The C-Test: A Valid Measure to Test Second Language Proficiency?

Fahimeh Khoshdel-Niyat

The C-Test: A Valid Measure to Test Second Language Proficiency?

Fahimeh Khoshdel-Niyat

English Department, Mashhad Branch, Islamic Azad University, Mashhad, Iran

E-mail: fahime_khoshdel@yahoo.com

Abstract

The C-Test is a gap-filling test belonging to the family of the reduced redundancy tests which is used as an overall measure of general language proficiency in a second or a native language. There is no consensus on the construct underlying the C-Test and many researchers are still puzzled by what is actually activated when examinees take a C-Test. The purpose of the present study is to cast light on this issue by examining the factors that contribute to C-Test item difficulty. A number of factors were selected and entered into regression model to predict item difficulty. Linear logistic test model was also used to support the results of regression analysis. Findings showed that the selected factors only explained 12 per cent of the variance in item difficulty estimates. Implications of the study for C-Test validity and application are discussed.

Keywords: *C-Test, Validation, Construct identification, LLTM*

Introduction

The C-Test is a text completion test and is based on the reduced redundancy principle. The noise test and oral cloze procedures are other kinds of reduced redundancy tests. These tests were developed on the basis of the assumption that natural languages are ‘redundant’. “This means that in natural communication messages contain elements which are not necessary” (Baghaei, 2011, p.7). According to information theory principles, redundancy can be reduced by eliminating words from a text and asking the learner to fill in the gap.

The C-Test is a variation of the cloze test and thus has the same basic theoretical assumptions as the cloze test **(**Grotjahn & Stemmer, 2002). The difference is that in C- Test parts of words are omitted not the whole words. Cloze test is an appropriate instrument for measuring general language proficiency as Oller (1976) concluded. The C-Test is based on the reduced redundancy principle (Spolsky, 1969), i.e., the assumption that natural languages are redundant, so advanced learners can be distinguished from beginners by their ability to deal with reduced redundancy (Beinborn, Zesch, & Gurevych, 2014).

Raatz and Klein-Braley (1981) suggested C-Tests because of several problems of cloze tests. These problems were:

(1) cloze tests should be long in order to have sufficient number of items; (2) cloze tests usually contain one longer text because of deletion principles. Therefore, it makes the test specific and also biased; (3) validity and reliability of cloze tests will be affected by some factors such as “text”, “deletion rate” and “starting point of deletion”; (4) most of cloze tests are less reliable

than what they are assumed to be; (5) scoring in cloze test is based on two methods: exact method that gaps should be filled by exact words and acceptable method that gaps can be filled by any appropriate words; so, it is more subjective and consumes much more time; (6) the difficulty of deleted words depends on the grammatical structures and content of words (Alderson, 1979, 1983; Klein-Braley, 1981).

A frequent question about C-Tests is that how behaviors of testees could be measured while filling the gaps (Klein-Braley, 2002). Many investigations have focused on mental processes while working on a C-test (Fledmann & Stemmer, 1987; Grotjahn & Stemmer, 1985). To identify test-taking processes, three possible approaches are suggested: *statistical item analysis, text linguistic item analysis* and *analysis of individual performance* (Fledmann et al, 1986; Grotjahn, 1986, as cited in Klein-Braley, 2002).

All of the above mentioned approaches seek for mental processes which examinees undertake while answering the test items. The mental processes which are taped by the test should be in line with the construct of the test; therefore, to validate a given test, these mental processes could be investigated.

Construct identification is concerned with factors that are involved in the test content and the methods that are used for subjects' scores (Sigott, 2004). So, it reveals validity of the test by studying characteristics which affect test difficulty. Test difficulty should be measured by learners' ability to answer any item that can be easy or difficult based on test content and method features that any particular test has. In other words, item difficulty is the proportion of wrong responses for every item of a given test (Farhady, Jafarpur, & Birjandi, 1994).

We are going to investigate the validity of the C-Test as a test of general language proficiency through analyzing difficulty of the items in the framework of construct identification.

Cloze tests belong to the family of reduced redundancy tests and were proposed as a measure of text difficulty by Taylor (1953). In cloze tests, every 7th word or higher is deleted from a text (Brown, 1989). In these tests, a reading passage should have a familiar topic for a learner to engage with a text and must not be very difficult (Brown, Yamashiro, & Ogane, 2001; Douglas, 1998; Sasaki, 2002). The ambiguity of the gaps is the main problem in cloze tests, so it cannot be anticipated what the set of potential solutions are (Horsmann & Zesch, 2014). Hence, because of some criticisms toward cloze tests, C-Tests were developed as a replacement for cloze tests in 1981 by Raatz and Klein-Braley from both theoretical and psychometric viewpoint (Babaii & Moghaddam, 2006). Notably, "the letter C stands for cloze to call to mind the relationship between the two tests" (Baghaei, 2008a, p. 33).

The next is the oral cloze procedure. Despite a cloze test, written cloze passage is not given to testees in this procedure. When a mutilated passage records, it will presents acoustically. The blanks are numbered or there is a pause where a blank occurs. While listening to the material, students should guess the missing part. It can be said that the advantage of this cloze technique is that it can be used for non-literate subjects (Oller, 1979).

Background to the Theoretical and Conceptual Views

A C-Test includes several authentic texts, usually between four and six with twenty to twenty five gaps in each text. Each text usually is eighty to one hundred words in length with different content. Five minutes is allocated for each text, so a test with four texts would take twenty minutes, the one with five texts would take twenty five and so on. In the literature 20 to 25 gaps in each passage are suggested (Raatz & Kelein-Braley, 2002), however, Baghaei (2011a, 2011b) demonstrate that C-Test with smaller number of gaps work as well as 25-gap C-Tests. In C-Tests, the exact answer should be given, however in some occasions as Raatz and Kelein-Braley (2002) deduced, the alternative solution would be accepted. Numbers and proper names are usually left without any changing (Raatz & Kelein-Braley, 2002). Furthermore, content of texts should be neutral without any special vocabulary from the general knowledge domain. In other words, texts can be selected from newspapers, brochures, magazines, non-fictional books and so forth (Grotjahn & Stemmer, 2002).

In C-Tests, first and last sentences should remain without any deletion. Beginning at word two in sentence two, the second half of every second word is deleted (Raatz & Kelein-Braley, 2002).

So, according to Raatz and Kelein-Braley (2002) a C-Test is a kind of reduced redundancy test, because it uses the cloze principles as mentioned earlier, which is derived from Information Theory. It means a redundant massage includes more information than is necessary for understanding the massage. Hence, when a massage is damaged, the other parts that are intact can help to find what the complete massage is.

However, Koberl and Sigott (1994) and Sigott and Koberl (1996) developed two other different variations of a German and an English C-Tests: 1) Deletion of two thirds of every second word; 2) Deletion of all but first letter of every second word.

Deleting the first halves of words was the most reliable test in German, but not in English. The larger “half” of words will be deleted in words with an odd number in letter. This technique is called “rule of two” in C-Tests texts (Klein-Braley & Raatz, 2002). It has to be noted that in this research we will work on the original C-Test principle on test difficulty.

Furthermore, a C-Test has its own rules for construction. These are as follow:

1. Target population and the test format should be defined,

2. Appropriate texts should be chosen more than needed and then the best ones would be selected,

3. After selecting the best texts, they should be brought into C-Test format (rule of two),

4. Analyzing the difficulty of the texts,

5. It should be decided the satisfactory of each text by changing, adding or removing some damaged words, because some are so difficult or easy,

6. Then good texts should be combined,

7. Item analysis, reliability and validity of the test would be performed,

8. Test should be improved if it is needed,

9. The final form of the test should be administered to a sample of the target population,

10. The test norms should be calculated (Klein-Braley & Raatz, 2002).

C-Test Validation Studies

Validating C-Tests has been researchers' concern for several decades of research on C-Tests. C-Tests have been developed and validated for different groups of learners whether L1 learners, L2 learners or foreign language learners (Baghaei, 2014). There is ample convincing evidence for the validity of C-Tests as measures of general language proficiency. For example, it is found that C-Tests have a high correlation with other language tests such as teacher ratings and students self-assessment or with composite scores of various language skills. Factorial structure and its fit to the Rasch model are another evidence of C-Test validity (Baghaei, 2008a, 2008b, 2010, 2011; Eckes, 2006, 2011; Eckes & Baghaei, 2015; Eckes & Grotjahn, 2006; Raatz, 1984, 1985).

Moreover, Borgards and Raatz (2002) examined German C-Tests' sensitivity to a construct-irrelevant attribute which is named coaching effect. In their study, there were control and experimental groups of 43 secondary level students which in pretest experimental group was exposed to 45 minutes of coaching for C-Test taking. The Posttest demonstrated that mean of both control and experimental groups similarly increased in comparison with the pretest. It means that coaching effect did not influence on C-Test scores. So, it can be found out C-Tests measure general language proficiency.

Also, Baghaei (2014) developed and validated a Persian C-Test. The result of his study showed that C-Tests can be used as a general language proficiency for ages 12-14 of Persian speakers. Hence, it can be used as measurement of general language proficiency in Persian as a second or foreign language. The validity of C-Tests was proved because the C-Test could fit the Rasch model.

Another study of C-test validity is the role of the text type or text genre in C-Tests. For example, Shohamy and Inbar (1991) found that in a listening comprehension test, test takers' performance depends on text genre. "When they constructed identical listening comprehension questions on the basis of texts which contained exactly the same information but only differed in their genre, examinees' performances were noticeably affected" (Baghaei & Grotjahn, 2014a, p.163; Baghaei & Grotjahn, 2014b). Based on Baghaei, Monshi Tousi and Boori's (2009) and Baghaei and Grotjahn's (2014a) research, the text type influences the construct of C-Tests, so they developed C-Tests from spoken discourse texts to tap test takers' oral abilities.

In addition, Baghaei and Grotjahn (2014a; 2014b) worked on the componential structure of an English C-Test which included two spoken-discourse passages and two written-discourse passages by using unidimensional and multidimensional Rasch models. Their C-Test fitted Adams, Wilson, and Wang's (1997) multidimensional partial credit model better than Masters' (1982) unidimensional partial credit model. Hence, it revealed two separate dimensions: spoken discourse and written-discourse C-Test passages. Note that the sample size of 99 is the limitation of their study and it can be considered as a pilot study. The result showed that different C-Test text types of C-Test can measure different constructs. For example, spoken discourse C-Test texts may be better in testing students' listening/speaking skills than written discourse texts. Moreover, if a C-Test includes both spoken and written discourse, it can be a better measurement for C-Test construct as a general language proficiency.

Furthermore, as Sigott (2004) stated, C-Tests can be multidimensional because of various interpretation of C-Test scores : "the same C-Test passage could well be different tests for subjects at different levels of proficiency…without [the test user] knowing to what extent different aspects of the construct are reflected in the individual test scores" (p.203).

Construct Validity

Construct validity is one of the most complicated aspects of test validation. Not only is construct validity based on analyzing the test scores, but it also analyzes test performance

(Sigott, 2004). Messick (1975, p. 975) states: "A measure estimates how much of something an individual displays or process and also he defines validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (1989, p. 13). He also writes that there was a shift from prediction to explanation in validity concept. It means that obvious interpretation of test scores is much more important than ambiguous prediction. Baghaei (2007) also has a similar idea toward importance of score meaning and its interpretation in construct validity.

Besides, Baghaei (2007) focused on another crucial aspect of construct validity that should be paid attention to, i.e., 'construct irrelevant variance'. In test scores, there are always some irrelevant sub-dimensions that are not our purpose to measure but they affect the construct. Baghaei indicated two reasons for 'construct irrelevant variance'; these are construct-irrelevant easiness' and 'construct-irrelevant difficulty'. Construct-irrelevant difficulty means inclusion of some tasks that make the construct difficult and results in 'invalidly low scores' for some people. Construct-irrelevant easiness, on the other hand, is the easiness of the test because of inclusion of faulty items that give clues to individuals (in the case of multiple-choice items) who are familiar with the test format and can benefit from this (Baghaei, 2011).

Construct Identification Studies

"Construct representation is concerned with identifying the theoretical mechanisms that underlie item responses, such as information processes, strategies, and knowledge stores" (Embretson, 1983, p 179). Also, it can be considered as 'construct identification', and 'construct specification' (Perkins & Linville, 1987; Stenner, Smith, & Burdick, 1983). When a person scores higher than another one, it indicates that he/she processes more of the construct in question or an item that score higher in difficulty presumably demands more in construct (Stenner et al., 1983). Stenner et al. (1983) believed the process reveals "something" that happens while examines provide responses to items and is called "construct definition". Based upon some research such as Klein-Braley (1996), focus of construct identification is twofold: first one is investigating C-Test takers' psycholinguistic strategies; another one is predicting the difficulty of C-Test passages from text characteristics. The ability of responding to the item and method features affect the difficulty of the test, subtest or items.

According to Previous works, Klein-Braley (1984, 1985) used German C-Tests for 9 and 11 year old L1 German speakers and English C-Test for L1 German-speaking English students at Duisburg University based on multiple regression equation. She listed text characteristics as follows:

(1) number of words in text, number of different type of words,

(2) number of sentences in the text,

(3) type token ratio,

(4) average sentence length in syllabus,

(5) average number of word in sentence and average number of syllables in word.

The type-token ratio and the average sentence length in syllables were the best predictors of scores for English students. For German students, the type token ratio and the average number of words in the sentences were the best predictors. These results are to predict the difficulty of C-Test passages for special groups and cannot be generalized for other groups. In our study we will focus on using the theory of construct identification to predict factors that influence item difficulty of C-Tests.

Remarkably, Eckes (2011) focused on a Rasch model to compare different C-Tests by constructing a calibrated item bank for C-tests. “In a calibrated item bank the parameter estimates for all items in the bank have been placed on the same difficulty scale” (Szabó, 2008; Vale, 2006; Wright & Stone, 1999, as cited in Eckes, 2011). Besides, fit of data to latent trait model is evidence of existence of a construct underlying the responses and hence validity (Baghaei & Tabatabaee-Yazdi, 2016).

The present study

C-Tests, like any other tests, consist of several items with different item difficulties, so we should find out the factors which make items easier or more difficult. To this end, various factors that affect test difficulties should be studied. The factors such as (1) the frequency of the mutilated word (Brown, 1989; Sigott, 1995) , (2) whether the words are content or function words, (3) the length of the mutilated word, (4) the length of the sentence where the gap is (Klein-Braley, 1984) , (5) the number of propositions in the sentence where the gap is, (6) the propositional density (of the sentence where the gap is) , (7) inflections (Beinborn et al., 2014) , (8) text difficulty (as measured by Lexile) (www.lexile.com) , (9) the frequency of the word before the mutilate word, (10) the frequency of the word after the mutilate word, (11) text difficulty (p-values of texts) (Beinborn et al, 2014),(12) dependency among items (Beinborn et al, 2014), and (13) word class (noun, verb, adjective, adverb, pronoun, preposition, conjunction, and determiner) (Sigott, 1995).

We should bear in mind that the researcher predicted some factors that may affect item difficulty and some of them are found in C-Test literature. So, we are going to study the factors which affect the item difficulty in a C-Test.

Method

Participants and Setting

The participants in the present study were 352 undergraduate EFL students at Islamic Azad University of Mashhad and Neyshabour, Ferdowsi, Khayyam, and Binalood universities. Both male (N=108) and female (N=244) students participated in this research with the age range of 20 to 35 (M=20, SD=10.33). They were assured that their information would be confidential and they were appreciated for their cooperation.

Instrumentation

The instrument employed in this study was a C-Test with four texts. Each text had 25 gaps with different general knowledge content. In this C-Test the first and the last sentences remained without any deletions. Beginning at word two, in sentence two, the second half of every second word was deleted (Raatz & Kelein-Braley, 2002). The texts were selected from CAE (Norris & French, 2008) and FCE books (Norris, 2008). Furthermore, online Collin dictionary was used to get the Frequency of each word.

Procedure

The test was given to 352 EFL students at Islamic Azad University of Mashhad and Neyshabour, Ferdowsi, Khayyam, and Binalood universities. They were informed that they should read the instructions carefully and fill in the 25 mutilated words in each text (in general, 100 items) based on the available textual information in the passages. By computing the difficulty of individual gaps based on their answers, we could explore the factors that make an item more difficult or easier.

Based on the available literature, as mentioned earlier, 13 factors were selected for investigation. To compute item difficulty, participants had 20 minutes to answer all 100 items (gaps). Item difficulty was computed as the proportion of wrong answers. Furthermore, the participants were asked to write down their email addresses if they would like to be informed of the result of their tests.

Study Design and Analysis

In this study correlational analysis, multiple regression, one-way analysis of variance (AVONA), and linear logistic test modeling (LLTM) (Fischer, 1973) were used to analyze the data. First of all, dependent and independent variables were clarified. Item difficulty of C-Test items was considered as the dependent variable and the 13 word-level and text-level factors mentioned above were independent variables. The correlation coefficient between these 13 factors and item difficulty were computed. Next, item difficulty was regressed on the independent variables.

At last, LLTM was used to cross check the results of regression analysis. LLTM is an extension of the Rasch (Rasch, 1960/1980) model which imposes linear constraints on item parameters (Baghaei & Kubinger, 2015). It postulates that several basic parameters affect item difficulty. Therefore, the difficulty of each item is obtained from the sum of the difficulty of basic parameters. If the Rasch model based item parameters and the LLTM reconstructed item parameters are close, LLTM will have a good fit (Baghaei & Ravand, 2015; Baghaei & Kubinger, 2015). For example, to solve a mathematics question that needs subtraction, addition, and division, LLTM considers these four operations as the basic operations which influence the difficulty of the item (question). The difficulty of this item should be recovered by the adding the difficulty of the three operations needed to solve the item.

Noteworthy, the difficulty of this item is calculated with the standard Rasch model, too, and then it is compared with the LLTM reconstructed item difficulty. If these two are close to each other, we will get that those three basic parameters (subtraction, addition, and division) influence the overall difficulty of the item.

SPSS software version 22 was employed for correlational analysis and regression, and eRm (Mair, Hatzinger, & Maier, 2014) package in R version 3.11 (R Core Development Team, 2015) was used for analyzing LLTM.

Results

Descriptive Statistics

Table 1 displays minimum, maximum, means, and standard deviations of the 12 independent variables that are chosen as predictors of C-Test item difficulty in the present study.

Table 1: *Descriptive statistics for the predicators in the analysis*

	N	Minimum	Maximum	Mean	Std. Deviation
1.Frequency	100	3	5	4.61	.601
2.F.C	100	0	1	.56	.499
3.L.Word	100	2	10	4.93	2.114
4.L.Sentece	100	5	74	29.48	23.808
5.Proposition	100	1	10	3.87	3.368
6.P.Density	100	.06	.50	.1346	.07455
7.Inflection	100	0	1	.18	.386
8.Lexile	100	700	1170	980.00	183.264
9.Dependancy	100	0	1	.20	.402
10.Frequency before mutilated word	100	2	5	4.71	.556
11.Frequency after mutilated word	100	3.00	5.00	4.8800	.38350
12.p value	100	9	13	11.25	1.486
Valid N (listwise)	100				

This section of result presents whether frequency of the mutilated words and C-Test item difficulty have any significant relationship or not. Correlational analysis showed that there was a negative correlation between the two variables, *r*= -.248, *n*=100, *p* < .05. That is, as the word becomes more frequent in the language difficulty decreases.

Table 2: *An example of word frequency and item difficulty for some words*

	Dependent variable	Independent variable
Item	Item difficulty	Word frequency
1(students)	.36	4
2(selected)	.29	5
3(first)	.11	5
4(into)	.50	5
5(just)	.14	5

Table 3: *An example of content/function words and item difficulty of some words*

	Dependent variable	Independent variable
Item	Item difficulty	Content /function
1 (students)	.36	1
2(selected)	.29	1
3(first)	.11	1
4(into)	.50	0
5(just)	.14	0

Content words are usually nouns, verbs, adjectives, and sometimes adverbs. They help us to focus on content story and tell the addressees where to focus their attention. Whereas, function words are the words we use to make our sentences grammatically correct. Pronouns, determiners, prepositions, and auxiliary verbs are examples of function words (Bell et al., 2009). If we use them incorrectly, we will be poor speakers of English but our listeners can get the main idea of our speaking. Furthermore, function words do not give us main information, so we do not use them to attract others' attention or sometimes we ignore them while speaking.

In this section of result, content words were coded '1' and function words were coded '0'. Correlational analysis revealed that there was a positive correlation between these two variables, *r*= .216, *n*= 100, *p*< .05. This means that content words are more difficult to reconstruct than function words and contribute more to C-Test item difficulty.

Table 4: *An example of length of each word and item difficulty of some words*

	Dependent variable	Independent variable
Item	Item difficulty	Length of word
1 (students)	.36	8
2(selected)	.29	8
3(first)	.11	4
4(into)	.50	4
5(just)	.14	4

Table 4 illustrates the length of each word and its difficulty (for all 100 items the length of words were counted). There was a weak positive relationship between these two variables, r= .013, n= 100, p = .70. The researcher hypothesized that the *length of word* is a factor that may affect item difficulty. But the analysis showed that it has no effect on difficulty of C-Test items.

Table 5: An example of sentence length and item difficulty of some words

	Dependent variable	Independent variable
Item	Item difficulty	Sentence length
1 (students)	.36	24
2(selected)	.29	24
3(first)	.11	24
4(into)	.50	24
5(just)	.14	24

We should count the number of all words in each sentence for any mutilated word. Correlational analysis showed that there was a low negative correlation between sentence length and item difficulty, *r*= -.012, *n*= 100, *p*= .91. So, this factor was not a good predictor for difficulty of items. Therefore, the fourth null hypothesis is accepted.

Table 6: *An example of number of propositions and item difficulty of some words*

	Dependent variable	Independent variable
Item	Item difficulty	Number of propositions
1 (students)	.36	2
2(selected)	.29	2
3(first)	.11	2
4(into)	.50	2
5(just)	.14	2

Number of proposition is another predictor that is chosen to predict item difficulty. It is based on the number of verbs in a given sentence. By performing correlational analysis, it is found out that there is no significant correlation between item difficulty and number of propositions, *r* =-.54, *n*=100, *p* = .61. The fifth null hypothesis is accepted too.

Table 7: *An example of propositional density and item difficulty of some words*

	Dependent variable	Independent variable		
Item	Item difficulty	Sentence length	Number of propositions	Propositional density
1 (students)	.36	24	2	2/24
2(selected)	.29	24	2	2/24
3(first)	.11	24	2	2/24
4(into)	.50	24	2	2/24
5(just)	.14	24	2	2/24

For calculating the propositional density, as another predictor, sentence length and the number of propositions in each sentence are needed. Propositional density is computed by dividing the number of propositions in a sentence by the number of words in the sentence. We hypothesized that the higher the density the more difficult the item. However, there was no significant correlation between these two variables, *r*=-.024, *n*=100, *p*= .84. So, propositional density was not a good predictor and the sixth null hypothesis is accepted as well.

Table 8: *An example of inflections and item difficulty of some words*

	Dependent variable	Independent variable
Item	Item difficulty	Inflections
1 students)	.36	1
2(selected)	.29	0
3(first)	.11	0
4(into)	.50	0
5(just)	.14	0

In grammar inflection is the modification of a word to show different grammatical categories such as number, gender, voice, tense, etc., inflected forms were coded ‘1’ and otherwise ‘0’. As Table 8 illustrates, the researcher selected seven types of inflections (ed, s (plural), s (3rdperson), al (adjective), ly (adverb), er, est (superlative)). Correlational analysis revealed that inflections had no significant correlation with item difficulty, *r* =.13, *n*=100, *p*=.17. So, the seventh null hypothesis is accepted.

Table 9: *An example of text difficulty and item difficulty of some words in the four C-Test texts*

	Dependent variable	Independent variable
Some items(text 1)	Item difficulty (text 1)	Text difficulty (text 1)
1 (students)	.36	940
2(selected)	.29	940
3(first)	.11	940
4(into)	.50	940
5(just)	.14	940
Some items(text 2)	Item difficulty (text 2)	Text difficulty (text 2)
1(severals)	.79	1110
2(down)	.34	1110
3(rebuilt)	.77	1110
4(only)	.75	1110
5(second)	.23	1110
Some items(text 3)	Item difficulty (text 3)	Text difficulty (text 3)
1(listen)	.19	1170
2(that)	.40	1170
3(any)	.95	1170
4(worry)	.78	1170
5(then)	.47	1170
Some items(text 4)	Item difficulty (text 4)	Text difficulty (text 4)
1(ever)	.20	700
2(maker)	.90	700
3(but)	.47	700
4(them)	.37	700
5(behaved)	.86	700

Text difficulty was computed according to Lexile framework (www.lexile.com). It was found that there was no significant correlation between these two variable, *r*= .12, *n*=100, *p*= 0.28. Therefore, the eighth null hypothesis is accepted.

Table 10: *An example of dependency and item difficulty of some words*

	Dependent variable	Independent variable
Item	Item difficulty	Dependency
1 (students)	.36	0
2(selected)	.29	0
3(first)	.11	0
4(into)	.50	0
5(just)	.14	0

Dependency among items was identified by examining the relationship between item residuals after a Rasch model (Rasch 1960/1980) analysis. Residuals are unexplained variance in the data after the latent trait is factored out. Correlations among residuals indicate dependency beyond the effect of the latent trait. A Rasch analysis using Winsteps (Linacre, 2014) identified 12 items with high residual correlations. These items were coded '1' and the rest '0'. Correlational analysis showed no significant correlation between these two variables, *r*= .11, *n*=100, *p*=.25.

Table 11: *An example of frequency of the word before the mutilate word and C-Test item difficulty*

	Dependent variable	Independent variable
Item	Item difficulty	frequency of the wor before the mutilated word
1(school)	.36	5
2(were)	.29	5
3(the)	.11	5
4(trip)	.50	4
4(have)	.14	5

The researcher postulated that the frequency of words *before* the mutilated words may affect item difficulty. So, by using Collins online dictionary, the frequency of those words were found. Correlational analysis revealed that there was no significant correlation between item difficulty and the frequency of the word before the mutilate word (*r*= -.05, *n*=100, *p*=.62). Thus, the tenth null hypothesis is accepted.

Table 12: *An example of frequency of the words after the mutilated word and C-Test item difficulty*

	Dependent variable	Independent variable
Item	Item difficulty	frequency of the word after th mutilated word
1(who)	.36	5
2(for)	.29	5
3(school)	.11	5
4(space)	.50	5
5(completed)	.14	5

In the example above, the researcher predicted that frequency of words *after* the mutilated word may affect item difficulty. So, by using Collins online dictionary, the frequency of those words were found. Correlational analysis revealed that there was no significant correlation between item difficulty and frequency of the word *after* the mutilate word. The frequency of the word immediately after the mutilated word was not a good predictor of item difficulty either (*r*=.05, *n*=100, *p*=.65). Thus, the eleventh null hypothesis is accepted.

Table 13: *An example of (text difficulty) p. value. And C-Test item difficulty of some words in the four C-Test texts*

	Dependent variable	Independent variable
Some items(text 1)	Item difficulty (text 1)	p (text 1)
1 (students)	.36	13
2(selected)	.29	13
3(first)	.11	13
4(into)	.50	13
5(just)	.14	13
Some items(text 2)	Item difficulty (text 2)	p (text 2)
1(severals)	.79	9
2(down)	.34	9
3(rebuilt)	.77	9
4(only)	.75	9
5(second)	.23	9
Some items (text 3)	Item difficulty (text 3)	p (text 3)
1(listen)	.19	11
2(that)	.40	11
3(any)	.95	11
4(worry)	.78	11
5(then)	.47	11
Some items(text4)	Item difficulty (text 4)	p (text 4)
1(ever)	.20	12
2(maker)	.90	12
3(but)	.47	12
4(them)	.37	12
5(behaved)	.86	12

Table 13 revealed that the p-value or difficulty of each passage was computed considering each passage as a super-item or testlet. This was considered another measure of difficulty for each passage alongside the Lexile estimates. The correlation between individual C-Test item difficulties and passage difficulties was *r*=-.24, *n*=100, *p*< .05. That is, as the text gets easier items get easier too. However, since this measure of text difficulty is based on the difficulty of individual C-Test items within a text a finding a correlation is obvious.

Correlational Analysis

In summary, Table 14 displays all coefficients of correlation between all the variables in this study. As it is shown in table below there are three significant correlations:

- Frequency of mutilated word and item difficulty, *r*= -.24, *n*=100, *p*< .05.
- Function/ content words and item difficulty, *r*= .21, *n*=100, *p*<.05.
- Text difficulty as estimated using passage difficulty as measured by super item p-value and item difficulty, *r*= -.246, *n*=100, *p*< .05.

Table 14: *Correlational Analysis*

	1	2	3	4	5	6	7	8	9	10	11	12	13
1.Difficulty	-	**-.248***	**.216***	.013	-.012	-.054	-.024	.130	.123	-.050	.052	**-.246***	.106
2.Frequency		-	-.573	-.524	-.056	-.022	.045	-.387	-.024	-.127	-.122	-.005	.040
3.F.C			-	.521	.060	-.004	-.131	.411	.016	.335	-.008	-.079	-.073
4.L.Word				-	-.115	-.141	-.073	.648	-.116	.210	.024	.058	-.070
5.L.Sentece					-	.937	-.108	-.234	.567	.114	.184	-.157	-.102
6.Proposition							.089	-.257	.489	.152	.052	-.151	-.133
7.P.Density							-	-.081	.073	.123	.053	-.083	-.118
8.Inflection								-	-.285	.197	-.123	.030	-.044
9.lexile									-	.133	.175	-.569	-.107
10.frequency before mutil word										-	-.021	-.060	-.059
11.frequency after mutil word											-	.028	.095
12. p value												-	-.85
13.Dependancy													-

Regression

Standard multiple regression was used to estimate the contribution of the 11 independent variables in explaining C-Test item difficulty. The assumptions of multicollinearity and independence of residuals were first checked. Number of propositions was deleted because of high correlation with sentence length (*r*=.94).

The independent variables all together explained 8% of variance in item difficulties which was not statistically significant, (*F* (11, 86) = 1.79, *p* =.06, R^2 = .18, R^2 Adjusted = .08). Table 15 shows the Beta weights for the independent variables, their statistical significance and part correlations. The square of part correlations show the unique contribution of each independent variable to explaining item difficulty.

Table 15: *Multiple Regression*

Independent variable	Beta	T	P	Part correlation
		.300	.765	
Inflection	.187	1.340	.184	.130
Frequency	**-.216**	-1.671	.098	**-.162**
F.C	**.227**	1.674	.098	**.163**
Length of Word	**-.262**	-1.800	.075	**-.175**
Length of Sentence	-.094	-.753	.454	-.073
Propositional Density	.014	.137	.892	.013
Frequency before mutilated word	-.144	-1.333	.186	-.130
Frequency after mutilated word	.033	.318	.752	.031
Text difficulty (super-item p-value)	.122	.998	.321	.097
Dependency	.115	1.130	.262	.110
Text difficulty (Lexile)	.142	.953	.343	.093

As Table 15 shows word length has the strongest contribution to item difficulty. Next is function/content and third comes word frequency. Also, the useful piece of information that is displayed in Table 15 is part correlation. The square of part correlation tells how much of the total of variance in the dependent variable is explained by each variable. In this study, frequency of word, function content word, the length of word have part correlation co-efficient of .16, .16, and -.17, respectively. If we square them we get .025, .25, and .030, indicating that frequency and content/function word explain 2.5 percent of item difficulty and length of word explains 3 percent of the variance of item difficulty. However, these three factors explain a few percentage of item difficulty but they are better predicators than the other factors.

One-way Analysis of Variance (ANOVA)

To answer the thirteenth research question, one-way analysis of variance (ANOVA) was run. As Table 16 displays, in this analysis there is one independent variable (word class) with eight levels.

Table 16: *Mean Item Difficulties of Different Word Classes*

	N	Mean
Noun	21	.5138
Verb	30	**.6135**
Adj	12	**.6139**
Adv	8	.5483
Pronoun	6	.4425
proposition	8	.4832
conjunction	10	.5170
determiner	5	**.3707**
Total	100	.5449

Table 16 shows ‘verbs’ and ‘adjectives’ are more difficult to answer in C-Tests and ‘determiners’ are easier. It can be said that word classes affect item difficulty in C-Test items. So, Hypothesis 13 that was ‘word classes’ have no effect on item difficulty is rejected.

One-way ANOVA showed that there was not a statistically significant difference at the level of p<.05 in word class for eight groups: F (7, 92) = 1.21, *p*=.30. So, Robust Tests of Equality of Means was used and it became significant: *p*=.27.

The Linear Logistic Test Model (LLTM) Analysis

The linear logistic test model (LLTM, Fischer, 1973) was also used to study the contribution of different factors to C-Test item difficulty Since LLTM is an extension of the Rasch model, the standard Rasch model (Rasch, 1960/1980) should fit the data first (Fischer, 1973; Baghaei & Kubinger, 2015; Baghaei & Ravand, 2015; Ghahramanlou, Zohoorian, & Baghaei, 2017; Hohensinn & Baghaei, 2017). Andersen’s likelihood ratio test (Andersen, 1973) showed that the 100 items do not fit the Rasch model, χ^2=533.16, *df*=99, *p=00.* Graphical model check (Figure 1) revealed that 36 out of 100 items were misfitting.

Figure 1: *Graphical Model Check for the 100 C-Test Gaps*

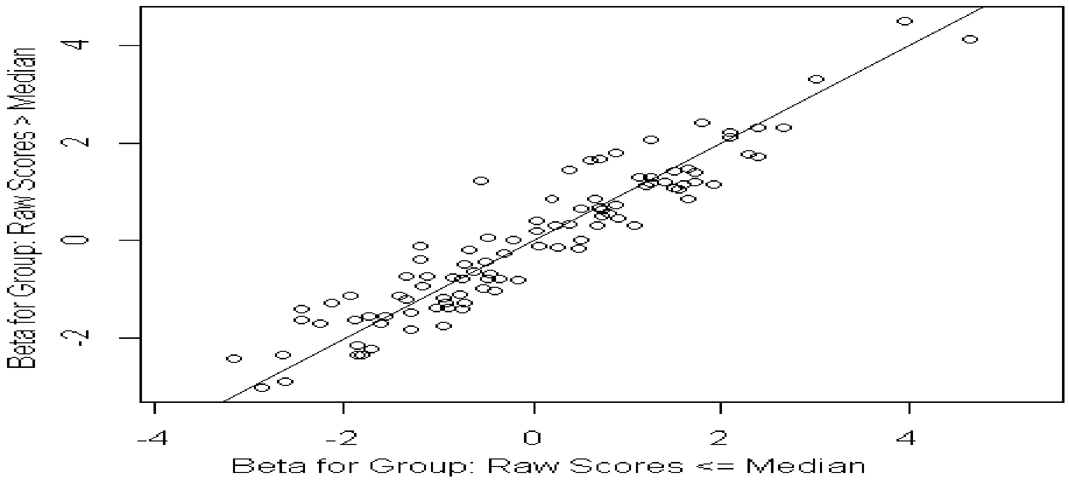
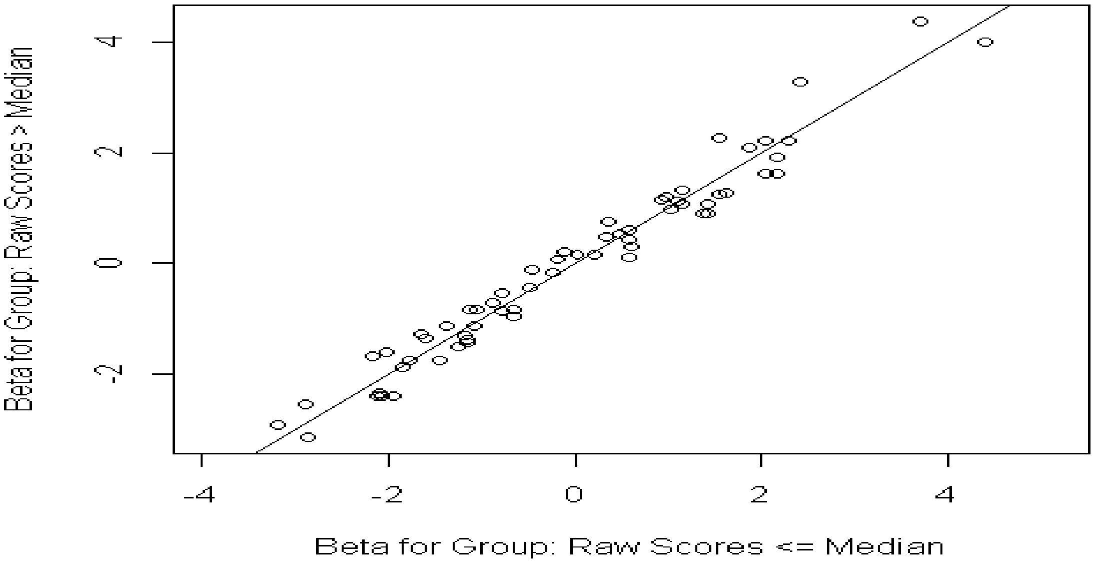


Figure 1 shows that after deleting the 36 misfitting items which fell far from 45 degree line, Rasch model was estimated again. Andersen's LR test showed that the 64 remaining items fit the Rasch model: $\chi 2 = 86.2$, $df = 63$, $p = 0.028$. Graphical model check showed that the items are close to the 45 degree line (Figure 2).

Figure 2: *Graphical Model Check for 64 C-Test Gaps*



Each word and text characteristic was considered a cognitive operation and a Q-matrix was constructed. The operations are as follows:

1.Frequency of mutilated word

2.Function/content words

3.Word length

4.Sentence length

5.Number of proposition
6.Propositional density
7.Text difficulty (Lexile)
8.Inflections
9.Frequency of word *before* the mutilate word
10. Frequency of word *after* the mutilated word
11. Text difficulty (P-value or difficulty of each super-item or passage)

Based on the results of graphical model check the 36 misftting items were deleted and a Q-Matrix for the 64 remaining items and the 11 basic parameters was developed. The 64 Rasch model fitting items and the Q-matrix were subjected to LLTM analysis using eRm (Mair, Hatzinger, & Maier, 2014) package in R version 3.11 (R Core Development Team, 2015). Table 20 shows the easiness parameters of the 11 operations, their standard errors, and their 95% confidence intervals.

Table 17: *Easiness of the basic parameters, standard errors and 95% confidence intervals for 11 operations*

	Estimate	Std. Error	lower CI	upper CI
1. Frequency of mutilated word	-0.592	0.049	0.688	-0.496
2. Content / Function word	-0.633	0.030	-0.692	-0.573
3. Word length	0.194	0.032	0.133	0.256
4. Sentence length	0.610	0.039	0.534	0.685
5. Number of proposition	-0.113	0.030	-0.172	-0.054
6. Propositional density	-0.131	0.033	-0.196	-0.066
7. Inflection	-0.252	0.039	-0.329	-0.175
8. Text difficulty (Lexile)	-0.332	NaN	NaN	NaN
9. Frequency of word before the mutilate Word	-0.592	0.074	-0.738	-0.446
10. Frequency of word after the mutilate word	0.825	0.107	0.616	1.035
11.Text difficulty (P-value or difficulty of each super-item or passage)	0.332	NaN	NaN	NaN

LLTM analysis revealed high errors for basic parameters 8, and 11, so they were omitted and LLTM was estimated again with 9 basic parameters (Table 17).

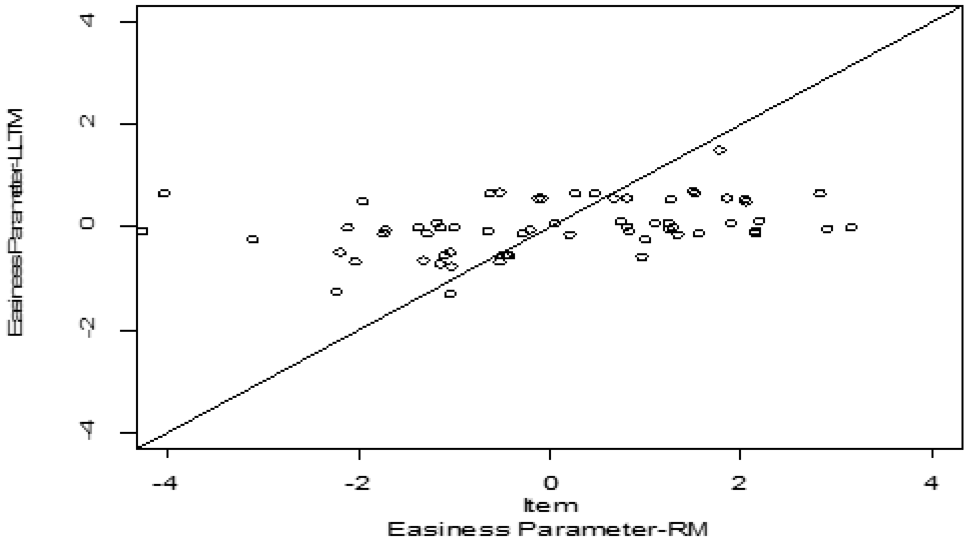
Table 18: *Easiness of the basic parameters, standard errors and 95% confidence intervals* for 9 *operations*

	Estimate	Error	lower CI	upper CI
	Std.			
1. Frequency of mutilated word (eta 1)	-0.573	0.049	-0.668	-0.477
2. Content / Function word (eta 2)	-0.689	0.029	-0.747	-0.632
3. Word length (eta 3)	0.118	0.029	0.061	0.176
4. Sentence length (eta 4)	0.652	0.038	0.578	0.727
5. Number of proposition (eta 5)	-0.094	0.030	-0.152	-0.035
6. Propositional density (eta 6)	-0.155	0.033	-0.220	-0.091
7. Inflection (eta 7)	-0.648	0.029	-0.704	-0.592
8. Frequency of word before the mutilated word (eta 8)	-0.566	0.074	-0.711	-0.420
9. Frequency of word after the mutilated word (eta 9)	0.834	0.107	0.625	1.044

LLTM imposes a linear constraint on the difficulty parameter. It means, we should be able to reconstruct Rasch model-based item parameters by adding the difficulty of the operations needed to solve each item.

Comparing the fit of LLTM and the Rasch model with the likelihood ratio test showed that the Rasch model fits significantly better than LLTM, χ^2=9856, *df* =54, *p =0.00*. The correlation between Rasch model-based items estimates and LLTM-reconstructed item estimates was .37; that is, we managed to explain 12% of the variance in item difficulties with the nine factors (Figure 3).

Figure 3: *RM Item Parameters vs. LLTM Item Parameters*



Results and Discussions

The purpose of this study was to establish whether 13 independent variables (1. The frequency of the mutilated word, 2. Whether there is content and function word, 3. The length of the mutilated word, 4. The length of the sentence where the gap is, 5. The number of propositional, 6. The propositional density (of the sentence where the gap is), 7. Inflections (ed, s(plural),s(3rdperson),al(adjective), ly (adverb),er, est (superlative)), 8. Text difficulty (lexile), 9.Frequency of the word *before* mutilate word, 10. Frequency of the word *after* mutilate word, 11. Text difficulty (p-values of texts), 12. Dependency, 13. Word classes (noun, verb, adjective, adverb, pronoun, preposition, conjunction, and determiner) have any significant effects on C-Test item difficulty.

After collecting the data, correlational analysis, regression, ANOVA, and LLTM were conducted. Based upon the response of participants, all the 13 hypotheses were tested and the results represented. By interpreting the results of correlational analysis, it is concluded that frequency of mutilated word and content/function word, and text difficulty as measured by p-values have a significant correlation with item difficulty in C-Test items. However, text difficulty as p-value has a significant correction because this measure of text difficulty is based on the difficulty of individual C-Test items within a text. Therefore, it should have correlation with item difficulty. The other variables had no significant correlation with item difficulty in this study.

ANOVA was used to analyze the effect of word class with eight levels as mentioned above, on item difficulty. It showed that there were significant differences among the mean difficulties of different word classes. Hence, Verbs and adjectives were harder to reconstruct for participants and determiners were easier.

Multiple regression was used to assess the ability of the 11 independent variables (Number of propositions was deleted because of high correlation with sentence length (r=.94) and also word class was not used in this analysis) to predict the item difficulty of the C-Test items. The result revealed that the 11 independent variable just explain 8% of variance in item difficulties. The Beta weights showed that word length has the strongest contribution to item difficulty. Next was function/content and third was word frequency.

The linear logistic test modeling (LLTM) was used to be sure about the explanation of the variance in item difficulty that acquired in the analysis of multiple regression. First, Rasch model was used for 100 items to determine whether they fit this model or not. Results showed that 36 items did not fit, so after deletion of these 36 items, for 64 items Rasch model was run again. After developing the Q-Matrix for 64 items and 11 basic operations (dependency and word class were omitted in this analysis), LLTM was run. Two basic operations were deleted due to high errors and LLTM was rerun. The 9 parameters as mentioned earlier were: 1. Frequency of mutilated word, 2. Content/function word, 3. Word length, 4. Sentence length, 5. Number of propositions, 6. Propositional density, 7. Inflections, 8. Frequency of word before the mutilated word, 9. Frequency of word after the mutilated word.

The result of this analysis showed that content words, inflections, and the frequency of the mutilated word had the greatest impact on item difficulty. Although, there were some other parameters but they did not have remarkable effect on item difficulty. Also, LLTM explained 12 % of variance in item difficulties.

Conclusions

As mentioned before, in this study the researcher hypothesized that 13 factors contribute to C-Test item difficulty. These factors were entered into regression analysis as independent variables to predict C-Test item difficulties. Also, correlational analysis, ANOVA, and LLTM were used. So, in the present study we are going to identify which factors make items more difficult or easier.

To determine the effect of the 13 independent variable on item difficulty, the current study was conducted. 352 students of several universities of Mashhad and Neyshabour were selected to answer 100 items of four C-Test texts. The texts were chosen from CAE (Norris & French, 2008) and FCE books (Norris, 2008).

The results of correlation analysis were as follow: First, the result of this study showed that frequency of mutilated word had a significant relationship with the item difficulty. It means that if the mutilated word has a high frequency, it will help test takers to answer it better than a low frequency word. For instance, the mutilated word 'sch……..'= 'school' with a high frequency (with frequency of '5') was easier to answer, but the mutilated word 'instr………' = 'instructed' with frequency of '3'was more difficult to answer. Therefore, word frequency affects the difficulty of each item.

Moreover, whether the mutilated words are function or content words can affect the item difficulty. If the mutilated word is a content word, it is harder to answer. For example, 'students' as a content word was more difficult to answer than 'into' as a function word. In addition, there was a significant correlation between text difficulty as measured by p-values with item difficulty in C-Test items. Text difficulty as p-value is based on the difficulty of individual C-Test items within a text, and it is clear that it has a significant correlation with the item difficulty. Finally, analyzing eight word classes illustrated that 'verbs' and 'adjectives' were more difficult to answer in C-Tests and 'determiners' were easier.

Whereas, the other 9 independent variables had no significant correlation with item difficulties in C-Tests. These variables were: (1) word length, (2) sentence length, (3) the number of propositions in the sentence where the C-Test item is, (4) the propositional density of the sentence where the C-Test item is, (5) inflections, (6) text difficulty as measured by Lexile, (7) item dependency, (8) frequency of the word before the mutilated word, (9) the frequency of the word after the mutilated word.

And the result of the linear logistic test model (LLTM, Fischer, 1973) was almost the same as correlational analysis. LLTM showed that Andersen's likelihood ratio test (Andersen, 1973) showed that all 100 items do not fit the Rasch model. As mentioned in chapter 4, graphical model check revealed that 36 out of 100 items were misfitting. So, 36 misfitting item were deleted, and then the Rasch model was estimated again. A Q-Matrix for the 64 remaining items and the 11 basic parameters was developed. They were: (1) frequency of mutilated word, (2) function/content words, (3) word length, (4) sentence length, (5) number of proposition, (6) propositional density, (7) text difficulty (Lexile), (8) inflections, (9) frequency of word before the mutilate word, (10) frequency of word after the mutilated word, (11) text difficulty (P-value or difficulty of each super-item or passage).

LLTM analysis illustrated high errors for two parameters named 'Inflections' and 'text difficulty' (P-value or difficulty of each super-item or passage). Hence, they were omitted and LLTM was estimated again with the other 9 basic parameters. The results revealed that the 9 independent variables all together explained 12% of variance in item difficulties. In general, based on findings of correlational analysis and LLTM it is concluded that frequency of mutilated word, content/function word, and text difficulty as measured by p-values had a significant contribution to item difficulty in C-Test items. According to the result of the LLTM model, content words, inflections, and the frequency of the mutilated word had the greatest impact on item difficulty.

The findings of this study revealed that the 13 factors that we selected only explained a small portion of the variance in C-Test item difficulties. Some of these factors were already in the literature and some were added by the researcher i.e., whether the words are content or function words, the length of the mutilated word, the number of propositions in the sentence where the gap is, the propositional density (of the sentence where the gap is), text difficulty (as measured by Lexile) (www.lexile.com), the frequency of the word before the mutilate word, the frequency of the word after the mutilate word. The researcher included all the possible factors which deemed to affect C-Test item difficulty. No construct identification study on C-Test has so far covered as many number of factors as included in this study. Nevertheless, the portion of the variance explained, i.e. 12%, is rather small considering the number of factors that were entered into the analysis.

One reason for the observed findings is that test takers may use different skills and strategies to answer C-Test items. Therefore, explaining item difficulties with a number of factors for all the test takers is not possible. According to Sigott (2004), C-Tests have a fluid construct. He argued that the construct underlying the C-Test changes as a function of person ability and text difficulty; That is, a C-Test could measure different things for different examinees. If the fluid construct phenomenon is true, explaining and modeling item difficulty in C-Test is very difficult if not impossible. However, researchers in future must think of other additional relevant factors that might contribute to item difficulty.

Another issue that must be given attention is that correlation is sensitive to restrictions of range. That is, when the range of the measured variables is small, the correlation coefficients are depressed. Our analysis suffered from this problem. Almost all of our independent variables such as word frequency, content/function, etc. suffered from range restrictions. Frequency was measured on scale from 1 to 5 and content/function was dichotomous with only two values, 1 and 2. Therefore, the small correlations we observed in this study are partly due to the small range of the variable.

Thus, different levels of proficiency cause different interpretations for C-Test scores because “the same C-Test passage could well be different tests for subjects at different levels of proficiency…without [the test user] knowing to what extent different aspects of the construct are reflected in the individual test scores” (Sigott, 2004, p.203). If FCP is true then it is very difficult to understand what factors make C-Test items hard. Consequently, while answering the C-Test item different factors may influence the difficulty of each item and it would be hard to find out the exact reason why an item becomes easy or hard.

The findings of this study may have some hints and implications for the other researchers. In the present study, the effect of 13 independent variables on item difficulty in C-Test items were investigated. In material and test development, it is crucial to find out which factors make an item easier or more difficult. In fact, what makes a test or task hard can guide teachers and material developers for ideal use of the tasks.

According to Grotjahn and Stemmer (2002), C-Tests are based on a variation of the cloze principle and thus have the same basic theoretical assumptions, but many of EFL teachers and educators do not know the similarities and differences of cloze tests and C-Tests. The results of the current research showed that a C-Test can be used to test knowledge of vocabulary and we cannot measure reading or grammar through such tests. So, C-Tests can be used as a vocabulary test at schools for different levels and as a vocabulary tasks in course books. It is important that C-Tests measures the exact thing that we want. For instance, if it is a vocabulary test, it should test the knowledge of vocabulary or if it is a grammar test, it should test the knowledge of grammar, etc.

Moreover, by reviewing the literature, it is found that C-Tests can measures crystalized intelligence (Baghaei & Tabatabaee, 2015). Crystallized intelligence is the ability to use “culturally approved, previously acquired problem solving methods” (Hunt, 2000, p. 127) and represents abilities that result from education and experience (Baghaei & Tabatabaee, 2015, p. 47). In general, it should be noted that different abilities and intelligences can affect the way test takers answer the test and based on their ability their answers would be different.

Suggestions for Further Research

The researcher suggests the following areas for further research related to the difficulty of C-Test items.

- According to Sigott (2004), fluid nature of C-Test construct may influence the C-Test scores interpretations. So, it should be considered as a crucial aspect of C-Tests for test scoring.
- Also, further studies should deal with the effect of paragraphs and text on C-Test item difficulty with the vast number of variables because in the present study the focus was on the gap-level.
- Moreover, The 13 independent variables all together explained 12% of variance in item difficulties. For this reason, more research is needed to identify what the exact factors are and how they affect item difficulty based on the gap-level.
- Other important consideration is that some researchers believe C-Test is a kind of general language proficiency test (Eckes & Grotjahn, 2006; Klein-Braley, 1985; Sigott, 1995) and the other point out that it is a vocabulary test (Chapelle, 1994; Singleton and Little, 1991) or a grammar test (Babaii and Ansary, 2001). With regard to the present study, the researcher concluded that C-Test can be a vocabulary test. However it still needs more investigation.
- Besides, the current study did not deal with ESL or ESP students so they can be examined using the same design procedures as same as EFL learners.

References

Adams, R. J., Wilson, M., & Wang, W. (1997). The Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement, 21*(1), 1-23. doi:10.1177/0146621697211001

Agarwal, M., & Mannem, P. (2011). Automatic Gap-fill Question Generation from Text Books. *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, 56–64.

Alderson, J.C. (1979). The effect on the cloze test of changes in deletion frequency. *Journal of Research in Reading,* 2, 18-108.

Alderson, J.C. (1983). The cloze procedure and proficiency in English as a foreign language.In J.W. Oller (Ed.). *Issues in language testing research*, Rowley, Mass., Newbury House,17-205.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38*(1), 123-140. doi:10.1007/bf02291180

Babaii, E., & Ansary, H. (2001). The C-test: a valid operationalization of reduced redundancy principle? *System*, 29(2), 209–219.

Babaii, E., & Jalali Moghaddam, M. (2006). On the interplay between test task difficulty and macro-level processing in the C-test. *System*, 34(4), 586-600.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Baghaei, P, Monshi Tousi, M., & Boori, A.A. (2009). An investigation into the validity of conversational C-Tests as a measure of oral abilities. *Iranian EFL Journal, 4,* 94–109.

Baghaei, P. & Grotjahn, R. (2014). Establishing the construct validity of conversational C-Tests using a multidimensional Rasch model. *Psychological Test and Assessment Modeling,* 56(1), 60-82.

Baghaei, P. (2011). *C-Test construct validation: a Rasch modeling approach*. Saarbrücken: VDM Verlag Dr. Müller.

Baghaei, P. (2008a). An attempt to fit the Rasch model to a C-Test. *Iranian EFL Journal*, 2, 6-21. Retrieved from http://www.iranian-efl-journal.com/2008-Editions.php.

Baghaei, P. (2008b). The effects of the rhetorical organization of texts on the C-test construct: A Rasch modelling study. *Melbourne Papers in Language Testing*, 3(2), 32-51.

Baghaei, P. (2014). Construction and validation of a C-Test in Persian. In R. Grotjahn (Ed.), *Der C-Test: Aktuelle Tendenzen/The C-Test: Current Trends*, 301-314. Frankfurt am Main: Lang.

Baghaei, P., & Kubinger, K. D. (2015). Linear logistic test modeling with R. *Practical Assessment, Research & Evaluation, 20,* 1-11. Retrieved from http://pareonline.net/getvn.asp?v=20&n=1.

Baghaei, P., & Ravand, H. (2015). A cognitive processing model of reading comprehension in English as a foreign language using the linear logistic test model. *Learning and Individual Differences, 43,* 100-105. doi:10.1016/j.lindif.2015.09.001

Baghaei, P., & Tabatabaee, M. (2015). The C-Test: An Integrative Measure of Crystallized Intelligence. *Journal of Intelligence, 3*(2), 46-58. doi:10.3390/jintelligence3020046

Baghaei, P. (2010). An investigation of the invariance of Rasch item and person measures in a C- Test. In R. Grotjahn (Ed.). *Der C-Test: Beiträge aus der aktuellen Forschung/ The C-Test: Contributions from Current Research* (pp.100-112). Frankfurt/M.: Lang.

Baghaei, P. (2011a). Do C-Tests with different number of gaps measure the same construct? *Theory and Practice in Language Studies, 1,* 688-693.

Baghaei, P. (2011b). Optimal number of gaps in C-Test passages. *International Education Studies, 4,* 166-171.

Baghaei, P., & Grotjahn, R. (2014a). The validity of C-Tests as measures of academic and everyday language proficiency: A multidimensional item response modeling study. In R. Grotjahn (Ed.). *Der C-Test: Aktuelle Tendenzen/The C-Test: Current trends* (pp. 163-171.). Frankfurt/M.: Lang.

Baghaei, P., & Grotjahn, R. (2014b). Establishing the construct validity of conversational C-Tests using a multidimensional Item Response Model. *Psychological Test and Assessment Modeling, 56,* 60-82.

Baghaei, P., & Tabatabaee-Yazdi, M. (2016). The logic of latent variable analysis as validity evidence in psychological measurement. *The Open Psychology Journal, 9,* 168-175.

Baur, R.S. & Meder, G. (1994). C-Tests zur Ermittlung der globalen Sprachfähigkeit im Deutschen und in der Muttersprache Bei ausländischen Schülern in der Bundesrepublik Deutschland. In R. Grotjahn (Ed.), *Der C- Test. Theoretische Grundlagen und praktische Anwendungen*. Bochum: Brockmeyer. 151-178. Retrieved from http://www.c-test.de/deutsch/index.php?lang=de&section=originalia.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*(1), 92-111. doi:10.1016/j.jml.2008.06.003

Beinborn, L., Zesch, T., & Gurevych, I. (2014). Predicting the Difficulty of Language Proficiency Tests. *Transactions of the Association for Computational Linguistics*, 2, 517–529.

Bolten, J. (1992). Wie schwierig ist ein C-Test? Erfahrungen mit dem C-Test als Einstufungstest in Hochschulkursen Deutsch als Fremdsprache. In R. Grotjahn (Ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen*, 193-203. Bochum: Brockmeyer.

Borgards, S. & Raatz, U. (2002). Sind C-Tests trainierbar? In R. Grotjahn (Ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* .Bochum: AKS-Verlag. 157–174.

Brown, J. D., Yamashiro, A. D., & Ogane, E. (2001). The emperor’s new cloze: strategies for revising cloze tests. In T. Hudson & J. D. Brown (Eds.), *A focus on language test development: Expanding the language proficiency construct across a variety of test*, 143-161. Honolulu, HI: University of Hawai’I Press.

Brown, J. C., Frishkoff, G. A., & Eskenazi, M. (2005). Automatic question generation for vocabulary assessment. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05*. doi:10.3115/1220575.1220678

Brown, J.D. (1984). “A cloze test is a cloze is a cloze?”. J. Handscombe, R.A. Orem, & B.P. Taylor (Eds.), *on TESOL’83: the question of control.* 109-119. Washington, D.C.: TESOL.

Brown, J.D. (1989). Cloze item difficulty. *JALT Journal*, 11, 46–67.

Brown, J. D. (2005). *Testing in language programs: a comprehensive guide to English language assessment*. New York: McGraw Hill.

Chapelle, C. A. (1994). Are C-tests valid measures for L2 vocabulary research? *Second Language Research, 10*(2), 157-187. doi:10.1177/026765839401000203

Cronbach, L.J. (1980). Validity on parole: how can we go straight? *New directions for testing and measurement, 5, 99-108.*

Cronbach, L.J., & Meehl, P.E. (1995). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281-302.

Douglas, D. (1998). Testing methods in context-based second language research. *Interfaces between Second Language Acquisition and Language Testing Research,* 141-155. doi:10.1017/cbo9781139524711.008.

Eckes, T. & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing*, 23(3), 290–325. doi:10.1191/0265532206lt330oa

Eckes, T. (2006). Rasch-Modelle zur C-Test-Skalierung. In Rüdiger Grotjahn (Ed.), *Der C-Test: Theorie, Empirie, Anwendungen/The C-Test: theory, empirical research applications.* Frankfurt am Main: Lang. 1-44.

Eckes, T., & Baghaei, P. (2015). Using testlet response theory to examine local dependency in C-Tests. *Applied Measurement in Education, 28,* 85–98.

Eckes, T. (2011). Item banking for C-tests: A polytomous Rasch modeling approach. *Psychological Test and Assessment Modeling*, 53, 414–439.

Embretson, S. (1983). Construct Validity: Construct Representation Versus Nomothetic Span. *Psychological Bulletin*, 93(1), 179-197.

Farhady, H., Jafarpur, A., & Birjandi, P. (1994). Test language Skills: From Theory to Practice. The organization for researching and composing university textbooks in the humanities (SAMT), The Center for research and Development in Humanities. Tehran.

Feldmann, U., & Stemmer, B. (1987). Thin____ aloud a___ retrospective da___ in C-te___ taking: diffe___ languages – diff___ learners – sa___ approaches? In Claus Færch & Gabriele Kasper (Hrsg.), *Introspection in second language research,* 251-267.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37(6), 359-374. doi:10.1016/0001-6918(73)90003-6

Germann, U. & Grotjahn, R. (1994). Das Lösen von C-Tests auf dem Computer: Eine Pilotuntersuchung zu den Bearbeitungsprozessen. In R. Grotjahn (Ed.), Der C-Test. *Theoretische Grundlagen und praktische Anwen-dungen,* 2, 279-304. Germany: Lang.

Ghahramanlou, M., Zohoorian, Z., & Baghaei, P. (2017). Understanding the cognitive processes underlying performance in the IELTS listening comprehension test. *International Journal of Language Testing,* 7, 1-15.

Grotjahn, R. & Stemmer, B. (2002). C-Tests and language processing. In J. A. Coleman, R. Grotjahn, & U. Raatz (Eds.) , *University language testing and the C-Test,* AKS-Verlag, Bochum. *Handbook of test development*, 261-285.Mahwah, NJ: Erlbaum.

Grotjahn, R., & Stemmer, B. (1985). On the development and evaluation of a C-Test for French. In C. Klein-Braley & U. Raatz (Eds.), *Fremdsprachen und Hochschule, 13/14: Thematischer Teil: C-Tests in der Praxis,* 101-120.

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

Horsmann.T. & Zesch. (2014). Towards automatic scoring of cloze items by selecting low ambiguity contexts. *NEALT Proceedings Series*, 22, 33–42.

Hunt, E. (2000). Let's hear it for crystallized intelligence. *Learning and Individual Differences, 12*(1), 123-129. doi:10.1016/s1041-6080(00)00033-9

Hohensinn, C. & Baghaei, P. (2017). Does the position of response options in multiple-choice tests matter? *Psicológica , 38*, 93-109.

Kamimoto, T. (1993). Tailoring the Test to Fit the Students: Improvement of the C-Test through Classical Item Analysis. *Language Education and Technology,* 30, 48-61.

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

Klein-Braley, C. (1984). Advance Prediction of Difficulty with C-Tests. In Terry Culhane, Christine Klein-Braley, and Douglas K. Stevenson, editors, *Practice and problems in language testing*, 7, 97–112.

Klein-Braley, C. (1985). A cloze-up on the C-Test: a study in the construct validation of authentic tests. *Language Testing, 2*(1), 76-104. doi:10.1177/026553228500200108

Klein-Braley, C. (1994). Language testing with the C-Test. *A linguistic and statistical investigation into the strategies used by C-Test takers, and the prediction of C-Test difficulty.* Unpublished post-doctoral thesis (Habilitationsschrift), Universität-Gesamthochschule Duisburg: Fachbereich 3, Sprach-und Literaturwissenschaften.

Klein-Braley, C. (1996). Towards a theory of C-Test processing. In R. Grotjahn (Ed.), Der C-Test. *Theoretische Grundlagen und praktische Anwendungen,* 3, 23-94. Bochum: Brockmeyer.

Klein-Braley, C. (1997). C-Tests in the context of reduced redundancy testing: an appraisal. *Language Testing, 14*(1), 47-84. doi:10.1177/026553229701400104

Klein-Braley, C. (2002). Psycholinguistics of C-Test taking. In J. A. Coleman, R. Grotjahn, & U. Raatz (Eds.), *University language testing and the C-test*, 131-142. Bochum:AKS-Verlag.

Koberl, J., & Sigott, G. (1994). Adjusting C-test difficulty in German. In R. Grotjahn (Ed.), Der C-Test. *Theoretische Grundlagen und praktische Anwendungen,* 2, 179-192. Bochum: Brockmeyer.

Li, Y. (2012). Applying Multidimensional Item Response Theory Models in Validating Test Dimensionality: An Example of K–12 Large-scale Science Assessment. *Journal of Applied Testing Technology*. 13(2), 1-27.

Linacre, J. M. (2014). WINSTEPS Rasch Measurement [Computer program]. Chicago, IL: Winsteps.

Mair, P., Hatzinger, R & ,.Mair, M. J. (2014). eRm: extended Rasch modeling [Computer software]. R package version 0.15-4 .http://CRAN.R-project.org/package=eRm.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174. doi:10.1007/bf02296272

Messick, S. (1989).Validity. In R.L. Linn (Ed.), *Educational measurement.* New York: Macmillan.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*(10), 955-966. doi:10.1037//0003-066x.30.10.955

Michael J. H., M.J. Collins-Thompson, K., Callan, J., & Eskenazi, M. (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. *In Proceedings of NAACL-HLT,* 460–467.

Mitkov, R., Ha, L.A., & Karamanis, N. (2006). A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering*, 12, 177–194.

Mitkov, R., Ha, L. A., & Karamanis, N. (2006). A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering, 12*(02), 177.

Mostow, J., & Jang, H. (2012). Generating Diagnostic Multiple Choice Comprehension Cloze Questions. *In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, 136–146. *Association for Computational Linguistics.*

Norris, R., & French, A. (2008). *Ready for CAE: coursebook*. Oxford, UK: Macmillan.

Norris, R. (2008). *Ready for FCE: coursebook*. Oxford: Macmillan.

Oller, J. W., Jr. (1976). Evidence of a general language proficiency factor: An expectancy grammar. *Die Neuren Sprachen*, 76, 165-74.

Oller, J. W. (1979). *Language tests at school: a pragmatic approach*. London: Longman.

Perez-Beltrachini, L., Gardent, C., & Kruszewski. G. (2012). Generating Grammar Exercises. *English Language Teaching* ,147–156.

Perkins, K., & Linnville, S. E. (1987). A construct definition study of a standardized ESL vocabulary test. *Language Testing, 4*(2), 125-141. doi:10.1177/026553228700400201

R CORE TEAM. (2012). R: a language and environment for statistical computing [Computer program]. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, URL http://www.R-project.org.

Raatz, U. (1984). The factorial validity of C-Tests. In T. Culhane, C. Klein- Braley, & D. K. Stevenson (Eds.), *Practice and problems in language testing 7. Proceedings of the Seventh International Language Testing Symposium of the IUS, Colchester*, 124-139. Colchester (England): University of Essex, Department of Language and Linguistics.

Raatz, U. (1985). Investigating dimensionality of language tests – a new solution to an old problem. In Viljo Kohonen, Hilkka von Essen & Christine Klein-Braley (Eds.), *Practice and problems in language testing 8.* 123–136.Tampere: AFinLA.

Raatz, U., & Kelein-Braley C. (2002). Introduction to the language and the C-Test. *University Language Testing and the C-Test*, 75-86.

Raatz, U., & Klein-Braley. C. (1981), The C-Test – a modification of the cloze procedure. In T. Culhane, C. Klein-Braley and D. K. Stevenson (Eds.), 113-38.

Rasch, G. (1960/1980). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research, 1960. (Expanded edition, Chi-cago: The university of Chicago Press, 1980).

Sasaki, M. (2002). Effects of cultural schemata on students' test-taking processes for cloze tests: a multiple data source approach. *Language Testing, 17*(1), 85-114.

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational Measurement*. 307-353. Westport, CT: American Council on Education and Praeger Publishers.

Shohamy, E., & Inbar, O. (1991). Validation of listening comprehension tests: the effect of text and question type. *Language Testing, 8*(1), 23-40.

Sigott, G. (1995). The C-test: some factors of difficulty. *AAA: Arbeiten aus Anglistik und Amerikanistik*, 43-53.

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

Sigott, G. (2006). How fluid is the C-Test construct. In Rudiger Grotjahn and Gunther Sigott, editors, *Der C-Test: Theorie, Empirie, Anwendungen .The C-Test: Theory, Empirical Research, Applications*, 139–146. Peter Lang.

Sigott, G., & Köberl, J . (1996). Deletion patterns and C-Test difficulty across languages. In R. Grotjahn (ed.), Der C-Test. *Theoretische Grundlag, und praktische Anwendungen*, 3,159-172. Bochum: Brockmeyer.

Singleton, D., & Little, D. (1991). The second language lexicon: some evidence from university-level learners of French and German. *Second Language Research, 7*(1), 61-81. doi:10.1177/026765839100700103

Skory, A., & Eskenazi, M. (2010, June). Predicting cloze task quality for vocabulary training. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, 49-56. Association for Computational Linguistics.

Spolsky, B. (1969). Reduced Redundancy as a Language Testing Tool. In G.E. Perren and J.L.M. Trim, editors, *Applications of linguistics*, 383-390. Cambridge University Press.

Spolsky, B. (2001). Closing the cloze. In H. Pürschel & U. Raatz (Eds.), *Tests and Translation. Papers in memory of Christine Klein-Braley*. Bochum: AKS-Verlag.

Spolsky, B. Bengt, S., Sato, M., Walker, E., & Aterburn, C. (1968) Preliminary studies in the development of techniques for testing overall second language proficiency. *Language Learning Special Issue*, 3, 79-101.

Stemmer, B. (1991). *What's on a C-test taker's mind: Mental processes in C-test taking*? .Bochum: Brockmeyer.

Stenner, A.J., Smith, M., & Burdick, D.S. (1983). Toward a theory of construct definition. Department of Mathematics, Duke University, Durham, North Carolina, *United States Journal of Educational Measurement* , 20, 305-316. DOI: 10.1111/j.1745-3984.1983.tb00209.x

Tadamitsu, K. (1993). Tailoring the Test to Fit the Students: Improvement of the C-Test through Classical Item Analysis. *Language Laboratory*, 30, 47–61.

Taylor, W. L. (1953). Cloze procedure: a new tool for measuring readability. *Journalism and Mass Communication Quarterly, 30*(4), 415.

Tong, K. (1987). An investigation into the 'coaching effect' on students' Performance in C-tests. *Language Testing Update, 4* - 25.