

Detecting Differential Item Functioning in Multidimensional Tests: A Comparison of the Relative Efficiency of Restricted Factor Analysis and Logistic Regression Methods

Piyathip Tinnaworn

M.R. Somporn Sudhasani and Seree Chadcham

Burapha University, Thailand

ABSTRACT

The purpose of this research was to compare the relative efficiency of Restricted Factor Analysis and Logistic Regression methods for detecting differential item functioning (DIF) in multidimensional tests under the 18 conditions ($3 \times 3 \times 2$), 3 sizes of samples (2,000 samples, 1,000 samples, 300 samples), 3 sizes of test length (40 items, 30 items, 20 items) and 2 matching criterions (total test score = TS, subtest score = SS), by using SIBTEST method for the criteria. The sample consisted of 2,000 Grade-six students under the Basic Education Commission Office, in academic year 2003 who entered the national achievement test on the Thai Language. Descriptive statistics were derived by using SPSS, while a second-order confirmatory factor analysis for construct validity was conducted by applying LISREL 8.50. Detecting DIF by using the SIBTEST, LISREL 8.50, and SPSS.

The results were as follow:

1. The Logistic Regression (LR) method was found having the same efficiency as Restricted Factor Analysis (RFA) method in detecting DIF for multidimensional test. In these two conditions, it was 1,000 samples $\times$ 20 items $\times$ SS and 300 samples $\times$ 20 items $\times$ SS.

2. The LR method was found more efficient than RFA method for detecting DIF for multidimensional test in the 16 conditions: 1) 2,000 $\times$ 40 $\times$ TS 2) 2,000 $\times$ 40 $\times$ SS 3) 2,000 $\times$ 30 $\times$ TS 4) 2,000 $\times$ 30 $\times$ SS 5) 2,000 $\times$ 20 $\times$ TS 6) 2,000 $\times$ 20 $\times$ SS 7) 1,000 $\times$ 40 $\times$ TS 8) 1,000 $\times$ 40 $\times$ SS 9) 1,000 $\times$ 30 $\times$ TS 10) 1,000 $\times$ 30 $\times$ SS 11) 1,000 $\times$ 20 $\times$ TS 12) 300 $\times$ 40 $\times$ TS 13) 300 $\times$ 40 $\times$ SS 14) 300 $\times$ 30 $\times$ TS 15) 300 $\times$ 30 $\times$ SS and 16) 300 $\times$ 20 $\times$ TS. This was because of LR method had less the Type I Error Rate than RFA method.

Based on a master's thesis completed at the Department of Educational Research and Measurement, Burapha University, under the supervision of Assoc. Prof. M.R. Somporn Sudhasani, Ph.D., and Assist. Prof. Seree Chadcham, Ph.D.

การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ: การเปรียบเทียบประสิทธิภาพระหว่างวิธีการวิเคราะห์องค์ประกอบจำกัด กับวิธีถดถอยโลจิสติก

ปิยะทิพย์ ดินวร

ม.ร.ว. สมพร สุทัศน์ีย์ และ เสรี ชัดเข้ม

มหาวิทยาลัยบูรพา

บทคัดย่อ

การวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติระหว่างวิธีการวิเคราะห์องค์ประกอบจำกัดกับวิธีถดถอยโลจิสติก ภายใต้เงื่อนไข 18 เงื่อนไข ($3 \times 3 \times 2$) คือ ขนาดของกลุ่มตัวอย่าง 3 ขนาด (2,000 คน 1,000 คน และ 300 คน) ความยาวของแบบทดสอบ 3 ขนาด (40 ข้อ 30 ข้อ และ 20 ข้อ) และ เกณฑ์การจับคู่ 2 เกณฑ์ (คะแนนรวมทั้งฉบับ = TS และ คะแนนแบบทดสอบย่อย = SS) เมื่อใช้ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยวิธีซิปเทสท์เป็นเกณฑ์สำหรับการเปรียบเทียบประสิทธิภาพ กลุ่มตัวอย่างเป็นนักเรียนชั้นประถมศึกษาปีที่ 6 สังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน ปีการศึกษา 2546 ที่เข้าสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติวิชาภาษาไทย จำนวน 2,000 คน วิเคราะห์ค่าสถิติพื้นฐานโดยใช้โปรแกรม SPSS ตรวจสอบความตรงเชิงโครงสร้างด้วยการวิเคราะห์องค์ประกอบเชิงยืนยันอันดับสองโดยใช้โปรแกรม LISREL 8.50 และ ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้โปรแกรม SIBTEST LISREL 8.50 และ SPSS ผลการวิจัยปรากฏว่า

1. วิธีถดถอยโลจิสติก (LR) มีประสิทธิภาพไม่แตกต่างกันกับวิธีวิเคราะห์องค์ประกอบจำกัด (RFA) ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ 2 เงื่อนไข (ขนาดของกลุ่มตัวอย่าง $\times$ ความยาวของแบบทดสอบ $\times$ เกณฑ์การจับคู่) คือ 1,000 คน $\times$ 20 ข้อ $\times$ SS และ 300 คน $\times$ 20 ข้อ $\times$ SS

2. วิธี LR มีประสิทธิภาพมากกว่าวิธี RFA ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ 16 เงื่อนไข คือ 1) 2,000 คน $\times$ 40 ข้อ $\times$ TS 2) 2,000 คน $\times$ 40 ข้อ $\times$ SS 3) 2,000 คน $\times$ 30 ข้อ $\times$ TS 4) 2,000 คน $\times$ 30 ข้อ $\times$ SS 5) 2,000 คน $\times$ 20 ข้อ $\times$ TS 6) 2,000 คน $\times$ 20 ข้อ $\times$ SS 7) 1,000 คน $\times$ 40 ข้อ $\times$ TS 8) 1,000 คน $\times$ 40 ข้อ $\times$ SS 9) 1,000 คน $\times$ 30 ข้อ $\times$ TS 10) 1,000 คน $\times$ 30 ข้อ $\times$ SS 11) 1,000 คน $\times$ 20 ข้อ $\times$ TS 12) 300 คน $\times$ 40 ข้อ $\times$ TS 13) 300 คน $\times$ 40 ข้อ $\times$ SS 14) 300 คน $\times$ 30 ข้อ $\times$ TS 15) 300 คน $\times$ 30 ข้อ $\times$ SS และ 16) 300 คน $\times$ 20 ข้อ $\times$ TS เนื่องจาก วิธี LR มีอัตราความคลาดเคลื่อนประเภทที่ 1 น้อยกว่าวิธี RFA อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

ส่วนหนึ่งของวิทยานิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีการวัดทางการศึกษา มหาวิทยาลัยบูรพา ปีการศึกษา 2549 อาจารย์ที่ปรึกษาหลัก รศ.ดร. ม.ร.ว. สมพร สุทัศน์ีย์ และ อาจารย์ที่ปรึกษาร่วม ผศ.ดร. เสรี ชัดเข้ม

ความนำ

แบบทดสอบเป็นเครื่องมือหลักที่ใช้ในการวัดและประเมินผลการศึกษา เพื่อตรวจสอบว่าผู้เข้าสอบมีคุณลักษณะแฝงหรือความสามารถอยู่ในระดับใด ดังนั้น การสร้างและการตรวจสอบคุณภาพของแบบทดสอบต้องคำนึงถึงความตรง (Validity) เป็นสำคัญ เนื่องจากความตรงเป็นคุณสมบัติของแบบทดสอบที่แสดงว่า คะแนนจากแบบทดสอบสามารถสรุปอ้างอิงไปยังสิ่งที่ต้องการวัดได้อย่างเหมาะสม มีความหมาย และเป็นประโยชน์ (Wiersma & Jurs, 1990) ถ้าผลการวัดได้ค่าที่ใกล้เคียงกับคุณลักษณะที่แท้จริงมากเพียงใด ก็ถือว่าการวัดมีความตรงมากขึ้นเพียงนั้น (ศิริชัย กาญจนวาสี, 2545: 103) นักวัดผลการศึกษาตรวจสอบหลักฐานที่แสดงถึงความตรงของแบบทดสอบ 3 ด้าน คือ (1) ความตรงเชิงเนื้อหา (Content validity) (2) ความตรงเชิงเกณฑ์สัมพันธ์ (Criterion related validity) และ (3) ความตรงเชิงโครงสร้าง (Construct validity) ส่วนการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ (Differential Item Functioning) เป็นการตรวจสอบคุณภาพของแบบทดสอบด้านความตรงเชิงโครงสร้าง (Mazor et al., 1992; Kim & Cohen, 1994)

การทำหน้าที่ต่างกันของข้อสอบ (Differential item functioning: DIF) มีผู้ให้นิยามไว้หลายความหมายด้วยกัน แต่ที่ยอมรับกันอย่างกว้างขวางก็คือ ข้อสอบทำหน้าที่ต่างกันภายใต้เงื่อนไขผู้เข้าสอบมีความสามารถเท่ากัน แต่มาจากกลุ่มผู้เข้าสอบย่อยที่มีลักษณะต่างกัน มีความน่าจะเป็นในการตอบข้อสอบข้อนั้นไม่เท่ากัน (เสรี ชัดแจ้ง, 2540: 42) หลักการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบใช้การเปรียบเทียบผลการตอบข้อสอบระหว่างกลุ่มผู้เข้าสอบย่อย 2 กลุ่ม ที่มีความสามารถระดับเดียวกัน โดยผู้เข้าสอบกลุ่มหนึ่งเป็นตัวแทนกลุ่มหลักในประชากรเรียกว่า “กลุ่มอ้างอิง” (Reference group: R) ซึ่งเป็นกลุ่มพื้นฐาน ส่วนผู้เข้าสอบอีกกลุ่มหนึ่งเป็นตัวแทนกลุ่มรองในประชากรเรียกว่า “กลุ่มเปรียบเทียบ” (Focal group: F) ซึ่งตามปกติแล้วเป็นกลุ่มผู้เข้าสอบที่สนใจจะศึกษาการทำหน้าที่ต่างกันของข้อสอบ (Angoff, 1993) ถ้าข้อสอบทำหน้าที่ต่างกันแล้วโอกาสตอบข้อสอบถูกของผู้เข้าสอบทั้งสองกลุ่มจะไม่เท่ากัน โดยคาดว่าผู้เข้าสอบกลุ่มอ้างอิงจะได้เปรียบในการตอบข้อสอบ ส่วนผู้เข้าสอบกลุ่มเปรียบเทียบจะเสียเปรียบในการตอบข้อสอบ

เงื่อนไขสำคัญในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ คือ การใช้เกณฑ์การจับคู่ (Matching criteria) เปรียบเทียบระดับความสามารถของผู้เข้าสอบทั้งสองกลุ่มที่นิยมใช้กันมี 2 แนวทางใหญ่ คือ แนวทางแรกเป็นการตรวจสอบโดยใช้เกณฑ์ภายนอก (External criteria) โดยการนำคะแนนจากการตอบข้อสอบไปสัมพันธ์กับเกณฑ์ที่ได้จากภายนอกแบบทดสอบ เช่น เกรดเฉลี่ยสะสมหรือคะแนนจากชิ้นงานที่ทำ เป็นต้น แต่จุดอ่อนของแนวทางนี้คือการยากที่จะหาเกณฑ์ภายนอกที่มีความตรงและยุติธรรม ซึ่งถ้าเกณฑ์การจับคู่ขาดคุณสมบัติเรื่องนี้จะทำให้ผลการตรวจสอบเกิดความคลาดเคลื่อนได้ ส่วนแนวทางที่สองเป็นการตรวจสอบโดยใช้เกณฑ์ภายใน (Internal criteria) โดยการนำวิธีการทางสถิติมา

ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เป็นการพิจารณาจากโครงสร้างภายในของแบบทดสอบเป็นหลัก (Rudner et al., 1980) ซึ่งในปัจจุบันเป็นแนวทางที่มีผู้นิยมใช้กันมาก

วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้พัฒนามาเรื่อย ๆ ปัจจุบันวิธีการวิเคราะห์องค์ประกอบจำกัด (Restricted Factor Analysis Method: RFA) ซึ่งพัฒนาขึ้นโดยออร์ท (Oort, 1998) เป็นวิธีการใหม่ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ วิธีนี้จัดอยู่ในกลุ่มวิธีไค-สแควร์ ซึ่งออร์ท (Oort, 1998: 107-124) ได้ศึกษาเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธีการตอบสนองข้อสอบแบบหนึ่งพารามิเตอร์ กับวิธีการวิเคราะห์องค์ประกอบจำกัด โดยใช้สถานการณ์จำลองสร้างข้อสอบ 3 แบบ คือ ข้อสอบทำหน้าที่ต่างกันมาก ข้อสอบทำหน้าที่ต่างกันน้อย และข้อสอบทำหน้าที่ไม่ต่างกัน และกำหนดขนาดกลุ่มตัวอย่าง 2 ขนาด คือ กลุ่มตัวอย่างขนาดใหญ่ (2,000 คน) และขนาดเล็ก (200 คน) ผลปรากฏว่า วิธีการวิเคราะห์องค์ประกอบจำกัด มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ใกล้เคียงกับวิธีการตอบสนองข้อสอบแบบหนึ่งพารามิเตอร์ แต่ในกรณีกลุ่มตัวอย่างขนาดเล็ก วิธีการวิเคราะห์องค์ประกอบจำกัด มีประสิทธิภาพดีกว่าวิธีการตอบสนองข้อสอบแบบหนึ่งพารามิเตอร์ ในทุกเงื่อนไขของการทดสอบ นอกจากนี้ วิธีการวิเคราะห์องค์ประกอบจำกัดยังประหยัดเวลาและค่าใช้จ่ายในการวิเคราะห์ข้อมูล ส่วนนิคม กิรติวางกูร (2542) ได้ศึกษาผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีการวิเคราะห์องค์ประกอบจำกัด วิธีแมนเทิล-แฮนส์เซล และวิธีการตอบสนองข้อสอบแบบสองพารามิเตอร์โดยใช้สถานการณ์จำลอง พบว่า โดยภาพรวมวิธีการวิเคราะห์องค์ประกอบจำกัดมีประสิทธิภาพในการตรวจสอบสูงที่สุด รองลงมาวิธีแมนเทิล-แฮนส์เซล และวิธีการตอบสนองข้อสอบแบบสองพารามิเตอร์ และวิธีการตอบสนองข้อสอบแบบสองพารามิเตอร์ มีอัตราความคลาดเคลื่อนประเภทที่ 1 สูงกว่าวิธีแมนเทิล-แฮนส์เซล และวิธีการวิเคราะห์องค์ประกอบจำกัด

สำหรับการศึกษาเกี่ยวกับปัจจัยที่ส่งผลต่อประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ปรากฏว่า ข้อค้นพบที่ได้มีความสอดคล้องกัน กล่าวคือ เมื่อกลุ่มตัวอย่างมีขนาดใหญ่ขึ้น อัตราการตรวจพบข้อสอบทำหน้าที่ต่างกันจะเพิ่มขึ้น (Roger & Swaminathan, 1993) แต่ในเรื่องขนาดของกลุ่มตัวอย่างยังไม่ชัดเจนว่า ต้องใช้กลุ่มตัวอย่างขนาดเท่าใดจึงจะเพียงพอ เมเซอร์และคณะ (Mazor et al., 1992) กล่าวว่า ขนาดของกลุ่มตัวอย่างที่เหมาะสมสำหรับ วิธีแมนเทิล-แฮนส์เซล ควรใช้ระหว่าง 100 ถึง 300 คน สำหรับกลุ่มใดกลุ่มหนึ่งหรือทั้งสองกลุ่ม (กลุ่มอ้างอิง และ กลุ่มเปรียบเทียบ) นารายันัน และสวามินาธาน (Narayanan & Swaminathan, 1994) ได้เสนอแนะว่า โดยทั่วไปใช้กลุ่มตัวอย่างขนาดกลุ่มละ 300 คน ก็เพียงพอที่จะตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้อย่างมีประสิทธิภาพ และพบว่าอัตราการตรวจพบข้อสอบทำหน้าที่ต่างกันระหว่างวิธีแมนเทิล-แฮนส์เซล กับวิธีชิปเทสต์ จะได้รับผลกระทบจากกลุ่มเปรียบเทียบที่มีขนาดเล็กมากกว่ากลุ่มอ้างอิงที่มีขนาดเล็ก นอกจากนี้ เคลาเซอร์ และคณะ (Clauser et al., 1993) ยังเสนอแนะว่า กลุ่มตัวอย่างที่ใช้ในการวิเคราะห์ด้วยวิธีแมนเทิล-แฮนส์เซล ควรอยู่ระหว่าง 200 ถึง 1,000 คน ยกเว้นในการวิเคราะห์บางเงื่อนไข การใช้กลุ่ม

ตัวอย่าง 200 คน ในกลุ่มใดกลุ่มหนึ่งอาจจะไม่เพียงพอ แต่ถ้าใช้กลุ่มตัวอย่างขนาดใหญ่จะทำให้ได้ผลการตรวจสอบที่ดีกว่า จากที่กล่าวมาอาจกล่าวได้ว่าขนาดกลุ่มตัวอย่างน่าจะมีผลกระทบต่อประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

การศึกษาของคิมและโคเฮน (Kim & Cohen, 1994) และ อัททาโรและมิลล์แซพ (Uttaro & Millsap, 1994) พบว่า ความยาวของแบบทดสอบมีผลกระทบต่ออัตราการตรวจพบข้อสอบทำหน้าที่ต่างกัน และจากงานวิจัยของนารายานัน และสวามินาทาน (Narayanan & Swaminathan, 1996) พบว่า แบบทดสอบขนาด 40 ข้อ เป็นแบบทดสอบที่มีความยาวเพียงพอสำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ แต่ โรเจอร์ และสวามินาทาน (Roger & Swaminathan, 1993) พบว่า เมื่อเพิ่มขนาดกลุ่มตัวอย่างจะทำให้อัตราการตรวจพบข้อสอบทำหน้าที่ต่างกันจะเพิ่มขึ้น ส่วนขนาดของแบบทดสอบและการกระจายของคะแนนไม่มีผลกระทบต่ออัตราการตรวจพบข้อสอบทำหน้าที่ต่างกัน ดังนั้นเพื่อเป็นการยืนยันข้อค้นพบให้มีความชัดเจนมากยิ่งขึ้น จึงควรศึกษาความยาวของแบบทดสอบว่ามีผลต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบหรือไม่

นอกจากนี้ยังมีการศึกษาปัจจัยอื่นๆ ที่มีผลกระทบต่อประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ได้แก่ เกณฑ์การจับคู่ (Matching criteria) เป็นการหาเกณฑ์การจับคู่เปรียบเทียบที่เหมาะสมสำหรับตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ที่ผ่านมามีการศึกษาวិธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ใช้ทั้งเกณฑ์ภายนอกและเกณฑ์ภายใน เช่น นพมาศ พิพัฒน์สุข (2541) ได้เปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธีแมนเทิล-แฮนส์เซล กับวิธีดออยโลจิสติก เมื่อใช้เกณฑ์ภายในเป็นเกณฑ์การจับคู่ พบว่า เมื่อใช้เกณฑ์การจับคู่คะแนนรวมทั้งฉบับ วิธีแมนเทิล-แฮนส์เซล มีประสิทธิภาพมากกว่าวิธีดออยโลจิสติก และเมื่อใช้เกณฑ์การจับคู่คะแนนแบบทดสอบย่อย วิธีทั้งสองมีประสิทธิภาพไม่แตกต่างกัน สำหรับวิธีดออยโลจิสติกมีประสิทธิภาพมากกว่าวิธีแมนเทิล-แฮนส์เซล เมื่อใช้เกณฑ์การจับคู่คะแนนจากแบบทดสอบย่อยหลายฉบับ ซึ่งสอดคล้องกับผลการศึกษาของเคลาเซอร์ และคณะ (Clauser et al., 1996: 202-214) ที่ศึกษาเกณฑ์การจับคู่ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ โดยเปรียบเทียบผลการตรวจสอบข้อสอบทำหน้าที่ต่างกันเมื่อใช้เกณฑ์การจับคู่ 3 เกณฑ์ ได้แก่ คะแนนรวมทั้งฉบับ คะแนนแบบทดสอบย่อย และคะแนนจากแบบทดสอบย่อยหลายฉบับ ระหว่างวิธีแมนเทิล-แฮนส์เซลกับวิธีดออยโลจิสติก ทั้งในข้อมูลจริงและข้อมูลจำลอง พบว่า ในแบบทดสอบพหุมิติ การใช้คะแนนรวมทั้งฉบับเป็นเกณฑ์การจับคู่ไม่เหมาะสม แต่เมื่อใช้คะแนนจากแบบทดสอบย่อยหลายฉบับเป็นเกณฑ์การจับคู่จะมีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ดีกว่า โดยพิจารณาจากจำนวนข้อสอบทำหน้าที่ต่างกันที่ตรวจพบน้อยที่สุดเมื่อเปรียบเทียบกับการใช้เกณฑ์การจับคู่อื่น

จากการศึกษางานวิจัยที่ผ่านมา มีประเด็นปัญหาที่ผู้วิจัยสนใจ ดังนี้

1. เท่าที่ศึกษา ยังไม่มีการเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธีการวิเคราะห์องค์ประกอบจำกัด กับวิธีดออยโลจิสติก ว่าวิธีใดมีประสิทธิภาพดีกว่ากัน

2. ปัจจัยที่ส่งผลต่อประสิทธิภาพการตรวจพบข้อสอบทำหน้าที่ต่างกัน คือ ขนาดของกลุ่มตัวอย่าง และความยาวของแบบทดสอบ ดังนั้น น่าจะศึกษาว่าขนาดของกลุ่มตัวอย่าง และความยาวของแบบทดสอบ มีผลต่อการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติอย่างไร

3. ภายใต้เงื่อนไขการใช้เกณฑ์ภายในเป็นเกณฑ์การจับคู่ ได้แก่ คะแนนรวมทั้งฉบับ และคะแนนแบบทดสอบย่อย วิธีการวิเคราะห์หองค์ประกอบจำกัดกับวิธีถดถอยโลจิสติก มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบต่างกันหรือไม่

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

วัตถุประสงค์การวิจัย

1. เพื่อตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างวิธีการวิเคราะห์หองค์ประกอบจำกัดกับวิธีถดถอยโลจิสติก ภายใต้เงื่อนไข 18 เงื่อนไข ($3 \times 3 \times 2$) คือ ขนาดของกลุ่มตัวอย่าง 3 ขนาด (2,000 คน 1,000 คน และ 300 คน) ความยาวของแบบทดสอบ 3 ขนาด (40 ข้อ 30 ข้อ และ 20 ข้อ) และเกณฑ์การจับคู่ 2 เกณฑ์ (คะแนนรวมทั้งฉบับ และคะแนนแบบทดสอบย่อย)

2. เพื่อเปรียบเทียบประสิทธิภาพของวิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างวิธีการวิเคราะห์หองค์ประกอบจำกัดกับวิธีถดถอยโลจิสติก ภายใต้เงื่อนไข 18 เงื่อนไข ($3 \times 3 \times 2$) คือ ขนาดของกลุ่มตัวอย่าง 3 ขนาด (2,000 คน 1,000 คน และ 300 คน) ความยาวของแบบทดสอบ 3 ขนาด (40 ข้อ 30 ข้อ และ 20 ข้อ) และเกณฑ์การจับคู่ 2 เกณฑ์ (คะแนนรวมทั้งฉบับ และคะแนนแบบทดสอบย่อย) เมื่อใช้ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยวิธีชิปเทสต์ (Stout, 1999) เป็นเกณฑ์สำหรับการเปรียบเทียบในเรื่อง อัตราความถูกต้อง และอัตราความคลาดเคลื่อน

สมมติฐานการวิจัย

โรเจอร์ และสวามินาทาน (Roger & Swaminathan, 1993: 105-116) ได้ศึกษาพบว่าวิธีแมนเทิล-แฮนส์เซล มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป (Uniform DIF) เท่าเทียมกับวิธีถดถอยโลจิสติก และวิธีถดถอยโลจิสติก มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป (Nonuniform DIF) ดีกว่าวิธีแมนเทิล-แฮนส์เซล ออร์ต (Oort, 1998: 107-124) ได้เปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธีการตอบสนองข้อสอบแบบหนึ่งพารามิเตอร์ กับวิธีการวิเคราะห์หองค์ประกอบจำกัด ในข้อสอบแบบให้คะแนนเป็น 2 ค่า (ตอบถูกได้ 1 และ

ตอบผิดได้ 0) เมื่อกำหนดเงื่อนไขให้ข้อสอบทำหน้าที่ต่างกันมาก น้อย และไม่ต่างกัน กลุ่มตัวอย่างมีขนาด 2 ขนาด คือ ขนาดใหญ่ 2,000 คน และขนาดเล็ก 200 คน ใช้แบบทดสอบเดียวกัน(40 ข้อ) พบว่า โดยภาพรวมการตรวจสอบด้วยวิธีการวิเคราะห์องค์ประกอบจำกัด ให้ผลการตรวจสอบดีกว่าวิธีการตอบสนองข้อสอบแบบหนึ่งพารามิเตอร์ สำหรับกลุ่มตัวอย่างขนาดเล็ก (200 คน) การตรวจสอบด้วยวิธีการวิเคราะห์องค์ประกอบจำกัด ให้ผลการตรวจสอบดีกว่า แต่เมื่อกลุ่มตัวอย่างมีขนาดใหญ่ ทั้งสองวิธีจะให้ผลดีขึ้น นอกจากนี้ผลการศึกษาของ นิคม กิรติวรังกูร (2542) ยังสนับสนุนว่า โดยภาพรวมวิธีการวิเคราะห์องค์ประกอบจำกัด มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบดีกว่าวิธีแมนเทิล-แฮนส์เซล และวิธีการตอบสนองข้อสอบแบบสองพารามิเตอร์ และวิธีการตอบสนองข้อสอบแบบสองพารามิเตอร์ มีอัตราความคลาดเคลื่อนประเภทที่ 1 สูงกว่าวิธีแมนเทิล-แฮนส์เซล และวิธีการวิเคราะห์องค์ประกอบจำกัด ผู้วิจัยจึงตั้งสมมติฐานการวิจัยดังนี้

1. เมื่อกลุ่มตัวอย่างขนาดใหญ่ (2,000 คน) แบบทดสอบขนาด 40 ข้อ 30 ข้อ และ 20 ข้อ และเกณฑ์การจับคู่คะแนนรวมทั้งฉบับ วิธีการวิเคราะห์องค์ประกอบจำกัดมีประสิทธิภาพดีกว่าวิธีถดถอยโลจิสติก
2. เมื่อกลุ่มตัวอย่างขนาดใหญ่ (2,000 คน) แบบทดสอบขนาด 40 ข้อ 30 ข้อ และ 20 ข้อ และเกณฑ์การจับคู่คะแนนแบบทดสอบย่อย วิธีการวิเคราะห์องค์ประกอบจำกัดมีประสิทธิภาพดีกว่าวิธีถดถอยโลจิสติก
3. เมื่อกลุ่มตัวอย่างขนาดกลาง (1,000 คน) แบบทดสอบขนาด 40 ข้อ 30 ข้อ และ 20 ข้อ และเกณฑ์การจับคู่คะแนนรวมทั้งฉบับ วิธีการวิเคราะห์องค์ประกอบจำกัดมีประสิทธิภาพดีกว่าวิธีถดถอยโลจิสติก
4. เมื่อกลุ่มตัวอย่างขนาดกลาง (1,000 คน) แบบทดสอบขนาด 40 ข้อ 30 ข้อ และ 20 ข้อ และเกณฑ์การจับคู่คะแนนแบบทดสอบย่อย วิธีการวิเคราะห์องค์ประกอบจำกัดมีประสิทธิภาพดีกว่าวิธีถดถอยโลจิสติก
5. เมื่อกลุ่มตัวอย่างขนาดเล็ก (300 คน) แบบทดสอบขนาด 40 ข้อ 30 ข้อ และ 20 ข้อ และเกณฑ์การจับคู่คะแนนรวมทั้งฉบับ วิธีการวิเคราะห์องค์ประกอบจำกัดมีประสิทธิภาพดีกว่าวิธีถดถอยโลจิสติก
6. เมื่อกลุ่มตัวอย่างขนาดเล็ก (300 คน) แบบทดสอบขนาด 40 ข้อ 30 ข้อ และ 20 ข้อ และเกณฑ์การจับคู่คะแนนแบบทดสอบย่อย วิธีการวิเคราะห์องค์ประกอบจำกัดมีประสิทธิภาพดีกว่าวิธีถดถอยโลจิสติก

นิยามศัพท์เฉพาะ

อัตราความถูกต้อง (Correct rates) หมายถึง สัดส่วนของจำนวนข้อสอบที่ตรวจสอบพบว่าถูกต้องตรงกับวิธีชิปเทสต์ มี 2 ประเภท ได้แก่ อัตราความถูกต้องประเภทที่ 1 และอัตราความถูกต้องประเภทที่ 2

อัตราความถูกต้องประเภทที่ 1 (Type I correct rate) หมายถึง สัดส่วนของจำนวนข้อสอบที่ตรวจพบว่าทำหน้าที่ไม่ต่างกันได้ถูกต้องตรงกับวิธีชิปเทสต์ ซึ่งคำนวณจากสัดส่วนของจำนวนข้อสอบทำหน้าที่ไม่ต่างกันถูกต้องต่อจำนวนข้อสอบทำหน้าที่ไม่ต่างกันทั้งหมดในแบบทดสอบ

อัตราความถูกต้องประเภทที่ 2 (Type II correct rate) หมายถึง สัดส่วนของจำนวนข้อสอบที่ตรวจพบว่าทำหน้าที่ต่างกันได้ถูกต้องตรงกับวิธีชิปเทสต์ ซึ่งคำนวณจากสัดส่วนของจำนวนข้อสอบทำหน้าที่ต่างกันถูกต้องต่อจำนวนข้อสอบทำหน้าที่ต่างกันทั้งหมดในแบบทดสอบ

อัตราความคลาดเคลื่อน (Error rates) หมายถึง สัดส่วนของจำนวนข้อสอบที่ระบุผิดพลาด มี 2 ประเภท ได้แก่ ความคลาดเคลื่อนประเภทที่ 1 และความคลาดเคลื่อนประเภทที่ 2

ความคลาดเคลื่อนประเภทที่ 1 (Type I error) หมายถึง การระบุผิดพลาดว่าข้อสอบทำหน้าที่ต่างกัน (False positive) ทั้ง ๆ ที่ในความเป็นจริงแล้วข้อสอบทำหน้าที่ไม่ต่างกัน ซึ่งในการวิจัยนี้คำนวณได้จากสัดส่วนของจำนวนข้อสอบที่ระบุผิดพลาดว่าทำหน้าที่ต่างกันต่อจำนวนข้อสอบที่ทำหน้าที่ไม่ต่างกันทั้งหมดในแบบทดสอบ

ความคลาดเคลื่อนประเภทที่ 2 (Type II error) หมายถึง การระบุผิดพลาดว่าข้อสอบทำหน้าที่ไม่ต่างกัน (False negative) ทั้ง ๆ ที่ในความเป็นจริงแล้วข้อสอบทำหน้าที่ต่างกัน ซึ่งในการวิจัยนี้คำนวณได้จากสัดส่วนของจำนวนข้อสอบที่ระบุผิดพลาดว่าทำหน้าที่ไม่ต่างกันต่อจำนวนข้อสอบที่ทำหน้าที่ต่างกันทั้งหมดในแบบทดสอบ

วิธีดำเนินการวิจัย

กลุ่มตัวอย่าง

งานวิจัยนี้ใช้ข้อมูลทุติยภูมิ ซึ่งเป็นผลการตอบข้อสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ ปีการศึกษา 2546 วิชาภาษาไทย ชั้นประถมศึกษาปีที่ 6 โดยขอความร่วมมือจากเจ้าหน้าที่สำนักทดสอบทางการศึกษา กระทรวงศึกษาธิการ ข้อมูลกลุ่มตัวอย่างได้มาโดยวิธีการสุ่มแบบแบ่งชั้น (Stratified random sampling) ใช้ระดับความสามารถของผู้เข้าสอบเป็นชั้น แบ่งเป็นกลุ่มดี กลุ่มพอใช้และกลุ่มปรับปรุง โดยมีนักเรียนเป็นหน่วยการสุ่ม สุ่มแยกตามตัวแปรเพศ ได้กลุ่มตัวอย่าง จำนวน 2,000 คน เพราะเป็นขนาดกลุ่มตัวอย่างที่สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้อย่างถูกต้อง แม่นยำ และมีความพอเพียงสำหรับประมาณค่าพารามิเตอร์ได้อย่างเหมาะสม (จิตติมา วรรณศรี, 2539) จำแนกเป็นผลการตอบของผู้เข้าสอบเพศชาย จำนวน 1,000 คน และผลการตอบของผู้เข้าสอบเพศหญิง จำนวน 1,000 คน

เครื่องมือที่ใช้ในการวิจัย

แบบทดสอบวัดผลสัมฤทธิ์ทางการเรียนระดับชาติ ชั้นประถมศึกษาปีที่ 6 ปีการศึกษา 2546 วิชาภาษาไทย ประกอบด้วยข้อสอบแบบหลายตัวเลือกชนิด 4 ตัวเลือก จำนวน 40 ข้อ คະแนนเต็ม 40 คะแนน สร้างโดยสำนักทดสอบทางการศึกษา กระทรวงศึกษาธิการ โครงสร้างของแบบทดสอบมี 2 องค์ประกอบ ดังนี้

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

แบบทดสอบทั้งฉบับมีค่าความยากประมาณ .50 ค่าอำนาจจำแนกทั้งฉบับมากกว่า .40 และค่าความเที่ยงของทั้งฉบับ เท่ากับ .86

การเก็บรวบรวมข้อมูล

การเก็บรวบรวมข้อมูลผลการตอบของนักเรียนที่เป็นกลุ่มตัวอย่าง ซึ่งเป็นข้อมูลทุติยภูมิ จากสำนักทดสอบทางการศึกษา กระทรวงศึกษาธิการ โดยเจ้าหน้าที่สุ่มผลการตอบของนักเรียนจำแนกตามระดับความสามารถและตามตัวแปรเพศ เพื่อให้ได้กลุ่มตัวอย่าง จำนวน 2,000 คน

การจัดเตรียมข้อมูลสำหรับวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ

1. ขนาดกลุ่มตัวอย่าง

กำหนดกลุ่มตัวอย่างเป็น 3 ขนาด คือ ขนาดใหญ่ (2,000 คน) ขนาดกลาง (1,000 คน) และขนาดเล็ก (300 คน) โดยผู้วิจัยสุ่มจากกลุ่มตัวอย่างขนาดใหญ่ (2,000 คน) ด้วยวิธีสุ่มอย่างง่ายโดยใช้คอมพิวเตอร์ จำแนกเป็นนักเรียนชาย และหญิงกลุ่มละเท่า ๆ กัน

2. ความยาวของแบบทดสอบ

กำหนดความยาวของแบบทดสอบเป็น 3 ขนาด คือ แบบทดสอบ 40 ข้อ 30 ข้อ และ 20 ข้อ โดยผู้วิจัยสุ่มจากแบบทดสอบ 40 ข้อ ด้วยวิธีสุ่มอย่างง่ายโดยใช้คอมพิวเตอร์ ตามโครงสร้างของแบบทดสอบ ดังนี้

2.1 แบบทดสอบขนาด 30 ข้อ

2.1.1 วัดโครงสร้างความรู้ จำนวน 11 ข้อ

2.1.2 วัดกระบวนการ จำนวน 19 ข้อ

2.2 แบบทดสอบขนาด 20 ข้อ

2.2.1 วัดโครงสร้างความรู้ จำนวน 7 ข้อ

2.2.2 วัดกระบวนการ จำนวน 13 ข้อ

การวิเคราะห์ข้อมูล

วิเคราะห์ค่าสถิติพื้นฐานโดยใช้โปรแกรม SPSS ตรวจสอบความตรงเชิงโครงสร้างของแบบทดสอบทั้งฉบับด้วยการวิเคราะห์องค์ประกอบเชิงยืนยันอันดับสอง โดยใช้โปรแกรม LISREL 8.50 และตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยใช้โปรแกรม SIBTEST LISREL 8.50 และ SPSS

ผลการวิจัย

ผู้วิจัยนำเสนอผลการวิจัยเป็น 2 ส่วน ได้แก่ ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ และผลการเปรียบเทียบประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติระหว่างวิธีการวิเคราะห์องค์ประกอบจำกัด (RFA) กับวิธีถดถอยโลจิสติก (LR) เมื่อใช้ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยใช้วิธีซิปเทสต์ (SIBTEST) เป็นเกณฑ์ ดังนี้

1. ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ดังตารางที่ 1

ตารางที่ 1 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ

ขนาดกลุ่ม ตัวอย่าง	ความยาวของ แบบทดสอบ	เกณฑ์การจับคู่	วิธี SIBTEST		วิธี RFA		วิธี LR	
			DIF	NO DIF	DIF	NO DIF	DIF	NO DIF
			(ข้อ)	(ข้อ)	(ข้อ)	(ข้อ)	(ข้อ)	(ข้อ)
2,000 คน	40 ข้อ	คะแนนรวมทั้งฉบับ	17	23	24	16	18	22
		คะแนนแบบทดสอบย่อย	17	23	26	14	18	22
	30 ข้อ	คะแนนรวมทั้งฉบับ	15	15	17	13	15	15
		คะแนนแบบทดสอบย่อย	15	15	22	8	15	15
	20 ข้อ	คะแนนรวมทั้งฉบับ	9	11	10	10	10	10
		คะแนนแบบทดสอบย่อย	8	12	6	14	6	14
1,000 คน	40 ข้อ	คะแนนรวมทั้งฉบับ	13	27	20	20	10	30
		คะแนนแบบทดสอบย่อย	11	29	23	17	10	30
	30 ข้อ	คะแนนรวมทั้งฉบับ	8	22	17	13	9	21
		คะแนนแบบทดสอบย่อย	10	20	21	9	9	21
	20 ข้อ	คะแนนรวมทั้งฉบับ	7	13	9	11	6	14
		คะแนนแบบทดสอบย่อย	7	13	6	14	8	12
300 คน	40 ข้อ	คะแนนรวมทั้งฉบับ	3	37	19	21	5	35
		คะแนนแบบทดสอบย่อย	3	37	18	22	5	35
	30 ข้อ	คะแนนรวมทั้งฉบับ	3	27	14	16	5	25
		คะแนนแบบทดสอบย่อย	4	26	14	16	5	25
	20 ข้อ	คะแนนรวมทั้งฉบับ	2	18	8	12	2	18
		คะแนนแบบทดสอบย่อย	2	18	5	15	4	16

หมายเหตุ DIF หมายถึง ข้อสอบทำหน้าที่ต่างกัน
NO DIF หมายถึง ข้อสอบทำหน้าที่ไม่ต่างกัน

จากตารางที่ 1 ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ปรากฏว่า เมื่อกลุ่มตัวอย่าง 2,000 คน แบบทดสอบ 40 ข้อ และเกณฑ์การจับคู่ใช้คะแนนรวมทั้งฉบับ วิธี RFA ตรวจพบข้อสอบทำหน้าที่ต่างกัน (DIF) จำนวน 24 ข้อ และวิธี LR ตรวจพบจำนวน 18 ข้อ และเมื่อเกณฑ์การจับคู่ใช้คะแนนแบบทดสอบย่อย วิธี RFA ตรวจพบข้อสอบทำหน้าที่ต่างกันจำนวน 26 ข้อ และวิธี LR ตรวจพบจำนวน 18 ข้อ

เมื่อกลุ่มตัวอย่าง 2,000 คน แบบทดสอบ 30 ข้อ และเกณฑ์การจับคู่ใช้คะแนนรวมทั้งฉบับ วิธี RFA ตรวจพบข้อสอบทำหน้าที่ต่างกัน จำนวน 17 ข้อ และวิธี LR ตรวจพบจำนวน 15 ข้อ และเมื่อเกณฑ์การจับคู่ใช้คะแนนแบบทดสอบย่อย วิธี RFA ตรวจพบข้อสอบทำหน้าที่ต่างกันจำนวน 22 ข้อ และวิธี LR ตรวจพบจำนวน 15 ข้อ

เมื่อกลุ่มตัวอย่าง 2,000 คน แบบทดสอบ 20 ข้อ และเกณฑ์การจับคู่ใช้คะแนนรวมทั้งฉบับ วิธี RFA ตรวจพบข้อสอบทำหน้าที่ต่างกัน จำนวน 10 ข้อ และวิธี LR ตรวจพบจำนวน 10 ข้อ และเมื่อเกณฑ์การจับคู่ใช้คะแนนแบบทดสอบย่อย วิธี RFA ตรวจพบข้อสอบทำหน้าที่ต่างกันจำนวน 6 ข้อ และวิธี LR ตรวจพบจำนวน 6 ข้อ

เมื่อกลุ่มตัวอย่าง 1,000 คน แบบทดสอบ 40 ข้อ และเกณฑ์การจับคู่ใช้คะแนนรวมทั้งฉบับ วิธี RFA ตรวจพบข้อสอบทำหน้าที่ต่างกัน จำนวน 20 ข้อ และวิธี LR ตรวจพบจำนวน 10 ข้อ และเมื่อเกณฑ์การจับคู่ใช้คะแนนแบบทดสอบย่อย วิธี RFA ตรวจพบข้อสอบทำหน้าที่ต่างกันจำนวน 23 ข้อ และวิธี LR ตรวจพบจำนวน 10 ข้อ

เมื่อกลุ่มตัวอย่าง 1,000 คน แบบทดสอบ 30 ข้อ และเกณฑ์การจับคู่ใช้คะแนนรวมทั้งฉบับ วิธี RFA ตรวจพบข้อสอบทำหน้าที่ต่างกัน จำนวน 17 ข้อ และวิธี LR ตรวจพบจำนวน 9 ข้อ และเมื่อเกณฑ์การจับคู่ใช้คะแนนแบบทดสอบย่อย วิธี RFA ตรวจพบข้อสอบทำหน้าที่ต่างกันจำนวน 21 ข้อ และวิธี LR ตรวจพบจำนวน 9 ข้อ

เมื่อกลุ่มตัวอย่าง 1,000 คน แบบทดสอบ 20 ข้อ และเกณฑ์การจับคู่ใช้คะแนนรวมทั้งฉบับ วิธี RFA ตรวจพบข้อสอบทำหน้าที่ต่างกัน จำนวน 9 ข้อ และวิธี LR ตรวจพบจำนวน 6 ข้อ และเมื่อเกณฑ์การจับคู่ใช้คะแนนแบบทดสอบย่อย วิธี RFA ตรวจพบข้อสอบทำหน้าที่ต่างกันจำนวน 6 ข้อ และวิธี LR ตรวจพบจำนวน 8 ข้อ

เมื่อกลุ่มตัวอย่าง 300 คน แบบทดสอบ 40 ข้อ และเกณฑ์การจับคู่ใช้คะแนนรวมทั้งฉบับ วิธี RFA ตรวจพบข้อสอบทำหน้าที่ต่างกัน จำนวน 19 ข้อ และวิธี LR ตรวจพบจำนวน 5 ข้อ และเมื่อเกณฑ์การจับคู่ใช้คะแนนแบบทดสอบย่อย วิธี RFA ตรวจพบข้อสอบทำหน้าที่ต่างกันจำนวน 18 ข้อ และวิธี LR ตรวจพบจำนวน 5 ข้อ

เมื่อกลุ่มตัวอย่าง 300 คน แบบทดสอบ 30 ข้อ และเกณฑ์การจับคู่ใช้คะแนนรวมทั้งฉบับ วิธี RFA ตรวจพบข้อสอบทำหน้าที่ต่างกัน จำนวน 14 ข้อ และวิธี LR ตรวจพบจำนวน 5 ข้อ และเมื่อเกณฑ์การจับคู่ใช้คะแนนแบบทดสอบย่อย วิธี RFA ตรวจพบข้อสอบทำหน้าที่ต่างกันจำนวน 14 ข้อ และวิธี LR ตรวจพบจำนวน 5 ข้อ

เมื่อกลุ่มตัวอย่าง 300 คน แบบทดสอบ 20 ข้อ และเกณฑ์การจับคู่ใช้คะแนนรวมทั้งฉบับ วิธี RFA ตรวจพบข้อสอบทำหน้าที่ต่างกัน จำนวน 8 ข้อ และวิธี LR ตรวจพบจำนวน 2 ข้อ และเมื่อเกณฑ์การจับคู่ใช้คะแนนแบบทดสอบย่อย วิธี RFA ตรวจพบข้อสอบทำหน้าที่ต่างกันจำนวน 5 ข้อ และวิธี LR ตรวจพบจำนวน 4 ข้อ

2. ผลการเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติระหว่างวิธีการวิเคราะห์องค์ประกอบจำกัด (RFA) กับวิธีถดถอยโลจิสติก (LR) เมื่อใช้ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยวิธีชิปเทสต์ (Stout, 1999) เป็นเกณฑ์สำหรับการเปรียบเทียบในเรื่อง อัตราความถูกต้อง และอัตราความคลาดเคลื่อน

2.1 การทดสอบนัยสำคัญของอัตราความถูกต้องและอัตราความคลาดเคลื่อนของผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธี RFA กับวิธี LR ดังตารางที่ 2 และ 3

ตารางที่ 2 ผลการทดสอบนัยสำคัญของอัตราความถูกต้องประเภทที่ 1 และอัตราความคลาดเคลื่อนประเภทที่ 1 ของผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธี RFA กับวิธี LR

ขนาดกลุ่มตัวอย่าง	ความยาวของแบบทดสอบ	เกณฑ์การจับคู่	อัตราความถูกต้องประเภทที่ 1				อัตราความคลาดเคลื่อนประเภทที่ 1			
			วิธี RFA	วิธี LR	Z test	สรุป	วิธี RFA	วิธี LR	Z test	สรุป
2,000 คน	40 ข้อ	คะแนนรวมทั้งฉบับ	0.609	0.870	-2.663*	RFA < LR	0.391	0.130	2.663*	RFA > LR
		คะแนนแบบทดสอบย่อย	0.522	0.913	-3.910*	RFA < LR	0.478	0.087	3.910*	RFA > LR
	30 ข้อ	คะแนนรวมทั้งฉบับ	0.600	0.933	-3.040*	RFA < LR	0.400	0.067	3.040*	RFA > LR
		คะแนนแบบทดสอบย่อย	0.467	0.933	-3.938*	RFA < LR	0.533	0.067	3.938*	RFA > LR
	20 ข้อ	คะแนนรวมทั้งฉบับ	0.545	0.909	-2.574*	RFA < LR	0.455	0.091	2.574*	RFA > LR
		คะแนนแบบทดสอบย่อย	0.833	1.000	-1.867*	RFA < LR	0.167	0.000	1.867*	RFA > LR
1,000 คน	40 ข้อ	คะแนนรวมทั้งฉบับ	0.593	0.926	-3.510*	RFA < LR	0.407	0.074	3.510*	RFA > LR
		คะแนนแบบทดสอบย่อย	0.483	0.897	-3.947*	RFA < LR	0.517	0.103	3.947*	RFA > LR
	30 ข้อ	คะแนนรวมทั้งฉบับ	0.364	0.955	-4.825*	RFA < LR	0.636	0.045	4.825*	RFA > LR
		คะแนนแบบทดสอบย่อย	0.400	0.900	-4.083*	RFA < LR	0.600	0.100	4.083*	RFA > LR
	20 ข้อ	คะแนนรวมทั้งฉบับ	0.538	0.923	-2.722*	RFA < LR	0.462	0.077	2.722*	RFA > LR
		คะแนนแบบทดสอบย่อย	0.769	0.923	-1.351	RFA = LR	0.231	0.077	1.351	RFA = LR
300 คน	40 ข้อ	คะแนนรวมทั้งฉบับ	0.541	0.892	-3.510*	RFA < LR	0.459	0.108	3.510*	RFA > LR
		คะแนนแบบทดสอบย่อย	0.541	0.942	-4.010*	RFA < LR	0.459	0.054	4.010*	RFA > LR
	30 ข้อ	คะแนนรวมทั้งฉบับ	0.519	0.889	-3.127*	RFA < LR	0.481	0.111	3.127*	RFA > LR
		คะแนนแบบทดสอบย่อย	0.538	0.923	-3.377*	RFA < LR	0.462	0.077	3.377*	RFA > LR
	20 ข้อ	คะแนนรวมทั้งฉบับ	0.556	0.944	-2.815*	RFA < LR	0.444	0.056	2.815*	RFA > LR
		คะแนนแบบทดสอบย่อย	0.722	0.889	-1.320	RFA = LR	0.278	0.111	1.320	RFA = LR

* $p < .05$

หมายเหตุ อัตราความถูกต้องประเภทที่ 1

RFA < LR หมายความว่า วิธี LR มีประสิทธิภาพในการตรวจสอบดีกว่าวิธี RFA

RFA = LR หมายความว่า วิธีทั้งสองมีประสิทธิภาพในการตรวจสอบเท่ากัน

อัตราความคลาดเคลื่อนประเภทที่ 1

RFA > LR หมายความว่า วิธี LR มีประสิทธิภาพในการตรวจสอบดีกว่าวิธี RFA

RFA = LR หมายความว่า วิธีทั้งสองมีประสิทธิภาพในการตรวจสอบเท่ากัน

จากตารางที่ 2 เมื่อทดสอบนัยสำคัญของอัตราความถูกต้องประเภทที่ 1 และอัตราความคลาดเคลื่อนประเภทที่ 1 ของผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธีการวิเคราะห์องค์ประกอบจำกัด (RFA) กับวิธีถดถอยโลจิสติก (LR) ปรากฏว่า

เมื่อใช้กลุ่มตัวอย่าง 1,000 คน แบบทดสอบ 20 ข้อ และเกณฑ์การจับคู่ใช้คะแนนแบบทดสอบย่อย ปรากฏว่า วิธีทั้งสองมีประสิทธิภาพในการตรวจสอบไม่แตกต่างกัน

เมื่อใช้กลุ่มตัวอย่าง 300 คน แบบทดสอบ 20 ข้อ และเกณฑ์การจับคู่ใช้คะแนนแบบทดสอบย่อย ปรากฏว่า วิธีทั้งสองมีประสิทธิภาพในการตรวจสอบไม่แตกต่างกัน

โดยภาพรวมวิธี LR มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ดีกว่าวิธี RFA เนื่องจากวิธี LR มีอัตราความถูกต้องประเภทที่ 1 มากกว่า และมีอัตราความคลาดเคลื่อนประเภทที่ 1 น้อยกว่าวิธี RFA อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

ตารางที่ 3 ผลการทดสอบนัยสำคัญของอัตราความถูกต้องประเภทที่ 2 และอัตราความคลาดเคลื่อนประเภทที่ 2 ของผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธี RFA กับวิธี LR

ขนาดกลุ่ม ตัวอย่าง	ความยาวของ แบบทดสอบ	เกณฑ์การจับคู่	อัตราความถูกต้องประเภทที่ 2				อัตราความคลาดเคลื่อนประเภทที่ 2			
			วิธี RFA	วิธี LR	Z test	สรุป	วิธี RFA	วิธี LR	Z test	สรุป
2,000 คน	40 ข้อ	คะแนนรวมทั้งฉบับ	0.882	0.882	0.000	<u>RFA = LR</u>	0.118	0.118	0.000	<u>RFA = LR</u>
		คะแนนแบบทดสอบย่อย	0.882	0.941	-0.933	<u>RFA = LR</u>	0.118	0.059	0.933	<u>RFA = LR</u>
	30 ข้อ	คะแนนรวมทั้งฉบับ	0.733	0.933	-2.108*	RFA < LR	0.267	0.067	2.108*	RFA > LR
		คะแนนแบบทดสอบย่อย	0.933	0.933	0.000	<u>RFA = LR</u>	0.067	0.067	0.000	<u>RFA = LR</u>
	20 ข้อ	คะแนนรวมทั้งฉบับ	0.556	1.000	-3.405*	RFA < LR	0.444	0.000	3.405*	RFA > LR
		คะแนนแบบทดสอบย่อย	0.500	0.750	-1.648*	RFA < LR	0.500	0.250	1.648*	RFA > LR
	40 ข้อ	คะแนนรวมทั้งฉบับ	0.692	0.615	0.734	<u>RFA = LR</u>	0.308	0.385	-0.734	<u>RFA = LR</u>
		คะแนนแบบทดสอบย่อย	0.727	0.636	0.868	<u>RFA = LR</u>	0.273	0.364	-0.868	<u>RFA = LR</u>
1,000 คน	30 ข้อ	คะแนนรวมทั้งฉบับ	0.375	1.000	-5.282*	RFA < LR	0.625	0.000	5.282*	RFA > LR
		คะแนนแบบทดสอบย่อย	0.900	0.700	1.907*	RFA > LR	0.100	0.300	-1.907*	RFA < LR
	20 ข้อ	คะแนนรวมทั้งฉบับ	0.429	0.714	-1.821*	RFA < LR	0.571	0.286	1.821*	RFA > LR
		คะแนนแบบทดสอบย่อย	0.429	1.000	-4.038*	RFA < LR	0.571	0.000	4.038*	RFA > LR
	40 ข้อ	คะแนนรวมทั้งฉบับ	0.667	0.333	2.929*	RFA > LR	0.333	0.667	-2.929*	RFA < LR
		คะแนนแบบทดสอบย่อย	0.333	1.000	-6.360*	RFA < LR	0.667	0.000	6.360*	RFA > LR
	30 ข้อ	คะแนนรวมทั้งฉบับ	0.333	0.667	-2.562*	RFA < LR	0.667	0.333	2.562*	RFA > LR
		คะแนนแบบทดสอบย่อย	0.500	0.750	-1.976*	RFA < LR	0.500	0.250	1.976*	RFA > LR
300 คน	20 ข้อ	คะแนนรวมทั้งฉบับ	0.000	0.500	-3.627*	RFA < LR	1.000	0.500	3.627*	RFA > LR
		คะแนนแบบทดสอบย่อย	0.000	1.000	-6.325*	RFA < LR	1.000	0.000	6.325*	RFA > LR

* $p < .05$

หมายเหตุ อัตราความถูกต้องประเภทที่ 2

RFA < LR หมายความว่า วิธี LR มีประสิทธิภาพในการตรวจสอบดีกว่าวิธี RFA

RFA = LR หมายความว่า วิธีทั้งสองมีประสิทธิภาพในการตรวจสอบเท่ากัน

RFA > LR หมายความว่า วิธี RFA มีประสิทธิภาพในการตรวจสอบดีกว่าวิธี LR

อัตราความคลาดเคลื่อนประเภทที่ 2

RFA > LR หมายความว่า วิธี LR มีประสิทธิภาพในการตรวจสอบดีกว่าวิธี RFA

RFA = LR หมายความว่า วิธีทั้งสองมีประสิทธิภาพในการตรวจสอบเท่ากัน

RFA < LR หมายความว่า วิธี RFA มีประสิทธิภาพในการตรวจสอบดีกว่าวิธี LR

จากตารางที่ 3 เมื่อทดสอบนัยสำคัญของการเปรียบเทียบอัตราความถูกต้องประเภทที่ 2 และอัตราความคลาดเคลื่อนประเภทที่ 2 ของผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบระหว่างวิธีการวิเคราะห์องค์ประกอบจำกัด (RFA) กับวิธีถดถอยโลจิสติก (LR) ปรากฏว่า

เมื่อใช้กลุ่มตัวอย่าง 2,000 คน แบบทดสอบ 40 ข้อ และเกณฑ์การจับคู่ใช้คะแนนรวมทั้งฉบับและคะแนนแบบทดสอบย่อย ปรากฏว่า วิธีทั้งสองมีประสิทธิภาพในการตรวจสอบไม่แตกต่างกัน

เมื่อใช้กลุ่มตัวอย่าง 2,000 คน แบบทดสอบ 30 ข้อ และเกณฑ์การจับคู่ใช้คะแนนแบบทดสอบย่อย ปรากฏว่า วิธีทั้งสองมีประสิทธิภาพในการตรวจสอบไม่แตกต่างกัน

เมื่อใช้กลุ่มตัวอย่าง 1,000 คน แบบทดสอบ 40 ข้อ และเกณฑ์การจับคู่ใช้คะแนนรวมทั้งฉบับและคะแนนแบบทดสอบย่อย ปรากฏว่า วิธีทั้งสองมีประสิทธิภาพในการตรวจสอบไม่แตกต่างกัน

เมื่อใช้กลุ่มตัวอย่าง 1,000 คน แบบทดสอบ 30 ข้อ และเกณฑ์การจับคู่ใช้คะแนนแบบทดสอบย่อย ปรากฏว่า วิธี RFA มีประสิทธิภาพในการตรวจสอบดีกว่าวิธี LR เนื่องจากวิธี RFA มีอัตราความถูกต้องประเภทที่ 2 มากกว่า และมีอัตราความคลาดเคลื่อนประเภทที่ 2 น้อยกว่าวิธี LR อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

เมื่อใช้กลุ่มตัวอย่าง 300 คน แบบทดสอบ 40 ข้อ และเกณฑ์การจับคู่ใช้คะแนนรวมทั้งฉบับ ปรากฏว่า วิธี RFA มีประสิทธิภาพในการตรวจสอบดีกว่าวิธี LR เนื่องจากวิธี RFA มีอัตราความถูกต้องประเภทที่ 2 มากกว่า และมีอัตราความคลาดเคลื่อนประเภทที่ 2 น้อยกว่าวิธี LR อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

โดยภาพรวมวิธี LR มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติดีกว่าวิธี RFA เนื่องจากวิธี LR มีอัตราความถูกต้องประเภทที่ 2 มากกว่า และมีอัตราความคลาดเคลื่อนประเภทที่ 2 น้อยกว่าวิธี RFA อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

การอภิปรายผลการวิจัย

ผลการวิจัยนี้ มีประเด็นสำคัญที่นำมาอภิปราย ดังนี้

ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างวิธีการวิเคราะห์องค์ประกอบจำกัด (RFA) กับวิธีถดถอยโลจิสติก (LR) ภายใต้เงื่อนไขที่ศึกษา 18 เงื่อนไข ปรากฏว่า ทั้งสองวิธีมีประสิทธิภาพในการตรวจสอบไม่แตกต่างกัน 2 เงื่อนไข คือ 1) เมื่อกลุ่มตัวอย่างขนาดกลาง (1,000 คน) แบบทดสอบมีความยาว 20 ข้อ และเกณฑ์การจับคู่ใช้คะแนนแบบทดสอบย่อย และ 2) เมื่อกลุ่มตัวอย่างขนาดเล็ก (300 คน) แบบทดสอบมีความยาว 20 ข้อ และเกณฑ์การจับคู่ใช้คะแนนแบบทดสอบย่อย ซึ่งไม่สอดคล้องกับสมมติฐานการวิจัย ทั้งนี้อาจเป็นเพราะว่าในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ เมื่อขนาดของกลุ่มตัวอย่าง และความยาวของแบบทดสอบต่างกันไม่มากนัก ผลการตรวจสอบจะไม่ต่างกัน สอดคล้องกับผลการศึกษาของออร์ท (Oort, 1998) ที่ว่า อำนาจการตรวจสอบและอัตราความคลาดเคลื่อนของวิธี RFA ไม่ขึ้นอยู่กับขนาดของกลุ่มตัวอย่างและความยาวของแบบทดสอบ และสอดคล้องกับผลการศึกษาของเคลาเซอร์และคณะ (Clauser et al., 1996) ที่ว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบที่มีมิติการวัดซับซ้อน ด้วยวิธีแมนเทิล-แฮนส์เซล และวิธีถดถอยโลจิสติก จะให้ผลการ

ตรวจสอบคล้ายกัน นั่นคือ เมื่อเปรียบเทียบทั้งสองวิธีโดยใช้คะแนนแบบทดสอบย่อยเป็นเกณฑ์การจับคู่เปรียบเทียบสามารถตรวจสอบได้ดีเช่นเดียวกับการใช้คะแนนรวมทั้งฉบับเป็นเกณฑ์การจับคู่ ซึ่งชี้ให้เห็นว่าประสิทธิภาพในการตรวจสอบของทั้งสองวิธีไม่แตกต่างกัน

ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างวิธีการวิเคราะห์องค์ประกอบจำกั (RFA) กับวิธีถดถอยโลจิสติก (LR) ภายใต้เงื่อนไขที่ศึกษา 18 เงื่อนไข ปรากฏว่า วิธี LR มีประสิทธิภาพในการตรวจสอบดีกว่าวิธี RFA 16 เงื่อนไข คือ เมื่อกลุ่มตัวอย่างขนาดใหญ่ (2,000 คน) มี 6 เงื่อนไข (3×2) คือ แบบทดสอบทั้ง 3 ขนาด และเกณฑ์การจับคู่ทั้ง 2 เกณฑ์ สำหรับกลุ่มตัวอย่างขนาดกลาง (1,000 คน) มี 5 เงื่อนไข คือ แบบทดสอบทั้ง 3 ขนาด และเกณฑ์การจับคู่ทั้ง 2 เกณฑ์ ยกเว้นแบบทดสอบขนาด 20 ข้อ และเกณฑ์การจับคู่คะแนนแบบทดสอบย่อย ส่วนกลุ่มตัวอย่างขนาดเล็ก (300 คน) มี 5 เงื่อนไข คือ แบบทดสอบทั้ง 3 ขนาด และเกณฑ์การจับคู่ทั้ง 2 เกณฑ์ ยกเว้นแบบทดสอบขนาด 20 ข้อ และเกณฑ์การจับคู่คะแนนแบบทดสอบย่อย ซึ่งข้อค้นพบดังกล่าวไม่สอดคล้องกับสมมติฐานการวิจัยที่ตั้งไว้ เพราะว่าการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ วิธี LR มีประสิทธิภาพในการตรวจสอบดีกว่าเกือบทุกเงื่อนไข เมื่อใช้กลุ่มตัวอย่างขนาดใหญ่ ขนาดกลาง และขนาดเล็ก โดยพิจารณาจากอัตราความถูกต้องประเภทที่ 1 และอัตราความคลาดเคลื่อนประเภทที่ 1 ทั้งนี้อาจจะเป็นเพราะว่าจากการศึกษางานวิจัยที่ผ่านมา ออร์ต (Oort, 1998) ศึกษาโดยการจำลองข้อมูลในแบบทดสอบเอกมิติ แต่ในงานวิจัยนี้ผู้วิจัยศึกษาจากข้อมูลจริงในแบบทดสอบพหุมิติ จึงทำให้ผลการตรวจสอบไม่สอดคล้องกับสมมติฐานการวิจัยซึ่งตั้งไว้ตามผลการวิจัยที่ผ่านมา แต่ผลการวิจัยนี้สอดคล้องกับผลการศึกษาของ เมเซอร์ และคณะ (Mazor et al., 1992) สวามินาทัน และโรเจอร์ (Swaminathan & Roger, 1990) ที่พบว่า เมื่อกลุ่มตัวอย่างมีขนาดใหญ่ขึ้นจะทำให้อัตราการตรวจพบข้อสอบทำหน้าที่ต่างกันมากขึ้น สอดคล้องกับผลการศึกษาของ อัททาโร และมิลแซพ (Uttaro & Millsap, 1994) ที่พบว่าแบบทดสอบที่มีความยาวและกลุ่มตัวอย่างขนาดใหญ่ ทำให้มีโอกาสตรวจพบข้อสอบทำหน้าที่ต่างกันได้มากขึ้น และผลการศึกษาของ คลาเซอร์ และคณะ (Clauser et al., 1996) ที่พบว่า แบบทดสอบที่มีการวัดในหลายมิติ การใช้คะแนนรวมทั้งฉบับเป็นเกณฑ์การจับคู่ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทำให้เกิดความผิดพลาดของผลการตรวจสอบ เนื่องจากคะแนนผลการสอบที่ได้ไม่สัมพันธ์กับคุณลักษณะแฝงที่ต้องการทั้งหมด แต่เมื่อใช้คะแนนแบบทดสอบย่อยจะให้ผลการตรวจสอบดีกว่า นอกจากนี้ เมเซอร์ และคลาเซอร์ (Mazor & Clauser, 1995) ได้เสนอแนะให้นำตัวแทนความสามารถที่หลากหลาย (Multiple ability) เข้ามาวิเคราะห์ด้วย เพราะจะทำให้การตรวจสอบมีความถูกต้องมากขึ้น ซึ่งวิธีที่เหมาะสมในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในกรณีนี้ คือ วิธีถดถอยโลจิสติก เนื่องจากมีความยืดหยุ่นในการนำตัวแทนความสามารถที่หลากหลายเข้าสู่การวิเคราะห์สูง สอดคล้องกับผลการศึกษาของ เมเซอร์ และคลาเซอร์ (Mazor & Clauser, 1995) ที่พบว่า วิธีถดถอยโลจิสติกสามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ดีในแบบทดสอบที่มีตัวแปรความสามารถสองตัวหรือมากกว่า ทั้งนี้เป็นเพราะความยืดหยุ่นของวิธีถดถอยโลจิสติก

ที่มีพื้นฐานอยู่บนสมการถดถอยโลจิสติก ซึ่งสามารถเพิ่มตัวแปรความสามารถหลาย ๆ ตัวเข้าไปในกระบวนการวิเคราะห์ จึงทำให้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้ดียิ่งขึ้น

ข้อเสนอแนะ

การนำผลการวิจัยไปใช้

1. จากผลการวิจัยปรากฏว่า การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ ระหว่างวิธีการวิเคราะห์องค์ประกอบจำกัดกับวิธีถดถอยโลจิสติก ภายใต้เงื่อนไข 18 เงื่อนไข (3×3×2) คือ ขนาดของกลุ่มตัวอย่าง 3 ขนาด (2,000 คน 1,000 คน และ 300 คน) ความยาวของแบบทดสอบ 3 ขนาด (40 ข้อ 30 ข้อ และ 20 ข้อ) และใช้เกณฑ์การจับคู่ 2 เกณฑ์ (คะแนนรวมทั้งฉบับและคะแนนแบบทดสอบย่อย) เมื่อใช้ผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบโดยวิธีชิปเทสต์เป็นเกณฑ์สำหรับการเปรียบเทียบประสิทธิภาพของวิธีตรวจสอบทั้งสองวิธี โดยภาพรวมวิธีถดถอยโลจิสติกมีประสิทธิภาพในการตรวจสอบมากกว่าวิธีการวิเคราะห์องค์ประกอบจำกัด เนื่องจากวิธีถดถอยโลจิสติกมีอัตราความถูกต้องประเภทที่ 1 และ 2 มากกว่า และมีอัตราความคลาดเคลื่อนประเภทที่ 1 และ 2 น้อยกว่าวิธีการวิเคราะห์องค์ประกอบจำกัด ดังนั้นวิธีที่ควรนำไปใช้ตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ คือ วิธีถดถอยโลจิสติก

2. เกณฑ์การจับคู่ สำหรับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติที่เหมาะสม คือ เกณฑ์การจับคู่ใช้คะแนนแบบทดสอบย่อย เพราะในแบบทดสอบพหุมิติการใช้คะแนนรวมทั้งฉบับเป็นเกณฑ์การจับคู่ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทำให้เกิดความผิดพลาดของผลการตรวจสอบ เนื่องจากคะแนนผลการสอบที่ได้ไม่สัมพันธ์กับคุณลักษณะแฝงที่ต้องการทั้งหมด แต่เมื่อใช้คะแนนแบบทดสอบย่อยจะให้ผลการตรวจสอบดีกว่า

การวิจัยต่อไป

1. การตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติกับข้อมูลจริง มีปัญหาในเรื่องการหาวิธีที่เหมาะสมที่สุด เพื่อใช้เป็นวิธีมาตรฐานสำหรับระบุว่าข้อสอบข้อใดทำหน้าที่ต่างกัน ซึ่งในงานวิจัยนี้ผู้วิจัยใช้วิธีชิปเทสต์ จึงน่าจะศึกษาเพิ่มเติมโดยการนำวิธีอื่น หรืออาจใช้หลายวิธีร่วมกัน เพื่อจะได้นำผลการตรวจสอบการทำหน้าที่ต่างกันมาเปรียบเทียบกัน

2. ควรมีการศึกษาเพิ่มเติมเกี่ยวกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติ เมื่อจำแนกประเภทของข้อสอบทำหน้าที่ต่างกัน ได้แก่ ข้อสอบทำหน้าที่ต่างกันแบบเอกรูป (Unifrom DIF) และแบบอนเอกรูป (Nonunifrom DIF) เพื่อศึกษาว่าวิธีใด ภายใต้การใช้เกณฑ์จับคู่แบบใดสามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบในแบบทดสอบพหุมิติได้ดีที่สุด

บรรณานุกรม

- จิตติมา วรรณศรี. (2539). การเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ที่ต่างกันของข้อสอบด้วยวิธีแมนเทิล-แฮนส์เชลกับวิธีชิปเทสท์ เมื่อความยาวของแบบสอบ ขนาดกลุ่มตัวอย่างและอัตราส่วนกลุ่มอ้างอิงและกลุ่มเปรียบเทียบแตกต่างกัน. วิทยานิพนธ์ปริญญาครุศาสตรมหาบัณฑิต, สาขาวิชาการวัดและประเมินผลการศึกษา, บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย.
- นพมาศ พิพัฒน์สุข. (2541). การเปรียบเทียบประสิทธิภาพระหว่างวิธีแมนเทิล-แฮนส์เชลกับวิธีถดถอยโลจิสติกในการตรวจสอบการทำหน้าที่ที่ต่างกันของข้อสอบ เมื่อใช้เกณฑ์จับคู่เปรียบเทียบแตกต่างกันในแบบสอบชนิดพหุมิติ. วิทยานิพนธ์ปริญญาครุศาสตรมหาบัณฑิต, สาขาวิชาการวัดและประเมินผลการศึกษา, บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย.
- นิคม กิรติวรังกุล. (2542). การเปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ที่ต่างกันของข้อสอบระหว่างวิธีวิเคราะห์องค์ประกอบจำกัด แมนเทิล-แฮนส์เชลและการตอบสนองข้อสอบ. วิทยานิพนธ์ปริญญาครุศาสตรมหาบัณฑิต, สาขาวิชาการวัดและประเมินผลการศึกษา, บัณฑิตวิทยาลัย, จุฬาลงกรณ์มหาวิทยาลัย.
- ศิริชัย กาญจนวาสี. (2545). ทฤษฎีการทดสอบแนวใหม่ (พิมพ์ครั้งที่ 2). กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.
- เสรี ชัดเข้ม. (2540). วิธีการทางสถิติที่ใช้ตรวจสอบข้อสอบทำหน้าที่ที่ต่างกัน. *วารสารมหาวิทยาลัยบูรพา*, 2(1), 41-53.
- Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland, and H. Wainer (Eds.), *Differential Item Functioning*. 3-23. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Clauser, B. E., Mazor, K. M., & Hambleton, R. K. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel. *Applied Measurement in Education*, 6(4), 269-279.
- Clauser, B. E., Nungester, R. J., Mazor, K. M., & Ripkey, D. (1996). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional. *Journal of Educational Measurement*, 33(3), 202-214.
- Kim, S. H., & Cohen, A. S. (1994). An investigation of Lord's procedure for the detection of differential item functioning. *Applied Psychological Measurement*, 18(3), 217-228.
- 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.
- Mazor, K. M., & Clauser, B. E. (1995). Using logistic regression and the Mantel-Haenszel with Multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement*, 32(2), 131-144.
- 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.

- Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling*, 5(2), 107-124.
- Roger, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, 17(2), 105-116.
- Rudner, L. W., Getson, P. R., & Knight, D. L. (1980). A monte carlo comparison of seven biased item detection techniques. *Journal of Educational Measurement*, 17(1), 1-10.
- Swaminathan, H., & Roger, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27(4), 361-370.
- Stout, W. F. (1999). *SIBTEST user manual [Computer program manual]*. Urbana-Champaign: University of Illinois, Department of Psychological Measurement.
- Uttaro, T., & Millsap, R. E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. *Applied Psychological Measurement*, 18(1), 15-225.
- Wiersma, W., & Jurs, S.G. (1990). *Educational measurement and testing*. (2 nd ed.). Boston: Allyn & Bacon.