Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 14, No. 1; 2024, March

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

Examining Local Item Dependence in a Cloze Test with the Rasch Model

Diyorjon Abdullaev[1]*, Djuraeva Laylo Shukhratovna[2], Jamoldinova Odinaxon Rasulovna[3], Jumanazarov Umid Umirzakovich[4], Olga V. Staroverova[5]

ARTICLE INFO

Article History:
Received: July 2023
Accepted: August 2023

KEYWORDS
Cloze test
Item fit
Local item dependence
Rasch model
Residual correlation

ABSTRACT

Local item dependence (LID) refers to the situation where responses to items in a test or questionnaire are influenced by responses to other items in the test. This could be due to shared prompts, item content similarity, and deficiencies in item construction. LID due to a shared prompt is highly probable in cloze tests where items are nested within a passage. The purpose of this research is to examine the occurrence and magnitude of LID in a cloze test. A cloze test was analyzed with the Rasch model and locally dependent items were identified with the residual correlations. Findings showed that three pairs of items were locally dependent. When these items were removed from the analysis, test reliability dropped but item fit and unidimensionality improved. Removing the three locally dependent items did not affect person ability mean and standard deviation, though. The findings are discussed in terms of LID detection and modeling in the context of cloze test and language testing.

1. Introduction

Local item dependence (LID) refers to the phenomenon in which the responses to certain items in a test or questionnaire are influenced by the responses to other items, even after controlling for the underlying construct being measured. In other words, it means that the likelihood of endorsing a particular item is not solely determined by an individual's level on the construct being measured, but also by their response pattern on other items (Yen, 1984).

LID can occur due to various reasons, such as item wording, context effects, or shared response biases. It can lead to inflated estimates of the reliability and validity of a test or questionnaire if not properly accounted for. Therefore, it is important to assess and address local item dependence when analyzing and interpreting data from tests or questionnaires. The consequences of LID can include:

1. Inflated reliability: Local item dependence can lead to artificially high estimates of internal consistency reliability (e.g., Cronbach's alpha) because items are interrelated. This can give a false sense of measurement precision and reliability.

*[1] Doctor of Science, Associate Professor, Department of Uzbek History, Tashkent State Pedagogical University (Nizami), Bunyodkor Street 27, Tashkent, Uzbekistan.
[2]Ph.D, Philosophy Sciences, Department of Social Sciences, New Uzbekistan University, Tashkent, Uzbekistan.
[3]DSc, Professor, Vice-Rector for Research and Innovation, Alisher Navo'i Tashkent State University of Uzbek Language and Literature.
[4]Associate Professor, Doctor of Pedagogical Sciences, Deputy Dean of the Faculty of Foreign Languages for Academic Affairs, Jizzakh State Pedagogical University, Jizzakh, Uzbekistan.
[5]Plekhanov Russian University of Economics, Moscow, Russia.

Cite this paper as: Abdullaev, D., Shukhratovna, D. L., Rasulovna, J. O., Umirzakovich, J. U., & Staroverova, O. V. (2024). Examining Local Item Dependence in a cloze test with the Rasch Model. *International Journal of Language Testing, 14*(1), 75–81. https://doi.org/10.22034/IJLT.2023.409812.1273

2. Distorted factor structure: LID can distort the factor structure of a test or survey. It may create spurious factors or collapse true factors, leading to inaccurate interpretations of the underlying constructs being measured.
3. Reduced discriminant validity: When items are locally dependent, they may measure similar aspects of the construct rather than distinct dimensions. This can reduce the ability to differentiate between different aspects of the construct and compromise discriminant validity.
4. Biased parameter estimates: LID violates the assumption of local independence in many statistical models used for analysis (e.g., factor analysis, item response theory). This violation can result in biased parameter estimates, leading to incorrect conclusions about relationships between variables.
5. Decreased generalizability: LID may limit the generalizability of findings beyond the specific sample or context in which it was observed (Marais & Andrich, 2008). The interrelationships between items may be unique to a particular group, making it difficult to generalize findings to other populations.
6. Increased response burden: If LID is present, respondents may find it more challenging and time-consuming to complete a test or survey due to redundant or overlapping items.
Overall, LID can have significant implications for measurement validity and reliability, factor structure interpretation, and generalizability of findings. Researchers should be aware of this phenomenon and take appropriate steps to address it during test development and data analysis (Yen, 1993).

Local item dependence may be handled with the following approaches:
1. Item analysis: Conducting a thorough item analysis to identify items that exhibit local item dependence. This can be done by examining inter-item correlations or conducting factor analyses.
2. Item modification: If LID is identified, consider modifying or removing problematic items from the test. This could involve rephrasing items, changing response options, or replacing them with new items that do not exhibit local item dependence.
3. Randomize item order: Randomizing the order of items can help reduce the influence of local item dependence. By presenting items in a different order for each participant, any potential dependencies between specific items are less likely to impact overall scores.
4. Control for local dependency statistically: If it is not feasible to modify or remove dependent items, statistical techniques such as Item Response Theory (IRT) models can be used to account for LID in scoring and interpretation of test results.
5. Develop parallel forms: Creating multiple versions of a test with different sets of items can help minimize local item dependence. By using parallel forms, participants are randomly assigned to different versions, reducing the likelihood of dependencies between specific items.
6. Pilot testing: Before administering a test on a large scale, conduct pilot testing with a smaller sample size to identify and address any issues related to LID.
7. Provide clear instructions: Ensure that participants understand how each item should be answered independently and that there is no right or wrong answer based on previous responses.
Handling LID requires careful attention during test development and administration. By implementing these strategies, researchers can minimize its impact on test scores and enhance the validity of their assessments (Zenisky et al., 2002).

Previous research has examined different strategies for modeling LID in tests like the cloze test where items are nested in a passage. For example, many researchers (Alabdallah et al., 2023; Dhyaaldian et al., 2022a; Dhyaaldian et al., 2022b; Hussein, 2022; Syman, 2023) following the tradition of accounting LID in C-Tests (Eckes & Baghaei, 2015; Forthmann et al., 2020) suggest using each stimulus or passage as a unit of analysis and employing polytomous IRT models. However, this modeling strategy only works when there are several passages and becomes obsolete when there is only one passage or prompt. In the current study, we aim to examine and handle LID in a cloze test. We used the first strategy presented above, i.e., employing item analysis to identify locally dependent items. We specifically used the Rasch models' residual correlations to identify LID items.

2. Methodology

2.1. Participants and Setting

Participants of the study were 256 undergraduate students (183 girls and 73 boys) of English as a foreign language at the Department of English at Jizzakh State Pedagogical University, Jizzakh, Uzbekistan. The age range was 18 to 37 with a mean of 23.79 and a standard deviation of 4.09. The

cloze test was administered along with a multiple-choice reading comprehension test as participants' final exam in a reading comprehension course in English.

2.2. Instrument

A standard traditional English language cloze test containing 20 items (gaps) was constructed by the researchers by removing every 7th word in a passage. The first and the last sentences remained intact for lead in. Proper nouns and numbers were not removed. The passage contained 210 words and was about social media influencers. The passage was taken from the British Council graded reading comprehension exercises at the B1 level of the Common European Framework of Reference (Council of Europe, 2020).

3. Results

The 20 cloze items were analyzed with the Rasch model ignoring the potential LID that might exist among the items. WINSTEPS Rasch model computer program version 5.6.0 (Linacre, 2023a) was used to estimate the parameters. Table 1 shows the item parameters, their standard errors, and their infit and outfit values. Infit and outfit values are local fit measures which show the conformity of portions of the data to the Rasch model specifications (Baghaei, et al., 2017). If only the data fit the Rasch model, one can use the raw total scores as indicators of examinees' latent trait locations (Baghaei et al., 2018). As Table 1 shows, Items 10, 14, and 18 have outfit mean square values greater than .130 and do not fit the Rasch model (Linacre, 2023b). The separation reliability of the test with 20 items is .66. The mean and standard deviation of person ability parameters were -.12 and 1.12, respectively.

Table 1.
Item Measures and Fit Values for the 20 Cloze Items

Item	Measure	SE	Infit MNSQ	Outfit MNSQ	Point-Meas. Corr.
1	.26	.14	.95	.89	.46
2	-2.24	.19	1.00	1.09	.27
3	.98	.15	1.04	1.07	.38
4	1.90	.19	.95	1.02	.43
5	4.18	.42	1.00	.71	.30
6	-5.07	.59	.94	.49	.16
7	-2.95	.24	.96	.96	.26
8	-.65	.14	1.04	1.15	.34
9	.96	.15	1.07	1.09	.35
10	1.51	.17	1.03	1.37	.37
11	-3.07	.25	1.00	.78	.25
12	-.91	.14	1.12	1.09	.28
13	2.45	.22	1.09	1.05	.34
14	3.51	.32	.94	3.86	.32
15	-2.51	.21	.95	.77	.32
16	-2.43	.20	.98	.81	.31
17	1.18	.16	.98	.96	.43
18	1.76	.18	.94	1.38	.43
19	.18	.14	.90	.83	.50
20	.96	.15	.99	.97	.42

Residual correlations were evaluated to identify locally dependent items. Residual correlations show the correlations between pairs of items after the influence of the latent trait is factored out. Items are supposed to be correlated only through the latent factor. When the impact of the latent factor is removed, they are expected to be independent (Baghaei & Ravand, 2019; Baghaei & Ravand, 2016). However, if items remain correlated after the impact of the latent trait is accounted for, it means that

they are locally dependent. Residual correlations are in fact the correlations between items after the impact of the latent factor is accounted for and is a standard method for identifying LID (Baghaei & Christensen, 2023; Linacre, 2023b). As Table 2 shows, three items in three pairs have noticeable residual correlations. That is Items 10, 14, and 18 were locally dependent. Interestingly, these are the items which do not fit the Rasch model due to large outfit values.

Table 2.
Item Pairs with High Residual Correlations

Item	Item	Residual Corr.
14	18	.59
10	14	.52
10	18	.34

Principal components analysis (PCA) of standardized residuals is a method of evaluating unidimensionality. That is, when PCA is applied to the residuals, it is expected not to find a factor (because they are expected to be uncorrelated). However, if a factor is extracted from the residuals, it is evidence for multidimensionality. PCA of standardized residuals for the 20 cloze items showed that the strength of the first factor is 2.2 eigenvalues which indicates multidimensionality (Linacre, 2023b; Smith, 2002).

In the next step, the three locally dependent items were removed and the data were reanalyzed. Table 3 shows the item measures and fit values after removing the three locally dependent items. As the table shows, all the items have acceptable infit and outfit mean square values. Examination of residual correlations showed none of the items were locally dependent. PCA of standardized residuals showed that the strength of the first contrast is 1.6 which is clear evidence that the data are unidimensional. The separation reliability of the test with 17 items was 57. The reason for the smaller reliability coefficient is that there are fewer items and there is no local dependence in the items. Local dependence increases internal consistency and spuriously increases test reliability (Zenisky et al., 2002). The mean and the standard deviation of person ability parameters with 17 items were .30 and 1.07, respectively. The correlation between the person parameters from the two analyses was .997.

Table 3.
Item Measures and Fit Values for the Cloze Items after Deleting the Locally Dependent Items

Item	Measure	SE	Infit MNSQ	Outfit MNSQ	Point-Meas. Corr.
1	.69	.14	.95	.92	.45
2	-1.84	.19	1.01	1.07	.29
3	1.42	.15	1.00	1.06	.40
4	2.34	.19	.95	1.10	.40
5	4.46	.43	1.04	1.29	.26
6	-5.04	.71	1.01	.74	.26
7	-2.56	.24	.97	.92	.30
8	-.23	.14	1.03	1.08	.36
9	1.40	.15	1.10	1.12	.33
11	-2.69	.25	1.03	.80	.27
12	-.50	.14	1.12	1.12	.29
13	2.89	.22	1.04	1.02	.34
15	-2.12	.21	.96	.79	.34
16	-2.03	.20	.97	.81	.34
17	1.62	.16	.98	.98	.41
19	.61	.14	.87	.81	.51
20	1.40	.15	.99	.94	.42

4. Discussion and Conclusion

The present study aimed to investigate the presence of local item dependence (LID) in a cloze test and its impact on the reliability and unidimensionality of the measure. LID refers to the situation where responses to certain items in a scale are influenced by responses to other items, leading to a violation of the assumption of item independence and unidimensionality. In this study, residual correlations were used as an indicator of LID.

The analysis revealed that three items in the cloze test exhibited significant residual correlations, indicating the presence of LID. This finding is consistent with previous research highlighting the potential occurrence of LID in cloze tests (Baghaei & Ravand, 2016; Zhang, 2010). The identification of LID is crucial as it can lead to biased estimates and affect the validity and reliability of the scale (Zenisky et al., 2002).

To address this issue, we decided to delete the three items that exhibited significant residual correlations. Deleting these items was deemed necessary to ensure that each item measured a unique aspect of the construct under investigation and that responses were not influenced by other items. However, it is important to note that deleting items can have consequences for scale properties such as unidimensionality and reliability.

After deleting the dependent items, we found that the remaining set of items formed a unidimensional scale as shown by the PCA of residuals. This suggests that removing the locally dependent items successfully eliminated their influence on other items and allowed for a clearer representation of the underlying construct.

However, it is worth noting that deleting dependent items had an impact on scale reliability. Reliability coefficients decreased after removing these items, indicating a reduction in internal consistency. This decrease in reliability can be attributed to the fact that by removing the dependent items, we reduced redundancy within the scale which may have contributed to higher internal consistency estimates; that is, spurious reliability dropped and a better representation of the test accuracy was provided. The decrease in reliability raises concerns about the precision and consistency of the scale scores. Lower reliability implies increased measurement error and reduced power to differentiate between individuals on the construct of interest. Researchers should be cautious when interpreting scores obtained from the revised scale, as they may be less reliable than those obtained prior to item deletion.

The findings of the current study suggest that local item dependence can have a significant impact on the measurement properties of a scale or questionnaire. By identifying and deleting the dependent items, the researchers were able to improve the unidimensionality of the data and enhance item fit. This suggests that local item dependence can introduce noise or bias into the measurement process, and removing such items can lead to a more accurate assessment. Nevertheless, it is essential to consider whether deleting dependent items is an appropriate solution for handling local item dependence. While it may improve certain measurement properties, it may also introduce other issues such as reduced content coverage or construct representation. It is crucial to carefully evaluate whether alternative approaches like statistical modeling techniques (e.g., testlet response theory) could be employed to account for local item dependence while preserving valuable information.

This study's findings may be limited by its specific context or sample characteristics. The impact of local item dependence on measurement properties could vary across different populations or settings. Therefore, caution should be exercised when generalizing these findings to other contexts. This study highlights the importance of addressing local item dependence in measurement instruments but also emphasizes the need for careful consideration of potential trade-offs and alternative approaches when handling such dependencies.

Future research should explore alternative approaches to handling LID that minimize the impact on scale reliability. One potential approach is to use statistical techniques such as testlet item response theory (IRT) models, which can account for LID while preserving scale properties (Bradlow et al., 1999; Wang & Wilson, 2005). Additionally, qualitative methods such as cognitive interviews could provide valuable insights into the nature and sources of LID, aiding in the development of more robust measurement instruments.

In this study, we identified local item dependence using residual correlations and addressed it by deleting dependent items. Although this resulted in a unidimensional structure, it also led to a

decrease in scale reliability. These findings highlight the importance of considering LID in measurement instrument development and suggest that further research is needed to explore alternative strategies for handling LID without compromising scale properties.

References

Alabdallah, Z. A., Ismail, I. A., Mutar, H. K., Mohammed, A., Alghazali, T., Mansoor, M. S., Anber, A. A., Ali, Y. M., Ghaleb, M. H., & Georgievna, G. V. (2023). Analysis of C-Tests with the equidistance and the dispersion models. *International Journal of Language Testing, 13*(Special Issue), 142-148. doi: 10.22034/IJLT.2023.403640.1264

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153–168. DOI: 10.1007/BF02294533

Baghaei, P., & Christensen, K. B. (2023). Modelling local item dependence in C-Tests with the Loglinear Rasch Model. *Language Testing Journal, 40*(3), 820-827. doi: 10.1177/02655322231155109

Baghaei, P., Ravand, H., & Nadri, M. (2019). Is the d2 test of attention Rasch scalable? Analysis with the Rasch Poisson Counts Model. *Perceptual and Motor Skills*, *126*, 70-86. doi: 10.1177/0031512518812183.

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

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

Council of Europe (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing. Available at www.coe.int/lang-cefr.

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*(2), 95-106. doi: 10.22034/IJLT.2022.157127

Dhyaaldian, S. M. A., Kadhim, Q. K., Mutlak, D. A., Neamah, N. R., Kareem, Z. H., Hamad, D. A., Tuama, J. H., & Qasim, M. S. (2022). A comparison of polytomous Rasch models for the analysis of C-Tests. *International Journal of Language Testing, 12*(2), 107-117. doi: 10.22034/IJLT.2022.157128

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

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*(6), 692–705. doi: 10.1177/0734282919889262

Hussein, R. A., Sabit, S. H., Alwan, M. G., Wafqan, H. M., Baqer, A. A., Ali, M. H., Hachim, S. K., Sahi, Z. T., AlSalami, H. T., & Sulaiman, B. F. (2022). Psychometric evaluation of dictations with the Rasch model. *International Journal of Language Testing, 12*(2), 118–127. doi: 10.22034/IJLT.2022.157129

Linacre, J. M. (2023a). *Winsteps® Rasch measurement computer program* (Version 5.6.0). Portland, Oregon: Winsteps.com.

Linacre, J. M. (2023b). *Winsteps® Rasch measurement computer program User's Guide*. Version 5.6.0. Portland, Oregon: Winsteps.com.

Marais, I., & Andrich, D. (2008). Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement, 9*(3), 200–215.

Smith, E. V., Jr. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement, 3*, 205–231.

Syman, K., Alallo, H. M. I., Mohammed, A., Hassan, A. Y., Suleiman, O. W., Ali, Y. M., Ghaleb, M. H., Mikhailovna, K. A., John, E. A., & Georgievna, G. V. (2023). Psychometric modelling of

reading aloud with the Rasch model. *International Journal of Language Testing, 13*(Special Issue), 62-68. doi: 10.22034/IJLT.2023.386577.1235

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29*(2), 126–149. doi: 10.1177/0146621604271053

Yen, W. (1984). Effects of local item dependence on the fit and equating performance of the three parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125–145. doi: 10.1177/014662168400800201

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187–213. doi: 10.1111/j.1745-3984.1993.tb00423.x

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*, 291–309. doi: 10.1111/j.1745-3984.2002.tb01144.x

Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. *Language Testing*, *27*, 119–140. doi: 10.1177/02655322093473