doi:10.5128/ERYa12.14

DEVELOPING A C-TEST TO MEASURE LANGUAGE ABILITY AS AN ALTERNATIVE TO A SKILLS-BASED TEST

Ingrid Sarapuu, Ene Alas

Abstract. The article investigates the properties of a c-test and its ability to measure test-takers' overall English language proficiency in the Estonian context. For this purpose, prior research concerning c-test validity and reliability is consulted, and the c-test's advantages as compared to a skills-based proficiency evaluation instrument are sought. The article then discusses the process of developing a c-test following the procedure recommended by Grotjahn (1987) and Raatz and Klein-Braley (2002), and piloting it among Estonian secondary school students who simultaneously took the skills-based national examination in the English language. Statistical analysis displays very strong correlations between the c-test results and those of the national examination, as well as with teacher evaluation of the test subjects' proficiency, substantiating the c-test's viability as an economical language ability measure in contexts where quick appraisal of the respective ability is required. The study reveals implications for language proficiency assessment practices as well as for the process of c-test development.

Keywords: language assessment, overall language ability, test development, validity, reliability, national examination

1. Introduction. Research questions

Nowadays, there are many free online c-test apps, also called c-test creators or generators, developed by different interested parties, that make the claim of creating tests that measure students' language ability with a high degree of precision and with a fraction of the cost spent on test development, administration and marking, compared to traditional skills-based tests that take months to develop and hours to complete and mark. C-test developers maintain that following a careful, informed test-development procedure, it is possible to develop an instrument that takes 20 to 30 minutes for students to complete and allows just as accurate language proficiency appraisal as is achieved by a 3-hour-and-55-minute-long – this is the time required to complete the current English language national examination in Estonia (Inglise keele riigieksami eristuskiri) – skills-based proficiency test (Eckes, Grotjahn 2006: 290). If true, such a test could prove very useful in situations where administering a full-fledged proficiency test would be problematic because of test development, time and administrative constraints, but where quick appraisal of candidates' language ability would still be necessary. Previous research in the European context (cf. Grotjahn 1987, Coleman 1996, Raatz 2002, Sigott 2004, Linnemann, Wilbert 2014) seems to suggest a high correlation between c-test and traditional proficiency test results, thus a question arises whether there is a similar link between the c-test and the Estonian national examination in the English language as well.

With the above in mind, a study was developed to investigate the differences in construct that the skills-based proficiency test and the c-test represent, what the advantages and problems of both test-types are, what constraints c-test development poses, and above all, to what extent c-test results correlate with those of the English language national examination.

2. Background to the study

There have been two basic approaches to defining language proficiency during the last 40 years. One, proposed by Oller (1979), maintains that there is one underlying language competence, which cannot be divided into separate proficiency components, i.e. it is unitary, and may be related to the general factor of intelligence. This idea seems to be supported in the work of Raatz and Klein-Braley (2002: 81) who say that "all language behaviour is related and thus integrative".

There is also a more widespread view that superseded Oller's hypothesis. Chomsky's distinction between competence and performance led Hymes to coin the term communicative competence, seeing competence as the most general term for the capabilities of a person which are dependent upon both knowledge and use (Hymes 1972: 282–283). Communicative competence is thus seen as the interaction of grammatical (what is formally possible), psycholinguistic (what is feasible in terms of human information processing), sociocultural (the social meaning or value of what is said), and probabilistic (what actually occurs) systems of competence (Canale, Swain 1980: 16). Canale and Swain suggested minimally three main competencies: grammatical (knowledge of lexical items and of rules of morphology, syntax, sentence-grammar semantics, and phonology), sociolinguistic (sociocultural rules of use and rules of discourse) and strategic (verbal and non-verbal communication strategies that help compensate for breakdowns in communication) (*ibid.* 27). This model was later extended by Bachman and Palmer, distinguishing three levels – organizational language knowledge (grammatical and textual knowledge), pragmatic language knowledge (functional and sociolinguistic knowledge), and strategic competence (metacognitive components and strategies) (1996: 66–68). Bachman prefers the term communicative competence to proficiency, considering it more inclusive than the latter, which had been used in the context of oral language testing. According to him, communicative language ability consists of language competence, strategic competence and psychophysiological mechanisms. Language competence, in turn, is divided into the categories of organisational competence, the subcategories of which are grammatical and textual competence, and pragmatic competence consisting of illocutionary and sociolinguistic competence (1990: 107–108). As can be seen from the above, language competence is defined as a set of numerous identifiable competencies.

Harsch (2014) brings the two approaches together by maintaining that "the question [of unitary or divisible nature of language competence can be] nowadays regarded as a 'nonquestion', as language proficiency can be conceptualised as unitary and divisible, depending on the level of abstraction and the purpose of the assessment and score reporting" (*ibid.* 152–153), giving factor analysis as support (*ibid.* 153–154).

The definition of language ability as a construct will determine the way it is assessed. Supporting the view of language proficiency being comprised of a number of clearly identifiable skills has resulted in a widespread use of skills-based language tests that require test-takers to engage in various different task types to display their respective competencies. Although these tests have a number of assets – "they aim to reflect the most scientifically credible ways in which learners represent L2 knowledge and the ability to use this knowledge for communication" (Purpura 2008: 53) – they also present problems. As "no consensus has been reached as to what exact components constitute a comprehensive model of communicative language ability, how the components might interact, how [they] are acquired and develop, [etc.]" (*ibid.* 63), inferences made on the basis of the skills-based tests about the test-takers' overall language ability should be made with caution, keeping the above in mind. Another key concern related to skills-based language testing, as mentioned above, is the time necessary to develop, administer and mark such tests.

Attempts to assess language proficiency as a unitary skill have led to the creation of various alternative assessment procedures, one of which is a c-test. The c-test was developed by Raatz and Klein-Braley in 1981 as a variation of the cloze test, proposing to integrate "all levels of language from letters through words, sentences, paragraphs to texts, but also the lexicon, the semantics and the pragmatics of a language" (Raatz, Klein-Braley 2002: 76). The c-test procedure, satisfying the above criteria, is based on the so-called rule of 2 (*ibid.*), which means that beginning from the second sentence, the second half of every second word is deleted until the required number of mutilations is reached. The test-taker would have to restore the missing part.

If a c-test is an instrument of assessing overall language proficiency, the question arises what exactly is being assessed, i.e. what the c-test construct is and what inferences can be made on the basis of its results; otherwise stated, if the instrument is a valid language ability assessment tool. The c-test construct has been investigated by numerous linguists (Sigott 2004, Eckes, Grotjahn 2006, Wilmes, 2007, Linnemann, Wilbert 2014). Eckes and Grotjahn (2006) summarise c-test construct-related research by first maintaining that "c-tests provide an integrative assessment of a construct often referred to as general language proficiency" (*ibid.* 291), which they define as "an underlying ability comprising both knowledge and skills and manifesting itself in all kinds of language use" (*ibid.*). They go on to demonstrate that c-test construct research, conducted with the help of a wide range of methods, among them correlation and factor analysis, has found a close link between c-test success and close reading ability, knowledge of lexis and grammar (*ibid*. 292–293). In addition, although weaker and perhaps slightly less investigated, a moderate to high link between c-tests and test takers' speaking and listening ability has been documented (*ibid*. 298). Research thus seems to suggest that c-test results reflect both the test-taker's micro-skills (e.g. control of subject-verb agreement) as well as the level of receptive and productive skills in general.

The validity of a c-test, like any test, has to do with the purpose for which it is used, whether it is useful in terms of the inferences we make on the basis of its scores, and to what extent those inferences are justified (Fulcher 2010: 20). The c-test is deemed useful both as a placement measure or an anchor test as well as a language proficiency measurement instrument (Eckes, Grotjahn 2006: 290). The test's purpose will determine its level of difficulty in terms of vocabulary and grammar, cohesion and genre. During the validation process, care will also have to be taken when decisions are made concerning text quality and the deletion starting point when the test is developed, as well as score interpretation once the results are in.

The c-test has a number of advantages. Besides being relatively easy, fast and inexpensive to develop, administer and score (Eckes, Grotjahn 2006: 290), research (Grotjahn, Stemmer 2002, Klein-Braley 1996) reports general high c-test reliability – "in virtually all the studies thus far reported, the c-tests have been shown to be highly reliable, with alpha coefficients very often higher than .9, and to have high correlations with whatever other measure was used to represent language proficiency" (Klein-Braley 1996: 24). Coleman (1994) maintains that in order to take a c-test, the test-takers need to call on their entire language processing competence. Klein-Braley (2002) supports this by saying that c-test completion is more demanding than simple reading or writing because for solving the items, both active and passive processes must be relied on. For example, she argues that incorrect response behaviour during test-taking and crossings out can be regarded as examples of reprocessing, which takes place at a very high level and requires pragmatic knowledge of the text. For Raatz and Klein-Braley (2002), the general language proficiency tested by the c-test seems to be similar to Bachman's operational competence – "the superordinate category for lexical, morphological, syntactical, graphological knowledge on the sentence level, and for knowledge of cohesion and rhetorical organisation on the text level" (*ibid*. 83). Köberl and Sigott (1994) point out that since the c-test consists of several texts on different topics, it enables better sampling of content, which means that the test-takers who happen to have field specific knowledge have no advantage over other test-takers. Due to its deletion rate (the second half of every second word), the probability of sampling all word classes is higher (Klein-Braley, Raatz 1984). Klein-Braley (1996) used item discrimination indices to discriminate between items and draw conclusions on their applicability. The texts showed more than 95% positive discrimination indices, which were interpreted as advantages of the c-test, as each item "measures the response behaviour in the same direction as all the other items and the whole scale" (*ibid*. 60). On the basis of inter-item correlations, she showed that items in proximity formed clusters and positive correlations between items that were further away from each other proved the need for high-level comprehension. The power to discriminate has also been pointed out by Coleman (1994: 218) who finds that a c-test offers better discrimination than for example a cloze test. An important positive feature is the c-test's integrative nature. Klein-Braley (1996: 24) proposes that since correlation between the c-test with some other measures, such as teacher judgments, self-assessments and other tests that are considered integrative, have regularly reached .7 or higher, the integrative nature of the test could be inferred. An added value of a c-test is that it can be used to measure language proficiency of both native and non-native speakers of language (Coleman 1994, Klein-Braley 1996). Raatz and Klein-Braley (2002) caution against using the test with educated native adults, though, since they should display near maximum results and thus the test would not discriminate very well. Raatz (2002) proposes a solution for this by suggesting that with educated native speakers, the c-test be speeded or the level of difficulty adjusted. A further advantage identified is that c-tests can easily be automated, again saving time and money and minimising the number of mistakes and intentional vagueness in handwriting when the test-taker is not certain about the correctness of his or her answer (Coleman 1994: 218, Koller, Zahn 1996: 416).

Although the c-test has been claimed to measure overall language proficiency effectively, not all scholars consider it unproblematic. The c-test has been mostly criticised for its lack of face validity due to the fact that the test blanks seem unnatural. For non-experts they are rather reading comprehension tests or some form of intelligence tests. Another area of contention is wondering what the test actually measures. Chapelle and Abraham (1990: 127) say that the c-test assesses more grammatical than textual competence and is therefore not an instrument for measuring overall proficiency. Eckes and Grotjahn (2006) oppose those who see the c-test as a measure of reading ability only, as there may be test-takers with high reading ability who, due to their lack of productive skills, may have very low scores on the c-test. A further controversy they see is related to test-takers' achievement of high scores on the c-test due to successful lexico-morphological processing, while having poor understanding of the text. The problematic construct of the c-test has also been highlighted by Wilmes (2007) who draws on the results of Sigott (2004), saying that test-takers with different L2 proficiency required different amounts of context. It appeared that more proficient test subjects needed less context for making decisions on the gaps, whereas less proficient students relied more on context. He concluded that since the test measured different constructs for different test takers, it could not be a valid measure of proficiency. Wilmes (2007: 13) is also critical of attempts to validate a c-test against other measures, warning that a mere presumption that another test is valid is insufficient for test validation purposes. He reminds readers that a high correlation coefficient between two measures does not signify that both measure the same construct. According to him, it is conceivable that language tests assess other psychological variables, such as concentration and intelligence, along with language proficiency.

All the above considered, it is clear that different instruments of language proficiency measurement each come with their own set of features that should prompt test users to exercise caution while interpreting their results and making inferences about test takers' language ability. Irrespective of the challenges related to c-test implementation, Eckes and Grotjahn (2006) maintain that a c-test can be used for screening large numbers of applicants before administering an expensive and time-consuming language test, or as a means of self-assessment for foreign language learners to obtain feedback on their progress (*ibid.* 290). It could be used when one needs a quick and efficient estimation of "a candidate's ability to function in a wide range of target language use situations irrespective of his or her language learning history" (*ibid.* 291), such as in placement testing, university admission or job application processes.

3. Method

As stated above, the current study looks at the process of creating and validating a c-test in the English language as a language ability measure in the Estonian context that could either supplement or serve as an alternative to a skills-based proficiency test. For that purpose, two different versions of a c-test were developed to see if different test versions relying on the same texts would produce a difference in test results. Once the c-test was ready, it was piloted among 43 Estonian form 12 secondary school students and two teachers of the same school. Its results were then correlated with those of the Estonian national examination in English 2013 (a skills-based test), using the statistical tool PSPP (no official acronymic expansion available). Correlations were also computed between the c-test and teacher evaluation of the test takers' proficiency as further validation of the c-test.

C-test development followed the procedure suggested by Raatz and Klein-Braley (2002: 84), including defining the target group, choosing suitable texts and determining their level of difficulty, bringing the tests into c-format and testing the task on educated adults. Once the results had been analysed and modifications made, the test was administered to the target group. That was followed by calculating the scores and the analyses of the test's reliability and validity.

In the process of text selection, several principles were followed. As the target group included form 12 students only, a conscious decision was made to exclude texts with specialised vocabulary and content, but to include texts that reflect the topics found in the national curriculum for Estonian secondary schools. Text difficulty was aimed at being roughly at level B, as required by the National Curriculum, and was judged considering the difficulty level of vocabulary on the one hand and text readability on the other. The level of vocabulary difficulty of the texts was estimated by means of the English Vocabulary Profile and can be seen in Table 1 below.

Table 1. Analysis of text difficulty based on vocabulary and CEFR levels

Text	Number of words	B-level	C-level	NA
1	94	31 (32.98%)	2 (2.13%)	4 (4.26%)
2	105	23 (21.90%)	6 (5.71%)	0 (0.00%)
3	99	26 (26.26%)	2 (2.02%)	2 (2.02%)

Thus, there are 94 words in the first text, 105 in the second, and 99 in the third. As regards the Common European Framework for Reference (CEFR) level, there are 31 (32.98%) B-level and 2 (2.13%) C-level words in the first text, 23 (21.90%) B-level and 6 (5.71%) C-level words in the second text, and 26 (26.26%) B-level and 2 (2.02%) C-level words in the third text. The rest of the vocabulary is on level A. The vocabulary labelled NA are words not available on the English Vocabulary Profile, but seem to belong to level C: 4 (4.26%) in the first text and 2 (2.02%) in the third. The distribution of vocabulary in the texts shows that the level of difficulty of the texts is not similar – the first text seems to be the most difficult, whereas the other two are of almost the same level. It is noteworthy, though, that the level referred to here is only related to vocabulary, and does not reflect other features like grammar or cohesion for example. As regards grammar, an attempt was made to ensure that the texts would include a wide variety of grammar structures, but since there are no objective criteria to measure their level of difficulty, a separate grammar analysis was not carried out.

A further means to verify the initial prediction of text difficulty was to use online readability scores/tools, which take into account several factors, the length of the words and sentences included. They do not help to measure the level of difficulty of the c-test as a whole but can give some insight into the extent to which the chosen texts differ, thus also helping to compare the difficulty levels of the texts for the c-test. The readability scores of the texts are presented in Table 2.

Table 2. Readability scores of the texts

The score	Text 1	Text 2	Text 3
Flesch-Kincaid Reading Ease	54.5	63	61
Flesch-Kincaid Grade Level[1]	9.6	9.8	9.7
Words per sentence	15.8	21.6	20.4

We can see that the second and third texts are easier to read than the first one, but the grade level is almost the same. And as the scales for measuring reading ease measure readability from 0 to 100, the indices show that the texts, indeed, are neither easy nor difficult.

All the texts were shortened but no other changes were made, to ensure that the texts' internal integrity was not inadvertently distorted. The texts were brought into c-test format by manually deleting the second half of every second word starting from the second sentence. Where the word consisted of an odd number of letters, the bigger part was deleted. Words consisting of one letter remained intact. As a result, the process yielded two c-tests (c-test 1 and c-test 2) both comprising three texts, containing 31, 36 and 36 gaps respectively. The starting point of the deletions in the two tests was set at different places in order to observe if the starting point would affect test results.

The two versions of the c-test, subjected to piloting can be seen below:

C-test 1

Task 0. (TIME: 20 minutes) version I
This reading task is comprised of three unrelated texts. Starting from the second sentence, the second half of every second word has been deleted. Fill in the gaps with missing letters. In case a word has an odd number of letters, the bigger part has been deleted.
An example (0) has been done for you.

If only you could enjoy flying without fear – you are not alone. Many peo_*ple*_ (0) develop fe______ as th______ mature a______ life se______ more prec______, while oth______ may ha______ experienced a b______

[1] https: //readability-score.com/ The Flesch–Kincaid Grade Level Formula presents a score as a U.S. grade level, making it easier for interested parties to evaluate the readability level of various texts. It can also mean the number of years of education generally required to understand this text, relevant when the formula results in a number greater than 10.

flight. Y______ must ha______ a strong imagi______ or rece______ started a fam______. But regar______ of h______ fears dev______, those w______ suffer c______ experience slee______ nights, elev______ anxiety, a______ panic att______. Concerns m______ include wea______, turbulence, take-______, flying ov______ water, claustr______, crowds, los______ control, hijac_______, and fe______ of hei______. You can overcome your fear right now using my online Fear of Flying Help Course. (*Adapted from* http://www.fearofflyinghelp.com)

Recent research has shown that having a pet can strengthen children's immune system, and make them less likely to have days off school with illnesses than those without animals in the home. Researchers disco______ that chil______ of fami______ who we______ either c______ or d______ owners h______ more hea______ problems, b______ as th______ grew ol______, their imm______ systems we______ given a bo______. They atte______ an ave______ of ni______ days mo______ school th______ those w______ didn't ha______ pets. Th______ theory sugg______ that be______ too cl______ in ea______ childhood wea______ the imm______ system. How______, despite contri______ to bet______ school atten______, pets c______ also p______ children's hea______ at ri______. (*Adapted from Click On 4*)

You don't need a book to tell you what it's like looking for a job in a tough market – unemployment levels are rarely out of the news. There a______ plenty o______ people comp______ for jo______, and empl______ have th______ pick fr______ a consid______ number o______ candidates. Fri______ and fam______ will alm______ gleefully te______ you ab______ many peo______ who ha______ been for______ to ta______ poorly pa______ jobs, o______ people w______ have app______ for ov______ a thou______ jobs wit______ success. B______ news ma______ us s______ difficulty rat______ than oppor______. That te______ us th______ the min______ adopted i______ just a______ important a______ planning. (*Adapted from John Lees "Just a Job! A smart and fast strategies to get the perfect job"*)

C-test 2

Task 0. (TIME: 20 minutes) version II
This reading task is comprised of three unrelated texts. Starting from the second sentence, the second half of every second word has been deleted. Fill in the gaps with missing letters. In case a word has odd number of letters, the bigger part has been deleted.
An example (0) has been done for you.

If only you could enjoy flying without fear – you are not alone. Many peo_*ple*_ (0) may dev______ fear a______ they mat______ and li______ seems mo______ precious, wh______ others m______ have exper______ a b______ flight. Y______ may ha______ a strong imagi______ or recently sta______ a fam______. Regardless o______ how fe______ develop, th______ who suf______ can exper______ sleepless nig______, elevated anx______, and pa______ attacks. Conc______ may inc______ weather, turbu______, take-offs, fly______ over wa______, claustrophobia, cro______, losing con______, hijackings, a______ fear o______ heights. You can overcome your fear right now using my online Fear of Flying Help Course. (*Adapted from* http://www.fearofflyinghelp.com)

Recent research has shown that having a pet can strengthen children's immune system, and make them less likely to have days off school with illnesses than those without animals in the home. Researchers ha______ discovered th______ children o______ those fami______ who we______ cat o______ dog own______ had mo______ health prob______, but a______ they gr______ older, th______ immune sys______ were gi______ a boost. Th______ attended a______ average o______ nine da______ more sch______ than th______ who di______ have pe______. This the______ suggests th______ being t______ clean i______ early chil______ weakens t______ immune sys______. However, des______ contributing t______ better sch______ attendance, pe______ can al______ put chil______ health a______ risk. (*Adapted from Click On 4*)

You don't need a book to tell you what it's like looking for a job in a tough market – unemployment levels are rarely out of the news. There are ple_______ of peo______ competing f______ jobs, a______ employers ha______ their pi______ from a qui______ considerable num______ of candi______. Friends a______ family wi______ almost glee______ tell y______ about ma______ people w______ have be______ forced t______ take poo______ paid jo______, or peo______ who ha______ applied f______ more th______ a thousand jo______ without suc______. Bad ne______ makes u______ see diffi______ rather th______ opportunity. Th______ tells u_____ that t______ mindset ado______ is ju______ as impo______ as plan______. (*Adapted from John Lees "Just a Job! A smart and fast strategies to get the perfect job"*)

Although it is suggested that c-tests be pretested by native adults, it was not considered relevant for this study, since native speakers' performance could not be considered an appropriate measure of the desired proficiency. Instead, the versions of the c-test were pretested by two educated non-native adults, the teachers of English of the students comprising the test population, each responding to a different variant. The scores obtained exceeded 95%, which corresponds to the suggested level of difficulty for educated adults (Raatz, Klein-Braley 2002). In addition to determining the appropriate level of difficulty, the pretesting by the teachers served as a means of editing the test for errors.

The c-test was administered along with the skills-based 2013 national examination paper, with the whole procedure serving as a mock exam, preparing students for their national examination. The c-test was attached to the reading paper of the national examination as it was here that the students were most likely to treat it as a plausible task, with the different versions of it distributed among students randomly. Students were aware that the c-test did not form part of the national exam. The number of both versions of the c-test for piloting among the students was near equal – 22 for one group and 21 for the other. When administering the c-test with two versions, care was taken (to increase test reliability) to ensure that the students were not able to see each other's versions as blanks in one test version might be undeleted words in the other. The time assigned for the examination was generally the same as allowed at the national examination in 2013, but because of the added c-test, the overall time was extended by 20 minutes. The c-test was completed by all students, but as it was administered within the national examination paper, it was not possible to measure the time actually spent on it by the test-takers.

The tests were marked by one assessor, following the evaluation guidelines for national examination papers: writing and speaking were marked subjectively using the respective marking scales for letters, essays and speaking; listening and reading was conducted objectively relying on the answer key (cf. Riigieksamite materjalid 2013). In line with the suggestions in c-test literature, exact scoring was used with the c-test, which meant that non-responses and responses with misspelling resulted in no points. The scoring also took account of the suggestion to form superitems, the number of which corresponded to the number of texts in the c-test. Each superitem was attributed the same weight and the results were combined. For combining the results, the score for each text was expressed as a percentage of the maximum for that text. That was followed by finding the average for the three texts. This resulted in an overall score expressed on a scale of 100, offering the same degree of discrimination as scoring one point for each correct item would have allowed.

For validation purposes, correlations were established between the c-test results and the national examination results and also with the teacher evaluation of the students' language ability, estimated subjectively by their English teacher on a 100-point scale.

4. Results and discussion

PSPP freeware, similar to SPSS (Statistical Package for Social Sciences), was used for statistical analysis, including the calculation of descriptive statistics to express the distribution of the results, comparison of means and bivariate correlations. Descriptive statistics of all three texts were calculated to investigate if the text difficulty based on the analysis of the vocabulary was correctly estimated in the test-construction stage. The results are shown in Tables 3 and 4.

Table 3. Text difficulty of the three texts in c-test 1 expressed by means

Variable	N	M	SD	Minimum	Maximum
C1a	22	18.55	6.29	7.00	26.00
C1b	22	24.45	6.82	11.00	34.00
C1c	22	19.86	6.75	10.00	32.00

Table 4. Text difficulty of the three texts in c-test 2 expressed by means

Variable	N	M	SD	Minimum	Maximum
C2a	21	18.05	5.53	9.00	27.00
C2b	21	25.81	5.30	17.00	34.00
C2c	21	28.19	3.30	22.00	33.00

The mean for the three tests, expressed in percentage (c1a = 59.8%, c1b = 67.9%, c1c = 55.8%; c2a = 58.2%, c2b = 71.7%, c2c = 78.3%) suggests that text difficulty varied slightly in the two variants. In test 1, text c proved to be the most challenging, whereas in test 2, it was text a that was the most difficult of the three, displaying the lowest mean, which is in line with the prediction made on the basis of vocabulary analysis. The same can be found during the comparison with Flesch-Kincaid scores (cf. p. 6) regarding text difficulty. Test reliability was investigated with the help of Cronbach's alpha, where the coefficient of .90 or more was sought. For c-test 1, the respective coefficient was .91 and for c-test 2, it was .92. Thus, all three texts in both tests appear to measure language ability reliably.

In order to achieve more reliable results by analysing a larger sample, the intent was to combine the individual results of the two tests. To do that, it was investigated if the two versions of the c-test measured the construct similarly, as only then would combining be justified. This meant calculating the mean scores and distribution. The outcome is presented in Table 5 below.

Table 5. Statistical means of the two c-tests

Measure		**C1 total**	**C2 total**
N	Valid	22	21
	Missing	0	0
Mean		60.98	70.80
Standard deviation		17.79	13.50
Range		56.39	41.51
Minimum		28.97	46.69
Maximum		85.36	88.20

The means of the two versions of the c-test varied by almost 10 points (mean 1 = 60.98 vs. mean 2 = 70.80). This might have stemmed from the differences in minimum results. Thus, c-test 1 was either more difficult than c-test 2 or the test subjects who took c-test 1 represented students with lower abilities. A further consideration may be the deletion starting point, i.e. it may have been that the point where deletion started in c-test 1 rendered it more difficult. In addition to being either easier or having been taken by mostly high-ability students, reflected by a higher mean, c-test 2 also showed more homogenous results indicated by a smaller standard deviation and range.

It was decided to continue establishing correlations between c-test 1 and c-test 2 with the results of the national examination paper by using paired sample statistics, suitable for a small sample size. It was hypothesised that if the results of both c-tests show a high correlation with the students' national examination result, the results of the two c-tests could be combined to form a bigger sample for the analysis and to draw more substantiated conclusions of students' language proficiency on the basis of the results of the c-test. By comparing c-test results to those of a skills-based test (the national examination), conclusions could be ventured about the level of different language skills being reflected in the c-test results. That in turn could prove useful for estimating the c-test's viability as proficiency measurement instrument. The correlations are shown in Table 6 and 7 below.

Table 6. Correlation between the results of c-test 1 and national examination

Paired Sample Statistics					
		Mean	**N**	**Std. Deviation**	**S. E. Mean**
Pair 1	Year 12	68.06	22	19.94	4.06
	C1Total	60.98	22	17.79	3.79
Paired Samples Correlation			**N**	**Correlation**	**Sig.**
Pair 1	Year 12 & C1Total		22	.88	.000

Table 7. Correlation between the results of c-test 2 and national examination

Paired Sample Statistics					
		Mean	**N**	**Std. Deviation**	**S. E. Mean**
Pair 1	Year 12	70.69	21	14.86	3.24
	C2Total	70.80	21	13.50	2.95
Paired Samples Correlation			**N**	**Correlation**	**Sig.**
Pair 1	Year 12 & C2Total		21	.90	.000

As we see, there is a strong correlation (.88) between the results of c-test 1 and the national examination results (significant at p < .001), and the respective correlation with c-test 2 is even stronger (.90), significant at p < .001. On the basis of very similarly strong correlation levels, it was decided that the results of the two c-test versions can be combined to form a bigger and a more reliable sample.

The next step in the analysis was to correlate the new sample with the results of the individual skills papers on the national examination, and teacher evaluation. The correlations can be seen in Table 8 below.

Table 8. Correlation between individual skill tests' results, overall skills-based test score, the c-test and teacher evaluation

		Listening	**Reading**	**Lang**	**Writing**	**Speaking**	**TOTALyear12**	**TOTALC**	**Teacher ev**
Listening	Pearson Correlation	1.00	.86	.89	.68	.82	.95	.81	.89
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000	.000
Reading	Pearson Correlation	.86	1.00	.80	.60	.78	.91	.81	.84
	Sig. (2-tailed)	.000		.000	.001	.000	.000	.000	.000
Lang	Pearson Correlation	.89	.80	1.00	.72	.81	.94	.83	.89
	Sig. (2-tailed)	.000	.000		.000	.000	.000	.000	.000
Writing	Pearson Correlation	.68	.60	.72	1.00	.62	.78	.62	.75
	Sig. (2-tailed)	.000	.001	.000		.000	.000	.000	.000
Speaking	Pearson Correlation	.82	.78	.81	.62	1.00	.91	.79	.82
	Sig. (2-tailed)	.000	.000	.000	.000		.000	.000	.000
TOTAL year12	Pearson Correlation	.95	.91	.94	.78	.91	1.00	.86	.93
	Sig. (2-tailed)	.000	.000	.000	.000	.000		.000	.000
TOTALC	Pearson Correlation	.81	.81	.83	.62	.79	.86	1.00	.85
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000		.000
Teacher ev	Pearson Correlation	.89	.84	.89	.75	.82	.93	.85	1.00
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	

There is a strong positive correlation between the results of the c-test and all the skills tests, and all the correlations are highly significant (p < .001). The c-test correlates best with language structures (.83), listening (.81), reading (.81), and speaking (.79). The correlation with writing is somewhat weaker, but still meaningful at .62. The correlation with the total score of the national examination in English is .86, which can be considered very strong. The correlation between the overall score of the national examination and teacher evaluation was .93 (p < .001). Teacher evaluation and c-test correlation was slightly weaker but still strong at .85 (p < .001). The above correlations seem to strongly suggest that the c-test can be used as a tool for measuring language proficiency, as the results obtained by it are very similar to those obtained by a skills-based test as well as teachers' evaluation of the students' proficiency.

5. Conclusion and implications

The current article investigated the construct of language proficiency and the approaches to measure it depending on the view taken. Even though most high-stakes contexts subscribe to the tried and tested skills-based approach to language proficiency assessment, there are other contexts where valid and reliable judgements about candidates' language ability need to be made without having the financial, temporal or administrative resources to administer such tests. The current study proposed the c-test as a more economical alternative and investigated the process of developing and administering it, providing statistical analysis with regard to the test's validity and reliability.

The results demonstrated that if careful c-test development procedures are observed (consideration of the level of text length and difficulty, observing the presence of both content and function words in the items while deciding the start of the deletion cycle, etc.), a high level of inner consistency can be achieved in the c-test, resulting in the test working as a reliable instrument of discriminating between stronger and weaker candidates. A very high positive correlation between the c-test results and the individual skills test results (with the writing skill correlating somewhat less), as well as the teachers' evaluation of the students' proficiency, seems to serve as proof that c-test results can be trusted to indicate candidates' language proficiency appropriately. Given the somewhat weaker correlation with the writing test, it could be suggested that if c-tests were to be used as language level indicators, they could be supplemented by a writing task where candidates show their writing ability more directly. The results obtained should be treated with caution, though, as the number of people involved in the study is small, representing one teaching context, restricting its generalisability.

Challenges related to c-test implementation concern its validation: choosing good quality texts of appropriate complexity, deciding the deletion starting point and interpreting the scores, i.e. how the proficiency level is decided based on the score obtained. For the latter, c-test validation probably has to include comparing its results against those of a skills-based test, deemed valid for the context and purpose for which the c-test is developed. This, incidentally, will reduce the c-test's value as a quick proficiency appraisal instrument, as such comparison inevitably takes time. It will, however, not diminish the test's value as a tool for quick selection purposes. As regards the test's lack of face validity, this can only be overcome if the task type is more consistently included in the test preparation process where test developers, teachers, students and proficiency evaluation agencies are familiarised with the task's nature and peculiarities. With the c-test principle applied in more than 20 language contexts, including English, French, German, Japanese and Turkish (Eckes, Grotjahn 2006: 290), the tool could be more boldly included in the language ability measurement repertoire in Estonia, too.

References

Bachman, Lyle 1990. Fundamental Considerations in Language Testing. Oxford, UK: Oxford University Press.

Bachman, Lyle; Palmer, Adrian 1996. Language Testing in Practice. Oxford, UK: Oxford University Press.

Canale, Michael; Swain, Merrill 1980. Theoretical bases of communicative approaches to language teaching and testing. – Applied Linguistics, 1 (1), 1–47. http://dx.doi.org/10.1093/applin/1.1.1

Chapelle, Carol A.; Abraham, Roberta G. 1990. Cloze method: what difference does it make? – Language Testing, 7 (2), 121–146. http://dx.doi.org/10.1177/026553229000700201

Coleman, James A. 1994. Profiling the advanced language learner: The c-test in British further and higher education. – Rüdiger Grotjahn (Ed.), Der C-Test. Theoretische Grundlagen und Praktische Anwendungen, Bd. 2. Bochum: Brockmeyer, 217–237.

Coleman, James A. 1996. A comparative survey of the proficiency and progress of language learners in British universities. – Rüdiger Grotjahn (Ed.), Der C-Test. Theoretische Grundlagen und praktische Anwendungen, Bd. 3. Bochum: Brockmeyer, 367–399.

Eckes, Thomas; Grotjahn, Rüdiger 2006. A closer look at the construct validity of c-tests. – Language Testing, 23 (3), 290–325. http://dx.doi.org/10.1191/0265532206lt330oa

Fulcher, Glenn 2010. Practical Language Testing. Hodder Education.

Grotjahn, Rüdiger; Stemmer, Brigitte 2002. C-tests and language processing. – J. Coleman, R. Grotjahn, U. Raatz (Eds.), University Language Testing and the C-Test. Bochum: AKS-Verlag, 115–130.

Grotjahn, Rüdiger 1987. How to construct and evaluate a c-test: A discussion of some problems and some statistical analyses. – R. Grotjahn, C. Klein-Braley, D. K. Stevenson (Eds.), Taking Their Measure: The Validity and Validation of Language Test. Bochum: Brockmeyer, 219–253.

Gümnaasiumi riiklik õppekava. http://www.oppekava.ee/index.php/G%C3%BCmnaasiumi_riiklik_%C3%B5ppekava (10.3.2016).

Harsch, Claudia 2014. General language proficiency revisited: Current and future issues. – Language Assessment Quarterly, 11 (2), 152–169. http://dx.doi.org/10.1080/15434303.2014.902059

Hulstijn, Jan H. 2007.The shaky ground beneath the CEFR: quantitative and qualitative dimensions of language proficiency. – Modern Language Journal, 91 (4), 663–667. http://dx.doi.org/10.1111/j.1540-4781.2007.00627_5.x

Hulstijn, Jan H. 2015. Language Proficiency in Native and Non-Native Speakers: Theory and Research. John Benjamins Publishing Company. http://dx.doi.org/10.1075/lllt.41

Hymes, Dell H. 1972. On communicative competence. – J. B. Pride, J. Holmes (Eds.), Sociolinguistics. Selected Readings. Harmondsworth: Penguin, 269–293.

Inglise keele riigieksami eristuskiri. http://www.innove.ee/UserFiles/erituskirjad_2015/RE%20inglise%20keel%20eristuskiri%202015.pdf (3.4.2015).

Klein-Braley, Christine 1996. Towards a theory of c-test processing. – R. Grotjahn (Ed.), Der C-Test. Theoretische Grundlagen und Praktische Anwendungen, Bd. 3. Bochum: Brockmeyer, 23–94.

Klein-Braley, Christine 2002. Psycholinguistics of c-test taking. – J. Coleman, R. Grotjahn, U. Raatz (Eds.), University Language Testing and the C-Test. Bochum: AKS-Verlag, 131–142.

Klein-Braley, Christine; Raatz, Ulrike 1984. A survey of research on the c-test. – Language Testing, 1 (2), 134–146. http://dx.doi.org/10.1177/026553228400100202

Koller, Gerhard; Zahn, Rosemary 1996. Computer based construction and evaluation of C-tests. – R. Grotjahn (Ed.), Der C-Test. Theoretische Grundlagen und praktische Anwendungen, Vol. 3. Bochum: Brockmeyer, 401–418.

Köberl, Johann; Sigott, Günther 1994. Adjusting c-test difficulty in German. – R. Grotjahn (Ed.), Der C-Test. Theoretische Grundlagen und Praktische Anwendungen, Vol. 2. Bochum: Brockmeyer, 179–192.

Linneman, Markus; Wilbert, Jürgen 2010. The C-Test: A valid instrument for screening language skills and reading comprehension of children with learning problems? – R. Grotjahn (Ed.), The C-Test: Contributions from Current Research. Frankfurt a.M.: Lang, 113–124.

Oller, John W. 1979. Language Tests at School: A Pragmatic Approach. London: Longman.

PSPP. https://www.gnu.org/software/pspp/ (15.12.2015).

Purpura, James E. 2008. Assessing communicative language ability: Models and their components. – Elana Shohamy, Nancy H. Hornberger (Eds.), Language Testing and Assessment. Encyclopedia of Language and Education. Springer, 53–68.

Raatz, Ulrich; Klein-Braley, Christine 2002. Introduction to language testing and to c-tests. – J. Coleman, R. Grotjahn, U. Raatz (Eds.), University Language Testing and the C-Test. Bochum: AKS-Verlag, 75–91.

Riigieksamite materjalid 2013. http://www.innove.ee/et/riigieksamid/riigieksamite-materjalid/riigieksamite-materjalid-2013 (15.12.2015).

Sigott, Günther; Köberl, Johann 1996. Deletion patterns and c-test difficulty across languages. – R. Grotjahn (Ed.), Der C-Test. Theoretische Grundlagen und praktische Anwendungen, Bd. 3. Bochum: Brockmeyer, 159–172.

Sigott, Günther 2004. Towards Identifying the C-Test Construct. Peter Lang Verlag.

Wilmes, Carsten 2007. Validation of a German Language Placement Test Based on a Modified C-test Procedure. Dissertation.

Ingrid Sarapuu (Rapla Ühisgümnaasium) teadushuviks on keeleoskuse mõõtmine ja riigieksami arendus.
Keskkooli 2, 79512 Rapla, Estonia
sarapuu.ingrid@gmail.com

Ene Alase (Tallinna Ülikool) teadushuvid on keeletestimine, testide koostamine ja nende kvaliteedi hindamine, õpetajakoolitus ja õppekirjanduse hindamine.
Narva mnt 25, 10120 Tallinn, Estonia
enealas@tlu.ee

C-TEST KEELEOSKUSE MÕÕTMISEKS OSAOSKUSTESTI ALTERNATIIVINA

Ingrid Sarapuu[1], Ene Alas[2]
Rapla Ühisgümnaasium[1], Tallinna Ülikool[2]

Artiklis vaadeldakse keelepädevust kui konstrukti ja näidatakse, kuidas konstrukti erinev defineerimine on viinud keeleoskuse testide erineva ülesehituseni, lähtudes sellest, kas keelt nähakse osaoskuste kogumina või ühtse jagamatu oskusena. Vaadeldes keelt jagamatu oskusena on osaoskustel põhinevate testide kõrvale tekkinud teisi alternatiivseid testi-tüüpe, millest üheks on c-test. Artiklis analüüsitakse c-testi omadusi ning püütakse selgusele jõuda, kas sellist testitüüpi on võimalik kasutada üldise inglise keele pädevuse hindamiseks Eesti kontekstis. Lähtudes Grotjahni (1987) ning Raatzi ja Klein-Braley (2002) mudelist, vaadeldakse kõigepealt c-testi koostamise ja administreerimise protsessi gümnaasiumi kontekstis paralleelselt riigieksamiga. C-testi katsetamise tulemusi võrreldakse seejärel samade gümnaasiumiõpilaste poolt sooritatud osaoskustel põhineva keelepädevustesti tulemustega ning õpilaste õpetajapoolsete pädevushinnangutega. Uurimusest selgub, et kuigi c-testil on olulisi puudusi, on c-testi korrelatsiooniindeksid nii osaoskustel põhineva pädevustestiga kui ka õpetajapoolsete hinnangutega võrreldes väga kõrged, lubades pidada c-testi usaldusväärseks keeleoskuse hindamise instrumendiks, mille peamisteks eelisteks on testi koostamise suhteline kiirus ja madal administreerimiskulu.

Võtmesõnad: keeleoskuse hindamine, üldine keelepädevus, eksamiarendus, valiidsus, reliaablus, riigieksam