บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

การทดสอบเป็นการดำเนินการที่ตั้งอยู่บนพื้นฐานการวัดคุณลักษณะแฝงภายในตัวบุคคล (Traits) โดยใช้ข้อสอบเป็นสิ่งเร้าให้ผู้ทดสอบแสดงความสามารถออกมาตอบสนอง หากมีข้อมูลที่สามารถยืนยันได้ว่าข้อสอบที่สร้างขึ้นมีคุณสมบัติวัดได้ตรงตามสิ่งที่ต้องการวัด (Validity) และผลการวัดมีความคงเส้นคงวา (Reliability) ก็ย่อมมั่นใจได้ระดับหนึ่งว่าข้อสอบที่สร้างขึ้นมีคุณภาพเพียงใดนั้น ผู้พัฒนาข้อสอบต้องมีความรู้ถึงแก่นแท้ของเนื้อหาวิชาที่จะวัด ประกอบกับความสามารถทักษะการเขียนข้อสอบ และต้องวางแผนการสร้างข้อสอบอย่างรอบคอบ ครอบคลุมเนื้อหาที่ต้องการวัด รวมทั้งมีการตรวจสอบคุณภาพของข้อสอบ ต้องนำข้อสอบที่สร้างขึ้นมาไปทดลองสอบกับกลุ่มตัวอย่าง แล้วนำผลการตอบของผู้สอบมาวิเคราะห์หาคุณภาพของข้อสอบเป็นรายข้อ ผลการวิเคราะห์คุณภาพข้อสอบเป็นรายข้อนี้จะทำให้ทราบว่าข้อสอบแต่ละข้อสามารถทำหน้าที่ได้ตรงตามที่ผู้พัฒนาข้อสอบต้องการหรือไม่ เพื่อเป็นข้อมูลพื้นฐานสำหรับการจัดทำเป็นแบบสอบที่เหมาะสมต่อไป (ศิริชัย กาญจนวาสี, 2548ข; Murphy, Charles, & Davidshofer, 2001)

ปัจจุบันการใช้แบบสอบแบบเลือกตอบ ยังคงใช้ประเมินความสามารถของผู้เรียนอย่างแพร่หลายทั้งผลการประเมินการเรียนรู้ในสถานศึกษาระดับชาติหรือแบบสอบคัดเลือกเข้าศึกษาต่อในระดับอุดมศึกษาเนื่องจากแบบสอบเลือกตอบหรือหลายตัวเลือกมีข้อดีหลายประการด้วยกัน คือ ประการที่1 เป็นแบบสอบที่เหมาะสำหรับการวัดความรู้ความสามารถตั้งแต่ขั้นต่ำไปจนถึงขั้นสูง ประการที่ 2 ใช้เวลาในการตรวจค่อนข้างน้อย เหมาะสำหรับผู้สอบจำนวนมาก ประการที่3 มีความตรงตามเนื้อหาและความเที่ยงค่อนข้างสูง ประการที่ 4 เหมาะสำหรับการพัฒนาเป็นแบบสอบมาตรฐาน ประการที่ 5 ให้สารสนเทศด้านการวินิจฉัยการเรียนรู้ของผู้เรียนได้ (ศิริชัย กาญจนวาสี, 2548ข; Mehrens & Lehnman, 1984) นอกจากนี้ในการทดสอบแห่งชาติ (National Test) ของกระทรวงศึกษาธิการ ยังได้ใช้เป็นแบบวัดประเมินผลสัมฤทธิ์ทางการเรียนในระดับช่วงชั้นของสำนักงานการศึกษาขั้นพื้นฐาน ซึ่งประกอบด้วยการสอบระดับชาติ (National Test) หรือ NT จัดสอบโดยสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน กระทรวงศึกษาธิการ (สพฐ.) เป็นการสอบเพื่อประเมินคุณภาพการศึกษาขั้นพื้นฐานของนักเรียนแต่ละโรงเรียน เพื่อนำข้อมูลมาเป็นแผนพัฒนานักเรียนให้สามารถอ่านออกเขียนได้

รู้จักคิดวิเคราะห์ โดยจะทดสอบตามมาตรฐานการเรียนรู้ของหลักสูตรการศึกษาขั้นพื้นฐาน ทั้ง 8 กลุ่มสาระเป็นการสอบความรู้สำหรับนักเรียนชั้นประถมศึกษาปีที่ 3 และใช้ข้อสอบมาตรฐานเดียวกันทั้งประเทศ

การพัฒนาวิธีการตอบและการให้คะแนนแบบทดสอบเลือกตอบจำเป็นต้องคำนึงถึงคุณภาพของแบบสอบ ซึ่งเป็นหัวใจของการวัดและประเมินผล คุณภาพที่สำคัญที่สุดของแบบทดสอบ คือ ความเที่ยงตรง (Validity) เพราะการนำแบบทดสอบที่ขาดความเที่ยงตรงมาใช้ในการวัดเป็นการไม่ยุติธรรมสำหรับผู้สอบ ผลที่ได้จากการวัดไม่สามารถสะท้อนถึงความสามารถที่แท้จริงของผู้สอบ การตรวจสอบความเที่ยงตรงของแบบทดสอบสามารถทำได้หลายวิธี วิธีการหนึ่งคือการตรวจการทำหน้าที่ต่างกันของข้อสอบ ข้อสอบที่ไม่ได้วัดเฉพาะคุณลักษณะแฝง (Latent Trait) ที่ต้องการวัด แต่วัดคุณลักษณะแฝงอื่นของผู้สอบที่ไม่ต้องการวัด แสดงว่า ข้อสอบข้อนั้นขาดความเที่ยงตรง (Shealy & Stout, 1993, p. 197) ถ้าผู้สอบกลุ่มย่อยใดมีคุณลักษณะแฝงอื่นสูงกว่าย่อมมีโอกาสที่จะตอบข้อสอบได้ถูกต้องมากกว่าทั้ง ๆ ที่มีคุณลักษณะแฝงที่ต้องการวัดเท่ากันกับกลุ่มผู้สอบย่อยกลุ่มอื่น จึงทำให้เกิดการได้เปรียบเสียเปรียบกันระหว่างกลุ่มผู้สอบย่อย ลักษณะนี้เดิมใช้คำว่า ความลำเอียงของข้อสอบ (Item Bias) ต่อมาระยะหลังเกิดความคลุมเครือในการที่ใช้เกณฑ์ในการตัดสินใจเรื่องความลำเอียง จึงนิยมใช้สารสนเทศทางสถิติมาเป็นเกณฑ์ในการตัดสินใจ และใช้คำว่าการทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) เนื่องจากเห็นว่าเป็นคำที่มีความหมายกลาง ๆ มีความเหมาะสมในเชิงวิชาการมากกว่าคำว่า ความลำเอียง (Holland & Wainer, 1993, pp. 4 – 5)

การทำหน้าที่ต่างกันของข้อสอบนั้น ขนาดและทิศทางของการทำหน้าที่ต่างกันของข้อสอบจะแปรเปลี่ยนไปตามระดับความสามารถที่แตกต่างกันของผู้สอบ โดยแบ่งลักษณะของข้อสอบที่ทำหน้าที่ต่างกันเป็น 2 ประเภท (Mellenbergh, 1982) ได้แก่ ข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอ (Uniform DIF) ซึ่งจะเกิดขึ้นเมื่อไม่มีปฏิสัมพันธ์ (Interaction) ระหว่างระดับความสามารถของผู้สอบกับการเป็นสมาชิกของกลุ่มย่อยนั่นคือ โอกาสของการตอบข้อสอบได้ถูกต้องของผู้สอบกลุ่มย่อยกลุ่มหนึ่ง สูงกว่าผู้สอบกลุ่มย่อยอีกกลุ่มหนึ่งตลอดช่วงความสามารถ และข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ (Nonuniform DIF) ซึ่งจะเกิดขึ้นเมื่อมีปฏิสัมพันธ์ระหว่างระดับความสามารถของผู้สอบกับการเป็นสมาชิกของกลุ่มย่อย นั่นคือ โอกาสของการตอบข้อสอบได้ถูกต้องของผู้สอบกลุ่มย่อยกลุ่มหนึ่งสูงกว่าผู้สอบกลุ่มย่อยอีกกลุ่มหนึ่งไม่ตลอดช่วงความสามารถ ซึ่งตามทฤษฎีการตอบสนองข้อสอบ (IRT) สามารถพิจารณาปฏิสัมพันธ์

ได้จากความแตกต่างของค่าพารามิเตอร์อำนาจจำแนกของข้อสอบระหว่างผู้สอบกลุ่มย่อย 2 กลุ่ม ถ้าข้อสอบระหว่างผู้สอบ 2 กลุ่มย่อยมีค่าพารามิเตอร์อำนาจจำแนกของข้อสอบระหว่างผู้สอบกลุ่มย่อย 2 กลุ่ม ถ้าข้อสอบระหว่างผู้สอบ 2 กลุ่มย่อย มีค่าพารามิเตอร์อำนาจจำแนกเท่ากันแล้ว โค้งคุณลักษณะข้อสอบ (Item Characteristic Curves: ICC) ของผู้สอบ 2 กลุ่มจะขนานกัน แสดงว่าข้อสอบที่ทำหน้าที่ต่างกันแบบสม่ำเสมอ แต่ถ้าข้อสอบระหว่างผู้สอบ 2 กลุ่มย่อย มีค่าพารามิเตอร์อำนาจจำแนกไม่เท่ากันแล้ว โค้งคุณลักษณะข้อสอบ (Item Characteristic Curves: ICC) ของผู้สอบ 2 กลุ่มจะไม่ขนานกัน แสดงว่าข้อสอบที่ทำหน้าที่ต่างกันแบบไม่สม่ำเสมอ

วิธีการในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนแบบสองค่ามีหลายวิธีจำแนกเป็น 2 กลุ่มวิธีใหญ่ ๆ คือ กลุ่มวิธี IRT และกลุ่มวิธีที่ไม่ใช่ IRTกลุ่มวิธี IRT เป็นกลุ่มที่วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบใช้คะแนนที่สังเกตไม่ได้ (Unobserved Score) หรือใช้ตัวแปรแฝง (Latent Variable) วิเคราะห์ตามทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) และใช้การประมาณค่าความสามารถของผู้สอบเป็นเกณฑ์การจับคู่กลุ่มผู้สอบ วิธีการตรวจในกลุ่มนี้แบ่งเป็น 3 วิธีใหญ่ ๆ คือ วิธีการวัดพื้นที่ วิธีการเปรียบเทียบค่าพารามิเตอร์ และวิธีซิบเทสท์ (SIBTEST) (Shealy & Stout, 1993) วิธีการวัดพื้นที่ ซึ่งแบ่งออกเป็นวิธีย่อย ๆ อีกหลายวิธี เช่น วิธีการวัดพื้นที่ของรัดเนอร์ (Rudner, 1977) วิธีการวัดพื้นที่ของลินน์, เลวิน, ฮาส์ติงส์ และวาร์ดรัพ (Linn, Levine, Hastings, & Wardrup, 1981) วิธีการวัดพื้นที่ของเชพพาร์ด, คามิลไล และวิลเลียมส์ (Shepard, Camilli, & Williams, 1984) วิธีการวัดพื้นที่ของราชู (Raju, 1990) และวิธีการวัดพื้นที่ของคิมและโคเฮน (Kim & Cohen, 1991) เป็นต้น ส่วนวิธีการเปรียบเทียบค่าพารามิเตอร์ แบ่งออกเป็นวิธีย่อย ๆ เช่น วิธีการทดสอบไค – สแควร์ของลอร์ด (Lord, 1986) และวิธีการทดสอบอัตราส่วนความน่าจะเป็น (Likelihood Ratio Test: LR) (Thissen, Steinberge, & Wainer, 1993) เป็นต้น ส่วนกลุ่มวิธีที่ไม่ใช่ IRT เป็นกลุ่มที่วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบใช้คะแนนที่สังเกตได้ (Observed Score) วิเคราะห์ตามทฤษฎีการทดสอบมาตรฐานเดิม (Classical Test Theory: CTT) และใช้คะแนนรวมของผู้สอบเป็นเกณฑ์การจับคู่กลุ่มผู้สอบซึ่งเป็นเกณฑ์ภายใน วิธีการที่สำคัญในกลุ่มนี้ ได้แก่ วิธีแปลงค่าความยากของข้อสอบ(Transformed Item Difficulty: TID) (Angoff & Sharon, 1972) วิธีการวิเคราะห์ความแปรปรวน (Analysis of Variance: ANOVA) (Cleary & Hillton, 1968) วิธีการทำให้เป็นมาตรฐาน (Standardization: STND) (Dorans & Kulick, 1986) วิธีลอก – ลิเนียร์ (Log – Linear: LL) (Mellenbergh, 1982) วิธีแมนเทล – แฮนส์เซล (Mantal – haenszel: MH) (Holland & Thayer, 1988) และวิธีการถดถอยโลจิสติก (Logistic Regression: LR) (Swaminathan & Rogers, 1990) เป็นต้น

สำหรับการวิเคราะห์คุณภาพของข้อสอบตามทฤษฎี IRT นั้น โมเดลซับซ้อนน้อยที่สุด คือ 1 – parameter Logistic Measurement Model (1PL) ผลลัพธ์ที่ได้จากการประมาณค่าพารามิเตอร์ คือ ค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) และค่าพารามิเตอร์ของข้อสอบ (β_i) ซึ่งก็มีนักวัดผลตั้งคำถามต่อไปว่า ค่าพารามิเตอร์ความสามารถของผู้สอบมีความผันแปรระหว่างกลุ่มผู้สอบหรือไม่และความผันแปรที่เกิดขึ้นมีสาเหตุมาจากตัวแปรอิสระใดบ้าง นักวัดผลจึงนำผลการตอบข้อสอบของผู้สอบมาประมาณค่าความสามารถของผู้สอบแต่ละคน ตามทฤษฎี IRT จากนั้นนำค่าพารามิเตอร์ความสามารถของผู้สอบที่ประมาณค่าได้ไปเป็นตัวแปรตาม โดยมีตัวแปรคุณลักษณะของผู้สอบ (Person Characteristics) เป็นตัวแปรทำนายด้วยวิเคราะห์ถดถอยพหุ (Multiple Regression) เพื่อศึกษาว่าตัวแปรอิสระใดบ้างที่มีอิทธิพลต่อค่าพารามิเตอร์ความสามารถของผู้สอบ และสามารถอธิบายความผันแปรที่เกิดขึ้นได้อย่างไรบ้าง ซึ่งนักวัดผลเรียกกระบวนการวิเคราะห์นี้ว่า การวิเคราะห์แบบ 2 ขั้นตอนซึ่งวิธีการนี้ก็ยังมีปัญหาในด้านของความเหมาะสมของการวิเคราะห์ข้อมูล Maier (2001) ได้อธิบายว่า ผลการวิเคราะห์แบบสองขั้นตอนจะทำให้เกิดผลการประมาณค่าของความแปรปรวนของความคลาดเคลื่อนแบบสุ่มในระดับที่ 1 สูงกว่าความเป็นจริง (Overestimate) ส่วนค่าของความแปรปรวนของความคลาดเคลื่อนแบบสุ่มในระดับที่ 2 ต่ำกว่าความเป็นจริง (Underestimate) ซึ่งทำให้ผลการวิเคราะห์ขาดความแม่นยำ ส่วน Hambleton & Swaminathan (1985) ได้อธิบายว่า ค่าความสามารถของผู้สอบ (θ) ที่ได้จากการประมาณค่าตามทฤษฎีการตอบสนองข้อสอบจะมีขนาดของค่าความคลาดเคลื่อนมาตรฐาน (Standard Error) ที่แตกต่างกันในแต่ละระดับความสามารถ ดังนั้นการนำค่าคะแนนความสามารถของผู้สอบมาเป็นตัวแปรตามจะเกิดความคลาดเคลื่อนจากการวัดไม่คงที่ (Heteroscedusticity Measurement Errors) รวมทั้งค่าประมาณตัวแปรตามที่ได้รับจากการใช้วิธี Modified Maximum Likelihood Estimation (MMLE) ก็จะเกิดความลำเอียง (Biased) และความไม่คงที่ของการประมาณค่า หากทำการวิเคราะห์แบบ 2 ขั้นตอนด้วยการวิเคราะห์ถดถอยพหุคูณ จะเกิดปัญหาในการวิเคราะห์ และมีแนวทางการแก้ปัญหาที่น่าสนใจ คือ การวิเคราะห์แบบขั้นตอนเดียว (One – step Analysis)

การวิเคราะห์แบบขั้นตอนเดียว โดยการใช้โมเดลทางสถิติที่สามารถวิเคราะห์แบบรวมตัวแปรคุณลักษณะของผู้สอบ ให้ทำหน้าที่เป็นตัวแปรทำนายความสามารถของผู้สอบร่วมกับการประมาณค่าพารามิเตอร์ข้อสอบ และค่าความสามารถของผู้สอบได้ไปพร้อม ๆ กัน

ตามโมเดลการตอบสนองข้อสอบ (Zwinderman, 1991; Mellenbegh, 1982; Rijman et al., 2005) เพราะการประมาณค่าอิทธิพลของคุณลักษณะของผู้สอบที่ถูกประมาณค่าพร้อม ๆ กับค่าพารามิเตอร์ของข้อสอบและค่าพารามิเตอร์ความสามารถของผู้สอบ ทำให้ไม่มีปัญหาที่เกิดขึ้นเหมือนกับการวิเคราะห์แบบสองขั้นตอน (Two – step Analysis)นักการศึกษาหลายท่านได้พยายามพัฒนาเทคนิคทางสถิติสำหรับการวิเคราะห์แบบขั้นตอนเดียวขึ้นมาในระยะแรก โดยเฉพาะ Fischer ที่ได้พัฒนาสถิติวิเคราะห์ในลักษณะนี้ (Embretson & Reise, 2000) ซึ่งโมเดลการวิเคราะห์ดังกล่าว เป็นที่แพร่หลายในกลุ่มนักวัดผลอย่างรวดเร็ว และได้รับการพัฒนาต่อมาโดย Linacre (1989) ที่เสนอโมเดล Many – facet Rasch Model ซึ่งเป็นโมเดลการวิเคราะห์ที่สามารถเพิ่มตัวแปรบ่งชี้ไปในสมการรวมกันเชิงเส้นตรง (Linear Combination) ให้ทำหน้าที่ตรวจสอบคุณลักษณะของผู้ตรวจหรือผู้ประเมินที่มีอิทธิพลต่อการให้คะแนน วิธีการนี้จึงสามารถตรวจสอบระดับความเข้มงวดของการให้คะแนนจากผู้ประเมินที่แตกต่างกันได้ (อิทธิฤทธิ์ พงษ์ปิยะรัตน์, 2551) Adam, Wilson, & Wu (1997) ได้พัฒนาเทคนิควิธีการวิเคราะห์ที่สามารถใช้ตัวแปรคุณลักษณะของผู้สอบ เป็นตัวแปรทำนายความสามารถของผู้สอบจากการประมาณค่าตามแนวทางการวิเคราะห์ของโมเดลราสช์ (Rasch Model) โดยเสนอโมเดลการวิเคราะห์ที่เรียกว่า Random Coefficient Multinomial Logit Model (RCMLM) และต่อมา Adam, Wilson, & Wu (1997) ก็ได้ขยายแนวคิดของโมเดลการวิเคราะห์แบบ RCMLM ต่อให้สามารถวิเคราะห์แบบทดสอบในลักษณะพหุมิติ (Multidimensional) ได้ด้วยโมเดลการวิเคราะห์แบบ Multidimensional Random Coefficient Multinomial Logit Model (MRCMLM) ซึ่ง Adam, Wilson, & Wu กล่าวว่าวิธีการประมาณค่าด้วยโมเดล MRCMLM สามารถประยุกต์ใช้ในการวิเคราะห์ข้อมูลหลายลักษณะ เช่น โมเดลราสช์แบบโลจิสติกอย่างง่าย การวิเคราะห์ที่ข้อมูลมีการกระจายแบบเอ็กโปแนนเชี่ยล และที่สำคัญคือ สามารถวิเคราะห์ข้อมูลลักษณะที่สอดแทรกเป็นพหุระดับ (Multilevel Data) ซึ่งการวิเคราะห์พหุระดับเป็นเทคนิคทางสถิติที่ใช้ในการวิเคราะห์อิทธิพลของตัวแปรทำนายหลายระดับที่มีต่อตัวแปรตาม ซึ่งคุณลักษณะที่สำคัญของตัวแปรทำนายจะต้องมีโครงสร้างเป็นข้อมูลระดับลดหลั่น (Hierarchical Data Structure) อย่างน้อย 2 ระดับ โดยตัวแปรทำนายและตัวแปรตามที่อยู่ระดับล่างต่างมีความสัมพันธ์ซึ่งกันและกัน ได้รับอิทธิพลร่วมกันจากตัวแปรทำนายที่อยู่ระดับบน (ศิริชัย กาญจนวาสี, 2548ก; Cronbach, 1976; Burstein, 1978; Goldstein, 1997; Aitkin & Longford, 1986; Raudenbush & Bryk, 1986)

จากการศึกษาโมเดลเชิงเส้นตรงทั่วไประดับลดหลั่น (Hierarchical Generalized Linear Model: HGLM) ค่าอิทธิพลของตัวแปรภายนอกต่อโอกาสในการตอบข้อสอบในการวิเคราะห์ ระดับที่ 2 (ระดับผู้สอบ) สามารถดำเนินการวิเคราะห์ได้จากโปรแกรม HLM ด้วยโมเดลเชิงเส้นตรงทั่วไประดับลดหลั่น (Hierarchical Generalized Linear Model: HGLM) ได้ และทำการวิเคราะห์การประมาณค่าพารามิเตอร์ความยากของข้อสอบ (δ) ค่าพารามิเตอร์ความสามารถของผู้สอบ (θ) จากโปรแกรม HLM ซึ่งมีลักษณะเป็นพารามิเตอร์แบบสุ่ม (Random Parameter) การดำเนินการวิเคราะห์ สามารถดำเนินการวิเคราะห์ในขั้นตอนเดียวตามโมเดล HGLM ด้วยโปรแกรมโมเดลเชิงเส้นตรงระดับลดหลั่น (HLM) ที่ผ่านมาส่วนใหญ่นักวิจัยได้ดำเนินการวิเคราะห์ในลักษณะแยกส่วน ซึ่งในการวิจัยครั้งนี้ผลการวิเคราะห์ข้อสอบ นอกจากจะให้ค่าพารามิเตอร์ข้อสอบค่าพารามิเตอร์ผู้สอบแล้ว ยังจะทราบต่อไปได้ว่าตัวแปรคุณลักษณะของผู้สอบตัวแปรใด สามารถอธิบายความแปรปรวนในค่าความสามารถของผู้สอบได้ และจะนำไปสู่การศึกษาในรายละเอียดเชิงลึกของการพัฒนาการทดสอบ โดยประโยชน์จากสารสนเทศที่ได้จากกระบวนการวิเคราะห์ที่น่าเชื่อถือ เพื่อการวางแผนกำหนดนโยบายในการพัฒนาคุณภาพการศึกษาให้เกิดประสิทธิภาพต่อไป (อิทธิฤทธิ์ พงษ์ปิยะรัตน์, 2551) ซึ่งโปรแกรม HLM เป็นโปรแกรมที่มีประโยชน์ต่อการพัฒนาการทดสอบแล้ว โปรแกรม Mplus ก็มีความสำคัญต่อการพัฒนาการทดสอบเช่นเดียวกัน

Muthén & Muthén (2007, 2010) พัฒนาโปรแกรม Mplus เพื่อให้นักวิจัยมีเครื่องมือสำหรับวิเคราะห์ข้อมูลด้วยสถิติวิเคราะห์ขั้นสูง ที่ให้ผลการวิเคราะห์ข้อมูลที่มีความถูกต้องมากกว่าสถิติวิเคราะห์แบบเดิม การพัฒนาโปรแกรมMplus ดำเนินการเป็นกระบวนการที่มีการพัฒนาปรับปรุงโปรแกรมมาจนถึงปัจจุบัน โดยมีการเผยแพร่ โปรแกรม Mplus Version 1 ปี 1998, Version 2 ปี 2001, Version 3 ปี 2004, Version 4 ปี 2006, Version 5 ปี 2007 และ Version 6 ปี 2010 โปรแกรม Mplus Version ใหม่ล่าสุด คือ โปรแกรม Mplus Version 7 โปรแกรม Mplus ได้รับการพัฒนาให้เป็นโปรแกรมที่ใช้งานได้ง่ายและสะดวก และได้รับการปรับปรุงให้ดีขึ้น สามารถวิเคราะห์ข้อมูลได้หลายประเภท

วิธีมิมิค (MIMIC) ก็เป็นวิธีหนึ่งที่ใช้ในโปรแกรม Mplus สำหรับการวิเคราะห์คุณภาพของข้อสอบตามทฤษฎี IRT ซึ่งวิธี MIMIC เป็นโมเดลลิสเรลที่มีตัวแปรแฝงเพียงตัวแปรเดียว โดยที่ตัวแปรแฝงนั้นได้รับอิทธิพลจากตัวแปรภายนอกสังเกตได้หลายตัวแปรและส่งอิทธิพลไปยังตัวแปรภายในสังเกตได้หลายตัวแปรกล่าวอีกอย่างหนึ่งคือเป็นโมเดลลิสเรลของคุณลักษณะแฝงที่มีหลายสาเหตุและวัดได้จากตัวบ่งชี้หลายตัวลักษณะโมเดลจะเห็นว่า

การวัดตัวแปรภายนอกสังเกตได้ต้องมีข้อตกลงข้างต้นว่าไม่มีความคลาดเคลื่อนในการวัด โมเดลมิมิคนี้เป็นประโยชน์มากในการตรวจสอบความเป็นเอกมิติ (Unidimensionality) ในการวิจัยสาขาในการวัดผลการศึกษาสามารถวิเคราะห์ค่าพารามิเตอร์คุณลักษณะข้อสอบ และค่าความสามารถของผู้สอบไม่สามารถสังเกตโดยตรงจึงต้องประมาณจากการตอบข้อสอบ การประมาณค่าพารามิเตอร์ในทฤษฎีการตอบสนองข้อสอบ

วิธี MIMIC มีข้อดีหลายประการของการใช้โมเดล MIMIC ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) (Muthén et al., 1991) วิธีนี้แสดงขนาดของการทำหน้าที่ต่างกันของข้อสอบ (DIF) โดยใช้หลักทฤษฎีการตอบสนองข้อสอบ (IRT) ประมาณค่าการทำหน้าที่ต่างกันของข้อสอบ (DIF) จากค่าพารามิเตอร์ตามทฤษฎีการตอบสนองข้อสอบ (IRT) ซึ่งมีประโยชน์ต่อการวิเคราะห์ข้อมูล จากการศึกษาของ Finch (2005) เปรียบเทียบประสิทธิภาพของโมเดล MIMIC กับการทดสอบโดยวิธีแมนเทลเฮนเซล (Mantel & Haenszel, 1959) และวิธี SIBTEST (Shealy & Stout, 1993) และวิธีการทดสอบ IRT Likelihood Ratio (Thissenet al., 1986) กับความคลาดเคลื่อนประเภทที่ 1 และอำนาจการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ซึ่งได้แสดงให้เห็นว่าวิธี MIMIC มีค่าสูงขึ้นและความคลาดเคลื่อนประเภทที่ 1 มีค่าลดลง เมื่อจำนวนข้อสอบมีจำนวน 50 ข้อ นอกจากนี้วิธี MIMIC ยังสามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบ Uniform DIF ได้เพียงอย่างเดียว

นอกจากนี้ในการนำทฤษฎีการตอบสนองข้อสอบมาใช้ ผู้วิจัยจำเป็นต้องเลือกวิธีที่เหมาะสมแล้ว วิธีประมาณค่าพารามิเตอร์ของข้อสอบ และความสามารถของผู้เข้าสอบก็เป็นอีกกระบวนการหนึ่งที่จำเป็นต้องเลือกใช้ให้เหมาะสมกับสภาพการวัดแต่ละครั้ง สำหรับทฤษฎีการตอบสนองข้อสอบนั้น วิธีการประมาณค่าพารามิเตอร์ของข้อสอบและความสามารถของผู้เข้าสอบมีหนึ่งวิธีที่น่าสนใจ คือ วิธีของเบส์ (Bayesian Estimation) (Swaminathan & Gifford, 1985, pp. 349 – 364)

จากการศึกษางานวิจัยเกี่ยวกับการนำทฤษฎีการตอบสนองข้อสอบไปประยุกต์ใช้ โดยการนำผลคะแนนที่ได้จากแบบสอบนั้นมาประมาณค่าความสามารถเพื่อให้เกิดความถูกต้องยุติธรรมในการประเมินผลการวัดในปัจจุบันได้นำทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) มาใช้ในการประมาณค่าความสามารถ ค่าความเที่ยง ค่าพารามิเตอร์ของข้อสอบ (สุนทร เทียนงาม, 2551) ไม่ว่าจะเพื่อวิเคราะห์หาคุณภาพของข้อสอบ แบบทดสอบหรือวิธีการสอบก็ตาม พบว่า มีการประมาณค่าพารามิเตอร์ของข้อสอบและความสามารถของผู้สอบด้วยวิธีแมกซิมัมไลค์ลิฮูดเป็นส่วนมากทั้ง ๆ ที่วิธีแมกซิมัมไลค์ลิฮูดมีข้อจำกัดที่ไม่สามารถประมาณค่า

ความสามารถของผู้เข้าสอบที่ได้คะแนนเต็มหรือศูนย์ได้ จำเป็นต้องมีการกำจัดผู้สอบเหล่านี้ออกไปจากการประมาณค่า ทำให้เหลือผู้เข้าสอบที่มีระดับความสามารถใกล้เคียงกันดังนั้น หากต้องการทำการเปรียบเทียบความสามารถของนักเรียนเมื่อทำการสอบด้วยแบบทดสอบ 2 ฉบับใด ๆ จะมีแนวโน้มว่าความสามารถของนักเรียนเฉลี่ยแล้วไม่แตกต่างกันส่งผลให้สรุปได้ว่าแบบสอบ 2 ฉบับนั้น มีคุณภาพไม่แตกต่างกัน ซึ่งข้อสรุปดังกล่าวอาจเป็นข้อค้นพบที่คลาดเคลื่อนอันเนื่องมาจากความจำกัดของวิธีประมาณค่าแบบแมกซิมัมไลค์ลิฮูดก็ได้ Swaminathan & Gifford (1985) กล่าวว่าการประมาณค่าพารามิเตอร์ด้วยวิธีของเบส์ ไม่มีข้อจำกัด ดังเช่นวิธีแมกซิมัมไลค์ลิฮูด กล่าวคือ สามารถประมาณค่าพารามิเตอร์ของผู้สอบที่ผู้เข้าสอบทุกคนตอบถูกหรือตอบผิดได้ และประมาณค่าความสามารถของผู้เข้าสอบที่ทำข้อสอบถูกหรือผิดทุกข้อได้ด้วย โดยเฉพาะในกลุ่มตัวอย่างขนาดเล็กประมาณ 200 – 300 คน วิธีของเบส์สามารถประมาณค่าสถิติของข้อสอบได้ใกล้เคียงกับค่าพารามิเตอร์มากกว่าวิธีแมกซิมัมไลค์ลิฮูด

ค่าพารามิเตอร์ของแบบทดสอบที่ประมาณค่าด้วยวิธีแมกซิมัมไลค์ลิฮูด วิธีฮิวริสติก และวิธีของเบส์ พบว่า ค่าความยากเฉลี่ยของแบบทดสอบและค่าส่วนเบี่ยงเบนมาตรฐานของค่าความยากของแบบทดสอบที่ประมาณค่าด้วยวิธีของเบส์ มีค่าสูงกว่าค่าความยากเฉลี่ยของแบบทดสอบและค่าส่วนเบี่ยงเบนมาตรฐานของค่าความยากของแบบทดสอบที่ประมาณค่าด้วยวิธีฮิวริสติกและวิธีแมกซิมัมไลค์ลิฮูด (รัตนา ศรีเหรัญ, 2539) และทฤษฎีการทดสอบแบบดั้งเดิม (Classical Test Theory: CTT) เป็นแนวคิดพื้นฐานสำหรับการพัฒนาแบบทดสอบทางการศึกษาและจิตวิทยา แต่ด้วยข้อจำกัดหลายประการของการวัดตามทฤษฎีการทดสอบแบบดั้งเดิม (CTT) ทำให้มีนักทฤษฎีทางการทดสอบหลายท่าน ต้องการแก้ไขจุดอ่อนของทฤษฎีแบบดั้งเดิม โดยการสร้างการวัดคุณลักษณะภายในของบุคคลแนวใหม่ เริ่มด้วย Thurstone (1927) เป็นผู้เสนอแนวคิดการวัดคุณลักษณะภายในของบุคคล และพัฒนาเทคนิคการวิเคราะห์ตัวประกอบสำหรับศึกษาคุณลักษณะทางจิตวิทยา ซึ่งเป็นการวางรากฐานความคิดที่สำคัญเกี่ยวกับทฤษฎีการทดสอบแนวใหม่ (Modern Test Theory: MTT) Cronbach et.al. (1963, 1972) เสนอแนวคิดเกี่ยวกับโมเดลความเที่ยงทั่วไปของแบบทดสอบภายใต้เงื่อนไขต่าง ๆ ของการทดสอบ Lord & Novice (1968) เสนอหลักการวัดแบบอิงโมเดล (Model – based Measurement) นับเป็นแนวคิดสำคัญที่ปฏิรูประบบความคิดของการวัดสู่ทฤษฎีการทดสอบแนวใหม่ ได้แก่ทฤษฎีการสรุปอ้างอิงความน่าเชื่อถือของผลการวัด (Generalizability Theory) หรือ G – theory และทฤษฎีการตอบสนองข้อสอบ (Item Response Theory) (ศิริชัย กาญจนวาสี, 2550) ทฤษฎีการตอบสนองข้อสอบ (IRT) เป็นทฤษฎีที่เสนอแนวคิดว่า ความน่าจะเป็น

ของการตอบสนองข้อสอบได้ถูกต้อง ขึ้นอยู่กับความสามารถจริงของผู้ตอบ และคุณลักษณะของข้อสอบ อันประกอบด้วย พารามิเตอร์ความยาก อำนาจจำแนกและโอกาสการเดาข้อสอบได้ถูกต้องซึ่งความสัมพันธ์ดังกล่าว สามารถแสดงด้วยโมเดลการตอบสนองข้อสอบ อาจเป็นโมเดล 1 พารามิเตอร์ โมเดล 2 พารามิเตอร์ และโมเดล 3 พารามิเตอร์ ซึ่งถือว่าค่าพารามิเตอร์ของข้อสอบและความสามารถจริงของผู้สอบมีความสัมพันธ์กัน (Andrich, 1978)

จากการศึกษาประสิทธิภาพของวิธีการประมาณของวิธีการสรุปอ้างอิงความน่าเชื่อถือของโมเดลการตอบสนองข้อสอบ (GIRM) 4 รูปแบบ ได้แก่รูปแบบที่ 1 Original GIRM ซึ่ง พัฒนาโดย Briggs & Wilson (2007) รูปแบบที่ 2 AGIRMA, รูปแบบที่ 3 AGIRMB และรูปแบบที่ 4 Numerical Bayesian GIRM พบว่า ความลำเอียงในการประมาณค่ารูปแบบที่ 1 กับ รูปแบบที่ 4 ให้ค่าประสิทธิภาพสูงที่สุด โดยรูปแบบที่ 4 เป็นรูปแบบที่ประมาณค่าพารามิเตอร์ได้เฉพาะลักษณะการแจกแจงเริ่มแรกของผู้สอบและข้อสอบแบบปกติสำหรับความไม่แน่นอนในการประมาณค่า พบว่า รูปแบบที่ 4 ให้ค่าประสิทธิภาพสูงที่สุดสำหรับลักษณะการแจกแจงเริ่มแรกของผู้สอบและข้อสอบแบบปกติส่วนลักษณะการแจกแจกเริ่มแรกของค่าพารามิเตอร์ตัวใดตัวหนึ่งที่ไม่มีลักษณะการแจกแจงเริ่มแรกแบบปกติพบว่า รูปแบบที่ 1 ให้ค่าประสิทธิภาพสูงที่สุดและเมื่อพิจารณาในด้านประสิทธิภาพขององค์ประกอบความแปรปรวนยุคลิด พบว่า รูปแบบที่ 2 ให้ค่าประสิทธิภาพสูงที่สุด (ชนะศึก นิชานนท์, 2553)

Rijiman et al. (2005) ได้วิเคราะห์วิธีการประมาณค่าโมเดลผสมสำหรับโมเดลราสช์เพื่อประเมินผลของการประมาณค่าโมเดลผสมสำหรับโมเดลเส้นตรงทั่วไป และโมเดลผสมไม่เป็นเส้นตรง ซึ่ง Rijman และคณะได้กล่าวว่า โมเดลผสม (Mixed Models) เป็นชุดของเครื่องมือทางสถิติที่มีความเหมาะสมสำหรับการวิเคราะห์ข้อมูลที่มีลักษณะเป็นกลุ่ม ๆ และสอดแทรกอยู่ด้วยกัน ในการศึกษาครั้งนี้ มีการจำลองข้อมูล เพื่อศึกษาลักษณะการประมาณค่าที่ต่างกันของโมเดลราสช์ตามเงื่อนไข ดังนี้กลุ่มคนมี 2 กลุ่ม คือ 100 คน และ 500 คน จำนวนข้อสอบมี 2 กลุ่ม คือ 5 ข้อ และ 25 ข้อ และค่าพารามิเตอร์ของข้อสอบมีค่าระหว่าง – 2 ถึง 2 โดยที่ในการวิเคราะห์แต่ละครั้งจะประกอบด้วยข้อมูลจำนวน 30 ชุดทั้ง 8 เงื่อนไข ซึ่งของการประเมินความเหมาะสมของการประมาณค่าพารามิเตอร์นั้น Rijman และคณะได้สร้างดัชนีสำหรับตรวจสอบและประเมินความสามารถในการประมาณค่า (GOR: Goodness of Recovery) จำนวน 3 ดัชนี ดังนี้ 1) BIAS: Different Between the Average Estimated Parameter 2) RMSD: Root Mean Square Deviation 3) MCSE: Monte Carlo Standard Error โดยดัชนีทั้ง 3 มีความสัมพันธ์กัน คือ $RMSD^2 = MCSE^2 + BIAS^2$

การประมาณค่าพารามิเตอร์จะดำเนินการผ่านวิธีการคำนวณ 4 วิธีการ ดังนี้ 1) วิธีการ Gaussian Quarture โดยใช้โปรแกรม SAS: NLMIXED 2) วิธีการ Sixth – order Laplace โดยใช้โปรแกรม HLM (V.5.04) 3) วิธีการ PQL2 โดยใช้โปรแกรม MLwiN (V.1.10) และ 4) วิธีการ MCMC โดยใช้โปรแกรม WinBUGS (V1.2) ซึ่งโปรแกรม WinBUGS มีความสามารถในวิเคราะห์ด้วยวิธี BAYESIAN โดยมีวิธีการวิเคราะห์ข้อมูลได้หลายแบบ เช่น การเขียนคำสั่งในโปรแกรม WinBUGS กรณีใช้ข้อมูลจริง (Real Data) เพื่อประมาณค่าพารามิเตอร์ของข้อสอบ (δ) พารามิเตอร์ความสามารถของผู้สอบ (θ) และการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) และการใช้ข้อมูลจำลอง (Simulation) จากโปรแกรม R และทำการประมวลผลภายใต้จากการเขียนคำสั่งการประมวลผลด้วยโปรแกรม WinBUGS ด้วย Package R2 WinBUGS

Saengla Chaimongkol et al. (2007) ได้ศึกษาเรื่องตัวแบบอธิบายการทำหน้าที่ต่างกันของข้อสอบโดยใช้ WinBUGS 1.4 เสนอตัวแบบพหุระดับการถดถอยโลจิสติกที่ใช้ในการตรวจสอบสาเหตุของการทำหน้าที่ต่างกันของข้อสอบ (DIF) โดยที่ตัวแบบที่เสนอจะพิจารณาโครงสร้างของข้อมูลที่มีการซ้อนทับกัน 3 ระดับ ที่มีการรวมผลลัพธ์ที่ได้จากการวิเคราะห์การถดถอยโลจิสติก เพื่อที่จะระบุลักษณะตัวแปรของข้อมูลระดับ 3 ที่สามารถใช้อธิบายสาเหตุการผันแปรของ DIF ได้ การศึกษาครั้งนี้จะใช้วิธีการจำลองข้อมูลในการตรวจสอบความถูกต้องและความเหมาะสมของตัวแบบ โดยที่ค่าพารามิเตอร์ต่าง ๆ ในตัวแบบจะถูกประมาณโดยการใช้หลักการของเบส์ที่ใช้โปรแกรม WinBUGS นอกจากนี้ วิธีเบส์เซียน มีความแตกต่างจากวิธีดั้งเดิมในการอนุมานรูปแบบความน่าจะเป็นของโมเดลสำหรับตัวแปรสังเกตและพารามิเตอร์ที่ไม่ทราบค่าเบส์เซียนจะอนุมานโดยการตรวจสอบเงื่อนไขของพารามิเตอร์ที่สอดคล้องกับข้อมูลที่สังเกตได้ จึงสามารถใช้งานได้ง่าย วิธีเบส์เซียนมีความยืดหยุ่นสูง สามารถแก้ปัญหาทั้งง่ายและซับซ้อนได้ดี มีการกำหนดการแจกแจงเริ่มต้นของค่าพารามิเตอร์ (Prior Distribution) ที่ใช้ในการกำหนดช่วงของค่าพารามิเตอร์ที่ต้องการประมาณค่า เป็นประโยชน์อย่างยิ่งสำหรับการวิเคราะห์ข้อมูล นอกจากนั้นโมเดลที่มีความซับซ้อนยิ่งในปัจจุบันการพัฒนาเทคนิคการสุ่มตัวอย่างในการจำลองข้อมูล Markov Chain Monte Carlo (MCMC) มีโปรแกรมสนับสนุนการประมาณค่ามากมาย เช่น โปรแกรม WinBUGS ซึ่งเป็นซอฟแวร์ที่ใช้งานได้ง่าย

จะเห็นได้ว่าการวิเคราะห์ข้อสอบ ด้วยทฤษฎีการตอบสนองข้อสอบสามารถให้ทั้งสารสนเทศที่เป็นค่าพารามิเตอร์ของข้อสอบเป็นรายข้อ (Item Parameter) พารามิเตอร์ของผู้สอบเป็นรายบุคคล (Person Parameter) รวมทั้งความสามารถในตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในการวิจัยนี้จึงมุ่งศึกษาการวิเคราะห์คุณภาพของข้อสอบตามทฤษฎี

การตอบสนองข้อสอบ (IRT) 1 – parameter Logistic Measurement Model (1PL) ทั้งนี้ด้วยข้อจำกัดของวิธีการวิเคราะห์ด้วยวิธี HGLM สามารถวิเคราะห์ข้อสอบได้เพียง 1PL ส่วนวิธี MIMIC สามารถวิเคราะห์ได้ 2PL และวิธี BAYESIAN สามารถวิเคราะห์ได้ 3PL เพื่อให้สามารถเปรียบเทียบผลการประมาณค่าพารามิเตอร์ความยากของข้อสอบได้ ในการวิจัยครั้งนี้ ผู้วิจัยจึงศึกษาเพียง 1PLจากนั้นจึงตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยดำเนินการวิเคราะห์ด้วยวิธี HGLM ประยุกต์ใช้โปรแกรม HLM วิธี MIMIC ประยุกต์ใช้โปรแกรม Mplus และวิธี BAYESIAN ประยุกต์ใช้โปรแกรม WinBUGS ซึ่งโปรแกรมดังกล่าวสามารถวิเคราะห์สถิติขั้นสูงได้ดี และเป็นที่นิยมของนักสถิติและนักวัดผลในขณะนี้โดยศึกษาจากการสอบวัดผลสัมฤทธิ์ทางการเรียนเพื่อประเมินคุณภาพการศึกษาระดับชาติ (NT) ปีการศึกษา2553 ชั้นประถมศึกษาปีที่ 3 ได้แก่ วิชาภาษาไทย คณิตศาสตร์ และวิทยาศาสตร์ เพื่อเป็นแนวทางสำหรับผู้ที่เกี่ยวข้องในการออกข้อสอบระดับชาติ ในการนำไปปรับปรุงและพัฒนาข้อสอบต่อไป

คำถามการวิจัย

1. ผลการประมาณค่าพารามิเตอร์ข้อสอบ (δ_i) และพารามิเตอร์ความสามารถของผู้สอบ (θ_j) ด้วยวิธี HGLM ประยุกต์ใช้โปรแกรม HLM วิธี MIMIC ประยุกต์ใช้โปรแกรม Mplus และวิธี BAYESIAN ประยุกต์ใช้โปรแกรม WinBUGS มีความสอดคล้องกันหรือไม่

2. การทำหน้าที่ต่างกันของข้อสอบ (DIF) จำแนกตามเพศและสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียนด้วยวิธี HGLM ประยุกต์ใช้โปรแกรม HLM วิธี MIMIC ประยุกต์ใช้โปรแกรม Mplus และวิธี BAYESIAN ประยุกต์ใช้โปรแกรม WinBUGS มีความสอดคล้องหรือไม่

3. ลักษณะของข้อสอบที่เกิดการทำหน้าที่ต่างกันของข้อสอบ (DIF) ที่ได้จากการวิเคราะห์การทำหน้าที่ต่างกัน โดยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN มีลักษณะและเนื้อหาของคำหรือข้อความที่ใช้ในการเขียนข้อสอบอย่างไร

วัตถุประสงค์ของการวิจัย

1. เพื่อเปรียบเทียบผลการประมาณค่าพารามิเตอร์ข้อสอบ (δ_i) และพารามิเตอร์ความสามารถของผู้สอบ (θ_j) ระหว่างวิธี HGLM วิธี MIMIC และวิธี BAYESIAN

2. เพื่อเปรียบเทียบผลการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ (DIF) จำแนกตามเพศและสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน ระหว่างวิธี HGLM วิธี MIMIC และวิธี BAYESIAN

3. เพื่อศึกษาลักษณะของข้อสอบที่เกิดการทำหน้าที่ต่างกันของข้อสอบ (DIF) ที่ได้จากการวิเคราะห์การทำหน้าที่ต่างกัน โดยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN ด้วยการวิเคราะห์ลักษณะและเนื้อหาของคำหรือข้อความที่ใช้ในการเขียนข้อสอบ

สมมุติฐานในการวิจัย

วิธีการประมาณค่าการประมาณค่าพารามิเตอร์ข้อสอบ (δ_i) และพารามิเตอร์ความสามารถของผู้สอบ (θ_j) ด้วยวิธี HGLM วิธี MIMIC และวิธี BAYESIAN ทำให้ค่าพารามิเตอร์ที่ได้จากการประมาณค่าทั้ง 3 วิธี แตกต่างกันและมีการทำหน้าที่ต่างกันของข้อสอบ (DIF) ด้วย ดังนั้น ผู้วิจัยจึงได้ตั้งสมมุติฐานในการวิจัยไว้ดังนี้

1. ค่าพารามิเตอร์ข้อสอบ (δ_i) และพารามิเตอร์ความสามารถของผู้สอบ (θ_j) ระหว่างวิธี HGLM วิธี MIMIC และวิธี BAYESIAN มีค่าสอดคล้องกัน

2. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) ระหว่างวิธี HGLM วิธี MIMIC และวิธี BAYESIAN เมื่อเพศและสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียนต่างกัน มีความสอดคล้องกัน

ประโยชน์ที่คาดว่าจะได้รับ

1. ประโยชน์ด้านวิชาการ

1.1 เป็นแนวทางในการระบุลักษณะของข้อสอบที่ทำหน้าที่ต่างกันและเป็นแนวทางปรับปรุงข้อสอบให้มีความยุติธรรมสำหรับผู้สอบเมื่อมีการพิจารณาหรือจำแนกกลุ่มผู้สอบตามตัวแปรอื่น ๆ

1.2 ผลการวิจัยจะเป็นประโยชน์ในเชิงวิชาการในด้านการประยุกต์วิธีการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ (DIF) ตามทฤษฎีการตอบสนองข้อสอบ ซึ่งจะเป็นแนวทางในการนำไปใช้วิเคราะห์ข้อสอบระดับชาติอื่น ๆ ต่อไป

2. ประโยชน์ด้านการนำไปใช้

2.1 เป็นแนวทางในการศึกษาเรื่องวิธีการประมาณค่าพารามิเตอร์ และการทำหน้าที่ต่างกันของข้อสอบ โดยใช้โปรแกรม HLM โปรแกรม Mplus และโปรแกรม WinBUGS

2.2 ทำให้ทราบว่าข้อสอบลักษณะใดทำหน้าที่ต่างกันสำหรับกลุ่มผู้สอบจำแนกตามเพศและสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน เพื่อศูนย์ทดสอบทางการศึกษาและสำนักทดสอบต่าง ๆ หรือผู้ที่เกี่ยวข้องในการสร้างแบบสอบระดับชาตินำสารสนเทศที่ให้ไปใช้พัฒนาแบบสอบที่มีความยุติธรรมต่อไป

ขอบเขตของการวิจัย

1. ประชากรที่ใช้ในการวิจัยครั้งนี้เป็นนักเรียนชั้นประถมศึกษาปีที่ 3 ที่เข้าสอบวัดผลสัมฤทธิ์ทางการเรียนเพื่อประเมินคุณภาพการศึกษาระดับชาติปีการศึกษา 2553 ซึ่งเป็นนักเรียนที่เข้าสอบวิชาภาษาไทย คณิตศาสตร์ และวิทยาศาสตร์ จำนวน 592,525 คน

2. การวิจัยครั้งนี้เป็นการศึกษาในบริบทของประเทศไทย โดยมุ่งศึกษาในกลุ่มตัวแปรเพศและสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน โดยไม่ศึกษาในตัวแปรอื่น เช่น ภาษาพูด เชื้อชาติ ศาสนา ประสบการณ์

3. การวิจัยครั้งนี้เป็นการประมาณค่าพารามิเตอร์ของข้อสอบ ด้วยวิธี HGLM – 2L วิธี MIMIC และวิธี BAYESIAN เป็นการเปรียบเทียบผลการวิเคราะห์ข้อสอบ และพารามิเตอร์ความสามารถของผู้สอบ

4. การวิจัยครั้งนี้มุ่งตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) สำหรับผู้สอบจำแนกตามเพศและสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน โดยประยุกต์ใช้โปรแกรม HLM และโปรแกรม Mplus และโปรแกรม WinBUGS ซึ่งทั้ง 3 วิธีนี้ต่างก็เป็นวิธีการที่ใช้หลักการวิเคราะห์บนพื้นฐานของทฤษฎีการตอบสนองข้อสอบ หากผลการทดสอบด้วยสถิติ พบว่ามีความแตกต่างกันแสดงถึงการมีเพศและสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียนแตกต่างกัน มีโอกาสในการตอบข้อสอบได้ถูกต้องไม่เท่ากันจึงเกิดการทำหน้าที่ต่างกันของข้อสอบ

นิยามศัพท์เฉพาะ

การทำหน้าที่ต่างกันของข้อสอบ (DIF) หมายถึง ข้อสอบที่ทำให้ผลการตอบของผู้สอบที่มีความสามารถเท่ากันในสิ่งที่ต้องการวัด มีโอกาสตอบข้อสอบข้อนั้นได้ถูกต้องไม่เท่ากันเนื่องจากอยู่ในกลุ่มย่อยต่างกัน ซึ่งในการวิจัยนี้ ศึกษาการทำหน้าที่ต่างของข้อสอบและแบบสอบ สำหรับกลุ่มผู้สอบจำแนกตาม เพศ และสถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน

เพศ หมายถึง เพศของนักเรียนที่เป็นกลุ่มตัวอย่างในเขตกรุงเทพมหานครและปริมณฑล และนอกเขตกรุงเทพมหานครและปริมณฑล

สถานที่ตั้งทางภูมิศาสตร์ของโรงเรียน หมายถึง เขตพื้นที่ที่โรงเรียนตั้งอยู่ โดยจำแนกเป็น 2 กลุ่ม ได้แก่ 1) เขตกรุงเทพมหานครและปริมณฑล และ 2) นอกเขตกรุงเทพมหานครและปริมณฑล

ผลการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกัน หมายถึง จำนวนข้อสอบที่ทำหน้าที่ต่างกันความสอดคล้องในการตรวจสอบความสัมพันธ์ระหว่างวิธีตรวจสอบ อัตราความไม่สอดคล้องในการตรวจสอบและอัตราความสอดคล้องในการตรวจสอบ

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) หมายถึง วิธีการวิเคราะห์ทางสถิติที่ใช้เพื่อบ่งบอกว่ามีการทำหน้าที่ต่างกันของข้อสอบระหว่างผู้สอบกลุ่มอ้างอิงและกลุ่มสนใจที่มีความสามารถระดับเดียวกัน ซึ่งวิธีการในงานวิจัยครั้งนี้เป็นวิธีการตรวจสอบที่ใช้กับข้อสอบที่มีการให้คะแนนแบบสองค่า

วิธี BAYSIAN หมายถึง วิธีประมาณค่าพารามิเตอร์ของข้อสอบ และความสามารถของผู้เข้าสอบ ด้วยวิธีของเบส์ที่มีการกำหนด Prior Distribution ของค่าความยากของข้อสอบไว้

วิธี MIMIC หมายถึง โมเดลลิสเรลที่มีตัวแปรแฝงเพียงตัวแปรเดียวโดยที่ตัวแปรแฝงนั้นได้รับอิทธิพลจากตัวแปรภายนอกสังเกตได้หลายตัวแปรและส่งอิทธิพลไปยังตัวแปรภายในสังเกตได้หลายตัวแปร

กลุ่มอ้างอิง (Reference Group: R) หมายถึง กลุ่มผู้สอบที่คาดว่าจะได้รับประโยชน์จากการตอบข้อสอบที่ทำหน้าที่ต่างกัน คือเป็นกลุ่มที่มีความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องสูงกว่าผู้สอบอีกกลุ่มหนึ่งทั้ง ๆ ที่มีความสามารถเท่ากัน

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

โมเดลเชิงเส้นตรงทั่วไประดับลดหลั่น (Hierarchical Generalized Linear Model: HGLM) หมายถึง รูปแบบหรือลักษณะการวิเคราะห์ข้อมูลเชิงเส้นทั่วไป ที่มีการประยุกต์ปรับให้ดำเนินการวิเคราะห์ข้อมูลร่วมกับโมเดลการวิเคราะห์ข้อมูลแบบอื่น ๆ และการวิเคราะห์พหุระดับที่มีข้อมูลสอดแทรกเป็นระดับลดหลั่นได้ โดยในระดับการวิเคราะห์ที่ 1 เป็นการวิเคราะห์ตามโมเดลเชิงเส้นทั่วไป (Generalized Linear Model: GLM) แล้วใช้ฟังก์ชั่นโยง (Link Function) ที่เป็นฟังก์ชั่นโยงแบบโลจิท (Logit Link Function) ในการปรับค่าเฉลี่ยของระดับการวิเคราะห์ที่ 1 นำมาสู่การวิเคราะห์ในระดับต่อไปได้โดยใช้โมเดลการวิเคราะห์พหุระดับด้วยโมเดลเชิงเส้นตรงระดับลดหลั่น (HLM) โดยการวิเคราะห์ระดับที่ 1 ตัวแปรตามจึงเป็น Log – odds ของความน่าจะเป็นในการตอบข้อสอบได้ถูกต้อง

การวิเคราะห์ข้อสอบพหุระดับ หมายถึง ขั้นตอนหรือวิธีการตรวจสอบคุณภาพของข้อสอบที่พิจารณาลักษณะข้อมูลของการตอบข้อสอบเป็นระดับลดหลั่น ซึ่งลักษณะการวิเคราะห์นี้กำหนดโมเดลการวิเคราะห์เป็น 2 ระดับ โดยการวิเคราะห์ในระดับที่ 1 ระดับข้อสอบเป็นการจัดให้ข้อสอบสอดแทรกในตัวบุคคล (Between Item within Person) การวิเคราะห์ระดับที่ 2 ระดับผู้สอบเป็นการจัดให้บุคคลสอดแทรกในโรงเรียน (Between Person within School)

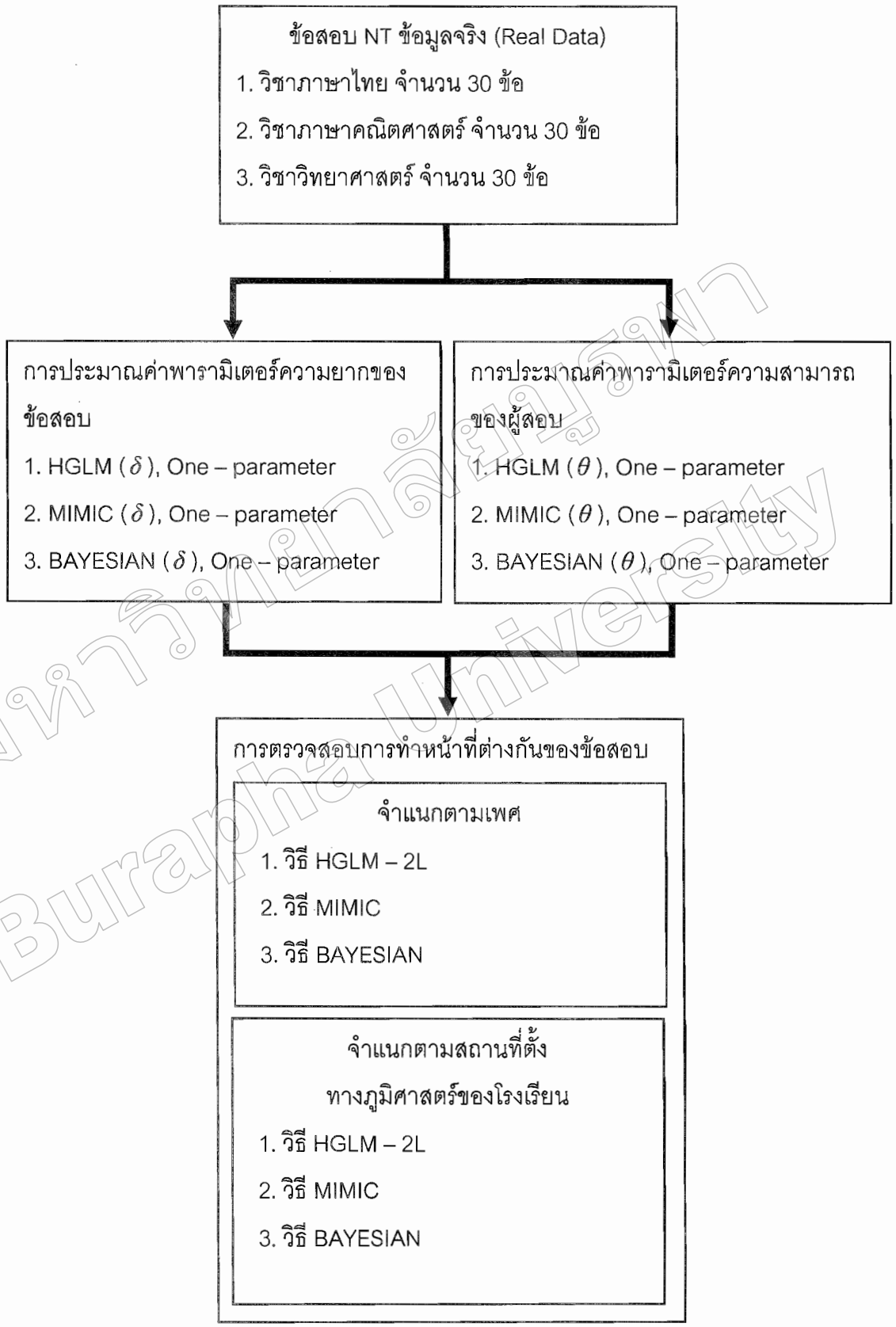
โอกาสในการตอบข้อสอบถูก หมายถึง ค่าสัดส่วนหรือค่าความน่าจะเป็นที่ผู้ตอบซึ่งมีความสามารถ θ จะตอบข้อสอบข้อที่ *i* ได้ถูกต้อง ซึ่งได้มาจากการประมาณค่าโดยใช้โมเดลทฤษฎีการตอบสนองข้อสอบแบบ 1 พารามิเตอร์ (One – parameter Model) หรือ Rasch Model พิจารณาจากค่าพารามิเตอร์ความยากของข้อสอบข้อที่ *i* ซึ่งเป็นค่าที่แสดงตำแหน่งของ ICC ณ จุด θ ที่มีโอกาสตอบข้อสอบถูก .50 และค่าคงที่ฐานของลอการิทึมธรรมชาติ (Natural Log: *e*) มีค่าเท่ากับ 2.718 โดยทำให้อยู่ในรูปคะแนนมาตรฐาน มีค่าอยู่ระหว่าง $-\infty$ ถึง $+\infty$

ค่าพารามิเตอร์ข้อสอบ (δ_i) หมายถึง ค่าที่แสดงตำแหน่งของโค้งลักษณะข้อสอบ (ICC) ณ จุด θ ที่มีโอกาสตอบข้อสอบถูก .50 ของข้อสอบผลสัมฤทธิ์ทางการเรียนวิชาภาษาไทย คณิตศาสตร์ และ วิทยาศาสตร์ของนักเรียนชั้นประถมศึกษาปีที่ 3

ค่าพารามิเตอร์ความสามารถของผู้สอบ (θ_j) หมายถึง การประมาณค่าได้จากการตอบแบบสอบวัดผลสัมฤทธิ์ทางการเรียนวิชาภาษาไทย คณิตศาสตร์ และวิทยาศาสตร์ของนักเรียนชั้นประถมศึกษาปีที่ 3 เพื่อประเมินคุณภาพการศึกษาระดับชาติปีการศึกษา 2553 แบบ 1 พารามิเตอร์ (IRT – 1 Parameter Logistic Measurement Model)

กรอบแนวคิดการวิจัย

จากการศึกษาแนวคิดทฤษฎีและงานวิจัยต่าง ๆ ที่เกี่ยวข้อง ผู้วิจัยได้สรุปรวบรวมโดยการวิเคราะห์และสังเคราะห์ เพื่อกำหนดกรอบแนวคิดการวิจัย ซึ่งกรอบแนวคิดในการวิจัยแสดงให้เห็นว่าการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธี HGLM วิธี MIMIC และ BAYESIAN ซึ่งตั้งอยู่บนพื้นฐานของทฤษฎีการตอบสนองข้อสอบ ส่งผลต่อประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ตรวจให้คะแนนแบบสองค่า (Dichotomous) โดยทั้งสามวิธีมีความสามารถในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบสองค่า (Dichotomous) ดังต่อไปนี้



ภาพที่ 1 – 1 กรอบแนวคิดการวิจัย