International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:5, No:2, 2011

Do C-Test and Cloze Procedure Measure what they Purport to be Measuring? A Case of Criterion-Related Validity

Masoud Saeedi, Mansour Tavakoli, Shirin Rahimi Kazerooni, and Vahid Parvaresh

***Abstract*—**This article investigated the validity of C-test and Cloze test which purport to measure general English proficiency. To provide empirical evidence pertaining to the validity of the interpretations based on the results of these integrative language tests, their criterion-related validity was investigated. In doing so, the test of English as a foreign language (TOEFL) which is an established, standardized, and internationally administered test of general English proficiency was used as the criterion measure. Some 90 Iranian English majors participated in this study. They were seniors studying English at a university in Tehran, Iran. The results of analyses showed that there is a statistically significant correlation among participants' scores on Cloze test, C-test, and the TOEFL. Building on the findings of the study and considering criterion-related validity as the evidential basis of the validity argument, it was cautiously deducted that these tests measure the same underlying trait. However, considering the limitations of using criterion measures to validate tests, no absolute claims can be made as to the construct validity of these integrative tests.

***Keywords*—**Integrative testing, C-test, Cloze test, the TOEFL, Validity.

I. Introduction

TESTING in general and language testing in particular is an indispensable part of any educational program. It is regarded as a thorny area in that it influences individuals' lives in varying ways and to different extents. The importance of testing is even more conspicuous when it is a high-stakes one i.e. some crucial decisions made on the basis of test results. Consequently, educators have always been concerned with developing appropriate tests. They have brought their endeavors to bear on the development of tests which on the one hand provide us with as accurate information on test takers' skill being measured as possible and, on the other hand are in keeping with the latest developments in other testing-related areas. They efforts, therefore, have made for the emergence of disparate approaches to testing, each claiming superiority over other competing testing approaches.

Masoud Saeedi is with the University of Isfahan, Isfahan, Iran. (Corresponding author email: Saeedi.tefl@gmail.com)
Mansour Tavakoli is with the University of Isfahan, Isfahan, Iran.
Shirin Rahimi Kazerooni is with Khorasgan Azad University, Isfahan, Iran.Vahid Parvaresh is with the University of Isfahan, Isfahan, Iran.

One of the dominant approaches to language testing is the integrative approach. This view of testing involves the testing of language in context. It is concerned, therefore, with overall meaning and proficiency, the total communicative effect of discourse and the underlying linguistic competence of which it is argued that all learners possess [1]. The adherents of integrative testing are of the opinion that natural language processing and production requires making a highly complex series of decisions, which will involve knowledge of a number of crucial elements such as grammatical structure, lexis, pronunciation and intonation, discourse structure and conventions. Hence, the argue that tests should not separate language skills into neat and ordered divisions, rather, should seek to gauge the test taker's ability to use two or more skills simultaneously [2], [1].One of the most common types of integrative tests is the Cloze test. The principle underpinning cloze tests rests on Gestalt Psychology and the information processing theory of "closure" which pertains to the inclination of individuals to complete a pattern once they have understood the its general significance [3]. Cloze tests are intended to assess the test taker's ability to decode interrupted or mutilated message by making the most acceptable substitutions from all the contextual clues available. There are several methods for deleting words on cloze tests. Some researchers have preferred a random deletion of words and others have opted for a selective deletion. Cloze tests, however, have traditionally consisted of the regular or systematic deletions of words from a text (usually every 5 to 10 words) and their replacement by even-length blank lines. The test takers are then supposed to guess the deleted words. The proponents of Cloze tests have contended that they provide a superior means of arriving at an overall picture of proficiency since they are indicative of the degree to which language skills are used in as meaningful context, but a number of researchers have also found them to be specially useful tools for gauging reading comprehension skill. Brown [4], Oller and Jonz [5], and Sampson and Briggs [6] held the idea that the major reason for this is the fact that the Cloze procedure assumes the reading is an interactive process and these tests are designed in such a way as to show whether the reader is familiar enough with the author's language and context to interact with the text in a way that preserves that author's meaning. Furthermore, Cloze tests measure the reader's ability to use contextual clues to derive meaning. Theorists have suggested that the ability to use contextual clues in order to derive meaning is a crucial step in the development of overall reading comprehension [7].The Cloze procedure enjoys several advantages over other

types of reading assessments. First, they are very easily created and administered. Moreover, they are based on silent reading, which is the predominant and most natural form of reading. Also, they can be constructed from materials that teachers use for instructional purposes or from authentic texts and they do not require the writing of particular comprehension questions. Finally, Cloze test often exhibits a high degree of consistency; though this consistency may vary considerably ,depending on the text selected, the deletion starting point, and gap rates that are utilized [8], [9], [10] ,and [3]. The literature abounds with research on Cloze procedure; researchers have considered and investigated Cloze tests from different perspectives, each focusing on a particular aspect of the test. Alderson [11], for instance, showed that changes in deletion frequency sometimes resulted in significant differences between tests. However, the change was not as expected, since less frequent deletion sometimes actually resulted in more difficult tests. When only those items common to both frequencies in any comparison were considered, no significant differences were found. He concluded that increasing the amount of context on either side of a Cloze gap beyond five words had no effect on the ease with which that gap would be closed. In another study Alderson [12], investigated the effect of certain methodological variables on the validity of Cloze test. These variables were: deletion rate, text, and the scoring procedure. Correlating test takers' performance on easy, medium and difficult Cloze tests with a test of proficiency in English as a foreign language (ELBA) test, Alderson [12] found that differences between texts are not very great when looking at the correlations with the total, but the correlations with individual parts of the ELBA vary. Furthermore, he found that scoring for any semantically acceptable word (SEMAC) produced among the highest correlations with the ELBA total. In particular, it almost always correlated higher than the exact word scoring procedure. In other words, changing the scoring procedure, results in different validity of the Cloze; the SEMAC appears to be the most valid procedure of EFL testing. As for the effect of deletion rate, the results of Alderson's study clinched the idea that it exerts a drastic effect on the validity of Cloze test. In a seminal research, Brown [13] found that Cloze blanks tend to provide a fairly representative sample of the language in the passages regardless of the starting point for the deletion pattern. It was, he believed, reasonable to assume that even a semi-random sampling of words from a passage will be reasonably representative of the words in that passage (especially if there are sufficient blanks, as in a 50 item cloze test). However, at the same time he noticed, quite reasonably, that some items were testing at the sentential level while others were testing at the inter-sentential level. What he came to realize was that only some of the items on a Cloze test may be functioning well for a given population of students, so regardless of the fact that the blanks may provide a representative sample of the language in the passage, the variance produced by those items may only be coming from those few items that are functioning well. Thus, the test variance may not be representative of the sampled items, and in turn may not be representative of the passage. For that reason, he hypothesized (as did Alderson[8]) that samples of items that delete different words, even in the same passage, may produce Cloze tests that are quite differentIn fact, the Cloze test was originally intended to measure the reading difficulty of a text. According to Cohen [2], it is a reliable means of determining whether or not certain texts are at appropriate level for particular groups of students. It also measures textual knowledge. However, perhaps the most common purpose of the Cloze test is to measure reading comprehension. A true Cloze test is generally used to measure "global" reading comprehension. However, it is also argued that it gauges an underlying global linguistic ability rather than supply those skills associated with reading comprehension [14]. A number of studies, including Oller [15], Irvine, Atai, and Oller [16], Sttubbs and Tucker [17], and Aitken [18] have proved that Cloze correlates well with measures of EFL proficiency. Alderson [8] also showed that Cloze in general relates more to tests of grammar and vocabulary than to tests of reading comprehension. Although Cloze procedure is such an important feature in language testing, it has certain shortcomings. Klein- Braley and Raatz [19] discussed some flaws of the classical Cloze procedure. One of the most serious problems of classical Cloze tests, they contended, is that "the systematic nth word deletion does not necessarily produce a random sample of the text." A related problem is the differences in difficulty, reliability, and validity of Cloze tests derived from the same text. Moreover, the common practice of deriving a Cloze test from a single text introduces bias in favor of a specific topic. Scoring poses still another problem. Exact scoring is quick and easy; but it imposes an arbitrary standard of correctness that is sometimes impossible for a test taker to meet. Acceptable scoring, on the other hand, involves a trade-off: the arbitrariness of the exact scoring is partially remedied much of the ease and speed of Exact scoring is lost [19]. Alderson [12], also concluded that the Cloze procedure is not a "unitary technique", since it results in tests which are markedly different; different tests give unpredictably different measures, at least of EFL proficiency. The above mentioned problems of Cloze procedure led to the development of an alternative test.

II. C-Test: An Alternative To Cloze Tests

The C-test developed in 1981 by Klein-Braley and Raatz was proposed as an alternative to the Cloze test procedure. Drawing on the underpinning principles of Cloze procedure, the C-test is claimed to have "several advantages over the classical Cloze test." [20]. Since 1981, it has been empirically investigated. Developed as a modification of the Cloze procedure, it is meant for "testing comprehension of the more specifically linguistic principle elements in a text." [21]. Hinging on the principle of reduced redundancy, it operates on "the rule of two". The C-test comprises at least four texts, whereby starting with the second sentence of a text, the second half of every other word is deleted and the ending sentence left intact. While retaining the concept of internalized grammar, the theoretical rationale behind classical Cloze procedure which is used in all language operations, the following criteria for the procedure were included to accommodate the necessary modifications: (a) much shorter texts should be used to make up at least 100 items; (b) no problems should arise in the choice of deletion rate and starting point;(c) the deletion should be an absolutely representative sample of the elements the texts; (e) the texts should not favor the examinees; (f) only exact

scoring should be used to foster score reliability; (g) native speakers are expected to achieve virtually perfect scores; and (h) the tests should have high reliability and validity.

C-test and its developers' claims have been extensively investigated by researchers. They have scrutinized the test from different perspectives and with different criteria. The literature on C-test, however, abounds with conflicting results. Although less investigated than its elder sibling, i.e., the cloze, the C-test has been put into test from different perspectives [19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31,32, 33, 34, 35, and 36]. With regard to establishing the validity coefficient, the developers decided on an empirical validity coefficient of at least 0.5 [19]. In their studies on validity, they used three different validation criteria, namely teacher assessment, self assessment, and some other already established psychometric-structuralist tests. In a study of the C-test in Hebrew, Cohen [2] reported high correlation of 0.87 with his test of grammar and a 0.69 correlation coefficient with the performance on the selected portion of the standardized reading comprehension tests. Likewise, a 0.69 correlation coefficient was reported when correlating C-test with the Cloze version of the same passages. Still adding to the findings on C-test is a research project carried out amongst Hungarian EFL learners [28]. Besides being a reliable instrument, C-test, the researchers reported, is also valid amongst Hungarian EFL learners. Acceptable validity coefficients were obtained even when the test proved too easy or too difficult for target groups.Despite a great deal of evidence in its support, however, C-testing came under attack by Jafarpur [31] on the grounds that it fails to deliver on its fundamental claims. In the conclusion of his article, Jafarpur [31] argued that

> It is easy to construct and score but native speakers do not achieve perfect scores. Different deletion starts and deletion ratios produce different test, hence the test is not valid. Previously untried material shows acceptable reliability but does not show acceptable validity against cloze testing, and finally c-tests do not enjoy face validity.

In a critique of Jafarpur's study, Hasting [37] rejected Jafarpur's claims on the grounds that he had not only mis-constructed the claims for C-testing, but had also failed to put those claims into a fair test. The basic claims of C-testing, he concluded, were well established. Despite all their pros and cons, however, Cloze procedure and C-tests continue to be basic testing techniques, frequently used worldwide. Although non-experts (some teachers, students, and parents), tend to view them as reading comprehension tests or even as a special form of IQ tests, Cloze procedure and C-test can measure general language proficiency [19].

III. The Study

The reason behind giving a language test and obtaining a test score is interpreting that score as an indicator of what a test taker knows or what he can do with that knowledge. Furthermore, our interpretation of that test score forms the basis for decision making. As such, when using a test score, we make an implicit link between test performance and a domain of language knowledge the test taker has or something the test taker can do with language in some language use domain beyond the test itself. In other words, when we use test scores, we are essentially reasoning from evidence, using the test score as the evidence for inferences or interpretations and decisions we want to make [38]. Yet, we cannot simply draw on test score to make inferences and decisions without efficient justification. If we want to use a test score for a particular purpose, we must justify it through a rationale and supporting evidence. As Bachman [39] put it, "We need to demonstrate, with logical argumentation and empirical evidence, that the intended interpretations and uses are valid." Validity in testing and assessment has traditionally been understood to mean "discovering whether a test measures accurately what it is intended to measure"[40] or uncovering the appropriateness of a given test or any of its component parts as a measure of what it is purposed to measure"[41]. Validation in language assessment is ominously important, judging educational and linguistic policies, institutional decisions, pedagogical practices, as well as underpinnings of language theory and research. However, establishing validity in language assessment is by all accounts problematic, conceptually challenging, and difficult to achieve [42], [43]. Test validation is the process of generating evidence to support the well-foundedness of inferences concerning trait from test scores, i.e., essentially, testing should be concerned with evidence-based validity. Test developers need to provide a clear argument for a test's validity in measuring a particular trait with credible evidence to support the plausibility of this interpretative argument [44]. This process entails necessarily providing data pertaining to *context-based, theory-based* and *criterion-related* validities, together with the various *reliabilities*, or *scoring validity*. Educational measurement and language testing offer an elaborate set of procedures for conducting validation research, but rather than a prescribed invariant path – or a menu of equally-appropriate choices – the tools of validation require context-specific decisions about what and how to validate.

As was pointed out above, Cloze procedure and C-tests were developed as integrative tests of general language proficiency. As such they can provide educators with a very convenient way of gauging learners' general language proficiency in a holistic, simultaneous manner, that is, instead of designing several separate sub-tests each assessing one trait at a time, these tests allow testers to gauge several skills simultaneously through a single administration of only one test. However, their related literature abounds with contradictory results as to their validity. A group of researchers support Cloze tests [31], while the other camp suggests C-tests as a superior alternative [12]. In fact, it can be said that c-tests and cloze procedure belong to the same family, i.e., "integrative reduced redundancy" tests, and derive from the same theoretical assumptions, that is, Gestalt theory and cognitive teaching. They share the same underlying principles, it can be concluded, and differ only in terms of the implementation of those shared principles. So a fundamental question is which procedure is actually superior in fulfilling its purported advantages? Put in a nutshell, which one is superior in gauging learners' general language proficiency? If a test purports to measure general language proficiency it should "go together" with other procedures gauging the same trait.

In attempt to investigate the validity of these integrative tests as measures of general English proficiency, this study addressed the following research questions.

1) Is there any significant correlation between participants' performance on the C-test and their scores on the TOEFL?

2) Is there any significant correlation between participants' performance on the Cloze test and their scores on the TOEFL?

3) Is there any significant difference between Cloze test and C-test in terms of their correlation coefficient with the TOEFL?

The following null hypotheses were also entertained:

H_1: There is no statistically significant relationship between participants' performance on C-test and their scores on the TOEFL.

H_2: There is no statistically significant relationship between participants' performance on the Cloze test and their scores on the TOEFL.

H_3: There is no statistically significant difference between C-test and Cloze test in terms of their correlation coefficient with the TOEFL.

IV. Methodology

A. Participants

To ensure that all subjects share approximately the same background knowledge, hence alleviate, if not eliminate, test bias as much as possible, the participants of the study are selected from among the population of English majors, including both translation majors and literature majors. The sample comprises 90 seniors. They are selected from both male and female population of English majors studying at a university in Iran.

B. Instruments

In this study a Cambridge Test of English as a Foreign Language, the TOEFL comprising the listening comprehension, structure and written expressions, and reading comprehension sub-sections was used. The test included 140 questions: 50 listening comprehension questions, 40 structure and written expression questions, and 50 questions assessing test takers' listening comprehension skill. A Cloze test, comprising 25 items was also used as the second data collection instrument (See Appendix B). A C-test, including four short thematically distinct segments of connected discourse, was used as another data collection instrument. The test featured 100 questions, including 25 questions for each of its component segments (See Appendix A).

C. Procedure and data analysis

Cloze procedure and C-test were developed as integrative tests assessing test takers' general language proficiency. As such, they must be fairly capable of providing testers with a general picture of test takers' language proficiency. If these tests measure what they purport to measure, they should give testers with values which are representative of the test takers' general mastery of language being tested. In order to investigate the validity of these integrative language tests, the participants' scores on Cloze test and C-test must be compared with their performance on a "criterion" test which breaks down the language into its component parts and can, therefore, provide us with an atomistic, structural view of language proficiency. Such a test gives us separate indices indicating individual test taker's mastery of the component parts of the language concerned. Drawing on the principles of structural approach to language testing, the Test of English as a Foreign Language, the TOEFL, incorporating a set of sub-sections each measuring test taker's knowledge of a component part of language in an atomistic way, was developed as a test of general English proficiency. The test is already standardized and validated and is widely administered throughout the world. Therefore it can serve as the "criterion measure" against which our tests are validated.The participants were divided into two groups; 45 subjects were given the TOEFL and a C-test (group A), and 45 participants took the TOEFL and a Cloze-test (group B). After collecting the data, the calculated descriptive statistics pertaining to the administered tests to each group, i.e., the Cloze test, C-test and the TOEFL tests, were estimated. In order to calculate the reliability coefficients of the tests the Cronbach Alpha was used. As the next step, using the Pearson Product-moment correlation formula, the correlation coefficients between tests were estimated. In doing so, first the correlation coefficient between the C-test and individual sub-sections of the TOEFL were calculated, and, after that the global correlation coefficient, that is, the correlation coefficient between the C-test and the total TOEFL was estimated. The same procedure was replicated for the Cloze test and the TOEFL. Regarding C-test scores and Cloze test scores as the dependant variables and the TOEFL scores as the independent variable, simple linear regression analysis was done to determine how much of the variation of the subjects' scores on the C-test and Cloze test could be accounted for by their scores on the TOEFL. As the next step, the one way analysis of variance (ANOVA) was used for both groups to determine the statistical significance of the contribution of independent variable to the dependent variables.It is important to note that in order to foster objectivity, subjectively scored sub-sections of the TOEFL (composition and interview) were not considered for correlational analyses. The C-test can not adopt the multiple-choice format; however it is fairly objective in terms of scoring procedure. Test takers were required to supply the missing letters of the words and each correct restoration of missing letters is given one score. Furthermore, spelling mistakes were penalized. The C-test comprised four short thematically distinct texts. Around five minutes was allowed for each text so that the whole C-test takes 20 minutes to complete. As mentioned in [42], test performance is affected by the characteristics of the methods used to elicit test performance and the characteristic of the expected response is one of the many test facets that affect performance on language tests. Allowing for the above-mentioned factors, the researcher decided to adopt the "constructed"-as opposed to "selected"- type of expected response for the Cloze test, too. So both the Cloze test and the C-test adopted the "constructed" response type. Therefore, test takers were required to supply the missing items of the Cloze test. Likewise, since the Exact Word scoring procedure is fairly objective it was used for scoring the Cloze. To summarize, both Cloze test and C-test adopted the constructed response type and were objectively scored. The Cloze test included 25 missing items and each correct restoration was given one point and twenty minutes was allocated for its completion.

The TOEFL was administered according to its standard, recommended procedure. The test took 115 minutes to complete.

D. Design

Since the researcher has no control over the independent variable, the design of the study was correlational. Most research in applied linguistics in correlational. It is one of the most commonly used sub-sets of the so called ex post facto design[46].

V. Results

The data collection tools comprising the C-test, Cloze test, and the TOEFL were administered to the participants and the results were inputted to some statistical procedures to arrive at answers to the research questions. The descriptive statistics of the administered tests were tabulated in table 1. As can be seen in the table, subjects' performance on the criterion test, the TOEFL, was almost the same in both groups. As reported in table, participants' mean score on the TOEFL was 71.33 and 71.98 for group A and group B, respectively. Moreover, the estimated standard deviation index for group A (SD= 14.24) was slightly different from that of group B (SD= 12.77). The maximum score on the TOEFL was 99 for group A and 95 for group B, which, on the whole, are not so significantly different. Also, group A had a minimum score of 48 which was the same as that of group B. These findings may be accounted for by the random selection of participants from among the population of senior English majors.

TABLE I
DESCRIPTIVE STATISTICS: C-TEST, CLOZE TEST AND THE TOEFL

SD	Mean	Max.	Min.	No.	
Group A					
12.3667	64.44	80	38	45	C-test
5.6298	22.18	32	13	45	TOEFL-Structure
4.4955	24.80	34	17	45	TOEFL-Reading
4.5776	25.00	33	14	45	TOEFL-Listening
14.2446	71.33	99	48	45	TOEFL-Total
Group B					
3.5262	13.56	21	8	45	Cloze
4.9785	23.38	30	5	45	TOEFL-Structure
3.6581	24.60	30	18	45	TOEFL-Reading
4.2832	23.13	32	16	45	TOEFL-Listening
12.7751	71.98	95	48	45	TOEFL-Total

A. Reliability indices

In order to determine the reliability indices of data collection tools Cronbach Alpha was utilized. The results indicated that both tests had acceptable reliability indices (α:0.95 and α: 0.65 for C-test and Cloze test, respectively).However, compared with Cloze test, the C-test possessed a higher reliability index (0.65 < 0.95). Since the reliability of a test is in part a function of the length of the test, to a certain point, the longer the test the greater the reliability. So, the greater the length of the test, the more representative it should be of the true scores of persons who take it [42]. The higher reliability index of the C-test, therefore, can be attributed to its length. In other words, since the C-test incorporated a larger number of items, it would be said that it is a more reliable test than the Cloze test. It should be noted that the estimated reliability index for C-test in this study was well beyond its developers' minimum coefficient of 0.80.

As for the TOEFL, since the test is a standardized test of general English proficiency and is used by several institutions of higher education all over the world, its validity and reliability are already established. In fact, in determining the criterion-related validity of a test, the criterion test should be an already valid and reliable one.

B. Correlations

Having collected the data, the researcher used the Pearson Product-moment formula was used to estimate the correlation coefficients between C-test scores and the TOEFL scores. The results of correlational analyses for group A were reported in table 2. As displayed in the table, all correlations were significant at 0.01 level. Among the calculated correlation coefficients between the C-test and the TOEFL, however, the strongest relationship was that between the C-test and the structure sub-section of the TOEFL (r: 0.87), and the weakest correlation was that between C-test and the listening comprehension section of the TOEFL (r: 0.72). Also, there was a correlation index of 0.91 between the C-test and the TOEFL as a whole, which was quite significant, in other words there was quite a strong relationship between the C-test scores and the TOEFL scores as a whole.As for the correlation between sub-sections of the TOEFL, a correlation coefficient of 0.71 was reported between the reading comprehension and the structure sub-sections, which was the most significant one. The weakest, relationship (r:0.57), however, held between the listening and reading comprehension sub-sections. Considering the correlations between subjects' total TOEFL scores and their performance on its individual component parts, the most noticeable correlation was that between the TOEFL and the structure sub-section (r:0.90), while the weakest relationship (r:0.82) pertained to that between the TOEFL scores and listening comprehension scores.

TABLE II
CORRELATIONS FOR GROUP A: C-TEST AND THE TOEFL

TOEFL-Total	TOEFL-Listening	TOEFL-Reading	TOEFL-Structure	C-test	Group
.917	.724	.773	.875	1.000	C-test
.000	.000	.000	.000	.	Sig.
.903	.592	.713	1.000	.875	TOEFL-Structure
.000	.000	.000	.	.000	Sig.
.873	.578	1.000	.713	.773	TOEFL-Reading
.000	.000	.	.000	.000	Sig.
.822	1.000	.578	.592	.724	TOEFL-Listening
.000	.	.000	.000	.000	Sig.

1.000	.822	.873	.903	.917	TOEFL-Total
.	.000	.000	.000	.000	Sig.

The same statistical procedures, correlational and regression analyses, were replicated for participants in group B. Subjects in this group were given a Cloze test and the same TOEFL which had been administered to group A. Having collected the data, using Pearson Product-moment formula, the researcher ran a correlational analysis to calculate the correlation coefficients between the Cloze test and the TOEFL scores. The results were shown in table 3.

TABLE III
CORRELATIONS FOR GROUP B: CLOZE TEST AND THE TOEFL

TOEFL-Total	TOEFL-Listening	TOEFL-Reading	TOEFL-Structure	Cloze	Group
.820	.625	.588	.799	1.000	Cloze
.000	.000	.000	.000	.	Sig.
.864	.544	.535	1.000	.799	TOEFL-Structure
.000	.000	.000	.	.000	Sig.
.820	.502	1.000	.535	.588	TOEFL-Reading
.000	.000	.	.000	.000	Sig.
.795	1.000	.502	.544	.625	TOEFL-Listening
.000	.	.000	.000	.000	Sig
1.000	.795	.820	.864	.820	TOEFL-Total
.	.000	.000	.000	.000	Sig.

As presented in the table, there was a statistically significant correlation between the Cloze test scores and the TOEFL scores. The estimated correlation coefficients were reported as 0.82, 0.79, 0.58, and 0.62 between Cloze test scores and TOEFL scores as a whole, structure, reading comprehension, and listening comprehension scores, respectively. As can be seen in the table, all correlations were significant at 0.01 level. Considering the correlation between subjects' scores on the TOEFL sub-sections, there was the highest correlation coefficient between the structure and listening comprehension sub-tests (r: 0.54), while the weakest relationship was found to be that between listening comprehension and reading comprehension (r:0.50). Finally, the total TOEFL scores showed the strongest correlation with the structure scores (r:0.86), and the weakest relationship with the listening scores (r:0.79).

C. Regression analyses

As the next step in the research process, considering C-test scores as the dependent variable (Y) and the TOEFL scores as the independent variable (X), the researcher ran a simple linear regression analysis to determine the contribution of the TOEFL scores to C-test scores. In other words, the purpose was to determine how much of the variation of the participants' scores on the C-test can be accounted for by their scores on the TOEFL test. The results of the regression analysis for the first dependent variable, C-test scores, were reported in table 4. The findings of the regression analysis showed that 0.84 percent of variation of subjects' scores on the C-test can be predicted on the basis of their scores on the TOEFL.

TABLE IV
REGRESSION ANALYSIS: C-TEST AND THE TOEFL

Std. Error of the Estimate	Adjusted R Square	R Square	R	Model
4.9857	.837	.841	.917	1

In order to test the statistical significance of the contribution of (X) to (Y) a one way analysis of variance (ANOVA) was used (see table 5). As shown in table 8, the contribution of the independent variable (the TOEFL scores) to C-test scores was statistically significant.

TABLE V
ANOVA- GROUP A

Sig.	F	MS	df	SS		Model
.000	227.72	5660.273	1	5660.273	Regression	1
		24.857	43	1068.838	Residual	
			44	6729.111	Total	

The results of the simple linear regression analysis for the first dependant variable were summarized in table 6.

TABLE VI
COEFFICIENTS FOR GROUP A

Sig.	t	Standardized Coefficients Beta	Unstandardized Coefficients B	Std. Error		Model
.900	.126		4.299	.541	(Constant)	1
.000	15.090	.917	.059	.888	A_TOEFL	

In the table, B stands for the slope or the predicted change in Y (dependant variable) for a unit of change in (X). Constant is the value of Y when X is zero, BETA is the standardized regression coefficient , which is the number of standard deviation change in Y for a unit standard deviation change in X. In simple regression, however, BETA equals r_{xy}. It should be noted that multiple regression R is really a simple R, since we have a simple regression with only two variables.Considering the second dependent variable as Cloze test scores (Y) and the TOEFL scores as the independent variable (X), the researcher replicated the simple linear regression for the Cloze test scores and the TOEFL scores. The results were reported in table 7. As displayed in the table, 0.67 percent of variation in subjects' scores on the Cloze test could be predicted on the basis of their scores on the TOEFL. This value, however, was lower than the value of R square between C-test and the TOEFL (0.67< 0.84).

TABLE VII
REGRESSION ANALYSIS: CLOZE TEST AND THE TOEFL

Std. Error of the Estimate	Adjusted R Square	R Square		Model
2.0428	.664	.672	.820	1

In order to determine the statistical significance of the contribution of the independent variable (the TOEFL scores) to the dependant variable (Cloze scores), one way ANOVA, was used. As shown in table 8, the contribution was also significant at 0.01 level of probability.

TABLE VIII
ANOVA FOR GROUP B

Sig.	F	MS	df	SS		Model
.000	88.107	367.671	1	367.671	Regression	1
		4.173	43	179.440	Residual	
			44	547.111	Total	

The results of simple linear regression analysis for the second dependent variable (Cloze test) and the independent variable (TOEFL scores) were reported in table 9.

TABLE IX
COEFFICIENTS FOR GROUP B

Sig.	t	Standardized Coefficients Beta	Unstandardized Coefficients Std. Error	Unstandardized Coefficients B		Model
.009	-2.722		2.066	-5.624	(Constant)	1
.000	9.387	.820	.029	.270	A_TOEFL	

VI. Discussion

The obtained results of this study rejected all null research hypotheses posed at the outset:

H_1: There is no statistically significant relationship between participants' performance on C-test and their scores on the TOEFL.

H_2: There is no statistically significant relationship between participants' performance on the Cloze test and their scores on the TOEFL.

H_3: There is no statistically significant difference between C-test and Cloze test in terms of their correlation coefficient with the TOEFL.

As mentioned above, the correlation coefficient between the C-test scores and total TOEFL scores was statistically significant at 0.01 level (r: 0.91). Consequently, the first null hypothesis of this study was rejected. In other words, there is a statistically significant relationship between C-test sores and total TOEFL scores. Regarding the validity of Cloze test, the obtained correlation coefficient between the test scores and the criterion test scores, the TOEFL, was 0.82 which was also significant at the 0.01 level. Hence, the second null hypothesis of the study was also rejected, meaning that there is a statistically significant relationship between the Cloze test scores and total TOEFL scores. Comparing the obtained correlation coefficients between our tests and the criterion test, the researcher deduced that there is a stronger correlation between the C-test scores and total TOEFL scores, the implication being that C-test is more valid a test as an integrative measure of general English proficiency. Drawing on the results, the researcher came to the conclusion that there is a significant difference between C-test and Cloze test in terms of their correlation with the TOEFL. So, the third null hypothesis of the study was also rejected. As indicated by the obtained results, the researcher deducted that as the correlation matrix displays, the obtained correlation coefficients among the integrative tests (i.e., C-test and the Cloze test) and the criterion test (i.e., the TOEFL) were statistically significant. This piece of evidence can be brought to bear on our validity argument. Since the integrative tests significantly correlated with the criterion measure which in the related literature is reported by many researchers to be a valid test of general English proficiency, it can be *cautiously* claimed that these tests measure what they purport to gauge. However, considering the limitations of using criterion measures to validate tests, we are not on the right track for making any absolute validity claims. The most serious limitation is that this evidence only considers the extent to which measures of the same ability tend to agree. It does not allow for the equally overriding consideration of the extent to which scores on the test are different from indicators of different abilities. Furthermore, language tests used as the criterion whose use for this purpose may be supported by considerable experience and empirical evidence, cannot, on these grounds alone, be interpreted as valid measures of any particular ability. As such, information about criterion-relatedness is by itself insufficient for validation [42].The results of this study add further to the promising findings in the related literature of C-test and Cloze test as to the validity of these integrative language testing tools. The findings of this research are in keeping with the outcomes of several other studies reporting high validity coefficients for C-test and Cloze test as integrative language tests gauging test takers' general language proficiency. As concluded above, Cloze test and C-test are both valid tests of general English proficiency. The same results, however, were reported by several other researches[12],[2],[9],[28], and[3]. Nonetheless, as mentioned above, in this study, C-test showed a stronger correlation with the total TOEFL score than the Cloze test. Therefore, it can be considered as a more valid integrative test of general English proficiency. According to Donyei and Katona [28] "not only the C-test is a reliable and valid measure of general language proficiency, but it is also one of the most efficient language testing instruments in terms of the ratio between resources invested and measurement accuracy obtained." The aforementioned findings together with promising results reported in its related literature, justify a wider use of C-tests in language programs for diverse purposes. So, it may be postulated that C-tests might be useful in schools as achievement, diagnostic and placement tests [34]. Furthermore, the respectable correlation coefficient between C-test scores and the TOEFL structure sub-test (r:0.87) shows that C-test can be used to test certain grammar areas (e.g., tenses or word formation) by including texts incorporating several cases of the structure in question. And last but not least, the researcher believes that C-test is one of the most versatile testing tools capable of serving different functions. So, incorporating C-test in language programs can be a worthwhile pedagogical experience for pupils and teachers alike.The findings reported in this research, it should be acknowledged, brought forth some questions and possible avenues for further research. Far from being

exhausted, still there is room for further research in the field of integrative testing in general, and the Cloze procedure and C-test in particular. As cited above, due to some administration problems, and the subjective nature of their scoring system, the researcher dispensed with the interview and composition sub-tests of the TOEFL and did not input them into correlational analyses. Yet, future research projects, may allow for those sections of the TOEFL, hence having a more thorough data to draw on. Furthermore, future research may validate integrative tests against other criterion measures, e.g. IELTS, and compare he results with the findings of this study. Likewise, the tests may be validated against teacher's informal classroom assessments or other integrative measures of language ability, say, dictation. Replicating this study in another different context with different participants with the purpose of investigating its effects on the validity of C-test and Cloze test can be still another suggestion for research. Investigating the effects on the validity of integrative tests of test takers' proficiency levels can also be put forward as possible topic for research. And finally, considering such learner variables as age, learning style, sex, and IQs, research may be carried out to examine the potential effects of these attributes on the validity of C-test or Cloze test.

APPENDIX A

The C-test with answers

*Read the passages bellow and fill in the missing letters. Half of the letters of each missing word have been left out. For example, if the word is three letters long, then two letters are missing. You should spend no more than 5 minutes on each passage. (Allotted time:20minutes)*Nothing beats the heat like a refreshing dip in a swimming pool. But wh_ _ it co_ _ _ to wa_ _ _, both ki_ _ and adu_ _ _ need t_ be car_ _ _ _. Susan King's daug_ _ _ _ _ - Alison , 12, a_ _ Christy, 9, a_ _ _ in th_ _ _ grandparents' po_ _ every d_ _. King's gi_ _ _ have ma_ _ pool ru_ _ _, including n_ _ being all_ _ _ _ in t_ _ pool ar_ _ without a_ adult , n_ jumping i_ the sha_ _ _ _ end, n_ running around the pool and no holding each other under water. "Kids drown quickly and quietly" caution Jen Costello of the National safe kids campaign. Even less than an inch of water can be enough. " Parents need to actively supervise children at all times." ,she says. "Don't take your eyes off them to answer the phone, to serve food or even to watch another child."The global dominance in word processing software held by Microsoft is under threat from a new coalition. The Cillicion-Valley ba_ _ _ Google and Micro systems ha_ _ announced a formi_ _ _ _ alliance. The_ _ plan t_ make wo_ _ processing a_ _ spreadsheet prog_ _ _ _ available o_ the Inte_ _ _ _, in a dir_ _ - challenge t_ Microsoft. Indu_ _ _ _ observers s _ _ increased compe_ _ _ _ _ _ in t_ _ global soft_ _ _ _ market wi_ _ be g_ _ for cons_ _ _ _ _. The comp_ _ _ _ _ could n_ _ say wh_ _ Google wo_ _ _

Begin carr_ _ _ _ Sun's technology, including open office which was launched in 2000.There are many possible causes of insomnia. Sometimes th_ _ _ is o_ _ main ca_ _ _, but of_ _ _ several fac_ _ _ _ interacting toge_ _ _ _ will ca_ _ _ a sl_ _ _ disturbance. T_ _ causes o_ insomnia inc_ _ _ _: Psychological, phys_ _ _ _ or temp_ _ _ _ _ factors. A la_ _ of go_ _ night's sl_ _ _ can le_ _ to var_ _ _ _ problems a_ _ _ _ _ _ circle co_ _ _ develop. Profess_ _ _ _ _ counseling fr_ _ a doc_ _ _, therapist o_ sleep specialist can help individuals cope with these conditions.A popular form of recreation in Britain is attendance at dog racing. The fi_ _ _ impression o_ the ar_ _ _ is attar_ _ _ _ _. However, t_ _ races thems_ _ _ _ _ are uninte_ _ _ _ _ _ _. - a f_ _ dogs cha_ _ _ _ a tin h_ _ _- but thi_ _ _ - two mill_ _ _ people att_ _ _ them annu_ _ _ _. Out o_ two ho_ _ _, barely fi_ _ to t_ _ minutes a_ _ usually dev_ _ _ _ to t_ _ actual rac_ _ _ . There wo_ _ _ be n_ interest i_ it if it were not for the betting. Many of the audience pay little attention to the racing, but have their eyes fixed on a board which gives the number of winners.Nothing beats the heat like a refreshing dip in a swimming pool. But whEN it coMESto waTER, both kids and adULTS need tO be carEFUL. Susan King's daugHTERS Alison , 12, aND Christy, 9, aRE in thEIR grandparents' poOL every dAY. King's giRLS have maDE pool ruLES, including nOT being allOWED in t HE pool arEA without aN adult , nO jumping iN the shaLLOW end, nO running around the pool and no holding each other under water. "Kids drown quickly and quietly" caution Jen Costello of the National safe kids campaign. Even less than an inch of water can be enough. " Parents need to actively supervise children at all times." ,she says. "Don't take your eyes off them to answer the phone, to serve food or even to watch another child."The global dominance in word processing software held by Microsoft is under threat from a new coalition. The Cillicion-Valley baSED Google and Micro systems haVE announced a formiDABLE alliance. TheIR plan tO make woRD processing aND spreadsheet progRAMS available oN the InteRNET , in a dirECT - challenge tO Microsoft. InduSTRY observers sEE increased compeTITION in tHE global softWARE market wiLL be gOOD for consUMERS. The compETITION could nOT say whEN Google woULD Begin carrYING Sun's technology, including open office which was launched in 2000.There are many possible causes of insomnia. Sometimes thERE is oNE main caUSE , but ofTEN several facTORS interacting togeTHER will caUSE a slEEP disturbance. THE causes oF insomnia incLUDE: Psychological, physICAL or tempORARY factors. A laCK of goOD night's slEEP can leAD to varIOUS problems aND a vicIOUS circle coULD develop. ProfessIONAL counseling frOM a docTOR, therapist oR sleep specialist can help individuals cope with these conditions.A popular form of recreation in Britain is attendance at dog racing. The fiRST impression oF the arENA is attarACTIVE. However, tHE races themsELVES are uninteRESTING. – a fEW dogs chaSING a tin hARE- but thiRTY-two millION people attEND them annuALLY. Out oF two hoURS, barely fiVE to tEN minutes aRE usually devOTED to tHE actual racING . There woULD be nO interest iN it if it were not for the betting. Many of the audience pay little attention to the racing, but have their eyes fixed on a board which gives the number of winners.

APPENDIX B

The Cloze test with answer keys

Read the following text and fill in the blanks. (Allotted time: 20 minutes)

The cat has a _____1_____ as fascinating and mysterious as the creature itself. The true beginnings of the domestic cat are unknown, but the cat may have first appeared around3000 B.C. in a _____2_____ called Nubia, which

bordered Egypt. By 2500 B.C., the cat was domesticated in Egypt. The cat's first _____3_____ in Egypt was Mau. The mau's _____4_____ in Egypt grew rapidly; she was eventually considered guardian of the temple and was worshipped as a goddess. Besides being worshipped as goddesses, cats also had a practical _____5_____: they kept _____6_____ from overrunning the
Egyptian grain store-houses.
The Greeks were probably the first _____7_____ to recognize cats for their
mouse- catching talents. When Egyptians refused to sell or trade any of their cats, the Greeks _____8_____ several of the Egyptian cats and sold the _____9_____ of these stolen cats to Romans. The cat became the _____10_____ of liberty in ancient Rome. By the end of the eleventh _____11_____ cats were popular among sailors because of their rat-catching skills. Sailors admired cats because they _____12_____disease-infested rats which lived on ships. Many sailors believed that cats possessed
special powers that could _____13_____ them at sea.
Although the cat was held in high regard and fancied during _____14_____ times, the cat didn't fare will in Europe in the Middle Ages. Cats were associated with evil, witchcraft, and black _____15_____. Many people believed that _____16_____ regularly transformed themselves into cats. Men and women were killed for helping a_____17_____ or injured cat. During the witch-hunts in Europe many innocent people were accused of witchcraft simply because they owned cats. Black cats were especially
feared. Some legends and _____18_____ about cats exist today, like that about the nine lives of cats. Another legend that survived from Europe's Middle Ages into the present states that a black cat crossing one's path brings bad _____19_____.
Today the elegant, graceful cat has become a popular house _____20_____ throughout the _____21_____. The cat is one of the smartest of tame animals, but they are independent and harder to train. Cats are valued for their gentle, affectionate natures. They have _____22_____ memories; they _____23_____ who treats them well and who treats them badly. A cat's loyalty is earned; a cat won't stay where it is _____24_____.They respond to loving owners with loyalty, affection, and respect. Cats are noted for their keen senses: their sharp hearing, sense of smell, and ability to _____25_____ in near darkness. Perhaps Leonardo DaVinci summed it up best when he referred to the cat as "Nature's Masterpiece."

status	4	name	3	country	2	history	1
stole	8	European	7	mice	6	function	5
Destroyed	12	century	11	symbol	10	kitten	9
witches	16	magic	15	ancient	14	protect	13
Pet	20	luck	19	superstitious	18	sick	17
Mistreated	24	remember	23	good	22	world	21
						see	25

REFERENCES

[1] J. W. Oller, Language tests at School: A Pragmatic Approach, London: Longman, 1979.
[2] A. D. Cohen, "On taking tests: what the students report," Language Testing, vol. 1, pp. 70-81, 1984.
[3] C. J. Weir, Communicative Language Testing, Hemel Hempstead: Prentice Hall, Inc., 1998.
[4] J. D. Brown, "Correlational study of four methods for scoring Cloze tests," Unpublished master's thesis, University of California, USA, 1978.
[5] J. W. Oller, and J. Jonz, Eds. Cloze and Coherence, Cranbury, NY: Bucknell University Press, 1994.
[6] M. R. Sampson, and L. D. Briggs, "A new technique for Cloze scoring: A semantically consistent method," Clearing House, vol. 57, no. 4, pp. 177-79, 1983.
[7] M. Pikulski, and A. W. Tobin, "Cloze procedure as an informal assessment technique," In J. J. Rikulsky and T. Shanahan, Eds. Approaches to the informal evaluation of reading, Newark: International Reading Association, 1982.
[8] J. C. Alderson, "The effect on the Cloze test of changes in deletion frequency," Journal of Research in Reading, vol. 2, pp. 108-18, 1979.
[9] A. G. Sciarone, and J. J. Schrool, "The Cloze test: or why small isn't always beautiful," Language Learning, vol. 39, no. 3, 415-38, 1989.
[10] M. Soudek, and L. Soudek, "Cloze after thirty years: New uses in language teaching," ELT Journal, Vol. 37, no. 4, pp. 335-40, 1983.
[11] J. C. Alderson, "The effect of certain methodological variables on Cloze test performance and its implications for the use of Cloze procedure in EFL testing," Paper presented at the fifth International Congress of Applied Linguistics, Montreal, 1978.
[12] J. C. Alderson, "The Cloze Procedure and Proficiency in English as a Foreign Language," In J. W. Jr. Oller, Ed. Issues in Language Testing Research, Rowly, Mass: Newbury House, 1983, pp. 205-217.
[13] J. D. Brown, "A closer look at Cloze: validity and reliability," In J. W. Oller, Ed. Issues in Language Testing Research, Rowley, Mass: Newbury House, 1983, pp. 237-250.
[14] J. B. Heaton, Writing English Language Tests. London: Longman, 1990.
[15] J. W. Oller, "Cloze tests of second language proficiency and what they measure," .Language Learning, vol. 23, pp. 101-105, 1973.
[16] P. Irvine, P. Atai, and J.W. Oller, "Cloze, dictation, and the Test of English as a Foreign Language," Language Learning, vol. 24, pp. 245-52, 1974.
[17] J. B. Stubbs, and G. R. Tucker, "The Cloze test as a measure of English proficiency," The Modern Language Journal, vol. 58, pp. 239-41, 1974.
[18] K. G. Aitken, "Using Cloze procedure as an overall language proficiency test," TESOL Quarterly, vol. 11, pp. 59-68, 1977.
[19] C. Klein-Braley, and U. Raatz, "A survey of research on the C-test," Language Testing, vol. 1, no. 2, pp. 134-46, 1984.
[20] U. Raatz, and C. Klein-Braley, "Ein neuer Ansatz zur Messung der Sprachleistung. Der C-test: Theorie und Praxis," In R. Horn, K. Ingencamp, and R. Jager, Eds. Tests und Trends, Jahrbuch der Padagogischen Diaanostik. Weinheim: Beltz: 1983, 107-38.
[21] C. J. Weir, Understanding and Developing Language Tests, Prentice Hall, Inc., 1993.
[22] A. Cohen, M. Segal, and R. Weiss, "The C-test in Hebrew," Language Testing, vol. 1, pp. 221–225, 1984.
[23] R. Grotjahn, "Test validation and cognitive psychology: Some methodological considerations," Language Testing, vol. 3, pp. 159–185, 1986.
[24] R. Grotjahn, "How to construct and evaluate a C-test: A discussion of some problems and some statistical analyses," In: R. Grotjahn, C. Klein-Braley, and D. K. Stevenson, Eds. Taking their Measure: The Validity and Validation of Language Tests, Brockmeyer, Bochum, Germany, 1987, pp. 219– 253.
[25] R. Grotjahn, Ed. "Der C-test: Theoretische Grundlagen und praktische Anwendungen," AKS-Verlag, Bochum, 2002.
[26] U. Feldmann, and B. Stemmer, "Thin____ aloud a____ retrospective da____ in c-te___ taking: Diffe_____languages - diff_____ learners - sa____ approaches?" In: C. Faerch, C. Kasper, Eds. Introspection in Second Language Research, Clevedon: Multilingual Matters, 1987.

[27] C. Cleary, "The C-test in English: Left-hand deletions," RELC Journal, vol. 19, pp. 26–38, 1988.
[28] C. Chappelle, and R. Abraham, "Cloze method: What difference does it make?" Language Testing, vol. 7, pp. 121–146, 1990.
[29] M. Hood, "The C-test: A viable alternative to the use of the cloze procedure in testing?" In L. Arena, Ed. Language Proficiency. Plenum Press, New York, 1990, pp. 173–189.
[30] Z. Dornyei, and L. Katona, "Validation of C-test among Hungarian EFL learner," Language Testing, vol. 9, pp. 187–206, 1992.
[31] T. Kamimoto, "An inquiry into what a C-test measures," Fukuoka Women's Jr. College Studies, vol. 44, pp. 67–79, 1992.
[32] T. Kamimoto, "Tailoring the test to fit the students: Improvement of the C-test through classical item analysis," Fukuoka Women's Junior College Studies, vol. 30, pp. 47–61, 1993.
[33] A. Jafarpur, "Is C-test superior to Cloze?" Language Testing, vol. 12, pp. 194–216, 1995.
[34] A. Jafarpur, "Can the C-test be improved with classical item analysis?" System, vol. 27, pp. 79–89, 1999.
[35] G. Sigott, J. Kobrel, "Deletion patterns and C-test difficulty across languages," In: R. Grotjahn, Ed. Der C-test, Theoretische grundlagen und praktische anwwndungen, vol. 3, pp. 159–172, Brockmeyer, Bochum, 1996.
[36] C. Klein-Braley, "C-tests in the context of reduced redundancy testing: An appraisal," Language Testing, vol. 14, pp. 47–84, 1997.
[37] C. Ikeguchi, "Do different C-tests discriminate proficiency levels of EL2 learners?" JALT Testing & Evaluation SIG Newsletter, vol. 2, pp. 3–8, 1998.
[38] E. Babaii, and H. Ansary, "The C-test: a valid operationalization of reduced redundancy principle?" vol. 29, pp. 209–219, 2001.
[39] A. J. Hastings, "The Focal Skills Approach: an Assessment," In F. Ekman, et al., Eds. Second Language acquisition Theory and Pedagogy, Hillsdale, NJ: Lawrence Erlbaum, 1995, pp. 29-43.
[40] R. J. Mislevy, "Test theory reconceived," Journal of Educational Measurement, vol. 33, no. 4, pp. 379-416, 1996.
[41] L. F. Bachman, Statistical Analysis for Language Assessment, Cambridge: Cambridge University Press, 2003.
[42] A. Hughes, Testing for language teachers, 1st ed. Cambridge: Cambridge University Press, 1989.
[43] G. Henning, A guide to language testing. Cambridge, MA: Newbury House, 1987.
[44] L. F. Bachman, Fundamental Considerations in Language Testing, Oxford: Oxford University Press, 1990.
[45] P. Groot, "Language testing in research and education: The need for standards," In J. De Jong, Ed. Standardization in Language Testing, AILA Review, vol. 7, pp. 9-23, 1990.
[46] M. T. Kane, "An argument-based approach to validity," Psychological Bulletin, vol. 122, no. 3, pp. 527–35, 1992.
[47] E. Hatch, and H. Farhady, Research Design and Statistics for Applied Linguistics. Rowley, Massachusetts: Newbury House, 1982.