“The Difference of Speed”Using a data mining system to model aspects of writing fluency in ESL students of HuachiewChalearmprakiet University

“ความแตกต่างของความเร็วในการเขียน” การใช้ระบบเหมืองข้อมูลในการวางรูปแบบลักษณะความคล่องในการเขียนของนักศึกษามหาวิทยาลัยหัวเฉียวเฉลิมพระเกียรติ ที่เรียนภาษาอังกฤษเป็นภาษาที่สอง

ADAM GARDINER

This research is supported by Huachiew Chalearmprakiet University

Academic Year 2011

Research Title “The Difference of Speed.” Using a data mining system to model aspects of writing fluency in ESL students of Huachiew Chalermprakiet University

Researcher Adam Gardiner

Institution Huachiew Chalermprakiet University

Year of Publication 2016

Publisher Huachiew Chalermprakiet University

Sources Huachiew Chalermprakiet University

No. of pages 70 pages

Keywords data mining, linear regression, writing fluency, metric,semi-structured writing

ABSTRACT

This research study aimed to develop an optimal metric of writing fluency and determine whether regular practice in semi-structured writing about journal topics could help L2 students at HCU to write more fluently. The study used a data mining software platform called RapidMiner, and applied the statistical method of linear regression. The data collection was from a class of fourth year English-Chinese majors, who were studying Report Writing in English. The pre-test on 28 October 2013 required writing two paragraphs about journal topics, chosen from two lists.After the pre-test, the students followed an 8-week program that involved weekly writing about other journal topics from the same two lists. Finally, in the post-test on 6 January 2014, they were given two new lists, which they had not seen previously. The following attributes were obtained from the students’ writing: the speed of writing (words per minute), total length (word count) of both paragraphs, time taken, the number of errors of each type, the error rate per 100 words for each type, the “lexical richness or range” and “FLO1” rating. “FLO1” and “FLO3” wereterms created by the researcher.FLO1 refers to: “The rating of an L2 writer’s writing by an L1 writer[i], who uses various attributes to measure that writing’s quality, *excluding* the attribute of writing speed.” FLO3 is a metric of writing quality that, like FLO1, uses various attributes to measure writing quality. However, unlike FLO1, the attributes

used by FLO3 include the actual writing speed. The value of FLO3 in the pre-test and post-test was determined for each student, by following a process of three steps. The first step was using a formula to adjust the word count of those students who wrote for longer than10 minutes. The second step was applying linear regression to refine and develop the model of FLO1. The result of this step was an optimal model of FLO1. The label or target attribute ("FLO1 = x/200") remained unchanged as the label in the optimal model. Six regular attributes, including "TOTAL LENGTH (word count)" and five different types of errors, were retained in the optimal model. The third step was applying this model to data that had been adjusted to reflect the speed of writing, namely an adjusted word count as explained in the first step. The output of applying the model was predictions of the "label." These predictions reflected the speed of writing and thus became the values of FLO3.Then the results of the pre-test and post-test were compiled, using the values of the FLO3 predictions calculated by the linear regression operator, the FLO1 ratings (x/200) from the raters, speeds (wpm) and total effects (= error rate multiplied by its regression coefficient) forfive different types of errors. For each attribute, in the pre- and post-test, means were calculated and compared. The following values decreased in the post-test: the class's average overall error rate, the class's average total effect for these five error types, and the class's average FLO1 rating. The following values increased in the post-test: the class's average writing speed and the class's average FLO3 rating. Finally, the quartile results were analyzed in order to compare the three metrics. FLO3 was shown to be a more useful metric than both FLO1 and pure speed for assessing the semi-structured writing of the pre- and post-test since it better captured the two most critical aspects of the change in the students' writing from the pre- to the post-test, namely speed and accuracy.

---

[i] An "L1 writer"in this study means anyone who writes exactly like an L1 writer, regardless of their birthplace.

ชื่อเรื่อง	"ความแตกต่างของความเร็วในการเขียน" การใช้ระบบเหมืองข้อมูลในการวางรูปแบบลักษณะความคล่องในการเขียนของนักศึกษามหาวิทยาลัยหัวเฉียวเฉลิมพระเกียรติที่เรียนภาษาอังกฤษเป็นภาษาที่สอง
ผู้วิจัย	อาดัม การ์ดิเนอร์
สถาบัน	มหาวิทยาลัยหัวเฉียวเฉลิมพระเกียรติ
ปีที่พิมพ์	2559
สถานที่พิมพ์	มหาวิทยาลัยหัวเฉียวเฉลิมพระเกียรติ
แหล่งที่เก็บรายงานฉบับสมบูรณ์	มหาวิทยาลัยหัวเฉียวเฉลิมพระเกียรติ
จำนวนหน้างานวิจัย	70 หน้า
คำสำคัญ	เหมืองข้อมูล การวิเคราะห์การถดถอยเชิงเส้น ความคล่องในการเขียน เกณฑ์การวัดผล การเขียนแบบกึ่งโครงสร้าง
ลิขสิทธิ์	มหาวิทยาลัยหัวเฉียวเฉลิมพระเกียรติ

บทคัดย่อ

งานวิจัยนี้มีจุดมุ่งหมายที่จะพัฒนาการวัดผลที่ดีที่สุด สำหรับความคล่องในการเขียนและกำหนดว่าการฝึกเขียนเป็นประจำแบบกึ่งโครงสร้างเกี่ยวกับหัวข้อการเขียนที่กำหนดให้สามารถช่วยผู้เรียนภาษาอังกฤษเป็นภาษาที่สองเขียนได้คล่องขึ้นหรือไม่ งานวิจัยนี้ใช้ข้อมูลที่เป็น Mining software platform ซึ่งมีชื่อเรียกว่า Rapid Miner และประยุกต์ สถิติการวิเคราะห์การถดถอยเชิงเส้นผู้วิจัยเก็บข้อมูลจากงานเขียนของนักศึกษาชั้นปีที่ 4 สาขาวิชาภาษาอังกฤษ-ภาษาจีน ที่ลงทะเบียนเรียนในรายวิชาการเขียนรายงาน โดยกำหนดให้นักศึกษาแต่ละคนเขียนย่อหน้าคนละ 2 ย่อหน้า มีการทดสอบก่อนเรียนเมื่อวันที่ 28 ตุลาคม 2556 โดยกำหนดให้เลือกหัวข้อการเขียนที่ผู้วิจัยกำหนดให้หลังจากนั้น 8 สัปดาห์ ผู้เรียนเขียนเรื่องเดิมอีกครั้ง และทดสอบหลังเรียนในวันที่ 6 มกราคม 2557 นักศึกษาได้รับรายชื่อหัวข้อการเขียนใหม่ที่ ไม่เคยฝึกเขียนมาก่อนจำนวน 2 หัวข้อ จากงานเขียนของนักศึกษาผู้วิจัยพบความเร็วในการเขียน (คำต่อนาที) ความยาวในการเขียน (จำนวนคำ) ของทั้ง 2 ย่อหน้า การใช้เวลาในการเขียนข้อผิดที่พบในแต่ละประเภท จำนวนข้อผิดต่อจำนวนคำหนึ่งร้อยคำ การใช้คำศัพท์ที่เพิ่มมากขึ้น และการให้คะแนนจากเจ้าของภาษาในการวัดคุณภาพงานเขียนเพิ่มขึ้น

ในงานวิจัยนี้ ผู้วิจัยได้กำนด รหัสFL01 และ FL03 ซึ่ง FL01 หมายความว่าเจ้าของภาษาใช้เกณฑ์ที่หลากหลาย ในการวัดคุณภาพงานเขียนโดยไม่รวมความเร็วในการเขียนส่วน FL03 หมายความว่าเจ้าของภาษาใช้เกณฑ์ที่หลากหลาย ในการวัดคุณภาพงานเขียน โดยรวมความเร็วในการเขียน โดยมีการกำหนดค่า FL03 สำหรับนักศึกษาแต่ละคนใน การทดสอบก่อนเรียนและการทดสอบหลังเรียนตามกระบวนการซึ่งประกอบด้วย 3 ขั้นตอนคือ ขั้นตอนแรกเป็นการใช้สูตร เพื่อปรับการนับคำ สำหรับนักศึกษาที่ใช้เวลาในการเขียนเกิน 10 นาที ขั้นตอนที่สองคือ การประยุกต์การวิเคราะห์ การถดถอยเชิงเส้น เพื่อนำไปปรับปรุงรูปแบบ

ของ FL01 เป็น FL01=X/200 ขั้นตอนที่สามคือ การประยุกต์รูปแบบนี้กับข้อมูลที่มีการปรับเพื่อสะท้อนความเร็วในการเขียน

ผู้วิจัยได้รวบรวมผลการทดสอบก่อนเรียน และผลการทดสอบหลังเรียนมาวิเคราะห์คำนวณโดยการวิเคราะห์ ถดถอยเชิงเส้น นอกจากนี้ผู้วิจัยได้เปรียบเทียบลักษณะข้อผิดที่แตกต่างกัน 5 รูปแบบในการทดสอบก่อนและหลังเรียน ซึ่งพบว่ามีจำนวนลดลงในการทดสอบหลังเรียน ส่วนค่าเฉลี่ยของงานเขียนของนักศึกษาทั้งชั้นเรียนในด้านความเร็ว ในการเขียนเพิ่มขึ้นจะเห็นได้ว่า FL03 เป็นประโยชน์ต่อการประเมินงานเขียน แบบกึ่งโครงสร้างเนื่องจากสามารถระบุ ลักษณะที่สำคัญสองประการในงานเขียนของนักศึกษาซึ่งก็คือความเร็วและความถูกต้องในการเขียนในการทดสอบทั้งก่อนเรียนและหลังเรียน

Acknowledgements

As a teacher of writing, I became interested in exploring this elusive concept of “writing fluency,”and ways that it can be measured and improved. Before we can improve something, we need to know what it is and what factors affect it. Due to my background and interest in computer science, I decided to use a data mining operation called linear regression to construct a model of the relationship between the target attribute (a rating of writing) and somevariables that may affect it. I hope that the results obtained from applying this model will be useful to some researchers of L2 writing, teachers of L2 writing, and students who want to improve their writing fluency in a second language. Further, I would like to thank Ajarn Sureerat Marapo, and the research committee, for giving me the opportunity to pursue this research study.

Adam Gardiner

Table of Contents

Chapter 1
Introduction

Background and Significance

Relatively little attention has been given in recent decades to the definition of writing fluency, especially in the context of students writing in a second language, compared to the definition of 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 theimproved speed and reliability of data communications, and the flourishing of social media applications on the web, it is arguable that ESL students are writing muchmore than their peers of twenty years ago.[1] Although some traditional forms of writing such as the handwritten letterhave declined, various kinds of online writing have surged to take their place. These include the established outletsof email, blogs and internet forums, as well asthe more recent outlets of tweeting, status updates and messaging. So while ESL students are 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 in the case of a Facebook status update. So being able to write quickly has become an increasingly important skill. Therefore, being able to write fluently, which in the researcher's view requires both speed and accuracy, has become a more pressing goal for ESL students than it was for their pre-digital peers.Consequently, giving attention to the definition of writing fluency and methods to improve it shouldbe an equally pressing goal for ESL researchers.

Objectives

It was hypothesized (Hypothesis 1) that of the attributes used by the L1 writer to determine a rating (for semi-structured writing about journal topics by L2 students at HCU), the most influential[2] one would be accuracy, that some types of errors would have more effect on the rating (in a negative direction) than other types, and that knowledge of this variation of effects among the error types would reveal an interesting pattern relating to writing fluency. Moreover, it would result in a more accurate and powerful model[3] of writing fluency. Therefore, all of the errors made in the students' paragraphs were assigned to eleven different types. These types were input to a linear regression model, together with the other attributes, in order to discover the variation in effects among different types. This variation in effects would provide the basis for a working definition of writing fluency. The effects were multiplied by the error rates for different types, so that a value of the cumulative effects was obtained. A paragraph that had a greater cumulative effect would receive a lower fluency rating than a paragraph that had a smaller cumulative effect, even if the overall error rate of the paragraphs was identical. So different error types were weighted differently according to the effect they had on the rating.

Hypothesis 2 is 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).This hypothesis emerged from the researcher's experience of noticing an improvement in students' writing fluency on various occasions in the past, and this improvement seemed to be a result of practice in semi-structured writing such as journal topics. So the researcher wanted to determine to what extent this apparent relationship had an objective basis. The method chosen was data mining, especially the operation of linear regression. This method was very data-intensive, as it required collecting the maximum possible amount of detail about each sample of writing. However, only meta-data would be stored in the data mining software. None of the actual content of the writing would be stored. This method was chosen because the researcher was also interested in the wider questions - what do we mean by fluency, and what factors affect it? The linear regression operation is ideal for building a model that would show the relative effects on fluency of different factors; moreover, alternative definitions of fluency could be applied to the same data and compared. The researcher also has the long-term objective of developing an app that would make suggestions based on patterns of errors in a student's writing; in order to do this, it was first necessary to record data about the errors such as their type and frequency.

Therefore, the data mining method was chosen as the most suitable method for testing Hypothesis 1 and Hypothesis 2. While testing these two hypotheses, three alternative metrics of writing fluency would also be compared at the same time. These metrics were called “FLO1”, “FLO3,” and “pure speed.” Two further hypotheses were proposed by the researcher so that the metrics could be compared in this study.

Hypothesis 3 is “that a metric (henceforth referred to as “FLO3”) of writing quality that includes the attribute of writing speed is more useful for rating semi-structured writing of a general and subjective nature than a metric of writing quality that does not include speed. That latter metric is henceforth referred to as “FLO1.”

Hypothesis 4 is: “That a metric (henceforth referred to as “FLO3”) of writing fluency that includes the attributes of both writing speed and accuracy is more useful for rating semi-structured writing of a general and subjective nature than a metric of pure speed.” Both Hypothesis 3 and 4 were tested by applying the three metrics to the semi-structured writing about the journal topics in the pre-test and post-test. In this study, FLO1 includes the same attributes as FLO3, with the exception of speed. “Pure speed” is, of course, speed by itself.

To sum up, this study is similar to some previous studies in that it regards speed as an essential component of writing fluency, but to the best of the researcher’s knowledge this study is the first to model the factors affecting the rating of L2 writing and use this model as a starting point for the definition of writing fluency. Above all, the researcher’s objective is to develop the best possible metric of writing fluency, within the constraints of this particular data collection, and demonstrate some aspects of its usefulness.

Scope of Research

The data collection was from a class of fourth year English-Chinese majors, who were studying Report Writing in English at Huachiew Chalermprakiet University between October 2013 and February 2014. All of the writing for the pre-test and the post-test was done in a supervised classroom environment. The time taken to complete a specified writing task was recorded, as speed of writing is generally agreed to be an essential component of writing fluency. To ensure that the writing output was not affected by factors unrelated to their writing ability (such as having to write about an unfamiliar topic), there was a choice of topics and these were general and subjective in nature.

Definition of Terms

The ACM defines *data mining* as “the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.” (“Data Mining Curriculum: A Proposal”). For this research project I intend to use the data mining operations of a software platform called RapidMiner.

Benefits of the Research

The data mining application will expand our understanding of the problems experienced by ESL learners at Huachiew Chalermprakiet University when engaged in semi-structured writing, and reveal the effects ofsemi-structured writing on their fluency. Further, it will develop a model of writing fluency which will be useful for assessing the progress of ESL writers. Finally, the information extracted by the data mining process can be used to generate new hypotheses for further research.

Chapter 2
Literature Review

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 ‘‘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. Chenowith and Hayes (2001) also used words written per minute to measure fluency.

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” (pg.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”).

Like some of the researchers mentioned above, this researcher also regards writing speed as an essential component of writing fluency. However, this study recorded the speed with more precision than some previous studies. The researcher of this study verified the time that was recorded in the pre- and post-test. He checked that they had recorded an accurate time on the test paper, as each student submitted their completed work.

Regarding lexical factors, this study 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, which refers to the lexical frequency of individual words. However, the linear regression method used did not find support for a dependent relationship between the target attribute and the lexical richness attribute. The lexical richness attribute was also intended as a control variable, to prevent the possibility of a fluency gain arising from the deliberate repetition of sentences or groups of words. An abnormally low value of this attribute would indicate such repetition, but all the values remained within an expected range.

Conceptual Framework

Rather than using any kind of lexical metric in combination with speed as a measure of fluency, the researcher chose the more exacting attribute of accuracy.[4] This attribute 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 for this study, so this program involved weekly writing about journal topics over a longer period of ten weeks.Further, this study used different criteria to define fluency, namely writing speed and accuracy.Nevertheless, writing a blog post and writing about a journal topic are similar in that both are semi-structured[5], subjective and non-technical in nature. Therefore this study's second hypothesis (Hypothesis 2), 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), is similar to Fellner and Apple's study, except for the criterion of FLO3.

Chapter 3
Methodology

Preprocessing

Choice of Software and Operations

The data mining software used for this project is called RapidMiner (see Fig. 1). This application provides a GUI that allows us to 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 two were selected for this study: linear regression and the correlation matrix.

Correlation Matrix The correlation matrix operator can help us to find correlations in the data. By correlations, I mean some kind of statistical relationship that shows *dependence* between two datasets or two attributes within a dataset (Wikipedia, “Correlation”). Unfortunately, it was not possible to complete the correlation matrix analysis due to limitations of the data collection. However, the results of the first stage and some provisional analysis may be perused in the Appendix.

Linear Regression The linear regression operator can help us to model the relationship between attributes and use this model to predict the value of a label attribute.

“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*).

Design of the Classroom Tests

Identification of the Students The data collection, as shown in Table 1, was from a class of fourth year English-Chinese majors, who were studying Report Writing in English between October 2013 and February 2014. Note that “No. of students” refers to the number of students who took both the pre-test and post-test, and were not excluded for other reasons.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 (see page 8, “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.

Selecting the Attributes to be Determined 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.

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

The researcher hypothesized 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, the most influential 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. To test the hypothesis, the first step taken was to record every single error in 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 hypothesized 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 2. 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 (as explained on pages 14-15, "Step 2: Using linear regression to develop a model of FLO1").

5) Lexical Richness

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

6) FLO1 Rating

The FLO1 rating is a measurement of writing quality[6] (and see the discussionon pages 11-12, "Definition and Limitations of FLO1"). The study hypothesizes that this is influenced by various factors, especially the total length and error rates for different types of errors as explained above.

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

The actual word "journal" may be dated, but many of the journal topics are similar to those that might be gleaned from a random sampling of various personal blogs and social media posts that are accessible today.The topics were intended to relate 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 her or his interests. There were four different lists of topics (see Figs. 3 & 4). Two of the lists were given for the pre-test assessment. Then 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 notifications, or problems with the word processing software being used. It also eliminated the risk that someone might copy and paste content from another document or the internet. Further, students vary considerably in their typing speed, when using a computing device. This could cause misleading results for the average writing speed, which is an essential attribute for this study. Students may also vary in how fast they are physically able to write by hand, but this variation is likely to be smaller than it is for typing speed.

4) Use of books and internet

Students were permitted to use dictionaries, either in book or electronic form, to check vocabulary during the assignment.However, reference to any other printed materials was not permitted. Internet access was also not permitted. This was to eliminate the possibility of a student copying content from an online source. Even if a student was able to access the internet, without the instructor's knowledge, the fact that the assignment had to be handwritten would make such copying impractical. Moreover, the subjective and general nature of the topics made it unlikely that any student would want to copy any external information.

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. So they could write freely without the stress of thinking about grades.

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 (Fig. 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). They could use any method for writing their paragraphs, but I recommended that they use an app called Evernote. Some students were already using it as a note-taking tool for their information search. Evernote is supported on most operating systems, including Windows and OS X on desktop computers, and Android and iOS on mobile devices. Finally, in the post-test they were given Lists 3 & 4 (Fig. 4), which they had not seen previously. The conditions of the post-test were identical to those in the pre-test (see page 6, “Conditions of the Classroom Tests”), except that the lists of topics were different.

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 were supposed 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 2) 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 original student paragraphs were handwritten on two pages, as below.

A = Paragraph on the left page	*B = Paragraph on the right page*

A separate table was used to copy selected words from the two paragraphs. The words were copied from the following columns, in this exact order. The reason for alternating between paragraphs A and B was to ensure that approximately an equal number of words were selected from each paragraph.

AR = Rightmost column of paragraph A

BL = Leftmost column of paragraph B

AL = Leftmost column of paragraph A

BR = Rightmost column of paragraph B

AR2 = Second Rightmost column of paragraph A

BL2 = Second Leftmost column of paragraph B

AL2 = Second Leftmost column of paragraph A

BR2 = Second Rightmost column of paragraph B

First, a sample of 50 words was mapped from the paragraph to an empty table (see Table 3) as follows:

Column AR

The rightmost word of the top line of paragraph A was inserted into the top cell of column AR. Then the rightmost word of the second line of paragraph A was inserted into the next cell down of column AR. I continued descending the right edge of the paragraph until I had reached the bottom line, and putting each word into column AR.

Column BL

The leftmost word of the top line of paragraph B was inserted into the top cell of column BR. I continued descending the left edge of the paragraph until I had reached the bottom line, putting each word into column BL.

Columns AL & BR

I repeated the above process for columns AL and BR.

If a total of 50 words had not yet been mapped to the table, then I continued as follows:

Columns AR2, BL2, AL2, BR2

The same process was repeated with these columns, except that the second rightmost or second leftmost words were mapped. The process was stopped when a total of 50 words had been inserted into the table. Depending on the length of the paragraphs, some of the columns might not be filled.

Table 4 is an example of a table after the richness attribute had been calculated. Note that the richness attribute is a score out of 50. It represents the number of unique words in that sample.

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 (see page 5, "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" (page 11).

Creation of New Variables During the processing described above the following new variables were created. These names correspond to the column headings in Excel. Note that "No. of errors/ no of words * 100 (Bigger is less accurate)" is the overall error rate.

SPEED (words per minute)

TOTAL LENGTH (word count)

No. of errors/ no of words * 100 (Bigger is less accurate)

No. of A type errors / no. of words * 100 (= no. of errors per 100 words)

No. of B type errors / no. of words * 100 (= no. of errors per 100 words)

No. of C type errors / no. of words * 100 (= no. of errors per 100 words)
No. of D type errors / no. of words * 100 (= no. of errors per 100 words)
No. of E type errors / no. of words * 100 (= no. of errors per 100 words)
No. of F type errors / no. of words * 100 (= no. of errors per 100 words)
No. of G type errors / no. of words * 100 (= no. of errors per 100 words)
No. of H type errors / no. of words * 100 (= no. of errors per 100 words)
No. of I type errors / no. of words * 100 (= no. of errors per 100 words)
No. of J type errors / no. of words * 100 (= no. of errors per 100 words)
No. of K type errors / no. of words * 100 (= no. of errors per 100 words)
No. of Z type errors / no. of words * 100 (= no. of errors per 100 words)
Vocab x/50
FLO1 = x/200

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 data collection was divided into various tables as explained in 1) and 2) below. Also, two of the tables were appended, as explained in 3) below.

1) The data collection was divided into two tables (the Excel files named "28 Oct 2013 EG 3173 CLEAN WITH 1800 AND 600 SECONDS" and "6 Jan 2014 EG 3173 CLEAN WITH 1800 AND 600 SECONDS"). These tables were only required to extract the value in the column titled "TOTAL TIME TAKEN (seconds)."

2) The data collection was divided into four tables. The pre-test data (28 October 2013) was divided into two tables: one table containing the labelled data (the Excel file named "28 Oct 2013 EG 3173 CLEAN 2 COMPLETE BY RATE ALL ROWS LABELLED FLO3" – see Table 5) and another containing the unlabelled data (the Excel file named "28 Oct 2013 EG 3173 CLEAN 2 COMPLETE BY RATE ONLY ROWS THAT ARE UNLABELLED FLO3"). Also, the post-test data (6 January 2014) was divided into two tables: one table containing the labelled data (the Excel file named "6 Jan 2014 EG 3173 CLEAN 2 COMPLETE BY RATE ALL ROWS LABELLED FLO3" – see Table 6) and another containing the unlabelled data (the Excel file named "6 Jan 2014 EG 3173 CLEAN 2 COMPLETE BY RATE ONLY ROWS THAT ARE UNLABELLED FLO3").

3) Also, the labelled post-test data ("6 Jan 2014 EG 3173 CLEAN 2 COMPLETE BY RATE ALL ROWS LABELLED FLO3") was *appended* to the labelled pre-test data ("28 Oct 2013 EG 3173 CLEAN 2 COMPLETE BY RATE ALL ROWS LABELLED FLO3"), to form a single data table of 44 rows (the Excel file named "28 Oct 2013 AND 6 Jan 2014 EG 3173 CLEAN 2 COMPLETE BY RATE ALL ROWS LABELLED FLO3"). Note that only the labelled data was appended in this way, and not the unlabelled data.

Data Mining

Definition and Limitations of "FLO1"

The term "FLO1" was created by the researcher to refer to:

"The rating of an L2 writer's writing by an L1 writer[7], who uses various attributes to measure that writing's quality, *excluding* the attribute of writing speed." Regarding the first part of this definition, the researcher hypothesized (Hypothesis 1) that of the attributes used by the L1 writer to measure the quality of the L2 writer's writing (for semi-structured writing about journal topics by L2 students at HCU), the most influential one would be accuracy, and that some types of errors would have more effect on the rating (in a negative direction) than other types. Accuracy in this context has both negative and positive aspects. For the former, accuracy implies the avoidance of grammatical errors that may obscure the intended meaning to varying degrees. For the latter, accuracy implies the skillful use of language to convey the intended meaning economically and precisely. However, as it is much easier to measure accuracy in terms of error avoidance, this study will focus on the negative aspect. For this study, the value of FLO1 was decided independently by two native

English-speaking instructors at the same university. As explained on page 10 ("FLO1 Rating"), the instructors were directed to give each set of two paragraphs a score out of 100. A perfect score (100%) should only be given to someone who wrote as an educated native speaker would. Regarding the second part of the definition ("but *excluding* the attribute of writing speed"), the determination of the FLO1 score is influenced by various factors including the length of the writing, but not by the actual speed of writing. This is because the instructors who decided the rating (FLO1) were not aware of the actual speed of writing. For example, supposing Writer A took 10 minutes to write a paragraph of 200 words, while Writer B took 40 minutes to write a paragraph of 200 words, the raters were not aware of this time difference. Supposing that all the other attributes for these two paragraphs were identical, then Writers A and B should receive an identical FLO1 rating. The speed of writing is invisible to the rater and therefore has no effect on the FLO1 rating. Therefore, the rating called "FLO1" is a useful indicator of a student's writing ability, but it is limited because we do not know the actual speed of writing. As mentioned before, speed is a vital component of speaking ability, and I argue that it should also be regarded as a vital component of writing ability. All other factors being equal (such as accuracy of the writing), it is asserted that someone who writes 100 words in 5 minutes is a more proficient writer than someone who writes 100 words in 10 minutes.

Using linear regression to determine "FLO3"

Background to "FLO3" The term "FLO3" was created by the researcher of this report. FLO3 is a metric of writing quality that uses various attributes to measure writing quality, like FLO1. However, unlike FLO1, the attributes used by FLO3 include the actual writing speed. So FLO3 uses the attribute of speed to measure writing quality, in addition to other attributes of the writing.

Methodology used 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. The steps of determining FLO3 are described in detail below.

STEP 1: Adjusting the word count for writers in the pre-test (28 October 2013) and post-test (6 January 2014).

The first step of determining FLO3 was reducing the word count for all those who wrote for longer than 10 minutes, in both the pre-test and post-test. In this case, all of the students in both the pre-test and post-test wrote for longer than 10 minutes, except for one student in the post-test who took exactly 10 minutes.

For step 1, it was necessary to refer to the value stored in the column titled "TOTAL TIME TAKEN (seconds)" (in the Excel files named "28 Oct 2013 EG 3173 CLEAN WITH 1800 AND 600 SECONDS" and "6 Jan 2014 EG 3173 CLEAN WITH 1800 AND 600 SECONDS"). For those students with a value in this column that was greater than 600, the following formula was applied:

Divide 600 seconds by the total time taken (in seconds), and use the result to multiply the word count

In effect, assuming that those writers wrote at a constant speed, this formula determined the number of words that those writers would have written after 600 seconds.

EXAMPLE 1: Supposing a writer wrote 300 words (= the word count) over 2400 seconds. 600/2400 = 0.25, so that writer's word count was multiplied by 0.25. 300 X 0.25 = 75. So the adjusted word count is 75. A value of 75 was entered in the "adjusted word count" column.

EXAMPLE 2: Supposing a writer wrote 100 words (= the word count) over 500 seconds. In this case, the writer had stopped writing before 600 seconds had passed, so this formula was not applied. So the word count was not increased. It remained the same, and a value of 100 was entered in the "adjusted word count" column.

For the data table of 28 Oct 2013, the application of the above formula had the effect of reducing the word count for all of the 22 students, since all of them wrote for a longer time than 600 seconds (see Table 5).

For the data table of 6 Jan 2014, the application of the above formula had the effect of reducing the word count for 21 of the 22 students, namely those who wrote for a longer time than 600 seconds, and not changing the word count for just one of the 22 students, namely one student who wrote for a time that was equal to 600 seconds (see Table 6).

A possible objection is that the formula is unfair to students who wrote for a much longer period of time than the maximum time limit, on the grounds that the longer the time that a student wrote for, his or her word count was reduced by a greater proportion. However, these students had a longer time to write and were therefore likely to have a higher original word count, so this was not an unfair consequence. Supposing Writer A wrote for a longer time than 600 seconds, while Writer B wrote for a time less than or equal to the 600 seconds. If Writer A's writing speed was greater than Writer B's, then his or her word count would still be higher than Writer B's, even after Writer A's word count was reduced according to the formula. The formula simply calculated the number of words that those students (who wrote for longer than 10 minutes) would have written after 600 seconds, assuming a constant speed.

Another possible objection is that the students would tend to write more slowly during the first 10 minutes, because they had to think more about their topics. However, in both the pre- and post-test the students were given 10 minutes preparation time, to think about the topics and prepare for their paragraphs. The timing only started after that preparation stage. The Researcher noticed that all of the students appeared ready to start writing by the end of the preparation stage, in both the pre- and post-test. The topics were all general and subjective, so 10 minutes was considered to be adequate preparation time.

STEP 2: Using linear regression to develop a model of FLO1

1. Overview

The second step of creating FLO3 required using the linear regression operator to develop a model of FLO1 (a rating of writing quality determined by a rater who is unaware of the factor of speed). Linear regression is a statistical method that "attempts to model the relationship between a scalar variable and one or more explanatory variables by fitting a linear equation to observed data" (RapidMiner documentation). This study used the standard linear regression model which used the "ordinary least squares" technique to estimate the relationship between the variables. Developing the linear regression model was a complex and time-consuming process which involved multiple iterations of the linear regression operator with varying sets of inputs, in order to arrive at the best possible model within the constraints of the data sample. Only the two principal iterations before the final model are described here.

2. Objective

To develop the best possible (most accurate) linear regression model, it was necessary in the beginning to input as many explanatory variables as possible to the linear regression operator. Then the detailed results of the linear regression operator had to be analyzed. The most important parts of the results are the *regression coefficients* and the *p-Values.* The coefficients show the size of the effects that the explanatory variables have on the dependent variable, as calculated by the linear regression operator. A larger value of the coefficient means a larger effect. Effects can be either positive or negative. A negative effect means that the value of the dependent variable goes down as the value of the explanatory variable goes up, while a positive effect means that the value of the dependent variable goes up as the value of the explanatory variable goes up. The meaning of the effect is illustrated by the output from RapidMiner's vector linear regression operator.

However, the coefficient in the linear regression output may not indicate a *real* effect of the explanatory variable. Therefore, the coefficient for each variable must be considered in conjunction with that variable's p-Value, to find out the probability of the effect being real instead of a random occurrence:

"The *p*-value is the probability of observing an effect given that the null hypothesis is true whereas the significance or alpha (α) level is the probability of rejecting the null hypothesis given that it is true." (Schlotzhauer 166)

The null hypothesis states that the explanatory variable has absolutely no effect on the dependent variable, and that the "effect" in the output of the linear regression operator does not really exist - it is just a sampling error.

Following common practice in the social science research community, 0.05 was set as the "significance level" (or "alpha level") in this research study. Therefore, only explanatory variables with a p-Value of below 0.05 were considered to be "statistically significant." Consequently, only explanatory variables with a p-Value of below 0.05 were retained in the model. Moreover, the lower the p-Value the better - a p-Value of 0.01 is preferable to a p-Value of say 0.04. The lower the p-Value, the lower the probability that we will reject a true null hypothesis, as shown in Table 7.

It was expected that compromises would be required, due to the constraints of the data sample. Unfortunately, some explanatory variables might have to be excluded from the model, not because it was proven that those variables did not have any effects on the dependent variable, 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 dependent variable.

Using a small sample size for the regression model leads to limitations in the results, especially the limitation of being able to see fewer statistically significant effects from the explanatory variables and these are more likely to be the bigger effects.

Jeff Sauro gives a useful analogy to explain this limitation:

"... statistical analysis with small samples is like making astronomical observations with binoculars. You are limited to seeing big things: planets, stars, moons and the occasional comet." ("Best Practices For Using Statistics On Small Sample Sizes"). However, this does not mean that useful results cannot be obtained from a small sample size: "Galileo, in fact, discovered Jupiter's moons with a telescope with the same power as many of today's binoculars."

Looking at the problem the other way, the higher our expectations from the regression model, in terms of effect size, p-Value and number of explanatory variables, the greater the sample size required. The sample size calculator for multiple regression (see Fig. 2) illustrates this relationship. Specifying a lower effect size, lower p-Value and larger number of predictors (explanatory variables) will all increase the required sample size.

Due to the limitations discussed above, some explanatory variables may have to be excluded from the model. However, it was necessary to begin with the maximum number of explanatory variables, in order to determine which ones were most useful for the model and which ones could be discarded.

3. Initial selection of input variables for the linear regression model

In this section, the initial selection of the input variables to the linear regression model will be explained. During the iterative process of developing the model, some changes were made to the initial selection. Those changes will be explained in the following section.

Scalar variable (or dependent variable)

The FLO1 rating was selected as the scalar variable (which is called the target attribute or label in RapidMiner) because it is hypothesized that the rating of writing quality is influenced by various factors, especially the total length and error rates for different types of errors as explained below.

Explanatory variables (or independent variables or predictor variables)

The following explanatory variables (called regular attributes in RapidMiner) were selected.

1.Total length

The first variable to be selected was the total length (in number of words). All other factors being equal, it was assumed that a higher word count of the two paragraphs would receive a higher FLO1 rating than a lower one.

Note that the speed of writing (words per minute) was not selected as an explanatory variable for modeling FLO1, because in this case the person who decided the rating (FLO1) was not aware of the writer's speed. For example, supposing Writer A took 10 minutes to write a paragraph of 200 words, while Writer B took 40 minutes to write a paragraph of 200 words, the rater was not aware of this time difference. Therefore, when modeling FLO1, the total length was selected as a variable but not the writing speed.

2. Vocab x/50

This is the number of unique words found in the first 50 words (see 1.4.2.5).This explanatory variable was included for two reasons. First, the researcher estimated that the number of unique words would approximately reflect the lexical richness of the paragraphs, and that this attribute would have an effect on the rating. Second, it was included as a control variable. A very low value for this attribute could mean that the writer was deliberately repeating the same sentences or the same words, in which case the paragraphs would be excluded.

3. Rates for types of errors

There was an overall error rate variable (number of errors per 100 words). However, this error rate was further sub-divided into error rates for different types of errors, namely A, B, C, D, E, F, G, H, I, K, Z (see Table 2: Types of errors with descriptions and examples),as it was hypothesized that different types of errors would affect the FLO1 rating to varying degrees (see Hypothesis 1), and that this variation of effects would help to refine the model of fluency. For this reason, these various types of errors (A, B, C, D, E, F, G, H, I, K, Z) were initially selected as explanatory variables, to input to the linear regression operator.

Note that Type "J" was originally included in the list and counted in the students' paragraphs, but due to its very low incidence it was decided to exclude this particular type from the linear regression process.

To sum up, in order to develop the best possible linear regression model, it is desirable to input as many explanatory variables (= regular attributes in RapidMiner) as possible to the linear regression operator initially, even if they have to be excluded later.

Therefore, thirteen explanatory variables were selected for input to the linear regression operator in RapidMiner for the first iteration. The thirteen explanatory variables are: total length, the error rate per 100 words for types A, B, C, D, E, F, G, H, I, K, and Z, and vocab x/50 (the richness attribute).

However, as an additional check, another iteration of the model was run with just three explanatory variables selected as inputs: the total length, no. of errors per 100 words (the overall error rate) and vocab x/50. This iteration allowed the overall error rate to be input, so that its coefficient and p-Value could be compared with that of the individual error types. Also, the results for the total length and vocab x/50 variables would be compared with those from the first iteration. The following section will look at the results that were output for both iterations of the model, and explain the revisions that were made as a consequence.

4. Refining the linear regression model

Before running the iterations described above, the data had to be prepared and input to the linear regression operator as follows.

First, a new file was created so that a single regression model would be applied to both the pre-test data and the post-test data. So the Excel files named "28 Oct 2013

EG 3173 CLEAN 2 COMPLETE BY RATE ALL ROWS LABELLED FLO3" and "6 Jan 2014 EG 3173 CLEAN 2 COMPLETE BY RATE ALL ROWS LABELLED FLO3" were appended, thus creating one table of 44 rows named "28 Oct 2013 AND 6 Jan 2014 EG 3173 CLEAN 2 COMPLETE BY RATE ALL ROWS LABELLED FLO3."

For each iteration in succession, all 44 rows of data were imported from the Excel file named "28 Oct 2013 AND 6 Jan 2014 EG 3173 CLEAN 2 COMPLETE BY RATE ALL ROWS LABELLED FLO3" into RapidMiner.

For iteration 1, the following attributes were selected:

A. The label or target attribute:

"FLO1 = x/200"

B. Thirteen regular attributes:

1. "TOTAL LENGTH (word count)"
2. "Vocab x/50"
3. "No. of A type errors ... per 100 words"
4. "No. of B type errors ... per 100 words"
5. "No. of C type errors ... per 100 words"
6. "No. of D type errors ... per 100 words"
7. "No. of E type errors ... per 100 words"
8. "No. of F type errors ... per 100 words"
9. "No. of G type errors ... per 100 words"
10. "No. of H type errors ... per 100 words"
11. "No. of I type errors ... per 100 words"
12. "No. of K type errors ... per 100 words"
13. "No. of Z type errors ... per 100 words"

The target attribute was not blank for any of the 44 rows. Therefore, this data was named "3075 linear regression data 28 Oct 13 and 6 Jan with all rows labelled (attribute FLO1 in here = to predict FLO3) 13 pvs" in RapidMiner.

For iteration 2, the following attributes were selected:

A. The label or target attribute:

"FLO1 = x/200"

B. Three regular attributes:

1. "TOTAL LENGTH (word count)"
2. "Vocab x/50"
3. "No. of errors/ no of words * 100 (Bigger is less accurate)"

Note that no. 3 above is the *overall* error rate.

The target attribute was not blank for any of the 44 rows. Therefore, this data was named "3074 linear regression data 28 Oct 13 and 6 Jan with all rows labelled (attribute FLO1 in here = to predict FLO3) 3 pvs" in RapidMiner.

The results that were output by the linear regression operator, for iteration 1 (13 regular attributes) and iteration 2 (3 regular attributes) can be seen in Tables 8 and 9 respectively. The results of both iterations will be compared for each attribute.

1. Total Length

As can be seen from Table 8 above (for iteration 1), the regression coefficient of "TOTAL LENGTH" is +0.252, which means that every time this variable increases by 1, the scalar variable (the FLO1 rating) increases by +0.252.

For the second iteration (see Table 9) the regression coefficient is nearly the same, +0.254. RapidMiner's vector linear regression operator (applied to exactly the same variables and data) shows the meaning of the effect in simple terms:

Vector Regression (for second iteration)

FLO1 = x/200 = 0.254 * TOTAL LENGTH (word count) - 0.040 * Vocab x/50 - 1.291 * No. of errors/ no of words * 100 (Bigger is less accurate) + 98.055

For both iterations (see Table 8 and Table 9), the p-Value of "TOTAL LENGTH" according to RapidMiner is 0, which means that the probability of this particular effect occurring with a true null hypothesis is 0. Note that a p-Value of 0 in RapidMiner may not indicate a genuine zero.[8] However, even if it is not a genuine zero, then it is so extremely small that it can safely be assumed to be zero for this study. So the null hypothesis can confidently be rejected for TOTAL LENGTH. This variable was retained for the final regression model. Of course, it was critical for this study that the TOTAL LENGTH variable had a very low p-Value, as an adjusted value of this variable was used to calculate predictions for FLO3. If the value had not been below 0.05, it would have been necessary to reject this method of calculating FLO3.

2. Vocab x/50

For the first iteration (see Table 8), the p-Value of "Vocab x/50" is 0.417, which is above the statistical significance level of 0.05. For the second iteration (see Table 9), it is even higher at 0.877. Therefore this variable was removed from the regression model.

3. Error Type rates AND Overall Error Rate

Note that there are 11 individual error type variables, which were input to the first iteration (see Table 8). There is also an overall error rate variable ("No. of errors/ no of words * 100 (Bigger is less accurate)")that was input to the second iteration (see Table 9).

Result from iteration 1 (11 error type variables): As can be seen fromTable 8, the following error type variables have a p-Value higher than the statistical significance level of 0.05: No. of H Type errors, No. of D Type errors, No. of K. Type errors, No. of B Type errors, No. of Z Type errors, No. of F Type errors, so they are not statistically significant. Therefore, these six error type variables were removed from the regression model. The other error type variables (for types A, I, C, G and E) have a p-Value lower or equal to 0.05, so were retained in the model.Note that this particular result does not prove that there exists no causal relationship between the scalar variable and any of the excluded error type variables. It just means that the linear regression operation on this particular sample of data is insufficient to disprove the null hypothesis (that there is no effect) for them. It is possible that running exactly the same operation on a larger sample of data would yield lower p-Values for these error type variables. In this case, the part of the data that we are concerned with is the number of errors belonging to each type. For the five error types that were retained in the model, the regression coefficients range from -1.643 (for A type errors) to -8.192 (for C type errors which are adjective/participle errors – such as "I am boring with this movie"). That the coefficients are negative is expected, as the FLO1 rating is expected to decrease as the error rate (for any type) increases. So for every time the C type error rate increases by 1, the scalar variable (the FLO1 rating) decreases by 8.192. It is interesting that this type of error should have the largest effect (out of the five statistically significant effects), as it is one of the more conspicuous errors made by Thai students writing in English. Possible reasons for the variation in the effect will be discussed in the conclusions.

Result from iteration 2 (just a single overall error rate variable)

As can be seen from Table 9, the p-Value of "No. of errors/ no of words * 100 (Bigger is less accurate)" is "0.000." This is even lower than the p-Values of the error types A, I, C, G and E (see Table 8), although not quite as low as 0 (in RapidMiner numerical syntax). It shows that the null hypothesis can be confidently rejected for this variable, just as it was for TOTAL LENGTH. The regression coefficient is negative, as expected, -1.291. However, it was decided to use a model that contained the five error types A, I, C, G and E, as one of the hypotheses was that the different error types affected fluency to varying degrees. Writers who had a higher error rate for those error types which had a greater effect on the dependent variable (which is FLO1), would receive a lower FLO3 rating than writers who had a lower error rate for those error types, even if their overall error rate was the same. A model that only used a single error rate variable would assume that all error types had equal effects, which is not consistent with the results of the process "3075 linear regression process 28 Oct 13 (to predict FLO3) 13 regular attributes." Using the five error types instead of the overall error rate would return more precise results. Writers who had a higher error rate for those error types which had a greater effect would receive a lower fluency rating than writers who had a lower error rate for those error types.

5. Outcome of the Refinement

At the end of the refinement process, the label or target attribute ("FLO1 = x/200") remained unchanged as the label. Seven regular attributes were excluded for reasons explained in the previous section. The following six regular attributes were retained in the regression model:

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”

Like the previous iterations, all 44 rows of data were imported from the Excel file named “28 Oct 2013 AND 6 Jan 2014 EG 3173 CLEAN 2 COMPLETE BY RATE ALL ROWS LABELLED FLO3” into RapidMiner. Unlike the previous iterations, a different set of attributes was selected for the import, namely the target attribute and the six regular attributes that were retained in the regression model. Therefore the following attributes were selected for the import:

A. The label or target attribute:

“FLO1 = x/200”

B. 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”

The target attribute was not blank for any of the 44 rows. Then, this data was named “3076 + 3086 linear regression data 28 Oct 13 and 6 Jan with all rows labelled (attribute FLO1 in here = to predict FLO3) SIX pvs” in RapidMiner. Note that the data set name begins with “3076 + 3086” because this data was input to a single model that would subsequently (in Step 3) be applied to both the pre-test data (by the process “3076 linear regression process 28 Oct 13 (to predict FLO3) SIX regular attributes”) and post-test data (by the process “3086 linear regression process 6 Jan 14 (to predict FLO3) SIX regular attributes”). However, here in Step 2, we are still developing the model and not applying it yet. Below, the process “3076 linear regression process 28 Oct 13 (to predict FLO3) SIX regular attributes” is only being used to output the model and not apply it yet. Actually, the process “3086 linear regression process 6 Jan 14 (to predict FLO3) SIX regular attributes" could also be used to output the model and the results would be the same. Tables 10 and 11 show the results that were output by the linear regression operator, when the process "3076 linear regression process 28 Oct 13 (to predict FLO3) SIX regular attributes" was run.

1. Total Length

As can be seen in Table 10, the p-Value of "TOTAL LENGTH" according to RapidMiner is still 0, which means that the probability of this particular effect occurring with a true null hypothesis is 0. The regression coefficient of "TOTAL LENGTH" is now +0.233, which means that every time this variable increases by 1, the scalar variable (the FLO1 rating) increases by +0.233. It is slightly smaller than the coefficient output from processes 3074 (0.254) and 3075 (0.252). This is to be expected due to the adjustment of the regression model.

2. Error Type rates

As can be seen in Table 10, the p-Values of error types A, I, C, G and E are not identical to their p-Values from processes 3074 and 3075. However, this is to be expected due to the adjustment of the regression model. The differences are only slight. Most importantly, all five error types still have p-Values lower than the significance level of 0.05. Their regression coefficients are also slightly different from those output by process 3075, but their relative positions are still the same. The coefficients range from -1.773 for A type errors (-1.643 in process 3075) to -7.14 for C type errors (-8.192 in process 3075).

To sum up Step 2, the linear regression operator in RapidMiner was used to develop the best possible model of FLO1 within the constraints of the data sample. In the next step, this model will be applied to predict the values of FLO3 (a metric of writing quality that uses various attributes including speed).

STEP 3: Applying the final model of FLO1 to predict the values of FLO3.

1. Overview

The final model output by linear regression in Step 2 can now be used to predict values of the "label" (the scalar variable) for instances where the value is unknown. In this case, the model is applied to data that has been adjusted to reflect the speed of writing, namely an adjusted word count as explained in Step 1. The outcome of applying the model will be predictions of the "label" (the scalar variable). These predictions will reflect the speed of writing and thus become the values of FLO3. The following sections will explain exactly how RapidMiner's linear regression operator was used to apply the model of FLO1 to unlabeled data of the pre-test and post-test and then make predictions for FLO3.

2. Applying the Linear Regression method to the data from 28 October 2013 (pre-test), to predict the values of FLO3

The process "3076 linear regression process 28 Oct 13 (to predict FLO3) SIX regular attributes" has already been used to output the model. Now it is used to both output the model and apply it to the pre-test data. First, the unlabeled pre-test data (22 rows) was imported from the Excel file ("28 Oct 2013 EG 3173 CLEAN 2 COMPLETE BY RATE ONLY ROWS THAT ARE UNLABELLED FLO3") into RapidMiner. Then, the same attributes were selected as for the labeled data:

A. The label or target attribute:

"VALUE OF FLO3 = left blank because it is the LABEL"

Note that this attribute has a different title from the labeled file (where its title is "FLO1 = x/200"), but it is the same Excel column. First FLO1 is modeled, then the model is applied to predict the values of FLO3 which will appear in the same column.

B. 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"

However, the unlabeled pre-test data had two differences from the pre-test data (22 rows) within the labeled data set (44 rows). First, the value for the total length was replaced with the adjusted word count that was obtained by applying the formula, as explained in Step 1. Second, the target attribute was left blank in all 22 rows. Therefore, this data was named "3076 linear regression data 28 Oct 13 with all rows

UNlabelled (attribute FLO1 in here = to predict FLO3) SIX pvs" in RapidMiner. So the linear regression model is being applied to the same group of students again, but with a different value for one of the input attributes (the word count). Effectively, the linear regression operator is being used to answer the following question: What would the target attribute (FLO3) have been if the word count was equal to the adjusted word count, or in other words if all the students had stopped writing after exactly 10 minutes what would their target attribute (FLO3) have been? Then the process "3076 linear regression process 28 Oct 13 (to predict FLO3) SIX regular attributes" was run in RapidMiner. First, the Retrieve Operator retrieved the labeled data set ("3076 + 3086 linear regression data 28 Oct 13 and 6 Jan with all rows labelled (attribute FLO1 in here = to predict FLO3) SIX pvs") from the repository. Then, the same data set was input to the Linear Regression operator as an Example Set or Training Set. The Linear Regression operator applied the linear regression algorithm to the data set, and output coefficients for the selected attributes (see Table 10). It also output a regression model, which was input to the Apply Model operator. This operator applies an already learnt or trained model to an Example Set. So the Apply Model operator applied the regression model to the unlabeled data set "3076 linear regression data 28 Oct 13 with all rows UNlabelled (attribute FLO1 in here = to predict FLO3) SIX pvs" which is input to the Apply Model operator. The unlabeled data set was updated by the Apply Model operator in the following way: the predicted values of the label attribute (FLO3 in this case) were added, as shown in Table 12. The workflow of the process "3076 linear regression process 28 Oct 13 (to predict FLO3) SIX regular attributes" is shown in Fig. 5.

3. Applying the Linear Regression method to the data from 6 January 2014 (post-test), to predict the values of FLO3

Next, the process "3086 linear regression process 6 Jan 14 (to predict FLO3) SIX regular attributes" is used to both output the model and apply it to the post-test data. First, the unlabeled post-test data was imported from the Excel file (named "6 Jan 2014 EG 3173 CLEAN 2 COMPLETE BY RATE ONLY ROWS THAT ARE UNLABELLED FLO3") into RapidMiner. Then, the same attributes were selected as for the labeled data:

A. The label or target attribute:
"VALUE OF FLO3 = left blank because it is the LABEL"
Note that this attribute has a different title from the labeled file (where its title is "FLO1 = x/200"), but it is the same Excel column. First FLO1 is modeled, then the model is applied to predict the values of FLO3 which will appear in the same column.
B. 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"

However, the unlabeled post-test data had three differences from the post-test data (22 rows) within the labeled data set (44 rows). First, unlike the labeled post-test data which consisted of 22 rows, the unlabeled post-test data consisted of 21 rows. These rows represented the 21 students whose word count was changed after the formula was applied. There was one student who wrote for exactly 600 seconds, whose word count was not changed. Second, the value for the total length was replaced with the adjusted word count that was obtained by applying the formula, as explained in STEP 1 (page 13). Third, the target attribute was left blank in all 21 rows. Therefore, this data was named "3086 linear regression data 6 Jan 14 with all rows UNlabelled (attribute FLO1 in here = to predict FLO3) SIX pvs" in RapidMiner. So the linear regression model is being applied to the same group of students again, but with a different value for one of the input attributes (the word count). Effectively, the linear regression operator is being used to answer the following question: What would the target attribute (FLO3) have been if the word count was equal to the adjusted word count, or in other words if all the students had stopped writing after exactly 10 minutes what would their target attribute (FLO3) have been? Then the process “3086 linear regression process 6 Jan 14 (to predict FLO3) SIX regular attributes” was run in RapidMiner. First, the Retrieve Operator retrieved the labeled data set (“3076 + 3086 linear regression data 28 Oct 13 and 6 Jan with all rows labelled (attribute FLO1 in here = to predict FLO3) SIX pvs”) from the repository. Then, the same data set was input to the Linear Regression operator as an Example Set or Training Set. The Linear Regression operator applied the linear regression algorithm to the data set, and output coefficients for the selected attributes (see Table 10). It also output a regression model, which was input to the Apply Model operator. This operator applies an already learnt or trained model to an Example Set. So the Apply Model operator applied the regression model to the unlabeled data set “3086 linear regression data 6 Jan 14 with all rows UNlabelled (attribute FLO1 in here = to predict FLO3) SIX pvs” which is input to the Apply Model operator. The unlabeled data set was updated by the Apply Model operator in the following way: the predicted values of the label attribute (FLO3 in this case) were added, as shown in Table 13. The workflow of the process “3086 linear regression process 6 Jan 14 (to predict FLO3) SIX regular attributes” is shown in Fig. 6.

Chapter 4
Results

Results and Findings

Comparison of the Pre-test and Post-test

The values of the FLO3 predictions (see Tables 12 and 13)were used to create a newExcel table (Table 14), together with the FLO1 ratings and speeds already obtained. Then the five error rates (for types A,C,E,G,I) were used for calculating the effects, as will be explained in the following section.

1) FLO1 (x/200)

As can be seen from Table 14, 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 coefficient (see Table 10). 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 14,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. The effect size is a more useful result than the overall error rate, as it gives more weighting to errors that have more effect on the rating. However, it is also interesting 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.

Review of the results

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

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 14**,** 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.

Using Analysis by Quartiles to Compare the Different Metrics (FLO1, FLO3 and pure speed)

OverviewUnlike the linear regression and correlation matrixoperators, “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 (seeTable 14)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 forinteresting patterns in the quartile distribution rather thanstatistical trends.Thus the second part of Hypothesis 1 will be tested.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, and thus provide a way to test Hypotheses 3 and 4. First, six new tables were created from the table showing the results of the linear regression (see Table 14). These new tables contained exactly the same data, but were ordered differently. The original table of results was ordered by the reference code (from B1to 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.

Creating a Quartiles Table for FLO1 The first two new tables were ordered by FLO1. One of the tables was ordered by the column named “PRETEST FLO1 SCORE (x/200)” and the other was ordered by the column named

"POSTEST FLO1 SCORE (x/200)."So, in both the pre-test and post-test table, a sort was applied to data in the column containing the FLO1 score, so that the students were ordered by FLO1– with the highest value of FLO1 at the top and lowest value of FLO1 at the bottom.Then the data table was divided into quartiles, using the following method (also known as Tukey's hinges):

- *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")

After applying the above method to both data tables, the first quartile or lower quartile (Q_1) consisted of the six students whose value of FLO1 was lowest and the third quartile or upper quartile (Q_3) consisted of the six students whose value of FLO1 was highest.Therefore, the upper quartile contained the writers who wrote the paragraphs with the highest FLO1 ratings and the lower quartile contained the writers who wrote the paragraphs with the lowest FLO1 ratings.

Finally, both tables were appended horizontally, so that the left-hand side showed the results of the pre-test as ordered by FLO1, and the right-hand side showed the results of the post-test ordered by FLO1. The quartiles table for FLO1 can be seen in Table 15.

Creating a Quartiles Table for FLO3 Another two new tables were ordered by FLO3. One of the tables was ordered by the column named "PRETEST FLO3 SCORE (x/200)" and the other was ordered by the column named "POSTEST FLO3 SCORE (x/200)."So, in both the pre-test and post-test table, a sort was applied to data in the column containing the FLO3 score, so that the students were ordered by FLO3– with the highest value of FLO3 at the top and lowest value of FLO3 at the bottom.Then the data table was divided into quartiles, using the method known as Tukey's hinges (described in "Creating a Quartiles Table for FLO1"above).After applying this method to both data tables, the first quartile or lower quartile (Q_1) consisted of the six students whose value of FLO3 was lowest and the third quartile or upper quartile (Q_3) consisted of the six students whose value of FLO3 was highest.Therefore, the upper quartile contained the writers who wrote the paragraphs with the highest FLO3 ratings and the lower quartile contained the writers who wrote the paragraphs with the lowest FLO3 ratings.Finally, both tables were appended horizontally, so that the left-hand side showed the results of the pre-test as ordered by FLO3, and the right-hand side showed the results of the post-test ordered by FLO3. The quartiles table for FLO3 can be seen in Table 16.

Creating a Quartiles Table for Speed Another two new tables were ordered by pure speed. One of the tables was ordered by the column named “PRETEST SPEED (wpm)” and the other was ordered by the column named “POST TEST SPEED (wpm).” So, in both the pre-test and post-test table, a sort was applied to data in the column containing the speed, so that the students were ordered by speed – with the highest value of words per minutes at the top and lowest value of words per minute at the bottom.Then the data table was divided into quartiles, using the method known as Tukey’s hinges (described in “Creating a Quartiles Table for FLO1” above). After applying this method to both data tables, the first quartile or lower quartile (Q_1) consisted of the six students whose speed was lowest and the third quartile or upper quartile ($Q_{3)}$ consisted of the six students whose speed was highest.Therefore, the upper quartile contained the writers who wrote the paragraphs at the fastest rate and the lower quartile contained the writers who wrote the paragraphs at the slowest rate.Finally, both tables were appended horizontally, so that the left-hand side showed the results of the pre-test as ordered by speed, and the right-hand side showed the results of the post-test ordered by speed. The quartiles table for speed can be seen in Table 17.

Implications

The following discussionuses 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 (Table 15), 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 (Table 16), 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 (Table 17), 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.

First, the above results will be analyzed in relation to the second part ofHypothesis 1. This part states that "knowledge of variation of effects among the error types would reveal an interesting pattern relating to writing fluency." The variation of effects amongthe error types has already been incorporated into the quartile charts (the effects are the result of multiplying the error rates of five error types by their regression coefficients and then adding them together, as explained earlier in "Comparison of the Pre-test and Post-test"). The quartile charts suggest that regular semi-structured writing may have benefitted the accuracy (in relation to these 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 theleast 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.

The improvement in accuracy was especially noticeable when ordered by speed (average effects decreased from 0.63 to 0.42), while there was a slight improvement when ordered by FLO3 (average effects decreased from 0.72 to 0.68). Of course, the members of the lower quartiles were not exactly the same in the post-test as in the pre-test, so there could be two factors behind the lower quartile's improvement in accuracy:several of the slowest writers improving their accuracy in the post-test (staying in the lower quartile for both tests), and several of the more accurate writers getting slower in the post-test (moving down to the lower quartile in the post-test).

In fact, if we look at the results again (Table 17) it can be seen that what really happened is not quite so straightforward. Two of the slowest writers (B3 and

B19) stayed in the lower quartile for both tests, but they both (B3) became less accurate. However, despite these two, there was an overall decrease in effects (=improvement in accuracy) for the lower quartile. Of the other four writers (B17,B15, B14, B7) in the lower quartile in the post-test, all four had indeed moved down to the lower quartile in the post-test. And three were indeed slower in the post-test, but B17 was actually faster – yetB7 still moved down to the lower quartile because the average speed of the whole class had increased for the post-test. Three of the writers (B17, B15, B17) actually became less accurate in the post-test, while only B14 gained in accuracy. Nevertheless those four writers, who moved down to the lower quartile, were still more accurate than the four writers (B13, B16, B20, B21) who were in the lower quartile for the pre-test and moved up in the post-test.Consequently, the average total effects still decreased significantly for the lower quartile in the post-test.Therefore, we have eventually found the main reason for the improvement in the accuracy of the lower quartile when ordered by speed–four of the slowest and least accurate writers in the pre-test (B13, B16, B20, B21) have become faster in the post-test and so moved out of the lower quartile. Three of these gained in accuracy too (only B13 declined in accuracy). They were replaced by four writers(B17, B15, B14, B7) who became slower in the post-test (except for one) and so moved down to the lower quartile. These four writers all had a lower value for total effects than those who they replaced in the lower quartile, even though three of them declined in accuracy in the post-test.

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 (see Table 15).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 indeed 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” (pages 23-24)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 additionalobservation 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.

Next, the results will be briefly analyzed in relation to Hypothesis 3 and 4. When comparing the usefulness of different metrics, the key question is which metric is most useful for capturing the important aspects of the kind of writing being assessed. In this case, the study assumes that the important aspects are the improvements in speed and accuracy that occurred between the pre- and post-test. If one accepts this study's underlying assumption, then FLO3 is clearly more useful than FLO1 (re: hypothesis 3). If we referred only to the FLO1 chart, then one would only be able to conclude that the overall quality of the writing had declined, despite the accuracy gain of the upper quartile.One might also conclude that the least able students, the ones in the lower quartile, had not improved in any way. By contrast, the FLO3 chart shows a slight "evening out" of the quartiles that reveals the improvement of the lower quartile. It captures an important aspect that is hidden by the FLO1 chart, since FLO1 is not influenced by the attribute of speed. By contrast, FLO3 integratesspeed with accuracy and thus can capture both important aspects of the writers' performance.

FLO3 also appears to be a more useful metric than pure speed (re: hypothesis 4) since obviously pure speed can only capture one of the important aspects. The most useful point of the pure speed chart is that it does show an extreme "evening out" (i.e. the average effects for the lower and upper quartiles are almost the same in the post-test (0.42 and 0.43 respectively), so it shows the improvement in the lower quartile even more than FLO3. Further, in this case at least, it shows thatthe fastest writers are not less accurate than the slowest writers, thus dispelling a fear often expressed by critics of semi-structured writing. The detailed analysis carried out earlier (pages 27-28), of the exact changes that occurred in the lower quartile between the pre- and post-test when ordered by speed, also adds weight to dispel this fear.Again, the quartile sizes are too small to establish a statistical trend here.

However, this extreme “evening out” is also the weakest point of pure speed. A very accurate writer is just as likely to end up in the lower quartile as a very inaccurate writer is likely to end up in the upper quartile, since speed is the only criterion of fluency in this chart. In fact, table 17 shows several writers in the upper quartile with higher total effects than several writers in the lower quartile. Likewise, any writers who improved in accuracy in the post-test, but got slower, would probablydescend the chart, while any writers who got faster in the post-test, but declined in accuracy, would probablyascend the chart. So, the important aspect of accuracy is hidden by the pure speed chart.

Chapter 5
Conclusion

The first part of Hypothesis 1 states that of the attributes used by the L1 writer to determine a rating (for semi-structured writing about journal topics by L2 students at HCU), the most influential one would be accuracy, and that some types of errors would have more effect on the rating (in a negative direction) than other types.

The results of the linear regression from process "3076 linear regression process 28 Oct 13 (to predict FLO3) SIX regular attributes" show that the five error type variables with a p-Value of less than 0.05 did have a greater effect on the rating than the TOTAL LENGTH attribute. Not surprisingly, accuracy is the most important attribute of those that were input to the regression model. These effects can be seen again in Table 10. So accuracy has been shown to be the most influential attribute out of those attributes that were input to the model, for this specific data collection with these specific raters.

Next, as predicted by Hypothesis 1, 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. What factor determined this variation of the effects? Answering this question is outside the scope of this study, so it is discussed below in "Further Research."

So the first part of Hypothesis 1 has been partially substantiated. The second part of Hypothesis 1 states that knowledge of this variation of effects among the error types would reveal an interesting pattern relating to writing fluency. After the effects of five error types were used (together with the Total Length attribute) by the regression operator to calculate the predictions for FLO3, the researcher then manually calculated the combined effects of these five error types to reveal the pre to post-test changes in accuracy levels for the lower and upper quartiles, and enable a comparison of the usefulness of three different metrics. Analysis of these quartiles charts did indeed reveal an interesting pattern, namely that the least fluent writers (the lower quartile as measured by FLO3 or pure speed) accomplished a greater improvement in either speed or accuracy (or both) from the pre-test to the post-test than the writers outside the lower quartile, although this pattern is much more emphatic when ordered by speed. However, the quartile sizes are too small to establish the existence of a statistical trend.

Hypothesis 2is 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 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”(pages 23-24), the students showed an overall improvement in both writing speed and accuracy (as defined by the total effects for error types A,C,E,G and I) from the pre- to the post-test.

Hypotheses 3 and 4 assert the superior usefulness of the metric called FLO3 for rating semi-structured writing of a general and subjective nature, compared to FLO1 (which excludes speed) and pure speed respectively. Here the study assumes that the best metric is the one that most accurately reflects the changes that have occurred in the writing between the pre- and the post-test. FLO1 is very useful for assessing some kinds of writing, such as a homework assignment for a structured essay, where speed is not important. Pure speed may also be a very useful metric in certain situations. However, in this study, FLO3 is shown to be more useful than both FLO1 and pure speed for assessing the semi-structured writing of the pre-test and post-test since it better captures the two most critical aspects of the change in the students' writing from the pre-test to the post-test.That speed is one of those critical aspects is an assumption based on the observed nature of semi-structured writing, which in practice is usually time-driven and interactive. That accuracy is the other critical aspect has already beenshown by the results obtained from testingthe first part of Hypothesis 1.

Therefore, the researcher considers that this study has achieved its objective of developing the best possible metric of writing fluency, within the constraints of this particular data collection, and some aspects of its usefulness have been demonstrated.

In the following section, these useful aspects will be explored in greaterdetail.First, in relation to hypothesis 1, the results show the usefulness of applying a metric of writing fluency that gives primacy to the attribute of accuracy and assigns varying weights to different types of writing errors.Such a metric is called FLO3 in this study.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. Although the quartile sizes are small, the above results suggest that regular practice in semi-structured writing of a general and subjective nature is likely to improvethe accuracy of most L2 writers at the university level, especially the accuracy of the least fluent writers. A possible explanation for this disproportionate effect on the least fluent writers is that this group of writers is more affected by anxiety during writing than their more fluent colleagues, and that this anxiety has a negative effect on both their writing speed and accuracy. However, they experience less anxiety when engaged in the subjective and more flexible nature of semi-structured writing, than they experience when engaged inmore objectiveorstructured writing tasks. When writing about journal topics, they are less likely to run out of ideas about what to write next, since the topics relate to their own experience, and less likely to worry about whether they are conforming to the required structure. So, when engaged in writing about journal topics, the least fluent writers are less anxious than they normally are when doing writing tasks. It is suggested that regular practice in this less anxiety-inducing type of writing initiates a feedback loop.The cumulative reduction in their level of anxiety causes an improvement in their self-perceived writing output (more words at a faster rate)which in turn furtherreduces their anxiety and increases their self-confidence as writers. In the researcher’s experience, writers become less likely to make errors as their anxiety-level decreases.Therefore, this anxiety factor would explain why regular practice in semi-structured writing would benefit the accuracy of most L2 writers, especiallythat of the leastfluent writers.Second, in relation to hypothesis 2, the pre- and post-test results show the usefulness of regular practice in semi-structured writing in terms of improving the writing fluency (as measured by FLO3) of all quartiles of L2 students at HCU. Since FLO3 measures both speed and accuracy, one would expect to see improvements in both speed and accuracy, which areindeed shown by the pre- and post-test results.Third, in relation to hypotheses 3 and 4, the results show that the metric called FLO3 is more useful than both FLO1 and pure speed for assessing student performance in semi-structured writing.

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. Nevertheless, the researcher considers it unlikely that some kind of lexical attribute or any other attribute would be shown to be more influential on the rating than accuracy, for this particular type of writing.

Further Research

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 seek a conclusive answer to this question, and besides a lot more data is needed including some samples of writing by native speakers, but two possible explanations will be suggested by the researcher to account for this variation in effects.

One possible explanation that the rater gives a weighting to errors according to their degree of *apparent divergence*[9] from the writing of a native speaker. "Apparent" is a key word here, as in this study and in most work situations, the rater is working under pressure and only has limited time to check each piece of writing. Some types of errors are more noticeable and hard to miss, while other types of errors may be equally divergent but less noticeable.

Another possible explanation is that the rater gives a weighting to errors according to the degree that they affect the comprehensibility of the writing.[10] One problem here is that comprehensibility is a complex and multi-layered attribute depending not just on the type of error, but on the context. Some types of errors may have little impact on comprehensibility in one sentence, but a large impact in another sentence. Also, a trivial error occurring alone in a sentence may have little impact on the meaning, but two trivial errors following each other could have a big impact. Comprehensibility may also be affected by logical errors. Two successive sentences may be totally accurate from a linguistic point of view, but not logically related. Consequently, those sentences are incomprehensible when considered together.

Of course, the above explanations are not incompatible. The variation in effects may be accounted for by multiple factors. 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 attribute with the lowest p-Value (see Table 10). 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.

Recommendations

Consequently, the researcher recommends that more recognition and attention be given to semi-structured writing in English-Chinese major writing courses, especially in the classroom environment, and that a metric like FLO3 is used to assess classroom-based semi-structured writing. This would complement the existing teaching and assessment of structured writing forms such as various types of essays. To support these recommendations, the advantages and disadvantages of the following three writing assessment strategieswill be discussed: homework-based structured essays assessed by FLO1, homework-based semi-structured writing tasks assessed by FLO1, and classroom-based semi-structured writing tasks assessed by FLO3.

In the first strategy of structured essays that are assigned as homework and assessed by a metric like FLO1, students can choose from many topics, although the number of topics given depends on the instructor. The main benefit of this type of writing is that the students learn how to write in a logical, structured way. They also learn how to write many kinds of essay such as descriptive, comparison/contrast, process, cause-effect and opinion. However, the time taken to write the essay is not recorded, so we do not know how quickly the essay was written. For example, two structured essays that are similar in length may contain very few errors, but one was written quickly by a student with excellent grammar, while the other was written very slowly and corrected many times by a student with weak grammar.Using a non speed-based metric like FLO1 to assess fluency, we might conclude that both students were equally fluent writers and not be aware of various problems affecting the slow writer. Another disadvantage is that somestudentsmay write very short essays. This short lengthwill result in their receiving a low score using the FLO1 metric, but we will have limited information about their specific problems. A very short essay could have beenwritten by a fast writer who wrote only a little because he or she was not interested in the topic, or by a very slow writer who ran out of time. In either case, we have limited information about the specific problems of the writer, because so little has been written. Yet another disadvantage is that students may resort to plagiarism for part or all of the essay, especially if they are not interested in the topic or find it too difficult. As a consequence, they will either receive a very high score using the FLO1 metric, if the plagiarism is not detected, or a very low score if it is detected. In either case, we will have at best limited information about their problems, or absolutely nothing if the entire essay is copied.

In the second strategy of semi-structured writing tasks (such as journal topics) that are assigned as homework and assessed by a metric like FLO1, students usually have more choice of topics than fora structured essay. The topics are usually easierto write about than those of a structured essay, since they usually relate to the student's own ideas or experiences. Also, although semi-structured writing is required to be coherent and logical, structure is not so important as in a structured essay. In the researcher's experience,even the weakest students write more in semi-structured writing tasks than they do for structured essays, due to the nature of the topics, and the less rigid structure. They are less likely to stop writing because they are bored by the topic or find it too demanding. Consequently, there is more writing to be assessed and specific problems are more apparent. Further, in the researcher's experience, it is rare for students to plagiarize when they are writing about subjective topics. It is certainly less likely than when they are writing about objective topics. However, semi-structured writing tasks assigned as homework shares the disadvantage of structured essays assigned as homework, namely that the time taken to write the task is not recorded, so we do not know the speed of the writing. Two samples of semi-structured writing that are similar in length and contain very few errorsmay receive the same FLO1 score, even though one took only ten minutes to write, while the other took an hour. We might conclude that both writers were equally fluent, and not be aware of the problems affecting the slow writer.

In the third strategy of classroom-based semi-structured writing tasks assessed by a metric like FLO3, combines the advantages of homework-based semi-structured writing tasks, namely that the students write more and are less likely to copy material, with the advantage ofusing FLO3 as the metric for assessment. It is not practical to expect students to record the exact time taken for their writing tasks outside the classroom, so homework-based semi-structured writing tasks have to be assessed by a non speed-based metric with its inherent disadvantages as already mentioned. In contrast, it is practical to record the time taken for classroom-based semi-structured writing tasks, so they can be assessed by a speed-based metric like FLO3. The exact time can be recorded for classroom-based semi-structured writing tasks, so writing fluency can be measured more precisely. The student's average writing speed is measured, which is not measuredat all in homework-based semi-structured writing. Also, the error rate obtained from classroom-based semi-structured writing tasks more accurately reflects the students' authentic writing, compared to the error rate obtained from homework-based semi-structured writing tasks.This is because the writers have less time to accomplish the classroom tasks, compared to tasks done outside the classroom, so they have less time to correct their errors by checking external sources such as websites or apps. Besides, the instructor can impose restrictions on students' access to technology, which is not practical outside the classroom. Thus, the strategy of classroom-based semi-structured writing assessed by FLO3 allows a more accurate and reliable assessment of the quality of a student's writing, compared to the strategy of homework-based semi-structured writing.

To sum up, the strategy of classroom-based semi-structured writing assessed by FLO3 benefits the students directly by providing them with ideal conditions for writing productively in a second language. It also benefits the students indirectly by giving the instructor a metric of writing fluency that is ideally suited to measuring student performance in semi-structured writing and provides a deeper understanding of the factors that limit an individual student's writing. It is therefore recommended that this strategy be implemented as soon as possible, in a way that complements existing FLO1-based writing assessment strategies, such as those discussed above.

References

"Interpreting Regression Output" (2007) **DSS - Princeton University Library.** Princeton University Library.

"Linear Regression" **RapidMiner Documentation.** RapidMiner. <http://docs.rapidminer.com/studio/operators/modeling/classification_and_regression/function_fitting/linear_regression.html>.

Soper, Dr. Daniel. (2008) "The Free Statistics Calculators Website - Home" **The Free Statistics Calculators Website - Home.**<http://www.danielsoper.com/statcalc3/>.

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.

Fellner, Terry and Matthew Apple. (2006) "Developing Writing Fluency and Lexical Complexity with Blogs" **The JALT CALL Journal.** 2(1) pages 15-26.

Latif, Muhammad M. Abdel. (2009) "Toward a New Process-Based Indicator for Measuring Writing Fluency: Evidence from L2 Writers' Think-Aloud Protocols" **Canadian Modern Language Review/ La Revue Canadienne Des Langues Vivantes**. 65(4) pages 531-58.

Laufer, B. (1995) "Vocabulary Size and Use: Lexical Richness in L2 Written Production" **Applied Linguistics.** 16(3) pages 307-22.

Polio, Charlene G. (1997) "Measures of Linguistic Accuracy in Second Language Writing Research" **Language Learning**. 47(1) pages 101-43.

Sauro, Jeff. (2013) "Best Practices For Using Statistics On Small Sample Sizes" **MeasuringU: Usability, Customer Experience & Statistics.**<http://www.measuringu.com/blog/small-n.php>.

Schlotzhauer, Sandra D. (2007) **Elementary Statistics Using JMP.** Cary, NC: SAS Institute.

Sellke, Thomas, M. J. Bayarri and James O. Berger. (2001) “Calibration of p Values for Testing Precise Null Hypotheses” **The American Statistician.** 55(1).

Trubek, Anne. (2015) "We Are All Writers Now" **More Intelligent Life.** The Economist Newspaper Limited.

Sugita, Yoshihito. (2012) "Enhancing Students’ Fluency in Writing: Learning to Use Transition Words" **Open Journal of Modern Linguistics** 02.01: pages 18-25.

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] 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 frequently advanced by commentators on the web. Anne Trubek's article "We are all writers now" is an eloquent example. According to Trubek, 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 - no judgment is being made regarding any other aspect of that writing.

[2] Again, the researcher is referring to general, subjective writing about everyday topics, where there is no "right answer" to be given to a question. Therefore, the L1 writer is concerned with the quality of the writing itself. Other attributes, such as the range of vocabulary and logical coherence may also be influential, but this study asserts that the most influential one is accuracy.

[3] In this study, the term "model" refers to the linear regression model. The linear regression model fits a linear equation to a set of data, in order to show the relationship between a scalar variable and one or more explanatory variables. In this study, the linear regression model is calculated by the RapidMiner linear regression operator, which uses the Akaike criterion for model selection. This criterion is explained in the RapidMiner documentation.

[4] 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.

[5] 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. However, the Researcher considers it unlikely that this attribute would be as influential as accuracy, for this type of writing.

[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] An "L1 writer" in this study means anyone who writes exactly like an L1 writer, regardless of their birthplace.

[8] Here is the reply I received from David A., Global Moderator of the RapidMiner forum:
It can be both.

For very small numbers RapidMiner just shows a 0 in the result view, but the actual value is used for further calculations (for example you can sort according to the p-values).
But for very small values it can happen that the p-value becomes a genuine 0, this depends on the underlying distribution functions and their parameters, so a fixed threshold cannot be given.
For all practical concerns the differences between a genuine 0 and a very small, non-zero, p-value should not matter.

[9] What follows is purely speculative, so it is included as an endnote. Here follows an example of "apparent divergence." Error type C (adjective/participle errors such as "I am boring with this movie"), is very divergent, since it is a type of error that is rarely if ever made in writing by a native speaker, and also very noticeable. Consequently, type C errors are given greater weight by the rater, and therefore have more effect on the rating (in a negative direction). By contrast, error type A (verb errors such as "she play the guitar very well") may be very divergent, but not in every case. An error such as "she play the guitar very well" is very divergent, since it is the kind of error that would never be made in writing by a native speaker (except by carelessness). However, there are some rarely used verb tenses or constructions in English, where an L2 writer may make the same kind of errors that are occasionally made by L1 writers, in which case these errors cannot be regarded as divergent. So, it is arguable that in general C type errors are more divergent then type A errors. They also differ in their noticeability. In the example of "she play the guitar very well," the error consists in the omission of just one letter. It is the kind of error that might be overlooked by a rater who is reading quickly. Admittedly, some cases of error type A are more noticeable, especially those requiring "to be." For example, the error in "she playing the guitar" is harder to miss, because it involves the omission of an entire word, "is", compared to the omission of a mere "s." However, the Report-writer would suggest that C type errors are generally more noticeable than A type errors. Therefore, the greater overall divergence and noticeability of the C type error would account for its greater effect (-7.14) in this study, compared to that of the A type error (-1.773).

[10] "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.

APPENDIX

Excel and RapidMiner terms used in this report

(RM) refers to a RapidMiner term

Data table

The data when it is stored in an Excel file, either before import into RapidMiner or after export from RapidMiner.

Data set (RM)

The data when it is stored in RapidMiner and imported into a RapidMiner process.

Model (RM)

In this study, the term “model” refers to the linear regression model. The linear regression model fits a linear equation to a set of data, in order to show the relationship between a scalar variable and one or more explanatory variables. In this study, the scalar variable is a rating given to a sample of writing, and the explanatory variables are attributes of the writing such as total length, etc.

Attribute (RM)

An attribute is a characteristic of a person or object which we are interested in. For example, a student may have the following attributes: Name, student code, subject major code, etc. In this study, attributes relate to characteristics of a sample of writing, such as total length, etc.

Target attribute (or label) *(RM)*

A target attribute is a characteristic whose value we are trying to find, but is currently unknown.

Process (RM)

A RapidMiner process is a “workflow” that consists of a sequence of operators.

Operator (RM)

Each operator performs one task within the process. The output of one operator forms the input of the next one.

Using the Correlation Matrix to Reveal Correlations

Overview

The Researcher assumed that there are certain correlations between the ability of the writer and the frequency distribution of types of errors made (when writing about general, everyday topics) and that these correlations vary only slightly from year to year for Thai students studying English-Chinese major. To support this assumption, the Researcher has statistical evidence, available on request.Following this assumption, these correlations should be useful for developing a model of writing fluency that will be applicable to future classes of similar students.

The objective was to find significant correlations between different types of errors. Significant correlations could be either positive (if a greater quantity of errors of type x was associated with a greater quantity of error of type y) or negative (if a greater quantity of errors of type x was correlated with a smaller quantity of error of type y) correlations. However, correlations where one or both of the error types had a very low frequency were ignored. Therefore, correlations relating to either type J or type K were ignored as all of the students had an error rate of less than 1 per 100 words for both these types.

Using the Correlation Matrix to Reveal Correlations

First, the pre and post-test data was *combined* into a single data table of 22 rows. For example, the total word counts in the pre and post-test were added together, and likewise the number of errors for each type in the pre and post-test were added. Then the combined error totals were divided by the combined word counts and multiplied by 100 to get new error rates for each type.

This combined data table was named "28 Oct 2013 AND 6 Jan 2014 EG3173 CLEAN BY RATE" in Excel. Then a data set was created by importing this file into RapidMiner. The following columns were selected as (numeric data) attributes:

No. of A type errors / no. of words * 100 (= no. of errors per 100 words)

No. of B type errors / no. of words * 100 (= no. of errors per 100 words)

No. of C type errors / no. of words * 100 (= no. of errors per 100 words)

No. of D type errors / no. of words * 100 (= no. of errors per 100 words)

No. of E type errors / no. of words * 100 (= no. of errors per 100 words)

No. of F type errors / no. of words * 100 (= no. of errors per 100 words)

No. of G type errors / no. of words * 100 (= no. of errors per 100 words)

No. of H type errors / no. of words * 100 (= no. of errors per 100 words)
No. of I type errors / no. of words * 100 (= no. of errors per
100 words)
No. of J type errors / no. of words * 100 (= no. of errors per
100 words)
No. of K type errors / no. of words * 100 (= no. of errors per 100 words)
No. of Z type errors / no. of words * 100 (= no. of errors per 100 words)

Then the data set named "correlation matrix data for error types Oct 13 AND Jan 14" was saved in RapidMiner's local repository.

Then RapidMiner was used to create a new process ("0030 Correlation matrix Oct 13 AND Jan 14 for error types A to Z"), which contained a Retrieve operator and a Correlation Matrix operator. The Retrieve operator retrieved the "correlation matrix data for error types Oct 13 AND Jan 14" data set from the repository, then output the data set to the Correlation Matrix operator. This operator then calculated the correlations of all attributes in the Example Set and output a correlation matrix that is shown in Table 18. The workflow of the process "0030 Correlation matrix Oct 13 AND Jan 14 for error types A to Z" is shown in Fig. 7.

Review of the Correlations

First, the greatest positive correlations were identified. Note that the correlation of 0.655 between errors of I type (article errors) and K type (adverb errors) was excluded due to the low frequency of K type errors for all students in these data tables.

1. The greatest positive correlation is between H type errors (missing or incorrect conjunction used) and B type errors (pronoun-related errors), namely 0.700. Note that a correlation coefficient of 1 indicates a perfect positive correlation, whereas a correlation coefficient of 0 indicates a complete absence of correlation between the two variables. So 0.7 indicates a high degree of correlation.
2. The next greatest positive correlation is between errors of Z type (sentence level errors) and E type (inappropriate word for context), namely 0.595.
3. Another large positive correlation is between errors of A type (verb errors) and F type (singular/plural errors), namely 0.528.
4. Another large positive correlation is between errors of B type (pronoun errors) and D type (preposition errors), namely 0.49.

5. Another large positive correlation is between errors of A type (verb errors) and B type (pronoun errors), namely 0.487.

6. Another large positive correlation is between D type errors (preposition errors) and G type errors (noun errors), of 0.432.

So far then, six potentially useful correlations have been identified: one between B and H type errors, another between Z and E type errors, another between A and F type errors, another between B and D type errors, another between A and B type errors, and a sixth between D and G type errors.

Regarding the H and B types correlation, a possible explanation could be that both pronoun errors and conjunction errors are quite basic errors, and therefore writers who make H errors are also likely to make B errors. The same explanation could account for the A and F types correlation, the B and D types correlation, the A and B types correlation and the D and G types correlation. Regarding the Z and E types correlation, a possible explanation could be that the more capable writers tend to experiment more with new words and constructions; also, they attempt more complicated sentences and thus make more sentence level errors.

To confirm these possible explanations or find alternative explanations for these correlations, more data needs to be collected and analyzed. Unfortunately, this was not feasible during the present study.

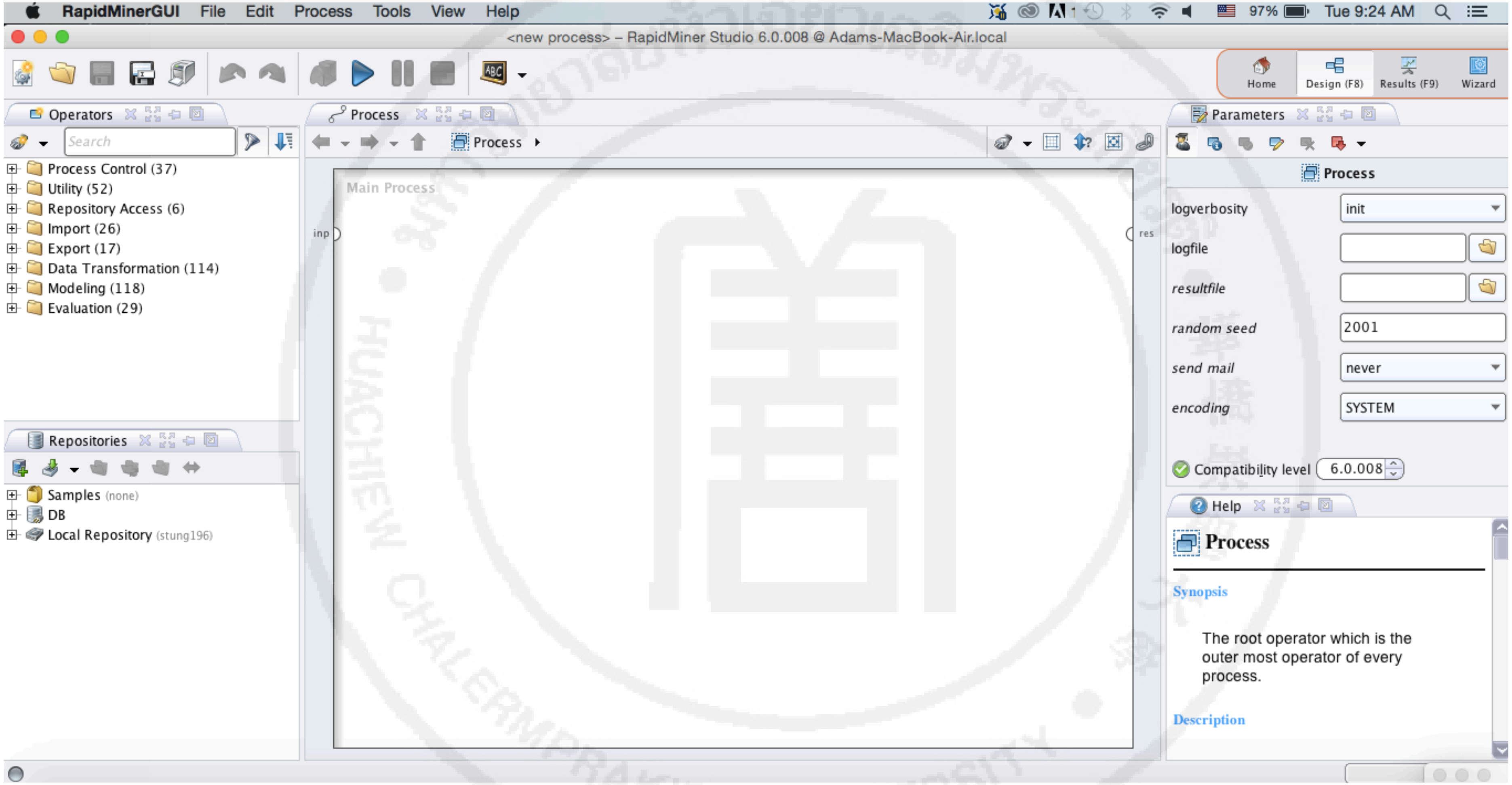


Fig 1: The main screen of RapidMiner.

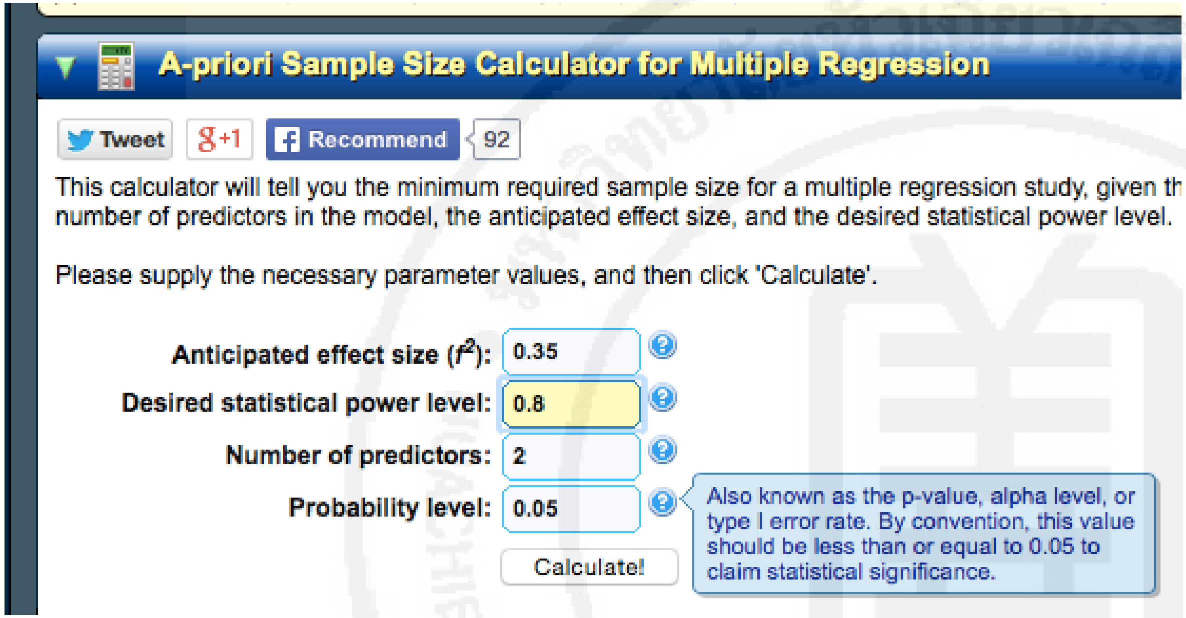


Fig. 2. A-priori Sample Size Calculator for Multiple Regression (Statistics Calculators, danielsoper.com)

Topic 2: ARTS, FASHION AND ENTERTAINMENT

1. Can you paint or draw?
2. What is your favorite painting? Why do you like it?
3. Have you ever been to a live music concert?
4, What is your favorite music band?
5. What is your favorite clothing store or brand?
6. Do you have a favorite TV program?
7. If you dyed your hair, what color would it be?
8. What is the coolest website you have ever seen?
9. Would you like to be a fashion model?
10. Do you know anyone who has had plastic surgery?
11. Do you have a friend who spends too much money on beauty and cosmetics?
12. What is good or bad about being a beautiful girl?
13. If HCU could change the student uniform, what should it look like?
14. Supposing you had a time machine, which time would you like to visit?
15. Which insect do you think is beautiful?
16. Which building in Bangkok do you think is the ugliest?
17. If you could design a new campus, what would it look like?
18. What do you think of young people wearing a lot of colors?
19. Is there another country's fashion that you really like?
20. If you could play any musical instrument, what would it be?
21. If you could write a book, what would it be about?
22. If you wanted to write a novel, where would you want to live?
23. Do you think that schools should teach more arts subjects or less?
24. Did you ever write something in code? Why?
25. If you had the chance to act in a new movie, what kind of movie would it be, and what role would you play?
26. Supposing you had magical powers, what would you do?
27. How would your life be different if you couldn't listen to music?

Topic 3: SPORTS, FOOD AND HEALTH

1. What sports do you like?
2. What food could you never give up in your life?
3. When was the first time you went swimming?
4, What was the longest time you were ever sick?
5. Do you prefer sweet or bitter foods?
6. What food do you dislike most?
7. Have you ever been on a diet?
8. If you could be in an Olympic event, what would it be?
9. What is the longest you have ever gone without sleep?
10. Have you ever done a bungee jump? Describe what it was like.
11. When did you feel most tired?
12. What is the furthest you have ever run?
13. Do you know someone who never exercises?
14. What was the most unusual food you ever tasted?
15. Have you ever followed a vegetarian diet?
16. Do you want to improve your hearing, sight, sense of smell, taste, touch?
17. Are you shy at parties?
18. How do you cope with stress during exam week?
19. Have you ever gone to watch a live sports event?
20. Is it healthier to live in the countryside or the city?
21. Do you know someone who has tried traditional Chinese medicine?
22. If you could choose, how long would you want to live?
23. Do you know someone who engages in an extreme sport?
24. Do you have to commute a long distance to HCU? Does it affect your health?
25. Invent a new sport.
26. How much can you know about someone from their blood type?
27. Where is the healthiest place to live in Samut Prakarn?
28. Should sports be compulsory at high school?

Fig. 3: Journal topics lists 2 &3

Topic 4: EDUCATION, WORK AND FAMILY

1. What do you like about HCU?
2. What did you like about Primary School?
3. What did you like about High School?
4. How is studying at university different from studying at High School?
5. Would you prefer to work for yourself or for an employer?
6. Would you prefer to work in an office or outside?
7. What advice would you give a freshie English-Chinese major student?
8. Have you ever done a part-time job?
9. What does the library at HCU need?
10. When your parents were the age you are now, how was their life different from yours?
11. Describe your life when you are 30 years old.
12. Describe your ideal job.
13. Imagine you are a kindergarten teacher. Describe a day of class.
14. Describe your ideal boyfriend or girlfriend.
15. Do boys and girls study English differently?
16. What would your life be like if you spoke fluent English and Chinese now?
17. Do you know someone who does an unusual job?
18. Do you think that new technology will make work easier or harder?
19. Did you ever have a relative who was very strict?
20. What are the advantages of being an eldest daughter or son?
21. What are the advantages of being a youngest daughter or son?
22. Did you ever argue with your sister or brother?
23. If you had an identical twin sister or brother, how would your life be different?

Topic 5: FRIENDS

1. Describe your ideal friend.
2. Do you have a best friend?
3. What do you like to do best with friends?
4. What was the best vacation trip you ever had with a friend?
5. Do you play any sports with your friends?
6. Do you have a rich friend?
7. Do you have an artistic friend?
8. Do you have a friend with a car?
9. Do you and your friend like the same music?
10. Do you play any sports with your friends?
11. Does your friend shop differently from you?
12. Do you have a friend who is very different from you?
13. How are your friends in college different from the friends you had at high school?
14. What are the differences between having male friends and having female friends?
15. Are your friends from Bangkok different from those from the countryside?
16. Do you have a friend from high school who didn't go to college?
17. When were you angriest with a friend?
18. When were you the most happy with a friend?
19. When were you the most disappointed with a friend?
20. Do you have friends who are friends of your sister or brother?
21. When did you meet your first friend?
22. Do you have a friend who you didn't like when you first met her/him?
23. Have you ever lied to a friend, or has a friend ever lied to you?
24. Do you prefer to live with your friends or to live alone?
25. What kinds of friends will you have when you are 30 years old?
26. Do you ever wish you had more friends?
27. Do you have any friends who you have never met in real life?
28. Do you think it is easier to make friends now, than twenty years ago?

Fig.4: Journal topics lists 4 & 5

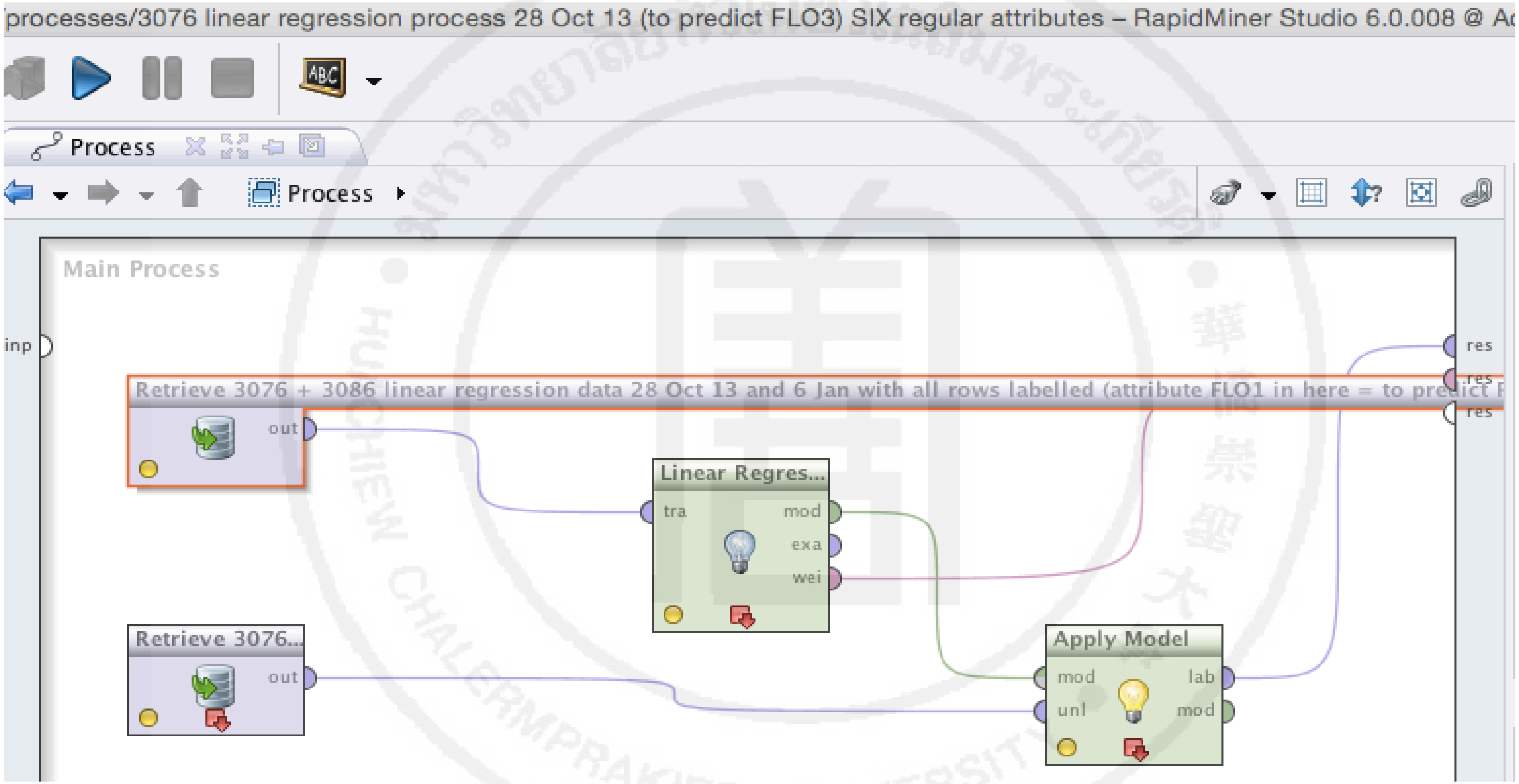


Fig. 5. Workflow of "3076 linear regression process 28 Oct 13 (to predict FLO3) SIX regular attributes

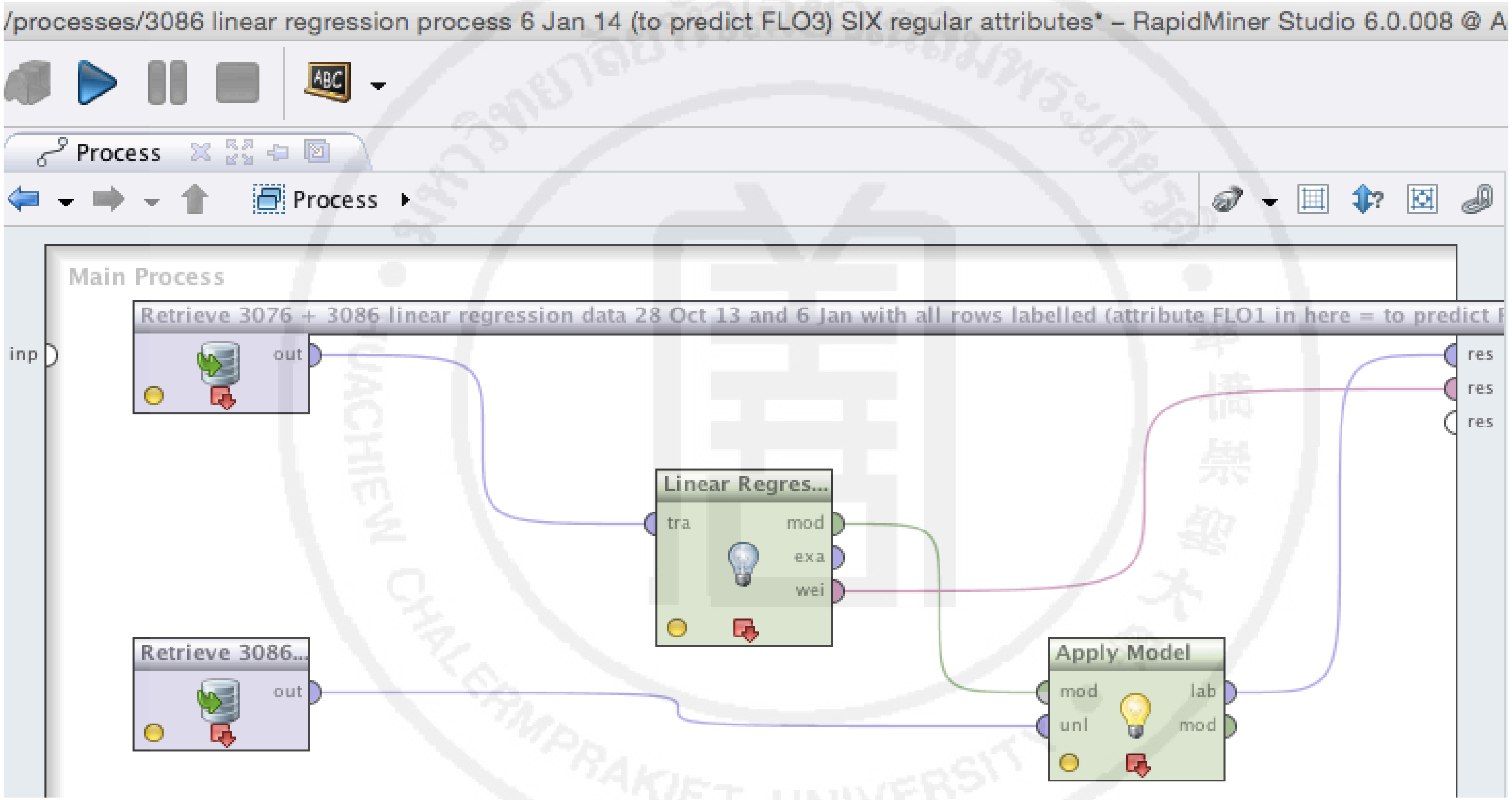


Fig. 6. Workflow of "3086 linear regression process 6 Jan 14 (to predict FLO3) SIX regular attributes"

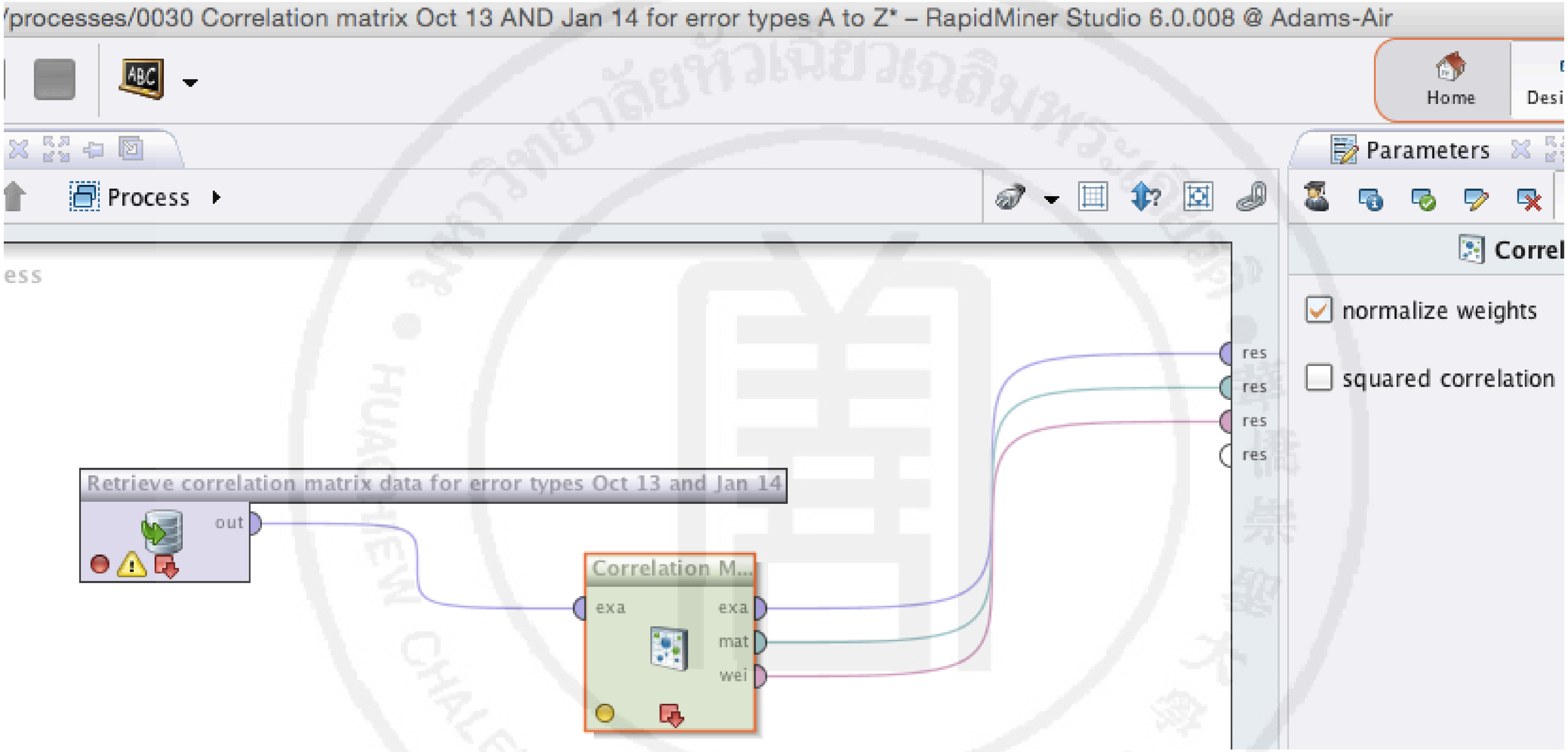


Fig. 7. Workflow of "0030 Correlation matrix Oct 13 AND Jan 14 for error types A to Z"

	Date of Pre-test	Date of post-test	Subject code	No. of students who took the pre-test & post-test	Total number of words written	Total time taken (seconds)
Data collection	28 Oct, 2013	6 Jan, 2014	EG 3173	22	8544	68353

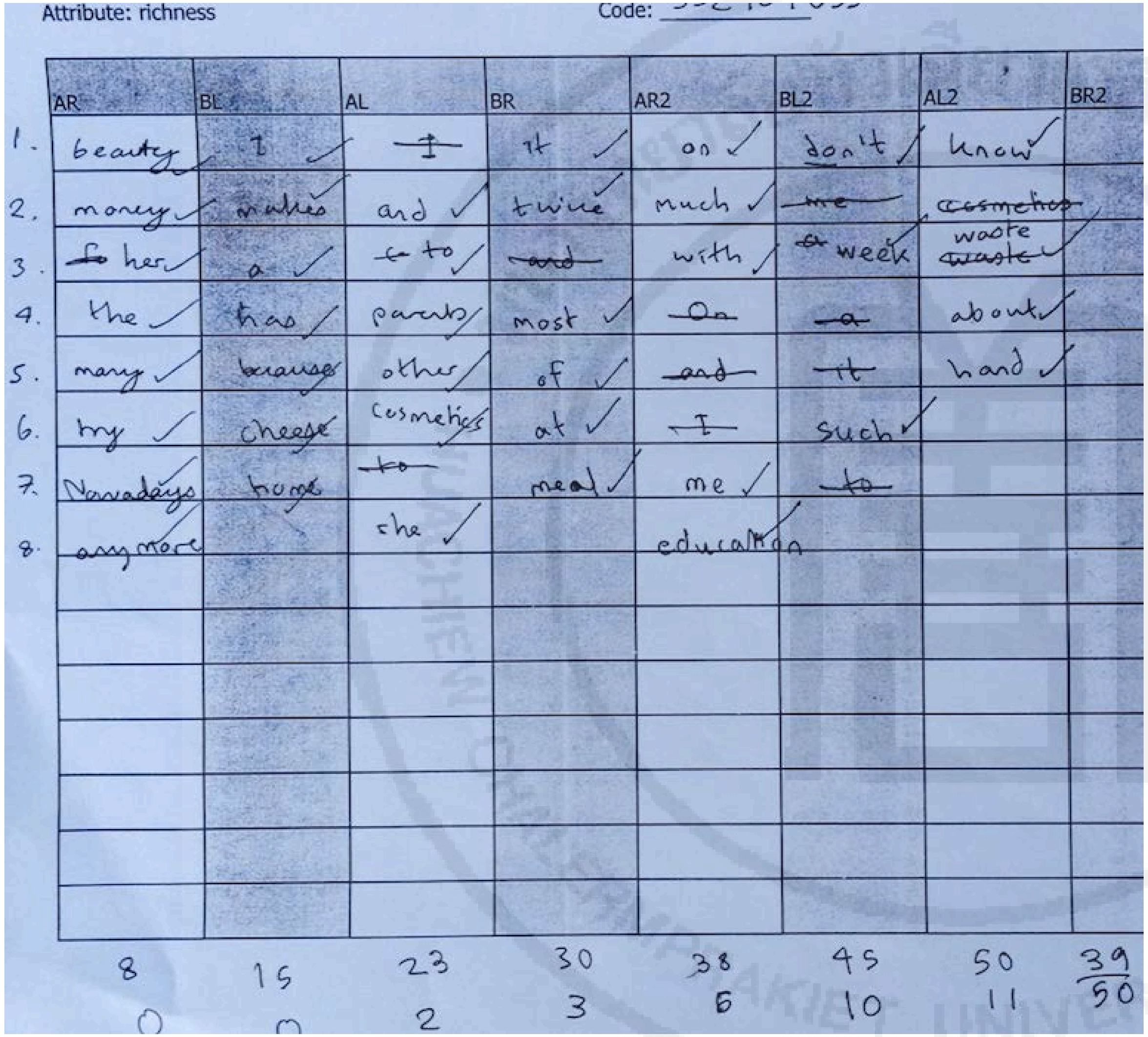
Table 1: Details of the data collection

Letter code	DESCRIPTION OF ERROR	EXAMPLE
A	Verb error (doesn't agree with the subject or the tense is incorrect) OR missing verb	She play the guitar very well
B	Pronoun error OR missing pronoun	I bought she a pair of shoes.
C	Adjective error (including particple error) OR missing adjective (the sentence requires	EX: I am boring with this subject
D	Preposition error OR missing preposition	Almost of tourists love this island OR I listen Korean pop music every day
E	Inappropriate word for context	I migrate to work by train
F	Singuar / plural error	There are a thousand factory in Samutprakarn
G	*Where a noun should have been used, but another part of speech was used*	"Funny" instead of "fun"
H	Conjunction error OR missing conjunction	I don't like sphagetti but pizza
I	Article error - incorrect article used or missing	I bought colorful sweater in Asiatique
J	Countable / uncountable error	There are too much car on this highway.
K	Adverb-related errors	The turtle moved slow across the beach.
Z	Errors of sentence format	Words in wrong order / run-on sentence / sentence fragment

Table 2: Types of errors with descriptions and examples

AR	BL	AL	BR	AR2	BL2	AL2	BR2	Totals

Table 3: Table for calculating the richness attribute



Attribute: richness Code:

	AR	BL	AL	BR	AR2	BL2	AL2	BR2
1.	beauty	I	~~I~~	it	on	don't	know	
2.	money	makes	and	twice	much	~~me~~	~~cosmetics~~	
3.	~~to~~ her	a	~~a~~ to	~~and~~	with	~~a~~ week	waste ~~waste~~	
4.	the	has	parents	most	~~On~~	~~a~~	about	
5.	many	because	other	of	~~and~~	~~it~~	hand	
6.	my	cheese	cosmetics	at	~~I~~	such		
7.	Nowadays	home	~~to~~	meal	me	~~to~~		
8.	anymore		she		education			
	8	15	23	30	38	45	50	39/50
	0	0	2	3	6	10	11	

Table 4: Example of table after calculating the richness attribute

Reference code	FLO1 = x/200	TOTAL LENGTH (word count)	ADJUSTED WORD COUNT (FROM SPEED)	SPEED (words per minute)	Vocab x/50	No. of A type errors / no. of words * 100 (= no. of errors	No. of B type errors / no. of words * 100 (= no. of errors	No. of C type errors / no. of words * 100 (= no. of errors	No. of D type errors / no. of words * 100 (= no. of errors	No. of E type errors / no. of words * 100 (= no. of errors	No. of F type errors / no. of words * 100 (= no. of errors	No. of G type errors / no. of words * 100 (= no. of errors	No. of H type errors / no. of words * 100 (= no. of errors	No. of I type errors / no. of words * 100 (= no. of errors	No. of J type errors / no. of words * 100 (= no. of errors	No. of K type errors / no. of words * 100 (= no. of errors	No. of Z type errors / no. of words * 100 (= no. of errors	No. of errors/ no of words * 100 (Bigger is less accurat
B1	150	311	62	6	36	1.93	0.00	0.32	0.96	4.50	0.64	0.64	0.64	0.32	0.00	0.00	1.29	11.3
B2	150	237	132	13	35	2.53	0.00	0.00	0.00	3.38	0.84	0.42	0.00	1.69	0.00	0.00	0.42	9.3
B3	100	109	57	6	31	1.83	0.00	0.92	2.75	3.67	0.92	1.83	0.00	5.50	0.00	0.92	2.75	21.1
B4	122	215	72	7	29	8.84	2.79	0.00	1.40	2.79	0.00	0.00	2.79	0.47	0.00	0.00	2.79	21.9
B5	160	323	140	14	44	2.17	0.62	0.62	0.00	0.93	2.17	0.00	0.00	1.86	0.00	0.00	0.62	9.0
B6	134	285	77	8	42	6.32	3.16	0.35	2.11	3.86	3.86	0.70	1.40	2.11	0.00	0.00	2.81	26.7
B7	154	227	63	6	38	1.76	0.44	0.00	0.44	0.00	0.00	0.00	0.88	0.88	0.00	0.00	0.00	4.4
B8	136	190	63	6	31	2.11	2.11	0.00	1.05	0.53	0.53	0.00	0.00	2.11	0.00	0.53	0.53	9.5
B9	120	131	81	8	39	3.82	2.29	0.00	2.29	1.53	2.29	0.76	0.76	0.76	0.00	0.76	0.00	15.3
B10	160	315	90	9	37	2.54	1.90	0.00	0.32	1.90	0.63	0.63	1.59	2.22	0.00	0.63	0.63	13.0
B11	156	232	122	12	41	1.29	0.00	0.00	0.43	0.43	0.43	0.00	0.00	1.29	0.00	0.00	0.86	4.7
B12	134	179	60	6	36	1.68	2.23	0.56	0.56	3.35	1.12	0.00	0.00	1.68	0.56	0.00	1.12	12.8
B13	106	145	58	6	36	6.21	0.69	0.00	0.00	3.45	2.76	0.69	0.69	1.38	0.69	0.00	1.38	17.9
B14	110	185	62	6	41	1.62	1.62	0.00	2.16	1.08	1.08	0.54	1.08	2.70	0.00	0.00	1.08	13.0
B15	125	211	106	11	38	3.32	0.47	0.95	0.47	0.95	0.47	0.47	0.00	0.00	0.00	0.00	0.00	7.1
B16	144	201	56	6	41	5.97	0.00	1.00	0.00	1.99	1.99	0.00	0.00	0.00	0.00	0.00	1.49	12.4
B17	148	197	56	6	39	0.51	1.02	0.51	0.51	0.51	1.52	0.00	0.51	0.51	0.00	0.00	0.51	6.1
B18	144	214	71	7	40	7.01	1.40	0.00	1.87	1.87	2.34	0.47	0.47	0.47	0.00	0.00	1.87	17.8
B19	130	165	40	4	39	0.00	0.61	0.00	0.00	3.03	0.61	0.00	0.00	0.00	0.00	0.61	1.21	6.1
B20	106	126	58	6	37	5.56	2.38	0.79	0.79	3.17	3.17	0.00	0.79	3.97	0.00	0.79	0.00	21.4
B21	84	118	57	6	40	6.78	2.54	2.54	0.00	0.00	2.54	0.00	0.00	0.85	0.00	0.85	2.54	18.6
B22	121	176	98	10	42	8.52	2.27	0.00	1.70	1.70	1.14	0.00	0.00	3.98	0.00	1.70	0.57	21.6

Table 5: Contents of "28 Oct 2013 EG 3173 CLEAN 2 COMPLETE BY RATE ALL ROWS LABELLED FLO3"

	Reference code	FLO1 = x/200	TOTAL LENGTH (word count)	ADJUSTED WORD COUNT (FROM SPEED)	SPEED (words per minute)	Vocab x/50	No. of A type errors / no. of words * 100 (= no. of errors	No. of B type errors / no. of words * 100 (= no. of errors	No. of C type errors / no. of words * 100 (= no. of errors	No. of D type errors / no. of words * 100 (= no. of errors	No. of E type errors / no. of words * 100 (= no. of errors	No. of F type errors / no. of words * 100 (= no. of errors	No. of G type errors / no. of words * 100 (= no. of errors	No. of H type errors / no. of words * 100 (= no. of errors	No. of I type errors / no. of words * 100 (= no. of errors	No. of J type errors / no. of words * 100 (= no. of errors	No. of K type errors / no. of words * 100 (= no. of errors	No. of Z type errors / no. of words * 100 (= no. of errors	No. of errors/ no of words * 100 (Bigger is less accurat
1	B1	118	125	125	12	39	1.60	0.00	0.00	0.80	4.00	0.80	0.00	0.80	0.00	0.80	0.00	1.60	10.4
2	B2	149	195	98	10	31	3.59	0.00	0.00	2.05	0.51	0.00	0.00	0.00	1.54	0.00	0.00	0.00	7.7
3	B3	74	97	65	6	36	5.15	0.00	0.00	0.00	4.12	1.03	4.12	0.00	2.06	0.00	0.00	0.00	16.5
4	B4	130	172	86	9	37	5.23	2.91	0.58	1.16	1.74	2.91	1.74	0.58	0.00	0.00	0.00	2.91	19.8
5	B5	144	155	107	11	42	2.58	0.00	0.00	0.65	0.65	0.65	0.00	0.00	0.65	0.00	0.00	0.00	5.2
6	B6	110	142	95	9	39	5.63	2.11	0.00	1.41	2.82	4.23	2.11	0.00	0.70	0.00	0.00	2.82	21.8
7	B7	147	258	40	4	39	0.78	0.00	0.00	0.00	1.55	0.39	0.00	0.39	0.78	0.00	0.00	1.16	5
8	B8	140	232	89	9	39	3.02	2.59	0.43	1.29	0.00	1.29	0.43	1.29	2.16	0.00	0.43	0.43	13.4
9	B9	116	168	105	11	32	5.95	1.19	0.60	1.79	3.57	0.60	0.60	0.60	0.60	0.00	0.00	1.79	17.3
10	B10	134	163	102	10	41	1.84	1.23	0.61	0.00	1.23	0.00	1.23	0.61	0.61	0.00	0.00	1.84	9.2
11	B11	133	210	90	9	33	3.81	0.95	0.00	0.00	0.48	0.95	0.00	0.00	0.95	0.00	0.00	0.48	7.6
12	B12	152	255	80	8	41	3.53	0.39	0.00	0.00	1.18	0.78	0.00	0.00	0.78	0.00	0.00	0.78	7.5
13	B13	100	134	96	10	37	7.46	2.24	0.00	0.75	5.22	0.75	0.00	0.00	2.24	0.00	0.00	2.99	21.6
14	B14	113	124	51	5	38	2.42	0.81	0.00	0.81	0.81	0.00	0.00	0.00	2.42	0.00	0.00	2.42	9.7
15	B15	110	131	57	6	34	9.92	0.00	0.00	0.00	0.00	0.76	0.76	0.76	0.00	0.76	0.00	0.00	13
16	B16	140	255	91	9	44	3.92	1.57	0.78	0.78	0.78	1.18	0.39	0.39	0.78	0.00	0.00	0.39	11
17	B17	153	274	78	8	41	1.82	0.36	0.36	0.36	1.46	1.09	0.36	0.00	1.46	0.00	0.00	1.09	8.4
18	B18	140	233	86	9	36	3.86	0.43	0.00	0.43	4.72	0.86	0.00	0.86	0.43	0.00	0.00	1.72	13.3
19	B19	140	221	49	5	37	2.26	0.45	0.00	0.00	0.45	2.26	0.00	0.00	1.36	0.00	0.00	0.00	6.8
20	B20	110	141	76	8	41	2.84	1.42	0.00	2.13	0.71	1.42	0.00	1.42	2.13	0.00	0.00	0.00	12.1
21	B21	132	233	118	12	40	6.44	0.43	0.00	0.86	2.58	1.72	0.00	0.43	3.43	0.00	0.86	0.43	17.2
22	B22	104	134	117	12	38	2.99	0.00	0.75	0.75	2.99	2.99	0.75	0.00	2.24	0.00	0.00	0.00	13.4

Table 6: Contents of "6 Jan 2014 EG 3173 CLEAN 2 COMPLETE BY RATE ALL ROWS LABELLED FLO3"

P value	Probability of incorrectly rejecting a true null hypothesis
0.05	At least 23% (and typically close to 50%)
0.01	At least 7% (and typically close to 15%)

Table 7. Source: Thomas SELLKE, M. J. BAYARRI, and James O. BERGER, "Calibration of p Values for Testing Precise Null Hypotheses"

//Local Repository/processes/3075 linear regression process 28 Oct 13 (to predict FLO3) 13 regular attributes

ExampleSet (//Local Repository/data/3075 [...] e FLO1 in here = to predict FLO3) 13 p

ult Overview | AttributeWeights (Linear Regression)

Attribute	Coefficient	Std. Error	Std. Coeffi...	Tolerance	t-Stat	p-Value	Code
(Intercept)	109.224	8.388	?	?	13.021	0	****
TOTAL LEN(	0.252	0.023	0.714	0.811	10.717	0	****
No. of A typ	-1.643	0.556	-0.191	0.897	-2.953	0.007	***
No. of I type	-3.762	1.297	-0.221	0.967	-2.900	0.007	***
No. of C typ	-8.192	2.864	-0.191	0.982	-2.861	0.008	***
No. of G typ	-4.643	1.728	-0.174	0.868	-2.687	0.013	**
No. of E typ	-2.107	0.999	-0.151	0.946	-2.109	0.050	**
No. of H typ	-4.472	2.563	-0.129	0.991	-1.745	0.110	
No. of D typ	2.136	1.895	0.081	0.925	1.127	0.380	
No. of K typ	4.631	4.167	0.085	0.919	1.111	0.392	
Vocab x/50	-0.253	0.234	-0.043	0.932	-1.080	0.417	
No. of B typ	-1.045	1.758	-0.051	0.960	-0.594	0.562	
No. of Z typ	0.643	1.609	0.031	0.925	0.400	0.696	
No. of F typ	0.500	1.401	0.026	0.912	0.357	0.727	

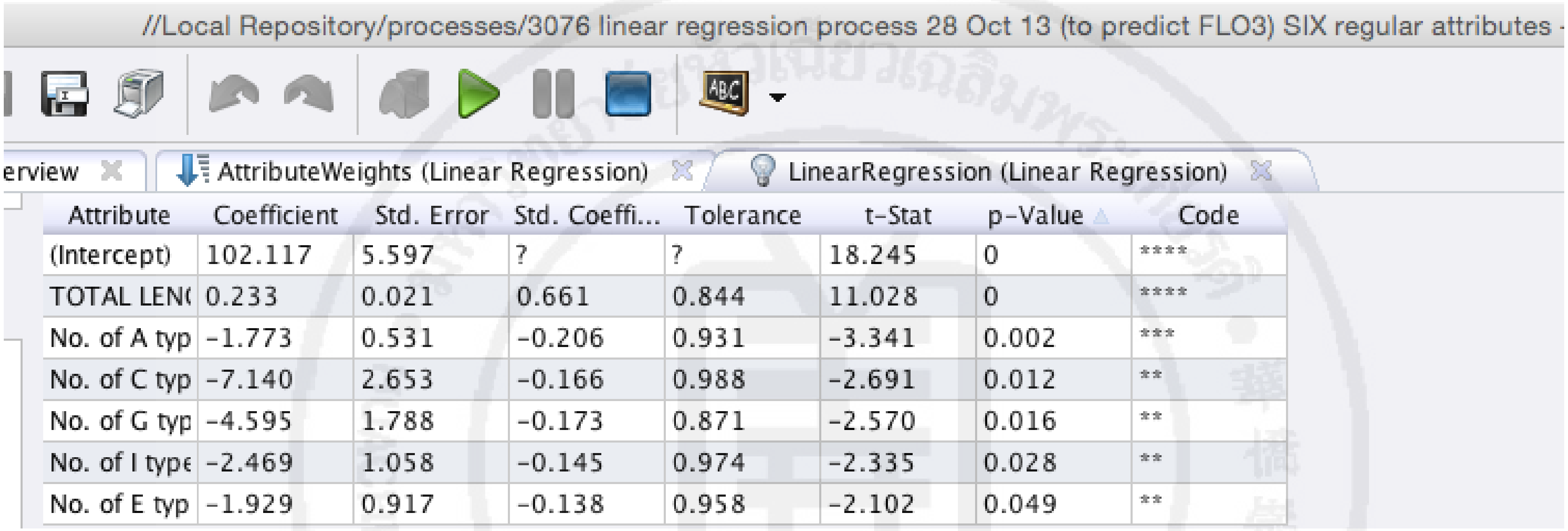
Table 8. Iteration 1: Output from "3075 linear regression process 28 Oct 13 (to predict FLO3) 13 regular attributes" (ordered by p-Value). NOTE: this process was used during the development stage and modified before the final model.

//Local Repository/processes/3074 linear regression process 28 Oct 13 (to predict FLO3) THREE regular attributes*

view | AttributeWeights (Linear Regression) | LinearRegression (Linear Regression)

Attribute	Coefficient	Std. Error	Std. Coeffi...	Tolerance	t-Stat	p-Value	Code
TOTAL LEN(	0.254	0.025	0.721	0.895	9.975	0	****
Vocab x/50	-0.040	0.255	-0.007	0.941	-0.157	0.877	
No. of error	-1.291	0.244	-0.362	0.915	-5.289	0.000	****
(Intercept)	98.055	8.241	?	?	11.898	0	****

Table 9. Iteration 2: Output from "3074 linear regression process 28 Oct 13 (to predict FLO3) THREE regular attributes" (ordered by p-Value). NOTE: this process was used during the development stage and modified before the final model.



//Local Repository/processes/3076 linear regression process 28 Oct 13 (to predict FLO3) SIX regular attributes -

erview | AttributeWeights (Linear Regression) | LinearRegression (Linear Regression)

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 type	-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 10: Output from "3076 linear regression process 28 Oct 13 (to predict FLO3) SIX regular attributes" (ordered by p-Value). NOTE: this process is running the final model.

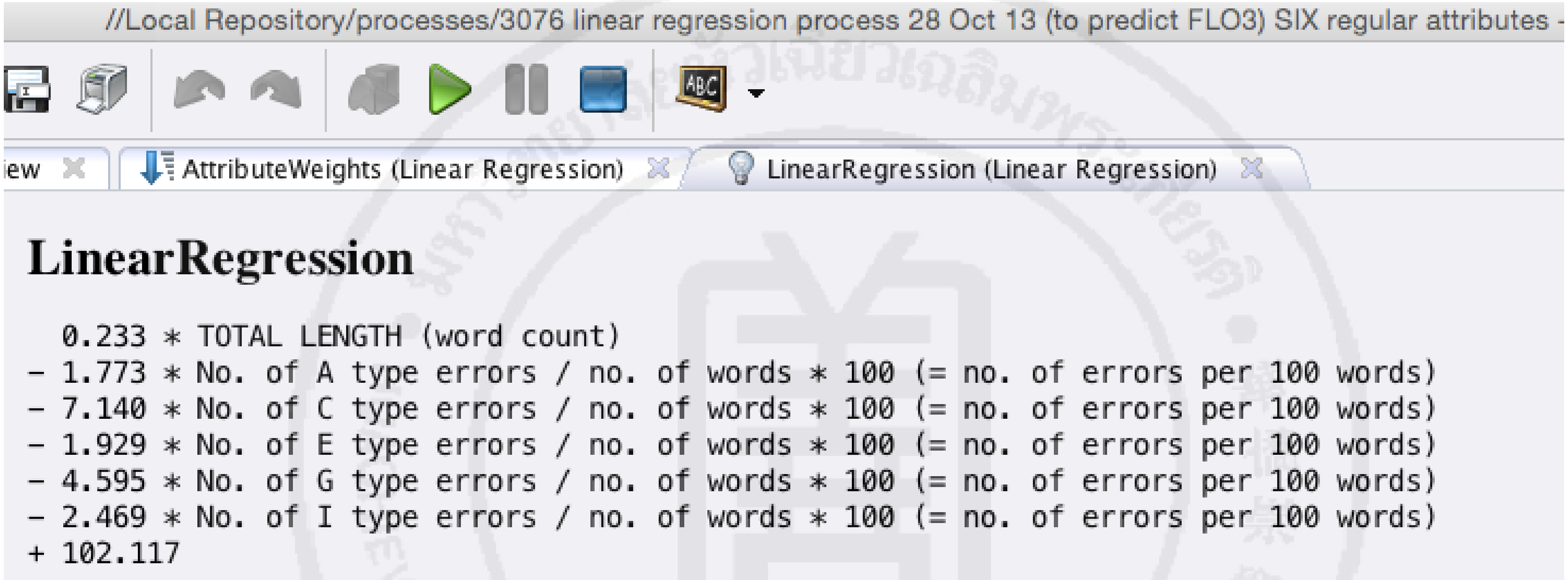


Table 11: Descriptive output of "3076 linear regression process 28 Oct 13 (to predict FLO3) SIX regular attributes"

//Local Repository/processes/3076 linear regression process 28 Oct 13 (to predict FLO3) SIX regular attributes

iew ExampleSet (Retrieve 3076 linear regress [...] FLO1 in here = to predict FLO3) SIX pvs)

ExampleSet (22 examples, 2 special attributes, 6 regular attributes)

Ro...	V...	prediction(...	TOTAL LE...	No. of A ty...	No. of C ty...	No. of E ty...	No. of G ty...	No. of I ty...
1	?	98.438	62	1.930	0.320	4.500	0.640	0.320
2	?	115.748	132	2.530	0	3.380	0.420	1.690
3	?	76.509	57	1.830	0.920	3.670	1.830	5.500
4	?	96.668	72	8.840	0	2.790	0	0.470
5	?	120.058	140	2.170	0.620	0.930	0	1.860
6	?	90.471	77	6.320	0.350	3.860	0.700	2.110
7	?	111.495	63	1.760	0	0	0	0.880
8	?	106.815	63	2.110	0	0.530	0	2.110
9	?	105.886	81	3.820	0	1.530	0.760	0.760
10	?	106.530	90	2.540	0	1.900	0.630	2.220
11	?	124.225	122	1.290	0	0.430	0	1.290
12	?	98.503	60	1.680	0.560	3.350	0	1.680
13	?	91.380	58	6.210	0	3.450	0.690	1.380
14	?	102.451	62	1.620	0	1.080	0.540	2.700
15	?	110.140	106	3.320	0.950	0.950	0.470	0
16	?	93.595	56	5.970	1	1.990	0	0
17	?	108.370	56	0.510	0.510	0.510	0	0.510
18	?	99.294	71	7.010	0	1.870	0.470	0.470
19	?	105.588	40	0	0	3.030	0	0
20	?	84.208	58	5.560	0.790	3.170	0	3.970
21	?	83.136	57	6.780	2.540	0	0	0.850
22	?	96.725	98	8.520	0	1.700	0	3.980

Table 12: Predictions output from "3076 linear regression process 28 Oct 13 (to predict FLO3) SIX regular attributes"

//Local Repository/processes/3086 linear regression process 6 Jan 14 (to predict FLO3) SIX regular attributes

iew ExampleSet (Retrieve 3086 linear regress [...] FLO1 in here = to predict FLO3) SIX pvs)

ExampleSet (21 examples, 2 special attributes, 6 regular attributes)

Ro...	V...	prediction(...	TOTAL LE...	No. of A ty...	No. of C ty...	No. of E ty...	No. of G ty...	No. of I ty...
1	?	113.787	98	3.590	0	0.510	0	1.540
2	?	76.156	65	5.150	0	4.120	4.120	2.060
3	?	97.378	86	5.230	0.580	1.740	1.740	0
4	?	119.601	107	2.580	0	0.650	0	0.650
5	?	97.394	95	5.630	0	2.820	2.110	0.700
6	?	105.134	40	0.780	0	1.550	0	0.780
7	?	107.108	89	3.020	0.430	0	0.430	2.160
8	?	100.610	105	5.950	0.600	3.570	0.600	0.600
9	?	108.721	102	1.840	0.610	1.230	1.230	0.610
10	?	113.049	90	3.810	0	0.480	0	0.950
11	?	110.286	80	3.530	0	1.180	0	0.780
12	?	95.646	96	7.460	0	5.220	0	2.240
13	?	102.165	51	2.420	0	0.810	0	2.420
14	?	94.309	57	9.920	0	0	0.760	0
15	?	105.566	91	3.920	0.780	0.780	0.390	0.780
16	?	106.408	78	1.820	0.360	1.460	0.360	1.460
17	?	105.134	86	3.860	0	4.720	0	0.430
18	?	105.295	49	2.260	0	0.450	0	1.360
19	?	108.151	76	2.840	0	0.710	0	2.130
20	?	104.731	118	6.440	0	2.580	0	3.430
21	?	103.962	117	2.990	0.750	2.990	0.750	2.240

Table 13: Predictions output from "3086 linear regression process 6 Jan 14 (to predict FLO3) SIX regular attributes

	Code	PRETEST FLO1 SCORE (x/200)	POSTTEST FLO1 SCORE (x/200)	Change in FLO1 (posttest - pretest)	PRETEST FLO3 SCORE (x/200)	POSTEST 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 14: Final results from the pre and post test, showing the changes for FLO1, FLO3, speed and effects (for types A,C,E,G,I)

	Code	PRETEST FLO1 SCORE (x/200)	PRE TEST TOTAL EFFECTS (TYPES A+C+E+G+I)	TOTAL EFFECTS (A + C + E + G + I) RELATIVIZED (1.0 = 41.10)	Code	POSTTEST FLO1 SCORE (x/200)	POST TEST TOTAL EFFECTS (TYPES A+C+E+G+I)	TOTAL EFFECTS (A + C + E + G + I) RELATIVIZED (1.0 = 41.10)
1	B5	160	14.66	0.36	B17	153	13.87	0.34
2	B10	160	16.54	0.40	B12	152	10.46	0.25
3	B11	156	6.30	0.15	B2	149	11.15	0.27
4	B7	154	5.29	0.13	B7	147	6.30	0.15
5	B2	150	17.11	0.42	B5	144	7.43	0.18
6	B1	150	18.12	0.44	B8	140	15.73	0.38
				0.32				0.26
7	B17	148	6.79	0.17	B16	140	17.74	0.43
8	B18	144	19.36	0.47	B19	140	8.23	0.20
9	B16	144	21.56	0.52	B18	140	17.01	0.41
10	B8	136	9.97	0.24	B10	134	17.15	0.42
11	B12	134	17.59	0.43	B11	133	10.03	0.24
12	B6	134	29.58	0.72	B21	132	24.86	0.60
13	B19	130	5.84	0.14	B4	130	24.77	0.60
14	B15	125	16.66	0.41	B1	118	10.55	0.26
15	B4	122	22.22	0.54	B9	116	25.96	0.63
16	B22	121	28.21	0.69	B14	113	11.83	0.29
				0.43				0.41
17	B9	120	15.09	0.37	B20	110	11.66	0.28
18	B14	110	14.10	0.34	B6	110	26.85	0.65
19	B13	106	24.24	0.59	B15	110	21.08	0.51
20	B20	106	31.42	0.76	B22	104	25.40	0.62
21	B3	100	38.88	0.95	B13	100	28.83	0.70
22	B21	84	32.26	0.78	B3	74	41.10	1.00
				0.63				0.63

Table 15: Data from the pre-test (Oct 28) and post-test (Jan 6), ordered by FLO1: showing the average total effect (of types A, C, E, G and I) for the upper and lower quartiles

	Code	PRETEST FLO3 SCORE (x/200)	PRE TEST TOTAL EFFECTS (TYPES A+C+E+G+I)	TOTAL EFFECTS (A + C + E + G + I) RELATIVIZED (1.0 = 41.10)		Code	POSTEST FLO3 SCORE (x/200)	POST TEST TOTAL EFFECTS (TYPES A+C+E+G+I)	TOTAL EFFECTS (A + C + E + G + I) RELATIVIZED (1.0 = 41.10)
1	B11	124.23	6.30	0.15		B5	119.60	7.43	0.18
2	B5	120.06	14.66	0.36		B1	118.00	10.55	0.26
3	B2	115.75	17.11	0.42		B2	113.79	11.15	0.27
4	B7	111.49	5.29	0.13		B11	113.05	10.03	0.24
5	B15	110.14	16.66	0.41		B12	110.29	10.46	0.25
6	B17	108.37	6.79	0.17		B10	108.72	17.15	0.42
				0.27					0.27
7	B8	106.81	9.97	0.24		B20	108.15	11.66	0.28
8	B10	106.53	16.54	0.40		B8	107.11	15.73	0.38
9	B9	105.89	15.09	0.37		B17	106.41	13.87	0.34
10	B19	105.59	5.84	0.14		B16	105.57	17.74	0.43
11	B14	102.45	14.10	0.34		B19	105.29	8.23	0.20
12	B18	99.29	19.36	0.47		B7	105.13	6.30	0.15
13	B12	98.50	17.59	0.43		B18	105.13	17.01	0.41
14	B1	98.44	18.12	0.44		B21	104.73	24.86	0.60
15	B22	96.73	28.21	0.69		B22	103.96	25.40	0.62
16	B4	96.67	22.22	0.54		B14	102.16	11.83	0.29
				0.41					0.37
17	B16	93.59	21.56	0.52		B9	100.61	25.96	0.63
18	B13	91.38	24.24	0.59		B6	97.39	26.85	0.65
19	B6	90.47	29.58	0.72		B4	97.38	24.77	0.60
20	B20	84.21	31.42	0.76		B13	95.65	28.83	0.70
21	B21	83.14	32.26	0.78		B15	94.31	21.08	0.51
22	B3	76.51	38.88	0.95		B3	76.16	41.10	1.00
				0.72					0.68

Table 16: Data from the pre-test (Oct 28) and post-test (Jan 6), ordered by FLO3: showing the average total effect (of types A, C, E, G and I) for the upper and lower quartiles

	Code	PRE TEST SPEED (wpm)	PRE TEST TOTAL EFFECTS (TYPES A+C+E+G+I)	TOTAL EFFECTS (A + C + E + G + I) RELATIVIZED (1.0 = 41.10)	Code	POST TEST SPEED (wpm)	POST TEST TOTAL EFFECTS (TYPES A+C+E+G+I)	TOTAL EFFECTS (A + C + E + G + I) RELATIVIZED (1.0 = 41.10)
1	B5	14	14.66	0.36	B1	12	10.55	0.26
2	B2	13	17.11	0.42	B21	12	24.86	0.60
3	B11	12	6.30	0.15	B22	12	25.40	0.62
4	B15	11	16.66	0.41	B5	11	7.43	0.18
5	B22	10	28.21	0.69	B9	11	25.96	0.63
6	B10	9	16.54	0.40	B2	10	11.15	0.27
				0.40				0.43
7	B9	8	15.09	0.37	B10	10	17.15	0.42
8	B6	8	29.58	0.72	B13	10	28.83	0.70
9	B18	7	19.36	0.47	B11	9	10.03	0.24
10	B4	7	22.22	0.54	B8	9	15.73	0.38
11	B7	6	5.29	0.13	B16	9	17.74	0.43
12	B17	6	6.79	0.17	B18	9	17.01	0.41
13	B8	6	9.97	0.24	B6	9	26.85	0.65
14	B14	6	14.10	0.34	B4	9	24.77	0.60
15	B12	6	17.59	0.43	B12	8	10.46	0.25
16	B1	6	18.12	0.44	B20	8	11.66	0.28
				0.38				0.44
17	B16	6	21.56	0.52	B17	8	13.87	0.34
18	B13	6	24.24	0.59	B15	6	21.08	0.51
19	B20	6	31.42	0.76	B3	6	41.10	1.00
20	B21	6	32.26	0.78	B19	5	8.23	0.20
21	B3	6	38.88	0.95	B14	5	11.83	0.29
22	B19	4	5.84	0.14	B7	4	6.30	0.15
				0.63				0.42

Table 17: Data from the pre-test (Oct 28) and post-test (Jan 6), ordered by speed: showing the average total effect (of types A, C, E, G and I) for the upper and lower quartiles

//Local Repository/processes/0030 Correlation matrix Oct 13 AND Jan 14 for error types A to Z – RapidMiner Studio 6.0.008 @ Adams-MacB

ExampleSet (Retrieve correlation matrix data for error types Oct 13 and Jan 14)

ult Overview | AttributeWeights (Correlation Matrix) | Correlation Matrix (Correlation Matrix)

Attributes	No. of A ty...	No. of B ty...	No. of C ty...	No. of D ty...	No. of E ty...	No. of F ty...	No. of G ty...	No. of H ty...	No. of I ty...	No. of J ty...	No. of K ty...	No. of Z ty...
No. of A typ	1	0.487	0.336	0.348	0.382	0.528	0.166	0.264	0.035	0.200	0.235	0.428
No. of B typ	0.487	1	0.004	0.490	0.064	0.534	0.026	0.700	0.111	-0.167	0.170	0.493
No. of C typ	0.336	0.004	1	-0.014	-0.124	0.225	0.158	-0.107	0.094	-0.052	0.355	-0.160
No. of D typ	0.348	0.490	-0.014	1	0.388	0.386	0.432	0.405	0.359	-0.308	0.219	0.364
No. of E typ	0.382	0.064	-0.124	0.388	1	0.346	0.486	0.143	0.095	0.336	0.005	0.595
No. of F typ	0.528	0.534	0.225	0.386	0.346	1	0.064	0.124	0.220	-0.137	0.180	0.383
No. of G typ	0.166	0.026	0.158	0.432	0.486	0.064	1	0.107	0.357	-0.051	0.157	0.393
No. of H typ	0.264	0.700	-0.107	0.405	0.143	0.124	0.107	1	-0.162	-0.136	-0.082	0.498
No. of I type	0.035	0.111	0.094	0.359	0.095	0.220	0.357	-0.162	1	-0.274	0.655	-0.095
No. of J type	0.200	-0.167	-0.052	-0.308	0.336	-0.137	-0.051	-0.136	-0.274	1	-0.302	0.072
No. of K typ	0.235	0.170	0.355	0.219	0.005	0.180	0.157	-0.082	0.655	-0.302	1	-0.253
No. of Z typ	0.428	0.493	-0.160	0.364	0.595	0.383	0.393	0.498	-0.095	0.072	-0.253	1

Table 18: The Correlation Matrix output from the Rapid Miner process called "0030 Correlation matrix Oct 13 AND Jan 14 for error types A to Z" (all students included) (NORMAL correlation)

Researcher Profile

Researcher Adam Gardiner

Education BA (Humanities)
University of Bristol, UK
MSc (Software Engineering)
University of Northumbria, UK

Contact English department, the Faculty of Liberal Arts
Huachiew Chalermprakiet University
02-312-6300 ext. 1431