ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 5, No. 6, pp. 1353-1362, November 2014
© 2014 ACADEMY PUBLISHER Manufactured in Finland.
doi:10.4304/jltr.5.6.1353-1362

Construct Validity of C-tests: A Factorial Approach

Ebrahim Khodadady
Ferdowsi University of Mashhad, Iran

Abstract—**This study explored the latent variables underlying C-Tests by analyzing the performance of 416 undergraduate and graduate university students majoring in English Language and Literature, English Language Teaching and English Translation in Mashhad, Iran. The C-Tests designed by Klein-Braley (1997) were changed into two other types of tests called Spelling Test and Decontextualised C-Test and administered along with the disclosed Test of English as a Foreign Language (TOEFL), a semantic schema-based cloze multiple choice item test (S-Test) and a lexical knowledge test (LKT). The application of principal component analysis (PCA) to the responses of the participants on the three tests, i.e., C-Tests, Spelling Test and Decontextualised C-Test, revealed two components called *language proficiency* and *direction specificity* in this study. While the inclusion of the S-Test, TOEFL and the LKT in the PCA yielded the same two components, their rotation brought about the highest loadings of the included tests as well as the moderate loadings of the C-Tests on the first component, validating them as proficiency measures of language. However, they loaded the highest on the second component along with the Spelling Test and Decontextualised C-Test and thus confirmed their spelling and direction specificity. The implications of the study are discussed and suggestions are made for future research.**

Index Terms—**C-tests, language proficiency, schema theory, construct validity**

I. Introduction

Theoretical discussion of measuring language proficiency has been a largely mute topic in the field of applied linguistics due to the lack of consensus on defining language proficiency in an operationalized and universally accepted manner. Kelly (1978), for example, asserted that “it is the purpose of a proficiency test to assess whether or not candidates are indeed capable of participating in typical communication events from the specified communication situations(s)” (p. 350). The assertion reflects the era in which little room, if any, was given to any foreign language teaching methods other than communicative approach.

Wilkins (1976) declared that communicative tests should seek answers to such questions as the test takers’ ability to perform certain functions in appropriate social environment. Based on several typologies of language functions outlined by scholars such as Austin (1961), Halliday (1973), Searle (1966), Sinclair and Coulthard (1975), and van Ek (1976) some attempts were made to develop communicative tests. Farhady (1980, 1983), for example, developed and validated the first functional test to measure the English proficiency of students who used English as their second language. However, these tests never gained popularity because translating functions into items measuring linguistic competence within social contexts proved to be too difficult.

Although communicative approach has paved the way for a fairly large number of teaching methods such as Task-Based Language Teaching (e.g., Belgar & Hunt 2002), it has *not* resulted in developing any new and widely accepted language proficiency test in the “Post-Method” era suggested by Brown (2002). Still the tests are developed on structures and vocabulary as language components and are presented either orally or in writing to measure skills such as listening and reading. These components and skills are, for example, measured in the Test of English as a Foreign Language (TOEFL). According to the official site of Educational Testing Service (ETS 2010), the TOEFL is “*the most widely respected* English-language test in the world, recognized by more than 7,500 colleges, universities and agencies in more than 130 countries.”

C-Tests (Klein-Braley, 1981; Raatz & Klein-Braley 1981) are, however, among the few alternative testing methods which differ from the proficiency measures such as the TOEFL in terms of their underlying construct validity. As a type of cloze tests, they are said to be based on reduced redundancy theory which approaches proficiency in a given language as the ability to understand a distorted message by formulating valid guesses about a certain percentage of omitted elements (Spolsky, 1973). C-Tests present test takers with some carefully chosen short texts in which the second half of every second word is removed from its second sentence onwards so that they can restore the mutilated part by activating their learned proficiency of a foreign language.

In 1997 Klein-Braley administered C-Tests with 1) a security language proficiency test called DELTA, 2) two cloze elides (Manning, 1986) requiring test takers to find some extra words inserted in the texts intentionally, 3) two cloze multiple choice item tests and 4) a dictation test to a large sample of university applicants and applied factor analysis to her data. Since three out of the four C-Tests loaded higher than .70 on the ***unrotated*** first factor she concluded that “the best test to select to represent general language proficiency as assessed by reduced redundancy testing would be the C-Test” (p. 71).

Khodadady (1997) designed and validated another type of cloze multiple choice item tests (MCITs) and called them schema-based cloze MCITs [henceforth S-Tests (Khodadady, 2012)]. These tests are developed on the assumption that ***the words comprising written texts employed in language tests are schemata*** whose understanding by themselves and in relation to other schemata comprising the texts under comprehension depends on test takers’ personally acquired background knowledge of the concepts they represent. Since these schemata are continuously met in everyday life in whatever forms and modes possible, they go through constant modification.

The schemata comprising written texts are classified into the three main domains of semantic, syntactic and parasyntactic, which are in turn broken into their constituting genera, species and types. The *semantic* schema domain, for example, consists of adjectives, adverbs, nouns and verbs as its genera. Similarly, the genus of adjectives forming the semantic domain includes agentive, comparative, complex, dative, derivational, nominal, simple and superlative adjectives as its species. In a hierarchical fashion, the agentive adjectives, as a species of adjective genus, consist of schema *types* such as interesting and fascinating used in describing the noun genus of semantic domain. Khodadady, Pishghadam and Fakhar (2010) employed these domains, genera, species and types to establish the content validity of their achievement tests and then employed them to explore the relationship among grammar, vocabulary and reading comprehension ability. [Interested readers may consult Khodadady (2008, 2013) for the analysis and classification of schemata comprising texts.]

In addition to treating all the *words* comprising the texts as *schemata*, the choices selected in developing and taking S-Tests are viewed as schemata if they bear syntactic, semantic and discoursal relationships with the deleted schemata given as keyed responses. It is hypothesized that the relationships present in the alternatives of the S-Tests provide the readers with concepts which *compete* with the keyed response in terms of their syntactic, semantic and discoursal appropriateness and thus distinguish them from their traditional counterparts called distracters.

The traditional cloze multiple choice item below was, for example, developed by Hale, Stansfield, Rock, Hicks, Butler and Oller (1988). As can be seen, distracter A, *inspecting,* is syntactically and semantically different from the keyed response *contrast*. Similarly, distracter B, *knowledge*, lacks semantic relationship with the keyed response. Distracter D, *medicine*, is discoursally related with some of the schemata forming the sentence such as *chest* and *lung* but has no semantic relationship with the keyed response.

By ….. (46) conventional x-rays generally differentiate only between bone and air, as in pictures of the chest and lungs.

46 A. inspecting B. knowledge C. contrast* D. medicine

In contrast to traditional cloze multiple choice items, S-Test items offer their takers three competitives which bear syntactic and semantic relationships with the keyed response and discoursal relationships with the schemata comprising the text as described in the item below developed by Khodadady (2004). In order to answer the item, test takers must know what the four choices mean individually. They should then focus on all the words, or schemata, used in the sentence in order to decide which alternative fits the blank best. The keyed response, i.e., *attack*, 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 syntactically related, the alternatives share the semantic feature of *assault* and must therefore be equally attractive to test takers.

Fears over access to medical records

Privacy campaigners in the US have launched a fierce ... (1) on a bill that they believe will expose medical records to too many prying eyes.

1 A. raid B. slander C. attack* D. ambush

However, in order for test takers to choose the keyed response *attack*, 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 schemata preceding and following the deleted schema in order to make an informed choice.

Khodadady (2004) administered an S-Test with the C-Tests (Klein-Braley, 1977), text-driven cloze test (Farhady & Keramati, 1994) and traditional cloze MCIT (Hale *et al* 1988) to 34 senior undergraduate Iranian students. He also administered the disclosed TOEFL test 1 (Educational Testing Service, 1991, pp. 75-100) as an internationally accepted measure of English language proficiency to explore the construct validity of C-Tests.

Similar to Klein-Braley’s (1997) findings, the C-Tests and C-test 2 had the highest loadings on the first factor, i.e., 0.93 and 0.78, respectively. These loadings could not, however, show language proficiency as Klein-Braley claimed because the TOEFL had the second lowest loading on this factor, i.e., 0.69. Furthermore, the TOEFL loaded on the second factor on which the C-test and its subtests all had negative loadings. Due to these unexpected loadings, Khodadady (2004) ran a rotated factor analysis on the data.

When the loadings were rotated, the TOEFL test did not load on the first factor any more. Only the C-Tests (0.95) and their subtests, i.e., C-Test 1 (0.72), C-Test 2 (0.80), C-Test 3 (0.67) and C-Test 4 (0.80) had the highest loadings on this factor. Khodadady (2004) concluded that since the TOEFL differs from the four methods of reduced redundancy in terms of its construction, it does not load on the first factor and therefore C-Tests have their own unique effect on the loadings. The TOEFL had, however, the highest loading on the second factor (0.90), upon which the S-Test, text-driven cloze test, traditional cloze MCIT and even C-Tests loaded, indicating that the second factor represents English language proficiency.

In order to find out what the nature of the first factor was, Khodadady (2007) developed two decontextualised and spelling tests on the C-Tests developed by Klein-Braley (1997) and administered them along with the TOEFL and two vocabulary tests to 63 senior undergraduate Iranian students majoring in English language and literature at Ferdowsi University of Mashhad. (The tests will be described in details in the instrumentation section).

Table 1 presents the varimax rotated factor matrix using principal component analysis of the C-Tests, decontextualised C-Test, matching vocabulary test, spelling test, disclosed TOEFL and its subtests. Based on the results shown in the table, Khodadady (2007) concluded that "since neither the C-Tests nor the decontextualised and spelling tests load on the first factor, they must measure method-specific abilities."

TABLE 1
VARIMAX WITH KAISER ROTATED FACTOR MATRIX

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	1.48	1.44
Variance Explained:	51.93%	11.39%	11.10 %

* Loadings less than .30

As will be discussed below, the tests employed by Khodadady (2007) were distinctly different from each other. And there was no other measure of language proficiency based on reduced redundancy to find out whether it would load with the C-Tests on the second factor. Furthermore, as can be seen in Table 1, the decontextualised C-Test and spelling tests loaded on the third factor rather unexpectedly. The present study was therefore designed to find out whether three rotated factors would be extracted if an S-Test was added to the list of the tests and whether the C-Tests and decontextualised C-test and spelling test would still load on two separate factors if all the tests were administered to a larger and more representative sample. These objectives were explored via the following research questions.

Research Questions

1. What is the factor structure for the C-Tests, the Decontextualised C-Test and the Spelling Test? Do they load on two factors as they did in Khodadady's (2007) study?

2. What is the factor structure if the TOEFL, the lexical knowledge test and S-Test are included? Do the C-Tests and S-Test load on the same factor?

II. METHOD

A. Participants

Four hundred and sixteen university students took part in the study in the course of two academic semesters in 2009. However, the scores of 402 participants were analyzed because 14 of them missed one or more of the tests for reasons beyond the researcher's control. Out of 402 participants, 327 (81.3%) were studying English Language and Literature, 86 (16.4%) Teaching English as a Foreign Language, and nine (2.2%) English Translation at undergraduate (n = 327) and graduate levels (n= 75) at Ferdowsi University of Mashhad in Iran. The majority of participants were female, i.e., 301 (75%), and only one fourth were male, i.e., 101 (25%). They had enrolled as freshman (147), sophomore (49), junior (93) and senior (113) full time students when the project started and conducted in the period specified. The participants' age ranged between 18 and 45 (Mean = 22.92, SD = 3.21) and they spoke Persian (395) and Turkish (7) as their mother languages. They took all the tests voluntarily, however, in order for the participants to take the tests seriously and be rewarded for their time and participation, it was announced that the researcher will add 10% to their final score in whatever courses they took with him.

B. Instruments

Six tests along with their various subtests were employed in this study, i.e., C-Tests consisting of four texts, a decontextualised C-Test, a spelling test, the disclosed TOEFL test consisting of structure, written expressions, sentential vocabulary and traditional multiple choice item reading comprehension test subtests, a lexical knowledge test, and a semantic schema-based cloze multiple choice item test.

C-Tests. The C-Tests developed by Klein-Braley (1997, pp. 79-80) were used in this study. It consisted of 99 items developed on four texts. Since they are different in content, each text is considered as a C-Test in its own right and specified as C-Test 1, C-Test2, C-Test 3 and C-Test 4. 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. The alpha reliability coefficient obtained by Khodadady (2007) for C-Tests was .89 and they correlated significantly with the TOEFL, i.e., r = .62, *p* < .01.

Decontextualised C-Test. This test was developed by Khodadady (2007). He took the 99 mutilated words comprising C-Test 1, C-Test 2, C-Test 3, and C-Test 4 out of their linguistic context, numbered them from 1 to 99 and presented them to 63 senior undergraduate students of English. The participants were instructed to restore the mutilated half of the words by considering the number of letters given. They could add the same number of letters given or one more. If, for example, there was one letter such as h__, they could add one or two letters to produce the English words h**e** and h**is** as appropriate responses. They were also told that only words having acceptable spelling will be scored correct. The restored words were then scored twice. As a decontextualised C-Test, only the exact words comprising the texts of the C-Tests were scored correct. For example, item six on the decontextualised C-Test requires the participants to supply the letters ***e*** and ***w*** to restore the mutilated word ***few*** as the exact answer. The reported alpha reliability coefficient for this test was .60 and it correlated significantly only with the written expressions section of the TOEFL, i.e. r = .62, *p* < .05.

Spelling Test. The researcher scored the restored mutilated words on the decontextualised C-Test for the second time by accepting whatever words the participants produced on the basis of the directions given, i.e., adding the same number of letters given or one more. For example, all the restored words for item six, i.e., *far*, *fat*, *few*, *fit*, *fix*, *for*, *fun* and *fur* were scored correct. In order to ensure the validity of scoring, various references such as *Collins Dictionary of the English Language* (Hanks, 1986) were consulted. Khodadady (2007) argued that since this method measures the test takers' knowledge of letters comprising English words irrespective of their meaning and context, it is a test of spelling knowledge. The reported alpha reliability coefficient for this test was .89 and it correlated significantly with the TOEFL, i.e. r = .37, *p* < .01.

Structure Test. Following Khodadady and Herriman (2000), the structure subtest of the disclosed written TOEFL test (Educational Testing Service, 1991) was employed in order to measure the participants' structure competence and its relationship with C-Tests. It comprises of 30 cloze multiple choice items developed on 30 isolated and unrelated sentences addressing a discrete grammatical point. The structure test had an alpha reliability coefficient of 0.89 and correlated significantly with the C-Tests, i.e., r = .55, *p* < .01.

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 underlined part of sentences. It had an alpha reliability coefficient of 0.74 and correlated significantly with the C-Tests, i.e., r = .49, *p* < .01.

Sentential 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' *sentential* vocabulary knowledge. It is labeled *sentential* in the present study because each item is developed on a single underlined word in an *isolated sentence* which has no thematic relationship with the other sentences comprising the sub test. From among the four alternatives given below each item, test takers must choose the keyed response replaceable with the underlined word. It had the alpha reliability coefficient of 0.79 and correlated significantly with the C-Tests, i.e., r = .53, *p* < .01.

Lexical Knowledge Test. Mirêlis (2004) defined mental lexicon as "the collection of words one speaker knows and the relationships between them" (p. 2). Since the matching vocabulary test designed by Nation (see Schmitt, Schmitt, & Clapham, 2001) is a collection 160 words, it was employed to measure the participants' lexical knowledge. Although the original lexical knowledge test (LKT) consisted of 60 items presented in 20 groups of three words having six words opposite to be selected by writing the number of question in front of the words given in each group, its format was changed to 6-choice items in the present study to save space and sheets and do away with writing.

In the new format, the three key words are numbered and presented in a single box. Six other words marked A, B, C, D, E and F are given in front of the three numbered words as shown below. The participants were instructed to select the choice which best fitted the meaning of each word and mark their choice on the answer sheet by filling in the corresponding box. For example, in the box below, the participants had to select choice **C** as the best meaning for the word number 1 and fill box C for item 1 on their answer sheet. The LKT proved to be a highly reliable test, i.e., α = .89, in Khodadady's (2007) study and correlated significantly with the C-Tests, i.e., r = .42, *p* < .01.

Example

1. Assert	**A**	cast	**D**	detest
2. Ban	**B**	confide	**E**	falter
3. Throw away	**C**	state	**F**	forbid

Traditional Multiple Choice Reading Comprehension Test. The multiple-choice reading comprehension subtest of the disclosed written TOEFL test (Educational Testing Service, 1991) was chosen as a *traditional* reading comprehension test (RCT) because it was developed by testing specialists who used their expertise and intuition to write some 30 items on five passages (Khodadady 1997, 1999; Khodadady & Herriman 2000). Both the stem and the choices of the items were developed by the ETS specialists and participants had to read and understand the passages so that they could choose the best alternative. Khodadady (2007) reported the alpha reliability coefficient of 0.79 for the traditional RCT correlating significantly with the C-Tests, i.e., r = .49, $p < .01$.

S-Test. The 60-item semantic S-Test developed by Gholami (2006) on the adjectives, adverbs, nouns and verbs of an authentic text was used in order to find out whether it would show strong relationships with the C-Tests and load with them on the same factor because both are developed on the schemata comprising texts. Gholami developed the S-Test on the article *why don't we just kiss and make up* (Dugatkin, 2005) published in *NewScientist* magazine having passages "more academic than articles in quality newspapers" (Clapham 1996, p.145). The reported alpha reliability coefficient for the S-Test was 0.64 and it correlated significantly with the TOEFL, i.e., r = .84, $p < .01$.

C. Procedure

With the exception of the decontextualised C-Test which was given as the first test, the other eight tests were combined and counterbalanced in four other sessions of administration. The former was given first because the administration of C-Tests would have disclosed its items and thus invalidated its decontextualised version. After the administration of the first test in the first session, the participants were seated on every other chair in the other four sessions and each received one of the tests different from the one given to the participant sitting nearby. Structural, written expressions and traditional reading comprehension tests were administered as a single test in one session as were the sentential vocabulary test and lexical knowledge tests in another. The C-Tests and S-Test were administered in two separate sessions and thus the whole project required five sessions in all. Since the participants wrote their names on the answer sheets, the researcher could easily check what test was given before and which had to be given next.

D. Statistical Analysis

After having the SPSS version 16.0 calculate the descriptive statistics of the tests administered in this study, their internal consistency reliability was estimated via Cronbach's alpha. The difficulty level and the discrimination power of the tests were 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 (Baker, 1989) and the r_{pbi} coefficients were estimated by correlating each individual item with the total test score. Finally factor analysis was run to determine what latent variables the tests administered in the study would load on.

III. Results

Table 2 presents the descriptive statistics of the tests administered in this study. As can be seen, with the exception of the decontextualised C-Tests, all other tests enjoyed high reliability levels. The highest and lowest reliable measures of language proficiency were lexical knowledge test, i.e., 0.93, and the decontextualised C-Tests, i.e., 0.46, respectively. The low reliability of the latter was expected because the restoration of its items did not depend on any specific language ability other than the test takers' personal and random selection of certain words from among all the possible words they were familiar with on the basis of their constituting letters fitting the directions given.

TABLE 2
DESCRIPTIVE STATISTICS OF THE TESTS AND THEIR SUBTESTS ADMINISTERED IN THE STUDY

Tests	# of item	Mean	Std. Deviation	Kurtosis	Mean *p*-value	Mean r_{pbi}	α
C-Tests	99	54.85	13.397	-.062	.55	.32	.91
C-Test 1	25	15.25	4.183	-.365	.61	.33	.77
C-Test 2	24	12.42	3.903	-.243	.52	.30	.72
C-Test 3	25	14.50	4.223	-.251	.58	.35	.78
C-Test 4	25	12.67	3.508	-.085	.51	.29	.70
Spelling	99	85.44	9.501	4.840	.86	.30	.91
Decontextualised C-Tests	99	18.75	4.878	3.430	.19	.13	.46
TOEFL	115	77.90	14.380	.458	.68	.29	.91
Structure	30	22.57	4.311	2.478	.75	.32	.79
Written Expressions	25	17.29	4.671	.328	.69	.37	.82
Sentential Vocabulary	30	20.09	5.170	.512	.66	.14	.82
Reading Comprehension	30	17.94	5.912	-.561	.60	.36	.86
S-Test	60	28.62	8.661	.061	.47	.31	.85
Lexical Knowledge Test	60	25.94	12.581	-.757	.43	.45	.93

Based on the mean *p*-values presented in Table 1, the most difficult and easiest tests were the spelling (.89) and Decontextualised C-Test (.19), respectively. The very difficulty of the latter and its being based on the participants' personal preference to restore the mutilated words only on the basis of the directions given has made the Decontextualised C-Test the most difficult and the least reliable. However, it is a valid measure of language ability because its mean r_{pbi} (.13), i.e., discriminatory power, is almost the same as the sentential vocabulary test (.14) and correlates significantly with all the tests employed in this study.

Table 3 presents the correlation coefficients among the 14 tests and their sub tests. As can be seen, they all correlate significantly with each other. With the exception of the spelling test, the decontextualised C-Test shows the highest and lowest significant correlations with the C-Tests, i.e., .32, and S-Test, i.e., .18, respectively. Similarly, the spelling test has the highest and lowest significant correlations with the C-Tests, i.e., 0.53, and S-Test, i.e., 0.31, indicating that among the TOEFL, C-Tests and Lexical Knowledge test, the S-Test is the least related to the directions given and the spelling proficiency of test takers.

TABLE 3
CORRELATION COEFFICIENTS OBTAINED AMONG THE TESTS

	TOEFL	**Str**	**WExp**	**SVK**	**Read**	**Ctest**	**CT1**	**CT2**	**CT3**	**CT4**	**Spell**	**DCt**	**S-Test**
TOEFL	1	.855**	.872**	.755**	.852**	.679**	.627**	.504**	.558**	.610**	.402**	.216**	.580**
Str	.855**	1	.772**	.490**	.651**	.595**	.532**	.462**	.507**	.515**	.335**	.168**	.417**
Wexp	.872**	.772**	1	.567**	.654**	.611**	.565**	.454**	.503**	.549**	.382**	.193**	.503**
SVT	.755**	.490**	.567**	1	.507**	.524**	.505**	.352**	.415**	.507**	.309**	.134**	.550**
Read	.852**	.651**	.654**	.507**	1	.534**	.482**	.400**	.454**	.471**	.288**	.211**	.461**
Ctest	.679**	.595**	.611**	.524**	.534**	1	.860**	.831**	.866**	.826**	.526**	.318**	.524**
CT1	.627**	.532**	.565**	.505**	.482**	.860**	1	.620**	.627**	.642**	.458**	.268**	.485**
CT2	.504**	.462**	.454**	.352**	.400**	.831**	.620**	1	.638**	.553**	.446**	.291**	.376**
CT3	.558**	.507**	.503**	.415**	.454**	.866**	.627**	.638**	1	.645**	.401**	.262**	.398**
CT4	.610**	.515**	.549**	.507**	.471**	.826**	.642**	.553**	.645**	1	.483**	.257**	.525**
Spell	.402**	.335**	.382**	.309**	.288**	.526**	.458**	.446**	.401**	.483**	1	.492**	.312**
DCt	.216**	.168**	.193**	.134**	.211**	.318**	.268**	.291**	.262**	.257**	.492**	1	.177**
S-Test	.580**	.417**	.503**	.550**	.461**	.524**	.485**	.376**	.398**	.525**	.312**	.177**	1
LKT	.569**	.412**	.472**	.607**	.431**	.451**	.438**	.329**	.308**	.465**	.244**	.105*	.579**

* All correlations are significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

Table 4 presents the unrotated and rotated components extracted via Principal Component Analysis. As can be seen, the C-Tests, Spelling test and decontextualised C-Test load on two components and thus confirm the results obtained by Khodadady (2007). The most interesting result of the present study is, however, the fact that when the components are rotated, the decontextualised C-Test loads the highest on the second component only and thus illuminates its nature. Since the test takers could get the exact responses of the C-Tests by following the directions given and without having any context to guide them, it reveals how direction-specific the C-Tests are. This finding, therefore, challenges Eckes and Grotjahn's (2006) claim regarding the nature of what C-tests measure. The success on C-tests partly depends on test takers' ability to comply with their directions.

TABLE 4
UNROTATED AND ROTATED COMPONENTS EXTRACTED VIA PRINCIPAL COMPONENT ANALYSIS FROM THE C-TESTS AND THE TWO TESTS DEVELOPED ON ITS ITEMS

Tests	Unrotated Components		Rotated Components**	
	1	2	1	2
C-Tests	.981	*	.971	*
C-Test 1	.841	*	.831	*
C-Test 2	.818	*	.793	*
C-Test 3	.837	*	.851	*
C-Test 4	.826	*	.809	*
Spelling Test	.661	.514	.391	.741
Decontextualised C-test	.455	.795	*	.912
Eigenvalue:	4.37	1.03	3.80	1.60
Variance Explained:	62.36%	14.76%	54.3%	22.81%

* Loadings less than .30
** Varimax with Kaiser Normalization

In addition to establishing C-Tests as direction specific, the results presented in Table 3 reveal another latent feature of C-Tests. As can be seen, even rotating the components does not bring about any drastic changes in the loadings of the spelling test as it does with the decontextualised C-Test. Since it loads on the first factor as well even after being rotated as the C-Tests do, it can be concluded that C-Tests are spelling specific as well.

Table 5 presents the unrotated and rotated components extracted via Principal Component Analysis from the tests and their subtests administered in this study. As can be seen, even when the S-Test, the TOEFL, and Lexical Knowledge Test are included in the Principal Component Analysis and the extracted components are rotated, two components emerge. These results thus disconfirm the existence of the third factor found by Khodadady (2007).

TABLE 5
UNROTATED AND ROTATED COMPONENTS EXTRACTED VIA PRINCIPAL COMPONENT ANALYSIS FROM THE TESTS AND THEIR SUBTESTS

Tests	**Unrotated Components**		**Rotated Components****	
	1	2	1	2
C-Tests	.906	.318	.513	.812
C-Test 1	.805	*	.508	.656
C-Test 2	.714	.385	.321	.745
C-Test 3	.754	.311	.398	.712
C-Test 4	.797	*	.516	.632
Spelling Test	.568	.453	*	.708
Decontextualised C-test	.350	.521	*	.626
S-Test	.673	*	.655	*
TOEFL	.905	-.335	.917	*
Structure	.774	*	.761	*
Written Expressions	.810	*	.801	*
Sentential Voc. Test	.715	-.342	.773	*
Reading Comprehension	.736	-.305	.767	*
Lexical Knowledge Test	.626	-.347	.706	*
Eigenvalue:	7.599	1.513	5.253	3.859
Variance Explained:	54.281	10.804	37.523	27.562

* Loadings less than .30
** Varimax with Kaiser Normalization

The results presented in Table 5 above also emphasize the distortion of latent variables when they are presented as unrotated components. As can be seen, all the tests and their subtests load acceptably on the first unrotated component, giving the rather queer impression that a spelling test based on some 99 words taken out of their context is as much a measure of language proficiency as a lexical knowledge test consisting of 160 semantic words simply because they load 0.57 and 0.63 on the first component, respectively. However, the high and acceptable loading of 0.57 disappears as soon as the components are rotated. It is therefore suggested that the findings reported on the basis of unrotated components be treated cautiously because they are distorted at their face value and logically fallacious on a theoretical basis.

IV. DISCUSSION AND CONCLUSION

Klein-Braley (1996) believed the question, "what exactly do C-tests measure in terms of language processing?" (p. 24) was *irrelevant* because C-Tests do function as measures of language proficiency. The results of the present study support her belief by revealing significant relationships between C-tests and the TOEFL (r = .68, p < .01). This relationship was further emphasized by Babaii and Ansari (2001) who found correlations as high as .88 between the two and thus provided statistical evidence to treat the C-tests and TOEFL as interchangeable measures of language proficiency.

In contrast to the TOEFL, C-Tests, however, show the highest significant correlations with the words whose restoration is solely based on the number of letters specified by their directions (r = .53, p < .01), indicating that 28 percent of test takers' scores on C-Tests can be explained just by their spelling knowledge. Similarly, the significant relationship between C-Tests and their decontextualised version (r = .32, p < .01) reveals the fact that ten percent of variance in scores on C-Tests is context-independent. These findings question the validity of employing retrospective verbal protocols to study test takers' performance on C-Tests (e.g., Babaii and Ansari, 2001).

Based on Ericson and Simon's (1984) understanding of verbal protocols (VP) as the direct verbalizations of specific cognitive processes, Babaii and Ansari (2001) asked their participants to verbalize how they completed the task after they took C-Tests. Based on the 6.1% of participants' retrospective VPs, they concluded that top-down cues are used to exploit the relationship among contextual words such as *police*, *theft*, *gang*, …, to restore 'missing' in (the mis___vehicles) by moving back and forth all through a text. The conclusion seems to be questionable on two grounds. First, in their primer on VPs Trickett and Trafton (2009) cited Nisbett and Wilson (1977) who argued that test takers do not necessarily have access to what they did or why they did it after they complete a given task. Babaii and Ansari's few participants, therefore, might have not based the restoration of the second half of the word *missing* on the specified contextual words but claimed to have done so in order to be treated more academically.

Secondly, the findings of the present study show that the majority of Babaii and Ansari's (2001) participants who restored the mutilated part of the word *missing* correctly might have done so by exploiting the number of letters constituting the word only. The responses of 402 participants on the decontextualised C-Test administered in this study, for example, showed that 367 (91%) and 271 (67%) restored the words *mixture* and *matter*, respectively, without having any top-down cues. In other words neither the immediate context of the words, i.e., (a mix___ of) and (important mat___, between) nor the words comprising other parts of the paragraph in which they appear were available to provide top-down cues for the respondents as Babaii and Ansari claimed.

It is argued in the present paper that the application of top-down cues are best captured in the performance of test takers on the disclosed TOEFL, S-Test and Lexical Knowledge Test (LKT) administered in this study because these measures of language proficiency neither depend on their directions as the C-Tests do nor load acceptably on the second rotated factor upon which only C-Tests and their decontextualised version as well as the spelling test do. If the C-Tests do require top-down processing they must load acceptably on the rotated factor upon which these tests load as well.

The factor analysis of the C-Tests along with the disclosed TOEFL, S-Test and LKT, however, revealed two components as the latent variables underlying these tests. While the unrotated components yielded logically unsound high loadings on the first factor due to initial extraction, the second unrotated factor revealed negative loadings not only on the TOEFL but also on the other two tests taken along with the C-Tests implying that whatever C-Tests measure, the other tests measure in the opposite direction.

The rotation of components, nonetheless, reveals the fact that C-Tests measure language proficiency as the TOEFL, S-Test and LKT do. But as a measure of language proficiency, C-Tests do not load on the first rotated component as highly as the other three measures do. Instead they reveal their dependency on the test takers' spelling proficiency and understanding of their directions by having their highest loadings on the second rotated component. Interestingly enough, neither the TOEFL nor S-Test and LKT had any acceptable loadings on the second rotated component, implying that the very dependence of C-Tests on the spelling and its directions makes it a unique language proficiency test.

Further research is therefore needed to specify the nature of spelling proficiency as a part of general factor contributing to the proficiency tests employed in this study. Future studies must, for example, show whether changing LKT into a spelling test by following the directions given in the C-Tests will bring about their loading on the second rotated component or not. Replicating the study with a different language proficiency test such as the International English Language Testing System may also shed some light on the nature of C-Tests because it requires a limited amount of writing in its being answered as the C-Tests do.

Furthermore, the findings of the present study highlight the importance of spelling and thus call for further research in terms of designing and including spelling tests as part of language proficiency measures. Although the Spelling Test employed in this study does not load on the first *rotated* component, it does contribute to whatever the C-Tests measure and thus load on both the first and second unrotated factors as the other measures employed in this study do. The inclusion of a more comprehensive spelling test in future research projects must indicate whether accommodating the spelling variable in language proficiency tests such as the International English Language System stands to fairness when it penalizes test takers for misspelling in their written responses.

ACKNOWLEDGEMENT

This study was funded by Ferdowsi University of Mashhad, i.e., Grant coded 2/15117 and dated June 22, 2010.

REFERENCES

[1] Austin, J. L. (1961). How to do things with words. Cambridge, Massachusetts: Harvard University.

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

[3] Belgar, D., Hunt, A. (2002). Implementing task-based langauge teaching. In J. C., Richards & Renandya, W. A. (Eds.). *Methodology in language teaching: An anthology of current practice* (pp. 96-106). Cambridge: CUP.

[4] Brown, H. D. (2002). English language teaching in the "Post-Method" era: Toward better diagnosis, treatment, and assessment. In J. C., Richards & Renandya, W. A. (Eds.). *Methodology in language teaching: An anthology of current practice* (pp. 9-18). Cambridge: CUP.

[5] Clapham, C. (1996). The development of IELTS: A study of the effect of background knowledge on reading comprehension. Cambridge: Cambridge University Press.

[6] Dugatkin, L. (May 07, 2005). Why don't we just kiss and make up? NewScientist 2498, p. 35. Retrieved May23, 2005, from http://www. newscientist.com/channel/life/mg18624981.300.

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

[8] Educational Testing Service. (1991). Reading for TOEFL. Princeton, NJ: ETS.

[9] Educational Testing Service. (2010). TOEFL. Accessed November 12, 2010 at http://www.ets.org/toefl.

[10] Ericson, K., & Simon, H. (1984). Protocol Analysis: Verbal Reports as Data. Cambridge: CUP.

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

[12] Farhady, H. (1980). Justification, development, and validation of functional language tests. Unpublished PhD dissertation, UCLA.

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

[14] Gholami, M. (2006). The effect of content schema type on Iranian test takers' performance. Unpublished MA thesis, Ferdowsi University of Mashhad, Iran.

[15] 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.

[16] Halliday, M. A. K. (1973). Explorations in the functions of language. London: Edward Arnold.

[17] Hanks, P. (Ed.). (1986). Collins Dictionary of the English Language (2^{nd} ed.). London: Collins.

[18] Kelly, R. (1978). On the construct validation of comprehension tests: an exercise in applied linguistics. Unpublished PhD thesis, University of Queensland.

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

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

[21] Khodadady, E. (2007). C-Tests method specific measures of language proficiency. *Iranian Journal of Applied Linguistics* (IJAL), 10(2), 1-26.

[22] Khodadady, E. (2008). Schema-based textual analysis of domain-controlled authentic texts. *Iranian Journal of Language Studies* (IJLS), 2(4), 431-448.

[23] Khodadady, E. (2012). Validity and tests developed on reduced redundancy, language components and schema theory. *Theory and Practice in Language Studies*, 2(3), 585-595.

[24] Khodadady, E. (2013). Research Principles and Methods and Statistics in Applied Linguistics. Mashhad: Hamsayeh Aftab.

[25] Khodadady, E. Pishghadam, R., & Fakhar, M. (2010). The relationship among reading comprehension ability, grammar and vocabulary knowledge: An experimental and schema-based approach. *Iranian EFL Journal*, 6(2), 7-49.

[26] 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.

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

[28] Klein-Braley, C. (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.

[29] Klein-Braley, C. (1996). Towards a theory of C-Test processing. In R. Grotjahn (Ed.). *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (23-94). Bochum: Brockmeyer.

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

[31] Klein-Braley, C., & Raatz, U. (1985). Fremdsprachen und Hochschule 13/14: Thematischer Teil: C-Tests in der *Praxis*. Bochum: AKS.

[32] 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.

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

[34] Mirâlis, M. T. M. (2004). Exploring the Adaptive Structure of the Mental Lexicon. Unpublished PhD dissertation, University of Edinburgh, England. Retrieved November 10, 2010 from http://www.ling.ed.ac.uk/~monica/tamariz_thesis.pdf.

[35] Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 231–259.

[36] 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.

[37] Searle, J. R. (1969). Speech acts: An essay in the philosophy of language. Cambridge: Cambridge University Press.

[38] Sinclair, J. M., & Coulthard, R. M. (1975). Towards an analysis of discourse: The English used by teachers and pupils. London: Oxford.

[39] 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.

[40] Trickett, S. B., & Trafton, J. G. (2009). A Primer on Verbal Protocol Analysis. In D. Schmorrow, J., Cohn, & D. Nicholson (Eds.). *The PSI Handbook of Virtual Environ- ments for Training and Education, Volume 1* (pp. 332-346). Westport, CT: Praeger Security International.

[41] van Ek, J. A. (1976). The threshold level for modern language teaching in the schools. London: Longman.

[42] Wilkins, D. A. (1976). Notional syllabuses. London: Oxford.

Ebrahim Khodadady was born in Iran in 1958. He obtained his PhD in Applied Linguistics from the University of Western Australia in 1998. He holds TESL Ontario and Canadian Language Benchmarks Placement Test (CLPBPT) certificates and has taught English as a first, second and foreign language to high school and university students in Australia, Canada and Iran.

Dr. Khodadady is currently an academic member of English Language and Literature Department at Ferdowsi University of Mashhad, Iran. He was invited as a VIP by Brock University in Canada in 2004 and served as the Associate Director of Assessment Center at George Brown College in Toronto for almost a year. His published books are *Multiple-Choice Items in Testing: Practice and Theory* (Tehran, Rahnama, 1999), *Reading Media Texts: Iran-America Relations* (Sanandaj, Kurdistan University, 1999), *English Language Proficiency Course: First Steps* (Sanandaj, Kurdistan University, 2001) and *Research Principles and Methods and Statistics in Applied Linguistics* (Mashhad, Hamsayeh Aftab, 2013). His main research interests are Language Learning and Teaching, Psycholinguistics, Sociolinguistics, Testing and Translation.