Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 7, No. 1, March 2017
ISSN 2476-5880

Examining the Underlying Components of C-Test Performance Using a Cognitive Process Model

Fahimeh Khoshdel [1]

Received: 02 November 2016 Accepted: 28 February 2017

Abstract

In the current study, the validity of C-Test is investigated using the construct identification approach. Based on construct identification approach, the factors which are deemed to affect item difficulty in C-Test items were identified. To this aim, 11 factors were selected to enter into Linear Logistic Testing Model (LLTM) analysis to reconstruct C-Test item difficulties using the difficulty of the underlying factors. The 11 factors explained only 12% of the variance in item difficulties. Findings revealed that content words, inflections, and the frequency of the mutilated words had the greatest impact on C-Test item difficulty.

Keywords: *C-Test, validation, construct identification, linear logistic test model*

1. Introduction

The C-Test is a variation of the cloze test with the same basic theoretical assumptions. The C-Test consists of four to six authentic texts in which the first and the last sentences remains intact and the deletions start from word two in sentence two where the second half of every second word should be omitted. (Baghaei, 2011a; Raatz & Kelein-Braley, 2002). In the literature 20 to 25 gaps in each passage are suggested (Raatz & Kelein-Braley, 2002), however, Baghaei (2011b, 2011c) demonstrate that C-Test with smaller number of gaps work as well as 25-gap C-Tests.
Raatz and Kelein-Braley (2002) claimed that a C-Test is a kind of reduced redundancy test which is derived from Information Theory. It means a redundant massage includes more information than is necessary for understanding the massage. Hence, when a massage is damaged, the other parts that are intact can help to find what the complete massage is. A proficient user of a language should not have difficulty in reconstructing the damaged messages. Validating C-Tests in different languages has been researchers' concern for several decades (Baghaei, 2014).

Baghaei (2014) developed and validated a Persian C-Test. The result of his study showed that C-Tests can be used as a general language proficiency test for ages 12-14 of Persian speakers. So, it

[1]*English Department, Islamic Azad University, Mashhad Branch, Mashhad, Iran. Email: Fahime_khoshdel@yahoo.com*

can be used as a measure of general language proficiency in Persian as a second or foreign language too.

The validity of C-Test has also been demonstrated by showing its fit to the Rasch model. Fit of data to latent trait model is evidence of existence of a construct underlying the responses and hence validity (Baghaei & Tabatabaee-Yazdi, 2016). In other studies Baghaei (2008a, 2008b) demonstrated that English C-Tests developed for tertiary students fit the Rasch model too.

Moreover, Borgards and Raatz (2002) examined German C-Tests' sensitivity to a construct-irrelevant attribute referred to as coaching effect. In their study, there were control and experimental groups of 43 secondary level students which in pretest experimental group was exposed to 45 minutes of coaching for C-Test taking. The Posttest demonstrated that the mean of both control and experimental groups similarly increased in comparison with the pretest. That is, coaching effect did not influence C-Test scores. Therefore, it was concluded that C-Tests measure general language proficiency and are not affected by practice. Over the years, different kinds of evidence has been provided for validity of C-Test as a measure of foreign and second language proficiency and even crystallized intelligence (Baghaei, Monshi-Toussi, & Boori, 2009; Baghaei & Grotjahn, 2014a; Baghaei & Grotjahn, 2014b; Baghaei & Tabatabaee, 2015; Baghaei, 2010; Eckes & Baghaei, 2015; Schipolowski, Wilhelm, & Schroeders, 2014)

Construct validity is one of the most complicated aspects of test validation. Not only is construct validity based on analyzing the test scores, but it also requires analysis of test performance (Sigott, 2004). One method of investigating validity is through analyzing difficulty of the items in the framework of construct identification or construct representation. "Construct representation is concerned with identifying the theoretical mechanisms that underlie item responses, such as information processes, strategies, and knowledge stores" (Embretson, 1983, p. 179). By identifying the factors which contribute to item difficulty we are, in fact, identifying what the test requires to be answered, hence, we are identifying the construct underlying the test and establishing construct validity.

2. Literature Review

In this study the validity of C-Test is investigated using construct identification approach. This is a follow up to Khoshdel, Baghaei, and Bemani (2016) who investigated the same issue using correlation and regression analysis. In this study the same data is analyzed using modern measurement theory, i.e., the linear logistic test model (Fischer, 1973).

According to the C-Test literature and the researcher's predication, 13 independent variables were considered to exert influence on C-Test item difficulty. Reviewing the literature showed that some researchers worked on several factors that affect C-Test item difficulty including:

- Klein-Braley (1984, 1985) concluded that the average of sentence length in syllabus was the best predicator of scores for English Students and the average number of words was the best for German Students.
- Sigott (1995) investigated that word frequency can affect C-test item difficulty but word class does not have any influence on item difficulty in C-test items.

- Beinborn, Zesch, & Gurevych' (2014) research revealed that micro level and macro level processing affect C-Test item difficulty.

3. Method

3.1 Participants and setting

In the present study, 352 undergraduate EFL students of Islamic Azad University of Mashhad and Neyshabour, Ferdowsi, Khayyam universities, and Binalood Institute of Higher Education were chosen as the participants. Both male (N=108) and female (N=244) students participated in this research with the age range of 20 to 35 (M=20, SD=10.33).

3.2 Instrumentation

A C-Test with four texts was employed as the instrument. Each text had 25 gaps with different general knowledge content. In this C-Test the first and the last sentences remained without any deletions. Beginning at word two, in sentence two, the second half of every second word was deleted (Raatz & Kelein-Braley, 2002). The texts were selected from CAE (Norris & French, 2008) and FCE books (Norris, 2008). Furthermore, online Collin dictionary was used to get the Frequency of each word.

3.3 Procedure

Based on the available literature and researcher's prediction 13 factors were selected. They are as follow:

- the frequency of the mutilated words (Brown, 1989; Sigott, 1995) as indicated by Collin's
- Cobuild Dictionary.
- whether the words are content or function words
- the length of the mutilated words
- the length of the sentence where the gap is (Klein-Braley, 1984)
- the number of propositions in the sentence where the gap is
- the propositional density (of the sentence where the gap is)
- inflections (Beinborn, Zesch, & Gurevych, 2014)

- text difficulty (as measured by Lexile) (www.lexile.com)
- the frequency of the word before the mutilate word
- the frequency of the word after the mutilate word
- text difficulty (p-values of texts) (Beinborn et al, 2014)

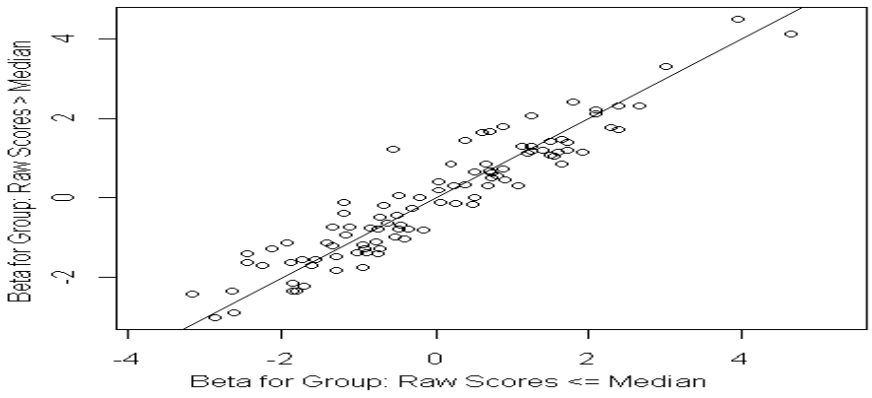
To compute item difficulty, participants had 20 minutes to answer all 100 items (gaps). Item difficulty was computed as the proportion of wrong answers. Linear Logistic Testing Model (LLTM) analysis was employed to predict item (gap) difficulties.

3. Analyses and Results

Khoshdel, Baghaei, and Bemani (2016) using regression concluded that the variables mentioned above could explain only 8% of the variance in C-Test item difficulties. The linear logistic test model (LLTM) (LLTM, Fischer, 1973) was used to ascertain the results of multiple regression in Khoshdel et al. (2016).

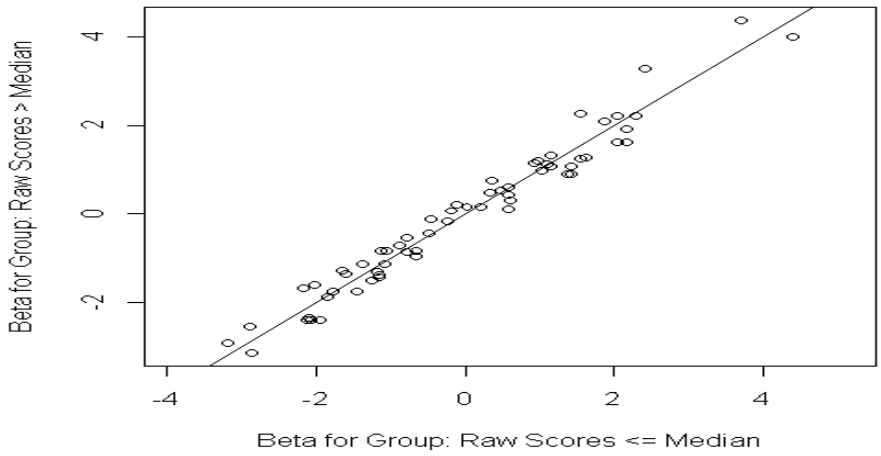
Since LLTM is an extension of the Rasch model, the standard Rasch model (Rasch, 1960/1980) should fit the data first (Fischer, 1973; Baghaei& Kubinger, 2015; Baghaei & Ravand, 2015; Hohensinn & Baghaei, 2017). Andersen's likelihood ratio test (LR test, Andersen, 1973) showed that the 100 items do not fit the Rasch model, χ 2=533.16, df=99, p=00. Graphical model check (Figure 1) revealed that 36 out of 100 items were misfitting.

Figure1.Graphical Model Check for the 100 C-Test Gaps



After deleting the 36 misfitting items which fell far from 45 degree line, Rasch model was estimated again. Andersen's LR test showed that the 64 remaining items fit the Rasch model: $\chi 2 = 86.2$, $df = 63$, $p = 0.028$. Graphical model check showed that the items are close to the 45 degree line (Figure 2).

Figure 2.Graphical Model Check for 64 C-Test Gaps



Each word and text characteristic was considered a cognitive operation and a Q-matrix was constructed. The operations were as follows:

- Frequency of mutilated word
- Function/content words
- Word length
- Sentence length
- Number of proposition
- Propositional density
- Text difficulty (Lexile)
- Inflections
- Frequency of word before the mutilate word
- Frequency of word after the mutilated word
- Text difficulty (P-value or difficulty of each super-item or passage)

For example, if a word was longer than four letters it was considered a long word and 1 was entered into the Q-matrix because it was assumed that processing longer words should be more difficult and if a word is long it need long word processing operation. If a mutilated word was a content word 1 was entered into the Q-matrix and if it was a function word 0 was entered. The assumption was that content words were more difficult to reconstruct than function words and they need a especial cognitive process. For sentences the average sentence length in the four passages, i.e., 20 words, was considered a criterion. Sentences with more than 20 words were considered long and for gaps within these sentences 1 was entered into the Q-matrix, otherwise 0 was inserted.

Based on the results of graphical model check the 36 misftting items were deleted and a Q-Matrix for the 64 remaining items and the 11 basic parameters was developed. The 64 Rasch model fitting items and the Q-matrix were subjected to LLTM analysis using eRm (Mair, Hatzinger, & Maier, 2014) package in R version 3.11 (R Core Development Team, 2015).Table1showstheeasinessparametersofthe11 operations, their standard errors, and their 95% confidence intervals.

Table 1. Easiness of the basic parameters, standard errors and 95% confidence intervals for 11 operations

	Estimate	Std. Error	lower CI	upper CI
1. Frequency of mutilated word	-0.592	0.049	0.688	-0.496
2. Content / Function word	-0.633	0.030	-0.692	-0.573
3. Word length	0.194	0.032	0.133	0.256
4. Sentence length	0.610	0.039	0.534	0.685
5. Number of proposition	-0.113	0.030	-0.172	-0.054
6. Propositional density	-0.131	0.033	-0.196	-0.066
7. Inflection	-0.252	0.039	-0.329	-0.175
8. Text difficulty (Lexile)	-0.332	NaN	NaN	NaN
9. Frequency of word before the mutilate word	-0.592	0.074	-0.738	-0.446
10. Frequency of word after the	0.825	0.107	0.616	1.035

 ISSN 2476-5880

mutilate word				
11.Text difficulty (P-value or difficulty of each super-item or passage)	0.332	NaN	NaN	NaN

LLTM analysis revealed high errors for basic parameters 8, and 11, so they were omitted and LLTM was estimated again with 9 basic parameters. Table 2 shows the difficulty estimates of the 9 basic parameters along with their standard errors and confidence intervals.

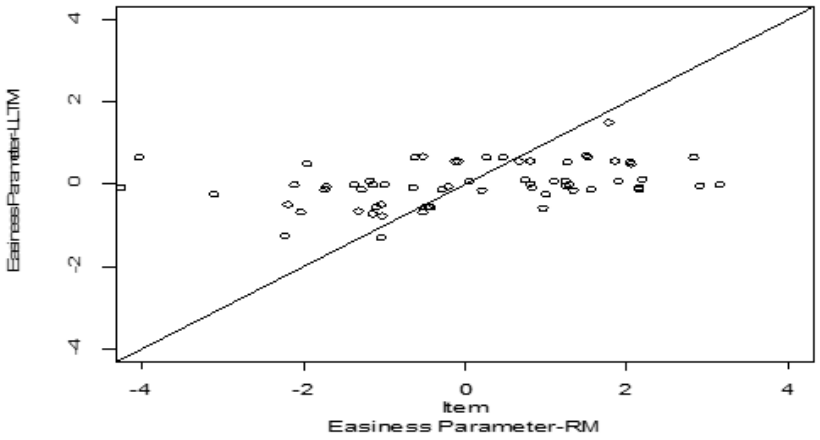
Table 2. Easiness of the basic parameters, standard errors and 95% confidence intervals for 9 operations

	Estimate Std.	Error	lower CI	upper CI
1. Frequency of mutilated word (eta 1)	-0.573	0.049	-0.668	-0.477
2. Content / Function word (eta 2)	-0.689	0.029	-0.747	-0.632
3. Word length (eta 3)	0.118	0.029	0.061	0.176
4. Sentence length (eta 4)	0.652	0.038	0.578	0.727
5. Number of proposition (eta 5)	-0.094	0.030	-0.152	-0.035
6. Propositional density (eta 6)	-0.155	0.033	-0.220	-0.091
7. Inflection (eta 7)	-0.648	0.029	-0.704	-0.592
8. Frequency of word before the mutilated word (eta 8)	-0.566	0.074	-0.711	-0.420
9. Frequency of word after the mutilated word (eta 9)	0.834	0.107	0.625	1.044

Comparingthefit of LLTM and the Raschmodel with the likelihood ratio test showed that the Rasch model fits significantly better than LLTM, χ^2=9856, *df* =54, *p =0.00*.

LLTM imposes a linear constraint on the difficulty parameter. It means that we should be able to reconstruct Rasch model-based item parameters by adding the difficulty of the operations needed to solve each item. The correlation between Rasch model- based items estimates and LLTM-reconstructed item estimates was .37; that is, we managed to explain 12% of the variance in item difficulties with the nine factors.

Figure 3.Rasch Model Item Parameters vs. LLTM Item Parameters



4. Discussion

The purpose of this study was to establish whether 11 independent variables have any significant effects on C-Test item difficulty or not. The study was a reanalysis of Khoshdel, Baghaei, and Bemani (2016) who investigated the factors that contribute to C-Test item difficulty using mainstream statistical methods, i.e., correlation and regression. In this study the same data were analyzed using an item response theory model, namely, the linear logistic test model (Fischer, 1973).

First, Rasch model was used for 100 items to determine whether they fit the model. Results showed that 36 items did not fit. After deletion of these 36 items, for 64 items Rasch model was run again. After developing a Q-Matrix for 64 items and 11 basic operations (dependency and word class were omitted in this analysis because they could not be entered into a Q-matrix), LLTM was run. Two basic operations were deleted due to high standard errors for their parameter estimates and LLTM was rerun. The nine basic parameters as mentioned earlier were: 1. Frequency of mutilated word, 2. Content/function word, 3. Word length, 4. Sentence length, 5. Number of propositions, 6. Propositional density, 7. Inflections, 8. Frequency of word before the mutilated word, 9. Frequency of word after the mutilated word.

The result of this analysis showed that content words, inflections, and the frequency of the mutilated word had the greatest impact on item difficulty. Although, there were some other parameters but they did not have remarkable effect on item difficulty and LLTM explained only 12 % of variance in item difficulties. The regression analysis in Khoshdel et al.'s (2016) study resulted in an $R^2$of 12 and an adjusted R^2 of 8.

Based on Beta weights in multiple regression analysis, word length, function/content, and the frequency of the mutilated word had the highest impact on C-Test item difficulty, respectively (Khoshdel et al. 2016). Whereas, based on the LLTM item easiness parameters in this study, inflections, function/content, and the frequency of the mutilated word had the highest effect on C-test item difficulty, respectively. There is a slight discrepancy between the two analyses. Otherwise, the results of the present study are in line with those of Khoshdel et al. (2016).

References

Alderson, J.C. (1983). The cloze procedure and proficiency in English as a foreign language. In J.W. Oller (Ed.). *Issues in language testing research*, Rowley, Mass., Newbury House,17-205.

Baghaei, P. (2008a). The effects of the rhetorical organization of texts on the C-Test construct: A Rasch modelling study. *Melbourne Papers in Language Testing, 13*: 2, 32-51.

Baghaei, P. (2008b). An attempt to fit the Rasch model to a C-Test. *Iranian EFL Journal, 2,* 6-15.

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.

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/M.: Lang.

Baghaei, P. (2011a). *C-Test construct validation: A Rasch modeling approach*. Saarbrücken: VDM Verlag Dr Müller.

Baghaei, P. (2011b). Do C-Tests with different number of gaps measure the same construct? *Theory and Practice in Language Studies, 1*, 688-693.

Baghaei, P. (2011c). Optimal number of gaps in C-Test passages. *International Education Studies, 4*, 166-171.

Baghaei, P. (2014). Construction and validation of a C-Test in Persian. In R. Grotjahn (Ed.), *Der C-Test: Aktuelle Tendenzen/The C-Test: Current Trends*, 301-314. Frankfurt am Main: Lang.

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 R. Grotjahn (Ed.). *Der C-Test: Aktuelle Tendenzen/The C-Test: Current trends* (pp. 163-171.). Frankfurt/M.: 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., & Kubinger, K. D. (2015). Linear logistic test modeling with R. *Practical Assessment, Research & Evaluation, 20,* 1-11. Retrieved from http://pareonline.net/getvn.asp?v=20&n=1.

Baghaei, P., & Ravand, H. (2015). A cognitive processing model of reading comprehension in English as a foreign language using the linear logistic test model. *Learning and Individual Differences,43,* 100-105. doi:10.1016/j.lindif.2015.09.001

Baghaei, P., & Tabatabaee, M. (2015). The C-Test: An integrative measure of crystallized intelligence. **Journal of Intelligence**, **3**, 46-58. Available: http://www.mdpi.com/2079-3200/3/2/46

Baghaei, P., & Tabatabaee-Yazdi, M. (2016). The logic of latent variable analysis as validity evidenceinpsychological measurement. *The Open Psychology Journal, 9*,168-175.

Beinborn, L., Zesch, T., & Gurevych, I. (2014). Predicting the Difficulty of Language Proficiency Tests.*Transactions of the Association for Computational Linguistics*, 2, 517–529.

Borgards, S. & Raatz, U. (2002). Sind C-Tests trainierbar? In R. Grotjahn (Ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* .Bochum: AKS-Verlag. 157–174.

Brown, J.D. (1989). Cloze item difficulty. *JALT Journal*, 11, 46–67.

Eckes, T., & Baghaei, P. (2015). Using testlet response theory to examine local dependency in C-Tests. *Applied Measurement in Education, 28,*85–98.

Embretson, S. (1983). Construct Validity: Construct Representation Versus Nomothetic Span. *Psychological Bulletin*, 93(1), 179-197.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37(6), 359-374.doi:10.1016/0001-6918(73)90003-6

Hohensinn, C. & Baghaei, P. (2017). Does the position of response options in multiple-choicetests matter? *Psicológica , 38,* 93-109.

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. (1984). Advance Prediction of Difficulty with C-Tests. In Terry Culhane, Christine Klein-Braley, and Douglas K. Stevenson, editors, *Practice and problems in language testing*, 7, 97–112.

Klein-Braley, C. (1985). A cloze-up on the C-Test: a study in the construct validation of authentic tests. *Language Testing,2*(1), 76-104. doi:10.1177/026553228500200108

Mair, P., Hatzinger, R,&Mair, M. J. (2014). eRm: extended Rasch modeling [Computer software]. R package version 0.15-4 .http://CRAN.R-project.org/package=eRm.

Norris, R. (2008). *Ready for FCE: coursebook*. Oxford: Macmillan.

Norris, R., & French, A. (2008). *Ready for CAE: coursebook*. Oxford, UK: Macmillan.

R CORE TEAM. (2012). R: a language and environment for statistical computing [Computer program]. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, URLhttp://www.R-project.org.

Raatz, U., & Kelein-Braley C. (2002). Introduction to the language and the C-Test. *University Language Testing and the C-Test*, 75-86.

Rasch, G. (1960/1980). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research, 1960. (Expanded edition, Chi-cago: The university of Chicago Press, 1980).

Schipolowski, S., Wilhelm, O. & Schroeders, U. (2014). On the nature of crystallized intelligence: The relationship between verbal ability and factual knowledge. *Intelligence*, *46*, 156–168.

Sigott, G. (1995). The C-test: some factors of difficulty. *AAA: Arbeiten aus Anglistik und Amerikanistik*, 43-53.

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