C-Tests
Four Kinds of Texts, Their Reliability and Validity

Akihiko Mochizuki
Aichi University of Education

This study examines test appropriateness for constructing the most effective C-Test. The following tests were administered to 42 college freshmen from April 1992 through May 1992: the Second Grade Test of the Society of Testing English Proficiency (STEP), *A Comprehensive English Language Test for Learners of English* (CELT) Listening, a Dictation test, and four C-Tests whose tests used Narration, Explanation, Description, and Argumentation. Results indicate the following: First, the reliability of the Narration C-Test is the highest ($r = 0.928$). Second, there is a fairly high correlation between the scores of the STEP test and the Narration C-Test. Third, there is a very low correlation between the scores of C-Tests and the CELT Listening Comprehension test. Fourth, there is a very low correlation between the scores of the C-Tests and the Dictation test. The study therefore indicates that a C-Test which uses a long narration text seems to be a promising means of measuring a language learner's overall language proficiency, and what a C-Test measures seems different from what a listening test and a Dictation test measure. Further research is needed to investigate the correlations between the Narration-based C-Test and a more reliable criterion test like the TOEFL.

4種のテクストによるクローズ・テスト：信頼性と妥当性

この研究は、最も効果的なクローズ・テスト（C-テスト）を開発するために、使用するテクストの適切さを検証する。物語、説明、記述、論述の4種のテクストを使用したC-テストと、STEP、CELT、およびディクテーション・テストを42名の大学1年生に実施した結果、長い物語文を使用したC-テストは学習者の総合的言語能力を測るのに適しているらしいこと、C-テストが測っている能力は、聞き取りテストとディクテーション・テストの測る能力とは異なるらしいことがわかった。今後は、物語文を使ったC-テストとTOEFLなどのより信頼性の高い基準との相関について研究する必要がある。

In 1953 Taylor (1956) invented the cloze test, in which every *n*th word is deleted mechanically from a passage of appropriate difficulty and the examinee is asked to fill in the blank with a suitable word. The cloze test, which is sometimes modified into a rationally deleted cloze test or a multiple-choice cloze test, has been gaining steady popularity in both ESL and EFL programs as a measure of a student's overall proficiency in the target language. The use of a cloze procedure has been suggestion as a possible alternative method of ESL placement as Hinofotis (1983) and Oller and Conrad (1971) report, and as an alternative or supplementary method of EFL placement, as Heilenman (1983) reports.

Problems with Cloze Tests

Recently, however, problems with the cloze test have been pointed out by several researchers. These problems are roughly divided into the following categories:

1. Types of measure. Lado (1986) administered the Oller-Conrad (1971) 50-item cloze test to 54 graduate and undergraduate ENL (English as a native language) students at Georgetown University. As a result of his analysis, he supported Carroll et al.'s (1959) conclusion that the cloze procedure may be a suitable testing device to assess group differences in ESL, but it is inadequate as a measure of individual differences. He argues as follows: that only one subject out of 54 (2%) obtained above 70 percent (the minimum passing score in academic courses); that the cloze test deals with a narrow sample of a single register of English because it lacks the formation of yes/no questions, requests, and so forth; that the cloze procedure does not encourage high level thinking, since unlike ordinary reading for meaning, it requires the examinee to use the context to search for specific words missing in the test; that the cloze procedure precludes the application of such good psychometric practice as the arrangement of items from easy to difficult; that the revision and rational selection of test items on the basis of experimental administration of the initial forms of the test of typical students are lacking; and that the majority of the examinees (87%) reacted negatively to the test.

2. Scoring methods. Brown (1980) administered the UCLA ESL Placement Examination, a 50-item cloze test, and a multiple choice cloze test to 112 UCLA students. After comparing the four scoring methods-multiple-choice, exact answer, acceptable answer, and clozentropy, his conclusion was that the best overall scoring method was acceptable answer

scoring. However, when acceptable scoring is adopted, the advantages of easy preparation and scoring are canceled.

3. Reliability and validity. Porter (1978) states that subjects' scores may vary according to where the deletion starts in the cloze test. Alderson (1979) and Klein-Braley (1981) are in general accord with the findings of Klein-Braley and Raatz (1984):

> Systematic *n*th-word deletion does not necessarily produce a random sample of the elements of the test. Different deletion rates and starting points applied to the same text produce tests which can differ very considerably in difficulty, reliability, and validity, and particularly for homogenous samples (classroom groups or monolingual groups) cloze tests tend to have unsatisfactory reliability coefficients. (p. 135)

Klein-Braley (1983) examined a total of 22 cloze test and could not find a single pair of reliable and parallel tests, which leads her to conclude that there may be no such things as cloze equivalent across tests.

4. What cloze tests measure. There are three main views on what cloze tests measure: (a) cloze tests and integrative tests cannot be distinguished statistically from discrete-point tests because of lack of statistically difference between the former and the latter (Farhady, 1979); (b) cloze tests measure only basic skills, because they are correlated more closely with grammar tests than with reading tests (Alderson, 1979); and (c) cloze tests measure overall proficiency, because they are closely correlated to dictation, reading tests, and essay writing besides standardized proficiency tests (Chavez-Oller et al., 1985).

C-Tests

In response to these problems, a new type of cloze test, the C-Test, was developed by Raatz and Klein-Braley (1981).

Format of the C-Test. In the C-Test, the second half of every second word is deleted instead of whole word, or according to the "rule of 2."

Advantages of the C-Test. First, in cloze tests, unpredictable results are obtained by various fixed-ratio deletion procedures, whereas the C-Test procedure meets the random sampling requirement because the deletion of "half words" results in a random, representative sample of parts of speech being so affected (Klein-Braley, 1985). Second, the test performance of the cloze test is affected by the text topic and difficulty, whereas in the C-Test it is minimized by the use of several different short texts. Third in cloze tests, even native speakers can hardly achieve

a maximum score, whereas, regarding the C-Test, Klein-Braley (1985) says "Adult educated native speakers achieve virtually perfect scores" (p. 84). Fourth, students find C-Tests less frustrating than cloze tests.

Disadvantages of the C-Test. What the C-Test measures is open to questions. Klein-Braley (1985, p. 100) states that "recognition of syntactical relationship comes first" in completing a given word, though semantic processing is indispensable for perfect performance. The C-Test is claimed to be a measure of overall language proficiency. However, it appears to indicate more grammatical than contextual competence, as Carroll (1986) states that the C-Test "harks back in many ways to the form of word completion test devised by the German psychologist Ebbinghaus (1987)," and further that "it seems to be limited to the measurement of general proficiency, chiefly at lower levels of ability, in written language" (p. 128).

Klein-Braley and Raatz (1984) suggest a list of criteria which the C-Test should meet: (a) it should use several different texts; (b) it should have at least 100 deletions; (c) adult native speakers should obtain virtually perfect scores; (d) the deletions should affect a representative sample of the text; (e) only exact scoring should be possible; and (f) the test should have high reliability and validity. They consider the test to be satisfactory if its reliability by Cronbach's alpha reaches 0.8 or higher, and its empirical validity (correlation with the criterion) is at least 0.5 (p. 136). They state that the C-Test is norm-oriented, which means the subject group should score an average of 50 percent on the test (p. 144).

Klein-Braley (1985) states that reduced redundancy tests, like cloze tests and C-Tests, have as a cornerstone in their underlying theory random sampling of the elements (p. 180). Further, she claims that *n*th-word deletion and random-word deletion in cloze tests are not the same, whereas "C-Test deletion (at least in the 100 English and 100 German tests we examined) does produce random samples of the word classes of the text involved" (p. 84).

The C-Test has a short history and many things remain to be explored. The research so far has not dealt with what kind of text produces a higher reliability and validity, and what exactly the C-Test measures. As for the former problem, Mochizuki (1984) shows that the correlation between the Narration M-C (Multiple-Choice) cloze test and a criterion test is higher than that between any other kind (i.e., Explanation, Discourse, or Description) of M-C cloze and the criterion test. As for the latter problem, Darnell (1968) shows that the highest correlation of cloze test scores with a various parts of the TOEFL is with the listening comprehension section (0.73). Oller and Conrad (1971) show that

there is a high correlation (0.82) between the cloze test and Dictation. Therefore, I assume that the Narration C-Test will produce a higher reliability and concurrent validity, and that the correlation between the score of the C-Test and the scores of Dictation and the Listening Comprehension might be high. The following hypotheses were set up:

H1. The reliability of the Narration C-Test will be the highest of the four kinds of texts (i.e., Narration, Explanation, Argumentation, and Description).
H2. There will be a high correlation between the score of the Narration C-Test and that of a placement test.
H3. There will be a high correlation between the score of the Narration C-Test and that of a listening test.
H4. There will be a high correlation between the score of the Narration C-Test and that of a Dictation test.

It is hoped that the C-Test will be "efficient" in terms of reliability and concurrent validity on the one hand, and in terms of practicality (being easy to write and conduct) on the other. For this to be true, the reliability of the C-Test should exceed the critical threshold level of 0.8, and the test should correlate with the reliable discrete point test at 0.5 or higher, as claimed by Klein-Braley and Raatz (1984, p. 136). At the same time the test writers, especially secondary school teachers who are very busy with day-to-day activities, should be exempt from the fairly difficult burden or selecting several kinds of texts for the C-Test, and should only have to use one kind of text for the C-Test.

Method

Purpose: The purposes of the study were to determine the most appropriate kind of text for making the C-Test most effective, and to define more exactly what the C-Test measures.

Subjects: The experiment was conducted from April, 1992 through May, 1992 at Naruto University of Education. The subjects were 42 first-year students who were enrolled in an undergraduate class in general English.

Materials: For this study, passages which were longer than Klein-Braley's suggestions (approximately 400 words) were used in the C-Tests. Four kinds of texts were represented, as in Mochizuki (1991):

1. *Explanation*: a passage which explains things as they are and in which the author does not express personal feelings, such as objective explanations of the activities performed on Thanksgiving Day.
2. *Argumentation*: a passage in which the author tries to convince the readers to adopt a particular point of view. The purpose is overtly persuasive and the subject matter may deal with issues such a criticism of art or literature.
3. *Description*: a passage which describes things, persons, or places in detail, in accordance with the author's impressions and feelings, and does not so much inform the readers as appeal to their feelings.
4. *Narration*: a passage that narrates something which happened either in reality or in the imaginary world, for example, excerpts from newspaper articles or novels.

The following were the materials used in this experiment:

1. A 66-item placement examination (STEP) composed of an assortment of items from past second grade STEP written examinations, which the subjects were allowed 55 minutes to complete.
2. A 50-item listening comprehension test (CELT, Harris & Palmer, 1986) which the subjects were allowed approximately 25 minutes to complete.
3. A 104-word dictation test (Dictation), made up of approximately 20-word passages from parts of the Shizuoka Prefecture Standardized Tests for second- and third-year high school students.
4. Four 120-item C-Tests[1] (C-Tests), in each of which the first few and the last few sentences were left intact and the second half of every second word was deleted, and for each of which on the basis of pretesting experience, 35 minutes were allowed for completion.

The four C-Tests were constructed and marked using the following principles:

1. The second half of every second word was deleted. In blanks composed of odd-numbered words (the number of the deleted in n), the subject is required to fill in the blanks with ([n - 1]/2) and ([n + 1]/2) numbered words alternately; for example, st*out*(1)... ph*one*(2)...mou*th*(3)...ov*er*(4). In word (1) two letters are deleted, in word (2) three letters, in word (3) two letters, and in word (4) three letters.
2. Difficult words/phrases were explained in easier English or Japanese to facilitate the understanding of the passage.
3. Numerals/proper nouns (e.g., *5100* km, Mr. *James Stewart*) were

disregarded in counting every second word.

4. A misspelled word was regarded as correct, as long as the scorer realized that the subject understood the targeted word.

Procedure

In order to study the concurrent validity of the C-Tests in question, a criterion test had to be specified. Therefore, CELT tests of vocabulary and structure and the STEP placement examination were administered to all the subjects in early April, 1992. The result revealed that the reliability of CELT was 0.638, whereas that of the STEP placement examination was 0.778 by the split-half method. Therefore, the STEP placement examination was chosen as a placement test for this study. The subjects were tested on the CELT listening test, the Dictation test, and the four C-Tests from April through the end of May, 1992.

After the test was administered, the test papers were exchanged between students and scored in unison following the teacher's comments. After the test papers had been collected, they were looked over and the miscalculations of the points of those tests were corrected by the teacher. It must be noted that because the C-Test was a new procedure for the students, their performance improved over time. Hence, the difference in their scores on the different types of passages may also reflect their familiarity with the testing format.

Results

The reliability coefficients were calculated as shown in Table 1. In this study, the split-half method was used for their calculation. In assessing the reliability of individual C-Test texts, the use of the split-half method or the KR-20, or the KR-21, or Cronbach's alpha assumes that the items are independent (i.e., that the test may be split into two independent halves). Klein-Braley and Raatz (1984) used each of the various texts as "superitem" (p. 136) without analyzing individual items, thereby avoiding the problem of item independence. Indeed there is some argument about whether cloze tests are sensitive to language constraints across sentences and can be completed considering only the context of the sentence. However, it seems that the cloze test as a measure of higher level skills and overall proficiency is finding approval (Brown, 1989). Likewise the C-Test appears to be a measure of grammatical competence rather than of textual competence. However, validity research suggests that the C-Test is a measure of overall language profi-

ciency, as is shown in Stansfield and Hansen (1983). So, the blanks could be considered to be independent, which means that the use of the split-half method is permissible as a measure of the reliability of the C-Test.

The reliability coefficients of all the tests except the CELT Listening Comprehension test were high or very high (Table 1). The four C-Tests are placed in the order of reliability coefficients, from highest to lowest: Narration (r = 0.928); Explanation (r = 0.904); Description (r = 0.899); and Argumentation (r = 0.860). The differences in mean scores among the four kinds of C-Tests show that the man score of the Narration C-Test is the highest (X = 70.095), and that of the Explanation C-Test the second highest (X = 65.881); in other words, learners perform better on passages which have a temporally ordered sequence of events. Only Narration fully meets this requirement, although Explanation does involve a sequential element.

Table 1.
Reliability Coefficients by Split-Half Method (n = 42)

Test	r	Mean	Full Score	*SD*
STEP Placement	0.780	70.619	160	18.789
CELT Listening	0.527	19.238	50	4.741
Dictation	0.949	62.095	104	15.173
C-Test No. 1	0.928	70.095	120	14.890
C-Test No. 2	0.904	65.881	120	14.026
C-Test No. 3	0.899	47.786	120	12.342
C-Test No. 4	0.860	54.952	120	10.148

Oller and Conrad (1971) constructed a 50-item cloze test by deleting every seventh word from the roughly 350-word passage, "What is a College?" (McCall & Crabbs, 1961) to attempt to partially determine the discriminative power of the cloze test (scored by the exact word method), and its validity. They conducted the cloze test and *Form 2C* of the *UCLA ESL Placement Examination* on beginning, intermediate, and advanced ESL students, along with two control groups of native (ENL) speakers (freshmen and graduates, respectively). Their analysis of the differentiation of proficiency shows that although the data yield no significant spread between advanced ESL students and ENL college freshmen, or between ENL college freshmen and ENL graduate students, the differences between beginning ESL, intermediate ESL, and ENL graduates are significant. In their experiment, the passage "What Is a College?" which is categorized as

Argumentation, works well with a discriminator, whereas in this experiment Narration and Explanation seem to work well as such. The two data sets have one thing in common: in their experiment beginning ESL students did poorly in the Argumentation cloze test (X = 7.00 out of 50, the lowest of all the seven groups) and my students did not do well either (X = 54.952 out of 120, the second lowest of the four C-Tests).

Clearly, in each pair there was always a fairly high or low correlation between the score of the STEP placement examination and the C-Test. The correlation procedure used in Table 2 is the Pearson product-moment procedure, and the correlations given in the table are values for *r*. The four pairs of STEP placement examinations and C-Tests are placed in the order of correlation coefficients, from highest to lowest:

Table 2
Correlation Between C-Tests and STEP Placement (n = 42)

Tests	*r*	*p*
1. STEP Placement and C-Test No. 1	0.438	<0.005
2. STEP Placement and C-Test No. 2	0.357	<0.005
3. STEP Placement and C-Test No. 3	0.267	<0.1
4. STEP Placement and C-Test No. 4	0.213	<0.2

1. Narration C-Test No. 1 + STEP; 2. Explanation C-Test No. 2 + STEP; 3. Description C-Test No. 3 + STEP; 4. Argumentation C-Test No. 4 + STEP. It must be noted that reliability coefficients and correlations are lower when the mean is low and standard deviation is small, as is the case for the listening test reliability and correlation.

Table 3
Correlation Between C-Tests
and CELT Listening Comprehension (n = 42)

Tests	*r*	*p*
1. CELT Listening and C-Test No. 1	0.030	<0.9
2. CELT Listening and C-Test No. 2	0.125	<0.5
3. CELT Listening and C-Test No. 3	0.105	<0.6
4. CELT Listening and C-Test No. 4	0.043	<0.8

Table 3 reveals a very low correlation between the scores of the C-Test and the CELT Listening Comprehension test in each pair. The highest correlation was between the Explanation C-Test No. 2 and the CELT

Listening Comprehension test. The lowest correlation was between the Narration C-Test No. 1 and the CELT Listening Comprehension test.

Table 4
Correlation Between C-Tests and Dictation (n = 42)

Tests	*r*	*p*
1. Dictation and C-Test No. 1	0.298	<0.1
2. Dictation and C-Test No. 2	0.278	<0.1
3. Dictation and C-Test No. 3	0.180	<0.3
4. Dictation and C-Test No. 4	0.073	<0.7

Table 4 shows that there was always a low or very low correlation between the scores of the C-Test and the Dictation test in each pair. The four pairs are placed in the order of correlation coefficients, from highest to lower: 1. Narration C-Test No. 1 + Dictation; 2. Explanation C-Test No. 2 + Dictation; 3. Description C-Test No. 3 + Dictation; 4. Argumentation C-Test No. 4 + Dictation.

Discussion

In the introduction four hypotheses were set up. Let us examine whether each was supported or not.

Hypothesis 1, of the highest reliability of the Narration C-Test among the four kinds of C-Tests, was supported. As can be inferred from Mochizuki (1984), the Narration C-Test was found to show the highest reliability (r = 0.928). The C-Tests conducted in this experiment all showed very high or high reliability. Two of them, Narration and Explanation C-Tests, were highly reliable (0.928 and 0.904), while Description and Argumentation C-Tests were reliable (0.899 and 0.860).

Hypothesis 2, of a high correlation between the score of the Narration C-Test and that of the placement test, was not supported clearly. The fairly high or low correlations between the C-Tests and the STEP placement examinations show that what the C-Test measures is different from what the STEP placement test measures. This result fell short of expectations. As described in the introduction, before this experiment the subjects had taken CELT vocabulary and structure tests; the reliability of the combined test turned out to be fairly high but not as high as expected (r = 0.642 by split-half method). In order to determine accurate correlations between the Narration C-Test and a criterion test, a very reliable discrete-point type criterion test with 0.8 or higher reliabil-

ity coefficient is needed. This part must be further researched in the near future by using a reputable TOEFL-like test which is easily available to secondary school teachers in Japan.

Hypothesis 3, of a high correlation between the score of the Narration C-Test and that of a Listening test, was not supported at al. I had expected a fairly high correlation, but the results showed the reverse. This means that what the C-Test measures seems to be quite different from what the Listening Comprehension test measures. However, again, the problems of the low mean score and small standard deviation of the Listening test must be addressed. The low correlation between the score of the Listening test and that of the Narration C-Test might have been caused by that. Before this experiment, the 50-item SONY Aural Comprehension Test (1980) was administered, but the reliability was found to be very low ($r = 0.289$ by the split-half method). The reliability of the CELT Listening Comprehension test is moderate ($r = 0.527$ by the split-half method). A more reliable Listening Comprehension test is urgently needed. A further investigation of the correlation between the Narration C-Test and that of the Listening Comprehension test will be undertaken when a highly reliable Listening Comprehension test is obtained.

Hypothesis 4, of a high correlation between the score of the Narration C-Test and that of a Dictation, was not supported. What the C-Test measures seems to be very different from what the Dictation test measures. However, the result of this analysis must take into consideration the fact that the data presented in this study is extremely limited.

Conclusion

In this search I investigated what is the most appropriate kind of passage for making the C-Test most effective and what the C-Test measures. The results showed that, first, the C-Tests which used a long passage, especially Narration and Explanation texts were found to be very reliable (0.928 and 0.904). Second, the C-Test seems to measure something different from what the Listening Comprehension and Dictation tests measure. What the Narration-based C-Test measures, to a moderate degree, seems to be the same as what the placement test measures ($r = 0.438$). However, in order to confirm the correlations between the C-Tests, whether those tests are Narration, Explanation, Description and Argumentation, and criterion tests, more reliable discrete-point criterion and Listening Comprehension tests with a reliability of 0.8 or higher are needed.

What is noteworthy is that this study revealed that C-Tests with a long passage, especially the Narration text, were able to overcome the

critical threshold level of 0.8 and were correlated with a reliable discrete-point placement test at close of 0.5, which is what is claimed by Klein-Braley and Raatz (1984). The C-Test with a long passage and with a text of only one kind might work in secondary school classes in Japan because it is reliable and could turn out to be valid with the use of more reliable discrete-point tests, and because it requires less time to write than several short C-Tests. Finally, because of the far-reaching potential C-Tests, further research is needed on their effectiveness.

The author sincerely thanks Suteo Kimura, Naruto University of Education, and Ken'ichi Ohtomo, University of Tsukuba, for their comments on the handling of the statistical data.

Akihiko Mochizuki (M.A., Michigan State University) is an assistant professor of TEFL in the Department of English at Aichi University of Education. His research interests are cloze tests, listening comprehension, writing, foreign language teaching methodology, and English usage.

Note

1. C-Test No. 1, which used Narration, "The Lock Keeper" (413 words) (Kaneda, et al., 1971), C-Test No. 2, which used Explanation, "The Pony Express" (367 words) (Kaneda et al., 1971), C-Test No. 3, which used Description, "Alan Shepard Gets Set for the Moon" (391 words), adapted from the 1970 *Life* magazine article, and C-Test No. 4, which used Argumentation, "Anger" (346 words), are part of a college entrance examination workbook. Copies of the C-Tests used in this experiment are available from the author on request.

References

Alderson, C. (1979). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly, 13*, 219-226.

Brown, J. D. (1980). Relative merits of four methods for scoring cloze tests. *Modern Language Journal, 64*, 311-317.

Brown, J. D. (1989). Cloze item difficulty. *JALT Journal, 11*, 46-62.

Carroll, B. (1986). LT + 25, and beyond? Comments. *Language Testing, 3*, 123-129.

Carroll, B., Carton, S., & Wilds, C. (1959). An investigation of "cloze" items in the measurement of achievement in foreign languages. College Entrance Examination Board Research and Development Reports. Laboratory for Research in Instruction, Graduate School of Education, Harvard University. (ERIC Document Reproduction Service No. ED 021-513)

Chavez-Oller, M., Chihara, T., Weaver, K., & Oller, J. (1985). When are cloze items sensitive to constraints across sentences? *Language Learning, 35*, 181-203.

Darnell, D. K. (1968). The development of an English language proficiency test of foreign students using a clozentropy procedure. (ERIC Document Reproduction Service No. ED 024-039)

Ebbinghaus, H. (1987). Uber eine neue Methode zur Prufung geistiger Fahigkeiten und ihre Anwendung bei Schulkindern. (On a new method of testing mental abilities and its use with schoolchildren). *Zeitschrift fur Psychologie und Physiologie der Sinnesorgane, 13*, 401-453.

Farhady, H. (1979). The disjunctive fallacy between discrete point and integrative tests. *TESOL Quarterly, 13*, 347-457.

Harris, D., & Palmer, L. (1986). *A comprehensive English language test for learners of English (CELT).* Listening section. New York: McGraw-Hill.

Hinofotis, F. (1983). Cloze as an alternative method of ESL placement and proficiency testing. In J. Oller & K. Perkins (Eds.), *Language Testing* (pp. 121-128). Rowley, MA: Newbury House.

Heilenman, L. (1983). The use of a cloze procedure in foreign language placement. *Modern Language Journal, 67*(2), 121-126.

Kaneda, M., Horiuchi, K., Yamaguchi, S., Shimizu, K., Ohta, H., & Ohkawara, R. (Eds.). The lock-keeper & The Pony Express. *Multi-level Reading Program, Yellow* (pp. 9-12). Tokyo: Goken.

Klein-Braley, C. (1981). *Empirical investigation of cloze tests: An examination of the validity of cloze tests as tests of general language proficiency in English for German university students.* Unpublished doctoral dissertation. University of Duisberg, Germany.

Klein-Braley, C. (1983). A cloze is a cloze is a question. In J. Oller, Jr. (Ed.). *Issues in language testing research* (pp. 281-288). Rowley, MA: Newbury House.

Klein-Braley, C. (1985). A cloze-up on the C-Test: A study in the construct validation of authentic tests. *Language Testing, 2*, 76-104.

Klein-Braley, C. & Raatz, U. (1984). A survey of research on the C-Test. *Language Testing, 1*, 134-146.

Lado, R. (1986). Analysis of native speaker performance on a cloze test. *Language Testing, 3*(2), 130-146.

McCall, W.A., & Crabbs, L.M. (1961). *Standard tests lessons in reading*. New York: Columbia University Teachers College

Mochizuki, A. (1984). Effectiveness of multiple-choice (M-C) cloze tests (2). *Research Bulletin 13*, 159-164. Shizuoka: Chubu English Language Education Society.

Mochizuki, A. (1991). Multiple choice (M-C) cloze tests. *ARELE 2*, 31-40. Tokyo: The Federation of English Language Education Societies in Japan.

Oller, J., & Conrad, C. (1971). The cloze technique and ESL proficiency. *Language Learning, 21*, 183-195.

Porter, D. (1978). Cloze procedure and equivalence. *Language Learning, 12*,

334-341.

Raatz, U., & Klein-Braley, C. (1981). The C-Test-A Modification of the cloze procedure. In T. Culhane, C. Klein-Braley, & D.K. Stevenson (Eds.), *Practice and problems in language testing*. University of Essex Department of Language and Linguistics Occasional Papers No. 26. Colchester: University of Essex.

SONY Language Laboratory. (1980). *SONY aural comprehension test*. Tokyo: SONY Language Laboratory.

Stansfield, C., & Hansen, J. (1983). Field dependence-independence as a variable in second language cloze test performance. *TESOL Quarterly, 17*, 29-38.

Taylor, W. (1956). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly, 33*, 42-48.