TELL, Vol. 2, No. 8, 2008

The effects of changing the deletion direction and deletion ratio on the validity and reliability of the C-test

Rahman Sahragard
Assistant Professor, Shiraz University
Jalal Rahimian
Associate Professor, Shiraz University
Reza Rahmani Anaraki,
MA, Shiraz University

Abstract

This study looks at the effects of changing the deletion direction and ratio on the validity and reliability of the C-test. For this purpose, a cluster sampling procedure was used to select 324 English students with different levels of proficiency. The subjects were then randomly assigned to three experimental groups. Each group took a TOEFL test and a C-test version. The three C-test versions were all based on the same texts. Immediately after completing the C-test, subjects received a ten-item questionnaire about issues such as face validity, suitability, and possible functions of the C-test. The findings indicated that changing the deletion ratio from 2 to 3 can result in an easier C-test with a higher discrimination power and a better criterion-related validity. On the other hand, changing the deletion direction (from right to left) can result in a more difficult C-test with a lower discrimination power and a worse concurrent validity –a C-test which might be a completely different test that measures a different construct. However, the changes to the deletion technique probably do not result in any changes to the reliability.

Key words: C-test, deletion direction, deletion ratio, validity, reliability

1. Introduction

The C-test is a fairly established member of integrative tests. Klein-Braley and Raatz (1984) developed the C-test as an attempt to remedy the shortcomings of the cloze test. A C-test consists of four or five short texts that are taken from authentic sources. The first sentence is left intact in each text. Beginning from word two of sentence two, the second half of every other word is deleted: This is called the ‘rule of two’ or the ‘C-principle’. In the canonical C-test each text will have either 20 or 25 blanks. So the whole C-test will have at least 100 gaps (i.e., missing parts).

As Eckes and Grotjahn (2006) suggest, C-tests are measures of general language proficiency rather than measures of specific skills (e.g. reading). Also, C-tests have the ability to differentiate between subjects of a lower and a higher proficiency level (Ikeguchi, 1998). Therefore, the C-test offers many advantages in the field of language testing over tests of the same caliber.

According to its proponents, the C-test is a short, easy-to-use, and qualified test of language proficiency. But this test has not yet become very popular with language test users. This is perhaps partly because of the lack of evidence in support of it. Therefore this study will try to shed more light on the issue.

One problem of the C-test is with its deletion procedure. The procedure in a normal C-test is that the right hand side of the words are deleted and only every second word is deleted. The founders of the C-test do not provide any clear reason why they mutilate every other words or why they delete the right part of the words and not the left part. The question is what happens if we change the deletion direction from the right to the left of the word. Or what happens if we mutilate every third word in the text instead of every second word. If these changes are used in the C-test (i.e. if the *rule of two* is violated), it is not clear whether the new test measures the same ability as the canonical C-test measures.

Except for a few studies related to this subject, no other study to date has tackled the object of the present study. Furthermore, even those few studies that have dealt with the issue (e.g. Cleary, 1988; Jafarpur, 1995, 1999; Farhady & Jamali, 2006) have had limitations that could substantially distort the results. For example, Cleary (1988) constructed a C-test by deleting the first half of the words and called it 'left-hand' C-test. He compared this left-hand C-test with a natural C-test and found that the left-hand C-test was more reliable, had a higher discrimination index and a higher correlation with an achievement test. He decided that left-hand deletions improve the testing qualities of the C-test. However, in his left-hand C-test, Cleary only deleted grammatically unmarked words on their left; other words were deleted on the right. Moreover, his C-tests were both based on a single text written by the researcher. Accordingly, McBeath (1989) indicated that Cleary had used a single 'purpose-written' text, instead of five authentic texts. Cleary (1988, p. 28) himself had noted that his passage was written "so as to sample the course of study that the subjects had followed at SOAF [Sultanate of Oman's Air Force] language schools" and that the resulted passage "was undoubtedly artificial and not altogether coherent above paragraph level". McBeath questioned Cleary's results and believed that his "adaptation of the C-test method" violated the authenticity requirement of the C-test. The founders of the C-test maintain that four to six authentic texts should be used for C-tests (Klein-Braley & Raatz, 1984; Raatz & Klein-Braley, 1995; Klein-Braley, 1997). Therefore, one might wonder whether Cleary's deviations from the *C-principle* have distorted his results.

Similarly, Heidari (1999) compared a left-hand C-test with a natural C-test. The results showed that the left hand C-test had more items with item discrimination (ID) values beyond .40. So, he concluded that left-hand deletions enhance the discrimination power of the C-test. However, this finding was only based on the number of items with ID values beyond .40. But he did not report any investigations of how efficiently his C-tests could differentiate between subjects with different levels of proficiency.

On the other hand, Jafarpur (1995) investigated the effect of changing the deletion ratio on the C-test. He developed twenty C-test versions of the same text by using different deletion starts and deletion ratios. He found that changing the deletion start and/or deletion ratio produces different C-tests, which tap different abilities. However, the number of the subjects who took each version of his C-tests was few. Each native speaker group had between 9 to 11 subjects and each non-native speaker had between 15 to 18 subjects. Moreover, like Cleary's (1988) study, Jafarpur's C-tests were based on one single text. Also, the numbers of mutilated words in Jafarpur's C-tests were between 14 to 45 items, which violate the minimum number of 100 items that is recommended by Raatz and Klein-Braley (1995). Therefore, Jafarpur's results can only be considered suggestive and not compelling.

In another study, Jafarpur (1999) accepted the shortcomings of his previous study (1995) and constructed two C-tests with a deletion ratio of 2 with 100 mutilations, and three C-tests with a deletion ratio of 3 with 77 mutilations. His results indicated no significant changes with regard to the reliability and discrimination power. However, his ratio-three C-tests violated the minimum number of 100 items that is recommended by Raatz and Klein-Braley (1995). Nonetheless, Jafarpur's (1999) results contradicted his previous findings.

Farhady and Jamali (2006) investigated the effect of using various deletion procedures on the C-test. They constructed 10 C-tests with different deletion ratios and different deletion direction (left or right). They used correlation and factor analysis. The results indicated that their different C-test versions measured different undelying abilities. However, only two of their C-tests had 100 mutilated words. Other C-tests had between 34 to 68 mutilated words. It is clear that a test with only 34 items cannot be considered a C-test. This limitation suggests that their results cannot be considered definite.

As it is observed, in all the studies referred to in this section, there are some shortcomings and deviations that could have influenced and distorted the results obtained. Therefore, the present study tries to investigate the issue of deletion technique of the C-test while controlling the other factors that are irrelevant to the matter in question.

However, as for the deletion frequency of the C-test, different ratios were possible (e.g. 3, 4, 5, 6, etc.). It was decided to use a ratio of 3 for this study because the results of previous studies showed some advantage for a ratio of 3 compared to other ratios (see for example, Jafarpur, 1995, 1999).

2. Method

2.1 Participants

The participants who took the tests were EFL learners with different proficiency levels. These participants were mostly females (89 percent were females) and were aged between 16 to 53. However, most participants were in their twenties. The total number of participants was 324.

Because it was difficult to use a purely random procedure to choose the subjects, it was decided to use a cluster sampling procedure. In all the educational centers there were several different classes. It was decided to randomly choose between three to five classes in proportion to the total number of classes in each center. This resulted in a total of 20 classes. Then, the students in these classes were randomly assigned to three experimental groups. Each group was given a C-test version: a standard C-test, a C-test with left-hand deletions, and a C-test with a deletion ratio of three (i.e. instead of deleting every other word, every third word is deleted).

2.2 Instruments

Four tests were used in this study, namely, a TOEFL test and three C-tests. The TOEFL test is a tailored version of the original TOEFL. This test includes 60 multiple-choice items and is divided into two subtests: Structure and Written Expression (S-WE) which includes 30 items, Reading Comprehension, and Vocabulary (RC-V) which includes another 30 items. According to Rahimi (2004), this test has a criterion-related validity of 0.76, a test re-test reliability of 0.92 and a KR-21 reliability of 0.85.

In this study, the TOEFL test was used to measure the proficiency level of the subjects. Table 1 presents the reliability estimates that were found in this study for the TOEFL and its subtests. These reliability estimates have been calculated by Kuder-Richardson Formula 21 (KR-21) formula.

Table 1: Reliability Indices for TOEFL and Its Subtests

	TOEFL	S-WE	RC-V
(N = 324) KR-21 estimate	.85	.77	.80

The following procedures were followed to construct the C-tests. First, ten passages were chosen from different English textbooks and sources. Since the books were written by native speakers, the selected texts were regarded as authentic. These texts were about different subjects. Flesch Readability Ease value was measured for each text. The readability values for the ten texts were in the range between 14.0 (very difficult) and 83.5 (very easy), which indicates different readability levels of the passages. Dörnyei and Katona (1992) recommend the use of texts with various difficulties for the purpose of better measurement accuracy.

For pretesting, a standard C-test with ten texts was constructed by the researchers, using the *rule of two*. Each text had 20 mutilations (total of 200 mutilations). Then these ten passages were pretested with 25 English students in Payam-Noor University. The C-tests were scored with the exact word method and item analysis was performed on the results. The passages with item discrimination indices of .30 or higher and item facility indices between .20 and .80 were considered acceptable. According to Raatz and Klein-Braley (1995), this range is quite reasonable. Five texts with the highest item discrimination and best item facility were chosen. The other passages were discarded.

The five selected texts varied in difficulty with Flesch Reading Ease values of 73.3, 83.5, 60.6, 60.5, and 38.4 respectively. With these five passages, three different versions of C-test were constructed: Canonical C-test, Left-hand C-test, and Ratio-three C-test. For constructing the Canonical C-test, the *rule of two* was applied. For the Left-hand C-test, again, the same steps were used but the deletions were on the left side of each word. And for the Ratio-three C-test, steps like the Canonical C-test were used but the deletion ratio was three. Therefore, the second half of every third word was deleted on the right side of the word. Each text in each C-test had twenty blanks. The texts in each C-test were arranged in the order of difficulty from easy to difficult. Each of the three versions had 100 items as recommended by Raatz and Klein-Braley (1995). The instructions were given in Persian with a short English C-test example and the answer.

2.3 Procedures

In all the classes, first the researcher explained about the tests in Persian in order to make sure that all the subjects clearly understand the points. Then, for each group, the tests were administered in a counterbalanced way: in some classes, first the C-tests were given and in other classes, first the TOEFL and then the C-test were administered. The subjects were told that they would be informed of their grades, and the top examinees would receive a prize. In addition, their lecturers were asked to tell them that their high scores on the test would influence their final term grades but their low scores would not. This strategy was followed to make them motivated and take the tests seriously. Except for one lecturer, all other lecturers agreed to this and therefore they followed the procedure.

The different versions of the C-test were scored with the exact word method with two options: one with spelling errors as incorrect and one without spelling errors. The first one was called the *Exact* method and the second one the *Exact Lenient* method.

To score the C-tests with acceptable word method, as native speakers were not available, some English teachers and lecturers with high proficiency were asked about their ideas about some items. Based on their ideas and the researcher's own experience in scoring the C-test papers, all the C-test papers were scored again with two acceptable scoring methods: one with spelling errors as incorrect and one without spelling errors. The first method was called the *Acceptable* method and the second one the *Acceptable Lenient* method.

3. Analysis

3.1 Item analysis

Item facility and item discrimination indexes were calculated for each C-test text. When these indexes were calculated, the item analysis was performed on the texts used in the C-tests (C-text). For this reason, each C-text was considered a 'super-item', as suggested by Klein-Braley and Raatz (1984), Raatz, and Klein-Braley (1995). Then, Bachman's (2004, p. 127-128) formulas were used for 'partial credit' items, because these super-items (C-texts) are scored with 20 points.

3.2 Validity and Reliability

To investigate the concurrent validity of the C-tests, the scores of the subjects on the C-tests were correlated with their TOEFL scores using Pearson product-moment correlation formula. Reliability coefficients for all the C-tests were estimated using Kuder-Richardson Formula 21 (KR-21) and Cronbach's alpha formula. Both these formulas are measures of internal consistency.

Bachman (1990) and Farhady (1983) have argued that internal consistency reliability coefficients are not suitable for cloze tests and C-tests because the test items in these measures are not independent. However, Raatz and Klein-Braley (1995) and Raatz (1985) suggest that it is possible to perform an inner consistency analysis on C-tests. They state that it is not acceptable to define the individual blanks in the C-test as items, since they are dependent on each other because of text structure and content. But they suggest the solution of considering each C-test text as a super-item and analyzing with four or five super-item. "With such scaled items Cronbach's Alpha formula should be used" (Raatz & Klein-Braley, 1995). Therefore, Cronbach's alpha and KR-21 formulas were used in this study to calculate the reliability of all the C-tests.

3.3 Discrimination and classification power

The discrimination power of the C-tests was investigated by three analyses of variance, three Scheffe post hoc tests, and a decision consistency analysis. For this reason, the subjects in each experimental group were first classified into three proficiency groups based on their TOEFL scores. The top 27% of the subjects were placed in one group called "High group", the bottom 27% of the subjects were placed in another group called "Low group", and the rest of the subjects were place in the group called "Middle group". Those subjects in the Middle group who had the same TOEFL scores as the subjects in the High group were omitted from the Middle group. In the same way, those subjects in the Middle group who had the same TOEFL scores as the subjects in the Low group were omitted from the Middle group.

Then, for each C-test version, an analysis of variance (ANOVA) was performed on the means of the three proficiency groups on their C-test scores. This ANOVA was performed to see whether the three C-test versions were able to discriminate between different levels of language ability.

Another analysis, which was performed on the scores of the three proficiency groups in each C-test version, was a decision consistency analysis. Decision consistency, according to Jafarpur (2002, p. 42-43), refers to “the percent classification of subjects by the experimental tests that correspond correctly to those by the criterion”. For this study, decision consistency here is what percentages of examinees are placed in their correct proficiency level if a C-test (instead of the TOEFL) is used as the criterion of placement.

It was decided that the average percentage of correct placements predicted by each C-test could be considered as an index of the classification power of that C-test and an index of the suitability of that C-test for placement purposes.

3.4 Analysis of different scoring methods

Four different scoring methods for the C-test were compared. For this reason, the following statistics were calculated: descriptive statistics, the correlation coefficients between four scoring methods on all C-tests, and the correlation coefficients between all the C-tests and the TOEFL test. In the end, an analysis of variance (ANOVA) was performed for the differences among means of four scoring methods on all C-tests.

4. Results and discussion

4.1 Descriptive & Inferential statistics

A descriptive statistical analysis was run on the data. For space purposes this is not shown in the article, though. This analysis showed that the largest range of scores belongs to the Ratio-three C-test (81 score points) followed by the Left-hand C-test (80) and the lowest range belongs to the Canonical C-test (71). This suggests that the Ratio-three C-test produced a better dispersion among the subjects' scores.

The Ratio-three C-test also produced the highest mean (52.2) followed by the Canonical C-test (48.18). The Left-hand C-test produced the lowest mean (39.2) of all. Tentatively, this can be indicative of the relative difficulty of the Left-hand C-test compared to the other two C-test forms.

Regarding the C-tests, the analysis indicates that the Canonical C-test is a little positively skewed (.21) and the Ratio-three C-test is a little negatively skewed (-.10). However, in both of these C-tests, the value of the skewness is not very large and does not make the distribution asymmetrical.

The Left-hand C-test is also positively skewed (1.2) and when this skewness statistic is divided by its standard error, it can be found that the value ratio 5.22 (1.2/.23) does not fall between -2 and +2. This means that the distribution is asymmetrical. The relatively larger value of skewness in the Left-hand C-test shows that this test has been more difficult for most of the test takers. This indicates that Left-hand deletion can produce C-tests that are more difficult than the standard deletion method (i.e. the *rule of two*) which is right-hand.

However, one cannot be confident of this result unless one is sure that these three experimental groups are at the same level of proficiency. In order to be confident of the equality of the three experimental groups, an analysis of variance (ANOVA) was performed on their scores on the TOEFL test.

Table 2 reports the mean scores of the subjects in each group on the TOEFL test.

Table 2: Means of the Three Experimental Groups on the TOEFL

Group	Mean on the TOEFL
Canonical C-test group	29.10
Left-hand C-test group	28.22
Ratio-three C-test group	29.22

The results of the ANOVA test is presented in Table 3 below. As the table indicates, *F*-ratio is not significant at the level of p<.05. This proves that the means of the three experimental groups are not significantly different. Therefore, it can be concluded that the three groups of participants who took the three C-test types had a similar proficiency level. This proves that the random sampling procedures were effective and the three samples are representations of the same population. This would also mean that the three experimental groups are homogeneous with regard to their language ability.

Table 3: ANOVA Results for the Differences among Means of the Three Experimental Groups on the TOEFL

	Source of Variance	Sum of Squares	*df*	Mean Square	*F*	Sig.
TOEFL	Between Groups	64.416	2	32.208	.351	.705
	Within Groups	29485.173	321	91.854		

4.1.1 Comparison of the means of the three C-tests

In order to understand whether there is any significant difference between the means of the experimental groups on their respective C-tests, another ANOVA was performed. Table 4 shows the results of this ANOVA. The figures in the table show that the obtained *F* ratio is highly significant. This indicates that the means of the participants belonging to each of these three experimental groups are significantly different from each other.

However, it should be mentioned that the significance of the *F* ratio in an analysis of variance only indicates that there is a significant difference among the means of the compared groups as a whole, that is, it indicates that there is at least *one* significant difference between the means of at least *one* pair of the groups compared (Brown, 1988; Delavar, 2002; Bachman, 2004). Nevertheless, it cannot be said where exactly this difference is, i.e., exactly which two means are different. In order to determine exactly which means are different the *multiple comparisons* is needed to perform, which are also called *post hoc* or *follow-up* tests (Hatch & Farhady, 1982; Delavar, 2002; Bachman, 2004). The only requirement for these tests is that the overall *F* in the ANOVA is statistically significant (Hatch & Farhady, 1982)

Table 4: ANOVA Results for the Differences among Means of the Three Experimental Groups on Their Respective C-tests

Independent Variable	Source of Variance	Sum of Squares	*df*	Mean Square	*F*	Sig.
C-test Version	Between Groups	9530.196	2	4765.098	17.799	.000
	Within Groups	85936.727	321	267.716		

The follow-up test that was used in this study is a Scheffe's test. Scheffe's test is a commonly used multiple comparison test which reveals the precise location of differences by analyzing every two means separately (Hatch & Farhady, 1982; Brown, 1988; Delavar, 2002).

The results of a Scheffe test performed on the means of the three experimental groups on their respective C-tests showed that the mean performance of the subjects on the Canonical and the Ratio-three C-test is so similar that the mean difference is not statistically significant. However, the Left-hand C-test shows significant mean difference with both the Canonical C-test and the Ratio-three C-test. Notice that the ability level of the subjects was found to be similar. Therefore, the subjects' poorer performance on the Left-hand C-test implies that Left-hand C-test is a completely different test from the other two C-test versions.

The descriptive statistics revealed that the Left-hand C-test had the largest positive value of skewness and the lowest mean among the three C-test versions. The ANOVA results found here confirm our previous tentative judgments about the relative difficulty of the Left-hand C-test.

4.1.2 The results of item analysis

As Brown (1996, p. 50) has noted, “sometimes …, item analysis is performed simply to investigate how well the items on a test are working with a particular group of students”. Therefore, an item analysis was applied to determine the difficulty and discrimination indices of the items, and thereby, identify the items that function well and the malfunctioning items.

The result of item analysis showed that the Ratio-three C-test has the highest average IF value (.52) which means the C-texts on this C-test version were in general easier than the C-texts of the other two C-test versions. The Canonical C-test had an average IF value (.49). And finally the Left-hand C-test had the lowest average IF value (.42) which means that the C-texts of the Left-hand C-test were in general more difficult than the other two C-test versions.

And for item discrimination, all the three C-tests have average item discrimination indexes in the range of .39 to .40.

According to the acceptability index proposed by Ebel(1979), it can be concluded that the average ID index for all of the three C-test types are acceptable as the figures above fall between the acceptable range. However, the C-text 2 in the Ratio-three C-test has an ID index of .27, which is considered a marginal item and needs modification according to Ebel. Though, Jafarpur (1997) believes that ID indexes as low as .20 is acceptable. All other C-texts have ID indexes that are above .30 and therefore, according to both Ebel and Jafarpur, their ID indexes are reasonably good.

4.1.3 Reliability

Table 5 tabulates reliability coefficients of all the C-tests. The reliability of all these measures was computed by the Kuder-Richardson Formula 21 (KR-21) and also by the Cronbach’s alpha formula, as recommended by Raatz and Klein-Braley (1995).

Table 5: Reliability Indices for All C-test Versions

Test	KR-21	Cronbach's alpha
(N = 109) Canonical C-test:	.91	.84
(N = 108) Left-hand C-test:	.93	.84
(N = 107) Ratio-three C-test:	.91	.86

As Table 5 indicates, scores from all the C-tests show very high KR-21 reliability coefficients in the range between .91 and .93. The coefficient alphas for the C-tests are a little lower and in the range of .84 to .86. With KR-21 method, the Left-hand C-test is the most reliable (.93), and with the Cronbach's alpha method, the Ratio-three C-test is the most reliable (.86).

The reliability indexes of the C-tests in this study, i.e. those estimated by the Cronbach's alpha formula, are very similar to the coefficient alpha that Klein-Braley (1997) found for her C-test (.85). The high reliability indexes for the C-tests in this study support the other research findings about the reliability indexes of the C-test (e.g. Klein-Braley & Raatz, 1984; Dörnyei & Katona, 1992; Kamimoto, 1993; Mochizuki, 1994; Babaii & Ansary, 2001; Jafarpur, 2002; Sigott, 2004; Rahimi & Saadat, 2006).

4.1.4 Criterion-related validity

Table 6 provides Pearson product-moment correlations (r_{xy}) among the scores from the C-tests and the TOEFL. The correlation corrected for attenuation (r_{CA}) is also presented in this table. According to Mousavi (1999, p. 67), correction for attenuation may be used "to determine what the correlation would be between two tests if both were perfectly reliable". For more information about correction for attenuation, see Bachman (2004).

Table 6 demonstrates that the Ratio-three C-test has the highest correlation with the TOEFL test (.73). The Canonical C-test follows it with a slightly lower correlation coefficient (.72) with the TOEFL. Finally, the Left-hand C-test has the lowest correlation with the TOEFL scores (.58).

Table 6: Correlation Coefficients between All C-tests and the TOEFL

Test	TOEFL	
	r_{xy}	r_{CA}
Canonical C-test	.61	.72
Left-hand C-test	.49	.58
Ratio-three C-test	.63	.73

All correlations are significant at $p<.01$ level (2-tailed).

The fact that the Ratio-three and the Canonical C-tests show reasonably high correlations (.73 and .72) with the TOEFL test provides good concurrent validity for these two C-tests. On the other hand, the low correlation of the Left-hand C-test and the TOEFL (.58) indicates that the Left-hand C-test is not a good test for measuring general language proficiency.

4. 2 Analysis of variance

Table 7 shows the ANOVA results for the test of differences among the means of the three proficiency groups on all the three

C-test versions. The obtained *F* ratios are all significant at p<.000 level, which suggests that there is a significant difference among the means. This is clear evidence that all the three C-tests have been able to differentiate among the subjects who are at different levels of proficiency in English.

Table 7: ANOVA Results for the Differences among Means of Three Proficiency Groups on All C-tests

Test	Source of Variance	Sum of Squares	*df*	Mean Square	*F*	Sig.
Canonical C-test	Between Groups	8626.164	2	4313.082	25.607	.000
	Within Groups	17348.780	103	168.435		
Left-hand C-test	Between Groups	7662.733	2	3831.366	16.339	.000
	Within Groups	23683.229	101	234.487		
Ratio-three C-test	Between Groups	9349.992	2	4674.996	25.787	.000
	Within Groups	18672.951	103	181.291		

Three followed-up (post hoc) tests were used to determine exactly which means differ from each other on each C-test version. Tables 3.10 through 3.12 indicate the results of three Scheffe's test performed on the means of the three proficiency groups on each C-test.

The result showed that there is significant difference between the means of every combination of two proficiency groups on the Canonical C-test. This means that the difference between test performances of the subjects in each group is large enough to be statistically meaningful. Therefore, it can be concluded that the Canonical C-test could discriminate between the subjects in the three proficiency groups successfully.

The results of another Scheffe's test performed on the means of the three proficiency groups on the Left-hand C-test showed that there is again significant difference between the means of all possible combinations of two proficiency groups on the Left-hand C-test. This means that the difference between test performances of the subjects in each group on the Left-hand C-test is large enough to be statistically meaningful. Therefore, it can be concluded that the Left-hand C-test could discriminate between the subjects in the three proficiency groups successfully.

The last Scheffe's test performed on the means of the three proficiency groups on the Ratio-three C-test showed that there is significant difference between the means of all possible combinations of two proficiency groups on the Ratio-three C-test. This means that the difference between test performances of the subjects in each group on the Ratio-three C-test is large enough to be statistically meaningful. Therefore, the Ratio-three C-test could discriminate between the subjects in all the three groups successfully.

The outcomes from two other multiple comparison tests were the same. These two tests were a Bonferroni test and a Tukey's HSD (honestly significant difference) test (cf. Delavar, 2002; Bachman, 2004; SPSS, 1989-2003).

However, one interesting fact was found from these three Scheffe tests. That is the Left-hand C-test has difficulty discriminating between the subjects in the Low and Middle groups. This result cannot be obtained with the other two C-test forms. One explanation for the weaker discrimination power of the Left-hand C-test is that its distribution is very peaked and very positively skewed.

4.2.1 Decision consistency

The scores from each C-test version were also studied for decision consistency. Table 8 shows the correct classifications that are made if the C-test was used as the placement criterion in each experimental group. As can be observed, the Canonical C-test can, on the average, correctly place only 51.9 percent of the subjects in appropriate proficiency groups. This means that the canonical C-test could not place about 48 percent of the subjects in their appropriate proficiency levels. This is not a good sign for a test if it cannot classify almost half of the examinees in their proper levels.

Table 8: Percent of Correct Classification Predicted by Each C-test Version

Criterion for Placement	Low	Middle	High	Average
Canonical C-test	60%	45.7%	50%	51.9%
Left-hand C-test	60%	56.8%	56.7%	57.8%
Ratio-three C-test	55.2%	56.3%	58.6%	56.7%

The other two C-test versions are a little better in this regard. The Ratio-three C-test can, on the average, correctly place 56.7 percent of the subjects in their appropriate proficiency levels. And the Left-hand C-test can correctly place 57.8 percent of the subjects in appropriate proficiency levels.

4.3 Scoring methods

As mentioned before in the method above, four scoring methods were used for scoring the C-test papers. Descriptive statistics showed the means of the subjects on each C-test change only very slightly with each scoring procedure. An analysis of variance was performed to see if these slight changes are statistically significant. Table 9 shows the results of the ANOVA for the differences among means of four scoring procedures on all C-test versions. As it is observed from the table, none of the *F* ratios obtained are significant at a desired level (e.g. $p < 0.5$). This indicates that the means of the subjects on all C-test types do not differ significantly with different scoring methods.

Table 9: ANOVA Results for the Differences among Means of Four Scoring Methods

Test	Source of Variance	Sum of Squares	*df*	Mean Square	*F*	Sig.
Canonical C-test	Between Groups	267.752	3	89.251	.361	.781
	Within Groups	106669.835	432	246.921		
Left-hand C-test	Between Groups	71.824	3	23.941	.078	.972
	Within Groups	130792.611	428	305.590		
Ratio-three C-test	Between Groups	409.495	3	136.498	.504	.680
	Within Groups	114822.804	424	270.808		

To further investigate the results of using different scoring methods with the C-test, two correlation analyses were performed. The results of these two analyses are reported in Tables 10 and 11.

Table 10: Correlation Coefficients between Four Scoring Methods on All C-tests

Test	Scoring Procedure	Exact r_{xy}	Exact Lenient r_{xy}	Acceptable r_{xy}	Acceptable Lenient r_{xy}
Canonical C-test	Exact	-	-	-	-
	Exact	.997	-	-	-
	Lenient	.998	.995	-	-
	Acceptable	.996	.998	.997	-
	Acceptable				
	Lenient				
Left-hand C-test	Exact	-	-	-	-
	Exact	.998	-	-	-
	Lenient	1.00	.997	-	-
	Acceptable	.998	1.00	.998	-
	Acceptable				
	Lenient				
Ratio-three C-test	Exact	-	-	-	-
	Exact	.997	-	-	-
	Lenient	.998	.993	-	-
	Acceptable	.996	.998	.997	-
	Acceptable				
	Lenient				

All correlations are significant at *p*<.01 level (2-tailed).

Table 10 above, shows that the four scoring methods used with each of the C-test versions have a nearly perfect correlation with each other. And all correlations are highly significant.

Table 11, on the other hand, shows that the four scoring procedure have almost the same correlation with the criterion measure. These results show that the use of any of these methods for scoring C-tests provides more or less the same results.

Table 11: Correlation Coefficients between All Tests with the Four Scoring Methods

Test	Scoring Procedure	TOEFL r_{xy}
Canonical C-test	Exact	.61
	Exact Lenient	.61
	Acceptable	.61
	Acceptable Lenient	.61
Left-hand C-test	Exact	.49
	Exact Lenient	.49
	Acceptable	.49
	Acceptable Lenient	.49
Ratio-three C-test	Exact	.63
	Exact Lenient	.62
	Acceptable	.63
	Acceptable Lenient	.62

All correlations are significant at *p*<.01 level (2-tailed).

However, a very interesting fact was observed from the investigation of the scoring method; i.e. the number of the individual scores that changed when using the Acceptable Lenient method of scoring compared with the Exact scoring. Table 12 shows the number and percentage of individual scores, which changed when using the Acceptable Lenient method with each C-test version. As can be observed, in the Canonical and Ratio-three C-tests, the majority of scores have changed when using the Acceptable Lenient method. However, in the Left-hand C-test, only 18 percent of the scores have changed.

A reanalysis of the C-test papers showed that the majority of the changes in the Canonical and Ratio-three C-tests were related to the inflectional errors. It means, the subjects had guessed the base form of the word but could not provide the correct inflectional morpheme or misspelled the inflectional morpheme. Note that the Canonical and Ratio-three C-tests use right-hand deletion method. But the Left-hand C-test uses a left-hand deletion method and therefore it does not delete the part of the words that carry inflectional morphemes. Since in the English language, all inflectional morphemes are suffixes (Yule, 2006), the inflectional errors were absent in the Left-hand C-test. This explains the lower percentage of score change in the Left-hand C-test.

Table 12: Number and Percentage of Individual Scores, Which Changed When Using the Acceptable Lenient Method

C-Tests	Freq.	%
Canonical (N = 109)	75	69
Left-Hand (N = 108)	19	18
Ratio-three (N = 107)	84	79
Total (N = 324)	178	55

These findings accord with Cleary’s (1988) results that the majority of his subjects’ errors on a right-hand deletion C-test were related to the inflectional morphemes while his left-hand deletion C-test had no errors related to the inflectional morphemes.

5. Conclusions

The results of this study are summarized and enumerated here:The Ratio-three C-test produced the best dispersion, was the easiest C-test with the highest mean score and the highest mean item facility, was the most discriminating test with the highest mean item discrimination, and produced the highest correlation with the criterion measure; the Left-hand C-test was the most difficult C-test with the lowest mean score and the lowest mean item facility, its distribution was asymmetrical, and produced the lowest correlation with the criterion; left-hand deletion did not improve the discrimination power of the C-test (as judged by the mean item discrimination value); the Left-hand C-test was a completely different test from the other two C-test versions and did not appear to tap the same construct; the Cronbach's alpha and KR-21 reliability of all three C-tests were high; the mean scores of different proficiency groups on all the C-tests increased progressively from lower levels to higher levels. Statistical significance of these means proved that all the three C-tests were able to discriminate among the subjects with different levels of language ability; all the three C-test versions could differentiate among the subjects at different levels of proficiency at p<.05 level of significance. However, The Left-hand C-test could not discriminate between the subjects at the lower end of the proficiency spectrum at p<.01 level of significance. This indicates a lower discrimination power for the Left-hand C-test; all the C-test versions failed to classify almost half of the examinees in their proper levels. The Canonical C-test was the most problematical version in this regard, indicating that the changes to the deletion procedure of the C-test had improved the measurement accuracy of the C-test; different scoring procedures for the C-test produced more or less the same results. This indicates all versions are not not sensitive to spelling errors; the majority of the subjects' errors on right-hand C-tests were related to inflectional morphemes; and the C-test as a whole was found to possess good face validity.

Finally, as a way of summarizing the main results, we conclude that the modification of the C-principle can certainly affect the qualities of the C-test. Changing the deletion ratio from 2 to 3 can result in an easier C-test with a higher discrimination power and a better criterion-related validity. On the other hand, changing the deletion direction from the right side of the words to the left side can result in a more difficult C-test with a lower discrimination power and a worse concurrent validity –a C-test which might be a completely different test that measures a different construct. However, the changes to the deletion technique probably do not produce any changes to the reliability.

References

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

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

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. London: Longman.

Brown, J. D. (1988). *Understanding research in second language learning: a teacher's guide to statistics and research design*. Cambridge: Cambridge University Press.

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

Cleary, C. (1988). The C-Test in English: left-hand deletions. *RELC Journal*, *19*(2), 26-35.

Connelly, M. (1997). Using C-Tests in English with post-graduate students. *English for Specific Purposes, 16*(2), 139-150.

Delavar, A. (2003). *Probability and applied statistics in education* (in Persian). Tehran: Roshd Publications.

Dörnyei, Z. & Katona, L. (1992). Validation of the C-test amongst Hungarian EFL learners. *Language Testing, 9*, 187-206.

Ebel, R. L. (1979). *Essentials of educational measurement* (3rd Ed.). Englewood Cliffs, NJ: Prentice-Hall.

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

Farhady, H. (1983). New directions for ESL proficiency testing. In J. W. Oller, Jr, (Ed.), *Issues in language testing research* (pp. 253-269). Rowley, MA: Newbury House.

Farhady, H. & Jamali, F. (2006). Varieties of C-test as measures of general language proficiency. In H. Farhady (Ed.), *Twenty-five years of living with applied linguistics: collection of articles* (pp. 287-302). Tehran: Rahnama.

Farhady, H., Jafarpur, A., & Birjandi, P. (1994). *Testing language skills: from theory to practice*. Tehran: SAMT.

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

Heidari, A. (1999). *C-test: Solving strategies, changing the position of deleted letters and discriminatory power*. Unpublished master thesis. Shiraz University, Iran.

Hughes, A. (2003). *Testing for language teachers* (2nd Ed.). Cambridge: Cambridge University Press.

Huhta, A. (1996). Validating an EFL C-test for students of English philology. In R. Grotjahn (Ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen [The C-Test: theoretical foundations and practical applications]* (Vol. 3, pp. 197-234). Bochum: Brockmeyer.

Ikeguchi, C. B. (1998). Do different C-tests discriminate proficiency levels of EL2 learners? *JALT Testing & Evaluation SIG Newsletter 2*(2), 3-8. Retrieved March 24, 2007, from http://www.geocities.com/ College Park/ Field/1087/test/ike_1.htm

Jafarpur, A. (1995). Is C-testing superior to Cloze? *Language Testing*, *12*(2), 194-216.

Jafarpur, A. (1999). What's magical about the rule-of-two for constructing C-Tests? *RELC Journal*, *30*(2), 86-100.

Jafarpur, A. (2002). A comparative study of a C-Test and a cloze test. In R. Grotjahn (Ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen [The C-Test: theoretical foundations and practical applications]* (Volume 4, pp. 31-51). Bochum: AKS-Verlag.

Kamimoto, T. (1992). An inquiry into what a C-Test measures. *Fukuoka Women's Junior College Studies*, *44*, 67-79.

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

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

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

McBeath, N. (1989). C-Tests in English: Pushed beyond the original concept? *RELC Journal, 20*(2), 36-41.

McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press.

Mochizuki, A. (1994). C-tests: four kinds of texts, their reliability and validity. *JALT Journal*, *16*(1), 41-54.

Mousavi, S. A. (1999). *A dictionary of language testing* (2nd Ed.). Tehran: Rahnama.

Raatz, U. (1985). Better theory for better tests? *Language Testing*, *2*(1), 60-75.

Raatz, U., & Klein-Braley, C. (1995). *Introduction to language testing and C-tests*. Retrieved December, 2006, from http://www.uniduisburg.de/fb3/angling/forschung/howtodo.htm

Rahimi, M. (2004). *An investigation into the factors affecting the Iranian EFL students' perceived use of language learning strategies.* Unpublished PhD Thesis. Shiraz University, Iran.

Rahimi, M., & Saadat, M. (2006). The application of item analysis to a C-test. *Language Forum, 23*(1-2), 207-218.

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

SPSS (1989-2003). *SPSS for windows, release 12.0.0*. SPSS Inc.

Weir, C. J. (1990). *Communicative language testing*. Hemel Hempstead: Prentice Hall.

Yule, G. (2006). *The study of language* (3rd Ed.). Cambridge: Cambridge University Press.