Comparing the Narration-based C-test and Klein-Braley's C-test : Their Validity and Reliability

Comparing the Narration-based C-test and Klein-Braley's C-test
—Their Validity and Reliability—

望 月 昭 彦
Akihiko MOCHIZUKI

（平成9年10月6日受理）

This study examines the appropriateness of the C-test with the narration text. The following tests were administered to 185 to 237 college freshmen and sophomores from October, 1994 through February, 1995 ; Test of English as a Foreign Language Practice test IV (TOEFL), TOEFL Listening Comprehension test (TOEFL Listening), Narration C-Test (Na. C-Test), and Klein-Braley C-Test (KB C-Test). Results indicate the following : First, the reliability of the Na.C-Test is very high (r=0.900), and higher than that of the KB C-Test. Second, there is a low correlation between the score of the Na. C-Test and that of TOEFL and a moderate one between the score of the KB C-Test and TOEFL. Third, there are very low correlations between the Na. C-Test and TOEFL Listening and between the KB C-Test and TOEFL Listening. The study indicates that a C-Test which uses a long narration text seems to work well as a measure of a learner's overall language proficiency, and what a C-Test measures seems different from what a listening test measures. A more reliable criterion test and a more reliable listening test are needed for further research.

Since the cloze test was developed by Taylor in 1953 as a measure of the readability of passages of prose, it has been used for many purposes. Caulfield, J. & Smith, V. C. (1981), Hinofotis (1983), Oller & Conrad (1971) suggested using the cloze test as a possible alternative method of ESL placement testing. Dizney, H. & Gromen, L. (1967), Piper, A. (1983), Heilenman (1983) and Fotos, S. (1991) suggested the use of it for EFL placement purposes.

Problems with cloze tests

Although the cloze test was believed at one time to be a panacea for language testing, in recent years it has been pointed out that it has several problems. Those problems can be summarized as follows:

1. Random sampling. Klein-Braley, C. (1985) states that tests of reduced redundancy aim at obtaining a random sample of the examinee's performance, and that *n*th word deletion and random word deletion were not equivalent (p. 82)

2. Scoring methods. The scoring methods for the cloze test are roughly divided into the

following two broad categories: an exact answer method and an acceptable answer method.

With regard to those scoring methods, there are two views: one which regards the two methods as producing no significant difference, and one which regards them as producing some discrimination. Research by Taylor (1953), Rankin (1957), Ruddell (1963) and Bormuth (1964, 1965a, 1965b) belongs to the former, showing that the simplest and most reliable way of scoring is an exact answer method and other scoring methods, like an acceptable answer method, are almost equivalent and do not produce significantly superior discrimination. The latter view is represented by Oller (1972) and Brown (1980). Oller (1972) reported that the acceptable answer method is better than the exact answer method in ESL contexts. Brown (1980) reported that the acceptable answer method is the best overall scoring method of four methods: the exact answer, the acceptable answer, clozentropy, and multiple-choice methods. The acceptable answer method, however, reduces advantages of easy scoring and preparation.

3. Deletion rates and starting points. Alderson (1980, 1983) and Klein-Braley (1981) showed that performance on a cloze test is affected by the nature of the text and by the deletion rate. Porter (1978) states that "The relatively low correlations obtained with either scoring method (i. e. the exact scoring and the acceptable scoring) indicate that students' achievement may vary markedly according to where the deletions begin, that is, according to what is deleted" (p. 336).

4. Reliability and validity. Brown (1993) showed that 50 natural cloze tests (i. e. cloze procedures developed without intercession based on the test writer's knowledge and intuitions about passage difficulty, suitable topics, etc.) were not necessarily reliable (ranging from 0.172 to 0.869 by Split Half method) and valid (ranging from 0.04 to 0.71). Coleman reported that most of the results have ranged between fairly high reliability coefficients (e. g. 0.76-0.94), but occasionally they were moderate (e. g. 0.52). Klein-Braley and Raatz (1984) state "particularly for homogeneous samples (classroom groups or monolingual groups) cloze tests tend to have unsatisfactory reliability and validity coefficients" (p.135).

5. What cloze tests measure. There are three main views on what cloze tests measure: (a) Cloze tests cannot be distinguished from discrete-point tests (Farhady, 1979): (b) Cloze tests measure only basic skills, because of stronger correlation with grammar tests than with reading tests (Alderson,1983): and (c) cloze tests measure overall proficiency, because of strong correlation with dictation, reading tests, and essay writing, in addition to standardized proficiency tests (Chavez-Olller et al. 1985) (Brown, 1993).

The C-Test

As a remedy for solving the above mentioned problems with the cloze test, the C-Test was developed by Raatz and Klein-Braley (1981). In this test, the second half of every second word is deleted with the first and the last sentences left intact. The subjects are

required to fill in the blanks.

Advantage of the C-Test. The C-test has the following advantages: First, the C-Test procedure meets the random sampling requirement (Klein-Braley, 1985), which is a prerequisite for a reliable and valid test. Second, the use of several different short text (usually five or six) minimizes the effect of text topic and text difficulty on test performance. Third, "adult educated native speakers achieve virtually perfect scores" in the C-test (Klein-Braley, 1985, p.84). Fourth, C-tests are less frustrating than cloze tests.

Problems with the C-Test. Recently the following problems with the C-test have been pointed out.

1. The same words were deleted. With its frequent deletion, the C-test deletes the same item several times, which means it adds no new information about the examinee particularly in texts with a limited range of lexical items, as Piper (1983) points out (p-49).

2. Content words vs. structure words. In cloze tests, it has been claimed that structure words are easier to replace than content words, and as Raatz and Klein-Braley (1981) pointed out, "the difficulty of any cloze test is related to the ratio of structure and content words deleted from the text". The same can be said of the C-test. Dornyei and Katona (1992) reported in their analysis of the experiment on Hungarian EFL learners that for the college students, content words are a better measure of language proficiency, whereas for secondary school students structure words are a better estimate of language proficiency.

3. Interpretation. Klein-Braley and Raatz (1984) state that "The great virtue of the C-test is that it spreads out examinees along a continuum and that the rankings it produces show high agreement with teacher judgments and with the results of other more complex language tests" (p.145). However, Piper (1983) concludes after an analysis of her experiment, "It seems that the C-test is more reliable as an item with the more advanced groups, and the Cloze test with the lower groups, if one takes concurrent validity as a measure of reliability" (p.49). It would seem that the C-test is effective in assessing advanced groups..

4. What the C-test measures. This is open to question and there are two main views: (a) The C-Test measures component language skills. Carroll (1986) states that the C-test "harks back in many ways to the form of word completion tests devised by the German psychologist Fbbinghaus (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). (b)The C-Test measures overall language proficiency. Chapelle, C. A. & Abraham, R. G. (1990) reported that the C-test, correlating most strongly with the vocabulary test, produced on average the highest correlations with the placement test among the fixed-ratio cloze test, the rational cloze test, the M-C cloze test and the C-test. Dornyej and Katona (1992) stated after they analyzed five different language tests (the department proficiency test, TOEIC, the oral interview, the cloze test and the C-test) conducted on Hungarian students that "the C-test is a highly integrative language test which measures global language proficiency" although it appeared to be less efficient in testing grammar (p.191).

Criteria which the C-test should meet. Klein-Braley and Raatz (1984) set up the criteria for the C-test as follows: (a) several different texts, (b) 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) exact scoring, (f) high reliability (0.8 or higher by Cronbach's alpha) and validity (at least 0.5) (p. 136).

A good test should be valid, reliable, easy to score and easy to administer. The C-Test should meet these requirements. In secondary schools teachers who serve as test writers are very busy with day-to-day activities and should be spared the burden of selecting several kinds of texts for the C-Test and should greatly benefit from the C-Test which uses just one kind of text as long as the C-Test is reliable and valid.

The C-Test which Klein-Braley and Raatz developed seems to meet what is required of a good test. However, what prevents the C-Test from being accepted widely in Japan seems to be that it is difficult to select five or six short texts.

Thus, emerges the C-Test which uses one kind of text. Mochizuki (1994) reports that the C-Test whose text uses Narration shows the reliability coefficient 0.928 and it seems to be a promising means of measuring a learner's overall language proficiency. In this article and the Narration-based C-Test, or, a "modified C-Test", in terms of reliability, and concurrent validity, and the absolute difficulty of the C-Test.

Method

Purpose : The purposes of the study were to determine whether the reliability of the Narration C-Test will be as high as the Klein-Braley C-Test, to determine whether the concurrent validity of the Narration-based C-Test is high, and to determine whether the correlation between C-Tests and listening Comprehension tests will be low.

Subjects : The experiments were conducted from October, 1994, through February, 1995. The subjects were 237 students at Aichi University of Education—148 first-year students who were enrolled in the courses of art, social studies, science and music, and 89 second-year students who were enrolled in the courses of art and Japanese.

Materials: For this study, the passage which was longer than Klein-Braley's suggestion (approximately 400 words) was used in the Narration C-Test. Narration stands for, as in Mochizuki (1991), a passage that narrates something which happened either in reality or in the imaginary world, for example, excerpts from newspaper articles or novels. Klein-Braley (1994) introduces five different C-Test versions for advanced German students of English, each of which is made up of two anchor items entitled "Car" and "Literature" and three other texts. Out of the five C-Test versions, the one whose mean (87.16) was higher than any other version, was selected for the study in view of the fact that the easiest test version for advanced German students would be suitable for non-advanced Japanese students. This test had a total of 125 items, with 25 items in each of the five texts.

The number of the items was reduced to 120, with 20 items in each text, so that it would be completed by average level college students within 30 minutes and that statistical calculation would be carried out in a split-half method, although the direction in the KB C-test says that around 5 minutes should be allotted for the completion of each text, with 25 minutes in all for 5 texts in the C-Test.

The following were the materials used in this experiment:
1. A 100-item TOEFL Practice Test 1V (TOEFL) composed of a 40-item Structure and Written Expressions part and a 60-item Reading Comprehension part, which the subjects were allowed 70 minutes to complete.
2. A 50-item TOEFL Practice Test 1V of Listening Comprehension (TOEFL Listening), which the subjects were allowed approximately 26 minutes to complete.
3. A 120-item Narration-based C-Test (Na. C-Test), in which the first few and the last few sentences were left intact and the second half of every second word was deleted, and for which 30 minutes were allowed for completion.
4. Klein-Braley 120-item C-Test (KB C-Test) which used 5 different 20-item texts, entitled "Car," "Literature," "Work," "Cuisine," and "Free Speech," and for which 30 minutes were allowed for completion.

The Narration C-Test was 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, stout(1)..... phone(2)..... mouth (3)... overt(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., 5100km, 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 KB C-Test and the Narration C-Test in question, a criterion test had to be specified. Therefore, TOEFL and TOEFL Listening were administered to 185 subjects, (96 freshmen and 89 sophomores) between October, 1994, 3nd February, 1995. A total of 237 subjects, (148 freshmen and 89 sophomores), were tested with the Narration C-Test and the KB C-Test from October, 1994, through the beginning of February.

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.

Results

The mean scores of TOEFL were calculated as shown in Table 1. A glance at the mean scores of the first-year and the second-year students gave me a hunch that there might be no difference between those two groups. The Z test shows that there is no significant difference in mean score between the first -year and second-year students. This means that the two groups are considered as one with the same proficiency level. Second-year students should have done better on TOEFL, but actually they did not. Half of the second-year students were enrolled in the course of art, which might contribute to the score decrease in TOEFL. As a result, first-year and seond-year students were regarded as one group of 237 subjects in this study.

Table 1

TOEFL Between 1st year and 2nd year students

Group	Mean	Full Score	SD	N
1. 1st year students	31.354	100	7.997	96
2. 2nd year students	32.180	100	7.094	89

Z-test $z = 0.774 < z_{0.05}$ (1.96) $P > 0.05$

The reliability coefficients and P values were calculated as shown in Table 2. With regard to the forms, in this study, the Split-Half Method was used for their calculation. In the assessment of the reliability of the Narration C-Test and Klein-Braley's C-Test, 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 (1983) used each of the various texts as "super items" (p.136) without analyzing individual items, thereby avoiding the issue of item independence. Whether cloze tests are sensitive to language constraints across sentences and can be completed only from the context of the sentence is argued about. But the cloze test as a measure of higher level skills and overall proficiency is being accepted (Brown, 1989). Like wise, the C-Test appears to be a measure of grammatical competence rather than of textual competence. However, the C-Test is a measure of overall language proficiency, as is shown in Stansfield and Hansen (1983). Therefore, the blanks could be considered to be independent, which means that the use of the Split-Half Method is acceptable as a measure of the reliability of the C-Test.

The reliability coefficients of the Narration C-Test and the KB C-test were very high or high (Table 2), the Narration C-Test ($r=0.900$) a lot higher than KB C-Test ($r=0.834$). In other words, learners perform better on passages which have a temporally ordered

sequence of events. The Narration C-Test meets this requirement, whereas the KB C-Test does not necessarily involve a sequential element.

For reference, Klein-Braley (1994) reports that the mean score of her standard C-Test was 87.16 out of 125 full score points, that is, 83.674 out of 120 points, and that the reliability coefficient by Cronbach alpha was 0.88. The wide gap in mean scores of the KB C-Test (83.674 vs. 44.730) between advanced German and non-advanced Japanese level students draws attention. What kind of factor contributes to the widening of the score difference between Japanese college students and their German counterparts?

The reliability coefficients of TOEFL and TOEFL Listening Comprehension were moderate or very low. The reliability coefficient of TOEFL was lower than expected. Statistics show that low mean scores yield low reliability. Although generally TOEFL is regarded as the most reliable measure of overall language proficiency, it can be inadequate for non-advanced students. The reliability coefficient of TOEFL Listening Comprehension was low as was expected. The same tendency was shown in Mochizuki (1994).

Table 2
Reliability Coefficients by Split Half Method

Tests	N	r	Mean	Full Score	SD
1. Na. C-Test	237	0.900	78.080	120	14.025
		P<0.001			
2. KB C-Test	237	0.834	44.730	120	12.395
		P<0.001			
3. TOEFL	185	0.669	31.751	100	7.567
		P<0.001			
4. TOEFL Listening	185	0.209	13.638	50	3.368
		P<0.05			

Na.=Narration, KB=Klein-Braley

The P-values showing the absolute difficulty of each test which can be obtained by dividing full score by mean score were calculated as shown in Table 3. Klein-Braley (1994) reports that the P-value of her standard C-Test for advanced German students was 69. The P-value of her C-Test conducted on Japanese college students was 37, which means that the KB C-Test was too difficult for them. On the other hand, the P-value of the Narration C-test was 65, which means that the level of the test was more adequate. What is noteworthy about the difference between the Narration C-Test and the KB C-test is that in addition to the number of texts used in those tests, the Narration C-Test carries notes which explain several difficult words/phrases in a bilingual way, whereas the Klein-Braley C-Test does not. The existence of the notes in the Narration C-Test might to some degree help the subjects to enhance their comprehensibility and raise their score, which may lead to the increase of the reliability and the P-value of the Narration C-Test. Some

examinees say that the KB C-Test looked difficult and that it was difficult to tackle, whereas the Narration C-Test was easy to understand as it was a story.

The P-value of TOEFL and TOEFL Listening Comprehension were low, which means that those tests were too difficult for Japananese college students. No test can match TOEFL as a reliable and valid measure of the learner who wants to be admitted into a US college, but the same TOEFL might not work well as a measure of an average-level Japanese learner's overall proficiency. The same might be said about TOEFL Listening Comprehension.

Table 3
P Values

Tests	N	P	Mean	Full Score	SD
1. Na. C-Test	237	65	78.080	120	14.025
2. KB C-Test	237	37	44.730	120	12.395
3. TOEFL	185	32	31.751	100	7.567
4. TOEFL Listening	185	27	13.638	50	3.368

Na.=Narration, KB=Klein-Braley

In each pair there were moderate or low correlations between the score of TOEFL and the C-Tests and between the score of TOEFL Listening Comprehension and the C-Tests. The correlation procedure used in Table 4 & 5 is the Pearson product-moment procedure, and the correlations given in the tables are values r. The correlation coefficients between TOEFL and C-Tests were expected to reach 0.5 but actually they did not. The correlation between the KB C-Test and TOEFL was moderate and that between the Narration C-Test and TOEFL was low.

Table 4
Correlation Between C-Tests and TOEFL Practice tests (n=185)

Tests	r	p
1. Na. C-Test and TOEFL	0.347	$p < 0.001$
2. KB C-Test and TOEFL	0.430	$p < 0.001$

Table 5 reveals a very low correlation between the scores of the C-Test and the TOEFL. Listening Comprehension test in each pair. 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 5
Correlation Between C-Tests and TOEFL Practice tests (n=185)

Tests	r	p
1. Na. C-Test and TOEFL Listening	0.114	$p<0.1$
2. KB C-Test and TOEFL Listening	0.085	$p<0.2$

Table 6 shows that there was a significant difference in the scores of the Narration C-Test between first-year students and second-year students, although the TOEFL score did not show any significant difference between them as shown in Table 1.

Table 6
Difference in Na. C-Test score between 1st year & 2nd year students

Subjects	N	Mean	SD
1. First-year students	148	76.358	14.634
2. Second-year students	89	80.944	12.510
z test $z=2.562>z_{0.03}$ (2.17)	$P<0.03$		

Table 7 shows that there was a significant difference in the scores of the KB C-Test between first-year and second-year students, whereas TOEFL scores did not show any significant difference between them as shown in Table 1.

Table 7
Difference in KB C-Test score between 1st year & 2nd year students

Subjects	N	Mean	SD
1. First-year students	148	42.906	12.918
2. Second-year students	89	47.281	11.910
z test $z=2.725>z_{0.01}$ (2.58)	$P<0.01$		

Discussion

As stated in the beginning in the Method section, let's compare Narration C-Test with Klein-Braley (KB) C-Test in terms of reliability, concurrent validity and their correlation with TOEFL Listening Comprehension tests.

First, the Narration C-Test turned out to be a reliable test whose reliability ($r=0.900$) was higher than that of the KB C-Test ($r=0.834$). The high reliability of the Narration C-Test can be inferred from Mochizuki (1994).

Second, although the concurrent validity of the Narration C-Test was expected to be high, actually it was not. The low correlation between the Narration C-Test, and TOEFL and the moderate one between the KB C-Test and TOEFL show that what the C-Tests measure is different from what TOEFL measures. This result fell short of what I had

expected. The reliability of TOEFL turned out to be moderate (r=0.669) by the Split-Half Method, far lower than I had expected. 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 reliability coefficient which reveals Japanese students' overall language proficiency is urgently needed. This part must be further investigated in the future by using a STEP-like test which is easily available to secondary school teachers in Japan.

Third, the correlation between C-Tests and listening comprehension tests turned out to be low. As can be inferred from Mochizuki (1994), I had expected a low correlation and the results confirmed what I had expected. This means what the C-Test measures seems to be quite different from what the Listening Comprehension test measures. However again the problem of the low mean score and small standard deviation must be paid attention to : the very low correlations between the scores of the C-Test and TOEFL Listening Comprehension might have been brought about by it. A reliable Listening Comprehension test is also urgently needed. I have examined the TOEFL Listening Comprehension, the CELT Listening Comprehension test, and the SONY Aural Comprehension test. The reliability coefficients of those tests were all moderate, not high. A further investigation of the correlation between the C-Test and the Listening Comprehension test will be carried out when a highly reliable Listening Comprehension test is obtained.

Conclusion

In this research, I compared the Narration C-Test and the Klein Braley C-Test in relation to TOEFL and TOEFL Listening Comprehension test. The results showed that, first, the Narration C-Test was found to be very reliable (0.900), and as reliable as Klein-Braley C-Test. Second, the concurrent validity of the Narration C-Test did not reach the threshold level of 0.5. What the Narration C-Test measures seems to be the same as what the criterion test measures (r=0.347) and it seems to measure something different from what the Listening Comprehension test measures. However, in order to confirm the correlations between the C-Tests and criterion tests, more reliable discrete-point criterion and Listening Comprehension tests with a reliability of 0.80 or higher are needed.

What is noteworthy is that this study revealed that the C-Test whose text used the Narration was a critical threshold level of 0.8 and more reliable than Klein-Braley's, and it was more sensitive to the continuum of the English proficiency level of the subjects than TOEFL. The C-Test with a long Narration passage might work in secondary schools in Japan, because it is reliable and could turn out to be valid with the use of more reliable discrete-point tests, and further because it requires less time to write than several short texts. Further research should be conducted in secondary schools in Japan to determine whether the Narration C-Test can work as suggested in college students.

Note

1. The Narration C-Test, which used Narration, "The Lock Keeper" (413 words) (Kaneda, et al., 1971), Klein-Braley C-Test, which used 5 short texts from the booklet entitled *Experimental C-Tests* (Klein-Braley, 1994). Copies of the C-Tests used in this experiment are available from the author on request.

REFERENCES

Alderson, J.C. (1980). Native and nonnative speaker performance on cloze tests. *Language Learning, 30,* 59-76.

Alderson, J. C. (1983). The cloze procedure and proficiency in English as a foreign language. In J.W. Oller (Ed.) *Issues in language testing research* (pp.205-217). Rowley, Massachusetts: Newbury House Publishers.

Brown, J. D. (1980). Relative merits of four methods for scoring cloze tests. *Modern Language Journal, 64* (3) , 311-317.

Brown, J. D. (1993) What are the characteristics of natural cloze tests? *Language Testing 10,* (2), 93-116.

Bormuth, J. R. (1964). Experimental applications of cloze tests. *International Reading Association Conference Proceedings 9*. 303.

Bormuth, J. R. (1965a). Optimum sample size and cloze test length in readability measurement. *Journal of Educational Measurement 2.* 111.

Bormuth, J. R.(1965b). *Validities of grammatical and semantic classifications of cloze scores. International Reading Association Conference Proceedings. 10.* 283.

Carroll, J. B (1986). LT+25, and beyond? Comments. *Language Testing, 3* (2), 123-129.

Caulfield, J. & Smith, U.C. (1981). The reduced redundancy test and the cloze procedure as measures of global language proficiency. *Modern Language Journal 65,* pp-54-58.

Chavez-011er, M., Chihara, T., Weaver, K., & Oller, J. (1985). When are cloze items sensitive to constraints across sentences? *Language Learning,* 35, 181-203.

Coleman, E. B. (1971). Developing a technology of written instruction: some determiners of the complexity of written prose. In *Verbal learning research and the technology of written instruction.* E. Z. Rothkopf and P. E. Johnson (Eds.) New York: Teachers College Press. Columbia University.

Dizney, H. & Gromen, L. (1967). Predictive validity and differential achievement on three MLA-cooperative foreign language tests. *Educational and Psychological Measurement. 27.* pp. 1127-30.

Dornyei, Z. and Lucy Katona. (1992). Validation of the C-test amongst Hungarian EFL learners. *Language Testing, 9* (2), 187-206.

Farhady, H. (1979). The disjunctive fallacy between discrete point and integrative tests. *TESOL Quarterly,* 13, 347-357.

Fotos, S. S. (1991). The cloze test as an integrative measure of EFL proficiency: a substitute for essays on college entrance examinations? *Language Learning 41,* 3, pp.

313-336.

Friedman, M. (1964). The use of the cloze procedure for improving the reading comprehension of foreign students at the University of Florida. Unpublishcd doctoral dissertation, University of Florida.

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

Hill, L. A. (1965). *Advanced stories for reproduction 1.* Oxford : Oxford University Press. 5-11, & 14-15.

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, NA : Newbury House.

Hunt, K. W. (1970). Syntactic maturity in school children and adults. *Monographs of the Society for Research in Child Development 35,* 1 (serial no.134).

Klein-Braley, C. (1981). Empirical investigations of cloze tests. Unpublished doctoral dissertation, Universitat Duisburg, Duisburg, Federal Republic of Germany.

_____. (1985). A cloze-up on the C-test : a study in the construct validation of an authentic test, *Language Testing, 2,* 76-104.

_____. (1994). *Experimental C-Tests.* Unpublished booklet. Universitat Duisburg, Duisburg, FB3 Angewandt Linguistik, D47048 Duisburg, Germany.

Mochizuki, A. (1991). Multiple choice (M-C) cloze tests. *ARELE 2,* 31-401 Tokyo : The Federation of English Language Education Societies in Japan.

Mochizuki, A. (1994). C-tests—four kinds of texts, their reliability and validity. *JALT Journal 16,* (1), 41-54.

Oller, J. (1972). Scoring methods and difficulty levels for cloze tests of proficiency in English as a second language. *Modern Language Journal, 56,* 151-157.

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

Piper, A. (1983). A comparison of the cloze and C-test as placement test items. *The British Journal of Language Teaching, 21* (1), 45-51.

Raatz, U. & Klein-Braley, C. (1981). The C-test—a modification of the cloze procedure. In *Practice and problems in language testing.* T Culhane, C. Klein-Braley, & D. K. Stevenson (Eds.) University of Essex.

Rankin, E. F., Jr. (1957). An evaluation of the cloze procedures a technique for measuring reading comprehension. Unpublished doctoral dissertation, University of Michigan.

Ruddell, R. B. (1963). The effect of oral and written patterns of language structure on reading comprehension. Unpublished doctoral dissertation, University of Indiana.

Steinberg, R. (1987) *Prentice Hall's practice tests for the TOEFL.* New Jersey : Prentice Hall Regents. 75-89.

Appendix A

A Part from C-Tests by Klein-Braley

Fill in the blanks with one or several letters so that the sentences will make sense. The second half of every second word is deleted.

Example : There are usually five men in the crew of a fire engine. One o_ them dri ____ the eng ____. 答。of, drives, engine

1. Car

Be particularly careful when buying a used car from a private individual —— you have fewer rights than uhen buying from a trader. Your (1)rights will (2)largely depend (3)on what (4)is said (5)between you (6)and the (7)seller —— that (8)is, what (9)you are (10)told about (11)the .....