Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 12, No. 2, October 2022
ISSN 2476-5880

A Comparison of Polytomous Rasch Models for the Analysis of C-Tests

Safa Mohammed Abdulridah Dhyaaldian[1], Qasim Khlaif Kadhim[2]*, Dhameer A. Mutlak[3], Nour Raheem Neamah[4], Zaidoon Hussein Kareem[5], Doaa A. Hamad[6], Jassim Hassan Tuama[7], Mohammed Saad Qasim[8]

Received: March 2022 Accepted: May 2022

Abstract
A C-Test is a gap-filling test for measuring language competence in the first and second language. C-Tests are usually analyzed with polytomous Rasch models by considering each passage as a super-item or testlet. This strategy helps overcome the local dependence inherent in C-Test gaps. However, there is little research on the best polytomous Rasch model for C-Tests. In this study, the Rating Scale Model (RSM) and the Partial Credit Model (PCM) for analyzing C-Tests were compared. To this end, a C-Test composed of six passages with both RSM and PCM was analyzed. The models were compared in terms of overall fit, individual item fit, dimensionality, test targeting, and reliability. Findings showed that, although the PCM has a better overall fit compared to the RSM, both models produce similar test statistics. In light of the findings of the study, the choice of the best Rasch model for C-Tests will be discussed.

Keywords: C-test, local item dependence, partial credit model, rating scale model, unidimensionality

1. Introduction

As an integrative method of assessment, C-test is viewed as a kind of testing procedure for the operationalization of the reduced redundancy principle (RRP; Spolsky, 1969; Spolsky et al., 1968). C-test was originally developed by Klein-Braley and Raatz (1984) as an improvement over classic Cloze test (Klein-Braley, 1997). Similar to Cloze tests, C-tests are pragmatic language tests. As Oller (1979) argued:

1 University of Warith Al-Anbiyaa, Karbala, Iraq.
2 English Language Department, Al-Mustaqbal University College, Babylon, Iraq. Email: qasim.khlaif@mustaqbal-college.edu.iq
3 Al-Nisour University College, Baghdad, Iraq.
4 Al-Manara College for Medical Sciences (Maysan), Iraq.
5 Department of Medical Laboratory Technics, Al-Zahrawi University College, Karbala, Iraq.
6 Nursing Department, Hilla University College, Babylon, Iraq.
7 Altoosi University College, Najaf, Iraq.
8 Al-Esraa University College, Baghdad, Iraq.

> Any procedure or task that causes the learner to process sequences of elements in a language that conform to the normal contextual constraints of that language, and which requires the learner to relate sequences of linguistic elements via pragmatic mappings to extra-linguistic context. (p. 38)

In cloze tests, a word is completely deleted in every *n*th word whereas in C-tests the second half of every second word is omitted. The first and the last sentences in each passage are kept intact to provide test takers with adequate context to process the text. Test takers have to read and comprehend the passages and then rebuild the distorted words within each passage. For every correct word restored by test takers, one point is awarded, and the total score on each passage is considered as a global language ability of test takers. Because C-tests, as gap-filling tests, are easily administered within a short period of time and efficiently scored, they are widely used in various educational contexts, especially in large-scale educational assessment, as a means to monitor, place, and provide feedback and remedial teaching to test takers in terms of their language ability level (Eckes, 2010; Eckes & Baghaei, 2015; Lee-Ellis, 2009). C-tests have been developed in different languages, such as English, French, German, Japanese, Korean, Persian, Russian, and Turkish. Researchers have demonstrated that C-tests are reliable and valid indicators of general language ability in both first (L1) and second language (L2) (Raatz & Klein-Braley, 1981).

2. Review of Literature

A great deal of research has been conducted over the last several decades to examine the psychometric quality of C-tests using different quantitative methods, including correlational approaches (Arras et al., 2002; Boonsathorn, 1987; Chapelle & Abraham, 1990; Dörnyei & Katona, 1992; Farhady & Jamali, 1999; Jafarpur, 2002; Negishi, 1987) and factor analysis (Grotjahn, 1992; Grotjahn & Allner, 1996). These studies provided different kinds of strong reliability and validity evidence for C-tests (Norris, 2018; Sigott, 2004). The results of factor analyses showed that C-tests load on a single factor, e.g., a general language proficiency factor, in conjunction with other language proficiency tests (Fadaeipour & Zohoorian, 2017; Grotjahn & Allner, 1996; Grotjahn, 1992; Rasoli, 2021). Correlational analyses also showed that there is a significant correlation between C-tests and productive (speaking and writing) and receptive skills (listening and reading), ranging from 0.37 to 0.97.

Along the same lines, item response theory (IRT) models have been frequently used to analyze the structure of C-tests. IRT models are mathematical models which describe the relationship between latent traits and their manifestations with respect to one or more item parameters. All IRT models include three basic assumptions: (a) *unidimensionality* which indicates that all the items of an instrument should only measure a single latent trait; (b) *monotonicity* indicates that as the level of the latent trait increases, the probability of giving a correct response to a set of test items should increase as well; and (c) *local item dependence* (LID) indicates that items should be independent given a certain level of the expected latent trait (Baghaei, 2021). In fact, LID states that the items should be uncorrelated after conditioning out the effect of the latent

trait. When these assumptions are violated, (the latent variables and item) parameter estimates and test statistics will be biased and misleading, respectively (Wang & Wilson, 2005). Numerous researchers have investigated the application of IRT models on C-tests and demonstrated that the total raw scores of test takers can be utilized to locate them on an ordinal scale (Eckes & Grotjahn, 2006; Grotjahn, 1992; Grotjahn & Allner, 1996; Moosbruger & Mueller, 1982; Raatz, 1984).

Although gaps in C-tests are generally taken as individual items, Forthmann et al. (2020) argued that this method inaccurately increases the number of item parameters, and due to the structure of C-tests in which gaps are interdependent and nested within passages, the assumption of local item independence, as an important assumption of IRT and Rasch models, is violated and makes serious problems for analyzing C-tests with IRT models. When the local independence is infringed, item difficulty and item discriminating estimates would be biased and the precision of persons' ability estimates and reliability coefficients would be overestimated (Thissen, Steinberg, & Mooney, 1989; Yen & Fitzpatrick, 2006).

As a strategy to resolve the problem of LID in C-tests, researchers usually use the super-item (Grotjahn, 1987; Raatz, 1984) or item bundle approach (Rosenbaum, 1988). In this approach, all the gaps within each text are summed, and the scores are then entered into the IRT analysis. In other words, each text is viewed as a polytomous item with a number of categories, e.g., between 20 to 25 response categories. In this case, (ordinal) polytomous Rasch models such as the rating scale model (RSM; Andrich, 1978), the partial credit model (PCM; Masters, 1982), and the continuous rating scale model (CRSM; Müller, 1987) can be used to model and analyze C-tests, especially the assumption of local item independence. Over the past few years, several researchers have applied polytomous Rasch models to C-tests (Baghaei, 2008, 2011; Eckes, 2006, 2007, 2011; Lee-Ellis, 2009; Norris, 2006; Schroeders, Robitzsch, & Schipolowski, 2014). The results of these studies showed the effectiveness of polytomous Rasch models in modeling C-tests. For example, Eckes (2006) compared the performance of RSM, PCM, and CRSM to discover the suitability of Rasch models for analyzing C-tests. The results revealed that although the three modes produced highly similar item parameter estimates, the CRSM had the best performance compared to the PCM and the RSM, and the RSM outperformed the PCM. In another study, Eckes (2007) analyzed the performance of two polytomous Rasch models, including the RSM and CRSM. The comparison of the two models indicated that the RSM is the most appropriate model for constructing and evaluating C-tests. Despite the fact that the previous studies on the comparison of polytomous Rasch models provided invaluable insight into the appropriateness of such models for the analysis of C-tests, there is a paucity of research on the comparison of several polytomous Rasch models. In fact, it is still unclear which polytomous Rasch model can better describe the functioning of C-tests when the super-item approach is used. The present study attempts to address this gap by comparing the performance of the RSM and PCM, as two popular (ordinal) polytomous Rasch models, for analyzing C-tests scored based on the super-item approach.

The rating scale model (RSM; Andrich, 1978) and the partial credit model (PCM; Masters, 1982) are polytomous generalizations of the Rasch model (RM; Rasch, 1960/1980). Both models assume that, in polytomous items including several ordered categories (e.g., 'strongly disagree',

ISSN 2476-5880

'disagree', 'moderately agree', 'agree', and 'strongly agree'), the adjacent response options or categories are two dichotomous categories, similar to the dichotomous IRT models (Fischer, 1995). However, the models are different in terms of estimating category boundaries or thresholds. As Andrich (1978) argued, thresholds are the locations on the latent trait continuum where the probability of endorsing two adjacent categories is equal. RSM is suitable for instruments in which all the items have the same structural response format, that is, the latent trait level to exceed to endorse a category is the same across all the items. For that reason, one set of category thresholds is estimated for all the items, which have equal distances across the items.

On the other hand, the PCM assumes that polytomously scored items include multiple ordered response options. In the PCM, thresholds do not require to have the same order as response options, that is, a unique set of thresholds is estimated for each item. The RSM is generally considered a restricted form of PCM and unlike the RSM, thresholds in the PCM are not on the same scale. It should be further pointed out that despite the fact that the RSM and PCM are extensions of the RM, both models maintain the unique characteristics of the RM, including the sufficiency of raw scores, separate person and item parameters, and objective comparison of persons and items. The purpose of this research is to compare the performance of RSM and PCM for analyzing C-tests.

3. Method

3.1 Participants

A sample of 203 students (115 female) studying English at Al-Nisour University College, Baghdad took the C-Test. Their age ranged from 21 to 33 (M=23.92, SD=3.76). The cloze tests were administrated as a mid-term exam in a reading comprehension course in six parallel classes.

3.2 *Instrument*

A C-Test battery containing six independent passages was employed in this study. To construct the C-Test, reading compression passages from the British Council website were used (https://learnenglish.britishcouncil.org/). For the purpose of this study, three passages were selected from the B1 level and three passages were selected from the B2 level. Considering the fact that our target group is composed of lower intermediate and intermediate learners, texts from other levels were deemed inappropriate. To construct the C-Tests, the rule-of-2 (Raatz & Klein-Barely, 2002) was applied. That is, starting from the second sentence, the second half of every second word was deleted. In the case of words with an odd number of letters, the bigger half was deleted. For example, in a word with nine letters, the second five letters were deleted and in a word with seven letters, the second four letters were deleted. There were no deletions in the first and the last sentences to provide some context to help examinees process the texts. There were 20 gaps in each passage.

ISSN 2476-5880

3.3 Procedures

To compare the performance of RSM and PCM for the analysis of C-Tests, both models were applied to the data. As explained above, to solve the problem of local item dependence in C-Tests, each passage is considered as a polytomous item (super-item) with 21 response categories. Winsteps Rasch model computer program version 5.2.2 (Linacre, 2022) was used for the analyses.

4. Results and Discussion

Table 1 shows the item difficulty parameters and their infit and outfit mean square values in both models. Item difficulty estimates ranged from -.37 to .27 logits in the RSM and from -.24 to .17 in the PCM. This shows that item parameters have a wider spread in the RSM compared to the PCM. The fit of data to the Rasch model is an essential requirement to have the appealing properties promised by the model including interval scaling (Baghaei et al., 2017). Following Bond and Fox (2007), infit and outfit mean square (MNSQ) values lower than 1.30 are acceptable. The fit statistics show that the items have a good fit in both models. However, they tend to fit the PCM slightly better. Item 4 with an infit mean square value of 1.30 and an outfit means square value of 1.26 has a better fit in the PCM. The precision of the item parameter estimates is the same in both models as shown by their standard errors. The last column in Table 1 shows the point-measure correlations. They are the correlations between the performance of individuals on the items and their overall person measures in logits. They are equivalent to item-total correlations in classical test theory and is an indication of the relationship between the item and the entire scale. They are also a measure of item discrimination. Table 1 shows that all the point-measure correlations are very high which shows that the items are strongly related to the overall scale scores.

Table 1.
Item measures and fit statistics for the six C-Test passages across RSM and PCM

	RSM					PCM				
Item	Diff.	SE	Infit MNSQ	Outfit MNSQ	Pt. Meas. Cor.	Diff.	SE	Infit MNSQ	Outfit MNSQ	Pt. Meas. Cor.
1	.04	.03	1.13	1.11	.92	.14	.03	1.21	1.19	.91
2	-.37	.03	1.02	.98	.93	-.24	.03	.96	1.01	.92
3	-.11	.03	.88	.91	.94	-.09	.03	.75	.87	.93
4	-.03	.03	1.30	1.26	.91	-.13	.03	1.13	1.11	.91
5	.20	.03	1.03	.99	.92	.16	.03	1.11	1.05	.92
6	.27	.03	1.11	1.09	.92	.17	.03	1.10	1.11	.92

Note. Diff=Difficulty Parameter; SE=Standard Error; Pt. Meas. Cor. =Point-Measure Correlation

Table 2 shows the deviances (a global fit statistic equal to -2loglikelihood of the model) for RSM and PCM models. Models with smaller deviances have a better fit. The deviances show that the PCM has an overall better fit than the RSM. Principal components analysis of standardized

ISSN 2476-5880

residuals which is a method of evaluating unidimensionality and global model fit was examined in both models. The eigenvalue of the first contrast in the RSM was 1.6 and in the PCM was 1.4. Although both values are smaller than 2 and indicate unidimensionality (Linacre, 2022), PCM had a better overall fit which corroborates the result from comparing model deviances. Person and item separation values indicate the number of statistically different ability and difficulty strata that the test can identify in the sample and the sample can identify in the test (Wright, 1996). The minimum separation index is 2 (Linacre, 2022). A person separation of 5 means that the test has identified 5 different performance strata in the sample. An item separation of 7 means that the sample has identified 7 strata of difficulty in the instrument.

Table 2.
Global model fit and precision for the two models

Model	Deviance	Reliability	Person Separation	Item Separation	Mean (SD)	Range
RSM	14150.26	.96	5.08	7.28	.65 (1.19)	6.54
PCM	13647.28	.96	4.99	5.45	.51 (1.24)	7.35

Both models have the same reliability but RSM has produced better person and item separation indices. The person ability parameters from the RSM were slightly higher than those in the PCM and had a correlation of .999. The mean of person parameters in the RSM (M=.65) was slightly higher than in the PCM (M=.51 logits). The absolute differences between the person parameters from the two models ranged from 0 to .47 with a mean of .15 and a standard deviation of .06. If the PCM is considered as the "true" model (since it has a better fit), it means that the person parameters are overestimated in the RSM by as much as .47 logit and on average by .15 logit. The person parameter estimates have a higher range in the PCM than the person parameters in the RSM which is an indication of a wider spread that the partial credit model produces. This is a sign that the PCM better distinguishes the examinees.

4. Conclusion

C-Tests as integrative overall tests of proficiency in the first and second language are commonly used in research in the second language and as test instruments in large-scale assessments. Due to the interdependence of C-Test gaps (items) analyzing them with Rasch and IRT models is problematic. Therefore, researchers like Raatz (1984) suggested to compute the passage scores and consider each passage as a unit of analysis or a super item. Then a polytomous IRT model may be used to analyze the polytomous items or passages. Over the past decades, the Rasch rating scale model (Andrich, 1978) has been mostly used for the analysis of C-Tests (for

example, Baghaei, 2010, 2011, Eckes, 2011, Eckes & Grotjahn, 2006). Studies comparing the performance of different polytomous Rasch or IRT models for the analysis of C-Tests are scarce. To the best of our knowledge, only Eckes (2006) compared the performance of several polytomous Rasch models for C-tests. He compared the application and performance of RSM, PCM, and continuous rating scale model (CRSM, Mueller, 1987) for the C-Test. His findings showed that the CRSM performed better than the RSM and the PCM, and the RSM performed better than the PCM. He also showed that all the models produced highly correlated person parameter estimates. The item parameters from the three models were also very similar. He also examined the invariance of item parameters across subsets of examinees and found that the CRSM produces the most invariant results while RSM produces the least invariant results. Infit and outfit statistics were very similar across the three models. He also reported that the items had the same precision across the three models.

In this study, the performance of RSM and PCM for analyzing C-tests was compared. Our findings showed that although both models yield comparable results, the PCM has better global and local items fit values. However, there are many more parameters in the PCM and large sample sizes are required to precisely estimate them. The findings of our study partly agree with those of Eckes (2006) as it was also found that person parameters are highly correlated and the precision of item parameters is similar in the three models. Our findings diverge from those of Eckes as it was found found that PCM is a better model than the RSM. The invariance of item parameters across subsamples was not checked because the sample size was relatively small, and partitioning a small sample leads to inaccurate parameter estimates. But the item infit and outfit values showed a better fit for the PCM. Besides, the global fit as indicated by -2loglikelihood of the models showed that the PCM has a better fit. The results also revealed highly correlated person parameters across RSM and PCM, but Eckes (2006) did not check the difference between individual parameters across the models. Our findings showed that person parameters in the RSM are slightly higher which could be an overestimation since the PCM has a better fit.

Another comparable study is Baghaei (2010) in which he compared the performance of RSM, PCM and the equidistant model (Andrich, 1982) for a reading comprehension test composed of six independent passages each containing six dichotomous items. The dichotomous items nested within each passage were summed and each passage was entered into the analysis as a 7-category rating scale (0 to 6). His findings showed that the three models had very similar outputs in terms of item fit, reliability, and precision of parameter estimates. The information criteria AIC and BIC did not agree on the best fitting model. While according to the AIC, PCM was the best fitting model, based on BIC, RSM was the best fitting model and the PCM was the worst.

Future research should examine the fit of other polytomous IRT models such as the graded response model (Samejima, 1969) and continuous response model (Samejima, 1973) for scaling C-tests. Rasch Poisson Counts Model (Baghaei & Doebler, 2019) should also be examined for speeded C-Tests. Multidimensional Rasch models including higher-order model, testlet model, and bifactor models should also be examined for modeling LID in C-tests (Baghaei, 2013; Baghaei & Ravand 2016, 2019). The linear logistic test model (Fischer, 1973) can be used to examine

components of difficulty in C-tests and enter LID as a factor of difficulty in the Q-matrix (Baghaei & Hohensinn, 2017).

References

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

Andrich, D. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters, *Psychometrika, 47*(1), 105-113.

Arras, U., Eckes, T., & Grotjahn, R. (2002). C-Tests im Rahmen des 'Test Deutsch als Fremdsprache' (TestDaF): Erste Forschungsergebnisse [C-tests within the 'Test of German as a foreign language' (TestDaF): Preliminary research findings]. In R. Grotjahn (Ed.), *Der C-Test: Theoretische Grundlagen und praktische Anwendungen [The C-test: theoretical foundations and practical applications]* (Volume 4, pp. 175-209). Bochum: AKS-Verlag.

Baghaei, P. (2021). *Mokken scale analysis.* Muenster: Waxmann.

Baghaei, P., & Ravand, H. (2019). Method bias in cloze tests as reading comprehension measures. *Sage Open, 9*, 1-8.

Baghaei, P. & Doebler, P. (2019). Introduction to the Rasch Poisson Counts Model: An R tutorial. *Psychological Reports, 122* (5), 1967-1994.

Baghaei, P., Yanagida, T., & Heene, M. (2017). Development of a descriptive fit statistic for the Rasch model. *North American Journal of Psychology, 19*, 155-168.

Baghaei, P., & Hohensinn, C. (2017). A method of Q-matrix validation for the linear logistic test model. *Frontiers in Psychology, 8*, 897.

Baghaei, P., & Ravand, H. (2016). Modeling local item dependence in cloze and reading comprehension test items using testlet response theory. *Psicológica, 37*, 85-104.

Baghaei, P. (2013). Development and psychometric evaluation of a multidimensional scale of willingness to communicate in a foreign language. *European Journal of Psychology of Education, 28*, 1087-1103.

Baghaei, P. (2010). A comparison of three polychotomous Rasch models for super-item analysis. *Psychological Test and Assessment Modeling, 52*, 313-323.

Baghaei, P. (2008). The effects of the rhetorical organization of texts on the C-test construct: A Rasch modeling study. *Melbourne Papers in Language and Testing, 13*, 32-51.

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

Bond, T. G., & Fox, C. M. (2007) (2nd Ed.) *Applying the Rasch model: Fundamental measurement in the human sciences*. New York: Lawrence Erlbaum.

Boonsathorn, S. (1987). *C-Tests, proficiency, and reading strategies in ESL* (Unpublished doctoral dissertation). University of Alberta, Canada.

Chapelle, C. A., & Abraham, R. G. (1990). Cloze method: What difference does it make? *Language Testing, 7*(2), 121-146.

Dörnyei, Z., & Katona, L. (1992). Validation of the C-test amongst Hungarian EFL learners. *Language Testing, 9*(2), 187-206.

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

Eckes, T. (2007). Konstruktion und Analyse von C-Tests mit Ratingskalen-Rasch-Modellen [Construction and analysis of C-tests with rating scale Rasch models]. *Diagnostica, 53*, 68-82.

Eckes, T. (2010). Der Online-Einstufungstest Deutsch als Fremdsprache (onDaF): Theoretische Grundlagen, Konstruktion und Validierung [The online placement test of German as a foreign language: Theoretical foundations, construction, and validation]. In R. Grotjahn (Ed.), *Der C-Test: Beiträge aus der aktuellen Forschung [The C-test: Contributions from current research]* (pp. 125-192). Frankfurt, Germany: Lang.

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

Eckes, T., & Baghaei, P. (2015). Using testlet response theory to examine local dependence in C-tests. *Applied Measurement in Education, 28*(2), 85-98.

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

Fadaeipour, A., & Zohoorian, Z. (2017). Comparing the psychometric characteristics of speeded and standard C-Tests. *International Journal of Language Testing, 7,* 40-50.

Farhady, H., & Jamali, F. (1999). Varieties of C-test as measures of general proficiency. *Journal of the Faculty of Foreign Languages,* 3, 23-42.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374.

Fischer, G. H. (1995). The derivation of polytomous Rasch models. In G. H. Fischer, & I. W. Molenaar (Eds.), *Rasch Models: Foundations, recent developments, and applications* (293-305). Springer: New York, NY.

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*(6), 692-705.

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.

Grotjahn, R. (1992). Der C-Test im Französischen. Quantitative Analy-sen. In R. Grotjahn (Ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Vol. 1, pp. 205-255). Bochum: Brockmeyer.

Grotjahn, R., & Allner, B. (1996). Der C-Test in der Sprachlichen Aufnahmeprüfung an Studienkollegs für ausländische Studierende an Universi-täten in Nordrhein-Westfalen. In

Rüdiger G. (Ed.), *Der C-Test. Theoreti-sche Grundlagen und praktische Anwendungen* (Vol. 3, pp. 279-342). Bochum: Brockmeyer.

Hohensinn, C., & Baghaei, P. (2017). Does the position of response options in multiple choice tests matter? *Psicológica, 38,* 93-109.

Jafarpur, A. (2002). A comparative study of a C-Test and a cloze test. In R. Grotjahn (Ed.), *Der C-Test: theoretische Grundlagen und praktische Anwendungen [The C-test: Theoretical foundations and practical applications]* (Volume 4., pp. 31-51). Bochum: AKS-Verlag.

Klein-Braley, C. (1997). C-tests in the context of reduced redundancy testing: An appraisal. *Language Testing*, *14*(1), 47-84.

Klein-Braley, C., & Raatz, U. (1984). A survey of research on the C-test. *Language Testing, 1*(2), 134-146.

Lee-Ellis, S. (2009). The development and validation of a Korean C-test using Rasch analysis. *Language Testing*, *26*(2), 245-274.

Linacre, J. M. (2022). *Winsteps® Rasch measurement computer program (Version 5.2.2)*. Portland, Oregon: Winsteps.com.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

Moosbrugger, H., & Muller, H. (1982). A classical latent additive test model (CLA model). *German Journal of Psychology*, *6*(2), 145-149.

Müller, H. (1987). A Rasch model for continuous ratings. *Psychometrika*, *52*, 165-181.

Negishi, M. (1987). The C-test: An integrative measure? *IRLT Bulletin, 1*, 3-26.

Norris, J. M. (2006). Development and evaluation of a curriculum-based German C-test for placement purposes. In R. Grotjahn (Ed.), *Der C-Test: Beiträge aus der aktuellen*

Forschung [The C-test: Theory, empirical research, applications] (pp. 45-83). Frankfurt, Germany: Lang.

Norris, J. M. (Ed.). (2018). *Developing C-tests for estimating proficiency in foreign language research*. Frankfurt am Main: Peter Lang.

Oller, J. W. Jr. (1979). *Language tests at school: A pragmatic approach*. London: Longman.

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

Raatz, U., & Klein-Braley, C. (1981). 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* (pp. 113–145). Colchester, UK: University of Essex.

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). Bochum: AKS-Verlag.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests* (expanded Ed.). Chicago, IL: University of Chicago Press.

Rasoli, M. K. (2021). Validation of C-test among Afghan students of English as a foreign language. *International Journal of Language Testing, 11*(2), 109-121.

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika, 53*, 349-359.

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika, 38*(2), 203-219.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika,* Monograph Supplement No. 17.

Schroeders, U., Robitzsch, A., & Schipolowski, S. (2014). A comparison of different psychometric approaches to modeling testlet structures: An example with C-tests. *Journal of Educational Measurement, 51*(4), 400-418.

Sigott, G. (2004). *Towards identifying the C-test construct*. Frankfurt, Germany: Lang.

Spolsky, B. (1969). *Reduced redundancy as a language testing tool* (pp. 1-18). Presented at the Second International Congress of Applied Linguistics, Cambridge, England, September, 8-12.

Spolsky, B., Bengt, S. M., Sato, E. W., & Aterburn, C. (1968). Preliminary studies in the development of techniques for testing overall second language proficiency. *Language Learning, 18*(3), 79-101.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: a use of multiple categorical-response models. *Journal of Educational Measurement, 26*(3), 247-260.

Wang, W.-C., & Wilson, M. (2005). Assessment of differential item functioning in testlet-based items using the Rasch testlet model. *Educational and Psychological Measurement, 65*(4), 549-576.

Wright, B. D. (1996). Reliability and separation. *Rasch Measurement Transactions, 9*(4), 472.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111-153). Westport, CT: American Council on Education/Praeger.