WHAT MAKES A CLOZE ITEM DIFFICULT?

JAMES DEAN BROWN
University of Hawai'i at Manoa

This pilot study explores the link between linguistic characteristics of cloze test items and individual item difficulty estimates. Five reading passages were randomly selected from a public library, developed into 30-item cloze tests and randomly administered to 179 Japanese EFL students. Each of the 150 resulting items was analyzed for item facility and for various linguistic characteristics. Multiple-regression analysis indicates that linguistic characteristics, in various combinations, can account for large proportions of the variation in cloze item facility. These results are discussed in terms of their implications for language testing research and plans for future research on a larger scale.

Cloze procedure initially surfaced when Taylor (1953) investigated its effectiveness as a tool for measuring the readability of materials for American school children. Cloze research next focused on its utility as a measure of native-speaker reading proficiency (Ruddell 1964; Bormuth 1965; Gallant 1965; Crawford 1970). In the sixties, studies also began on cloze as a measure of overall ESL proficiency, and dozens 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 careful review of the literature on cloze as a measure of overall ESL proficiency reveals that the results are far from consistent. For instance in Brown (1984), it was noted that the relative reliability and validity of cloze tests have varied considerably within and among the investigations.

Reliability estimates 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, 1988; Bachman 1985). In other words, there are a variety of results indicating that different cloze tests in different situations may vary from exceptionally

This paper is a much revised version of a study presented at the 1988 Second Langauge Research Forum in Honolulu, Hawai'i, March 4, 1988.

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 between 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, 1988; 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 ESL 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) deletion patterns (e.g., every 5th word, every 7th word, etc.), 3) blank lengths, 4) passage difficulties, 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, i.e., 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 essentially 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 pilot 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 further explore just what it is that cloze items test, it is necessary to make every effort 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. The goal of this pilot project is to investigate the feasibility of letting the data guide the researcher (rather than the other way around) in examining the relationships among items, as well as between items and the text as a whole. Because this is just a first step in trying to discover what linguistic elements cloze items tap, the resulting 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 facility levels?
2. What variables are significantly and meaningfully related to item facility in a cloze environment?
3. What combination of variables best predicts item facility in a cloze environment?

If the results of this pilot 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 study, the alpha level for all statistical decisions was set at $\alpha < .05$.

METHOD

Subjects

This study attempts to control variables that literally remains 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 in this study were at the Junior college or university level in Japan (N = 179), were Japanese nationals and had Japanese as their first language. In addition, all of the students were enrolled in EFL courses in their respective institutions. They ranged in age from 18 to 23 and included 98 females and 81 males. The five cloze tests used here (see Materials below) were administered to random samples of these students such that the performances of the resulting groups can reasonably be assumed to be approximately equal across the five tests.

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. 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.

Each of these passages was then modified so that every 12th word was deleted and replaced by a blank for a total of thirty items. Two sentences were left unmutilated at the beginning of the 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 each passage 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.

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 this study will depend on the notion that the five, thirty-item cloze tests constitute a collection of 150 items representative of all items that could have been created from the books in the Leon County Library.

Procedures

With these cloze tests in hand, data gathering began with the cooperation of six EFL teachers at universities in Japan. 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 thirty 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 Hawai'i.

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 procedures (see Alderson 1979 and Brown 1980 for more on this). More importantly, exact-answer scoring was adopted here because it was considered essential that a correct answer be interpretable as a single possible choice.

Analyses

To understand the central analyses in this study, it is important to first recognize that the DEPENDENT VARIABLE is item facility. Item facility (IF) is defined here as the percentage 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 facility for that item would be .50 (18 ÷ 36 = .50). The IF is the focus of analysis here because it gives an estimate of how difficult (or easy) the students found each item to be.

Table 1: Independent Variable Definitions

VARIABLE LABEL	DEFINITION
ITEM DIS	Item discrimination (IF for upper third of students minus IF for lower third)
CON/FUNC	Dichotomous variable indicating whether the correct answer for a blank was a content word or a function word
PAS FREQ	The frequency with which the same word as the correct answer appeared elsewhere in the passage
TOT FREQ	The frequency with which the same word as the correct answer appeared elsewhere in all five passages
LOG PFRQ	A log transformation (to linearize relationship with IF) of PAS FREQ above
LOG TFRQ	A log transformation (to linearize relationship with IF) of TOT FREQ above
SYLL/T-U	The number of syllables in the T-unit in which the blank was found (see Hunt 1965; Gaies 1980)
SYLL/SEN	The number of syllables in the sentences in which the blank was found
WRDS/T-U	The number of words in the T-unit in which the blank was found
WRDS/SEN	The number of words in the sentence in which the blank was found
CHRS/WRD	The number of characters in the word which was the correct answer
READLTY1	Flesch-Kincaid readability index for the passage in which the blank was found (as described in Klare 1984)
READLTY2	Fry readability index for the passage in which the blank was found (see Fry 1985)

All of the other analyses in this study were used to examine the relationships between various independent variables and the IF dependent variable. Thus Pearson product-moment correlations and multiple-regression analyses were conducted between various independent variables (and combinations of these variables) and the dependent variable. The independent variables used here were chosen because they are item characteristics which are quantifiable and logically have the potential to explain variations in item facility. In other words, these are variables which might help to explain what makes individual cloze items easy or difficult. The independent variables are defined in Table 1. All but three of these variables should be clear as shown there. Further explanation of the other three variables follows:

1. 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 correlation coefficient rather that the Pearson product-moment coefficient.
2. The LOG PFRQ is a log transformation of the PAS FREQ defined just above it in the table. The log transformations here were necessitated by the fact that both of these variables were found to form a nonlinear relationship when plotted against the item facility values. However, a linear relationship could be obtained with this simple transformation and, as you will see in Table 5, the transformed data formed a stronger relationship.
3. Similarly, LOG TFRQ is a log transformation of the TOT FREQ above it.

All of the analyses were performed using the Quattro spreadsheet program (Borland 1987) on an IBM AT 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 algorithms.

RESULTS

Description of the results of this study begins in Table 2, 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 2: 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 3 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 3: Descriptive Statistics for Item Facility (Dependent Variable)

CLOZE	k	$\bar{X}_{IF}$	S_{IF}	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 3 focuses on the statistical characteristics related to the dependent variable, item facility. For each test and for all tests combined, it shows the number of items (k), the mean item facility ($\bar{X}_{IF}$), the standard deviation of the item facility indices (S_{IF}), the minimum (MIN) and maximum (MAX) IFs 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 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 IFs would no doubt have been considerably higher.

More importantly for this type of project, the tests appear to have generated a wide variety of item facility indexes, as indicated by the MIN and MAX columns, which show IFs ranging from as low as .00 to as high as .97, and the standard deviations of these IFs (S_{IF}), which are all large. Since the purpose of this study is to investigate what causes such items to be difficult or easy a wide variety of facility levels will help. However, one possible problem appears in this table.

Notice that the S_{IF} for each test is as large or larger than the $\overline{X}_{IF}$. This is a potential problem in that such a situation indicates that the distribution of IF indices may be skewed, i.e., 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 3 is that the passages vary considerably in overall difficulty. This is of course indicated by the $\overline{X}_{IF}$ discussed above, but also by the two readability indexes. 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.

Similar descriptive statistics (k, $\overline{X}$, S, MIN and MAX) are given in Table 4 for each of the independent variables. The first column labels the variable being described. For ease of interpretation, these independent variables are presented in the same order as their definitions shown in Table 1. 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 interpret the correlational results that follow.

Table 4: Descriptive Statistics for Independent Variables

VARIABLE	k	$\bar{X}$	S	MIN	MAX
ITEM DIS	150	.20	.22	-.31	.83
CON/FUNC	150	1.63	.48	1.00	2.00
PAS FREQ	150	7.37	9.67	1.00	44.00
TOT FREQ	150	23.04	34.04	1.00	124.00
LOG PFRQ	150	.56	.51	.00	1.64
LOG TFRQ	150	.87	.69	.00	2.09
SYLL/T-U	150	28.43	14.02	4.00	67.00
SYLL/SEN	150	31.41	13.60	4.00	67.00
WRDS/T-U	150	19.01	9.17	3.00	41.00
WRDS/SEN	150	21.37	8.62	4.00	41.00
CHRS/WRD	150	4.26	2.16	1.00	11.00
READLTY1	150	8.42	2.24	4.63	11.20
READLTY2	150	10.86	2.40	6.70	13.90

Table 5 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 magnitude 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 has only a five percent probability of occurring by chance alone. [See Brown 1988 for more on interpreting these statistics.]

Table 5: Correlation Matrix for All Variables*

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. IF	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	–.46	– 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	.O2	.02	–.13	–.12	.42	.48	.36	.44	.10	.99	1.00

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

The single strongest relationship in the Table 5 is between the two readability indices (variables 14 and 13) which correlate at .99. This makes sense upon reexamination of Table 3 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 logical 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. All of these relationships are matters of common sense in the context of this study.

Perhaps more interesting is the relationship 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, the more likely the item is to be easy (1), to discriminate 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 make sense.

Since the focus of this study was on the degree to which each of the independent variables predict item facility, 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 degree of difficulty (IF) of the 150 items on these cloze tests. This may not at first seem particularly remarkable until you consider that these 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, i.e., item facility. Clearly, some of the independent variables are more highly related to the IF 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 related to IF.

Table 6: Multiple Regression Analyses (best fits)

DEPENDENT = VARIABLE	INDEPENDENT VARIABLES	MR	MR^2
IF =	LOG PFRQ	.51	.26
IF =	LOG PFRQ + CHRS/WRD	.53	.28
IF =	LOG PFRQ + CHRS/WRD + SYLL/SEN	.56	.31
IF =	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 IF dependent variable. The most productive multiple-regression analyses for this study are shown in Table 6. 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 IF 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 in those same blanks.

DISCUSSION

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

1) Are randomly selected cloze tests reliable and valid tools for gathering data on variables that are related to their own item facility 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 estimate of .90 for the internal consistency reliability of the 5 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 reflection. These modest reliability estimates may be due in part to the relatively homogeneous nature of the samples. The samples are fairly uniform because they are made up of students at the college level who by definition have all studied many years of English. Thus the range of possible scores is restricted by the fact that there are no students at the lower end of the scale of ability levels and perhaps few at the very top of the same scale. This may also be 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 passages which had been randomly selected from a public library and by further selecting items on a semi-random basis (i.e., every *n*th 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 which are a representative sample of the universe of all possible items if that universe is defined as written receptive and productive language as it is found in an American public library and as it is tapped by single word blanks (after Cronbach 1970).

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 facility levels found within them. In addition, it is felt that the test development methodology used in this pilot study is sufficiently effective to continue its use in any large scale study that might follow.

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

The results above also indicate that a number of relatively simple and countable variables are related to the item facility (i.e., the degree to which individual cloze items are difficult or easy). Most striking and meaningful are the relationships between IF and those counts associated with the frequency of the word in its passage and in the five passages taken together. Also striking is the relationship between IF 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 possibly related to item facility were correlated with it either negatively or positively at the $p < .05$ significance level. 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 the lines. For instance, at the clausal level, the distinction between words of Latinate or Germanic origin might be related to item facility. At a more global level, it might prove profitable to examine the IFs 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 expand this pilot study into a full-blown research project.

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

The single most striking finding encountered in trying to fit an effective prediction model to these data was the apparent multicollinearity of these variables. In lay terms, this means that these variables appear to be interrelated to such a degree that entering one of the variables into a regression model as the first predictor variable leaves little unique variance for other variables to add to the prediction. For example, consider Table 6 were the LOG PFRQ is entered first into the multiple-regression prediction. LOG PFRQ seems appropriate as a first variable because it is the single most highly correlated with the dependent variable (see Table 5). Yet once the variance due to LOG PFRQ is accounted for, CHRS/WRD (which is also fairly highly related in a negative direction to IF) 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 variables seem to be interrelated to a degree that limits the degree to which either one can explain variance in the dependent variable that the other one has not already explained. This also appears to be true for many of the other variables. The degree of multicollinearity will no doubt be a factor that must be considered in any larger scale study that follows.

Implications and Future Directions

It seems clear that the overall results of this pilot study are encouraging enough to pursue this research direction on a larger scale. The implications for language testing alone seem important enough to warrant expansion of the research. Such expansion will not only allow further examination of the complex set of linguistic variables (sentential and intersentential) that are contributing to the relative difficulty of performing on cloze items, but also afford an opportunity to examine the statistical properties of a large number of tests all administered to comparable groups under similar conditions.

The present pilot study used 5 passages for a total of 150 items administered to 179 students. The proposed research will necessarily use many more passages and many more items with a much larger sample of students. To that end, a study is presently being designed, which will include 50 randomly selected passages with thirty 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 be needed (30 students x 50 tests = 1500 students).

Based on the experience gained in conducting this pilot 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 of this ten-item "anchor cloze" will help control for sampling errors. Such control will make the assumption of equality across the 50 samples much more tenable.

The 50 passages have already been randomly selected and modified into cloze tests. Negotiations for large scale data gathering have also been initiated with a number of Japanese universities. Once such cooperation is established, it is believed that the research can proceed relatively quickly.

As is often the case, more questions were raised than settled in the process of doing this pilot research project, the following general questions are offered as indications of some of the directions in which the larger study might head:

1. Are randomly selected cloze tests reliable and valid tools for gathering data when 50 passages are used?
2. Are the means, standard deviations, reliability estimates and other descriptive statistics for 50 randomly selected cloze tests normally distributed as would be predicted by classical test theory?
3. How do latent trait sample free estimates of item facility compare classical theory estimates of item facility?
4. What variables (including many others not explored here) are significantly and meaningfully related to item facility when 1500 items are included?
5. What combination of variables best predicts item facility in these 1500 items?
6. What combinations of variables best predict the overall passage difficulties in terms that might help define and construct a new readability index for non-native speakers of English?
7. Would similar results be obtained if this larger scale study were replicated with other nationalities and language groups?
8. Are there any hierarchies of difficulty for any of the linguistic variables taken separately or together that would have implications for second language acquisition research?

Received October 1, 1988.

Author's address for correspondence:

J.D. Brown
Department of English as a Second Language
1890 East–West Road
University of Hawai'i at Manoa
Honolulu, HI 96822

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 Yorio, C.A., Perkins, K. and Schachter, J., editors, *On TESOL '79*, Washington, D.C.:TESOL.

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

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

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 Figurel, J.A., editor, *Reading and inquiry*, Newark, Delaware: International Reading Associates, 283-285.

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: part I – validity. In Oller, J.W. Jr., editor, *Issues in Language Testing*, Rowley, MA: Newbury House.

Brown, J.D. 1983b: A closer look at cloze: part II – reliability. In Oller, J.W. Jr., editor, *Issues in Language Testing*, Rowley, MA: Newbury House.

Brown, J.D. 1983c: An exploration of morpheme-group interactions. In K.M. Bailey, M.H. Long & S. Peck (Eds.) *Second Language Acquisitions Studies*, Rowley, MA: Newbury House, pp. 25-40.

Brown, J.D. 1984: A cloze is a cloze is a cloze? In Handscombe, J., Orem, R.A. and Taylor, B.P., editors, *On TESOL '83*, Washington, D.C.: TESOL.

Brown, J.D. 1986: Cloze procedure: a tool for teaching reading. *TESOL Newsletter*, 20, 5, pp. 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, 1.

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.

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

Conrad, C. 1970: The cloze procedure as a measure of English proficiency (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 (doctoral dissertation, University of California Los Angeles).

Cronbach, L.J. 1970: *Essentials of psychological testing*. New York: Harper and Row, 145-146.

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 Figurel, J.A., editor, *Reading and inquiry*, Newark, Delaware: International Reading Associates, 286-287.

Guilford, J.P. and Fruchter, B. 1979: *Fundamental statistics in psychology and education*, fifth edition, New York: McGraw-Hill.

Hinofotis, F.B. 1980: Cloze as an alternative method of ESL placement and proficiency testing. In Oller, J.W. Jr. and Perkins, K., editors, *Research in language testing*, Rowley, Massachusetts: Newbury House.

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

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

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

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

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

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 Briere, E.J. and Hinofotis, F.B., editors, *Concepts in language testing: some recent studies*, Washington, D.C.: TESOL.

Oller, J.W. Jr. 1972a: Dictation as a test of ESL proficiency. In Allen, H.B. and Campbell, R.N., editors, *Teaching English as a second language: a book of readings*, 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. and N. Inal. 1971: A cloze test of English prepositions. *TESOL Quarterly* 5, 315-326.

Pike, L.W. 1973: *An evaluation of present and alternative item formats for use in the test of English as a foreign language*. Princeton, New Jersey: 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 and D. Porter (Eds.) *Current developments in language testing*. London: Academic Press, 63-74.

Richards, J.C., J. Platt and H. Weber. 1985: *Longman Dictionary of Applied Linguistics*. London: Longman.

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. and 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. and I. Lorge. 1959: *The teacher's word book of 30,000 words.* New York: Columbia Teachers College.