The Study on the Rater Reliability of Three Scoring Methods in Assessing Argumentative Essays: Holistic, Analytic, and Multiple-Trait Scoring Methods

Jonggeum Park
(Seoul National University)

Abstract

Various studies have been conducted to minimize the subjectivity and increase the accuracy in assessing written texts, and the present study focused on the scoring rubrics which were the basic criteria for evaluating writing. Three different scoring rubrics (holistic, analytic and multiple-trait scoring method) were compared in evaluating argumentative essays written by Korean high school students. The present study aims to investigate the rater-reliability of the three scoring methods, holistic, analytic, and multiple-trait scoring methods. Scores of the five raters which were obtained from using the three scoring methods were compared. It was found that there were significant mean differences in the three scoring methods. Raters gave the relatively low scores when they used the holistic scoring. Next, the highest inter-rater reliability was found in the multiple-trait scoring. All the three scoring methods showed an acceptable level of reliability above .07. However, raters showed the highest reliability when they used a multiple-trait scoring rubric. Also, high correlation was found among components of analytic and multiple-trait scoring methods, indicating that the multiple-trait scoring rubric can replace the analytic scoring rubric. Finally, raters expressed a favor over the multiple-trait scoring. The result of this study suggests some implications for writing assessment in Korean secondary English classes.

I. INTRODUCTION

1. The Background and Purpose of the Study

In Korean English classrooms, writing has been considered the least important compared with other skills. In this respect, Kwon (2004) called for a balanced development of English skills in Korean secondary school students. Also, Kwon (2006) suggested the need for developing English production skills test and introduced several attempts to develop production tests in Korea. The most difficult aspect in assessing production skills test is its inevitable inclusion of raters' subjective decisions. Various mechanisms have been studied for minimizing the subjectivity and improving the accuracy of scoring such as use of explicit scoring rubrics, use of a long scoring scale, use of augmentation of holistic grades, cross-checking or moderation of marking, systematic scoring processes, rater training (Brown et al., 2004). However, there have

been few studies concerning the accuracy of various scoring methods. To increase the reliability of any performance test, the development and use of a proper scoring method is very important. In this sense, developing a proper scoring rubric for evaluating written texts should be considered importantly. Therefore, comparison studies between different scoring methods will show which scoring method is more reliable.

There are three scoring methods in assessing written texts; holistic, analytic and multiple-trait scoring. Holistic and analytic scoring methods have been used widely across various settings while multiple-trait scoring method is quite a new way to evaluate texts of different genres.

Several studies attempted to investigate the effectiveness of various scoring methods. However, compared with several studies on the relationship between holistic and analytic scoring method (Breland, 1983), there have been little researches on the rater-reliability of multiple-trait scoring method and its comparison with holistic and analytic scoring methods. Therefore, this paper investigates the rater-reliability of three different scoring methods, holistic, analytic, multiple-trait, and seeks to find out which scoring method will be the most effective in secondary school English classrooms in Korea.

2. Research Hypothesis

1) Will there be any difference in the mean scores of the three scoring methods?
2) Will there be any difference in the inter-rater reliability of the three scoring methods?
3) Will there be a correlation among components of analytic and multiple-trait scoring methods?
4) How will raters feel about the three scoring methods?

II. LITERATURE REVIEW

1. Approaches to Scoring

According to Hyland (2003) there are three scoring methods in assessing writing; holistic, analytic, and trait-based scoring methods. Some scholars assume trait-based scoring method as a part of the analytic scoring method (Weigle, 2002), but the trait-based scoring model is clearly different from the analytic scoring method, in that it provides a clear picture of the basic genre requirements rather than vague descriptors often found in the analytic scoring model (Hyland, 2003).

1) Holistic Scoring

In holistic scoring, each text is read quickly and judged according to a scoring rubric that describes the scoring criteria. A rater assigns a single score to a text based on the overall impression of the text (Hyland, 2003; Weigle, 2002). This scoring method reflects the idea that writing is a unidimensional entity and can be captured by a single scale which integrates the inherent qualities of the writing (Hyland, 2003).

There are some advantages of holistic scoring (Hyland, 2003; Weigle, 2002; White, 1984). First, it is faster and less expensive since the text is read only once quickly and assigned a single score. Also, it focuses on the strength of the writing, not on the deficiencies, by emphasizing what the writer can do well. Finally, it is more valid than analytic scoring method because it reflects the reader's whole reaction, not focusing on too many details as in the analytic scoring methods. Homburg (1984) claimed that holistic evaluation of ESL compositions, with training to familiarize readers with the types of features present in ESL compositions, can be considered to be adequately reliable and valid.

On the other hand, some disadvantages of holistic scoring are found. First of all, a single score can not provide a useful diagnostic feedback about a writer's writing ability (Hamp-Lyons, 1995). In particular, this is very problematic for second language writers because different aspects of writing ability develop at different rates for different writers. For example, some writers have good skills in dealing with content and organization, but have poor skills in grammar and vice versa. The holistic scoring can not provide these ESL writers with the necessary feedback they need since it can not distinguish writing components each other. Second, composite holistic scores are not always easy to interpret because raters do not always use the same criteria to give the same score to the text. For example, one rater may give a point 5 focusing mainly on the content and organization while the other rater may give a point 5 focusing on the linguistic features. Moreover, raters may overlook subskills, and finally much rater training is required to reach a mutual agreement on the specific criteria.

The best-known holistic rubric in ESL is TWE scoring guide in TOEFL writing test. The rubric identifies six levels of descriptors which describe the syntactic and rhetorical qualities of the writing.

2) Analytic Scoring

In analytic scoring, texts are read more than once, each time focusing on several different categories which are considered to be features of good writing. Therefore, it

provides more specific, detailed information about the different aspects of writing.

The primary advantage of analytic scoring is that it can provide more useful diagnostic information about students' writing abilities. In addition, it is more useful in rater training because inexperienced raters can understand and apply the criteria more easily. The analytic scoring is also proper for ESL writers who show an uneven profile across different aspects of writing. Finally, it is more reliable than holistic scoring because writers get several scores for several different categories (Hyland, 2003; Weigle, 2002).

One of the disadvantages of analytic scoring is that it takes longer time, and thus costs more money than holistic scoring. Also, descriptors may overlap or be ambiguous. In addition, if scores on the different scales are added to make a composite score, a good deal of the information provided by the analytic scale is lost. Most seriously, however, the analytic scoring has the danger of the halo effect where results in rating one scale can influence the rating of others.

The well-known analytic scoring rubric is ESL Composition Profile developed by Jacobs et al. (1981). It has five distinguished and differentially weighted categories; content (30 points), language use (25 points), organization, vocabulary (20 points respectively) and mechanics (5 points).

Weigle (2002) presents a comparison of holistic and analytic scales based on the six qualities of test usefulness suggested by Bachman and Palmer (1996): reliability, construct validity, practicality, impact, authenticity and interactiveness. It is presented in Table 1.

TABLE 1
A Comparison of Holistic and Analytic Scales on Six Qualities of Test Usefulness (Weigle, 2002, p. 121)[1]

Quality	Holistic Scale	Analytic Scale
Reliability	lower than analytic but still acceptable	higher than holistic
Construct Validity	holistic scale assumes that all relevant aspects of writing ability develop at the same rate and can thus be captured in a single score; holistic scores correlate with superficial aspects such as length and handwriting	analytic scales more appropriate for L2 writers as different aspects of writing ability develop at different rates
Practicality	relatively fast and easy	time-consuming; expensive
Impact	single score may mask an uneven writing profile and may be misleading for placement	more scales provide useful diagnostic information for placement and/or instruction; more useful for rater training
Authenticity	White (1995) argues that reading holistically is a more natural process than reading analytically	raters may read holistically and adjust analytic scores o match holistic impression
Interactiveness *	n/a	n/a

*Interactiveness, as defined by Bachman and Palmer, relates to the interaction between the test taker and the test. It may be that this interaction is influenced by the rating scale if the test taker knows how his/her writing will be evaluated; this is an empirical question.

[1] Weigle does not distinguish multiple trait scales from analytic scales.

3) Multiple-Trait Scoring

In the multiple-trait scoring, raters provide separate scores for different writing features as in the analytic scoring. However, the difference with the analytic scoring is that the writing features that are assessed are related to the specific assessment task. Multiple-trait scoring is based on the context for which the scoring is used, and is developed with a specific purpose of a specific writing context (Hamp-Lyons, 1991). Thus, it can be said that multiple-trait scoring treats writing as a multifaceted construct which is situated in particular contexts and purposes, so "scoring rubrics can address traits that do not occur in more general analytic scales" (Hyland, 2003, p. 230). It is very flexible as each task can be related to its own scale with scoring adjusted to the context, purposes of each genre.

Multiple-trait scoring can be an ideal compromise by teachers since it judges a text based on the writing features, while at the same time considering the specific writing task in the classroom. Therefore, it can provide rich data that can be used for remedial action and for course content. However, multiple-trait scoring requires enormous amount of time to devise and administer. One way of handling this can be to modify a basic "Content, Structure, Language" analytic template to the specific demands of each task. One more problem is that even though traits are task-specific, teachers may still depend on traditional general categories in their scoring rather than using genre-specific traits (Hyland, 2003).

Hamp-Lyons (1991) who suggested the multiple-trait scoring for the first time identified six advantages of multiple-trait instruments:

- Salience: features which can be assessed can be determined by different writing contexts whose focuses on writing qualities are different
- Reality and community: the scoring is based on the readers' compromise on the construct of what writing is.
- Reliability: multiple-trait scoring enhances the reliability of single composite number scores built from its components
- Validity: multiple-trait scoring satisfies the construct and content validity since it reflects the accurate measurement of the behavior which defines the construct, and also the traits in the multiple-trait scoring derive from concrete expectations in the specific writing context.
- Increased information: performance on different components of writing is assessed and reported.
- Backwash: the increased accuracy and the details of the information provided by the multiple-trait scoring can bring about the positive effect on teaching.

In the same respect, Hamp-Lyons and Henning (1991) claimed that multiple-trait scoring can be useful to obtain communicative writing profile which is "a description of the writer's demonstrated ability in writing on a set of text features, with the writer's level of competence reported separately for each text feature (p. 339)".

One of the examples of multiple-trait scoring rubrics is asTTle[2] Writing Scoring Rubric developed by Ministry of Education in New Zealand. It identified six major functions or genres, and it includes contextual features as well as linguistic features (Glasswell. et al., 2001).

2. Previous Studies

Vacc (1989) examined the concurrent validity of holistic scores and analytic scores. Four classroom teachers' holistic scores of texts written by low-ability male eighth graders were highly correlated with the analytic scores of the same texts. However, a regression of the analytic scoring features on the holistic scores showed that "quality and development of ideas" was the only analytic feature that accounted for a significant amount of variance in holistic scores for all teachers. This indicates that raters arrived at similar holistic scores through different writing features, which support the concern expressed by White (1984) that little agreement exists about the subskills that constitute writing.

In another comparison study of holistic and analytic rating scale types, Carr (2000) found that changing the composition rating scale would have a potential to fundamentally change the overall emphasis of the test even though the two scales were designed to measure the same construct. Carr (2000) concluded that test scores derived using the two rating scales are not comparable.

Also, Bacha (2001) found that the EFL program would benefit from more analytic measures after comparing the holistic and analytic scores of the same texts using ESL Composition Profile. It was claimed that although high inter- and intra-reliability coefficients were found, the holistic scoring revealed little about the performance of the students in the different components of the writing skill. This was proven by the fact that when the analytic scores were compared with each other, high significant differences among the different writing components were found, and this means that students performed significantly differently in the various aspects of the writing skill. EFL Students may have different proficiency levels in different writing components, therefore analytic scoring which provides feedback on different components can be more helpful

[2] The Assessment Tools for Teaching and Learning (asTTle)

to the students.

Finally, Nakamura (2002) compared holistic and analytic scoring methods in the assessment of writing using the Rasch model. He found that in holistic scoring one among three raters was not within the acceptable range while in analytic scoring all the raters were within the acceptable range. He concluded that analytic assessment with several items is strongly recommended to avoid risky idiosyncratic ratings and warned that it is very risky for one classroom teacher to judge students using a holistic rating scoring.

Recent approaches to assessing writing seek to "develop multiple-trait scoring instruments to fit a particular view or construct of what writing is in this context, and to reflect what it is important that writers can do" (Hamp-Lyons, 1991, p. 248). It also reflect the concept of genre and context, in other words, different genres have different social contexts where they are used (Glasswell. et al., 2001). Therefore, scoring rubrics for different genres should reflect different features of each genre.

In this regard, Brown et al. (2004) studied the reliability and validity of a New Zealand writing assessment scoring rubric which contains curriculum-based multiple-trait rating scales. They found pretty high reliability levels in terms of consensus, consistency and measurement in spite of a short rater training. Their conclusion was that genre-specific multiple-trait rubrics can be used in making instructional decisions. Meanwhile, Hamp-Lyons and Henning (1991) investigated the validity of a multiple-trait scoring procedure in contexts other than that for which it was developed. They claimed that the scoring method taken as a whole seemed to be highly reliable in composite assessment, appropriate for writings from different contexts; however, little psychometric support for reporting scores on five or seven components of writing was found. They suggested that the issue of transferring the existing multiple-trait scoring rubric to new contexts would be educational rather than statistical. Finally, Gearhart et al. (1995) developed a genre-specific narrative rubric in an attempt to combine the large-scale and classroom assessment perspectives, and found a reliability and validity of a newly developed narrative rubric.

III. METHODOLOGY

1. Participants

Five female graduate students majoring in English Education in S university participated in the study as raters. They were considered to form a homogeneous group based on their educational background, major and age. Students who don't have any

form of rating experience were chosen to avoid any influence of previous rating experiences because the present study aims to explore the accuracy of various scoring methods for the possible use of secondary school English classes where extensive rater training is very hard due to the teacher's burden of overwork. To minimize the effect of rater training, a brief rater training focused on the understanding of the use and application of descriptors in the scoring rubrics.

2. Materials and Procedure

16 argumentative essays written by Korean sophomore high school students were used in the present study. Essays were typed in order to avoid the effect of handwriting ability on the evaluation (Vaughan, 1991). One argumentative essay was used as a sample text in training raters, and the other essays were evaluated by raters.

As for the rubrics, TWE scoring guide for holistic, ESL Composition Profile for analytic, and multiple-trait scoring rubric for an argumentative essay which was designed by the researcher were used (Appendix 1). TWE scoring guide and ESL Composition Profile scoring rubric were chosen because these are the two most widely used rubrics in scoring essays. ESL Composition Profile has five categories whose scale weights are differentiated to emphasize different categories, but it was adjusted to 4-point scale to balance with the multiple-trait scoring rubric which has 4-point scale in each category.

At the beginning of each rating, raters were trained with one sample text and they discussed how to apply the criteria correctly. To reduce the impact from scoring the same writing three times, raters scored one writing in five days' distance.

3. Data Analysis

For a research question 1, one-way repeated ANOVA was used to see any mean difference between three scoring methods. For a research question 2, Pearson correlation coefficients and Chronbach's alpha coefficient were calculated by SPSS to see the rater reliabilities of the three scoring methods. For a research question 3, Pearson correlation coefficients, Chronbach's alpha coefficient and Friedman's Two-way ANOVA were used to examine the correlation level in the three scoring methods. For a research question 4, a questionnaire was used.

IV. RESULTS AND DISCUSSION

1. The Differences in the Mean Scores

Before the experiment, no significant difference among the three scoring methods was expected. However, one-way ANOVA result shows that there was a significant mean difference. Table 2 shows the raw scores and percent scores of each scoring method, and Table 3 presents the result of one-way ANOVA. Scores of each section were added as for the raw scores of analytic and multiple-trait scoring methods. Since the total scores were different in each scoring method, the raw score was changed into the percent score, and then one-way ANOVA was conducted.

TABLE 2
Averaged Mean Scores of Holistic, Analytic and Multiple-Trait

	Holistic		Analytic		Multiple-Trait	
	Mean	SD	Mean	SD	Mean	SD
Raw	3	0.92	13.97	2.69	11.27	2.56
Percent	50.00	15.35	69.87	13.44	70.42	16.02

*raw score - Holistic (6), Analytic (20, composite), Multiple-trait (16, composite)

TABLE 3
One-way ANOVA by Percent Mean Scores

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	4059.405	2	2029.702	14.132	.000
Within Groups	6032.095	42	143.621		
Total	10091.499	44			

As can be seen in Table 3, there was a significant difference in percent mean scores (F=14.132, p=.000). To find out where the difference came from, post-hoc test (Scheffe) was done and Table 4 shows the result.

TABLE 4
Post Hoc Tests (Scheffe)

(I) Method	(J) Method	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Holistic	Analytic	-19.86733(*)	4.37601	.000	-30.9723	-8.7624
	Multiple-Trait	-20.41733(*)	4.37601	.000	-31.5223	-9.3124
Analytic	Holistic	19.86733(*)	4.37601	.000	8.7624	30.9723
	Multiple-Trait	-.55000	4.37601	.992	-11.6550	10.5550
Multiple-Trait	Holistic	20.41733(*)	4.37601	.000	9.3124	31.5223
	Analytic	.55000	4.37601	.992	-10.5550	11.6550

* The mean difference is significant at the .05 level.

The difference came from the scores between holistic versus analytic and multiple-trait scores. The shaded area in the Table shows that the mean differences of holistic vs analytic, and holistic vs multiple-trait were -19.86733 (p=.000), -20.41733 (p=.000) respectively. This means that holistic scores were significantly lower than analytic and multiple-trait scores. It can be seen that raters gave lower scores by holistic scoring in rating the same texts. This is probably due to the fact that the holistic scoring rubric used in the present study, TWE scoring guide, was for evaluating college level texts, and the texts used in the present study were written by high school students. Some raters answered in the questionnaires that since TWE scoring guide is for TOEFL writing test, it was hard for them to give high scores to the average high school students' writings which are in the early developmental stage. It implies that the use of a widely-used holistic scoring guide can not guarantee the correct measurement of students' writings because it does not consider different students' levels in different academic levels. Also, this shows that a single scale can not provide information about EFL students' varying skills in writing ability.

2. The Inter-Rater Reliability of the Three Scoring Methods

The Cronbach's alpha coefficient was calculated for each scoring method to obtain the rater reliability, and Table 5 shows the result.

TABLE 5
Cronbach's Alpha Coefficient for Reliability

Holistic	Analytic	Multiple-Trait
.749	.861	.894

The reliability increased in the order of holistic, analytic and multiple-trait scoring method; holistic (.749) < analytic (.861) < multiple-trait (.894). The alpha coefficients of all the three scoring methods were acceptable above .7 which is generally considered to be a reliable level (Stemler, 2004). However, the multiple-trait scoring method showed the highest reliability which is .894. When holistic and analytic, and multiple-trait scoring were compared, the analytic and multiple-trait scoring method showed higher reliability than the single holistic scoring. This means that measuring with several categories rather than one single score can contribute to the increase of the reliability. On the other hand, in the comparison of analytic and multiple-trait scorings, multiple-trait scoring showed a little higher reliability than the analytic scoring. This indicates that the use of simplified context-reflective genre-specific rubric can make raters understand the criteria of the rubric more clearly.

One thing to note is that as Nakamura (2002) mentioned, the use of holistic scoring can be very risky when one teacher has to evaluate the writings independently. In the present study, one rater showed an idiosyncratic rating in the holistic scoring even though all the raters were equal in terms of the rater training and scoring procedure. Table 6 shows the Pearson correlation among the five raters between the three scoring methods.

TABLE 6
Pearson Correlations among Raters' Holistic, Analytic and Multiple-Trait Scores

		R2	R3	R4	R5
R1	Holistic	.714(**)	.562(*)	.517(*)	-.025
	Analytic	.393	.835(**)	.805(**)	.663(**)
	Multiple-Trait	.544(*)	.865(**)	.698(**)	.625(*)
R2	Holistic		.461	.626(*)	.271
	Analytic		.288	.588(*)	.219
	Multiple-Trait		.729(**)	.794(**)	.361
R3	Holistic			.863(**)	.014
	Analytic			.874(**)	.799(**)
	Multiple-trait			.807(**)	.733(**)
R4	Holistic				-.077
	Analytic				.801(**)
	Multiple-trait				.506

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

R5 showed no correlation with the other raters in the holistic scoring. Even R5 showed the negative correlation with R1 and R4 in the holistic scoring. However, in the analytic and multiple-trait scoring methods, R5 showed the correlation with three raters. The negative correlation was never expected because the raters were considered to be homogeneous and they were trained equally. R5 noted in the questionnaire that she was confused over how to apply the scoring criteria of the holistic scoring because the description of each scale was confusing and difficult to understand. Probably this caused the R5's discrepancy with other raters, and the use of analytic and multiple-trait scoring method reduced this discrepancy to a certain degree.

3. The Correlation among Components of Analytic and Multiple-trait Scoring Methods

To examine the comparability of multiple-trait scoring method with the analytic scoring method, first, components which are considered to measure the same construct in the analytic and multiple-trait scoring were compared by correlation and then the

Friedman test was conducted to find out the difficult level of each component in two scoring methods.

Table 7 shows the result of Pearson correlation of each component. The components of analytic scoring are listed in the left column, and components of multiple-trait scoring method are listed in the upper right column (AC- Analytic, Content / AO- Analytic, Organization / AV- Analytic, Vocabulary / AL- Analytic, Language use/ AM- Analytic, Mechanics / MC- Multiple-trait, Content / MO- Multiple-trait, Organization / ML- Multiple-trait, Language / MCO- Multiple-trait, Connection)

TABLE 7
Pearson Correlations among Components of Analytic, Multiple-Trait Scoring Rubrics

	MC	MO	ML	MCO
AC	.965(**)	.826(**)	.833(**)	.760(**)
AO	.866(**)	.862(**)	.730(**)	.793(**)
AV	.935(**)	.842(**)	.802(**)	.809(**)
AL	.677(**)	.500	.824(**)	.471
AM	.365	.263	.493	.578(*)

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

High correlation was found between AC/MC (.965**), AO/MO (.862**), AV, AL/ML (.802**, .824** respectively). The components of Content, Organization in the two scoring methods showed the high correlation with each other. In addition, Vocabulary and Language Use in the analytic scoring is combined into language in the multiple-trait scoring, and there was high correlation among these components. Therefore, the conclusion can be drawn from this finding that analytic and multiple-trait scoring can be said to be compatible with each other and further it can set the ground for claiming that multiple-trait scoring method can replace the analytic scoring method which does not reflect the context and specific writing task. The further advantages of multiple-trait scoring methods over the analytic one were found in the raters' answers from the questionnaires, and it will be dealt in the next section.

In order to find out the relative difficulty of each component, the Friedman Test was conducted. The result of mean rank comparison by Friedman Test is like the following[3]:

[3] The component with the highest difficulty came first.

- Analytic: language use – vocabulary – organization – content – mechanics (p=.000)
- Multiple-Trait: language – organization – content – connection (p=.000)

The same order of difficulty level was found among components of two scoring methods. As mentioned before, language use and vocabulary in the analytic scoring were merged together into language in the multiple-trait scoring. It was found that in both scoring methods, language section showed the highest difficulty. In other words, students received the lowest scores in this section. Organization came next, and students received fairly high scores in the content section than other sections. Mechanics in the analytic and connection in the multiple-trait scoring were the easiest ones to the students. However, these two sections measure the different aspect, thus can not be compared. This result indicates that the multiple-trait scoring can measure the writing construct with at least equal reliability with the analytic scoring. Considering the fact that the multiple-trait scoring showed the highest rater reliability, it can be said that the use of multiple-trait scoring method can improve the accuracy of scoring.

In conclusion, the multiple-trait scoring rubric developed by the researcher for evaluating Korean high school students' writings showed the similar results with, and sometimes high reliability than the ESL Composition Profile which was developed for evaluating ESL compositions. Also, it showed the comparability with the analytic scoring rubric.

4. Raters' Attitudes

After finishing each scoring, raters were required to answer the questionnaire asking how difficult or comfortable it was to rate according to the specific scoring rubric. Since the present study aims to find out which is the most accurate scoring method, it will be worthwhile to discuss how raters felt toward each scoring method. What raters said about the holistic and analytic scoring was similar with the general conceptions of each scoring method. It is summarized in Table 8[4].

[4] Answers were translated into English by the researcher.

TABLE 8
Raters' Attitudes Regarding Each Scoring Method

<table>
<tr><th colspan="2">Holistic</th></tr>
<tr><th>Strength</th><th>Weakness</th></tr>
<tr><td>- A text with rich content than with grammatical errors can get high scores.
- Since it is based on the overall impression, it is proper for evaluating the general degree of perfection and ranking texts.
- Raters are comfortable.
- Less time-consuming</td><td>- Since the description of the criteria is vague, it takes longer time to decide a point.
- Rater training is required since it depends on the rater's subjective decision
- It cannot provide feedback.
- Different students can get the same point though their strength and weakness in writing is different with each other.</td></tr>
<tr><th colspan="2">Raters' words</th></tr>
<tr><td colspan="2">- They were confused over the description of the criteria. They compared scores of the previous students in evaluating the text rather than following the rubric.
- They were confused when a rhetorical level and syntactic level of a text did not match. (e.g., when a text had a good content but low grammaticality, and a text had a good grammaticality but poor content)
- The holistic scoring is good for deciding Pass/Fail, but not good for dividing students into different levels.</td></tr>
<tr><th colspan="2">Analytic</th></tr>
<tr><th>Strength</th><th>Weakness</th></tr>
<tr><td>- Possible to find out the weaknesses of students' writing ability
- It increases the correctness of evaluating since a rater read a text several times.
- It can provide feedback.
- It is less dependent on the raters' subjective decision than the holistic since it has several</td><td>- The rubric had 4 points, so mostly raters gave middle scores.
- Time-consuming
- The rubric is only for evaluating.
- Possible not to see the whole
- Raters felt that they were doing the holistic scoring several times over and over.</td></tr>
</table>

<table>
<tr><td colspan="2">components</td></tr>
<tr><td colspan="2">Raters' words</td></tr>
<tr><td colspan="2">- Difficult to distinguish content and organization, especially in evaluating the short texts.
- Some raters considered content and organization the most important; therefore they decided other sections based on these two sections.
- Higher construct validity than holistic scoring.
- Necessary to include more scale levels, for example, use of the decimals.</td></tr>
<tr><td colspan="2">Multiple-Trait</td></tr>
<tr><td>Strength</td><td>Weakness</td></tr>
<tr><td>- The criteria description was simpler, thus easier to refer to.
- Proper to decide that a particular genre includes features of a particular genre well
- Proper for high school students' levels
- the rubric combined the strengths of holistic and analytic scoring well.
- More concrete and objective than the analytic scoring rubric
- It was fair to students since what they were taught were evaluated.</td><td>- The distinction of organization and connection was unclear.</td></tr>
<tr><td colspan="2">Raters' words</td></tr>
<tr><td colspan="2">Since the rubric is genre-specific, the descriptions of criteria were much easier to understand and apply.</td></tr>
</table>

Regarding the multiple-trait scoring method, raters said that it was more convenient for them to refer to when evaluating, since the description of the criteria was simpler and clearer than the analytic rubric. Also, they thought that the description is simpler than the analytic, but more concrete and objective. Therefore it can increase the reliability of scoring among raters. In addition, since the rubric is based, thus dependent on the specific genre and context, in this study, the argumentative essay writing of Korean high school students, it is more proper to evaluate students' writings since students' writings

were evaluated based on what they had learned to write in the classroom. Finally, raters said that the multiple-trait scoring rubric can be more appropriate to measure a particular genre of writing since the features to be measured reflect those of a particular genre.

In brief, all the raters said that they felt the multiple-trait scoring rubric was more convenient and clearer to use. In addition, the rubric was considered to be more objective and reflect the features of argumentative essays written by Korean high school students.

V. CONCLUSION

1. Major Findings and Pedagogical Implications

The present study aims to investigate the rater-reliability of the three scoring methods, holistic, analytic, and multiple-trait scoring methods. Scores of the five raters which were obtained from using the three scoring methods were compared. The major findings can be summarized as follows. First, significant mean differences were found between holistic, analytic, and multiple-trait scoring. Raters gave lower scores by the holistic scoring rubric than by the analytic and multiple-trait scoring rubrics. Second, the highest inter-rater reliability was found in the multiple-trait scoring. All the three scoring methods showed an acceptable level of reliability above .07. However, raters showed the highest reliability when they used a multiple-trait scoring rubric. Third, high correlation was found among components of analytic and multiple-trait scoring methods. This indicates that the multiple-trait scoring rubric can replace the analytic scoring rubric. Finally, raters expressed a favor over the multiple-trait scoring. They said that the multiple-trait scoring rubric reflects the features of an argumentative essay and characteristics of a writing context such as the writers (Korean high school students) and their level (novice).

The result of this study suggests some implications for writing assessment in Korean EFL English classes. First, the development of a proper rubric is necessary in assessing writing. Popham (1997) said that a rubric has three essential features: evaluative criteria, quality definitions, and a scoring strategy. In addition, a specific writing context should also be considered in developing a new rubric. In particular, Korea is a special context where English is a foreign language and the writing proficiency level of ordinary high school students is pretty low. The rubric which reflects this unique situation should be developed and used rather than just accepting the existing rubric which was developed mostly for assessing ESL college level compositions. Second, the more raters are involved the more the reliability increases. It is necessary to train English teachers for assessing secondary school students' writings properly. The special care is needed in

using the holistic scoring rubric when one English teacher has to evaluate all the students' writings. Finally, more attention to secondary school students' writings is needed. Investigating the development of students' writing ability in various genres will also enhance the development of different genre-specific rubrics.

2. Limitations and Suggestions for the Further Research

The limitations of this study suggest several helpful directions for further research on Korean EFL writing assessment in secondary schools. First, the present study did not investigate the intra-reliability of raters in using the three different scoring rubrics. The study of how raters evaluate the same texts after some time, using the three different scoring methods is worthwhile. Also, investigating of the rating process will be able to show more information about what is going on in the rater's mind (Connor-Linton, 1995). Second, raters evaluated texts in five day's distance in the present study. The study with longer terms will show the more exact nature of the change of reliability. Finally, an argumentative essay was used in the present study. The study of reliability in assessing different genres using a genre-specific multiple-trait scoring rubric will be able to tell how a multiple-trait scoring rubric can be used in the writing classroom.

References

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

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Barrs, M. (2004). Writing and Thinking. *www.teachingthinking.com*, Spring, 2004.

Breland, H. (1983). *The direct assessment of writing skill: a measurement review.* Technical Report No. 83-6. Princeton, NJ: College Entrance Examination Board.

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*, 105-121.

Carr, N. T. (2000). A Comparison of the Effects of Analytic and Holistic Rating Scale Types in the Context of Composition Tests. *Issues in Applied Linguistics, 11*(2), 207-241.

Connor-Linton, J. (1995). Looking Behind the Curtain: What Do L2 Composition Ratings Really Mean? *TESOL Quarterly, 29*, 762-765.

Gearhart, M., Herman, J. L., Novak, J. R., & Wolf, S. A. (1995). Toward the Instructional Utility of Large-Scale Writing Assessment: Validation of a New Narrative Rubric. *Assessing Writing, 2*(2), 207-242.

Glasswell, K., Parr. J., & Aikman. M. (2001). *Development of the asTTle Writing Assessment Rubrics for Scoring Extended Writing Task* (asTTle Tech. Rep. No. 6). Auckland, NZ: University of Auckland / Ministry of Education.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241-278). Norwood, NJ: Multilingual Matters.

Hamp-Lyons, L. (1995). Rating Nonnative Writing: The Trouble with Holistic Scoring. *TESOL Quarterly, 29*, 759-762.

Hamp-Lyons, L., & Henning, G. (1991). Communicative Writing Profiles: An Investigation of the Transferability of a Multiple-Trait Scoring Instrument Across ESL Writing Assessment Contexts. *Language Learning, 41*(3), 337-373.

Homburg, T. J. (1984). Holistic Evaluation of ESL Compositions: Can It Be Validated Objectively? *TESOL Quarterly, 18*(1), 87-107.

Hyland, K. (2003). *Second Language Writing.* Cambridge: Cambridge University Press.

Hyland, K. (2004). *Genre and second language writing.* Ann Arbor: The University of Michigan Press.

Jacobs, H. J. et al. (1981). *Testing ESL Compositions: A Practical Approach.* Rowley, MA: Newbury House

Johns, A. M. (2002). *Genre in the Classroom: Multiple Perspectives.* Mahwah: Lawrence Erlbaum Associates.

Kwon, O. R. (2006). *Developing English Production Skills Tests for Korean Students.* Paper presented at the 9th AFELTA Conference.

Kwon, O. R., Yoshida K., Watanabe Y., Negishi M., & Naganuma N. (2004). A Comparison of English Proficiency of Korean, Japanese and Chinese High School Students. *English Teaching*, *59*(4), 3-21.

Nakamura, Y. (2002). A comparison of holistic and analytic scoring methods in the assessment of writing. *The Interface Between Interlanguage, Pragmatics and Assessment: Proceedings of the JALT Testing Conference 2002.*

Popham, W. J. (1997). What's Wrong - and What's Right - with Rubrics. *Educational Leadership, 55*(2).

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *9*(4)

Vacc, N. N. (1989). Writing Evaluation: Examining Four Teachers' Holistic and Analytic Scores. *The Elementary School Journal*, *90*(1), 87-95.

Vaughan, C. (1991). 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: Multilingual Matters.

Weigle, S., C. (2002). *Assessing Writing.* Cambridge, UK: Cambridge University Press

White, E. M. (1984). Holisticism. *College Composition and Communication, 35,* 400-410.

Appendix 1. Multiple-Trait Scoring Rubric for an Argumentative Essay

	SCORE	CRITERIA
CONTENT	4	knowledgeable, substantive development of thesis, relevant to assigned topic
	3	sure knowledge of subject, adequate range, limited development of thesis, mostly relevant to topic but lacks detail
	2	limited knowledge of subject, little substance, inadequate development of topic
	1	does not show knowledge of subject, non-substantive, not pertinent
ORGANIZATION	4	introduction contains clear thesis statement, ideas clearly stated/supported, well-organized, logical sequencing, cohesive, well summing up
	3	introduction contains thesis statement, loosely organized but main ideas stand out, mostly supported, mostly logical sequencing, summing up
	2	introduction contains unclear and weak thesis statement, ideas confused or disconnected, lacks logical sequencing and development, weak summing up
	1	introduction contains no thesis statement, no organization, no logical sequencing, no summing up
LANGUAGE	4	excellent control of language, excellent use of vocabulary, excellent choice of grammar, appropriate tone and style
	3	good control of language, adequate vocabulary choices, varied choice of grammar, mainly appropriate tone
	2	inconsistent language control, lack of variety in choice of grammar and vocabulary, inconsistent tone and style
	1	little language control, reader seriously distracted by grammar errors, poor vocabulary and tone
CONNECTION	4	full control of connections
	3	generally adequately connected but occassionally awkward
	2	some awkward, missing connections
	1	connections usually missing or unsuccessful