A Study of four C-tests

A Study of four C-tests

望 月 昭 彦
Akihiko MOCHIZUKI

（平成8年10月7日受理）

This study is intended to confirm the result of the experiment on appropriateness for constructing the most effective C-test, which was conducted on 42 college freshmen in 1992. In this study the following tests were administered to 727 second-year high school students from September,1995, through February, 1996: the Pre-Second Grade Test of the Society of English Proficiency (STEP), and four C-tests whose texts used Narration, Explanation, Description, and Argumentation. Results confirmed what the 1992 experiment revealed. They indicate the following: First, the reliability of the Narration C-test is very high ($r = 0.960$) and the highest among the four C-tests. Second, there is a high correlation between the scores of Narration C-test and the STEP test. Third, the face validity does not look good in view of the survey conducted on 12 senior high school teachers and six EFL instructors interested in testing and also because of the fact that several intermediate and lower level students scored zero points in each C-test. It seems that further research is necessary to study the content validity of a Narration C-test.

The cloze test was developed by Taylor in 1953 as a means of measuring readability,and since then it has been regarded as a prime example of integrative testing. It has been used to measure overall proficiency of language learners. But several problems with the cloze test such as random sampling, scoring method, deletion rates and starting points have been pointed out by Klein-Braley (1981), Alderson (1980,1983), Porter (1978), Brown (1993), Klein-Braley and Raatz (1984). As a remedy for the solution to those problems 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 gaps.

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).

The C-Test has not established a legitimate position as another testing technique.

Jafarpur,A.(1995) claimed that C-testing is not superior to standard cloze testing by saying (a) that although it is easy to construct and to score C-tests, native speakers do not achieve perfect scores, (b) that the deletions do not affect a representative sample of the text; different deletion starts and deletion ratios produce different tests, which is suggestive of the invalidity of the procedure; (c) that previously untried material demonstrates satisfactory reliability but does not show acceptable validity against cloze testing; and (d) that C-tests do not possess face validity. The C-Test has much to explore.

The C-Test has not dealt with what kind of text will generate a higher reliability and validity. It is rather hard to deal with several different kinds of texts in one C-test, for Klein-Braley and Raatz (1984) did not specify what kinds of texts should be used nor how many texts. If one kind of text were determined to produce a higher reliability and validity, it would relieve the test writer of the big burden of finding appropriate texts for the test. Mochizuki (1994), after administering four C-Tests whose texts used Narration, Explanation, Description and Argumentation, and a criterion test etc. to college students, reported that a Narration C-Test seems to be a promising means of measuring a learner's overall language proficiency. Therefore, the intent of this research is to confirm the results by conducting the four C-Tests whose texts used Narration, Explanation, Description and Argumentation on a large number of high school students.

Method

Purpose

1. To compare four C-Tests whose texts used Explanation, Argumentation, Description, and Narration in terms of concurrent validity and reliability.
2. To investigate the face validity of C-Tests.

Subjects:

A total of 727 second-year high school students at Shizuoka High School, Numazu Higashi High School, Shizuoka Higashi High School, Tajimi High School, Taiseiji High School, Aichi University Attached High School, and Tsushima Kita High School. The subjects took a criterion test, and four C-Tests, but only those who took all the five tests were dealt with for this study so that a statistical calculation could be carried out with ease. After all the number of the subjects who took all five tests turned out to be 600.

Materials:

a. What was used as a criterion test was the Pre-Second Grade Test of the Society of Testing English Proficiency (STEP).
b. C-Tests. Four kinds of texts were represented, as in Mochizuki (1991):

1. 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 as a criticism of art or literature.
2. 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.
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 happpened 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. The 36-item criterion test (STEP) composed of an assortment of items from past pre-second grade STEP written examinations, which the subjects were allowed 50 minutes to complete.
2. Four 100-item C-Tests, in each of which the first 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, 25 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 ($[n\text{-}1]/2$) and ($[n+1]/2$) numbered words alternately; for example, sto*ut*(1)... ph*one* (2)... mou*th*(3) ...ov*ert*(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. Numbers/proper nouns (e.g., *511* 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.

c. Questionnaire: The attitudes of Japanese high school teachers and also EFL non-native instructors toward C-tests were gathered through the same questionnaire as Jafarpur (1995) used. Twelve high school teachers were the overseers who conducted the five tests at their high schools. They were the aged from 30 to 61, with at least 8 years of teaching experience at a high school. They answered the

same questions on completeness, representativeness, appearance and what the test measured. Six EFL instructors were Japanese faculty members teaching at a 4-year university or a 2-year junior college who are interested in testing.

Procedure

In order to investigate the concurrent validity of the C-Tests in question, a criterion test had to be specified. Therefore, the STEP placement examination was administered to all the subjects between September, 1995, and February, 1996. The result revealed that the reliability of the STEP examination was 0.874 by the split half method ($\overline{X}$ =73.25, SD = 25.98). The calculation was carried out on the basis of a total of 102 subjects, that is, the top 33 from the upper level group, the middle 39 from the intermediate level group, and the bottom 30 from the lower level group. All the subjects were tested on the four C-Tests beginning with C-Test No.1 and ending with C-Test No.4. Before these tests were administered, I explained in a letter or by talking with the senior high school teachers, how I had carried out research on cloze tests including C-tests and what the C-test meant in testing in English education. Therefore, they probably asked their students to cooperate for my project.

After the tests were administered, all the test papers were scored by my part-time student assistants using model answers which I had approved. Each assistant scored one kind of test. 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 to some degree their familiarity with the testing format.

Results

1. Reliability.

The reliability coefficients were calculated as shown in Table 1. In this study, in assessing the reliability of individual C-test texts, the KR-21 was used so that a large number of test papers could be processed quickly and easily.

Table 1 Reliability Coefficients by the KR-21 (n = 600)

Test	r	Mean	SD
C-Test No.1	0.935	42.24	18.147
C-Test No.2	0.939	33.94	17.908
C-Test No.3	0.861	28.12	11.705
C-Test No.4	0.960	47.68	22.311

Note:

* Mean scores reported out of 100.

The reliability coefficients of all the four C-Tests were high or very high. The four C-Tests were placed in the order of reliability coefficients, from highest to lowest; Narration C-Test No.4 (r=0.960); Explanation C-Test No.2 (r = 0.939); Argumentation C-Test No.1 (r = 0.935); Description C-Test No.3 (r = 0.861). The difference in mean scores among the four kinds of C-Tests show that the mean score of Narration C-Test No.4 is the highest ($\bar{X}$ = 47.68), which shows that learners perform well on passages which have a temporally ordered sequence of events. Narration meets this requirement. The same tendency is found with college students (Mochizuki 1994).

In order to know the reliability in relation to proficiency level, I divided 600 subjects into three levels, that is, the upper level (n=200), the intermediate level (n= 200), and the lower level (n=200). The reliability coefficients of the four C-Tests in three level groups were high or very high as shown in Table 2.

Table 2 Reliability coefficients in the upper level group by the KR-21 (n = 200)

Test	r	Mean	SD
C-Test No.1	0.911	56.94	15.81
C-Test No.2	0.914	48.06	16.17
C-Test No.3	0.739	36.98	9.32
C-Test No.4	0.924	66.62	16.17

Note:

* Mean scores reported out of 100.

The four C-Tests were placed in the order of reliability coefficients, from highest to lowest; Narration C-Test No.4(r =0.924);Explanation C-Test No.2 (r=0.914); Argumentation C-Test No.1 (r=0.911);Description C-Test No.3 (r=0.739). In the upper level group, Narration C-Test No.4 ranked first and the rank order was the same as in Table 1.

Now I would like to show the order of reliability coefficients in the case of intermediate level group.

Table 3 Reliability coefficients in the intermediate level group by the KR-21 (n= 200)

Test	r	Mean	SD
C-Test No.1	0.882	40.09	13.75
C-Test No.2	0.904	31.20	14.32
C-Test No.3	0.794	27.57	9.66
C-Test No.4	0.930	46.72	17.76

Note:

* Mean scores reported out of 100.

The four C-Tests were placed in the order of reliability coefficients, from highest to lowest:Narration C-Test No.4 (r=0.930); Explanation C-Test No.2 (r=0.904); Argumentation C-Test No.1 (r=0.882); Description C-Test No.3 (r=0.794). Again Narration C-Test No.4 came in first and the rank order was equivalent to Table 1.

Table 4 shows the reliability coefficients in the case of the lower group.

Table 4 Reliability coefficients in the lower group by the KR-21 (n=200)

Test	r	Mean	SD
C-Test No.1	0.894	29.69	13.06
C-Test No.2	0.899	22.55	12.61
C-Test No.3	0.817	19.83	9.13
C-Test No.4	0.920	29.70	15.28

Note:

* Mean scores reported out of 100.

The four C-Tests were placed in the order of reliability coefficients, from highest to lowest: Narration C-Test No.4 (r=0.920); Explanation C-Test No.2 (r=0.899); Argumentation C-Test No.1 (r=0.894); Description C-Test No.3 (r=0.817). Narration C-Test No.4 topped the list and the other C-Tests ranked in the same order as in Table 1.

In Oller and Conrad's (1971) experiment, the passage "What's a college?" which is categorized as Argumentation worked well with a discriminator with beginning ESL students doing poorly in the Argumentation cloze test ($\bar{X}$ =7.00 out of 50, the lowest of all the seven groups). However in my experiment my students did well ($\bar{X}$ =42.24 out of 100, the second highest of the four C-Tests). The student's good performance might have much to do with the topic of the passage. The students are familiar with the concept of democracy in Argumentation C-Test No.1.

2 . Concurrent validity

Table 5 shows the concurrent validity of C-Tests. In each pair there was always a moderate or high correlation between the score of the STEP examination and the C-Test. The correlation procedure which was used in Table 5 is the Pearson product-moment procedure, and the correlations shown in the table are values for r. The four pairs of STEP examinations and C-Tests are placed in the order of correlation coefficients, from highest to lowest: 1. Narration C-Test No.4 + STEP; 2. Argumentation C-Test No.1 +STEP; 3. Description C-Test No.3 + STEP; 4. Explanation C-Test No.2 + STEP. Table 5 shows that the correlation between Narration C-Test No.4 and STEP was the highest (r= 0.7333).

Table 5 Correlations between C-Tests and STEP placement (n = 600)

Test	r	p
STEP placement and C-Test No.1	0.6782	0.000
STEP placement and C-Test No.2	0.6429	0.000
STEP placement and C-Test No.3	0.6478	0.000
STEP placement and C-Test No.4	0.7333	0.000

3 . Content Validity

The content analysis was made on the basis of the distribution of content and function words in each C-test as in Table 6.

Table 6 Content and function words in C-tests

C-test	Content	Function	% Content	Mean
C-test No.1	47	53	47	42.24
C-test No.2	57	43	57	33.94
C-test No.3	57	43	57	28.12
C-test No.4	51	49	51	47.68

Analysis of this table shows that the distribution of content and function words has almost the same density in the 4 kinds of C-tests.

The text of C-test No.1 is Argumentation , a passage on democracy. Although the text is a formal style, most of the subjects have already studied the philosophy and political system of democracy through social studies class and know them in their daily life. The passage is to some degree similar to Explanation. Therefore, the subjects did very well on this test. I should have chosen a more persuasive passage which is more suited to the definition of Argumentation. The text of C-test No. 2 is Explanation, a passage on the Pony Express. It dealt with a mail system which took place in the U.S. in the 1860s. The subjects were probably unfamiliar with the topic. Therefore, they did rather poorly on the test. The text of C-test No.3 was Description, an adaptation from the Los Angeles Times dealing with a present day U.S. politician named Robert Reich. The person in question was quite unfamiliar to the subjects. Therefore, they did the poorest on it of all the C-tests. The text of C-test No.4 was Narration, a passage on a ghost who appeared in the canal. Just as in Mochizuki (1994), the subjects performed the best on this test of all the 4 C-tests. The reason for a very good performance is open to discussion. It seems to me that the text of Narration is more capable of eliciting real overall proficiency of the subject regardless of his/her level than any other type, and of putting it along the continuum of the students' level from highest to lowest.

As to the content validity, I should have analyzed the readability of the text by using

a readability formula. At any time the text of a test should fit the level of the testees.

3 Correlations between C-Tests

Table 6 presents correlations between Narration C-Test No.4 and the other C-tests.

Table 7 Correlations between C-Tests (n =600)

	C-Test 1	C-Test 2	C-Test 3	C-Test 4
C-Test 1	1.0000	.8023	.7308	.7567
	p= .	p=.000	p=.000	p=.000
C-Test 2	.8023	1.0000	.7394	.7852
	p=.000	p= .	p=.000	p=.000
C-Test 3	.7308	.7394	1.0000	.7888
	p=.000	p=.000	p=.	p=.000
C-Test 4	.7567	.7852	.7888	1.0000
	p=.000	p=.000	p=.000	p=.

The average correlation coefficients between Argumentation C-Test No.1 and the other three C-Tests, between Explanation C-Test No.2 and the other three C-Tests, between Description C-Test No.3 and the other three C-tests and between Narration C-Test No.4 and the other three C-Tests are, 0.7633, 0.7756, 0.7530, and 0.7769 respectively. We observe that correlation coefficients between any one C-Test and the other three C-Tests were high (lowest 0.7308, highest 0.8023) and that correlation between Naration C-Test No.4 and the other three C-tests was the highest (r=0.7769). It implies that Narration C-Test No.4 is highly correlated to each of the other three C-tests.

4 . Face validity

A Summary of the view of Japanese high school teachers of English on how a C-test looks appears in Table 8. The majority of the high school teachers responded in positive terms (75%). Thirty-three percent of them thought that the C-Test measured English proficiency only and the rest thought that the test measured, besides English proficiency, inference, patience, motivation toward studying English, concentration, background knowledge, logical thinking and knowledge about vocabulary.

More than half of the teachers (59 %) gave positive responses to Question 5. Yet the teachers split into half as to whether to accept the C-test as the criterion for student selection in the university entrance examination. This indicates they were hesitant about the validity and reliability of the C-Test. However, it is noteworthy that several teachers said that this C-Test should be included as a part of the university entrance examination.

Table 8 Veteran Japanese teachers' responses to the questionnaire on C-tests (n=12)

Question	Responses (%) Positive	Negative	Neutral
1. What do you think of this as a test of English?	75	25	
2. Do you think it is a good test?	67	33	
3. Do you think this test measures English proficiency only?	33	59	8

4. If not, what else does it measure?
   1. Semantic,morphological and syntactic inference. The past cloze tests focused on syntactic inference too much.
   2. The C-tests are effective for the students in an elementary level but they are not for those in an intermediate or advanced level.
   3. Inference and patience.
   4. Patience and motivation toward studying English.
   5. Ability to comprehend the logic of the essay, how to organize it and to predict a story, and imaginative power.
   6. Background knowledge, reading rather than speaking ability, logical thinking and knowledge about vocabulary.
   7. Inference.
   8. Inference and concentration.
   9. Ability to put into an image what a passage in question expresses.

5. Do you think this is a fair test of English?	59	33	8

6. Why so, or why not?

(Positive)
   1. C-tests are better than the past cloze tests in that the former asks the testee to make semantic inference.
   2. C-tests are good as a type of proficiency test, but they cannot be used for everyday practice in a class.
   3. C-tests can measure vocabulary and grammatical competence since the testees must complete the sentence by considering parts of the speech in the context .

4. C-tests can measure the testees' genuine English performance in that they are not allowed to use Japanese in answering the questions.
5. C-tests can measure reading comprehension and grammatical competence.
6. When I administered cloze tests to my students, I found that returnees were able to do well, whereas a type of students depending on rote-memorization, or the students who are not really proficient did not do well. Therefore, I think that C-tests can measure the learner's true overall English proficiency.

(Negative)

1. Too much emphasis on vocabulary. (2 people)
2. C-tests are good in examining the students' reading comprehension but questionable in examining their language performance. I am afraid that this type of tests may produce some answers which are difficult for the testees to answer or ambiguous.
3. Too difficult.

(Neutral)

1. C-testing is a good way to measure vocabulary, and grammatical competence.

7. What do you think of the representativeness of this test?	58	25	17
8. What do you think of the completeness of this test?	42	58	
9. Would you want your students' acceptance at universities to depend on this test?	50	42	8

10. Why?

(Positive)

1. C-tests should be included as part of a university entrance examination. They are effective in measuring linguistic inference which plays an important role in linguistic competence and also in measuring writing ability.(2 people)
2. I like new things.
3. I suggest using C-tests together with other types of tests.
4. C-tests seem to be able to measure the learner's real overall proficiency. It will do him/her good to start to get accustomed to this type of test at a high school or a junior high school.

(Negative)

1. The testees, while taking the C-tests didn't show any sign of enjoyment. The texts of the C-tests should be the ones which will be attractive to the students and the vocabulary should be that which is familiar to them.
2. This is not a type of achievement test. I want to evaluate how much the students achieved what they had learned.
3. The students will have much difficulty preparing for this type of tests. The C-tests are difficult for the lower-level students, who I am afraid will not have a willingness to tackle the questions in those tests.
4. Judgement criterion may depend on the text which the C-test uses.
5. I have some anxiety about the use of the C-tests since their reliability and validity are not fully proved yet. The C-tests alone are not sufficient as a university entrance examination in terms of measuring the students' listening, speaking, and writing abilities. When students are in elementary level, the teacher must measure their basic knowledge such as vocabulary, tenses and word order. I presume that there might be another test with fewer items to measure students' multifacet abilities other than C-tests. Tests which have the examinees fill in the blanks are not natural language activities. C-tests might not be able to measure the ability to comprehend a long passage rapidly or to grasp the logic or to pronounce words correctly.

(Neutral)

1. C-tests are good in writing and scoring and they can measure students' overall proficiency in that they include the indispensable ingredient for ordinary language activities such as tapping the students' ability to infer what the other person wants to say. However, the university entrance examination should include questions for reading comprehension and knowledge about grammar and vocabulary.

As shown in Table 9, the views of the college/university level instructors are very conservative especially when compared to those of the high school teachers discussed above. The instructors gave positive responses to 36 percent of the questions excluding explanation-type. They thought that the procedure tapped not language competence but something else, like spelling regardless of the context, knowledge about vocabulary, and guessing. Only two of the instructors responded that they would accept the test as the criterion for student selection in a university entrance examination. This indicates their lack of confidence in the procedure.

Table 9 Responses of college/university level instructors to the questionnaire on C-tests (n=6)

Question	Responses (%) Positive	Negative	Neutral
1. What do you think of this as a test of English?	66	17	17
Reasons for Neutral:1. What do you mean by "a test of English?" English structures or proficiency or …?			
2. Do you think it is a good test?	50	17	33
Reasons for Neutral :			
1. This test is a good test for students who like English or who have basic knowledge about and skills in English. But it might be more demanding than a cloze test or an ordinary type of test for students who are poor at English or who lack basic knowledge about and skills in English, because the distance between blanks is very short.			
2. Yes or no.			
3. Do you think this test measures English proficiency only?	17	66	17
Reasons for Neutral: It depends on how you define "proficiency."			
4. If not, what else does it measure?			
1. Ambiguity tolerance, field independence, physiological/psychological conditions.			
2. Knowledge about vocabulary.			
3. Guessing, grasping the context, and schema.			
4. Spelling regardless of the context.			
5. Do you think this is a fair test of English?	50	0	50
Reasons for Neutral :			
1. If you can control such non-linguistic factors as ambiguity tolerance, this test can be called a fair test of English.			
2. The meaning of "a fair test of English" is ambiguous.			
3. I can't say either "Yes" or "No."			
6. Why so, or why not?			
(Positive)			
1. The subject's task is to restore the second half of the word, so only one			

answer to each question can be obtained.

2. Because it is an integrative type of test in every meaning of the term.

7. What do you think of the representativeness of this test? 33 17 50

Reasons for Neutral:

1. Representativeness of general proficiency? If so, positive.
2. We need more research on this type of test.

8. What do you think of the completeness of this test? 0 50 50

Reasons for Neutral :

1. There are no "complete" tests in any kind of tests; tests are expected to function as complementary.
2. Extensive research would be required for its completion; especially on the issue of an appropriate interval of blanking (every 2 words, every 3 words, every 4 words,……).

9. Would you want your students' acceptance at universities to depend on this test? 33 50 17

Reasons for Neutral:

1. I can't figure out what the question means. If you make a blank in every second word, you will find more function words in those blanks. The function words are easy to infer regardless of whatever the context is. So a certain length of text is needed for this test. Furthermore, regarding content words, if you find "peo____" on your test, you will easily predict "people"or if you find "candi_____", then you can get "candidate" without thinking of the content.

10. Why?

(Positive)

1. This test could be used as a part of comprehensive test battery so that it will give variation to university entrance examinations. This test could measure some productive aspect of linguistic competence very objectively which multiple choice tests could not probe.
2. Because of the very nature of the C-Test.

(Negative)

1. Depends on what kinds of English abilities you think are required for

university entrance exams. This C-test gives an impression of a kind of vocabulary test. At the moment it is not clear how much this test can measure reading comprehension, or creative performance in writing etc.

2. This test mesures knowledge about vocabulary.

3. You should follow the principles of the C-test Klein-Braley and Raatz developed, namely, one C-test should be composed of 4 different kinds of texts, each 25 questions totaling 100 questions. Then it will keep a good balance between kinds of the texts and the number of question items.

4. How well the subjects will do on this test depends on the text style of a reading selection. Although a test whose text used Narration is correlated with the learner's overall proficiency, just as Mr.Mochizuki reported, the other text styles might not be. Next, the definitions of the text styles should be clear. Third,the difficulty level of the reading selection is important. You should see to it that reading selection fits the English ability of the subjects.

While I was taking a record of scores for each of the five tests, my attention was drawn to the fact that several subjects scored zero points on C-tests from No.1 to No.4, though not on the STEP examination. Table 10 presents the number of the subjects who scored zero points in each of the three proficiency level groups.

Table 10 Number of subjects who scored zero points

	n	C-Test 1	C-Test 2	C-Test 3	C-Test 4
Upper level	200	0	0	0	0
Intermediate level	200	0	4	1	2
Lower level	200	2	9	7	9
Total	600	2	13	8	11

Although I did not ask the subjects about what they thought about C-Tests after they took them, Table 10 was part of their answer. On each of the C-Tests,two to thirteen subjects scored zero points. They were in the intermediate or lower level. Especially the lower level students accounted for 79 % of all the subjects who scored zero points. The texts of C-Tests No.1-4 may have been too difficult for them to answer, and their first glance at them may have discouraged them from replying. There may have been something that prevented them from tackling C-Tests.

Conclusion

The four kinds of C-Tests whose texts used Argumentation, Explanation, Description, and Narration were investigated in terms of reliability, concurrent validity, content validity, and face validity. The results of the investigation indicate that

1) the reliability of Narration C-Test No.4 was very high (r=0.960) and when the subjects were divided into three levels, that is, upper, intermediate and lower levels, the reliability of Narration C-Test No.4 was the highest (r=0.924, 0.930, 0.920 respectively).
2) the concurrent validity between Narration C-Test No.4 and STEP examination was high (r=0.7333).
3) the face validity of C-Tests was slightly favored by Japanese teachers of English at high schools. However, college/university level instructors' responses reflect their cautious and conservative attitudes toward C-tests. Several intermediate and lower level students scored zero points in each C-Test.

This study confirmed the results of the experiment conducted on 42 college freshmen in 1992, which showed 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 (Mochizuki 1994). The author demonstrated some possibility of a Narration C-test being used as part of a placement examination such as the Pre Second Grade STEP examination. Further research is needed to study the content validity of a Narration C-Test.

Note

1. C-Test No.1, Which used Argumentation, "What Democracy Means. "(336 wods) (Shimizu, et al., 1983), C-Test No.2, which used Explanation, "The Pony Express"(367 words)(Kaneda et al., 1971), C-Test No.3, which used Description, "Robert Reich"(353 words), adapted from the 1995 *Los Angeles Times* newspaper article, and C-Test No.4, which used Narration, "The Lock Keeper"(413 words), (Kaneda, et al. 1971).

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

Jafarpur,A. (1995). Is C-testing superior to cloze? *Language Testing*, 12, 194-216.

Kaneda,M., Horiuchi,K., Yamaguchi, S., Shimizu,K., Ohta, H. & Ohkawara,R. (Eds). (1971). The lock-keeper & The pony express., *Multi-level reading program, yellow.* (pp.9-12)

Klein-Braley,C. (1981). Empirical investigations of cloze tests. Unpublished doctoral

dissertation, Universitat Duisburg, Duisburg, Federal Republic of Germany.

Klein-Braley,C. & Raatz,U. (1984). A survey of research on the C-Test. *Language Testing*, 2, 134-146.

Mochizuki,A.(1991). Multiple-choice (M-C) cloze tests. *ARELE* 2, 31-40. Tokyo: The Federaion 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. & Conrad,C. (1971). The cloze technique and ESL proficiency. *Language Learning*, 21, 183-195.

Porter,D. (1978). Cloze procedure and equivalence. *Language Learning*, 28,333-341.

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.

Shimizu, T.(Ed) (1983). Democracy. *The international English II.* 146-148. Tokyo:Kaitakusha.

Appendix

A part from C test No.4 Narration "The Lock-Keeper"

Notes

sign ____がいる気配	weeds 雑草	body 死体
edge 端、へり	closely じっと	phoning phoneは「電話する」
search を探す	sergeant 巡査部長	shudder 身震いする

It was evening as I walked along a path near the canal.

Ahead of me was a lock, its wooden gates were closed. The lock-keeper stood outside his small, grey house, watching the canal.

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

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

"Quick, " I said, "There's a woman in the water."

"Where?" he said.

There was no sign of the woman.

"She fell in just a moment ago," I said. "You must have seen her." I was standing right on the edge of the canal, looking into the water. She couldn't have disappeared so quickly.

"She's not there," said the lock-keeper. "You can(1) search as(2) long as(3) you like(4) —you'll never(5) find her(6). "Perhaps she(7) was caught(8) in the(9) weeds under(10) the water(11), or perhaps(12) she had(13) come