Optimal Number of Gaps in C-Test Passages

Purya Baghaei

English Department, Islamic Azad University, Mashhad Branch

91886- Mashhad, Iran

E-mail: pbaghaei@mshdiau.ac.ir

The research is financed by Alexander von Humboldt Foundation, Germany.

Abstract

This study addresses the issue of the optimal number of gaps in C-Test passages. An English C-Test battery containing four passages each having 40 blanks was given to 104 undergraduate students of English. The data were entered into SPSS spreadsheet. Out of the complete data with 160 blanks seven additional datasets were constructed. In the first dataset the scores on the first five gaps in each passage were aggregated and the rest of the gaps were ignored, as if each passage had only five gaps. In the second dataset the scores on the first ten gaps were aggregated. In each subsequent dataset five more gas were added. The eight datasets were analyzed and their psychometric properties were compared. The results showed that as the number of gaps in each passage increases item discrimination, reliability and factorial validity of the test increase accordingly. The implications for C-Test application are discussed.

Keywords: C-Test, Reliability, Item discrimination, Factorial validity

1. Introduction

In studies in second language, researchers usually need to measure the general language proficiency of learners and use it as a moderator or control variable. The assessment of language proficiency is only a side activity and is done along with the testing of other variables which are selected for the study. This is usually a cumbersome process as the testing time and resources at the researchers' disposal are very limited.

C-Test (Raatz & Klien-Barley, 1982) as an economical, practical, quick, valid and reliable measure of language proficiency has found its way in second language research. A number of researchers have advocated and used C-Test as a general proficiency measure in second language research (see Chapelle, 1994; Coleman, 1994a, 1994b, 1995a, 1995b, 1996a, 1996b, 1996c; Djiwandono, 1998; Hopp, 2006; Krekeler, 2006; Lee-Elis, 2009; Read & Chapelle, 2001; Ridley & Singleton, 1995; Schmid & Dusseldorp, 2010; Singleton, 1990; Singleton, 1999; Singleton & Little, 1991).

Considering its ease of application and scoring and the short time needed to administer it, C-Test can be a valuable proficiency measure in second language research. Besides, three decades of research on C-Test has demonstrated its validity and reliability as a measure of first and second language proficiency (see Dörnyei & Katona, 1992; Eckes & Grotjahn, 2006; Grotjahn, 1992, 1994, 1996, 2002, 2006, 2010; Sigott, 2004).

Apart from its application in research in second language, C-Test has widely been used in large scale testing contexts for selection and placement purposes in Germany (see Grotjahn, Klein-Braley & Raatz, 2002).

Optimal number of response categories

In psychology literature there is ample research on the optimal number of response categories in Likert-type scales. Numerous studies have shown that the psychometric properties of scales are affected by the number of response categories (see Bending 1954a/1954b; Cicchetti, Showalter & Tyrer, 1985; Cox, 1980; Dolan, 1994; Ferrando, 2000; Finn, 1972; Hofmans, Theuns & Mairesse, 2007; Jenkings & Taber, 1977; Lozano, Garcia-Cueto & Muniz, 2008; McKelvie, 1978; Neumann, 1979; Nunnaly, 1970; Preston & Colman, 2000; Ramsay, 1973; Weng, 2004). These studies have attempted to find the optimal number of response options by considering the reliability and validity of scales as criteria.

C-Test is a general language proficiency test that is comprised of 4-6 short independent passages. Stating from the second word in the second sentence the second half of every second word is deleted. To avoid problems of *local item dependence* in statistical analyses of C-Tests, each passage is considered a polytomous or super-item. The total scores on each passage is entered into analysis as if each passage is an independent Likert item (Baghaei, 2007, 2010; Eckes & Grotjahn, 2006; Sigott, 2004).

C-Test proponents, Raatz and Klein-Braley (1985, 2002), suggested either 20 or 25 blanks in each C-Test passage, but never provided a psychometric grounding for their suggestion. There is no research in C-Test literature that shows the optimal number of gaps in C-Test passages from a psychometric point of view. The purpose of the present

study is to systematically investigate and monitor the psychometric characteristics of a C-Test with different number of gaps in each passage. The characteristics which are considered here are interval consistency reliability, item discrimination and factorial validity.

2. Method

2.1 Instruments, participants and procedure

For the purpose of this study a C-Test battery comprising four passages, each passage containing 40 blanks was constructed. The test was given to 104 Iranian undergraduate EFL students in two universities. As was mentioned before in statistical analyses of C-Tests the data for the single gaps are not entered into the analysis. That is, the gaps are not considered as items; each passage is considered a super-item or testlet. A C-Test battery that contains say, four short passages in fact has four polytomously scored items. The number of gaps which have been correctly filled in each passage are counted and that is the score for that super-item. This is done in order to avoid local item dependence problem in C-Tests. Under classical test theory and Item Response Theory the items should be locally independent from each other. This means that a correct or wrong reply to an item should not lead to a correct or wrong reply to another item. We know that this assumption is violated in C-Tests and cloze tests if we treat each single gap as an independent item. This is the reason why the number of correct gaps in each passage is aggregated and passage total scores are entered into analysis. In other words, C-Test passages are analysed as Likert items which can have as many as 20-25 response options.

The data were entered into SPSS spreadsheet for analysis. Out of the complete data with 160 blanks seven other datasets were constructed. In the first dataset the scores on the first five gaps in each passage were aggregated and the rest of the gaps were ignored, as if each passage had only five gaps. In the second dataset the scores on the first ten gaps were aggregated. In each subsequent dataset five more gas were added. This resulted in eight datasets, Dataset 1 to 8, with five, ten, fifteen, twenty, twenty-five, thirty, thirty-five and forty gaps in each C-Test passage respectively.

Along with the C-Test a reading comprehension test comprising 14 multiple choice items was also administered to the sample. The test comprised two passages and was taken from CAE practice tests. The Cronbach's alpha reliability of the reading tests was .57 which is due to the small number of items. The datasets were analysed separately, considering each passage as a polytomous item, and their psychometric properties were compared.

Results

In each dataset classical item facilities or p-values and item discrimination indices for passages or super-items were calculated. Test level comparisons were made by calculating Cronbach's alpha reliability, mean of the sample on each test form, standard deviations, correlation with reading and factorial validity, i.e., the percentage of variance explained by the first factor. Tables 1 and 2 show the results.

Table 1 about here

Table 1 shows that as the number of gaps in passages increases the discriminations of the super-items (passages) increase accordingly. However, the p-values of the passages do not show a pattern which is reasonable. There is no reason why the difficulty indices of the first five gaps should be greater or smaller than the difficulty indices of the first ten gaps, unless we assume that test-takers get fatigued and lose concentration when they get to the gaps at the end of the passages or they become familiar with the test as they work through the texts and this makes the subsequent items easier for them. The fact that a falling or rising pattern in p-values is not observed means that forty gaps and four passages are well within the concentration span of the test-takers and also three is no learning effect to influence item difficulty indices.

Table 2 about here

Table 2 shows the test-level statistics for test datasets with different number of gaps in each passage. It is evident from the table that Cronbach's alpha reliability increases as the number of gaps in each passage increase. In order to make the means and standard deviations of the sample comparable across datasets, the percentage of correct replies for each examinee was computed rather than the sum scores. The means and standard deviations of these percent corrects were then computed and compared. As the table shows, there is no pattern in the means and standard deviations. Observing no pattern in means is parallel with previous finding of no pattern in the p-values. This also indicates that the test-takers do not get tired as they work through the passages.

The eight datasets were analyzed with principal component analysis (PCA). Before performing PCA the factorability of the data was checked for all the datasets. The Kaiser-Meyer-Oklin value for the dataset with five gaps in each passage was .69 which exceeds the recommended value of .60 (Kaiser, 1974). The value increases as

the number of gaps increase. Bartlett's Test of Sphericity was significant for all eight datasets ($p<.001$). These suggest the suitability of the data for factor analysis.

Principal component analysis resulted in one factor solutions for all datasets. The fifth column "% Variance" shows that the variance first factor explains increases as the number of gaps increases. The varimax rotation does not change the percent of variance explained by the first factor in the datasets.

As the last column shows the correlation between C-Tests with varying number of gaps and the reading comprehension test does not show a specific pattern and does not increase with the number of gaps. In fact, the highest correlation is observed with 15 gaps. This indicates that the criterion-related validity of C-Test is somewhat independent of the number of gaps in passages and the preciseness of predictions does not depend on the length of the passages.

Discussion and Conclusions

The results of the present study clearly show that as the number of gaps in C-Test passages increases the reliability and factorial validity increase accordingly up to a certain point. However, as the number of gaps increased from 35 to 40 no change in reliability was observed. Furthermore, the improvement in reliability in tests with 20 to 35 gaps was very small. The same pattern was also observed for factorial validity. There is a great leap in the improvement of factorial validity from five to 20 gaps but from 20 gaps onward the change is very little. This indicates that as the number of gaps in each passage increases we get more information about examinees'. However, after a certain point the information that extra gaps add become repetitive and we do not learn more about the respondents.

Grotjahn (1987) argues that with small number of gaps we cannot measure text-level macro-skills. He states that 25 or even 30 blanks are needed to measure these skills. The results of the present study suggest that with 25 gaps in each passage the reliability and factorial validity is reasonably high for most purposes. However, for placement and research purposes where we need to have some rough proficiency groupings we can have C-Tests with 15 gaps. This will save a lot of administration and scoring time and leaves the researcher with plenty of time for the testing of the other variables of interest. However, for high-stakes testing contexts where more precise measures are required more gaps are needed.

One limitation to the study is that the datasets of five to 35-gap passages were not actually separate test forms. That is, the data for these datasets contained the entire passages. The additional items were just ignored in the analyses. Test-takers had the chance to take advantage of the entire passages for answering the items which may have affected the results. The results might be different if we construct C-Tests out of very short passages that contain only ten or fifteen items without further text to help students in text-processing. Future research in this line should focus on constructing separate C-Test forms which have passages with different number of gaps and compare their psychometric properties. Especially comparing the results of such a study with the results of the present study can be very informative as regards the effects of larger context on C-Test processing.

It is important to note that the results of this study can only be generalized to C-Tests which contain four passages. The number of passages can also have significant effects on the psychometric properties of C-Test. It would be interesting to study the conjoint effect of the number of passages and the number of gaps in each passage. In other words, what number of passages and with what numbers of gaps optimize the psychometric properties of the C-Test.

References

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* (101-112). Frankfurt/M: Peter Lang.

Baghaei, P. (2007). C-Test construct validation: A Rasch modeling approach. Unpublished PhD dissertation. Klagenfurt University.

Bendig, A.W. (1954a). Reliability and the number of rating-scale categories. *Journal of Applied Psychology, 38*, 38–40.

Bendig, A.W. (1954b). Reliability of short rating scales and the heterogeneity of the rated stimuli. *Journal of Applied Psychology, 38*, 167–170.

Chapelle, C.A. **(**1994). Are C-tests valid measures for L2 vocabulary research? *Second Language Research, 10*, 157–87.

Cicchetti, D. V., Showalter, D. & Tyrer, P. J. (1985). The effect of number of rating scale categories on levels of inter-rater reliability: a Monte-Carlo investigation. *Applied Psychological Measurement, 9*, 31-36.

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). Profiling the advanced language learner: the C-Test in British further and higher education. In R. Grotjahn (Ed.), *Der C-Test. Theoretische Grundlagen und Praktische Anwendungen*. Vol. 2. (pp. 217-237). Bochum: Brockmeyer.

Coleman, J. A. (1995a). *Progress, proficiency and motivation among British university language learners.* Trinity College Dublin, Centre for Language and Communication Studies, CLCS Occasional Paper 40.

Coleman, J. A. (1995b). The evolution of language learner motivation in British universities, with some international comparisons. In R. G. Wakely. (Ed.), *Issues and perspectives in language study* (pp. 1-16). Edinburgh: University of Edinburgh/London: CILT.

Coleman, J. A. (1996a). 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. 367-399). Vol.3. 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 Proficiency Survey: an overview of findings. In J. A. Coleman (Ed.), *University language testing and the C-Test*. Proceedings of a conference held at the University of Portsmouth in April 1995. Portsmouth: University of Portsmouth Occasional Papers in Linguistics.

Cox, E. P. (1980). The optimal number of response alternatives for a scale: a review. *Journal of Marketing Research, 17,* 407-422.

Djiwandono, P. (1998). *The relationship between EFL learning strategies, degree of extroversion, and oral communication proficiency*. Unpublished Ph.D dissertation. State University of Malang.

Dolan, C.V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimator using simulated data. *British Journal of Mathematical and Statistical Psychology, 47,* 309–326.

Dörnyei, Z. & Katona. L. (1992). Validation of C-Test among Hungarian EFL learners. *Language Testing*, 9, 187–206.

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

Ferrando, P. J. (2000). Testing the equivalence among different item response formats in personality measurement: A structural equation modeling approach. *Structural Equation Modeling, 7*, 271–286.

Finn, R. H. (1972). Effects of some variations in rating-scale characteristics on the means and reliabilities of ratings. *Educational and Psychological Measurement, 34*, 885–892.

Grotjahn, R. (Ed.) (1992). *Der C-Test. Theoretische Grundlagen und praktische Anwendungen*. Vol.1. Bochum: Brockmeyer.

Grotjahn, R. (Ed.) (1994). *Der C-Test. Theoretische Grundlagen und praktische Anwendungen*. Vol.2. Bochum: Brockmeyer.

Grotjahn, R. (Ed.) (1996). *Der C-Test. Theoretische Grundlagen und praktische Anwendungen*. Vol.3. Bochum: Brockmeyer.

Grotjahn, R. (Ed.) (2002). *Der C-Test. Theoretische Grundlagen und praktische Anwendungen*. Vol.4. Bochum: AKS-Verlag.

Grotjahn, R. (Ed.) (2006). *Der C-Test: Theorie, Empirrie, Anwendungen/The C-test: theory, empirical research, applications*. Frankfurt/ M: Peter Lang.

Grotjahn, R. (Ed.) (2010). *Der C-Test: Beiträge aus der aktuellen Forschung/The C-Test: Contributions from Current Research.* Frankfurt am Main: Peter Lang.

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

Grotjahn, R. (1987). How to construct and evaluate a C-Test: A discussion of some problems and some statistical analyses. In R. Grotjahn, C. Klein-Braley & D.K. Stevenson (Eds.), *Taking their measure: The validity and validation of language tests* (pp. 219-253). Bochum: Brockmeyer.

Hofmans, J., Theuns, T. & Mairesse, O. (2007). Impact of the number of response categories on linearity and sensitivity of Self-Snchoring Scales: A Functional Measurement approach. *Methodology, 3,* 160-169.

Hopp, H. (2006). Syntactic features and reanalysis in near-native processing. *Second language Research, 22*, 369-397.

Jenkings, C. D., & Taber, T.A. (1977). A Monte Carlo study of factors affecting three indices of composite scale reliability. *Journal of Applied Psychology, 62*, 392–398.

Kaiser, H. (1974). An index of factorial simplicity. *Psychometrika, 39*, 31-36.

Krekeler, C. (2006). Language for special academic purposes (LSAP) testing: the effect of background knowledge revisited. *Language Testing, 23*, 99-130.

Lee-Ellis, S. (2009). The development and validation of a Korean C-Test using Rasch Analysis. *Language Testing, 26*, 245–274.

Lozano, L. M., García-Cueto, M. & Muñiz, J. (2008) Effect of the number of response categories on the reliability and validity of rating scales. *Methodology, 4*, 73-79.

McKelvie, S. J. (1978). Graphic rating scales: How many categories? *British Journal of Psychology, 69,* 185-202.

Neumann, L. (1979). Effects of categorization on relationships in bivariate distributions and applications to rating scales. *Dissertation Abstracts International, 40*, 2262-B.

Nunnally, J. C. (1970). *Psychometric theory*. New York: McGraw-Hill.

Preston, C. C. & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica 104*, 1-15.

Ramsay, J. O. (1973). The effect of number of categories in rating scales on precision of estimation of scale values. *Psychometrika, 38*, 513-533.

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.

Raatz, U. & Klein-Braley, C. (1985). How to develop a C-Test. In C. Klein-Braley, & U. Raatz (Eds.), *Fremdsprachen und Hochschule 13/14: Thematischer Teil: C-Tests in der Praxis.* (pp. 20-22). Bochum: AKS.

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). AKS-Verlag.

Read, J. & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing, 18,* 1–32.

Ridley, J. & Singleton, D. (1995). Strategic L2 lexical innovation: case study of a university-level ab initio learner. *Second language Research, 11*, 137-148.

Schmid, M. S. & Dusseldorp, E. (2010) Quantitative analyses in a multivariate study of language attrition: the impact of extralinguistic factors. *Second Language Research*, 26, 125-160.

Sigott, G. (2004). *Towards identifying the C-Test construct*. Frankfurt/M: Peter Lang.

Singleton, D. (1990). *The TCD Modern Languages Research Project: objectives, instruments and preliminary results. CLCS Occasional Paper 26.* Dublin: Trinity College (ERIC Reports ED 333 723).

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

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

Weng, L. (2004). Impact of the number of response categories and anchor labels on coefficient α and test-retest reliability. *Educational and Psychological Measurement, 64*, 956–972.

Table 1.Passage difficulty and discrimination for the datasets

	Passage 1		Passage 2		Passage 3		Passage 4	
# of Gaps	P-value	Discr.	P-value	Discr.	P-value	Discr.	P-value	Discr.
5	.47	.45	.70	.43	.75	.43	.68	.25
10	.55	.68	.64	.57	.67	.73	.59	.58
15	.60	.70	.66	.71	.65	.75	.63	.70
20	.57	.74	.65	.74	.66	.77	.60	.75
25	.55	.73	.60	.75	.71	.79	.58	.80
30	.55	.70	.63	.77	.69	.83	.60	.80
35	.53	.70	.62	.82	.70	.85	.61	.82
40	.55	.72	.60	.82	.68	.86	.61	.84

Table 2. Test level statistics for the datasets

# of Gaps	Alpha	Mean %	S.D. %	% Var.	Cor.
5	.61	65	17.79	46.80	.56*
10	.81	61.25	18.45	65.08	.58*
15	.86	63.40	18.19	71.82	.62*
20	.88	62.15	17.26	74.93	.60*
25	.89	60.82	16.83	76.72	.58*
30	.90	60.85	16.33	77.18	.58*
35	.91	61.48	16.78	79.09	.59*
40	.91	60.95	17.14	80.66	.59*