FLUENCY IN ESL WRITING: LENGTH OF WRITING AND LENGTH OF T-UNIT AS MEASURES

SAERHIM OH

University of Hawai'i at Mānoa

This study investigated 67 writing samples from a placement test which were placed into two different levels. The purpose of the study was threefold. One was to see if there is a difference among the two levels in terms of fluency. Second, the study examined how much the fluency measures can predict students' placement in different levels. Third, it attempted to see what kind of relationship exists between the two measures used as gauge of fluency: the number of words and the mean length of T-unit. The analysis of the study reveals that there is a significant difference in fluency between the two levels, and that the percentage of the prediction of the fluency measures in students' placement differs for the two levels. Additionally the correlation between the two measures suggests that the number of words and the mean length of the T-unit both might not be measuring the same construct.

INTRODUCTION

The use of a placement test is a crucial matter for both English as Second Language (ESL) learning students and teachers (Bachman, 1990; Brown, 1996; Crusan, 2002; Hughes, 2003). For students, placement determines the amount of time and money spent in a given language program along with the content they learn while they are in the

placed class. Also for teachers, it influences the range of the proficiency of the students who are in their classes. Thus, it is important to investigate placement tests. Among the many components that are included in a placement test, this study focuses on the writing component.

Writing is especially important for ESL students in an English as an Academic Purpose (EAP) setting because many classes in universities require students to show their work through writing. In addition, assessment of writing is a considerable issue. To be specific, raters are not like computers, they do not internalize a predetermined frame that they apply to every essay in the same way. Also, different raters focus on different essay elements and have individual approaches to reading essays because rating essays is a very subjective matter (Carlson & Bridgeman, 1986; Vaughan, 1991). Accordingly, this study is mainly focused on the holistic assessment of writing, especially looking at the fluency measure and the rater's decision of placement.

Holistic assessment

Writings are rated and placed holistically in many language programs. Holistic assessments are used in almost 90 percent of the English departments across the country (Purnell, 1982; CCCC Committee on Testing, cited in White, 1984). It is effective for making selection or placement decisions as well as in a ranking or scoring procedure. However it may not be suitable in offering correction, feedback, or diagnosis information (Carlson et al., 1986; Charney, 1984). In Cooper (1977), holistic assessment is defined as "any procedure which stops short of enumerating linguistic, rhetorical, or informational features of a piece of writing. […] but the reader is never required to stop and count or

tally incidents of the feature" (p. 4). In other words, raters assign a single score based on the overall impression of the writing when using this method. A rating scale or a scoring rubric that provides a guideline of the scoring criteria is used in a typical holistic assessment (Weigle, 2002). For example, the Test of Written English (TWE) section in the Test of English as a Foreign Language (TOEFL) is assessed holistically according to a 6-scale rubric.

Weigle (2002) described the disadvantages and the advantages of holistic scoring comparing it to analytical scoring. One of the disadvantages is that holistic assessments are not easy to interpret because raters do not necessarily use the same criteria though they may have given the same score. An overall score of 5 could be given to different students due to different reasons. Also, because it is the overall impression that takes into account, the score may be influenced by superficial characters, such as handwriting, word choice, and spelling errors (Charney, 1984; Stewart & Grobe, 1979). In some cases, these features may not even be mentioned in the rubrics (Vaughan, 1991). In addition, a single score that a writing sample receives does not help raters distinguish various aspects of wiring, such as the control of syntax, depth of vocabulary, and organization. This is claimed by Weigle (2002) to be especially crucial for second language writers because different writers develop different aspects of writing abilities at different speed. Ellis (2005) suggested that an analytic scale designed carefully to assess the second language writing proficiency will be more useful than assessing the writings holistically.

However, there are reasons why holistic assessments are widely used. First, it is because they are more practical compared to analytical assessments (Weigle, 2002). When assessing writings holistically, raters spend less time reading each writing samples

compared to reading to assign scores for each category. Also, it is more authentic than analytical assessments because it reflects the personal reaction of a reader to the text, which analytical assessment does not. By assessing writings holistically, readers can pay attention to the strengths of the writing, so that writers are rewarded for what they do well. Furthermore, the scoring rubric can be designed to focus on certain aspects of writing (White, 1984, 1985).

As mentioned earlier, holistic assessment could be less reliable than analytical assessment (Breland, 1983; Ellis, 2005; Hamp-Lyons, 1991b; Huot, 1990). However, White (1984) claimed that the reliability of holistic assessment is still acceptable and that the following practices and procedures are important as a way to maintain high reliability. First, a scoring rubric should be used and it should explicitly describe the criteria that assist when assessing the writings. When training the raters beforehand, a writing sample should be shown with the rubrics to raters, in order to give them a taste of the points of the scoring guideline. Additionally, at least two raters should score each writings independently, with a third rater and they should do it at the same place and time. Lastly, there should be a reading leader who is responsible for checking the agreement of the raters, and all the records of the raters should be kept to keep track of the raters.

In Penny, Johnson, and Gordon (2000), rating with augmentation was examined to see its effect on reliability. When raters feel that the writing represents the description of the level in the rubric, they would assign that level. However, for instance, if raters think that a writing sample is a bit higher than the description of level 4 but not adequately a 5, they would assign a 4+. Also, if it seems like it is slightly lower than a 5 but not as low as a 4 or a 4+, it will be assigned a 5-. The results of Penny et al. (2000)

indicate that the use of rating augmentation can improve the inter-rater reliability of holistic assessments.

In order to investigate whether holistic evaluations can be validated objectively in assessing second language writings, Homburg (1984) studied if a reliable and valid subjective grading scheme exists and if this scheme relates to certain objective measures that is present in second language writings. He concludes that holistic evaluation of second language writings can be considered to be adequately reliable and valid with rater training to help them get familiar with the types of features present in the writings.

Fluency

Lennon (1990) distinguished the term fluency of spoken language into the broad sense and the narrow sense. Fluency in the broad sense covers the term for the overall oral proficiency. Being fluent in this case indicates that someone has a high oral proficiency and is in the highest point on the scale that measures oral proficiency. In the narrow sense, fluency is merely one of the components of the oral proficiency, such as "correctness, idiomaticness, relevance, appropriateness, pronunciation, and lexical range" (Lennon, 1990, p. 389). As for this specific term, fluency refers to the "native-like rapidity" (p.390). A fluent speaker according to the narrow definition has a speed of a native speaker whereas a less fluent speaker's speech is slow, stammering, and confusing.

Along with the length of the talk without pause, Fillmore (1979) verified three other criteria that are used to categorize fluent production: the coherence and the complexity of the speech, the appropriateness of the talk, and the creativity of the language use.

As for fluency in writing, Polio (2001) stated that, one way to define it is examining how native like the writing sounds. The other way is looking at the amount of production in a writing sample. Tarone, Downing, Cohen, Gillette, Murie, Dailey (1993) used a holistic scale to compare Southeast Asian-American immigrant children's English writing skills according to their level, and also with international students, and native-speaking undergraduate students. In the holistic scale, fluency is referred to "nativeness, standardness, length, ease of reading, idomaticity" (p.170). This kind of holistic scale is an example of the first way of defining fluency. In order to measure the amount of production in writings, number of words, clauses, and T-units are counted, as well as clauses per T-unit, average length of T-unit, and type-token ratio, which are more commonly used for complexity or lexical quality, are also examined (Polio, 2001).

Larsen-Freeman (1978) looked at the average number of words per composition of EFL students. The study showed that there is an increase in the number of words per composition as the group level goes up. It is stated that this may be because of their fluency, their expressiveness or their increased self-confidence of their ability. However, the length of the composition drops for the group with the highest proficiency. The researcher explained that this is because the measures dependent on length are less discriminatory at the upper levels of proficiency than the lower.

Using the think-aloud method, Kaufer, Hayes, & Flower (1986) observed graduate and undergraduate students' writing looking at the amount of production, also. They concluded that the experienced writers proposed longer length of burst than less experienced writers.

Wolfe-Quintero, Inagaki, & Kim (1998) however, claimed that the frequency measures such as the number of words mentioned above are not a valid measure of fluency. Instead they suggested fluency ratios, such as words per minute, words per clause, words per sentence, and words per T-unit, are better measures.

In contrast to Wolfe-Quintero et al.'s view, Polio (2001) questioned the relationship between how quickly writers can write (words per minute) and the quality of the writing. It is stated that there might not be a relationship between them at all, or that it could even be negative. Also, it is shown in many first language acquisition literatures that length of T-unit is a good measure of complexity (Hunt, 1965; Loban, 1976 cited in Larsen-Freeman, 1978; Mellon, 1976 cited in Larsen-Freeman, 1978). Additionally, Ortega (2003) used words per sentence, words per clause, and words per sentence as measures of syntactic complexity and found meaningful relationship with proficiency. In the same vein, Henry (1996) used the length of t-unit as to measure syntactic complexity of Russian, while the length of the essay is used to measure fluency.

Chenoweth & Hayes (2001) emphasize the importance of fluency especially for second language learners and state that being less fluent than others, as second language learners are, can be a serious barrier to educational achievement. When students need to write a term paper that is due shortly, or when they are taking a writing test and have to write it for a limited amount of time, it would be beneficial if they could write quickly without hesitance.

In addition to emphasizing fluency as a term of writing quickly during a certain given time, in many rubrics of writing tests, fluency is one of the components that are looked at (Tarone et al., 1993; Jacob et al., 1981; Wesche, 1987).

In this study, fluency is examined in writing samples assigned to two different levels. Number of words is counted to see the amount of production of the writing. In addition, the length of t-unit will also be examined and compared with the overall length of the writing and with the placement level to test the claim by Wolfe-Quintero et al.

Research Questions

This study focuses answering the following question:

1) What is the relationship between EAP learners' fluency on a writing placement exam?
2) How much of the placement decision can be predicted by the fluency variables: the overall length of the essay and the length of T-unit?
3) What is the relationship between the length of the writings and the mean length of the T-unit?

METHOD

Placement Test

The placement test that is investigated in this study is the Mānoa Writing Placement Test administered in the beginning of the Fall semester in 2005. It is a 3 hour test which all the undergraduate students admitted to University of Hawai‘i at Mānoa has to take. There are three prompts that students can choose to write about. One is about

drug testing (Prompt 1), another about whether sports are for everyone (Prompt 2), and the last about elitism (Prompt 3). The direction in the writing prompts asks the test takers to decide to agree or disagree with the article dealing with one of the three topics mentioned above, and write an essay describing and defending the position with supports.

The international and immigrant students take this test in order to be placed in the two courses (73 and 100) in the English Language Institute (ELI) or get exempted from ELI and placed in ENG100 offered by the English department.

ELI 73 is the intermediate writing class. Writing samples are placed into this level when it "needs to develop L2 proficiency; [has] notable unfamiliarity with and general lack of control of academic writing; would benefit from at least two semesters of ELI writing instruction" (Writing Hallmark, 2002, see Appendix A). Writings that are placed into ELI 100 are placed in the advanced writing class. According to the Hallmark, they "show some knowledge and control of academic writing; needs to develop L2 proficiency, writing ability, awareness of genres/conventions common in U.S. universities; will benefit from ELI rather than English department instruction". Finally the ENG 100 is an expository writing class, which the non-international students are mainly placed into. The writing placed in this course "shows high proficiency in L2, but need for instruction in rhetoric, organization, support, and argumentation; will benefit from English department rather than ELI writing instruction".

Writing Samples

The data for this study are 67 writing samples written for the placement test in order to get placed into the ELI. These samples were obtained with the approval of the

ELI to disclose the test scores and the rating of the test. The number of samples for each prompt and level is displayed in Table 1.

Table 1. *Number of writing samples for each prompt and level*

	Prompt 1	**Prompt 2**	**Prompt 3**	**Total**
ELI 73	11	12	4	27
ELI 100	20	12	8	40
Total	31	24	12	67

The writings are rated holistically according to the Hallmark developed by the graduate assistants in Fall 2002. The Hallmarks are used as a guideline to help the raters' judgments. They are divided into the categories of Content, Organization, Vocabulary, Grammar, and Fluency. Though it describes these different categories, raters are encouraged to make the decision of the course placement with the overall qualities of the paper in mind.

Participants

The students who wrote the 67 writing samples are new coming international or immigrant undergraduate students who are admitted to the university and who do not speak English as their native language.

Procedures

Though the international students and the students who speak English as their native language take the same Mānoa Writing Placement Test, the test is administered by different departments, thus are also graded separately. The writings of students who speak English as the first language are graded by the English department, and those of the international and immigrant students are graded by the ELI.

These ELI writings are rated by the graduate assistants who teach in the ELI. They use the Hallmark as a guide and are shown sample essays for each level at the beginning of the training section. Each writing samples are read by three different raters, not necessarily by the same three raters. These three different raters read each writings and decide the class to place the student. When there are disagreements on the placement, they discuss the reasons why they think the writing should be placed in that certain level. If the three cannot agree on a level, a fourth reader reads the sample and gives another opinion to it. In addition, there is a leader responsible for checking the agreement of all the readers and he or she keeps the record of the decisions made by the three raters.

Though it is not required, raters often put a plus (+) or a minus (-) sign after the level when they think that the writing seems to be a little higher or lower than the given description of the levels.

Thus, based on the views of reliability of Homburg (1984), Penny et al. (2000), and White (1984), the holistic assessment of the placement test has the potential to be reliable.

Analyses

The researcher examined the 67 writing samples that were placed in the two levels according to the procedure described above. In this research, because the placement test was a paper based test, the number of words and the mean length of T-unit were manually counted twice by the researcher. For the number of words, there was 92.5% agreement and for the length of T-unit, there was 89.56% of agreement of the two times that were counted. When the calculations were different, they were counted again to make sure of the result. In order to see the difference between the two levels based on the number of words and the mean length of the T-unit, an independent t-test was conducted. Also, to see how much of the placement decision can be predicted by the fluency variables, a discriminant analysis was conducted. Last, a Spearman rank order correlation was conducted using the ranking order of the measures, instead of a Pearson correlation because of the big difference of the scale of the two measures.

RESULTS

Research Question 1: What is the relationship between EAP learners' fluency on a writing placement exam?

Number of words. The frequency of the number of words in level 73 and level 100 is plotted in Figures 1 and 2, respectively. The skewness of level 73 is 1.39, and that of level 100 is 1.66, which are both a little positively skewed. The measures of central tendency of level 73 in Table 2 has the distribution of a typical positively skewed

distribution (Mode<Median<Mean). However, for level 100, it has the order of Mean < Median = Mode. This is because it is almost a bimodal distribution as shown in Figure 2.

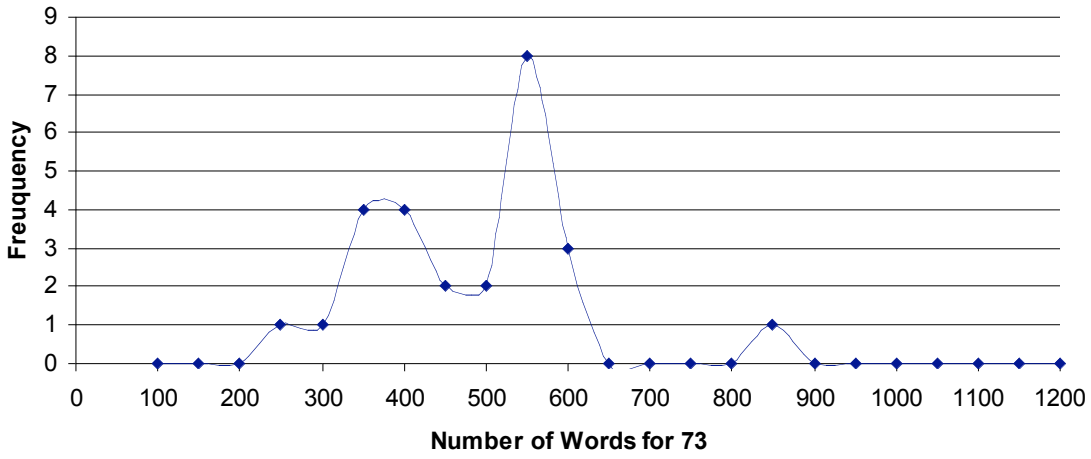


Figure 1. Frequency distribution of number of words in a writings in level 73

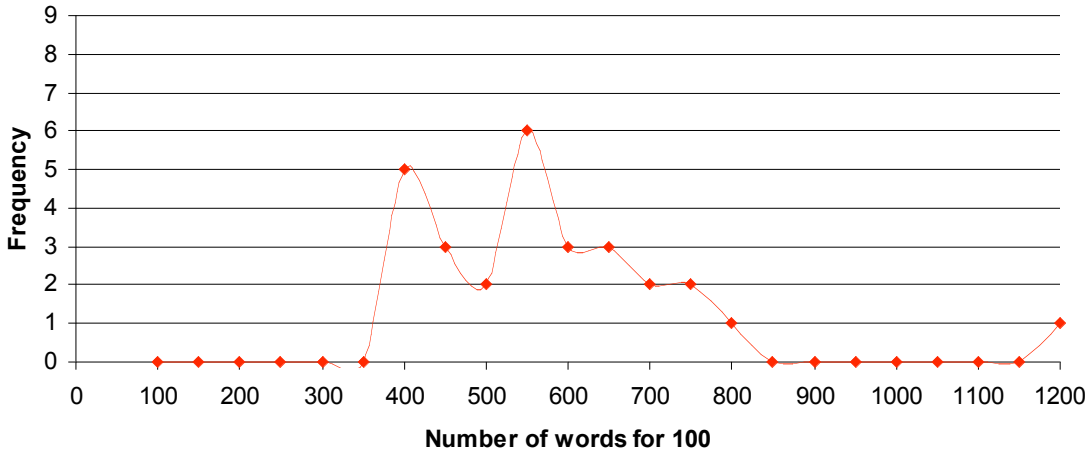


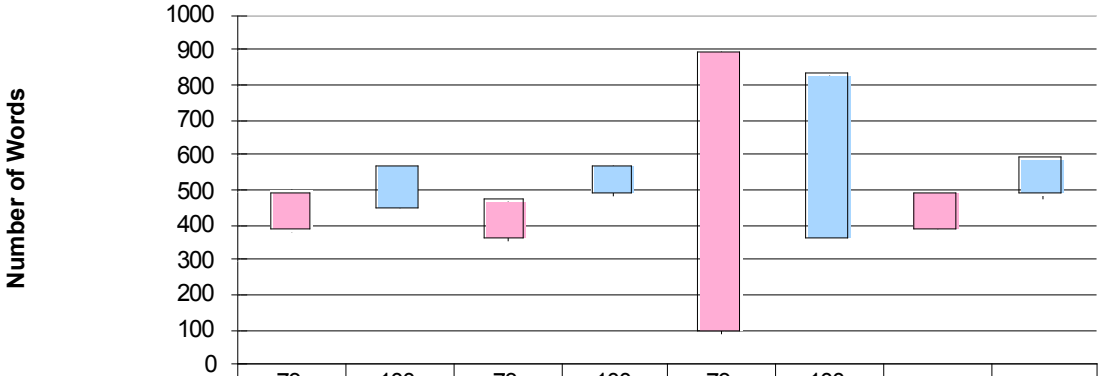
Figure 2. Frequency distribution of number of words in writings in level 100

Table 2. *Descriptive statistics for number of words in writing samples*

Statistic	ELI 73				ELI 100			
	Prompt 1	**Prompt 2**	**Prompt 3**	**Total**	**Prompt 1**	**Prompt 2**	**Prompt 3**	**Total**
N	11	12	4	27	20	12	8	40
Mean	437.18	411.58	489.75	433.59	502.6	526	597.75	528.65
Median	423	398	436	421	531	539.5	500	531
Mode	N/A	N/A	N/A	375	531	N/A	N/A	531
SD	88.25	90.11	252.14	120.33	107.31	138.73	272.07	160.17
Min	314	308	254	254	216	353	383	216
Max	605	580	833	833	654	702	1186	1186

Table 2 also shows the descriptive statistics for number of words as the measure for the three different prompts of each level. The N size for prompts one and two are similar but that of prompt three is extremely small. At a glance, it seems like the writings in the two levels are different in terms of the number of words. As a matter of fact, the difference between the writing samples in ELI 73 and ELI 100 according to the average overall length of the writing was found to be statistically significant, $t(65) = -2.622$, $p=.011$.

Interestingly, the shortest writing sample (number of words = 216) as well as the longest sample (number of words = 1186) were both placed in the upper level. Thus, this may be revealing that the length of writing is not a big matter for the placement. This will be discussed in more detailed while answering the second research question.



	73 prompt1 N=11	100 prompt1 N=20	73 prompt2 N=12	100 prompt2 N=12	73 prompt3 N=4	100 prompt3 N=8	73 N=27	100 N=40
CI upper (95%)	495.76	552.82	468.84	566.91	890.91	825.28	481.21	579.84
mean	437.18	502.6	411.58	526	489.75	597.75	433.59	528.65
CI lower (95%)	378.61	452.38	354.33	485.09	88.59	370.22	385.97	477.46

Figure 3. 95% confidence intervals for number of words in writings in different levels and those with different prompts

Figure 3 displays the upper and lower 95% confidence boundaries of the mean length of the writing samples for the two levels and those levels with three different writing prompts. As the confidence intervals with the same colors present, the average length of the writing for the three prompts overlap with each other at the 95% confidence level for both 73 and 100. Thus, although 41% of the students wrote about drug testing, 44% about sports, and 15% about elitism in the 73 level, and 50%, 30%, and 20% respectively in the 100 level, there was no significant difference among the three prompts in each level. This could be the case because the students had the choice of which prompt they would like to write about. The result could have been different if they were assigned to write about different topics.

Length of T-unit. Frequency of the mean length of T-unit for each level is distributed in Figures 4 and 5. The skewness for level 73 and 100 are 0.55 and 0.46, respectively. Thus, the distributions of the mean length of both levels are fairly normally distributed.

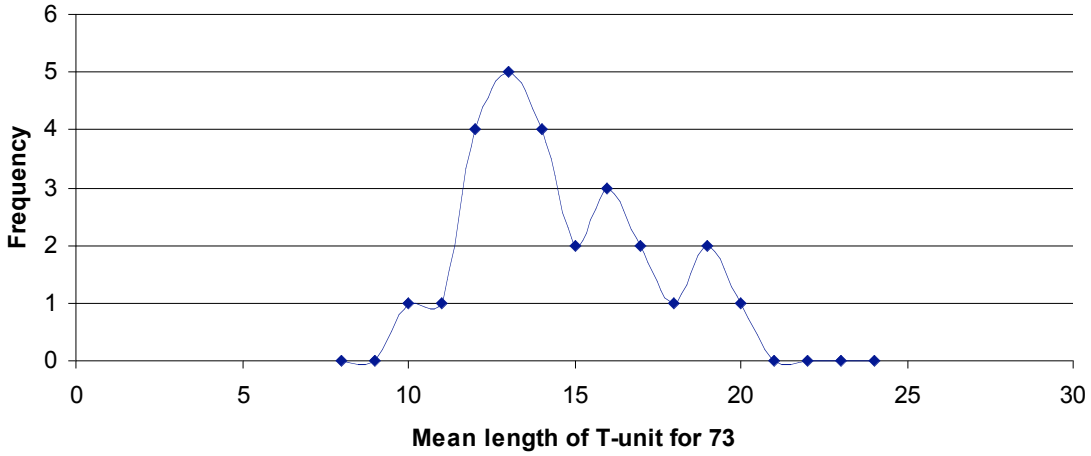


Figure 4. Frequency of mean length of T-unit in writings in level 73

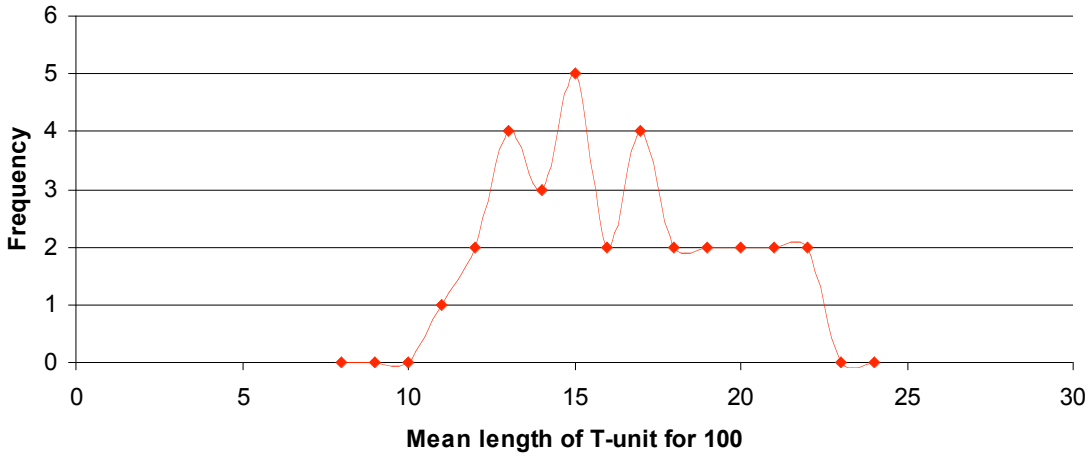


Figure 5. Frequency of mean length of T-unit in writings in level 100

The descriptive statistics for the mean length of T-unit is displayed in Table 3. It is not easy to interpret the relationship between the measures for the two levels merely

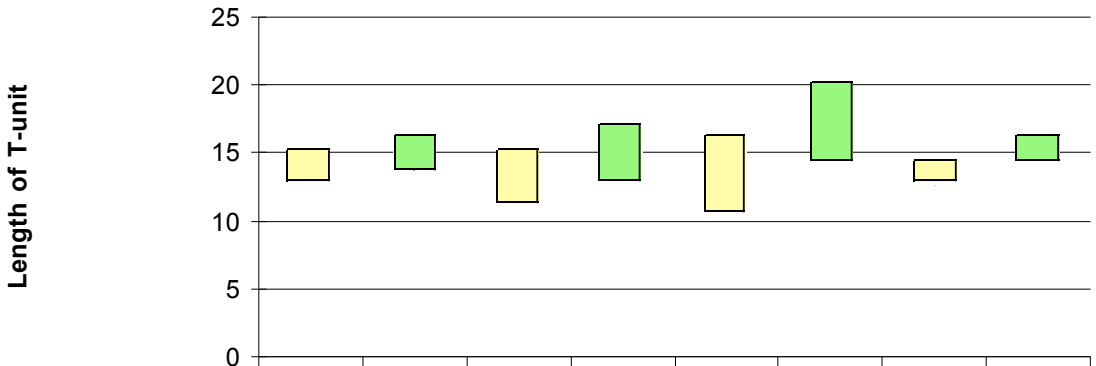
through the descriptive statistics. Thus, the confidence interval of 95% is graphically shown in Figure 6.

Table 3. *Descriptive statistics for length of T-unit in writing samples*

Statistic	ELI 73				ELI 100			
	Prompt 1	**Prompt 2**	**Prompt 3**	**Total**	**Prompt 1**	**Prompt 2**	**Prompt 3**	**Total**
N	11	12	4	27	20	12	8	40
Mean	13.88	13.19	13.46	13.51	14.88	15.16	17.32	15.45
Median	13.49	12.18	13.25	12.92	14.52	14.71	18.34	15
Mode	12.88	N/A	N/A	12.88	N/A	N/A	N/A	N/A
SD	1.50	3.02	1.70	2.27	2.56	3.05	3.32	2.95
Min	11.47	9.69	11.61	9.69	11.13	10.27	11.28	10.27
Max	16.35	18.3	15.72	18.3	21.15	21.84	20.54	21.84

Figure 6 shows that there is no overlap between the two levels with the total N size (CI upper for 73: 14.41 and CI lower for 100: 14.72). This indicates that the difference between the two measures in terms of the mean length of the T-unit can be interpreted as a trustworthy difference. In other words, the difference between the writing samples in ELI 73 and ELI 100 according to the mean length of the T-unit was found to be statistically significant, $t(65) = -2.99$ at $p = .005$

Additionally, the result of the comparison between the three prompts for the mean length of the T-unit is identical to that for the length of the essay; there is no significant difference for both levels. It is interpreted this way because the three prompts overlap with each other in each level in Figure 6.



	73 prompt1 N=11	100 prompt1 N=20	73 prompt2 N=12	100 prompt2 N=12	73 prompt3 N=4	100 prompt3 N=8	73 N=27	100 N=40
CI upper (95%)	14.89	16.08	15.1	17.1	16.16	20.09	14.41	16.18
mean	13.88	14.88	13.19	15.16	13.46	17.32	13.51	15.45
CI lower (95%)	12.87	13.68	11.28	13.22	10.76	14.55	12.61	14.72

Figure 6. Length of T-unit and 95% confidence intervals for placements and placements with different prompts

Research Question 2: How much of the placement decision can be predicted by the fluency variables: the overall length of the essay and the mean length of T-unit?

The following three tables show how much the overall length of an essay, the mean length of T-unit in an essay, and these two together predict a student's placement. Table 4 shows that the number of words in a writing sample was able to correctly classify 88.9% of the students to level 73, and 52.5% to level 100.

Table 4. *Classification results for number of words*

Level	Predicted Level Membership	
	73	100
73	88.9%	11.1%
100	47.5%	52.5%

For the mean length of T-unit, as displayed in Table 5, the mean length of the T-unit predicted 77.8% of the students' membership to level 73 and 47.5% to level 100.

Table 5. *Classification results for length of T-unit*

Level	Predicted Level Membership	
	73	100
73	77.8%	22.2%
100	52.5%	47.5%

A similar result came out for the prediction of the number of words and the mean length of T-unit together in Table 5: 77.8% for level 78 and 57.5% for level 100.

Table 6. *Classification results for number of words & length of T-unit*

Level	Predicted Level Membership	
	73	100
73	77.8%	22.2%
100	42.5%	57.5%

These results show that the length of an essay and the mean length of a T-unit predict highly for a student's membership in the lower level, but not so much for the higher level. This indicates that there may be other effects that influence a student's placement to the higher level rather than just the length of an essay or the length of a T-unit.

Research Question 3. What is the relationship between the length of an essay and the mean length of the T-units?

The Spearman correlation between the two measures is displayed in Table 7. Because the number of pair of the two measures is more than 30 (N>30), the values of the correlation can be converted into an approximation to Student's t-distribution with N-2 degrees of freedom with the following equation (McCall, 2001):

$$t = \frac{\rho\sqrt{N-2}}{\sqrt{1-\rho 2}}$$

After converting it to a Student's t-distribution, the correlation for level 73 was found to be statistically significant, $t(65) = 3.01$, at $p < .05$. However, the correlation for level 100 and that of both levels together were found not to be statistically significant, $t(65)=.57$ and $t(65)=1.86$ at $p < .05$. Therefore, this indicates that though there might be a correlation that is statistically significant in the lower, level, this is not the case for the higher level. This could mean that these two measures might be gauging different measures.

Table 7. *Spearman rank order correlation of the two measures*

	Number of words		
T-unit length	73	100	Total
73	.35*		
100		-.07	
Total			.22

* p < .05

CONCLUSION

In conclusion, the overall length and the mean length of T-unit of the writing show differences in terms of the level they are placed in. Additionally, they both predict highly for the lower level but not for the higher level of the placement. Last, the correlation of the two measures shows that they may not be measuring the same measure, in this case, fluency.

The following are some of the questions that occurred when conducting this study, which might inform future research:

1. Would a larger N size influence the result of the study?
2. What would the results be if there were different kinds of measures, such as, syntactic complexity and accuracy?
3. What could be found out with the comparison of writing samples of students whose native language is English?

REFERENCES

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University.

Berland, H. M., & Jones, R. J. (1982). ESP: Benefits for all of ESL. *English for Special Purposes, 64,* 6-7

Breland, H. (1983). The direct assessment of writing skill: a measurement review. Technical Report No. 83-86. Princeton, NJ: College Entrance Examination Board.

Brown, J. D. (1996). *Testing in language programs.* Upper Saddle River, NJ: Prentice Hall Regents.

Carlson, S., & Bridgeman, B. (1986). Testing ESL student writers. In K. L. Greenberg, H. S. Wiener, & R.A. Donovan (Eds.). *Writing assessment.* (pp.126-152). White Plains, NY: Longman

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

Chenoweth, N. A., & Hayes, J. R. (2001). Fluency in writing: Generating text in L1 and L2. *Written Communication, 18,* 80-98.

Cooper, C. R. (1977). Holistic evaluation of writing. In C. R. Copper & L. Odell (Eds.). *Evaluating writing: Describing, measuring, judging* (pp.3-32). Urbana, IL: NCTE.

Crusan, D. (2002). An assessment of ESL writing placement assessment. *Assessing Writing, 8,* 17-30.

Ellis, D. (2005). *Holistic, analytic, and linguistic measures of second language writing placement test decisions.* Unpublished master's thesis, University of Hawai'i at Mānoa, Honolulu, Hawai'i.

Fillmore, C. J. (1979). On fluency. In C. Fillmore, D. Kempler, & W. S.Y. Wang (Eds.), *Individual differences in language ability and language behavior* (pp. 85-101). New York: Academic Press.

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

Grobe, C. (1981). Syntactic maturity, mechanics and vocabulary as predictors of quality ratings. *Research in the Teaching of English, 15,* 75-86.

Hamp-Lyons, L. (1991a). Basic concepts. In L. Hamp-Lyons (Ed.). *Assessing second language writing in academic contexts.* (pp. 5-15). Norwood, NJ: Ablex, Publishing Corporation.

Hamp-Lyons, L., (1991b). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.). *Assessing second language writing in academic contexts.* (pp. 87-107). Norwood, NJ: Ablex, Publishing Corporation.

Henry, K. (1996). Early L2 writing development: A study of autobiographical essays by university-level students of Russian. *Modern Language Journal, 80,* 309-326.

Hughes, A. (2003). *Testing for language teachers.* Cambridge, NY: Cambridge University.

Huot, B. (1990). The literature of direct writing assessment: major concerns and prevailing trends. *Review of Educational Research, 60* (2), 237-263.

Jacobs, H. L., Zingraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J.B. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.

Kaufer, D. S., Hayes, J. R., & Flower, L. (1986). Composing written sentences. *Research in the Teaching of English, 20,* 121-140.

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

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning, 40*,387-417

Loban, W. (1976). *Language development: kindergarten through grade twelve.* NCTE Research Report No. 18. Urbana, IL: National Council of Teachers of English.

McCall, R. B. (2001). *Fundamental statistics for behavior sciences.* Belmont, CA: Wadsworth

Mellon, J. (1969). *Transformational sentence-combining: A method for enhancing the development of syntactical fluency in English composition.* NCTE Research Report No. 10. Urbana, IL: National Council of Teachers of English.

Ortega, L. (2003). Syntactic complexity measure and their relationship to L2 Proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics, 24,* 492-518.

Penny, J., Johnson, R. L., & Gordon, B. (2000). The effect of rating augmentation on inter-rater reliability: An empirical study of a holistic rubric. *Assessing Writing, 7,* 143-164.

Polio, C. (2001). Research methodology in second language writing research: The case of text-based studies. In T. Silva & P. K. Matsuda (Eds.). On second language writing (pp. 91-115). Mahwah, NJ: Lawrence Erllllbaum.

Purnell, R. B. (1982). *A Survey of the Testing of Writing Proficiency in College: A Progress Report.* College Composition and Communication 33, 407-410.

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

Tarone, E., Downing, B., Cohen, A., Gillette, S., Murie, R., & Dailey, B. (1993). The writing of Southeast Asian-American students in secondary school and university. *Journal of Second Language Writing, 2,* 149-172.

Vaughan, C. (1991). Holistic assessment: What goes on in the raters' minds? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts*. (pp. 111-126). Norwood, NJ: Ablex Publishing Corporation.

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

Wesche, M. B. (1987). Second language performance testing: the Ontario Test of ESL as an example. *Language Testing, 37*, 28-47.

White, E. M. (1984). Holisticism. *College Compsition and Communication, 35,* 400-409.

White, E. M. (1985). *Teaching and assessing writing.* San Francisco, CA: Joosey Bass.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy & complexity* (Technical Report #17). Honolulu: University of Hawaii, Second Language Teaching and Curriculum Center.

APPENDIX A

	Content	Organization	Vocabulary	Grammar	Fluency
English Shows high proficiency in L2, but need for instruction in rhetoric, organization, support, argumentation. Will benefit from	Paper shows evidence of: • Clear, developed argument, but may be simplistic • Some insight on the topic, but may lack depth • Effective support, but evidence and examples may be general or vague	Paper is: • Cohesive • Somewhat formulaic (e.g. 5-para essay format) • Marked by appropriate transitions	Paper is: • Varied vocabulary • Few problems with collocations • Few problems with word choice	Paper has: • Few errors • Complex sentence structure (e.g. complex coordination, subordination, embedded questions, etc.)	Amount of writing is: • Suitable for level of analysis and/or amount of time provided to write the paper
ELI 100 Shows some knowledge and control of academic writing; needs to develop L2 proficiency, writing ability, awareness of genres/conventions common in US universities. Will benefit from ELI rather than English Dept instruction.	Paper shows evidence of: • Clear, developed argument, but may e simplistic • Some insight on the topic, but may lack depth • Effective support, but evidence and examples may be general or vague	Paper is: • Cohesive • Somewhat formulaic (e.g. 5-para essay format) • Marked by appropriate transitions, with some misuse/overuse of transitional phrases	Paper is: • Varied vocabulary • Some problems with collocations • Some problems with word choice	Paper has: • Several errors (e.g. verb tense/aspect, word form, articles, pre-positions), but typically do not interfere with comprehension Paper has: • Some correct complex sentence structure; evidence of other (incorrect) attempts	Amount of writing is: • Suitable for level of analysis and/or amount of time provided to write the paper
ELI 73 Needs to develop L2 proficiency; notable unfamiliarity with and general lack of control of academic writing; would benefit from at least two semesters of ELI writing instruction.	Paper shows evidence of: • Underdeveloped or unclear argument • Simple topic description/restatement, but with little insight • A general lack of supporting evidence, detail, examples • Redundancy of ideas, argumentation	Paper is: • Not cohesive • Formulaic (e.g. 5-para essay format), or lacking organization • Marked by absence of clear transitions between ideas, or simple sentence-level transitions used at paragraph level (e.g. first, next, then)	Paper is: • Notably limited vocabulary • Repetition/overuse of certain lexical items • Numerous problems with word choice • Incorrect collocations	Paper has: • Numerous errors that typically interfere with comprehension • General lack of sentence complexity	Amount of writing is: • Unsuitable for level of analysis and/or amount of time provided to write paper