Do C-Test and Cloze Procedure Measure what they Purport to be Measuring? A Case of Criterion-Related Validity
Masoud Saeed, 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 Saeed 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 Khorasan 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 a 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 tests 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 different. In fact, the
we make an implicit link between test performance and a basis for decision making. As such, when using a test score, further, our interpretation of that test score forms the test taker knows or what he can do with that knowledge.

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 misconstrued 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, scoring should be used to foster score reliability; (g) 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 purported 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:

H₁: There is no statistically significant relationship between participants’ performance on C-test and their scores on the TOEFL.
H₂: There is no statistically significant relationship between participants’ performance on the Cloze test and their scores on the TOEFL.
H₃: 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 expression, 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 correlation 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 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 test 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 is 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 inputed into 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 1 | DESCRIPTIVE STATISTICS: C-TEST, CLOZE TEST AND THE TOEFL |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| SD Group A      | Mean            | Max.            | Min.            | No.             | C-test          | TOEFL-Structure |
| 12.3667         | 64.44           | 80              | 38              | 45              | C-test          | TOEFL-Structure |
| 5.6298          | 22.18           | 32              | 13              | 45              | C-test          | TOEFL-Structure |
| 4.9455          | 24.80           | 34              | 17              | 45              | C-test          | TOEFL-Reading   |
| 4.5776          | 25.00           | 33              | 14              | 45              | C-test          | TOEFL-Listing   |
| 14.2446         | 71.33           | 99              | 48              | 45              | TOEFL-Totall   |
| Group B         | 3.5262          | 13.56           | 21              | 8               | 45              | TOEFL-Structure |
| 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-Listing   |
| 12.7751         | 71.98           | 95              | 48              | 45              | TOEFL-Totall   |

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, ( 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           |
| Sig.             | Sig.             | Sig.             | Sig.             | Sig.             | Sig.             | Sig.             |
| -.947            | .000             | -.744            | .000             | -.723            | .000             | .875             | .000             |
| .903             | .592             | .713             | 1.000            | .875             | .000             | .000             | .000             |
| .873             | .578             | 1.000            | .713             | .773             | .000             | .000             | .000             |
| .822             | 1.000            | .578             | .592             | .724             | .000             | .000             | .000             |
| .000             | .000             | .000             | .000             | .000             | .000             | .000             | .000             |

World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
International Scholarly and Scientific Research & Innovation 5(2) 2011
194
ISNI:0000000091950263
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>
<thead>
<tr>
<th>TOEFL-</th>
<th>Cloze</th>
<th>TOEFL-</th>
<th>TOEFL-</th>
<th>TOEFL-</th>
<th>Cloze</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>.820</td>
<td>.620</td>
<td>.880</td>
<td>.790</td>
<td>1.000</td>
</tr>
<tr>
<td>Sig.</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
</tbody>
</table>

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 replicated the 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>
<thead>
<tr>
<th>TOEFL-</th>
<th>C-Test</th>
<th>TOEFL-</th>
<th>TOEFL-</th>
<th>TOEFL-</th>
<th>C-Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>.820</td>
<td>.620</td>
<td>.880</td>
<td>.790</td>
<td>1.000</td>
</tr>
<tr>
<td>Sig.</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
</tbody>
</table>

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.

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 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).

The results of the simple linear regression analysis for the first dependant variable were summarized in table 6.
In order to determine the statistical significance of the contribution of the independent variable (the TOEFL scores) to the dependent variable (Cloze scores), one way ANOVA, was used. As shown in table 8, the contribution was also significant at 0.01 level of probability.

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.

VI. DISCUSSION

The obtained results of this study rejected all null research hypotheses posed at the outset:
H1: There is no statistically significant relationship between participants’ performance on C-test and their scores on the TOEFL.
H2: There is no statistically significant relationship between participants’ performance on the Cloze test and their scores on the TOEFL.
H3: 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 scores 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
APPENDIX A

The C-test with answers

Read the passages below 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: 20 minutes) Nothing beats the heat like a refreshing dip in a swimming pool. But wh_it co_t to wa_ _ _ _ , both ki_and adu__ need t_e be car__ _ _ Susan King’s dau_ghter _ _ _ - Alison, 12, a__ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
bordered Egypt. By 2500 B.C., the cat was domesticated in Egypt. The cat's first name in Egypt was Mau. The maid's four 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: they kept mice from overrunning the Egyptian grain store-houses.

The Greeks were probably the first to recognize cats for their mouse-catching talents. When Egyptians refused to sell or trade any of their cats, the Greeks eight several of the Egyptian cats and sold the nine of these stolen cats to Romans. The cat became the 10 of liberty in ancient Rome. By the end of the eleventh century, cats were popular among sailors because of their rat-catching skills. Sailors admired cats because they recognized cats for their unique use in near darkness. Perhaps Leonardo DaVinci summed it up best when he referred to the cat as “Nature’s Masterpiece.”

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 well in Europe in the Middle Ages. Cats were associated with evil, witchcraft, and black magic. 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 disliked.

Some legends and 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 luck. Today the elegant, graceful cat has become a popular house 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 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 see in 25 near darkness. Perhaps Leonardo DaVinci summed it up best when he referred to the cat as “Nature’s Masterpiece.”

### REFERENCES


[34] A. Jafarpur, "Can the C-test be improved with classical item analysis?" System, vol. 27, pp. 79–89, 1999.