Title	Construct Validity of Analytic Rating Scales Used in EFL Essay Writing Assessment : Reconsidering Components in Rhetorical Features
Sub Title	エッセイライティング分析的評価の修辞面構成要素に関する考察
Author	宮崎, 啓(MIYAZAKI, Kei)
Publisher	慶應義塾大学外国語教育研究センター
Publication year	2008
Jtitle	慶應義塾外国語教育研究 (Journal of foreign language education). Vol.5, (2008.) ,p.1- 22
JaLC DOI	
Abstract	When using analytic rating scales to assess EFL essay writing, each component of the scale is usually understood to reflect a single distinct rhetorical features (RF), such as Content or Organization. However, it is possible for a component to have traits in common with other components in the scale, making scoring of some of these components redundant and therefore inefficient. The present study examines the construct validity of analytic rating scales by investigating the interrelationships among the components associated with the rhetorical features: Content, Organization, Cohesion, and Voice. 70 essays, written by Keio High School students, were scored and analyzed in this study. Multiple regression analyses were performed to investigate the extent to which scores of Content can be predicted from the scores of the other RF. The findings indicate that Voice is the most significant feature contributing to the prediction of Content scores. The study also illustrates a high correlation between Organization and Cohesion. It is suggested that basing an analytic rating scale on two components alone — Content and Organization — is sufficient to provide an accurate and more efficient assessment of RF in EFL essay writing.
Notes	研究論文
Genre	Departmental Bulletin Paper
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=AA12043414-20080000-0001

Construct Validity of Analytic Rating Scales Used in EFL Essay Writing Assessment: Reconsidering Components in Rhetorical Features

Kei Miyazaki

Abstract

When using analytic rating scales to assess EFL essay writing, each component of the scale is usually understood to reflect a single distinct rhetorical features (RF), such as Content or Organization. However, it is possible for a component to have traits in common with other components in the scale, making scoring of some of these components redundant and therefore inefficient. The present study examines the construct validity of analytic rating scales by investigating the interrelationships among the components associated with the rhetorical features: Content, Organization, Cohesion, and Voice. 70 essays, written by Keio High School students, were scored and analyzed in this study. Multiple regression analyses were performed to investigate the extent to which scores of Content can be predicted from the scores of the other RF. The findings indicate that Voice is the most significant feature contributing to the prediction of Content scores. The study also illustrates a high correlation between Organization and Cohesion. It is suggested that basing an analytic rating scale on two components alone — Content and Organization — is sufficient to provide an accurate and more efficient assessment of RF in EFL essay writing.

1. Introduction

In assessing English as a Foreign Language (EFL) essay writing, analytic rating schemes have long been used to assess learners' writing abilities within a single modality. The analytic rating scheme separates and weights different features of the learners' performance on a writing task and assigns separate scores to each feature. The major advantage of analytic rating is to give learners feedback and provide useful diagnostic information about learners' writing abilities (H. Brown, 2004; Hamp-Lyons, 2003; Raimes, 1990; Weigle, 2002). On the

other hand, there are a number of problems associated with the criteria of the scoring scales. There is little assurance that each analytic scale is used properly without influence from criteria of other scales. The results of rating one scale may influence the rating of another (Cohen, 1994).

With regard to the criteria for evaluating writing skills using the analytic rating scale, some researchers within the language testing field have been paying much attention to the effect of raters on each trait of the rating scales (K. Brown, 2002; Elder et al., 2007; Lumley, 2002; Schoonen, 2005; Turner & Upshur, 2002). Much research has been conducted to investigate the potential bias or different interpretations of criteria among raters using the analytic scale.

However, relatively little research has been done on the interrelationships among the components of the analytic rating. Of the few studies that dealt with the components of the analytic rating, Astika (1993) investigated the extent to which each component of the essay writing contributed to the total score variance. The study found that Vocabulary accounted for the largest amount of variance in the total scores. Sawaki's (2007) study on speaking assessment, showed high inter-correlations between Vocabulary and Cohesion, indicating that the learners with good scores in Cohesion could have received high ratings in Vocabulary as well. The previous studies above treated all rating scale components including both rhetorical features (RF), such as Content or Organization, and linguistic features, such as Grammar or Mechanics (Weigle, 2002).

The present study focuses on RF and examines whether RF components are linked to each other. More specifically, it investigates the relationships between Content and other components: Organization, Cohesion, and Voice. The reason for this is that Content seems to have a broad definition which might involve the elements of other components. That is, there is a possibility that some components should be subsumed within Content. Therefore, this study aims to examine the degree to which RF components contribute to predicting Content scores, and then reconsider the analytic scoring scales which enable examiners to properly assess writing ability of EFL learners.

2. Literature Review

Among numerous analytic rating scales that have been used in assessing essay writing ability in EFL academic contexts, probably the most well-known and widely used scale is *ESL Composition Profile* by Jacobs et al. (1981). The Profile, which was one of the first attempts to develop an L2 analytic type of scale, is divided into five major writing

components: Content, Organization, Vocabulary, Language, and Mechanics. Each component in a level has clear descriptors of the writing proficiency required for that particular level as well as a numerical scale. Each of the five components in the Profile is weighted differently with Content receiving the most weight. It seems reasonable to suppose that the weightings of these components in the Profile reflect the importance of RF (see Appendix A). Another example is a scale which Schoonen (2005) used in his research on the estimation of variance components in the writing scores. In his study, RF components were integrated as a whole: "Content and Organization."

There are other scales which divide their RF into components such as Content, Organization, Cohesion, Coherence, and Voice (or Audience). *Test in English for Educational Purposes* (TEEP) by Weir (1990), in addition to Content and Organization, provides Cohesion as one of the RF. In a similar vein, *Michigan Merit Exam Persuasive Writing Scoring Guide* (Hamp-Lyons, 1997) emphasizes the need for assessing cohesion and logic. Regarding RF, the Guide sets up three components: Position, Complexity, and Organization. The first two components correspond to the elements of Content, Cohesion, and Coherence. This scoring guide was established based on Hamp-Lyons's (1991) study that demonstrated the significance of assessment for specific purposes.

Sasaki & Hirose (1999), Spandel & Calham (1993), and Witt (1995), in their research on analytic scale development, suggest that the element of reader awareness be incorporated into the components as one of the key factors measuring writing ability. This type of component is, in most cases, rated on such features as Voice or Audience.

AIMS Six Trait Analytic Writing Rubric Official Scoring Guide (2006) includes all sorts of RF components: Idea & Content, Organization, Voice, Cohesion, and Fluency (Coherence). This scoring guide is research-based, provides specific information about learners' performance, and is supported with classroom instructional activities. Moreover, the guide is designed to provide a consistent scoring method based on recognized characteristics of effective writing common to all genres.

These studies mentioned above show that analytic scoring scales require raters to judge the quality of learners' written language relative to such various types of component. That is, the analytic scores assigned to the learners' responses are assumed to reflect the underlying abilities being measured. However, there are two further points which need to be clarified. The first point concerns inter-rater reliability: the way a rater interprets the description of an analytic rating ability may affect scores. The second point concerns construct validity:

whether each analytic scale is a true reflection of the trait being measured can be questioned.

As for the first point, the past decade has witnessed a steady growth in research on inter-rater reliability on analytic rating. Shi (2001) made scoring comparison in raters with different language backgrounds, showing that both the native and non-native raters tend to use their own criteria value. Lumley (2002) and Schoonen (2005) investigated the process by which raters of ESL learners' essays make their scoring using an analytic rating scale. The results showed that the raters appeared to differ in the emphases they give to the various scale descriptors of each RF component. Kondo-Brown's (2002) study, in the field of JFL, also showed that raters were influenced by their own experiences as much as by the variation in quality of learners' essays. Similarly, Elder et al's (2007) study, comparing levels of rater agreement and bias in analytic rating, revealed limited overall gains in inter-rater reliability.

On the other hand, consistency of analytic ratings among raters can be seen in some other research. Bacha (2001) compared holistic rating scores with analytic scores by two raters. The result showed high correlation between two types of score as well as between two raters, indicating that a combination of holistic and analytic evaluation is needed to better evaluate learners' essay writing proficiency. One of Turner & Upshur's (2002) studies, on the process of analytic scale development, demonstrated that inter-rater agreement was high within empirically different scales and concluded that ratings of scale developers are not substantially influenced by their interpretations of the scale descriptors. The findings of this body of research highlight the need for rater training and more efficient analytic scoring development.

Regarding the second point: construct validity, little information is available. Nevertheless, several studies have attempted to explore the relationships among components in analytic rating. McNamara (1990) and Astika (1993) investigated the relationship of analytic rating scales to an overall score. McNamara's findings suggest that grammar scale plays the most important role in giving the overall score. In the case of Astika's study, the multiple regression analysis indicated that vocabulary scale profoundly influenced the total score. Sawaki (2007) examined the construct validity of analytic rating scales in terms of speaking ability. From the findings of her research showing that correlation between Vocabulary and Cohesion was high, she claims there can be overlap of constructs across the analytic scoring scales.

Some researchers propose that new RF components should be incorporated into the analytic scales. Rogers (2004) attempted to supply a definition of Coherence and investigated whether or not Coherence can be regarded as a single measurable component. The result of

comparing Coherence with overall writing quality indicated that it is possible for Coherence to be scored as an independent component. Sasaki & Hirose (1999), in their L1 Japanese analytic scale development, compared the *ESL Composition Profile* with their empirical scoring scales, asserting that a scale which includes "Reader's Awareness" (Voice or Audience) is more valid and reliable in assessing essay writing. Sano (2007) also suggests that the more rhetorical features are emphasized in the EFL setting, the more readers' awareness should be included in language production assessment.

Rogers, Sano, and Sasaki & Hirose take the position that a certain number of RF components should be assessed to provide detailed information about learners' rhetorical performance. Similarly, *AIMS* (2006) mentioned earlier, has no less than 5 RF traits in the analytic scale on the belief that "multiple traits allow a high score in one trait to compensate for a low score in another" (para. 3). However, as Cohen (1994) states, if the results of rating one component influence the rating another, some components should be eliminated or modified. Likewise, Weigle (2002) points out that in the general foreign language instructions on low and intermediate levels, it may be more appropriate to have separate components for linguistic features, but not appropriate for RF. Polio (2003) also describes that RF is more difficult to operationalize than linguistic features.

Difficulty in RF scale settings appeared in author's empirical knowledge. In fact, when the author was evaluating EFL learners' essays, the question arose as to the relevancy of Content to other components of RF. More specifically, the author thought it might be pointed out that the learners who received high scores on Organization, Cohesion, or Voice tended to receive high scores on Content. If elements of Content or any other component for that matter overlap with elements of other components, a composite score derived from the sum of scores in the scales assigned as independent features may lead to an inaccurate assessment of learners' writing abilities. In other words, it is possible that scales and criteria are sometimes improperly grouped. Indeed, as Polio (2003) states, Content is generally a matter of quality, and a kind of "holistic scale assessing the entire piece of writing" (p. 42). Furthermore, Shi (2001) reports that a large number of raters give weight to Content when assessing essay writing. In order to compensate for the shortcomings of analytic rating, some researchers (Bacha, 2001; Hughes, 2003) suggest that a combination of holistic and analytic evaluation is required to better evaluate learners' essay writing proficiency. However, due to time-consuming work and considerable burden, the reality is that teachers give feedback to learners only by calculating a sum total of scores for each component in an analytic scale.

Therefore, the construct validity of analytic rating scales should be further explored with a careful examination of the inter-relationship of rhetorical features.

3. Purpose of the Study

The purpose of the study is to investigate the interrelationships among the components associated with RF. By considering the previous studies and based on the author's experience in EFL essay writing evaluations, it appears highly probable that some RF component scores overlap with Content scores. Thus, it is assumed that some RF component scores can predict Content scores. Furthermore, although the previous studies mentioned above have affected the RF scale settings, the interrelationships among RF components have not yet been investigated experimentally. Thus, the present study was performed due to evidence showing the extent to which RF component scores are involved in Content scores. The following hypothesis and research question were constructed:

HS: Organization, Cohesion, Coherence, or Voice scores can predict the Content score in essay writing.

RQ: If the hypothesis is supported then, which component score — Organization, Cohesion, Coherence, or Voice — is the best predictor of the Content score?

4. Definition of Each Component

This research used analytic rating scales containing all common RF components. Content is defined as the degree to which "ideas and opinions are clear, complete and well developed; writing is relevant to the topic." Organization is defined as the degree to which "the structure suits the topic with a planned opening and closing, and supporting details that enrich the theme." Cohesion is defined as "transitions that tie the details together." Voice consists of three elements: "a clear sense of writing to be read, individual way of writing, and effective message involved in the topic." The above definitions were adopted from *AIMS Six Trait Analytic Writing Rubric Official Scoring Guide* (2006). This scoring guide was chosen primarily because it has all sorts of RF components and all 5 RF components are scored with equal weight. The score is based on the total of the individual component score. This type of analytic scale was suitable for use in the present research which would examine the interrelationship of the components. To save scoring time, the original form was slightly

revised (see Appendix B). Since the scoring guide mentioned above did not have clear definitions of Coherence, the author regarded Coherence as "making a series of sentences a connected set and linking all the meaning." This definition was cited from the research paper written by Rogers (2004), who conducted an empirical study to analyze written discourse according to the principles of coherence.

5. Method

5.1 Participants

The participants in the current study were 74 second-year learners at Keio High School in Japan. Most of them had studied English in Japan, yet a few of them had studied in countries such as the United States, Singapore, and China. The academic level of the school is relatively high, and all of the participants, evaluated by their records in the previous semester and by a common English test, were assigned to the intermediate level group. The common English test had been conducted in the final term exam period during the 2006 academic year. Due to the class size issues, these learners were divided into three English classes, with approximately 25 learners per class. To examine whether these three groups were equivalent in English language proficiency, a one-way analysis of variance (ANOVA) was conducted on their results from the common English test. The result of the ANOVA showed that the three groups were not statistically different in terms of their English language proficiency. ($F(2, 71) = 0.04$, $p = .96$ n.s.).

5.2 Materials

The instruments used in this study were an academic persuasive essay and an analytic scoring scheme. The prompt prepared for the essay task was one which intended to generate ideas and give the learners a starting point and direction for writing. The topic of the essay "clothing" was familiar to the learners and thus a relatively easy topic on which to write. The familiarity of the topic allowed learners to write more detail on the subject thus giving the study more in the way of useable data. The prompt and topic used in this study was adopted from Langan (2007), as follows:

> Do you agree or disagree with the following statement? "People behave differently when they wear different clothes." Do you agree that different clothes influence the way people behave? Use specific examples to support your opinion.

Analytic scoring scales used in the current study had five components: Content, Organization, Cohesion, Coherence, and Voice. All of these rating scales had five rating points representing: "Very Poor (1)," "Unsatisfactory(2)," "Moderate(3)," "Good(4)," and "Excellent(5)." Since this research focused only on RF, linguistic features such as Grammar, Mechanics, and Vocabulary were omitted.

In order to analyze the data and answer the research questions, SPSS 16.0 and Amos 7.0 were utilized. These commonly used statistical analysis software packages provide enough guidance for researchers to adapt the research design to general educational purpose (Storey, 2004).

5.3 Procedure

The persuasive essay writing test was conducted in February 2008. The learners were given a blank sheet of paper and 50 minutes in which to complete the assignment in the classroom. Eight months prior to this test, from May to early December 2007, the learners had been exposed to much formal instruction in fundamental essay writing by means of author-created handouts. They had been given sufficient opportunities to practice writing skills such as deciding on a title, using topic sentences, making paragraphs with indentations, organizing the essay, and including supporting details and specific examples. Moreover, the learners had been given the analytic scoring scale for their essays in order to understand the way of assessing their essay writing. Before this study was conducted in February, the learners had written seven persuasive essays, showing gradual improvement in rhetorical features.

The essay test in February required the learners to write approximately 200 – 250 words on the topic of "clothing." They were expected to write an essay that had an introduction, at least two supporting paragraphs, and a conclusion.

The learners were allowed to use their dictionary so that their abilities as they relate to RF could be demonstrated in the essay without any detrimental linguistic effects. The course syllabus of the English class told the learners that the essay writing would be 20 % of their course grade. Thus, the learners were all equally motivated.

Each essay was then collected, read, and scored independently by two raters using the same analytic rating scales mentioned earlier.

5.4 Scoring and Data analyses

The analyses for the essay writing were administered in the following several steps. First,

the number of words in each essay was counted in order to collect valid data in the study. Since the learners had been required to write at least 200 words in the essay, the essays comprising less than 200 words were eliminated. This resulted in 70 valid essays. These 70 essays were then scored independently by the author (rater A) and one experienced native TESOL teacher who works at a Japanese university (rater B).

Next, due to the fact that "rating on writing tests in academic contexts vary considerably" (Hamp-Lyons, 2003, p. 174), inter-rater reliability correlation coefficients were calculated using the Pearson correlation coefficient for each component of the essays. The results of the inter-rater reliability and *t*-test appear in Table 1.

Table 1. Inter-rater Reliability and *t*-test

	Rater A	Rater B		
	Mean	Mean	Correlation	*t*-value
Content	3.129	3.514	0.669**	1.946
Organization	3.80	3.820	0.733**	0.698
Cohesion	3.671	3.686	0.726**	0.829
Coherence	3.013	3.50	0.417**	3.750**
Voice	3.30	3.343	0.747**	0.516

**$p < .01$

Content, Organization, Cohesion, and Voice had relatively high agreements (r = 0.67, 0.73, 0.73, and 0.75 respectively) (J. Brown, 1996, p. 153). Coherence, however, showed less agreement (r = 0. 42). Since the low reliability was not suitable for this study, Coherence was removed. Additionally, since the mean values of both rater A's scores and rater B's scores would be used in descriptive statistics, a *t*-test was conducted to examine whether the scores by the two raters were equivalent. Before conducting the *t*-test, it was ensured that distribution of scores for the population was approximately normal. According to the t-test, there was no statistical difference between rater A's scores and rater B's scores in Content, Organization, Cohesion, and Voice. In other words, both raters were consistent in the scoring of these four categories.

Next, inter-correlation between all the components was checked to explore the relationships among components. At the same time, linearity between the components was examined. Additionally, Mahalanobis Distances were utilized in order to detect multivariate outliers. These two analyses were carried out to see whether the assumptions for the subsequent

regression analysis were met.

Finally, a multiple regression analysis was performed in order to see whether it would be possible to predict the Content score based on the other RF components. Content was set up as a dependent variable and the other components — Organization, Cohesion, and Voice — were placed as independent variables.

6. Results

6.1 Descriptive Statistics

Table 2 reports the descriptive statistics for the 70 learners for four RF components scored by two raters. As mentioned earlier, the scores were averaged across two raters. Skewness and Kurtosis were within ±2, suggesting that distributions were normal.

Table 2. Descriptive Statistics for Four Rhetorical Feature Components

	Mean	S.D.	Skew	Kurt	Min	Max	Total
Content	3.32	0.692	−0.441	0.019	1.50	4.50	70
Organization	3.81	0.781	−0.246	−0.224	2.00	5.00	70
Cohesion	3.68	0.692	−0.336	0.137	2.00	5.00	70
Voice	3.32	0.722	0.181	0.077	2.00	5.00	70

Note: Skew = Skewness; Kurt = Kurtosis; Min = Minimum; Max = Maximum (full score = 5.00)

6.2 Inter-correlations

Based on the descriptive statistics, inter-correlations between all the variables were examined to explore the relationships between variables. Results are shown in Table 3.

Table 3. Inter-correlations among Variables by Pearson Correlation Coefficients

	Content	Organization	Cohesion	Voice
Content	1.00	0.642**	0.522**	0.769**
Organization		1.00	0.827**	0.672**
Cohesion			1.00	0.543**
Voice				1.00

**$p < .01$

As can be seen in Table 3, the inter-correlation had a moderate margin, ranging from 0.52 to 0.83. Focusing on Content, the results indicate that there was considerably high correlation

between Content and Voice (r = 0.77), suggesting that a learner that obtained high score on Voice tended to obtain high score on Content as well. Although this result demonstrates the high inter-correlations among some scoring components, correlation coefficients do not allow one to systematically verify the extent to which score of Content can be predicted by other components. Thus, this issue was explored further by means of multiple regression analysis.

As a preliminary step, the linearity of the relationships between the dependent and independent variables was examined by scatter plot data. The plotted graphs are not shown here due to space constraints. The results show that the relationships between the dependent variable and independent variables were all adequately linear.

Furthermore, in order to detect multivariate outliers, Mahalanobis Distance for the four variables was calculated. The largest distance (8.83) was lower than the x^2value at four degrees of freedom (18.5). Therefore, it can be said that there were no cases in multivariate outliers (Molloy & Newfields, 2005).

Finally, it must be noted here that the correlation between Organization and Cohesion was extremely high. In such a case, the independent variables' contributions to the dependent variable overlap, and it is impossible to examine the contribution of each independent variable through the results of multiple regression analysis (Sasaki & Hirose, 1999). Thus, it was expected that either Organization or Cohesion, one of which would be unsuitable for use as a predictor in multiple regression analysis, would have to be excluded from the independent variable lists.

6.3 Regression Analyses

Following the administration of inter-correlation analysis, multiple regression analyses were performed based on the average scores of the two raters. Two types of multiple regression analysis were utilized: standard entry and stepwise selection. In the standard entry multiple regression, all of the independent variables were entered together into the regression equation model simultaneously (Table 4). According to the adjusted R-square, about 60 % of the variation in the dependent variable (Content) can be explained by the regression model with all the three independent variables ($F(3, 66) = 35.859$, $p<.01$). Of the three predictors, only Voice resulted in the highest standardized coefficient with statistically significant t-ratio. This result indicates that Organization and Cohesion were not appropriate for use in the regression model. It can be said that since the two predictor variables, Organization and Cohesion, are highly correlated (r = 0.827), Cohesion adds relatively little in prediction when Organization

is in the regression equation.

Table 4. Standard Entry Multiple Regression Summary for Three Variables Predicting Content

Variable	B	SEB	*β*	*t*	p	VIF	(Multico-index)
(Constant)	0.536	0.304		1.964	0.084		
Organization	0.204	0.136	0.231	1.505	0.137	4.071	} (26.502)
Cohesion	−0.003	0.135	−0.003	−0.023	0.982	3.164	
Voice	0.589	0.098	0.616	6.001	0.000**	1.827	

Note: n = 70; R^2= 0.620; ⊿R^2= 0.602; R^2change = 0.620; *D.W.* = 1.783 **p < .01

A further stepwise selection regression analysis was also performed in order to confirm whether Organization and Cohesion are superfluous variables in the regression analysis. The result was the construction of two models. Model 1 summary appears in Table 5 and Table 6.

Table 5. Model 1. Stepwise Multiple Regression Summary for One Variable Predicting Content

Variable	B	SEB	*β*	*t*	p	Partial Correlation	Part Correlation
(Constant)	0.876	0.252		3.473	0.001**		
Voice	0.736	0.074	0.769	9.919	0.000**	0.769	0.769

Note: n = 70; R^2= 0.591; ⊿R^2= 0.585; R^2change = 0.591; *D.W.* = 1.784 **p < .01 *p < .05

Table 6. Model 1. Stepwise Excluded Variables

Variable	*β*	*t*	p	Partial Correlation
Organization	0.228	2.239	0.028	0.264
Cohesion	0.148	1.620	0.110	0.194

Note: n = 70; R^2= 0.591; ⊿R^2= 0.585; *D.W.* = 1.784 **p < .01

The two predictors: Organization and Cohesion were excluded from Model 1. About 59 % of the variation in the dependent variable (Content) can be explained by the regression model 1 with one predictor (Voice). According to ANOVA, the regression model with one predictor was significantly related to the dependent variable ($F(1, 68) = 98.382$, $p<.01$). In order to examine the absolute values of the partial correlations for variables not in the equation, excluded variables were checked. The beta value associated with Organization is larger

(β=0.228), indicating Organization would make the greater contribution of the two excluded predictors. The partial correlation between Organization and Content is 0.264 after the effect of Voice was removed from both Organization and Content. The observed significance level associated with Organization is 0.028, which is significant at the 95% level (p<.05). On the other hand, the significance level with Cohesion is 0.110, which is not statistically significant (p>.05). Thus, it would seem likely that Organization is a more suitable second predictor for the equation.

In Model 2, only Cohesion was excluded. The results are shown in Table 7 and Table 8.

Table 7. Model 2. Stepwise Multiple Regression Summary for Two Variables Predicting Content

Variable	B	SEB	β	t	p	Partial Correlation	Part Correlation
(Constant)	0.593	0.276		2.152	0.035*		
Voice	0.590	0.097	0.616	6.050	0.000**	0.594	0.456
Organization	0.202	0.090	0.228	2.239	0.028*	0.246	0.169

Note: n = 70; R^2= 0.620; ΔR^2= 0.608; R^2change = 0.028; *D.W.* = 1.784 **p < .01 *p < .05

Table 8. Model 2. Stepwise Excluded Variable

Variable	β	t	p	Partial Correlation
Cohesion	−0.003	−0.023	0.982	−0.003

Note: n = 70; R^2= 0.620; ΔR^2= 0.608; *D.W.* = 1.784 **p < .01 *p < .05

Judging from the data, about 61 % of the variation of Content can be explained by the regression model with two predictors: Voice and Organization. According to R-square change, an additional 2.8 % of the variance of Content is contributed by Organization. Part correlation between Content and Organization after removing the effect of Voice from Organization is 0.246. Observed significance level associated with the excluded variable, Cohesion, is 0.982, which is too large to be accepted (p>.05). As a result, in the stepwise selection regression analysis, the best regression equation might be one that contains two predictor variables: Voice and Organization. However, since it was also debatable whether Organization and Cohesion have impact on Content, two additional regression analyses were performed between Content and Organization, and between Content and Cohesion respectively. The tables of the results are not shown here due to space constraints. As a result of the two regression analyses, adjusted R-square showed only 40 % and 26 % for each equation,

suggesting that Organization and Cohesion are not significant for the entry of the regression equations.

6.4 Path Analysis

The results clarify that Voice is the best predictor that can account for Content score, yet there is still not a quite clear evidence as to whether Organization and Cohesion make a strong impact on Content. The vagueness is caused by the fact that Organization was barely selected as a predictor in the second multiple regression model; on the other hand, the correlation between Organization and Cohesion is considerably high. This research is therefore needed, allowing for the rating scale constructions, to accurately verify the relationships among Content, Organization, and Cohesion. In order to explain the relationships among the three components, a path analysis was performed. In the path model (Figure 1), the rectangles represent observed variables and the circles represent measurement errors. The arrows and numbers indicate the degree of impact. The coefficients range from −1.00 to 1.00 in the standardized solution. As shown in Table 9, all the coefficients of the regression weights and correlations were significant, indicating that the path model fit the data properly. The path model demonstrates that there are theoretical directions where the impact of Cohesion on Organization (0.93) is stronger than that of Organization on Content (0.57).

Figure 1. Path Model.

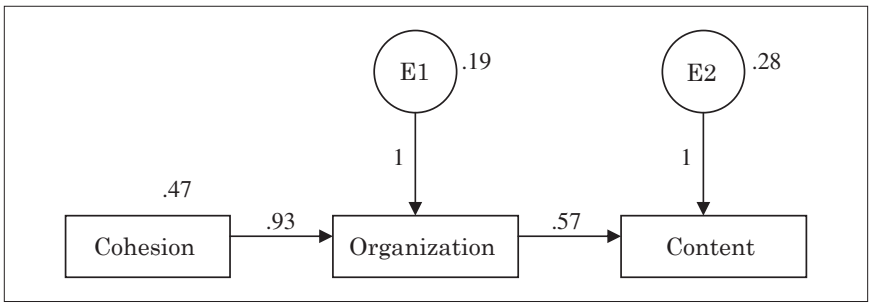

Table 9. Fit Statistics for Model

	x^2(df)	GFI	AGFI	RMR	RMSEA
Model 1	.029(1)	1.00	.998	.002	.000
Criteria	P > .05	> 0.9	> 0.9	< .05	< 0.1

7. Discussion

With the hypothesis in the current study, the research dealt with whether Content score can be predicted by other RF components. This relationship was examined through the correlation analysis and the multiple regression analyses. The result showed that the correlations between Content and each of the RF components were moderately high, particularly high between Content and Voice ($r = 0.769$). This result implied that some RF component scores might be good indicators of learners' Content scores. It then followed the multiple regression analyses. Judging from the adjusted R-Square, about 60 % of Content score can be explained by the regression models. Therefore, the hypothesis was supported.

A possible explanation for the observed outcomes may lie in the raters' scoring procedures and the traits that were rated. Such a fairly good fit to the regression model may be due to the fact that Content was the final scale the raters scored and Content may have been seen as the most important scale. The raters' comments suggest that the raters took a wider, more overall view on Content than on other components. More precisely, both the raters moved their attention from Organization or Cohesion to Content in ascending rank order of importance for each essay. Furthermore, raters' interpretations of the scale descriptors must have been a factor affecting the result. Rater B stated that the phrases such as "clear opinion" and "development of idea" in the Content descriptor can slightly urge the raters to score the essays holistically. Thus, as Cohen (1994) points out, it is probable that Content has several traits in common with other RF components.

Another suggested explanation for the result pertaining to the hypothesis may relate to the washback effects of the essay writing test on instruction. That is, the instructional emphasis underlying the author's classroom teaching of EFL writing might be reflected in the way the essay writing test would be scored. In the instructions, the learners were advised to start with taking their position and establishing a clear opinion before writing. The next step was to develop and fix the main idea in order to lead their readers through their organized thoughts on the topic. Basically, all these writing processes can be manifest in Content in analytic rating. However, the processes also include some elements of Organization and Voice. Under these writing instructions, which had been affected by the writing test evaluation, the learners believed these writing processes were the best ways to gain good scores. The learners might have been aware that Content is a part of a comprehensive guide to be rated.

As for the research question — "which of the RF components has unique contribution to predicting the score of Content?" — it was found, through the standard entry and the

stepwise regression models, that Voice is the most significant. Thus, it seems reasonable to conclude that Voice is the best predictor in examining learners' abilities in the area of Content. An explanation for this finding may be that the definition of Voice overlaps with a considerable part of Content. For instance, the expression "a clear sense of writing to be read" in Voice is closely related to the expression "ideas and opinions are clear" in Content. In the same way, "effective message involved in the topic" in Voice is indissolubly connected with "relevant to topic" in Content. Moreover, "depending on audience and purpose" in Voice and "well-suited to audience and purpose" in Content are almost the same point of assessment.

With respect to other RF components, Organization turned out to make considerably smaller contribution (2.8 %) to Content in the stepwise regression model 2. The additional regression analyses showed that Organization and Cohesion were not significant for the entry of the equations. It reveals that Organization and Cohesion provide only a little predictability that can account for Content score. Furthermore, the path analysis demonstrates that there is a stronger relationship between Organization and Cohesion than between Content and Organization. The interpretation of the finding is that the data from the regression models and the path model did not actually indicate a significant causal relationship between Content and Organization. In other words, it appears that Content ability is not substantially reflected in Organization scores. Although interpretations may be divergent, it is worth pointing out the fact that Organization and Cohesion were highly correlated with each other. As can be seen in Table 4, when Cohesion was entered into the regression model, Organization was not a significant variable. This occurred because these two components are too intimately related, as indicated in the multi-collinearity index (26.502 > 15.0). The Variance Inflation Factor value (VIF) also indicated that multi-collinearity would be a concern for Organization and Cohesion. This is not a surprising result because the raters must have scored Organization partly by checking transition words and because the instructions encouraged the learners to use many transition words to indicate their organized thoughts. Therefore, it seems reasonable to suppose that we should emphasize in the discussion the strong relationship between Organization and Cohesion rather than the distant relationship between Content and Organization.

The above issues concerning the relationships among RF components lead to several implications for developing analytic rating scales. Two suggestions for analytic scale construction emerge from the findings.

In the first place, RF components should be divided into two categories: "Content" and

"Organization." Content would be scored including Voice, and Organization would be scored including Cohesion. In this sense, *ESL Composition Profile* may have been providing the best possible set of indicators that can be used to assign scales to essay writing assessment. Concerning the descriptors in the components, it is suggested that a keyword of Voice, such as "well-suited to audience" should be added in the descriptor in Content. Moreover, "ideas clearly stated" in Organization should be written in the Content category. (See Appendix A).

Second, it also seems to be the case that we interpret the relationship between Content and Organization as a remote but slight connection. In this case, another suggestion is that the RF components should be combined as one scale: "Content and Organization." This suggestion is compatible with the analytic rating construction used in the research conducted by Schoonen (2005). He defines this component as "the propositional content of the text, as well as the ordering and coherence of the propositions and their illocution" (p. 9). This definition is totally general and broad, but it is reasonable to provide this definition as a RF component of scoring scales because the definition contains all the keywords to express the elements of Content, Organization, Cohesion, and Voice.

8. Conclusion

The present study showed that Content could be predicted by other RF components. More specifically, it was found that there was a strong relationship between Content and Voice, and little relationship between Content and Organization. However, an additional finding, which was not the original purpose of the study, revealed that Organization and Cohesion were highly correlated. Conjunctive links between Organization and Cohesion are also pointed out by Cohen (1994) and Lumley (2002). These empirical and theoretical studies imply that RF components should be divided into two categories: "Content" and "Organization."

We might not have established requirements for characterizing analytic rating that yields invariant scales producing consistent writing evaluation. Indeed, "decisions on which features of the texts are to be scored should be determined by the construct one wants to assess" (Schoonen, 2005, p. 18). However, in part, the traits are intertwined, dependent on one another regarding rhetorical features. When the composite score of each analytic scale reflects on learners' grades, the scales should be appropriately constructed and grouped. Especially within the classroom context, in most cases, one instructor is usually in charge of the rating and tends to grade learners on the basis of the sum total of the analytic scoring. Therefore, construct validity should be carefully considered when designing the elements of

essay writing assessment. In other words, the role of the appropriately grouped components and their descriptors would make the accurate and more efficient assessment.

Finally, it should be noted that there are at least two limitations in the present study. A major limitation is that only two raters were used to score the 74 essays. Bachmen (1995) and Carlsen (2003) suggest that inter-rater reliability can be improved by having more than two raters. Performance could be evaluated with more reliability when essays are scored by greater number of raters. Indeed, more research is needed to qualitatively examine the difference in raters' interpretations of scoring descriptors.

Another limitation is that Coherence in analytic rating scales was excluded from the analyses. In the present study, rater B gave higher marks in Coherence than rater A. This implies that rater B had a biased view between Coherence and other components, while rater A tended to draw a clear distinction between them. It may be assumed that Coherence bears some relevance to Content or Organization ability, as Rogers (2004) points out that "forming consistent topic strings within paragraphs contributes to the overall coherence of discourse"(p. 144). If Coherence had been included in the independent variables in the present study, the results of the analyses might have been different. It may be necessary to use other statistical devices, such as the structural equation modeling (SEM), in order to examine the interrelationships of RF component skills more accurately.

Acknowledgements

An earlier version of this paper was presented by the author at the 12th Japan Language Testing Association Annual Conference on September 14, 2008, at Tokiwa University. I would like to express my gratitude to Dax Thomas at Keio High School for proofreading and valuable comments. I am also thankful to John Pulasky at Tokyo Woman's Christian University for helping this research.

References

AIMS Six Trait Analytic Writing Rubric. (2006). U. S. Arizona Department of Education, 2006. Retrieved December 18, 2007, from http://www.ade.state.az.us/standards/AIMS/SampleTests 6Trait/

Astika, G. G. (1993). Analytic Assessments of Foreign Students' Writing. *RELC Journal*, 24 (1), 61 – 70.

Bacha, N. (2001). Writing Evaluation: What Can Analytic Versus Holistic Essay Scoring Tell Us? *SYSTEM*, 29 (3), 371 – 383.

Bachman, L. F. (1995). *Fundamental Constructions in Language Testing* (3rd ed.). Oxford, UK: Oxford University Press.

Brown, H. D. (2004). *Language Assessment: Principles and Classroom Practices.* NY: Longman.

Brown, J. D. (1996). *Testing in Language Program.* NJ: Prentice-Hall, Inc.

Carlsen, C. (2004). *Guarding the Guardians: Rating Scale and Rater Training Effects an Reliability and Validity of Scores of an Oral Test of Norwegian as a Second Language*. Bergen, Norge: Nordic Institute University.

Cohen, A. D. (1994). *Assessing Language Ability in the Classroom* (2nd ed.). Boston, Massachusetts: Heinle & Heinle Publishers.

Cumming, A. (2001). ESL/EFL Instructors' Practices for Writing Assessment: Specific Purposes or General Purposes? *Language Testing*, 18 (2), 207 – 224.

Elder, C., Barkhuizen, G., Knoch, U., & Randow, J. (2007). Evaluating Rater Responses to an Online Training Program for L2 Writing Assessment. *Language Testing*, 24 (1), 37 – 64.

Hamp-Lyons, L. (1991). Scoring Procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing Scond Language Writing in Academic Contexts* (pp. 241 - 276). Norwood: Ablex.

Hamp-Lyons, L. (1997). Second Language Writing Assessment Issues. In B. Kroll (Ed.), *Second Language Writing: Research Insights for the Classroom* (pp. 69 - 87). Cambridge, UK: Cambridge University Press.

Hamp-Lyons, L. (2003). Writing Teachers as Assessors of Writing. In B. Kroll (Ed.), *Exploring the Dynamics of Second Language Writing* (pp. 162 – 189). Cambridge, UK: Cambridge University Press.

Hughes, A. (2003). *Testing for Language Teachers* (2nd ed.). Cambridge, UK: Cambridge University Press.

Jacobs, H., Zingraf, S., Wormuth, D., Hartfiel, V., & Hughey, J. (1981). *Testing ESL Composition: A Practical Approach.*Roweley, MA: Newbury House.

Kondo-Brown, K. (2002) A FACETS Analysis of Rater Bias in Measuring Japanese Second Language Writing Performance. *Language Testing*, 19 (1), 3 – 31.

Langan, J. (2007). *College Writing Skills*. (7th ed.). McGraw-Hill, Inc.

Lumley, T. (2002). Assessment Criteria in a Large-scale Writing Test: What Do They Really Mean to the Rater? *Language Testing*, 19 (3), 246 – 276.

McNamara, T. (1990). Item Response Theory and the Validation of an ESP Test for Health Professionals.

Language Testing, 7, 52 – 75.

Molloy, H. P. L., & Newfield, T. (2005). Some Preliminary Thoughts on Statistics and Background Information on SPSS (Part 3). *JALT Testing & Evaluation SIG Newsletter*, 4 (2), 2 – 7.

Polio, C. (2003). Research on Second Language Writing: An Overview of What We Investigate and How. In B. Kroll (Ed.), *Exploring the Dynamics of Second Language Writing* (pp. 35 – 65). Cambridge, UK: Cambridge University Press.

Raimes, M. (1990). The TOEFL Test of Written English Causes for Concern. *TESOL Quarterly*, 24 (2), 427 – 442.

Rogers, S. H. (2004). Evaluating Texual Coherence: A Case Study of University Busoness Writing by EFL and Native English-Speaking Students in New Zealand. *RELC Journal*, 35 (2), 135 – 147.

Sano, F. (佐野　富士子, Ed.). (2007).「ライティング力をつけるための授業」『三省堂高校英語教育』2007, Summer, 2 – 5.

Sasaki, M. & Hirose, K. (1999). Development of an Analytic Rating Scale for Japanese L1 Writing. *Language Testing*, 16 (4), 457 – 478.

Sawaki, Y. (2007). Construct Validation of Analytic Rating Scales in a Speaking Assessment: Reporting a Score Profile and a Composite. *Language Testing*, 24 (3), 355 – 390.

Schoonen, R. (2005). Generalizability of Writing Scores: An Application of Stractural Equation Modeling. *Language Testing*, 22 (1), 1 – 30.

Shi, L. (2001). Native- and Nonnative- speaking EFL Teachers' Evaluation of Chinese Students' English Writing. *Language Testing*, 18 (3), 303 – 325.

Spandel, V. & Culham, R. (1993). Analytic Trait Scoring Guide. Portland, OR: Northwest Regional Educational Laboratory.

Spandel, V. & Culham, R. (1993). *Problems and Pitfalls Encountered by Raters*. Developed at the Northwest Regional Educational Laboratory for the Oregon Department of Education.

Storey, C. W. (2006). Doing Quantitative Research in Education with SPSS. *The Language Teacher*, 10, 36.

Turner, C. E. & Upshur, J. A. (2002). Rating Scales Derived from Student Samples: Effects of the Scale Maker and the Student Sample on Scale Content and Student Scores. *TESOL Quarterly*, 36 (1), 49 – 69.

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

Weir, C. J. (1999). *Communicative Language Testing*. London, UK: Prentice Hall International Ltd.

Witt, E. A. (1995). Issues in Constructing an Analytic Scoring Scale for a Writing Assessment. *Educational Resources Information Center ED387500*. Retrieved November 10, 2007, from http://eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/14/30/09.pdf

Appendix A

ESL Composition Profile Descriptors of the rhetorical feature traits (Jacobs, *et al*., 1981)

Content	30-27	EXCELLENT TO VERY GOOD knowledgeable, substantive, thorough development of thesis, relevant to assigned topic
	26-22	GOOD TO AVERAGE some knowledgeable of subject, adequate range, limited development of thesis, mostly relevant to assigned topic, but lacks details
	21-17	FAIR TO POOR Limited knowledgeable of subject, little substance, inadequate development of topic
	16-13	VERY POOR Does not show knowledge of subject, non-substantive, not pertinent, not enough to evaluate
Organization	20-18	EXCELLENT TO VERY GOOD fluent expression, ideas clearly stated / supported, succinct, well-organized, logical sequencing, cohesive
	26-22	GOOD TO AVERAGE somewhat choppy, loosely organized but main ideas stand out, limited support, logical but incomplete sequencing
	21-17	FAIR TO POOR Non-fluent, ideas confused or disconnected, lacks logical sequencing and development
	16-13	VERY POOR Does not communicate, no organization, not enough to evaluate
Vocabulary	20-7	point
Language Use	25-5	point
Mechanics	5-2	point

Appendix B

Revised Version of *AIMS* Analytic Writing Scoring Scale Descriptors of the Rhetorical Feature Traits

Content: Ideas and Opinions are clear, complete and well-developed; writing is relevant to the topic. One clear focus should be apparent but development and details should be thorough, balanced, and well-suited to the audience and purpose.
Excellent(5) Good(4) Moderate(3) Unsatisfactory(2) Very Poor(1)

Organization: A well-thought out order of ideas is apparent. The structure suits the topic with a planned opening and closing, and supporting details that enrich the theme.
Excellent(5) Good(4) Moderate(3) Unsatisfactory(2) Very Poor(1)

Cohesion: Transitions that tie the details together. Paragraphs are logically connected.
Excellent(5) Good(4) Moderate(3) Unsatisfactory(2) Very Poor(1)

Coherence: It makes a series of sentences a connected set and linking all the meaning.
Excellent(5) Good(4) Moderate(3) Unsatisfactory(2) Very Poor(1)

Voice: A clear sense of writing to be read, individual way of writing, and effective message involved in the topic. It should be appropriately written depending on the audience and purpose.
Excellent(5) Good(4) Moderate(3) Unsatisfactory(2) Very Poor(1)