Developing A Cloze Procedure As A Reading Comprehension Achievement Test

I Ketut Seken
IKIP Negeri Singaraja

Abstract: The project was concerned with developing a cloze procedure as a reading comprehension achievement test. The subjects were students of the English Education Department of the Faculty of Letters, State University of Malang, who were halfway in the semester to complete Reading II course. The test was planned and constructed on the foundation of existing theory of cloze test construction. A review of theory concerning reading comprehension, testing reading comprehension, and cloze testing led to the construction of the test, including the decision concerning how to score the test and to interpret the scores. Using a class of 28 students, the test was tried out a week after the mid-semester test was administered by the Reading II teacher. It was found that the test is sufficiently reliable on the basis of a reliability coefficient of .79 through split-half procedure and a coefficient value of .78 by K-R 20. The test also showed high inter-section correlation. The validity of the test was viewed in terms of face, content, and construct. The test scores correlate moderately with those obtained from the mid-semester test by the teacher. Some problems are discussed and a suggestion made with regard to a possible solution to these problems.

Key words: cloze test, achievement test, reading comprehension

Reading is one of the major skill courses in the curriculum of the English Education Department of the Faculty of Letters, State University of Malang. There are six reading courses for the students to complete throughout their undergraduate (S1) program, including extensive reading. The

importance of these reading courses is evident not merely in association with the language-related skills that they have to acquire in order to graduate, but also in relation to the requirement for them in the program to be able to read content books written in English, such as books on linguistics, teaching methodology, and literature. It certainly follows that evaluating the students' reading achievement is a major and necessarily frequent activity for the teachers involved in these courses to do in order to monitor their progress in reading ability as well as to enhance their learning in the rest of the courses.

Ideally, assessment of reading ability should cover all of the subskills that together define reading ability. However, this is not an easy and straightforward matter to deal with. A number of studies cited by Lumley (2000), for example, indicate that research has not yet been able to present a clear concept of reading subskills. Besides, the extent to which such subskills are assessable is still largely under scrutiny.

There are several ways in which the students' achievement in reading comprehension can be assessed. One way to measure reading comprehension is to use the 'cloze' technique, which is commonly referred to as the cloze procedure. Heaton (1988) maintains that the most common purpose of the cloze procedure is to measure global reading comprehension. The procedure involves deleting a given number of words from a text and then having the subject attempt to guess and supply the words that have been deleted. The proportion of correctly-guessed words gives an indication of the extent to which the subject has understood the text concerned.

This paper reports on the result of development of a cloze procedure as a test to measure the reading comprehension achievement of EFL learners. The EFL learners in question were undergraduate students of the English Education Department of the Faculty of Letters, State University of Malang, specifically those who, when the project was carried out, were taking the Reading II course. The test, *Cloze Reading Comprehension Test* (CRCT) was planned and constructed by the writer, tried-out to a group of students whose achievement in reading comprehension the test is intended to measure, and the result analyzed.

The test development undertaken is intended to accomplish two purposes. First, it is meant as an effort to establish a valid and reliable test of the students' reading comprehension achievement, specifically their

achievement in the Reading II course, by means of a cloze procedure. Second, it attempts to reveal the merit and weaknesses of a cloze procedure when used to assess EFL students' achievement in a reading course. Besides being used as general reading proficiency tests, cloze tests have now been widely used in the classroom, such as in achievement, placement, and diagnostic tests (Heaton, 1988).

CLOZE TEST AND READING COMPREHENSION

Reading comprehension is basically an interactive process of meaning making between the reader and the author through the text, which involves mental activities and background knowledge (Weir, 1993; Singhal, 1999). Reading comprehension test, like any other trait, should always start from a clear understanding of what reading comprehension really is. The creation of a reading comprehension test, in other words, must always be based on a construct, upon which it can be justified to be a test as such. A reading comprehension test, whatever forms are used and however the problems are designed to realize it, should test the ability within a scope as broad as its construct allows. There are a handful of ways commonly used to test reading comprehension (see, for example, Djiwandono, 1996; Heaton, 1988). To use cloze tests to assess reading comprehension has also become a common practice in both L1 and L2.

Anderson (1976) has referred to cloze procedure as working on the basis of two very important characteristics of language: redundancy and sequential constraint. He views redundancy as the excess of rules of syntax in a language. Carroll (1964) sees redundancy as a property of language that allows the language user to predict missing symbols from the context. Redundancy reduces the possibility of errors and misunderstanding and allows communication where there is interference in the communication channel (Aitken, 1977). However, redundancy only works to the extent that the receiver of the message, in this case the reader, is capable of taking advantage of it. Associated with this notion, cloze procedure can be seen as assessing the reader's capability of making use of language redundancy contained in written text, for only when he/she is able to benefit from the redundancy will he/she be able to understand the mutilated passage.

By sequential constraint is meant the predictability of elements in a message by virtue of their statistical characteristics (Anderson, 1976). Associated with this is what is commonly referred to as 'grammatical expectancy' which allows a language user to predict the appearance of certain linguistic elements in a particular context of communication on the basis of given clues. In reading comprehension, the reader who possesses high capability in grammatical expectancy will be able to predict the occurrence of certain words or even phrases in the passage, given sufficient clues. Viewed in relation to this, cloze procedure can be thought of as measurement of reading comprehension on the basis of the reader's grammatical expectancy level.

PLANNING AND CONSTRUCTING THE TEST

The subjects to be assessed by the test were semester-3 students of the English Education Department of Faculty of Letters, State University of Malang. They were part of a body of undergraduate (S1) students majoring in English education preparing for a qualification to teach English at high schools. These students had been intensively trained in English in the two semesters that they had taken and were about halfway in completion of their semester-3 courses. One of the courses they took in semester 3 was Reading II, a four-credit course to be completed in at least 16 weeks (32 class meetings). It was their achievement in this course the test under scrutiny was to measure. In order to be allowed to take the course in their third semester study program, they must first have passed in Reading I, a prerequisite course taken in semester 2, which naturally serves as the foundation for the expected development of their reading ability through the Reading II course. The test was intended to measure their half-semester achievement (comparable to the mid-semester test conducted by their Reading II teacher).

The content of the test was made to adhere to the content of the course the achievement of which the test was to measure. This should mean that the test takes the course objectives as its source for the content to be tested. As a reading course, Reading II puts emphasis on reading as its sole content, the level of which is adjusted to the level of the students' proficiency at their present stage of English learning. Various reading passages from different sources have been selected for use in the course relating to the requirements in the accomplishment of the objectives of the

course. These reading texts are read, analyzed, and discussed with the students doing various exercises towards the acquisition of reading skills and strategies that will lead them to the competence for understanding texts of the level prescribed by the course. It is the overall ability as the reflection of the achieved competence that the test is intended to measure. To do this, the test uses texts taken from the same sources from which the course texts have been taken.

The CRCT is an achievement test packed in a cloze format. It is of the fixed-ratio deletion type of cloze test, in which every-nth word deletion method is applied. This choice was made considering that this is the most commonly used and the best researched type (Oller, 1979) as well as the purest (Anderson, 1976). The test employs every *tenth* deletion method since, according to research, this method gives the best result of cloze tests on non-native speakers of English (Klare, *et al.*, 1972; Heaton, 1988). A blank of standardized length replaces each deleted word.

The test consists of three parts (Part I, Part II, and Part III), each posing slightly different problems to the testees. The purpose of designing different parts of the test is basically motivational, that is, in order to motivate the testees to give their best efforts to do the test for test problems which are graded in difficulty may motivate the subjects and thus increase their confidence in doing the test (Weir, 1993). The test is graded in difficulty in the sense that the first part of it (Part I) is easier than the second part (Part II), and the second part is easier than the third part (Part III). With this grading the subjects are expected to have more facility in coping with the problems in Part I so that they may be motivated to do the next. Part I of the test consists of *multiple-choice* cloze problems following the model used by Djiwandono (1990). In this part, the testees are to find the correct answer to each problem out of four options provided. Part II poses cloze problems with *first-letter clues*, a technique of cloze testing suggested by Oller (1979) and Heaton (1988). Part II is expected to pose more difficult problems to the testees since they have to choose from many words known to them with the same first letter as that of the deleted word as compared to choosing from four options in the case of the problems in Part I. Part III poses cloze problems purely in their original sense, *without a clue* whatsoever, and should, therefore, be the most difficult part of the test.

Each gap and the context in which the gap occurs in the cloze texts is counted as an item of the test. Physically, the items consist of numbered blanks of the same length with specification according to the type of cloze problems posed to the testees as described in the discussion on the test format above. Each item in Part I has a numbered blank followed by four options (A, B, C, and D) which are put in parentheses and are printed in italic. One of these options is the correct answer to the problem, that is, it is the word that has been deleted from the text that has to be restored by the testees. Thus, to answer the problems in Part I the subjects are required to pick one of the options that they think is the one word that has been deleted from the text. In Part II each item has a numbered blank with the first letter of the deleted word printed on the left end of the blank. To answer the problems in this part the testees are required to print the word in question on the blank using the first letter clue that has already been printed there. In Part III each item has a numbered blank only, on which the subjects are required to print the word that they think has been deleted from the text completely on their own. There are 35 items in Part I, 17 items in Part II, and 21 items in Part III, mounting to a total of 73 items in the whole test. Table 1 shows the test sections with their respective item numbers and specifications.

The scoring of the test is based on the exact word scoring method for at least two reasons. First, the exact word method of scoring cloze tests correlates so strongly with all of the other proposed methods (Oller, 1979), meaning that whichever method of scoring used will give statistically the same result. Second, it is much easier to use than the other methods. The synonym method or the contextually acceptable method, for example, can be extremely difficult, time consuming, and subjective, with only a few advantages to make it worthwhile (Eanes, 1997; Alderson, *et. al*, 1995).

Besides, subjective judgments about which synonyms are acceptable can vary and the results can, therefore, be inconsistent.

As an achievement test, CRCT is meant to measure the students' achievement in reading comprehension after some period of learning. As mentioned earlier, the test is intended to measure the students' achieve-

Table 1. The Test Sections and Item Format Specification

TEST SECTIONS	ITEMS INCLUDED	FORMAT SPECIFICATION
Part I	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35.	Multiple-Choice
Part II	36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48,49, 50, 51, 52.	First-Letter Clue
Part III	53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73.	No Clue

ment after completing the first half of the course time so that it is comparable to a mid-semester test. This being the case, the test scores are interpreted in a criterion-based interpretation. As stated above, the highest or ideal grade of the students is 100. The criterion for the students' minimum success in reading comprehension achievement is determined on the basis of the assumption that the students have achieved at least 56% of the instructional objectives, taken globally in this case. This assumption is made on the basis of the passing grade criteria used by the Reading II teacher (as well as the other teachers in the English Education Department). This means that a grade of 56 reached by a student puts him/her on a position as a minimally successful achiever. Students whose grades are less than 56 are categorized as 'below criterion' or 'unsuccessful'. A set of success criteria can then be established on the basis of which the students' grades can be interpreted. Using four success descriptors: excellent, good, high average, and average, the range of grades and their descriptors can be shown as follows (adapted from Hopkins and Antes, 1990).

Table 2. Established Criteria for Interpretation

Grade Ranges	Descriptor Set A	Descriptor Set B
86 – 100	Excellent	A
76 – 85	Good	B
66 – 75	High Average	C+
56 – 65	Average	C
55 and lower	Below Criterion	Unsuccessful

THE TRYOUT

There were actually two steps of tryout done in the development of CRCT. A preliminary tryout was done in relation to the construction of the test, which was meant to accomplish three necessary tasks: (1) to guarantee the texts used in the test are of comparable level of difficulty to the ones used in the reading course concerned; (2) to find suitable distractors for the multiple-choice cloze items; and (3) to obtain the maximum criterion score. It is after this preliminary tryout was carried out that planning and constructing the test as described in 3.1 could be done. The inputs received from the results of the preliminary tryout were maximally utilized to produce the best possible form of the test to be further tried out (main tryout).

There were two classes of semester-3 students taking the Reading II course that were used in relation to the development of the test. The first group consisting of 25 students was used for the preliminary tryout. The second group was used for the main tryout of the test. There were 28 students in this group and all participated in the tryout. Like the preliminary tryout, the main tryout was also carried out within the scheduled time for the Reading II course as arranged on the timetable. The Reading II teacher was the one who supervised the test, without the test developer being present in the room where the test was taking place. Though the test direction is already clearly printed on the test paper, the teacher was advised to give the direction again orally so that none of the students would be stranded in doing the test because of not getting the direction clearly. This is

Table 3. The Scores of Three Subjects to Be Used as Maximum Criterion

SUBJECTS	RESULT							
	Part I		Part II		Part III		Total	
	RS	%	RS	%	RS	%	RS	%
01	26	74	11	65	15	71	52	71
02	29	83	13	76	16	76	58	79
03	30	86	11	65	13	62	54	74
Average	28.3	81	11.7	69	14.7	70	54.7	75

RS: Raw Score, % : Percentage Score

necessary to emphasize since it was the first time the subjects were exposed to a cloze test. After completing the test, the students filled in a brief seven-item questionnaire, which mainly aims at getting data on the face validity of the test.

THE TEST RESULT ANALYSIS

Table 4 shows the scores obtained by the students on CRCT, which then were converted into grades shown on Table 5. The interpretation of the students' grades as shown on Table 6 gives indication about how many students are successful achievers and how many are not. The table shows that 23 students (82%) are above the criterion and are therefore successful achievers. The result of the test tryout thus described is subject to the 'quality' of the test as the instrument by which such result has been yielded. The essential characteristics of the test are discussed below to see the extent to which the test measures what it is meant to measure, the degree of its consistency as a measuring instrument, and the degree of effectiveness and efficiency of the items used in the test to reveal the strength or the weaknesses of the testee in relation to the ability measured.

The Validity of the Test

Four types of validity are examined in order to see whether or not CRCT is a valid measuring instrument. They are *face* validity, *content* validity, *concurrent* validity, and *construct* validity. Face validity simply concerns the 'look' of the test and involves 'lay' people's intuitive judgement about the content of the test (Alderson, *et al.*, 1995). The face validity of CRCT was first commented by the Reading II teacher, who looked at the test for the first time when the test was about to be tried out. To her, though it looks difficult, the test is a suitable instrument to assess the students' reading comprehension ability in a general sense. The students who did the test were mostly of the same opinion. From the questionnaire distributed to them after doing the test, it can be concluded that the test does have face validity, judging from the fact that 73% out of 26 students say that the test is a suitable test to use as a reading comprehension test.

The degree of content validity for an achievement test is determined by comparing the content of the test with the content of classroom in-

struction (Hopkins and Antes, the content of the test is adjusted to the

Table 4. The Subjects' Scores

SUBJECTS	RESULT							
	PART I		PART II		PART III		TOTAL	
	RS	%	RS	%	RS	%	RS	%
01	12	34	7	41	7	33	26	36
02	14	40	7	41	5	24	26	36
03	20	57	10	59	10	48	40	55
04	21	60	10	59	13	62	44	60
05	18	51	8	47	10	48	36	49
06	24	69	12	71	14	67	50	68
07	21	60	12	71	12	57	45	65
08	18	51	8	47	13	62	39	53
09	14	40	8	47	7	33	29	40
10	22	63	12	71	10	48	44	60
11	17	49	8	47	11	52	36	49
12	22	63	14	82	13	62	49	67
13	21	60	10	59	14	67	45	62
14	20	57	11	65	12	57	43	59
15	19	54	8	47	12	57	39	53
16	14	40	6	35	8	38	28	38
17	18	51	9	53	7	33	34	47
18	15	43	4	24	8	38	27	37
19	21	60	9	53	13	62	43	59
20	23	66	11	65	13	62	47	64
21	17	49	8	47	10	48	35	48
22	19	54	10	59	13	62	42	58
23	16	46	8	47	10	48	34	47
24	16	46	10	59	7	33	33	45
25	13	37	10	59	7	33	33	45
26	19	54	5	29	7	33	31	42
27	18	51	11	65	11	52	40	55
28	22	63	13	76	14	67	49	67

RS: Raw Score % : Percentage Score

Part I : Mean = 52.43: SD = 9.26; Variance = 85.81; Range = 35; Max. = 69; Min. = 34; Sum = 1486.
Part II : Mean = 54.46; SD = 13.89; Variance = 192.99; Range = 58; Max. = 82; Min. = 24; Sum = 1525.
Part III: Mean = 50.04; SD = 12.65; Variance = 159.96; Range = 43; Max. = 67; Min. = 24; Sum = 1401.
Total : Mean = 52.18; SD = 9.98: Variance = 99.63; Range = 32; Max. = 68; Min. = 36; Sum = 1461.

content of the Reading II course, the achievement of which it is to measure. The objectives of the course, which covers four components: language, text content, text structure, and reading skills, constitute the basis on which the content of Reading II course is to be determined. These objectives are also the source for the content of the test, though not precisely in the sense adopted in the more

Table 5. The Students' Grades and Their Interpretation

Subjects	Score (%)	Grades	Interpretation
06	68	91	Excellent
28	67	89	
12	67	89	
20	64	85	Good
13	62	83	
07	62	83	
04	60	80	
10	60	80	
14	59	79	
19	59	79	
22	58	77	
27	55	73	High Average
03	55	73	
08	53	71	
15	53	71	
11	49	65	Average
05	49	65	
21	48	64	
23	47	63	
17	47	63	
24	45	60	
25	45	60	
26	42	56	
09	40	53	Below Criterion
16	38	51	(Unsuccessful)
18	37	49	
01	36	48	
02	36	48	

specific-objective oriented achievement tests. Being a cloze test, the present achievement test is more oriented in its content to the global objective

of the course rather than its componential objectives. It measures the overall reading comprehension ability of the students after completing half of the course time. This overall ability that the test measures is the reflection of the achieved competence in reading comprehension as the result of learning through the half time of the course. It is in this sense that the test can be regarded as having content validity.

Table 6. Percentages of Achievers under Each Descriptor

Descriptors	Number of Achievers	Percentages
Excellent	3	11%
Good	8	29%
High Average	4	14%
Average	8	29%
Below Criterion	5	18%

Seeing the concurrent validity of a test essentially concerns comparing the test scores with some other measure for the same subjects at roughly the same time as the test. In relation to this Alderson, *et al.* (1995) state that the test scores can be compared to other measures, such as scores from another parallel version of the same test or from some other test; the candidates' self assessments of their language abilities; or the candidates' ratings on relevant dimensions made by teachers, subject specialists, or other informants. Similarly Heaton (1988) refers to such criterion measure as (a) an existing test, known or believed to be valid given at the same time; (b) the teacher's ratings or any other such form of independent assessment given at the same time or later; or (c) the subsequent validly measured performance of the testees.

There were no scores from a reading comprehension test of known validity that could be drawn from the students so that a concurrent validity coefficient using a criterion measure of known validity could not be obtained. The measure that was used to concurrently validate the test was in the form of scores obtained by the teacher from the mid-semester test given a week before the tryout was carried out. The computation of correlation between the students' grades on CRCT and their mid-semester test scores yields a correlation coefficient of .50 ($p = .01$). This means that

the two measures are significantly correlated though moderately, which therefore indicates that the CRCT does have some concurrent validity. Table 7 shows the two measures compared. Construct validity of a test indicates the relationship between what a theory predicts and what test scores show and, therefore, establishment of construct validity is in principle theory validation.

Table 7. The CRCT and the Mid-Semester Test Compared

	Mean	SD	Variance	Range	Max.	Min.
CRCT	69.57	13.32	177.44	43	91	48
Mid-Sem. Test	73.21	10.95	119.80	39	88	49

r = .50 p = .01

There are two questions that require solution in relation to the construct validity of CRCT. First, 'How appropriate is it for the subjects tested?' and, second, 'How appropriate is it to be a test of language proficiency? The answer to the first question lies on the test's being an achievement test. It was administered after an instructional period had been completed, in which the students were trained and taught with materials as required by the curriculum. The test was an instrument to find out the gain of the instruction by revealing individual students' overall achievement and the overall achievement of the class. To the subjects the scores obtained on the test provide indication concerning their progress. What the test did is essentially appropriate for the students so that the test can be said to have construct validity.

The Reliability of the Test

Test reliability is the degree of consistency of measurement that a test yields in measuring what it is intended to measure; it is obtained and used as an index of measurement consistency the test performs. The reliability of CRCT was obtained through three methods, all being concerned with the internal consistency of the test. These are (1) internal correlation method, (2) split-half method, and (3) K-R 20. Computation using SPSS/PC+ yields correlation coefficients between parts and between the part and whole of the CRCT which indicate moderate to high correlation

between them. These correlations indicate that all parts and the whole of the test measure the same trait and that they put the subjects on relatively the same ranks. This reflects the internal consistency of the test.

Table 8. Inter-part correlations of the CRCT

Test Components		Correlation	
		r	p
Part I	Part II	.68	.00
Part I	Part III	.76	.00
Part II	Part III	.63	.00
Part I	Whole	.92	.00
Part II	Whole	.85	.00
Part III	Whole	.89	.00

The computation of the split-half scores of the test yields a correlation coefficient of .65 (p = .00) as a correlation coefficient of half on the test scores. The Spearman-Brown Prophecy formula further shows the adjusted full-test reliability of CRCT, which is .79. With this reliability coefficient, the test can be said to have moderate to high reliability. To confirm this status of reliability, K-R 20 was further used, which yields a reliability coefficient of .78.

Item Analysis

Though item analysis (such as item facility analysis or item discrimination analysis) is more commonly associated with norm-referenced tests, to some extent it is nevertheless of significance to undertake for a test developer dealing with a criterion-referenced test. This is so considering that whatever the kind of test used in a particular assessment activity the items of the test must qualify to be ones that serve the purpose of testing. An analysis was done to the items of the CRCT. The analysis covers (1) IF analysis, (2) item discrimination (ID) analysis, and (3) distractor-efficiency analysis (relevant only to the multiple-choice items). The result of the IF analysis shows 33% of the items are unacceptable, 17% being too easy and 16% too difficult. The ID analysis yields information that only 20% of the items are well qualified as test items as such, 39% must undergo improvement in order to qualify as good measuring items, and 41% must be rejected completely for the poor discriminating ability asso-

ciated with them.

Analyzing the distractors of a multiple-choice item chiefly aims at finding out the degree to which the distractors really distract the testees who are not certain of the correct answer to the item. This is done by analyzing the percentages of the subjects who chose each option to answer the item. To see whether or not distractors are functioning efficiently, it is more informative and useful to see what the distribution of responses was for the upper, middle, and lower groups (Oller, 1979; Brown, 1996). This is accomplished by devising a response frequency distribution, which shows the proportion (usually in percentage or decimal) of the upper, middle, and lower group subjects that chose each option for each item of the test. By observing the response distribution frequency, one can clearly see the 'problematic' option(s) of each item of the test. As for the items of the CRCT, a number of options (including some correct options) of a number of items were found to be functioning badly. There are 17 items out of the 35 multiple-choice items of the CRCT that have at least one 'zero' option, that is, one which was chosen by none of the subjects. The distractors of the other items can be said to be working, taking the stance that a distractor is functioning if it was chosen by at least one test-taker.

DISCUSSION

From the analysis of the tryout result some broad views may develop. First, the test seems to have reasonable validity, though not empirically. Its face validity, content validity, and construct validity do look good. Its concurrent validity, however, may require rechecking. Correlated to a teacher-made test with a coefficient value of .50, though significant, the test still largely lacks evidence to be concurrently valid. That it is comparable to some extent to the mid-semester test used by the teacher is of course illuminating. At least it should mean that the test measured the same trait as did the mid-semester test. The slightly lower mean score obtained by the CRCT as compared to the mid-semester test can be related to the relative newness of the cloze procedure to the students, while they are already very familiar with the kind of test used by their own teacher.

Second, the test also looks good in terms of reliability, particularly its internal consistency. A reliability coefficient of .79 obtained through

split-half method and .78 from K-R 20 indicate that the test is convincingly reliable internally. The high correlation coefficients obtained between its parts and between the part and whole give further indication that the test has solid internal consistency.

Third, the test does not however look good enough in terms of the items that compose it. Since items are the basic units of a test (Brown, 1996), this needs serious attention. The results of the item analysis suggest that there is much that needs to be done in relation to the items if the test is to improve in its ability to measure accurately. It is true that 68% of its items have acceptable IF value. However, it must be remembered that this calculation is based on Oller's (1979) method, which is as lenient as allowing IF ranging from .15 to .85 to be acceptable. If a more severe method had been used, more items would have been categorized as 'unacceptable'. The ID analysis reveals even a more worrying picture of the items. Only 21% of the items fulfill the requirement to be good items in terms of ID value, while 41% should definitely be rejected. A number of 28 items (38%) must in theory receive treatment to improve in order for them to stay in the test. The problem might not be so complicated if the test being developed was not a cloze test. The problem with cloze tests in this respect has long been realized, yet solution to it has never come to satisfaction. A clear disadvantage of cloze test is that it is not easily amended (Alderson, *et al.*, 1995). It is often the case that the weakness of a cloze test is known before the test is used, yet a test constructor may not find it easy to handle. Many times the whole test has to be completely dropped for the sake of a few items that simply do not qualify as test items. This problem usually confronts a test constructor adopting a strict exact-word method.

If weak items such as this is to be improved, the possible improvement is perhaps to be done by providing more clue for the testees. However, it is not possible for the test developer to improve only one or a few items. If a cloze test should undergo item alteration to improve, then all of its items are to be treated as such. This can be done, for example, by providing "two first letters" instead of one, so that in the case of the word *altered* above the clue would be *al*, which would rule out all of the words like *affected, abolished,* and *annoyed.* The problem with this is that for some other items this might make them 'too easy', which will certainly

pose another problem for the test constructor.

For the items with no-clue format, the problem is similarly complicated, if not more seriously so. Items 57 and 58, for example, are problematic in precisely opposite nature. Item 57 is too easy because the context has the word *instance* following the deleted word so that none of the students missed this item. Quite the contrary from this is item 58, which requires the testees to recover a name word or proper noun. Both of these items are therefore a 'waste' in the test. The question is how do we improve them? The answer to this question is apparently unavailable for the moment. What is readily available is a suggestion to improve the method, that is, to alter it for example by replacing the every-nth deletion method by some other method through which cloze test items can be planned and rationalized in a more reasonable way.

In the lights of the results of the item analysis undertaken in this project, the suggestion concerning altering the method of cloze testing may be worth considering, especially when the cloze test is used for instructional concern. Modified versions of cloze tests may be more suitable to use in the classroom in the sense that they are more adjustable to the specific instructional objectives that the classroom testing is most logically concerned with. Similarly, the 'severe' and somewhat illogical exact-word scoring method may not be suitable to use in an instructional context. Where cloze tests are used to promote instruction and student learning, a more adaptable scoring method may be preferred.

CONCLUSION

The above discussion and the entire experience in developing a cloze procedure as an achievement test can be concluded in a few statements. In a broad sense it can be said that the test at issue is essentially worth its purpose and the entire process of development it underwent within the time available for the project. The idea to develop a cloze procedure as an achievement test was in fact felt queer at first, yet it was intriguing as well. The result of the tryout, which showed that the test is in essence valid and reliable and could put the students in relatively the same ranks as did an existing achievement measure really used to determine their grades, should to some degree clear out any doubt that a cloze test can possibly serve as an achievement test. Some problems concerning this

status, however, did make themselves occur especially in matters that concern item qualification and improvement. Based on the result of the study, it is strongly suggested, in view of the use of cloze test for instructional concern, that the test method be improved, especially in terms of its construction and scoring method. It is also perhaps timely to question the statement that constructing a cloze test is an easy job, for what matters is not how to construct the test per se but how the constructed test can serve its purpose well either as a measuring instrument or as an instructional means.

REFERENCES

Aitken, K. G. 1977. "Using cloze procedure as an overall language proficiency test". *TESOL Quarterly*, 11, 1, pp. 59-67.

Alderson, J. C., C. Clapham, and D. Wall. 1995. *Language Tests Construction and Evaluation.* Cambridge: Cambridge University Press.

Anderson, J. 1976. *Psycholinguistic Experiments in Foreign Language Testing.* University of Queensland Press.

Brown, J. D. 1996. *Testing in Language Programs.* Upper Saddle River, N. J.: Prentice Hall Regents.

Carroll, J. B. 1964. *Language and Thought.* Englewood Cliffs, N. J.: Prentice-Hall.

Djiwandono, M. S. 1990. Laporan Penelitian: Pengembangan Tes Kemampuan Berbahasa Indonesia. Pusat Penelitian IKIP Malang.

Djiwandono, M. S. 1996. *Tes Bahasa dalam Pengajaran.* Bandung: Penerbit ITB.

Eanes, R. 1997. *Content Area Literacy: Teaching for Today and Tomorrow.* New York: Delmar Publishers.

Heaton, J. B. 1988. *Writing English Language Tests.* New Edition. London: Longman.

Henning, G. 1987. *A Guide to Language Testing: Evolution, Evaluation, Research.* Cambridge: Newbury House Publishers.

Hopkins, C. D. and R. L. Antes. 1990. *Classroom Measurement and Evaluation.* Third Edition. Itasca, Illinois: F. E. Peacock Publishers, Inc.

Klare, G. R., H. W. Simaiko, and L. M. Stolurow. 1972. "The cloze procedure: a convinient readability test for training materials and translations". *International Review for Applied Psychology* 21(2), pp. 77-106.

Lumley, T. 2000. "The notion of subskills in reading comprehension tests: an EAP example." *Language Testing Journal*, Vol. 17, No. 1, pp. 211-35.

Oller, J. W. Jr. 1979. *Language Tests at School: A Pragmatic Approach.* London: Longman.

Singhal, Meena. 1999. "A comparison of L1 and L2 reading: cultural differences and schema". *The Internet TESL Journal* 3(2), pp. 123-41.

Weir, C. J. 1993. *Understanding and Developing Language Tests.* New York: Prentice Hall.