Tabaran Institute of Higher Education
International Journal of Language Testing
In press

ISSN 2476-5880
https://www.ijlt.ir

Validation of C-Test among Iraqi EFL University Students

Aalaa Yaseen Hassan[1], Eman Adil Jaafar[2], Marwah Firas Abdullah Al-Rawe[3], Shaden Shamel Abdullah[4]

ARTICLE INFO

Article History:
Received: April 1st
Accepted: May 17

KEYWORDS
C-Test
Language Ability
Reliability
Validity

ABSTRACT

Determining students' language proficiency is essential for successful instruction and learning objectives in many educational settings. To this end, one of the most efficient assessment tools for measuring language proficiency is the C-test. Thus, the primary aim of this research is to assess the performance of university students through the C-Test and to analyze the extent to which this test is valid in measuring language ability. A standardized C-Test has been created with four brief passages, each containing 20 gaps. The length of each passage varied from 95 to 109 words. Throughout each passage, only the first and last sentences were not changed. The test was taken by 100 students; 39 were male and 61 were female at Al-Nisour University/Department of English in Baghdad, Iraq. The sample consists of two groups. Both groups come from the same school and would receive similar educational input in both cases based on their grade level. The validity and reliability of the C-Test were investigated using various techniques. The study analyzed the performance of fourth- and third-year students on the Common Language Proficiency Test in Iraq. The results showed that the test discriminates well between high-ability and low-ability examinees, with no significant difference between the two groups. The Rasch model separation reliability was relatively high, and the data were one-dimensional. The students faced difficulties in guessing the most appropriate words due to their limited English proficiency. The results suggest that developing and implementing this test could significantly improve students' academic achievements in basic foreign language classes in Iraq.

1. Introduction

The study of C-Tests is viral and widely used because many studies have been presented on this topic since the 1980s, which are varied in decisive ways (Eckes & Grotjahn, 2006). Various decisions have been supported by C-Test scores where the construction of this test developed differently. Researchers and developers of tests applied C-Tests to language programs to assess students' learning achievements (e.g., Mozgalina & Ryshina-Pankova, 2015; Norris, 2006; Rasoli, 2021). Karimi (2011) tried to use a C-Test to assess vocabulary knowledge in the second language. On the other hand, Grujić and Danilović (2012) showed the importance of pedagogical values by examining students; thus, the C-Test was used as a means of learning and teaching tools in most studies. Another kind of test is applied to disabled students by Linnemann and Wilbert (2010), who confirmed that this test has not been designed for these students. However, the results proved that this test is applicable because it showed good qualities. The researchers, Daller and Phelan (2006), tried to use the C-Test as a partially replaceable version of the Test of English for International Communication (TOEIC®). They proved

[1] English Department, College of Education for Women, University of Baghdad, Iraq, Email: alaa.y@coeduw.uobaghdad.edu.iq
[2] English Department, College of Education for Women, University of Baghdad, Iraq, Email: eman_jafer@coeduw.uobaghdad.edu.iq
[3] Department of English, College of Education for Humanities, University of Anbar, Iraq, Email: marwa.feras@uoanbar.edu.iq
[4] Journalism Department, College of Media, Aliraqia University, Iraq, Email: shadan.s.abdullah@aliraqia.edu.iq

that the C-Test is more suitable than TOEIC® to measure language ability. Ikeguchi (1998) used two C-Tests to measure university students' proficiency; one of these tests is long, and the others are short. His study revealed that the learner's responses in the four short passages were better than the long ones. As a result, all of these studies have been conducted on this test; each has its impact and ways of examining the student's performance in the second or foreign languages.

These variations pose challenges to researchers, especially in the assessment of the proficiency of a second language. Instruments for testing language proficiency have been utilized with new approaches, which have been tried and tested by experts who are unsatisfied with these instruments (Rasoli, 2021). Spolsky (1973) said that "in searching for a test of overall proficiency, we must try to find some way to get beyond the limitations of testing a sample of surface features and seek rather to tap into the underplaying of linguistic competence" (p. 175). Thus, he gave two choices: firstly, it is an interview, which is difficult to establish and administer the reliability of scoring the measure of proficiency, and this has been restricted in its implementation; secondly, it is the reduced redundancy principle (RRP). As mentioned by Klein-Braley (1997), the operationalization of the last option was well-received by a large number of specialists as testing procedures of different types: cloze test, C-Test, dictation, rational dictation cloze, the noise test, partial dictation, multiple-choice cloze, and the cloze elide.

Both cloze tests and C-Tests were developed to measure language ability in written texts to help students restore missing or deleted parts of words. Thus, the researchers sometimes referred to C-Tests as cloze tests in literature reviews (Chapelle & Abraham, 1990). For cloze tests, the test takers can give correct responses when they work on short and long contextual constraints. The inferential ideas, social setting, facts, states of affairs, and relationships help students to restore this information in the passage, which is considered an example of pragmatic mappings (Oller, 1983). Klein-Braley and Raatz (1984) and Klein-Braley (1997) presented different assertions about the C-Test: it is not difficult to establish it; it is possible to give different items to shorter texts; it has one possible solution. For this reason, scoring is objective; native speakers or teachers would not take time to read the text; thus, scoring is quick; it is not difficult for native speakers; the C-Test has several different texts.

The C-Test is an economic test that measures language ability and is simple to use, score, and create from different texts. In comparison to the cloze test, it is shorter and has more deletions. In this study, the researchers try to measure the performance of Iraqi students through the C-Test and analyze the extent to which this test is valid in measuring language ability hypothesizing that fourth-year students will achieve better proficiency levels on the C-Test than third-year students, and this is due to their superior linguistic skills and analytical abilities in the English language. It also hypothesizes that this kind of test is valid to measure the achievements of examinees, which improved after careful analysis of the students' answers to both stages. The researchers developed two broad questions based on the problem statement, which will then be answered:

Q1: Is the C-Test a reliable tool for assessing language ability among Iraqi students learning English as a second language?
Q2: Is the C-Test a valid tool for assessing Iraqi students' language proficiency?

2. Review of Literature

Both test experts and test users have found that the C-Test is an acceptable assessment instrument to measure proficiency in foreign languages and in the first languages, as well it has also been used in several studies, as published dissertations (Klein-Braley, 1981; McKay, 2019; Roos, 1994; Stemmer, 1991) and articles (Babaii & Ansary, 2001; Cohen et al., 1984; Klein-Braley, 1994; Klein-Braley, 1996; Peppé & McCann, 2003). It is effective in measuring the range of students' abilities in languages such as English, Arabic, Spanish, etc. (Coleman, 1996). C-Tests are also associated with various measures of vocabulary knowledge (e.g., Singleton & Little, 1991; Singleton, 1999). A comparison of native and non-native English speakers' performance on C-Tests found that language proficiency was highly related to C-Test scores, with higher scores indicating greater proficiency (Park, 1998).

One of the theories that is behind the C-Test is Redundancy in Language. According to this theory, a redundant message implies more information beyond what is essential for understanding the

message. Adult-educated native speakers of a language use their competence to restore damaged messages. The learners, lacking fully developed competence, may struggle to employ redundancies to the restoration process. In 1968, Spolsky and other researchers used this idea to justify the Noise Test, whereas Oller (1983) justified its use for close tests in foreign language assessments. The Noise test works by increasing levels of electronically generated hissing sounds to each sentence, which the test takers must write. On the other hand, cloze tests, initially developed by Taylor (1953) as readability measures, work by systematically deleting every seventh word from a written text, where n is a number between 5 and 10. To ensure adequate sampling, it is advised to use tests with fifty blanks. Cloze tests have high reliability and validity (Coleman et al., 2002).

Alderson (1979a & 1979b) and Klein-Braley (1981) found that although cloze tests are theoretically sound, they suffer from technical flaws, such as being too long, having only one longer text, affecting reliability and validity coefficients, being overly challenging for adult-educated native speakers, requiring a significant subjective component, and consuming a considerable amount of time. Furthermore, cloze tests are difficult to perform depending on how many structure and content words are removed, and many reported tests have proven to be less reliable than originally expected. In 1981, Raatz and Klein-Braley introduced C-Tests as a technical advancement over cloze tests. The new test format was designed to be shorter, comprising a minimum of 100 items with a fixed deletion rate and a representative sample of text elements. Examinees with specialized knowledge were excluded from the test, and exact scoring was adopted. Native speakers had to achieve excellent scores of 90% or higher on the C-Tests. Notably, the test was intended to be reliable, valid, and easily developed. The preliminary findings in 1982 confirmed that these requirements had been met.

Tests must meet criteria such as standardization, objectivity, reliability, and validity for accurate findings. Standardized tests require identical administration, material, conditions, and scoring procedures for all subjects, ensuring comparable results and meeting quality criteria. If the administration, scoring, or interpretation of a test cannot be affected by the test user, it is considered objective. The reliability of a test, which indicates its measurement accuracy, is measured by a coefficient that falls within the range of 0 (for errors) to 1 (for accurate answers). There are four methods commonly employed to assess this coefficient: retest, parallel test, split-half, and analysis of inner consistency. In the case of C-Tests, consistency analysis is conducted using super-items, employing Cronbach's Alpha Formula instead of the Kuder-Richardson formula. A test is considered valid if it accurately measures the intended construct or fulfills its intended purpose. Content validity relies on the relevance and representativeness of the test content, while construct validity is more user-oriented and practical. C-Tests demonstrate content validity by utilizing authentic texts and reflecting real-life language usage. Notably, there is a significant correlation between C-Test results and teacher grades, indicating that teachers' evaluations are comprehensive and not merely based on isolated grading procedures (Coleman et al., 2002). It has been used several methodologies to assess the validity and reliability of the C-Test here.

In this study, the "rule of two" is used, starting from the second word in the second sentence and deleting the second half of each subsequent word. This rule originated with the suggestion made by Klein-Braley and Raatz (1984) that every second word, beginning with the second sentence, be mutilated rather than certain words being deleted. Klein-Braley (1997) added that, in addition to following this guideline, four to six brief texts should be carefully selected and arranged according to difficulty in an intuitive manner. English language instructors or native speakers with adult education should learn the recently created C-Test to try out. Before finalizing standard C-tests to be ready for the target population, the tests must ensure that native speakers correctly restore 90% of the mangled words. These procedures make creating C-tests extremely difficult, if not very laborious; this is particularly true in the case of foreign languages, where it is nearly impossible to find cooperating native speakers.

3. Method

3.1 Research Design

This study adapts Coleman et al. (2002) methodology. A C-Test is comprised of several brief authentic texts, typically ranging from five to six, each text can stand alone as a self-contained unit of meaning. The first sentence is left intact in these passages. The "rule of two" is used, starting from the second word in the second sentence and deleting the second half of each subsequent word. The

undamaged parts in this test are the numbers and proper names. For the examinees, the instructions guide them to replace the damaged or missing parts in the test. The test starts with the easiest text and moves on to the most difficult ones. Five minutes were given for each text, for example, a test with six parts takes thirty minutes, and one point is given to each correct restoration. The C-Test gives an approximate measure of general language proficiency, the interpretation of this can be either norm-oriented (comparing the examinee's performance to other test takers) or criterion-oriented (determining if a specific proficiency level has been attained). To this end, a standardized C-Test battery with four brief passages, each containing 20 gaps, was employed. The length of each passage varied from 95 to 109 words. Throughout each passage, only the first and last sentences were not changed. There were four passages (see Appendix A) included in the test (Engelhardt, 2013). The four passages covered different subjects as biology, conversation, linguistics, and sleep science.

3.2. Participants and Setting

During regular class periods, students were given 30 minutes to finish the test. The test was taken by 100 students; 39 were male and 61 were female at Al-Nisour University/Department of English in Baghdad, Iraq during the academic year 2023-2024. All of the participants who took part in this research are native speakers of Arabic, and English is a foreign language to them. The sample consisted of two age groups: third-year students (female, M= 23.97, SD =3.66; male, M= 24, SD = 3.55) and fourth-year students (female, M= 21.92, SD =2.31; male, M= 21.95, SD = 2.14). Both groups come from the same school and would receive similar educational input in both cases based on their grade level.

3.3. Procedures

The study's procedural steps included the following: Initially, a comprehensive literature review was conducted to provide a theoretical foundation, focusing on the efficacy of assessing language abilities, particularly in English, and examining previous studies related to the problem. Subsequently, the research sample is specified, comprising 100 fourth-year and third-year university students from Al-Nisour University. Following this, the study entails designing, executing, and evaluating a pilot test to assess its suitability. The Winsteps program was used to analyze the data from the four passages after participants had completed them (Linacre, 2023b). Statistical techniques are then applied to determine the test's reliability. Lastly, the obtained test results are subjected to further analysis utilizing these statistical techniques.

3.4. Data Analysis

A variety of measures were used in this paper to assess the reliability and validity of the test. Descriptive statistics, including means, standard deviations, and range were utilized to show inter-item correlation within the C-Test. T-Tests were conducted to compare means between 3rd-year and 4th-year students to identify significant differences. The Rasch model, encompassing Differential Item Functioning (DIF), was used to ascertain the validity of the C-Test items across different groups. The reliability of a test is usually assessed by using a single test to determine its accuracy (McCowan & McCowan, 1999). According to Heaton (1975), a reliable test has stable scores. Cronbach's Alpha was used to demonstrate the reliability of the test, with a coefficient of 0.88 for both groups.

4. Results

After careful analysis of the students' answers in both stages (see Appendix B, 2024), results through Table 1 show the means, standard deviations, range, minimum, and maximum for each text.

Table 1
Descriptive Statistics for the Texts Used in the Study

	Range	Minimum	Maximum	Mean	SD
Text1	18	2	20	12.96	4.61
Text2	20	0	20	12.20	5.31
Text3	20	0	20	15.12	4.81
Test4	20	0	20	12.75	4.80

Table 2 shows the inter-item correlation within the text of C-Test. The inter-item correlations among passages are above 0.50, which suggests statistical significance. As can be seen from the corrected item-total correlations, or discriminations, all items have very high coefficient correlations with the total score. In other words, the items serve as a useful instrument for distinguishing between examinees with high and low skill levels.

Table 2
Inter-Item Correlation Matrix and Discrimination

	Text1	Text2	Text3	Text4	Discrimination
Text1	1	.526	.624	.663	.671
Text2		1	.639	.714	.705
Text3			1	.776	.780
Text4				1	.840

An independent sample t-test was conducted to compare the mean scores of third-year (M= 49.86, SD= 18.55) and fourth-year (M = 55.91, SD = 14.44) students on the total C-Test scores. The t-test revealed that there is no significant difference between the two groups (t(97) = -1.81, p =.07). However, the mean scores show that fourth-year students have outperformed third-year students by 6 points. Similarly, an independent sample t-test was conducted to compare the mean scores of females (M = 52.80, SD = 16.26) and males (M = 53.40, SD = 17.99) on total C-Test scores. The t-test revealed that there is no significant difference between the groups (t(98) = -1.70, p = .86). Therefore, male and female students performed equally on the C-Test.

4.1 Rasch Model Analysis

To further examine the validity of the C-Test, a Rasch model was applied to the data. Each passage or text was considered a polytomous item with 21 categories, and the rating scale model was used (Andrich, 1978). The Winsteps computer program was employed to estimate the model (Linacre, 2023a). Table 3 presents the item difficulty parameters, fit statistics, and point-measure correlations for each item. Based on the table, all items have acceptable infit and outfit mean square values, falling within the range of .50 to 1.50 (Linacre, 2023b), and the point-measure correlations are positive and very high. The Rasch model separation reliability was found to be .87, indicating relatively high precision. Principal components analysis of standardized residuals revealed that the strength of the first contrast extracted from the residuals is 1.50, which is smaller than the criterion of 2 set by Linacre (2023b). It shows that the four C-Test passages measure a single latent trait, namely general language proficiency, which indicates that the data are unidimensional.

Table 3
Item Measures and Fit Statistics

	Measure	SE	Infit MNSQ	Outfit MNSQ	Pt-Measure Cor.
Text1	.06	.04	1.33	1.19	.76
Text2	.19	.04	1.32	1.14	.77
Text3	-.35	.05	1.18	.98	.72
Text4	.10	.04	.64	.61	.82

The findings of the differential item analysis for each of the four C-Test passages across the sexes are displayed in Table 4. The DIF measure indicates the difficulty of the items in each sex group, while the 'DIF contrast' illustrates the difference between the two measures. Ideally, the DIF contrasts should be zero, signifying identical item difficulty in both groups and no DIF. One can use a t-test to determine the statistical significance of DIF contrasts. The columns 't' and 'Prob.' display the t-values and their statistical significance. The table indicates that the DIF contrasts are very small and non-significant for the remaining two items, and they are zero for the other two. This suggests that none of the C-Test passages contain gender DIF.

Table 4

Differential Item Functioning Statistics

	DIF measure (M)	DIF measure (F)	Joint SE	DIF contrast	T	Prob.
Text1	.06	.06	.09	.00	.00	1
Text2	.19	.19	.08	.00	.00	1
Text3	-.31	-.37	.10	-.06	-.65	.51
Text4	.04	.13	.09	.08	.96	.33

5. Discussion

The assessment of the performance of Iraqi students through the C-Test and analysis of the extent to which the test is valid to measure language ability. Fourth-year students achieved better proficiency levels than third-year students, and this is due to their superior linguistic skills and analytical abilities in the English language. According to the analysis, the achievements of the test-takers varied according to their performance. The inter-item correlations between passages showed the items discriminate well between the high-ability and low-ability examinees. The t-test showed that there was no significant difference between the two groups. However, the mean scores showed that fourth-year students outperformed better than third-year students by 6 points. To compare the mean scores of males and females on the total test, an independent sample t-test was conducted. The t-test revealed that there is no significant difference between the female and male groups. Therefore, male and female students have performed equally on the C-Test.

The item difficulty parameters fit statistics and point-measure correlations for each item. All the items have acceptable in-fit and outfit mean square values, and the point-measure correlations are positive and very high. The Rasch model separation reliability was relatively high. The principal components analysis of standardized residuals showed that the strength of the first contrast extracted from the residuals is a sign that the data are unidimensional, and the four C-Test passages measure a single latent trait, which is general language proficiency. The results of differential item analysis for the four C-Test passages across sexes measured DIF, which showed the difficulty of the items in each sex group. While ‘DIF contrast’ showed the difference between the two measures. The DIF contrasts were zero, small, and non-significant; therefore, the item has identical difficulty in both groups, and there is no DIF. Numerous scientists have supported and used the C-Test as a broad measure of proficiency in second language acquisition, as Alabdallah et al. (2023) used Polytomous Rasch models to measure the underlying construct of the test and local item dependence. The study showed the adopted models used in their studies were applicable, making understanding Likert questions and language tests easier for developers and testers. A second study was carried out to evaluate the examinees' proficiency in a second language. The study discovered that the two-parameter logistic (2PL) IRT model was the most effective for scoring and calibrating the C-Test (Alpizar et al., 2023). Daller et al. (2021) related the C-Test to memory function, in which a student's processing speed in English as a foreign language is affected by their vocabulary knowledge. Since the C-test measures these language proficiency characteristics correctly, it has strong predictive validity for international student's academic success. The purpose of Drackert and Timukova's (2020) C-test format study was to find out if learners who are not foreign language learners capture language proficiency aspects similarly. The analysis of the error for a few biased items revealed that, despite the learners' apparent advantage in reconstructing the meaning of the gaps, they failed to convert this recognition ability into the production of the correct form. The C-Test is a reliable tool for assessing student competency in this study. It also suggests that

many Iraqi educators have not been trained in test development and lack the skills and knowledge necessary to create assessments. Thus, this test could serve in teaching reading and comprehension of English texts in Iraq, so it should be developed and carried out to improve testers' academic achievements.

6. Conclusion

To sum up, these differences showed that fourth-year students are better than third-year students according to the analysis of their responses, and the reason behind this is due to the superior linguistic skills and analytical analyses of the first group; however, no significant differences have appeared apparently via t-test. Furthermore, there was no gender difference in the performance of the two groups, indicated by an independent sample t-test. As a result, the hypothesis proved that fourth-year students achieved better proficiency levels on the C-Test. The C-Test was used as an assessment to show the level the students reached in the English language and the differences between the abilities of the two groups. This kind of test helps measure the ability of the students to use English words correctly; however, the students faced some difficulties in guessing the most appropriate words and this is due to their weakness in the English language. For examiners, some paragraphs seemed to be more difficult than others, such as paragraphs 2 and 3. Finally, the results indicate that developing and carrying out this test could improve testers' academic achievements in basic foreign language classes; thus, this test should be in teaching reading and comprehension of English texts in Iraq.

The current study's findings have several implications for educators, test administrators, and students. This kind of test has a good impact on the students in basic foreign language lessons, where they may achieve better academic levels by using such assessments, which would enhance their overall achievements despite the different tests employed in educational settings around the world, such as TOEFL and IELTS. The C-Test is widely accepted by scientists as a comprehensive assessment tool for second language acquisition (e.g., Baghaei & Grotjahn, 2014; Connelly, 1997; Fadaeipour & Zohoorian, 2017; Forthmann, 2020; Gogolin et al., 2021; Grujić & Danilović, 2012; Harsch & Hartig, 2016; Hiser & Ho 2016; Hood, 1990; Khodadady, 2007; Khodadady & Ghergloo, 2013; Khodadady, 2014; Khoshdel et al., 2016; Lee-Ellis, 2009; McKay, 2019; Rasoli1, 2021; Yang & Osborne, 2023). The findings also prove that the C-Test is suitable to be used as a kind of assessment to evaluate students' achievements in reading. The study also suggests that many Iraqi teachers lack the skills and knowledge necessary to create assessments and have not been given any test-development training. It is recommended that all language instructors in Iraq make use of the C-Test due to its low cost and ease of development.

7. Implications and Suggestions for Further Studies

As appeared clearly in this research, this test is validated by a C-Test among Iraqi EFL university students. Although the sample for this paper was limited to students at Al-Nisour University in Baghdad, the population it represents is EFL learners in the private sector. To get considerably more credible and trustworthy results, the researchers advise individuals to undertake the same research to choose a sample from the University of Baghdad, Al-Mustansiriyah University, and Al Iraqia University. The second important suggestion is a longitudinal study to monitor students' advancement from lower to higher levels of English language competency over time. The last suggestion, using qualitative techniques like focus groups and interviews to learn more about the particular difficulties students have comprehending and finishing the C-Test to examine the differences in difficulty levels between test passages and items to have a more comprehensive of which language skills should pose. Examining additional tests of decreased redundancy is the other recommendation for additional research.

Declaration of Conflicting Interests: We declare that this manuscript is original, has not been published before, and is not currently being considered for publication elsewhere. We know of no conflicts of interest associated with this publication.

Funding: We declare that there has been no financial support for this work.

References

Alabdallah, Z. A., Ismail, I. A., Mutar, H. K., Ahmed, A., Alghazali, T., Mansoor, M. S., & Georgievna, G. V. (2023). Analysis of C-Tests with the equidistance and the dispersion models. *International Journal of Language Testing, 13* (Special Issue), 142-148. DOI: 10.22034/IJLT.2023.403640.1264

Alderson, J. C. (1979a). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly*, *13*(2), 219–227. DOI: 10.2307/3586211

Alderson, J. C. (1979b). The effect on the doze test of changes in deletion frequency. *Journal of Research in Reading*, *2*(2), 108-119.

Alpizar, D., Li, T., Norris, J. M., & Gu, L. (2023). Psychometric approaches to analyzing C-tests. *Language Testing*, *40*(1), 107-132. DOI: 10.1177/02655322211062138

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

Baghaei, P., & Grotjahn, R. (2014). Establishing the construct validity of conversational C-Tests using a multidimensional Rasch model. *Psychological Test and Assessment Modeling*, *56*(1), 60–82.

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

Cohen, A. D., Segal, M., & Weiss Bar-Siman-Tov, R. (1984). The C-Test in Hebrew. *Language Testing*, *1(*2), 221–225. DOI: 10.1177%2F026553 228400100206.

Coleman, J. A. (1996). A comparative survey of the proficiency and progress of language learners in British universities. In R. Grotjahn (Ed.), *Der C-Test: Theoretische grundlagen und praktische anwendungen* (pp. 367-99). Bochum, Brockmeyer.

Raatz, U., & Klein-Braley, C. (2002). Introduction to language testing and to C-Tests. In J. A. Coleman, R. Grotjahn & U. Raatz (Eds.), *University language testing and the C-test* (pp. 75–91). Bochum: AKS-Verlag.

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

Daller, H., & Phelan, D. (2006). The C-test and TOEIC® as measures of students' progress in intensive short courses in EFL. In R. Grotjahn (Ed.), *Der C-Test: Theorie, empirie, anwendungen/the c-test: theory, empirical research, applications* (pp. 101-119). Peter Lang.

Daller, M., Müller, A., & Wang-Taylor, Y. (2021). The C-Test as predictor of the academic success of international students. *International Journal of Bilingual Education and Bilingualism*, *24*(10), 1502-1511. DOI: 10.1080/13670050.2020.1747975

Drackert, A., & Timukova, A. (2020). What does the analysis of C-test gaps tell us about the construct of a C-test? A comparison of foreign and heritage language learners' performance. *Language Testing*, *37*(1), 107-132. DOI: 10.1177/0265532219861042

Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-Tests. *Language Testing*, *23*(3), 290-325. DOI: 10.1191/0265532206lt330oa

Engelhardt, D. (2013). *Intermediate English reading and comprehension.* McGraw Hill.

Fadaeipour, A., & Zohoorian, Z. (2017). Comparing the psychometric characteristics of speeded and standard C-Tests. *International Journal of Language Testing*, *7*(1), 40-50.

Forthmann, B., Grotjahn, R., Doebler, P., & Baghaei, P. (2020). A comparison of different item response theory models for scaling speeded C-Tests. *Journal of Psychoeducational Assessment*, *38*(6), 692-705.

Gogolin, I., Schnoor, B., & Usanova, I. (2021). Crossing the bridge to literacy in foreign languages: C-Test as a measure of language development. *Multilingua*, *40*(6), 771-790.

Grujić, T., & Danilović, J. (2012, April 27-28). *The value of C-Tests in English language testing and teaching*. Multidisciplinary Conferences Language, Literature, Values, Faculty of Philosophy, University of Niš, Serbia.

Harsch, C., & Hartig, J. (2016). Comparing C-tests and Yes/No vocabulary size tests as predictors of receptive language skills. *Language Testing*, *33*(4), 555-575. DOI: 10.1177/0265532215594642

Heaton, J. B. (1975). *Writing English Language Tests.* Longman.

Hiser, E. A., & Ho, K. S. T. (2016). C-Tests in Vietnam: An exploratory study of English proficiency. *Electronic Journal of Foreign Language Teaching*, *13*(2), 184-202.

Hood, M. A. G. (1990). The C-Test: A viable alternative to the use of the cloze procedure in testing? In L. A. Arena (Ed.), *Language proficiency: Defining, teaching, and testing* (pp. 173-189). Springer.

Hughes, A. (1989). *Testing for language teachers*. Cambridge University Press.

Ikeguchi, C. B. (1998). Do different C-Tests discriminate proficiency levels of EL2 learners? *JALT Testing & Evaluation SIG Newsletter*, *2*(1), 2-10.

Karimi, N. (2011). C-Test and vocabulary knowledge. *Language Testing in Asia*, *1*(4), 7. DOI: 10.1186/2229-0443-1-4-7

Khodadady, E. (2007). C-Tests method specific measures of language proficiency. *Iranian Journal of Applied Linguistics*, *10*(2), 1-26.

Khodadady, E., & Hashemi, M. (2011). Validity and C-Tests: The role of text authenticity. *International Journal of Language Testing*, *1*(1), 30-41.

Khodadady, E., & Ghergloo, E. (2013). S-Tests and C-Tests: Measures of content-based achievement at grade four of high schools. *American Review of Mathematics and Statistics*, *1*(1), 1-16.

Khodadady, E. (2014). Construct validity of C-Tests: A factorial approach. *Journal of Language Teaching and Research*, *5*(6), 1353. DOI:10.4304/jltr.5.6.1353-1362

Khoshdel, F., Baghaei, P., & Bemani, M. (2016). Investigating factors of difficulty in C-Tests: A construct identification approach. *International Journal of Language Testing*, *6*(2), 113-122.

Klein- Braley, C. (1981). *Empirical investigations of cloze tests: An examination of the validity of cloze tests as tests of general language proficiency in English for German university students* [Doctoral dissertation, University of Duisburg].

Klein-Braley, C. (1994). *Language testing with the C-Test: A linguistic and statistical investigation into the strategies used by C-Test takers, and the prediction of C-Test difficulty* [Unpublished post-doctoral thesis]. University of Duisburg

Klein-Braley, C. (1996). Towards a theory of C-Test processing. In R. Grotjahn (Ed.), *Der C-Test: Theoretische grundlagen und praktische anwendungen* (pp. 23-94). Bochum, Brockmeyer.

Klein-Braley, C. (1997). C-Tests in the context of reduced redundancy testing: An appraisal. *Language Testing*, *14*(1), 47-84. DOI: 10.1177/026553229701400104

Klein-Braley, C., & Raatz, U. (1984). A survey of research on the C-Test1. *Language Testing*, *1*(2), 134-146. DOI:10.1177/026553228400100202

Lee-Ellis, S. (2009). The development and validation of a Korean C-Test using Rasch Analysis. *Language Testing*, *26*(2), 245-274. DOI: 10.1177/0265532208101007

Linacre, J. M. (2023b). *Winsteps® Rasch measurement computer program User's Guide* (Version 5.6.0). Portland, Oregon: Winsteps.com.

Linacre, J. M. (2023a). *Winsteps®* (Version 5.6.0) [Computer Software]. Portland, Oregon: Winsteps.com. Available from https://www.winsteps.com/

Linnemann, M., & Wilbert, J. (2010). The C-Test: A valid instrument for screening language skills and reading comprehension of children with learning problems. In R. Grotjahn (Ed.), *C-Test: Contributions from current research*, (pp.113-124). Frankfurt/M.: Lang.

McCowan, R. J., & McCowan, S. C. (1999). *Item analysis for criterion-referenced tests*. Center for Development of Human Services.

McKay, T. (2019). *More on the validity and reliability of C-Test scores: A meta-analysis of C-Test studies*. Georgetown University.

Mehrens, W.A., & Lehmann, I.J. (1973). *Measurement and evaluation in education and psychology*. Rinehart & Winston.

Mozgalina, A., & Ryshina–pankova, M. (2015). Meeting the challenges of curriculum construction and change: Revision and validity evaluation of a placement test. *The Modern Language Journal*, *99*(2), 346-370. DOI: 10.1111/modl.12217

Norris, J. M. (2006). Development and evaluation of a curriculum-based German C-Test for placement purposes. In R. Grotjahn (Ed.), *Der C-Test: theorie, empirie, anwendungen/the C-Test: Theory, empirical research, applications* (pp. 45-83). Frankfurt/M.: Lang.

Oller, J. W. (1983). Evidence for a general language proficiency factor: An expectancy grammar. In J. W. Oller (Ed.), *Issues in language testing research* (pp. 3-10). Rowley, MA; Newbury House

Park, J. (1998). *The C-Test: usefulness for measuring written language ability of non-native speakers of English in high school* [Master's thesis, Iowa State University]. https://core.ac.uk/download/pdf/38925471.pdf

Peppé, S., & McCann, J. (2003). Assessing intonation and prosody in children with atypical language development: The PEPS-C test and the revised version. *Clinical Linguistics & Phonetics, 17*(4-5), 345-354. DOI: 10.1080/0269920031000079994

Raatz, U., & Klein-Braley, C. (1981). The C-Test--A modification of the cloze procedure. In T. Culhane, C. Klein-Braley & D. K. Stevenson (Eds.), *Practice and problems in language testing: Proceedings of the fourth international language testing symposium of the IUS* (pp. 113–138). Peter Lang GmbH.

Rasoli, M. K. (2021). Validation of C-Test among Afghan students of English as a foreign language. *International Journal of Language Testing, 11*(2), 109-121. https://www.ijlt.ir/article_138060.html

Roos, U. (1994). The C-Test in Japanese. In R. Grotjahn (Ed.), *Der C-Test: Theoretische grundlagen und praktische Anwendungen*, (pp. 61–113). Bochum, Brockmeyer

Singleton, D. (1999). *Exploring the second language mental lexicon*. Cambridge University Press.

Singleton, D., & Little, D. (1991). The second language lexicon: Some evidence from university-level learners of French and German. *Second Language Research, 7*(1), 61-81.

Spolsky, B. (1973). What does it mean to know a language, or how do you get someone to perform his competence? In J. W. Oller & Jack C. Richards (Eds.), *Focus on the learner: Pragmatic perspectives for the language teacher*, (pp. 164-76). Newbury House Pub.

Spolsky, B., Sigurd, B., Sato, M., Walker, E., & Arterburn, C. (1968). Preliminary studies in the development of techniques for testing overall second language proficiency. *Language Learning, 18* (3), 79-101. DOI: 10.1111/j.1467-1770.1968.tb00224.x

Stemmer, B. (1991). *What's on a C-Test taker's mind? Mental processes in C-Test taking*. N. Brockmeyer.

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly, 30*(4), 415-433. DOI: 10.1177/107769905303000401

Testbook. (2022, January 8). CTET Paper 2 Maths & Science. Retrieved April 30, 2023, from https://testbook.com/question-answer/which-of-the-following-adaptations-help-camels-sur--62732ec6fe2aeb3ad172ac08.

Yang, X., & Osborne, C. (2023). The development and validation of a C-Test and a pseudo-character test for online CFL assessment. *Journal of China Computer-Assisted Language Learning, 3*(1), 101-131. DOI: 10.1515/jccall-2022-0019

Appendix A

Read the passages carefully and fill in the missing letters.

Passage 1

Camels live in desert where it is hot and dry. Camels ha------ adaptations th------ help th------ live i------ deserts. Th------ have a th------ coat o------ hair th------ protects th------ from t------ sun. Th------ have b------ and so------ feet, s------ they c------ walk a lo------ time i------ the h------ sands. Sev------ adaptations he------ a camel lives in a desert. When there is food and water, a camel can eat and drink large amounts and store it as fat in the hump. Then, when there is no food or water, the camel uses the fat for energy, and the hump becomes small and soft.

Passage 2

Endangered means to be under threat or near extinction. When a spe------ / animal i------ endangered i------ means th------ they a------ disappearing fa------ or ha------ a ve------ small popul------ – not la------ enough t------ survive. Th------ are so------ endangered ani------ in Afr------. We c------ find on------ a f------ of th------. Some exam------ are zebras, pandas, and elephants. Humans must not destroy the natural homes of the animals in the forests. They must not hunt animals and hurt the nature.

Passage 3

For many years people have been trying to create a simple universal language that would serve all over the world as a common means of communication. In t------ last th------ hundred ye------, more th------ seven hun------ such lang------ have be------ suggested. T------ most succe------ and t------ most pop------ of th------ is a lang------ called Esperanto. I------ was inve------ by Ludwig Zamenhof, w------ lived i------ Poland. Wh------ he w------ growing u------, he saw that people from different backgrounds who lived in Poland had lots of difficulties communicating with each other. This often led to disagreements.

Passage 4

In the first hour of a normal night's sleep, you go into a deep sleep. In fa------, this i------ the ti------ your sl------ is dee------. Then la------ in t------ night, t------ mind go------ into a parad------ sleep wh------ means "lig------ sleep". I------ is dur------ this ty------ of sl------ that y------ have yo----- sweet dre------. In a nor------ night, most people go from deep sleep to paradoxical sleep about four or five times. Each period of deep sleep becomes less deep and shorter, and each period of paradoxical sleep becomes longer and lighter.

Key Answers

Passage 1 "Adaptations of Camels to Survive in Deserts"

Camels live in desert, where it is hot and dry. Camels have adaptations that help them live in deserts. They have a thick coat of hair that protects them from the sun. They have broad and soft feet, so they can walk a long time in the hot sands. Several adaptations help a camel lives in a desert. When there is food and water, a camel can eat and drink large amounts and store it as fat in the hump. Then, when there is no food or water, the camel uses the fat for energy, and the hump becomes small and soft.

Passage 2 "Endangered Animals and Challenges in Survival"

Endangered means to be under threat or near extinction. When a species / animal is endangered, it means that they are disappearing fast or have a very small population – not large enough to survive. There are so many endangered animals in Africa. We can find only a few of them. Some examples are zebras, pandas, and elephants. Humans must not destroy the natural homes of the animals in the forests. They must not hunt animals and hurt the nature.

Passage 3 "Historical Attempts at a Universal Language for Global Communication"

For many years people have been trying to create a simple universal language that would serve all over the world as a common means of communication. In the last three hundred years, more than seven hundred such languages have been suggested. The most successful and the most popular of these is a language called Esperanto. It was invented by Ludwig Zamenhof, who lived in Poland. When he was growing up, he saw that people from different backgrounds who lived in Poland had lots of difficulties communicating with each other. This often led to disagreements.

Passage 4 "Understanding Human Sleep: Adaptations and Transitions"

In the first hour of a normal night's sleep, you go into a deep sleep. In fact, this is the time your sleep is deepest. Then later in the night, the mind goes into a paradoxical sleep which means "light sleep". It is during this type of sleep that you have your sweet dreams. In a normal night, most people go from deep sleep to paradoxical sleep about four or five times. Each period of deep sleep becomes less deep and shorter, and each period of paradoxical sleep becomes longer and lighter.

Students' Responses

1. Third-year students' answers pdf. Google Docs. https://drive.google.com/file/d/1JcP-IL7Hqz4ChAyLc0qlHmA9ObJhtqhi/view?usp=sharing.
2. Fourth-year students' answers pdf. Google Docs. https://drive.google.com/file/d/1Gk4sfuG2QmVf0ziAHc1fKC4xaGmJdgWP/view