

การใช้ระบบเหมืองข้อมูลเป็นแบบจำลองลักษณะความคล่อง ในการเขียนของนักศึกษา มหาวิทยาลัยหัวเฉียวเฉลิมพระเกียรติ

Adam Gardiner

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์ที่จะประเมินการฝึกเขียนที่กำหนดโครงสร้างเกี่ยวกับความคล่องในการเขียนของผู้เรียนภาษาที่ 2 โดยใช้แบบจำลองที่ผู้วิจัยได้พัฒนาขึ้น การศึกษาครั้งนี้เก็บข้อมูลจากกลุ่มตัวอย่าง ซึ่งเป็นนักศึกษามหาวิทยาลัยหัวเฉียวเฉลิมพระเกียรติชั้นปีที่ 4 สาขาวิชาภาษาอังกฤษ-ภาษาจีน จำนวน 1 ห้อง ผู้วิจัยกำหนดให้นักศึกษา ทำแบบทดสอบก่อนเรียนและหลังเรียน ซึ่งนักศึกษามีเวลา 8 สัปดาห์ ในการฝึกเขียนที่กำหนดโครงสร้างในหัวข้อต่าง ๆ ลักษณะที่พบจากงานเขียนคือความเร็วในการเขียนความยาวของแต่ละย่อหน้า ระยะเวลาที่ใช้ จำนวนข้อผิดพลาดในแต่ละประเภท วงคำศัพท์ รวมถึงอัตราความเร็ว FLO1 ลักษณะต่าง ๆ เหล่านี้ถูกนำมาสร้างเป็นแบบจำลอง ปัจจัยที่กระทบต่อความคล่องในการเขียนภาษาที่สอง แบบจำลองนี้ถูกนำมาประยุกต์ เพื่อวัดผลกระทบจากการฝึกเขียนที่กำหนดโครงสร้างของผู้เรียนภาษาที่สอง ความคล่องในการเขียนถูกนำมาวัดและกำหนดค่าเรียกใหม่เป็น FLO3 ในทางตรงกันข้ามกับ FLO1 , F LO3 รวมอัตราความเร็วที่แท้จริงในการเขียนค่าของ FLO3 ถูกนำมาคำนวณทางสถิติการถดถอยเชิงเส้นและผลจากการศึกษาในครั้งนี้ สามารถนำไปพิจารณาปรับใช้กับ วิธีการสอน เพื่อเสริมสร้างความคล่องในการเขียน

คำสำคัญ : ความคล่องในการเขียน การเขียนที่กำหนดโครงสร้าง การถดถอยเชิงเส้น FLO3

Using a data mining system to model aspects of writing fluency in ESL students of Huachiew Chalermprakiet University

Adam Gardiner

Abstract

The objective of the study was to assess the effects of practice in semi-structured writing on the writing fluency of L2 writers, using a model that was developed by the researcher. First, the data was collected from a class of fourth year English-Chinese major students at Huachiew Chalermprakiet University. Between the pre-test and the post-test, the students followed an 8-week program that involved regular semi-structured writing about journal topics of a subjective nature. The following attributes were obtained from the students' writing: the speed of writing, total length of both paragraphs, time taken, the number of errors of each type, the error rate for each type, the "lexical richness or range" and "FLO1" rating. "FLO1" refers to: "The rating of an L2 writer's writing by an L1 writer¹, who uses various attributes to measure that writing's quality, *excluding* the attribute of writing speed." Then these attributes were used to construct a model of the factors affecting the rating of L2 writing fluency. Next this model was applied to measure the impact of practice in semi-structured writing about journal topics on the fluency of L2 writers. Fluency was measured by creating a new attribute called "FLO3." In contrast to the attribute of FLO1, FLO3 includes the actual writing speed. The values of FLO3 were calculated by applying the statistical method of linear regression. The values of the following attributes decreased in the post-test: the mean overall error rate, the mean total effect for the five statistically significant error types, and the mean FLO1 rating. The values of the following attributes increased in the post-test: the mean writing speed and the mean FLO3 rating. Finally, some implications of the results for the construct of writing fluency were considered.

Keywords: Writing fluency, semi-structured writing, linear regression, FLO3

I: Introduction

1.1. Background and Significance

Relatively little attention has been given in recent decades to the definition of writing fluency in the context of studying a second language, compared to speaking fluency. This is not surprising, given that most ESL students would rate becoming a fluent speaker as a more urgent goal than becoming a fluent writer. However, since the advent of the worldwide web, the nature of writing has been changing in various ways. Due to the improved speed and reliability of data communications, and the flourishing of social media applications on the web, the author estimates that students are writing much more than their peers of twenty years ago.² Although some traditional forms of writing such as the handwritten letter have declined, various kinds of online writing have surged to take their place. So while students may be writing more than their pre-digital peers, the nature of their writing has changed. There are various aspects of this change, but the most important one for this study is their writing's relationship to time. Most online writing takes place in the context of an online community and expects a response, ranging from hours for a blog post to minutes for a Facebook status update. So being able to write quickly has become an increasingly valuable skill. Therefore, being able to write fluently has become a more pressing goal for ESL students than it was for their pre-digital peers.

Depending on the objective of the research study, writing fluency has been defined in various ways, but usually either in terms of the number of words written or the time taken to write a certain number of words. Wolfe-Quintero, Inagaki, and Kim (1998) defined fluency as the “rapid production of language” (p. 117). In 2003, Jean Chandler (“The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing”) measured writing fluency by asking the students to record the amount of time spent writing an assignment, and then calculating the time taken per 100 words. Chenoweth and Hayes (2001) also used words written per minute to measure fluency. Like the researchers mentioned above, this researcher also accepts writing speed as an essential component of writing fluency. As each student submitted their completed work in the pre- and post-test, the researcher checked that they had recorded an accurate time on the test paper.

Some researchers have added a lexical component to speed or quantity. In 2006, Fellner and Apple (“Developing writing fluency and lexical complexity with blogs”) defined

writing fluency as “the number of words produced in a specified time frame, together with lexical frequency, irrespective of spelling and content, provided that the writer’s meaning is readily understandable” (p.19). The less frequently a word appears in normal written English, the more difficult it was considered to be. Therefore, the students’ fluency was measured by their word count over time, and the proportion of low-frequency words in a student’s writing. An increase in the proportion of low-frequency words used, together with increased word count, would indicate an increase in their fluency. Another study by Sugita, in 2012 (“Enhancing Students’ Fluency in Writing: Learning to Use Transition Words”) measured fluency by the number of words written and successful connections (using transition words such as “moreover”). Regarding lexical factors, this study also tested an attribute that reflected the “lexical richness” of the student’s paragraph. The number of unique words per 50 words was counted. Therefore this study’s lexical richness attribute refers to the lexical range of a paragraph, in contrast to Fellner and Apple’s attribute of lexical frequency. However, the linear regression method used did not find support for a dependent relationship between the target attribute and the lexical richness attribute.

In addition to testing a lexical metric in combination with speed as a measure of fluency, the researcher also included the more exacting attribute of accuracy.³ In this way, this study diverged from those mentioned above. The attribute of accuracy was approached from the viewpoint of the L1 writer assessing a paragraph written by the L2 writer. Written work by L2 students at a university is commonly assessed and given a rating by an instructor who is a native speaker of that language. The researcher decided to develop a model of the factors affecting the rating process and use that model as a starting point for the definition of writing fluency.

Fellner and Apple found that according to their definition of fluency, based on word counts and lexical frequency, the students showed an improvement in their writing fluency following an intensive seven-day CALL-based program. This program required daily posting of messages to a class blog. However, resources for an intensive CALL-based program were not available to the researcher, so this study involved weekly writing about journal topics over a longer period of ten weeks. Writing a blog post and writing about a journal topic are both semi-structured⁴, subjective and non-technical in nature. Therefore this study’s hypothesis, that regular practice in semi-structured writing about topics of a general and subjective nature can help L2 students to write more fluently, shares common ground with

Fellner and Apple's study, but this study's model of fluency diverges from that study and others by including the attribute of accuracy.

2. Objectives of the study

The main objective of the study was to assess the effects of practice in semi-structured writing on these students' writing fluency, and thereby test the hypothesis of the researcher that regular practice in semi-structured writing about topics of a general and subjective nature can help L2 students at HCU to write more fluently (as measured by "FLO3," a metric of writing quality which includes the attribute of writing speed).

A secondary objective was to compare three alternative metrics of writing assessment during the process of testing. These metrics were called "FLO1", "FLO3," and "pure speed."

3. Research Methodology

3.1 Preprocessing

3.1.1 Choice of Software and Operations

The data mining software used for this project is called RapidMiner. This application provides a graphical user interface that allows us to conveniently perform various data mining operations on the data that we obtained from the student writing. In the case of this study, all of the data obtained is "meta-data," which is data about the student writing. The content of the writing is not imported into RapidMiner. There are many operations available in RapidMiner, but the following was selected for this study: linear regression. The linear regression operator can help us to model the relationship between regular attributes and a label attribute. The regression model can then be used to predict the value of a label attribute. In this particular study, the regular attributes (also known as independent variables) refer to various data about the student writing and the label attribute (also known as the dependent variable) refers to the rating of the L2 writing, so the regression model will be used to predict the value of the rating.

"Regression is a technique used for numerical prediction. Regression is a statistical measure that attempts to determine the strength of the relationship between one dependent variable (i.e. the label attribute) and a series of other changing variables known as independent variables (regular attributes)" ("Linear Regression," *RapidMiner Documentation*).

3.1.2 Design of the Classroom Tests

The data collection was from a class of fourth year English-Chinese majors (studying Report Writing in English) between October 2013 and February 2014. For the purpose of this study, each student was assigned a unique reference code containing a letter and a number. The codes were assigned in the same order as their official student code. All students who attended both the pre-test and post-test were given a reference code starting with B, except for two students who had to be excluded because they wrote only one paragraph (as explained below in “Exclusions”). After the exclusions, there remained 22 students who attended both the pre-test and post-test, so their codes ranged from B1 to B22.

The following section explains why each of the following attributes was selected to be determined from the paragraphs. Note that the values of some attributes would be available immediately after the test, such as the total time taken. Other attributes would require further processing, such as the error rates for different types of errors. Finally, the value of “FLO1” would only be determined after the paragraphs had been read by the raters. After the attributes had been determined, a model of factors affecting the rating of L2 writing would be tested and used as the starting point for the definition of writing fluency called “FL03.”

1) Speed of Writing

Speed is an essential component of speaking fluency, and this study proposes it as an essential component of writing fluency. All other factors being equal, (such as the topic, etc.) someone who writes 100 words in 5 minutes is more fluent than someone who writes 100 words in 10 minutes.

2) Total Length (Word Count)

It was necessary to record the number of words in order to calculate the writing speed. However, this attribute was also used independently of the speed to develop a model of FLO1, as will be explained later.

3) Time Taken

It was also necessary to record the time taken to write two paragraphs, in order to calculate the speed (words per minute).

4) Types of Errors

For the purpose of building the model, the researcher started from the assumption that of

the attributes used by the L1 writer to rate the quality of writing (meaning general writing about everyday topics) by the L2 writer, an important one is accuracy, and that some types of errors would have more effect on the rating than other types. On the one hand, accuracy implies the lack of errors that may obstruct the intended meaning to varying degrees. On the other hand, accuracy implies the proficient use of language to express meaning with economy and precision. The former definition was used, as this would be easier to quantify. To measure the attribute of accuracy, the first step taken was to record every single error in all of the paragraphs and allocate them to different types. The different types of errors related to different aspects of grammar. The errors were divided in this way because all of the students (whose first language was Thai) were writing in English as a second language. Since the grammatical rules of Thai and English are quite different, the types of errors made would most likely relate to grammar. It was assumed that these types would all affect the value of “FLO1” (given by the L1 rater), to varying degrees. The types of errors are shown in Table 1 below.

Error type	
A	Verb error (missing verb, agreement error, or tense error)
B	Pronoun error (missing pronoun, or incorrect pronoun)
C	Adjective error (missing adjective, or incorrect ending)
D	Preposition error (missing preposition, or incorrect preposition)
E	Inappropriate word for context (meaning does not accord with context)
F	Singular/plural error
G	Noun error (missing noun, or another part of speech used instead)
H	Conjunction error (missing conjunction or incorrect conjunction)
I	Article error (missing article or incorrect article)
J	Countable/uncountable error
K	Adverb error (missing adverb or another part of speech used instead)
Z	Sentence structure errors (fragments, run-ons, etc.)

Table 1: Types of errors used for the accuracy attribute

The second step taken was to determine the relative effects of different types of errors on the rating given by the L1 writer, using linear regression to develop a model of FLO1. These relative effects were calculated by multiplying the error rate for each type by its coefficient, as explained later in “Comparison of the Pre-Test and Post-Test.”

5) Lexical Richness

This refers to the number of unique words used. It was assumed that more fluent writers tend to use a greater range of vocabulary.

6) FLO1 Rating

This refers to “the rating of an L2 writer’s writing by an L1 writer⁵, who uses various attributes to measure that writing quality⁶, *excluding* the attribute of writing speed.” For this study, the value of FLO1 was decided independently by two native English-speaking instructors at the same university.

3.1.3 Conditions of the Classroom Tests The classroom assignments were given in a relaxed environment, designed to provide the best possible conditions for free writing. Next, the specific conditions of the classroom writing assignments will be described, with reasons where appropriate.

1) Journal Topics

These topics mainly related to the students’ own experience, so that their writing was not slowed by having to look up reference information. Also, most of the topics required imagination and creativity, which encouraged the students to write freely without worrying about making mistakes. The topics were wide-ranging, so that every student could find a topic that matched their interests. There were four different lists of topics. Two of the lists were given for the pre-test assessment, and the other two lists were given for the post-test. This was to eliminate the possibility of students memorizing the paragraph they had written in the pre-test. One topic had to be chosen from each list, and one paragraph had to be written for each topic, making a total of two paragraphs.

2) Time limit

The students were given 10 minutes before the test started. During these 10 minutes, they were directed to browse the lists of topics (which they had not seen before), choose their two topics and do any kind of prewriting technique to prepare their paragraphs, but not to start writing the paragraphs. After the ten minutes preparation stage, they were allowed to start writing the paragraphs. The students were asked to write two paragraphs at their natural pace. Each student recorded the time taken to write a paragraph for each

topic (the total time for two paragraphs was calculated later). Note that this time did *not* include the 10 minutes preparation stage. They were directed to start recording the time only when they started writing the first paragraph. Regarding the question of whether there should be a time limit, it was important for the study that students could write freely without pressure, in contrast to exam conditions. It was also important that students were able to write an integral paragraph that covered the topic, and different writers take differing lengths of time to achieve that. For these reasons, a single time limit was not imposed. Instead, a minimum and maximum time limit were set (not including the 10 minutes preparation time). Specifically, a minimum time limit of 10 minutes and maximum time limit of 30 minutes were set for the pre-test and post-test. In both tests, all of the students wrote for 10 minutes or longer. In both tests, most of the students had stopped writing before or at the 30 minutes cutoff. However, the maximum time limit was not enforced. This was because a few students were very reluctant to stop writing after 30 minutes, being so immersed in their topic, so I decided to let them continue until they had finished. Finally, the assignment only took place in the classroom, so I was able to check the accuracy of the time that was recorded by the student. Each student handed their paragraphs to me when they had finished, and I checked that they had recorded a time that was accurate.

3) Handwriting

The assignment had to be handwritten. Handwritten paragraphs allowed the students to write freely without any distractions such as pop-up notifications. It also eliminated the risk that someone might copy and paste content from another document or the internet.

4) Use of books and internet

Students were permitted to use dictionaries, either in book or electronic form. However, reference to any other printed materials was not allowed. Internet access was also not permitted. This was to exclude the possibility of a student copying content from an online source.

5) Assessment

Students were informed that their writing in the pre-test and post-test would not have any effect on their final grade in the subject that they were studying.

3.1.4. Conditions of the Program

After the pre-test, the students were directed to continue writing about topics from the two lists that had been given in the pre-test, except for the two topics that they had already written about. They were directed to write two paragraphs weekly, choosing a

different topic each time, and follow the same time limit as for the pre-test (minimum of 10 and maximum of 30 minutes). However, it was not necessary to record the time taken. Lists 2 & 3 contained a total of 55 topics, so they had to choose a total of 16 (8 X 2) topics from the remaining 53. Week 1 ended on November 4th and Week 8 ended on December 23rd (December 30th was excluded, as it was in the New Year holiday). Finally, in the post-test they were given Lists 3 & 4, which they had not seen previously. The conditions of the post-test were identical to those in the pre-test (see “Conditions of the Classroom Tests”), except that the lists of topics were different.

3.1.5 Preparing the Data

In this section, the various steps of preparing the data (for data mining operations) are described, starting from immediately following the classroom test, up to the point of inputting the data to the RapidMiner software.

Exclusions Some students had to be excluded from the study due to missing or invalid data as follows:

Pre-test (28 October 2013)

Two students were excluded from the study due to absence from the pre-test.

Post-test (6 January 2014)

One student (who had attended the pre-test) was excluded from the study due to absence from the post-test. Two students were excluded for only choosing one topic and writing one paragraph. They did not follow the given directions, to choose two topics and write a paragraph about each topic.

Processing of the Attributes

1) Speed of Writing

The times taken to write the two paragraphs were added, and converted to seconds. Then the total number of seconds was divided by the number of words, to give a value of seconds per word. Then sixty was divided by this value to give a value of words per minute.

2) Total Length (Word Count)

For each paragraph, the number of words was counted. A combined total was produced for both paragraphs.

3) Time Taken

The time taken to write two paragraphs was used in combination with the word count to calculate the speed (words per minute) as described above.

4) Types of Errors

For both paragraphs written by the student, all errors were identified. Each error was allocated to a type (see Table 1) and counted. Then the number of errors for each type was divided by the total number of words (of both paragraphs combined) and multiplied by 100, thus giving the number of errors of that type occurring every 100 words.

5) Lexical Richness

The richness attribute is a score out of 50. It represents the number of unique words in a sample of 50 words that were randomly extracted from the two paragraphs.

6) FLO1 Rating

Identifying information such as the student code was concealed on all the paragraphs. The raters were only able to see the reference code (as explained in “Design of the Classroom Tests”). They were also shuffled, so that the raters did not know which paragraphs were in the pre-test and which were in the post-test. Separate copies were given to the two raters, so that each could not see the other’s score. The raters were two native English-speaking instructors at the same university. The instructors were directed to give each set of two paragraphs a score out of 100, as if they were the written part of a university exit exam. A perfect score (100%) should only be given to someone who wrote as an educated native speaker would. After the rating process, each set of two paragraphs had two scores (each out of 100). These scores were added together to give a score out of 200. This score was entered in the column “FLO1 = $x/200$ ” in various Excel files of the data collection, as detailed below in “Formatting of Tables.”

Formatting of Attributes

Next, all the attributes had to be correctly formatted for import into RapidMiner. In particular, the data types of the attributes in the labeled data set had to exactly match those of the attributes in the unlabeled data set, in order for the linear regression operator to execute completely. For example, if an attribute was defined as “numerical” in the labeled dataset and “integer” in the unlabeled dataset, the operator would terminate execution as soon as it tried to read the unlabeled dataset.

Formatting of Tables

Finally, before it could be imported into RapidMiner, the data collection (namely, the data collection from a class of fourth year English-Chinese majors, studying Report Writing in English between October 2013 and February 2014) had to be arranged to suit the requirements of the methodology used to determine FLO3. Specifically, the pre-test data

(28 October 2013) was divided into two tables: one table containing the labelled data and another containing the unlabelled data. Likewise, the post-test data (6 January 2014) was divided into two tables: one table containing the labelled data and another containing the unlabelled data.

3.2 Data Mining

3.2.1 Background

As already explained in “Objectives of the study”, FLO3 is a metric of writing quality that uses various attributes to measure writing quality, like FLO1. However, unlike FLO1, FLO3 uses the attribute of speed to measure writing quality, in addition to other attributes of the writing.

3.2.2 Using linear regression to determine “FLO3”

The methodology used to determine “FLO3” consisted of three steps. The first step was using a formula to adjust the word count of those students who wrote for longer than 10 minutes (in this case, all of the students in both pre-test and post-test, except for one student in the post-test who took exactly 10 minutes). Their word count was reduced according to a formula which calculated the number of words they would have written after 10 minutes, assuming a constant speed. The second step was applying the statistical method of multiple linear regression (the standard ordinary least squares version) to develop a model of FLO1. The third step was applying that model to the adjusted data of these writers, to obtain predictions for the values of FLO3. After applying these three steps, those students who wrote for more than 10 minutes received a FLO3 rating that was less than their FLO1 rating. For any writers who wrote for exactly 10 minutes or less (only one in this case) the values of FLO1 and FLO3 were identical. A much more detailed description of these steps can be obtained from the author.

4. Research Results

The following analysis of results is dependent on the final linear regression model that was applied during the data mining stage.

4.1 Comparison of the Pre-test and Post-test

The values of the FLO3 predictions were used to create a new Excel table (Table 2), together with the FLO1 ratings and speeds already obtained. Then five error rates (for types A,C,E,G,I) were used for calculating the total effects, as will be explained in the

following section. These five error types were used because these were the only error types that had a statistically significant p-Value of less than 0.05 and therefore the only error types that were retained in the final regression model.

	Code	PRETEST FLO1 SCORE (x/200)	POSTTEST FLO1 SCORE (x/200)	Change in FLO1 (posttest - pretest)	PRETEST FLO3 SCORE (x/200)	POSTTEST FLO3 SCORE (x/200)	Change in FLO3 (posttest - pretest)	PRE TEST SPEED (wpm)	POST TEST SPEED (wpm)	Change in Speed (posttest - pretest)	PRE TEST TOTAL EFFECTS (TYPES A+C+E+G+I)	POST TEST TOTAL EFFECTS (TYPES A+C+E+G+I)	CHANGE IN TOTAL EFFECTS (TYPES A+C+E+G+I)
1	B1	150	118	-32	98.44	118.00	19.56	6	12	6	18.12	10.55	-7.57
2	B2	150	149	-1	115.75	113.79	-1.96	13	10	-3	17.11	11.15	-5.96
3	B3	100	74	-26	76.51	76.16	-0.35	6	6	0	38.88	41.10	2.22
4	B4	122	130	8	96.67	97.38	0.71	7	9	2	22.22	24.77	2.55
5	B5	160	144	-16	120.06	119.60	-0.46	14	11	-3	14.66	7.43	-7.23
6	B6	134	110	-24	90.47	97.39	6.92	8	9	1	29.58	26.85	-2.73
7	B7	154	147	-7	111.49	105.13	-6.36	6	4	-2	5.29	6.30	1.01
8	B8	136	140	4	106.81	107.11	0.30	6	9	3	9.97	15.73	5.76
9	B9	120	116	-4	105.89	100.61	-5.28	8	11	3	15.09	25.96	10.87
10	B10	160	134	-26	106.53	108.72	2.19	9	10	1	16.54	17.15	0.61
11	B11	156	133	-23	124.23	113.05	-11.18	12	9	-3	6.30	10.03	3.73
12	B12	134	152	18	98.50	110.29	11.79	6	8	2	17.59	10.46	-7.13
13	B13	106	100	-6	91.38	95.65	4.27	6	10	4	24.24	28.83	4.59
14	B14	110	113	3	102.45	102.16	-0.29	6	5	-1	14.10	11.83	-2.27
15	B15	125	110	-15	110.14	94.31	-15.83	11	6	-5	16.66	21.08	4.42
16	B16	144	140	-4	93.59	105.57	11.98	6	9	3	21.56	17.74	-3.82
17	B17	148	153	5	108.37	106.41	-1.96	6	8	2	6.79	13.87	7.08
18	B18	144	140	-4	99.29	105.13	5.84	7	9	2	19.36	17.01	-2.35
19	B19	130	140	10	105.59	105.29	-0.30	4	5	1	5.84	8.23	2.39
20	B20	106	110	4	84.21	108.15	23.94	6	8	2	31.42	11.66	-19.76
21	B21	84	132	48	83.14	104.73	21.59	6	12	6	32.26	24.86	-7.40
22	B22	121	104	-17	96.73	103.96	7.23	10	12	2	28.21	25.40	-2.81
	Avg =>	131.55	126.77	-4.77	101.19	104.48	3.29	7.68	8.73	1.05	18.72	17.64	-1.08

Table 2: Final results from the pre and post test, showing the changes for FLO1, FLO3, speed and effects (for types A,C,E,G,I)

1) FLO1 (x/200)

As can be seen from Table 2, the class's average FLO1 rating decreased from 131.55 in the pre-test to 126.77 in the post-test.

2) FLO3 (x/200)

The same table shows that the class's average FLO3 rating increased from 101.19 in the pre-test to 104.48 in the post-test.

3) Speed (wpm)

The same table shows that the class's average writing speed increased from 7.68 wpm in the pre-test to 8.73 wpm in the post-test.

4) Total Effects (of error types A, C, E, G and I)

For each of the error types in the final regression model (A, C, E, G and I), that attribute's error rates in the pre-test and post-test were multiplied by its regression coefficient, as seen in the second column of Table 3 below. The coefficient indicates the amount of change occurring in the dependent variable (the rating) in response to the change in one unit of an independent variable. Multiplying an attribute's error rate by its coefficient gives a "total effect" which is more useful than the overall error rate, as it gives more weighting to types of errors that have more effect on the rating.

Attribute	Coefficient	Std. Error	Std. Coeffi...	Tolerance	t-Stat	p-Value	Code
(Intercept)	102.117	5.597	?	?	18.245	0	****
TOTAL LEN	0.233	0.021	0.661	0.844	11.028	0	****
No. of A typ	-1.773	0.531	-0.206	0.931	-3.341	0.002	***
No. of C typ	-7.140	2.653	-0.166	0.988	-2.691	0.012	**
No. of G typ	-4.595	1.788	-0.173	0.871	-2.570	0.016	**
No. of I typ	-2.469	1.058	-0.145	0.974	-2.335	0.028	**
No. of E typ	-1.929	0.917	-0.138	0.958	-2.102	0.049	**

Table 3: Output from “3076 linear regression process 28 Oct 13 (to predict FLO3) 6 regular attributes” (ordered by p-Value). NOTE: this process is running the final model.

For example, in the pre-test, the error rate for type A for writer B1 was 1.93. This rate was multiplied by 1.773 which is the coefficient for type A, to return a value of 3.42. Next, the sum of the five values was calculated for each writer, for both pre- and post-test. The sums are shown in the Excel columns named “PRE TEST TOTAL EFFECTS (TYPES A+C+E+G+I)” and “POST TEST TOTAL EFFECTS (TYPES A+C+E+G+I).” Finally, the means of each column were calculated and compared. The difference between the pre and post-test is shown in the column named “CHANGE IN TOTAL EFFECTS (TYPES A+C+E+G+I).” As can be seen in Table 2, the class's average total effect for these five error types decreased from 18.72 in the pre-test to 17.64 in the post-test. As already mentioned, the total effect for the error types is more useful than the overall error rate, as it gives more weighting to types of errors that have more effect on the rating. However, it is also notable that the overall error rate decreased from 13.7 per 100 words in the pre-test to 12.2 per 100 words in the post-test.

1) Why did the FLO1 rating decrease?

The main reason for the decrease in FLO1 must be that the average word count was higher for the pre-test (204) than for the post-test (184). The error rates can be excluded as a cause, since average total effects for types A, C, E, G and I decreased in the post-test, as did the overall error rate. Of course, it is possible that the decrease in FLO1 was also caused by changes in one or more explanatory variables that were not included in the model. However, if the decrease (20) in the average word count is multiplied by its regression coefficient (0.233), the result is 4.66 which is almost equal to 4.77 (the actual decrease in the FLO1 rating). So the decrease in average word count appears to be the main cause.

2) Why did the FLO3 rating increase?

The FLO3 rating was the result of applying a model containing the following label or target attribute: “VALUE OF FLO3 = left blank because it is the LABEL,” and the following six regular attributes:

1. “TOTAL LENGTH (word count)”
2. “No. of A type errors ... per 100 words”
3. “No. of C type errors ... per 100 words”
4. “No. of E type errors ... per 100 words”
5. “No. of G type errors ... per 100 words”
6. “No. of I type errors ... per 100 words”

Therefore, the increase in the FLO3 rating must be due to changes in at least one of these regular attributes. For calculating the FLO3 rating, the TOTAL LENGTH variable was equivalent to the adjusted word count. The average adjusted word count for the pre-test is 76, while the average adjusted word count for the post-test is 86. So the increase in the FLO3 rating can be partially accounted for by the increase in the average adjusted word count. Moreover, the increase in the average adjusted word count must have been caused by an increase in average speed, since we already know that the average unadjusted word count decreased in the post-test. As can be seen in Table 2, the average speed did indeed increase (by 1.05 wpm) in the post-test. So the increase in the average adjusted word count is one factor accounting for the increase in FLO3. Another factor must be the decrease in the average total effect for these five error types from 18.72 in the pre-test to 17.64 in the post-test.

4.2 Using Analysis by Quartiles to Compare the Results by Groups and Metrics

Unlike the linear regression and correlation matrix operators, “Analysis by Quartiles” is not a statistical method, but simply a different way of presenting the results from the linear regression. In the following section, the results obtained (see Table 2) from the linear regression will be ordered by quartiles to show the differences between the upper quartile of writers (also referred to by the term Q_3) and the lower quartile of writers (also referred to by the term Q_1). This method of ordering will reveal the effects of regular semi-structured writing on different groups by ability. The quartiles are small, so the researcher is looking for interesting patterns in the quartile distribution rather than statistical trends. The results will also be ordered by three different metrics: FLO1, FLO3 and pure speed. This method of ordering will enable a comparison of the different metrics. First, six new tables were created from the table showing the results of the linear regression (see Table 2). These new tables contained exactly the same data, but were ordered differently. The original table of results was ordered by the reference code (from B1 to B22). Two new tables were ordered by FLO1, one for the pre-test and another for the post-test. Two new tables were ordered by FLO3, one for the pre-test and another for the post-test. Two new tables were ordered by pure speed, one for the pre-test and another for the post-test. Then, for each metric the pre-test and post-test tables were appended horizontally, resulting in three tables – for FLO1, FLO3 and pure speed. These three quartile tables are not shown here, but can be obtained from the author.

The following method (also known as Tukey’s hinges) was used to divide the data into quartiles.

- *Use the median to divide the ordered data set into two halves. If the median is a datum (as opposed to being the mean of the middle two data), include the median in both halves.*
- *The lower quartile value is the median of the lower half of the data. The upper quartile value is the median of the upper half of the data.*

(Wikipedia, “Quartiles”)

4.3 Implications

The following discussion uses the relativized values of the total effects (for error types A,C,E,G,I), in which the highest value (41.1) is mapped to 1.0. Theoretically, it would be possible for a writer’s absolute value of total effects to increase, while their corresponding relativized value decreased, or vice versa, but in fact the total scores did not vary enough

between pre and post-test for this discrepancy to occur. So the researcher decided to use the relativized totals in this analysis, as it is easier to compare the consequences of applying the different metrics.

When ordered by FLO1, the average total effects for the lower quartile remained the same (0.63) for the pre- and post-test. The average total effects for the upper quartile decreased from 0.32 to 0.26. By contrast, when ordered by FLO3, the average total effects for the lower quartile decreased from 0.72 to 0.68. They remained the same (0.27) for the upper quartile. When ordered by speed, the average total effects for the lower quartile decreased from 0.63 to 0.42. They increased from 0.40 to 0.43 for the upper quartile.

Further, the quartile charts suggest that regular semi-structured writing may have benefitted the accuracy (in relation to the five error types) of the lower quartile more than the upper quartile when ordered by FLO3 or pure speed. In the following analysis, the researcher will look more deeply into the changes that took place between the pre- and post test to investigate this apparent phenomenon, namely that the least fluent writers (the lower quartile as measured by FLO3 or pure speed) generally achieved a greater improvement in either speed or accuracy (or both) in the post-test than the writers outside the lower quartile.

Only when ordered by FLO1 did the upper quartile show an improvement in accuracy, and only when ordered by FLO1 did the lower quartile not show an improvement in accuracy. This is not surprising, considering that FLO1 is the only metric of the three that is not influenced by speed.

Therefore, from the provisional analysis above, the researcher concludes that the quartile charts do reveal an interesting pattern, namely that the least fluent writers (the lower quartile as measured by FLO3 or pure speed) generally achieved a greater improvement in either speed or accuracy (or both) in the post-test than the writers outside the lower quartile, although this pattern is much more pronounced when ordered by speed. However, the size of the quartiles is too small to establish the existence of a statistical trend.

So we already know from “Comparison of the Pre-test and Post-test” that the average speed increased and average total effects (for A,C,E,G,I) decreased (= accuracy improved) from the pre to the post-test. From the quartile analysis, we can make an additional observation that the least fluent writers have made a greater improvement than the remaining writers, if the least fluent writers are defined as in the preceding paragraph.

5. Discussion and Recommendations

The study's hypothesis was that regular practice in semi-structured writing of a general and subjective nature could help L2 students at HCU to write more fluently (as measured by FLO3). This study indeed found that the students showed an improvement in their writing fluency (as measured by FLO3), following regular practice in semi-structured writing about journal topics between 28 October 2013 and 6 January 2014. As described in "Comparison of the Pre-test and Post-test", the students showed an overall improvement in both writing speed and accuracy (as defined by either the simple overall error rate or the total effects for error types A,C,E,G and I) from the pre- to the post-test. Further, when the combined effects of the five error types, instead of the simple overall error rate, are compared for the pre- and post-test, the results show that the least fluent writers (the lower quartile as ordered by FLO3) accomplished a greater improvement in accuracy from the pre-test to the post-test than the writers outside the lower quartile. The improvement in accuracy of the lower quartile, compared to the other quartiles, is especially emphatic when the quartiles are ordered by pure speed.

5.1 Research Limitation

It was not practical to build a model that tested every single attribute that might possibly affect the rating of writing quality. To do that would have required a much larger volume of data and a few more years to analyze the results. The researcher considers it likely that some kind of lexical attribute has some effect on the rating and one kind of lexical attribute (lexical range) was tested in this study's model. However, the null hypothesis could not be rejected for lexical range in this model and so this attribute had to be removed. This does not necessarily mean that lexical range does not have any effect; using a much larger sample of data might allow the null hypothesis to be rejected for this attribute.

5.2 Further Research

The results showed that some types of errors did have more effect on the rating than others. A very interesting phenomenon here is the variation in the sizes of the effects. So what factor determined this variation of the effects (of different types of errors on the rating)? It is outside the scope of this study to answer this question, but one possible explanation is that the rater gives a weighting to errors according to the degree that they affect the comprehensibility of the writing.⁷ It would be interesting to extend this study by

using a greater range of writing samples, including some writing samples from native speakers, a larger number of raters, and a more precise classification of error types.

Further, only five error types (A, C, E, G and I) had a statistically significant p-Value of less than 0.05. So only these error types could be retained in this study's regression model. Again, this reflects the constraints of the data sample. In this case, the limitation lay in the quantity of errors for each type. It may not be coincidental that the most frequent type of error in the pre- and post-test (the A type) is also the error type with the lowest p-Value (see Table 3). Unfortunately, the remaining error types had to be excluded from the model, not because it was proven that those variables did not have any effects on the rating, but because the sample size was not large enough to indicate with a high enough level of probability that they did have effects on the rating. A larger data sample, with a greater quantity of errors, might allow the effects of a greater range of error types to be measured and incorporated into the model.

References

- Chandler, Jean. (2003) "The Efficacy of Various Kinds of Error Feedback for Improvement in the Accuracy and Fluency of L2 Student Writing." **Journal of Second Language Writing** 12 (3): pages 267-96.
- Chenoweth, Ann; Hayes, John (2001) Fluency in writing: Generating text in L1 and L2. *Written Communication* 18, 80-98.
- Fellner, Terry, and Matthew Apple. (2006) "Developing Writing Fluency and Lexical Complexity with Blogs." **The JALT CALL Journal** 2 (1): pages 15-26.
- Sauro, Jeff. (2013) "Best Practices For Using Statistics On Small Sample Sizes." **MeasuringU: Usability, Customer Experience & Statistics**.
- Sugita, Yoshihito. (2012) "Enhancing Students' Fluency in Writing: Learning to Use Transition Words." **Open Journal of Modern Linguistics**. 2 (1): pages 18-25.
- Trubek, Anne. (2015) "We Are All Writers Now." **More Intelligent Life**. The Economist Newspaper Limited.
- Wolfe-Quintero, Kate, Shunji Inagaki, and Hae-Young Kim. (1998) **Second Language Development in Writing: Measures of Fluency, Accuracy, & Complexity**. Honolulu: Second Language Teaching & Curriculum Center, University of Hawaii at Manoa.

Endnotes

1 An “L1 writer” in this study means anyone who writes exactly like an L1 writer, regardless of their birthplace.

2 Due to the diversity of contexts of online writing, it is difficult to verify this statistically. However, the argument that people are writing more than they used to is often advanced by commentators on the web. For example, according to Trubek (“We are all writers now”) not only is more being written, but also more people are writing. At this point, we are simply concerned with the *quantity* of writing that is taking place.

3 The researcher is referring to the L1 writer’s assessment of writing about a general topic, where no specialist knowledge is required. Accuracy in this context refers solely to the accurate use of language, and not to any facts about the real world.

4 Whenever this study refers to “semi-structured writing” it is referring to writing that does not follow a formal structure comprising an attention getter, thesis statement, body paragraphs and conclusion. However, “semi-structured writing” is still focused on a single topic and should be logically coherent. Hence, it is called “semi-structured.” Note that logical coherence was not selected as an attribute, due to the difficulty of measuring it.

5 An “L1 writer” in this study means anyone who writes exactly like an L1 writer, regardless of their birthplace.

6 Whenever this study refers to “quality,” it is referring to the quality of general writing about everyday topics, not intensely creative writing such as a short story or novel, or technical writing about specialist topics. It is also referring to brief writing activities, which are completed in a single session, rather than sustained writing activities over multiple sessions.

7 “Comprehensibility” here refers to the rater’s estimation of the comprehensibility of the text. It does not refer to the L1 writer’s degree of comprehension of what the L2 writer has written.