A Study of the Effect of Three Different Types of Feedback on Writing: Part 2 − Data Analysis

By
Peter Duppenthaler

Introduction

This is the second of a three-part series describing a one-year study involving the use of journals with a group of 99 second-year students at a Japanese girls' high school. As we have seen in Part 1, the students in the study are very similar in both ability level and background to those at our school, and therefore the findings should be applicable to our students. Part 1 included a brief introduction, the research questions, and information on the participants, site, materials, and procedures used in the study. Part 2 will be devoted to the statistical analyses of the data (for more detailed analyses, including additional charts, tables, and graphs see Duppenthaler, 2002). Part 3 will include a discussion of the findings, suggestions for further study, and conclusions.

As mentioned in Part 1, the study investigated the effect of three different types of teacher feedback on students' journal entries, and possible positive transfer effects on their in-class compositions. The three types of feedback were (a) meaning-focused feedback, (b) positive comments, and (c) error-focused feedback. Each treatment group, based on type of feedback, was formed by blocking the students, and each group contained an almost equal number of students from each of three original intact classes. Therefore, each treatment group was made up of a similar portion of students who were enrolled in classes which were taught by each of the teachers who taught the students in this study, and the only difference was the type of treatment they received.

Data Analysis

The main statistics in this study were run on a Macintosh LC III computer using the statistical software package Statistica 4.1 for Macintosh by StatSoft (1994). The data files were generated in Excel version 5.0 and then imported into Statistica.

The major research question was to determine if there were significant group differ-

ences. The goal of the analysis was to create a linear combination of eight dependent variables to maximize mean group differences. The dependent variables consisted of: (a) the total number of words, (b) total number of error-free clauses, (c) total number of clauses, (d) four vocabulary indices (i.e., Token%1, Token%2, Token%3, and TokenNot), and (e) the Flesch-Kincaid Readability Index. The independent variable was group assignment to one of three feedback treatment groups: Group 1 (meaning-focused feedback), Group 2 (positive comments), and Group 3 (error-focused feedback).

Assumptions

Procedures related to the identification of possible outliers, the evaluation of the assumptions of normality of sampling distributions, homogeneity or variance-covariance matrices, linearity, and multicollinearity were carried out following recommendations found in Tabachnick and Fidell (1996), and examined through various Statistica programs. The overall alpha level of this study was set at .05; however, a Bonferroni type adjustment was made in order to guard against inflated Type I error (i.e., rejection of a true null hypothesis). The adjusted alpha for all Multivariate Analysis of Variance (MANOVA) tests was set at .005 (the original alpha level of .05 divided by 10, the total number of MANOVA tests in the study). In addition an adjustment was made in the alpha level for all Univariate *F* tests. In this case, the adjusted alpha for the MANOVA tests (.005) was divided by the number of dependent variables (DVs). In the case of the pretreatment and posttreatment questionnaires the adjusted alpha was .0005 (.005/10 DVs), and in the case of the in-class writing samples and journal entries the adjusted alpha was .000625 (.005/8 DVs). As will be shown below, the results of the identification of possible outliers, and evaluation of the assumptions of normality of sampling distributions, homogeneity of variance-covariance matrices, linearity, and multicollinearity were found to be satisfactory.

Outliers

"Outliers are cases with such extreme values on one variable or a combination of variables that they distort statistics" (Tabachnick & Fidell, 1996, p. 65). "Univariate outliers are cases with an extreme value on one variable; multivariate outliers are cases with an unusual combination of scores on two or more variables" (Tabachnick & Fidell, 1996, p. 66).

Screening for univariate and multivariate outliers was carried out using the Statistica Outlier procedures whereby any case that is greater than ± 2.5 times the standard deviation is identified as being an outlier. Only one multivariate and one univariate case were found. Both of these cases were found to be in the in-class writing samples. "A rule of thumb is that one needs to be concerned if an observation (or observations) falls outside the mean ± 3 times the standard deviation" (StatSoft, 1994, p. 344). As neither case was greater than ± 3.0 times the standard deviation, neither was deleted from subsequent analyses. As a result, no cases were deleted. The total *N* was 99 with an equal number of 33 cases in each of the three treatment groups.

The balanced-design and large sample size were beneficial because, as cited in Tabachnick and Fidell (1996), Mardia (1971) shows that MANOVA, which was used later in the study, is robust to violations of normality if a sample size is larger than 20 degrees of freedom for error in the univariate case and if sample sizes are equal and two tailed-tests are used. In addition, the *F* test has been shown to be fairly robust to violations of the assumption of normality if there are no outliers.

Normality

Normality of the cloze test, in-class writing samples, and journal entries was checked by using Statistica Fit procedures. In a normal distribution the values of skewness and kurtosis are zero. A skewed variable is one whose mean is not in the center of the distribution. "If there is positive skewness, there is a pileup of cases to the left and the right tail is too long; with negative skewness, there is a pileup of cases to the right and the left tail is too long" (Tabachnick & Fidell, 1996, p. 71). "Kurtosis values above zero indicate a distribution that is too peaked with long tails, while kurtosis values below zero indicate a distribution that is too flat (also with too many cases in the tails). Nonnormal kurtosis produces an underestimate of the variance of a variable" (Tabachnick & Fidell, 1996, p. 71). "Because the standard error for both skewness and kurtosis decreases with larger *N*, with large samples the null hypothesis is likely to be rejected when there are only minor deviations in normality" (Tabachnick & Fidell, 1996, p. 73). An examination of the data indicate that not all of the variables were normally distributed.

Normality was also checked with normal probability plots and detrended probability plots. These also indicated that some variables were not normally distributed. However,

if the *n* per cell is fairly large, deviations from normality do not matter much

at all because of the *central limit theorem*, according to which the sampling distribution of the mean approximates the normal distribution, regardless of the distribution of the variable in the population (StatSoft, 1994, p. 389).

Homogeneity of Variances

In MANOVA, homogeneity of variance for each of the dependent variables is assumed. Homogeneity of variance-covariance matrices was examined through Box's *M* test. The results indicated that the group variance-covariance matrices were not equal. However, Tabachnick and Fidell (1996) state that, "if sample sizes are equal, robustness of significance tests is expected; disregard the outcome of Box's *M* test, a notoriously sensitive test of homogeneity of variance-covariance matrices" (p. 382).

Linearity

MANOVA assumes linear relationships among all pairs of dependent variables in each cell. Violations were expected because the variables were skewed. Linearity was checked using scatterplots. Several of the plots showed nonlinear relationships. Deviations from linearity reduce the power of statistical tests and this point should be taken into consideration when discussing results.

Multicollinearity

When correlations among dependent variables are high, one dependent variable is a near-linear combination with other dependent variables. In order to examine multicollinearity, correlations between all dependent variables were checked with Pearson product-moment correlations. "If variables are used and they contain, in part, the same items, correlations are inflated. Don't over interpret a high correlation between two measures composed, in part, of the same items" (Tabachnick & Fidell, 1996, p. 58). With this in mind it was assumed that some of the variables would show high correlations (e.g., in the 80s) and this was found to be the case. However, according to Tabachnick and Fidell (1996) "The statistical problems created by singularity and multicollinearity occur at much higher correlations (.90 and higher)" (p. 86). None of the correlations was .90 or higher.

Cloze Test

During the first week of school, the three intact classes of 99 second-year students were given a 40-item multiple-choice cloze test. All of the 99 students recorded their answers on computer mark-sheet cards. These cards are routinely used at the school and all of the students were familiar with them. I machine scored the cards and analyzed the data for reliability. "Reliability is usually defined as the extent to which a test produces consistent, accurate results when administered under similar conditions" (Hatch & Lazaraton, 1991, p. 530). There are several ways to estimate reliability. "Internal consistency methods are used when it is not convenient to collect data twice or to use parallel tests" (Hatch & Lazaraton, 1991, p. 535); as was the case in this study. Two methods for calculating internal consistency were used: Kuder-Richardson 20 and split-half adjusted. Kuder-Richardson 20 reliability for the cloze test was Cronbach alpha .77. The Split-half adjusted reliability for the cloze test was .82.

All second-year students are divided into five classes: one higher-level class, one middle-level class and three lower-level classes, based on their performance during their first year of high school. The students in the three lower-level classes (i.e., those who took part in the study) are then assigned to their three respective classes on the basis of alphabetical order. These students are therefore a rather homogeneous group of individuals. Reliability can be depressed by a number of different factors such as: a small number of items in the test, setting, time span, history, the homogeneity of the group being tested, and so on. As noted by Ary, Jacobs, and Razavieh, (1990), "The reliability coefficient increases as the spread, or heterogeneity, of the subjects who take the test increases. Conversely, the more homogeneous the group is with respect to the trait being measured, the lower will be the reliability coefficient" (p. 280). This is because the type of statistical analysis used in estimating reliability indexes the mean correlation among the items. Given the extreme likelihood of this being a very homogeneous group, it was felt that the level of reliability was acceptable for blocking purposes.

Students were blocked into the three treatment groups mentioned earlier. This was done by using the sort function in Microsoft Excel to sort the participants in descending order based on their cloze scores. Starting at the top and working through to the bottom of the list, the participants were assigned a number (one, two, or three) to represent the three treatment groups. This was done by assigning the number one to the first student on the list, number two to the second, number three to the third, number one to the fourth, number

two to the fifth, number three to the sixth, and then repeating this ordering sequence until all of the participants had been assigned to a treatment group. Each group consisted of exactly 33 students per group (for more on block design see Kirk, 1995).

In order to further check that the three groups were similar, a one-way Analysis of Variance (ANOVA) was performed using cloze test scores as the dependent variable and treatment group assignment, with 33 students in each treatment group, as the independent variable. There were no significant differences among the three treatment groups at $p = .9973$.

Pretreatment and Posttreatment Questionnaires

Pretreatment Questionnaire (Bilingual English and Japanese)

The Pretreatment Questionnaire consisted of ten questions designed to establish the students' degree of extracurricular exposure to English. It was given during the second week of school--treatment began in the fifth week--in order to check for any pretreatment differences among the three treatment groups. Questions 1, 4, 5, 8, and 9 were Yes/No questions. Questions 2, 3, 6, 7, and 10 asked for more detailed information relating to questions 1, 5, and 9. They were only answered by those students who had answered "Yes" to any of these three questions (i.e., 1, 5, and 9) and were skipped by those who had answered "No" to these questions. For example, Question 1 was, "Have you ever been to an English-speaking country?" while Question 2 had to do with length of time spent there. The Pretreatment Questionnaire questions were as follows:

Question 1: Have you ever been to an English-speaking country?

Question 2: How long were you there?

Question 3: How old were you at the time?

Question 4: Did you study English while you were there?

Question 5: Do you study English outside of school?

Question 6: Where do you study?

Question 7: How long have you studied in the place you circled in No. 6?

Question 8: Do you ever speak English with your family?

Question 9: Do you have pen pals in foreign countries?

Question 10: How often do you write to them?

Because questions 1, 4, 5, 8, and 9 were Yes/No questions they were coded using

"one" for yes and "zero" for no. The dichotomous nature of these questions meant that logistic regression, rather than ANOVA or Linear Regression, was the preferred method of analysis. This was because unlike ANOVA and Regression, in which the dependent variable should be continuous, "Logistic [Regression] is relatively free of restrictions, and with the capacity to analyze a mix of all types of predictors (continuous, discrete, and dichotomous)" (Tabachnick & Fidell, 1996, p. 578). In this type of analysis, if the chi-square is small, "then one concludes that the two variables are independent; a poor fit leads to a large chi-square . . . and the conclusion that the two variables are related" (Tabachnick & Fidell, 1996, p. 56). Logistic regression for questions 1, 4, 5, 8, and 9 in the Pretreatment Questionnaire showed small chi-square and *p* values which indicated that there were no significant differences among the three groups.

Pretreatment Questionnaire questions 2, 3, 6, 7, and 10 were on a scale, which meant that a MANOVA, rather than ANOVA, was the preferred method of analysis because the research design included more than one dependent variable. Like ANOVA, MANOVA is a statistical procedure for testing whether the difference among the means of two or more groups is significant. However,

> MANOVA has a number of advantages over ANOVA. First, by measuring several DVs [Dependent Variables] instead of only one, the researcher improves the chance of discovering what it is that changes as a result of different treatments and their interactions. . . . A second advantage of MANOVA over a series of ANOVAs when there are several DVs is protection against inflated Type I error [i.e., rejection of a true null hypothesis] due to multiple tests of (likely) correlated DVs (Tabachnick & Fidell, 1996, pp. 375-376).

A one-way MANOVA for Pretreatment Questionnaire questions 2, 3, 6, 7, and 10 also showed no significant differences among the three treatment groups at $p = .3999$.

An examination of the raw data showing how many students selected each option in the Pretreatment Questionnaire indicated that slightly less than half of the students in each group had been abroad, and that those who had been, were there for short periods of time between the ages of 11 and 18. About half of those who had been abroad studied English while they were there. As you will recall from Part 1, the school has a short-term, summer, study abroad program.

About half of each group studied English outside of school. Those who were studying English outside of school were fairly evenly distributed between: foreign language schools,

private Japanese English teacher, and "other" (in the blank under "other" the students wrote either "prep school" or "*juku*" [preparatory or cram school]). Most of those who had studied English outside of school had done so for over a year. The vast majority of the students never spoke English with either family or friends. Very few of the students had pen pals in foreign countries, and most of those who did, did not write often.

The analyses of the Pretreatment Questionnaire showed that there were no significant differences among the three treatment groups with regard to their degree of extracurricular exposure to English prior to the study.

Posttreatment Questionnaire (Bilingual English and Japanese)

The Posttreatment Questionnaire was given during the 37th week of school. The first ten questions were exactly the same in both the Pretreatment and Posttreatment Questionnaires. Questions 11 through 20, discussed below, appeared only in the Posttreatment Questionnaire. Logistic regression for questions 1, 4, 5, 8, and 9 in the Posttreatment Questionnaire showed small chi-square and p values which indicated that there were no significant differences among the three groups. A one-way MANOVA for Posttreatment Questionnaire questions 2, 3, 6, 7, and 10 also showed no significant differences among the three treatment groups at $p = .2611$.

An examination of the raw data showing how many students selected each option in the Posttreatment Questionnaire indicated that the numbers had remained almost exactly the same as in the case of the Pretreatment Questionnaire. In other words, there had been no changes with regard to extracurricular English activities during the course of the year.

Questions 11 through 20, which did not appear in the Pretreatment Questionnaire, were designed to determine (a) the degree of either positive or negative feelings the students had toward writing in their journals and (b) whether they felt the experience had been a positive one. A 5-point Likert scale was used for each question (1 = strongly agree, 2 = agree, 3 = neither agree nor disagree, 4 = disagree, 5 = strongly disagree). Question 20 included space for a written response. Students were free to write in either Japanese or English. All of the students wrote comments. I coded these using the same 5-point Likert scale used for the other questions. The questions were as follows:

Question 11: I enjoyed writing in my journal.

Question 12: I think writing in my journal had a positive effect on my English.

Question 13: I would like to continue writing in a journal next year.

Question 14: I enjoy writing in English more now than I did a year ago.

Question 15: I think my writing is better now than a year ago.

Question 16: I can express myself in writing more easily now than a year ago.

Question 17: I think writing in my journal was a good experience for me.

Question 18: Writing in my journal made me want to study English more.

Question 19: I looked forward to getting my journal back each week.

Question 20: Has writing a journal changed your attitude toward English?

A one-way MANOVA of questions 11 through 20 indicated no significant differences among the three groups at $p = .0007$; however, in order to interpret the results of the post-treatment questionnaire let us look at Table 1 which shows the group averages for questions 11 through 20. Once again, the three groups are: Group 1 (meaning-focused feedback), Group 2 (positive comments), and Group 3 (error-focused feedback).

Table 1. Posttreatment Questionnaire Questions 11 through 20 Averages

Question	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20
Group 1 average (n = 30)	2.3	1.9	2.2	2.2	2.8	2.6	1.8	2.9	2.0	2.5
Group 2 average (n = 30)	2.6	2.0	2.9	2.3	2.3	2.6	2.0	2.8	3.0	2.6
Group 3 average (n = 29)	2.8	2.0	3.2	2.4	2.8	3.0	1.9	3.3	2.2	2.7

For purposes of interpretation, the following standard was used: 1.8 - 2.3 = agree, 2.4 - 2.8 = less strongly, 2.9 - 3.3 = neutral. Using this standard we can interpret the above averages for each question as follows:

Question 11: Group 1 most enjoyed writing in their journals, with the other two groups enjoying the journals, but to a lesser degree.

Question 12: All groups agreed that writing in their journals had a positive effect on their English.

Question 13: Group 1 agreed they would like to continue writing in a journal next year, while

Groups 2 and 3 neither agreed nor disagreed.

Question 14: Groups 1 and 2 enjoyed writing in English more at the end of the year than a year earlier. This sentiment was shared to a slightly lesser degree by Group 3.

Question 15: Group 2 felt that their writing was better than a year earlier, and the other two groups agreed to a lesser extent that their writing had improved.

Question 16: Group 3 neither agreed nor disagreed that they could express themselves in writing more easily than a year earlier, but Groups 1 and 2 agreed to a slightly higher degree that they could do so.

Question 17: All groups felt that writing in a journal had been a good experience.

Question 18: Group 2 felt, to some extent, that writing in a journal made them want to study English more, but Groups 1 and 3 were neutral.

Question 19: Groups 1 and 3 looked forward to getting their journals back each week, but Group 2 was neutral on this.

Question 20: All groups agreed to some extent that writing journals had changed their attitude toward English. An examination of the written comments accompanying this question indicated that almost all of the comments were positive.

First In-class Writing Sample

The first of three, 35-minute, in-class writing assignments was given during the second week of school prior to any journal writing. This assignment was used (a) to establish a base-line estimate of the students' in-class writing ability, and (b) to further determine if there were any group differences prior to the outset of the study.

The assignments were collected and passed to me. Each assignment was typed and the total number of words was counted using Microsoft Word's word count function. The number of clauses and error-free clauses in each sample were counted by hand. Two raters, working independently, were used in order to establish interrater reliability. Prior to counting clauses and error-free clauses, the two raters met and agreed on a set of criteria to be used (discussed below). I was the first rater. I was responsible for all of the material and then selected a random sample of approximately 30% of the students' first in-class writing samples, copied them, and passed them to the second rater for analysis. The second rater, a native speaker of English with an M. Ed. in TESOL and several years of teaching experience in Japan, counted the number of clauses and error-free clauses in each sample, recorded the

scores, and then returned the scores and materials to the first rater. The first and second raters' scores for the same items were entered into an Excel file, imported into Statistica, and analyzed for interrater reliability using Pearson *r*. The same identical procedures and raters were used for all the clause and error-free clause counts used in the study (i.e., first, second, and third in-class writing samples, and all journal entries).

The first in-class writing samples' interrater reliability (Pearson *r*) for the two variables that were counted by hand were as follows: number of clauses (.96) and number of error-free clauses (.98). As all other items were computer-generated only a small random sample, consisting of five essays from each treatment group, was checked for accuracy by rerunning these through the computer programs. No inaccuracies were found.

Clauses are introduced in the second year of junior high school and were selected rather than T-units for a number of reasons. According to Richards, Platt, and Weber (1985) a T-unit is "the shortest unit which a sentence can be reduced to, and consisting of one independent clause together with whatever dependent clauses are attached to it" (pp. 299-300). Although a number of studies have been carried out using T-units, several researchers (Gaies, 1980; Pery-Woodley, 1991, Laufer & Nation, 1995; Ishikawa, 1995), have pointed out problems with the use of T-units. In addition, in a study with low-level Japanese learners, Ishikawa (1995) recommended against their use when attempting to examine "the efficacy of different experimental treatments for low-proficiency EFL writing" (p. 68).

Researchers often disagree about what constitutes correctness. Some (especially Larsen-Freeman & Strom, 1977) would consider a writing sample to be error-free only if correct in every respect. However, low-proficiency writers such as those who took part in this study often make mistakes, and requiring perfectly correct samples would amount to holding learners to impossibly high native-speaker standards. As a result, a number of concessions were made (a) punctuation, misplaced or omitted commas, misplaced or omitted punctuation used in or with direct quotations, and misplaced or missing apostrophes in plural possessives (e.g., women's' college) were disregarded. Misplaced or missing apostrophes in contractions were counted as mistakes; (b) capitalization, uncapitalized proper nouns and sentences not beginning with a capital letter were not counted as mistakes; (c) spelling errors were disregarded; (d) blanks or missing words other than articles rendered the clause or sentence incorrect; (e) use of the native language which either rendered the clause or sentence incorrect when there was a common, one-word equivalent in English, did not render the clause or sentence incorrect when expressing the concept in English would have required complex sen-

tence-structure(s) or sophisticated cultural knowledge, did not render the clause or sentence incorrect when the word (e.g., typhoon, tatami) was already fairly widely used in English-speaking countries, and did not render the clause or sentence incorrect when it was a proper name (e.g., Mr. Suzuki, Umeda [a geographical area within Osaka City]); (f) when two clauses were incorrectly joined, one clause was counted as incorrect; (g) a sentence beginning with a conjunction was counted as two clauses, but the first one was counted as incorrect. Several examples of this type of concession are: (1) correlative conjunctions (either...or, neither...nor) when only one was used in an otherwise grammatically correct sentence; (2) subordinate conjunctions followed by an otherwise grammatically correct sentence; (3) conjunctive adverbs (after, still) followed by an otherwise grammatically correct sentence (I went to school. After I had breakfast). In these cases the remainder of the sentence was grammatically correct. In order not to invalidate that, the unit was counted as two clauses, with the first one being counted as incorrect.

In the case of the first in-class writing sample, "quality" was measured by a battery of eight variables: (a) number of clauses (b) Token%1, (c) Token%2, (d) Token%3, (e) TokenNot, (f) Flesch-Kincaid, (g) Flesch, (h) Fog. (Variables g and h, two of the three readability indices, were dropped from later analyses for reasons which will be discussed below.) As stated above, clauses were counted by hand. Variables b through e (Token%1, Token%2, Token%3, and TokenNot), were obtained by running a vocabulary computer program called the VocabProfile (downloadable from http://www.vuw.ac.nz/lals/staff/nation.aspx). This program "shows the percentage of words a learner uses at different vocabulary frequency levels" (Laufer & Nation, 1995, p. 307). The program compares "a text against vocabulary lists to see what words in the text are and are not in the lists, and to see what percentage of the items in the text are covered by the lists" (VocabProfile manual, n. d., p. 1). The "lists" in this case are (a) the first most frequent 1000 words of English, (b) the second most frequent 1000 words, and (c) "words not in the first 2000 words of English but which are frequent in upper secondary school and university texts from a wide range of subjects" (VocabProfile manual, n. d., p. 3). All three of these lists include the base and derived forms of the words. The word forms in the base lists are grouped into word families under a headword (e.g., aid [the headword], aided, aiding, aids, unaided). For more information on word families see Bauer and Nation (1993). The sources of these lists are *A General Service List of English Words* (West, 1953) for the first 2000 words (i.e., Token%1 and Token%2), and "The University Word List" (Nation, 1990) for words in Token%3.

In short, the variable Token%1 is the percentage of words in the sample that is found in the list of the first 1000 most frequently used words in English, the variable Token%2 is the percentage of words in the sample that is found in the list of the next thousand words, the variable Token%3 is the percentage of words in the sample that is frequent in upper secondary school and university texts on a wide range of subjects, and the variable TokenNot is the percentage of words in the sample that is not found in any of the other three lists.

Given the level of the students' English, it was expected that the majority of the words contained in the students' writing would be within the first 1000 words (Token%1) list. "The first thousand words of *A General Service List of English Words* are usually those in the list with a frequency higher than 332 occurrences per five million words, plus months, days of the week, numbers, titles (Mr., Mrs., Miss, Ms., Mister), and frequent greetings (Hello, Hi, etc.)" (VocabProfile manual, n. d., p. 3).

In order to investigate the level of the vocabulary contained in the students' in-class writing samples, the samples were run through the VocabProfile program and the four Token numbers for each sample recorded. As stated above, the Token option counts each occurrence of the words in a given text, by base word lists, and provides the percentage of words in the given text found in each of the three lists, plus those not contained in the three lists (i.e., TokenNot). One is thus able to determine "the percentage of words a learner uses at different vocabulary frequency levels in her writing--or, put differently, the relative proportion of words from different frequency levels" (Laufer & Nation, 1995, p. 311). This information was used as an indication of the level of vocabulary contained in each sample.

The remaining three variables (i.e., variables f through h), were Readability Indices generated by RightWriter (version 3.1), a commercially available computer grammar/style program. A Readability Index is designed to indicate the level of education a reader will need in order to understand a given text. (For more on readability see Duppenthaler, 2000). RightWriter reports three readability indices: the Flesch-Kincaid, the Flesch, and the Fog. "Extensive testing of RightWriter's readability calculation shows an average error of less than 2%. This is usually lower than the error rate for calculations made by human operators. Further, RightWriter measures the readability for the entire document, not just samples. This eliminates sampling error" (RightWriter User's Manual, 1990, p. 7-5).

All three readability indices are based on the average number of words per sentence and the average number of syllables per word. The Flesch-Kincaid formula is the United States Government Department of Defense standard. The Flesch formula is widely used in

the insurance industry to check the readability of insurance policies. The Fog formula is used mainly in education.

All of the students' writing (both in-class writing samples and journal entries) was typed up and run through the spell checker program in Microsoft Word. Any spelling errors were corrected. This was done so that accurate estimates of vocabulary and readability could be obtained from the two computer programs (i.e., VocabProfile and RightWriter). Any words in Japanese were left unchanged. An examination of the raw data showed that these were very few, and either names of persons and places or words and phrases for which the students did not know the English and wrote the Japanese rather than guess.

Each student's in-class writing sample was run through the three computer programs in order to obtain the battery of variables, and the results were entered into an Excel spreadsheet. As all of the items (except for the clause and error-free clause counts) were computer generated, only a small random sample was checked for accuracy. No inaccuracies were found.

An examination of the data showed that no student had included any Token%3 vocabulary items in the first in-class writing sample. As there was no variance for this item, it had to be dropped from the analysis because the statistical program will not run when such a variable is present. Dropping this item had no effect on the overall analysis because it contributed nothing to indicating any group difference; all of the groups were exactly the same in that not one of the students had written any Token%3 vocabulary items.

A one-way MANOVA, with exactly 33 students in each group, was performed using total number of words, total number of clauses, total number of error-free clauses, three of the four vocabulary indices (i.e., Token%1, Token%2, and TokenNot), and the three readability indices (i.e., Flesch-Kincaid, Flesch, and Fog) as the nine dependent variables, and group assignment as the independent variable. There were no significant differences among the three treatment groups at $p = .0702$.

The results of the analysis of the cloze test, Pretreatment Questionnaire, and first in-class writing sample served to indicate that there were no significant group differences among the three treatment groups with regard to (a) prior or current English-language experience as assessed by the Pretreatment Questionnaire and (b) initial writing ability as assessed by the first in-class writing samples. It was therefore felt that no adjustments needed to be made in the make-up of the three treatment groups. During the remainder of the study, material was collected as it became available; however, due to time constraints, further analysis was delayed until after the end of the treatment period.

After gathering and processing all the raw data, a factor analysis was carried out on the journal entries. Journal entries were selected as the most appropriate for the purposes of a factor analysis because they represented the largest body of data and factor analysis requires several variables and a large number of cases to work properly.

Factor Analysis

Principle components analysis (PCA) and factor analysis (FA) are "statistical techniques applied to a single set of variables where the researcher is interested in discovering which variables in the set form coherent subsets that are relatively independent of one another" (Tabachnick & Fidell, 1996, p. 635). In addition, these techniques can be used

> to determine whether it is possible to reduce a large number of variables to one or more values that will still let us reproduce the information found in the original variables. These new values are called components or factors. They do not exist on the surface of the observed data but they can be captured by these analytic techniques (Hatch & Lazaraton, 1991, p. 490).

> Both PCA and FA include the following steps:
> selecting and measuring a set of variables, preparing the correlation matrix (to perform either PCA or FA), extracting a set of factors from the correlation matrix, determining the number of factors, (probably) rotating the factors to increase interpretability, and, finally, interpreting the results (Tabachnick & Fidell, 1996, p. 636).

In this study, these procedures were used both to confirm that the variables under consideration grouped together as they had been expected to and to investigate the possibility of reducing the overall number of variables while retaining those that offered the best picture of the overall processes involved. (For more on factor analysis see Duppenthaler, 1999.)

The default setting of the computer program automatically selects any factor with an Eigenvalue of greater than one. Using the default setting resulted in the identification of three factors with Eigenvalues of greater than one. These three factors fit nicely with those predicted to occur and accounted for 76% of the variance.

Variables are said to "load" on a particular factor when they have values of greater than plus or minus three, and the higher the number, regardless of sign, the more likely it is that the variable is part of the factor. Table 2 shows the unrotated solution.

Table 2. Factor Analysis Unrotated Solution Using Ten Variables

variable	Factor 1	Factor 2	Factor 3	Communalities
Token%1	-.61	-.68	-.30	.92
Token%2	.23	.55	.05	.36
Token%3	.28	-.02	-.56	.39
TokenNot	.57	.51	.50	.84
Flesch-Kincaid	.86	.23	-.28	.87
Flesch	-.84	-.31	.21	.85
Fog	.42	-.12	-.74	.74
Total Words	.71	-.56	.21	.86
Total Clauses	.59	-.75	.22	.96
Error-free Clause	.56	-.71	.23	.87
Variance	.36	.26	.15	.77

In order to determine more precisely which variables loaded on which factors, the data were submitted to a varimax rotation (see Table 3).

Table 3. Factor Analysis Varimax Rotated Solution Using Ten Variables

variable	Factor 1	Factor 2	Factor 3	Communalities
Token%1	-.96	-.02	.01	.92
Token%2	.55	.24	-.06	.36
Token%3	-.02	-.00	-.62	.39
TokenNot	.88	-.14	.18	.83
Flesch-Kincaid	.61	-.27	-.65	.87
Flesch	-.68	.22	.58	.85
Fog	-.06	-.10	-.85	.74
Total Words	.14	-.91	-.15	.87
Total Clauses	-.07	-.97	-.08	.95
Error-free Clauses	-.05	-.93	-.06	.87
Variance	.29	.28	.19	.76

The varimax rotated solution seemed to indicate that Factor 1 might be called "lexis," in that it is mainly composed of the vocabulary items generated by the VocabProfile program. The exception to this is the Flesch Readability Index which also loaded on this factor.

However, the Flesch Readability Index also loaded fairly highly on Factor 3, although not as highly as on Factor 1. Factor 2 might be viewed as "volume," which includes the total number of words, total number of clauses, and total number of error-free clauses; and Factor 3 as "syntactic," which includes both the Fog and Flesch-Kincaid Readability Indices and to a lesser extent the Flesch Readability Index. The exception to items loading on Factor 3 is the Token%3, words not found in the first 2000 words of English but frequent in upper secondary school and university texts.

At this point two decisions were made regarding which variables would be the most appropriate for later analyses. First, it was decided to drop the Flesch Readability Index for three reasons: (a) both the Flesch Readability Index and the Flesch-Kincaid Readability Index use the same basic formula; (b) the Flesch Readability Index did not load with the other syntactic factors; and (c) the Flesch Readability Index is widely used in the insurance industry to check the readability of insurance policies, while the students' writing would be at a much simpler level.

The second decision had to do with a problem that only became apparent after the first in-class writing sample. While gathering more information on the Fog Readability Index, I discovered that, "For the Fog Readability Index writing samples must consist of 100 words or more" (RightWriter User's Manual, 1990, p. 7-5). Although students had been asked to write between 200 and 250 words, several of them wrote fewer than 100 words. In short, the 100 word minimum for the Fog Index could not be met with any real assurance.

Finally, although Token%3 did not load with the other "lexis" items, it was a variable of interest because of its potential to indicate the use of higher level vocabulary and was therefore retained. I was therefore left with a total of eight variables: (a) total number of words, (b) total number of error-free clauses, (c) total number of clauses, (d) Token%1, (e) Token%2, (f) Token%3, (g) TokenNot, and (h) Flesch-Kincaid Readability Index.

A subsequent factor analysis using these eight variables resulted in the same variable groupings as before and accounted for 78 percent of the variance (see Tables 4 and 5).

Table 4. Factor Analysis Unrotated Solution Using Eight Variables

variable	Factor 1	Factor 2	Factor 3	Communalities
Token%1	-.01	.99	.06	.98
Token%2	-.17	-.63	-.33	.54
Token%3	.18	-.05	-.71	.54
TokenNot	.09	-.84	.31	.81
Flesch-Kincaid	.29	.07	-.75	.85
Total Words	.92	-.15	.05	.87
Total Clauses	.97	.07	.13	.96
Error-free Clauses	.93	.03	.10	.88
Variance	.35	.27	.16	.78

Table 5. Factor Analysis Varimax Rotated Solution Using Eight Variables

variable	Factor 1	Factor 2	Factor 3	Communalities
Token%1	-.02	.99	.04	.98
Token%2	-.21	-.67	-.27	.57
Token%3	.05	-.06	-.73	.54
TokenNot	.16	-.82	.31	.79
Flesch-Kincaid	.15	.06	-.79	.65
Total Words	.92	-.12	-.12	.88
Total Clauses	.97	.09	-.06	.95
Error-free Clauses	.93	.05	-.07	.87
Variance	.34	.27	.17	.78

Table 6 shows the final interpretation with Factor 1, "volume" being comprised of the total number of words, total number of error-free clauses, and total number of clauses; Factor 2, "lexis," being made up of the vocabulary items generated by the VocabProfile program, with the exception of Token%3, words in the sample that are frequent in upper secondary school and university texts; and Factor 3, "syntactic," including both the Flesch-Kincaid Readability Index and Token%3.

Table 6. Final Interpretation of Factor Analysis

Factor 1	Factor 2	Factor 3
	Token%1	
	Token%2	
		Token%3
	TokenNot	
		Flesch-Kincaid
Total Words		
Total Clauses		
Total Error-free Clauses		

Three In-class Writing Samples

Reanalysis of the First In-class Writing Sample

Taking into account the information obtained from the factor analysis and the decision to delete both the Flesch and Fog Readability Indices, it seemed prudent to reanalyze the first in-class writing sample using the new set of eight dependent variables. Therefore, another one-way MANOVA was performed using the eight variables of interest (minus Token%3 which, as mentioned earlier, lacked any variance). Group assignment was used as the independent variable. No significant differences were found among the three groups at $p = .6825$. Therefore this analysis was stopped.

Second In-class Writing Sample

The second-in class writing assignment was given during the 24th week of school, which meant that the participants had just completed half of the treatment time. One student in each of the three treatment groups was absent; however, these students wrote their samples after school during the same week, under the same conditions as the other students. As in the case of the first in-class writing sample, each sample was typed up and the total number of words were counted using Microsoft Word's word count function. The number of clauses and the number of error-free clauses in each sample were independently counted by hand by the two raters. The interrater reliability for these two variables was as follows: number of clauses (.99), and number of error-free clauses (1.0). As the other items were com-

puter-generated, only a small random sample, consisting of five essays from each treatment group, was checked for accuracy. No inaccuracies were found.

A one-way MANOVA was performed using the eight variables of interest as the eight dependent variables and group assignment as the independent variable. No overall significant difference was found at $p = .0176$. Therefore this analysis was stopped.

Third In-class Writing Assignment

The third in-class writing assignment was given during the 37th week of school, the last week of the treatment period. This time two students were absent from each treatment group; however, they wrote their samples, under the same conditions, during the same week, one day after school. The interrater reliability for the two variables that were counted by hand was as follows: number of clauses (.99), and number of error-free clauses (.99). As the other items were computer-generated, only a small random sample, consisting of five essays from each treatment group, was checked for accuracy. No inaccuracies were found.

A one-way MANOVA was performed using the eight variables of interest as the dependent variables and group assignment as the independent variable. No overall significant differences were found at $p = .0146$. Therefore this analysis was stopped.

The above analyses of the three in-class writing samples were carried out as the data became available. There were no significant differences among the three treatment groups with regard to the three in-class writing samples. However, one should keep in mind that journal writing and in-class writing are two very different types of writing tasks and that it may be unrealistic to expect to find any transfer effect. I will have more to say about this in Part 3.

Journals

As with the in-class writing samples, all of the journal entries were typed up, any spelling mistakes were corrected, the typescripts were checked against the originals for accuracy, and the total number of words was counted using Microsoft Word's word count function. The number of clauses and the number of error-free clauses in each sample were independently counted by hand by the two raters. The interrater reliability for the two variables that were counted by hand was as follows: number of clauses (.98), and number of error-free

clauses (.99). As the other items were computer-generated, only a small random sample, consisting of five journals from each treatment group, was checked for accuracy. No inaccuracies were found.

On average, students tended to write between 40 and 50 words each week. The occasional student who failed to write anything in her journal for a particular week received a zero word count for that week.

All of the entries the students in one treatment group wrote during the course of one week were combined into a single group entry for that week. This resulted in 22 entries for each treatment group, one for each week the students wrote in their journals. All of the data on the eight variables of interest, the same as those in the in-class writing samples, were entered into an Excel file and then imported into Statistica for analysis.

A one-way MANOVA was performed using the eight variables of interest as the dependent variables and group assignment as the independent variable. An overall significant difference was found at $p = .0000$.

Univariate F tests, with an adjusted alpha of .0006, indicated that there were three significant differences. They were (a) total number of words at $p = .0000$, (b) total number of clauses at $p = .0000$, and (c) total number of error-free clauses at $p = .0000$ (see Table 7).

Table 7. Univariate F Tests with Degrees of Freedom (2, 63) Table of Specific Effects for Journals (22 Entries)

Depend. Variable	Mean Sqr Effect	Mean Sqr Error	f(df1,2) 2,63	p-level
Token%1	3.	2.5	1.04400	.3580549
Token%2	1.	.7	1.25751	.2914000
Token%3	0.	.1	.20267	.8170776
TokenNot	2.	1.6	1.22515	.3006144
Flesch-Kincaid	0.	.1	3.41960	.0389135
Total Words	2634792.	119854.0	21.98335	.0000001*
Total Clauses	18340.	869.4	21.09421	.0000001*
Error-free Clauses	3515.	291.3	12.06881	.0000365*

* $p < .0006$

In order to determine more precisely where significant differences were, post hoc comparisons were conducted using the Scheffé test. The result of the Scheffe test on the

Total Number of Words revealed a significant difference between Groups 1 and 2 at p = .0000, and between Groups 1 and 3 at p = .0000, but no other significant differences (see Table 8). An examination of the means (see Table 8) showed that Group 1 (2070) had a significantly higher mean than either Group 2 (1407) or Group 3 (1565). Group 1 wrote significantly more words than either Groups 2 or 3, and Group 3 wrote more words than Group 2 (i.e., in descending order the three groups are Group 1, Group 3, Group 2).

Table 8. Scheffé Test for Journals (22 Entries) Total Number of Words

GROUP (mean)		{1} (2070.13)	{2} (1407.63)	{3} (1565.36)
1	{1}		.0000002	.0000480
2	{2}		.0000002*	.3258153
3	{3}	.0000480*	.3258153	

* p < .0006

Post hoc analysis (Scheffé test) on the total number of clauses revealed a significant difference between Groups 1 and 2 at p = .0000, but no other significant differences (see Table 9). An examination of the means showed that Group 1 (168) had a significantly higher mean than Group 2 (110). Group 1 wrote significantly more clauses than Groups 2, and Group 3 wrote more clauses than Group 2 (i.e., in descending order the three groups are again Group 1, Group 3, Group 2).

Table 9. Scheffé Test for Journals (22 Entries) Total Number of Clauses

GROUP (mean)		{1} (168.636)	{2} (110.954)	{3} (137.454)
1	{1}		.0000001	.0036297
2	{2}	.0000001*		.0156723
3	{3}	.0036297	.0156723	

* p < .0006

Post hoc analysis (Scheffé test) on the total number of error-free clauses revealed a significant difference between Groups 1 and 2 at p = .0000, but no other significant differences (see Table 10). An examination of the means showed that Group 1 (82) had a signifi-

cantly higher mean than Group 2 (56) and that Group 3 (67) had a higher mean than Group 2 (i.e., in descending order the three groups are again Group 1, Group 3, Group 2).

Table 10. Scheffé Test for Journals (22 Entries) Total Number of Error-free Clauses

GROUP (mean)		{1} (82.0909)	{2} (56.9545)	{3} (67.1818)
1	{1}		.0000404	.0194435
2	{2}	.0000404*		.1472480
3	{3}	.0194435	.1472480	

* $p < .005$

In summary, the above journal analyses indicated that there were significant differences in the groups but did not indicate if these differences existed at the outset of the study or had developed over time. In order to determine if there had been any change over time, a pre/post design was adopted using the first four entries (i.e., first month of treatment) as the "pre" and the last four entries (i.e., the last month of treatment) as the "post."

Pre/Post Analysis Using the First Four Weeks and Last Four Weeks of Journal Entries

In order to determine if any significant differences had existed among the three treatment groups at the beginning of the treatment, a one-way MANOVA was performed using the "pre" (i.e., first four weeks of journal entries) eight dependent variables of interest and group assignment as the independent variable. There were no significant differences among the three treatment groups at $p = .7536$.

As in the case of the first in-class writing sample, there were no significant differences among the three treatment groups with regard to their first four (i.e., first month) of journal entries. In addition, although there were significant differences in the total number of journal entries (22-entries) analysis there were no significant differences, with an adjusted alpha of .005, found in the students' last four journal entries (i.e., last month) at $p = .0124$. Therefore this analysis was stopped.

As can be seen from the above, no significant differences existed among the three treatment groups during either the first or last four weeks of journal writing but significant differences were found in the analysis of journal entries 1 through 22 (i.e., during the course of

the treatment period). This would seem to indicate that, as others (Kreeft, 1984; Casanave, 1994) have stated, meaning-focused feedback in journals may be more effective than other types of feedback. However, there was no indication of a pre-post difference making interpretation somewhat difficult in this case--the meaning focused group outperformed the other two groups but their progress did not show a steady increase over time as measured by the first and last four weeks of journal entries.

In order to be able to supplement the preceding analyses of both the students' in-class and journal writing, a time series analysis of both the in-class writing samples and the journal entries for each of the eight variables of interest was carried out using Statistica's Time Series Module. The advantage to doing this is that the resulting graphs provided a visual representation of change over time which made it much easier to get a feel for both significant and nonsignificant, but no less interesting, changes which occurred in the students' in-class writing samples and journal entries during the course of the academic year. Due to space limitations, it is impossible to include these graphs here. Interested readers should refer to Duppenthaler (2002).

Time Series

"The analysis of time series is based on the assumption that successive values in the data file represent consecutive measurements taken at equally-spaced intervals" (StatSoft, 1984, p. 849). In order to run the Statistica's Time Series Module, there must be at least five "times" for each variable. In the case of the journals, this did not present any problem; however, only three in-class writing samples were gathered (i.e., there were only three "times"). In order to overcome this problem, two dummy times were added to the beginning of the series, followed by the three actual in-class samples. In order to avoid visual scaling problems that might have occurred if the two dummy times were set at zero, the lowest mean in each variable for the three groups was used as the dummy code for that variable's graph. As far as the journals were concerned, there was no need to include any dummy entries as there was a sufficient number of "times" in the material.

In-class Time Series

As mentioned above, there were no significant differences in the three in-class writ-

ing samples; however, there was, one interesting difference in the second in-class writing sample (the number of error-free clauses) and one interesting difference in the third in-class writing sample (TokenNot, words not found in any of the other vocabulary lists). Let us now look at each of these in turn.

In the second in-class writing sample, Groups 1 wrote far more error-free clauses than Group 3. An examination of the total number of error-free clauses graph showed that Group 1 (meaning-focused feedback) was the only group that made a steady increase in the number of error-free clauses over time.

Comparing this to the in-class writing Total Clauses graph showed that all three groups made progress in the number of clauses they wrote, but that Groups 1 and 3 outperformed Group 2 in the third in-class sample. Group 2's progress was also not as dramatic as that of the other two groups. Again, this may indicate that positive comments are not a particularly effective type of feedback.

In the case of the third in-class writing sample, Group 3 had also used far more TokenNot words than Group 1, even when taking into account Japanese vocabulary items and TokenNot words which were the same for each group. This may indicate a tendency on the part of Group 3 to use the basic words they know and to supplement them with words that they look up in a dictionary, while Groups 1 and 2 tend to use slightly more Token%2 vocabulary items than Group 3. It is tempting to interpret this difference as an indication that Groups 1 and 2 have acquired a higher level (i.e., Token%2 vocabulary) of working vocabulary than Group 3. However, this might indicate either a lack of desire for risk-taking, which might be a result of Group 3's treatment (i.e., error-focused feedback), or simply a smaller working vocabulary. A longer study might be able to shed some light on this point.

Journal Time Series

In the case of the journal entries, two types of graphs were generated: graphs showing the entire 22 entries, and pre/post graphs (i.e., first four weeks/last four weeks). As discussed above, there were no significant differences in either the "pre" or "post" journal entries. However, there were three significant differences in the 22-entries analysis. These were: the total number of words, the total number of clauses, and the total number of error-free clauses.

An examination of the pre/post graph for the Total Number of Words showed a general tendency for the three groups to move together in a pattern of a quick increase from the

first to second entries, followed by a gradual decline (perhaps indicating that the initial novelty of writing in a journal quickly wears off), followed by an increase toward the end of the treatment period. This same pattern was also noted by Casanave (1994) in her study.

The graph showing the Total Number of Words for the entire 22 weeks of journal writing indicated that topic had a lot to do with the number of words written. It is interesting to note that Brunette (1994) found a significant relationship between the number of words per entry and topic assignment on the individual level, but not on the group level. The students in this study were free to write on any topic they wanted to and no topics were assigned. However, the peak that could be seen in entry eight corresponded to the first journal entry following the school's Culture Festival, and all of the students had a lot to say about their preparations for and participation in the event. Likewise, the peak in entry 12 corresponded to the students' three-day school trip to Nagasaki. However, regardless of these two peaks, the general pattern is the same: after starting out almost together in entry 1, Group 1 consistently wrote more than either of the other two groups while Group 3 had a tendency to outperform Group 2. Again, this may indicate that positive comments are not a particularly motivating type of feedback.

The pre/post graphs for the Total Number of Clauses showed that all groups made some early progress followed by a decline, but then the differences among the three groups become much more pronounced. Group 1 made steady progress while Groups 2 and 3 make only moderate progress at the very end of the study, finishing far below Group 1.

The graph showing the Total Number of Clauses in the 22-entries showed that although Group 1 started out between Groups 2 and 3, by the time of the third entry Group 1 had overtaken the other two groups and maintained this position for the remainder of the treatment, with the exception of weeks 14 and 17 in which Group 3 had the highest mean.

The same general pattern can be seen in the pre/post graph and 22-entries graph for the total number of Error-free Clauses. We can see that all groups make some early progress followed by a decline, but then Group 1 has a tendency to outperform the two other groups.

This concludes Part 2, the various analyses of the data. In Part 3, I will discuss the findings, make suggestions for further study, and draw some conclusions based on the analysis of the data.

References

Ary, D., Jacobs, L. C., & Razavieh, A. (1990). *Introduction to research in education* (4th ed.). Fort Worth: Harcourt Brace College Publishers.

Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journals of Lexicography, 6* (3), 1-27.

Brunette, K. (1994). Adult ESL writing journals: A case study of topic assignment. *The ORTESOL Journal, 15*, 60-71.

Casanave, C. P. (1994). Language development in students' journals. *Journal of Second Language Writing, 3* (3), 179-201.

Duppenthaler, P. (1999). Test reliability and validity: Using statistical analysis to establish upper and lower boundaries of reliability and to estimate validity. *Pacifica, 9/10*, 1-16.

Duppenthaler, C. (2000). Readability measurements of some English readers used in Japanese high schools. *Baika Review, 33*, 35-45.

Duppenthaler, P. (2002). *Feedback and Japanese high school English language journal writing*. Unpublished doctoral dissertation, Temple University, PA.

Gaies, S. J. (1980). T-unit analysis in second language research: Applications, problems and limitations. *TESOL Quarterly, 14*, 53-60.

Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. Boston: Heinle and Heinle.

Ishikawa, S. (1995). Objective measurement of low-proficiency EFL narrative writing. *Journal of Second Language Writing, 4* (1), 51-69.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole Publishing Company.

Kreeft, J. E. (1984). *Dialogue journal writing and the acquisition of grammatical morphology in English as a second language*. Unpublished doctoral dissertation, Georgetown University, Washington, DC.

Larsen-Freeman, D., & Strom, V. (1977). The construction of a second language acquisition index of development. *Language Learning, 27* (1), 123-134.

Laufer, B, & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*(3), 307-322.

Mardia, K. V. (1971). The effect of nonnormality on some multivariate tests and robustness to nonnormality in linear models. *Biometrika, 58* (1), 105-121.

Nation, I. S. P. (1990). A University Word List. In I. S. P. Nation (Ed.), *Teaching and Learning Vocabulary* (pp. 235-239). Boston, Heinle and Heinle.

Pery-Woodley, M. (1991). Writing in L1 and L2: Analyzing and evaluating learners' texts. *Language Teaching, 24* (2), 69-83.

Richards, J., Platt, J., & Weber, H. (1985). *Longman dictionary of applied linguistics*. London: Longman.

RightWriter User's Manual. (1990). *User's manual*. Carmel, IN: Que Software, a Division of Macmillian Computer Company.

StatSoft. (1994). *STATISTICA for the Macintosh*. Tuls, OK: StatSoft, Inc.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: HarperCollins College Publishers.

VocabProfile User's Manual. (n. d.). Wellington, New Zealand: Victoria University, English Language Institute. (downloadable from http://www.vuw.ac.nz/lals/staff/nation.aspx).

West, M. (1953). *A General Service List of English Words*. London: 1953.