Shiken: JALT Testing & Evaluation SIG Newsletter, 2 (1) October 1998 (p. 2 - 10)

Do different C-tests discriminate proficiency levels of EL2 learners?

様々なC-テストは、EL2学習者のレベルを判別できるか？

Cecilia B. Ikeguchi (Tsukuba Women's University)

Abstract

A study was conducted to determine the ability of two C-tests to discriminate the English s levels of two groups of university students in Japan. A C-test based short passages from different texts and another C-test based on one long narrative passage was used. The data was analyzed by an analysis of variance, t-tests, and other statistical measures. The results were statistically significant and this paper suggests that C-tests can discriminate levels of English proficiency among EL2 students in this sample group. It also suggests the superiority of C-tests constructed from several short segments over those made from only one long passage.

Keywords: C-tests, cloze tests, ESL assessment, ESL proficiency levels, ESL testing

概念

この研究は、C-テストが英語の学習レベルを判別するために有効であるかを実施したものです。 2 種類の C-テストが二つのグループで試されました。一方の C-テストは、いくつかの短文から成っています。もう一方の C-テストは、長文で構成されています。これらの結果は、いろいろな方法によってその信頼性を確認しました。その結果、統計学的に重要なことが判明し、この調査では、C-テストが日本の学習者の英語レベルを識別することに有効であるといえ、更に、C-テストの短文のものが、長文に比べて判別に適しているといえます。

キーワード： C-テスト、クローズ・テスト、英語能力のテスト、外国語のアセスメント-、主要記事へ戻る

Since the introduction of the cloze procedure as a measure of readability by Wilson Taylor (1953), it has been employed as one way of measuring the reading ability of native speakers (Bormuth, 1967; Crawford, 1970). Other researchers later investigated the effectiveness of cloze testing as a measure of ESL/EFL proficiency (Darnell, 1968; Brown, 1983, 1988, 1993; Irvine, Atai, and Oller, 1974; Oller, 1972, 1983) to name a few. The results have indeed been widely varied across studies and a number of defects have been found with cloze procedures. In the light of these criticisms, Klein-Braley and Raatz (1984) proposed a modification known as C-testing. The procedure, developed to answer the psychometric problems of cloze testing, has been purported as an empirically and theoretically valid measure of language proficiency (Raatz and Klein-Braley, 1981; Klein-Braley, 1985; Klein-Braley and Raatz, 1984, 1985; Raatz, 1985). This was later proposed by other researchers as a substitute for cloze tests (Mc Beath, 1990; Cohen, Segall, and Weiss, 1984).

Originally, the C-testing procedure involved making a test from four or

> *"[Mochizuki (1994) contends that] long passages, especially narratives, ... [are] the most appropriate for making the C-test effective in terms of reliability and concurrent validity."*"

five thematically distinct segments of a connected discourse in which the second half of every second word (usually 100 words in all) were deleted. Examinees got credit for exact word restorations. The use of several different short texts minimized the effect of text topic familiarity or difficulty. Nevertheless, researches did not explore what kind of text produces higher reliability or validity until Mochizuki (1994) experimented with four kinds of texts for classroom C-tests: narratives, explanations, arguments, and descriptions. His study - which is later counter-indicated in this paper - suggested that long passages, especially narratives, were the most appropriate for making the C-test effective in terms of reliability and concurrent validity.

Klein-Braley and Raatz basically utilized teacher judgments or school grades as a criterion for validating C-tests, while other researchers have supplied evidence grounded on other kinds of criteria. For example, Nigishi (1987) reports correlation coefficients of .80 and .76 between C-tests and the reading subtest of ELBA and total ELBA, respectively; while the studies of Ikeguchi (1994) indicate the C-test responses to correlate highest with the grammar results of TOFEL exams. Still other studies in support of this test procedure include the validation of C tests among ESL/EFL learners. For instance, Feldman and Stemmer (1987) found C test validation through verbal reports, while Doornyei and Katona (1992) studied C tests against different language tests, including oral interviews. They found further support for C-testing, reporting that this procedure gives a random and representative sample of an original text. That supports an earlier assertion that the every-other-word deletion in the C test produces a large number of 'random samples of the word classes of the text involved' (Klein-Braley, 1985, 1984). Other recent SLA researchers suggested that C-tests could also be useful for L2 vocabulary research. For instance, Singleton and Little (1991) found the responses of L2 learners to C-tests as a source of evidence about second language lexical development.

On the other hand, criticisms have been leveled against the C-test procedure. Some common criticisms meriting further investigation are:

1. Do C-tests accurately reflect students' ability to process discourse for general proficiency (Cleary, 1988)?
2. Do C-tests encourage only microlevel processing rather than macrolevel processing (Cohen, Segal, and Weiss, 1984)?
3. Do C-tests have adequate face validity (Weir, 1988)? and
4. Since C-tests have reduced redundancy, are they valid tests of language competence (Carroll, 1987)?

Although C-tests may tap into a measure of grammatical competence (Klein-Braley,1985), there is not enough validity research regarding the specific traits they measure (Chapelle and Abraham, 1990). Moreover, according to Jafarpur (1995) 'assumptions of random sampling of the basic elements of a text are doubtful'.

The use of C-tests since their introduction (Klein-Braley and Raatz, 1984) as a means of constructing norm-referenced measures for proficiency and placement testing, and to solve problems concerning the cloze procedures, has been extended to certain indefinite limits such as 'measure of language creativity' (Carroll, 1987), and has yielded results contrary to the researchers' expectations that were not the purpose for which this test was originally intended. Furthermore, the empirical evidence in support of C-tests is scanty (Weir, 1988) and warrants further investigation in the context of second language instruction.

Methodology

Purpose of the study

The objectives of this study are to investigate whether C-tests, using two procedures of construction, can discriminate levels of language proficiency between ESL learners in Japan and to determine the superiority of a C-test using several passages (C-test 1) over a C-test constructed from only one long passage, a narrative type (C-test 2), in terms of reliability and correlation with an external criterion.

Subjects

Two groups of first year university students in Japan were chosen for the investigation: one group consisted of 60 undergraduate students enrolled in a general English course, the second group was made up of 30 students in a class of English for returnees. Students from the first group were picked randomly from an intact class, while those from the latter group belonged to one English class for returnees. To qualify for that class, the students must have stayed in an English speaking country for at least a period of one year, and have passed the qualifying exam administered by the university. In terms of proficiency level, most of the students in that class had advanced listening and oral production skills, but post-intermediate writing and reading skills (Tschirner, 1996).

Materials

Two kinds of C-tests were used in this study: one type was constructed using four short passages from different texts, while the other was constructed using only one long narrative text. The use of several short segments of different texts has been shown by the researches above, to have a satisfactory reliability above .80. According to Klein-Braley and Raatz (1984) it is also empirically valid. For this study, the four short passages were chosen from different texts within similar readability and interest levels using the Fry (1985) and Flesch (as described in Klare, 1984) indices. The readability estimates of the texts where segments were chosen for this study had a 6 - 8 level by the Fry index, and a 6.7 - to 9.6 level by the Flesch index. These numbers which appear to be quite different scales are remarkable only in that they indicate variations in the readability levels of the passages used (Brown, 1993). C-test 1 was constructed using 25 items from different passages, making a total of 100 items. Every first and last sentence of each passage were left intact to provide contextual clues.

The second type of C-test was adopted for use in this study based on Mochizuki's (1994) research on different types of discourse: the description, the exposition, the narration and argumentation to construct C-tests for classroom use. Among these four types of texts, the narrative type was found to be the most reliable - .92 . The narrative text "The Lock Keeper" consisting of 120 items which was found to be the most reliable and with the highest concurrent validity (Mochizuki, 1995).

This study is an attempt to investigate which of these two types of C-test constructions would yield higher reliability and concurrent validity. The external criterion used was the STEP-Eiken exam. The STEP-Eiken exam consists of 66 written test questions on vocabulary, grammar and reading comprehension. The STEP-Eiken has been established in previous investigations as resulting in high reliability as well as high coefficients as an external validating criterion with Japanese university students (Kimura, 1995). In a previous study using the STEP-Eiken and CELT results to investigate the external validity of C-tests constructed from different types of discourse, STEP-Eiken was found to have a higher reliability (.778) than the CELT (.638), and other C-tests (Mochizuki, 1994).

Procedures

Each student from the two groups of students took the two versions of the C-tests and the STEP-Eiken. To control for a potential order effect, the order of administering the C-test and the STEP-Eiken was counterbalanced: half the subjects in the non-returnees group and the returnees group took the two C-tests first, and the STEP-Eiken during the English class the

following week. The other half of each group took the STEP-Eiken test first, and then the English test.

Analyses

The students' responses for both C-test 1 and C-test 2 were scored for exact replacements. Descriptive statistics for the scores of the C-tests were obtained. Reliability coefficients were obtained by the KR-20 method. The use of KR-20 has been questioned in the past. For instance, Faraday (1983) and Bachman (1990) claim that the internal consistency reliability coefficients are inappropriate for cloze and C-tests because of the interdependence of items. On the other hand, Woods (1984), Henning (1987) and (Jafarpur, 1995) claimed that the KR-20 method yields the same results as Cranach's alpha. Moreover, Brown (1983) provided evidence that the differences between reliability coefficients from KR-20 and Cronbach's are negligible.

To address the first research question, that of determining the discriminative power of the C-tests, a comparison of the subjects' scores among groups was obtained, based on the results of the group t-tests. The subjects' mean scores within each group for each test was obtained and subjected to an analysis of variance test and t-tests.

For the second research objective which is to determine the reliability and correlation of C-tests and STEP-Eiken, the Pearson product moment correlation coefficients were computed.

Results and discussion

Table 1. *Basic descriptive statistics for non-returnees' scores on the C-test 1, C-test 2 and the STEP-Eiken tests.*

Test type	N	No. of Items	Mean	Reliability *	
C-test 1	60	100	61	.67	.73
C-test 2	60	120	98	.70	.83
STEP-Eiken	60	160	109	.75	.85

* Raw score reliabilities (KR 20) appear on the right and reliabilities that would be observed if all the tests contained 100 items appear on the left.

Table 2. *Basic descriptive statistics for returnees' scores on the C-test 1, C-test 2 and the STEP-Eiken tests.*

Test type	N	No. of items	Mean	Reliability *	
C-test 1	30	100	74	.65	.76
C-test 2	30	120	109	.71	.89
STEP	30	160	124	.87	.91

* Raw score reliabilities (K-R 20) appear on the left and reliabilities that would be observed if all the tests contained 100 items on the right.

An analysis of means for all tests for the non-returnees indicates the highest means obtained by the STEP-Eiken, and the lowest mean scores by the C-test 2, indicating the former to be the easiest, and the Narration C-test to be the most difficult. An ANOVA was conducted to find the statistical significance in these scores, and the obtained results were: F = 176.18 (2, 179), p <.00.

A similar analysis was conducted on the mean scores obtained by the returnees for the two types of C-tests and the STEP-Eiken. The results indicated the highest mean scores for the latter and the lowest means for the first type of C-test. This shows a similar pattern as that observed for the non-returnees. These score differences were checked by an ANOVA and the results were found to be highly significant : F = 56.94 (2,75), p <.00 as summarized in Table 3.

Table 3. *Results of an ANOVA analysis for the scores of all subjects on all tests.*

Group	Source of variance	SS	*df*	*MS*	*F*
Non-returnees	Between groups	3581.4	2	1.790	233.3
	Within group	6677.13	57	76.75	
	Total	10258.53			
Returnees	Between groups	3416.5	2	17082.3	88.303
	Within group	1458.8	27	193.5	
	Total	4875.3			

p = <.001

A cursory glance at these tables shows that the returnees group obtained a consistently higher set of mean scores for both the C-test using different short segments from different texts and the C-test using only one narrative passage. These differences show that the C-test types were much easier for the returnees than for the other group. To further determine the extent to which C-tests of different types can discriminate English proficiency levels among the students, *t*-tests were conducted between the scores of each group for each.

The results of t-test analyses indicate that C-test 2 using different short texts was easier for the returnees than for the other group at a significant level: t=.86 df= 29, p=.00. In the same manner, the narrative C-test proved to be much easier for the returnees than for the non-returnees, and the difference level was found to be highly significant: t = 3.21 df=59, p =.005. The returnees outperformed the non-returnees in both C-test 1 and C-test 2. These results indicate the two C-test types used in this study can discriminated levels of English proficiency of the Japanese university students who took part in this study.

In addition, there is also the question of which of these two C-test types is superior to the other in terms of reliability, and in terms of concurrent validity. To permit comparison among the reliability estimates of the different tests used in this study, 'corrected' reliabilities', the reliabilities that would be observed if all the test types had contained 100 items, were applied to

all the cloze tests and STEP-Eiken test items (Gordon, 1989 and Chapelle, 1990). Higher reliability results were observed for the C-test using several segments than the narrative type for both sample groups.

Criterion related validity

To determine how well C-tests relate to an outside criterion, both C-test scores for both groups were correlated with their scores STEP-Eiken scores. Moreover, since these correlations are based on tests with different number of items, correlations were adjusted corrected for attenuation (Jafarpur, 1995) as shown in Table 4.

Table 4. *Correlations among the C-test types and STEP-Eiken scores*.

Group	C-test1 (different texts) and STEP	C-test 2 (narrative) and STEP
Returnees	.58	.29
Non-returnees	.51	.26

The table shows only moderate correlations, of at least .50, (Klein-Braley, 1984) between C-test 1 and STEP-Eiken test scores. The C-test 2 did not correlate much with the STEP-Eiken scores. The C-test that was based on short texts was superior to the one based on a long narrative, counter-indicating Mochizuki's (1994) claim that single narratives make the best C-tests.

"The C-test that was based on short texts was superior to the one based on a long narrative, counter-indicating Mochizuki's (1994) claim that single narratives make the best C-tests."

More importantly, the moderate correlations between the C-test from various texts against a single criterion suggests that it is possible for C-tests to tap different language abilities of ESL learners (Jafarpur, 1995). Finally, texts carefully chosen according to their similarities in terms of interest and readability level lead to the superiority of a C-test constructed using several short passages over a C-test using only one text.

Summary and conclusion

Three points can be made from this research: (1) The C-test procedure does discriminate moderately between the levels of English proficiency for the Japanese university students in this sample. (2) The C-test using several short segments from different texts appears to be superior to the one using only one long narrative text. (3) The two C-tests differ in terms of their criterion-related validity.

The writer acknowledges the fact that the number of samples and tests included in the study was small. It appears quite possible that random variation alone could account for the variability in the results of statistical analysis. Notwithstanding, the results of this investigation suggest that C-tests have the ability to differentiate ESL levels the Japanese university students in this sample. Furthermore, the C-test constructed from different passages has been shown to have more validity against a reference criterion than a narrative type C-test. Because of the far-reaching potential of C-tests in the field of empirical research and classroom testing, further research on their application and effectiveness is warranted.

References

Bormuth, J. R. (1967). Comparable cloze and multiple-choice comprehension test scores. *Journal of Reading 10,* 291-299.

Brown, J. D. (1983). A closer look at cloze: validity and reliability. In Oller, J. W. Jr. (Ed.) *Issues in Language Testing*. Rowley, MA: Newbury House, 237-250.

Brown, J. D. (1988). Tailored cloze: improved with classical item analysis and techniques. *Language Testing, 5* (1) 19-31.

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

Carroll, J.B. (1987). Review of Klein-Braley and Raatz. C-tests in der praxis. *Language Testing, 4,* 99-106.

Chapelle, A. and Abraham, R. (1990). Cloze Method: what difference does it make? *Language Testing, 7* (2) 121-146.

Chapelle, C. (1994). Are C-tests valid measures for L2 vocabulary research? *Second Language Research, 10* (2) 157 -187.

Cohen, A.D., Segal, M, and Weiss, R. (1984). The C-tests in Hebrew. *Language Testing, 1* (2) 221- 225.

Darnell, D.K. (1970). Clozentropy: A procedure for testing English language proficiency of foreign students. *Speech monographs*. 37, 36-46.

Dornjei, Z. and Katona, L. (1992). Validation of C-tests among Hungarian EFL learners. *Language Testing*. *2*, 187-206.

Harris, D. & Palmer, L. (n.d.) *A Comprehensive English language test for learners of English (CELT)*. New York: Mc Graw Hill.

Henning, J. (1987). *A guide to language testing: Development, evaluation, measurement*. Cambridge, MA: Newbury House.

Ikeguchi, C. (Unpublished ms.) The four cloze types: To each its own. Tsukuba Women's, University, Japan

Jafarpur, A. (1995). Is C testing superior to Cloze? *Language Testing, 12* (2) 194-215.

Jonz, J. (1990). Another turn in the conversation: what does the cloze measure? *TESOL Quarterly, 24* (1) 61-63.

Kimura, K. & Visgatis, B. (1996). High school English textbooks and college entrance examinations: A comparison of reading passage difficulty. *JALT Journal, 18* (1) 81-95.

Kimura, Y. (1995). Investigating the English competence of students returned from overseas. in K. Kitao, et al. *Culture and Communication*. Kyoto: Yamaguchi Shoten.

Klare, G.R. (1984). Readability. In P. D. Pearson (Ed.), *Handbook of Reading Research (*pp. 681-738). New York: Longman.

Klein-Braley, C. (1985). A close-up on the C test: A study in the construct validation of authentic tests. *Language Testing, 2* (1) 76-104.

Klein-Braley, C. & Raatz, E. (1984). A survey on the C test-1. *Language Testing, 1* (2) 134-146.

McBeath, N. (1990). C-tests: Some words of caution. *English Teaching Forum, 28*, 45-46.

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

Negishi, M. (1987). The C-test: An integrative measure? *IRLT Bulletin 1,* 3-26.

Oller, J. W. Jr. (1972). Scoring methods and difficulty levels for cloze tests of proficiency in English as a second language. *Modern Language Journal 56,* 151-158.

Oller, J. W. Jr. (1983). *Issues in Language Testing*. Rowley, MA: Newbury House.

Raatz, U. (1985). Better theory for better tests? *Language Testing, 2* (1) 60-75.

Raatz, U. and 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. Paper 26. Colchester: University of Essex.

Taylor, W.L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly, 30,* 414-438.

Tschirner. E. (1996). Rethinking beginning FL instruction. *Modern Language Journal*. *80*, 1-13.

HTML: http://www.jalt.org/test/ike_i.htm **PDF**: http://www.jalt.org/test/PDF/Ikeguchi1.pdf