JRM JOURNAL OF RESEARCH METHODOLOGY

Journal of Research Methodology
Vol. 36 No. 2 (May–August 2023)
ISSN 0857-2933 ISSN online 2697-4835
Research Article https://doi.org/10.14456/jrm.2023.9

The Appropriate Sample Size for Test Length on Estimating Item Parameters in Item Response Theory

Pramote Sopa[1], Prakittiya Tuksino[2], Pattrawadee Makmee[3]

Received: February 22, 2022 – Revised: September 1, 2022 – Accepted: September 30, 2022

Abstract

An important drawback in item response theory (IRT) is that the sample size must be sufficiently large to accurately estimate item parameters. Therefore, to save budget, time, and manpower, it is necessary to determine the minimum sample size suitable for different test lengths. Thus, the purpose of this research was to compare the results of different test lengths and sample sizes on estimating item parameters in IRT. The data used in this study were secondary data from The National Institute of Educational Testing Service (Public Organization). These data consisted of results from individual exams in the O-NET test, grade nine, which included only multiple-choice exams with binary item scores. Three test lengths were considered: Mathematics with 20 items, Thai language with 40 items, and English language with 50 items. A sample size of 5,000 students was selected to represent the true item parameter values for each test. Six conditional sample sizes (200, 300, 400, 500, 700, and 1,000 students) were obtained through systematic random sampling. The data were analyzed using R and SPSS programs. To compare the test parameter estimates with the actual parameters, a 3-parameter model was employed, with the criterion that the correlation coefficient (r) should be ≥ .70 and the root mean square difference (*RMSD*) ≤ 0.33 across all parameters (a, b, c). The results showed that for a 20-item test length, a sample size greater than 1,000 is recommended. A minimum sample size of 700 is appropriate for a 40-item test length, and for a 50-item test length with a first component variance of less than 10%, a sample size greater than 1,000 is recommended.

Keywords: Sample Size, Test Length, Item Parameters, Item Response Theory

[1] *Corresponding Author,*
Educational Measurement and Evaluation Program, Faculty of Education, Khon Kaen University, Khon Kaen, 40000, Thailand. pramotesopa@kkumail.com

[2] Educational Measurement and Evaluation Program, Faculty of Education, Khon Kaen University, Khon Kaen, 40000, Thailand. praktu@kku.ac.th

[3] Cognitive Science and Innovation Research Unit, College of Research Methodology and Cognitive Science, Burapha University, Chonburi, 20131, Thailand. pattrawadee@gmail.com

JRM JOURNAL OF RESEARCH METHODOLOGY
วารสารวิธีวิทยาการวิจัย
ปีที่ ๓๖ ฉบับที่ ๒ (พฤษภาคม–สิงหาคม ๒๕๖๖)
ISSN 0857-2933 ISSN online 2697-4835
บทความวิจัย https://doi.org/10.14456/jrm.2023.9

ขนาดตัวอย่างที่เหมาะสมกับความยาวแบบสอบสำหรับการประมาณค่าพารามิเตอร์ของข้อสอบตามทฤษฎีการตอบสนองข้อสอบ

ปราโมทย์ โสภา[1], ประกฤติยา ทักษิโณ[2] และ ภัทราวดี มากมี[3]

รับต้นฉบับ : 22 กุมภาพันธ์ 2565 – รับแก้ไข : 1 กันยายน 2565 – ตอบรับตีพิมพ์ : 30 กันยายน 2565

บทคัดย่อ

ปัญหาสำคัญของทฤษฎีการตอบสนองข้อสอบ (item response theory: IRT) คือต้องมีขนาดตัวอย่างที่ใหญ่เพียงพอจึงจะประมาณค่าพารามิเตอร์ได้อย่างแม่นยำ ด้วยเหตุนี้เพื่อเป็นการประหยัดงบประมาณ เวลา และกำลังคน จึงจำเป็นต้องมีการศึกษาหาขนาดตัวอย่างต่ำสุดที่เหมาะสมกับความยาวแบบสอบขนาดต่าง ๆ ดังนั้นการวิจัยครั้งนี้จึงมีวัตถุประสงค์เพื่อเปรียบเทียบผลของความยาวแบบสอบและขนาดตัวอย่างที่มีต่อการประมาณค่าพารามิเตอร์ของข้อสอบใน IRT ข้อมูลที่ใช้ในการวิจัยเป็นข้อมูลทุติยภูมิจากสถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน) ซึ่งเป็นข้อมูลผลการตอบข้อสอบรายข้อจากการทดสอบ O-NET ชั้นมัธยมศึกษาปีที่ 3 เฉพาะข้อสอบแบบเลือกตอบที่ให้คะแนนแบบ 2 ค่า จำนวน 3 ความยาวแบบสอบ ได้แก่ วิชาคณิตศาสตร์ จำนวน 20 ข้อ วิชาภาษาไทย จำนวน 40 ข้อ และวิชาภาษาอังกฤษ จำนวน 50 ข้อ กำหนดขนาดตัวอย่างที่เป็นตัวแทนค่าพารามิเตอร์ที่แท้จริง แบบสอบละ 5,000 คน ส่วนขนาดตัวอย่างสำหรับศึกษาตามเงื่อนไข จำนวน 6 ขนาด ได้แก่ 200, 300, 400, 500, 700 และ 1,000 คน โดยใช้แผนการสุ่มตัวอย่างแบบมีระบบ วิเคราะห์ข้อมูลด้วยโปรแกรม R และ SPSS กำหนดเกณฑ์เปรียบเทียบผลการประมาณค่าพารามิเตอร์ของข้อสอบกับค่าพารามิเตอร์ที่แท้จริงด้วยโมเดลแบบ 3 พารามิเตอร์ โดยต้องมีค่าสัมประสิทธิ์สหสัมพันธ์ (r) ≥ .70 และค่ารากที่สองของความแตกต่างกำลังสองเฉลี่ย (*RMSD*) ≤ 0.33 ในทุกพารามิเตอร์ (a, b, c) ผลการวิจัย พบว่าแบบสอบที่มีความยาว 20 ข้อ ควรใช้ขนาดตัวอย่างมากกว่า 1,000 คน ส่วนแบบสอบที่มีความยาว 40 ข้อ ควรใช้ขนาดตัวอย่างขั้นต่ำ 700 คน และแบบสอบที่มีความยาว 50 ข้อ ที่มีความแปรปรวนขององค์ประกอบแรกน้อยกว่าร้อยละ 10 ควรใช้ขนาดตัวอย่างมากกว่า 1,000 คน

คำสำคัญ : ขนาดตัวอย่าง, ความยาวแบบสอบ, ค่าพารามิเตอร์ของข้อสอบ, ทฤษฎีการตอบสนองข้อสอบ

[1] *ผู้รับผิดชอบบทความหลัก,*
สาขาวิชาการวัดและประเมินผลการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยขอนแก่น ขอนแก่น 40000, pramotesopa@kkumail.com

[2] สาขาวิชาการวัดและประเมินผลการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยขอนแก่น ขอนแก่น 40000, praktu@kku.ac.th

[3] หน่วยวิจัยวิทยาการปัญญาและนวัตกรรม วิทยาลัยวิทยาการวิจัยและวิทยาการปัญญา มหาวิทยาลัยบูรพา ชลบุรี 20131, pattrawadee@gmail.com

ความเป็นมาและความสำคัญ

ปัจจุบันทฤษฎีการตอบสนองข้อสอบ (item response theory: IRT) ถูกนำมาใช้กันอย่างแพร่หลายในการพัฒนาแบบสอบ การปรับเทียบคะแนนระหว่างแบบสอบ (test equating) รวมถึงการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (differential item functioning: DIF) เป็นต้น (Hambleton et al., 1991; Şahin & Anıl, 2017) อย่างไรก็ตาม ปัญหาสำคัญในการนำ IRT มาใช้สำหรับพัฒนาแบบสอบคือข้อกำหนดเกี่ยวกับขนาดตัวอย่างที่ใหญ่เพียงพอ (ประมาณ 1,000 คน) จึงจะประมาณค่าพารามิเตอร์ของข้อสอบได้อย่างแม่นยำ (Hambleton, 1989 as cited in Şahin & Anıl, 2017) ซึ่ง Stone & Yumoto (2004) และ Swaminathan et al. (2003) กล่าวว่า ขนาดตัวอย่างที่เหมาะสมหรือใหญ่เพียงพอเป็นปัจจัยหลักสำหรับความคงที่ของค่าพารามิเตอร์ใน IRT โดยมีความสำคัญอย่างยิ่งกับโมเดล IRT แบบ 2 - 3 พารามิเตอร์ นอกจากนี้ Narayanan & Swaminathan (1994) กล่าวว่า กระบวนการทำงานของ IRT ต้องใช้ตัวอย่างขนาดใหญ่ ซึ่งเป็นเงื่อนไขที่มักจะพบได้ยากในทางปฏิบัติ นอกจากนี้ Swaminathan & Gifford (1979) พบว่าการเพิ่มจำนวนผู้สอบและข้อสอบมีผลกับความแม่นยำของการประมาณค่าพารามิเตอร์ของข้อสอบ โดยเริ่มแรกในปี 1968 Lord ได้เสนอแนะขนาดตัวอย่าง 1,000 คน คือขนาดต่ำสุดที่เหมาะสมสำหรับการประมาณค่าพารามิเตอร์แบบ 3 พารามิเตอร์ ที่มีความแม่นยำสูงสำหรับความยาวแบบสอบ 50 ข้อ (Chuah et al., 2006; Drasgow, 1989) ซึ่งข้อเสนอแนะนี้ได้รับการสนับสนุนต่อ ๆ มาจากนักวิจัยหลายคน ทำให้ขนาดตัวอย่าง 1,000 คน เป็นขนาดต่ำสุดที่ได้รับการยอมรับสำหรับการประมาณค่าพารามิเตอร์ข้อสอบในทฤษฎี IRT (Şahin & Anıl, 2017; Thissen, 1982 as cited in Drasgow, 1989)

อย่างไรก็ตาม เป็นเรื่องที่สำคัญมากในการใช้ขนาดตัวอย่างต่ำสุดที่เหมาะสม เพราะจะทำให้ประหยัดงบประมาณ เวลา และกำลังคนในการวิจัย (Chuah et al., 2006) ในช่วงเวลาต่อมาจึงมีการศึกษาเกี่ยวกับการใช้ขนาดตัวอย่างน้อยกว่า 1,000 คน ในการประมาณค่าพารามิเตอร์ข้อสอบ ซึ่งข้อค้นพบที่ได้มีความแตกต่างกันค่อนข้างมาก โดยในการศึกษาโมเดลแบบ 3 พารามิเตอร์ มีข้อค้นพบที่สำคัญ และสรุปผลของขนาดตัวอย่างต่ำสุดที่เหมาะสมกับจำนวนข้อสอบในการประมาณค่าพารามิเตอร์ที่แม่นยำ ดังนี้ แบบสอบ 20 ข้อ กับขนาดตัวอย่าง 1,000 คน (Patsula & Gessaroli, 1995; Swaminathan & Gifford, 1979; Yen, 1987) แต่ Şahin & Anıl (2017) กลับพบว่าขนาดตัวอย่างต่ำสุดที่เหมาะสมของแบบสอบ 20 ข้อ คือ 750 คน แบบสอบ 25 ข้อ กับขนาดตัวอย่าง 200 คน (Weiss & Minden, 2012) แบบสอบ 30 ข้อ กับขนาดตัวอย่าง 500 คน (Akour & Al-Omari, 2013) แต่ Şahin & Anıl (2017) กลับพบว่าขนาดตัวอย่างต่ำสุดที่เหมาะสมของแบบสอบ 30 ข้อ คือ 350 คน แบบสอบ 40 ข้อ กับขนาดตัวอย่าง 1,000 คน (Patsula & Gessaroli, 1995; Tang et al., 1993; Yen, 1987) แบบสอบ 50 ข้อ กับขนาดตัวอย่าง 1,000 คน (Lord, 1968 as cited in Drasgow, 1989) แต่ Chuah et al. (2006) กลับพบว่าแบบสอบ 50 ข้อ กับขนาดตัวอย่าง 300 คน แบบสอบ 60 ข้อ กับขนาดตัวอย่าง 1,000 คน (Hulin et al., 1982) แบบสอบ 75 ข้อ กับขนาดตัวอย่าง 1,000 คน (Yoes, 1995 as cited in Şahin & Anıl, 2017) และแบบสอบ 80 ข้อ กับขนาดตัวอย่าง 500 คน (Ree & Jensen, 1980 as cited in Şahin & Anıl, 2017)

จากข้อค้นพบที่แตกต่างกันในเรื่องของความยาวแบบสอบและขนาดตัวอย่างนี้ มีข้อสังเกตว่างานวิจัยส่วนใหญ่ที่ผ่านมาจะใช้ข้อมูลจากการจำลองข้อมูล ซึ่งข้อมูลที่ได้จากการจำลองจะมีข้อเสียเปรียบที่สำคัญคือจะมีคุณภาพหรือสถานการณ์การทดสอบที่แตกต่างจากข้อมูลจริง (Sireci, 1991; Swaminathan et al., 2003) นอกจากนี้ Lord (1975, as cited in Swaminathan & Gifford, 1979) กล่าวว่าผลการศึกษาที่ได้จากการจำลองข้อมูลจะไม่สามารถสรุปอ้างอิงไปยังสถานการณ์จริงได้ และจากการค้นคว้าพบว่า บางงานวิจัยถึงแม้จะใช้ข้อมูลจริง แต่ก็ใช้ข้อมูลจริงจากข้อมูล 1 ชุด แล้วทำการสุ่มหรือใช้วิธีทางสถิติเลือกแบบสอบให้มีขนาดสั้นลงเพื่อสร้างเงื่อนไขในการศึกษาทำให้ละเลยโครงสร้างที่แท้จริงของแบบสอบ ดังนั้นในการวิจัยครั้งนี้ ผู้วิจัยจึงแก้ไขปัญหาดังกล่าวด้วยการใช้ข้อมูลจริงจากแบบสอบที่มีมาตรฐานและมีจำนวนข้อสอบที่แตกต่างกันของสถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน) หรือ สทศ. ซึ่งเป็นข้อมูลทุติยภูมิที่ได้จากการทดสอบ O-NET ชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2562 จำนวน 3 วิชา ที่มีความยาวแบบสอบแตกต่างกัน คือ วิชาคณิตศาสตร์ จำนวน 20 ข้อ วิชาภาษาไทย จำนวน 40 ข้อ และวิชาภาษาอังกฤษ จำนวน 50 ข้อ ผลที่ได้จากการวิจัยครั้งนี้จะเป็นการขยายองค์ความรู้และเป็นประโยชน์ต่อวงการวัดและประเมินผลทางการศึกษาของประเทศไทยทั้งในปัจจุบันและอนาคต เพราะทำให้ทราบขนาดตัวอย่างต่ำสุดที่เหมาะสมกับความยาวแบบสอบขนาดต่าง ๆ จากการวิเคราะห์ด้วยข้อมูลจริงทำให้ข้อค้นพบที่ได้สอดคล้องกับสถานการณ์จริง สามารถนำผลการวิจัยที่ได้ไปใช้ประโยชน์ในงานวิจัยที่เกี่ยวข้องต่อไปได้

วัตถุประสงค์วิจัย

เพื่อเปรียบเทียบผลของความยาวแบบสอบและขนาดตัวอย่างที่มีต่อการประมาณค่าพารามิเตอร์ของข้อสอบในทฤษฎีการตอบสนองข้อสอบ

เอกสารและงานวิจัยที่เกี่ยวข้อง

แนวคิดเกี่ยวกับทฤษฎีการตอบสนองข้อสอบ (item response theory: IRT)

ทฤษฎีการตอบสนองข้อสอบ (IRT) นำเสนอแนวคิดที่ว่าความน่าจะเป็นของการตอบข้อสอบได้ถูกต้องขึ้นอยู่กับความสามารถจริงของผู้ตอบ และคุณลักษณะของข้อสอบซึ่งประกอบด้วย พารามิเตอร์ความยาก (*b*) อำนาจจำแนก (*a*) และโอกาสการเดาข้อสอบได้ถูก (*c*) จะแสดงระบบความสัมพันธ์ด้วยโมเดลการตอบสนองข้อสอบ ซึ่งอาจเป็นโมเดล 1 พารามิเตอร์ (*b*) โมเดล 2 พารามิเตอร์ (*a*, *b*) หรือโมเดล 3 พารามิเตอร์ (*a*, *b*, *c*) โดยทฤษฎี IRT ถือว่าค่าพารามิเตอร์ของข้อสอบและความสามารถจริงของผู้สอบมีความสัมพันธ์กัน ดังนั้นการประมาณค่าพารามิเตอร์ของข้อสอบ (ความยาก อำนาจจำแนก โอกาสการเดาข้อสอบได้ถูก) จึงต้องพิจารณาร่วมกับความสามารถจริงของผู้ตอบ ถ้ากลุ่มผู้ตอบมีขนาดใหญ่ที่เป็นตัวแทนของประชากร การประมาณค่าพารามิเตอร์ของข้อสอบจะต้องกระทำไปพร้อม ๆ กับการประมาณค่าความสามารถจริงของผู้สอบจึงจะทำให้ได้ค่าพารามิเตอร์ที่มีความน่าเชื่อถือและไม่แปรผันไปตามความสามารถของกลุ่มผู้สอบ (ศิริชัย กาญจนวาสี, 2555)

แนวคิดเกี่ยวกับ IRT สามารถจำแนกได้เป็น 2 ประเภท ได้แก่ ทฤษฎีการตอบสนองข้อสอบแบบตรวจให้คะแนน 2 ค่า หรือทวิวิภาค (binary or dichotomous IRT) ซึ่งเป็นโมเดล

การตอบสนองข้อสอบที่ใช้กับการตรวจคะแนนรายข้อแบบ 2 ค่า เช่น ข้อสอบหรือข้อคำถามที่ตรวจให้คะแนนแบบ 0,1 (ตอบผิดได้ 0, ตอบถูกได้ 1) แบบถูก/ผิด ใช่/ไม่ใช่ เป็นต้น และทฤษฎีการตอบสนองข้อสอบแบบตรวจให้คะแนนมากกว่า 2 ค่า หรือพหุวิภาค (polytomous IRT) ซึ่งเป็นโมเดลการตอบสนองข้อสอบที่ใช้กับการตรวจให้คะแนนรายข้อแบบมากกว่า 2 ค่า เช่น ข้อสอบหรือข้อคำถามมาตรประมาณค่า (rating scale) การตรวจข้อสอบแบบให้คะแนนความรู้บางส่วน (partial credit) เป็นต้น (ศิริชัย กาญจนวาสี, 2555) แต่เนื่องจากการวิจัยครั้งนี้ศึกษาเฉพาะข้อสอบแบบตรวจให้คะแนน 2 ค่า ดังนั้นผู้วิจัยจึงจะกล่าวถึงเฉพาะทฤษฎีการตอบสนองข้อสอบแบบตรวจให้คะแนน 2 ค่าหรือทวิวิภาค (dichotomous item response theory) โดยทฤษฎี IRT มีข้อตกลงเบื้องต้นว่า แบบสอบมุ่งวัดคุณลักษณะเดียว (unidimensionality) มีความเป็นอิสระระหว่างข้อสอบ (independence) โมเดลการตอบสนองข้อสอบมีรูปแบบเป็นฟังก์ชันโลจิสติค (logistic function) และแบบสอบที่ใช้ต้องไม่เป็นแบบสอบประเภทความเร็ว (speed test) (ศิริชัย กาญจนวาสี, 2555)

แนวคิดเกี่ยวกับทฤษฎีการตอบสนองข้อสอบแบบตรวจให้คะแนนแบบทวิวิภาค (dichotomous item response theory)

ทฤษฎีการตอบสนองข้อสอบจะอธิบายความสัมพันธ์ระหว่างความสามารถที่มีอยู่ภายในตัวบุคคลกับพฤติกรรมการตอบสนองข้อสอบของบุคคลนั้นว่ามีโอกาสตอบข้อสอบถูกมากน้อยเพียงใด โดยมีพื้นฐานความเชื่อว่าพฤติกรรมการตอบสนองต่อข้อสอบของผู้สอบซึ่งเป็นสิ่งที่สังเกตได้โดยตรงว่าถูกหรือผิด จะถูกกำหนดโดยคุณลักษณะภายในหรือความสามารถที่อยู่ภายในตัวบุคคลซึ่งเป็นสิ่งที่ไม่สามารถสังเกตได้โดยตรง โดยอธิบายความสัมพันธ์ดังกล่าวในรูปของฟังก์ชันคณิตศาสตร์ที่เรียกว่าฟังก์ชันการตอบสนองข้อสอบ ซึ่งมีลักษณะความสัมพันธ์เป็นแบบฟังก์ชันโลจิสหรือฟังก์ชันปกติสะสม

ฟังก์ชันการตอบสนองข้อสอบสามารถนำมาใช้ศึกษาความสัมพันธ์ระหว่างความน่าจะเป็นในการตอบข้อสอบแต่ละข้อได้ถูก [$P_i(\theta)$] กับระดับความสามารถของผู้สอบที่วัดได้โดยแบบสอบฉบับนั้น (θ) เมื่อนำมาเขียนกราฟจะได้โค้งลักษณะข้อสอบ (item characteristic curve: ICC) ซึ่งมีหลายลักษณะขึ้นอยู่กับโมเดล หรือแบบจำลองที่ใช้อธิบายความสัมพันธ์ดังกล่าว โมเดลที่นิยมใช้กัน คือ โมเดลแบบหนึ่งพารามิเตอร์ (one-parameter model) โมเดลแบบสองพารามิเตอร์ (two-parameter model) และโมเดลแบบสามพารามิเตอร์ (three-parameter model) โดยโมเดลการตอบสนองข้อสอบประกอบด้วยค่าพารามิเตอร์และค่าคงที่ดังนี้

1) พารามิเตอร์ผู้สอบ

θ = ระดับความสามารถของผู้สอบ ซึ่งประมาณได้จากโมเดลตามทฤษฎี IRT นิยมปรับให้เป็นคะแนนมาตรฐานที่มีค่าเฉลี่ยเป็น 0 และส่วนเบี่ยงเบนมาตรฐานเป็น 1 และมีพิสัยอยู่ระหว่าง $-\infty$ ถึง $+\infty$ แต่ผลการวิเคราะห์ส่วนใหญ่มักให้ค่าอยู่ระหว่าง -3 ถึง +3 โดย $P_i(\theta)$ คือความน่าจะเป็นที่ผู้ตอบซึ่งมีความสามารถ θ จะตอบข้อสอบ i ได้ถูกต้อง

2) พารามิเตอร์ของข้อสอบ

b_i = ค่าพารามิเตอร์ความยากของข้อสอบข้อที่ i (difficulty parameter) เป็นตำแหน่งของโค้งบนสเกลของความสามารถ (θ) ที่ทำให้มีโอกาสตอบข้อสอบได้ถูกต้อง สำหรับ

โมเดล 1 พารามิเตอร์ และ 2 พารามิเตอร์ $P_i(\theta) = 0.50$ ส่วนโมเดล 3 พารามิเตอร์ $P_i(\theta) = \frac{1+c_i}{2}$ ในทางทฤษฎี b_i มีค่าอยู่ระหว่าง -∞ ถึง +∞ โดย ศิริชัย กาญจนวาสี (2555) กล่าวว่าในทางปฏิบัติ นิยมใช้ข้อสอบที่มีค่า b_i อยู่ระหว่าง -2.50 ถึง +2.50 ค่า b_i ที่อยู่ใกล้ -2.50 แสดงว่าเป็นข้อสอบที่ง่าย ส่วนค่า b_i ที่อยู่ใกล้ +2.50 แสดงว่าเป็นข้อสอบที่ยาก

a_i = ค่าพารามิเตอร์อำนาจจำแนกของข้อสอบข้อที่ i (discrimination parameter) คือ การจำแนกค่าความต่างของ $P_i(\theta)$ ระหว่างผู้สอบที่มีความสามารถ $\leq \theta$ กับ $> \theta$ ซึ่งมีค่าเป็นสัดส่วนโดยตรงของค่าความชันของ ICC ที่ตำแหน่ง b_i โดยค่า a_i ที่สูงแสดงถึงการจำแนกผู้สอบที่มีความสามารถแตกต่างกันได้ดี ในทางทฤษฎีมีค่าระหว่าง -∞ ถึง +∞ โดย ศิริชัย กาญจนวาสี (2555) กล่าวว่าค่า a_i ควรมีค่าเป็น + ตามปกติมีค่าไม่เกิน +2.50 และในทางปฏิบัตินิยมใช้ข้อสอบที่มีค่า a_i อยู่ระหว่าง +0.50 ถึง +2.50

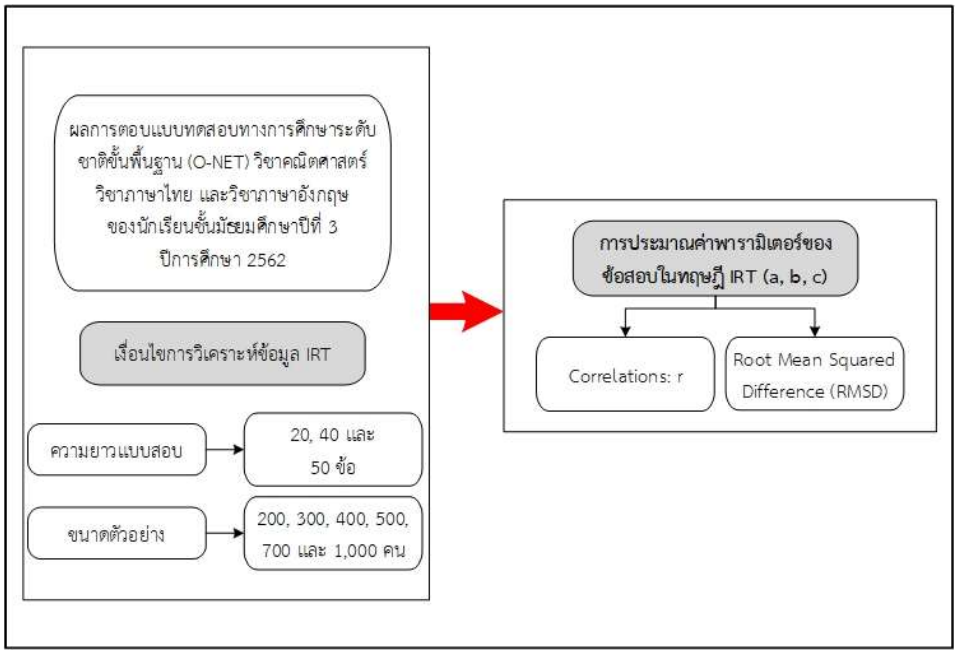
c_i = ค่าพารามิเตอร์โอกาสในการเดาข้อสอบได้ถูก (guessing parameter) คือ โอกาสในการตอบถูกของผู้สอบที่มีความสามารถต่ำ ในทางทฤษฎีมีค่าอยู่ระหว่าง 0 ถึง 1 โดย ศิริชัย กาญจนวาสี (2555) กล่าวว่าโดยทั่วไปนิยมใช้ข้อสอบที่มีค่า c_i ไม่เกิน .30 และตามปกติควรมีค่าต่ำกว่าโอกาสในการตอบถูกโดยการเดาตามทฤษฎี CTT ในขณะที่ Baker (2001) กล่าวว่าในทางปฏิบัติค่าพารามิเตอร์โอกาสในการเดาข้อสอบได้ถูกที่ยอมรับได้จะอยู่ระหว่าง 0 ถึง .35

3) ค่าคงที่

e = 2.71828 คือ ค่าคงที่ของลอการิทึมธรรมชาติ (natural log)

D = 1.70 คือ ค่าองค์ประกอบของการปรับสเกล (scaling factor) เพื่อทำให้ logistic function กับ normal ogive function ใกล้เคียงกัน หรือมีค่าประมาณ θ ต่างกันไม่เกิน 0.01

กรอบแนวคิดการวิจัย



ภาพ 1 กรอบแนวคิดการวิจัย

วิธีดำเนินการวิจัย

ผู้วิจัยดำเนินการวิจัยตามขั้นตอน ดังนี้

ประชากรและตัวอย่างวิจัย

ประชากร

ประชากรที่ใช้ในการวิจัยครั้งนี้ คือ นักเรียนชั้นมัธยมศึกษาปีที่ 3 ที่เข้ารับการทดสอบ O-NET ปีการศึกษา 2562 สังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน (สพฐ.) จำนวน 4 วิชา 3 ความยาวแบบสอบ เฉพาะแบบเลือกตอบที่ให้คะแนนแบบ 2 ค่า ดังนี้

1. วิชาคณิตศาสตร์ ความยาวแบบสอบ 20 ข้อ จำนวน 486,823 คน
2. วิชาภาษาไทย ความยาวแบบสอบ 40 ข้อ จำนวน 486,937 คน
3. วิชาวิทยาศาสตร์ ความยาวแบบสอบ 40 ข้อ จำนวน 486,594 คน
4. วิชาภาษาอังกฤษ ความยาวแบบสอบ 50 ข้อ จำนวน 486,681 คน

ตัวอย่างวิจัย

ตัวอย่างที่ใช้ในการวิจัยครั้งนี้ คือ นักเรียนชั้นมัธยมศึกษาปีที่ 3 สังกัด สพฐ. ที่เข้ารับการทดสอบ O-NET ปีการศึกษา 2562 จำนวน 3 วิชา ที่มีความยาวแบบสอบที่แตกต่างกัน ได้แก่ วิชาคณิตศาสตร์ (20 ข้อ) วิชาภาษาไทย (40 ข้อ) และวิชาภาษาอังกฤษ (50 ข้อ) ซึ่งผู้วิจัยมีวิธีการเลือกวิชาที่นำมาวิเคราะห์ และวิธีกำหนดขนาดตัวอย่าง และแผนการสุ่มตัวอย่าง ดังนี้

1. ในการสอบ O-NET ของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2562 สทศ. ได้กำหนดวิชาสอบ จำนวน 4 วิชา ซึ่งมีความยาวแบบสอบเฉพาะแบบเลือกตอบที่ให้คะแนนแบบ 2 ค่า จำนวน 3 ขนาดความยาว คือ 20, 40 และ 50 ข้อ โดยผู้วิจัยมีวิธีการเลือกวิชาที่จะนำมาวิเคราะห์ข้อมูล ดังนี้

1) ความยาวแบบสอบ 20 ข้อ คือ วิชาคณิตศาสตร์ เนื่องจากมีเพียงวิชาเดียว

2) ความยาวแบบสอบ 40 ข้อ มี 2 วิชา ได้แก่ วิชาภาษาไทย และวิชาวิทยาศาสตร์ ซึ่งจากการวิจัยของ ณัฏฐภรณ์ หลาวทอง และคณะ (2563) ที่ได้วิเคราะห์องค์ประกอบผลสัมฤทธิ์ทางการเรียนของนักเรียนในโรงเรียนสังกัด สพฐ. โดยใช้คะแนนสอบ O-NET ปี พ.ศ. 2560 ของ สทศ. จำนวน 4 วิชา ได้แก่ คณิตศาสตร์ วิทยาศาสตร์ ภาษาไทย และภาษาอังกฤษ พบว่า สามารถสกัดองค์ประกอบได้ 2 องค์ประกอบ โดยองค์ประกอบแรกประกอบด้วย วิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาไทย ส่วนองค์ประกอบที่สอง ได้แก่ วิชาภาษาอังกฤษ ซึ่งจากการวิจัยดังกล่าว พบว่า วิชาภาษาไทย และวิชาวิทยาศาสตร์ ที่มีความยาวแบบสอบเท่ากัน คือ วิชาละ 40 ข้อ ทั้งสองวิชาเป็นกลุ่มวิชาเดียวกัน ดังนั้นผู้วิจัยจึงทำการสุ่มอย่างง่ายเพื่อเลือกแบบสอบความยาว 40 ข้อ เพียงวิชาเดียว คือ วิชาภาษาไทย

3) ความยาวแบบสอบ 50 ข้อ คือ วิชาภาษาอังกฤษ เนื่องจากมีเพียงวิชาเดียว

2. ขนาดตัวอย่างสำหรับการประมาณค่าพารามิเตอร์ที่แท้จริง (true values)

ในการกำหนดขนาดตัวอย่าง ผู้วิจัยได้ศึกษางานของ Swaminathan et al. (2003) ซึ่งในฐานข้อมูลขนาดใหญ่ได้ใช้ขนาดตัวอย่าง จำนวน 5,000 คน เพื่อเป็นตัวแทนของค่าพารามิเตอร์ของข้อสอบที่แท้จริง นอกจากนี้ Lord (1968) กล่าวว่า ขนาดตัวอย่างประมาณ 3,000 คน อยู่ในระดับที่ยอมรับได้จากปัญหาความคลาดเคลื่อนที่เกิดจากการสุ่มตัวอย่าง

(sampling error) ดังนั้นในการวิจัยครั้งนี้ ผู้วิจัยจึงเลือกขนาดตัวอย่างที่เป็นตัวแทนของค่าพารามิเตอร์ที่แท้จริงแบบสอบละ 5,000 คน

3. ขนาดตัวอย่างที่ศึกษาตามเงื่อนไขต่าง ๆ

ในการกำหนดขนาดตัวอย่างที่ศึกษาตามเงื่อนไขต่าง ๆ ผู้วิจัยได้ศึกษางานวิจัยที่เกี่ยวข้องกับขนาดตัวอย่างต่ำสุดที่เหมาะสมสำหรับประมาณค่าพารามิเตอร์ข้อสอบที่มีความยาว 20, 40 และ 50 ข้อ ผู้วิจัยจึงเลือกเงื่อนไขขนาดตัวอย่างที่จะทำการศึกษา 6 เงื่อนไข ดังนี้

1) 200 คน เนื่องจากในแบบสอบ 50 ข้อ ขนาดตัวอย่างต่ำสุดที่เคยศึกษาและได้รับการแนะนำ คือ 300 คน (Chuah et al., 2006) แต่จากการศึกษาของ Weiss & Minden (2012) ในความยาวแบบสอบที่สั้นกว่า คือ แบบสอบ 25 ข้อ ได้แนะนำขนาดตัวอย่างที่เหมาะสม คือ 200 คน ดังนั้นในความยาวแบบสอบ 50 ข้อ จึงควรศึกษากับขนาดตัวอย่าง 200 คน

2) 300 คน เป็นขนาดตัวอย่างต่ำสุดที่ได้รับการแนะนำโดย Chuah et al. (2006) ในแบบสอบ 50 ข้อ

3) 400 คน เป็นขนาดตัวอย่างที่ยังไม่เคยมีการศึกษามาก่อน ในทุกความยาวแบบสอบ

4) 500 คน เป็นขนาดตัวอย่างปานกลาง ที่ควรศึกษากับทุกเงื่อนไขความยาวแบบสอบ

5) 700 คน เป็นขนาดตัวอย่างที่ยังไม่เคยมีการศึกษามาก่อน ในทุกความยาวแบบสอบ และเป็นเงื่อนไขระหว่าง 500 และ 750 คน ที่ศึกษาโดย Şahin & Anıl (2017) ซึ่งผลการศึกษาพบว่า ขนาดตัวอย่าง 750 คน เหมาะสำหรับความยาวแบบสอบ 20 ข้อ

6) 1,000 คน เป็นขนาดตัวอย่างต่ำสุดที่ได้รับการแนะนำสำหรับแบบสอบ 40 ข้อ (Patsula & Gessaroli, 1995; Tang et al., 1993; Yen, 1987)

4. แผนการสุ่มตัวอย่าง

การวิจัยครั้งนี้ ผู้วิจัยใช้แผนการสุ่มตัวอย่างแบบมีระบบ (systematic random sampling) โดยมีขั้นตอนการสุ่มตัวอย่างในแต่ละขั้น ดังนี้

1) ผู้วิจัยนำไฟล์ข้อมูลที่ได้มาจัดเรียงตามภูมิภาค จังหวัด ที่ตั้งโรงเรียน ขนาดโรงเรียน และเพศ

2) ใช้การสุ่มตัวอย่างแบบมีระบบ (systematic random sampling) โดย

- เลือกหน่วยตัวอย่างมา 1 หน่วย จากทุก ๆ k หน่วย
- k เป็นจำนวนเต็มที่ใกล้เคียง N/n เช่น กรณีขนาดตัวอย่าง 200 คน k = 5,000/200 = 25
- สุ่มตัวอย่างจนครบจำนวนตามเงื่อนไขที่ต้องการ

เครื่องมือที่ใช้ในการวิจัย

เครื่องมือที่ใช้ในการวิจัยครั้งนี้ ได้แก่ ผลการตอบข้อสอบรายข้อจากแบบทดสอบ O-NET ชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2562 จำนวน 3 วิชา ที่เป็นข้อสอบแบบเลือกตอบซึ่งให้คะแนนแบบ 2 ค่า วิชาคณิตศาสตร์ จำนวน 20 ข้อ วิชาภาษาไทย จำนวน 40 ข้อ และวิชาภาษาอังกฤษ จำนวน 50 ข้อ

การเก็บรวบรวมข้อมูล

ข้อมูลที่ใช้ในการวิจัยครั้งนี้เป็นข้อมูลทุติยภูมิ โดยผู้วิจัยทำหนังสือขอความอนุเคราะห์ข้อมูลจากสถาบันทดสอบทางการศึกษาแห่งชาติ (องค์การมหาชน) หรือ สทศ.

การวิเคราะห์ข้อมูล

1. ตรวจสอบความสมบูรณ์ของข้อมูลผลการตอบข้อสอบ O-NET ที่ได้รับจาก สทศ. ก่อนการวิเคราะห์

2. ทำการสุ่มตัวอย่างที่ใช้ในการวิจัยตามเงื่อนไขที่ศึกษาจำแนกตามความยาวแบบสอบ

3. ตรวจสอบข้อตกลงเบื้องต้นของทฤษฎี IRT ด้วยการตรวจสอบความเป็นเอกมิติ (unidimensionality) และความเป็นอิสระ (local independence) อย่างไรก็ตาม Warm (1978) กล่าวว่าถ้าแบบสอบมีความเป็นเอกมิติแล้ว แสดงว่าแบบสอบนั้นมีความเป็นอิสระต่อกัน แต่ถ้าแบบสอบมีความเป็นอิสระต่อกันแล้วยังไม่เพียงพอที่จะสรุปได้ว่าแบบสอบนั้นมีความเป็นเอกมิติหรือไม่ ในขณะที่ Dorans (1985, as cited in Sireci, 1991), Hambleton (1989, as cited in Sireci, 1991) และ Hambleton et al. (1991) กล่าวว่าข้อตกลงเบื้องต้นหลักของ IRT คือ ความเป็นเอกมิติ หากข้อสอบมีความเป็นเอกมิติแล้วก็แสดงว่าข้อตกลงเบื้องต้นอื่น ๆ ก็จะได้รับการยอมรับด้วย เช่น ความเป็นอิสระ เป็นต้น

ดังนั้นในการวิจัยครั้งนี้ ผู้วิจัยจึงทำการตรวจสอบเฉพาะความเป็นเอกมิติเท่านั้น โดยใช้วิธีการวิเคราะห์องค์ประกอบเชิงสำรวจ (exploratory factor analysis) ด้วยวิธีวิเคราะห์องค์ประกอบหลัก (principal component analysis) โดยใช้โปรแกรม SPSS เวอร์ชัน 26 และมีเกณฑ์พิจารณาความเป็นเอกมิติ 2 เกณฑ์ คือ เกณฑ์ของ Reckase (1979) และ Reeve et al. (2007) ที่ให้พิจารณาค่าความแปรปรวนขององค์ประกอบแรก ซึ่งถ้าสามารถอธิบายความแปรปรวนทั้งหมดได้อย่างน้อย 20% จะบ่งบอกถึงความเป็นเอกมิติ และเกณฑ์พิจารณาความเป็นเอกมิติจากค่าอัตราส่วนระหว่างค่าไอเกน (eigenvalue: λ) ขององค์ประกอบแรกต่อค่าไอเกนขององค์ประกอบที่สอง (λ_1/λ_2) หากมีค่ามากกว่า 4.00 จะบ่งบอกถึงความเป็นเอกมิติ (John et al., 2014; Reeve et al., 2007)

4. เปรียบเทียบความเหมาะสมระหว่างโมเดล 1, 2 และ 3 พารามิเตอร์ ด้วยการตรวจสอบความสอดคล้องของข้อมูลเชิงประจักษ์กับโมเดลต่าง ๆ ซึ่งจากการศึกษางานวิจัยของ Chen et al. (2018) ที่ได้ทำการเปรียบเทียบความสอดคล้องกลมกลืนระหว่างโมเดลด้วยค่า Akaike information criterion (*AIC*) และ Bayesian information criterion (*BIC*) โดยโมเดลที่มีค่า *AIC* และ *BIC* น้อยกว่า จะเป็นโมเดลที่ดีกว่า สอดคล้องกับ Hooper et al. (2008) ที่กล่าวว่าโมเดลที่มีค่า *AIC* น้อยกว่า คือ โมเดลที่มีความสอดคล้องกลมกลืนมากกว่า ส่วน Kang & Cohen (2007) กล่าวว่าโมเดล IRT ที่มีค่า *AIC* น้อยที่สุดคือโมเดลที่ถูกเลือก และโมเดลที่มีค่า *-2log.Lik* หรือ ค่า deviance (*d*) น้อยกว่า คือ โมเดลที่มีความสอดคล้องกลมกลืมมากกว่า

5. วิเคราะห์เปรียบเทียบผลของความยาวแบบสอบและขนาดตัวอย่างที่มีต่อการประมาณค่าพารามิเตอร์ของข้อสอบในทฤษฎีการตอบสนองข้อสอบตามความเหมาะสมของโมเดลหรือความสอดคล้องกลมกลืมกับข้อมูลเชิงประจักษ์ที่วิเคราะห์ได้จากข้อ 4 โดยผู้วิจัยใช้เกณฑ์สำหรับเปรียบเทียบ จำนวน 2 เกณฑ์ ตามเกณฑ์ของ Şahin & Anıl (2017) ดังนี้

5.1 ค่าสัมประสิทธิ์สหสัมพันธ์ (product-moment correlations: *r*) คือ ค่าความสัมพันธ์ระหว่างค่าประมาณค่าพารามิเตอร์ของข้อสอบที่ได้จากการประมาณค่ากับค่าพารามิเตอร์ที่แท้จริง โดยค่าที่ยอมรับได้ คือ $r \geq .70$

5.2 ค่ารากที่สองของความแตกต่างกำลังสองเฉลี่ย (root mean squared difference: *RMSD*) คือ ค่าความแตกต่างระหว่างการประมาณค่าพารามิเตอร์ของข้อสอบที่ได้จากการประมาณค่ากับค่าพารามิเตอร์ที่แท้จริง โดยค่าที่ยอมรับได้คือ $RMSD \leq 0.33$ โดย *RMSD* มีสูตรในการคำนวณ คือ

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n}(\bar{\pi}_i - \pi_i)^2}{n}}$$

โดยที่ $\bar{\pi}_i$ คือ ค่าพารามิเตอร์ของข้อสอบที่ได้จากการประมาณค่า (*a*, *b*, หรือ *c*) ของข้อสอบข้อที่ i

π_i คือ ค่าพารามิเตอร์ของข้อสอบข้อที่ i ที่แท้จริงหรือค่าฐาน

n คือ จำนวนข้อสอบทั้งหมด

6. สรุปผลการวิเคราะห์ข้อมูลและนำเสนอผลการวิจัย

ผลการวิจัย

1. ผลการตรวจสอบข้อตกลงเบื้องต้นของ IRT

การตรวจสอบข้อตกลงเบื้องต้นของ IRT ครั้งนี้ ผู้วิจัยทำการตรวจสอบเฉพาะความเป็นเอกมิติ (unidimensionality) ของแบบทดสอบ O-NET ชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2562 พบว่าแบบสอบทั้ง 3 ชุด ไม่ผ่านเกณฑ์ข้อตกลงเบื้องต้นด้านความเป็นเอกมิติของการวัดทั้ง 2 เกณฑ์ โดยกรณีใช้เกณฑ์ค่าความแปรปรวนขององค์ประกอบแรกอย่างน้อย 20% จะบ่งบอกถึงความเป็นเอกมิติ พบว่าไม่ผ่านเกณฑ์ทุกความยาวแบบสอบ ซึ่งแบบสอบที่มีร้อยละความแปรปรวนขององค์ประกอบแรกมากที่สุด เรียงตามลำดับ คือ แบบสอบ 40 ข้อ (15.77%) แบบสอบ 20 ข้อ (15.70%) และแบบสอบ 50 ข้อ (9.58%) ตามลำดับ

ส่วนกรณีใช้เกณฑ์ค่าอัตราส่วนระหว่างค่าไอเกนขององค์ประกอบแรกต่อค่าไอเกนขององค์ประกอบที่สอง (λ_1/λ_2) หากมีค่ามากกว่าหรือเท่ากับ 4.00 จะบ่งบอกถึงความเป็นเอกมิติ พบว่าไม่ผ่านเกณฑ์ทุกความยาวแบบสอบ โดยแบบสอบที่มีค่า λ_1/λ_2 มากที่สุดเรียงตามลำดับ คือ แบบสอบ 50 ข้อ ($\lambda_1/\lambda_2 = 3.12$) แบบสอบ 40 ข้อ ($\lambda_1/\lambda_2 = 2.95$) และแบบสอบ 20 ข้อ ($\lambda_1/\lambda_2 = 2.10$) ตามลำดับ โดยผลการวิเคราะห์ข้อมูล แสดงดังตาราง 1

ตาราง 1 ผลการตรวจสอบข้อตกลงเบื้องต้นของทฤษฎีการตอบสนองข้อสอบ (IRT)

ความยาวแบบสอบ (ข้อ)	ผลการวิเคราะห์ข้อมูล	ผลการพิจารณา
20 (คณิตศาสตร์)	1. ร้อยละความแปรปรวนขององค์ประกอบที่ 1 = 15.70	ไม่ผ่านเกณฑ์
	2. λ_1/λ_2 = 2.10	ไม่ผ่านเกณฑ์
40 (ภาษาไทย)	1. ร้อยละความแปรปรวนขององค์ประกอบที่ 1 = 15.77	ไม่ผ่านเกณฑ์
	2. λ_1/λ_2 = 2.95	ไม่ผ่านเกณฑ์
50 (ภาษาอังกฤษ)	1. ร้อยละความแปรปรวนขององค์ประกอบที่ 1 = 9.58	ไม่ผ่านเกณฑ์
	2. λ_1/λ_2 = 3.12	ไม่ผ่านเกณฑ์

2. ผลการเปรียบเทียบความสอดคล้องกลมกลืนระหว่างโมเดลแบบ 1, 2 และ 3 พารามิเตอร์

ผู้วิจัยได้ทำการเปรียบเทียบความสอดคล้องกลมกลืนระหว่างโมเดล 1, 2 และ 3 พารามิเตอร์ ของแบบทดสอบ O-NET ชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2562 ทั้ง 3 ความยาวแบบสอบ พบว่าแบบสอบทั้ง 3 ชุด มีความสอดคล้องกลมกลืนกับโมเดลแบบ 3 พารามิเตอร์ (3PL) มากที่สุด โดยแสดงรายละเอียด ดังนี้

2.1 แบบสอบ 20 ข้อ (วิชาคณิตศาสตร์)

ผลการเปรียบเทียบความสอดคล้องกลมกลืนของโมเดลการตอบสนองข้อสอบระหว่างโมเดล 1, 2 และ 3 พารามิเตอร์ ของแบบทดสอบ O-NET ชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2562 วิชาคณิตศาสตร์ เพื่อเลือกโมเดลที่มีความสอดคล้องกลมกลืนมากที่สุด พบว่า ผลการเปรียบเทียบค่า *AIC*, *BIC* และ *-2log.Lik* เป็นไปในทิศทางเดียวกัน คือ โมเดล 3PL มีค่า *AIC*, *BIC* และ *-2log.Lik* น้อยที่สุด จึงเป็นโมเดลที่ดีที่สุด นอกจากนี้เมื่อทำการทดสอบนัยสำคัญของความสอดคล้องกลมกลืนของโมเดลทีละคู่ด้วยการทดสอบอัตราส่วนความเป็นไปได้ พบว่าโมเดล 3PL แตกต่างจากโมเดล 1PL และ 2PL อย่างมีนัยสำคัญทางสถิติ ($p < .001$) ดังนั้น โมเดล 3PL จึงเป็นโมเดลที่สอดคล้องกลมกลืนกับข้อมูลเชิงประจักษ์มากที่สุด ดังตาราง 2

ตาราง 2 ผลการเปรียบเทียบความสอดคล้องกลมกลืนระหว่างโมเดล 1, 2 และ 3 พารามิเตอร์ของแบบทดสอบ O-NET ชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2562 วิชาคณิตศาสตร์

วิชาคณิตศาสตร์	โมเดล 1PL	โมเดล 2PL	โมเดล 3PL
จำนวนตัวอย่าง	5,000	5,000	5,000
จำนวนข้อสอบ	20	20	20
AIC	120,446.2	119,342.2	118,354.7
BIC	120,576.5	119,602.9	118,745.8
-2log.Lik	120,406.2	119,262.2	118,234.8
1PL vs. 2PL	*LRT* = 1,144.0, *df* = 20, $p < .001$		
1PL vs. 3PL	*LRT* = 2,171.4, *df* = 40, $p < .001$		
2PL vs. 3PL	*LRT* = 1,027.4, *df* = 20, $p < .001$		

2.2 แบบสอบ 40 ข้อ (วิชาภาษาไทย)

ผลการเปรียบเทียบความสอดคล้องกลมกลืนของโมเดลการตอบสนองข้อสอบระหว่างโมเดล 1, 2 และ 3 พารามิเตอร์ ของแบบทดสอบ O-NET ชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2562 วิชาภาษาไทย เพื่อเลือกโมเดลที่มีความสอดคล้องกลมกลืนมากที่สุด พบว่าผลการเปรียบเทียบค่า *AIC*, *BIC* และ *-2log.Lik* เป็นไปในทิศทางเดียวกัน คือ โมเดล 3PL มีค่า *AIC*, *BIC* และ *-2log.Lik* น้อยที่สุด จึงเป็นโมเดลที่ดีที่สุด นอกจากนี้เมื่อทำการทดสอบนัยสำคัญของความสอดคล้องกลมกลืนของโมเดลทีละคู่ด้วยการทดสอบอัตราส่วนความเป็นไปได้ พบว่า โมเดล 3PL แตกต่างจากโมเดล 1PL และ 2PL อย่างมีนัยสำคัญทางสถิติ ($p < .001$) ดังนั้น โมเดล 3PL จึงเป็นโมเดลที่สอดคล้องกลมกลืนกับข้อมูลเชิงประจักษ์มากที่สุด ดังตาราง 3

ตาราง 3 ผลการเปรียบเทียบความสอดคล้องกลมกลืนระหว่างโมเดล 1, 2 และ 3 พารามิเตอร์ของแบบทดสอบ O-NET ชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2562 วิชาภาษาไทย

วิชาภาษาไทย	โมเดล 1PL	โมเดล 2PL	โมเดล 3PL
จำนวนตัวอย่าง	5,000	5,000	5,000
จำนวนข้อสอบ	40	40	40
AIC	241,620.6	238,764.4	238,124.8
BIC	241,881.3	239,285.8	238,906.8
-2log.Lik	241,540.6	238,604.4	237,884.8
1PL vs. 2PL	*LRT* = 2,936.2, *df* = 40, $p < .001$		
1PL vs. 3PL	*LRT* = 3,655.9, *df* = 80, $p < .001$		
2PL vs. 3PL	*LRT* = 719.7, *df* = 40, $p < .001$		

2.3 แบบสอบ 50 ข้อ (วิชาภาษาอังกฤษ)

ผลการเปรียบเทียบความสอดคล้องกลมกลืนของโมเดลการตอบสนองข้อสอบระหว่างโมเดล 1, 2 และ 3 พารามิเตอร์ ของแบบทดสอบ O-NET ชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2562 วิชาภาษาอังกฤษ เพื่อเลือกโมเดลที่มีความสอดคล้องกลมกลืนมากที่สุด พบว่า ผลการเปรียบเทียบค่า *AIC*, *BIC* และ *-2log.Lik* เป็นไปในทิศทางเดียวกัน คือ โมเดล 3PL มีค่า *AIC*, *BIC* และ *-2log.Lik* น้อยที่สุด จึงเป็นโมเดลที่ดีที่สุด นอกจากนี้เมื่อทำการทดสอบนัยสำคัญของความสอดคล้องกลมกลืนของโมเดลทีละคู่ด้วยการทดสอบอัตราส่วนความเป็นไปได้ พบว่า โมเดล 3PL แตกต่างจากโมเดล 1PL และ 2PL อย่างมีนัยสำคัญทางสถิติ ($p < .001$) ดังนั้น โมเดล 3PL จึงเป็นโมเดลที่สอดคล้องกลมกลืนกับข้อมูลเชิงประจักษ์มากที่สุด ดังตาราง 4

ตาราง 4 ผลการเปรียบเทียบความสอดคล้องกลมกลืนระหว่างโมเดล 1, 2 และ 3 พารามิเตอร์ของแบบทดสอบ O-NET ชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2562 วิชาภาษาอังกฤษ

วิชาภาษาอังกฤษ	โมเดล 1PL	โมเดล 2PL	โมเดล 3PL
จำนวนตัวอย่าง	5,000	5,000	5,000
จำนวนข้อสอบ	50	50	50
AIC	306,267.8	301,167.0	298,164.7
BIC	306,593.7	301,818.7	299,142.3
-2log.Lik	306,167.8	300,967.0	297,864.6
1PL vs. 2PL	LRT = 5,200.9, df = 50, p < .001		
1PL vs. 3PL	LRT = 8,303.1, df = 100, p < .001		
2PL vs. 3PL	LRT = 3,102.3, df = 50, p < .001		

3. ผลการเปรียบเทียบความยาวแบบสอบและขนาดตัวอย่างที่มีต่อการประมาณค่าพารามิเตอร์ของข้อสอบ (*a*, *b*, *c*) ในทฤษฎี IRT

การวิเคราะห์เปรียบเทียบความยาวแบบสอบและขนาดตัวอย่างที่มีต่อการประมาณค่าพารามิเตอร์ของข้อสอบในทฤษฎี IRT ของการทดสอบ O-NET ชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2562 ประกอบด้วยความยาวแบบสอบ 3 ขนาด ได้แก่ ความยาวแบบสอบ 20, 40 และ 50 ข้อ และศึกษาขนาดตัวอย่างตามเงื่อนไข 6 ขนาด ได้แก่ 200, 300, 400, 500, 700 และ 1,000 คน ตามลำดับ ผู้วิจัยนำเสนอผลการวิเคราะห์ข้อมูลตามลำดับ ดังนี้

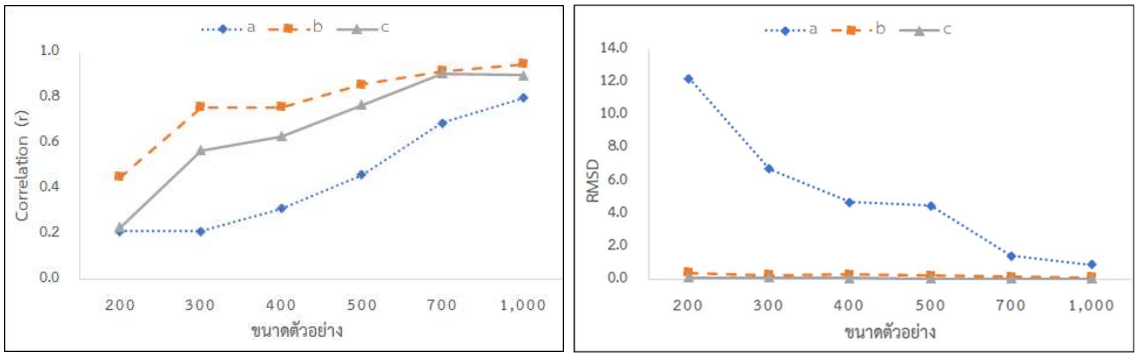
4.1 ความยาวแบบสอบ 20 ข้อ

จากภาพ 2 ผลการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ (r) และค่ารากที่สองของความแตกต่างกำลังสองเฉลี่ย (*RMSD*) ของการประมาณค่าพารามิเตอร์ของข้อสอบภายใต้เงื่อนไขความยาวแบบสอบ 20 ข้อ และขนาดตัวอย่างต่างกัน โดยเมื่อพิจารณาค่าอำนาจจำแนก (*a*) พบว่า r มีค่าอยู่ระหว่าง .21 (ขนาดตัวอย่าง 200 คน) ถึง .80 (ขนาดตัวอย่าง 1,000 คน) และ *RMSD* มีค่าอยู่ระหว่าง 0.91 (ขนาดตัวอย่าง 1,000 คน) ถึง 12.27 (ขนาดตัวอย่าง 200 คน) ส่วนค่าความยาก (*b*) พบว่า r มีค่าอยู่ระหว่าง .45 (ขนาดตัวอย่าง 200 คน) ถึง .95 (ขนาดตัวอย่าง 1,000 คน) และ *RMSD* มีค่าอยู่ระหว่าง 0.12 (ขนาดตัวอย่าง 1,000 คน) ถึง 0.40 (ขนาดตัวอย่าง 200 คน) และค่าโอกาสการเดา (*c*) พบว่า r มีค่าอยู่ระหว่าง .23 (ขนาดตัวอย่าง 200 คน) ถึง .91 (ขนาดตัวอย่าง 700 คน) และ *RMSD* มีค่าอยู่ระหว่าง 0.04 (ขนาดตัวอย่าง 700 และ 1,000 คน) ถึง 0.13 (ขนาดตัวอย่าง 200 คน)

สรุปขนาดตัวอย่างที่ผ่านเกณฑ์ $r \geq .7$ ของพารามิเตอร์ *a* คือ 1,000 คน พารามิเตอร์ *b* คือ 300, 400, 500, 700 และ 1,000 คน และพารามิเตอร์ *c* คือ 500, 700 และ 1,000 คน ส่วนขนาดตัวอย่างที่ผ่านเกณฑ์ *RMSD* ≤ 0.33 ของพารามิเตอร์ *a* คือ ไม่มีขนาดตัวอย่างที่ผ่านเกณฑ์ พารามิเตอร์ *b* คือ 300, 400, 500, 700 และ 1,000 คน และพารามิเตอร์ *c* คือ 200, 300 400, 500, 700 และ 1,000 คน

จากผลการประมาณค่าพารามิเตอร์ของข้อสอบ (*a*, *b*, *c*) ภายใต้เงื่อนไขความยาวแบบสอบ 20 ข้อ และขนาดตัวอย่างที่แตกต่างกันตั้งแต่ 200 - 1,000 คน พบว่าไม่มีขนาดตัวอย่าง

ที่ผ่านเกณฑ์ $r \geq .7$ และ *RMSD* ≤ 0.33 ในทุกพารามิเตอร์ของข้อสอบ (*a*, *b*, *c*) ดังนั้นสำหรับแบบสอบที่มีความยาว 20 ข้อ จึงพบว่าขนาดตัวอย่างขั้นต่ำ 1,000 คน ยังไม่เหมาะสมเพียงพอที่จะทำให้การประมาณค่าพารามิเตอร์ของข้อสอบมีความแม่นยำ จึงควรใช้ขนาดตัวอย่างมากกว่า 1,000 คน



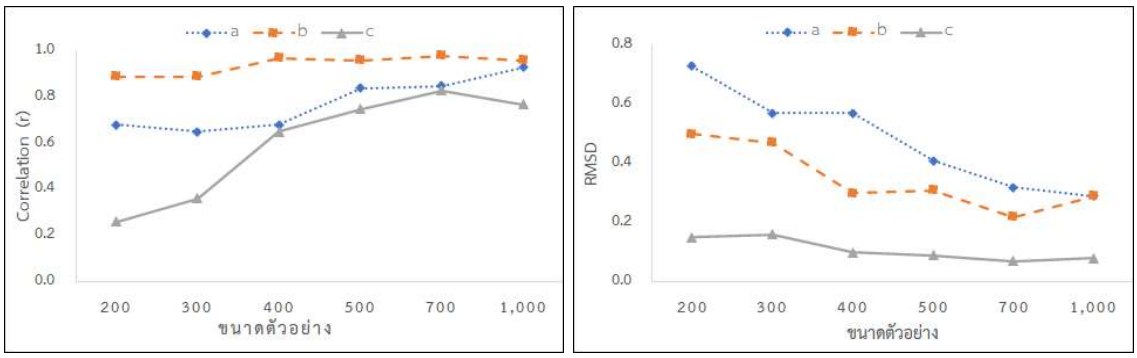
ภาพ 2 ค่าสัมประสิทธิ์สหสัมพันธ์ (*r*) และค่ารากที่สองของความแตกต่างกำลังสองเฉลี่ย (*RMSD*) ของการประมาณค่าพารามิเตอร์ของข้อสอบ ภายใต้เงื่อนไขความยาวแบบสอบ 20 ข้อ และขนาดตัวอย่างต่างกัน

4.2 ความยาวแบบสอบ 40 ข้อ

จากภาพ 3 ผลการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ (*r*) และค่ารากที่สองของความแตกต่างกำลังสองเฉลี่ย (*RMSD*) ของการประมาณค่าพารามิเตอร์ของข้อสอบ ภายใต้เงื่อนไขความยาวแบบสอบ 40 ข้อ และขนาดตัวอย่างต่างกัน โดยเมื่อพิจารณาค่าอำนาจจำแนก (*a*) พบว่า *r* มีค่าอยู่ระหว่าง .65 (ขนาดตัวอย่าง 300 คน) ถึง .93 (ขนาดตัวอย่าง 1,000 คน) และ *RMSD* มีค่าอยู่ระหว่าง 0.29 (ขนาดตัวอย่าง 1,000 คน) ถึง 0.73 (ขนาดตัวอย่าง 200 คน) ส่วนค่าความยาก (*b*) พบว่า *r* มีค่าอยู่ระหว่าง .89 (ขนาดตัวอย่าง 200 และ 300 คน) ถึง .98 (ขนาดตัวอย่าง 700 คน) และ *RMSD* มีค่าอยู่ระหว่าง 0.22 (ขนาดตัวอย่าง 700 คน) ถึง 0.50 (ขนาดตัวอย่าง 200 คน) และค่าโอกาสการเดา (*c*) พบว่า *r* มีค่าอยู่ระหว่าง .26 (ขนาดตัวอย่าง 200 คน) ถึง .83 (ขนาดตัวอย่าง 700 คน) และ *RMSD* มีค่าอยู่ระหว่าง 0.07 (ขนาดตัวอย่าง 700 คน) ถึง 0.16 (ขนาดตัวอย่าง 300 คน)

สรุปขนาดตัวอย่างที่ผ่านเกณฑ์ $r \geq .7$ ของพารามิเตอร์ *a* คือ 500, 700 และ 1,000 คน พารามิเตอร์ *b* คือ 200, 300, 400, 500, 700 และ 1,000 คน และพารามิเตอร์ *c* คือ 500, 700 และ 1,000 คน ส่วนขนาดตัวอย่างที่ผ่านเกณฑ์ *RMSD* ≤ 0.33 ของพารามิเตอร์ *a* คือ 700 และ 1,000 คน พารามิเตอร์ *b* คือ 400, 500, 700 และ 1,000 คน และพารามิเตอร์ *c* คือ 200, 300, 400, 500, 700 และ 1,000 คน

จากผลการประมาณค่าพารามิเตอร์ของข้อสอบ (*a*, *b*, *c*) ภายใต้เงื่อนไขความยาวแบบสอบ 40 ข้อ และขนาดตัวอย่างที่แตกต่างกันตั้งแต่ 200 - 1,000 คน พบว่า ขนาดตัวอย่างต่ำสุดที่ผ่านเกณฑ์ $r \geq .7$ และ *RMSD* ≤ 0.33 ในทุกพารามิเตอร์ของข้อสอบ (*a*, *b*, *c*) คือ 700 และ 1,000 คน ดังนั้นแบบสอบที่มีความยาว 40 ข้อ จึงควรใช้ขนาดตัวอย่างขั้นต่ำ 700 คน



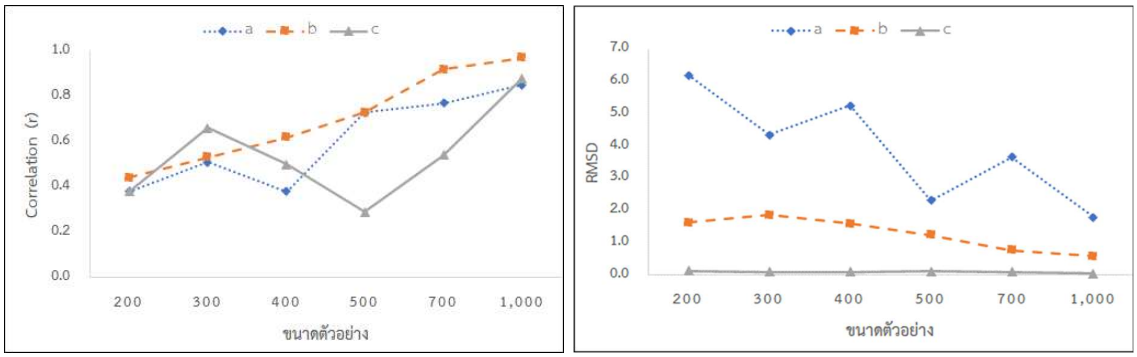
ภาพ 3 ค่าสัมประสิทธิ์สหสัมพันธ์ (*r*) และค่ารากที่สองของความแตกต่างกำลังสองเฉลี่ย (*RMSD*) ของการประมาณค่าพารามิเตอร์ของข้อสอบ ภายใต้เงื่อนไขความยาวแบบสอบ 40 ข้อ และขนาดตัวอย่างต่างกัน

4.3 ความยาวแบบสอบ 50 ข้อ

จากภาพ 4 ผลการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ (*r*) และค่ารากที่สองของความแตกต่างกำลังสองเฉลี่ย (*RMSD*) ของการประมาณค่าพารามิเตอร์ของข้อสอบภายใต้เงื่อนไขความยาวแบบสอบ 50 ข้อ และขนาดตัวอย่างต่างกัน โดยเมื่อพิจารณาค่าอำนาจจำแนก (*a*) พบว่า *r* มีค่าอยู่ระหว่าง .38 (ขนาดตัวอย่าง 200 และ 400 คน) ถึง .85 (ขนาดตัวอย่าง 1,000 คน) และ *RMSD* มีค่าอยู่ระหว่าง 1.78 (ขนาดตัวอย่าง 1,000 คน) ถึง 6.18 (ขนาดตัวอย่าง 200 คน) ส่วนค่าความยาก (*b*) พบว่า *r* มีค่าอยู่ระหว่าง .44 (ขนาดตัวอย่าง 200 คน) ถึง .97 (ขนาดตัวอย่าง 1,000 คน) และ *RMSD* มีค่าอยู่ระหว่าง 0.57 (ขนาดตัวอย่าง 1,000 คน) ถึง 1.84 (ขนาดตัวอย่าง 300 คน) และค่าโอกาสการเดา (*c*) พบว่า *r* มีค่าอยู่ระหว่าง .29 (ขนาดตัวอย่าง 500 คน) ถึง .88 (ขนาดตัวอย่าง 1,000 คน) และ *RMSD* มีค่าอยู่ระหว่าง 0.04 (ขนาดตัวอย่าง 1,000 คน) ถึง 0.11 (ขนาดตัวอย่าง 200 คน)

สรุปขนาดตัวอย่างที่ผ่านเกณฑ์ $r \geq .7$ ของพารามิเตอร์ *a* คือ 500, 700 และ 1,000 คน พารามิเตอร์ *b* คือ 500, 700 และ 1,000 คน และพารามิเตอร์ *c* คือ 1,000 คน ส่วนขนาดตัวอย่างที่ผ่านเกณฑ์ $RMSD \leq 0.33$ ของพารามิเตอร์ *a* และ *b* คือ ไม่มีขนาดตัวอย่างที่ผ่านเกณฑ์ ส่วนพารามิเตอร์ *c* คือ 200, 300, 400, 500, 700 และ 1,000 คน

จากผลการประมาณค่าพารามิเตอร์ของข้อสอบ (*a*, *b*, *c*) ภายใต้เงื่อนไขความยาวแบบสอบ 50 ข้อ และขนาดตัวอย่างที่แตกต่างกันตั้งแต่ 200 - 1,000 คน พบว่าไม่มีขนาดตัวอย่างที่ผ่านเกณฑ์ $r \geq .7$ และ $RMSD \leq 0.33$ ในทุกพารามิเตอร์ของข้อสอบ (*a*, *b*, *c*) ดังนั้นสำหรับแบบสอบที่มีความยาว 50 ข้อ จึงพบว่าขนาดตัวอย่างขั้นต่ำ 1,000 คน ยังไม่เหมาะสมเพียงพอที่จะทำให้การประมาณค่าพารามิเตอร์ของข้อสอบมีความแม่นยำ จึงควรใช้ขนาดตัวอย่างมากกว่า 1,000 คน



ภาพ 4 ค่าสัมประสิทธิ์สหสัมพันธ์ (*r*) และค่ารากที่สองของความแตกต่างกำลังสองเฉลี่ย (*RMSD*) ของการประมาณค่าพารามิเตอร์ของข้อสอบ ภายใต้เงื่อนไขความยาวแบบสอบ 50 ข้อและขนาดตัวอย่างต่างกัน

อภิปรายผลการวิจัย

ผู้วิจัยอภิปรายผลการวิจัย ดังนี้

1. ผลการตรวจสอบข้อตกลงเบื้องต้นของ IRT ของแบบสอบ O-NET ทั้ง 3 ความยาวแบบสอบ พบว่า ไม่ผ่านเกณฑ์ข้อตกลงเบื้องต้นด้านความเป็นเอกมิติ (unidimensionality) ซึ่งสอดคล้องกับคำกล่าวของ Hambleton et al. (1991) และ Kose & Demirtasli (2012) ที่ว่าในแบบสอบวัดผลสัมฤทธิ์และความสามารถทางด้านการศึกษาและจิตวิทยานั้นเป็นเรื่องยากที่จะผ่านข้อตกลงเบื้องต้นด้านความเป็นเอกมิติ เนื่องจากหลาย ๆ ปัจจัย เช่น สติปัญญา บุคลิกลักษณะ หรือปัจจัยส่วนตัวของผู้สอบ อันได้แก่ แรงจูงใจ ความเครียด หรือการเดาข้อสอบ เป็นต้น ซึ่งสิ่งเหล่านี้จะทำให้เกิดปัญหาด้านความตรงของการประมาณค่าพารามิเตอร์ของข้อสอบได้ ดังนั้นเพื่อให้ผลการวิจัยมีความตรงและความเที่ยงจึงอาจประยุกต์ใช้ทฤษฎีการตอบสนองข้อสอบแบบพหุมิติ (multidimensional item response theory: MIRT)

2. ผลการเปรียบเทียบความสอดคล้องกลมกลืนกับข้อมูลเชิงประจักษ์ระหว่างโมเดลแบบ 1, 2 และ 3 พารามิเตอร์ ของแบบทดสอบ O-NET ชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2562 ทั้ง 3 ความยาวแบบสอบ ที่มีรูปแบบข้อสอบเป็นแบบปรนัย 4 ตัวเลือก 1 คำตอบ พบว่า แบบสอบทั้ง 3 ชุด มีความสอดคล้องกลมกลืนกับโมเดลแบบ 3 พารามิเตอร์ (3PL) มากที่สุด ซึ่งสอดคล้องกับ DeMars (2008) และ ศิริชัย กาญจนวาสี (2555) ที่กล่าวว่าสำหรับโมเดลการตอบสนองข้อสอบแบบ 3 พารามิเตอร์ (three-parameter logistic model: 3PL) ข้อสอบแต่ละข้อมีความแตกต่างกันได้ทั้งพารามิเตอร์ *a*, *b* และ *c* โมเดลนี้จึงเหมาะสำหรับใช้กับข้อสอบแบบเลือกตอบทั่วไป ข้อสอบแบบหลายตัวเลือก เนื่องจากผู้สอบสามารถเดาคำตอบได้

3. ผลการเปรียบเทียบความยาวแบบสอบและขนาดตัวอย่างที่มีต่อการประมาณค่าพารามิเตอร์ของข้อสอบ (*a*, *b*, *c*) ในทฤษฎี IRT ในแบบสอบ 20 ข้อ พบว่าขนาดตัวอย่างขั้นต่ำ 1,000 คน ยังไม่เหมาะสมเพียงพอที่จะทำให้การประมาณค่าพารามิเตอร์ของข้อสอบมีความแม่นยำ ซึ่งไม่สอดคล้องกับผลการศึกษาของ Patsula & Gessaroli (1995), Swaminathan & Gifford (1979) และ Yen (1987) ที่พบว่าขนาดตัวอย่างขั้นต่ำ 1,000 คน เหมาะสมกับแบบสอบ 20 ข้อ และ Şahin & Anıl (2017) ที่พบว่าขนาดตัวอย่างต่ำสุดที่เหมาะสมของแบบสอบ 20 ข้อ คือ 750 คน

ทั้งนี้เพราะ Patsula & Gessaroli (1995) ใช้การจำลองข้อมูลและข้อสอบหลายข้อมีค่าพารามิเตอร์โอกาสการเดา (c) เป็นศูนย์ ส่วน Swaminathan & Gifford (1979) ใช้การจำลองข้อมูลเช่นกันและไม่มีเกณฑ์ในการพิจารณาความแม่นยำของการประมาณค่าพารามิเตอร์ แต่ใช้การเปรียบเทียบจากวิธีการประมาณค่า 2 วิธี (urry และ maximum likelihood) ส่วน Yen (1987) ก็ใช้การจำลองข้อมูลและเปรียบเทียบจาก 2 โปรแกรม คือ BILOG v2.2 และ LOGIST 5.0 v2.5 ในขณะที่ Şahin & Anıl (2017) ใช้ข้อมูลจริงจากแบบสอบ 50 ข้อ และใช้การวิเคราะห์องค์ประกอบเชิงสำรวจแล้วเลือกข้อสอบที่มีน้ำหนักองค์ประกอบในปัจจัยที่ 1 สูงที่สุด 20 ข้อแรกเพื่อเป็นแบบสอบ 20 ข้อ ซึ่งทำให้ละเลยโครงสร้างที่แท้จริงของแบบสอบ ดังนั้นในการวิจัยครั้งนี้ซึ่งใช้ข้อมูลจริงจะพบว่า แบบสอบ 20 ข้อ ควรใช้ขนาดตัวอย่างมากกว่า 1,000 คน เพื่อให้การประมาณค่าพารามิเตอร์มีความแม่นยำ

ในแบบสอบ 40 ข้อ พบว่าขนาดตัวอย่างต่ำสุดที่เหมาะสม คือ 700 คน ซึ่งไม่สอดคล้องกับผลการศึกษาของ Patsula & Gessaroli (1995) และ Yen (1987) ที่พบว่าขนาดตัวอย่าง 1,000 คน เหมาะกับแบบสอบ 40 ข้อ ทั้งนี้เพราะ Patsula & Gessaroli (1995) ศึกษาจากการจำลองข้อมูลขนาดโดยตัวอย่างที่ใกล้เคียง คือ 500 และ 1,000 คน ส่วน Yen (1987) ใช้การจำลองข้อมูลและศึกษาเฉพาะขนาดตัวอย่าง 1,000 คน เท่านั้น

ในแบบสอบ 50 ข้อ พบว่าขนาดตัวอย่างขั้นต่ำ 1,000 คน ยังไม่เหมาะสมเพียงพอที่จะทำให้การประมาณค่าพารามิเตอร์ของข้อสอบมีความแม่นยำ ซึ่งไม่สอดคล้องกับข้อเสนอแนะของ Lord ในปี 1968 ที่แนะนำว่าแบบสอบ 50 ข้อ เหมาะสมกับขนาดตัวอย่าง 1,000 คน (Chuah et al., 2006; Drasgow, 1989) แต่ในขณะที่ Chuah et al. (2006) กลับพบว่าแบบสอบ 50 ข้อ เหมาะสมกับขนาดตัวอย่าง 300 คน ทั้งนี้อาจเป็นเพราะร้อยละของความแปรปรวนขององค์ประกอบที่ 1 ของแบบสอบ 50 ข้อ ที่ใช้ในการวิจัยมีเพียงร้อยละ 9.58 ซึ่งฝ่าฝืนข้อตกลงเบื้องต้นด้านความเป็นเอกมิติค่อนข้างมาก โดย Reckase (1979) กล่าวว่าขนาดขององค์ประกอบที่ 1 มีความสำคัญกับการประมาณค่าพารามิเตอร์มาก โดยเฉพาะหากร้อยละของความแปรปรวนน้อยกว่าร้อยละ 10 จะทำให้การประมาณค่าพารามิเตอร์ของข้อสอบไม่คงที่ (unstable) ดังนั้นในกรณีแบบสอบที่มีความยาว 50 ข้อ ที่มีความแปรปรวนขององค์ประกอบแรกน้อยกว่าร้อยละ 10 จึงควรใช้ขนาดตัวอย่างมากกว่า 1,000 คน

ข้อเสนอแนะ

ข้อเสนอแนะในการนำผลการวิจัยไปใช้

ผู้สนใจสามารถนำผลการวิจัยไปใช้ประโยชน์ได้ เนื่องจากการวิจัยครั้งนี้ใช้ข้อมูลจริงทำให้ผลการวิจัยที่ได้สอดคล้องกับความเป็นจริง ขนาดตัวอย่างต่ำสุดจากผลการวิจัยที่ได้จึงเป็นขนาดที่เหมาะสมซึ่งจะทำให้การประมาณค่าพารามิเตอร์ของข้อสอบมีความแม่นยำ

ข้อเสนอแนะในการวิจัยครั้งต่อไป

1. ผู้ที่สนใจสามารถนำไปทำการวิจัยครั้งต่อไปด้วยการเพิ่มเงื่อนไขของขนาดตัวอย่างที่มากขึ้น เช่น 2,000, 2,500 หรือ 3,000 เป็นต้น เพื่อตรวจสอบความแม่นยำของการประมาณค่าพารามิเตอร์ของข้อสอบในกรณีการใช้ข้อมูลจริงที่ไม่ผ่านข้อตกลงเบื้องต้นของ IRT

2. ผู้ที่สนใจสามารถนำไปทำการวิจัยครั้งต่อไปด้วยการเปลี่ยนประชากรที่ศึกษา เนื่องจากนักเรียนที่เข้าสอบ O-NET มาจากหลายสังกัด เช่น สำนักงานคณะกรรมการส่งเสริมการศึกษาเอกชน (สช.) กรมส่งเสริมการปกครองท้องถิ่น (สถ.) กระทรวงการอุดมศึกษา วิทยาศาสตร์ วิจัยและนวัตกรรม (อว.) สำนักการศึกษา กรุงเทพมหานคร (กทม.) หรือ สำนักงานพระพุทธศาสนาแห่งชาติ (พศ.) เป็นต้น ซึ่งสามารถนำผลการวิจัยที่ได้มาเปรียบเทียบกันเพื่อให้ได้ผลการวิจัยที่ถูกต้องชัดเจนมากยิ่งขึ้น

3. ผู้ที่สนใจสามารถนำไปทำการวิจัยครั้งต่อไปด้วยการใช้ข้อมูลจริงจากผลการสอบ O-NET ของนักเรียนชั้นประถมศึกษาปีที่ 6 หรือชั้นมัธยมศึกษาปีที่ 6 ซึ่งจะทำให้มีความยาวแบบสอบที่หลากหลายจากข้อมูลจริง ทำให้ได้ผลการวิจัยที่หลากหลายและมีประโยชน์มากยิ่งขึ้น

รายการอ้างอิง

ณัฏฐภรณ์ หลาวทอง, สิวะโชติ ศรีสุทธิยากร, และ ฟราย เจอราร์ด วอลตัน. (2563). *การพัฒนาดัชนีวัดความเสมอภาคและความเท่าเทียมทางการศึกษาของประเทศไทย*. สำนักงานคณะกรรมการส่งเสริมวิทยาศาสตร์ วิจัยและนวัตกรรม.

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

Akour, M., & Al-Omari, H. (2013). Empirical investigation of the stability of IRT item-parameters estimation. *International Online Journal of Educational Sciences*, *5*(2), 291-301.

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation. https://eric.ed.gov/?id=ED458219

Chen, S. L., Chen, J. H., & Lee, Y. (2018). A comparison of competing models for understanding industrial organization's acceptance of cloud services. *Sustainability*, *10*(3), 1-20. http://dx.doi.org/10.3390/su10030673

Chuah, S. C., Drasgow F., & Luecht, R. (2006). How big is big enough? Sample size requirements for cast item parameter estimation. *Applied Measurement in Education, 19*(3), 241-255. http://dx.doi.org/10.1207/s15324818ame1903_5

DeMars, C. E. (2008, March). *Scoring multiple choice items: A comparison of IRT and classical polytomous and dichotomous methods* [Paper presentation]. Annual Meeting of the National Council on Measurement in Education, New York.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement, 13*(1), 77-90. https://doi.org/10.1177/014662168901300108

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. SAGE.

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, *6*(1), 53-60. https://doi.org/10.21427/D7CF7R

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two and three-parameter logistic item characteristic curves: A monte carlo study. *Applied Psychological Measurement*, *6*(3), 249-260. http://dx.doi.org/10.1177/014662168200600301

John, M. T., Reissmann, D. R., Feuerstahler, L., Waller, N., Baba, K., Larsson, P., Celebic, A., Szabo, G., & Rener-Sitar, K. (2014). Exploratory factor analysis of the oral health impact profile. *Journal of Oral Rehabilitation*, *41*(9), 635-643. https://doi.org/10.1111/joor.12192

Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement, 31*(4), 331-358. https://doi.org/10.1177/0146621606292213

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. https://doi.org/10.1016/j.sbspro.2012.05.082

Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28*(4), 989-1020. http://dx.doi.org/10.1177/001316446802800401

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement*, *18*(4), 315-328. https://doi.org/10.1177/014662169401800403

Patsula, L. N., & Gessaroli M. E. (1995, June). *A comparison of item parameter estimates and ICCs produced with TESTGRAF and BILOG under different test lengths and sample sizes* [Paper presentation]. Annual Meeting of the National Council on Measurement in Education, San Francisco.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*(3), 207-230. https://doi.org/10.2307/1164671

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., Thissen, D., Revicki, D. A., Weiss, D. J., Hambleton, R. K., Liu, H., Gershon, R., Reise, S. P., Lai, J. S., Cella, D., & PROMIS Cooperative Group (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, *45*(5), S22-S31. https://doi.org/10.1097/01.mlr.0000250483.85507.04

Şahin, A., & Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory and Practice*, *17*, 321-335. http://dx.doi.org/10.12738/estp.2017.1.0270

Sireci, S. G. (1991, June). *Sample-independent item parameters? An investigation of the stability of IRT item parameters estimated from small data sets* [Paper presentation]. Annual Conference of Northeastern Educational Research Association, New York.

Stone, M., & Yumoto, F. (2004). The effect of sample size for estimating Rasch/IRT parameters with dichotomous items. *Journal of Applied Measurement*, *5*(1), 48-61.

Swaminathan, H., & Gifford, J. A. (1979). *Estimation of parameters in the three-parameter latent trait model*. School of Education, University of Massachusetts.

Swaminathan, H., Hambleton, R. K., Sireci, S. G., Xing, D., & Rizavi, S. M. (2003). Small sample estimation in dichotomous item response models: Effect of priors based on judgmental information on the accuracy of item parameter estimates. *Applied Psychological Measurement, 27*(1), 27-51. http://dx.doi.org/10.1177/0146621602239475

Tang, K. L., Way, W. D., & Carey, P. A. (1993). *The effect of small calibration sample sizes on TEOFL IRT-based equating.* Educational Testing Service.

Warm, T. A. (1978). *A primer of item response theory*. Coast Guard Institute. https://apps.dtic.mil/sti/pdfs/ADA063072.pdf

Weiss, D. J., & Minden, S. V. (2012). *A comparison of item parameter estimates from Xcalibre 4.1 and Bilog-MG.* Assessment Systems Corporation.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika, 52*(2), 275-291. https://doi.org/10.1007/BF02294241

Translated Thai References

Kanjanawasee, S. (2012). *Modern test theories* (4th ed.). Chulalongkorn University Press.

Lawthong, N., Srisutiyakorn, S., & Fry, G. W. (2020). *Development of index of measures educational equity and equality in Thailand*. Thailand Science Research and Innovation.