การศึกษาบุคลิกภาพของผู้ตรวจให้คะแนนความสามารถในการเขียนเรียงความของนักเรียนชั้นประถมศึกษาปีที่ 3

The Study of Rater's Characteristics on Writing Ability by The Elementary Schooling Grade 3

บุษวรรษ์ แสนปลื้ม[1],องอาจ นัยพัฒน์[2]

Butsawan Saenpluem[1],Ong-Art Naiyapatana[2]

[1]นิสิตระดับปริญญาเอก สาขาการทดสอบและวัดผลการศึกษา มหาวิทยาลัยศรีนครินทรวิโรฒ ได้รับทุนอุดหนุน การวิจัยจากบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ [2]รองศาสตราจารย์ คณะศึกษาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ

ABSTRACT

The purposes of this study were to develop and improve the scoring rubrics on 3 types : analytic, holistic and annotated scoring rubrics in order to identify and study agreement of the rater's characteristics (severity/leniency, gender differential rater functioning (DRF) and Differential Rater Functioning Over Time (DRIFT)) on scoring the writing Ability differently. The instruments used in the study were the essay-type test with 3 types of scoring rubrics (analytic, holistic and annotated scoring rubrics). The samples of this study were selected from 50 students from the elementary schooling grade 3 and 130 students majoring Thai language subject from Faculty of Education, 3 Rajabhat Universities through simple sampling. The results of this study revealed that:The 3 types of scoring rubrics identified the reliability between .813 - .921 and rater agreement index between .820 - .833. According to the Rater Characteristic, it indicated : The severity/leniency were 1) 17 raters with the severity, 8 raters with leniency and 90 raters with centrality for the analytic scoring rubrics. 2) 10 raters with the severity, 16 raters with leniency and 83 raters with centrality for the holistic scoring rubrics. 3) 17 raters with the severity, 15 raters with leniency and 79 raters with centrality for the annotated scoring rubrics. DRF were 1) 8 raters with bias relative to boy, 14 raters with bias relative to girl and 71 raters with centrality for the analytic scoring rubrics. 2) 5 raters with bias relative to boy, 8 raters with bias relative to girl and 96 raters with centrality for the holistic scoring rubrics. 3) 10 raters with bias relative to boy, 11 raters with bias relative to girl and 90 raters with centrality for the annotated scoring rubrics. DRIFT were 1) 85 raters with inaccuracy and 32 raters with accuracy for the analytic scoring rubrics. 2) 41 raters with inaccuracy and 68 raters with accuracy for the holistic scoring rubrics. 3) 71 raters with inaccuracy and 41 raters with accuracy for the annotated scoring rubrics. For the agreement of identifying the Rater Characteristic on writing ability, it was found that: The severity/leniency possessed no difference (by using χ^2 =4.000, p<.01). DRF possessed no difference (by using χ^2 =1.032, p<.01). And DRIFT possessed difference (by using χ^2 =27.875, p>.01).

Keywords : *Essays, Rubrics, Severity/Leniency, Differential Rater Functioning, Differential Rater Functioning Over Time*

บทคัดย่อ

งานวิจัยนี้มีจุดประสงค์เพื่อสร้างและพัฒนาเกณฑ์การตรวจให้คะแนนเรียงความ 3 รูปแบบ คือ แบบแยกองค์ประกอบ แบบรวมองค์ประกอบ และแบบผสมผสาน ระบุบุคลิกภาพของผู้ตรวจให้คะแนน (ความเข้มงวด/ใจดี ความลำเอียงด้านเพศ และการทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป) และศึกษาความสอดคล้องของการระบุบุคลิกภาพของผู้ตรวจให้คะแนน เมื่อใช้เกณฑ์การตรวจให้คะแนนที่แตกต่างกัน เครื่องมือที่ใช้ในการวิจัย คือ เกณฑ์การตรวจให้คะแนน ที่ผู้วิจัยสร้างขึ้น กลุ่มตัวอย่าง คือ นักเรียนชั้นประถมศึกษาปีที่ 3 จำนวน 50 คน และนักศึกษา ครุศาสตร์ที่เรียนวิชาเอกภาษาไทย มหาวิทยาลัยราชภัฏ 3 แห่ง จำนวน 130 คน ที่ได้มาจากการสุ่มอย่างง่าย ผลการวิจัยพบว่า เกณฑ์การให้คะแนน 3 รูปแบบ ที่ผู้วิจัยสร้างขึ้น มีค่าความเชื่อมั่นอยู่ระหว่าง .813 - .921 และค่าความสอดคล้องของผู้ตรวจให้คะแนนมีค่าอยู่ระหว่าง .820 - .833 การระบุบุคลิกภาพของผู้ตรวจให้คะแนน ความเข้มงวด/ใจดี 1) แบบแยกองค์ประกอบ ผู้ตรวจให้คะแนนมีความเข้มงวด 17 คน ใจดี 8 คน และเป็นกลาง 90 คน 2) แบบรวมองค์ประกอบ เข้มงวด 10 คน ใจดี 16 คน และเป็นกลาง 83 คน 3) แบบผสมผสาน เข้มงวด 17 คน ใจดี 15 คน และเป็นกลาง 79 คน ความลำเอียงด้านเพศ 1) แบบแยกองค์ ประกอบ มีผู้ตรวจให้คะแนนลำเอียงเข้าข้างเพศชาย 8 คน เพศหญิง 14 คน และเป็นกลาง 71 คน 2) แบบรวมองค์ประกอบ เพศชาย 5 คน เพศหญิง 8 คน และเป็นกลาง 96 คน 3) แบบผสมผสาน เพศชาย 10คน เพศหญิง 11 คน และเป็นกลาง 90 คน การทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป 1) แบบแยกองค์ประกอบ มีผู้ตรวจให้คะแนนไม่คงที่ 85 คน และคงที่ 32 คน 2) แบบรวมองค์ประกอบ มีผู้ตรวจให้คะแนนไม่คงที่ 41 คน และคงที่ 68 คน 3) แบบผสมผสาน มีผู้ตรวจให้คะแนนไม่คงที่ 71 คน และคงที่ 41 คน และ ความสอดคล้องของการระบุบุคลิกภาพของผู้ให้คะแนน การระบุความเข้มงวด/ใจดีไม่แตกต่างกัน (χ^2=4.000, p<.01) การระบุความลำเอียงด้านเพศไม่แตกต่างกัน (χ^2=1.032, p<.01) และ การระบุการทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไปแตกต่างกัน (χ^2=27.875, p>.01)

คำสำคัญ: *เรียงความ, เกณฑ์การตรวจให้คะแนน, ความเข้มงวด/ใจดี, ความลำเอียงด้านเพศ, การทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป*

บทนำ

การเขียนเรียงความเป็นวิธีการขั้นพื้นฐานที่สำคัญยิ่งของการเรียนการสอนภาษาไทย เพราะเป็นวิธีการสำคัญในการวัดการอ่านออกเขียนได้ นอกจากนี้หลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน พุทธศักราช 2551 ได้กำหนดคุณภาพการเขียนเรียงความของนักเรียน ไว้ในมาตรฐานการเรียนรู้ในกลุ่มสาระการเรียนรู้ภาษาไทยสาระที่ 2 การเขียน มาตรฐาน ท 2.1 คือ ใช้กระบวนการเขียน เขียนสื่อสาร เขียนเรียงความ ย่อความ และเขียนเรื่องราวในรูปแบบต่างๆ เขียนรายงานข้อมูลสารสนเทศ และรายงานการศึกษาค้นคว้าอย่างมีประสิทธิภาพอีกด้วย (กระทรวง-ศึกษาธิการ, 2552) แต่จากค่าสถิติพื้นฐานผลการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET) พบว่า สาระและมาตรฐานการเรียนรู้ภาษาไทย มาตรฐาน ท 2.1 จากคะแนนเต็ม 12.50 คะแนน นักเรียนทั่วประเทศ มีค่าเฉลี่ย 5.11 และในปี 2552 จากคะแนนเต็ม10 คะแนน นักเรียนทั่วประเทศมีค่าเฉลี่ยเพียง 2.76 ซึ่งมีค่าเฉลี่ยไม่ถึงกึ่งหนึ่งของคะแนนเต็ม จะเห็นได้ว่าคะแนนในส่วนการเขียนนั้นมีแนวโน้มลดลง (สถาบันทดสอบทางการศึกษาแห่งชาติ, 2553) ปัญหาที่เกิดขึ้นไม่ได้เกิดจากนักเรียนขาดการฝึกหัดหรือขาดทักษะการเขียนเรียงความเพียงอย่างเดียวเท่านั้น แต่อาจเกิดจากคุณภาพของการตรวจให้คะแนน ที่ขาดมาตรฐาน ไม่มีความคงเส้นคงวาและความเชื่อถือได้

การที่ตรวจให้คะแนนขาดคุณภาพไม่ได้มาตรฐาน สาเหตุอาจมาจากความลำเอียงเฉพาะตัวของผู้ตรวจ เช่น ความเข้มงวด ความใจดี ความคล้อยตามกัน การให้คะแนนที่มีแนวโน้มเป็นกลาง หรืออาจมาจากความประทับใจส่วนตัวต่อนักเรียน นอกจากนี้สิ่งแวดล้อมขณะการตรวจให้คะแนนก็อาจมีผลต่อการตรวจให้คะแนนเช่นกัน เช่น บรรยากาศขณะตรวจให้คะแนน เกณฑ์ที่ใช้ในการตรวจให้คะแนนไม่มีความเป็นปรนัย ซึ่งวิธีการหนึ่งในแก้ปัญหาดังที่กล่าวมาคือต้องมีเกณฑ์การตรวจให้คะแนนที่ชัดเจน มีความเป็นปรนัย มีความเชื่อมั่นและความเที่ยงตรงสูง โดยส่วนใหญ่แล้วผู้เชี่ยวชาญได้แบ่งเกณฑ์การตรวจให้คะแนนออกเป็น 2 รูปแบบ คือ 1) แบบแยกองค์ประกอบ (Analytic Scoring Rubrics) เกณฑ์แบบนี้ใช้เวลาตรวจค่อนข้างนาน สร้างความเมื่อยล้าหากชิ้นงานที่ต้องตรวจมีจำนวนมาก แต่ข้อดีคือมีสารสนเทศที่ครบถ้วนส่งกลับไปยังนักเรียนเพื่อพัฒนาการเขียนให้ดียิ่งขึ้น และ 2) แบบรวมองค์ประกอบ (Holistic Scoring Rubrics) แม้ว่ารูปแบบนี้ตรวจง่าย ใช้เวลาน้อยและมีความเชื่อมั่นสูง แต่ก็ยากที่จะอธิบายเหตุผลของคะแนนที่นักเรียนได้รับ ส่วนรูปแบบที่ 3) แบบผสมผสาน (Annotated Scoring Rubrics) เป็นการตรวจให้คะแนนที่เพิ่มขึ้นมาใหม่ (Nitko, 1996) เป็นการรวมเกณฑ์การให้คะแนนแบบรวมองค์ประกอบและแบบแยกองค์ประกอบไว้ด้วยกัน เริ่มด้วยการให้คะแนนในภาพรวมด้วยเกณฑ์แบบรวมองค์ประกอบ แล้วเลือกให้คะแนนอีกเพียงบางคุณลักษณะของการให้คะแนนแบบแยกองค์ประกอบที่ผู้ตรวจคิดว่าสำคัญ เพื่อเป็นการประหยัดเวลาในการตรวจและมีสารสนเทศที่จำเป็นย้อนกลับไปให้นักเรียนเพื่อพัฒนาการเขียนเรียงความต่อไป ดังผลการวิจัยของ นิศารัตน์ คงสวัสดิ์ (2544) และปวีณา ปีอาทิตย์ (2545) ที่พบว่า เมื่อใช้รูปแบบเกณฑ์การตรวจให้คะแนนแตกต่างกัน ผู้ตรวจให้คะแนนจะให้คะแนนแตกต่างกัน

คะแนนที่ได้จากเกณฑ์การตรวจให้คะแนนดังที่กล่าวมาข้างต้นควรจะเป็นตัวแทนที่ดีของความสามารถในการเขียนเรียงความของนักเรียน ซึ่งทฤษฎีการตอบสนองข้อสอบ (Item-Response Theory: IRT) เป็นการนำใช้โมเดลตอบสนองรายข้อมาใช้ให้เป็นประโยชน์ จากการพัฒนางานวิจัยที่ผ่านมาพบว่าโมเดลการวัดที่เหมาะสมในการตรวจให้คะแนนเรียงความคือ พาเชียลเครดิตโมเดล (Partial Credit Model: PCM) และจากการศึกษาเอกสารเพิ่มเติมของผู้วิจัยในเรื่องการวิเคราะห์ตามโมเดลของราส์ชแบบหลายองค์ประกอบ (Many-Facet Rasch Measurement: MFRM) พบว่าการวิเคราะห์ส่วนใหญ่มุ่งเน้นการระบุบุคลิกภาพของผู้ตรวจให้คะแนน (Rater's Characteristics) เช่น ความเข้มงวด/ใจดี (Severity/Leniency) โดยที่ผู้ตรวจให้คะแนนที่มีแนวโน้มให้คะแนนที่ต่ำจะถูกระบุว่ามีความเข้มงวด และผู้ตรวจให้คะแนนที่มีแนวโน้มให้คะแนนที่สูงจะถูกระบุว่ามีความใจดี ซึ่งจะพิจารณาจากค่าศูนย์ (0) เครื่องหมายบวก (+) ลบ (-) ของค่า Logit ในโปรแกรม FACETS 3.70.1 ของ Linacre John M. (2012) ดังเช่นงานวิจัยของ ยูจิ นากามูระ (Nakamura Yuji, 2002) ที่ศึกษาการใช้งาน MFRM ในการวิเคราะห์การสอบการเขียน นอกจากนี้ได้มีการต่อยอดการใช้งาน MFRM โดยใช้หาความลำเอียงด้านเพศของผู้ตรวจให้คะแนน (Differential Rater Functioning: DRF) โดยใช้หลักการของการวิเคราะห์ความเข้มงวด/ใจดีของผู้ตรวจให้คะแนนข้ามกลุ่มเพศ ดังเช่นงานวิจัยของ โทมัส เอคเคส (Eckes Thomas, 2005) ที่ศึกษาการวิเคราะห์ MFRM ในการตรวจสอบผลกระทบจากผู้ตรวจให้คะแนนในการสอบ TestDaF และในปี 2009 ได้มีการนำ MFRM มาระบุคุณลักษณะผู้ตรวจให้คะแนน คือ การทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป (Differential Rater Functioning Over Time: DRIFT) โดยใช้หลักการของการวิเคราะห์ความเข้มงวด/ใจดีของผู้ตรวจให้คะแนนข้ามเวลา (Time Facet) ดังเช่นงานวิจัยของ เมฟอร์ดและวูฟ (Myford and Wolfe 2009) ที่ได้ศึกษาการเฝ้าตรวจสอบการทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป

จากปัญหาดังที่กล่าวมาข้างต้นผู้วิจัยจึงสนใจสร้างและพัฒนาเกณฑ์การตรวจให้คะแนนเรียงความ 3 รูปแบบ คือ แบบรวมองค์ประกอบ แบบแยกองค์ประกอบ และแบบผสมผสาน โดยใช้กลุ่มตัวอย่างเป็นนักเรียนชั้นประถม ศึกษาปีที่ 3 เพื่อ

เป็นการพัฒนาวิธีการวัด เพื่อพัฒนาครูผู้ตรวจให้คะแนนในอนาคต จากนั้นจึงนำเกณฑ์การตรวจให้คะแนนที่ได้ไปเป็นเครื่องมือ เพื่อศึกษาบุคลิกภาพของผู้ตรวจให้คะแนน ด้วยการระบุว่าผู้ตรวจให้คะแนนที่เป็นกลุ่มตัวอย่างมีความเข้มงวด/ใจดี ความลำเอียงด้านเพศ และการทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไปหรือไม่ และหาความสอดคล้องของการระบุบุคลิกภาพเพื่อตอบปัญหาที่ว่าเกณฑ์การตรวจให้คะแนนที่แตกต่างกันจะทำให้บุคลิกภาพของผู้ตรวจให้คะแนนแตกต่างกันหรือไม่ โดยมีสมมติฐานของการวิจัย คือ ความเข้มงวด/ใจดี ความลำเอียงด้านเพศ และการทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป มีความสอดคล้องกันเมื่อใช้เกณฑ์การตรวจให้คะแนนเรียงความแตกต่างกัน

วิธีการวิจัย

1. กลุ่มตัวอย่าง กลุ่มตัวอย่างที่ใช้ในการวิจัยมี 2 กลุ่ม คือ 1) นักเรียนชั้นประถมศึกษาปีที่ 3 ภาคการเรียนที่ 2 ปีการศึกษา 2553 ของโรงเรียนในสังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน สำนักงานเขตพื้นที่การศึกษาประถมศึกษาสุรินทร์ เขต 1 ที่ได้มาจากการสุ่มอย่างง่าย จำนวน 50 คน และ 2) นักศึกษา ครุศาสตร์ ที่เรียนวิชาเอกภาษาไทย ชั้นปีที่ 4 ภาคการเรียนที่ 1 ปีการศึกษา 2555 ในมหาวิทยาลัยราชภัฏสุรินทร์ บุรีรัมย์ และร้อยเอ็ด ที่ได้มาจากการสุ่มอย่างง่าย จำนวน 130 คน

2. เครื่องมือที่ใช้ในการวิจัย คือ 1) กระดาษคำตอบในส่วนของการเขียน จากข้อสอบการอ่านออกเสียง การเขียน และการคิดคำนวณ ชั้นประถมศึกษาปีที่ 3 ในโครงการประเมินคุณภาพการศึกษาขั้นพื้นฐาน เพื่อประกันคุณภาพผู้เรียน ปีการศึกษา 2553 จำนวน 50 ชุด 2) เกณฑ์การตรวจให้คะแนนการเขียนเรียงความ และคู่มือการใช้เกณฑ์การตรวจให้คะแนนเรียงความที่ผู้วิจัยสร้างขึ้น มีลักษณะเป็นมาตรประมาณค่า 5 ระดับ โดยที่ 1) แบบแยกองค์ประกอบ มีหัวข้อย่อยที่ต้องให้คะแนน 4 หัวข้อ คือ ประเด็นสำคัญ รูปแบบการเขียน การนำเสนอ และการใช้ภาษา 2) แบบรวมองค์ประกอบ มีการให้คะแนนเพียง 1 หัวข้อแบบรวมองค์ประกอบ และ 3) แบบผสมผสาน มีหัวข้อย่อยที่ต้องให้คะแนน 3 หัวข้อ คือ การให้คะแนนแบบรวมองค์ประกอบ รูปแบบการเขียน และการใช้ภาษา

3. วิธีการดำเนินการทดลอง หลังจากคัดเลือกกระดาษคำตอบการเขียนเรียงความของนักเรียน จำนวน 50 ชุดแล้ว ผู้วิจัยได้ดำเนินการดังต่อไปนี้

3.1 สร้างและพัฒนาเกณฑ์การตรวจให้คะแนนเรียงความ 3 รูปแบบ คือ การตรวจให้คะแนนแบบรวมองค์ประกอบ แบบแยกองค์ประกอบ และแบบผสมผสาน จากนั้นนำเกณฑ์ทั้ง 3 รูปแบบมาหาความเที่ยงตรงเชิงประจักษ์โดยผู้เชี่ยวชาญด้านการวัดผลการศึกษา ผู้เชี่ยวชาญด้านภาษาไทย และอาจารย์โรงเรียนสาธิต จำนวน 5 ท่าน และหลังจากปรับปรุงตามคำแนะนำของผู้เชี่ยวชาญแล้ว ผู้วิจัยได้สร้างคู่มือเพื่อใช้ในการอบรมผู้ตรวจให้คะแนน แล้วนำไปประเมินระดับความเหมาะสมโดยผู้เชี่ยวชาญด้านการวัดผลการศึกษา ผู้เชี่ยวชาญด้านหลักสูตรและการสอน ผู้เชี่ยวชาญด้านภาษาไทย และครูชั้นประถมศึกษา จำนวน 5 ท่าน

3.2 นำกระดาษคำตอบของนักเรียนที่คัดเลือกไว้แล้ว 50 ชุด และคู่มือการใช้เกณฑ์การ ตรวจให้คะแนนที่ปรับปรุงแล้วจากข้อ 1 มาทดลองใช้กับนักศึกษา วิชาเอกภาษาไทย ชั้นปีที่ 4 มหาวิทยาลัยราชภัฏเลย จำนวน 24 คน เพื่อหาคุณภาพของเกณฑ์การตรวจให้คะแนน ในวันที่ 27 กุมภาพันธ์ 2555 หลังจากอบรมการใช้เกณฑ์การตรวจให้คะแนน ผู้วิจัยแบ่งนักศึกษาออกเป็น 3 กลุ่ม ด้วยวิธีการสุ่มอย่างง่าย แล้วดำเนินการตามรูปแบบการทดลองแบบหมุนเวียนสมดุล (Counterbalanced Designs) ซึ่งเป็นเทคนิคหนึ่งสำหรับความเท่าเทียมกันของกลุ่มทดลอง ในเทคนิคนี้นักศึกษาแต่ละกลุ่มจะต้องให้คะแนนนักเรียน 50 คนโดยใช้เกณฑ์ทั้ง 3 รูปแบบ คือ การตรวจให้คะแนนแบบรวม

องค์ประกอบ แบบแยกองค์ประกอบ และแบบผสมผสาน แต่จะแตกต่างกันในลำดับที่ที่ได้รับมอบหมายให้ตรวจ (องอาจ นัยพัฒน์, 2548)

3.3 วิเคราะห์คุณภาพของเกณฑ์การให้คะแนนทั้ง 3 รูปแบบ โดยหาค่าความเชื่อมั่น และความสอดคล้องของผู้ตรวจให้คะแนน (Rater Agreement Index: RAI)

3.4 จัดอบรมและทดลองการตรวจให้คะแนนเรียงความแก่นักศึกษาที่เป็นกลุ่มตัวอย่างที่มหาวิทยาลัย ราชภัฏสุรินทร์ บุรีรัมย์ และร้อยเอ็ด ระหว่างวันที่ 8 มิถุนายน 2555 - 17 สิงหาคม 2555

3.5 วิเคราะห์หาค่าพารามิเตอร์โดยใช้โมเดลPCM จากนั้นระบุบุคลิกภาพของผู้ตรวจให้คะแนนเป็นรายบุคคล โดยใช้ โปรแกรม FACETS 3.70.1 แล้ววิเคราะห์ความสอดคล้องของการระบุบุคลิกภาพของผู้ตรวจให้คะแนน เมื่อใช้เกณฑ์การตรวจให้คะแนนแตกต่างกัน

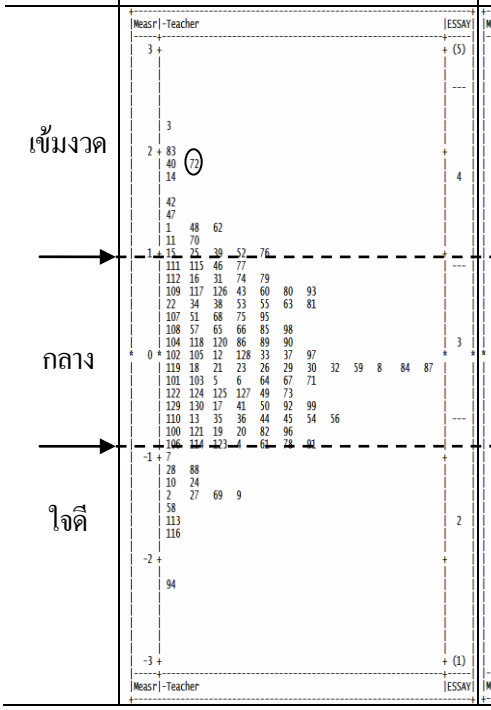
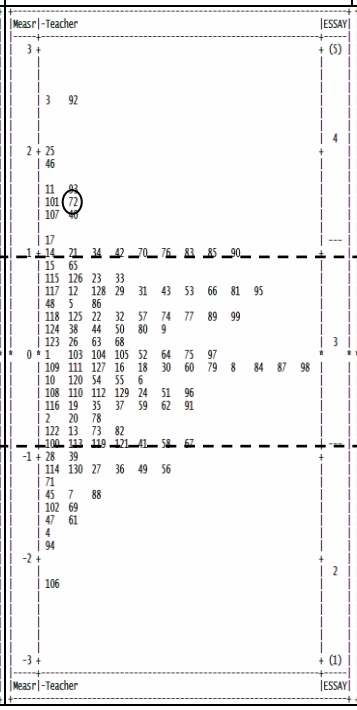
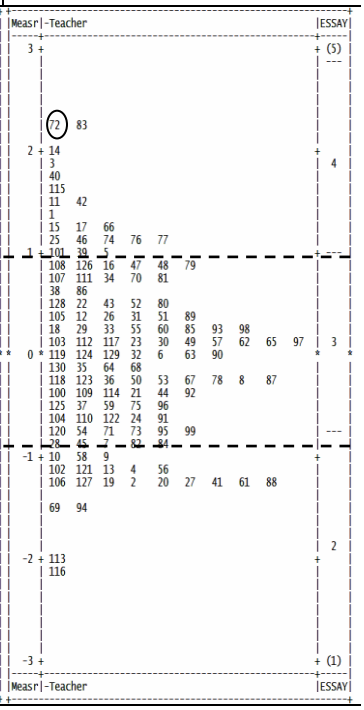
ผลของการวิจัย

1. ผลจากการสร้างและพัฒนาเกณฑ์การตรวจให้คะแนนเรียงความ 3 รูปแบบ พบว่า 1) แบบแยกองค์ประกอบ มีดัชนีความสอดคล้องของผู้ตรวจให้คะแนน เท่ากับ .820 และค่าความเชื่อมั่นเท่ากับ .813 2) แบบรวมองค์ประกอบ มีดัชนีความสอดคล้องของผู้ตรวจให้คะแนน เท่ากับ .833 และค่าความเชื่อมั่นเท่ากับ .921 และ3) แบบผสมผสาน มีดัชนีความสอดคล้องของผู้ตรวจให้คะแนน เท่ากับ .830 และค่าความเชื่อมั่นเท่ากับ .844

2. ผลการระบุบุคลิกภาพของผู้ตรวจให้คะแนน

2.1 การระบุความเข้มงวด/ใจดี ในการการวิจัยครั้งนี้ผู้วิจัยคำนวณแยกตามรูปแบบเกณฑ์การตรวจให้คะแนน โดยในขั้นแรกนั้นต้องพิจารณาว่าผู้ตรวจให้คะแนนแต่ละคนให้คะแนนเชิงประจักษ์เหมาะสมกลมกลืนกับโมเดล PCM หรือไม่ โดยพิจารณาค่า Weighted mean square residual: INFIT และ Unweighted mean square residual: OUTFIT ที่กำหนดขอบเขตของการยอมรับได้อยู่ระหว่าง 0.5 ถึง 1.5 จากนั้น จึงระบุความเข้มงวด/ใจดี จากค่าศูนย์ (0) เครื่องหมายบวก (+) ลบ (-) ของค่า Logit ผู้ตรวจให้คะแนนที่เข้มงวดจะมีค่า Logit มากกว่า +1.00 ให้คะแนน เป็นกลาง -1< Logit <+1 และให้คะแนนใจดีจะมีค่า Logit น้อยกว่า -1.00

ตารางที่ 1 แสดงเส้นภาพแสดงความเข้มงวด/ใจดีของผู้ตรวจให้คะแนน จากโปรแกรม FACETS

เกณฑ์	1) แบบแยกองค์ประกอบ	2) แบบรวมองค์ประกอบ	3) แบบผสมผสาน
เข้มงวด กลาง ใจดี			

หมายเหตุ เส้นประ - - - แสดงเส้นแบ่งการระบุความเข้มงวด,ความเป็นกลาง และความใจดี

วงกลม ◯ แสดงตัวอย่างผู้ตรวจให้คะแนนที่มีความเข้มงวดสอดคล้องกันทั้ง 3 รูปแบบ

จากตาราง 1 พบว่า จากผู้ตรวจ 130 คน 1) แบบแยกองค์ประกอบ มีผู้ตรวจให้คะแนนเชิงประจักษ์เหมาะสมกลมกลืนกับโมเดล 115 คน เป็นผู้ตรวจให้คะแนนที่เข้มงวด 17 คน เป็นกลาง 90 คน และใจดี 8 คน 2) แบบรวมองค์ประกอบ มีผู้ตรวจให้คะแนนเชิงประจักษ์เหมาะสมกลมกลืนกับโมเดล 109 คน เป็นผู้ตรวจให้คะแนนที่เข้มงวด 10 คน เป็นกลาง 83 คน และใจดี 16 คน 3) แบบผสมผสาน มีผู้ตรวจให้คะแนนเชิงประจักษ์เหมาะสมกลมกลืนกับโมเดล 111คน เป็นผู้ตรวจให้คะแนนที่เข้มงวด 17 คน เป็นกลาง 79 คน และใจดี 15 คน

2.2 การระบุความลำเอียงด้านเพศของผู้ตรวจให้คะแนน ในขั้นแรกนั้นต้องพิจารณาว่าผู้ตรวจให้คะแนน แต่ละคนให้คะแนนเชิงประจักษ์เหมาะสมกลมกลืนกับโมเดล PCM หรือไม่ ต่อมาจึงระบุความลำเอียงด้านเพศของผู้ตรวจให้คะแนนจากค่า Z ที่มีนัยสำคัญทางสถิติ สุดท้ายพิจารณาว่าผู้ตรวจให้คะแนน ลำเอียงเข้าข้างเพศชายหรือเพศหญิง โดยที่คะแนนที่มีแนวโน้มใจดี (Logit ต่ำกว่า) กับเพศใด แสดงว่าลำเอียงเข้าข้างเพศนั้น

จากการวิเคราะห์พบว่า จากผู้ตรวจ 130 คน 1) แบบแยกองค์ประกอบ มีผู้ตรวจให้คะแนนเชิงประจักษ์เหมาะสมกลมกลืนกับโมเดล 115 คน มีผู้ตรวจให้คะแนนลำเอียงเข้าข้างเพศชาย 8 คน เป็นกลาง 71 คน และเข้าข้างเพศหญิง 14 คน 2) แบบรวมองค์ประกอบ มีผู้ตรวจให้คะแนนเชิงประจักษ์เหมาะสมกลมกลืนกับโมเดล จำนวน 109 คน มีผู้ตรวจให้คะแนนลำเอียงเข้าข้างเพศชาย 5 คน เป็นกลาง 96 คน และเข้าข้างเพศหญิง 8 คน 3) แบบผสมผสาน มีผู้ตรวจให้คะแนนเชิงประจักษ์เหมาะสมกลมกลืนกับโมเดล 111 คน มีผู้ตรวจให้คะแนนลำเอียงเข้าข้างเพศชาย 10 คน เป็นกลาง 90 คน และเข้าข้างเพศหญิง 11 คน

2.3 การระบุการทำหน้าที่ต่างกันของผู้ให้คะแนนเมื่อเวลาผ่านไป ในขั้นแรกนั้นต้องพิจารณาว่าผู้ตรวจให้คะแนนแต่ละคนให้คะแนนเชิงประจักษ์เหมาะสมกลมกลืนกับโมเดล PCM หรือไม่ ต่อมาจึงระบุการทำหน้าที่ต่างกันของผู้ให้คะแนนเมื่อเวลาผ่านไป จากค่า Z-Score โดยที่ค่า Z-Score ที่มีนัยสำคัญทางสถิติแสดงว่าผู้ตรวจให้คะแนนให้คะแนนไม่คงที่ ถ้า Z-Score ไม่มีนัยสำคัญทางสถิติแสดงว่าผู้ตรวจให้คะแนนให้คะแนนคงที่

จากการวิเคราะห์ พบว่า จากผู้ตรวจ 130 คน 1) แบบแยกองค์ประกอบ มีผู้ตรวจให้คะแนนเชิงประจักษ์เหมาะสมกลมกลืนกับโมเดล 117 คน มีผู้ตรวจให้คะแนนไม่คงที่ 85 คน ให้คะแนนคงที่ 32 คน 2) แบบรวมองค์ประกอบ มีผู้ตรวจให้คะแนนเชิงประจักษ์เหมาะสมกลมกลืนกับโมเดล 109 คน มีผู้ตรวจให้คะแนนไม่คงที่ 41 คน ให้คะแนนคงที่ 68 คน 3) แบบผสมผสาน มีผู้ตรวจให้คะแนนเชิงประจักษ์เหมาะสมกลมกลืนกับโมเดล 112 คน มีผู้ตรวจให้คะแนนไม่คงที่ 71 คน ให้คะแนนคงที่ 41 คน

3. ผลการศึกษาความสอดคล้องของการระบุบุคลิกภาพของผู้ตรวจให้คะแนน เมื่อใช้เกณฑ์การตรวจให้คะแนนเรียงความที่แตกต่างกัน ขั้นแรกพิจารณาว่าบุคลิกภาพของผู้ให้คะแนนเมื่อใช้เกณฑ์การตรวจให้คะแนนต่างกัน ผู้ตรวจให้คะแนนคนใดมีบุคลิกภาพตรงกันบ้าง จากนั้นการทดสอบความแตกต่างด้วยการใช้ χ^2 จาก Friedman Test ถ้าพบว่าค่า χ^2 แตกต่างกันอย่างมีนัยสำคัญทางสถิติ แสดงว่าการระบุบุคลิกภาพมีความแตกต่างกัน

3.1 การระบุความเข้มงวด/ใจดี เมื่อใช้เกณฑ์การตรวจให้คะแนนเรียงความแตกต่างกัน มีผู้ตรวจให้คะแนนแสดงความเข้มงวดตรงกัน 5 คน เป็นกลางเหมือนกัน 54 คน และใจดีเหมือนกัน 2 คน รวมแสดงความเข้มงวด/ใจดีตรงกันทั้งหมด 61 คน คิดเป็นร้อยละ 46.92 เมื่อทดสอบความแตกต่างโดย พบว่า ระบุความเข้มงวด/ใจดีไม่แตกต่างกัน (χ^2=4.000, p<.01)

3.2 ความลำเอียงด้านเพศ เมื่อใช้เกณฑ์การตรวจให้คะแนนเรียงความแตกต่างกัน มีผู้แสดงความลำเอียงเข้าข้างเพศชายตรงกัน 1 คน ไม่ลำเอียงด้านเพศตรงกัน 57 คน และลำเอียงเข้าข้างเพศหญิงตรงกัน 3 คน รวมแสดงความลำเอียงด้านเพศตรงกันทั้งหมด 61 คน คิดเป็นร้อยละ 46.92 เมื่อทดสอบความแตกต่าง พบว่า ระบุความลำเอียงด้านเพศไม่แตกต่างกัน (χ^2=1.032, p<.01)

3.3 การทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป เมื่อใช้เกณฑ์การตรวจให้คะแนนเรียงความแตกต่างกัน มีผู้ให้คะแนนไม่คงที่ตรงกัน 24 คน และคงที่ตรงกัน 8 คน รวมแสดงการทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไปตรงกันทั้งหมด 32 คน คิดเป็นร้อยละ 24.62 เมื่อทดสอบความแตกต่าง พบว่า ระบุการทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไปแตกต่างกัน (χ^2=27.875, p<.01)

การอภิปรายผล

1. เกณฑ์การตรวจให้คะแนนทั้ง 3 รูปแบบที่ผู้วิจัยสร้างและพัฒนาขึ้นมีคุณภาพที่เหมาะสมสำหรับการนำไปใช้ ซึ่งเกณฑ์การตรวจให้คะแนนที่ผู้วิจัยสร้างขึ้นสอดคล้องกับงานวิจัยของ อัญญารัตน์ เจริญพฤฒินาถ (2546) ที่ได้พัฒนาแบบประเมินทักษะการอ่าน คิดวิเคราะห์ เขียน ของนักเรียนชั้นประถมศึกษา โดยได้จัดการสนทนากลุ่ม กลุ่มครู นักเรียน ผู้ปกครอง เพื่อสร้างตัวบ่งชี้และเกณฑ์การให้คะแนนการประเมินทักษะการเขียน ได้ตัวบ่งชี้ 4 ตัว คือ เนื้อเรื่อง ลำดับเรื่อง ไวยากรณ์ กลไกการเขียน แต่เนื่องจากเกณฑ์ ที่ผู้วิจัยสร้างขึ้นไม่ได้สร้างมาจากการสนทนากลุ่มแต่มาจากการผสมผสานกันระหว่างเกณฑ์ที่มาจากงานวิจัยและเกณฑ์ที่มาจากต่างประเทศ ดังนั้นเกณฑ์ที่ได้จึงมี 4 หัวข้อ คือ ประเด็นสำคัญ รูปแบบการเขียน การนำเสนอ และการใช้ภาษา

2. การระบุบุคลิกภาพของผู้ตรวจให้คะแนน

2.1 ความเข้มงวด/ใจดี พบว่า ผู้ตรวจให้คะแนนส่วนใหญ่มีความเป็นกลาง และผู้ตรวจให้คะแนนที่มีความเข้มงวดมีจำนวนมากกว่าใจดี ซึ่งไม่สอดคล้องกับงานวิจัยของ เอ็ดเวิร์ด วูฟ (Wolfe Edward W. 2004) ที่ได้ศึกษาการระบุผลกระทบจากผู้ตรวจให้คะแนนโดยใช้โมเดลตัวแปรแฝง ซึ่งพบว่าผู้ตรวจให้คะแนนที่มีความใจดีมีจำนวนมากกว่าผู้ตรวจที่มีความเข้มงวด (เข้มงวด 76 คนและใจดี 14 คน) อาจเป็นเพราะนักศึกษาที่เป็นกลุ่มตัวอย่างขาดประสบการณ์ในการตรวจเรียงความจึงมีความคาดหวังผลว่านักเรียนชั้นประถมศึกษาปีที่ 3 จะเขียนเรียงความได้ดีกว่านี้

2.2 ความลำเอียงด้านเพศ พบว่า ผู้ตรวจให้คะแนนส่วนใหญ่มีความเป็นกลาง และมีผู้ตรวจให้คะแนนลำเอียงเข้าข้างเพศหญิงมากกว่าลำเอียงเข้าข้างเพศชาย ซึ่งสอดคล้องกับงานวิจัยของ โทมัส เอคเคส (Eckes Thomas. 2005) ที่ได้ศึกษาการวิเคราะห์ MFRM ในการตรวจสอบผลกระทบจากผู้ตรวจให้คะแนนในการสอบ TestDaF ซึ่งพบว่าผู้ตรวจให้คะแนนส่วนใหญ่ให้คะแนนในส่วนของการพูดและการเขียนลำเอียงเข้าข้างเพศหญิง สาเหตุอาจมาจากงานเขียนของนักเรียนหญิงอ่านง่ายและมีความสะอาดเรียบร้อยมากกว่า

2.3 การทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป เมื่อใช้เกณฑ์การตรวจให้คะแนนแบบแยกองค์ประกอบ และแบบผสมผสาน พบว่ามีผู้ตรวจให้คะแนนให้คะแนนไม่คงที่เป็นจำนวนมาก ซึ่งไม่สอดคล้องกับงานวิจัยของ เมฟอร์ดและวูฟ (Myford and Wolfe 2009) ที่ได้ศึกษาการเฝ้าตรวจสอบการทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป ที่พบว่าผู้ตรวจให้คะแนนให้คะแนนไม่คงที่มีจำนวนน้อยมาก สาเหตุอาจมาจากกลุ่มตัวอย่างที่ใช้ในงานวิจัยของเมฟอร์ดและวูฟมีประสบการณ์การตรวจข้อสอบอย่างต่ำ 4 ปี จึงมีความคุ้นเคยกับเกณฑ์การตรวจให้คะแนนมากกว่า ดังนั้นเมื่อต้องตรวจให้คะแนนเรียงความจึงมีความแม่นยำและคงเส้นคงวาในการให้คะแนนมากกว่านักศึกษาปี 4 มหาวิทยาลัยราชภัฏที่มีประสบการณ์การตรวจให้คะแนนเรียงความเพียง 2 ครั้ง

3. ความสอดคล้องของการระบุบุคลิกภาพของผู้ตรวจให้คะแนน เมื่อใช้เกณฑ์การตรวจให้คะแนนเรียงความแตกต่างกัน

3.1 การระบุความเข้มงวด/ใจดี เมื่อใช้เกณฑ์การตรวจให้คะแนนเรียงความแตกต่างกัน พบว่า ระบุความเข้มงวด/ใจดีตรงกันทั้งหมด 61 คน (เข้มงวด 5 คน, ใจดี 2 คน, เป็นกลาง 54 คน) คิดเป็นร้อยละ 46.92 เมื่อทดสอบความแตกต่าง พบว่า ระบุความความเข้มงวด/ใจดีสอดคล้องกัน (χ^2=8.340, p<.01) แสดงว่าความเข้มงวด/ใจดีของผู้ตรวจให้คะแนนไม่แปรเปลี่ยนไปแม้ว่าจะเปลี่ยนรูปแบบเกณฑ์การตรวจให้คะแนน ซึ่งไม่สอดคล้องกับงานวิจัยของนิศารัตน์ คงสวัสดิ์ (2544) และปวีณา ปีอาทิตย์ (2545) ที่พบว่า เมื่อใช้รูปแบบเกณฑ์การตรวจให้คะแนนแตกต่างกัน ผู้ตรวจให้คะแนนจะให้คะแนนแตกต่างกัน สาเหตุอาจเกิดจากกลุ่มตัวอย่างมีบุคลิกภาพความเข้มงวด/ใจดีที่คงที่ ส่งผลต่อการตรวจให้คะแนนอย่างเป็นระบบ

3.2 การระบุความลำเอียงด้านเพศ เมื่อใช้เกณฑ์การตรวจให้คะแนนเรียงความแตกต่างกัน พบว่า ระบุความลำเอียงด้านเพศตรงกันทั้งหมด 61 คน (เข้าข้างเพศชาย 1 คน, เข้าข้างเพศหญิง 3 คน, ไม่ลำเอียงด้านเพศ 57 คน) คิดเป็นร้อยละ 46.92 เมื่อทดสอบความแตกต่าง พบว่า ระบุความลำเอียงด้านเพศสอดคล้องกัน (χ^2=9.783, p<.01) แสดงว่าความลำเอียงด้านเพศของผู้ตรวจให้คะแนนไม่แปรเปลี่ยนไปแม้ว่าจะเปลี่ยนรูปแบบเกณฑ์การตรวจให้คะแนน ซึ่งไม่สอดคล้องกับงานวิจัยของ พรรณี เจียมสุบุตร (2543) ที่ได้ศึกษาเปรียบเทียบความเชื่อมั่นของแบบทดสอบวัดความสามารถในการแก้โจทย์ปัญหาทางคณิตศาสตร์ที่มีจำนวนผู้ตรวจและวิธีการตรวจต่างกัน โดยพบว่าเมื่อใช้

รูปแบบเกณฑ์การตรวจให้คะแนนแตกต่างกัน ผู้ตรวจให้คะแนนจะให้คะแนนแตกต่างกัน สาเหตุอาจเกิดจากผู้ตรวจให้คะแนนมีความเข้าใจในการใช้เกณฑ์การตรวจให้คะแนนเรียงความเป็นอย่างดี ดังนั้นไม่ว่าจะใช้เกณฑ์รูปแบบใดคะแนนที่ได้จึงค่อนข้างคงที่ไม่เปลี่ยนแปลง

3.3 การทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป เมื่อใช้เกณฑ์การตรวจให้คะแนนเรียงความแตกต่างกัน พบว่า ระบุตรงกันทั้งหมด 32 คน คิดเป็นร้อยละ 24.62 และระบุไม่ตรงกันทั้งหมด 64 คน คิดเป็นร้อยละ 49.23 เมื่อทดสอบความแตกต่าง พบว่าระบุการทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไปแตกต่างกัน (χ^2=27.875, p<.01) ซึ่งไม่สอดคล้องกับงานวิจัยของเมฟอร์ดและวูฟ (Myford and Wolfe, 2009) ที่ได้ศึกษาการเฝ้าตรวจสอบการทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป ที่พบว่าจากช่วงเวลาที่เปรียบเทียบ 8 ช่วงเวลา (ช่วงเช้าและบ่าย รวม 4 วัน) มีผู้ตรวจให้คะแนนจำนวนน้อยมาก (17%) ที่พบว่าให้คะแนนไม่คงที่ สาเหตุอาจเกิดจากผู้ตรวจให้คะแนนลืมว่าเคยให้คะแนนเรียงความนักเรียนคนใดเท่าใดและไม่เข้าใจการใช้เกณฑ์การตรวจให้คะแนนเรียงความ

สรุปได้ว่าการระบุความเข้มงวด/ใจดี และความลำเอียงด้านเพศของผู้ตรวจให้คะแนนมีความสอดคล้องกันตามสมมติฐานที่ว่า การระบุความเข้มงวด/ใจดี และความลำเอียงด้านเพศของผู้ตรวจให้คะแนน เมื่อใช้เกณฑ์การตรวจให้คะแนนเรียงความแตกต่างกัน 3 รูปแบบ คือ แบบรวมองค์ประกอบ แบบแยกองค์ประกอบ และแบบผสมผสาน มีความสอดคล้องกัน แต่การทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป ไม่สอดคล้องกับสมมิฐานที่ตั้งไว้

ข้อเสนอแนะ

1. เมื่อผู้ตรวจให้คะแนนใช้รูปแบบเกณฑ์การให้คะแนนแตกต่างกัน 3 รูปแบบ ทำให้การระบุบุคลิกภาพของผู้ตรวจให้คะแนน คือ ความเข้มงวด/ใจดี และความลำเอียงด้านเพศไม่แตกต่างกัน ดังนั้นผู้ตรวจให้คะแนนใช้เกณฑ์อะไรก็ได้ให้เหมาะกับบริบทในการสอบ เช่น แบบแยกองค์ประกอบใช้เมื่อมีการเรียนการสอนเรียงความ แบบรวมองค์ประกอบใช้เมื่อสอบปลายภาค และแบบผสมผสานใช้เมื่อสอบย่อยหรือสอบกลางภาค เป็นต้น

2. ควรมีการศึกษาคุณภาพของการตรวจให้คะแนนและบุคลิกภาพของผู้ตรวจให้คะแนนโดยใช้ทฤษฎีการสรุปอ้างอิง

บรรณานุกรม

กระทรวงศึกษาธิการ. 2552. ตัวชี้วัดและสาระการเรียนรู้แกนกลาง กลุ่มสาระการเรียนรู้ภาษาไทยตามหลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน พุทธศักราช 2551. กรุงเทพฯ: โรงพิมพ์คุรุสภาลาดพร้าว.

นิศารัตน์ คงสวัสดิ์. 2544. ผลการวิเคราะห์ค่าความเชื่อมั่นของผู้ตรวจแบบทดสอบความเรียงที่มีจำนวนผู้ตรวจ และวิธีการตรวจต่างกัน. ปริญญานิพนธ์ ปริญญาการศึกษามหาบัณฑิต สาขาวิชาการวัดผลการศึกษา มหาวิทยาลัยศรีนครินทรวิโรฒ.

ปวีณา ปีอาทิตย์. 2545. การศึกษาจำนวนผู้ประเมินและจำนวนงานเขียนที่เหมาะสมเมื่อใช้เกณฑ์การให้คะแนนที่ต่างกัน. วิทยานิพนธ์ ปริญญาครุศาสตร์มหาบัณฑิต สาขาวิชาการศึกษาและการสอน จุฬาลงกรณ์มหาวิทยาลัย.

พรรณี เจียมสุบุตร. 2543. การเปรียบเทียบความเชื่อมั่นของแบบทดสอบวัดความสามารถในการแก้โจทย์ปัญหาทางคณิตศาสตร์ที่มีจำนวนผู้ตรวจและวิธีการตรวจต่างกัน. ปริญญานิพนธ์ ปริญญาการศึกษามหาบัณฑิต สาขาวิชาการวัดผลการศึกษา มหาวิทยาลัยศรีนครินทรวิโรฒ.

สถาบันทดสอบทางการศึกษาแห่งชาติ. 2553. ค่าสถิติพื้นฐานผลการทดสอบ O-NET ป.6 จำแนกรายมาตรฐานการเรียนรู้ระดับประเทศ ปีการศึกษา 2551. เข้าถึงจาก http://www.niets.or.th. (ค้นวันที่ 29 มีนาคม 2553)

Eckes Thomas. 2005. Examining Rater Effects in TestDaF Writing and Speaking Performance Assessments: A Many-Facet Rasch Analysis. Language Assessment Quarterly 2(3), 197–221.

Linacre John M. 2012. A User's Guide to FACETS Rasch-Model Computer Programs. USA: Winsteps.

Myford Carrol M. and Wolfe Edward W. 2009. Monitering Rater Performance Over Time: A Framework for Detecting Differential Accuracy and Differential Scale Category Use. Journal of Education Measurement 46 (4) : 371-389.

Nakamura Yuji. 2002. An application of a many-faceted Rasch model of writing Test analysis. Curriculum Innovation, Testing and Evaluation: Proceedings of the 1st Annual JALT Pan-SIG Conference, May 11-12, 2002. Kyoto: Kyoto Institute of Technology.

Nitko Anthony J. 1996. Educational Assessment of Students. N.J.: Merrill.

Wolfe Edward W. 2004 . Identifying rater effects using latent trait models. Psychology Science 46 (1), 35-51.