บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

การวัดและประเมินผลทางการศึกษานั้น ข้อสอบที่ใช้วัดจะต้องมีคุณภาพให้ผลที่เชื่อถือได้ คุณสมบัติของข้อสอบที่ดีได้แก่ ความตรง (Validity) ความเที่ยง (Reliability) ความยาก (Difficulty) ความเป็นปรนัย (Objectivity) อำนาจจำแนก (Discrimination) ความจำเพาะเจาะจง (Specificity) และความยุติธรรม (Fairness) (ชูศรี แสงดารารัตน์, 2546, หน้า 1)

จากคุณสมบัติที่ดีของข้อสอบข้างต้น ความยุติธรรมเป็นคุณลักษณะที่สำคัญอย่างหนึ่งของข้อสอบ ในการสอบแต่ละครั้งผู้สอบอาจมีคุณลักษณะแตกต่างกัน เช่น เชื้อชาติ ศาสนา วัฒนธรรม ภูมิลำเนา สังคม เพศ ภาษา อายุ และประสบการณ์ เป็นต้น ผู้สอบดังกล่าวอาจไม่ได้รับความยุติธรรมในการทำข้อสอบ โดยข้อสอบบางข้อมีความลำเอียงเข้าข้างกลุ่มผู้สอบย่อยบางกลุ่ม ซึ่งอาจทำให้เกิดการได้เปรียบเสียเปรียบ ทั้ง ๆ ที่ทำข้อสอบเดียวกัน แสดงว่าข้อสอบที่ขาดความยุติธรรมทำให้ขาดความตรง สาเหตุดังกล่าวอาจเนื่องมาจากข้อสอบไม่ได้วัดความสามารถเป้าหมายที่ต้องการวัด (Target Ability: θ) เพียงอย่างเดียว แต่ยังวัดความสามารถแทรกซ้อนที่ไม่ต้องการวัด (Nuisance Ability: η) อีกด้วย ข้อสอบทุกข้อจะวัดความสามารถเป้าหมาย แต่ข้อสอบทำหน้าที่ต่างกันจะวัดทั้งความสามารถเป้าหมายและความสามารถแทรกซ้อน (Nandakumar, 1993) แสดงว่าถ้าผู้สอบกลุ่มย่อยกลุ่มใดมีความสามารถแทรกซ้อนสูงกว่าก็มีโอกาสตอบข้อสอบถูกมากกว่า ทั้ง ๆ ที่มีความสามารถเป้าหมายที่ต้องการวัดเท่ากัน จึงมีผลทำให้ข้อสอบทำหน้าที่ต่างกัน ข้อสอบทำหน้าที่ต่างกันเมื่อผู้สอบที่มีความสามารถระดับเดียวกันแต่อยู่ต่างกลุ่มกัน มีโอกาสในการตอบข้อสอบถูกได้ต่างกัน (Li & Stout, 1993 cited in Narayanan & Swaminathan, 1996 อ้างถึงใน วลีมาศ แซ่อึ้ง, 2543, หน้า 3) ซึ่งขนาดและทิศทางของการทำหน้าที่ต่างกัน จะแปรเปลี่ยนตามระดับความสามารถที่ต่างกัน เช่น ผู้สอบที่ต้องการวัดความสามารถด้านคณิตศาสตร์ (ความสามารถเป้าหมาย) แต่ข้อสอบมีเนื้อหาเกี่ยวกับกีฬาฟุตบอล (ความสามารถแทรกซ้อน) เมื่อกลุ่มผู้สอบมีความแตกต่างทางเพศ (กลุ่มผู้ชายกับกลุ่มผู้หญิง) กลุ่มผู้สอบเพศชายจะมีความรู้เรื่องกีฬาฟุตบอลสูง ในขณะที่มีความรู้ด้านคณิตศาสตร์เท่ากันกับกลุ่มผู้สอบเพศหญิง ผู้สอบทั้งสองกลุ่มย่อยจะตอบข้อสอบได้ถูกต้องแตกต่างกัน โดยกลุ่มผู้สอบเพศชายจะตอบข้อสอบได้ถูกต้องมากกว่ากลุ่มผู้สอบเพศหญิง

โดยทั่วไปในแบบทดสอบมาตรฐานวัดผลสัมฤทธิ์ทางการเรียน ถ้ามีสัดส่วนของข้อสอบทำหน้าที่ต่างกันร้อยละ 10 ถึง 15 ถือว่าไม่ผิดปกติ แต่ถ้าสัดส่วนของข้อสอบทำหน้าที่ต่างกันร้อยละ 20 ถือว่าเป็นเรื่องผิดพลาดอย่างมาก (Clauser, 1993 cited in Narayanan & Swaminathan, 1994 อ้างถึงใน วลีมาศ แซ่อึ้ง, 2543, หน้า 1) ปัจจุบันนักวิจัยได้ให้ความสนใจการตรวจสอบสัดส่วนของข้อสอบทำหน้าที่ต่างกันมากยิ่งขึ้น ทั้งนี้เนื่องจากแบบทดสอบมาตรฐานที่ใช้ตรวจสอบคุณภาพของผู้เรียน ไม่เพียงเพื่อการนำไปใช้ในการวัดผลสัมฤทธิ์ทางการเรียนเท่านั้น ยังนำไปใช้ด้านอื่น ๆ เช่น คัดเลือกบุคคลเข้าศึกษาต่อ บรรจุเข้าทำงาน เลื่อนขั้นหรือเลื่อนตำแหน่ง เป็นต้น แบบทดสอบที่มีข้อสอบทำหน้าที่ต่างกันจำนวนมาก จะทำให้แบบทดสอบมีคุณภาพต่ำ 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) และข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป (Nonuniform DIF) ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปยังแบ่งได้ 2 ลักษณะ ได้แก่ ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปแบบมีปฏิสัมพันธ์ไม่เป็นลำดับ (Disordinal Interaction) และข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปแบบมีปฏิสัมพันธ์เป็นลำดับ (Ordinal Interaction) (Swaminathan & Rogers, 1990)

กระบวนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้วิธีทางสถิติสามารถตรวจสอบได้ 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) จำแนกตามลักษณะการตรวจให้คะแนนได้เป็น 2 ประเภท คือ ข้อสอบที่ให้คะแนนแบบทวิภาคหรือสองค่า (Dichotomous Scoring) และข้อสอบที่ให้คะแนนแบบพหุวิภาค หรือหลายค่า (Polytomous Scoring) วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแต่ละประเภทยังสามารถจำแนกได้อีก 2 มิติ ได้แก่ มิติลักษณะของตัวแปรเกณฑ์ ซึ่งแบ่งเป็นกลุ่มวิธีที่ใช้คะแนนสังเกตได้ (Observed Score) และกลุ่มวิธีที่ใช้คุณลักษณะแฝงหรือคะแนนของตัวแปรแฝง (Latent Variable) และมิติลักษณะของสถิติวิเคราะห์ ซึ่งแบ่งเป็นกลุ่มที่ใช้สถิติพาราเมตริก (Parametric Approach) และกลุ่มวิธีที่ใช้สถิตินันพาราเมตริก (Nonparametric Approach) (Potenza & Dorans, 1995; Feinstien, 1995)

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนแบบทวิภาค กลุ่มวิธีที่ใช้คะแนนสังเกตได้ ส่วนใหญ่วิเคราะห์ตามทฤษฎีการทดสอบแบบดั้งเดิม (CCT) หรือเรียกว่ากลุ่มที่ไม่ใช้ทฤษฎีการตอบสนองข้อสอบ (Non-IRT Approach) โดยใช้คะแนนรวมของผู้สอบเป็นเกณฑ์ในการจับคู่กลุ่มผู้สอบ วิธีการตรวจสอบที่สำคัญในกลุ่มนี้ที่ใช้สถิติพาราเมตริกได้แก่ การวิเคราะห์ความแปรปรวน (ANOVA) วิธีวิเคราะห์การถดถอยโลจิสติก (Logistic Regression: LR) และวิธีที่ใช้สถิตินันพาราเมตริก ได้แก่ วิธีแปลงค่าความยากของข้อสอบ (Transformed Item Difficulty: TID) วิธีดัชนีมาตรฐาน (Standardization: STND) และวิธีตารางการณ์จร (Contingency Table: CT) ซึ่งในวิธีตารางการณ์จรประกอบด้วย วิธีลอก – ลิเนียร์ (Log – Linear: LL) วิธีไค – สแควร์ (Chi – Square: χ^2) และวิธีแมนเทล – แฮนส์เซล (Mantel – Haenszel: MH) กลุ่มวิธีที่ใช้คุณลักษณะแฝงจะวิเคราะห์บนพื้นฐานของทฤษฎีการตอบสนองข้อสอบ (IRT) สำหรับใช้เป็นเกณฑ์ในการจับคู่กลุ่มผู้สอบ วิธีการตรวจสอบที่สำคัญในกลุ่มนี้ที่ใช้สถิติพาราเมตริก ได้แก่ วิธีวัดพื้นที่ความแตกต่างระหว่างโค้งการตอบสนองข้อสอบ (IRT-D^2) โดยวิธีวัดพื้นที่ของ Raju และวิธีการวัดพื้นที่ของ Kim & Cohen (1991) วิธีเปรียบเทียบค่าพารามิเตอร์ซึ่งในวิธีนี้ประกอบด้วยวิธีการเปรียบเทียบค่าความยาก (Difficulty Shift) วิธีการทดสอบ F วิธีการทดสอบไค-สแควร์ของลอร์ด (Lord's Chi – Square Test) วิธี IRT แบบเทียม (Pseudo - IRT) และวิธีทดสอบอัตราส่วนไลค์ลิฮูดทั่วไป (Likelihood Ratio Test: LR) รูปแบบที่ใช้สถิตินันพาราเมตริก ได้แก่วิธีซิปเทสท์ (SIBTEST) (Millsap & Everson, 1993; Holland & Wainer, 1993; Potenza & Dorans, 1995; Feinstein, 1995)

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ให้คะแนนแบบพหุวิภาค กลุ่มที่ใช้คะแนนสังเกตได้ วิธีที่ใช้ตรวจสอบโดยใช้สถิติพาราเมตริกได้แก่ วิธีการวิเคราะห์ความแปรปรวน วิธีการวิเคราะห์การถดถอยโลจิสติกพหุวิภาค (Polytomous Logistic Regression) วิธีตรวจสอบโดยใช้สถิตินันพาราเมตริกได้แก่ วิธีดัชนีมาตรฐานพหุวิภาค (Polytomous Standardization) และวิธี

แมนเทล – แฮนส์เซล ทั่วไป (General Mantel – Haenszel: GMH) กลุ่มวิธีที่ใช้คุณลักษณะแฝง วิธีที่ใช้ตรวจสอบโดยใช้สถิติพาราเมตริกได้แก่ วิธีอัตราส่วนไลค์ลิฮูดในรูปทั่วไป (General IRT Likelihood Ratio) วิธีการให้คะแนนบางส่วน (Partial Credit Model: PCM) วิธีทดสอบโดยใช้สถิตินันพาราเมตริกได้แก่ วิธีซิปเทสท์พหุวิภาค (Polytomous SIBTEST) และวิธีการให้คะแนนบางส่วนทั่วไป (Generallized Partial Credit Model: GPCM) (Potenza & Dorans, 1995; Feinstein, 1995)

วิธีการทางสถิติที่ใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ที่นิยมใช้อย่างมากคือ วิธีแมนเทล-แฮนส์เซล วิธีซิปเทสน์ วิธีทำให้เป็นมาตรฐาน และวิธีลอจิสติก (David & Lori, 2000, pp. 263-280)

วิธีซิปเทสท์ (SIBTEST: Simulatancous Item Bias Test) พัฒนาโดยเชียลีและสเตาส์ (Shealy & Stout, 1993) ใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ การทำหน้าที่ต่างกันของแบบสอบ (Differential Test Functioning: DTF) และการทำหน้าที่ต่างกันของหมวดข้อสอบ (Differential Bundle Functioning: DBF) วิธีนี้สามารถวิเคราะห์ได้ทั้งแบบทดสอบเอกมิติ (Unidimensional Test) และแบบทดสอบพหุมิติ (Multidimensional Test) วิธีซิปเทสท์ใช้สถิติทดสอบแบบนันพาราเมตริก ซึ่งพัฒนาบนพื้นฐานของทฤษฎี IRT ชนิดพหุมิติ แต่ไม่ต้องใช้ฟังก์ชันการตอบสนองข้อสอบหรือการประมาณค่าความสามารถแฝง วิธีซิปเทสท์ได้รับการออกแบบมาสำหรับตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบมีทิศทางเดียว (Unidirectional DIF) ดังนั้นจึงไม่ไวในการตรวจสอบการทำหน้าที่ด่างกันของข้อสอบที่ไม่มีทิศทาง (Nondirectional DIF) (Li & Stout, 1996 cited in Shealy & Stout, 1993) จุดเด่นของวิธีซิปเทสท์ คือ สามารถคำนวณได้ง่าย ไม่ซับซ้อน ประหยัดค่าใช้จ่าย ไม่จำเป็นต้องใช้กลุ่มตัวอย่างที่มีขนาดใหญ่ และใช้สถิติทดสอบนัยสำคัญ นอกจากนี้ยังสามารถนำไปประยุกต์ใช้กับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่มีการให้คะแนนแบบพหุวิภาคได้

ศุภวัฒน์ มะลิเผือก (2548) ได้เปรียบเทียบค่าความเที่ยงของแบบทดสอบก่อนตัดข้อสอบทำหน้าที่ต่างกัน และหลังตัดข้อสอบทำหน้าที่ต่างกันออก 30 % แล้วปรับขยายสัดส่วนจำนวนข้อให้เท่ากับก่อนตัดข้อสอบทำหน้าที่ต่างกัน โดยใช้จำนวนกลุ่มตัวอย่าง 2,000 คน คำนวณค่าความเที่ยงของแบบทดสอบโดยใช้สูตรของ สเปียร์แมน บราว (Spearman Brown) แบบทดสอบที่มีข้อสอบทำหน้าที่ต่างกัน 30% และเมื่อดัดข้อสอบทำหน้าที่ต่างกันออก พบว่าค่าความเที่ยงของแบบทดสอบทั้งสองฉบับแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05

จากการศึกษางานวิจัยส่วนใหญ่ปรากฏว่าจะเป็นการเปรียบเทียบวิธีการตรวจสอบข้อสอบทำหน้าที่ต่างกันว่าวิธีการใดมีประสิทธิภาพมากที่สุด แต่ยังไม่พบงานวิจัยเกี่ยวกับคุณภาพ

ของแบบทดสอบด้านความเที่ยงเมื่อมีข้อสอบทำหน้าที่ต่างกันระดับต่าง ๆ จะมีผลต่อคุณภาพของแบบทดสอบด้านความเที่ยงอย่างไร เพราะจากการศึกษาเกี่ยวกับคุณภาพของแบบทดสอบ พบว่าหัวใจสำคัญของแบบทดสอบ คือ ความเที่ยงและความตรง

จากเหตุผลดังกล่าวข้างต้นผู้วิจัยจึงสนใจศึกษาการทำหน้าที่ต่างกันของข้อสอบโดยวิธีซิปเทสท์ เพราะเป็นวิธีที่เป็นมาตรฐาน เชื่อถือได้และยังเป็นการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่พิจารณาจากความแตกต่างของคะแนนจริงระหว่างผู้สอบที่มีความสามารถระดับเดียวกัน (นพมาศ พิพัฒนสุข, 2541) ในการวิจัยครั้งนี้ผู้วิจัยได้เลือกศึกษาตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ วิชาภาษาไทย ชั้นประถมศึกษาปีที่ 6 (กระทรวงศึกษาธิการ, 2546 อ้างถึงใน ปิยะทิพย์ ดินวร, 2549, หน้า 5) ซึ่งเป็นข้อสอบที่มีการให้คะแนนแบบทวิภาคหรือสองค่า (Dichotomous Scoring) ที่เลือกศึกษาวิชาภาษาไทยเพราะเป็นวิชาที่มักพบความลำเอียงต่อเพศอย่างชัดเจน โดยส่วนใหญ่จะลำเอียงเข้าข้างผู้สอบเพศหญิง และข้อสอบวิชาภาษาไทย เป็นวิชาที่มีวัฒนธรรมทางภาษาเข้ามาเกี่ยวข้องด้วย จึงทำให้ข้อความและถ้อยคำลำเอียงได้ง่าย ซึ่งสอดคล้องกับ ลิน (Linn, 1993) ซึ่งได้ศึกษาพบว่าในแบบทดสอบ SAT จำนวน 7 ฉบับ พบข้อสอบทำหน้าที่ต่างกัน 29 ข้อ เป็นข้อสอบทางภาษาจำนวน 25 ข้อ นอกจากนี้ กาญจนา วัธนสุนทร (2537) พบว่าข้อสอบทำหน้าที่ต่างกันมากกว่าครึ่งในวิชาภาษาอังกฤษลำเอียงเข้าข้างเพศหญิง และเกอเรย์ (Giray, 1995) พบว่าผู้สอบเพศหญิงจะมีความสามารถในด้านการใช้ภาษา หรือเกี่ยวกับคำศัพท์ได้ดีกว่าเพศชาย

ในการวิจัยครั้งนี้ผู้วิจัย สนใจที่จะเปรียบเทียบค่าความเที่ยงของแบบทดสอบ 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 มาตรวจสอบการทำหน้าที่ต่างกันโดยวิธีซิปเทสท์ เมื่อพบข้อสอบที่ทำหน้าที่ต่างกันแล้ว นำข้อสอบดังกล่าวมารวมกับข้อสอบที่ไม่ทำหน้าที่ต่างกัน ให้ได้สัดส่วนข้อสอบที่ทำหน้าที่ต่างกัน 7 ระดับ แล้วทำการวิเคราะห์หาค่าความเที่ยงของแบบทดสอบ

ผลการวิจัยจะทำให้ทราบถึงความเที่ยงของแบบทดสอบว่าเป็นอย่างไร ซึ่งข้อค้นพบที่ได้นี้จะเป็นประโยชน์ในการตรวจสอบคุณภาพของข้อสอบ ของแบบทดสอบ และปรับปรุงคุณภาพของแบบทดสอบ นอกจากนี้ยังสามารถนำไปเป็นสารสนเทศในการปรับปรุงแบบทดสอบให้มีคุณภาพดียิ่งขึ้น

วัตถุประสงค์ของการวิจัย

1. เพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ จากแบบทดสอบผลสัมฤทธิ์ทางการเรียนระดับชาติ วิชาภาษาไทย ชั้นประถมศึกษาปีที่ 6 ปีการศึกษา 2546

2. เพื่อเปรียบเทียบค่าความเที่ยงของแบบทดสอบ ภายใต้เงื่อนไขจำนวนข้อสอบทำหน้าที่ต่างกันแตกต่างกัน 7 เงื่อนไข ได้แก่ 0%, 5 %, 10%, 15%, 20 %, 25% และ 30%

คำถามการวิจัย

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

สมมติฐานของการวิจัย

การเปรียบเทียบค่าความเที่ยงของแบบทดสอบที่มีจำนวนข้อสอบทำหน้าที่ต่างกันแตกต่างกัน ผู้วิจัยได้ศึกษางานวิจัยเกี่ยวกับคุณภาพของแบบทดสอบด้านความเที่ยงจากงานวิจัยของ อารี วัชรโสคถิกุล(2543) ที่เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยใช้รูปแบบและวิธีการที่ต่างกัน พบว่า ค่าความเที่ยงของแบบทดสอบหลังจากตัดข้อสอบทำหน้าที่ต่างกันออก แล้วทำการปรับขยายค่าความเที่ยงของแบบทดสอบให้มีจำนวนข้อสอบเท่ากัน เมื่อตรวจสอบด้วยวิธีซิปเทสน์ พบว่า ค่าความเที่ยงแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ซึ่งสอดคล้องกับผลงานวิจัยของ รักชนก ยี่สุ่นศรี (2544) ได้วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบและแบบสอบด้วยกระบวนการ ดี เอฟ ไอ ที สำหรับแบบสอบคัดเลือกบุคคลเข้าศึกษาในระดับอุดมศึกษาวิชาภาษาอังกฤษและวิชาคณิตศาสตร์ พบว่าเมื่อเปรียบเทียบคุณภาพของแบบสอบก่อนและหลังตัดข้อสอบทำหน้าที่ต่างกันออก แบบทดสอบฉบับหลังที่ตัดข้อสอบทำหน้าที่ต่างกันออกส่วนใหญ่มีค่าความเที่ยงลดลง และยังสอดคล้องกับงานวิจัยของ ศุภวัฒน์ มะลิเผือก (2548) ที่ศึกษาเกี่ยวกับอิทธิพลการทำหน้าที่ต่างกันของข้อสอบ ที่ส่งผลต่อคุณภาพของแบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ ชั้นประถมศึกษาปีที่ 6 พบว่าค่าความเที่ยงของแบบทดสอบฉบับก่อนและหลังคัดข้อสอบทำหน้าที่ต่างกันออกจากแบบทดสอบ มีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ดังนั้นการวิจัยนี้จึงตั้งสมมติฐานการวิจัย ดังนี้

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

ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

1. ได้ข้อค้นพบเกี่ยวกับคุณภาพของแบบทดสอบด้านความเที่ยงของแบบทดสอบ เมื่อไม่มี ข้อสอบทำหน้าที่ต่างกันและมีข้อสอบทำหน้าที่ต่างกัน 5 %, 10%, 15%, 20 %, 25% และ 30%

2. ได้สารสนเทศในการปรับปรุงแบบทดสอบให้มีคุณภาพดียิ่งขึ้น รวมทั้งเป็นแนวทางในการปรับปรุงข้อสอบให้มีความยุติธรรม

ขอบเขตของการวิจัย

1. ประชากรและกลุ่มตัวอย่าง

ในการวิจัยครั้งนี้ใช้ข้อมูลทุติยภูมิ เป็นผลการตอบข้อสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ วิชาภาษาไทย ชั้นประถมศึกษาปีที่ 6 ปีการศึกษา 2546 ของนักเรียนสังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน กระทรวงศึกษาธิการ

ประชากร เป็นนักเรียนชั้นประถมศึกษาปีที่ 6 สังกัดคณะกรรมการการศึกษาขั้นพื้นฐาน กระทรวงศึกษาธิการ ปีการศึกษา 2546 ทั้งประเทศ ที่เข้าสอบวิชาภาษาไทย จำนวน 750,978 คน

กลุ่มตัวอย่าง เป็นนักเรียนชั้นประถมศึกษาปีที่ 6 สังกัดคณะกรรมการการศึกษาขั้นพื้นฐาน กระทรวงศึกษาธิการ ได้มาโดยการสุ่มแบบแบ่งชั้นไม่กำหนดสัดส่วน ใช้ระดับความสามารถ (คะแนนรวม) เป็นชั้นในการแบ่ง แบ่งเป็น 3 ระดับ คือ ดี พอใช้ และปรับปรุง ใช้นักเรียนเป็นหน่วยในการสุ่ม สุ่มมาจำนวน 2,000 คน แบ่งเป็นนักเรียนชาย 1,000 คน และนักเรียนหญิง 1,000 คน

2. วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบใช้วิธีซิบเทสท์

3. ตัวแปรที่ใช้ในการวิจัยประกอบด้วย

ตัวแปรอิสระ มี 1 ตัว ได้แก่ จำนวนข้อสอบทำหน้าที่ต่างกันในแบบทดสอบมี 7 เงื่อนไข ได้แก่ 0%, 5 %, 10%, 15%, 20 %, 25% และ 30%

ตัวแปรตาม มี 1 ตัว ได้แก่ ค่าความเที่ยงของแบบทดสอบ

ข้อจำกัดของการวิจัย

งานวิจัยนี้เป็นการศึกษาข้อมูลจริงจากแบบทดสอบผลสัมฤทธิ์ทางการเรียนระดับชาติ วิชาภาษาไทย ชั้นประถมศึกษาปีที่ 6 ปีการศึกษา 2546 ซึ่งมีจำนวนข้อสอบในแบบทดสอบเพียง 40 ข้อ ดังนั้น ผลการวิจัยในกรณีเช่นนี้จึงไม่สามารถสรุปไปยังแบบทดสอบที่มีความยาวแตกต่างกันออกไป และไม่สามารถกำหนดจำนวนข้อสอบทำหน้าที่ต่างกันได้ตามเปอร์เซ็นต์ที่ระบุใน

ตัวแปรอิสระ แต่ใช้ค่าประมาณที่ใกล้เคียง ผลการวิจัยนี้จึงเป็นแนวทางในการทำวิจัยให้ละเอียดลึกซึ้งในครั้งต่อไป

นิยามศัพท์เฉพาะ

1. **การทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning: DIF)** หมายถึง ข้อสอบที่ทำให้ผู้สอบที่มีความสามารถหรือคุณลักษณะที่มุ่งวัดเท่าเทียมกันมีโอกาสในการตอบข้อสอบถูกต้องแตกต่างกัน เนื่องจากผู้สอบอยู่ในกลุ่มย่อยต่างกัน

2. **ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ** หมายถึง ข้อสอบที่ตรวจสอบพบว่ามีการทำหน้าที่ต่างกัน โดยใช้วิธีวิธีซิปเทสท์ในการตรวจสอบ

3. **กลุ่มอ้างอิง (Reference Group)** หมายถึง กลุ่มผู้สอบที่น่าจะได้เปรียบในการตอบข้อสอบทำหน้าที่ต่างกัน ในการวิจัยครั้งนี้กลุ่มอ้างอิง คือ กลุ่มผู้สอบเพศหญิง

4. **กลุ่มเปรียบเทียบ (Focal Group)** หมายถึง กลุ่มผู้สอบที่น่าจะเสียเปรียบในการตอบข้อสอบทำหน้าที่ต่างกัน ในงานวิจัยครั้งนี้กลุ่มเปรียบเทียบ คือ กลุ่มผู้สอบเพศชาย

5. **ความเที่ยงของแบบทดสอบ (Test Reliability)** หมายถึง คุณภาพของแบบทดสอบที่ให้ผลการวัดมีความคงที่หรือคงเส้นคงวา ไม่ว่าจะวัดกี่ครั้งภายใต้สภาวะการแบบใด ก็ให้ผลการวัดคงเดิม สำหรับการวิจัยในครั้งนี้หาค่าความเที่ยงแบบความสอดคล้องภายใน (Internal Consistency) โดยใช้วิธีการคำนวณค่าความเที่ยงแบบสัมประสิทธิ์แอลฟาของครอนบัค

6. **วิธีซิปเทสท์ (SIBTEST: Simulataneous Item Bias Test)** หมายถึง วิธีการที่ใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (DIF) และการทำหน้าที่ต่างกันของกลุ่มข้อสอบ (Differential Bundle Functioning: DBF) พัฒนาโดยเชียลี และสเตาส์ (Shealy & Stout, 1993) วิธีนี้สามารถวิเคราะห์ได้ทั้งแบบสอบเอกมิติ (Unidimensional Test) และแบบสอบพหุมิติ (Multidimensional Test) วิธีซิปเทสท์ใช้สถิติทดสอบแบบนันพาราเมตริก (Nonparametric) ซึ่งพัฒนาบนพื้นฐานของทฤษฎี IRT

7. **โปรแกรมซิปเทสท์** หมายถึง โปรแกรมที่ใช้วิเคราะห์การทำหน้าที่ต่างกันของข้อสอบด้วยวิธีซิปเทสท์ โดยวิเคราะห์ข้อสอบมีลักษณะการให้คะแนนแบบ 0,1

8. **แบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ (National Achievement Test) ชั้นประถมศึกษาปีที่ 6** หมายถึง แบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนที่สร้างขึ้นตามหลักสูตรประถมศึกษา พ.ศ. 2521 (ปรับปรุง 2533) และดำเนินการสอบในปีการศึกษา 2546 การวิจัยครั้งนี้ใช้เฉพาะแบบทดสอบวิชาภาษาไทย

9. **การตัดข้อสอบออกจากแบบทดสอบ** หมายถึง การไม่นำเอาข้อสอบ ข้อที่พบว่าทำหน้าที่ต่างกันมาคิดคะแนนให้กับผู้สอบ

10. **จำนวนข้อสอบที่ทำหน้าที่ต่างกัน** หมายถึง เปอร์เซ็นต์ของการทำหน้าที่ต่างกันของข้อสอบ โดยเทียบจากจำนวนของแบบทดสอบทั้งหมด