Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 14, No. 2; 2024, October

ISSN 2476-5880
https://www.ijlt.ir

Modeling Local Item Dependence in Cloze Tests with the Rasch Model: Applying a New Strategy

Barno S. Abdullaeva[1], Diyorjon Abdullaev[2], Nurislom I. Khursanov[3], Khurshida B. Kadirova[4], Laylo Djuraeva[5,6]

ARTICLE INFO

Article History:
Received: January 2024
Accepted: March 2024

KEYWORDS
Cloze test
Local independence
Partial credit model
Rasch model

ABSTRACT

Cloze tests are commonly used in language testing as a quick measure of overall language ability or reading comprehension. A problem for the analysis of cloze tests with item response theory models is that cloze test items are locally dependent. This leads to the violation of the conditional or local independence assumption of IRT models. In this study, a new modeling strategy is suggested to circumvent the problem of local item dependence in cloze tests. This strategy involves identifying locally dependent items in the first step and combining them into polytomous items in the second step. Finally, the partial credit model is applied to the combination of dichotomous and polytomous items. Our findings showed that the new strategy results in a better model-data fit than the dichotomous model where dependence is ignored but with a lower reliability. Results also indicated that the person and item parameters from the two models highly correlate. The findings are discussed in light of the literature on managing local dependence on educational tests.

1. Introduction

A cloze test is a type of language assessment or evaluation where a portion of a text is removed, and participants are required to fill in the blanks with appropriate words (Oller, 1979). The goal of a cloze test is to measure a person's understanding of context, grammar, vocabulary, and reading comprehension. The removed words are typically every *nth* word, where *n* depends on the difficulty of the test. Participants need to use their language comprehension skills to infer the missing words based on the surrounding context. Cloze tests are commonly used to assess learners' reading and overall language proficiency (Hughes & Hughes, 2020).

The cloze test has its roots in the research and work of Taylor (1953), who introduced the concept. Taylor developed the cloze procedure as a way to assess text readability Taylor's idea was to create a test format where words were systematically deleted from a passage, and participants were required to fill in the blanks with the most appropriate words. The deleted words were often every fifth or tenth word, depending on the desired difficulty level. The objective was to measure the difficulty of a text by

[1] Professor of Pedagogical Sciences, Vice-Rector for Scientific Affairs, Tashkent State Pedagogical University, Tashkent, Uzbekistan.
[2] Department of Scientific Affairs, Innovations and Training of Scientific Pedagogical Personnel, Urganch State Pedagogical Institute, Urganch, Uzbekistan.
[3] Doctor of Philosophy (Ph.D) in Philological Sciences, Department of Pedagogy and Philology, Renaissance University of Education, Tashkent, Uzbekistan.
[4] Assistant Professor, Department of Teaching the Uzbek Language in Foreign Language Groups, Alisher Navo'i Tashkent State University of the Uzbek Language and Literature, Tashkent, Uzbekistan.
[5] Ph.D, Philosophy Sciences, Department of Innovation and Sciences, New Uzbekistan University, Tashkent, Uzbekistan.
[6] Department of Science and Innovation, Tashkent State Pedagogical University named after Nizami, Tashkent, Uzbekistan.

Cite this paper as: Abdullaeva, B., Abdullaev, D., Khursanov, N. I., Kadirova, B. K., & Djuraeva, L. (2024). Modeling Local Item Dependence in Cloze Tests with the Rasch Model: Applying a New Strategy. *International Journal of Language Testing*, *14*(2), 97–103. https://doi.org/ 10.22034/IJLT.2024.435779.1319

evaluating readers' ability to use context clues and linguistic knowledge to infer missing words and comprehend the overall meaning of a text. Later on, the cloze test gained popularity as an effective tool for evaluating language proficiency and reading comprehension skills (Alderson, 2000). It was widely adopted in educational research and language assessment, becoming a standard feature in language testing, particularly in the evaluation of English as a second language (ESL) learners.

2. Review of Literature

In recent decades, cloze tests have gained widespread use in both large-scale assessments and classroom evaluations, serving as indicators of general language proficiency and reading comprehension (Abraham & Chapelle, 1992; Alderson, 1980). Various studies and researchers have presented diverse forms of validity evidence for cloze tests. Through correlational analyses, it has been consistently demonstrated that cloze tests exhibit correlations with other language proficiency assessments, encompassing reading, writing, speaking, listening, vocabulary, and grammar (Oller, 1983).

Criterion validity evidence has further substantiated the validity of cloze tests, revealing substantial coefficients ranging from .71 to .89 when compared to standardized ESL proficiency tests (Brown, 2013). A recent validation study among Iraqi university learners of English as a foreign language (EFL) by Sattar (2022) reported notable correlation coefficients between cloze tests and assessments of grammar (r = .70), vocabulary (r = .60), reading comprehension (r = .68), and the combined score of vocabulary, grammar, and reading comprehension (r = .78).

In a study conducted by Zare and Boori (2018), a robust correlation of .81 was found between cloze tests and reading comprehension, whereas Yazdinejad and Zeraatpishe (2019) reported a moderate correlation of .48 between reading comprehension and cloze tests. They attributed the relatively low correlation to the diminished reliability of their tests. Upon correcting the correlation coefficient for attenuation, the correlation increased to .81. Additionally, reliability coefficients for cloze tests have consistently shown high magnitudes, ranging from .80 to .90 (Brown, 1983, 2013).

Studies utilizing exploratory factor analysis have indicated that cloze tests are associated with a broader factor of general language proficiency, alongside other specific language ability skills (Oller, 1983; Sattar, 2022). Researchers have sought to determine whether cloze tests assess language competence beyond mere knowledge of sentence-level grammatical structures, delving into the macro-level textual competence of examinees. For instance, Ramanauskas (1972) conducted a comparison of native English speakers' performance on a cloze test with sentences in their original order versus a version with randomly rearranged sentences. The subjects demonstrated significantly better performance in the cloze test with intact sentences.

Chihara et al., (1977) replicated Ramanauskas' study with both native and non-native English speakers, finding that both groups performed better on passages with intact sentences. Oller (1975) demonstrated that as the contextual information around gaps increased from 5 to 50 words, the average scores on cloze items for native English speakers also increased. These outcomes suggest that cloze tests are responsive to linguistic contexts extending beyond a single sentence, indicating that they measure macro-level language abilities. In essence, cloze tests go beyond being mere assessments of micro-level grammar and vocabulary; they tap into higher-order skills that operate beyond individual sentences. These findings also contribute to the construct validity of cloze tests by elucidating what the tests actually measure and the underlying abilities influencing performance on cloze items.

Despite the accumulation of various types of validity evidence for cloze tests, there has been a lack of evidence based on the fit to item response theory (IRT) models. This absence stems from the interdependence of cloze items, where gaps are nested within passages, violating the local independence assumption of IRT models (Baghaei & Ravand, 2016, 2019). IRT models hinge on two fundamental assumptions: unidimensionality and local independence (Hambleton & Swaminathan, 1985). Unidimensionality posits that all items should measure a single latent trait, while local independence asserts that, once the impact of the latent trait is accounted for, the items of a test should be uncorrelated. If items remain correlated after removing the latent trait's influence, it suggests that the test measures an additional dimension irrelevant to the intended measurement, constituting a violation of unidimensionality and a piece of evidence against validity.

Following the C-Test literature (Eckes & Baghaei, 2015), to make cloze tests analyzable with IRT models, i.e., to circumvent the LID problem, Dhyaaldian et al. (2022) suggested using several cloze

passages and consider each passage as a polytomous super-item. This strategy allows using polytomous IRT models for the analysis of cloze tests. Using this strategy, the problem of the dependence of gaps does not arise. This modeling strategy has some drawbacks, though. First, in large-scale studies where numerous items and skills are tested, using several cloze passages is very difficult, if not impossible. Tests usually contain only one longer cloze test instead of several shorter cloze tests. Including several cloze tests in an assessment, takes the time of other skills and test types and has little practicality. The other problem with this method is that information about individual gaps or items is lost and information will be available only about passages (Forthmann et al., 2019). Baghaei and Christensen (2023) used the loglinear Rasch model to account for local dependency in a C-Test which is a variation of the cloze test. Their findings showed that when LID is accounted for the model fits better.

In this current study, we follow Baghaei and Effatpanah (2023) who suggested an alternative strategy for modeling LID in C-Tests using the Rasch model. The Rasch model is extensively used in second and foreign language data analysis (see Afsharrad, 2023; Askari & Tabatabaee-Yazdi, 2023; Effatpanah & Baghaei, 2024). We employ the same strategy for the problem of LID in cloze tests. Baghaei and Effatpanah stated that instead of considering each passage as a super item, we can identify the items which have local dependence first and then only combine the items which exhibit LID into polytomies. In other words, instead of assuming that all the gaps within passages are locally dependent a priori, we test this assumption and aggregate only those gaps which are empirically shown to be dependent. The purpose of the present study is to apply this method to a cloze test and compare the results with the analysis where items are assumed to be independent.

3. Method

3.1 Participants

Participants were 256 EFL students at the Department of Pedagogy and Philology, Renaissance University of Education in Tashkent, Uzbekistan. Out of these students, 161 were female and 95 were male. The average age of the participants was 22.57 with SD = 3.94.

3.2 Instrument

An English cloze test containing 30 blanks was given to the participants as a section of their final exam in a reading comprehension course. The test was given alongside other multiple-choice reading comprehension test items. The passage contained 243 words where every 7th word was deleted (proper nouns were not deleted). A sentence at the beginning and two sentences at the end of the passage remained intact to provide some contextual clues.

3.3 Procedure

The cloze test was administered to the participants and the data of the 30 cloze items were analyzed with the Rasch unidimensional model using the Winsteps program (Linacre, 2022a). Locally dependent items were identified by checking the correlations between their residuals (Linacre, 2022b). Correlations above .20 were considered as indicative of LID (Christensen et al., 2017). In the following step, locally dependent items were combined into polytomous items, and the Rasch model was applied to the combination of dichotomous and polytomous items. The two models were compared with regard to fit, reliability, and unidimensionality.

4. Results

Table 1 shows the fit statistics and item parameter estimates for the 30 cloze items. This analysis ignores LID and treats items as locally independent. Properties of the Rasch model can be obtained if the data fit the model (Baghaei et al., 2017). As can be seen in Table 1, Items 1, 7, 14, 15, 23, 25, and 27 had large infit or outfit mean square values greater than 1.30 (Bond et al., 2020) and did not fit the Rasch model. Furthermore, there were several other items with acceptable fit values but with very small point-measure correlations, including Items 13, 16, and 19. Item 19 has a very small outfit mean square value and overfits the Rasch model. Overfitting items have less variation than expected by the model and are generally benign. The Rasch separation reliability was .68, while Cronbach's alpha reliability of the raw scores was .86.

Table 1
Item Parameters and Fit Statistics for the 30 Cloze Items

Item entry	Measure	S.E.	Infit MNSQ	Outfit MNSQ	Pt-measure cor.
1	1.80	.24	1.20	1.46	.29
2	-.33	.25	1.01	.89	.36
3	.54	.25	.84	.77	.56
4	1.13	.25	.87	.82	.56
5	-.68	38	.82	.80	.49
6	.65	.29	.98	1.10	.45
7	-2.10	.47	.97	2.97	.12
8	.38	.26	1.20	1.12	.30
9	-2.19	.43	.97	.88	.21
10	-.34	.29	1.13	1.14	.23
11	-.77	.29	1.04	.86	.31
12	1.81	.27	.95	.94	.51
13	-3.19	.72	1.03	.81	.09
14	.11	.24	.99	1.64	.35
15	-3.33	.72	1.04	1.62	.03
16	6.28	1.02	1.04	1.21	.04
17	.16	.26	.86	.73	.53
18	.52	.26	.92	.94	.50
19	-3.96	1.01	.99	.37	.15
20	-.09	.24	.97	.82	.44
21	-2.19	.43	.99	.62	.30
22	.20	.32	.98	.92	.39
23	.92	.28	1.33	1.44	.18
24	61	.28	.84	.76	.55
25	2.86	.32	1.26	1.98	.12
26	5.02	1.07	1.08	.44	.26
27	-4.02	1.01	1.04	2.58	-.04
28	.02	.24	.91	.79	.49
29	-1.99	.40	.99	.96	.20
30	2.17	.26	.90	.64	.63

Note: SE: standard error, MNSQ: mean square, pt-measure cor.; point-measure correlation

In the next step, LID was evaluated by examining residual correlations. Examination of residual correlations showed that five item pairs have large correlations, indicating local dependence (Linacre, 2022b). Table 2 shows the item pairs and their magnitudes of correlations. Note that three of the items with high fit statistics and low point-measure correlations were among the items with local dependence.

Table 2
Items with the Highest Residual Correlations

Item	Item	Residual Cor.
15	27	.88
9	29	.87
20	28	.72
21	29	.40
13	21	.36

Locally dependent items were combined into polytomous items. Item pairs 15/27 and 20/28 were combined into two separate 3-point items. The remaining three item pairs, containing Items 9, 13, 21, and 29, were combined into a 5-point polytomous item. This was done because these item pairs had one shared item which connected them together. This scheme resulted in a combination of 22 dichotomous

and three polytomous items. The resulting 25 items were analyzed again with Rasch partial credit model (Masters, 1982). Table 3 shows the item parameters and their fit values.

Table 3
Item Measures and Fit Statistics after Merging Locally Dependent Items

Item entry	Measure	SE	Infit MNSQ	Outfit MNSQ	PT measure correlation
1	1.22	.24	1.15	1.22	.32
2	-.87	.24	.96	.84	.40
3	-.02	.25	.82	.75	.57
4	.55	.25	.87	.82	.55
5	-1.23	.37	.81	.76	.48
6	.09	.28	.95	1.05	.46
7	-2.61	.47	1.02	2.66	.07
8	-.17	.26	1.14	1.05	.32
10	-.91	.29	1.11	1.09	.23
11	-1.30	.29	1.01	.83	.31
12	1.23	.27	.93	.92	.51
14	-.45	.23	.97	1.56	.34
16	5.66	1.02	1.04	1.23	.04
17	-.39	.25	.85	.73	.53
18	-.05	.26	.93	.95	.47
19	-4.43	1.01	.98	.40	.14
22	-.40	.32	.95	.89	.40
23	.35	.28	1.30	1.39	.17
24	.03	.28	.84	.76	.55
25	2.26	.32	1.23	1.87	.12
26	4.40	1.06	1.08	.45	.26
30	1.57	.26	.80	.65	.63
31	-2.46	.46	1.08	4.36	.01
32	-.32	.14	1.06	.94	.51
33	-1.75	.22	1.05	1.37	.20

As Table 3 shows, Items 7, 14, 23, 25, the combination of Items 15/27, and the combination of Items 9/13/21/29 have large infit or outfit mean square values, and Item 16, which has acceptable fit values, has a very low point-measure correlation. Examination of correlation residuals shows that none of the items have large correlation residuals which indicates that the assumption of conditional independence holds for the data in the second analysis. If an analyst decides to delete the misfitting items, this decision leads to the deletion of locally dependent items as well. In other words, LID is a form of misfit, and items that violate conditional independence assumption should be deleted along with those which have unacceptable infit and outfit statistics. This means that many of the dependent items should be deleted because of poor infit and outfit statistics, if fit to the Rasch model is desired. Our findings show that by combining the items, we managed to save only item pairs 20/29, and the rest of the dependent items had to be deleted due to poor fit after merging.

Evaluation of dimensionality by means of principal components analysis of residual showed that the second analysis where locally dependent items were merged is closer to unidimensionality. The strength of the first contrast in the analysis where LID was ignored was 3.2 which is greater than the maximum value of 2 (Linacre, 2002b), while in the second analysis where locally dependent items were merged, this value dropped to 2.6. Comparison of global fit statistics, i.e., deviance (-2loglikelihood) also showed that the model that addresses LID by combining items has a better fit. The model which ignores LID had a deviance of 1909.44, while the model that accounts for LID had a deviance of 1742.86. A lower deviance is an indication of a better model fit. Rasch person separation reliability for the second analysis was .66 which was slightly smaller than the analysis where LID was ignored. The

higher reliability in the former analysis was due to ignoring LID and high item dependence which spuriously increases reliability (Zenisky et al., 2002).

5. Discussion

The study addresses a significant issue in the analysis of cloze tests, a commonly used tool in language testing. Cloze tests are frequently used to assess overall language proficiency or reading comprehension. However, the local dependence among cloze test items poses a challenge when applying IRT models. The violation of the conditional or local independence assumption in IRT models can compromise the accuracy of the results and the validity of the inferences drawn from the test.

To overcome this challenge, following Baghaei and Effatpanah (2023), the study proposed a novel modeling strategy. This strategy involves a two-step process: first, identifying the locally dependent items, and second, combining them into polytomous items. The application of a partial credit model to the combined set of dichotomous and polytomous items is the final step in this new approach.

The findings of the study reveal that the new modeling strategy yields a better fit between the model and the data, addressing the issue of local item dependence. Results also showed that when the dependent items are combined, reliability drops. This is due to the disappearance of the undue dependence between the items. When LID is modeled, reliability drops and gets closer to the true reliability. This improvement is a significant contribution to the field of language testing, i.e., closing the gap between the empirical reliability and actual reliability in assessments based on cloze tests. The identification and handling of locally dependent items through polytomous item combinations demonstrate the efficacy of the proposed strategy in mitigating the challenges associated with cloze test analysis.

6. Conclusion

This study introduced a new modeling strategy to address the issue of local item dependence in cloze tests when applying IRT models. The two-step process of identifying and combining locally dependent items into polytomous items, followed by the application of a partial credit model, improves the fit between the model and the data. This advancement is crucial for enhancing the accuracy and validity of language assessments that utilize cloze tests.

Because the application of testlet response theory (Bradlow et al., 1999) models is complex and requires large sample sizes, the suggested strategy in this study is an effective and simple solution to overcome the LID problem in cloze tests and other tests which have LID. Future researchers can explore ways to optimize the balance between model fit and reliability or propose alternative strategies that address the challenges posed by LID in cloze tests. Overall, this study contributes significantly to the refinement of assessment methodologies in language testing, providing a valuable framework for handling local item dependence in the analysis of cloze test data. Future studies should also consider other models such as the loglinear Rasch model for modeling LID in cloze tests (Baghaei & Christensen, 2023)

References

Abraham, R. G., & Chapelle, C. A. (1992). The meaning of cloze test scores: An item difficulty perspective. *Modern Language Journal, 76*, 468–479. https://doi.org/10.1111/j.1540-4781.1992.tb05394.x

Afsharrad, M. (2023). A Rasch model analysis of the Persian translation of the EFL listening strategy inventory. *Educational Methods & Practice, 1*, 1–14.

Alderson, J. C. (2000). *Assessing reading*. Cambridge University Press.

Alderson, J. C. (1980). Native and non-native speaker performance on cloze tests. *Language Learning, 30*, 59–76. https://doi.org/10.1111/j.1467-1770.1980.tb00151.x

Askari, A., & Tabatabaee-Yazdi, M. (2023). The development and validation of an inventory to measure EFL teachers' collegiality using item response theory. *Educational Methods & Practice, 1*, 1–16.

Baghaei, P., & Effatpanah, F. (2023). An alternative strategy for modeling local item dependence in C-Tests. In Dobrić, N., Cesnik, H., & Harsch, C. (Eds.), *Festschrift in honour of Günther Sigott: Advanced methods in language testing* (pp. 153–163). Peter Lang.

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., & Ravand, H. (2019). Method bias in cloze tests as reading comprehension measures. *Sage Open, 9*(1). doi: 10.1177/2158244019832706

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

Bond, T. G., Yan, Z., Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th Ed.). Routledge.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153–168. https://doi.org/10.1007/bf02294533

Brown, J. D. (2013). My twenty-five years of cloze testing research: So what? *International Journal of Language Studies, 7*(1), 1–32.

Chihara, T. J., Oller, J. W., Weaver, K., & Chavez-Oller, M. (1977). Are cloze items sensitive to constraints across sentences? *Language Learning, 27*(1), 63–73. https://doi.org/10.1111/j.1467-1770.1977.tb00292.x

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q_3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement*, *41*(3), 178–194. https://doi.org/10.1177/0146621616677520

Dhyaaldian, S. M. A., Al-Zubaidi, S. H., Mutlak, D. A., Neamah, N. R., Albeer, M. A., Hamad, D. A., Al Hasani, S. F., Jaber, M. M., & Maabreh, H. G. (2022). Psychometric evaluation of cloze tests with the Rasch model. *International Journal of Language Testing, 12*, 95–106. doi: 10.22034/IJLT.2022.157127

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

Effatpanah, F., & Baghaei, P. (2024). Examining the dimensionality of linguistic features in L2 writing using the Rasch measurement model. *Educational Methods & Practice, 2*, 1–22.

Forthmann, B., Grotjahn, R., Doebler, P., & Baghaei, P. (2019). 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

Hughes, A., & Hughes, J. (2020). *Testing for language teachers* (3rd Ed.). Cambridge University Press.

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

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-–174. https://doi.org/10.1007/bf02296272

Oller, J. W., Jr. (1983). Evidence for a general language proficiency factor: An expectancy grammar. In J. W. Oller, Jr. (Ed.), *Issues in language testing research* (pp. 3–10). Newbury House.

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

Oller, J. W. Jr. (1975). Cloze, discourse, and approximations to English. In M. K. Burt & H. C. Dulay, H. C. (Eds.), *New directions in TESOL* (pp. 345–356). TESOL.

Sattar, A. (2022). Validation of the cloze test as an overall measure of English language proficiency among Iraqi EFL learners. *North American Journal of Psychology, 23*(1), 147–154.

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly, 30*, 415–453. https://doi.org/10.1177/107769905303000401

Yazdinejad, A., & Zeraatpishe, M. (2019). Investigating the validity of partial dictation as a test of overall language proficiency. *International Journal of Language Testing, 9*, 44–56.

Zare, S., & Boori, A. A. (2018). Psychometric evaluation of the speeded cloze elide test as a general test of proficiency in English as a foreign language. *International Journal of Language Testing, 8*, 33–43.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2002). Identification and evaluation of local item dependencies in the Medical College Admissions Test. *Journal of Educational Measurement, 39*(4), 291–309. https://doi.org/10.1111/j.1745-3984.2002.tb01144.x