การใช้จีอาร์เอ็ม จีพีซีเอ็มและโมเดลโลจิสติก ในการเปรียบเทียบฟังก์ชัน สารสนเทศของแบบวัด ที่มีวิธีการให้คะแนนต่างกัน*

(THE USE OF GRM, GPCM AND LOGISTIC MODELS IN THE COMPARISONS OF INFORMATION FUNCTIONS OF TESTS WITH DIFFERENT SCORING METHODS)

ธนวัฒน์ แสนสุข

ABSTRACT

The purpose of this research was to study the use of GRM, GPCM and logistic models to compare test information functions of tests employing dichotomous (0, 1) and polytomous (1, 2, 3, 4) scoring methods. Data obtained from dichotomous and polytomous scoring methods were analyzed based on 1, 2, 3 parameter logistic models, and based on GRM and GPCM, respectively. Two data sets were conducted in this study : (1) general affective scale data of 6,300 Prathom Suksa 6 students from the Office of Educational Assessment and Testing Service, Department of Curriculum and Instruction Development, Ministry of Education (2) mathematics achievement data of 470 Prathom Suksa 6 students collected by the researcher. Detections of person and item fits were performed through RSM using BIGSTEPS

*วิทยานิพนธ์ครุศาสตรมหาบัณฑิต สาขาการวัดและประเมินผลการศึกษา ปีการศึกษา 2538 โดยมีรองศาสตราจารย์ ดร.สุวิมล ว่องวาณิช และผู้ช่วยศาสตราจารย์ ดร.นงลักษณ์ วิรัชชัย เป็นอาจารย์ที่ปรึกษา

program. Test information functions were determined by MULTILOG and PARSCALE programs. Confirmatory factor analysis was used to examine construct validity of general affective scale through LISREL program.

Results showed that polytomous scoring method based on GRM provided higher test information function than dichotomous scoring method. Nevertheless, there was no clear evidence whether the polytomous scoring method based on GPCM yielded higher test information function than dichotomous scoring method.

บทคัดย่อ

การวิจัยนี้มีจุดมุ่งหมายเพื่อศึกษาการใช้ GRM, GPCM และโมเดลโลจิสติกในการเปรียบเทียบค่าฟังก์ชันสารสนเทศของแบบวัดที่มีการตรวจให้คะแนนแบบทวิวิภาค (0, 1) และแบบพหุวิภาค (1, 2, 3, 4) ข้อมูลจากการตรวจให้คะแนนแบบทวิวิภาควิเคราะห์ตามโมเดลโลจิสติก 1, 2 และ 3 พารามิเตอร์ ส่วนการตรวจให้คะแนนแบบพหุวิภาควิเคราะห์ตาม GRM และ GPCM การวิจัยครั้งนี้ทำการศึกษากับข้อมูล 2 ชุด (1) ข้อมูลผลการวัดคุณลักษณะทั่วไป ของนักเรียนชั้นประถมศึกษาปีที่ 6 จำนวน 6,300 คน จากสำนักงานทดสอบทางการศึกษา กรมวิชาการ กระทรวงศึกษาธิการ (2) ข้อมูลผลสัมฤทธิ์ทางคณิตศาสตร์ของนักเรียนชั้นประถมศึกษาปีที่ 6 จำนวน 470 คน ที่ผู้วิจัยเก็บเอง ทำการตรวจสอบความสมบูรณ์ของข้อมูลโดยวิเคราะห์ความเหมาะสมของผู้ตอบและข้อกระทงตาม RSM ด้วยโปรแกรม BIGSTEPS และวิเคราะห์เพื่อหาค่าฟังก์ชันสารสนเทศของแบบวัดโดยใช้โปรแกรม MULTILOG และ PARSCALE จากนั้นตรวจสอบความตรงเชิงโครงสร้างของแบบวัดคุณลักษณะทั่วไปโดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยัน ด้วยโปรแกรม LISREL

ผลการวิจัยพบว่าการตรวจให้คะแนนแบบพหุวิภาคเมื่อวิเคราะห์ตาม GRM ให้ค่าฟังก์ชันสารสนเทศสูงกว่าการตรวจให้คะแนนแบบทวิวิภาค สำหรับการตรวจให้คะแนนแบบพหุวิภาควิเคราะห์ตาม GPCM

และแบบทวิวิภาคยังไม่สามารถสรุปได้ว่าวิธีใดให้ค่าฟังก์ชันสารสนเทศสูงกว่ากัน

ความเป็นมาและความสำคัญของปัญหา

แบบวัด (tests) เป็นเครื่องมือวัดผลทางการศึกษาประเภทหนึ่งที่นิยมใช้กันมาก เนื่องจากมีความสะดวกในการนำไปใช้ คือช่วยกำหนดระดับของคุณลักษณะออกมาในรูปคะแนน ซึ่งเป็นข้อมูลเชิงปริมาณ แบบวัดดังกล่าวแบ่งได้เป็น 2 ชนิดตามลักษณะการนำไปใช้ คือ (1) แบบวัดคุณลักษณะ โดยทั่วไปมักเป็นมาตรประมาณค่า (rating scale) (2) แบบสอบวัดผลสัมฤทธิ์ทางการเรียน (achievement tests) แบบวัดทั้ง 2 ชนิดดังกล่าวมีวิธีให้คะแนน 2 แบบ คือ การให้คะแนนแบบทวิวิภาค (0, 1) และการให้คะแนนแบบพหุวิภาค อย่างไรก็ดีในการนำแบบวัดคุณลักษณะที่เป็นมาตรประมาณค่าแบบตัวเลือกบังคับตอบ (force - choice ratings) และแบบสอบวัดผลสัมฤทธิ์ไปใช้พบว่ามีการตรวจให้คะแนน 2 แบบ คือ แบบทวิวิภาค โดยให้ตัวเลือกที่ถูกต้องหรือเหมาะสมที่สุดเป็น 1 คะแนน ส่วนตัวเลือกอื่น ๆ เป็น 0 คะแนน (วารี นิยมธรรม, 2536; บุษรินทร์ บุญรอด, 2536) ซึ่งวิธีนี้ทำได้สะดวกและรวดเร็ว และอีกวิธีหนึ่งคือการให้คะแนนแบบพหุวิภาค โดยกำหนดค่าคะแนนให้กับตัวเลือกทุกตัวแตกต่างกันตามระดับของคุณลักษณะ ซึ่งการใช้วิธีนี้พบได้ในงานวิจัยทั่วไป (สุพัตรา เทียนอุดม, 2536; สุดใจ ขันทองคำ, 2534; บุญทอง บุญทวี, 2534)

การให้คะแนนแบบวัดทั้ง 2 วิธีดังกล่าว เคยมีการศึกษาในเชิงเปรียบเทียบคุณภาพโดยเปรียบเทียบค่าฟังก์ชันสารสนเทศตามแนวทฤษฎีการตอบสนองข้อสอบ (IRT) ซึ่งได้ข้อสรุปที่ขัดแย้งกัน คือ โดนัฟ (Donoghue, 1994) สมิธ (Smith, 1987) เฟรรี และฮัทชินสัน (Frary and Hutshinson, 1982) พบว่าแบบวัดที่ใช้วิธีให้คะแนนแบบพหุวิภาคให้ค่าฟังก์ชันสารสนเทศสูงกว่าการตรวจให้คะแนนแบบทวิวิภาค ส่วนผลการศึกษาของ ยามาโมโต และคูลิก (Yamamoto and Kulick, 1992 cited by Donoghue, 1994) ได้ข้อสรุปว่าการให้คะแนนแบบทวิวิภาคให้ค่าฟังก์ชันสารสนเทศสูงกว่าการให้คะแนนแบบพหุวิภาค นอกจากนี้ เวนเนอร์ และทิซเซน (Wainer and Thissen, 1993 cited by Donoghue, 1994) ได้ให้ข้อสังเกตว่า การตรวจให้คะแนนแบบพหุวิภาคจะต้องใช้เวลาและค่าใช้จ่ายสูงกว่าการตรวจให้คะแนนแบบทวิวิภาค ซึ่งเขาได้พิจารณาถึงความคุ้มค่ามากกว่าคุณภาพด้านความตรง และความเที่ยง จึงเป็นประเด็นว่าการตรวจให้คะแนนในแบบวัดคุณลักษณะ และแบบสอบวัดผลสัมฤทธิ์ด้วย 2 วิธีดังกล่าว วิธีใดจะเหมาะสมกว่ากัน

การวัดที่ตรงและเที่ยงนอกจากจะพิจารณาถึงวิธีการให้คะแนนที่เหมาะสมแล้ว การวิเคราะห์เพื่อประมาณค่าความสามารถของผู้สอบ และตรวจสอบคุณภาพของข้อกระทงก็เป็นอีกกระบวนการหนึ่งที่จะทำให้ได้สารสนเทศเกี่ยวกับบุคคลและข้อกระทงที่มีความตรงและเที่ยง การวิเคราะห์ดังกล่าวจำแนกเป็น 2 แนว คือ (1) การวิเคราะห์ตามแนวทฤษฎีการทดสอบแบบดั้งเดิม (classical test theory : CTT) วิธีนี้มีข้อจำกัดเนื่องจากค่าสถิติที่ได้มักแปรเปลี่ยนไปตามกลุ่มตัวอย่าง (2) การวิเคราะห์ตามแนวทฤษฎีการตอบสนองข้อสอบ (item response theory : IRT) วิธีนี้จะให้ค่าพารามิเตอร์ที่สามารถอ้างอิงได้กับกลุ่มตัวอย่างทั่วไป ซึ่งจำแนกการวิเคราะห์เป็น 2 แนวทางตามวิธีการให้คะแนน คือ (1) การวิเคราะห์ตาม Dichotomous IRT Model ซึ่งใช้กับแบบวัดที่ตรวจให้คะแนนแบบทวิวิภาค ได้แก่การวิเคราะห์ตามโมเดลโลจิสติก 1, 2 และ 3 พารามิเตอร์ เป็นต้น (2) การวิเคราะห์ตาม Polytomous IRT Model ซึ่งใช้กับแบบวัดที่ตรวจให้คะแนนพหุวิภาค มีด้วยกันหลายโมเดล (Donoghue, 1994; Muraki, 1993) เช่น Graded Response Model (GRM), Nominal Response Model (NRM), Rating Scale Model (RSM), Partial Credit Model (PCM), Successive Interval Model (SIM) และ Generalized Partial Credit Model (GPCM) สำหรับโมเดลที่มีความยืดหยุ่นและมีการนำไปใช้กันมากในปัจจุบัน คือ GRM ซึ่งพัฒนามาจากโมเดลโลจิสติก 2 พารามิเตอร์ และ GPCM ซึ่งพัฒนามาจากโมเดลราส์ช 1 พารามิเตอร์ (Donoghue, 1994; Muraki, 1993) จากวิธีการวิเคราะห์ดังกล่าวจึงเป็นประเด็นสำหรับการวิจัยตามมาว่าวิธีให้คะแนนทั้ง 2 วิธี คือ แบบทวิวิภาคและแบบพหุวิภาค ที่วิเคราะห์ตามโมเดลโลจิสติก GRM และ GPCM เมื่อใช้กับแบบวัดคุณลักษณะและแบบสอบวัดผลสัมฤทธิ์จะให้ค่าฟังก์ชันสารสนเทศต่างกันหรือไม่ ถ้าต่างกัน วิธีใดจะให้ค่าฟังก์ชันสารสนเทศสูงกว่ากัน

วัตถุประสงค์ของการวิจัย

1. เพื่อเปรียบเทียบค่าฟังก์ชันสารสนเทศของแบบวัดคุณลักษณะและแบบสอบ (TIF) สำหรับการตรวจให้คะแนนแบบทวิวิภาคและแบบพหุวิภาคเมื่อวิเคราะห์ตาม GRM และ GPCM

2. เพื่อเปรียบเทียบอัตราส่วนสารสนเทศเฉลี่ย (ratio of average information : RAI) ของแบบวัดคุณลักษณะและแบบสอบสำหรับการตรวจให้คะแนนแบบทวิวิภาคและแบบพหุวิภาค เมื่อวิเคราะห์ตาม GRM และ GPCM

3. เพื่อตรวจสอบความสอดคล้องของค่าฟังก์ชันสารสนเทศของแบบวัดคุณลักษณะและแบบสอบระหว่างการวิเคราะห์ตาม GRM และ GPCM

สมมติฐานของการวิจัย

1. ผลการวิเคราะห์การตรวจให้คะแนนทั้ง 2 วิธี เมื่อวิเคราะห์ตาม GRM และ GPCM ทั้ง 2 โมเดลจะให้ผลที่สอดคล้องกัน คือ การตรวจให้คะแนนแบบพหุวิภาคที่วิเคราะห์ตาม GRM และ GPCM จะให้ค่าฟังก์ชันสารสนเทศ (TIF) สูงกว่าการตรวจให้คะแนนแบบทวิวิภาคที่วิเคราะห์ตามโมเดลโลจิสติก

2. ค่าสารสนเทศเฉลี่ย (RAI) ของแบบวัดที่ใช้วิธีการตรวจให้คะแนนแบบพหุวิภาคเมื่อวิเคราะห์ตาม GPCM จะให้ค่าฟังก์ชันสารสนเทศสูงกว่าการวิเคราะห์ตาม GRM

วิธีดำเนินการวิจัย

ประชากรและกลุ่มตัวอย่าง

ประชากรในการศึกษาครั้งนี้แบ่งเป็น 2 กลุ่ม คือ (1) นักเรียนชั้นประถมศึกษาปีที่ 6 ของทุกเขตการศึกษาที่เรียนในปีการศึกษา 2537 และเข้ารับการทดสอบโดยการใช้แบบวัดคุณลักษณะของสำนักทดสอบ กรมวิชาการ รวมทั้งสิ้น จำนวน 102,117 คน (2) นักเรียนชั้นประถมศึกษาปีที่ 6 ในสังกัดสำนักงานการประถมศึกษาอำเภอพระนครศรีอยุธยาที่กำลังเรียนในภาคเรียนที่ 2 ปีการศึกษา 2538 จำนวน 1,567 คน

กลุ่มตัวอย่างที่ศึกษามี 2 กลุ่ม คือ (1) นักเรียนที่เข้ารับการทดสอบโดยสำนักงานทดสอบทางการศึกษา กรมวิชาการ จำนวน 6,300 คน (2) นักเรียนในสังกัดสำนักงานการประถมศึกษาอำเภอพระนครศรีอยุธยา จำนวน 470 คน

ตัวแปรที่ศึกษา

1. ตัวแปรต้น

1.1 วิธีการตรวจให้คะแนนแบ่งเป็น 2 วิธี คือ การตรวจให้คะแนนแบบทวิวิภาคและการตรวจให้คะแนนแบบพหุวิภาค

1.2 โมเดลที่ใช้วิเคราะห์เปรียบเทียบ ในการศึกษาครั้งนี้มี 3 โมเดล คือ โมเดลการวิเคราะห์ GRM, GPCM และโมเดลโลจิสติก

2. ค่าฟังก์ชันสารสนเทศ

2.1 อัตราส่วนสารสนเทศเฉลี่ย (RAI)

2.2 ความสอดคล้องของค่าฟังก์ชันสารสนเทศของแบบวัดระหว่างการวิเคราะห์ตาม GRM และ GPCM

แบบวัดและแบบสอบ

เครื่องมือที่ใช้สำหรับเก็บรวบรวมข้อมูลในการศึกษาครั้งนี้มี 2 ชุด คือ

1. แบบวัดคุณลักษณะทั่วไป ที่พัฒนาขึ้นโดยสำนักทดสอบ กรมวิชาการ กระทรวงศึกษาธิการ ซึ่งมีลักษณะเป็นมาตรประมาณค่าแบบสถานการณ์บังคับเลือกตอบ มี 4 ตัวเลือก กำหนดให้คะแนนตามระดับของคุณลักษณะด้านจิตพิสัย (affective domain) เป็น 1, 2 3 และ 4 คะแนน จากต่ำไปสูงตามลำดับ

2. แบบสอบคณิตศาสตร์ ใช้วัดการแก้โจทย์ปัญหาคณิตศาสตร์สำหรับนักเรียนชั้นประถมศึกษาปีที่ 6 พัฒนาขึ้นโดย วลี เฉลยสมัย (2539) แบบสอบดังกล่าวมี 6 ข้อ แต่ละข้อแบ่งเป็น 3 ตอน คือ ตอนที่ 1 วัดความเข้าใจวิธีการแก้โจทย์ปัญหา ตอนที่ 2 วัดการคิดคำนวณ และตอนที่ 3 วัดความสามารถในการวิเคราะห์ตามสภาพที่เป็นจริง โดยทั้ง 3 ตอน มีสถานการณ์และตัวเลขเหมือนกันหมด การตรวจให้คะแนนแบบพหุวิภาคกับแบบสอบคณิตศาสตร์ในแต่ละข้อของแบบสอบทั้ง 3 ตอน กำหนดเป็น 4 ลำดับขั้น คือ นักเรียนที่ไม่ตอบหรือตอบไม่ถูกเลยทั้ง 3 ตอน ให้ 1 คะแนน นักเรียนที่ตอบถูกเพียง 1 ตอน ให้ 2 คะแนน นักเรียนที่ตอบถูก 2 ตอน ให้ 3 คะแนน และนักเรียนที่ตอบถูกทั้ง 3 ตอน ให้ 4 คะแนน

การวิเคราะห์เพื่อคัดเลือกผู้ตอบและข้อกระทงที่เหมาะสม

ในการวิจัยครั้งนี้ ได้ดำเนินการศึกษาจากข้อมูล 2 ชุด ข้อมูลชุดที่ 1 เป็นข้อมูลจากการใช้แบบวัดทั้งหมดและผู้ตอบทั้งหมด ส่วนข้อมูลชุดที่ 2 เป็นข้อมูลจากผู้ตอบที่เหมาะสม (person-fit) และข้อกระทงที่เหมาะสม (item-fit) ตาม Rating Scale Model (RSM) คัดเลือกโดยใช้โปรแกรม BIGSTEPS และหลังจากคัดเลือกข้อกระทงที่เหมาะสมแล้วพบว่าแบบวัดคุณลักษณะ มีจำนวนข้อกระทงลดลง จึงได้ทำการตรวจสอบความตรงเชิงโครงสร้างโดยวิธีวิเคราะห์องค์ประกอบเชิงยืนยัน (confirmatory factor analysis : CFA) โดยใช้โปรแกรม LISREL หลังจากการวิเคราะห์พบว่าโครงสร้างของแบบวัดคุณลักษณะในข้อมูลชุดที่สองยังมีโครงสร้างของข้อกระทงที่สามารถวัดคุณลักษณะได้ไม่แตกต่างจากข้อมูลชุดที่ 1

การวิเคราะห์ข้อมูล

ดำเนินการวิเคราะห์ข้อมูลเบื้องต้นของคะแนนนักเรียน จากการใช้แบบวัดคุณลักษณะทั่วไป และแบบสอบคณิตศาสตร์ วิเคราะห์เพื่อคัดเลือกผู้ตอบที่เหมาะสม (person fit) และข้อกระทงที่เหมาะสม (item fit) โดยใช้โปรแกรมสำเร็จรูป BIGSTEPS วิเคราะห์องค์-ประกอบเชิงยืนยัน (confirmatory factor analysis : CFA) เพื่อตรวจสอบความตรงเชิงโครงสร้าง (construct validity) ของแบบวัดคุณลักษณะจากข้อมูลชุดที่ 1 และ 2 ในแบบวัดย่อยที่ 1 และ 2 ตามลำดับโดยใช้สถิติทดสอบไค-สแควร์ (chi-square) ด้วยโปรแกรมสำเร็จรูป LISREL วิเคราะห์หาค่าฟังก์ชันสารสนเทศของแบบวัดคุณลักษณะและแบบสอบจากข้อมูลชุดที่ 1 และ 2 ด้วย GRM, GPCM และโมเดลโลจิสติก ด้วยโปรแกรม MULTI-LOG และ PARSCALE คำนวณค่าสารสนเทศเฉลี่ย (average information : AI) และอัตราส่วนสารสนเทศเฉลี่ย (RAI) โดยคำนวณมือ

สรุปผลการวิจัย

1. ค่าฟังก์ชันสารสนเทศ (TIF) ของแบบวัดคุณลักษณะและแบบสอบคณิตศาสตร์ที่มีการตรวจให้คะแนนแบบทวิวิภาคและแบบพหุวิภาค

การวิเคราะห์ข้อมูลจากแบบวัดคุณลักษณะชุดที่ 1 และ 2 ตาม GRM ให้ผลการวิเคราะห์ที่สอดคล้องกัน คือ ในช่วง θ ตั้งแต่ -2 ถึง 0 การตรวจให้คะแนนแบบพหุวิภาคที่วิเคราะห์ตาม GRM ให้ค่าฟังก์ชันสารสนเทศสูงกว่าการตรวจให้คะแนนแบบทวิวิภาค ที่วิเคราะห์ตามโมเดลโลจิสติก 1, 2 และ 3 พารามิเตอร์ แต่ในช่วง θ ตั้งแต่ 0 ถึง 1.5 การตรวจให้คะแนนแบบทวิวิภาคที่วิเคราะห์ตามโมเดลโลจิสติก 3 พารามิเตอร์ ให้ค่าฟังก์ชันสารสนเทศสูงกว่าการตรวจให้คะแนนแบบพหุวิภาคที่วิเคราะห์ตาม GRM การวิเคราะห์ข้อมูลแบบสอบคณิตศาสตร์ชุดที่ 1 และ 2 ให้ผลที่สอดคล้องกัน คือ ในช่วง θ ตั้งแต่ -2 ถึง 0.5 การตรวจให้คะแนนแบบพหุวิภาคที่วิเคราะห์ตาม GRM ให้ค่าฟังก์ชันสารสนเทศสูงกว่าการตรวจให้คะแนนแบบทวิวิภาคที่วิเคราะห์ตามโมเดลโลจิสติก 1, 2 และ 3 พารามิเตอร์ แต่ในช่วง θ ตั้งแต่ 0.5 ถึง 2 การตรวจให้คะแนนแบบทวิวิภาคที่วิเคราะห์ตามโมเดลโลจิสติก 1, 2 และ 3 พารามิเตอร์ ให้ค่าฟังก์ชันสารสนเทศสูงกว่าการตรวจให้คะแนนแบบพหุวิภาคที่วิเคราะห์ตาม GRM นอกจากนี้ยังพบว่าข้อมูลจากแบบวัดคุณลักษณะชุดที่ 1 และ 2 ที่ใช้วิธีตรวจให้คะแนนแบบทวิวิภาคให้ผลที่สอดคล้องกัน คือในช่วง θ ตั้งแต่ -2 ถึง -0.5 การวิเคราะห์ตามโมเดลโลจิสติก 1 พารามิเตอร์ ให้ค่าฟังก์ชันสารสนเทศสูงสุด รองลงมาคือการวิเคราะห์ตาม

โมเดลโลจิสติก 2 และ 3 พารามิเตอร์ตามลำดับ

2. การเปรียบเทียบอัตราส่วนสารสนเทศเฉลี่ยของแบบวัดและแบบสอบ สำหรับวิธีตรวจให้คะแนนแบบทวิวิภาคและแบบพหุวิภาคเมื่อวิเคราะห์ตาม GRM และ GPCM

การเปรียบเทียบอัตราส่วนสารสนเทศเฉลี่ยจากข้อมูลแบบวัดคุณลักษณะชุดที่ 1 และ 2 ให้ผลการวิเคราะห์ที่สอดคล้องกัน คือ การตรวจให้คะแนนแบบพหุวิภาคที่วิเคราะห์ตาม GRM มีประสิทธิภาพสัมพัทธ์สูงกว่าการตรวจให้คะแนนแบบทวิวิภาค ที่วิเคราะห์ตามโมเดลโลจิสติก 1, 2 และ 3 พารามิเตอร์ และพบว่าการตรวจให้คะแนนแบบทวิวิภาคที่วิเคราะห์ตามโมเดลโลจิสติก 1, 2 และ 3 พารามิเตอร์ มีประสิทธิภาพสัมพัทธ์สูงกว่าการตรวจให้คะแนนแบบพหุวิภาคที่วิเคราะห์ตาม GPCM การวิเคราะห์ดังกล่าว เมื่อใช้กับแบบสอบคณิตศาสตร์ทั้งจากข้อมูลชุดที่ 1 และ 2 ก็ให้ผลที่สอดคล้องกัน คือ การตรวจให้คะแนนแบบพหุวิภาคที่วิเคราะห์ตาม GRM และ GPCM มีประสิทธิภาพสัมพัทธ์สูงกว่าการตรวจให้คะแนนแบบทวิวิภาค ที่วิเคราะห์ตามโมเดลโลจิสติก 1, 2 และ 3 พารามิเตอร์ ส่วนการเปรียบเทียบผลการตรวจให้คะแนนแบบพหุวิภาค การวิเคราะห์ตาม GRM มีประสิทธิภาพสัมพัทธ์สูงกว่าการวิเคราะห์ตาม GPCM เล็กน้อย

3. การพิจารณาความสอดคล้องของผลการตรวจให้คะแนนแบบทวิวิภาคและแบบพหุวิภาคจากการวิเคราะห์ตาม GRM และ GPCM

ผลการวิเคราะห์ตาม GRM และ GPCM ที่สอดคล้องกัน คือ ผลจากการนำไปใช้วิเคราะห์กับแบบสอบคณิตศาสตร์ โดยพบว่าในข้อมูลแบบสอบคณิตศาสตร์ ชุดที่ 1 และ 2 ที่วิเคราะห์ตาม GRM และ GPCM มีประสิทธิภาพสัมพัทธ์สูงกว่าการตรวจให้คะแนนแบบทวิวิภาคที่วิเคราะห์ตามโมเดลโลจิสติก 1, 2 และ 3 พารามิเตอร์ แต่ผลที่ขัดแย้งกัน คือ การนำไปใช้กับแบบวัดคุณลักษณะ โดยพบว่าการตรวจให้คะแนนแบบพหุวิภาคที่วิเคราะห์ตาม GRM มีประสิทธิภาพสัมพัทธ์สูงกว่าการตรวจให้คะแนนแบบทวิวิภาคที่วิเคราะห์ตามโมเดลโลจิสติก 1, 2 และ 3 พารามิเตอร์ แต่การตรวจให้คะแนนแบบพหุวิภาคที่วิเคราะห์ตาม GPCM มีประสิทธิภาพสัมพัทธ์ต่ำกว่าการตรวจให้คะแนนแบบทวิวิภาคที่วิเคราะห์ตามโมเดลโลจิสติก 1, 2 และ 3 พารามิเตอร์

อภิปรายผลการวิจัย

ผลจากการศึกษาครั้งนี้สรุปได้ว่าการตรวจให้คะแนนแบบพหุวิภาคเป็นวิธีที่เหมาะกว่าการตรวจให้คะแนนแบบทวิวิภาค โดยเฉพาะการตรวจให้คะแนนวิธีดังกล่าวเมื่อวิเคราะห์ตาม

GRM มีความเหมาะสมกว่าการวิเคราะห์ตาม GPCM และโมเดลโลจิสติก 1, 2 และ 3 พารามิเตอร์ ซึ่งสอดคล้องกับผลงานวิจัยของโดนัฟ (Donoghue, 1994) และซาเมจิมา (Samejima, 1976 cited by Donoghue, 1994) ที่ศึกษาเปรียบเทียบค่าฟังก์ชันสารสนเทศของแบบสอบประเภทเขียนตอบ โดยใช้วิธีตรวจให้คะแนนแบบพหุวิภาคและแบบทวิวิภาค และพบว่าการตรวจให้คะแนนแบบพหุวิภาคให้ค่าฟังก์ชันสารสนเทศสูงกว่าการตรวจให้คะแนนแบบทวิวิภาค นอกจากนี้ บอค ซิมสัน และทิซเซน (Bock, Sympson and Tissen cited by Drasgow, F, 1995) ยังได้แนะนำให้ใช้การตรวจให้คะแนนแบบพหุวิภาคกับแบบวัดทางจิตวิทยามากกว่าการตรวจให้คะแนนแบบทวิวิภาค เพราะคะแนนที่ได้ให้สารสนเทศมากกว่าการใช้การตรวจให้คะแนนแบบทวิวิภาค

ผลการวิจัยครั้งนี้ไม่เป็นไปตามสมมติฐานทั้ง 2 ข้อ คือการวิเคราะห์ตาม GRM และ GPCM ให้ผลไม่สอดคล้องกันเมื่อใช้กับเครื่องมือที่เป็นแบบวัดคุณลักษณะทั่วไป โดยพบว่าการตรวจให้คะแนนแบบพหุวิภาคที่วิเคราะห์ตาม GRM ให้ค่าสารสนเทศเฉลี่ยสูงกว่าการตรวจให้คะแนนแบบทวิวิภาคเมื่อวิเคราะห์ตามโมเดลโลจิสติกแบบ 1, 2 และ 3 พารามิเตอร์ แต่การตรวจให้คะแนนแบบพหุวิภาคที่วิเคราะห์ตาม GPCM พบว่าให้ค่าสารสนเทศเฉลี่ยต่ำกว่าการตรวจให้คะแนนแบบทวิวิภาคที่วิเคราะห์ตามโมเดลโลจิสติก 1, 2 และ 3 พารามิเตอร์ เมื่อพิจารณาจากงานวิจัยที่ผ่านมา (Donoghue, 1994 : Samejima, 1976 cited by Donoghue, 1994) พบว่างานวิจัยดังกล่าวล้วนใช้เครื่องมือที่เป็นแบบสอบประเภทวัดผลสัมฤทธิ์ทั้งสิ้น เช่น การศึกษาของโดนัฟ (Donoghue, 1994) ใช้เครื่องมือที่เป็นแบบสอบวัดความสามารถในการอ่าน (reading assessment) วิเคราะห์ด้วย GPCM ได้ข้อค้นพบที่สอดคล้องกันว่าการตรวจให้คะแนนแบบพหุวิภาคให้ค่าฟังก์ชันสารสนเทศสูงกว่าการตรวจให้คะแนนแบบทวิวิภาค การที่ผลการวิจัยครั้งนี้ไม่สอดคล้องกับสมมุติฐานดังกล่าว เป็นเพราะลักษณะของเครื่องมือที่ใช้มีอิทธิพลต่อการวิเคราะห์ด้วย GPCM เพราะเมื่อพิจารณาจากผลการใช้กับแบบสอบคณิตศาสตร์ การวิเคราะห์ตาม GPCM ให้ค่าฟังก์ชันสารสนเทศสูงกว่าการตรวจให้คะแนนแบบทวิวิภาคที่วิเคราะห์ตามโมเดลโลจิสติก ทั้งนี้เพราะนักเรียนที่ตอบแบบวัดคุณลักษณะมีการตอบแบบหลอก (faking) สูงกว่าการตอบแบบเดา (guessing) ในการทำแบบสอบ แบบวัดคุณลักษณะที่ใช้ในการศึกษาครั้งนี้เป็นมาตรประมาณค่าแบบตัวเลือกบังคับตอบ 4 ตัวเลือก ส่วนแบบสอบคณิตศาสตร์ประกอบด้วยข้อสอบชนิดเลือกตอบ 5 ตัวเลือก โดยรวมตัวเลือกปลายเปิดอยู่ด้วย ทำให้นักเรียนมีโอกาสการเดาถูกต่ำกว่าแบบวัดคุณลักษณะ

ผลการศึกษาที่ไม่เป็นไปตามสมมติฐานอีกประการหนึ่ง คือ การตรวจให้คะแนนแบบพหุวิภาคที่ใช้กับแบบวัดคุณลักษณะและแบบสอบคณิตศาสตร์ เมื่อวิเคราะห์ตาม GRM ให้ค่าสารสนเทศเฉลี่ยสูงกว่าการวิเคราะห์ตาม GPCM แสดงให้เห็นว่าการตรวจให้คะแนนแบบพหุวิภาคที่วิเคราะห์ตาม GRM มีความเหมาะสมกว่าการวิเคราะห์ตาม GPCM ซึ่งจะเห็นได้จากการวิจัยของ ดี อะยาลา De-Ayala (1992) ที่ศึกษาเปรียบเทียบ ผลการวิเคราะห์การตรวจให้คะแนนแบบพหุวิภาคตาม GRM และ PCM (pertial credit model) กับข้อมูลจำลอง (simulated examinees) พบว่าการตรวจให้คะแนนแบบพหุวิภาคที่วิเคราะห์ตาม GRM มีความแม่นยำในการประมาณค่าพารามิเตอร์ของผู้สอบมากกว่าการวิเคราะห์ตาม PCM ซึ่งเป็นโมเดลพื้นฐานที่ยังไม่ได้พัฒนาเป็น GPCM เหมือนในการศึกษาในครั้งนี้ อย่างไรก็ดีการที่ GPCM ให้ค่าฟังก์ชันสารสนเทศที่ต่ำกว่า GRM ก็ไม่ได้หมายความว่า GPCM เป็นโมเดลที่ไม่เหมาะสม มูรากิ (Muraki, 1993) ได้เคยศึกษาฟังก์ชันสารสนเทศของ GPCM ในการนำไปใช้วิเคราะห์แบบสอบของ NAEP (National Assessment of Educational Progress) และให้ข้อเสนอแนะว่าการวิเคราะห์ตาม GPCM สามารถทำให้ค่าฟังก์ชันสารสนเทศสูงขึ้นได้โดยต้องมีการยุบรวม (collapsing) ลำดับขั้นคะแนนที่มีค่าความยากใกล้เคียงกันเข้าด้วยกันหรือเรียงลำดับขั้นคะแนน (reordering categorical response) ใหม่ตามลำดับขั้นของความยากเพื่อให้มีความเหมาะสมกับ GPCM ซึ่งผลการวิเคราะห์แบบวัดคุณลักษณะตาม GPCM ในการวิจัยครั้งนี้พบว่ามีการเรียงลำดับขั้นคะแนนที่ไม่เหมาะสมด้วยเช่นกัน (ตารางที่ 7 และภาพที่ 5-6) ดังนั้นจึงควรศึกษาเพิ่มเติมในข้อมูลชุดนี้ โดยมีการยุบรวมหรือเรียงลำดับขั้นคะแนนใหม่แล้ววิเคราะห์ตาม GPCM และเปรียบเทียบกับผลการวิเคราะห์ตาม GRM ที่ไม่ต้องปรับลำดับขั้นคะแนน ว่าจะให้ผลสอดคล้องกับการศึกษาครั้งนี้หรือไม่

จากผลการศึกษาที่ผ่านมาและการศึกษาในครั้งนี้จะเห็นได้ว่า ไม่ว่าจะเป็นการศึกษากับเครื่องมือประเภทแบบวัดคุณลักษณะหรือแบบวัดผลสัมฤทธิ์ก็ตาม การตรวจให้คะแนนแบบพหุวิภาคที่วิเคราะห์ตาม GRM เป็นวิธีที่เหมาะสมกว่า และได้ข้อสังเกตประการหนึ่งว่าหากนำเครื่องมือดังกล่าวไปใช้กับนักเรียนที่มีความสามารถสูงมาก ๆ การตรวจให้คะแนนและการวิเคราะห์ตามโมเดลใดก็ตามจะให้ผลการวิเคราะห์ที่ไม่แตกต่างกันมากนัก แต่หากใช้กับกลุ่มที่มีความสามารถต่ำมาก การใช้วิธีการตรวจให้คะแนนแบบพหุวิภาคจะเป็นวิธีที่เหมาะสมกว่า

ข้อเสนอแนะ

1. ข้อเสนอแนะในการนำไปใช้

1.1 การใช้แบบวัดคุณลักษณะกับนักเรียนที่มีระดับของคุณลักษณะปานกลางค่อนไปทางต่ำ (θ ตั้งแต่ -2 ถึง 0) การตรวจให้คะแนนแบบพหุวิภาคที่วิเคราะห์ตาม GRM เป็นวิธีที่เหมาะสมกว่า หากนำไปใช้กับนักเรียนที่มีระดับของคุณลักษณะปานกลางค่อนไปทางสูง (θ ตั้งแต่ 0 ถึง 1.5) การตรวจให้คะแนนแบบทวิวิภาคที่วิเคราะห์ตามโมเดลโลจิสติก 3 พารามิเตอร์ จะเป็นวิธีการที่เหมาะสม หากใช้กับนักเรียนที่มีระดับของคุณลักษณะสูง (θ ตั้งแต่ 2 ขึ้นไป) การตรวจให้คะแนนและวิเคราะห์ตามโมเดลใดก็ตาม จะให้ผลที่ไม่แตกต่างกันมากนัก อย่างไรก็ดีหากต้องการนำแบบวัดคุณลักษณะดังกล่าวไปใช้ แต่ไม่ทราบระดับของคุณลักษณะกลุ่มตัวอย่าง การตรวจให้คะแนนแบบพหุวิภาคที่วิเคราะห์ตาม GRM จะเหมาะสมกว่า

1.2 การใช้แบบสอบคณิตศาสตร์กับนักเรียนที่มีความสามารถระดับใดก็ตาม การตรวจให้คะแนนแบบพหุวิภาคที่วิเคราะห์ตาม GRM และ GPCM จะให้ค่าฟังก์ชันสารสนเทศที่ใกล้เคียงกัน โดยเฉพาะในช่วงความสามารถของนักเรียนปานกลางค่อนไปทางต่ำ (θ ตั้งแต่ -2 ถึง 0.5) การตรวจให้คะแนนและการวิเคราะห์ดังกล่าวเหมาะสมกว่าการตรวจให้คะแนนแบบทวิวิภาคที่วิเคราะห์ตามโมเดลโลจิสติก 1, 2 และ 3 พารามิเตอร์ หากนำไปใช้กับนักเรียนที่มีความสามารถค่อนไปทางสูง (θ ตั้งแต่ 0.5 ถึง 2) การตรวจให้คะแนนแบบทวิวิภาคที่วิเคราะห์ตามโมเดลโลจิสติก 1, 2 และ 3 พารามิเตอร์ จะเป็นวิธีที่เหมาะสม แต่ถ้าหากไม่ทราบความสามารถของกลุ่มตัวอย่างการตรวจให้คะแนนแบบพหุวิภาคที่วิเคราะห์ตาม GRM และ GPCM จะเหมาะสมกว่า

2. ข้อเสนอแนะในการวิจัย

2.1 โปรแกรมคอมพิวเตอร์ที่ใช้วิเคราะห์ข้อมูลในการวิจัยครั้งนี้มีข้อจำกัดในการวิเคราะห์ตาม GRM ด้วยโปรแกรม MULTILOG ให้ค่าฟังก์ชันสารสนเทศของแบบวัดเฉพาะในช่วง θ ตั้งแต่ 2 ถึง +2 แต่ภายหลังผู้วิจัยพบว่ามีผู้พัฒนาโปรแกรม GR-GRAPH สำหรับใช้ร่วมกับ MULTILOG ซึ่งสามารถวิเคราะห์ให้ค่าฟังก์ชันสารสนเทศในช่วง θ ที่กว้างขึ้น คือตั้งแต่ -5 ถึง +5 ดังนั้นจึงควรมีการศึกษาในทำนองเดียวกันนี้ แต่ศึกษาในช่วง θ ที่กว้างขึ้นกว่านี้

2.2 การวิจัยครั้งนี้ไม่ได้กำหนดให้เครื่องมือที่ใช้เป็นตัวแปรต้นในการวิจัย ซึ่งจากผลการวิจัยพบว่าค่าฟังก์ชันสารสนเทศของแบบวัดเมื่อใช้เครื่องมือที่ต่างกันให้ค่าที่ต่างกันด้วย

ดังนั้นจึงควรมีการศึกษาโดยกำหนดให้เครื่องมือที่ใช้เป็นตัวแปรต้นในการวิจัย แล้วดูผลว่าเมื่อเครื่องมือที่ใช้ต่างกันจะให้ค่าฟังก์ชันสารสนเทศต่างกันด้วยหรือไม่

2.3 แบบวัดคุณลักษณะที่ศึกษาในครั้งนี้ใช้วิธีการตรวจให้คะแนนแบบพหุวิภาคที่มีลักษณะเป็นมาตรประมาณค่า (1, 2, 3, 4) อยู่แล้ว และผู้วิจัยได้นำวิธีดังกล่าวมาเปรียบเทียบกับการให้คะแนนแบบทวิวิภาคโดยปรับให้คะแนนในลำดับขั้นที่ 4 เป็น 1 คะแนน และลำดับขั้นอื่น ๆ เป็น 0 คะแนน ซึ่งเป็นข้อจำกัดที่ผู้วิจัยใช้ข้อมูลที่มีผู้รวบรวมไว้แล้ว ดังนั้นควรศึกษาว่าหากนำแบบวัดที่ใช้มาสร้างให้มีการให้คะแนนแบบทวิวิภาค (0-1) จริง ๆ ไม่ใช่เกิดจากการยุบรวมคะแนนจากมาตรประมาณค่าดังที่ได้ปรากฏในการวิจัยครั้งนี้ จะให้ผลที่สอดคล้องกันหรือไม่

เอกสารอ้างอิง

เอกสารภาษาไทย

บุญทอง บุญทวี. 2534. ***การสร้างแบบสอบวัดจิตพิสัยกลุ่มสร้างเสริมลักษณะนิสัยระดับชั้นประถมศึกษา.*** วิทยานิพนธ์ปริญญาครุศาสตรมหาบัณฑิต จุฬาลงกรณ์มหาวิทยาลัย.

บุษรินทร์ บุญรอด. 2536. ***การศึกษาความสัมพันธ์ระหว่างความท้อถอยกับบุคลิกภาพของช่างสำรวจ สำนักงานชลประทานที่ 1, 2 และ 3 สังกัดกองสำรวจภูมิประเทศ.*** วิทยานิพนธ์ปริญญาการศึกษามหาบัณฑิต มหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร.

วลี เฉลยสมัย. 2539. ***การพัฒนาวิธีการวินิจฉัยการคำนึงถึงสภาพที่เป็นจริงในการแก้โจทย์ปัญหาคณิตศาสตร์ของนักเรียนชั้นประถมศึกษาปีที่ 6.*** วิทยานิพนธ์ปริญญาครุศาสตรมหาบัณฑิต จุฬาลงกรณ์มหาวิทยาลัย. อยู่ในระหว่างการจัดพิมพ์.

วารี นิยมธรรม. 2536. ***ผลการใช้เทคนิคแม่แบบจากนิทานชาดกเพื่อพัฒนาเชิงจริยธรรมในนักเรียนชั้นอนุบาลปีที่ 2.*** วิทยานิพนธ์ปริญญาการศึกษามหาบัณฑิต มหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร.

สุดใจ ขันทองคำ. 2534. ***การพัฒนาแบบวัดความซื่อสัตย์ของนักเรียนชั้นประถมศึกษาปีที่ 6.*** วิทยานิพนธ์ปริญญาครุศาสตรมหาบัณฑิต จุฬาลงกรณ์มหาวิทยาลัย.

สุพัตรา เทียนอุดม. 2536. ***การพัฒนาแบบวัดความมีระเบียบวินัยสำหรับนักเรียนชั้นประถมศึกษาปีที่ 6.*** วิทยานิพนธ์ปริญญาครุศาสตรมหาบัณฑิต. จุฬาลงกรณ์มหาวิทยาลัย.

ภาษาอังกฤษ

De Ayla, R.J. (1990). ***A Comparision of the Partial Credit and Graded Response Models in Computerized Adaptive Testing.*** Boston : [ERIC ED 320920]

Donoghue, J.R. (1994). An Empirical Examination of the IRT Information of Polytomous Scored Reading Item Under The Generalized Partial Credit Model. ***Journal of Educational Measurement.*** 31 (4) : 295-311.

Drasgow, F. 1995. Introduction to the Polytomous IRT Special Issue. ***Applied Psychological Measurement.*** 19 (1) : 1-3.

Frary, R.B. and Huthchison, T.P. 1982. Willingness to Answer Multiple-Choice Questions as Mainfested Both in Genuine and Nonsense Item. ***Educational and Psychological Measurement.*** 42 (3) : 815-821.

_____. and Zimmerman, D.W. 1970. Effect of Variation in Probability of Guessing Correctly on Reliability of Multiple-choice Tests. ***Educational and Psychological Measurement.*** 30 : 596-605.

Muraki, E. 1983. ***Marginal Maximum Likelihood Estimation For Three-Parameter Polytomous Item Response Models : Application of an EM Algorithm.*** (DAI 44/074 : 2121).

_____. 1992. A Generalized Partial Credit Model : Application of an EM Algorithm. ***Applied Psychological Measurement.*** 16 (2) : 159-176.

_____. 1993. Information Functions of the Generalized Partial Credit Model. ***Applied Psychological Measurement.*** 17 (4) : 351-363.

_____. and Bock, R.D. 1991. A User's Guide to PARSCALE [Computer Program]. Chicago : Scientific Software, Inc.

Smith, R.M. 1987. Assessing Partial Knowledge in Vocabulary. ***Journal of Educational Measurement.*** 24 (3) : 217-231.

Thissen, D. 1976. Information in Wrong Response to Raven Progressive Matrices. ***Journal of Educational Measurement.*** 13 : 201-214.

_____. 1991. MULTILOG User's Guide. Chicago, USA. : Scientific Software, Inc.