AN EFL READABILITY INDEX

JAMES DEAN BROWN
University of Hawai'i

This study explores readability and its relationship to the cloze passage performance of EFL students. Fifty reading passages were randomly selected from an American public library and made into 30-item cloze passages by deleting every 12th word. The subjects were 2298 EFL students from 18 university level institutions in Japan. Each student was randomly selected to take one of the 30-item cloze passages. Any differences between the cloze passages were therefore assumed to be due to other than sampling differences.

The result was a set of 50 cloze passages the means of which served as the dependent variable: EFL Difficulty. Each passage was then analyzed for two sets of independent variables chosen to investigate how well they predict the EFL Difficulty: the first set was made up of various first language readability indices (including the Flesch, Flesch-Kincaid, Fry, Gunning, Fog, and modified Gunning-Fog indices); the second set was made up of quantifiable linguistic characteristics of the passages (e.g., the percent of function words, number of syllables per sentence, number of words per paragraph, frequencies of words in the passages, and many others).

Correlational, factor and multiple-regression analyses indicated that the first language readability indices were only weakly related to EFL Difficulty (23 to 30 percent). However, the analysis of linguistic characteristics indicated clear groupings among the variables. In addition, four of the linguistic characteristics (number of syllables per sentence, the average frequency of lexical items elsewhere in the passage, percent of words with seven or more letters, and percent of function words) when combined were more highly related to EFL Difficulty (55 percent). These results are discussed in terms of their implications for the development of an EFL readability index.

INTRODUCTION

The cloze procedure first appeared in the literature when Taylor (1953) investigated its value as a device for estimating the readability of materials used in public education. Research has also investigated the effectiveness of cloze procedure as a measure of reading ability for native speakers of English, and, in the seventies, a number of studies also explored the effectiveness of cloze as a measure of overall ESL/EFL proficiency (for overviews on cloze research, see Alderson, 1978; Oller, 1979). After brief discussion of these developments this paper will review efforts that have gone on in both the first and second language readability literatures.

University of Hawai'i Working Papers in ESL, Vol. 15, No. 2, Spring 1997, pp. 85-119.

What is Cloze Measuring?

In the first language literature, numerous studies indicate that cloze scores are moderately to highly correlated with various standardized reading comprehension tests (Weaver & Kingston, 1963; Ruddell, 1964; Bormuth, 1965, 1967; Gallant, 1965; Ransom, 1968; Crawford, 1970) with correlation coefficients ranging from .25 to .95 (see Brown 1978, for a more detailed summary). These results indicate that cloze scores can provide reasonable estimates of reading comprehension ability, at least as measured by standardized reading comprehension tests.

In the second language literature, equally numerous studies indicate that cloze, if carefully developed, can prove to be a sound measure of overall English language proficiency (Conrad, 1970; Darnell, 1970; Oller & Inal, 1971; Oller, 1972a & b; Irvine et al, 1974; Stubbs & Tucker, 1974; Mullen, 1979; Alderson, 1979, 1980; Hinofotis, 1980; Brown, 1980, 1984, 1988b; Bachman, 1985; Revard, 1990) with coefficients ranging from .43 to .91. These studies indicate that cloze scores can provide a reasonable estimate of overall ESL/EFL proficiency (as measured by standardized proficiency tests).

Cloze and Readability

As mentioned above, the connection between cloze and readability was an issue when cloze procedure was first introduced by Taylor (1953). Other studies including Taylor (1957), Rankin (1965), Bormuth (1966, 1968), Miller and Coleman (1967), Bickley, Ellington, & Bickley (1970), Moyle (1970), and Ransom (1968) all indicated that cloze was to some degree related to readability.

However, other researchers have criticized the use of cloze procedures, especially as a criterion-measure in readability studies. As Carrell (1987) pointed out,

> ...cloze procedure can be, and often is, misused as a criterion. The most common abuse is to use only one of the *n* forms of a fixed-ratio, every *n*th deletion, format, to collect criterion data. Studies have shown that all *n* forms of and every *n*th fixed-ratio deletion cloze are seldom equal in difficulty.

Carrell's article does not make clear which studies have shown that different *n*th word deletion patterns seldom produce equal difficulties. Indeed, based on sampling theory, it would be reasonable to expect variations in difficulty such that the difficulties would only rarely be the same. The issue is not if they will differ but rather the degree to which they will differ beyond expectations within statistical sampling theory—an issue that, to my knowledge, has not been addressed in the literature.

Another critique, Carver (1977-1978), felt that cloze was not a good criterion measure for readability indices because it depended on the ability level of the particular group of students involved. As he put it:

> Superficially, it may appear that cloze would provide an acceptable estimate of material difficulty level (L_d). Yet the cloze measure has an inherent disadvantage which precludes its being used as standard for measuring language-knowledge difficulty of the material (L_d). Cloze is a rubber yardstick because the cloze difficulty estimate depends both upon the ability level of the particular group which was administered the cloze test, as well as the difficulty level of the material.

Carver's view is condemning the value of cloze to pinpoint actual grade level difficulty of passages. However, it ignores the benefits to be derived from basing readability estimates on human performance and, in fact, does not condemn the usefulness of cloze to estimate the relative difficulty of passages.

Kintsch and Vipond (1979) offer further criticism when they state that:

> The cloze procedure ... is probably actually misleading. It measures the statistical redundancy of a text, which is a far cry from its comprehensibility. By that score, a high-order statistical approximation of English that nevertheless constitutes incomprehensible gibberish would be preferred to a well-organized text with less predictable local patterns.

In fact, if a cloze passage were based on highly redundant "incomprehensible gibberish," as suggested by Kintsch and Vipond, it would be reasonable to expect students to score relatively poorly on it. Cziko (1978) provided evidence of this when he showed that, in French, students performed significantly better on a normal cloze passage than they did on one that had the sentences scrambled.

Furthermore, Kintsch and Vipond provide no support for their contention that the cloze procedure only measures statistical redundancy. Indeed, as noted at the top of this article, research indicates that cloze assesses general reading comprehension for native speakers and overall English language proficiency for ESL/EFL students. However, little indication exists in the literature on cloze that researchers have any more specific ideas on what cloze is measuring—redundancy or otherwise. The point is that, even if one accepts the notion that cloze is principally assessing the students' abilities to deal with redundancy, it can be argued (as I have elsewhere, see Brown, 1986) on the basis of the work of Goodman (1967) and Smith (1975; 1978) that the use of redundancy and prediction in taking a cloze test may be very similar to what goes on in the reading process.

Readability Indices

First language readability. Literally hundreds of readability indices have been created over the years. For overviews of the first language readability literature see Chall (1958), Klare (1963; 1984), or Zakaluk & Samuels (1988). For a review of the many uses to which readability indices have been put, see Fry (1987).

An entire literature discusses the effectiveness of these first language readability indices. However, one study (Brown, Chen, & Wang, 1984) was particularly influential in making me think that such readability indices might work. That study indicated a strong degree of relationship between the Fry readability estimates and grade levels as determined by native-speaker performance. In that study, the Fry scale for SRA kit cards was compared with the grade levels previously established by the author of the kits (based on the performance of North American elementary school children). Table 1 shows the results of this comparison.

Table 1
The Accuracy of First Language Readability Estimates Using the Fry Scale (Adapted From Brown, Wang, & Chen, 1984)

SRA KIT	PASSAGE GRADES ESTABLISHED BY STUDENT PERFORMANCE	FRY SCALE ESTIMATE MEAN	SD	LOW - HIGH
3A	3.5	3.22	1.20	2 - 6
	4.0	4.56	1.42	3 - 6
	4.5	5.56	0.88	4 - 7
	5.0	6.44	0.73	5 - 7
	6.0	7.11	0.93	6 - 8
	7.0	8.22	2.17	6 - 13
	8.0	8.67	1.50	6 - 10
	9.0	9.56	1.67	6 - 12
	10.0	10.22	1.48	7 - 12
	11.0	10.11	2.15	6 - 12
4A	8.0	8.56	1.13	6 - 10
	9.0	9.44	0.88	8 - 10
	10.0	10.44	1.74	9 - 14
	11.0	11.11	1.83	7 - 13
	12.0	12.56	1.51	11 - 16
	13.0	13.11	3.30	9 - 17+
	14.0	13.25	1.98	9 - 15

Notice, in Table 1, that the results are given for the 3A and 4A SRA kits as labeled to the left. Also note that the grade levels for each color within the kits are given in the

second column. Each color designates the cards in one grade (or half grade) level as established by the performance of native-speaker students on those cards. Each color contains 12 to 14 cards. The statistics for the Fry scale readability estimates for the cards in each color are given in the four columns to the right. Notice that the mean Fry index for each color/grade level is fairly close to the actual grade level of the cards as established by student performance. Clearly, a strong relationship can be seen between the mean grade levels as estimated using the Fry scale and the grade levels as established on the basis of students' performances.

However, note that the Fry scale estimates shown in Table 1 are averages across 12 to 14 cards in each color and that considerable variation exists in Fry readability indices among the cards within any given color/grade level as indicated by the standard deviation (SD), as well as by the low and high statistics given to the right of the table. Nonetheless, these results clearly indicate that a readability index like the Fry scale does have a striking relationship with the difficulty level of the materials for native speakers of English.

The first language readability indices of focus in this study are the Flesch reading ease formula (Flesch, 1948), the Flesch-Kincaid readability index (as described in Klare, 1984), the Fry readability index (see Fry, 1985), as well as the Gunning index, the Fog count, and a modified version of the Gunning-Fog readability index (see Larson, 1987).

Second language readability. In contrast to the vast amount of work that has been done on first language readability indices, very little has been done with regard to readability indices specifically designed for second language students. [For an excellent overview of readability issues directly related to ESL/EFL teaching, see Carrell (1987).]

I was able to find only a few studies wherein readability was investigated in languages other than English. A readability formula was developed for Vietnamese (Nguyen & Henkin, 1982). The Fry formula was applied to the readability of Spanish texts (Gilliam, Peña, & Mountain, 1980). In addition, Klare (1963. pp. 98-99, 272-274) surveys nine other early studies of readability indices for French, German, Japanese, and Spanish.

In the ESL field, Haskell (1973) found that cloze successfully differentiated passages regardless of variations in passage length, scoring method deletion rate, etc. Hamsik (1984) studied the relationships between four different readability indices and student performance on cloze tests developed from the passages found in the Miller-Coleman Readability Scale (Miller & Coleman, 1967). [It should be noted that Miller and Coleman had themselves ranked the passages on the basis of the cloze scores of 479 American college students.] She found that the readability formulas were appropriate for measuring ESL readability levels with rank order correlation coefficients ranging from .78 to .82 between the readability estimates and students' cloze performances.

However, on the whole, very little work has been done to establish any indices specifically tailored to second language learners' needs. Is such an index desirable? It seems to me that many occasions arise in which second language materials developers do need to sequence reading and other materials according to readability difficulty level just like first language materials developers do. Often when that need has come up in my work, like other ESL/EFL specialists, I have fallen back on the first language readability indices and made the assumption that they worked equally well in my setting because the texts that I was judging for readability were first language texts.

In reading Carrell (1987), however, I began to realize that the first language readability indices might *not* be appropriate for ESL/EFL settings. As she rightly pointed out, a number of factors are left out of the first language indices that might be crucial to judging the readability of texts for second language learners. For one thing, reader-based variables are totally ignored by such first language formulas. Consequently, differences in readability that might arise from differences in learners' characteristics (in terms of language differences, education, age, or learning style, for instance) are not taken into account.

Even in considering text-based factors alone, Carrell (1987) pointed out that first language indices typically include no measures of syntactic complexity (e.g., the T-Unit, Hunt, 1965), rhetorical organization, or "propositional density" (after Kintsch & Keenan, 1973). It occurred to me that additional factors might usefully be included in a second language readability index. From a lexical stand-point, several factors have seldom been considered in the first language readability indices: perhaps the type, function, and frequency of the words in a passage would be important factors in a second language index. For example, the type of vocabulary (e.g., the proportion of words of Latin origin as opposed to Germanic) might be an important consideration for ESL/EFL readers, particularly for students from Germanic or Latin language backgrounds, or even for students from other language groups. The frequency of the lexis within the passage itself (as redundancy), or the frequency of the lexis in the language might also be important factors in second language readability. What about the type of passage? For instance, could important differences exist in the readability of straight prose passages in contrast to dialogs, or other types of texts? What about extra-textual factors? Do accompanying illustrations, diagrams, and charts make a passage more readable for second language students? What about language specific factors like the number of words in the language of the students that are loan words from English?

Purpose of This Study

These and may other questions ultimately lead to the study that is being reported here. To answer such questions, I decided to focus on two central issues in doing this research. One purpose was to investigate the relationship between first language readability estimates and actual passage difficulties as established by EFL learners. In other words, the first purpose was to find out whether those indices were adequate for distinguishing EFL readability levels. A second purpose was to explore a wide range of textual, and extra-textual characteristics which might help to predict the relative difficulty that EFL students have with different passages. In the process, every effort was made to keep an open mind so that the data would guide me into discovering any existing patterns rather than the other way around. Nevertheless, the following exploratory and open-ended research questions were posed at the outset of this study:

1. Are randomly selected cloze tests reliable and valid tools for gathering data on the linguistic text variables that may be related to passage difficulty?
2. To what degree are traditional first language readability indices related to the average cloze scores for the same passages (when they are administered to EFL students)?
3. What combination of linguistic text variables best predicts passage difficulty for EFL students?
4. How can this combination of linguistic text variables be used as an EFL Difficulty Estimate?
5. How does the EFL Difficulty Estimate compare to other existing first language indices?

Since this research was exploratory in nature, the alpha level for all statistical decisions was set at a conservative $\alpha < .01$.

METHOD

Subjects

This study focused on the performance of 2298 Japanese university students who were all native speakers of Japanese. The subjects were selected as intact EFL classes from 18 different colleges and universities across Japan.[1] The subjects ranged in age from 18 to 24 and included 880 females and 1418 males. A total of fifty cloze procedures were administered such that all students were randomly assigned across all testing sessions to their particular cloze passages. This was done so that the results of the different groups

could reasonably be assumed to be equivalent across the fifty cloze procedures. An average of 45.96 students took each cloze, with a range of 42 to 50.

One problem with this study is that it focuses entirely on the performance of university students in Japan. Thus the results can only be generalized to Japanese university students. However, the fact that only one nationality was used can also be considered a strength of the study. In many studies in North America and other ESL settings, students with a variety of language backgrounds are mixed together. The results of such studies are difficult to interpret, at best, and cannot reasonably be generalized beyond the single institution in which the data were gathered. In addition, while the subjects in this study are not a random sample of all Japanese university students, the sample can at least be viewed as homogeneous with regard to the nationality, language background and educational level of the students.

Materials

The cloze procedures used here were based on texts which had been randomly selected from all the books in the adult reading section of the Leon County Public Library in Tallahassee, Florida. Fifty such books were randomly selected. Next, a page was randomly chosen from each book, and the actual passages were isolated by backing up to a logical starting point for a 400 to 450 word passage. Thus the passages were not 100 percent arbitrary. They were selected so that they would form sensible semantic units. Some passages were somewhat longer than 450 words because the stopping point was also determined by logical stopping points. In fact, the fifty passages ranged in length from 366 to 478 words with an average of 412.1 words per passage. The result was a set of fifty passages selected such that they can be assumed to represent the passages that would be encountered in the books found in a U.S. public library.

Once selected, every 12th word was deleted from each passage (for a total of thirty blanks) in order to create cloze procedures. The 12th word deletion pattern was used instead of the more traditional 7th word deletion pattern to make the items far enough apart so that performance on one item would minimally affect performance on other items. Generally, one sentence was left intact at the beginning of each passage and one or more sentences were unmodified at the end of each passage. Blanks were then created at the top of each passage for the students' name, sex, age, native language, and country of passport. Directions were supplied to explain what the students must do in filling in the blanks and how the blanks would be scored. The net result was a set of fifty cloze procedures (see Appendix A for example directions and 12 items taken from TEST A in the pilot study reported in Brown, 1989).

The reliability estimates for the cloze tests used in this study indicate that most of the cloze tests were reasonably reliable in the .70 to .80 range. However, the reliability estimates ranged considerably from one exceptionally low one of .172 to a high of .869 (for more details, see Brown, 1992 or 1993). The average of all fifty reliability estimates (using the Fisher *z* transformation) was .70. These reliability estimates are important in that the results of the study can be no more reliable than the measures upon which they are based.

A second very short ten-item cloze procedure was also created on the basis of the pretesting reported in Brown (1989). This cloze was modified using procedures similar to those described in Brown (1988b) so that only blanks that had proven very effective from an item analysis point of view were deleted. The purpose of this short cloze was to provide a common measure for making comparisons across the fifty groups of students.

The Importance of Randomization

Before moving to a description of the procedures used in this study, I would like to briefly discuss the importance of the notion of randomization in this study. The passages in this study were selected randomly from a public library and the blanks were selected on a semi-random basis (every 12th word). Based on sampling theory, the remainder of this study depends on the notion that the fifty, thirty-item cloze procedures constitute a collection of 50 texts which are representative of all of the texts in the Leon County Public Library. The representativeness of these passages appears to be supported by study of the lexical frequencies. The lexical frequencies of the 50 passages were counted and compared to the frequencies published for the "Brown" corpus (Kucera & Francis, 1967; Francis & Kucera, 1982) and (after being logrithmically transformed, see Carroll, 1967) were found to correlate at .93. Thus based on sampling theory and comparison of the lexical frequencies, I feel reasonably safe in assuming that these passages and blanks are representative samples of the English language, at least the English language written in the books found in an a U.S. public library.

In addition, the fifty groups of students were randomly assigned to the cloze passages. As such, it can be assumed that the groups were about equal in overall proficiency. Additional support for this assumption is found in Brown (1993), where one-way analysis of variance results for a the single 10-item cloze test that was administered across all 50 of these groups were not statistically significant ($F = 1.195$; $df = 49, 2248$; $p > .10$).

Procedures

The data for this study were gathered with the cooperation of a large number of Japanese, American, and British EFL teachers at 18 universities in various locations throughout Japan (see footnote [1]). The cloze procedures were photocopied and randomly distributed such that all students had an equal chance of getting any one of the 50 passages. They were administered by the teachers to their own students. The directions were read aloud and clarified as necessary. A total of 25 minutes was allowed for completing both the thirty-item and ten-item cloze procedures. According to feedback from the teachers, the 25 minute time limit proved sufficient.

The exact-answer scoring method was used throughout this study, which means that only the original word that had occupied the blank was counted as correct. This was justified because the results were not being reported to the students and because research indicates high correlations between exact-answer scoring results and other scoring procedures (Alderson, 1979 and Brown, 1980).

Analyses

The analyses in this study were all based on two kinds of variables: a dependent variable and a number of independent variables. The discussion in this section will first cover these two categories of variables, then briefly list the statistical analyses that were used in this study.

Dependent variable. EFL Difficulty, as a variable, was operationally defined as the mean scores on the cloze tests normalized by converting them to *z* values (relative to each other) then to percentiles. EFL Difficulty was the *dependent* variable in this study because it was the variable of primary interest in answering questions like the following: To what degree are the traditional first language readability indices related to EFL difficulty? and What linguistic variables can best be combined to predict EFL difficulty? In other words, EFL Difficulty was the dependent variable because it was measured "to determine what effect, if any, the other types of variables may have on it" (Brown, 1988a, p. 10).

Independent variables. The *independent* variables in this study were chosen because, in one way or another, they were factors which were potentially related to the EFL Difficulty dependent variable and because they were quantifiable in some way or other. In other words, the independent variables were selected because they might statistically explain, at least in part, the varying difficulty levels of the cloze passages in this study. Only ten independent variables have survived to be part of this report; these fall into two subcategories: (a) six first language readability indices and (b) four second language

linguistic predictor variables (that is, those four linguistic variables that proved to have meaningful, yet non-redundant relationships with the dependent variable).

The clearest way to explain the ***first language readability indices*** is to provide the formulas that define them. For instance, the formula for the Flesch reading ease index is as follows:

1) Flesch Reading Ease Formula (Flesch, 1948)

= 206.835 - .846(syllables/words) - 1.015(words/sentences)

This formula simply means that you must calculate the average number of syllables per word (syllables/words) and the average number of words per sentence (words/sentences). Next, multiply the average number of syllables per word by .846 and subtract the result from 206.835. From that result, subtract 1.015 times the average number of words per sentence. The other readability indices work in similar manner:

2) Flesch-Kincaid Index (as cited in Klare, 1984)

= .39(words/sentences) + 11.8(syllables/words) - 15.59

3) Fry Grade Level (Fry, 1977, or 1985)

= on the Fry reading graph, the grade value at the point where the coordinates for sentences per 100 words and syllables per 100 words cross

4) Gunning Index (as cited in Carrell, 1987)

= .4(words/sentences + % of words over two syllables)

5) Fog Count (as cited in Carrell, 1987)

$$= \frac{\left[\frac{\text{easy words} + 3(\text{hard words})}{\text{sentences}} - 3\right]}{2}$$

6) Gunning-Fog Index (Larson, 1987)

$$= \text{words/sentences} + 100 \times \left[\frac{\text{long words/sentences}}{\text{words/sentences}} \right]$$

A large number of ***second language linguistic predictor variables*** were also investigated in this study. Some of the simplest counts were the number of characters per word, syllables per word, syllables per sentence, words per sentence, syllables per paragraph, words per paragraph, and sentences per paragraph. Two measures of syntactic complexity were also included: words per T-unit (see Hunt, 1965; Gaies, 1980) and syllables per T-unit. Some lexical frequency variables were also added (as average frequencies): average frequency of the deleted words elsewhere in the cloze blanks, average frequency of the deleted words elsewhere in the passage in which they were found, average frequency of deleted words elsewhere in the 50 passages of this study, and average frequency of the deleted words in the Brown corpus (see Kucera & Francis, 1967; Francis & Kucera, 1982). Other lexical variables were calculated as percents: the percent of long words (seven or more letters), percent of function words, percent of Germanic root words. In addition, several learner-related variables were calculated as percents: percent of loan words to Japanese (based on Miura, 1979), and percent of Japanese Ministry of Education basic 500 words. Rhetorical organization was not studied here, but passage type was (i.e., whether the passage was straight prose or included a dialog). Finally, the presence or absence of illustrations (including pictures or diagrams) was an extra-textual variable that was considered.

It should be noted that many of the variables and readability indices in this study were quantified and calculated by using three software programs: Scandinavian PC Systems (1988), Que Software (1990), and PC-Style by Button (1986).

Note also that, out of all of the variables examined in this study, only a small subset survived. These variables were selected on the basis of factor and regression analysis as being orthogonal and most important in predicting EFL Difficulty. This does not mean that the other variables had no value, but rather that, in comparison to those variables that remained, they were relatively less important in predicting passage difficulty for Japanese university students. In other words, the relative importance of the above listed variables might have been quite different if the students had been older, or if they had been Spanish speakers, etc.

Of the three types of variables suggested by Carrell (1987), syntactic complexity (using T-units) and learner-related variables did not turn out to be very strongly related to EFL Difficulty in this particular study (as they were operationalized here). However, syllables per sentence and the percent of long words, which are both factors that show up in many of the traditional indices, did prove to be useful predictors of the relative difficulty of the passages for Japanese university students. In addition, two other factors related to the frequency and type of lexis were introduced in this study; these two variables, passage frequency and percent of function words, are not variables associated with traditional readability indices, but they did turn out to be useful in predicting the relative difficulty that students had with the 50 passages involved here.

To be specific, the subset of variables which survived to be included in the ensuing analyses are the following:

1. Syll/Sent - The average number of syllables found in the sentences in each passage.

2. Pass Freq - The average frequency with which the correct answers in the 30 blanks appeared elsewhere in the passage.

3. % Long Words - The percent of words that contained seven or more letters in the passages.

4. % Func Words - The percent of function words among the 30 deleted words in each passage. The remaining words were content words. Function words included articles, prepositions, conjunctions, and auxiliaries. Content words included nouns, pronouns, verbs, adjectives and adverbs.

Statistical analyses. The statistical analyses in this study included descriptive statistics for the 50 cloze tests and for the dependent and independent variables just described. At certain points Pearson product-moment correlations coefficients were used to investigate the degree of relationship between various pairs of the variables in this study. Factor analysis techniques, including principal components analysis and Varimax rotation, were used to investigate the degree to which variables were orthogonal (independent of each other). Finally, multiple regression analysis was used to investigate the degree to which combinations of the independent variables listed above could be used to predict the EFL Difficulty dependent variable.

RESULTS

The descriptive statistics for the 50 sets of cloze passages are given in Table 2, which describes the overall test characteristics for all 50 cloze tests in terms of the mean, standard deviation (SD), minimum score obtained (MIN), maximum score (MAX), the number of subjects who took the particular cloze (N), and the internal consistency reliability of the test (using the split-half method adjusted by the Spearman-Brown formula). In addition, the EFL Difficulty levels are reported in the column furthest to the right. Recall that these EFL Difficulty levels are simply the means converted to standardized percentiles (for passages relative to each other).

One salient result which surfaces in Table 2 is that the means of the fifty cloze tests range from 1.020 to 9.918. For reasons that are explained above, the groups can be assumed to be about equal in overall proficiency. Therefore, the variation among the means reported in Table 2 surely indicates considerable variation in the difficulty of the passages rather than differences in proficiency among the groups. Note that, for a test with 30 items, these means are fairly low. However, such low means are common for cloze tests which have been scored by the exact-answer method.

Notice also the wide range of standard deviations, from a low of 1.247 to a high of 4.435. Such a range of standard deviations suggests considerable variation in the degree to which the students' scores were dispersed around the mean on these cloze tests. The minimum (MIN) and maximum (MAX) indicate similar variations with the minimum ranging from 0 to 4 and the maximum ranging from 3 to 21. The number of subjects on each cloze passage also ranged from 42 to 50. The reliability of the 50 cloze tests likewise varied considerably. Notice that the lowest internal consistency reliability was .172, while the highest was .869. Finally, the EFL Difficulty levels show the difficulty of each of the passages relative to all other passages.

Table 3 focuses on the statistical characteristics of the first language readability indices examined in this study. Notice that, rather than being arranged by passage number as they were in the previous table, the passages are arranged here from the most difficult to the easiest as indicated by the EFL Difficulty in the second column. The remaining columns give the readability estimates for each passage using the Flesh, Flesch-Kincaid, Fry, Gunning, Fog, and Gunning-Fog indices. Notice that all of the indices except the Gunning-Fog index are on scales that resemble the grades in U.S. public schools. Notice also that, in some cases, they are fairly comparable across indices. In addition, note that the indices indicate similar relative difficulties for the passages. In other words, a passage

Table 2

Descriptive Statistics for 50 Cloze Passages

PASSAGE	MEAN	STD	MIN	MAX	N	RELIABILITY	EFL DIFF.
01	5.229	3.164	0	15	48	0.708	30.15
02	4.208	3.421	0	13	47	0.858	48.40
03	2.021	2.126	0	10	48	0.735	83.40
04	7.543	3.866	2	16	46	0.803	5.59
05	3.979	2.787	0	13	47	0.734	52.39
06	5.106	3.230	0	14	47	0.803	32.28
07	6.140	3.407	0	16	43	0.825	17.36
08	3.156	2.270	0	8	45	0.457	67.00
09	2.848	2.458	0	11	46	0.773	71.90
10	2.543	2.310	0	8	46	0.825	76.42
11	5.935	3.358	0	16	46	0.742	20.05
12	8.980	3.967	0	21	47	0.789	1.22
13	2.870	1.714	0	8	46	0.503	71.57
14	3.234	2.503	0	9	47	0.682	65.91
15	9.180	3.416	4	18	49	0.683	0.96
16	1.360	1.411	0	6	48	0.650	89.80
17	1.383	1.247	0	5	46	0.348	89.62
18	1.020	1.086	0	3	50	0.500	92.36
19	4.760	2.881	0	10	50	0.701	38.21
20	4.375	3.238	0	15	47	0.855	45.22
21	9.918	4.435	0	19	48	0.840	0.37
22	3.702	2.858	0	11	47	0.841	57.53
23	3.638	2.401	0	11	43	0.646	58.71
24	2.957	2.259	0	9	47	0.436	70.19
25	5.362	2.740	0	12	46	0.627	28.10
26	2.681	1.559	0	5	47	0.172	74.54
27	2.340	2.723	0	13	47	0.869	79.39
28	2.581	2.170	0	8	43	0.574	76.11
29	2.318	1.768	0	7	44	0.640	79.67
30	9.563	3.284	3	16	48	0.715	0.59
31	3.783	3.078	0	15	46	0.832	55.96
32	3.833	2.525	0	9	42	0.770	55.17
33	2.136	1.866	0	6	44	0.633	81.86
34	5.867	2.918	0	13	45	0.819	20.90
35	6.630	3.662	0	17	45	0.719	12.30
36	5.000	2.054	0	9	46	0.505	34.09
37	5.458	3.657	0	13	48	0.767	26.76
38	1.708	1.567	0	8	48	0.746	86.65
39	2.511	1.977	0	9	47	0.648	77.04
40	3.488	1.897	0	9	43	0.659	61.41
41	2.870	2.507	0	10	43	0.764	71.57
42	4.409	3.099	0	18	44	0.811	44.43
43	1.432	1.452	0	7	44	0.190	89.25
44	3.239	2.521	0	10	46	0.673	65.54
45	6.548	3.874	0	16	42	0.788	12.92
46	2.163	1.816	0	7	47	0.307	81.59
47	3.791	2.328	0	11	43	0.685	55.96
48	2.690	2.121	0	11	42	0.738	74.54
49	4.564	2.808	0	11	49	0.748	41.48
50	2.488	2.697	0	12	45	0.774	77.34

Table 3
First Language Readability Estimates for 50 Passages

PASSAGE	EFL DIFF.	FLESCH	FLESCH-KINCAID	FRY	GUNNING	FOG	GUNNING-FOG
18	92.36	9.69	12.7	12	6.06	9.70	40
16	89.80	8.90	13.0	9	8.99	15.72	50
17	89.62	15.60	20.4	14	9.78	18.83	58
43	89.25	11.51	13.9	10	9.72	15.16	43
38	86.65	11.01	12.9	11	8.13	12.95	42
03	83.40	2.83	4.8	3	3.25	3.51	21
33	81.86	13.82	16.3	12	11.01	20.82	59
46	81.59	8.78	11.2	9	5.80	8.50	34
29	79.67	13.58	16.0	11	11.00	17.25	46
27	79.39	9.36	10.0	9	7.20	11.02	38
50	77.34	18.51	21.3	15	13.48	25.35	64
39	77.04	5.09	6.7	6	5.81	7.54	27
10	76.42	11.86	15.2	10	9.61	15.76	46
28	76.11	12.00	14.4	14	8.23	14.48	49
26	74.54	13.95	16.6	14	9.05	16.67	54
48	74.54	8.51	11.2	8	6.95	11.59	44
09	71.90	12.30	15.3	12	9.34	16.11	49
41	71.57	12.26	14.3	12	9.33	15.47	47
13	71.57	10.65	12.1	10	8.83	13.53	40
24	70.19	10.69	13.1	10	8.95	13.61	40
08	67.00	8.46	11.2	8	7.83	11.43	36
14	65.91	4.79	8.5	6	4.26	5.45	27
44	65.54	11.60	13.9	11	7.81	12.69	43
40	61.41	5.69	8.1	6	5.47	7.53	30
23	58.71	11.45	13.9	13	7.35	12.54	46
22	57.53	8.97	10.8	9	7.16	10.67	37
47	55.96	9.99	11.9	9	8.24	12.70	40
31	55.96	8.13	11.6	10	5.26	8.09	37
32	55.17	7.80	9.6	8	5.94	8.13	30
05	52.39	11.00	13.9	10	6.57	10.38	40
02	48.40	10.71	13.5	13	6.07	10.04	42
20	45.22	8.30	10.8	8	7.03	10.24	35
42	44.43	7.10	9.1	8	5.19	7.17	31
49	41.48	7.59	10.3	7	8.19	12.00	37
19	38.21	8.27	10.2	8	6.40	9.42	35
36	34.09	7.88	11.3	8	5.82	9.37	40
06	32.28	5.18	7.0	6	4.11	5.22	27
01	30.15	6.78	9.6	7	6.15	8.65	32
25	28.10	7.72	10.2	7	7.09	9.69	31
37	26.76	6.03	8.6	2	6.81	9.26	31
34	20.90	10.69	12.8	10	8.48	13.38	42
11	20.05	2.71	5.0	3	3.05	3.22	20
07	17.36	9.37	9.9	10	6.07	10.08	43
45	12.92	8.47	11.1	8	6.72	10.06	36
35	12.30	3.69	4.8	4	4.09	4.79	22
04	05.59	5.95	7.6	6	6.41	8.41	28
12	01.22	8.59	11.0	10	5.67	8.09	32
15	00.96	9.69	12.0	10	6.41	9.98	38
30	00.59	4.63	6.5	5	5.08	6.03	22
21	00.37	4.74	7.5	5	4.85	5.91	24

that appears to be relatively easy on one index is also relatively easy on the other ones, while a passage that appears to be relatively difficult on one index is also relatively difficult on the others.

Table 4 shows the simple correlation coefficients above the diagonal and coefficients of determination below the diagonal for all possible pairs of the first language readability estimates used in this study. The coefficients of determination are calculated by squaring the correlation coefficient, and they indicate the percent of overlapping variance between the two variables involved. Thus the correlation coefficient of .48 shown above the diagonal in Table 4 between the Fry index and Observed EFL Difficulty can be interpreted as indicating that 23 percent ($.48^2 \times 100 = .2304 \times 100 = 23.04$, or about 23 percent) of the variance in EFL Difficulty is accounted for by the Fry index. These squared values are shown below the diagonal.

Notice that the coefficients of determination within the triangle are mostly fairly high with the lowest being .49 and the highest being .96. These relatively high coefficients indicate that the first language readability indices (variables 1 through 6 in the table) are all fairly highly related to each other. In other words, they are lining up the relative difficulty of the passages in a very similar ways.

The long thin rectangle (on the left side) outlines those coefficients of determination which show the percent of relationship between the various first language readability estimates and the observed performance of Japanese students on the cloze passages, as represented by the observed EFL Difficulty percentiles (variable A). It turns out that the first language indices overlap between 23 and 30 percent (depending on which one is examined) with the variance in observed EFL Difficulties. In short, these first language readability indices account for less than thirty percent of the variance in the observed EFL Difficulty levels.

A large number of linguistic variables were also examined for relationship to EFL Difficulty. Four of these variables were selected on the basis of factor analysis as being orthogonal: syllables per sentence, average frequency elsewhere in the passage of the words that had been deleted, the percent of long words of seven letters or more, and the percent of function words. When combined, they proved to be the best predictors of observed EFL Difficulty. The descriptive statistics for these four independent (predictor) variables and the dependent (predicted) variable, EFL Difficulty, are shown in Table 5.

Table 4
Correlation Coefficients (Above the Diagonal) and Correlation Coefficients for First Language Readability Indices and EFL Difficulty

A. Observed EFL Difficulty	1.00	0.74	0.50	0.52	0.48	0.50	0.54	0.55
B. ESL Diff. Estimate	0.55	1.00	0.70	0.72	0.66	0.66	0.70	0.73
1. Flesch	0.25	0.49	1.00	0.98	0.92	0.89	0.93	0.95
2. Flesch-Kincaid	0.27	0.52	0.96	1.00	0.90	0.87	0.92	0.95
3. Fry	0.23	0.44	0.85	0.81	1.00	0.70	0.78	0.88
4. Gunning	0.25	0.44	0.79	0.76	0.49	1.00	0.98	0.87
5. Fog	0.29	0.49	0.86	0.85	0.61	0.96	1.00	0.95
6. Gunning-Fog	0.30	0.53	0.90	0.90	0.77	0.76	0.90	1.00
	A	B	1	2	3	4	5	6

Table 5
Descriptive Statistics for the Predicted and Predictor Variables

VARIABLE	MEAN	SD	MINIMUM	MAXIMUM
PREDICTED				
EFL Difficulty	53.02	28.12	.37	92.36
PREDICTOR				
Syll/Sent	36.95	12.62	15.57	76.63
Pass Freq	6.96	0.59	5.66	8.82
% Long Words	20.52	5.94	9.89	34.33
% Func Words	31.55	8.17	13.33	50.00

The degree to which the independent variables listed in the previous paragraph were collectively related to EFL Difficulty was investigated using multiple-regression analysis. The assumptions underlying multiple regression were checked and found to be met.[3] A forward-stepping multiple-regression analysis was calculated for the four variables regressed against EFL Difficulty. The results of this regression analysis are presented in Tables 6 and 7. Table 6 shows the technical results of the regression analysis including the progressive additivity of the multiple correlation (MR) and multiple coefficient of determination (MR^2). Note that the overall Analysis of Variance

Table 6

Stepwise Regression Analysis of Four Independent Variables Predicting the EFL Difficulty Dependent Variable

PROB VALUE TO ADD/REMOVE: 0.1000

DEPENDENT VARIABLE: EFL Difficulty
STEP 1 MR=.5506 MR^2=0.3032 ADDED Syll/Sent
STEP 2 MR=.6699 MR^2=0.4487 ADDED Pass Freq
STEP 3 MR=.7168 MR^2=0.5138 ADDED % Long Words
STEP 4 MR=.7418 MR^2=0.5502 ADDED % Func Words

VARIABLE	REGRESSION COEFFICIENT	STANDARDIZED COEFFICIENT	STANDARD ERROR	T	PROB
Syll/Sent	0.7823	0.351189	0.2793	2.8014	0.0075
Pass Freq	-126.1770	-0.520334	27.3129	-4.6197	0.0000
% Long Words	1.2878	0.272007	0.6117	2.1051	0.0409
% Func Words	0.7596	0.220810	0.3982	1.9076	0.0628

ESTIMATED CONSTANT TERM: 38.7469
STANDARD ERR OF ESTIMATE: 19.6800

TABLE 7: SUMMARY OF THE VARIABLES CONTRIBUTING TO THE STEPWISE REGRESSION COEFFICIENTS

DEPENDENT = VARIABLE	INDEPENDENT VARIABLES	MR	MR^2
Pass Diff =	Syll/Sent	.55	.30
Pass Diff =	Syll/Sent + Pass Freq	.67	.45
Pass Diff =	Syll/Sent + Pass Freq + % Long Words	.72	.51
Pass Diff =	Syll/Sent + Pass Freq + % Long Words + % Func Words	.74	.55

results were deleted for economy of space, but $F = 13.7618$, df 4, 45, $p < .00001$. For each independent variable, Table 6 also gives the regression coefficients, standardized coefficients, individual standard errors, t value at entry, and the probability associated with t. Finally, the constant, and an overall standard error of estimate for the predicted values of EFL Difficulty are given in the lower-left corner. Table 7 illustrates the progressive additivity of the variables and the associated multiple correlations (MR) and the multiple coefficients of determination (MR^2).

These results indicate that the combination of Syll/Sent + Pass Freq + % Long Words + % Func Words taken together produce a multiple-correlation (MR) of .74 and a corresponding MR^2 of .55. This means that the combination of simple countable independent variables taken together predicts about 55 percent of the variance in the performance of Japanese students on the 50 cloze passages in this study. In other words, the results here indicate that each of the independent variables separately is related to EFL Difficulty and that, taken together, they account for 55 percent of the variance in EFL Difficulty.

DISCUSSION

The discussion will now return to the original five research questions. The implications of these findings for second language readability estimation will then be covered in the **CONCLUSIONS** section.

1. Are Randomly Selected Cloze Tests Reliable and Valid Tools for Gathering Data on the Linguistic Text Variables That May Be Related to Passage Difficulty?

Based on Table 2, the cloze passages used in this study appear on average to be moderately reliable at .70 using the adjusted Split-half method, but also, individual tests can clearly vary considerably in reliability from .172 to .869. To some degree, such variation in reliability appears to be related to the magnitude of the means and standard deviations involved. However, all of these variations in descriptive statistics and reliability could conceivably have occurred by chance alone.

For the purposes of this study, the validity of the fifty cloze passages will be considered from a fairly common-sense point of view. First, the cloze passages were created from books which were randomly selected from a public library, and the items for each passage were selected semi-randomly (i.e., every 12th word deletion). Based on sampling theory, the passages can be said to be a representative sample of the language found in the books in the library from which they were taken, and the items can be said to provide a representative sample of the blanks that could be created in the language contained in the passages. Since the validity of a test can be defined as the degree to which it is measuring what it purports to be measuring, it seems reasonable to claim a high degree of content validity for these cloze passage items because they can be said to be representative samples of the universe of all possible items (after Cronbach, 1970) if that universe is defined as single-word blanks created in the written language which is found in a U.S.

public library. [For much more discussion of the reliability and validity of these passages, see Brown (1993); for an overview of test reliability and validity issues, see Brown, 1996.]

2. To What Degree Are Traditional First Language Readability Indices Related to the Average Cloze Scores for the Same Passages (When They Are Administered to EFL Students)?

Tables 3 and 4 both indicate that some degree of relationship exists between each of the first language readability indices and EFL Difficulty. More specifically, the first language readability indices used in this study are related to EFL Difficulty at between 23 and 30 percent—at least as EFL Difficulty is measured by the performance of Japanese university students on the cloze passages. The first language readability indices also appear to be highly interrelated with each other, producing coefficients of determination of .49 to .96, which indicate 49 to 96 percent overlapping variance.

Aside from the fact that first language readability indices are not very highly related to the EFL Difficulty, another problem with these first language readability indices is that they use grade levels (in American schools) as their yardstick. Such grade levels do not make sense for second language students. Grades are different from country to country. Even within the United States, the meaning of reading levels at different grades may have changed in recent years with fewer and fewer students reading at or above their own grade level. Instead, any EFL Difficulty Estimate should probably be referenced to a specific population in percentile terms. Such estimates will therefore be population specific, and that is perhaps as it should be.

3. What Combination of Linguistic Text Variables Best Predicts Passage Difficulty for EFL Students?

The variables that best predicted EFL Difficulty, at least for the population of Japanese university students, were Syllables/Sentence and Passage Frequency and % Long Words and % Function Words (see Table 6 or 7). This combination of independent variables produced a multiple correlation of .75 with the dependent variable. Its squared value, the multiple coefficient of determination, indicated that the four variables taken together account for about 55 percent of the variance in EFL Difficulty. Of course, such results must be interpreted very cautiously. For instance, these results do not necessarily mean that these same variables in the same order will be found to be the best predictors in a replication of this study. In addition, many of the other variables examined in this study might have been used in this formula. The fact that these particular variables were chosen was based on a factor analysis, which indicated that four orthogonal factors existed in the

correlation matrix of dozens of independent variables. For those four factors, the variables listed above were the ones most strongly correlated with the factor.

4. How Can This Combination of Linguistic Text Variables Be Used As an EFL Difficulty Estimate?

Like many of the first language readability indices, the EFL Difficulty Estimate can be calculated by using a regression equation. In this case, the regression equation for predicting a single dependent variable (Y) takes the form of a constant (a) and four independent variables (X_1 to X_4) with their associated slopes (b_1 to b_4). Such an equation would take the following general form:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$$

In more familiar terms, the regression equation for predicting, or estimating, the single dependent variable (EFL Difficulty) is formed by using the constant (shown to be 38.7469 shown in Table 6), as well as the four slopes (called regression coefficients in Table 6) and the values for each of the four independent variables (Syllables/Sentence, Passage Frequency, % Long Words, % Function Words). The equation in this case would take the following form:

EFL Difficulty Estimate = 38.7469 + (.7823 x Syll/Sent)
+ (-126.1770 x Pass Freq)
+ (1.2878 x % Long Words)
+ (.7596 x % Func Words)

For instance, the equation for the EFL Difficulty Estimate for Passage 43 (where Syll/Sent = 76.63; Pass Freq = .41; % Long Words = 19.22; and % Func Words = 23.33) would be as follows:

EFL Difficulty Estimate = 38.7469 + (.7823 x 76.63)
+ (-126.1770 x .41)
+ (1.2878 x 19.22)
+ (.7596 x 23.33)

EFL Difficulty Estimate = 38.7469 + (59.9476)
+ (-51.7326)
+ (23.7515)
+ (17.7215)

EFL Difficulty Estimate = 89.4349 ≅ 89.43

Obviously such an EFL Difficulty Estimate is not easy to calculate by hand. The counts that are necessary and the computations are not only laborious, but also very prone to calculation errors if done by hand. However, computer software could no doubt be

developed to do the job quickly and efficiently. Examples of similar software include Scandinavian PC Systems (1988), Que Software (1990) and PC-Style by Button (1986). All three of these software packages produce first language readability indices, and no doubt, a similar software package could easily be programmed to count the necessary linguistic elements and calculate an EFL Difficulty Estimate like the one shown here.

5. How Does the EFL Difficulty Estimate Compare to Other Existing First Language Indices?

The most variance in EFL Difficulty that was accounted for by any of the first language readability indices was 30 percent. In other words, these first language readability indices accounted for less than three tenths of the variance in observed EFL Difficulties. The ESL Difficulty Estimates, on the other hand, are correlated with the EFL Difficulties at .74, which indicates that 55 percent ($.74^2 = .5476 \times 100 \cong 55$) of the variance in Passage Difficulties was accounted for. In other words, the EFL Difficulty Estimates accounted for more than half of the variance in Passage Difficulties. Another way to look at this issue is that the EFL Difficulty Estimates accounted for nearly twice as much variance in Passage Difficulties as did the first language readability indices. In short, the EFL Difficulty Estimate is much more strongly related to Passage Difficulty than any of the first language readability indices.

However, the EFL Difficulty Estimate is not without its own problems. It is still only a moderately good predictor as indicated by the multiple coefficient of determination. Another way to think about the accuracy of predictions offered by the EFL Difficulty Estimate is to consider the standard error of estimate, which is shown to be 19.68 at the bottom of Table 6. This statistic indicates a confidence interval around the predicted values within which the estimates can be expected to fall 68 percent of the time. In practical terms, this means that the estimates can be expected to be inaccurate by as much as 19.68 points 68 percent of the time. In even more concrete terms, the results displayed in Table 8 will illustrate the accuracy of the EFL Difficulty Estimates in predicting the Actual EFL Difficulties. Table 8 presents the Actual EFL Difficulties and EFL Difficulty Estimates side-by-side for easy comparison. In addition, the differences between actual and estimated difficulties are shown in the column furthest to the right. Even though the EFL Difficulty Estimate is fairly highly related to the Actual EFL Difficulty, it is far from precise even when it is applied to the data from which it was derived.

Table 8
Actual and Predicted EFL Difficulty Values

CLOZE PASSAGE	PASSAGE MEAN	ACTUAL EFL DIFF.	EFL DIFF. ESTIM.	DIFFERENCE
18	1.02	92.36	78.75	13.61
16	1.36	89.80	81.00	8.80
17	1.38	89.62	78.89	10.73
43	1.43	89.25	89.43	-0.18
38	1.71	86.65	55.34	31.31
03	2.02	83.40	40.44	42.96
33	2.14	81.86	87.96	-6.10
46	2.16	81.59	41.51	40.08
29	2.32	79.67	64.18	15.49
27	2.34	79.39	50.65	28.74
50	2.49	77.34	92.05	-14.71
39	2.51	77.04	41.32	35.72
10	2.54	76.42	78.86	-2.44
28	2.58	76.11	72.22	3.89
26	2.68	74.54	57.06	17.48
48	2.69	74.54	55.65	18.89
9	2.85	71.90	65.78	6.12
41	2.87	71.57	68.42	3.15
13	2.87	71.57	63.15	8.42
24	2.96	70.19	54.34	15.85
8	3.16	67.00	58.58	8.42
14	3.23	65.91	37.66	28.25
44	3.24	65.54	76.69	-11.15
40	3.49	61.41	60.93	0.48
23	3.64	58.71	75.71	-17.00
22	3.70	57.53	60.55	-3.02
47	3.79	55.96	73.27	-17.31
31	3.78	55.96	55.46	0.50
32	3.83	55.17	50.00	5.17
5	3.98	52.39	61.46	-9.07
2	4.21	48.40	51.17	-2.77
20	4.38	45.22	45.59	-0.37
42	4.41	44.43	56.29	-11.86
49	4.56	41.48	31.92	9.56
19	4.76	38.21	49.96	-11.75
36	5.00	34.09	64.78	-30.69
6	5.11	32.28	37.95	-5.67
1	5.23	30.15	50.00	-19.85
25	5.36	28.10	36.03	-7.93
37	5.46	26.76	22.99	3.77
34	5.87	20.90	52.68	-31.78
11	5.93	20.05	42.19	-22.14
7	6.14	17.36	8.61	8.75
45	6.55	12.92	32.99	-20.07
35	6.63	12.30	29.65	-17.35
4	7.54	5.59	31.88	-26.29
12	8.98	1.22	32.75	-31.53
15	9.18	0.96	38.61	-37.65
30	9.56	0.59	-0.81	1.40
21	9.92	0.37	8.62	-8.25

CONCLUSIONS

In general terms, the results of this study indicate that a variety of first language readability indices for a set of 50 passages were only weakly correlated with the average performances of Japanese university students on cloze versions of those same passages. In other words, the first language indices were only weakly related to EFL Difficulty (no more than 30 percent related). The EFL Difficulty Estimate provided in this paper had a higher degree of association (about 55 percent related). Unfortunately, the EFL Difficulty Estimate is not easy to calculate. However, it does account for more of the variance in EFL Difficulty than the traditional first language readability formulas. Perhaps ESL/EFL readability formulas will necessarily be more complex. Perhaps, other higher order linguistic and student variables like those used in this study will be needed to account for this additional variance.

In addition, because of the controversy surrounding the cloze procedure as a criterion measure for readability indices, it might be better to think of the EFL Difficulty Estimate developed in this project as a sort of clozability index, or indication of the degree of proficiency needed to successfully fill in blanks in a cloze format. Surely, some association exists between the EFL Difficulty Estimate provided here and some aspect of the relative difficulty of the cloze passages used. Since cloze passages are well-established measures of overall ESL/EFL proficiency, the EFL Difficulty Estimate might best be viewed as a measure of the overall difficulty of passages with respect to the ESL/EFL proficiency needed to comprehend them.

The primary point is not that this particular index is the magical answer to determining the readability of passages for use in ESL/EFL curricula and materials, but rather that such an index can be created, one that is more highly related to the performance of second language learners than are the first language readability indices. A second point is that such an index may necessarily include some reference to lexical variables, in this case the average percent of long words (seven or more letters), the average percent of function words, the average frequency of the word elsewhere in the passage. A third point is that EFL/ESL readability might best be estimated separately for students from different language backgrounds. Perhaps different variables in different combinations with different weightings will work better or worse in predicting the readability of passages for speakers of different languages.

Thus a strategy similar to the one employed in this study could be used to constantly improve the readability estimates for speakers of different languages as we learn more and more about what makes text difficult for students to process.

Suggestions for Future Research

As is often the case in research of this sort, more questions were raised in the process of doing this study than were answered. The following research questions are provided in the hope that other researchers will pursue this line of inquiry:

1. What differences and similarities would occur if this study were replicated at other institutions in Japan? With students from other language groups? With students at other levels of study? Or other ages?
2. What other linguistic text or extra-textual variables might be included in such research? How well would they predict EFL Difficulty?
3. What hierarchies of difficulty are found at the passage level for any of the linguistic variables (separately or combined) that would have implications for second language acquisition research?

NOTES

[1] I would like to thank all of those colleagues who helped at various stages of this project by administering the cloze procedures at Dokkyo University, Fukuoka Teacher's College, Fukuoka University of Education, Fukuoka Women's University, International Christian University, International University of Japan, Kanazawa University, Kansei Gakuin University, Meiji University, Saga University, Seinan Gakuin University, Soai University, Sophia University, Tokyo University of Agriculture and Technology, Toyama University, Toyama College of Foreign Languages, Toyo Women's Junior College, and Waseda University. I would also like to thank Dr. Ian Richardson (currently a professor at King Saud University in Abha, Kingdom of Saudi Arabia) for his help in selecting and creating the cloze procedures used here. I must also thank Dr. Thom Hudson for his careful readings and comments on an earlier version of this paper.

[2] Note that the dependent variable, Passage Difficulty was normalized by transforming it to a percentile scale (using the areas under the curve in the *z* distribution). The Passage Frequency variable was transformed in all analyses using a standard log transformation (see Chatterjee & Price, 1977, pp. 27-38, or Neter & Wasserman, 1974, pp. 121-130). This was necessary to correct for a curvelinear relationship with the dependent variable. Further justification for these transformations is based on Carroll (1967), who found that word-frequency counts are lognormally distributed.

[3] One concern whenever performing regression analysis is that the rather rigorous assumptions and design conditions be met. One of these assumptions is that the dependent and independent variables must be normally distributed. In order to achieve normality and linearity two of the variables were transformed as pointed out in footnote [2]. Table 5 indicates that, as analyzed, all of the variables in the regression analysis were reasonably normal in distribution. In addition, the relationships of each of the independent variables was found to be linear with Passage Difficulty (the dependent variable). Multicollinerarity was avoided by using factor analysis in the selection process with the goal of maximizing the orthogonality of the dependent variables. The assumption of heteroscedasticity was checked by examining the scatterplots of each variable with residuals; it was not found to be a problem. In addition, the Durbin-Watson statistic turned out to be 1.4 indicating that autocorrelation was not an issue (Chatterjee & Price, 1977, p. 127). However, one final problem is more worrisome. The units of analysis, cloze passages, were only 50 in number. Thus the *N*-size for the regression was only 50,

and, with four dependent variables, this may not be large enough. No hard and fast rule exists about this matter, yet this is a problem that readers should keep in mind while interpreting the results of the present study.

APPENDIX A: EXAMPLE CLOZE PASSAGE (FROM BROWN 1989)

Name_____ Native Language_____
(Last) (First)

Sex_____ Age_____ Country of Passport_____

DIRECTIONS:

1. Read the passage quickly to get the general meaning.
2. Write *only one* word in each blank. Contractions (example: *don't*) and possessives (*John's* bicycle) are one word.
3. Check your answers.

NOTE: Spelling will *not* count against you as long as the scorer can read the word.

EXAMPLE: The boy walked up the street. He stepped on a piece of ice. He fell (1)_____, but he didn't hurt himself.

A FATHER AND SON

Michael Beal was just out of the service. His father had helped him get his job at Western. The (1)_____ few weeks Mike and his father had lunch together almost every (2)_____ . Mike talked a lot about his father. He was worried about (3)_____ hard he was working, holding down two jobs.

"You know," Mike (4)_____ , "before I went in the service my father could do just (5)_____ anything. But he's really kind of tired these days. Working two (6)_____ takes a lot out of him. He doesn't have as much (7)_____ . I tell him that he should stop the second job, but (8)_____ won't listen.

During a smoking break, Mike introduced me to his (9)_____ . Bill mentioned that he had four children. I casually remarked that (10)_____ hoped the others were better than Mike. He took my joking (11)_____ and, putting his arm on Mike's shoulder, he said, "I'll be (12)_____ if they turn out as well as Mike." (continues ...)

REFERENCES

Alderson, J. C. (1978). *A study of the cloze procedure with native and non-native speakers of English*. Unpublished doctoral dissertation, University of Edinburgh.

Alderson, J. C. (1979). Scoring procedures for use on cloze tests. In C. A. Yorio, K. Perkins, & J. Schachter (Eds.) *On TESOL '79* (pp. 193-205). Washington, D.C.: TESOL.

Alderson, J. C. (1980). Native and non-native speaker performance on cloze tests. *Language Learning, 30*, 59-76.

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly, 19*, 535-555.

Bickley, A. C., Ellington, B. J., & Bickley, R. T. (1970). The cloze procedure: A conspectus. *Journal of Reading Behavior, 2*, 232-234.

Bormuth, J. R. (1965). Validities of grammatical and semantic classifications of cloze test scores. In J. A. Figurel (Ed.), *Reading and inquiry* (pp. 283-285). Newark, Delaware: International Reading Associates.

Bormuth, J. R. (1966). Readability: A new approach. *Reading Research Quarterly, 1*, 79-132.

Bormuth, J. R. (1967). Comparable cloze and multiple-choice comprehension tests scores. *Journal of Reading, 10*, 291-299.

Bormuth, J. R. (1968). The cloze readability procedure. *Elementary English*, 36, 429-436.

Brown, J. D. (1978). *Correlational study of four methods for scoring cloze tests*. Unpublished MA thesis, University of California, Los Angeles.

Brown, J. D. (1980). Relative merits of four methods for scoring cloze tests. *Modern Language Journal, 64*, 311-317.

Brown, J. D. (1983a). A closer look at cloze: Part I - validity. In J. W. Oller Jr. (Ed.) *Issues in Language Testing* (pp. 237-243). Rowley, MA: Newbury House.

Brown, J. D. (1983b). A closer look at cloze: Part II - reliability. In J. W. Oller Jr. (Ed.) *Issues in Language Testing* (pp. 243-250). Rowley, MA: Newbury House.

Brown, J. D. (1984). A cloze is a cloze is a cloze? In J. Handscombe, R. A. Orem, & B. P. Taylor (Eds.), *On TESOL '83* (pp. 109-119). Washington, DC: TESOL.

Brown, J. D. (1986). Cloze procedure: A tool for teaching reading. *TESOL Newsletter, 20*(5), 1 & 7.

Brown, J. D. (1988a). *Understanding research in second language learning: A teacher's guide to statistics and research design*. London: Cambridge University Press.

Brown, J. D. (1988b). Tailored cloze: Improved with classical item analysis techniques. *Language Testing, 5*, 19-31.

Brown, J. D. (1989). Cloze item difficulty. *JALT Journal, 11*, 46-67.

Brown, J. D. (1992). What text characteristics predict human performance on cloze test items. In the *Proceedings of the Third Conference on Language Research in Japan* (pp. 1-26). Urasa, Japan: International University Japan.

Brown, J. D. (1993). What are the characteristics of "natural" cloze tests? *Language Testing, 10*(2), 93-116.

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

Brown, J. D., Chen, Y., & Wang, Y. (1984). An evaluation of native-speaker self-access reading materials in an EFL setting. *RELC Journal, 15*, 75-84.

Button, J. (1986). *PC-Style: The program that has a way with words. Grand Prairie*. TX: Lone Star Software.

Carrell, P. L. (1987). Readability in ESL. *Reading in a Foreign Language, 4*, 21-40.

Carroll, J. B. (1967). On sampling from a lognormal model of word-frequency distribution. In H. Kucera & W.N. Francis *Computational analysis of present-day English* (pp. 406-413). Providence, RI: Brown University.

Carver, R. P. (1977-1978). Toward a theory of reading comprehension and reading. *Reading Research Quarterly, 13*, 8-63.

Chall, J. S. (1958). *Readability: An appraisal of research and application*. Columbus, Ohio: Ohio State University.

Chatterjee, S., & Price, B. (1977). *Regression analysis by example*. New York: John Wiley & Sons.

Chavez-Oller, M. A., Chihara, T., Weaver, K. A., & Oller, J. W. Jr. (1985). When are cloze items sensitive to constraints across sentences? *Language Learning, 35*, 181-206.

Chihara, T., Oller, J. W. Jr., Weaver, K. A., & Chavez-Oller, M. A. (1977). Are cloze items sensitive to constraints across sentences? *Language Learning, 27*, 63-73.

Conrad, C. (1970). *The cloze procedure as a measure of English proficiency*. Unpublished master's thesis, University of California Los Angeles.

Crawford, A. (1970). *The cloze procedure as a measure of reading omprehension of elementary level Mexican-American and Anglo-American children*. Unpublished doctoral dissertation, University of California Los Angeles.

Cronbach, L. J. (1970). *Essentials of psychological testing* (pp. 145-146). New York: Harper & Row.

Darnell, D. K. (1970). Clozentropy: A procedure for testing English language proficiency of foreign students. *Speech Monographs, 37*, 36-46.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*, 221-233.

Francis, W. N., & Kucera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston, MA: Houghton Mifflin.

Fry, E. B. (1968). A readability formula that saves time. *Journal of Reading, 11*, 513-516, 575-578.

Fry, E. B. (1969). The readability graph validated at primary levels. *The Reading Teacher, 22*, 534-538.

Fry, E. B. (1977). Fry's readability graph: Clarifications, validity, and extension to level 17. *Journal of Reading, 21*, 242-252.

Fry, E. B. (1985). *The NEW reading teacher's book of lists*. Englewood Cliffs, NJ: Prentice-Hall.

Fry, E. B. (1987). The varied uses of readability measurement today. *Journal of Reading, 30*, 338-343.

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

Gallant, R. (1965). Use of cloze tests as a measure of readability in the primary grades. In J. A. Figurel (Ed.), *Reading and inquiry* (pp. 286-287). Newark, Delaware: International Reading Associates.

Gilliam, B., Peña, S. C., & Mountain, L. (1980). The Fry graph applied to Spanish readability. *The Reading Teacher, 33*, 426-430.

Goodman, K. S. (1967). Reading: A psychological guessing game. *Journal of the Reading Specialist, 6*, 126-135.

Hamsik, M. J. *Reading, readability, and the ESL reader*. Unpublished doctoral dissertation, Florida State University, Tallahassee, FL.

Haskell, J. (1973). Refining the cloze testing and scoring procedures for use with ESL students. Unpublished doctoral dissertation, Columbia University.

Hinofotis, F. B. (1980). Cloze as an alternative method of ESL placement and proficiency testing. In J. W. Oller Jr. & K. Perkins (Eds.), *Research in language testing* (pp. 121-128). Rowley, Massachusetts: Newbury House.

Hunt, K. W. (1965). *Grammatical structures written at three grade levels*. Champaign, IL: National Council of Teachers of English.

Irvine, P., Atai, P., & Oller, J. W. Jr. (1974). Cloze, dictation, and the test of English as a foreign language. *Language Learning, 24*, 245-252.

Jonz, J. (1976). Improving on the basic egg: The M-C cloze. *Language Learning, 26*, 255-256.

Jonz, J. (1987). Textual cohesion and second language comprehension. *Language Learning, 37*, 409-38.

Jonz, J. (1990). Another turn in the conversation: What does cloze measure? *TESOL Quarterly, 24*, 61-83.

Kintsch W., & Keenan, J. M. (1973). Reading rate and retention as a function of the number of propositions in the base structures of sentences. *Cognitive Psychology, 5*, 257-274.

Kintsch, W., & Vipond, D. (1979). Reading comprehension and readability in educational practice and psychological theory. In L.-G. Nilsson (Ed.), *Perspectives on memory research* (pp. 329-365). Hillsdale, NJ: Lawrence Erlbaum.

Klare, G. R. (1963). *The measurement of readability*. Ames: Iowa State University.

Klare, G. R. (1984). Readability. In R. Barr, M. L. Kamil, & P. Mosenthal (Eds.), *Handbook of reading research* (pp. 681-744). NY: Longman.

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day English*. Providence, RI: Brown University.

Larson, R. (1987). How Readability was created. In Scandinavian PC Systems. *Readability program for the IBM PC, XT and AT* (pp. 8-1 to 8-20). Rockville, MD: Scandinavian PC Systems.

Lorge, I. (1959). *The Lorge formula for estimating difficulty of reading materials*. New York: Columbia Teachers College.

Manzo, A. V. (1979). Readability: A postscript. *Elementary English, 47*, 962-965.

Markham, P. L. (1985). The rational deletion cloze and global comprehension in German. *Language Learning, 35*, 423-430.

Miller, G. R. & Coleman, E. B. (1967). A set of thirty-six prose passages calibrated for complexity. *Journal of Verbal Learning and Verbal Behavior, 6*, 851-854.

Miura, A. (1979). *English loanwords in Japanese*. Rutland, VT: Tuttle.

Moyle, D. (1970). Readability—the use of cloze procedure. In J. Merrit (Ed.), *Reading and the curriculum* (pp. 159-168). London: Ward Lock Educational.

Mullen, K. (1979). More on cloze tests as tests of proficiency in English as a second language. In E. J. Briere & F. B. Hinofotis (Eds.), *Concepts in language testing: Some recent studies* (pp.21-32). Washington, DC: TESOL.

Neter, J., & Wasserman, W. (1974). *Applied linear statistical models: Regression analysis, analysis of variance, and experimental design*. Homewood, IL: Irwin.

Nguyen, L. T., & Henkin, A. B. (1982). A readability formula for Vietnamese. *Journal of Reading, 26*, 243-251.

Oller, J. W. Jr. (1972a). Dictation as a test of ESL proficiency. In H. B. Allen & R. N. Campbell (Eds.), *Teaching English as a second language: A book of readings* (pp. 346-354). New York: McGraw-Hill.

Oller, J. W. Jr. (1972b). Scoring methods and difficulty levels for cloze tests of proficiency in English as a second language. *Modern Language Journal, 56*, 151-158.

Oller, J. W. Jr. (1979). *Language tests at school: A pragmatic approach*. London: Longman.

Oller, J. W. Jr., & N. Inal. (1971). A cloze test of English prepositions. *TESOL Quarterly, 5*, 315-326.

Pike, L. W. (1973). *An evaluation of present and alternative item formats for use in the test of English as a foreign language*. Princeton, NJ: Educational Testing Service.

Porter, D. (1983). The effect of quantity of context on the ability to make linguistic predictions: A flaw in a measure of general proficiency. In A. Hughes & D. Porter (Eds.), *Current developments in language testing* (pp. 63-74). London: Academic Press.

Que Software. (1990). *Right writer: Intelligent grammar checker*. Carmel, IN.

Ransom, P. E. (1968). Determining reading levels of elementary school children by cloze testing. In J. A. Figurel (Ed.), *Forging ahead in reading* (pp. 477-482). Newark, Delaware: International Reading Association.

Revard, D. (1990). Tailoring the cloze to fit: Improvement of cloze tests through classical item analysis. Unpublished scholarly paper. Honolulu, HI: University of Hawaii at Manoa.

Ruddell, R. B. (1964). A study of the cloze comprehension technique in relation to structurally controlled reading material. *Improvement of Reading Through Classroom Practice, 9*, 298-303.

Scandinavian PC Systems. (1988). *Readability program for the IBM PC, XT and AT*. Rockville, MD.

Smith, F. (1975). *Comprehension and learning*. New York: Holt, Rinehart and Winston.

Smith, F. (1978). *Understanding Reading* (3rd ed.). New York: Holt, Rinehart, & Winston.

Stevens, K. C. (1980). Readability formulae and McCall-Crabbs Standard Test Lessons in Reading, *The Reading Teacher*, 33, 413-415.

Stubbs, J. B., & Tucker, G. R. (1974). The cloze test as a measure of ESL proficiency for Arab students. *Modern Language Journal, 58*, 239-241.

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly, 30*, 414-438.

Taylor, W. L. (1957). "Cloze" readability scores as indices of individual differences in comprehension and aptitude. *Journal of Applied Psychology, 41*, 19-26.

Thorndike, E. L., Lorge, I. (1959). *The teacher's word book of 30,000 words*. New York: Columbia Teachers College.

Weaver, W. W., & Kingston, A. J. (1963). A factor analysis of the cloze procedure and other measures of reading and language ability. *Journal of Communication*, 13, 252-261.

Zakaluk, B. L., & Samuels, S. J. (1988). *Readability: Its past, present, and future*. Newark, DE: International Reading Association.

James Dean Brown
Department of ESL
1890 East-West Road
Honolulu, Hawai'i 96822

brownj@hawaii.edu