การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ: การเปรียบเทียบระหว่างรายข้อกับรายหมวดข้อสอบ โดยใช้วิธีซิปเทสท์

พีรญา สูงเนิน

เสรี ชัดแช้ม และ สมโภชน์ อเนกสุข

มหาวิทยาลัยบูรพา

บทคัดย่อ

การวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างการตรวจสอบเป็นรายข้อกับรายหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) กลุ่มตัวอย่างเป็นนักเรียนชั้นประถมศึกษาปีที่ 6 สังกัดเขตพื้นที่การศึกษานครศรีธรรมราช ปีการศึกษา 2546 ที่เข้าสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ จำนวน 2,000 คน ข้อมูลทุติยภูมิที่ใช้เป็นคะแนนจากแบบทดสอบวิชาภาษาไทย ชั้นประถมศึกษาปีที่ 6 จำนวน 40 ข้อ จำแนกเป็น 2 หมวดข้อสอบ คือ หมวดที่ 1 วัดด้านโครงสร้างความรู้ จำนวน 15 ข้อ และหมวดที่ 2 วัดด้านกระบวนการ จำนวน 25 ข้อ วิเคราะห์ค่าสถิติพื้นฐานโดยใช้โปรแกรม SPSS ตรวจสอบความตรงเชิงโครงสร้างด้วยการวิเคราะห์องค์ประกอบเชิงยืนยันอันดับสอง โดยใช้โปรแกรม LISREL 8.50 และตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยใช้โปรแกรม SIBTEST ผลการวิจัยปรากฏว่า

1. ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก ขนาดกลาง และขนาดใหญ่ ระหว่างการตรวจสอบการทำหน้าที่ต่างกันเป็นรายข้อกับรายหมวดข้อสอบ พบข้อสอบทำหน้าที่ต่างกันแตกต่างกัน

2. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นรายข้อ เมื่อกลุ่มตัวอย่างขนาดเล็ก พบข้อสอบทำหน้าที่ต่างกันจำนวน 4 ข้อ คิดเป็นร้อยละ 10 ขนาดกลางพบจำนวน 13 ข้อ คิดเป็นร้อยละ 32.5 และขนาดใหญ่พบจำนวน 15 ข้อ คิดเป็นร้อยละ 37.5

3. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นรายหมวดข้อสอบ เมื่อกลุ่มตัวอย่างขนาดเล็ก พบข้อสอบทำหน้าที่ต่างกันจำนวน 4 ข้อ คิดเป็นร้อยละ 10 ขนาดกลางพบจำนวน 8 ข้อ คิดเป็นร้อยละ 20 และขนาดใหญ่พบจำนวน 16 ข้อ คิดเป็นร้อยละ 40

4. การตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบ ชี้ให้เห็นว่า หมวดที่ 2 ภายใต้เงื่อนไขกลุ่มตัวอย่างขนาดกลาง มีนัยสำคัญทางสถิติที่ระดับ .05

ส่วนหนึ่งของวิทยานิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีการวัดทางการศึกษา มหาวิทยาลัยบูรพา อาจารย์ที่ปรึกษาหลัก รศ.ดร.เสรี ชัดแช้ม และอาจารย์ที่ปรึกษาร่วม ผศ.ดร.สมโภชน์ อเนกสุข

Detecting Differential Item Functioning in Multidimensional Tests: Comparing Single Items to Item Bundles Using SIBTEST

Peeraya Soongnern

Seree Chadcham and Sompoch Anegasukha

Burapha University, Thailand

Abstract

This study aimed to compare the detection of differential item functioning between single items and item bundles in multidimensional tests. The study applied SIBTEST in small, medium, and large samples, using datasets generated from 2,000 Nakornsritammarat sixth-grade students sitting for the National Achievement Test in the 2003 academic year. The secondary data was derived from the scores of the Grade 6 Thai Language achievement test which consists of 40 items, separated into two bundles. The first bundle contained 15 items and was structured to test knowledge, whereas the second bundle contained 25 items and was structured to assess process.

The results were as follows:

1. Among the three sample sizes, differences were found in DIF between single items and item bundles in the multidimensional tests.

2. The differential item functioning of the entire test showed that there were 4 DIF items (10%) in the small sample size, 13 DIF items (32.5%) in the medium sample size, and 15 DIF items (37.5%) in the large sample size.

3. The differential item functioning of individual bundles revealed that there were 4 DIF items (10%) in the small sample size, 8 DIF items (20%) in the medium sample size, and 16 DIF items (40%) in the large sample size.

4. The differential bundle functioning of the second bundle in the medium sample size was statistically significant at the .05 level.

Based on a master thesis in Educational Measurement Technology, Burapha University, under the supervision of Assoc. Prof. Seree Chadcham, Ph.D., and Assist. Prof. Sompoch Anegasukha, Ed.D.

ความนำ

การวัดผลการศึกษาเป็นการตรวจสอบผู้เรียนว่ามีความรู้หรือคุณลักษณะที่ต้องการวัดอยู่ในระดับใด ซึ่งผลที่ได้จากการวัดมีความสำคัญต่อการพัฒนาคุณภาพของการศึกษา การทดสอบเป็นวิธีการวัดผลการศึกษาวิธีหนึ่งที่นิยมใช้กันมาก เครื่องมือที่ใช้ในการทดสอบที่สำคัญก็คือแบบทดสอบชนิดต่าง ๆ ในการสร้างและการตรวจสอบคุณภาพของแบบทดสอบ ต้องคำนึงถึงคุณภาพด้านความตรงเป็นสำคัญ ซึ่งการตรวจสอบความตรงของแบบทดสอบที่นิยมกันมี 3 ประเภท คือ ความตรงตามเนื้อหา (Content Validity) ความตรงตามเกณฑ์ (Criterion Validity) และความตรงเชิงโครงสร้าง (Construct Validity) ส่วนการทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning) และการทำหน้าที่ต่างกันของหมวดข้อสอบ (Differential Bundle Functioning) เป็นการตรวจสอบในประเด็นของความไม่ยุติธรรมของข้อสอบ (Item Unfairness) ซึ่งเป็นอีกคุณลักษณะหนึ่งที่สำคัญของการตรวจสอบคุณภาพด้านความตรง โดยปกติแล้วในแบบทดสอบมาตรฐานวัดผลสัมฤทธิ์ทางการเรียน ถ้ามีสัดส่วนของข้อสอบทำหน้าที่ต่างกันร้อยละ 10 ถึง 15 ถือว่าไม่ผิดปกติ แต่ถ้ามีสัดส่วนของข้อสอบทำหน้าที่ต่างกันร้อยละ 20 ถือว่าเป็นเรื่องผิดพลาดอย่างมาก (Clauser, 1993)

การทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) และการทำหน้าที่ต่างกันของหมวดข้อสอบ (Differential Bundle Functioning: DBF) เป็นการเปรียบเทียบผลการตอบข้อสอบ ระหว่างผู้เข้าสอบ 2 กลุ่ม กลุ่มแรก เรียกว่า กลุ่มเปรียบเทียบ (Focal Group หรือ กลุ่ม F) เป็นกลุ่มที่สนใจศึกษาและคาดว่าจะเป็นกลุ่มที่เสียเปรียบในการตอบข้อสอบ กล่าวคือ มีโอกาสตอบข้อสอบถูกน้อยกว่าผู้เข้าสอบกลุ่มอ้างอิง ซึ่งเป็นกลุ่มที่สอง กลุ่มอ้างอิง (Reference Group หรือ กลุ่ม R) เป็นกลุ่มที่คาดว่าจะได้เปรียบจากการตอบข้อสอบ กล่าวคือ มีโอกาสตอบข้อสอบถูกมากกว่าผู้เข้าสอบกลุ่มเปรียบเทียบ เนื่องจากคุณลักษณะเฉพาะของบุคคลกับเนื้อหาของข้อสอบ ตัวอย่างเช่น ข้อสอบวัดความคิดเชิงตรรกศาสตร์ที่มีบริบทเกี่ยวกับการเล่นฟุตบอล อาจทำให้เพศชายซึ่งเป็นกลุ่มอ้างอิงได้รับประโยชน์มากกว่าเพศหญิงซึ่งเป็นกลุ่มเปรียบเทียบ เนื่องจากเพศชายมีความคุ้นเคยและมีความรู้เกี่ยวกับฟุตบอลมากกว่าเพศหญิง (Shealy & Stout, 1993)

การศึกษาเกี่ยวกับการทำหน้าที่ต่างกันของข้อสอบได้รับความสนใจอย่างมาก จากการศึกษางานวิจัยที่ผ่านมามีผู้ศึกษาค้นคว้าและเสนอวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบไว้หลายวิธี เช่น วิธีดีเอฟไอที (DFIT) และวิธีซิปเทสท์ (SIBTEST) เป็นต้น วิธีเหล่านี้ตั้งอยู่บนพื้นฐานของทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) ส่วนวิธีที่ไม่ตั้งอยู่บนพื้นฐานของทฤษฎีการตอบสนองข้อสอบ เช่น วิธีแมนเทล – แฮนส์เซล (Mantel – Haenszel) วิธีการวิเคราะห์องค์ประกอบจำกัด (Restricted Factor Analysis) และวิธีการถดถอยโลจิสติก (Logistic Regression) เป็นต้น วิธีที่กล่าวมาส่วนใหญ่จะตรวจสอบการทำหน้าที่ต่างกันในระดับข้อสอบ แต่มีวิธีที่สามารถตรวจสอบการทำหน้าที่ต่างกันในระดับหมวดข้อสอบ คือ วิธีดีเอฟไอที และวิธีซิปเทสท์ ซึ่งเป็นวิธีที่อยู่บนพื้นฐานของทฤษฎีการตอบสนองข้อสอบเหมือนกัน แต่ใช้สถิติทดสอบแตกต่างกัน คือ วิธีดีเอฟไอทีใช้สถิติทดสอบแบบพาราเมตริก (Parametric) ส่วนวิธีซิปเทสท์ใช้สถิติทดสอบแบบนันพาราเมตริก (Nonparametric) สำหรับการศึกษาเกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบ ยังพบไม่มากนัก

Shealy & Stout (1993) ได้พัฒนาวิธีซิปเทสท์ (Simulataneous Item Bias Test: SIBTEST) ที่สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ การทำหน้าที่ต่างกันของหมวดข้อสอบ และการทำหน้าที่ต่างกันของแบบทดสอบ (Differential Test Functioning: DTF) วิธีนี้สามารถวิเคราะห์ได้ทั้งแบบทดสอบเอกมิติ

(Unidimensional Test) และแบบทดสอบพหุมิติ (Multidimensional Tests) วิธีซิปเทสท์ใช้สถิติทดสอบแบบนันพาราเมตริก พัฒนาบนพื้นฐานของทฤษฎีการตอบสนองข้อสอบ ชนิดพหุมิติ แต่ไม่ต้องใช้ฟังก์ชันการตอบสนองข้อสอบ หรือการประมาณค่าความสามารถแฝง (Latent Ability) วิธีซิปเทสท์ได้รับการออกแบบให้ใช้กับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป (Uniform DIF) ดังนั้นจึงไม่มีความไวต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป (Nonuniform DIF) (Li & Stout, 1996) จุดเด่นของวิธีซิปเทสท์ คือ คำนวณได้ง่าย ไม่ซับซ้อน ประหยัดค่าใช้จ่ายและไม่จำเป็นต้องใช้กลุ่มตัวอย่างขนาดใหญ่ อีกทั้งใช้สถิติทดสอบนัยสำคัญ (Narayanan & Swaminathan, 1996) นอกจากนี้ยังนำไปใช้กับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนแบบพหุวิภาค (Polytomous DIF) ได้อีกด้วย (Chang, Mazzeo, & Roussos, 1996; Narayanan & Swaminathan, 1996)

ปัจจุบันนิยมใช้วิธีซิปเทสท์กันมาก ในระยะแรกใช้ตรวจสอบข้อมูลการตอบข้อสอบของแบบทดสอบเอกมิติ โดยตรวจสอบการทำหน้าที่ต่างกันของข้อสอบครั้งละ 1 ข้อ ต่อมา Douglas, Roussos, & Stout (1996, pp. 465 – 484) ได้นำไปใช้ตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบ จากข้อมูลการตอบข้อสอบของแบบทดสอบพหุมิติที่มีการให้คะแนนแบบสองค่า โดยศึกษากับข้อมูลเชิงประจักษ์ พบว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบครั้งละหลาย ๆ ข้อมีประสิทธิภาพสูงกว่าการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบครั้งละ 1 ข้อ นอกจากนั้น Nandakumar (1993: p. 294) ได้เสนอแนะว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบหลายข้อพร้อมกัน ทำให้สามารถศึกษาการทำหน้าที่ต่างกันของข้อสอบแบบขยายผล (DIF Amplification) และการทำหน้าที่ต่างกันของข้อสอบแบบหักล้างกัน (DIF Cancellation) ได้ ดังนั้นในบางครั้งการตรวจสอบเป็นรายข้อไม่พบข้อสอบทำหน้าที่ต่างกัน แต่เมื่อตรวจสอบเป็นรายหมวดข้อสอบ (Bundle of Items) อาจพบการทำหน้าที่ต่างกันของหมวดข้อสอบได้ หรือในทำนองกลับกัน เมื่อตรวจสอบเป็นรายข้อ พบข้อสอบบางข้อทำหน้าที่ต่างกันกับกลุ่มอ้างอิง และพบข้อสอบบางข้อทำหน้าที่ต่างกันกับกลุ่มเปรียบเทียบ แต่เมื่อพิจารณาพร้อม ๆ กันเป็นรายหมวดข้อสอบ อาจไม่พบการทำหน้าที่ต่างกันของหมวดข้อสอบ ก็ได้

การศึกษาเกี่ยวกับปัจจัยที่ส่งผลต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ และหมวดข้อสอบ ในเรื่องขนาดของกลุ่มตัวอย่าง Stout, Li, Nandakumar, & Bolt (1997: pp. 195 – 213) เสนอแนะว่า การวิเคราะห์โดยใช้โปรแกรม SIBTEST กลุ่มตัวอย่างขนาดเล็กสุดที่ควรใช้ คือ ขนาด 100 คน Narayanan & Swaminathan (1994) ได้เสนอแนะว่า โดยทั่วไปใช้กลุ่มตัวอย่างขนาดกลุ่มละ 300 คน ก็เพียงพอที่จะตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้อย่างมีประสิทธิภาพ นอกจากนี้ กาญจนา วัธนสุนทร (2538) ได้ศึกษาเพื่อพัฒนาเกณฑ์ตัดสินข้อสอบลำเอียงทางเพศ พบว่า วิธีซิปเทสท์ และวิธีแมนเทล – แฮนส์เซล ควรใช้กลุ่มตัวอย่าง 600 คนขึ้นไป ส่วนจิตติมา วรรณศรี (2539) ได้เปรียบเทียบประสิทธิภาพของวิธีแมนเทล – แ ฮนส์เซล และวิธีซิปเทสท์ พบว่า เมื่อกลุ่มตัวอย่างมีขนาด 200 คน และ 600 คน วิธีทั้งสองสามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ถูกต้องร้อยละ 50 แต่ถ้ากลุ่มตัวอย่างขนาด 1,000 คน สามารถตรวจสอบได้ถูกต้อง ร้อยละ 100 และ สิริรัตน์ วิภาสศิลป์ (2545) ได้เปรียบเทียบวิธีซิปเทสท์ กับวิธีดีเอฟไอที ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ หมวดข้อสอบ และแบบทดสอบ จากข้อมูลการตอบของแบบทดสอบพหุมิติ พบว่า กลุ่มตัวอย่างขนาด 500 คน และ 1,000 คน ส่งผลต่อความถูกต้องในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ และหมวดข้อสอบ ด้วยวิธีซิปเทสท์ สูงกว่ากลุ่มตัวอย่างขนาด 50 คน 100 คน และ 200 คน

จากผลการศึกษาข้างต้น จึงมีประเด็นที่น่าสนใจ ดังนี้

1. การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างเป็นรายข้อกับรายหมวดข้อสอบ โดยใช้วิธีซิปเทสท์

2. ปัจจัยที่ส่งผลต่อการตรวจพบข้อสอบทำหน้าที่ต่างกัน ได้แก่ ขนาดของกลุ่มตัวอย่างที่ต่างกัน คือ ขนาดเล็ก ขนาดกลาง และขนาดใหญ่

ผู้วิจัยจึงสนใจศึกษาเปรียบเทียบการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างเป็นรายข้อกับรายหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือขนาดเล็ก ขนาดกลาง และขนาดใหญ่ โดยใช้ผลการตอบข้อสอบในแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ วิชาภาษาไทย ชั้นประถมศึกษาปีที่ 6 ซึ่งเป็นแบบทดสอบที่จำแนกเป็นหมวดข้อสอบได้ 2 หมวด คือ หมวดข้อสอบด้านโครงสร้างความรู้ และหมวดข้อสอบด้านกระบวนการ มีลักษณะเป็นข้อสอบหลายตัวเลือกชนิด 4 ตัวเลือก (สำนักทดสอบทางการศึกษา, 2546 ก, 2546 ข) และใช้เพศหญิงเป็นกลุ่มอ้างอิง เพราะแบบทดสอบทางด้านภาษา เช่น วิชาภาษาไทย เป็นต้น ส่วนใหญ่จะลำเอียงเข้าข้างเพศหญิง (สุพัฒน์ สุกมลสันต์, 2534; กาญจนา วัธนสุนทร, 2538) ข้อค้นพบที่ได้จะนำไปใช้เป็นแนวทางในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ เป็นรายข้อและรายหมวดข้อสอบ และเป็นแนวทางในการเลือกขนาดของกลุ่มตัวอย่างที่เหมาะสมสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบต่อไป

วัตถุประสงค์การวิจัย

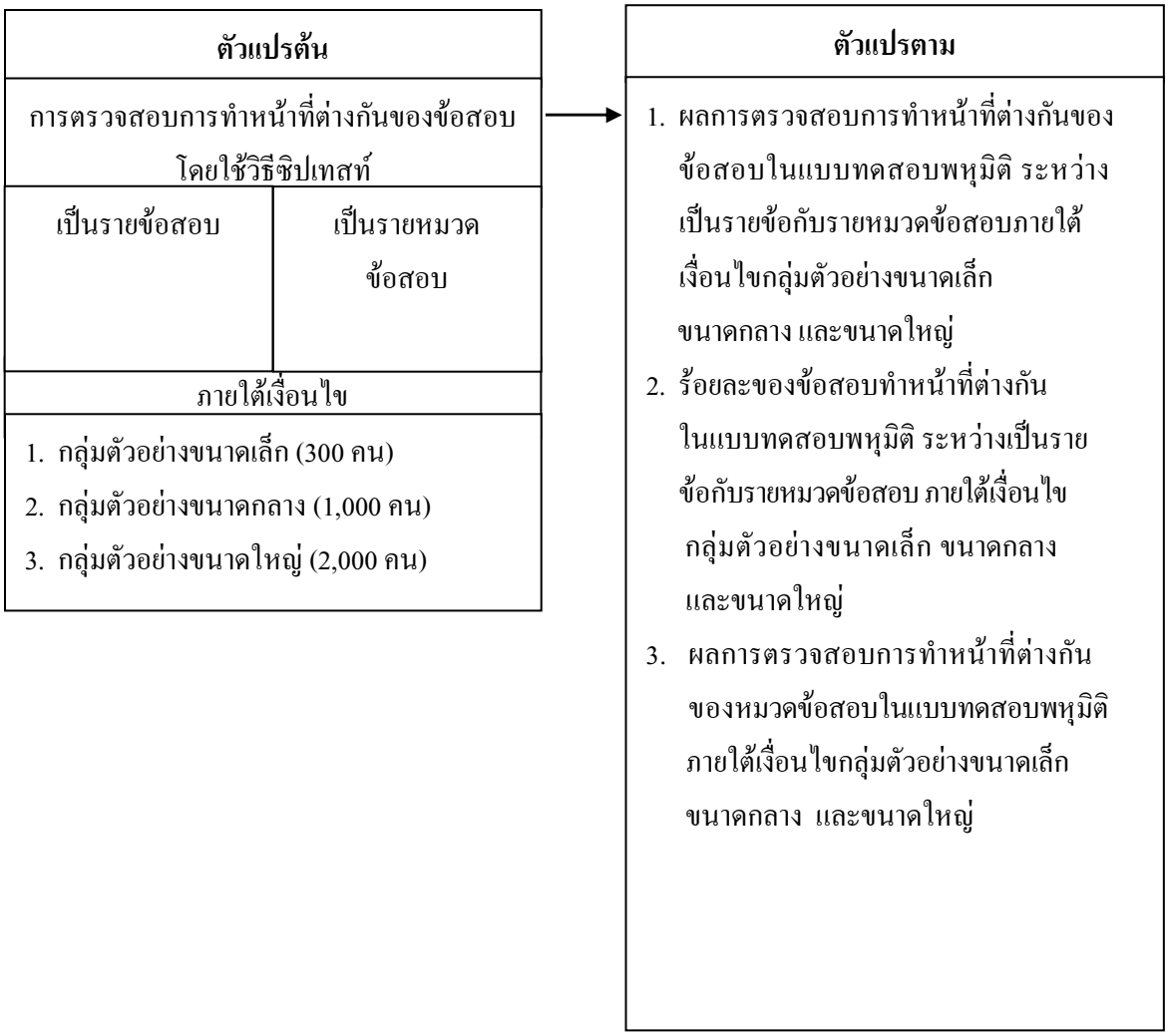
1. เพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ เป็นรายข้อและรายหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน)

2. เพื่อเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันในแบบทดสอบพหุมิติ ระหว่างเป็นรายข้อกับรายหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก ขนาดกลาง และขนาดใหญ่

3. เพื่อตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบในแบบทดสอบพหุมิติ โดยใช้วิธีซิปเทสท์ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก ขนาดกลาง และขนาดใหญ่

กรอบแนวคิดการวิจัย

การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างเป็นรายข้อกับรายหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก ขนาดกลาง และขนาดใหญ่ อาศัยแนวคิดของทฤษฎีการตอบสนองข้อสอบ (Item Response Theory: IRT) ชนิดพหุมิติ โดยมีกรอบแนวคิดในการวิจัย ดังนี้



ภาพที่ 1 กรอบแนวคิดการวิจัยการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ เป็นรายข้อและรายหมวดข้อสอบ

สมมติฐานการวิจัย

จากการศึกษางานวิจัยที่ผ่านมา ยังไม่พบการเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างเป็นรายข้อกับรายหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ เป็นการเฉพาะ แต่มีข้อค้นพบเกี่ยวกับปัจจัยที่ส่งผลต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ คือ ขนาดของกลุ่มตัวอย่าง Narayanan & Swaminathan (1994) ได้เสนอแนะว่า โดยทั่วไปใช้กลุ่มตัวอย่างขนาดกลุ่มละ 300 คน ก็เพียงพอ

ที่จะตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้อย่างมีประสิทธิภาพ กาญจนา วัธนสุนทร (2538) พบว่า วิธีซิปเทสท์ และวิธีแมนเทล-แฮนส์เซล ควรใช้กลุ่มตัวอย่าง 600 คนขึ้นไป ส่วน จิตติมา วรรณศรี (2539) พบว่า วิธีแมนเทล-แฮนส์เซล และวิธีซิปเทสท์ เมื่อใช้กลุ่มตัวอย่างขนาด 200 คน และ 600 คน ทั้งสองวิธีสามารถตรวจสอบได้ถูกต้องร้อยละ 50 แต่ถ้ากลุ่มตัวอย่างขนาด 1,000 คน สามารถตรวจสอบได้ถูกต้อง ร้อยละ 100 นอกจากนี้ผลการศึกษาของ Mazor, Clauser, & Hambleton (1992) ยังสนับสนุนว่า เมื่อใช้กลุ่มตัวอย่างขนาดใหญ่ (2,000 คน) ทำให้ตรวจพบข้อสอบทำหน้าที่ต่างกันได้ดีกว่าการใช้กลุ่มตัวอย่างขนาดเล็ก ดังนั้นผู้วิจัยจึงตั้งสมมติฐานการวิจัยไว้ดังนี้

1. เมื่อกลุ่มตัวอย่างขนาดเล็ก ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติระหว่างเป็นรายข้อกับรายหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ แตกต่างกัน

2. เมื่อกลุ่มตัวอย่างขนาดกลาง ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติระหว่างเป็นรายข้อกับรายหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ แตกต่างกัน

3. เมื่อกลุ่มตัวอย่างขนาดใหญ่ ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติระหว่างเป็นรายข้อกับรายหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ แตกต่างกัน

นิยามศัพท์เฉพาะ

1. การทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) หมายถึง ข้อสอบที่ผู้ตอบข้อสอบมีความสามารถหรือคุณลักษณะที่ต้องการวัดเท่ากัน มีโอกาสตอบข้อสอบข้อนั้นได้ถูกต้องไม่เท่ากัน เนื่องจากผู้ตอบอยู่ในกลุ่มผู้สอบย่อยที่มีลักษณะต่างกัน ในที่นี้คือ กลุ่มผู้ตอบเพศชายเป็นกลุ่มเปรียบเทียบ และกลุ่มผู้ตอบเพศหญิงเป็นกลุ่มอ้างอิง

2. การทำหน้าที่ต่างกันของหมวดข้อสอบ (Differential Bundle Functioning: DBF) หมายถึง ข้อสอบในหมวดข้อสอบตามโครงสร้างของแบบทดสอบพหุมิติ ซึ่งผู้ตอบข้อสอบมีความสามารถหรือคุณลักษณะที่ต้องการวัดเท่ากัน มีโอกาสตอบข้อสอบหมวดนั้นได้ถูกต้องไม่เท่ากัน เนื่องจากผู้ตอบอยู่ในกลุ่มผู้สอบย่อยที่มีลักษณะต่างกัน ในที่นี้ คือ กลุ่มผู้ตอบเพศชายเป็นกลุ่มเปรียบเทียบ และกลุ่มผู้ตอบเพศหญิงเป็นกลุ่มอ้างอิง

3. แบบทดสอบพหุมิติ (Multidimensional Tests) หมายถึง แบบทดสอบที่วัดคุณลักษณะเด่นตั้งแต่ 2 ลักษณะขึ้นไป ในที่นี้ คือ แบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ วิชาภาษาไทย ชั้นประถมศึกษาปีที่ 6 ซึ่งเป็นแบบทดสอบที่จำแนกเป็น 2 หมวด คือ หมวดที่ 1 วัดด้านโครงสร้างความรู้ และหมวดที่ 2 วัดด้านกระบวนการ มีลักษณะเป็นข้อสอบหลายตัวเลือก ชนิด 4 ตัวเลือก

วิธีดำเนินการวิจัย

กลุ่มตัวอย่าง

การวิจัยนี้ใช้ข้อมูลทุติยภูมิ ซึ่งเป็นผลการตอบข้อสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ ชั้นประถมศึกษาปีที่ 6 วิชาภาษาไทย ปีการศึกษา 2546 ของนักเรียนสังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน กระทรวงศึกษาธิการ ในเขตพื้นที่การศึกษานครศรีธรรมราช ได้มาโดยการสุ่มแบบแบ่งชั้น (Stratified Random Sampling) แบบจัดสรรเท่าเทียมกัน (Equal Allocation) โดยแบ่งชั้นตามระดับความสามารถ

เป็น 3 ระดับ คือ ดี พอใช้ และปรับปรุง ใช้นักเรียนเป็นหน่วยการสุ่ม สุ่มมาจำนวน 2,000 คน เป็นเพศชาย 1,000 คน และเพศหญิง 1,000 คน

เครื่องมือการวิจัย

เครื่องมือการวิจัยเป็นแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ วิชาภาษาไทย ชั้นประถมศึกษาปีที่ 6 ประกอบด้วย ข้อสอบแบบหลายตัวเลือก ชนิด 4 ตัวเลือก จำนวน 40 ข้อ คะแนนเต็ม 40 คะแนน สร้างโดยสำนักทดสอบทางการศึกษา กระทรวงศึกษาธิการ จำแนกเป็น 2 หมวดข้อสอบ ดังนี้

1. หมวดข้อสอบ หมวดที่ 1 วัดด้านโครงสร้างความรู้ จำนวน 15 ข้อ
2. หมวดข้อสอบ หมวดที่ 2 วัดด้านกระบวนการ จำนวน 25 ข้อ

การเก็บรวบรวมข้อมูล

ผู้วิจัยเก็บรวบรวมข้อมูลผลการตอบข้อสอบของนักเรียนที่เป็นกลุ่มตัวอย่าง ซึ่งเป็นข้อมูลทุติยภูมิจากสำนักทดสอบทางการศึกษา กระทรวงศึกษาธิการ โดยเจ้าหน้าที่สุ่มผลการตอบของนักเรียนแยกตามระดับความสามารถ และเพศ

การวิเคราะห์ข้อมูล

วิเคราะห์ค่าสถิติพื้นฐาน โดยใช้โปรแกรม SPSS ตรวจสอบความตรงเชิงโครงสร้างด้วยการวิเคราะห์องค์ประกอบเชิงยืนยันอันดับสอง โดยใช้โปรแกรม LISREL 8.50 และตรวจสอบการทำหน้าที่ต่างกันของข้อสอบและหมวดข้อสอบ โดยใช้โปรแกรม SIBTEST (Stout & Rousos, 1999)

ผลการวิจัย

ผู้วิจัยนำเสนอผลการวิจัยเป็น 2 ส่วน ได้แก่ ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ เป็นรายข้อและรายหมวดข้อสอบ และผลการตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ และเปรียบเทียบร้อยละของข้อสอบทำหน้าที่ต่างกันในแบบทดสอบพหุมิติ ระหว่างเป็นรายข้อกับรายหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ ดังนี้

1. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ เป็นรายข้อกับรายหมวดข้อสอบ และผลการตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ แสดงดังตารางที่ 1

ตารางที่ 1 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างเป็นรายข้อกับรายหมวดข้อสอบ และผลการตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบ โดยใช้วิธีซิปเทสท์

ขนาดกลุ่มตัวอย่าง	การตรวจ DIF เป็นรายข้อ		การตรวจ DIF เป็นรายหมวดข้อสอบ		การตรวจหมวดข้อสอบ (DBF)	
	NO DIF	DIF	NO DIF	DIF	หมวดที่ 1	หมวดที่ 2
ขนาดเล็ก (300 คน)	36 ข้อ	4 ข้อ (ข้อที่ 13, 23, 26, 28)	36 ข้อ	4 ข้อ (หมวด 1ข้อ 9, 13, หมวด 2 ข้อ 16, 23)	NO DBF	NO DBF
ขนาดกลาง (1,000 คน)	27 ข้อ	13 ข้อ (ข้อที่ 10, 13, 16, 21, 27, 28, 32, 34, 35, 36, 38, 39, 40)	32 ข้อ	8 ข้อ (หมวด1ข้อ 10, 13, หมวด 2 ข้อ 16, 21, 27, 28, 39, 40)	NO DBF	DBF
ขนาดใหญ่ (2,000 คน)	25 ข้อ	15 ข้อ (ข้อที่ 2, 9, 10, 13, 21, 23, 24, 27, 28, 34, 35, 36, 38, 39, 40)	24 ข้อ	16 ข้อ (หมวด 1 ข้อ 6, 9, 10, 13 หมวด 2 ข้อ 17, 21, 23, 24, 25, 27, 28, 34, 36, 38, 39, 40)	NO DBF	NO DBF

หมายเหตุ : DIF หมายถึง ข้อสอบทำหน้าที่ต่างกัน DBF หมายถึง หมวดข้อสอบทำหน้าที่ต่างกัน
NO DIF หมายถึง ข้อสอบทำหน้าที่ไม่ต่างกัน NO DBF หมายถึง หมวดข้อสอบทำหน้าที่ไม่ต่างกัน

จากตารางที่ 1 เมื่อกลุ่มตัวอย่างขนาดเล็ก ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นรายข้อ พบข้อสอบทำหน้าที่ต่างกัน 4 ข้อ ได้แก่ ข้อที่ 13, 23, 26, 28 ส่วนผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นรายหมวดข้อสอบ พบข้อสอบทำหน้าที่ต่างกัน 4 ข้อ ได้แก่ หมวดที่ 1 ข้อที่ 9, 13 และหมวดที่ 2 ข้อที่ 16, 23 สำหรับผลการตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบชี้ให้เห็นว่า ทั้งหมวดข้อสอบที่ 1 และที่ 2 ทำหน้าที่ไม่ต่างกัน (NO DBF)

เมื่อกลุ่มตัวอย่างขนาดกลาง ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นรายข้อ พบข้อสอบทำหน้าที่ต่างกัน 13 ข้อ ได้แก่ ข้อที่ 10, 13, 16, 21, 27, 28, 32, 34, 35, 36, 38, 39, 40 ส่วนผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นรายหมวดข้อสอบ พบข้อสอบทำหน้าที่ต่างกัน 8 ข้อ ได้แก่ หมวดที่ 1 ข้อที่ 10, 13 และหมวดที่ 2 ข้อที่ 16, 21, 27, 28, 39, 40 สำหรับผลการตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบชี้ให้เห็นว่า หมวดที่ 1 ทำหน้าที่ไม่ต่างกัน (NO DBF) แต่หมวดที่ 2 ทำหน้าที่ต่างกัน (DBF) มีนัยสำคัญทางสถิติที่ระดับ .05

เมื่อกลุ่มตัวอย่างขนาดใหญ่ ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นรายข้อ พบข้อสอบทำหน้าที่ต่างกัน 15 ข้อ ได้แก่ ข้อที่ 2, 9, 10, 13, 21, 23, 24, 27, 28, 34, 35, 36, 38, 39, 40 ส่วนผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นรายหมวดข้อสอบ พบข้อสอบทำหน้าที่ต่างกัน 16 ข้อ ได้แก่ หมวดที่ 1 ข้อที่ 6, 9, 10, 13 และหมวดที่ 2 ข้อที่ 17, 21, 23, 24, 25, 27, 28, 34, 36, 38, 39, 40 สำหรับผลการตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบชี้ให้เห็นว่า ทั้งหมวดข้อสอบที่ 1 และที่ 2 ทำหน้าที่ไม่ต่างกัน (NO DBF)

2. เปรียบเทียบร้อยละของข้อสอบทำหน้าที่ต่างกันในแบบทดสอบพหุมิติ ระหว่างการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นรายข้อ กับรายหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ แสดงดังตารางที่ 2

ตารางที่ 2 เปรียบเทียบร้อยละของข้อสอบทำหน้าที่ต่างกันในแบบทดสอบพหุมิติ ระหว่างการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นรายข้อ กับรายหมวดข้อสอบ โดยใช้วิธีซิปเทสท์

ขนาดกลุ่มตัวอย่าง	การตรวจ DIF เป็นรายข้อ			การตรวจ DIF เป็นรายหมวดข้อสอบ		
	NO DIF	**DIF**	**ร้อยละ (DIF)**	**NO DIF**	**DIF**	**ร้อยละ (DIF)**
ขนาดเล็ก (300 คน)	36 ข้อ	4 ข้อ	10	36 ข้อ	4 ข้อ	10
ขนาดกลาง (1,000 คน)	27 ข้อ	13 ข้อ	32.50	32 ข้อ	8 ข้อ	20
ขนาดใหญ่ (2,000 คน)	25 ข้อ	15 ข้อ	37.50	24 ข้อ	16 ข้อ	40

จากตารางที่ 2 เมื่อกลุ่มตัวอย่างขนาดเล็ก ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นรายข้อ พบข้อสอบทำหน้าที่ต่างกันจำนวน 4 ข้อ คิดเป็นร้อยละ 10 ส่วนผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นรายหมวดข้อสอบ พบข้อสอบทำหน้าที่ต่างกันจำนวน 4 ข้อ คิดเป็นร้อยละ 10

เมื่อกลุ่มตัวอย่างขนาดกลาง ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นรายข้อ พบข้อสอบทำหน้าที่ต่างกันจำนวน 13 ข้อ คิดเป็นร้อยละ 32.50 ส่วนผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นรายหมวดข้อสอบ พบข้อสอบทำหน้าที่ต่างกันจำนวน 8 ข้อ คิดเป็นร้อยละ 20

เมื่อกลุ่มตัวอย่างขนาดใหญ่ ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นรายข้อ พบข้อสอบทำหน้าที่ต่างกันจำนวน 15 ข้อ คิดเป็นร้อยละ 37.50 ส่วนผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นรายหมวดข้อสอบ พบข้อสอบทำหน้าที่ต่างกันจำนวน 16 ข้อ คิดเป็นร้อยละ 40

การอภิปรายผลการวิจัย

จากผลการวิจัยมีประเด็นสำคัญที่นำมาอภิปราย ได้ดังนี้

1. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างเป็นรายข้อกับรายหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ ภายใต้เงื่อนไขขนาดกลุ่มตัวอย่างต่างกัน คือ ขนาดเล็ก (300 คน) ขนาดกลาง (1,000 คน) และขนาดใหญ่ (2,000 คน) ปรากฏว่า สอดคล้องกับสมมติฐานการวิจัย 2 ขนาดกลุ่มตัวอย่างได้แก่ ขนาดกลางและขนาดใหญ่ กล่าวคือ จำนวนข้อสอบทำหน้าที่ต่างกันของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นรายข้อกับรายหมวดข้อสอบ แตกต่างกัน ส่วนในกลุ่มตัวอย่างขนาดเล็ก พบจำนวนข้อสอบทำหน้าที่ต่างกันเท่ากัน (4 ข้อ) แต่สลับข้อกัน ซึ่งในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้ง 2 วิธี จะสังเกตเห็นได้ว่าเมื่อกลุ่มตัวอย่างมีขนาดใหญ่ขึ้น จะทำให้ตรวจพบข้อสอบทำหน้าที่ต่างกันได้ดีกว่ากลุ่มตัวอย่างขนาดเล็ก สอดคล้องกับผลการศึกษาของ Mazor, Clauser,& Hambleton (1992) ที่ศึกษาพบว่า เมื่อใช้กลุ่มตัวอย่างขนาดใหญ่ขึ้น ทำให้ตรวจพบข้อสอบทำหน้าที่ต่างกันได้ดีกว่าการใช้กลุ่มตัวอย่างขนาดเล็ก

ผลการตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบ ปรากฏว่า ภายใต้เงื่อนไขกลุ่มตัวอย่างขนาดต่างกันไม่ส่งผลต่อการตรวจสอบพบการทำหน้าที่ต่างกันของหมวดข้อสอบ กล่าวคือ พบการทำหน้าที่ต่างกันของหมวดข้อสอบ (DBF) หมวดที่ 2 ด้านกระบวนการ ภายใต้เงื่อนไขกลุ่มตัวอย่างขนาดกลาง อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 เท่านั้น แต่เมื่อตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบ ภายใต้เงื่อนไขกลุ่มตัวอย่างขนาดเล็กและกลุ่มตัวอย่างขนาดใหญ่ กลับไม่พบการทำหน้าที่ต่างกันของหมวดข้อสอบ ทั้งนี้อาจจะเนื่องมาจากการตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบ เป็นการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบหลายข้อพร้อมกัน ซึ่งเป็นการทำหน้าที่ต่างกันของข้อสอบแบบขยายผลและหักล้างกัน (Steven, 2005) ดังนั้นในการตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบอาจจะพบหรือไม่พบการทำหน้าที่ต่างกันของหมวดข้อสอบก็ได้

2. เปรียบเทียบร้อยละของข้อสอบทำหน้าที่ต่างกันในแบบทดสอบพหุมิติ ระหว่างการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นรายข้อกับรายหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ ปรากฏว่า เมื่อกลุ่มตัวอย่างขนาดเล็ก ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นรายข้อ พบข้อสอบทำหน้าที่ต่างกันคิดเป็นร้อยละ 10 ส่วนเป็นรายหมวดข้อสอบ พบข้อสอบทำหน้าที่ต่างกันคิดเป็นร้อยละ 10 และพบข้อสอบทำหน้าที่ต่างกันตรงกันจำนวน 2 ข้อ เมื่อกลุ่มตัวอย่างขนาดกลาง ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นรายข้อ พบข้อสอบทำหน้าที่ต่างกัน คิดเป็นร้อยละ 32.50 ส่วนเป็นรายหมวดข้อสอบ พบข้อสอบทำหน้าที่ต่างกันคิดเป็นร้อยละ 20 และพบข้อสอบทำหน้าที่ต่างกันตรงกันจำนวน 8 ข้อ เมื่อกลุ่มตัวอย่างขนาดใหญ่ ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นรายข้อ พบข้อสอบทำหน้าที่ต่างกันคิดเป็นร้อยละ 37.50 ส่วนเป็นรายหมวดข้อสอบ พบข้อสอบทำหน้าที่ต่างกันคิดเป็นร้อยละ 40 และพบข้อสอบทำหน้าที่ต่างกันตรงกัน จำนวน 12 ข้อ เมื่อพิจารณาผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้ง 2 วิธี จะเห็นได้ว่าพบข้อสอบทำหน้าที่ต่างกันจำนวนไม่เท่ากันและไม่สอดคล้องกัน อย่างไรก็ตามภายใต้เงื่อนไขของกลุ่มตัวอย่างขนาดใหญ่ (2,000คน) พบข้อสอบทำหน้าที่ต่างกันตรงกัน จำนวน 12 ข้อ คิดเป็นร้อยละ 30 ซึ่งสอดคล้องกับผลการศึกษาของ ศุภวัฒน์ มะลิเผือก (2548) ที่ศึกษาข้อมูลชุดนี้แล้วพบว่า ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบของกลุ่มตัวอย่างขนาด 2,000 คน โดยใช้

วิธีการตรวจสอบสองวิธี คือวิธีการถดถอยโลจิสติก และวิธีซิปเทสท์ปรับใหม่ พบข้อสอบทำหน้าที่ต่างกัน จำนวน 12 ข้อ คิดเป็นร้อยละ 30

ข้อเสนอแนะ

การนำผลการวิจัยไปใช้

1. ในการวิจัยนี้ มุ่งศึกษาผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติเป็นรายข้อและรายหมวดข้อสอบ โดยใช้วิธีซิปเทสท์ เพื่อเป็นแนวทางในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ จากผลการวิจัย ปรากฏว่า ส่วนใหญ่ให้ผลแตกต่างกัน ส่วนผลการตรวจสอบการทำหน้าที่ต่างกันของหมวดข้อสอบ ปรากฏว่า หมวดที่ 2 ภายใต้เงื่อนไขกลุ่มตัวอย่างขนาดกลาง มีนัยสำคัญทางสถิติ ดังนั้นในทางปฏิบัติการนำผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบเป็นรายข้อสอบและรายหมวดข้อสอบในแบบทดสอบพหุมิติไปใช้ ควรใช้ทั้ง 2 วิธี เพื่อใช้ตัดสินใจเลือกข้อสอบไว้หรือตัดข้อสอบออก โดยใช้วิธีซิปเทสท์ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เนื่องจากเป็นวิธีที่มีประสิทธิภาพสูง และให้ผลการตรวจสอบที่เชื่อถือได้ รวมทั้งเป็นวิธีที่ใช้ได้สะดวก และประหยัดเวลาในการวิเคราะห์

2. ขนาดของกลุ่มตัวอย่างต่างกันมีผลต่อการตรวจพบข้อสอบทำหน้าที่ต่างกัน กล่าวคือ เมื่อขนาดของกลุ่มตัวอย่างใหญ่ขึ้น จะทำให้สามารถตรวจพบข้อสอบทำหน้าที่ต่างกันได้ดีกว่ากลุ่มตัวอย่างขนาดเล็ก แต่ขนาดของกลุ่มตัวอย่างต่างกัน ไม่ส่งผลต่อการตรวจพบการทำหน้าที่ต่างกันของหมวดข้อสอบ

การทำวิจัยต่อไป

1. ควรมีการเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติระหว่างเป็นรายข้อกับรายหมวดข้อสอบ ด้วยวิธีซิปเทสท์ กับวิธีอื่น ๆ เช่น วิธีดีเอฟไอที เป็นต้น โดยอาจใช้ผลการตอบข้อสอบวิชาอื่น ๆ เช่น วิชาภาษาอังกฤษ วิชาคณิตศาสตร์ เป็นต้น รวมทั้งศึกษาตัวแปรอื่นที่ส่งผลต่อประสิทธิผลของการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เช่น อัตราส่วนระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ เป็นต้น

2. ควรมีการศึกษาเพื่อปรับปรุงคุณภาพของแบบทดสอบพหุมิติ โดยการพิจารณาค่าความเที่ยง และความตรงเชิงโครงสร้างของแบบทดสอบ ภายหลังการตัดข้อสอบทำหน้าที่ต่างกันออกทีละข้อ จนกระทั่งในแบบทดสอบไม่มีข้อสอบทำหน้าที่ต่างกัน

3. ควรมีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างเป็นรายข้อกับรายหมวดข้อสอบ ในกรณีข้อสอบที่มีการให้คะแนนแบบหลายค่า โดยใช้วิธี Polytomous SIBTEST ภายใต้เงื่อนไขกลุ่มตัวอย่างขนาดต่างกัน

4. ควรมีการศึกษาการประมาณค่าความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างเป็นรายข้อกับรายหมวดข้อสอบ โดยอาจเปรียบเทียบระหว่างวิธีซิปเทสท์กับวิธีอื่น ๆ เช่น วิธีดีเอฟไอที เป็นต้น

เอกสารอ้างอิง

กาญจนา วัธนสุนทร. (2538). *การพัฒนาเกณฑ์ตัดสินข้อสอบลำเอียงทางเพศ.* วิทยานิพนธ์ปริญญาครุศาสตรดุษฎีบัณฑิต, สาขาวิชาการวัดและประเมินผลการศึกษา, บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย.

จิตติมา วรรณศรี. (2539). *การเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมนเทล-แฮนส์เซลกับวิธีซิปเทสท์ เมื่อความยาวของแบบสอบ ขนาดกลุ่มตัวอย่าง และอัตราส่วนกลุ่มอ้างอิงและกลุ่มเปรียบเทียบแตกต่างกัน.* วิทยานิพนธ์ปริญญาครุศาสตรมหาบัณฑิต, สาขาวิชาการวัดและประเมินผลการศึกษา, บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย.

ศุภวัฒน์ มะลิเผือก. (2549). อิทธิพลของการทำหน้าที่ต่างกันของข้อสอบ ที่ส่งผลต่อคุณภาพของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ ชั้นประถมศึกษาปีที่ 6. *วารสารวิจัยและวัดผลการศึกษา มหาวิทยาลัยบูรพา, 4*(1), 46-60.

สิริรัตน์ วิภาสศิลป์. (2545). *การเปรียบเทียบวิธีซิปเทส์และดีเอฟไอทีในการตรวจสอบการทำหน้าที่เบี่ยงเบนของข้อสอบ หมวดข้อสอบ และแบบทดสอบจากข้อมูลการตอบข้อสอบที่ใช้ความสามารถหลายมิติ.* วิทยานิพนธ์ปริญญาการศึกษาดุษฎีบัณฑิต, สาขาวิชาการวัดผลการศึกษา, บัณฑิตวิทยาลัย, มหาวิทยาลัยศรีนครินทรวิโรฒ.

สุพัฒน์ สุกมลสันต์. (2534). *การวิเคราะห์ความลำเอียงของข้อสอบภาษาอังกฤษเข้ามหาวิทยาลัย ปี 2531-2533.* กรุงเทพฯ: สถาบันภาษา จุฬาลงกรณ์มหาวิทยาลัย.

สำนักทดสอบทางการศึกษา. (2546 ก). *คู่มือการจัดสอบวัดผลสัมฤทธิ์ทางการเรียนชั้นประถมศึกษาปีที่ 6 ปีการศึกษา 2546.* วันที่ค้นข้อมูล 9 มกราคม 2551 เข้าถึงได้จาก http://bet.obec.go.th.

สำนักทดสอบทางการศึกษา. (2546 ข). *ผลการสอบวัดผลสัมฤทธิ์ทางการเรียน ระดับชั้นประถมศึกษาปีที่ 6 ปีการศึกษา 2546.* วันที่ค้นข้อมูล 9 มกราคม 2551 เข้าถึงได้จาก http://bet.obec.go.th.

Chang, H. H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33*(3), 333 – 353.

Clauser, B.E., Mazor, K.M., & Hambleton, R. K. (1993). The effect of purification of matching criterion on the Identification of DIF using the Mantel – Haenszel . *Applied Measurement in Education. 6*(4), 269 – 279.

Douglas, A., Roussos, A., & Stout, F. (1996). Item–bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement, 33*(4), 465 – 484.

Li, H., & Stout, W. (1996). Multidimensional DIF analyses: The effects of matching on unidimensional subtest scores. *Psychological Measurement, 22*(4), 357 – 367.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel – Haenszel statistic. *Educational and Psychological Measurement, 52*(3), 443 – 451.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's Test for DIF. *Journal of Educational Measurement, 30*(4), 293 – 311.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel – Haenszel and simultaneous item bias procedure for detecting differential item functioning. *Applied Psychological Measurement, 18*(4), 315 – 328.

Narayanan, R., & Swaminathan, S. (1996). Using statistical procedures to identify differentially functioning test items. *Education Measurement, 17*(1), 31 – 44.

Shealy, R., & Stout, W. (1993). A model – based standardization approach that separates true bias/ DIF form group ability difference and detect test bias/ DIF as well as item bias/ DIF. *Psychometrika, 58*(2), 159 – 194.

Steven, R. S. (2005). *Estimates of Type I Error and Power for Indices of Differential Bundle And Test Functioning.* Retrieved February 23, 2008, from http://proquest.umi.com /pqdweb?index.

Stout, W. F., Li, H., Nandakumar, R., & Bolt, D. (1997). MULTISIB: A procedure to investigate DIF when a test is intentionally two-dimensional. *Applied Psychological Measurement, 21*, 195 – 213.

Stout, W. F., & Rousos, L. A. (1999). *Dimensionality – based DIF/ DBF Package* [Computer program]. Champaign – Urbana, IL: William Stout Institute for Measurement, University of Illinois.