Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 6, No. 2, October 2016
ISSN 2476-5880

Investigating Factors of Difficulty in C-Tests: A Construct Identification Approach

Fahimeh Khoshdel[1], Purya Baghaei[*2] Masoumeh Bemani[3]

Received: 29 May 2016 Accepted: 16 September 2016

Abstract

In this paper we tried to demonstrate the validity of C-Test using construct identification approach. In this approach to construct validation the factors which contribute to item difficulty are identified. The assumption is that the factors which make items difficult are actually the construct underlying the test. For the purposes of this study, 11 item-level and sentence-level factors, deemed to affect item difficulty, were entered into a regression analysis to predict classical item p-values. The 11 factors explained only 8% of the variance in item difficulties. This finding shows that lexical and sentential factors explain only a very small portion of the variance in p-values. It seems that a great amount of variation in item difficulties should be attributed to above-sentence and text level factors. The implications of the study for C-Test construct validity are discussed.

Keywords: *C-Test, validation, construct identification*

1. Introduction

C-Test is a variation of the cloze test and thus has the same basic theoretical assumptions as the cloze test (Grotjahn, Klein-Braley, & Raatz, 2002). The difference is that in C- Test parts of words are omitted while in cloze tests whole words are deleted. The C-Test is based on the reduced redundancy principle (Spolsky, 1969), i.e., the assumption that natural languages are redundant, so advanced learners can be distinguished from beginners by their ability to deal with reduced redundancy (Beinborn, Zesch, & Gurevych, 2014).

Validating C-Tests has been researchers' concern for several decades of research on C-Tests (Baghaei, 2014). C-Tests have been developed and validated for different groups of learners whether L1 learners, L2 learners or foreign language learners. There is ample convincing evidence for the validity of C-Tests as measures of general language proficiency. For example, it is found that C-Tests have a high correlation with other language tests such as teacher ratings and students self-assessment or with composite scores of various language skills.

[1]*English Department, Islamic Azad University, Mashhad Branch, Mashhad, Iran*
2English Department, Islamic Azad University, Mashhad Branch, Mashhad, Iran
** Corresponding author, Email: baghaei@mshdiau.ac.ir*
[3]*English Department, Islamic Azad University, Mashhad Branch, Mashhad, Iran*

Factorial structure and its fit to the Rasch model is another evidence of C-Test validity (Baghaei, 2008a, 2008b; Eckes, 2006, 2011; Eckes & Grotjahn, 2006; Raatz, 1984, 1985).

One approach to test validation is 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). When a person scores higher than another one, it indicates that he/she possesses more of the construct in question or an item that score higher in difficulty presumably demands more in construct (Stenner, Smith, & Burdick, 1983). Based upon some research such as Klein-Braley (1996), focus of construct identification is twofold: first is investigating C-Test takers' psycholinguistic strategies; second is predicting the difficulty of C-Test passages from text characteristics.

Construct identification is concerned with factors that are involved in the test content that contribute to item difficulty (Sigott, 2004). So, it reveals the validity of the test by examining the characteristics which affect test difficulty. Factors which play a significant role in making the items difficult are in fact the test construct.

This kind of validation is based on information about test, subtest and difficulty of each item. It looks as if the difficulty of items depends on various linguistic features as mentioned above. Identifying the factors that contribute to item difficulty broadens our understanding of the construct underling the C-Test. The factors which make items difficult, in fact, constitute the test construct. By pinpointing factors of difficulty in C-Tests, we explicate construct validity at the item level.

Klein-Braley (1984, 1985) used multiple regression to predict passage difficulty in German C-Tests for 9 and 11 year old L1 German speakers and English C-Tests for L1 German-speaking English students at Duisburg University. She used the following text characteristics as predictors:

(1) number of words in text, number of different type of words,
(2) number of sentences in the text,
(3) type token ratio,
(4) average sentence length in syllabus,
(5) average number of word in sentence and average number of syllables in word.

The type-token ratio and the average sentence length in syllables were the best predictors of scores for English students. For German students, the type token ratio and the average number of words in the sentences were the best predictors. These results are more relevant to predict the difficulty of C-Test passages for special groups and may not be generalizable for other groups. In this study we will focus on using the theory of construct identification to predict factors that influence item difficulty of C-Tests.

2. Method

2.1 Participants and setting

The participants in the present study were 352 undergraduate EFL students at Islamic Azad University of Mashhad and Neyshabour, Ferdowsi, Khayyam, and Binalood universities. 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). They were assured that their information would be confidential and they were appreciated for their cooperation.

2.2 Instrumentation

The instrument employed in this study was a C-Test with four texts. 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).

2.3 Procedure

C-Tests, like any other tests, consist of several items with different item difficulties. The purpose of construct identification is to cast light on the factors which make items easier or more difficult. To this end, various factors that might affect the difficulty of individual gaps were taken into account. Specifically the following factors were examined in this research:

(1) the frequency of the mutilated words (Brown, 1989; Sigott, 1995) as indicated by Collin's Cobuild Dictionary.
(2) whether the words are content or function words
(3) the length of the mutilated words
(4) the length of the sentence where the gap is (Klein-Braley, 1984)
(5) the number of propositions in the sentence where the gap is
(6) the propositional density (of the sentence where the gap is)
(7) inflections (Beinborn et al, 2014)
(8) text difficulty (as measured by Lexile) (www.lexile.com)
(9) the frequency of the word before the mutilate word
(10) the frequency of the word after the mutilate word
(11) text difficulty (p-values of texts) (Beinborn et al, 2014)
(12) dependency among items (Beinborn et al, 2014)
(13) word class (noun, verb, adjective, adverb, pronoun, preposition, conjunction, and determiner) (Sigott, 1995).

Correlation and multiple regression was employed to predict item (gap) difficulties, i.e., p-values or raw incorrect proportions, using the above factors.

3. Analyses and Results

Table 1 displays minimum, maximum, means, and standard deviations of the 12 independent variables that are chosen as predictors of C-Test item difficulty in the present study. Due to the categorical nature of the 'word class' it was not included in the descriptive statistics.

ISSN 2476-5880

Table 1. *Descriptive statistics for the predictors in the C-Test*

	N	Minimum	Maximum	Mean	Std. Deviation
1.Frequency	100	3	5	4.61	.601
2.F.C	100	0	1	.56	.499
3.L.Word	100	2	10	4.93	2.114
4.L.Sentece	100	5	74	29.48	23.808
5.Proposition	100	1	10	3.87	3.368
6.P.Density	100	.06	.50	.1346	.07455
7.Inflection	100	0	1	.18	.386
8.Lexile	100	700	1170	980.00	183.264
9.Dependancy	100	0	1	.20	.402
10.Frequency before mutilated word	100	2	5	4.71	.556
11.Frequency after mutilated word	100	3.00	5.00	4.8800	.38350
12.p value	100	9	13	11.25	1.486

The correlation coefficient between all the independent variables and the dependent variable (item difficulty) was calculated. Item difficulty was computed as proportion wrong so that higher values indicates more difficult items. Table 2 displays all coefficients of correlation between all the variables in this study. As it is shown in table below there are only three significant correlations:

Frequency of the mutilated word and item difficulty, r= -.24, n=100, p< .05.

Function/ content words and item difficulty, r= .21, n=100, p<.05.

Text difficulty as measured by super-item (passage) p-value and item difficulty, r= -.24, n=100, p< .05.

The above findings indicate that as a mutilated word becomes more frequent in the language its reconstruction becomes easier. In the data coding the content words were coded 1 and function words were coded 0. The positive correlation between 'function/content' and item difficulty shows that content words are harder to reconstruct than function words. Passage difficulty had a negative correlation with item difficulty. That is, as p-value increases (easier text) item difficulty decreases (items become easier).

 ISSN 2476-5880

Table 2. *Correlations between all the variables*

	1	2	3	4	5	6	7	8	9	10	11	12	13
1.Difficulty	-	**-.248***	**.216***	.013	-.012	-.054	-.024	.130	.108	-.050	.052	**-.246***	.106
2.Frequency		-	-.573	-.524	-.056	-.022	.045	-.387	-.024	-.127	-.122	-.005	.040
3.F.C			-	.521	.060	-.004	-.131	.411	.016	.335	-.008	-.079	-.073
4.L.Word				-	-.115	-.141	-.073	.648	-.116	.210	.024	.058	-.070
5.L.Sentece					-	.937	-.108	-.234	.567	.114	.184	-.157	-.102
6.Proposition							.089	-.257	.489	.152	.052	-.151	-.133
7.P.Density							-	-.081	.073	.123	.053	-.083	-.118
8.Inflection								-	-.285	.197	-.123	.030	-.044
9.lexile									-	.133	.175	-.569	-.107
10.frequency before mutilated word										-	-.021	-.060	-.059
11.frequency after mutilated word											-	.028	.095
12. p value												-	-.85
13.Dependancy													-

Standard multiple regression was used to estimate the contribution of the 11 independent variables in explaining C-Test item difficulty (number of propositions was deleted because of high correlation with sentence length, r=.94). The assumptions of multicollinearity and independence of residuals were first checked. The independent variables all together explained 8% of variance in item difficulties which was not statistically significant, (F (11, 86) = 1.79, p =.06, R2 = .18, R2 Adjusted = .08). Table 3 shows the Beta weights for the independent variables, their statistical significance and part correlations.

Table 3. *Multiple Regression*

Independent variable	Beta	T	P	Part correlation
		.300	.765	
Inflection	.187	1.340	.184	.130
Frequency	**-.216**	-1.671	.098	**-.162**
F.C	**.227**	1.674	.098	**.163**
Length of Word	**-.262**	-1.800	.075	**-.175**
Length of Sentence	-.094	-.753	.454	-.073
Propositional Density	.014	.137	.892	.013

 ISSN 2476-5880

Frequency before mutilated word	-.144	-1.333	.186	-.130
Frequency after mutilated word	.033	.318	.752	.031
Text difficulty (super-item p-value)	.122	.998	.321	.097
Dependency	.115	1.130	.262	.110
Text difficulty (Lexile)	.142	.953	.343	.093

As the table shows word length (number of letters in a word) has the strongest contribution to item difficulty. Next is function/content and third comes word frequency. Also, the useful piece of information that is displayed in Table 3 is part correlation. The square of part correlation tells how much of the total variance in the dependent variable is explained by each variable.

In this study, frequency of words, function/content word, the word length have part correlations of .16, .16, and -.17, respectively. If we square them we get .025, .025, and .030, indicating that frequency and content/function word explain 2.5 percent of item difficulty and length of word explains 3 percent of the variance of item difficulty. However, these three factors explain a very small portion of variance in item difficulties.

Next, one-way analysis of variance (ANOVA) was run to compare the difficulty of different word classes. As Table 4 displays, in this analysis there is one independent variable (word class) with eight levels.

Table 4. *Mean Item Difficulties of Different Word Classes*

	N	Mean
Noun	21	.5138
Verb	30	**.6135**
Adj	12	**.6139**
Adv	8	.5483
Pronoun	6	.4425
proposition	8	.4832
conjunction	10	.5170
determiner	5	**.3707**
Total	100	.5449

Table 4 shows 'verbs' and 'adjectives' are more difficult to answer in C-Tests and 'determiners' are easier. One-way ANOVA showed that there was not a statistically significant difference at the level of $p<.05$ in word class for the eight word classes: $F(7, 92) = 1.21$, $p=.30$. It can be concluded that word classes do not affect item difficulty in C-Test items.

4. Discussion

As mentioned before, in this study the researcher hypothesized that 13 factors contribute to C-Test item difficulty. Eleven of these factors were entered into regression analysis as independent variables to predict C-Test item difficulties. In this study 352 students from several universities in Mashhad and Neyshabour were selected to answer the 100 items of a four-text C-Test battery.

The study showed that the frequency of the mutilated words had a significant relationship with item difficulty. That is, if the mutilated word has a high frequency, it will help test takers to answer it easier than a low frequency word. For instance, the mutilated word 'sch------' (school) with a frequency of 5 is easier to reconstruct than 'instr-----' (instructed) with a frequency of 3. Therefore, it was concluded that word frequency affects item difficulty.

Moreover, whether the mutilated words are function or content words can affect item difficulty. If the mutilated word is a content word, it is harder to answer. For example, 'student' as a content word was more difficult to reconstruct than 'the' as a function word. In addition, there was a significant correlation between text difficulty as measured by p-values with item difficulty in C-Test items. Text difficulty as p-value is based on the difficulty of individual C-Test items within a text. Thus, it is not surprising that it has a significant correlation with item difficulty. However, the correlation between item difficulty and passage difficulty as measured by Lexile, which is an independent measure of text difficulty, was very low and not statistically significant ($r=.10$, $n=100$, $p=.28$). Finally, analyzing eight word classes illustrated that 'verbs' and 'adjectives' were more difficult to answer in C-Tests and 'determiners' were easier.

The findings of this study revealed that the 11 factors that we selected only explained a small portion of the variance in C-Test item difficulties. Some of these factors were already in the literature and some were added by the researcher. The researcher included all the possible factors which deemed to affect C-Test item difficulty. No construct identification study on C-Test has so far covered as many factors as included in this study. Nevertheless, the portion of the variance explained, i.e., 8% is extremely small considering the number of factors that were entered into the analysis.

The small portion of the variance explained indicate that word level and sentence level factors have a very small impact on item difficulties. In other words, there must be other beyond sentence and text level factors which have a significantly greater impact on item difficulties than lower level word and sentential factors.

Note that one reason for the observed findings is that test takers may use different skills and strategies to answer C-Test items. Therefore, explaining item difficulties with one set of factors for all the test takers is not possible. According to Sigott (2004), C-Tests have a fluid construct. He argued that the construct underlying the C-Test changes as a function of person ability and text difficulty. That is, a C-Test could measure different things for different examinees. Thus, different levels of proficiency require different interpretations for C-Test scores because "the same C-Test passage could well be different tests for subjects at different levels of proficiency…without [the test user] knowing to what extent different aspects of the

ISSN 2476-5880

construct are reflected in the individual test scores" (Sigott, 2004, p.203). If the fluid construct phenomenon is true then it is very difficult to understand what factors make C-Test items hard. Consequently, while answering the C-Test items different factors may influence the difficulty of each item and it would be hard to find out the exact reason why an item becomes easy or hard. Nevertheless, researchers in future must think of other additional relevant factors that might contribute to item difficulty.

Another issue that must be given attention is that correlation is sensitive to restrictions of range. That is, when the range of the measured variables is small, the correlation coefficients are depressed. Our analysis suffered from this problem. Almost all of our independent variables such as word frequency, content/function, etc. suffered from range restrictions. Frequency was measured on scale from 1 to 5 and content/function was dichotomous with only two values, 1 and 0. Therefore, the small correlations we observed in this study are partly due to the small range of the measured variables.

The findings of this study may have some hints and implications for the other researchers. In the present study, the effect of 13 independent variables on item difficulty in C-Test items were investigated. Findings showed that C-Tests are basically measures of vocabulary as the only factors which affected item difficulties were vocabulary factors. However, this does not mean that C-Tests measure only vocabulary and nothing else. Because the independent variables explained only 8% of the variance in item difficulties and were are left with 92% unexplained variance. What we can conclude from this study is that C-Tests do measure vocabulary at the gap level. What accounts for the remaining 92% variance is an open question.

In material and test development, it is crucial to know exactly which factors make an item easier or more difficult. In fact, what makes a test or task hard can guide teachers and material developers for ideal use of the tasks. The results of the current research showed that a C-Test can be used to test knowledge of vocabulary So, C-Tests can be used as vocabulary tests at schools for different levels and as vocabulary tasks for class activities.

Future studies should deal with the effect of paragraphs and text characteristics on C-Test item difficulty because in the present study the focus was on the gap-level. The 11 independent variables all together explained 8% of variance in item difficulties. For this reason, more research is needed to identify other gap and test level factors that might affect item difficulties.

References

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-journal.com/2008-Editions.php.

Baghaei, P. (2008b). 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. Retrieved from http://www.ltrc.unimelb.edu.au/mplt/index.html.

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.

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

Brown, J.D. (1989). Cloze item difficulty. *Journal of the Japan Association of Language Teachers.* 11(1), 46–67.

Eckes, T. (2006). Rasch-Modelle zur C-Test-Skalierung. In Rüdiger Grotjahn (Ed.), *Der C-Test: Theorie, Empirie, Anwendungen/The C-Test: theory, empirical research, applications* (pp.1-44). Frankfurt am Main: Lang.

Eckes, T. (2011). Item banking for C-tests: A polytomous Rasch modeling approach. *Psychological Test and Assessment Modeling*, 53(4), 414–439. Retrieved from http://www.psychologie-aktuell.com/index.php?id=200.

Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing*, 23(3) , 290–325.

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

Grotjahn, R., Klein-Braley, C., & Raatz, U. (2002). C-Tests: An overview. In James A. Coleman, Rüdiger Grotjahn & Ulrich Raatz (Eds.), *University language testing and the C-test* (pp. 93-114). Bochum: AKS-Verlag.

Klein-Braley, C. (1984). Advance prediction of difficulty with C-Tests. In Terry Culhane, Christine Klein-Braley & Douglas K. Stevenson (Eds.), *Practice and problems in language testing 7. Proceedings of the Seventh International Language Testing Symposium of the IUS, Colchester, October 1983* (pp. 97-112). Colchester: University of Essex, Dept. of Language and Linguistics.

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.

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

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

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

Raatz, U. (1984). The factorial validity of C-Tests. In Terry Culhane, Christine Klein- Braley & Douglas K. Stevenson (Eds.), *Practice and problems in language testing 7. Proceedings of the Seventh International Language Testing Symposium of the IUS, Colchester, October 1983*, (pp.124–139). Colchester: University of Essex, Department of Language and Linguistics.

Raatz, U. (1985). Investigating dimensionality of language tests – a new solution to an old problem. In Viljo Kohonen, Hilkka von Essen & Christine Klein-Braley (Eds.), *Practice and problems in language testing 8.* (pp.123-136) .Tampere: AFinLA.

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

Sigott, G. (1995). The C-test: some factors of difficulty. *Arbeiten aus Anglistik und Amerikanistik*, 20(1), 43–54.

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

Spolsky, B. (1969). Reduced Redundancy as a Language Testing Tool. In G.E. Perren and J.L.M. Trim (Eds.), *Applications of linguistics* (pp. 383-390), Cambridge University Press.

Stenner, A.J., Smith, M., & Burdick, D.S. (1983). Toward a theory of construct definition. *Journal of Educational Measurement* , 20(4), 305-316.