บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

การจัดการศึกษาของชาติทุกระดับ หัวใจสำคัญก็คือ การวัดและประเมินผลการศึกษา เพื่อตรวจสอบว่าผู้เรียนมีความรู้หรือคุณลักษณะที่ต้องการวัดอยู่ในระดับใด ซึ่งผลที่ได้จากการวัดนั้นมีความสำคัญต่อการพัฒนาคุณภาพของการศึกษา การทดสอบจึงเป็นวิธีการวัดผลการศึกษาวิธีหนึ่งที่นิยมใช้กันมากที่สุด ซึ่งเครื่องมือที่ใช้ในการทดสอบที่สำคัญ ก็คือแบบทดสอบต่าง ๆ ดังนั้นในการสร้างและการตรวจสอบคุณภาพของแบบทดสอบ จะต้องคำนึงถึงคุณภาพด้านความตรงเป็นสำคัญ ซึ่งในการตรวจสอบคุณภาพด้านความตรงของแบบทดสอบที่นิยมใช้มี 3 ประเภทหลัก คือ ความตรงตามเนื้อหา (Content Validity) ความตรงตามเกณฑ์ (Criterion Validity) และความตรงเชิงโครงสร้าง (Construct Validity) (เสรี ชัดแช้ม, 2544, หน้า 137 – 139) ส่วนการทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) และการทำหน้าที่ต่างกันของหมวดข้อสอบ (Differential Bundle Functioning: DBF) ก็เป็นอีกคุณลักษณะหนึ่ง ที่สำคัญมากของการตรวจสอบคุณภาพด้านความตรง ซึ่งเป็นการตรวจสอบในประเด็นของความไม่ยุติธรรมของข้อสอบ (Item Unfairness) โดยทั่วไปแล้วในแบบทดสอบมาตรฐานวัดผลสัมฤทธิ์ทางการเรียนถ้ามีสัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 10 ถึง 15 ถือว่าไม่ผิดปกติ แต่ถ้ามีสัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันร้อยละ 20 ถือว่าเป็นเรื่องผิดพลาดอย่างมาก (Clauser, 1993 อ้างถึงในวลีมาศ แซ่อึ้ง, 2543, หน้า 1)

การทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) และการทำหน้าที่ต่างกันของหมวดข้อสอบ (Differential Bundle Functioning: DBF) เป็นการเปรียบเทียบผลการตอบข้อสอบระหว่างผู้เข้าสอบ 2 กลุ่ม กลุ่มแรก เรียกว่า กลุ่มเปรียบเทียบ (Focal Group หรือ กลุ่ม F) เป็นกลุ่มที่ผู้วิจัยสนใจศึกษาและคาดว่าจะเป็นกลุ่มที่เสียเปรียบในการตอบข้อสอบ กล่าวคือ มีโอกาสตอบข้อสอบถูกได้น้อยกว่าผู้เข้าสอบกลุ่มอ้างอิง และกลุ่มที่สอง เรียกว่า กลุ่มอ้างอิง (Reference Group หรือกลุ่ม R) เป็นกลุ่มที่คาดว่าจะได้เปรียบจากการตอบข้อสอบ กล่าวคือ มีโอกาสในการตอบข้อสอบถูกได้มากกว่าผู้เข้าสอบกลุ่มเปรียบเทียบ เนื่องจากคุณลักษณะเฉพาะของบุคคลกับเนื้อหาของข้อสอบนั้น ตัวอย่างเช่น ข้อสอบวัดความคิดเชิงตรรกศาสตร์ที่มีบริบทเกี่ยวกับการเล่นฟุตบอล อาจทำให้เพศชายซึ่งเป็นกลุ่มอ้างอิงได้รับประโยชน์มากกว่าเพศหญิงซึ่งเป็นกลุ่มเปรียบเทียบ เนื่องจากเพศชายมีความคุ้นเคยและความรู้เกี่ยวกับฟุตบอลมากกว่าเพศหญิง (Shealy & Stout, 1993)

การศึกษาเกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) ยังคงได้รับความสนใจจากนักวิจัยเป็นอย่างมาก ซึ่งจากการศึกษางานวิจัยที่ผ่านมา มีผู้ศึกษาค้นคว้าและเสนอวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบหลายวิธี และวิธีที่รู้จักกันโดยทั่วไป เช่น วิธีทดสอบค่าไคสแคว์ของ ลอร์ด (Lord's χ^2 Test) วิธีดีเอฟไอที (DFIT) และวิธีซิบเทสท์ (SIBTEST) เป็นต้น ซึ่งเป็นวิธีที่อยู่บนพื้นฐานทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) ส่วนวิธีที่ไม่อยู่บนพื้นฐานทฤษฎีการตอบสนองข้อสอบ เช่น วิธีแมนเทล – แฮนส์เซล (Mantel – Haenszel) วิธีการวิเคราะห์องค์ประกอบจำกัด (Restricted Factor Analysis) และวิธีถดถอยโลจิสติก (Logistic Regression) เป็นต้น ซึ่งวิธีเหล่านี้จะตรวจสอบการทำหน้าที่ต่างกันในระดับข้อสอบ แต่จะมี 2 วิธีที่สามารถตรวจสอบการทำหน้าที่ต่างกันในระดับหมวดข้อสอบ คือ วิธีดีเอฟไอที (DFIT) และวิธีซิบเทสท์ (SIBTEST) ซึ่งเป็นวิธีที่อยู่บนพื้นฐานทฤษฎีการตอบสนองข้อสอบเหมือนกัน แต่ความแตกต่างทั้ง 2 วิธีนี้ คือ วิธีดีเอฟไอที (DFIT) ใช้สถิติทดสอบแบบพาราเมตริก (Parametric) แต่วิธีซิบเทสท์ (SIBTEST) ใช้สถิติทดสอบแบบนันพาราเมตริก (Nonparametric) สำหรับการศึกษาเกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบ ยังพบไม่มากนัก

ในปี ค.ศ. 1993 เชียลีและสเตาท์ (Shealy & Stout, 1993) ได้พัฒนาวิธีซิบเทสท์ (Simulataneous Item Bias Test: SIBTEST) ที่สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) การทำหน้าที่ต่างกันของหมวดข้อสอบ (Differential Bundle Functioning: DBF) และการทำหน้าที่ต่างกันของแบบทดสอบ (Differential Test Functioning: DTF) วิธีนี้สามารถวิเคราะห์ได้ทั้งแบบทดสอบเอกมิติ (Unidimensional Test) และแบบทดสอบพหุมิติ (Multidimensional Tests) วิธีซิบเทสท์ใช้สถิติทดสอบแบบนันพาราเมตริก (Nonparametric) พัฒนาบนพื้นฐานของทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) ชนิดพหุมิติ แต่ไม่ต้องใช้ฟังก์ชันการตอบสนองข้อสอบ หรือการประมาณค่าความสามารถแฝง (Latent Ability) และวิธีซิบเทสท์ได้รับการออกแบบมาสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป (Uniform DIF) ดังนั้นจึงไม่มีความไวในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป (Nonuniform DIF) (Li & Stout, 1996) จุดเด่นของวิธีซิบเทสท์ คือ คำนวณได้ง่าย ไม่ซับซ้อน ประหยัดค่าใช้จ่ายและไม่จำเป็นต้องใช้กลุ่มตัวอย่างที่มีขนาดใหญ่ อีกทั้งใช้สถิติทดสอบนัยสำคัญ (Narayanan & Swaminathan, 1996) นอกจากนี้ยังสามารถนำไปประยุกต์ใช้กับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนแบบพหุวิภาค (Polytomous DIF) (Chang, Mazzeo, & Roussos, 1996; Narayanan & Swaminathan, 1996)

ปัจจุบันวิธีซิปเทสท์เป็นวิธีที่นิยมศึกษากันมาก ในระยะแรกวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ใช้ตรวจสอบข้อมูลการตอบข้อสอบจากแบบทดสอบเอกมิติ โดยตรวจสอบการทำหน้าที่ต่างกันของข้อสอบครั้งละ 1 ข้อ ต่อมา ดักกลัส รุสโซ และสเตาท์ (Douglas, Roussos, & Stout, 1996, pp. 465 – 484) ได้ศึกษาการตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบ จากข้อมูลการตอบข้อสอบพหุมิติ มีการให้คะแนนแบบสองค่าโดยใช้ข้อมูลเชิงประจักษ์ วิเคราะห์ข้อมูลด้วยโปรแกรม SIBTEST พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบครั้งละหลาย ๆ ข้อ มีประสิทธิภาพสูงกว่าการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบครั้งละ 1 ข้อ นอกจากนั้น แนนดาคูมาร์ (Nandakumar, 1993, p. 294) เสนอแนะว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบหลายข้อพร้อมกัน ทำให้สามารถศึกษาการทำหน้าที่ต่างกันของข้อสอบแบบขยายผล (DIF Amplification) และการทำหน้าที่ต่างกันของข้อสอบแบบหักล้างกัน (DIF Cancellation) ดังนั้นบางครั้งตรวจสอบไม่พบการทำหน้าที่ต่างกันของข้อสอบรายข้อ แต่เมื่อพิจารณาเป็นหมวดข้อสอบ (Bundle of Items) อาจพบการทำหน้าที่ต่างกันของหมวดข้อสอบได้ หรือในทำนองกลับกัน เมื่อตรวจสอบพบการทำหน้าที่ต่างกันของข้อสอบรายข้อต่อกลุ่มอ้างอิง และตรวจสอบพบข้อสอบข้ออื่น ๆ ทำหน้าที่ต่างกันต่อกลุ่มเปรียบเทียบ เมื่อพิจารณาพร้อม ๆ กันทั้งหมวดข้อสอบอาจไม่พบการทำหน้าที่ต่างกันของหมวดข้อสอบก็ได้

สำหรับการศึกษาเกี่ยวกับปัจจัยที่ส่งผลต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบและหมวดข้อสอบ เกี่ยวกับขนาดของกลุ่มตัวอย่าง สเตาท์ ลี แนนดาคูมาร์ และโบลท (Stout, Li, Nandakumar, & Bolt, 1997, pp. 195 – 213) ได้เสนอแนะว่า การวิเคราะห์โดยใช้โปรแกรม SIBTEST กลุ่มตัวอย่างขนาดเล็กสุดที่ควรใช้ คือ ขนาด 100 คน เมเซอร์ เคลาเซอร์ และแฮมเบิลตัน (Mazor, Clauser, & Hambleton, 1992, pp. 443 – 451) กล่าวว่า ขนาดของกลุ่มตัวอย่างที่เหมาะสมสำหรับวิธีแมนเทล – แฮนส์เซล ควรใช้ระหว่าง 100 คน และ 300 คน สำหรับกลุ่มใดกลุ่มหนึ่งหรือทั้งสองกลุ่ม (กลุ่มอ้างอิงและกลุ่มเปรียบเทียบ) แฮมเบิลตัน เคลาเซอร์ เมเซอร์ และโจนส์ (Hambleton, Clauser, Mazor, & Jones, 1993) เสนอแนะว่า กลุ่มตัวอย่างที่ใช้ในการวิเคราะห์ด้วยวิธีแมนเทล – แฮนส์เซล ควรอยู่ระหว่าง 200 คน ถึง 1,000 คน แต่ในการวิเคราะห์บางเงื่อนไข การใช้กลุ่มตัวอย่าง 200 คน ในกลุ่มใดกลุ่มหนึ่งอาจจะไม่เพียงพอ ถ้าใช้กลุ่มตัวอย่างขนาดใหญ่จะทำให้ได้ผลการตรวจสอบที่ดีกว่า นารายานัน และสวามินาทาน (Narayanan & Swaminathan, 1994) ได้เสนอแนะว่า โดยทั่วไปใช้กลุ่มตัวอย่างขนาดกลุ่มละ 300 คน ก็เพียงพอที่จะตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้อย่างมีประสิทธิภาพ นอกจากนี้ กาญจนา วัธนสุนทร (2538) ศึกษาเพื่อพัฒนาเกณฑ์ตัดสินข้อสอบลำเอียงทางเพศ พบว่า วิธีซิปเทสท์ และวิธีแมนเทล – แฮนส์เซล ควรใช้กลุ่มตัวอย่าง 600 คนขึ้นไป ส่วน จิตติมา วรรณศรี (2539) ได้เปรียบเทียบประสิทธิภาพของวิธีแมนเทล – แฮนส์เซล และวิธีซิปเทสท์

พบว่า เมื่อกลุ่มตัวอย่างมีขนาด 200 คน และ 600 คน ทั้งสองวิธีสามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ถูกต้องร้อยละ 50 แต่ถ้ากลุ่มตัวอย่างขนาด 1,000 คน สามารถตรวจสอบได้ถูกต้องร้อยละ 100 และ สิริรัตน์ วิภาสศิลป์ (2545) ศึกษาเปรียบเทียบวิธีซิปเทสท์ และดีเอฟไอที ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ หมวดข้อสอบ และแบบทดสอบ จากข้อมูลการตอบจากแบบทดสอบพหุมิติ พบว่า กลุ่มตัวอย่างขนาด 500 คน และ 1,000 คน ส่งผลต่อความถูกต้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ และหมวดข้อสอบ ด้วยวิธีซิปเทสท์สูงกว่ากลุ่มตัวอย่างขนาด 50 คน 100 คน และ 200 คน

จากผลการศึกษาข้างต้น จึงมีประเด็นปัญหาที่ผู้วิจัยสนใจจะศึกษา ดังนี้

1. ยังไม่พบการศึกษาเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างข้อสอบรายข้อกับหมวดข้อสอบ โดยใช้วิธีซิปเทสท์

2. ปัจจัยที่ส่งผลต่อการตรวจพบข้อสอบทำหน้าที่ต่างกัน คือ ขนาดของกลุ่มตัวอย่างที่มีขนาดต่างกัน คือ ขนาดเล็ก ขนาดกลาง และขนาดใหญ่ ส่งผลต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบอย่างไร

3. ร้อยละของข้อสอบที่ทำหน้าที่ต่างกัน ระหว่างข้อสอบรายข้อกับหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ คิดเป็นร้อยละเท่าไร

4. ผลกระทบในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบหลายข้อพร้อมกัน ส่งผลต่อการทำหน้าที่ต่างกันของหมวดข้อสอบอย่างไร

จากประเด็นปัญหาดังกล่าว ผู้วิจัยจึงสนใจการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ การเปรียบเทียบระหว่างข้อสอบรายข้อกับหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือขนาดเล็ก ขนาดกลาง และขนาดใหญ่ โดยใช้ผลการตอบข้อสอบในแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ วิชาภาษาไทย ชั้นประถมศึกษาปีที่ 6 ซึ่งเป็นแบบทดสอบที่แบ่งเป็นหมวดข้อสอบเป็น 2 หมวด คือ หมวดข้อสอบด้านโครงสร้างความรู้ และหมวดข้อสอบด้านกระบวนการ มีลักษณะเป็นข้อสอบหลายตัวเลือก ชนิด 4 ตัวเลือก (สำนักทดสอบทางการศึกษา, 2546 ก, 2546 ข) และใช้เพศหญิงเป็นกลุ่มอ้างอิง เพราะแบบทดสอบทางด้านภาษา คือ วิชาภาษาไทย ส่วนใหญ่จะลำเอียงเข้าข้างเพศหญิง (สุพัฒน์ สุกมลสันต์, 2534; กาญจนา วัธนสุนทร, 2538) ข้อค้นพบที่ได้ใช้เป็นแนวทางในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างข้อสอบรายข้อกับหมวดข้อสอบ และเป็นแนวทางในการเลือกขนาดของกลุ่มตัวอย่างที่ควรใช้ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบต่อไป

วัตถุประสงค์ของการวิจัย

1. เพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างข้อสอบรายข้อกับหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)

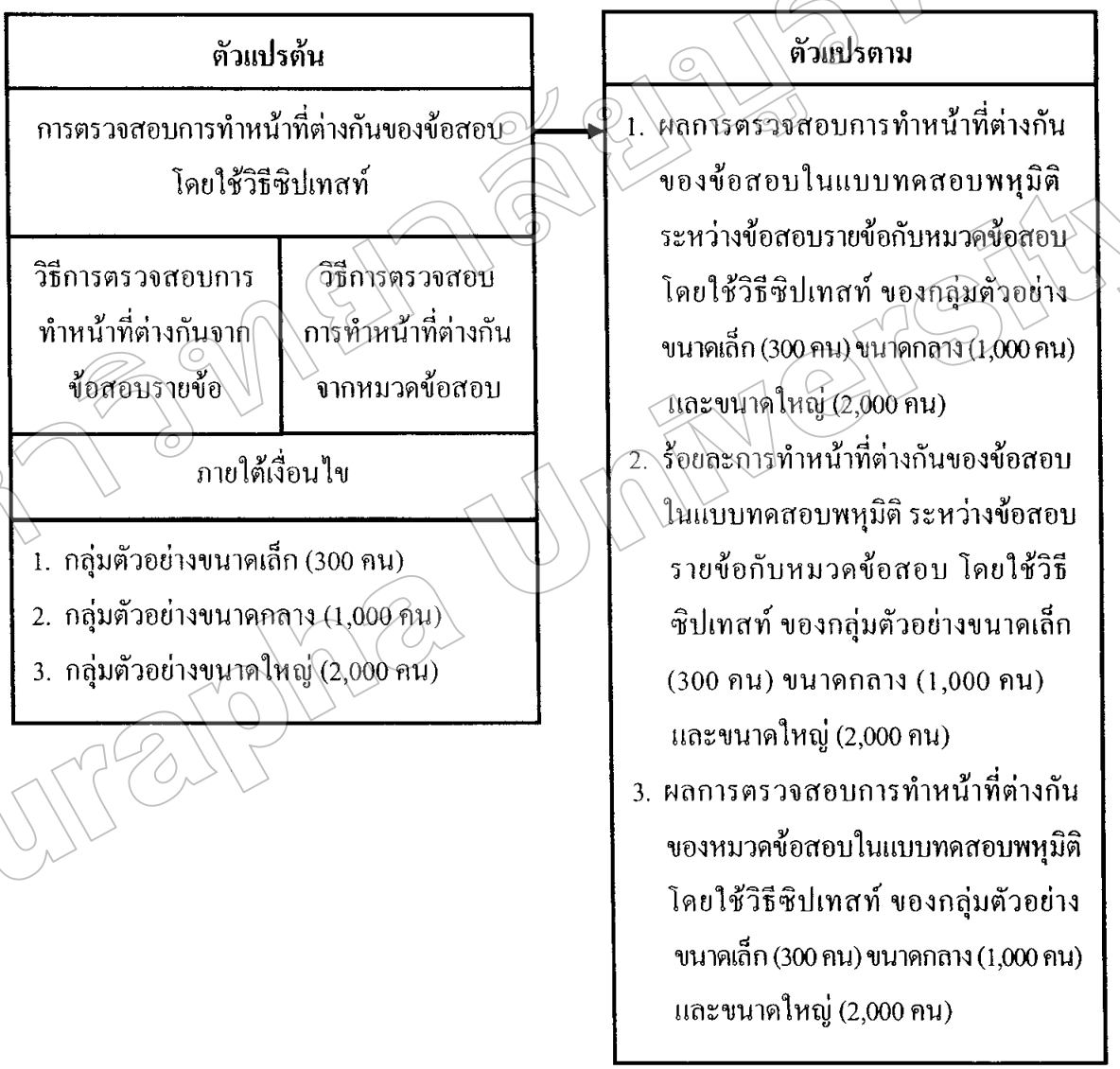
2. เพื่อเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างข้อสอบรายข้อกับหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)

3. เพื่อเปรียบเทียบร้อยละการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างข้อสอบรายข้อกับหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)

4. เพื่อตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบในแบบทดสอบพหุมิติ โดยใช้วิธีซิปเทสท์ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)

กรอบแนวคิดในการวิจัย

ประเด็นที่ผู้วิจัยจะศึกษา คือ การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ การเปรียบเทียบระหว่างข้อสอบรายข้อกับหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ ซึ่งอาศัยกรอบแนวคิดพื้นฐานของทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) ชนิดพหุมิติ โดยมีกรอบแนวคิดในการวิจัย ดังนี้



ภาพที่ 1 กรอบแนวคิดในการวิจัย

สมมติฐานของการวิจัย

จากการศึกษางานวิจัยที่ผ่านมา ยังไม่พบการศึกษาเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างข้อสอบรายข้อกับหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ แต่ได้ข้อค้นพบเกี่ยวกับปัจจัยที่ส่งผลต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ คือ ขนาดของกลุ่มตัวอย่าง นารายานัน และสวามินาทาน (Narayanan & Swaminathan, 1994) ได้เสนอแนะว่า โดยทั่วไปใช้กลุ่มตัวอย่างขนาดกลุ่มละ 300 คน ก็เพียงพอที่จะตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้อย่างมีประสิทธิภาพ กาญจนา วัธนสุนทร (2538) พบว่า วิธีซิปเทสท์ และวิธีแมนเทล – แฮนส์เซล ควรใช้กลุ่มตัวอย่าง 600 คนขึ้นไป ส่วน จิตติมา วรรณศรี (2539) พบว่า วิธีแมนเทล – แฮนส์เซล และวิธีซิปเทสท์ เมื่อใช้กลุ่มตัวอย่างขนาด 200 คน และ 600 คน ทั้งสองวิธีสามารถตรวจสอบได้ถูกต้อง ร้อยละ 50 แต่ถ้ากลุ่มตัวอย่างขนาด 1,000 คน สามารถตรวจสอบได้ถูกต้อง ร้อยละ 100 นอกจากนี้ ผลการศึกษาของ เมเซอร์ และคณะ (Mazor et al., 1992) ยังสนับสนุนว่า เมื่อใช้กลุ่มตัวอย่างขนาดใหญ่ (2,000 คน) ทำให้ตรวจพบข้อสอบทำหน้าที่ต่างกันได้ดีกว่าการใช้กลุ่มตัวอย่างขนาดเล็ก จากผลการศึกษาดังกล่าว จึงทำให้ผู้วิจัยตั้งสมมติฐานของการวิจัย ดังนี้

1. เมื่อกลุ่มตัวอย่างขนาดเล็ก (300 คน) ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างข้อสอบรายข้อกับหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ แตกต่างกัน

2. เมื่อกลุ่มตัวอย่างขนาดกลาง (1,000 คน) ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างข้อสอบรายข้อกับหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ แตกต่างกัน

3. เมื่อกลุ่มตัวอย่างขนาดใหญ่ (2,000 คน) ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างข้อสอบรายข้อกับหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ แตกต่างกัน

ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

การวิจัยครั้งนี้ มุ่งเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างข้อสอบรายข้อกับหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก ขนาดกลาง และขนาดใหญ่ โดยศึกษาจากข้อมูลจริง ซึ่งเป็นผลการตอบจากแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนชั้นประถมศึกษาปีที่ 6 วิชาภาษาไทย ปีการศึกษา 2546 ผู้วิจัยจึงคาดว่าจะเป็นประโยชน์ ดังนี้

1. ได้ข้อค้นพบในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ การเปรียบเทียบระหว่างข้อสอบรายข้อกับหมวดข้อสอบ โดยใช้วิธีซิปเทสท์

2. ได้ข้อค้นพบในการเลือกขนาดกลุ่มตัวอย่าง ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก ขนาดกลาง และขนาดใหญ่

3. ได้ข้อค้นพบในการตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบในแบบทดสอบพหุมิติ โดยใช้วิธีซิปเทสท์ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน

4. เป็นแนวทางในการศึกษาการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบรายข้อและหมวดข้อสอบในแบบทดสอบพหุมิติต่อไป

ขอบเขตของการวิจัย

1. กลุ่มตัวอย่าง

ใช้ข้อมูลทุติยภูมิ ที่เป็นผลการตอบข้อสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ ชั้นประถมศึกษาปีที่ 6 วิชาภาษาไทย ปีการศึกษา 2546 ของนักเรียนสังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน กระทรวงศึกษาธิการ ในเขตพื้นที่การศึกษานครศรีธรรมราช แบ่งเป็น 3 ระดับ คือ ระดับดี ระดับพอใช้ และระดับปรับปรุง ใช้นักเรียนเป็นหน่วยการสุ่ม สุ่มมาจำนวน 2,000 คน แบ่งเป็นนักเรียนชาย 1,000 คน และนักเรียนหญิง 1,000 คน

2. เนื้อหา

ศึกษาแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ ชั้นประถมศึกษาปีที่ 6 วิชาภาษาไทย ปีการศึกษา 2546

3. ตัวแปรที่ศึกษา

3.1 ตัวแปรต้น มี 2 ตัว ได้แก่

3.1.1 การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้วิธีซิปเทสท์ 2 วิธี ได้แก่

3.1.1.1 วิธีการตรวจสอบการทำหน้าที่ต่างกันจากข้อสอบรายข้อ

3.1.1.2 วิธีการตรวจสอบการทำหน้าที่ต่างกันจากหมวดข้อสอบ

3.1.2 ขนาดของกลุ่มตัวอย่าง 3 ขนาด ได้แก่

3.1.2.1 กลุ่มตัวอย่างขนาดเล็ก (300 คน)

3.1.2.2 กลุ่มตัวอย่างขนาดกลาง (1,000 คน)

3.1.2.3 กลุ่มตัวอย่างขนาดใหญ่ (2,000 คน)

3.2 ตัวแปรตาม มี 3 ตัว ได้แก่

3.2.1 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างข้อสอบรายข้อกับหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ ของกลุ่มตัวอย่างขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)

3.2.2 ร้อยละการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างข้อสอบรายข้อกับหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ ของกลุ่มตัวอย่างขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)

3.2.3 ผลการตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบในแบบทดสอบพหุมิติ โดยใช้วิธีซิปเทสท์ ของกลุ่มตัวอย่างขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)

4. เกณฑ์ที่ใช้แสดงการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างข้อสอบรายข้อกับหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ ได้แก่ ค่าดัชนี $\beta_{uni} > 0$ และการทดสอบนัยสำคัญทางสถิติที่ระดับ .05

ข้อตกลงเบื้องต้น

วิธีซิปเทสท์ เป็นการนำวิธีการทางสถิติมาใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ โดยพิจารณาโครงสร้างภายในของแบบทดสอบพหุมิติ (Multidimensional Tests) ใช้หลักการของทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) ชนิดพหุมิติ จึงเป็นวิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้ผลการตรวจสอบที่เชื่อถือได้ (Shealy & Stout, 1993) ดังนั้น ผู้วิจัยจึงเลือกวิธีซิปเทสท์ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ เพื่อเปรียบเทียบระหว่างข้อสอบรายข้อกับหมวดข้อสอบ

นิยามศัพท์เฉพาะ

1. การทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) หมายถึง ข้อสอบที่ผู้ตอบข้อสอบซึ่งมีความสามารถหรือคุณลักษณะที่ต้องการวัดเท่ากัน มีโอกาสตอบข้อสอบข้อนั้นได้ถูกต้องไม่เท่ากัน เนื่องจากอยู่ในกลุ่มผู้เข้าสอบย่อยที่มีลักษณะต่างกันในที่นี้คือ กลุ่มผู้ตอบข้อสอบเพศชาย กับกลุ่มผู้ตอบข้อสอบเพศหญิง

2. การทำหน้าที่ต่างกันของหมวดข้อสอบ (Differential Bundle Functioning: DBF) หมายถึง ข้อสอบในหมวดข้อสอบตามโครงสร้างของแบบทดสอบพหุมิติ ซึ่งผู้ตอบข้อสอบมีความสามารถหรือคุณลักษณะที่ต้องการวัดเท่ากัน มีโอกาสการตอบข้อสอบหมวดนั้นได้ถูกต้องไม่เท่ากัน เนื่องจากอยู่ในกลุ่มผู้เข้าสอบย่อยที่มีลักษณะต่างกันในที่นี้ คือ กลุ่มผู้ตอบข้อสอบเพศชาย กับกลุ่มผู้ตอบข้อสอบเพศหญิง

3. แบบทดสอบพหุมิติ (Multidimensional Tests) หมายถึง แบบทดสอบที่วัดคุณลักษณะเด่นตั้งแต่ 2 ลักษณะขึ้นไป ในที่นี้ คือ แบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ วิชาภาษาไทย ชั้นประถมศึกษาปีที่ 6 ซึ่งเป็นแบบทดสอบที่แบ่งเป็นหมวดข้อสอบตามโครงสร้างของแบบทดสอบพหุมิติออกเป็น 2 หมวด คือ หมวดที่ 1 ด้านโครงสร้างความรู้ และหมวดที่ 2 ด้านกระบวนการ มีลักษณะเป็นข้อสอบเลือกตอบ ชนิด 4 ตัวเลือก

4. วิธีซิปเทสท์ (SIBTEST) หมายถึง วิธีการทางสถิติมาใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ เพื่อเปรียบเทียบระหว่างข้อสอบรายข้อกับหมวดข้อสอบ ใช้สถิติทดสอบแบบนันพาราเมตริก (Nonparametric) พัฒนาบนพื้นฐานของทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) ชนิดพหุมิติ ที่พัฒนาโดย เชียลี และสเตาท์ (Shealy & Stout, 1993) ซึ่งจะเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ โดยแบ่งแบบทดสอบออกเป็น 2 ชุดย่อย คือ (1) ชุดแบบทดสอบที่มีความตรง (Valid Subtest) ใช้ในการจับคู่เปรียบเทียบระหว่างผู้สอบกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ (2) ชุดของแบบทดสอบที่ต้องการศึกษา (Studied Subtest) ใช้ในการคำนวณดัชนีการทำหน้าที่ต่างกันของข้อสอบ โดยคำนวณจากค่าเฉลี่ยสัดส่วนการตอบข้อสอบถูกระหว่างผู้สอบกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ แล้วทดสอบนัยสำคัญด้วยสถิติ Z-Test

5. วิธีการตรวจสอบการทำหน้าที่ต่างกันจากข้อสอบรายข้อ (Single Items) หมายถึง วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจากข้อสอบรายข้อในแบบทดสอบพหุมิติ ข้อสอบรายข้อที่ผู้ตอบข้อสอบซึ่งมีความสามารถหรือคุณลักษณะที่ต้องการวัดเท่ากัน มีโอกาสตอบข้อสอบข้อนั้นได้ถูกต้องไม่เท่ากัน เนื่องจากอยู่ในกลุ่มผู้เข้าสอบย่อยที่มีลักษณะต่างกันในที่นี้คือ กลุ่มผู้ตอบข้อสอบเพศชาย กับกลุ่มผู้ตอบข้อสอบเพศหญิง

6. วิธีการตรวจสอบการทำหน้าที่ต่างกันจากหมวดข้อสอบ (Bundle Items) หมายถึง วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจากหมวดข้อสอบในแบบทดสอบพหุมิติ ซึ่งผู้ตอบข้อสอบมีความสามารถหรือคุณลักษณะที่ต้องการวัดเท่ากัน มีโอกาสการตอบข้อสอบในหมวดข้อสอบนั้นได้ถูกต้องไม่เท่ากัน เนื่องจากอยู่ในกลุ่มผู้เข้าสอบย่อยที่มีลักษณะต่างกันในที่นี้คือ กลุ่มผู้ตอบข้อสอบเพศชาย กับกลุ่มผู้ตอบข้อสอบเพศหญิง

7. กลุ่มอ้างอิง (Reference Group: R) หมายถึง กลุ่มผู้ตอบข้อสอบที่คาดว่าจะได้ประโยชน์จากการทำหน้าที่ต่างกันของข้อสอบรายข้อกับหมวดข้อสอบ ในการวิจัยครั้งนี้ คือ กลุ่มเพศหญิง

8. กลุ่มเปรียบเทียบ (Focal Group: F) หมายถึง กลุ่มผู้ตอบข้อสอบที่คาดว่าจะเสียประโยชน์จากการทำหน้าที่ต่างกันของข้อสอบรายข้อกับหมวดข้อสอบ ในการวิจัยครั้งนี้ คือ กลุ่มเพศชาย

9. ขนาดของกลุ่มตัวอย่าง (Sample Size) หมายถึง จำนวนผู้ตอบข้อสอบในกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่ใช้ในการศึกษา มี 3 ขนาด คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)

10. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ระหว่างข้อสอบรายข้อกับหมวดข้อสอบ หมายถึง ข้อสอบที่ตรวจพบว่า ทำหน้าที่ต่างกันทั้งสองวิธี คือ วิธีการตรวจสอบการทำหน้าที่ต่างกันจากข้อสอบรายข้อ กับวิธีการตรวจสอบการทำหน้าที่ต่างกันจากหมวดข้อสอบ

11. ร้อยละของข้อสอบที่ทำหน้าที่ต่างกัน หมายถึง ค่าร้อยละของผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างข้อสอบรายข้อกับหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ของกลุ่มตัวอย่างขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)

12. ผลการตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบ หมายถึง หมวดข้อสอบที่ตรวจพบว่าทำหน้าที่ต่างกัน โดยใช้วิธีซิปเทสท์ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)