CLOZE TESTS AS INDICATORS OF GENERAL LANGUAGE PROFICIENCY

HENRYK KRZYŻANOWSKI

Adam Mickiewicz University, Poznań

THE EXPERIMENT

1.0. The aim of the experiment was to establish how a cloze test measures the language proficiency of a non-native learner. It consisted of the application of a battery of language tests (of which one was a cloze test) to a group of about one hundred Polish students from Poznań secondary schools. Correlations between the results of the cloze test and the traditional tests were then computed. In the second part of the experiment the same students were retested with four versions of another cloze test. The four versions differed as to the number and the places of deletions. The aim of the above procedure was to find some of the factors influencing test results. An analysis of errors from the two tests served as an insight into the problem of cloze test validity.

1.1. A structure test, a vocabulary test, a dictation and a cloze test formed the battery. Each of the four tests was accorded 50 points (though two of them had actually more items), so as to assure equal weight in the total score. The structure test was a multiple-choice recognition test of 80 items. It was based on three mixed structure tests taken from the collection of language tests by Bloor et al. (1970: vol. II. 29 - 64). The choice of items was made by the writer, who, in general, preserved the structural content of the original tests within its proportions. In a number of cases items were slightly altered for contrastive reasons. The vocabulary test, designed and written especially for the experiment, was a multiple-choice partial-production test of 60 items. It consisted of filling in the blanks in the stem sentences with one of the three words suggested by the initial and final letters.

The dictation was a passage of about 150 words about youth hostels in

Britain. For scoring the text was divided into twenty-five units. The syntactical structure of the text served as a criterion for the division. Each unit was scored twice. For the first scoring only syntactical accuracy was required (word order, main morphological features, etc.). For the second scoring other less important features as well as spelling were taken into consideration.

For both the cloze tests a passage of about 750 words was selected from the reader by Lapidus et al. (1969: 38 - 40). The selection was a narrative about a sea adventure; however, there was no specific marine vocabulary. The first part of the passage served as CLOZE I and the second as CLOZE II (see 4.0.). Both parts were reviewed by native speakers.

In CLOZE I every seventh word was deleted from the text. The first and last sentences were left untouched. In all, there were fifty deletions.

1.2. The tests were administered to pupils of three secondary schools in Poznań: No 1, No 8 and No 4. From now on the three groups will be referred to as Group I, Group II and Group III respectively. The whole population of the test will be referred to as P. There were 31 persons in both Group I and Group II. They were third year students, which means that their level of proficiency could be defined as intermediate. In Group III there were 32 fourth year students, who had undergone a special program of English. The level of this group was then much higher than of Group I and Group II and can be defined as advanced.

All the test subjects were told at the beginning of each series of tests that the results would not serve as a basis of any kind of grading for school purposes.

THE RESULTS OF THE BATTERY OF TESTS

2.0. For each of the tests the following data were calculated: range, mean, standard deviation, reliability (with Kuder-Richardson formula) and the coefficient of test facility (proportion of right answers). The same statistics were computed for the battery as a whole. Computing was made separately for the three groups and for the whole population of the test (P).

2.1. On the whole, the results, as presented in Table 1, show one feature common to all the tests — a very clear discrimination between Groups I and II on the one hand and Group III on the other. All the tests were very easy for Group III and fairly difficult for Groups I and II.

The scoring system of the dictation (see 1.1.) proved, in general, to be very good, though some improvement could be made so as to minimize the effect of spelling mistakes on the final score.

As for the scoring system of the cloze test, it has been shown that the count of any contextually acceptable word is the best method (Oller 1972a). Such a method was used to score CLOZE I and CLOZE II. However, in a

Table 1

Results of the Battery of Tests

TEST	N	GROUP	Range	Mean	Standard Deviation	Reliability	Facility
STRUCT	80	I	39	43,7	11,4	.85	.55
		II	40	48,5	9,5	.79	.60
		III	41	62,3	9,6	.85	.78
		P	52	51,6	12,9	.89	.64
VOCAB	60	I	48	28,7	13,9	.92	.48
		II	43,5	21,9	12	.90	.36
		III	23	51,4	6,0	.79	.85
		P	46	26,6	17,2	.95	.60
DICTAT	50	I	36	21,4	9,9	.87	.43
		II	35	19,3	9,9	.88	.39
		III	30	38,1	7,9	.86	.76
		P	46	26,6	12,7	.92	.53
CLOZE I	50	I	36	22,1	9,5	.86	.44
		II	33	24,4	9,7	.87	.49
		III	22	39,1	5,5	.72	.78
		P	44	28,4	11,4	.90	.57
TOTAL	200	I	125,8	94,7	37	.96	.47
		II	126,3	92,2	33	.95	.46
		III	86,1	159,3	22	.93	.80
		P	163,7	115,9	44	.975	.58

N = number of items

number of cases it was extremely difficult to decide whether or not a word is contextually acceptable.

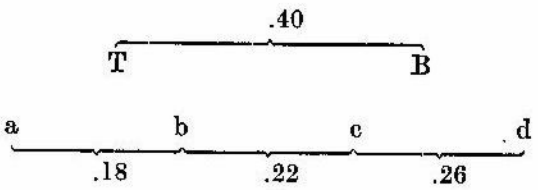
The results of CLOZE I were quite similar to those of traditional tests. In Group I all the statistics resemble very closely those of the other integrative test, i. e., DICTATION. For Group II, CLOZE I was clearly easier than DICTATION. Most assuredly, the cause for this lies in different teaching methods used by the teachers of the two groups[1]. This, however, does not affect reliability, which is very similar in the two groups for both tests. In Group III the rate of test facility and the mean of CLOZE I are similar to those of DICTATION. However, the reliability is much lower, which co-occurs with a lesser standard deviation (5,5 compared with 7,9 of DICTATION) and a narrower range (22 as compared to 30). Such a narrow range means that the scores of the whole group are very much concentrated and that a

[1] In an oral interview the teacher of Group II confirmed that during the two previous years the pupils had had very little practice in listening, whereas Group I had had special listening classes with the use of a tape for at least half a year.

large part of the items contribute very little to discriminate among the test subjects (being either too difficult or, as was the case, too easy). In fact, out of 50 items only 24 discriminated at a rate of .20 or more between top and bottom halves of Group III.

The item analysis made for the whole population showed that the test discriminated sufficiently between the test subjects at all levels of proficiency. This is shown in Fig. 1:

Fig. 1. Discriminatory Power of CLOZE I



T=top half of P
B=bottom half of P
a=top half of T
b=bottom half of T
c=top half of B
d=bottom half of B

The discriminatory power of the test tends to increase with the increase of the test difficulty. This is quite natural as the reliability of CLOZE I is much higher in Groups I and II than in Group III. In other words, this means that this test measured the language proficiency at the intermediate level better than at the advanced level (though it was efficient at both).

2.2. The main tendencies of the particular tests were reinforced in the results of the TOTAL. The distance between Group III and Groups I and II was maintained. For all the groups, the reliability of the TOTAL was visibly higher than the reliability of any of the four tests (which is quite natural, considering the length of the TOTAL). The reliability of the whole battery is very high (.975) and it has a good standard deviation (44, that is, one fifth of N). On the whole, the battery seems to be a good instrument of assessing language proficiency.

CORRELATIONS BETWEEN THE TESTS

3.0. For the estimation of the validity of a test Lado (1961), Valette (1967) and Harris (1969) recommend a method consisting of correlating its results with the results of another test (or, better yet, a battery of tests), of which the validity has been evaluated and confirmed previously. A high coefficient of correlation between the criterion test and the examined test would prove that both measure the same component of a language skill. With the use of this method, Oller (1972) came to the conclusion that a cloze test and a dictation measure basically the same element of a language skill, that is, the ability of anticipating linguistic elements in the process of communication. In another work, Oller and Inal (1971) showed that a cloze test of prepositions should be considered as a grammar test as it correlates the highest with a grammar test. On the other hand, Rand (1972) seems quite sceptical about a real value of such a statistical procedure. He thinks that, in most cases, even if we ascertain a high correlation between two tests, we remain unable to say what specific component of a skill (if the same) the two tests measure. He may be correct in saying that with such a procedure the reliability is unduly taken for validity. It also seems quite possible that a high coefficient of correlation may represent two different elements of a skill that are joined together under the conditions of a given test and with a given group of test subjects, but which will diverge when the conditions are changed.

In spite of these reservations the method applied by Oller was used in this work and coefficients of correlation between particular tests forming the battery were calculated. It is true that the battery had not been validated before with any external criterion. Yet, on the other hand, the analysis showed its very good reliability and at the same time its validity was confirmed by its high correlation with the teachers' classification of the students (see 3.2. below). For that reason we are justified to assume that regularities (if any) found in the set of coefficients may prove useful in the further investigation of the test validity.

3.1. In all cases, the correlation was calculated with Pearson Product-Moment of Correlation. All the coefficients are shown in Table 2. The data presented in Table 2 are summarized in Table 2A (the highest coefficients of correlation for each of the tests in each group are shown in this table).

Table 2A shows clearly that the assumption that a cloze test correlates the best with the dictation (Oller and Conrad 1971) does not apply here except for Group III. It may be that this assumption is true only for more advanced learners, but there is not sufficient material to prove this a fact. The highest correlation of CLOZE I with VOCABULARY could be related to a similar technique of answering the test items (filling-in-the-blanks), but the fact that both DICTATION for Group III and VOCABULARY for the rest of the test subjects had the highest reliability out of the three tests correlated seems to be the most important here.

As for VOCABULARY, it correlated differently in different groups. Except for Group II, the tests which correlated the highest were also the most reliable ones (the same as for CLOZE I). In Group II the most reliable test out of the three was DICTATION. However, both VOCABULARY and DICTATION were visibly more difficult for this group than the other two tests and this may be the reason why they correlated lower.

Table 2

Coefficients of Correlation Between the Tests

Group	Reliability	CLOZE I	VOCAB	STRUCT	DICTAT	TOTAL
I	.86		.89	.88	.86	.945
II	.87	CLOZE I	.90	.79	.75	.95
III	.72		.65	.49	.78	.89
P	.905		.89	.88	.88	.95
I	.92			.85	.89	.98
II	.90		VOCAB	.91	.79	.98
III	.79			.76	.76	.875
P	.95			.85	.90	.97
I	.85				.90	.95
II	.79			STRUCT	.71	.95
III	.85				.78	.91
P	.89				.855	.93
I	.87					.96
II	.88				DICTAT	.88
III	.86					.935
P	.92					.96
I	.96					
II	.95					TOTAL
III	.93					
P	.975					

Table 2A

The Highest Correlations for Particular Tests

Group	CLOZE I	VOCAB	STRUCT	DICTAT	TOTAL
I	VOCAB	CL/DICT	DICTAT	STRUCT	VOCAB
II	VOCAB	STRUCT	VOCAB	VOCAB	VOCAB
III	DICTAT	STR/DICTAT	DICT	CL/STR	DICTAT
P	VOCAB	DICTAT	CLOZE	VOCAB	VOCAB

Note: For particular tests correlations with the TOTAL were excluded — in all cases they were the highest

For Groups I, II and P STRUCTURE was the least reliable test. In Groups II and III the tests which correlated the best with STRUCTURE were those with the highest coefficient of reliability. This was not so for Groups I and P, where DICTATION and CLOZE were not the most reliable of all the tests correlated.

DICTATION correlated differently, maintaining, however, the regularity of correlating the best with the most reliable of the three tests in Groups II, III and P. In Group I DICTATION itself was the most difficult test and this may have been the cause of its not following the pattern.

As for the correlations of particular tests with the TOTAL, they were quite naturally much higher than the correlations with other tests. For all the groups the scores of the TOTAL correlated the highest with the most reliable of the four tests and not with the test which contributed in the greatest proportion to the TOTAL. For instance, in Group II only 20% of the TOTAL was produced by VOCABULARY and about 33% by STRUCTURE; the reliability of the two tests was .90 and .79 respectively; and the coefficient of correlation with the TOTAL was .98 for VOCABULARY and .95 for STRUCTURE.

On the whole, in 16 cases out of 20, tests correlated the best with the most reliable of the three or four (in case of TOTAL) tests. Apparently, there is no other principle which might explain the data presented in Table 2 and 2A. To say, for example, that as CLOZE and VOCABULARY correlate the best, a cloze test at the intermediate level of language proficiency measures basically the knowledge of vocabulary would be inconsistent with other data from the same table as well as with those from Table 1 (see the coefficients of test facility for both CLOZE and VOCABULARY for Groups I and II in Table 1). On the other hand, these data do not exclude the possibility of joining the concepts of correlation between tests and reliability of particular tests with that of test validity. Knowing that the test subjects have received basically the same kind of language instruction, and knowing the objectives of this instruction to be fairly general, we may assume that the language proficiency of an average test subject represents various components of a language skill in the same proportion for each learner. There is no reason to believe that, at this level of language learning, some students would absorb only the vocabulary material of the course and ignore the structural part or vice versa. Besides, it can be observed in classroom practice that the growth of the language proficiency of a learner takes place through gradual absorbing of all the main components of a skill and any differences in this respect are never substantial. (The above applies to components of one language skill and not to language skills as such.) Consequently, once a test has the general validity of a language test, operating within the frame of a given language skill, its reliability becomes much more important than its specific content validity (that of being a vocabulary test, a grammar test, etc.). This last statement well explains the data from Table 2 and the fact that the four tests correlated according to their reliability and not to their content validity. It also justifies a legitimate use of integrative tests, such as cloze or dictation, in language testing, on condition, however, of making them sufficiently reliable (which, in general, will be more difficult in case of these tests than in case of multiple-choice discrete-point tests).

3.2. Before the results of the battery were known, each of the teachers classified his pupils according to their overall language proficiency (all skills being included). The three lists were then correlated with the results of the tests. The rank-difference method was used for the calculation. The coefficients of correlation for all the groups and all the tests are shown in Table 3.

Table 3

Correlation Between Test Results and the Teachers' Classification

Tests	Group I	Group II	Group III
CLOZE I	.87	.88	.77
VOCABULARY	.895	.91	.76
STRUCTURE	.83	.88	.86
DICTATION	.87	.76	.73
TOTAL	.90	.915	.89

These data are very important for the estimation of the general validity of the whole battery of tests. However, we should remember that, though aiming at the same objective, the two sources of data (i. e., teachers' lists and test statistics) are of apparently different character and, as such, are not fully comparable. On the one hand, the teachers' classifications represent a much deeper and more integrated knowledge of learners' proficiency than that offered by the tests, but on the other hand, they are not formalized, being thus less precise and less reliable. Therefore, we shall not analyse each of the coefficients separately, but instead, we shall note main regularities of the data. The most important, for the evaluation of the test validity, is the fact that, for all the groups, the correlation with the TOTAL is the highest. This means that with regard to general validity, each of the tests forming the battery contributes to it, operating on different areas of a linguistic skill. This complementary character of the four tests is best perceived in Group III, where coefficients for particular tests are clearly lower than in Groups I and II (except for STRUCTURE), the coefficient for the TOTAL being nevertheless almost the same. The coefficients for CLOZE I are similar to those for the other tests. On the whole, the coefficients of correlation are high, which proves the general validity of the battery of tests.

INTERNAL FACTORS INFLUENCING THE RESULTS OF A CLOZE TEST

4.0. We shall see now how some changes in the structure of a cloze test influence its results. To examine this problem four versions of the same cloze test were used. As a text, CLOZE II was a continuation of the story told in CLOZE I (see 1.1). Being a whole, the two texts had the advantage of representing the same level of difficulty as to vocabulary, style and syntactical complexity.

In CLOZE IIA every seventh word was deleted according to the same principle as in CLOZE I, that is, in a random way. CLOZE IIB had the same rate of deletions but preference was given to content words (see 4.1. below). The two other versions were CLOZE IIC with every sixth word deleted and CLOZE IID in which the deletion fell on every fifth word. Similarly to CLOZE I and CLOZE IIA, the deletion was made at random with no regard to grammatical categories.

The four versions were given to four groups of learners, each of whom had been tested with the whole battery of tests beforehand. The groups were carefully formed so as to make them fully comparable in regard to language proficiency. To assure this a twofold criterion had been used. The sums of the total scores of the battery and of the cloze scores of all group members were practically equal for each of the four groups. The mean score of CLOZE I was 30,3 in each group. The distribution of pupils with high, medium and low scores was similar in all the groups. There were 21 persons in each group.

4.1. Out of 50 items of CLOZE I, 9 were nouns, 10 were verbs, 8 adverbs, 7 adjectives, 6 pronouns, 2 articles, 3 conjunctions and 2 were particles. For all versions of the cloze tests, nouns, verbs (without auxiliaries), adverbs and adjectives, as they carried most of the narrative load of the text, were called content words, in contrast to all the others called function words. (However inaccurate such a distinction is linguistically, it seems quite valid for the purposes of the present analysis of the functioning of cloze items). There were then 33 content words among 50 items of CLOZE I (one verb was an auxiliary). Apart from this external division of the items, another one, based on an internal criterion, was made. It divided the items according to whether an item could be answered within the limits of the clause from which it was taken, on the basis of the information furnished in this clause (short-range choice category), or whether some knowledge of the text was necessary in answering it (long-range choice category). For instance, of these two items, "21)_____ and his men watched the ship 22) _____ some time", the second belongs to short-range choice category, whereas some external information is needed to answer the first item. In CLOZE I there were 22 items belonging to the long-range choice category, though there were cases where categorizing was difficult.

In CLOZE IIA every seventh word was deleted in a random way. Out of 50 items 26 were content words (11 nouns, 9 verbs, 4 adverbs and 2 adjectives). Among function words there were 10 pronouns, 4 articles, 5 conjunctions, 2 prepositions, 1 particle and 2 auxiliaries. As for the other division, there were 24 long-range items. It should be noted, however, that some words were repeated in the text several times because of the character of the story

(e. g. the word "found"). The items where such a word was deleted were obviously easier than other long-range items.

In CLOZE IIB the deletions were made so as to change the proportion of content words vs. function words. Every seventh word was deleted and out of 50 items 39 were content words (16 nouns, 13 verbs, 4 adverbs, 6 adjectives) and 11 were function words (5 pronouns, 3 articles, 2 prepositions and 1 particle). The proportion of long-range items to short-range items was also different than in the case of the two previous tests; more items belonged to the long-range category (33 items as compared to 22 in CLOZE I, and 24 in CLOZE IIA).

In CLOZE IIC every sixth word was deleted in a random way. There were 53 items, of which 31 were content words (11 nouns, 13 verbs, 3 adverbs and 4 adjectives) and 22 were function words (9 pronouns, 2 articles, 1 conjunction, 5 prepositions, 3 particles and 2 auxiliaries). There were 22 long-range items altogether.

In CLOZE IID every fifth word was deleted, which resulted in 66 items. There were 32 content words (14 nouns, 6 verbs, 4 adverbs and 8 adjectives) and 34 function words (8 pronouns, 9 articles, 3 conjunctions, 11 prepositions, 1 particle and 2 auxiliaries). Out of 66, only 23 items belonged to the long-range category.

The results of the four versions of CLOZE II are shown in Table 4. Some of the statistics of CLOZE IIC and CLOZE IID have been reduced proportionally in order to make them fully comparable with those of 50-item versions. In Table 4 the data reduced in that way are in parentheses.

Table 4

The Results of CLOZE II

Version	N	Range	Mean	Standard Deviation	Reliability	Facility	Discrim Power
IIA	50	31	37,2	8,2	.86	.74	.28
IIB	50	39	30,9	10,3	.89	.62	.41
IIC	53	39 (37)	36,8 (34,7)	10,7	.90	.69	.34
IID	66	47 (35)	46,5 (35,2)	11,5	.90	.70	.28
CLOZE 1	50	40	30,3	11,4	.90	.61	.40

Note: The data of CLOZE I presented here for the sake of comparison are calculated with the exclusion of the test subjects who were not given CLOZE II.

4.2. Before a more detailed analysis of the results of CLOZE II can be made, it should be noted that the results were influenced by two factors. One of them was the effect of practice in taking this kind of test, which the test subjects had gained during the administration of CLOZE I and which certainly helped them to write CLOZE II. The other, perhaps still more important, was the fact

that CLOZE II was a continuation of the story told in CLOZE I. Bearing this in mind, we may assume that, under normal conditions, CLOZE II should prove less difficult than CLOZE I. Only when this assumption has been made are the comparisons between the two tests really valid.

The data presented in Table 4 show that of the four versions of CLOZE II the easiest was IIA and the most difficult IIB (both had the same rate of deletions). The versions where the rate of deletions varied kept the medium position of difficulty between CLOZE IIB and CLOZE IIA and did not differ between each other (the difference of .01 not being pertinent). Compared with CLOZE I all the versions except CLOZE IIB were easier. At this point, remembering the assumption made before, we can safely conclude that CLOZE IIB was the most difficult of all the versions used in the experiment, including CLOZE I. Conversely, CLOZE IIA was probably the easiest of all (as it differed visibly from IIC and IID). This means that the rate of deletions taken separately is not a good criterion of cloze test difficulty. To find such a criterion the quality of the deletions must therefore be taken into consideration. In other words, it is not enough to know how many deletions have been made; information as to which words have been deleted is also needed.

When analysing the quality of the deletions, we shall refer to the division of cloze items made in the previous paragraph. The number of content words in the deletions was similar in CLOZE I and CLOZE IIC, content words were less numerous in CLOZE IIA and IID and the most numerous in CLOZE IIB. As for the distinction between short-range and long-range choice items, the proportions were about the same in CLOZE I, CLOZE IIA and IIC. In CLOZE IIB the number of long-range items was the greatest and in CLOZE IID the smallest. One could conclude that the items with content words deleted and the long-range items were more difficult than the others. However, the item analysis showed that the items with content words were on the same level of difficulty as the items with function words, both in CLOZE I and in the four versions of CLOZE II (though these last data are not fully reliable because of the small number of subjects in each of the groups). The same was also true for long-range and short-range items.

Seemingly then, there is a contradiction in the data collected. On the one hand CLOZE IIB, where the deletions were made so as to have the greatest possible number of content words deleted, was clearly the most difficult of all the versions and on the other, the content words and the function words showed the same coefficient of difficulty in all the tests. The solution to this problem lies in the fact that the amount of information carried by content words is by no means greater than that relying on function words. The more content words are deleted, the more cues to the content are missing and the less the whole text becomes comprehensible. And as the text is made more difficult, all the deletions in it become more difficult independently of the category to

(e. g. the word "found"). The items where such a word was deleted were obviously easier than other long-range items.

In CLOZE IIB the deletions were made so as to change the proportion of content words vs. function words. Every seventh word was deleted and out of 50 items 39 were content words (16 nouns, 13 verbs, 4 adverbs, 6 adjectives) and 11 were function words (5 pronouns, 3 articles, 2 prepositions and 1 particle). The proportion of long-range items to short-range items was also different than in the case of the two previous tests; more items belonged to the long-range category (33 items as compared to 22 in CLOZE I, and 24 in CLOZE IIA).

In CLOZE IIC every sixth word was deleted in a random way. There were 53 items, of which 31 were content words (11 nouns, 13 verbs, 3 adverbs and 4 adjectives) and 22 were function words (9 pronouns, 2 articles, 1 conjunction, 5 prepositions, 3 particles and 2 auxiliaries). There were 22 long-range items altogether.

In CLOZE IID every fifth word was deleted, which resulted in 66 items. There were 32 content words (14 nouns, 6 verbs, 4 adverbs and 8 adjectives) and 34 function words (8 pronouns, 9 articles, 3 conjunctions, 11 prepositions, 1 particle and 2 auxiliaries). Out of 66, only 23 items belonged to the long-range category.

The results of the four versions of CLOZE II are shown in Table 4. Some of the statistics of CLOZE IIC and CLOZE IID have been reduced proportionally in order to make them fully comparable with those of 50-item versions. In Table 4 the data reduced in that way are in parentheses.

Table 4

The Results of CLOZE II

Version	N	Range	Mean	Standard Deviation	Reliability	Facility	Discrim Power
IIA	50	31	37,2	8,2	.86	.74	.28
IIB	50	39	30,9	10,3	.89	.62	.41
IIC	53	39 (37)	36,8 (34,7)	10,7	.90	.69	.34
IID	66	47 (35)	46,5 (35,2)	11,5	.90	.70	.28
CLOZE I	50	40	30,3	11,4	.90	.61	.40

Note: The data of CLOZE I presented here for the sake of comparison are calculated with the exclusion of the test subjects who were not given CLOZE II.

4.2. Before a more detailed analysis of the results of CLOZE II can be made, it should be noted that the results were influenced by two factors. One of them was the effect of practice in taking this kind of test, which the test subjects had gained during the administration of CLOZE I and which certainly helped them to write CLOZE II. The other, perhaps still more important, was the fact

that CLOZE II was a continuation of the story told in CLOZE I. Bearing this in mind, we may assume that, under normal conditions, CLOZE II should prove less difficult than CLOZE I. Only when this assumption has been made are the comparisons between the two tests really valid.

The data presented in Table 4 show that of the four versions of CLOZE II the easiest was IIA and the most difficult IIB (both had the same rate of deletions). The versions where the rate of deletions varied kept the medium position of difficulty between CLOZE IIB and CLOZE IIA and did not differ between each other (the difference of .01 not being pertinent). Compared with CLOZE I all the versions except CLOZE IIB were easier. At this point, remembering the assumption made before, we can safely conclude that CLOZE IIB was the most difficult of all the versions used in the experiment, including CLOZE I. Conversely, CLOZE IIA was probably the easiest of all (as it differed visibly from IIC and IID). This means that the rate of deletions taken separately is not a good criterion of cloze test difficulty. To find such a criterion the quality of the deletions must therefore be taken into consideration. In other words, it is not enough to know how many deletions have been made; information as to which words have been deleted is also needed.

When analysing the quality of the deletions, we shall refer to the division of cloze items made in the previous paragraph. The number of content words in the deletions was similar in CLOZE I and CLOZE IIC, content words were less numerous in CLOZE IIA and IID and the most numerous in CLOZE IIB. As for the distinction between short-range and long-range choice items, the proportions were about the same in CLOZE I, CLOZE IIA and IIC. In CLOZE IIB the number of long-range items was the greatest and in CLOZE IID the smallest. One could conclude that the items with content words deleted and the long-range items were more difficult than the others. However, the item analysis showed that the items with content words were on the same level of difficulty as the items with function words, both in CLOZE I and in the four versions of CLOZE II (though these last data are not fully reliable because of the small number of subjects in each of the groups). The same was also true for long-range and short-range items.

Seemingly then, there is a contradiction in the data collected. On the one hand CLOZE IIB, where the deletions were made so as to have the greatest possible number of content words deleted, was clearly the most difficult of all the versions and on the other, the content words and the function words showed the same coefficient of difficulty in all the tests. The solution to this problem lies in the fact that the amount of information carried by content words is by no means greater than that relying on function words. The more content words are deleted, the more cues to the content are missing and the less the whole text becomes comprehensible. And as the text is made more difficult, all the deletions in it become more difficult independently of the category to

which they belong. This dependence of cloze scores on the textual information was stressed by Carrol, who called the whole mechanism "local redundancy" (in Johansson 1973).

According to what has been said about the quality of deletions in a cloze test, CLOZE IID should be the easiest, as it has the smallest proportion of "difficult items". However, it is on the same level of difficulty as CLOZE IIC. No doubt, the explanation lies in the fact that a distinctly larger number of deletions (in spite of their relative facility) in this version prevented it from being too easy. The conclusion can be made that the difficulty of a cloze test depends on both the quality and the quantity of deletions in a text.

A final remark here should be made concerning the reliability of the four versions. CLOZE IIA was slightly less reliable than the other versions, which is, no doubt, a result of its greater facility. CLOZE IID, as having the greatest number of items, should be the most reliable of all the versions. However, this was not the case and both CLOZE IIB and IIC were equally reliable. Again the explanation lies in the fact that quantitative changes in one version of a test can be balanced by qualitative changes in another.

4.3. The data presented in Fig. 1 show that with an increase of test difficulty the discriminatory power of the test becomes greater. The same is shown in Table 4, where CLOZE IIB, the most difficult version, has the highest coefficient of discrimination and CLOZE IIA, the easiest of all, the lowest (sharing it, however, with CLOZE IID). It has been said that, on the one hand, test difficulty depends partly on the quality of deletions and, on the other, that within one test difficulty of particular items is not concerned with any grammatical category and is the same for short-range and long-range items. Both the criteria for the division of cloze items (i.e., grammatical and contextual) were taken into consideration while examining the discriminatory power of the items. In both cases the results were the same; no regularity was found. The discriminatory power of long-range items was, on the whole, equal to that of short-range ones. Similarly, the arrangement of grammatical categories, in order of the discriminating power of the items, was totally different for CLOZE I and all the versions of CLOZE II. There was no stable difference between content words and function words, either. These data lead to the conclusion that the discriminatory power of the cloze items does not depend on any formal criterion of classification of the deletions, external to the text. The immediate context of the deleted words thus becomes the main factor that determines the functioning of cloze items. Typical examples of this are items 20) and 23) in CLOZE I and 3) and 27) in CLOZE IID. In CLOZE I in both cases the adverbial "very" was deleted from the following sentences: "because something 20) _____ strange happened on the Mary Celeste", "They did not feel 23) _____ happy about it." Item 20) was only slightly more difficult than item 23) for all the test subjects but it had a much higher coefficient of discrimination

(.73 against .22 for item 23). In CLOZE IID the auxiliary "did" was deleted from the following sentences: "Why 3) _____ they leave their ship?", "What 27) _____ this mean?". Item 3) was very easy and it did not discriminate at all (coefficient .00), while item 27) was rather difficult and its coefficient of discrimination was .60.

Both, the two examples and the data discussed before, show the importance of the immediate context of the deletions to the discriminatory power of the cloze test items. To summarize, we may say that the context of the deleted words is the main factor that determines the functioning of the cloze test items.

4.4. A typical multiple-choice recognition test does not differentiate among the subjects who have given wrong answers to particular items. This is because of the very character of the test in which not only the right answer but also the wrong ones have been programmed by the test writer. In other words, test subjects are not allowed to freely make their "own" mistakes. Also in most cases it is not possible (nor does it seem advisable) to make the distractors operate on various levels of difficulty. Such is not the case for a cloze test, where the error-analysis shows the existence of several levels of incorrectness of the answers. The following example from CLOZE I illustrates the above: "and he 49) *sent* some men in the boat to 50) _____ Captain Morehouse from the Dei Gratia." Filling in "take" in item 50) is a vocabulary mistake, though it proves that the test subject fully understands the context. The answer "the" is a grammatical mistake, though it is fully acceptable as far as meaning is concerned. The answer "with" is both ungrammatical and lacking in sense and as such it shows that the test subject does not understand the context. The three answers obviously represent different levels of incorrectness and therefore should be treated separately in a diagnostic analysis. (Though, for reasons of practicality, such a differentiation should be avoided in scoring. Cf. Oller 1972a).

The classification of errors can be useful in two ways; first, by providing the teacher with a list of productive errors of his class in order of their frequency, and secondly, by offering an insight into the problem of cloze test validity (as it shows on what grounds the discrimination of test subjects occurs).

The error-analysis made for CLOZE I showed the occurrence of the following types of errors:

(a) lexical errors — a wrong choice of a lexical unit, e.g., "take" for "bring" in item 50);

(b) structure errors — different types of violations of the structural system of the language;

(c) errors caused by partial misinterpretation of the context, e.g., "Morehouse called 37) *some* of his officers and sent him ...". Errors of this type occur when a test subject has not taken into consideration some details given in the text; they are more textual than structural;

(d) errors caused by total misunderstanding of the immediate context, e.g., "Morehouse called 37) *and* of his officers and sent ...". By analogy to the distinction made in 4.1., these errors could be called short-range choice errors, whereas errors belonging to the previous category would be long-range choice errors;

(e) blanks left after the completion of the test. (For other classifications of errors see Oller et al. 1972, Oller and Inal 1971).

In many cases categories overlap. Most often such is the case for lexical or long-range errors accompanied by some structural violations, e.g., "Morehouse and his men 13) *meet* another ship far away." (It should be "saw" here.) The treatment of blanks is a matter of question. The detailed analysis showed that they are especially numerous when many short-range errors occur. This observation might show the similar character of the two categories. (A test subject puts a nonsensical word in the blank or leaves the blank untouched when he does not understand the context.) However, such an interpretation would be an oversimplification. There are many cases where a test subject, though he understands the context, remains unable to fill in the blank because of his lack of productive resources.

As was said in 2.1., CLOZE I discriminated on different levels of language proficiency with a tendency to discriminate better along with the increase of the test difficulty. In what follows an analysis of errors will show whether the differences between levels of proficiency are only quantitative or whether the type of discrimination changes as well. The discrimination on three levels was considered: between top and bottom halves of the top half of P, between top and bottom halves of P and between top and bottom halves of the bottom half of P. Not all the items were considered, but only those with a discrimination coefficient higher than or equal to the mean coefficient of discrimination for the given group. In all the groups the items selected in such a way represented about 70% of the discrimination. In the first group there were 15 such items and 159 errors. Out of these, 66 were structural errors, 31 long-range errors, 29 lexical errors, 29 blanks and only 4 short-range errors. In the second group (embracing the whole population of the test), there were 25 items discriminating above the mean and 967 errors (402 of them were blanks, 185 structure errors, 173 long-range errors, 136 short-range errors and 71 lexical errors). In the third group (the bottom half of P), 23 items had the coefficient above the average and 377 errors were analysed (156 of them were blanks, 90 were structural errors, 82 short-range errors, 37 long-range errors and only 12 lexical errors). To summarize these data, it is possible to say that at all levels of proficiency CLOZE I was a test of structure (as it discriminated between those who made structural mistakes and those who did not). At a higher level of proficiency it was also a test of comprehension and a lexical test. Less proficient test subjects were not tested on the comprehension of the whole text, but rather on the understanding of fragments of sentences. (Note that this corresponds to the natural sequence of language learning, especially as the forming of language retention is concerned.).

GENERAL CONCLUSIONS

5.0. Bearing in mind that, in order to be fully reliable, the results of the experiment presented in this paper should be confirmed by some further experimentation, we may conclude as follows:

1) A cloze test can be a very good instrument of assessing language proficiency, especially as a part of a battery of tests. It is comparable to other tests in reliability (though it would be very difficult to make it as reliable as a multiple-choice test). It has an advantage of being a test of production and, therefore, can be very useful as a diagnostic test. On the other hand, as the functioning of cloze items depends mainly on the immediate context of deletions, its applicability as an achievement test seems very limited.

2) The validity of a cloze test can be best perceived through the analysis of errors. Such an analysis shows that a cloze test operates at various levels of language proficiency and that it measures different elements of language skills at different proficiency levels.

3) It is possible to manipulate the items of a cloze test so as to make the test easier or more difficult. Similarly, it is possible to use it as an exercise in teaching different components of language skills.

REFERENCES

Bloor, M., Bloor, T., Forrest, R., Laird, E. and H. Relton. 1970. *Objective tests in English as a foreign language.* London: Macmillan.

Harris, D. P. 1969. *Testing English as a second language.* New York: McGraw-Hill Book Company.

Johansson, S. 1973. "Partial dictation as a test of foreign language proficiency". *Swedish-English contrastive studies. Report 3.* Department of English, University of Lund.

Lado, R. 1961. *Language testing.* London: Longmans.

Lapidus, B. A., Shelkova, T. G. and S. D. Lysko. 1969. *English through practice.* Moscow: IMO.

Oller, J. W. 1972a. "Scoring methods and difficulty levels for cloze tests of ESL proficiency". *Modern language journal 56.* 151 - 157.

Oller, J. W. 1972b. "Assessing competence in ESL: reading". *TESOL 6.* 313 - 325.

Oller, J. W. and C. Conrad. 1971. "The cloze technique and ESL proficiency". *Language learning 21.* 183 - 195.

Oller, J. W. and N. Inal. 1971. "A cloze test of English prepositions". *TESOL 5.* 315 - 327.

Oller, J. W., Bowen, B. T., Dion, T. and V. W. Mason. 1972. "The cloze tests in English, Thai and Vietnamese". *Language learning 22.* 1 - 15.

Rand, E. J. 1972. "Integrative and discrete-point tests at UCLA". *Workpapers. Teaching English as a second language 6.* UCLA.

Valette, R. M. 1967. *Modern language testing: a handbook.* New York: Harcourt, Brace and World.