Do analytic measures of content predict scores assigned for content in timed writing?

Rachael Ruegg and Yuko Sugiyama

Macquarie University and
Kanda University of International Studies

Abstract

In the instruction of second language writing, the emphasis has shifted over the last few decades from a focus on linguistic accuracy to a focus on the communication of ideas. Consequently, the focus of assessment criteria has undergone a corresponding shift. Nowadays, rating criteria for writing, whether they involve analytic rating scales or holistic rating, invariably evaluate the content of the writing being assessed. However, it is unclear what raters are sensitive to when rating writing for content. The qualities often sought are: the quantity of main ideas, the logical connection between the thesis statement and the main ideas, the use of examples to support the main ideas and the level of development of the main ideas.

The current study used multiple regression to ascertain what raters of the writing section of the Kanda English Proficiency Test (KEPT) were sensitive to when rating writing for content using analytic rating scales. The number, logical connection, support and development of the main ideas within 116 timed essays were evaluated in order to see what raters are more sensitive to when rating writing for content. It was found that none of the qualities evaluated were predictive of content scores. The two variables that were predictive of content scores were organisation scores and essay length. Implications of the findings will be discussed and suggestions for further research will be outlined.

Melbourne Papers in Language Testing 2010 Volume 15.1 pp. 70-91.
The Language Testing Research Centre, The University of Melbourne.

Introduction

Although for decades writing instructors did not go beyond grammatical correctness in their evaluation of writing, in the teaching of writing in recent years the focus has shifted and is now more on content. It is easy to see the advantages of focusing on content rather than form. Second language writers, especially in EFL contexts, tend to focus excessively on sentence level errors, while not placing enough emphasis on meaning (Tsui & Ng, 2000). This is particularly the case in countries with a traditional approach to language education, such as Japan, where most of the focus in the language classroom at the high school level is on accuracy of form. There are also some disadvantages that may not be so easily apparent. While grammatical correctness is easier to quantify when it comes to assessing writing, assessments of content are fundamentally more subjective. This issue not only affects those working in the area of testing but is also an important consideration for writing teachers when carrying out classroom assessment. Initially, it needs to be decided whether the number of ideas present or the quality of the ideas is valued more. What is meant by quality is also highly disputed. Currently, many rating criteria for writing take relevance of ideas, logical connection between ideas, support provided for ideas and development of ideas as the ideal qualities.

The current study examines the ratings given for content on a timed essay task. The content scale ignores the number of ideas and specifies connection, support and development of the ideas presented as the qualities to consider when assigning a score for content. These qualities are subjective and therefore raters may have difficulty quantifying them. The purpose of this study is to clarify which of the qualities predict higher content scores in practice.

Review of literature

The importance of content

In both analytic and holistic scoring in the assessment of academic writing, raters have become more concerned about how the ideas are communicated in an essay than the sentence level structure. Several studies have shown that the quality of the content of an essay is valued. For example, a study by Harris (1977) had teachers rank 12 papers, half of which were strong on content and organisation and half of which were strong on mechanics and sentence structure, based on five criteria and ranked the papers based on their perception of the overall quality. There was a tendency for teachers to value content and organisation more than sentence structure and mechanics although she also concluded that errors in mechanics and sentence structure may affect organization and content scores. In a questionnaire the teachers also stated that they valued content and organisation above sentence structure and mechanics, yet most of the comments the teachers made about the papers were on sentence structure and mechanics.

Freedman (1979) also conducted a study on raters' perceptions of a piece of academic writing in the L1 setting. Essays were rewritten to make them stronger or weaker in content, organisation, sentence structure and mechanics. The ratings of the rewritten essays were compared to the ratings of the original and it was found that the changes made in the content category significantly affected the holistic ratings. Breland and Jones (1984) correlated holistic scores from the College Board's English Composition Test and detailed comments made by the raters and also found that raters' judged the writing quality on the basis of content and organisation. Content is not only significant in essays written in L1, but also in ESL/EFL writing. A study by Vaughan (1991) examined recordings of raters commenting on six essays written by native speakers and non-native speakers of English. It was found that raters commented on the content of the essays most frequently. Another study (Connor & Carrell, 1993) analyzed think aloud protocols from writers who wrote

academic essays and from the raters who scored those essays. Results showed that both writers and raters were concerned about the language use, content and the development of the topic, but not much about organization. From these studies that used holistic ratings of essays, it can be assumed that "holistic raters are most influenced by the content and organization of a student's writing" (Huot, 1990, p. 207). However, this tendency can be applied to raters who use analytic rating scales as well. Lumley (2002) studied how raters understood analytic scales of writing assessment and how they applied the scales in their rating. He stated that "one significant feature missing from the Special Test of English Proficiency (STEP) scale criteria – but which clearly forms part of the construct for raters – relates to the content of the writing" (Lumley, 2002, p. 263), which implies that content was a significant factor for the raters.

The difficulty of assessing content

The raters in Freedman's (1979) study said that assessing content was the most difficult. Although raters find content to be important, assessing and measuring it is a challenge. In reality, the interpretation of content may vary for each rater. As Erdosy (2004) states, "constructs such as 'content' and 'organisation' have as many manifestations as there are raters" (p. 10). Often, raters complain "that the exact nature of the construct[s] they assess remains uncertain" (Cummings, Kantor & Powers, 2001, p. 3). Additionally, raters respond to and interpret the scales differently (MacIntyre, 2007), which also can affect the reliability and validity of the resulting scores (Lumley, 2002). Despite this, there are some common features that raters focus on when rating writing for content.

Defining content in writing

The definition of content varies between studies. First, think aloud research on the Kanda English Proficiency Test (KEPT) investigated whether the use of analytic rating scales is effective in writing assessment (MacIntyre, 2007). Six raters of the KEPT were asked to verbalize their thoughts while they were rating essays using the

analytic rating scales, which had a category for content. The raters mentioned variables related to content, such as the connection between ideas, as well as between the thesis statement and the main ideas, support or justification for the ideas as well as ideas being explained clearly, strongly and explicitly.

A similar study was conducted by Erdosy (2004) who gathered think-aloud protocols with four experienced raters of the TOEFL TWE and compared the ways they evaluated a written text using a holistic scale. Raters in this study came up with various aspects of content such as the development of ideas, argumentation, reasoning, logic and topic development.

Other studies have shown similar responses from raters or teachers in their studies. In Vaughan's (1991) study, the criteria for a highly rated essay included "a pattern of development [and] explanations or illustrations to support assertions" (p. 114). In Ballard and Clanchy's study (1991), comments from lecturers on 500 essays written in their respective disciplines were examined and classified. The comments from the lecturers who were assessing essays written by ESL undergraduates and graduate students were mostly "logicality" or "elegance of the argument", relevance, development of the argument and providing multiple perspectives (pp. 30-31). Another study that investigated twenty four scripts of think-aloud protocols of four experienced raters found three common features from the essays from the STEP test: relevance to the topic and clarity, quantity of ideas and cohesive devices, which are closely related to content (Lumley, 2002). Freedman's (1979) definition of strong content was an essay in which its content is relevant, developed, has logical connections and clarity. To conclude, the common features of the construct of content include the following: logical connection between the thesis statement and the main ideas, the use of examples to support the main ideas and the level of development of the main ideas.

The research question for the current study is: Are raters more sensitive to the number of ideas, the connection of ideas to the thesis

statement, support provided for ideas or development of ideas when rating writing for content using an analytic rating scale?

Methodology

The test

The writing samples used for the current study were 115 essays, written for the writing section of the Kanda English Proficiency Test (KEPT) in March 2009. The writing section of the test comprises a single prompt, about which examinees have 30 minutes to write an essay. Essays need to be a minimum of 80 words in length in order to receive a score. The essays are double rated using a set of four analytic rating scales and the ratings are scaled using Rasch modeling, to control for the comparative strictness and leniency of raters as well as the comparative difficulty of the rating scales. This study focused on the content scale, which rates the extent to which the test taker is able to support and develop his/her main ideas and the extent to which the ideas are logically connected to the main idea of the essay (see Appendix A).

The writers were 115 incoming freshman students, entering either the department of English or the department of International Languages and Cultures of a university of foreign studies in Japan. The writers' English language levels ranged from pre-intermediate to advanced. However, coming from the Japanese school system, most students have little, if any, experience of writing. Therefore, their writing ability is particularly weak.

The raters were 45 native or near-native English speaking lecturers or learning advisors at the university, holding Masters Degrees in TESOL, Linguistics or a related field. They were from: The USA (16), England (9), Australia (5), Canada (4), Japan (3), Scotland (2), New Zealand (2), Bulgaria, Ireland and Jamaica. One third of the raters were women (15) and two thirds were men (30). The raters take part in rater training either in January (for existing staff) or immediately before the administration of the test in March (for incoming staff).

The data

Through the literature, it was found that the aspects of academic essay writing most commonly associated with content were connection of ideas to the thesis statement, support provided for ideas and development of ideas. These are also the qualities mentioned in the rating scale for content. Therefore, the researchers evaluated and gave each essay a score for the logical connection of ideas to the thesis statement, the amount of support provided and the extent to which the ideas were developed by the writer. Anecdotal evidence from the KEPT test rater training sessions suggests that, although the quantity of ideas is not mentioned in the rating scale, many raters take this into account, seeing it as a fundamental part of the rating of content. For this reason, the number of main ideas was also included in the analysis in order to ascertain what role the number of main ideas plays in ratings for content.

In this study, the analysis was done collaboratively rather than independently as it was considered important that the figures could be agreed upon by both researchers, therefore no inter-rater reliabilities were calculated. (For a similar method see Ferris, 2006.)

Evaluation of connection

To determine a score for the logical connection of ideas to the thesis statement, first of all it was necessary to locate the thesis statement in each essay. After having done this, the researchers met to compare their findings. Essays were given a score on a three-point scale for their thesis statements. A score of 0 indicated that no thesis statement could be located by the researchers. A score of 0.5 indicated that a thesis statement could be located but the meaning of the thesis statement was unclear. An example of an unclear thesis statement is: "I don't think so."

A score of 1 indicated that the thesis statement could be both located and understood by the researchers. Initially, the raters had chosen different thesis statements on just a small minority of the essays and

ultimately, there was no disagreement about which sentence was intended as the thesis statement in each essay nor about whether or not they were clear. For the essays that received a score of 0 or 0.5, it was impossible to ascertain the logical connection between the thesis statement and the main ideas in the essay.

Next, the number of main ideas in each essay was counted and agreed upon by the researchers. For the essays that received a score of 1 for their thesis statements, subsequently each main idea was considered in terms of its logical connection with the thesis statement. An example of a main idea which is logically connected to the thesis statement is:

> "I think it is not good that people get married before 30 years old. If the young couple have a child, it is difficult to be grown up. Because also child, and they don't have ample experience to be grow up the child."

An example of a main idea that is not logically connected to the thesis statement is:

> "I disagree with this opinion…. Even if people who have enough wisdom, they can spend good life. So we have to learn many things in society, for example language, culture and more."

The overall score an essay received for connection constituted the proportion of main ideas that were logically connected with the thesis statement. For example; if an essay had 3 main ideas and 2 of them were logically connected to the thesis statement, the essay would receive a score of 0.67 indicating that two-thirds of the main ideas were logically connected to the thesis statement. The lowest resulting proportion was 0.33. Therefore, essays whose thesis statements could be located and understood but whose main ideas were entirely unrelated to the thesis statement were given the score of 0.22. Essays whose thesis statements could be located but not understood were given a score of 0.11. Essays whose thesis statements could not be

located remained with a score of 0. In this way, even if none of the main ideas related to the thesis statement, the writer got credit for the thesis statement itself.

Evaluation of support

To determine the score for support, the number of main ideas that were supported by examples was agreed upon by raters. A supported idea was defined as: *An idea that has specific examples to illustrate it.* In cases where the writer had used examples to support the counter argument but not the argument itself, a score of 0 was given as it was considered by the researchers that this would be likely to be detrimental to their content score. The following is an example of a supported main idea:

> "First, we can experience so many important things. For example, it's a part time job. During student life, many people do part-time job which tell us difficulties of relation of other people."

An unsupported idea would be an idea that did not provide any examples. An example is:

> "Second, they don't have enough knowledge how to living. They should learn society rules and need more experience."

As with the connection scores, the support scores constituted the proportion of main ideas that were supported by examples. For example, if an essay had 3 main ideas and 2 of them were supported by examples, the essay would receive a score of 0.67 indicating that two-thirds of the main ideas were supported. The lowest resulting proportion was 0.

Evaluation of development

In order to determine scores for the development of ideas, the number of main ideas that were developed was agreed upon by the researchers. The development scores also constituted a proportion.

An undeveloped main idea was defined as: *An idea that is simply stated and not explained at all.* An example of an undeveloped main idea is:

> "When have a child can teach my child. So this child very good people."

A developed main idea was defined as: *An idea that is stated and explained to the reader.* An example of such a developed idea is:

> "First, when we get married, maybe we have some children. However, most of the parents in their twenties still want to play with their friends then grow their children. So they give up their children and go out to play. I think babys whos parents such as young are very poor."

As with the connection scores and the support scores, the development scores constituted the proportion of main ideas that were developed. For example, if an essay had 4 main ideas and 2 of them were developed, the essay would receive a score of 0.5 indicating that half of the main ideas were developed. The lowest resulting proportion was 0.

Two example essays are included in appendices. Appendix B is the essay which was determined to have the lowest quality in terms of the content measures evaluated by the researchers. Appendix C is the essay which was determined to have the highest quality in terms of the content measures.

Analysis

In order to address the research question, a multiple regression analysis was performed with content scores as the dependant variable and length, organisation scores, number of main ideas, connection, support and development as the independent variables.

Anecdotal evidence suggests that ratings given for the organisation scale and the content scale in the KEPT are often the same and thus it

may be difficult for raters to distinguish between the quality of the organisation of ideas and the quality of the content itself. For this reason, the organisation scores were included in the analysis to see what role they play in the prediction of content scores.

Essay length may inflate writing test scores regardless of the quality of the essay. Weigle (2002) states that "A number of L1 studies have demonstrated that length…is a significant predictor of holistic scores" (p. 69). Moreover, a recent study has shown that essay length impacts both holistic and analytic scores (Lee, Gentile & Kantor, 2010). For this reason, length is often included in text analytic studies related to writing assessment. In the current study, the length of each essay was included in the analysis to ascertain the extent to which the quantity of writing predicts higher content scores. Since the length or the amount of text produced was a predictor of holistic scoring (i.e. Breland & Jones, 1984; Wiegle, 2002), the length was also taken into account. Length scores were determined by the number of words in the essay. The number of words in an essay is a common measure of essay length (e.g. Intaraprawat & Steffensen, 1995).

Results

The descriptive statistics in Table 1 show the mean, standard deviation and the number of instances of each variable that was included in the analyses. The possible scores for content and organisation ranged from 0 to 4. The actual scores for content ranged from 0.20 to 4.00, while those for organisation also ranged from 0.20 to 4.00. The lengths of the essays ranged from 80 words to 271 words. The number of main ideas ranged from 1 to 4. The connection scores, support scores and development scores, analysed by the researchers, all ranged from 0 to 1. All skewness and kurtosis measures fell between -2 and 2, thus the data can be determined to be sufficiently standardly distributed to employ regression analysis.

In addition, Pearson correlation was used to determine how strong the relationships between the variables were. There was a small but significant (at the 0.05 level) correlation between essay length and the

number of main ideas (Pearson r = 0.247), as well as between the connection and development of ideas (Pearson r = 0.208) and a small negative correlation between the number of main ideas and support (Pearson r = -0.236). There was a medium correlation (significant at the 0.05 level) between essay length and connection (Pearson r = 0.327), as well as essay length and development (Pearson r = 0.434). There was no significant correlation between any of the other variables and none of the variables had a correlation of more than 0.5, showing that none of the variables were strongly correlated.

Table 1. Descriptive Statistics

Variable	Mean	SD	Skewness	Kurtosis
Content	2.1687	0.99617	-0.018	-1.287
Organisation	2.1139	1.00985	0.259	-1.133
Main ideas	1.6261	0.74293	0.862	-0.192
Length	140.6140	50.67710	0.939	0.003
Connection	0.7527	0.37066	-1.004	-0.778
Support	0.4275	0.44816	0.310	-1.705
Development	0.4681	0.46720	0.132	-1.871

N=115

Multiple regression shows to what extent changes in each independent variable can predict changes in the dependent variable (content). The multiple regression results can be seen in Table 2.

Table 2. Multiple regression

Variable	B	Std. Error	Beta	t	Sig.
Constant	.033	.198	-	.168	.867
Organisation	.722	.067	.729	10.828	.000*
Main ideas	-.050	.070	-.038	-.718	.474
Length	.004	.001	.199	3.330	.001*
Connection	.063	.125	.024	.507	.613
Support	.041	.103	.019	.403	.688
Development	.112	.110	.053	1.025	.308

R^2=.792
Dependent Variable: Content
*Significant at the 0.05 level.

The R squared figure of 0.792 indicates that all of these variables together account for a large proportion of the variance in content scores. As can be seen from the results in Table 2, the two variables that predict variance in content scores are organisation scores and essay length. Increases in the level of connection of main ideas to the thesis statement, support provided for ideas and the development of ideas do not predict higher content scores.

Discussion and conclusion

Previous studies based their findings on data from think aloud protocols (i.e. Erdosy, 2004; Lumley, 2002; MacIntyre, 2007) or on written comments (i.e. Ballard & Clanchy, 1991). This current study attempted to quantify variables of content and further pinpoint

which variables influence analytical scores, but the results showed that it was mostly organisation scores that predicted content scores. In addition to organization scores, the only other variable which was found to predict content scores was essay length.

It is clear from the results of this study that the scores students receive for content on their timed writing represent a different construct of content than is intended by the analytic rating scale used for the rating. The rating scale specifies three qualities that should make up the content scores: logical connection between ideas, support provided for ideas and development of ideas. In the essays analysed for this research, none of these three qualities predicted content scores assigned. There are various possible reasons that these three qualities are not being evaluated by the raters when they assign a score for content.

It seems that there may be a fundamental connection between organisation and content which may prevent the two constructs from ever being completely distinct from each other. There was no particular mention in previous literature on the possible overlap between content and organization (i.e. Connor & Carrell, 1993; Vaughan, 1991) but anecdotal evidence from KEPT rater training sessions suggests that the line distinguishing the construct of organisation from that of content is difficult for many raters to define. This is despite the two rating scales being discussed distinctly during rater training sessions and raters generally rating organisation as the first of the four scales and content as the fourth (with vocabulary and grammar rated in between). The extent of this connection is difficult to ascertain clearly because examinees usually develop both skills in a parallel way. This connection means that it may be beneficial to either collapse these two scales into a single rating scale, or focus more on the distinction between the two constructs during the rater training procedure.

The wording of the rating scales for organisation and content may be somewhat confusing for raters. Specifically, the organisation scale mentions *connection between sentences* and *connection between ideas* while the content scale mentions *logical connection* and *ideas being connected.*

Two different concepts are supposed to be described in the different rating scales; the organisation scale is supposed to take semantic connection into consideration, while the content scale should encompass logical connection between the thesis statement and the main ideas and logical connection between the main ideas. However, the use of the same word in the two scales may be confusing for some raters because it may bring to mind the same concept. The finding shows clearly that it is necessary to change the wording in either the content or the organization scale in order to make a clearer distinction between content and organization.

Length, however, was found to predict content scores. Numerous L1 studies (i.e. Breland & Jones, 1984; Weigle, 2002) have shown that the length of an essay contributes to the holistic score assigned. A previous study (Ruegg, Fritz, & Holland, 2011) found that essay length did not predict variation in lexis scores on the KEPT test. In addition to this study, a study by Ruegg and Sugiyama (forthcoming) also found that essay length did not predict variation in organisation scores on the KEPT test. In the present study, on the other hand, it was found that essay length does predict variation in content scores on the KEPT test. This seems to be a justified result as it is likely that longer essays have better content whereas the same cannot be said of lexical quality or organisation. Indeed, essay length had a medium correlation with both connection and development, two of the three variables that were intended to make up the construct of content according to the scale.

It is important to consider that the KEPT test is administered to incoming first-year students as well as to first and second year students at the end of the academic year. The data for this study was collected from the March administration of the test, which is when the test is administered to incoming first-year students. Many of the essays were short and this made it difficult to analyse some of the essays in terms of some of the variables. Especially, the number of main ideas was sometimes difficult to ascertain as sometimes a main idea literally constituted one sentence whereas other times an entire paragraph was used, or several paragraphs. Essays at lower levels may be more difficult to rate for content than those at higher levels. These are the lowest level students

who take the KEPT test. Furthermore, many incoming first year students have never had any instruction in writing essays in English, whereas all other groups of students who take the KEPT test have had at least one year of instruction. Therefore, the findings may have been different if other students had been used.

Furthermore, all data used in this study was based on the evaluations of just two researchers. If a larger group of evaluators had participated in the discussion, it may have resulted in different evaluations and subsequently, different results. Nevertheless, if the ratings assigned for content do not represent the construct of content as defined by the analytic rating scale used in the rating, clearly this is cause for concern.

References

Ballard, B. & Clanchy, J. (1991). Assessment by misconception: Cultural influences and intellectual traditions. In L. Hamp-Lyons (Ed.), *Assessing in second language writing in academic contexts* (pp. 19-35). Norwood, NJ: Ablex.

Breland, H. M. & Jones, R. J. (1984). Perceptions of writing skills. *Written Communication*, *1*, 101-109.

Connor, U. & Carrell, P. (1993). The interpretation of tasks by writers and readers in holistically rated direct assessment of writing. In J. G. Carson & I. Leki (Eds.), *Reading in the composition classroom* (pp. 141-160). Boston: Heinle and Heinle.

Cummings, A., Kantor, R. & Powers, D. (2001). Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making, and development of a preliminary analytic framework. *TOEFL Monograph Series*. Princeton, NJ: Educational Testing Service.

Erdosy, M. U. (2004). Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions. *TOEFL Research Report 70.*

Ferris, D. (2006). Does error feedback help student writers? In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 81-104). New York: Cambridge University Press.

Freedman, S. (1979). How characteristics of student essays influence teachers' evaluations. *Journal of Educational Psychology, 71*(3), 328-338.

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

Huot, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication, 41*(3), 201-213.

Intaraprawat, P. & Steffensen, M. S. (1995). The use of metadiscourse in good and poor ESL essays. *Journal of Second Language Writing, 4*(3), 253-272.

Lee, Y., Gentile, C. & Kantor, R. (2010). Toward automated multi-trait scoring of essays: Investigating links among holistic, analytic, and text feature scores. *Applied Linguistics, 31*(3), 391-417.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246-276.

MacIntyre, R. (2007). Revision of a criterion-referenced rating scale used to assess academic writing. *Studies in Linguistics and Language Teaching*, *18,* 203-219.

Ruegg, R., Fritz, E. & Holland, J. (2011). Rater sensitivity to qualities of lexis in writing. *TESOL Quarterly, 45*(1), 63-80.

Ruegg, R. & Sugiyama, Y. (forthcoming). *Organisation of ideas in writing: What are raters sensitive to?*

Tsui, A. & Ng, M. (2000). Do Secondary L2 Writers Benefit from Peer Comments? *Journal of Second Language Writing*, *9*, 147-171.

Vaughan, D. (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: Ablex.

Weigle S. C. (2002). *Assessing writing*. New York: Cambridge University Press.

Appendix A

KEPT Essay Rating Scales

	Organisation Think about: • **Coherence** • **Structure**	**Lexis** Think about: • **Variety** • **Control**	**Grammar** Think about: • **Range** • **Accuracy**	**Content** Think about: • **Relevancy to the main idea** • **Ideas that are supported and developed**
0	No coherence or organisation, unconnected sentences which communicate little.	Demonstrates minimal word knowledge.	Phrases or sentences produced, but many inaccuracies make message/ writing difficult to understand.	A list of sentences with no logical connection and/or are irrelevant.
1	Some attempts to organise information but with little connection between ideas apparent.	A limited variety of vocabulary with little control.	Inadequate range of grammar used repetitively or inaccurately.	Ideas are connected but not relevant, developed or supported.
2	Obvious attempts to organize information though sometimes the	Uses an adequate variety of vocabulary with moderate	An adequate range of grammar used, with inaccuracies that impede the	Ideas are connected, relevant, but not supported or developed.

	lack of coherence creates ambiguity.	control.	understanding of sentences.	
3	The writing displays an organizational structure which enables the message to be followed although sometimes the lack of coherence might create ambiguity.	Uses a wide variety of vocabulary but there are inaccuracies in word choice and formation.	An adequate range of grammar but occasionally accuracy affects the understanding of sentences.	Ideas are connected and relevant. They are supported but the main idea is not developed.
4	The writing displays a coherent organizational structure which enables the message to be followed effortlessly.	Uses a wide variety of vocabulary, accurately and with control.	A wide range of grammar used accurately.	The ideas are relevant, well supported and developed.

Appendix B

Essay with the lowest overall content quality

I think one of the reasons that many people before the age of 30 don't get married is women's proceeding of society. In these days, many women have job and they like to work.

But if they got married and had children, they would have to quit their jobs. It is difficult for them to take a long vacation to raise children in Japan.

Now woman can live themselve.

But some people want to get married before the age of 30. My friend told me so. So I think when the peorson I want to marry existed I want to get married.

Appendix C

Essay with the highest overall content quality

I almost agree with this essay. We should not get married so early. I have two reasons to support my idea.

First, when we get married, maybe we have some children. However, most of the parents in their twenties still want to play with their friends then grow their children. So they give up their children and go out to play. In fact, I have seen many young couples for example stoped growing their child or killed them on the TV programs and nonfiction books. I think babys whos parents such as young are very poor.

Second, nowadays the people in twenties don't have a lot of money than before. In the past, many people get their jobs earlier then now such as 15 years old, but many young people in Japan these days, they don't get the job even they graduated from university. Even if they get the job, they couldn't have enough money to get married. So I think we should get more money before we get married.

In conclusion, we shouldn't get married in twenties and we have to consider about our future before marriage. This is my opinion.