

**การเปรียบเทียบวิธีโพลีโทมัสซิบเทสต์ วิธีการวิเคราะห์ฟังก์ชันการจำแนกโลจิสติก
และวิธีการถดถอยโลจิสติกแบบจัดอันดับ ในการตรวจสอบการทำหน้าที่เบี่ยงเบน
ของข้อสอบในแบบทดสอบที่มีการให้คะแนนแบบหลายค่า**

**A Comparison of The Polytomous Sibtest, Logistic Discriminant Function Analysis
and Ordinal Logistic Regression procedures on detecting differential item
functioning in polytomously scored items**

สุชาติ สิริมินนันท์¹

Suchat Siriminnunte¹

บทคัดย่อ

การศึกษาครั้งนี้มีจุดมุ่งหมายเพื่อศึกษา และเปรียบเทียบอัตราความถูกต้องและอัตราความคลาดเคลื่อนประเภทที่ 1 จากการตรวจสอบการทำหน้าที่เบี่ยงเบนของข้อสอบในแบบทดสอบที่มีการให้คะแนนแบบหลายค่าด้วยวิธีโพลีโทมัสซิบเทสต์ วิธีการวิเคราะห์ฟังก์ชันการจำแนกโลจิสติก และวิธีการถดถอยโลจิสติกแบบจัดอันดับ ในการศึกษาครั้งนี้ใช้ข้อมูลจำลองโดยจำลองภายใต้โมเดลพาร์เซียลเครดิททั่วไปแบบมิติเดียวของมูรากิ ซึ่งแต่ละข้อจะมีรายการตอบ 5 รายการ ให้คะแนนเป็น 1, 2, 3, 4 หรือ 5 คะแนน ข้อมูลดังกล่าวจำลองภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือรูปแบบของข้อสอบทำหน้าที่เบี่ยงเบน 2 รูปแบบ ความยาวของแบบทดสอบ 3 ขนาด สัดส่วนของข้อสอบทำหน้าที่เบี่ยงเบน 3 ขนาด และขนาดตัวอย่าง 3 ขนาด รวมข้อมูลที่ใช้ศึกษาทั้งหมด 54 เงื่อนไข ในแต่ละเงื่อนไขจำลองซ้ำ 50 ครั้ง

ผลการวิจัยพบว่า 1) อัตราความถูกต้องในการตรวจสอบการทำหน้าที่เบี่ยงเบนของข้อสอบในแบบทดสอบที่มีการให้คะแนนแบบหลายค่า วิธีการวิเคราะห์ฟังก์ชันการจำแนกโลจิสติกสูงกว่าวิธีโพลีโทมัสซิบเทสต์และวิธีการถดถอยโลจิสติกแบบจัดอันดับในทุกเงื่อนไข 2) อัตราความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบการทำหน้าที่เบี่ยงเบนของข้อสอบในแบบทดสอบที่มีการให้คะแนนแบบหลายค่า วิธีการวิเคราะห์ฟังก์ชันการจำแนกโลจิสติกต่ำกว่าวิธีโพลีโทมัสซิบเทสต์และวิธีการถดถอยโลจิสติกแบบจัดอันดับในทุกเงื่อนไข 3) โดยภาพรวมวิธีการวิเคราะห์ฟังก์ชันการจำแนกโลจิสติกมีประสิทธิภาพในการตรวจสอบการทำหน้าที่เบี่ยงเบนของข้อสอบสูงกว่าวิธีโพลีโทมัสซิบเทสต์และวิธีการถดถอยโลจิสติกแบบจัดอันดับ

คำสำคัญ: การทำหน้าที่เบี่ยงเบนของข้อสอบ วิธีโพลีโทมัสซิบเทสต์ วิธีการวิเคราะห์ฟังก์ชันการจำแนกโลจิสติก วิธีการถดถอยโลจิสติกแบบจัดอันดับ

¹ สาขาการทดสอบและวัดผลการศึกษา บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

Abstract

The purpose of this research is to study and compare the power and type I error rates from the detection of differential item functioning (DIF) with Polytomous SIBTEST, Logistic discriminant function analysis and Ordinal logistic regression procedures. In this study, I used simulated data under the unidimensional generalized partial credit model according to Muraki. Type of all items were five response categories scoring as 1, 2, 3, 4 or 5. This data was simulated under a variety of factors: 2 differing levels form of DIF, 3 differing levels of test length, 3 differing levels of proportion of DIF items, and 3 differing levels of sample sizes. A total of 54 conditions was studied. The data were replicated 50 times for each condition.

The major findings are as follow: 1) The Power rate of Logistic discriminant function analysis on procedure detecting differential item functioning in polytomously scored items is higher than Polytomous SIBTEST and Ordinal logistic regression procedures in every condition. 2) The Type I error rate of Logistic discriminant function analysis procedure on detecting differential item functioning in polytomously scored items is lower than Polytomous SIBTEST and Ordinal logistic regression procedure in every condition. 3) The Logistic discriminant function analysis is more efficient than Polytomous SIBTEST and Ordinal logistic regression procedures in detecting differential item functioning.

Keywords: differential item functioning, Polytomous SIBTEST, Logistic discriminant function analysis, Ordinal logistic regression

บทนำ

ในการวัดผลถ้าจะวัดได้ถูกต้องตรงกับสิ่งที่ต้องการวัด จะต้องวัดด้วยเครื่องมือวัดที่ดีมีคุณภาพ ดัชนีบ่งชี้คุณภาพของเครื่องมือวัดที่สำคัญแบ่งเป็นสองกลุ่มคือ ดัชนีบ่งชี้คุณภาพรายข้อกับดัชนีบ่งชี้คุณภาพทั้งฉบับ ดัชนีบ่งชี้คุณภาพรายข้อประกอบด้วย ความยากง่าย และอำนาจจำแนก ส่วนดัชนีบ่งชี้คุณภาพทั้งฉบับ ประกอบด้วย หลักฐานแสดงความเที่ยงตรง และหลักฐานแสดงความเชื่อมั่น ความเที่ยงตรงถือเป็นลักษณะสำคัญที่สุดของเครื่องมือวัดที่ดีมีคุณภาพ นอกจากดัชนี

บ่งชี้คุณภาพของเครื่องมือวัดดังกล่าวมาแล้ว เครื่องมือวัดที่ดียังประกอบด้วยลักษณะที่สำคัญซึ่งได้แก่ ความยุติธรรม

ความยุติธรรมของแบบทดสอบ (Test Fairness) มีความสำคัญต่อผู้สร้างและผู้ใช้แบบทดสอบ ทั้งนี้เพราะปัจจุบันภาคส่วนต่างๆ อาทิภาคการศึกษา ธุรกิจ อุตสาหกรรม อื่นๆ ได้ใช้การสอบด้วยแบบทดสอบในการตรวจสอบความรู้ความสามารถของบุคคล เพื่อเป็นพื้นฐานสำหรับการตัดสินใจคัดเลือกบุคคลเข้าศึกษาต่อ บรรจุนาน เลื่อนตำแหน่ง หรือออกใบอนุญาต

การตัดสินใจเหล่านี้มีความสำคัญต่อบุคคล สังคม และนโยบายของรัฐอย่างเกี่ยวข้องกัน ดังนั้นแบบทดสอบที่นำมาใช้ควรเป็นแบบทดสอบที่เชื่อถือได้ วัดได้ตามคุณลักษณะที่ต้องการวัดเกิดความคลาดเคลื่อนในการวัดน้อยที่สุด รวมทั้งสามารถจำแนกความสามารถของผู้สอบได้ นอกจากนี้ควรเป็นแบบทดสอบที่มีความยุติธรรมระหว่างกลุ่มผู้สอบกลุ่มต่างๆที่มีลักษณะบางอย่างต่างกัน เช่น เพศ เชื้อชาติ วัฒนธรรม หรือภูมิหลังอื่นๆของผู้สอบ เพื่อไม่ให้เกิดการบิดเบือนในผลการสอบของผู้สอบกลุ่มเฉพาะเหล่านี้ ซึ่งจะช่วยให้ผู้มีหน้าที่เกี่ยวข้องกับการคัดเลือกได้นำไปใช้เป็นข้อมูลในการพิจารณาตัดสินผลการคัดเลือกให้มีความถูกต้องและเป็นธรรม ในทางตรงกันข้าม ถ้าแบบทดสอบที่นำมาใช้ในการคัดเลือกขาดความยุติธรรม มีความลำเอียงเกิดขึ้น โดยข้อสอบหรือแบบทดสอบเข้าข้างผู้สอบกลุ่มหนึ่งมากกว่าอีกกลุ่มหนึ่ง จะส่งผลให้คะแนนสอบแตกต่างกันระหว่างกลุ่ม ทั้งๆที่ทั้งสองกลุ่มมีความสามารถที่ต้องการวัดเท่ากัน ทำให้เกิดการได้เปรียบเสียเปรียบกัน การตรวจสอบความลำเอียงของข้อสอบ (Item Bias) ซึ่งถือเป็นส่วนหนึ่งของการพัฒนาแบบทดสอบจึงมีความจำเป็นสำหรับนักพัฒนาแบบทดสอบ เพื่อใช้เป็นหลักฐานแสดงความยุติธรรมของแบบทดสอบ โดยคัดเลือกข้อที่มีความลำเอียงออกจากแบบทดสอบ

ปัจจุบันนักวิจัยส่วนใหญ่ใช้คำว่า “การทำหน้าที่เบี่ยงเบนของข้อสอบ” (Differential Item Functioning) เรียกสั้นๆ ว่า “DIF” แทนคำว่า “ความลำเอียงของข้อสอบ” (Item Bias) คำสองคำนี้ไม่เหมือนกันแต่มีความเกี่ยวข้องกัน กล่าวคือการทำหน้าที่เบี่ยงเบนของข้อสอบ เป็นคุณลักษณะทางสถิติของข้อสอบที่ตรวจได้ด้วยวิธีการวิเคราะห์ทางสถิติ ซึ่งเป็นส่วนประกอบหนึ่งของสิ่งที่แสดงถึงความลำเอียงของข้อสอบ วิธีการทางสถิติที่ใช้

ในการตรวจสอบการทำหน้าที่เบี่ยงเบนของข้อสอบเป็นเงื่อนไขที่จำเป็นในการตัดสินความลำเอียงของข้อสอบ แต่ไม่ใช่เงื่อนไขที่เพียงพอในการตัดสินความลำเอียงของข้อสอบ

ในการตรวจสอบDIF จะทำการเปรียบเทียบผลการตอบข้อสอบระหว่างผู้สอบ กลุ่มย่อยสองกลุ่มที่มีความสามารถระดับเดียวกัน โดยที่ผู้สอบกลุ่มหนึ่งเรียกว่า “กลุ่มอ้างอิง” เป็นกลุ่มสำหรับเปรียบเทียบกับอีกกลุ่มหนึ่ง เรียกว่า “กลุ่มสนใจ” ถือเป็นกลุ่มเป้าหมายที่ต้องการศึกษา ข้อสอบที่ใช้ในการตรวจสอบเรียกว่า “ข้อสอบที่ศึกษา” ถ้าข้อสอบทำหน้าที่เบี่ยงเบนแล้ว ความน่าจะเป็นในการตอบข้อสอบถูกต้องของผู้สอบแต่ละกลุ่มจะไม่เท่ากัน การตรวจสอบ DIF ในระยะแรกมุ่งเน้นการตรวจสอบข้อสอบที่ให้คะแนนแบบสองค่า โดยใช้คะแนนสังเกตภายใต้ทฤษฎีการทดสอบมาตรฐานเดิม (CTT) เป็นเกณฑ์ในการจับคู่กลุ่มผู้สอบ เช่น วิธีแมนเทิล-แฮนส์เซล วิธีการทำให้เป็นมาตรฐาน และ วิธีการถดถอยโลจิสติก เป็นต้น ต่อมามีการนำทฤษฎีการตอบข้อสอบ (IRT) มาประยุกต์ใช้ทำให้การใช้เกณฑ์การจับคู่กลุ่มผู้สอบเปลี่ยนไป โดยจะใช้ค่าประมาณระดับความสามารถของผู้สอบ ซึ่งเป็นตัวแปรคุณลักษณะแฝงเป็นเกณฑ์การจับคู่กลุ่มผู้สอบ เช่น วิธีการทดสอบไค-สแควร์ของลอร์ด วิธีการวัดพื้นที่ เป็นต้น จากนั้นมีการพัฒนาวิธีการตรวจสอบ DIF ของข้อสอบที่ให้คะแนนแบบหลายค่า ซึ่งส่วนมากปรับขยายมาจากวิธีเดิมของทั้ง 2 กลุ่ม เช่น วิธีโพลีโตมัสซิบเทสท์ (Poly – SIB) วิธีการวิเคราะห์ฟังก์ชันการจำแนกโลจิสติก (LDFA) และวิธีการถดถอยโลจิสติกแบบจัดอันดับ (OLR) เป็นต้น จะเห็นว่าวิธีการตรวจสอบ DIF ได้มีการคิดค้นและพัฒนาอย่างต่อเนื่องเป็นลำดับ ซึ่งมีประเด็นที่น่าสนใจ 3 ประเด็น กล่าวคือ

ประเด็นแรก การจัดการศึกษาในปัจจุบัน ครูจะต้องประเมินความก้าวหน้าในผลสัมฤทธิ์ทางการเรียนของนักเรียนที่มาจากการจัดกิจกรรมการเรียนการสอนที่เน้นนักเรียนเป็นศูนย์กลาง และปฏิบัติจริง โดยใช้การวัดและประเมินผลหลากหลายรูปแบบ เช่น วิธีการวัด ประเมินการปฏิบัติที่ใช้แบบทดสอบที่มีการให้คะแนนแบบหลายค่า เพราะเชื่อว่าการให้คะแนนข้อสอบแบบหลายค่าได้ให้สารสนเทศเกี่ยวกับแบบแผน การตอบและความเชื่อมั่นจากการตอบที่สูงกว่าการให้คะแนนข้อสอบแบบสองค่า ข้อสอบที่ให้คะแนนแบบหลายค่าจึงเป็นประเด็นหนึ่งที่น่าสนใจในครั้งนี้

ประเด็นที่สอง ความเที่ยงตรงของการประเมิน DIF ขึ้นอยู่กับคุณภาพของตัวแปรการจับคู่ นั่นคือ วิธีต่างๆ ที่ใช้คุณลักษณะแฝงเป็นเกณฑ์การจับคู่ มีความน่าเชื่อถือในคะแนนแบบทดสอบ สำหรับวิธีที่ใช้คะแนนรวมเป็นเกณฑ์การจับคู่ จะมีคุณค่าทางทฤษฎีเช่นกัน ถ้าข้อมูลเป็นแบบมิติเดียว และเป็นไปตามโมเดล PCM หรือ GPCM วิธีการตรวจสอบ DIF ในข้อสอบที่ให้คะแนนแบบหลายค่า และใช้คะแนนรวมเป็นเกณฑ์การจับคู่กลุ่มผู้สอบที่น่าสนใจได้แก่ วิธี Poly-SIB วิธี LDFA และวิธี OLR

ประเด็นที่สาม ในการเลือกใช้วิธีการตรวจสอบ DIF ของข้อสอบที่ให้คะแนนแบบหลายค่าเพื่อให้ผลการตรวจสอบมีประสิทธิภาพ และประสิทธิผล ควรคำนึงถึงปัจจัยร่วมที่นำมาใช้ในการตรวจสอบ เช่น รูปแบบของข้อสอบทำหน้าที่เบี่ยงเบน ความยาวของแบบทดสอบ สัดส่วนของข้อสอบทำหน้าที่เบี่ยงเบน และขนาดตัวอย่าง เป็นต้น เพื่อเป็นสารสนเทศส่วนหนึ่งในการตัดสินใจเลือกใช้วิธีการตรวจสอบวิธีใดวิธีหนึ่ง ดังนั้นปัจจัยที่แปรเปลี่ยน 4 ปัจจัยดังกล่าวเป็นประเด็นที่สงสัยว่ามีผลต่อประสิทธิภาพใน

การตรวจสอบของวิธี Poly-SIB วิธี LDFA และ วิธี OLR หรือไม่ทั้งสามวิธีสามารถตรวจสอบ DIF ได้ดีในเงื่อนไขใด

จากประเด็นดังกล่าวผู้วิจัยจึงนำวิธี Poly-SIB วิธี LDFA และวิธี OLR มาใช้ตรวจสอบ DIF ของข้อสอบที่ให้คะแนนแบบหลายค่า ภายใต้ปัจจัยรูปแบบของข้อสอบทำหน้าที่เบี่ยงเบน ความยาว ของแบบทดสอบ สัดส่วนของข้อสอบที่ทำหน้าที่เบี่ยงเบน และขนาดตัวอย่าง

วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาค่า P และค่า E ในการตรวจสอบ DIF ของข้อสอบที่ให้คะแนนแบบหลายค่า ด้วยวิธี Poly-SIB วิธี LDFA และวิธี OLR ภายใต้เงื่อนไขต่างกันของปัจจัยที่แปรเปลี่ยน 4 ปัจจัย ซึ่งประกอบด้วยรูปแบบของข้อสอบทำหน้าที่เบี่ยงเบน ความยาวของแบบทดสอบ สัดส่วนของข้อสอบทำหน้าที่เบี่ยงเบน และขนาดตัวอย่าง

2. เพื่อเปรียบเทียบค่า P และ ค่า E ในการตรวจสอบ DIF ของข้อสอบที่ให้คะแนนแบบหลายค่าด้วยวิธี Poly-SIB วิธี LDFA และ วิธี OLR ภายใต้เงื่อนไขต่างกันของปัจจัยที่แปรเปลี่ยน 4 ปัจจัย ซึ่งประกอบด้วยรูปแบบของข้อสอบทำหน้าที่เบี่ยงเบน ความยาวของแบบทดสอบ สัดส่วนของข้อสอบทำหน้าที่เบี่ยงเบน และขนาดตัวอย่าง

3. เพื่อเปรียบเทียบค่า P และ ค่า E ในการตรวจสอบ DIF ของข้อสอบที่ให้คะแนนแบบหลายค่าระหว่างวิธี Poly-SIB วิธี LDFA และวิธี OLR ภายใต้เงื่อนไขต่างกันของปัจจัยที่แปรเปลี่ยน 4 ปัจจัย ซึ่งประกอบด้วย รูปแบบของข้อสอบทำหน้าที่เบี่ยงเบน ความยาวของแบบทดสอบ สัดส่วนของข้อสอบทำหน้าที่เบี่ยงเบน และขนาดตัวอย่าง

ขอบเขตของการวิจัย

ประชากรและกลุ่มตัวอย่าง

ข้อมูลที่ศึกษาเป็นข้อมูลจำลอง โดยจำลองภายใต้โมเดลพาร์เซียลเครดิตทั่วไปแบบมิติเดียว ข้อสอบแต่ละข้อมีรายการตอบ 5 รายการ โดยให้คะแนนเป็น 1, 2, 3, 4 หรือ 5 ในการจำลองข้อมูลจะจำลองผลการตอบข้อสอบภายใต้เงื่อนไขของปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือ รูปแบบของข้อสอบทำหน้าที่เบี่ยงเบน 2 รูปแบบ ความยาวของแบบทดสอบ 3 ขนาด สัดส่วนของข้อสอบทำหน้าที่เบี่ยงเบน 3 ขนาด และขนาดตัวอย่าง 3 ขนาด รวมข้อมูลทั้งหมดที่ต้องจัดกระทำเพื่อตรวจสอบ DIF จำนวน 54 เงื่อนไข ($2 \times 3 \times 3 \times 3$) ในแต่ละเงื่อนไขจำลองข้อมูลซ้ำ 50 ครั้ง

ตัวแปรที่ใช้ในการวิจัย

1. ตัวแปรอิสระ มี 5 ตัวแปร ดังนี้

1.1 วิธีการตรวจสอบ DIF มี 3 วิธี คือ วิธี Poly – SIB วิธี LDFA และวิธี OLR

1.2 รูปแบบของข้อสอบทำหน้าที่เบี่ยงเบนมี 2 รูปแบบคือ ข้อสอบทำหน้าที่เบี่ยงเบนที่เป็นรูปแบบเดียวกัน และข้อสอบทำหน้าที่เบี่ยงเบนที่ไม่เป็นรูปแบบเดียวกัน

1.3 ความยาวของแบบทดสอบ มี 3 ขนาด คือ จำนวน 20 ข้อ, 40 ข้อ และ 60 ข้อ

1.4 สัดส่วนของข้อสอบทำหน้าที่เบี่ยงเบนมี 3 ขนาด คือ จำนวน 10%, 20% และ 30%

1.5 ขนาดตัวอย่างประกอบด้วยกลุ่มอ้างอิงและกลุ่มสนใจที่มีจำนวนผู้สอบในแต่ละกลุ่มเท่ากันมี 3 ขนาด คือ จำนวน 250 คน 500 คน และ 1000 คน

2. ตัวแปรตาม มี 2 ตัวแปร คือ อัตราความถูกต้อง (P) และอัตราความคลาดเคลื่อนประเภทที่ 1 (E)

สมมติฐานของการวิจัย

1. ปัจจัยที่แปรเปลี่ยน 4 ปัจจัย ประกอบด้วยรูปแบบของข้อสอบทำหน้าที่เบี่ยงเบน ความยาวของแบบทดสอบ สัดส่วนของข้อสอบทำหน้าที่เบี่ยงเบน และขนาดตัวอย่าง มีผลทำให้ค่า P และค่า E ในการตรวจสอบ DIF ด้วยวิธี Poly-SIB วิธี LDFA และวิธี OLR แตกต่างกัน

2. วิธี Poly-SIB วิธี LDFA และ วิธี OLR ในการตรวจสอบ DIF ภายใต้เงื่อนไขปัจจัยที่แปรเปลี่ยน 4 ปัจจัย ประกอบด้วยรูปแบบของข้อสอบทำหน้าที่เบี่ยงเบน ความยาวของแบบทดสอบ สัดส่วนของข้อสอบทำหน้าที่เบี่ยงเบน และขนาดตัวอย่าง มีค่า P และ ค่า E แตกต่างกัน

วิธีดำเนินการวิจัย

1. จำลองข้อมูล ภายใต้โมเดลพาร์เซียลเครดิตทั่วไป (GPCM) ของมูรากิ (Muraki, 1992) ที่มีขนาดความยาวแบบทดสอบ 20 ข้อ, 40 ข้อ และ 60 ข้อ ในข้อสอบแต่ละข้อจะวัดความสามารถเป้าหมายเพียงมิติเดียว และแต่ละข้อมีรายการตอบ 5 รายการ ให้คะแนนเป็น 1, 2, 3, 4 หรือ 5 สำหรับการประมาณค่าพารามิเตอร์ความสามารถของผู้สอบ และประมาณค่าพารามิเตอร์ข้อสอบจากความสามารถผู้สอบของข้อสอบที่มีการตรวจให้คะแนนแบบหลายค่าจะใช้โปรแกรม WinGen2 (Han and Hambleton, 2007) จำลองผลการตอบข้อสอบภายใต้เงื่อนไขของปัจจัย 4 ปัจจัย คือ

1.1 รูปแบบของข้อสอบทำหน้าที่เบี่ยงเบน 2 รูปแบบ (ข้อสอบทำหน้าที่เบี่ยงเบนที่เป็นรูปแบบเดียวกัน และข้อสอบทำหน้าที่เบี่ยงเบนที่ไม่เป็นรูปแบบเดียวกัน)

1.2 ความยาวของแบบทดสอบ 3 ขนาด (20 ข้อ, 40 ข้อ และ 60 ข้อ) 3) สัดส่วนของข้อสอบทำหน้าที่เบี่ยงเบน 3 ขนาด (10%,

20% และ 30%) 4) ขนาดตัวอย่าง 3 ขนาด (250 คน, 500 คน และ 1,000 คน) ซึ่งมีชั้นตอนหลัก 3 ชั้นตอน ดังนี้

ขั้นแรก เป็นการสร้างข้อมูลผู้สอบโดยระบุจำนวนผู้สอบของขนาดตัวอย่าง เช่น 250 คน จากนั้นเลือกชนิดของการแจกแจงคุณลักษณะหรือความสามารถของผู้สอบโดยเลือกการแจกแจงแบบโค้งปกติ

ขั้นที่สอง เป็นการสร้างข้อมูลข้อสอบโดยระบุจำนวนของข้อสอบในแบบทดสอบ เช่น 20 ข้อ พร้อมทั้งระบุจำนวนรายการคำตอบโดยเลือก '5' ในการประมาณค่าพารามิเตอร์ต่อจากนั้นเลือกการแจกแจงของพารามิเตอร์ข้อสอบและระบุคุณลักษณะของการแจกแจงโดยเลือกพารามิเตอร์ α และ b ที่มีการแจกแจงแบบโค้งปกติ ทั้งนี้ผู้วิจัยศึกษารูปแบบของข้อสอบทำหน้าที่เบี่ยงเบน 2 รูปแบบคือ ข้อสอบทำหน้าที่เบี่ยงเบนที่เป็นรูปแบบเดียวกัน และที่ไม่เป็นรูปแบบเดียวกัน รูปแบบแรกกำหนดให้พารามิเตอร์ความยาก(b)ของข้อสอบระหว่างกลุ่มอ้างอิงและกลุ่มสนใจมีค่าแตกต่างกัน 0.25 ส่วนพารามิเตอร์อำนาจจำแนก(α)ของข้อสอบระหว่างกลุ่มอ้างอิงและกลุ่มสนใจมีค่าเท่ากัน สำหรับรูปแบบหลังกำหนดให้อำนาจจำแนกข้อสอบระหว่างกลุ่มอ้างอิงและกลุ่มสนใจมีค่าต่างกัน 1.0 ส่วนพารามิเตอร์ความยากของข้อสอบระหว่างกลุ่มอ้างอิงและกลุ่มสนใจมีค่าเท่ากัน โดยมีสัดส่วนของข้อสอบทำหน้าที่เบี่ยงเบน 3 ขนาด คือ 10%, 20% และ 30% ของแต่ละขนาดความยาวของแบบทดสอบ

ขั้นสุดท้าย เป็นการสร้างข้อมูลการตอบข้อสอบ ในขั้นนี้เป็นการนำ เซตข้อมูลผู้ตอบที่สร้างขึ้นในขั้นแรก และเซตข้อมูลพารามิเตอร์ข้อสอบที่สร้างขึ้นในขั้นที่สอง มาจัดทำข้อมูลรายการคำตอบของกลุ่มตัวอย่างที่จะทำการศึกษา ในแต่

ละเงื่อนไขจะทำข้อมูลซ้ำ 50 ชุด ซึ่งได้ข้อมูลที่เป็นไปตามเงื่อนไขทั้งสิ้น 2,700 ชุด

2. นำข้อมูลที่ได้จากการจำลองทั้งหมดมาตรวจสอบ DIF 3 วิธี คือ วิธี Poly-SIB วิธี LDA และวิธี OLR โดยใช้โปรแกรมคอมพิวเตอร์แล้วตัดลิ้นข้อสอบทำหน้าที่เบี่ยงเบนที่ระดับนัยสำคัญ .05 จากนั้นนำผลการตรวจสอบ DIF ทั้งหมดมาคำนวณค่า P และ ค่า E

3. การเปรียบเทียบค่า P และ ค่า E ของวิธีการตรวจสอบ DIF ด้วยวิธี Poly-SIB วิธี LDA และวิธี OLR ภายใต้ปัจจัยที่แปรเปลี่ยน 4 ปัจจัยคือ รูปแบบของข้อสอบทำหน้าที่เบี่ยงเบน (ข้อสอบทำหน้าที่เบี่ยงเบนที่เป็นรูปแบบเดียวกัน และข้อสอบทำหน้าที่เบี่ยงเบนที่ไม่เป็นรูปแบบเดียวกัน) ความยาวของแบบทดสอบ (20 ข้อ, 40 ข้อ และ 60 ข้อ) สัดส่วนของข้อสอบทำหน้าที่เบี่ยงเบน (10%, 20% และ 30%) และขนาดตัวอย่าง (250 คน, 500 คน และ 1,000 คน) ใช้ t-test ทดสอบเงื่อนไขของปัจจัยรูปแบบของข้อสอบทำหน้าที่เบี่ยงเบน และ F-test ทดสอบเงื่อนไขของปัจจัยความยาวของแบบทดสอบ เงื่อนไขของปัจจัยสัดส่วนของข้อสอบทำหน้าที่เบี่ยงเบน และเงื่อนไขของปัจจัยขนาดตัวอย่าง ซึ่งเมื่อมีนัยสำคัญทางสถิติแล้ว จะทดสอบความแตกต่างรายคู่โดยใช้การทดสอบของเชฟเฟ่ สำหรับการเปรียบเทียบ ค่า P และ ค่า E ในการตรวจสอบ DIF ระหว่าง วิธีตรวจสอบทั้งสามวิธีภายใต้ปัจจัยดังกล่าว ใช้ F-test ทดสอบความแตกต่างของวิธีตรวจสอบทั้งสามวิธี ภายใต้เงื่อนไขของแต่ละปัจจัย เมื่อมีนัยสำคัญทางสถิติแล้ว ผู้วิจัยทำการทดสอบความแตกต่างรายคู่โดยใช้การทดสอบของเชฟเฟ่

ผลการวิจัย

ผลการตรวจสอบ DIF

1. ผลการตรวจสอบ DIF ด้วยวิธีที่ใช้ตรวจสอบ

1.1 การทดสอบปัจจัยรูปแบบของข้อสอบทำหน้าที่เบี่ยงเบน พบว่า วิธี Poly – SIB และวิธี LDFA มีค่า P ในการตรวจพบข้อสอบทำหน้าที่เบี่ยงเบนที่เป็นรูปแบบเดียวกัน สูงกว่าที่ไม่เป็นรูปแบบเดียวกัน ส่วนวิธี OLR มีค่า P ในการตรวจพบข้อสอบทำหน้าที่เบี่ยงเบนที่ไม่เป็นรูปแบบเดียวกันสูงกว่าที่เป็นรูปแบบเดียวกัน เมื่อพิจารณา ค่า E พบว่า วิธี LDFA และวิธี OLR มีค่า E ในการตรวจพบ DIF ของข้อสอบทำหน้าที่เบี่ยงเบนที่เป็นรูปแบบเดียวกันต่ำกว่าที่ไม่เป็นรูปแบบเดียวกัน ซึ่งตรงข้ามกับวิธี Poly-SIB

1.2 การทดสอบปัจจัยความยาวของแบบทดสอบพบว่าเมื่อความยาวของแบบทดสอบเพิ่มขึ้น ค่า P ของวิธี LDFA มีค่าเพิ่มขึ้น สำหรับวิธี Poly-SIB และวิธี OLR มีค่า P สูงขึ้นเมื่อความยาวของแบบทดสอบเพิ่มขึ้นจาก 20 ข้อเป็น 40 ข้อ และลดลงเมื่อความยาวของแบบทดสอบเพิ่มขึ้นจาก 40 ข้อ เป็น 60 ข้อ เมื่อพิจารณา ค่า E พบว่า เมื่อความยาวของแบบทดสอบเพิ่มขึ้นจาก 20 ข้อเป็น 40 ข้อ วิธี Poly-SIB มีค่า E เพิ่มขึ้นขณะที่วิธี LDFA และ วิธี OLR มีค่า E ลดลงและเมื่อความยาวของแบบทดสอบเพิ่มขึ้นจาก 40 ข้อ เป็น 60 ข้อ ทั้ง 3 วิธี มีค่า E ลดลง

1.3 การทดสอบปัจจัยสัดส่วนของข้อสอบทำหน้าที่เบี่ยงเบน พบว่า เมื่อขนาดสัดส่วนของข้อสอบทำหน้าที่เบี่ยงเบนเพิ่มขึ้น วิธี Poly-SIB และวิธี OLR มีค่า P สูงขึ้นในทุกเงื่อนไข สำหรับวิธี LDFA มีค่า P ในเงื่อนไขสัดส่วนของข้อสอบทำหน้าที่เบี่ยงเบน 10% สูงกว่าที่ระดับ 20% เมื่อพิจารณาค่า E พบว่าเมื่อขนาดสัดส่วนของข้อสอบทำหน้าที่เบี่ยงเบน เพิ่มขึ้นวิธี Poly-SIB

และวิธี OLR มีค่าสูงขึ้นในทุกเงื่อนไข ส่วนวิธี LDFA ในเงื่อนไขสัดส่วนของข้อสอบทำหน้าที่เบี่ยงเบน 10% และ 30% มีค่าใกล้เคียงกัน

1.4 การทดสอบปัจจัยขนาดตัวอย่างพบว่า เมื่อขนาดตัวอย่างเพิ่มขึ้น ค่า P ในการตรวจสอบของแต่ละวิธีมีค่าเพิ่มขึ้นในทุกเงื่อนไข เมื่อพิจารณาค่า E พบว่า วิธี Poly – SIB และวิธี OLR มีค่า E เพิ่มขึ้นเมื่อขนาดตัวอย่างเพิ่มขึ้น ซึ่งตรงข้ามกับการตรวจสอบด้วยวิธี LDFA ที่พบว่าเมื่อขนาดตัวอย่างเพิ่มขึ้นมีผลทำให้ค่า E ลดลง

2. ผลการตรวจสอบ DIF ภายใต้อัจฉริยะที่ศึกษา พบว่า ค่า P ระหว่างวิธี Poly – SIB วิธี LDFA และวิธี OLR ในการตรวจสอบ DIF ภายใต้อัจฉริยะที่แปรเปลี่ยน 4 ปัจจัยคือรูปแบบของข้อสอบทำหน้าที่เบี่ยงเบน ความยาวของแบบทดสอบ สัดส่วนของข้อสอบทำหน้าที่เบี่ยงเบนและขนาดตัวอย่าง พบว่าภายใต้อัจฉริยะต่างๆ วิธี Poly – SIB, วิธี LDFA และวิธี OLR มีค่า P แตกต่างกัน (ยกเว้นเงื่อนไขขนาดสัดส่วนของข้อสอบทำหน้าที่เบี่ยงเบน 20%, ขนาดตัวอย่าง 250 คนและ 500 คนของวิธี Poly – SIB และวิธี OLR) กล่าวคือ วิธี LDFA มีค่า P สูงกว่าวิธี Poly – SIB และวิธี OLR อย่างมีนัยสำคัญทางสถิติที่ระดับ.01

ส่วนผลการเปรียบเทียบค่า E พบว่าภายใต้อัจฉริยะต่างๆ วิธี Poly – SIB วิธี LDFA และวิธี OLR มีค่า E แตกต่างกัน กล่าวคือวิธี LDFA มีค่า E ต่ำกว่าวิธี Poly – SIB และวิธี OLR อย่างมีนัยสำคัญทางสถิติที่ระดับ.01

อภิปรายผลการวิจัย

1. ปัจจัยรูปแบบของข้อสอบทำหน้าที่เบี่ยงเบน

ผลการตรวจสอบ พบว่า ปัจจัยรูปแบบของข้อสอบทำหน้าที่เบี่ยงเบนมีผลต่อค่า P และค่า E ของทั้ง 3 วิธี ผลดังกล่าวเป็นไปตามสมมติฐานข้อ 3.1 โดยผลการตรวจสอบ DIF ภายใต้ปัจจัยข้อสอบทำหน้าที่เบี่ยงเบนที่เป็นรูปแบบเดียวกัน ทั้งสามวิธีมีค่า P เกิน 50% ขณะที่ผลการตรวจสอบ DIF ภายใต้เงื่อนไขข้อสอบทำหน้าที่เบี่ยงเบนที่ไม่เป็นรูปแบบเดียวกัน วิธี Poly – SIB มีค่า P 43.5% ซึ่งต่ำที่สุดและต่ำกว่าการตรวจสอบด้วยวิธี LDFA และวิธี OLR ที่มีค่า P เฉลี่ยประมาณ 97.8% และ 59.0% ตามลำดับ ทั้งนี้เป็นเพราะวิธี Poly – SIB ไม่มีความไวในการตรวจสอบข้อสอบทำหน้าที่เบี่ยงเบนที่ไม่เป็นรูปแบบเดียวกัน เนื่องจากวิธี Poly – SIB พัฒนามาจากวิธีซิบเทสท์ ซึ่งเป็นวิธีที่ใช้ตรวจสอบ DIF ของข้อสอบที่มีทิศทางเดียวโดยเฉพาะ ดังนั้นจึงมีข้อจำกัดในการตรวจสอบ DIF ที่ไม่มีทิศทางเดียวกัน (Li; & Stout. 1996)

2. ปัจจัยความยาวของแบบทดสอบ

ผลการตรวจสอบ พบว่าปัจจัยความยาวของแบบทดสอบมีผลต่อค่า P และค่า E ของทั้ง 3 วิธีผลดังกล่าวเป็นไปตามสมมติฐานข้อ 3.1 เมื่อเพิ่มความยาวแบบทดสอบจาก 20 ข้อ เป็น 40 ข้อ ค่า P ในการตรวจสอบ DIF ของวิธี Poly-SIB วิธี LDFA และวิธี OLR จะเพิ่มขึ้นประมาณ 7.4%, 3.2% และ 37.3% ตามลำดับ ซึ่งสอดคล้องกับสวามินาธานและโรเจอร์ส (Swaminathan;& Rogers. 1990) ที่ว่าแบบทดสอบที่ยาวและขนาดตัวอย่างที่มาก ทำให้โอกาสในการตรวจพบข้อสอบทำหน้าที่เบี่ยงเบนสูง แต่เมื่อความยาวแบบทดสอบเป็น 60 ข้อ อัตราการตรวจสอบของวิธี Poly – SIB และวิธี OLR

ลดลง ซึ่งขัดแย้งกับค่ากล่าวข้างต้น ทั้งนี้อาจเนื่องมาจากเมื่อความยาวแบบทดสอบเป็น 60 ข้อ ซึ่งเป็นแบบทดสอบที่ยาว จะทำให้ค่าความยากเฉลี่ยของข้อสอบระหว่างกลุ่มอ้างอิงและกลุ่มสนใจมีค่าใกล้เคียงกัน ทำให้ทั้งสองกลุ่มมีความสามารถใกล้เคียงกัน จึงทำให้ตรวจพบข้อสอบทำหน้าที่เบี่ยงเบนได้ถูกต้องลดลง

เมื่อพิจารณาค่า E พบว่าปัจจัยความยาวของแบบทดสอบ มีผลต่อค่า E กล่าวคือเมื่อแบบทดสอบมีความยาวมากขึ้นค่า E จะลดลง ดังเช่น วิธี LDFA มีค่า E ในแบบทดสอบที่มีความยาว 20 ข้อ 40 ข้อ และ 60 ข้อ เฉลี่ยประมาณ 4.7%, 2.2% และ 0.1% ตามลำดับ ขณะที่วิธี OLR มีค่า E เฉลี่ยประมาณ 37.1%, 23.2% และ 8.7% ตามลำดับ ซึ่งสอดคล้องกับผลการศึกษาของอุททาโรและมิลส์แซบ (Uttaro; & Millsap. 1994) สำหรับวิธี Poly – SIB มีค่า E ในแบบทดสอบที่มีความยาว 20 ข้อ และ 40 ข้อ เฉลี่ยประมาณ 19.0% และ 52.4% ตามลำดับ ซึ่งเป็นไปในทิศทางเพิ่มขึ้นผลดังกล่าวสอดคล้องกับ ผลการศึกษาของสวามินาธานและโรเจอร์ส (Swaminathan; & Rogers. 1990)

3. ปัจจัยสัดส่วนของข้อสอบทำหน้าที่เบี่ยงเบน

ผลการตรวจสอบ DIF พบว่า ปัจจัยสัดส่วนของข้อสอบทำหน้าที่เบี่ยงเบน มีผลต่อค่า P และค่า E ของทั้ง 3 วิธี ผลดังกล่าวเป็นไปตามสมมติฐานข้อ 3.1 กล่าวคือ เมื่อสัดส่วนของข้อสอบทำหน้าที่เบี่ยงเบนเพิ่มขึ้นจาก 10% ถึง 20% วิธี LDFA มีค่า P ลดลง และมีค่า E เพิ่มขึ้น ซึ่งสอดคล้องกับผลการศึกษานารายานาน และสวามินาธาน (Narayanan; & Swaminathan. 1996) ขณะที่วิธี Poly – SIB และวิธี OLR มีค่า P และค่า E เพิ่มขึ้นซึ่งไม่สอดคล้อง

กับผลการศึกษาของอรินทร์ น่วมถนอม (2549) ที่พบว่าสัดส่วนของข้อสอบทำหน้าที่เบี่ยงเบน ในช่วง 10% ถึง 20% ไม่มีผลต่อวิธี Poly – SIB และผลการศึกษาของโอชิม่าและมิลเลอร์ (Oshima; & Miller. 1992) ที่พบว่าเมื่อสัดส่วนของข้อสอบทำหน้าที่เบี่ยงเบนเพิ่มขึ้นแล้ว ค่า P จะลดลง และการศึกษาของโคเฮนและคิม (Cohen;& Kim. 1993) ซึ่งพบว่าเมื่อสัดส่วนของข้อสอบทำหน้าที่เบี่ยงเบนมีจำนวนลดลงแล้ว ค่า E จะมีค่าเพิ่มขึ้น

4. ปัจจัยขนาดตัวอย่าง

ผลการตรวจสอบ DIF พบว่า ปัจจัยขนาดตัวอย่าง มีผลต่อค่า P และค่า E ของทั้ง 3 วิธี ผลดังกล่าวเป็นไปตามสมมติฐานข้อ 3.1 กล่าวคือ เมื่อใช้กลุ่มตัวอย่างขนาด 250 คนต่อกลุ่ม ทั้ง 3 วิธีมีค่า P ต่ำสุดเมื่อเทียบกับขนาดตัวอย่างอื่น ๆ โดยวิธี Poly – SIB วิธี LDFA และวิธี OLR สามารถตรวจสอบได้ถูกต้อง 46.2%, 96.9% และ 47.2% ตามลำดับ และเมื่อใช้ขนาดตัวอย่างมากขึ้น เป็น 500 คน ค่า P ของทั้ง 3 วิธีจะยังคงเพิ่มขึ้น ในขณะที่ขนาดตัวอย่างเป็น 1000 คนต่อกลุ่ม ค่า P ของวิธี Poly – SIB วิธี LDFA และวิธี OLR สามารถตรวจสอบได้ถูกต้อง 60.40%, 98.90% และ 63.10% ตามลำดับ ซึ่งสอดคล้องกับผลการศึกษาของนารายานานและสวามินาธาน (Narayanan; & Swaminathan.1996) ที่พบว่าเมื่อขนาดตัวอย่างเพิ่มขึ้นจะมีผลทำให้ค่า P เพิ่มขึ้น สำหรับการตรวจสอบค่า E พบว่า วิธี LDFA สามารถควบคุมค่า E ให้อยู่ในเกณฑ์ที่กำหนดร้อยละ 10 ได้ทุกเงื่อนไข และจะมีค่าลดลงเมื่อขนาดตัวอย่างมีค่าเพิ่มขึ้น ขณะที่วิธี Poly – SIB และวิธี OLR มีแนวโน้มว่าเมื่อ ขนาดตัวอย่างเพิ่มขึ้นค่า E จะเพิ่มสูงขึ้นด้วย ซึ่งสอดคล้องกับ ผลการศึกษาของอุทัยวรรณ สายพัฒนา (2547) ที่พบว่า ความผิดพลาดของการตรวจสอบ DIF ด้วยวิธี

Poly-SIB เมื่อกลุ่มตัวอย่างมีขนาดใหญ่ขึ้น ความผิดพลาดในการตรวจสอบจะมีค่ามากขึ้น ด้วย แต่ขัดแย้งกับการศึกษาของนารายานานและสวามินาธาน (Narayanan; & Swaminathan. 1996) ที่พบว่าขนาดตัวอย่างไม่มีผลกระทบต่อค่า E

5. ผลการเปรียบเทียบค่า P และค่า E ของวิธีตรวจสอบ DIF

ผลการเปรียบเทียบค่า P และค่า E ระหว่างวิธี Poly – SIB วิธี LDFA และวิธี OLR ในการตรวจสอบ DIF ภายใต้เงื่อนไขของปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือรูปแบบของข้อสอบ ทำหน้าที่เบี่ยงเบน ความยาวของแบบทดสอบ สัดส่วนของข้อสอบทำหน้าที่เบี่ยงเบน และขนาดตัวอย่าง พบว่าภายใต้แต่ละเงื่อนไขทั้ง 3 วิธี มีค่า P และค่า E แตกต่างกัน ผลการเปรียบเทียบดังกล่าวสอดคล้องกับ สมมติฐานข้อ 3.2 โดยวิธี LDFA มีประสิทธิภาพในการตรวจสอบ สูงสุดเพราะแต่ละเงื่อนไขมีค่า P มากกว่า 90% ขึ้นไป ซึ่งถือว่ามีความถูกต้องเพียงพอ และค่า E อยู่ในขอบเขตไม่เกิน 5%

ข้อเสนอแนะการวิจัย

ข้อเสนอแนะสำหรับการปฏิบัติ

1. ผลการศึกษาค้นคว้าครั้งนี้พบว่า การตรวจสอบ DIF ภายใต้เงื่อนไขของปัจจัยที่แปรเปลี่ยน 4 ปัจจัย คือรูปแบบของข้อสอบ ทำหน้าที่เบี่ยงเบน ความยาวของแบบทดสอบ สัดส่วนของข้อสอบทำหน้าที่เบี่ยงเบน และขนาดตัวอย่าง โดยภาพรวมภายใต้ทุกเงื่อนไข วิธี LDFA ตรวจสอบได้ดีที่สุดเมื่อเทียบกับวิธี Poly – SIB และวิธี OLR ทั้งนี้เพราะวิธีดังกล่าว ให้ผลการตรวจสอบ DIF มีค่า P สูงเกินร้อยละ 90 และมีค่า E ต่ำไม่เกินร้อยละ 5 ในทุกเงื่อนไข

2. ผู้ปฏิบัติควรระลึกว่าการทำหน้าที่เบี่ยงเบนของข้อสอบไม่ใช่ความลำเอียงของ

ข้อสอบ ถ้าข้อสอบข้อใดข้อหนึ่งทำหน้าที่เบี่ยงเบน
ไม่ได้หมายความว่าข้อสอบข้อนั้นไม่ดีเสมอไป

ข้อเสนอแนะสำหรับการวิจัยครั้งต่อไป

1. ควรทำการศึกษาซ้ำด้วยวิธีการเดิม
คือเปรียบเทียบวิธี Poly – SIB วิธี LDFA และ
วิธี OLR ภายใต้ปัจจัยอื่นๆ เช่น ความแตกต่างของ

การแจกแจงความสามารถขนาดตัวอย่างระหว่าง
กลุ่มอ้างอิงและกลุ่มสนใจที่ไม่ใช่อัตราส่วน 1:1
หรือศึกษาที่ระดับนัยสำคัญต่าง ๆ กัน เช่น .001
หรือศึกษาจากข้อมูลจริง

2. ควรศึกษาเพื่อเปรียบเทียบ ผลการ
ตรวจสอบ DIF ของหมวดข้อสอบ

เอกสารอ้างอิง

- อรินทร์ น่วมถนอม. (2549). การเปรียบเทียบวิธีโพลี-ชิปเทสท์ วิธีการถดถอยโลจิสติกแบบจัดอันดับ
และวิธีการจัดอันดับหลายมิติ ในการตรวจสอบการทำหน้าที่เบี่ยงเบนของข้อสอบที่วัด
ความสามารถหลายมิติ และให้คะแนนหลายค่า. ปรินซ์นิพนธ์ กศ.ด. (การทดสอบและวัดผล
การศึกษา). กรุงเทพฯ: บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ.
- อุทัยวรรณ สายพัฒนะ. (2547). การเปรียบเทียบประสิทธิภาพของผลการตรวจสอบการทำหน้าที่
ต่างกันของข้อสอบ ในแบบทดสอบที่มีการให้คะแนนแบบหลายค่าระหว่างวิธี GMH และวิธี
Poly-SIBTEST. ปรินซ์นิพนธ์ กศ.ด. (การทดสอบและวัดผลการศึกษา). กรุงเทพฯ: บัณฑิต
วิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ.
- Cohen, A.S.; & Kim, S.H. (1993). A comparison of Lord's chi-square and Raju's area measures in
detection of DIF. *Applied Psychological Measurement*. 17(1): 39-52.
- Han, K.T.; & Hambleton, R.K. (2007). *User's Manual WinGen2*. Amherst, MA: University of
Massachusetts, Center for Educational Assessment.
- Li, H-H.; & Stout, W. (1996). A new procedure for detecting crossing DIF. *Psychometrika*. 61(4):
647- 677.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM Algorithm. *Applied
Psychological Measurement*. 16(2): 159-176.
- Narayanan, P.; & Swaminathan, H. (1996). Identification of items that show nonuniform DIF.
Applied Psychological Measurement. 20 (3): 257-274.
- Oshima, T.C.; & Miller, M.D. (1992). Multidimensional and item bias in item Response theory.
Applied Psychological Measurement. 16(3): 237-248.
- Swaminathan, H.; & Rogers, H.J. (1990) . Detecting differential item functioning using logistic
Regression Procedures. *Journal of Educational Measurement*. 27(4): 361-370.
- Uttaro, T.; & Millsap, R.E. (1994). Factors influencing the Mantel-Haenszel procedure in the
detection of differential item functioning. *Applied Psychological Measurement*. 18(1):15-25.