Educational Methods & Practice

https://emp-open.de
Volume 1 (2023)

Rapid Communication

A Comparison of Different Methods for Investigating the Reliability of C-Tests

Zahra Motallebzadeh*
Hakim Sabzevari University, Sabzevar, Iran

Abstract

C-Tests are widely used as tests of general language proficiency in foreign language assessment and research. C-Test is reported as a reliable measure with reliability indexes of 0.80 and above. However, in most studies only Cronbach's alpha reliability has been reported. In this research, we aim to examine the reliability of C-Test with other methods of estimating reliability. To achieve this, three different datasets from previous studies were employed to investigate the reliability of C-Test with different methods. Eight methods of estimating reliability, namely, Cronbach's Alpha, split-half, Guttman, parallel forms, omega, MS (*rho*), Lambda-2, and LCRC, were applied to evaluate the reliability of the C-Tests. Findings showed that C-Tests are highly reliable and different methods yield similar results. The findings of this study support the high reliability of C-Tests.

Keywords

C-Test, reliability, alpha, omega, split-half, parallel, Guttman, Lambda-2, MS, LCRC

1 | Introduction

C-Tests, which are formed by deleting the second half of every second word in a text, measure competency in a foreign or second language (Klein-Braley, 1985). A C-Test usually consists of 4 to 8 authentic texts which are hierarchically ordered by their difficulties. Each text contains 20 to 25 blanks where deletions begin from the second word in the second sentence. However, unlike cloze tests, the second half of every other word is deleted instead of whole words. The first and the last sentences in each passage are kept intact to provide some context, and the blanks are spread between these two sentences (Baghaei & Effatpanah, 2023a; Klein-Braley, 1997).

Some scholars argue that the C-Test scores are ambiguous in measuring the proficiency of foreign learners. On the other hand, others claim that C-Tests measure writing and reading abilities (see Sigott 2004 for a complete review). An advantage of C-Tests is that only exact word procedure of scoring is possible, and if the test takers reconstruct the word correctly, they will score one and otherwise zero; therefore, this makes C-Tests highly objective and reliable (Baghaei & Grotjahn, 2014a). In longer C-Tests, reliabilities often are greater than 0.90 like in the onDaf (see https://www.onset.de),

Corresponding author:
*Zahra Motallebzadeh, Department of English Language and Literature, Hakim Sabzevari University, 9617976487 Sabzevar, Iran.
Email: z.motallebzadeh@gmail.com
Zahra Motallebzadeh: https://orcid.org/0009-0008-2510-8575

Received 28 October 2023; Received in revised form 2 December 2023; Accepted 11 December 2023
Available online 17 December 2023

and even in shorter ones, they exceed 0.85 (Eckes & Grotjahn, 2006). C-Tests are highly recommended for measuring general language proficiency in different contexts (see Eckes & Baghaei, 2015).

C-Tests are frequently used in language testing and assessment. One of the distinguishing features of C-Tests is that they are economical measurement instruments. C-Tests are highly objective and, therefore, they could be considered as tests with high reliability (0.80s or higher) (Eckes & Grotjahn, 2006). C-Test is a well-researched test format and up to 2020, there were more than 503 entries in the C-Test bibliography (Grotjahn & Drackert, 2020).

C-Test as an alternative form of cloze test is one type of reduced redundancy test of language proficiency (Klein-Braley, 1985). In the beginning, it was believed that C-Tests measured general second language proficiency, but after 40 years, many studies related to the C-Test in different languages show that this alternative form of cloze test could measure both first and second language proficiency (Grotjahn, 2019) and even crystalized intelligence (Baghaei & Tabatabaee-Yazdi, 2015).

Many researchers and experts confirm that C-Tests are valid (See Norris, 2018). The construct validity of C-Test has been examined in many studies and results demonstrate that C-Tests are highly correlated with a variety of general language proficiency tests (Sigott, 2004). The investigations related to the C-Test reliability are not as many as those of validity. Therefore, the purpose of this research is to examine C-Test reliability using different methods.

C-Tests, as a type of reduced redundancy test of language proficiency, are applied in many settings (Grotjahn & Drackert, 2020). C-Tests are objectively scored and, therefore, are highly reliable (Eckes & Grotjahn, 2006). C-Tests reliability has been examined in several studies. Table 1 shows C-Test reliabilities reported by different researchers.

Table 1

Summary of Studies on the Reliability of C-Tests

Author	Number of passages	Reliability	Reliability formula type
Baghaei (2008)	4	0.91	Cronbach's alpha
Baghaei et al. (2009)	4	0.85	Cronbach's alpha
Lee-Ellis (2009)	5	0.97	Person separation
Baghaei (2011)	4	0.88	Cronbach's alpha
Eckes (2011)	10	0.96 – 0.97	Person separation
Baghaei (2014)	5	0.92	Person separation
Baghaei & Grotjahn (2014a)	2	0.87	EAP*
Baghaei & Grotjahn (2014b)	8	0.82	EAP
Eckes & Baghaei (2015)	8	0.91	Person separation

Note. *EAP: Expected A Posteriori (Adams, 2005)

2 | Method

In this study, three different C-Test datasets were evaluated. Dataset 1 came from a C-Test with four English passages. Two of them were spoken language texts taken from dialogues, and two passages were written text passages. Each passage contained 25 gaps and there were totally 100 items in the entire test. The data were originally collected and analyzed in Baghaei et al. (2009). The dataset contained the responses of 99 Iranian BA English students (23 males and 76 females aged between 21 and 33). The second C-Test battery contained five Persian passages, each containing 25 gaps. The texts were taken from scientific and general knowledge books written for teenagers. These data were used in Baghaei (2014). The test had been given to 158 male participants (aged between 12 and 17) who were Iranian junior and senior high school students.

The third dataset contained six German passages and 256 participants (129 native German speakers and 127 learners of German as a second language) and belonged to the study by Forthmann et al. (2020).

3 | Results

The reliability of the three C-Test batteries in this study was evaluated using different methods including Cronbach's alpha, split-half, Guttman, parallel, omega, MS, and LCRC (Latent Class Reliability Coefficient, van der Ark et al., 2011). Alpha is the most well-known estimator of reliability (Cronbach, 1951). It is a lower bound to the reliability which means that for large samples, the true reliability is larger than alpha. Lambda-2 (Guttman, 1945) is also a lower bound to the reliability. It is older than alpha but less well-known than it. However, lambda-2 is closer to the true reliability than alpha, in such a way that: alpha ≤ lambda-2 ≤ true reliability. MS is the Molenaar-Sijtsma statistic (also known as *rho*; Molenaar & Sijtsma, 1984, 1988; Sijtsma & Molenaar, 1987; van der Ark, 2010). If Mokken's double monotonicity model (Mokken, 1971) holds, MS is an (almost) unbiased estimate of the true reliability. LCRC is an estimate of the test-score reliability based on latent class analysis (van der Ark et al., 2011; van der Palm et al., 2014).

Tables 2, 3, and 4 show the reliabilities of the C-Test batteries calculated with these methods. In these analyses, each passage was considered a super-item. SPSS program, *mokken* package in R (van der Ark, 2012), and the *MBESS* package in R (Kelley, 2007) were used to estimate the reliability coefficients.

Table 2

C-Test Reliability in Dataset 1

Cronbach's Alpha	Split-half	Guttman	Parallel	Omega	MS	Lambda	LCRC
0.863	0.840	0.843	0.863	0.875	0.870	0.870	0.852

Table 3

C-Test Reliability in Dataset 2

Cronbach's Alpha	Split-half	Guttman	Parallel	Omega	MS	Lambda	LCRC
0.943	0.924	0.938	0.943	0.946	0.950	0.945	0.937

Table 4

C-Test Reliability in Dataset 3

Cronbach's Alpha	Split-half	Guttman	Parallel	Omega	MS	Lambda	LCRC
0.956	0.954	0.959	0.957	0.959	0.959	0.958	0.955

4 | Conclusion

This study set out to investigate the reliability of C-Tests using different methods. The results showed that Omega and MS estimate the highest reliability coefficients, whereas split-half yields the lowest estimate. In the first dataset, Omega with a value of 0.875 is the highest, MS with a value of 0.870 comes next, and split-half yields the lowest measure of C-Test reliability with a value of 0.84. In dataset 2, MS with a value of 0.950 gives the highest estimate of reliability and split-half with a value of 0.924 yields the lowest coefficient. The third dataset is very similar to the second dataset. That is, Omega, MS, and parallel yield the highest scale reliabilities and split-half method with a value of 0.954 gives the lowest coefficient. Through employing various methods for estimating C-Test reliability, the results illustrate that C-Tests are highly reliable. The findings also show that the difference between Cronbach'a alpha, which is mostly criticized for underestimating

reliability (see Baghaei & Effatpanah, 2023b), and Omega (and other more preferred methods of estimating reliability) is not much as far as C-Tests are concerned.

Funding

The author(s) received no specific funding for this work from any funding agencies.

Conflict of Interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Availability Statements

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

How to Cite:

Motallebzadeh, Z. (2023). A comparison of different methods for investigating the reliability of C-Tests. *Educational Methods & Practice, 1*:1. URL:https://emp-open.de/page?id=36

References

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31*(2), 162–172. https://doi.org/10.1016/j.stueduc.2005.05.008

Baghaei, P. (2008). The effects of the rhetorical organization of texts on the C-test construct: A Rasch modeling study. *Melbourne Papers in Language Testing, 13*(2), 32–51. URL:https://arts.unimelb.edu.au/__data/assets/pdf_file/0004/3518689/13_2_2-Baghaei.pdf

Baghaei, P. (2011). *C-test construct validation: A Rasch modeling approach*. VDM.

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* (pp. 301–314). Peter Lang.

Baghaei, P., & Effatpanah, F. (2023a). An alternative strategy for modelling local item dependence in C-Tests. In N. Dobric, H. Cesnik, & C. Harsch (Eds.), *Festschrift in honour of Guenther Sigott: Advanced methods in language testing* (pp. 153–164). Peter Lang.

Baghaei, P., & Effatpanah, F. (2023b). *Elements of psychometrics* (2nd Ed.). Sokhan Gostar Publishing.

Baghaei, P., & Grotjahn, R. (2014a). Establishing the construct validity of conversational C-Tests using a multi-dimensional Rasch model. *Psychological Test and Assessment Modeling, 56*(1), 60–82. URL:https://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2014_20140324/04_Baghaei.pdf

Baghaei, P., & Grotjahn, R. (2014b). 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. 165–173). Peter Lang.

Baghaei, P., & Tabatabaee-Yazdi, M. (2015). The C-Test: An integrative measure of crystallized intelligence. *Journal of Intelligence*, *3*(2), 46–58. https://doi.org/10.3390/jintelligence3020046

Baghaei, P., Monshi Tousi, M. T., & Boori, A. A. (2009). An investigation into the validity of conversational C-Tests as a measure of oral abilities. *Iranian EFL Journal, 4*, 94–109. https://profdoc.um.ac.ir/articles/a/1011729.pdf

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334. https://doi.org/10.1007/BF02310555

Eckes, T. (2011). Item banking for C-tests: A polytomous Rasch modeling approach. *Psychological Test and Assessment Modeling, 53*(4), 414–439.

URL:https://www.psychologie-aktuell.com/fileadmin/download/ptam/4-2011_20111217/02_eckes.pdf

Eckes, T., & Baghaei, P. (2015). Using testlet response theory to examine local dependence in C-tests. *Applied Measurement in Education, 28*(2), 85–98. https://doi.org/10.1080/08957347.2014.1002919

Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing*, *23*(3), 290–325. https://doi.org/10.1191/0265532206lt330oa

Forthmann, B., Grotjahn, R., Doebler, P., & Baghaei, P. (2020). A comparison of different item response theory models for scaling speeded C-Tests. *Journal of Psychoeducational Assessment, 38*, 692–705. https://doi.org/10.1177/0734282919889262

Grotjahn, R. (2019). C-tests. In S. Jeuk & J. Settinieri (Eds.), *Sprachdiagnostik Deutsch als zweitsprache: Ein handbuch* (pp. 579–603). De Gruyter Mouton.

Grotjahn, R., & Drackert, A. (2020). The electronic C-Test bibliography: Version October 2020. Available at:http://www.c-test.de & https://www.ruhr-uni-bochum.de/sprachetesten/index.html.de

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10,* 255–282. https://doi.org/10.1007/BF02288892

Kelley, K. (2007). Methods for the behavioral, educational, and social sciences: An R package. *Behavior Research Methods, 39*(4), 979–984. https://doi.org/10.3758/bf03192993

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. https://doi.org/10.1177/026553228500200108

Klein-Braley, C. (1997). C-tests in the context of reduced redundancy testing: An appraisal. *Language Testing*, *14*(1), 47–84. https://doi.org/10.1177/026553229701400104

Lee-Ellis, S. (2009). The development and validation of a Korean C-Test using Rasch analysis. *Language Testing, 26*(2), 245–274. https://doi.org/10.1177/0265532208101007

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Mouton. https://doi.org/10.1515/9783110813203

Molenaar, I. W., & Sijtsma, K. (1984). Internal consistency and reliability in Mokken's nonparametric item response model. *Tijdschrift voor Onderwijsresearch, 9*, 257–268. Retrieved from https://pure.uvt.nl/ws/portalfiles/portal/1030704/INTERNAL.PDF

Molenaar, I. W., & Sijtsma, K. (1988). Mokken's approach to reliability estimation extended to multicategory items. *Kwantitatieve Methoden, 9*(28), 115–126. Retrieved from https://pure.uvt.nl/ws/portalfiles/portal/1030575/MOKKEN__.PDF

Norris, J. M. (2018). Developing and investigating C-tests in eight languages: Measuring proficiency for research purposes. In J. M. Norris (Ed.), *Developing C-tests for estimating proficiency in foreign language research* (pp. 7–33). Peter Lang.

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

Sijtsma, K., & Molenaar, I. W. (1987). Reliability of test scores in nonparametric item response theory. *Psychometrika, 52*, 79–97. https://doi.org/10.1007/BF02293957

van der Ark, L. A. (2010). Computation of the Molenaar Sijtsma statistic. In A. Fink, B. Lausen, W. Seidel, & A. Ultsch (Eds.), Advances in data analysis, data handling and business intelligence (pp. 775–784). Springer. https://doi.org/10.1007/978-3-642-01044-6_71

van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software*, *48*(5), 1–27. https://doi.org/10.18637/jss.v048.i05

van der Ark, L. A., van der Palm, D. W., & Sijtsma, K. (2011). A latent class approach to estimating test-score reliability. *Applied Psychological Measurement, 35*(5), 380–392. https://doi.org/10.1177/0146621610392911

van der Palm, D. W., van der Ark, L. A., & Sijtsma, K. (2014). A flexible latent class approach to estimating test-score reliability. *Journal of Educational Measurement, 51*(4), 339–357. https://doi.org/10.1111/jedm.12053