

WHAT TEXT CHARACTERISTICS PREDICT HUMAN PERFORMANCE  
ON CLOZE TEST ITEMS?

James Dean Brown  
University of Hawaii at Manoa

ABSTRACT

This study explores the link between some of the linguistic characteristics of the text surrounding cloze test items and the corresponding item difficulty estimates. Fifty reading passages were randomly selected from an American public library and made into thirty-item cloze tests by deleting every 12th word. The subjects were 2298 EFL students from 18 university level institutions in Japan. Each student took one of the 30-item cloze tests. The 50 cloze tests were randomly assigned across all of the subjects. Any differences between the cloze tests or the individual test items were therefore assumed to be due to other than sampling differences. The result was a set of 1500 item difficulty estimates (50 tests times 30 items). Each item was also analyzed for linguistic characteristics, e.g., passage readability, number of words per sentence, frequencies of occurrence in passage(s), and others.

Correlational, factor and multiple-regression analyses of the linguistic characteristics indicated that there were clear groupings among the variables and showed that linguistic characteristics can be used to account for up to 31 percent of the variance in item difficulties. These results are discussed in terms of their implications for cloze testing research.

# WHAT TEXT CHARACTERISTICS PREDICT HUMAN PERFORMANCE ON CLOZE TEST ITEMS?

James Dean Brown  
University of Hawaii at Manoa

Cloze procedure first entered the literature when Taylor (1953) studied cloze as a device for estimating the readability of materials used in public education. Research next turned to the effectiveness of cloze as a measure of reading proficiency for native speakers of English (Ruddell, 1964; Bormuth, 1965, 1967; Gallant, 1965; Crawford, 1970). During the sixties and seventies, a number of studies were also done on the effectiveness of cloze as a measure of overall ESL proficiency (for excellent overviews on cloze research at that time, see Alderson, 1978; Oller, 1979). However, even a careful review of this work on cloze as an overall ESL proficiency test will reveal that the results are far from consistent across studies. As noted in Brown (1984, 1988b), even the relative reliability and validity of cloze tests have varied dramatically across studies and across the years.

In the ensuing decades, two distinct strands of cloze test research have surfaced. Both strands are concerned with the degree to which cloze test items are tapping students' abilities to manipulate linguistic elements. However, the two strands diverge when it comes to deciding which linguistic elements are indeed sampled. One group of researchers (Alderson, 1979; Porter, 1983; and Markham, 1985) argue that cloze test items are primarily assessing linguistic elements at the clause (or sentence) level, while another group (Chihara *et al*, 1977; Brown, 1983a; Bachman, 1985; Chavez-Oller *et al*, 1985; and Jonz, 1987, 1990) argue that cloze is primarily assessing students' abilities to deal with the intersentential components of the language.

The truth probably lies somewhere between these two extreme positions. It seems unlikely that cloze test items only assess clausal level skills; too many researchers have presented convincing arguments in support of the notion that cloze tests assess intersentential elements. It also seems unlikely that cloze items only measure intersentential elements. Top notch professionals like Alderson (1979), Porter (1983), Markham (1985) have all presented convincing arguments to the contrary.

Perhaps it is crucial to remember that the English language is complex and is governed by many different and overlapping rule systems. These systems range from morphemic and clausal level grammar rules to discourse and pragmatic level rules of cohesion and coherence. To make matters more knotty, these rule systems probably interact in intricate and, at present, unpredictable ways. Nevertheless, given that well-established sampling theory is accepted as the basis for much of the research in applied linguistics, it seems reasonable to accept the proposition that the semi-random selection procedures used to create cloze tests provide fairly representative samples of the written language including rule systems at the word, clause, sentence, discourse, and pragmatic levels. Naturally, such acceptance would require large samples as a prerequisite.

The central questions in cloze research today appear to revolve around the issue of how words, i.e., the units being sampled in a cloze test, are constrained by all of the levels of language. If there are many levels of rules which constrain the choices of words that writers make and if semi-random sampling creates a representative selection of these words, it must be concluded that cloze items assess the ability to use a complex combination of morpheme to pragmatic level rules (as they apply to individual words) in approximately the same proportions as they exist in the written language from which they were sampled. Therefore, taking the position that cloze items are essentially sentential, or the position that they are primarily intersentential, and then conducting studies to support either position only insures that the researcher will find what he/she is looking for. If both types of constraints are operating in the language, then both positions are equally tenable because researchers are likely to find what they are looking for. Unfortunately, both positions are also fundamentally wrong in that they tend to exclude the other position.

The overall purpose of this study was to explore the linguistic characteristics which make individual cloze items relatively easy or difficult. To that end, every effort was made to sift through the data and keep an open mind (while trying to dispassionately view cloze as a simple data gathering instrument) because it was hoped that the data would guide the researcher in investigating any existing patterns. To that end, the following exploratory and open-ended research questions were posed:

1. Are randomly selected cloze tests reliable and valid tools for gathering data on the linguistic text variables that may be related to their own item difficulty levels?
2. What linguistic text variables are significantly and meaningfully associated with item difficulty in a cloze environment?
3. What combination of linguistic text variables best predicts item difficulty in a cloze passage?

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 systematically controls variables that literally remain out of control in many ESL studies: the nationality and language background of the subjects. Whereas many studies report on students from a variety of institutions, countries and language groups, all of the subjects (N = 2298) in this project were: 1) studying at the university level, 2) Japanese nationals, and 3) first language speakers of Japanese. The subjects were selected as intact EFL courses from 18 different institutions. The subjects

ranged in age from 18 to 24 and included 880 females and 1418 males. During the administrations of the 50 cloze tests (see Materials below), the test forms that students received were randomly assigned across all of the universities and testing sessions. This was done so that the performances of the resulting groups could reasonably be assumed to be approximately equal across all 50 tests. There was an average of 45.96 students per test with a range of 42 to 50.

#### Materials

The cloze tests used here were based on passages randomly selected from books in the adult reading section of the Leon County Library in Tallahassee, Florida. Fifty such books were picked. Then a page was randomly selected from each book. The passages were chosen by backing up to the nearest logical starting point for a complete semantic unit and counting off about 400 to 450 words. Some passages were somewhat shorter (as short as 366 words) and some were longer (as long as 478 words) because the stopping point was also determined by logical stopping points, but the average length was about 412 words. The result was a set of 50 randomly selected passages which can probably be assumed to represent the types of passages that are encountered in American public library books.

A 12th word deletion pattern (for a total of thirty items on each test) was used to create cloze tests from the passages. The 12th word deletion pattern was used instead of a more traditional 7th word deletion pattern so that the items would have little if any affect on each other. In most cases, one sentence was left unmodified at the beginning of each passage and one or more sentences were unmodified at the end of each passage. There were also blanks for the students' biodata information (name, sex, age, native language, and country of passport), which were placed at the top of all passages along with directions for what the students must do in filling in the blanks and how the blanks would be scored. The final result was a set of 50 cloze tests (see Appendix A for example directions and 12 items taken from the pilot study reported in Brown, 1989).

An important issue in this study was the degree to which randomization was used throughout the passage selection processes including semi-random selection (every 12th word) for defining the blanks. Based on sampling theory, the remainder of this study depends heavily on the notion that the fifty 30-item cloze tests constitute a collection of 1500 items representative of all items that could have been created from the books in the Leon County Library. The representativeness of these passages is further supported by the finding that the lexical frequencies that exist in all 50 passages (taken together) correlated at .93 with the "Brown" corpus (Kucera & Francis, 1967; Francis & Kucera, 1982) and that even the relatively restricted sample of lexical frequencies for the 1500 cloze items correlated at .86 with that same corpus. Thus it is with some confidence that these passages and blanks are treated as representative samples of the English language, at least as it is written in the books found in an American library.

## Procedures

Because of concern that randomly selected cloze tests might be too difficult to provide meaningful results, a cautious approach was taken in the initial stages of this study. A pilot study (Brown, 1989) was conducted on the basis of five cloze tests representative of the entire set of 50 passages. This study was conducted at a combination of junior colleges and universities in Japan. It was found that the passages ranged in difficulty from mean percent scores of 15 percent to 40 percent with an average of 27 percent. It was decided that the full-blown study was justified because sufficiently meaningful results were found in the pilot study.

With that pilot study information in hand, data gathering began with the cooperation of a number of Japanese, American, and British EFL teachers at 18 universities in various locations throughout Japan (see Note 1). The tests were duplicated and randomly ordered such that all students had an equal chance of getting any one of the 50 tests. They were then shipped to Japan, where the tests were administered by the teachers to their own students. Explicit step-by-step directions were read aloud and clarified as necessary. A total of 25 minutes was allowed for completing the tests. According to the teachers, the 25 minute time limit proved sufficient.

The exact-answer scoring method was used throughout this study. Thus only the word that had occupied the blank in the original passage was counted as correct. This was justified by the fact that the results were not being reported to the students and by research which indicated a high correlation between exact-answer scoring results and other scoring procedures (for more on this, see Alderson, 1979 and Brown, 1980). Another substantial reason for adopting the exact-answer scoring method in this study was that a single correct answer for each blank was essential for some of the linguistic analyses. For instance, it was important that the lexical frequencies involved be traceable to a single possible lexical item, or that the number of characters per word be based on a single word, etc.

## Analyses

The analyses in this study were all based on two kinds of variables: one dependent variable and sixteen independent variables. The dependent variable was item difficulty (ITEM DIFF), which was defined as follows:

- 1) ITEM DIFF - the proportion of students who correctly answered each of the 1500 cloze items.

ITEM DIFF was calculated by dividing the number of students who correctly answered each item by the total number of students who took the test in which it was found. Hence, if 21 out of 42 students answered an item correctly, the item difficulty for that item would be .50 ( $21 \div 42 = .50$ ). In other words, ITEM DIFF gives an estimate of how difficult (or easy) the students found each item to be by estimating the proportion of students who answered correctly. ITEM DIFF was the dependent variable in this study because it was the

principal variable of interest and because it was measured "to determine what effect, if any, the other types of variables may have on it" (Brown, 1988a: 10).

All of the independent variables were chosen because they were linguistic characteristics that were potentially related to the ITEM DIFF dependent variable and because they were quantifiable. In other words, the independent variables were selected because they might statistically explain, at least in part, what makes cloze items difficult (or easy). The independent variables (numbered 2-17 in this study) were defined as follows:

- 2) ITEM CORR - Item correlation is the point-biserial correlation coefficient between the students performance on each item on the test and the total score for the same test
- 3) CONT-FUNC - This variable indicated whether the correct answer for a blank was a content word or a function word. Content words included nouns, verbs, adjectives and adverbs. Function words included articles, prepositions, conjunctions and auxiliaries. Thus CONT-FUNC was a dichotomous variable. [The importance of this last observation is that this variable, unlike all of the others, was treated as a "dummy" variable (i.e., coded with 0 for content words and 1 for function words).]
- 4) CHRS/WORD - The number of characters found in the correct answer for each blank
- 5) SYLL/WORD - The number of syllables found in the correct answer for each blank
- 6) ITEM FREQ - The frequency with which the correct answer appeared elsewhere among the items of the 50 cloze passages
- 7) PASS FREQ - The frequency with which the correct answer appeared elsewhere in the passage
- 8) STDY FREQ - The frequency with which the correct answer appeared elsewhere in all 50 passages in this study.
- 9) BRWN FREQ - The frequency with which the correct answer appeared in the "Brown" corpus (see Kucera & Francis, 1967; Francis & Kucera, 1982)
- 10) WRDS/SENT- The number of words in the sentence in which the blank was found
- 11) WRDS/T-UN- The number of words in the T-unit (see Hunt, 1965; Gaies, 1980) in which the blank was found
- 12) SYLL/SENT- The number of syllables in the sentence in which the blank was found
- 13) SYLL/T-UN- The number of syllables in the T-unit in which the blank was found
- 14) READBLTY1- The Flesch-Kincaid readability index (as described in Klare, 1984) for the passage in which the blank was found

- 15) READBLTY2- A modified version of the Gunning Fog readability index (see Larson, 1987) for the passage in which the blank was found
- 16) READBLTY3- The Fry readability index (see Fry, 1985) for the passage in which the blank was found
- 17) ITEM NUMB- The item number of the blank within the test (ranging from 1 for the first item to 30 for the last item)

The analyses in this study included descriptive statistics for the 50 cloze tests as well as for all of the variables studied here. Simple Pearson product-moment correlations were also calculated between all possible pairs of the 17 variables defined above in order to determine the degree of relationship involved in each pairing. Factor analyses techniques, including principal components analysis and Varimax rotation, were used to investigate the degree to which common factors existed in the correlation matrix. 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 ITEM DIFF dependent variable.

### RESULTS

The results of this study begin with descriptive statistics for the 50 sets of test results followed by similar statistics describing the dependent and independent variables. Table 1 describes the overall test characteristics for all 50 cloze tests in terms of the mean (MEAN), standard deviation (SD), minimum score obtained (MIN), maximum score (MAX), the number of subjects who took the particular cloze (N), the internal consistency reliability (RELIA.) of the test using the split-half method adjusted by the Spearman-Brown formula, and the three readability indices (described above as variables 14-16) for the 50 passages.

[Insert Table 1 about here]

Notice that the means of the 50 cloze tests range from 1.020 to 9.918. Since, the 50 groups of students were randomly assigned to the tests and can therefore reasonably be assumed to be about equal in overall proficiency, the variety among the means reported in Table 1 probably indicates that there is considerable variation in the difficulty of these passages. Examination of the readability indices shown in the last three columns leads to a similar conclusion. Notice that the standard deviations also range widely, from a low of 1.247 to a high of 4.435. This range of standard deviations indicates that there was considerable variation in the degree to which the students' scores were dispersed on these cloze tests. The MIN and MAX indicate similar variations in dispersion with the MIN ranging from 0 to 4 and the MAX ranging from 3 to 21. Due to sampling errors, the number of subjects also ranged from 42 to 50.

At first glance, the reliability estimates for the individual cloze tests seem to indicate that most of the cloze

tests were moderately reliable in the .70 to .80 range. However, these reliability estimates ranged considerably from one exceptionally low one of .172 to a high of .869. The average of these 50 reliability estimates (using the Fisher  $z$  transformation) turned out to be .71. However, since the results are based on the much longer 1500 item five cloze test results, the Spearman-Brown formula was also applied to explore the effects of adjusting for the difference in length between each of the 30 item tests and the 1500 item total. Based on the average reliability (.71), the adjusted reliability estimate turned out to be .99, which is interpreted here as a rough estimate of what the reliability of the whole set of tests would be if it were possible to administer them as one large 1500 item test. The magnitude of these reliability estimates is important in the sense that the results of the study can be no more reliable than the tests upon which they are based.

[Insert Table 2 about here]

Table 2 focuses on the statistical characteristics of the dependent and independent variables. For the dependent variable, ITEM DIFF, the mean was .1372. In other words, on the whole, the cloze tests were fairly difficult for the students with an average of only 13.72 percent of the students answering each item correctly. More importantly for this type of project, the tests appear to have generated a wide variety of item difficulty indices, as indicated by the MIN and MAX columns, which show that ITEM DIFF ranged quite widely from 0 percent to 96 percent of the students correctly answering individual items. Since the purpose of this study was to study what causes such items to vary in difficulty, the wide variety of item difficulties (0 to 96 percent) was a good sign. However, one possible problem appears in the results for this variable. Notice that the SD is considerably larger than the MEAN for ITEM DIFF. This fact poses a potential problem in that it indicates that the item difficulty indices were not normally distributed. Since the correlational analyses that follow assume normal distributions on the variables involved, this was felt to be a potentially serious problem. Similar situations arose for the word-frequency variables (numbers 6 - 9) as well. [For an explanation of how this was handled see Note 2.]

The same descriptive statistics (MEAN, SD, MIN and MAX) are given in Table 2 for each of the independent variables. In order to make interpretation easier, these independent variables are numbered in the same order as their definitions in the Analyses section. Note that the variables are being described as they occurred across all 1500 items in the 50 cloze tests.

[Insert Table 3 about here]

Table 3 shows the simple correlation coefficients for all possible pairs of the variables in this study. Notice that the abbreviation NS has been placed in all of those places



where the correlation coefficient was not significant (statistically at  $p < .01$ ; two-tailed;  $df = 1498$ ).

Note also that there are five triangles marked off along the diagonal on the right side of the table. Triangle A only includes the moderately high correlation coefficient of .76 between the two sets of item statistics (variables 1 & 2). Triangle B delineates the coefficients for the degree of relationship among those variables which can be said to describe the items at the word level (variables 3-5). Triangle C contains the coefficients for the four lexical frequencies (variables 6-9). Triangle D emphasizes those correlation coefficients for variables related to sentence/T-unit level counts (variables 10-13). Finally, Triangle E contains the correlation coefficients between the various possible pairs of the three passage level readability indices used in this study (variables 14-16).

With the exception of Triangle B, the coefficients within each triangle are consistently higher than the other coefficients elsewhere in the table. This consistency indicates that the associated variables in each triangle are more highly related to each other than they are to the other variables. As for the word level variables involved in Triangle B, the number of characters and syllables per word are fairly highly correlated .80. However, the content-function word distinction is less highly correlated with these character and syllable counts, and is negatively correlated with them because the number of characters and syllables per word is higher for content words (coded 0) than for function words (coded 1).

However, notice in Rectangle 2 that all three of these word level variables appear to be moderately correlated with all of the frequency count variables (variables 6-9). The negative correlations for variables 4 and 5 with the frequency variables 6-9 are logical in that the number of characters and syllables tends to be small for frequent words and large for rare words.

The correlation coefficients found in Rectangle 3 indicate that there is a low to moderate degree of relationship between the readability indices reported in this study and the words and syllable counts for sentences and T-units. Even a cursory examination of such readability indices will indicate that syllable or word counts per sentence are key elements in their make up.

Since the focus of this analysis was on the degree to which each of the independent variables predicted item difficulty, the correlation coefficients of central interest are those found in Rectangle 1. Notice that all of these correlation coefficients, whether negative or positive, were significant. In other words, all of the independent variables appear to be related to the proportion of students answering correctly (ITEM DIFF) on the 1500 cloze test items. This may not at first seem particularly remarkable until it is noted that the independent variables, which are all simple linguistic counts in the text of the 50 passages, are each predicting (at least to some degree) the performances of

living, breathing students on those items, as represented by the item difficulty estimates.

Clearly, some of the independent variables are more highly related to ITEM DIFF than others (e.g., 3-9 are more highly correlated than 10-17). In addition, a little common-sense reflection on how each of the variables is defined in this study will clarify why some of the correlation coefficients are negative while others are positive. None of these relationships are counter-intuitive in the context of this study. Also note that the same relationships (with slightly lower correlation coefficients) exist for the ITEM CORR variable listed in the second column of numbers. This suggests that, at least under the conditions of this study, the same variables that contribute to the magnitude of ITEM DIFF also contribute to a slightly lesser degree to the degree to which items discriminate between the high scoring and low scoring students on the whole test.

[Insert Table 4 about here]

The fact that all of the independent variables were significantly related to ITEM DIFF led to an investigation of the degree to which various combinations of these variables might logically be grouped. To that end a Principal Components Analysis was conducted on the correlation matrix and followed up with a Varimax rotation for the same matrix. The results of the Varimax rotation (with the Eigen value set at 1.0) are presented in Table 4. Notice that five factors were extracted. As indicated by the rectangle around the loadings for some of the variables in the first factor, the variables that loaded highest on the first factor were clearly those which might be called word level variables as well as those which could be labeled as lexical frequencies. The second factor received the highest loadings from the sentence/T-unit level variables. The third factor had high loadings only from the item statistic variables. The fourth factor received the highest loadings from those variables which would be labeled passage level readability variables. And finally, the fifth factor had only one variable loading heavily on it, item number. Clearly then, there are sensible groupings among the variables as indicated by the five factors, which together account for more than 86 percent of the variance in the correlation matrix.

[Insert Table 5 about here]

The degree to which the types of variables listed in the previous paragraph were collectively related to item difficulty was further investigated using multiple-regression analysis. A forward-stepping multiple-regression analysis was conducted such that all of the independent variables were entered as possible predictors of the dependent variable. The results were that one variable each from the word level, lexical frequency level, sentence/T-unit level, passage level, and item number make the best combination of predictors. In each case, that variable within each level which was most

highly correlated with the ITEM DIFF dependent variable was the survivor. The results are presented in Table 5, which illustrates the progressive additivity of the variables and presents the multiple correlations (MR) as well as the multiple coefficients of determination ( $MR^2$ ).

The results indicate that the combination of PASS FREQ + READBLTY2 + CHRS/WORD + SYLL/T-UN + ITEM NUMB taken together produce a multiple-correlation (MR) of .56 and a corresponding  $MR^2$  of .31. This means that this combination of simple countable independent variables taken together predicts about 31 percent of the variance in the performance of Japanese students on these 1500 items. In more specific terms, it indicates that item difficulty is influenced to some degree by each of the following text elements:

- 1) blanks with words that are frequently found elsewhere in the passage are easier, while less frequent words are harder
- 2) blanks found in passages with lower readability levels are easier, while blanks in high readability passages are more difficult
- 3) blanks which have short words as the answer are easier than blanks with long words
- 4) blanks which appear in T-units with fewer syllables are easier than blanks which appear in T-units with more syllables
- 5) blanks which appear late in the test will have lower item difficulty results than blanks that appear early in the test (this may be a result of fatigue, discouragement, or other motivational factors)

Again, all of this may not initially appear to be particularly interesting because it only explains 31 percent of the variance in ITEM DIFF values; there remains 69 percent of the variance in ITEM DIFF that is unexplained. However, given that these independent variables are based on simple linguistic counts related to the word in each cloze blank (in this case, the frequency of occurrences of a word in the passage, the readability of the passage, the number of characters in the word, the number of syllables in the T-unit in which it was found, and the item number), it is remarkable that they predict nearly one-third of the variance in the difficulty that Japanese students have in filling in those same blanks.

#### DISCUSSION.

The discussion in this section will now return to the original three research questions posed at the out set of this study, and then the CONCLUSION section will touch on the implications of these findings for cloze testing in language testing contexts.

1) Are randomly selected cloze tests reliable and valid tools for gathering data on variables that are related to their own item difficulty levels?

In answering this research question, it was necessary to consider the descriptive statistics for all of the variables,

as well as the degree to which the cloze tests were reliable and valid for the stated purposes of this study. That is why this research question was placed first. In a sense, positive results for this research question were prerequisite to answering either of the other two.

The descriptive results of this study indicate that the cloze tests did function reasonably well for observing at least the variables explored in this study. However, it is important to note that overall, the means were relatively low on these tests, i.e., the highest mean of 9.918 represented a percentage score of only 33 percent. As scored here (using the exact-answer method), the cloze passages appear to be difficult from a testing perspective. However, as a tool for observing language behavior, these cloze tests appeared to be more than adequate because they produced item difficulty estimates ranging from .00 to .96 for the dependent variable and a wide variety of descriptive statistics for the independent variables, as well.

In terms of reliability, the cloze passages reported in this study seem to have been reasonably sound. This is indicated by the average split-half reliability estimate of .71 for the 50 cloze tests. However, it must be recognized that the reliability indices varied considerably ranging from .172 to .869. Since the analyses here were based on the total sample of cloze 1500 items, it may also be worth noting that adjusting the average reliability for the 50 cloze tests as though they could be taken together resulted in a .99 overall estimate of reliability. However, this last estimate is largely a result of the extraordinary 1500 item length of the theoretical cloze test involved.

Special attention must be paid to the fact that some of the cloze tests had low reliabilities. These low reliability estimates may be due in part to the fact that the samples in this study were relatively homogeneous in ability levels. The samples were all made up of randomly selected students at roughly the same levels of study. This meant that they were fairly uniform in ability levels because they had, by definition, all studied many years of English before arriving in Japanese universities. Because of this uniformity, the range of possible scores may have been restricted (as reflected in the relatively low standard deviations). Such restrictions in range are often associated with relatively low reliability estimates. [See Brown, 1984 for more on the relationship between the standard deviation and reliability estimates.]

The validity of these 50 cloze passages for the purposes of this study can be argued in common-sense terms without resorting to elaborate statistical analyses. To begin with, consider the fact that the cloze passages were randomly selected and that the items were selected on a semi-random basis (i.e., every nth word deletion) within each passage. Based on sampling theory, the passages can be said to be representative samples of the language contained in the books in that library, and the items can be said to provide a representative sample of the words contained in the passages. Since the validity of a test can be defined as the degree to

which it is measuring what it claims 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 the written language which is found in an American public library (as it is tapped by single word blanks).

Based on the forgoing arguments, the cloze tests in this study were viewed with some confidence as being reliable and valid for the purposes of gathering data on the variables of interest in this study.

2) What variables are significantly and meaningfully related to item difficulty indices in a cloze environment?

It seems, then, that a number of relatively simple and countable variables are related to item difficulty. The most striking relationships are represented by the correlation coefficients between ITEM DIFF and those variables connected with the frequency of the word when it is compared to the words in the other items in the 50 cloze tests, the words elsewhere in the same passage, the words in the 50 passages of this study, and the words in the "Brown" corpus (Kucera & Francis, 1967; Francis & Kucera, 1982). Another striking set of relationships is the one found between ITEM DIFF and the word level variables like the content-function distinction, the number of characters per word, and the number of syllables per word. Other correlation coefficients between ITEM DIFF and the remaining independent variables are lower in magnitude, but nonetheless interesting. All of these remaining variables were correlated with ITEM DIFF either negatively or positively at the  $p < .01$  significance level. Thus none of these variables can easily be dismissed; they all appear to have non-chance relationships with ITEM DIFF.

Naturally, there is no end to the number of different types of variables that might be considered in a study such as this one. For example, at the word level, it might be useful to consider whether each word is of Latinate or Germanic origin. At the lexical frequency level, it might help to consider additional word-frequency lists like those found in Thorndike and Lorge (1959). At the sentence level, there may be other indicators of syntactic complexity that should have been considered. At the passage level, it might prove profitable to examine the item difficulties in terms of other readability scales like the Lorge (1959) scale. Perhaps cohesive devices should be brought into the model, or other discourse/pragmatic level elements.

The main point is that the results of this study are sufficiently encouraging in terms of the number and strength of the observed relationships to justify further analysis of these data and to promote the application of this type of research elsewhere.

### 3) What combination of variables best predicts item difficulty in a cloze environment?

The groupings shown in Table 4 were clearly related to the regression analysis in that two variables from the first factor and one each from the second, fourth and fifth factors were automatically selected by the step-wise procedure for inclusion in the regression model. Overall, the regression analysis indicated that the best combination of variables for predicting item difficulty (shown in Table 5) included: ITEM DIFF = PASS FREQ + READBLTY2 + CHRS/WORD + SYLL/T-UN + ITEM NUMB. This combination of independent variables produced a multiple correlation of .56 with the dependent variable. Naturally, such results must always be interpreted very carefully. For instance, these results do not necessarily mean that these same variables in this same order will be found in a replication of this study. Consider the fact that a similarly strong prediction was made in the pilot study for this project (Brown, 1989), but that the prediction was based on a different combination of variables which appeared in a somewhat different order. The results of that regression analysis indicated a multiple correlation of .57 based on only four variables: PASS FREQ + CHRS/WORD + SYLL/SENT + CONT-FUNC.

It seems then that various combinations of variables, when taken from different levels (i.e., the word level, lexical frequency level, sentence/T-unit level, and passage level) may provide a sufficiently high degree of prediction for ITEM DIFF to be meaningful. However, the variable within each of those categories that will be involved and the order in which they will appear in the regression model may differ from study to study. The important observation is that human performance on cloze items (as operationalized by ITEM DIFF) can be predicted by certain linguistic qualities of the texts involved (at least the ones that can be operationalized by simple counts, as in this study).

### CONCLUSION

One of the limitations of this study is that it focused only on the performance of Japanese students. Thus the results can only be responsibly generalized to Japanese students. However, the fact that only one nationality was used is also one of the strengths of this study. In most studies conducted in the United States and other ESL settings, various language backgrounds are typically mixed together. As such, the results are often difficult to interpret or apply because they cannot be generalized beyond the single situation in which the data were gathered. While the sample in this study cannot be said to be a random sample of all Japanese university students, it can be viewed as homogeneous with regard to the nationality, language background and educational level of the students.

[Insert Table 6 about here]

In general terms, the results of this study indicate that, for Japanese university students, a wide variety of variables were significantly correlated with the item

difficulty values on 50 cloze tests. These variables fell into categories that proved useful in looking for patterns in the results. Table 6 summarizes the correlation coefficients (of various independent variables with ITEM DIFF) after they have been reorganized so that those variables which are primarily LOCAL in nature (i.e., at the word level, sentence/T-unit level) are grouped together and those variables which are more GLOBAL in nature (i.e., lexical frequency level, and passage level) are also grouped together. Notice that the highest coefficients are those for the lexical frequency counts, and that somewhat lesser magnitudes were produced for the word level variables. This suggests that, for Japanese students, the lexical frequency and word level factors were more highly related to performance on individual items than the other factors included in this design. Nonetheless, the other variables do appear to contribute to the variance in item difficulty estimates as indicated by their statistically significant correlations with ITEM DIFF, as well as by the significant contribution made by some of these other variables in the regression analysis.

It would be dangerous to argue on the basis of these results that cloze tests are primarily measuring at the word, clause or sentence level, or for that matter, that cloze tests focus predominately on more global, intersentential elements. However, the evidence here seems to support the notion (suggested at the outset of this study) that, at least for Japanese students, performance on cloze test items is related to a wide variety of factors. True, student performance was shown to be most highly related to lexical frequency factors, but it also appears to be significantly correlated with a number of other factors at the word level, sentence/T-unit level and passage level. Hence cloze test items appear to be measuring at a number of different levels simultaneously, and, naturally, there may be a large number of possible interactions among these levels, as well. It is hoped that this study has at least provided a start in the direction of discovering what makes cloze test items easy or difficult, perhaps even a start in the direction of discovering what cloze tests are actually measuring.

#### Implications and Future Directions

The overall results of this study seem encouraging enough to continue doing research that explores the link between human performance on cloze tests and linguistic qualities of the passages involved. Further research should probably continue to examine the variables covered in this study as well as whatever more complex linguistic variables can be isolated and shown to be contributing to the relative difficulty of cloze items.

As is often the case in research, more questions were raised in the process of doing this study than were answered. The following general questions are provided in the hope that

other researchers will find this line of inquiry interesting enough to pursue in the future:

1. What differences and similarities would occur if this study were replicated at other institutions in Japan?
2. What differences and similarities would occur if this study were replicated in other countries to include students from other language backgrounds?
3. Are cloze tests naturally reliable and valid measures? What differences occur among randomly selected passages in terms of test characteristics?
4. What combinations of passage level variables best predicts the overall passage readability levels?
5. What other linguistic variables might be included in readability research, and how well would they predict cloze test difficulty?
6. What hierarchies of difficulty are found for any of the linguistic variables (separately or combined) that would have implications for second language acquisition research?



## NOTES

<sup>1</sup> The author would like to thank all of those colleagues who helped in this project by administering tests in the pilot study at Baika Junior College, Kobe Yamato Junior College, National University and Wakayama University. The author would also like to thank those who helped by administering tests for the main body of this project 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. The author would like to thank Dr. Ian Richardson for his help in selecting and creating the cloze tests used here. He is presently a professor at King Saud University in Abha, Kingdom of Saudi Arabia. Thanks are also due to Dr. Thom Hudson at the University of Hawaii at Manoa for his careful readings and comments on earlier versions of this paper.

<sup>2</sup> Please note that the ITEM FREQ, PASS FREQ, STDY FREQ and BRWN FREQ variables were transformed in the correlational, factor and regression analyses. Standard log transformations (see Chatterjee & Price, 1977, pp. 27-38, or Neter & Wasserman, 1974, pp. 121-130) were needed to correct for curvilinear relationships found when scatterplots were examined. After transformation, visual examination indicated that the relationships were approximately linear. It should be noted that Carroll (1967) found that the word-frequency distributions are lognormally distributed.

<sup>3</sup> The apparent overall difficulty of these cloze tests (indicated by the average ITEM DIFF OF 13.72) is probably due in large part to the fact that an exact-answer scoring method was used. Had an acceptable-answer scoring scheme been used instead, the mean item difficulties would have probably been much higher (for example, in Brown, 1980, the mean score for acceptable-answer scoring turned out to be 71 percent higher than the mean for exact-answer scoring).

<sup>4</sup> Note that STDY FREQ was also selected statistically for inclusion in the model, but it was eliminated in the final analysis because it was causing problems of multicollinearity in the overall analysis, and yet was adding little to the MR.

## APPENDIX A: EXAMPLE CLOZE PASSAGE

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."

## 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 and J. Schachter (Eds.) On TESOL '79 (pp. 193-205). Washington, D.C.: TESOL.
- Bachman, L.F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. TESOL Quarterly, 19, 535-555.
- 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. (1967). Comparable cloze and multiple-choice comprehension tests scores. Journal of Reading, 10, 291-299.
- 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: Validity and reliability. In J.W. Oller Jr. (Ed.) Issues in Language Testing (pp. 237-243). Rowley, MA: Newbury House.
- Brown, J.D. (1984). A cloze is a cloze is a cloze? In J. Handscombe, R.A. Orem and B.P. Taylor (Eds.) On TESOL '83 (pp. 109-119). Washington, D.C.: TESOL.
- 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.
- 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.
- Chatterjee, S. & B. Price. (1977). Regression analysis by example. New York: John Wiley & Sons.
- Chavez-Oller, M.A., T. Chihara, K.A. Weaver and J.W. Oller Jr. (1985). When are cloze items sensitive to constraints across sentences? Language Learning, 35, 181-206.
- Chihara, T., J.W. Oller Jr., K.A. Weaver and M.A. Chavez-Oller. (1977). Are cloze items sensitive to constraints across sentences? Language Learning, 27, 63-73.
- Crawford, A. (1970). The cloze procedure as a measure of reading comprehension 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 and Row.
- Francis, W.N. & H. Kucera. (1982). Frequency analysis of English usage: Lexicon and grammar. Boston, MA: Houghton Mifflin.
- Fry, E. (1985). The NEW reading teacher's book of lists. Englewood Cliffs, NJ: Prentice-Hall.

- 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.
- Hunt, K.W. (1965). Grammatical structures written at three grade levels. Champaign, IL: National Council of Teachers of English.
- 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.
- Klare, G.P. (1984). Readability. In P.D. Pearson (Ed.) Handbook of reading research (pp. 681-744). NY: Longman.
- Kucera, H. & W.N. Francis. (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.
- Markham, P.L. (1985). The rational deletion cloze and global comprehension in German. Language Learning, 35, 423-430.
- Neter, J. & W. Wasserman. (1974). Applied linear statistical models: Regression analysis, analysis of variance, and experimental design. Homewood, IL: Irwin.
- Oller, J.W. Jr. (1979). Language tests at school: A pragmatic approach. London: Longman.
- 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 and D. Porter (Eds.) Current developments in language testing (pp. 63-74). London: Academic Press.
- Ruddell, R.B. (1964). A study of the cloze comprehension technique in relation to structurally controlled reading material. Improvement of reading through Classroom Practice, 2, 298-303.
- Taylor, W.L. (1953). Cloze procedure: A new tool for measuring readability. Journalism Quarterly, 30, 414-438.
- Thorndike, E.L. and I. Lorge. (1959). The teacher's word book of 30,000 words. New York: Columbia Teachers College.

TABLE 1: DESCRIPTIVE STATISTICS FOR 50 CLOZE TESTS

TEST	MEAN	SD	MIN	MAX	N	RELIA.	READABILITY		
							1	2	3
1	5.229	3.164	0	15	48	0.708	9.6	32	7
2	4.208	3.421	0	13	47	0.858	13.5	42	13
3	2.021	2.126	0	10	48	0.735	4.8	21	3
4	7.543	3.866	2	16	46	0.803	7.6	28	6
5	3.979	2.787	0	13	47	0.734	13.9	40	10
6	5.106	3.230	0	14	47	0.803	7.0	27	6
7	6.140	3.407	0	16	43	0.825	9.9	43	10
8	3.156	2.270	0	8	45	0.457	11.2	36	8
9	2.848	2.458	0	11	46	0.773	15.3	49	12
10	2.543	2.310	0	8	46	0.825	15.2	46	10
11	5.935	3.358	0	16	46	0.742	5.0	20	3
12	8.980	3.967	0	21	47	0.789	11.0	32	10
13	2.870	1.714	0	8	46	0.503	12.1	40	10
14	3.234	2.503	0	9	47	0.682	8.5	27	6
15	9.180	3.416	4	18	49	0.683	12.0	38	10
16	1.360	1.411	0	6	48	0.650	13.0	50	9
17	1.383	1.247	0	5	46	0.348	20.4	58	14
18	1.020	1.086	0	3	50	0.500	12.7	40	12
19	4.760	2.881	0	10	50	0.701	10.2	35	8
20	4.375	3.238	0	15	47	0.855	10.8	35	8
21	9.918	4.435	0	19	48	0.840	7.5	24	5
22	3.702	2.858	0	11	47	0.841	10.8	37	9
23	3.638	2.401	0	11	43	0.646	13.9	46	13
24	2.957	2.259	0	9	47	0.436	13.1	40	10
25	5.362	2.740	0	12	46	0.627	10.2	31	7
26	2.681	1.559	0	5	47	0.172	16.6	54	14
27	2.340	2.723	0	13	47	0.869	10.0	38	9
28	2.581	2.170	0	8	43	0.574	14.4	49	14
29	2.318	1.768	0	7	44	0.640	16.0	46	11
30	9.563	3.284	3	16	48	0.715	6.5	22	5
31	3.783	3.078	0	15	46	0.832	11.6	37	10
32	3.833	2.525	0	9	42	0.770	9.6	30	8
33	2.136	1.866	0	6	44	0.633	16.3	59	12
34	5.867	2.918	0	13	45	0.819	12.8	42	10
35	6.630	3.662	0	17	45	0.719	4.8	22	4
36	5.000	2.054	0	9	46	0.505	11.3	40	8
37	5.458	3.657	0	13	48	0.767	8.6	31	2
38	1.708	1.567	0	8	48	0.746	12.9	42	11
39	2.511	1.977	0	9	47	0.648	6.7	27	6
40	3.488	1.897	0	9	43	0.659	8.1	30	6
41	2.870	2.507	0	10	43	0.764	14.3	47	12
42	4.409	3.099	0	18	44	0.811	9.1	31	8
43	1.432	1.452	0	7	44	0.190	13.9	43	10
44	3.239	2.521	0	10	46	0.673	13.9	43	11
45	6.548	3.874	0	16	42	0.788	11.1	36	8
46	2.163	1.816	0	7	47	0.307	11.2	34	9
47	3.791	2.328	0	11	43	0.685	11.9	40	9
48	2.690	2.121	0	11	42	0.738	11.2	44	8
49	4.564	2.808	0	11	49	0.748	10.3	37	7
50	2.488	2.697	0	12	45	0.774	21.3	64	15

TABLE 2: DESCRIPTIVE STATISTICS FOR ALL VARIABLES

VARIABLE	MEAN	SD	MIN	MAX
1) ITEM DIFF	.1372	.1925	0.00	.9600
2) ITEM CORR	0.24	0.23	-0.16	0.76
3) CONT-FUNC	[Not applicable for a dichotomous variable]			
4) CHRS/WORD	4.59	2.44	1.00	17.00
5) SYLL/WORD	1.53	0.90	1.00	11.00
6) ITEM FREQ	16.00	29.20	1.00	109.00
7) PASS FREQ	6.74	9.33	1.00	55.00
8) STDY FREQ	180.73	303.65	1.00	1024.00
9) BRWN FREQ	10118.20	19251.30	0.00	69971.00
10) WRDS/SENT	24.27	13.38	1.00	104.00
11) WRDS/T-UN	22.98	13.25	1.00	104.00
12) SYLL/SENT	36.96	21.31	1.00	148.00
13) SYLL/T-UN	35.04	21.17	1.00	148.00
14) READBLTY1	11.47	3.50	4.80	21.30
15) READBLTY2	38.11	9.80	20.00	64.00
16) READBLTY3	8.92	2.98	2.00	15.00
17) ITEM NUMB	NA	NA	1.00	30.00

TABLE 3: CORRELATION COEFFICIENTS FOR ALL VARIABLES

1 ITEM DIFF	1.00																
2 ITEM CORR	0.76	1.00															
3 CONT-FUNC	0.29	0.27	1.00														
4 CHRS/WORD	-0.39	-0.35	-0.52	1.00													
5 SYLL/WORD	-0.31	-0.30	-0.34	0.80	1.00												
6 ITEM FREQ	0.42	0.39	0.78	-0.66	-0.48	1.00											
7 PASS FREQ	0.48	0.43	0.62	-0.52	-0.39	0.80	1.00										
8 STDY FREQ	0.47	0.44	0.75	-0.75	-0.56	0.93	0.79	1.00									
9 BRWN FREQ	0.42	0.39	0.71	-0.74	-0.56	0.86	0.66	0.93	1.00								
10 WRDS/SENT	-0.16	-0.08	0.07	NS	NS	NS	NS	NS	NS	1.00							
11 WRDS/T-UN	-0.17	-0.10	0.07	NS	NS	NS	NS	NS	NS	0.94	1.00						
12 SYLL/SENT	-0.19	-0.10	0.07	0.07	0.07	NS	NS	NS	NS	0.96	0.92	1.00					
13 SYLL/T-UN	-0.19	-0.11	0.07	0.07	0.07	NS	NS	NS	NS	0.91	0.97	0.95	1.00				
14 READBLTY1	-0.19	-0.10	NS	0.11	0.12	NS	NS	NS	NS	0.36	0.38	0.48	0.50	1.00			
15 READBTY2	-0.20	-0.11	NS	0.11	0.12	NS	NS	NS	NS	0.34	0.37	0.46	0.48	0.95	1.00		
16 READBTY3	-0.17	-0.09	NS	0.10	0.11	NS	NS	NS	NS	0.30	0.32	0.42	0.44	0.90	0.88	1.00	
17 ITEM NUMB	-0.06	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	1.00
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17

Triangle A

Rectangle 1

Triangle B

Rectangle 2

Triangle C

Triangle D

Rectangle 3

Triangle E

TABLE 4: LOADINGS AND VARIANCE EXPLAINED FOR FACTOR ANALYSIS AFTER VARIMAX ROTATION

VARIABLE	FACTOR				
	1	2	3	4	5
1 ITEM DIFF	0.294	-0.098	-0.912	-0.091	-0.029
2 ITEM CORR	0.304	-0.047	-0.732	-0.021	0.118
3 CONT-FUNC	0.799	0.043	-0.076	0.089	0.049
4 CHRS/WORD	-0.823	0.015	0.135	0.134	0.082
5 SYLL/WORD	-0.675	-0.006	0.094	0.177	0.151
6 ITEM FREQ	0.914	0.014	-0.185	0.067	0.054
7 PASS FREQ	0.755	-0.006	-0.313	0.089	0.072
8 STDY FREQ	0.941	0.012	-0.221	0.012	0.034
9 BRWN FREQ	0.912	0.027	-0.167	0.002	0.017
10 WRDS/SENT	0.033	0.969	0.071	0.112	0.014
11 WRDS/T-UN	0.029	0.966	0.078	0.136	0.022
12 SYLL/SENT	0.008	0.944	0.080	0.251	0.025
13 SYLL/T-UN	0.004	0.937	0.087	0.270	0.030
14 READBLTY1	-0.014	0.256	0.083	0.938	-0.010
15 READBLTY2	-0.006	0.240	0.104	0.932	-0.009
16 READBLTY3	-0.014	0.195	0.067	0.928	-0.007
17 ITEM NUMB	-0.010	0.060	0.034	-0.028	0.974

86 PERCENT OF TOTAL VARIANCE EXPLAINED AS FOLLOWS:

1	2	3	4	5
27.010	20.209	18.028	15.123	5.353



TABLE 5: STEPWISE REGRESSION ANALYSIS OF  
ALL INDEPENDENT VARIABLES (EXCEPT ITEM  
CORRELATION) PREDICTING THE ITEM DIFFICULTY  
DEPENDENT VARIABLE

DEPENDENT = VARIABLE	INDEPENDENT VARIABLES	MR	MR <sup>2</sup>
ITEM DIFF = PASS FREQ		.4849	.2351
ITEM DIFF = PASS FREQ + READBLTY2		.5295	.2804
ITEM DIFF = PASS FREQ + READBLTY2 + CHRS/WORD		.5463	.2984
ITEM DIFF = PASS FREQ + READBLTY2 + CHRS/WORD + SYLL/T-UN		.5563	.3095
ITEM DIFF = PASS FREQ + READBLTY2 + CHRS/WORD + SYLL/T-UN + ITEM NUMB		.5587	.3121

TABLE 6: CORRELATIONS WITH ITEM DIFFICULTY  
(GROUPED BY LEVEL OF VARIABLES)

<u>LEVEL</u> VARIABLE	CORR WITH ITEM DIFF	<u>LEVEL</u> VARIABLE	CORR WITH ITEM DIFF
<b>LOCAL VARIABLES</b>		<b>GLOBAL VARIABLES</b>	
<u>WORD LEVEL</u>		<u>LEXICAL FREQUENCIES</u>	
CONT-FUNC	.29	ITEM FREQ	.42
CHRS/WORD	-.39	PASS FREQ	.48
SYLL/WORD	-.31	STDY FREQ	.47
		BRWN FREQ	.42
<u>SENTENCE/T-UNIT LEVEL</u>		<u>PASSAGE LEVEL</u>	
WRDS/SENT	-.16	READBLTY1	-.19
WRDS/T-UN	-.17	READBLTY2	-.20
SYLL/SENT	-.19	READBLTY2	-.17
SYLL/T-UN	-.19		