A Study on Rater-related Variables in the Evaluation of L2 Writing

Kahoko Matsumoto (Tokai University)

Tama Kumamoto (Nagoya Univ. of Foreign Studies)

(Abstract)

This study investigates how various rater-related factors influence the evaluation of writings produced by second language (L2) learners of English in Japan. This interest was generated by the previous questionnaire study done by the authors, which elucidated various rater-related differences seen in this field.

The study consists of two phases: the ①preliminary phase and ②experimental phase. In the preliminary phase (①), three raters were asked to evaluate 141 students' argumentative essays, out of which ten essays that exhibited larger differences in evaluation were selected for further inquiry. This pre-selection was done in order for us to directly tune into possibly problematic rater-related factors and evaluation categories. At the same time, we recruited twelve raters with various backgrounds for the experimental phase and checked the factors that might influence their rating through a questionnaire. In the experimental phase (②), these twelve raters were asked to evaluate the selected ten essays based on the well-accepted evaluation scale created by Jacobs, Zinkgraf, Wormuth, Hartfiel & Hughey (1981). Since the scale is made up of five categories (content, organization, vocabulary, language use and mechanics), we ended up with the correlation of 10 papers X 12 raters X 5 evaluation categories. Along with conducting the statistical analysis, we also interviewed the raters as to the points on which their evaluation showed conspicuous differences in order to seek the reasons for these differences.

The results show that rater-related variables such as nationality, the type of teacher training the raters had, their teaching experience, and their approaches and tenets as a teacher greatly influence the emphases they place in the evaluation of L2 writing. Greater differences caused by such factors were found in the categories of content, grammar and organization. The results will be presented with an eye toward establishing reliable and educational L2 writing evaluation criteria that fit and serve the needs of learners of English as a Foreign Language (EFL learners).

1 Introduction

Recently, English has been gaining importance as the most widely-used language in the world close to *Lingua Franca,* propelled by globalization and the advancement of computers and information technology. It is then quite natural that the skills of writing in English required for e-mails or other computer-mediated communication have become increasingly important (Warschauer, 2000; Warschauer & Ware 2006). This means that more emphasis should be given to teaching English writing as well as its evaluation at the time, as writing skills may function as an important gate-keeping factor leading to a successful career in this IT-driven world.

Since the late 1970's there has been a shift in the instruction of second language (L2) writing, from product-oriented to process-oriented instruction (Grabe & Kaplan, 1996). Along with this change in focus and approach, many studies have been conducted on the L2 learners' writing process and the types of effective feedback as related to the process-oriented approach (Silva & Matsuda, 2001). However, the research on evaluation is lagging because L2 writing has more complicated factors involved both with raters and learners than L1 writing. This has caused difficulty in establishing L2-proper criteria, and the existing body of research has at best inconclusive or mixed results (Hamp-Lyons, 2002; Cumming, 2002). Since evaluation is the other side of the coin to diagnostic feedback, classroom evaluation has always been done as a part of instruction. At the same time, the writing sections of gate-keeping tests such as TOEFL®, IELTS® and Michigan Test® have screened test-takers based on their own established criteria. Thus, we have to make an effort not only to increase the validity and reliability of classroom evaluation commensurate to such outside evaluation used as a qualification test, but also to revalidate the evaluation criteria of outside tests in order to find a fit between the two types of evaluation, which can be educational and beneficial to L2 learners (Crusan, 2002; Zak & Weiver, 1998).

With the increasing responsibility or accountability required for L2 English writing evaluation and its raters, we took up one of the less-researched areas of rater variables in this study and conducted an experiment on how rater-related variables affect evaluation.

2 Literature Review

As L2 English writing instruction has been directed toward a process-oriented approach, research on evaluation has been accumulated, which are roughly categorized into the following five areas.

①research on rating process
②research on rating scale/categories
③research on rater background
④research on rater training
⑤research on computer rating

First, research on rating process (①) and rating scale/categories (②) are comparatively common. The former mostly deals with the comparison between the holistic scale and analytic scale with advantages and problems of each method of evaluation (Bacha, 2001; Hamp-Lyons, 1995). Some studies explore the possibilities for alternative assessment or evaluation criteria that better reflect instructional purposes and learner needs (Turner & Upshur, 2002; Cumming, 2001). The latter, research on rating scale/categories, further looks into the validity or characteristics of different subcategories and the inter-rater reliability related to different subcategories (McNamara, 1996; Astika, 1993). Some studies have tried to find the components of writing skills which were most

influential in determining evaluation, where vocabulary and its variety turned out to be the biggest factor affecting the final score (Laufer, 1994; Laufer & Nation, 1995). On the other hand, Sasaki and Hirose's (1999) study using Japanese university students showed that a widely-accepted analytic scale such as that of Jacobs, et al. (1981) is not so reliable because the final score given to a writing depends on the arbitrary weighing of each category.

So long as rating is done by humans, it is inevitably influenced by ~~their~~ personal beliefs or preferences as practitioners. Thus, research on rater background (③) and rater training (④) accumulated so far has inconclusive results. Santos (1988) found that non-native raters are stricter than native-speaker raters, and Brown's study (1991) showed no significant differences between L1 teachers and ESL teachers in their evaluations of the same L2 writings. It also seems that inter-rater reliability differs in different rating subcategories (Schoonen, Vergeer & Eiting, 1997). In Japan, two studies exist with contradictory results; Kanatani & Takanashi (1978) found that American and Japanese raters focused on different aspects (content and form, respectively) in rating, while a JACET study (1989) showed no significant difference between native-speaker and non-native speaker raters. Related to these studies on rater backgrounds, the importance and effect of rater training have begun to be recognized (Shohamy, Gordon and Kraemer, 1992; Weigle, 1994); studies have shown that training has the effect to erase the differences or biases caused by differences in backgrounds and teaching beliefs of raters.

There is now an emerging area of research related to the effect and possibilities of computer rating, and the high correlations between human rating and computer rating were generally reported (Li, 2000; Pennington, 2003). It is a very promising area in that automated computer evaluation may reduce the formulaic part of teacher evaluation so that writing teachers can direct their energy towards more creative aspects of teaching effectively.

3 Purpose and Nature of Study

Considering the development and accumulation of research on writing evaluation, there seems to be a lack of qualitative inquiry: the detailed validation of what different raters focus on and how they actually evaluate each rating point. Thus, we thought of a study combining the areas where past research were not sufficient, namely a study on rater-related variables and rater training incorporating some qualitative validation, hoping to contribute to the field of evaluation research. Since the study framework is complicated, we decided to do a small-scale inquiry first for the future larger-scale study based on the insights gleaned from this research.

Our research questions are as follows.

(1) Do raters' backgrounds and beliefs/tenets affect their evaluation of EFL writing?
(2) If so, how does it manifest (in which evaluative areas and how)?
(3) Is rater training effective?

4 Methodology

This study has two phases: the preliminary phase in which we sought some statistical inter-rater reliability data based on Japanese college students' writing and the experimental phase which inquires into the influence of rater variables on rating and training effect both quantitatively and qualitatively.

4.1 Preliminary Phase

4.1.1 Participants

The subjects in this preliminary phase of our study were 141 students who took a TOEFL®-TWE mock test at one of the authors' university. The students' backgrounds were varied: they came from both undergraduate and graduate levels; their majors included both humanities/social sciences and natural sciences. The test was used to place students who had applied for the special in-house certification program that was created to support students with high motivation.

4.1.2 Design

The students took a forty-minute essay writing test, the topics of which were taken from the TOEFL®-TWE online site. 1) The tests were administered five times over the period of two and a half years beginning July 2003. The essays were rated on a TWE-type scale (from level 1 to 6 with 0.5 intervals) by two raters who had taken in-house training sessions. (When there were discrepancies between the two raters' evaluations, a third rater came in and made the final decision.) The correlation between the two raters' evaluations in each of the five administrations was calculated.

4.2 Experimental Phase

4.2.1 Participants

The subjects of this experimental phase were twenty-two teachers of English: eleven native speaker teachers (NST) and eleven Japanese teachers (JT). Among the eleven NSTs, seven teachers had a TESOL background (NST-TESOL), and four had the backgrounds other than TESOL, such as philosophy, literature and history (NST-NonTESOL). The NSTs' nationalities included American, Australian, British, Canadian, Indian and New Zealander. JTs were also divided into two groups: one with a TESOL background (JT-TESOL) and the other with backgrounds other than TESOL (JT-NonTESOL). All subjects had varied length of teaching experience from 4 to 30 years. To see the effect of training, the evaluation results of six native-speaker raters who had received training were added.

4.2.2 Design

Ten compositions whose evaluations had larger two-rater discrepancies compared to others were selected from the preliminary phase. The discrepancies of the two-rater evaluations of these essays were larger than 1. The teachers selected as raters in this study were asked to evaluate the ten

essays using an analytic scale slightly modified from that of Jacobs, et al. (1981). The scale included the following five components that are considered important for written communication: Content, Organization, Vocabulary, Language Use, and Mechanics. This time, the scores in each component were measured on a five-point scale: 1 (poor), 2 (fair), 3 (average), 4 (good), and 5 (excellent) (see Appendix). The results obtained from different rater groups were averaged to see whether any difference in the scores existed among the groups. Raters were also interviewed by the authors to explain the judgment behind the scores they gave.

5 Results and Discussion

5.1 Quantitative Results

The result of the preliminary phase, i.e., all the five administrations, is shown in Table 1. The overall rater correlation was 0.68. The agreement among raters generally increased presumably due to rater training. Overall correlation of 0.68 seems lower than those on ETS-issued reports probably because 0.5 intervals were used.

Table 1

Correlation

Administration	1st	2nd	3rd	4th	5th	Total
Number of Students	42	55	56	45	43	141
Correlation	0.58	0.69	0.72	0.68	0.74	0.68

The results of the experimental phase are shown in Table 2 and Figures 1 through 3. The rater variables, such as nationalities and majoring fields other than non-native teachers vs. NST and with or without TESOL background, did not have significant effects on the results. Table 2 shows the average scores of all of the four groups plus the trained group. Note that the mean score of the trained group was highest in all the five evaluative categories (3.70), followed by JT-NonTESOL (3.33), NST-TESOL (3.27), NST-NonTESOL (2.94), and JT-TESOL (2.66). Each group showed a similar tendency of rating content highest (3.97 for the trained group and 3.18 for the mean of the four groups), followed by organization (3.87 and 3.11), and vocabulary (3.62 and 3.04, respectively). The last two evaluative areas were rated lowest by the trained group (3.52 for both language use and mechanics), while language use was rated lowest (2.88) following mechanics (2.94) by all the four groups combined. Figure 1 demonstrates this general trend. This is contrary to Santos' finding (1988), in which university professors in general rated "content" lower than "language" when lexical mistakes were serious. This may well be because of his rather vague impressionistic definitions of the categories, "content" and "language". In his study, "content" is defined as "holistic impression, development, and sophistication" while "language" is defined as "comprehensibility, acceptability,

and irritation." Also, his participants included many science professors, who rated much more severely than professors in humanities. This might have caused the difference from the results of this study as our subjects were mostly with humanities backgrounds.

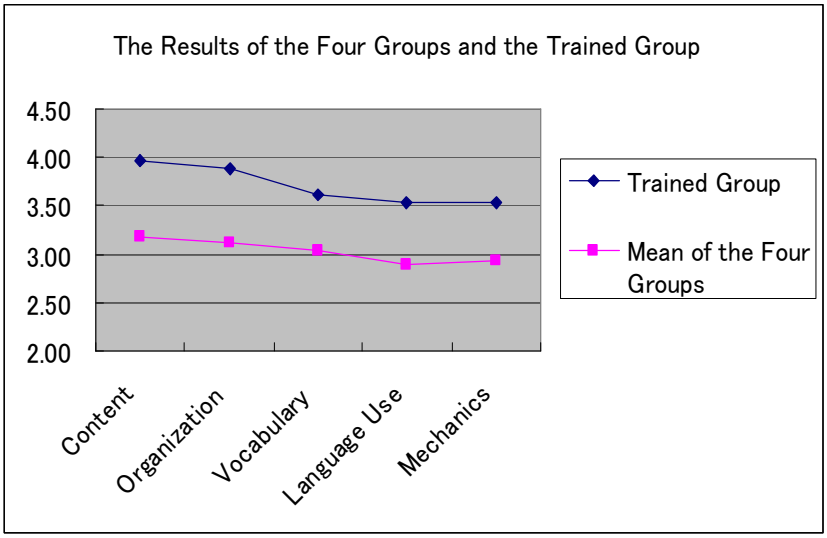
Table 2

Results of the four groups and the trained group

	Content	Organization	Vocabulary	Language Use	Mechanics	Mean of Four groups
NST-TESOL	3.39	3.40	3.30	3.14	3.14	3.27
JT-TESOL	2.87	2.61	2.67	2.53	2.61	2.66
NST-NonTESOL	3.03	3.15	3.08	2.68	2.78	2.94
JT-NonTESOL	3.50	3.43	3.18	3.25	3.30	3.33
Mean of Five Evaluative Areas	3.18	3.11	3.04	2.88	2.94	3.03
Trained	3.97	3.87	3.62	3.52	3.52	3.70

Figure 1

Results of the Four Groups and the Trained Group



Actually, there are differences among the four groups' rating behaviors. Figures 2 and 3 below demonstrate how two groups, NST (both TESOL and NonTESOL) and JT (both TESOL and NonTESOL) respectively, evaluated the students' writing in each of the five areas.

Figure 2

NST

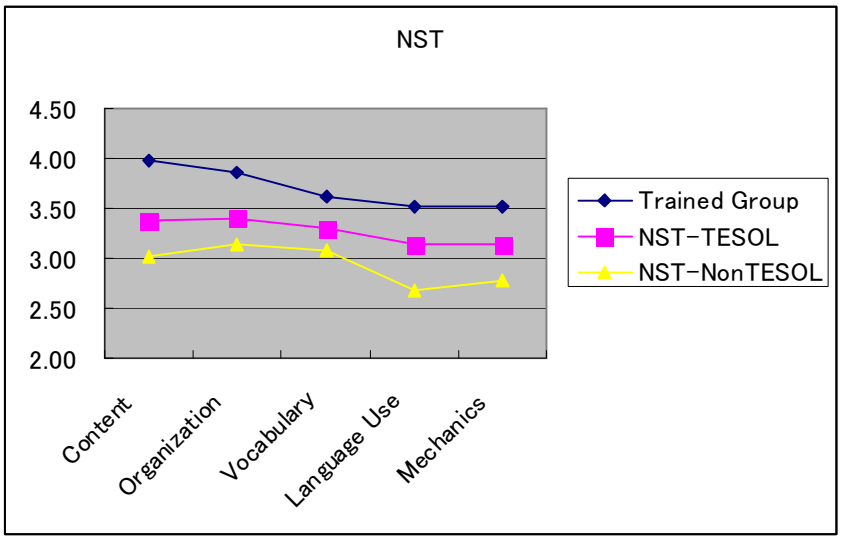
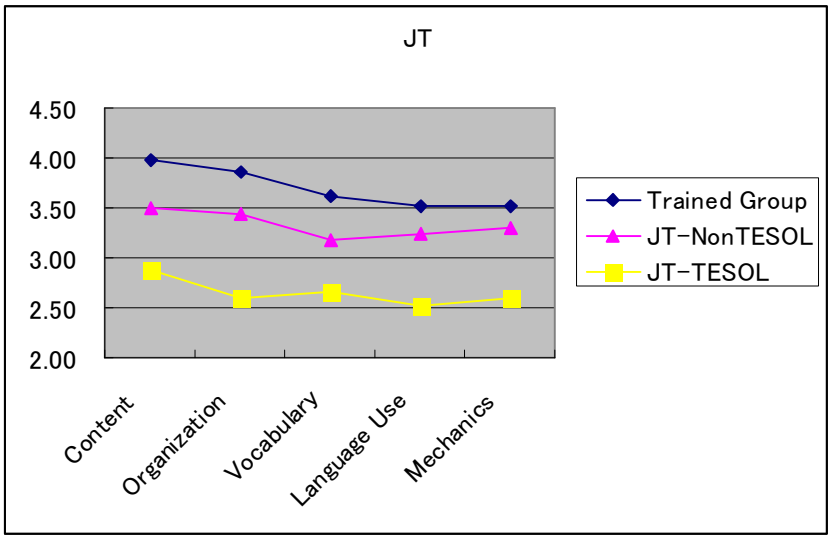


Figure 3

JT



By comparing the two groups, NST and JT, a few things can be pointed out: 1) the scores given by TESOL teachers were quite the opposite in these groups; among the NSTs, TESOL teachers' scores were higher than those of NonTESOL teachers, but in contrast, Japanese TESOL teachers scored much lower than their NonTESOL counterpart; 2) in either case, the trained group's scores were always higher than groups without training; 3) the JT-TESOL group was generally severe, particularly in organization and language use; and 4) the NST-NonTESOL group was strict on language use.

5.2 Qualitative Results

The group that gave the lowest scores in all the five categories was the JT-TESOL group (Figure 1). The teachers of this group are nonnative speakers of English with a TESOL degree; they have achieved the level of mastery through rigorous training, which could be the reason for their stricter scoring. Some of the remarks in this group indicate that the teachers looked at the mechanical aspects more severely than content and organization.

JT6: As nonnative speakers ourselves, we know how nonnative speakers learn grammar. For us it is clear how much grammar a student mastered from the writing sample, unlike native speaker teachers who tend to look only communicativeness. For example, one student sample uses such expressions as "Another example is that…," "It has been argued that…," and "This fact also indicate…," which my students can never write, so I gave high scores for vocabulary, language use and mechanics.

JT7: When there is no indentation, a line between paragraphs, or transitions, it gets difficult to read. Also, students should be careful not to soil the test paper when correcting.

The resulting severe evaluation agrees with Santos' study (1988), in which nonnative teachers' ratings were lower than native teachers'.

The NST-TESOL group, or native speaker teachers with a TESOL background, showed more lenient attitudes toward student writing samples; their scores were the closest to those of the trained group among the four groups. This is the group that seems to be most aware of the existing widely accepted criteria and accustomed to such standardization. In other words, they are more tuned to TOEFL-type of grading standards.

NST1: Content is the most important factor to me in deciding the quality of a writing, thus I'm drawn to a creative, self-expressive writing like (one sample) despite its poor grammar and language use.

NST3: It's quite natural that those teachers who have gone through rater training had more consistency among them. It was hard for me to find the benchmark for each score and I felt a bit insecure about my own evaluation.

The other Japanese subgroup, the JT-nonTESOL group also gave somewhat similar scores as the NST-TESOL group. However, the result of this group does not come from their awareness of TOEFL-type of grading standards; rather, the teachers in this group seem to have trouble in finding the anchor point on which their scoring should be based.

JT8: All these evaluation categories seem to be entangled, so I tend to be influenced by the overall impression I form.

JT11: Being objective and fair in this kind of evaluation is really hard. I think I overvalue organization while undervaluing grammar and mechanics.

The last group, NST-nonTESOL, shares the same tendency as the Japanese nonTESOL

counterpart; they are not sure about the norms, and seem to waver in their evaluation. The scores of this group were lower than the aforementioned two groups, but not as low as those of the JT-TESOL group.

NST9: I don't know how to grade the writings like (one sample) and (another sample), where high proficiency is exhibited, but I cannot follow the logic. Also, to me, content and organization are inseparable, so I had a hard time giving separate scores to them.

NST10: I think that the first holistic impression dictates when I try to give separate scores to each evaluation point, and such impression are colored by each rater's preferences and attitude toward teaching.

6 Conclusion and Implications

With the quantitative/qualitative investigations above, we conclude as follows:

(1) The teacher background (TESOL or NonTESOL, NST or JT) did make some difference in rating. The NST-TESOL group was aware of the TOEFL writing norms and their behavior was close to the trained group. Non-TESOL groups, both Japanese and native English speaker teachers, were not confident about the evaluation norms they relied on and wavered in their decisions. The JT-TESOL group had a very severe standard that might reflect their own learning history and possibly the awareness of the academic requirements they had experienced as graduate students in the West.

(2-1) Among the five categories, content and organization usually received higher scores than the rest of the categories in the ratings of four groups of teachers and the trained group.

(2-2) JT-TESOL and NST-NonTESOL groups tended to rate lower than JT-NonTESOL and NST-TESOL.

(3) Rater training was found effective.

These results lead to the following implications. Giving evaluation is a complicated process because writing itself is a combination of cognitive and communicative skills. When it is used for screening/placement or gate-keeping purposes, all of which being high-stakes occasions, rater training is absolutely necessary to render consistent results. It is necessary not only for the fair treatment of our students, but also for the raters' confidence building in their task. In our study, raters were merely shown the samples of the TOEFL evaluation norms with short explanations, but it was proven not sufficient. For consistent, accurate rating, raters should go through the training sessions like those trained raters used in this study, where they can discuss with other raters over their own evaluation results in order to establish a shared standard within an institution.

Also, the teacher difference based on their backgrounds and beliefs/tenets may give positive as

Contact: Kahoko Matsumoto (mkahoko@za.cyberhome.ne.jp)

well as negative effects on their daily diagnostic or instructional practice, and further study is necessary in what ways these variables affect their teaching and evaluation.

References

Astika, G. (1993). Analytical assessments of foreign students writing. *RELC Journal,* 24(1), 61-72.

Bacha, N. (2001). Writing evaluation: what can analytic versus holistic essay scoring tell us? *System*, 29, 371-383.

Cumming, A. (2001). ESL/EFL instructors' practices for writing assessment: specific purposes or general purposes? *Language Testing*, 18, 207-224.

Cumming, A. (2002). Assessing L2 writing: Alternative constructs and ethical dilemmas. *Assessing Writing*, 8, 73-83.

Crusan, D. (2002). An assessment of ESL writing placement assessment. *Assessing Writing*, 8, 17-31.

Grabe, W. and Kaplan, B. (eds.) (1996). Theory and Practice of writing. New York: Addison Wesley.

Hamp-Lyons, L. (1995). Rating nonnative writing: the trouble with holistic scoring. *TESOL Quarterly,* 29, 759-762.

Hamp-Lyons, L. (2002). The scope of writing assessment. *Assessing Writing*, 8, 5-16.

JACET Study Group on the Evaluation of EFL/ESL Writing. (1989). A study of the measurement of EFL writing…Can we specify an analytic scoring item which shows high correlation with impressionistic scoring. *JACET Bulletin,* 20, 17-36.

Jakobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., and Hughey, J. B. (1981). *Testing ESL Composition: A Practical Approach.* Rowley, Massachusetts: Newbury House Publishers.

Kanatani, K., and Takanashi, Y. (1978). 誤りの評価に関する語彙研究―日本人と米国人の評価の相違についてー」中部地区英語教育学会紀要、第８号, 81-89.

Laufer, B. (1994). The lexical profile of second language writing: Does it change over time? *RELC Journal,* 25(2), 21-33.

Laufer, B., and Nation, I. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics,* 16(3), 307-322.

Li, Y. (2000). Assessing second language writing: The relationship between computerized analysis and rater evaluation. *ITL: review of applied linguistics*, 127/128, 37-51.

McNamara, T. (1996). Measuring Second Language Performance. London/New York: Longman.

Pennington, M. C. (2003). The impact of the computer in second-language writing. In Matsuda, P. K., Cox, M., Jordan, J., & Ortmeier-Hooper, C. (Eds.). *Second-Language Writing in the Composition Classroom: A Critical Sourcebook.* 2006. NCTE. Originally appeared in Barbara

Kroll (Ed.), *Exploring the Dynamics of Second Language Writing.* UK: Cambridge Univ. Press, 2003. 283-310.

Santos, T. (1988). Professors' reactions to the academic writing of nonnative-speaking students. *TESOL Quarterly*, 22 (1), 69-90.

Schoonen, R., Vergeer, M. and Eiting, M. (1997). The assessment of writing ability: Expert readers versus lay readers. *Language Testing,* 14(2), 157-184.

Shohamy, E., Gordon, C. and Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal,* 76(1), 27-33.

Silva, T and Matsuda, P. K. (eds.) (2001). On Second Language Writing. New Jersey: Lawrence Erlbaum Associates.

Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly*, 36, 49-70.

Warschauer, M. & Ware, P. (2006). Automated writing evaluation: defining the classroom research agenda. *Language Teaching Research* 10(2), 1-24.

Warschauer, M. (2000). The changing global economy and the future of English teaching. *TESOL Quarterly,* 34(3), 511-535.

Weigle, S. (1994). Effects of training on raters of ESL composition. *Language Testing,* 11(2), 197-223.

Zak, F. and Weiver, C. C. (1998). The theory and practice of grading writing. State University of New York Press.

Appendix

Evaluation Form

Criteria	Excellent	Good	Average	Fair	Poor
Content: knowledgeable, substantive, thorough development of thesis, relevant to assigned topic					
Organization: fluent expression, ideas clearly stated/ supported, succinct, well-organized, logical sequencing, cohesive					
Vocabulary: sophisticated range, effective word/ idiom choice and usage, word form mastery, appropriate register					
Language Use: effective complex constructions, few errors of agreement, tense, number, word order/function, articles, pronouns, prepositions					
Mechanics: demonstrates mastery of conventions, few errors of spelling, punctuation, capitalization, paragraphing					