที่ต่างกันและความยาวของแบบสอบทั้งฉบับมีผลต่อประสิทธิภาพการตรวจสอบขนาดอิทธิพลทั้งสองเกณฑ์ อย่างมีนัยสำคัญทางสถิติที่ระดับ .001

สุภาภรณ์ แดงเพ็ง ชิดชนก เชิงเชาว์ และบุญญิสา แซ่หล่อ (2554) ได้ศึกษาการเปรียบเทียบความลำเอียงของแบบทดสอบคณิตศาสตร์ในการประเมินภาพการศึกษาระดับท้องถิ่นขั้นประถมศึกษาปีที่ 5 จังหวัดปัตตานี ระหว่างวิธีแมนเทล-แฮนส์เซล และโค้งลักษณะข้อสอบ 3 พารามิเตอร์ โดยมีวัตถุประสงค์ของการวิจัย 1) เปรียบเทียบจำนวนข้อของแบบทดสอบที่ตรวจพบความลำเอียงจากการวิเคราะห์ระหว่างวิธีแมนเทล-แฮนส์เซล และวิธีโค้งลักษณะข้อสอบ 3 พารามิเตอร์ 2) ศึกษาความสอดคล้องของดัชนีความลำเอียงจากการวิเคราะห์ระหว่างวิธีแมนเทล-แฮนส์เซลและวิธีโค้งลักษณะข้อสอบ 3 พารามิเตอร์ และ 3) เปรียบเทียบความเชื่อมั่นของแบบทดสอบหลังจากคัดเลือกข้อสอบข้อที่มีความลำเอียงออกระหว่างวิธีแมนเทล-แฮนส์เซล และวิธีโค้งลักษณะข้อสอบ 3 พารามิเตอร์ พบว่า จำนวนข้อสอบที่มีความลำเอียงในแบบทดสอบใน

กลุ่มสาระการเรียนรู้คณิตศาสตร์ เมื่อวิเคราะห์ความลำเอียงด้วยวิธี 2 วิธี คือ วิธีแมนเทล-แฮนส์เซล และวิธีโค้งลักษณะข้อสอบ 3 พารามิเตอร์ ซึ่งวิธีแมนเทล-แฮนส์เซลตรวจพบข้อสอบที่ทำหน้าที่ต่างกันจำนวน 15 ข้อ มีความลำเอียงเข้าหากลุ่มอ้างอิงจำนวน 7 ข้อ และลำเอียงเข้าหากลุ่มเปรียบเทียบจำนวน 8 ข้อ ส่วนวิธีโค้งลักษณะข้อสอบ 3 พารามิเตอร์ จะวัดความแตกต่างค่าพารามิเตอร์ความยาก พบว่า มีข้อสอบที่มีความลำเอียงจำนวน 7 ข้อ ลำเอียงเข้าหากลุ่มอ้างอิงจำนวน 4 ข้อ และลำเอียงเข้าหากลุ่มเปรียบเทียบจำนวน 3 ข้อ จำนวนข้อสอบที่มีความลำเอียงจากการวิเคราะห์ ด้วยวิธีแมนเทล-แฮนส์เซลและวิธีโค้งลักษณะข้อสอบ มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ.05 และค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างดัชนีความลำเอียงของข้อสอบจากการวิเคราะห์ ด้วยวิธีแมนเทล–แฮนส์เซล และวิธีโค้งลักษณะข้อสอบ 3 พารามิเตอร์ มีค่าพารามิเตอร์ มีค่าเท่ากับ 0.846 ซึ่งอยู่ในระดับสูง และเมื่อทดสอบนัยสำคัญทางสถิติ พบว่า ดัชนีความลำเอียงจากการวิเคราะห์ด้วยทั้ง 2 วิธี มีความสอดคล้องกันสูงอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ส่วนค่าความเชื่อมั่นของแบบทดสอบก่อนการคัดเลือกข้อสอบที่มีความลำเอียงออก มีค่าความเชื่อมั่นเท่า 0.646 ค่าความยากเฉลี่ย เท่ากับ 0.318 และค่าอำนาจจำแนกเฉลี่ยเท่ากับ 0.164 และหลังจากการคัดเลือกข้อสอบที่มีความลำเอียงออก ด้วยวิธีแมนเทล-แฮนส์เซล เป็นจำนวน 15 ข้อ เมื่อนำไปทดสอบด้วยทฤษฎีการทดสอบดั้งเดิม พบว่า ข้อสอบที่เหลือ จำนวน 25 ข้อ มีค่าความเชื่อมั่นเท่ากับ 0.462 ค่าความเชื่อมั่นลดลง .184 ถือว่ายังอยู่ในระดับปานกลางมีค่าความยากเฉลี่ยเท่ากับ 0.318 ค่าอำนาจจำแนกเฉลี่ยเท่ากับ 0.121 และเมื่อวิเคราะห์ความลำเอียงข้อสอบด้วยวิธีโค้งลักษณะข้อสอบ 3 พารามิเตอร์ แล้วเลือกข้อสอบที่มีความลำเอียงจำนวน 7 ข้อ ออกเมื่อประเมินคุณภาพของแบบทดสอบโดยทฤษฎีการทดสอบแบบดั้งเดิม พบว่า ข้อสอบที่เหลือ จำนวน 33 ข้อ ซึ่งค่าความเชื่อมั่นอยู่ในระดับปานกลางเท่ากับ 0.561 ซึ่งลดลงเพียงเล็กน้อย คือ ลดลง .085 มีค่าความยากเพิ่มขึ้น .008 คือ เท่ากับ 0.310 และค่าอำนาจจำแนกเฉลี่ย 0.147 ข้อสอบที่ลำเอียงทั้ง 2 วิธี จำนวน 7 ข้อ คือ ข้อที่ 7, 16, 19, และ 28 จะลำเอียงเข้าหากลุ่มผู้สอบที่ใช้ภาษาไทยในชีวิตประจำวันเป็นข้อสอบเกี่ยวกับตัวประกอบของจำนวน คุณสมบัติของรูปเรขาคณิต ทรงเรขาคณิต และโจทย์ปัญหาการค้าขาย และข้อที่ 21, 34 และ 40 ลำเอียงเข้าหากลุ่มผู้สอบที่ใช้ภาษามลายูถิ่นปัตตานี เป็นข้อสอบที่เกี่ยวกับการเรียงลำดับเหตุการณ์ การใช้แผนภูมิ ร้อยละ ใช้สัญลักษณ์และเน้นตัวเลข

Awuor (2008) ได้ศึกษาเรื่องผลของขนาดตัวอย่างไม่เท่ากันในพลังของการตรวจสอบ DIF กับ วิธี IRT-Based ศึกษา Monte Carlo กับวิธี SIBTEST และวิธี Mantel-แฮนส์เซล พบว่า การจำลองการศึกษานี้มุ่งเน้นไปที่การกำหนดผลของขนาดตัวอย่างที่ไม่เท่ากันในอัตราความถูกต้องทางสถิติของวิธี SIBTEST และวิธี Mantel-Haenzsel ในการตรวจหา DIF ของขนาดปานกลางและขนาดใหญ่ ข้อสอบเป็นแบบ พารามิเตอร์โดยสใช้แบบ 2 PLM โดยใช้โปรแกรม WinGen2 (Han, 2006) ถูกใช้ในการจำลองการประมาณความสามารถและการสร้างข้อมูลการตอบสนองที่ถูกวิเคราะห์โดย วิธี SIBTEST ขั้นตอนวิธี SIBTEST กับกระบวนการการถดถอยการตรวจความถูกต้องใช้ในการคำนวณสถิติ DIF คือขนาดผล DIF และสถิติ ความสำคัญของการมีลำเอียง วิธี SIBTEST เป็นวิธีที่ถูกนำมาใช้ในการคำนวณสถิติ DIF สำหรับ ขั้นตอนวิธี Mantel-Haenzsel โดยมีเงื่อนไขในสิ่งแวดล้อมที่พารามิเตอร์สามารถถูกจำลองข้อมูลการตอบสนองและการดำเนินการสร้าง DIF

วิเคราะห์ ข้อสอบที่ใช้ทดสอบเพื่อตรวจสอบพบว่ารายการที่แสดงให้เห็นถึงเบื้องต้นของการทำหน้าที่ต่างกันของข้อสอบ ผลการศึกษาชี้ให้เห็นว่ามีตัวอย่างที่ขนาดต่างกันในอัตราส่วนใด ๆ วิธี Mantel-Haenzsel มีข้อผิดพลาดประเภทที่ 1 ดีขึ้นกว่า วิธี SIBTEST ผลยังชี้ให้เห็นว่าไม่เพียงแต่สัดส่วนแต่ยังขนาดตัวอย่างและขนาดของการทำหน้าที่ต่างกันของข้อสอบที่มีอิทธิพลพฤติกรรมของวิธี SIBTEST และ วิธี Mantel-Haenzsel เกี่ยวกับการพฤติกรรมอัตราความผิดพลาด

Acar (2011) ได้ศึกษาขนาดของกลุ่มตัวอย่างของการทำหน้าที่ต่างกันของข้อสอบ โดยใช้วิธี Hierarchical Linear Modeling (HGLM) เนื่องมาจากวิธี Hierarchical Linear Modeling (HGLM) วิธี Mantel-Haenszel วิธี Logistics Regression มีลักษณะที่มีความคล้ายคลึงกันจึงเลือกใช้วิธี HGLM ในการหาความแตกต่างกันของกลุ่มตัวอย่างประชากรที่ทำการศึกษา คือ นักเรียนที่ทำการสอบ OKS ในปี 2006 จำนวน 798,307 คน โดยสุ่มตัวอย่างจำนวน 10,727 คน ทำการทดสอบในวิชาภาษาตุรกี วิชาวิทยาศาสตร์และวิชาสังคม วิชาละ 25 ข้อ พบว่า กลุ่มตัวอย่าง จำนวน 2,681 คน เริ่มส่งผลให้เกิดการทำหน้าที่ต่างกันของข้อสอบโดยเมื่อจำนวนกลุ่มตัวอย่างที่เพิ่มมากขึ้นส่งผลให้การทำหน้าที่ต่างกันของข้อสอบเพิ่มสูงขึ้น

Steinmayr, Bergold, and Stiksrud (2015) ได้ศึกษาเรื่อง เพศที่แตกต่างในการทดสอบความรู้ทั่วไปของ การทำงานที่แตกต่างกันของข้อสอบ พบว่า เพศที่แตกต่างในการทดสอบความรู้ความนิยมในหมู่คนที่มีความแตกต่างทางเพศมีเสถียรภาพมากที่สุดที่พบในมาตรการความสามารถทางปัญญาแม้ว่าหลายคนพยายามที่จะทำให้อธิบายการค้นพบนี้การศึกษาส่วนใหญ่ยังไม่ได้พิจารณาอย่างพอเพียง ด้านระเบียบวิธีการที่แตกต่างกันเช่นรายการที่ทำงาน (DIF) การศึกษาครั้งนี้ตรวจสอบว่ามีการทดสอบความรู้ทั่วไปชาวเยอรมันจะแสดงให้เห็นความแตกต่างทางเพศและไม่ว่าจะเป็นเพศที่แตกต่างเหล่านี้อาจจะอธิบายได้ด้วย DIF ด้วยเหตุนี้เราบริหารงานการทดสอบความรู้ เพื่อเป็นตัวอย่างของ N = 977 เยอรมันนักเรียนมัธยมปลายเราสังเกตเห็นความแตกต่างทางเพศขนาดใหญ่ในคะแนนรวมของทั่วไปการทดสอบความรู้ (|d| = 0.78) บนพื้นฐานของวิธีการตรวจสอบ nonparametric DIF พบว่า 40 จาก 84 รายการที่แสดงให้เห็นจำนวนมากของ DIF ขจัดรายการเหล่านั้นจากคะแนนการทดสอบความรู้โดยรวมลดความแตกต่างทางเพศที่จะสังเกต |d| = 0.32 ผลการค้นหาจะมีการหารือเกี่ยวกับการแตกต่างทางเพศในการทดสอบความรู้ในการพิจารณาทั่วไปและระเบียบวิธีการ

สรุปจากการศึกษางานวิจัยที่เกี่ยวข้อง การทำหน้าที่ต่างกันของข้อสอบโดยวิธีแมนเทล-แฮนส์เซล กลุ่มตัวอย่างของการทำหน้าที่ต่างกันของข้อสอบ โดยเมื่อจำนวนกลุ่มตัวอย่างที่เพิ่มมากขึ้นจะส่งผลให้การทำหน้าที่ต่างกันของข้อสอบเพิ่มสูงขึ้น

บทที่ 3
วิธีดำเนินการวิจัย

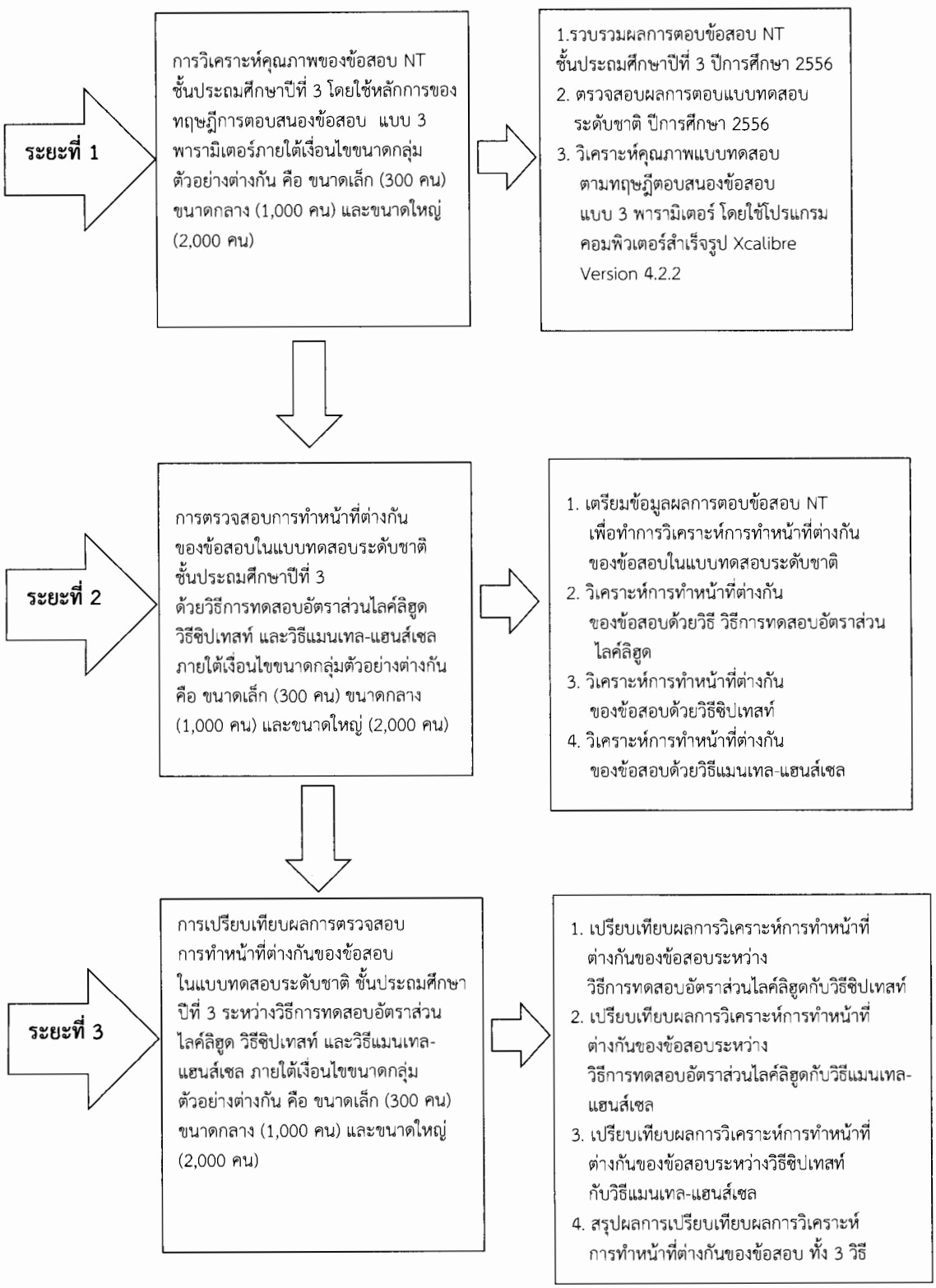
การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อวิเคราะห์คุณภาพของข้อสอบ NT ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ และเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ของนักเรียนชั้นประถมศึกษาปีที่ 3 จำนวน 3 ด้าน คือ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) ด้วยวิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล ผู้วิจัยเสนอวิธีการดำเนินการวิจัย แบ่งออกเป็น 3 ระยะ ดังนี้

ระยะที่ 1 การวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 โดยใช้หลักการของทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)

ระยะที่ 2 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้วยวิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)

ระยะที่ 3 การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้วยวิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)

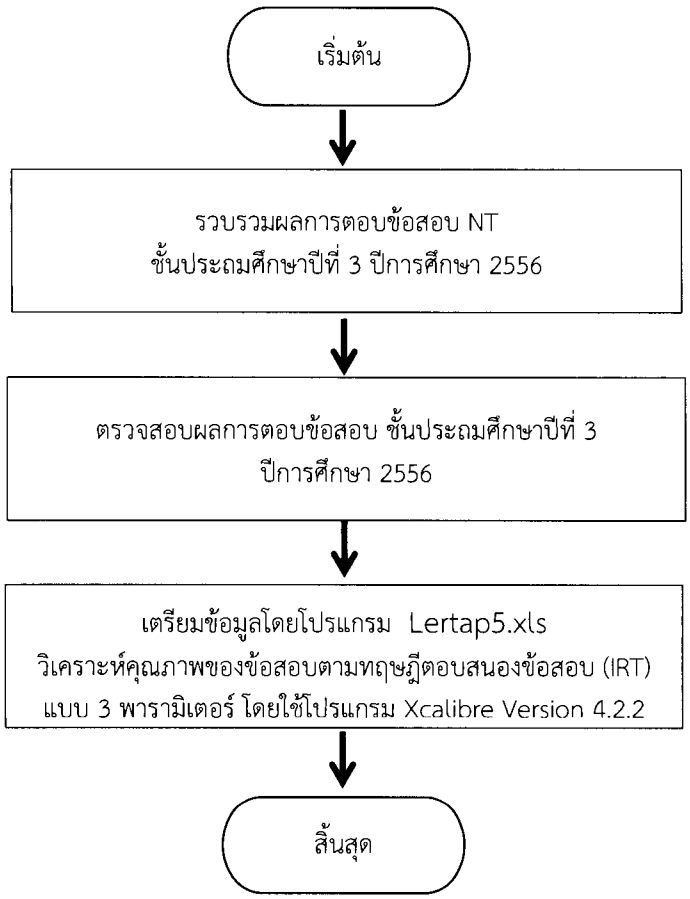
โดยมีวิธีดำเนินการวิจัยแบ่งเป็น 3 ระยะ ดังภาพที่ 3-1



ภาพที่ 3-1 ขั้นตอนการดำเนินงานวิจัย

ระยะที่ 1 การวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 โดยใช้หลักการของทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)

การวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) ดังภาพที่ 3-2



ภาพที่ 3-2 ขั้นตอนการวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 โดยใช้หลักการของทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์

จากภาพที่ 3-2 แสดงขั้นตอนการวิเคราะห์คุณภาพของ NT ชั้นมัธยมศึกษาปีที่ 3 โดยใช้หลักการของทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) มีขั้นตอนดังนี้

1. ผู้วิจัยได้ดำเนินการขอหนังสือขอความอนุเคราะห์ข้อมูลผลการตอบแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ปีการศึกษา 2556 จากวิทยาลัยวิทยาการวิจัยและวิทยาการปัญญา

ถึงสำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน (สพฐ.)

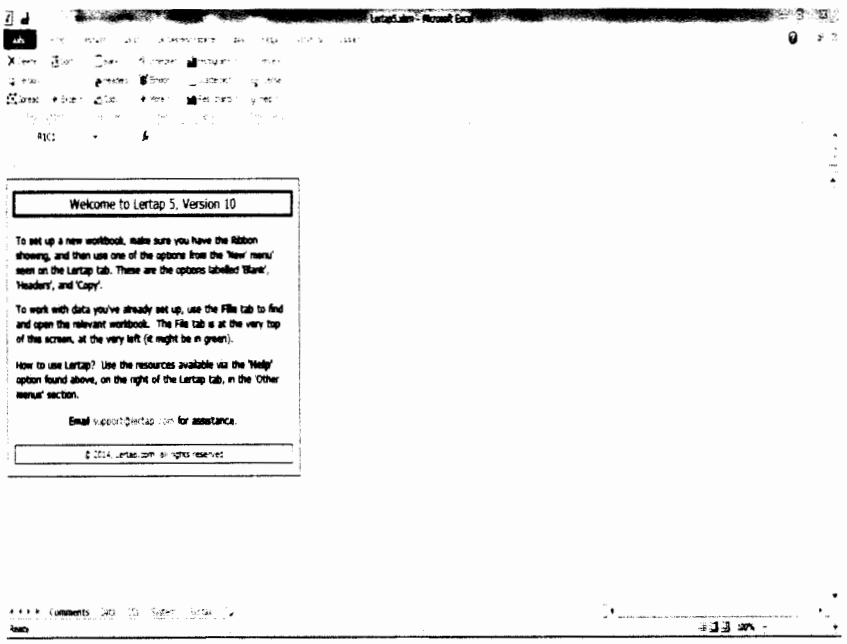
2. ตรวจสอบผลการตอบแบบทดสอบระดับชาติ ปีการศึกษา 2556 ทั้งโจทย์คำถาม ตัวเลือกคำตอบ เฉลยคำตอบ และคำตอบของผู้เข้าสอบมีครบถ้วนในการตอบแบบทดสอบ ทั้ง 3 ด้าน คือ ด้านภาษา ด้านคำนวณ และด้านเหตุผล

3. วิเคราะห์คุณภาพของข้อสอบ NT ตามทฤษฎีตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ โดยใช้โปรแกรม Lertap5.xls มีขั้นตอน ดังนี้

3.1 เริ่มต้นด้วยเวอร์ชัน 5 ไฟล์ Excel xls จะเรียกว่า "งาน" สมุดงานคือชุดของแผ่นงานจะเห็นตัวอย่างดังกล่าวเมื่อเปิดไฟล์ Lertap5.xls ดังภาพที่ 3-3

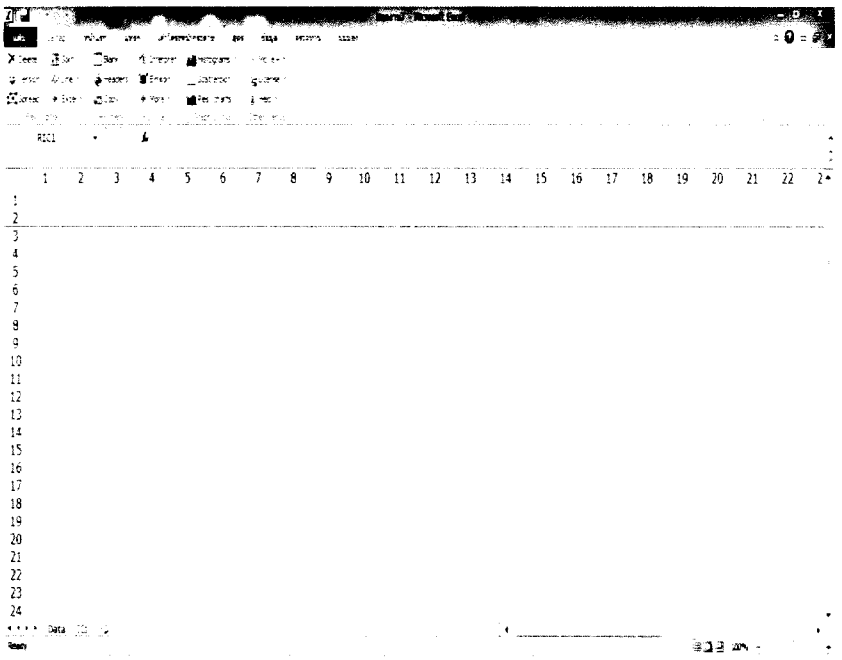
3.2 หนึ่งแถบเมนูและแถบเครื่องมือสองแถบจะปรากฏที่ด้านบนของหน้าจอ แถบเมนูเริ่มจากตัวเลือกที่ชื่อ File, Edit และ View

3.3 ไฟล์ Lertap5.xls มี 5 แผ่นงานที่มองเห็นได้ โดยมีชื่อจะแสดงอยู่ด้านบนเป็น- Comments, Data, CCs, System และ Syntax ในแท็บแผ่นงานที่เกี่ยวข้องซึ่งปรากฏที่ด้านล่างของหน้าจอ



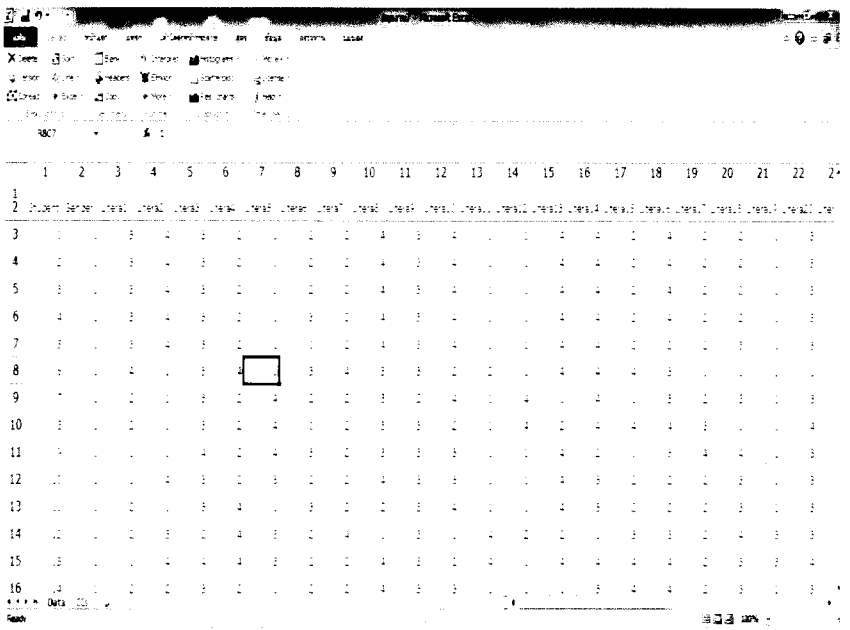
ภาพที่ 3-3 ตัวอย่างเปิดไฟล์ Lertap5.xls

3.4 คลิก Bank ดังภาพที่ 3-4



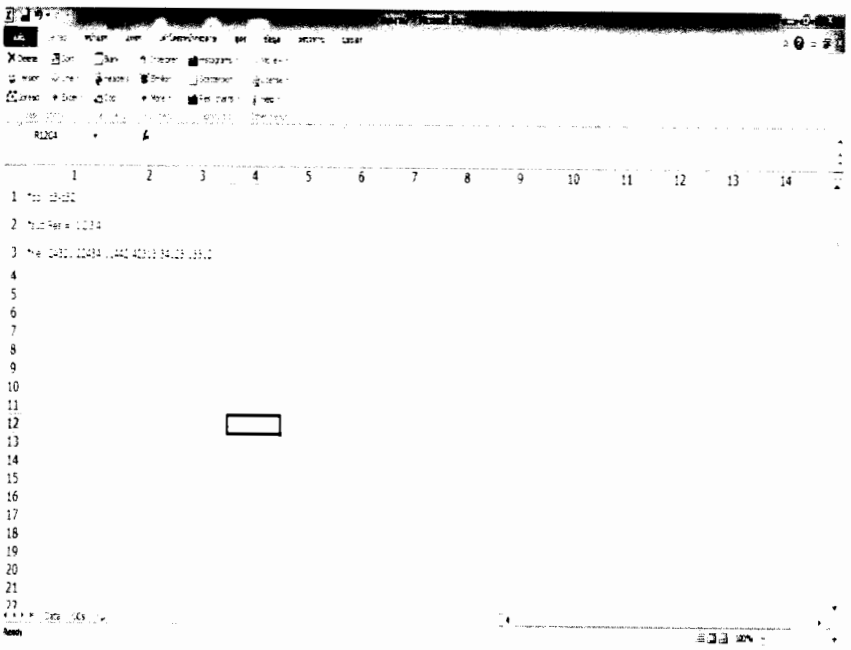
ภาพที่ 3-4 ตัวอย่างข้อมูลที่ต้องการจะวิเคราะห์ไฟล์ Lertap5.xls

3.5 เปิดไฟล์ที่บันทึกไว้ใน Excel แล้ว copy ข้อมูลมาวางไว้ใน Data ดังภาพที่ 3-5



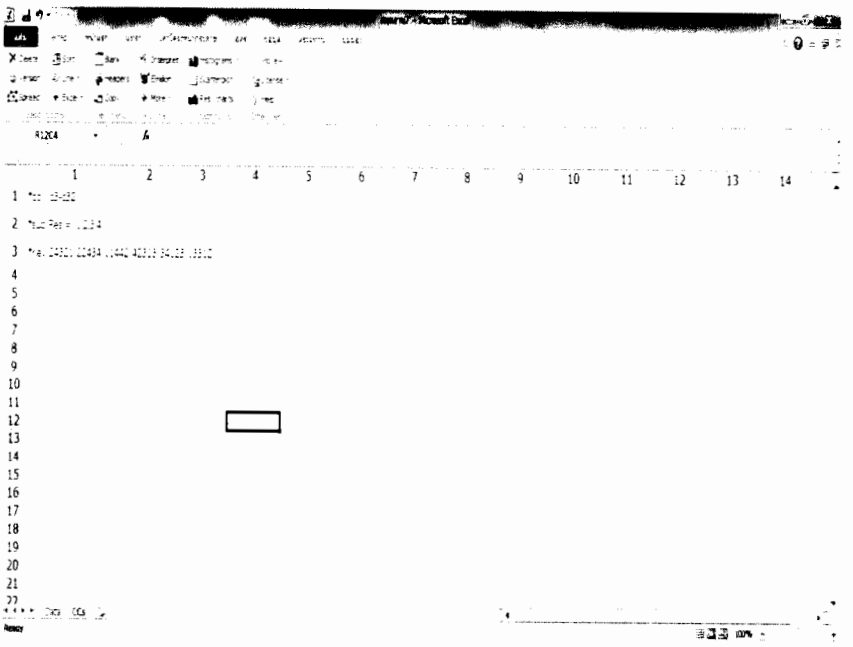
ภาพที่ 3-5 ตัวอย่างข้อมูลที่บันทึกไว้ใน Excel ที่จะวิเคราะห์ไฟล์

3.6 เปิดไฟล์ที่บันทึกไว้ใน Excel แล้ว copy เฉลยแบบทดสอบมาวางไว้ใน CCs ดังภาพที่ 3-6



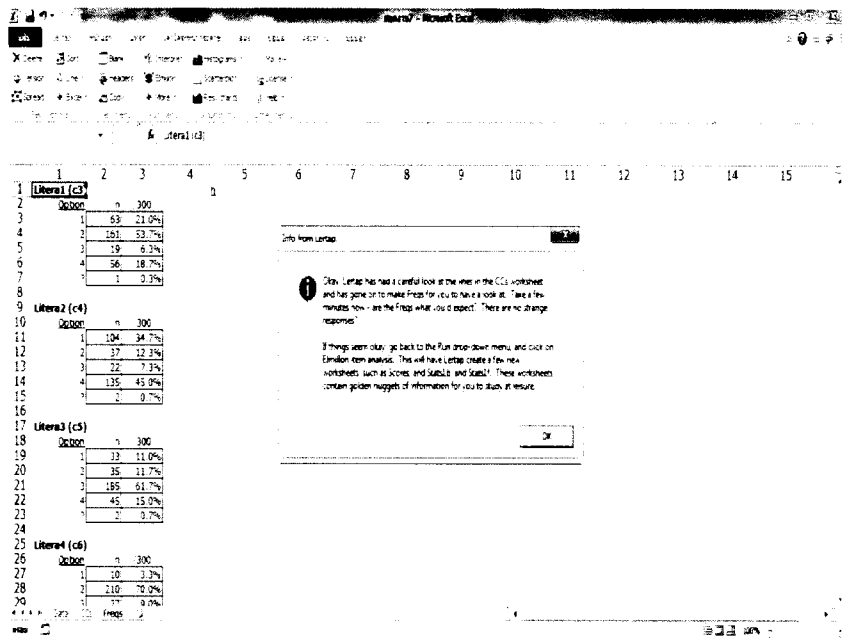
ภาพที่ 3-6 ตัวอย่างข้อมูลแผ่นงาน CCs ที่มีเฉลยคำตอบ

3.7 เลือก Interpret ดังภาพที่ 3-7



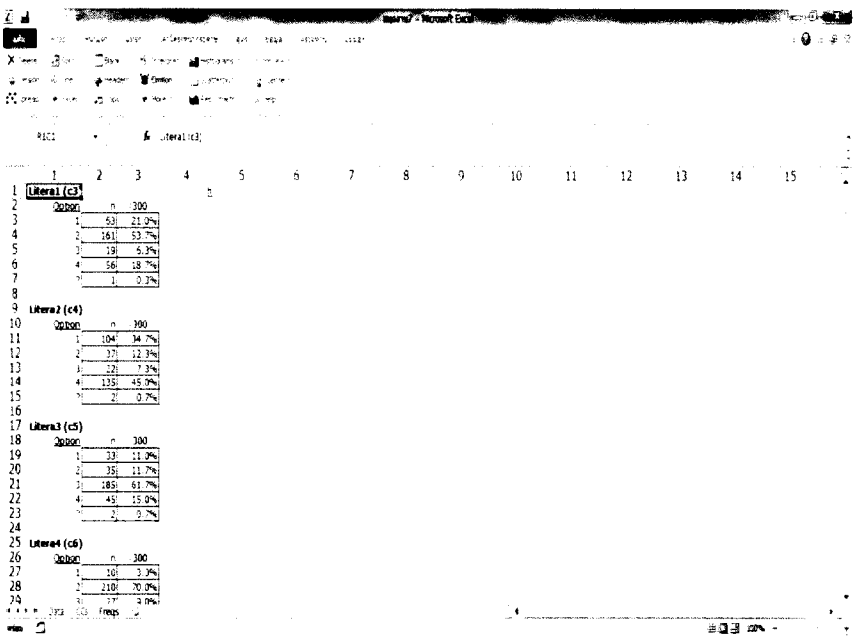
ภาพที่ 3-7 ตัวอย่างเลือกเมนู Interpret

3.8 แผ่นงานจะแสดงหน้าต่างขึ้นมา แล้วคลิก Ok ดังภาพที่ 3-8

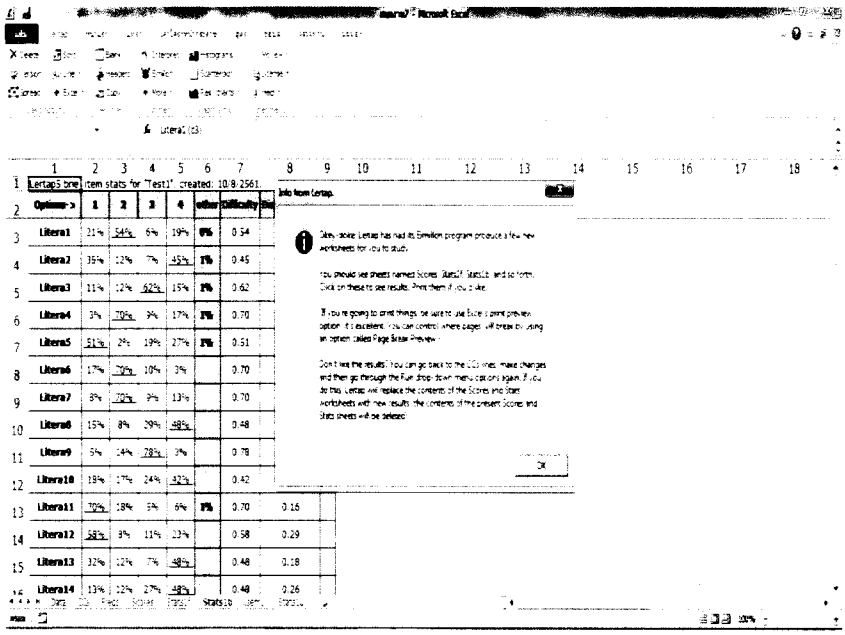


ภาพที่ 3-8 ตัวอย่างเลือกเมนู Ok

3.9 เมื่อคลิก Ok จะแสดงผล ดังภาพที่ 3-9

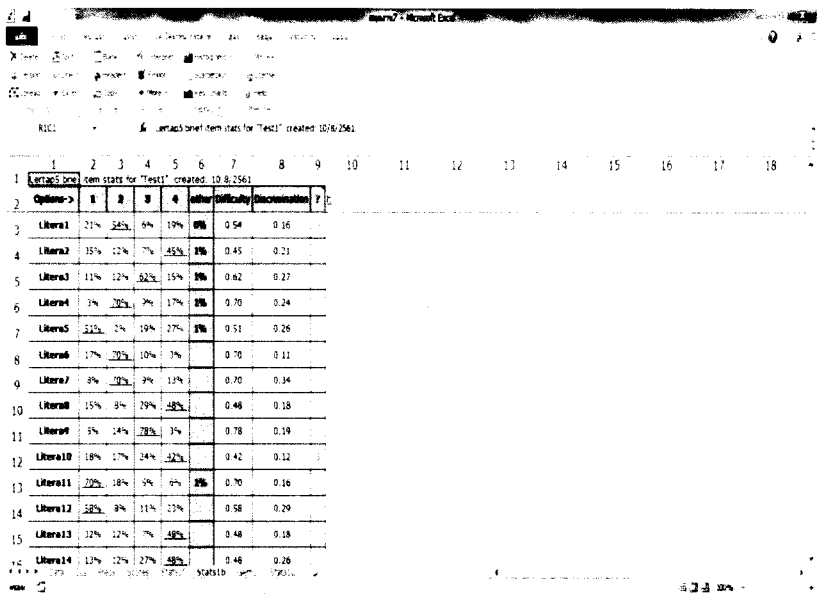


ภาพที่ 3-9 ตัวอย่างแสดงผลลัพธ์ Interpret



ภาพที่ 3-10 ตัวอย่างเลือกเมนู Elmillon และคลิก OK

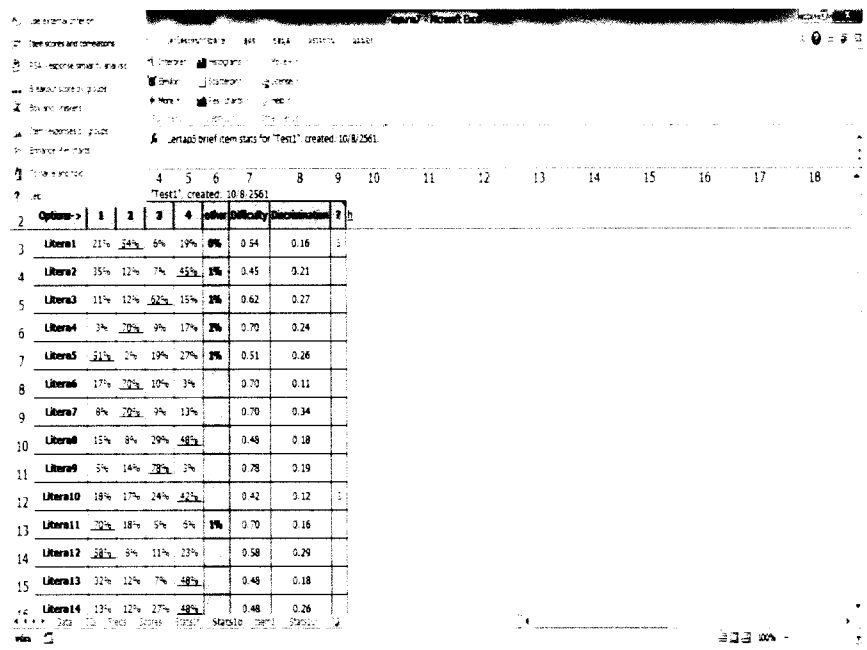
3.10 เมื่อคลิก OK จะแสดงผล แล้วคลิก OK อีกครั้ง ดังภาพที่ 3-11



ภาพที่ 3-11 ตัวอย่างแสดงภาพหน้าจอหลังจากเลือกเมนู OK

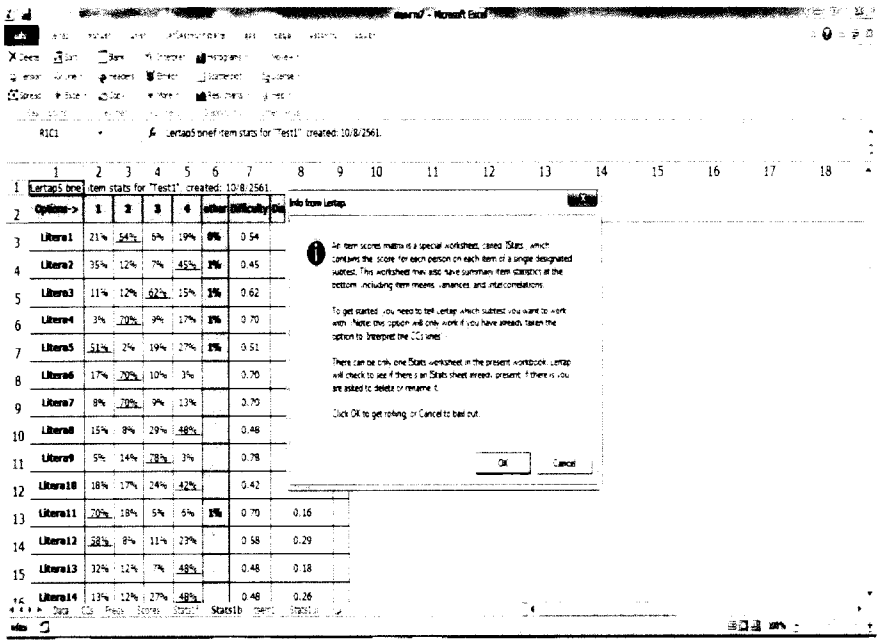
3.11 เลือก More ดังภาพที่ 3-11

3.12 เลือก Item scores and correlation ดังภาพที่ 3-12



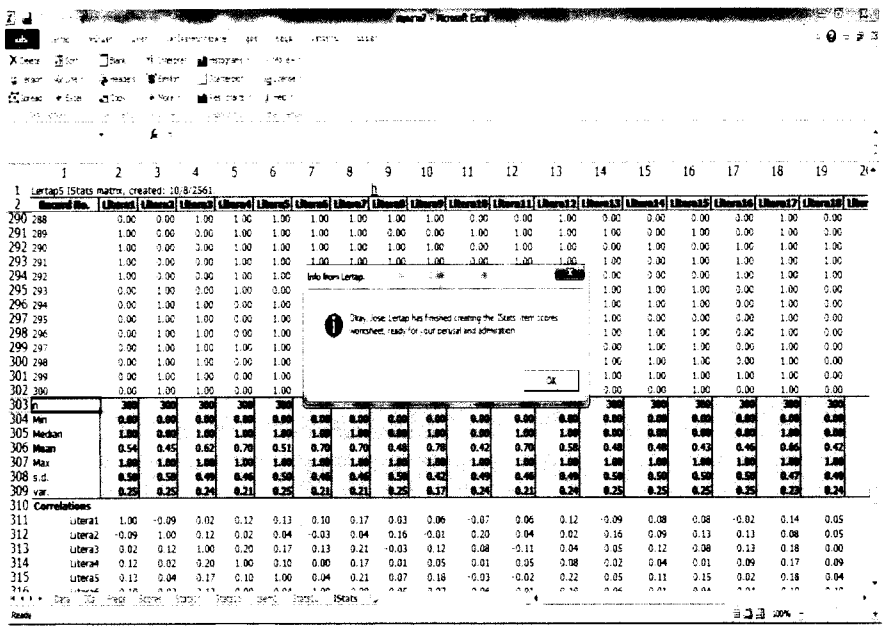
ภาพที่ 3-12 ตัวอย่างแสดงภาพหน้าจอหลังจากเลือกเมนู More เลือก Item Scores and Correlation

3.13 คลิก OK ดังภาพที่ 3-13



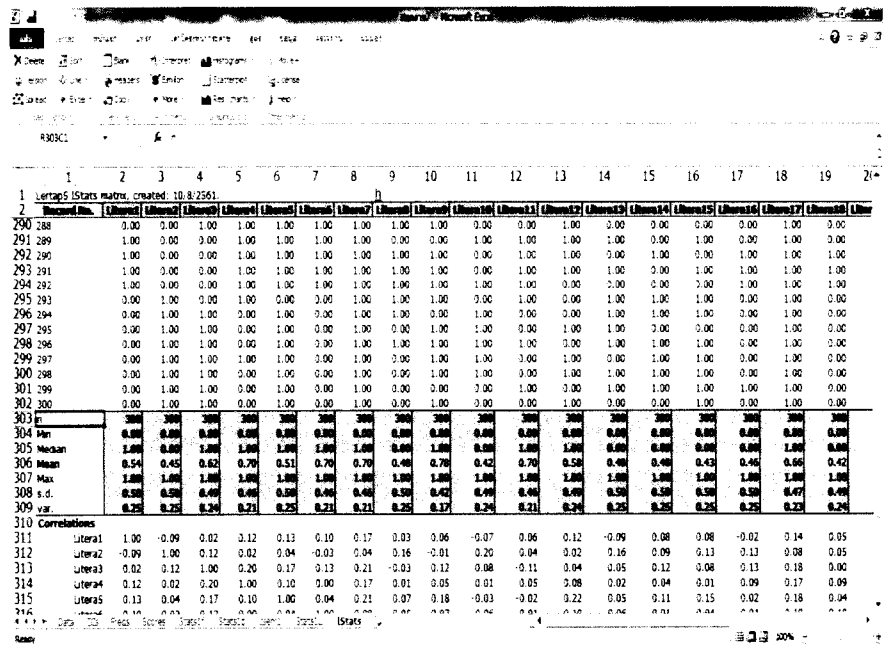
ภาพที่ 3-13 ตัวอย่างแสดงภาพหน้าจอหลังจากเลือกเมนู More เลือก Item Scores and Correlation และคลิก OK

3.14 คลิก OK ดังภาพที่ 3-14



ภาพที่ 3-14 ตัวอย่างแสดงภาพหน้าจอหลังจากคลิก OK และคลิก OK

3.15 คลิก OK ดังภาพที่ 3-15

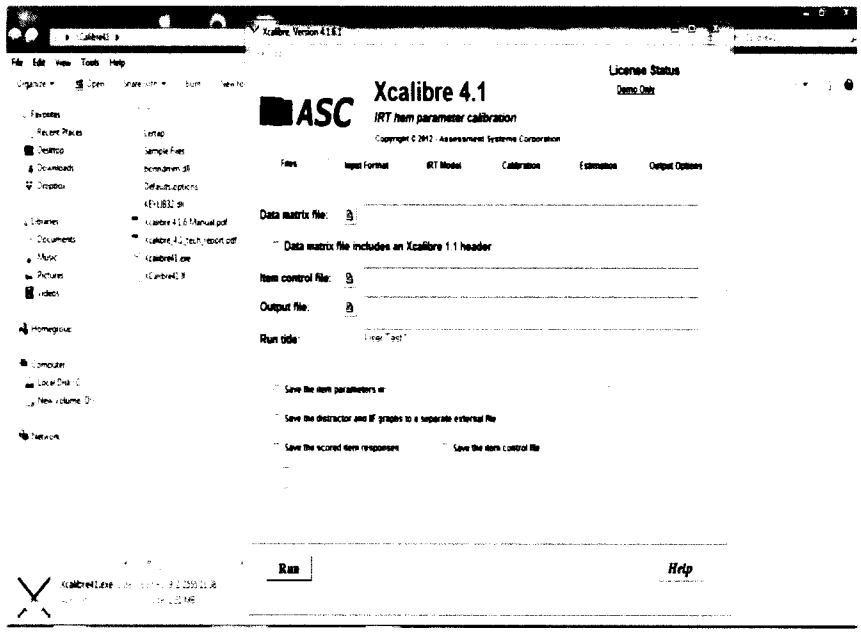


ภาพที่ 3-15 ตัวอย่างแสดงภาพหน้าจอ Output ของการ Run เมนู Item Scores and Correlation

3.16 เมื่อวิเคราะห์ Run แล้วจะได้ไฟล์ ดังนี้เป็น 2 ไฟล์ดังนี้

LrtpXcal_Data.txt LrtpXcal_ICF.txt และนำ 2 ไฟล์นี้เข้าโปรแกรม Xcalibre

4. การวิเคราะห์คุณภาพของข้อสอบ NT ตามทฤษฎีตอบสนองข้อสอบแบบ 3 พารามิเตอร์ โดยใช้โปรแกรม Xcalibre มีขั้นตอนดังนี้

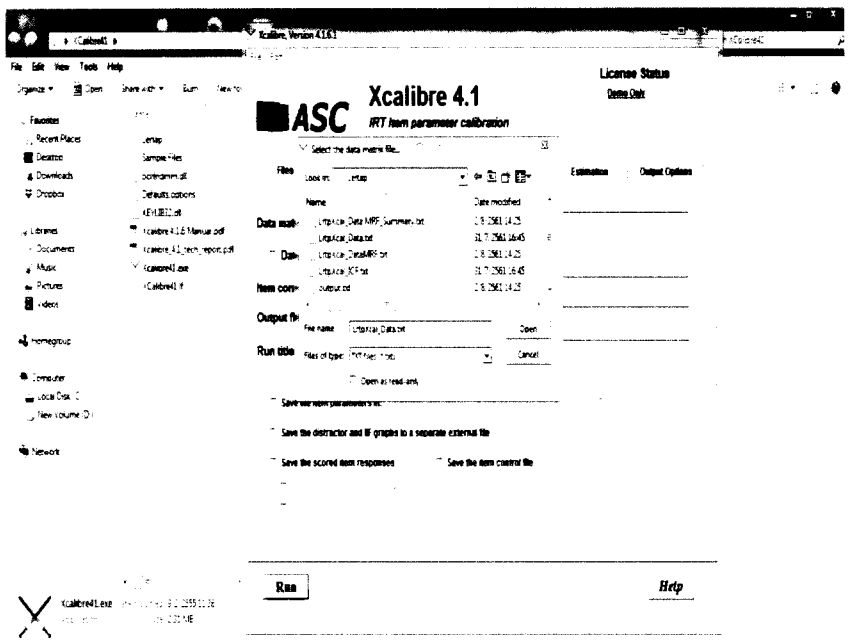


ภาพที่ 3-16 ตัวอย่างโปรแกรม Xcalibre

4.1 โปรแกรม Xcalibre ต้องมี 2 ไฟล์ ได้แก่ Data Matrix File และ Item Control File

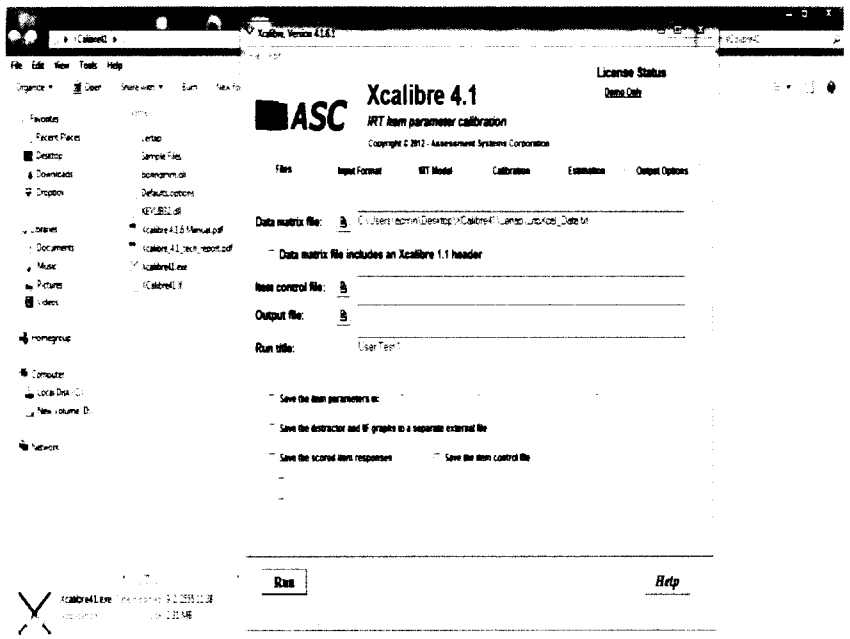
4.2 ระบุไฟล์ในแท็บ Files

4.3 แต่ละไฟล์ ซึ่งจะเปิดใช้หน้าต่างมาตรฐานเพื่อระบุ ชื่อของแต่ละไฟล์

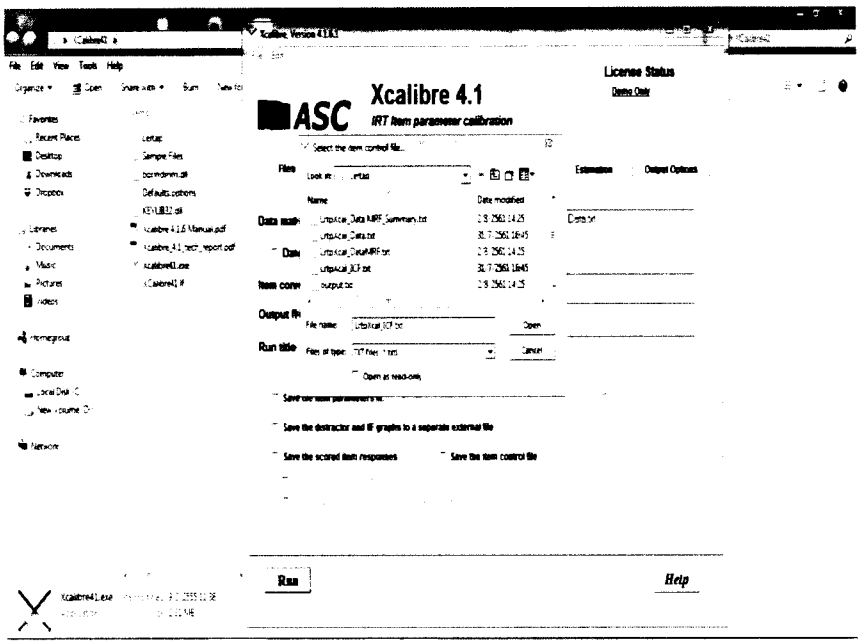


ภาพที่ 3-17 ตัวอย่างการเลือกไฟล์ Data.txt สำหรับวิเคราะห์โปรแกรม Xcalibre

4.4 ไฟล์ที่เรียกเข้า Data Matrix File คือ LrtpXcal_Data.txt ดังภาพที่ 3-18

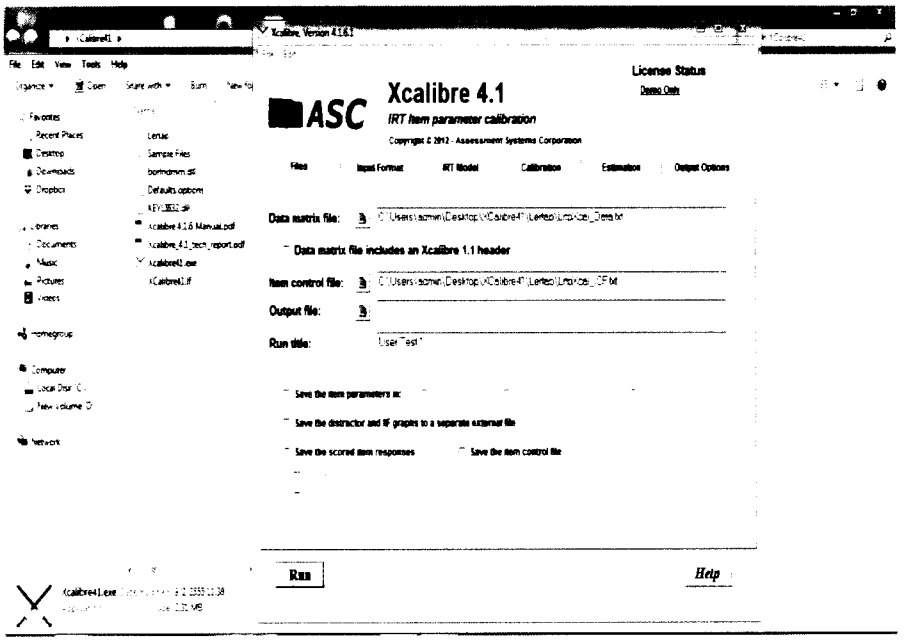


ภาพที่ 3-18 ตัวอย่างไฟล์ที่ต้องการเลือกวิเคราะห์โปรแกรม Xcalibre ใน Data Matrix File



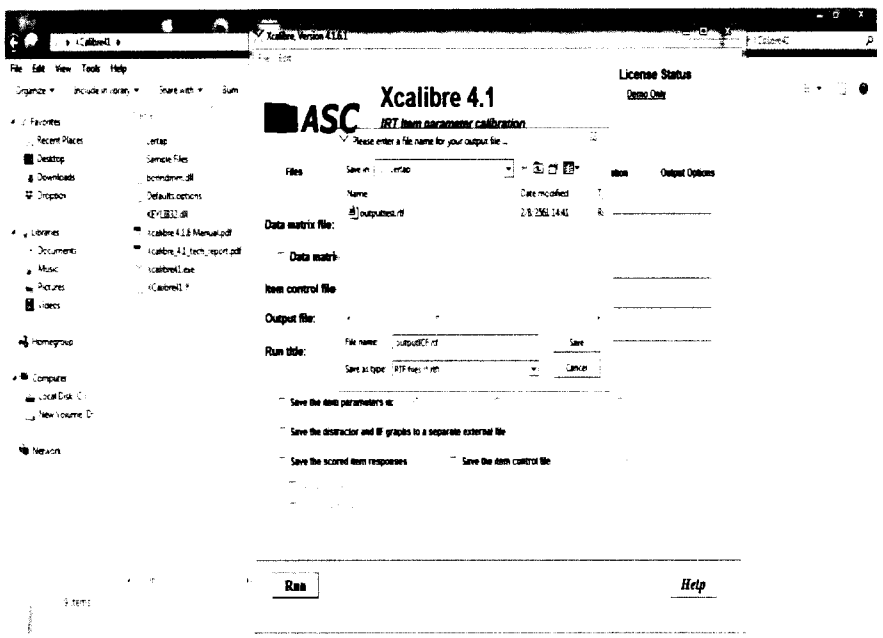
ภาพที่ 3-19 ตัวอย่างการเลือกไฟล์ ICF.txt วิเคราะห์โปรแกรม Xcalibre

4.5 ไฟล์ที่เรียกเข้า Item Control File คือ LrtpXcal_ICF.txt ดังภาพที่ 3-20



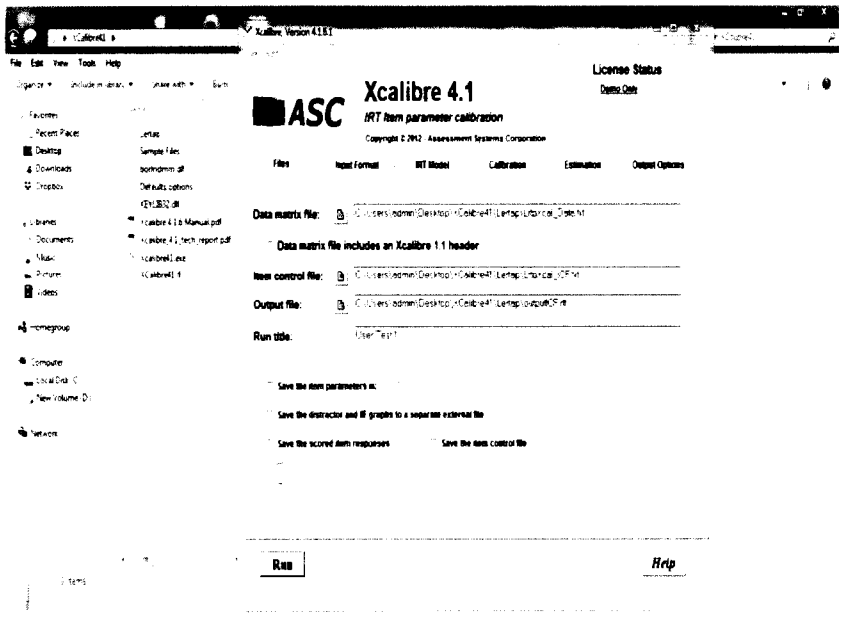
ภาพที่ 3-20 ตัวอย่างไฟล์ที่ต้องการเลือกวิเคราะห์โปรแกรม Xcalibre ใน Item Control File

4.6 ระบุชื่อไฟล์ที่ต้องการจะเก็บไว้ Output File คือ ดังภาพที่ 3-21

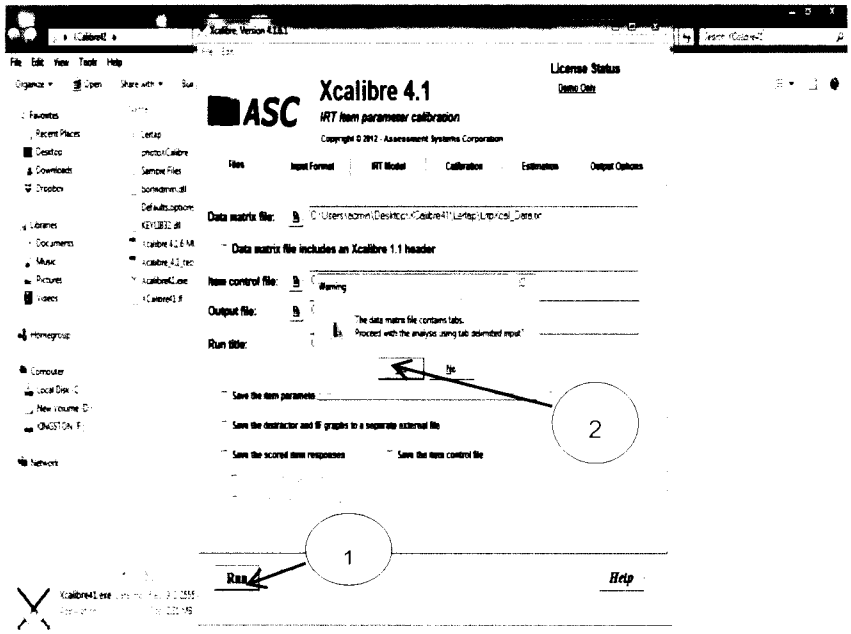


ภาพที่ 3-21 ตัวอย่างระบุไฟล์ที่ต้องการจะเก็บช่อง Output File

4.7 ไฟล์ที่ระบุต้องการจะเก็บไว้จะแสดงผลที่ช่อง Output File คือ ดังภาพที่ 3-22



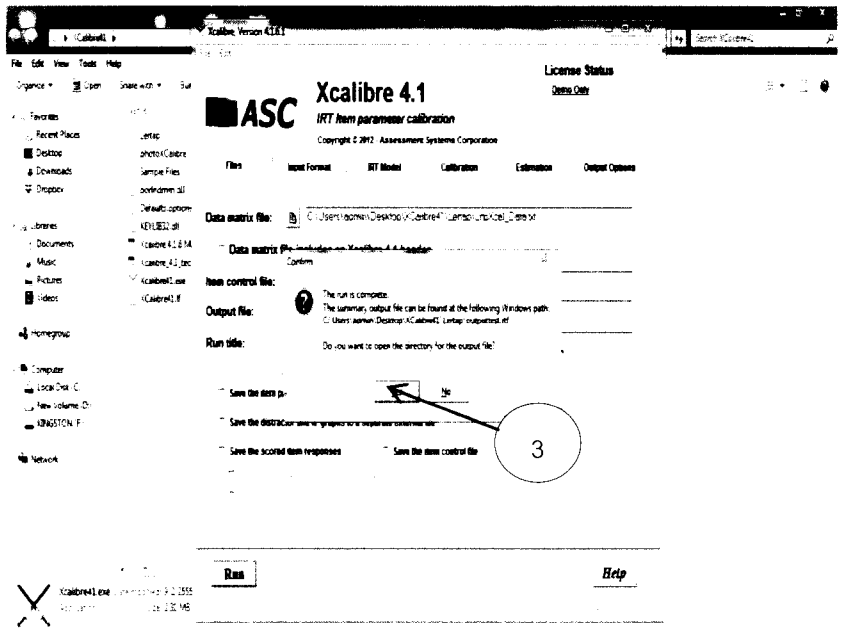
ภาพที่ 3-22 ตัวอย่างไฟล์ที่ต้องการเลือกวิเคราะห์โปรแกรม Xcalibre ในช่อง Output File



ภาพที่ 3-23 ตัวอย่างไฟล์ที่ต้องการเลือกวิเคราะห์โปรแกรม Xcalibre

4.8 คลิกที่ หมายเลข 1 เพื่อ Run ข้อมูลการวิเคราะห์ข้อมูล คือ ดังภาพที่ 3-23

4.9 คลิกที่ หมายเลข 2 เพื่อ Yes ดังภาพที่ 3-23



ภาพที่ 3-24 ตัวอย่างไฟล์ที่ต้องการเลือกวิเคราะห์โปรแกรม Xcalibre เลือกที่ Yes

4.10 คลิกที่ หมายเลข 3 เพื่อ Yes ดังภาพที่ 3-24

IRT Item Parameter Calibration Report

User Test 1

Report created on 3/8/2018

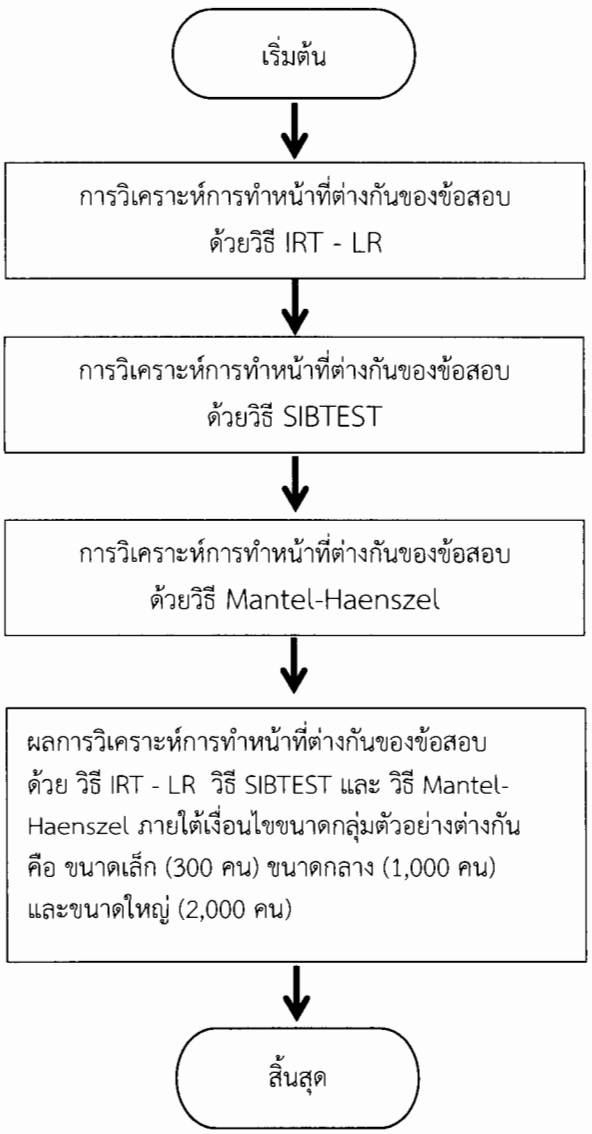
Xcalibre 4.1: IRT Item Parameter Estimation Software

ASC

ภาพที่ 3-25 แสดงตัวอย่างไฟล์ ผลการวิเคราะห์ด้วยโปรแกรม Xcalibre

ระยะที่ 2 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้วยวิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้วยวิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) ดังภาพที่ 3-26



ภาพที่ 3-26 ขั้นตอนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3

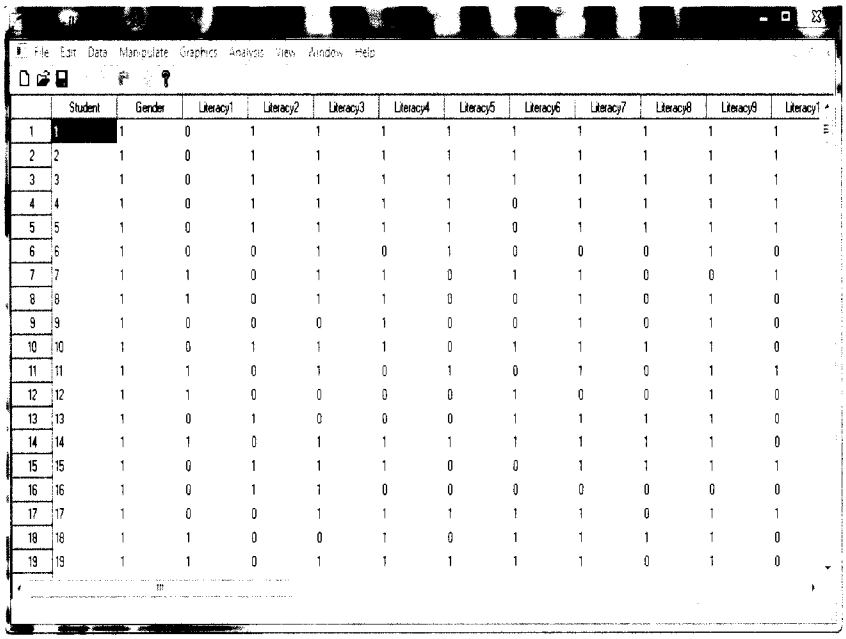
จากภาพที่ 3-26 แสดงขั้นตอนการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้วยวิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล ดังนี้

1. การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติด้วยวิธี IRT-LR โดยใช้โปรแกรม IRTPRO

การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธี IRT-LR โดยใช้โปรแกรม IRTPRO เป็นการวิเคราะห์ด้วยสถิติทางคณิตศาสตร์ โดยการเปรียบเทียบความแตกต่างของ ค่าพารามิเตอร์ในกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ผู้วิจัยได้เตรียมการวิเคราะห์ กำหนดค่าตัวแปร ด้านภาษา ด้านคำนวณ และด้านเหตุผล โดยกำหนดค่าตัวแปรเพศ ให้ตัวแปรเพศชายเป็น 1 และตัวแปรเพศหญิงเป็น 2 ดังนี้

ขั้นตอนที่ 1 เตรียมไฟล์ข้อมูลสำหรับวิเคราะห์

การวิเคราะห์ข้อมูลด้วยวิธี IRT–LR โดยใช้โปรแกรม IRTPRO เตรียมข้อมูล ในรูปแบบของไฟล์ .ssig วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยโปรแกรม IRTPRO ประกอบด้วยข้อมูล คือ (Studen) นักเรียนที่เข้าสอบ (Gender) เพศชาย เพศหญิง และผลการตอบแบบทดสอบของผู้สอบ (Response) ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) ดังภาพที่ 3-27



	Student	Gender	Literacy1	Literacy2	Literacy3	Literacy4	Literacy5	Literacy6	Literacy7	Literacy8	Literacy9	Literacy1
1	1	1	0	1	1	1	1	1	1	1	1	1
2	2	1	0	1	1	1	1	1	1	1	1	1
3	3	1	0	1	1	1	1	1	1	1	1	1
4	4	1	0	1	1	1	1	0	1	1	1	1
5	5	1	0	1	1	1	1	0	1	1	1	1
6	6	1	0	0	1	0	1	0	0	0	1	0
7	7	1	1	0	1	1	0	1	1	0	0	1
8	8	1	1	0	1	1	0	0	1	0	1	0
9	9	1	0	0	0	1	0	0	1	0	1	0
10	10	1	0	1	1	1	0	1	1	1	1	0
11	11	1	1	0	1	0	1	0	1	0	1	1
12	12	1	1	0	0	0	0	1	0	0	1	0
13	13	1	0	1	0	0	0	1	1	1	1	0
14	14	1	1	0	1	1	1	1	1	1	1	0
15	15	1	0	1	1	1	0	0	1	1	1	1
16	16	1	0	1	1	0	0	0	0	0	0	0
17	17	1	0	0	1	1	1	1	1	0	1	1
18	18	1	1	0	0	1	0	1	1	1	1	0
19	19	1	1	0	1	1	1	1	1	0	1	0

ภาพที่ 3-27 ตัวอย่างการจัดไฟล์ข้อมูลสำหรับวิเคราะห์ด้วยวิธี IRT-LR ในรูปแบบไฟล์ .ssig

ขั้นตอนที่ 2 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF)

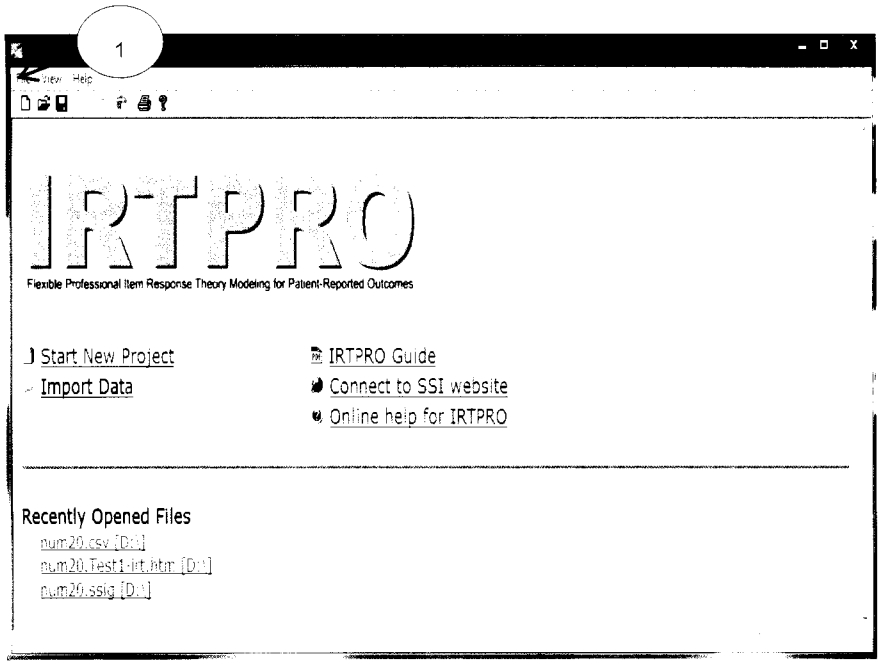
การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ด้วยวิธี IRT-LR โดยใช้โปรแกรม IRTPRO จะประเมินความสำคัญระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบในแบบทดสอบที่แตกต่างของแต่ละพารามิเตอร์ ดังนี้

1. เลือกวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยข้อสอบตามทฤษฎีการตอบทฤษฎีการตอบสนองข้อสอบแบบเอกมิติ (Unidimensional IRT)

2. กำหนดตัวแปร ในด้านภาษา และด้านเหตุผล ให้เพศหญิงเป็นกลุ่มอ้างอิง (Referent Group: R) และเพศชายเป็นกลุ่มเปรียบเทียบ (Focal Group: F) ส่วนด้านคำนวณ ให้เพศชายเป็นกลุ่มอ้างอิง (R) และเพศหญิงเป็นกลุ่มเปรียบเทียบ (F)

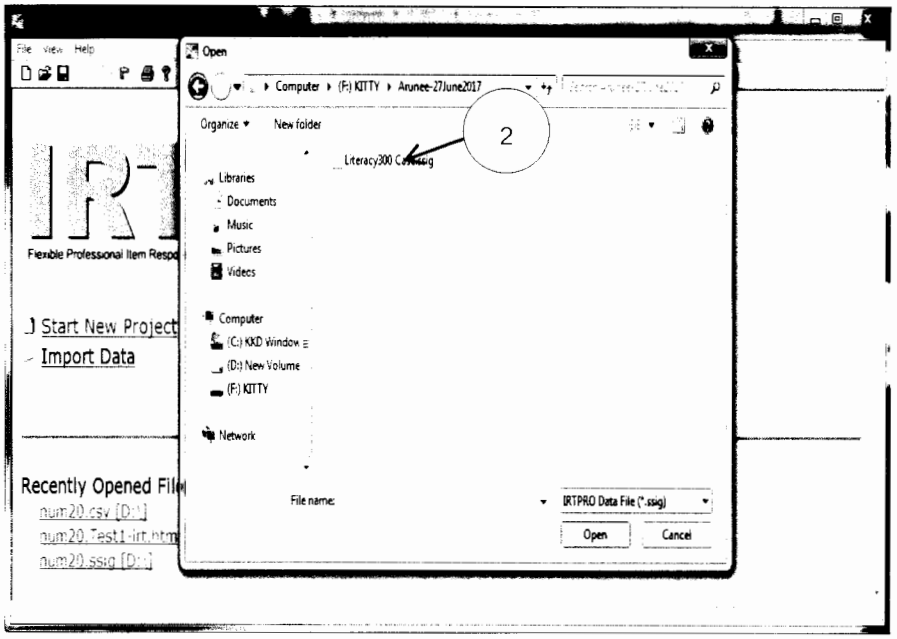
3. เลือกโมเดลสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เพื่อวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ ทำการวิเคราะห์แบบทดสอบทุกข้อจากนั้นพิจารณาว่าข้อใดมี DIF โดยดูจากค่า p-value ที่มีนัยสำคัญทางสถิติที่ระดับ .05 ดังภาพที่ 3-28

1. เปิดโปรแกรม IRTPRO ดังภาพที่ 3-28



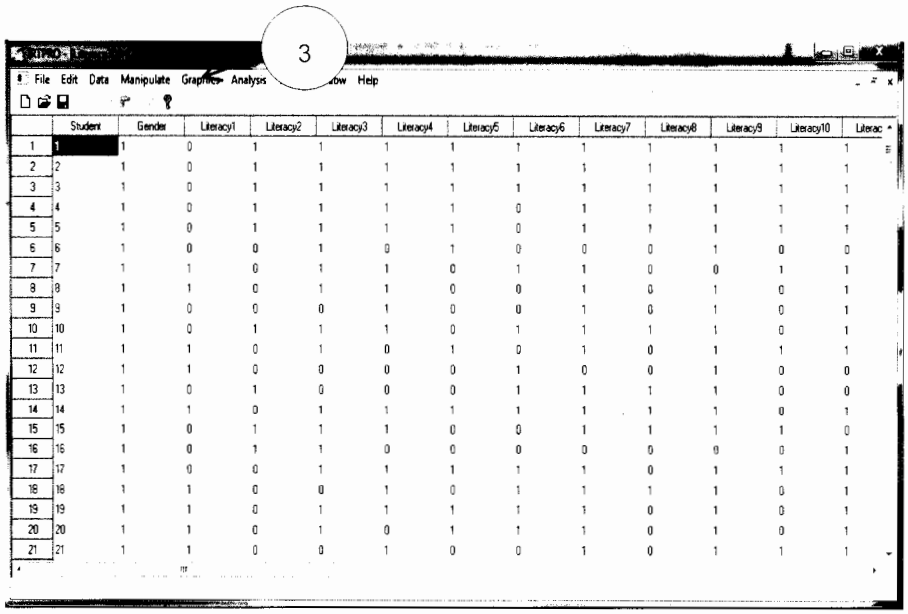
ภาพที่ 3-28 ตัวอย่างการเรียกไฟล์ข้อมูลสำหรับวิเคราะห์ด้วยวิธี IRT-LR ในรูปแบบไฟล์ .ssig

2. เลือกที่หมายเลข 1 คลิกที่ File แล้วเลือก Open ดังภาพที่ 3-29



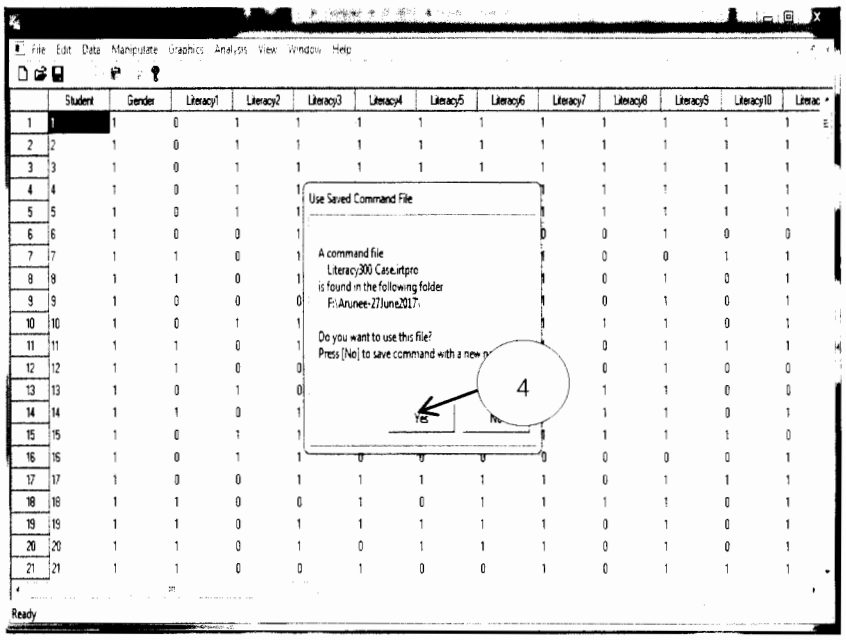
ภาพที่ 3-29 ตัวอย่างการเปิดไฟล์ข้อมูลสำหรับวิเคราะห์ด้วยวิธี IRT-LR ในรูปแบบไฟล์ .ssig

3. เลือกที่หมายเลข 3 เพื่อเลือก Analysis ดังภาพที่ 3-30



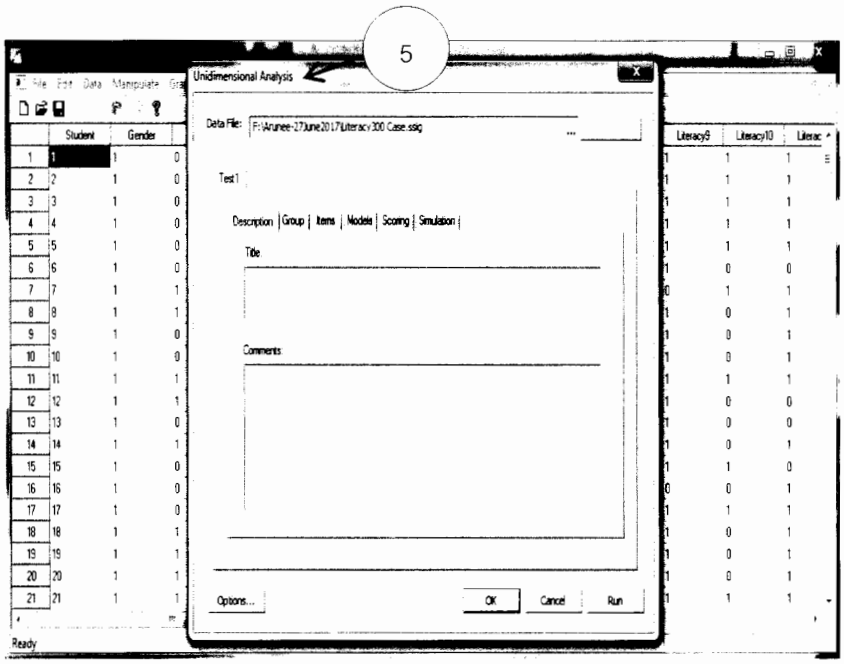
ภาพที่ 3-30 ตัวอย่างการเลือก Analysis การวิเคราะห์ข้อมูล

4. เลือกที่หมายเลข 4 เพื่อเลือก Yes ดังภาพที่ 3-31



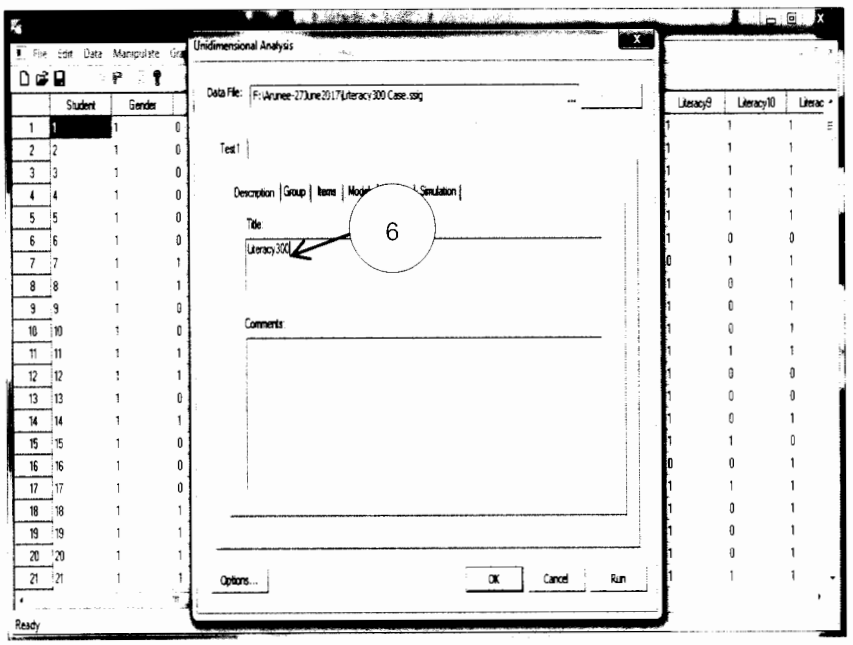
ภาพที่ 3-31 ตัวอย่างการเลือก Yes เพื่อการวิเคราะห์ข้อมูล

5. เลือกที่หมายเลข 5 เพื่อเลือก Unidimensional Analysis ดังภาพที่ 3-32



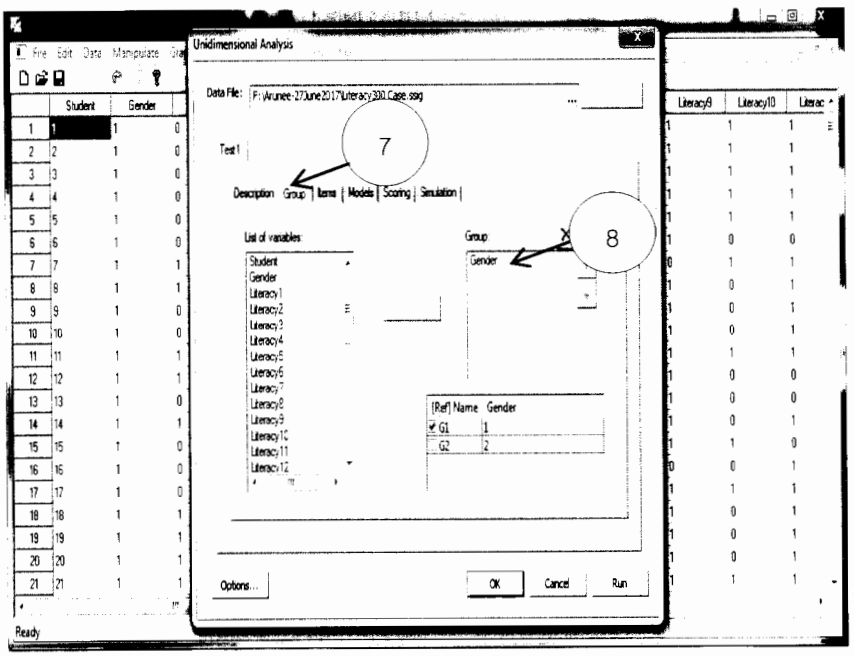
ภาพที่ 3-32 ตัวอย่างการเลือก Unidimensional Analysis เพื่อการวิเคราะห์ข้อมูล

6. เลือกที่หมายเลข 6 เพื่อเลือก ใส่ชื่อเรื่องไฟล์ ดังภาพที่ 3-33



ภาพที่ 3-33 ตัวอย่างการเลือกใส่ชื่อไฟล์เพื่อการวิเคราะห์ข้อมูล

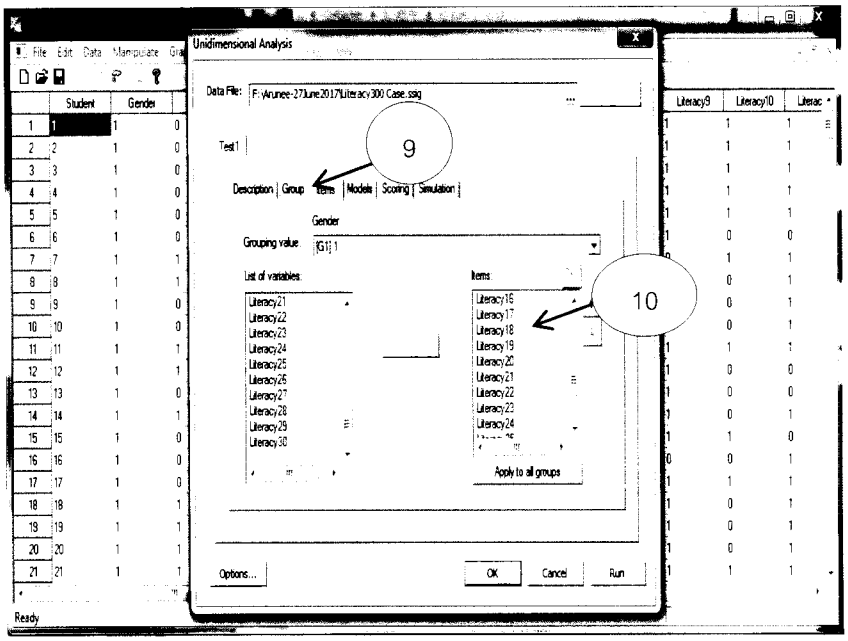
7. เลือกที่หมายเลข 7 เพื่อเลือก Group ดังภาพที่ 3-34
8. เลือกที่หมายเลข 8 เพื่อเลือก Gender ดังภาพที่ 3-34



ภาพที่ 3-34 ตัวอย่างการเลือก Group และเลือก Gender เพื่อการวิเคราะห์ข้อมูล

9. เลือกที่หมายเลข 9 เพื่อเลือก Item ดังภาพที่ 3-35

10. เลือกที่หมายเลข 10 เพื่อเลือก จำนวนข้อสอบทั้งหมด ดังภาพที่ 3-35



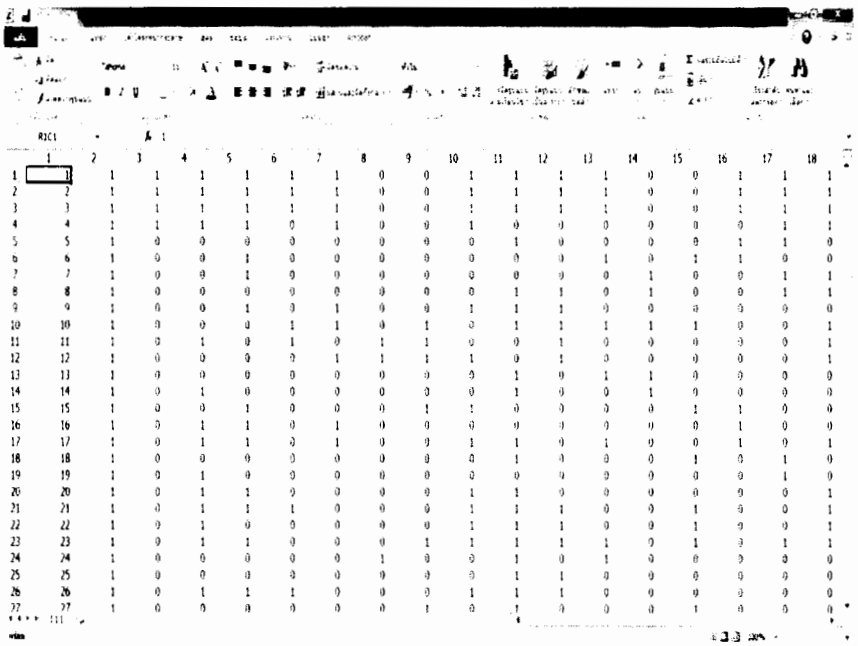
ภาพที่ 3-35 ตัวอย่างการเลือก Item และเลือกจำนวนข้อสอบทั้งหมดเพื่อการวิเคราะห์ข้อมูล

2. การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ด้วยวิธี SIBTEST

การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธี SIBTEST โดยใช้โปรแกรม SIBTEST ในการเปรียบเทียบความแตกต่างของ ค่าพารามิเตอร์ในกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ผู้วิจัยได้เตรียมการวิเคราะห์ กำหนดค่าตัวแปร ด้านภาษา ด้านคำนวณ และด้านเหตุผล โดยกำหนดค่าตัวแปรเพศ ให้ตัวแปรเพศชายเป็น 1 และตัวแปรเพศหญิงเป็น 2 ดังนี้

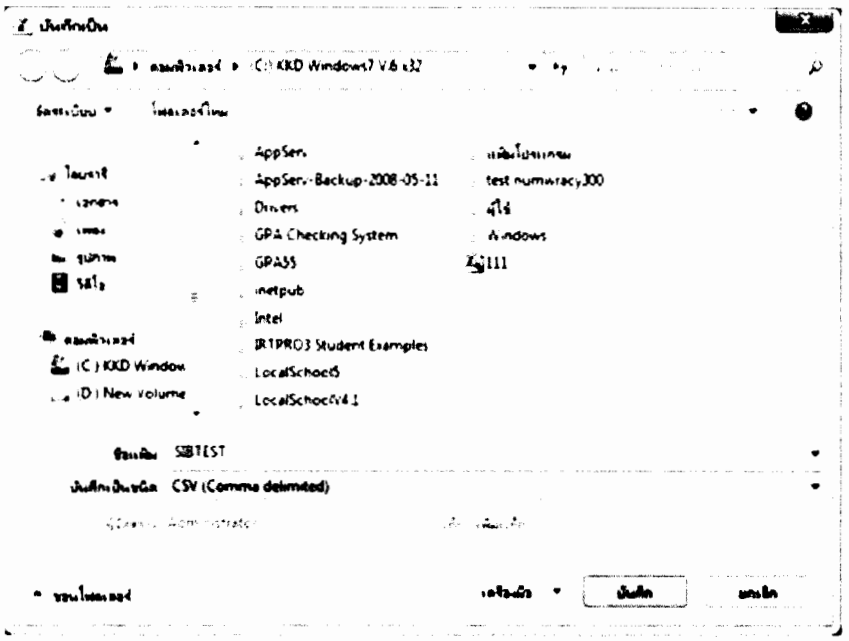
ขั้นตอนที่ 1 เตรียมไฟล์ข้อมูลสำหรับวิเคราะห์

การวิเคราะห์ข้อมูลด้วยวิธี SIBTEST ผู้วิจัยได้เตรียมข้อมูล วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยโปรแกรม SIBTEST ประกอบด้วยข้อมูล คือ นักเรียน (Studen) ที่เข้าสอบ (Gender) เพศชาย เพศหญิง และผลการตอบแบบทดสอบของผู้สอบ (Response) ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) ดังภาพที่ 3-36



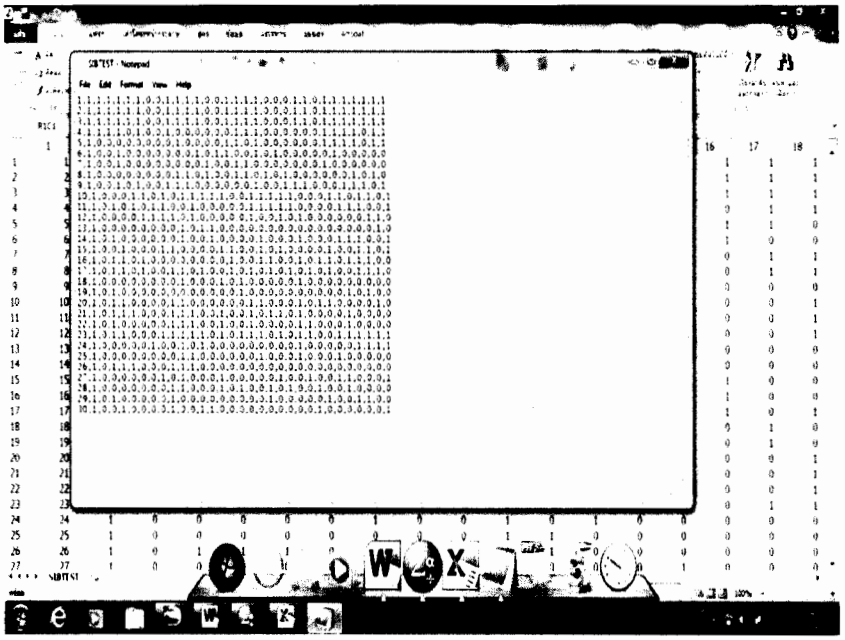
ภาพที่ 3-36 ตัวอย่างการจัดไฟล์ข้อมูลสำหรับวิเคราะห์ด้วยวิธี SIBTEST ในรูปแบบไฟล์ .dat

1. เปิดโปรแกรม Excel



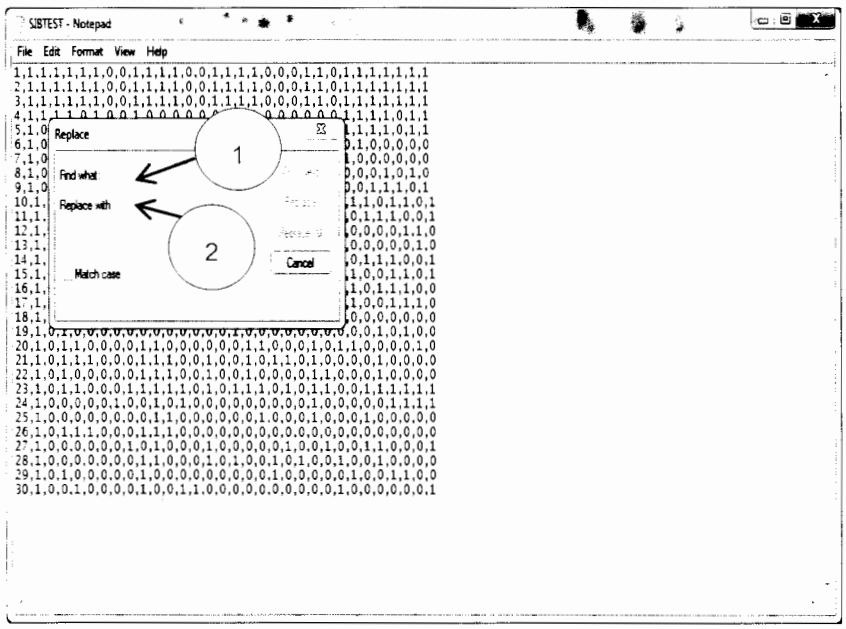
ภาพที่ 3-37 ตัวอย่างการบันทึกไฟล์ข้อมูลสำหรับวิเคราะห์ด้วยวิธี SIBTEST ในรูปแบบไฟล์ .dat

2. เลือก แฟ้ม แล้วเลือก บันทึกเป็น
3. ใส่ชื่อแฟ้ม
4. เลือก CSV (Comma delimited)
5. เลือกบันทึก เลือก Ok เลือก Yes



ภาพที่ 3-38 ตัวอย่างการจัดไฟล์ข้อมูลสำหรับวิเคราะห์ด้วยวิธี SIBTEST ในรูปแบบไฟล์ .dat ในรูปแบบ NotePad

6. เปิดโปรแกรม NotePad



ภาพที่ 3-39 ตัวอย่างการจัดไฟล์ข้อมูลที่ลบจุลภาค (,) ออก และชิดของข้อมูลสำหรับวิเคราะห์ด้วยวิธี SIBTEST ในรูปแบบไฟล์ .dat

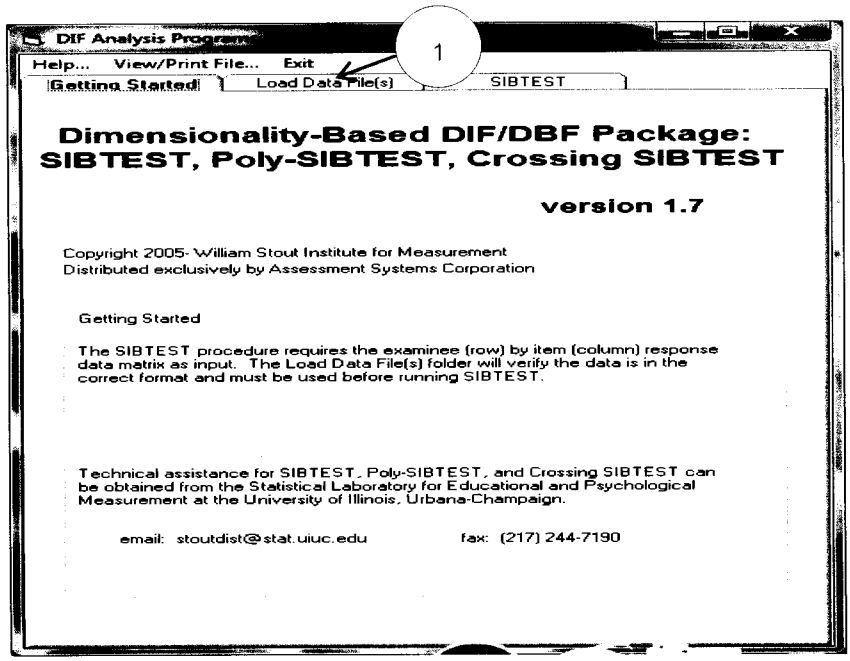
7. เลือกที่ Edit เลือก หมายเลข 1 ให้ใส่เครื่องหมาย (,) ช่องที่ 2 ไม่ต้องใส่อะไร แล้วไปเลือกที่ Replace เพื่อเอาจุลภาค (,) ออก และให้ข้อมูลชิดกัน

8. เลือกที่ File เลือก Save As

9. เลือกที่ All File แล้วไปตั้งชื่อ File ตามด้วย .dat แล้วบันทึก

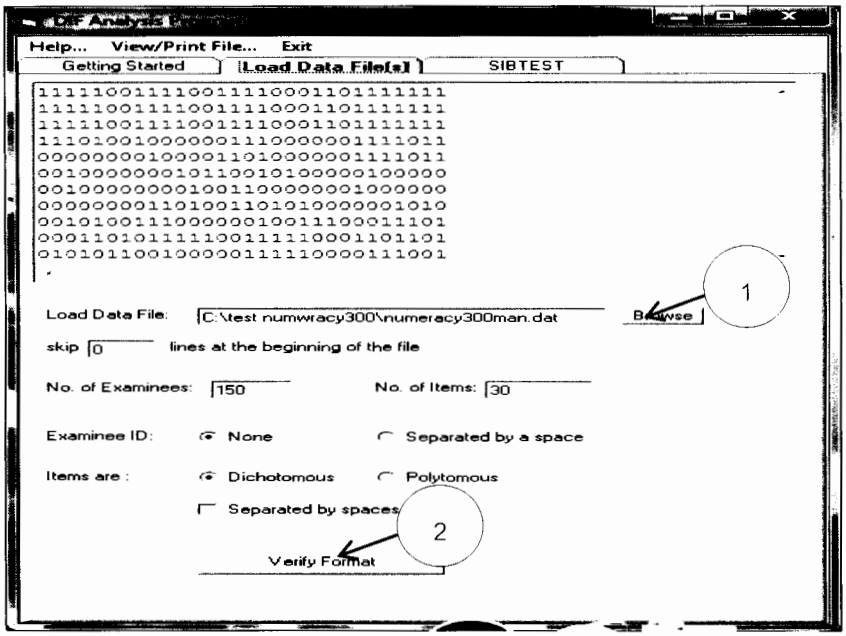
ขั้นตอนที่ 2 การวิเคราะห์การทำหน้าที่กันของข้อสอบด้วยวิธี SIBTEST มีดังนี้

1. เปิดโปรแกรม SIBTEST



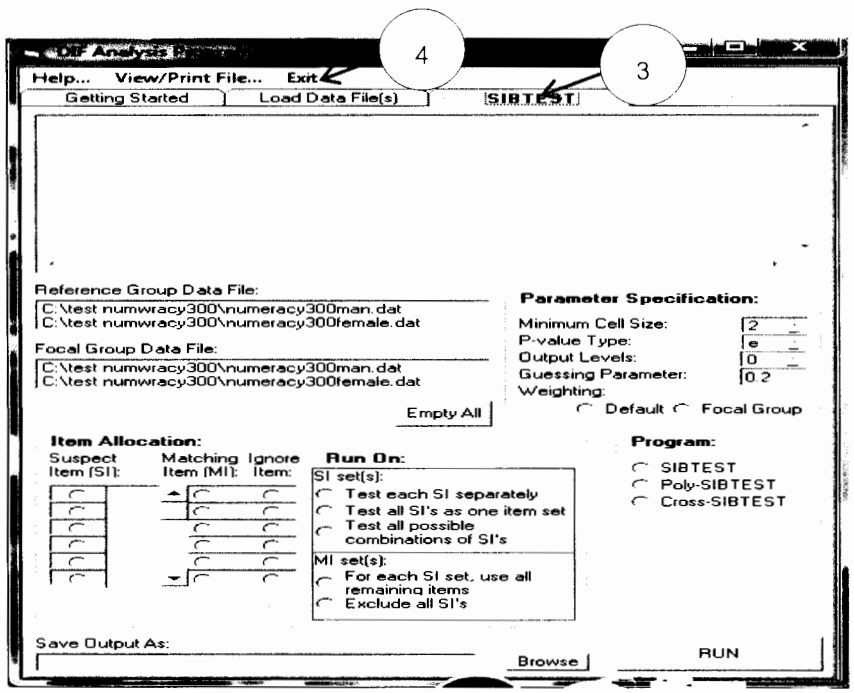
ภาพที่ 3-40 ตัวอย่างการวิเคราะห์ด้วยวิธี SIBTEST ในรูปแบบไฟล์ .dat เรียกไฟล์ข้อมูลวิเคราะห์

2. เลือกที่ Load Data File หมายเลข 1



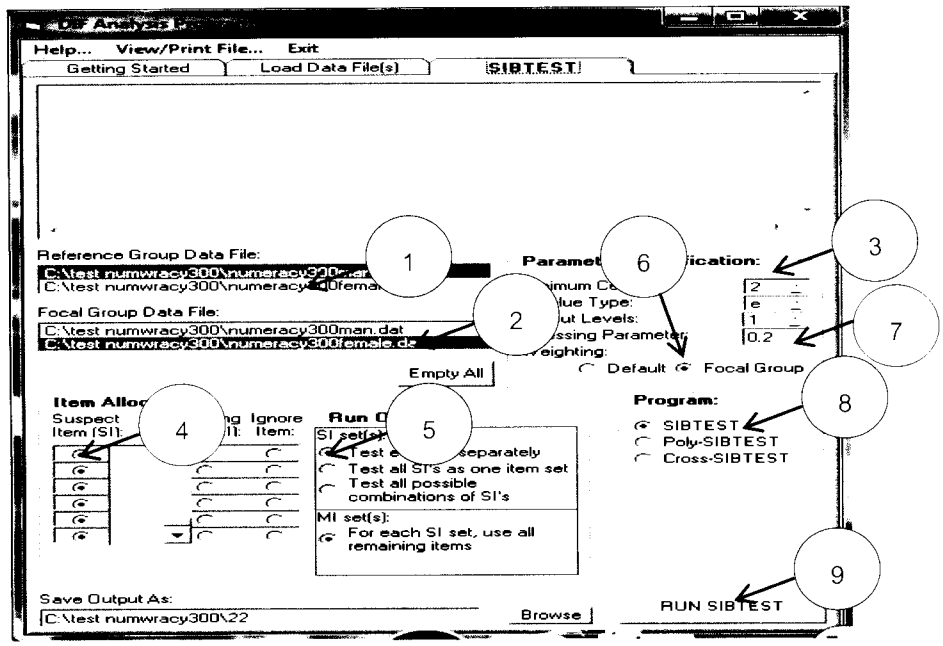
ภาพที่ 3-41 ตัวอย่างการวิเคราะห์ด้วยวิธี SIBTEST ในรูปแบบไฟล์ .dat เมื่อไฟล์ข้อมูลเข้าในโปรแกรม

3. เลือกที่ Browse หมายเลข 1
4. เลือกไฟล์ที่ Save ไว้ และตรวจเช็คข้อมูลให้เรียบร้อย
5. เลือกที่ Verify Fomat หมายเลข 2



ภาพที่ 3-42 ตัวอย่างการวิเคราะห์ด้วยวิธี SIBTEST ในรูปแบบไฟล์ .dat

6. เลือก SIBTEST เพื่อเตรียมไฟล์อีกไฟล์เรียกเข้ามา หมายเลข 3
7. เลือก Load Data File หมายเลข 4 และทำตาม ข้อ 3-5



ภาพที่ 3-43 ตัวอย่างการเลือกข้อมูลวิเคราะห์ด้วยวิธี SIBTEST ในรูปแบบไฟล์ .dat

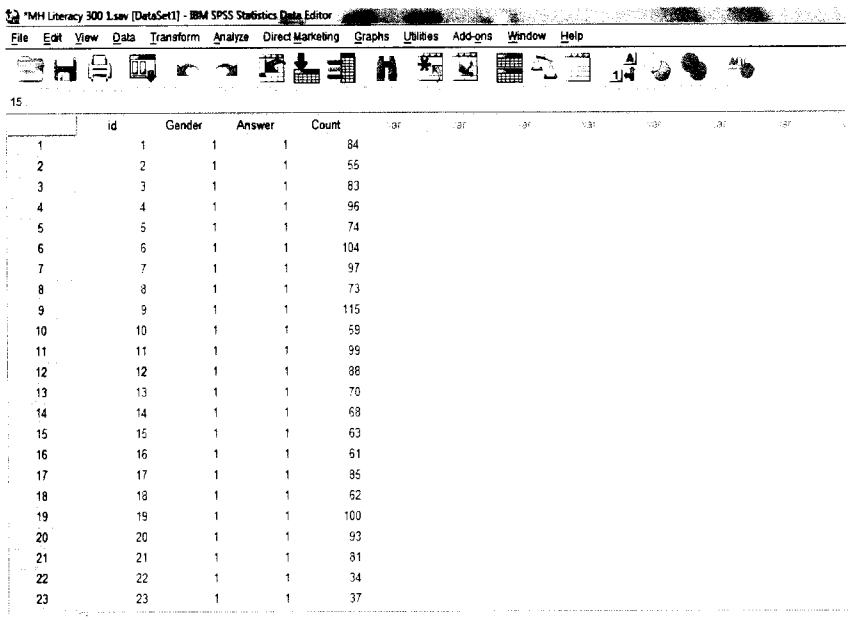
8. เลือก Reference Group Data File หมายเลข 1 และหมายเลข 2
9. เลือก ดูตรง หมายเลข 3 เพื่อตรวจเช็คความเรียบร้อย
10. เลือกตรง หมายเลข 4 เพื่อคลิกจำนวนข้อสอบให้ครบจำนวนข้อ
11. เลือกตรง หมายเลข 5 เพื่อ Test จำนวนข้อสอบ
12. เลือกตรง หมายเลข 6 เพื่อ เลือก Group
13. เลือกตรง หมายเลข 7 เพื่อ Save File Output
14. เลือกตรง หมายเลข 8 เพื่อเลือกโปรแกรม SIBTEST
15. เลือกตรง หมายเลข 9 เพื่อ Run โปรแกรม
16. พิจารณาว่าข้อใดมี DIF โดยดูจากค่า *p*-value ที่มีนัยสำคัญทางสถิติที่ระดับ .05

3. การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติด้วยวิธี Mantel-Haenszel

การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธี Mantel-Haenszel โดยใช้โปรแกรม SPSS เป็นการทดสอบอัตราส่วนเปรียบเทียบด้วยไค-สแควร์ (Chi-square Statistic) ในการเปรียบเทียบความแตกต่างของ ค่าพารามิเตอร์ในกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ผู้วิจัยได้เตรียมการวิเคราะห์ กำหนดค่าตัวแปร ด้านภาษา ด้านคำนวณ และด้านเหตุผล โดยกำหนดค่าตัวแปรเพศ ให้ตัวแปรเพศชายเป็น 1 และตัวแปรเพศหญิงเป็น 2 ดังนี้

ขั้นตอนที่ 1 เตรียมไฟล์ข้อมูลสำหรับวิเคราะห์

การวิเคราะห์ข้อมูลด้วยวิธี Mantel-Haenszel ผู้วิจัยได้เตรียมข้อมูลผลการตอบแบบทดสอบ ระหว่างกลุ่มอ้างอิง (R) และกลุ่มเปรียบเทียบ (F) มาจัดลงในตารางการณ์จรแบบ 2×2 (กลุ่มผู้สอบ 2 กลุ่ม × ผลการตอบ 2 แบบ) ประกอบด้วยข้อมูล คือ (ID) จำนวนข้อสอบ (Gender) เพศชาย เพศหญิง (Answer) คำตอบถูกเป็น 1 คำตอบผิดเป็น 0 และ (Count) จำนวนผลการตอบแบบทดสอบของผู้สอบ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) ดังภาพที่ 3-44



*MH Literacy 300 1.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

15

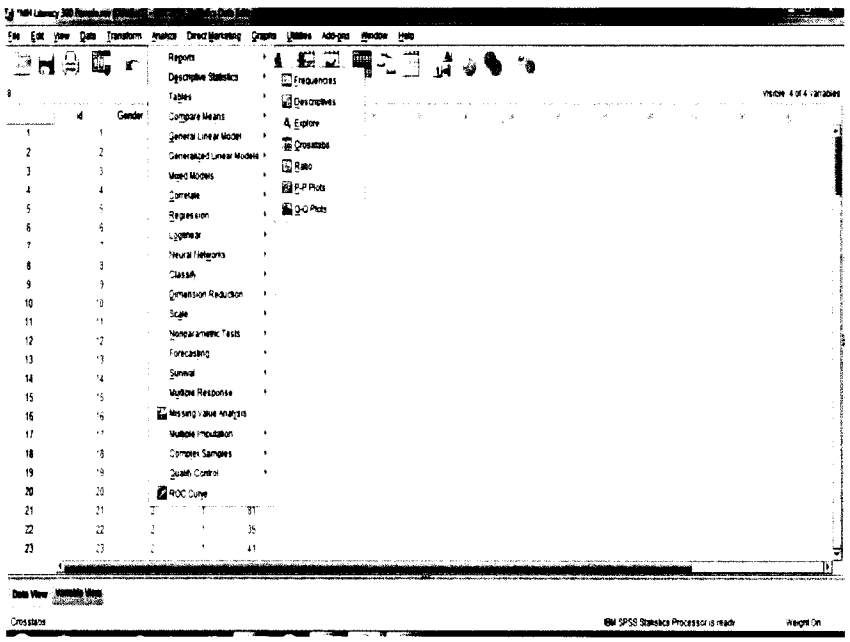
	id	Gender	Answer	Count
1	1	1	1	84
2	2	1	1	55
3	3	1	1	83
4	4	1	1	96
5	5	1	1	74
6	6	1	1	104
7	7	1	1	97
8	8	1	1	73
9	9	1	1	115
10	10	1	1	59
11	11	1	1	99
12	12	1	1	88
13	13	1	1	70
14	14	1	1	68
15	15	1	1	63
16	16	1	1	61
17	17	1	1	85
18	18	1	1	62
19	19	1	1	100
20	20	1	1	93
21	21	1	1	81
22	22	1	1	34
23	23	1	1	37

ภาพที่ 3-44 ตัวอย่างการจัดไฟล์ข้อมูลสำหรับวิเคราะห์ด้วยวิธี Mantel-Haenszel ตารางการณ์จรแบบ 2×2

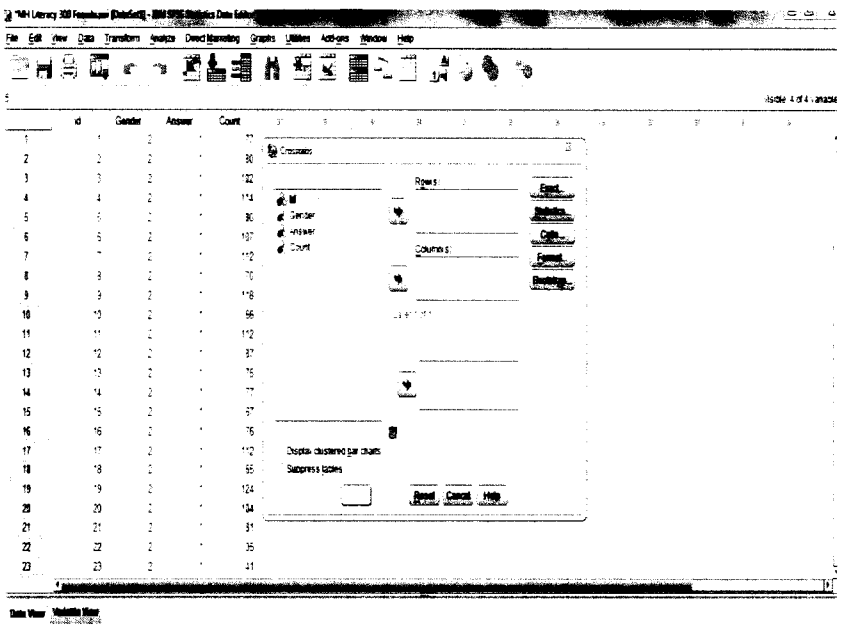
ขั้นตอนที่ 2 การตรวจสอบการทำหน้าที่กันของข้อสอบ

1. กำหนดตัวแปร ในด้านภาษา และด้านเหตุผล ให้เพศหญิงเป็นกลุ่มอ้างอิง (Referent Group: R) และเพศชายเป็นกลุ่มเปรียบเทียบ (Focal Group: F) ส่วนด้านคำนวณ ให้เพศชายเป็นกลุ่มอ้างอิง (R) และเพศหญิงเป็นกลุ่มเปรียบเทียบ (F)

2. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ด้วยวิธี Mantel-Haenszel เป็นการเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ในความแตกต่างของแต่ละพารามิเตอร์ โดยโปรแกรม SPSS คำสั่ง Analyze ➔ Descriptive Sttatistics ➔ Crosstabs ดังภาพที่ 3-43 ภาพที่ 3-44 และภาพที่ 3-45

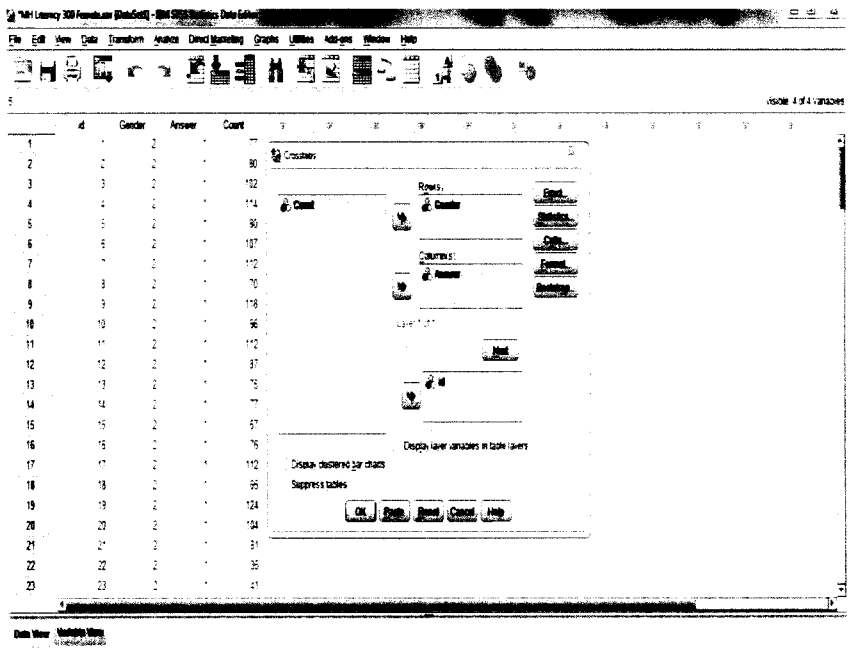


ภาพที่ 3-45 ตัวอย่างคำสั่งที่ Analysis การวิเคราะห์ด้วยวิธี Mantel-Haenszel โดยโปรแกรม SPSS



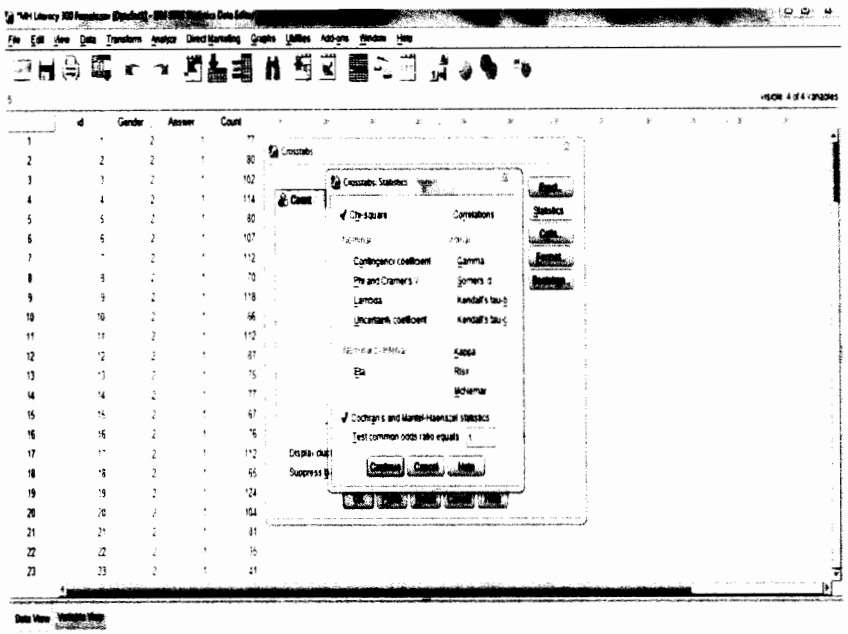
ภาพที่ 3-46 ตัวอย่างคำสั่งที่ Crosstabs การวิเคราะห์ด้วยวิธี Mantel-Haenszel โดยโปรแกรม SPSS

3. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ด้วยวิธี Mantel-Haenszel เป็นการเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ในความแตกต่างของแต่ละพารามิเตอร์ โดยโปรแกรม SPSS โดยเลือก Gender ไปที่ช่อง Row(s) ➔ Answer เลือกไปที่ช่อง Column(s) ➔ ID เลือกไปที่ช่อง Layer 1 of 1 ดังภาพที่ 3-47



ภาพที่ 3-47 ตัวอย่างคำสั่งเรียกไฟล์ที่ต้องการจะวิเคราะห์ด้วยวิธี Mantel-Haenszel โดยโปรแกรม SPSS

4. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ด้วยวิธี Mantel-Haenszel เป็นการเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ ในความแตกต่างของแต่ละพารามิเตอร์ โดยโปรแกรม SPSS โดยเลือก Statistics ➔ เลือกที่ Chi-square ➔ Cochram's and Mantel-Haenszel Statistics ดังภาพที่ 3-48



ภาพที่ 3-48 ตัวอย่างคำสั่งเลือก Chi-Square การวิเคราะห์ด้วยวิธี Mantel-Haenszel โดยโปรแกรม SPSS

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Sample * Response	270	100.0%	0	0.0%	270	100.0%
Subject * Response	270	100.0%	0	0.0%	270	100.0%
Methods * Response	270	100.0%	0	0.0%	270	100.0%

Sample * Response

Crosstab

Count

		Response		Total
		NO-DIF	DIF	
Sample	300 persons	242	28	270
Total		242	28	270

Chi-Square Tests

	Value
Pearson Chi-Square	.a
N of Valid Cases	270

a. No statistics are computed because Sample is a constant.

Subject * Response

Crosstab

Count

		Response		Total
		NO-DIF	DIF	
Subject	Literacy	80	10	90
	Numeracy	82	8	90
	Reasoning	80	10	90
Total		242	28	270

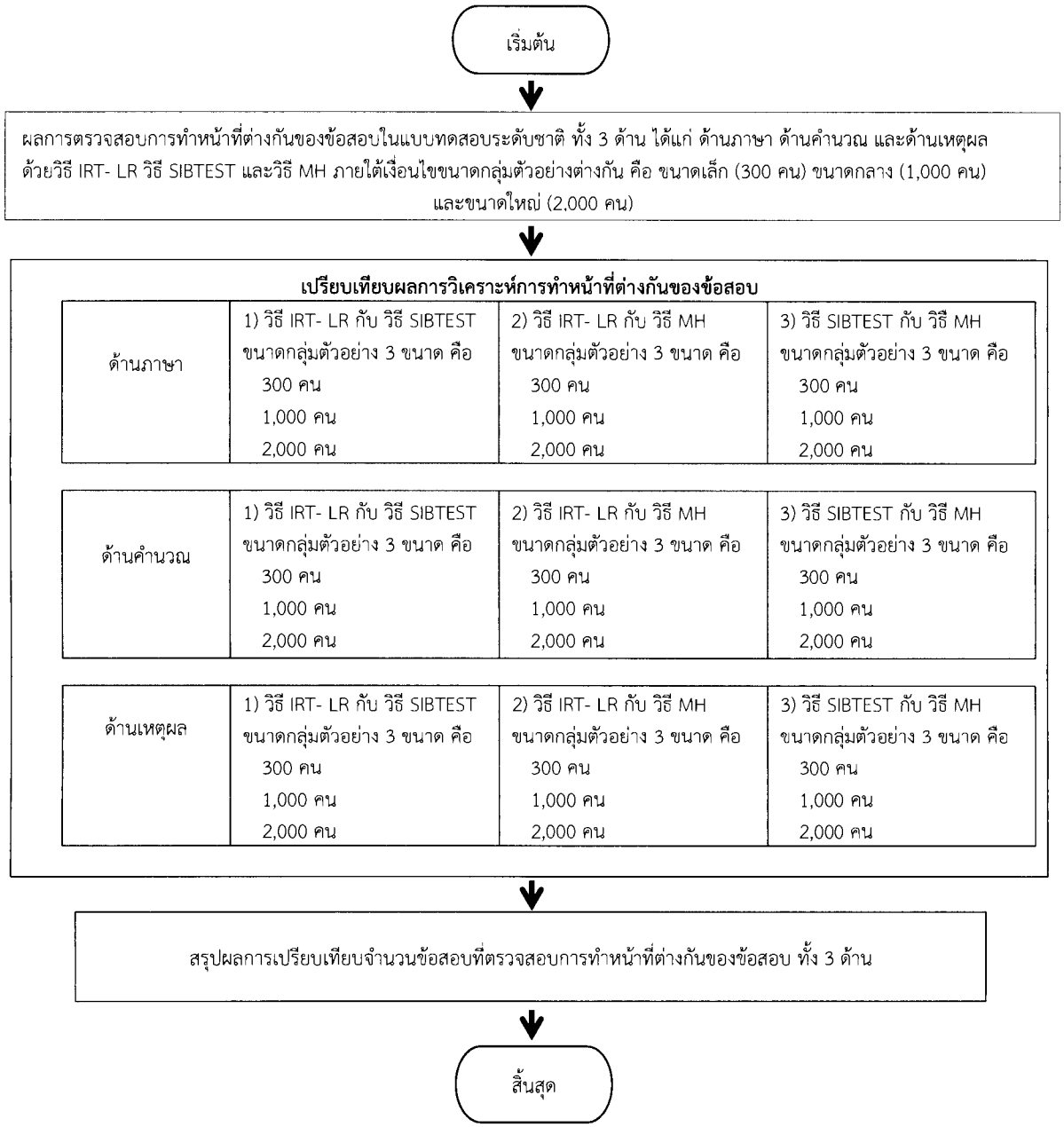
ภาพที่ 3-49 ตัวอย่างผลการวิเคราะห์ Chi-Square วิธี Mantel-Haenszel

5. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เพื่อวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ ทำการวิเคราะห์แบบทดสอบทุกข้อจากนั้นพิจารณาว่าข้อใดมี DIF โดยดูจากค่าไค-สแควร์ (Chi-square Statistic) ที่มีนัยสำคัญทางสถิติที่ระดับ .05

จากการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ ผู้วิจัยได้ดำเนินการรวบรวมผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ของวิธี IRT- LR วิธี SIBTEST และวิธี Mantel-Haenszel ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) และขั้นตอนการเปรียบเทียบผลการสอบการทำหน้าที่ต่างกันของข้อสอบมีดังนี้

ระยะที่ 3 การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้วยวิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)

การเปรียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ มีขั้นตอนดำเนินการ ดังนี้



ภาพที่ 3-50 ขั้นตอนการเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

ผู้วิจัยได้เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้วยวิธี IRT-LR วิธี SIBTEST และ วิธี MH ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) ตามสมมติฐานการวิจัย ดังนี้

1. เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา กลุ่มตัวอย่างขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) ระหว่างวิธี IRT-LR กับวิธี SIBTEST

2. เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา กลุ่มตัวอย่างขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) ระหว่างวิธี IRT-LR กับ วิธี MH

3. เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา กลุ่มตัวอย่างขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) ระหว่างวิธี SIBTEST กับวิธี MH

4. เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านคำนวณ กลุ่มตัวอย่างขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) ระหว่างวิธี IRT-LR กับวิธี SIBTEST

5. เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านคำนวณ กลุ่มตัวอย่างขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) ระหว่างวิธี IRT-LR กับ วิธี MH

6 .เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านคำนวณ กลุ่มตัวอย่างขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) ระหว่างวิธี SIBTEST กับวิธี MH

7. เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านเหตุผล กลุ่มตัวอย่างขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) ระหว่างวิธี IRT-LR กับวิธี SIBTEST

8. เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านเหตุผล กลุ่มตัวอย่างขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) ระหว่างวิธี IRT-LR กับ วิธี MH

9. เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านเหตุผล กลุ่มตัวอย่างขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) ระหว่างวิธี SIBTEST กับวิธี MH

บทที่ 4
ผลการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อวิเคราะห์คุณภาพของข้อสอบ ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ และเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ระหว่างวิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล ภายใต้เงื่อนไข ขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) ผู้วิจัยนำเสนอผลการวิจัยเป็น 3 ตอน ดังนี้

ตอนที่ 1 ผลการวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 โดยใช้หลักการของทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)

ตอนที่ 2 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้วยวิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)

ตอนที่ 3 ผลการเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้วยวิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)

สัญลักษณ์และอักษรย่อที่ใช้ในการเสนอผลการวิเคราะห์ข้อมูล มีดังนี้

n	หมายถึง	จำนวนตัวอย่าง
a	หมายถึง	ค่าอำนาจจำแนกของข้อสอบ
b	หมายถึง	ค่าความยากของข้อสอบ
c	หมายถึง	ค่าโอกาสการเดาของข้อสอบ

ตอนที่ 1 ผลการวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 โดยใช้หลักการของทฤษฎีการตอบสนองข้อสอบ แบบ 3 พารามิเตอร์ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)

การวิเคราะห์คุณภาพข้อสอบ NT ระดับชั้นประถมศึกษาปีที่ 3 เป็นการวิเคราะห์โดยใช้ทฤษฎีการตอบสนองข้อสอบแบบ 3 พารามิเตอร์ คือ ค่าอำนาจจำแนกของข้อสอบ (a) ค่าความยากของข้อสอบ (b) และค่าโอกาสการเดาของข้อสอบ (c) โดยใช้โปรแกรมสำเร็จรูป Xcalibre Version 4.2.2 ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) การประมาณค่าพารามิเตอร์ของข้อสอบจะมีเกณฑ์ในการคัดเลือกข้อสอบดังนี้

เกณฑ์การคัดเลือกข้อสอบ (Urry, 1977)

1. ค่าอำนาจจำแนกของข้อสอบ (a) มีค่าตั้งแต่ 0.50 ถึง 2.50
2. ค่าความยากของข้อสอบ (b) มีค่าตั้งแต่ -2.50 ถึง 2.50
3. ค่าการเดาของข้อสอบ (c) มีค่าไม่เกิน 0.30

เกณฑ์การประเมินค่าความยากของข้อสอบ (b) ระดับชั้นประถมศึกษาปีที่ 3 ดังนี้

ค่าความยากเฉลี่ยมากกว่า	2.0000	หมายถึง	ข้อสอบยากมาก
ค่าความยากเฉลี่ยตั้งแต่	1.0001 ถึง 2.0000	หมายถึง	ข้อสอบยาก
ค่าความยากเฉลี่ยตั้งแต่	0.5001 ถึง 1.0000	หมายถึง	ข้อสอบค่อนข้างยาก
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

ตารางที่ 4-1 ผลการวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา จำนวน 30 ข้อ โดยใช้ทฤษฎีการตอบสนองข้อสอบ (IRT) แบบ 3 พารามิเตอร์ กลุ่มตัวอย่างขนาดเล็ก (300 คน)

แบบทดสอบ NT	ข้อที่	ค่าพารามิเตอร์		
		a	b	c
ด้านภาษา	1	0.314	1.066	0.264
	2	0.377	1.709	0.255
	3	0.574	0.499	0.364
	4	0.549	-0.488	0.254
	5	0.646	0.736	0.250
	6	0.423	-0.538	0.259

ตารางที่ 4-1 (ต่อ)

แบบทดสอบ NT	ข้อที่	ค่าพารามิเตอร์		
		a	b	c
ด้านภาษา	7	0.680	-0.438	0.247
	8	0.558	1.209	0.258
	9	0.516	-1.120	0.252
	10	0.582	1.790	0.263
	11	0.472	-0.581	0.255
	12	0.628	0.280	0.248
	13	0.553	1.166	0.258
	14	0.612	1.007	0.251
	15	0.602	1.348	0.246
	16	0.579	1.257	0.252
	17	0.782	-0.189	0.248
	18	0.743	1.158	0.238
	19	0.839	-0.660	0.252
	20	0.654	-0.173	0.251
	21	0.478	0.802	0.258
	22	0.931	2.316	0.214
	23	0.695	2.829	0.231
	24	0.676	1.385	0.250
	25	0.652	2.153	0.278
	26	0.647	1.212	0.262
	27	0.656	1.289	0.251
	28	0.617	0.245	0.256
	29	0.611	0.910	0.244
	30	0.788	0.629	0.238
ค่าความเที่ยงทั้งฉบับ (Reliability)			0.712	

จากตารางที่ 4-1 ข้อสอบด้านภาษา จากจำนวนข้อสอบ 30 ข้อ มีข้อสอบที่ผ่านเกณฑ์การวิเคราะห์คุณภาพข้อสอบ จำนวน 30 ข้อ เป็นข้อสอบที่มีค่าอำนาจจำแนกของข้อสอบ (a) ตั้งแต่ 0.314 ถึง 0.931 ค่าความยากของข้อสอบ (b) ตั้งแต่ -1.120 ถึง 2.829 และค่าโอกาสการเดาของข้อสอบ (c) ตั้งแต่ 0.214 ถึง 0.364 และมีค่าความเที่ยงทั้งฉบับ 0.712

ตารางที่ 4-2 ผลการวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านคำนวณ จำนวน 30 ข้อ โดยใช้ทฤษฎีการตอบสนองข้อสอบ (IRT) แบบ 3 พารามิเตอร์ กลุ่มตัวอย่างขนาดเล็ก (300 คน)

แบบทดสอบ NT	ข้อที่	ค่าพารามิเตอร์		
		a	b	c
ด้านคำนวณ	1	0.731	1.430	0.079
	2	0.624	1.351	0.081
	3	0.563	2.395	0.248
	4	0.940	1.362	0.339
	5	0.901	1.315	0.225
	6	0.895	2.137	0.216
	7	0.795	2.408	0.229
	8	0.767	0.701	0.247
	9	0.678	0.735	0.254
	10	0.791	1.365	0.241
	11	0.699	0.893	0.260
	12	0.715	2.537	0.256
	13	0.791	1.281	0.250
	14	0.983	1.250	0.230
	15	0.738	1.732	0.251
	16	0.860	0.905	0.248
	17	0.540	0.271	0.256
	18	0.676	1.457	0.239
	19	0.648	1.636	0.246
	20	0.527	0.728	0.257
	21	0.807	1.772	0.254
	22	0.699	2.112	0.238
	23	0.921	3.503	0.222
	24	0.706	1.867	0.255
	25	0.733	1.718	0.239
	26	0.915	1.592	0.235
	27	0.807	0.662	0.241

ตารางที่ 4-2 (ต่อ)

แบบทดสอบ NT	ข้อที่	ค่าพารามิเตอร์		
		a	b	c
ด้านคำนวณ	28	1.024	1.004	0.242
	29	0.746	0.989	0.249
	30	1.073	1.180	0.222
ค่าความเที่ยงทั้งฉบับ (Reliability)		0.763		

จากตารางที่ 4-2 ข้อสอบด้านคำนวณ จากจำนวนข้อสอบ 30 ข้อ มีข้อสอบที่ผ่านเกณฑ์การวิเคราะห์คุณภาพข้อสอบ จำนวน 30 ข้อ เป็นข้อสอบที่มีค่าอำนาจจำแนกของข้อสอบ (a) ตั้งแต่ 0.527 ถึง 1.073 ค่าความยากของข้อสอบ (b) ตั้งแต่ 0.271 ถึง 3.503 และค่าโอกาสการเดาของข้อสอบ (c) ตั้งแต่ 0.079 ถึง 0.339 และมีค่าความเที่ยงทั้งฉบับ 0.763

ตารางที่ 4-3 ผลการวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านเหตุผล จำนวน 30 ข้อ โดยใช้ทฤษฎีการตอบสนองข้อสอบ (IRT) แบบ 3 พารามิเตอร์ กลุ่มตัวอย่างขนาดเล็ก (300 คน)

แบบทดสอบ NT	ข้อที่	ค่าพารามิเตอร์		
		a	b	c
ด้านเหตุผล	1	0.629	0.628	0.286
	2	0.806	0.400	0.275
	3	0.712	0.945	0.379
	4	0.787	-0.432	0.248
	5	0.455	0.390	0.251
	6	0.710	0.098	0.248
	7	0.733	0.327	0.245
	8	0.752	1.234	0.270
	9	0.679	0.789	0.251
	10	0.798	1.432	0.246
	11	0.609	0.497	0.252
	12	0.707	0.476	0.251
	13	0.552	1.053	0.254
	14	0.899	1.596	0.241

ตารางที่ 4-3 (ต่อ)

แบบทดสอบ NT	ข้อที่	ค่าพารามิเตอร์		
		a	b	c
ด้านเหตุผล	15	0.701	1.292	0.256
	16	0.954	2.022	0.254
	17	0.781	2.820	0.247
	18	0.821	0.850	0.243
	19	0.668	1.189	0.249
	20	0.772	2.384	0.270
	21	0.791	0.403	0.244
	22	0.787	0.597	0.248
	23	0.781	0.723	0.239
	24	0.821	0.331	0.247
	25	0.610	1.169	0.258
	26	0.786	0.637	0.251
	27	0.802	1.811	0.234
	28	0.737	2.031	0.236
	29	0.933	1.555	0.228
	30	1.289	0.886	0.220
ค่าความเที่ยงทั้งฉบับ (Reliability)		0.766		

จากตารางที่ 4-3 ข้อสอบด้านเหตุผล จากจำนวนข้อสอบ 30 ข้อ มีข้อสอบที่ผ่านเกณฑ์การวิเคราะห์คุณภาพข้อสอบ จำนวน 30 ข้อ เป็นข้อสอบที่มีค่าอำนาจจำแนกของข้อสอบ (a) ตั้งแต่ 0.455 ถึง 1.289 ค่าความยากของข้อสอบ (b) ตั้งแต่ -0.432 ถึง 2.820 และค่าโอกาสการเดาของข้อสอบ (c) ตั้งแต่ 0.220 ถึง 0.379 และมีค่าความเที่ยงทั้งฉบับ 0.766

ตารางที่ 4-4 ผลการวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา จำนวน 30 ข้อ โดยใช้ทฤษฎีการตอบสนองข้อสอบ (IRT) แบบ 3 พารามิเตอร์ กลุ่มตัวอย่างขนาดกลาง (1,000 คน)

แบบทดสอบ NT	ข้อที่	ค่าพารามิเตอร์		
		a	b	c
ด้านภาษา	1	0.337	0.483	0.188
	2	0.323	3.475	0.209
	3	0.644	0.443	0.328

ตารางที่ 4-4 (ต่อ)

แบบทดสอบ NT	ข้อที่	ค่าพารามิเตอร์		
		a	b	c
ด้านภาษา	4	0.503	-0.950	0.253
	5	0.575	0.087	0.242
	6	0.571	-0.749	0.254
	7	0.849	-0.710	0.244
	8	0.609	0.873	0.247
	9	0.587	-1.207	0.251
	10	0.645	1.972	0.283
	11	0.469	-0.572	0.262
	12	0.760	0.285	0.246
	13	0.523	1.859	0.269
	14	0.786	0.796	0.244
	15	0.481	1.965	0.270
	16	0.704	1.120	0.245
	17	0.783	-0.400	0.246
	18	0.648	1.100	0.241
	19	1.010	-0.903	0.246
	20	0.742	-0.408	0.243
	21	0.679	0.627	0.252
	22	0.993	2.133	0.200
	23	0.556	2.297	0.246
	24	0.618	0.882	0.243
	25	0.702	2.695	0.303
	26	0.440	1.510	0.280
	27	0.759	1.070	0.242
	28	0.685	0.516	0.262
	29	0.587	1.396	0.236
	30	0.766	0.302	0.236
ค่าความเที่ยงทั้งฉบับ (Reliability)			0.738	

จากตารางที่ 4-4 ข้อสอบด้านภาษา จากจำนวนข้อสอบ 30 ข้อ มีข้อสอบที่ผ่านเกณฑ์การวิเคราะห์คุณภาพข้อสอบ จำนวน 30 ข้อ เป็นข้อสอบที่มีค่าอำนาจจำแนกของข้อสอบ (a) ตั้งแต่ 0.323 ถึง 1.010 ค่าความยากของข้อสอบ (b) ตั้งแต่ -1.207 ถึง 3.475 และค่าโอกาสการเดาของข้อสอบ (c) ตั้งแต่ 0.188 ถึง 0.328 และมีค่าความเที่ยงทั้งฉบับ 0.738

ตารางที่ 4-5 ผลการวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านคำนวณ จำนวน 30 ข้อ โดยใช้ทฤษฎีการตอบสนองข้อสอบ (IRT) แบบ 3 พารามิเตอร์ กลุ่มตัวอย่างขนาดกลาง (1,000 คน)

แบบทดสอบ NT	ข้อที่	ค่าพารามิเตอร์		
		a	b	c
ด้านคำนวณ	1	0.619	1.672	0.080
	2	0.827	0.733	0.077
	3	0.796	2.426	0.277
	4	1.188	0.783	0.307
	5	0.843	1.444	0.214
	6	0.959	1.484	0.202
	7	1.044	1.771	0.197
	8	1.057	0.144	0.244
	9	0.958	0.358	0.248
	10	1.008	1.043	0.259
	11	0.804	1.019	0.271
	12	0.566	2.370	0.252
	13	0.593	0.748	0.250
	14	0.786	1.835	0.237
	15	0.777	1.720	0.262
	16	1.163	0.710	0.243
	17	0.750	0.016	0.253
	18	0.803	1.040	0.228
	19	0.745	1.005	0.241
	20	0.524	0.815	0.254
	21	0.654	1.817	0.238
	22	1.147	2.452	0.216
	23	1.165	3.688	0.203
	24	0.918	1.069	0.258

ตารางที่ 4-5 (ต่อ)

แบบทดสอบ NT	ข้อที่	ค่าพารามิเตอร์		
		a	b	c
ด้านคำนวณ	25	0.943	0.935	0.224
	26	0.822	1.529	0.250
	27	0.761	0.567	0.236
	28	0.994	1.384	0.229
	29	0.823	1.257	0.261
	30	0.978	1.127	0.212
ค่าความเที่ยงทั้งฉบับ (Reliability)			0.791	

จากตารางที่ 4-5 ข้อสอบด้านคำนวณ จากจำนวนข้อสอบ 30 ข้อ มีข้อสอบที่ผ่านเกณฑ์การวิเคราะห์คุณภาพข้อสอบ จำนวน 30 ข้อ เป็นข้อสอบที่มีค่าอำนาจจำแนกของข้อสอบ (a) ตั้งแต่ 0.524 ถึง 1.188 ค่าความยากของข้อสอบ (b) ตั้งแต่ 0.016 ถึง 3.688 และค่าโอกาสการเดาของข้อสอบ (c) ตั้งแต่ 0.077 ถึง 0.307 และมีค่าความเที่ยงทั้งฉบับ 0.791

ตารางที่ 4-6 ผลการวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านเหตุผล จำนวน 30 ข้อ โดยใช้ทฤษฎีการตอบสนองข้อสอบ (IRT) แบบ 3 พารามิเตอร์ กลุ่มตัวอย่างขนาดกลาง (1,000 คน)

แบบทดสอบ NT	ข้อที่	ค่าพารามิเตอร์		
		a	b	c
ด้านเหตุผล	1	0.686	0.177	0.247
	2	0.796	0.020	0.236
	3	0.655	0.317	0.336
	4	0.761	-0.533	0.246
	5	0.537	-0.088	0.243
	6	0.581	-0.067	0.250
	7	0.754	0.235	0.241
	8	0.806	0.801	0.273
	9	0.672	0.232	0.252
	10	0.833	1.119	0.237
	11	0.435	0.705	0.253
	12	0.633	0.728	0.238
	13	0.496	0.774	0.251

ตารางที่ 4-6 (ต่อ)

แบบทดสอบ NT	ข้อที่	ค่าพารามิเตอร์		
		a	b	c
ด้านเหตุผล	14	0.661	1.315	0.240
	15	0.646	1.548	0.256
	16	0.908	1.940	0.266
	17	0.999	2.908	0.231
	18	0.694	0.903	0.226
	19	0.530	0.965	0.235
	20	0.925	3.056	0.270
	21	0.806	0.675	0.254
	22	0.964	0.398	0.262
	23	0.934	0.142	0.227
	24	1.113	0.113	0.248
	25	0.516	1.744	0.271
	26	0.776	0.409	0.236
	27	0.901	1.144	0.217
	28	0.830	1.224	0.202
	29	0.884	1.677	0.208
	30	1.115	1.090	0.225
ค่าความเที่ยงทั้งฉบับ (Reliability)			0.784	

จากตารางที่ 4-6 ข้อสอบด้านเหตุผล จากจำนวนข้อสอบ 30 ข้อ มีข้อสอบที่ผ่านเกณฑ์การวิเคราะห์คุณภาพข้อสอบ จำนวน 30 ข้อ เป็นข้อสอบที่มีค่าอำนาจจำแนกของข้อสอบ (a) ตั้งแต่ 0.435 ถึง 1.115 ค่าความยากของข้อสอบ (b) ตั้งแต่ -0.533 ถึง 3.056 และค่าโอกาสการเดาของข้อสอบ (c) ตั้งแต่ 0.202 ถึง 0.336 และมีค่าความเที่ยงทั้งฉบับ 0.784

ตารางที่ 4-7 ผลการวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา จำนวน 30 ข้อ โดยใช้ทฤษฎีการตอบสนองข้อสอบ (IRT) แบบ 3 พารามิเตอร์ กลุ่มตัวอย่างขนาดใหญ่ (2,000 คน)

แบบทดสอบ NT	ข้อที่	ค่าพารามิเตอร์		
		a	b	c
ด้านภาษา	1	0.480	0.453	0.211
	2	0.585	2.064	0.216

ตารางที่ 4-7 (ต่อ)

แบบทดสอบ NT	ข้อที่	ค่าพารามิเตอร์		
		a	b	c
ด้านภาษา	3	0.545	-0.030	0.208
	4	0.625	-0.931	0.214
	5	0.539	0.153	0.210
	6	0.708	-0.772	0.215
	7	1.089	-0.432	0.211
	8	0.670	0.614	0.208
	9	0.627	-1.140	0.215
	10	0.503	1.508	0.231
	11	0.588	-0.582	0.222
	12	0.928	0.051	0.201
	13	0.415	1.520	0.216
	14	0.822	0.575	0.199
	15	0.496	1.260	0.218
	16	0.698	0.525	0.201
	17	1.001	-0.398	0.213
	18	0.733	0.733	0.203
	19	0.949	-0.902	0.211
	20	0.737	-0.384	0.213
	21	0.594	0.546	0.206
	22	0.984	1.606	0.179
	23	0.546	1.904	0.207
	24	0.559	0.698	0.206
	25	0.438	2.113	0.232
	26	0.318	1.339	0.220
	27	0.722	1.021	0.222
	28	0.689	0.193	0.220
	29	0.555	0.877	0.194
	30	0.741	0.102	0.203
ค่าความเที่ยงทั้งฉบับ (Reliability)			0.729	

จากตารางที่ 4-7 ข้อสอบด้านภาษา จากจำนวนข้อสอบ 30 ข้อ มีข้อสอบที่ผ่านเกณฑ์การวิเคราะห์คุณภาพข้อสอบ จำนวน 30 ข้อ เป็นข้อสอบที่มีค่าอำนาจจำแนกของข้อสอบ (a) ตั้งแต่ 0.318 ถึง 1.089 ค่าความยากของข้อสอบ (b) ตั้งแต่ -1.140 ถึง 2.113 และค่าโอกาสการเดาของข้อสอบ (c) ตั้งแต่ 0.179 ถึง 0.232 และมีค่าความเที่ยงทั้งฉบับ 0.729

ตารางที่ 4-8 ผลการวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านคำนวณ จำนวน 30 ข้อ โดยใช้ทฤษฎีการตอบสนองข้อสอบ (IRT) แบบ 3 พารามิเตอร์ กลุ่มตัวอย่างขนาดใหญ่ (2,000 คน)

แบบทดสอบ NT	ข้อที่	ค่าพารามิเตอร์		
		a	b	c
ด้านคำนวณ	1	0.695	1.620	0.080
	2	1.007	1.267	0.171
	3	0.693	2.681	0.276
	4	1.061	0.767	0.241
	5	0.855	1.335	0.206
	6	0.854	1.748	0.200
	7	1.037	1.640	0.203
	8	0.969	0.286	0.257
	9	0.839	0.449	0.243
	10	0.947	0.989	0.239
	11	0.746	1.023	0.255
	12	0.480	2.713	0.262
	13	0.638	0.590	0.234
	14	0.589	2.214	0.251
	15	0.711	1.607	0.251
	16	1.002	0.986	0.243
	17	0.744	0.061	0.255
	18	0.829	1.290	0.220
	19	0.903	1.340	0.234
	20	0.483	1.018	0.256
	21	0.605	2.024	0.244

ตารางที่ 4-8 (ต่อ)

แบบทดสอบ NT	ข้อที่	ค่าพารามิเตอร์		
		a	b	c
ด้านคำนวณ	22	1.125	2.886	0.213
	23	1.208	4.000	0.193
	24	0.960	1.341	0.271
	25	0.997	1.068	0.233
	26	0.921	1.445	0.239
	27	0.773	0.796	0.234
	28	0.714	1.755	0.214
	29	0.727	1.651	0.240
	30	0.973	1.345	0.192
ค่าความเที่ยงทั้งฉบับ (Reliability)		0.775		

จากตารางที่ 4-8 ข้อสอบด้านคำนวณ จากจำนวนข้อสอบ 30 ข้อ มีข้อสอบที่ผ่านเกณฑ์การวิเคราะห์คุณภาพข้อสอบ จำนวน 30 ข้อ เป็นข้อสอบที่มีค่าอำนาจจำแนกของข้อสอบ (a) ตั้งแต่ 0.480 ถึง 1.208 ค่าความยากของข้อสอบ (b) ตั้งแต่ 0.061 ถึง 4.000 และค่าโอกาสการเดาของข้อสอบ (c) ตั้งแต่ 0.080 ถึง 0.276 และมีค่าความเที่ยงทั้งฉบับ 0.775

ตารางที่ 4-9 ผลการวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านเหตุผล จำนวน 30 ข้อ โดยใช้ทฤษฎีการตอบสนองข้อสอบ (IRT) แบบ 3 พารามิเตอร์ กลุ่มตัวอย่างขนาดใหญ่ (2,000 คน)

แบบทดสอบ NT	ข้อที่	ค่าพารามิเตอร์		
		a	b	c
ด้านเหตุผล	1	0.681	0.413	0.257
	2	0.925	0.214	0.248
	3	0.746	0.309	0.315
	4	0.667	-0.442	0.241
	5	0.542	-0.107	0.237
	6	0.551	-0.100	0.252
	7	0.703	0.363	0.230
	8	0.822	0.758	0.264
	9	0.613	0.311	0.255

ตารางที่ 4-9 (ต่อ)

แบบทดสอบ NT	ข้อที่	ค่าพารามิเตอร์		
		a	b	c
ด้านเหตุผล	10	0.737	1.113	0.225
	11	0.503	0.455	0.237
	12	0.540	0.549	0.238
	13	0.584	0.900	0.241
	14	0.794	1.084	0.225
	15	0.646	1.604	0.252
	16	1.004	1.738	0.275
	17	0.898	2.323	0.227
	18	0.730	0.756	0.222
	19	0.628	0.702	0.221
	20	0.955	2.968	0.255
	21	0.793	0.751	0.244
	22	0.792	0.571	0.251
	23	0.833	0.286	0.216
	24	1.068	0.200	0.240
	25	0.508	1.613	0.260
	26	0.697	0.460	0.214
	27	0.843	1.188	0.210
	28	0.727	1.451	0.196
	29	0.822	1.802	0.205
	30	0.914	1.311	0.210
ค่าความเที่ยงทั้งฉบับ (Reliability)		0.791		

จากตารางที่ 4-9 ข้อสอบด้านเหตุผล จากจำนวนข้อสอบ 30 ข้อ มีข้อสอบที่ผ่านเกณฑ์การวิเคราะห์คุณภาพข้อสอบ จำนวน 30 ข้อ เป็นข้อสอบที่มีค่าอำนาจจำแนกของข้อสอบ (a) ตั้งแต่ 0.503 ถึง 1.068 ค่าความยากของข้อสอบ (b) ตั้งแต่ -0.442 ถึง 2.968 และค่าโอกาสการเดาของข้อสอบ (c) ตั้งแต่ 0.196 ถึง 0.315 และมีค่าความเที่ยงทั้งฉบับ 0.791

จากการวิเคราะห์คุณภาพของแบบทดสอบระดับชาติ โดยการวิเคราะห์ค่าพารามิเตอร์ตามทฤษฎีการตอบสนองข้อสอบ พบว่า

แบบทดสอบระดับชาติ ด้านภาษา ขนาดกลุ่มตัวอย่าง 300 คน มีค่าอำนาจจำแนกของข้อสอบ (a) มีค่าเฉลี่ยเท่ากับ 0.614 ค่าความยากของข้อสอบ (b) มีค่าเฉลี่ยเท่ากับ 0.760 และ

ค่าโอกาสการเดาของข้อสอบ (c) มีค่าเฉลี่ยเท่ากับ 0.255 สรุปได้ว่าแบบทดสอบด้านภาษา ขนาดกลุ่มตัวอย่าง 300 คน มีอำนาจจำแนกข้อสอบอยู่ในระดับดีมาก มีค่าความยากของข้อสอบอยู่ในระดับข้อสอบค่อนข้างยาก และค่าโอกาสการเดาของข้อสอบ ไม่เกิน 0.3

แบบทดสอบระดับชาติ ด้านคำนวณ ขนาดกลุ่มตัวอย่าง 300 คน มีค่าอำนาจจำแนกของข้อสอบ (a) มีค่าเฉลี่ยเท่ากับ 0.776 ค่าความยากของข้อสอบ (b) มีค่าเฉลี่ยเท่ากับ 1.476 และค่าโอกาสการเดาของข้อสอบ (c) มีค่าเฉลี่ยเท่ากับ 0.235 สรุปได้ว่าแบบทดสอบด้านคำนวณ ขนาดกลุ่มตัวอย่าง 300 คน มีอำนาจจำแนกข้อสอบอยู่ในระดับดีมาก มีค่าความยากของข้อสอบอยู่ในระดับข้อสอบยาก และค่าโอกาสการเดาของข้อสอบ ไม่เกิน 0.3

แบบทดสอบระดับชาติ ด้านเหตุผล ขนาดกลุ่มตัวอย่าง 300 คน มีค่าอำนาจจำแนกของข้อสอบ (a) มีค่าเฉลี่ยเท่ากับ 0.762 ค่าความยากของข้อสอบ (b) มีค่าเฉลี่ยเท่ากับ 1.004 และค่าโอกาสการเดาของข้อสอบ (c) มีค่าเฉลี่ยเท่ากับ 0.254 สรุปได้ว่าแบบทดสอบด้านเหตุผล ขนาดกลุ่มตัวอย่าง 300 คน มีอำนาจจำแนกข้อสอบอยู่ในระดับดีมาก มีค่าความยากของข้อสอบอยู่ในระดับข้อสอบยาก และค่าโอกาสการเดาของข้อสอบ ไม่เกิน 0.3

แบบทดสอบระดับชาติ ด้านภาษา ขนาดกลุ่มตัวอย่าง 1,000 คน มีค่าอำนาจจำแนกของข้อสอบ (a) มีค่าเฉลี่ยเท่ากับ 0.644 ค่าความยากของข้อสอบ (b) มีค่าเฉลี่ยเท่ากับ 0.733 และค่าโอกาสการเดาของข้อสอบ (c) มีค่าเฉลี่ยเท่ากับ 0.250 สรุปได้ว่าแบบทดสอบด้านภาษา ขนาดกลุ่มตัวอย่าง 1,000 คน มีอำนาจจำแนกข้อสอบอยู่ในระดับดีมาก มีค่าความยากของข้อสอบอยู่ในระดับข้อสอบค่อนข้างยาก และค่าโอกาสการเดาของข้อสอบไม่เกิน 0.3

แบบทดสอบระดับชาติ ด้านคำนวณ ขนาดกลุ่มตัวอย่าง 1,000 คน มีค่าอำนาจจำแนกของข้อสอบ (a) มีค่าเฉลี่ยเท่ากับ 0.867 ค่าความยากของข้อสอบ (b) มีค่าเฉลี่ยเท่ากับ 1.297 และค่าโอกาสการเดาของข้อสอบ (c) มีค่าเฉลี่ยเท่ากับ 0.231 สรุปได้ว่าแบบทดสอบด้านคำนวณ ขนาดกลุ่มตัวอย่าง 1,000 คน มีอำนาจจำแนกข้อสอบอยู่ในระดับดีมาก มีค่าความยากของข้อสอบอยู่ในระดับข้อสอบยาก และค่าโอกาสการเดาของข้อสอบ ไม่เกิน 0.3

แบบทดสอบระดับชาติ ด้านเหตุผล ขนาดกลุ่มตัวอย่าง 1,000 คน มีค่าอำนาจจำแนกของข้อสอบ (a) มีค่าเฉลี่ยเท่ากับ 0.762 ค่าความยากของข้อสอบ (b) มีค่าเฉลี่ยเท่ากับ 0.856 และค่าโอกาสการเดาของข้อสอบ (c) มีค่าเฉลี่ยเท่ากับ 0.246 สรุปได้ว่าแบบทดสอบด้านเหตุผล ขนาดกลุ่มตัวอย่าง 1,000 คน มีอำนาจจำแนกข้อสอบอยู่ในระดับดีมาก มีค่าความยากของข้อสอบอยู่ในระดับข้อสอบค่อนข้างยาก และค่าโอกาสการเดาของข้อสอบ ไม่เกิน 0.3

แบบทดสอบระดับชาติ ด้านภาษา ขนาดกลุ่มตัวอย่าง 2,000 คน มีค่าอำนาจจำแนกของข้อสอบ (a) มีค่าเฉลี่ยเท่ากับ 0.643 ค่าความยากของข้อสอบ (b) มีค่าเฉลี่ยเท่ากับ 0.476 และค่าโอกาสการเดาของข้อสอบ (c) มีค่าเฉลี่ยเท่ากับ 0.211 สรุปได้ว่าแบบทดสอบด้านภาษา ขนาดกลุ่มตัวอย่าง 2,000 คน มีอำนาจจำแนกข้อสอบอยู่ในระดับดีมาก มีค่าความยากของข้อสอบอยู่ในระดับข้อสอบปานกลาง และค่าโอกาสการเดาของข้อสอบ ไม่เกิน 0.3

แบบทดสอบระดับชาติ ด้านคำนวณ ขนาดกลุ่มตัวอย่าง 2,000 คน มีค่าอำนาจจำแนกของข้อสอบ (a) มีค่าเฉลี่ยเท่ากับ 0.836 ค่าความยากของข้อสอบ (b) มีค่าเฉลี่ยเท่ากับ 1.465 และค่าโอกาสการเดาของข้อสอบ (c) มีค่าเฉลี่ยเท่ากับ 0.228 สรุปได้ว่าแบบทดสอบด้านคำนวณ

ขนาดกลุ่มตัวอย่าง 2,000 คน มีอำนาจจำแนกข้อสอบอยู่ในระดับดีมาก มีค่าความยากของข้อสอบอยู่ในระดับข้อสอบยาก และค่าโอกาสการเดาของข้อสอบ ไม่เกิน 0.3

แบบทดสอบระดับชาติ ด้านเหตุผล ขนาดกลุ่มตัวอย่าง 2,000 คน มีค่าอำนาจจำแนกของข้อสอบ (a) มีค่าเฉลี่ยเท่ากับ 0.742 ค่าความยากของข้อสอบ (b) มีค่าเฉลี่ยเท่ากับ 0.851 และค่าโอกาสการเดาของข้อสอบ (c) มีค่าเฉลี่ยเท่ากับ 0.239 สรุปได้ว่าแบบทดสอบด้านเหตุผล ขนาดกลุ่มตัวอย่าง 2,000 คน มีอำนาจจำแนกข้อสอบอยู่ในระดับดีมาก มีค่าความยากของข้อสอบอยู่ในระดับข้อสอบค่อนข้างยาก และค่าโอกาสการเดาของข้อสอบ ไม่เกิน 0.3

ตารางที่ 4-10 สรุปผลการวิเคราะห์คุณภาพของแบบทดสอบรายข้อ และทั้งฉบับ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล

แบบทดสอบ	กลุ่มตัวอย่าง (คน)	ค่าอำนาจจำแนกของข้อสอบ (a)	ค่าความยากของข้อสอบ (b)	ค่าโอกาสการเดาของข้อสอบ (c)	ค่าความเที่ยงทั้งฉบับ
ด้านภาษา	300	0.314 - 0.931	-1.120 – 2.829	0.214 – 0.364	0.712
	1,000	0.323 – 1.010	-1.207 – 3.475	0.188 – 0.328	0.738
	2,000	0.318 – 1.089	-1.140 – 2.113	0.179 – 0.232	0.729
ด้านคำนวณ	300	0.527 – 1.073	0.271 – 3.503	0.079 – 0.339	0.763
	1,000	0.524 – 1.188	0.016 – 3.688	0.077 – 0.307	0.791
	2,000	0.480 – 1.208	0.061 – 4.000	0.080 – 0.276	0.775
ด้านเหตุผล	300	0.455 – 1.289	-0.432 – 2.820	0.220 – 0.379	0.766
	1,000	0.435 – 1.115	-0.533 – 3.056	0.202 – 0.336	0.784
	2,000	0.503 – 1.068	-0.442 – 2.968	0.196 – 0.315	0.791

จากตารางที่ 4-10 ปรากฏว่า แบบทดสอบด้านภาษา เมื่อกลุ่มตัวอย่างขนาดเล็ก (300 คน) มีค่าความเที่ยงทั้งฉบับเท่ากับ 0.712 กลุ่มตัวอย่างขนาดกลาง (1,000 คน) มีค่าความเที่ยงทั้งฉบับเท่ากับ 0.738 และกลุ่มตัวอย่างขนาดใหญ่ (2,000 คน) มีค่าความเที่ยงทั้งฉบับเท่ากับ 0.729 แบบทดสอบด้านคำนวณ เมื่อกลุ่มตัวอย่างขนาดเล็ก (300 คน) มีค่าความเที่ยงทั้งฉบับเท่ากับ 0.763 กลุ่มตัวอย่างขนาดกลาง (1,000 คน) มีค่าความเที่ยงทั้งฉบับเท่ากับ 0.791 และกลุ่มตัวอย่างขนาดใหญ่ (2,000 คน) มีค่าความเที่ยงทั้งฉบับเท่ากับ 0.775 ส่วนแบบทดสอบด้านเหตุผล เมื่อกลุ่มตัวอย่างขนาดเล็ก (300 คน) มีค่าความเที่ยงทั้งฉบับเท่ากับ 0.766 กลุ่มตัวอย่างขนาดกลาง (1,000 คน) มีค่าความเที่ยงทั้งฉบับเท่ากับ 0.784 และกลุ่มตัวอย่างขนาดใหญ่ (2,000 คน) มีค่าความเที่ยงทั้งฉบับเท่ากับ 0.791

ตอนที่ 2 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้วยวิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 แบบทดสอบแบ่งออกเป็น 3 ด้าน คือ ด้านภาษา 30 ข้อ ด้านคำนวณ 30 ข้อ ด้านเหตุผล 30 ข้อ ด้วยวิธี IRT-LR วิธี SIBTEST และวิธี MH ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)

ตารางที่ 4-11 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ระดับชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ ด้านเหตุผล ด้วยวิธี IRT-LR ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างขนาดเล็ก (300 คน)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี IRT-LR กลุ่มตัวอย่าง (300 คน)		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
1	NO-DIF	NO-DIF	NO-DIF
2	DIF	NO-DIF	NO-DIF
3	NO-DIF	NO-DIF	DIF
4	NO-DIF	NO-DIF	NO-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	NO-DIF
9	NO-DIF	NO-DIF	NO-DIF
10	NO-DIF	NO-DIF	NO-DIF
11	NO-DIF	DIF	NO-DIF
12	NO-DIF	DIF	DIF
13	NO-DIF	NO-DIF	NO-DIF
14	NO-DIF	NO-DIF	DIF
15	NO-DIF	NO-DIF	NO-DIF
16	NO-DIF	NO-DIF	NO-DIF
17	DIF	NO-DIF	NO-DIF
18	NO-DIF	NO-DIF	NO-DIF

ตารางที่ 4-11 (ต่อ)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี IRT-LR		
	กลุ่มตัวอย่าง (300 คน)		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
19	NO-DIF	NO-DIF	NO-DIF
20	NO-DIF	NO-DIF	NO-DIF
21	NO-DIF	DIF	NO-DIF
22	NO-DIF	NO-DIF	NO-DIF
23	NO-DIF	NO-DIF	NO-DIF
24	NO-DIF	NO-DIF	NO-DIF
25	NO-DIF	NO-DIF	NO-DIF
26	NO-DIF	NO-DIF	NO-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	NO-DIF
จำนวนข้อที่พบ DIF พบว่า	(3 ข้อ) 10%	(3 ข้อ) 10%	(3 ข้อ) 10%

หมายเหตุ DIF หมายถึง ข้อสอบที่ทำหน้าที่ต่างกัน
NO-DIF หมายถึง ข้อสอบที่ไม่พบว่าทำหน้าที่ต่างกัน

จากตารางที่ 4-11 แสดงผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ ด้านเหตุผล ด้วยวิธี IRT-LR ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่าง จำนวน 300 คน พบว่า

ด้านภาษา ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ จำนวน 3 ข้อ ได้แก่ ข้อที่ 2, 17, 30 คิดเป็นร้อยละ 10 ของข้อสอบทั้งหมด

ด้านคำนวณ ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ จำนวน 3 ข้อ ได้แก่ ข้อที่ 11, 12, 21 คิดเป็นร้อยละ 10 ของข้อสอบทั้งหมด

ด้านเหตุผล ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ จำนวน 3 ข้อ ได้แก่ ข้อที่ 3, 12, 14 คิดเป็นร้อยละ 10 ของข้อสอบทั้งหมด

ตารางที่ 4-12 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ระดับชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ ด้านเหตุผล ด้วยวิธี IRT-LR ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างขนาดกลาง (1,000 คน)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี IRT-LR กลุ่มตัวอย่าง (1,000 คน)		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
1	NO-DIF	NO-DIF	DIF
2	NO-DIF	NO-DIF	NO-DIF
3	NO-DIF	NO-DIF	DIF
4	NO-DIF	NO-DIF	NO-DIF
5	DIF	NO-DIF	NO-DIF
6	NO-DIF	NO-DIF	NO-DIF
7	NO-DIF	NO-DIF	NO-DIF
8	NO-DIF	NO-DIF	NO-DIF
9	NO-DIF	NO-DIF	NO-DIF
10	NO-DIF	DIF	NO-DIF
11	NO-DIF	NO-DIF	DIF
12	NO-DIF	DIF	DIF
13	NO-DIF	NO-DIF	NO-DIF
14	NO-DIF	NO-DIF	NO-DIF
15	DIF	NO-DIF	NO-DIF
16	NO-DIF	DIF	NO-DIF
17	NO-DIF	NO-DIF	NO-DIF
18	DIF	NO-DIF	NO-DIF
19	NO-DIF	DIF	NO-DIF
20	NO-DIF	NO-DIF	NO-DIF
21	NO-DIF	NO-DIF	NO-DIF
22	NO-DIF	NO-DIF	NO-DIF
23	NO-DIF	NO-DIF	NO-DIF
24	NO-DIF	NO-DIF	DIF
25	NO-DIF	NO-DIF	NO-DIF
26	NO-DIF	NO-DIF	NO-DIF
27	NO-DIF	NO-DIF	NO-DIF
28	NO-DIF	NO-DIF	NO-DIF

ตารางที่ 4-12 (ต่อ)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี IRT-LR		
	กลุ่มตัวอย่าง (1,000 คน)		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
29	NO-DIF	NO-DIF	NO-DIF
30	**DIF**	NO-DIF	NO-DIF
จำนวนข้อที่พบ DIF พบว่า	(4 ข้อ) 13%	(4 ข้อ) 13%	(5 ข้อ) 16%

หมายเหตุ DIF หมายถึง ข้อสอบที่ทำหน้าที่ต่างกัน
NO-DIF หมายถึง ข้อสอบที่ไม่พบว่าทำหน้าที่ต่างกัน

จากตารางที่ 4-12 แสดงผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านทางภาษา ด้านคำนวณ ด้านเหตุผล ด้วยวิธี IRT-LR ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่าง จำนวน 1,000 คน พบว่า

ด้านภาษา ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ จำนวน 4 ข้อ ได้แก่ ข้อที่ 5, 15, 18, 30 คิดเป็นร้อยละ 13 ของข้อสอบทั้งหมด

ด้านคำนวณ ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ จำนวน 4 ข้อ ได้แก่ ข้อที่ 10, 12, 16, 19 คิดเป็นร้อยละ 13 ของข้อสอบทั้งหมด

ด้านเหตุผล ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ จำนวน 5 ข้อ ได้แก่ ข้อที่ 1, 3, 11, 12, 24 คิดเป็นร้อยละ 16 ของข้อสอบทั้งหมด

ตารางที่ 4-13 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติระดับชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ ด้านเหตุผล ด้วยวิธี IRT-LR ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างขนาดใหญ่ (2,000 คน)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี IRT-LR		
	กลุ่มตัวอย่าง (2,000 คน)		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
1	NO-DIF	NO-DIF	**DIF**
2	NO-DIF	NO-DIF	NO-DIF
3	NO-DIF	NO-DIF	**DIF**
4	NO-DIF	**DIF**	NO-DIF
5	**DIF**	NO-DIF	NO-DIF
6	NO-DIF	NO-DIF	NO-DIF
7	NO-DIF	NO-DIF	NO-DIF
8	NO-DIF	NO-DIF	NO-DIF

ตารางที่ 4-13 (ต่อ)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี IRT-LR		
	กลุ่มตัวอย่าง (2,000 คน)		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
9	NO-DIF	NO-DIF	DIF
10	NO-DIF	NO-DIF	NO-DIF
11	NO-DIF	NO-DIF	DIF
12	NO-DIF	DIF	DIF
13	NO-DIF	NO-DIF	NO-DIF
14	NO-DIF	NO-DIF	DIF
15	NO-DIF	NO-DIF	NO-DIF
16	NO-DIF	DIF	NO-DIF
17	NO-DIF	NO-DIF	NO-DIF
18	NO-DIF	NO-DIF	NO-DIF
19	NO-DIF	NO-DIF	DIF
20	NO-DIF	NO-DIF	NO-DIF
21	NO-DIF	NO-DIF	DIF
22	NO-DIF	NO-DIF	NO-DIF
23	NO-DIF	NO-DIF	NO-DIF
24	NO-DIF	NO-DIF	NO-DIF
25	NO-DIF	NO-DIF	NO-DIF
26	NO-DIF	NO-DIF	NO-DIF
27	NO-DIF	NO-DIF	NO-DIF
28	NO-DIF	NO-DIF	NO-DIF
29	DIF	NO-DIF	DIF
30	DIF	NO-DIF	DIF
จำนวนข้อที่พบ DIF พบว่า	(3 ข้อ) 10%	(3 ข้อ) 10%	(10 ข้อ) 33%

หมายเหตุ DIF หมายถึง ข้อสอบที่ทำหน้าที่ต่างกัน
NO-DIF หมายถึง ข้อสอบที่ไม่พบว่าทำหน้าที่ต่างกัน

จากตารางที่ 4-13 แสดงผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ ด้านเหตุผล ด้วยวิธี IRT-LR ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่าง จำนวน 2,000 คน พบว่า

ด้านภาษา ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ จำนวน 3 ข้อ ได้แก่ ข้อที่ 5, 29, 30 คิดเป็นร้อยละ 10% ของข้อสอบทั้งหมด

ด้านคำนวณ ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ จำนวน 3 ข้อ ได้แก่ ข้อที่ 4, 12, 16 คิดเป็นร้อยละ 10% ของข้อสอบทั้งหมด

ด้านเหตุผล ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ จำนวน 10 ข้อ ได้แก่ ข้อที่ 1, 3, 9, 11, 12, 14, 19, 21, 29, 30 คิดเป็นร้อยละ 33% ของข้อสอบทั้งหมด

ตารางที่ 4-14 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ระดับชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ ด้านเหตุผล ด้วยวิธี SIBTEST ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างขนาดเล็ก (300 คน)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี SIBTEST กลุ่มตัวอย่าง (300 คน)		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
1	NO-DIF	NO-DIF	NO-DIF
2	NO-DIF	NO-DIF	NO-DIF
3	NO-DIF	NO-DIF	NO-DIF
4	NO-DIF	NO-DIF	NO-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	NO-DIF
9	NO-DIF	NO-DIF	NO-DIF
10	NO-DIF	NO-DIF	NO-DIF
11	NO-DIF	NO-DIF	NO-DIF
12	NO-DIF	DIF	NO-DIF
13	NO-DIF	NO-DIF	NO-DIF
14	NO-DIF	NO-DIF	NO-DIF
15	NO-DIF	NO-DIF	NO-DIF
16	NO-DIF	NO-DIF	NO-DIF
17	NO-DIF	NO-DIF	NO-DIF
18	NO-DIF	NO-DIF	DIF
19	DIF	NO-DIF	NO-DIF
20	NO-DIF	NO-DIF	DIF
21	NO-DIF	DIF	NO-DIF
22	NO-DIF	NO-DIF	NO-DIF
23	NO-DIF	NO-DIF	NO-DIF
24	NO-DIF	NO-DIF	NO-DIF

ตารางที่ 4-14 (ต่อ)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี SIBTEST		
	กลุ่มตัวอย่าง (300 คน)		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
25	NO-DIF	NO-DIF	NO-DIF
26	NO-DIF	**DIF**	NO-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	NO-DIF
จำนวนข้อที่พบ DIF พบว่า	(2 ข้อ) 7%	(3 ข้อ) 10%	(2 ข้อ) 7%

หมายเหตุ DIF หมายถึง ข้อสอบที่ทำหน้าที่ต่างกัน
NO-DIF หมายถึง ข้อสอบที่ไม่พบว่าทำหน้าที่ต่างกัน

จากตารางที่ 4-14 แสดงผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ ด้านเหตุผล ด้วยวิธี SIBTEST ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่าง จำนวน 300 คน พบว่า

ด้านภาษา ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ จำนวน 2 ข้อ ได้แก่ ข้อที่ 19, 30 คิดเป็นร้อยละ 7% ของข้อสอบทั้งหมด

ด้านคำนวณ ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ จำนวน 3 ข้อ ได้แก่ ข้อที่ 12, 21, 26 คิดเป็นร้อยละ 10% ของข้อสอบทั้งหมด

ด้านเหตุผล ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ จำนวน 2 ข้อ ได้แก่ ข้อที่ 18, 20 คิดเป็นร้อยละ 7% ของข้อสอบทั้งหมด

ตารางที่ 4-15 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ระดับชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ ด้านเหตุผล ด้วยวิธี SIBTEST ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างขนาดกลาง (1,000 คน)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี SIBTEST		
	กลุ่มตัวอย่าง (1,000 คน)		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
1	NO-DIF	NO-DIF	**DIF**
2	NO-DIF	NO-DIF	NO-DIF
3	NO-DIF	NO-DIF	NO-DIF
4	NO-DIF	NO-DIF	NO-DIF

ตารางที่ 4-15 (ต่อ)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี SIBTEST กลุ่มตัวอย่าง (1,000 คน)		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
5	DIF	NO-DIF	NO-DIF
6	NO-DIF	NO-DIF	NO-DIF
7	NO-DIF	NO-DIF	NO-DIF
8	NO-DIF	NO-DIF	NO-DIF
9	NO-DIF	NO-DIF	DIF
10	DIF	NO-DIF	NO-DIF
11	NO-DIF	NO-DIF	DIF
12	NO-DIF	NO-DIF	DIF
13	NO-DIF	NO-DIF	NO-DIF
14	NO-DIF	NO-DIF	NO-DIF
15	NO-DIF	NO-DIF	NO-DIF
16	NO-DIF	NO-DIF	NO-DIF
17	NO-DIF	NO-DIF	NO-DIF
18	NO-DIF	NO-DIF	NO-DIF
19	NO-DIF	DIF	DIF
20	NO-DIF	NO-DIF	DIF
21	NO-DIF	NO-DIF	NO-DIF
22	NO-DIF	NO-DIF	NO-DIF
23	NO-DIF	NO-DIF	NO-DIF
24	NO-DIF	NO-DIF	NO-DIF
25	NO-DIF	NO-DIF	NO-DIF
26	NO-DIF	NO-DIF	NO-DIF
27	NO-DIF	NO-DIF	NO-DIF
28	NO-DIF	NO-DIF	NO-DIF
29	NO-DIF	NO-DIF	NO-DIF
30	NO-DIF	NO-DIF	NO-DIF
จำนวนข้อที่พบ DIF พบว่า	(2 ข้อ) 7%	(1ข้อ) 3%	(6 ข้อ) 20%

หมายเหตุ DIF หมายถึง ข้อสอบที่ทำหน้าที่ต่างกัน

NO-DIF หมายถึง ข้อสอบที่ไม่พบว่าทำหน้าที่ต่างกัน

จากตารางที่ 4-15 แสดงผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ ด้านเหตุผล ด้วยวิธี SIBTEST ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่าง จำนวน 1,000 คน พบว่า

ด้านภาษา ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ จำนวน 2 ข้อ ได้แก่ ข้อที่ 5, 10 คิดเป็นร้อยละ 7% ของข้อสอบทั้งหมด

ด้านคำนวณ ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ จำนวน 1ข้อ ได้แก่ ข้อที่ 19 คิดเป็นร้อยละ 3% ของข้อสอบทั้งหมด

ด้านเหตุผล ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ จำนวน 6 ข้อ ได้แก่ ข้อที่ 1, 9, 11, 12, 19, 20 คิดเป็นร้อยละ 20% ของข้อสอบทั้งหมด

ตารางที่ 4-16 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ระดับชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ ด้านเหตุผล ด้วยวิธี SIBTEST ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่าง ขนาดใหญ่ (2,000 คน)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี SIBTEST		
	กลุ่มตัวอย่าง (2,000 คน)		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
1	NO-DIF	NO-DIF	DIF
2	NO-DIF	NO-DIF	NO-DIF
3	NO-DIF	NO-DIF	NO-DIF
4	NO-DIF	DIF	NO-DIF
5	DIF	NO-DIF	NO-DIF
6	NO-DIF	NO-DIF	NO-DIF
7	NO-DIF	NO-DIF	NO-DIF
8	NO-DIF	NO-DIF	NO-DIF
9	NO-DIF	DIF	DIF
10	NO-DIF	NO-DIF	NO-DIF
11	NO-DIF	NO-DIF	DIF
12	NO-DIF	NO-DIF	DIF
13	NO-DIF	DIF	NO-DIF
14	NO-DIF	NO-DIF	DIF
15	NO-DIF	NO-DIF	NO-DIF
16	NO-DIF	DIF	NO-DIF
17	NO-DIF	NO-DIF	NO-DIF
18	NO-DIF	NO-DIF	NO-DIF
19	NO-DIF	DIF	DIF

ตารางที่ 4-16 (ต่อ)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี SIBTEST		
	กลุ่มตัวอย่าง (2,000 คน)		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
20	NO-DIF	NO-DIF	DIF
21	NO-DIF	NO-DIF	DIF
22	NO-DIF	NO-DIF	NO-DIF
23	NO-DIF	NO-DIF	NO-DIF
24	DIF	NO-DIF	NO-DIF
25	NO-DIF	NO-DIF	DIF
26	NO-DIF	NO-DIF	DIF
27	NO-DIF	NO-DIF	NO-DIF
28	NO-DIF	DIF	NO-DIF
29	DIF	NO-DIF	DIF
30	DIF	NO-DIF	DIF
จำนวนข้อที่พบ DIF พบว่า	(4 ข้อ) 13%	(6 ข้อ) 20%	(12 ข้อ) 40%

หมายเหตุ DIF หมายถึง ข้อสอบที่ทำหน้าที่ต่างกัน
NO-DIF หมายถึง ข้อสอบที่ไม่พบว่าทำหน้าที่ต่างกัน

จากตารางที่ 4-16 แสดงผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ ด้านเหตุผล ด้วยวิธี SIBTEST ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่าง จำนวน 2,000 คน พบว่า

ด้านภาษา ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ จำนวน 4 ข้อ ได้แก่ ข้อที่ 5, 24, 29, 30 คิดเป็นร้อยละ 13% ของข้อสอบทั้งหมด

ด้านคำนวณ ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ จำนวน 6 ข้อ ได้แก่ ข้อที่ 4, 9, 13, 16, 19, 28 คิดเป็นร้อยละ 20% ของข้อสอบทั้งหมด

ด้านเหตุผล ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ จำนวน 12 ข้อ ได้แก่ ข้อที่ 1, 9, 11, 12, 14, 19, 20, 21, 25, 26, 29, 30 คิดเป็นร้อยละ 40% ของข้อสอบทั้งหมด

ตารางที่ 4-17 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ระดับชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ ด้านเหตุผล ด้วยวิธี MH ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างขนาดเล็ก (300 คน)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี MH กลุ่มตัวอย่าง (300 คน)		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
1	NO-DIF	NO-DIF	NO-DIF
2	DIF	NO-DIF	NO-DIF
3	DIF	NO-DIF	NO-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	NO-DIF
9	NO-DIF	NO-DIF	NO-DIF
10	NO-DIF	NO-DIF	NO-DIF
11	NO-DIF	NO-DIF	DIF
12	NO-DIF	DIF	DIF
13	NO-DIF	NO-DIF	DIF
14	NO-DIF	NO-DIF	NO-DIF
15	NO-DIF	NO-DIF	NO-DIF
16	NO-DIF	NO-DIF	NO-DIF
17	DIF	NO-DIF	NO-DIF
18	NO-DIF	NO-DIF	NO-DIF
19	DIF	NO-DIF	NO-DIF
20	NO-DIF	NO-DIF	NO-DIF
21	NO-DIF	DIF	NO-DIF
22	NO-DIF	NO-DIF	NO-DIF
23	NO-DIF	NO-DIF	NO-DIF
24	NO-DIF	NO-DIF	NO-DIF
25	NO-DIF	NO-DIF	NO-DIF
26	NO-DIF	NO-DIF	DIF
27	NO-DIF	NO-DIF	NO-DIF

ตารางที่ 4-17 (ต่อ)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี MH		
	กลุ่มตัวอย่าง (300 คน)		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
28	NO-DIF	NO-DIF	NO-DIF
29	NO-DIF	NO-DIF	NO-DIF
30	NO-DIF	NO-DIF	NO-DIF
จำนวนข้อที่พบ DIF พบว่า	(5 ข้อ) 17%	(2 ข้อ) 7%	(5 ข้อ) 17%

หมายเหตุ DIF หมายถึง ข้อสอบที่ทำหน้าที่ต่างกัน
NO-DIF หมายถึง ข้อสอบที่ไม่พบว่าทำหน้าที่ต่างกัน

จากตารางที่ 4-17 แสดงผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ ด้านเหตุผล ด้วยวิธี MH ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่าง จำนวน 300 คน พบว่า

ด้านภาษา ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ จำนวน 5 ข้อ ได้แก่ ข้อที่ 2, 3, 4, 17, 19 คิดเป็นร้อยละ 17% ของข้อสอบทั้งหมด

ด้านคำนวณ ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ จำนวน 2 ข้อ ได้แก่ ข้อที่ 12, 21 คิดเป็นร้อยละ 7% ของข้อสอบทั้งหมด

ด้านเหตุผล ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ จำนวน 5 ข้อ ได้แก่ ข้อที่ 4, 11, 12, 13, 26 คิดเป็นร้อยละ 17% ของข้อสอบทั้งหมด

ตารางที่ 4-18 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ระดับชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ ด้านเหตุผล ด้วยวิธี MH ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างขนาดกลาง (1,000 คน)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี MH		
	กลุ่มตัวอย่าง (1,000 คน)		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
1	NO-DIF	NO-DIF	NO-DIF
2	NO-DIF	NO-DIF	NO-DIF
3	**DIF**	NO-DIF	NO-DIF
4	**DIF**	NO-DIF	**DIF**
5	NO-DIF	NO-DIF	**DIF**
6	NO-DIF	NO-DIF	**DIF**
7	**DIF**	NO-DIF	NO-DIF

ตารางที่ 4-18 (ต่อ)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี MH		
	กลุ่มตัวอย่าง (1,000 คน)		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
8	NO-DIF	NO-DIF	NO-DIF
9	DIF	NO-DIF	NO-DIF
10	NO-DIF	NO-DIF	DIF
11	NO-DIF	NO-DIF	DIF
12	NO-DIF	NO-DIF	DIF
13	DIF	DIF	DIF
14	NO-DIF	NO-DIF	DIF
15	NO-DIF	NO-DIF	NO-DIF
16	DIF	DIF	NO-DIF
17	DIF	DIF	NO-DIF
18	DIF	NO-DIF	DIF
19	DIF	DIF	DIF
20	NO-DIF	NO-DIF	NO-DIF
21	NO-DIF	NO-DIF	NO-DIF
22	NO-DIF	NO-DIF	DIF
23	NO-DIF	NO-DIF	DIF
24	DIF	NO-DIF	DIF
25	NO-DIF	NO-DIF	NO-DIF
26	NO-DIF	NO-DIF	DIF
27	NO-DIF	NO-DIF	NO-DIF
28	NO-DIF	NO-DIF	DIF
29	DIF	NO-DIF	NO-DIF
30	NO-DIF	NO-DIF	NO-DIF
จำนวนข้อที่พบ DIF พบว่า	(11 ข้อ) 37%	(4 ข้อ) 13%	(15 ข้อ) 50%

หมายเหตุ DIF หมายถึง ข้อสอบที่ทำหน้าที่ต่างกัน

NO-DIF หมายถึง ข้อสอบที่ไม่พบว่าทำหน้าที่ต่างกัน

จากตารางที่ 4-18 แสดงผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ ด้านเหตุผล ด้วยวิธี MH ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่าง จำนวน 1,000 คน พบว่า

ด้านภาษา ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ จำนวน 11 ข้อ ได้แก่ ข้อที่ 3, 4, 7, 9, 13, 16, 17, 18, 19, 24, 29 คิดเป็นร้อยละ 37% ของข้อสอบทั้งหมด

ด้านคำนวณ ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ จำนวน 4ข้อ ได้แก่ ข้อที่ 13, 16, 17, 19 คิดเป็นร้อยละ 13% ของข้อสอบทั้งหมด

ด้านเหตุผล ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ จำนวน 15 ข้อ ได้แก่ ข้อที่ 4, 5, 6, 10, 11, 12, 13, 14, 18, 19, 22, 23, 24, 26, 28 คิดเป็นร้อยละ 50% ของข้อสอบทั้งหมด

ตารางที่ 4-19 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติระดับชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ ด้านเหตุผล ด้วยวิธี MH ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างขนาดใหญ่ (2,000 คน)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี MH กลุ่มตัวอย่าง (2,000 คน)		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
1	NO-DIF	NO-DIF	NO-DIF
2	NO-DIF	NO-DIF	DIF
3	DIF	NO-DIF	DIF
4	DIF	NO-DIF	DIF
5	DIF	NO-DIF	DIF
6	NO-DIF	NO-DIF	DIF
7	DIF	NO-DIF	DIF
8	NO-DIF	DIF	DIF
9	DIF	DIF	NO-DIF
10	NO-DIF	NO-DIF	DIF
11	DIF	NO-DIF	DIF
12	DIF	NO-DIF	DIF
13	NO-DIF	DIF	DIF
14	DIF	NO-DIF	DIF
15	DIF	NO-DIF	DIF
16	DIF	DIF	NO-DIF
17	DIF	DIF	DIF
18	DIF	NO-DIF	DIF
19	DIF	DIF	DIF

ตารางที่ 4-19 (ต่อ)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี MH		
	กลุ่มตัวอย่าง (2,000 คน)		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
20	DIF	DIF	NO-DIF
21	DIF	NO-DIF	NO-DIF
22	NO-DIF	NO-DIF	DIF
23	NO-DIF	NO-DIF	DIF
24	DIF	NO-DIF	DIF
25	NO-DIF	NO-DIF	NO-DIF
26	NO-DIF	NO-DIF	DIF
27	DIF	NO-DIF	DIF
28	DIF	DIF	DIF
29	DIF	NO-DIF	NO-DIF
30	NO-DIF	NO-DIF	NO-DIF
จำนวนข้อที่พบ DIF พบว่า	(19 ข้อ) 63%	(8 ข้อ) 27%	(22 ข้อ) 73%

หมายเหตุ DIF หมายถึง ข้อสอบที่ทำหน้าที่ต่างกัน
NO-DIF หมายถึง ข้อสอบที่ไม่พบว่าทำหน้าที่ต่างกัน

จากตารางที่ 4-19 แสดงผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ ด้านเหตุผล ด้วยวิธี MH ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่าง จำนวน 2,000 คน พบว่า

ด้านภาษา ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ จำนวน 19 ข้อ ได้แก่ ข้อที่ 3, 4, 5 , 7, 9, 11, 12, 14, 15, 16, 17, 18, 19, 21, 24, 27, 28, 29 คิดเป็นร้อยละ 63 ของข้อสอบทั้งหมด

ด้านคำนวณ ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ จำนวน 8 ข้อ ได้แก่ ข้อที่ 8, 9, 13, 16, 17, 19, 20, 28, คิดเป็นร้อยละ 27 ของข้อสอบทั้งหมด

ด้านเหตุผล ตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ จำนวน 22 ข้อ ได้แก่ ข้อที่ 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 17, 18, 19, 22, 23, 24, 26, 27, 28 คิดเป็นร้อยละ 73 ของข้อสอบทั้งหมด

ตารางที่ 4-20 ผลการวิเคราะห์คุณภาพของแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล

แบบทดสอบ	กลุ่มตัวอย่าง (คน)	จำนวนข้อสอบ (ข้อ)	วิธี IRT-LR		วิธี SIBTEST		วิธี MHl	
			จำนวนข้อที่พบ DIF	ร้อยละ	จำนวนข้อที่พบ DIF	ร้อยละ	จำนวนข้อที่พบ DIF	ร้อยละ
ด้านภาษา	300	30	3	10	2	7	5	17
	1,000	30	4	13	2	7	11	37
	2,000	30	3	10	4	13	19	63
ด้านคำนวณ	300	30	3	10	3	10	2	7
	1,000	30	4	13	1	3	4	13
	2,000	30	3	10	6	20	8	27
ด้านเหตุผล	300	30	3	10	2	7	5	17
	1,000	30	5	16	6	20	15	50
	2,000	30	10	33	12	40	22	73
รวมทั้ง 3 ด้าน		270	38	14	38	14	91	34

จากตารางที่ 4-20 แสดงผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ปรากฏว่า วิธี Mantel-Haenszel ตรวจพบ DIF จำนวน 91 ข้อ คิดเป็นร้อยละ 34 ของข้อสอบทั้งฉบับ รองลงมาคือ วิธี SIBTEST และวิธี IRT-LR ตรวจพบ DIF จำนวนข้อ 38 ข้อ เท่ากัน คิดเป็นร้อยละ 14

ตอนที่ 3 ผลการเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้วยวิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)

การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 คือ ด้านภาษา จำนวน 30 ข้อ ด้านคำนวณ จำนวน 30 ข้อ และด้านเหตุผล จำนวน 30 ข้อ โดยเปรียบเทียบระหว่างวิธี 2 วิธี วิธีใดตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบจำนวนมากกว่ากัน ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)

ตารางที่ 4-21 การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ด้านภาษา ระหว่างวิธี IRT-LR กับวิธี SIBTEST

ข้อที่	ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้านภาษา					
	วิธี IRT-LR			วิธี SIBTEST		
	300 คน	1,000 คน	2,000 คน	300 คน	1,000 คน	2,000 คน
1	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
2	DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-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	DIF	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	DIF	NO-DIF
11	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
12	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-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	DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
16	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
17	DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
18	NO-DIF	DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF

ตารางที่ 4-21 (ต่อ)

ข้อที่	ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้านภาษา					
	วิธี IRT-LR			วิธี SIBTEST		
	300 คน	1,000 คน	2,000 คน	300 คน	1,000 คน	2,000 คน
19	NO-DIF	NO-DIF	NO-DIF	DIF	NO-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	DIF
25	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
26	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-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	DIF	NO-DIF	NO-DIF	DIF
30	DIF	DIF	DIF	DIF	NO-DIF	DIF
จำนวนข้อที่พบDIF	(3 ข้อ)	(4 ข้อ)	(3 ข้อ)	(2 ข้อ)	(2 ข้อ)	(4 ข้อ)
	10%	13%	10%	7%	7%	13%

หมายเหตุ DIF หมายถึง ข้อสอบที่ทำหน้าที่ต่างกัน
NO-DIF หมายถึง ข้อสอบที่ไม่พบว่าทำหน้าที่ต่างกัน

จากตารางที่ 4-21 ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้านภาษา ระหว่างวิธี IRT-LR กับวิธี SIBTEST พบว่า เมื่อกลุ่มตัวอย่างขนาดเล็ก (300 คน) วิธี IRT-LR พบ DIF มากกว่า วิธี SIBTEST จำนวน 1 ข้อ คิดเป็นร้อยละ 3 ขนาดกลาง (1,000 คน) วิธี IRT-LR พบ DIF มากกว่า วิธี SIBTEST จำนวน 2 ข้อ คิดเป็นร้อยละ 7 ขนาดใหญ่ (2,000 คน) วิธี IRT-LR พบ DIF น้อยกว่า วิธี SIBTEST จำนวน 1 ข้อ คิดเป็นร้อยละ 3 ของข้อสอบทั้งหมด

ตารางที่ 4-22 การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ด้านคำนวณ ระหว่างวิธี IRT-LR กับ วิธี SIBTEST

ข้อที่	ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้านคำนวณ					
	วิธี IRT-LR			วิธี SIBTEST		
	300 คน	1,000 คน	2,000 คน	300 คน	1,000 คน	2,000 คน
1	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
2	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
3	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
4	NO-DIF	NO-DIF	DIF	NO-DIF	NO-DIF	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	DIF
10	NO-DIF	DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
11	DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
12	DIF	DIF	DIF	DIF	NO-DIF	NO-DIF
13	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	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	DIF	DIF	NO-DIF	NO-DIF	DIF
17	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
18	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
19	NO-DIF	DIF	NO-DIF	NO-DIF	DIF	DIF
20	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
21	DIF	NO-DIF	NO-DIF	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	NO-DIF	NO-DIF	NO-DIF	DIF	NO-DIF	NO-DIF
27	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF

ตารางที่ 4-22 (ต่อ)

ข้อที่	ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้านคำนวณ					
	วิธี IRT-LR			วิธี SIBTEST		
	300 คน	1,000 คน	2,000 คน	300 คน	1,000 คน	2,000 คน
28	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	**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
จำนวนข้อที่พบDIF	(3 ข้อ)	(4 ข้อ)	(3 ข้อ)	(3 ข้อ)	(1 ข้อ)	(6 ข้อ)
	10%	13%	10%	10%	3%	20%

หมายเหตุ DIF หมายถึง ข้อสอบที่ทำหน้าที่ต่างกัน
NO-DIF หมายถึง ข้อสอบที่ไม่พบว่าทำหน้าที่ต่างกัน

จากตารางที่ 4-22 ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้านคำนวณ ระหว่างวิธี IRT-LR กับวิธี SIBTEST พบว่า เมื่อกลุ่มตัวอย่างขนาดเล็ก (300 คน) วิธี IRT-LR พบ DIF เท่ากัน กับวิธี SIBTEST จำนวน 0 ข้อ คิดเป็นร้อยละ 0 ขนาดกลาง (1,000 คน) วิธี IRT-LR พบ DIF มากกว่า วิธี SIBTEST จำนวน 3 ข้อ คิดเป็นร้อยละ 10 ขนาดใหญ่ (2,000 คน) วิธี IRT-LR พบ DIF น้อยกว่า วิธี SIBTEST จำนวน 3 ข้อ คิดเป็นร้อยละ 10 ของข้อสอบทั้งหมด

ตารางที่ 4-23 การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ด้านเหตุผล ระหว่างวิธี IRT-LR กับวิธี SIBTEST

ข้อที่	ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้านเหตุผล					
	วิธี IRT-LR			วิธี SIBTEST		
	300 คน	1,000 คน	2,000 คน	300 คน	1,000 คน	2,000 คน
1	NO-DIF	**DIF**	NO-DIF	NO-DIF	**DIF**	**DIF**
2	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
3	**DIF**	**DIF**	**DIF**	NO-DIF	NO-DIF	NO-DIF
4	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-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	**DIF**	NO-DIF	**DIF**	**DIF**

ตารางที่ 4-23 (ต่อ)

ข้อที่	ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้านเหตุผล					
	วิธี IRT-LR			วิธี SIBTEST		
	300 คน	1,000 คน	2,000 คน	300 คน	1,000 คน	2,000 คน
10	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
11	NO-DIF	DIF	DIF	NO-DIF	DIF	DIF
12	DIF	DIF	DIF	NO-DIF	DIF	DIF
13	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
14	DIF	NO-DIF	DIF	NO-DIF	NO-DIF	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	NO-DIF	NO-DIF	DIF	NO-DIF	NO-DIF
19	NO-DIF	NO-DIF	DIF	NO-DIF	DIF	DIF
20	NO-DIF	NO-DIF	NO-DIF	DIF	DIF	DIF
21	NO-DIF	NO-DIF	DIF	NO-DIF	NO-DIF	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	DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
25	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF
26	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-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	DIF	NO-DIF	NO-DIF	DIF
30	NO-DIF	NO-DIF	DIF	NO-DIF	NO-DIF	DIF
จำนวนข้อที่พบDIF	(3 ข้อ)	(5 ข้อ)	(10 ข้อ)	(2 ข้อ)	(6 ข้อ)	(12 ข้อ)
	10%	16%	33%	7%	20%	40%

หมายเหตุ DIF หมายถึง ข้อสอบที่ทำหน้าที่ต่างกัน
NO-DIF หมายถึง ข้อสอบที่ไม่พบว่าทำหน้าที่ต่างกัน

จากตารางที่ 4-23 ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้านเหตุผล ระหว่างวิธี IRT-LR กับวิธี SIBTEST พบว่า เมื่อกลุ่มตัวอย่างขนาดเล็ก (300 คน) วิธี IRT-LR พบ DIF มากกว่าวิธี SIBTEST จำนวน 1 ข้อ คิดเป็นร้อยละ 3 ขนาดกลาง (1,000 คน) วิธี IRT-LR พบ DIF

น้อยกว่า วิธี SIBTEST จำนวน 1 ข้อ คิดเป็นร้อยละ 3 ขนาดใหญ่ (2,000 คน) วิธี IRT-LR พบ DIF น้อยกว่า วิธี SIBTEST จำนวน 2 ข้อ คิดเป็นร้อยละ 7 ของข้อสอบทั้งหมด

ตารางที่ 4-24 การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ด้านภาษา ระหว่างวิธี IRT-LR กับวิธี MH

ข้อที่	ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้านภาษา					
	วิธี IRT-LR			วิธี MH		
	300 คน	1,000 คน	2,000 คน	300 คน	1,000 คน	2,000 คน
1	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
2	DIF	NO-DIF	NO-DIF	DIF	NO-DIF	NO-DIF
3	NO-DIF	NO-DIF	NO-DIF	DIF	DIF	DIF
4	NO-DIF	NO-DIF	NO-DIF	DIF	DIF	DIF
5	NO-DIF	DIF	DIF	NO-DIF	NO-DIF	DIF
6	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
7	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF	DIF
8	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
9	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF	DIF
10	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
11	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF
12	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF
13	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF	NO-DIF
14	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF
15	NO-DIF	DIF	NO-DIF	NO-DIF	NO-DIF	DIF
16	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF	DIF
17	DIF	NO-DIF	NO-DIF	DIF	DIF	DIF
18	NO-DIF	DIF	NO-DIF	NO-DIF	DIF	DIF
19	NO-DIF	NO-DIF	NO-DIF	DIF	DIF	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	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	DIF	DIF
25	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
26	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF

ตารางที่ 4-24 (ต่อ)

ข้อที่	ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้านภาษา					
	วิธี IRT-LR			วิธี MH		
	300 คน	1,000 คน	2,000 คน	300 คน	1,000 คน	2,000 คน
27	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	**DIF**
28	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	**DIF**
29	NO-DIF	NO-DIF	**DIF**	NO-DIF	**DIF**	**DIF**
30	**DIF**	**DIF**	**DIF**	NO-DIF	NO-DIF	NO-DIF
จำนวนข้อที่พบDIF	(3 ข้อ)	(4 ข้อ)	(3 ข้อ)	(5 ข้อ)	(11 ข้อ)	(18 ข้อ)
	10%	13%	10%	17%	37%	60%

หมายเหตุ DIF หมายถึง ข้อสอบที่ทำหน้าที่ต่างกัน
NO-DIF หมายถึง ข้อสอบที่ไม่พบว่าทำหน้าที่ต่างกัน

จากตารางที่ 4-24 ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้านภาษา ระหว่างวิธี IRT-LR กับวิธี MH พบว่า เมื่อกลุ่มตัวอย่างขนาดเล็ก (300 คน) วิธี IRT-LR พบ DIF น้อยกว่าวิธี MH จำนวน 2 ข้อ คิดเป็นร้อยละ 7 ขนาดกลาง (1,000 คน) วิธี IRT-LR พบ DIF น้อยกว่า วิธี MH จำนวน 7 ข้อ คิดเป็นร้อยละ 23 ขนาดใหญ่ (2,000 คน) วิธี IRT-LR พบ DIF น้อยกว่า วิธี MH จำนวน 15 ข้อ คิดเป็นร้อยละ 50 ของข้อสอบทั้งหมด

ตารางที่ 4-25 การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ด้านคำนวณ ระหว่างวิธี IRT-LR กับวิธี MH

ข้อที่	ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้านคำนวณ					
	วิธี IRT-LR			วิธี MH		
	300 คน	1,000 คน	2,000 คน	300 คน	1,000 คน	2,000 คน
1	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
2	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
3	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
4	NO-DIF	NO-DIF	**DIF**	NO-DIF	NO-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	**DIF**

ตารางที่ 4-25 (ต่อ)

ข้อที่	ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้านคำนวณ					
	วิธี IRT-LR			วิธี MH		
	300 คน	1,000 คน	2,000 คน	300 คน	1,000 คน	2,000 คน
9	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF
10	NO-DIF	DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
11	DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
12	DIF	DIF	DIF	DIF	NO-DIF	NO-DIF
13	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF	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	DIF	DIF	NO-DIF	DIF	DIF
17	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF	DIF
18	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
19	NO-DIF	DIF	NO-DIF	NO-DIF	DIF	DIF
20	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF
21	DIF	NO-DIF	NO-DIF	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	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-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	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
จำนวนข้อที่พบDIF	(3 ข้อ)	(4 ข้อ)	(3 ข้อ)	(2 ข้อ)	(4 ข้อ)	(8 ข้อ)
	10%	13%	10%	7%	13%	27%

หมายเหตุ DIF หมายถึง ข้อสอบที่ทำหน้าที่ต่างกัน
NO-DIF หมายถึง ข้อสอบที่ไม่พบว่าทำหน้าที่ต่างกัน

จากตารางที่ 4-25 ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้านคำนวณ ระหว่างวิธี IRT-LR กับวิธี MH พบว่า เมื่อกลุ่มตัวอย่างขนาดเล็ก (300 คน) วิธี IRT-LR พบ DIF มากกว่าวิธี MH จำนวน 1 ข้อ คิดเป็นร้อยละ 3 ขนาดกลาง (1,000 คน) วิธี IRT-LR พบ DIF

เท่ากัน กับวิธี MH จำนวน 0 ข้อ คิดเป็นร้อยละ 0 ขนาดใหญ่ (2,000 คน) วิธี IRT-LR พบ DIF น้อยกว่า วิธี MH จำนวน 5 ข้อ คิดเป็นร้อยละ 17 ของข้อสอบทั้งหมด

ตารางที่ 4-26 การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ด้านเหตุผล ระหว่างวิธี IRT-LR กับวิธี MH

ข้อที่	ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้านเหตุผล					
	วิธี IRT-LR			วิธี MH		
	300 คน	1,000 คน	2,000 คน	300 คน	1,000 คน	2,000 คน
1	NO-DIF	DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
2	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF
3	DIF	DIF	DIF	NO-DIF	NO-DIF	DIF
4	NO-DIF	NO-DIF	NO-DIF	DIF	DIF	DIF
5	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF	DIF
6	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF	DIF
7	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF
8	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF
9	NO-DIF	NO-DIF	DIF	NO-DIF	NO-DIF	NO-DIF
10	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF	DIF
11	NO-DIF	DIF	DIF	DIF	DIF	DIF
12	DIF	DIF	DIF	DIF	DIF	DIF
13	NO-DIF	NO-DIF	NO-DIF	DIF	DIF	DIF
14	DIF	NO-DIF	DIF	NO-DIF	DIF	DIF
15	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	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	DIF
18	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF	DIF
19	NO-DIF	NO-DIF	DIF	NO-DIF	DIF	DIF
20	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
21	NO-DIF	NO-DIF	DIF	NO-DIF	NO-DIF	NO-DIF
22	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF	DIF
23	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF	DIF
24	NO-DIF	DIF	NO-DIF	NO-DIF	DIF	DIF
25	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
26	NO-DIF	NO-DIF	NO-DIF	DIF	DIF	DIF

ตารางที่ 4-26 (ต่อ)

ข้อที่	ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้านเหตุผล					
	วิธี IRT-LR			วิธี MH		
	300 คน	1,000 คน	2,000 คน	300 คน	1,000 คน	2,000 คน
27	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	**DIF**
28	NO-DIF	NO-DIF	NO-DIF	NO-DIF	**DIF**	**DIF**
29	NO-DIF	NO-DIF	**DIF**	NO-DIF	NO-DIF	NO-DIF
30	NO-DIF	NO-DIF	**DIF**	NO-DIF	NO-DIF	NO-DIF
จำนวนข้อที่พบDIF	(3 ข้อ)	(5 ข้อ)	(10 ข้อ)	(5 ข้อ)	(15 ข้อ)	(22 ข้อ)
	10%	16%	33%	17%	50%	73%

หมายเหตุ DIF หมายถึง ข้อสอบที่ทำหน้าที่ต่างกัน
NO-DIF หมายถึง ข้อสอบที่ไม่พบว่าทำหน้าที่ต่างกัน

จากตารางที่ 4-26 ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้านเหตุผล ระหว่างวิธี IRT-LR กับวิธี MH พบว่า เมื่อกลุ่มตัวอย่างขนาดเล็ก (300 คน) วิธี IRT-LR พบ DIF น้อยกว่าวิธี MH จำนวน 2 ข้อ คิดเป็นร้อยละ 7 ขนาดกลาง (1,000 คน) วิธี IRT-LR พบ DIF น้อยกว่า วิธี MH จำนวน 10 ข้อ คิดเป็นร้อยละ 33 ขนาดใหญ่ (2,000 คน) วิธี IRT-LR พบ DIF น้อยกว่า วิธี MH จำนวน 12 ข้อ คิดเป็นร้อยละ 40 ของข้อสอบทั้งหมด

ตารางที่ 4-27 การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติด้านด้านภาษา ระหว่างวิธี SIBTEST กับวิธี MH

ข้อที่	ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้านภาษา					
	วิธี SIBTEST			วิธี MH		
	300 คน	1,000 คน	2,000 คน	300 คน	1,000 คน	2,000 คน
1	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
2	NO-DIF	NO-DIF	NO-DIF	**DIF**	NO-DIF	NO-DIF
3	NO-DIF	NO-DIF	NO-DIF	**DIF**	**DIF**	**DIF**
4	NO-DIF	NO-DIF	NO-DIF	**DIF**	**DIF**	**DIF**
5	NO-DIF	**DIF**	**DIF**	NO-DIF	NO-DIF	**DIF**
6	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
7	NO-DIF	NO-DIF	NO-DIF	NO-DIF	**DIF**	**DIF**
8	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF

ตารางที่ 4-27 (ต่อ)

ข้อที่	ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้านภาษา					
	วิธี SIBTEST			วิธี MH		
	300 คน	1,000 คน	2,000 คน	300 คน	1,000 คน	2,000 คน
9	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF	DIF
10	NO-DIF	DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
11	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF
12	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF
13	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF	NO-DIF
14	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF
15	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF
16	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF	DIF
17	NO-DIF	NO-DIF	NO-DIF	DIF	DIF	DIF
18	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF	DIF
19	DIF	NO-DIF	NO-DIF	DIF	DIF	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	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	DIF	NO-DIF	DIF	DIF
25	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
26	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
27	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF
28	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF
29	NO-DIF	NO-DIF	DIF	NO-DIF	DIF	DIF
30	DIF	NO-DIF	DIF	NO-DIF	NO-DIF	NO-DIF
จำนวนข้อที่พบDIF	(2 ข้อ)	(2 ข้อ)	(4 ข้อ)	(5 ข้อ)	(11 ข้อ)	(18 ข้อ)
	7%	7%	13%	17%	37%	60%

หมายเหตุ DIF หมายถึง ข้อสอบที่ทำหน้าที่ต่างกัน
NO-DIF หมายถึง ข้อสอบที่ไม่พบว่าทำหน้าที่ต่างกัน

จากตารางที่ 4-27 ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้านภาษา ระหว่างวิธี SIBTEST กับวิธี MH พบว่า เมื่อกลุ่มตัวอย่างขนาดเล็ก (300 คน) วิธี SIBTEST พบ DIF น้อยกว่าวิธี MH จำนวน 3 ข้อ คิดเป็นร้อยละ 10 ขนาดกลาง (1,000 คน) วิธี SIBTEST พบ DIF

น้อยกว่า วิธี MH จำนวน 9 ข้อ คิดเป็นร้อยละ 30 ขนาดใหญ่ (2,000 คน) วิธี SIBTEST พบ DIF น้อยกว่า วิธี MH จำนวน 14 ข้อ คิดเป็นร้อยละ 47 ของข้อสอบทั้งหมด

ตารางที่ 4-28 การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ด้านคำนวณ ระหว่างวิธี SIBTEST กับวิธี MH

ข้อที่	ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้านคำนวณ					
	วิธี SIBTEST			วิธี MH		
	300 คน	1,000 คน	2,000 คน	300 คน	1,000 คน	2,000 คน
1	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
2	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
3	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
4	NO-DIF	NO-DIF	DIF	NO-DIF	NO-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	DIF
9	NO-DIF	NO-DIF	DIF	NO-DIF	NO-DIF	DIF
10	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
11	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
12	DIF	NO-DIF	NO-DIF	DIF	NO-DIF	NO-DIF
13	NO-DIF	NO-DIF	DIF	NO-DIF	DIF	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	DIF	NO-DIF	DIF	DIF
17	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF	DIF
18	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
19	NO-DIF	DIF	DIF	NO-DIF	DIF	DIF
20	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF
21	DIF	NO-DIF	NO-DIF	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	NO-DIF	NO-DIF

ตารางที่ 4-28 (ต่อ)

ข้อที่	ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้านคำนวณ					
	วิธี SIBTEST			วิธี MH		
	300 คน	1,000 คน	2,000 คน	300 คน	1,000 คน	2,000 คน
27	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF
28	NO-DIF	NO-DIF	DIF	NO-DIF	NO-DIF	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
จำนวนข้อที่พบDIF	(3 ข้อ)	(1 ข้อ)	(6 ข้อ)	(2 ข้อ)	(4 ข้อ)	(8 ข้อ)
	10%	3%	20%	7%	13%	27%

หมายเหตุ DIF หมายถึง ข้อสอบที่ทำหน้าที่ต่างกัน
NO-DIF หมายถึง ข้อสอบที่ไม่พบว่าทำหน้าที่ต่างกัน

จากตารางที่ 4-28 ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้านคำนวณ ระหว่างวิธี SIBTEST กับวิธี MH พบว่า เมื่อกลุ่มตัวอย่างขนาดเล็ก (300 คน) วิธี SIBTEST พบ DIF มากกว่าวิธี MH จำนวน 1 ข้อ คิดเป็นร้อยละ 3 ขนาดกลาง (1,000 คน) วิธี SIBTEST พบ DIF น้อยกว่า วิธี MH จำนวน 3 ข้อ คิดเป็นร้อยละ 10 ขนาดใหญ่ (2,000 คน) วิธี SIBTEST พบ DIF น้อยกว่า วิธี MH จำนวน 2 ข้อ คิดเป็นร้อยละ 7 ของข้อสอบทั้งหมด

ตารางที่ 4-29 การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบ ระดับชาติ ด้านเหตุผล ระหว่างวิธี SIBTEST กับวิธี MH

ข้อที่	ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้านเหตุผล					
	วิธี SIBTEST			วิธี MH		
	300 คน	1,000 คน	2,000 คน	300 คน	1,000 คน	2,000 คน
1	NO-DIF	DIF	DIF	NO-DIF	NO-DIF	NO-DIF
2	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF
3	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF
4	NO-DIF	NO-DIF	NO-DIF	DIF	DIF	DIF
5	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF	DIF
6	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF	DIF
7	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF
8	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF

ตารางที่ 4-29 (ต่อ)

ข้อที่	ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้านเหตุผล					
	วิธี SIBTEST			วิธี MH		
	300 คน	1,000 คน	2,000 คน	300 คน	1,000 คน	2,000 คน
9	NO-DIF	DIF	DIF	NO-DIF	NO-DIF	NO-DIF
10	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF	DIF
11	NO-DIF	DIF	DIF	DIF	DIF	DIF
12	NO-DIF	DIF	DIF	DIF	DIF	DIF
13	NO-DIF	NO-DIF	NO-DIF	DIF	DIF	DIF
14	NO-DIF	NO-DIF	DIF	NO-DIF	DIF	DIF
15	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	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	DIF
18	DIF	NO-DIF	NO-DIF	NO-DIF	DIF	DIF
19	NO-DIF	DIF	DIF	NO-DIF	DIF	DIF
20	DIF	DIF	DIF	NO-DIF	NO-DIF	NO-DIF
21	NO-DIF	NO-DIF	DIF	NO-DIF	NO-DIF	NO-DIF
22	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF	DIF
23	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF	DIF
24	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF	DIF
25	NO-DIF	NO-DIF	DIF	NO-DIF	NO-DIF	NO-DIF
26	NO-DIF	NO-DIF	DIF	DIF	DIF	DIF
27	NO-DIF	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF
28	NO-DIF	NO-DIF	NO-DIF	NO-DIF	DIF	DIF
29	NO-DIF	NO-DIF	DIF	NO-DIF	NO-DIF	NO-DIF
30	NO-DIF	NO-DIF	DIF	NO-DIF	NO-DIF	NO-DIF
จำนวนข้อที่พบDIF	(2 ข้อ)	(6 ข้อ)	(12 ข้อ)	(5 ข้อ)	(15 ข้อ)	(22 ข้อ)
	7%	20%	40%	17%	50%	73%

หมายเหตุ DIF หมายถึง ข้อสอบที่ทำหน้าที่ต่างกัน
NO-DIF หมายถึง ข้อสอบที่ไม่พบว่าทำหน้าที่ต่างกัน

จากตารางที่ 4-29 ผลการเปรียบเทียบการทำหน้าที่ต่างกันของข้อสอบ ด้านเหตุผล ระหว่างวิธี SIBTEST กับวิธี MH พบว่า เมื่อกลุ่มตัวอย่างขนาดเล็ก (300 คน) วิธี SIBTEST พบ DIF น้อยกว่าวิธี MH จำนวน 3 ข้อ คิดเป็นร้อยละ 10 ขนาดกลาง (1,000 คน) วิธี SIBTEST พบ DIF

น้อยกว่า วิธี MH จำนวน 9 ข้อ คิดเป็นร้อยละ 30 ขนาดใหญ่ (2,000 คน) วิธี SIBTEST พบ DIF น้อยกว่า วิธี MH จำนวน 10 ข้อ คิดเป็นร้อยละ 33 ของข้อสอบทั้งหมด

ตารางที่ 4-30 การเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วย วิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)

		เปรียบเทียบร้อยละของการตรวจพบ DIF		
แบบทดสอบ NT	กลุ่มตัวอย่าง (คน)	วิธี IRT-LR กับ วิธี SIBTEST	วิธี IRT-LR กับ วิธี MH	วิธี SIBTEST กับ วิธี MH
ด้านภาษา	300	วิธี IRT-LR > วิธี SIBTEST* (1 ข้อ ร้อยละ 3)	วิธี IRT-LR < วิธี MH* (2 ข้อ ร้อยละ 7)	วิธี SIBTEST < วิธี MH* (3 ข้อ ร้อยละ 10)
	1,000	วิธี IRT-LR > วิธี SIBTEST* (2 ข้อ ร้อยละ 7)	วิธี IRT-LR < วิธี MH* (7 ข้อ ร้อยละ 23)	วิธี SIBTEST < วิธี MH* (9 ข้อ ร้อยละ 30)
	2,000	วิธี IRT-LR < วิธี SIBTEST* (1 ข้อ ร้อยละ 3)	วิธี IRT-LR < วิธี MH* (16 ข้อ ร้อยละ 53)	วิธี SIBTEST < วิธี MH* (15 ข้อ ร้อยละ 50)
ด้านคำนวณ	300	วิธี IRT-LR = วิธี SIBTEST * (0 ข้อ ร้อยละ 0)	วิธี IRT-LR > วิธี MH* (1 ข้อ ร้อยละ 3)	วิธี SIBTEST > วิธี MH* (1 ข้อ ร้อยละ 3)
	1,000	วิธี IRT-LR > วิธี SIBTEST* (3 ข้อ ร้อยละ 10)	วิธี IRT-LR = วิธี MH* (0 ข้อ ร้อยละ 0)	วิธี SIBTEST < วิธี MH* (3 ข้อ ร้อยละ 10)
	2,000	วิธี IRT-LR < วิธี SIBTEST* (3 ข้อ ร้อยละ 10)	วิธี IRT-LR < วิธี MH* (5 ข้อ ร้อยละ 17)	วิธี SIBTEST < วิธี MH* (2 ข้อ ร้อยละ 7)
ด้านเหตุผล	300	วิธี IRT-LR > วิธี SIBTEST* (1 ข้อ ร้อยละ 3)	วิธี IRT-LR < วิธี MH* (2 ข้อ ร้อยละ 7)	วิธี SIBTEST < วิธี MH* (3 ข้อ ร้อยละ 10)
	1,000	วิธี IRT-LR < วิธี SIBTEST* (1 ข้อ ร้อยละ 3)	วิธี IRT-LR < วิธี MH* (10 ข้อ ร้อยละ 33)	วิธี SIBTEST < วิธี MH* (9 ข้อ ร้อยละ 30)
	2,000	วิธี IRT-LR < วิธี SIBTEST* (2 ข้อ ร้อยละ 7)	วิธี IRT-LR < วิธี MH* (12 ข้อ ร้อยละ 40)	วิธี SIBTEST < วิธี MH* (10 ข้อ ร้อยละ 33)

หมายเหตุ * $p < .05$

วิธี IRT-LR > วิธี SIBTEST หมายถึง วิธี IRT-LR ตรวจพบ DIF มากกว่า วิธี SIBTEST
วิธี IRT-LR < วิธี SIBTEST หมายถึง วิธี IRT-LR ตรวจพบ DIF น้อยกว่า วิธี SIBTEST
วิธี IRT-LR = วิธี SIBTEST หมายถึง วิธี IRT-LR ตรวจพบ DIF เท่ากับ วิธี SIBTEST
วิธี IRT-LR > วิธี MH หมายถึง วิธี IRT-LR ตรวจพบ DIF มากกว่า วิธี MH
วิธี IRT-LR < วิธี MH หมายถึง วิธี IRT-LR ตรวจพบ DIF น้อยกว่า วิธี MH
วิธี IRT-LR = วิธี MH หมายถึง วิธี IRT-LR ตรวจพบ DIF เท่ากับ วิธี MH
วิธี SIBTEST >วิธี MH หมายถึง วิธี SIBTEST ตรวจพบ DIF มากกว่า วิธี MH
วิธี SIBTEST< วิธี MH หมายถึง วิธี SIBTEST ตรวจพบ DIF น้อยกว่า วิธี MH

จากตารางที่ 4-30 ผลการเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) พบว่า

ด้านภาษา จำนวน 300 คน วิธี IRT-LR ตรวจพบข้อสอบทำหน้าที่ต่างกันของข้อสอบมากกว่า วิธี SIBTEST ร้อยละ 3 ของข้อสอบทั้งหมด วิธี IRT-LR ตรวจพบข้อสอบทำหน้าที่ต่างกันของข้อสอบน้อยกว่า วิธี MH ร้อยละ 7 ของข้อสอบทั้งหมด วิธี SIBTEST ตรวจพบข้อสอบทำหน้าที่ต่างกันของข้อสอบน้อยกว่า วิธี MH ร้อยละ 10 ของข้อสอบทั้งหมด

ด้านภาษา จำนวน 1,000 คน วิธี IRT-LR ตรวจพบข้อสอบทำหน้าที่ต่างกันของข้อสอบมากกว่า วิธี SIBTEST ร้อยละ 7 ของข้อสอบทั้งหมด วิธี IRT-LR ตรวจพบข้อสอบทำหน้าที่ต่างกันของข้อสอบน้อยกว่า วิธี MH ร้อยละ 23 ของข้อสอบทั้งหมด วิธี SIBTEST ตรวจพบข้อสอบทำหน้าที่ต่างกันของข้อสอบน้อยกว่า วิธี MH ร้อยละ 30 ของข้อสอบทั้งหมด

ด้านภาษา จำนวน 2,000 คน วิธี IRT-LR ตรวจพบข้อสอบทำหน้าที่ต่างกันของข้อสอบน้อยกว่า วิธี SIBTEST ร้อยละ 3 ของข้อสอบทั้งหมด วิธี IRT-LR ตรวจพบข้อสอบทำหน้าที่ต่างกันของข้อสอบน้อยกว่า วิธี MH ร้อยละ 53 ของข้อสอบทั้งหมด วิธี SIBTEST ตรวจพบข้อสอบทำหน้าที่ต่างกันของข้อสอบน้อยกว่า วิธี MH ร้อยละ 50 ของข้อสอบทั้งหมด

ด้านคำนวณ จำนวน 300 คน วิธี IRT-LR ตรวจพบข้อสอบทำหน้าที่ต่างกันของข้อสอบเท่ากับ วิธี SIBTEST ของข้อสอบทั้งหมด วิธี IRT-LR ตรวจพบข้อสอบทำหน้าที่ต่างกันของข้อสอบมากกว่า วิธี MH ร้อยละ 3 ของข้อสอบทั้งหมด วิธี SIBTEST ตรวจพบข้อสอบทำหน้าที่ต่างกันของข้อสอบมากกว่า วิธี MH ร้อยละ 3 ของข้อสอบทั้งหมด

ด้านคำนวณ จำนวน 1,000 คน วิธี IRT-LR ตรวจพบข้อสอบทำหน้าที่ต่างกันของข้อสอบมากกว่า วิธี SIBTEST ร้อยละ 10 ของข้อสอบทั้งหมด วิธี IRT-LR ตรวจพบข้อสอบทำหน้าที่ต่างกันของข้อสอบเท่ากับ วิธี MH ของข้อสอบทั้งหมด วิธี SIBTEST ตรวจพบข้อสอบทำหน้าที่ต่างกันของข้อสอบน้อยกว่า วิธี MH ร้อยละ 10 ของข้อสอบทั้งหมด

ด้านคำนวณ จำนวน 2,000 คน วิธี IRT-LR ตรวจพบข้อสอบทำหน้าที่ต่างกันของข้อสอบน้อยกว่า วิธี SIBTEST ร้อยละ 10 ของข้อสอบทั้งหมด วิธี IRT-LR ตรวจพบข้อสอบทำหน้าที่ต่างกันของข้อสอบน้อยกว่า วิธี MH ร้อยละ 17 ของข้อสอบทั้งหมด วิธี SIBTEST ตรวจพบข้อสอบทำหน้าที่ต่างกันของข้อสอบน้อยกว่า วิธี MH ร้อยละ 7 ของข้อสอบทั้งหมด

ด้านเหตุผล จำนวน 300 คน วิธี IRT-LR ตรวจพบข้อสอบทำหน้าที่ต่างกันของข้อสอบมากกว่า วิธี SIBTEST ร้อยละ 3 ของข้อสอบทั้งหมด วิธี IRT-LR ตรวจพบข้อสอบทำหน้าที่ต่างกันของข้อสอบน้อยกว่า วิธี MH ร้อยละ 7 ของข้อสอบทั้งหมด วิธี SIBTEST ตรวจพบข้อสอบทำหน้าที่ต่างกันของข้อสอบน้อยกว่า วิธี MH ร้อยละ 10 ของข้อสอบทั้งหมด

ด้านเหตุผล จำนวน 1,000 คน วิธี IRT-LR ตรวจพบข้อสอบทำหน้าที่ต่างกันของข้อสอบน้อยกว่า วิธี SIBTEST ร้อยละ 3 ของข้อสอบทั้งหมด วิธี IRT-LR ตรวจพบข้อสอบทำหน้าที่ต่างกันของข้อสอบน้อยกว่า วิธี MH ร้อยละ 33 ของข้อสอบทั้งหมด วิธี SIBTEST ตรวจพบข้อสอบทำหน้าที่ต่างกันของข้อสอบน้อยกว่า วิธี MH ร้อยละ 30 ของข้อสอบทั้งหมด

ด้านเหตุผล จำนวน 2,000 คน วิธี IRT-LR ตรวจพบข้อสอบทำหน้าที่ต่างกันของข้อสอบน้อยกว่า วิธี SIBTEST ร้อยละ 7 ของข้อสอบทั้งหมด วิธี IRT-LR ตรวจพบข้อสอบทำหน้าที่ต่างกันของข้อสอบน้อยกว่า วิธี MH ร้อยละ 40 ของข้อสอบทั้งหมด วิธี SIBTEST ตรวจพบข้อสอบทำหน้าที่ต่างกันของข้อสอบน้อยกว่า วิธี MH ร้อยละ 33 ของข้อสอบทั้งหมด

สรุปผลการเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) ด้วยวิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล พบว่าด้านภาษา วิธีการทดสอบอัตราส่วนไลค์ลิฮูด พบ DIF มากกว่า วิธีซิปเทสท์ ร้อยละ 7 ของข้อสอบทั้งหมด วิธีการทดสอบอัตราส่วนไลค์ลิฮูด พบ DIF น้อยกว่า วิธีแมนเทล- แฮนส์เซล ร้อยละ 80 ของข้อสอบทั้งหมด และ วิธีซิปเทสท์ พบ DIF น้อยกว่า วิธีแมนเทล-แฮนส์เซล ร้อยละ 87 ของข้อสอบทั้งหมด ส่วนด้านคำนวณ วิธีอัตราส่วนไลค์ลิฮูด พบ DIF เท่ากันกับ วิธีซิปเทสท์ ร้อยละ 0 ของข้อสอบทั้งหมด วิธีอัตราส่วนไลค์ลิฮูด พบ DIF น้อยกว่าวิธีแมนเทล-แฮนส์เซล ร้อยละ 13 ของข้อสอบทั้งหมด และวิธีซิปเทสท์ พบ DIF น้อยกว่าวิธีแมนเทล-แฮนส์เซล ร้อยละ 13 ของข้อสอบทั้งหมด และส่วนด้านเหตุผล วิธีอัตราส่วนไลค์ลิฮูด พบ DIF น้อยกว่าวิธีซิปเทสท์ ร้อยละ 7 ของข้อสอบทั้งหมด วิธีอัตราส่วนไลค์ลิฮูด พบ DIF น้อยกว่าวิธีแมนเทล-แฮนส์เซล ร้อยละ 80 ของข้อสอบทั้งหมด และวิธีซิปเทสท์ พบ DIF น้อยกว่าวิธีแมนเทล-แฮนส์เซล ร้อยละ 73 ของข้อสอบทั้งหมดอย่างมีนัยสำคัญทางสถิติที่ .05

บทที่ 5
สรุปและอภิปรายผล

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อวิเคราะห์คุณภาพของแบบทดสอบระดับชาติ ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ และเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ของนักเรียนชั้นประถมศึกษาปีที่ 3 จำนวน 3 ด้าน คือ ด้านภาษา ด้านคำนวณ และด้านเหตุผล โดยมีวิธีการดำเนินการวิจัยเป็น 3 ระยะ ดังนี้ ระยะที่ 1 การวิเคราะห์คุณภาพของแบบทดสอบระดับชาติ โดยใช้ทฤษฎีการตอบสนองข้อสอบ (IRT) แบบ 3 พารามิเตอร์ ระยะที่ 2 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธีการตรวจสอบ 3 วิธี คือ วิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิเคราะห์ด้วยโปรแกรมสำเร็จรูป IRTPRO วิธีซิปเทสท์ วิเคราะห์ด้วยโปรแกรมสำเร็จรูป SIBTEST และวิธีแมนเทล-แฮนส์เซล วิเคราะห์ด้วยโปรแกรมสำเร็จรูป SPSS และระยะที่ 3 การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติด้วยวิธีการตรวจสอบทั้ง 3 วิธี ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)

สรุปผลการวิจัย

1. ผลการวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และ ขนาดใหญ่ (2,000 คน) โดยการวิเคราะห์ตามทฤษฎีการตอบสนองข้อสอบแบบ 3 พารามิเตอร์ พบว่า กลุ่มตัวอย่างขนาดเล็ก (300 คน) ด้านภาษา มีค่าอำนาจจำแนกของข้อสอบ (a) เฉลี่ยเท่ากับ 0.614 ค่าความยากของข้อสอบ (b) เฉลี่ยเท่ากับ 0.760 และค่าการเดาของข้อสอบ (c) เฉลี่ยเท่ากับ 0.255 สรุปได้ว่าแบบทดสอบด้านภาษา กลุ่มตัวอย่างขนาดเล็ก (300 คน) มีค่าอำนาจจำแนกข้อสอบอยู่ในระดับดีมาก มีค่าความยากของข้อสอบอยู่ในระดับข้อสอบค่อนข้างยาก และมีค่าการเดาของข้อสอบไม่เกิน 0.3 สำหรับ ด้านคำนวณ มีค่าอำนาจจำแนกของข้อสอบ (a) เฉลี่ยเท่ากับ 0.776 ค่าความยากของข้อสอบ (b) เฉลี่ยเท่ากับ 1.476 และมีค่าการเดาของข้อสอบ (c) เฉลี่ยเท่ากับ 0.235 สรุปได้ว่าแบบทดสอบด้านคำนวณ กลุ่มตัวอย่างขนาดเล็ก (300 คน) มีค่าอำนาจจำแนกข้อสอบอยู่ในระดับดีมาก มีค่าความยากของข้อสอบอยู่ในระดับข้อสอบยาก และมีค่าการเดาของข้อสอบไม่เกิน 0.3 และด้านเหตุผล มีค่าอำนาจจำแนกของข้อสอบ (a) เฉลี่ยเท่ากับ 0.762 ค่าความยากของข้อสอบ (b) เฉลี่ยเท่ากับ 1.004 และมีค่าการเดาของข้อสอบ (c) เฉลี่ยเท่ากับ 0.254 สรุปได้ว่าแบบทดสอบด้านเหตุผล กลุ่มตัวอย่างขนาดเล็ก (300 คน) มีค่าอำนาจจำแนกข้อสอบอยู่ในระดับดีมาก มีค่าความยากของข้อสอบอยู่ในระดับข้อสอบยาก และมีค่าการเดาของข้อสอบไม่เกิน 0.3

กลุ่มตัวอย่างขนาดกลาง (1,000 คน) ด้านภาษา มีค่าอำนาจจำแนกของข้อสอบ (a) เฉลี่ยเท่ากับ 0.644 ค่าความยากของข้อสอบ (b) เฉลี่ยเท่ากับ 0.733 และมีค่าการเดาของข้อสอบ (c) เฉลี่ยเท่ากับ 0.250 สรุปได้ว่าแบบทดสอบด้านภาษา กลุ่มตัวอย่างขนาดกลาง (1,000 คน)

มีค่าอำนาจจำแนกข้อสอบอยู่ในระดับดีมากมีค่าความยากของข้อสอบอยู่ในระดับข้อสอบค่อนข้างยาก และ มีค่าการเดาของข้อสอบไม่เกิน 0.3 สำหรับด้านคำนวณ มีค่าอำนาจจำแนกของข้อสอบ (a) เฉลี่ยเท่ากับ 0.867 ค่าความยากของข้อสอบ (b) เฉลี่ยเท่ากับ 1.297 และมีค่าการเดาของข้อสอบ (c) เฉลี่ยเท่ากับ 0.231 สรุปได้ว่าแบบทดสอบด้านคำนวณ กลุ่มตัวอย่างขนาดกลาง (1,000 คน) มีค่าอำนาจจำแนกข้อสอบอยู่ในระดับดีมาก มีค่าความยากของข้อสอบอยู่ในระดับข้อสอบยาก และมีค่าการเดาของข้อสอบไม่เกิน 0.3 และด้านเหตุผล มีค่าอำนาจจำแนกของข้อสอบ (a) เฉลี่ยเท่ากับ 0.762 ค่าความยากของข้อสอบ (b) เฉลี่ยเท่ากับ 0.856 และมีค่าการเดาของข้อสอบ (c) เฉลี่ยเท่ากับ 0.246 สรุปได้ว่าแบบทดสอบ ด้านเหตุผล กลุ่มตัวอย่างขนาดกลาง (1,000 คน) มีค่าอำนาจจำแนกข้อสอบอยู่ในระดับดีมาก มีค่าความยากของข้อสอบอยู่ในระดับข้อสอบค่อนข้างยาก และมีค่าการเดาของข้อสอบไม่เกิน 0.3

กลุ่มตัวอย่างขนาดใหญ่ (2,000 คน) ด้านภาษา มีค่าอำนาจจำแนกของข้อสอบ (a) เฉลี่ยเท่ากับ 0.643 ค่าความยากของข้อสอบ (b) เฉลี่ยเท่ากับ 0.476 และมีค่าการเดาของข้อสอบ (c) เฉลี่ยเท่ากับ 0.211 สรุปได้ว่าแบบทดสอบ ด้านภาษา กลุ่มตัวอย่างขนาดใหญ่ (2,000 คน) มีอำนาจจำแนกข้อสอบอยู่ในระดับดีมาก มีค่าความยากของข้อสอบอยู่ในระดับข้อสอบปานกลาง และมีค่าการเดาของข้อสอบไม่เกิน 0.3 สำหรับ ด้านคำนวณ มีค่าอำนาจจำแนกของข้อสอบ (a) เฉลี่ยเท่ากับ 0.836 ค่าความยากของข้อสอบ (b) มีค่าเฉลี่ยเท่ากับ 1.465 และมีค่าการเดาของข้อสอบ (c) มีค่าเฉลี่ยเท่ากับ 0.228 สรุปได้ว่าแบบทดสอบ ด้านคำนวณ กลุ่มตัวอย่างขนาดใหญ่ (2,000 คน) มีอำนาจจำแนกข้อสอบอยู่ในระดับดีมาก มีค่าความยากของข้อสอบอยู่ในระดับข้อสอบยาก และมีค่าการเดาของข้อสอบไม่เกิน 0.3 และ ด้านเหตุผล มีค่าอำนาจจำแนกของข้อสอบ (a) เฉลี่ยเท่ากับ 0.742 ค่าความยากของข้อสอบ (b) เฉลี่ยเท่ากับ 0.851 และมีค่าการเดาของข้อสอบ (c) มีค่าเฉลี่ยเท่ากับ 0.239 สรุปได้ว่าแบบทดสอบ ด้านเหตุผล กลุ่มตัวอย่างขนาดใหญ่ (2,000 คน) มีค่าอำนาจจำแนกข้อสอบอยู่ในระดับดีมาก มีค่าความยากของข้อสอบอยู่ในระดับข้อสอบค่อนข้างยาก และมีค่าการเดาของข้อสอบไม่เกิน 0.3

2. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และ ด้านเหตุผล ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) พบว่า วิธีการทดสอบอัตราส่วนไลค์ลิฮูด ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน (DIF) กลุ่มตัวอย่างขนาดเล็ก (300 คน) ด้านภาษา จำนวน 3 ข้อ คิดเป็นร้อยละ 10 สำหรับ ด้านคำนวณ พบ DIF จำนวน 3 ข้อ คิดเป็นร้อยละ 10 และด้านเหตุผล พบ DIF จำนวน 3 ข้อ คิดเป็นร้อยละ 10 ส่วนกลุ่มตัวอย่างขนาดกลาง (1,000 คน) ด้านภาษา จำนวน 4 ข้อ คิดเป็นร้อยละ 13 สำหรับ ด้านคำนวณ พบ DIF จำนวน 4 ข้อ คิดเป็นร้อยละ 13 และ ด้านเหตุผล พบ DIF จำนวน 5 ข้อ คิดเป็นร้อยละ 16 และกลุ่มตัวอย่างขนาดใหญ่ (2,000 คน) ด้านภาษา จำนวน 3 ข้อ คิดเป็นร้อยละ 10 สำหรับ ด้านคำนวณ พบ DIF จำนวน 3 ข้อ คิดเป็นร้อยละ 10 และด้านเหตุผล พบ DIF จำนวน 10 ข้อ คิดเป็นร้อยละ 33 ส่วนวิธีชิปเทสท์ ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน (DIF) กลุ่มตัวอย่างขนาดเล็ก (300 คน) ด้านภาษา จำนวน 2 ข้อ คิดเป็นร้อยละ 7 ด้านคำนวณ พบ DIF จำนวน 3 ข้อ คิดเป็นร้อยละ 10 และด้านเหตุผล พบ DIF จำนวน 2 ข้อ คิดเป็นร้อยละ 7 กลุ่มตัวอย่างขนาดกลาง (1,000 คน)

ด้านภาษา จำนวน 2 ข้อ คิดเป็นร้อยละ 7 ด้านคำนวณ พบ DIF จำนวน 1 ข้อ คิดเป็นร้อยละ 3 และด้านเหตุผล พบ DIF จำนวน 6 ข้อ คิดเป็นร้อยละ 20 และกลุ่มตัวอย่างขนาดใหญ่ (2,000 คน) ด้านภาษาพบว่า จำนวน 4 ข้อ คิดเป็นร้อยละ 13 ด้านคำนวณ พบ DIF จำนวน 6 ข้อ คิดเป็นร้อยละ 20 และด้านเหตุผล พบ DIF จำนวน 12 ข้อ คิดเป็นร้อยละ 40 และวิธีแมนเทล-แฮนส์เซล ตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน (DIF) กลุ่มตัวอย่างขนาดเล็ก (300 คน) ด้านภาษา จำนวน 5 ข้อ คิดเป็นร้อยละ 17 สำหรับด้านคำนวณ พบ DIF จำนวน 2 ข้อ คิดเป็นร้อยละ 7 และด้านเหตุผลพบ DIF จำนวน 5 ข้อ คิดเป็นร้อยละ 17 ส่วนกลุ่มตัวอย่างขนาดกลาง (1,000 คน) ด้านภาษา จำนวน 11 ข้อ คิดเป็นร้อยละ 37 สำหรับด้านคำนวณ พบ DIF จำนวน 4 ข้อ คิดเป็นร้อยละ 13 และด้านเหตุผล พบ DIF จำนวน 15 ข้อ คิดเป็นร้อยละ 50 และกลุ่มตัวอย่างขนาดใหญ่ (2,000 คน) ด้านภาษา จำนวน 19 ข้อ คิดเป็นร้อยละ 63 ด้านคำนวณ พบ DIF จำนวน 8 ข้อ คิดเป็นร้อยละ 27 และด้านเหตุผลพบ DIF จำนวน 22 ข้อ คิดเป็นร้อยละ 73 อย่างมีนัยสำคัญทางสถิติที่ .05

3. ผลการเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) ด้วยวิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล พบว่าด้านภาษา วิธีการทดสอบอัตราส่วนไลค์ลิฮูด พบ DIF มากกว่า วิธีซิปเทสท์ ร้อยละ 7 ของข้อสอบทั้งหมด วิธีการทดสอบอัตราส่วนไลค์ลิฮูด พบ DIF น้อยกว่า วิธีแมนเทล- แฮนส์เซล ร้อยละ 80 ของข้อสอบทั้งหมด และ วิธีซิปเทสท์ พบ DIF น้อยกว่า วิธีแมนเทล-แฮนส์เซล ร้อยละ 87 ของข้อสอบทั้งหมด ส่วนด้านคำนวณ วิธีอัตราส่วนไลค์ลิฮูด พบ DIF เท่ากันกับ วิธีซิปเทสท์ ร้อยละ 0 ของข้อสอบทั้งหมด วิธีอัตราส่วนไลค์ลิฮูด พบ DIF น้อยกว่าวิธีแมนเทล-แฮนส์เซล ร้อยละ 13 ของข้อสอบทั้งหมด และวิธีซิปเทสท์ พบ DIF น้อยกว่าวิธีแมนเทล-แฮนส์เซล ร้อยละ 13 ของข้อสอบทั้งหมด และส่วนด้านเหตุผล วิธีอัตราส่วนไลค์ลิฮูด พบ DIF น้อยกว่าวิธีซิปเทสท์ ร้อยละ 7 ของข้อสอบทั้งหมด วิธีอัตราส่วนไลค์ลิฮูด พบ DIF น้อยกว่าวิธีแมนเทล-แฮนส์เซล ร้อยละ 80 ของข้อสอบทั้งหมด และ วิธีซิปเทสท์ พบ DIF น้อยกว่าวิธีแมนเทล-แฮนส์เซล ร้อยละ 73 ของข้อสอบทั้งหมด อย่างมีนัยสำคัญทางสถิติที่ .05

อภิปรายผล

ผลการวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล โดยใช้ทฤษฎีการตอบสนองข้อสอบแบบ 3 พารามิเตอร์ การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) และการเปรียบเทียบผลการตรวจพบข้อสอบทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ ทั้ง 3 วิธี มีประเด็นที่ควรอภิปราย ดังนี้

1. การวิเคราะห์คุณภาพของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ และด้านเหตุผล การวิเคราะห์คุณภาพของแบบทดสอบโดยใช้หลักการทฤษฎีการตอบสนองข้อสอบ (IRT) แบบ 3 พารามิเตอร์ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) ประกอบด้วยค่าอำนาจจำแนกของข้อสอบ (a)

ค่าความยากของข้อสอบ (b) และค่าโอกาสการเดาของข้อสอบ (c) แบบทดสอบระดับชาติ ด้านภาษา มีค่าอำนาจจำแนกของข้อสอบ (a) ทั้งฉบับอยู่ในระดับดีมาก มีค่าความยากของข้อสอบ (b) ทั้งฉบับอยู่ในระดับปานกลาง ถึงค่อนข้างยาก และมีค่าโอกาสการเดาของข้อสอบ (c) ทั้งฉบับไม่เกิน 0.3 สำหรับด้านคำนวณ มีค่าอำนาจจำแนกของข้อสอบ (a) ทั้งฉบับอยู่ในระดับดีมากมีค่าความยากของข้อสอบ (b) ทั้งฉบับอยู่ในระดับยาก และมีค่าโอกาสการเดาของข้อสอบ (c) ทั้งฉบับไม่เกิน 0.3 และส่วนด้านเหตุผล มีค่าอำนาจจำแนกของข้อสอบ (a) ทั้งฉบับอยู่ในระดับดีมาก มีค่าความยากของข้อสอบ (b) ทั้งฉบับอยู่ในระดับค่อนข้างยาก และมีค่าโอกาสการเดาของข้อสอบ (c) ทั้งฉบับไม่เกิน 0.3

2. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธีการตรวจสอบ 3 วิธี คือ วิธี IRT-LR วิธี SIBTEST และ วิธี Mantel-Haenszel ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก ขนาดกลาง และขนาดใหญ่ พบว่า วิธี Mantel-Haenszel สามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกัน (DIF) ได้มากที่สุด รองลงมา คือ วิธี IRT-LR และวิธี SIBTEST ตามลำดับ ซึ่งสอดคล้องกับงานวิจัยของ สุธาทิพย์ ตรีสิน และปิยะทิพย์ ประดุจพรม (2560) พบว่า วิธี IRT-LR จะมีประสิทธิภาพในการตรวจสอบ DIF ได้ดี นอกจากนี้ยังสอดคล้องกับงานวิจัยของ Awuor (2008) ที่ได้ศึกษาขนาดกลุ่มตัวอย่างไม่เท่ากันในการตรวจสอบ DIF ด้วยวิธี IRT-Based วิธี SIBTEST และวิธี Mantel-Haenszel พบว่า ขนาดตัวอย่างต่างกันจะส่งผลต่อการตรวจสอบ DIF โดยวิธี SIBTEST และวิธี Mantel-Haenszel มีประสิทธิภาพในการตรวจสอบ DIF ได้ดีเมื่อกลุ่มตัวอย่างขนาดกลาง และขนาดใหญ่ และสอดคล้องงานวิจัยของ พีรญา สูงเนิน เสรี ชัดแช้ม และ สมโภชน์ อเนกสุข (2552) เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบแบบพหุมิติระหว่างข้อสอบรายข้อกับหมวดข้อสอบ ด้วยวิธี SIBTEST ผลการศึกษาพบว่า ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก ขนาดกลาง และขนาดใหญ่ ระหว่างการตรวจสอบการทำหน้าที่ต่างกันเป็นรายข้อกับรายหมวดข้อสอบ พบข้อสอบทำหน้าที่ต่างกันแตกต่างกัน มีนัยสำคัญทางสถิติที่ระดับ .05 และขนาดของกลุ่มตัวอย่างต่างกันมีผลต่อการตรวจพบข้อสอบทำหน้าที่ต่างกัน เมื่อขนาดของกลุ่มตัวอย่างใหญ่ขึ้นจะทำให้สามารถตรวจพบข้อสอบทำหน้าที่ต่างกันได้ดีกว่ากลุ่มตัวอย่างขนาดเล็ก ซึ่งงานวิจัยของ Kabasakal, Arsan, Gok, and Kelecioglu (2014) ศึกษาประสิทธิภาพในการตรวจสอบ DIF ด้วยวิธี Mantel-Haenszel วิธี SIBTEST และวิธี IRT-LR พบว่า ประสิทธิภาพในการตรวจสอบ DIF ได้ดี และ Li, Hunter and Oshima (2013) ได้ศึกษาการตรวจสอบการทำหน้าที่ต่างกันในแบบทดสอบด้านการอ่าน และด้านเหตุผล ด้วยวิธีการทดสอบอัตราส่วนไลค์ลิฮูด และวิธีแมนเทล-แฮนส์เซล พบว่า มีประสิทธิภาพในการตรวจสอบ DIF ได้ดีในแบบทดสอบที่มีความยาวของข้อสอบตั้งแต่ 20 ข้อขึ้นไป

3. การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ทั้ง 3 วิธี ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก ขนาดกลาง และขนาดใหญ่ พบว่า วิธี Mantel-Haenszel ตรวจพบ DIF มากกว่าวิธี SIBTEST และวิธี IRT-LR ทั้ง 3 ด้าน คือ ด้านภาษา ด้านคำนวณ และด้านเหตุผล ส่วนวิธี Mantel-Haenszel ตรวจพบ DIF มากกว่าวิธี IRT-LR ในด้านภาษา และด้านเหตุผล ซึ่งสอดคล้องกับงานวิจัยของ Awuor (2008) ที่ได้ศึกษาขนาดกลุ่มตัวอย่างไม่เท่ากันในการตรวจสอบ DIF ด้วยวิธี IRT-Based วิธี SIBTEST และวิธี Mantel-Haenszel

พบว่าขนาดตัวอย่างที่แตกต่างกัน วิธี SIBTEST และวิธี Mantel-Haenszel ตรวจพบ DIF ได้ดีในกลุ่มตัวอย่างขนาดกลางและขนาดใหญ่ และตรวจพบได้ใกล้เคียงกัน และสอดคล้องกับงานวิจัยของ Kabasakal, Arsan, Gok, and Kelecioglu (2014) ได้ศึกษาประสิทธิภาพในการตรวจสอบ DIF ด้วยวิธี Mantel-Haenszel วิธี SIBTEST และวิธี IRT-LR ผลการศึกษาพบว่า มีประสิทธิภาพในการตรวจสอบ DIF ได้ดี ส่วนงานวิจัยของ Li, Hunter and Oshima (2013) ได้ศึกษาการตรวจสอบการทำหน้าที่ต่างกันในแบบทดสอบด้านการอ่าน และด้านเหตุผล ด้วยวิธีการทดสอบอัตราส่วนไลค์ลิฮูด และวิธีแมนเทล-แฮนส์เซล พบว่า มีประสิทธิภาพในการตรวจสอบ DIF ได้ดีในแบบทดสอบที่มีความยาวของข้อสอบตั้งแต่ 20 ข้อขึ้นไป และงานวิจัยของ Yildirim and Berberoglu (2009) ได้วิเคราะห์ข้อมูลการตัดสินใจและทางสถิติของวิชาคณิตศาสตร์ PISA-2003 พบว่า วิธี IRT-LR มีประสิทธิภาพในการตรวจสอบ DIF ได้ดีในด้านความสามารถทางคณิตศาสตร์ของโครงการประเมินผลนักเรียนนานาชาติ (PISA, 2003)

ข้อเสนอแนะในการนำผลการวิจัยไปใช้

จากผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ และการเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบระดับชาติ มีข้อเสนอแนะในการนำผลการวิจัยไปใช้ ดังนี้

1. สำนักทดสอบทางการศึกษาสามารถนำผลการวิเคราะห์คุณภาพของแบบทดสอบระดับชาติที่ผ่านเกณฑ์การวิเคราะห์คุณภาพโดยใช้หลักการของทฤษฎีการตอบสนองข้อสอบแบบ 3 พารามิเตอร์ ไปใช้สอบในครั้งต่อไป เพื่อใช้สำหรับวัดความสามารถของนักเรียนชั้นประถมศึกษาปีที่ 3 ของสำนักทดสอบทางการศึกษาสานักงานคณะกรรมการการศึกษาขั้นพื้นฐาน (สพฐ.)

2. นักวิจัยและนักวัดผลการศึกษาที่สนใจเกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธีการตรวจสอบที่อยู่บนพื้นฐานของทฤษฎีการตอบสนองข้อสอบ (IRT) ทั้ง 3 วิธี คือ วิธีอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล ขนาดของกลุ่มตัวอย่างต้องมีขนาดใหญ่จะทำให้สามารถตรวจพบข้อสอบทำหน้าที่ต่างกันได้ดีกว่ากลุ่มตัวอย่างขนาดเล็ก

ข้อเสนอแนะสำหรับการวิจัยต่อไป

1. วิธีอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล มีประสิทธิภาพในการตรวจพบข้อสอบที่ทำหน้าที่ต่างกันของข้อสอบ เมื่อกลุ่มตัวอย่างมีขนาดใหญ่ จะทำให้สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ดีกว่ากลุ่มตัวอย่างขนาดเล็ก จึงควรมีการเปรียบเทียบเพิ่มเติมกับวิธีการตรวจสอบอื่น ๆ และศึกษาตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบที่มีการตรวจให้คะแนนแบบมากกว่า 2 ค่า

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

บรรณานุกรม

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

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

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

ณรงค์ จันทรมหา. (2554). การเปรียบเทียบค่าความเที่ยงของแบบทดสอบผลสัมฤทธิ์ทางการเรียนที่มีจำนวนข้อสอบทำหน้าที่ต่างกันแตกต่างกัน. *วิทยาการวิจัยและวิทยาการปัญญา, 8*(2), 58-71.

ธเกียรติกมล ทองงอก, โชติกา ภาษี และศิริชัย กาญจนวาสี. (2556). ประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบภายใต้วิธีถดถอยโลจิสติกสำหรับข้อสอบที่ตรวจให้คะแนนแบบทวิวิภาค:เปรียบเทียบขนาดอิทธิพลสองเกณฑ์. *วารสารวิจัย มสด.*, 31-49.

ธีระวัฒน์ สุขีสาร, ดุษฎี โยเหลา, เสกสรรค์ ทองคำบรรจง และนิยะดา จิตต์จรัส. (2555). การศึกษาความเที่ยงตรงของการประมาณค่าในการวิเคราะห์โมเดลสมการโครงสร้างพหุระดับภายใต้เงื่อนไขวิธีการประมาณค่าและขนาดตัวอย่างที่ต่างกัน. *วารสารการวัดผลการศึกษา, 17*(1), 95-106.

นุภาพรรณ ปลื้มใจ ปิยะทิพย์ตินวร และโสพส สุขานนท์สวัสดิ์. (2558). การพัฒนาโปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์สำหรับการจัดสอบ O-NET ระดับชั้นมัธยมศึกษาปีที่ 6. *วิทยาการวิจัยและวิทยาการปัญญา, 13*(2), 109-125.

ปนัดดา หัสปราบ. (2557). แนวทางการนำผลการทดสอบทางการศึกษาไปใช้ในการพัฒนาคุณภาพผู้เรียน. *สถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน).* วันที่ค้นข้อมูล 1 สิงหาคม 2561, เข้าถึงได้จาก www.niets.or.th/th/content/download/262.

ปิยะทิพย์ ตินวร, ม.ร.ว.สมพร สุทัศน์ย์ และเสรี ชัดแช้ม. (2550). การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ: การเปรียบเทียบประสิทธิภาพระหว่างวิธีการวิเคราะห์องค์ประกอบจำกัดกับวิธีถดถอยโลจิสติก. *วารสารวิจัยและวัดผลการศึกษา, 5*(1), 63-80.

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

รุ่งนภา แสนอำนวยผล, ประกฤติยา ทักษิโณ และชนะศึก นิชานนท์ (2555). ประสิทธิภาพของแบบทดสอบ วัดผลสัมฤทธิ์ทางการเรียนรูปแบบผสม:การประยุกต์ ใช้ทฤษฎีการตอบสนองข้อสอบแบบ ตรวจให้คะแนนความรู้บางส่วน และทฤษฎีการตอบสนองข้อสอบแบบตรวจให้คะแนนความรู้บางส่วนแบบทั่วไป. *วารสารศึกษาศาสตร์ มหาวิทยาลัยขอนแก่น, 35*(1), 58-66.

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

ศาตรา แสนปัญญา. (2555). การศึกษาพัฒนาการความสามารถทางวิทยาศาสตร์โดยใช้วิธีปรับแนวตั้งตามทฤษฎีการตอบสนองข้อสอบ. ใน *การประชุมวิชาการแห่งชาติ ครั้งที่ 9 วันที่ 6-7 ธันวาคม 2555* (หน้า1493 - 1499) นครปฐม: มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตกำแพงแสน.

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

สำนักทดสอบทางการศึกษา. (2557). *ผลการประเมินคุณภาพผู้เรียนระดับชาติ ปีการศึกษา 2555 บทสรุปและข้อเสนอแนะเชิงนโยบาย.* กรุงเทพฯ: โรงพิมพ์ชุมชนสหกรณ์การเกษตรแห่งประเทศไทย.

สุชาดา สกลกิจรุ่งโรจน์ เสรี ชัดแช้ม และสมพร สุทัศนีย์.(2559). ประสิทธิภาพของโปรแกรมการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์สำหรับมาตรวัดความสุขของคนไทย. *วารสารการวัดผลการศึกษา มหาวิทยาลัยมหาสารคาม, 22*(2), 321-331.

สุทธิวรรณ พีรศักดิ์โสภณ, ปิยพงษ์ คล้ายคลึง และสมกิจ กิจพูนวงศ์. (2560). การศึกษาคุณภาพแบบทดสอบความถนัดทางการเรียน SWUSAT ปีการศึกษา 2553 -2555 ตามทฤษฎีการทดสอบแบบมาตรฐานเดิมและทฤษฎีการตอบข้อสอบ. *วารสารศึกษาศาสตร์ มหาวิทยาลัยทักษิณ, 17*(1), 65-73.

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

สุภาภรณ์ แดงเพ็ง, ชิดชนก เชิงเชาว์ และบุญญิสา แซ่หล่อ. (2554).การเปรียบเทียบความลำเอียงของแบบทดสอบคณิตศาสตร์ในการประเมินคุณภาพการศึกษาระดับท้องถิ่นชั้นประถมศึกษาปีที่ 5 จังหวัดปัตตานี ระหว่างวิธีแมนเทล-แฮนส์เซลและโค้งลักษณะข้อสอบ 3 พารามิเตอร์. *วารสารสงขลานครินทร์, 17*(2), 135-326.

อนุชิต กลิ่นกำเนิด, อรจิรา สิทธิศักดิ์ และทัศนวรรณ ศูนย์กลาง.(2555). ผลการประเมินระบบบริหารจัดการการเรียนรู้แบบปรับเหมาะ กรณีศึกษาเรื่อง องค์ประกอบของระบบสารสนเทศ. *วารสารวิชาการทางเทคโนโลยีคอมพิวเตอร์และระบบสารสนเทศประยุกต์, 1*(2), 1-7.

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

Acar, T., & Kelecioglu, H. (2010). Comparison of Differential Item Functioning Determination Techniques: HGLM, LR and IRT-LR. *Educational Sciences: Theory and Practice, 10*(2), 639-649.

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

Apinyapibal, S., Lawthong, N., Kanjanawasee, S. (2015). A Comparative Analysis of the Efficacy of Differential Item Functioning Detection For Dichotomously Scored Items among Logistic Regression SIBTEST and Raschtree Methods. *Procedia-Social and Behavioral Sciences, 191,* 21–25.

Atar, B. (2007). *Differential Item Functioning Analyses for Mixed Response data using IRT Likelihood-Ratio Test, Logistic Regression, and GLLMM Procedures*. Doctor of Philosophy, Department of Educational Psychology and Learning Systems.

Awuor, R. A. (2008). *Effect of Unequal Sample Sizes on the Power of DIF Detection: An IRT-Based Monte Carlo Study with SIBTEST and Mantel-Haenszel Procedures.* Doctor of Philosophy, Educational Research and Evaluation, Virginia Polytechnic Institute and State University, 1-111.

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

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

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

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

González-Betanzos, F., Abad, F. J., Barrada, J. R. (2014). Fixed item parameter calibration for assessing differential item functioning in computerized adaptive tests. *Psicológica, 35*, 331-359.

Kabasakal, K. A., Arsan, N., Gok, B., Kelecioglu, H. (2014). Comparing Performances (Type I error and Power)of IRT Likelihood Ratio SIBTEST and Mantel-Haenszel Methods in the Determination of Differential Item Functioning. *Educational Sciences: Theory & Practice, 14*(6), 2186-2193.

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

Kose, I. A., & Demirtasli, N. C. (2012). Comparison of unidimensional and multidimensional models based on item response theory in terms of both variables of test length and sample size. *Procedia - Social and Behavioral Sciences, 46,* 135–140.

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

Li, H., Hunter, C. V., & Oshima, T. C. (2013). *Gender DIF in Reading Tests: A Synthesis of Research.* In New Developments in Quantitative Psychology (pp. 489-506). Springer, New York: Springer Science Business Media.

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

Moghadamzadeh, A., Salehi, K., & Khodaie, E. (2011). A Comparison the Information Functions of the Item and Test in One, Two and Three Parametric Model of the Item Response Theory (IRT). *Procedia - Social and Behavioral Sciences, 29,* 1359–1367.

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 Internationnal Conference 2015: Moving Forword to a Prosperous and Sustainable Community, July 10-12, 2015 Bangsaen Heritafe Hotel Chonburi.* Thailand: Burapha University.

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

Rivas, G. E. L. (2012). *Detection and Classification of DIF Types Using Parametric and NonparametricMethods: A comparison of the IRT-Likelihood Ratio Test, Crossing-SIBTEST, and Logistic Regression Procedures*. University of South Florida, Department of Psychology College of Arts and Sciences.

Sterinmayr, R., Bergold, S., Stiksrud, J. M., & Freund, P. A. (2015). Gender differences on general knowledge tests: Are they due to Differential Item Functioning. *Inteligence, 50*, 164-174.

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

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

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99*(1), 118-128.

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

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

ภาคผนวก

ภาคผนวก ก
หนังสือรับรองผลการพิจารณาจริยธรรมการวิจัย

แบบรายงานผลการพิจารณาจริยธรรมการวิจัยในคน
วิทยาลัยวิทยาการวิจัยและวิทยาการปัญญา
มหาวิทยาลัยบูรพา

๑. ชื่อเรื่องวิทยานิพนธ์
ชื่อเรื่องวิทยานิพนธ์ (ภาษาไทย) การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ระดับชั้นประถมศึกษาปีที่ ๓: ระหว่างวิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิบเทสท์ และวิธีแมนเทล-แฮนส์เซล
ชื่อเรื่องวิทยานิพนธ์ (ภาษาอังกฤษ) COMPARISON OF THE DIFFERENTIAL ITEM FUNCTIONING OF THE NATIONAL TEST ITEM AT GRADE 3 LEVEL BETWEEN IRT LIKELIHOOD RATIO, SIBTEST AND MENTEL-HAENSZEL METHODS

๒. ชื่อนิสิต (นาย, นาง, นางสาว): อรุณี แปลงกาย
หลักสูตรวิทยาศาสตรมหาบัณฑิต (M.Sc.) สาขาวิชาการวิจัยและสถิติทางวิทยาการปัญญา
☑ ภาคปกติ ☐ ภาคพิเศษ
รหัสประจำตัว ๕๖๙๑๐๔๐๓ 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: ๒๘ เดือน มีนาคม พ.ศ. ๒๕๕๙

ลงนาม.....

(ผู้ช่วยศาสตราจารย์ ดร.สุชาดา กรเพชรปาณี)
ประธานกรรมการพิจารณาจริยธรรมการวิจัยในคน
คณบดีวิทยาลัยวิทยาการวิจัยและวิทยาการปัญญา
วันที่ ๒๘ มีนาคม พ.ศ. ๒๕๕๙

ภาคผนวก ข
หนังสือขอความอนุเคราะห์ข้อมูลเพื่อการวิจัย

ที่ ศธ ๖๖๒๘/๐๗๖๗

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

๒๐ เมษายน ๒๕๕๙

เรื่อง ขอความอนุเคราะห์ขอข้อมูลเพื่อการวิจัย

เรียน ผู้อำนวยการสำนักทดสอบทางการศึกษา

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

ด้วย นางสาวอรุณี แปลงกาย รหัสประจำตัว ๕๖๙๑๐๔๐๓ นิสิตหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาการวิจัยและสถิติทางวิทยาการปัญญา ได้รับอนุมัติให้ทำวิทยานิพนธ์เรื่อง "การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ระดับชั้นประถมศึกษาปีที่ ๓ ระหว่างวิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิบเทสท์ และวิธีแมนเทล-แอนส์เซล" ซึ่งอยู่ในความควบคุมดูแลของ ดร.ปิยะทิพย์ ประดุจพรม อาจารย์ที่ปรึกษาหลัก ในการนี้ ผู้วิจัยมีความประสงค์ขอความอนุเคราะห์ขอข้อสอบ NT (National Test) ระดับชั้นประถมศึกษาปีที่ ๓ พร้อมเฉลย รหัสโรงเรียน เพศ และผลการตอบข้อสอบ ในปีการศึกษา ๒๕๕๖ จำนวน ๓ ด้าน คือ ด้านภาษา ด้านการคิดคำนวณ และด้านเหตุผล

จึงเรียนมาเพื่อโปรดพิจารณา วิทยาลัยวิทยาการวิจัยและวิทยาการปัญญา หวังเป็นอย่างยิ่งว่าคงจะได้รับความอนุเคราะห์จากท่านด้วยดี และขอขอบคุณมา ณ โอกาสนี้

ขอแสดงความนับถือ

(ผู้ช่วยศาสตราจารย์ ดร.สุชาดา กรเพชรปาณี)
คณบดีวิทยาลัยวิทยาการวิจัยและวิทยาการปัญญา

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

ภาคผนวก ค
ตัวอย่างแสดงข้อมูลผลการตอบข้อสอบ NT
ปีการศึกษา 2556 ชั้นประถมศึกษาปีที่ 3

ตารางที่ ค-1 แสดงข้อมูลผลการตอบข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา จำนวน 300 คน

ข้อ																																
Student	Gender	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	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	0	0	0	1	1	0	1	22
2	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	0	0	0	1	1	0	1	22
3	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	0	0	0	1	1	0	1	22
4	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	0	1	0	1	1	0	1	22
5	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0	1	0	1	0	1	1	0	1	22
6	1	0	0	1	0	1	0	0	0	1	0	0	1	1	1	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	9
7	1	1	0	1	1	0	1	1	0	0	1	1	0	0	1	0	0	1	1	1	1	1	0	1	0	0	1	1	1	1	1	19
8	1	1	0	1	1	0	0	1	0	1	0	1	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	9
9	1	0	0	0	1	0	0	1	0	1	0	1	1	1	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	9
10	1	0	1	1	1	0	1	1	1	1	0	1	1	1	0	1	0	1	1	1	1	0	0	0	1	0	0	0	1	1	1	19
11	1	1	0	1	0	1	0	1	0	1	1	1	1	1	0	1	0	1	1	1	1	1	1	0	0	0	0	0	1	1	1	19
12	1	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	1	0	1	0	8
13	1	0	1	0	0	0	1	1	1	1	0	0	1	1	1	0	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0	12
14	1	1	0	1	1	1	1	1	1	1	0	1	1	0	0	0	1	1	1	1	1	1	0	0	1	0	1	1	1	0	0	20
15	1	0	1	1	1	0	0	1	1	1	1	0	1	1	0	0	1	0	1	1	1	0	0	0	1	1	0	0	0	1	0	16
16	2	1	1	1	1	0	1	1	1	1	0	1	1	0	0	0	1	1	0	1	1	0	0	0	1	1	1	1	1	1	0	20
17	2	1	1	0	0	1	1	1	0	1	1	1	0	1	1	1	1	1	0	1	1	0	0	0	0	1	0	1	1	1	0	19
18	2	0	1	1	1	1	1	0	0	1	0	1	0	1	0	1	1	0	0	1	1	0	0	0	0	1	0	1	1	0	1	16
19	2	0	1	0	1	1	1	1	0	0	0	1	0	1	0	0	0	0	0	1	1	0	0	1	1	1	0	1	0	1	1	15
20	2	0	1	1	1	0	0	0	1	1	0	0	1	0	0	0	1	1	1	1	1	0	0	0	1	1	1	1	0	1	1	17
21	2	0	1	1	1	0	1	1	1	1	0	1	1	1	0	0	1	1	1	1	1	1	0	0	1	0	1	0	1	1	1	21
22	2	0	0	1	0	1	0	0	0	1	0	0	1	1	1	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	9
23	2	1	0	1	1	0	1	1	0	0	1	1	0	0	1	0	0	1	1	1	1	1	0	1	0	0	1	1	1	1	1	19
24	2	1	0	1	1	0	0	1	0	1	0	1	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	9
25	2	1	1	1	1	1	0	1	0	1	0	1	1	1	1	1	0	1	0	1	1	1	0	0	0	1	0	0	1	0	0	18
26	2	0	1	1	1	0	1	1	1	1	0	1	1	0	0	0	1	1	1	1	1	1	0	0	1	0	1	0	1	1	1	20
27	2	1	0	1	1	0	1	1	0	0	1	1	0	0	1	0	0	1	1	1	1	1	0	1	0	0	1	1	1	0	1	18
28	2	1	0	1	1	0	1	1	0	0	1	1	0	0	1	0	0	1	0	1	0	1	1	1	0	0	1	1	1	1	1	18
29	2	1	0	1	1	0	1	1	0	0	1	1	0	0	1	0	0	1	1	1	1	1	0	1	0	0	1	1	1	0	1	18
30	2	0	0	0	1	1	0	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	0	0	1	0	1	0	1	1	19
300	2	0	1	1	0	1	0	1	0	1	0	0	1	0	0	1	0	1	0	1	1	1	0	0	0	0	0	0	1	1	0	13
รวม		161	135	185	210	154	211	209	143	233	125	211	175	145	145	130	137	197	127	224	197	162	69	78	127	120	140	131	179	145	151	

หมายเหตุ 1 คือ เพศชาย 2 คือ เพศหญิง

ตารางที่ ค-2 แสดงข้อมูลผลการตอบข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านคำนวณ จำนวน 300 คน

ข้อ																																
Student	Gender	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	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1	0	0	0	1	1	0	1	1	1	1	1	1	1	23
2	1	1	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1	0	0	0	1	1	0	1	1	1	1	1	1	1	23
3	1	1	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1	0	0	0	1	1	0	1	1	1	1	1	1	1	23
4	1	1	1	1	0	1	0	0	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	1	0	1	1	15
5	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	1	0	0	0	0	0	0	1	1	1	1	0	1	1	11
6	1	0	0	1	0	0	0	0	0	0	0	1	0	1	1	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	8
7	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	6
8	1	0	0	0	0	0	0	0	0	1	1	0	1	0	0	1	1	0	1	0	1	0	0	0	0	0	0	1	0	1	0	10
9	1	0	0	1	0	1	0	0	1	1	1	0	0	0	0	0	0	1	0	0	1	1	1	0	0	0	1	1	1	0	1	14
10	1	0	0	0	1	1	0	1	0	1	1	1	1	1	0	0	1	1	1	1	1	0	0	0	1	1	0	1	1	0	1	19
11	1	0	1	0	1	0	1	1	0	0	1	0	0	0	0	0	1	1	1	1	1	0	0	0	0	1	1	1	0	0	1	15
12	1	0	0	0	0	1	1	1	1	0	1	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	1	1	0	11
13	1	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	5
14	1	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	1	1	1	0	0	1	10
15	1	1	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1	0	0	0	1	1	0	1	1	1	1	1	1	1	23
16	2	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	8
17	2	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0	1	1	0	1	0	0	0	1	1	0	0	0	1	0	12
18	2	1	0	0	0	1	0	0	0	0	1	1	0	0	1	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	1	12
19	2	1	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1	0	0	0	1	1	0	0	0	0	1	1	0	1	20
20	2	1	0	0	0	0	0	0	1	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	9
21	2	0	1	1	0	0	0	0	1	1	1	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	1	1	0	1	15
22	2	0	0	1	0	0	0	0	0	0	0	1	0	1	1	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	9
23	2	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	7
24	2	0	0	0	0	0	0	0	0	1	1	0	1	0	0	1	1	0	1	0	1	0	0	0	0	0	0	1	0	1	0	11
25	2	0	0	0	0	1	0	0	0	0	1	1	0	0	1	1	1	1	1	1	0	0	0	0	0	0	1	1	0	1	1	15
26	2	0	0	0	0	0	0	0	0	1	1	1	1	0	1	0	0	1	1	1	0	1	0	0	0	0	0	1	1	0	1	14
27	2	0	0	1	0	0	0	0	1	1	0	1	1	0	0	0	1	0	0	0	1	0	0	1	0	1	1	0	0	0	0	12
28	2	0	0	0	0	1	1	0	1	1	0	1	1	0	0	0	1	1	1	0	1	1	0	0	1	0	0	1	0	0	0	15
29	2	0	0	1	0	0	0	0	1	1	0	1	1	0	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0	11
30	2	0	1	1	0	1	0	0	1	1	0	1	1	0	0	0	1	0	0	0	1	1	0	0	1	0	1	1	0	1	1	17
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
300	2	1	1	1	1	0	0	0	1	1	0	1	1	0	0	1	0	1	1	1	1	0	0	0	0	0	0	0	0	1	0	14
รวม		83	93	107	130	110	77	81	153	157	119	151	97	128	114	114	142	180	119	118	162	111	97	61	113	108	104	150	131	141	111	

หมายเหตุ 1 คือ เพศชาย 2 คือ เพศหญิง

ตารางที่ ค-3 แสดงข้อมูลผลการตอบข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านเหตุผล จำนวน 300 คน

ข้อ																																
Student	Gender	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	1	1	1	1	1	1	0	0	1	1	0	0	1	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	16
2	1	1	1	1	1	1	1	1	0	0	1	1	0	0	1	0	0	1	1	1	0	1	1	1	1	1	1	0	0	1	1	21
3	1	1	1	1	1	1	1	1	0	0	1	1	0	0	1	0	0	1	1	1	0	1	1	1	1	1	1	0	0	1	1	21
4	1	1	1	1	1	1	1	1	0	0	1	1	0	0	1	0	0	1	1	1	0	1	1	0	1	1	1	0	0	1	1	20
5	1	1	1	1	1	1	1	1	0	0	1	1	0	0	1	0	0	1	1	1	0	1	1	0	1	1	1	0	0	1	1	20
6	1	0	0	1	1	0	0	1	0	1	1	1	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	9
7	1	0	1	0	1	1	1	1	0	1	0	0	0	1	0	0	0	0	0	1	0	0	1	0	1	1	1	1	1	0	1	15
8	1	0	1	0	1	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	0	1	14
9	1	1	1	0	0	0	0	0	0	0	1	1	1	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	1	0	1	10
10	1	1	1	0	0	1	0	1	1	0	1	1	1	0	0	1	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	12
11	1	1	1	0	1	1	0	0	1	0	1	1	1	0	0	1	0	0	1	1	1	1	1	0	1	1	1	0	0	0	0	17
12	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	1	1	1	0	0	0	0	1	1	1	0	0	0	9
13	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	5
14	1	1	1	0	1	0	1	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	0	0	1	0	0	15
15	1	1	0	0	1	1	1	1	1	0	1	1	1	1	0	1	0	0	1	0	1	0	0	0	0	0	0	1	1	0	0	15
16	2	0	0	1	1	1	1	0	0	1	1	1	0	1	0	0	0	0	1	1	1	1	1	0	1	0	1	0	1	1	0	17
17	2	0	0	1	0	1	1	1	0	0	1	1	1	1	0	0	0	0	1	0	0	1	0	0	1	0	0	1	0	1	0	13
18	2	1	1	1	1	0	1	1	0	1	1	1	1	1	0	0	0	1	0	1	0	0	1	1	1	1	0	1	0	1	0	19
19	2	0	1	1	1	1	1	1	0	0	1	1	1	1	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	1	0	14
20	2	1	0	1	1	0	0	1	0	0	1	1	1	1	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	11
21	2	1	0	1	1	1	1	1	0	0	1	1	1	1	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	1	0	14
22	2	0	0	1	1	0	0	1	0	1	1	1	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	9
23	2	0	1	0	1	1	1	1	0	1	0	0	0	1	0	0	0	0	0	1	0	0	1	0	1	1	1	1	1	0	1	15
24	2	0	1	0	1	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	0	1	14
25	2	1	0	1	1	1	1	1	0	0	1	1	1	1	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	1	0	14
26	2	1	0	1	1	1	1	1	0	1	0	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	1	13
27	2	0	1	0	1	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	0	1	14
28	2	0	1	0	1	1	1	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	1	1	1	1	1	0	1	14
29	2	0	1	0	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	1	1	1	1	0	1	13
30	2	0	1	1	1	0	0	1	1	0	0	1	1	1	1	1	0	0	1	0	0	0	0	1	1	1	1	1	1	0	1	18
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
300	2	0	1	1	1	1	0	0	0	0	0	1	1	1	0	0	0	0	0	1	0	1	0	1	1	1	1	0	1	0	0	14
รวม		166	171	169	211	174	183	170	140	152	118	167	166	147	106	132	99	84	141	134	105	164	157	146	170	141	156	98	95	99	120	

หมายเหตุ 1 คือ เพศชาย 2 คือ เพศหญิง

ตารางที่ ค-4 แสดงข้อมูลผลการตอบข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา จำนวน 1,000 คน

ข้อ																															
Student	Gender	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	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	0	0	0	1	1	0	1	22
2	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	0	0	0	1	1	0	1	22
3	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	0	0	0	1	1	0	1	22
4	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	0	1	0	1	1	0	1	22
5	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0	1	0	1	0	1	1	0	1	22
6	1	0	0	1	0	1	0	0	0	1	0	0	1	1	1	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	9
7	1	1	0	1	1	0	1	1	0	0	1	1	0	0	1	0	0	1	1	1	1	0	1	0	0	1	1	1	1	1	1	19
8	1	1	0	1	1	0	0	1	0	1	0	1	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	9
9	1	0	0	0	1	0	0	1	0	1	0	1	1	1	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	9
10	1	0	1	1	1	0	1	1	1	1	0	1	1	1	0	1	0	1	1	1	1	0	0	0	1	0	0	0	1	1	1	19
11	1	1	0	1	0	1	0	1	0	1	1	1	1	1	0	1	0	1	1	1	1	1	1	0	0	0	0	0	1	1	1	19
12	1	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	1	0	1	0	8
13	1	0	1	0	0	0	1	1	1	1	0	0	1	1	1	0	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0	12
14	1	1	0	1	1	1	1	1	1	0	1	1	0	0	0	1	1	1	1	1	1	1	0	0	1	0	1	1	1	0	0	20
15	1	0	1	1	1	0	0	1	1	1	1	0	1	1	0	0	1	0	1	1	1	0	0	0	1	1	0	0	0	1	0	16
16	2	1	1	1	1	0	1	1	1	1	0	1	1	0	0	0	1	1	0	1	1	0	0	0	1	1	1	1	1	1	0	20
17	2	1	1	0	0	1	1	1	0	1	1	1	0	1	1	1	1	1	0	1	1	0	0	0	0	1	0	1	1	1	0	19
18	2	0	1	1	1	1	1	0	0	1	0	1	0	1	0	1	1	0	0	1	1	0	0	0	0	1	0	1	1	0	1	16
19	2	0	1	0	1	1	1	1	0	0	0	1	0	1	0	0	0	0	0	1	1	0	0	1	1	1	0	1	0	1	1	15
20	2	0	1	1	1	0	0	0	1	1	0	0	1	0	0	0	1	1	1	1	1	0	0	0	1	1	1	1	0	1	1	17
21	2	0	1	1	1	0	1	1	1	1	0	1	1	1	0	0	1	1	1	1	1	1	0	0	1	0	1	0	1	1	1	21
22	2	0	0	1	0	1	0	0	0	1	0	0	1	1	1	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	9
23	2	1	0	1	1	0	1	1	0	0	1	1	0	0	1	0	0	1	1	1	1	1	0	1	0	0	1	1	1	1	1	19
24	2	1	0	1	1	0	0	1	0	1	0	1	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	9
25	2	1	1	1	1	1	0	1	0	1	0	1	1	1	1	1	0	1	0	1	1	1	0	0	0	1	0	0	1	0	0	18
26	2	0	1	1	1	0	1	1	1	1	0	1	1	0	0	0	1	1	1	1	1	1	0	0	1	0	1	0	1	1	1	20
27	2	1	0	1	1	0	1	1	0	0	1	1	0	0	1	0	0	1	1	1	1	1	0	1	0	0	1	1	1	0	1	18
28	2	1	0	1	1	0	1	1	0	0	1	1	0	0	1	0	0	1	0	1	0	1	1	1	0	0	1	1	1	1	1	18
29	2	1	0	1	1	0	1	1	0	0	1	1	0	0	1	0	0	1	1	1	1	1	0	1	0	0	1	1	1	0	1	18
30	2	0	0	0	1	1	0	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	0	0	1	0	1	0	1	1	19
1,000	2	0	0	0	1	1	0	1	1	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	1	1	0	0	12
รวม		542	323	601	750	606	736	749	495	794	401	705	576	419	487	423	448	693	453	791	689	530	247	349	488	367	483	444	554	420	565	

หมายเหตุ 1 คือ เพศชาย 2 คือ เพศหญิง

ตารางที่ ค-5 แสดงข้อมูลผลการตอบข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านคำนวณ จำนวน 1,000 คน

ข้อ		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
Student	Gender																															
1	1	1	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1	0	0	0	1	1	0	1	1	1	1	1	1	1	23
2	1	1	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1	0	0	0	1	1	0	1	1	1	1	1	1	1	23
3	1	1	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1	0	0	0	1	1	0	1	1	1	1	1	1	1	23
4	1	1	1	1	0	1	0	0	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	1	0	1	1	15
5	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	1	0	0	0	0	0	0	1	1	1	1	0	1	1	11
6	1	0	0	1	0	0	0	0	0	0	0	1	0	1	1	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	8
7	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	6
8	1	0	0	0	0	0	0	0	0	1	1	0	1	0	0	1	1	0	1	0	1	0	0	0	0	0	0	1	0	1	0	10
9	1	0	0	1	0	1	0	0	1	1	1	0	0	0	0	0	0	1	0	0	1	1	1	0	0	0	1	1	1	0	1	14
10	1	0	0	0	1	1	0	1	0	1	1	1	1	1	0	0	1	1	1	1	1	0	0	0	1	1	0	1	1	0	1	19
11	1	0	1	0	1	0	1	1	0	0	1	0	0	0	0	0	1	1	1	1	1	0	0	0	0	1	1	1	0	0	1	15
12	1	0	0	0	0	1	1	1	1	0	1	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	1	1	0	11
13	1	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	5
14	1	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	1	1	1	0	0	1	10
15	1	1	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1	0	0	0	1	1	0	1	1	1	1	1	1	1	23
16	2	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	8
17	2	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0	1	1	0	1	0	0	0	1	1	0	0	0	1	0	12
18	2	1	0	0	0	1	0	0	0	0	1	1	0	0	1	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	1	12
19	2	1	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1	0	0	0	1	1	0	0	0	0	1	1	0	1	20
20	2	1	0	0	0	0	0	0	1	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	9
21	2	0	1	1	0	0	0	0	1	1	1	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	1	1	0	1	15
22	2	0	0	1	0	0	0	0	0	0	0	1	0	1	1	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	9
23	2	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	7
24	2	0	0	0	0	0	0	0	0	1	1	0	1	0	0	1	1	0	1	0	1	0	0	0	0	0	0	1	0	1	0	11
25	2	0	0	0	0	1	0	0	0	0	1	1	0	0	1	1	1	1	1	1	0	0	0	0	0	0	1	1	0	1	1	15
26	2	0	0	0	0	0	0	0	0	1	1	1	1	0	1	0	0	1	1	1	0	1	0	0	0	0	0	1	1	0	1	14
27	2	0	0	1	0	0	0	0	1	1	0	1	1	0	0	0	1	0	0	0	1	0	0	1	0	1	1	0	0	0	0	12
28	2	0	0	0	0	1	1	0	1	1	0	1	1	0	0	0	1	1	1	0	1	1	0	0	1	0	0	1	0	0	0	15
29	2	0	0	1	0	0	0	0	1	1	0	1	1	0	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0	11
30	2	0	1	1	0	1	0	0	1	1	0	1	1	0	0	0	1	0	0	0	1	1	0	0	1	0	1	1	0	1	1	17
.		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
1,000	2	1	1	1	1	0	0	0	1	1	0	1	1	0	0	1	0	1	1	1	1	0	0	0	0	0	0	0	0	1	0	14
รวม		263	383	344	496	354	324	276	595	556	446	477	350	517	342	385	482	625	431	455	519	364	240	192	448	432	386	517	366	433	386	

หมายเหตุ 1 คือ เพศชาย 2 คือ เพศหญิง

ตารางที่ ค-6 แสดงข้อมูลผลการตอบข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านเหตุผล จำนวน 1,000 คน

ข้อ		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
Student	Gender																															
1	1	1	1	1	1	1	1	1	0	0	1	1	0	0	1	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	16
2	1	1	1	1	1	1	1	1	0	0	1	1	0	0	1	0	0	1	1	1	0	1	1	1	1	1	1	0	0	1	1	21
3	1	1	1	1	1	1	1	1	0	0	1	1	0	0	1	0	0	1	1	1	0	1	1	1	1	1	1	0	0	1	1	21
4	1	1	1	1	1	1	1	1	0	0	1	1	0	0	1	0	0	1	1	1	0	1	1	0	1	1	1	0	0	1	1	
5	1	1	1	1	1	1	1	1	0	0	1	1	0	0	1	0	0	1	1	1	0	1	1	0	1	1	1	0	0	1	1	20
6	1	0	0	1	1	0	0	1	0	1	1	1	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	20
7	1	0	1	0	1	1	1	1	0	1	0	0	0	1	0	0	0	0	0	1	0	0	1	0	1	1	1	1	1	0	1	9
8	1	0	1	0	1	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	0	1	15
9	1	1	1	0	0	0	0	0	0	0	1	1	1	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	1	0	1	14
10	1	1	1	0	0	1	0	1	1	0	1	1	1	0	0	1	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	10
11	1	1	1	0	1	1	0	0	1	0	1	1	1	0	0	1	0	0	1	1	1	1	1	0	1	1	1	0	0	0	0	12
12	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	1	1	1	0	0	0	0	1	1	1	0	0	0	17
13	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	9
14	1	1	1	0	1	0	1	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	0	0	1	0	0	5
15	1	1	0	0	1	1	1	1	1	0	1	1	1	1	0	1	0	0	1	0	1	0	0	0	0	0	0	1	1	0	0	15
16	2	0	0	1	1	1	1	0	0	1	1	1	0	1	0	0	0	0	1	1	1	1	1	0	1	0	1	0	1	1	0	15
17	2	0	0	1	0	1	1	1	0	0	1	1	1	1	0	0	0	0	1	0	0	1	0	0	1	0	0	1	0	1	0	17
18	2	1	1	1	1	0	1	1	0	1	1	1	1	1	0	0	0	1	0	1	0	0	1	1	1	1	0	1	0	1	0	13
19	2	0	1	1	1	1	1	1	0	0	1	1	1	1	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	1	0	19
20	2	1	0	1	1	0	0	1	0	0	1	1	1	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	14
21	2	1	0	1	1	1	1	1	0	0	1	1	1	1	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	1	0	11
22	2	0	0	1	1	0	0	1	0	1	1	1	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	14
23	2	0	1	0	1	1	1	1	0	1	0	0	0	1	0	0	0	0	0	1	0	0	1	0	1	1	1	1	1	0	1	9
24	2	0	1	0	1	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	0	1	15
25	2	1	0	1	1	1	1	1	0	0	1	1	1	1	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	1	0	14
26	2	1	0	1	1	1	1	1	0	1	0	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	1	14
27	2	0	1	0	1	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	0	1	13
28	2	0	1	0	1	1	1	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	1	1	1	1	1	0	1	14
29	2	0	1	0	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	1	1	1	1	0	1	14
30	2	0	1	1	1	0	0	1	1	0	0	1	1	1	1	1	0	0	1	0	0	0	0	1	1	1	1	1	1	0	1	13
.		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
1.000	2	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	7
รวม		598	618	621	717	632	637	581	514	593	426	545	505	523	421	413	352	246	460	482	293	515	566	584	609	432	542	393	370	312	598	

หมายเหตุ 1 คือ เพศชาย 2 คือ เพศหญิง

ตารางที่ ค-7 แสดงข้อมูลผลการตอบข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา จำนวน 2,000 คน

ข้อ		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
Student	Gender																															
1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	0	0	0	1	1	0	1	22
2	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	0	0	0	1	1	0	1	22
3	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	0	0	0	1	1	0	1	22
4	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	0	1	0	1	1	0	1	22
5	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0	1	0	1	0	1	1	0	1	22
6	1	0	0	1	0	1	0	0	0	1	0	0	1	1	1	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	9
7	1	1	0	1	1	0	1	1	0	0	1	1	0	0	1	0	0	1	1	1	1	1	0	1	0	0	1	1	1	1	1	19
8	1	1	0	1	1	0	0	1	0	1	0	1	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	9
9	1	0	0	0	1	0	0	1	0	1	0	1	1	1	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	9
10	1	0	1	1	1	0	1	1	1	1	0	1	1	1	0	1	0	1	1	1	1	0	0	0	1	0	0	0	1	1	1	19
11	1	1	0	1	0	1	0	1	0	1	1	1	1	1	0	1	0	1	1	1	1	1	1	0	0	0	0	0	1	1	1	19
12	1	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	1	0	1	0	8
13	1	0	1	0	0	0	1	1	1	1	0	0	1	1	1	0	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0	12
14	1	1	0	1	1	1	1	1	1	1	0	1	1	0	0	0	1	1	1	1	1	1	0	0	1	0	1	1	1	0	0	20
15	1	0	1	1	1	0	0	1	1	1	1	0	1	1	0	0	1	0	1	1	1	0	0	0	1	1	0	0	1	0	0	16
16	2	1	1	1	1	0	1	1	1	1	0	1	1	0	0	0	1	1	0	1	1	0	0	0	1	1	1	1	1	1	0	20
17	2	1	1	0	0	1	1	1	0	1	1	1	0	1	1	1	1	1	0	1	1	0	0	0	0	1	0	1	1	1	0	19
18	2	0	1	1	1	1	1	0	0	1	0	1	0	1	0	1	1	0	0	1	1	0	0	0	0	1	0	1	1	0	1	16
19	2	0	1	0	1	1	1	1	0	0	0	1	0	1	0	0	0	0	0	1	1	0	0	1	1	1	0	1	0	1	1	15
20	2	0	1	1	1	0	0	0	1	1	0	0	1	0	0	0	1	1	1	1	1	0	0	0	1	1	1	1	0	1	1	17
21	2	0	1	1	1	0	1	1	1	1	0	1	1	1	0	0	1	1	1	1	1	1	0	0	1	0	1	0	1	1	1	21
22	2	0	0	1	0	1	0	0	0	1	0	0	1	1	1	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	9
23	2	1	0	1	1	0	1	1	0	0	1	1	0	0	1	0	0	1	1	1	1	1	0	1	0	0	1	1	1	1	1	19
24	2	1	0	1	1	0	0	1	0	1	0	1	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	9
25	2	1	1	1	1	1	0	1	0	1	0	1	1	1	1	1	0	1	0	1	1	1	0	0	0	1	0	0	1	0	0	18
26	2	0	1	1	1	0	1	1	1	1	0	1	1	0	0	0	1	1	1	1	1	1	0	0	1	0	1	0	1	1	1	20
27	2	1	0	1	1	0	1	1	0	0	1	1	0	0	1	0	0	1	1	1	1	1	0	1	0	0	1	1	1	0	1	18
28	2	1	0	1	1	0	1	1	0	0	1	1	0	0	1	0	0	1	0	1	0	1	1	1	0	0	1	1	1	1	1	18
29	2	1	0	1	1	0	1	1	0	0	1	1	0	0	1	0	0	1	1	1	1	1	0	1	0	0	1	1	1	0	1	18
30	2	0	0	0	1	1	0	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	0	0	1	0	1	0	1	1	19
.																																
.																																
.																																
2,000	2	1	0	0	0	1	0	0	0	0	1	0	1	0	0	1	0	0	0	1	0	0	0	0	1	1	1	0	1	0	0	10
รวม		1070	642	1200	1511	1147	1487	1403	969	1577	822	1401	1159	845	937	860	948	1380	903	1580	1346	1004	527	674	966	748	952	842	1133	895	1145	

หมายเหตุ 1 คือ เพศชาย 2 คือ เพศหญิง

ตารางที่ ค-8 แสดงข้อมูลผลการตอบข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านคำนวณ จำนวน 2,000 คน

ข้อ																																
Student	Gender	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	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1	0	0	0	1	1	0	1	1	1	1	1	1	1	23
2	1	1	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1	0	0	0	1	1	0	1	1	1	1	1	1	1	23
3	1	1	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1	0	0	0	1	1	0	1	1	1	1	1	1	1	23
4	1	1	1	1	0	1	0	0	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	1	0	1	1	15
5	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	1	0	0	0	0	0	0	1	1	1	1	0	1	1	11
6	1	0	0	1	0	0	0	0	0	0	0	1	0	1	1	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	8
7	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	6
8	1	0	0	0	0	0	0	0	0	1	1	0	1	0	0	1	1	0	1	0	1	0	0	0	0	0	0	1	0	1	0	10
9	1	0	0	1	0	1	0	0	1	1	1	0	0	0	0	0	0	1	0	0	1	1	1	0	0	0	1	1	1	0	1	14
10	1	0	0	0	1	1	0	1	0	1	1	1	1	1	0	0	1	1	1	1	1	0	0	0	1	1	0	1	1	0	1	19
11	1	0	1	0	1	0	1	1	0	0	1	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	1	1	1	0	0	15
12	1	0	0	0	0	1	1	1	1	0	1	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	1	1	0	11
13	1	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	5
14	1	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	1	1	1	0	0	1	10
15	1	1	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1	0	0	0	1	1	0	1	1	1	1	1	1	1	23
16	2	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	8
17	2	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0	1	1	0	1	0	0	0	1	1	0	0	0	1	0	12
18	2	1	0	0	0	1	0	0	0	0	1	1	0	0	1	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	1	12
19	2	1	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1	0	0	0	1	1	0	0	0	0	1	1	0	1	20
20	2	1	0	0	0	0	0	0	1	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	9
21	2	0	1	1	0	0	0	0	1	1	1	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	1	1	0	1	15
22	2	0	0	1	0	0	0	0	0	0	0	1	0	1	1	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	9
23	2	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	7
24	2	0	0	0	0	0	0	0	0	1	1	0	1	0	0	1	1	0	1	0	1	0	0	0	0	0	0	1	0	1	0	11
25	2	0	0	0	0	1	0	0	0	0	1	1	0	0	1	1	1	1	1	1	0	0	0	0	0	0	1	1	0	1	1	15
26	2	0	0	0	0	0	0	0	0	1	1	1	1	0	1	0	0	1	1	1	0	1	0	0	0	0	0	1	1	0	1	14
27	2	0	0	1	0	0	0	0	1	1	0	1	1	0	0	0	1	0	0	0	1	0	0	1	0	1	1	0	0	0	0	12
28	2	0	0	0	0	1	1	0	1	1	0	1	1	0	0	0	1	1	1	0	1	1	0	0	1	0	0	1	0	0	0	15
29	2	0	0	1	0	0	0	0	1	1	0	1	1	0	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0	11
30	2	0	1	1	0	1	0	0	1	1	0	1	1	0	0	0	1	0	0	0	1	1	0	0	1	0	1	1	0	1	1	17
2,000	2	1	1	1	1	0	0	0	1	1	0	1	1	0	0	1	0	1	1	1	1	0	0	0	0	0	0	0	0	1	0	14
รวม		511	663	665	951	735	621	618	1156	1081	884	934	715	1048	715	789	886	1236	778	779	1006	726	453	370	840	840	757	961	682	755	687	511

หมายเหตุ 1 คือ เพศชาย 2 คือ เพศหญิง

ตารางที่ ค-9 แสดงข้อมูลผลการตอบข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านเหตุผล จำนวน 2,000 คน

ข้อ		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
Student	Gender																															
1	1	1	1	1	1	1	1	1	0	0	1	1	0	0	1	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	16
2	1	1	1	1	1	1	1	1	0	0	1	1	0	0	1	0	0	1	1	1	0	1	1	1	1	1	1	0	0	1	1	21
3	1	1	1	1	1	1	1	1	0	0	1	1	0	0	1	0	0	1	1	1	0	1	1	1	1	1	1	0	0	1	1	21
4	1	1	1	1	1	1	1	1	0	0	1	1	0	0	1	0	0	1	1	1	0	1	1	0	1	1	1	0	0	1	1	
5	1	1	1	1	1	1	1	1	0	0	1	1	0	0	1	0	0	1	1	1	0	1	1	0	1	1	1	0	0	1	1	20
6	1	0	0	1	1	0	0	1	0	1	1	1	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	20
7	1	0	1	0	1	1	1	1	0	1	0	0	0	1	0	0	0	0	0	1	0	0	1	0	1	1	1	1	1	0	1	9
8	1	0	1	0	1	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	0	1	15
9	1	1	1	0	0	0	0	0	0	0	1	1	1	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	1	0	1	14
10	1	1	1	0	0	1	0	1	1	0	1	1	1	0	0	1	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	10
11	1	1	1	0	1	1	0	0	1	0	1	1	1	0	0	1	0	0	1	1	1	1	1	0	1	1	1	0	0	0	0	12
12	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	1	1	1	0	0	0	0	1	1	1	0	0	0	17
13	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	9
14	1	1	1	0	1	0	1	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	0	0	1	0	0	5
15	1	1	0	0	1	1	1	1	1	0	1	1	1	1	0	1	0	0	1	0	1	0	0	0	0	0	0	1	1	0	0	15
16	2	0	0	1	1	1	1	0	0	1	1	1	0	1	0	0	0	0	1	1	1	1	1	0	1	0	1	0	1	1	0	15
17	2	0	0	1	0	1	1	1	0	0	1	1	1	1	0	0	0	0	1	0	0	1	0	0	1	0	0	1	0	1	0	17
18	2	1	1	1	1	0	1	1	0	1	1	1	1	1	0	0	0	1	0	1	0	0	1	1	1	1	0	1	0	1	0	13
19	2	0	1	1	1	1	1	1	0	0	1	1	1	1	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	1	0	19
20	2	1	0	1	1	0	0	1	0	0	1	1	1	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	14
21	2	1	0	1	1	1	1	1	0	0	1	1	1	1	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	1	0	11
22	2	0	0	1	1	0	0	1	0	1	1	1	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	14
23	2	0	1	0	1	1	1	1	0	1	0	0	0	1	0	0	0	0	0	1	0	0	1	0	1	1	1	1	1	0	1	9
24	2	0	1	0	1	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	0	1	15
25	2	1	0	1	1	1	1	1	0	0	1	1	1	1	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	1	0	14
26	2	1	0	1	1	1	1	1	0	1	0	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	1	14
27	2	0	1	0	1	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	0	1	13
28	2	0	1	0	1	1	1	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	1	1	1	1	1	0	1	14
29	2	0	1	0	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	1	1	1	1	0	1	14
30	2	0	1	1	1	0	0	1	1	0	0	1	1	1	1	1	0	0	1	0	0	0	0	1	1	1	1	1	1	0	1	13
2,000	2	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	7
รวม		1132	1178	1216	1335	1263	1285	1107	1019	1168	856	1110	1082	976	851	805	740	558	959	993	550	986	1059	1106	1170	870	1047	782	709	618	732	1132

หมายเหตุ 1 คือ เพศชาย 2 คือ เพศหญิง

ภาคผนวก ง
ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ
ด้วยวิธี IRT-LR วิธี SIBTEST และวิธี MH

ตารางที่ ง-10 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ ด้านเหตุผล ด้วยวิธี IRT-LR ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างขนาดใหญ่ (300 คน)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี IRT-LR		
	กลุ่มตัวอย่าง (300 คน)		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
1	0.22(NO-DIF)	0.91(NO-DIF)	0.10(NO-DIF)
2	0.01(DIF)	0.84(NO-DIF)	0.89(NO-DIF)
3	0.30(NO-DIF)	0.79(NO-DIF)	0.01(DIF)
4	0.33(NO-DIF)	0.76(NO-DIF)	0.20(NO-DIF)
5	0.73(NO-DIF)	0.89(NO-DIF)	0.12(NO-DIF)
6	0.10(NO-DIF)	0.76(NO-DIF)	0.56(NO-DIF)
7	0.20(NO-DIF)	0.27(NO-DIF)	0.87(NO-DIF)
8	0.53(NO-DIF)	0.20(NO-DIF)	0.95(NO-DIF)
9	0.76(NO-DIF)	0.72(NO-DIF)	0.28(NO-DIF)
10	0.35(NO-DIF)	0.61(NO-DIF)	0.50(NO-DIF)
11	0.42(NO-DIF)	0.02(DIF)	0.17(NO-DIF)
12	0.34(NO-DIF)	0.04(DIF)	0.05(DIF)
13	0.54(NO-DIF)	0.92(NO-DIF)	0.11(NO-DIF)
14	0.88(NO-DIF)	0.79(NO-DIF)	0.05(DIF)
15	0.18(NO-DIF)	0.39(NO-DIF)	0.72(NO-DIF)
16	0.55(NO-DIF)	0.25(NO-DIF)	0.26(NO-DIF)
17	0.03(DIF)	0.30(NO-DIF)	0.72(NO-DIF)
18	0.21(NO-DIF)	0.44(NO-DIF)	0.06(NO-DIF)
19	0.07(NO-DIF)	0.30(NO-DIF)	0.56(NO-DIF)
20	0.77(NO-DIF)	0.13(NO-DIF)	0.08(NO-DIF)
21	0.76(NO-DIF)	0.00(DIF)	0.06(NO-DIF)
22	0.40(NO-DIF)	0.92(NO-DIF)	0.88(NO-DIF)
23	0.72(NO-DIF)	0.39(NO-DIF)	0.70(NO-DIF)
24	0.19(NO-DIF)	0.58(NO-DIF)	0.86(NO-DIF)
25	0.96(NO-DIF)	0.59(NO-DIF)	0.28(NO-DIF)
26	0.75(NO-DIF)	0.30(NO-DIF)	0.30(NO-DIF)
27	0.63(NO-DIF)	0.87(NO-DIF)	0.94(NO-DIF)
28	0.13(NO-DIF)	0.28(NO-DIF)	0.18(NO-DIF)
29	0.64(NO-DIF)	0.69(NO-DIF)	0.36(NO-DIF)
30	0.02(DIF)	0.28(NO-DIF)	0.13(NO-DIF)
จำนวนข้อที่พบ DIF	(3 ข้อ) 10%	(3 ข้อ) 10%	(3 ข้อ) 10%

ตารางที่ ง-11 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ ด้านเหตุผล ด้วยวิธี IRT-LR ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างขนาดใหญ่ (1,000 คน)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี IRT-LR		
	กลุ่มตัวอย่าง (1,000 คน)		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
1	0.97(NO-DIF)	0.22(NO-DIF)	0.00(DIF)
2	0.12(NO-DIF)	0.99(NO-DIF)	0.63(NO-DIF)
3	0.76(NO-DIF)	0.44(NO-DIF)	0.01(DIF)
4	0.42(NO-DIF)	0.56(NO-DIF)	0.90(NO-DIF)
5	0.00(DIF)	0.34(NO-DIF)	0.59(NO-DIF)
6	0.55(NO-DIF)	0.58(NO-DIF)	0.45(NO-DIF)
7	0.88(NO-DIF)	0.41(NO-DIF)	0.11(NO-DIF)
8	0.99(NO-DIF)	0.33(NO-DIF)	0.83(NO-DIF)
9	0.78(NO-DIF)	0.56(NO-DIF)	0.07(NO-DIF)
10	0.28(NO-DIF)	0.03(DIF)	0.29(NO-DIF)
11	0.78(NO-DIF)	0.96(NO-DIF)	0.00(DIF)
12	0.23(NO-DIF)	0.03(DIF)	0.00(DIF)
13	0.42(NO-DIF)	0.20(NO-DIF)	0.11(NO-DIF)
14	0.75(NO-DIF)	0.57(NO-DIF)	0.09(NO-DIF)
15	0.01(DIF)	0.96(NO-DIF)	0.74(NO-DIF)
16	0.37(NO-DIF)	0.05(DIF)	0.71(NO-DIF)
17	0.24(NO-DIF)	0.10(NO-DIF)	0.89(NO-DIF)
18	0.01(DIF)	0.72(NO-DIF)	0.23(NO-DIF)
19	0.50(NO-DIF)	0.04(DIF)	0.15(NO-DIF)
20	0.23(NO-DIF)	0.36(NO-DIF)	0.15(NO-DIF)
21	0.14(NO-DIF)	0.48(NO-DIF)	0.15(NO-DIF)
22	0.24(NO-DIF)	0.22(NO-DIF)	0.93(NO-DIF)
23	0.70(NO-DIF)	0.36(NO-DIF)	0.30(NO-DIF)
24	0.44(NO-DIF)	0.06(NO-DIF)	0.05(DIF)
25	0.97(NO-DIF)	0.48(NO-DIF)	0.20(NO-DIF)
26	0.60(NO-DIF)	0.27(NO-DIF)	0.76(NO-DIF)
27	0.99(NO-DIF)	0.97(NO-DIF)	0.85(NO-DIF)
28	0.79(NO-DIF)	0.57(NO-DIF)	0.98(NO-DIF)
29	0.07(NO-DIF)	0.90(NO-DIF)	0.86(NO-DIF)
30	0.01(DIF)	0.08(NO-DIF)	0.08(NO-DIF)
จำนวนข้อที่พบ DIF	(4 ข้อ) 13%	(4 ข้อ) 13%	(5 ข้อ) 17%

ตารางที่ ง-12 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ ด้านเหตุผล ด้วยวิธี IRT-LR ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างขนาดใหญ่ (2,000 คน)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี IRT-LR		
	กลุ่มตัวอย่าง (2,000 คน)		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
1	0.11(NO-DIF)	0.22(NO-DIF)	0.00(DIF)
2	0.07(NO-DIF)	0.82(NO-DIF)	0.15(NO-DIF)
3	0.21(NO-DIF)	0.18(NO-DIF)	0.01(DIF)
4	0.39(NO-DIF)	0.00(DIF)	0.39(NO-DIF)
5	0.00(DIF)	0.14(NO-DIF)	0.10(NO-DIF)
6	0.45(NO-DIF)	0.91(NO-DIF)	0.20(NO-DIF)
7	0.49(NO-DIF)	0.20(NO-DIF)	0.11(NO-DIF)
8	0.70(NO-DIF)	0.10(NO-DIF)	0.15(NO-DIF)
9	0.37(NO-DIF)	0.26(NO-DIF)	0.00(DIF)
10	0.95(NO-DIF)	0.06(NO-DIF)	0.29(NO-DIF)
11	0.68(NO-DIF)	0.98(NO-DIF)	0.00(DIF)
12	0.39(NO-DIF)	0.02(DIF)	0.00(DIF)
13	0.59(NO-DIF)	0.15(NO-DIF)	0.19(NO-DIF)
14	0.68(NO-DIF)	0.12(NO-DIF)	0.00(DIF)
15	0.06(NO-DIF)	0.54(NO-DIF)	0.79(NO-DIF)
16	0.20(NO-DIF)	0.04(DIF)	0.11(NO-DIF)
17	0.22(NO-DIF)	0.06(NO-DIF)	0.13(NO-DIF)
18	0.69(NO-DIF)	0.95(NO-DIF)	0.92(NO-DIF)
19	0.14(NO-DIF)	0.06(NO-DIF)	0.00(DIF)
20	0.24(NO-DIF)	0.52(NO-DIF)	0.24(NO-DIF)
21	0.79(NO-DIF)	0.06(NO-DIF)	0.01(DIF)
22	0.41(NO-DIF)	0.58(NO-DIF)	0.31(NO-DIF)
23	0.79(NO-DIF)	0.36(NO-DIF)	0.67(NO-DIF)
24	0.07(NO-DIF)	0.08(NO-DIF)	0.15(NO-DIF)
25	0.92(NO-DIF)	0.79(NO-DIF)	0.08(NO-DIF)
26	0.17(NO-DIF)	0.37(NO-DIF)	0.12(NO-DIF)
27	0.17(NO-DIF)	0.29(NO-DIF)	0.22(NO-DIF)
28	0.45(NO-DIF)	0.055(NO-DIF)	0.52(NO-DIF)
29	0.03(DIF)	0.68(NO-DIF)	0.03(DIF)
30	0.01(DIF)	0.37(NO-DIF)	0.00(DIF)
จำนวนข้อที่พบ DIF	(3 ข้อ) 10%	(3 ข้อ) 10%	(10 ข้อ) 33%

ตารางที่ ง-13 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ ด้านเหตุผล ด้วยวิธี SIBTEST ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างขนาดใหญ่ (300 คน)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี SIBTEST		
	กลุ่มตัวอย่าง (300 คน)		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
1	0.11(NO-DIF)	0.81(NO-DIF)	0.58(NO-DIF)
2	0.09(NO-DIF)	0.77(NO-DIF)	0.61(NO-DIF)
3	0.42(NO-DIF)	0.53(NO-DIF)	0.07(NO-DIF)
4	0.42(NO-DIF)	0.49(NO-DIF)	0.19(NO-DIF)
5	0.94(NO-DIF)	0.90(NO-DIF)	0.35(NO-DIF)
6	0.54(NO-DIF)	0.53(NO-DIF)	0.54(NO-DIF)
7	0.67(NO-DIF)	0.60(NO-DIF)	0.49(NO-DIF)
8	0.45(NO-DIF)	0.08(NO-DIF)	0.37(NO-DIF)
9	0.50(NO-DIF)	0.60(NO-DIF)	0.69(NO-DIF)
10	0.96(NO-DIF)	0.17(NO-DIF)	0.61(NO-DIF)
11	0.26(NO-DIF)	0.67(NO-DIF)	0.14(NO-DIF)
12	0.16(NO-DIF)	0.03(DIF)	0.10(NO-DIF)
13	0.86(NO-DIF)	0.18(NO-DIF)	0.14(NO-DIF)
14	0.93(NO-DIF)	0.97(NO-DIF)	0.09(NO-DIF)
15	0.71(NO-DIF)	0.45(NO-DIF)	0.90(NO-DIF)
16	0.64(NO-DIF)	0.10(NO-DIF)	0.19(NO-DIF)
17	0.10(NO-DIF)	0.88(NO-DIF)	0.95(NO-DIF)
18	0.52(NO-DIF)	0.79(NO-DIF)	0.05(DIF)
19	0.01(DIF)	0.36(NO-DIF)	0.11(NO-DIF)
20	0.94(NO-DIF)	0.77(NO-DIF)	0.05(DIF)
21	0.15(NO-DIF)	0.02(DIF)	0.07(NO-DIF)
22	0.77(NO-DIF)	0.86(NO-DIF)	0.43(NO-DIF)
23	0.57(NO-DIF)	0.36(NO-DIF)	0.75(NO-DIF)
24	0.62(NO-DIF)	0.80(NO-DIF)	0.46(NO-DIF)
25	0.35(NO-DIF)	0.30(NO-DIF)	0.30(NO-DIF)
26	0.46(NO-DIF)	0.04(DIF)	0.77(NO-DIF)
27	0.42(NO-DIF)	0.77(NO-DIF)	0.92(NO-DIF)
28	0.10(NO-DIF)	0.13(NO-DIF)	0.32(NO-DIF)
29	0.36(NO-DIF)	0.61(NO-DIF)	0.20(NO-DIF)
30	0.00(DIF)	0.16(NO-DIF)	0.18(NO-DIF)
จำนวนข้อที่พบ DIF	(2 ข้อ) 7%	(3 ข้อ) 10%	(2 ข้อ) 7%

ตารางที่ ง-14 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ ด้านเหตุผล ด้วยวิธี SIBTEST ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างขนาดใหญ่ (1,000 คน)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี SIBTEST		
	กลุ่มตัวอย่าง (1,000 คน)		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
1	0.88(NO-DIF)	0.71(NO-DIF)	0.00(DIF)
2	0.85(NO-DIF)	0.76(NO-DIF)	0.31(NO-DIF)
3	0.53(NO-DIF)	0.27(NO-DIF)	0.72(NO-DIF)
4	0.26(NO-DIF)	0.28(NO-DIF)	0.71(NO-DIF)
5	0.00(DIF)	0.47(NO-DIF)	0.72(NO-DIF)
6	0.46(NO-DIF)	0.74(NO-DIF)	0.56(NO-DIF)
7	0.88(NO-DIF)	0.91(NO-DIF)	0.21(NO-DIF)
8	0.99(NO-DIF)	0.49(NO-DIF)	0.83(NO-DIF)
9	0.95(NO-DIF)	0.19(NO-DIF)	0.02(DIF)
10	0.43(NO-DIF)	0.08(NO-DIF)	0.76(NO-DIF)
11	0.90(NO-DIF)	0.93(NO-DIF)	0.00(DIF)
12	0.09(NO-DIF)	0.33(NO-DIF)	0.00(DIF)
13	0.27(NO-DIF)	0.12(NO-DIF)	0.24(NO-DIF)
14	0.61(NO-DIF)	0.50(NO-DIF)	0.59(NO-DIF)
15	0.15(NO-DIF)	0.87(NO-DIF)	0.86(NO-DIF)
16	0.07(NO-DIF)	0.07(NO-DIF)	0.43(NO-DIF)
17	0.11(NO-DIF)	0.13(NO-DIF)	0.70(NO-DIF)
18	0.30(NO-DIF)	0.93(NO-DIF)	0.30(NO-DIF)
19	0.48(NO-DIF)	0.04(DIF)	0.01(DIF)
20	0.25(NO-DIF)	0.10(NO-DIF)	0.05(DIF)
21	0.33(NO-DIF)	0.71(NO-DIF)	0.17(NO-DIF)
22	0.09(NO-DIF)	0.12(NO-DIF)	0.52(NO-DIF)
23	0.62(NO-DIF)	0.76(NO-DIF)	0.45(NO-DIF)
24	0.23(NO-DIF)	0.65(NO-DIF)	0.58(NO-DIF)
25	0.80(NO-DIF)	0.30(NO-DIF)	0.06(NO-DIF)
26	0.29(NO-DIF)	0.28(NO-DIF)	0.25(NO-DIF)
27	0.91(NO-DIF)	0.52(NO-DIF)	0.53(NO-DIF)
28	0.98(NO-DIF)	0.24(NO-DIF)	0.93(NO-DIF)
29	0.07(NO-DIF)	0.22(NO-DIF)	0.88(NO-DIF)
30	0.00(DIF)	0.64(NO-DIF)	0.08(NO-DIF)
จำนวนข้อที่พบ DIF	(2 ข้อ) 7%	(1 ข้อ) 3%	(6 ข้อ) 20%

ตารางที่ ง-15 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ ด้านเหตุผล ด้วยวิธี SIBTEST ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างขนาดใหญ่ (2,000 คน)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี SIBTEST		
	กลุ่มตัวอย่าง (2,000 คน)		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
1	0.14(NO-DIF)	0.35(NO-DIF)	0.00(DIF)
2	0.35(NO-DIF)	0.87(NO-DIF)	0.21(NO-DIF)
3	0.10(NO-DIF)	0.61(NO-DIF)	0.26(NO-DIF)
4	0.63(NO-DIF)	0.04(DIF)	0.25(NO-DIF)
5	0.00(DIF)	0.45(NO-DIF)	0.08(NO-DIF)
6	0.28(NO-DIF)	0.84(NO-DIF)	0.34(NO-DIF)
7	0.68(NO-DIF)	0.23(NO-DIF)	0.14(NO-DIF)
8	0.46(NO-DIF)	0.24(NO-DIF)	0.45(NO-DIF)
9	0.10(NO-DIF)	0.00(DIF)	0.00(DIF)
10	0.78(NO-DIF)	0.42(NO-DIF)	0.29(NO-DIF)
11	0.86(NO-DIF)	0.85(NO-DIF)	0.00(DIF)
12	0.13(NO-DIF)	0.33(NO-DIF)	0.00(DIF)
13	0.87(NO-DIF)	0.02(DIF)	0.08(NO-DIF)
14	0.46(NO-DIF)	0.07(NO-DIF)	0.00(DIF)
15	0.14(NO-DIF)	0.26(NO-DIF)	0.41(NO-DIF)
16	0.07(NO-DIF)	0.04(DIF)	0.38(NO-DIF)
17	0.31(NO-DIF)	0.37(NO-DIF)	0.34(NO-DIF)
18	0.98(NO-DIF)	0.75(NO-DIF)	0.54(NO-DIF)
19	0.47(NO-DIF)	0.00(DIF)	0.00(DIF)
20	0.29(NO-DIF)	0.15(NO-DIF)	0.05(DIF)
21	0.64(NO-DIF)	0.41(NO-DIF)	0.02(DIF)
22	0.17(NO-DIF)	0.67(NO-DIF)	0.67(NO-DIF)
23	0.89(NO-DIF)	0.86(NO-DIF)	0.26(NO-DIF)
24	0.03(DIF)	0.96(NO-DIF)	0.74(NO-DIF)
25	0.58(NO-DIF)	0.78(NO-DIF)	0.02(DIF)
26	0.33(NO-DIF)	0.43(NO-DIF)	0.03(DIF)
27	0.38(NO-DIF)	0.41(NO-DIF)	0.29(NO-DIF)
28	0.51(NO-DIF)	0.02(DIF)	0.82(NO-DIF)
29	0.02(DIF)	0.95(NO-DIF)	0.02(DIF)
30	0.00(DIF)	0.19(NO-DIF)	0.00(DIF)
จำนวนข้อที่พบ DIF	(4 ข้อ) 13%	(6 ข้อ) 20%	(12 ข้อ) 40%

ตารางที่ ง-16 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ ด้านเหตุผล ด้วยวิธี MH ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างขนาดใหญ่ (300 คน)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี MH		
	กลุ่มตัวอย่าง (300 คน)		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
1	0.42(NO-DIF)	0.69(NO-DIF)	0.35(NO-DIF)
2	0.00(DIF)	0.53(NO-DIF)	0.41(NO-DIF)
3	0.02(DIF)	0.72(NO-DIF)	0.08(NO-DIF)
4	0.02(DIF)	0.82(NO-DIF)	0.01(DIF)
5	0.49(NO-DIF)	1.00(NO-DIF)	0.06(NO-DIF)
6	0.71(NO-DIF)	0.90(NO-DIF)	0.19(NO-DIF)
7	0.06(NO-DIF)	0.15(NO-DIF)	0.24(NO-DIF)
8	0.73(NO-DIF)	0.13(NO-DIF)	0.49(NO-DIF)
9	0.68(NO-DIF)	0.56(NO-DIF)	0.82(NO-DIF)
10	0.41(NO-DIF)	0.56(NO-DIF)	0.16(NO-DIF)
11	0.10(NO-DIF)	0.91(NO-DIF)	0.02(DIF)
12	0.91(NO-DIF)	0.04(DIF)	0.01(DIF)
13	0.56(NO-DIF)	0.82(NO-DIF)	0.02(DIF)
14	0.30(NO-DIF)	0.81(NO-DIF)	0.09(NO-DIF)
15	0.64(NO-DIF)	0.48(NO-DIF)	0.82(NO-DIF)
16	0.08(NO-DIF)	0.17(NO-DIF)	0.18(NO-DIF)
17	0.00(DIF)	0.35(NO-DIF)	1.00(NO-DIF)
18	0.73(NO-DIF)	0.41(NO-DIF)	0.20(NO-DIF)
19	0.00(DIF)	0.16(NO-DIF)	0.10(NO-DIF)
20	0.18(NO-DIF)	0.11(NO-DIF)	0.07(NO-DIF)
21	1.00(NO-DIF)	0.00(DIF)	0.64(NO-DIF)
22	0.89(NO-DIF)	0.90(NO-DIF)	0.30(NO-DIF)
23	0.60(NO-DIF)	0.20(NO-DIF)	0.36(NO-DIF)
24	0.60(NO-DIF)	0.55(NO-DIF)	0.16(NO-DIF)
25	1.00(NO-DIF)	0.47(NO-DIF)	0.30(NO-DIF)
26	0.25(NO-DIF)	0.15(NO-DIF)	0.02(DIF)
27	0.91(NO-DIF)	0.82(NO-DIF)	0.62(NO-DIF)
28	0.41(NO-DIF)	0.13(NO-DIF)	0.06(NO-DIF)
29	0.73(NO-DIF)	0.91(NO-DIF)	0.07(NO-DIF)
30	0.13(NO-DIF)	0.12(NO-DIF)	0.48(NO-DIF)
จำนวนข้อที่พบ DIF	(5 ข้อ) 17%	(2 ข้อ) 7%	(5 ข้อ) 17%

ตารางที่ ง-17 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ ด้านเหตุผล ด้วยวิธี MH ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างขนาดใหญ่ (1,000 คน)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี MH		
	กลุ่มตัวอย่าง (1,000 คน)		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
1	0.25(NO-DIF)	0.94(NO-DIF)	0.30(NO-DIF)
2	0.74(NO-DIF)	0.56(NO-DIF)	0.15(NO-DIF)
3	0.03(DIF)	0.29(NO-DIF)	0.26(NO-DIF)
4	0.02(DIF)	0.70(NO-DIF)	0.01(DIF)
5	0.24(NO-DIF)	0.36(NO-DIF)	0.04(DIF)
6	0.57(NO-DIF)	0.28(NO-DIF)	0.03(DIF)
7	0.01(DIF)	0.57(NO-DIF)	0.41(NO-DIF)
8	0.15(NO-DIF)	0.27(NO-DIF)	0.16(NO-DIF)
9	0.03(DIF)	0.45(NO-DIF)	0.95(NO-DIF)
10	0.11(NO-DIF)	0.13(NO-DIF)	0.00(DIF)
11	0.24(NO-DIF)	0.66(NO-DIF)	0.00(DIF)
12	0.80(NO-DIF)	0.23(NO-DIF)	0.00(DIF)
13	0.05(DIF)	0.04(DIF)	0.01(DIF)
14	0.23(NO-DIF)	0.89(NO-DIF)	0.04(DIF)
15	0.06(NO-DIF)	0.75(NO-DIF)	0.48(NO-DIF)
16	0.01(DIF)	0.01(DIF)	0.60(NO-DIF)
17	0.00(DIF)	0.02(DIF)	0.46(NO-DIF)
18	0.01(DIF)	0.23(NO-DIF)	0.00(DIF)
19	0.00(DIF)	0.01(DIF)	0.00(DIF)
20	0.31(NO-DIF)	0.18(NO-DIF)	0.06(NO-DIF)
21	1.38(NO-DIF)	0.60(NO-DIF)	0.85(NO-DIF)
22	0.61(NO-DIF)	0.14(NO-DIF)	0.01(DIF)
23	0.74(NO-DIF)	0.42(NO-DIF)	0.00(DIF)
24	0.00(DIF)	0.61(NO-DIF)	0.00(DIF)
25	0.74(NO-DIF)	0.61(NO-DIF)	0.44(NO-DIF)
26	0.11(NO-DIF)	0.36(NO-DIF)	0.00(DIF)
27	0.13(NO-DIF)	0.66(NO-DIF)	0.08(NO-DIF)
28	0.08(NO-DIF)	0.29(NO-DIF)	0.04(DIF)
29	0.00(DIF)	0.48(NO-DIF)	0.34(NO-DIF)
30	0.34(NO-DIF)	0.36(NO-DIF)	0.90(NO-DIF)
จำนวนข้อที่พบ DIF	(11 ข้อ) 37%	(4ข้อ) 13%	(15 ข้อ) 50%

ตารางที่ ง-18 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ชั้นประถมศึกษาปีที่ 3 ด้านภาษา ด้านคำนวณ ด้านเหตุผล ด้วยวิธี MH ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างขนาดใหญ่ (2,000 คน)

ข้อที่	ผลการตรวจสอบ DIF ด้วยวิธี MH		
	กลุ่มตัวอย่าง (2,000 คน)		
	ด้านภาษา	ด้านคำนวณ	ด้านเหตุผล
1	0.59(NO-DIF)	0.64(NO-DIF)	1.00(NO-DIF)
2	0.92(NO-DIF)	0.14(NO-DIF)	0.00(DIF)
3	0.00(DIF)	0.32(NO-DIF)	0.01(DIF)
4	0.08(DIF)	0.45(NO-DIF)	0.00(DIF)
5	0.01(DIF)	0.21(NO-DIF)	0.00(DIF)
6	0.11(NO-DIF)	0.09(NO-DIF)	0.00(DIF)
7	0.00(DIF)	0.85(NO-DIF)	0.01(DIF)
8	0.09(NO-DIF)	0.00(DIF)	0.01(DIF)
9	0.00(DIF)	0.00(DIF)	0.93(NO-DIF)
10	0.32(NO-DIF)	0.65(NO-DIF)	0.00(DIF)
11	0.01(DIF)	0.08(NO-DIF)	0.00(DIF)
12	0.03(DIF)	0.24(NO-DIF)	0.00(DIF)
13	0.14NO-DIF	0.00(DIF)	0.00(DIF)
14	0.02(DIF)	0.82(NO-DIF)	0.00(DIF)
15	0.01(DIF)	0.06(NO-DIF)	0.00(DIF)
16	0.00(DIF)	0.00(DIF)	0.20(NO-DIF)
17	0.00(DIF)	0.00(DIF)	0.01(DIF)
18	0.01(DIF)	0.05(NO-DIF)	0.00(DIF)
19	0.00(DIF)	0.00(DIF)	0.00(DIF)
20	0.01(DIF)	0.02(DIF)	0.19(NO-DIF)
21	0.04(DIF)	0.58(NO-DIF)	0.25(NO-DIF)
22	0.58(NO-DIF)	0.63(NO-DIF)	0.00(DIF)
23	0.30(NO-DIF)	0.25(NO-DIF)	0.00(DIF)
24	0.00(DIF)	0.32(NO-DIF)	0.00(DIF)
25	0.46(NO-DIF)	0.17(NO-DIF)	0.93(NO-DIF)
26	0.11(NO-DIF)	0.82(NO-DIF)	0.00(DIF)
27	0.00(DIF)	0.82(NO-DIF)	0.00(DIF)
28	0.00(DIF)	0.01(DIF)	0.00(DIF)
29	0.00(DIF)	0.62(NO-DIF)	0.85(NO-DIF)
30	0.89(NO-DIF)	0.60(NO-DIF)	0.78(NO-DIF)
จำนวนข้อที่พบ DIF	(19 ข้อ) 63%	(8 ข้อ) 27%	(22 ข้อ) 73%

ภาคผนวก จ
ตัวอย่าง Print Out การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี IRT-LR

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

Project:	Literacy 2000
Description:	
Date:	06 August 2017
Time:	02:15 PM

Table of Contents

2PL Model Item Parameter Estimates for Group 1, logit: $a\theta + c$ or $a(\theta - b)$ (Back to TOC)

Item	Label		*a*	*s.e.*		*c*	*s.e.*	*b*	*s.e.*
1	Litera1	2	0.63	0.09	1	0.35	0.08	-0.57	0.11
2	Litera2	4	0.44	0.08	3	-0.63	0.07	1.44	0.36
3	Litera3	6	0.67	0.09	5	0.49	0.08	-0.73	0.11
4	Litera4	8	0.92	0.11	7	1.50	0.11	-1.63	0.14
5	Litera5	10	0.82	0.10	9	0.77	0.09	-0.94	0.10
6	Litera6	12	0.93	0.11	11	1.50	0.11	-1.60	0.14
7	Litera7	14	1.23	0.12	13	1.28	0.11	-1.04	0.08
8	Litera8	16	0.65	0.09	15	0.09	0.07	-0.13	0.11
9	Litera9	18	0.80	0.11	17	1.56	0.11	-1.94	0.19
10	Litera10	20	0.33	0.08	19	-0.30	0.07	0.88	0.34
11	Litera11	22	0.64	0.09	21	1.03	0.09	-1.61	0.19
12	Litera12	24	1.07	0.11	23	0.67	0.09	-0.62	0.07
13	Litera13	26	0.38	0.08	25	-0.26	0.07	0.67	0.27
14	Litera14	28	0.87	0.10	27	0.05	0.08	-0.06	0.09
15	Litera15	30	0.51	0.08	29	-0.25	0.07	0.49	0.18
16	Litera16	32	0.75	0.09	31	0.02	0.07	-0.03	0.10
17	Litera17	34	1.09	0.11	33	1.11	0.10	-1.02	0.08
18	Litera18	36	0.64	0.09	35	-0.10	0.07	0.16	0.12
19	Litera19	38	1.11	0.12	37	1.77	0.13	-1.59	0.12
20	Litera20	40	0.80	0.10	39	0.98	0.09	-1.22	0.12
21	Litera21	42	0.61	0.09	41	0.13	0.07	-0.22	0.11
22	Litera22	44	0.56	0.09	43	-0.92	0.08	1.66	0.33

23	Litera23	[46]	0.32	0.08	[45]	-0.63	0.07	1.98	0.61
24	Litera24	[48]	0.57	0.09	[47]	-0.08	0.07	0.13	0.13
25	Litera25	[50]	0.18	0.08	[49]	-0.49	0.07	2.66	1.31
26	Litera26	[52]	0.32	0.08	[51]	-0.06	0.07	0.17	0.23
27	Litera27	[54]	0.45	0.08	[53]	-0.32	0.07	0.72	0.24
28	Litera28	[56]	0.65	0.09	[55]	0.38	0.08	-0.58	0.11
29	Litera29	[58]	0.60	0.09	[57]	-0.24	0.07	0.41	0.15
30	Litera30	[60]	0.97	0.10	[59]	0.69	0.09	-0.72	0.08

Summed-Score Based Item Diagnostic Tables and X^2s for Group 1 (Back to TOC)

S-X^2 Item Level Diagnostic Statistics

Item	Label	X^2	*d.f.*	Probability
1	Litera1	15.68	20	0.7369
2	Litera2	21.82	20	0.3523
3	Litera3	38.89	20	0.0069
4	Litera4	20.84	19	0.3477
5	Litera5	21.51	19	0.3084
6	Litera6	22.38	19	0.2648
7	Litera7	14.79	18	0.6773
8	Litera8	27.45	21	0.1561
9	Litera9	17.87	19	0.5323
10	Litera10	21.61	22	0.4850
11	Litera11	25.20	20	0.1934
12	Litera12	23.25	18	0.1806
13	Litera13	20.01	21	0.5219
14	Litera14	23.04	19	0.2351
15	Litera15	26.91	20	0.1375
16	Litera16	22.88	19	0.2418
17	Litera17	31.80	18	0.0231
18	Litera18	28.36	20	0.1009
19	Litera19	32.64	18	0.0184
20	Litera20	27.96	20	0.1100
21	Litera21	23.68	21	0.3081
22	Litera22	26.71	20	0.1432
23	Litera23	27.35	20	0.1254
24	Litera24	17.33	20	0.6326
25	Litera25	24.59	21	0.2645
26	Litera26	24.21	21	0.2821
27	Litera27	22.29	21	0.3844
28	Litera28	11.40	20	0.9353
29	Litera29	22.95	20	0.2903
30	Litera30	23.41	20	0.2683

2PL Model Item Parameter Estimates for Group 2, logit: $a\theta + c$ or $a(\theta - b)$ (Back to TOC)

Item	Label	*a*	*s.e.*	*c*	*s.e.*	*b*	*s.e.*
1	Litera1	[62] 0.45	0.08	[61] 0.17	0.07	-0.39	0.16
2	Litera2	[64] 0.20	0.08	[63] -0.75	0.07	3.80	1.55
3	Litera3	[66] 0.65	0.09	[65] 0.66	0.07	-1.00	0.16
4	Litera4	[68] 0.71	0.10	[67] 1.40	0.09	-1.98	0.26
5	Litera5	[70] 0.56	0.08	[69] 0.19	0.07	-0.35	0.13
6	Litera6	[72] 0.92	0.11	[71] 1.34	0.09	-1.46	0.16
7	Litera7	[74] 1.38	0.14	[73] 1.47	0.11	-1.06	0.09
8	Litera8	[76] 0.68	0.09	[75] 0.01	0.07	-0.02	0.10
9	Litera9	[78] 0.85	0.12	[77] 1.75	0.10	-2.06	0.24
10	Litera10	[80] 0.34	0.08	[79] -0.32	0.07	0.95	0.28
11	Litera11	[82] 0.76	0.10	[81] 1.09	0.08	-1.45	0.18
12	Litera12	[84] 1.07	0.11	[83] 0.52	0.08	-0.48	0.08
13	Litera13	[86] 0.27	0.08	[85] -0.25	0.06	0.92	0.35
14	Litera14	[88] 0.78	0.09	[87] -0.03	0.07	0.03	0.09
15	Litera15	[90] 0.29	0.08	[89] -0.16	0.06	0.56	0.27
16	Litera16	[92] 0.69	0.09	[91] 0.18	0.07	-0.26	0.10
17	Litera17	[94] 1.29	0.13	[93] 1.37	0.10	-1.06	0.09
18	Litera18	[96] 0.75	0.09	[95] -0.09	0.07	0.13	0.09
19	Litera19	[98] 1.45	0.16	[97] 2.14	0.14	-1.48	0.11
20	Litera20	[100] 1.05	0.11	[99] 1.03	0.09	-0.99	0.10
21	Litera21	[102] 0.67	0.09	[101] 0.11	0.07	-0.16	0.10
22	Litera22	[104] 0.56	0.09	[103] -1.07	0.08	1.91	0.31
23	Litera23	[106] 0.39	0.08	[105] -0.65	0.07	1.67	0.37
24	Litera24	[108] 0.55	0.08	[107] 0.14	0.07	-0.25	0.12
25	Litera25	[110] 0.23	0.08	[109] -0.49	0.07	2.13	0.76
26	Litera26	[112] 0.14	0.07	[111] -0.02	0.06	0.17	0.47
27	Litera27	[114] 0.63	0.09	[113] -0.20	0.07	0.31	0.11
28	Litera28	[116] 0.80	0.10	[115] 0.47	0.07	-0.59	0.10
29	Litera29	[118] 0.57	0.08	[117] 0.00	0.07	-0.00	0.12
30	Litera30	[120] 0.80	0.09	[119] 0.34	0.07	-0.43	0.09

Summed-Score Based Item Diagnostic Tables and X^2s for Group 2 (Back to TOC)

S-X^2 Item Level Diagnostic Statistics

Item	Label	X^2	*d.f.*	Probability
1	Litera1	23.75	20	0.2526
2	Litera2	30.56	20	0.0611
3	Litera3	18.53	19	0.4887
4	Litera4	18.60	18	0.4186
5	Litera5	20.78	20	0.4120
6	Litera6	14.29	18	0.7110
7	Litera7	21.18	18	0.2699
8	Litera8	23.44	20	0.2668
9	Litera9	19.97	19	0.3982
10	Litera10	25.42	20	0.1855
11	Litera11	15.38	19	0.6987
12	Litera12	13.58	18	0.7569
13	Litera13	21.96	20	0.3446

14	Litera14	26.26	19	0.1227
15	Litera15	18.50	20	0.5561
16	Litera16	26.63	19	0.1133
17	Litera17	26.12	18	0.0969
18	Litera18	18.41	19	0.4965
19	Litera19	20.65	17	0.2416
20	Litera20	19.75	18	0.3490
21	Litera21	16.95	19	0.5943
22	Litera22	26.40	19	0.1190
23	Litera23	18.91	19	0.4639
24	Litera24	22.33	20	0.3221
25	Litera25	25.45	20	0.1841
26	Litera26	17.25	20	0.6376
27	Litera27	36.60	19	0.0089
28	Litera28	13.37	19	0.8196
29	Litera29	14.74	20	0.7917
30	Litera30	22.12	19	0.2777

Group Parameter Estimates (Back to TOC)

Group	Label	μ	*s.e.*	σ^2	*s.e.*	σ	*s.e.*
1	G1	-0.36	-----	0.98	-----	0.99	-----
2	G2	0.00	-----	1.00	-----	1.00	-----

DIF Statistics for Graded Items (Back to TOC)

Item numbers in:

Group 1	Group 2	Total X^2	*d.f.*	*p*	X^2_a	*d.f.*	*p*	$X^2_{c\|a}$	*d.f.*	*p*
1	1	4.4	2	0.1124	2.3	1	0.1324	2.1	1	0.1489
2	2	5.2	2	0.0729	4.2	1	0.0411	1.1	1	0.3028
3	3	3.1	2	0.2122	0.0	1	0.9159	3.1	1	0.0782
4	4	1.9	2	0.3870	1.9	1	0.1713	0.0	1	0.8791
5	5	28.4	2	0.0001	4.3	1	0.0372	24.1	1	0.0001
6	6	1.6	2	0.4463	0.0	1	0.9466	1.6	1	0.2050
7	7	1.4	2	0.4880	0.6	1	0.4415	0.8	1	0.3590
8	8	0.7	2	0.6976	0.1	1	0.7877	0.6	1	0.4212
9	9	2.0	2	0.3684	0.1	1	0.7802	1.9	1	0.1663
10	10	0.1	2	0.9508	0.0	1	0.9450	0.1	1	0.7567
11	11	0.8	2	0.6802	0.7	1	0.3874	0.0	1	0.8802
12	12	1.9	2	0.3864	0.0	1	0.9658	1.9	1	0.1685
13	13	1.1	2	0.5869	1.0	1	0.3146	0.1	1	0.8177
14	14	0.8	2	0.6818	0.4	1	0.5289	0.4	1	0.5435
15	15	5.5	2	0.0639	3.8	1	0.0517	1.7	1	0.1913
16	16	3.2	2	0.2027	0.2	1	0.6404	3.0	1	0.0851
17	17	3.1	2	0.2166	1.4	1	0.2453	1.7	1	0.1912
18	18	0.7	2	0.6883	0.7	1	0.3890	0.0	1	0.9485
19	19	3.9	2	0.1411	2.8	1	0.0936	1.1	1	0.2940
20	20	2.9	2	0.2353	2.6	1	0.1048	0.3	1	0.6123
21	21	0.5	2	0.7866	0.3	1	0.5778	0.2	1	0.6803
22	22	1.8	2	0.4109	0.0	1	0.9977	1.8	1	0.1827

23	23	0.5	2	0.7925	0.4	1	0.5338	0.1	1	0.7807
24	24	5.4	2	0.0688	0.0	1	0.8382	5.3	1	0.0212
25	25	0.2	2	0.9178	0.2	1	0.6824	0.0	1	0.9504
26	26	3.5	2	0.1747	3.0	1	0.0826	0.5	1	0.4836
27	27	3.6	2	0.1704	2.4	1	0.1191	1.1	1	0.2927
28	28	1.6	2	0.4534	1.3	1	0.2523	0.3	1	0.6036
29	29	6.8	2	0.0332	0.1	1	0.7780	6.7	1	0.0095
30	30	9.7	2	0.0078	1.3	1	0.2466	8.4	1	0.0038

Marginal fit (X^2) and Standardized LD X^2 Statistics for Group 1 (Back to TOC)

Item	Label	Marginal X^2	1	2	3	4	5	6	7	8	9	10
1	Litera1	0.0										
2	Litera2	0.0	-0.7									
3	Litera3	0.0	3.0	0.1								
4	Litera4	0.0	-0.7	-0.7	1.2							
5	Litera5	0.0	1.3	-0.5	-0.7	-0.6						
6	Litera6	0.0	-0.2	2.3	0.5	-0.6	2.1					
7	Litera7	0.0	0.4	-0.5	-0.3	-0.5	-0.1	-0.5				
8	Litera8	0.0	-0.7	-0.1	0.6	0.0	-0.4	0.7	-0.4			
9	Litera9	0.0	0.5	3.0	-0.7	-0.5	-0.6	2.6	-0.7	1.7		
10	Litera10	0.0	-0.2	-0.1	-0.6	-0.6	-0.7	-0.7	-0.5	-0.7	-0.7	
11	Litera11	0.0	-0.1	-0.5	-0.4	-0.6	-0.3	-0.7	0.8	0.8	-0.5	-0.5
12	Litera12	0.0	-0.2	-0.4	-0.3	-0.7	-0.7	-0.3	-0.4	-0.7	-0.6	-0.6
13	Litera13	0.0	-0.4	1.2	2.6	0.3	0.4	-0.4	-0.7	-0.4	-0.2	2.3
14	Litera14	0.0	-0.7	-0.1	1.6	-0.6	0.1	0.6	-0.2	0.6	-0.6	0.6
15	Litera15	0.0	-0.7	0.1	-0.6	-0.5	-0.6	-0.7	-0.7	0.3	-0.7	-0.7
16	Litera16	0.0	2.4	-0.5	-0.7	-0.6	1.2	-0.1	-0.1	1.9	-0.6	-0.6
17	Litera17	0.0	-0.5	-0.7	2.8	-0.7	-0.7	-0.5	0.5	-0.5	-0.6	-0.1
18	Litera18	0.0	1.4	-0.7	7.0	-0.5	-0.6	-0.6	-0.7	-0.7	0.0	-0.7
19	Litera19	0.0	1.8	2.2	0.6	0.7	-0.7	-0.4	1.7	-0.7	-0.7	0.5
20	Litera20	0.0	-0.6	1.9	-0.5	-0.4	-0.3	-0.7	-0.2	0.8	-0.6	-0.0
21	Litera21	0.0	1.7	1.5	-0.4	0.2	-0.5	-0.7	-0.7	0.0	-0.5	0.4
22	Litera22	0.0	0.3	1.4	-0.6	-0.1	-0.7	-0.6	0.2	-0.7	1.2	-0.6
23	Litera23	0.0	0.4	1.2	1.2	-0.6	0.2	1.0	-0.6	-0.7	-0.7	-0.5
24	Litera24	0.0	0.7	0.2	3.7	0.1	-0.7	-0.5	-0.7	-0.7	1.0	0.0
25	Litera2	0.0	4.0	1.0	1.2	-0.5	0.5	1.2	-0.7	-0.1	-0.7	-0.6

	5										
26	Litera26	0.0	0.4	0.4	1.8	-0.7	-0.1	-0.7	-0.2	-0.5	-0.5	-0.6
27	Litera27	0.0	-0.7	1.6	0.1	0.9	1.1	-0.7	-0.2	-0.7	0.6	0.2
28	Litera28	0.0	3.2	4.8	-0.0	0.3	2.2	1.3	-0.7	-0.4	-0.7	0.2
29	Litera29	0.0	-0.6	0.5	-0.7	-0.7	0.6	-0.1	-0.2	1.4	-0.4	-0.7
30	Litera30	0.0	1.7	-0.7	-0.3	-0.3	-0.7	-0.5	-0.1	0.5	-0.5	1.1

Item	Label	Marginal X^2	11	12	13	14	15	16	17	18	19	20
11	Litera11	0.0										
12	Litera12	0.0	0.1									
13	Litera13	0.0	0.1	-0.3								
14	Litera14	0.0	-0.6	2.9	5.1							
15	Litera15	0.0	-0.7	0.8	0.5	2.4						
16	Litera16	0.0	-0.3	0.5	-0.3	-0.6	-0.5					
17	Litera17	0.0	0.7	-0.7	-0.2	-0.7	2.2	-0.7				
18	Litera18	0.0	0.3	-0.6	-0.2	-0.7	-0.2	-0.5	2.9			
19	Litera19	0.0	0.6	1.6	-0.7	0.9	-0.5	-0.7	-0.6	-0.7		
20	Litera20	0.0	-0.6	-0.3	-0.4	0.1	-0.7	-0.2	-0.5	-0.7	9.5	
21	Litera21	0.0	1.3	0.8	-0.1	-0.6	0.4	0.7	-0.5	-0.3	-0.6	-0.6
22	Litera22	0.0	-0.7	-0.5	-0.3	0.5	1.7	1.4	-0.0	-0.6	-0.1	-0.0
23	Litera23	0.0	-0.7	-0.5	-0.7	-0.7	3.5	-0.4	0.3	-0.7	-0.7	0.1
24	Litera24	0.0	-0.7	1.3	-0.4	-0.7	-0.7	-0.6	0.7	-0.6	-0.7	0.8
25	Litera25	0.0	-0.7	1.2	-0.5	-0.5	2.6	0.2	-0.4	-0.4	-0.4	0.3
26	Litera26	0.0	-0.2	0.8	-0.5	2.7	-0.0	-0.6	-0.6	-0.6	-0.5	-0.3
27	Litera27	0.0	-0.6	-0.6	-0.5	1.7	-0.7	-0.7	-0.7	0.1	-0.5	0.2
28	Litera28	0.0	0.3	-0.7	3.6	-0.7	-0.6	-0.7	-0.5	-0.5	-0.6	0.1
29	Litera29	0.0	-0.6	0.2	-0.6	-0.4	-0.7	-0.7	3.0	0.1	-0.5	-0.2
30	Litera30	0.0	-0.7	-0.7	2.7	-0.6	0.0	0.5	-0.7	-0.3	-0.4	-0.7

Item	Label	Marginal X^2	21	22	23	24	25	26	27	28	29
21	Litera21	0.0									
22	Litera22	0.0	1.5								
23	Litera23	0.0	0.1	-0.7							
24	Litera24	0.0	1.1	-0.7	-0.7						
25	Litera25	0.0	3.0	0.7	0.0	-0.2					
26	Litera26	0.0	-0.3	-0.6	-0.6	4.9	6.4				
27	Litera27	0.0	1.9	-0.6	1.6	2.0	-0.5	4.1			

28	Litera28	0.0	-0.7	-0.6	-0.3	-0.7	3.8	0.0	0.4		
29	Litera29	0.0	-0.7	-0.5	1.5	3.8	1.8	0.0	0.7	8.4	
30	Litera30	0.0	0.1	0.0	2.5	-0.6	-0.5	-0.3	-0.4	0.2	2.7

Marginal fit (X^2) and Standardized LD X^2 Statistics for Group 2 (Back to TOC)

Item	Label	Marginal X^2	1	2	3	4	5	6	7	8	9	10
1	Litera1	0.0										
2	Litera2	0.0	-0.6									
3	Litera3	0.0	3.1	-0.4								
4	Litera4	0.0	5.6	-0.6	0.2							
5	Litera5	0.0	-0.1	0.1	-0.2	-0.7						
6	Litera6	0.0	-0.7	0.7	-0.5	2.5	-0.6					
7	Litera7	0.0	0.0	0.7	-0.7	0.8	-0.4	1.4				
8	Litera8	0.0	-0.5	-0.3	3.7	-0.3	0.4	0.4	0.0			
9	Litera9	0.0	-0.1	-0.6	-0.6	-0.2	0.2	-0.6	-0.3	-0.6		
10	Litera10	0.0	0.1	2.1	-0.7	0.3	-0.7	-0.1	-0.4	-0.5	-0.7	
11	Litera11	0.0	0.9	-0.5	0.9	-0.6	-0.6	1.4	-0.6	-0.5	-0.3	0.1
12	Litera12	0.0	-0.7	-0.5	0.8	-0.7	-0.4	0.2	-0.3	-0.1	0.4	-0.2
13	Litera13	0.0	0.8	6.1	3.3	-0.1	1.4	-0.7	-0.7	-0.7	-0.7	-0.7
14	Litera14	0.0	-0.7	-0.6	-0.4	1.0	-0.2	-0.4	0.9	-0.7	-0.7	0.6
15	Litera15	0.0	1.3	1.7	-0.3	4.7	-0.7	3.9	-0.2	-0.5	-0.7	2.2
16	Litera16	0.0	-0.6	-0.6	-0.7	-0.7	-0.5	-0.3	-0.4	-0.6	-0.0	-0.4
17	Litera17	0.0	-0.7	0.2	-0.4	0.7	-0.3	-0.1	-0.6	-0.7	1.1	-0.7
18	Litera18	0.0	-0.3	-0.3	1.3	-0.6	-0.4	1.3	0.4	2.3	-0.3	-0.4
19	Litera19	0.0	0.2	-0.6	-0.6	-0.2	-0.6	1.5	-0.4	-0.0	-0.6	-0.6
20	Litera20	0.0	0.7	3.8	-0.5	-0.1	0.3	-0.7	-0.6	-0.7	-0.6	-0.6
21	Litera21	0.0	0.9	-0.7	-0.3	0.2	-0.7	0.2	-0.6	-0.7	-0.7	-0.2
22	Litera22	0.0	0.9	0.1	0.2	-0.6	2.2	-0.4	-0.4	0.8	-0.4	-0.6
23	Litera23	0.0	0.2	2.4	2.1	0.8	-0.6	-0.5	-0.5	1.0	-0.6	0.3
24	Litera24	0.0	-0.6	-0.4	1.3	-0.6	-0.1	-0.7	1.3	-0.7	-0.1	3.5
25	Litera25	0.0	-0.2	-0.5	2.1	2.4	6.2	-0.7	0.2	-0.4	-0.7	-0.5
26	Litera26	0.0	-0.7	2.0	6.6	0.4	-0.2	-0.5	-0.7	-0.7	-0.1	-0.4
27	Litera2	0.0	-0.6	-0.6	0.1	0.3	-0.6	-0.5	-0.7	-0.3	-0.5	-0.5

	7											
28	Litera28	0.0	2.8	-0.6	5.4	2.9	-0.5	3.2	0.1	0.2	-0.5	0.0
29	Litera29	0.0	-0.1	2.4	1.2	0.1	1.4	0.1	-0.7	-0.5	0.1	-0.7
30	Litera30	0.0	1.6	0.4	-0.5	-0.3	0.8	-0.4	-0.3	1.4	-0.4	-0.7

Item	Label	Marginal X^2	11	12	13	14	15	16	17	18	19	20
11	Litera11	0.0										
12	Litera12	0.0	-0.6									
13	Litera13	0.0	-0.6	-0.2								
14	Litera14	0.0	0.2	-0.5	0.5							
15	Litera15	0.0	-0.2	-0.0	-0.7	-0.7						
16	Litera16	0.0	-0.6	-0.5	0.5	0.1	-0.6					
17	Litera17	0.0	-0.2	-0.6	-0.7	-0.5	-0.6	1.2				
18	Litera18	0.0	-0.2	0.2	1.2	0.3	-0.7	1.4	3.7			
19	Litera19	0.0	-0.5	1.9	-0.6	-0.2	-0.7	0.0	-0.6	-0.6		
20	Litera20	0.0	1.8	-0.7	-0.7	1.0	-0.7	-0.2	-0.7	-0.6	0.4	
21	Litera21	0.0	-0.7	-0.7	-0.7	-0.2	-0.7	-0.7	-0.3	-0.5	-0.4	-0.1
22	Litera22	0.0	-0.7	-0.2	1.7	0.8	0.7	-0.7	-0.5	-0.5	-0.3	-0.7
23	Litera23	0.0	-0.5	-0.4	-0.2	-0.4	-0.7	-0.4	2.0	0.1	-0.3	-0.7
24	Litera24	0.0	-0.1	-0.1	0.4	-0.2	0.1	-0.4	-0.7	-0.6	3.5	-0.7
25	Litera25	0.0	-0.7	1.3	-0.5	1.7	2.8	0.4	-0.7	1.3	-0.6	-0.6
26	Litera26	0.0	-0.1	-0.3	-0.4	-0.3	0.2	-0.1	-0.7	1.7	0.0	0.3
27	Litera27	0.0	-0.6	-0.0	-0.3	-0.7	-0.5	0.1	0.6	-0.7	-0.5	-0.1
28	Litera28	0.0	-0.4	-0.7	0.3	1.9	0.1	-0.7	-0.6	-0.1	1.2	-0.6
29	Litera29	0.0	-0.6	-0.7	0.2	-0.5	-0.7	-0.7	0.1	1.3	0.2	-0.2
30	Litera30	0.0	-0.7	-0.5	-0.4	-0.2	-0.7	2.3	-0.3	-0.3	-0.6	-0.7

		Margina

Item	Label	I X^2	21	22	23	24	25	26	27	28	29
21	Litera21	0.0									
22	Litera22	0.0	4.2								
23	Litera23	0.0	-0.5	2.6							
24	Litera24	0.0	-0.5	-0.7	-0.1						
25	Litera25	0.0	1.2	-0.6	-0.5	-0.4					
26	Litera26	0.0	0.0	-0.1	-0.6	3.7	-0.3				
27	Litera27	0.0	0.2	-0.4	-0.4	-0.7	-0.5	-0.4			
28	Litera28	0.0	1.6	0.9	-0.7	-0.7	-0.7	1.2	-0.4		
29	Litera29	0.0	0.7	-0.7	1.5	-0.7	-0.6	-0.4	-0.4	8.8	
30	Litera30	0.0	-0.4	0.5	0.2	-0.3	0.4	0.3	-0.5	-0.7	-0.4

Item Information Function Values for Group 1 at 15 Values of θ from -2.8 to 2.8 (Back to TOC)

Item	Label	θ: -2.8	-2.4	-2.0	-1.6	-1.2	-0.8	-0.4	-0.0	0.4	0.8	1.2	1.6	2.0	2.4	2.8
1	Litera1	0.06	0.07	0.08	0.09	0.09	0.10	0.10	0.09	0.09	0.08	0.07	0.06	0.05	0.05	0.04
2	Litera2	0.02	0.03	0.03	0.03	0.03	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.05	0.05	0.04
3	Litera3	0.07	0.08	0.09	0.10	0.11	0.11	0.11	0.10	0.10	0.09	0.08	0.06	0.05	0.04	0.04
4	Litera4	0.16	0.19	0.20	0.21	0.20	0.18	0.16	0.13	0.10	0.07	0.05	0.04	0.03	0.02	0.01
5	Litera5	0.10	0.12	0.14	0.16	0.17	0.17	0.16	0.15	0.13	0.11	0.08	0.07	0.05	0.04	0.03
6	Litera6	0.16	0.19	0.21	0.22	0.21	0.19	0.16	0.13	0.10	0.08	0.06	0.04	0.03	0.02	0.01
7	Litera7	0.14	0.20	0.27	0.34	0.38	0.37	0.33	0.26	0.19	0.13	0.09	0.05	0.03	0.02	0.01
8	Litera8	0.05	0.06	0.07	0.08	0.09	0.10	0.10	0.11	0.10	0.10	0.09	0.08	0.07	0.06	0.05
9	Litera9	0.14	0.16	0.16	0.16	0.15	0.13	0.11	0.09	0.07	0.06	0.04	0.03	0.03	0.02	0.01
10	Litera10	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
11	Litera11	0.09	0.10	0.10	0.10	0.10	0.10	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.03	0.02
12	Litera12	0.09	0.13	0.17	0.22	0.26	0.28	0.28	0.26	0.21	0.17	0.12	0.09	0.06	0.04	0.03
13	Litera13	0.02	0.03	0.03	0.03	0.03	0.03	0.03	0.04	0.04	0.04	0.04	0.04	0.03	0.03	0.03
14	Litera14	0.06	0.08	0.10	0.12	0.15	0.17	0.18	0.19	0.18	0.16	0.14	0.12	0.09	0.07	0.05
15	Litera15	0.03	0.04	0.04	0.05	0.05	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.05	0.05
16	Litera16	0.06	0.07	0.09	0.10	0.12	0.13	0.14	0.14	0.14	0.13	0.12	0.10	0.08	0.07	0.05
17	Litera17	0.13	0.18	0.23	0.27	0.29	0.29	0.26	0.22	0.17	0.13	0.09	0.06	0.04	0.03	0.02
18	Litera18	0.05	0.06	0.07	0.08	0.08	0.09	0.10	0.10	0.10	0.10	0.09	0.08	0.07	0.06	0.05

19	Litera19	0.20	0.25	0.29	0.31	0.30	0.26	0.21	0.15	0.11	0.08	0.05	0.03	0.02	0.01	0.01
20	Litera20	0.11	0.13	0.15	0.16	0.16	0.16	0.14	0.13	0.11	0.09	0.07	0.05	0.04	0.03	0.02
21	Litera21	0.05	0.06	0.07	0.08	0.08	0.09	0.09	0.09	0.09	0.08	0.08	0.07	0.06	0.05	0.04
22	Litera22	0.02	0.03	0.03	0.04	0.04	0.05	0.06	0.06	0.07	0.07	0.08	0.08	0.08	0.07	0.07
23	Litera23	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.02
24	Litera24	0.04	0.05	0.06	0.06	0.07	0.08	0.08	0.08	0.08	0.08	0.08	0.07	0.06	0.06	0.05
25	Litera25	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
26	Litera26	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.02	0.02
27	Litera27	0.03	0.03	0.04	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.04	0.04
28	Litera28	0.07	0.08	0.09	0.10	0.10	0.11	0.11	0.10	0.10	0.09	0.08	0.07	0.06	0.05	0.04
29	Litera29	0.04	0.05	0.06	0.06	0.07	0.08	0.08	0.09	0.09	0.09	0.08	0.08	0.07	0.06	0.06
30	Litera30	0.10	0.13	0.16	0.20	0.22	0.23	0.23	0.21	0.18	0.14	0.11	0.08	0.06	0.04	0.03
Test Information:		3.19	3.66	4.12	4.49	4.71	4.74	4.57	4.26	3.87	3.47	3.09	2.75	2.46	2.22	2.01
Expected s.e.:		0.56	0.52	0.49	0.47	0.46	0.46	0.47	0.48	0.51	0.54	0.57	0.60	0.64	0.67	0.71

Marginal Reliability for Response Pattern Scores: 0.75

Item Information Function Values for Group 2 at 15 Values of θ from -2.8 to 2.8 (Back to TOC)

Item	Label	θ: -2.8	-2.4	-2.0	-1.6	-1.2	-0.8	-0.4	-0.0	0.4	0.8	1.2	1.6	2.0	2.4	2.8
1	Litera1	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.04	0.04	0.04	0.03	0.03
2	Litera2	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
3	Litera3	0.08	0.09	0.10	0.10	0.11	0.11	0.10	0.10	0.09	0.08	0.07	0.06	0.05	0.04	0.03
4	Litera4	0.12	0.12	0.13	0.12	0.12	0.11	0.09	0.08	0.07	0.05	0.04	0.03	0.03	0.02	0.02
5	Litera5	0.05	0.06	0.06	0.07	0.07	0.08	0.08	0.08	0.07	0.07	0.06	0.06	0.05	0.05	0.04
6	Litera6	0.15	0.18	0.20	0.21	0.21	0.19	0.17	0.14	0.11	0.08	0.06	0.05	0.03	0.02	0.02
7	Litera7	0.15	0.22	0.32	0.42	0.47	0.46	0.39	0.29	0.20	0.13	0.08	0.05	0.03	0.02	0.01
8	Litera8	0.05	0.06	0.08	0.09	0.10	0.11	0.11	0.12	0.11	0.11	0.10	0.09	0.07	0.06	0.05
9	Litera9	0.16	0.18	0.18	0.17	0.16	0.14	0.11	0.09	0.07	0.05	0.04	0.03	0.02	0.02	0.01
10	Litera10	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03

11	Litera11	0.11	0.13	0.14	0.14	0.14	0.13	0.12	0.11	0.09	0.07	0.06	0.05	0.04	0.03	0.02
12	Litera12	0.08	0.12	0.16	0.21	0.25	0.28	0.29	0.27	0.23	0.19	0.14	0.10	0.07	0.05	0.03
13	Litera13	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
14	Litera14	0.05	0.07	0.09	0.10	0.12	0.14	0.15	0.15	0.15	0.14	0.13	0.11	0.09	0.07	0.06
15	Litera15	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
16	Litera16	0.06	0.07	0.09	0.10	0.11	0.12	0.12	0.12	0.11	0.11	0.09	0.08	0.07	0.06	0.05
17	Litera17	0.14	0.21	0.29	0.37	0.41	0.40	0.35	0.27	0.19	0.13	0.08	0.05	0.03	0.02	0.01
18	Litera18	0.05	0.06	0.08	0.09	0.11	0.12	0.13	0.14	0.14	0.13	0.12	0.10	0.09	0.07	0.06
19	Litera19	0.23	0.35	0.46	0.52	0.51	0.42	0.30	0.20	0.12	0.07	0.04	0.02	0.01	0.01	0.00
20	Litera20	0.12	0.17	0.21	0.25	0.27	0.27	0.25	0.21	0.17	0.13	0.09	0.06	0.04	0.03	0.02
21	Litera21	0.06	0.07	0.08	0.09	0.10	0.11	0.11	0.11	0.11	0.10	0.09	0.08	0.07	0.06	0.05
22	Litera22	0.02	0.02	0.03	0.03	0.04	0.05	0.05	0.06	0.07	0.07	0.07	0.08	0.08	0.08	0.07
23	Litera23	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.03	0.04	0.04	0.04	0.04	0.04	0.04	0.04
24	Litera24	0.05	0.05	0.06	0.07	0.07	0.07	0.08	0.08	0.07	0.07	0.06	0.06	0.05	0.05	0.04
25	Litera25	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
26	Litera26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
27	Litera27	0.04	0.05	0.06	0.07	0.08	0.09	0.09	0.10	0.10	0.10	0.09	0.08	0.08	0.07	0.06
28	Litera28	0.08	0.10	0.12	0.14	0.15	0.16	0.16	0.15	0.14	0.12	0.10	0.08	0.06	0.05	0.04
29	Litera29	0.05	0.05	0.06	0.07	0.07	0.08	0.08	0.08	0.08	0.08	0.07	0.07	0.06	0.05	0.05

30	Litera30	0.07	0.09	0.11	0.13	0.15	0.16	0.16	0.16	0.15	0.13	0.11	0.09	0.07	0.05	0.04
	Test Information:	3.11	3.66	4.23	4.72	4.98	4.95	4.68	4.27	3.81	3.38	2.99	2.65	2.36	2.12	1.92
	Expected s.e.:	0.57	0.52	0.49	0.46	0.45	0.45	0.46	0.48	0.51	0.54	0.58	0.61	0.65	0.69	0.72

Marginal Reliability for Response Pattern Scores: 0.74

Likelihood-based Values and Goodness of Fit Statistics (Back to TOC)

Statistics based on the loglikelihood	
-2loglikelihood:	74708.63
Akaike Information Criterion (AIC):	74948.63
Bayesian Information Criterion (BIC):	75620.73

Statistics based on the full item x item x ... classification
The table is too sparse to compute the general multinomial goodness of fit statistics.

Statistics based on one- and two-way marginal tables
The M_2 statistics were not requested.

Summary of the Data and Control Parameters (Back to TOC)

Group:	Group 1	Group 2
Sample Size	1000	1000
Number of Items	30	30
Number of Dimensions	1	1

Group 1

Item	Label	Categories	Model
1	Litera1	2	2PL
2	Litera2	2	2PL
3	Litera3	2	2PL
4	Litera4	2	2PL
5	Litera5	2	2PL
6	Litera6	2	2PL
7	Litera7	2	2PL
8	Litera8	2	2PL
9	Litera9	2	2PL
10	Litera10	2	2PL
11	Litera11	2	2PL
12	Litera12	2	2PL
13	Litera13	2	2PL
14	Litera14	2	2PL
15	Litera15	2	2PL
16	Litera16	2	2PL
17	Litera17	2	2PL
18	Litera18	2	2PL
19	Litera19	2	2PL

20	Litera20	2	2PL
21	Litera21	2	2PL
22	Litera22	2	2PL
23	Litera23	2	2PL
24	Litera24	2	2PL
25	Litera25	2	2PL
26	Litera26	2	2PL
27	Litera27	2	2PL
28	Litera28	2	2PL
29	Litera29	2	2PL
30	Litera30	2	2PL

Group 2

Item	Label	Categories	Model
1	Litera1	2	2PL
2	Litera2	2	2PL
3	Litera3	2	2PL
4	Litera4	2	2PL
5	Litera5	2	2PL
6	Litera6	2	2PL
7	Litera7	2	2PL
8	Litera8	2	2PL
9	Litera9	2	2PL
10	Litera10	2	2PL
11	Litera11	2	2PL
12	Litera12	2	2PL
13	Litera13	2	2PL
14	Litera14	2	2PL
15	Litera15	2	2PL
16	Litera16	2	2PL
17	Litera17	2	2PL
18	Litera18	2	2PL
19	Litera19	2	2PL
20	Litera20	2	2PL
21	Litera21	2	2PL
22	Litera22	2	2PL
23	Litera23	2	2PL
24	Litera24	2	2PL
25	Litera25	2	2PL
26	Litera26	2	2PL
27	Litera27	2	2PL
28	Litera28	2	2PL
29	Litera29	2	2PL
30	Litera30	2	2PL

Parameter Estimation Control Values

Bock-Aitkin EM Algorithm	
Maximum number of cycles:	500
Convergence criterion:	1.00e-005
Maximum number of M-step iterations:	50

Convergence criterion for iterative M-steps:	1.00e-006	
Number of rectangular quadrature points:	49	
Minimum, Maximum quadrature points:	-6.00	6.00
SEM algorithm tolerance:	1.00e-003	
Standard error computation algorithm:	Supplemented EM	

DIF Analysis
All items are evaluated for DIF
(Conditional on population distribution estimates obtained with all items constrained equal)

Contrasts among groups:

Group:	1	2
Contrast 1	1.000	-1.000

Miscellaneous Control Values

Print parameter numbers?	Yes
Z tolerance, max. abs. logit value:	50.00
Number of processor cores used:	8
Number of cycles completed:	34
Maximum parameter change:	0.00e+000
Number of free parameters:	120

Processing times (in seconds)

E-step computations:	0.07
M-step computations:	0.05
Standard error computations:	1.65
Goodness-of-fit statistics:	0.12
Total:	1.89

Output Files

HTML results and control parameters:	F:\Run IRT 06-08-2560\Literacy\Literacy 2000\Literacy 2000.Test1-irt.htm

Convergence and Numerical Stability

Engine status:	Normal termination
SEM algorithm status:	Normal
First-order test:	Convergence criteria satisfied
Condition number of information matrix:	1.28e+001
Second-order test:	Solution is a possible local maximum

ภาคผนวก ฉ
ตัวอย่าง Print Out การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี SIBTEST

number of items on test = 30
name of file for Ref. grp. scores = d:\RMCS หมูอ้วน วิเคราะห์โปรแกรม\Run SIBTEST\Run SIBTEST 20-08-2560\Literacy\Literacy 2000\Literacy 2000 Female.dat
name of file for Focal grp. scores = d:\RMCS หมูอ้วน วิเคราะห์โปรแกรม\Run SIBTEST\Run SIBTEST 20-08-2560\Literacy\Literacy 2000\Literacy 2000 Male.dat
minimum no. of examinees per matching score cell = 2
number of runs for this data set = 30
number of examinees in Reference group = 1000
number of examinees in Focal group = 1000

Examinee Test Score Summary Statistics
Reference Group: Mean = 16.78
Standard deviation = 4.66
Focal Group: Mean = 15.33
Standard deviation = 4.87
Standardized Score Difference = 0.30

Item Statistics
# = item number
p = proportion right on item (Classical Test Theory p-value)
r = point biserial (item score-test score correlation)

#:	1	2	3	4	5	6	7	8	9	10
p:	0.535	0.321	0.600	0.756	0.573	0.743	0.701	0.484	0.789	0.411
r:	0.308	0.237	0.335	0.348	0.345	0.381	0.478	0.356	0.337	0.241
#:	11	12	13	14	15	16	17	18	19	20
p:	0.701	0.580	0.423	0.468	0.430	0.492	0.690	0.451	0.790	0.673
r:	0.344	0.445	0.244	0.402	0.274	0.373	0.454	0.364	0.428	0.387
#:	21	22	23	24	25	26	27	28	29	30
p:	0.502	0.264	0.337	0.483	0.374	0.476	0.421	0.567	0.447	0.572
r:	0.352	0.283	0.218	0.332	0.195	0.212	0.316	0.380	0.341	0.414

$$
OUTPUT FOR RUN NUMBER 1 OUTPUT FOR RUN NUMBER 1

Suspect subtest items:

1

Matching subtest items:

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

estimate of guessing on this matching subtest = 0.20

proportion of Ref. grp. examinees eliminated = 0.019

proportion of Focal grp. examinees eliminated = 0.036

KR-20 for Ref. grp. = 0.727

KR-20 for Foc. grp. = 0.757

Matching Subtest Summary Statistics

Reference Group: Mean = 16.24

Standard deviation = 4.54

Focal Group: Mean = 14.80

Standard deviation = 4.73

Standardized Score Difference = 0.31

SIBTEST-Focal weighting Results

Beta estimate	standard error	p-value for DIF against either grp.
-0.034	0.023	0.139065

SIBTEST error flag = 0

No errors. This SIBTEST run had a normal successful completion.

$$$

OUTPUT FOR RUN NUMBER 2 OUTPUT FOR RUN NUMBER 2

Suspect subtest items:

2

Matching subtest items:

1 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

estimate of guessing on this matching subtest = 0.20

proportion of Ref. grp. examinees eliminated = 0.017

proportion of Focal grp. examinees eliminated = 0.032

KR-20 for Ref. grp. = 0.727

KR-20 for Foc. grp. = 0.756
Matching Subtest Summary Statistics
Reference Group: Mean = 16.45
Standard deviation = 4.59
Focal Group: Mean = 15.01
Standard deviation = 4.76
Standardized Score Difference = 0.31

SIBTEST-Focal weighting Results

Beta estimate	standard error	p-value for DIF against either grp.
-0.020	0.021	0.348945

SIBTEST error flag = 0
No errors. This SIBTEST run had a normal successful completion.

$$

OUTPUT FOR RUN NUMBER 3 OUTPUT FOR RUN NUMBER 3
Suspect subtest items:
3
Matching subtest items:
1 2 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
estimate of guessing on this matching subtest = 0.20

proportion of Ref. grp. examinees eliminated = 0.018
proportion of Focal grp. examinees eliminated = 0.036
KR-20 for Ref. grp. = 0.725
KR-20 for Foc. grp. = 0.758

Matching Subtest Summary Statistics
Reference Group: Mean = 16.13
Standard deviation = 4.53
Focal Group: Mean = 14.78

Standard deviation = 4.73

Standardized Score Difference = 0.29

SIBTEST-Focal weighting Results

Beta estimate	standard error	p-value for DIF against either grp.
0.037	0.022	0.094296

SIBTEST error flag = 0

No errors. This SIBTEST run had a normal successful completion.

$$

OUTPUT FOR RUN NUMBER 4 OUTPUT FOR RUN NUMBER 4

Suspect subtest items:

4

Matching subtest items:

1 2 3 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

estimate of guessing on this matching subtest = 0.20

proportion of Ref. grp. examinees eliminated = 0.020

proportion of Focal grp. examinees eliminated = 0.040

KR-20 for Ref. grp. = 0.725

KR-20 for Foc. grp. = 0.756

Matching Subtest Summary Statistics

Reference Group: Mean = 16.00

Standard deviation = 4.55

Focal Group: Mean = 14.60

Standard deviation = 4.73

Standardized Score Difference = 0.30

SIBTEST-Focal weighting Results

Beta estimate	standard error	p-value for DIF against either grp.
0.010	0.020	0.629005

SIBTEST error flag = 0

No errors. This SIBTEST run had a normal successful completion.

$$$

OUTPUT FOR RUN NUMBER 5 OUTPUT FOR RUN NUMBER 5

Suspect subtest items:

5

Matching subtest items:

1 2 3 4 6 7 8 9 10 11

12 13 14 15 16 17 18 19 20 21

22 23 24 25 26 27 28 29 30

estimate of guessing on this matching subtest = 0.20

proportion of Ref. grp. examinees eliminated = 0.017

proportion of Focal grp. examinees eliminated = 0.037

KR-20 for Ref. grp. = 0.724

KR-20 for Foc. grp. = 0.755

Matching Subtest Summary Statistics

Reference Group: Mean = 16.23

Standard deviation = 4.52

Focal Group: Mean = 14.73

Standard deviation = 4.71

Standardized Score Difference = 0.33

SIBTEST-Focal weighting Results

Beta estimate	standard error	p-value for DIF against either grp.
-0.120	0.022	0.000000

SIBTEST error flag = 0

No errors. This SIBTEST run had a normal successful completion.

$$

OUTPUT FOR RUN NUMBER 6 OUTPUT FOR RUN NUMBER 6

Suspect subtest items:

6

Matching subtest items:

1 2 3 4 5 7 8 9 10 11

12 13 14 15 16 17 18 19 20 21

22 23 24 25 26 27 28 29 30

estimate of guessing on this matching subtest = 0.20

proportion of Ref. grp. examinees eliminated = 0.020

proportion of Focal grp. examinees eliminated = 0.040

KR-20 for Ref. grp. = 0.721

KR-20 for Foc. grp. = 0.755

Matching Subtest Summary Statistics

Reference Group: Mean = 16.02

Standard deviation = 4.52

Focal Group: Mean = 14.60

Standard deviation = 4.72

Standardized Score Difference = 0.31

SIBTEST-Focal weighting Results

Beta estimate	standard error	p-value for DIF against either grp.
-0.022	0.020	0.278894

SIBTEST error flag = 0

No errors. This SIBTEST run had a normal successful completion.

$$

OUTPUT FOR RUN NUMBER 7 OUTPUT FOR RUN NUMBER 7

Suspect subtest items:
7

Matching subtest items:
1 2 3 4 5 6 8 9 10 11
12 13 14 15 16 17 18 19 20 21
22 23 24 25 26 27 28 29 30
estimate of guessing on this matching subtest = 0.20

proportion of Ref. grp. examinees eliminated = 0.018
proportion of Focal grp. examinees eliminated = 0.039
KR-20 for Ref. grp. = 0.716
KR-20 for Foc. grp. = 0.750

Matching Subtest Summary Statistics
Reference Group: Mean = 16.03
Standard deviation = 4.47
Focal Group: Mean = 14.68
Standard deviation = 4.67

Standardized Score Difference = 0.30

SIBTEST-Focal weighting Results

Beta estimate	standard error	p-value for DIF against either grp.
0.008	0.021	0.679891

SIBTEST error flag = 0
No errors. This SIBTEST run had a normal successful completion.

$$
OUTPUT FOR RUN NUMBER 8 OUTPUT FOR RUN NUMBER 8

Suspect subtest items:
8

Matching subtest items:
1 2 3 4 5 6 7 9 10 11
12 13 14 15 16 17 18 19 20 21
22 23 24 25 26 27 28 29 30
estimate of guessing on this matching subtest = 0.20

proportion of Ref. grp. examinees eliminated = 0.016
proportion of Focal grp. examinees eliminated = 0.035
KR-20 for Ref. grp. = 0.721
KR-20 for Foc. grp. = 0.756

Matching Subtest Summary Statistics
Reference Group: Mean = 16.27
Standard deviation = 4.50
Focal Group: Mean = 14.87
Standard deviation = 4.72

Standardized Score Difference = 0.31

SIBTEST-Focal weighting Results

Beta estimate	standard error	p-value for DIF against either grp.
-0.017	0.022	0.460945

SIBTEST error flag = 0
No errors. This SIBTEST run had a normal successful completion.

$$$
OUTPUT FOR RUN NUMBER 9 OUTPUT FOR RUN NUMBER 9

Suspect subtest items:
9

Matching subtest items:
1 2 3 4 5 6 7 8 10 11
12 13 14 15 16 17 18 19 20 21
22 23 24 25 26 27 28 29 30
estimate of guessing on this matching subtest = 0.20

proportion of Ref. grp. examinees eliminated = 0.020
proportion of Focal grp. examinees eliminated = 0.041
KR-20 for Ref. grp. = 0.724
KR-20 for Foc. grp. = 0.758

Matching Subtest Summary Statistics
Reference Group: Mean = 15.95
Standard deviation = 4.55
Focal Group: Mean = 14.58
Standard deviation = 4.75

Standardized Score Difference = 0.30

SIBTEST-Focal weighting Results

Beta estimate	standard error	p-value for DIF against either grp.
0.031	0.019	0.103676

SIBTEST error flag = 0
No errors. This SIBTEST run had a normal successful completion.
$$
OUTPUT FOR RUN NUMBER 10 OUTPUT FOR RUN NUMBER 10

Suspect subtest items:
10

Matching subtest items:
1 2 3 4 5 6 7 8 9 11
12 13 14 15 16 17 18 19 20 21

22 23 24 25 26 27 28 29 30
estimate of guessing on this matching subtest = 0.20

proportion of Ref. grp. examinees eliminated = 0.016
proportion of Focal grp. examinees eliminated = 0.040
KR-20 for Ref. grp. = 0.727
KR-20 for Foc. grp. = 0.761

Matching Subtest Summary Statistics
Reference Group: Mean = 16.35
Standard deviation = 4.56
Focal Group: Mean = 14.93
Standard deviation = 4.78

Standardized Score Difference = 0.30

SIBTEST-Focal weighting Results

Beta estimate	standard error	p-value for DIF against either grp.
-0.006	0.023	0.781364

SIBTEST error flag = 0
No errors. This SIBTEST run had a normal successful completion.

$$$
OUTPUT FOR RUN NUMBER 11 OUTPUT FOR RUN NUMBER 11

Suspect subtest items:
11

Matching subtest items:
1 2 3 4 5 6 7 8 9 10
12 13 14 15 16 17 18 19 20 21
22 23 24 25 26 27 28 29 30
estimate of guessing on this matching subtest = 0.20

proportion of Ref. grp. examinees eliminated = 0.021
proportion of Focal grp. examinees eliminated = 0.041
KR-20 for Ref. grp. = 0.723
KR-20 for Foc. grp. = 0.759

Matching Subtest Summary Statistics
Reference Group: Mean = 16.05
Standard deviation = 4.52
Focal Group: Mean = 14.66
Standard deviation = 4.74

Standardized Score Difference = 0.30

SIBTEST-Focal weighting Results

Beta estimate	standard error	p-value for DIF against either grp.
0.004	0.021	0.856078

SIBTEST error flag = 0
No errors. This SIBTEST run had a normal successful completion.

$$
OUTPUT FOR RUN NUMBER 12 OUTPUT FOR RUN NUMBER 12

Suspect subtest items:
12

Matching subtest items:
1 2 3 4 5 6 7 8 9 10
11 13 14 15 16 17 18 19 20 21
22 23 24 25 26 27 28 29 30
estimate of guessing on this matching subtest = 0.20

proportion of Ref. grp. examinees eliminated = 0.017
proportion of Focal grp. examinees eliminated = 0.032

KR-20 for Ref. grp. = 0.716
KR-20 for Foc. grp. = 0.751

Matching Subtest Summary Statistics
Reference Group: Mean = 16.17
Standard deviation = 4.46
Focal Group: Mean = 14.78
Standard deviation = 4.68

Standardized Score Difference = 0.31

SIBTEST-Focal weighting Results

Beta estimate	standard error	p-value for DIF against either grp.
-0.033	0.022	0.131220

SIBTEST error flag = 0
No errors. This SIBTEST run had a normal successful completion.

$$$

OUTPUT FOR RUN NUMBER 13 OUTPUT FOR RUN NUMBER 13

Suspect subtest items:
13

Matching subtest items:
1 2 3 4 5 6 7 8 9 10
11 12 14 15 16 17 18 19 20 21
22 23 24 25 26 27 28 29 30
estimate of guessing on this matching subtest = 0.20

proportion of Ref. grp. examinees eliminated = 0.022
proportion of Focal grp. examinees eliminated = 0.040
KR-20 for Ref. grp. = 0.730
KR-20 for Foc. grp. = 0.759

Matching Subtest Summary Statistics

Reference Group: Mean = 16.34

Standard deviation = 4.58

Focal Group: Mean = 14.93

Standard deviation = 4.77

Standardized Score Difference = 0.30

SIBTEST-Focal weighting Results

Beta estimate	standard error	p-value for DIF against either grp.
0.004	0.023	0.866235

SIBTEST error flag = 0

No errors. This SIBTEST run had a normal successful completion.

$$

OUTPUT FOR RUN NUMBER 14 OUTPUT FOR RUN NUMBER 14

Suspect subtest items:

14

Matching subtest items:

1 2 3 4 5 6 7 8 9 10

11 12 13 15 16 17 18 19 20 21

22 23 24 25 26 27 28 29 30

estimate of guessing on this matching subtest = 0.20

proportion of Ref. grp. examinees eliminated = 0.016

proportion of Focal grp. examinees eliminated = 0.032

KR-20 for Ref. grp. = 0.719

KR-20 for Foc. grp. = 0.751

Matching Subtest Summary Statistics

Reference Group: Mean = 16.28

Standard deviation = 4.49
Focal Group: Mean = 14.89
Standard deviation = 4.69

Standardized Score Difference = 0.30

SIBTEST-Focal weighting Results

Beta estimate	standard error	p-value for DIF against either grp.
-0.016	0.022	0.457467

SIBTEST error flag = 0
No errors. This SIBTEST run had a normal successful completion.

$$

OUTPUT FOR RUN NUMBER 15 OUTPUT FOR RUN NUMBER 15

Suspect subtest items:
15

Matching subtest items:
1 2 3 4 5 6 7 8 9 10
11 12 13 14 16 17 18 19 20 21
22 23 24 25 26 27 28 29 30
estimate of guessing on this matching subtest = 0.20

proportion of Ref. grp. examinees eliminated = 0.020
proportion of Focal grp. examinees eliminated = 0.032
KR-20 for Ref. grp. = 0.729
KR-20 for Foc. grp. = 0.757

Matching Subtest Summary Statistics
Reference Group: Mean = 16.32
Standard deviation = 4.57
Focal Group: Mean = 14.93

Standard deviation = 4.75

Standardized Score Difference = 0.30

SIBTEST-Focal weighting Results

Beta estimate	standard error	p-value for DIF against either grp.
0.034	0.023	0.137777

SIBTEST error flag = 0

No errors. This SIBTEST run had a normal successful completion.

$$$

OUTPUT FOR RUN NUMBER 16 OUTPUT FOR RUN NUMBER 16

Suspect subtest items:

16

Matching subtest items:

1 2 3 4 5 6 7 8 9 10

11 12 13 14 15 17 18 19 20 21

22 23 24 25 26 27 28 29 30

estimate of guessing on this matching subtest = 0.20

proportion of Ref. grp. examinees eliminated = 0.015

proportion of Focal grp. examinees eliminated = 0.036

KR-20 for Ref. grp. = 0.722

KR-20 for Foc. grp. = 0.754

Matching Subtest Summary Statistics

Reference Group: Mean = 16.24

Standard deviation = 4.51

Focal Group: Mean = 14.89

Standard deviation = 4.71

Standardized Score Difference = 0.29

SIBTEST-Focal weighting Results

		p-value
Beta estimate	standard error	for DIF against either grp.
0.040	0.022	0.071020

SIBTEST error flag = 0

No errors. This SIBTEST run had a normal successful completion.

$$

OUTPUT FOR RUN NUMBER 17   OUTPUT FOR RUN NUMBER 17

Suspect subtest items:

17

Matching subtest items:

1 2 3 4 5 6 7 8 9 10

11 12 13 14 15 16 18 19 20 21

22 23 24 25 26 27 28 29 30

estimate of guessing on this matching subtest = 0.20

proportion of Ref. grp. examinees eliminated = 0.018

proportion of Focal grp. examinees eliminated = 0.041

KR-20 for Ref. grp. =   0.716

KR-20 for Foc. grp. =   0.752

Matching Subtest Summary Statistics

Reference Group: Mean = 16.04

Standard deviation = 4.48

Focal Group: Mean = 14.69

Standard deviation = 4.68

Standardized Score Difference = 0.29

SIBTEST-Focal weighting Results

Beta estimate	standard error	p-value for DIF against either grp.
0.021	0.021	0.307216

SIBTEST error flag = 0

No errors. This SIBTEST run had a normal successful completion.

$$

OUTPUT FOR RUN NUMBER 18 OUTPUT FOR RUN NUMBER 18

Suspect subtest items:

18

Matching subtest items:

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

estimate of guessing on this matching subtest = 0.20

proportion of Ref. grp. examinees eliminated = 0.017
proportion of Focal grp. examinees eliminated = 0.033
KR-20 for Ref. grp. = 0.719
KR-20 for Foc. grp. = 0.755

Matching Subtest Summary Statistics

Reference Group: Mean = 16.30
Standard deviation = 4.49
Focal Group: Mean = 14.91
Standard deviation = 4.73

Standardized Score Difference = 0.30

SIBTEST-Focal weighting Results

p-value

Beta estimate	standard error	for DIF against either grp.
0.001	0.022	0.979689

SIBTEST error flag = 0

No errors. This SIBTEST run had a normal successful completion.

$$$

OUTPUT FOR RUN NUMBER 19 OUTPUT FOR RUN NUMBER 19

Suspect subtest items:

19

Matching subtest items:

1 2 3 4 5 6 7 8 9 10

11 12 13 14 15 16 17 18 20 21

22 23 24 25 26 27 28 29 30

estimate of guessing on this matching subtest = 0.20

proportion of Ref. grp. examinees eliminated = 0.018

proportion of Focal grp. examinees eliminated = 0.038

KR-20 for Ref. grp. = 0.718

KR-20 for Foc. grp. = 0.754

Matching Subtest Summary Statistics

Reference Group: Mean = 15.95

Standard deviation = 4.51

Focal Group: Mean = 14.58

Standard deviation = 4.71

Standardized Score Difference = 0.30

SIBTEST-Focal weighting Results

p-value

Beta estimate standard error for DIF against either grp.

0.014 0.019 0.471486

SIBTEST error flag = 0

No errors. This SIBTEST run had a normal successful completion.

$$

OUTPUT FOR RUN NUMBER 20 OUTPUT FOR RUN NUMBER 20

Suspect subtest items:

20

Matching subtest items:

1 2 3 4 5 6 7 8 9 10

11 12 13 14 15 16 17 18 19 21

22 23 24 25 26 27 28 29 30

estimate of guessing on this matching subtest = 0.20

proportion of Ref. grp. examinees eliminated = 0.021

proportion of Focal grp. examinees eliminated = 0.037

KR-20 for Ref. grp. = 0.719

KR-20 for Foc. grp. = 0.757

Matching Subtest Summary Statistics

Reference Group: Mean = 16.08

Standard deviation = 4.49

Focal Group: Mean = 14.69

Standard deviation = 4.72

Standardized Score Difference = 0.30

SIBTEST-Focal weighting Results

Beta estimate	standard error	p-value for DIF against either grp.
-0.023	0.022	0.290775

SIBTEST error flag = 0

No errors. This SIBTEST run had a normal successful completion.

$$$

OUTPUT FOR RUN NUMBER 21 OUTPUT FOR RUN NUMBER 21

Suspect subtest items:

21

Matching subtest items:

1 2 3 4 5 6 7 8 9 10

11 12 13 14 15 16 17 18 19 20

22 23 24 25 26 27 28 29 30

estimate of guessing on this matching subtest = 0.20

proportion of Ref. grp. examinees eliminated = 0.018

proportion of Focal grp. examinees eliminated = 0.036

KR-20 for Ref. grp. = 0.721

KR-20 for Foc. grp. = 0.756

Matching Subtest Summary Statistics

Reference Group: Mean = 16.25

Standard deviation = 4.51

Focal Group: Mean = 14.85

Standard deviation = 4.73

Standardized Score Difference = 0.30

SIBTEST-Focal weighting Results

Beta estimate	standard error	p-value for DIF against either grp.
-0.011	0.022	0.640146

SIBTEST error flag = 0

No errors. This SIBTEST run had a normal successful completion.

$$$

OUTPUT FOR RUN NUMBER 22 OUTPUT FOR RUN NUMBER 22

Suspect subtest items:
22

Matching subtest items:
1 2 3 4 5 6 7 8 9 10
11 12 13 14 15 16 17 18 19 20
21 23 24 25 26 27 28 29 30
estimate of guessing on this matching subtest = 0.20

proportion of Ref. grp. examinees eliminated = 0.018
proportion of Focal grp. examinees eliminated = 0.034
KR-20 for Ref. grp. = 0.721
KR-20 for Foc. grp. = 0.753

Matching Subtest Summary Statistics
Reference Group: Mean = 16.51
Standard deviation = 4.56
Focal Group: Mean = 15.07
Standard deviation = 4.76

Standardized Score Difference = 0.31

SIBTEST-Focal weighting Results

Beta estimate	standard error	p-value for DIF against either grp.
-0.026	0.019	0.173378

SIBTEST error flag = 0
No errors. This SIBTEST run had a normal successful completion.

$$$

OUTPUT FOR RUN NUMBER 23 OUTPUT FOR RUN NUMBER 23
Suspect subtest items:

23

Matching subtest items:
1 2 3 4 5 6 7 8 9 10
11 12 13 14 15 16 17 18 19 20
21 22 24 25 26 27 28 29 30
estimate of guessing on this matching subtest = 0.20

proportion of Ref. grp. examinees eliminated = 0.019
proportion of Focal grp. examinees eliminated = 0.037
KR-20 for Ref. grp. = 0.726
KR-20 for Foc. grp. = 0.761

Matching Subtest Summary Statistics
Reference Group: Mean = 16.43
Standard deviation = 4.57
Focal Group: Mean = 15.01
Standard deviation = 4.80

Standardized Score Difference = 0.30

SIBTEST-Focal weighting Results

Beta estimate	standard error	p-value for DIF against either grp.
-0.003	0.021	0.885020

SIBTEST error flag = 0
No errors. This SIBTEST run had a normal successful completion.

$$
OUTPUT FOR RUN NUMBER 24 OUTPUT FOR RUN NUMBER 24

Suspect subtest items:
24

Matching subtest items:
1 2 3 4 5 6 7 8 9 10
11 12 13 14 15 16 17 18 19 20
21 22 23 25 26 27 28 29 30
estimate of guessing on this matching subtest = 0.20

proportion of Ref. grp. examinees eliminated = 0.017
proportion of Focal grp. examinees eliminated = 0.035
KR-20 for Ref. grp. = 0.725
KR-20 for Foc. grp. = 0.756

Matching Subtest Summary Statistics
Reference Group: Mean = 16.25
Standard deviation = 4.53
Focal Group: Mean = 14.90
Standard deviation = 4.73

Standardized Score Difference = 0.29

SIBTEST-Focal weighting Results

Beta estimate	standard error	p-value for DIF against either grp.
0.050	0.023	0.028911

SIBTEST error flag = 0
No errors. This SIBTEST run had a normal successful completion.

$$
OUTPUT FOR RUN NUMBER 25 OUTPUT FOR RUN NUMBER 25

Suspect subtest items:
25

Matching subtest items:
1 2 3 4 5 6 7 8 9 10

11 12 13 14 15 16 17 18 19 20
21 22 23 24 26 27 28 29 30
estimate of guessing on this matching subtest = 0.20

proportion of Ref. grp. examinees eliminated = 0.019
proportion of Focal grp. examinees eliminated = 0.038
KR-20 for Ref. grp. = 0.729
KR-20 for Foc. grp. = 0.763

Matching Subtest Summary Statistics
Reference Group: Mean = 16.40
Standard deviation = 4.58
Focal Group: Mean = 14.97
Standard deviation = 4.81

Standardized Score Difference = 0.30

SIBTEST-Focal weighting Results

Beta estimate	standard error	p-value for DIF against either grp.
-0.012	0.022	0.580680

SIBTEST error flag = 0
No errors. This SIBTEST run had a normal successful completion.

$$$
OUTPUT FOR RUN NUMBER 26 OUTPUT FOR RUN NUMBER 26

Suspect subtest items:
26

Matching subtest items:
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 27 28 29 30

estimate of guessing on this matching subtest = 0.20

proportion of Ref. grp. examinees eliminated = 0.018
proportion of Focal grp. examinees eliminated = 0.043
KR-20 for Ref. grp. = 0.734
KR-20 for Foc. grp. = 0.762

Matching Subtest Summary Statistics
Reference Group: Mean = 16.28
Standard deviation = 4.60
Focal Group: Mean = 14.87
Standard deviation = 4.78

Standardized Score Difference = 0.30

SIBTEST-Focal weighting Results

Beta estimate	standard error	p-value for DIF against either grp.
0.022	0.023	0.334398

SIBTEST error flag = 0
No errors. This SIBTEST run had a normal successful completion.
$$$
OUTPUT FOR RUN NUMBER 27 OUTPUT FOR RUN NUMBER 27

Suspect subtest items:
27

Matching subtest items:
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 28 29 30
estimate of guessing on this matching subtest = 0.20

proportion of Ref. grp. examinees eliminated = 0.017

proportion of Focal grp. examinees eliminated = 0.038
KR-20 for Ref. grp. = 0.721
KR-20 for Foc. grp. = 0.758

Matching Subtest Summary Statistics
Reference Group: Mean = 16.32
Standard deviation = 4.51
Focal Group: Mean = 14.94
Standard deviation = 4.76

Standardized Score Difference = 0.30

SIBTEST-Focal weighting Results

Beta estimate	standard error	p-value for DIF against either grp.
0.019	0.022	0.378562

SIBTEST error flag = 0
No errors. This SIBTEST run had a normal successful completion.

$$$
OUTPUT FOR RUN NUMBER 28 OUTPUT FOR RUN NUMBER 28

Suspect subtest items:
28

Matching subtest items:
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 29 30
estimate of guessing on this matching subtest = 0.20

proportion of Ref. grp. examinees eliminated = 0.018
proportion of Focal grp. examinees eliminated = 0.036
KR-20 for Ref. grp. = 0.720

KR-20 for Foc. grp. = 0.755

Matching Subtest Summary Statistics

Reference Group: Mean = 16.18

Standard deviation = 4.49

Focal Group: Mean = 14.80

Standard deviation = 4.71

Standardized Score Difference = 0.30

SIBTEST-Focal weighting Results

Beta estimate	standard error	p-value for DIF against either grp.
0.015	0.023	0.512855

SIBTEST error flag = 0

No errors. This SIBTEST run had a normal successful completion.

$$

OUTPUT FOR RUN NUMBER 29 OUTPUT FOR RUN NUMBER 29

Suspect subtest items:

29

Matching subtest items:

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 30

estimate of guessing on this matching subtest = 0.20

proportion of Ref. grp. examinees eliminated = 0.015

proportion of Focal grp. examinees eliminated = 0.032

KR-20 for Ref. grp. = 0.723

KR-20 for Foc. grp. = 0.755

Matching Subtest Summary Statistics

Reference Group: Mean = 16.28

Standard deviation = 4.52

Focal Group: Mean = 14.94

Standard deviation = 4.73

Standardized Score Difference = 0.29

SIBTEST-Focal weighting Results

Beta estimate	standard error	p-value for DIF against either grp.
0.053	0.022	0.018511

SIBTEST error flag = 0

No errors. This SIBTEST run had a normal successful completion.

$$$

OUTPUT FOR RUN NUMBER 30 OUTPUT FOR RUN NUMBER 30

Suspect subtest items:

30

Matching subtest items:

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

estimate of guessing on this matching subtest = 0.20

proportion of Ref. grp. examinees eliminated = 0.018

proportion of Focal grp. examinees eliminated = 0.034

KR-20 for Ref. grp. = 0.720

KR-20 for Foc. grp. = 0.751

Matching Subtest Summary Statistics

Reference Group: Mean = 16.20

Standard deviation = 4.49
Focal Group: Mean = 14.76
Standard deviation = 4.67

Standardized Score Difference = 0.31

SIBTEST-Focal weighting Results

Beta estimate	standard error	p-value for DIF against either grp.
-0.073	0.022	0.000939

SIBTEST error flag = 0
No errors. This SIBTEST run had a normal successful completion.
$$$

SUMMARY OF THE RUNS

p-value notation:
R denotes p-value for test of DIF/DBF against Ref. group
F denotes p-value for test of DIF/DBF against Foc. group
E denotes p-value for test of DIF/DBF against either the Ref. or Foc. group.

NOTES:
MS/SSD = Matching Subtest Standardized Score Difference. Standardized difference in mean observed scores between Reference group and Focal group on the matching subtest.

p-elim = proportion of Reference (R) and Focal (F) groups eliminated (not used) in SIBTEST calculations.

Positive Beta estimate indicates DIF/DBF favoring Ref. grp.
Negative Beta estimate indicates DIF/DBF favoring Foc. grp.

FLAG = error flag indicator. FLAG=0 indicates a normal successful completion of a SIBTEST run. All other values of FLAG come with short error messages.

SIBTEST-Focal weighting

Run no.	Suspect Subtest Item Numbers	Beta estimate	standard error	p-value	p-elim R	p-elim F	MS SSD	FLAG
1	1	-0.034	0.023	0.139	E .02	.04	0.31	0
2	2	-0.020	0.021	0.349	E .02	.03	0.31	0
3	3	0.037	0.022	0.094	E .02	.04	0.29	0
4	4	0.010	0.020	0.629	E .02	.04	0.30	0
5	5	-0.120	0.022	0.000	E .02	.04	0.33	0
6	6	-0.022	0.020	0.279	E .02	.04	0.31	0
7	7	0.008	0.021	0.680	E .02	.04	0.30	0
8	8	-0.017	0.022	0.461	E .02	.04	0.31	0
9	9	0.031	0.019	0.104	E .02	.04	0.30	0
10	10	-0.006	0.023	0.781	E .02	.04	0.30	0
11	11	0.004	0.021	0.856	E .02	.04	0.30	0
12	12	-0.033	0.022	0.131	E .02	.03	0.31	0
13	13	0.004	0.023	0.866	E .02	.04	0.30	0
14	14	-0.016	0.022	0.457	E .02	.03	0.30	0
15	15	0.034	0.023	0.138	E .02	.03	0.30	0
16	16	0.040	0.022	0.071	E .02	.04	0.29	0
17	17	0.021	0.021	0.307	E .02	.04	0.29	0
18	18	0.001	0.022	0.980	E .02	.03	0.30	0
19	19	0.014	0.019	0.471	E .02	.04	0.30	0
20	20	-0.023	0.022	0.291	E .02	.04	0.30	0
21	21	-0.011	0.022	0.640	E .02	.04	0.30	0
22	22	-0.026	0.019	0.173	E .02	.03	0.31	0
23	23	-0.003	0.021	0.885	E .02	.04	0.30	0
24	24	0.050	0.023	0.029	E .02	.04	0.29	0
25	25	-0.012	0.022	0.581	E .02	.04	0.30	0
26	26	0.022	0.023	0.334	E .02	.04	0.30	0
27	27	0.019	0.022	0.379	E .02	.04	0.30	0

28	28	0.015	0.023	0.513	E	.02 .04	0.30	0
29	29	0.053	0.022	0.019	E	.02 .03	0.29	0
30	30	-0.073	0.022	0.001	E	.02 .03	0.31	0

ภาคผนวก ช
ตัวอย่าง Print Out การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ด้วยวิธี MH

```
GET
  FILE='C:\Users\KITTY\Desktop\MH Literacy 300 1.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.

SAVE OUTFILE='C:\Users\KITTY\Desktop\MH Literacy 2000 Literacy
Female.sav'
  /COMPRESSED.
CROSSTABS
  /TABLES=Gender BY Answer BY id
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ RISK CMH(1)
  /CELLS=COUNT
  /COUNT ROUND CELL.
```

Crosstabs

[DataSet1] C:\Users\KITTY\Desktop\MH Literacy 2000 Literacy Female.sav

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Gender * Answer * id	60000	100.0%	0	0.0%	60000	100.0%

Gender * Answer * id Crosstabulation

Count

id			Answer		Total
			0	1	
1	Gender	1	471	529	1000
		2	459	541	1000
	Total		930	1070	2000
2	Gender	1	680	320	1000
		2	678	322	1000
	Total		1358	642	2000
3	Gender	1	445	555	1000
		2	355	645	1000
	Total		800	1200	2000
4	Gender	1	270	730	1000
		2	219	781	1000
	Total		489	1511	2000
5	Gender	1	398	602	1000
		2	455	545	1000

	Total		853	1147	2000
6	Gender	1	272	728	1000
		2	241	759	1000
	Total		513	1487	2000
7	Gender	1	345	655	1000
		2	252	748	1000
	Total		597	1403	2000
8	Gender	1	534	466	1000
		2	497	503	1000
	Total		1031	969	2000
9	Gender	1	246	754	1000
		2	177	823	1000
	Total		423	1577	2000
10	Gender	1	600	400	1000
		2	578	422	1000
	Total		1178	822	2000
11	Gender	1	325	675	1000
		2	274	726	1000
	Total		599	1401	2000
12	Gender	1	444	556	1000
		2	397	603	1000
	Total		841	1159	2000
13	Gender	1	594	406	1000
		2	561	439	1000
	Total		1155	845	2000
14	Gender	1	557	443	1000
		2	506	494	1000
	Total		1063	937	2000
15	Gender	1	601	399	1000
		2	539	461	1000
	Total		1140	860	2000
16	Gender	1	556	444	1000
		2	460	540	1000
	Total		1016	984	2000
17	Gender	1	359	641	1000
		2	261	739	1000
	Total		620	1380	2000
18	Gender	1	576	424	1000
		2	521	479	1000
	Total		1097	903	2000
19	Gender	1	249	751	1000

		2	171	829	1000
	Total		420	1580	2000
20	Gender	1	353	647	1000
		2	301	699	1000
	Total		654	1346	2000
21	Gender	1	520	480	1000
		2	476	524	1000
	Total		996	1004	2000
22	Gender	1	742	258	1000
		2	731	269	1000
	Total		1473	527	2000
23	Gender	1	674	326	1000
		2	652	348	1000
	Total		1326	674	2000
24	Gender	1	566	434	1000
		2	468	532	1000
	Total		1034	966	2000
25	Gender	1	634	366	1000
		2	618	382	1000
	Total		1252	748	2000
26	Gender	1	542	458	1000
		2	506	494	1000
	Total		1048	952	2000
27	Gender	1	613	387	1000
		2	545	455	1000
	Total		1158	842	2000
28	Gender	1	468	532	1000
		2	399	601	1000
	Total		867	1133	2000
29	Gender	1	605	395	1000
		2	500	500	1000
	Total		1105	895	2000
30	Gender	1	429	571	1000
		2	426	574	1000
	Total		855	1145	2000
Total	Gender	1	14668	15332	30000
		2	13223	16777	30000
	Total		27891	32109	60000

Chi-Square Tests

id		Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
	Pearson Chi-Square	.289[c]	1	.591		
	Continuity Correction[b]	.243	1	.622		
	Likelihood Ratio	.289	1	.591		
1	Fisher's Exact Test				.622	.311
	Linear-by-Linear Association	.289	1	.591		
	N of Valid Cases	2000				
	Pearson Chi-Square	.009[d]	1	.924		
	Continuity Correction[b]	.002	1	.962		
	Likelihood Ratio	.009	1	.924		
2	Fisher's Exact Test				.962	.481
	Linear-by-Linear Association	.009	1	.924		
	N of Valid Cases	2000				
	Pearson Chi-Square	16.875[e]	1	.000		
	Continuity Correction[b]	16.502	1	.000		
	Likelihood Ratio	16.903	1	.000		
3	Fisher's Exact Test				.000	.000
	Linear-by-Linear Association	16.867	1	.000		
	N of Valid Cases	2000				
	Pearson Chi-Square	7.040[f]	1	.008		
	Continuity Correction[b]	6.767	1	.009		
	Likelihood Ratio	7.050	1	.008		
4	Fisher's Exact Test				.009	.005
	Linear-by-Linear Association	7.037	1	.008		
	N of Valid Cases	2000				
	Pearson Chi-Square	6.642[g]	1	.010		
	Continuity Correction[b]	6.411	1	.011		
	Likelihood Ratio	6.646	1	.010		
5	Fisher's Exact Test				.011	.006
	Linear-by-Linear Association	6.638	1	.010		
	N of Valid Cases	2000				
	Pearson Chi-Square	2.520[h]	1	.112		
6	Continuity Correction[b]	2.360	1	.125		
	Likelihood Ratio	2.521	1	.112		

	Fisher's Exact Test				.124	.062
	Linear-by-Linear Association	2.518	1	.113		
	N of Valid Cases	2000				
	Pearson Chi-Square	20.652[i]	1	.000		
	Continuity Correction[b]	20.210	1	.000		
	Likelihood Ratio	20.716	1	.000		
7	Fisher's Exact Test				.000	.000
	Linear-by-Linear Association	20.642	1	.000		
	N of Valid Cases	2000				
	Pearson Chi-Square	2.741[j]	1	.098		
	Continuity Correction[b]	2.594	1	.107		
	Likelihood Ratio	2.741	1	.098		
8	Fisher's Exact Test				.107	.054
	Linear-by-Linear Association	2.739	1	.098		
	N of Valid Cases	2000				
	Pearson Chi-Square	14.274[k]	1	.000		
	Continuity Correction[b]	13.864	1	.000		
	Likelihood Ratio	14.326	1	.000		
9	Fisher's Exact Test				.000	.000
	Linear-by-Linear Association	14.267	1	.000		
	N of Valid Cases	2000				
	Pearson Chi-Square	1.000[l]	1	.317		
	Continuity Correction[b]	.911	1	.340		
	Likelihood Ratio	1.000	1	.317		
10	Fisher's Exact Test				.340	.170
	Linear-by-Linear Association	.999	1	.318		
	N of Valid Cases	2000				
	Pearson Chi-Square	6.199[m]	1	.013		
	Continuity Correction[b]	5.958	1	.015		
	Likelihood Ratio	6.204	1	.013		
11	Fisher's Exact Test				.015	.007
	Linear-by-Linear Association	6.196	1	.013		
	N of Valid Cases	2000				
12	Pearson Chi-Square	4.533[n]	1	.033		
	Continuity Correction[b]	4.342	1	.037		

	Likelihood Ratio	4.534	1	.033		
	Fisher's Exact Test				.037	.019
	Linear-by-Linear Association	4.530	1	.033		
	N of Valid Cases	2000				
	Pearson Chi-Square	2.232[o]	1	.135		
	Continuity Correction[b]	2.098	1	.147		
	Likelihood Ratio	2.232	1	.135		
13	Fisher's Exact Test				.147	.074
	Linear-by-Linear Association	2.230	1	.135		
	N of Valid Cases	2000				
	Pearson Chi-Square	5.223[p]	1	.022		
	Continuity Correction[b]	5.020	1	.025		
	Likelihood Ratio	5.225	1	.022		
14	Fisher's Exact Test				.025	.013
	Linear-by-Linear Association	5.220	1	.022		
	N of Valid Cases	2000				
	Pearson Chi-Square	7.842[q]	1	.005		
	Continuity Correction[b]	7.591	1	.006		
	Likelihood Ratio	7.847	1	.005		
15	Fisher's Exact Test				.006	.003
	Linear-by-Linear Association	7.838	1	.005		
	N of Valid Cases	2000				
	Pearson Chi-Square	18.437[r]	1	.000		
	Continuity Correction[b]	18.055	1	.000		
	Likelihood Ratio	18.465	1	.000		
16	Fisher's Exact Test				.000	.000
	Linear-by-Linear Association	18.428	1	.000		
	N of Valid Cases	2000				
	Pearson Chi-Square	22.450[s]	1	.000		
	Continuity Correction[b]	21.994	1	.000		
	Likelihood Ratio	22.521	1	.000		
17	Fisher's Exact Test				.000	.000
	Linear-by-Linear Association	22.439	1	.000		
	N of Valid Cases	2000				
18	Pearson Chi-Square	6.107[t]	1	.013		

	Continuity Correction[b]	5.887	1	.015		
	Likelihood Ratio	6.111	1	.013		
	Fisher's Exact Test				.015	.008
	Linear-by-Linear Association	6.104	1	.013		
	N of Valid Cases	2000				
	Pearson Chi-Square	18.336[u]	1	.000		
	Continuity Correction[b]	17.869	1	.000		
	Likelihood Ratio	18.422	1	.000		
19	Fisher's Exact Test				.000	.000
	Linear-by-Linear Association	18.327	1	.000		
	N of Valid Cases	2000				
	Pearson Chi-Square	6.143[v]	1	.013		
	Continuity Correction[b]	5.909	1	.015		
	Likelihood Ratio	6.148	1	.013		
20	Fisher's Exact Test				.015	.008
	Linear-by-Linear Association	6.140	1	.013		
	N of Valid Cases	2000				
	Pearson Chi-Square	3.872[w]	1	.049		
	Continuity Correction[b]	3.698	1	.054		
	Likelihood Ratio	3.873	1	.049		
21	Fisher's Exact Test				.054	.027
	Linear-by-Linear Association	3.870	1	.049		
	N of Valid Cases	2000				
	Pearson Chi-Square	.312[x]	1	.577		
	Continuity Correction[b]	.258	1	.612		
	Likelihood Ratio	.312	1	.577		
22	Fisher's Exact Test				.612	.306
	Linear-by-Linear Association	.312	1	.577		
	N of Valid Cases	2000				
	Pearson Chi-Square	1.083[y]	1	.298		
	Continuity Correction[b]	.987	1	.321		
	Likelihood Ratio	1.083	1	.298		
23	Fisher's Exact Test				.321	.160
	Linear-by-Linear Association	1.083	1	.298		
	N of Valid Cases	2000				

	Pearson Chi-Square	19.230[z]	1	.000		
	Continuity Correction[b]	18.840	1	.000		
	Likelihood Ratio	19.261	1	.000		
24	Fisher's Exact Test				.000	.000
	Linear-by-Linear Association	19.221	1	.000		
	N of Valid Cases	2000				
	Pearson Chi-Square	.547[aa]	1	.460		
	Continuity Correction[b]	.481	1	.488		
	Likelihood Ratio	.547	1	.460		
25	Fisher's Exact Test				.488	.244
	Linear-by-Linear Association	.546	1	.460		
	N of Valid Cases	2000				
	Pearson Chi-Square	2.598[ab]	1	.107		
	Continuity Correction[b]	2.456	1	.117		
	Likelihood Ratio	2.599	1	.107		
26	Fisher's Exact Test				.117	.059
	Linear-by-Linear Association	2.597	1	.107		
	N of Valid Cases	2000				
	Pearson Chi-Square	9.485[ac]	1	.002		
	Continuity Correction[b]	9.208	1	.002		
	Likelihood Ratio	9.493	1	.002		
27	Fisher's Exact Test				.002	.001
	Linear-by-Linear Association	9.480	1	.002		
	N of Valid Cases	2000				
	Pearson Chi-Square	9.693[ad]	1	.002		
	Continuity Correction[b]	9.415	1	.002		
	Likelihood Ratio	9.702	1	.002		
28	Fisher's Exact Test				.002	.001
	Linear-by-Linear Association	9.689	1	.002		
	N of Valid Cases	2000				
	Pearson Chi-Square	22.296[ae]	1	.000		
	Continuity Correction[b]	21.873	1	.000		
	Likelihood Ratio	22.339	1	.000		
29	Fisher's Exact Test				.000	.000
	Linear-by-Linear Association	22.285	1	.000		

	N of Valid Cases	2000				
	Pearson Chi-Square	.018[af]	1	.892		
	Continuity Correction[b]	.008	1	.928		
	Likelihood Ratio	.018	1	.892		
30	Fisher's Exact Test				.928	.464
	Linear-by-Linear Association	.018	1	.892		
	N of Valid Cases	2000				
	Pearson Chi-Square	139.893[a]	1	.000		
	Continuity Correction[b]	139.699	1	.000		
	Likelihood Ratio	139.949	1	.000		
Total	Fisher's Exact Test				.000	.000
	Linear-by-Linear Association	139.891	1	.000		
	N of Valid Cases	60000				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 13945.50.

b. Computed only for a 2x2 table

c. 0 cells (.0%) have expected count less than 5. The minimum expected count is 465.00.

d. 0 cells (.0%) have expected count less than 5. The minimum expected count is 321.00.

e. 0 cells (.0%) have expected count less than 5. The minimum expected count is 400.00.

f. 0 cells (.0%) have expected count less than 5. The minimum expected count is 244.50.

g. 0 cells (.0%) have expected count less than 5. The minimum expected count is 426.50.

h. 0 cells (.0%) have expected count less than 5. The minimum expected count is 256.50.

i. 0 cells (.0%) have expected count less than 5. The minimum expected count is 298.50.

j. 0 cells (.0%) have expected count less than 5. The minimum expected count is 484.50.

k. 0 cells (.0%) have expected count less than 5. The minimum expected count is 211.50.

l. 0 cells (.0%) have expected count less than 5. The minimum expected count is 411.00.

m. 0 cells (.0%) have expected count less than 5. The minimum expected count is 299.50.

n. 0 cells (.0%) have expected count less than 5. The minimum expected count is 420.50.

o. 0 cells (.0%) have expected count less than 5. The minimum expected count is 422.50.

p. 0 cells (.0%) have expected count less than 5. The minimum expected count is 468.50.

q. 0 cells (.0%) have expected count less than 5. The minimum expected count is 430.00.

r. 0 cells (.0%) have expected count less than 5. The minimum expected count is 492.00.

s. 0 cells (.0%) have expected count less than 5. The minimum expected count is 310.00.

t. 0 cells (.0%) have expected count less than 5. The minimum expected count is 451.50.

u. 0 cells (.0%) have expected count less than 5. The minimum expected count is 210.00.

v. 0 cells (.0%) have expected count less than 5. The minimum expected count is 327.00.

w. 0 cells (.0%) have expected count less than 5. The minimum expected count is 498.00.

x. 0 cells (.0%) have expected count less than 5. The minimum expected count is 263.50.

y. 0 cells (.0%) have expected count less than 5. The minimum expected count is 337.00.

z. 0 cells (.0%) have expected count less than 5. The minimum expected count is 483.00.

aa. 0 cells (.0%) have expected count less than 5. The minimum expected count is 374.00.

ab. 0 cells (.0%) have expected count less than 5. The minimum expected count is 476.00.

ac. 0 cells (.0%) have expected count less than 5. The minimum expected count is 421.00.

ad. 0 cells (.0%) have expected count less than 5. The minimum expected count is 433.50.

ae. 0 cells (.0%) have expected count less than 5. The minimum expected count is 447.50.

af. 0 cells (.0%) have expected count less than 5. The minimum expected count is 427.50.

Risk Estimate

id		Value	95% Confidence Interval	
			Lower	Upper
	Odds Ratio for Gender (1 / 2)	1.049	.880	1.251
1	For cohort Answer = 0	1.026	.934	1.127
	For cohort Answer = 1	.978	.901	1.061
	N of Valid Cases	2000		
	Odds Ratio for Gender (1 / 2)	1.009	.836	1.218
2	For cohort Answer = 0	1.003	.944	1.065
	For cohort Answer = 1	.994	.875	1.129
	N of Valid Cases	2000		
	Odds Ratio for Gender (1 / 2)	1.457	1.217	1.744
3	For cohort Answer = 0	1.254	1.125	1.397
	For cohort Answer = 1	.860	.801	.925
	N of Valid Cases	2000		
	Odds Ratio for Gender (1 / 2)	1.319	1.075	1.619
4	For cohort Answer = 0	1.233	1.056	1.440
	For cohort Answer = 1	.935	.889	.983
	N of Valid Cases	2000		
	Odds Ratio for Gender (1 / 2)	.792	.663	.946
5	For cohort Answer = 0	.875	.790	.969
	For cohort Answer = 1	1.105	1.024	1.192
	N of Valid Cases	2000		
	Odds Ratio for Gender (1 / 2)	1.177	.962	1.439
6	For cohort Answer = 0	1.129	.972	1.311
	For cohort Answer = 1	.959	.911	1.010

	N of Valid Cases	2000		
	Odds Ratio for Gender (1 / 2)	1.563	1.289	1.897
7	For cohort Answer = 0	1.369	1.194	1.570
	For cohort Answer = 1	.876	.827	.928
	N of Valid Cases	2000		
	Odds Ratio for Gender (1 / 2)	1.160	.973	1.382
8	For cohort Answer = 0	1.074	.987	1.170
	For cohort Answer = 1	.926	.846	1.014
	N of Valid Cases	2000		
	Odds Ratio for Gender (1 / 2)	1.517	1.221	1.885
9	For cohort Answer = 0	1.390	1.170	1.651
	For cohort Answer = 1	.916	.875	.959
	N of Valid Cases	2000		
	Odds Ratio for Gender (1 / 2)	1.095	.916	1.309
10	For cohort Answer = 0	1.038	.965	1.117
	For cohort Answer = 1	.948	.853	1.053
	N of Valid Cases	2000		
	Odds Ratio for Gender (1 / 2)	1.276	1.053	1.546
11	For cohort Answer = 0	1.186	1.037	1.357
	For cohort Answer = 1	.930	.878	.985
	N of Valid Cases	2000		
	Odds Ratio for Gender (1 / 2)	1.213	1.015	1.449
12	For cohort Answer = 0	1.118	1.009	1.240
	For cohort Answer = 1	.922	.856	.994
	N of Valid Cases	2000		
	Odds Ratio for Gender (1 / 2)	1.145	.959	1.367
13	For cohort Answer = 0	1.059	.982	1.141
	For cohort Answer = 1	.925	.835	1.025
	N of Valid Cases	2000		
	Odds Ratio for Gender (1 / 2)	1.228	1.030	1.464
14	For cohort Answer = 0	1.101	1.014	1.195
	For cohort Answer = 1	.897	.817	.985
	N of Valid Cases	2000		

	Odds Ratio for Gender (1 / 2)	1.288	1.079	1.538
15	For cohort Answer = 0	1.115	1.033	1.204
	For cohort Answer = 1	.866	.782	.958
	N of Valid Cases	2000		
	Odds Ratio for Gender (1 / 2)	1.470	1.233	1.753
16	For cohort Answer = 0	1.209	1.108	1.319
	For cohort Answer = 1	.822	.752	.900
	N of Valid Cases	2000		
	Odds Ratio for Gender (1 / 2)	1.586	1.310	1.920
17	For cohort Answer = 0	1.375	1.204	1.571
	For cohort Answer = 1	.867	.818	.920
	N of Valid Cases	2000		
	Odds Ratio for Gender (1 / 2)	1.249	1.047	1.490
18	For cohort Answer = 0	1.106	1.021	1.197
	For cohort Answer = 1	.885	.803	.975
	N of Valid Cases	2000		
	Odds Ratio for Gender (1 / 2)	1.607	1.292	1.999
19	For cohort Answer = 0	1.456	1.224	1.733
	For cohort Answer = 1	.906	.866	.948
	N of Valid Cases	2000		
	Odds Ratio for Gender (1 / 2)	1.267	1.051	1.528
20	For cohort Answer = 0	1.173	1.034	1.331
	For cohort Answer = 1	.926	.871	.984
	N of Valid Cases	2000		
	Odds Ratio for Gender (1 / 2)	1.193	1.001	1.421
21	For cohort Answer = 0	1.092	1.000	1.193
	For cohort Answer = 1	.916	.839	1.000
	N of Valid Cases	2000		
	Odds Ratio for Gender (1 / 2)	1.058	.867	1.291
22	For cohort Answer = 0	1.015	.963	1.070
	For cohort Answer = 1	.959	.828	1.111
	N of Valid Cases	2000		

	Odds Ratio for Gender (1 / 2)	1.104	.917	1.328
23	For cohort Answer = 0	1.034	.971	1.100
	For cohort Answer = 1	.937	.828	1.059
	N of Valid Cases	2000		
	Odds Ratio for Gender (1 / 2)	1.482	1.243	1.768
24	For cohort Answer = 0	1.209	1.110	1.317
	For cohort Answer = 1	.816	.744	.894
	N of Valid Cases	2000		
	Odds Ratio for Gender (1 / 2)	1.071	.893	1.283
25	For cohort Answer = 0	1.026	.959	1.098
	For cohort Answer = 1	.958	.855	1.073
	N of Valid Cases	2000		
	Odds Ratio for Gender (1 / 2)	1.155	.969	1.377
26	For cohort Answer = 0	1.071	.985	1.165
	For cohort Answer = 1	.927	.846	1.017
	N of Valid Cases	2000		
	Odds Ratio for Gender (1 / 2)	1.322	1.107	1.580
27	For cohort Answer = 0	1.125	1.043	1.212
	For cohort Answer = 1	.851	.767	.943
	N of Valid Cases	2000		
	Odds Ratio for Gender (1 / 2)	1.325	1.110	1.582
28	For cohort Answer = 0	1.173	1.061	1.297
	For cohort Answer = 1	.885	.820	.956
	N of Valid Cases	2000		
	Odds Ratio for Gender (1 / 2)	1.532	1.283	1.829
29	For cohort Answer = 0	1.210	1.117	1.310
	For cohort Answer = 1	.790	.716	.872
	N of Valid Cases	2000		
	Odds Ratio for Gender (1 / 2)	1.012	.848	1.209
30	For cohort Answer = 0	1.007	.910	1.115
	For cohort Answer = 1	.995	.922	1.073
	N of Valid Cases	2000		

	Odds Ratio for Gender (1 / 2)	1.214	1.175	1.253
Total	For cohort Answer = 0	1.109	1.090	1.129
	For cohort Answer = 1	.914	.900	.928
	N of Valid Cases	60000		

Tests of Homogeneity of the Odds Ratio

	Chi-Squared	df	Asymp. Sig. (2-sided)
Breslow-Day	86.936	29	.000
Tarone's	86.936	29	.000

Tests of Conditional Independence

	Chi-Squared	df	Asymp. Sig. (2-sided)
Cochran's	152.058	1	.000
Mantel-Haenszel	151.772	1	.000

Under the conditional independence assumption, Cochran's statistic is asymptotically distributed as a 1 df chi-squared distribution, only if the number of strata is fixed, while the Mantel-Haenszel statistic is always asymptotically distributed as a 1 df chi-squared distribution. Note that the continuity correction is removed from the Mantel-Haenszel statistic when the sum of the differences between the observed and the expected is 0.

Mantel-Haenszel Common Odds Ratio Estimate

Estimate			1.234
ln(Estimate)			.210
Std. Error of ln(Estimate)			.017
Asymp. Sig. (2-sided)			.000
Asymp. 95% Confidence Interval	Common Odds Ratio	Lower Bound	1.194
		Upper Bound	1.276
	ln(Common Odds Ratio)	Lower Bound	.177
		Upper Bound	.244

The Mantel-Haenszel common odds ratio estimate is asymptotically normally distributed under the common odds ratio of 1.000 assumption. So is the natural log of the estimate.

ผลงานวิจัย

อรุณี แปลงกาย และปิยะทิพย์ ประดุจพรม. (2561). การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ NT ระดับชั้นประถมศึกษาปีที่ 3 ระหว่างวิธีการทดสอบอัตราส่วนไลค์ลิฮูด วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล. ใน การประชุมวิชาการระดับชาติครั้งที่ 3 ประจำปี 2561 “นวัตกรรมการจัดการ;vการขับเคลื่อนธุรกิจชุมชนสู่เศรษฐกิจดิจิทัล”, *วันที่ 1 มิถุนายน พ.ศ. 2561.* (หน้า 720-730). ปทุมธานี: มหาวิทยาลัยราชภัฏวไลยอลงกรณ์ ในพระบรมราชูปถัมภ์.