Castledown

OPEN ACCESS

Language Education & Assessment
ISSN 2209-3591
https://www.castledown.com/journals/lea/

Language Education & Assessment, 8(1), 2279 (2025)
https://doi.org/10.29140/lea.v8n1.2279

Comparative Evaluation of C-test Reliability Using Classical and Modern Psychometric Methods

NEDA KIANINEZHAD[a]

MOHSEN KIANINEZHAD[b]

[a]*Hakim Sabzevari University, Iran*
kianinezhad.neda@gmail.com

[b]*University of Tehran, Iran*
mohsenkianinezhad.41@gmail.com

Abstract

This study presents a comparative analysis of classical reliability measures, including Cronbach's alpha, test-retest, and parallel forms reliability, alongside modern psychometric methods such as the Rasch model and Mokken scaling, to evaluate the reliability of C-tests in language proficiency assessment. Utilizing data from 150 participants across two testing sessions, the results indicate high internal consistency (Cronbach's α = .91, .88), strong parallel forms reliability (r = .82), and moderate test-retest reliability (r = .76). Rasch analysis provided nuanced insights into item difficulty and person reliability, while Mokken scaling confirmed a reliable ordinal hierarchy among test items. This combined methodological approach offers a more comprehensive view of C-test performance and its psychometric properties. Future research should explore scalability across diverse linguistic contexts to enhance the test's generalizability and practical applicability.

Keywords: C-test, reliability, classical and modern psychometric methods

Introduction

Reliability is a cornerstone of test validation, ensuring that assessments yield consistent results across different contexts and populations (McMillan, 2017). The C-test, as a result, designed to measure language proficiency through the restoration of systematically deleted text segments, is recognized for

Data Availability Statement: *All relevant data are within this paper.*

its efficiency and high correlation with general language competence (Klein-Braley & Raatz, 1984). Traditionally, C-tests have been evaluated using classical psychometric approaches, including internal consistency, test-retest reliability, and parallel forms reliability (Schmitt, 1996; Thorndike et al., 1991).

With the advent of modern psychometrics, models such as Rasch analysis and Mokken scaling allow for more granular, item-level assessments of reliability. The Rasch model, moreover, grounded in item response theory (IRT), offers detailed evaluations of item difficulty and individual ability, enhancing interpretability and providing stronger support for construct validity (Amirian et al., 2015; Bond & Fox, 2015; Kazemi et al., 2020; Kianinezhad, 2024). Meanwhile, Mokken scaling, a non-parametric method, assesses whether items create a reliable ordinal scale, which is valuable in determining hierarchical structures within tests (Sijtsma & Molenaar, 2002; Shoahosseini et al., 2024). However, empirical studies comparing classical and modern psychometric approaches for C-tests remain scarce. Therefore, this study aims to address this gap by applying both classical reliability measures and modern psychometric methods to provide a comprehensive assessment of C-test reliability at the item and test levels.

Literature Review

The C-test, introduced by Raatz and Klein-Braley (1981), is widely used for language proficiency assessments. This test format, which requires restoring systematically deleted segments of text, measures linguistic knowledge and comprehension, making it an effective indicator of language proficiency (Klein-Braley & Raatz, 1984). A C-test typically consists of a series of short texts, where the second half of every second word is deleted. Test-takers must reconstruct the missing parts by supplying the correct letters, requiring them to rely on syntactic, morphological, and lexical knowledge. For example, the sentence "Language testing is an essential comp___ of applied ling___" would require test-takers to complete the words as "component" and "linguistics." This method ensures that participants engage with the linguistic structure of the text rather than relying on rote memorization or superficial comprehension.

In fact, reliability, a fundamental concept in language assessment, refers to the consistency and stability of test scores across different administrations and conditions. In psychometric research, reliability ensures that a test measures a construct dependably and without excessive measurement error, thus increasing confidence in score interpretations (Messick, 1995). Therefore, in language testing, high reliability is crucial, as inconsistent scores can undermine the validity of proficiency decisions (Bachman & Palmer, 2010).

Traditional C-test evaluations have predominantly relied on classical test theory (CTT), focusing on internal consistency and test-retest reliability. For instance, Cronbach's alpha, a popular measure of internal consistency, evaluates the degree to which test items measure the same construct (Tavakol & Dennick, 2011). Additionally, test-retest reliability examines the stability of scores over time, an essential factor in confirming the robustness of language assessments (Thorndike, 2005).

Recent advancements in psychometrics offer more detailed item-level insights. Rasch analysis, a cornerstone of IRT, models the interaction between item difficulty and individual ability, providing valuable diagnostics on item fit and test reliability (Bond & Fox, 2015). Similarly, Mokken scaling, grounded in non-parametric IRT, assesses the scalability and hierarchical ordering of test items, thus identifying if test items form a reliable ordinal scale (Sijtsma & Molenaar, 2002). While these models have been applied in other language assessments, there is limited research on their application to C-tests, particularly in multilingual or EFL contexts such as Iran. By integrating classical and modern

psychometric approaches, this study aims to provide a more comprehensive and nuanced evaluation of C-test reliability, addressing existing gaps in the literature.

Methodology

Participants

The sample consisted of 150 participants (80 females, 70 males) enrolled in a university language proficiency program in Iran. All participants were native Persian (L1) speakers, ensuring linguistic homogeneity in the sample. The age range was 18 to 45 (M = 26.5, SD = 6.7). The test of competency in English, C-test, was one of the instruments employed to gauge the language index of the registrants; being the key factor. This sample was covered by the ones who have their proficiency level in between intermediate and advanced on the basis of their pre-placement test results and academic performances and so a convenience sampling was used, selecting students from a university English language proficiency program. Thus, this type of sampling was adopted because the study population was easily approachable and at the same time it made the results and goals of the research be closely related to that particular group. Those who joined the program of learning English language were favored, and thus, they could interact with the C-test reflecting their best levels. This study, however, was diversified in terms of age and level of education and it was hence, more valuable in the context of teaching English to students as a second language.

Instruments

Two versions of the C-test were constructed, each comprising five passages with systematic deletions of the second half of every second word. Each test was scored out of 50 points. The test passages covered all possible general academic subjects, such as education, health, technology, history, and social issues, providing content validity for the assessment in various domains of language use. Version 2 was developed as a parallel form of Version 1, maintaining equivalent difficulty while using different content to allow for parallel forms reliability analysis. The two versions of the test were created by the researchers, following standard C-test construction principles (Klein-Braley, 1981). To ensure equivalence, the tests were developed with matching text length, word frequency distributions, and syntactic complexity levels. Additionally, analyses were conducted to confirm their parallelism.

> For example, in a passage about technological advancements, the original sentence:
> *"Artificial intelligence is rapidly changing the w___ we interact w___ technology"*
> would appear as:
> *"Artificial intelligence is rapidly changing the w___ __ interact w___ t___"*
> where test-takers must correctly complete the words as "way," "with," and "technology."

Procedure

The tests were administered in a controlled classroom environment. Each participant was given 30 minutes to complete each version of the test, ensuring standardized administration conditions. To assess test-retest reliability, Version 1 was administered again two weeks after the initial test session. Version 2 was only taken once, immediately after Version 1 in the initial session, to assess parallel forms reliability. Test administrators provided instructions before each session, and no external reference materials were allowed. Ethical approval was secured from the university's ethics committee, and all participants provided informed consent.

Data Analysis

Five psychometric techniques were utilized to assess the C-test's reliability:

1. **Internal Consistency**: Cronbach's alpha was used to assess the consistency of items in measuring language proficiency. This metric is widely accepted as an indicator of reliability for homogeneous test constructs (Tavakol & Dennick, 2011).
2. **Test-Retest Reliability**: A Pearson correlation coefficient was calculated between participants' scores on Version 1 across the two testing sessions (initial and two-week retest) to evaluate score stability over time (Thorndike, 2005).
3. **Parallel Forms Reliability**: A Pearson correlation was computed between the scores of Versions 1 and Version 2 from the initial test session to assess the equivalence of the two test forms, an important consideration for alternate form reliability (Hambleton & Swaminathan, 2013).
4. **Rasch Analysis**: Conducted via Winsteps software, Rasch analysis yielded item difficulty parameters, person reliability, and fit statistics, providing insights into the interaction between items and examinees—essential for validating test construct (Bond & Fox, 2015).
5. **Mokken Scaling**: Using the R package 'Mokken,' this analysis evaluated the hierarchical structure of test items, confirming that the C-test items form a reliable ordinal scale. This non-parametric approach is particularly valuable for detecting scalability within assessments (Sijtsma & Molenaar, 2002).

Results

This section presents the comparative findings of the reliability analyses conducted on the C-tests, based on classical and modern psychometric approaches.

Internal Consistency

The internal consistency of both C-test versions, as measured by Cronbach's alpha **(**$\alpha = .91$ for Version 1 and $\alpha = .88$ for Version 2), reflects a high level of reliability, which is essential in language proficiency assessments (Table 1). Cronbach's alpha values above .80 are considered indicative of strong reliability, as suggested by Nunnally (1970), who proposed that an acceptable threshold for educational and psychological tests should be at least .70, with .80 or higher reflecting good internal consistency. Cronbach's alpha is widely used to assess the degree to which items on a test consistently measure a single construct—in this case, language proficiency (Tavakol & Dennick, 2011). The high internal consistency suggests that the C-test items effectively measure shared aspects of language proficiency, supporting the cohesive structure of the test.

Test-retest Reliability

The test-retest reliability, indicated by a Pearson correlation of $r = .76$, provides a measure of score stability over time (Table 1). This correlation, which falls into the moderate range, suggests that the C-test yields relatively consistent scores when administered across a two-week interval (Nunnally, 1970). However, the moderate stability may imply that certain external factors influence score variation, such as increased familiarity with the test format and minor improvements in participants' language proficiency over the two-week period**.** For language assessments, test-retest correlations around .70 to .80 are typically considered acceptable, though higher correlations are ideal for proficiency measures

Table 1 *Reliability statistics for C-tests using classical and modern methods*

Method	Statistic	Version 1	Version 2	Version 1 & Version 2
Internal Consistency	Cronbach's alpha (α)	.91	.88	—
Test-Retest Reliability	Pearson correlation (r)	.76	—	—
Parallel Forms Reliability	Pearson correlation (r)	—	—	.82
Rasch Analysis	Item reliability	.89	.87	—
	Person reliability	.84	.81	—
	Misfitting items (Infit MNSQ > 1.4)	3	2	—
Mokken Scaling	Scalability coefficient (H)	.43	.41	—

intended for longitudinal tracking (Nunnally & Bernstein, 1994). The moderate stability here suggests that, while the test is generally reliable, repeated administrations over short intervals might show slight variation due to these extraneous factors.

Parallel Forms Reliability

As shown in Table 1, parallel forms reliability, indicated by a Pearson correlation of r = .82 between the two versions**,** reflects strong agreement and consistency across different test forms. This measure evaluates whether two versions of a test produce similar scores, supporting their interchangeability (Hambleton & Swaminathan, 1985). A high parallel forms reliability is especially useful in contexts where multiple test versions are required, as it ensures that participants can take different versions without one version being inherently easier or harder. This agreement enhances the C-test's practical utility in repeated testing scenarios, as it minimizes practice effects and reduces the likelihood of content memorization influencing test performance.

Rasch Analysis

Rasch analysis, a modern psychometric approach, provides item-level insights beyond classical measures. As shown in Table 1, item reliability was .89 for Version 1 and .87 for Version 2, while person reliability was .84 for Version 1 and .81 for Version 2. These values indicate that the test items align well with participant ability levels, meaning the C-test effectively differentiates between varying proficiency levels (Bond & Fox, 2015).

However, the presence of misfitting items in both versions—three items in Version 1 and two items in Version 2 (infit MNSQ > 1.4)—suggests that certain items did not conform to the expected difficulty-response patterns. Misfit can occur due to ambiguous wording, unexpected participant responses, or inconsistencies in item difficulty across different proficiency levels. Addressing these misfit items through revision or pilot testing could improve alignment with the Rasch model and enhance the overall accuracy of the test (Linacre, 2002).

Mokken Scaling

Mokken scaling analysis yielded scalability coefficients of H = .43 for Version 1 and H = .41 for Version 2**,** indicating a moderate hierarchical structure among items (Table 1). According to Mokken scale classification, a medium scalability level falls within the range of .40 < H < .50

(Sijtsma & Molenaar, 2002; Mota et al., 2023). The Mokken H coefficient evaluates the degree to which test items conform to an ordinal, difficulty-based hierarchy, with values above .30 generally indicating that a scale reliably orders participants according to their ability level. A moderate hierarchy suggests that while the test items align somewhat sequentially in terms of difficulty, there is potential to strengthen this ordering to better capture incremental proficiency stages; hence, optimizing item hierarchy could enhance the test's diagnostic value by clarifying which items align most closely with specific proficiency levels, thereby supporting the C-test's use as a scalable and precise measure of language skill.

Discussion

The findings of this study provide robust support for the reliability of C-tests when evaluated using both classical and modern psychometric methods. The high internal consistency indices, as evidenced by Cronbach's alpha ($\alpha = .91$ and .88 for the two versions), indicate that the test items consistently measure language proficiency, aligning with prior research highlighting C-tests as reliable tools for linguistic competence assessment (Schmitt, 1996; Tavakol & Dennick, 2011). Additionally, the moderate test-retest reliability ($r = .76$) confirms score stability across time, a critical component in longitudinal language assessments. However, this finding also suggests room for improvement, as C-tests may benefit from enhancements that increase score stability over extended periods, especially for proficiency measures that may be administered intermittently over a learner's language development journey.

Parallel forms reliability ($r = .82$) provides strong evidence that alternate versions of the C-test maintain equivalency in difficulty and scoring, which enhances the C-test's applicability across diverse testing conditions. This equivalence supports the practical use of different test forms interchangeably, a valuable feature for settings where repeated assessments are necessary or for institutions seeking to prevent test familiarity effects. The work of Hambleton and Swaminathan (1985) marks the beginning of the concept of parallel forms reliability, however, recent years have seen some research focusing on a new interpretation. Kolen and Brennan (2014) and other works like McCarthy et al. (2021), claim that not only should a well-constructed test be adjusted to match item exposure, but it should also be difficult and yield highly precise data if the language test is to be a very good one. Studies like this support the present-day idea that using multiple test forms in high-stakes language results will keep examinee's scores fair and at the same time will diminish the exposure effect of particular items.

The Rasch model provided nuanced insights that extended beyond classical test theory (CTT) measures. The item and person reliability indices (.89 and .84, respectively, for Version 1, and .87 and .81 for Version 2) suggest that the C-test items function effectively across different levels of language proficiency, a testament to the test's discriminatory power in distinguishing between high and low proficiency. However, the identification of three misfitting items in Version 1 and two misfitting items in Version 2 (Infit MNSQ > 1.4) highlights the value of Rasch diagnostics for pinpointing items that do not conform to expected performance patterns. Besides, misfitting items in Version 2 may have resulted from inconsistencies in item difficulty or ambiguities in lexical cues, which could lead to varied response behaviors across participants. Addressing these issues through targeted revisions, as recommended by Smith (2020), would enhance the precision of the test and ensure more stable measurement properties. In fact, the revision or replacement of these items could improve the measurement accuracy of the C-test and reduce potential biases or ambiguities that may affect lower-ability test-takers, aligning with Bond and Fox's (2015) recommendations for Rasch-based refinement of language assessments.

Mokken scaling provided additional support for the internal hierarchy among C-test items, with moderate scalability coefficients ($H = .43$ and .41), confirming a reliable ordering of item difficulty. According to Sijtsma and Molenaar (2002) and Mota et al. (2023), a moderate scalability level falls

within the range of .40 < H < .50, indicating that test items follow a logical progression but may benefit from further refinement to enhance their hierarchical ordering. This finding implies that the items possess a meaningful structure that corresponds with progressive language proficiency levels, which is valuable for language assessments that aim to identify incremental stages of learning. However, the moderate values suggest that while an ordinal structure exists, additional refinement, i.e., selecting items with clearer stepwise increases in difficulty could improve the precision of the test's hierarchical organization. Future work should explore whether increasing item difficulty variation or restructuring item placement can strengthen this scalability.

Theoretical Implications and Critical Comparison of Psychometric Methods

In the harmonization of classical and modern psychometric methods in this study highlights the mutual attributes of both approaches to better understanding of reliability of C-test. Although classical test theory (CTT) is very widely used, it just tells about the total test reliability; modern psychometric methods such as Rasch analysis and Mokken scaling provide richer item-level diagnostics. CTT methods such as Cronbach's alpha and test-retest reliability provide valuable insights into the overall consistency and stability of the test, while modern techniques like Rasch analysis contribute to refining item properties by identifying misfitting items and ensuring construct validity. Mokken scaling further enhances the analysis by confirming whether test items form a logical hierarchy, making it particularly useful in proficiency-based assessments.

Next, a critical examination of each technique identifies that although CTT has the benefits of being an accessible and interpretable measure of reliability, it lacks precision in diagnosis like item response models. Hence, Rasch analysis is the best basis for an item, difficulty; person ability control target evaluation especially in the development of language tests (Engelhard 2013). Rasch analysis is longitudinal and thus, more difficult then classical approaches but depends on the assumption of unidimensionality as well as item-fit validation. Mokken scaling is a non-parametric IRT method that is more flexible for ordinal language assessments, however its moderate scalability in optimization of items stimulated further that better test items have to be developed to allow clearer progression of difficulty, but establishment has not occurred.

Thus, the combined approach used in this study aligns with emerging trends in psychometric evaluation, where classical and modern methods are jointly employed to maximize reliability, validity, and diagnostic power (Sijtsma, 2012). Future studies should build upon this integrated framework to refine C-test design, ensuring optimal test performance across diverse linguistic and educational settings.

Limitations and Future Research Directions

Although this study made important contributions, one should bear in mind the limitations of it as well. One limitation was the single university from which the sample was drawn, which may not be representative of more general populations. The participants in future research should be from diverse participant pools, i.e., varying types of learners and different educational backgrounds/language processing abilities.) In addition, the study was based on a two-week test-retest interval, that is, although a standard in reliability studies may not fully capture long-term score stability. Additional research could increase the time between retests to determine how long C-test performance sustains over weeks or months.

A further limitation Mokken scaling, is that the moderate scalability coefficients emerged in this study. Although the results provide a reasonable hierarchical structure, further item difficulty classification is needed to support item order for diagnostic test refinement. Further research efforts should examine other procedures like automated IRT-based item selection methods, in order to improve the order of

items and make proficiency stages more interpretable. Besides, the study also did not provide information on differential item functioning (DIF) that would be useful to know whether test items function equally for different groups (e.g., gender, background). Incorporating DIF analysis in future studies would ensure that test performance is not biased.

Conclusion

This study demonstrates the utility of using both classical and contemporary psychometric approaches in assessing reliability of C-test. As a starting point for understanding the overall test consistency, classical methods such as Cronbach's alpha and test-retest reliability. Rasch analysis and Mokken scaling, on the other hand offer finer grain item-level insights from this hierarchy of internal test structure. Together, these methods enhance the psychometric properties of C-test so it can be used as a valid and reliable language proficiency test. Moreover, prospective studies maybe conducted to investigate other advanced techniques such as generalizability theory and advanced IRT (e.g., 2PL, 3PL models) which improve item measure estimation enabling better skill judgments for test takers, respectively. Second, C-test must be cross-culturally validated in order to have this testing instrument be fair and useable in culturally diverse contexts. In the end proper psychometric evaluation of C-test reliability is necessary to advance more inclusive, accurate and fairer language proficiency assessment.

References

Amirian, S. M. R., Salari, S., Heshmatifar, Z., & Rahimi, J. (2015). A validation study of the newly developed version of the vocabulary size test for Persian learners. *International Journal of Education and Research, 3*(8), 359–380.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world.* Oxford University Press.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences.* Routledge.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Kluwer-Nijhoff.

Kazemi, S., Ashraf, H., Motallebzadeh, K., & Zeraatpishe, M. (2020). Development and validation of a null curriculum questionnaire focusing on 21st-century skills using the Rasch model. *Cogent Education, 7*(1), 1736849. https://doi.org/10.1080/2331186X.2020.1736849

Kianinezhad, N. (2024). Validation of the adapted attitudes toward online teaching scale for Iranian EFL teachers: A Rasch model analysis. *Education Mind, 3*(1), 31–45.

Klein-Braley, C., & Raatz, U. (1984). A survey of C-test research. *Language Testing.*

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer.

Linacre, J. M. (2002). Understanding Rasch measurement: Optimal rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85–106.

McCarthy, A. D., Yancey, K. P., LaFlair, G. T., Egbert, J., Liao, M., & Settles, B. (2021, November). Jump-starting item parameters for adaptive language tests. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 883–899. https://doi.org/10.18653/v1/2021.emnlp-main.67

McMillan, J. H. (2017). *Classroom assessment: Principles and practice for effective standards-based instruction* (7th ed.). Pearson.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Mota, J., Martins, J., & Onofre, M. (2023). Portuguese Physical Literacy Assessment Questionnaire (PPLA-Q) for adolescents: Validity and reliability of the psychological and social modules using Mokken scale analysis. *Perceptual and Motor Skills, 130*(3), 958–983.

Nunnally Jr, J. C. (1970). Introduction to psychological measurement.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Raatz, U., & Klein-Braley, C. (1981, September). *The C-Test--A Modification of the Cloze Procedure*. Presented at the Language Testing Symposium, Colloquium at the University of Essex, England.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350–353.

Shoahosseini, R., Baghaei, P., Khodabakhshzadeh, H., & Ashraf, H. (2024). C-test construct validity: Evidence from nonparametric item response theory. *Language Testing in Asia, 14*(1), 10.

Sijtsma, K. (2012). Future of psychometrics: Critical discussion and perspectives on modern methods. *Psychometrika, 77*(4), 674–688. https://doi.org/10.1007/s11336-012-9283-y

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory.* Sage.

Smith, R. M. (2020). Item fit analysis in Rasch measurement: Practical implications for test development. *Measurement and Evaluation in Counseling and Development, 53*(2), 99–114. https://doi.org/10.1080/07481756.2019.1667245

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., & Hagen, E. P. (1991). *Measurement and evaluation in psychology and education.* Macmillan Publishing Co.