บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

แบบทดสอบเป็นเครื่องมือหลักที่ใช้ในการวัดและประเมินผลการศึกษา เพื่อตรวจสอบว่าผู้เข้าสอบมีคุณลักษณะแฝงหรือความสามารถอยู่ในระดับใด ดังนั้น การสร้างและการตรวจสอบคุณภาพของแบบทดสอบจะต้องคำนึงถึงความตรง (Validity) เป็นสำคัญ ทั้งนี้เพราะว่าความตรงเป็นคุณสมบัติของแบบทดสอบที่แสดงว่า คะแนนจากแบบทดสอบสามารถสรุปอ้างอิงไปยังสิ่งที่วัดได้อย่างเหมาะสม มีความหมาย และเป็นประโยชน์ (เสรี ชัดแช้ม, 2544, หน้า 137) ถ้าผลการวัดได้ค่าที่ใกล้เคียงกับคุณลักษณะที่แท้จริงมากเพียงใด ก็ถือว่าการวัดมีความตรงมากขึ้นเพียงนั้น (ศิริชัย กาญจนวาสี, 2545, หน้า 103) นักวัดผลตรวจสอบหลักฐานที่แสดงความตรงของแบบทดสอบได้ 3 ด้าน คือ (1) ความตรงเชิงเนื้อหา (Content Validity) (2) ความตรงเชิงเกณฑ์สัมพันธ์ (Criterion Related Validity) และ (3) ความตรงเชิงโครงสร้าง (Construct Validity) ส่วนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (Differentail Item Functioning) เป็นการตรวจสอบคุณภาพของแบบทดสอบด้านความตรงเชิงโครงสร้าง (Mazor, Clauser, & Hambleton, 1992; Kim, Kim, & Cohen, 1994 อ้างถึงใน วลีมาศ แซ่อึ้ง, 2543, หน้า 1)

การทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) มีผู้ให้นิยามไว้หลายความหมายด้วยกัน แต่ที่ยอมรับกันอย่างกว้างขวางก็คือ ข้อสอบทำหน้าที่ต่างกันภายใต้เงื่อนไขผู้เข้าสอบมีความสามารถเท่ากัน แต่มาจากกลุ่มผู้เข้าสอบย่อยที่มีลักษณะต่างกัน มีความน่าจะเป็นในการตอบข้อสอบข้อนั้นไม่เท่ากัน (เสรี ชัดแช้ม, 2540, หน้า 42)

หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบใช้การเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มผู้เข้าสอบย่อย 2 กลุ่ม ที่มีความสามารถระดับเดียวกัน โดยผู้เข้าสอบกลุ่มหนึ่งเป็นตัวแทนกลุ่มหลักในประชากรเรียกว่า "กลุ่มอ้างอิง" (Reference Group: R) ซึ่งเป็นกลุ่มพื้นฐาน ส่วนผู้เข้าสอบอีกกลุ่มหนึ่งเป็นตัวแทนกลุ่มรองในประชากรเรียกว่า "กลุ่มเปรียบเทียบ" (Focal Group: F) ซึ่งตามปกติแล้วเป็นกลุ่มผู้เข้าสอบที่สนใจจะศึกษาการทำหน้าที่ต่างกันของข้อสอบ (Angoff, 1993) ถ้าข้อสอบทำหน้าที่ต่างกันแล้วโอกาสตอบข้อสอบถูกของผู้เข้าสอบทั้งสองกลุ่มจะไม่เท่ากัน โดยคาดว่าผู้เข้าสอบกลุ่มอ้างอิงจะได้เปรียบในการตอบข้อสอบ ส่วนผู้เข้าสอบกลุ่มเปรียบเทียบจะเสียเปรียบในการตอบข้อสอบ

เงื่อนไขสำคัญในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ ใช้เกณฑ์การจับคู่ (Matching Criteria) ที่นิยมใช้กันมี 2 แนวทางใหญ่ คือ แนวทางแรกเป็นการตรวจสอบโดยใช้เกณฑ์ภายนอก (External Criteria) โดยการนำคะแนนจากการตอบข้อสอบไปสัมพันธ์กับเกณฑ์ที่ได้จากภายนอกแบบทดสอบ เช่น เกรดเฉลี่ยสะสมหรือคะแนนจากชิ้นงานที่ทำเป็นต้น แต่จุดอ่อนของแนวทางนี้คือ เป็นการยากที่จะหาเกณฑ์ภายนอกที่มีความตรงและยุติธรรม ซึ่งถ้าเกณฑ์การจับคู่ขาดคุณสมบัติเรื่องนี้จะทำให้ผลการตรวจสอบเกิดความคลาดเคลื่อนได้ ส่วนแนวทางที่สอง เป็นการตรวจสอบโดยใช้เกณฑ์ภายใน (Internal Criteria) โดยการนำวิธีการทางสถิติมาตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เป็นการพิจารณาจากโครงสร้างภายในของแบบทดสอบเป็นหลัก (Rudner, Getson & Knight, 1980) ซึ่งในปัจจุบันเป็นแนวทางที่มีผู้นิยมศึกษากันมาก

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบมีหลายวิธี แฮมเบิลตัน และคณะ (Hambleton et al., 1993 อ้างถึงใน เสรี ชัดแช้ม, 2539, หน้า 1-4) ได้จำแนกออกเป็น 3 กลุ่มใหญ่ ๆ ดังนี้

1. กลุ่มวิธีที่ใช้ทฤษฎีการทดสอบแบบดั้งเดิม (Methods Using Classical Test Theory: CTT) วิธีในกลุ่มนี้พัฒนามาจากหลักการของทฤษฎีการทดสอบแบบดั้งเดิม วิธีการในกลุ่มนี้ได้แก่ การวิเคราะห์ความแปรปรวน (Analysis of Variance) วิธีสหสัมพันธ์ (Correlational Methods) (Green & Draper, 1972 cited in Scheuneman & Bleistein, 1989) วิธีแปลงค่าความยากของข้อสอบ (Transformed Item Difficulty Method, TID) ฯลฯ

2. กลุ่มวิธีที่ใช้ทฤษฎีการตอบสนองข้อสอบ (Methods Using Item Response Theory: IRT) วิธีในกลุ่มนี้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ตามกรอบแนวคิดของทฤษฎีการตอบสนองข้อสอบ วิธีการในกลุ่มนี้ได้แก่ วิธี Analysis of fit (Durovic, 1975, cited in Hambleton & Others, 1993) วิธี Difficulty shift (Wright, Mead & Draba, 1976, cited in Hambleton & Others, 1993) ฯลฯ

3. กลุ่มวิธีที่ใช้วิธีไค-สแควร์ (Methods Using Chi-Square Methods) วิธีการในกลุ่มนี้ได้แก่ วิธีแมนเทล-แฮนส์เซล (Mantel-Haenszal Method: MH) (Holland & Thayer, 1986, 1988) และวิธีถดถอยโลจิสติก (Logistic Regression Methods: LR) (Swaminathan & Rogers, 1990) ฯลฯ

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้พัฒนาขึ้นเรื่อย ๆ ปัจจุบันวิธีการวิเคราะห์องค์ประกอบจำกัด (Restricted Factor Analysis Methods: RFA) ซึ่งได้รับการพัฒนาขึ้นโดย ออร์ท (Oort, 1998) เป็นวิธีการใหม่ล่าสุดในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ซึ่งอาจจัดอยู่ในกลุ่มวิธีไค-สแควร์ และออร์ท (Oort, 1998, pp. 107-124) ได้ศึกษาเปรียบเทียบ

ประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธีการตอบสนองข้อสอบแบบ 1 พารามิเตอร์ กับวิธีการวิเคราะห์องค์ประกอบจำกัด โดยใช้สถานการณ์จำลองสร้างข้อสอบ 3 แบบ คือ ข้อสอบทำหน้าที่ต่างกันมาก ข้อสอบทำหน้าที่ต่างกันน้อย และข้อสอบทำหน้าที่ไม่ต่างกัน และกำหนดขนาดกลุ่มตัวอย่าง 2 ขนาด คือ กลุ่มตัวอย่างขนาดใหญ่ (2,000 คน) และขนาดเล็ก (200 คน) ผลปรากฏว่า วิธีการวิเคราะห์องค์ประกอบจำกัด มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ใกล้เคียงกับวิธีการตอบสนองข้อสอบ แบบ 1 พารามิเตอร์ แต่ในกรณีกลุ่มตัวอย่างขนาดเล็ก วิธีการวิเคราะห์องค์ประกอบจำกัด มีประสิทธิภาพดีกว่าวิธีการตอบสนองข้อสอบ แบบ 1 พารามิเตอร์ ในทุกเงื่อนไขของการทดสอบ นอกจากนี้ วิธีการวิเคราะห์องค์ประกอบจำกัดยังประหยัดเวลาและค่าใช้จ่ายในการวิเคราะห์ ส่วน นิคม กีรติวรางกูร (2542) ได้ศึกษาผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการวิเคราะห์องค์ประกอบจำกัด วิธีแมนเทล-แฮนส์เซล และวิธีการตอบสนองข้อสอบ แบบ 2 พารามิเตอร์ พบว่า โดยภาพรวม วิธีการวิเคราะห์องค์ประกอบจำกัด มีประสิทธิภาพในการตรวจสอบสูงที่สุด รองลงมาวิธีแมนเทล-แฮนส์เซล และวิธีการตอบสนองข้อสอบ แบบ 2 พารามิเตอร์ และวิธีการตอบสนองข้อสอบ แบบ 2 พารามิเตอร์ มีอัตราความคลาดเคลื่อนประเภทที่ 1 สูงกว่าวิธีแมนเทล-แฮนส์เซล และวิธีการวิเคราะห์องค์ประกอบจำกัด

สำหรับการศึกษาเกี่ยวกับปัจจัยที่ส่งผลต่อประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ปรากฏว่า ข้อค้นพบที่ได้มีความสอดคล้องกัน กล่าวคือ เมื่อขนาดกลุ่มตัวอย่างใหญ่ขึ้น อัตราการตรวจพบการทำหน้าที่ต่างกันของข้อสอบจะเพิ่มขึ้น (Roger & Swaminathan, 1993) แต่ในเรื่องขนาดของกลุ่มตัวอย่างยังไม่ชัดเจนว่า ต้องใช้กลุ่มตัวอย่างขนาดเท่าใดจึงจะเพียงพอ เมเซอร์ และคณะ (Mazor et al., 1992) กล่าวว่า ขนาดของกลุ่มตัวอย่างที่เหมาะสมสำหรับ วิธีแมนเทล-แฮนส์เซล ควรใช้ระหว่าง 100 และ 300 คน สำหรับกลุ่มใดกลุ่มหนึ่งหรือทั้งสองกลุ่ม (กลุ่มอ้างอิง และกลุ่มเปรียบเทียบ) นารายานัน และสวามินาทาน (Narayanan & Swaminathan, 1994) ได้เสนอแนะว่า โดยทั่วไปใช้กลุ่มตัวอย่างขนาด กลุ่มละ 300 คน ก็เพียงพอที่จะตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้อย่างมีประสิทธิภาพ และพบว่าอัตราการตรวจพบข้อสอบทำหน้าที่ต่างกันระหว่างวิธีแมนเทล-แฮนส์เซล กับวิธีซิปเทสท์ จะได้รับผลกระทบจากกลุ่มเปรียบเทียบที่มีขนาดเล็กมากกว่ากลุ่มอ้างอิงที่มีขนาดเล็ก นอกจากนี้ แฮมเบิลตัน และคณะ (Hambleton et al., 1993) ยังเสนอแนะว่า กลุ่มตัวอย่างที่ใช้ในการวิเคราะห์ด้วยวิธีแมนเทล-แฮนส์เซล ควรอยู่ระหว่าง 200 ถึง 1,000 คน แต่ในการวิเคราะห์บางเงื่อนไข การใช้กลุ่มตัวอย่าง 200 คน ในกลุ่มใดกลุ่มหนึ่งอาจจะไม่เพียงพอ ถ้าใช้กลุ่มตัวอย่างขนาดใหญ่จะทำให้ได้ผล

การตรวจสอบที่ดีกว่า จากที่กล่าวมาอาจกล่าวได้ว่าขนาดกลุ่มตัวอย่างน่าจะมีผลกระทบต่อประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

การศึกษาของ คิมและโคเฮน (Kim & Cohen, 1994) และ อัททาโรและมิลล์แซพ (Uttaro & Millsap, 1994) พบว่า ความยาวของแบบทดสอบมีผลกระทบต่ออัตราการตรวจพบข้อสอบทำหน้าที่ต่างกัน และจากงานวิจัยของ นารายานัน และสวามินาทาน (Narayanan & Swaminathan, 1996) พบว่า แบบทดสอบขนาด 40 ข้อ เป็นแบบทดสอบที่มีความยาวเพียงพอสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ แต่ โรเจอร์ และสวามินาทาน (Roger & Swaminathan, 1993) พบว่า ความยาวของแบบทดสอบไม่มีผลกระทบต่ออัตราการตรวจพบข้อสอบทำหน้าที่ต่างกัน ดังนั้นเพื่อเป็นการยืนยันข้อค้นพบให้มีความชัดเจนมากยิ่งขึ้น จึงควรศึกษาความยาวของแบบทดสอบว่ามีผลต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบหรือไม่

นอกจากนี้ยังมีการศึกษาปัจจัยอื่น ๆ ที่มีผลกระทบต่อประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ คือ เกณฑ์การจับคู่ (Matching Criteria) เป็นการหาเกณฑ์การจับคู่เปรียบเทียบที่เหมาะสมสำหรับตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ที่ผ่านมามีการศึกษาวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ใช้ทั้งเกณฑ์ภายนอก และเกณฑ์ภายใน เช่น นพมาศ พิพัฒนสุข (2541) ได้เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธีแมนเทล-แฮนส์เซล กับวิธีถดถอยโลจิสติก เมื่อใช้เกณฑ์ภายในเป็นเกณฑ์การจับคู่ พบว่า เมื่อใช้เกณฑ์การจับคู่คะแนนรวมทั้งฉบับ วิธีแมนเทล-แฮนส์เซล มีประสิทธิภาพมากกว่าวิธีถดถอยโลจิสติก และเมื่อใช้เกณฑ์การจับคู่คะแนนแบบทดสอบย่อย วิธีทั้งสองมีประสิทธิภาพไม่แตกต่างกัน สำหรับวิธีถดถอยโลจิสติกมีประสิทธิภาพมากกว่าวิธีแมนเทล-แฮนส์เซล เมื่อใช้เกณฑ์การจับคู่คะแนนจากแบบทดสอบย่อยหลายฉบับ ซึ่งสอดคล้องกับผลการศึกษาของเคลาเซอร์ และคณะ (Clauser et al., 1996, pp. 202-214) ที่ศึกษาเกณฑ์การจับคู่ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ โดยเปรียบเทียบผลการตรวจสอบข้อสอบทำหน้าที่ต่างกันโดยใช้เกณฑ์การจับคู่ 3 เกณฑ์ ได้แก่ คะแนนรวมทั้งฉบับ คะแนนแบบทดสอบย่อย และคะแนนจากแบบทดสอบย่อยหลายฉบับ ระหว่างวิธีแมนเทล-แฮนส์เซลกับวิธีถดถอยโลจิสติก ทั้งในข้อมูลจริง และข้อมูลจำลอง พบว่า ในแบบทดสอบพหุมิติ การใช้คะแนนรวมทั้งฉบับเป็นเกณฑ์การจับคู่ไม่เหมาะสม แต่เมื่อใช้คะแนนจากแบบทดสอบย่อยหลายฉบับเป็นเกณฑ์การจับคู่จะมีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ดีกว่า โดยพิจารณาจากจำนวนข้อสอบทำหน้าที่ต่างกันที่ตรวจพบน้อยที่สุด และเมื่อเปรียบเทียบกับเกณฑ์การจับคู่คะแนนรวมทั้งฉบับ และคะแนนแบบทดสอบย่อย การนำคะแนนรวมทั้งฉบับมาใช้เป็นเกณฑ์การจับคู่เปรียบเทียบนั้นไม่เหมาะสม เพราะจะทำให้อัตราความคลาดเคลื่อนประเภทที่ 1 สูงขึ้น

จากการศึกษางานวิจัยที่ผ่านมา มีประเด็นปัญหาที่ผู้วิจัยสนใจ ดังนี้

1. ยังไม่มีการเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธีการวิเคราะห์องค์ประกอบจำกัด กับวิธีถดถอยโลจิสติก ว่าวิธีใดมีประสิทธิภาพดีกว่ากัน

2. ปัจจัยที่ส่งผลกระทบต่อประสิทธิภาพการตรวจพบข้อสอบทำหน้าที่ต่างกัน คือ ขนาดของกลุ่มตัวอย่าง และความยาวของแบบทดสอบ ดังนั้น น่าจะศึกษาว่าขนาดของกลุ่มตัวอย่างและความยาวของแบบทดสอบ มีผลต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบอย่างไร

3. ภายใต้เงื่อนไขการใช้เกณฑ์ภายในเป็นเกณฑ์การจับคู่ ได้แก่ คะแนนรวมทั้งฉบับ และคะแนนแบบทดสอบย่อย วิธีการวิเคราะห์องค์ประกอบจำกัดกับวิธีถดถอยโลจิสติก มีประสิทธิภาพในการตรวจสอบต่างกันหรือไม่

การวิจัยครั้งนี้ผู้วิจัยตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ วิชาภาษาไทย ชั้นประถมศึกษาปีที่ 6 ซึ่งวัดองค์ประกอบด้านโครงสร้างความรู้ และด้านกระบวนการ (กระทรวงศึกษาธิการ, 2546) จึงเป็นแบบทดสอบพหุมิติ

สำหรับตัวแปรจำแนกกลุ่มประชากรที่เลือกศึกษา ได้แก่ ตัวแปรเพศ เนื่องจากพิจารณาเห็นว่าเป็นตัวแปรที่สามารถแบ่งได้โดยปราศจากความคลาดเคลื่อน (Millsap & Everson, 1993; Zieky, 1993 อ้างถึงใน เสรี ชัดแช้ม, 2539, หน้า 12) จึงเหมาะสมที่จะนำมาใช้ศึกษาในเชิงวิธีการกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยข้อมูลจริง ในงานวิจัยนี้เพศหญิงเป็นกลุ่มอ้างอิง เพราะแบบทดสอบพหุมิติวิชาภาษาไทยส่วนใหญ่ จะลำเอียงเข้าข้างเพศหญิง (สุพัฒน์ สุกมลสันต์, 2534; กาญจนา วัธนสุนทร, 2537)

จากประเด็นปัญหาดังกล่าวข้างต้น ผู้วิจัยจึงสนใจที่จะเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างวิธีการวิเคราะห์องค์ประกอบจำกัดกับวิธีถดถอยโลจิสติก ภายใต้เงื่อนไขขนาดของกลุ่มตัวอย่าง ความยาวของแบบทดสอบ และเกณฑ์การจับคู่ต่างกัน ข้อค้นพบที่ได้จะใช้เป็นแนวทางในการเลือกวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ว่าควรเลือกวิธีใด ใช้กลุ่มตัวอย่างขนาดเท่าใด แบบทดสอบยาวเท่าใด และเกณฑ์การจับคู่แบบใด จึงจะมีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ

วัตถุประสงค์ของการวิจัย

1. เพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างวิธีการวิเคราะห์องค์ประกอบจำกัดกับวิธีถดถอยโลจิสติก ภายใต้เงื่อนไข 18 เงื่อนไข (3×3×2) คือ ขนาดของกลุ่มตัวอย่าง 3 ขนาด (2,000 คน 1,000 คน และ 300 คน) ความยาวของแบบทดสอบ

3 ขนาด (40 ข้อ 30 ข้อ และ 20 ข้อ) และเกณฑ์การจับคู่ 2 เกณฑ์ (คะแนนรวมทั้งฉบับ และคะแนนแบบทดสอบย่อย)

2. เพื่อเปรียบเทียบประสิทธิภาพของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างวิธีการวิเคราะห์องค์ประกอบจำกัดกับวิธีถดถอยโลจิสติก ภายใต้เงื่อนไข 18 เงื่อนไข (3×3×2) คือ ขนาดของกลุ่มตัวอย่าง 3 ขนาด (2,000 คน 1,000 คน และ 300 คน) ความยาวของแบบทดสอบ 3 ขนาด (40 ข้อ 30 ข้อ และ 20 ข้อ) และเกณฑ์การจับคู่ 2 เกณฑ์ (คะแนนรวมทั้งฉบับ และคะแนนแบบทดสอบย่อย) เมื่อใช้ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยวิธีซิปเทสท์เป็นเกณฑ์สำหรับการเปรียบเทียบในเรื่อง อัตราความถูกต้อง และอัตราความคลาดเคลื่อน

สมมติฐานของการวิจัย

โรเจอร์ และสวามินาทาน (Roger & Swaminathan, 1993, pp. 105-116) ได้ศึกษาพบว่า วิธีแมนเทล-แฮนส์เซล มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปเท่าเทียมกับวิธีถดถอยโลจิสติก และวิธีถดถอยโลจิสติก มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปมากกว่าวิธีแมนเทล-แฮนส์เซล ออร์ท (Oort, 1998, pp. 107-124) ได้เปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธีการตอบสนองข้อตอบ แบบ 1 พารามิเตอร์ กับวิธีการวิเคราะห์องค์ประกอบจำกัด ในข้อสอบแบบให้คะแนนเป็น 2 ค่า (ตอบถูกได้ 1 และตอบผิดได้ 0) เมื่อกำหนดเงื่อนไขให้ข้อสอบทำหน้าที่ต่างกันมาก ปานกลาง และน้อย กลุ่มตัวอย่างมีขนาด 2 ขนาด คือ ขนาดใหญ่ 2,000 คน และขนาดเล็ก 200 คน ใช้แบบทดสอบเดียวกัน(40 ข้อ) พบว่า โดยภาพรวมการตรวจสอบด้วยวิธีการวิเคราะห์องค์ประกอบจำกัด ให้ผลการตรวจสอบที่ดีกว่าวิธีการตอบสนองข้อตอบ แบบ 1 พารามิเตอร์ สำหรับกลุ่มตัวอย่างขนาดเล็ก (200 คน) การตรวจสอบด้วยวิธีการวิเคราะห์องค์ประกอบจำกัด ให้ผลการตรวจสอบดีกว่า แต่เมื่อกลุ่มตัวอย่างมีขนาดใหญ่ ทั้งสองวิธีจะให้ผลดีขึ้น นอกจากนี้ผลการศึกษาของ นิคม กีรติวรางกูร (2542) ยังสนับสนุนว่า โดยภาพรวมวิธีการวิเคราะห์องค์ประกอบจำกัด มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบดีกว่าวิธีแมนเทล-แฮนส์เซล และวิธีการตอบสนองข้อสอบ แบบ 2 พารามิเตอร์ และวิธีการตอบสนองข้อสอบ แบบ 2 พารามิเตอร์ มีอัตราความคลาดเคลื่อนประเภทที่ 1 สูงกว่าวิธีแมนเทล-แฮนส์เซล และวิธีการวิเคราะห์องค์ประกอบจำกัด จึงทำให้ผู้วิจัยตั้งสมมติฐานของการวิจัยดังนี้

1. เมื่อกลุ่มตัวอย่างขนาดใหญ่ (2,000 คน) แบบทดสอบ (40 ข้อ 30 ข้อ และ 20 ข้อ) และเกณฑ์การจับคู่คะแนนรวมทั้งฉบับ วิธีการวิเคราะห์องค์ประกอบจำกัดมีประสิทธิภาพดีกว่าวิธีถดถอยโลจิสติก

2. เมื่อกลุ่มตัวอย่างขนาดใหญ่ (2,000 คน) แบบทดสอบ (40 ข้อ 30 ข้อ และ 20 ข้อ) และเกณฑ์การจับคู่คะแนนแบบทดสอบย่อย วิธีการวิเคราะห์องค์ประกอบจำกัดมีประสิทธิภาพดีกว่าวิธีถดถอยโลจิสติก

3. เมื่อกลุ่มตัวอย่างขนาดกลาง (1,000 คน) แบบทดสอบ (40 ข้อ 30 ข้อ และ 20 ข้อ) และเกณฑ์การจับคู่คะแนนรวมทั้งฉบับ วิธีการวิเคราะห์องค์ประกอบจำกัดมีประสิทธิภาพดีกว่าวิธีถดถอยโลจิสติก

4. เมื่อกลุ่มตัวอย่างขนาดกลาง (1,000 คน) แบบทดสอบ (40 ข้อ 30 ข้อ และ 20 ข้อ) และเกณฑ์การจับคู่คะแนนแบบทดสอบย่อย วิธีการวิเคราะห์องค์ประกอบจำกัดมีประสิทธิภาพดีกว่าวิธีถดถอยโลจิสติก

5. เมื่อกลุ่มตัวอย่างขนาดเล็ก (300 คน) แบบทดสอบ (40 ข้อ 30 ข้อ และ 20 ข้อ) และเกณฑ์การจับคู่คะแนนรวมทั้งฉบับ วิธีการวิเคราะห์องค์ประกอบจำกัดมีประสิทธิภาพดีกว่าวิธีถดถอยโลจิสติก

6. เมื่อกลุ่มตัวอย่างขนาดเล็ก (300 คน) แบบทดสอบ (40 ข้อ 30 ข้อ และ 20 ข้อ) และเกณฑ์การจับคู่คะแนนแบบทดสอบย่อย วิธีการวิเคราะห์องค์ประกอบจำกัดมีประสิทธิภาพดีกว่าวิธีถดถอยโลจิสติก

ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

งานวิจัยครั้งนี้ มุ่งเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างวิธีการวิเคราะห์องค์ประกอบจำกัดกับวิธีถดถอยโลจิสติก ภายใต้เงื่อนไข 18 เงื่อนไข (3×3×2) คือ ขนาดของกลุ่มตัวอย่าง 3 ขนาด (2,000 คน 1,000 คน และ 300 คน) ความยาวของแบบทดสอบ 3 ขนาด (40 ข้อ 30 ข้อ และ 20 ข้อ) และเกณฑ์การจับคู่ 2 เกณฑ์ (คะแนนรวมทั้งฉบับ และคะแนนแบบทดสอบย่อย) โดยศึกษากับข้อมูลจริง ผู้วิจัยจึงคาดว่าจะเป็นประโยชน์ ดังนี้

1. เป็นแนวทางในการเลือกวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ที่เหมาะสมต่อการปฏิบัติ โดยคำนึงถึงประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

2. เป็นแนวทางในการเลือกใช้ขนาดของกลุ่มตัวอย่าง ความยาวของแบบทดสอบ และเกณฑ์การจับคู่ ว่าควรใช้กลุ่มตัวอย่างเท่าใด ความยาวของแบบทดสอบขนาดเท่าใด และเกณฑ์การจับคู่แบบใด จึงจะทำให้มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติดีที่สุด

3. เป็นแนวทางในการศึกษาเกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติต่อไป

ขอบเขตของการวิจัย

1. ประชากรและกลุ่มตัวอย่าง

1.1 ประชากร เป็นนักเรียนชั้นประถมศึกษาปีที่ 6 สังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน กระทรวงศึกษาธิการ ปีการศึกษา 2546 ทั้งประเทศ ที่เข้าสอบวัดผลสัมฤทธิ์ทางการเรียนเพื่อประเมินคุณภาพการศึกษาระดับชาติ จำนวน 29,204 โรงเรียน และมีนักเรียนที่เข้าสอบวิชาภาษาไทย จำนวน 750,978 คน

1.2 กลุ่มตัวอย่าง เป็นนักเรียนชั้นประถมศึกษาปีที่ 6 สังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน ได้มาโดยการสุ่มแบบแบ่งชั้นไม่กำหนดสัดส่วน โดยใช้ระดับความสามารถ (คะแนนรวม) แบ่งเป็น 3 ระดับ คือ ดี พอใช้ และปรับปรุง ใช้นักเรียนเป็นหน่วยการสุ่ม สุ่มมาจำนวน 2,000 คน แบ่งเป็น นักเรียนชาย 1,000 คน และนักเรียนหญิง 1,000 คน

2. ตัวแปรที่ศึกษา

2.1 ตัวแปรต้น มี 4 ตัว ได้แก่

2.1.1 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ 2 วิธี ได้แก่

2.1.1.1 วิธีการวิเคราะห์องค์ประกอบจำกัด

2.1.1.2 วิธีถดถอยโลจิสติก

2.1.2 ขนาดของกลุ่มตัวอย่าง 3 ขนาด ได้แก่

2.1.2.1 กลุ่มตัวอย่างขนาดใหญ่ (2,000 คน)

2.1.2.2 กลุ่มตัวอย่างขนาดกลาง (1,000 คน)

2.1.2.3 กลุ่มตัวอย่างขนาดเล็ก (300 คน)

2.1.3 ความยาวของแบบทดสอบ 3 ขนาด ได้แก่

2.1.3.1 แบบทดสอบ 40 ข้อ

2.1.3.2 แบบทดสอบ 30 ข้อ

2.1.3.3 แบบทดสอบ 20 ข้อ

2.1.4 เกณฑ์การจับคู่ 2 เกณฑ์ ได้แก่

2.1.4.1 เกณฑ์การจับคู่คะแนนรวมทั้งฉบับ

2.1.4.2 เกณฑ์การจับคู่คะแนนแบบทดสอบย่อย

2.2 ตัวแปรตาม คือ ประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในเรื่อง

2.2.1 อัตราความถูกต้อง

2.2.2 อัตราความคลาดเคลื่อน

3. เกณฑ์ที่ใช้แสดงว่าข้อสอบทำหน้าที่ต่างกัน มีดังนี้

3.1 เกณฑ์สำหรับวิธีซิปเทสท์ ได้แก่ ค่า $\beta_u > 0$ และการทดสอบค่าสถิติ Z ที่ระดับนัยสำคัญ .05

3.2 เกณฑ์สำหรับวิธีการวิเคราะห์องค์ประกอบจำกัด ได้แก่ ค่าดัชนีดัดแปรโมเดล (MI) และค่าพารามิเตอร์ที่คาดหวัง (EPC) แตกต่างจาก 0 หรือไม่ ที่ระดับนัยสำคัญ .05

3.3 เกณฑ์สำหรับวิธีถดถอยโลจิสติก ได้แก่ ค่าอิทธิพลจากปฏิสัมพันธ์ระหว่างกลุ่มผู้เข้าสอบกับความสามารถของผู้เข้าสอบหรือค่าอิทธิพลจากกลุ่มผู้เข้าสอบ และการทดสอบค่าสถิติ χ^2 ที่ระดับนัยสำคัญ .05

ข้อตกลงเบื้องต้น

วิธีซิปเทสท์ เป็นการนำวิธีการทางสถิติมาใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยพิจารณาโครงสร้างภายในของแบบทดสอบพหุมิติ (Multidimensional Test) ใช้หลักการของทฤษฎีการตอบสนองข้อสอบแบบพหุมิติ จึงเป็นวิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ที่ให้ผลการตรวจสอบที่เชื่อถือได้ (Shealy & Stout, 1993) ดังนั้น ผู้วิจัยจึงเลือกวิธีซิปเทสท์เป็นเกณฑ์สำหรับเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธีการวิเคราะห์องค์ประกอบจำกัดกับวิธีถดถอยโลจิสติก ในการศึกษาครั้งนี้ผู้วิจัยวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยใช้โปรแกรม SIBTEST ที่สั่งซื้อมาจาก Assessment Systems Corporation ในประเทศสหรัฐอเมริกา

นิยามศัพท์เฉพาะ

การทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) หมายถึง ข้อสอบที่ผู้เข้าสอบซึ่งมีความสามารถเท่ากันในเนื้อหาที่ต้องการวัด มีโอกาสตอบข้อสอบข้อนั้นได้ถูกต้อง

ไม่เท่ากัน เนื่องจากอยู่ในกลุ่มผู้เข้าสอบย่อยที่มีลักษณะต่างกันในที่นี้คือ กลุ่มผู้เข้าสอบเพศหญิงกับกลุ่มผู้เข้าสอบเพศชาย

แบบทดสอบพหุมิติ (Multidimensional Test) หมายถึง แบบทดสอบที่วัดคุณลักษณะเด่นตั้งแต่ 2 คุณลักษณะขึ้นไป ในที่นี้คือแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ วิชาภาษาไทย ชั้นประถมศึกษาปีที่ 6 ซึ่งแบบทดสอบวัดองค์ประกอบ 2 ด้าน ได้แก่ ด้านโครงสร้างความรู้ และด้านกระบวนการ มีลักษณะเป็นข้อสอบเลือกตอบ ชนิด 4 ตัวเลือก

วิธีซิปเทสท์ (SIBTEST) หมายถึง วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในกลุ่มวิธี IRT ที่พัฒนาโดย เชียลี และสเตาท์ (Shealy & Stout, 1993) ซึ่งคำนวณดัชนีการทำหน้าที่ต่างกันของข้อสอบจากค่าเฉลี่ยสัดส่วนการตอบข้อสอบถูกระหว่างผู้เข้าสอบกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบในชุดแบบทดสอบที่ศึกษา (Studied Subtest) โดยใช้คะแนนการจับคู่เปรียบเทียบระหว่างกลุ่มผู้เข้าสอบจากชุดแบบทดสอบที่มีความตรง (Valid Subtest) แล้วทดสอบนัยสำคัญด้วยค่าสถิติ Z ในการศึกษาครั้งนี้ผู้วิจัยวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยใช้โปรแกรม SIBTEST

วิธีการวิเคราะห์องค์ประกอบจำกัด (Restricted Factor Analysis: RFA) หมายถึง วิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในกลุ่มวิธีไค-สแควร์ ที่พัฒนาโดย ออร์ท (Oort, 1998) เป็นวิธีที่ใช้คะแนนรวมของการสอบแทนความสามารถของผู้เข้าสอบ เกณฑ์การพิจารณาการทำหน้าที่ต่างกันของข้อสอบ ได้แก่ ค่าดัชนีดัดแปรโมเดล (Modification Indices: MI) และค่าพารามิเตอร์ที่คาดหวัง (Expected Parameter Change: EPC) แตกต่างจาก 0 หรือไม่ ที่ระดับนัยสำคัย .05 ในการศึกษาครั้งนี้ผู้วิจัยวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยใช้โปรแกรม LISREL 8.50

วิธีถดถอยโลจิสติก (Logistic Regression: LR) หมายถึง วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในกลุ่มวิธีไค-สแควร์ ที่พัฒนาโดย สวามินาธาน และโรเจอร์ (Swaminathan & Rogers, 1990) ซึ่งคำนวณดัชนีการทำหน้าที่ต่างกันของข้อสอบจากผลการตอบข้อสอบถูกระหว่างผู้เข้าสอบกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ เกณฑ์การพิจารณาการทำหน้าที่ต่างกันของข้อสอบ ได้แก่ ค่าอิทธิพลจากปฏิสัมพันธ์ระหว่างกลุ่มผู้เข้าสอบกับความสามารถของผู้เข้าสอบหรือค่าอิทธิพลจากกลุ่มผู้เข้าสอบ และการทดสอบค่าสถิติ χ^2 ที่ระดับนัยสำคัญ .05 ในการศึกษาครั้งนี้ผู้วิจัยวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยใช้โปรแกรม SPSS

กลุ่มอ้างอิง (Reference Group: R) หมายถึง กลุ่มผู้เข้าสอบที่คาดว่าจะได้ประโยชน์จากการตอบข้อสอบทำหน้าที่ต่างกัน เป็นกลุ่มที่มีความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องสูงกว่าผู้เข้าสอบอีกกลุ่มหนึ่งทั้ง ๆ ที่มีความสามารถเท่ากัน ในการวิจัยนี้กลุ่มอ้างอิง คือ เพศหญิง

กลุ่มเปรียบเทียบ (Focal Group: F) หมายถึง กลุ่มผู้เข้าสอบที่คาดว่าจะเสียประโยชน์จากการตอบข้อสอบทำหน้าที่ต่างกัน เป็นกลุ่มที่มีความน่าจะเป็นในการตอบข้อสอบได้ถูกต้องต่ำกว่าผู้เข้าสอบอีกกลุ่มหนึ่งทั้ง ๆ ที่มีความสามารถเท่ากัน ในการวิจัยนี้กลุ่มเปรียบเทียบ คือ เพศชาย

ขนาดกลุ่มตัวอย่าง (Sample Size) หมายถึง จำนวนผู้เข้าสอบในกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่ใช้ในการศึกษา มี 3 ขนาด คือ ขนาดใหญ่ (2,000 คน) ขนาดกลาง (1,000 คน) และขนาดเล็ก (300 คน)

ความยาวของแบบทดสอบ (Test Length) หมายถึง จำนวนข้อสอบในแบบทดสอบที่ใช้ในการศึกษา มี 3 ขนาด คือ 40 ข้อ 30 ข้อ และ20 ข้อ

เกณฑ์การจับคู่ (Matching Criteria) หมายถึง คะแนนที่ใช้แทนความสามารถที่แท้จริงของผู้เข้าสอบสองกลุ่ม ซึ่งใช้ในการจับคู่เปรียบเทียบกลุ่มผู้เข้าสอบย่อย เพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ มี 2 เกณฑ์ คือ คะแนนรวมทั้งฉบับ และคะแนนแบบทดสอบย่อย

เกณฑ์การจับคู่คะแนนรวมทั้งฉบับ (Total Matching Criteria) หมายถึง คะแนนที่ใช้แทนความสามารถที่แท้จริงของผู้เข้าสอบ ซึ่งเป็นผลรวมของคะแนนที่ได้จากผลการสอบแบบทดสอบวิชาภาษาไทยทั้งฉบับ ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบครั้งนี้ใช้คะแนนรวมของผลการสอบทั้งฉบับ เป็นเกณฑ์จับคู่ระหว่างกลุ่มผู้เข้าสอบเพศหญิงกับกลุ่มผู้เข้าสอบเพศชาย

เกณฑ์การจับคู่คะแนนแบบทดสอบย่อย (Subtest Matching Criteria) หมายถึง คะแนนที่ใช้แทนความสามารถที่แท้จริงของผู้เข้าสอบ ซึ่งเป็นผลรวมของคะแนนที่ได้จากผลการสอบแบบทดสอบวิชาภาษาไทย ของแต่ละองค์ประกอบของการวัดแบบทดสอบย่อย ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบครั้งนี้ ใช้คะแนนรวมของผลการตอบข้อสอบ 2 องค์ประกอบ คือ ด้านโครงสร้างความรู้ และด้านกระบวนการ เป็นเกณฑ์จับคู่ระหว่างกลุ่มผู้เข้าสอบเพศหญิงกับกลุ่มผู้เข้าสอบเพศชาย โดยแยกวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบทีละองค์ประกอบ

ประสิทธิภาพของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (Efficiency of Detecting DIF Items) หมายถึง ความถูกต้องของการระบุข้อสอบทำหน้าที่ต่างกัน จากการตรวจสอบโดยวิธีการวิเคราะห์องค์ประกอบจำกัด และวิธีถดถอยโลจิสติก ซึ่งพิจารณาได้จากอัตราความถูกต้องและอัตราความคลาดเคลื่อน เมื่อเปรียบเทียบกับผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีซิบเทสท์

อัตราความถูกต้อง (Correct Rates) หมายถึง สัดส่วนของจำนวนข้อสอบ ที่ตรวจสอบพบว่าถูกต้องตรงกับวิธีซิบเทสท์ มี 2 ประเภท คือ อัตราความถูกต้องประเภทที่ 1 และอัตราความถูกต้องประเภทที่ 2

อัตราความถูกต้องประเภทที่ 1 (Type I Correct Rate) หมายถึง สัดส่วนของจำนวนข้อสอบที่ตรวจพบว่าทำหน้าที่ไม่ต่างกันได้ถูกต้องตรงกับวิธีซิปเทสท์ ซึ่งคำนวณจากสัดส่วนของจำนวนข้อสอบทำหน้าที่ไม่ต่างกันถูกต้องต่อจำนวนข้อสอบทำหน้าที่ไม่ต่างกันทั้งหมดในแบบทดสอบ

อัตราความถูกต้องประเภทที่ 2 (Type II Correct Rate) หมายถึง สัดส่วนของจำนวนข้อสอบที่ตรวจพบว่าทำหน้าที่ต่างกันได้ถูกต้องตรงกับวิธีซิปเทสท์ ซึ่งคำนวณจากสัดส่วนของจำนวนข้อสอบทำหน้าที่ต่างกันถูกต้องต่อจำนวนข้อสอบทำหน้าที่ต่างกันทั้งหมดในแบบทดสอบ

อัตราความคลาดเคลื่อน (Error Rates) หมายถึง สัดส่วนของจำนวนข้อสอบที่ระบุผิดพลาด มี 2 ประเภท คือ ความคลาดเคลื่อนประเภทที่ 1 และความคลาดเคลื่อนประเภทที่ 2

ความคลาดเคลื่อนประเภทที่ 1 (Type I Error) หมายถึง การระบุผิดพลาดว่าข้อสอบทำหน้าที่ต่างกัน (False Positive) ทั้ง ๆ ที่ในความเป็นจริงแล้วข้อสอบทำหน้าที่ไม่ต่างกัน ซึ่งในการวิจัยนี้คำนวณได้จากสัดส่วนของจำนวนข้อสอบที่ระบุผิดพลาดว่าทำหน้าที่ต่างกันต่อจำนวนข้อสอบที่ทำหน้าที่ไม่ต่างกันทั้งหมดในแบบทดสอบ

ความคลาดเคลื่อนประเภทที่ 2 (Type II Error) หมายถึง การระบุผิดพลาดว่าข้อสอบทำหน้าที่ไม่ต่างกัน (False Negative) ทั้ง ๆ ที่ในความเป็นจริงแล้วข้อสอบทำหน้าที่ต่างกัน ซึ่งในการวิจัยนี้คำนวณได้จากสัดส่วนของจำนวนข้อสอบที่ระบุผิดพลาดว่าทำหน้าที่ไม่ต่างกันต่อจำนวนข้อสอบที่ทำหน้าที่ต่างกันทั้งหมดในแบบทดสอบ