Exploratory Data Analysis of Japanese EFL Students' Extensive Reading Form Responses: An Action Research Project

### *By Chris Elvin*

# Abstract

The author uses a Google form in conjunction with R programming to analyse students’ reviews of graded readers as well as biodata and concludes that digital is more advantageous than paper. By accepting reviews online, there is no deadline to interrupt the pleasure of extensive reading. Furthermore, by using a bespoke script their homework submissions could be monitored using a heat map, the students received feedback of their progress and preferences in the form of charts and tables, and data analytics facilitated insight about their learning which lead to informed decisions about teaching as action research. Topics covered include academic integrity, corpus analysis, plagiarism detection, authorship attribution, grouping, and grading. Statistical testing includes ANOVA, t-tests, and chi-squared analysis, and R packages comprised ggplot for graphs, quanteda for duplicate detection, caTools to split data into training and testing sets, rpart for authorship attribution, and ggdendro for class grouping.

# Key Words

action research, academic integrity, corpus studies, exploratory data analysis, extensive reading, grouping, grading, plagiarism, R programing, writing

# Introduction

If you have been teaching English as a foreign language (EFL) for a number of years, have you ever had the feeling that surprises are unlikely and that every year you can confidently predict the incoming students will be very similar to those of the previous cohort?

This is how I felt having taught at one university in Japan for more than a decade. I was used to highly motivated intelligent people whose independence I respected, and for certain tasks whom I could entrust to complete without the need for much monitoring or even interferement.

If my schedule were to change and I was asked to teach a different subject, my feelings would be the same, and this was the case when I was asked to prepare a reading and writing course in 2017 and administer an extensive reading component. This entailed students reading as many graded readers as they liked outside of class.

The benefits of extensive reading (ER) are well known; they become better readers (Day & Bamford, 1998), there are improvements in reading rates (Kusanagi, 2005), vocabulary acquisition (Horst, 2005), oral fluency (Cho and Krashen, 1994), writing proficiency (Janopoulos, 1986), reading motivation (Elley, 1991), and learner autonomy (Benson, 2011).

ER works well when learners can choose what they want to read from a wide variety of topics, when the reading material is easy, when learners have the opportunity to read as much as possible, and when it’s a pleasurable experience (Day & Bamford, 2002).

However, allowing students to read for pleasure as an ideal with little to show for it was potentially a recipe for disaster as some of them may not comply, so I asked my students to write short reports. Such summary writing would hopefully improve their comprehension (Champeau de Lopez, 1989) and also their writing (Robb & Susser, 1989).

I told them that they needn’t write any more than was necessary to prove they had read the book, and when the course was finished I would rubber stamp the summaries based on whether or not I thought the book had actually been read.

In class every week for the first five or ten minutes, students discussed their readers in pairs or small groups. This made reading social and it also gave me the opportunity to monitor their progress indirectly. At one campus the activity was invariably successful while at the other it was only a partial success. Some were enthusiastic but others looked as if they were biding for time, even for the lesson to begin, and In hindsight I should have realized that I had a problem on my hands.

In one lesson late into the semester, two members of my class approached me to complain that it was unfair that they had been asked to write summaries every week while other teachers were apparently only asking for three such reports, so we discussed this in class. I pointed out to them that the absolute minimum for extensive reading to be effective should be one book every two weeks (Nation & Wang, 1999), that their manuals recommended at least twenty books per semester (Durand *et al*., 2016), and that there was no maximum limit (Day & Bamford, 2002). I also suggested that if the writing took up too much of their time, that they could do the remaining reports in Japanese if they so wished — the important thing was to do the reading.

At the end of the semester, all of my class at the suburban Saitama campus handed in completed portfolios as did two thirds of my students from the Tokyo location. The other six, who were the silent ones during graded reader discussion time and who included the two who had complained, handed in very minimal portfolios of well-known stories and movies such as “Alice in Wonderland” and “Pirates of the Caribbean. Johnny Depp, he’s so cool!”

It was a great pity that they had not taken their reading more seriously, as you can learn a lot from books. If they had read *The Diary of a Young Girl*, for example, they could have learned that “laziness may appear attractive, but work gives satisfaction (Frank, 1993, p. 256).

The reason why I had not asked them to hand in assignments periodically was because I thought that the imposition of regular deadlines may have made extensive reading troublesome rather than the pleasure that it was supposed to be. In addition, some of them found it difficult to develop a regular reading habit and I also wanted to abide by the university’s philosophy of allowing freedom and responsibility for the individual to pursue his or her own self-realization (“History & Mission”, n.d.), so it was up to them when and how hard they worked.

The problem was not just that some had made a nominal effort, it was also that others had put in a great deal of work, and I was not able to reward them appropriately due to my transparent yet rigid grading system. Consequently, for the following year I decided to monitor their’ homework assignments more closely, give them guidance about academic responsibility, and develop a more honorable grading scheme.

If I asked for their homework digitally, there would be no need for periodic collections and I would be able to check them as soon as they arrived, so my research question was to find out if it were better for the class to change to digital, and how would that be for me.

I used a Google form for collection and R programming (R Core Team, 2013) to analyze their responses, and I reverted to English only to avoid scuppering their writing progress, because some of the previous handwritten Japanese was illegible, and because data analysis of just one language would be easier to manage and interpret.

# Action Research and Exploratory Data Analysis

According to Bradbury, “Action research is a democratic and participative orientation to knowledge creation. It brings together action and reflection, theory and practice, in the pursuit of practical solutions to issues of pressing concern (2015, p. 1).

Exploratory data analysis (EDA), on the other hand, employs graphical representations and statistical methods to look at and interpret data in order to discover patterns and trends. Sometimes the starting point for EDA may be an observation in class, which leads to a visualization, a hypothesis, a statistical test, perhaps, and then a decision to modify the question or take action. On other occasions a query may emerge from the data itself, such as a plot that reveals something unexpected, and as old questions are answered new ones arise.

If AR seeks to improve the methods and approaches of those involved, EDA can assist by representing data evidence that in turn may herald a change in teaching practice. EDA is therefore an ideal companion to AR as it guides practitioners to solutions in the form of visualizations and statistical proof and both are iterative and cyclical.

# Data Collection

The minimum number of fields a response form may require would be two, an ID field and a summary text box. I wanted to analyse my students’ responses in greater detail so my form had twelve: an email address, nickname, student number, book title, author, publisher, genre, summary, opinion, star ranking, campus, and gender.

Below are some suggestions to instigate effective learning and to ensure clean data:

1. Allow learners to edit their responses so that they can improve their writing.
2. A separate field for opinion encourages critical thinking and discourages plagiarism.
3. To mitigate against empty fields, make questions compulsory.
4. Multiple choice, checkbox and dropdown options safeguard cleaner data.
5. For privacy and fun, ask for students’ nicknames.
6. For gender inclusivity include a “prefer not to say” option.

# Academic Integrity

According to the International Center for Academic Integrity, there are six fundamental values that promote personal integrity in learning. They are honesty, trust, fairness, respect, responsibility and courage (Fishman, 2013). These are the core qualities that I strive to inculcate among my class to inspire then to be good citizens of this world.

Dealing with plagiarism is an important aspect of academic integrity. It’s not that I want to detect it but rather by maintaining its avoidance I can help to concomitate a fair and relaxed working environment.

First I tell the class my plagiarism rules and then I make sure that they know that I am vigilant in confirming its non-appearance. Fallout from failure to do so may mean that my grades would become worthless in the eyes of my students; If one person could get a good grade easily by deception, then everyone else’s may be questioned and therefore devalued.

My rules are not to copy sentences, nor copy them and change the words, and that identical strings of five words or more may be construed as cheating unless it is a collocation. I think that these guidelines are reasonably clear, concise and comprehensive to me, but how about for others?

According to Pennycook, plagiarism is rarely a simple black-and-white issue, in part because it requires an understanding of cultural differences about learning and literacy (1996). The ideal scenario for everyone would be to have no true positives (correctly identified plagiarists), no false positives (wrongful accusations) and no false negatives (successful cheats), but if we live in a less than ideal complex world where the interpretation of plagiarism is often gray, then mistakes will be made.

If the cost to the community of failure to detect a false negative is high in terms of devaluing grading, of equal if not more importance would be the meting out a false positive as it is unjust and damages morale.

Consequently, I practice the well-known maxim that it is better that ten guilty persons escape than that one innocent suffer (Blackstone, 1978, p. 358). This ratio may have been debated over time, and indeed inflated tenfold by Benjamin Franklin (Smyth, 1970, p. 293) but its validity ought not to be in doubt. My students are not criminals, of course, yet they need justice too and teachers should be seen to be just. My policy, therefore, was to have a very high threshold for type I errors at the expense of type II which would have to be condoned.

If learners copy whole texts or sentences, even replacing words with synonyms, it hardly shows learning and it violates the rules. When they borrow new vocabulary from their readers and spend time synthesising their summaries, of which there are always several cases, this sweat of the brow is sufficient enough to indicate endeavour or even scholarship and therefore for me to accept it.

This semester, there were just two cases of plagiarism that I was able to detect; one had copied her summary directly from the internet and the other had transcribed sentences directly from his reader. One more had written “Matilda’s mother was soaked in bingo” which at first aroused suspicion of software translation but was later revealed to be L1 interference.

After telling the class that my computer kept on discovering duplicate sentences, I received an apologetic email confession from someone who said that she had done so due to her lack of confidence in writing and that she was asking for consideration, so I told her to rewrite and resubmit.

She had somewhat randomly copied plain sentences from her graded reader, and the reason why I think I had failed to detect it was that her writing, lacking the cohesion of the original, looked uncannily like the work of a grammatically competent high intermediate.

Therein lay another problem. Those who copied exotically were more likely to be flagged than those who wrote in plain vanilla. At least my telling her about the computer issue had worked as a deterrent, but I wondered how many more cases I had failed to notice. How could I detect these ordinary duplicates?

## Duplicate Detection and Diagnosis

It would be very time consuming to identify every case of plagiarism using conventional methods such as pasting a string into a search engine. Time is limited and I believe I should be employed more usefully. If I had a corpus of student writing, however, I could at least distinguish those who were copying from the same source or from each other or from themselves.

I used *quanteda* (Benoit, 2018) to tokenize their summaries at the sentence level, counted the sentence types, of which the vast majority were singular, sorted the table of counts, and printed out the duplicates. Then I matched these copies with the summaries using a grep command (globally search a regular expression and print) to discover who was writing them and from which source. The final thing to do was to decide if the copies were really plagiarism or not, and what to do about it if they were.

After eliminating identical double postings, my corpus of roughly four thousand sentences had just forty-eight duplicates, nearly all of which were identified as coming from a single person who had mistakenly posted two similar assignments for the same book. One of the copies by two different learners referred to the same graded reader, but since it was not a natural expression it was considered a coincidence. As for the other sentences, they were short and not atypical of student writing. Consequently, I was able to conclude that copying from the book or from each other was either very rare or non-existent and that they deserved praise for their honesty and responsibility.

This tactic of distinguishing duplicates to curb cheating worked primarily as a disincentive, perhaps. Even if two people copied from the same book, they could miss each other by selecting different sentences, and those who were determined to avoid detection could either insert synonyms or errors into their work deliberately.

Thankfully, my students were generally earnest so I wasn’t particularly concerned. If it becomes an issue, I will create a corpus of five-grams and seek out these five-word strings in their summaries. The corpus will be several times larger than the present one and it will generate more noise and take longer to sift through, but if necessary it will have to be done. In the meantime, I would like to have faith in them as they have shown me their spirit.

There was just one more thing I wanted to try. Could it be possible to identify a student by analyzing the stylometric features of her writing?

## Authorship Attribution

Everybody’s writing must be different in some way or other, in terms of how much they write for their summary or opinion, or the variety of words they use, for example.

I wanted to find out if it were possible to use basic word frequencies in order to classify my students and thereby help indirectly to safeguard against plagiarism.

The problem of differentiating between fifty-eight dependent variables, the learners, exacerbated by the limited data of their writing, and a far from exhaustive list of features meant that it may have been challenging to build a practical model. Moreover, the summaries were not just samples of the their writing, they were also about specific books which were often shared. If two of them wrote about the same story, for instance, this may cause the algorithm to model the random noise of the training data and thereby miss the signal in the testing set.

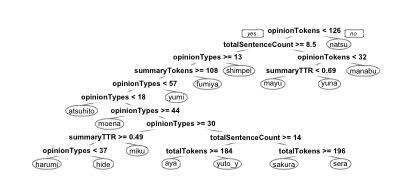
It may seem unlikely that such a bag-of-words approach to text analysis where context was ignored would yield meaningful results, but there were numerous real-world applications in which word frequency alone contained sufficient information for many types of analysis (Grimmer & Stewart, 2013). I wasn’t confident, but I decided anyway to see if it were possible to use classification and regression trees (CART) to identify individuals in class.

### Classification and Regression Trees (CART)

I used *quanteda* (Benoit, 2018) to create sentence counts for the students’ summaries and opinions, and similarly I generated word counts (tokens), different word counts (types) and type token ratios (TTR) to give me eight features. After totalling sentence counts, tokens and types, I used *dplyr* (Wickham & Francois, 2014) to select these categories as a new data frame of one dependent variable, their names, and eleven regressors.

Then *caTools* (Tuszynski, 2014) was used to split the data into training and testing sets, of which two-thirds were chosen to train, and finally *rpart* (Therneau, Atkinson, & Ripley, 2015) built the model and plotted the decision tree.

#### Tree Diagram of Student Writing



Starting from the top, decision trees represent a path with every node a binary option and every leaf an outcome. For example, If a student’s opinion had 126 words or more, predict Natsu. If it were less than 126 words, and the total sentence count was less than 8.5, and opinion tokens were equal or greater than 32, predict Manabu, and so on down the tree.

#### Results

When I ran the model on the testing set, it had an accuracy of twenty-one per cent, which was twelve times better and nineteen percentage points more than the base model of arbitrarily guessing. Thirteen people in the decision tree were identified correctly by their writing features with three being mistakes. Despite a positive predictive value of eighty-one per cent, the fact that forty-two were unclassified meant that it was a weak classifier, and since it had limited immediate utility I decided to give up.

I can imagine that it may be possible with more sophisticated feature extraction than simple sentence count, word token count and word type count, and with a cleverer algorithm such as random forests, that it may be possible to identify more people by their writing. In the long term, such knowledge may not only help to defeat plagiarism but also identify aspects of student writing that may need to be improved.

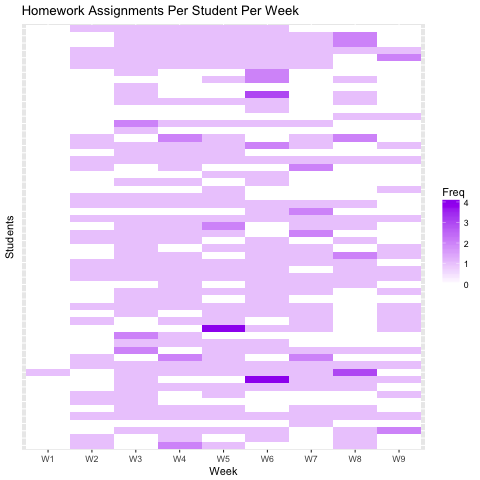
# Monitoring Homework Assignments

I remember many years ago at a teachers’ conference in Japan, the applied linguist Diane Larsen Freeman telling me how her presence in class as an observer had so visibly shaken the teacher that she had had to stop what she was doing and put down her pencil. This told me two things; that an ethnographer’s job is never easy, as one’s presence changes the landscape, and that few people enjoy being monitored, even teachers.

I wanted to oversee my class but I was wary of possibly being overbearing or intrusive. They knew that I was following them but I hoped that they didn’t mind and could see that I was doing it for altruistic reasons.

To visualize how often they did their homework assignments, I created a heatmap of individual postings per week using *ggplo*t (Wickham, 2009).

## Homework Heatmap



The graph was useful as it gave me a quick and easy way of visualizing their submissions. One thing to note was that only students who posted showed up on the chart, so I had to be alert to those who were missing. The other thing I should mention was that those who were editing their work via a copy from an automatic email reply had their timestamps updated, so it might appear as if some were posting four times in one week but the likelihood was that they were amending.

## Writing Volume

How much they wrote was an issue which emanated from a concern of the previous academic year when one campus appeared to write more than the other. I had no explanation as to why, so I decided to run a two-tailed t-test of total word count by campus rather than a one-sided one. I also investigated the differences between genders and plotted histograms using *ggplot* (Wickham, 2009).

### Campus

The result of a Welch two sample t-test showed that Saitama students (*M* = 225, *SD* = 132) wrote significantly more than those at Tokyo (*M* = 164, *SD* = 62), *t*(298) = 5.49, *p*<.001, *d* = 0.54, with Cohen’s d being calculated using the default settings of the R package *effsize* (Torchiano, 2017).

The diplomatic explanation for this difference would probably be to state that there was no explanation for it, but if pressed to do so this is what I may say. Firstly, even though all of my students had been placed in band level three as a result of a TOEIC ITP test, it was impossible to assign them to exactly equivalent classes as they were divided not only by location but also by department (social studies, psychology and welfare), and according to the Academic Affairs Division, the Tokyo class was at the lower end of the stream whereas the two others in Saitama were central.

Casual observation alone may not be so scientific, so I couldn’t say for sure, but if some had chosen Tokyo purposely because of its attractions and others had selected Saitama to avoid these distractions, if it were true, it could partially explain the difference in attitude to study as reflected by writing volume.

### Gender

Similarly, the result of a Welch two sample t-test indicated that the girls (*M* = 227, *SD* = 138) wrote significantly more than the boys (*M* = 174, *SD* = 65), *t*(260) = 4.48, *p*<.001, *d* = 0.47.

If pressed for an explanation for this difference, did I detect an air of entitlement in some of the boys at such a distinguished school? I’m not sure, but it was noticeable that the girls tended to enjoy reading more than the boys, and they were more social and talkative in class. Consequently their intrinsic and extrinsic motivation was greater.

As for the number of postings, a chi-squared test for given probabilities revealed that there was no difference in the number of postings per campus, *X2*(1) = .02, *p* = .89, nor between genders, *X2*(1) = 1.75, *p* = .19.

### Summary and Opinion

In the previous year, I had asked for summaries only, but there was a tendency to be mechanical and a shortage of self-expression. Therefore, this year I also included opinion.

The mean length of a student’s summary was 164 words whereas the average count for opinion was forty. A one sample t-test indicated that summaries were significantly longer than opinions, *t*(303) = 27.62, *p*<0.001, *d* = 1.61 (large).

The reason why summaries were longer than opinions was partly because they were expected to be and also that they were easier to do (Anderson & Krathwohl, 2001). My students were relatively unnaccustomed to expressing a point of view, as critical thinking was new to many, and coupled by its negative connotation it took time to for some of them to be galvanized to express more than a throwaway remark.

According to Bloom’s taxonomy of the cognitive domain (Anderson & Krathwohl, 2001), remembering and understanding were level one and two reasoning skills respectively, whereas applying, analyzing, evaluating and creating were at level three, four five and six. I therefore emphasized that they should be focussing on the more complex learning domains of application, analysis, evaluation and creation when writing their opinions. I suggested that they could apply the stories that they had read to their own lives or a new situation, for instance, or analyze the plot, or empathize with a character, or suggest a different ending.

However, It was perhaps not until the mid-term test when many dropped points for opinion that they realized what was expected of them. After the test, both the length of their opinions and the quality of them improved, so I hope that this will continue.

# The Learner Experience

There were two main purposes of providing feedback to them. One was to report their homework assignment progress, and the other was to inform them of their reading preferences.

Most lessons I handed individual homework assignment feedback to them if they had posted that week. This meant marking and updating the script on the eve of the lesson and then accessing it in class via my phone.

For other information, I sometimes showed data to them in class, but but most often I used the course blog. I posted their preferences for books, including star ratings, and also genres, authors, and publishers. I informed them about grading criteria, tests, grouping arrangements, warned them about copying, and offered them words of encouragement. I had hoped that such information would be informative and motivational, and furthermore, it was extra authentic reading practice for them. With over fifteen hundred page views per month, the website was being accessed on average around once per day per student.

## Books

A chi-squared goodness of fit test of frequency of borrowed books by genre, *X2*(12) = 139, *p* <0.001, was statistically significant. An inspection of the standardized residuals showed that mystery, action adventure, and fantasy were significantly more borrowed than other genres, while that of young adult, sport, classical literature, science fiction, and biography were significantly underrepresented. There was no need to tell my students this when the bar chart sufficed.

Some of these differences may be explained in part by the fact that not all genres were equally represented in the library, and that there was a propensity to mislabel, often selecting “fantasy” for example, for books which were not.

When they were asked to rate the books they had borrowed, there was a tendency to choose neutrality. The average rating for a book was 2.97 points out of five and a one-way analysis of variance (ANOVA) for such rating was not significant, *F*(12, 291) = 1.18, *p* =.29.

Neither a two-way analysis of variance (ANOVA) comparing genre selection with campus nor genre selection with gender was carried out, because the assumption that the value of the cells should be more than five in eighty per cent of them was not met.

## Groups

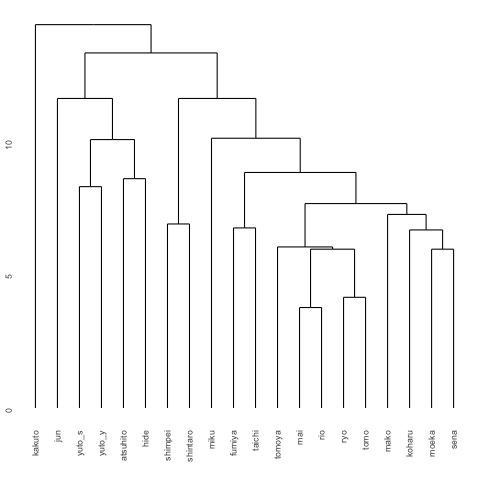
According to Dornyei and Ushioda, talking about books in class can turn the solitary extensive reading task into a socio-dynamic one that matches the students’ academic and social needs and may in some cases reinforce their intrinsic motivation to read (2011).

I was aware of an element of reticence for some to start extensive reading so I endeavoured to make my classes social. We always started off with group discussions about graded readers and ended with a class activity that related to the intensive reading of the lesson, for example student generated surveys. In this way, I tried to engender a community spirit that would make the lessons fun even for those who weren’t particularly enamoured about reading in a foreign language.

The two usual ways to arrange grouping in class would be for the students to choose where to sit for themselves or for the teacher to designate. In order to facilitate and maximize meaningful discussions about graded readers, I used a hierarchical clustering algorithm to compute pairwise distances between each student in the same class using the variables *book title*, *author*, *publisher*, *genre* and *stars*. Then I plotted a dendrogram using *ggdendro* (de Vries & Ripley, 2014), and assigned groups.

I was slightly apprehensive about using *stars* as a feature, even as a factor variable, as I had feared that I might end up with a group sharing a low opinion of their readers. After inspecting the clusters, however, that did not appear to be the case.

### A Dendrogram of Readers



When I organised the class into such groups during week seven, the discussions they had about their graded readers was no change in one class, the Tokyo class dominated by boys to the ratio of 7:3, but between four and five times longer than usual in the other two classes, and it would have been longer had I not interrupted them to start a new activity. I thought that most of them enjoyed the company of their kindred spirits, and if incoming data about their book preferences generates a new arrangement I hope they will like that, too.

# Homework Assessment

## Grading Criteria

The first semester of the reading and writing course was expected to be with an emphasis on reading and the latter on writing. However, the fact that they were writing so much meant that it behoved becoming part of their grades. A criterion-referenced test for such a course should therefore measure reading and writing (Glaser, 1963). I decided that I would assess their writing by how well they wrote their summaries, and for reading by how intelligently they expressed an opinion.

For each homework assignment, these were my bandscales:

0 plagiarism, or a mistake

1 a minimal effort

2 presents a more than minimal summary or opinion

3 a well written summary OR an intelligent opinion

4 BOTH a well written summary and an intelligent opinion

In order to spur extensive reading, there was no limit as to how many books they may read. They could write assignments as much as they liked until they attained the target of forty points. Even those who were neither strong writers nor advanced readers could reach this goal by reviewing the minimum recommendation set by the school of twenty books (Durand *et al*., 2016). They were encouraged to continue reading after reaching this target, and indeed they would still need a reader for class, but there was no need to continue to post. If there is a weakness in this grading system it is that those who score highly and obtain forty points earlier may appear to earn the option to read fewer books.

## Homework Assignment Scores

The average score for an assignment was 2.86 points out of four with a mean of 5.24 books per student halfway through week nine. This projected to an average of twenty-five points out of forty by the end of the semester, and with absences rare and thirty points for attendance almost guaranteed, most should have passed the course before their final intrinsic reading test at the end of the semester.

### Correlation of Assignment Scores and Writing Variables

It may have been possible to glean correlation of assignment scores with writing variables by introspection, since I was the one who was grading and knew somewhat subconsciously how I was marking. Statistical correlation saved time and would be more accurate.

All eleven of the variables measured for homework assignment writing correlated significantly with teacher assessment. The highest correlation was for total types, *r*(302) = 0.44, *p*<0.001, and the lowest was for summary sentence count, *r(*302) = 0.28, *p*<0.001. Therefore, a wide variety of words was valued more than the number of sentences even though both were important. The reason why sentence count could have been of lesser importance may be explained by the fact that the key criterion of error density was moderately independent of the length of their summaries. Overall, types were associated more strongly with teacher assessment than tokens, and opinion variables more so than summaries.

Linear regression of these eleven writing features followed by the step function revealed that summary tokens, summary types, opinion types and summary type-token-ratio were all influential variables and that these features taken together accounted for thirty per cent of their assignment scores. The negative sign of the coefficients of some of these similar variables suggested that multicollinearity was present.

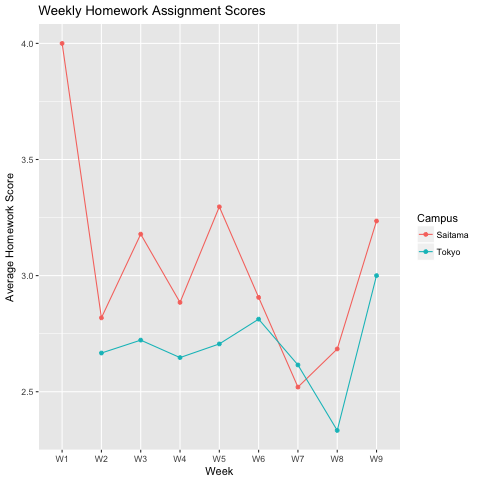
### Gender and Campus

There was a significant difference between females’ scores (*M* = 2.95, SD = .76) and males’ (*M* = 2.72, *SD* = .77), *t*(276) = 2.61, *p* = .010, *d* = 2.7. Similarly, Saitama students (*M* = 2.94, *SD* = .76) scored higher than Tokyo (*M* = 2.69, *SD* = .76), *t*(215) = 2.79, *p* = .006, *d* = .34. A two-way analysis of variance yielded a main effect for gender, *F*(1, 300) = 6.90, *MSE* = .58, *p* = .009. A Tukey’s HSD post hoc test revealed that females at Saitama scored significantly higher than males at Tokyo (*p* = .004).

## Homework Assignment Scores & Time

I was concerned about whether I was setting too much homework because of an incident in the previous year. Ironically, my course seemed to be even more demanding now. Would their work deteriorate over time due either to fatigue or a looming deadline, or both, or neither? Contrarily, would they become accustomed to writing and also respond to teacher feedback and gradually improve?

I plotted a line chart of the average assignment scores for each campus and followed it week by week to see if there was any pattern. By week five, I became perturbed by parallel lines.



The problem with such periodicity and its subsequent emergence was that it may have been revealing an underlying cause, and that was when I realized I may have overlooked something.

Rather than assume that the marking was consistent so that I could observe a trend in their writing over time, was it not also equally as possible if not more so that the their writing was consistent but my marking was fluctuating?

I will concede that since I could now check digitally, I often did so, on my cell phone on the train on the way home, for example. Did I deviate from the band descriptors and vacillate between objectivity and subjectivity, and why would I do so anyway?

The answer to that may lie partly in the transparency of my grading. If it is true that our personalities can be distilled down to five broad factors, one of which is openness (McCrae & John, 1992), then my lack of concealment concerning grading perhaps couldn’t be helped. Others may prefer a last-minute formula fixed to a favorite or villain perhaps but that would not fit for me.

To estimate a class’s capabilities before meeting them was not often straightforward given fairly broad bands, and this year the level was lower than last. With thirty percent of assignments failing by scoring two points or less, was I under pressure unconsciously to increment them?

A one-way analysis of variance (ANOVA) between assignment scores and week revealed a significant difference, *F*(8, 295) = 2.73, *p* = 0.006. A Tukey’s HSD post hoc test indicated that only the mean scores for week seven and week nine were different significantly, if only just (*p* = 0.05).

On reflection, the probability of the lines being parallel for three consecutive weeks between week two and week five was not unlikely at one in eight, and it was fairly subtle parallelism at that, so perhaps there was no cause for alarm. What I was made to realize, however, was that I should have been more circumspect in seeking out conceivably confounding variables. If I was marking any time, my grading became a function of time, so I was not controlling for it.

The scores that students had received up till now were temporary because I had been challenging them to polish their writing by editing and sharpen their critical thinking, too. The final score will be the one that I calculate after the course has finished, and this will also control for time. If the periodicity disappears after marking again, then I may have to concede that it was due to inconsistency, and if that is the case I might have to cease marking on the train.

The issue now was why weeks seven and nine were significantly different. The improvement may have been as a result of feedback about opinions after the poor performance during the mid-term test in week eight, or it could be that the more competent also had better time management skills so refrained from posting during this busy exam period, or it might also just be a coincidence.

There were many competing variables that may be underlying causes behind the assignment score chart; their natural development, fatigue, editing, the ups and downs of normal life, my environment, mood, and weekly interventions, other events, and other unknowns, so its interpretation may always be ambiguous. As a teacher, I cared less for such experiments than I did about their development. The data might have been illuminating but my only desire was that they can continue to make progress.

# Conclusion

There were three main advantages of using a digital form to collect students’ responses in conjunction with R programming; I was able to monitor their progress more closely, the learner experience was enhanced both inside and outside of class, and data analytics gave me knowledge of their learning and provided me with insight to make informed decisions about teaching, of which visualizations often helped.

A heatmap was plotted to follow homework assignment progress including their editing. Histograms were used to show their summary and opinion writing volume prior to running t-tests. CART was used to explore whether or not authorship attribution was attainable, and dendrograms were drawn to generate groups. ANOVA allowed me to observe whether or not there was an interaction between gender and campus, for example, and bar charts and chi-squared analysis gave me information about book preferences to pass on in class. Finally, correlation and linear regression were employed to determine which writing variables were most pertinent to assignment scores, albeit as an extrapolation of my subconscious mind.

The advantage of an analytics approach was that I could obtain facts about my students whenever I wanted to and that this could be used for immediate benefit to the class. I was able to be attentive about academic integrity, I kept them up-to-date about their favourite books or genres, the dendrogram was good for bonding, and I kept them abreast of their scores, which conversely could have been calculated by conventional means, if not quite as quick.

Some of what I learned this term was not because of data analysis, however, it was regular professional development. The reviews exposed my shortcomings as a writing teaching, for instance, so that will have to pick up. Specifically, I need to tell the class earlier what good writing is and what critical thinking entails. Therefore, I plan to use some of their outstanding examples as a teaching resource and also allow online access to them. In addition, I will start to itemize assignment feedback so that they will know for sure whether their scores were due to good writing or critical thinking, for example.

In a few week time, the semester will finish and everyone will go home. Might I then be able to generalize my findings to imaginary people whom I will never meet and predict that girls will write more than boys, for example, or the suburbs more than the city, given the evidence? Some people might argue that it would be irrational to do so whereas others could disagree. I would say that it is now largely irrelevant for me as it is simple, rational and pragmatic to discard the old data, input the new, and rerun the script.

Whether data analysis becomes a new direction for AR waits to be seen. It may even become an essential tool for all in education, but that might depend on how willing educators are prepared to embrace programming languages such as R or Python, for instance, or employ others to write bespoke scripts. If teachers convert to digital and become comfortable with code, its adoption will surely spread, but it will not be hasty and resistance likely. Without data I am lost, but like my students and many teachers, I prefer paper books, and there’s nothing wrong with that. The more likely scenario may be that in the not too distant future when paper and pencils are almost things of the past, the information teachers may need will be readily available and built into a portable device.

As for the previous procrastinators, had I accepted their homework digitally, I would either have been able to identify them earlier, or they would have worked anyway knowing that I was watching them, so I wonder what they would have done.

If the probability that a student is work averse is less than one per class, as I believe it is at this university, then it is inevitable that every twenty classes or so, there will be two indolent types per classroom. In my reading classes, which are always social, it is certain that such people would meet and in some cases have the capability of bringing down otherwise good folk with them. Would such learners have done any differently with stricter observation? I don’t know.

The work-shy can be resourceful when it comes to doing nothing and spending too much time dealing with it when there are others waiting to learn is neither wise nor good time management. I would have spoken to them and asked them to try extensive reading just for fun knowing that perhaps my words were falling on deaf ears, so in that sense there would be no difference between digital or paper.

Even if digital does have its merits, that does not mean my problems will be alleviated, as the students will presumably remain analogue. I am still waiting for one student to post his first homework assignment even though we are now in week nine. He told me that he has read four books and will post soon, so I hope he does. I am also concerned that the student who had copied from the internet has not posted since. She emailed me recently to say that she will upload soon so I am looking forward to that, too.

Pennycook (1996) reminds us about our roles as educators and reflect on whether we are setting good examples. Do we not sometimes use resources without attribution, for instance, or bring copyrighted material to class such as photocopied prints? We can’t expect learners to live up to standards we don’t abide by ourselves, can we?

Recently, I was interviewed for an academic writing post at one of Tokyo’s prestigious universities. I had brought along my own teaching materials for the course but they weren’t interested as they already had their own syllabus. Instead, I was requested to photocopy any of the commercially published textbooks that they had prescribed, meaning that these bright young undergrads would get three for the price of none.

I told the professor not unreasonably that he was roughly twenty-five years behind the times, thinking back to when I was a Master’s student having to wait in line at the photocopier to reproduce a teacher’s pack and thereby avoid copyright infringement.

His immediate reply was that the nineties were worse! “Oh God!”, he cried, clasping his face in his hands and muffling a laughter.

I was not invited back, unsurprisingly, and nor did I want to be, but it is an episode that may be on my mind if I ever get a chance to speak to my student in private. If there is a credible and excusable explanation, I could reconsider and remove my C ceiling for cheating but in all likelihood I will not. Whatever she may divulge, I might have to just smile and nod politely, Japanese style, and say to her, “I understand.”

# Acknowledgements

To Rob Waring for his advice about extensive reading.

# Notes

To view the annotated R script, visit [http://rpubs.com/celvin/](http://rpubs.com/celvin/400500)

To access the class blog, visit <http://www.elvinrw.blogspot.com>

For questions or comments, email [chriselvin@gmail.com](mailto:chriselvin@gmail.com).

# References

Anderson, L.W., & Krathwohl, D.R. (Eds.). (2001). A taxonomy for learning, teaching and assessing: A revision of Bloom’s taxonomy of educational objectives: Complete edition. New York, NY: Longman.

Benoit, K. (2018). quanteda: Quantitative analysis of textual data [computer software manual] (R package version 1.1.1). Retrieved from http://quanteda.io

Benson, P. (2011). *Teaching and Researching Autonomy*. London, England: Longman/Pearson.

Blackstone, W. (2016). The Oxford Edition of Blackstone’s: Commentaries on the laws of England: Book IV: Of public wrongs. Oxford, England: Oxford University Press. (Original work published 1783).

Bradbury, H. (2015). *The SAGE handbook of action research*. London, England: SAGE Publications.

Champeau de Lopez, C.L. (1989). Improved reading through writing. *TESL Reporter 22*(2), 23-26.

Cho, K., & Krashen, S. D. (1994). Acquisition of vocabulary from the Sweet Valley Kids series: Adult ESL acquisition. *Journal of Reading, 37*, 662–667.

Day, R, R., and Bamford, J. (1998). *Extensive reading in the second language classroom*. Cambridge, England: Cambridge University Press.

Day, R. R., & Bamford, J. (2002). Top ten principles for teaching extensive reading. *Reading in a Foreign Language, 14*, 136–141.

de Vries, A., & Ripley, D. (2014). ggdendro: Tools for extracting dendrogram and tree diagram plot data for use with ggplot2. (R package version 0.1-15). Retrieved from http://CRAN.R-project.org/package=ggdendro

Dornyei, Z., & Ushioda, E. (2011). *Teaching and researching motivation (2nd ed.)*. Harlow, England: Longman.

Durand, J.K., Hadingham, O.E., Kato, C.T., Mori, S., Nomura, Y., Rabbini, R., Saito, Y., Shrosbree, M., Yamamoto, Y., Yokomoto, K. (2016). *Reading the future: Developing reading skills in the college classroom (2nd Edition)*. Tokyo, Japan: Rikkyo University.

Elley, W. B. (1991). Acquiring literacy in a second language: The effect of book-based programs. *Language Learning, 41*.

Fishman, T. (Ed.). (2013). The fundamental values of academic integrity (second edition). *International Center for Academic Integrity*. Retrieved from https://academicintegrity.org/wp-content/uploads/2017/12/Fundamental-Values-2014.pdf

Frank, A. (1993). Anne Frank: *The diary of a young girl*. New York, NY: Bantam Books.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18*(8): 519-521. doi:10.1037/h0049294

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic analysis methods for political texts. *Political Analysis, 21*(3), 267-297. doi: 10.1093/pan/mps028

History & mission. (n.d.). Rikkyo University Official Website. Retrieved from http://english.rikkyo.ac.jp/about/history/index.html

Horst, M. (2005). Learning L2 vocabulary through extensive reading: A measurement study. *The Canadian Modern Language Review, 61*, 355–382.

Janopoulos, M. (1986). The relationship of pleasure reading and second language writing proficiency. *TESOL Quarterly, 20*(4). 763-768.

Kusanagi, Y. (2005). A class report 2: Course evaluation of pleasure reading course. *The Journal of Rikkyo University Language Center (14)*, 29–42.

McCrae, R.R., & John, O.P. (1992). An introduction to the five-factor model and its applications. Retrieved from <https://doi.org/10.1111/j.1467-6494.1992.tb00970.x>

Nation, P., & Wang, K. (1999). Graded readers and vocabulary. *Reading in a foreign language (12)* 355-380.

Pennycook, A. (1996). Borrowing others’ words: Text, ownership, memory and plagiarism. *TESOL Quarterly 30*, (2).

R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from [http://www.R-project.org](http://www.r-project.org)

Robb, T.N., & Susser, B. (1989). Extensive reading vs. skills building in an EFL context. *Reading in a Foreign Language 5*, 239-251.

Smyth, A.H. (1970). *The writings of Benjamin Franklin*. New York, NY: Haskell House Publishers. (Original work published 1906).

Therneau, T., Atkinson, B., & Ripley, B. (2015). rpart: Recursive partitioning and regression trees. (R package version 4.1-9). Retrieved from [http://CRAN.R-project.org/package=rpart](http://cran.r-project.org/package=rpart)

Torchiano, M. (2017). effsize: Efficient Effect Size Computation. (R package version 0.7.1). https://CRAN.R-project.org/package=effsize

Tuszynski, J. (2014). caTools: moving window statistics, GIF, Base64, ROC AUC, etc. R package version 1.17.1. Retrieved from [http://CRAN.R-project.org/package=caTools](http://cran.r-project.org/package=caTools)

Wickham, H. (2009). ggplot2: elegant graphics for data analysis. New York, NY: Springer.

Wickham, H., & Francois, R. (2014). dplyr: dplyr: a grammar of data manipulation. R package version 0.2. Retrieved from [http://CRAN.R-project.org/package=dplyr](http://cran.r-project.org/package=dplyr)