THE JOURNAL OF ASIA TEFL Vol. 5, No. 4, pp. 49-84, Winter 2008

Investigating EFL Writing Assessment in a Classroom Setting: Features of Composition and Rater Behaviors

Dongwan Cho
Pohang University of Science and Technology, South Korea

This study compares how English faculty members at the same college in an EFL setting rate their students' essays. Despite the diversity of their rating behaviors, some special features of EFL composition assessment emerged. First, language use attracted a very high proportion of comments (47%). Second, content was rarely commented on. Third, half of the statements on rhetorical organization were positive, whereas comments on other major components of a composition were largely negative. Last, the teacher-raters showed acceptable agreement of ratings without being given any rater training. Some pedagogical implications applied to the teaching of EFL composition are suggested.

Key words: EFL writing assessment, Classroom setting

INTRODUCTION

Literature on L2 writing assessment shows that research in this field has mostly been conducted in an ESL setting where writers are likely to have more opportunities to be exposed to a target language and to have better language proficiency than those in an EFL setting. Differences in writing proficiency of writers would lead educators to apply different criteria of the assessment of their writing. It seems that EFL writing assessment focuses on linguistic factors such as sentence structure, grammar, vocabulary and mechanics due to lack of English proficiency. In contrast, ESL writing assessment appears to be more concerned with rhetorical features such as overall organization and paragraph and sentence cohesion because language related errors would be not frequently observed in ESL writing. This general observation, however, has not been confirmed yet.

Another issue to discuss related to L2 writing assessment is that much research has been conducted in large scale proficiency or placement test settings, mainly for purely research purposes. Large scale L2 proficiency writing assessment, however, would fail to deliver the characteristics of writing assessment taking place in a classroom setting. Revealing and confirming peculiar features of EFL writing assessment carried out in a classroom setting is pedagogically significant because they will help direct the teaching of EFL writing. Considering the relationship between assessment and its effect on teaching, more investigation of L2 writing assessment needs to be done in classroom settings.

Assessment of the quality of compositions varies among individual raters, rating settings and purposes and these factors cause variability in rating results. Among these factors, the rater has been considered the most responsible for rating discrepancies. Differences in ratings are a result of raters' diverse expectations of writing, their backgrounds (Milanovic, Saville, & Shuhong, 1996), responses to writers' language backgrounds (Hamp-Lyons, 1989) and differences in raters' academic specialties, sex and amount of exposure to ESL writing (Vann, Lorenz, & Meyer, 1993). Even when the same assessment guidelines are provided, raters tend to apply their own philosophies and perceptions when judging the quality of compositions.

Training raters clearly reduces variability in their assessments. Carlson, Bridgeman, Camp and Waanders (1985) reported that inter-rater reliability was high after raters had been given a training session. Rater training also reduces variability in writing assessments by helping raters to understand rating criteria more clearly (Charney, 1984) and to modify their expectations of good writing (Huot, 1990). Several studies have also shown that training raters reduces assessment variability in large-scale ESL writing assessment. Weigle (1994) reported that after a training process, or norming process, inexperienced raters, who were involved in their first writing assessment task, could provide assessments comparable to expert raters. However, in normal circumstances assessments made by raters or teachers are usually variable. This observation leads to the fundamental question of “What makes a good composition?” and attempts to answer this question require us to revisit and reconsider rating criteria and their significance in a rated composition.

The rating criteria of the early ESL composition research conducted in the 1970s and the 1980s (Grobe, 1981; Homburg, 1984; Larsen-Freeman, 1978; Perkins, 1980) were mainly concerned with such factors as essay length, the number of T-units, error free T-units, and clause per T-unit ratio. Perkins (1980), for example, employed words per composition, sentences per composition, T-units, total errors and other language factors as indicators to assess the quality of the composition. The rating components applied in his research considered only language-related facets, disregarding meta-linguistic factors such as content and development of ideas. In a similar context, the rating criteria in Grobe’s research (1981) centered on such factors as composition length, the number of spelling errors, and words per T-unit. No components in his research considered content or rhetorical organization, which are now major components in many ESL writing assessment profiles. Homburg (1984) was not an exception to this trend since content was not adopted as a component to be assessed. The essays examined in his research were from the writing test results of the Michigan Test of English Language Proficiency, in which assessment focused mainly on the measures of length, subordination and relativization, sentence connectors and number of errors. The assessment criteria applied in these studies were claimed to be objective measures to assess compositions since they can be numerically counted for a quality judgment. In fact, Larsen-Freeman and Storm (1977) reported a noticeable correlation between the number of words in a composition and its holistic evaluation.

In contrast, in L1 writing assessment research performed in the late 1970s

(Freedman, 1979; Harris, 1977) meta-linguistic factors such as content and organization, along with sentence structure and mechanics, were major constituents to rate compositions. Harris (1977) studied high school teachers' responses to student papers in the US and found that the teachers tended to give the most weight to content and organization. Likewise, in her research with L1 essays, Freedman (1979) set up content, organization, sentence structure and mechanics as the four major categories to be evaluated, claiming that these features were more pedagogically meaningful and significant than the prevalent categories of L1 writing assessment of that time such as essay length and the number of spelling errors. Her interpretation of content was concerned with whether ideas stated in an essay were relevant to the topic given and how well they were developed and argued. In other words, she included relevance to a topic as well as logical development of ideas as the assessment guides for judging the quality of content.

It was in the 1990s when meta-linguistic factors such as content and structure in L2 composition started to attract research attention. The assessment of content in L2 compositions has puzzled writing teachers and raters, since there seems to be no consensus of the definition and application of the component. Content in some research (Hamp-Lyons, 1993; Lumley, 2002; Sakyi, 2000; Shi, 2001; Vaughn, 1993; Weigle, 1994) is assessed in terms of how a topic or a task assigned is addressed. This standpoint centers on how ideas stated in a composition are developed and processed, viewing content as the argumentation of ideas. The second perspective of content in ESL/EFL writing (Hamp-Lyons, 1993; Sakyi, 2000; Weigle, 1994) focuses on how appropriately and relevantly a composition addresses a given topic. Following this definition, raters are likely to assess the degree of relevance of a composition to a topic or task. The third view of content concerns what is actually stated in a composition, or the actual content of the writing. This approach to content sometimes entails the judgment of opinions and values expressed in writing: opinions differing from those of raters may negatively affect the rating of a composition. Raters in Vaughn (1993) and Sakyi (2000) seemed to have adopted this view of content. Which of the three concepts is applied to actual assessment depends on the raters, the purpose, and the environment of rating.

Cumming (1990) investigated rating behaviors of novice and expert raters assessing ESL compositions. In his study, the raters rated twelve compositions, adopting the criteria of ‘language use,’ ‘rhetorical organization,’ and ‘substantive content.’ Their rating behaviors were coded and categorized into self-control focus, content focus, language focus and organization focus, which included twenty-eight detailed sub-categories. Among these, “classifying errors” was the most common behavior (19%) for the expert group, while “editing phrases” was the most common behavior in the novice group. Both of these behaviors are concerned with language use problems, broadly speaking.

Hamp-Lyons (1993) researched the rating behaviors of IELTS raters, focusing on whether they were entitled to judge or respond to student writing on tests designed to assess academic writing proficiency. Each rater had an MA or Ph D in TESOL or applied linguistics. When assessing IELTS compositions, they had difficulty understanding the content of those written on social studies and general academic questions. The rating criteria they applied reflected those for the IELTS, which consisted of ‘communicative quality,’ ‘organization,’ ‘argumentation,’ ‘linguistic accuracy’ and ‘linguistic appropriacy.’ Of these, the first three concern meta-linguistic factors, while the others concern linguistic ones. Because this study was designed to look into content-related concerns or comments, language factors were not discussed in it.

Vaughn (1993) examined the assessment behaviors of raters, who assessed six essays collected at a university, two written by native speakers and four by non-native speakers. They were given criteria to assess each composition such as response to topic; language; pattern of development; explanations or illustrations to support assertions; vocabulary; command of syntax and grammar; punctuation, and spelling. They raters commented most frequently on content (30% of comments). Handwriting problems were next most frequently mentioned. Based on the reading styles of the raters, she characterized them into "the single-focus approach," "the first impression dominates approach," "the two-category strategy," "the laughing rater," and "the grammar-oriented rater."

With an intention to examine the process of decision-making behaviors of raters and composition characteristics, Milanovic, Saville and Shuhong (1996) examined decision-making process of raters and found that markers participating in their study adopted four different approaches to the reading of the compositions: the "principled two-scan/read," "pragmatic two-scan/read," "read through" and "provisional mark" approaches. Detailed composition features of length, legibility, grammar, structure, communicative effectiveness, tone, vocabulary, spelling, content, task realization and punctuation were illustrated. Along with these two major findings, they also found that assessments were affected by the proficiency level of the scripts rated and by the raters' different rating experiences and backgrounds.

Cumming, Kantor and Powers (2002) performed a study on raters' decision-making behaviors, asking experienced raters of essays to point to the elements considered to be important while rating. Two groups of raters, who were experienced ESL/EFL raters and experienced English-mother-tongue composition raters, showed different approaches to the assessment of writing. ESL/EFL writers were more attentive to language factors than rhetorical organization, while English-mother-tongue composition raters paid similar attention to language factors as well as meta-linguistic factors. It was also found that the raters' teaching experience had affected their rating behaviors and conceptions of good composition.

Lumley (2002) looked into rating behaviors of raters and the application of features of a rating scale in a study of ESL assessment criteria in a large-scale writing test. He observed that Conventions of Presentation (CoP), Cohesion and Organisation (C&O) and Grammatical Control (GC) attracted almost the same amount of frequency of comments with 27%, 26%, and 27%, respectively, while Task Fulfillment and Appropriacy (TFA) embraced about 21%. Analysis of the rating styles of the raters revealed that they sometimes interpreted and applied the rating scales in quite different ways. On the basis of these findings, he argued that rater training should be carried out to obtain reliable ratings.

Quite recently, Eckes (2008), pointing out that few raters participated in the previous research on L2 writing, conducted a large scale writing assessment in an L2 context with 64 experienced teachers and specialists working in the field of German as a foreign language. The findings showed similar results of the previous research in that the raters differed much in their perceptions of scoring criteria importance and they seemed not to have any common ground for the quality of good writing. Then he categorized raters' rating styles with respect to their views of criterion importance: the syntax type, the correctness type, the structure type, the fluency type, the non-fluency type and the non-argumentation type.

RESEARCH QUESTIONS

The literature on L2 writing shows that research in this field has mostly been conducted in an ESL setting and has investigated ratings made in large scale proficiency or placement test settings. EFL writing assessment in a classroom setting needs to be more conducted for pedagogical purpose. With this concern in mind, this research aims to first investigate

1) What features of essays do teacher-raters focus on in EFL writing assessment performed in a classroom setting?

The second research question addresses rating behaviors of teacher-raters teaching in the same language program. Several studies designed to examine rating styles have been carried out in ESL proficiency test environments, but little research on this topic has been conducted in an EFL classroom setting. The second research question is:

2) How do teacher-raters teaching in an EFL program differ in the evaluation of the

same compositions?

RESEARCH METHODOLOGY

This section describes the characteristics of teacher-raters in this research and writing samples read by them.

Teacher-Raters and Program

The teacher-raters were faculty members teaching EFL in the Division of Humanities and Social Sciences at the researcher’s institution. Four were native speakers of English and one was Korean. The native speakers have their MAs in ESL, education or history and have taught EFL composition for at least five years in Korea, the USA and other countries. The Korean professor holds a Ph.D. in Linguistics and had taught EFL composition for several years. All of the teacher-raters have taught for several years in the freshman English program where this experiment was carried out. With their EFL teaching and grading experience, they were believed to be well qualified to evaluate the compositions in this research.

TABLE 1
Characteristics of Teacher-Raters

Teacher-Rater	Nationality	EFL/ESL Teaching Experience (years)	Degree
1	USA	5	M.A ., Education
2	USA	10	M.A ., TESOL
3	Canada	8	M.A ., History
4	Korea	4	Ph. D., Linguistics
5	USA	6	M.A ., Education

Students who score in the top 1% on the Korean SAT are admitted to the university. However, students had not learned how to write English essays in high school, so this course offered their first practice in writing formal essays in English. Their writings showed problems with grammar, sentence structure, mechanics and vocabulary. The students' paper-based TOEFL scores ranged from 450 to 530. The freshman English program at the school is geared to develop the four skills of English, with particular emphasis on essay writing. It is a two-semester program, in which students learn the basics of composition and then compose several essays of three to five paragraphs as a final product in the second semester. The program hosts regular faculty meetings every two weeks where the instructors of the program discuss the curriculum, classroom management, grading, general affairs, etc. All courses use the same textbook and require the same number of writing assignments. However, the program does not have any detailed grading guides or topics for the student writing.

Writing Samples Assessed

Compositions to be assessed consisted of seven essays with the topic of "How could science and technology contribute to the development of a country?" These essays were assignments written by students in a Freshman English class taught by the researcher. The essays rated varied in length from about 200 words to 330 words. There were no time constraints for writing the essays and the students reported that they had spent two or three hours writing them. The students were believed to have done their best because the essays were to be graded by the instructor of the class, who conducted this research. In fact, seventeen essays on the above-mentioned topic were originally collected and seven were chosen based on the level of the essays: two upper intermediate, two intermediate and three poor level essays. The seven essays selected by the researcher were representative samples of the class' English proficiency.

Data-gathering Procedures

The teacher-raters were instructed to grade the essays as: Very Good,

Good+, Good, Fair, or Poor. Because this study was designed to compare the rating processes and behaviors of the teacher-raters teaching in the same English program, neither rater training nor detailed rating guidelines were provided. Here it must be noted that the English program where the teacher-raters were teaching had not set up grading criteria for essay assignments even though they shared their grading methods and schemes through regular faculty meetings. Reflecting this reality, the teacher-raters were free to assess the compositions with their own assessment styles as if they were reading writing assignments given to their own students.

Data were collected using talk aloud verbal protocol, which was conceived to disclose heeded information and cognitive structures of the mind (Ericsson & Simon, 1993). This method has been used in several studies on ESL compositions (Cumming, 1990; DeRember, 1998; Milanovic, Saville, & Shuhong, 1996; Vaughn, 1993; Weigle, 1994, 1999), despite some defects as a data gathering method (Nisbett & Wilson, 1977). Weigle (1999) mentioned some disadvantages of this method but admitted that it was one of the most convincing ways to identity raters' thought processes while rating. Green (1998) argued that this data-gathering technique in language testing is more direct than *a posteriori* data gathering methods such as oral interviews or questionnaires. Teacher-raters were instructed to approach the essays as if they were grading them for their own class. While reading the essays, the teacher-raters recorded their mental processes by speaking into a tape recorder. They were instructed to say whatever occurred to them about the essay.

Teacher-raters' comments were transcribed and grouped into fourteen categories: Interpreting Language Use, (2) Judging Language Use, (3) Interpreting Rhetorical Organization, (4) Judging Rhetorical Organization, (5) Interpreting Content, (6) Judging Content, (7) Interpreting Vocabulary, (8) Judging Vocabulary, (9) Interpreting Mechanics, (10) Judging Mechanics, (11) Interpreting Writing Style, (12) Judging Writing Style, (13) Summarizing Judgments, and (14) Comments on Ambiguous Phrases. These categories are derived from Cumming's 28-category system (Cumming, 1990), but some categories that seem indistinct or redundant were eliminated. For example, significant overlap occurs between “to establish personal responses to qualities of items” and “to define, assess, or revise one’s own criteria and strategies” which Cumming’s (1990) categorizes under “Self-control Strategies to Guide Judgments.” Furthermore, some criteria categorized into different strategies seem to deal with similar rating behaviors. For example, “to assess development of topics,” which is one of the coding schemes of “Strategies to Judge Content,” is very similar to “to rate organization overall” in “Strategies to Judge Rhetorical Organization.”

In this research, along with the major composition components such as language use, rhetorical organization, content, vocabulary, and mechanics, three other categories of Writing Styles, Ambiguous Phrases and Summarizing Judgments were included due to their peculiar traits. For example, Interpreting Writing Styles and Judging Writing Styles deserved to be separate categories since they assess meta-linguistic factors which are not the main components of composition such as rhetorical organization and content. Similarly, comments on Ambiguous Phrases were categorized separately in case it was not clear what major categories they belonged to. Summarizing Judgments was also assigned to a separate coding category because in many cases it combined components from more than one component of writing.

Interpreting Strategies often involved paraphrasing or summarizing content as a way to confirm the reading or correcting of errors:

> All right, so the first paragraph is . . . begins with a kind of anecdote about the current situation in Korea. (Interpreting Rhetorical Organization)
> Ok, next sentence. “A rice” is an improper use of the article and “hadn’t.” “Did not have” is a more correct tense form. (Interpreting Language Use)

In contrast, Judging Strategies are more direct: raters typically made a simple statement of assessment.

> “So, how can Korea use science and technology to advance its economy” is not a bad introductory paragraph. (Judging Rhetorical Organization)

> This essay does not at all address the question of writing task. (Judging Content)

The Appendix shows more examples of the coded data.

RESULTS

Characteristics of EFL Writing Assessment Performed in a Classroom Setting

The summary of the frequency of comments below includes the percent of negative and positive comments of each category, which indicate the teacher-raters' perception and judgment of the components analyzed.

TABLE 2
Summary of All TRs' Comments by Category

	TR 1	TR 2	TR 3	TR 4	TR 5	Total/(%)	N(%)	P(%)
ILU	93	67	3	36	7	206(41.7%)		
JLU	5	7	8	2	5	27(5.5%)	21(77.8%)	6(22.2%)
IRO	1	2	1		1	5(1.0%)		
JRO	3	17	14	9	3	46(9.3%)	23(50%)	23(50%)
IC		1	9		1	11(2.2%)		
JC		4	6			10(2.0%)	10(100%)	
IV	22	7	3	7	1	40(8.1%)		
JV		1				1(0.2%)	1(100%)	
IM	19	17		10	1	47(9.5%)		
JM								
IWS		3		2		5(1.0%)		
JWS	1	6	5	1	8	21(4.3%)	21(100%)	
SJ	7	7	7	7	7	35(7.1%)		
AP	17	11	2	10		40(8.1%)		
TOTAL	168	150	58	84	34	494(100%)		

Note TR: Teacher-rater, ILU: Interpreting Language Use, JLU: Judging Language Use, IRO: Interpreting Rhetorical Organization, JRO: Judging Rhetorical Organization, IC: Interpreting Content, JC: Judging Content, IV: Interpreting Vocabulary, JV: Judging Vocabulary, IM: Interpreting Mechanics, JM: Judging Mechanics, IWS: Interpreting Writing Style, JWS: Judging Writing Style, SJ: Summarizing Judgment, AP: Awkward Phrases, N: Negative Comments, P: Positive Comments

Language use The first conspicuous finding was that 47% of comments concerned language use. This was probably because the writers' limited language proficiency resulted in problems with language use. However, it is surprising that the component attracted such a high proportion of comments even though language use here was used as a broad concept encompassing grammar, sentence structure and sentence level expressions. The predominance of language use over the other components was observed in all individual Teacher-Raters (TR), except for TR 3: 58.4% for TR 1, 49.4% for TR 2, 19.0% for TR 3, 45.3% for TR 4 and 35.3% for TR 5. The most general rating behaviors of this component were to correct grammatically incorrect phrases and sentences while reading.

> "Because US army," The US and army should be capitalized. "has very advanced technology and their weapons are very effective." "In international," in an international society, delete, the "military power has an effect on every situation." [TR 1, Essay (E) 2]
>
> Next sentence, "And" is capitalized. It should not be. Engineer should be plural. The sentence "But is it desirable to make our country, Korea, wealthy and powerful?" All right. The sentence is not saying what the author wants to say. He should be saying, but this is desirable? [TR 2, E 2]

On the other hand, some of the comments just pinpoint words and phrases with grammatical problems:

> "Today people use many tools which is made by the modern science engineering." Subject-verb agreement, non-idiomatic phrasing here. [TR 2, E 3]

About 78% of statements concerning language use were negative;

> "So they invest that research which has economical things." The idea is clear but the structure is completely awful. [TR 3, E 3]
>
> "Otherwise," hmm, "if they remain there, these people spread in the

> developed countries will be . . .” Hmm, this is not sentence at all, just awful. [TR 5, E 7]

Rhetorical organization The second finding of interest was that about 10% of comments concerned rhetorical organization or structure. This relatively high proportion may occur because rhetorical organization in essay writing is a significant component of EFL classes; this emphasis could have led TRs to pay attention to this factor. All TRs except TR 1 commented on this factor frequently: 2.4% for TR1, 12.6% for TR 2, 26.0% for TR 3, 10.7% for TR 4 and 11.7% for TR 5. Of fifty-one comments concerning rhetorical organization, forty-six were related to judging the factor. Most of the statements on rhetorical organization concerned general organization of the essays, rather than specific aspects of rhetorical organization. Half of these comments were positive. Given that 77.8% of the comments on Judging Language Use and 100% of the comments on Judging Content and Writing Style were negative, this high proportion of positive comments was extraordinary. This illustrates that the students even with limited English proficiency can be relatively good at constructing essays.

The normal approach to assessing rhetorical organization was to judge the component from the perspective of the typical structure of an essay. Factors mentioned include use of well-written topic sentences, the introduction, supporting paragraphs and conclusion, and connecting skills and phrases:

> “Now we will think about the role of scientists for the economic development of a country.” That’s an attempt to write a topic sentence. Not very good but Ok. [TR 2, E 3]
>
> The overall structure is pretty good. I’ll rate this paper high on structures since it uses a topic sentence and controlling ideas. [TR 2, E 4]

Content In this research, content drew only 4.2% of all the comments, implying that the TRs did not pay much attention to this component. In fact, this proportion of comments on content was mostly attributable to TR 3, whose comments on content (25.8%) outnumbered those on language use (19.0%). For the other teacher-raters, the statements on content accounted for 3.4% for TR 2, and 2.9% for TR 5. TR 1 and TR 4 made no comments on content. Typically, comments on content simply noted that essays failed to address the topic properly:

> This essay does not at all address the question of writing task. [TR 5, E 2]
>
> And again the question did not specially ask about Korea. [TR 3, E 1]

Only TR 3 actually interpreted and judged content:

> It starts talking about how the Korean War destroyed almost all infrastructure but now Korea is the leader of the IT industry. What do those two things have to do with each other? I mean it has been 50 years since the Korean War ended and so what? What is the point he is trying to make? [TR 3, E 3]

Mechanics Of all the comments, 9.5% referred to mechanics. TRs commented on this component with noticeably different frequencies. Generally, comments on this component corrected use of awkward and incorrect mechanics:

> Here, they have a period and capital B. We need to get rid of the period and change the capital B to a lower case b. [TR 1, E 2]

However, some TRs just referred to incorrect mechanics:

> Second sentence, “for example,” punctuation problem. “3years” no space between 3 and years. That seems to be something that comes from Korean. [TR 2, E 3]

Writing styles Comments in this category typically noted that a rhetorical question is not an appropriate form for a topic sentence:

> I don’t really like a rhetorical question as sort of a topic sentence, whatever. I think it’s a chief way of writing. But if they have been taught that, then they can do about that. [TR 3, E 1]
>
> The other is a rhetorical question in the introduction, which is bad. [TR 5, E 5]

TR 2, on the other hand, commented on beginning with “Nowadays” as an improper way to begin an essay:

> Beginning with “Nowdays,” ok, improper spelling or . . . I don’t think it’s a good way to begin an essay. [TR 2, E 6]

These examples illustrate that comments on Writing Style appear to be influenced by TRs’ subjective perceptions of factors in the essay. The view of using a rhetorical question as a topic sentence and starting an essay with “nowadays” may differ greatly among EFL composition teachers.

Judging essays All teacher–raters included a final judgment for every essay. The most common judgment style was to summarize previous comments and assign a grade:

> The weakness of this essay uses, that is, a lot of repetition. It has a lot of sentential problems. So here the overall structure’s pretty good but the sentence level structure is very poor. So good essay structure, poor sentence structure and poor use of vocabulary. So I can give this paper. Hmm, I would say a Good. It couldn’t be better. I’d say this paper is Very Good on structure, but Fair or Poor on sentence structure. [TR 2, E 4]
>
> Grammatically, vocabulary-wise no real problem except for that one sentence. The grammar and vocabulary are pretty good but the content, since it is off-topic. It is more about how to improve science and technology. It isn’t that good at all. I have to say that makes this fall in the middle. I think at best this is a good plus paper. This is the end. [TR 3, E 7]

> This essay is a bit weak at structure. It seems like that there is no planning in advance how to construct it and there is no point. No suggestion was made on how science and technology had affect on a country. This essay, from my own judgment, is only a Poor. [TR 4, E 6]
>
> This is not a bad paper, fairly good. It just has mechanical problems, but reasoning is logical except for one sentence. The reasoning is pretty good. Just grammatical problems [TR 1, E 2]

In one extreme case, on the other hand, all comments were devoted to judgment of the essay:

> This is all the best of the essays I've seen so far. It actually is an activity of the essays. Introduction is quite nice. The supporting paragraphs are underdeveloped. That's the main problem with this. There are grammatical problems particularly with some verb usage and couple of periods where periods go. But on the other hand, he actually had tried to divide the essay and what's he talking about is quiet clear. Umm. What's my choice here? I'm gonna say right that I'd give this a Very Good. [TR 5, E 4]

Rating Behaviors of Individual Teacher-raters

Along with the general rating behaviors mentioned above, the rating styles of the individual teacher-raters are reported below.

TR 1

The rating style of TR 1 is characterized by careful and meticulous reading: He read almost every sentence in the essays and tried to correct awkward words, expressions and mechanics. Consequently, 82.8% of the comments he made concerned linguistic factors such as language use, vocabulary and mechanics, which accounted for 58.4%, 13.1% and 11.3% of all comments, respectively (Table 3). The first part of the remarks on Essay 3 clearly illustrates his style:

> Instead of "Today," "should people are" currently "using many tools which is made," are made "by the modern science engineering," modern science and engineering. "For example" missing a comma. "Korea has a Korean war during 3 years," should be Korea has a war for 3 years "so almost infrastructures was destroyed by Korean war," should be so most infrastructures were destroyed during the Korean war. The Korean War is a proper noun, was and were, we have wrong verbs here. During and for, we have preposition problem. [TR 1, E 3]

TABLE 3
Summary of TR 1's Comments by Category

	Essay1	Essay2	Essay3	Essay4	Essay5	Essay6	Essay7	Total(%)	N(%)	P(%)
ILU	5	12	14	21	22	6	13	93(55.4%)		
JLU			1	1	1	1	1	5(3%)	4(80%)	1(20%)
IRO		1						1(0.6%)		
JRO		1				1	1	3(1.8%)	1(33.3%)	2(66.7%)
IC										
JC										
IV	4	3	7	2	1	5		22(13.1%)		
JV										
IM	8	7	3	1				19(11.3%)		
JM										
IWS										
JWS					1			1(0.6%)	1(100%)	
SJ	1	1	1	1	1	1	1	7(4.2%)		
AP	5	2	5	2	3			17(10.1%)		
Total	23	27	31	28	29	14	16	168(100%)		

Note: ILU: Interpreting Language Use, JLU: Judging Language Use, IRO: Interpreting Rhetorical Organization, JRO: Judging Rhetorical Organization, IC: Interpreting Content, JC: Judging Content, IV: Interpreting Vocabulary, JV: Judging Vocabulary, IM: Interpreting Mechanics, JM: Judging Mechanics, IWS: Interpreting Writing Style, JWS: Judging Writing Style, SJ: Summarizing Judgment, AP: Awkward Phrases, N: Negative Comments, P: Positive Comments

Since his rating behaviors mainly addressed language factors, he seemed to ignore meta-linguistic components such as content and rhetorical organization, which attracted no comments and 2.4% of all the comments, respectively. His rating style resembles that of the grammar-oriented rater in Vaughn's (1993) research, who reacted to almost all grammatical items. As noted above, TR 1 read sentence by sentence, commenting on factors related to language use. His reading style is also similar to the "read through" (Milanovic et al., 1996) in which markers adopting this approach read a script only once to pick up its good and bad points.

TR 2

TR 2 read the essays with a balanced approach by referring to the major components of a composition such as language use, rhetorical organization and content all together, even though language use alone drew 49.4% of al the comments. The factor that he mentioned second most-frequently was rhetorical organization with a proportion of 12.6%. Given the fact that the compositions rated here were formal essays, which required a well-structured organization, his reading style appeared to be appropriate. The following transcription is typical of his view of rhetorical organization:

> It has an introduction and a conclusion. The second paragraph states a kind of theoretical statement. The third paragraph is a concrete illustration of the theory which is pretty good for the overall structure for an essay. [TR 2, E 5]

TABLE 4
Summary of TR 2's Comments by Category

	Essay1	Essay2	Essay3	Essay4	Essay5	Essay6	Essay7	Total(%)	N(%)	P(%)
ILU	12	8	10	13	9	9	6	67(44.7%)		
JLU	1	1	1	1	1	1	1	7(4.7%)	5(71%)	2(29%)
IRO	1	1						2(1.3%)		
JRO	3	2	2	6	1		3	17(11.3%)	6(35.3%)	11(64.7%)
IC							1	1(0.7%)		
JC		2	1				1	4(2.7%)	4(100%)	
IV	1	1		1	3	1		7(4.7%)		
JV				1				1(0.7%)	1(100%)	
IM	3	1	5	1	3	3	1	17(11.3%)		
JM										
IWS				1	1		1	3(2.0%)		
JWS		1		3		2		6(4.0%)	6(100%)	
SJ	1	1	1	1	1	1	1	7(4.7%)		
AP	2		1		1	4	3	11(7.3%)		
Total	24	18	21	28	20	21	18	150(100%)		

Note: ILU: Interpreting Language Use, JLU: Judging Language Use, IRO: Interpreting Rhetorical Organization, JRO: Judging Rhetorical Organization, IC: Interpreting Content, JC: Judging Content, IV: Interpreting Vocabulary, JV: Judging Vocabulary, IM: Interpreting Mechanics, JM: Judging Mechanics, IWS: Interpreting Writing Style, JWS: Judging Writing Style, SJ: Summarizing Judgment, AP: Awkward Phrases, N: Negative Comments, P: Positive Comments

In the meantime, he often pointed to the transfer of the Korean language to the construction of the whole essay and sentence structure and the writing of phrasal expressions, as seen below:

> I would say this paper follows a very Korean style structure. Well, it lists many different examples and then makes them a conclusion in the end. Giving the topic and main idea at the end of essay is not a very good structure for English writing. [TR 2, E 2]
>
> The third sentence is again . . . needs another clause in there and begins with "because," which is transfer error from Korean, "Oenyahameon." [TR 2, E 1]

He also comments on direct translation of Korean into English.

> It seems to almost be translated words from Korean. [TR 2, E 2]

Mentioning the language transfer and direct translation of Korean may be attributable to his many years of teaching English in Korea, which in turn has led to his familiarity with the typical errors of English made by Korean students.

TR 2's marking style could be termed a "three category" strategy (Vaughn, 1993) because he mostly focused on language use, rhetorical organization and mechanics. His reading style is also quite similar to the "provisional mark" approach (Milanovic et al., 1996) in that he read the essays only once with several pauses to evaluate what he had read to that point, then resumed his reading with some expectations.

> Ok, so far so good. Especially the structure seems to be OK. Hmm, I'm kind of interested in how the ideas will be developed. [TR 2, E 7]

TR 3

This rater was the only teacher-rater whose comments on content (25.8%) outnumbered those on language use (19.0%). His statements on content can be assigned to two groups: one concerning ideas per se stated in the essays, and the other about topic relevance.

> Hmm, the first body paragraph doesn't really have any details and examples. It's not really a paragraph. It's a sentence, rather than a paragraph. It doesn't have any details and doesn't have any examples. The second one is talking about unreal future, unreal condition, which again I guess it could be ungrounded reality, right? And the third body paragraph is the probably one that makes most sense, hmm, content-wise. [TR 3, E 4]
>
> They're talking about the economy and making money. Hmm, there is not clear connection to how entrepreneurs wanting to invest in science lead to improvement in science, which leads to improvement and the economy and anything like that. [TR 3, E 3]

Along with comments on the content stated in the essays, the teacher-rater noted failure to address the topic:

> Content-wise this essay doesn't actually do with the question. [TR 3, E 7]

It should also be noted that rhetorical organization was responsible for 26.0% of this rater's comments. When referring to this component, TR 3 applied the typical structure of an essay. The standard structure was viewed as a requirement, not as something to be evaluated highly.

> The conclusion is a pretty standard conclusion. It is just statement of the idea they are supposed to be talking. But since the introduction is kind of off-topic, and the second paragraph doesn't make much sense, it's not a really conclusion into the paper. [TR 3, E 6]

TABLE 5
Summary of TR 3's Comments by Category

	Essay1	Essay2	Essay3	Essay4	Essay5	Essay6	Essay7	Total(%)	N(%)	P(%)
ILU	3							3(5.2%)		
JLU	2	1	2	1	1		1	8(13.8%)	6(75%)	2(25%)
IRO				1				1(1.9%)		
JRO	2	1	1	3	3	1	3	14(24.1%)	5(35.7%)	9(64.3%)
IC	1	1	1	1	1	2	2	9(15.5%)		
JC	1	2	1		1		1	6(10.3%)	6(100%)	
IV	1		1		1			3(5.2%)		
JV										
IM										
JM										
IWS										
JWS	1		1		2	1		5(8.6%)	5(100%)	
SJ	1	1	1	1	1	1	1	7(12.1%)		
AP	2							2(3.4%)		
Total	14	6	8	7	10	5	8	58(100%)		

Note: ILU: Interpreting Language Use, JLU: Judging Language Use, IRO: Interpreting Rhetorical Organization, JRO: Judging Rhetorical Organization, IC: Interpreting Content, JC: Judging Content, IV: Interpreting Vocabulary, JV: Judging Vocabulary, IM: Interpreting Mechanics, JM: Judging Mechanics, IWS: Interpreting Writing Style, JWS: Judging Writing Style, SJ: Summarizing Judgment, AP: Awkward Phrases, N: Negative Comments, P: Positive Comments

> Ok. Looking at no 4, it has a very basic five paragraph essay structure, right? The introduction, three body paragraphs, and the conclusion is the fifth paragraph. Hmm, structurally it is very basic and there is nothing wrong with it. It is just a sort of standard, right? [TR 3, E 4]

Taken all together, TR 3's rating behaviors centered on meta-linguistic factors such as content and rhetorical organization, rather than on language factors. It is interesting to speculate that this focus may be a consequence of his education in History, rather than in ESL or Linguistics.

TR 3's reading style, like TR 2, would be the "three category" approach (Vaughn, 1993). One very peculiar characteristic is that this rater approached the content of the essays with very negative viewpoints. His reading behaviors also resembles the "provisional mark" approach (Milanovic et al., 1996) on the grounds that he read the essays with some pauses in order to judge what had been read so far, as shown in the comments on Essays 4 and

6 above.

TR 4

This rater, who is a non-native speaker of English, was very similar to TR 2 in terms of frequency of the comments on composition: 68% of her statements concerned language use, mechanics and rhetorical organization (Table 6). On the other hand, her reading style resembled that of TR 1: Reading sentence-by-sentence, she corrected grammatically awkward and incorrect expressions and made some suggestions for corrections. Compared to the reading of TR 1, who tried to amend almost all mistakes, TR 4 sometimes showed hesitation in her suggestions for corrections and questioned herself as a way to confirm them:

> "Business of people" looks awkward. Hmm, instead of that, people's work or other seems to be. [TR 4, E 1]
>
> I don't know what that means. Does it mean science and technology doesn't affect the economy of a country? Right? [TR 4, E 3]
>
> It would be much better put "undoubtedly" before the 21st century. Oh, really? Is that new information? [TR 4, E 7]

Similarly, instead of suggesting specific corrections, she often stated "Rewrite this," (twice in Essay 1, three times in Essay 2, twice in Essay 3, once in Essay 4, three times in Essay 6 and five times in Essay 7).

Her reading style, which was characterized by frequent statements of "Rewrite this" and hesitant corrections, was probably due to the fact that she is not a native speaker of English, even though she has been teaching EFL composition for many years and has attained native-like proficiency in English. TR 4's reading behaviors resemble the "provisional mark" approach (Milanovic et al., 1996); she read the essays only once with occasional pauses in order to clarify and double-check her own judgments, as shown in the comments above.

TABLE 6
Summary of TR 4’s Comments by Category

	Essay1	Essay2	Essay3	Essay4	Essay5	Essay6	Essay7	Total(%)	N(%)	P(%)
ILU	1	2	5	6	10	10	2	36(42.9%)		
JLU		1					1	2(2.4%)	2(100%)	
IRO										
JRO	1	1	1	2	1	2	1	9(10.7%)	9(100%)	
IC										
JC										
IV	2	1		1	2	1		7(8.3%)		
JV										
IM	4	1	5					10(11.9%)		
JM										
IWS				1		1		2(2.4%)		
JWS						1		1(1.2%)	1(100%)	
SJ	1	1	1	1	1	1	1	7(8.3%)		
AP	1	2	1	1	1	3	1	10(11.9%)		
Total	10	9	13	12	15	19	6	84(100%)		

Note: ILU: Interpreting Language Use, JLU: Judging Language Use, IRO: Interpreting Rhetorical Organization, JRO: Judging Rhetorical Organization, IC: Interpreting Content, JC: Judging Content, IV: Interpreting Vocabulary, JV: Judging Vocabulary, IM: Interpreting Mechanics, JM: Judging Mechanics, IWS: Interpreting Writing Style, JWS: Judging Writing Style, SJ: Summarizing Judgment, AP: Awkward Phrases, N: Negative Comments, P: Positive Comments

TR 5

Of the teacher-raters, TR5 made the fewest comments: only 34, compared to the average of 115 for the other raters (Table 7). He commented fewer than five times on Essays 6 and 7. The comments of Essay 5 are noted here for their clear representation of his reading style:

> This is not an essay. This is something that he wrote in Korean and then translated. He might use a computer program. I don’t know, but this isn’t an essay. It doesn’t make sense. This is poor. [TR 5, E 5]

This low frequency of comments was probably due to his unique rating manner of first scanning the whole text, followed by the evaluation of the other essay components. Looking over the text quickly, TR 5 seemed to have judged the quality of the essays in advance:

> He's trying to develop two points here in these supporting paragraphs. [TR 5, E 1]
> This is a difficult essay that I give a grade because all the problems are actually language problems. [TR 5, E 2]
> First of all, the message is too short. There are lots of problems in proper usage. [TR 4, E 3]
> This is all for the best of the essays I've seen so far. It actually is an activity of the essays. [TR 5, E 4]
> This is not an essay. This is something that he wrote in Korean and then translated. [TR 5, E 5]
> This is the last paper that I'm holding here. I really want to compare this to essays 2 and 5 because I think it's quite similar in some ways. [TR 5, E 7]

TR 5's reading style was much different from other teacher-raters in that he seemed to first scan the whole text and then went into details, mainly focusing on writing style and final judgments. In this respect, his reading resembles the "pragmatic two-scan/read" approach (Milanovic et al., 1996). Markers who adopt this reading style may read the script twice before making the final judgment; in this approach the second reading is done when markers faced difficulties in interpreting texts. After the first reading, TR 5 appeared to comment in detail on problems in the essay components. In addition, TR 5 was the only teacher-rater who compared essays when deciding on the final mark, as shown below:

> So I'm going to initially assign this a grade of good that a . . .This would definitely be an essay I'd want to review after looking over all essays. [TR 5, E 5]

TABLE 7
Summary of TR 5's Comments by Category

	Essay1	Essay2	Essay3	Essay4	Essay5	Essay6	Essay7	Total(%)	N(%)	P(%)
ILU	1	2	1	1	2			7(20.6%)		
JLU	1	1	1	1	1			5(14.7%)	4(80%)	1(20%)
IRO	1							1(2.9%)		
JRO			1	1	1			3(8.8%)	1(33.3%)	2(66.7%)
IC		1						1(2.9%)		
JC										
IV					1			1(2.9%)		
JV										
IM							1	1(2.9%)		
JM										
IWS										
JWS	1	1	1	1	1	2	1	8(23.5%)	8(100%)	
SJ	1	1	1	1	1	1	1	7(20.6%)		
AP										
Total	5	6	5	5	7	3	3	34(100%)		

Note: ILU: Interpreting Language Use, JLU: Judging Language Use, IRO: Interpreting Rhetorical Organization, JRO: Judging Rhetorical Organization, IC: Interpreting Content, JC: Judging Content, IV: Interpreting Vocabulary, JV: Judging Vocabulary, IM: Interpreting Mechanics, JM: Judging Mechanics, IWS: Interpreting Writing Style, JWS: Judging Writing Style, SJ: Summarizing Judgment, AP: Awkward Phrases, N: Negative Comments, P: Positive Comments

Summary of rating behaviors of Teacher-Raters

Simple comparison of the rating behaviors of the frequency of comments and of the rating styles of the teacher-raters do not reveal common marking features. The teacher-raters of this study used a variety of rating styles: meticulous reading of TR 1, well-balanced reading of TR 2, content-focused reading of TR 3, and hesitating reading of TR 4 and quick scan reading of TR 5.

Inter-rater Reliability

The rating behaviors of the teacher-raters of this study differed greatly. To examine how these differences influenced the final judgment of the essays, descriptive statistics and inter-rater reliability between the teacher-raters were calculated. All essays were given grades of Very Good, Good+, Good, Fair, or Poor (Table 8). Essays 3 and 5 rated by rater TR 1 are exceptions, because he did not provide assessments for these essays.

Grades assigned by the TRs present an acceptable degree of agreement. Ratings agreed perfectly for Essays 2, 3 and 7 if Good and Good+ are to be considered the same. Ratings of Essays 1, 5 and 6 showed a scale difference at most. For example, in Essay 1, two teacher-raters rated Good or Good+, while the other three gave a grade of Fair. On the other hand, the ratings of Essay 4 varied widely, from Fair to Very Good.

Descriptive statistics support the conclusion of acceptable inter-rater reliability. Ratings of Poor, Fair, Good (Good+), and Very Good were converted into the numeric values 1, 2, 3 and 4, respectively and the results were analyzed using Spearman Rank-order correlation. The Spearman's correlation coefficients between the teacher-raters ranged from .52 to .93 (Table 9). Except for inter-rater reliability between TR 4 and TR 5, which is .52, the others fell into an acceptable range (Landis & Koch, 1977). Overall agreement between the teacher-raters adopting the suggestion of Hatch and Lazaraton (1991) was .77. This finding suggests the teacher-raters of this research seemed judge the quality of the essays similarly.

TABLE 8
Essay Ratings by Teacher-Raters

	Essay 1	Essay 2	Essay 3	Essay 4	Essay 5	Essay 6	Essay 7
Rater 1	Fair	Good+		Fair		Poor	Good
Rater 2	Good	Good+	Poor	Good	Very Good	Fair	Good
Rater 3	Fair	Good+	Poor	Good+	Very Good	Poor	Good+
Rater 4	Good+	Good	Poor	Fair	Good+	Poor	Good+
Rater 5	Fair	Good	Poor	Very Good	Good+	Poor	Good

TABLE 9
Spearman Rank-Order Correlation Coefficients of Teacher-Raters

	TR 2	*TR 3*	*TR 4*	*TR 5*
TR 2	1.00	.93**	.82*	.72
TR 3	.93	1.00	.73	.84*
TR 4	.82	.73	1.00	.52
TR 5	.72	.84	.52	1.00

* Correlation is significant at the 0.05 level.
** Correlation is significant at the 0.01 level.

DISCUSSION

The results of this research present some peculiar features of evaluation of EFL compositions: TRs commented most frequently on language use and rarely on content. This study also reveals a moderate inter-rater agreement among the TRs, despite the differences in their rating styles.

About 47% of comments addressed language use. This component was most frequently indicated by four of the five teacher-raters with proportions of 58.4% for TR 1, 49.4% for TR 2, 45.3% for TR 4 and 35.3% for TR 5. In general, EFL writings demonstrate problems with language use due to the writers' lack of skill in the target language, particularly regarding sentence structure, grammar and vocabulary. This trait could have led the teacher-raters to comment frequently on the factor. However, the high frequency of comments on language use does not necessarily mean that this component was the most decisive factor in assessing the quality of writing. Nevertheless, the finding that about 78% of the comments on language use were negative suggests that the writing component seemed to have been important in determining essay quality. This result, however, is not consistent with findings of other research on L1 and L2 writing assessment. In an early study of L1 composition, Stewart and Grobe (1979) reported that composition length and correct spelling had major effects on the judgment of writing quality. Grobe (1981) found that vocabulary was the most important factor considered in assessment of writing quality. Bridgeman and Carlson (1983) determined that meta-linguistic features such as organization, development of ideas, paragraph organization, addressing the topic and overall writing were more important factors than linguistic features such as sentence structure and vocabulary usage.

On the other hand, about half of the comments related to rhetorical organization were positive. This finding is well contrasted with the fact that

78% of the comments on language use and 100% on those of content and writing style were negative. The high proportion of positive comments on rhetorical organization signifies that teacher-raters were generally satisfied with the quality of the structure of the essays and that the students were good at constructing essays, despite their limited English proficiency. In fact, in one survey (Cho, 2006) the freshmen of the school where this investigation was performed responded that they felt most comfortable with structure when writing, followed by content, grammar and vocabulary. In other words, they indicated that structure and content were easy for them and that language use gave them the greatest difficulty.

Next, only 4.2% of TR comments concerned content, which was consistent among the TRs; it was only 3.4% for TR 2 and 2.9% for TR 5; TRs 1 and 4 never commented on content. Only TR 3 made frequent (25.8%) comments on the factor. This low frequency of comments on content contrasts with many other studies of composition assessment, in which it was the most decisive factor (Freedman, 1979), or the most frequently mentioned factor (Vaughn, 1993) or attracted as many comments as other major components of compositions (Lumley, 2002). However, our finding corresponds to one study of an undergraduate English faculty’s view of content (Bridgeman & Carlson, 1983), in which the importance of content was ranked ninth, which implies that the component was not considered to be as important as other major components of writing such as organization, development of ideas, and sentence structure. This relative disregard for content found in this research is consistent with the actual ratings of student essays written as assignments in the freshman program of the college where this research was done. In the data collected for a study (Cho, 2006), the teacher-raters did not mention content in the feedback to student essays.

The last issue addresses an attribute of assessment carried out in a classroom setting. Even though the interpretation of inter-rater reliability varies, it is generally agreed that inter-rater agreement over 0.7 is acceptable (Landis & Koch, 1977). Summarizing rater consistency of L1 and L2 writing assessment, Brown, Glasswell and Harland (2004) reported that inter-rater correlation coefficients of standardized tests ranged from .70 to .80, but that for portfolio writing samples it varied from .40 to .60. Except for cases in which experienced and professional raters trained for a specific large scale test were involved, it was difficult to find studies in which inter-rater reliability of ESL/EFL composition assessment was over .80. In an investigation performed in Korea (Lee, 1998), the inter-rater consistency of three groups assembled according to raters' teaching and grading experience was very low: .28, .36, and .47, respectively. Even in a study on a large scale assessment test using highly experienced raters who taught and rated the Vermont Writing Assessment, rater agreement showed a moderate range of .40 to .60 (DeRember, 1998). Similarly, experiments in rater training have shown that it does not eliminate severe variations in assessments (Lumley & McNamara, 1995; Weigle, 1994). The overall inter-rater reliability of this study, .77, which is similar to the inter-rater agreement of standardized writing tests, was somewhat unexpected because the teacher-raters participating in this research did not have any rater training specific to this study. This inter-rater agreement may occur because they had taught in the same English program for several years and had had meetings twice a month to discuss a variety of concerns about the program. Referring to a term adopted in social science (Lave & Wenger, 1991), the teacher-raters of this study appear to have formed a "community of practice" by working together in the same teaching setting, thereby leading them to share similar concepts about the quality of writings. This community of practice may have contributed to promoting the consistency of ratings made by the teacher-raters of this research.

IMPLICATIONS FOR THE TEACHING OF COMPOSITION IN AN EFL SETTING

The results of this study have implications to the teaching of composition in an EFL setting, particularly in Korea. First, that the language components attracted the highest number of comments implies that the compositions rated had numerous problems with English language use. This leads to a suggestion that teaching of language related factors such as grammar, sentence level structure and expressions should be emphasized, especially to low level students. This suggestion may lack sufficient theoretical background and empirical data since the essays rated here represent a group of students who were taking a freshman English class at a university in Korea. If the writings of advanced students were collected and rated, different results might emerge. Irrespective of this possibility, the observation made here could be applied, to a great extent, to the teaching of English composition at universities in Korea. Because English composition is rarely taught in high schools in Korea, English writing skills of high school students are generally very poor, and this leads to poor English composition skills in university students. Considering that the students whose essays were rated for this research represent the top 1% of students based on the Korean SAT, it is not difficult to imagine the general English writing proficiency of Korean university students. Given that language problems diminish the quality of compositions, the teaching of compositions in an EFL setting should be done with more emphasis on English grammar and general language use.

Next, that rhetorical organization presents the highest proportion of positive comments suggests how to teach the component. Since the component was judged to be relatively good, it could be less weighted, compared to other language-related components such as sentence structure, grammar and vocabulary. Generally, as the writing experience of students accumulates, the weight of the components can be adjusted. For example, the weight of rhetorical organization can be reduced as students become more proficient at this skill. On the other hand, if some factors of composition remain undeveloped, the weight of these factors can be increased over time. This way of adjusting the weight of the components of a composition can lead students to concentrate more on the factors emphasized by teachers and may in turn boost overall writing ability.

Lastly, the Freshman English program discussed here should set up common assessment criteria for the instructors. This task may require a lot of work but it is worthwhile considering the fairness and washback effect of testing. Having explicit assessment guides would lead the students and instructors to have a clear picture of the writing tasks and the goals of the program. At the same time, however, the instructors' teaching philosophy and perceptions of good writing should be respected.

THE AUTHOR

Dr. Dongwan Cho is Associate Professor in the Division of Humanities and Social Sciences of the Pohang University of Science and Technology, Pohang, Korea. His research interests include EFL writing assessment and TEFL.
Email: dongwanc@postech.ac.kr

REFERENCES

Bridgeman, B., & Carlson, S. (1983). *Survey of academic writing tasks required of graduate and undergraduate foreign students.* Princeton, NJ: Educational Testing Service.

Brown, G. T. L., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing, 9*(2), 105-121.

Carlson, S., Bridgeman, B., Camp, R., & Waanders, J. (1985). *Relationship of admission test scores to writing performance of native and nonnative speakers of English* (TOEFL Research Report No. 19). Princeton, NJ: Educational Testing Service.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English, 18*(1), 65-81.

Cho, D. (2006). A case study of a college freshman English program designed to boost writing skills. *Foreign Languages Education, 13*(2), 69-91.

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

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal, 86*(1), 67-96.

DeRember, M. L. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing, 5*(1), 7-29.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155-185.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol Analysis: Verbal Reports as Data.* Cambridge, MA: MIT Press.

Freedman, S. (1979). How characteristics of student essays influence teachers' evaluations. *Journal of Educational Psychology, 71*, 328-338.

Green, A. (1998). *Verbal protocol analysis in language testing research*. Cambridge: Cambridge University Press.

Grobe, C. (1981). Syntactic maturity, mechanics and vocabulary as predictors of quality ratings. *Research in the Teaching of English, 15*, 75-85.

Hamp-Lyons, L. (1989). Raters respond to rhetoric in writing. In H. Dechert & G. Raupach (Eds.), *Interlingual processes* (pp. 229-244). Tubingen, Germany: Gunter Narr Verlag.

Hamp-Lyons, L. (1993). Reconstructing academic writing proficiency. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 127-153). Norwood, NJ: Ablex Publishing Corporation.

Harris, W. (1977). Teacher respond to student writing: a study of the response patterns of high school English teachers to determine the basis for teacher judgment and student writing. *Research in the Teaching of English, 11*, 175-185.

Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. New York: Newbury House.

Homburg, T. J. (1984). Holistic evaluations of ESL compositions: can it be validated objectively? *TESOL Quarterly, 18*, 87-107.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60*, 237-263.

Landis, J, R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Larsen-Freeman, D. (1978). An ESL index of development. *TESOL Quarterly, 12*, 439-448.

Larsen-Freeman, D., & Storm, V. (1977). The construction of a second language index of development. *Language Learning, 27*, 123-134.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.*

New York: Cambridge University Press.

Lee, Y. (1998). An investigation into Korean markers' reliability for English writing assessment. *English Teaching, 53*(1), 54-71.

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.

Lumely, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing, 12*(1), 54-71.

Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision-making behavior of composition markers. In M. Milanovic & N. Saville (Eds.), *Studies in language testing 3* (pp. 92-114). Cambridge: Cambridge University Press.

Nisbett, R. E., & Wilson, W. T. D. (1977). Telling more than we know: Verbal reports on mental processes. *Psychological Review, 84*, 231-259.

Perkins, K. (1980). Using objective methods of attained writing proficiency to discriminate among holistic evaluators. *TESOL Quarterly, 14*, 61-69.

Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: how raters evaluate compositions. In A. J. Kunnan (Ed.), *Studies in language testing 9* (pp. 129-152). Cambridge: Cambridge University Press.

Shi, L. (2001). Native- and non-native speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing, 18*, 303-325.

Stewart, M., & Grobe, C. (1979). Syntactic maturity, mechanics of writing and teachers' quality ratings. *Research in the Teaching of English, 13*, 207-215.

Vann, R. J., Lorenz, F., & Meyer, D. (1993). Error gravity: Faculty response to errors in the written discourse of nonnative speakers of English, In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 181-195). Norwood, NJ: Ablex Publishing Corporation.

Vaughn, C. (1993). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111-125). Norwood, NJ: Ablex Publishing Corporation.

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

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing writing, 6*, 145-178.

APPENDIX

Examples of Coded Data

Comment on Interpreting Language Use: ILU

Ok, next sentence. “a rice” is an improper use of the article and “hadn’t.” “Did not have” is a more correct tense form.

Comment on Judging Language Use: JLU

“For the last ten years”, Ok. . . that’s good.

Comment on Interpreting Rhetorical Organization: IRO

All right, so the first paragraph is . . . begins with a kind of anecdote about the current situation in Korea.

Comment on Judging Rhetorical Organization: JRO

“So, how can Korea use science and technology to advance its economy?” is not a bad introductory paragraph.

Comment on Interpreting Content: IC

They are talking about the economy and making money. Hmm, but there is no clear connection to how entrepreneurs wanting to invest in science lead to improvement in science which leads to improvement and the economy and anything like that.

Comment on Judging Content: JC

This essay does not at all address the question of writing task.

Comment on Interpreting Mechanics: IM

I’ll have a comma here and a lower case, ‘s’ and Korea in Korean is not capitalized for some reason so they have problems here.

Comment on Interpreting Vocabulary: IV

Word choice, scientific technique . . . should be science and technology.

Comments on Summarizing Judgments: SJ

So it's not very helpful in content-wise but structurally and grammatically it's Ok. So this one would be a good or good+.

Comments on writing style: WS

And this student is perhaps trying to be a little but more creative than his ability allows. Using the rhetorical question, I think it'll be much better to stick to a very clear topic sentence with a controlling idea.

Comments on Ambiguous Phrases: AP

The first sentence in the second paragraph is very unclear. It doesn't make much sense. Especially, the last part "by not those but preoccupying of overseas market using the production capability, technology and originality."