A Study of the Relationship between Rater Experience and Rater Bias in a Mock National English Ability Test in Korea

Sara Choo
(Chonbuk National University)
Myo-young Park
(Chonbuk National University)

ABSTRACT

This study investigated differences between experienced and inexperienced raters in terms of their rating performance with a mock National English Ability Test (NEAT). Five raters participated in this study assessing23 student writing samples. The relationship of four domains-examinee, rater, task type, and scoring categories- were analyzed by Multi-faceted Rasch Analysis. The results revealed: (1) All the raters showed intra-rater reliability regardless of their previous rating experience. (2) The experienced professional raters exhibited less bias toward types of tasks than the novice raters did. (3) The novice raters showeddistinct bias patterns in certain categories when rating. These findingsshow the implication of rater training and indicate further study for efficient methods of rater training.

Key words: writing assessment, rater experience, rater bias, FACETS

I.Introduction

This year the Korean government launched the National English Ability Test (NEAT), a test that has been in development for years. With the release of NEAT, interest in the test is escalating. Even though the decision whether or not the test will completely replace English as a subject on the Korean Scholastic Aptitude Test (KSAT) has not been finalized, the effects of NEAT on English education in Korea will be grand regardless of the final decision.

One of the main distinguishing features of NEAT is that it includes both writing and speaking sections separate from reading and listening sections, but having equal importance. Of course, “the best way to test people’s writing ability is get them to write” (Hughes, 2003, p.83).This assertion can be true for speaking as well. However, becausethe stakes of NEATare high, concerns regarding its speaking and writing componentsrun deep. One of the main concerns is if the grading results can be trusted. Since performance assessment has hardly been evaluated and graded on high stake standardized tests in Korea before, this issue is worth closely examining.

II. Literature Review

1.Rater Expertise and Rater Bias

An extensive body of research has been accumulated which examines the various rater characteristics which might affect their rating ability. One category of rater characteristic israter experience or expertise. With the issue of rater training or development, rater experience or expertise has been one center of

attention to a number of researchers (Cummings, 1990; Weigle, 1998; Wolfe, Kao, and Ranney, 1998; Huot, 1993; Barkaoui, 2010; Lim, 2011). According to Lim (2011), experience refers to the length of time and amount of rating a rater has been involved in rating while expertise is synonymous with a good rating quality such as consistent marking performance. (Two of the raters as the subjects of the current study were professional raters and their expertise will also be examined with other issues involved in the study.)

Wolfe, Kao, and Ranney (1998) categorize raters into three groups depending on their expertise: competent, intermediate, and proficient. They further specify the characteristics of each rater group. In their findings, proficient raters make more general comments on writing samples. They also distribute their attention equally to all criteria of the given rubric.

In another study, Weigle (1998) evaluated the differences in rater severity and consistency between experienced and inexperienced rater groups using FACETS analysis. What she found out was that the inexperienced group initially tended to be more severe and less consistent in their rating than the other group. However, after training, the gap between the two groups was reduced.

From the rater-criteria point of view, Kondo-Brown (2002) investigated how judgments of trained teacher raters were biased towards certain types of criteria in assessing Japanese second language writing. The results of the study were that the raters scored certain criteria more leniently or more harshly, and that every rater's bias pattern was different.

Tom Lumley (2002) attempted to determine the cause for rater bias in his study using Think-aloud method. According to him, raters are highly influenced by their intuitive impression of the writing sample even ifthey refer closely to a given rubric or a scale.

II. Research Questions

In this study, the following questions are investigated:

1. Are there differences between novice raters and experienced raters in terms of rating results? If so, what are the differences?
2. Are there rater biases toward task types or scoring categories? If so, in what areas do the rater biases appear?

III. Methods

1.Participants

23 Korean middle school students and five Korean raters participated in this study. The students are in the 1st to 3rd grade in the various schools in Jeonju, Korea. They share Korean as their first language, and theirEnglish proficiency level is advanced compared to their grade peers.

Offive raters who participated in the current study, two of the raters were professional raterswhile three of the raters were first time raters. The novice raters are English teachers who have varyinglengths of teaching experience. They also reside in Jeonju, Korea.

2. Instrument

The test used in this research was a set of the mock NEAT (MNEAT) developed and hosted by Enter-test, one of the major NEAT prep test developers in Korea. The test's organization strictly followed the standards of the Korea Institution of Curriculum and Evaluation (KICE) for NEAT question items, thus the questions types are believed to be identical to the ones in the actual NEAT. The test consists of four sections: reading, listening, speaking, and writing. For the writing section, test takers are given four types of question: 1. writing a short sentence for a given situation, 2. completing a sentence in picture which describes a picture, 3. writing a letter, and 4. describing a picture and writing a paragraph length essay related to the picture. Of these four question types, the current study investigated type 3 and 4, which are considered more challenging to most students due to the length and complexity of the tasks.Both type 3 and 4 require comprehensive writing ability from examinees', so there may be more possibility that raters practice their subjective decisions. The format of the MNEAT is was both timed and impromptu.

3.Procedure

The evaluated data were 23 sets of writing samples collected from 23 students who took the MNEAT conducted in August, 2012. The students were given 35 minutes to finish the writing sections. 10 minutes was allocated for task 3, and 15 minutes for task 4.The collected writing samples were then scored by five raters. Two of the raters were professional raters who were working for the test developers, and three of them were English teachers who had no previous experience with composition scoring. Before the scoring session, the novice raters had a brief training session which included a test orientation, a review of the task questions (see Appendix A), a discussion of the scoring rubric (see Appendix B), and a quick rating practice using a piece of sample writing.

The scoring rubric for both tasks consisted of four domains: task completion (TC), language use (LU), contents (C), and organization (O). The scores for each domain range from 1 to 5for both tasks.

IV. Analysis

The Many-facet Rasch measurement (FACETS, Linacre, 1989) was employed for analysis.

V. Results

[FIGURE 1] FACETS summary

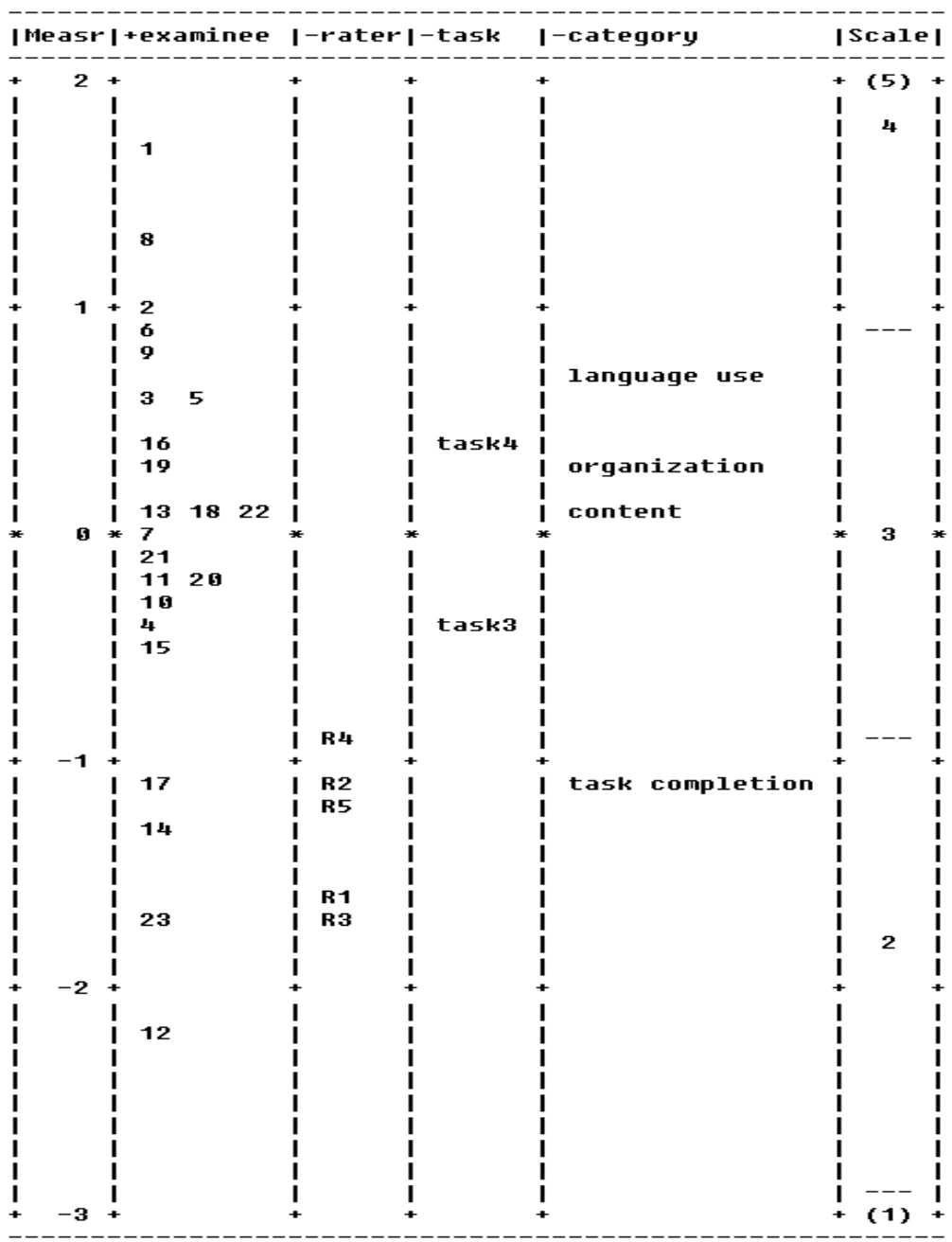


Figure 1 shows the measured examinee ability, rater severity, and the difficulty of task types and scoring categories. The column of 'examinees' indicates the 23 examinee's overall ability, higher ability examinees being positioned higher and lower ability examinees located lower. The column of 'raters' indicates the raters'severity. According to the table, five participants in this study are considered lenient raters since they are all located under the logit 0. Among the raters, rater 4 (R4) rated most severely in comparison to the other raters, and rater 3 (R3) evaluated most leniently.

In terms of 'task type' in the fourth column, the examinees indicate that task 3 (-.44 logits) was more challenging than task 4 (.44 logits). Column 5 demonstrates category difficulty variability among the scale categories. The most leniently scored category was Task Completion, and the most strictly evaluated section was Language Use.

[FIGURE 2] The probability curves

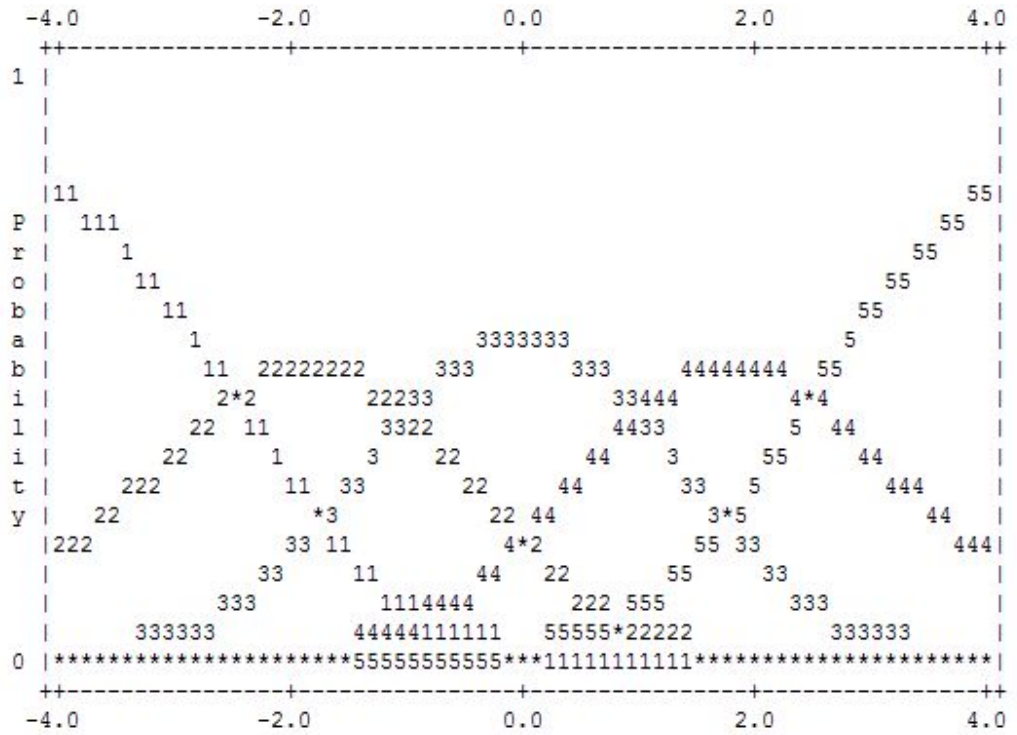


Figure 2 demonstrates that each rating scale is appropriate to be applied in assessment. The curves from the scale 1 on the left to the scale 5 on the right are independent, and each curve forms a hill-shape (Linacre, 2011). Thus, there existvalid differences between the scales.

[TABLE 1] Rater Measurement Report

Ob Aver	Fair-M Aver	Model Measure	S.E.	Infit MnSq	ZStd	Exact Agree Obs %	Exp %	Rater
3.9	3.93	-1.68	.11	1.06	.6	49.2	39.1	R3
3.8	3.88	-1.58	.11	.92	-.8	55.8	39.4	R1
3.6	3.64	-1.16	.10	1.18	1.7	49.7	39.3	R5
3.6	3.61	-1.10	.10	.93	-.6	58.2	39.2	R2
3.5	3.49	-88	.10	.96	-.3	46.3	38.1	R4
3.7	3.7	-1.28	.10	1.01	.1		Mean (count:5)	
.2	.19	.34	.00	.11	1.1		S.D.	

*RMSE (Model) .10 Adj, S. D. .32 Separation 3.11 Reliability .91
Fixed (all same) chi-square: 42.3 df: 4 significance: .00
Inter-Rater agreement opportunities:1840 Exact agreement:939=51.8% Expect agreement:718.3=39%

Table 1 provides the information about the raters' rating tendency. Measure indicates the raters' severity: the higher the number, the more severethe severity tends to be. Since the five raters are located under logit 0, showing minus (-) measure, all of them are considered lenient raters. However, there is some relativity:the most severe rater is R4 (-.88) while R3 (-1.68) is the most lenient rater.

Another important statisticis Infit, whichindicatesintra-rater reliability. The Infit of all the five raters is situated between .75 to 1.3, providing z scores of +2 to -2 (McNamara, 1996). This indicates that the raters have intra-reliability.

[TABLE 2] Category Measurement Report

Obsvd Score	Obsvd Count	Obsevd Average	Fair-M Average	Model Measure	S.E.	Infit MnSq	ZStd	Rating Category
973	230	4.2	4.30	-1.13	.10	1.34	3.4	Task completion
833	230	3.6	3.63	.14	.09	.91	-.9	Content
818	230	3.6	3.57	.26	.09	1.01	.0	Organization
761	230	3.3	3.31	.72	.09	.85	-1.7	Language use
846.3	230.0	3.7	3.70	.00	.09	1.03	.2	Mean (count:4)
90.0	.0	.4	.42	.79	.01	.22	2.3	S.D. (Sample)

* RMSE (Model) .09 Adj, S. D. .79 Separation 8.42 Reliability .99
Fixed (all same) chi-square: 195.6 df: 3 significance: .00

Table 2 summarized the information concerning the validity of the rating categories. Measure indicates

rater severityfor each test category. The raters scored Task Completion (-1.13) most leniently and Language Use (0.72) most severely.

[TABLE 3] Rater-Category Bias Interaction Report

Obsrvd .Score	Exp Score.	Obs-Exp Average	Bias Size	Model S.E.	t	Rater	Category
152	162.2	-.22	-.41	.20	-2.05	R3	Language use
205	192.4	.27	.70	.25	2.78	R5	Task completion
160	142.0	.39	.72	.20	3.56	R4	Language use

Fixed (all=0) chi-square: 40.4 d.f.: 20 significance (probability): .00

Table 3 shows meaningful examples of rater bias patterns in the categories. In this table, t-value under -2.0 indicates that the rater rated more leniently than the model expected, and t-value over +2.0 means that the rater rated more severely. The raters who displayed biases are R3, R4, and R5, who were identified as novice raters. R3 and R4 showed bias for Language Use: R3 most leniently (-2.05) and R4 most severely (3.56). R5 rated Task Completion severely (2.78).

[TABLE 4] Rater-Task type Bias Interaction Report

Obsrvd .Score	Exp Score.	Obs-Exp Average	Bias Size	Model S.E.	t	Rater	Task
289	298.1	-.10	-.18	.14	-1.30	R4	Task 4
370	376.4	-.07	-.15	.15	-1.00	R3	Task 3
350	353.8	-.04	-.08	.15	-.56	R5	Task 3
308	309.5	-.02	-.03	.14	-.21	R2	Task 4
372	372.2	.00	-.01	.15	-.04	R1	Task 3
333	332.6	.00	.01	.14	.06	R1	Task 4
353	351.4	.02	.04	.15	.24	R2	Task 3
316	312.1	.04	.08	.14	.56	R5	Task 4
344	337.4	.07	.14	.15	.96	R3	Task 4
353	340.7	.10	.20	.15	1.35	R4	Task 3

Fixed (all=0) chi-square: 6.2 d. f.: 10 significance (probability): .80

Table 4 provides examples of rater-task type bias patterns. A negative t-value indicates that the rater rated a certain task type leniently, and the opposite is true for a positive t-value. For example, R4 rated more leniently for task type 4 than for task type 3.

Table 4 provides information of t-values of the possible 10 cases (5 raters × 2 task types). The t-values are within ±2, and Bias Sizes are under 1. Thus, the raters are not considered to be biased toward a certain task. Yet, it is noticeable that R1 and R2 are less biased in comparison with R3, R4, and R5.

VI.Conclusion

The current study investigated rater reliability and rater biases inthe writing section ofa MNEAT. From the results, we can draw severalconclusions First, the raters exhibited meaningful intra-rater reliability regardless of their previous rating experience of expertise. Thus, concerns about NEAT rating outcome,which are supposed to be rated by secondary school English teachers, should not be overemphasized.
The results also showed rater bias toward a certain type of the tasks tested. However, the professional raters showed less of this bias than novice raters. Therefore, to decrease rater bias regardingspecific task types, it will be crucial to provide ample rating opportunities for the raters.

With regard to rater-category bias interaction, all the raters displayed the tendency to rate Task Completion more leniently and Language Use more severely. However, the novice raters showed distinct bias patterns in the categories of Language Use and Task Completion. Interestingly, for the category of Language Use, one rater (R3) rated most leniently, and another (R4) most severely. This should be examined more closely to determine what factors account for this discrepancy.

English testing and, subsequently, the direction of English education in Korea are facing a big change with the advent of NEAT. With reliable test scoring ofthe performance components, the change will be willingly embraced by all.

REFERENCES

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*(1), 54-74

Cummings, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*(1), 31-51.

Hughes, A. (2008). *Testing for Language Teachers*. Cambridge: Cambridge University Press

Huot, B. A. (1993). The influence of holistic scoring procesures on reading and rating student essays. In M. M. Williamson & B. A. Huot (Eds.), Validating holistic scoring for writing assessment (pp. 206-232). Cresskill, NJ: Hampton Press.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*(1), 3-31.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experience raters. *Language Testing, 28(*4), 543-560.

Linacre, J. M. (1989). *Many-facet Rasch measurement*, Chicago: MESA Press.
Linacre, J. M. (2011). A User's Guide to FACET: Rasch-Model Computer Programs [Program Manual 3.68].http://www.winsteps.com/a/facets-manual.pdf.
Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? Language Testing, 19(3), 246-276.
McNamara, T. F. (1996). *Measuring Second Language Performance*. London: Longman.
Weigle, S. C. (1998). Using FACETS to model rater training effects. Language Testing, 15, 263-287.
Wolfe, E. W., Kao, C. W., & Ranny, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication, 15*(4), 465-492.

About the Presenters

Sara Choo received an MA in TESOL from California State University, US and is a doctorial student in English education at Chonbuk National University. She has been involved in teaching for over 15 years at the primary and secondary levels both in the US and in Korea. Her academic interests include performance assessment in English teaching and development of tools for performance assessment. Currently, she is working as a director at Avalon English in Jeonju.

Myo-young Park received an MA in English Education from Chonbuk National University. She completed the doctoral course work in English education at the same university. She has taught English at high schools for two years and at collages for two years. Her academic interests are performance assessment, specially focused on rater's reliability and the validity of the rating scale. Currently, she is teaching English at Vision Collage in Jeonju.

Appendix A
Task Questions

Task 3

Part 3

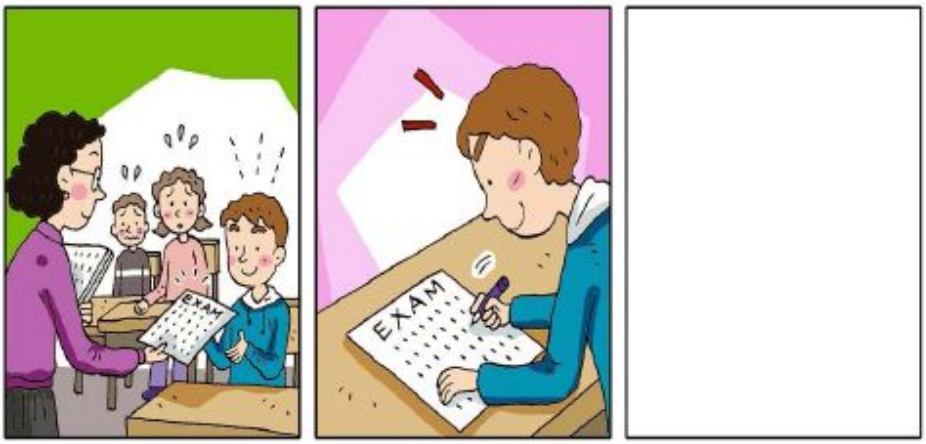
여러분이 친구의 물건을 깨뜨린 것에 대해 용서를 구하는 편지를 써야 합니다. 다음에 주어진 세 가지 정보를 포함하여 글을 쓰세요. (40~50단어)

- Write a letter to your friend apologizing for breaking something of his/hers.
- Mention what you broke.
- Explain what you will do for your friend.

Task 4

Part 4

다음 그림은 순서대로 일어난 일입니다. 첫 번째 그림과 두 번째 그림에 나타난 상황을 각각 묘사하고, 이에 따른 세 번째 그림의 내용을 추론하여 작성하세요. (40~50단어)



Appendix B
Scoring Rubric

Task 3

Task Completion	5(over 40 words writing in details including all 3 given conditions) 4(over 40 words writing in details, 1 condition is not fully satisfied) 3 (21-39 words, including 1-2 conditions) 2 (less than 20 words) 1 (Korean/No answer)
Language Use	5(proper vocab & 1-2 minor grammar mistakes) 4 (mostly proper vocab & 1-2 major and several minor grammar mistakes) 3(partly improper vocab & more than 3 major grammar mistakes) 2(many improper vocab & major grammar mistakes) 1(Korean/all wrong answer)
Content	5(highly relevantly and adequately developed with supporting details) 4(overall relevant and adequately but little insufficient or unclear) 3(relevant ideas but insufficient or unclear for many parts) 2(mostly irrelevant and inadequate) 1(Korean/No answer)
Organization	5(highly organized and coherent) 4(overall organized and coherent but little insufficient or unclear) 3(partly coherent) 2(mostly incoherent) 1(Korean/No answer)

Task 4

Task Completion	*-Testees should describe 1st picture and 2nd picture in order.* *-Testee should infer what could happen in the picture 3 and describe it.* *-Do not degrade when the WC is over 50 though the directions of the task state WC 40-50.* 5(described 3 pictures in details over WC 40) 4(described 3 in details with WC 30-39/or described 3 but not in details over WC 40) 3(described only 2/or described 3 with WC 20-29 2(described only 1 over 10WC) 1(less than 10WC)
Language Use	5(proper vocab & 1-2 minor grammar mistakes) 4(mostly proper vocab & 1 major and several minor grammatical mistakes) 3(overall proper vocab & 2-3 major and several minor grammar mistakes) 2(many improper vocab & 4 or more major grammar mistakes) 1(mostly improper vocab & many major grammar mistakes)
Content	*-Assess the relevancy of the content, not the creativity of the content.* 5(highly relevantly and logically developed with supporting details) 4(overall relevant and logical but little insufficient of unclear) 3(relevant ideas but unclear of illogical for some parts) 2(relevant ideas but unclear of illogical for some parts) 1(mostly irrelevant and illogical)

Organization	5(highly well-organized in details & followed sequential order using transitional words) 4(overall organized and sequenced but little weak at using transitional words) 3(overall organized but did not follow sequential order clearly) 2 (weak at organizing and mostly did not follow sequential order) 1(disorganized and irrelevant)