Can a Multiple-Choice Rational Cloze Really Measure Reading Comprehension?

by *Tadashi Noro*

1. INTRODUCTION

Cloze tests are not new to English instructors in Japan. At the university level, however, they are not so frequently used as they are at the high school level. Cloze tests have been used as exercises for promoting careful reading, training vocabulary building, etc. as well as testing devices for measuring English proficiency. Oller (1979: 348–63) listed the following as applications of cloze procedure: 1) judging readability of textual materials, 2) estimating reading comprehension, 3) studying the nature of contextual constraints, 4) estimating overall language proficiency (especially in bilingual and second language learners), and 5) evaluating teaching effectiveness. Among those applications I have been interested in cloze tests in terms of reading comprehension because some English Testing Syndicates, such as STEP (The Society for Testing English Proficiency), employ cloze procedure presumably for measuring reading comprehension.

In a cloze test, blanks are placed in a given text where every *n* th word (usually 5th–12th) has been deleted. The examinees are required to restore the missing word from all the contextual clues available by filling in the blanks. M. L. Taylor invented the cloze technique to measure the readability of texts for native readers. Oller (1979: 341) discloses that "he is also responsible for coining the word 'cloze' which is rather obviously a spelling corruption of the word 'close' ... The term is a mnemonic or perhaps humorless pun intended to call to mind the prospect of closure celebrated by Gestalt psychologists." That is, his principle is based on the Gestalt theory of 'closure', which means "the perception of incomplete figures or situations as though complete by ignoring the missing parts or by compensating for them by projection based on past experience (Webster's Third International Dictionary 1981)." This original cloze test procedure can be modified or developed for specific purposes by presenting three or four alternate responses for each blank, and/or by deleting the words which a test designer wants to test.

The purpose of this paper is to discuss advantages and disadvantages of conventional multiple-choice reading comprehension tests, random cloze tests, multiple-choice random cloze tests and multiple-choice rational cloze tests, and to compare a multiple-choice reading comprehension test with a multiple-choice rational cloze test on the basis of statistical evidence. A number of empirical research studies have shown the close relationship between individual scores on cloze tests and reading ability. Oller and Conrad (1971:187) found the high correlation (0.80) between individual scores on a cloze and a reading section on the UCLA ESL Placement Examination. Stubbs and Tucker (1974: 239–42) obtained the significant correlation (0.67 by exact word scoring; 0.70 by acceptable word scoring) between the two subsets (cloze and reading) of the English Entrance Examination given to the applicants to American University in Beirut. Bensoussan and Ramraz (1984:235) reported that the Pearson correlations between multiple-choice reading comprehension and multiple-choice rational cloze scores were 0.75 for the first battery and 0.79 for the second. In Japan, too, Shimizu (1989:109) demonstrated that the correlations between reading scores of STEP tests and those of random cloze tests were 0.63 (exact word scoring) and 0.68 (acceptable word scoring) and that the correlation between reading comprehension and multiple-choice random cloze was 0.68.

2. Conventional Reading Comprehension Tests

Two main conventional reading comprehension tests are true/false tests and multiple-choice tests. The merits of true/false tests lie in the easiness of selecting suitable test passages and in the easiness and speed of test construction. In addition to that, scoring them is straightforward and quick, but the scores obtained can be unreliable because they encourage guessing and there may not be enough well constructed items. Therefore, true/false tests are best used as class progress tests or as teaching devices for directing the students' attention to the salient points in the passage.

Multiple-choice reading comprehension tests also have some demerits. First, the ratio of text to items is inefficient; a testee is required to read quite a lot of lines of texts to answer relatively few questions. Second, it is rather difficult and time-consuming to construct flawless multiple-choice items. Third, whether or not a multiple-choice reading test is a good test of reading comprehension depends on the difficulty level of a text and the type of items. A difficult text can encourage random guessing. Some items may closely be related to the testing of vocabulary or of comprehension of grammatical structures. In spite of these demerits, multiple-choice reading tests are useful because scoring them is very easy.

3. Multiple-Choice Rational Cloze Tests

Cloze tests can be categorized into two types according to the system of deletion: "random cloze" tests, in which every *n* th word in the text is omitted, and "rational cloze" tests, in which a test designer decides which words and how many words to delete on the basis of some rationale. Researchers have different opinions as to whether the random cloze or the rational cloze is a better test of reading comprehension. The random cloze procedure "has a good text-item ratio and requires relatively little re-adjustment time between items (Bensoussan and Ramraz 1984:231)." However, you have to stick to the permitted span between gaps. If you changed it, it would be no longer a random cloze test. On the other hand, in the rational cloze you can delete the words you want to test, such as function words or content words, for specific purposes. Unlike the random cloze test, it is possible that a blank space in the rational cloze test can take the place of more than one word, such as an idiomatic expression.

Deletion should be based on some criteria. Bensoussan and Ramraz (1984:231) proposed three criteria for deleting items according to discourse analysis theory:

1) the *micro-level,* focusing on the lexical choice of words and their interaction with other words in the context;
2) the *pragmatic-level,* which is extra-textual and draws on the reader's general knowledge of the world; and
3) the *macro-level,* dealing with the function of the sentences and the structure of the text as a whole.

To put it more concretely, macro-level items would test "writer's opinion, words showing comprehension of key concepts, function words signaling contrast / opposition, and main idea of paragraph (Bensoussan and Ramraz 1984:231)." Backman also suggested four types of deletions:

1) *syntactic,* which depend only on clausal-level context;
2) *cohesive,* which depend upon the interclausal or intersentential cohesive context;

3) *strategic*, which depend on parallel patterns of coherence (Backman 1982:63);

4) *extra-textual*, which depend on extra-textual schematic context (Backman 1985: 538).

When we aim at measuring reading comprehension, it might be advisable to focus on macro-level, cohesive or strategic items rather than on micro-level or syntactic items. However, in rational cloze tests as well as in random cloze tests, it is almost impossible for testees to get full marks, and marking is awkward and time-consuming. For this reason, we can't use such cloze tests in English entrance examinations or in school achievement tests.

The high correlation with reading tests does not necessarily guarantee that the cloze test measures reading ability, because what general English proficiency cloze tests assess seems to comprise various components of testees' English ability. When they fill in the blanks of a cloze test, apparently they need production ability besides comprehension ability. So, a multiple-choice format is a kind of solution to facilitate the measuring of true reading comprehension ability. This format also can overcome the demerit of time-consuming scoring.

So far, I have explained that the multiple-choice rational cloze test offers an effective and efficient way of testing reading comprehension in terms of designing and scoring reading tests. I wished to find statistical proof that this type of cloze test would do its job as well as the frequently used, conventional type of multipul-choice reading test. Bensoussan and Ramraz (1984:230–39) conducted extensive experiments on effectiveness of a multiple-choice rational cloze, so my following experiment is a sort of simplified replication of their study.

4. Experiments on Effectiveness of M-C Rational Cloze Tests

4.1. Purpose

The present study aims to give statistical evidence for the effectiveness of multiple-choice rational cloze test (R C) by comparing them with multiple-choice reading comprehension test (M-C), and to find whether or not similar results Bensoussan and Ramraz had would be obtained from a limited study using Japanese college students.

4.2. Hypotheses

It was hypothesized as follows:

(1) There are significant correlations between R C scores and M-C scores.

(2) The correlation between the scores of M-C tests and those of R C tests becomes higher as the number of items of the cloze tests increases.

4.3. Subjects

33 second-year students of Soai University from April, 1993 through October 1993.

4.4. Materials

3 M-C tests and 2 R C tests were administered.

(1) The 10-item M-C reading comprehension test taken from the second passage in *Reading Comprehension Test Papers*, 1975, Oxford University Press.

(2) The 8-item M-C reading comprehension test taken from Green III a Rate Builder No.1 in *SRA Reading Laboratory.*

(3) The 5-item M-C reading comprehension test taken STEP 2nd-class test held in October, 1992.

(4) The 5-item M-C rational cloze test taken STEP 2nd-class test held in July, 1985.

(5) The 25-item M-C rational cloze test taken from "Television" in *A New Way to Proficiency in English* 1974, Kenkyusha.

4.5. Procedure

The M-C reading comprehension test No.1 and No.2 were conducted with two classes of second-year students, totaling 53, at the end of April, 1993. Since reliability seems to be affected by the number of items in the test, another M-C reading comprehension test No. 3 was administered to the same class, totaling 55, at the end of July, 1993. M-C in the tables stands for the total scores on M-C No.1, M-C No.2 and M-C No.3.

The M-C rational cloze test No.4 (RCl) was conducted with the same classes of students, totaling 55, at the end of July, 1993. In order to ensure that higher correlation between the score of M-C tests and R C tests will be obtained as the number of items of cloze test increases, another 25-item R C test (RC2) was administered to the same classes, totaling 39, in the middle of October, 1993. RC3 is the total of RCl and RC2. The net number of 33 subjects means that 33 students out of 58 students in the two classes took all of the five tests.

In constructing M-C rational cloze tests, I kept the following points in mind as Bensoussan and Ramraz suggest (1984:232). The first point is to use words for alternative responses focusing on a particular point, either in terms of content or structure. That is, students will find it easy to choose a correct response from among four similar types of distractors, such as four adjectives, rather than from among different type alternative responses, such as two adjectives and two conjunctions. The second point is to delete a variety of words, from content words to function words. The third is to avoid synonymous distractors, which may cause the correct choice to be ambiguous (even for native speakers). The fourth is to avoid asking about detailed grammatical points, because the RC test is essentially a test of reading comprehension.

4.6. Results and Discussions

As Table 1 shows, the Pearson correlation between M-C and RC1 scores was 0.54; between M-C and RC2 scores being 0.57; between M-C and RC3 being 0.67. The two-tailed tests of significance at the level of $p<0.05$ indicate that these correlation coefficients are significant and meaningful, so it can be safely be said that there were considerable correlations in three cases, and Hypothesis 1 can be proved. However, the problem is that in this study the number of subjects was very small, compared with more than 7,000 subjects in Bensoussan and Ramraz (1984). They had high correlations between M-C and RC scores ranging from 0.748 to 0.798. There is a strong possibility that we will obtain higher correlations if the number of subjects increases.

Table 1 Pearson Correlations between M-C Tests and RC Test Scores

	RC1	RC2	RC3 (total)
M-C	0.54* (p=3.75)	0.57* (p=3.61)	0.67** (p=5.02)

* $p<0.05$ ** $p<0.01$

Hypothesis 2 can be proved. As the number of items increases, the correlation coefficients become a little higher. Henning (1987:78) estimates the probable relationship between reliability and the number of items in a test at approximately 0.50 reliability with 25 items. Actually, however, further systematic studies are needed to decide at least how many items are necessary for a reliable rational cloze test for measuring reading comprehension. RC1 test, which consists of only 5 items, shows 0.54. How do I interpret this figure? Is it possible to say that only a 5-item cloze test can measure reading ability if it is carefully constructed? This point is another challenge for my further study.

Table 2 A Comparison among M-C and RC Tests

	M-C	RC1	RC2	RC3
Number of Items	23	5	25	30
Number of Subjects	33	33	33	33
Mean Scores (%)	10.63 (46.2)	2.24 (44.8)	11.21 (44.8)	13.45 (44.8)
Standard Deviation	2.434	1.061	2.583	3.042
K-R Reliability	0.04	−1.22	0.08	0.21
Discrimination Indices (average)	0.22		0.25	

The problems are that all of the tests show quite low reliability coefficients and that the average of discrimination indices (or easiness indices) indicate that many of the items of each test fail to discriminate effectively since "items showing a discrimination index of below 0.30 are of doubtful use (Heaton 1975:177)". There are some reasons for these figures. All of the test materials may have been difficult for the subjects, as the average score of each test indicates. This led to the rather narrow range between the highest and lowest marks. All of the standard deviations show very small spread of scores, too. It is also partly because many students didn't take all five tests, especially students with poor English proficiency and partly because our English students are rather homogeneous in English proficiency.

CONCLUSION

In this research I discussed some conventional reading comprehension tests and cloze tests, and concluded that a multiple-choice rational cloze test is one of the most useful reading comprehension tests. Next I examined whether the scores of multiple-choice reading comprehension tests correlated with those of multiple-choice rational cloze tests.

The results of the experiment shows that there are significant correlations between the two, though there are serious problems about the reliability of each test, and that the more items, the higher the correlations between the two kinds of tests. To overcome these problems, it is necessary to choose suitable materials and appropriate items by administering pre-tests. What is more important, we should try to conduct additional research to obtain larger samples from greater number of subjects.

This study is quite limited, but significant correlations between the two types of reading tests imply that this multiple-choice rational cloze test can be used as one type of reading comprehension tests for entrance examinations.

BIBLIOGRAPHY

Backman, Lyle F. "The Trait Structure of Cloze Test Scores." *TESOL Quarterly*, 30, No.1 (1980) , 59–76.

———. "Performance on Cloze Tests with Fixed-Ratio and Rational Deletion." *TESOL Quarterly*, 19, No.3 (1985), 535–56.

Bensoussan, Marsha and Rachel Ramraz. "Testing EFL Reading Comprehension Using a Multiple-Choice Rational Cloze." *The Modern Language Journal*, 68, No.3 (1984), 230–39.

Heaton, J. B. *Writing English Language Tests*. Essex: Longman, 1975.

Henning, Grant. *A Guide to Language Testing: Development, Evaluation, Research*. Cambridge, Mass.: Newbury House, 1987.

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

———, and C. A. Conrad. "The Cloze Technique and ESL Proficiency." *Language Learning*, 21, No.2 (1971), 183–95.

Sato, Shiro. *The Role of Cloze Testing in English Teaching*. Tokyo: Nanundo, 1988.

Shimizu, Yuko. "A Study on Correlation between Cloze Tests and STEP Written Tests." *STEP Bulletin*, Vol.1 (1989), 103–16.

Stubbs, Joseph B., and G. Richard Tucker. "The Cloze Test as a Measure of English Proficiency." *The Modern Language Journal*, 58, No.5–6 (1974), 293–41.

Appendix A Sample of the 25-item M-C Rational Cloze Text and Answers

CHOOSE THE MOST SUITABLE WORD(S) TO FILL IN EACH BLANK:

Television now plays such an important part in so many people's lives that it is essential for us to try to decide whether it is a blessing or a curse. Obviously television has both advantages and disadvantages. But do the former outweigh the latter?

In the first place, television is not only a convenient source of entertainment, but also a comparatively (1) one. For a family of four, (2), it is more convenient as well as cheaper to sit comfortably at home, with practically (3) entertainment avaiable, than to go out (4) amusement elsewhere. There is no transport to arrange. They do not have to find a baby-sitter. They do not have to (5) for expensive seats at the theater, the cinema, the opera or the ballet, only to discover, perhaps, that the show is a (6) one. (7) they have to do is turn a knob, and they can see plays, films, the opera, and shows of every kind, (8) political discussions and the latest exciting football match. Some people, however, maintain that this is precisely where the (9) lies. The television viewer need do (10). He does not even use his legs. ...

Answer Sheet

No._______ Name_______________________

	A	B	C	D	Answer
1	expensive	cheap	time-saving	domestic	(B)
2	for example	however	also	in fact	(A)
3	unlimited	complete	popular	consistent	(A)
4	in favor of	in return for	in honor of	in search of	(D)
5	arrange	pay	buy	work	(B)
6	appropriate	rotten	interesting	curious	(B)
7	How	What	That	All	(D)
8	so to speak	not to mention	not to say	not only	(B)
9	excitement	comfort	danger	merit	(C)
10	nothing	something	anything	whatever	(A)