JALT JOURNAL, VOLUME 11, NO. 1

CLOZE ITEM DIFFICULTY

James Dean Brown

Abstract

This study explores the link between some of the linguistic characteristics of cloze test items and the corresponding item difficulty estimates.[1] Five reading passages were randomly selected from an American public library and made into 30-item cloze tests by deleting every 12th word. EFL students (N=179) at the post-secondary level in Japan each took one of the resulting 30-item cloze tests. The five cloze tests were randomly administered 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 150 item difficulty estimates (5 tests times 30 items), which served as the dependent variable: cloze item difficulty. Each item was also analyzed for linguistic characteristics, which served as the independent variables, e.g., the content/function word distinction, passage readability, number of words per sentence, frequencies of occurrence in passage(s) and many others. Multiple-regression analysis of the linguistic characteristics as predictors of the item difficulty estimates showed that characteristics such as frequency of occurrence, number of characters per word, and number of syllables per sentence account for up to 32 percent of the variation in item difficulties. These results are discussed in terms of their implications for language testing research and plans for future research on a larger scale.

1. Introduction

Cloze procedure initially surfaced when Taylor (1953) investigated its effectiveness as a tool for measuring the readability of materials for American school children. Research next focused on the utility of cloze as a measure of native-speaker reading proficiency (Ruddell, 1964; Bormuth, 1965, 1967; Gallant, 1965; Crawford, 1970). In the sixties, studies also began on cloze as a measure of overall ESL proficiency, and dozens

James Dean Brown is director of the English Language Institute, Department of English as a Second Language, University of Hawaii at Manoa. He is the author of *Understanding Research in Second Language Learning: A Teacher's Guide to Statistics and Research Design* and numerous articles on cloze testing and the interpretation of cloze tests.

of studies on this use for cloze have surfaced since (for excellent overviews on cloze research, see Alderson, 1978; Oller, 1979; Cohen, 1980). However, a care-ful review of the literature on cloze as a measure of overall ESL proficiency reveals that the results are far from consistent. For instance, Brown (1984) noted that the relative reliability and validity of cloze tests have varied considerably within and among the investigations.

Reliability indices indicate the degree to which a test produces consistent results. Such indices can range from a low of 0.0 (completely unreliable) to a high of 1.0 (perfectly reliable). Studies to date show reliabilities for cloze ranging from .31 to .96 (Darnell, 1970; Oller, 1972b; Pike, 1973; Jonz, 1976; Alderson, 1979; Mullen, 1979; Hinofotis, 1980; Brown, 1980, 1983b, 1984, 1988b; Bachman, 1985). In other words, there are a variety of results indicating that different cloze tests in different situations may vary from exceptionally weak to very strong in terms of reliability.

Similarly disparate results have been obtained for the validity of cloze tests. Validity coefficients are an indication of the degree to which a test is measuring what it claims to be measuring—in this case, overall ESL proficiency. The problem is commonly approached by calculating a correlation coefficient between the results on a cloze test and parallel results on some well-established criterion measure of ESL proficiency such as TOEFL. The squared value of such a correlation coefficient indicates the percentage of shared, or overlapping, variance betweeen the cloze test and the criterion measure. This type of validity is most often referred to as criterion-related validity. The studies reviewed here (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), reported correlation coefficients ranging from .43 to .91. The corresponding squared values, ranging from .19 to .83, indicate that various cloze tests were related to the criterion measures of EFL proficiency in a variety of ways: from very weak relationships (19 percent) to fairly strong ones (83 percent).

Many of the studies cited above were designed to discover which procedures were most efficient for developing and interpreting cloze tests in terms of reliability, validity, and other test characteristics. In the process, different combinations of the following variables were manipulated: (1) scoring methods, (2) frequency of deletions (e.g., every 5th word, every 7th word, etc.), (3) length of blanks, (4) textual difficulty, (5) native versus non-native performance, and (6) number of items. Over time, there has been some controversy, but a degree of consensus has also formed that certain

scoring methods, deletion patterns, etc. may be more effective than others.

Another strain of research has investigated the degree to which cloze test items are primarily tapping students' abilities to manipulate linguistic elements at the clause or sentence level, as opposed to predominately focusing on intersentential elements. The truth probably lies somewhere between the two positions or rather will be found in some combination of them. It seems unlikely that cloze items only assess clausal level skills; Chihara et al. (1977), Brown (1983a), Bachman (1985), Chavez-Oller et al. (1985) and Jonz (1987) have all presented arguments to the contrary. It seems equally absurd that cloze items measure exclusively at the intersentential level; Alderson (1979), Porter (1983), Markham (1985) have all come to the opposite conclusion. The point is that most linguists would concede that the English language is complex and is made up of a variety of constraints ranging at least from morphemic and clausal level grammar rules to discourse and pragmatic level rules of cohesion and coherence, all of which interact in intricate ways. Based on sampling theory, it is also a safe assumption that semi-random selection procedures like those used in creating a cloze test will create a representative sample of whatever is being selected as long as the samples are large enough. This assumption is the basis of much of the research done in the world today.

The question appears to hinge on the degree to which words, that is the units being sampled in a cloze test, are constrained by all of the levels of rules that operate in the language. If there are indeed different levels operating in the language which constrain the choices of words that writers make, and if semi-random sampling creates a representative selection of these words, there is no alternative but to conclude that cloze items tap a complex combination of morpheme to discourse level rules in approximately the same proportions as they exist in the language from which they were sampled. Thus taking either of the positions above (i.e., that cloze items are essentially sentential, or primarily intersentential) and then conducting studies to support either position is to insure that the investigators will find what they are looking for. If both types of constraints are in operation, then both schools of thought are correct in finding what they are looking for and fundamentally wrong in excluding the other possibility.

The project reported here expands on the views expressed by others that cloze tests are a "family of item types" (Mullen, 1979) and "merely a technique for producing tests, like any other technique" (Alderson, 1979). Since the overall purpose is to explore just what it is that makes cloze items easy or difficult, every effort has been made to actually explore (in the sense

of keeping an open mind) without gratuitously excluding possibilities, while remaining relatively dispassionate with regard to cloze as a data gathering instrument. Thus it is hoped that the data are guiding the researcher (rather than the other way around) in examining any existing patterns. Because this is just a first step in trying to discover some of the linguistic elements that cloze items tap, the intial research questions will necessarily remain very exploratory and open-ended throughout the study and the results will be important largely insofar as they point to useful directions for future research. To those ends, let's begin with the following set of research questions:

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

If the results of this study are encouraging in the sense that the data gathering methodology works and relationships of interest emerge, a much larger investigation may be pursued in the future. Because of the exploratory nature of this research, the alpha level for all statistical decisions was set at $\alpha < .05$.

2. Method

2.1 Subjects

This study attempts to control 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 countries and language groups, all of the subjects (N=179) in this project were studying at one of four post-secondary institutions;[2] they were all Japanese nationals and had Japanese as their first language. In addition, all of the students were intact groups enrolled in EFL courses in their respective institutions. They ranged in age from 18 to 23 and included 118 females and 61 males. During the administrations of the five cloze tests used here (see *Materials* below), the particular test that each student received was randomly assigned so that the performances of the resulting groups could reasonably be assumed to be approximately equal across the five tests.

2.2 *Materials*

The cloze tests were based on passages found in books randomly selected from the adult reading section of the Leon County Library in Tallahassee, Florida.[3] Five such books were collected. A page was randomly picked from each book; then a passage was selected by backing up to the nearest logical starting point for a complete semantic unit and counting off about 450 words. Some passages were somewhat longer because the stopping point was also determined by semantically logical stopping points. The result was a set of five randomly selected passages which are assumed to represent the types of passages that would be encountered in American public library books. They were entitled as follows: *A Father and Son* (fiction), *Terror in the Red Sea* (historical piece on piracy), *Visitors to James Cave* (about a cave in Kentucky), *A Short History of Ammunition* (about the development of gunpowder), *Most Problems Are Just Events* (fiction).

Each of these passages was then modified so that every 12th word was deleted and replaced by a blank for a total of 30 items. Two sentences were left intact at the beginning of each passage as were two or more sentences at the end of the passages. Blanks for the students' biodata information 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 five cloze tests (see Appendix A for example directions and 12 items taken from TEST A in this study).

It is important to note that randomization was used throughout the passage selection process and that semi-random selection (every 12th word) was used to define the blanks. Based on sampling theory, the remainder of his study depends on the notion that the five 35-item cloze tests constitue a collection of 150 items representative of all items that could have been created from the books in the Leon Country Library.

2.3 *Procedures*

With these cloze tests in hand, data gathering began in cooperation with six EFL teachers at post-secondary institutions in Japan (see Note 3). The five tests were duplicated and randomly stacked such that all students had an equal chance of getting any one of the five passages. They were then sent to Japan, where the tests were distributed by the teachers to their students and the directions were read and clarified as necessary. The students were allowed 25 minutes to complete the 30 items. The cloze tests were administered under comfortable conditions familiar to all of the students. The 25-minute time limit proved sufficient for all students. The tests were collected

and then sent to one of the teachers for consolidation and shipment back to Hawaii.

Scoring was done entirely by the exact-answer scoring method, which means that only the word found in the original passage was counted as correct. This was justified because the results were not being reported to the students and because there is typically a very high correlation between exact-answer scoring results and the other seemingly fairer scoring procedure (for more on this, see Alderson, 1979, and Brown, 1980). Perhaps most crucially, the exact-answer scoring method was adopted here because it was considered essential that a correct answer be interpretable as a single possible choice.

2.4 Analysis

To understand the central analyses in this study, it is important to understand that it is dealing with a number of different variables. Brown (1988a, p. 7) defines a *variable* as "something that may vary, or differ." For instance, the first variable of interest in this study is item difficulty (ITEM DIF), which is defined as follows:

1. ITEM DIF — the proportion of students who correctly answered each of the 150 cloze test items.

In this case, it 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. Thus if 18 out of 36 students answered an item correctly, the item difficulty for that item would be .50 (18 ÷ 36 = .50).

ITEM DIF is considered a variable because it gives an estimate of how difficult (or easy) the students found each item to be, and this is something that may vary, or differ, from item to item. ITEM DIF is considered the *dependent* variable in this study because it was measured "to determine what effect, if any, the other types of variables may have on it" (Brown, 1988a, p. 10).

All of the other variables in this study (called *independent* variables) were chosen because of their potential relationships with the ITEM DIF dependent variable. These relationships were explored using Pearson product-moment correlations and multi-regression analyses, which were conducted between various independent variables (and combinations of these variables) and the dependent variable. The independent variables used here were selected because they are item characteristics which are quantifiable and have the potential to explain variation in item difficulties. In other words, these are variables which might help to explain what makes individ-

ual cloze items easy or difficult. The independent variables (which are variables 2-14 in this study) are defined as follows:

2. ITEM DIS — Item discrimination (Item difficulty for the upper third of students on the whole test scores minus the item difficulty for the lower third on the whole test scores)
3. CON/FUNC — Dichotomous variable indicating 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.
4. PAS FREQ — The frequency with which the same word as the correct answer appeared elsewhere in the passage
5. TOT FREQ — The frequency with which the same word as the correct answer appeared elsewhere in all five passages. This is assumed to be a rough estimate of the frequency of the word in the library as a whole.
6. LOG PFRQ — A log transformation (to linearize relationship with ITEM DIF) of PAS FREQ above
7. LOG TFRQ — A log transformation (to linearize relationship with ITEM DIF) of TOT FREQ above
8. SYLL/T-U — The number of syllables in the T-unit in which the blank was found (see Hunt, 1965; Gaies, 1980)
9. SYLL/SEN — The number of syllables in the sentences in which the blank was found
10. WRDS/T-U — The number of words in the T-unit in which the blank was found
11. WRDS/SEN — The number of words in the sentence in which the blank was found
12. CHRS/WRD — The number of characters in the word which was the correct answer
13. READLTY1 — Flesch-Kincaid readability index for the passage in which the blank was found (as described in Klare, 1984)
14. READLTY2 — Fry readability index for the passage in which the blank was found (see Fry, 1985)

All but three of the independent variables should be clear as described above. The three exceptions are clarified as follows:

3. The CON/FUNC variable is different from all of the other variables in that it is dichotomous rather than continuous. In other words, a word is either a content word or a function word, one or the other. This is unlike the other variables which are all on interval scales from 0 to 1, 1 to 124, etc. The importance of this fact is that this variable, unlike all of the others, was necessarily analyzed using the point-biserial correction coefficient rather that the Pearson product-moment co-effficient.
6. The LOG PFRQ is a log transformation of the PAS FREQ defined just above it in the table. The log transformations here and in (7) below were necessitated by the fact that both of these variables were found to form a curvelinear relationship when plotted against the item difficulty values. However, a linear relationship could be obtained with this simple transformation and, as you will see in Table 4, the transformed data formed a stronger correlation.
7. Similarly, LOG TFRQ is a log transformation of the TOT FREQ above it.

All of the analyses were peformed using the *Quattro* spreadsheet program (Borland, 1987) on an IBMAT computer. The multiple-regression algorithms were cross-verified by recalculating them using *Lotus 1-2-3* (Lotus, 1985). There were only minor differences found in the results of the two sets of analyses.

3. Results

Description of the results of this study begins in Table 1, which shows the overall cloze test characteristics in terms of the following descriptive statistics: the number of subjects who took the particular cloze (N), the number of items on it (k), as well as the mean ($\bar{X}$), standard deviation (S), Kuder-Richardson formula 20 (K-R20) and standard error of measurement (SEM).

Table 1: Cloze Test Characteristics

CLOZE	N	k	$\bar{X}$	S	K-R20	SEM
TEST A	35	30	12.06	3.41	.68	1.93
TEST B	33	30	7.52	2.65	.53	2.65
TEST C	37	30	9.68	3.72	.73	1.94
TEST D	38	30	7.24	2.97	.62	1.82
TEST E	36	30	4.58	2.39	.62	1.49
TOTAL (A-E)	179	150	8.20	–	(.90)	–

Notice that the means of the five cloze tests range from 4.58 to 12.06. Since, based on sampling theory, the five groups of students can be assumed to be about equal in overall proficiency, these differences in means probably indicate that there is considerable variation in the difficulty of these passages. The readability indices reported below in Table 2 reflect differences of similar magnitude. The standard deviations also range considerably, from a low of 2.39 to a high of 3.72.

At first glance, the reliability estimates for the individual cloze tests seem to indicate that these procedures are only moderately reliable. The average of these five reliability estimates is only .636. However, since the results are based on the much longer 150-item five cloze test results, the Spearman-Brown formula was applied to adjust for the difference in length between each of the 30-item tests and the 150-item total. Based on the average reliability (.636), the adjusted reliability estimate is .8973, or about .90, which is interpreted here as a rough estimate of the reliability of the whole set of tests taken together. The magnitude of this reliability estimate is encouraging because logically the results of this study can be no more reliable than the tests upon which they are based.

Table 2: Descriptive Statistics for ITEM DIF (Dependent Variable)

CLOZE	k	$\bar{X}_{ID}$	S_{ID}	MIN	MAX	READLTY1	READLTY2
TEST A	30	.4019	.3349	0	.97	4.63	6.70
TEST B	30	.2505	.2773	0	.85	11.21	13.90
TEST C	30	.3225	.2942	0	.87	9.33	11.50
TEST D	30	.2413	.2645	0	.90	7.49	10.20
TEST E	30	.1529	.2331	0	.83	9.46	12.00
TOTAL (A-E)	150	.2738	.2913	0	.97	8.04	10.86

Table 2 focuses on the statistical characteristics related to the dependent variable, item difficulty. For each test and for all tests combined, it shows the number of items (k), the mean item difficulty ($\bar{X}_{ID}$), the standard deviation of the item difficulty indices (S_{ID}), the minimum (MIN) and maximum (MAX) item difficulties that were found on each of the cloze tests, as well as the Flesch-Kincaid readability index for the passage (READLTY1) and the Fry readability index (READLTY2). Notice that the cloze tests, on the whole, were fairly difficult for the students with 15.29 to 40.19 percent of the students (i.e., $\bar{X}_{ID}$ of .1529 to .4019) filling in the blanks correctly on average. This is probably due in large part to the use of the exact-answer scoring method. Had an acceptable-answer scoring scheme been used instead, the mean item difficulties would no doubt have been considerably

higher (e.g., in Brown, 1980, the mean score for acceptable-answer scoring turned out to be 71 percent higher than the mean for exact-answer scoring).

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 DIF ranges from as low as .00 to as high as .97, and has standard deviations (S_{ID}), which are all reasonably large. Since the purpose of this study is to investigate what causes such items to be difficult or easy, the wide variety of item difficulties (.00 to .97) was felt to be desirable. However, one possible problem appears in this table. Notice that the S_{ID} for each test is as large or larger than the $\bar{X}_{ID}$. This is a potential problem in that such a situation indicates that the distribution of item difficulty indices may be skewed, that is not normally distributed. Since the correlation coefficients calculated elsewhere in this study assume normal distributions on the variables involved, this skewing must be included in the interpretation of results.

Another pattern that once again emerges in Table 2 is that the passages vary considerably in overall difficulty. This is of course indicated by the $\bar{X}_{ID}$ discussed above, but also by the two readability indices. The Flesch-Kincaid index ranges from a low of grade 4.63 for Test A to a high of 11.21 for Test B. The Fry scale appears to be exactly parallel, but several grades higher for each test, with a low of 6.7 and a high of 13.9.

Table 3: Descriptive Statistics for Independent Variables

VARIABLE	k	$\bar{X}$	S	MIN	MAX
2. ITEM DIS	150	0.20	0.22	–0.31	0.83
3. CON/FUNC	150	1.63	0.48	1.00	2.00
4. PAS FREQ	150	7.37	9.67	1.00	44.00
5. TOT FREQ	150	23.04	34.04	1.00	124.00
6. LOG PFRQ	150	0.56	0.51	0.00	1.64
7. LOG TFRQ	150	0.87	0.69	0.00	2.09
8. SYLL/T-U	150	28.43	14.02	4.00	67.00
9. SYLL/SEN	150	31.41	13.60	4.00	67.00
10. WRDS/T-U	150	19.01	9.17	3.00	41.00
11. WRDS/SEN	150	21.37	8.62	4.00	41.00
12. CHRS/WRD	150	4.26	2.16	1.00	11.00
13. READLTY1	150	8.42	2.24	4.63	11.20
14. READLTY2	150	10.86	2.40	6.70	13.90

Similar descriptive statistics (k, $\bar{X}$, S, MIN and MAX) are given in Table 3 for each of the independent variables. The first column labels the variable being described. For ease of interpretation, these independent variables are

numbered and presented in the same order as their definitions in the *Analysis* section. Note, in the second column (k), that the variables are being described as they occurred across all 150 items in the five cloze tests. These descriptive statistics are presented here to help the reader cloze tests. These descriptive statistics are presented here to help the reader interpret the correlational results that follow.

Table 4: Correlation Matrix for All Variables

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. ITEM DIF	1.00													
2. ITEM DIS	.32	1.00												
3. CON/FUNC	-.19	-.14	1.00											
4. PAS FREQ	.38	.27	-.41	1.00										
5. TOT FREQ	.27	.18	-.62	.85	1.00									
6. LOG PFRQ	.51	.32	-.50	.87	.79	1.00								
7. LOG TFRQ	.45	.31	-.66	.76	.84	.91	1.00							
8. SYLL/T-U	-.19	-.29	-.08	-.13	-.07	-.16	-.11	1.00						
9. SYLL/SEN	-.17	-.18	-.11	-.02	.01	-.05	-.01	.86	1.00					
10. WRDS/T-U	-.15	-.27	-.14	-.09	-.03	-.12	-.06	.94	.81	1.00				
11. WRDS/SEN	-.14	-.15	-.14	.00	.04	-.03	.01	.84	.96	.84	1.00			
12. CHRS/WRD	-.45	-.29	.50	-.44	-.48	-.62	-.71	.02	-.05	-.05	-.13	1.00		
13. READLTY1	-.19	-.08	-.06	.04	.03	-.11	-.11	.41	.47	.35	.44	.09	1.00	
14. READLTY2	-.20	-.09	-.05	.02	.02	-.13	-.12	.42	.48	.36	.44	.10	.99	1.00

*CRITICAL VALUE (ONE-TAILED, $p < .05$, df = 148) = +/- .13487
df = 148

Table 4 shows the simple correlations between all variables in this study. Notice (below the table) that the critical value is given for the conditions of this study (i.e., one-tailed; df=148; $p < .05$). In all cases, directionality was predictable based on common sense so only one-tailed (directional) decisions were made. This footnote indicates that all correlation coefficients higher in magnitiude than +.13487, or lower than –.13487 occurred for other than chance reasons (with 95 percent probability). Put another way, any correlation coefficient larger in magnitude (either positive or negative) than .13487 had only a 5 percent probability of occurring by chance alone. (See Brown, 1988a, for more on interpreting these statistics.)

The single strongest relationship in Table 4 is between the two readability indices (variables 13 and 14) which correlate at .99. This makes sense upon reexamination of Table 2 because, though they appear to disagree by about two grade levels in their assessment of the readability of the passages, they rank the passages in exactly the same order. Likewise, the relatively high correlations among the two frequency counts and their log transformations (variables 4, 5, 6, and 7) are obvious at a common sense level. Other

correlations that are both high and logical are those which occur between the counts of words or syllables per sentence or T-unit (variables 8, 9, 10, and 11). Those same counts (8-11) also appear to be moderately correlated with the passage readability indices (13 and 14) which are, of course, based in part on such counts. None of these relationships are counter-intuitive in the context of this study.

Perhaps more interesting is the realtionship between characters per word (12) and variables 1 through 7. This series of moderate negative correlations indicates relationships between the length of the word required to fill in a blank and the seven other factors. In other words, the shorter a word (12), the more likely the item is to be easy (1), to disciminate well between students (2), to be a function word (3), as well as to be found frequently in the passage (4), total passages (5), and two frequency count log transformations (6 and 7). This simple letter count appears to be a better predictor of other characteristics than was expected at the beginning of this study. However, in retrospect, these relationships also make sense.

Since the focus of this analysis was on the degree to which each of the independent variables predict item difficulty, the correlation coefficients of most interest are those found in the second column (labeled 1). Notice that all of these correlation coefficients, whether negative or positive, were significant (i.e., higher than the critical value of .13487). In other words, all of these independent variables appear to be related to the proportion correct (ITEM DIF) on each of the cloze test 150 items. This may not at first seem particularly remarkable until you consider that the independent variables, which are all simple countables in the text of five passages, are each predicting to some degree the performance of living, breathing students on those items, that is, the item difficulty estimates. Clearly, some of the independent variables are more highly related to ITEM DIF than others (e.g., 2, 4, 6, 7, and 12). This observation led to investigating the degree to which various combinations of these variables might be most highly related to ITEM DIF.

Table 5: Multiple Regression Analyses (best fits)

DEPENDENT VARIABLE		INDEPENDENT VARIABLES	MR	MR^2
ITEM DIF	=	LOG PFRQ	.51	.26
ITEM DIF	=	LOG PFRQ + CHRS/WRD	.53	.28
ITEM DIF	=	LOG PFRQ + CHRS/WRD + SYLL/SEN	.56	.31
ITEM DIF	=	LOG PFRQ + CHRS/WRD + SYLL/SEN + CON/FUNC	.57	.32

Various mixtures of independent variables were analyzed to determine which set would best predict the ITEM DIF dependent variable. The most productive multiple-regression analyses for this study are shown in Table 5. Notice that the combination of LOG PFRQ + CHRS/WRD + SYLL/SEN + CON/FUNC taken together produce a multiple-correlation (MR) of .57 and a corresponding MR^2 of .32. This means that this combination of simple countable independent variables taken together predict about 32 percent of the variation in the performance of Japanese students on these items. Again, this may not initially appear to be particularly interesting; there is still 68 percent of the variation in ITEM DIF that remains unexplained. However, if you consider that these independent variables are based on different simple counts related to the word in each cloze blank (i.e., the frequency of occurrences of a word in the passage, the number of characters in the word, the number of syllables in the sentence in which it is found and whether it is a content or function word), it is remarkable that they predict 32 percent of the variation in the difficulty that Japanese students have in filling those same blanks.

4. Discussion

The discussion will now return to the original three research questions (which serve as subheadings) and then touch on the implications of these findings especially as they relate to future research along the same lines.

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

It appears from the results above that these cloze tests do function well for observing at least the variables explored in this study. As with any tool for observing language behavior, it is important to consider the degree to which these cloze tests are reliable and valid for the stated purposes before investing too much faith in any results obtained with them. That is why this research question was placed first. In a sense, a positive answer to this research question is prerequisite to answering either of the other two.

In terms of reliability, the cloze passages used here appear to be reasonably consistent. This is indicated by the Spearman-Brown estimate of .90 for the internal consistency reliability of the five cloze tests taken together. However, it is important to recognize that the reliability indices for the individual passages were considerably lower, ranging from .53 to .73 with an average of .636. Since the analyses here are based on the total sample of

cloze 150 items, the .90 overall estimate will be taken as the more appropriate estimate.

Nevertheless, the lower passage reliabilities bear some consideation. These modest reliability estimates may be due in part to the relatively homogeneous nature of the samples. The samples may be fairly uniform because they are made up of students at roughly the same level of study who, by definition, have all studied many years of English. Thus the range of possible scores may be restricted as reflected in the relatively low standard deviations which are in turn directly associated with reliability estimates. (See Brown, 1984, for more on the relationship between the standard deviation and reliability estimates.)

The validity of these five cloze passages when used for the purposes of this study can be argued in simple logical terms without recourse to elaborate statistics. Consider the fact that these cloze tests were developed from randomly selected passages and that the items were selected on a semi-random basis (i.e., every *nth* word deletion). Based on sampling theory, it is arguable that the passages are a representative sample of the language contained in the books in that library and, in turn, that the items provide a representative sample of the language contained in the passages. Since the validity of a measure may be defined as the degree to which it is measuring what it claims to be measuring, it seems safe to claim a high degree of content validity for these cloze passage items because they can be said to be a representative sample of the universe of all possible items (after Cronbach, 1970). Such an interpretation presupposes that the universe is defined as that written language which is found in an American public library as it is tapped by single word blanks.

Based on all of the above, it is with some confidence that the cloze tests in this study are viewed as reliable and valid for the purposes of gathering data on variables that are related to the item difficulty levels found within them. In addition, it is felt that the test development methodology used in this study is sufficiently effective to continue its use in any large-scale study that might follow.

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

The results above also indicate that a number of relatively simple and countable variables are related to the item difficulty (i.e., the degree to which individual cloze items are difficult or easy). Most striking are the magnitudes of the correlation coefficients between ITEM DIF and those counts

associated with the frequency of the word in its passage and in the five passages taken together. Also striking is the degree of relationship between ITEM DIF and the word length in terms of characters per word. Somewhat less meaningful but also interesting, however, is the fact that all of the variables identified as independent variables that might possibly be related to item difficulty were indeed correlated with it either negatively or positively at the $p < .05$ signficance level (i.e., there is only a 5 percent probability that these correlation coefficients occurred by chance alone). (Note that this is true even though some of the distributions were skewed [which would tend to depress any resulting correlation coefficients].) Thus none of these variables should be casually dismissed because they all appear to represent non-chance relationships.

After completing this study, it became clear that there are a number of additional variables that should be considered in any other research that is done along the same lines. For instance, at the clausal level, the distinction between words of Latinate or Germanic origin might be related to item difficulty. At a more global level, it might prove profitable to examine the item difficulties in terms of other readability scales like the Lorge (1959) scale or word frequency lists like those found in Thorndike and Lorge (1959). Perhaps cohesive devices should even be brought into the model.

Nevertheless, the results as they stand are sufficiently encouraging in terms of the number and strength of the observed relationships to encourage the expansion of this study into a large-scale research project.

4.3 What combination of variables best predicts item difficulty in a cloze environment?

The single best combination of variables for predicting item difficulty (see Table 5) was the combination of LOG PFRQ + CHRS/WRD + SYLL/SEN + CON/FUNC, which had a multiple correlation of .57 with the dependent variable. Related to this finding, an apparently high degree of multicollinearity was observed. In simple terms, this means that these variables appear to be interrelated among themselves to such a degree that entering one of them into a multiple-regression model as the first predictor variable leaves little unique variance for other variables to add to the prediction.

For example, consider Table 5 where the LOG PFRQ is entered first into the multiple-regression prediction. LOG PFRQ seems appropriate as a first variable because it is the variable most highly correlated with the ITEM DIF (see Table 4). Yet once the variance due to LOG PFRQ is accounted for, CHRS/WRD (which is also fairly highly related in a negative direction to ITEM DIF)

only adds .02 to the multiple correlation (MR). A quick look at the correlation of –.62 between LOG PFRQ and CHRS/WRD helps to understand this effect. In short, these two variables seem to be interrelated to a magnitude that limits the degree to which either of them can explain variance in the dependent variable that is not also explained by the other. This appears to be true for many of the other variables as well. The degree of multicollinearity will no doubt be a factor that must be considered in any future research along these lines.

5. Conclusion

One of the distinct advantages of this study over much of the other research on cloze procedure is that it is focused on Japanese students, and Japanese students only. Other studies, primarily based on ESL institutions at universities and colleges in the United States and Great Britain, have commonly included a variety of languages mixed together. As such, the results of such studies are difficult to interpret because they cannot be generalized beyond the situation in which the data were gathered. While the sample here cannot be said to be a random sample of all Japanese post-secondary students, it is at very least homogeneous with regard to nationality, language background, and educational level of the students. The results here pertain to Japanese students, and Japanese students only.

In general terms, the results here indicate that, for Japanese post-secondary students, a wide variety of variables were significantly correlated with the item difficulty values on the five cloze tests investigated. These variables fall into categories that might prove useful in looking for patterns in the results. Table 6 summarizes the correlation coefficients (with ITEM DIF), but they are reorganized so that those variables which operate primarily at the word level are grouped together, while others which would more accurately be classified as T-unit or sentence level variables are grouped separately. Still others are grouped under lexical frequencies, and the remaining variables appear to be most appropriately classified as passage level variables. Notice that the highest correlation coefficients are those for one of the word level variables and for the lexical frequency counts (especially when logarithmically transformed). This suggests that, for Japanese students, lexical factors are more highly related to performance on individual items than the other factors. However, this does not mean that the other variables make no significant contribution to the variation in item difficulty estimates.

Table 6: Correlations with Item Difficulty (grouped by variable type)

LEVEL VARIABLE	CORR w/ IF	*LEVEL* VARIABLE	CORR w/ IF
WORD LEVEL		***LEXICAL FREQUENCIES***	
CHRS/WRD	−.45	PAS FREQ	.38
CON/FUNC	−.19	TOT FREQ	.27
		LOG PFRQ	.51
		LOG TFRQ	.45
T-UNIT/SENTENCE LEVEL		***PASSAGE LEVEL***	
SYLL/T-U	−.19	ITEM DIS	.32
SYLL/SEN	−.17	READLTY1	−.19
WRDS/T-U	−.15	READLTY2	−.20
WRDS/SEN	−.14		

It would be impossible to argue on the basis of these results that cloze tests are primarily measuring at the clause or sentence level, or for that matter, that cloze tests focus predominately on intersentential elements. As proposed at the end of the **Introduction** section, the evidence here suggests that, at least for Japanese students, performance on cloze test items is related to a wide variety of factors. True, it is most highly related to lexical frequency factors, but it is also significantly correlated with a number of factors at the word level, T-unit/sentence level, and passage level. Thus cloze tests appear to be assessing at a number of levels simultaneously, and of course there are a large number of potential interactions among all of the variables investigated here. In addition, there are no doubt many linguistic variables (particularly discourse and pragmatic variables) that have not yet been isolated and studied.

5.1 Implications and Future Directions

It seems clear that the overall results of this study are encouraging enough to continue pursuing this research direction. Further research should generally 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 test items. Such research would also allow for investigation of the statistical properties of a large numbers of tests all administered to comparable groups under similar conditions.

The present study used five passages for a total of 150 items administered to 179 students. Research is presently being conducted that will use many more passages and many, many more items with a much larger sample of

students. To that end, a study has been designed to include 50 randomly selected passages with 30 items each for a total of 1500 items (50 tests x 30 items = 1500 items). Since it is also desirable for statistical reasons that at least 30 students be randomly assigned to take each test, a total of at least 1500 subjects will participate (30 students x 50 tests = 1500 students).

Based on the experience gained in conducting the present study, a number of changes will be made in the research design. The first and most important of these is that latent trait analysis will be built into the design. Each of the 50 cloze tests will include an additional ten-item cloze passage which is exactly the same across all 50 of the tests. The use of latent trait analysis based on this ten-item "anchor cloze" will help control sampling error. Such control will make the assumption of equality across the 50 samples even more tenable. The 50 passages have already been randomly selected and modified into cloze tests.

As is often the case, more questions were raised than settled in the process of doing the present research project, so the following general questions are offered as indications of some of the directions in which the future research might usefully head:

1. Are cloze tests reliable and valid tools for gathering data when 50 randomly selected passages are used? What differences occur among passages?
2. Do the test statistics for 50 randomly selected cloze tests vary as would be predicted by classical test theory?
3. To what degree do latent trait sample free estimates of item difficulty compare to classical theory estimates?
4. Which linguistic variables are significantly and meaningfully related to item difficulty when all 1500 cloze test items are analyzed as a set?
5. What combinations of variables best predict item difficulty in these 1500 items?
6. What combinations of variables best predict the overall passage readability levels?
7. What differences and similarities would occur if this large-scale study were replicated with students from other countries and language backgrounds?
8. What hierarchies of difficulty are found for any of the linguistic variables (taken separately or combined) that would have implications for second language acquisition research?

Appendix A

EXAMPLE CLOZE PASSAGE (TEST A)

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."

Notes

1. This paper is a much revised version of a study presented at the 1988 Second Language Research Forum in Honolulu, Hawaii. The author would like to thank Keiichi Orikasa for his fine Japanese translation of the abstract (earlier in this volume). Mr. Orikasa is a recent graduate of the Department of ESL at the University of Hawaii at Manoa. He teaches at Keio Senior High School in Yokohama.
2. The author would like to thank Gary Buck for helping to distribute these tests to sites in Japan. Thanks are also given to those colleagues who helped by administering tests at Baika Junior College, Kobe Yamato Junior College, Kobe University, and Wakayama University. Unfortunately, their names are presently unknown to me, but their efforts are nevertheless appreciated.
3. 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.

References

Alderson, J. C. (1978). *A study of the cloze procedure with native and non-native speakers of English*. 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, DC: 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.

Borland. (1987). *Quattro: The professional spreadsheet*. Scotts Valley, CA: Borland International.

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 DE: International Reading Associates.

Bormuth, J. R. (1967). Comparable cloze and multiple-choice comprehension test 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: 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.

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.

Cohen, A. D. (1980). *Testing language ability in the classroom.* Rowley, MA: Newbury House.

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 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.* New York: Harper & Row.

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

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, DE: International Reading Associates.

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, MA: 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*, 409-438.

Klare, G. P. (1984). Readability. In P. D. Pearson (Ed.), *Handbook of reading research* (pp. 681-744). New York: Longman.

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

Lotus. (1985). *1-2-3.* Cambridge, MA: Lotus Development.

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

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.

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., & Inal, N. (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.

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.

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.

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