



**การศึกษาความสอดคล้องของการระบุความเข้มงวด/ ใจดี และการทำหน้าที่ต่างกัน
ของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป ของผู้ตรวจให้คะแนนความสามารถ
ในการเขียนเรียงความของนักเรียนชั้นประถมศึกษาปีที่ 3**

**The Study of Agreement of Rater's Severity/ Leniency and Differential Rater
Functioning Over Time on Essay Composition by
The Elementary Schooling Grade 3**

บุษวรรธน์ แสนปลื้ม*

Butsawan Saenplue

งามอาจ นัยพัฒน์**

Ong-art Naiyapatana

บทคัดย่อ

งานวิจัยนี้มีจุดประสงค์เพื่อสร้างและพัฒนาเกณฑ์การตรวจให้คะแนนเรียงความ 3 รูปแบบ ระบุและศึกษาความสอดคล้อง ของความเข้มงวด/ ใจดีและการทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป เครื่องมือที่ใช้ในการวิจัย คือ เกณฑ์การตรวจให้คะแนนที่ผู้วิจัยสร้างขึ้น กลุ่มตัวอย่าง คือ นักเรียนชั้นประถมศึกษาปีที่ 3 จำนวน 50 คน และนักศึกษามหาวิทยาลัยราชภัฏ จำนวน 130 คน ที่ได้มาจากการสุ่มอย่างง่าย ผลการวิจัยพบว่า 1. เกณฑ์การตรวจให้คะแนน 3 รูปแบบ ที่ผู้วิจัยสร้างขึ้น มีค่าความเชื่อมั่นอยู่ระหว่าง .813 - .921 2. การระบุความเข้มงวด/ ใจดี และการทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป พบว่า 2.1 การระบุความเข้มงวด/ ใจดี 1) แบบแยกองค์ประกอบ มีเข้มงวด 11 คน ใจดี 9 คน และเป็นกลาง 74 คน 2) แบบรวมองค์ประกอบ มีเข้มงวด 9 คน ใจดี 11 คน และเป็นกลาง 74 คน 3) แบบผสมผสาน มีเข้มงวด 15 คน ใจดี 13 คน และเป็นกลาง 66 คน 2.2 การทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป 1) แบบแยกองค์ประกอบ มีผู้ตรวจให้คะแนนไม่คงที่ 71 คน และคงที่ 25 คน 2) แบบรวมองค์ประกอบ มีผู้ตรวจให้คะแนนไม่คงที่ 37 คน และคงที่ 59 คน 3) แบบผสมผสาน มีผู้ตรวจให้คะแนนไม่คงที่ 59 คน และคงที่ 37 คน 3. ความสอดคล้องของการระบุความเข้มงวด/ ใจดี และการทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป พบว่า 3.1 การระบุความเข้มงวด/ ใจดี ไม่แตกต่างกัน ($X^2 = 4.000$, $p > .01$) 3.2 การระบุการทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป แตกต่างกัน ($X^2 = 27.875$, $p < .01$)

คำสำคัญ: เรียงความ/ เกณฑ์การตรวจให้คะแนน/ ความเข้มงวด/ใจดี/ การทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป

* นิสิตปริญญาการศึกษาศาสตรบัณฑิต สาขาวิชาการทดสอบและวัดผลการศึกษา ได้รับทุนอุดหนุนการวิจัย จากบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

** อาจารย์ที่ปรึกษา รองศาสตราจารย์ ดร. ภาควิชาการวัดผลและวิจัยทางการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ



Abstract

The purposes of this study were to create and develop the scoring rubrics on 3 types in order to identify and study agreement of the severity/leniency and Differential Rater Functioning Over Time (DRIFT) on scoring the essay composition differently. The instruments used in the study were the essay-type test with 3 types of scoring rubrics. The samples of this study were selected from 50 students from the elementary schooling grade 3 and 130 students of 3 Rajabhat Universities through simple sampling. The results of this study revealed that: 1. The 3 types of scoring rubrics identified the reliability between .813 - .921 and rater agreement index between .820 - .833. 2. According to the severity/leniency and DRIFT, it indicated : 2.1 The severity/leniency were 1) 11 raters with the severity, 9 raters with leniency and 74 raters with centrality for the analytic scoring rubrics. 2) 9 raters with the severity, 11 raters with leniency and 74 raters with centrality for the holistic scoring rubrics. 3) 15 raters with the severity, 13 raters with leniency and 66 raters with centrality for the annotated scoring rubrics. 2.2 The DRIFT were 1) 71 raters with inaccuracy and 25 raters with accuracy for the analytic scoring rubrics. 2) 37 raters with inaccuracy and 59 raters with accuracy for the holistic scoring rubrics. 3) 59 raters with inaccuracy and 37 raters with accuracy for the annotated scoring rubrics. 3. For the agreement of identifying the rater's severity/leniency and DRIFT on essay composition, it was found that: 3.1 the severity/leniency possessed no difference (by using $X^2 = 4.000$, $p > .01$) 3.2 the DRIFT possessed difference (by using $X^2 = 27.875$, $p < .01$)

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

บทนำ

การเขียนเรียงความเป็นวิธีการขั้นพื้นฐานที่สำคัญยิ่งของการเรียนการสอนภาษาไทย เพราะเป็นวิธีการสำคัญในการวัดการอ่านออกเขียนได้ นอกจากนี้หลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน พุทธศักราช 2551 ได้กำหนดคุณภาพการเขียนเรียงความของนักเรียน ไว้ในมาตรฐานการเรียนรู้ในกลุ่มสาระการเรียนรู้ภาษาไทยสาระที่ 2 การเขียน มาตรฐาน ท 2.1 คือ ใช้กระบวนการเขียน เขียนสื่อสาร เขียนเรียงความ ย่อความ และเขียนเรื่องราวในรูปแบบต่างๆ เขียนรายงานข้อมูลสารสนเทศ และรายงานการศึกษาค้นคว้าอย่างมีประสิทธิภาพอีกด้วย (กระทรวงศึกษาธิการ,

2552: 1) แต่จากค่าสถิติพื้นฐานผลการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-NET) ช่วงชั้นที่ 2 ปีการศึกษา 2551 พบว่า สาระและมาตรฐานการเรียนรู้ภาษาไทย มาตรฐาน ท 2.1 จากคะแนนเต็ม 12.50 คะแนน นักเรียนทั่วประเทศมีค่าเฉลี่ย 5.11 และในปี 2552 จากคะแนนเต็ม 10 คะแนน นักเรียนทั่วประเทศมีค่าเฉลี่ยเพียง 2.76 ซึ่งมีค่าเฉลี่ยไม่ถึงกึ่งหนึ่งของคะแนนเต็ม จะเห็นได้ว่าคะแนนในส่วนการเขียนนั้นมีแนวโน้มลดลงเรื่อยๆ (สถาบันทดสอบทางการศึกษาแห่งชาติ, 2553) ปัญหาที่เกิดขึ้นไม่ได้เกิดจากนักเรียนขาดการฝึกหัดหรือขาดทักษะการเขียนเรียงความเพียงอย่างเดียวเท่านั้น แต่อาจเกิด



จากคุณภาพของการตรวจให้คะแนน ที่ขาดมาตรฐาน ไม่มีความคงเส้นคงวาและความเชื่อถือได้

การที่ตรวจให้คะแนนขาดคุณภาพไม่ได้มาตรฐาน สาเหตุอาจมาจากความลำเอียงเฉพาะตัวของผู้ตรวจ เช่น ความเข้มงวด ใจดี ใจร้าย คล้อยตามกัน การตรวจให้คะแนนที่มีแนวโน้มเป็นกลาง หรืออาจมาจากความประทับใจส่วนตัวต่อนักเรียน นอกจากนี้ สิ่งแวดล้อมขณะการตรวจให้คะแนนก็อาจมีผลต่อการตรวจให้คะแนนเช่นกัน เช่น บรรยากาศขณะตรวจให้คะแนน เกณฑ์ที่ใช้ในการตรวจให้คะแนนไม่มีความเป็นปรนัย ซึ่งวิธีการหนึ่งในแก้ปัญหาดังที่กล่าวมาคือต้องมีเกณฑ์การตรวจให้คะแนนที่ชัดเจน มีความเป็นปรนัย มีความเชื่อมั่นและความเที่ยงตรงสูง โดยส่วนใหญ่แล้วผู้เชี่ยวชาญได้แบ่งเกณฑ์การตรวจให้คะแนนออกเป็น 2 รูปแบบ คือ 1) แบบแยกองค์ประกอบ (Analytic Scoring Rubrics) เพื่อให้การมองคุณภาพของเรียงความหรือความสามารถของนักเรียนได้อย่างชัดเจน จึงได้มีการแยกองค์ประกอบของการตรวจให้คะแนน และให้คะแนนอธิบายคุณภาพของเรียงความในแต่ละองค์ประกอบเป็นระดับต่างๆ เกณฑ์แบบนี้ใช้เวลาตรวจค่อนข้างนาน สร้างความเมื่อยล้าหากชิ้นงานที่ต้องตรวจมีจำนวนมาก แต่ข้อดีคือมีสารสนเทศที่ครบถ้วนส่งกลับไปยังนักเรียน เพื่อพัฒนาการเขียนให้ดียิ่งขึ้น และ 2) แบบรวมองค์ประกอบ (Holistic Scoring Rubrics) คือ การตรวจให้คะแนนเรียงความ โดยดูภาพรวมของชิ้นงานว่านักเรียนมีความเข้าใจในความคิดรวบยอด การสื่อความหมาย กระบวนการที่ใช้ และผลงานเป็นอย่างไร แม้ว่ารูปแบบนี้ตรวจง่าย ใช้เวลาน้อยและมีความเชื่อมั่นสูง แต่ก็ยากที่จะอธิบายเหตุผลของคะแนนที่นักเรียนได้รับ ส่วนรูปแบบที่ 3) แบบผสมผสาน (Annotated Scoring Rubrics) เป็นการตรวจให้คะแนนที่เพิ่มขึ้นมาใหม่ (Nitko, 1996: 171) เป็นการรวมเกณฑ์การตรวจให้คะแนนแบบรวมองค์ประกอบ

และแบบแยกองค์ประกอบไว้ด้วยกัน เริ่มด้วยการตรวจให้คะแนนในภาพรวมด้วยเกณฑ์แบบรวมองค์ประกอบ แล้วเลือกให้คะแนนอีกเพียงบางคุณลักษณะของการตรวจให้คะแนนแบบแยกองค์ประกอบที่ผู้ตรวจคิดว่าสำคัญ เพื่อเป็นการประหยัดเวลาในการตรวจและมีสารสนเทศที่จำเป็นย้อนกลับไปให้นักเรียนเพื่อพัฒนาการเขียนเรียงความต่อไป ดังผลการวิจัยของ นิศาตร์ณ์ คงสวัสดิ์ (2544) และ ปวีณา ปิอาทิศย์ (2545) ที่พบว่า เมื่อใช้รูปแบบเกณฑ์การตรวจให้คะแนนแตกต่างกัน ผู้ตรวจให้คะแนนจะให้คะแนนแตกต่างกัน

คะแนนที่ได้จากเกณฑ์การตรวจให้คะแนน ดังที่กล่าวมาข้างต้นควรจะเป็นตัวแทนที่ดีของความสามารถในการเขียนเรียงความของนักเรียน ซึ่งทฤษฎีการตอบสนองข้อสอบ (Item-Response Theory: IRT) เป็นการนำใช้โมเดลตอบสนองรายข้อมาใช้ให้เป็นประโยชน์ต่อการวัดผลการศึกษา จากการพัฒนางานวิจัยที่ผ่านมาพบว่าโมเดลการวัดในทฤษฎีการตอบสนองข้อสอบที่เหมาะสมในการตรวจให้คะแนนเรียงความคือ พาเชียลเครดิตโมเดล (Partial Credit Model: PCM) ที่ได้รับการพัฒนาให้ดีขึ้นโดย Wright และ Masters (1982: 38) เนื่องจากพาเชียลเครดิตโมเดล มีลักษณะคล้ายโมเดลของราล์ฟ ใช้หลักการคำนวณความน่าจะเป็นของการตอบแต่ละระดับขึ้น การตอบโดยตรงแบบขั้นตอนเดียว (Direct IRT Method) ลักษณะข้อคำถามแต่ละข้อนั้นจะถูกอธิบายด้วยค่าพารามิเตอร์ระดับความยากประจำข้อ (Item Step Difficulty, δ_j) ในทุกข้อคำถามจะมีค่าพารามิเตอร์ความชันเท่ากัน (Equal Slope) โดยอยู่บนพื้นฐานที่ว่าข้อคำถามทุกข้อจะต้องมีความสัมพันธ์ที่เท่าเทียมกันกับคุณลักษณะ θ ดังนั้นโมเดลนี้จึงเหมาะสมสำหรับการวัดความสามารถในการเขียนเรียงความที่ผู้วิจัยสร้างขึ้น นอกจากการเลือกใช้โมเดลการวัดผลให้เหมาะสมกับลักษณะของคะแนนแล้ว



จากการศึกษาเอกสารเพิ่มเติมของผู้วิจัยในเรื่องการวิเคราะห์ตามโมเดลของราส์ชแบบหลายองค์ประกอบ (Many-Facet Rasch Measurement: MFRM) พบว่าในงานวิจัยต่างประเทศการวิเคราะห์ส่วนใหญ่มุ่งเน้นการตรวจหาบุคลิกภาพความลำเอียงของผู้ตรวจให้คะแนน (Rater's Characteristics) เช่น ความเข้มงวด (Severity) ความใจดี (Leniency) ความคล้อยตามกัน (Halo Effect) การให้คะแนนที่มีแนวโน้มเป็นกลาง (Central Tendency) ความลำเอียงด้านเพศ (Gender Differential Rater Functioning: DRF) และการทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป (Differential Rater Functioning Over Time: DRIFT) ดังนั้นผู้วิจัยจึงเลือกบุคลิกภาพผู้ตรวจให้คะแนนที่สนใจมาวิเคราะห์ในบทความนี้ 2 บุคลิกภาพ คือ ความเข้มงวด/ใจดี (Severity/ Leniency) และการทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป (Differential Rater Functioning Over Time: DRIFT) ความเข้มงวด/ใจดีของผู้ตรวจให้คะแนน (Severity/ Leniency) จะหมายถึงผู้ตรวจที่ให้คะแนนสูงกว่าหรือต่ำกว่าคะแนนเฉลี่ยอย่างเป็นรูปแบบ โดยผู้ตรวจให้คะแนนที่มีแนวโน้มให้คะแนนที่ต่ำจะถูกระบุว่ามีความเข้มงวด และผู้ตรวจให้คะแนนที่มีแนวโน้มให้คะแนนที่สูงจะถูกระบุว่ามีความใจดี ซึ่งจะพิจารณาจากค่าศูนย์ (0) เครื่องหมายบวก (+) ลบ (-) ของค่า Logit ในโปรแกรม FACETS 3.70.1 ของลินาเคอร์ (Linacre, 2012) ดังเช่นงานวิจัยของนาคุมุระ (Nakamura, 2002) ที่ศึกษาการใช้งาน MFRM ในการวิเคราะห์การสอบการเขียน งานวิจัยของ วูล์ฟ (Wolfe, 2004) ที่ศึกษาการระบุผลกระทบจากผู้ตรวจให้คะแนนโดยใช้โมเดลตัวแปรแฝง นอกจากนี้ได้มีการต่อยอดการใช้งาน MFRM โดยในปี 2009 ได้มีการนำ MFRM มาระบุคุณลักษณะผู้ตรวจให้คะแนน คือ การทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป (Differential Rater Functioning

Over Time: DRIFT) โดยใช้หลักการของการวิเคราะห์ความเข้มงวด/ใจดีของผู้ตรวจให้คะแนนข้ามเวลา (Time Facet) การทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไปจะถูกระบุจากสถิติ Z ที่มีนัยสำคัญทางสถิติ ดังเช่นงานวิจัยของเมฟอร์ดและวูล์ฟ (Myford and Wolfe, 2009) ที่ได้ศึกษาการเฝ้าตรวจสอบการทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป

จากปัญหาดังที่กล่าวมาข้างต้นผู้วิจัยจึงสนใจสร้างและพัฒนาเกณฑ์การตรวจให้คะแนนเรียงความ 3 รูปแบบ คือ แบบรวมองค์ประกอบ แบบแยกองค์ประกอบ และแบบผสมผสาน โดยใช้กลุ่มตัวอย่างเป็นนักเรียนชั้นประถมศึกษาปีที่ 3 เพื่อเป็นการส่งเสริมความสามารถในการเขียนเรียงความของนักเรียนที่เรียนในระดับพื้นฐานโดยการพัฒนานิเวศการวัดเพื่อพัฒนาครูผู้ตรวจให้คะแนนในอนาคต จากนั้นจึงนำเกณฑ์การตรวจให้คะแนนที่ได้ไปเป็นเครื่องมือเพื่อศึกษาบุคลิกภาพของผู้ตรวจให้คะแนน คือ ความเข้มงวด/ใจดี และการทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป และหาความสอดคล้องของการระบุบุคลิกภาพเพื่อตอบปัญหาที่ว่าเกณฑ์การตรวจให้คะแนนที่แตกต่างกันจะทำให้บุคลิกภาพของผู้ตรวจให้คะแนนแตกต่างกันหรือไม่

จุดประสงค์ของการวิจัย

1. เพื่อสร้างและพัฒนาเกณฑ์การตรวจให้คะแนนเรียงความ 3 รูปแบบ คือ แบบรวมองค์ประกอบ แบบแยกองค์ประกอบ และแบบผสมผสาน
2. เพื่อระบุความเข้มงวด/ใจดี และการทำหน้าที่ต่างกันของผู้ให้คะแนนเมื่อเวลาผ่านไป จำแนกตามเกณฑ์การตรวจให้คะแนน
3. เพื่อศึกษาความสอดคล้องกันของการระบุความเข้มงวด/ใจดี และการทำหน้าที่ต่างกันของผู้ให้คะแนนเมื่อเวลาผ่านไป เมื่อใช้เกณฑ์การตรวจให้คะแนนแตกต่างกัน 3 รูปแบบ



สมมติฐานของการวิจัย

การระบุความเข้มงวด/ใจดี และการทำหน้าที่ต่างกันของผู้ให้คะแนนเมื่อเวลาผ่านไป มีความสอดคล้องกัน เมื่อใช้เกณฑ์การตรวจให้คะแนนเรียงความแตกต่างกัน 3 รูปแบบ

ขอบเขตการวิจัย

1. ตัวแปรที่ใช้ในการวิจัย

1.1 ตัวแปรอิสระ เกณฑ์การตรวจให้คะแนน คือ 1) แบบแยกองค์ประกอบ 2) แบบรวมองค์ประกอบ และ 3) แบบผสมผสาน

1.2 ตัวแปรตาม 1) การระบุความเข้มงวด/ใจดี และการทำหน้าที่ต่างกันของผู้ให้คะแนนเมื่อเวลาผ่านไป 2) ความสอดคล้องกันของการระบุความเข้มงวด/ใจดี และการทำหน้าที่ต่างกันของผู้ให้คะแนนเมื่อเวลาผ่านไป

2. กลุ่มตัวอย่าง กลุ่มตัวอย่างที่ใช้ในการวิจัย มี 2 กลุ่ม คือ 1) นักเรียนชั้นประถมศึกษาปีที่ 3 ภาคการเรียนที่ 2 ปีการศึกษา 2553 ของโรงเรียนในสังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน สำนักงานเขตพื้นที่การศึกษาประถมศึกษาสุรินทร์ เขต 1 ที่ได้มาจากการสุ่มอย่างง่าย จำนวน 50 คน เพื่อใช้เป็นผู้สอบในการสร้างและพัฒนาคุณภาพของเครื่องมือวัดความสามารถในการเขียนเรียงความ และ 2) นักศึกษาคณะครุศาสตร์ ที่เรียนวิชาเอกภาษาไทย ชั้นปีที่ 4 ภาคการเรียนที่ 1 ปีการศึกษา 2555 ในมหาวิทยาลัยราชภัฏสุรินทร์ บุรีรัมย์ และร้อยเอ็ด ที่ได้มาจากการสุ่มอย่างง่าย จำนวน 130 คน เพื่อใช้เป็นผู้ตรวจให้คะแนนความสามารถในการเขียนเรียงความของนักเรียนชั้นประถมศึกษาปีที่ 3

เครื่องมือที่ใช้ในการวิจัย

เครื่องมือที่ใช้ในการวิจัย คือ 1) กระดาษคำตอบในส่วนของการเขียน จากข้อสอบการอ่านออกเสียง การเขียน และการคิดคำนวณ ชั้นประถมศึกษาปีที่ 3 ในโครงการประเมินคุณภาพการศึกษาขั้นพื้นฐาน เพื่อประกันคุณภาพผู้เรียน ปีการศึกษา 2553 จำนวน 50 ชุด 2) เกณฑ์การตรวจให้คะแนนการเขียนเรียงความ 3 รูปแบบ และคู่มือการใช้เกณฑ์การตรวจให้คะแนนเรียงความที่ผู้วิจัยสร้างขึ้น โดยที่

1) เกณฑ์การตรวจให้คะแนนแบบแยกองค์ประกอบ มีหัวข้อย่อยที่ต้องให้คะแนน 4 หัวข้อ โดยแต่ละหัวข้อย่อยมีลักษณะเป็นมาตราประมาณค่า 5 ระดับ มีรายละเอียดดังนี้คือ (1) ประเด็นสำคัญของเรียงความ โดยนักเรียนจะต้องแสดงความสามารถในการนำแนวคิดของรูปภาพมาตั้งชื่อเรื่องให้น่าสนใจ มีความคิดแปลกใหม่ สื่อความหมายได้ถูกต้องชัดเจนตามประเด็นสำคัญของภาพ ใช้ถ้อยคำกะทัดรัด สอดคล้องกับประเด็นสำคัญของรูปภาพ และนำแนวคิดจากภาพและชื่อเรื่องที่ตั้งไว้มาเขียนเรื่องที่มีขอบข่ายครอบคลุม (2) รูปแบบการเขียนเรียงความ โดยนักเรียนจะต้องมีงานเขียนที่มีแบบแผนตามลักษณะการเขียนเรียงความ ประกอบด้วย คำนำ เนื้อหาสรุป และมีจำนวนบรรทัดในสัดส่วนที่เหมาะสม (3) การนำเสนอ นักเรียนจะต้องเรียบเรียงเนื้อเรื่องเป็นลำดับต่อเนื่องสัมพันธ์กันชัดเจนตลอดตั้งแต่การเปิดเรื่อง การดำเนินเรื่อง และการปิดเรื่อง เสนอความคิดในแง่มุมต่างๆ ทุกตอนของเรื่องสัมพันธ์กัน มีเนื้อเรื่องน่าสนใจชวนติดตาม มีการเน้นข้อคิดที่สำคัญให้เห็นอย่างเด่นชัด สอดแทรกแนวคิดที่เป็นประโยชน์ต่อสังคม แสดงเหตุผลสนับสนุนความคิดเห็นอย่างถูกต้อง และ (4) การใช้ภาษา นักเรียนจะต้องเขียนสะกดคำถูกต้องทุกคำ เลือกใช้ถ้อยคำได้ตรงกับความหมาย ใช้คำสุภาพ ไม่กำกวม มีความหมายชัดเจนเข้าใจง่าย ไม่ใช้ภาษาพูดในภาษาเขียน ไม่มีการเขียน



ฉีกคำหรือแยกคำ และลักษณะของชิ้นงานโดยภาพรวมมีความสะอาดเรียบร้อย ไม่มีรอยลบขีดฆ่า ลายมือเป็นระเบียบสวยงาม อ่านง่าย และมีจำนวนบรรทัดทั้งหมดของงานเขียนไม่ต่ำกว่าที่โจทย์กำหนดไว้ 10 บรรทัด

2) เกณฑ์การตรวจให้คะแนนแบบรวมองค์ประกอบ มีการตรวจให้คะแนนเพียง 1 หัวข้อ แบบรวมองค์ประกอบ

3) เกณฑ์การตรวจให้คะแนนแบบผสมผสาน เป็นการรวมเกณฑ์การตรวจให้คะแนนแบบรวมองค์ประกอบและแบบแยกองค์ประกอบไว้ด้วยกัน มีหัวข้อย่อยที่ต้องให้คะแนน 3 หัวข้อ คือ (1) การตรวจให้คะแนนแบบรวมองค์ประกอบ, (2) รูปแบบการเขียนเรียงความ และ (3) การใช้ภาษา (หัวข้อย่อยที่ (2) และ (4) จากเกณฑ์การให้คะแนนแบบแยกองค์ประกอบ)

การดำเนินการวิจัย

หลังจากคัดเลือกกระดาษคำตอบการเขียนเรียงความของนักเรียน จำนวน 50 ชุดแล้ว ผู้วิจัยได้ดำเนินการดังต่อไปนี้

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

2. นำกระดาษคำตอบของนักเรียน 50 ชุด ที่คัดเลือกไว้แล้วและคู่มือการใช้เกณฑ์การตรวจให้คะแนนที่ปรับปรุงแล้วจากข้อ 1 มาทดลองใช้กับนักศึกษา วิชาเอกภาษาไทย ชั้นปีที่ 4 มหาวิทยาลัยราชภัฏเลย จำนวน 24 คน เพื่อหาคุณภาพของเกณฑ์การตรวจให้คะแนน ในวันที่ 27 กุมภาพันธ์ 2555 หลังจากอบรมการใช้เกณฑ์การตรวจให้คะแนนเรียบร้อยแล้ว ผู้วิจัยแบ่งนักศึกษาออกเป็น 3 กลุ่ม ด้วยวิธีการสุ่มอย่างง่าย แล้วดำเนินการตามรูปแบบการทดลองแบบหมุนเวียนสมดุล (Counterbalanced Designs) ซึ่งเป็นเทคนิคหนึ่งสำหรับความเท่าเทียมกันของกลุ่มทดลอง ในเทคนิคนี้นักศึกษาแต่ละกลุ่มจะต้องใช้เกณฑ์ทั้ง 3 รูปแบบ แต่จะแตกต่างกันในลำดับที่ที่ได้รับมอบหมายให้ตรวจ (อองอาจ นัยพัฒน์, 2548)

3. วิเคราะห์คุณภาพของเกณฑ์การตรวจให้คะแนนทั้ง 3 รูปแบบ โดยหาค่าความเชื่อมั่น และความสอดคล้องของผู้ตรวจให้คะแนน (Rater Agreement Index: RAI)

4. จัดอบรมการตรวจให้คะแนนเรียงความแก่นักศึกษาที่เป็นกลุ่มตัวอย่างจากมหาวิทยาลัยราชภัฏสุรินทร์ บุรีรัมย์ และร้อยเอ็ด มหาวิทยาลัยละ 2 ครั้ง ตามวันและเวลาที่ได้นัดหมาย ระหว่างวันที่ 8 มิถุนายน 2555 - 17 สิงหาคม 2555 โดยทดลองครั้งที่ 1 เพื่อการระบุความเข้มงวด/ใจดี และทดลองครั้งที่ 2 เพื่อการระบุการทำหน้าที่ต่างกันของผู้ให้คะแนนเมื่อเวลาผ่านไป

5. วิเคราะห์หาค่าพารามิเตอร์โดยใช้ Partial Credit Model จากนั้นระบุความเข้มงวด/ใจดี และการทำหน้าที่ต่างกันของผู้ให้คะแนนเมื่อเวลาผ่านไป เป็นรายบุคคลโดยใช้ โปรแกรม FACETS 3.70.1 แล้ววิเคราะห์ความสอดคล้องของการระบุความเข้มงวด/ใจดี และการทำหน้าที่ต่างกันของผู้ให้คะแนนเมื่อเวลาผ่านไป



ผลของการวิจัย

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

2. การระบุความเข้มงวด/ใจดี ในขั้นแรกนั้น ต้องพิจารณาว่าผู้ตรวจให้คะแนนแต่ละคนให้คะแนน

เชิงประจักษ์เหมาะสมกลมกลืนกับโมเดล Partial Credit Model หรือไม่ โดยพิจารณาค่า Weighted mean square residual: INFIT และ Unweighted mean square residual: OUTFIT ที่กำหนดขอบเขตของการยอมรับได้ของโมเดลที่เหมาะสมกับข้อมูลอยู่ระหว่าง 0.5 ถึง 1.5 จากนั้นจึงระบุความเข้มงวด/ใจดี จากค่าศูนย์ (0) เครื่องหมายบวก (+) ลบ (-) ของค่า Logit ผู้ตรวจให้คะแนนที่เข้มงวดจะมีค่า Logit มากกว่า +1.00 ให้คะแนนเป็นกลาง $-1 < \text{Logit} < +1$ และให้คะแนนใจดีจะมีค่า Logit น้อยกว่า -1.00



ตาราง 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 มีผู้ตรวจให้คะแนนเชิงประจักษ์ไม่เหมาะสมกลมกลืนกับโมเดล จำนวน 34 คน คิดเป็นร้อยละ 27.69 การระบุความเข้มงวด/ ใจดี 1) แบบแยกองค์ประกอบ ผู้ตรวจให้คะแนนมีความเข้มงวด 11 คน คิดเป็นร้อยละ 8.46 ใจดี 9 คน คิดเป็นร้อยละ 6.92 และเป็นกลาง 74 คน คิดเป็นร้อยละ 56.92 2) แบบรวมองค์ประกอบ ผู้ตรวจให้คะแนนมีความเข้มงวด 9 คน คิดเป็นร้อยละ 6.92 ใจดี 11 คน คิดเป็นร้อยละ 8.46 และเป็นกลาง 74 คน คิดเป็นร้อยละ 56.92 3) แบบผสมผสาน ผู้ตรวจให้คะแนนมีความเข้มงวด 15 คน คิดเป็นร้อยละ 11.54 ใจดี 13 คน คิดเป็นร้อยละ 10.00 และเป็นกลาง 66 คน คิดเป็นร้อยละ 50.77

2.2 การระบุการทำหน้าที่ต่างกันของผู้ให้คะแนนเมื่อเวลาผ่านไป ในขั้นแรกนั้นต้องพิจารณาว่าผู้ตรวจให้คะแนนแต่ละคนให้คะแนนเชิงประจักษ์เหมาะสมกลมกลืนกับโมเดล Partial Credit Model หรือไม่โดยพิจารณาค่า INFIT และ OUTFIT ต่อมาจึงระบุการทำหน้าที่ต่างกันของผู้ให้คะแนนเมื่อเวลาผ่านไป จากค่า Z-Score โดยที่ค่า Z-Score ที่มีนัยสำคัญทางสถิติแสดงว่าผู้ตรวจให้คะแนนให้คะแนนไม่คงที่ ถ้า Z-Score ไม่มีนัยสำคัญทางสถิติแสดงว่าผู้ตรวจให้คะแนนให้คะแนนคงที่

จากการวิเคราะห์ พบว่า จากผู้ตรวจ 130 คน มีผู้ตรวจให้คะแนนเชิงประจักษ์ไม่เหมาะสมกลมกลืนกับโมเดล จำนวน 34 คน คิดเป็นร้อยละ 26.15



การระบุการทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป พบว่า 1) แบบแยกองค์ประกอบ มีผู้ตรวจให้คะแนนไม่คงที่ 71 คน คิดเป็นร้อยละ 54.62 และคงที่ 25 คน คิดเป็นร้อยละ 19.23 2) แบบรวมองค์ประกอบ มีผู้ตรวจให้คะแนนไม่คงที่ 37 คน คิดเป็นร้อยละ 28.46 และคงที่ 59 คน คิดเป็นร้อยละ 45.38 3) แบบผสมผสาน มีผู้ตรวจให้คะแนนไม่คงที่ 59 คน คิดเป็นร้อยละ 45.38 และคงที่ 37 คน คิดเป็นร้อยละ 28.46

3. ผลการศึกษาความสอดคล้องของการระบุความเข้มงวด/ใจดี และการทำหน้าที่ต่างกันของผู้ให้คะแนนเมื่อเวลาผ่านไป เมื่อใช้เกณฑ์การตรวจให้คะแนนเรียงความแตกต่างกัน 3 รูปแบบ โดยทดสอบความแตกต่างจากการใช้ X^2 จาก Friedman Test ถ้าพบว่าค่า X^2 ไม่แตกต่างกันอย่างมีนัยสำคัญทางสถิติ แสดงว่าผู้ตรวจให้คะแนนมีการระบุความเข้มงวด/ใจดี และการทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไปมีความสอดคล้องกัน

3.1 การระบุความเข้มงวด/ใจดี มีผู้แสดงความเข้มงวด ทั้ง 3 รูปแบบ จำนวน 5 คน คิดเป็นร้อยละ 3.85 แสดงความใจดี ทั้ง 3 รูปแบบ จำนวน 2 คน คิดเป็นร้อยละ 1.54 แสดงความเป็นกลาง ทั้ง 3 รูปแบบ จำนวน 54 คน คิดเป็นร้อยละ 41.54 รวมแสดงความเข้มงวด/ใจดีตรงกันทั้งหมด จำนวน 61 คน คิดเป็นร้อยละ 46.92 และแสดงความเข้มงวด/ใจดีต่างกันทั้งหมด จำนวน 33 คน คิดเป็นร้อยละ 25.38 เมื่อทดสอบความแตกต่างโดย พบว่า ระบุความเข้มงวด/ใจดีไม่แตกต่างกัน ($X^2 = 4.000$, $p > .01$)

3.2 การทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป มีผู้ให้คะแนนไม่คงที่ ทั้ง 3 รูปแบบ จำนวน 24 คน คิดเป็นร้อยละ 18.46 และคงที่ ทั้ง 3 รูปแบบ จำนวน 8 คน คิดเป็นร้อยละ

6.15 รวมแสดงการทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไปตรงกันทั้งหมด จำนวน 32 คน คิดเป็นร้อยละ 24.62 และแสดงการทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไปไม่ตรงกันทั้งหมด จำนวน 64 คน คิดเป็นร้อยละ 49.23 เมื่อทดสอบความแตกต่าง พบว่า ระบุการทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไปแตกต่างกัน ($X^2 = 27.875$, $p < .01$)

สรุปผลการวิจัยและอภิปรายผลการวิจัย

1. เกณฑ์การตรวจให้คะแนนทั้ง 3 รูปแบบ ที่ผู้วิจัยสร้างและพัฒนาขึ้นมีคุณภาพที่เหมาะสมสำหรับการนำไปใช้ ซึ่งเกณฑ์การตรวจให้คะแนนที่ผู้วิจัยสร้างขึ้นสอดคล้องกับงานวิจัยของ อัญญรัตน์ เจริญพุดนิลา (2546) ที่ได้พัฒนาแบบประเมินทักษะการอ่าน คิดวิเคราะห์ เขียน ของนักเรียนชั้นประถมศึกษา โดยได้จัดการสนทนากลุ่ม กลุ่มครู นักเรียน ผู้ปกครอง เพื่อสร้างตัวบ่งชี้และเกณฑ์การตรวจให้คะแนนการประเมินทักษะการเขียน ได้ตัวบ่งชี้ 4 ตัว คือ เนื้อเรื่อง ลำดับเรื่อง ไวยากรณ์ กลไกการเขียน แต่เนื่องจากเกณฑ์ที่ผู้วิจัยสร้างขึ้นไม่ได้สร้างมาจากการสนทนากลุ่มแต่มาจากการผสมผสานกันระหว่างเกณฑ์ที่มาจากงานวิจัยและเกณฑ์ที่มาจากต่างประเทศ ซึ่งมีจำนวนหัวข้อในการตรวจให้คะแนนน้อย แต่ครอบคลุมความสามารถในการวัดความสามารถในการเขียนของนักเรียนชั้นประถมศึกษาปีที่ 3 และสอดคล้องกับวุฒิภาวะของเด็กชั้นประถมศึกษาปีที่ 3 ดังนั้นเกณฑ์ที่ได้จากการวิจัยครั้งนี้ จึงมี 4 หัวข้อ คือ ประเด็นสำคัญ รูปแบบการเขียน การนำเสนอ และการใช้ภาษา

2. การระบุความเข้มงวด/ใจดี และการทำหน้าที่ต่างกันของผู้ให้คะแนนเมื่อเวลาผ่านไป จำแนกตามเกณฑ์การตรวจให้คะแนน พบว่า



2.1 ความเข้มงวด/ ใจดี พบว่า ผู้ตรวจให้คะแนนส่วนใหญ่มีความเป็นกลาง และผู้ตรวจให้คะแนนที่มีความเข้มงวดมีจำนวนมากกว่าใจดี ซึ่งไม่สอดคล้องกับงานวิจัยของ เอ็ดเวิร์ด วูล์ฟ (Wolfe, 2004) ที่ได้ศึกษาการระบุผลกระทบจากผู้ตรวจให้คะแนนโดยใช้โมเดลตัวแปรแฝง ซึ่งพบว่าผู้ตรวจให้คะแนนที่มีความใจดีมีจำนวนมากกว่าผู้ตรวจที่มีความเข้มงวด (เข้มงวด 76 คน และใจดี 14 คน) อาจเป็นเพราะนักศึกษาที่เป็นกลุ่มตัวอย่างขาดประสบการณ์ในการตรวจเรียงความ จึงมีความคาดหวังผลว่านักเรียนชั้นประถมศึกษาปีที่ 3 จะเขียนเรียงความได้ดีกว่านี้ เป็นผลทำให้เกิดการให้คะแนนที่มีค่าเฉลี่ยต่ำกว่าที่ควรและมีผู้ตรวจให้คะแนนที่มีคุณลักษณะเข้มงวดเป็นจำนวนมาก สอดคล้องกับงานวิจัยของ อุษณีย์ บัวศิริพันธ์ (2543) ที่ได้ศึกษาเปรียบเทียบค่าสัมประสิทธิ์การสุปอ้างอิงของแบบทดสอบวิชาคณิตศาสตร์ที่มีวิธีการตรวจ จำนวนผู้ตรวจ และประสบการณ์การตรวจแตกต่างกัน โดยกลุ่มแรกมีประสบการณ์การตรวจให้คะแนนต่ำกว่า 5 ปี เปรียบเทียบกับกลุ่มที่ 2 ที่มีประสบการณ์การตรวจสูงกว่า 5 ปี ซึ่งผลการวิจัยพบว่าผู้ตรวจให้คะแนนที่มีประสบการณ์ต่างกันส่งผลให้คะแนนมีค่าแตกต่างกัน ดังนั้นเมื่อนักศึกษามีประสบการณ์มากขึ้นอาจจะทำให้นักศึกษาในการให้คะแนนเปลี่ยนแปลงไปในทางที่ดีขึ้น

2.2 การระบุการทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป เมื่อใช้เกณฑ์การตรวจให้คะแนนแบบแยกองค์ประกอบ และแบบผสมผสาน พบว่ามีผู้ตรวจให้คะแนนให้คะแนนไม่คงที่เป็นจำนวนมาก ซึ่งไม่สอดคล้องกับงานวิจัยของ เมฟอร์ดและวูล์ฟ (Myford and Wolfe, 2009) ที่ได้ศึกษาการเฝ้าตรวจสอบการทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป ที่พบว่าผู้ตรวจให้คะแนนให้คะแนนไม่คงที่มีจำนวนน้อยมาก สาเหตุอาจมาจาก

กลุ่มตัวอย่างที่ใช้ในงานวิจัยของเมฟอร์ดและวูล์ฟ มีประสบการณ์การตรวจข้อสอบอย่างต่ำ 4 ปี ดังนั้นเมื่อต้องตรวจให้คะแนนเรียงความจำนวนมากจึงมีความแม่นยำและคงเส้นคงวาในการตรวจให้คะแนนมากกว่านักศึกษาปี 4 มหาวิทยาลัยราชภัฏที่มีประสบการณ์การตรวจให้คะแนนเรียงความเพียง 2 ครั้ง ประกอบกับการทิ้งระยะการเก็บข้อมูลเป็นเวลานานกว่างานวิจัยของเมฟอร์ดและวูล์ฟ ซึ่งเป็นการเก็บข้อมูลโดยใช้องค์ประกอบของเวลา 8 ช่วงภายใน 4 วัน ดังนั้นโอกาสที่ผู้ตรวจให้คะแนนที่เป็นนักศึกษาราชภัฏที่เป็นกลุ่มตัวอย่างในการวิจัยครั้งนี้จะให้คะแนนไม่คงที่ข้ามเวลาจึงมีสูงกว่า

3. ความสอดคล้องของการระบุความเข้มงวด/ใจดี และการทำหน้าที่ต่างกันของผู้ให้คะแนนเมื่อเวลาผ่านไป เมื่อใช้เกณฑ์การตรวจให้คะแนนเรียงความแตกต่างกัน 3 รูปแบบ

3.1 การระบุความเข้มงวด/ ใจดี เมื่อใช้เกณฑ์การตรวจให้คะแนนเรียงความแตกต่างกัน 3 รูปแบบ พบว่า ระบุความเข้มงวด/ ใจดีตรงกันทั้งหมด จำนวน 61 คน (เข้มงวด 5 คน, ใจดี 2 คน, เป็นกลาง 54 คน) คิดเป็นร้อยละ 46.92 และระบุต่างกัน จำนวน 33 คน คิดเป็นร้อยละ 25.38 เมื่อทดสอบความแตกต่าง พบว่า ระบุความความเข้มงวด/ ใจดีสอดคล้องกัน ($X^2 = 4.000$, $p > .01$) แสดงว่าความเข้มงวด/ใจดีของผู้ตรวจให้คะแนนไม่แปรเปลี่ยนไปแม้ว่าจะเปลี่ยนรูปแบบเกณฑ์การตรวจให้คะแนน ซึ่งไม่สอดคล้องกับงานวิจัยของนิคาร์ตัน คงสวัสดิ์ (2544) และปวีณา ปิอาทิตย์ (2545) ที่พบว่า เมื่อใช้รูปแบบเกณฑ์การตรวจให้คะแนนแตกต่างกัน ผู้ตรวจให้คะแนนจะให้คะแนนแตกต่างกัน สาเหตุอาจเกิดจากกลุ่มตัวอย่างมีบุคลิกภาพความเข้มงวด/ ใจดีที่คงที่ ส่งผลต่อการตรวจให้คะแนนอย่างเป็นระบบ



3.2 การทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป เมื่อใช้เกณฑ์การตรวจให้คะแนนเรียงความแตกต่างกัน 3 รูปแบบ พบว่า ระบุตรงกันทั้งหมด จำนวน 32 คน คิดเป็นร้อยละ 24.62 และระบุไม่ตรงกันทั้งหมด จำนวน 64 คน คิดเป็นร้อยละ 49.23 เมื่อทดสอบความแตกต่าง พบว่า ระบุการทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไปแตกต่างกัน ($X^2 = 27.875, p < .01$) ซึ่งไม่สอดคล้องกับงานวิจัยของเมฟอร์ดและวูล์ฟ (Myford and Wolfe, 2009) ที่ได้ศึกษาการเฝ้าตรวจสอบการทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไป ที่พบว่าจากช่วงเวลาเปรียบเทียบ 8 ช่วงเวลา (ช่วงเช้าและบ่าย รวม 4 วัน) มีผู้ตรวจให้คะแนนจำนวนน้อยมาก (17%) ที่พบว่าให้คะแนนไม่คงที่ สาเหตุอาจเกิดจากการทิ้งระยะในการเก็บข้อมูลทีนานเกินไป จนทำให้ผู้ตรวจให้คะแนนลืมนึกว่าเคยให้คะแนนเรียงความนักเรียนคนใดเท่าใด และผู้ตรวจไม่เข้าใจการใช้เกณฑ์การตรวจให้คะแนนเรียงความ

สรุปได้ว่าการระบุความเข้มงวด/ใจดี และการทำหน้าที่ต่างกันของผู้ให้คะแนนเมื่อเวลาผ่านไป ไม่สอดคล้องกันตามสมมติฐานที่ว่า การระบุความเข้มงวด/ใจดี และการทำหน้าที่ต่างกันของผู้ให้คะแนนเมื่อเวลาผ่านไป เมื่อใช้เกณฑ์การตรวจให้คะแนนเรียงความแตกต่างกัน 3 รูปแบบ คือ แบบรวมองค์ประกอบ แบบแยกองค์ประกอบ และแบบผสมผสาน มีความสอดคล้องกัน

ข้อเสนอแนะ

1. เกณฑ์การตรวจให้คะแนนและคู่มือการตรวจให้คะแนนเรียงความที่ผู้วิจัยสร้างขึ้น มีความเชื่อมั่นและความเที่ยงตรงสูงสามารถนำไปประยุกต์ใช้กับการสอบการเขียนเรียงความชั้นประถมศึกษาปีที่ 3 โดยเฉพาะโรงเรียนในสังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน สำนักงานเขต

พื้นที่การศึกษาประถมศึกษาสุรินทร์ เขต 1 ถ้านำไปใช้ในบริบทอื่นหรือกลุ่มตัวอย่างอื่นอาจต้องนำไปทดลองใช้ใหม่เนื่องจากวิธีการให้คะแนนนี้สร้างขึ้นให้เหมาะสมกับกลุ่มตัวอย่างที่เป็นนักเรียนชนบท และควรอ่านและทำความเข้าใจคู่มือการตรวจให้คะแนนอย่างละเอียด โดยเฉพาะขั้นตอนในการให้คะแนนในแต่ละวิธีการตรวจให้คะแนน เนื่องจากผู้ตรวจให้คะแนนต้องทำตามลำดับขั้นตอนและใช้วิจารณญาณในการพิจารณาคะแนนในแต่ละประเด็นอย่างเคร่งครัด คู่มือฉบับนี้จึงจะมีประโยชน์สูงสุด

2. เมื่อผู้ตรวจให้คะแนนใช้รูปแบบเกณฑ์การตรวจให้คะแนนแตกต่างกัน 3 รูปแบบ ทำให้การระบุบุคลิกภาพของผู้ตรวจให้คะแนน คือ ความเข้มงวด/ใจดี สอดคล้องกัน ดังนั้นผู้ตรวจให้คะแนนใช้เกณฑ์อะไรก็ได้ให้เหมาะกับบริบทในการสอบ เช่น แบบแยกองค์ประกอบใช้เมื่อมีการเรียนการสอนเรียงความแบบรวมองค์ประกอบใช้เมื่อสอบปลายภาค และแบบผสมผสานใช้เมื่อสอบย่อยหรือสอบกลางภาค เป็นต้น

3. ควรมีการศึกษาคุณภาพของการตรวจให้คะแนนและบุคลิกภาพของผู้ตรวจให้คะแนนโดยใช้ทฤษฎีการสรุปอ้างอิง

4. ควรมีการทำวิจัยเชิงคุณภาพถึงภูมิหลังที่ทำให้ผู้ตรวจให้คะแนนมีคุณลักษณะที่เข้มงวดหรือใจดี เช่น ผลสัมฤทธิ์ทางการเรียนของผู้ตรวจให้คะแนน ประสบการณ์ในการสอน การสังกัดในโรงเรียนที่มีขนาดต่างกัน

5. ควรมีการศึกษาการทำหน้าที่ต่างกันของผู้ตรวจให้คะแนนเมื่อเวลาผ่านไปโดยใช้องค์ประกอบของเวลา (Time Facet) ในรูปแบบอื่นๆ เช่น แบ่งออกเป็น ช่วงเช้า-ช่วงบ่าย เพื่อศึกษาว่าเมื่อองค์ประกอบของเวลาเปลี่ยน ผลการระบุคุณลักษณะการให้คะแนนไม่คงที่ข้ามเวลาจะเปลี่ยนแปลงหรือไม่



6. ควรมีการศึกษาเพื่อพัฒนาการ จะทำให้คุณลักษณะบุคลิกภาพการตรวจให้คะแนน (Longitudinal Study) ของการตรวจให้คะแนน ผันแปรไปหรือไม่ อย่างไร ของผู้ตรวจให้คะแนนแต่ละคนว่าคะแนนที่ได้นั้น

เอกสารอ้างอิง

- กระทรวงศึกษาธิการ. (2552). **ตัวชี้วัดและสาระการเรียนรู้แกนกลาง กลุ่มสาระการเรียนรู้ภาษาไทย ตามหลักสูตรแกนกลางการศึกษาขั้นพื้นฐาน พุทธศักราช 2551**. กรุงเทพมหานคร: โรงพิมพ์คุรุสภา.
- นิศารัตน์ คงสวัสดิ์. (2544). ผลการวิเคราะห์ค่าความเชื่อมั่นของผู้ตรวจแบบทดสอบความเรียงที่มีจำนวนผู้ตรวจ และวิธีการตรวจต่างกัน. วิทยานิพนธ์ปริญญาการศึกษามหาบัณฑิต สาขาวิชาการวัดผลการศึกษา มหาวิทยาลัยศรีนครินทรวิโรฒ.
- ปวีณา ปิอาทิตย์. (2545). การศึกษาจำนวนผู้ประเมินและจำนวนงานเขียนที่เหมาะสมเมื่อใช้เกณฑ์การตรวจให้คะแนนที่ต่างกัน. วิทยานิพนธ์ปริญญาครุศาสตรมหาบัณฑิต สาขาวิชาการศึกษาและการสอน จุฬาลงกรณ์มหาวิทยาลัย.
- พรรณี เจียมสุบุตร. (2543). การเปรียบเทียบความเชื่อมั่นของแบบทดสอบวัดความสามารถในการแก้โจทย์ปัญหาทางคณิตศาสตร์ที่มีจำนวนผู้ตรวจและวิธีการตรวจต่างกัน. วิทยานิพนธ์ ปริญญาการศึกษามหาบัณฑิต สาขาวิชาการวัดผลการศึกษา มหาวิทยาลัยศรีนครินทรวิโรฒ.
- สถาบันทดสอบทางการศึกษาแห่งชาติ. (2553). **ค่าสถิติพื้นฐานผลการทดสอบ O-NET ป.6 จำแนกรายมาตรฐานการเรียนรู้ ระดับประเทศ ปีการศึกษา 2551**. [ออนไลน์]. สืบค้นเมื่อ 29 มีนาคม 2553. จาก <http://www.niets.or.th>,
- อัญญารัตน์ เจริญพฤตนิมิต. (2546). การพัฒนาแบบประเมินทักษะการอ่าน คติวิเคราะห์ เขียน ของนักเรียนชั้นประถมศึกษา. วิทยานิพนธ์ปริญญาครุศาสตรมหาบัณฑิต สาขาวิชาการศึกษาและการสอน จุฬาลงกรณ์มหาวิทยาลัย.
- อุษณีย์ บัวศิริพันธุ์. (2543). การเปรียบเทียบค่าสัมประสิทธิ์การสรุปอ้างอิงของแบบทดสอบวิชาคณิตศาสตร์ที่มีวิธีการตรวจจำนวนผู้ตรวจและประสบการณ์ของผู้ตรวจแตกต่างกัน. วิทยานิพนธ์ปริญญาการศึกษามหาบัณฑิต สาขาวิชาการวัดผลการศึกษา มหาวิทยาลัยศรีนครินทรวิโรฒ.
- งามอาจ นัยพัฒน์. (2548). **วิธีวิทยาการวิจัยเชิงปริมาณและเชิงคุณภาพทางพฤติกรรมศาสตร์และสังคมศาสตร์**. กรุงเทพมหานคร: ศูนย์หนังสือจุฬาลงกรณ์มหาวิทยาลัย.
- 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. Japan : Kyoto Institute of Technology.
- Nitko, Anthony J. (1996). **Educational assessment of students.** USA:Merrill.
- Wolfe, Edward W. (2004). **Identifying rater effects using latent trait models.** Psychology Science 46 (1): 35-51.
- Wright, Benjamin D. and Master, Geoffrey N. (1982). **Rating Scale Analysis.** USA : MESA Press.