S-Tests and C-Tests: Measures of Content-Based Achievement at Grade Four of High Schools

Ebrahim Khodadady[1]

Elmira Ghergloo[1]

Abstract

This study aims to explore the internal, empirical as well as content validity of C-Tests and S-Tests developed on the passages of the textbook Learning to Read English for Pre-University Students (Birjandi, Sarab, & Samimi, 2011) taught to grade four high school students in Dargaz, Iran. To this end, all the single and phrasal words of the textbook were treated as schemata and assigned to semantic, syntactic and parasyntactic domains, genera and species. By choosing a number of paragraphs from the passages of the textbook, a 100-item C-Test and 90-item S-Test were designed and administered to 283 students. The reliability estimates showed that the C-test was more reliable than the S-Test. When the two tests were subjected to correlational analyses they correlated very highly with each other (r=.99, p<.01), establishing themselves as empirically valid measures of English achievement. The analysis of types and tokens comprising the two tests, however, showed that the S-Test enjoyed higher content validity because it measured more schema types than the C-Test did and its difficulty level was, therefore, higher. The results are discussed and suggestions are made for future research.

Keywords: Testing, content vs. objective based achievement, schema theory, fairness

1. Introduction

Along with subjects such as history and biology, English has been offered at four grades in Iranian high schools for a long time. To the best knowledge of the present researchers there are, however, few academic studies available to show what type of reliable and valid language tests can be designed to measure students' English achievement in these grades in general and grade four or pre-university level in particular. There is just one study available in which Ghaffari (2000) explored whether schema-based cloze multiple choice item tests or "S-Tests" (Khodadady, 2012, p. 587) could be developed to fulfill the function. His findings showed that English teachers themselves could write and employ S-Tests to measure pre-university high school (PUHS) students' English achievement successfully.

The present study extends what Ghaffari (2000) did and provides more comprehensive procedures to develop the S-Tests in the light of what recent studies have found regarding the schema-based analysis of the passages on which the tests are designed (e.g., Khodadady, Alavi & Khaghaninejad, 2011; Khodadady & Javadi Mehr, 2012; Khodadady & Khaghaninejad, 2012). It also aims at exploring whether C-Tests can be developed and utilized as measures of English achievement.

---

[1] Ferdowsi University of Mashhad, Iran.

1.1 S-Tests

S-Tests are cloze multiple choice item tests which employ the micro structural approach of schema theory to justify and establish their theoretical foundation (Khodadady, 1997, 1999). The approach treats the single and phrasal words comprising texts as schemata whose activation in their readers' minds in isolation and in connection with each other results in their comprehension of texts (Khodadady et al, 2011; Khodadady & Elahi, 2012). Instead of adopting sentences as the basic units of language (e.g., Chomsky, 1957, 1965), schemata are viewed as its building blocks whose juxtopositioning with each other brings about languages and texts alike.

Treating schemata as the basic units of language plays a very important role in language testing for several reasons. First, they entail identifying and classifying all the schemata comprising texts and thus secure the content validity of S-Tests, Secondly, they help appreciate the hierarchical nature of language learning through semantic, syntactic and parasyntactic domains and their genera, species, types and tokens (Khodadady & Herriman, 2000). Thirdly, the number of items to be developed is determined by the percentage of semantic, syntactic and parasyntactic schemata comprising the texts to be tested. Fourthly, the S-Test alternatives are chosen not only on the basis of their types and tokens but also by the discoursal relationships they enter with other schemata comprising the text. And finally the three semantic, syntactic and parasyntactic domains enable researchers to conduct research projects dealing with language translation (e.g., Khodadady, 2008) and teaching (e.g., Khodadady, Alavi, & Khaghaninejad, 2011).

To evaluate the content validity of a Persian translation, Lagzian (2012), for example, employed 70 pages (10%) comprising the dentistry textbook *Radiology: Principles and Interpretation* (White & Pharoah, 2004), parsed and codified its schemata and compared them with their Persian equivalents as translated by Valizadeh (1384). A domain-based analysis of her data shows that the randomly chosen pages consist of 27926 English schemata, i.e., 15670 (56.1%) semantic, 10503 (37.6%) syntactic and 1753 (6.3%) parasyntactic. The statistical analyses of the domains show that they differ from each other significantly in terms of their schema types and tokens.

Semantic domain schemata are open set words (Quirk, Greenbaum, Leech & Svartvik, 1985) which represent the concepts readers, and in the context of the present study, test takers ought to activate in their minds and connect to each other in order to understand what they read, i.e., adjectives, adverbs, nouns and verbs. To fulfill their pivotal function, these four genera of semantic domain schemata have to be many in type but relatively few in their token or frequency. The 70-randomly selected pages of *Radiology: Principles and Interpretation* (White & Pharoah, 2004) textbook, for example, consists of 3703 semantic schema types of which 1841 (49.7%) and 587 (15.9%) have been used just once or two times, respectively. In other words, 2428 (65.6%) out of 3703 semantic schemata have a token of one or two.

In contrast to semantic domain schemata, syntactic ones are few in type but many in tokens. They consist of conjunctions, determiners, prepositions, pronouns, and syntactic verb genera. These genera fulfill a linguistic function and must attach to semantic schemata to find their specific meaning. The 70-randomly selected pages of dentistry textbook, *Radiology: Principles and Interpretation* (White & Pharoah, 2004), for example, contains only 220 syntactic genera of which 89 (40.5%) have tokens of one or two. Due to their linguistic function some syntactic schemata are fairly frequently utilized by authors to describe and connect their main semantic schemata together into coherent sentences. The determiner "the", preposition "of" and conjunction "and" have, for example, the tokens of 2571, 1385, and 801, in the textbook, respectively.

Similar to the schemata constituting the semantic domain of texts, parasyntactic schemata are usually many in type but few in tokens. They comprise abbreviations, interjections, names, numerals,

para-adverbs, particles, and symbols. These genera have, however, a basically linguistic function and must of necessity attach themselves to semantic schemata in order to express a specific meaning. For example, out of 433 parasyntactic schema types comprising the seventy pages of *Radiology: Principles and Interpretation* (White & Pharoah, 2004), 334 (77.1) had tokens of one or two. The particle "to", the abbreviation "fig." and negation para-adverb "not" had the highest tokens in the 70 pages analysed, i.e., 217, 116, 100, respectively.

In developing S-Tests, all the schema types constituting given texts are taken into consideration and their items are developed on the basis of their being used once and the percentage with which they comprise the whole text. These important criteria require employing more semantic schema types and fewer syntactic ones as the statistics reported in the previous paragraphs show. Khodadady and Alaee (2012), for example, developed an S-Test to explore the relationship between social capital and English language achievement. For designing their S-Test they chose one paragraph from each of six lessons comprising *English book 3* (Birjandi, Nouroozi, & Mahmoodi, 2010) taught at grade three in Iranian high schools. The S-Test comprised thirty five semantic schemata (81.4%) and eight syntactic and parasyntactic schemata (18.6%). The very dependence of S-Tests on the proportionate selection of syntactic and semantic schemata differentiates them from C-Tests as measures developed on the microstructural approach of schema theory as well (see Khodadady, 2013).

1.2 C-Tests

C-Tests are variants of cloze tests which were originally designed by Taylor (1953) to measure readability. They were, however, adopted as measures of language proficiency in applied linguistics. For designing cloze tests a reading passage is usually given to test takers in which the first sentence is left intact "to set the scene" (Vallete, 1977, p. 212) and from the second sentence every nth, e.g., fifth or seventh, word is deleted and replaced with a blank. The test takers are required to restore the exact or acceptable words to reveal their proficiency. Research findings, however, show that cloze tests are extremely difficult when they are scored on the basis of exact-word method and as a result suffer from lower reliability levels.

Khodadady (2004), for example, administered a C-Test, S-Test, text-driven cloze test, traditional cloze multiple choice item test (MCIT) and Test of English as a Foreign Language (TOEFL) to 34 non-native speakers to find out what factors the tests loaded on. His results showed that the cloze test had the lowest mean p-value (.45) as compared to the C-Test (.54), S-Test (.59), cloze MCIT (.61) and TOEFL (.60). Its alpha reliability coefficient (.55) was also lower than these tests, i.e., .86, .71, .78, and .88, respectively. In addition to these two defects, Klein-Braley and Raaz (1984) and Klein-Braley (1997) brought up six other shortcomings which seem to question the acceptability of cloze tests.

In order to provide a justified replacement for cloze tests from both theoretical and psychometric perspectives (Babaii & Moghaddam, 2006), Klein-Braley (1981) and Raatz and Klein-Braley (1981) developed C-Tests on the basis of a new deletion procedure called the C-Principle or the Rule of Two. In this procedure instead of deleting the whole nth words, every second word from the second sentence is mutilated by deleting the second half. The test takers are required to restore the second half in order to get the item right.

The nine C-Test items below which were developed by Khodadady and Hashemi (2011) on the second words of the second sentence of the authentic text "Why don't we just kiss and make up?" (Dugatkin, (2005) provide representative samples. As can be seen, the first mutilated schema is o_ (1) whose second half, i.e., f, must be restored by test takers. Klein-Braley (1997) claimed that the items are selected on the basis of reduced redundancy principle (RRP) put forward by Spolsky (1973), i.e.,

"The non-native's inability to function with reduced redundancy, evidence that he cannot supply from his knowledge of the language the experience on which to base his guesses as to what is missing" (p. 17).

Sample C-Test items

LOOK at the world's worst trouble spots and you can't fail to notice they have one thing in common: tit-for-tat attacks between warring parties. Escalation o_____ (1) violence i_____ (2) incredibly destr_____ (3), yet w_____ (4) humans fi_____ (5) it ve_____ (6) difficult t_____ (7) break t_____ (8) vicious cy_____ (9).

Khodadady (2013), however, argued that no language test has been developed on the basis of RRP so far simply because any part of a given message, i.e., words, phrases, clauses and even complete sentences of a text, can go missing under real circumstances such as unexpected interruptions encountered in contexts such as listening to a lecturer or watching news broadcast. The selection and mutilation of every second schema in C-Tests is, however, systematic and thus defies randomness as the basic principle of reduced redundancy. He further argued that C-Tests are developed on single-word schemata and are, therefore, microstructural by their very nature. The present study is, therefore, designed to explore whether C-Tests can be employed as achievement measures as are S-Tests. It also aims to explore the content and empirical validities of both C-Tests and S-Tests.

2. Methodology

2.1 Participants

Two hundred eighty three PUHS students participated voluntarily in the present study. However, since thirty three of them failed to take one of the two tests or missed most of their constituting items they were excluded from the study. The remaining 250 students, 169 (67.6%) female and 81 (62.4%) male, had enrolled in state (n=220, 88%) and private (n=30, 12%) schools. They were majoring in human sciences (n=74, 29.6%) math (n=66, 26.4%) and science (n=110, 44.0) in Almahdi (n=53, 21.2%) Emam (n=41, 16.4%), Gholami (n=52, 20.8%), Parvin (n=49, 19.6%), Reyhane (n=15, 6.0%), Sayad (n=25, 10.0%), and Sina (n=15, 6.0%) high schools in Dargaz. The participants' age ranged between 17 and 18 (mean=17.9, SD=.29). They spoke Persian (n=200, 80.0%), Kurdish (n=18, 7.2%) and Turkish (n=32, 12.8%) as their mother language.

2.2 Materials

The textbook *Learning to Read English for Pre-University Students* (Birjandi, Sarab & Samimi, 2012) taught to PUHS students was employed to develop the C-Test and S-Test in this study. As can be seen in Table 1, its eight reading passages and vocabulary sections consist of 5790 schema tokens and 1578 schema types. In terms of domains, they comprise 2915 (50.3%) semantic, 2256 syntactic (39.5%), and 586 parasyntactic (10.2%) tokens. As it can also be seen, when schema types are taken into consideration, the number and percentage of semantic (n=1174, 74.4%), syntactic (n=209, 13.2%) and parasyntactic (n=195, 12.4%) schemata change drastically.

Table 1: The schema domain tokens and types comprising the pre-university English textbook

Domains	Tokens		Types	
	Frequency	Percent	Frequency	Percent
Semantic	2915	50.3	1174	74.4
Syntactic	2286	39.5	209	13.2
Parasyntactic	589	10.2	195	12.4
Total	5790	100.0	1578	100.0

2.3 Instruments

Four instruments were designed and employed in this study, i.e., a demographic scale, C-Test, and S-Test. (Another scale was also employed in the study. The data and results related to that scale will, however, be reported in a separate paper.)

2.3.1 Demographic Scale

The demographic scale contained seven short answer questions and multiple choice items dealing with the participants' age, gender, name and type of school, field of study and mother language.

2.3.2 C-Test

Following Klein-Braley (1997) five paragraphs were chosen from the eight lessons comprising the textbook *Learning to Read English for Pre-University Students*. The title of each paragraph along with its first sentence were left intact. Starting from the second sentence of each paragraph every second word was mutilated by deleting its second half. If the word consisted of odd number of letters, the larger half was deleted. This procedure resulted in deleting twenty words from each paragraph totaling 100 for the whole C-Test (Appendix A). No study has employed C-Tests to measure achievement.

Following Babaii and Ansary (2001), each letter deled from the second half of every second word was replaced with a dash (-) to provide the test takers with extra clue and visual help to restore the deleted half by focusing on the number of required letters. All the instructions were given in Persian, the test takers' mother language, to make sure that they understood exactly what they were required to do. In order to help them review their written responses, each mutilated word was also numbered in superscript form. Numbering the items also helped tabulate the data more accurately. The title, the first intact sentence and the second mutilated sentence are given as examples below.

Aerobics

Aerobics is a word for needing oxygen, and aerobic exercise is any kind of activity that makes your muscles use oxygen. Aerobic ^{1}**exer**_ _ _ _ is^2**repet**_ _ _ _ _, meaning ^{3}**i**_ is ^{4}**a**_ activity ^{5}**th**_ _ you ^{6}**d**_ over ^{7}**a**_ _ over, ^{8}**t**_ keep ^{9}**brin**_ _ _ _ in ^{10}**oxy**_ _ _ to ^{11}**a**_ _ of ^{12}**yo**_ _ muscles.

2.3.3 S-Test

A schema-based cloze multiple choice item test (S-Test) was developed by choosing at least one paragraph from each of the eight lessons comprising the textbook *Learning to Read English for Pre-University Students* (Birjandi, Sarab, & Samimi, 2012) totaling to fourteen. From these paragraphs ninety single/phrasal schemata, were chosen, deleted and offered as the keyed responses (Appendix B). In order to find out whether the test takers could activate their knowledge of the keyed response and relate it to schemata comprising the paragraphs, three alternatives called "competetives" (Khodadady, 1997) were chosen from among the 1578 schema types comprising the content of the whole textbook. As the first and second S-Test items given below show, the competitives have syntactical, semantic and discoursal relationships with the keyed response.

Aerobics is a word for … (1) oxygen, and aerobic exercise is any kind of activity that … (2) your muscles use oxygen.

		lacking		needing*		avoiding		missing
		makes*		builds		produces		creates

The three competitives, *lacking*, *avoiding* and *missing*, of the first item, for example, are all syntactically related to the keyed response *needing* in being verbs. Semantically, *needing* shares the semantic feature of "being essential" with the competitives *lacking* and *missing*. However, the

contextual schema *aerobics* does not qualify as an animate agent to *miss* or *lack* a substance such as oxygen and thus *lacking* and *missing* become discoursally inappropriate to be chosen as the keyed response. Similarly, *avoiding* indicates the necessity of keeping away from something during aerobics, however, it does not relate to the keyed response as semantically as the other two competitives do.

2.4 Procedures

Similar to Khodadady and Lagzian (2013) the reading passages, new words and their definitions were typed and broken into their single word and phrasal schemata. Following Khodadady (2008), parsed schemata were assigned to three domains, i.e., semantic, syntactic and parasyntactic. The 16 genera of these domains were then specified, i.e., adjective, adverbs, nouns, verbs, conjunctions, determiners, prepositions, pronouns, syntactic verbs, abbreviations, interjections, names, numerals, para-adverbs, particles, and symbols. And finally the 122 species of these genera were codified. The present researchers believe that the codification of all schema species helps not only determine the syntactic and semantic load of all texts but also secures the content validity of achievement tests such as C-Tests and S-Tests.

Upon determining the percentage of semantic, syntactic and parasyntactic schema species, the S-Test was developed. As will be discussed shortly, the codification of schema species allowed researchers to develop their test items proportionately. However, this procedure could not be followed for the C-Test since its development is rather set, i.e., mutilation of every second word. While 100 C-Test items could be easily designed on five paragraphs, 14 paragraphs had to be included in the S-Test in order to develop ninety schema-based cloze multiple choice items.

Since the 100-item C-Test and 90-item S-Test shared five paragraphs they were treated as tests and retests to be administered to the same test takers, respectively. To overcome the learning and practice effects (Bachman, 1990), Rubin and Babbie's (2011) suggestion regarding an interval of two weeks was followed. The C-Test was administered first and S-Test was held after two weeks. The answers given to the C-Test were manually marked by accepting the exactly restored second halves only.

2.5 Data Analysis

The reliability of C-Test and S-Test was estimated by employing Cronbach's Alpha. The internal validity of the two tests was determined by utilizing the item facility (IF) and item discrimination (ID) indices. While the IF index was obtained as the proportion of correct responses given to each item, each individual item was correlated by the total score to obtain the biserial correlation coefficients as ID indices. Items having the IF indices falling between 0.25 to 0.75 (Baker, 1989) and ID indices equal to and higher than 0.25 were considered well functioning. The C-Test and S-Test as well as their subtests were correlated by employing Pearson correlation formula. All the statistical analyses were carried out via the IBM SPSS statistics to test the three hypotheses below

H1. The C-Test will be internally as valid as the S-Test.

H2. The C-Test and S-Test will correlate significantly with each other.

3. Results

Table 2 presents the descriptive statistics as well as the reliability estimates of the C-Test and S-Test developed in this study. As can be seen, both schema-based tests are reliable. The alpha reliability coefficient of C-Test (.83) is, however, higher than that of the S-Test, i.e., .75. The superior reliability level of the former might partly be attributed to its having ten more items as well as its being easier than the latter.

The mean IF index of the C-Test is .56 whereas that of the S-Test is .44, indicating that PUHS students could answer more C-Test items. The less challenging nature of C-Test becomes more obvious when the mean IF of syntactic items on the C-Test and S-Test are compared with each other, i.e., 0.61 vs. 0.44, respectively. As it can also be seen in Table 2, the syntactic subtest of C-Test is as reliable as the whole S-Test, i.e., 0.75.

Table 2: Descriptive statistics and reliability estimate of C-Test, S-Test and their subtests

Test	Subtest	# of items	% of items	Mean	SD	Mean IF	Mean ID	# of WFI	Alpha
C-Test	Semantic	47	47.0	24.36	5.683	.5183	.2211	20	.66
	Syntactic	48	48.0	29.06	6.376	.6055	.2511	27	.75
	Parasyntactic	5	5.0	2.40	1.189	.4808	.2064	0	.17
	Total	100	100.0	55.83	11.402	.5583	.2348	47	.83
S-Test	Semantic	55	61.1	24.64	5.892	.4479	.2035	19	.64
	Syntactic	35	38.9	15.26	4.144	.4361	.2090	10	.53
	Total	90	100.0	39.90	9.066	.4433	.2056	29	.75

In addition to superior internal consistency, the C-Test enjoys higher internal validity because out of 100 items, 47 have functioned well, i.e., almost fifty percent of C-Test items have had acceptable levels of difficulty and discrimination power. However, the percentage falls to 32%, i.e., 29 out of 90 items in the case of S-Test. These results thus *reject* the first hypothesis and establish C-Test as achievement measures whose internal validity is higher than that of S-Test.

Table 3 presents the correlation coefficient obtained among the C-Test, S-Test and their subtests. As can be seen, the C-Test correlates almost perfectly with the S-Test, i.e., r=.99, *p*<.01, explaining 98% of its variance. Similarly C-Test correlates significantly with the semantic and syntactic subtests of the S-Test, i.e., r=.92 and .85 (*p*<.01), explaining approximately 85% and 72% of variance in them, respectively. These results not only *confirm* the second hypothesis that *the C-Test and S-Test will correlate significantly with each other* but also measure the same ability, i.e., EFL achievement, because their r is "in the .80 to 1 range" (Hatch & Lazaraton, 1991, p. 442)

Table 3: Correlations among the C-Test, S-Test and their subtests

Tests	C-Test				S-Test		
	Total	Semantic	Syntactic	Parasyntactic	Total	Semantic	Syntactic
C-Test	1	.900**	.906**	.429**	.986**	.917**	.854**
Semantic	.900**	1	.646**	.382**	.888**	.837**	.752**
Syntactic	.906**	.646**	1	.240**	.891**	.819**	.786**
Parasyntactic	.429**	.382**	.240**	1	.435**	.398**	.386**
S-Test	.986**	.888**	.891**	.435**	1	.934**	.860**
Semantic	.917**	.837**	.819**	.398**	.934**	1	.620**
Syntactic	.854**	.752**	.786**	.386**	.860**	.620**	1

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

Along with the C-Test itself, its semantic subtest correlates significantly with the semantic subtest of S-Test at such a high level which makes them interchangeable, i.e., r=.84, *p*<.01. Surprisingly, however, the syntactic subtest of C-Test shows more significant relationship with the

semantic subtest of S-Test, i.e., r=.82, *p*<.01, than with the semantic subtest of C-Test, i.e., r=.64, *p*<.01. Similarly, the parasyntactic subtest of C-Test correlates more with the syntactic subtest of the S-Test, i.e., r=.39, *p*<.01, than with the syntactic subtest of C-Test, i.e., r=.24, *p*<.01, indicating that S-Tests provide better measures of the PUHS students' syntactic and parasyntactic schemata than C-Tests do.

4. Discussions

To the best knowledge of the present researchers no study has explored the validity and reliability of C-Tests as measures of language achievement in general and achievement at secondary education in particular. This study is, therefore, the first in developing these tests on seen passages to fulfill content-based achievement testing within a secondary education context. It is a pioneering study because it bases C-Tests on schema theory which approaches language learning as an ever-developing process in which learners develop and modify their schemata as they lead their daily lives.

The results of the present study show that C-Tests correlate at the highest possible level with S-Tests and explain 97.2 percent of variance in each other. Although C-Tests and S-Tests reveal the highest possible level of relationship with each other and seem to be measuring the same variable (e.g., Hatch & Lazaraton, 1991), they should be employed together at best in order to help learners develop and improve their schemata as comprehensively as possible. Because of their nature, conventional C-Tests suffer from a number of shortcomings which make their exclusive use as measures of content-based achievement questionable.

The first shortcoming of C-Tests is their dependence on schema tokens rather than schema types as shown in Table 4. Out of 100 items comprising the C-Test, six, five and four have been developed on three syntactic schemata "and", "that" and "in", respectively, indicating that 15 out of 100 items are testing students' knowledge of three most frequent schemata. Thus the analysis of C-Test items in terms of their types provides a totally different picture of what constitutes C-Tests. As it can also be seen in Table 4, 69 schema types constitute the conventional C-Test out of which only 37 are semantic in nature, i.e., Is, Bring(s), Do, Exercise, Winters, Average, Be, Becomes, Body's, Burns, Calories, Change, Changing, Child, City, Climate, Cold, Countries, Deep, Devote, Energy, Give, Growing, Include, Labor, Need, Nutrients, Observed, Oxygen, Permission, Poor, Poverty, Repetitive, School, Suffering, Tells and Walls.

Table 4: Schema types and tokens (T) restored as C-Test items

Schema	T	Schema	T	Schema	T	Schema	T	Schema	T
And	6	Winters	2	Calories	1	Give	1	Or	1
That	5	You	2	Change	1	Growing	1	Oxygen	1
In	4	Africa	1	Changing	1	Has	1	Permission	1
Is	4	All	1	Child	1	Her	1	Poor	1
The	4	America	1	City	1	Include	1	Poverty	1
To	3	Among	1	Climate	1	It	1	Repetitive	1
Your	3	An	1	Cold	1	Labor	1	School	1
About	2	Asia	1	Countries	1	Mainly	1	She	1
Bring(s)	2	Average	1	Deep	1	Million	1	Suffering	1
But	2	Be	1	Devote	1	Mostly	1	Tells	1
Do	2	Becomes	1	Energy	1	Need	1	Walls	1
Exercise	2	Between	1	Europe	1	Nutrients	1	When	1
Like	2	Body's	1	Example	1	Observed	1	Whether	1
Of	2	Burns	1	For	1	On	1		

In contrast to the C-Tests whose semantic items make up only 37% of the whole measure, 55 out of 90 items (61.1%) comprising the S-Test are semantic in nature, i.e., Amounts, Average, Body, Breathing, Brushing, Common, Crossing, Devote, Distant, Distracting, Easily, Efficiently, Estimated, Exactly, Extinction, Extreme, Fields, Fresh, Functions, Gestures, Heart, Height, Include, Invention, Knowing, Light, Look, Look after, Makes, Mean, Mild, Mostly, Move, Necessary, Needing, Number, Officials, Poverty, Practice, Presentation, Rattle, React, Received, Relaxed, Religious, Repetitive, Rise, Task, Taught, Time, Tired, Twice, Various, Voluntary, and Weigh. The high percentage of S-Test items requiring the selection of 55 keyed responses from among 220 semantic schemata secures the representativeness of the measure. Out of the 90 items comprising the test, only one syntactic schema, i.e., when, has a token of two, i.e., items 5 and 74.

The second shortcoming of C-Tests is the context-independence of some of their items. This was first brought up by Khodadady (2012) who took out the 99-items comprising Klein-Braley's (1997) C-Test and presented them as a list of 99 mutilated words to 430 undergraduate and graduate students of English. He provided the test takers with the directions given to C-Test, namely, to restore the second half of the words by adding the same number of letters given or one more. His results showed that 13 items functioned well, indicating that the test takers could restore those items without having any context to base their restoration on.

C-Tests require test takers to restore their mutilated items by writing down their answers. As the third shortcoming of C-Tests, writing down the restored parts of the mutilated words questions the construct validity of measuring reading comprehension ability as a receptive process. In other words, C-Tests measure not only reading comprehension ability but also spelling knowledge of test takers. These deficiencies may contribute to the loading of C-Tests on a factor other than the rotated latent variable upon which S-Tests and TOEFL (Educational Testing Service, 1991) load, leading Khodadady (2007) to call them method-specific measures of language proficiency.

While C-Tests are developed on every second word of the second sentence onwards, S-Tests are designed on schema types which are specified by analyzing the syntactic, semantic and parasyntactic domains of the texts upon which they are developed. Khodadady (2009) was the first researcher who developed an S-Test on some passages suggested for outside reading to measure the objective-based achievement of undergraduate university students. The present study shows that S-Tests can also be developed on the passages taught during school years to provide PUHS teachers with valid and reliable measures of their students' content-based achievement.

As the fifth shortcoming, C-Tests are developed only on single-word schemata whereas S-Test items can also be developed on phrasal word schemata such as "nor do we know" measuring the test takers' syntactic knowledge as regards their application of the syntactic-verb "do" after the conjunction "nor" when it connects two sentences, e.g., item 45 of the S-Test developed in this study. The analysis of statistics on this item shows only 26 percent of test takers could get this item right leading to its very low ID index of .10 and the relatively low alpha coefficient of S-Test compared to C-Test. In other words, S-Tests are superior to C-Tests in terms of measuring syntactic structures and helping the instructors diagnose their learners' linguistic weaknesses. The same argument holds true for the superiority of S-Tests in terms of their text representativeness.

While C-Tests are developed on a few, mostly four or five, paragraphs as their last shortcoming, S-Tests can be developed on as many paragraphs as necessary. In the present study, for example, one paragraph from each of the eight lessons taught during the school year was included in the S-Test. Since the S-Tests are usually developed based on schematic analysis of texts, they provide a representative sample of schemata taught as well and can therefore be utilized to distinguish low-ability achiever from those of high-ability. This very distinctive feature of S-Tests secures their fairness and provides instructors with objective diagnostic measures to base their future educational programs on.

5. Conclusions

The findings of the present study show that C-Tests can be developed on the passages taught during school years to provide valid and reliable measures of content-based achievement of PUHS students. Future research must, however, show whether they fulfill the same function in other grades of high schools as well. Although they correlate almost perfectly with S-Tests and thus establish themselves as schema-based measures of language achievement, they are suggested to be used along with other measures because they are limited by their very nature. One hundred C-Test items are usually developed on every second word of the second sentences of a few paragraphs and thus test designers have no choice but to limit their selection of passages by choosing some randomly. In contrast to the C-Tests, S-Tests, however, are designed on schema types rather than tokens, enabling teachers to choose at least one paragraph from each lesson covered and thus secure the representativeness of what they measure as their learners' achievement.

For specifying the types and number of schemata to be chosen as S-Test items, the teachers need to identify and codify the schemata constituting the texts they teach during school years to determine their domains, genera and species. This process by itself provides the language teachers not only with the most empirical criterion to define their course objectives but also with the schemata upon which they spend most of their class time. The findings of the present study, for example, show that 1578 schemata are taught to PUHS students 1174 (74.4%) of which are semantic in nature. The inclusion of what they have taught in the content of what they test secures fairness and allows the teacher to evaluate their teaching outcomes and plan for future instruction. It is, therefore, suggested that the schemata comprising the English textbooks be specified and S-Tests as well as C-Test be developed on these textbooks to measure content-based achievement as objectively and as comprehensively as possible.

References

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

Babaii, E., & Moghaddam, M. J. (2006). On the interplay between test task difficulty and macro-level processing in the C-test. *System*, 34, 586-600.

Birjandi, P., Sarab, A. M. R., & Samimi, D. (2012/1391). *Learning to read English for pre-university students*. Tehran: CHAP VA NASHRE KETABHAYEH DARSI IRAN.

Birjandi, P., Nouroozi, M., & Mahmoodi, G. (2010). *English book 3*. Tehran: CHAP VA NASHRE KETABHAYEH DARSI IRAN.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, Massachusetts: The MIT Press.

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

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

Ghaffary, C. M. (2000). *Schemata-based multiple choice item test and measuring achievement at pre-university centers*. Unpublished MA thesis, Islamic Azad University of Garmsar.

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

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

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

Khodadady, E. (2009). Objective-based achievement testing in the context of schema theory. *Iranian Journal of Language Studies* (IJLS), 3(1), 1-30.

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.

Khodadady, E. (2013). Authenticity and sampling in C-Tests: A schema-based and statistical response to Grotjahn's critique. *The International Journal of Language Learning and Applied Linguistics World* (IJLLALW), 2(1), 1-17.

Khodadady, E., & Alaee, F. F. (2012). Exploring the relationship between social capitals and English language achievement within a specific grade and context. *English Language Teaching*, 5 (2), 701-717.

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.

Khodadady, E., & Hashemi, M. (2011). Validit and C-Tests: The role of text authenticity. *Iranian Journal of Language Testing*, 1(1), 30-41.

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.

Khodadady, E., & Javadi Mehr, S. (2012). Schema-based analysis of gendered self-disclosure in Persian: Writing for dating context. *English Language Teaching*, 5(1), 20-31, doi:10.5539/elt.v5n1p20.

Khodadady, E., & Khaghaninejad, M. S. (2012). Acquisition of French polysemous vocabularies: schema-based instruction versus translation-based instruction. *Porta Linguarum*, 17, 29-46.

Khodadady, E., Alavi, S. M., & Khaghaninejad, M. S. (2011). Schema-based instruction: A novel approach to teaching English to Iranian university students. *Ferdowsi Review*, 2(1), 3-21.

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

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

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

Lagzian, M. (2012). *Textual analysis of an English dentistry text and its translation in Persian: A schema-based approach*. Unpublished MA thesis, Ferdowsi University of Mashhad.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. London: Longman.

Raatz, U., & Klein-Braley, C. (1981). The C-test- a modification of the cloze procedure. In Culhane, T., C. Klein-Braley, & D. Stevenson, (Eds.), *Practice and problems in language testing: University of Essex Occasional Paper* (113-148). Colchester: University of Essex.

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.

Taylor, W. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, 30, 415-33.

Valizadeh, M. (trans.). (1384). RADIOLOGY DAHAN: OSOOL WA TAFSIR. Tehran: Teimourzadeh.

Vallete, R. M. (1977). *Modern language testing* (2nd ed.). New York: Harcourt Jovanovich Brace.

White, S. C., & Pharoah, M. J. (2004). *Oral radiology: Principles and interpretation* (5th ed.). St. Louis, Missouri: Mosby.

Appendix A

C-Test

نام و نام خانوادگی : ....

دستورالعمل: در پنج متن زیر حروف نیمه دوم صد کلمه حذف شده است و به جای آنها خطوطی به تعداد حروف حذف
شده داده شده است. مثلا دو خط در _ _ Bo در جمله The bo _ _ _is on the shelf نشان می دهد که شما باید دو حرف ok را به
آن اضافه کنید تا کلمه درست book کامل شود. دقت کنید که :
1. اضافه کردن حروف باید با توجه به متن انجام شود و بنابراین در مثال بالا کلمه body یا boil بدلیل بی ارتباط بودن به
معنی متن درست نمی باشد.
2. فقط یک کلمه باید برای هر جواب نوشته شود.
3. جواب داده شده باید از لحاظ املایی درست باشد.

Aerobics

Aerobics is a word for needing oxygen, and aerobic exercise is any kind of activity that makes your muscles use oxygen. Aerobic [1]**exer**_ _ _ _ is[2]**repet**_ _ _ _ _, meaning [3]**i**_ is [4]**a**_ activity [5]**th**_ _ you [6]**d**_ over [7]**a**_ _ over, [8]**t**_ keep [9]**brin**_ _ _ _ in [10]**oxy**_ _ _ to [11]**a**_ _ of [12]**yo**_ _ muscles.[13]**Wh**_ _ you [14]**d**_ aerobic [15]**exer**_ _ _ _ and [16]**br**_ _ _ in[17]**th**_ _ oxygen, [18]**yo**_ _heart[19]**bec**_ _ _ _ stronger ([20]**a**_ _ even a bit bigger!).

Nutrition

Every time you eat food, your body does the same thing: it uses some of the nutrients in the food as fuel. It [21]**bu**_ _ _ these [22]**nutr**_ _ _ _ _ to[23]**gi**_ us[24]**ene**_ _ _ or [24]**calo**_ _ _ _. You [26]**ne**_ _ calories [27]**f**_ _ all [28]**o**_ your [29]**bo**_ _ _'_functions,[30]**whe**_ _ _ _it [31]**i**_things [32]**y**_ _ think [33]**ab**_ _ _ doing, [34]**li**_ _brushing[35]**yo**_ _ teeth, [36]**o**_ things[37]**y**_ _ never think [38]**ab**_ _ _ doing,[39]**li**_ _breathing.[40]**B**_ _ if the body is not able to use all the calories that comes from food, it stores them as fat.

Climate

Climate is the average of a region's weather over a period of time. For [41]**exa**_ _ _ _, it [42]**i**_ possible [43]**th**_ _ a [44]**win**_ _ _ day [45]**i**_ a [46]**ci**_ _ could [47]**b**_ sunny[48]**a**_ _mild,[49]**b**_ _ the [50]**ave**_ _ _ _ weather [51]**te**_ _ _ us [52]**th**_ _its[53]**win**_ _ _ _ will [54]**mai**_ _ _ be [55]**co**_ _ and [46]**inc**_ _ _ _ snow [57]**a**_ _ rain. [58]**Cli**_ _ _ _ change [59]**i**_ a [60]**cha**_ _ _ inthese general weather patterns.

Child Labor

Child labor is more common in developing countries, but it also exists in industrialized nations. While [61]**chi**_ _ labor [62]**mos**_ _ _ exists [63]**i**_ South [64]**a**_ _ Southeast [65]**As**_ _, South [66]**Ame**_ _ _ _, and [67]**Afr**_ _ _, it [68]**i**_ also a [69]**gro**_ _ _ _ concern [70]**i**_ Eastern [71]**Eur**_ _ _where[72]**count**_ _ _ _ are [73]**chan**_ _ _ _ economically. [74]**T**_ _ International [75]**La**_ _ _ Organization [76]**h**_ _estimated[77]**th**_ _ about 250 [78]**mil**_ _ _ _ children, [79]**bet**_ _ _ _ the [80]**a**_ _of five and fourteen, work in poor countries- at least 120 million on a full time basis.

Mother Teresa

From 1929 to 1948 Mother Teresa taught at St. Mary's High School in Calcutta. The [81]**suff**_ _ _ _ and [82]**pov**_ _ _ _ she [83]**obse**_ _ _ _ outside [84]**t**_ _ convent [85]**wa**_ _ _ made a[86]**de**_ _ impression [87]**o**_ her. [88]**I**_ 1046, [89]**s**_ _received[90]**permi**_ _ _ _ _ from [91]**h**_ _ superiors [92]**t**_ leave [93]**t**_ _ convent [94]**sch**_ _ _ and [95]**dev**_ _ _ herself[96]**t**_ working [97]**am**_ _ _ the [98]**po**_ _ in [99]**t**_ _slums [100]**o**_ Calcutta.

Appendix B

S-Test

<u>Direction</u>: Choose the best answer and mark it on the answer sheet

Aerobics is a word for … (1) oxygen, and aerobic exercise is any kind of activity that … (2) your muscles use oxygen. Aerobic exercise is … (3), meaning it is an activity that you do over and over, to keep bringing … (4) oxygen to all of your muscles. … (5) you do aerobic exercise and bring in that oxygen, your … (6) becomes stronger (and even a bit bigger!). The … (7) of blood cells in your blood increases, … (8) the blood can carry even more oxygen. The blood in your body even moves more … (9) through the blood vessels. All these things … (10) that your body works more … (11) to keep you … (12), and you can do a lot of exercise without getting … (13).

1	A	lacking	B	needing	C	avoiding	D	missing
2	A	makes	B	builds	C	produces	D	creates
3	A	competitive	B	boring	C	repetitive	D	tiring
4	A	different	B	additional	C	juicy	D	fresh
5	A	When	B	Where	C	What	D	Which
6	A	lung	B	heart	C	muscle	D	chest
7	A	size	B	amount	C	number	D	degree
8	A	so	B	as	C	such	D	though
9	A	purely	B	clearly	C	simply	D	easily
10	A	imply	B	mean	C	intend	D	mention
11	A	capably	B	emotionally	C	efficiently	D	completely
12	A	healthfully	B	healthful	C	health	D	healthy
13	A	tired	B	bored	C	finished	D	ended

Every … (14) you eat food, your body does the same thing: it uses some of the nutrients in the food as fuel. It burns these nutrients to give … (15) energy or calories. You need calories for all of your body's … (16), whether it is things you think about doing, like … (17) your teeth, or things you never think about doing, like … (18). But if the body is not able to use … (19) the calories that … (20) from food, it stores them as fat. Exercise helps keep you at a … (21) that is right for your … (22), by burning up extra calories. When you exercise, your body uses that extra fuel to keep you … (23).

14	A	season	B	time	C	period	D	stage
15	A	our	B	we	C	us	D	ours
16	A	functions	B	positions	C	characters	D	roles
17	A	scraping	B	touching	C	sweeping	D	brushing
18	A	breaking	B	breathing	C	blowing	D	blending
19	A	several	B	some	C	all	D	any
20	A	is coming	B	comes	C	coming	D	are coming
21	A	weight	B	weigh	C	weighty	D	weightily
22	A	width	B	length	C	height	D	sight
23	A	strength	B	strong	C	strongly	D	strengthen

Good eye contacts help your audience feel more … (24) and builds confidence in your speaking ability. Keep eye contact by … (25) your speech so well that you need to have a quick … (26) at your notes only from time to time.

24	A	comforted	B	admired	C	noticed	D	relaxed
25	A	recognizing	B	understanding	C	knowing	D	familiarizing
26	A	look	B	watch	C	observe	D	stare

Find a few … (27) faces in the audience that … (28) to your message and concentrate on giving your speech to … (29). Keep eye contact for four to five seconds at a time, and then … (30) to someone else.

27	A	befriend	B	friends	C	friendship	D	friendly
28	A	contact	B	react	C	act	D	detract
29	A	theirs	B	they	C	them	D	themselves
30	A	transport	B	travel	C	walk	D	move

Be aware of habits you might have like … (31) your arms, leaning … (32) a wall, or tapping a pen. This might also be … (33) to your audience or might tell them that you are uninterested or unconfident.

31	A	resting	B	stretching	C	crossing	D	washing
32	A	against	B	on	C	upon	D	above
33	A	contracting	B	distracting	C	retracting	D	protracting

… (34) your speech in front of a mirror to check your way of … (35) and body movements and change them if … (36). Try many different ways to find a comfortable balance of … (37) to use in front of an audience.

34	A	Practice	B	Exercise	C	Drill	D	Perform
35	A	presenting	B	representation	C	presentation	D	representing
36	A	critical	B	necessary	C	crucial	D	obligatory
37	A	tokens	B	motions	C	gestures	D	signals

Climate is the … (38) of a region's weather over a period of time. For example, it is possible that a winter day in a city could be sunny and … (39), but the average weather tells us that … (40) winters will mainly be cold and … (41) snow and rain. Climate change is a change in … (42) general weather patterns. They can become warmer or colder; … (43) of rainfall or snowfall can increase or decrease each year.

38	A	average	B	normal	C	usual	D	custom
39	A	gentle	B	calm	C	quiet	D	mild
40	A	its	B	it's	C	it	D	itself
41	A	comprise	B	include	C	consist	D	involve
42	A	their	B	theirs	C	these	D	them
43	A	sizes	B	qualities	C	numbers	D	amounts

Global warming brings with it no guarantees. We don't know … (44) what will happen– what the effects will be– … (45) exactly where or when they will make problems. But it should … (46) that scientists have a … (47) idea of what is going to happen. Scientists and researchers from different … (48) tell us that the possible effects of climate change could be big and, in some cases, would cause serious problems. … (49) the possible effects are increased number of human deaths, … (50) of groups of animals and plants, and a dangerous … (51) in sea levels. With this in mind, we have … (52) of the costs of action and … (53) them against the risks of inaction.

44	A	carefully	B	closely	C	exactly	D	faithfully
45	A	do we nor know	B	nor do we know	C	we do know nor	D	know do we nor
46	A	be said	B	say	C	are said	D	be saying
47	A	pretty and good	B	good and pretty	C	good pretty	D	pretty good
48	A	fields	B	subjects	C	matters	D	topics
49	A	Between	B	Within	C	Among	D	Middle
50	A	destruction	B	extinction	C	elimination	D	abolition
51	A	raise	B	arise	C	reprise	D	rise
52	A	think	B	thought	C	to think	D	been thought
53	A	weigh	B	reflect	C	assess	D	meditate

Earthquakes happen every day. Though millions of people … (54) an earthquake, it is a very … (55) happening on this planet. So today, … (56), an earthquake will occur. It may be so … (57) that only special instruments can record its movement; it may shake houses, … (58) windows, and change the place of small things. It may be … (59) to cause damage, injury, and … (60)

54	A	never may experience	B	may never experience	C	experience may never	D	may experience never
55	A	average	B	frequent	C	popular	D	common
56	A	somewhere	B	whereon	C	anywhere	D	whereas
57	A	shiny	B	bright	C	light	D	intense
58	A	bang	B	rattle	C	crash	D	collapse
59	A	strong so	B	too strong	C	strong enough	D	very strong
60	A	death	B	die	C	dead	D	deadly

Most child laborers around the world are busy doing … (61) forms of work that are dangerous for their health. They are also … (62) of their rights, including not only the right to develop to the … (63) level through education, but also the right to the … (64). They often work as … (65) as 12 hours a day, and use tools and machines … (66) are not designed for them.

61	A	severe	B	extreme	C	ultimate	D	excessive
62	A	robbing	B	having robbed	C	been robbed	D	being robbed
63	A	high	B	highest	C	highly	D	higher
64	A	childhood	B	children	C	childish	D	childishly
65	A	most	B	much	C	many	D	more
66	A	when	B	who	C	what	D	which

Child labor is more common in … (67) countries, but it also exists in industrialized nations. While child labor … (68) exists in South and Southeast Asia, South America, and Africa, it is also a growing concern in Eastern Europe … (69) countries are changing economically. The International Labor Organization has … (70) that about 250 million children, between the age of five and fourteen, work in poor countries- at least 120 million on a full time basis.

67	A	developed	B	developing	C	developmental	D	development
68	A	amply	B	adequately	C	mostly	D	abundantly
69	A	whose	B	which	C	when	D	where
70	A	valued	B	estimated	C	rated	D	evaluated

In 1906, the Italian scientist, Galileo, was probably the first person to use a new … (71) - the telescope- to observe the sky. A telescope is an instrument that magnifies, or makes larger, … (72) objects. With this telescope Galileo observed the moon and some mountains, valleys, and craters that … been seen (73) before.

71	A	invention	B	discovery	C	innovation	D	detection
72	A	isolated	B	distant	C	vague	D	obscure
73	A	have never	B	never have	C	had never	D	never had

The modern age of space exploration began in 1957, … (74) the former Soviet Union launched Sputnik *I*, an artificial satellite. A satellite is any natural … (75), like the moon, or any artificial object that orbits … (76) object. Sputnik, which was about … (77) the size of a soccer ball, carried instruments … (78) the density and temperature of the Earth's upper atmosphere.

74	A	whose	B	when	C	which	D	where
75	A	body	B	figure	C	character	D	outline
76	A	the other	B	other	C	another	D	others
77	A	twice	B	twin	C	double	D	identical
78	A	measure	B	measured	C	measuring	D	to measure

From 1929 to 1948 Mother Teresa … (79) at St. Mary's High School in Calcutta. The suffering and … (80) she observed outside the convent walls made a … (80) impression on her. In 1046, she … (82) permission from her superiors to leave the convent school and … (83) herself to working among the … (84) in the slums of Calcutta. Although she had no money, she started an open-air school for homeless children. Soon … (85) helpers joined her, and financial support came from … (86) church organizations, as well as from the city … (87). In 1950, she was permitted to start her own … (88) community "The Missionaries of Charity". Its … (89) was to care for those persons nobody was prepared to … (90).

79	A	taught	B	instructed	C	trained	D	coached
80	A	deficiency	B	poverty	C	scarcity	D	shortage
81	A	deeply	B	depth	C	deep	D	deepen
82	A	deceived	B	received	C	conceived	D	perceived
83	A	enforce	B	apply	C	allocate	D	devote
84	A	poor	B	poorly	C	poorness	D	poverty
85	A	amateur	B	intentional	C	deliberate	D	voluntary
86	A	changeable	B	various	C	dissimilar	D	fluctuating
87	A	approvals	B	officers	C	officials	D	authors
88	A	religious	B	political	C	scientific	D	social
89	A	function	B	career	C	profession	D	task
90	A	look for	B	look down	C	look after	D	look up

This is the end of the test.