TELL, Vol. 1, No. 4, 2007

Validity and Discriminatory Power of the C-Test as a Measure of General Language Proficiency

Mahmood Rouhani*
M.A. in TEFL, English Teacher

Abstract

This study reports on an investigation of reliability, different aspects of validity, and discrimination power of the C-Test as a measure of overall language ability. A C-Test developed by the researcher and a Michigan Test of English Language Proficiency (MTELP) were administered concurrently to 144 university students. The reliabilty coefficients found for the C-Test were high. The C-Test also proved to have acceptable content relevance and fairly high criterion-related validity. Results of two factor analyses confirmed that the C-Test texts measure, to a large extent, the same underlying trait as the MTELP –significant evidence of construct validity for the C-Test. However, the C-Test texts did not prove to behave consistently with examinees of different proficiency levels. Also it came out that the C-Test could not consistently classify the subjects in their appropriate proficiency levels. This finding was further affirmed by an ANOVA whose results demonstrated that the C-Test had difficulty discriminating between participants of lower and upper intermediate levels.

* My thanks are due to Dr. Mansour Koosha for his precious reviewing and constructive comments.

Keywords: language proficiency, reliability, disriminatory power, construct validity, criterion-related validity, content validity, cloze test, C-Test.

1. Introduction

Cloze test is now a well-known and widely-used integrative language test.Taylor (1953) first introduced the cloze procedure as a device for estimating the *readability* of a text. However, what brought the cloze procedure widespread popularity was the investigations with the cloze test as a measure of *ESL proficiency* (Jonz, 1976, 1990; Hinofotis, 1980; Bachman, 1982, 1985; Brown, 1983, 1993; Laesch & van Kleek, 1987; Chapelle & Abraham 1990; see also Oller, 1979 for an overview). The results of the substantial volume of research on cloze test have been extremely varied. Furthermore, major technical defects have been found with the procedure. Alderson (1979, 1980, 1983), for instanse, showed that changes in the starting point or deletion rate affect reliability and validity coefficients. Other researchers like Carroll (1980), Klein-Braley and Raatz (1984), Klein-Braley (1983, 1985), Farhady (1983b), and Brown (1993) have questioned the reliability and different aspects of validity of cloze tests. In view of all the criticisms made against the cloze procedure, Klein-Braley and Raatz proposed the C-Test as a modified form of the cloze test.

The C-Test consists of four or five short texts in each of which the first sentence is left standing, then the ***C-principle*** (or ***the rule of two***) is applied: the second half of every second word is deleted, beginning with the second word of the second sentence. If a word has an odd number of letters, the 'larger' half is omitted. Numbers, proper names, abbreviations, and one-letter words such as 'I' are ignored in the counting. In the canonical C-Test each text will have either 20 or 25 blanks. The students' task is to restore the missing parts. Only entirely correct restorations are counted as correct (i.e., spelling problems are considered errors). The testees would have roughly five minutes to answer each text, so that a test with five parts would take twenty five minutes to complete.

The C-Test is believed to have a number of advantages over the cloze test (Klein-Braley & Raatz, 1984; Klein-Braley, 1997). Some of the most important rewards of the C-Test are as follows:

> The use of a variety of passages allows for a better sampling and representation of the language and content. Also, a person with special knowledge in a certain field cannot have an unfair advantage all through the test.
> Since every second word is damaged, it is possible to obtain a better sampling of all the different language elements in a text.
> C-Tests are very easy for native speakers. But someone who doesn't know the language at all normally scores zero or close to zero.
> C-Tests are easy to construct, administer, and score.
> As there is only one acceptable solution in most cases, the scoring is more objective.

Ever since it was introduced, the C-Test has been the subject of many research studies and scholarly controversies. On one hand, some researchers have found the C-Test a highly integrative, reliable and valid measure of overall language ability (Klein-Braley & Raatz, 1984; Cohen, Segal & Weiss, 1984; Klein-Braley, 1985, 1997; Chapelle & Abraham, 1990; Dörnyei & Katona, 1992; Huhta, 1996; Connelly, 1997; Ikeguchi, 1998; Babaii & Ansary, 2001; Eckes & Grotjahn, 2006; see Sigott, 2004 for an extensive review). On the other hand, researchers like McBeath (1989, 1990), Hughes (2003), Weir (1990), and Jafarpur (1995, 1999a, 1999b, 2002) have doubted some of the claims made on the part of the C-Test. These researchers seriously questioned the face validity of the C-Test, its content coverage and relevance, and variability of test results as a function of deletion start and deletion ratio.

In the light of the variability and inconsistency of the results obtained with the C-Test, it seemed to the researcher that replicative investigations of the qualities of this testing device are in order before definitive decisions can be made as to its credibility for the assessment of overall language ability. Therefore, the current study

set out to empirically explore aspects of validity and discriminatory power of the C-Test among Iranian EFL learners.

2. Method

2.1. Instrumentation

a. The C-Test: To construct the C-Test, thirteen texts were chosen from various EFL/ESL materials. The excerpts were authentic and self-contained and they varied in subject matter. The texts were of different levels of difficulty as judged by the Flesch Reading Ease readability scale (Microsoft Word, 1983–99) and a group of eight university EFL instructors. Every first sentence of each passage was left intact to provide a complete context. Beginning from word two of sentence two, the second half of every other word was deleted. In each mutilation, exactly half of the word was omitted, but if the number of letters was uneven, one extra letter was left out. Numbers, proper names and one-letter words were ignored in the counting and thus were not mutilated either. In this way, thirteen mutilated texts were produced with each one containing 20 gaps.

To facilitate pretesting, the extracts were randomly divided into two C-Tests which were, then, randomly given to 49 Iranian foreign language learners of English, 6 Iranian EFL teachers, and 3 native speakers. The completed test papers were scored giving one point for each exact restoration. The scores were item analyzed and five texts with superior discriminability and facility values were chosen. These texts were about culture, education, listening, bees, and underwater discoveries. They varied in difficulty with Flesch Reading Ease values of 62, 40, 75, 82, and 64, respectively. Dörnyei and Katona (1992) recommend the use of extracts with various difficulties in order to obtain equal measuement accuracy in both tails of a sample distribution.

The C-Test thus prepared comprised 100 gaps, fulfilling the recommended minimum number of mutilations (Klein-Braley, 1997; Raatz & Klein-Braley, 1995). The instructions were given in Persian along with a short English C-Test example and its restored answer. The final version of the C-Test can be found in Appendix I.

b. The criterion measure: The Form Q of the Michigan Test of English Language Proficiency (MTELP) (Corrigan, Dobson,

Kellman, Spaan, & Tyma, 1979) was used as the criterion for determining concurrent validity coefficients. This test is a retired component of the Michigan English Language Assessment Battery (MELAB) which is a discrete point language proficiency measure. The MTELP lasts 75 minutes to administer and comprises three subtests: 'Grammar', 'Vocabulary', and 'Reading comprehension'. The subtests contain 40, 40, and 20 four-choice items, respectively. The total score is the sum of the subtest scores. The manual reports reliability estimates of over .90 for the test and its subtests.

2.2. Participants

A total of 144 university students participated in this study. From these, 101 subjects took both the C-Test and the MTELP. They include: (a) 14 freshman, 22 sophomore, 23 junior, and 31 senior English majors studying at Khurasgan Azad University and the University of Isfahan, and (b) 11 engineering majors enrolled at an ESP course at Isfahan University of Technology. The other 43 subjects were all MA students of TEFL. They include 23 students at Najafabad Azad University, 14 students at Khurasgan Azad University, and 6 students at the University of Isfahan. These examinees took the C-Test only. The participants (mostly in their twenties) were of both sexes and enjoyed different levels of proficiency.

2.3. Test administration and scoring

In neither of the two tests had the participants been informed beforehand; so there was no preparation of any kind for the exam. The MTELP was first administered to the testees within the time limit of 75 minutes. The subjects were told that they would be informed of their grades, that their high scores on the test would affect their final term grades, and that high-ranking students would receive a prize. They were all informed that marks would be taken away for their wrong answers. The answer sheets were scored by the researcher. The MTELP scores were corrected for guessing in order to reduce the effect of chance (cf. Harris, 1969; Jafarpur, 1997). However, to remove the effect of practice, the subjects were not told that they were going to be tested again.

The C-Test was administered to the same subjects. However, since the subjects studied at different universities, the administration date varied from 10 to 14 days to cope with some limitations. It was assumed that the examinees' level of language proficiency had not changed significantly over the period. The completed C-Test papers were scored using the more convenient exact word scoring and counting spelling mistakes as incorrect. Alternative scoring procedures (acceptable word scoring, and tolerating spelling mistakes) have been shown to produce practically the same results as the one adopted in this study (Dörnyei & Katona, 1992; Huhta, 1996).

3. Results and discussions

The scores of the participants on all the tests and subtests were processed using the Statistical Package for the Social Sciences, Release 9.0.0 (SPSS, 1989-99). Table 1 shows descriptive statistics obtained from the C-Test, the MTELP, and their respective subparts, along with item facility (IF) and item discrimination (ID) indices of each C-Test text (C-Text, hereafter). In computing item facility and item discrimination indices each C-Text was considered a 'super-item' (see below). A sample seperation procedure was adopted for computing item discrimination indices (Henning, 1987; Farhady, Jafarpur, & Birjandi, 1994).

Table 1: Descriptive statistics for the scores of the subjects on all measures.

Test	No. of Items	Mean	SD	Min.	Max.	IF	ID
(N = 144)							
C-test:	100	54.69	14.70	16	93		
C-Text 1	20	14.25	3.09	5	20	.70	.29
C-Text 2	20	11.79	3.49	3	20	.58	.30
C-Text 3	20	12.65	3.52	3	20	.62	.32
C-Text 4	20	9.45	4.36	0	20	.49	.47
C-Text 5	20	6.60	4.35	0	18	.35	.42
(N = 101)							
Michigan:	100	28.08	14.96	-3.33	70.33		
Grammar	40	15.94	8.84	-4.33	36		
Vocabulary	40	7.39	5.87	-.33	33.33		
Reading	20	4.75	4.18	-2.33	18.66		

The item discrimination values are in the range of .29 to .47 with mean value of .36 for the whole test. Jafarpur (1997, 2002) believes that item discrimination indices higher than .20 are acceptable. On this basis, texts in our C-Test demonstrate fairly low, yet acceptable item discriminability indices. The most attractive item facility and the highest item discrimination goes to C-Text 4 with an IF value of .49 and an ID value of .47.

The item facility indices for the five texts of the C-Test are in the acceptable range of .35 to .70 (cf. Raatz & Klein-Braley, 1995). The mean item facility for the whole C-Test is thus .55, which is very desirable (Henning, 1987; Farhady et al., 1994). As far as item facility is concerned, except for C-Text 3, the other texts are arranged in an ascending order of difficulty.

As another index of relative difficulty, mean scores of the participants on each extract show the same pattern. They drop from 14.25 on C-Text 1, to 11.79 on C-Text 2, and after a slight increase to 12.65 on C-Text 3 continue a descending route to 9.45 on C-Text 4 and then 6.60 on C-Text 5.

3.1. Reliability

In order to allow better comparison, reliability coefficients for all the tests and subtests were estimated by the Kuder-Richardson Formula 21 (KR-21). The reliability estimate for the C-Test was also computed by the Cronbach's alpha formula. Both these formulas are measures of internal consistency.

Raatz and Klein-Braley (1995) suggest that it is possible to perform an inner consistency analysis on C-Tests. They agree that it is not permissible to define the individual blanks in the C-Test as items, since they are dependent on each other as a result of text structure and content. But they propose a practical solution: to consider each C-Test text as a super-item and then enter these four or five super-items into the Cronbach's alpha formula to estimate the reliability. Raatz (1985, p. 64) states:

> Assuming that all the parts are independent of each other, but are equivalent and measure the same thing, then the total test score is the sum of the part scores. These parts can be viewed as superitems. In this case one can calculate intercorrelations

and discrimination indices for the superitems without going inside the test parts. The reliability of the whole test can be calculated using Cronbach's alpha.

Table 2: Reliability indices for all tests

Test	KR-21	Cronbach's Alpha
C-test:	.90	.85
C-Text 1	.65	
C-Text 2	.63	
C-Text 3	.68	
C-Text 4	.78	
C-Text 5	.79	
Michigan:	.92	
Grammar	.90	
Vocabulary	.85	
Reading	.83	

Table 2 shows reliability coefficients of the two tests and their subparts computed by KR-21 formula. It also shows the reliability estimate for the C-Test computed by the Cronbach's Alpha formula. In doing so, each C-Test text was regarded as a 'super-item' and accordingly the alpha coefficient was calculated with five items.
Scores from both the C-Test and the MTELP show very high KR-21 reliability coefficients (.90 and .92, respectively). The reliability of the C-Test as estimated by the Cronbach's Alpha formula is also reasonably high (.85). The reliability coefficients of the scores obtained from the components of the MTELP are quite acceptably high too, the coefficients for each being over .83. However, only two subparts of the C-Test show satisfactory reliability indices, namely C-Text 4 (.78) and C-Text 5 (.79). The other three C-Texts demonstrate only moderately acceptable reliability with coefficients of .65 for C-Text 1, .63 for C-Text 2, and .68 for C-Text 3.

The fact that the whole C-Test is almost as reliable as the criterion (MTELP) appears to support claims concerning the high reliability of the C-Test (e.g. Klein-Braley & Raatz, 1984; Klein-Braley, 1985, 1997; Dörnyei & Katona, 1992; Connelly, 1997, to name a few).

3.2. Validity

The primary concern for any test is that the interpretations and the uses we make from the test scores are valid. The evidence that we collect in support of the validity of a particular test can be of three general types: content relevance, criterion relatedness, and meaningfulness of construct (Bachman, 1990). These categories have been separately discussed below with regard to the data presented in this study and the interpretations that can be legitimately made on their basis.

3.2.1. Content validity

A necessary stage in test validation is to investigate whether the test is relevant to a given area of content or ability. In the case of language tests, one principal concern of content validity is with the extent to which a test measures a representative sample of the language in question (Weir, 1990).

Table 3 represents the number and percentage of content and function words in the whole C-Test and each of its texts. In addition, it shows the number, percentage, and type of words mutilated in the same texts. In this analysis, auxiliary verbs, prepositions, conjunctions, pronouns, determiners, numbers, and adverbs (other than manner adverbs) have been counted as function words. The other words in the texts belong to categories of nouns, verbs, adjectives, and adverbs of manner, which are typically considered content words.

The truncated words in each C-Text represent different parts of speech. In C-Text 1, as an example, four prepositions, three adverbs, two determiners, one pronoun, one auxiliary verb, and one numeric expression are mutilated. As for content words, there are five nouns, two verbs, and one adjective mutilated.

Table 3: Number and percentage of content and function words(mutilated) in each C-Text and in the whole C-Test

	Total (343 words)				Mutilated (100 words)			
	Content		Function		Content		Function	
	Freq.	%	Freq.	%	Freq.	%	Freq.	%
C-Test	160	47	183	53	46	29	54	30
C-Text 1	27	42	37	58	8	30	12	32
C-Text 2	39	53	35	47	12	31	8	23
C-Text 3	25	42	35	58	7	28	13	37
C-Text 4	29	37	50	63	10	34	10	20
C-Text 5	40	61	26	39	9	23	11	42

As is evident from the table, the percentage of content words mutilated in the whole C-Test (29%) is almost equal to the percentage of the function words mutilated (30%). Hence, the truncated words in the C-Test conform to the demands of content validity as they represent 'a slice of reality' (Raatz 1985, p. 63). Although, this finding does not accord with Jafarpur's (1995) results, it compares very favorably with those of Dörnyei and Katona (1992) and Klein-Braley (1985) for it reveals that the C-principle is capable of obtaining a reasonably representative sample of all the word classes in a text.

3.2.2. Criterion validity

Exploring the validity of a test by means of external criteria is seen as essential by many scholars (Weir, 1990; Bachman, 1990). Criterion-related evidence demonstrates a relationship between test scores and some criterion which is believed to be also an indicator of the ability tested. Concurrent validity is a kind of criterion-related validity which is obtained through concurrent administration of a newly developed test with another well-known standardized test of which the validity is already established (Hatch & Farhady, 1982; Brown, 1988).

Table 4 provides product-moment correlations among the scores from the C-Test and the MTELP. The table delineates that total C-Test scores correlate comparatively highly with total scores from the criterion (.72). The correlation coefficients between the C-Test and each of its C-Texts are quite high (.71, .73, .80, .85, and .77, respectively). There is also considerable correlation between the MTELP and the five C-Texts (.54, .63, .63, .63, and .45, respectively).

Table 4: Correlation coefficients among the scores of the two measure

Test	MTELP	C-Test
	(N = 101)	(N = 144)
C-Test:	.72	
C-Text 1	.54	.71
C-Text 2	.63	.73
C-Text 3	.63	.80
C-Text 4	.63	.85
C-Text 5	.45	.77
	(N = 101)	(N = 101)
MTELP:		.72
Grammar	.88	.70
Vocabulary	.81	.47
Reading	.59	.46

All correlations are significant at $p<.01$ level (2-tailed).

The C-Test shows a reasonably high correlation with the grammar subtest (.70). However, its correlations with the vocabulary and reading subtests are not very much promising (.47, and .46, respectively). These coefficients seem to contradict Dörnyei and Katona (1992) who found that the C-Test is less efficient in testing grammar. By contrast, these results are comparable with Chapelle and Abraham (1990) who concluded that the C-Test is more of a grammatically based test. Also Babaii and Ansary's (2001) finding that their subjects mostly utilized their grammatical judgments to reconstruct the text is supported here.

Notice that the correlation of the C-Test with the MTELP was only moderately high (.72). A reasonable hypothesis is that the low

face validity associated with the C-Test (Hughes, 2003; Weir, 1990; Jafarpur, 1995) could most probably have affected the subjects' performance. If a test does not appear to the testees as face valid, then their adverse reaction to it results in a performance which is not a true reflection of their abilities. Weir (1990, p. 26) quotes Anastasi (1982, p. 136) who has argued:

> Certainly if test content appears irrelevant, inappropriate, silly or childish, the result will be poor co-operation, regardless of the actual validity of the test. Especially in adult testing, it is not sufficient for a test to be objectively valid. It also needs face validity to function effectively in practical situations.

3.2.3. Construct validity

The main concern of language test makers is whether test performance truly reflects language abilities. Construct validation helps to substantiate the extent to which a testee's performance on a particular test can be indicative of his/her underlying competence.

Construct validity, as characterized by Bachman (1990, p. 254), refers to 'the extent to which performance on tests is consistent with predictions that we make on the basis of a theory of abilities, or constructs'. In investigations of construct validity, therefore, we are concerned with empirically testing hypotheses about the relationships between test scores and underlying traits. Below there are reports on several analytical procedures conducted on the data obtained in this study to examine the construct validity of the C-Test.

3.2.3.1. C-Test and staged development of L2 competence

One theory in second language learning holds that there is an orderly progress in L2 learning and learners go through a number of developmental stages, "from very primitive and deviant versions of the L2, to progressively more elaborate and target-like versions" (Mitchell & Myles, 1998, p. 10). In an attempt to establish the construct validity of the C-Test, Klein-Braley (1985) provides evidence that C-Tests support the theory of a regular progression in language learning. That is, since language competence increases

progressively, if "the same C-Test is administered to the subjects at different stages of language development, then the C-Test scores will become successively higher as the subjects become more proficient in the language" (Klein-Braley, 1985, p. 84).

To investigate the plausibility of this claim a special kind of subject grouping was required. Therefore, the undergraduate subjects were first classified into four proficiency groups based on the distance of their MTELP scores from the mean of the whole sample (the MA students were not included for they had not taken the MTELP). The subjects whose scores were lower than -2/3 SD below the mean were operationally classified as elementary level. Similarly, the lower intermediate level comprised examinees with scores between the mean and -2/3 SD. Those whose scores were between the mean and +2/3 SD were placed in the upper intermediate level. And finally, the advanced level contained examinees with scores more that 2/3 SD higher than the mean.

Table 5: Raw means and standard deviations for four proficiency groups

	Elementary (N = 26)		Lower Intermediate (N = 28)		Upper Intermediate (N = 23)		Advanced (N = 24)	
Test	Mean	SD	Mean	SD	Mean	SD	Mean	SD
C-test:	36.38	10.59	51.79	7.78	56.39	8.84	65.21	15.44
C-Text1	11.08	3.46	13.75	3.92	14.35	2.60	16.04	3.16
C-Text 2	7.54	2.49	11.25	2.82	12.74	2.70	13.00	3.18
C-Text 3	8.88	3.22	12.11	2.48	13.83	2.55	14.83	3.41
C-Text 4	5.35	2.86	8.43	2.53	9.74	2.83	12.88	5.30
C-Text 5	3.38	2.84	6.25	2.82	6.22	4.28	8.50	4.86
MTELP:	10.40	4.51	23.41	2.91	32.36	3.03	48.58	9.81

Table 5 presents means and standard deviations for the scores of the undergraduate subjects. As it is observed, the mean scores of both the criterion and experimental measures for the four groups increase progressively. Specifically, the mean scores on the C-Test become increasingly higher from a mean of 36.38 to 51.79 to 56.39

to 65.21, respectively. The mean scores on each C-Text behave in the same fashion, i.e., they become successively higher as the level of proficiency increases. Though these means speak of validity for the C-Test, they should be subjected to further scrutiny to ensure their credibility. One way to do this is to examine the differences among the means of the four proficiency groups through an analysis of variance (ANOVA).

Table 6: ANOVA results for the differences among means of four proficiency groups on the C-Test

Source of Variance	Sum of Squares	*df*	Mean Square	*F*	Sig.
Between Groups	10971.339	3	3657.113	30.4	.000
Within Groups	11638.305	97	119.983	80	

Table 6 shows the ANOVA results for the test of differences among the means obtained by the four proficiency groups on the C-Test. The obtained *F* ratio is significant at $p<.000$ level suggesting that there *is* a difference among the means. However, it has to be noted that the significance of the *F* ratio in an analysis of variance merely indicates that there is a significant difference among the means of the compared groups as a whole; that is, it indicates that there is at least *one* significant difference between the means of at least *one* pair of the groups compared (Brown, 1988). All the same, it does not tell us where exactly this difference lies, i.e., exactly which two means are different. In order to determine exactly which means differ one has to resort to *pairwise multiple comparisons*, which are considered *post hoc* or *follow-up* tests (Hatch & Farhady, 1982). The only requirement for these tests is that the overall *F* in the ANOVA is statistically significant.

Table 7 represents the results of a Tukey's honestly significant difference (HSD) test conducted on the means of the four proficiency groups. Tukey's HSD test is a commonly used multiple comparison test which reveals the precise location of differences by analyzing every two means separately (Brown, 1988; Delavar, 2002). Table 7 denotes that there is significant difference between

the means of every combination of two proficiency groups except for one: the upper and the lower intermediate groups. That is, the performances of the upper and the lower intermediate groups on the C-Test are not so much different that can be statistically acceptable.

Table 7: Results of Tukey's HSD multiple comparisons on the means of the four proficiency groups

	Elementary	Lower Inter.	Upper Inter.	Advanced
Elementary	-----			
Lower Inter.	7.31*	-----		
Upper Inter.	9.02*	2.12	-----	
Advanced	13.16*	6.25*	3.90*	-----

* Significant mean difference at $p<.05$ level

The fact that the C-Test has not been able to produce significant distinction between the two middle groups in this study is indicative of a lucid shortcoming for the C-Test, namely a low classification power. These results are not only in clear contrast to claims about the measurement accuracy of the C-Test (Dörnyei & Katona, 1992) but they also challenge the dependability of using C-Tests for placement purposes (Klein-Braley, 1997). This interpretation is further supported by an investigatin of decision consistency described below.

3.2.3.2. Decision consistency

The scores from the C-Test were also studied for decision consistency. Decision consistency refers to the agreement between the classifications of the same examinees based on two tests of the same ability (Livingston & Lewis, 1995). In more practical terms, decision consistency is "the percent classifications of subjects by the experimental test that correspond correctly to those by the criterion" (Jafarpur, 2002, p. 42). Table 8 shows the percent correct classifications that are made if the C-Test was used as the criterion. As can be observed, the C-Test can on the average correctly place just over fifty percent of the subjects in their appropriate proficiency

groups. It is by no means a promising quality for a test to fail to classify almost half of the examinees in their proper levels.

Table 8: Percent of correct classification predicted by the C-test

Criterion for Placement	Elementary	Lower Inter.	Upper Inter.	Advanced	Average
C-test	69%	39%	35%	62.5%	51.5%

3.2.3.3. C-Test and text difficulty

In an attempt to establish the construct validity of the C-Test, Klein-Braley (1985, p. 88) claims that it is possible to show that while their empirically measured difficulty (as C-Test texts) varies according to the subject group involved, the group of texts used in any one C-Test remains more or less constant in terms of relative difficulty.

Therefore, one construct validity concern is to see whether C-Test texts (or C-Texts) function similarly across proficiency levels. In order to explore how similarly subjects from different levels of proficiency perform on each C-Text an ANOVA was performed on the scores obtained from the five C-Texts for the four proficiency groups and the MA students. It was assumed that the mean performance of a group of subjects on a C-Text can be a good index of the difficulty of that C-Text for that particular group.

Table 9 provides the outcome statistics of the ANOVA. The significance of the *F* ratio found for each group (at $p<.000$ level) denotes that there are statistically meaningful differences among the means (i.e., average performances) of each group on the five C-Texts. Again, a Tukey's HSD test was carried out to specify on exactly which C-Texts the performances of each of the groups differ. Table 10 depicts the significant mean differences found among the five C-Texts for the four proficiency groups and the MAs.

Table 9: ANOVA results for differences among means of five C-Texts for five proficiency groups

Proficiency Group	Source of Variance	Sum of Squares	*df*	Mean Square	*F*	Sig. (*p*)
Elementary (N = 26)	Between texts Residual	935.123 1119.000	4 125	233.781 8.952	26.115	.000
Lower Inter. (N = 28)	Between texts Residual	1006.857 897.286	4 135	251.714 6.647	37.871	.000
Upper Inter. (N = 23)	Between texts Residual	1056.617 1033.304	4 110	264.404 9.394	28.147	.000
Advanced (N = 24)	Between texts Residual	788.783 1918.917	4 115	197.196 16.686	11.818	.000
MA (N = 43)	Between texts Residual	1506.400 2480.233	4 210	376.600 11.811	31.887	.000

As observed, the mean performance of both the upper intermediate and the MA groups on the first three C-Texts are not significantly different. However, their means on C-Text 4 and C-Text 5 not only show a significant difference from the other three C-Texts but from each other as well. While the table indicates a similar pattern for the lower intermediate group, it is however visible that the performance of this group has changed noticeably from C-Text 1 to C-Text 2, too. On the other hand, the results obtained for the means of the advanced group on the five C-Texts represent a completely different pattern. For them it is simply the C-Text 5 which is significantly different from the other four C-Texts. The pattern of mean differences for the elementary group, however, is so complicated that it is almost impossible to interpret.

Table 10: Results of Tukey's HSD for differences among the means of each proficiency group on the five C-Texts

		Elementary	Lower Inter.	Upper Inter.	Advanced	MA
C-Text 1	C-Text 2	*	*			
	C-Text 3					
	C-Text 4	*	*	*		*
	C-Text 5	*	*	*	*	*
C-Text 2	C-Text 1	*	*			
	C-Text 3					
	C-Text 4		*	*		*
	C-Text 5	*	*	*	*	*
C-Text 3	C-Text 1					
	C-Text 2					
	C-Text 4	*	*	*		*
	C-Text 5	*	*	*	*	*
C-Text 4	C-Text 1	*	*	*		*
	C-Text 2		*	*		*
	C-Text 3	*	*	*		*
	C-Text 5		*	*	*	*
C-Text 5	C-Text 1	*	*	*	*	*
	C-Text 2	*	*	*	*	*
	C-Text 3	*	*	*	*	*
	C-Text 4		*	*	*	*

* Significant mean difference at $p<.05$ level

What is evident is that there are no less than *four patterns* of mean difference among these five groups. The fact that these five groups have performed differentially on the five C-Texts can be interpreted as a counter evidence to Klein-Braley's (1985) claim concerning relative constancy of C-Test texts' difficulty independent of the subjects' proficiency level. These results are suggestive of the point that the C-Test suffers from one of the same problems as the cloze test does, namely the unpredictably variable nature of the cloze procedure (cf. Brown, 1993; see also Alderson, 1983; Klein-Braley, 1983). Jafarpur (1995) arrived at a similar

conclusion as a result of comparing 20 C-Test versions developed based on the same text.

3.2.3.4. Factorial validity

One of the most extensively used approaches in construct validation of language tests is factor analysis (Bachman, 1990). Factor analysis attempts to identify underlying variables, or factors, that explain the pattern of correlations within a set of observed variables (Farhady, 1983a; see also Oller & Hinofotis, 1980). Therefore, in order to further investigate the construct validity of the C-Test the scores of the subjects on the two measures were subjected to a factor analysis. To ensure higher precision, a principal axis factoring (PAF), as opposed to a principal components factoring (PCF), was employed to extract the initial factors (cf. Sharma, 1996; see also Carroll, 1983; Farhady, 1983a; Baker, 1989).

In order to determine the number of factors to be extracted, the eigenvalue-greater-than-one rule was utilized (Sharma, 1996). The eigenvalue-greater-than-one rule suggests that those factors whose eigenvalues (sum of squared loadings) are less than unity be excluded from the analysis. It appeared that only the eigenvalue for the first factor exceeded unity. Accordingly, the one-factor solution was adopted as the most reasonable.

Table 11: Results of factor analysis (subtests only)
Factor structure converged after 5 iterations

Test	Subtest	Factor 1
C-Test:	C-Text 1 C-Text 2 C-Text 3 C-Text 4 C-Text 5	.67 .75 .80 .80 .60
MTELP:	Grammar Vocabulary Reading	.77 .57 .49
Eigenvalue		4.283
Percent of total variance explained by the factor		53.539

Table 11 presents the results of the factor analysis with loading patterns on the first factor (Factor 1). Almost all measures have high loadings on Factor 1 (i.e., have high correlations with it). The highest belongs to C-Text 3 and C-Text 4 (.80) and the lowest to the reading comprehension test (.49). Also Factor 1 explains 53.5% of the total variance, that is, more than half of the variance produced by the eight measures entered into the analysis is due to Factor 1, which probably can be best interpreted as accounting for overall proficiency of the subjects in English. These results can also be regarded as evidence that the tests to a large extent measure the same construct.

Table 12: Results of factor analysis with MTELP as a single variable

Test (Subtest)	Factor 1
MTELP:	.80
C-Text 1 C-Text 2 C-Text 3 C-Text 4 C-Text 5	.66 .75 .81 .81 .61
Eigenvalue	3.732
Percent of total variance explained by the factor	62.199

Factor structure converged after 5 iterations

Another factor analysis was conducted, with MTELP entered as one single variable, so as to substantiate the results found above. The same methods were applied for factor extraction and for deciding on the number of factors. Table 12 shows the results of the second factor analysis where it was again only one factor whose eigenvalue was greater than one. That factor (once again termed Factor 1) could explain 62% of the variance with the MTELP heavily loading on it (.80) and with the C-Test texts demonstrating nearly the same loadings pattern as above. These figures serve to further confirm our conjecture that probably this first factor pertains to general proficiency in English. Given this conjecture is sustained,

the comparatively high correlation of the C-Texts with the first factor can be regarded as evidence that each of the C-Texts has a good claim to measuring language proficiency even on their own.

In view of the results of the two factor analyses just reported, it can be argued that the two experimental and criterion measures, to a great extent, tap the same underlying construct. Therefore, if what the MTELP measures is general language proficiency, then it is most probably what the C-Test measures as well. These results compare favorably with those of Jafarpur (2002) and Eckes and Grotjahn (2006), and provide support for Klein-Braley (1997) and Sigott's (2004) claims concerning the factorial validity of the C-Test.

4. Conclusions

The present study has primarily dealt with exploring the validity of the C-Test as reflected in domains of content relevance, criterion-relatedness, and construct meaningfulness. The reliability of the C-Test was also examined in the process. Reliability estimates found in this study confirmed earlier reports of high reliability coefficients associated with the C-Test.

A content/function word analysis was used to investigate the content validity of the C-Test. The C-principle showed a satisfactory method of sampling the linguistic elements in the text; hence, the claims of content validity made on its part are supported in this study.

As far as criterion-related validity is concerned, the C-Test scores correlated fairly highly with those of the MTELP. Not only that, but the C-Test's correlation coefficient with the MTELP was higher than with the grammar, vocabulary, and reading comprehension tests. This is to be considered further evidence in favor of the claims that C-Tests measure general language ability.

The C-Test also was capable of fulfilling many of the requirements of a suitable test in terms of construct validity. The most important finding in this view was the factorial validity found with the C-Test scores. The subparts of the C-Test manifested the highest loadings on the same factor as the MTELP suggesting that

not only the C-Test itself, but even the subparts of it had a substantial claim to measurement of general language proficiency.

The texts used in the C-Test, however, did not function uniformly with all proficiency groups. Each proficiency group found a different text or combination of texts more difficult. This finding reveals that when a text is turned into a C-Test, the C-Test text may unpredictably become more or less difficult for different proficiency levels. This is indicative of an unpredictable variability in C-Test results, a defficiency which has frequently been levelled against the cloze test as well.

As for the discrimination power of the C-Test, it came out that the C-Test did not perform very satisfactorily in differentiating subjects with different levels of linguistic ability. Specifically, the C-Test could not successfully discriminate between the subjects in lower and upper intermediate levels. In addition, a decision consistency analysis substantiated that the C-Test functioned poorly in classifying the participants in their appropriate proficiency levels. Therefore, contrary to C-Test proponents' claim (Klein-Braley & Raatz 1984; Klein-Braley 1997; Katona & Dörnyei 1993), the C-Test used in this study did not prove a very accurate and satisfactory placement test.

References

Alderson, J.C. (1979). The effect on the cloze test of changes in deletion frequency. *Journal of Research in Reading, 2,* 108-118.

Alderson, J. (1980). Native and non-native 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, Jr., (Ed.), *Issues in language testing research* (pp. 205-217). Rowley, MA: Newbury House.

Anastasi, A. (1982). *Psychological testing*. London: Macmillan.

Babaii, E. & Ansary, H. (2001). The C-test: a valid operationalization of reduced redundancy principle? *System, 29,* 209-219.

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. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

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

Brown, J. D. (1988). *Understanding research in second language learning: a teacher's guide to statistics and research design*. Cambridge: Cambridge University Press.

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

Carroll, B. J. (1980). *Testing communicative performance: an interim study*. London: Pergamon Institute of English.

Carroll, J. B. (1983). Psychometric theory and language testing. In J.W. Oller, Jr., (Ed.), *Issues in language testing research* (pp. 80-107). Rowley, MA: Newbury House.

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

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

Connelly, M. (1997). Using C-Tests in English with post-graduate students. *English for Specific Purposes, 16,* 139-150.

Corrigan, A, Dobson, B., Kellman, E., Spaan, M., & Tyma, S. (1979). *Michigan test of English language proficiency (Form Q).* Ann Arbor: Testing and Certification Division, the University of Michigan.

Delavar, A. (2002). *Probabilities and applied statistics in psychology and educational sciences.* Tehran: Roshd.

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

Eckes, T. & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing, 23,* 290-325.

Farhady, H. (1983a). On the plausibility of the unitary language proficiency factor. In J.W. Oller, Jr., (Ed.), *Issues in language testing research* (pp. 11-28). Rowley, MA: Newbury House.

Farhady, H. (1983b). New directions for ESL proficiency testing. In J. W. Oller, Jr, (Ed.), *Issues in language testing research* (pp. 253-269). Rowley, MA: Newbury House.

Farhady, H., Jafarpur, A., & Birjandi, P. (1994). *Testing language skills: from theory to practice*. Tehran: SAMT.

Harris, D. P. (1969). *Testing English as a second language*. New York: McGraw-Hill.

Hatch, E., & Farhady, H. (1982). *Research design and statistics for applied linguistics*. Rowley, MA: Newbury House.

Henning, G. (1987). *A guide to language testing*. Cambridge, Mass.: Newbury House.

Hinofotis, F. B. (1980). Cloze as an alternative method of ESL placement and proficiency testing. In J. W. Oller, Jr. & K. Perkins (Eds.), *Research in language testing* (pp. 121-34). Rowley, MA: Newbury House.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

Huhta, A. (1996). Validating an EFL C-test for students of English philology. In R. Grotjahn (Ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen [The C-Test: theoretical foundations and practical applications]* (Vol. 3, pp. 197-234). Bochum: Brockmeyer.

Ikeguchi, C. B. (1998). Do different C-tests discriminate proficiency levels of EL2 learners? *JALT Testing & Evaluation SIG Newsletter, 2*, 3-8. Retrieved November 20th, 2006, from: http://www.geocities.com/CollegePark/Field/1087/test/ike_1.htm

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

Jafarpur, A. (1997). *An introduction to language testing*. Shiraz: Shiraz University Press.

Jafarpur, A. (1999a). Can the C-test be improved with classical item analysis? *System, 27,* 79-89.

Jafarpur, A. (1999b). What's magical about the rule-of-two for constructing C-Tests? *RELC Journal, 30,* 86-100.

Jafarpur, A. (2002). A comparative study of a C-Test and a cloze test. In R. Grotjahn (Ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen [The C-Test: theoretical foundations and practical applications]* (Vol. 4, pp. 31-51). Bochum: AKS-Verlag.

Jonz, J. (1976). Improving the basic egg: The multiple-choice cloze. *Language Learning, 26,* 255-265.

Jonz, J. (1990). Another turn in the conversation: What does cloze measure? *TESOL Quarterly, 24,* 61-83.

Katona, L. & Dörnyei, Z. (1993). The C-test: A teacher-friendly way to test language proficiency. *English Teaching Forum, 31,* 34-35.

Klein-Braley, C. (1983). A cloze is a cloze is a question. In J.W. Oller, Jr., (Ed.), *Issues in language testing research* (pp. 218-228). Rowley, MA: Newbury House.

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

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

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

Laesch, K. B., & van Kleek, A. (1987). The cloze test as an alternative measure of language proficiency of children considered for exit from bilingual education programs. *Language Learning, 37,* 171-189.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32,* 179-197.

McBeath, N. (1989). C-Tests in English: Pushed beyond the original concept? *RELC Journal, 20,* 36-41.

McBeath, N. (1990). C-Tests – some words of caution. *English Teaching Forum, 28,* 45-46.

Mitchell, R., & Myles, F. (1998). *Second language learning theories*. London: Arnold.

Oller, J.W., Jr. & Hinofotis, F. B. (1980). Two mutually exclusive hypotheses about second language ability: Indivisible and partly divisible competence. In J. W. Oller, Jr. & K. Perkins (Eds.), *Research in language testing* (pp. 13-23). Rowley, MA: Newbury House.

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

Raatz, E. & Klein-Braley, C. (1995). Introduction to language testing and C-tests. In Coleman (Ed.), *University language testing and the C-test*. Proceedings of a conference held at the University of Portsmouth in April 1995. Also retrievable from: http://www.uni-duisburg.de/FB3/ANGLING/FORSCHUNG/HOWTODO.HTM

Sharma, S. (1996). *Applied multivariate techniques*. USA: John Wiley & Sons Inc.

Sigott, G. (2004). *Towards identifying the C-Test construct*. Frankfurt am Main: Peter Lang.

SPSS (1989-99). *SPSS for windows, release 9.0.0*. SPSS Inc.

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

Appendix

The C-Test developed by the researcher is seen below:

در هر یک از متون زیر جمله ی اول دست نخورده باقی مانده است. اما از جمله ی دوم به بعد از هر دو واژه یکی دستکاری شده است؛ به این معنا که نیمی از حروف آن حذف شده است. در هر واژه تعداد حروف قسمت حذف شده یا مساوی قسمت باقی مانده است یا یکی بیشتر. به عنوان مثال به متن زیر توجه کنید:

There are usually five men in the crew of a fire engine. One of them drives the engine. The leader sits beside the driver. The other firemen sit inside the cabin of the fire engine.

این متن به شکل زیر دستکاری می شود:

There are usually five men in the crew of a fire engine. One o____ them dri_____ the eng_____. The lea_____ sits bes_____ the

dri_____. The ot_____ firemen s_____ inside t_____ cabin o_____ the fi_____ engine.

اکنون با توجه به محتوای هر متن واژه های دستکاری شده را کامل کنید. غلطهای املایی لحاظ (موجب کسر نمره) خواهند شد.

Text One

In some cultures around the world, polygamy is recognized and accepted. This me_____(1) that a m_____(2) may ha_____(3) more th_____(4) one wi_____(5) or, i_____(6) some ca_____(7), a woman m_____(8) have mo_____(9) than o_____(10) husband a_____(11) the sa_____(12) time. Some_____(13) polygamous soci_____(14) occur wh_____(15), for so_____(16) reason, th_____(17) is a_____(18) imbalance bet_____(19) the num_____(20) of men and women, perhaps due to war, famine, or disease.

Text Two

The way teachers teach is often a personal interpretation of what they think works best in a given situation. For ma_____(21) teachers, a teac_____(22) approach i_____(23) something uniq_____(24) personal, wh_____(25) they dev_____(26) through exper_____(27) and ap_____(28) in diff_____(29) ways acco_____(30) to t_____(31) demands o_____(32) specific situa_____(33). Teachers cre_____(34) their o_____(35) roles wit_____(36) the clas_____(37) based o_____(38) their theo_____(39) of teac_____(40) and learning and the kind of classroom interaction they believe best supports these theories.

Text Three

How can you learn to focus your attention better while listening? The mo_____(41) important th_____(42) is t_____(43) concentrate o_____(44) what t_____(45) speaker i_____(46) saying. Y_____(47), your da_____(48) tomorrow ni_____(49) is ve_____(50) important, b_____(51) right n_____(52) you mu_____(53) listen. B_____(54) firm wi_____(55) yourself. I_____(56) your mi_____(57) wanders, br_____(58) it ba_____(59) to list_____(60). You cannot concentrate and daydream at the same time.

Text Four

A family of bees is called a hive or a colony. It h_____(61) been sa_____(62) that i_____(63) a beehive, t_____(64) king i_____(65) actually a qu_____(66). The wor_____(67) are dev_____(68) to h_____(69), and th_____(70) wait o_____(71) her a_____(72) all ti_____(73) and sat_____(74) every ne_____(75) that s_____(76) might ha_____(77). They br_____(78) her t_____(79) best fo_____(80), and they lick her body whenever she passes by them in the hive. They guard her constantly so that no enemy can attack and harm her.

Text Five

Underwater archaeologists have it easy. Wrecks l_____(81) undisturbed f_____(82) centuries a_____(83) are pres_____(84) in go_____(85) condition. B_____(86) there i_____(87) one b_____(88) problem-i_____(89) is da_____(90) down th_____(91). Powerful lig_____(92) do n_____(93) help: ne_____(94) the bot_____(95), clouds o_____(96) tiny part_____(97) scatter t_____(98) light li_____(99) fog. A n_____(100) underwater TV system from Westinghouse Oceanics uses a fine beam of blue-green laser light to quickly scan the depths.