บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

หลักสูตรการศึกษาขั้นพื้นฐาน พุทธศักราช 2544 ที่สถานศึกษานำไปจัดทำเป็นหลักสูตรสถานศึกษา เพื่อใช้ในการจัดการเรียนการสอนสำหรับแต่ละสถานศึกษานั้น มีข้อกำหนดให้สถานศึกษาต้องดำเนินการเกี่ยวกับการวัดและประเมินผลการเรียนของผู้เรียนทุกคนในปีสุดท้ายของแต่ละช่วงชั้น คือ ชั้นประถมศึกษาปีที่ 3 ชั้นประถมศึกษาปีที่ 6 ชั้นมัธยมศึกษาปีที่ 3 และชั้นมัธยมศึกษาปีที่ 6 ให้เข้ารับการประเมินคุณภาพการศึกษาระดับชาติ ในกลุ่มสาระการเรียนรู้ต่าง ๆ ซึ่งดำเนินการโดยกระทรวงศึกษาธิการ การประเมินผลดังกล่าวมีจุดมุ่งหมายเพื่อตรวจสอบคุณภาพการศึกษาของสถานศึกษาและคุณภาพการศึกษาของชาติ ข้อมูลจากการประเมินนำไปใช้ในการพัฒนาคุณภาพผู้เรียน คุณภาพการจัดการศึกษาของสถานศึกษา และคุณภาพการศึกษาของชาติ (กรมวิชาการ, 2545, หน้า 24)

การประเมินผลสัมฤทธิ์ทางการเรียนระดับชาติ มีประโยชน์ ดังนี้

1. ทำให้สามารถเปรียบเทียบผลการประเมินคุณภาพระหว่างระดับชั้นเรียน ระดับสถานศึกษา ระดับเขตพื้นที่การศึกษา และระดับชาติ ตลอดจนการประเมินภายนอกได้อย่างสมเหตุสมผล

2. สามารถประเมินได้ทั้งผลสัมฤทธิ์ทางวิชาการตามหลักสูตร และความถนัดทางการเรียน

3. ส่งเสริมและกระตุ้นให้สถานศึกษาให้ความสนใจอย่างจริงจังในการพัฒนาผลสัมฤทธิ์ที่สำคัญของหลักสูตร

4. สามารถใช้ผลการประเมินให้เป็นประโยชน์ทั้งในระดับผู้เรียน ระดับชั้นเรียน ระดับสถานศึกษา ระดับเขตพื้นที่การศึกษา และระดับชาติ

5. สร้างแรงจูงใจ กระตุ้นและท้าทายให้ผู้เรียนทุกคนตั้งใจใฝ่หาสัมฤทธิ์ผลทางการเรียนและด้านอื่น ๆ

6. เพื่อเป็นข้อมูลสร้างความมั่นใจเกี่ยวกับคุณภาพของนักเรียน แก่ผู้เกี่ยวข้องทั้งภายในและภายนอกสถานศึกษา

จากประโยชน์ของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ (National Test) ที่มีความสำคัญทั้งต่อระดับผู้เรียน ระดับสถานศึกษา ระดับเขตพื้นที่การศึกษา และระดับชาติ

การพัฒนาแบบทดสอบจึงต้องคำนึงถึงคุณภาพของข้อสอบรายข้อ และแบบทดสอบทั้งฉบับอย่างรอบด้าน โดยเฉพาะประเด็นด้าน "ความตรง" ซึ่งเป็นหัวใจสำคัญของคุณภาพแบบทดสอบทั้งฉบับในการสร้าง และการตรวจสอบคุณภาพของแบบทดสอบ จะต้องคำนึงถึงคุณภาพด้านความตรงเป็นสำคัญ ทั้งนี้เพราะว่าความตรงเป็นความสามารถของแบบทดสอบในการอธิบายธรรมชาติหรือโครงสร้างของคุณลักษณะทางจิตวิทยาที่อยู่เบื้องหลังกลุ่มข้อสอบ หรือสิ่งที่แบบทดสอบมุ่งวัดเป็นสิ่งที่เราไม่สามารถสังเกตได้โดยตรง แต่เป็นสิ่งที่มีอยู่ตามทฤษฎีทางจิตวิทยา (เสรี ชัดแช้ม, 2544, หน้า 139) ถ้าผลการวัดที่ได้มีค่าที่ใกล้เคียงกับค่าคุณลักษณะที่แท้จริงเพียงใด ก็ถือว่าการวัดมีความตรงมากขึ้นเท่านั้น

นักการศึกษานิยมตรวจสอบความตรงของแบบทดสอบ 3 ประเภท คือ 1) ความตรงเชิงเนื้อหา (Content Validity) 2) ความตรงเชิงเกณฑ์สัมพันธ์ (Criterion Related Validity) และ 3) ความตรงเชิงโครงสร้าง (Construct Validity) ส่วนประเภทหนึ่งที่ใช้ตรวจสอบคุณภาพด้านความตรงของแบบทดสอบ คือ การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning; DIF) (Mazor, Clauser, & Hambleton, 1992; Kim, Kim, & Cohen, 1994 อ้างถึงในวลีมาศ แซ่อึ้ง, 2543, หน้า 1) โดยเป็นการตรวจสอบในด้านความไม่ยุติธรรมของข้อสอบ (Item Unfaire) ข้อสอบข้อใดทำหน้าที่ต่างกันจะถูกคัดออกจากแบบทดสอบ โดยทั่วไปแล้วในแบบทดสอบมาตรฐานที่ใช้วัดผลสัมฤทธิ์ทางการเรียน ถ้ามีสัดส่วนของข้อสอบทำหน้าที่ต่างกันร้อยละ 10 ถึง 15 ถือว่าไม่ผิดปกติ แต่ถ้ามีสัดส่วนของข้อสอบทำหน้าที่ต่างกันร้อยละ 20 ถือว่าเป็นเรื่องผิดพลาดอย่างมาก (Clauser, 1993 cited in Narayanan & Swaminathan, 1994 อ้างถึงในวลีมาศ แซ่อึ้ง, 2543, หน้า 1)

ในการทดสอบแต่ละครั้ง ผู้สอบอาจมีลักษณะแตกต่างกัน เช่น เชื้อชาติ ศาสนา วัฒนธรรม ภูมิลำเนา สังคม เพศ ภาษา อายุ และประสบการณ์ เป็นต้น ผู้สอบดังกล่าวอาจไม่ได้รับความยุติธรรมในการทำข้อสอบ โดยข้อสอบบางข้ออาจมีความลำเอียงเข้าข้างผู้สอบกลุ่มย่อยบางกลุ่มของผู้เข้าสอบทั้งหมด ซึ่งทำให้เกิดการได้เปรียบเสียเปรียบระหว่างผู้สอบกลุ่มย่อยด้วยกันทั้ง ๆ ที่สอบด้วยข้อสอบข้อเดียวกันหรือแบบทดสอบฉบับเดียวกัน แสดงว่าแบบทดสอบหรือข้อสอบขาดความตรง สาเหตุดังกล่าวอาจเนื่องมาจากแบบทดสอบไม่ได้วัดความสามารถเป้าหมายที่ต้องการวัด (Target Ability; θ) เพียงอย่างเดียว แต่ยังวัดความสามารถแทรกซ้อนที่ไม่ต้องการวัด (Nuisance Ability; η) อีกด้วย ตัวอย่างเช่น แบบทดสอบวัดความสามารถด้านคำศัพท์วิชาภาษาไทยฉบับหนึ่ง ข้อสอบบางข้ออาจถามความรู้สำหรับผู้หญิงโดยเฉพาะ เช่น ความรู้เกี่ยวกับงานในบ้าน ในขณะที่ข้อสอบบางข้ออาจถามความรู้สำหรับผู้ชายเป็นพิเศษ เช่น ความรู้เรื่องกีฬา จากสถานการณ์ดังกล่าว แบบทดสอบวัดความสามารถคำศัพท์ ในวิชาภาษาไทยเป็นความสามารถ

เป้าหมายที่ต้องการวัด (θ) ส่วนความสามารถทางด้านกีฬา และงานในบ้านเป็นความสามารถแทรกซ้อน (η) ข้อสอบทุกข้อในแบบทดสอบจะวัดความสามารถเป้าหมาย ส่วนข้อสอบบางข้อที่ทำหน้าที่ต่างกันจะวัดทั้งความสามารถเป้าหมาย และความสามารถแทรกซ้อน (Nandakumar, 1993) นั่นแสดงว่าถ้าผู้สอบกลุ่มย่อยกลุ่มใดมีความสามารถแทรกซ้อนสูงกว่าก็มีโอกาสในการตอบข้อสอบได้ถูกต้องมากกว่า ทั้ง ๆ ที่ระดับความสามารถเป้าหมายที่ต้องการวัดเท่ากัน จึงมีผลทำให้ข้อสอบทำหน้าที่ต่างกัน

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นการเปรียบเทียบผลการตอบข้อสอบระหว่างผู้สอบ 2 กลุ่ม ได้แก่ กลุ่มแรก เรียกว่า "กลุ่มเปรียบเทียบ" (Focal Group) เป็นกลุ่มที่ผู้วิจัยสนใจศึกษา และคาดว่าจะเป็นกลุ่มที่เสียประโยชน์ในการตอบข้อสอบ และกลุ่มที่สอง เรียกว่า "กลุ่มอ้างอิง" (Reference Group) ซึ่งเป็นกลุ่มที่คาดว่าจะได้ประโยชน์จากการตอบข้อสอบได้ถูกต้อง นักวัดผลได้ให้ความสนใจกับการนำเทคนิคการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมาใช้ตรวจสอบคุณภาพของแบบทดสอบกันมากขึ้น แต่ส่วนใหญ่เน้นศึกษาและวิจัยเกี่ยวกับวิธีการในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ภายใต้เงื่อนไขต่าง ๆ ว่าวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวิธีใดมีความถูกต้องในการตรวจสอบมากที่สุด หรือดีที่สุด และมีความคลาดเคลื่อนในการตรวจสอบน้อยที่สุด โดยวิธีการตรวจสอบที่กำลังอยู่ในความสนใจ เช่น วิธีโค้งลักษณะข้อสอบ 3 พารามิเตอร์ (Item Characteristic Curve 3 Parameter: ICC-3) วิธีการทดสอบอัตราส่วนไลค์ลิฮูด (Likelihood Ratio Test) และวิธีซิปเทสท์ (SIBTEST) เหตุที่วิธีเหล่านี้ได้รับความสนใจ เนื่องจากค่าสถิติของข้อสอบที่ได้จากวิธีการตรวจสอบในกลุ่มนี้ไม่แปรเปลี่ยนไปตามกลุ่มตัวอย่างที่สุ่มมาจากประชากรเดียวกัน ค่าพารามิเตอร์ของข้อสอบตามทฤษฎีการตอบสนองข้อสอบมีความแกร่ง ผลที่ได้จึงน่าเชื่อถือมากกว่าวิธีการตรวจสอบในกลุ่มทฤษฎีการทดสอบแบบดั้งเดิม (Classical Test Theory: CTT) ที่มีข้อบกพร่องหลายอย่าง สำหรับเงื่อนไขที่ใช้ในการศึกษาเกี่ยวกับวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ มีการศึกษาทั้งในสภาพของข้อมูลจริง หรือการจำลองข้อมูลขึ้นมา ศึกษากับแบบทดสอบที่สร้างขึ้นเองหรือแบบทดสอบมาตรฐานที่สร้างโดยหน่วยงานอื่น ๆ ของประเทศ โดยมีการศึกษาตัวแปรต้นและตัวแปรตามพอจะสรุปได้ ดังนี้ ตัวแปรต้น ได้แก่ วิธีการตรวจสอบ กลุ่มเปรียบเทียบที่ศึกษา เช่น เพศ เชื้อชาติ ศาสนา พื้นที่ทางภูมิศาสตร์ เป็นต้น ขนาดของกลุ่มตัวอย่าง ความยาวของแบบทดสอบ, สัดส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบ, การใช้คะแนนที่ได้จากแบบทดสอบ เช่น ใช้คะแนนรวม คะแนนแบบทดสอบย่อย คะแนนแยกตามพฤติกรรม คะแนนแยกตามเนื้อหา เป็นต้น, สัดส่วนของข้อสอบทำหน้าที่ต่างกัน เป็นต้น ตัวแปรตาม ได้แก่ ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ, ความสอดคล้องของผลการตรวจสอบการทำหน้าที่ต่างกันของ

ข้อสอบระหว่างวิธีการตรวจสอบ, ความแตกต่างของผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธีการตรวจสอบ, อำนาจการทดสอบ (Power of Test), ความคลาดเคลื่อนประเภทที่ 1 (Type I Error), ความคลาดเคลื่อนประเภทที่ 2 (Type II Error), ค่าความเที่ยง (Reliability), ความตรง (Validity), สหสัมพันธ์อันดับของผู้สอบ เป็นต้น

โดยปกติเมื่อผู้พัฒนาแบบทดสอบตรวจสอบพบว่าข้อสอบข้อใดทำหน้าที่ต่างกันแล้วก็จะตัดข้อสอบข้อนั้นออกจากแบบทดสอบ ไม่นำมาคิดคะแนนให้กับผู้สอบ เพื่อให้เกิดความยุติธรรมกับผู้เข้าสอบอย่างเท่าเทียมกัน แต่การตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบนั้นอาจทำให้โครงสร้างในการวัด และคุณภาพของแบบทดสอบเปลี่ยนแปลงไปจากเดิม ดังนั้นควรมีการศึกษาเพิ่มเติมเกี่ยวกับผลที่เกิดขึ้นจากการตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบว่าส่งอิทธิพลต่อคุณภาพของแบบทดสอบทั้งฉบับอย่างไรบ้าง แบบทดสอบมีคุณภาพด้านต่าง ๆ เพียงพอที่จะนำไปใช้ได้หรือไม่ สำหรับการศึกษาเกี่ยวกับอิทธิพลของการทำหน้าที่ต่างกันของข้อสอบ หลังจากตัดข้อสอบที่ทำหน้าที่ต่างกันออกจากแบบทดสอบ เช่น เกษร หว่างจิตร (2539) ได้ศึกษาพบว่า ค่าความเที่ยง และความตรงเชิงโครงสร้าง ของแบบทดสอบฉบับก่อนกับฉบับหลังตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบ ส่วนใหญ่ไม่แตกต่างกัน เรวดี อินทสะระ (2539) พบว่า ค่าสหสัมพันธ์อันดับของผู้สอบไม่ว่าจะคิดคะแนนมาตรฐานทีปกติหรือคิดคะแนนน้ำหนักความสามารถ และใช้ข้อสอบจำนวนทั้งหมดหรือใช้เฉพาะข้อสอบที่ไม่มีการทำหน้าที่ต่างกัน ต่างมีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติ โรโซลสกี และไรส์ (Rozowski & Reith, 1999) พบว่าแบบทดสอบที่มีข้อสอบทำหน้าที่ต่างกันกับแบบทดสอบที่ไม่มีข้อสอบทำหน้าที่ต่างกัน ไม่ได้ทำให้แบบทดสอบมีคุณภาพในการวัดหรือความตรงต่ำลงมากนัก ส่วนค่าสหสัมพันธ์อันดับของผู้สอบที่มีข้อสอบทำหน้าที่ต่างกันกับแบบทดสอบที่ไม่มีข้อสอบทำหน้าที่ต่างกันมีความสัมพันธ์กันสูง รักชนก ยี่สุ่นศรี (2544) พบว่า แบบทดสอบฉบับก่อนกับฉบับหลังตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบ มีความตรงเชิงโครงสร้างของแบบทดสอบไม่แตกต่างกัน และแบบทดสอบฉบับหลังตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบส่วนใหญ่มีค่าความเที่ยงลดลง ส่วนสหสัมพันธ์อันดับของผู้สอบ พบว่า ในทุกกรณีมีความสัมพันธ์ในทางบวกซึ่งกันและกันอย่างมีนัยสำคัญ

จากผลการศึกษาข้างต้น ทำให้ผู้วิจัยสนใจที่จะศึกษาอิทธิพลของการทำหน้าที่ต่างกันของข้อสอบเพิ่มเติมในประเด็นด้าน ค่าความเที่ยง ความตรงเชิงโครงสร้าง ความคงที่ของโครงสร้างองค์ประกอบ และสหสัมพันธ์อันดับของผู้สอบ ของแบบทดสอบฉบับก่อนกับฉบับหลังตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบ โดยใช้วิธีการตรวจสอบที่แตกต่างไปจากผู้ที่เคยศึกษาไว้ จำนวน 2 วิธี คือ วิธีซิปเทสท์ปรับใหม่ (Modify SIBTEST) และวิธีถดถอยโลจิสติก

(Logistic Regression) ซึ่งจากงานวิจัยของ วลีมาศ แซ่อึ้ง (2543) พบว่า วิธีทั้งสองมีอำนาจการทดสอบในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เท่าเทียมกันภายใต้เงื่อนไขเกือบทุกเงื่อนไขของการตรวจสอบ และวิธีทั้งสองมีอำนาจการทดสอบสูงกว่าวิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล (Mantel-Haenszel) ภายใต้เงื่อนไขเกือบทุกเงื่อนไข ทั้งสองวิธีมีคุณลักษณะที่สำคัญ พอจะสรุปได้ดังนี้

วิธีซิปเทสท์ปรับใหม่ ได้ปรับปรุงขั้นตอนการวิเคราะห์ที่แตกต่างไปจากวิธีซิปเทสท์ (SIBTEST) แบบเดิม มีประสิทธิภาพมากกว่า พัฒนาอยู่บนพื้นฐานของทฤษฎีการตอบสนองข้อสอบชนิดพหุมิติ เป็นวิธีทดสอบแบบนันพาราเมตริก ได้ออกแบบมาเพื่อใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป (Uniform) โดยเฉพาะ คำนวณได้ง่าย ไม่ซับซ้อน ประหยัดค่าใช้จ่าย และไม่จำเป็นต้องใช้กลุ่มตัวอย่างที่มีขนาดใหญ่ นอกจากนี้ยังตัดสินข้อสอบทำหน้าที่ต่างกันโดยใช้สถิติทดสอบนัยสำคัญ

วิธีถดถอยโลจิสติก เป็นวิธีที่อยู่บนพื้นฐานของแบบจำลองการวิเคราะห์สมการถดถอยโลจิสติก เป็นแบบจำลองที่สามารถเพิ่มตัวแปรความสามารถ และปฏิสัมพันธ์เข้าไปในสมการได้ และสอดคล้องกับธรรมชาติของการทำหน้าที่ต่างกันของแบบทดสอบได้ดี ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป (Uniform) และแบบอเนกรูป (Nonuniform) ได้ ใช้โมเดลที่มีความยืดหยุ่น ตัดสินข้อสอบทำหน้าที่ต่างกันโดยใช้สถิติทดสอบนัยสำคัญ

การวิจัยครั้งนี้ผู้วิจัยเลือกศึกษา วิชาภาษาไทย เนื่องจากเป็นวิชาที่มักจะพบว่ามีความลำเอียงต่อเพศอย่างชัดเจน โดยส่วนใหญ่จะลำเอียงเข้าข้างผู้สอบเพศหญิง และข้อสอบวิชาภาษาไทย เป็นวิชาที่มีวัฒนธรรมทางภาษาเข้ามาเกี่ยวข้องด้วย จึงทำให้ข้อความ และถ้อยคำลำเอียงได้ง่าย ซึ่งสอดคล้องกับ ลิน (Linn, 1993) ที่ระบุว่าข้อสอบทำหน้าที่ต่างกันมักจะพบในแบบทดสอบด้านภาษา โดยเขาศึกษาพบว่าในแบบทดสอบ SAT จำนวน 7 ฉบับ พบข้อสอบทำหน้าที่ต่างกันจำนวน 29 ข้อ เป็นข้อสอบทางด้านภาษา จำนวน 25 ข้อ สุพัฒน์ สุกมลสันต์ (2534) พบว่า ข้อสอบในแบบทดสอบวิชาภาษาอังกฤษมีความลำเอียงต่อเพศโดยลำเอียงเข้าข้างผู้สอบเพศหญิง และกาญจนา วัธนสุนทร (2537) พบว่าข้อสอบทำหน้าที่ต่างกันมากกว่าครึ่งในวิชาภาษาอังกฤษลำเอียงเข้าข้างผู้สอบเพศหญิง เกอเรย์ (Giray, 1995) พบว่าผู้สอบเพศหญิงจะมีความสามารถในด้านการใช้ภาษา หรือเกี่ยวกับคำศัพท์ได้ดีกว่าผู้สอบเพศชาย

ความยาวของแบบทดสอบในการศึกษา คือ จำนวน 40 ข้อ ซึ่งจะไม่ส่งผลกระทบต่ออำนาจการทดสอบในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ดังผลการศึกษาของ โรเจอร์ และสวามินาธาน (Rogers & Swaminathan, 1993) ที่พบว่า ความยาวของแบบทดสอบไม่มีผลต่ออำนาจการทดสอบของวิธีแมนเทล-แฮนส์เซลและวิธีถดถอยโลจิสติก และกาญจนา วัธนสุนทร

(2537) ที่พบว่า ความยาวของแบบทดสอบไม่มีผลกระทบต่ออัตราการตรวจสอบของวิธีแมนเทล-แฮนส์เซล และวิธีซิปเทสท์

สัดส่วนกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบที่ใช้ในการศึกษา คือ เพศชาย 1,000 คน ต่อ เพศหญิง 1,000 คน (รวม 2,000 คน) ซึ่งเป็นขนาดกลุ่มตัวอย่างที่เหมาะสมและมีขนาดใหญ่เพียงพอที่จะสามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้อย่างถูกต้อง แม่นยำ สอดคล้องกับผลการวิจัยของ จิตติมา วรรณศรี (2539) ที่พบว่า เมื่อใช้ขนาดกลุ่มตัวอย่างขนาด 1,000 คน สามารถตรวจสอบพบข้อสอบที่ทำหน้าที่ต่างกันได้ถูกต้องร้อยละ 100 และอัตราความคลาดเคลื่อนอยู่ในเกณฑ์ที่เหมาะสม เมเซอร์ คลาสเซอร์ และแฮมเบิลตัน (Mazor, Clauser & Hambleton, 1992 อ้างถึงใน จิตติมา วรรณศรี, 2539) ที่พบว่า เมื่อใช้กลุ่มตัวอย่างขนาดใหญ่ ทำให้ตรวจพบข้อสอบทำหน้าที่ต่างกันได้มากกว่าการใช้กลุ่มผู้สอบขนาดเล็ก คือ เมื่อใช้ขนาดกลุ่มตัวอย่าง 2,000 คน จะเกิดความถูกต้องในการตรวจสอบ ร้อยละ 70 ถึง 75 นารายานาน และสวามินาธาน (Narayanan & Swaminathan, 1994; 1996) พบว่า ขนาดกลุ่มตัวอย่าง และอัตราส่วนของกลุ่มอ้างอิงต่อกลุ่มเปรียบเทียบมีผลต่ออำนาจการทดสอบและความคลาดเคลื่อนประเภทที่ 1 คือ เมื่อเพิ่มขนาดกลุ่มตัวอย่าง จะทำให้อำนาจการทดสอบมีค่าเพิ่มมากขึ้นและเมื่อใช้กลุ่มตัวอย่างขนาดเล็ก อัตราความคลาดเคลื่อนประเภทที่ 1 จะต่ำกว่าใช้กลุ่มตัวอย่างขนาดใหญ่ ภายใต้เงื่อนไขเกือบทุกเงื่อนไขของการตรวจสอบ

การศึกษาเกี่ยวกับอิทธิพลของการทำหน้าที่ต่างกันของข้อสอบ จึงเป็นการสะท้อนให้เห็นคุณภาพของแบบทดสอบในด้านที่เป็นหัวใจสำคัญของแบบทดสอบ ได้แก่ ค่าความเที่ยง ความตรงเชิงโครงสร้าง และความคงที่ของโครงสร้างองค์ประกอบ ว่าเปลี่ยนแปลงไปในทิศทางใด หลังจากตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบ ข้อค้นพบที่ได้จะเป็นประโยชน์ต่อผู้ที่นำวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบไปใช้ในการตรวจสอบคุณภาพของแบบทดสอบ ใช้ตัดสินใจว่าสมควรที่จะตัดข้อสอบที่พบว่ามีการทำหน้าที่ต่างกันออกจากแบบทดสอบหรือควรปรับปรุงข้อสอบข้อนั้นเสียใหม่ เพื่อรักษาโครงสร้าง และคุณภาพของแบบทดสอบฉบับนั้นไว้ และผลที่ได้จากการวิจัยสามารถนำไปเป็นสารสนเทศในการปรับปรุงแบบทดสอบให้มีคุณภาพดียิ่งขึ้น

วัตถุประสงค์ของการวิจัย

1. เพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ ชั้นประถมศึกษาปีที่ 6 วิชาภาษาไทย ในกรณีจำแนกกลุ่มผู้สอบตามตัวแปรเพศ

2. เพื่อเปรียบเทียบค่าความเที่ยงของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ

ชั้นประถมศึกษาปีที่ 6 ฉบับก่อนกับฉบับหลังตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบ

3. เพื่อเปรียบเทียบความตรงเชิงโครงสร้างของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ ชั้นประถมศึกษาปีที่ 6 ฉบับก่อนกับฉบับหลังตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบ

4. เพื่อเปรียบเทียบความคงที่ของโครงสร้างองค์ประกอบ กรณีจำแนกกลุ่มผู้สอบตามตัวแปรเพศ ฉบับก่อนกับฉบับหลังตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบ

5. เพื่อวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์อันดับของผู้สอบ ฉบับก่อนกับฉบับหลังตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบ

คำถามของการวิจัย

1. แบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ ชั้นประถมศึกษาปีที่ 6 วิชาภาษาไทย มีข้อสอบทำหน้าที่ต่างกันหรือไม่ คิดเป็นร้อยละเท่าไร

2. ค่าความเที่ยงของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ ชั้นประถมศึกษาปีที่ 6 ฉบับก่อนกับหลังตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบ เป็นอย่างไร แตกต่างกันอย่างมีนัยสำคัญหรือไม่

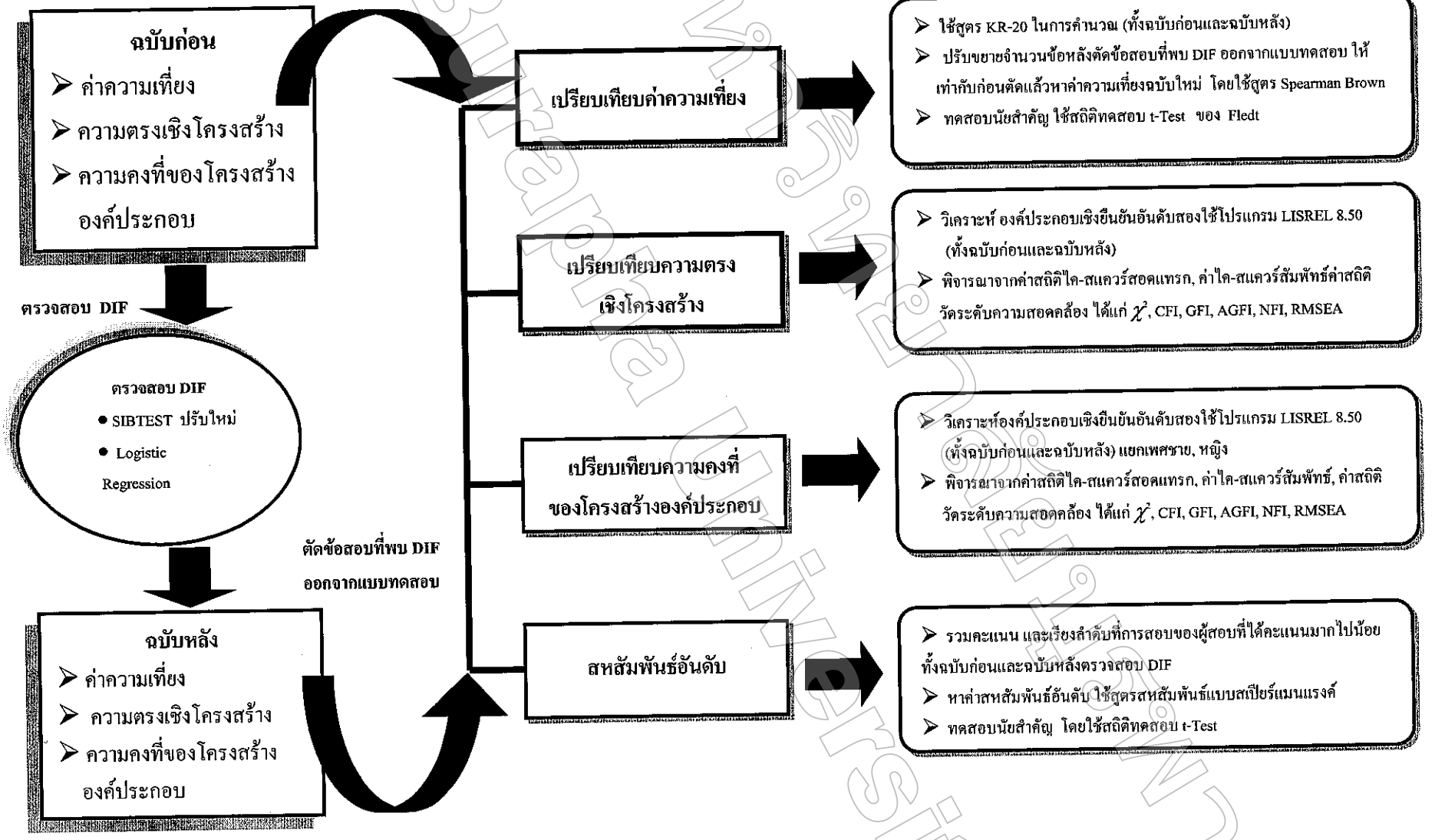
3. ความตรงเชิงโครงสร้างของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ ชั้นประถมศึกษาปีที่ 6 ฉบับก่อนกับฉบับหลังตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบ เป็นอย่างไร แตกต่างกันหรือไม่

4. แบบทดสอบฉบับก่อนกับฉบับหลังตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบในกรณีจำแนกกลุ่มผู้สอบตามตัวแปรเพศ มีโครงสร้างองค์ประกอบแตกต่างกันหรือไม่

5. ลำดับที่การสอบของผู้สอบ จากแบบทดสอบฉบับก่อนตัดข้อสอบกับฉบับหลังตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบ มีความสัมพันธ์กันหรือไม่

กรอบแนวคิดในการวิจัย

ผู้วิจัยได้ศึกษาเอกสาร และงานวิจัยที่เกี่ยวข้องกับวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ 2 วิธี คือ วิธีซิปเทสท์ปรับใหม่กับวิธีถดถอยโลจิสติก ซึ่งเป็นวิธีการตรวจสอบที่มีอำนาจการทดสอบสูง และกำลังอยู่ในความสนใจของนักวัดผล รวมทั้งศึกษางานวิจัยที่เกี่ยวข้องกับอิทธิพลที่เกิดขึ้นจากการตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบ กำหนดเป็นกรอบแนวคิดในการวิจัย ดังภาพที่ 1



ภาพที่ 1 กรอบแนวคิดในการวิจัย

สมมติฐานของการวิจัย

จากผลการศึกษาของ เกษร หว่างจิตร (2539) ที่ศึกษาอิทธิพลของการทำหน้าที่ต่างกันของข้อสอบโดยใช้วิธีแมนเทล-แฮนส์เซลในการตรวจสอบ พบว่า ความตรงเชิงโครงสร้างของแบบทดสอบ ฉบับก่อนกับฉบับหลังตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบ ไม่แตกต่างกัน สอดคล้องกับ โรโซลสกี และไรส์ (Rozowski & Reith, 1999) ที่ศึกษาโดยใช้วิธีแมนเทล-แฮนส์เซลในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ พบว่า แบบทดสอบที่มีข้อสอบทำหน้าที่ต่างกันกับแบบทดสอบที่ไม่มีข้อสอบทำหน้าที่ต่างกัน ไม่ได้ทำให้แบบทดสอบมีคุณภาพในการวัด หรือความตรงต่ำลงมากนัก ส่วนค่าสหสัมพันธ์อันดับของผู้สอบ ระหว่างแบบทดสอบทั้งสองชุดมีความสัมพันธ์กันสูง อารี วัชรโสตถิกุล (2543) พบว่า ค่าความเที่ยงของแบบทดสอบฉบับหลังตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบ แล้วทำการปรับขยายค่าความเที่ยงโดยใช้สูตรสเปียร์แมน-บราวน์ พบว่า ค่าความเที่ยงที่ได้มีค่าแตกต่างจากแบบทดสอบฉบับก่อนตรวจสอบ DIF อย่างมีนัยสำคัญทางสถิติ อีกทั้ง รักชนก ยี่สุ่นศรี (2544) ศึกษาอิทธิพลของการตัดข้อสอบทำหน้าที่ต่างกัน โดยใช้กระบวนการดีเอฟไอทีในการตรวจสอบ พบว่า แบบทดสอบฉบับก่อนกับฉบับหลังตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบ มีค่าความตรงเชิงโครงสร้างของแบบทดสอบไม่แตกต่างกัน และแบบทดสอบฉบับหลังตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบส่วนใหญ่มีค่าความเที่ยงลดลง ส่วนค่าสหสัมพันธ์อันดับของผู้สอบพบว่า มีความสัมพันธ์ในทางบวกซึ่งกันและกันอย่างมีนัยสำคัญ ดังนั้นในการวิจัยจึงกำหนดสมมติฐาน ดังนี้

1. ค่าความเที่ยงของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ ชั้นประถมศึกษาปีที่ 6 วิชาภาษาไทย ฉบับก่อนกับฉบับหลังตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบแตกต่างกัน

2. ความตรงเชิงโครงสร้างของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ ชั้นประถมศึกษาปีที่ 6 ฉบับก่อนกับฉบับหลังตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบไม่แตกต่างกัน

3. แบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ วิชาภาษาไทย กรณีจำแนกกลุ่มผู้สอบตามตัวแปรเพศ ฉบับก่อนกับฉบับหลังตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบ มีโครงสร้างองค์ประกอบไม่แตกต่างกัน

4. ลำดับที่ของผู้สอบ จากแบบทดสอบฉบับก่อนกับฉบับหลังตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบ มีความสัมพันธ์กัน

ประโยชน์ที่คาดว่าจะได้รับ

งานวิจัยครั้งนี้ มุ่งศึกษาอิทธิพลของการทำหน้าที่ต่างกันของข้อสอบ โดยเปรียบเทียบคุณภาพของแบบทดสอบในด้านค่าความเที่ยง ความตรงเชิงโครงสร้าง ความคงที่ของโครงสร้างองค์ประกอบ และหาค่าสหสัมพันธ์อันดับของผู้สอบ ฉบับก่อนกับฉบับหลังตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบโดยใช้ข้อมูลเชิงประจักษ์ ผู้วิจัยจึงคาดว่าจะเป็นประโยชน์ ดังนี้

1. ได้ข้อค้นพบเกี่ยวกับคุณภาพของแบบทดสอบว่า คุณภาพด้านค่าความเที่ยง ความตรงเชิงโครงสร้าง และความคงที่ของโครงสร้างองค์ประกอบ มีความคงที่หรือไม่ เมื่อตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบ

2. เป็นข้อมูลในการตัดสินใจของผู้ที่จะนำวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบไปใช้ในการตรวจสอบคุณภาพของแบบทดสอบ ว่าสมควรที่จะตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบหรือควรจะปรับปรุงข้อสอบข้อนั้นเสียใหม่ เพื่อไม่ให้แบบทดสอบมีคุณภาพด้อยลงไป

3. ได้ข้อเสนอแนะต่อหน่วยงานที่เกี่ยวข้องกับการสร้างแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ ชั้นประถมศึกษาปีที่ 6 ใช้เป็นข้อมูลในปีต่อ ๆ ไป เพื่อหาทางปรับปรุงแก้ไข ข้อสอบทำหน้าที่ต่างกัน ให้สามารถนำไปใช้ได้อย่างยุติธรรมสำหรับผู้สอบต่างกลุ่ม อันจะส่งผลให้ผลการสอบมีความถูกต้องมากยิ่งขึ้น

ขอบเขตของการวิจัย

1. กลุ่มตัวอย่าง

ใช้ข้อมูลทุติยภูมิ ที่เป็นผลการตอบข้อสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ ชั้นประถมศึกษาปีที่ 6 วิชาภาษาไทย ปีการศึกษา 2546 ของนักเรียนสังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน กระทรวงศึกษาธิการ ในเขตพื้นที่การศึกษานครศรีธรรมราช จำนวน 2,000 คน ได้มาโดยการสุ่มแบบแบ่งชั้น (Stratified Random Sampling) แบบไม่กำหนดสัดส่วน โดยแยกเป็นผู้สอบเพศชาย จำนวน 1,000 คน และผู้สอบเพศหญิง จำนวน 1,000 คน

2. เนื้อหา

ศึกษาแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ ชั้นประถมศึกษาปีที่ 6 ปีการศึกษา 2546 วิชาภาษาไทย

3. ตัวแปรที่ศึกษา

3.1 ตัวแปรอิสระ มี 2 ตัว ได้แก่

3.1.1 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ 2 วิธี ได้แก่

3.1.1.1 วิธีซิปเทสท์ปรับใหม่

3.1.1.2 วิธีถดถอยโลจิสติก

3.1.2 ลักษณะของแบบทดสอบ มี 2 ลักษณะ ได้แก่

3.1.2.1 แบบทดสอบฉบับก่อนตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

3.1.2.2 แบบทดสอบฉบับหลังตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบ

3.2 ตัวแปรตาม มี 5 ตัว ได้แก่

3.2.1 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

3.2.2 ค่าความเที่ยงฉบับก่อนกับฉบับหลังการตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบ

3.2.3 ความตรงเชิงโครงสร้างฉบับก่อนกับฉบับหลังตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบ

3.2.4 ความคงที่ของโครงสร้างองค์ประกอบ ในกรณีจำแนกกลุ่มผู้สอบตามตัวแปรเพศ ทั้งฉบับก่อนกับฉบับหลังตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบ

3.2.5 สหสัมพันธ์อันดับของผู้สอบ ฉบับก่อนกับฉบับหลังตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบ

นิยามศัพท์เฉพาะ

การทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) หมายถึง การที่ข้อสอบทำให้ผู้สอบที่มีความสามารถหรือคุณลักษณะที่มุ่งวัดเท่าเทียมกัน มีโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกัน เนื่องจากผู้สอบอยู่ในกลุ่มย่อยต่างกัน การวิจัยครั้งนี้จำแนกกลุ่มย่อยโดยใช้ตัวแปรเพศ

ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ หมายถึง ข้อสอบที่ตรวจสอบพบว่ามีการทำหน้าที่ต่างกัน โดยผลที่ได้มาจากวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ที่พบตรงกันสองวิธี คือ วิธีซิปเทสท์ปรับใหม่กับวิธีถดถอยโลจิสติก

อิทธิพลของการทำหน้าที่ต่างกันของข้อสอบ หมายถึง การเปลี่ยนแปลงของคุณภาพแบบทดสอบในด้านค่าความเที่ยง ความตรงเชิงโครงสร้าง ความคงที่ของโครงสร้างองค์ประกอบ และสหสัมพันธ์อันดับของผู้สอบ ซึ่งเกิดจากการตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบฉบับเดิม

แบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ (National Achievement Test) ชั้นประถมศึกษาปีที่ 6 ปีการศึกษา 2546 หมายถึง แบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนที่สร้างตามหลักสูตรประถมศึกษา พ.ศ. 2521 (ปรับปรุง 2533) และดำเนินการสอบในปีการศึกษา 2546 การวิจัยครั้งนี้ใช้เฉพาะแบบทดสอบวิชาภาษาไทย

กลุ่มอ้างอิง (Reference Group) หมายถึง กลุ่มผู้สอบที่คาดว่าน่าจะได้เปรียบในการตอบข้อสอบเมื่อข้อสอบทำหน้าที่ต่างกัน โดยมีโอกาสในการตอบข้อสอบได้ถูกต้องมากกว่าผู้สอบกลุ่มเปรียบเทียบ การวิจัยครั้งนี้ กลุ่มอ้างอิง คือ เพศหญิง

กลุ่มเปรียบเทียบ (Focal Group) หมายถึง กลุ่มผู้สอบที่คาดว่าน่าจะเสียเปรียบในการตอบข้อสอบเมื่อข้อสอบทำหน้าที่ต่างกัน โดยมีโอกาสในการตอบข้อสอบได้ถูกต้องน้อยกว่าผู้สอบกลุ่มอ้างอิง การวิจัยครั้งนี้ กลุ่มเปรียบเทียบ คือ เพศชาย

วิธีซิปเทสท์ (SIBTEST) หมายถึง วิธีการในตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในกลุ่มทฤษฎีการตอบสนองข้อสอบ (Item Response Theory; IRT) ที่ พัฒนาโดย เชียลลี่และสเตาท์ (Shealy & Stout, 1993) ซึ่งในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบจะแบ่งแบบทดสอบออกเป็น 2 ชุดย่อย คือ 1) ชุดของแบบทดสอบที่มีความตรง (Valid Subtest) ใช้ในการจับคู่เปรียบเทียบระหว่างผู้สอบกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ 2) ชุดของแบบทดสอบที่ต้องการศึกษา (Studied Subtest) ใช้ในการคำนวณดัชนีการทำหน้าที่ต่างกันของข้อสอบ โดยคำนวณจากค่าเฉลี่ยสัดส่วนการตอบข้อสอบถูก ระหว่างผู้สอบกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ แล้วทดสอบนัยสำคัญด้วยสถิติ Z – Test ในการตัดสินว่าข้อสอบทำหน้าที่ต่างกัน

วิธีซิปเทสท์ปรับใหม่ (Modify SIBTEST) หมายถึง วิธีซิปเทสท์ที่ปรับปรุงขั้นตอนการวิเคราะห์โดยแบ่งกลุ่มผู้สอบออกเป็นสองกลุ่ม ตามระดับความสามารถ คือ กลุ่มผู้สอบที่มีความสามารถต่ำ และกลุ่มผู้สอบที่มีความสามารถสูง โดยใช้คะแนนเฉลี่ยของผู้สอบทั้งหมดเป็นเกณฑ์ในการแบ่งกลุ่ม แล้วคำนวณดัชนีการทำหน้าที่ต่างกันของข้อสอบแยกกันในแต่ละกลุ่ม

วิธีถดถอยโลจิสติก (Logistic Regression) หมายถึง วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในกลุ่มวิธีที่ไม่ใช้ทฤษฎีการตอบสนองข้อสอบที่พัฒนาโดย สวามินาธาน และโรเจอร์ (Swaminathan & Rogers, 1990 cited in Rogers & Swaminathan, 1993) ซึ่งคำนวณดัชนีการทำหน้าที่ต่างกันของข้อสอบจากผลการตอบข้อสอบถูกระหว่างผู้สอบกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบโดยใช้โมเดลการถดถอยโลจิสติก แล้วทดสอบนัยสำคัญด้วยสถิติ χ^2

การตัดข้อสอบออกจากแบบทดสอบ หมายถึง การไม่นำเอาข้อสอบ ข้อที่พบว่าทำหน้าที่ต่างกัน มาคิดคะแนนให้กับผู้สอบ

ความเที่ยง (Reliability) หมายถึง คุณภาพของแบบทดสอบที่ให้ผลการวัดมีความคงที่หรือคงเส้นคงวา ไม่ว่าจะวัดกี่ครั้งหรือภายใต้สภาวะการวัดแบบใดก็ให้ผลการวัดคงเดิม สำหรับการวิจัยในครั้งนี้หาค่าความเที่ยงแบบความสอดคล้องภายใน (Internal Consistency) โดยใช้วิธีการคำนวณค่าความเที่ยงแบบสัมประสิทธิ์แอลฟาของครอนบาค

ความตรงเชิงโครงสร้าง (Construct Validity) หมายถึง ความสามารถของแบบทดสอบในการอธิบายธรรมชาติหรือโครงสร้างของคุณลักษณะทางจิตวิทยาที่อยู่เบื้องหลังกลุ่มข้อสอบ หรือสิ่งที่แบบทดสอบมุ่งวัดเป็นสิ่งที่เราไม่สามารถสังเกตได้โดยตรง แต่เป็นสิ่งที่มีตามทฤษฎีทางจิตวิทยา การวิจัยในครั้งนี้ใช้วิธีวิเคราะห์องค์ประกอบเชิงยืนยันอันดับสอง (Second-Order Confirmatory Factor Analysis)

ความคงที่ของโครงสร้างองค์ประกอบ (Invariance of Factor Construct) หมายถึง ค่าพารามิเตอร์ของแบบทดสอบมีความคงที่ ไม่แปรเปลี่ยนไปตามกลุ่มประชากร (กลุ่มประชากรในการวิจัยจำแนกตามตัวแปรเพศ คือ กลุ่มเพศชาย กับ กลุ่มเพศหญิง) ในการวิจัยครั้งนี้ใช้วิธีวิเคราะห์องค์ประกอบเชิงยืนยันอันดับสอง แล้วเปรียบเทียบค่าพารามิเตอร์ที่ได้จากทั้งสองกลุ่ม ถ้าค่าพารามิเตอร์ที่ได้จากทั้งสองกลุ่มไม่แตกต่างกัน แสดงว่า แบบทดสอบมีความคงที่ของโครงสร้างองค์ประกอบ

สหสัมพันธ์อันดับของผู้สอบ (Spearman Rank Correlation) หมายถึง ค่าที่แสดงความสัมพันธ์ของลำดับที่การสอบจากแบบทดสอบสองชุด คือ ชุดที่มีข้อสอบทำหน้าที่ต่างกันอยู่ในแบบทดสบกับชุดที่ไม่มีข้อสอบทำหน้าที่ต่างกันอยู่ในแบบทดสอบ การวิจัยในครั้งนี้ใช้วิธีหาค่าสหสัมพันธ์อันดับของสเปียร์แมน (Spearman Rank: r_s)