

## Validation of C-Test among Afghan Students of English as a Foreign Language

Mohammad Kabir Rasoli<sup>1</sup>

Received: 14 March 2021

Accepted: 13 July 2021

### Abstract

This study reports on the results of a research program carried out to validate the C-Test amongst Afghan EFL learners. One hundred advanced English majors were administered two different language tests namely a language proficiency test (composed of listening, reading, and writing) and a C-Test composed of four different texts to measure the overall language ability of the participants. Various analyses were used to examine the validity and reliability of the C-Test. The C-Test correlated highly with the criterion measures. The results of the study confirmed that the C-Test is a reliable and valid test that can be used as a general English language proficiency test among Afghan students of English as a foreign language.

*Keywords:* Afghan Learners; C-Test, Proficiency test; Reduced Redundancy Principle; Reliability; Validation.

### 1. Introduction

In the domain of second language testing, testing language proficiency has always been a major issue regardless of any specific set of instructional materials, examinees educational and linguistic background. Testing specialists have always tested and tried various approaches to invent new procedures for testing language proficiency because they are not satisfied with the existing instruments testing language proficiency. Spolsky (1973) stated that: “In searching for a test of overall proficiency, then, we must try to find some way to get beyond the limitations of testing a sample of surface features and seek rather tap underlying linguistics competence” (p. 175).

To this end, Spolsky has offered two options: The first one is an interview which has been used continuously as a valid measure of language proficiency. But difficulties in administration and establishing the reliability of scoring limited its implementation. The second one is the reduced redundancy principle (RRP). The second option has been well-received by many professionals who have been trying to operationalize this principle through a variety of testing procedures: dictation, the noise test, partial dictation, cloze test, cloze elide, multiple-choice cloze, rational dictation cloze, and the C-Test (Klein-Braley, 1997).

<sup>1</sup> English Department, Islamic Azad University, Mashhad Branch, Mashhad, Iran. Email: [m.kabirrasoli@gmail.com](mailto:m.kabirrasoli@gmail.com)

Naturally, languages are redundant. Based on RRP native speakers of languages can fix the missing parts of broken sentences, or they have the capability of understanding the meaning of a sentence without having whole information. This characteristic of the human verbal communication system decreases the possibility of making errors and permits communication where there is some interference in the communication channel (Spolsky, 1973).

According to Hughes (2003), native speakers of a language can cope well when this redundancy is reduced. For example, they can understand what someone is saying although there are noises in the environment that prevent them from hearing every sound that is made. Likewise, they can infer the meaning of the text of a newspaper that has been left outside in the rain, causing the print to become blurred. Because non-native speakers generally find it more difficult to cope with reduced redundancy, the deliberate reduction of redundancy has been used as a means of estimating foreign language ability. Learners' overall ability has been estimated by measuring how well they can restore a reduced text to its original form. Tests that are constructed under the basis of the reduced redundancy principle (RRP) use authentic materials. Authentic materials are those which are not written for learning purposes. These authentic materials are damaged in some way and then examinees are asked to restore the materials. On the other hand, Klien-Barley (1985) believes that no language test is authentic, if what we mean by authentic is normal language in everyday use: normal language is not produced to be assessed. Therefore, the question of test authenticity cannot be solved directly.

A large number of test developers and researchers (Baghaei, 2008a, 2008b, 2010, 2011, 2014; Eckes, 2006, 2011; Eckes & Baghaei, 2015; Eckes & Grotjahn, 2006; Forthmann, Grotjahn, Doebler, & Baghaei, 2020; Raatz, 1984, 1985) in the field of language testing confirmed that C-Test is a valid, reliable, economical, and easy to construct, but in the Afghan context of education, the utility, application, validity, and reliability of C-Test have not been researched so far. The purpose of this study is to build validity evidence for the C-Test as a measure of general language proficiency among Afghan students of English. English language proficiency tests used in Afghanistan are made with several components like listening, speaking, reading, writing, vocabulary, and grammar; therefore, they are time-consuming both to the examinees and the examiners. Examinees should spend 120 or 150 minutes doing all the items. And finally, it is difficult to score because it consists of a large number of questions.

C-Test is an economic test of overall language proficiency with many advantages. It is not only easy to design but also to score and a variety of texts can be used to construct a complete test. Compared to the cloze test, C-Test is shorter and contains more deletions (items). Based on the stated problem the researcher formulated two general questions. They are as follows:

Q1: Is the C-Test valid for measuring overall language proficiency among Afghan students of English as a foreign language?

Q2: Is the C-Test reliable for measuring overall language proficiency among Afghan learners?

Language teachers do not have the expertise and time to write items for examinations. In fact, developing language tests is a very demanding task for teachers and many of them have not received any training for this. The family of reduced redundancy tests including C-Tests is very easy to develop and does not require much expertise. Thus, they can be very useful for

language teachers (Coniam, 2008). For measuring general language proficiency, Afghans English teachers use teacher-produced tests. According to Gronlund (1985), there are several salient differences between standardized tests and teacher produced test. Standardized tests are much more valid and reliable to the target test-takers as long as they are well constructed and the quality of test items on standardized tests is higher because each item is written by specialists. In contrast, the quality of a teacher-produced test is lower and with more uncertain reliability. Likewise, teacher-produced tests are full of problems since teachers are not skilled at principles of educational measurement (Popham, 2001). If the study shows that C-Test can successfully be used in Afghanistan then it can be very useful in the Afghan English language teaching context as well as test developers, teachers, and learners.

## 2. Review of Literature

C-Tests and cloze tests are pragmatic language tests. Oller (1979) stated that pragmatic language test refers to any “task that causes the learner to process a sequence of elements in language that conforms to the normal contextual constraints of that language, and which requires the learner to relate sequences of linguistic elements via pragmatic mappings to extra-linguistic context” (p. 38).

As the basis of test construction both Cloze Test and C-Test, use authentic materials. Cloze test is known as a pragmatic test because it meets the naturalness criteria for the language test. The learner must operate based on both immediate and strong range contextual constraints to give correct responses. Learners should use data that is inferred from the fact, ideas, events, relationships, states of affairs, social settings, and the like that are pragmatically mapped by the passage's linguistic sequences (Oller, 1979). All the stated information and comments are applied equally to C-Tests.

Spolsky (1979) claimed, “Knowledge of a language necessarily requires the ability to function even when there is reduced redundancy” (P. 13) and C-Test uses this idea. In reduced redundancy tests like C-Test, segments are deleted and test-takers are asked to fix and restore either the original text or a possible text by calling on all he or she knows of the language or about the language. “They assume that as the learner’s control of the language increases he or she will be able to make more successful use of the redundancy provided by natural language, and will thus achieve a higher score on the test” (Klein-Braley, 1985, p. 78).

In second language studies, the C-Test (Raatz & Klien-Barley, 1982) has found its way as an economic, realistic, quick, accurate, and reliable measure of language skills. A variety of scientists have advocated and used C-Test as a general indicator of competence in the study of second languages (see Chapelle, 1994; Coleman, 1994a, 1994b, 1995a, 1995b, 1996a, 1996b, 1996c; Djiwandono, 1998; Hopp, 2006; Krekeler, 2006; Lee-Ellis, 2009; Read & Chapelle, 2001; Ridley & Singleton, 1995; Schmid & Dusseldorp, 2010; Singleton, 1990; Singleton, 1999; Singleton & Little, 1991).

C-Tests often consist of several short passages in which the second half of every second word is deleted (most commonly four to six). In each passage, there are typically 20–25 mutilated words. Each word that the test takers correctly reconstructed is scored one and otherwise zero. C-Tests that have been specifically planned and pre-tested are as a rule,

extremely accurate. In particular, the reliability coefficients often surpass 0.9 for longer C-Tests, such as on DaF (Test Deutsch als Fremdsprache), and often exceed 0.85 also for shorter C-Tests. C-Tests are known to be general proficiency tests and are commonly used in language testing. Since there is only one passage in a cloze test, the chances that certain examinees would be familiar with the passage content are greater compared to a C-Test with 4-6 passages (Baghaei & Grotjahn, 2014a; Baghaei & Grotjahn, 2014b).

As already stated, C-Tests are regarded as general language proficiency tests, and the scores of test-takers on C-Tests are interpreted as measures of their overall foreign language skill. However, as C-Tests do not involve students in oral/aural skills, such an interpretation can only be partly justified (cf. Shohamy, 1982, p. 162; and also Eckes & Grotjahn, 2006, p. 297). Therefore, one might argue that C-Test scores could be better measures of reading/writing skill than listening/speaking ability.

Alderson (2002) stated, in line with this point, that the C-Test is not an appropriate measure of general language proficiency because it does not tap into oral-aural abilities. Concurrent validation studies to a certain degree corroborate this argument (Baghaei, Monshi, & Boori, 2009). For example, a correlation coefficient of 0.47 (corrected for attenuation) between an English C-Test and the listening section of the Iowa State English Placement Test was reported by Chapelle and Abraham (1990). Correlation coefficients of 0.33 and 0.51 were found by Dörnyei and Katona (1992) between an English C-Test and both the listening parts of a university department test and the Test of English for International Communication (TOEIC). A correlation coefficient of 0.43 between the C-Test and an oral interview was also found by the authors. Grotjahn (1992) discovered a correlation coefficient of 0.24 between a French C-Test and the self-ratings of their speaking capacity by students. However, Coleman (1994), for listening comprehension, showed a much higher correlation coefficient of 0.76 and a moderate coefficient of 0.47 for speaking (Cambridge A-level examination). A correlation coefficient of 0.64 between a German C-Test and both the Test DaF (Test Deutsch also Fremdsprache; Test of German as a Foreign Language) speaking and listening subtests were found in Arras, Eckes, and Grotjahn (2002), while Eckes found correlation coefficients of 0.62 and 0.48 between the onDaF C-Test method and the Test DaF listening and speaking subtests. For the English C-Test and the TOEIC listening section, Daller and Phelan (2006) reported a correlation coefficient of 0.45. The findings of correlational studies of C-Tests and speaking and listening tests are also contradictory; however, low correlations are reported more often than high correlations, and the correlation coefficients reported for speaking are lower than those for listening (see Eckes & Grotjahn, 2006 for more detail on correlational studies).

According to Norris (2006), overly technical, bizarre, or infrequent texts should be avoided, as should texts with excessive use of proper nouns. In each text, the first sentence is left unchanged to give the test taker a general understanding of the test's content. Then, beginning from the second word in the second sentence, the second half of every second word is omitted. The second half of the word plus one letter would be omitted if words have an odd number of letters. Acronyms, proper nouns, words with one letter, numbers, and numerically written dates are left unchanged. The words are no longer removed after the 20th or 25th deletion in a text. The last sentence is left intact to permit the text to come to a natural end.

In addition, the C-Test has its own construction rules. These will be as follows:

- The target population should be specified as well as the test format,
- More than enough, sufficient texts should be chosen and then the best ones should be selected,
- They should be brought into the C-Test format after choosing the best texts (rule of two),
- Analyzing the texts' difficulties,
- It should be decided the satisfaction of each text by changing, adding, or removing some damaged words because some are so difficult or easy,
- Then good texts should be combined,
- Item analysis, reliability, and validity of the test would be performed,
- The test should be improved if necessary,
- The final test form should be administered to a target population sample,
- It is important to measure the test norms (Klein-Braley & Raatz, 2002).

### 3. Method

#### 3.1. Participants and Setting

The participants in the present study were 100 undergraduate EFL Afghan learners at Kahkashan-e-Sharq University, Lincoln Learning Center, Professor Sakeena Yaqobi English Language Institute, Nusrat English Language Academy, and Zarifi English Language Academy. Both male (N= 66) and female (N= 34) students participated in this research with the age range of 16 to 40. All the participants in this study were non-native speakers whose first language was Persian. They were assured that their information would be confidential and they were appreciated for their cooperation.

#### 3.2. Instrumentation

To collect the required data for the present study, two instruments were employed. The instruments which were used are a C-Test and PET English Language proficiency test which included listening, reading, and writing skills to measure the overall language ability of the participants.

*3.2.1 C-Test.* For constructing the C-Test, the researcher used four authentic short passages taken from the Listen-In series by (Nunan, 2003) and the Great Writing 2 by *Folse, Muchmore-Vokoun, and Solomon (2010)* with 25 damaged words in each passage. The texts were ordered according to their difficulty. In this C-Test, the first and the last sentences remained without any deletions. The deletion technique was used in this study is that the researcher deleted the second half of every second word beginning from word two in sentence two. The following rules were established for constructing the C-Tests. If the word has an odd number of letters then the larger half is deleted. If a word has only one letter, for example, the words **I** and **a** then this word is ignored in the counting.

*3.2.2 PET Proficiency Test.* The PET (Preliminary English Test) proficiency test consists of 39 questions that measure the language proficiency of the participants. Questions are divided into three different sections. Each section measures one language skill like

listening, reading, and writing. This proficiency test is taken from the British Council website. Cambridge English: Preliminary is also known as the Preliminary English Test (PET) and Preliminary English Test for Schools (PETfS). This exam shows that you can communicate in English in practical, everyday situations. It will give you a good foundation if you want to study for a professional English qualification. The participants had 75 minutes to do the three different sections of the test. They had 30 minutes for the listening section and 45 minutes for reading and writing sections.

Table 1  
*Descriptive Statistics*

|                | Listening | Reading | Writing | C-Tests | Maximum Scores |
|----------------|-----------|---------|---------|---------|----------------|
| Mean           | 7.97      | 5.21    | 2.01    | 48.01   | C-Tests 100    |
| Median         | 7         | 4       | 2       | 48      | Listening 21   |
| Mode           | 6         | 4       | 0.0     | 58      | Reading 15     |
| Std. Deviation | 4.07      | 2.86    | 1.56    | 22.160  | Writing 5      |
| Variance       | 16.59     | 8.20    | 2.45    | 491.08  |                |
| Range          | 17        | 12      | 5       | 93      |                |
| Minimum        | 2         | 1       | 0.0     | 0.0     |                |

#### 4. Results

Table 1 shows the Mean, Median, Mode, Standard Deviation, Variance, Range, Minimum, and Maximum for each of the variables in the study. Since the nature and the number of items in each test are different, the researcher cannot directly compare the means. For scoring the tests, the number of correct replies was counted.

##### 4.2 Checking the Reliability of the Tests

The Cronbach's alpha reliability of the listening test, reading comprehension test, writing test, and C-Test were 0.78, 0.67, 0.66, and 0.91, respectively (Table 3). All the reliability indices for listening, reading, writing, and C-Test are above 0.65 and they are acceptable. According to Farhady, Jafarpur, & Birjandi, (2019):

“The magnitude of reliability can range from zero to one. The reliability of zero, which is the minimum, means that all observed variation is due to error. That is, the test is completely unreliable. On the other hand, the reliability of one (1) indicates that there is no error in measurement and the test is perfectly reliable.” (p. 116).



Table 2  
*The Cronbach's Alpha Reliability*

| Tests     | No. Items | Reliability (Alpha) |
|-----------|-----------|---------------------|
| Listening | 19        | 0.78                |
| Reading   | 15        | 0.67                |
| Writing   | 5         | 0.66                |
| C-Tests   | 4         | 0.91                |

#### 4.3 Correlational analysis

Table 3 depicts the correlations between listening, reading, writing, and the C-Tests. Table 3 shows performance on the C- test is significantly correlated with the listening test ( $r=.607$ ), the reading test ( $r=.574$ ), and the writing ( $r=.627$ ). The C- test is highly correlated with a composite score of listening plus reading comprehension plus writing which is named proficiency test in this study ( $r=.70$ ). This combined score is a more accurate measure of general language proficiency.

Table 3  
*Pearson Correlation Coefficient*

|           | Listening | Reading | Writing | C-Tests |
|-----------|-----------|---------|---------|---------|
| Listening | 1         | .600**  | .595**  | .607**  |
| Reading   |           | 1       | .571**  | .574**  |
| Writing   |           |         | 1       | .627**  |

Note. \*.Correlation is significant at the 0.01 level (2-tailed).

#### 4.4 Factor Analysis

Factor analysis gives information about how each of the individual items or tests does in terms of measuring the intended construct. The researcher has prepared seven tests that start from C-Test passages 1, 2, 3, 4, then listening, reading, and writing parts for data collection. Then there is a component column with different values. These are called component loadings or factor loadings and they tell us how strong the relationship is between the item and the component. These are the Pearson correlations between the tests and the component or factor.

Table 4 shows that C-Test 1 correlates or loads 0.88 on the component which is the highest one. C-Test 2 correlates at 0.84 on the component. C-Test 3 correlated 0.85 with the component. C-Test 4 loads 0.80 on the component. Likewise, listening, reading, and writing tests loadings are 0.76, 0.73, and 0.77 on the component. It is worth mentioning that all of these items correlate or load very highly on the component.

Table 4

*Component Matrix*

|           | Component |
|-----------|-----------|
|           | 1         |
| C-Test_1  | .884      |
| C-Test_2  | .842      |
| C-Test_3  | .855      |
| C-Test_4  | .801      |
| Listening | .762      |
| Reading   | .735      |
| Writing   | .770      |

## 5. Discussion and Conclusion

This study aimed to examine the validity of the C-Test as a measure of overall language proficiency in English as a foreign language among Afghan EFL learners. One hundred advanced students of English were selected using convenience sampling. Four C-Tests passages along with a multiple-choice listening test, a multiple-choice reading test, and indirect writing tests were administered to the participants. Three of the C-Tests were constructed based on written discourse and the other was based on spoken discourse. There was a very strong correlation between the C-Test and a combined score of listening, reading, and writing designated as proficiency. Table 4 depicts the correlations between listening, reading, writing, and C-Tests. Among listening, reading, and writing the C-Test is highly correlated with writing. This strong correlation between C-Test and writing suggests that examiners could use C-Test for measuring writing skill which needs plenty of time to ask learners to write a letter, paragraph, or an essay. The C- test is also highly correlated with the score of listening rather than reading comprehension. This strong correlation between C-Test and listening was also found by Jafarpur (2002), but the correlation between C-Test and reading found by Jafarpur (2002) is much stronger than the correlation found in this study. Similar correlations were obtained by Babaii and Ansary (2001) who compared the results on C-Test and TOEFL. The total scores of the two tests correlated at  $r = 0.88$ , while structure and vocabulary correlated at 0.88 and 0.79, respectively. Finally, Dörnyei and Katona (1992) studied the correlation between the C-Test and the TOEIC. They found a slightly lower correlation with the total score (0.62), as well as correlations with listening 0.51 and reading 0.54.

The correlations between the C-Test and the criterion measures were strong and significant. If we consider listening, reading, and writing as valid measures of the components of language proficiency, the high correlations observed between C-Test and these measures support the validity of C-Test as a measure of foreign language general proficiency. Furthermore, C-Test has a strong correlation with both receptive (listening, reading) and productive skills (writing) which is evidence that the C-Test is also a valid measure of both receptive and productive skills.

The Cronbach's alpha reliability of the listening test, reading comprehension test, writing test, and C-Test were 0.78, 0.67, 0.66, and 0.91, respectively (Table 3). All the



reliability indices for listening, reading, writing, and C-Test are above 0.65 and they are acceptable. As table 3 shows the reliability of all tests (listening, reading, writing, and C-Tests) are more than 0.65 which can be claimed that the tests are reliable.

Table 4 shows the loadings of each variable on the single factor extracted from the data. All the loadings are very high (above .73) which supports the unidimensionality of the data. The fact that all the C-Test passages highly load on a single factor on which all other measures load supports the validity of C-Test as a measure of overall language ability.

To conclude, the results of descriptive statistics, reliability of C-Test plus listening, reading and writing tests, Pearson correlation coefficient, and factor analysis suggest that the C-Test has a satisfactory correlation with general English language proficiency and is a reliable measure. Furthermore, the significant correlation between the C-Test and the proficiency test ( $r=.70$ ) as well as all its sections used for the study seems to prove that the C-Test holds acceptable concurrent validity and measures the same thing as what the test and the C-Test are measuring, i.e. general language proficiency. Finally, the results of construct validity analysis lend further support to the claim that the C-Test is a valid and reliable measure of general language proficiency amongst Afghan EFL learners.

Obtained information from the current study can provide several implications for test administrators, teachers, and students. Although English language proficiency tests like TOFEL and ILETS (a test which includes all language skills as well as its components) is accepted as a valid and reliable test all over the world, still it is not economical, and fast. As for its pedagogical implications, we propose that the C-Test be utilized as general English language proficiency and a placement test in the classroom environment. For C-Test (Raatz & Klien-Barley, 1982) has found its way as an economic, realistic, quick, accurate, and reliable measure of language skills. A variety of scientists have advocated and used C-Test as a general indicator of competence in the study of second languages (see Chapelle, 1994; Coleman, 1994a, 1994b, 1995a, 1995b, 1996a, 1996b, 1996c; Djwandono, 1998; Hopp, 2006; Krekeler, 2006; Lee-Ellis, 2009; Read & Chapelle, 2001; Ridley & Singleton, 1995; Schmid & Dusseldorp, 2010; Singleton, 1990; Singleton, 1999; Singleton & Little, 1991) (cited in Baghaei, 2011). The findings of the present study also revealed that C-Test can replace the general English language proficiency test. Furthermore, C-Test has been used and suggested to be applied for measuring cognitive abilities such as verbal intelligence (Baghaei & Tabatabaee, 2015). Therefore, this is another line of research that can be pursued in the future. The other implication of this study is that a large number of teachers in Afghanistan do not have the knowledge and the expertise in developing tests and they do not receive any training in test development. As C-Test, construction is easy and economical to construct; therefore, it is recommended for all language trainers in Afghanistan to use it.

As it is stated in the entire research, C-Test is validated among Afghan EFL learners. The population of this thesis is Afghan EFL learners but the sample is only selected in the Herat province of Afghanistan. Due to some security problems, the researcher could not travel to other provinces of Afghanistan to select a sample among all EFL learners. Therefore, the researcher suggests those who want to conduct the same research it would be great to select the sample from major provinces of Afghanistan to have much more valid and reliable findings.

The second important suggestion for future research is to apply a larger sample size for validating C-Test among Afghan EFL learners. Since all the Universities, schools, and other educational institutes were closed due to the Corona pandemic virus, the researcher could invite only one hundred participants to this study.

The other suggestion for further study is to investigate other tests of reduced redundancy. The present study focused on only one type of reduced redundancy test which is the C-Test. The researchers can research to find if other reduced redundancy tests (dictation, the noise test, partial dictation, cloze test, cloze elide, multiple-choice cloze, rational cloze, and C-Test (Fadaeipour & Zohoorian, 2017; Klein-Barely, 1997; Yazdinejad & Zeraatpishe, 2019; Zare & Boori, 2018) are valid and reliable among Afghans' students of English as a foreign language.

### **Acknowledgment**

I would like to thank Dr. Purya Baghaei, Islamic Azad University, Mshhad Branch, Iran, for his generous help and invaluable suggestions in the preparation of the paper.

### **References**

- Alderson, J. C. (2002). Testing proficiency and achievement: Principles and practice. In Coleman, J.A., Grotjahn, R., & Raatz, U. (Eds.), *University language testing and the C-Test* (pp. 15-30). Bochum: AKS-Verlag.
- Arras, U. Eckes, T., & Grotjahn, R. (2002). C-Test in Rahmen des "Test Deutsch als Fremdsprache". In R. Grotjahn, (Ed), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen. Vol. 4*, (pp. 175-209). Bochum: AKS-Verlag
- Babaii, E. & Ansari, H. (2001). The C-Test: a valid operationalization of reduced redundancy principle? *System*, 29, 209-219.
- Baghaei, P. (2011). C-Test Construct Validation: A Rasch modeling approach. Saarbrücken: VDM Verlag.
- Baghaei, P. (2010). An investigation of the invariance of Rasch item and person measures in a C- Test. In R. Grotjahn (Ed.). *Der C-Test: Beiträge aus der aktuellen Forschung/The C-Test: Contributions from Current Research* (pp.100-112). Frankfurt/am: Lang.
- Baghaei, P. (2008a). An attempt to fit the Rasch model to a C-Test. *Iranian EFL Journal*, 2, 6-21. Retrieved from <http://www.iranian-efl.com/2008-Editions.php>.
- Baghaei, P. (2008b). The effects of the rhetorical organization of texts on the C-Test construct A Rasch modeling study. *Melbourne Papers in Language Testing*, 3(2), 32-51.
- Baghaei, P. & Grotjahn, R. (2014a). The validity of C-Tests as measures of academic and everyday language proficiency: A multidimensional item response modeling study. In Grotjahn, R. (Ed.), *Der C-Test: Aktuelle Tendenzen/The C-Test: Current trends* (pp. 165-173). Frankfurt am Main: Lang.
- Baghaei, P., & Grotjahn, R. (2014b). Establishing the construct validity of conversational C-Tests using a multidimensional Item Response Model. *Psychological Test and Assessment Modeling*, 56, 60-82.

- Baghaei, P., & Tabatabaee, M. (2015). The C-Test: An integrative measure of crystallized intelligence. *Journal of Intelligence*, 3, 46-58.
- Baghaei, P. (2014). Development and validation of a C-Test in Persian. In R. Grotjahn (Ed.), *Der C-Test: Aktuelle Tendenzen/The C-Test: Current trends* (pp. 299-312). Frankfurt/M.: Lang.
- Baghaei, P., Monshi Toussi, M.T., & Boori, A. A. (2009). An Investigation into the validity of conversational C-Test as a measure of oral abilities. *Iranian EFL Journal*, 4, 94-109.
- Chapelle, C. A. (1994). Are C-Tests valid measures for L2 vocabulary research? *Second Language Research*, 10, 157-187.
- Chapelle, C. A. & Abraham, R. G. (1990). Cloze method: what difference does it make? *Language Testing*, 7 (2), 121-146.
- Coleman, J. A. (1994a). Degrees of proficiency: Assessing the progress and achievement of university language learners. *French Studies Bulletin*, 50, 11-16.
- Coleman, J. A. (1994b). Language teaching, language learning, language testing. University of Portsmouth, Inaugural Lecture, 17, February 1994. C-Test: especially (pp. 32-21.)
- Coleman, J. A. (1995a). *Progress, proficiency, and motivation among British university language learners* (CLS Occasional Paper No. 40). Trinity College Dublin: Centre for Language and Communications Studies.
- Coleman, J. A. (1995b). The evolution of language learner motivation in British universities, with some international comparisons. In Wakely, R. G. (ed.), *Issues and perspectives in language study*. University of Edinburgh/CILT.
- Coleman, J. A. (1996a). A comparative survey of the proficiency and progress of language learners in British universities. In Grotjahn, R. (Ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen*, 3, (pp. 367-399). Bochum: Brockmeyer.
- Coleman, J. A. (1996b). *Studying languages. A survey of British and European students. The proficiency, background, attitudes, and motivations of students of foreign languages in the United Kingdom and Europe*. London: CILT.
- Coleman, J. A. (1996c). The European Language Proficiency Survey: an overview of findings. In Coleman, J. A., Grotjahn, R., & Raatz, U. (Eds.), *University language learning and the C-Test* (pp. 201-206). Bochum, Germany: AKS-Verlag.
- Coniam, D. (2009). Investigating the quality of teacher-product tests for EFL students and the effects of training in the test development principles and practices on improving test quality. *System*, 37, 226-242.
- 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*. Volume 6, (pp. 101-119). Frankfurt/am: Peter Lang.
- Djiwandono, P. (1998). The relationship between EFL learning strategies, degree of extroversion, and oral communication proficiency. Unpublished Ph.D. dissertation. The State University of Malang.
- Dörnyei, Z. & Katona, L. (1992). Validation of the C-Test amongst Hungarian EFL learners. *Language Testing*, 9 (2), 187-206.

- Eckes, T., & Baghaei, P. (2015). Using testlet response theory to examine local dependency in C-Tests. *Applied Measurement in Education*, 28, 85–98.
- Eckes, T. & Grotjahn, R. (2006). A closer look at the construct validity of C-Tests. *Language Testing*, 23(3), 290–325.
- Fadaeipour, A., Zohoorian, Z. (2017). Comparing psychometric characteristics of speed and standard C-Test. *International Journal of Language Testing*, 7, 40–50.
- Farhady, H. Jafarpur, A. Birjandi, P. (2019). *Testing language skill: From theory to practice*. Tehran: SAMT.
- Folse, K. S., Muchmore-Vokoun, A., & Solomon, E. V. (2010). *Great Writing 2: Great Paragraphs* (3<sup>rd</sup> Ed.). Boston: Heinle Cengage Learning.
- Forthmann, B., Grotjahn, R., Doeblner, P., & Baghaei, P. (2020). A comparison of different item response theory models for scaling speeded C-Tests. *Journal of Psychoeducational Assessment*, 38, 692-705.
- Grotjahn, R. (1992). Der C-Test. Einleitende Bemerkungen. In R. Grotjahn (Ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Vol. 1, pp. 1-18). Bochum: Brockmeyer.
- Gronlund N. E. (1985). *Measurement and evaluation in Teaching*. New York: McMillan.
- Hughes, A. (2003). *Testing for language teachers* (2nd Ed.). Cambridge: Cambridge University Press.
- Hopp, H. (2006). Syntactic features and reanalysis in near-native processing. *Second language Research*, 22, 369-397.
- Jafarpur, A. (2002). A comparative study of a C-Test and a cloze test. In Grotjahn, R. (Ed.), *Der C-Test: Theoretische Grundlagen und prakti-sche Anwendungen* (Vol. 4, pp. 31-51). Bochum: AKS-Verlag.
- Klein-Braley, C. (1997). C-Tests in the context of reduced redundancy testing: An appraisal. *Language Testing*, 14(1), 47-84.
- Krekeler, C. (2006). Language for special academic purposes (LSAP) testing: the effect of background knowledge revisited. *Language Testing*, 23(1), 99-130.
- Lee-Ellis, S. (2009). The development and validation of a Korean C-Test using Rasch analysis. *Language Testing*, 26(2), 245-274.
- 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 am Main: Lang.
- Nunan, D. (2003). *Listen In* (Second Edition Ed.). Singapore: Global ELT: Christopher Wenger.
- Oller, J. (1979). *Language Tests at School: A Pragmatic Approach*. London: Longman.
- Popham, W.J., (2001). *The Truth about Testing: An Educator's Call to Action*. Association for Supervision and Curriculum Development, 1703 N. Beauregard St., Alexandria, VA 22311-1714.

- Raatz, U. (1985). Investigating dimensionality of language tests – a new solution to an old problem. In V. Kohonen, H. von Essen, & C. Klein-Braley (Eds.), *Practice and problems in language testing 8*, (pp. 123–136). Tampere: AFinLA.
- Raatz, U. (1984). The factorial validity of C-Tests. In T. Culhane, C. Klein- Braley, & D. K. Stevenson (Eds.), *Practice and problems in language testing 7. Proceedings of the Seventh International Language Testing Symposium of the IUS, Colchester*, (pp. 124-139). Colchester (England): University of Essex, Department of Language and Linguistics.
- Raatz, U. & Klein-Braley, C. (1982). 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 IV* (pp. 113-138). Colchester: University of Essex, Dept. of Language and Linguistics.
- Read, J. & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing*, 18, 1–32.
- Ridley, J. & Singleton, D. (1995). Contrastivity and individual learner contrasts. *Fremdsprachen Lehren und Lernen*, 24, 123-137.
- Singleton, D. (1990). *The TCD Modern Languages Research Project. Objectives, instruments, and preliminary results* (CLCS Occasional Paper No. 26). Trinity College Dublin: Centre for Language and Communication Studies.
- Singleton, D. (1999). *Exploring the second language mental lexicon*. Cambridge: 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.
- Schmid, M. S. & Dusseldorp, E. (2010). Quantitative analyses in a multivariate study of language attrition: the impact of extralinguistic factors. *Second Language Research*, 26(1), 125-160.
- Shohamy, E. (1982). Affective Considerations in Language Testing. *The Modern Language Journal*. 66 (1), 13-17.
- Spolsky, B. (1973). What does it mean to know a language, or how do you get someone to perform his competence? In Oller, J. W. Jr., & Richards, J. C. (Eds.), *Focus on the learner: Pragmatic perspectives for the language teacher*, (pp. 164–76). Rowley, MA: Newbury House.
- Yazdinejad, A., & Zeraatpishe, M. (2019). Investigating the validity of partial dictation as a test of overall language proficiency. *International Journal of Language Testing*, 9, 44-56.
- Zare, S., & Boori, A. A. (2018). Psychometric evaluation of the speeded cloze-elide test as a general test of proficiency in English as a foreign language. *International Journal of Language Testing*, 8, 33-43.