ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 1, No. 6, pp. 688-693, June 2011

doi:10.4304/tpls.1.6.688-693

Do C-Tests with Different Number of Gaps Measure the Same Construct?

Purya Baghaei
Islamic Azad University, Mashhad Branch, Iran
Email: puryabaghaei@gmail.com

***Abstract*—Constructing C-Tests with 20-25 blanks in each passage is the standard practice among C-Test users and researchers. The present study explores if C-Tests with different number of blanks measure the same construct as those with standard 25 gaps. Out of data from a four-passage C-Test with 40 blanks in each passage, two other datasets with 10 and 25 blanks in each passage were constructed. The three datasets were separately analysed with the Rasch model and the psychometric characteristics of the three test forms were compared. Results showed that the three forms have similar fit indices and are equally unidimensional. But C-Tests with more gaps turned out to be more reliable. Cross plotting students' ability measures from pairs of C-Test forms on x and y axes revealed that test-takers have identical ability measures regardless of the number of gaps in the C-Tests. This was interpreted as evidence of the invariance of the C-Test construct when the number of gaps in the passages is altered. Implications of the study for C-Test applications with an important word of caution are given.**

***Index Terms*—C-Test, Rasch model, invariance, fit statistics**

I. Introduction

Raatz and Klein-Braley (2002) and Grotjahn (1987) in their guidelines for C-Test construction suggest four to six passages with 20-25 gaps in each passage. Almost all C-Test researchers and practitioners have followed this recommendation and have developed C-Tests which contain 20-25 gaps (Babaii&Shahri, 2010, Baghaei, 2007; Daller & Phelan, 2006; Koller & Zahn, 1996; Kontra & Kormos, 2006; Linnemann & Wilber, 2010; Norris, 2006; Traxel & Dresemann, 2010). The issue of the optimal number of gaps in C-Test passages has not received much attention in the literature. Among few researchers who have touched on the topic is Grotjahn (1987). He states that with small number of gaps we cannot measure macro text-level skills and argues that 25 or even 30 blanks are needed to measure these skills. Baghaei (2011) systematically monitored and compared the psychometric properties of C-Tests with different number of gaps. He compared eight C-Tests which contained 5, 10, 15, 20, 25, 30, 35, and 40 gaps in each passage and concluded that as the number of gaps in each passage increased the internal consistency reliability, item discrimination and factrorial validity increased accordingly.

Nevertheless, Baghaei's (2011) results indicated that the boost in reliability and factorial validity of the C-Tests after 15 gaps was very small.The findings showed that C-Tests which contained 10 and 15 gaps also enjoyed acceptable indices of discrimination and reliability and correlated as high as C-Tests which contained 30, 35 and 40 gaps with a reading comprehension tests which was used as anexternal criterion. The direct implication of these findings is that for low stakes and medium stakes tests where very precise measures are not required, practitioners can construct C-Tests with 10-15 blanks in each passage to save on administration and scoring time.

Research on optimal number of response categories in Likert-type items, which are extensively used in psychology, also shows that Likert scales which contain seven to nine points result in highest psychometric characteristics for psychological instruments and increasing the number of response categories above nine does not improve scale characteristics(Bendig 1954; 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; Nunnally, 1970; Preston & Colman, 2000; Ramsay, 1973; Weng, 2004). Since C-Tests are analysed like Likert scales by considering each passage as a super-item and entering the aggregated passage scores into analysis, 10-gap C-Tests produce psychometrically adequate results. But psychometric equivalence which is investigated by comparing internal consistency reliability and item discrimination does not guarantee construct invariance.

The important question which is raised here is whether C-Tests with different number of gaps measure the same construct. It is extremely important to ascertain that C-Tests which contain say 10 gaps measure the same construct as those which contain 25 gaps, before one embarks on using C-Tests with smaller number of gaps.The purpose of the present study is to empirically show that this is in fact the case.

II. Method

A. Participants

Participants of the study were 104 Iranian undergraduate students of English at two universities in Mashhad, Iran. There were 73 girls and 31 boys among them; mean age of the subjects was 21.6 with a standard deviation of 3.8.

B. Instrument

For the purposes of this study data from Baghaei (2011) were reanalyzed. A C-Test battery containing four passages each having 40 blanks was the instrument used in Baghaei (2011) and the present study. The internal consistency reliability of the C-Test considering each passage as a super-item was .91.

C. Procedures

Out of the dataset for the C-Test battery with 160 blanks two other datasets were constructed. In the first dataset the scores on the first 10 gaps in each passage were aggregated as if each passage had only 10 gaps to complete; this C-Test battery was named C-Test-10. In the second dataset the scores on the first 25 gaps were aggregated (C-Test-25) and in the third dataset the scores on all the 40 gaps were aggregated (C-Test-40). In this way data for three C-Test batteries with 10, 25 and 40 gaps in each passage were created.

D. Data Analysis

In order to investigate the construct invariance of C-Tests with 10, 25 and 40 blanks, Rasch rating scale model (Andrich, 1978) as implemented in WINSTPES (Linacre, 2009) was utilized. The three datasets were separately analysed with rating scale model by considering each passage as a polychotomous super-item. The difficulty estimates of the items (passages), their associated standard errors, and fit statistics were computed. Table 1 shows the results for the four items in each C-Test.

TABLE 1:
ITEM STATISTICS FOR C-TESTS WITH DIFFERENT NUMBER OF GAPS

Test	Measure				Standard Error				Infit MNSQ			
	Item 1	Item 2	Item 3	Item 4	Item 1	Item 2	Item 3	Item 4	Item 1	Item 2	Item 3	Item 4
C-Test-10	.32	-.15	-.28	.11	.07	.07	.07	.07	.84	1.29	.77	1.06
C-Test-25	.29	.03	-.48	.16	.04	.04	.04	.04	1.09	1.17	1.02	.84
C-Test-40	.23	.04	-.29	.02	.03	.03	.03	.03	1.39	1.14	1.01	.82

Table 1 shows that the difficulty measures of the items only slightly change across the three forms. The greatest change is for Item 3 which has a change of .20 logits, from C-Test-10 to C-Test-25. These changes in item estimates are negligible although one should not expect stability of itemsin this case because the nature of C-Test items changes when more blanks are added to them. When 15 more blanks are added to a C-Test item the item is in fact a new item and expecting it to maintain it difficulty measure is unrealistic. The other point is that to compare the difficulties of items across forms the forms should be equated and a common origin should be set. The results show that even without equating the item estimates have remained invariant across forms. One reason for this is that the items share part of their content. It is worth mentioning that changes of less than .50 logits are considered negligible in the literature (Bond & Fox, 2007).

The fits of the items have also remained stable across the forms, except for item 1 which slightly over fits in C-Test-40, while it fits well in C-Test-10 and C-Test-25.The table indicates that as the number of gaps increases the standard errors of the item estimates decrease since there is more information in scales with more categories.

The statistics in Table 1 clearly indicate that C-Tests with different number of gaps are equally unidimensional and the C-Test construct has remained stable across the three forms. In other words, regardless of the number of gaps C-Tests measure the same construct.

TABLE 2:
SAMPLE STATISTICS FOR C-TESTS WITH DIFFERENT NUMBER OF GAPS (N=104)

Test	Error		Infit MNSQ		Outfit MNSQ		Measure	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
C-Test-10	.39	.11	.97	.74	.97	.73	.70	1.04
C-Test-25	.22	.03	.95	1.29	.94	1.29	.37	.79
C-Test-40	.16	.03	1	1.44	1	1.46	.34	.64

Table 2 shows summary statistics for the sample. C-Test-40 is the most precise test with the smallest average error. The mean of infit and outfit indices are very close across the three forms. The ideal value for infit and outfit mean square values is one but values in the range of .7 to 1.3 are acceptable (Linacre, 2009). Although the means of infit and outfit values are one in C-Test-40 their standard deviationsare larger in this form. Acceptable infit and outfit statistics for persons show that persons have behaved as expected and they are located on an interval scale which is independent of the distribution of items.

TABLE 3:
SEPARATION, RELIABILITY AND RMSE FOR PERSONS AND ITEMS IN C-TESTS WITH DIFFERENT NUMBER OF GAPS

Test	Separation		Reliability		RMSE	
	Persons	Items	Persons	Items	Persons	Items
C-Test-10	2.13	2.92	.82	.90	.44	.08
C-Test-25	2.94	6.51	.90	.98	.25	.04
C-Test-40	3.27	5.69	.91	.97	.19	.03

Table 3 reveals that person reliability and separation are highest in C-Test-40 although for items these indices are higher in C-Test-25.Root mean square error (RMSE) for persons and items has the smallest value for C-Test-40, meaning that asthe number of gaps increases measurement becomes more precise.

E. Checking Person Measure Invariance

Under the Rasch model two tests measure the same construct if test-takers have the same ability estimates on them-after their estimates are brought onto the same scale. "The invariance of relative person ability … is good evidence to suggest that these tests [BLOT and PRTIII] can be used interchangeably. Invariance of person estimates and item estimates within the modeled expectations of measurement error over time, across measurement contexts, and so on, is a key Rasch strategy" (Bond & Fox, 2007, p. 88).

If two tests measure the same construct then persons' ability measures on the two tests should be equivalent within measurement error (Baghaei, 2009). In traditional correlational studies non-linear raw scores are used for this purpose. Cross-plotting the Rasch linear transformations of the non-linear raw scores which are in the form of person measures are more informative. These linear measures contain more information about the locations of individuals on the ability continuum and the precision of the locations, i.e. the extent to which the location is blurred. The precision of measures is embodied in the standard errors of the person parameter estimates.

If the measures from the two tests are identical, after plotting them on x and y axes the dots would exactly fall on a diagonal line with the slope of unity (45 °). This happens under ideal error-free conditions which is not achievable in practice. When cross-plotting traditional raw scores, dispersion of the points from the 45 ° diagonal line is considered as unmodeled variance (Bond & Fox, 2007). That is, we do not know how much of dispersion is tolerable under practical measurement conditions to sustain that the two scores are similar enough to consider the tests as measures of the same dimension. "Consequently, test development approaches based on the inspection of correlations generally are less than adequate" (Bond & Fox, 2001: 59).

In Rasch measurement, conversely, we have the standard errors of estimates which can be exploited to model the deviations of the points from the perfect diagonal line. On the basis of the standard errors of ability estimates from both test 95% quality control lines are drawn (cf. Baghaei, 2009/2010). If the dispersion of scores is within the modeled 95% quality control lines, the test-developer can argue that the two tests are measuring the same dimensions even if the correlation between the two tests is "disastrously" low (Masters & Beswick, 1986, cited from Bond & Fox, 2007, p.88). Therefore, if the deviations of scores from the 45 ° diagonal line are within modeled measurement error, which is indicated by the two parallel quality control lines, one can argue that the measures from the two tests are invariant enough to use them interchangeably.

Figure 1 shows the cross plot of person measures from C-Test-10 against C-Test-25. The plot shows that all the persons are clustered around the best fit line and no person falls outside the parallel quality control lines. In other words, whether we administer a C-Test battery with 10 gaps in each passage or 25 gaps in each passage, test-takers will have the same ability estimates. The coefficient of correlation between measures from the two tests is .94.

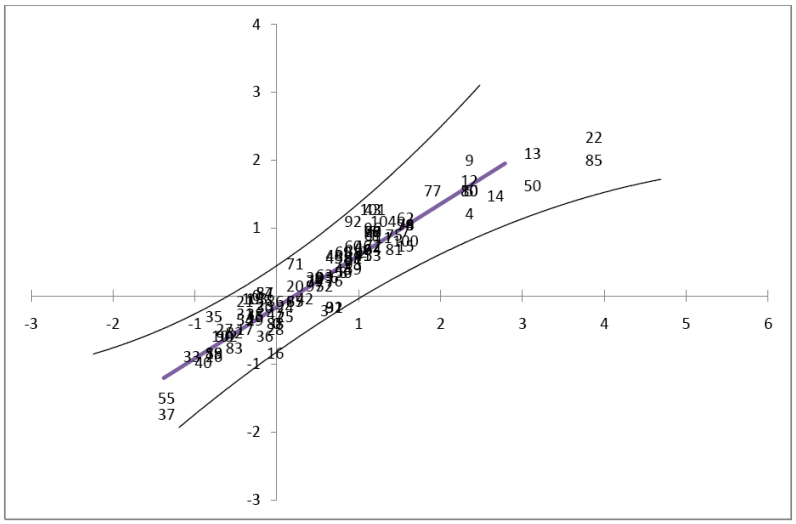

Figure 1: Cross plot of person measures from C-Test-10 against C-Test-25

Figure 2 shows the cross plot of person measures from C-Test-10 against C-Test-40. This plot also shows that all the persons are clustered around the best fit line and no person falls outside the parallel quality control lines. That is, whether we administer a C-Test battery with 10 gaps in each passage or 40 gaps in each passage, test-takers will have the same ability estimates. The coefficient of correlation between measures from the two tests is .92.

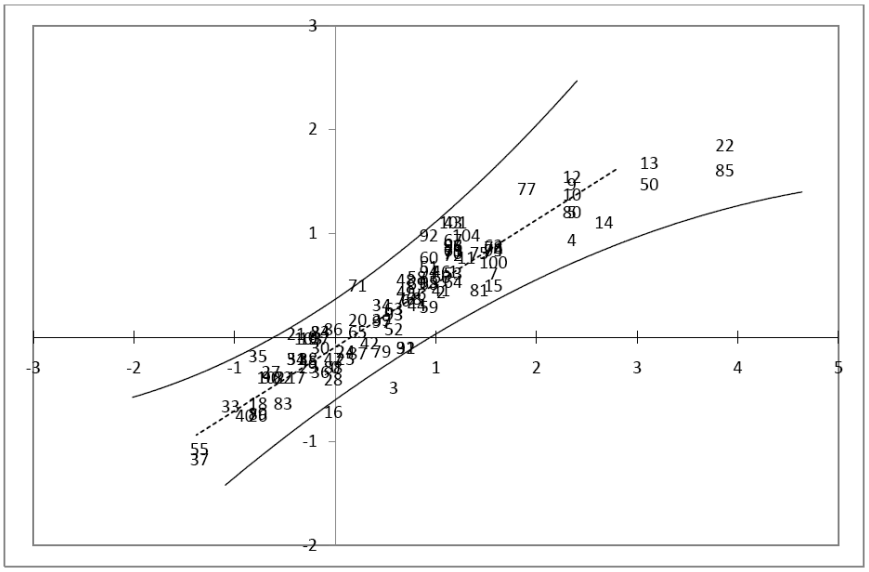

Figure 2: Cross plot of person measures from C-Test-10 against C-Test-40

Figure 3 shows the cross plot of person measures from C-Test-25 against C-Test-40. This plot also shows that all the persons are clustered around the best fit line and no person falls outside the parallel quality control lines. That is, whether we administer a C-Test battery with 25 gaps in each passage or 40 gaps in each passage, test-takers will have the same ability estimates. The coefficient of correlation between measures from the two tests is .98.

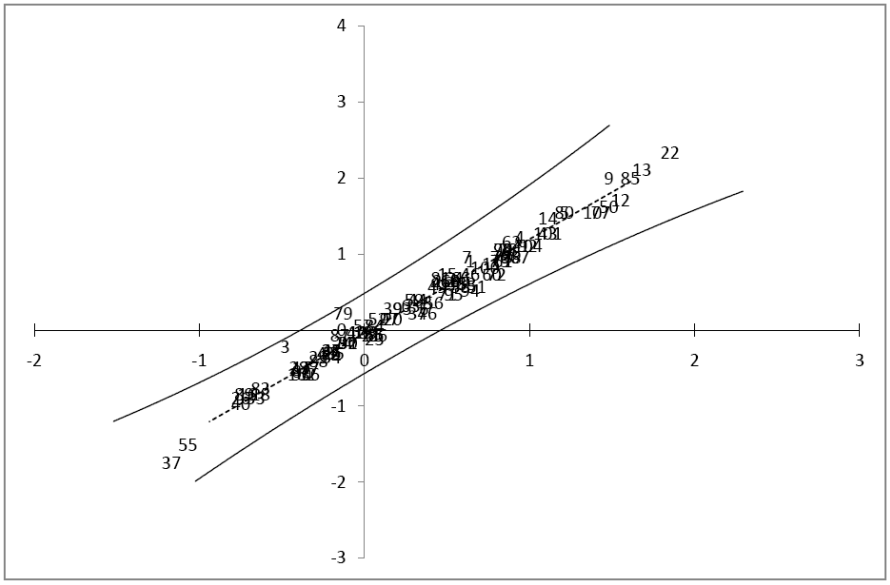

Figure3: Cross plot of person measures from C-Test-25 against C-Test-40

III. Results and Discussion

Graphical investigation of the invariance of person measures estimated from C-Tests with 10, 25 and 40 gaps in each passage revealed thatexaminees have the same ability estimates regardless of the number of blanks in each passage. That is, the construct of C-Test remains invariant across C-Tests with different number of gaps. Fit analyses of the tests showed that C-Test-10, C-Test-25 and C-Test-40 are equally unidimensional. Results showed that, however, with more gaps in each passage the measurement of the latent trait is more precise as C-Tests with more gaps turned out to be more reliable with smaller RMSE's.

The results of this study are in line with Baghaei (2011). He demonstrated that the correlations between C-Tests with different number of gaps (5, 10, 15, 20, 25, 30, 35, 40) and a reading comprehension test, which was used as an external criterion, were almost the same for all the eight C-Tests. This is evidence that C-Test construct is independent of the number of gaps in each passage. Comparing this finding with those of Alderson (1979, 1983) about cloze, which showed that the correlations of cloze tests with different number of deletions and external criteria drastically change, demonstrates the stability of C-Test construct.

The implication of the study is that C-Tests with 10 to 15 gaps are equivalent to C-Tests with 25 to 40 gaps in terms of the construct they measure. For practical testing purposes, when there are constraints of testing time and when not very precise measures of proficiency will do, C-Test with 10-15 gaps in each passage can be employed.

A note of caution is necessary though. The data from C-Test with 10 and 25 gaps were not from separate C-Tests. As was stated in the Method section there was only one C-Test in this study with 40 gaps in each passage given to one sample. Data for C-Tests with 10 and 25 gaps were created out of this dataset by adding the scores on the first 10 and 25 gaps in each passage respectively and ignoring the rest. This means that when performing on C-Test-10 and C-Test-25, test-takers had the entire passage with 40 blanks to resort to. That is, they had all the contextual clues available in the larger context. The results might be different if we construct short C-Tests which contain only 10 gaps without further context to help test-takers, except leaving the first and last sentences intact, which is standard in C-Test construction. The reason why C-Tests with different number of gaps turned out to be measuring the same construct could well be that the psychological processes underlying the performance on C-Test-10,C-Test-25 and C-Test-40 weremade similar by the larger context which was at the test-takers' disposal to the same degree. Further research is needed with separate C-Tests with small number of items to ascertain that C-Tests with fewer numbers of gaps are measuring the same construct as those with more gaps.

REFERENCES

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

[2] 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*(pp. 205-217). Rowley MA: Newbury House.

[3] Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561-573.

[4] Babaii, E. &Shahri, S. (2010). Psychometric rivalry: The C-Test and the cloze test interacting with test takers' characteristics. In R. Grotjahn (Ed.), *Der C-Test: Beiträgeaus der aktuellen Forschung/The C-Test: contributions from current research* (pp. 41-56). Frankfurt/am: Peter Lang.

[5] Baghaei, P. (2011).Optimal number of gaps in C-Test passages. *International Education Studies*, *4*(1), 166-171.

[6] Baghaei, P. (2010). An investigation of the invariance of Rasch item and person measures in a C-Test. InR. Grotjahn (Ed.), *Der C-Test: Beiträgeaus der aktuellen Forschung/The C-Test: contributions from current research* (pp. 101-112). Frankfurt/am: Peter Lang.

[7] Baghaei, P. (2009).Understanding the Rasch model. Mashhad: Mashhad Islamic Azad University Press.

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

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

[10] Bond, T.G. & Fox, C.M. (2007). (2nd ed.) Applying the Rasch model: fundamental measurement in the human sciences. Lawrence Erlbaum.

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

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

[13] 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, Emprie, Anwendungen/The C-Test: Theory, empirical research, applications* (pp. 101-119). Frankfurt/am: Peter Lang.

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

[15] 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.

[16] 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.

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

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

[19] 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.

[20] Koller, G. & Zahn, R. (1996). Computer based construction and evaluation of C-Tests. In R. Grotjahn (Ed.),*Der C-Test. Theoretische Grundlagen und Praktische Anwendungen*. Vol.3. Bochum: Brockmeyer.

[21] Kontra, E.H. & Kormos, J. (2006). Strategy use and the construct of C-Test.In R. Grotjahn (Ed), Der C-Test: *Theorie, Emprie, Anwendungen/The C-Test: Theory, empirical research, applications* (pp. 121-138). Frankfurt/am: Peter Lang.
[22] Linacre, J.M. (2009). WINSTEPS [Computer Software]. Version 3.69.1.10, Chicago, IL: winsteps.com.
[23] Linacre, J. M. (2009). A user's guide to WINSTEPS-MINISTEP: Rasch-model computer programs. Chicago, IL: winsteps.com.
[24] Linnemann, M. & Wilbert, J. (2010).The C-Test: A valid instrument for screening language skills and reading comprehension of children with learning problems? In R. Grotjahn(Ed.), *Der C-Test: Beiträgeaus der aktuellenForschung/The C-Test: contributions from current research* (pp. 113-124). Frankfurt/am: Peter Lang.
[25] Lozano, L. M., García-Cueto, M. &Muñiz, J. (2008). Effect of the number of response categories on the reliabilityand validity of rating scales. *Methodology, 4*, 73-79.
[26] McKelvie, S. J. (1978). Graphic rating scales: How many categories? *British Journal of Psychology, 69*, 185-202.
[27] Neumann, L. (1979). Effects of categorization on relationships in bivariate distributions and applications to ratingscales. *Dissertation Abstracts International, 40*, 2262-B.
[28] Norris, J. M. (2006).Development and validation of a curriculum-based German C-Test for placement purposes. In R. Grotjahn (Ed.), *Der C-Test: Theorie, Emprie, Anwendungen/The C-Test: Theory, empirical research, applications* (pp. 45-83). Frankfurt/am: Peter Lang.
[29] Nunnally, J. C. (1970). Psychometric theory. New York: McGraw-Hill.
[30] 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.
[31] Ramsay, J. O. (1973). The effect of number of categories in rating scales on precision of estimation of scale values. *Psychometrika, 38*, 513-533.
[32] 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.
[33] Traxel, O. & Dresemann, B. (2010). Collect, calibrate, compare: A practical approach to estimating the difficulty of C-Test items. In R. Grotjahn(Ed.), *Der C-Test: Beiträgeaus der aktuellen Forschung/The C-Test: contributions from current research* (pp. 113-124). Frankfurt/am: Peter Lang.
[34] 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.

Purya Baghaei is an assistant professor in the English Department of Mashhad Islamic Azad University in Iran. He holds a PhD in applied linguistics from Klagenfurt University, Austria. His major research interests are language testing and the application of Rasch models in education and psychology. He has published a book on fundamentals of Rasch model and has published several articles on language testing and Rasch measurement.