ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 2, No. 3, pp. 585-595, March 2012
© 2012 ACADEMY PUBLISHER Manufactured in Finland.
doi:10.4304/tpls.2.3.585-595

Validity and Tests Developed on Reduced Redundancy, Language Components and Schema Theory

Ebrahim Khodadady
Ferdowsi University of Mashhad, Iran
Email: ekhodadady@ferdowsi.um.ac.ir

***Abstract*—This study reports the performance of 430 undergraduate and graduate students of English on standard C-Tests, the disclosed TOEFL, lexical knowledge test (LKT) and semantic schema-based cloze multiple choice item test (S-Test). Among the four tests, the first has been shown to be affected by their context. Two other versions were, therefore, developed on the standard C-Tests, i.e., Context Independent (CI)C-Test, and Context Dependent (CD)C-Tests from which the CIC-Test items were removed. Inspired by Kamimoto (1993), the well functioning items of standard C-Tests developed by Klein-Braley (1997) and their CD and CI versions were employed to study their underlying factors. The factorial analysis of the three C-Tests revealed two latent variables representing reduced redundancy and another unknown construct. The inclusion of the TOEFL, LKT and S-Test in the analysis to illuminate the nature of the second factor, however, resulted in the extraction of three latent variables. Since the standard C-Tests and CDC-Tests loaded the highest on the first factor along with other tests, it represented reduced redundancy. However, since the S-Test had its highest loading on the second factor upon which the other two measures of language proficiency loaded as well, it represented schema theory. The vocabulary subtest of the TOEFL loaded the highest on the third factor upon which the TOEFL itself and its structure subtest loaded as well revealing the componential nature of the TOEFL. The results are discussed and suggestions are made for future research.**

***Index Terms*—testing, language proficiency, language components, schema theory, reduced redundancy**

I. Introduction

The development and validity of written language proficiency measures such as the Test of English as a Foreign Language (TOEFL) depends on the theory upon which proficiency is defined. This study divides them into three distinct categories, i.e., componential, reduced redundancy and schema and addresses their underlying rationale empirically and factorially by administering their representative measures to 430 learners of English in Iran. While the first category is best represented by the TOEFL accommodating the structure, written expressions, vocabulary and reading components of language, the second and third are claimed to be materialized in standard C-Tests and schema-based cloze multiple choice item tests (henceforth S-Tests), respectively.

Upon reviewing the literature related to the theories of language testing, the present author developed context independent (CI) C-Tests by removing the items of standard C-Tests from their context and administering it as a single test. The participants’ responses on the CIC-Test were then *tailored* by keeping the items which had functioned well, i.e., had acceptable item facility and discrimination indices. Since the standard C-Tests shared some well functioning items with the CIC-Test, these items were removed from the standard C-Tests to develop the context dependent (CD) C-Tests.

The scores on the tailored standard C-Test, the CIC-Test and CDC-Tests were then subjected to principal component analysis (PCA) to determine what constructs underlie C-Tests as measures of reduced redundancy in language proficiency. To further illuminate the nature of latent variables extracted from the three versions of C-Tests, the tailored TOEFL, S-Test as well as the LKT were added to the PCA analysis to find out whether their addition would result in the extraction of two other latent variables representing language components and schema theory. The measures were also correlated with each other to have a better understanding of the three latent variables extracted in the study.

A. *Proficiency Tests Based on Language Components*

Internationally recognized language tests such as the TOEFL treat written proficiency as a psychological construct consisting of four components or abilities, i.e., structure, written expressions, vocabulary and reading comprehension (Khodadady, 1997, 1999). The more recent versions of the TOEFL, however, have reduced the components to two, i.e., 1) structure which includes the written expressions and 2) reading comprehension which includes some items developed on short passages as wholes and some vocabulary items constructed on the words constituting the same passages (see ETS, 2003). It is assumed that the proficiency of English language learners can best be measured by the responses they give to traditionally designed multiple choice items addressing each of these four components separately. An earlier and disclosed TOEFL test along with its subtests measuring the four components have, therefore, been used in the present study as a representative measure of language components.

B. Proficiency Tests Based on Reduced Redundancy

Reduced redundancy measures of language proficiency derive their theory from closure in Gestalt psychology highlighting the tendency of the human brain to complete figures even when *part of* the information is *missing*. Taylor (1953) employed the psychology to develop *cloze tests* by deleting a number of words in given texts in order to estimate the readability level of the texts. He hypothesized that if readers have achieved reading comprehension ability at a given grade, they must be able to activate their acquired ability successfully to restore the words deleted from the texts taught at that grade. Cloze tests were thus employed originally as measures of readability.

Spolsky (1973), however, resorted to closure to formulate his reduced redundancy theory of language proficiency. He seems to have argued that since a given function such as *past tense* may be brought up in *various* linguistic devices, i.e., “words,” (Bardovi-Harlig & Comajoan, 2008, p. 391) such as verbs and adverbs within a given text, proficient language users can restore a missing past tense verb or adverb by focusing on other verbs or adverbs which carry the same function in the same text. The very materialization of the same function in *various* words comprising texts establishes the link between cloze tests and reduced redundancy theory. According to Bardovi-Harlig and Comajoan (2008), however, the linguistic devices have varying *functional loads*.

For example, if an adverb such as *yesterday* is the only indicator that an event happened in the past (*yesterday I go*), then it has high functional load. If the past tense also indicates the time frame (*yesterday I went*), the functional load of both the adverb and the verbal morphology is less than for either one occurring alone (p. 391)

Although the alleged theoretical foundation of cloze tests on reduced redundancy theory brought about their widespread application in testing as integrative measures of language proficiency, *the varying degrees of words' functional loads were not taken into account* in their development. Standard cloze tests, for instance, depend on deleting every nth word, e.g., 7th, of text regardless of the deleted words' functional load. For this very reason, research findings showed that they are extremely difficult and therefore suffer from low reliability. Along with several tests, Khodadady (2004), for example, administered the 35-item close test developed by Farhady and Keramati (1994) on the verb phrases of a university text to 34 senior undergraduate students of English and found it the least reliable, i.e., alpha = .55, and the most difficult test among others in terms of its mean IF, i.e. .45.

In order to remedy the “technical defects” of cloze tests Klein-Braley (1997, p.63) developed C-tests. After employing the Duisburg placement test DELTA as a validation criterion, Klein-Braley administered the C-tests consisting of four short 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. Since the C-tests had the highest loadings on the first *unrotated* factor, 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).

Khodadady (2007), however, took the 99 C-test items developed by Klein-Braley (1997) out of their context and presented them as single words to 63 university students majoring in English and asked them to restore the second mutilated half by adding the same number of letters given, i.e., words consisting of even number of letters, or one more, i.e., words comprising odd number of letters. He scored the restored words in two manners. First, the words meeting the direction requirement were scored correct and the resulting score was considered as an index of spelling ability. In the second scoring, only the exactly restored C-Test items were scored correct. Since the restoration of these exact items is independent of the contexts in which they appeared, the second procedure is also adopted in the present study and named CIC-Test.

When the scores of the participants on the standard C-tests, spelling test, CIC-Test, TOEFL, matching vocabulary test or lexical Knowledge test (LKT) were analysed via PCA method and the extracted factors were rotated, the standard C-Tests loaded the highest on the second factor whereas the TOEFL and the LKT loaded on the first and the spelling and CIC-Test on the third. Based on these results Khodadady (2007) concluded that standard C-Tests are method specific measures of language proficiency whose results depend on the way they are constructed.

C. Proficiency Tests Based on Schema Theory

While proficiency tests developed on language components do discretely specify what abilities in language should be measured one at a time, they fail to address the choices with which most of their representative items are designed (Khodadady, 1997, 1999, Khodadady & Herriman, 2000). They depend on experts and specialized institutions such as Educational Testing Service (ETS) to design well functioning items and subject them to item analysis before they are administered to test takers. Moreover, tests designed on components suffer from not having a sound theory as regards their selection and utilization of authentic texts, i.e., passages written to be read rather than tested (Khodadady, 1995). In order to have well functioning items, the reading passages of the TOEFL are usually *written* by testing specialists.

Although cloze tests can be developed on authentic texts found in newspapers, magazines and books, they are too difficult to answer (Khodadady, 2004). C-tests do overcome the difficulty involved in taking cloze tests by providing the first half of the deleted words as the missing parts of some carefully chosen short texts (Klein-Braley, 1997); however, they fail to specify what they really measure. In other words, the very dependence of the C-test items on the provision of the first part of the mutilated words, violates the theory behind closure, i.e., there are no *missing* parts but *mutilated* words whose constituting letters give away part of their missing information by specifying their number.

In contrast to two componential and reduced redundancy theories of language proficiency, schema theory views all the *words* constituting written texts as *schemata* which have syntactic, semantic, and discoursal relationships with each other. Since they have already been produced by writers under real conditions, in real places, at real times and for real purposes, i.e., being read or heard, the constituting schemata of authentic texts enjoy pragmatic relationships among themselves by their very being chosen to convey the message (Khodadady & Elahi, 2012). This microstructural approach towards viewing and accepting schemata as constituting units of texts differs sharply from its macrostructural perspective.

Macrostructurally, schema is defined as "a *conventional knowledge structure* that exists in memory" (Yule, 2006 p. 132). Since the presumed *conventional knowledge structure* [CKS] exists in isolation, i.e., dictionaries, and has little to do with any given text, it has failed to explain, as Grabe (2002) put it, "how it would work for reading comprehension" (p. 282). Similarly, Spiro, Vispoel, Schmitz, Samarapungavan, and Boerger (1987) emphasized the unproductive nature of detextuaslized macrostructural approach to schema theory by asserting that in spite of substantial agreement on the application of schema as pre-existing background knowledge, "we know very little about the organization of background knowledge and the method of its application to the understanding of new situations" (p. 177).

Microstructurally, a schema is, however, approached as any word which constitutes a text and is understood by its readers in terms of its syntactic, semantic, and discoursal relationships with the other schemata comprising the text. While macro structural schemata owe their existence to CKS given in dictionaries, they lack the potential to explain new situations brought up in various texts whereas the micro structural schemata are comprehended as a result of the readers' personal and background knowledge with each and all the schemata constituting those texts. In other words, *there is virtually no macro schema to explain what the whole text is about*. Rather, *these are the constituting words of a text or schemata and their relationships with each other which explain what it is composed to convey*.

Yule (2006), for example, offered the macro schema *classroom* as a CKS which explains Sanford and Garrod's (1981) interesting experiment. They presented the two sentences, "*John was on his way to school last Friday*. *He was really worried about the math lesson*," to some readers and asked them what they thought John did. As it can be read, there is nothing related to a *classroom* as Yule indirectly insists there is. While the *invisible* and *presumed* macro schema *classroom* provides no clues to the readers, the readers do have direct access to the 17 schemata comprising the two sentences in general and the schemata *school*, *math*, *lesson* and *worry* in particular to infer that John is *a school boy* because he is going to *school* and his teacher may ask him some *math* questions to which he might not have any answers for whatever reasons and he is therefore *worried*. Not surprisingly, they did say that they thought John was *a school boy*. In other words, the two sentences have nothing to do with *classroom* but *John* and his being understood in terms of the schemata with which he has been mentioned.

When Sanford and Garrod (1981) added the third sentence, "*Last week he had been unable to control the class*," the readers said that John was a school teacher. In other words, the schema related to John is discoursally related to new schemata given in the third sentence, the readers immediately modified their comprehension of John in terms of his job in order to accommodate the newly introduced schema *control*. Through personal experience, the readers have learned to assign the responsibility of controlling the *class* to teachers, not to students. They also modified their understanding of *worry* by changing its cause from facing possible math questions to controlling the *math class*. In other words, it is not the CKS of classroom which helps the readers change John's job from student to teacher, it is the juxtopositioning of the schemata forming the newly introduced sentence with the previous sentences which led the readers to change John's job.

Microstructural approach to schema theory not only explains the process of understanding written texts as establishing syntactic, semantic and discoursal relationships among the schemata comprising the texts but also specifies the type of alternatives with which the schemata deleted from the texts should be presented. In other words, instead of employing their intuition to develop traditional multiple choice items (see Khodadady, 1999), English language teachers and test designers alike can select certain schemata of given texts as keyed responses and present them with the choices which bear semantic, syntactic and disoursal relationships not only with the keyed responses but also other schemata comprising the whole texts. (Sample schema-based cloze multiple choice item will be given in the instrumentation section.)

If the componential tests, i.e., TOEFL, reduced redundancy tests, i.e., C-Tests, CIC-Test and CDC-Test, and schema-based tests, i.e., S-Tests, do measure language proficiency as common to all theories, the tests developed on these theories must all load *highly* on a single factor after *they have all been administered to the same sample* and their malfunctioning items have been removed, i.e., the tests have been tailored. However, if the TOEFL and its subtests, the standard C-tests along with its versions, and S-Test employed in this study derive their rationale from different theories, they must load on three distinct latent variables after they have been tailored and subjected to the PCA analysis.

II. Methodology

A. Participants

Four hundred thirty university students, 323 female and 107 male, took part in the study. Their age ranged between 18 and 45 (Mean = 22.95, SD = 3.18) and were majoring in English language and literature (N=352, 81.9%), teaching English as a foreign language (N=69, 16.0%), and English translation (N=9, 2.1%) for a bachelor (N=352, 81.9%), master (N=74, 17.2%) and PhD (N=4, 0.9%) degree. One hundred fifty seven (36.5%), 49 (11.4%), 97 (22.6%), and 127 (29.5%) were freshman, sophomore, junior and senior students at Ferdowsi University of Mashhad, Iran, respectively. While the majority spoke Persian (N=424, %=98.6), only six (1.4%) conversed in Turkish as their mother languages. They all participated in the project voluntarily. Since 16 participants did not take one of the tests, they were excluded from study. In order to compensate for their time and externally motivate the participants, it was announced that whoever took a course with the researcher; he would add an extra 10% to their final score and consider their participation as a part of their class activity.

B. Instruments

Three English language proficiency tests were employed in this study, i.e., disclosed Test of English as a Foreign Language (TOEFL), standard C-Tests and their context dependent and independent versions, and schema-based cloze multiple choice item test (S-Test). Since the vocabulary was also measured by the three proficiency tests, the Lexical Knowledge Test (LKT) was employed to explore its effect on the componential structure of tests.

1. Test of English as a Foreign Language

The disclosed TOEFL consisting of structure, written expressions, sentential vocabulary, and reading comprehension subtests was used as a representative test of the componential theory of language testing. The structure subtest consisted of 30 cloze multiple choice items which addressed discrete grammatical points in isolated sentences. The test takers' knowledge of the conjunction *because* is, for example, assessed in the example sentence below.

Geysers have often been compared to volcanoes … they both emit hot liquids from below the Earth's surface (ETS, 2003, p. 45).

A due to **B** because **C** in spite of **D** regardless of

The twenty five isolated and unrelated sentences of the written expressions subtest are divided into four underlined parts. The test takers were required to choose the erroneous part as the correct answer. In the example below, for instance, the test takers must choose B as the correct answer because after the preposition *for*, the verb of the sentence shows a process, i.e., *creating*, rather than a trait, i.e., creativity.

Animation is a technique for creativity the illusion of life in inanimate things (ETS, 2003, p. 49).
A B C D

The third subtest contained 30 isolated sentences in which some words and phrases had been underlined and their synonyms had to be found among the four choices offered below the sentences. The phrase *arrive at*, for example, is underlined in the sentence below so that the test takers can choose "its closest in meaning to the original sentence" (ETS, 1987, p. 13).

Most sound vibrations arrive at the eardrum by way of the auditory canal.

A search for **B** reach **C** tickle **D** whisper to

And finally 30 traditional multiple choice items addressing referential, factional and inferential points brought up in five short passages formed the reading comprehension subtest. The alpha reliability coefficients reported by Khodadady (2012) were .91, .79, .82, .82, and .86 for the total TOEFL, the structure, written expressions, sentential vocabulary, and reading comprehension sections, respectively.

2. Standard C-Tests

After careful selection of four short and separate passages, Klein-Braley (1997, pp. 79-80) left their first sentences intact and removed the second half of the second word to develop four C-Tests. With the exception of the C-Test 2 comprising 24 items, the other three C-Tests consisted of 24 items and thus formed the 99-item C-Tests employed in this study. The sixteen out of 25 items below, for example, form the standard C-Test 1.

Within the last twenty years the blood group of peoples in all parts of the world have been studied. The mo________ (1) interesting res________ (2) of th________ (3) studies h________ (4) been th________ (5), with f________ (6) exceptions, nea________ (7) every hu________ (8) group exam________ (9) has be________ (10) found t________ (11) consist o________ (12) a mix________ (13) of t________ (14) same fo________ (15) blood gro________ (16); …

Following Khodadady and Hashemi (2011) the C-Test 1, a part of which was given above, along with the three other C-Tests developed by Klein-Braely (1997) and piloted on either native speakers or second language teachers are referred to as standard C-Tests in this study. Khodadady (2012) reported the alpha reliability coefficients of .91, .77, .72, .78, and .70, for the total standard C-Tests and C-Test 1, 2, 3 and 4, respectively.

The 16 mutilated items comprising the standard C-Test 1, i.e., Mo ..., res..., th..., h..., th..., f..., nea..., hu..., exam.., be..., t..., o..., mix..., t..., fo..., gro..., along with the other items totaling 99 were taken out of their context and administered as a list of mutilated words to be restored by the participants. By receiving the same directions given for the standard C-Tests, they had to restore the *exact* mutilated words, e.g., *most*, *result*, *these*, *has*, *this*, *few*, *nearly*, *human*, *examined*, *been*, *to*, *of*, *mixture*, *the*, *four*, and *groups*, by restoring to their background knowledge. The exactly restored words were then treated as the constituents of CIC-Test.

Among the tailored CIC-Test items, some had acceptable facility and discrimination indices on the standard C-Tests as well. These well-functioning items were, therefore, removed from the tailored Standard C-Tests to form the third version of the C-Tests. Since this version did not include the items whose mutilated parts could be restored without having any context, they were treated as constituents of context-dependent (CD) C-Test.

3. S-Test

The articles published in *NewScientist* seem to be “more academic than … articles in quality newspapers" (Clapham, 1996, p. 145), Gholami (2006), therefore, chose “why don’t we just kiss and make up” (Dugatkin, 2005) from this magazine and developed four versions (V) of S-Test on its 60 adjectives (V1), 60 adverbs (V2), 60 nouns (V3), and 60 verbs (V4). She also developed the fifth version on the schemata comprising the semantic domain, i.e., 14 adjectives, seven adverbs, 24 nouns and 15 verbs of the same text and administered them to 92 undergraduate students of English language and literature at Ferdowsi University of Mashhad.

The first item of the semantic domain S-Test is given below. As can be seen, the schema ***attacks*** has been deleted and presented as choice A along with the three schemata *inroads*, *raids* and *ambushes* which share the syntactic feature of being nouns with each other. Since the choices B, C and D share the semantic feature of *trying to destroy* with the keyed response and thus compete with it in being selected, they are called *competitives* in order to differentiate them from their traditional counterparts, i.e., distracters (Khodadady, 1997). The readers must read all the choices and relate them to other schemata comprising the text in order to select the keyed response.

Why don't we just kiss and make up?

LOOK at the world's worst trouble spots and you can't fail to notice they have one thing in common: tit-for-tat … (1) between warring parties. …

1 **A**. attacks **B**. inroads **C**. raids **D**. ambushes

Table 1 presents the results obtained by Gholami (2006). As can be seen, V5, i.e., semantic domain S-Test proved to be the most challenging because the mean score (20.3) obtained on this version was the lowest as was the alpha reliability coefficient, i.e., .64. However, it showed the highest correlation coefficients with the TOEFL (r = .84, *p* <.01), its written expressions (r = .77, *p* <.01) and reading comprehension subtests (r = .74, *p* <.01), respectively. It also correlated significantly with the structure subtest of the TOEFL (r = .56, *p* <.05) and thus established itself as a valid measure of language proficiency. Based on these results, this version was taken as a representative S-Test and employed in this study.

TABLE 1
DESCRIPTIVE STATISTICS AND CORRELATIONS AMONG THE FIVE VERSIONS OF THE S-TEST

S-Test	# of items	Mean	SD	Alpha	Correlation Coefficients			
					TOEFL	Structure	Written expressions	Reading
Adjectives	60	21.2	9.1	.87	.55*	.57*	.27	.61*
Adverbs	60	24.7	10.7	.90	.70**	.50*	.63**	.63**
Nouns	60	24.2	9.6	.87	.19	.13	.05	.26
Verbs	60	22.0	12.1	.92	.19	-.13	.28	.21
Semantic domain	60	20.3	5.8	.64	.84**	.56*	.77**	.74**

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

4. Lexical Knowledge Test

Nation (see Schmitt, Schmitt, & Clapham, 2001) developed a vocabulary test consisting of 60 items presented in 20 groups of three numbered words/phrases. For the three items in each group, six words are offered from which the test takers have to choose their synonyms by writing the number of the synonymous item in their front. Since this test reflected mental lexicon defined as “the collection of words one speaker knows and the relationships between them” (Mir ěis, 2004, p. 2), it was named Lexical Knowledge Test (LKT) in this study. In order to save space and dispense with writing, the format of the test was also changed into six choices to be selected as the synonyms of the three numbered groups of words/phrases as shown below. The LKT is a highly reliable and valid measure of lexical knowledge because its alpha reliability coefficient was .89 in Khodadady’s (2007) study and it 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

C. Procedure

Following Khodadady (2007), CIC-Test was administered first because the administration of the standard C-Test would have given its items away. The standard C-Test was administered to the participants along with the TOEFL, S-

Test and the LKT in a counterbalanced manner in several testing sessions. Since the participants were known to the researcher, they were divided into four groups every session of test administration. For example, the first group took the standard C-Tests while the other three answered the TOEFL, S-Test and LKT. In the second session, the first group took the S-Test whereas the other three answered the standard C-Test, LKT and the TOEFL. Similarly, in the third session, the first, second, third and fourth groups of participants answered the LKT, S-Test, TOEFL and the C-Tests, respectively.

After all the tests were administered under standard conditions, the CIC-Test was tailored by keeping its well functioning items. The same procedure was followed for the standard C-Test. For developing the context dependent (CD) C-Tests, the items which had functioned well on the CIC-Test, were specified among the well functioning items on the standard C-Tests and removed. The fulfillment of these procedures resulted in the inclusion of three versions of C-Tests in statistical analyses, i.e., 1) Standard C-Tests, 2) CIC-Test, and 3) CDC-Test. The tailoring procedure was also followed for the TOEFL, its subtests, S-Test and LKT.

D. Data Analysis

Point biserial correlation coefficients were estimated as the item discrimination (ID) indices by correlating each individual item with the total score obtained on a given test. Following Thorndike (2005), items whose IDs fell below 0.20 were considered as malfunctioning and removed from validity analyses. The number of correct responses given to each item was also divided by the total number of answers to obtain facility (IF) indices. The FIs falling below 0.25 and above 0.75 were also excluded from validity analyses (Baker, 1989). Cronbach's Alpha was also estimated to determine the internal consistency of the tests containing well functioning items only. Furthermore, Kuder-Richardson Formula 21 (KR21) was employed to estimate the reliability of all tests and to compare them with the Alpha. Principal Component Analysis along with Varimax with Kaiser Normalization were utilized to extract rotated factors. All statistical analyses were performed by using IBM SPSS statistics 19.0 to address the questions below:

1. How reliable are the standard C-Tests, CIC-Test, CDC-Test, TOEFL, S-Test and LKT when they are tailored?

2. What is the factor structure when the tailored standard C-Tests, CDC-Tests and CIC-Test are studied? Will they load on a single factor representing reduced redundancy?

3. What is the factor structure when the tailored standard C-Tests, CDC-Tests and CIC-Test are studied with the tailored TOEFL, S-Test and LKT? Will they load on three factors representing reduced redundancy, schema and componential theories of language testing?

III. Results and Discussion

Table 2 presents the descriptive statistics as well as the reliability estimates of the proficiency tests administered in the study. As can be seen, the LKT is the most reliable test administered in the study (α =.93). The Standard C-Tests (α =.88) measuring reduced redundancy enjoy the same level of reliability as does the TOEFL (α =.88) as a measure of componential theory. The reliability of the S-Test (α =.85) is, however, slightly lower than the standard C-Tests and the TOEFL because of two reasons. First, it was relatively more difficult (mean IF = .47) than the standard C-Tests (mean IF = .51) and the TOEFL (mean IF = .59). Secondly, its well functioning items were fewer than the two tests. These results not only establish tailored tests as reliable measures and thus answer the first research question but also show that the fewness of well functioning items on a language proficiency test and its higher difficulty level lower its reliability.

TABLE 2
DESCRIPTIVE STATISTICS OF THE TESTS AND THEIR SUBTEST CONSISTING OF WELL FUNCTIONING ITEMS

	# of items	Mean	Variance	Std. Deviation	Mean IF	Mean ID	KR21	Alpha
Standard C-Tests	47	23.1	73.941	8.599	.51	.37	.86	.88
C-Test 1	15	7.9	10.350	3.217	.53	.36	.68	.71
C-Test 2	13	6.7	7.737	2.782	.48	.37	.63	.67
C-Test 3	13	6.7	8.857	2.976	.51	.39	.69	.73
C-Test 4	6	2.3	2.140	1.463	.48	.41	.41	.51
CDC-Tests	41	19.9	57.533	7.585	.50	.38	.84	.86
CDC-Test 1	14	7.3	9.156	3.026	.52	.36	.67	.70
CDC-Test 2	12	5.7	7.070	2.659	.48	.38	.63	.68
CDC-Test 3	11	5.5	6.546	2.558	.50	.40	.64	.69
CDC-Test 4	4	1.4	1.183	1.088	.50	.52	.32	.45
CIC-Test	13	7.1	4.679	2.163	.45	.27	.34	.52
TOEFL	47	27.6	78.330	8.850	.59	.37	.87	.88
Structure	10	6.5	5.194	2.279	.65	.39	.63	.66
Written Expressions	13	7.6	9.845	3.138	.59	.39	.74	.76
Sentential Voc.	4	2.6	1.301	1.141	.66	.25	.41	.44
Reading	20	10.8	19.476	4.413	.54	.37	.78	.81
S-Test	42	19.8	55.036	7.419	.47	.37	.83	.85
Lexical Knowledge Test	47	22.1	118.640	10.892	.47	.47	.92	.93

The second interesting finding of the present study is the closeness of alpha and KR21 coefficients and the relatively but consistently low magnitude of the latter to the former. As can be seen in Table 2, the alpha obtained on the standard C-Tests (.88), for example, is slightly higher than KR21 (.86). It is not, however, clear why some scholars studying C-Tests utilize either the KR21 (e.g., Babaii & Ansari, 2001) or KR20 (e.g., Jafarpur, 1999) when the alpha can easily be estimated by employing statistical packages such as the SPSS.

Table 3 presents the unrotated and rotated factors extracted via PCA and Varimax with Kaiser Normalization. As can be seen, standard C-Test 4 and CDC-test 4 both load on the unrotated factor two. However, when the two factors are rotated, not only Standard C-Tests but also CDC-Tests load acceptably on the second factor as well. If we assume the first factor to represent the reduced redundancy theory of language proficiency as measured by Standard C-Tests, what does the second factor represent? For finding the answer, the S-Test, LKT and the TOEFL with its subtests were included in the statistical analysis.

TABLE 3
FACTOR MATRIX OF THE STANDARD C-TESTS AND THEIR CONTEXT DEPENDENT AND INDEPENDENT VERSIONS

Tests	Unrotated Factors		Rotated Factors	
	1	2	1	2
Standard C-Tests	.997	*	.896	.440
C-Test 1	.844	*	.756	.378
C-Test 2	.846	*	.861	*
C-Test 3	.818	*	.810	*
C-Test 4	.740	.624	.332	.909
CDC-tests	.988	*	.900	.415
CDC-test 1	.832	*	.750	.365
CDC-test 2	.841	*	.859	*
CDC-test 3	.805	*	.808	*
CDC-test 4	.641	.704	*	.929
CIC-Test	.774	*	.608	.495
Eigenvalue:	7.671	1.140	6.060	2.750
% of Variance:	69.733	10.363	55.095	25.00

* Loadings less than .30

Table 4 presents the factors extracted from the three proficiency tests representing reduced redundancy, schema and componential theories of language testing. As can be seen, the Standard C-Tests have the highest loading on the first unrotated factor (.96) followed by CDC-Tests (.95). This factor provides further support for Khodadady (2007) labeling of C-Tests as method-specific measures of language proficiency because only its standard and content-dependent versions have the highest loading on the first factor even when the factors are rotated. As a matter of fact, rotating the factors emphasizes the method-specificity of C-Tests because both the standard C-Tests and CDC-Tests reveal almost the same loading on the first factor, i.e., .92, and thus highlight the distinct contribution of C-Tests to reduced redundancy testing in terms of their unique context.

TABLE 4
FACTOR MATRIX OF THE TESTS ADMINISTERED IN THE STUDY

Tests	**Unrotated Factors**			**Rotated Factors**		
	1	2	3	1	2	3
Standard C-Tests	.961	*	*	.911	.407	*
C-Test 1	.814	*	*	.775	.369	*
C-Test 2	.799	-.357	*	.840	*	*
C-Test 3	.783	*	*	.777	*	.305
C-Test 4	.757	*	-.394	.483	.664	*
CDC-Tests	.947	*	*	.917	.377	*
CDC-Test 1	.803	*	*	.766	.360	*
CDC-Test 2	.791	-.368	*	.841	*	*
CDC-Test 3	.764	-.315	*	.779	*	.315
CDC-Test 4	.656	*	-.445	.391	.626	-.349
CIC-Test	.776	*	*	.617	.477	*
S-Test	.678	.334	*	.308	.698	*
TOEFL	.825	.455	*	.333	.832	.383
Structure	.724	.363	*	.316	.704	.328
Written Expressions	.761	.366	*	.345	.745	*
Sentential Voc.	*	*	.572	*	*	.596
Reading	.682	.446	*	*	.747	*
LKT	.566	.448	*	*	.721	*
Eigenvalue:	10.266	1.765	1.078	6.760	5.148	1.200
% of Variance:	57.034	9.803	5.988	37.556	28.602	6.666

The second rotated factor extracted in this study represents schema theory because the S-Test has its highest loading (.70) on this latent variable as does the LKT, i.e., .72. Furthermore, not only the TOEFL itself but also its written expressions and reading comprehension ability subscales have the highest loading on this factor, i.e., .83, .75, and .75, respectively. The TOEFL does not, however, represent schema theory because it loads acceptably on the third factor as well (.38) as do its structure and sentential vocabulary subtests, i.e., .33 and .60, respectively. The reading comprehension subtest of the TOEFL as well as the LKT depend, nonetheless, on schema theory because they call for establishing discoursal and semantic relationships between the questions and the texts on the one hand and the words given as choices on the other.

Along with the first factor representing the reduced redundancy as measured by C-Tests, and the second factor representing schema theory as measured by S-Test, a third factor appears upon which the vocabulary subscale of the TOEFL has the highest loading, i.e., .60. Since the TOEFL itself and its structure subscale load acceptably on the third factor, i.e., .38 and .33, respectively, along with the CDC-Test 3 (.32) and standard C-Test 3 (.31), it represents the componential theory of language testing because answering both the TOEFL and some C-Tests depend *solely* on discrete rather than discoursal comprehension of words.

Among the three rotated factors extracted in this study, the second represents language proficiency in general due to the explanatory power of schema theory. It shows the test takers' ability to activate their syntactic knowledge of schemata to answer the structure subtest of the TOEFL, mobilize their semantic knowledge of schemata to select their synonyms from among the six offered in the LKT and establish syntactic, semantic and discoursal relationship among the schemata to choose the keyed responses on the S-Test and reading comprehension subtest of the TOEFL. These arguments are further supported when the correlations among these measures are taken into account.

Table 5 presents the correlation coefficients obtained among the English language proficiency tests and their subtests. As can be seen, the correlations among the Standard C-Tests, TOEFL and the S-Test do not reach .80 so that they can *replace* each other (Hatch & Lazaraton, 1991, p. 442), indicating that the dependence of language proficiency tests on different theories bring about different indicators of test takers' proficiency. While the standard C-Tests explain 44 percent of variance in the TOEFL (r = .66, *p*<.01), for example, it drops to 41 and 37 percent for the CDC-Tests (r = .64, *p*<.01), and the S-Test (r = .61, *p*<.01), respectively.

TABLE 5
CORRELATIONS AMONG THE TESTS ADMINISTERED IN THE STUDY

Tests and Subtests	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1 TOEFL	1	.84**	.86**	.28**	.89**	.66**	.55**	.53**	.54**	.58**	.64**	.55**	.52**	.52**	.50**	.58**	.61**	.54**
2 Structure	.84**	1	.71**	.15**	.62**	.59**	.49**	.46**	.50**	.49**	.57**	.48**	.45**	.49**	.41**	.53**	.48**	.47**
3 Written Expressions	.86**	.71**	1	.16**	.60**	.63**	.55**	.49**	.49**	.55**	.61**	.55**	.48**	.47**	.47**	.54**	.58**	.50**
4 Sentential Voc.	.28**	.15**	.16**	1	.12*	.17**	.10	.17**	.17**	.13**	.18**	.10*	.17**	.18**	.13**	.13**	.12*	.11*
5 Reading	.89**	.62**	.60**	.12*	1	.53**	.44**	.43**	.43**	.49**	.52**	.43**	.42**	.41**	.42**	.48**	.53**	.45**
6 Standard (S) C-Tests	.66**	.59**	.63**	.17**	.53**	1	.85**	.85**	.83**	.70**	.99**	.84**	.84**	.81**	.59**	.75**	.56**	.43**
7 SC-Test 1	.55**	.49**	.55**	.10	.44**	.85**	1	.63**	.54**	.54**	.86**	.99**	.62**	.54**	.45**	.60**	.48**	.37**
8 SC-Test 2	.53**	.46**	.49**	.17**	.43**	.85**	.63**	1	.61**	.48**	.85**	.62**	.98**	.60**	.39**	.62**	.44**	.32**
9 SC-Test 3	.54**	.50**	.49**	.17**	.43**	.83**	.54**	.61**	1	.49**	.81**	.53**	.61**	.97**	.41**	.63**	.43**	.31**
10 SC-Test 4	.58**	.49**	.55**	.13**	.49**	.70**	.54**	.48**	.49**	1	.67**	.53**	.49**	.48**	.88**	.65**	.52**	.46**
11 CDC-Tests	.64**	.57**	.61**	.18**	.52**	.99**	.86**	.85**	.81**	.67**	1	.86**	.85**	.82**	.59**	.68**	.54**	.41**
12 CDC-Test 1	.55**	.48**	.55**	.10*	.43**	.84**	.99**	.62**	.53**	.53**	.86**	1	.61**	.52**	.44**	.56**	.47**	.36**
13 CDC-Test 2	.52**	.45**	.48**	.17**	.42**	.84**	.62**	.98**	.61**	.49**	.85**	.61**	1	.60**	.39**	.58**	.43**	.33**
14 CDC-Test 3	.52**	.49**	.47**	.18**	.41**	.81**	.54**	.60**	.97**	.48**	.82**	.53**	.60**	1	.40**	.55**	.40**	.28**
15 CDC-Test 4	.50**	.41**	.47**	.13**	.42**	.59**	.45**	.39**	.41**	.88**	.59**	.44**	.39**	.40**	1	.48**	.46**	.41**
16 CIC-Test	.58**	.53**	.54**	.13**	.48**	.75**	.60**	.62**	.63**	.65**	.68**	.56**	.58**	.55**	.48**	1	.50**	.42**
17 S-Test	.61**	.48**	.58**	.12*	.53**	.56**	.48**	.44**	.43**	.52**	.54**	.47**	.43**	.40**	.46**	.50**	1	.61**
18 LKT	.54**	.47**	.50**	.11*	.45**	.43**	.37**	.32**	.31**	.46**	.41**	.36**	.33**	.28**	.41**	.42**	.61**	1

** Correlation is significant at the 0.01 level (2-tailed)
* Correlation is significant at the 0.05 level (2-tailed)

The explanation of a larger variance in the TOEFL, i.e., 41%, by the Standard C-Tests, is the result of their inclusion of a noticeable number of items developed on syntactic schemata. (As can be seen in Table 5, the correlation of the standard C-Tests with the structure, i.e., r = .59, *p*<.01, is higher than its correlation with reading, i.e., r = .53, *p*<.01. If the English language proficiency is defined as an ability to "read and understand short passages … in North American colleges and universities" (ETS, 2003, p. 11), then the same amount of variance in the reading comprehension ability of participants, i.e., 28% is explained by both standard C-Tests and S-Tests (r = .53, *p*<.01), indicating that the selection of a single authentic text does away with choosing a number of texts by specialists and submitting them to pilot studies as suggested by designers not only of standard C-Tests but also of the TOEFL.

IV. CONCLUSIONS

The mutilated words comprising the four texts of standard C-Tests were taken out of their context and presented as isolated items to 430 undergraduate and graduate students majoring in various fields related to English as a foreign language in order to establish two new versions of C-Tests. The exact and well functioning answers given to the decontexualized items were treated as Context Independent (CI) C-Test to establish the first version. The well functioning items of the CIC-Test were then removed from the standard C-Tests to establish the Context Dependent (CD) C-Tests as the second version. Following Brown (1984) and Kamimoto (1993) not only the standard C-Tests, CIC-Test and CDC-Tests but also other proficiency tests employed in this study were tailored on the basis of the argument that whatever constructs these tests measured were reflected in their items having acceptable IF and ID indices only.

When the standard C-Tests, CIC-Test and CDC-Tests were submitted to factor analysis two latent variables emerged. Since the three versions of C-Tests had the highest loading on the first factor, it was taken as the construct representing reduced redundancy theory of language proficiency measured by C-Tests. The rotation of the factors showed that both standard C-Tests and its CI and CD versions loaded acceptably on the second factor as well, indicating that they were measuring a construct other than reduced redundancy. When the tailored TOEFL, schema-based cloze multiple choice item test (S-Test) and the Lexical Knowledge Test (LKT) were included in the factorial analysis, three factors emerged.

With the exception of the sentential vocabulary subtest of the TOEFL, the three versions of C-Tests as well as the TOEFL, S-Test and the LKT loaded acceptably on the first factor. The loadings of the C-Tests were, however, much *higher* than not only the TOEFL and S-Test but also the LKT measuring lexical knowledge of participants. The noticeably *different* loadings on the part of the three proficiency tests and LKT thus necessitated the rotation of three extracted factors resulting in the highest loadings of the three versions of C-Tests on the first factor *only*. These findings provided further support for the method-specificity of C-Tests in measuring reduced redundancy.

In contrast to the standard C-Tests, CIC-Test and CDC-Tests, not only the S-Test but also the TOEFL and LKT loaded the highest on the *second* rotated factor. Since S-Test had its highest loading on this factor as did the written expressions and reading comprehension subtest of the TOEFL and the LKT, it represented schema theory of language proficiency. Although among the three proficiency tests the TOEFL had the highest loading on the second factor, it did not represent schema theory because its structure and sentential vocabulary subtests loaded on the third factor which represents the componential theory of language proficiency. Future research must, however, show whether the same three factors will be extracted if more representative measures of reduced redundancy, schema and componential theories of language testing are administered, tailored and analysed factorially.

ACKNOWLEDGEMENT

Dr. Ali Shariati Faculty of Letters and Humanities at Ferdowsi University of Mashhad funded this study via grant coded 2/15117. The author wishes to acknowledge the support of its authorities and express his gratitude for their approval of the grant on June 22, 2010.

REFERENCES

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

[2] Bardovi-Harlig, K., & Comajoan, L. (2008). Order of Acquisition and Developmental Readiness. In B. Spolsky and F. M. Hult (Eds.). *The Handbook of Educational Linguistics* (pp. 383-397. Malden, MA: Blackwell.

[3] Brown, J. D. (1984). A cloze is a cloze is a cloze? In J. Handscombe, R. A. Orem, & B. P. Taylor (Eds.), *On TESOL '83* (pp. 109-119). Washington, DC: TESOL.

[4] Educational Testing Service (1987). Reading for TOEFL: An official TOEFL study aid. Princeton, NJ.: ETS.

[5] Educational Testing Service (2003). TOEFL Test preparation kit (2nd ed.). Princeton, NJ.: ETS.

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

[7] Grabe, W. (2002). Dilemmas for the development of second language reading abilities. In J. C., Richards & Renandya, W. A. (Eds.). *Methodology in language teaching: An anthology of current practice* (pp. 276-286). Cambridge: CUP.

[8] Jafarpur, A. (1999). Can the C-test be improved with classical item analysis? *System* 29, 79-89.

[9] Kamimoto, T. (1993). Tailoring the test to fit the students: Improvement of the C-test through classical item analysis. *Fukuoka Women's Junior College Studies*, 30, 47-61.

[10] Khodadady, E. (1995, September). Textual analysis of testing materials: Validity readdressed. Paper presented at the 8th Educational conference of the ELICOS Association of Australia, Fremantle, Western Australia.

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

[12] Khodadady, E. (1999). Multiple-choice items in testing: Practice and theory. Tehran: Rahnama.

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

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

[15] Khodadady, E. (2012). Construct Validity of C-Tests: A factorial approach. Manuscript submitted for publication.

[16] Khodadady, E., & Elahi, M. (2012). The effect of schema-vs-translation-based instruction on Persian medical students' learning of general English. *English Language Teaching*, 5(1), 146-165. URL: http://dx.doi.org/10.5539/elt.v5n1p146.

[17] Khodadady, E., & Hashemi, M. (2011). Validity and C-Tests: The role of text authenticity. *Iranian Journal of Language Testing*, 1(1), 1-12.

[18] Khodadady, E., & Hashemi, M. (2011). Validity and C-Tests: The role of text authenticity. *Iranian Journal of Language Testing*, 1(1), 1-12.

[19] 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.

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

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

[22] Mir ə̉is, 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

[23] Sanford, A., & Garrod, S. (1981). Understanding written language. New York: J. Wiley & Sons.

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

[25] Spiro, P., Vispoel, W., Schmitz, J., Samarapungavan, A., & Boerger, A. (1987). Cognitive flexibility and transfer in complex cognitive domains. In B. Britton & S. Slynn (Eds.). *Executive control processes in reading* (pp. 177-199). Hillsdale, NJ: Lawrence Erlbaum.

[26] 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.

[27] Taylor, W.L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, 30, 415–433.

[28] Thorndike, R. M. (2005). Measurement and evaluation in psychology and education (7th ed.). Upper Saddle River, NJ: Pearson Education.

[29] Yule, G. (2006). The study of language (3rd ed.). Cambridge: CUP.

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.

He 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) and *English Language Proficiency Course: First Steps* (Sanandaj, Kurdistan University, 2001). His main research interests are Testing, Language Learning and Teaching.

Dr. Khodadady is currently a member of Teaching English Language and Literature Society of Iran (TELLSI), TESL Ontario and European Society for Translation Studies. He is on the editorial board of *Ferdowsi Review: An Iranian Journal of TESL, Literature and Translation Studies* and has reviewed some research papers for *Iranian Journal of Applied Linguistics* and *TESL Canada Journal* as a guest reviewer.