

การเปรียบเทียบค่าความเที่ยงของแบบทดสอบผลสัมฤทธิ์ทางการเรียน ที่มีจำนวนข้อสอบทำหน้าที่ย่อยต่างกันแตกต่างกัน

ณรงค์ จันทรมหา

โรงเรียนบ้านค่าย จังหวัดระยอง

บทคัดย่อ

การวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบค่าความเที่ยงของแบบทดสอบผลสัมฤทธิ์ทางการเรียนที่มีจำนวนข้อสอบทำหน้าที่ย่อยต่างกันแตกต่างกัน 7 เงื่อนไข คือ 0 % 5 % 10 % 15 % 20 % 25 % และ 30% กลุ่มตัวอย่างเป็นนักเรียนชั้นประถมศึกษาปีที่ 6 สังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน ปีการศึกษา 2546 จำนวน 2,000 คน ที่ทำแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ วิชาภาษาไทย จำนวน 40 ข้อ วิเคราะห์ค่าสถิติพื้นฐานโดยใช้โปรแกรม SPSS ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้โปรแกรมชิปเทสต์ และทดสอบความแตกต่างของค่าความเที่ยงของแบบทดสอบ โดยใช้การทดสอบสถิติซี

ผลการวิจัยปรากฏว่า ค่าความเที่ยงของแบบทดสอบที่มีจำนวนข้อสอบทำหน้าที่ย่อยต่างกัน 0 % 5 % 10 % 15 % 20 % 25 % และ 30 % ไม่แตกต่างกัน

คำสำคัญ: ค่าความเที่ยงของแบบทดสอบ/ การทำหน้าที่ย่อยต่างกันของข้อสอบ

A Comparison of Achievement Test Reliabilities under Increasing Levels of DIF

Narong Chanmaha

Bankhai School, Rayong, Thailand

Abstract

The purpose of this study was to compare reliability figures for test having a varying number of items with DIF, differential item functioning. Seven categories were used: test with no DIF items (0% DIF), tests with 5 % DIF, 10 % DIF, 15 % DIF, 20 % DIF, 25 % DIF, and 30 % DIF. The sample involved 2,000 Grade 6 students who sat a 40-item national achievement test on the Thai Language in a school under the jurisdiction of the Office of the Basic Education Commission. Descriptive statistics were derived by using SPSS; DIF was detected by use of the SIBTEST program. Gender was the classifying variable. The Z-test was used to compare test Reliabilities.

The result was as follow:

No significant differences were found among the Reliability coefficients as the number of DIF items changed over the seven levels mentioned.

Keywords: reliability coefficient of test, differential item functioning

ความนำ

ความยุติธรรมเป็นคุณลักษณะที่สำคัญอย่างหนึ่งของข้อสอบ ในการสอบแต่ละครั้งผู้สอบอาจมีคุณลักษณะแตกต่างกัน เช่น เชื้อชาติ ศาสนา วัฒนธรรม ภูมิสำเนา สังคม เพศ ภาษา อายุ และประสบการณ์ เป็นต้น ผู้สอบดังกล่าวอาจไม่ได้รับความยุติธรรมในการทำข้อสอบ โดยข้อสอบบางข้อมีความลำเอียงเข้าข้างกลุ่มผู้สอบย่อยบางกลุ่ม ซึ่งอาจทำให้เกิดการได้เปรียบเสียเปรียบ ทั้ง ๆ ที่ทำข้อสอบเดียวกัน สาเหตุดังกล่าวอาจเนื่องมาจากข้อสอบไม่ได้วัดความสามารถเป้าหมายที่ต้องการวัด (Target Ability: θ) เพียงอย่างเดียว แต่ยังวัดความสามารถแทรกซ้อนที่ไม่ต้องการวัด (Nuisance Ability: η) อีกด้วย ข้อสอบทุกข้อจะวัดความสามารถเป้าหมาย แต่ข้อสอบทำหน้าที่ต่างกันจะวัดทั้งความสามารถเป้าหมายและความสามารถแทรกซ้อน (Nandakumar, 1993) แสดงว่าถ้าผู้สอบกลุ่มย่อยกลุ่มใดมีความสามารถแทรกซ้อนสูงกว่าก็มีโอกาสตอบข้อสอบถูกมากกว่า ทั้ง ๆ ที่มีความสามารถเป้าหมายที่ต้องการวัดเท่ากัน จึงมีผลทำให้ข้อสอบทำหน้าที่ต่างกัน ข้อสอบทำหน้าที่ต่างกันเมื่อผู้สอบที่มีความสามารถระดับเดียวกัน แต่อยู่ต่างกลุ่มกัน มีโอกาสในการตอบข้อสอบถูกได้ต่างกัน (Li & Stout, 1993 cited in Narayanan & Swaminathan, 1996) ซึ่งขนาดและทิศทางของการทำหน้าที่ต่างกัน จะแปรเปลี่ยนตามระดับความสามารถที่ต่างกัน เช่น ผู้สอบที่ต้องการวัดความสามารถด้านคณิตศาสตร์ (ความสามารถเป้าหมาย) แต่ข้อสอบมีเนื้อหาเกี่ยวกับกีฬาฟุตบอล (ความสามารถแทรกซ้อน) เมื่อกลุ่มผู้สอบมีความแตกต่างทางเพศ (กลุ่มผู้ชายกับกลุ่มผู้หญิง) กลุ่มผู้สอบเพศชายจะมีความรู้เรื่องกีฬาฟุตบอลมาก ในขณะที่มีความรู้ด้านคณิตศาสตร์เท่ากันกับกลุ่มผู้สอบเพศหญิง ผู้สอบทั้งสองกลุ่มย่อยจะตอบข้อสอบได้ถูกต้องแตกต่างกัน โดยกลุ่มผู้สอบเพศชายจะตอบข้อสอบได้ถูกต้องมากกว่ากลุ่มผู้สอบเพศหญิง

โดยทั่วไปในแบบทดสอบมาตรฐานวัดผลสัมฤทธิ์ทางการเรียน ถ้ามีสัดส่วนของข้อสอบทำหน้าที่ต่างกันร้อยละ 10 ถึง 15 ถือว่าไม่ผิดปกติ แต่ถ้าสัดส่วนของข้อสอบทำหน้าที่ต่างกันร้อยละ 20 ถือว่าเป็นเรื่องผิดพลาดอย่างมาก (Clauser, 1993 cited in Narayanan & Swaminathan, 1996) ปัจจุบันนักวิจัยได้ให้ความสนใจการตรวจสอบสัดส่วนของข้อสอบทำหน้าที่ต่างกันมากขึ้น ทั้งนี้เนื่องจากแบบทดสอบมาตรฐานที่ใช้ตรวจสอบคุณภาพของผู้เรียน ไม่เพียงเพื่อการนำไปใช้ในการวัดผลสัมฤทธิ์ทางการเรียนเท่านั้น ยังนำไปใช้ด้านอื่นๆ เช่น คัดเลือกบุคคลเข้าศึกษาต่อ บรรจุเข้าทำงาน เลื่อนขั้นหรือเลื่อนตำแหน่ง เป็นต้น แบบทดสอบที่มีข้อสอบทำหน้าที่ต่างกันจำนวนมากจะทำให้แบบทดสอบมีคุณภาพต่ำ ขาดความน่าเชื่อถือ ความสามารถที่วัดได้ไม่ใช่ความสามารถที่แท้จริง การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ถ้าวิธีที่ใช้ตรวจสอบมีประสิทธิภาพสูง จะทำให้ข้อสอบมีความเป็นมาตรฐานมากขึ้น สามารถวัดความสามารถของผู้สอบได้อย่างถูกต้องแม่นยำ

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF) เป็นการเปรียบเทียบผลการตอบข้อสอบของผู้สอบ 2 กลุ่ม กลุ่มแรก เรียกว่า “กลุ่มเปรียบเทียบ (Focal Group)” เป็นกลุ่มที่ผู้วิจัยสนใจศึกษาและคาดว่าเป็นกลุ่มที่เสียประโยชน์ในการตอบข้อสอบ และกลุ่มที่สอง เรียกว่า “กลุ่มอ้างอิง (Reference Group)” ซึ่งเป็นกลุ่มที่คาดว่าจะได้รับประโยชน์ในการตอบข้อสอบ โดยข้อสอบทำหน้าที่แตกต่างกันแบ่งเป็น 2 ประเภท คือ ข้อสอบทำหน้าที่ต่างกันแบบเอกรูป (Uniform DIF) และข้อสอบทำหน้าที่ต่างกันแบบอนเอกรูป (Nonuniform DIF) ข้อสอบทำหน้าที่ต่างกันแบบอนเอกรูปยังแบ่งได้ 2 ลักษณะ ได้แก่ ข้อสอบทำหน้าที่ต่างกันแบบอนเอกรูปแบบมีปฏิสัมพันธ์ไม่เป็นลำดับ (Disordinal Interaction) และข้อสอบทำหน้าที่ต่างกันแบบอนเอกรูปแบบมีปฏิสัมพันธ์เป็นลำดับ (Ordinal Interaction) (Swaminathan & Rogers, 1990)

กระบวนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้วิธีการทางสถิติสามารถตรวจสอบได้ 2 วิธี ได้แก่

1. วิธีใช้เกณฑ์ภายนอก (External Method) โดยใช้ความสัมพันธ์กับเกณฑ์ภายนอก เช่น เกรดเฉลี่ย ผลการปฏิบัติงาน เป็นต้น แต่วิธีนี้ไม่นิยมนัก เพราะถ้าเกณฑ์ภายนอกที่นำมาหาความสัมพันธ์ไม่มีมาตรฐานแล้ว ผลการตรวจสอบจะขาดความถูกต้อง

2. วิธีใช้เกณฑ์ภายใน (Internal Method) เป็นวิธีที่สามารถตรวจสอบโครงสร้างภายในของข้อสอบ โดยพิจารณาคะแนนที่ได้จากผลการตอบของผู้เข้าสอบแต่ละกลุ่มว่าวัดในคุณลักษณะที่ต้องการวัดตามโครงสร้างเดียวกันหรือไม่ ซึ่งเป็นการสนับสนุนการตรวจสอบความตรงเชิงโครงสร้าง

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF Detection) จำแนกตามลักษณะการตรวจให้คะแนน ได้เป็น 2 ประเภท คือ ข้อสอบที่ให้คะแนนแบบทวิภาคหรือสองค่า (Dichotomous Scoring) และข้อสอบที่ให้คะแนนแบบพหุภาค หรือหลายค่า (Polytomous Scoring) วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ แต่ละประเภทยังสามารถจำแนกได้อีก 2 มิติ ได้แก่ มิติลักษณะของตัวแปรเกณฑ์ ซึ่งแบ่งเป็นกลุ่มวิธีที่ใช้คะแนนสังเกตได้ (Observed Score) และกลุ่มวิธีที่ใช้คุณลักษณะแฝงหรือคะแนนของตัวแปรแฝง (Latent Variable) และ มิติลักษณะของสถิติวิเคราะห์ ซึ่งแบ่งเป็นกลุ่มที่ใช้สถิติพารามетริก (Parametric Approach) และกลุ่มวิธีที่ใช้สถิตินั้นพารามетริก (Nonparametric Approach) (Potenza & Dorans, 1995; Feinstein, 1995)

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนแบบทวิภาค กลุ่มวิธีที่ใช้คะแนนสังเกตได้ ส่วนใหญ่วิเคราะห์ตามทฤษฎีการทดสอบแบบดั้งเดิม (CCT) หรือเรียกว่ากลุ่มที่ไม่ใช้ทฤษฎี การตอบสนองข้อสอบ (Non-IRT Approach) โดยใช้คะแนนรวมของผู้สอบเป็นเกณฑ์ในการจับคู่กลุ่มผู้สอบ วิธีการตรวจสอบที่สำคัญในกลุ่มนี้ที่ใช้สถิติพารามетริกได้แก่ การวิเคราะห์ความแปรปรวน (ANOVA) วิธีการวิเคราะห์ถดถอยโลจิสติกส์ (Logistic Regression: LR) และวิธีที่ใช้สถิตินั้นพารามетริก ได้แก่ วิธีแปลงค่าความยากของข้อสอบ (Transformed Item Difficulty: TID) วิธีดัชนีมาตรฐาน (Standardization: STND) และวิธีตารางการณัจร (Contingency Table: CT) ซึ่งในวิธีตารางการณัจรประกอบด้วย วิธีล็อก – ลิเนียร์ (Log – Linear: LL) วิธีไค – สแควร์ (Chi – Square: χ^2) และ วิธีแมนเทล – แฮนส์เซล (Mantel – Haenszel: MH) กลุ่มวิธีที่ใช้คุณลักษณะแฝงจะวิเคราะห์บนพื้นฐานของทฤษฎี การตอบสนองข้อสอบ (IRT) สำหรับใช้เป็นเกณฑ์ในการจับคู่กลุ่มผู้สอบ วิธีการตรวจสอบที่สำคัญในกลุ่มนี้ที่ใช้ สถิติพารามетริก ได้แก่ วิธีวัดพื้นที่ความแตกต่างระหว่างโค้งการตอบสนองข้อสอบ (IRT-D²) โดยวิธีวัดพื้นที่ของ Raju และวิธีการวัดพื้นที่ของ Kim & Cohen (1991) วิธีเปรียบเทียบค่าพารามетอร์ซึ่งในวิธีนี้ประกอบด้วยวิธีการ เปรียบเทียบค่าความยาก (Difficulty Shift) วิธีการทดสอบ F วิธีการทดสอบไค-สแควร์ของลอร์ด (Lord's Chi – Square Test) วิธี IRT แบบเทียม (Pseudo – IRT) และวิธีทดสอบอัตราส่วนไลค์ลิฮูดทั่วไป (Likelihood Ratio Test: LRT) รูปแบบที่ใช้สถิตินั้นพารามетริก ได้แก่ วิธีซิปเทสท์ (SIBTEST) (Millsap & Everson, 1993; Holland & Wainer, 1993; Potenza & Dorans, 1995; Feinstein, 1995)

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนแบบพหุภาค กลุ่มที่ใช้คะแนน สังเกตได้ วิธีที่ใช้ตรวจสอบโดยใช้สถิติพารามетริกได้แก่ วิธีการวิเคราะห์ความแปรปรวน วิธีการวิเคราะห์ถดถอยโลจิสติกส์ พหุภาค (Polytomous Logistic Regression) วิธีตรวจสอบโดยใช้สถิตินั้นพารามетริกได้แก่ วิธีดัชนีมาตรฐานพหุภาค (Polytomous Standardization) และวิธีแมนเทล – แฮนส์เซล ทั่วไป (General Mantel – Haenszel: GMH) กลุ่ม

วิธีที่ใช้คุณลักษณะแฝง วิธีที่ใช้ตรวจสอบโดยใช้สถิติพารามเมตริกได้แก่ วิธีอัตราส่วนโลคลิฮูดในรูปทั่วไป (General IRT Likelihood Ratio) วิธีการให้คะแนนบางส่วน (Partial Credit Model: PCM) วิธีทดสอบโดยใช้สถิติพารามเมตริกได้แก่ วิธีชิปเทสพหุภาค (Polytomous SIBTEST) และวิธีการให้คะแนนบางส่วนทั่วไป (Generalized Partial Credit Model: GPCM) (Potenza & Dorans, 1995; Feinstein, 1995)

วิธีการทางสถิติที่ใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่นิยมใช้มากคือ วิธีแมนเทิล-แฮนส์เซล วิธีชิปเทส วิธีทำให้เป็นมาตรฐาน และวิธีถดถอยโลจิสติก (David & Lori, 2000)

วิธีชิปเทส (SIBTEST: Simultaneous Item Bias Test) พัฒนาโดยเชียลีและสตาตส์ (Shealy & Stout, 1993) ใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ การทำหน้าที่ต่างกันของแบบทดสอบ (Differential Test Functioning: DTF) และการทำหน้าที่ต่างกันของหมวดข้อสอบ (Differential Bundle Functioning: DBF) วิธีนี้สามารถวิเคราะห์ได้ทั้งแบบทดสอบเอกมิติ (Unidimensional Test) และแบบทดสอบพหุมิติ (Multidimensional Test) วิธีชิปเทสที่ใช้สถิติทดสอบแบบพารามเมตริก ซึ่งพัฒนามาบนพื้นฐานของทฤษฎี IRT ชนิดพหุมิติ แต่ไม่ต้องใช้ฟังก์ชันการตอบสนองข้อสอบหรือการประมาณค่าความสามารถแฝง วิธีชิปเทสที่ได้รับการออกแบบมาสำหรับตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบมีทิศทางเดียว (Unidirectional DIF) ดังนั้นจึงไม่ไวต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ไม่มีทิศทาง (Nondirectional DIF) (Li & Stout, 1996 cited in Shealy & Stout, 1993) จุดเด่นของวิธีชิปเทส คือ สามารถคำนวณได้ง่าย ไม่ซับซ้อน ประหยัดค่าใช้จ่าย ไม่จำเป็นต้องใช้กลุ่มตัวอย่างที่มีขนาดใหญ่ และใช้สถิติทดสอบน้อยสำคัญ นอกจากนี้ยังสามารถนำไปประยุกต์กับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนแบบพหุภาคได้

งานวิจัยทางด้านนี้ส่วนใหญ่ เป็นการเปรียบเทียบวิธีการตรวจสอบข้อสอบทำหน้าที่ต่างกันว่า วิธีการใดมีประสิทธิภาพมากที่สุด แต่ยังไม่พบงานวิจัยเกี่ยวกับคุณภาพของแบบทดสอบด้านความเที่ยง เมื่อมีข้อสอบทำหน้าที่ต่างกันระดับต่าง ๆ จะมีผลต่อคุณภาพของแบบทดสอบด้านความเที่ยงอย่างไร

จากเหตุผลดังกล่าวข้างต้นผู้วิจัยจึงสนใจศึกษาการทำหน้าที่ต่างกันของข้อสอบโดยวิธีชิปเทส เพราะเป็นวิธีที่เป็นมาตรฐาน เชื่อถือได้และยังเป็นการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่พิจารณาจากความแตกต่างของคะแนนจริงระหว่างผู้สอบที่มีความสามารถระดับเดียวกัน (นพมาศ พิพัฒน์สุข, 2541) ในการวิจัยครั้งนี้ได้เลือกศึกษาตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ วิชาภาษาไทย ชั้นประถมศึกษาปีที่ 6 (กระทรวงศึกษาธิการ, 2546 อ้างถึงใน ปิยะทิพย์ ดินวรรและคณะ, 2550) ซึ่งเป็นข้อสอบที่มีการให้คะแนนแบบทวิภาคหรือสองค่า (Dichotomous Scoring) ที่เลือกศึกษาวิชาภาษาไทยเพราะเป็นวิชาที่ส่วนใหญ่พบความลำเอียงต่อเพศอย่างชัดเจน โดยส่วนใหญ่จะลำเอียงเข้าข้างผู้สอบเพศหญิง และข้อสอบวิชาภาษาไทยเป็นวิชาที่มีวัฒนธรรมทางภาษาเข้ามาเกี่ยวข้องด้วย จึงทำให้ถ้อยคำและข้อความลำเอียงได้ง่าย ซึ่งสอดคล้องกับลิน (Linn, 1993) ซึ่งได้ศึกษาพบว่าในแบบทดสอบ SAT จำนวน 7 ฉบับ พบข้อสอบทำหน้าที่ต่างกัน 29 ข้อ เป็นข้อสอบทางภาษาจำนวน 25 ข้อ นอกจากนี้ กาญจน วรณสุนทร (2537) พบว่าข้อสอบทำหน้าที่ต่างกันมากกว่าครึ่งในวิชาภาษาอังกฤษลำเอียงเข้าข้างเพศหญิง

การวิจัยนี้ผู้วิจัยจะเปรียบเทียบค่าความเที่ยงของแบบทดสอบ โดยสุ่มข้อมูลผลการตอบข้อสอบของผู้เข้าสอบจำนวน 2,000 คน เป็นชาย 1,000 คน และหญิง 1,000 คน จัดข้อมูลด้วยการกระทำปัจจัยที่เปลี่ยนแปลง คือ จำนวนของข้อสอบทำหน้าที่ต่างกัน 7 เงื่อนไข ได้แก่ 0 %, 5 %, 10 %, 15 %, 20 %, 25 % และ 30 %

การจัดกระทำข้อมูลดังกล่าวให้มีระดับการทำหน้าที่ต่างกัน 7 ระดับ ทำได้โดยนำผลการตอบข้อสอบ มาตรวจสอบ การทำหน้าที่ต่างกันโดยวิธีซิปเทสท์ เมื่อพบข้อสอบที่ทำหน้าที่ต่างกันแล้ว นำข้อสอบดังกล่าวมารวมกับข้อสอบ ที่ไม่ทำหน้าที่ต่างกัน ให้ได้สัดส่วนข้อสอบที่ทำหน้าที่ต่างกัน 7 ระดับ แล้วทำการวิเคราะห์หาค่าความเที่ยงของ แบบทดสอบ ซึ่งข้อค้นพบที่ได้นี้จะเป็นประโยชน์ในการตรวจสอบคุณภาพของข้อสอบ และปรับปรุงคุณภาพของ แบบทดสอบ นอกจากนี้ยังสามารถนำไปเป็นสารสนเทศในการปรับปรุงแบบทดสอบให้มีคุณภาพดียิ่งขึ้นต่อไป

วัตถุประสงค์ของการวิจัย

1. เพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบจากแบบทดสอบผลสัมฤทธิ์ทางการเรียนระดับชาติ วิชาภาษาไทย ชั้นประถมศึกษาปีที่ 6 ปีการศึกษา 2546
2. เพื่อเปรียบเทียบค่าความเที่ยงของแบบทดสอบ ภายใต้เงื่อนไขจำนวนข้อสอบทำหน้าที่ ต่างกัน แตกต่างกัน 7 เงื่อนไข ได้แก่ 0 %, 5 %, 10 %, 15 %, 20 %, 25 % และ 30 %

กรอบแนวคิดการวิจัย

การเปรียบเทียบค่าความเที่ยงของแบบทดสอบที่มีจำนวนข้อสอบทำหน้าที่ต่างกันแตกต่างกัน ผู้วิจัยได้ ศึกษางานวิจัยเกี่ยวกับคุณภาพของแบบทดสอบด้านความเที่ยงจากงานวิจัยของ อารี วัชรโสติกุล (2543) ที่เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยใช้รูปแบบและวิธีการที่ต่างกัน พบว่า ค่าความเที่ยง ของแบบทดสอบหลังจากตัดข้อสอบทำหน้าที่ต่างกันออก แล้วทำการปรับขยายค่าความเที่ยงของแบบทดสอบให้มี จำนวนข้อสอบเท่ากัน เมื่อตรวจสอบด้วยวิธีซิปเทสท์ ปรากฏว่า ค่าความเที่ยงแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ ระดับ .05 ซึ่งสอดคล้องกับผลงานวิจัยของรักชนก ยี่สุนศรี (2544) ได้วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบและ แบบสอบด้วยกระบวนการ ดี เอฟ ไอ ที่ สำหรับแบบสอบคัดเลือกบุคคลเข้าศึกษาในระดับอุดมศึกษาวิชาภาษาอังกฤษ และวิชาคณิตศาสตร์ พบว่าเมื่อเปรียบเทียบคุณภาพของแบบสอบก่อนและหลังตัดข้อสอบทำหน้าที่ต่างกัน ออกแบบทดสอบฉบับหลังที่ตัดข้อสอบทำหน้าที่ต่างกันออกส่วนใหญ่มีค่าความเที่ยงลดลง และยังสอดคล้องกับงาน วิจัยของศุภวัฒน์ มะลิเผือก (2549) ที่ศึกษาเกี่ยวกับอิทธิพลการทำหน้าที่ต่างกันของข้อสอบที่ส่งผลต่อคุณภาพของ แบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ ชั้นประถมศึกษาปีที่ 6 พบว่า ค่าความเที่ยงของแบบทดสอบ ฉบับก่อนและหลังตัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติ ที่ระดับ .05 ดังนั้นจึงตั้งสมมติฐานการวิจัยไว้ดังนี้

แบบทดสอบที่มีจำนวนข้อสอบทำหน้าที่ต่างกัน 7 เงื่อนไข ได้แก่ 0 %, 5 %, 10 %, 15 %, 20 %, 25 % และ 30 % มีค่าความเที่ยงของแบบทดสอบแตกต่างกัน

วิธีดำเนินการวิจัย

กลุ่มตัวอย่าง

กลุ่มตัวอย่าง เป็นนักเรียนชั้นประถมศึกษาปีที่ 6 สังกัดคณะกรรมการการศึกษาขั้นพื้นฐาน กระทรวงศึกษาธิการ ปีการศึกษา 2546 ที่เข้าสอบวัดผลสัมฤทธิ์ทางการเรียนวิชาภาษาไทย จำนวน 2,000 คน แบ่งเป็นนักเรียนชาย 1,000 คน และนักเรียนหญิง 1,000 คน ได้มาโดยการสุ่มแบบแบ่งชั้นไม่กำหนดสัดส่วน ใช้ระดับความสามารถ (คะแนนรวม) เป็นชั้นในการแบ่งเป็น 3 กลุ่ม ได้แก่ กลุ่มดี กลุ่มพอใช้ และกลุ่มปรับปรุง

เครื่องมือการวิจัย

แบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ ชั้นประถมศึกษาปีที่ 6 วิชาภาษาไทย ประกอบด้วยข้อสอบแบบ 4 ตัวเลือก จำนวน 40 ข้อ คะแนนเต็ม 40 คะแนน สร้างโดยสำนักทดสอบทางการศึกษา กระทรวงศึกษาธิการ คณะกรรมการออกข้อสอบ ประกอบด้วย ครูผู้สอน ศึกษานิเทศก์ นักวิชาการ และผู้เชี่ยวชาญที่เกี่ยวข้อง มีโครงสร้างของแบบทดสอบ ดังนี้

1. วัดโครงสร้างความรู้ จำนวน 15 ข้อ
2. วัดกระบวนการ จำนวน 25 ข้อ

แบบทดสอบทั้งฉบับมีค่าความยากง่ายประมาณ .50 ค่าอำนาจจำแนกทั้งฉบับมากกว่า .40 และค่าความเที่ยงของทั้งฉบับประมาณ เท่ากับ .80

การเก็บรวบรวมข้อมูล

ผู้วิจัยรวบรวมข้อมูลผลการตอบของนักเรียนที่เป็นกลุ่มตัวอย่าง ซึ่งเป็นข้อมูลทุติยภูมิจากฐานข้อมูลเก่าของวิทยาลัยวิทยาการวิจัยและวิทยาการปัญญา มหาวิทยาลัยบูรพา ซึ่งได้นำมาจากสำนักทดสอบทางการศึกษา กระทรวงศึกษาธิการ

การวิเคราะห์ข้อมูล

วิเคราะห์ค่าสถิติพื้นฐานของแบบทดสอบโดยใช้โปรแกรม SPSS และวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธีซิปเทสท์ โดยใช้โปรแกรมซิปเทสท์ ซึ่งเป็นโปรแกรมที่ใช้วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบที่มีลักษณะของข้อสอบแบบ Dichotomous (ข้อสอบที่มีการให้คะแนน 0,1) โดยใช้วิธีการทดสอบการทำหน้าที่ต่างกันโมเดลพหุมิติของ Shealy and Stout (1993)

ผลการวิจัย

1. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ จากข้อมูลทฤษฎีซึ่งเป็นผลการตอบข้อสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ วิชาภาษาไทย 40 ข้อ ของกลุ่มตัวอย่างจำนวน 2,000 คน โดยโปรแกรมชิปเทสท์ พบข้อสอบทำหน้าที่ต่างกัน จำนวน 13 ข้อ ได้แก่ ข้อที่ 2, 9, 13, 21, 23, 24, 27, 28, 34, 36, 38, 39 และ 40 และข้อสอบทำหน้าที่ไม่ต่างกัน 27 ข้อ เมื่อจำแนกกลุ่มผู้สอบตามตัวแปรเพศ แสดงดังตารางที่ 1

ตารางที่ 1 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ จำนวน 40 ข้อ ผู้เข้าสอบจำนวน 2,000 คน

ข้อที่	β_{uni}	$p - value$	หมายเหตุ
1	.02	.50	
2	.05	.01*	ลำเอียงเข้าข้างเพศหญิง
3	.00	.89	
4	.00	.98	
5	-.02	.29	
6	-.04	.07	
7	.01	.76	
8	-.02	.30	
9	.06	.00*	ลำเอียงเข้าข้างเพศหญิง
10	-.04	.07	
11	.01	.76	
12	.00	.95	
13	.06	.01*	ลำเอียงเข้าข้างเพศหญิง
14	-.02	.29	
15	-.01	.73	
16	-.03	.12	
17	-.03	.13	
18	.00	.97	
19	.03	.19	
20	-.03	.25	
21	-.07	.00*	ลำเอียงเข้าข้างเพศชาย
22	-.03	.20	
23	-.04	.04*	ลำเอียงเข้าข้างเพศชาย
24	-.06	.01*	ลำเอียงเข้าข้างเพศชาย
25	-.03	.14	
26	-.01	.66	
27	.05	.02*	ลำเอียงเข้าข้างเพศหญิง
28	.05	.03*	ลำเอียงเข้าข้างเพศหญิง
29	.02	.36	
30	-.03	.11	

ตารางที่ 1 (ต่อ)

ข้อที่	β_{uni}	$p - value$	หมายเหตุ
31	.02	.29	
32	.02	.34	
33	.02	.28	
34	.07	.00*	ลำเอียงเข้าข้างเพศหญิง
35	.04	.06	
36	.10	.00*	ลำเอียงเข้าข้างเพศหญิง
37	.00	.89	
38	.07	.00*	ลำเอียงเข้าข้างเพศหญิง
39	.07	.00*	ลำเอียงเข้าข้างเพศหญิง
40	-.04	.04*	ลำเอียงเข้าข้างเพศชาย

หมายเหตุ. * $p < .05$, β_{uni} = ดัชนีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

2. ผลการเปรียบเทียบค่าความเที่ยงของแบบทดสอบระหว่างแบบทดสอบฉบับที่ไม่มีข้อสอบทำหน้าที่ต่างกัน หรือมีข้อสอบทำหน้าที่ต่างกัน 0 % ซึ่งมีค่าความเที่ยงของแบบทดสอบ 0.7990 กับ แบบทดสอบฉบับที่มีข้อสอบทำหน้าที่ต่างกันระดับต่างๆ คือฉบับที่มีข้อสอบทำหน้าที่ต่างกัน 3.70 % (เป็นตัวแทนเงื่อนไข 5 %) ซึ่งมีค่าความเที่ยงของแบบทดสอบ 0.8029 ฉบับที่มีข้อสอบทำหน้าที่ต่างกัน 11.11 % (เป็นตัวแทนเงื่อนไข 10 %) มีค่าความเที่ยงของแบบทดสอบ 0.8086 ฉบับที่มีข้อสอบทำหน้าที่ต่างกัน 14.81 % (เป็นตัวแทนเงื่อนไข 15 %) มีค่าความเที่ยงของแบบทดสอบ 0.8086 ฉบับที่มีข้อสอบทำหน้าที่ต่างกัน 22.22 % (เป็นตัวแทนเงื่อนไข 20 %) มีค่าความเที่ยงของแบบทดสอบ 0.8041 ฉบับที่มีข้อสอบทำหน้าที่ต่างกัน 25.93 % (เป็นตัวแทนเงื่อนไข 25 %) มีค่าความเที่ยงของแบบทดสอบ 0.8019 และฉบับที่มีข้อสอบทำหน้าที่ต่างกัน 29.63 % (เป็นตัวแทนเงื่อนไข 30 %) มีค่าความเที่ยงของแบบทดสอบ 0.8018 ผลการเปรียบเทียบพบว่า ค่าความเที่ยงไม่แตกต่างกันอย่างมีนัยสำคัญทางสถิติ และเมื่อวิเคราะห์เปรียบเทียบค่าความเที่ยงของแบบทดสอบ ระหว่างฉบับที่ไม่มีข้อสอบทำหน้าที่แตกต่างกันกับฉบับที่มีข้อสอบทำหน้าที่ต่างกันและลำเอียงเข้าหากลุ่มผู้สอบกลุ่มใดกลุ่มหนึ่ง ค่าความเที่ยงของแบบทดสอบไม่แตกต่างกันอย่างมีนัยสำคัญทางสถิติ แสดงดังตารางที่ 2 และ 3

ตารางที่ 2 ค่าความเที่ยงของแบบทดสอบที่ปรับตามเงื่อนไขจำนวนข้อสอบทำหน้าที่ต่างกัน

จำนวนข้อสอบ NO DIF	จำนวนข้อสอบ DIF	รวมข้อสอบ ทั้งหมด	% ข้อสอบ DIF ในแบบทดสอบ	ค่าความเที่ยง
27	0	27	0	.7990
26	1	27	3.70	.8029
24	3	27	11.11	.8086
23	4	27	14.81	.8086
21	6	27	22.22	.8041
20	7	27	25.93	.8019
19	8	27	29.63	.8018

ตารางที่ 3 การเปรียบเทียบค่าความเที่ยงของแบบทดสอบ ที่มีจำนวนข้อสอบทำหน้าที่ต่างกันแตกต่างกันตามเงื่อนไขที่กำหนด

การเปรียบเทียบ	ค่าความเที่ยงของแบบทดสอบ	Z-test	p-value
NO DIF	.7990	.038	.970
DIF 3.70 %	.8029		
NO DIF	.7990	.097	.918
DIF 11.11 %	.8086		
NO DIF	.7990	.048	.966
DIF 14.81 %	.8086		
NO DIF	.7990	.048	.966
DIF 22.22 %	.8041		
NO DIF	.7990	.031	.976
DIF 25.93 %	.8019		
NO DIF	.7990	.031	.976
DIF 29.63 %	.8018		

อภิปรายผลการวิจัย

ผลการวิจัยปรากฏว่า แบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ วิชาภาษาไทย ของนักเรียน ชั้นประถมศึกษาปีที่ 6 สังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน กระทรวงศึกษาธิการ ปีการศึกษา 2546 มีข้อสอบทำหน้าที่ย่อยต่างกัน 13 ข้อ คิดเป็นร้อยละ 32.5 ค่าความเที่ยงของแบบทดสอบฉบับที่ตัดข้อสอบทำหน้าที่ย่อยต่างกันออก 13 ข้อ ซึ่งเหลือข้อสอบในแบบทดสอบ 27 ข้อ วิเคราะห์ค่าความเที่ยงของแบบทดสอบฉบับที่ไม่มีข้อสอบทำหน้าที่ย่อยต่างกัน เปรียบเทียบค่าความเที่ยงของแบบทดสอบกับฉบับที่มีข้อสอบทำหน้าที่ย่อยต่างกัน ระดับต่าง ๆ คือ 3.70 % , 11.11 % , 14.81 % , 22.22 % , 25.93 % และ 29.63 % ซึ่งผลการเปรียบเทียบพบว่าค่าความเที่ยงของแบบทดสอบไม่แตกต่างกัน จากผลการวิจัยครั้งนี้ มีประเด็นสำคัญที่นำมาอภิปราย ดังต่อไปนี้

1. ผลการวิเคราะห์การทำหน้าที่ย่อยต่างกันของข้อสอบจากแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ วิชาภาษาไทยโดยโปรแกรมชิปเทสต์ ปรากฏว่าข้อสอบทำหน้าที่ย่อยต่างกันจำนวน 13 ข้อ คิดเป็นร้อยละ 32.5 ซึ่งเป็นข้อสอบในกลุ่มวัดโครงสร้างความรู้ 3 ข้อ คือข้อที่ 2, 9 และ 13 และกลุ่มวัดกระบวนการ 10 ข้อ คือข้อที่ 21, 23, 24, 27, 28, 34, 36, 38, 39 และ 40 เมื่อพิจารณาในรายละเอียด ปรากฏว่า ในจำนวนข้อสอบ 13 ข้อ ซึ่งทำหน้าที่ย่อยต่างกัน มีค่า β_{uni} มากกว่า 0 จำนวน 9 ข้อ คือข้อที่ 2, 9, 13, 27, 28, 34, 36, 38 และ 39 ซึ่งแสดงว่าข้อสอบทั้ง 9 ข้อ ลำเอียงเข้าข้างกลุ่มอ้างอิงหรือผู้สอบเพศหญิง และมีค่า β_{uni} น้อยกว่า 0 อยู่ 4 ข้อ คือข้อที่ 21, 23, 24 และ 40 ซึ่งเป็นข้อสอบที่ลำเอียงเข้าข้างกลุ่มเปรียบเทียบหรือผู้สอบเพศชาย เมื่อพิจารณาจำนวนข้อสอบและร้อยละของข้อสอบทำหน้าที่ย่อยต่างกัน แสดงว่าแบบทดสอบมีความผิดปกติมาก เพราะมีจำนวนข้อสอบทำหน้าที่ย่อยต่างกันมากกว่าร้อยละ 20 คือมีถึงร้อยละ 32.50 โดยความผิดปกติของข้อสอบที่เกิดขึ้นสอดคล้องกับข้อค้นพบของคลอสเซอร์ ที่ระบุว่า แบบทดสอบมาตรฐานที่ใช้วัดผลสัมฤทธิ์ทางการเรียน ถ้ามีสัดส่วนของข้อสอบที่ทำหน้าที่ย่อยต่างกันร้อยละ 10 – 15 ถือว่าไม่ผิดปกติ แต่ถ้ามีสัดส่วนของข้อสอบทำหน้าที่ย่อยต่างกันร้อยละ 20 ถือว่าผิดพลาดอย่างมาก (Clouser, 1993 citing Narayanan & Swaminathan, 1994) ทั้งนี้อาจเนื่องมาจาก ข้อสอบมีเนื้อหาที่ผู้สอบกลุ่มเพศชายซึ่งเป็นกลุ่มเปรียบเทียบไม่คุ้นเคย หรือข้อสอบวัดในสิ่งที่ไม่สอดคล้องกับสิ่งที่กำหนดไว้ในหลักสูตร หรือข้อสอบมีเนื้อหาทักษะ หรือสารสนเทศที่ไม่เป็นการทั่วไปกับผู้สอบทุกคน และเมื่อวิเคราะห์อย่างละเอียด พบว่าข้อสอบในกลุ่มที่วัดทักษะกระบวนการ มีข้อสอบทำหน้าที่ย่อยต่างกันถึง 10 ข้อ จาก 25 ข้อ ซึ่งอาจเนื่องมาจากผู้สอบเพศชายมีทักษะกระบวนการด้านภาษาน้อย หรือข้อสอบวัดไม่ตรงกับทักษะที่หลักสูตรกำหนด ส่งผลให้เกิดการได้เปรียบเสียเปรียบกันในการตอบข้อสอบ โดยกลุ่มผู้สอบเพศหญิงซึ่งเป็นกลุ่มอ้างอิงได้เปรียบในการตอบข้อสอบส่วนกลุ่มผู้สอบเพศชายเสียเปรียบในการตอบข้อสอบ

2. ผลการเปรียบเทียบค่าความเที่ยงของแบบทดสอบระหว่างฉบับที่ไม่มีข้อสอบทำหน้าที่ย่อยต่างกันกับฉบับที่มีข้อสอบทำหน้าที่ย่อยต่างกัน 6 เงื่อนไข คือ 3.70 % , 11.11 % , 14.81 % , 22.22 % , 25.93 % และ 29.63 % ปรากฏว่าค่าความเที่ยงของแบบทดสอบมีค่าไม่สูงมากและไม่แตกต่างกัน ซึ่งไม่เป็นไปตามสมมติฐานที่ตั้งไว้ ทั้งนี้อาจเนื่องจากแบบทดสอบฉบับที่ไม่มีข้อสอบทำหน้าที่ย่อยต่างกัน มีปริมาณข้อสอบน้อยเพียง 27 ข้อ และเมื่อจัดข้อสอบทำหน้าที่ย่อยต่างกันเข้าไปแทนข้อสอบทำหน้าที่ย่อยไม่ต่างกัน เพื่อให้แบบทดสอบมีจำนวนข้อสอบ 27 ข้อเท่าเดิม จัดกระทำข้อมูลเพิ่มขึ้นประมาณ ครึ่งละ 5 เปอร์เซนต์ ซึ่งเมื่อคิดเป็นจำนวนข้อสอบเป็นเพียง 1 – 2 ข้อ จึงอาจมีผลทำให้ค่าความแปรปรวนของคะแนนมีค่าแตกต่างจากเดิมเพียงเล็กน้อย เนื่องจากจำนวนข้อสอบทั้งหมดในแบบทดสอบมีจำนวนน้อย ส่งผลให้ค่าความเที่ยงของแบบทดสอบมีค่าแตกต่างจากเดิมเพียงเล็กน้อย ซึ่งแสดงให้เห็นว่าเมื่อข้อสอบในแบบทดสอบมีจำนวนน้อย ข้อสอบทำหน้าที่ย่อยต่างกันจำนวนน้อยจะไม่มีผลต่อค่าความเที่ยงของแบบทดสอบ

ข้อเสนอแนะ

การนำผลการวิจัยไปใช้

1. ในการเปรียบเทียบค่าความเที่ยงของแบบทดสอบระหว่างฉบับที่ไม่มีข้อสอบทำหน้าที่ย่อยกับฉบับที่มีข้อสอบทำหน้าที่ย่อยในปริมาณต่าง ๆ เป็นแบบทดสอบที่มีข้อสอบเพียง 27 ข้อ ซึ่งถือว่ามีย่านน้อยทำให้ค่าความเที่ยงที่นำมาเปรียบเทียบมีค่าไม่แตกต่างกัน ดังนั้นถ้าต้องการนำผลการเปรียบเทียบค่าความเที่ยงไปใช้ไม่ควรตัดข้อสอบทำหน้าที่ย่อยออกทั้ง 13 ข้อ เนื่องจากแบบทดสอบที่มีข้อสอบ ทำหน้าที่ย่อยต่างกัน 10 - 15 % ถือว่าไม่ผิดปกติ ดังนั้นควรตัดข้อสอบออกเพียงบางข้อ โดยดูเนื้อหาของข้อสอบเป็นเกณฑ์ว่ามีความเหมาะสมหรือไม่เพื่อจะได้ค่าความเที่ยงที่ถูกต้องแม่นยำมากยิ่งขึ้น

2. จากผลการวิเคราะห์ข้อสอบ พบว่าในแบบทดสอบมีข้อสอบทำหน้าที่ย่อยต่างกันถึง 13 ข้อ เพื่อให้เกิดความยุติธรรมแก่ผู้เข้าสอบ ควรคิดคะแนนเฉพาะข้อสอบทำหน้าที่ย่อยไม่ต่างกัน เพราะจะทำให้ได้ข้อมูลด้านความสามารถที่แท้จริงของผู้เข้าสอบ

การทำวิจัยต่อไป

1. จากการทดสอบการทำหน้าที่ย่อยต่างกันของข้อสอบ ผู้วิจัยใช้วิธีชิปเทสท์ซึ่งเป็นวิธีที่เป็นมาตรฐานและเชื่อถือได้ รวมทั้งใช้โปรแกรมชิปเทสท์รุ่นล่าสุด ทำให้ตรวจพบข้อสอบทำหน้าที่ย่อยต่างกันหลายข้อ เพื่อให้ผลการตรวจสอบมีความน่าเชื่อถือมากขึ้น ควรมีการตรวจสอบด้วยวิธีต่าง ๆ อีกหลายวิธี แล้วพิจารณาเฉพาะข้อสอบทำหน้าที่ย่อยที่ตรงกัน เพื่อลดความคลาดเคลื่อนประเภทที่ 1

2. งานวิจัยครั้งนี้ใช้ข้อมูลจริง จึงมีข้อจำกัดในด้านจำนวนข้อสอบ และลักษณะของแบบทดสอบ ที่ไม่สามารถระบุให้มีจำนวนข้อสอบทำหน้าที่ย่อยต่างกัน แตกต่างกันตามเงื่อนไขที่กำหนด คือ 0 %, 5 %, 10 %, 15 %, 20 %, 25 % และ 30 % ได้ ดังนั้น ในการวิจัยต่อไป ถ้าใช้วิธีการจำลองข้อมูลจะทำให้ได้ผลการวิจัยที่แม่นยำมากยิ่งขึ้น

เอกสารอ้างอิง

- กาญจนา วัฒนสุนทร. (2537). *การพัฒนาเกณฑ์ตัดสินข้อสอบลำเอียงทางเพศ*. วิทยานิพนธ์ปริญญาครุศาสตรดุษฎีบัณฑิต, สาขาวิชาการวัดและประเมินผลการศึกษา, บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย.
- นพมาศ พิพัฒน์สุข. (2541). *การเปรียบเทียบประสิทธิภาพระหว่างวิธีแมนเทิล-แฮนส์เซล กับวิธีถดถอยโลจิสติกส์ในการตรวจสอบการทำหน้าที่ย่อยต่างกันของข้อสอบ เมื่อใช้เกณฑ์จับคู่เปรียบเทียบแตกต่างกันในแบบสอบชนิดพหุมิติ*. วิทยานิพนธ์ปริญญาครุศาสตรมหาบัณฑิต, สาขาวิชาการวัดและประเมินผลการศึกษา, บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย.
- ปิยะทิพย์ ดินวร, ม.ร.ว.สมพร สุทัศน์ และ เสรี ชัดแจ้ง. (2550). การตรวจสอบการทำหน้าที่ย่อยต่างกันของข้อสอบในแบบทดสอบพหุมิติ : การเปรียบเทียบประสิทธิภาพระหว่างวิธีการวิเคราะห์องค์ประกอบจำกักับวิธีถดถอยโลจิสติกส์. *วารสารวิจัยและวัดผลการศึกษา มหาวิทยาลัยบูรพา*, 5(1), 63 - 80.

- รักชนก ยี่สุนศรี. (2544). การวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบและแบบสอบด้วยกระบวนการตีเอฟไอไอที สำหรับแบบสอบคัดเลือกบุคคลเข้าศึกษาในสถาบันอุดมศึกษา วิชาภาษาอังกฤษและวิชาคณิตศาสตร์. วิทยานิพนธ์ปริญญาครุศาสตรมหาบัณฑิต, สาขาวิชาการวัดและประเมินผลการศึกษา, บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย.
- ศุภวัฒน์ มะลิเผือก. (2549). อิทธิพลของการทำหน้าที่ต่างกันของข้อสอบที่ส่งผลต่อคุณภาพของแบบทดสอบ วัดผลสัมฤทธิ์ทางการเรียนระดับชาติ ชั้นประถมศึกษาปีที่ 6. วารสารวิจัยและวัดผลการศึกษา มหาวิทยาลัยบูรพา, 4(1), 46 - 60.
- อารี วัชรโสติกุล. (2543). การเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้รูปแบบและวิธีการต่างกัน. วิทยานิพนธ์ปริญญาการศึกษามหาบัณฑิต, สาขาวิชาการวัดผลการศึกษา, บัณฑิตวิทยาลัย, มหาวิทยาลัยศรีนครินทรวิโรฒ.
- David S., J., & Lori M., D. (2000). An Expected response function approach to graphical differentiation item function. *Journal of Educational Measurement*, 37(3), 263-280.
- Feinstein, Z. S. (1995). Effect of differing item parameters on closed-interval DIF statistics. *Applied Psychological Measurement*, 19(2), 131-142.
- Holland, P. W., & Wainer, H. (1993). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Kim, S. H., & Cohen, A. S. (1991). A comparison of two area measures for detecting differential item functioning. *Applied Psychological Measurement*, 15(3), 269-278.
- Linn, R. L. (1993). The use of differential item functioning statistics: A discussion of current practice and future implementation. In P. W., Holland, & H., Wainer, (Eds.), *Differential item functioning* (pp. 349-364). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Millsap, R.E., & Everson, H. T. (1993). Methodology Review: Statistical Approaches for Assessing Measurement Bias. *Applied Psychological Measurement*, 17(1), 297-334.
- Nandakumar, R. (1993). Assessing essential unidimensionality of real data. *Applied Psychological Measurement*, 17, 29-38.
- 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, P., & Swaminathan, H. (1996). Identification of item that show nonuniform DIF. *Applied Psychological Measurement*, 20(3), 257-274.

- Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement*, 19(1), 23-37.
- Shealy, R., & Stout, W. (1993). A model based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/ DIF. *Psychometrika*, 58(2) (June 1993), 159-194.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27(4), 361-370.