C-Tests: Method Specific Measures of Language Proficiency

Ebrahim Khodadady
Ferdowsi University

Abstract

The present study reports the performance of 63 senior non-native undergraduate university students on C-Tests, a decontextualised C-Test, a spelling test, a matching vocabulary test (MVT) and a disclosed Test of English as a Foreign Language (TOEFL). It was postulated that the availability of the first half of deleted words might render the items independent from the context in which they appear. It was also postulated that the directions given on the C-Tests might call for the test takers' spelling knowledge rather than their vocabulary knowledge and language proficiency. To test the hypotheses, the C-Tests designed by Klein-Braley (1997) were administered along with their decontexualised versions. The analysis of the data showed that the C-Tests correlated significantly with all the tests except the decontextualised C-Test, implying that context plays an indispensable role in answering C-Tests. Although the C-Tests correlated significantly with the MVT and TOEFL, their coefficients were not high enough to establish them as independent measures of vocabulary and language proficiency. The initial principal component analysis revealed three factors. While the C-Tests had the highest loadings both on the first and second factors, the decontextualised C-Test and the spelling test loaded on the third. Varimax rotation showed that the MVT and TOEFL had the highest loadings on the first or G-factor whereas the C-Tests loaded on the second. The decontextualised C-tests had the highest loading on the third factor. These findings show that performing on the C-Tests requires an ability other than spelling ability, vocabulary knowledge and language proficiency. This ability is method specific and depends on understanding the context in which C-Test items appear.

Key Words: Spelling Ability, Vocabulary Knowledge, Language Proficiency, C-Tests, Context

Introduction

C-Tests were developed for the first time in 1981 by Raatz and Klein-Braley. Klein-Braley (1981) found that the way classical cloze tests are constructed, i.e., deleting every *n*th word in a reading text, brings about unsatisfactory performance on the part of test takers. Recent research confirms her findings. Khodadady (2007), for example, administered three cloze tests to eight intermediate adult ESL learners to explore the relationship between listening comprehension ability and vocabulary knowledge. The written cloze tests were developed on tape scripts to which learners had listened one week before taking the test. The four sample cloze items below form the beginning paragraph of the tape script dealing with a surefire way of predicting hurricanes.

Nature has its own way of … (1) us know that she's about to … (2) her fury. I have weathered many … (3) hurricanes in my time. Never have … (4) come unannounced.

Table 1 presents the descriptive statistics of four sample cloze items above. As can be seen in Table 1, the majority of participants have failed to restore even the first cloze item. The obtained item difficulty index of 0.29 reveals the fact that only 29% of participants could use the context of the test to restore the lexical verb *let*. None of them could even restore the subject pronoun *they* though they had heard it twice on the audio tape. Due to their extreme difficulty, cloze items fail not only to measure the test takers' ability of whatever they measure but also to discriminate high ability learners from their low ability counterparts.

Table 1

Descriptive statistics of four sample cloze items

Items	**Deleted words**	**Participants**	**Item Difficulty index**	**Item discrimination index**
1	Letting	2	0.29	-0.64
2	Unleash	1	0.14	-0.21
3	Fierce	0	0	0
4	They	0	0	0

To overcome the difficulty faced in answering standard cloze tests, Raatz and Klein-Braley (1981) decided to modify them by developing a deletion technique called the C-principle and named these new versions C-Tests. According to Klein-Braley (1997), between four and six short and authentic texts should be chosen by a test writer to design a well functioning C-Test. Since test writers usually overestimate text difficulty (Klein-Braley, 1994), they should begin with more texts than will be needed and order the texts intuitively according to their difficulty level.

After selecting the texts carefully, the second half of every second word beginning from word two in sentence two is deleted. If the word consists of odd numbered letters then the larger half is deleted. If a word has only one letter, it should be ignored in counting and deleting. Mutation of words should continue till its number reaches 100. Since test takers have not read the text before, its first and last sentences have been left untouched.

The mutilated words are usually piloted by administering them to a control group of adult educated native speakers or teachers of the language under assessment. The restored words should be around 90% correct. According to Klein-Braley (1997), when the native speakers score lower, the text should be examined very carefully for eccentricities. If necessary, it should be discarded. Alternative solutions should however be accepted.

Once C-Tests are constructed and piloted, they can be used to measure language proficiency through reduced redundancy. The idea of measuring language proficiency through reduced redundancy was first explained by Spolsky (1973). It was based on the assumption that knowing a language involves the ability to understand a distorted message by formulating valid guesses about a certain percentage of omitted elements.

After employing the Duisburg placement test DELTA, i.e., a high security test, as a validation criterion, Klein-Braley (1997) administered four C-Test texts along with two cloze tests, two multiple choice cloze tests, two cloze-elide tests requiring test takers to find some randomly added words in a text and cross them out (Manning, 1986) and a dictation test as measures of reduced redundancy to 81 university students.

Table 2 presents the results obtained by Klein-Braley (1997). As can be seen, three of the four C-Test texts have loadings higher than 0.70. Based on these results Klein-Braley concluded: “The best test to select to represent general language proficiency as assessed by reduced redundancy testing would be the C-Test” (p. 71).

Table 2
Factor analysis (unrotated solution): tests of reduced redundancy with DELTA
[adapted from Klein-Braley (1997, p.70)]

Tests	Factor 1	Factor 2
DELTA	.87	*
Cloze 1	.79	*
Cloze 2	.67	*
C-Test 1	.76	*
C-Test 2	.77	*
C-Test 3	.70	*
C-Test 4	.63	-.37
Multiple choice cloze 1	.60	*
Multiple choice cloze 2	.67	*
Cloze-elide 1	.60	.50
Cloze-elide 2	.56	.45
Dictation	.75	*
	Eigenvalue: 6.41	Eigenvalue: 1.33
	Variance: 53.4%	Variance: 11.1%

* Loadings less than .30

Do the results presented in Table 2 above provide enough evidence to support Klain-Braley’s (1997) claim regarding the best representativeness of C-Tests? Do C-Tests provide valid indicators of non-native speakers’ ability to function within a reduced redundancy context? In other words, do the C-Tests show that “the key thing missing is the richness of knowledge of probabilities – on all levels, phonological, grammatical, lexical, and semantic – in the language” (Spolsky 1973, p. 170).

In order to explore the ability of C-Tests to measure language proficiency, Khodadady (2004) developed a schema-based cloze multiple choice item test (MCIT) on an unmodified article excerpted from *NewScientist* magazine (5 August 1995, No. 1989). Similar to rational cloze tests, in schema-based cloze MCITs a number of words are rationally chosen and deleted completely from the text and presented along with three syntactically, semantically and discoursally related words as item altnerantives. These tests are based on the assumption that each and all words comprising a given text represent the text writer’s knowledge and experiences with the concepts expressed individually and collectively and should therefore activate the same, similar and/ or even idiosyncratic knowledge and experiences in the minds of text readers or test takers (Khodadady, 2001).

For example, the two cloze multiple choice items below are schema-based because they expose test takers to a testing situation which requires their knowledge and comprehension of and experiences with each and all words used in the sentence, i.e., individually and collectively, respectively.

Fears over access to medical records

Privacy campaigners in the US have launched a fierce ... (1) on a bill

1	A. raid	B. slander	C. attack*	D. ambush
2	A. inquiring	B. prying*	C. interfering	D. probing

that they believe will expose medical records to too many … (2) eyes.

In order to answer item 1, test takers must know what the four choices mean individually. They should then focus on all the words used in the sentence in order to decide which alternative fits the blank best. The keyed response, i.e, *attack*, of the first item and its alternatives, i.e, *raid*, *slander* and *ambush*, have syntactic and semantic relationships with each other. Since they are syntactically nouns by nature, they can all fill the same slot. In addition to being syntacticlly related, the alternatives share the semantic feature of *assault* and must therefore be equally attractive to test takers.

However, in order for test takers to choose the keyed response, they must activate their discoursal knowledge and relate it to the contextual expressions of *privacy campaigners* and *bill*, which dictate what type of assault should be launched. *Raid* and *ambush* are not what the writer has used because they involve physical assault. Since *attack* shares the semantic feature of *physical assault* with *raid* and *ambush*, a test designer can rationally predict that they will appeal to the test takers more than *slander*. They will have no choice but to read all the words preceding and following the deleted word in order to make an informed choice.

Khodadady (2004) administered the schema-based cloze MCIT with the C-Test (Klein-Braley, 1977), text-driven cloze test (Farhady & Keramati, 1994) and traditional cloze MCIT (Hale, Stansfield, Rock, Hicks, Butler, & Oller, 1988) to 34 senior undergraduate Iranian students. The disclosed TOEFL test 1 (Educational Testing Service, 1991, pp. 75-100) was used as an internationally accepted measure of English language proficiency.

Table 3 presents the results obtained by Khodadady (2004). As can be seen, similar to Klein-Braley's (1997) findings, the C-Test and C-test 2 have the highest loadings on the first factor, i.e., 0.93 and 0.78, respectively. These loadings can not, however, show language proficiency as klein-Braley claims. This is because the TOEFL has the second lowest loading on this factor, i.e, 0.69. Furthermore, the TOEFL loads on the second factor on which the C-test and its subtests all have negative loadings. Due to these unexpected loadings, Khodadady ran a rotated factor analysis on the data.

Table 3

Unrotated Factor Matrix using principle factor with iteration for the redundancy tests and TOEFL [from Khodadady (2004, p. 234)]

Test	**Factor 1**	**Factor 2**
TOEFL	.69	.60
C-Test 1	.81	*
C-Test 2	.85	-.31
C-Test 3	.70	*
C-Test 4	.65	-.46
C-Test (Total)	.93	-.36
Schema-based cloze MCIT	.73	.50
Text-driven cloze test	.74	*
Traditional cloze MCIT	.73	.44
	Eigenvalue: 5.13	Eigenvalue: 1.30
	Variance: 57.32%	Variance: 14.43%

* Loadings less than .30

Table 4 presents varimax rotation of factors for the reduced redundancy tests and the TOEFL. As can be seen, the TOEFL test does not load on the first factor any more. Only the C-Test (0.95) and its subtests, i.e., C-Test 1 (0.72), C-Test 2 (0.80), C-Test 3 (0.67) and C-Test 4 (0.80) have the highest loadings on this factor. According to Khodadady (2004), these results indicate that C-Tests have their own effect. Since the TOEFL differs from the four methods of reduced redundancy in terms of its construction, it does not load on the first factor. The TOEFL has, however, the highest loading on the second factor (0.90), upon which schema-based cloze MCIT, text-driven cloze test, traditional cloze MCIT and even C-Test load, indicating that the second factor represents English language proficiency.

Table 4

Varimax with Kaiser rotated factor matrix using principal component analysis for the reduced redundancy tests with the TOEFL [from Khodadady (2000, p. 64)]

Tests	Factor 1	Factor 2
TOEFL	*	.90
C-Test 1	.72	.39
C-Test 2	.80	*
C-Test 3	.67	*
C-Test 4	.80	*
C-Test	.95	.30
Schema-based cloze MCIT	*	.85
Text-driven cloze test	.56	.49
Traditional cloze MCIT	*	.81
	Eigenvalue: 3.62 Variance: 40.27%	Eigenvalue: 2.81 Variance: 31.21%

* Loadings less than .30

The present study was, therefore, conducted to determine the nature of the first factor upon which C-Tests have the highest loadings. Since the second half of every second word from the second sentence is deleted, it was postulated that *test takers perform well on the C-Tests just because they resort to their vocabulary and spelling knowledge independently of the context in which they occur*.

Method

Participants

The tests employed in the present study were administered to 63 senior undergraduate Iranian students majoring in English language and literature at Ferdowsi University of Mashhad in Iran in 2008. They participated in the research project voluntarily in order to obtain a valid and reliable measure of their own English language proficiency. Out of 63 students, 44 (70%) were female, and the rest were male. The participants' age ranged between 21 and 32 and most of them were between 22 and 24 years old (69%). They spoke Turkish (68%), Kurdish (18%) and Persian (14%) as their mother languages.

Instrumentation

In the present study eight tests were employed: C-Tests, a decontextualised C-Test, a spelling test, a structure test, a written expression test, a traditional multiple choice vocabulary test, a reading comprehension test, and a matching vocabulary test. The structure, written expression, traditional multiple choice vocabulary and reading comprehension tests formed the subtests of a disclosed TOEFL test (Educational Testing Service, 1991).

C-Tests

The C-Tests developed by Klein-Braley (1997, pp. 79-80) were used in this study. According to Klein-Braley and Raatz (1985, 1990), C-Tests measure test takers' language proficiency and should have at least 100 items. The C-Tests employed in this study consisted of 99 items. They comprised four texts upon which C-Test 1, C-Test2, C-Test 3 and C-Test 4 were designed. With the exception of C-Test 2, which had 24 items, the other three C-Tests had 25 items each. The reliability coefficient (KR-21) reported for the C-Test was 0.85.

Decontextualised C-Test

Based on the C-Tests designed by Klein-Braley (1977) a decontextualised C-Test was constructed by the present researcher. In this test the 99 mutilated words comprising the texts on which C-Test 1, C-Test 2, C-Test 3, and C-Test 4 were developed were taken out of their linguistic context and numbered from 1 to 99 (Appendix 1). The test takers were required to restore the mutilated half of the words by chance and on the basis of the directions. They had to restore the larger "half" of the mutilated word if it consisted of an odd number of letters. The restored word had to be only one word and its spelling had to be correct. In this test only the exact correct words comprising the texts of the C-Tests were scored correct. For example, item 6 on the decontextualised C-Test requires test takers to supply the letter ***e*** and ***w*** to restore the mutilated word ***few*** as the exact answer.

Spelling Test

The restored mutilated words on the decontextualised C-Test were scored for the second time. Rather than accepting only the exactly restored mutilated words which comprised the C-Tests, all restored mutilated words having the specified number of letters, i.e., the second even or odd half, were scored correct on second scoring. For this purpose various references such as *Collins Dictionary of the English Language* (Hanks, 1986) were consulted to find out whether the restored words having the specified number of letters did exist in English. For example, the words *far*, *fat*, *few*, *fit*, *fix*, *for*, *fun* and *fur* were scored correct for item 6. As this example shows, in this method both the exactly restored mutilated words and words having the specified number of letters were marked correct. Since this method measures the test takers' knowledge of letters comprising English words irrespective of their meaning and context, it was labeled the spelling test.

Structure Test

Following Khodadady and Herriman (2000) the structure subtest of the disclosed written TOEFL test (Educational Testing Service, 1991) was employed to measure the grammar proficiency of the participants. It consisted of 30 cloze multiple choice items developed on 30 isolated and unrelated sentences, which address a discrete grammatical point.

Written Expressions Test

The written expressions subtest of the disclosed written TOEFL test (Educational Testing Service, 1991) was also employed to measure the grammar proficiency of the participants. It consisted of 25 isolated and unrelated sentences whose four parts had been underlined and numbered. In contrast to the structure subtest, the written expressions subtest of the TEOFL requires test takers to identify the erroneous part of sentences.

Multiple Choice Vocabulary Test

The multiple choice vocabulary subtest of the disclosed written TOEFL test (Educational Testing Service, 1991) was used in the present study to measure test takers' global vocabulary knowledge. It consisted of 30 items developed on isolated and unrelated sentences. Each item presents one single sentence in which one word or phrase has been underlined. From among the four alternatives given below each item, test takers must choose the keyed response which can be replaced by the underlined word.

Reading Comprehension Test

The multiple choice reading comprehension subtest of the disclosed written TOEFL test (Educational Testing Service, 1991) was employed to measure test takers' reading comprehension ability. It consisted of 30 items developed on five short passages. Test takers had to read and understand the passages so that they could answer the questions.

Matching Vocabulary Test

The matching vocabulary test designed by Paul Nation (see Schmitt, Schmitt, & Clapham, 2001) was also used to determine test taker's lexical knowledge. It consisted of 60 items presented in 20 groups of three words having six words opposite to select from. The test takers had to choose the correct answer from the six words which best fitted each meaning. Then they had to put the meaning number next to the word which suited it best as shown below.

Example		**Your Answer**	
1. assert	______ cast	1. assert	___3__ cast
	______ confide		______ confide
2. ban	______ state	2. ban	___1__ state
	______ detest		______ detest
3. throw away	______ falter	3. throw away	______ falter
	______ forbid		___2__ forbid

Procedure

The participants took the decontextualised C-Test in the first session. Their answers were scored once according to the context in which they appeared in the C-Tests. The responses elicited on the decontextualised C-Test were marked once again in a different method to yield the spelling test. In this method all restored English words meeting the criteria, i.e., having the second even or odd half number of letters, were marked correct. For example, the following responses given to items 1 of the spelling test were marked correct. (English dictionaries were consulted to mark the spelling test manually.)

Item	Responses Marked Correct for the Spelling Test
1	Mock, Mode, Money, Month, Mood, Moon, Moral, More, Morn, **Most,** Motif, Motif, Motor, Mouse, Mouth, Move

After one week, the C-Tests were administered under standard conditions. Test takers were required to write their answers on the test booklet. All the restored mutilated words were marked manually. According to its designer's guidelines, only the exact words having no spelling errors were marked correct.

The structure, written expressions and reading comprehension subtests of the TOEFL were administered together in one session one week later. Test takers were required to mark their answers on answer sheets administered along with the test booklet. The answer sheets were marked manually.

And finally the matching vocabulary test and the multiple choice vocabulary subtest of the TOEFL were administered together in a session in the fourth week. No answer sheets were used. Test takers had to mark their answers on the test booklet. The booklets were scored manually.

Data analysis

For estimating the internal consistency reliability of the C-Test, decontextualised C-Test and spelling test, Cronbach's $\propto$ was employed. In addition to reliability, the internal validity of the three tests was determined through item difficulty and item discrimination estimated by employing *p*-values and point biserial correlation coefficients (r_{pbi}). *P*-values were calculated as the proportion of correct responses given to each item. *P*-values falling within the range of 0.25 to 0.75 were considered acceptable (Baker, 1989). The r_{pbi} coefficients were estimated by correlating each individual item with the total test score. Items having r_{pbi} coefficients of 0.25 or higher were accepted as well-functioning items.

The external validity of the C-Test, decontextualised vocabulary test and spelling test, was determined by correlating them with the matching vocabulary test and the TOEFL and its subtests as external criteria. Principal component and factor analyses were also run to determine the factors upon which the tests would load. All statistical analyses were performed by using SPSS Release 11.5 for Windows, standard version. These analyses were carried out to answer the following questions:

1. How reliable are the C-Tests, the Decontextualised C-Test and the Spelling Test?

2. How valid are the C-Tests, the Decontextualised C-Test and the Spelling Test?

3. What is the factor structure for the C-Tests, the Decontextualised C-Test and the Spelling Test? Do they load on the same factor?

4. What is the factor structure if the matching vocabulary test (MVT) and TOEFL are included? Do the C-Tests load on the same factor as the MVT and TOEFL do?

Results and Discussion

Table 5 presents the descriptive statistics for the C-test, Decontextualised C-Test, Spelling test, Matching Vocabulary Test and the TOEFL and its subtests. In terms of difficulty, as judged by mean *p*-value, the decontextualised vocabulary test (.18) and the matching vocabulary test (.25) were the most difficult. Both tests lacked context. In contrast to the decontextualised C-Test, however, test takers had six alternatives to choose the best response from on the matching vocabulary test. While the spelling test posed the easiest items (*p*-value = .81), the C-Tests enjoyed an acceptable level of difficulty, i.e., *p*-value = .53.

Table 5

Basic descriptive statistics for the C-test, Decontextualised C-Test, matching vocabulary test, spelling test, TOEFL and its subtests administered to 63 test takers

Tests	**No. of items**	**Mean**	**Sd**	**Kurtosis**	**Mean *p*-value**	**Mean r_{pbi}**	**α**
C-Tests	99	53.0	12.3	.44	.53	.28	.89
C-Test 1	25	14.5	3.8	.32	.57	.30	.71
C-Test 2	24	12.2	3.7	-.34	.51	.29	.71
C-Test 3	25	14.4	3.6	.80	.57	.31	.71
C-Test 4	25	12.0	3.3	-.42	.47	.23	.65
Decontextualised C-Test	99	17.5	5.0	.22	.18	.14	.60
Spelling test	99	80.2	11.8	2.94	.81	.36	.89
Matching vocabulary test	60	14.7	8.9	2.36	.25	.36	.89
TOEFL	115	55.5	10.6	-.35	.63	.28	.90
Structure	30	20.7	4.1	-.61	.69	.27	.74
Written expressions	25	15.7	4.1	.59	.63	.29	.74
Vocabulary	30	17.5	4.2	.61	.63	.26	.79
Reading comprehension	30	19.0	4.8	-.36	.58	.28	.79

Reliability

As can be seen in Table 5, both the C-Test and the spelling test are highly reliable (α = .89). This high level of reliability is shared by the matching vocabulary test, i.e., α = .89). Among the various tests administered in the study, the decontextualised C-Test had the lowest reliability coefficient (α = .60). This degree of moderate reliability was, nonetheless, expected because test takers did not have any context or alternatives to help them restore the mutilated words comprising the texts of the C-Tests.

Internal Validity

The internal validity of decontextualized C-Tests, C-Tests, the TOEFL test and matching vocabulary test was determined by estimating the percentage of their well functioning items, i.e., items whose *p*-values fall within 0.25 and 0.75 and their r_{pbi} is 0.25 or higher. Table 6 shows the percentage of these items. Table 6 shows the percentage of well functioning items.

Table 6

Percentage of well functioning items comprising the four tests

Tests	Total number of of items	No of well functioning items	Percentage
Decontextualised C-Tests	99	11	11
Matching vocabulary test	60	23	38
C-Tests	99	37	37
TOEFL	115	48	42

As shown in Table 6, C-Tests were internally valid measures of language proficiency because 37 percent of their items had educationally appropriate levels of difficulty and could discriminate between high and low ability language learners. In contrast, decontextualized C-Tests had very low internal validity due to their low percentage of well functioning items, i.e., 11%. The spelling test totally lacked internal validity because no criterion could be established to accept the restored words. For example, the test takers had given 16 semantically and orthographically different answers to item 94, i.e. Dad, Dam, Day, Dew, Did, Die, Dig, Dim, Din, Dip, **Do,** Dog, Dot, Dry, Due, Dye. The difference in the meaning of the restored words makes assigning the same score to different answers questionable if not unfair.

Empirical Validity

Table 7 presents the correlation coefficients of the C-Test with the TOEFL and its subtests, decontextualised C-Test, spelling test and matching vocabulary test. As it can be seen, the C-Test shows *no* relation with the decontextualised C-Test. It does, however, correlate significantly with the spelling test, i.e., 0.38, and thus provides spelling tests with empirical validity.

Table 7

Correlation coefficients of the C-test, Decontextualised C-Test, Spelling and Matching vocabulary Test with the TOEFL and its subtests

Tests	**C-Tests**	**TOEFL**	**Str**	**WE**	**Voc**	**Read**	**DCT**	**Spell**
TOEFL	.62**	1						
Structure (Str)	.55**	.85**	1					
Written expressions (WE)	.49**	.83**	.68**	1				
Vocabulary (Voc)	.53**	.65**	.53**	.49**	1			
Reading (Read)	.49**	.78**	.46**	.40**	.57**	1		
Decontextualised C-Test (DCT)	.23	.16	.18	.28*	.19	-.03	1	
Spelling (Spell)	.38**	.37**	.37**	.34**	.41**	.22	.66**	1
Matching Vocabulary test	.42**	.58**	.50**	.44**	.63**	.48**	.12	.29*

* Correlation is significant at the 0.05 level (2-tailed)

** Correlation is significant at the 0.01 level (2-tailed)

As shown in Table 7, the C-Test correlates significantly with the TOEFL (0.62). This finding is compatible with that of Chihara, Cline and Sakruai's (1996) finding whose Japanese junior college students' scores on the TEOFL correlated 0.57 to 0.65 with the C-Tests. Farhady and Jamali's (1999) study, however, shows a lower correlation between the C-Test and the TOEFL, i.e., 0.46. In spite of being significant, these levels of correlation do not support the claim that the C-Tests measure language proficiency. According to Hatch and Lazaraton (1991, p. 442), in order for two tests to measure the same ability they should correlate 0.80 or higher. The correlation coefficient of 0.62 between the C-Tests and the TOEFL does *not* validate C-Tests as measures of English language proficiency.

Similarly, although the C-tests employed in this study correlated significantly with the TOEFL vocabulary subtest (0.53) and matching vocabulary test (0.42), their correlation coefficients did not establish them as measures of vocabulary knowledge. These findings are compatible with those of Chihara, Cline, and Sakurai (1996) whose study showed that the C-Tests correlated significantly from 0.38 to 0.50 with vocabulary tests. The C-tests have, however, shown a very high correlation, i.e., 0.84, with 'grammatically based" vocabulary tests (Chapelle & Abraham, 1990, p. 146).

Factorial Validity

A principle component analysis was run in order to explore the research questions: What is the factor structure for the C-Tests, the Decontextualised C-Test and the Spelling Test? Do they load on the same factor? Table 8 presents the results. As it can be seen, the C-Tests load the highest on the first factor (0.98) whereas the decontextualized C-Test and the Spelling Test have the highest loading on the second factor, i.e., 0.83 and 0.74, respectively. These results indicate that the C-Tests measure an ability other than spelling and vocabulary knowledge.

Table 8

Unrotated Factor Matrix using principle factor with iteration for the C-Test, decontextualised C-Test and spelling test

Tests	**Factor 1**	**Factor 2**
C-Test	.98	*
C-Test 1	.89	*
C-Test 2	.83	*
C-Test 3	.85	*
C-Test 4	.75	*
Decontextualised C-Test	.39	.83
Spelling test	.53	.74
	Eigenvalue: 4.17 Variance: 59.58%	Eigenvalue: 1.39 Variance: 19.83%

In order to find out what ability the C-Tests measure, the matching vocabulary test (MVT) and the TOEFL were included in the principal component analysis. Table 9 presents the results. As it can be seen, the C-Tests and the TOEFL load the highest on the first factor, i.e., 0.90 and 0.80, respectively. Since all the tests load on this factor, it represents English language proficiency. However, the C-Tests and the Decontextualised and spelling tests also load on the second and third factors whose inferred nature calls for employing rotation to achieve "a simpler factor structure, preferably with each variable loading primarily on only one factor" (Farhady 1983, p. 19).

Table 9

Unrotated Factor Matrix using principle factor with iteration for the C-Test, decontextualised C-Test, matching vocabulary test, spelling test, TOEFL and its subtests

Tests	**Factor 1**	**Factor 2**	**Factor 3**
Matching vocabulary test	.64	-.39	*
TOEFL	.88	-.39	*
Structure	.77	*	*
Written expressions	.72	*	*
Vocabulary	.73	-.30	*
Reading comprehension	.67	-.42	*
C-Tests	.90	.36	*
C-Test 1	.79	.41	*
C-Test 2	.74	.38	*
C-Test 3	.77	.34	*
C-Test 4	.75	*	*
Decontextualised C-Test	.32	.40	.78
Spelling test	.52	*	.69
	Eigenvalue: 6.75 Variance: 51.93%	Eigenvalue: 1.48 Variance: 11.39%	Eigenvalue: 1.44 Variance: 11.10 %

* Loadings less than .30

Table 10 presents the varimax rotated factor matrix using principle component analysis of the C-Tests, Decontextualised C-Test, matching vocabulary test, spelling test, TOEFL and its subtests. As it can be seen, this time the TOEFL loads the highest on the first factor upon which matching vocabulary test and its structure as well as reading comprehension subtests have the second highest loadings. These results provide further support for labeling the first factor as English language proficiency. Since neither the C-Tests nor the Decontextualised and spelling tests load on the first factor, they must measure method-specific abilities. Although it sounds reasonable to expect the spelling test to load on the first factor, it loads highly on the third factor to reveal its very unique method of construction. The method-specific nature of spelling test is further revealed when the highest loading of the Decontextualised C-Test is taken into account. None of the items comprising these two tests were selected randomly from any authentic text to reveal the possible role of spelling and decontextualised vocabulary knowledge in measuring language proficiency. Neither do the loadings of the decontextualised and spelling test support the assumption that since the first half of the mutilated words in the C-Tests are given, the test takers' knowledge of the spelling structure of these words would help them restore the deleted half without reading the context provided in the C-Tests. These loadings in turn highlight the method-specific nature of the C-Tests themselves because they load on a separate factor.

Table 10

Varimax with Kaiser rotated factor matrix using principal component analysis for the C-Tests, decontextualised C-Test, matching vocabulary test, spelling test, TOEFL and its subtests.

Tests	**Factor 1**	**Factor 2**	**Factor 3**
Matching vocabulary test	.73	*	*
TOEFL	.90	.33	*
Structure	.75	.30	*
Written expressions	.70	*	*
Vocabulary	.74	*	*
Reading comprehension	.75	*	*
C-Tests	.35	.93	*
C-Test 1	*	.86	*
C-Test 2	*	.83	*
C-Test 3	*	.85	*
C-Test 4	.47	.58	*
Decontextualised C-Test	*	*	.93
Spelling test	*	*	.83
	Eigenvalue: 6.75 Variance: 51.93%	Eigenvalue: 1.48 Variance: 11.39%	Eigenvalue: 1.44 Variance: 11.10 %

* Loadings less than .30

Conclusion

Although C-Tests were invented in 1981, they have received little attention in *English* language testing literature. For example, in his fairly comprehensive review of correlational studies conducted on C-Tests so far, Sigott (2004, pp 61-65) could tabulate 28 among which only 11, i.e., 39%, have been in English. This is reflected in textbooks written for teacher training programs, e.g., Madsen (1983), Heaton (1988), Baker (1989), where C-Tests are not even mentioned. Similarly, there is no entry for C-Tests in the *Dictionary of Language, Teaching and Applied Linguistics* (Richards, Platt, & Platt, 1992). Bachman (1990), however, referred to C-Tests as variants of cloze tests in passing.

The overall public and expert inattention might be attributed partly to a lack of research on C-Tests and partly to their nature. Davies (1990) dubbed them 'a particular and rather recondite use of the cloze test' (p. 94) and thus obliged researches like the present one to contribute to those studies which have already shed some light on their internal, empirical and factorial validity.

This study showed that C-Tests reveal a significant correlation with established language proficiency tests such as the TOEFL. They should not, however, be used *alone* to measure test takers' language proficiency because they fail to correlate with standardized tests highly enough to replace them. The C-Tests nonetheless measure a language proficiency ability which is lacking in traditional multiple choice items measuring structure, vocabulary, and reading comprehension abilities.

The results of the present study call for a more comprehensive research to find out whether C-Tests load on a specific factor when they are administered along with different language proficiency tests such as IELTS and the TOEFL. Future research may focus on the question whether the C-Tests constructed on authentic and unmodified texts can function as well as those developed by testing specialists. As Eckes and Grotjahn (2006, p. 316) concluded "the exact nature of what

C-tests measure depends to some extent on the ability level of the examinees and on the difficulty of the C-test."

Received 25 March, 2006
Accepted 31 August 2007

Acknowledgements

I am grateful to the anonymous reviewers of the *Iranian Journal of Applied Linguistics* for their fruitful comments. I would also like to thank Dr. Hossein Farhady for his reading and inputs on the final draft of the present paper.

References

Baker, D. (1989). *Language testing: a critical survey and practical guide*. London: Edward Arnold.

Chapelle, C.A., Abraham, R.G. (1990). Cloze method: what difference does it make? *Language Testing*, 7, 121–46.

Chihara, T., Cline, W.D., & Sakurai, T. (1996). If the cloze test is a question, is the C-test the answer? In R. Grotjahn (Ed.). *Der C-Test: theoretische Grundlagen und praktische Anwendungen [The C-test: theoretical foundations and practical applications]* (Vol. 3) [183–95]. Bochum: Brockmeyer.

Davies, A. (1990). *Principles of language testing*. Oxford: Basil Blackwell.

Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing*, 23/3, 290-325.

Educational Testing Service. (1991). *Reading for TOEFL*. Princeton, NJ: ETS.

Farhady, H. & Keramati, M. N. (1994). A text-driven method for the deletion procedure in cloze passages. *Language testing*, 191-207.

Farhady, H., & Jamali, F. (1999). Varieties of C-test as measures of general proficiency. *Journal of the Faculty of Foreign Languages* (Tehran University Press), 3, 23–42

Grotjahn, R. (1986). Test validation and cognitive psychology: Some methodological considerations. *Language Testing* 3/2, 159-85.

Hale, G. A., Stansfield, C. W., Rock, D. A., Hicks, M. M., Butler, F. A., & Oller, J. W. Jr. (1988). *Multiple-choice items and the Test of English as a Foreign Language* (TOEFL Research Report No. 26). Princeton, NJ: Educational Testing Service.

Hanks, P. (Ed.). (1986). *Collins Dictionary of the English Language* (2nd ed.). London: Collins.

Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statisitics for applied linguistics*. Boston, Mas.: Heinle & Heinle.

Heaton, J. B. (1988). *Writing English language tests* (new edition). Essex: Longman.

Khodadady, E. (1997). *Schemata theory and multiple choice item tests measuring reading comprehension*. Unpublished PhD thesis, the University of Western Australia.

------------, (2001). Schema: A theory of translation. In S. Cunico (ed.). *Training Translators and Interpreters in the New Millennium, Portsmouth 17th March 2001 Conference Proceedings* (pp. 107-123). University of Portsmouth: School of Languages and Areas Studies.

------------, (2004). Schema-based cloze multiple choice item tests: Measures of reduced redundancy and language proficiency. *ESPecialist*, **25**/**2**, 221-243.

------------, (2007, March). *Knowledge of recently taught words and listening comprehension ability*. Paper presented at the annual TESL Niagara Conference, Welland, Canada.

Khodadady, E., & Herriman, M. (2000). Schemata Theory and Selected Response Item Tests: From Theory to Practice. In A. J. Kunnan (Ed.), *Fairness and validation on language assessment* (pp. 201-222). Cambridge: CUP.

Klein-Braley, C. (1981). *Empirical investigations of cloze tests*. Unpublished PhD dissertation, University of Duisburg.

------------, (1994). *Language testing with the C-Test. A linguistic and statistical investigation into the strategies used by C-Test takers, and the prediction of C-Test difficulty*. Unpublished higher doctoral thesis (Habilitationsschrift), University of Duisburg.

------------, (1997). C-Tests in the context of reduced redundancy testing: an appraisal. *Language Testing*, 14/1, 47-84.

Klein-Braley, C., & Raatz, U. (1990). Die objective Erfassung des Sprachstands im mutter- und fremdsprachlichen Unterricht durch C-Tests. In A. Wolff, & H. Rössler (Eds.). *Deutsch als Fremdsprache in Europa* (pp. 239-50). Regensburg: Arbeitskreis Deutsch als Fremdsprache.

------------, (1985). *Fremdsprachen und Hochschule 13/14: Thematischer Teil: C-Tests in der Praxis*. Bochum: AKS.

Madsen, H., S. (1983). *Techniques in testing*. Oxford: Oxford University Press.

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

Raatz, V., & Klein-Braley, C. (1981). The c-test- a modification of the cloze procedure. In T. Culhane, C., Klein-Braley, & D. K. Stevenon (eds.). *Practice and problems in language testing. University of Essex Occasional Papers* (pp. 113-38). Colchester, Essex: Department of Language and Linguistics, University of Essex.

Richards, J. C., Platt, J., & Platt, H. (1992). *Longman dictionary of language, teaching and applied linguistics* (3rd ed.). Essex: Longman.

Schmitt, N., Schmitt, D. & Clapham, C. (2001) Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing* **18**,1, 55-88.

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

Spolsky, B. (1973). What does it mean to know a language; or how do you get somebody to perform his competence? In J. W. Oller J. and J. R. Richards (Eds.). *Focus on the learner* (pp.164-76). Rowley, MA: Newbury House.