Validation of the C-test amongst Hungarian EFL learners

Zoltán Dörnyei and Lucy Katona *Eötvös University, Budapest*

This article reports on the results of a research programme carried out to validate the C-test amongst Hungarian EFL learners. One hundred and two university English majors were administered four different language tests (including an oral interview) to form a General Language Proficiency measure against which the C-test was evaluated. Various analyses were made, partly to replicate the results of the earlier studies to see to what extent these could be generalized, and partly to shed light on controversial issues. The same C-test was then administered to four groups of secondary school pupils (N=53) to examine whether the findings amongst university students were also true in another proficiency range. The results of the programme confirmed that the C-test is a reliable and valid instrument, and detailed information was obtained about issues such as text difficulty and text appropriateness, the role of content and structure words, and the use of different scoring methods.

I Introduction

The popularity of the cloze test has proved that language examiners and testers are very much in need of a written measuring instrument which is easy to design and administer, and which assesses overall language proficiency reliably. This popularity is all the more significant as cloze tests are known to be rather difficult and problematic to score objectively unless 'exact word scoring' is followed, which is not easy to justify and makes the test very difficult. There have also been more theoretical criticisms of cloze testing, concerning the representatives of the deleted words, the fact that the choice of a particular text, deletion rate, starting point for deletion, and scoring method considerably affects the score, and that even native speakers can rarely obtain a maximum score, which is rather surprising about a test which is supposed to measure language abilities (for an overview see Alderson, 1979; Klein-Braley and Raatz, 1984).

In 1981, Christine Klein-Braley and Ulrich Raatz introduced a new deletion technique, *'the rule of 2'*, which was believed to remedy most of the shortcomings of the classical cloze procedure. According to *'the rule of 2'*, the second half of every second word should be deleted in a text, starting and ending with an intact sentence. This resulted in what the two researchers termed the *'C-test'* (for an overview see Klein-Braley and Raatz, 1984; Raatz and Klein-Braley, 1985; Raatz, 1985; Klein-Braley, 1985).

The C-test is easy both to design and to score and several different texts can be used to make up a complete test, which is shorter and contains more deletions than a cloze test. Native speakers can score up to 100%, and Raatz (1985) presents evidence that the truncated words tend to be representative of the texts they occur in. What is more, students also find C-tests less frustrating than cloze tests. Almost too good to be true, and yet the empirical results all attest to the fact that the C-test is a reliable and valid instrument, doing 'everything that the cloze test promised' (Klein-Braley and Raatz, 1984: 145).

One shadow on the positive features of the C-test has been cast by the fact that no one seems to be quite sure what it actually measures, but this ambiguity exists with regard to the cloze test as well and, after all, in the history of psychometrics there is a surprisingly large number of tests which are known to work reliably whereas nobody knows exactly why and how (see, for example, aptitude tests in general, including language aptitude batteries).

The great potential of C-testing led us to launch a research programme at Eötvös University, Budapest aiming at validating the test amongst Hungarian EFL learners. We intended to achieve two things: 1) to replicate the very positive results reported in the literature to see to what extent they could be generalized, and 2) to look into some uncharted or controversial areas in C-testing. We were especially interested in the following issues:

a) the concurrent validity of the C-test and the aspects of language proficiency it measures;
b) a comparison of the cloze with a C-test;
c) the relationship between text difficulty, reliability and measurement accuracy;
d) the role of content and structure words;
e) measurement accuracy in samples of different homogeneity and proficiency level;
f) the effect of different scoring methods.

In order to investigate these issues, we needed a very reliable and valid criterion measure to serve as a reference point and we took every care to produce one. Through the co-operation of 24 colleagues and graduate students, we formed a composite criterion score from the results of four different language tests/test batteries: our department's regular test battery (three subtests), a standardized American test battery (two subtests), an oral interview and a cloze test. After we had completed the main study with university students, we re-administered the C-test to secondary school pupils to check whether our results still held true for a different sample with a lower proficiency level.

II Method

1 Subjects

In the first (and main) study the subjects were 102 first-year English majors at the Department of English, Eötvös University, Budapest. They represented more than two-thirds of the whole first year. As they had all had to take an entrance exam to be accepted at the Department, they formed a rather homogeneous sample with an average language proficiency of above the intermediate level. The follow-up study involved four groups of secondary pupils (N=53) in two different schools; their proficiency level was well under that of the first sample.

2 The language tests comprising the criterion measure

Four different language tests were administered to all the subjects in the first study to form a composite criterion measure; for the secondary school sample only the second battery in the list below was administered:

Department Proficiency Test: which is developed every year for the purpose of testing first year students in order to screen out those who are not proficient enough to enter the second year. This battery takes about 90 minutes to complete and consists of three parts:

a) vocabulary test (20 multiple-choice items);
b) grammar test - sentence-transformation tasks, similar to the ones in the Cambridge First Certificate test (20 items);
c) listening comprehension test (two dialogues, 20 multiple-choice and true-or-false items).

Test of English for International Communication (TOEIC): which is a standardized multiple-choice test offered by the Educational Testing Service, Princeton as a nonacademic counterpart of TOEFL; it is considered to be a 'highly reliable and valid measure of English' (Perkins, 1987: 82) and it has been validated for Hungarian EFL learners (Dornyei, 1990). TOEIC consists of two sections, each containing 100 items, and students have two and a half hours to complete it. These sections are:

a) listening comprehension;
b) reading, which contains varied items focusing also on language accuracy.

Oral interview: conducted by two examiners, one of whom is always a native speaker of English, lasting about 10–15 minutes. The students' achievement was rated according to accuracy, vocabulary and fluency, on five-point scales where each point could be modified by '+' and '–' marks, and these are then summarized to form a composite score.

Cloze test: (39 gaps with a deletion rate of five - see Appendix).

3 *The development of the C-test*

The development of the C-test was carried out as part of a specialization course for graduate students. Here we had the same experience as Grotjahn (1987), who pointed out that it could prove difficult to find texts which could discriminate amongst advanced learners, especially if we are aiming at the ideal 50% solution rate. We considered more than 15 texts, and finally four were included in the final C-test, containing 24, 17, 21 and 19 gaps respectively (see Appendix). The internal consistency reliability of the four extracts regarded as superitems, assessed by means of Cronbach a was .75 in the university student sample and .77 in the secondary pupil sample.

4 *Procedures*

The Department Proficiency Test, TOEIC and the Oral Interview were administered centrally as part of the university students' official end-of-term exams. The cloze and the C-test were administered six weeks later, at the beginning of the following term, during the students' language practice classes. In the secondary school sample, the C-test and TOEIC were administered in the pupils' regular English language classes. The data were processed using the Statistical Package for the Social Sciences (SPSS). The criterion measure was created by factor analysing the four different language tests and taking the regression-method factor score generated by SPSS from the one-factor solution (see below for more details). The various data procession analyses (described below in detail) involved correlation and factor analyses, and *t*-test procedures for various subgroups.

III Results and discussion

1 Concurrent validity

Using a minimum-eigenvalue criterion of 1.0, the principal component factor analysis of the four different language tests yielded a one-factor solution accounting for over 50% of the variance (Table 1, first column). This was indeed expected, as the subjects in the survey were all preselected university English majors whose different language skills were fairly evenly developed and thus could be sufficiently represented by a unitary factor. As can be seen in the table, the various language measures loaded on this factor evenly, thus it was termed *General Language Proficiency.*

Table 2 (first two columns) presents the correlations of General Language Proficiency and the administered language tests with the C-test in the university and secondary school samples. Two composite scores, Department Proficiency Test Total and TOEIC Total, were computed by adding up the standardized component scores of the batteries in question.

As can be seen in Table 2, the C-test has highly sufficient positive correlations in all the composite language proficiency measures (General Language Proficiency, Department Proficiency Test Total and TOEIC Total), and has significant positive correlations with all the language tests. This indicates that the C-test is a highly integrative language test which measures global language proficiency. The only area in our study where the C-test appeared to be less efficient is in the testing of grammar; this can be explained by the fact that deletions in the C-test are not performed at sentence level but at word level. Little and Singleton had similar findings:

> There is some evidence that in filling C-test slots our subjects tended to give priority to a ready lexical solution over morpho-syntactic and more general semantic issues. Intuitively this is in line with the priorities that we exercise in spontaneous communication (Little and Singleton, 1990: 14).

Table 1 Factor analyses of the language tests administered (Analysis 1: all the language tests except the C-test have been entered; Analysis 2: all the language tests except the C-test and the cloze test have been entered)

Type of test	Analysis 1 Factor 1	Analysis 2 Factor 1
Department Proficiency Test		
vocabulary section	.72	.75
grammar section	.67	.70
listening section	.63	.63
TOEIC Listening Section	.80	.80
Reading Section	.73	.73
Oral Interview	.75	.75
Cloze Test	.66	—

Table 2 Correlations of General Language Proficiency and the administered language tests with the C-test in both samples, and with the cloze test in the university sample

Type of test	C-test		Cloze test
	Univ. sample (N=102)	Sec. sample (N=53)	Univ. sample (N=102)
General Language Proficiency	**.57*****	—	—
General Language Proficiency **(B)**[1]	**.56*****	—	**.53*****
Department Proficiency Test			
Vocabulary Section	**.38*****	—	.33***
Grammar Section	.25*	—	.26**
Listening Section	**.33*****	—	.38***
Total[2]	**.43*****	—	.43***
TOEIC Listening Section	**.51*****	**.53*****	.45***
Reading Section	**.54*****	**.58*****	.44***
Total[2]	**.62*****	**.62*****	.52***
Oral Interview	**.43*****	—	.46***
Cloze Test	**.38*****	—	—

Notes: [1] General Language Proficiency (B) has been computed in the same way as General Language Proficiency except that the cloze scores have not been entered in the factor analysis (see Table **1,** Column 2).

[2] **The** Total scores were computed by adding the standardized component scores.

******p*<**.05; *****p*<**.01; ******p*<**.001**

On the other hand, as confirmed by the high correlations with the more general proficiency measures, the C-test does call the testee's expectancy grammar into action in a similar way to the cloze test. Indeed, based on the highly significant correlations with the two most general measures - General Language Proficiency and TOEIC Total (.57 and .62 respectively) - we may conclude that the C-test is a very efficient **integrative** language test. Furthermore (as will be described later) after piloting the C-test and considering only the most efficient extracts (1, 2 and 3), these correlations became even higher (.61 and .65 - see Table 5).

2 *Comparing the results of the cloze and the C-test*

Table 2 (third column) presents correlations between the cloze test and the other language measures. In this case, however, the composite measure termed General Language Proficiency could not be used as it actually contained the cloze results and therefore the correlation would have been disproportionately high. In order to be able to compare the cloze and the C-test we computed a second general language proficiency measure, termed General Proficiency Language (B), in the same way as we computed General Language Proficiency, but this time the cloze results were not entered into the factor analysis (see Table 1, Column 2). The correlations in Table 2

Table 3 Mean scores, standard deviations and the difficulty rates of the C-test and its four extracts in the two **samples**[1]

Extracts	University sample			Secondary school sample		
	$\bar{x}$	*SD*	*Diff. rate*	$\bar{x}$	*SD*	*Diff. rate*
C-test	58.5	8.4	.72	29.9	7.8	.37
Extract 1	19.2	2.2	.80	13.4	3.4	.56
Extract 2	11.9	2.9	.70	2.6	1.9	.15
Extract 3	14.0	3.4	.67	8.2	2.8	.39
Extract 4	13.3	2.6	.70	5.8	2.0	.31

Note: [1] **As** could be expected all the intersample differences in the mean scores and the difficulty rates are highly significant.

suggest that the C-test provides as good estimates of language proficiency as the cloze, **if** not better. In fact, the differences in the correlations with **TOEIC** imply superior discriminatory power on the part of the C-test. (This will be confirmed later when analysing results obtained in more homogeneous **subgroups.)**

3 *Text difficulty, reliability and measurement accuracy*

Klein-Braley and Raatz state that '**it** has been repeatedly shown that even tests which are far too difficult or far too easy for the target groups will still produce acceptable reliability and validity coefficients' **(1984: 140).** Our results support this claim. Table 3 presents the mean difficulty rates of the C-test in the two samples (as obtained by dividing the mean scores by the number of **gaps).** These show that the test proved to be fairly easy for the university sample and rather difficult for the secondary school sample. However, as has been mentioned before, the internal consistency coefficients found in the two samples were very consistent, .75 and .77 respectively.

Let us consider now the difficulty of the extracts. Klein-Braley (1985) reports attempts to define objective predictor measures for text difficulty: The two measures which emerged as the best predictors were the type-token ratio (number of different **words/** total number of words) and average sentence length. In a review of these findings, Carroll **states:** '**it** is not clear, however, that these estimates are really generalizable over varied samples of examinees and texts; one wonders whether the extensive investigation that might be required would be worth the **effort' (1987: 102).** Our findings support Carroll's doubts. Table 4 presents ***t*-test** statistics testing the significance of the differences between the extract difficulties in the two **samples.** For the university **students,** Extract 1 proved to be significantly easier than the other three **texts,** Extract 3

Table 4 Paired samples f-test statistics comparing the difficulty rates of the four extracts

Extract	$\bar{x}$		*SD*		***t-value***		*Probability*	
pairs	UnS	SecS	UnS	SecS	UnS	SecS	UnS	SecS
Extract 1	.80	.56	.09	.14	6.61	24.74	.000	.000
Extract 2	.70	.15	.17	.11				
Extract 1	.80	.56	.09	.14	10.33	8.82	.000	.000
Extract 3	.67	.39	.16	.14				
Extract 1	.80	.56	.09	.14	7.58	12.07	.000	.000
Extract 4	•70	.31	.14	.11				
Extract 2	.70	.15	.17	.11	1.93	-13.88	.057	.000
Extract 3	.67	.39	.16	.14				
Extract 2	.70	.15	.17	.11	.01	-9.91	.992	.000
Extract 4	.70	.31	.14	.11				
Extract 3	.67	.39	.16	.14	-1.98	4.59	.051	.000
Extract 4	.70	.31	.14	.11				

Notes: **UnS = University student sample; SecS = Secondary pupil sample**

was the most difficult, and Extracts 2 and 4 were equally difficult. This is, however, not the case for the secondary school sample: there, Extract 2 proved to be by far the most difficult and Extract 3, which the university students found most difficult, turned out here to be the second easiest. This seems to be at odds with Klein-Braley's conclusion, namely that 'the group of texts used in any C-test remains more or less constant in terms of relative difficulty' (1985: 88).

Another and perhaps more important question is whether the measuring capacity/discrimination index of the texts remains constant in different samples. Table 5 (first column) presents the correlations of the scores on the extracts and their various combinations with General Language Proficiency, as obtained in the university student sample. In the secondary school sample only the two TOEIC subtests were administered besides the C-test and therefore their combination (TOEIC Total) will be treated as the language criterion measure. The fifth column of Table 5 presents the correlations between TOEIC Total and the various extract-combinations; in order to be able to compare the results obtained in the two samples, the fourth column of Table 5 contains the same extract-TOEIC Total correlations for the university sample.

As can be seen in Table 5, there is a very consistent pattern of extract-criterion measure correlations across the two samples and the two criterion measures. Of the four extracts, Extract 1 appears to measure language proficiency best, followed by Extract 2 and Extract 3; adding Extracts 2 and 3 to Extract 1 increases the accuracy of the measurement. Extract 4, however, seems to be a

Table 5 Correlations of the four extracts and their various combinations a) with General Language Proficiency in the whole university sample and in the high and low ability subsamples, and b) with TOEIC Total (the sum of the standardized scores of the two TOEIC subtests) in both samples

Extracts	General Language Proficiency (University student sample)			TOEIC Total	
	Whole sample (N=102)	Top half (N=51)	Bottom half (N=51)	Univ. sample (N=102)	Secondary sample (N=53)
Extract 1	.58***	.36**	.53***	.61***	.62***
Extract 2	.46***	.22	.32*	.54***	.52***
Extract 3	.46***	.49***	.20	.46***	.44**
Extract 4	.25*	.02	.16	.27**	.24
Extracts 1 + 2	.59***	.33*	.49***	.66***	.65***
Extracts 1 + 3	.57***	.50***	.39**	.58***	.62**
Extracts 2 + 3	.55***	.46***	.33*	.60***	.55**
Extracts 3 + 4	.46***	.37**	.23	.47***	.43**
Extracts 1 +4	.48***	.18	.45***	.51***	.59**
Extracts 2 + 4	.42***	.12	.30*	.48***	.44**
Extracts 1 + 2 + 3	.61***	.47***	.44***	.65***	.64**
Extracts 1 + 3 + 4	.54***	.39**	.38**	.56***	.59**
Extracts 1 + 2 + 4	.53***	.22	.44***	.59***	.61**
Extracts 2 + 3 + 4	.52***	.35*	.32*	.56***	.51***
Extracts 1 + 2 + 3 + 4 (complete test)	.57***	.38**	.42**	.62***	.62***

Notes: *p<.05; **p<.01; ***p<.001

bad predictor of language proficiency, and its inclusion actually reduces the measurement accuracy of the instrument (thus our C-test would be more efficient with only the first three extracts). These results point to the fact that even though the relative difficulty of an extract varies according to the proficiency level of the testees, the measuring ability of an extract is fairly stable. This implies that once an extract has proved to be efficient with one sample, it is likely to work well with different samples too; the results also confirm Klein-Braley and Raatz's (1984) claim that even if a C-test is too easy or difficult for a particular target group, it will produce acceptable validity.

How can differences in measuring capacity be explained? Table 6 summarizes the main parameters of each of the four extracts: the mean sentence length (both for the extract as a whole and for the mutilated sentences only), the type-token ratio, the difficulty rate (for the two samples), and finally the rank order of measuring capacity (for the two samples) as defined by the extract-criterion measure correlations. We have already seen that the difficulty rate is not directly related to the measuring capacity and, as the table shows, neither is the type-token ratio. The two sentence length measures, however, produce a pattern which is very consistent with the measuring capacity: Extract 1, which measures language pro-

Table6 Summary of different parameters of the four extracts

Extract	Mean sentence length		Type-token ratio	Difficulty rate		Rank order of meas. capacity	
	Whole extracts	Mutilated sentences		Univ. sample	Secondary sample	Univ. sample	Secondary sample
Extract 1	9.8	10.0	.79	.80	.56	1	1
Extract *2*	11.8	17.5	.73	.70	.15	*2-3*	*2*
Extract 3	13.3	11.3	.78	.67	.39	*2-3*	*3*
Extract 4	18.3	19.5	.73	.70	.31	4	4

ficiency best, contains the shortest sentences, whereas Extract 4, the least efficient of the four texts, contains the longest sentences. The two interim texts do not show a clear pattern with respect to the two sentence length measures: even though Extract 3 has longer sentences as a whole, the actual mutilated sentences are shorter than those in Extract 2; this may be the reason why the two texts turned out to be more or less equal measures of language proficiency in the university student sample.

Thus we may conclude that in our two samples mean sentence length was the best predictor of the efficiency of a text as a measure of language proficiency: the shorter the sentences were, the better the text measured general language proficiency. **Klein-Braley** (1985) gives a detailed overview of studies which concluded that average sentence length was a good index of syntactical complexity and also of the text's cognitive loading. This is certainly true about our extracts. In trying to find a fourth, that is, most **difficult**, text for our **C-test**, we selected a highly complex, abstract passage which was difficult to understand even without any mutilations - and, as it turned out, the text did not work very well. As for the other **extreme**, the first and easiest text was a straightforward narrative in the simple past - and this simplicity worked beautifully. All this points to the fact that texts which have a relatively **simpie** structure both in terms of syntax and content complexity are best suited to being used for C-testing. This does not mean, however, that they cannot be difficult, but the difficulty of a text should be a function of the language and especially of the vocabulary load the text **contains**. This issue will be further investigated below with regard to the structure and content words in the extracts.

4 *Content words and structure words*

It has been shown repeatedly that truncated structure words are significantly easier to restore in a C-test than truncated content words (Klein-Braley, 1985). This was true in our investigation too, as demonstrated by *t*-**test** statistics ($t = 23.35$, $p < .001$). The difference in word type difficulty was assumed to have important bearings on the measuring capacity of the C-test and the individual **extracts**, and therefore we analysed to what extent the two word types relate to the accuracy of measurement in the two samples; separate analyses were also computed for the high and low ability halves of the university student sample based on the General Language Proficiency measure. Table 7 contains the correlations of the scores on the truncated structure and content words (broken down by the extracts) with the language criterion measures and the total C-test score.

The correlations presented in the table help to shed light on why the extracts behaved as they did (i.e., which were their weak parts and which proficiency range in particular was affected by these), but the most important feature of the pattern is that content and structure words behave strikingly differently in the two samples. Amongst the university students, for whom the C-test was fairly easy, the content words were a better measure of language proficiency than structure words. In fact, the latter actually reduced the measuring capacity of the test (i.e., if we ignore structure words, we get higher correlations with General Language Proficiency). On the other hand, amongst the secondary school pupils, for whom the C-test was too difficult, we find the opposite tendency: the structure words separately give a better estimate of the testees' language proficiency than the whole test, and here it is the content words which reduce the measurement accuracy of the test. In the low ability half of the university sample, which is somewhere in between the university and the secondary school samples in terms of average proficiency level, the increasing importance of the structure words can already be noticed, especially in Extract 2. The last four columns of Table 7 present correlations between the whole C-test and its parts. Here again we find that in the secondary school sample the structure words contribute more to the total C-test score than the content words.

We believe that the tendency described above is a typical example of the remarkably versatile character of the C-test: we could say that it has 'something to offer to everybody', whether university language majors or average secondary school learners. The special deletion technique in C-test appears to allow for many different approaches to achieving item solutions at various testee proficiency levels, tapping on a wide range of language areas and cognitive processing skills. A good example of this was provided by Cohen, Segal and Bar-Siman-Tov, who found that 'because half the word was given, students who did not understand the macro-context could still mobilize their vocabulary skills adequately to fill in the appropriate discourse connector without indulging in higher-level processing' (1984: 225). In a study using thinking-aloud and retrospective methods, Feldmann and Stemmer (1987) also found various problem-solving student behaviours and succeeded in defining several different bottom-up and top-down processing strategies which are called into action by subjects doing C-tests. This range of possible strategies to be applied explains why the test works in samples of various proficiency levels, and this is also the reason

Table 7 Correlations of scores on the truncated structure and content words (broken down by the four extracts) with the language criterion measures and the total C-test scores as obtained in the two samples

	Gen. Lang. Proficiency (University sample)				C-test			
	Whole sample (N=102)	Top half (N=51)	Bottom half (N=51)	TOEIC Total (sec. sch. sample) (N=53)	UnWh	UnTop	UnBot	SecS
Whole C-test								
Content Words (N=44)	.61***	.47**	.42**	.44**	.97***	.96***	.97***	.88***
Structure words (N=37)	.34***	.11	.30*	.67***	.82***	.83***	.78***	.91***
Extract 1								
Content Words (N=10)	.62***	.40**	.48***	.25	.72***	.64***	.66***	.64***
Structure words (N=14)	.34***	.17	.49***	.66***	.56***	.62***	.48***	.69***
Extract 2								
Content Words (N=11)	.48***	.26	.40**	.27*	.77***	.74***	.74***	.52***
Structure words (N=6)	.23*	-.05	.05	.53***	.54***	.38**	.56***	.71***
Extract 3								
Content Words (N=12)	.53***	.56***	.22	.40**	.78***	.71***	.75***	.70***
Structure words (N=9)	.20*	.20	.11	.36**	.51***	.63***	.38**	.66***
Extract 4								
Content Words (N=11)	.27**	.04	.21	.30*	.67***	.64***	.66***	.62***
Structure words (N=8)	.01	-.18	-.04	.04	.31**	.40**	.18	.32*

Notes: *$p<.05$; **$p<.01$; ***$p<.001$; UnWh = Whole university sample; UnTop = High ability university subsample; UnBot = Low ability university subsample; SecS = Secondary school pupil sample.

why the test is more 'user-friendly' than, for example, the cloze test, which is known to be consistently difficult and frustrating (Scott and Madsen, 1983).

5 *Student proficiency range and sample homogeneity*

The proficiency level of the university student sample ranged from intermediate to advanced. To examine how well the C-test discriminated in both tails of the distribution, the sample was divided into a high and a low ability group, based on the General Language Proficiency measure. This also allowed us to see how the C-test worked in extremely homogeneous samples: by dividing the university students into two proficiency level groups we further reduced the within-group variation in a sample which was already preselected to start with.

The second and the third columns of Table 5 present the correlations of the four extracts and their various combinations with General Language Proficiency found in the two subsamples. As can be seen in the table, Extract 4 again does not work very well with either subsample. Extract 1, which is the easiest text, appears to be the most efficient with the lower level group, and Extract 3, the most difficult text, works best with the high proficiency group. The various combinations of the extracts produce a wide range of correlations, with considerable differences between the correlations in the two groups. The only combination which has correlations over .40 in both groups is the sum of Extracts 1, 2 and 3, which indicates that in order to cater for variations amongst language testees, and thus provide equal measurement accuracy in both tails of a sample distribution, we should select extracts of various difficulties for a C-test.

The correlations in the two subsamples are depressed by the reduced variance, but it is worth noting how much more efficiently the C-test measures language proficiency in the two subgroups than the cloze test, as shown in Table 8. In fact, the rather low reliability and validity of the classical cloze test in homogeneous samples was one criticism put forward by Klein-Braley and Raatz (1984). The C-test does indeed remedy this shortcoming to a certain extent.

6 *Scoring methods*

There are three obvious ways of scoring the C-test: 1) only the exact solutions are accepted; 2) every appropriate word is accepted; and 3) either of the methods above combined with spelling mistakes being or not being tolerated. A rationale for allowing spelling mistakes would be partly the wish to maintain face validity (students who fill a gap but commit a minor spelling mistake will feel that their attempt is worth more than that of others who leave the gap unfilled), and partly the fact that integrative tests are trying to measure the ability to cope with reduced language redundancy and to activate one's expectancy grammar, and spelling is only marginally related to these complex skills.

We believe that the first option, 'exact word scoring' is hard to justify theoretically and therefore we decided to apply the second method. However, we were soon forced to realize that the role of spelling mistakes was more significant than we had originally assumed. Table 9 contains the frequencies of the spelling errors in the two samples: Almost 60% of the university students did commit spelling mistakes and more than 22% made more than one mistake. The secondary school pupils committed fewer spelling errors, but even in this sample 32.1% of the students made at least one spelling mistake.

Table 10 presents the correlations between the language criterion measures (General Language Proficiency and TOEIC Total) and the C-test score with and without the spelling mistakes. As can be seen, whether or not we tolerate spelling mistakes does not significantly affect the discriminatory power of the C-test, although some minor improvement can be achieved by tolerating spelling mistakes (in the university student sample there was an improvement of .0053, which disappeared when the coefficients were rounded). It is

Table 8 Correlations of General Language Proficiency (B) with the C-test and the cloze test in the top and bottom proficiency halves of the university sample

	Top half (N=51)		Bottom half (N=51)	
	C-test	Cloze	C-test	Cloze
General Language Proficiency (B)	.44***	.25	.34*	.25

Notes: **p*<.05; ***p*<.01; ****p*<.001

Table 9 Frequencies of spelling mistakes in the two samples

No. of spelling mistakes made by one student	Frequency of students		Percent	
	Univ. sample	Sec. sample	Univ. sample	Sec. sample
0	41	36	40.2%	67.9%
1	38	14	37.3%	26.4%
2	18	1	17.6%	1.9%
3	3	2	2.9%	3.8%
4	1	—	0.9%	—
5	1	—	0.9%	—

Table 10 Correlation between the language criterion measures and the C-test score with and without the spelling **mistakes**

Language criterion measures	C-test			
	Spelling mistakes tolerated		Spelling mistakes marked wrong	
	Univ. **samp.**	Sec. **samp.**	Univ. **samp.**	Sec. **samp.**
General Language Proficiency	.57	—	.57	—
TOEIC Total	**—**	**.62**	**—**	**.61**

Note: All the coefficients are highly significant

therefore up to the examiner to decide which scoring method is more appropriate for the sample and purpose in question. In the analyses reported in this study we have marked every appropriate word with a minor spelling mistake correct.

IV Conclusion

The various analyses reported in this study produced the following results:

1) The C-test proved to be a reliable and valid instrument amongst Hungarian EFL learners. The high concurrent validity coefficients were especially noteworthy for two reasons: a) the C-test we used was previously untried, and by submitting it to statistical evaluation and selecting only the most efficient extracts, the coefficients could be further improved; b) the criterion measure we used was very carefully developed and involved several language proficiency measures including an oral measure.
2) Satisfactory reliability and validity coefficients were obtained even when the test was too easy or too difficult for the target group, which confirmed earlier results in other countries.
3) The C-test proved to be a highly integrative language testing instrument, assessing general language proficiency very efficiently. This again is consistent with previous findings.
4) The C-test appeared to be a somewhat better measure of general language proficiency than the cloze test; this superiority became especially featured in more homogeneous samples.
5) In our study, the relative difficulty of the individual extracts comprising the C-test varied according to the target group; the measuring ability/discrimination index of the extracts, however, appeared to be fairly stable across samples.
6) The best predictor of the measuring ability of a text in our study was mean sentence length; it was argued that tests which have a

relatively simple structure both in terms of syntax and content complexity are best suited to being used for C-testing.

7) Similarly to other studies, our results also suggested that a C-test should include extracts of different difficulties to provide an accurate measurement in both tails of the sample distribution.
8) Truncated structure words were confirmed to be easier to reconstruct than truncated content words.
9) We found that with samples of higher proficiency levels the measurement of language proficiency was primarily due to the role played by content words, and structure words actually reduced measurement accuracy. In lower level samples, however, structure words gained an increasing importance to a point when the role of the two word types was completely reversed.
10) The tolerance of spelling errors, which increases the C-test's face validity, did not reduce the concurrent validity of the test; in fact, if anything, a slight improvement in measurement accuracy was found when the spelling mistakes were not marked wrong.

In sum, the results we obtained, exceeded our original expectations. The C-test proved to be a highly integrative and versatile measuring instrument, working well in samples of various difficulty and homogeneity levels. We have argued that this is due to the special deletion technique, which samples a representative set of language items and allows for several different approaches to achieving item solutions at various testee proficiency levels, thus tapping on a wide range of language areas and cognitive skills.

With respect to more practical questions, it was difficult to believe that a test which is significantly easier to design and score than most other testing instruments, did as good a job - in some ways better - than the latter. A major objective of research on language testing is to increase the cost-effectiveness of the assessment; our conclusion about the C-test is that not only is it a reliable and valid measure of general language proficiency, but it is also one of the most efficient language testing instruments in terms of the ratio between resources invested and measurement accuracy obtained.

V References

Alderson, J.C. 1979: The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly* 13, 219–26.

Carroll, J.B. 1987: Review of ‘Klein-Braley, C. and Raatz, U., editors, *C-Tests in der Praxis. Fremdsprachen und Hochschule, AKS-Rundbrief 13/14,* Bochum: Arbeitskreis Sprachenzentrum [AKS]'. *Language Testing* 4, 99-106.

Cohen, A.D., Segal, M. and Bar-Siman-Tov, R. 1984: The C-Test in Hebrew. *Language Testing* 1, 221-25.

Dörnyei, Z. 1990: Local validation of TOEIC in Hungary. *The Reporter, (TOEIC News International)* 4, Princeton: Educational Testing Service, 4.

Feldmann, U. and Stemmer, B. 1987: Thin___ aloud a__ retrospective da__ in C-te__ taking: diffe languages - diff___learners - sa__ approaches? In Færch, C. and Kasper, G., editors, *Introspection in second language research,* Clevedon: Multilingual Matters, 251–67.

Grotjahn, R. 1987: How to construct and evaluate a C-Test: a discussion of some problems and some statistical analyses. In Grotjahn, R., Klein-Braley, C. and Stevenson, D.K., editors, *Taking their measure: the validity and validation of language tests,* Bochum: Brockmeyer, 219–53.

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

Klein-Braley, C. and Raatz, U. 1984: A survey of research on the C-Test. *Language Testing* 1, 134–46.

Little, D. and Singleton, D. 1990: The C-test as an elicitation instrument in second language research. Paper presented at AILA 1990, Thessaloniki, Greece, 16–22 April.

Perkins, K. 1987: Review of the 'Test of English for International Communication'. In Alderson, J.C., Krahnke, K.J. and Stansfield, C.W., editors, *Reviews of English language proficiency tests,* Washington, DC: TESOL, 81–83.

Raatz, U. 1985: Tests of reduced redundancy - The C-Test, a practical example. In Klein-Braley, C. and Raatz, U., editors, *C-Tests in der Praxis. Fremdsprachen und Hochschule, AKS-Rundbrief 13/14*, Bochum: Arbeitskreis Sprachenzentrum [AKS], 14–19.

Raatz, U. and Klein-Braley, C. 1985: How to develop a C-test. In Klein-Braley, C. and Raatz, U., editors, *C-Tests in der Praxis. Fremdsprachen und Hochschule, AKS-Rundbrief 13/14*, Bochum: Arbeitskreis Sprachenzentrum [AKS], 20–22.

Scott, M.L. and Madsen, H.S. 1983: The influence of retesting on test affect. In Oiler, J.W. Jr., editor, *Issues in language testing research,* Rowley, MA: Newbury House, 270–79.

VI Appendix

The cloze test used in the study

In olden days when a glimpse of stocking was looked upon as something far to shocking to distract the serious work of an office, secretaries were men. Then came the first . . . . . War and the male . . . . . were replaced by women. . . . . . man's secretary became his . . . . . servant, charged with remembering . . . . . wife's birthday and buying . . . . . presents; taking his suits . . . . . the dry-cleaners;

telling lies ..... the telephone to keep ..... he did not wish ..... speak to at bay; ....., of course, typing and ..... and taking shorthand. Now ..... this may be changing ...... The microchip and high ..... is sweeping the British ..... taking with it much ..... the routine clerical work ..... secretaries did. 'Once office ..... takes over generally, the ..... of the job will ..... again because it will ..... only the high-powered work-..... then men will want ..... do it again'. That ..... said by one of ..... male executives of one ..... the biggest secretarial agencies ..... this country. What he ..... predicted is already under ..... in the U.S. One ..... described to me a ..... temporary job placing men ..... secretarial jobs in San ....., she noted that all ..... men she dealt with ..... to be gay so ..... that is just a ..... twist to the old ...... Over here, though, there are men coming onto the job market as secretaries.

The C-test used in the study

Extract 1 One cool autumn evening, Bob L., a young professional, returned home from a trip to the supermarket to find his computer gone. Gone! all sorts of crazy thoughts raced through his mind: Had it been stolen? Had it been kidnapped? He searched his house for a clue until he noticed a small piece of printout paper stuck under a magnet on his refrigerator door. His heart sank as he read this simple message: CAN'T CONTINUE, FILE CLOSED, BYE.

Extract 2 There is a third factor besides farming and herding in the spread of man-made deserts: deforestation. The progressive destruction of the Third World's stock of trees is damaging not only in dry regions: everywhere it occurs it can accelerate the decay of the soil and reduce its capacity to feed people. It can reduce rainfall and lead to drought.

Extract 3 There are certain things which no student can do without and others which may not be as necessary as you thought. It may be worth[1] considering some small hints. You may find yourself in need of electrical appliances such as light bulbs, adaptors or plugs. These can be obtained from many places. GILL is a good hardware shop and trying to find it is a challenge. It is hidden in a little alley leading off High Street called Wheatsheaf Yard.

[1] Due to a copying fault the letter 'o' was not legible on some copies and therefore the restoration of this word was not included in the final score.

Extract 4 The private conscience of the leader - father than his public responsibilities - becomes the focal point of politics. Internal criteria - possession of, devotion to, and standing up for private principles - become the standards of political judgement. Constituents disappear, and we are left with a political leader determining policy on the basis of compatibility with his private principles. From this perspective we can better understand why Goldwater voted against the Civil Rights Act of 1964.