Multiple Choice (M-C) Cloze Tests

Akihiko Mochizuki

Naruto University of Education

This study examines whether that which M-C cloze tests measure is equivalent to that which listening comprehension tests or dictation tests measure. The following tests were administered: an M-C cloze test and a dictation test to 82 college freshmen in November, 1989; a cloze test and a dictation test to 44 college sophomores in November, 1989; an M-C cloze test, a listening test and a dictation test to 43 college sophomores in April, 1990; an M-C cloze test, a listening test and a dictation test to the same 43 college sophomores in June, 1990. Results indicate the following things: First, what the M-C cloze tests measure is different from what the listening comprehension tests measure. Second, what the M-C cloze tests measure does not seem to be equivalent to what the dictation tests measure. Third, what the cloze test measures is to a considerable degree but not a larger degree, equivalent to what the dictation test measures. Fourth, an analysis of reliability in M-C cloze tests makes it necessary to be careful about the choice of the texts for the tests in constructing M-C cloze tests. This study reveals that an M-C cloze test has great potential as an efficient means of measuring a language learner's overall language proficiency.

1. Modified Cloze Tests

John Oller divided tests into two kinds. One is discrete-point tests, which test only one discrete point, namely one point of grammar, phonology, vocabulary, etc. at a time; e.g. Auditory Comprehension test (i) (iy) in ship, sheep. The other is tests of integrative skills, which test full-scale language use, and language use in communication; e.g. Dictation, Cloze Tests.

It is 37 years since Wilson Taylor developed a cloze test, a type of test which is constructed by deleting every nth word from a passage and requiring the subject to fill in the blanks. Initially he used cloze tests to measure readability of prose. Now cloze tests are regarded as a measure of various aspects of language usage: reading compre-
hension, listening comprehension, the quality of translation of technical training manuals used by the military, knowledge of vocabulary, the reader's I.Q. and even oral ability².

Cloze tests are divided into two types: "standard cloze" tests and "modified cloze" tests. In "standard cloze" tests, words are deleted mechanically regardless of their function or meaning. In "modified cloze" tests words are deleted rationally on the basis of some criteria of the test writer. The latter type of tests draw much attention in terms of easier scoring and less testing time. D. Porter³ introduced the modified cloze procedure. It provided the subject with multiple choice alternatives, the construction of which depends on the depth of linguistic attainment and fineness of stylistic discrimination of the subject. For example, at a more advanced level the incorrect alternatives may be grammatically but not semantically appropriate, or vice versa. J. Jonz⁴ developed the M-C cloze and showed many merits in regard to testing time, scoring and reliability, compared with an established placement exam. Oscar Ozete⁵ administered adaptations of Carver's (1973) reading-input tests, a modified version of the cloze procedure, which required the subject to select systematically between correct answers and distractors, and he showed that reliabilities (0.68 and 0.57) were found to be good to moderate. Marsha Bensoussan and Rachel Ramraz⁶ developed a fill-in test, a multiple-choice rational cloze test focusing on reading comprehension, not grammar, and showed that the fill-in test measured up to the traditional multiple-choice test statistically.

I have been interested in multiple-choice (M-C) cloze tests since I read articles on cloze tests at Michigan State University in 1977. I have written four articles on effectiveness of M-C cloze tests, each of which I will summarize. In "TESL in the U. S. and its Application to English Teaching in Japan"⁷, I stated that there was a higher correlation between my 91-item M-C cloze test and a standardized exam (r=0.617) conducted on high school students than between Jonz's 33-item M-C cloze test and another standardized test (r=0.442).

In "Effectiveness of Multiple-Choice (M-C) Cloze Tests"⁸, I argued that there was a high correlation between my M-C cloze score and an English Proficiency Exam on 89 first-year high school students (r=0.709) and that the reliability of my M-C cloze test was high (r=0.708).

In "Effectiveness of Multiple-Choice (M-C) Cloze Tests(2)"⁹, I wrote that the correlation between the score of the M-C cloze based on Narration (one of 4 kinds of compositions: Explanation, Argumentation, Description, and Narration) and the score of the exam which extended over the work of the whole year was higher (r=0.756) than that between any other kind of M-C cloze and the exam which extended over the work of the whole year.

In "Effectiveness of Multiple-Choice (M-C) Cloze Tests(3)"¹⁰, I concluded that what
the M-C cloze tests measure is to a considerable but not larger degree equivalent to that which the Cloze tests measure, especially in the pair based on Narration, \( r = 0.675 \), and the M-C cloze tests are more reliable than Cloze tests.

It is necessary to explain in more detail the above-mentioned four kinds of compositions: (1) Explanation means a composition which explains things as they are and in which the author does not express his own feelings, such as objective explanations on the activities performed on Thanksgiving Day as opposed to personal reminiscences. (2) Argumentation refers to a composition in which the author tries to convince the readers of his point of view. The purpose is overtly persuasive and the subject matter may deal with issues such as criticisms on art, literature, etc. (3) Description means a composition which describes things or persons, places, and so on in detail, in accordance with the author's impressions and feelings, and does not so much inform the readers as appeal to their feelings, e.g. a composition in which the author observes persons, places, things and so on closely, and describes them in detail. (4) Narration means a composition that narrates what happened either in reality or in the imaginary world, e.g. newspaper articles that inform the readers of the facts, or novels. Up to 1988, I had conducted experiments about effectiveness of M-C cloze tests on senior high school students, but from now on I intend to deal mainly with adult learners of college level as subjects of my experiments. In this article, I examined whether that which M-C cloze tests measure is equivalent to that which listening comprehension tests or dictation tests measure.

2. Experiments on Effectiveness of M-C Cloze tests.

(1) Purpose:
To examine whether that which M-C cloze tests measure is equivalent to that which listening comprehension tests or dictation tests measure.

(2) Hypotheses:
1. The correlation between the scores of M-C cloze tests and those of listening comprehension tests is low.
2. The correlation between the scores of M-C cloze tests and those of dictation tests is high.
3. The correlation between the score of a cloze test and that of a dictation test is high.
4. The reliability of Jonz's M-C cloze test is high.

(3) Method:
i) Subjects:
44 second-year students of Naruto University of Education in Nov., 1989
43 second-year students of Naruto University of Education from April through June, 1990
82 first-year students of Naruto University of Education in Nov., 1989
ii) Materials:

I administered 11 tests as follows:

1. The 50-item M-C cloze test No.1 which used Narration, "The Lock Keeper" in
   *MULTI-LEVEL Reading Program*, 1971, YELLOW, Goken. Every 8th word was
   deleted in the 413-word text.

2.3. The 106-word dictation tests No.1a and No.1b. (*Dictation test No.1a is the
   same as Dictation test No.1b*).

4. The 50-item cloze test No.1 which used Narration, "The Lock keeper" from the
   same text as 1. Every 8th word was deleted in the 413-word text.

5. The 58-item M-C cloze test No.2 which used Narration, "Big Buster's House" in
   *MULTI-LEVEL Reading Program*, 1971, OLIVE, Goken. Every 9th word was
   deleted in the 568-word text.

6. The 40-item listening comprehension test No.1 which used Part I in TOEFL

7. The 48-item dictation test No.2 which used Monologue I, Unit 7 in *Listening

8. The 40-item listening comprehension test No.2 which used Part II in the same
   material as 6.

9. The 58-item M-C cloze test No.3 which used Narration, "The Tree" in the same
   book as 5. Every 9th word was deleted in the 563-word text.

10. The 51-item dictation test No.3 which used Monologue II, Unit 10 in the same
    book as 7.


The above-mentioned M-C cloze tests No.1, No.2 and No.3 were constructed upon the
same principles as in *Research Bulletin English Language Education Society*, Vol.10
(1980, p.155): 1. There are four alternatives for each word. 2. Of the four alternatives,
two words are alike syntactically and semantically. Only one of them is the correct
answer, and the other is incorrect since it does not fit the context. 3. One out of the four
is completely wrong syntactically and semantically, and can easily be found to be
incorrect. The remaining choice is incorrect from the syntactic point of view although
it is apparently associated with the sentence semantically. In establishing the above
principles, I relied upon the advice of Prof. Paul Munsell at Michigan State University.
I am grateful to him for his help.

The cloze test No.1 was written by deleting every 8th word mechanically.

(4) Procedure:

I administered the following tests:

The M-C cloze test No.1 and the 7-minute dictation test No.1a to two classes of
first-year students totaling 82 at the end of November, 1989.
The cloze test No.1 and the 7-minute dictation test No.1b to a class of second-year students totaling 44 at the end of November, 1989.

The M-C cloze test No.2, the 30-minute listening comprehension test No.1, and the 5-minute dictation test No.2 to a class of second-year students totaling 43 at the end of April, 1990.

The M-C cloze test No.3, the 30-minute listening comprehension test No.2, the 5-minute dictation test No.3 and Jonz’s M-C cloze test to the same 43 second-year students at the end of June, 1990.

The time it took for the subjects to complete the M-C cloze test was consistently the number of test items divided by two measured in minutes, e.g. the 58-item M-C cloze test No.2 took about 29 minutes. In contrast, the minutes required to complete the cloze test were about equal to the number of test items, i.e. the 50-item cloze No.1 took about 50 minutes.

In constructing M-C cloze tests, I kept in mind that the level of these tests should remain constant with each other. Therefore, I saw to it ①that the number of words used in the text would be approximately the same in M-C cloze No.2 and M-C cloze No.3, and ②that on the test paper there would be several notes of unfamiliar words or phrases which were difficult for the subjects to understand.

(5) Results:

**Table 1** Correlation between M-C Cloze tests and Listening Comprehension Tests

<table>
<thead>
<tr>
<th>Test Combination</th>
<th>Correlation Coefficient</th>
<th>Significance Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>M-C cloze No.2 and Listening No.1</td>
<td>0.142</td>
<td>p&lt;0.4</td>
</tr>
<tr>
<td>M-C cloze No.3 and Listening No.2</td>
<td>0.050</td>
<td>p&lt;0.8</td>
</tr>
</tbody>
</table>

(N = 43)

The average correlation coefficient was 0.096. It shows that there was almost no correlation between the M-C cloze tests and the listening comprehension tests.

**Table 2** Correlation between M-C cloze tests and Dictation tests

<table>
<thead>
<tr>
<th>Test Combination</th>
<th>Correlation Coefficient</th>
<th>Significance Level</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>M-C cloze No.1 and Dictation No.1a</td>
<td>0.332</td>
<td>p&lt;0.005</td>
<td>82</td>
</tr>
<tr>
<td>M-C cloze No.2 and Dictation No.2</td>
<td>0.142</td>
<td>p&lt;0.4</td>
<td>43</td>
</tr>
<tr>
<td>M-C cloze No.3 and Dictation No.3</td>
<td>0.249</td>
<td>p&lt;0.1</td>
<td>43</td>
</tr>
</tbody>
</table>

There was a low correlation between the scores of the M-C cloze test and the Dictation test in each pair. The highest correlation was between the M-C cloze test No. 1 and the Dictation test No.1a; the second highest was between the M-C cloze No.3 and the Dictation test No.3; the lowest was between the M-C cloze test No.2 and the Dictation test No.2. The average correlation coefficient was 0.256.
Table 3 Correlation between a Cloze test and a Dictation test

<table>
<thead>
<tr>
<th>Cloze Test No.1 and Dictation No.1</th>
<th>N=44</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.472 (p&lt;0.002)</td>
<td></td>
</tr>
</tbody>
</table>

There was a considerable correlation between the score of the Cloze test and the Dictation test, although it was not high.

Table 4 Reliability Coefficients by K.R.21

<table>
<thead>
<tr>
<th></th>
<th>r</th>
<th>Mean</th>
<th>SD</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jonz's M-C cloze</td>
<td>0.435</td>
<td>24.186</td>
<td>3.343</td>
<td>43 (S)</td>
</tr>
<tr>
<td>M-C cloze No.1</td>
<td>0.559</td>
<td>32.585</td>
<td>5.010</td>
<td>82 (F)</td>
</tr>
<tr>
<td>M-C cloze No.2</td>
<td>0.375</td>
<td>33.279</td>
<td>4.741</td>
<td>43 (S)</td>
</tr>
<tr>
<td>M-C cloze No.3</td>
<td>0.506</td>
<td>38.116</td>
<td>5.095</td>
<td>43 (S)</td>
</tr>
<tr>
<td>Cloze No.1</td>
<td>0.402</td>
<td>19.977</td>
<td>4.449</td>
<td>44 (S)</td>
</tr>
<tr>
<td>Dictation No.1a</td>
<td>0.865</td>
<td>62.659</td>
<td>13.382</td>
<td>82 (F)</td>
</tr>
<tr>
<td>Dictation No.1b</td>
<td>0.892</td>
<td>65.386</td>
<td>14.666</td>
<td>44 (S)</td>
</tr>
<tr>
<td>Dictation No.2</td>
<td>0.606</td>
<td>24.209</td>
<td>5.433</td>
<td>43 (S)</td>
</tr>
<tr>
<td>Dictation No.3</td>
<td>0.690</td>
<td>25.442</td>
<td>6.274</td>
<td>43 (S)</td>
</tr>
<tr>
<td>Listening No.1</td>
<td>0.233</td>
<td>16.372</td>
<td>3.344</td>
<td>43 (S)</td>
</tr>
<tr>
<td>Listening No.2</td>
<td>0.156</td>
<td>15.465</td>
<td>3.182</td>
<td>43 (S)</td>
</tr>
</tbody>
</table>

* (S) = Sophomores, (F) = Freshmen. The number of () in Mean = a full score

The above tests are placed in the order of reliability coefficients, from highest to lowest: ①Dictation No.1b (r=0.892), ②Dictation No.1a (r=0.865), ③Dictation No.3 (r=0.690), ④Dictation No.2 (r=0.606), ⑤M-C cloze No.1 (r=0.559), ⑥M-C cloze No.3 (r=0.506), ⑦Jonz's M-C cloze (r=0.435), ⑧Cloze No.1 (r=0.402), ⑨M-C cloze No.2 (r=0.375), ⑩Listening No.1 (r=0.233), ⑪Listening No.2 (r=0.156).

It might be difficult to give a definite conclusion, since the numbers of subjects are different and the ages of the subjects are different—first-year college students and second-year students. However, Table 4 shows several tendencies: ①The dictation tests were the most reliable of all the eleven tests. ②Of the four M-C cloze tests, three were higher in the reliability coefficient than the cloze test. ③The listening tests were the lowest in the reliability coefficients. What is most noteworthy is that although Jonz claimed that the reliability coefficient of his M-C cloze “Poor Simpleton” was 0.76 my experiment showed that it stood at 0.435.

Considerations:

Hypothesis 1 can be proved. As Jonz (1976) remarked, the correlation between the scores of M-C cloze tests and those of listening comprehension tests was low. It means that that which M-C cloze tests measure is different from that which listening comprehension tests measure.
Hypothesis 2 cannot be proved. Cloze tests and dictation tests are tests of integrative skills, which measure full-scale language use and language use in communication. An M-C cloze test is a modified type of a cloze test. Therefore, I expected that that which M-C cloze tests measure is equivalent to that which dictation tests measure. However, what Table 2 displayed was contrary to my expectation. The correlations between M-C cloze tests and dictation tests were low: r=0.333, r=0.294, or almost non-existent: r=0.142.

Hypothesis 3 can be half proved. Just as I mentioned earlier, cloze tests and dictation tests are measures of the learners' overall proficiency. I expected a high correlation between a cloze test and a dictation test. But what Table 3 displayed fell short of my expectation. The correlation between the cloze test and the dictation test was not high (r=0.472).

Hypothesis 4 cannot be proved. Jonz (1976) showed that with only thirty-three items, his M-C cloze test maintained what was comparable to the reliability of an approximately three-hour placement examination (r=0.76 by K.R.29). In his test, the number of subjects was 33 (college students of ESL courses). The mean score was 21.39 out of 33 possible. I conducted the same 33-item M-C cloze test on 43 sophomores of my college in June, 1990. As a result, the reliability coefficient turned out to be 0.435 (by K.R.29). I wonder where this sizeable difference in reliability coefficients comes from, although they were calculated in slightly different ways, that is, by K.R.29 on one hand and by K.R.29 on the other hand.

This experiment leads me to consider a way of improving reliability coefficients in M-C cloze tests. As Table 4 displays, out of three M-C cloze tests which I constructed, two were higher in reliability coefficients than Jonz's 33-item M-C cloze test—namely, 0.559, and 0.506, as opposed to 0.435. Jonz's M-C cloze test was constructed on the basis of rational deletion and item and error analyses, whereas mine was constructed on the basis of mechanical deletion and the principles that I mentioned previously. Next, although I (1990) conducted the same M-C cloze test which used Narration, a short 413-word story, on 95 second-year high school students and 82 first-year college students, the reliability coefficients were different—that is, 0.584 for high school students, and 0.559 for college students.

In order to see what kind of M-C cloze tests would show high reliability for college students, I constructed two M-C cloze tests, both of which used Narration, a 568-word story for M-C cloze No.2 and a 563-word story for M-C cloze No.3, in the same college level reading book, and in both of which every 9th word was deleted. I used longer and probably more difficult passages in the M-C cloze tests No.2 and No.3 than in the M-C cloze No.1, in order to make M-C cloze tests more reliable. However, the reliability coefficients were not high, that is, 0.375 for M-C cloze No.2, and 0.506 for M-C cloze No.
3. On the contrary, the M-C cloze test No.1 which used a shorter and easier (413-word) story than the M-C cloze tests No.2 and No.3 showed a higher reliability coefficient, which was 0.559.

Our attention is drawn to two facts: First, the reliability coefficients of the dictation tests No.1a and No.1b were high. Both dictation tests used the same 5 short passages made up of a total of 106 words. I scored only correct words out of 106 possible. The subjects listened to the pre-recorded tape three times; the first time, they were instructed only to listen, the second time, to transcribe the passages, and the third time, to correct their mistakes, which took 7 minutes in all. Second, the reliability coefficients of the listening tests No.1 and No.2 were low. Since I used TOEFL practice listening comprehension tests which took 30 minutes for each test, I expected their reliability coefficients to be a little higher, but they were not—that is, 0.233 for Listening No.1 and 0.156 for Listening No.2. I wonder whether the listening comprehension tests were too difficult for the subjects, or whether the subjects were poor in listening. It comes to my mind that in Farhady’s study (1979), Japan came in 11th out of 12 countries in listening comprehension of the UCLA English as a Second Language Placement Examination (ESLPE) which was conducted on 800 foreign students.

Conclusion

In this research I examined whether that which M-C cloze tests measure is equivalent to that which listening comprehension tests or dictation tests measure. The results of the experiment showed the following things: First, what the M-C cloze tests measured was different from what the listening comprehension tests measured. Second, what the M-C cloze tests measured did not seem to be equivalent to what the dictation tests measured. Third, what the cloze test measured was to a considerable degree but not a larger degree, equivalent to what the dictation test measured. Fourth, there is some room for improving the way of constructing M-C cloze tests. In view of the facts that the M-C cloze tests No.1 and No.3 that I constructed indicated higher reliability than Jonz’s M-C cloze test, and that two M-C cloze tests (No.1 and No.3) showed higher reliability than the cloze test No.1, M-C cloze tests deserve further research for proof that they are a useful and valid means of measuring a language learner’s overall language proficiency.

Appendix A Sampling of the 50-item M-C Cloze No.1 Text and Answers

The Lock-Keeper (The 50-item M-C Cloze No.1)

Word Building

<table>
<thead>
<tr>
<th>Lock 「運河の水門」</th>
<th>sergeant 「巡査部長」</th>
</tr>
</thead>
<tbody>
<tr>
<td>as well 「〜もまた」</td>
<td>body 「死体」</td>
</tr>
</tbody>
</table>

38
It was evening as I walked along a path near the canal.
Ahead of me was a lock, its gates were closed. The lock-keeper stood outside a small, grey house, watching the canal.

Then, I saw a woman in a white coat, walking across the lock gates. As I went, she slipped and with a small cry, into the water below the gate.

I was running, shouting as I went. The lock-keeper, ran forward from his house - but then stopped. He was still standing when reached the lock.

Notes:
This is a revised version of a paper which is to be printed in Research Bulletin of Educational Sciences, Naruto University of Education, Vol.6, 1991.
6 Bensoussan, M. & Rachel Ramraz, “Testing EFL Reading Comprehension Using a


**References:**


