Effectiveness of the Cloze Test for Beginning Japanese EFL Learners in Comparison with the Discrete-point Test and 4 Communicative Tests

Akihiko MOCHIZUKI

As part of a 1996-98 Education Ministry sponsored grant-in-aid research, this study examines the effectiveness of the Cloze Test for beginning EFL learners in Japan. The following tests were administered to about 460 third-year students aged 14 to 15 at three public middle schools in Central Japan from October 1997 through mid-March 1998: the Third Grade Test of the Society of Testing English Proficiency (STEP), a Cloze Test (consisting of a 26-item Narration and a 26-item Conversation text) and four Communicative Tests (Listening (L), Speaking (S), Reading (R) and Writing (W) Tests). A questionnaire about students' perception of the Cloze Test was also given to 58 students. Results indicate: First, the Cloze Test is effective in terms of item difficulty, item discrimination, reliability, validity, ease of construction, ease of scoring and ease of interpretation. Second, the Cloze Test turns out to be strongly related to a Discrete-point Test, STEP. The test was also found to be strongly related to the W, R, and L Tests, with the R Test the highest. However, it was weakly correlated with the S Test. Third, the Cloze Test seems to be capable of discriminating between upper-level and lower-level groups. Fourth, the Cloze Test does not seem to be high in face validity. The study thus indicates that a 50-or-more item Cloze Test can be used as a placement test in the event that a teacher is too busy to construct a regular full-fledged test.

I Introduction

Since the cloze test was developed by Taylor (1953) to study the readability of prose, it has been regarded as a measure of various aspects of language usage: reading comprehension, listening comprehension, the quality of translation of technical training manuals used by the military, knowledge of vocabulary, the reader's I.Q. and even oral ability. Many researchers have conducted experiments on cloze tests and they have revealed various findings on the basis of the tests conducted on university students (Klein-Braley,1997, Fotos, 1991, Sasaki,2000, Brown, 1980, Chapelle and Abraham, 1990, Dorynyei and Katona, 1992, Bachman, 1985, Brown, 1983, Sciarone and Schoorl, 1989[1]. However, very few studies have dealt with middle school students (Piper, 1983[2], Klein-Braley, 1985[3]). In this paper, I therefore explore how beginning Japanese EFL learners perform on the cloze test and how the cloze test is related to the discrete-point test and tests of communicative competence, namely, communicative tests. A further point of inquiry is concerned with how beginning learners of English perceive the Cloze Test. This study forms part of the results from a 1996-1998 grant-in-aid research (Kiban Kenkyu (C) (2) No. 08680285) " Measurement and Evaluation of English Communicative Competence -Integrative vs. Discrete-point Tests & Analytic vs. Holistic Evaluations ".

II Cloze Tests, Discrete-point Tests and Communicative Tests

1 *Kinds of Cloze Test*

The cloze test which Taylor (1953) invented is a type of test which is constructed by deleting every *n*th words mechanically regardless of their function or meaning from a passage and requiring the subject to fill in the blanks. Cloze can be divided into two kinds: standard cloze tests and modified cloze tests. Standard cloze tests are as explained above and they are sometimes called fixed-ratio cloze tests (e.g.Carol et al.,1990). Modified cloze tests can be divided into (1) rational cloze tests, in which words are deleted rationally on the basis of test-writer's criteria (Bachman, 1985, Chappelle & Abraham ,1990, etc.); (2) multiple-choice cloze tests as developed by Jonz (1976) and Porter(1976), which provide multiple choice alternatives, where the construction depends on the depth of linguistic attainment and fineness of stylistic discrimination of the subject; (3) C-tests as developed by Raatz and Klein-Braley (1981), in which the second half

[1] The subjects were 74 native speakers of Indonesian who were seeking admission to Delft University of Technology, the Netherlands, aged between 17 and 21.

[2] Piper conducted the cloze and C-test on Dutch students, with an age range of 13-81, 44 in number, two-thirds being women.

[3] Klein-Braley administered the German C-Test to 398 German L1 speakers attending the three different types of German secondary schools.

of every second word is deleted instead of a whole word, or according to the "rule of 2;" (4) cloze-elide tests as developed by Manning (1986), which insert words that do not make sense in the context of the passage and in which the subject points out which inserted words are irrelevant. Of those kinds of cloze tests, a standard cloze test or a fixed-ratio cloze test was used for this research.

2 *Problems of the cloze test*

Oller (1973) divided tests into two kinds. The first is discrete-point tests, which test only one discrete point, namely one point of grammar, phonology, vocabulary, etc. at a time; for example, an Auditory Comprehension test might ask candidates to discriminate between [i]and [i:] in sh<u>i</u>p and sh<u>ee</u>p. The other kind involves tests of integrative skills, which test full-scale language use, and language use in communication; for example, dictation. Cloze tests are regarded as a typical integrative test. Since cloze tests were developed, they have gained much popularity among researchers as if they were a language testing panacea; indeed, researchers and teachers have tried using them for many purposes. However, in recent years, certain problems with the cloze test have been pointed out. These problems can be roughly divided into the five following areas:

a *Scoring methods*: The scoring methods for the cloze test can be roughly divided into the following two broad categories: an exact answer method and an acceptable answer method. Both of these are problematic. 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), Ruddlel (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, such as 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 (Darnell,1968), and multiple-choice methods. The acceptable answer method, however, reduces advantages of easy scoring and preparation.

So, which scoring method should be adopted? Johnson (2001) argues that either will do, stating "Whichever method [the 'exact' method or the 'acceptable' method] is used does not seem to make much difference as far as *ranking* the learners by result (putting them in order) is concerned" (p.296). Therefore, in this research, an exact answer method was adopted in terms of scoring time.

b *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 deletion begins, that is, according to what is deleted" (p.336). In this research, the cloze test used two passages, a Narration and a Conversation. The cloze test using the Narration text keeps the first paragraph intact and deletes every 7 th words. The Conversation-text cloze test keeps the first turn-taking part, namely, a greeting and its reply, intact and deletes every 7th words.

c *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 the Split Half method) and valid (ranging from 0.04 to 0.71). Colemen (1971) 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, p.135) state, "particularly for homogeneous samples (classroom groups or monolingual groups) cloze tests tend to have unsatisfactory reliability and validity coefficients" . In this research, homogeneous samples, namely, 460 third-year junior high school students took the cloze test.

d *Kinds of texts*: In order to solve the above-mentioned problems, Raatz and

Klein-Braley (1981) developed a new type of cloze test, the C-Test. One of the reasons which they cited was that the test performance of the cloze test is affected by the text topic and difficulty, whereas in the C-Test these factors are minimized by the use of several different short texts. Text selection remains however a difficult task for text constructors. Mochizuki (1984) conducted M-C cloze tests by using 4 different kinds of texts, namely, Narration, Description, Explanation, and Argumentation on 96 second-year high school students and found out that the correlation between the score of the M-C cloze based on Narration (one of four kinds of texts which involves a passage that narrates something which happened either in reality or in the imaginary world and which could draw on, for example, excerpts from newspaper articles or novels) and the score of the exam which extended over the work of the whole year was higher (r =0.756) than that between any other kind of M-C cloze and the exam which extended over the work of the whole year. Mochizuki (1994) further indicates that, as a result of conducting four kinds of texts in the C-Tests on 42 college freshmen, the reliability of the Narration C-Test is the highest (r =0.928) and that there is a fairly high correlation between the scores of the STEP test and the Narration C-Test. He confirmed that finding by conducting the C-Tests with four different kinds of texts on 727 second-year high school students. For this reason, a Narration text is used for one of the two part-cloze test. The other part of the cloze test uses a Conversation text, the purpose of which is to investigate the relationship between the conversation-text cloze test and a speaking test for future analysis, not for the present study.

e *What cloze tests measure*: There are three main views as to 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 their stronger correlation with grammar tests than with reading tests (Alderson, 1983) ; and, (c) cloze tests measure overall proficiency, because of the strong correlation with dictation, reading tests and essay writing, in addition to standardized proficiency tests (Chavez-Oller et.al., 1985; Brown,1993). However the current consensus among most researchers is that cloze tests work as an instrument for a measure of overall language ability.

This study is intended to study the relationship among the cloze test, the Discrete-point Test and four kinds of Communicative Tests. Therefore, characteristics of Communicative Tests, the ways of assessing four Communicative Tests — Holistic, and Analytic Evaluations, and how to evaluate tests will be dealt with.

1 *Communicative Tests*

As Johnson (2001) points out, communicative testing was given birth to together with Communicative Language Teaching. As noted by researchers such as Savignon (1983, p.254), Brown(1994), Bachman (1991, p.678) and Weir (1993, p.167), Communicative Tests, tests of communicative competence have the following characteristics: (1) Communicative Tests are integrative tests, (2) Communicative Tests are direct tests, (3) Communicative Tests use real life situations as tasks, (4) Communicative Tests set up situations, and (5) Communicative Tests measure functions (Mochizuki 2000). The author and co-researcher constructed four kinds of Communicative Tests following the framework proposed by Weir (1990).

2 *Traditional, holistic and analytic evaluations*

Four kinds of Communicative Tests, Listening, Speaking, Reading, Writing Tests, were conducted. Of the four tests, Speaking and Writing Tests, Speaking and Writing Tests were scored in three ways of evaluation, that is, traditional, holistic, and analytic evaluations. Listening and Reading Tests were marked in a traditional way. The following are explanations about three kinds of evaluation.

a *Traditional Evaluation*: Traditional Evaluation (abbreviated to TE) is a method used to grade a test mechanically by counting the errors made by each subject and deducing the number from a given total score. At Japanese middle and high schools as well as colleges, this way of scoring is the most prevalent. Heaton (1975, p.175) remarks that "Although this (mechanical accuracy or error-count method) is the most mechanical of all methods, it is the least valid and is not recommended." In terms of the present study, the author and his partner Yamada, N. thus decided to mark the Writing and Speaking Tests in terms of communicability, that is, whether each sentence can succeed in communication or not.

b *Holistic Evaluation*: Holistic Evaluation (HE) is a method in which "the teachers use a single general scale to give a single global rating for each student's language production" (Brown, 1996, p.61). This has also been called "impressionistic scoring", "global rating" and "the impression method". The merits of HE are discussed by researchers such as Perkins (1983, p.652), Hughes (1989), Heaton (1975) and Weir (1990). The drawbacks of HE are mentioned by researchers such as Hughes (1989, p.86), Homburg (1984), Polio (1997), Bachman and Palmer (1996) and Weir (1990).

c *Analytic Evaluation* : Analytic Evaluation (AE) is a method in which "the teachers rate various aspects of each student's language production separately" (Brown, 1996, p.61). This procedure has been gaining more popularity in recent years in middle and high schools in Japan. The merits and demerits of AE are discussed by Hughes (1989), Weir (1990), Bachman and Palmer (1996). With regard to how we should adopt either the HE and AE, Mochizuki (2001, pp.3-4) shows some suggestions that might lead to a solution: "First, as Hughes (1989) recommends, it is wise to use both methods with one as a check on the other. Second, it is necessary to train raters before the administration of the test. Third, multiple scoring (scoring a few times) increases reliability. Hughes (1989, pp.86-87) states that scoring four times makes the reliability high.

3 *How to compare and evaluate tests*

This research aims at assessing the cloze test in relation to the discrete-point test and Communicative Tests. Therefore, the author would like to briefly review the ways of evaluating tests themselves.

Most teachers construct classroom tests without being too conscious of how to construct them properly. In order to construct a reliable and valid test, teachers need to pay attention to the test construction from the beginning. Advice from Brown (1996) and Bachman and Palmer (1996) can be useful in this regard. Also in terms of what teachers should take into account in constructing tests, Weir (1993)'s "three-part framework," which explains what activities to assess in constructing communicative tests, performance conditions, and criteria for assessment, respectively, deserves to be recommended.

However, once teachers have constructed tests, it is usual for them not to evaluate them. Most test constructors as well as students are concerned with the results of the tests alone. In fact, the test must be evaluated in terms of such aspects as washback effects and interpretation of scores. Here are some suggestions for criteria for evaluating tests. Bachman and Palmer (1996) proposed the concept of "usefulness" as the main criterion. This criterion is composed of six elements: reliability, construct validity, authenticity, interactiveness, impact and practicality. Of several kinds of validity, they also emphasize construct validity, meaning "the extent to which we can interpret a given test score as an indicator of the ability(ies), or construct(s), we want to measure" (p.21). Another element which must be paid attention to is authenticity, which they think is "the degree of correspondence of the characteristics of a given language test to the features of a TLU [Target Language Use] task" (p.23). "Impact" means an influence "on society and educational systems and upon the individuals within those systems" as is inferred by the term. Bachman and Palmer thus identify several important aspects for evaluating tests.

Guerrero (2000) adopted Messick's framework (1989) for analyzing the Four Skills Exam of Spanish (FSE) which comprised 4 facets: (1) construct validity (To what extent, statistically, is the FSE a reliable and valid measure? Did the test performance of the examinees vary as a function of their Spanish-language background?). (2) relevance and utility (What evidence is there of content relevance and coverage? What evidence is there to support the utility of the scores (e.g. pass/fail) for the applied purposes (i.e. teaching in a bilingual setting)?). (3) value implications (What was the pass and fail rate of the examinees on the different parts of the test and the test as a whole? To what extent is score interpretation valid?). (4) social consequences (Did the FSE fulfill its intended purposes? Were there any unintended social consequences related to its use?) Guerrero usefully showed how to apply Messick's framework in the FSE.

In comparing two cloze tests, three C-Tests, two Multiple-choice cloze tests, two cloze-elide tests, dictation and the external criterion test called DELTA, Klein-Braley (1997) used usability criteria, which are made up of five factors: (1) difficulty (2) reliability (3) validity (4) ease of construction and (5) scoring. With regard to validity, she examined concurrent validity in her study.

In this article, the author would like to evaluate the standard cloze test conducted on beginning Japanese learners of English and to consider the relationship of the cloze test with a discrete-point test and communicative tests. The criteria proposed here are "effectiveness" criteria, which are composed of the following seven facets: (1) difficulty, (2) discrimination, (3) reliability, (4) validity, (5) ease of construction, (6) ease of scoring, and (7) ease of interpretation. Thus, the following four research questions (RQs) were derived:

RQ1. How the Cloze Test can be evaluated in terms of Effectiveness, which is made up of (1) item difficulty, (2) item discrimination, (3) reliability, (4) validity, (5) ease of construction, (6) ease of scoring, and (7) ease of interpretation.

RQ2. How the Cloze Test is related to the Discrete-point Test and Communicative Tests.

RQ3. Whether high scorers on the Cloze Test will acquire high scores on the STEP also.

RQ4. How beginning Japanese EFL learners of English perceive the Cloze Test.

III Method

1 *Purpose*

The purposes of the study were to investigate the effectiveness of standard cloze tests; to investigate the relationship between a Cloze Test, on one hand, and a Discrete-point Test and 4 Communicative Tests, on the other hand; to investigate the relationship between scores on the Cloze Test with the STEP in 3 groups — upper, middle, and lower groups; and to investigate the beginning EFL learners' perception toward the Cloze Test.

2 *Subjects*

About 460 third-year students aged 14 to 15 at three public middle schools (into high, intermediate and low levels) in Shizuoka Prefecture. Of the 460 total subjects, the data from 273 subjects at two middle schools, which were at intermediate and low levels, were used for the Writing Test and 106 subjects from one school equipped with a language laboratory, the level of which was intermediate, were used for the Speaking Test. The data from 369 students were used for the Reading and Listening Tests, the 52-item Standard Cloze Test, and the Society for Testing English Proficiency 3 rd Grade Test (STEP).

The author would like to summarize the number of the subjects whose data were adopted for this research in the following table.

Table 1 Breakdown of the subjects and kinds of tests.

Schools	Level	N	STEP	Cloze	S	W	L	R	Quest
JH 1	High	96	96	96	0	0	96	96	0
JH 2	Intermediate	106	106	106	106	106	106	106	58
JH 3	Low	167	167	167	0	167	167	167	0
Total		369	369	369	106	273	369	369	58

Notes. JH=Junior High School, Quest =Questionnaire.

3 *Materials*

The following materials were used in this experiment:

1) A 52-item Cloze Test in which every 7 th word was deleted was made up of a 26-item Cloze 1 [4] (Cloze 1) using a Narration text, and a 26-item Cloze 2 [5] (Cloze 2) using a Conversation text.

2) A 42-item STEP[6] composed of five subtests (stress finding for 6 items,

[4] Cloze 1 passage is a 244-word story from a collection of Entrance Exams for Kyoto Public High Schools (1990), Fuji Kyoiku. p.16. Flesch-Kincaid Grade Level = 2.8,

[5] Cloze 2 passage is a 227-word conversation from a collection of Entrance Exams for Wakayama Public High Schools (1990),Fuji Kyoiku. p.16. Flesch-Kincaid Grade Level =1.9,

[6] An 8-part 42 item external criterion STEP was an assortment of items from past Third Grade STEP written examinations.

vocabulary for 4 items, grammar for 12 items, composition for 6 items, and reading comprehension for 14 items), was constructed by using an assortment of items from the past Third Grade STEP written examinations. The reason why the STEP was chosen as a Discrete-point Test and an external criterion as well is that it is a high-stakes test with a high reliability and validity and whose rationale is in line with the course of study issued by the Ministry of Education, Science and Technology of our country.

3) In accordance with Weir's "three-part frameworks" , four kinds of Communicative Tests (namely, Speaking[7], Writing[8], Reading[9], and Listening[10] Tests) were constructed, keeping in mind the fact that communicative tests are integrative direct tests, using real life situations as tasks that set up situations and measure functions (see Mochizuki, 2000, pp.261-262).

4 *Procedure*

The above-mentioned cloze test, STEP and four kinds of Communicative Tests were administered to the students from October 1997, through mid-March 1998 (Mochizuki, 1999, 2000).

With regard to scoring the four kinds of Communicative Tests, the Reading and Listening tests were marked by TE, which reduces one or two points out of three in accordance with the selected criterion. For the Speaking and Writing tests, which are both productive tests, three methods were employed, namely, TE, HE and AE. TE for Writing and Speaking Tests is slightly different from an ordinary TE in that it does not ignore grammatical accuracy but instead puts more emphasis on communicability, that is, whether each sentence is communicable. It

[7] Speaking Test: Akagawa (1991) *Cambridge Eiken 2kyu Goukaku Enshu*. Tokyo: Kenkyusha. 50-73.

[8] Writing Test: one item was from Ohta, R.,Ito, K.and Kusaka, T. (1984). "In the library" , *New Horizon Revised Edition English Course* 2. Tokyo: Tokyo Shoseki. P.71. The other two items were original ones.

[9] Reading Test: Ohta, R., Ito, K. & Kusaka, T. (1983). *New Horizon English Course* 2. Tokyo:Tokyo Shoseki. 42-45, Ito, K. (1982) *New Scope English Course* I. Tokyo: Tokyo Shoseki. 42-42, & Iamura, M., Noya, T, & Torii, T. (1985) *New Prince English Course*. Tokyo: Kairyudo. 11-15.

[10] Listening Test: Mochizuki, A, & Yamada, N. (1996) *Watashino Eigo Jugyo*. Tokyo:Taishu Kan.134-276.

marks each sentence using a 3-point scale with the highest point being 3, a grammatically incorrect but communicable sentence, the middle point being 2, a grammatically incorrect and rarely communicable sentence, the lowest point being 1. HE is the way to mark the test globally on the basis of the rater's overall impressions by using a 5-point scale, with the highest being five, and the average being 3 and the lowest being 1. AE is the way to mark the test analytically through a multi-trait profile of a script in terms of analytic features, such as organization, grammatical accuracy, and attitudes. Each analytical feature is marked on the basis of a 3-point scale.

After the four kinds of Communicative Tests were completed by the students, the same tests of each kind were scored twice again by the same rater at an interval of about one month. The tests were later exchanged between the two raters and marked again from August, 1997 through mid-March, 1998. Each answer sheet was scored four times by two raters.

IV Results

1 *Item Difficulty*

Item Difficulty (or Item Facility, IF) was calculated in the formula (n total/n correct) as shown in Table 2. By following suit of Anderson et al. (1991), the author categorized the item difficulty (p) values equivalent to item facility into three groups: easy items $p > .67$;average items $.33 \leq p \leq .67$; and difficult items $< p$ .33 .

Table 2 Average Item Facility of Cloze Test (1,2, and Total) N=369

Tests	IF	Mean	SD	Range	Max	Min	Full Score
Cloze 1	0.542	14.10	6.73	26	26	0	26
Cloze 2	0.555	14.49	6.75	26	26	0	26
Cloze Total	0.550	28.62	13.12	52	52	0	52

Note: Cloze Total = (Cloze 1) + (Cloze 2)

Table 2 shows that item difficulty indices of Cloze 1, 2 and Cloze Total are close to 0.5. It means that items of these two cloze tests are average items.

2 *Item discrimination*

Item discrimination was here calculated by the formula (IF upper - IF lower) as shown in Table 3. In this experiment, the author took the upper 27% and lower 27% respectively from the whole group.

Table 3 Item discrimination (ID) of Cloze Test (1,2, and Total)

Test	ID	Test	ID	Test	ID
Cloze 1	0.629	Cloze 2	0.621	Cloze Total	0.625

Table 3 shows that item discrimination indices of Cloze 1,2,and Total are all over 0.6. According to Ebel (1972, p.399), the index of 0.40 and up indicates "very good items." Therefore, items in subtests and the whole Cloze Test were found to be very good items. Of both subtests, Cloze 1 seems to have a slightly more discriminatory power.

3 *Reliability*

The overall reliability of STEP calculated by the Split Half Method for 369 students was r=0.90. The reliability coefficients of the subtests of the STEP for just 100 students at one public middle school were calculated by Cronbach α because of the limitation of time as shown in Table 4.

Table 4 Statistics of STEP (n =369)

	Reliability	Mean	SD	Range	Max	Min	Full Score
Stre	0.65	7.42	3.18	12	12	0	12
Voca	0.49	5.30	2.46	8	8	0	8
Gram	0.54	21.79	7.17	30	30	0	30
Comp	0.70	9.72	4.89	18	18	0	18
Read	0.70	24.98	10.14	36	36	0	36
Total	0.90	69.23	23.60	98	104	6	104

Notes: Stre = Stress finding (6 items x 2 pts =12 pts)
Voca = Vocabulary (4 items x 2 pts =8 pts)
Gram = Grammar (6 items x 2 pts+6 items x 3 pts=30 pts)

Comp = Composition (6 items x 3pts =18 pts)

Read = Reading Comprehension (6 items x 2 + 4 items x 3 +4 items x 3 = 36 pts)

With regard to the way to calculate cloze reliability by using K-R 21, Klein-Braley (1997, p.67) states that "this is not legitimate since the items in the tests are not statistically independent." Therefore, reliabilities of the subtests of the Cloze test were calculated by Cronbach α. Since the overall reliability coefficient was 0.90, STEP was found to be a very reliable test.

Table 5 Reliabilities of Cloze test (1,2, and Total) (n=369)

Test	r	Test	r	Test	r
Cloze 1	0.920	Cloze 2	0.920	Cloze Total	0.958

Table 5 shows reliabilities of the Cloze Test calculated by Cronbach α. The reliabilities of Cloze 1, 2 and Total were all found to be very high. Cloze 1 and Cloze 2 have the same reliability.

Table 6. Reliabilities of Four Communicative Tests (S,W,R.& L)

Kinds of	Intra-rater Reliability			Inter-rater Reliability			N
Tests	TR	HE	AE	TR	HE	AE	
Speaking	0.992	0.952	0.963	0.969	0.894	0.915	106
Writing	0.996	0.971	0.974	0.985	0.939	0.953	273
Reading	0.997			0.993			369
Listening	0.998			0.996			369

*TR =Traditional Evaluation, HE = Holistic Evaluation, AE = Analytic Evaluation

Table 6 shows the reliability measures of the four Communicative Tests (S, W, R & L). These were calculated not by KR 20 nor KR21 nor Cronbach α but by Intra-rater and Inter-rater Reliability, since each of the items of the four tests was scored not in the form of binary (0-1) data but in three-way evaluations: Traditional Evaluation, Holistic Evaluation, and Analytic Evaluations (see Mochizuki, 2001). In this study two raters rated each test two times in each of the three ways of evaluation, 12 scores for each test were generated (2 raters x 2 iterations x 3 ways of evaluations yields 12 scores). In calculating intra-and inter-rater reliabilities, the author used the first time scoring of 2 iterations for each way of evaluating and the author adopted Cronbach α following Bachman (1990)'s explanation (pp.178-181) to calculate the correlation coefficients. As a result, the reliability of the four Communicative Tests was shown in each case to be high or very high in all three types of evaluation.

4 *Validity*

With regard to validity, as Hughes (1989,p.27) points out, it has been divided into four kinds, face, content, criterion-related, construct validities. However, Messick (1989) stated that validity is a "unitary concept (p.13), and Bachman (1990a) introduced validity as a unitary concept regarding test interpretation and use. Chappelle (1999) summarized the contrasts between past and current conceptions about validity as follows. In the past (1) "validity was considered a *characteristic of* a test: the extent to which a test measures what it is supposed to measure" (2) "Reliability was seen as distinct from and a necessary *condition for validity,* " (3) Construct validity was seen as one of *three types of validity* (the three validities were content, criterion-related, and construct) " ,whereas currently (4) "validity is considered an *argument* concerning test interpretation and use: the extent to which test interpretations and uses can be justified" , (5) "Reliability can be seen as *one type of validity evidence*, and (6) "Validity is a *unitary concept* with construct validity as central (content and criterion-related evidence can be used as evidence about construct validity) " (p.258) (The author numbered (1)～(6) by quoting them from Chappelle's Table 1.) All this literature about validity requires us to collect "almost all forms of validity evidence" (Messick, 1989, p.17). Also consequences must be taken into consideration as Messick (1989) included "social consequences of test interpretations and use in construct validation" (p.17) and Chappelle mentioned "the consequences of testing" in her table (p.258).

In regard to validity, Klein-Braley (1997) used criterion-related, namely, concurrent validity, to evaluate several kinds of cloze tests. This study investigated content, construct, criterion-related (concurrent) , face validities but not social consequences which are hard to assess.

a *Content validity*: First, concerning content validity, the content of the cloze test is composed of a Narration test and a Conversation text as already mentioned in 1. *Kinds of cloze test ,d Kinds of texts*. Mochizuki (1984,1994) showed that the cloze test with a Narration text is assumed to reflect Japanese EFL learners' proficiency very well. In this study the author conducted a preliminary cloze test with a Narration text and a Conversation text on a small number of third -year junior high school students. The result showed that it discriminated them in terms of their level. Brown (1983) revealed that the cloze test is valid in terms of cohesive devices since cloze tests that are based on every *n*th word deletions whether they start at different points tend to tap equally well any cohesive devices that may exist in the prose. Therefore, the author considered this study to be content valid.

Concerning four kinds of Communicative Tests, the author constructed them by following Weir's (1990) three-part framework and the content validity was judged to be high by the author and his co-researcher Yamada,N.

b *Criterion-related validity*: In regard to criterion-related ,that is, concurrent validity, in order to investigate concurrent validity, the correlation coefficients were calculated between the Cloze Test and STEP, as shown in Table 7.

Table 7 Correlations between Cloze Test (1,2 and Total) and STEP (n =369)

Correlation	r
Cloze 1 — STEP	r=0.837**
Cloze 2 — STEP	r=0.839**
Cloze Total — STEP	r=0.861**

** = Significant at the 0.01 level (2-tailed)

The Cloze Test (1,2,and Total) was found to be strongly correlated to STEP, as shown in Table 7. It can be said that Cloze 1 and Cloze 2 are almost the same in validity. Cloze Test 1 shares about 70.1% ($.837^2$ x 100) of variance with STEP, while Cloze Test 2 shares about 70.4 % ($.839^2$ x 100) with STEP. Both the Cloze Test and STEP seem to measure the overall English proficiency of the subjects.

Table 8 Correlations between Cloze Test (1,2,and Total) and Subtests of STEP (n =369)

	Stre	Voca	Gram	Comp	Read	Total
Cloze 1	0.546	0.618	0.751	0.691	0.760	0.837
Cloze 2	0.532	0.619	0.757	0.704	0.757	0.839
Cloze Total	0.557	0.634	0.776	0.718	0.777	0.861

Cloze 1 with a Narration text is found to be highly related to Grammar (r =0.751) and, most strongly, to Reading Comprehension (r =0.760). Cloze 2 with a Conversation text is found to be strongly correlated to Grammar and Reading Comprehension (both r =0.757) and Composition (r =0.704).

Table 9 Correlations between Cloze and four Communicative Tests

	TR	HE	AE	n
Cloze 1 :S	0.376**	0.364**	0.354**	106
Cloze 2 :S	0.378**	0.340**	0.338**	106
Cloze Total :S	0.396**	0.369**	0.365**	106
Cloze 1 :W	0.727**	0.690**	0.683**	273
Cloze 2 :W	0.736**	0.700**	0.696**	273
Cloze Total :W	0.756**	0.716**	0.713**	273
Cloze 1 :R	0.806**			369
Cloze 2 :R	0.792**			369
Cloze Total :R	0.817**			369
Cloze 1 :L	0.780**			369
Cloze 2 :L	0.761**			369
Cloze Total :L	0.791**			369

Table 9 reveals the following three things. First, with regard to the correlation between the Cloze Test (1,2 & Total) and Speaking Test, the correlation coefficients were all less than .40 in any of the three types of evaluation, namely Traditional, Holistic and Analytic. Second, regarding the correlation between the Cloze Test (Cloze Total) and the Writing Test, the correlation coefficients in

Traditional, Holistic, Analytic Evaluations were high (r= .0756, 0.716, 0.713 respectively). Third, concerning the correlation between the Cloze Test and the Reading Test, the Cloze Test was found to be strongly related to the Reading Test (r = 0.806,0.792,0.817 in Cloze 1,2 and Total, by Traditional Evaluation). Fourth, regarding the correlation between the Cloze Test and the Listening Test, the Cloze Test was shown to be closely related to the Listening Test (r = 0.780,0.761,0.791 in Cloze 1,2 and Cloze Total) by Traditional Evaluation.

c **Face validity**: As to face validity, this means how good or right a test in question looks to the test takers. In order to know how beginning Japanese EFL learners perceived the cloze test (results shown in Table 10), the author conducted a questionnaire on 58 students out of 460 participants, just that small portion who submitted the responses voluntarily after they were asked to fill it in at one intermediate level public middle school.

Table 10 Perception of beginning EFL learners toward the Cloze Test n = 58

Q.1 Do you think that this Cloze Test is a good test as an English test?

Yes. (37.9%) , No.(13.8 %), Don't know. (48.3 %)

Q.2 What abilities do you think this Cloze Test measures? Check as many options as you think suitable. The top five responses are shown here.

1. the ability to predict and infer (81.0%),
2. grammatical knowledge (77.6 %),
3. English reading comprehension (62.1%),
4. English spelling (55.2%),
5. English vocabulary (48.3%) (Since the respondents were asked to check as many options as they think suitable, the sum of the percentage points totals over 100 %.)

Q.3 Do you agree to this Cloze Test being introduced into an entrance examination to high schools?

Agree.(55.2%), Disagree.(27.6%), Don't know.(17.2%)

(Only To Those who answered "Agree") To what extent should this Cloze Test be introduced into the entrance examination to high schools?

As part of the entrance examination to high schools. (96.9 %)

This Cloze Test can replace the current entrance examination to high schools. (3.1%).

(Translated into English by the author.)

Question 1 is posed to investigate the face validity of the Cloze Test. The results showed that 48.3% answered “Don't know” , whereas 37.9 % of the students “ Yes” and 13.8 % responded negatively to the question "Is this cloze test a good test as an English test?" It means that about half of the students showed reservations. Therefore, it cannot be said that this cloze test is high in terms of face validity. With regard to what abilities the Cloze Test measures, the subjects' perception about the Cloze Test measuring grammatical knowledge and reading comprehension matches the statistical analysis of this test, as shown in Table 8. Concerning the adoption of the Cloze Test as part of the high school entrance examination, more than half of the respondents favored it, although this result does not necessarily go with Q.1. This shows that beginning learners seemed to be very interested in such a 'new' type of test although they have some doubt about whether the cloze test is a good test. It cannot be said that the cloze test is face valid.

d *Construct validity*: Lastly, construct validity of the cloze test requires us to collect as much evidence as possible from content, criterion-related, and face validities. Concurrent and content validities were deemed to be acceptable. However, the cloze test was not highly face valid. Overall , the cloze test can be judged to be acceptable as to construct validity.

V Discussion

In II Cloze Tests, Discrete-point Tests and Communicative Tests, four research questions were set up. Let us examine each research question.

With regard to Research Question (RQ) 1 of the Effectiveness of the Cloze Test, the author investigated the Cloze Test in terms of seven criteria as follows. First, item difficulty indices of Cloze 1, 2 and Total are close to 0.5, average difficulty level ($.33 \leq p \leq .67$). Second, the item discrimination indices of Cloze 1,2 and Total can be regarded as indicating best items, bearing in mind that “0.4 and up “in item discrimination is the optimum level. Third, the reliabilities of Cloze 1, 2 and Cloze Total are very high (r= 0.920, r =0.920, r =0.958). Fourth, the concurrent validity of the Cloze Test was very high (r = 0.837, r = 0.839, r

=0.861 respectively). Fifth, concerning ease of construction, as already pointed out by many researchers, cloze tests are easy to construct, a merit for test constructors. This was the case with this cloze test featuring two subtests, one with a narration text and the other with a conversation text. Sixth, regarding ease of scoring, the Cloze Test is easy to score. In this study, exact word scoring was adopted, since such researchers as Taylor (1953) and Bormuth (1964, 1965a,1965b) and Johnson (2001) pointed out that there is no significant difference between the exact word scoring and the acceptable word scoring. A further reason for choosing this scoring method was economy of time. The final criterion is ease of interpretation. If scorers have difficulty interpreting the scores of the subjects, the test will not be favored by them. Also the scores of the test must be easily understood by the subjects after they have received their result. The Cloze test provides the scorers and subjects with a single score, which shows overall proficiency. This is easy to interpret. In view of the fact that there are a tiny number[11] of middle school teachers in Japan who know about the cloze test, cloze testing, its nature, strengths and weaknesses should be taught to them at various seminars and study meetings.

Research Question 2 deals with the relationship of the Cloze Test with the Discrete-point test and also with four Communicative Tests (S, W, R & L). First, with regard to the relationship of the Cloze Test with the Discrete-point test (STEP) previously discussed in terms of concurrent validity (See Table 7), the Cloze Test is closely related to STEP; above all, among the subtests of STEP, it is closely related to Grammar ($r = 0.751$) and Reading Comprehension ($r = 0.760$). It should be noted that the focus here is on concurrent validity (i.e. the relationship between the Cloze Test and the external criterion, namely, STEP). RQ 3 further explores that relationship. The high correlation between the Cloze Test with Narration and Reading Comprehension reminds us that the Narration text Cloze test requires integrative thinking. This is similar to grasping the meaning of a whole passage in the Reading Comprehension section of the STEP.

[11] The author gives a lecture on communicative testing to middle and high school teachers of English at the Open Lecture at the University of Tsukuba every July and asks them whether they have heard about the cloze test. However, very few teachers have.

Cloze 2 is also strongly correlated to Grammar and Reading Comprehension (both r = 0.757) and Composition (r = 0.704). The Conversation text Cloze 2 encouraged the author to hypothesize a strong correlation with Stress finding, as both of them have something in common in respect of association with sounds, but in fact there is no bigger relationship between Cloze 2 and Stress finding than that between Cloze 1 and Stress finding. Cloze 2 is composed of a 26-line dialogue between a Japanese middle school boy, Jiro, and a boy from an English-speaking country, Tom. This dialogue is a daily conversation in which the Japanese boy invites Tom to join a trip to Kyoto together with foreign children of Jiro's father's employees. It can be categorized as a kind of Narration in a dialogue form. This requires participants to think of the plot of the whole story as well as the dynamic human relations revealed in the conversation, while at the same time keeping an eye on grammar and structure of sentences. In contrast, Stress finding in STEP deals with what word carries stress in short two-sentence dialogues. Subjects are required to pay attention to a single place in a response. Therefore the complexity of the task in Cloze 2 is different from that in Stress finding. Overall it may be safe to say that the Cloze Test is strongly related to Grammar (r = 0.776), Composition (r = 0.718) and Reading Comprehension (r = 0.777). The reason for the strong correlation between Composition and the Cloze Test seems to be that both are similar in eliciting productive ability in that the former asks them to attend to word order and the latter to write down answers of their own in creating in the context.

Second, we shall consider the relationship of the Cloze Test with four Communicative Tests. The correlation between the Speaking Test and the Cloze Test was low. Noteworthy was the correlation between Cloze 2 using a Conversation text and the Speaking Test. Since the Conversation cloze was reported to measure the oral ability of the subjects (Hughes, 1981), a high correlation between them had been expected. However, the results fell short of such expectations. As Hughes (1989, p.67) says, "cloze procedures, since they produce purely pencil and paper tests, cannot tell us anything about the oral component of overall proficiency." The correlation between the Writing Test and the Cloze Test was high in the three types of evaluation (TE, HE and AE). This high correlation can be attributed to the fact that both elicit productive ability of the subjects. The Cloze Test asks the subjects to recover the deleted words by guessing from the context, whereas the Writing Test directly asks them to complete a letter or write a class diary, or summarize a story in English.

The correlation between the Reading Test and the Cloze Test was high. The reason for this seems to be that both types of the tests are similar in terms of requiring the subjects to exercise the ability to grasp the meaning of the whole passage and guess the meaning from the context.

The correlation between the Listening Test and the Cloze Test was high. Both tests have something in common, that is, checking productive ability. The Listening Test provides the subjects with three kinds of tasks such as summarizing a story in Japanese, classifying given information, and identifying things on a sketch map of a room which the speaker referred to. Both tests require productive ability of the subjects.

Concerning Research Question 3 of high scorers on the Cloze Test and the STEP, in order to know whether high scorers on the STEP will gain high scores on the Cloze Test as well, the author sought the corrected correlation of scores between the Cloze Test and STEP by using the formula for correction for attenuation. The formula for correction for attenuation is: Correlation coefficient

$$r = (\text{Cloze Total - STEP})/\sqrt{(\text{Reliability of STEP})\ (\text{Reliability of Cloze})}.$$

This produces a very high correlation coefficient of 0.927. It means that the higher the scores a subject obtains on STEP, the higher the scores he/she gains on the Cloze test. The Cloze Test thus discriminates between subjects very efficiently.

Regarding Research Question 4 of beginning EFL learners' perceptions of the Cloze Test, the face validity of the Cloze Test was not high. What the subjects think the Cloze Test measures coincides with what the statistical results show in Table 8. The response validity of the Cloze Test is high in that the subjects' perception that it measures grammatical knowledge and Reading Comprehension is supported by statistical analysis.

Limitation: After the STEP test was administered to 467 students, it was scored not in the form of 1-0 data but in the form of allotting certain points to each of 42 question items totaling 104 points. Therefore, the reliability coefficients for the

STEP for the whole 369 subjects were not calculated. Just the data from 100 subjects were calculated.

VI Conclusion

This study was conducted to find the answers to the four research questions: First, how can the Cloze Test be evaluated in terms of effectiveness, which is made up of (1) item difficulty, (2) item discrimination, (3) reliability, (4) validity, (5) ease of construction, (6) ease of scoring, and (7) ease of interpretation? The results showed that the Cloze Test is effective with regard to all seven criteria. Just as shown in Tables 2, 3, 5, 7 and 8, the Cloze Test with a Narration text can be regarded as effective since the single 26-item Cloze Test shows strong reliability and validity.

Second, how is the Cloze Test related to the Discrete-point Test and Communicative Tests? It turned out that the Cloze Test was strongly related to the STEP, a discrete-point test, ($r = 0.861^{**}$). Another finding was that of four Communicative Tests (Speaking, Writing, Reading, and Listening), what was strongly correlated to the Cloze Test was Writing, Reading, and Listening, with Reading the highest ($r = 0.817$). What was contrary to the author's expectation was that Cloze 2 with a Conversation text was weakly correlated to Speaking ($r = 0.378$). With regard to the effectiveness of a Conversation Cloze Test for Japanese beginning learners of English, some doubt is cast on the claim by Hughes (1981) about the strong correlation between the conversation cloze and oral ability.

Third, the Cloze Test seems capable of discriminating upper and lower proficiency groups efficiently.

Fourth, since approximately half of the beginning learners of English showed reservations about whether to judge the Cloze test a good test, it cannot be said that this cloze test is high in terms of face validity, although they did seem interested in this new type of test. Concerning what the Cloze Test measures, their perception matches what past research reported.

Overall, this study shows evidence in favor of the reliability ,validity and practicality of the cloze test in this particular usage context. It seems to work as a proficiency test of writing and reading abilities. It also appears to be a rough indicator of the subject's listening ability (r = 0.791) as well, although a cloze test using a conversation as a text seems unable to measure speaking ability. As an implication for testing in the classroom, the 50-or more item cloze test using a Narration text could be used as a placement test when the teacher is too busy to construct a regular long placement test.

Acknowledgements

I would like to express my gratitude to Prof. Noboru Yamada at Shizuoka Sangyo University, my co-researcher who dealt with data from high school students for this 1996-98 Education Ministry sponsored grant-in-aid research for his cooperation in constructing the Speaking, and Listening Communicative Tests for middle school students. I am very grateful to three teachers of English, Mrs. Yuriko Kuronuma at a middle school in Shizuoka City, Mr. Tomohisa Kaneda at a middle school in Hamamatsu City, and Mr. Hiroaki Sugimoto at a middle school in Ito City in Shizuoka Prefecture for their permitting me to conduct all these tests.

References

Alderson, J.C. (1980). Native and nonnative speaker performance on cloze tests. *Langua*, 313-336.

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.

Anderson,N.J. Bachman,L., Perkins,K. and Cohen,A. An exploratory study into the construct validity of a reading comprehension test: triangulation of data sources. *Language testing* 8, 41-66.

Bachman,L.F. (1982). The trait structure of cloze test scores. *TESOL Quarterly 16*, 61-70.

Bachman,L.F. (1985) Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly 19*, 535-556.

Bachman,L.F. (1990a). *Fundamental considerations in language testing*. Oxford

University Press.

Bachman,L.F. (1991). What does language testing have to offer? *TESOL Quarterly 25*, 671-704.

Bachman, L.F. and Palmer, A.S. (1996). *Language testing in practice*. Oxford University Press.

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.

Brown, H.D. (1994) *Principles of language learning and teaching*. New Jersey: Prentice Hall Regents.

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

Brown, J.D. (1983). A closer look at cloze: validity and reliability. *Language testing issues in research*. 237-250.

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

Brown, J.D. (1996). *Testing in language programs*. New Jersey: Prentice Hall Regents.

Carol A, C. and Roberta G. A. (1990). Cloze method: what difference does it make? *Language Testing 7,* 121-146.

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

Chapelle,C.A. and Abraham,R.G. (1990). Cloze method :what difference does it make? *Language Testing 7,* 121- 146.

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

Darnell,D.K. (1968). The development of an English language proficiency test of foreign students using a clozentropy procedure. U.S. DHEW Project No.7-H-010, *ERIC EDO*, 024-039. Boulder: University of Colorado,

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

Ebel, R. (1972). *Essentials of educational measurement* (2 nd ed.).New Jersey: Prentice Hall, Inc.

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

Guerrero, M.D. (2000). The unified validity of the four skills exam applying Messick's framework. *Language Testing 17*, 397-421.

Heaton, J.B. (1975). *Writing English Language Tests*. Essex, England: Longman Group Limited.

Homburg, T.J. (1984). Holistic evaluation of ESL compositions: can it be validated objectively? *TESOL Quarterly 18*, 87-107.

Hughes, A. (1981). Conversational cloze as a measure of oral ability. *ELT Journal 35*, 161-167.

Hughes, A. (1989). *Testing for language teachers*. Cambridge University Press.

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

Jonz, J. (1976). Improving on the basic egg: the M-C cloze. *LanguageLearning 26*, 255-265.

Johnson, K. (2001) *Introduction to foreign language learning and teaching*. Edinburgh Gate, U.K.: Pearson Education Limited.

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

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. (1997). C-Tests in the context of reduced redundancy testing: an appraisal. *Language Testing 14*, 47-84.

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

Manning, W.H. (1986). *Development of cloze-elide tests of English as a second* *language*. Princeton, NJ: Educational Testing Service.

Messick, S. (1989). Meaning and values in test validation: the science and ethics of assessment. *Educational Researcher, 18,* 5-11.

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

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

Mochizuki, A. (2000). Reliability and validity of communicative tests based on Weir's three-part framework. *Research Bulletin 29,* 259-266. Chubuchiku English Language Education Society.

Mochizuki, A. (2001). Three ways to evaluate results of communicative tests based on Weir's 3-part framework. *Foreign Language Education Bulletin, 23,* 1-18. University of Tsukuba.

Oller, J. (1972). Scoring methods and difficulty levels for cloze tests of proficiency in English as a second language. *Modern Language Journal,LVI,* 151-158.

Oller, J. (1973). Discrete-point tests versus tests of integrative skills. in Oller, J.W. & Richards, *Focus on the learner. Rowley,* Massachusetts: Newbury House. 184-199.

Perkins, K. (1983). On the use of composition scoring techniques, *TESOL Quarterly 17,* 651-71.

Porter, D. (1976) Modified cloze procedure, *English language teaching, 30,* 151-155.

Porter, D. (1978) Cloze procedure and equivalence. *Language Learning, 12,* 334-341.

Polio, C.G. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning 47,* 101-143.

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.

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.

Sasaki,M. (2000). Effects of cultural schemata on students' test-taking processes for cloze tests: a multiple data source approach. *Language Testing 17*, 85-114.

Savignon S.J. (1983). *Communicative Competence Theory and Classroom Practice*. Massachusetts: Addison — Wesley.

Sciarone,A.g. and Schoorl, (1989). The cloze test: or why small isn't always beautiful. *Language Learning, 39*, 415-438.

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*, 313-336.

Taylor, W.L. (1953).Cloze procedure: a new tool for measuring readability. *Journalism Quarterly, 30*, 415-433.

Weir, C.J. (1993). *Understanding and Developing Language Tests*. Hertfordshire: Prentice Hall.

Weir, C.J. (1990). *Communicative language testing*. Hempstead: Prentice Hall International (UK) Ltd.

Appendix 1 Part from the Cloze Test 1.

I It was about five in the afternoon. I was still working at the office then. An old friend called Mike telephoned me from the airport. He is a teacher of French of an American high school.

He said, "I have just arrived in Japan. (1) Can I see you right now?"

"Well, (2) I am too busy to meet you (3) now," I said. "Will you come to (4) my house and wait for me? I'll (5) be back soon." He said, "I don't (6) know how to get to your house." (7) I said,

" I think you have the map I sent you two months ago. Look at the map, and you will (10) be able to find my house easily. (11) I've left the door key under the (12) stone on the left side of the (13) door ." Then I told him to go (14) into the kitchen and find something to (15) eat or to drink. (The rest is omitted.)

Note. (translated from Japanese into English) luckily =as a result of good luck, neighbor = someone who lives in the house or apartment next to you or near you

Appendix 2 Part from the Cloze Test 2.

II. J= Jiro, T= Tom

J: Hi, Tom.

T: Good morning, Jiro. How are you?

J: Fine, thank you, and how are (27) you ?

T: I'm fine, too, thank you. Are (28) you busy this summer?

J: Yes, I am. (29) Last night we had a dinner party (30) with foreign students.

Their fathers are working (31) for my father's factory.

T: Have you known (32) them for a long time?

J: No, I (33) haven't. I saw them for the first (34) time at the party.

T: Are they from (35) the same country?

J: No, they aren't. They (36) are from different countries. As soon as (37) the __ summer ⋯. (The rest is omitted.)

Appendix 3 Part from the Speaking Test . (Conducted in an LL at a middle school by having the students talk to the microphone following the directions over the Public Speaking System and tape-record their responses on the tape in unison.)

1. Introduce yourself in English for one minute. Include your name, academic year, school, family and interests (hobbies) in your talk. Think for 2 minutes and start.

2. You are now in New York and you are going to a concert on the Broadway. Ask a clerk at a ticket office of your hotel Questions 1～ 6 in English in two minutes.

Questions you have to ask the clerk: The following are translations from Japanese.

(1) (Who will sing tonight?)

(2) (What time will the concert start?)

(3) ～ (6) (Omitted)

3. Read the following passage in two minutes. Then turn your paper upside

down and summarize the content in two minutes.

Tom liked to play baseball very much. After school he enjoyed playing baseball with his friends in the playground. He was very good at hitting the ball. ……(The rest is omitted.)

Appendix 4 Part from the Writing Test.

1. You have received the following letter from your friend. Write your reply, keeping in mind that you will provide all information that the sender wants.

St. Clair, Michigan
November 10,1997

Dear friend,

I hope everything is going well with you and your family. We would like to visit Shizuoka City for a week at the end of March next year. Can we see you then? What's the weather like in that season? Do we have to bring our warm coats? My brother Tom would like you to tell us what we could visit during our stay. What kind of food will we be able to eat? Would you let me know a good hotel for us to stay in?

Love,
Anne

2. (The teacher uses a tape-recorder and has his or her students write their answers.) Listen to a conversation between Akio and Mrs. Brown and write down the content within 80 words by paying attention to where this dialogue took place and what they were talking about. After you have written down the summary, count the number of words and put it into the parentheses at the end. The dialogue will be played twice.

3. In your class you are supposed to write the class diary by turns. December 10 is the day when you are to write in it. The following memo is what you have taken notes of so that you can write in the class diary. On the basis of these memos, write the class diary in English. You are required to write at least 6 sentences. The initial part is shown for you.

(Translations from Japanese.)

Thursday, December 10.

(In the morning, classes in social studies, Japanese, and English.

Social studies ··· studied about pollution.) (The rest is omitted.)

Appendix 5 Part from the Listening Test . (The teacher uses a tape recorder and has his or her students write their answers.)

1. Listen to the dialogues (1) ～(6). Fill in the blanks with suitable Japanese so that your replies will match the content of those dialogues.

Dialogue 1. (Script A: Where are you living now? B:In Oakland. A: Oh really? Where's that? B: It's north of San Francisco.)

Answer : (Translated from Japanese.) ( ) is located ( ) San Francisco.

(The rest is omitted.)

Appendix 6 Part from the Reading Test.

1. Read the following passage and fill in the blanks by answering the questions. (Open-ended forms.)

(The passage and answer columns are omitted.)

2. Passages (1) ～ (7) below pictures (a) ～(g) are in correct order. Put the pictures in order so that each picture will match its content of the passage. Passage 1 matches the content of Picture (a).

(The passages and pictures are omitted.)

3. Read the three-paragraph passage and summarize the content of each paragraph in Japanese. Write down in 140 or so letters in total.

(The passage is omitted.)