# The effects of context-dependent and context-independent test design on Iranian EFL learners' performance on vocabulary tests

Los efectos del diseño de exámenes dependientes e independientes del contexto sobre el desempeño de estudiantes ILE iraníes en los exámenes del vocabulario

**Mahmoud ABDI TABARI**
Oklahoma State University
(Stillwater, OK, USA)

## Abstract

*This study explores the role of context in vocabulary assessment. In addition, it examines how learners from groups of slightly different proficiency levels performed on 2 context-dependent and context-independent tests. Performances of 40 L2 English university students on identical items on two tests, the matching test (context-independent test) and the C-test (context-dependent test), were compared. The result showed that all participants performed slightly better on the matching test than the C-test. This suggests that context did not play a major role in their performance in the C-test. Moreover, high-intermediate learners performed much better on both tests than participants at the intermediate level. It is therefore concluded that higher proficiency learners use more context in response to items in the test than do lower proficiency learners. These results help identifying the best format for vocabulary assessment.*

**Key Words**: context-dependent test; context-independent test; vocabulary tests.

## Resumen

*Este estudio explora el papel del contexto en la evaluación de vocabulario. Además, examina cómo los alumnos de los grupos de niveles de competencia ligeramente diferentes se desempeñaron en 2 exámenes, dependiente e independiente del contexto. Se compararon los desempeños de 40 estudiantes universitarios cono inglés L2 en artículos idénticos en dos exámenes, el examen de correspondencia (un examen independiente del contexto) y el examen de C (un examen dependiente del contexto). El resultado mostró que todos los participantes se desempeñaron un poco mejor en el examen de correspondencia que en el examen de C. Esto sugiere que el contexto no jugó un papel importante en sus desempeños del examen de C. Por otra parte, los alumnos de alto intermedio realizan mucho mejor en las 2 pruebas que los participantes del nivel intermedio. Por tanto, se concluyó que los estudiantes de competencia superiores utilizan más contexto en respuesta a los elementos en el examen que hacer los alumnos de competencia más bajas. Estos resultados ayudan a identificar el mejor formato para la evaluación de vocabulario.*

**Palabras Claves**: examen dependiente del contexto; examen independiente del contexto; examen del vocabulario.

## INTRODUCTION

The assessment of vocabulary learning often has a significant place in classroom tests. In this context, the design features of the test rely on a concise definition of the construct of vocabulary. Issues such as what a "word" means, whether vocabulary includes multi-word lexical items, and how words are influenced by context all arise when one seeks to define what any particular vocabulary test is measuring. One thing that makes construct definition rather difficult in the area of second language vocabulary is the variety of theoretical concepts and frameworks that scholars propose to account for vocabulary acquisition, knowledge, and use (Bachman & Palmer, 2000). Although certain tools, such as the concept of vocabulary size, and Nation's analysis (1999; cited in Read, 2000, p. 57) of what it means to know a word are widely known and used, there is no comprehensive, generally accepted conceptual framework for L2 vocabulary knowledge. Schwartz's model of the vocabulary knowledge (1984; cited in Read, 2000, p. 23) provides a broad framework to distinguish two dimensions of vocabulary knowledge:

> The test may be called "breadth" of knowledge, by which we mean the number of words for which the person knows at least some of the significant aspects of meaning …. [There] is a second dimension of vocabulary knowledge namely the quality or "depth" of knowledge. We shall assume that, for most purposes, a person has a sufficiently deep understanding of a word if it conveys to him or her all of the distinctions that would be understood by an ordinary adult under normal circumstances'. (Schwartz 1984; cited in Read, 2000, p. 23)

One practical difficulty with "breadth" test is that word forms are presented out of context. When the word form is presented by itself, there is no indication as to which meaning or use of the word the test-writer intends to assess. Hence, more recent tests have emphasized the value of the quality or depth of knowledge of vocabulary. In such tests, items are presented in text to assess the ability of learners to make inferences about unfamiliar words that they encounter in their reading or listening.

There has been little research that explicitly addresses the role of context in assessing knowledge of individual words, although testing vocabulary in context is chiefly pursued in the shift away from de-contextualized discrete-point tests of the 1970s. One type of test that has received a lot of attention has been the cloze procedure. Different formats of this test were designed, and each claims

that it is a panacea for the deficiencies inherent in the previous one. The C-test, as the latest version of cloze tests, is perhaps the most valid one; nonetheless, it has not passed unchallenged, either.

## Research background

Few research studies would dispute that receptive vocabulary is probably larger than productive vocabulary (Goldestein, 2004; Laufer, 2002; Meara, 2000; Nation, 2000), and some level of receptive knowledge of a word must exist in order for the word to be produced. Nonetheless, Meara and Jones (1988; cited in Meara & Fitzpatrick, 2000, p. 21) claim that the difficulty at the conceptual level is to find the criteria for distinguishing words that have receptive status from those that are part of a person's productive ability, and the problem is to locate the threshold at which the word passes from receptive to productive status. They further stipulate that measuring the control of productive vocabulary is very difficult because the implicit assumption is that productive vocabulary knowledge can reasonably be inferred from measures of receptive knowledge. Moreover, the vocabulary produced by a learner, whether in written or spoken form, tends to be so context-specific that is difficult to calculate the true size or range of the learner's productive vocabulary from any small sample. It is also possible that test takers do not use the vocabulary that the researcher intended to assess. They might resort to the strategy of avoidance; that is, the ability to ignore the use of unknown words in order to overcome limitations in language knowledge (Oxford, 1990, p.15). Thus, the tendency in vocabulary research has been to prefer very controlled, selective tasks, in which the target words are presented in a limited context. Such tests are the most cost-effective and efficient way of eliciting data from testers to make a rough estimate about their productive vocabulary skills.

Therefore, controlled, productive tests allow modern testers to deal with problems in free productive vocabulary tests. One creatively attractive alternative is the C-test, in which a series of short texts are prepared for testing by deleting the second half of selected content words. It has been considered the most appropriate test design for vocabulary assessment (Read, 2000). Testers can elicit target words from each of the 5 word-frequency bands: 2000, 3000, 5000, University Word List, and 10,000. This test is a refinement of cloze tests. Some testers indeed simply treat it as a variant of the now widely used cloze-test

because the testers mutilate the selected particular words or apply the fixed-ratio deletion in a series of short texts (Chapelle & Abraham, 1990; cited in Connelly, 2006, p. 130). If C-tests are chosen by teachers and researchers specifically as a vocabulary measure for achievement and diagnostic tests, then content words can be chosen and, in particular, words that could be restored by making use of contextual clues. The advantages of such a test are that, because it is a discrete, selective, and controlled test, it can easily identify what has not been learnt by test-takers. It is also possible to control the characteristics of every item at the word level and, subsequently, the total test for facility and discrimination. This test is easy to construct and score, and relatively short texts yield good internal consistency and validity (Klein-Braley & Raatz 1984; cited in Babaii & Ansary, 2001, p. 220). However, "analyzing the errors cannot show how they arrive at the answer they write in each blank: do they automatically come up with the response, or is conscious thinking involved?" (Read, 2000, p. 113).

What a C-test actually measures has been an issue of debate for many years, though these tests have been applied to more than 20 languages. C-tests have been used in numerous contexts and for various purposes; for example, as a placement test, as an anchor test in test equating, and as a research instrument in cognitive and applied linguistics (Ekes & Grotjahn, 2006). In support of C-tests, Dorney and Katona (1992; cited in Ekes & Grotjahn, 2006, p. 292) point out: "It [the C-test] is one of the most efficient language testing instruments in terms of the ratio between resources invested and measurement accuracy obtained".

It is worth mentioning that when test-takers are completing the C-test, they merely consider short texts in the C-test as a particular context of use and employ them as a tool for eliciting evidence of the mental processes dealing with linguistic resources. The reason lies in the fact that these test-takers are lacking necessary literacy skills to iron out linguistic resources and work on short texts as a context of use. They see many words as isolated items in most cases, without looking at the text. Stalnaker and Kurath (1935; cited in Read, 2000) compared two methods of testing knowledge of German vocabulary at the elementary level. The first method they used for testing knowledge of vocabulary was the *best-answer* test, including multiple-choice items, with each target word is shown in isolation. The second methodology they employed was the *context* test, including a reading passage with all 100 of the same target words. The test takers had to provide the English equivalent of each underlined word. When the two tests were

administered to students of German at the University of Chicago, they showed similar performances on both tests. These two tests had a high correlation with each other and showed very close correlations with other measures of the students' ability. Therefore, researchers reached a conclusion that the two tests were equally valid measures of the same ability. Although the researchers did not mention which type of test should be preferably chosen, the implication was that there was no real advantage in testing words in context.

## METHOD

### Participants

A selection of participants drawn from a group about 60 freshmen students of English studying at Islamic Azad University, North Tehran Branch (Iran) participated in this study. The participants were finishing their first semester, and included both male and female students between 18 and 25 years of age. The results of a pre-test, which had been developed by the researcher based on the book series *Reading for Ideas* (Vols. 1 & 2) (Khoii, Fotovat Ahmadi, & Shokouhian, 2007) helped divide the participants into intermediate and high-intermediate level groups. Students who scored above the median were assigned to the high intermediate group while those who scored below the median were assigned to the intermediate group. Eventually, 40 students were selected as participants in this study: 20 students were assigned to the high-intermediate group and 20 to the intermediate group.

### Instrumentation

Three tests were utilized in this study: A teacher-made general proficiency test used as a placement test, a matching test, and a C-test. These tests are explained in the subsequent sections.

### *Placement Pretest*

This test was a test of general proficiency consisting of 30 grammar items, 30 vocabulary items developed by the researcher based on the general English books they used to study, and 30 reading comprehension items adopted from CBT TOEFL (2004 version). The general proficiency test went through item analysis in order to make sure that it was a reliable test. Items with an Item Facility (IF) that

fell within a range between 0.30 and 0.70 were considered acceptable. Items that had the highest Item Discrimination (ID) indexes were also selected for the test. Overall, 26 items were discarded from the test; as a result, the number of items was reduced from 90 to 64, and the proficiency scores of the students were based on the remaining items. The reliability coefficient of the test was 0.84, which indicated that the test was reliable. The placement pretest was used to divide participants into two groups, namely, intermediate and higher intermediate.

### C-test

A C-test was constructed according to the instructions put forward by Read and Klein-Braley (2006). Based on the readability formula, texts that were suitable for the proficiency level of the participants in the intermediate group were selected. Six texts that were authentic, self-contained, and varied in subject matter were selected to avoid test bias. The researcher chose them on the basis of the vocabulary in the first 3 units of the series, *Reading for Ideas* (Vols. 1 & 2) (Khoii, Fotovat Ahmadi, & Shokouhian, 2007). It should be noted that the difficulty levels of the texts were estimated using the Fog Formula, a simple way to measure the level of reading ease or difficulty for any piece of writing. The Fog Formula is built on the premise that short words and sentences are easy to understand, but long words and sentences are difficult to grasp. The formula does not determine directly if the writing is too basic or advanced for the audience. It does not even measure reading comprehension, but rather is a mathematic calculation of the ease or difficulty of reading. To indicate this ease or difficulty, the formula reflects the equivalent of a grade level.

In this study, the readability indexes of the texts according to the Fog Formula were about 11, which was of almost the same difficulty level as the texts in the subjects' course books. The texts were relatively of different lengths due to semantic considerations. In each text, the first and last sentences were left intact. Then a rational method of deletion of content words was used to mutilate the texts. The deletion principle was applied somewhat flexibly, so that the selected words would appear in different parts of the texts. Throughout the texts, the missing part of each word was indicated by a single underline of constant length (see Appendix 1: The C-test). Test-takers were provided with enough information about how to complete the missing parts at the outset of the test. The C-test consisted of 6 texts with 30 blanks in total, representing the items deleted. The

vocabulary selected for the C-test was then used for constructing the matching test. The number of items in both tests was the same.

In the process of the C-test development, the issue of how to score acceptable variants and orthographic errors was also considered. Since orthographic correctness is important in academic contexts, and since the scoring of spelling errors as incorrect made the C-test more reliable, spelling errors were counted as incorrect in the final version. The reliability of the C-test was 0.81, indicating that the test was relatively reliable.

### Matching test

On a matching test, test takers are required to match the words to the definitions. The matching tests consisted of 30 items based on the same vocabulary used in the C-test. The researcher constructed the C-test based on the texts in the subjects' textbooks. The matching test was then constructed using the mutilated words in the C-test but in their complete forms. The matching test was developed following the guidelines put forward by Read (2000). One or two extra definitions were used in the right-hand column to avoid the process of the elimination of words when the actual knowledge of the meaning of words was lacking by test takers. The standard practice was to group a number of words together with a large number of definitions. As mentioned previously, the words selected for the construction of the test were the same as the deleted words in the C-test (see Appendix 2: Matching test). Moreover, all of the target items belonged to the same word class (for example, nouns or adjectives) to avoid the use of structural clues for answering. The reliability of the matching test (0.90) indicated that the test was reliable.

### Procedure

The purpose of this study was to determine whether learners performed differently on vocabulary C-tests and matching tests. It was also intended to determine the role of proficiency level in the students' performance on these tests. Thus in the first step, a pretest (which was a general proficiency test developed by the researcher) was administered to 60 students in order to determine their proficiency level. Based on their scores on the pretest, the subjects were divided into two groups, intermediate and high intermediate. Those

who scored above the median were classified as high intermediate and those who scored below the median were classified as intermediate learners.

The matching test was administered to the same subjects at the same time as the pretest. To minimize the possible learning effect, especially on the vocabulary measures, the vocabulary C-test was given to the participants after 2 weeks. The whole administration took about 2 weeks.

## RESULTS

In this study, the researcher tried to investigate the effect of context dependent and context independent vocabulary test formats on the performance of EFL freshmen students of English translation studying at the Islamic Azad University, North Tehran Branch. In other words, he intended to see whether learners perform differently on a matching test (context independent) and a C-test (context dependent). The following research questions were put forward by the researcher at the beginning of the study:

1. Is there any significant difference between a C-test and a matching test measuring the vocabulary knowledge of Iranian foreign language learners?
2. Do learners at different proficiency levels perform differently on C-tests and matching tests?

To answer the abovementioned questions, the researcher devised a general proficiency test of English to assess the learners' knowledge of English and to place them in two groups, intermediate and high intermediate. First, he administered the general proficiency test to the learners. As previously noted, the reliability index of this test, calculated by the KR-21 formula, was 0.84, which indicated that the teacher-made test was reliable. The item characteristics of the tests are presented in the remainder of this section. To see how test takers performed on the general proficiency test, the descriptive statistics of subjects' performance on this test are presented in Table 1.

As shown in Table 1 the mean, median and mode of the proficiency test were calculated in order to determine the general language proficiency of the students. The mean, median, and mode of the proficiency test were 36.27, 35, and 36, respectively. The fairly close figures for these three statistics indicate that the students' scores on the proficiency test enjoyed an almost normal distribution.

**Table 1. Descriptive statistics for general proficiency test.**

| PROFICIENCY | | |
|---|---|---|
| N | Valid | 40 |
| | Missing | 0 |
| Mean | | 36.2750 |
| Median | | 35.0000 |
| Mode | | 36.00 |
| Std. Deviation | | 13.76541 |
| Skewedness | | 0.448 |
| Std. Error of Skewedness | | 0.374 |
| Kurtosis | | 0.390 |
| Std. Error of Kurtosis | | 0.733 |
| Minimum | | 15.00 |
| Maximum | | 67.00 |

**Note**: Multiple modes exist. The largest value is shown.

Based on these results, and in order to probe the second research question, it was decided to divide the students into two groups of high and low general language proficiency groups on the basis of their median score on the proficiency test: that is, 35. In other words, the students whose scores were equal to or above 36 comprised the high proficiency group, while the students who scored 35 and below formed the low proficiency group. Employing this method yielded 2 groups of students of equal size, with 20 cases in each group: 20 students were assigned to the high intermediate group and 20 to the intermediate group.

To answer the first research question, a paired-sample t-test was run to compare the mean scores of the students on the matching test and C-test. The t-observed value was 2.28 (see Table 2). This t-value was higher than the critical value of t at 39 degrees of freedom: that is, 2.02.

**Table 2. Paired-sample t-test comparing the matching test with the C-test.**

| | Paired Differences | | | | | | | |
| | | | | 95% Confidence Interval of the Difference | | | | |
| | Mean | Std. Deviation | Std. Error Mean | Lower | Upper | t | df | Sig. (2-tailed) |
|---|---|---|---|---|---|---|---|---|
| **MATCHING vs. C-TEST** | 2.52500 | 6.99812 | 1.10650 | .28689 | 4.76311 | 2.282 | 39 | .028 |

Since the observed value of t was higher than its critical value, it can be inferred that there is a statistically significant difference between the students' mean scores on the matching test and C-test. Thus, as no significant difference between the students' mean scores on the matching test and C-test, the null-hypothesis is rejected. As displayed in Table 3, the students performed better on the matching test, with a mean score of 12.77.

**Table 3. Descriptive statistics of the matching test and the C-test.**

| | Mean | N | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|
| **MATCHING TEST** | 12.7750 | 40 | 4.59368 | 0.72633 |
| **C-TEST** | 10.2500 | 40 | 6.96419 | 1.10114 |

To investigate the second research question, a multivariate analysis of variances (MANOVA) was run to investigate the effect of the proficiency level of the students on their performance on the matching test and C-test. Table 4 displays the descriptive statistics for the high and low proficiency groups on the matching test and C-test. As the results indicate, the high proficient students performed better than the low proficiency group on both matching test and C-test. The mean scores for the high and low proficient students on the matching test were 15 and 10.55. Their mean scores on the C-test also were 13.75 and 6.75 respectively.

**Table 4. Descriptive statistics of the matching test and the C-test by proficiency levels.**

|  | PROFLEVEL | Mean | Std. Deviation | N |
|---|---|---|---|---|
| **MATCHING TEST** | LOW PROFICIENCY | 10.55 | 4.35 | 20 |
|  | HIGH PROFICIENCY | 15.00 | 3.72 | 20 |
|  | Total | 12.77 | 4.59 | 40 |
| **C-TEST** | LOW PROFICIENCY | 6.75 | 3.91 | 20 |
|  | HIGH PROFICIENCY | 13.75 | 7.64 | 20 |
|  | Total | 10.25 | 6.96 | 40 |

The F-observed value for the effect of the proficiency level of the students was 11.24 (see Table 5). This amount of F-value at 2 and 37 degrees of freedom was higher t value of 3.25. Since the observed value of F exceeded its critical value, it can be claimed that there was a significant difference between the overall high and low proficiency groups on the matching test and C-test. In addition, the SPSS produced four F values the first of which (that is, Pillai's Trace) was the most robust (Field, 2005); even in conditions where the groups were not balanced, there were absent cases and the assumptions of normality were violated. In other words, when using Pillai's Trace, one does not need to worry about any of the assumptions of MANOVA.

**Table 5. Effect of proficiency on the matching test and the C-test.**

|  | Value | F | Hypothesis df | Error df | Sig. |
|---|---|---|---|---|---|
| **Pillai's trace** | 0.378 | 11.242 | 2 | 37 | 0.000 |
| **Wilks' lambda** | 0.622 | 11.242 | 2 | 37 | 0.000 |
| **Hotelling's trace** | 0.608 | 11.242 | 2 | 37 | 0.000 |
| **Roy's largest root** | 0.608 | 11.242 | 2 | 37 | 0.000 |

Therefore, as learners at different proficiency levels do not perform differently on C-test and matching test, the second null-hypothesis is rejected. As depicted in

table 6, there is a significant difference between the mean scores of the high and low proficiency groups on the matching test (P= .001< .05). The results also indicate that there is a significant difference between the mean scores of the high and low proficiency groups on the C-test (P= .05).

**Table 6. Post-hoc comparison of the matching test and the C-test by proficiency.**

| Dependent Variable | (I) PROFLEVEL | (J) PROFLEVEL | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval of the Difference | |
|---|---|---|---|---|---|---|---|
| | | | | | | Lower Bound | Upper Bound |
| MATCHING TEST | HIGH PROFICIENCY | LOW PROFICIENCY | 4.450* | 1.282 | .001 | 1.854 | 7.046 |
| C-TEST | HIGH PROFICIENCY | LOW PROFICIENCY | 7.000* | 1.920 | .001 | 3.112 | 10.888 |

* The mean difference is significant at the .05 level.

In general, the performances of students in the intermediate and high-intermediate groups on the 2 tests of vocabulary (that is, the C-tests and the matching tests) were significantly different; that is, students in both groups performed differently on two distinctive tests of vocabulary. However, both groups had better performance on the matching test than on the C-test. In particular, the high-intermediate group outperformed the intermediate group on both vocabulary tests. The high-intermediate group showed a consistency in their performance on the matching test (M=15.00) and the C-test (M=13.75), whereas the intermediate group, with limited English proficiency, had an inconsistency in their performance on the matching test (M=10.55) and the C-test (M=6.75), and performed differently on these 2 tests. To sum up, the changes in performance on these 2 tests suggest the intermediate group with limited English proficiency had more difficulty on the C-test than the higher intermediate group, who were better able to perform on both tests, particularly in the case of the C-test.

## DISCUSSION

The major development that has been advocated in recent years is a rethinking of vocabulary assessment from a three-component perspective on test purpose and an internationalist perspective to construct definition when it is appropriate. An internationalist approach to inferences requires that vocabulary knowledge and

use should be defined in relation to particular context. In fact, the adequacy of the learners' lexical knowledge is assessed through their ability to use a word in relation to its particular context. This approach has implications not only for the selection of lexical items to be assessed but also the nature of the task the learners are asked to perform. Therefore, a "contextualized" test design has little value unless the test takers are required to engage with the contextual features in a meaningful way (Read, 2001).

It is important to identify the sources of difficulty that make the results of the 2 tests non-parallel, in spite of the fact that both tests used the same vocabulary. The enduring popularity of matching test (Vocabulary Level Test) for a whole range of uses must owe much to its availability, simplicity and convenience to measure receptive knowledge of lexical items, but the C-test (a new version of cloze test) to its convenience, high reliability, and validity. The deletion procedure of this test taps into the test-taker's productive knowledge of vocabulary. The first research question concerned the significant difference between the students' mean scores on the matching test and the C-test. As displayed in the result, the students performed better on the matching test with a mean score of 12.77.

The present investigation reveals that the participants treated most words as standalones even when they were embedded in a context-dependent test. Therefore, it seems to be justified to say that the test takers did not consider short texts in the C-test as a particular context of use; they were merely employed like a vehicle for eliciting evidence of general mental processes involved in lexical access. This fact can be linked to their weak literacy skills, such as reading. It is in line with the findings of Stalnaker and Kurath (1935, cited in Read, 2000) who compared 2 methods of testing knowledge of German vocabulary at the elementary level.

The second research question aimed to explore the relationship between proficiency level and performance; these 2 tests revealed that learners at different proficiency levels performed differently on the C-test and the matching test. In this study, higher-ability students had better reading skills than their lower-ability counterparts, which would explain the results and would render C-tests usable only from certain levels of proficiency and up. A host of reasons can support this finding.

First, the production involves a higher level of knowledge than reception knowledge, because it involves the capacity not only to make use of but also to recognize a word when it is met in a new context or in isolation (Backman & Palmer, 1997; cited in Cameron, 2002, p. 63). Research has shown that the Vocabulary Level Test requires little grammatical knowledge or reading ability, and it is effective with elementary and intermediate level learners who lack sufficient knowledge of reading skill. It is probable that literacy skill interacts with the measure of vocabulary knowledge.

Additionally, the logic behind constructing the C-test is that half of a word is given in each mutilation in the C-test, and the examinees who do not understand the macro-context could mobilize their vocabulary skills to fill the gap without indulging in higher-level processing. It can be safely argued that C-tests tap into the macro-level and micro-level strategies, because the same items in the C-test that seemed difficult for the test-takers were answered correctly in the matching test. This fact shows that they did not resort to the micro-level strategy in the C-test, guessing the meaning of the word without the use of context. The test-takers were involved meaningfully with the text to find clues to answer the items, in spite of the fact that their reading ability was relatively weak. Therefore, more research with higher proficiency learners is vital to understand the role of macro-strategy in answering the items in vocabulary testing.

Finally, the researcher realized during the test-administration that the test-takers were unfamiliar with the format of the C-test. The appearance of the mutilated text was initially off-putting for some test-takers and resembles nothing that appears in the real world of the language learner. Therefore, further research is needed to adjust C-test difficulty to the test-takers' level of language proficiency by changing the format factors.

## REFERENCES

Babaii, E., & Ansari, H. (2001). The C-test: A valid operationalization of reduced redundancy principle. *System, 2*9, 209-219.

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

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice.* Oxford, UK: Oxford University Press.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing Journal, 17*(1), 1-42.

Batia, L., Elder, C., Hill, K, & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing, 21*(2), 202-220.

Brown, J. D. (1985). Tailored cloze: Improved with classical item analysis technique. *Language Testing, 4,*(1), 19-31.

Brown, J. D. (1996). *Testing in language programs.* Upper Saddle River, NJ: Prentice Hall Regents.

Cameron, L. (2002). Measuring vocabulary size in English as an additional language. *Language Testing Research, 6*(2), 145-173.

Chapelle, C. A., & Abraham, R. G. (1990). Cloze method: What difference does it make? *Language Testing, 7*(2), 121-146.

Connelly, M. (1997). Using C-test in English with postgraduate students. *Pergomon, 16*(2), 139-150.

De La Funente, M. J. (2002). *Acquisition of L2 vocabulary: The role of input and output in the receptive and productive acquisition of words.* Cambridge, UK: Cambridge University Press.

Eckes, Th., & Grotjahn, R. (2006). A closer look at the construct validity of C-Tests. *Language Testing, 23*(3), 291-319.

Farhadi, H., & Keramati, M. N. (1996). A text-driven method for the deletion procedure in cloze passages. *Language Testing, 13*(1), 191- 207.

Hatch, E., & Farhady, H. (1981). *Research design and statistics for applied linguistics.* Tehran, Iran: Rahnama Publication.

Hedge, T. (2000). *Teaching and learning in the language classroom.* Oxford, UK: Oxford University Press.

Hughes, A. (2003). *Testing for language teachers.* Cambridge, UK: Cambridge University Press.

Jafarpour, A. (1999). Can the C-test be improved with classical item analysis? *System, 27*, 79-89.

Klein-Braley, Ch., & Raatz, U. (1984). A survey of research on the C-test. *Language Testing, 2*(1), 135-146.

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

Khoii, R., Fotovat Ahmadi, P., & Shokouhian, M. (2007). *Reading for ideas* (Vol. 1). Tehran, Iran: Rahnama Publication.

Khoii, R., Fotovat Ahmadi, P., & Shokouhian, M. (2007). *Reading for ideas* (Vol. 2). Tehran, Iran: Rahnama Publication.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge; size, strength, and computer adaptiveness. *Language Learning, 54*(3), 399-436.

Laufer, B., C., Hill, K., & Congdon, P. (2002). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing, 21*, 202-226.

Meara, P. (2000). The rediscovery of vocabulary. *Second Language Research, 18*(4), 393-407.

Meara, P., & Buxton, B. (1983). An alternative to multiple choice vocabulary tests. *Language Testing, 2*(3), 143-154.

Meara, P., & Fitzpatrick, T. (2000). Lex 30: An improved method of assessing productive vocabulary in an L2. *System, 28*(3), 19-30.

Meara, P., & Nation, I. S. P. (2002). Vocabulary. In N. Schmitt (Ed), *An introduction to applied linguistics* (pp. 35-54). London, UK: Arnold Press.

Nation, I. S. P. (1982). Beginning to learn foreign vocabulary: A review of the research. *RELC Journal, 13*, 14-22.

Nation, I. S. P. (2001). *A study of the most frequent Learning vocabulary in another language.* Cambridge, UK: Cambridge University Press.

Oxford, R. L. (1990). *Language learning strategies: What every teacher should know.* Boston, MA: Heinle & Heinle.

Perkin, K., & Lineville, S. (1984). A construct definition study of a standardized vocabulary test. *Language Testing, 1*(2), 125-141.

Quian, D. D., & Schedle, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing, 21*(1), 28-52.

Read, J. (1988). Measuring the vocabulary knowledge of second language learners. *RELC Journal, 19*(2), 12-25.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing Research, 10*, 355-371.

Read, J. (2000). *Assessing vocabulary.* Cambridge, UK: Cambridge University Press.

Read, J., & Chapelle, C. A. (2001). A framework for second language measure for assessing reading performance. *Language Testing, 18*(1), 1-32.

Sasaki, M. (2000). Effects of cultural schemata on students' test taking processes for cloze tests: A multiple data source approach. *Language Testing, 17*(1), 85-114.

Schmitt, N. (1999). The relationship between TOEFL vocabulary items and meaning, association, collocation and word-class knowledge. *Language Testing, 16*(2), 189-216.

Schmitt, N. (2000). *Vocabulary in language teaching.* Cambridge, UK: Cambridge University Press.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behavior of two new versions of the vocabulary level test. *Language Testing, 18*(1) 55–88.

Shitotsu, T., & Weir, C. J. (2007). The relative significance of vocabulary breath in the prediction of reading comprehension test performance. *Language Testing, 12,* 24-99.

**APPENDIX 1: THE C-TEST**

Complete the mutilated words with appropriate letters.

**Passage 1**

There was once a wise man, a sage, who set out a journey that would take him through many lands. He had been travelling for some times and passing many hur----   when he was suddenly taken ill. Unable to continue on his way, the old man found shelter with a widow and his son. The widow was desperately poor but nursed the sage with great devo----, and he recov---- his strength in a short while. Before taking his leave, the old man glanced at the delap---- hut and then said," I've heard, when you wors----- god. I know your fears. "You see there is a little pro---- to provide food for your son. From now on there is a food ration of wheat and oil for you which will be handed out every month to your house".

**Passage 2**

The connection between the amount of cholesterol in the body and heart disease has not as yet, been definitely proved. Much research is being conducted to dete-----whether this fat- hormone-like subs-----  that flo--- around in the blood str--- is a factor causing some forms of heart trouble and serious ailm---. All that can be said is that cholesterol can be made from food products already stored in the body.

**Passage 3**

The best amateur athletes in the world match endure in a series of contests called the Olympic Games. The goal of the Olympic Game is to safe-----  the slo--- of a "sound mind in a sound body", and there are in fact some passio--sportsman and woman who have ventured beyond the pain barr--- in pursuit of their dream. But more remarkable are those champions who over---- injury, tragedy, and oppression to asc---   to the top. Their presence in the competition is a source of inspiration to everyone who has encountered them.

**Passage 4**

In Nepal environmentalists are delighted at the prospect of the return of cloth and paper bags in place of plastic. The use of plastic carrier bags is prohi----- due to the huge quantities of disca---- plastics that are littered the cities, clogging drains and polluting culti----- lands. Not only has it released toxic fumes in to the atmosphere, but also posing a health haz--- to the public and, more so, to workers in plastic units.

**Passage 5**

It may be comforting for many parents to know that it is an inhe---- behavior of human to express his strongest feelings toward his children. However, parents are amazed how something triv--- and mundane can trigger off their intense rage. Sometimes the anger occurs over more serious issues such as lying, and stea---- money from their pockets but most of the time, it is a resp---- to their helplessness in exerting their autho----  . It is necessary for adults not to turn the house into a battleground and ongo---- war. They need to manage their anger in order to settle peace.

**Passage 6**

The worst thing that can happen to couples is not interfering in-laws but conflicting lifestyles, which is a major cause of str--- in many marriages.
Living with a partner who has opposing interests and opinions is very hard. Some couples find it difficult to adjust  to their partner's habits and try to attr----- their disagreements to other issues while others, are enthusi------ to adopt

themselves to marital situation to maintain normal relationship. This group accepts to live with each other's shortcomings.

## APPENDIX 2: MATCHING TEST

Write the letter of the words in the second column next to their synonyms in the first column. There are two extra words.

| A | B |
|---|---|
| Strife | insignificant |
| enthusiastic | innate |
| disagreement | keen |
| inherent | destruction |
| Trivial | continuing |
| authority | conflict |
| ongoing | power |
| discarded | something that is disposed of |
| cultivated | necessary |
| response | reflection |
| | discord |

| A | B |
|---|---|
| hazard | excitable |
| hurdles | motto |
| dilapidated | flood |
| prospect | ruined |
| substance | encouraging |
| ailment | barrier |
| passionate | material |
| slogan | harm |
| devotion | loyalty |
| stream | expected |
| | sickness |
| | reality |

| A | B |
|---|---|
| attribute | rise |
| maintain | rescue |
| worship | rob |
| recover | drift |
| steal | ban |
| float | think of as caused by |
| safeguard | praise |
| ascend | struggle |
| overcome | welcome |
| prohibit | defend |
|  | support |
|  | defeat |

## BIODATA

**Mahmoud Abdi Tabari** is a Ph.D. student in Education at Oklahoma State University, USA, and a member of the Young Researchers' Club. His main research interests are second language acquisition, second language writing, and vocabulary assessment.